The Electromagnetic Radiation Spectrum

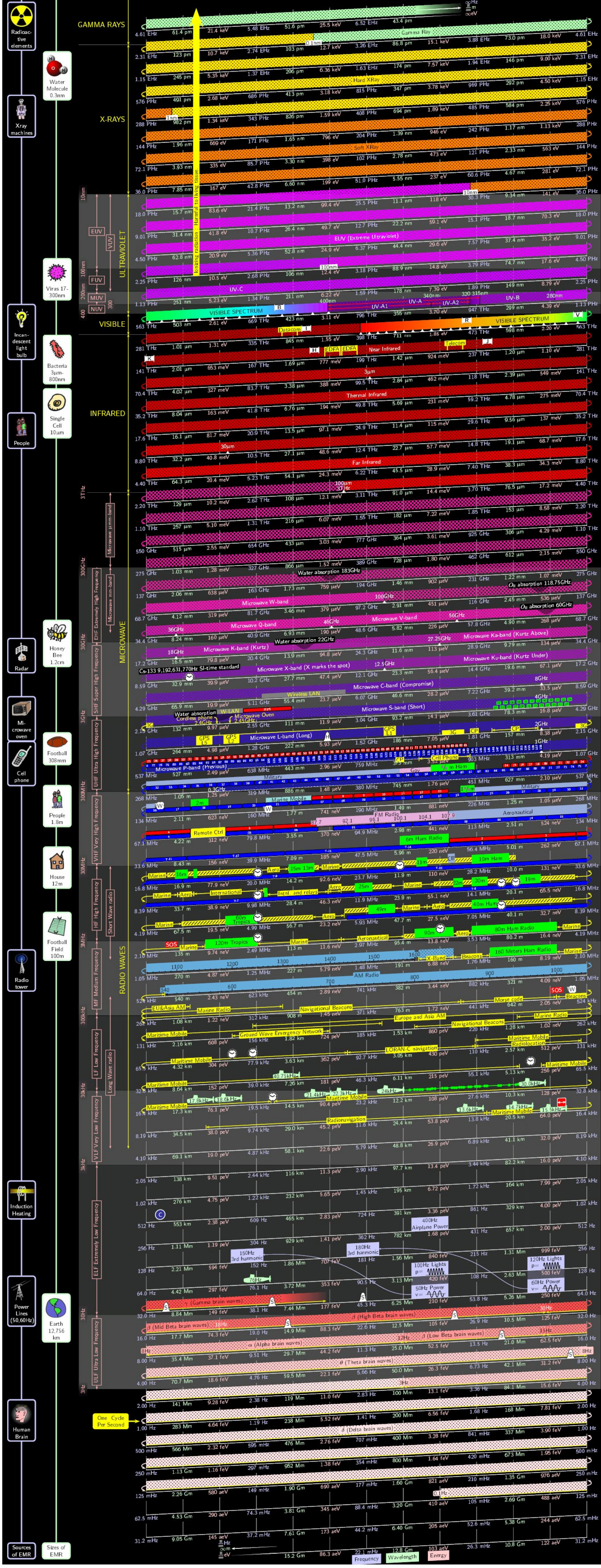

How to read this chart

- This chart is organized in octaves (frequency doubling/halving) starting at 1Hz and going higher (2,4,8, etc) and lower (1/2, 1/4, etc). The octave is a natural way to represent frequency.
- Frequency increases on the vertical scale in the upward direction.
- The horizontal bars wrap around from far right to far left as the frequency increases upwards.
- There is no limit to either end of this chart, however, due to limited space, only the "known" items have been shown here. A frequency of 0Hz is the lowest possible frequency but the method of depicting octaves used here does not allow for ever reaching 0Hz, only approaching it. Also, by the definition of frequency (Cycles per second), there is no such thing as negative frequency.
- Values on the chart have been labelled with the following colours: Frequency measured in Hertz, Wavelength measured in meters, Energy measured in electronVolts.

Ultraviolet Light

- Ultraviolet light is beyond the range of human vision.
- Physicists have divided ultraviolet light ranges into Vacuum Ultraviolet (VUV), Extreme Ultraviolet (EUV), Far Ultraviolet (FUV), Medium Ultraviolet (MUV), and Near Ultraviolet (NUV).
- UV-A, UV-B and UV-C were introduced in the 1930's by the Commission Internationale de l'Éclairage (CIE, International Commission on Illumination) for photobiological spectral bands.
- Short-term UV-A exposure causes sun-tanning which helps to protect against sunburn. Exposure to UV-B is beneficial to humans by helping the skin produce vitamin D. Excessive UV exposure causes skin damage. UV-C is harmful to humans but is used as a germicide.
- The CIE originally divided UVA and UVB at 315nm, later some photo-dermatologists divided it at 320nm.
- UVA is subdivided into UVA1 and UVA2 for DNA altering effects at 340nm.
- The sun produces a wide range of frequencies including all the ultraviolet light, however, UVB is partially filtered by the ozone layer and UVC is totally filtered out by the earth's atmosphere.
- A bumblebee can see light in the UVA range which helps them identify certain flowers.

Emission and Absorption

- As EMR passes through elements, certain wavelength bands get absorbed and some new ones get emitted. This absorption and emission produces characteristic spectral lines for each element which are useful in determining the makeup of distant stars. These lines are used to prove the red-shift amount of distant stars.
- When a photon hits an atom it may be absorbed if the energy is just right. The energy level of the electron is raised – essentially holding the radiation. A new photon of specific wavelength is created when the energy is released. The jump in energy is a discrete step and many possible levels of energy exist in an atom.
- Johann Balmer created this formula defining the photon emission wavelength (λ); where m is the initial electron energy level and n is the final electron energy level:

$$\lambda = 364.56nm\left(\frac{m^2}{m^2 - n^2}\right)$$

- Much of the interstellar matter is made of the simplest atom hydrogen. The hydrogen visible-spectrum emission and absorption lines are shown below:

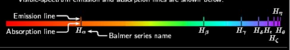

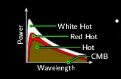

- Max Planck determined the relationship between the temperature of an object and its radiation profile; where R_λ is the radiation power, λ is the wavelength, T is the temperature:

$$R_\lambda = \frac{37418}{\lambda^5 \epsilon^{\left(\frac{14388}{\lambda T} - 1\right)}}$$

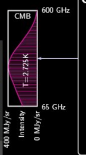

Cosmic Microwave Background Radiation

- CMB radiation is the leftover heat from the hot early universe, which last scattered about 400,000 years after the Big Bang.
- CMB permeates the entire universe at a temperature of 2.725 ± 0.001K.
- CMB was predicted in the 1940's by Ralph Alpher, George Gamow and Robert Herman.
- Arno Penzias and Robert Wilson accidentally discovered CMB while working for Bell Telephone Laboratories in 1965.
- The intensity is measured in Mega Jansky (Jy) per steradian. $1Jy = 10^{-26}W/m^2/Hz$

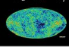

Close examination of slight CMB intensity variations in different parts of the sky help cosmologists study the formation of galaxies.

WMAP photo by NASA

Television

- Television is transmitted in the VHF and UHF ranges (30MHz - 3GHz).
- TV channels transmitted over the air are shown as TV.
- TV channels transmitted through cable (CATV) are shown as TV. CATV channels starting with "T-" are channels fed back to the cable TV station (like news feeds).
- Air and cable TV stations are broadcast with the separate video, colour, and audio frequency carriers grouped together in a channel band as follows:

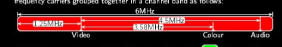

- Satellite channels broadcast in the C-Band are depicted as TV. These stations are broadcast in alternating polarities (Ex. Ch 1 is vertical and 2 is horizontal and vice versa on neighbouring satellites).
- The 15.7 kHz horizontal sweep signal produced by a TV can be heard by some young people. This common contaminant signal to VLF spectra listening is depicted as ↔.

Radio Bands

- The radio spectrum (ELF to EHF) is populated by many more items than can be shown on this chart, only a small sampling of bands used around the world have been shown.
- Communication using EMR is done using either:
  - Amplitude Modulation (AM)

    OR
  - Frequency Modulation (FM)

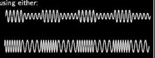

- Each country has its own rules and regulations for allotting bands in this region. For more information, look up the radio communications authority in your area (Ex. FCC in the US, DOC in Canada).
- Not all references agree on the ULF band range, the HAARP range is used here.
- **RA**dio **D**etecting **A**nd **R**anging (RADAR) uses EMR in the microwave range to detect the distance and speed of objects.
- **C**itizens **B**and Radio (CB) contains 40 stations between 26.965-27.405MHz.
- Schumann resonance is produced in the cavity between the Earth and the ionosphere. The resonant peaks are depicted as S.
- Hydrogen gas emits radio band EMR at 21cm H.
- Some individual frequencies are represented as icons:

xxHz	Submarine communications
🕒	Time and frequency standards
xxm	Ham radio and international meter bands
▨	Miscellaneous short wave radio
W	Weather stations
CP	Cellular and PCS Phones (including; FDMA, TDMA, CDMA ranges)

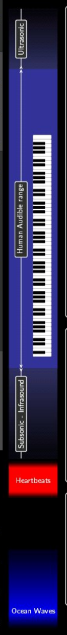

Sound

- Although sound, ocean waves, and heartbeats are not electromagnetic, they are included on this chart as a frequency reference. Other properties of electromagnetic waves are different from sound waves.
- Sound waves are caused by an oscillating compression of molecules. Sound cannot travel in a vacuum such as outer space.
- The speed of sound in air is 1240kph (770mph).
- Humans can only hear sound between ≈20Hz to ≈20kHz.
- Infrasound (below 20Hz) can be sensed by internal organs and touch. Frequencies in the 0.2Hz range are often the cause of motion sickness.
- Bats can hear sound up to ≈50kHz.
- The 88 piano keys of the Equal Temperament scale are accurately located on the frequency chart.
- Over the ages people have striven to divide the continuous audio frequency spectrum into individual musical notes that have harmonious relationships. Microtonal musicians study various scales. One recent count lists 4700 different musical scales.
- Middle C is depicted on the chart as C

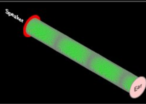

This image depicts air being compressed as sound waves in a tube from a speaker and then travelling through the tube towards the ear.

Gravity Waves

- Gravity is the mysterious force that holds large objects together and binds our planets, stars and galaxies together. Many people have unsuccessfully theorized about the details of gravity and its relationship to other forces. There have been no links between gravity waves and electromagnetic radiation.
- Gravity is theorized to warp space and time. In fact, gravity is responsible for bending light as observed by the gravity-lens example of distant galaxies.
- "Gravity waves" would appear as ripples in space-time formed by large objects moving through space that might possibly be detected in the future by very sensitive instruments.
- The speed that gravity propagates through space has been theorized to be the same as the speed of light.

Brain Waves

- By connecting electrodes from the human head to an electroencephalograph (EEG), it is possible to measure very small cyclic electrical signals.
- There has been much study on this topic, but like all effects on humans, the science is not as exact as the science of materials.
- Generally, lower brain wave frequencies relate to sleep, and the higher frequencies relate to alertness.
- Devices have been made for measuring and stimulating brain waves to achieve a desired state.

Electromagnetic Radiation (EMR)

- EMR is emitted in discrete units called photons but has properties of waves as seen by the images below. EMR can be created by the oscillation or acceleration of electrical charge or magnetic field. EMR travels through space at the speed of light (2.997 924 58 ×10^8 m/s). EMR consists of an oscillating electrical and magnetic field which are at right angles to eachother and spaced at a particular wavelength. There is some controversy about the phase relationship between the electrical and magnetic fields of EMR, one of the theoretical representations is shown here:

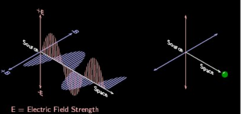

E = Electric Field Strength
B = Magnetic Field Strength
Wave Nature

Particle Nature

- The particle nature of EMR is exhibited when a solar cell emits individual electrons when struck with very dim light.
- The wave nature of EMR is demonstrated by the famous double slit experiment that shows cancelling and addition of waves.
- Much of the EMR properties are based on theories since we can only see the effects of EMR and not the actual photon or wave itself.
- Albert Einstein theorized that the speed of light is the fastest that anything can travel. So far he has not been proven wrong.
- EMR can have its wavelength changed if the source is receding or approaching as in the red-shift example of distant galaxies and stars that are moving away from us at very high speeds. The emitted spectral light from these receding bodies appears more red than it would be if the object was not moving away from us.
- We only have full electronic control over frequencies in the microwave range and lower. Higher frequencies must be created by waiting for the energy to be released from elements as photons. We can either pump energy into the elements (ex. heating a rock with visible EMR and letting it release infrared EMR) or let it naturally escape (ex. uranium decay).
- We can only see the visible spectrum. All other bands of the spectrum are depicted as hatched colours ▒.

Système International d'unité prefixes (SI unit prefixes)

Symbol	Name	Exp.	Multiplier
Y	yotta	10^{24}	1,000,000,000,000,000,000,000,000
Z	zetta	10^{21}	1,000,000,000,000,000,000,000
E	exa	10^{18}	1,000,000,000,000,000,000
P	peta	10^{15}	1,000,000,000,000,000
T	tera	10^{12}	1,000,000,000,000
G	giga	10^{9}	1,000,000,000
M	mega	10^{6}	1,000,000
k	kilo	10^{3}	1,000
		10^{0}	1
m	milli	10^{-3}	0.001
µ	micro	10^{-6}	0.000 001
n	nano	10^{-9}	0.000 000 001
p	pico	10^{-12}	0.000 000 000 001
f	femto	10^{-15}	0.000 000 000 000 001
a	atto	10^{-18}	0.000 000 000 000 000 001
z	zepto	10^{-21}	0.000 000 000 000 000 000 001
y	yocto	10^{-24}	0.000 000 000 000 000 000 000 001

Measurements on this chart

Symbol	Name	Value
c	Speed of Light	2.997 924 58 ×10^8 m/s
h	Planck's Constant	6.626 1 ×10^{-34} J·s
$\hbar$	Planck's Constant (freq)	1.054 592 ×10^{-34} J·s
f	Frequency (cycles / second)	Hz
λ	Wavelength (meters)	m
E	Energy (Joules)	J

Conversions

$E = h \cdot f$
$\lambda = \frac{c}{f}$
$1Å = 0.1nm$
$1nm = 10Å$
$1Joule = 6.24 \times 10^{18} eV$

Gamma Rays

- Gamma radiation is the highest energy radiation (up to ≈ 10^{20} eV) that has been measured. At this energy, the radiation could be from gamma-rays, protons, electrons, or something else.
- Alpha, beta, and delta radiation are not electromagnetic but are actually parts of the atom being released from a radioactive atom. In some cases this can cause gamma radiation. These are not to be confused with brain waves of similar names.

Visible Spectrum

- The range of EMR visible to humans is also called "Light". The visible spectrum also closely resembles the range of EMR that filters through our atmosphere from the sun.
- Other creatures see different ranges of visible light, for example bumble-bees can see ultraviolet light and dogs have a different response to colours than do humans.
- The sky is blue because our atmosphere scatters light and the shorter wavelength blue gets scattered the most. It appears that the entire sky is illuminated by a blue light but in fact that light is scattered from the sun. The longer wavelengths like red and orange move straight through the atmosphere which makes the sun look like a bright white ball containing all the colours of the visible spectrum.
- Interestingly, the visible spectrum covers approximately one octave.
- Astronomers use filters to capture specific wavelengths and reject unwanted wavelengths, the major astronomical (visual) filter bands are depicted as X

Infrared Radiation

- Infrared radiation (IR) is sensed by humans as heat and is below the range of human vision. Humans (and anything at room temperature) are emitters of IR.
- IR remote control signals are invisible to the human eye but can be detected by most camcorders.
- Night vision scopes/goggles use a special camera that senses IR and converts the image to visible light. Some IR cameras employ an IR lamp to help illuminate the view.
- IR LASERs are used for burning objects.
- A demonstration of IR is to hold a metal bowl in front of your face. The IR emitted by your body will be reflected back using the parabolic shape of the bowl and you will feel the heat.
- Fiber-optic based infrared communication signals are sometimes amplified with Erbium-Doped Fiber Amplifiers EDFA

LASER

- LASER is an acronym for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation.
- A LASER is a device that produces monochromatic EMR of high intensity.
- With proper equipment, any EMR can be made to operate like a LASER. For example, microwaves are used to create a MASER.

Polarization

- As a photon (light particle) travels through space, its axis of electrical and magnetic fluctuations does not rotate. Therefore, each photon has a fixed linear polarity of somewhere between 0° to 360°. Light can also be circularly and elliptically polarized.
- Some crystals can cause the photon to rotate its polarization.
- Receivers that expect polarized photons will not accept photons that are in other polarities. (ex. satellite dish receivers have horizontal and vertical polarity positions).
- A polarized filter (like Polaroid™ sunglasses) can be used to demonstrate polarized light. One filter will only let photons that have one polarity through. Two overlapping filters at right angles will almost totally block the light that exits, however, a third filter inserted between the first two at a 45° angle will rotate the polarized light and allow some light to come out the end of all three filters.
- Light that reflects off an electrical insulator becomes polarized. Conductive reflectors do not polarize light.
- Perhaps the most reliably polarized light is a rainbow.
- Moonlight is also slightly polarized. You can test this by viewing the moonlight through a Polaroid™ sunglass lens, then rotate that lens, the moonlight will dim and brighten slightly.

Refraction

- Refraction of EMR is dependent on wavelength as can be seen by the prism example below.

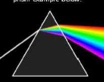

By using a glass prism, white light can be spread by refraction into a spectrum of its composite colours. All wavelengths of EMR can be refracted by using the proper materials.

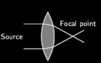

Convex lenses make objects appear closer and are used to correct far-sitedness.

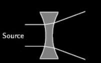

Concave lenses make objects appear farther away and are used to correct near-sitedness.

Photo by STScI

Heavy objects like dense galaxies and large planets cause light to bend due to gravitational lensing.

Reflection

- Reflection of EMR is dependent on wavelength as demonstrated when visible light and radio waves bounce off objects that X-Rays would pass through. Microwaves, which have a large wavelength compared to visible light, will bounce off metal mesh in a microwave oven whereas visible light will pass through.

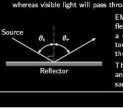

EMR of any wavelength can be reflected, however, the reflectivity of a material depends on many factors including the wavelength of the incident beam.

The angle of incidence (θ_i) and angle of reflection (θ_r) are the same.

© unihedron.com 2004-12-12

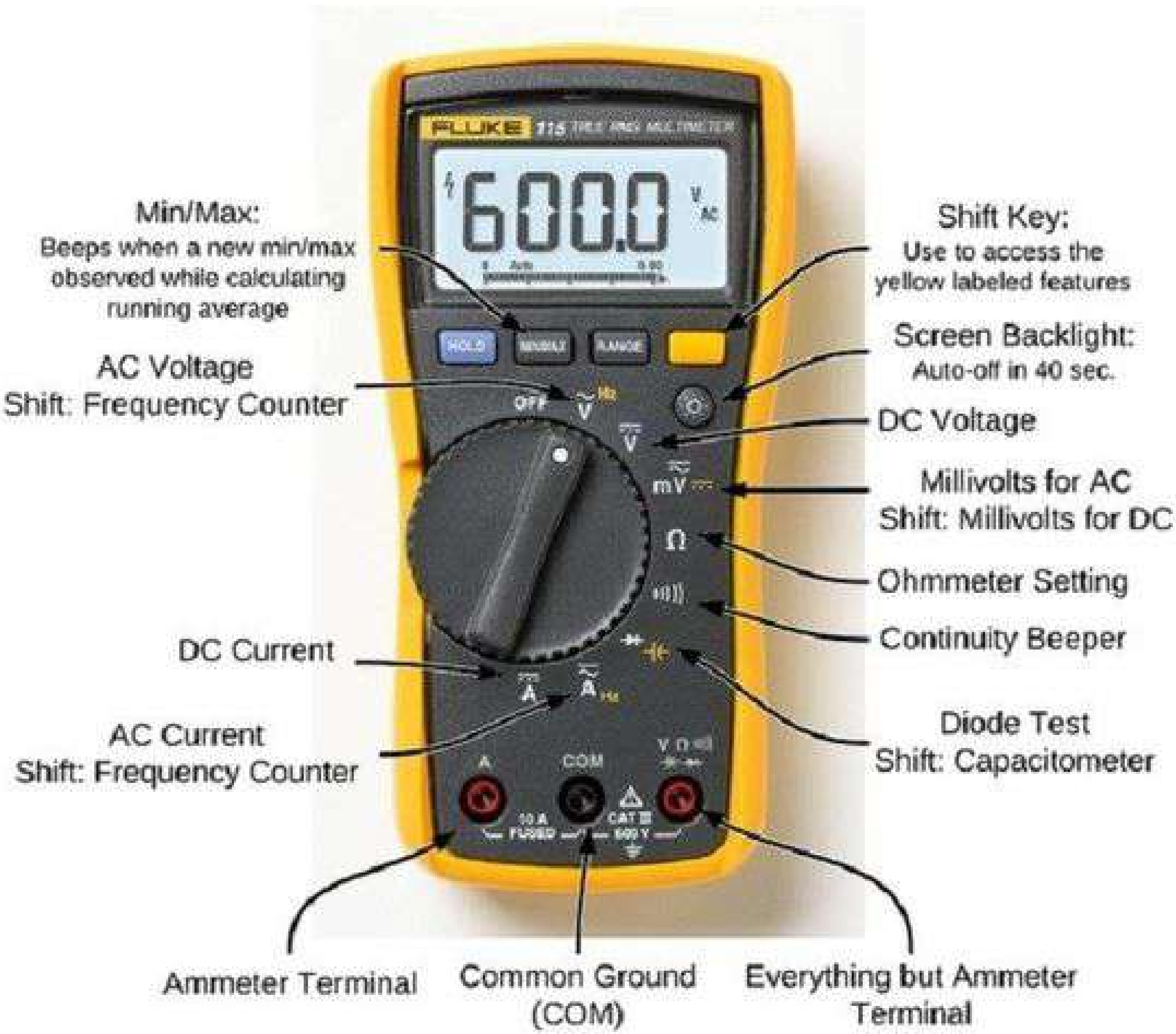
FLUKE 115 TRUE RMS MULTIMETER
600.0
Min/Max:
Beeps when a new min/max observed while calculating running average
Shift Key:
Use to access the yellow labeled features
Screen Backlight:
Auto-off in 40 sec.
AC Voltage
Shift: Frequency Counter
DC Voltage
Millivolts for AC
Shift: Millivolts for DC
Ohmmeter Setting
Continuity Beeper
DC Current
Diode Test
Shift: Capacitometer
AC Current
Shift: Frequency Counter
HOLD
MIN MAX
RANGE
OFF
A
COM
10 A FUSED
CAT III 600 V
Ammeter Terminal
Common Ground (COM)
Everything but Ammeter Terminal

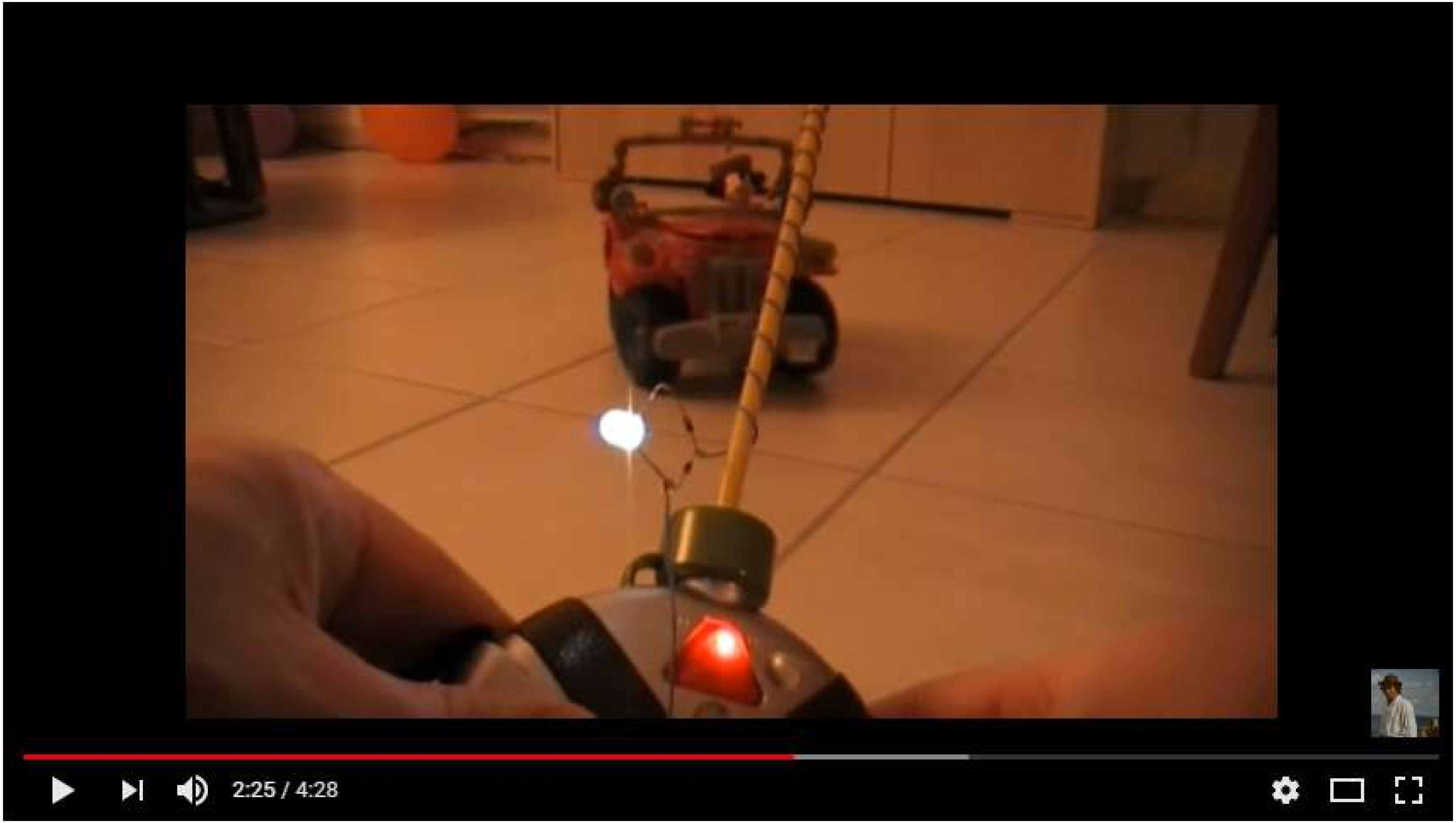

electricity from radiowaves 3

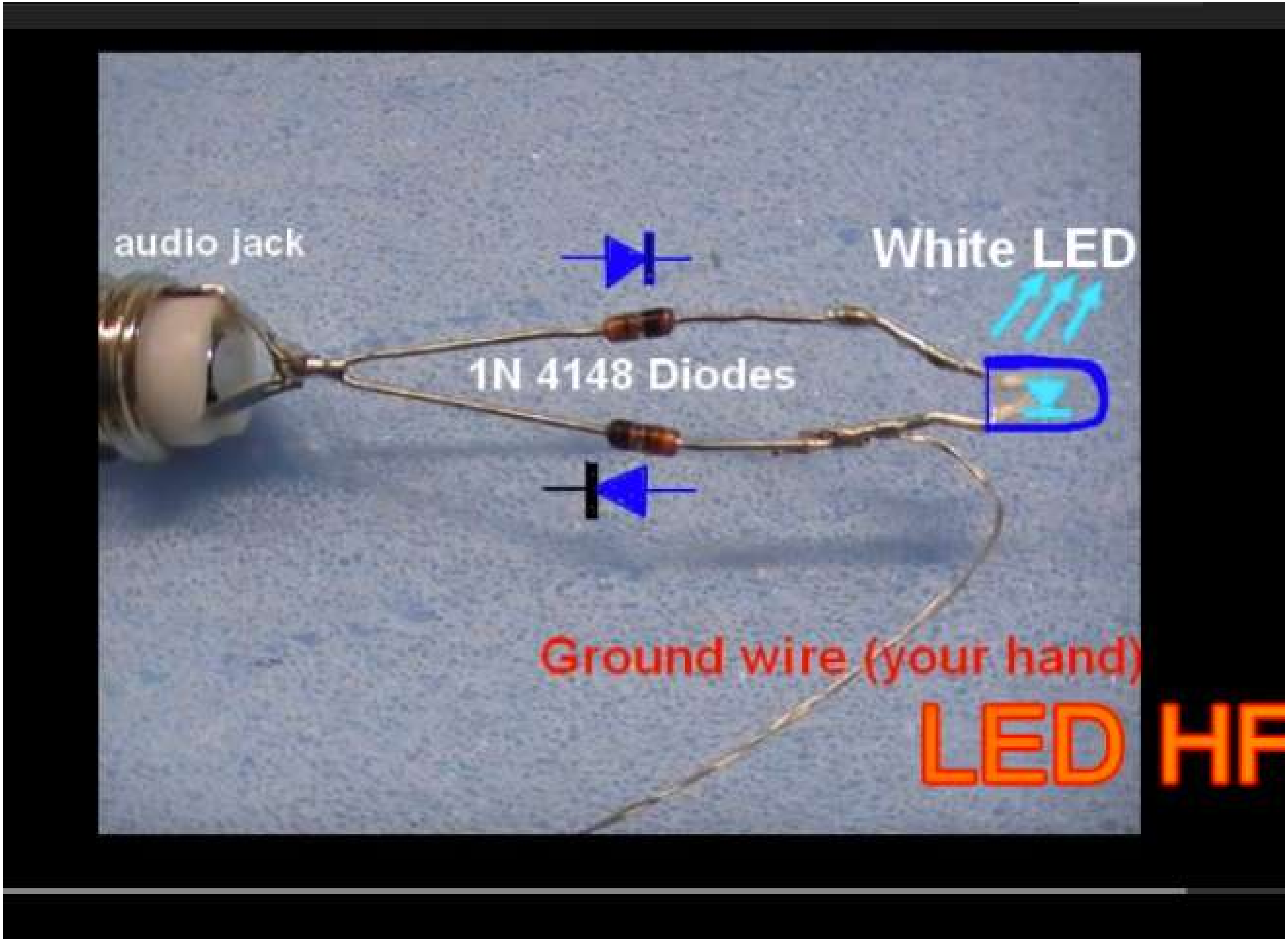
audio jack
White LED
1N 4148 Diodes
Ground wire (your hand)
LED HF

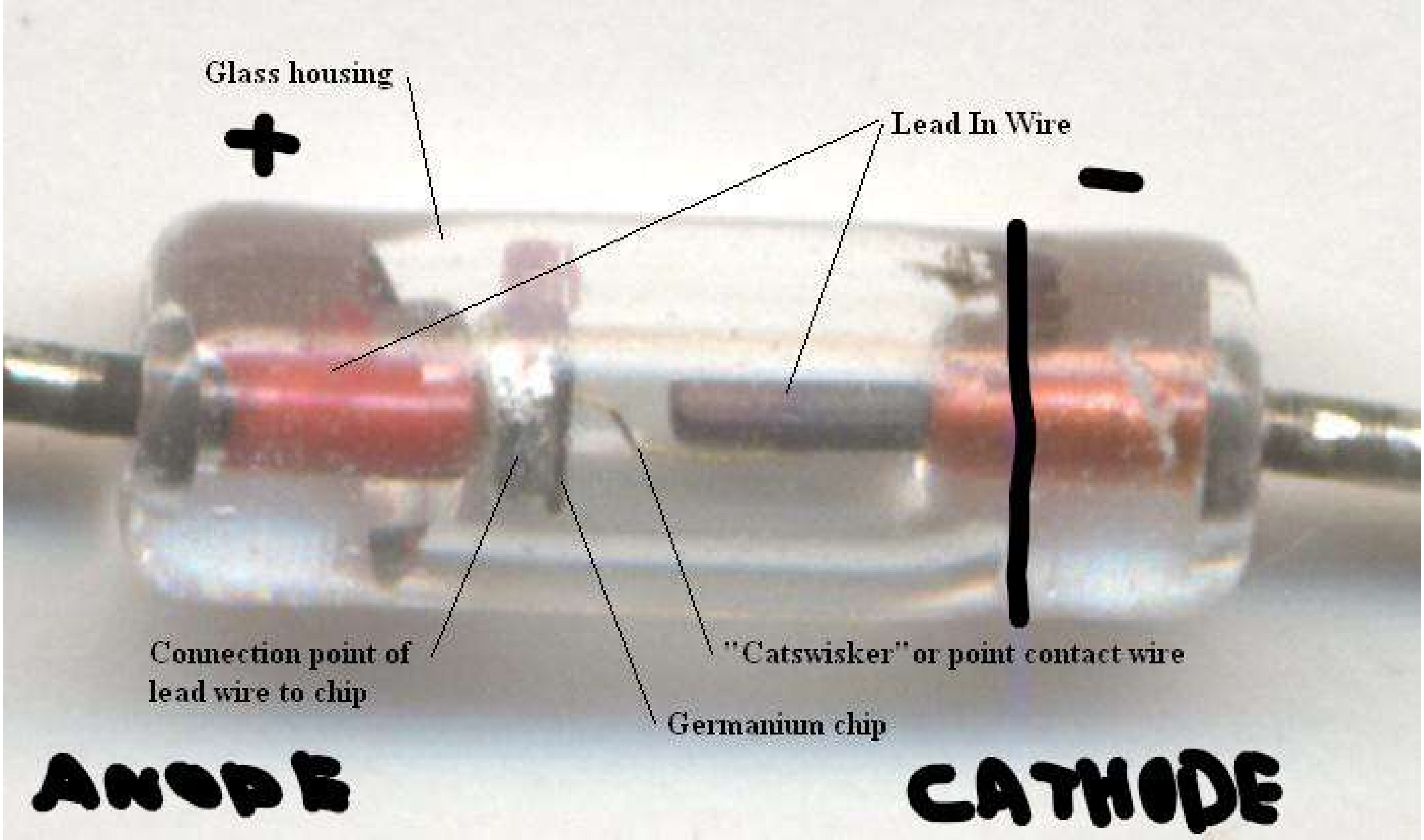
Glass housing
+
Lead In Wire
-
Connection point of lead wire to chip
"Catswisker" or point contact wire
Germanium chip
ANODE
CATHODE

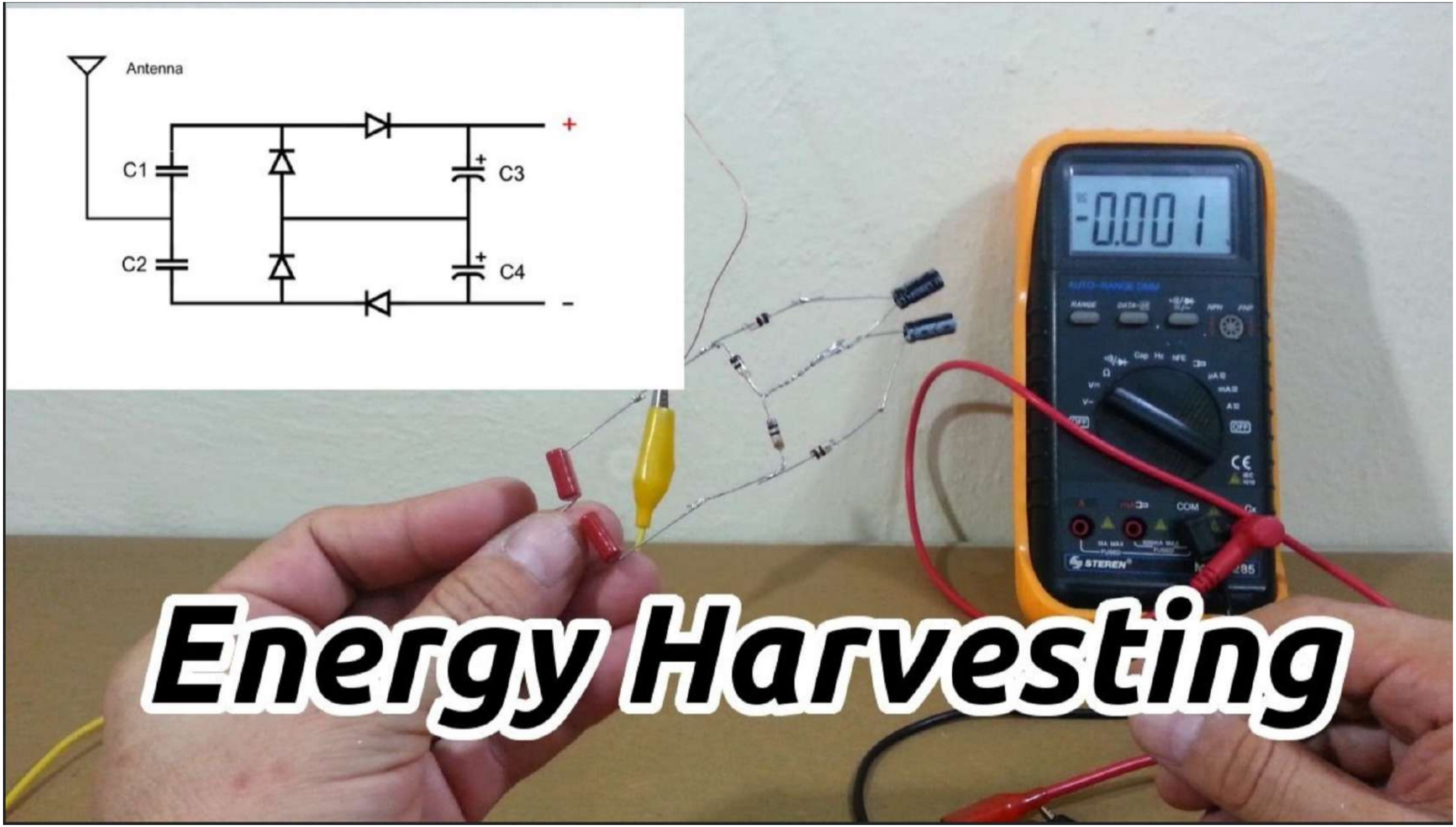
Antenna
C1
C2
C3
C4
+
-
-0.001
STEREN
COM
Energy Harvesting

feluri (fig. III.9), schemele fiind echivalente cu reprezentarea de bază din fig. III.8. Diodele sînt legate în serie, în formă de patrulater, două avînd comun anodul (punctul *2*), iar celelalte două catodul (punctul *4*). Tensiunea alternativă de intrare se aplică pe diagonala *1—3*, iar consumatorul se conectează pe diagonala *2—4*.

Pentru a urmări funcționarea punții, să presupunem că prima alternanță sosită în nodul *1* este pozitivă. Ea blochează dioda D_2 și o deschide pe D_1, debitînd prin R_S un curent I_1 (săgețile pline), care se întoarce la

tice sau cu parametri cît mai apro-

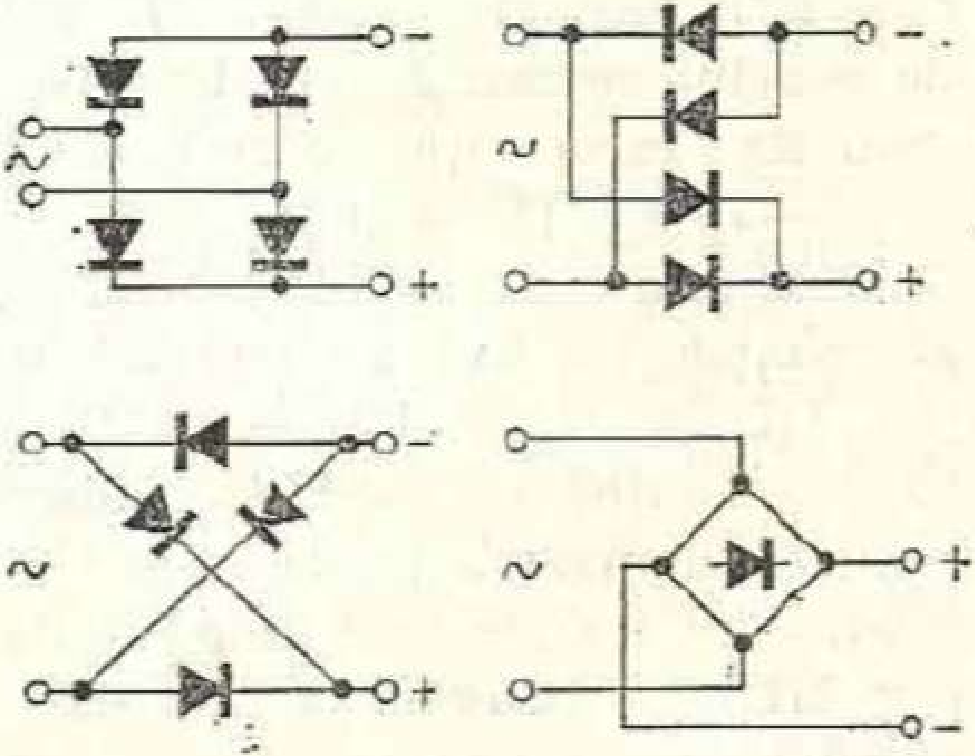

III.9. Puntea redresoare în diferite reprezentări.

circuits used on LC "RadioWaves"

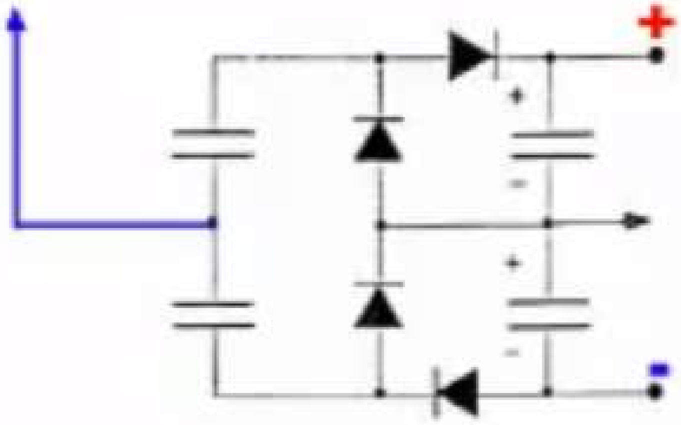

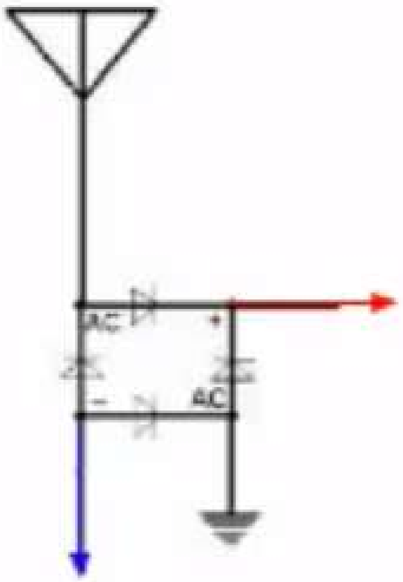

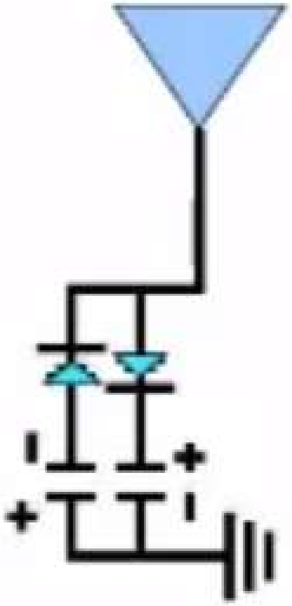

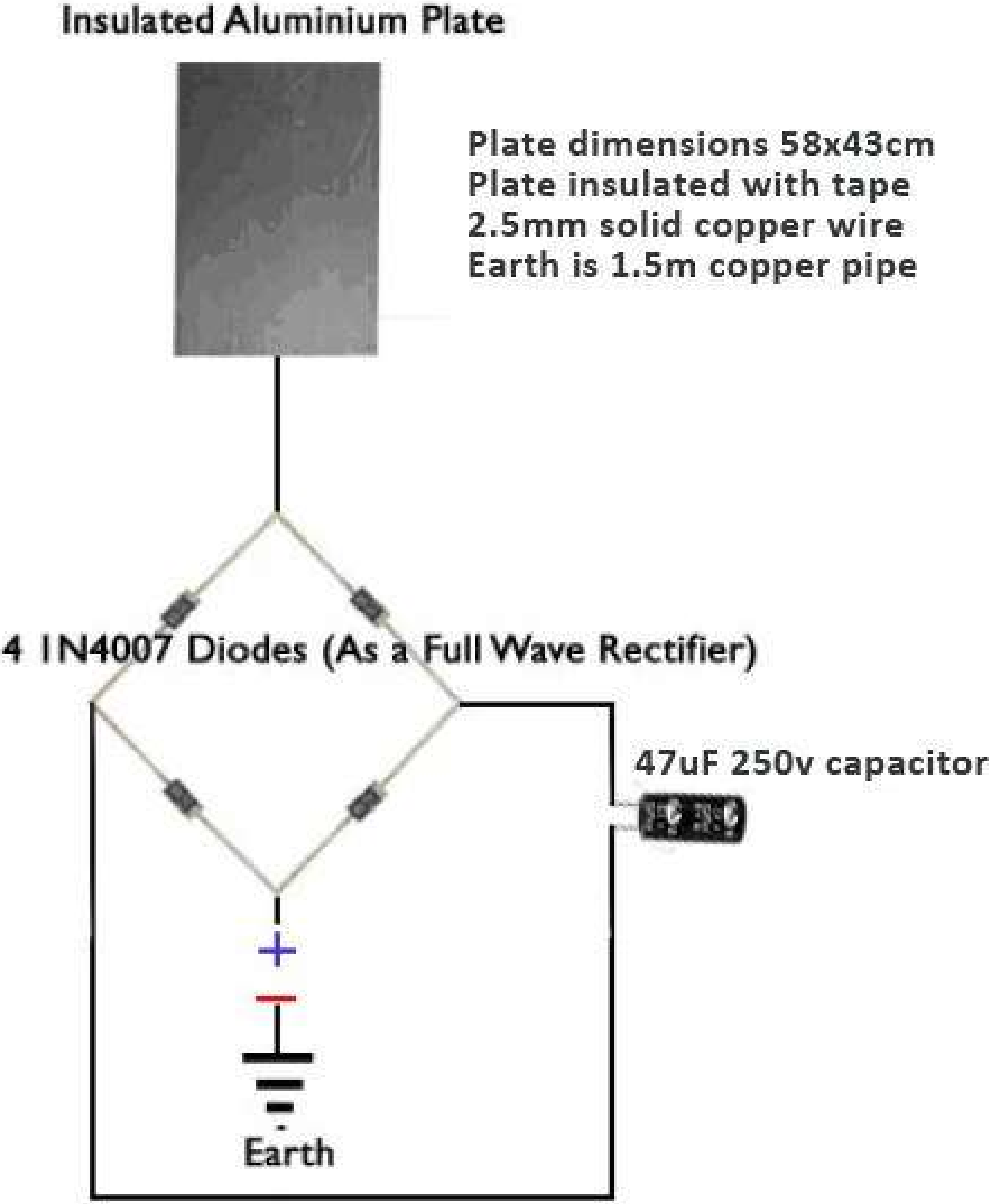
Insulated Aluminium Plate
4 1N4007 Diodes (As a Full Wave Rectifier)
47uF 250v capacitor
+
−
Earth

Plate dimensions 58x43cm
Plate insulated with tape
2.5mm solid copper wire
Earth is 1.5m copper pipe

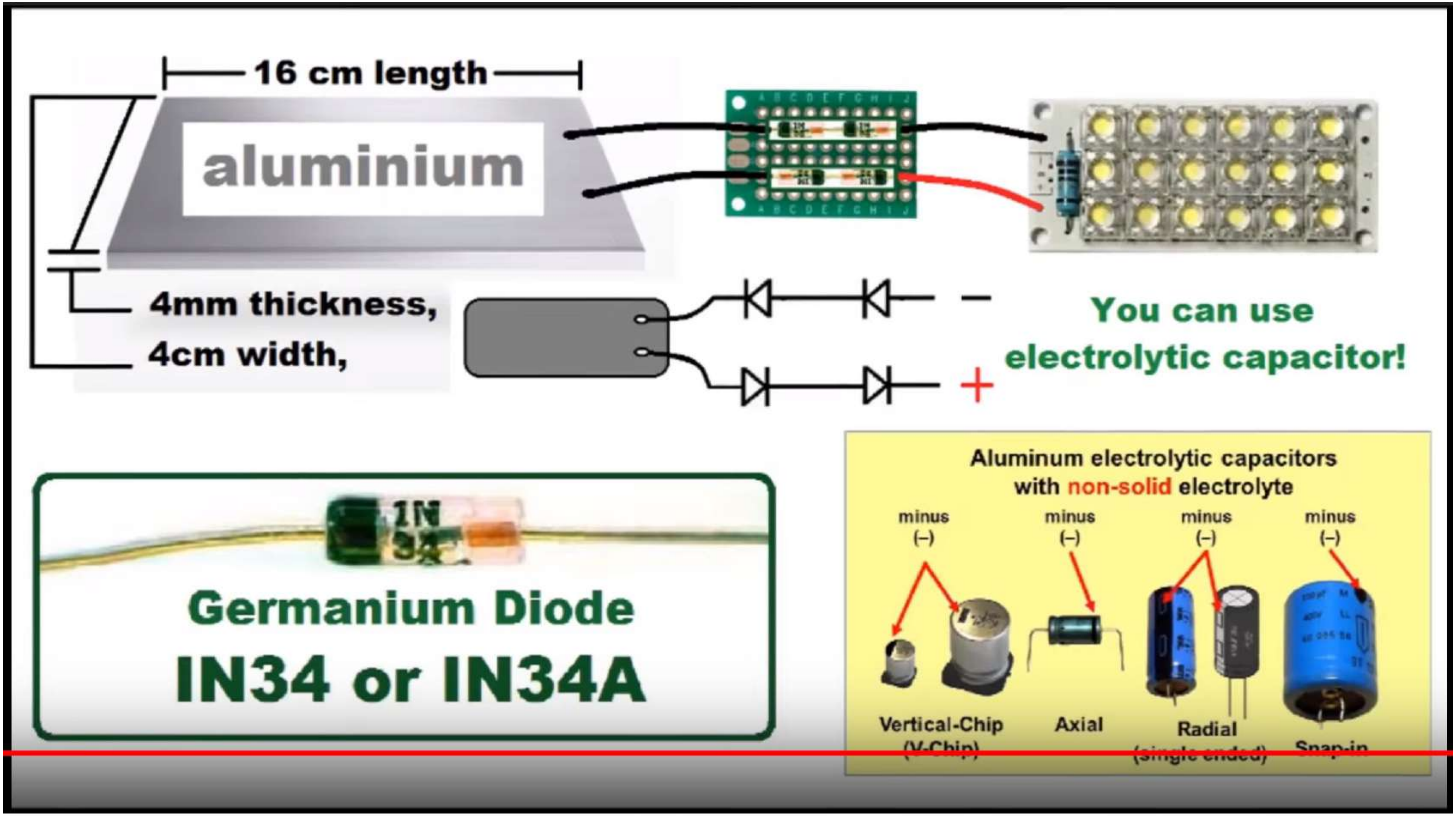

16 cm length
aluminium
4mm thickness,
4cm width,
−
+

You can use electrolytic capacitor!

Germanium Diode

IN34 or IN34A

Aluminum electrolytic capacitors with non-solid electrolyte

minus (–)
minus (–)
minus (–)
minus (–)
Vertical-Chip (V-Chip)
Axial
Radial (single ended)
Snap-in

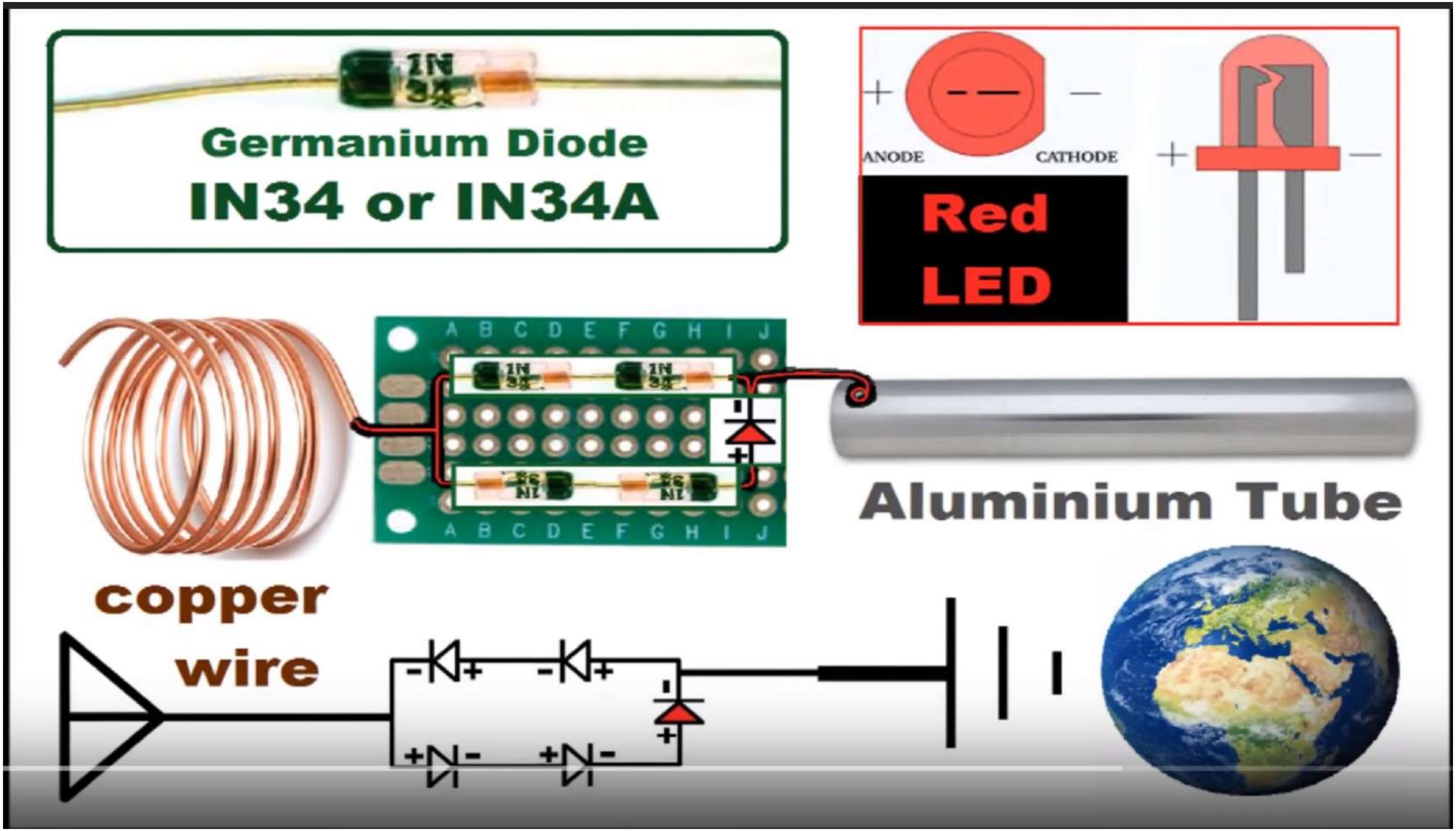

Germanium Diode

IN34 or IN34A

Red LED

Aluminium Tube

copper wire

1N
34
+
−
ANODE
CATHODE
A B C D E F G H I J
+
−

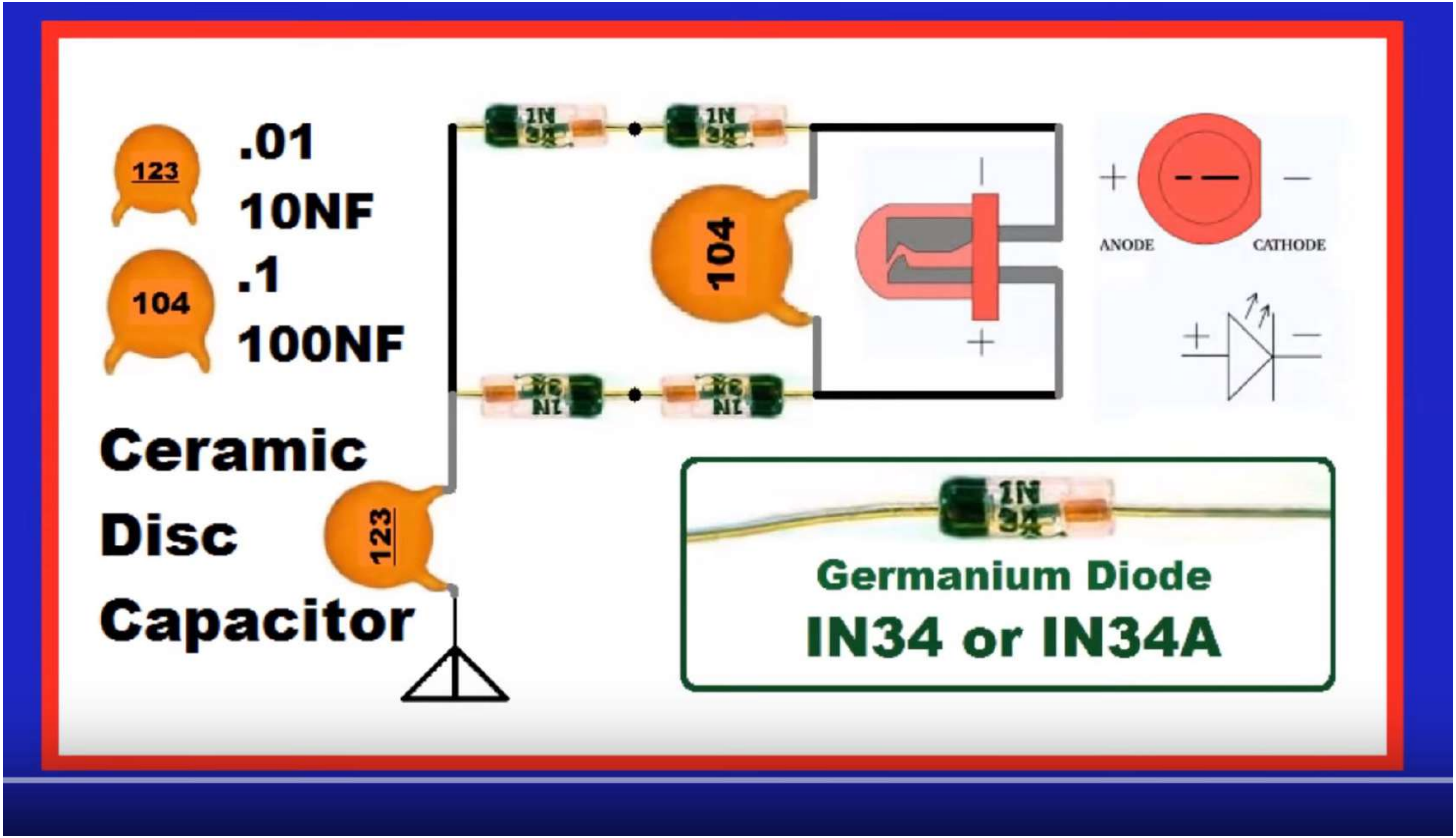
123
104
1N 34
ANODE
CATHODE
+
−

.01
10NF

.1
100NF

Ceramic Disc Capacitor

Germanium Diode

IN34 or IN34A

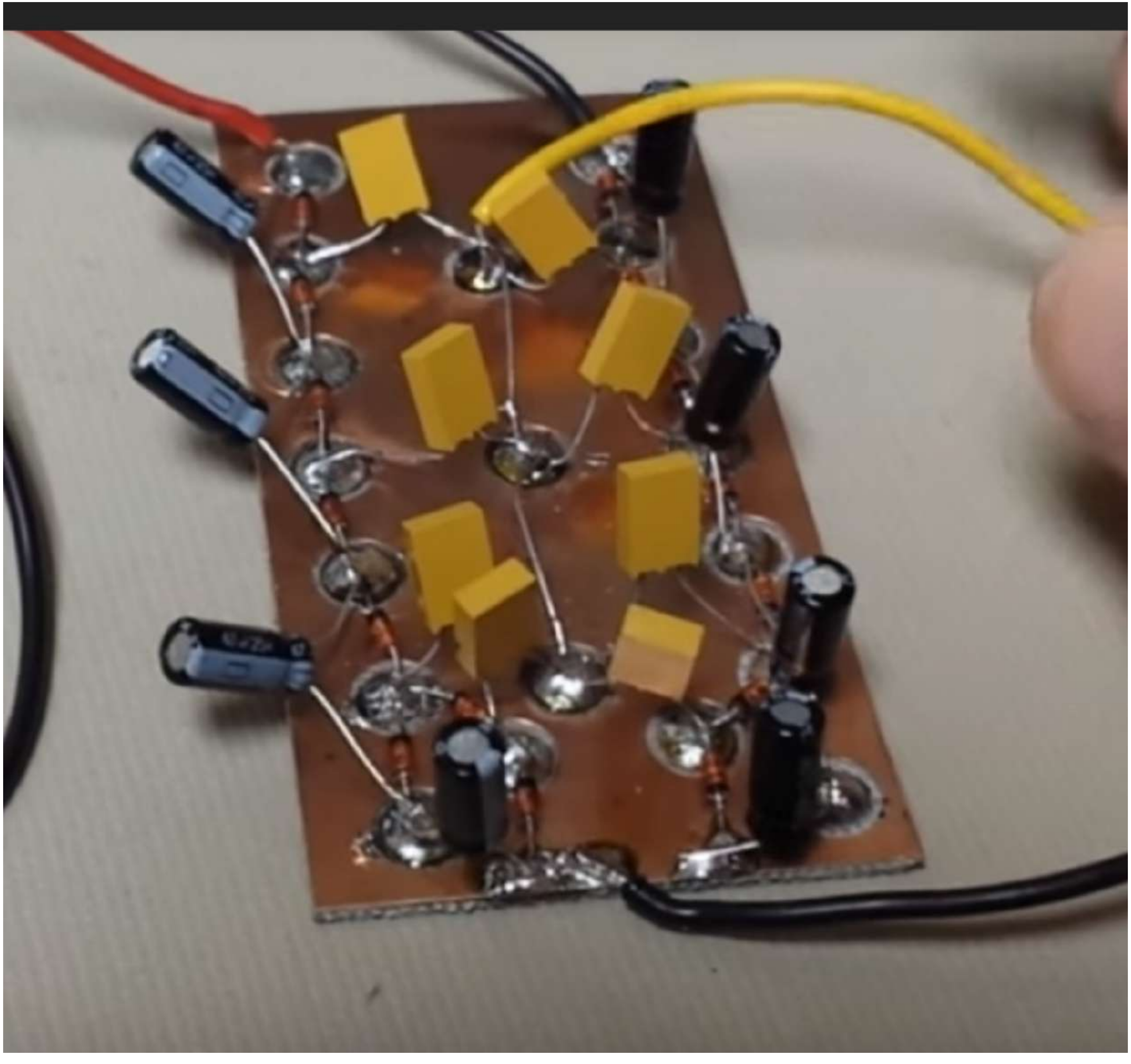

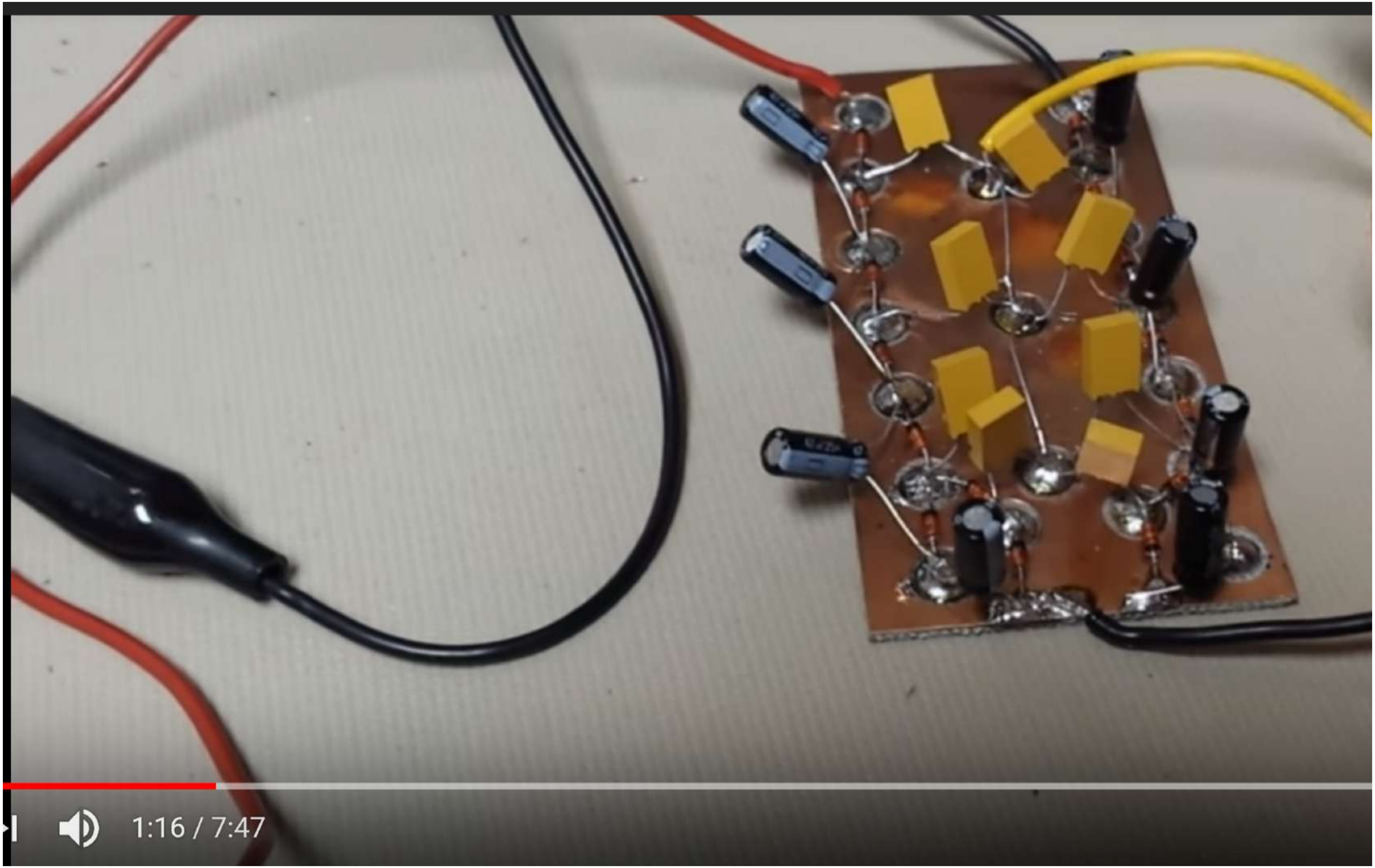
1:16 / 7:47

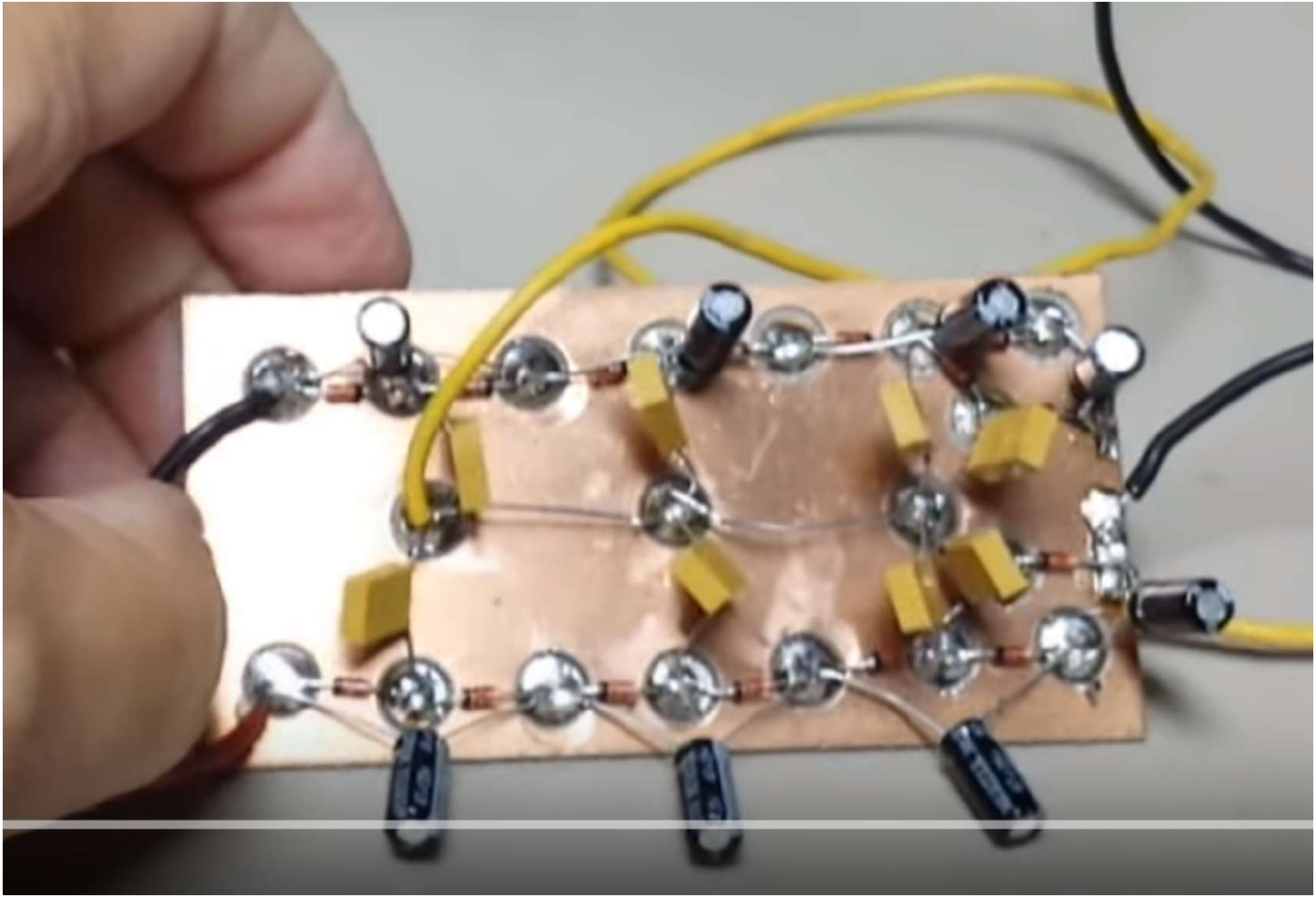

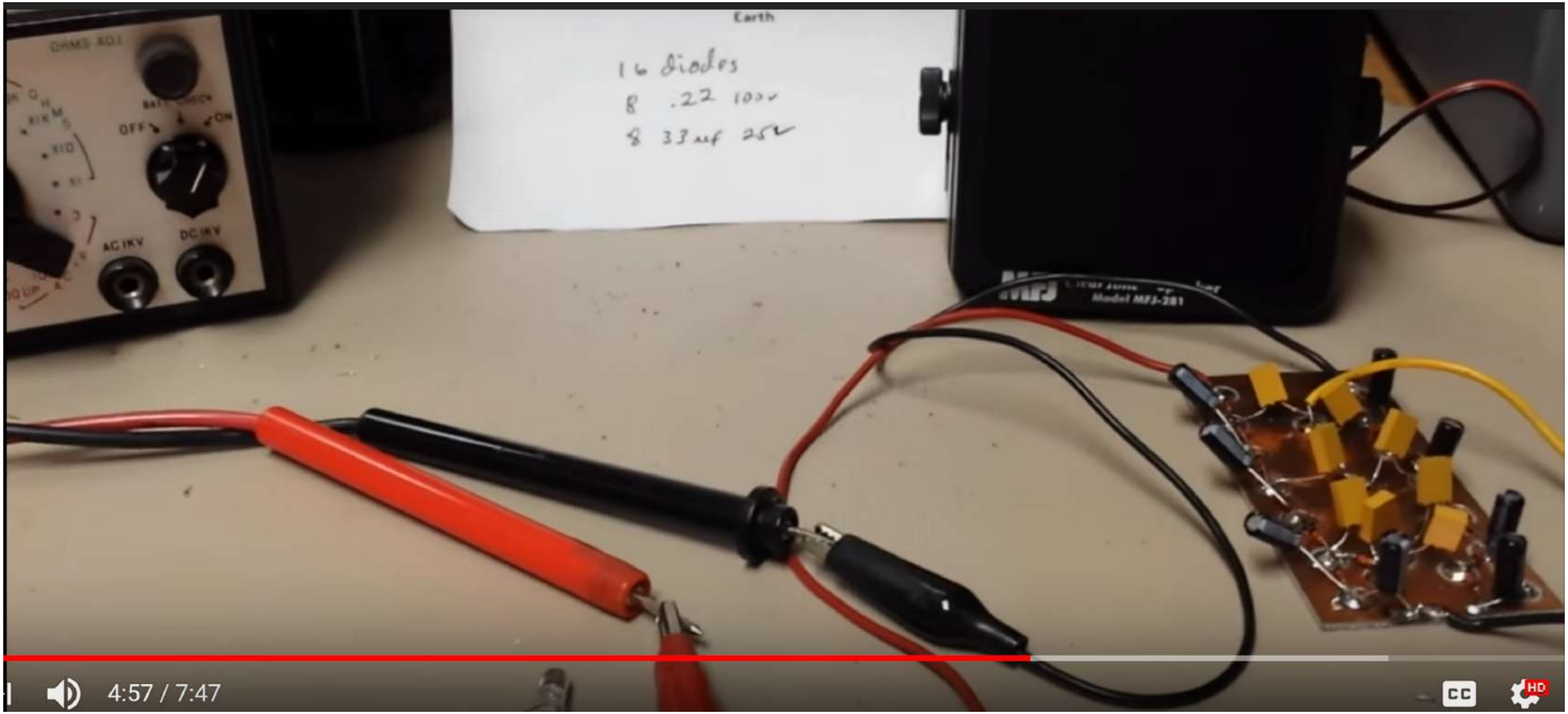

Earth

16 diodes
8 .22 100v
8 33uf 25v

OHMS ADJ
BATT CHECK
OFF
ON
AC 1KV
DC 1KV
MFJ
Model MFJ-281
4:57 / 7:47
CC
HD

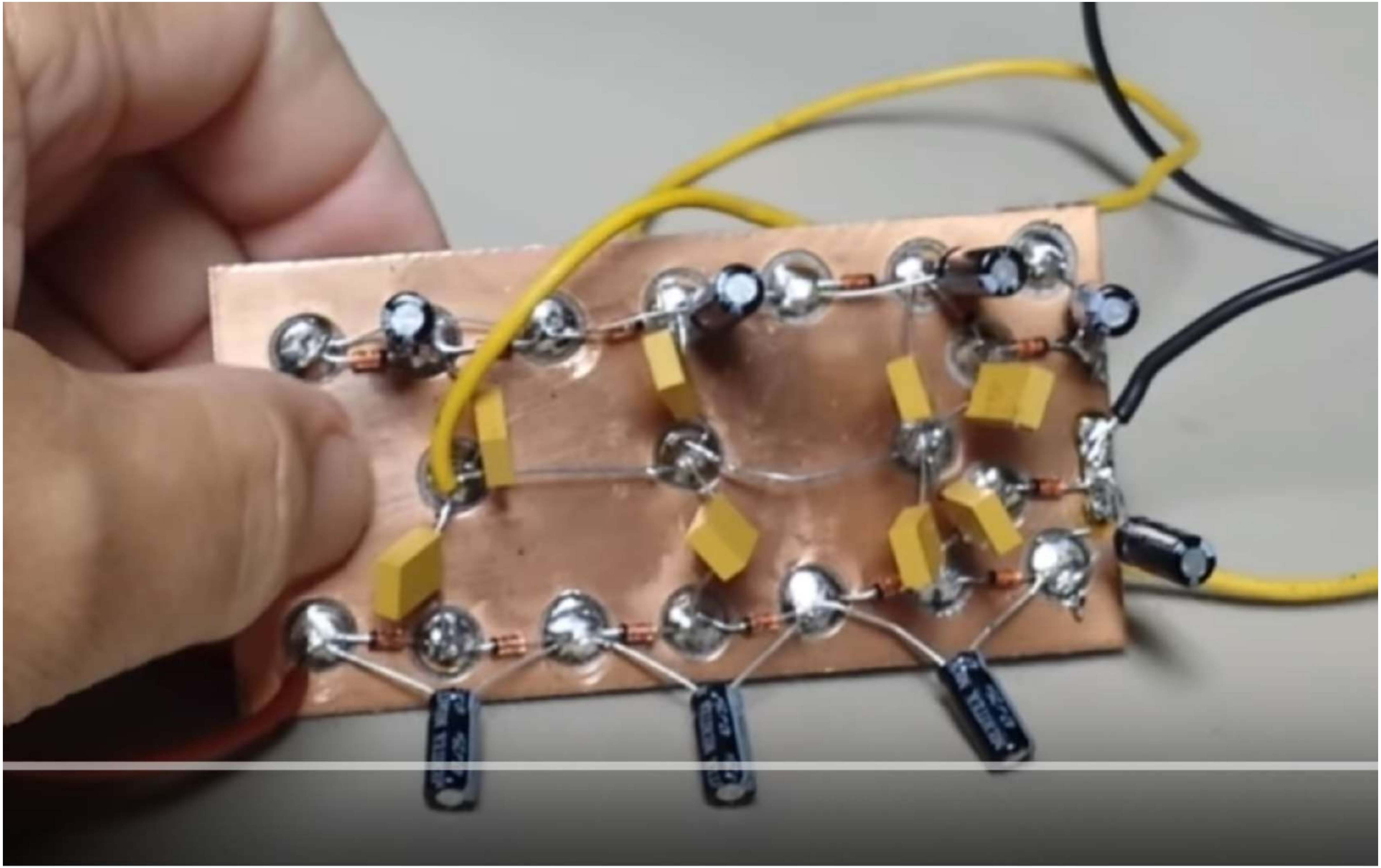

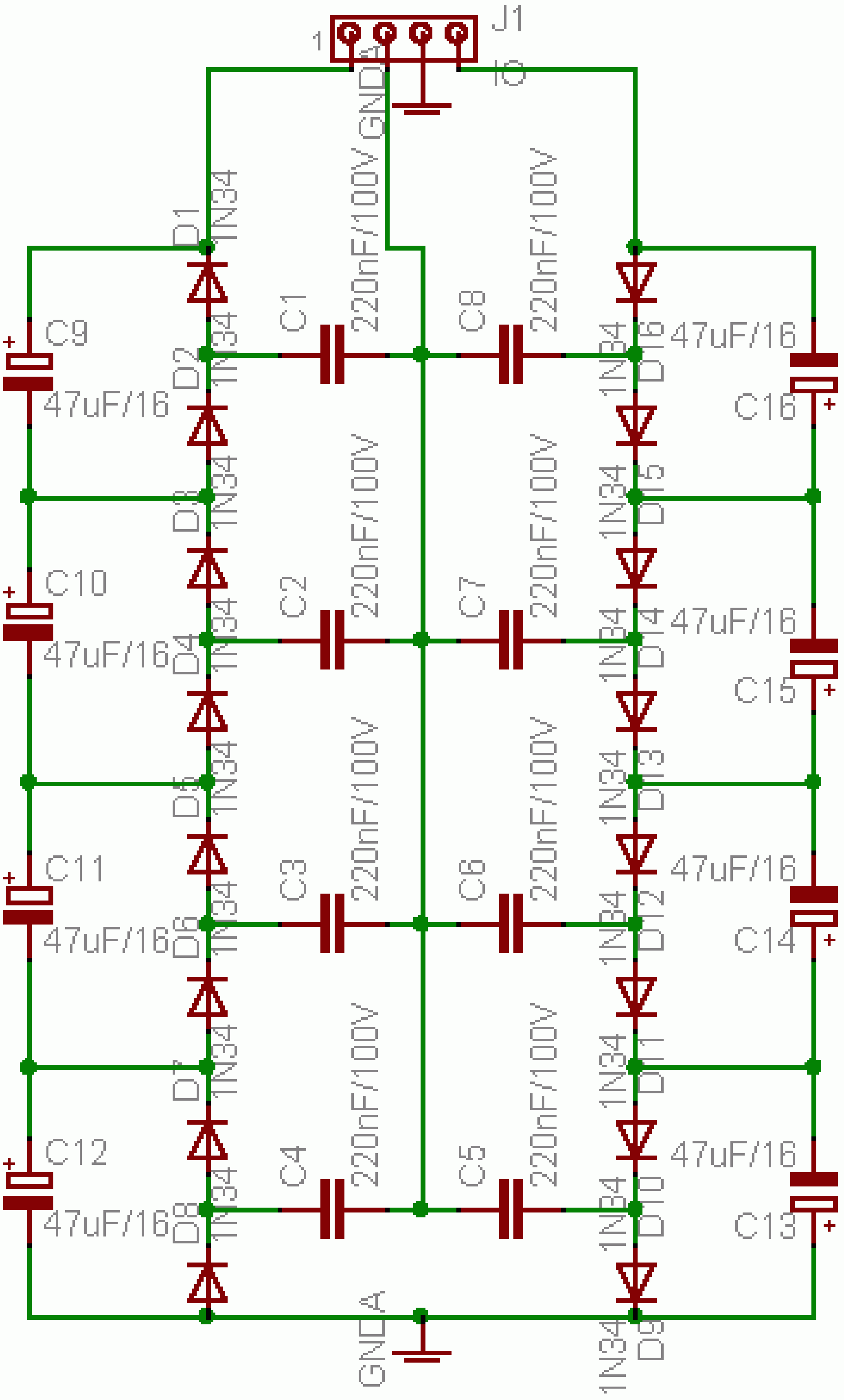
J1
1
GNDA
IO
D1
1N34
D2
1N34
D3
1N34
D4
1N34
D5
1N34
D6
1N34
D7
1N34
D8
1N34
D9
1N34
D10
1N34
D11
1N34
D12
1N34
D13
1N34
D14
1N34
D15
1N34
D16
1N34
C1
220nF/100V
C2
220nF/100V
C3
220nF/100V
C4
220nF/100V
C5
220nF/100V
C6
220nF/100V
C7
220nF/100V
C8
220nF/100V
C9
47uF/16
C10
47uF/16
C11
47uF/16
C12
47uF/16
C13
47uF/16
C14
47uF/16
C15
47uF/16
C16
47uF/16
+
GNDA

AMBIENT POWER MODULE

(C) - GROUNDLOOP 2012

Circuit Diagram

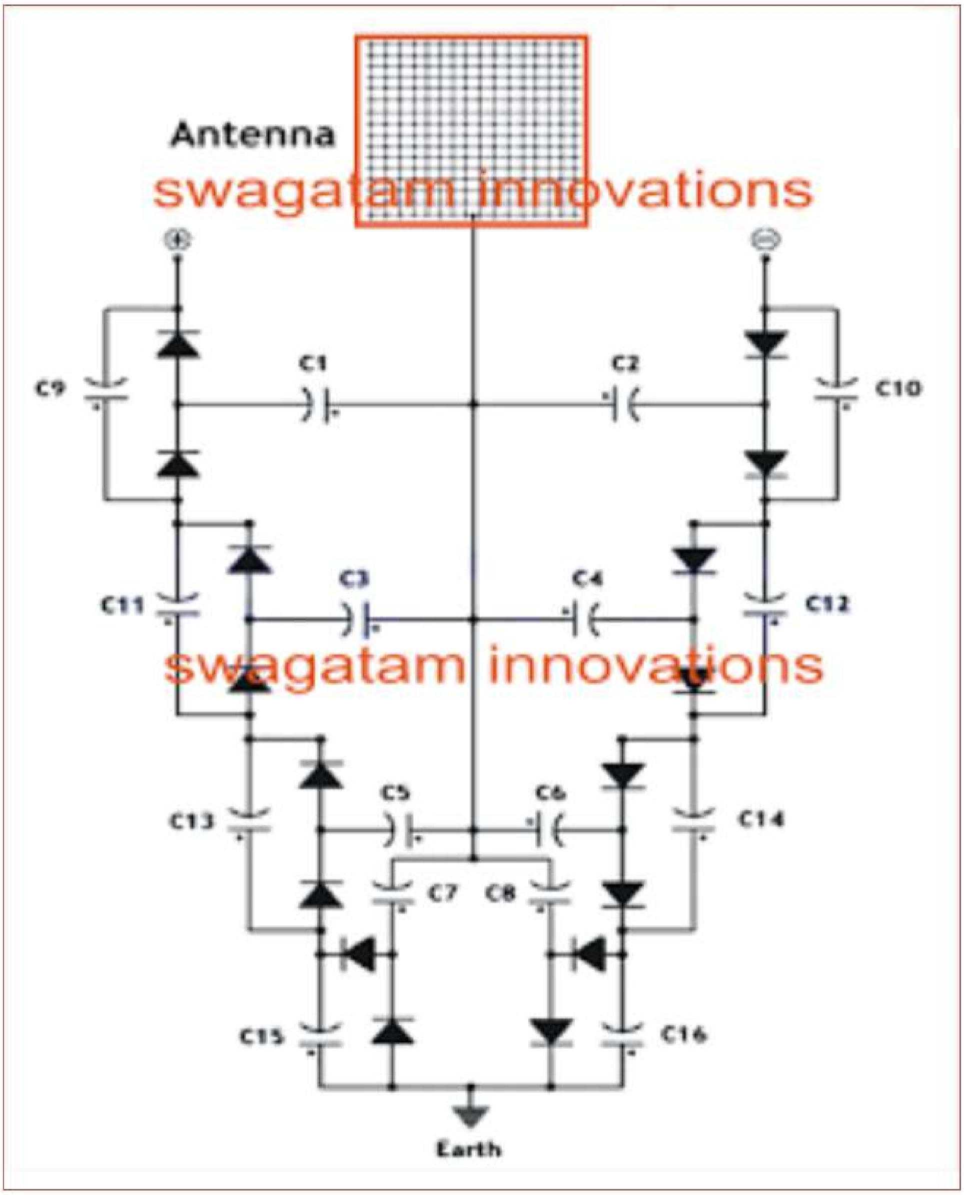
Antenna
swagatam innovations
⊕
⊖
C1
C2
C9
C10
C3
C4
C11
C12
swagatam innovations
C5
C6
C13
C14
C7
C8
C15
C16
Earth

Parts List

All Diodes are 1N4148

C1---C8 = 0.22uF/100V mylar

C9----C16 = 33uF/25V electrolytic

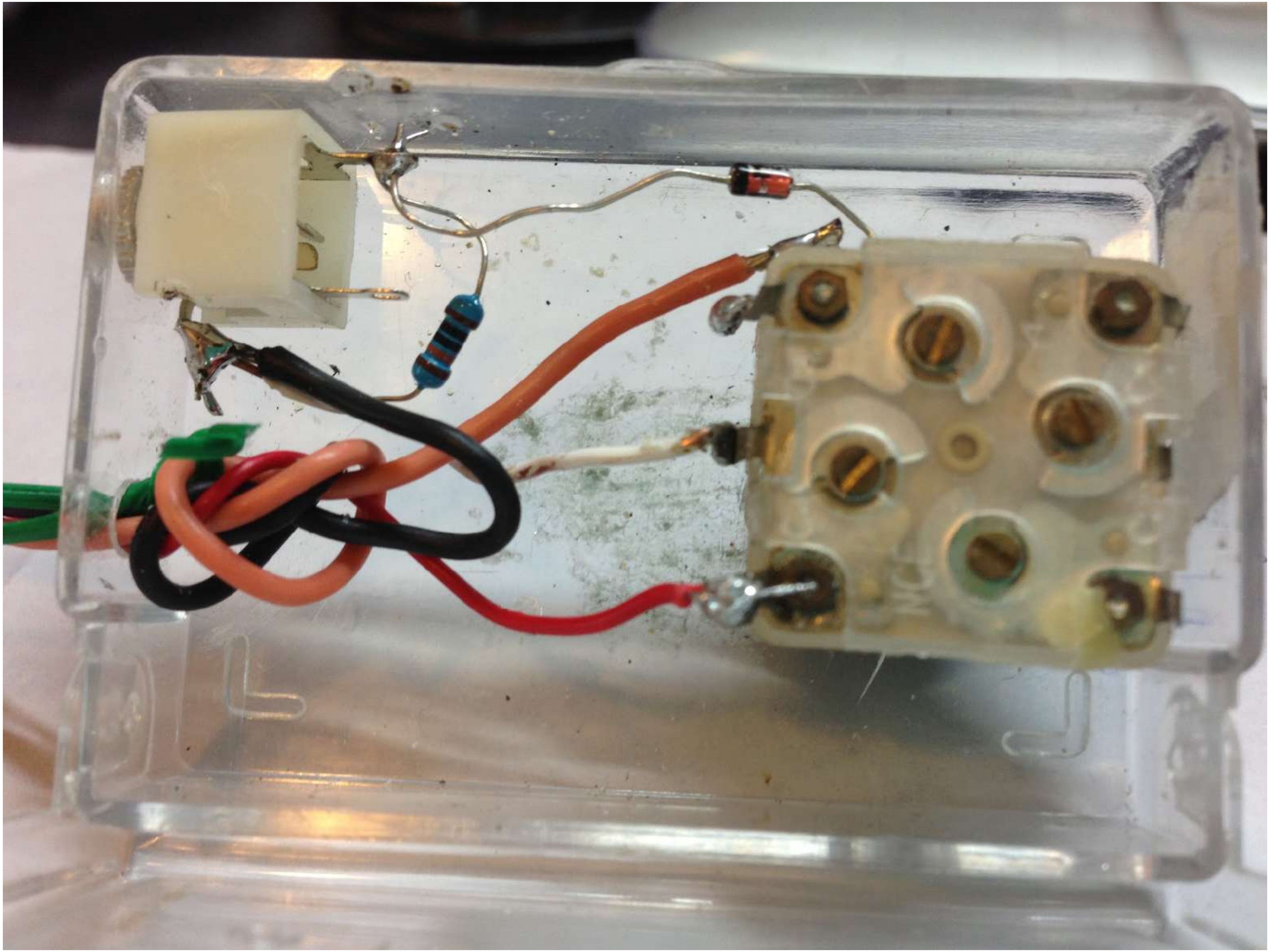

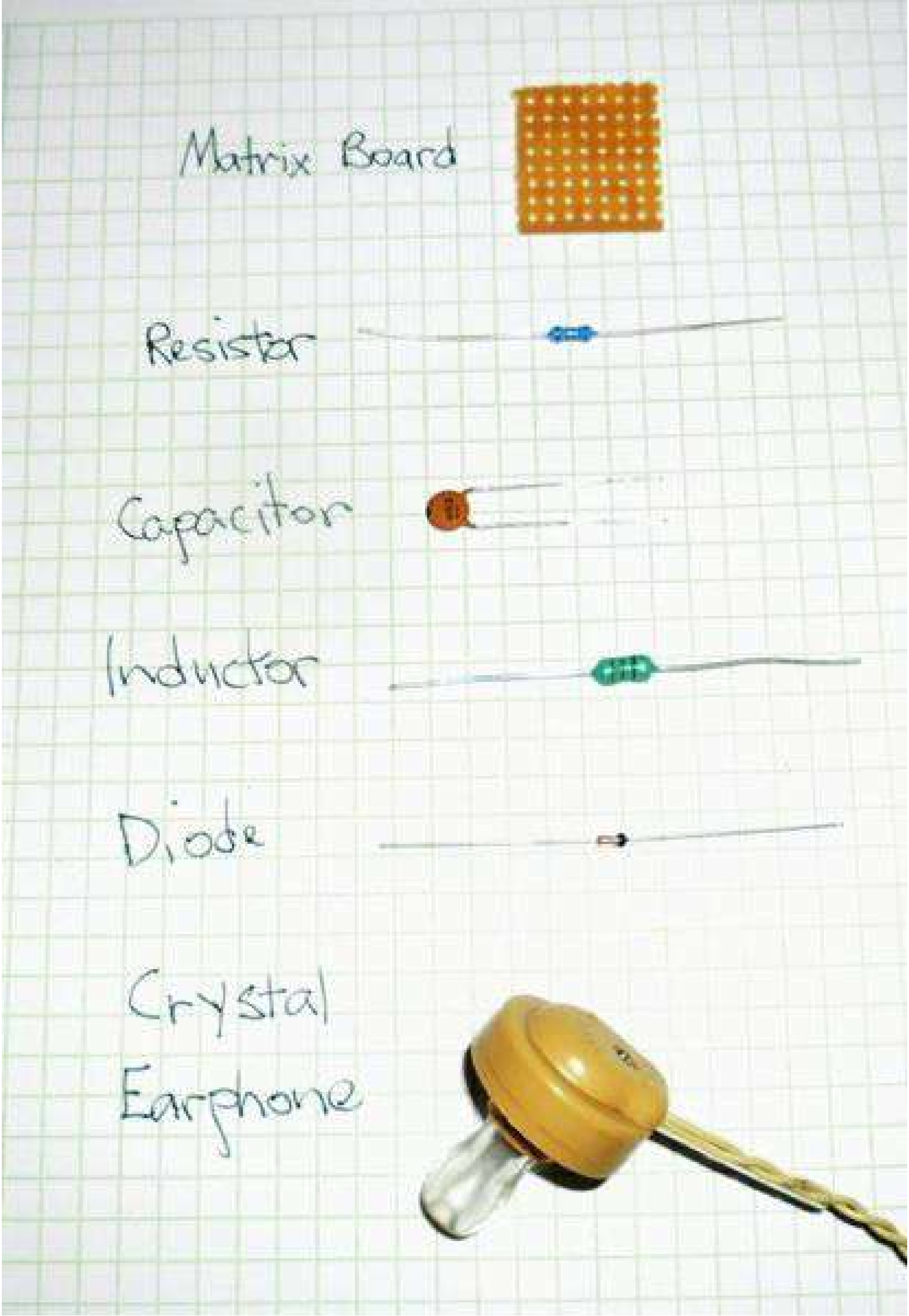
Matrix Board
Resistor
Capacitor
Inductor
Diode
Crystal Earphone

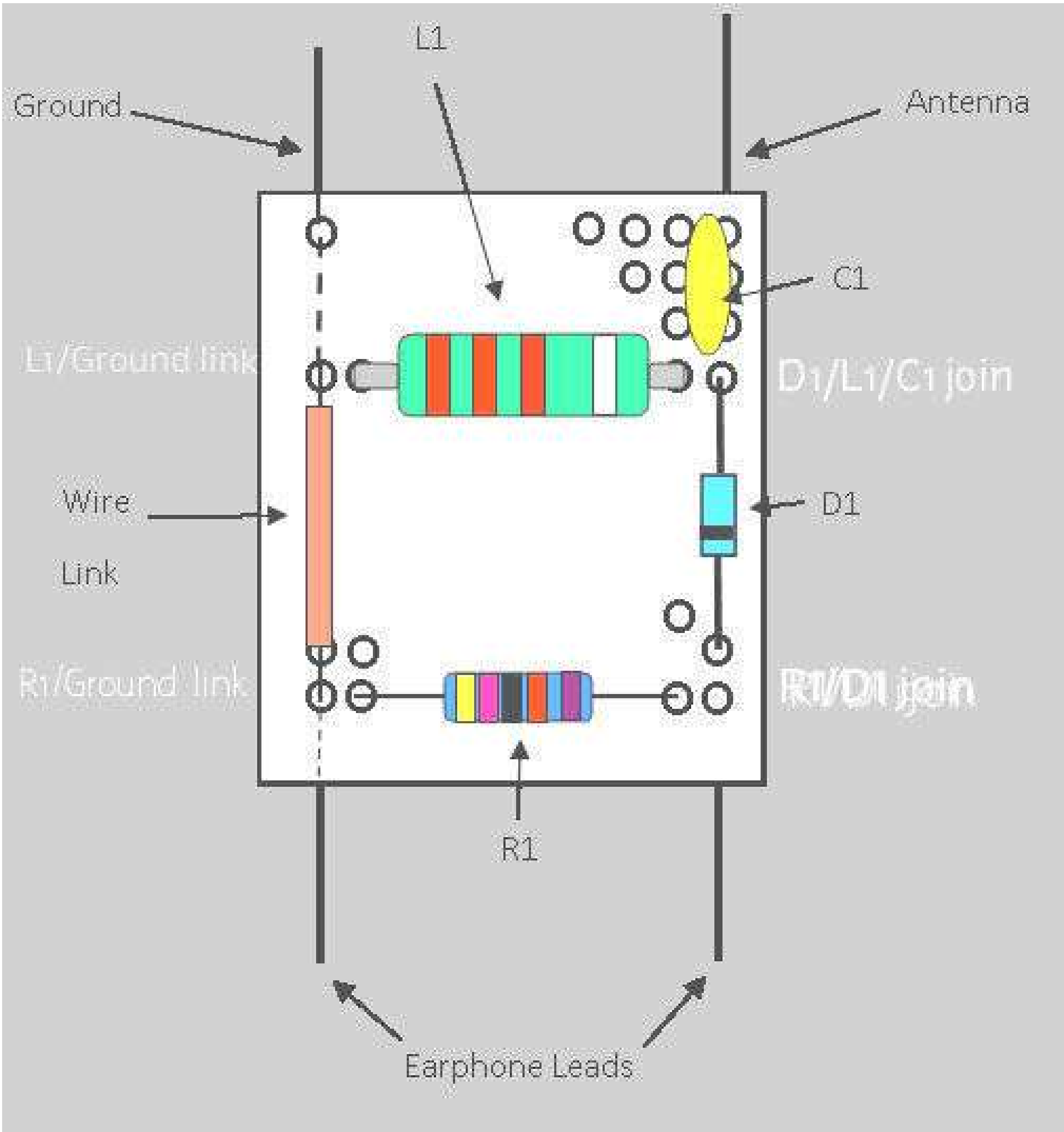
Ground
L1
Antenna
C1
L1/Ground link
D1/L1/C1 join
Wire
Link
D1
R1/Ground link
R1/D1 join
R1
Earphone Leads

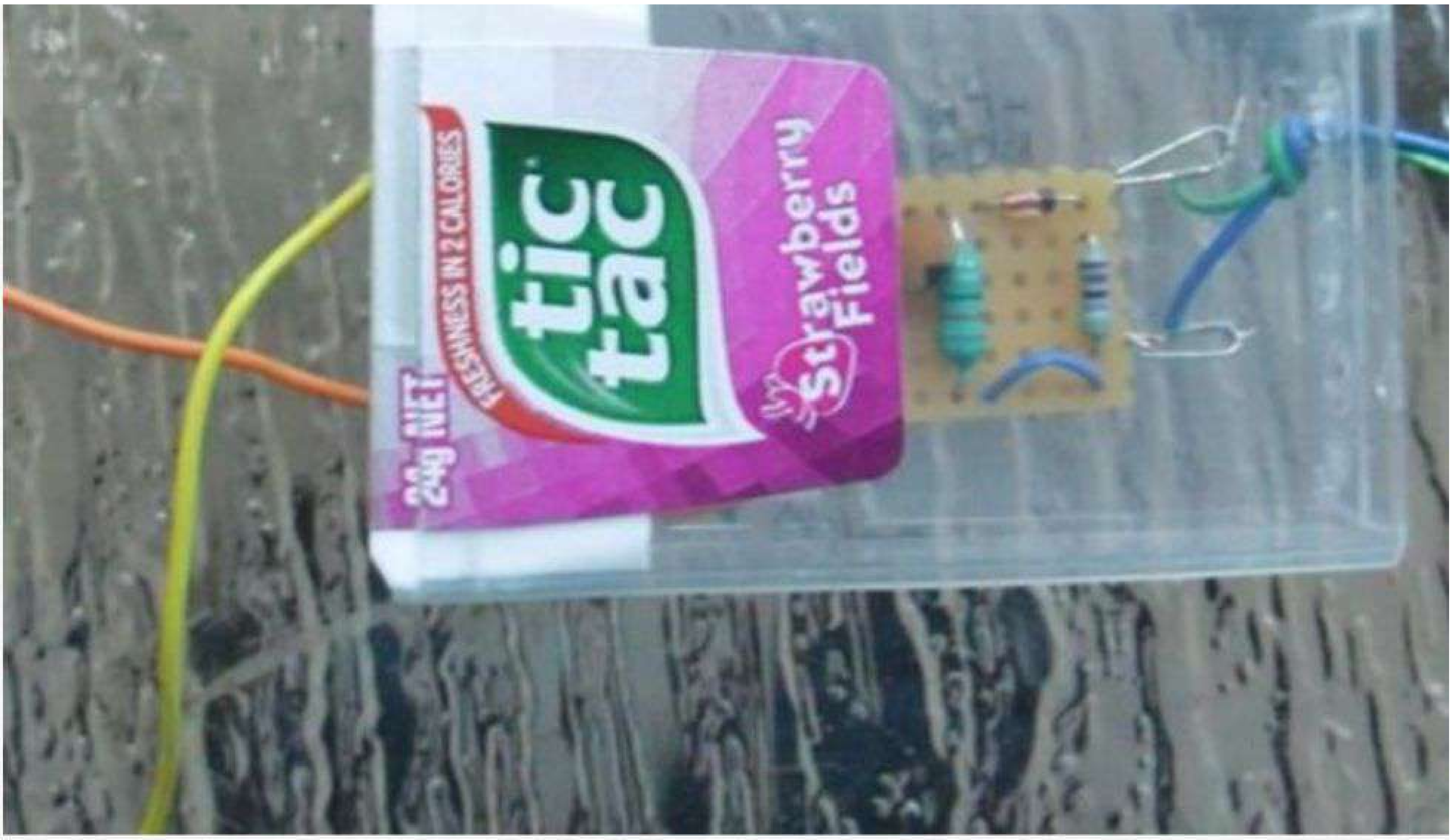
tic tac
FRESHNESS IN 2 CALORIES
24g NET
Strawberry Fields

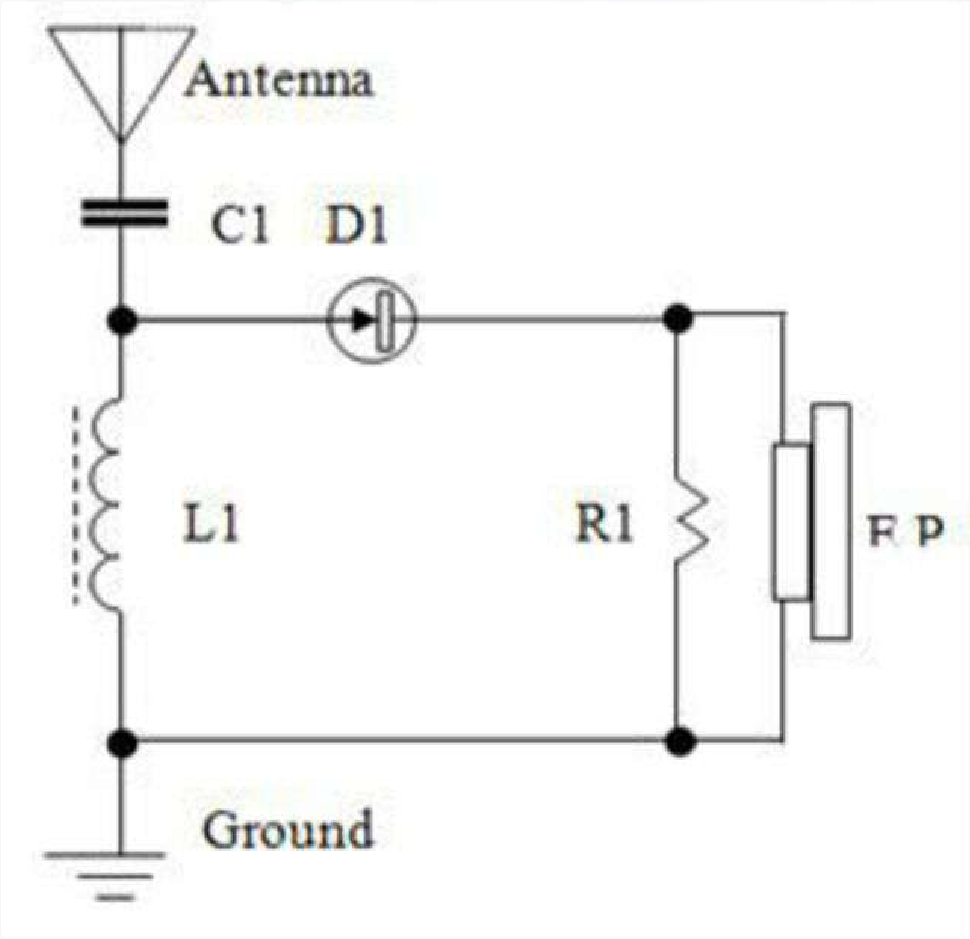
Antenna
C1
D1
L1
R1
EP
Ground

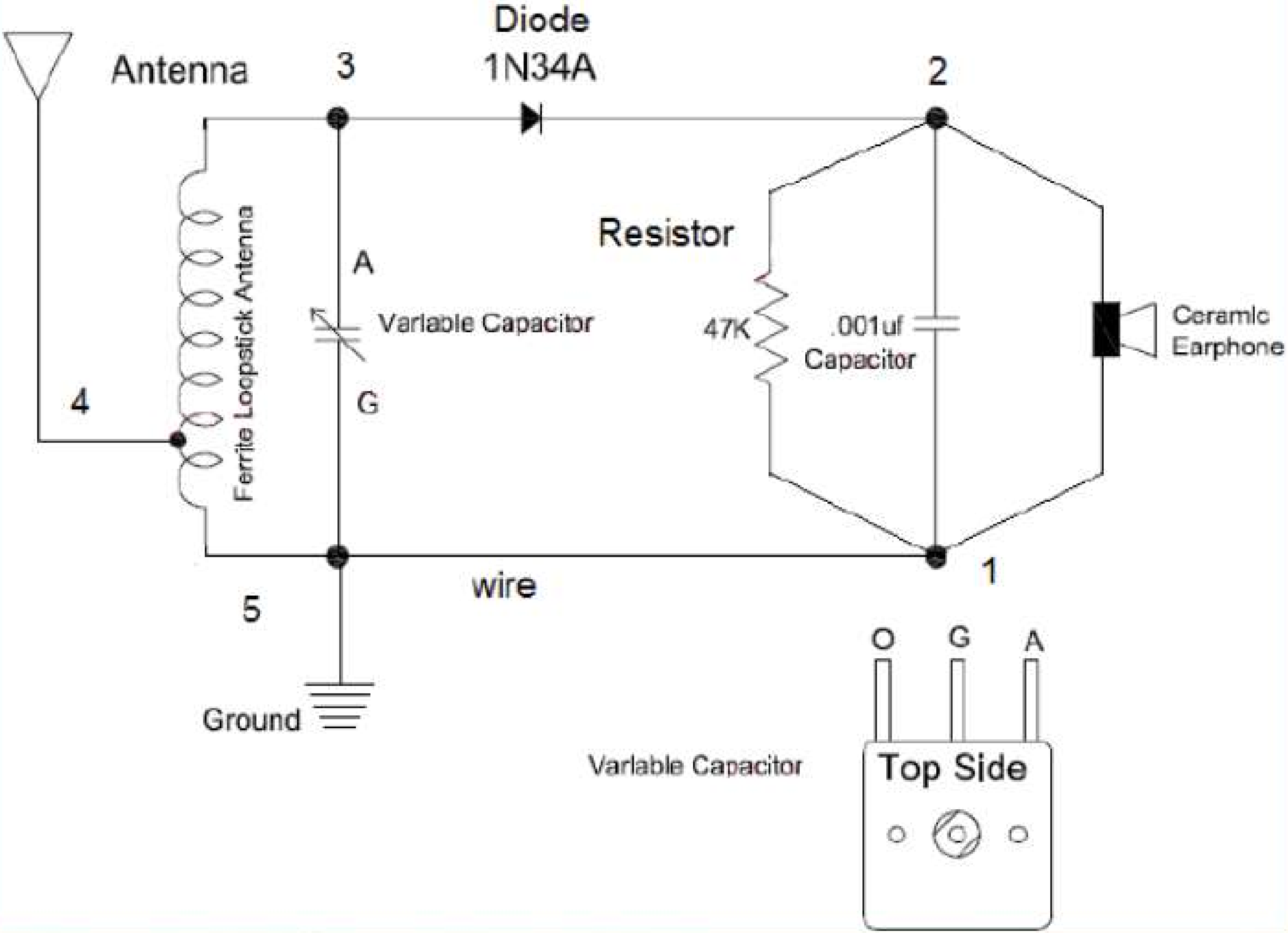
Diode
1N34A
Antenna
3
2
Resistor
Ferrite Loopstick Antenna
A
Variable Capacitor
47K
.001uf
Capacitor
Ceramic
Earphone
G
4
1
wire
5
Ground
O
G
A
Variable Capacitor
Top Side

[So
http://www.angelfire.com/electronic2/index1/DmCrystalSet.html]

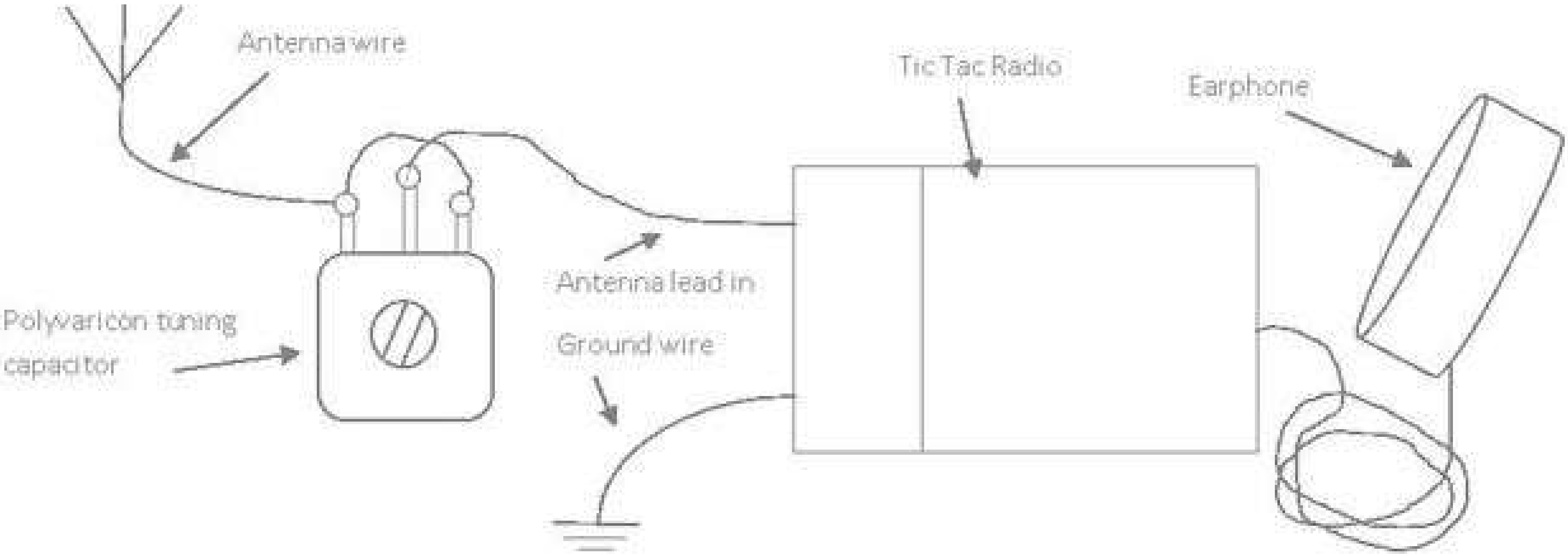
Antenna wire
Tic Tac Radio
Earphone
Antenna lead in
Polyvaricon tuning capacitor
Ground wire

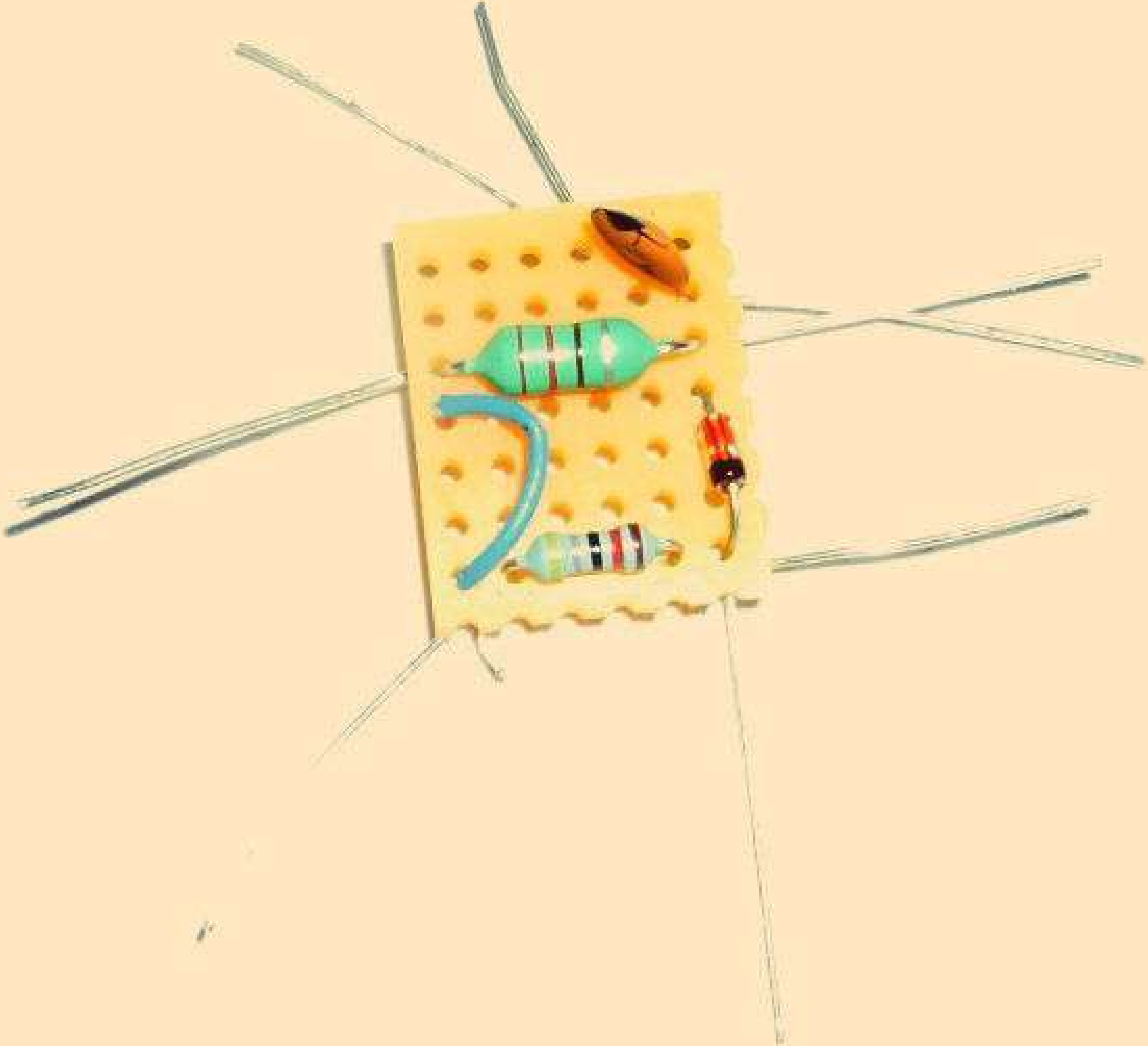

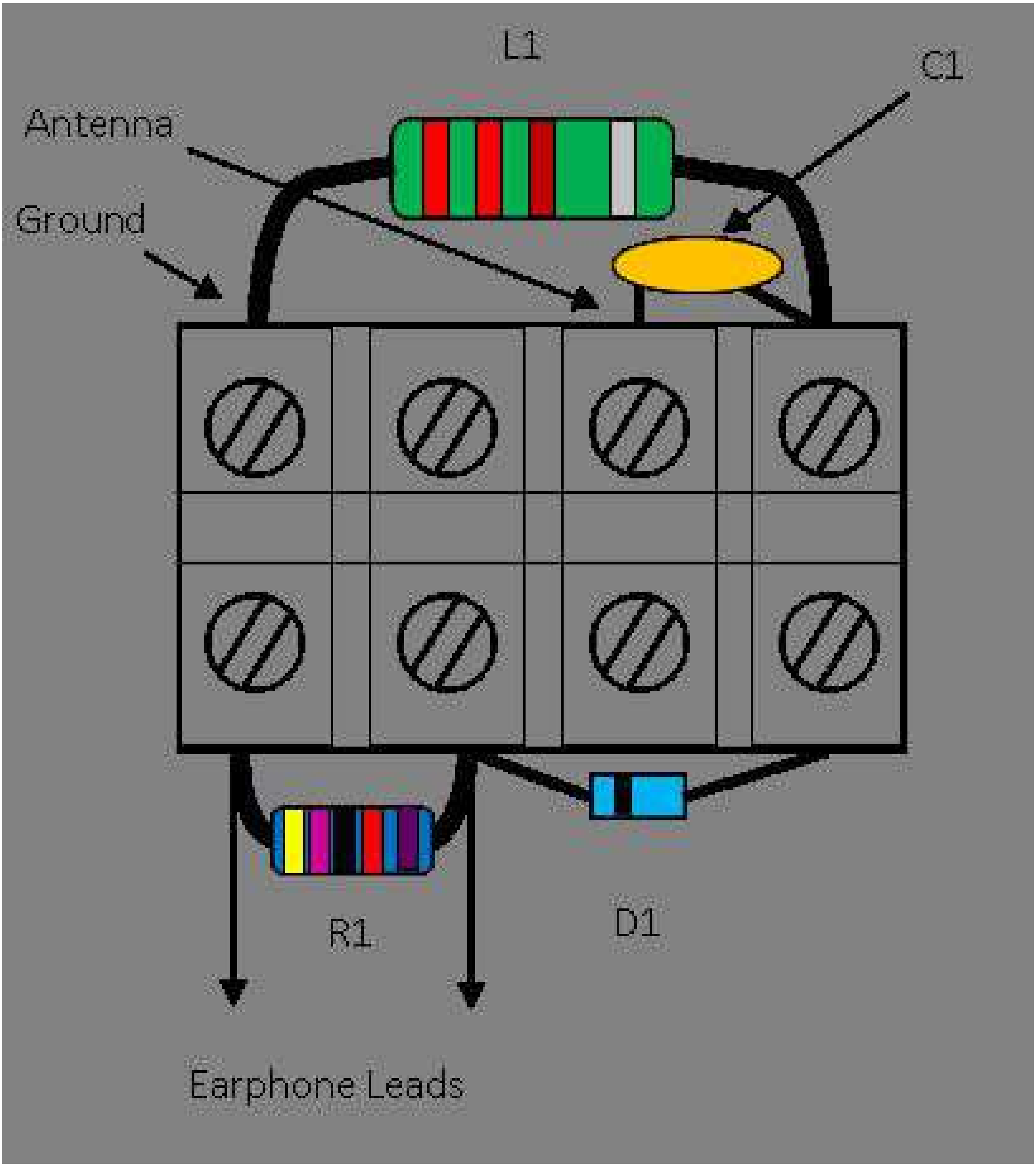
L1
C1
Antenna
Ground
R1
D1
Earphone Leads

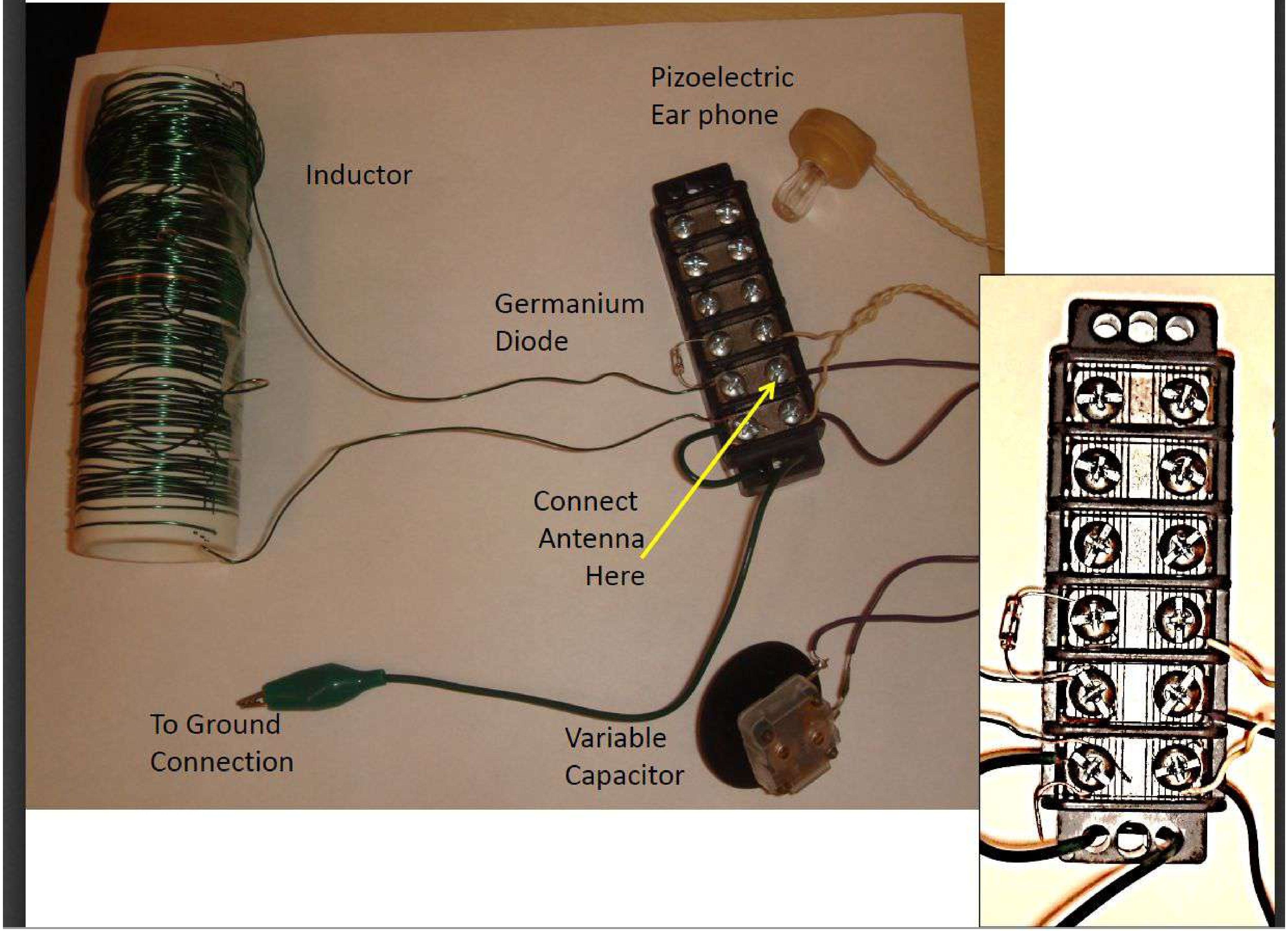

Pizoelectric
Ear phone

Inductor

Germanium
Diode

Connect
Antenna
Here

To Ground
Connection

Variable
Capacitor

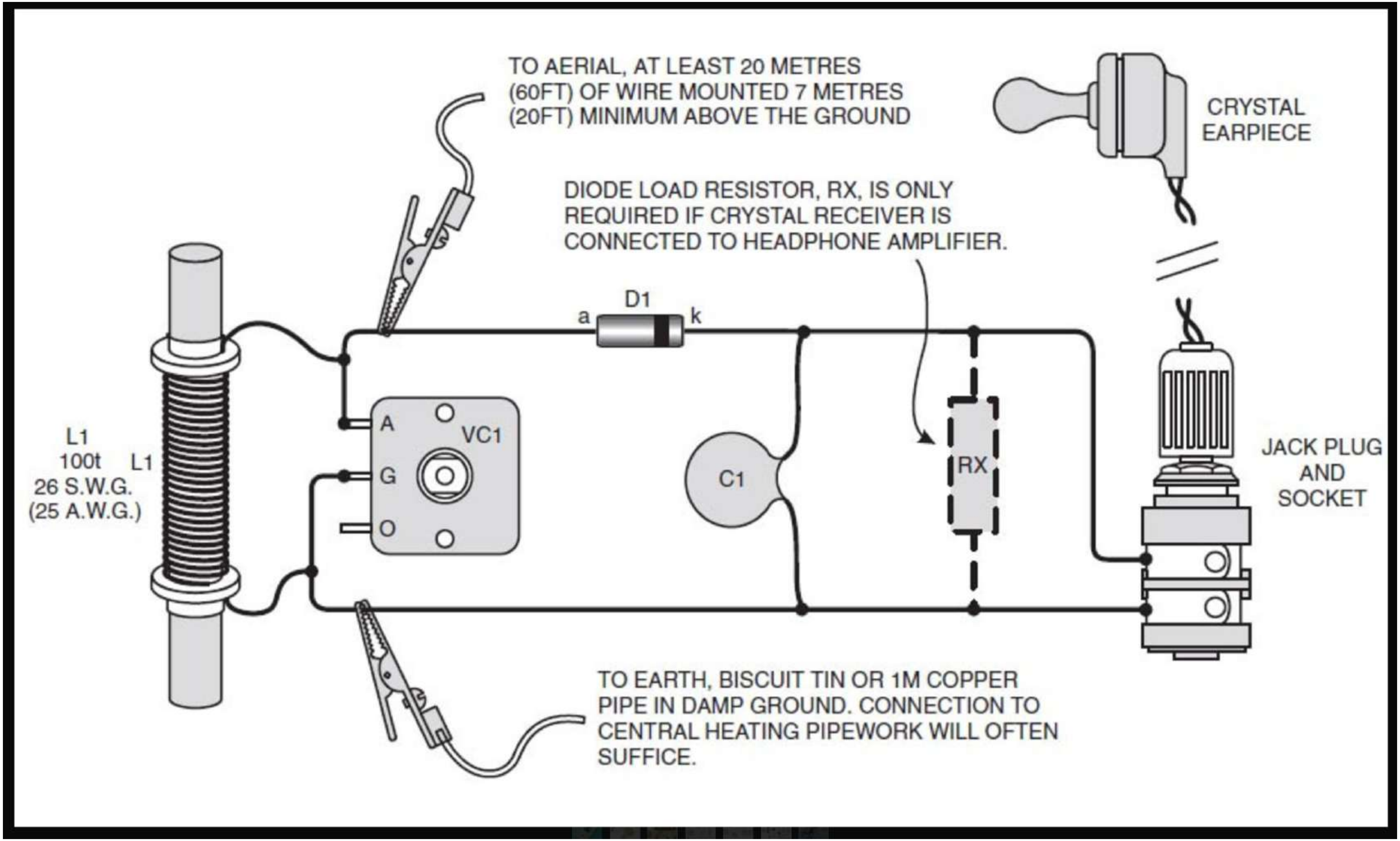TO AERIAL, AT LEAST 20 METRES (60FT) OF WIRE MOUNTED 7 METRES (20FT) MINIMUM ABOVE THE GROUND
CRYSTAL EARPIECE
DIODE LOAD RESISTOR, RX, IS ONLY REQUIRED IF CRYSTAL RECEIVER IS CONNECTED TO HEADPHONE AMPLIFIER.
D1
a
k
L1
100t
L1
26 S.W.G.
(25 A.W.G.)
A
VC1
G
O
C1
RX
JACK PLUG AND SOCKET
TO EARTH, BISCUIT TIN OR 1M COPPER PIPE IN DAMP GROUND. CONNECTION TO CENTRAL HEATING PIPEWORK WILL OFTEN SUFFICE.

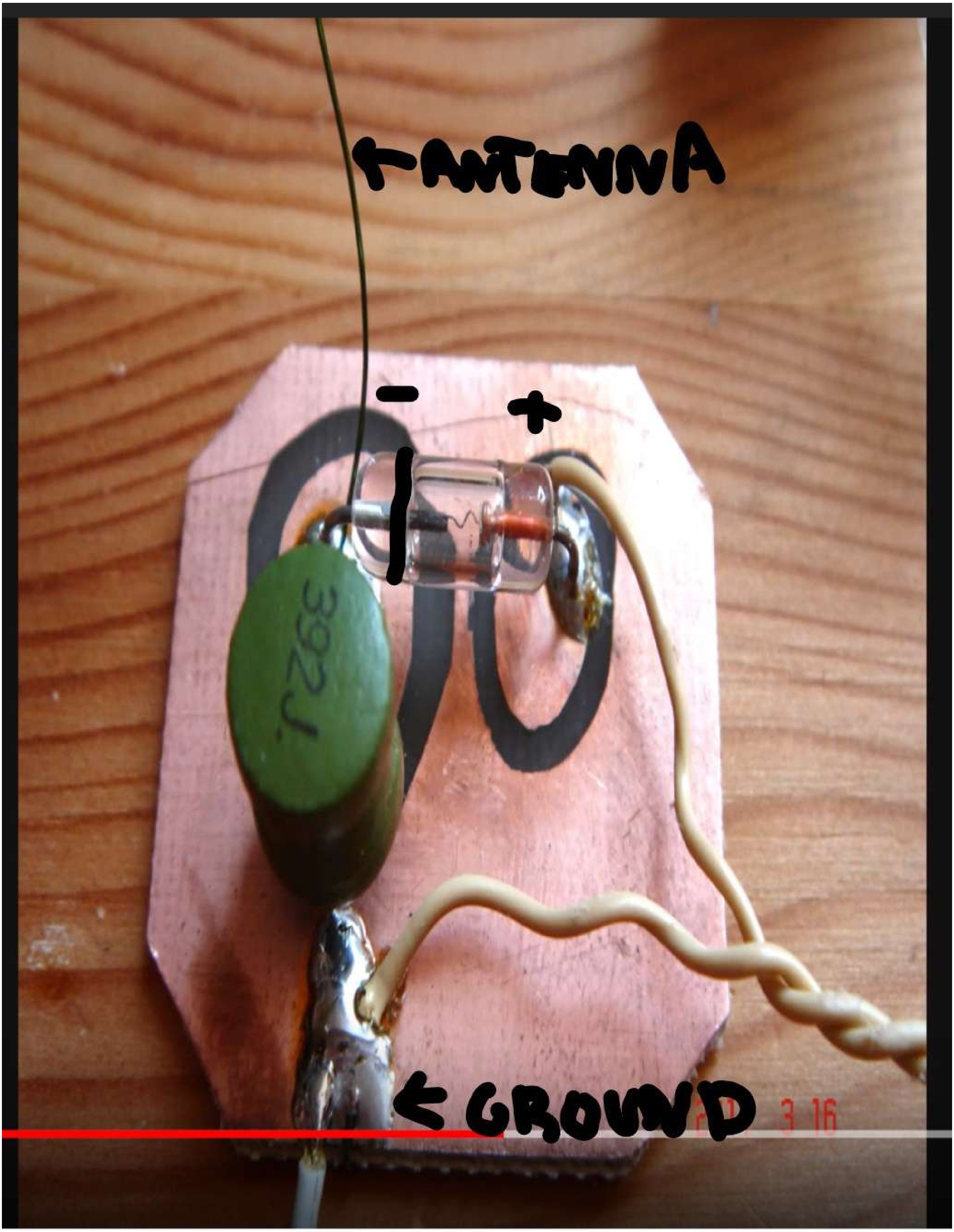
ANTENNA
−
+
392J.
GROUND
3 16

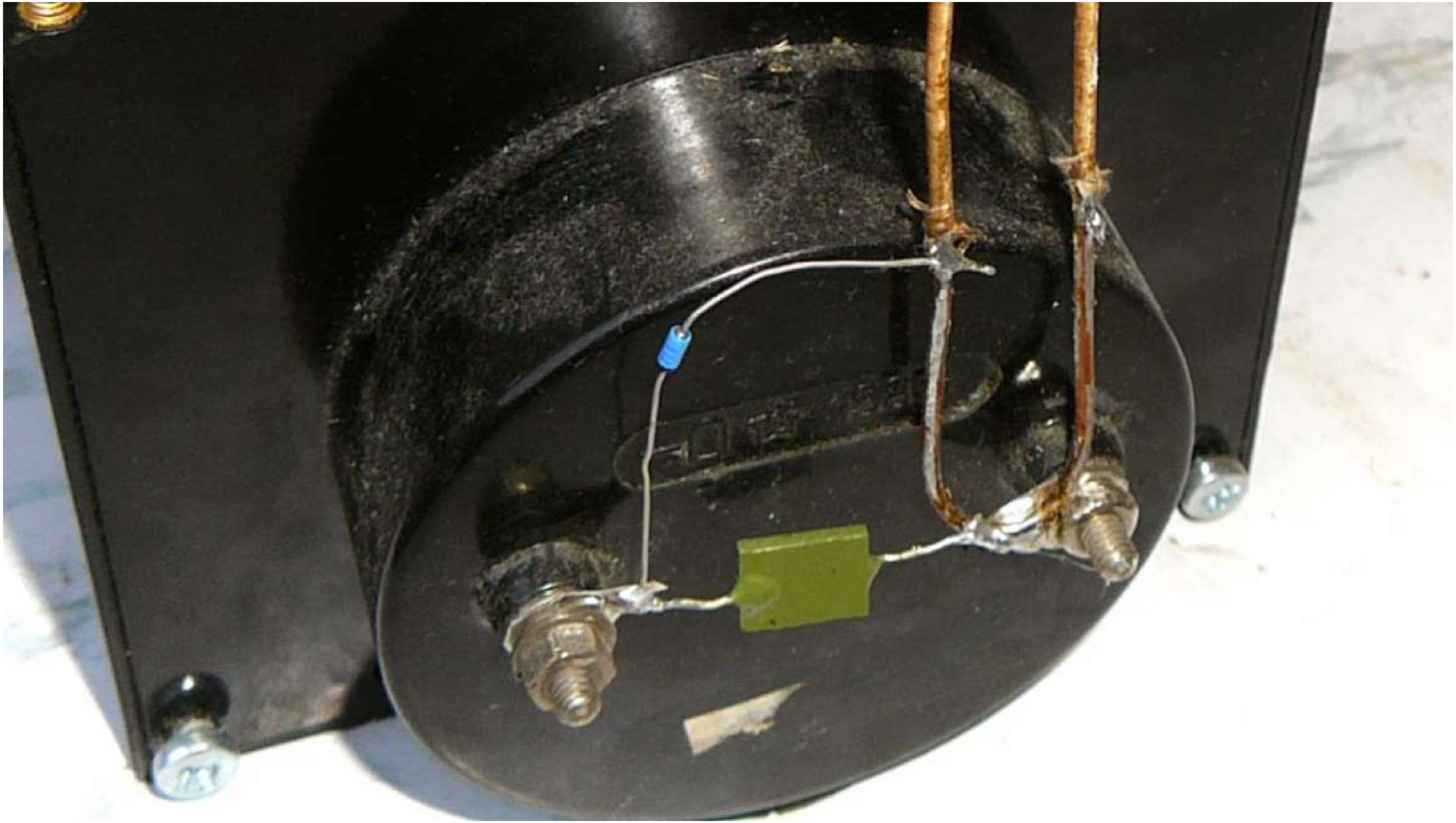

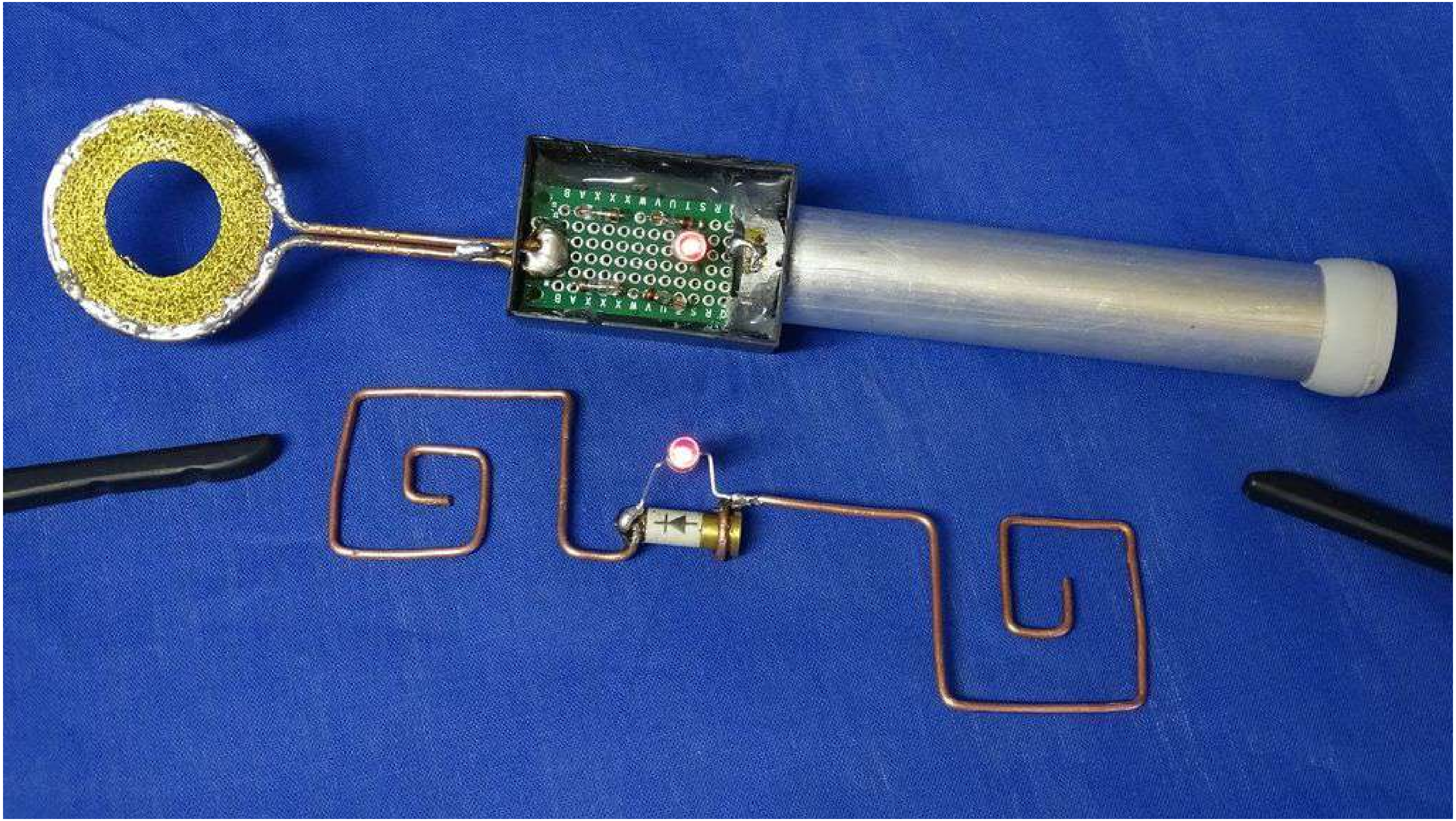

Step 1: LED+RF Diode

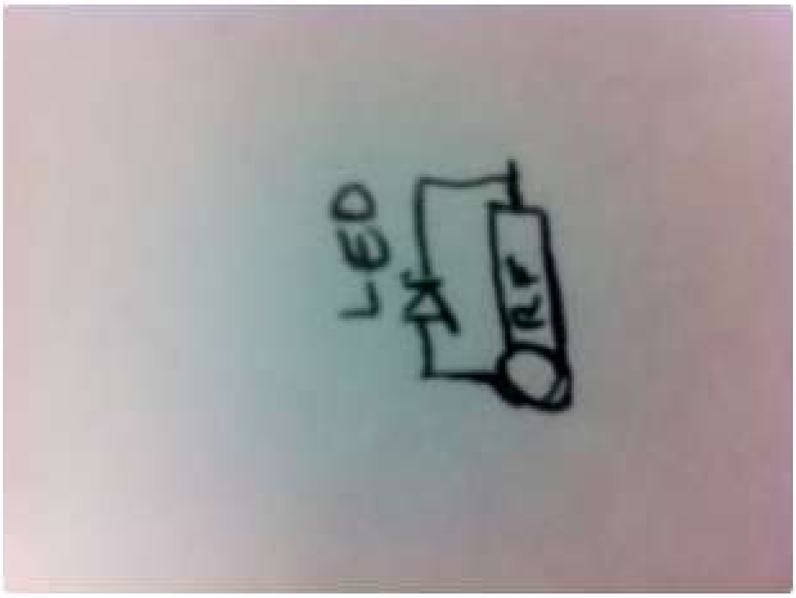

First Solder the led parallel to the Rf diode

Add Tip Ask Question Comment Download

Step 2: RF Diode+ LED+ Wires

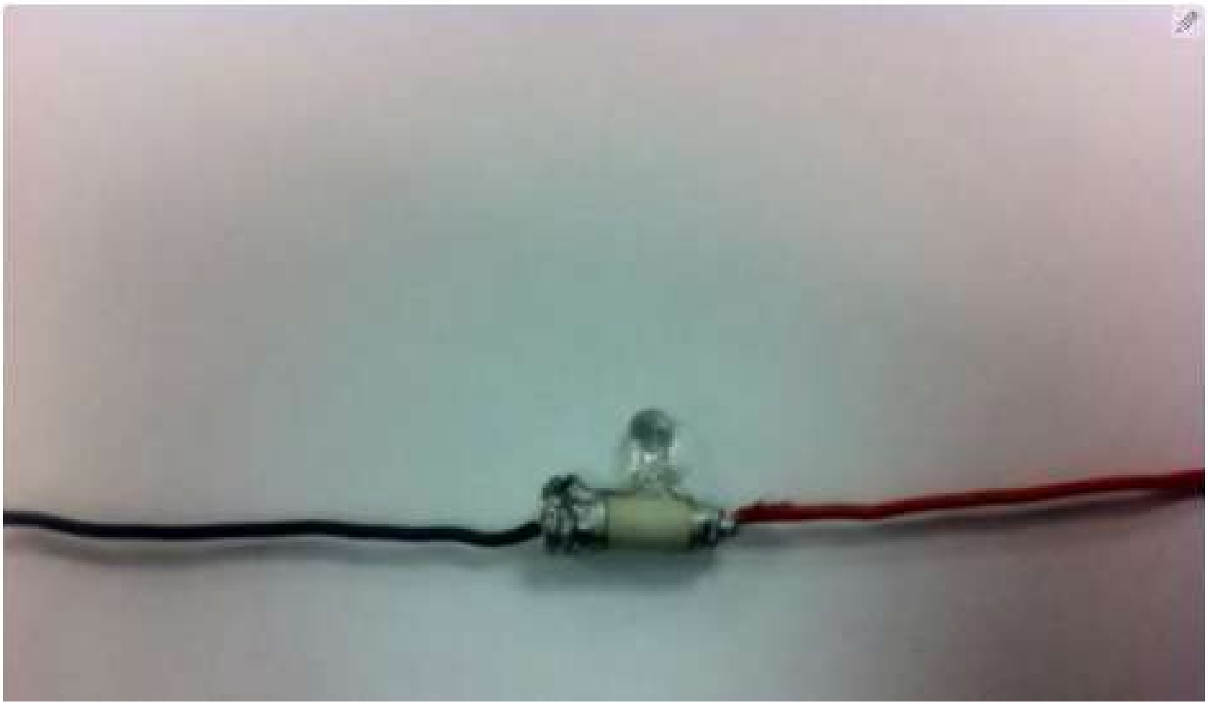

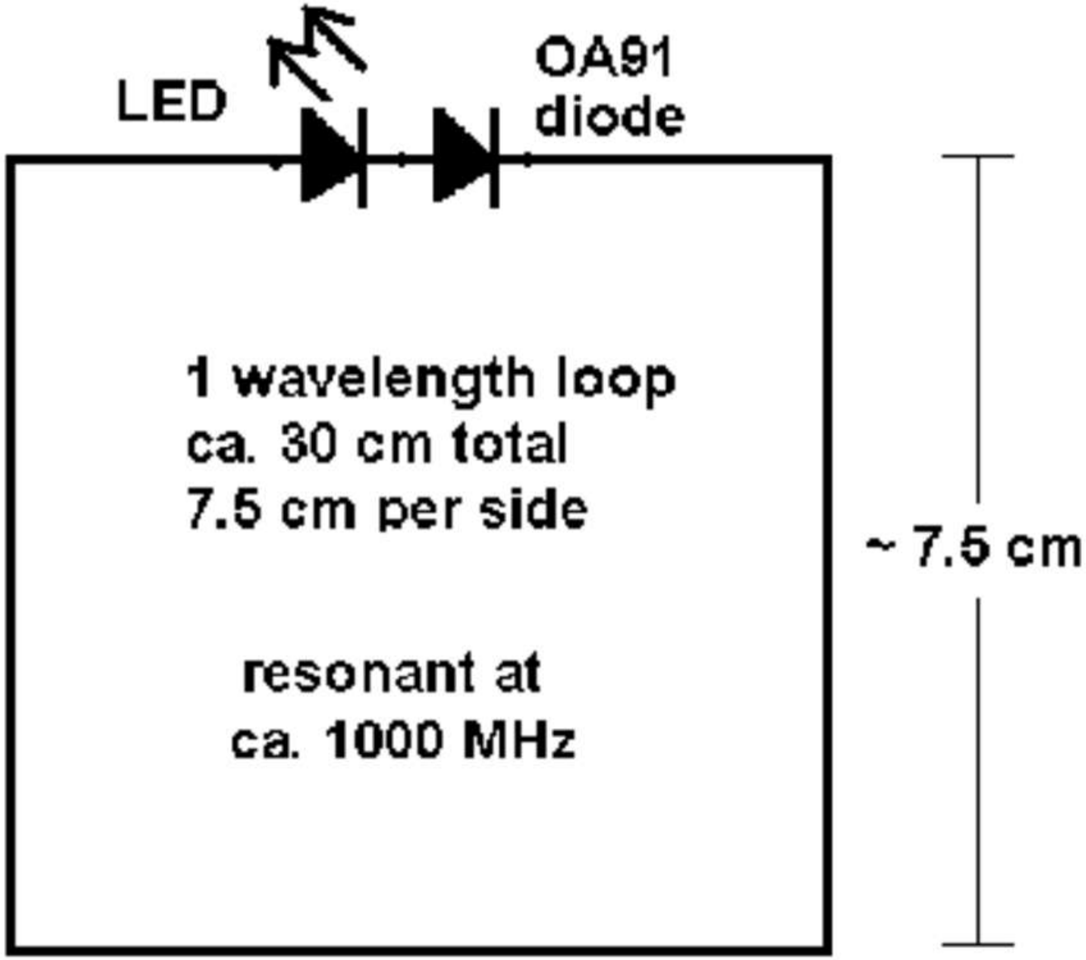
LED
OA91
diode
~ 7.5 cm

1 wavelength loop
ca. 30 cm total
7.5 cm per side

resonant at
ca. 1000 MHz

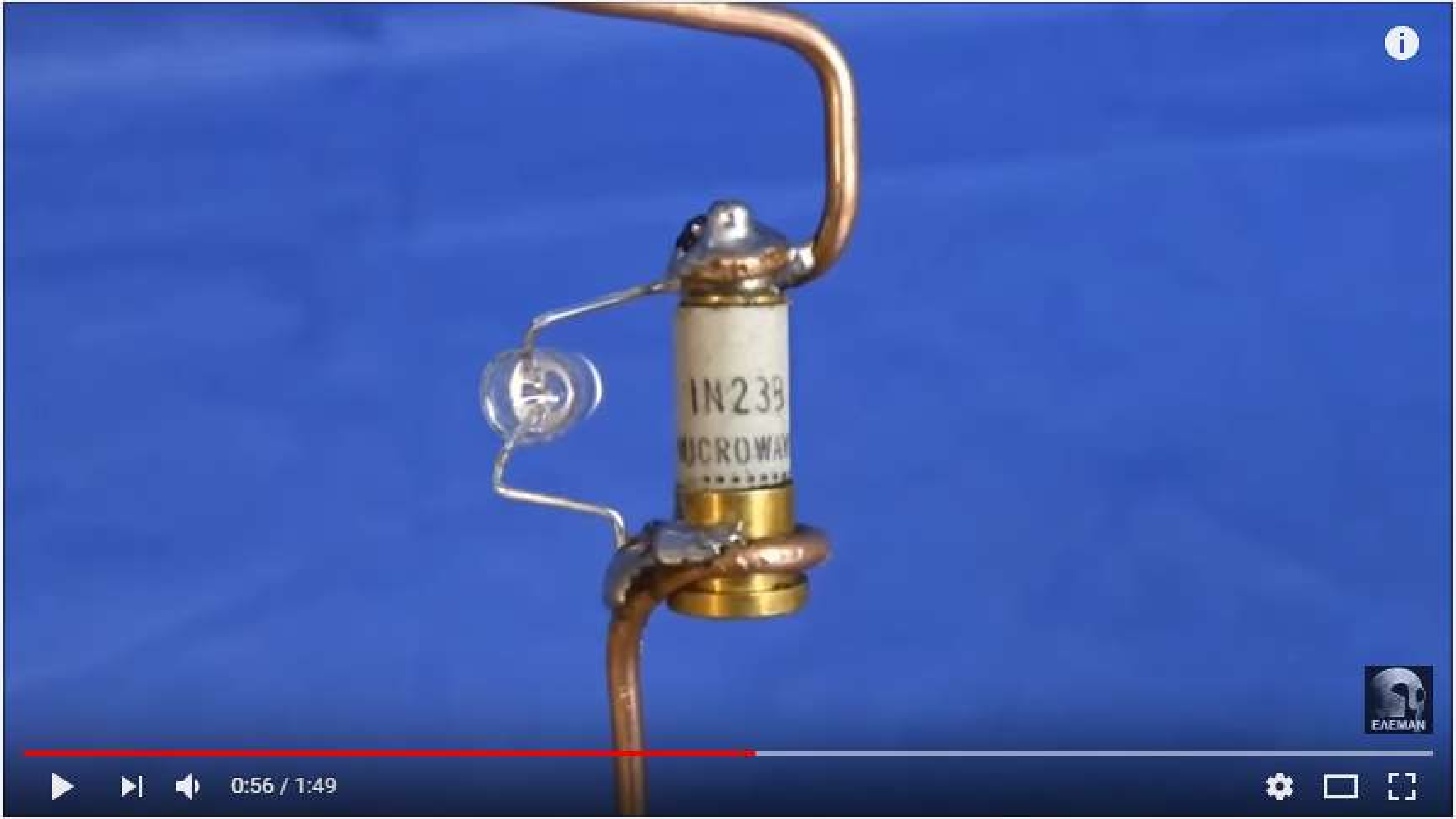
IN23B
ICROWA
ΕΛΕΜΑΝ
0:56 / 1:49

COM LOGIC

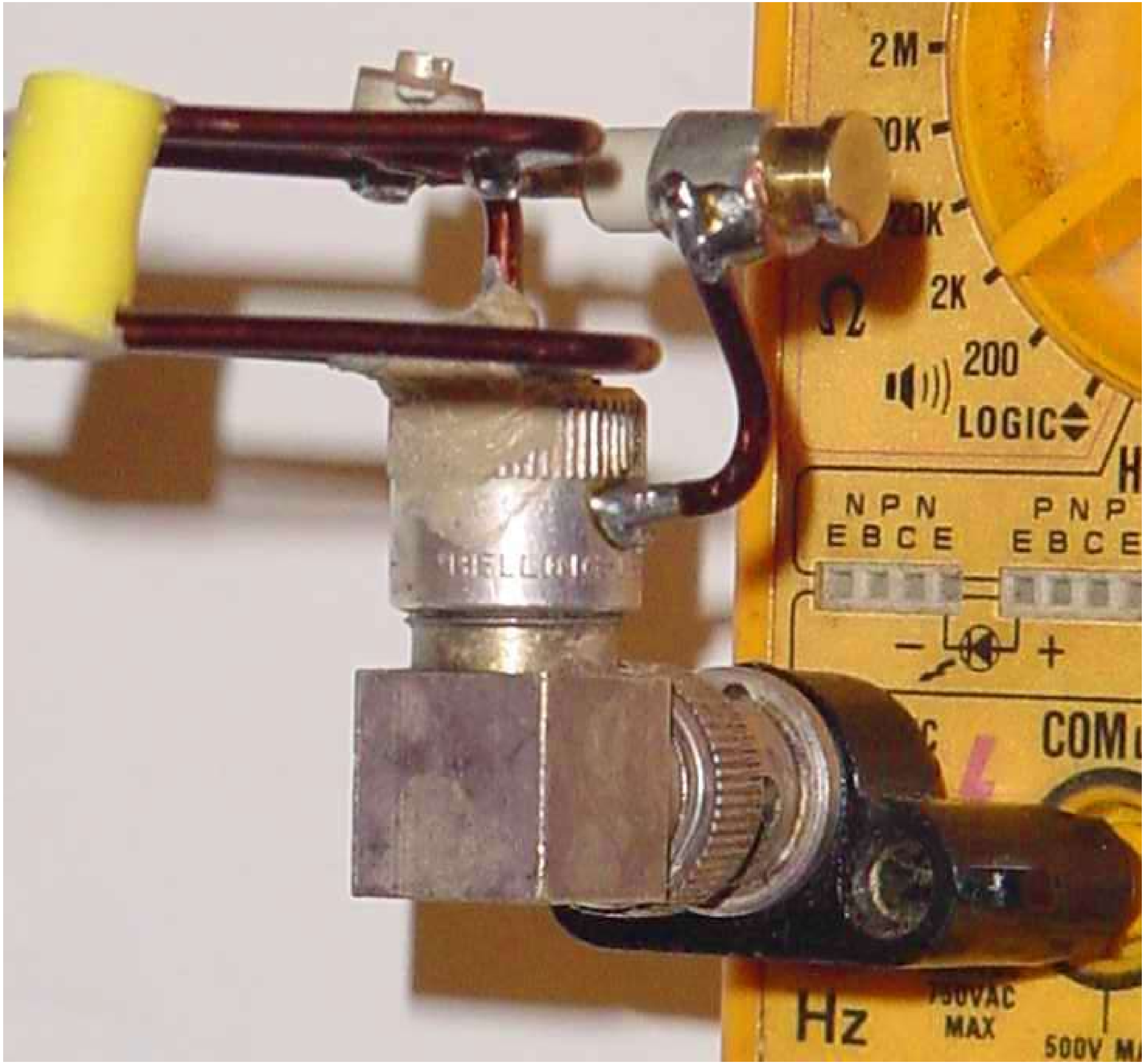
2M
2K
200
Ω
LOGIC
NPN
EBCE
PNP
EBCE
–
+
COM
Hz
MAX
500V

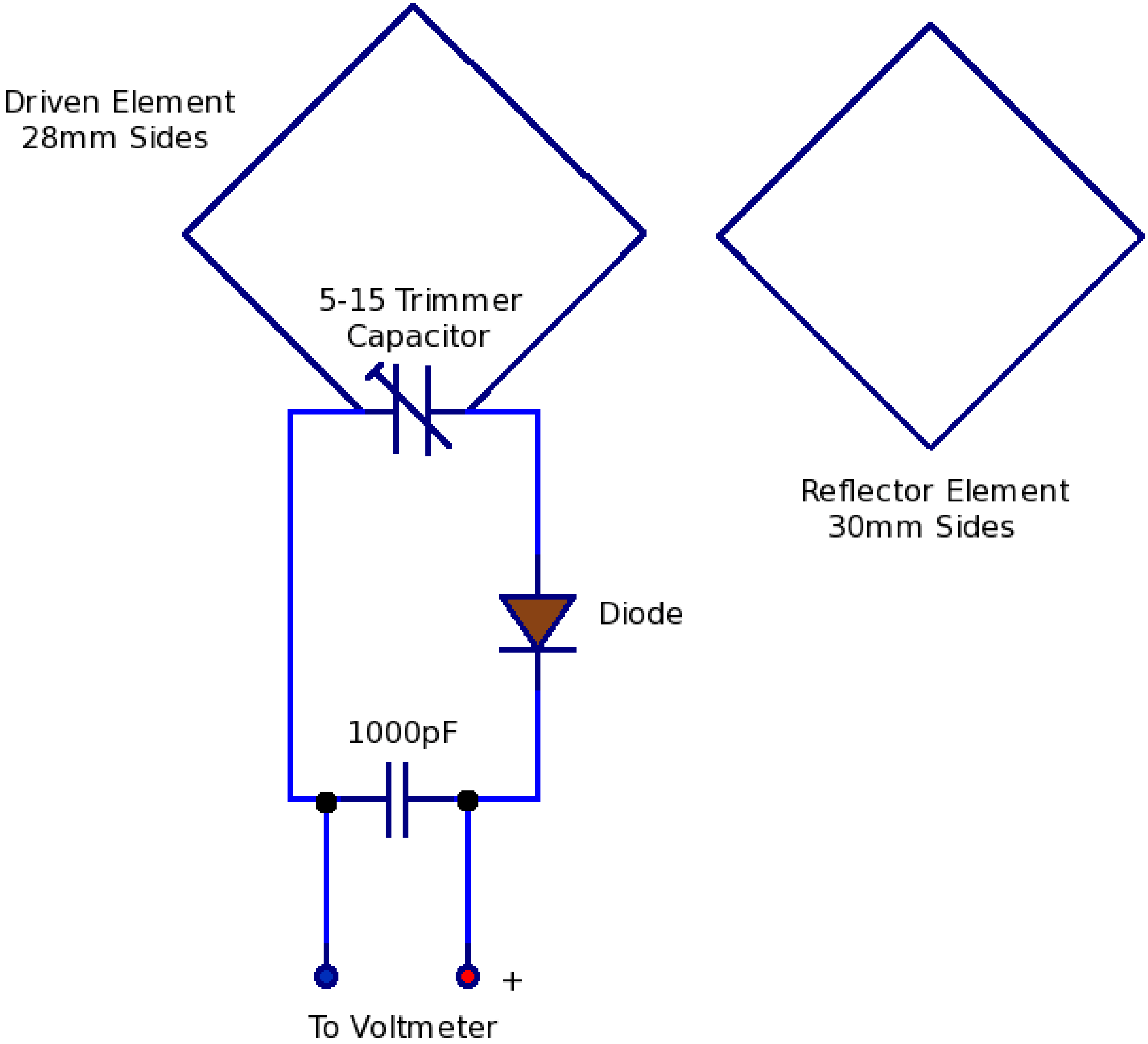
Driven Element
28mm Sides
5-15 Trimmer
Capacitor
Reflector Element
30mm Sides
Diode
1000pF
+
To Voltmeter

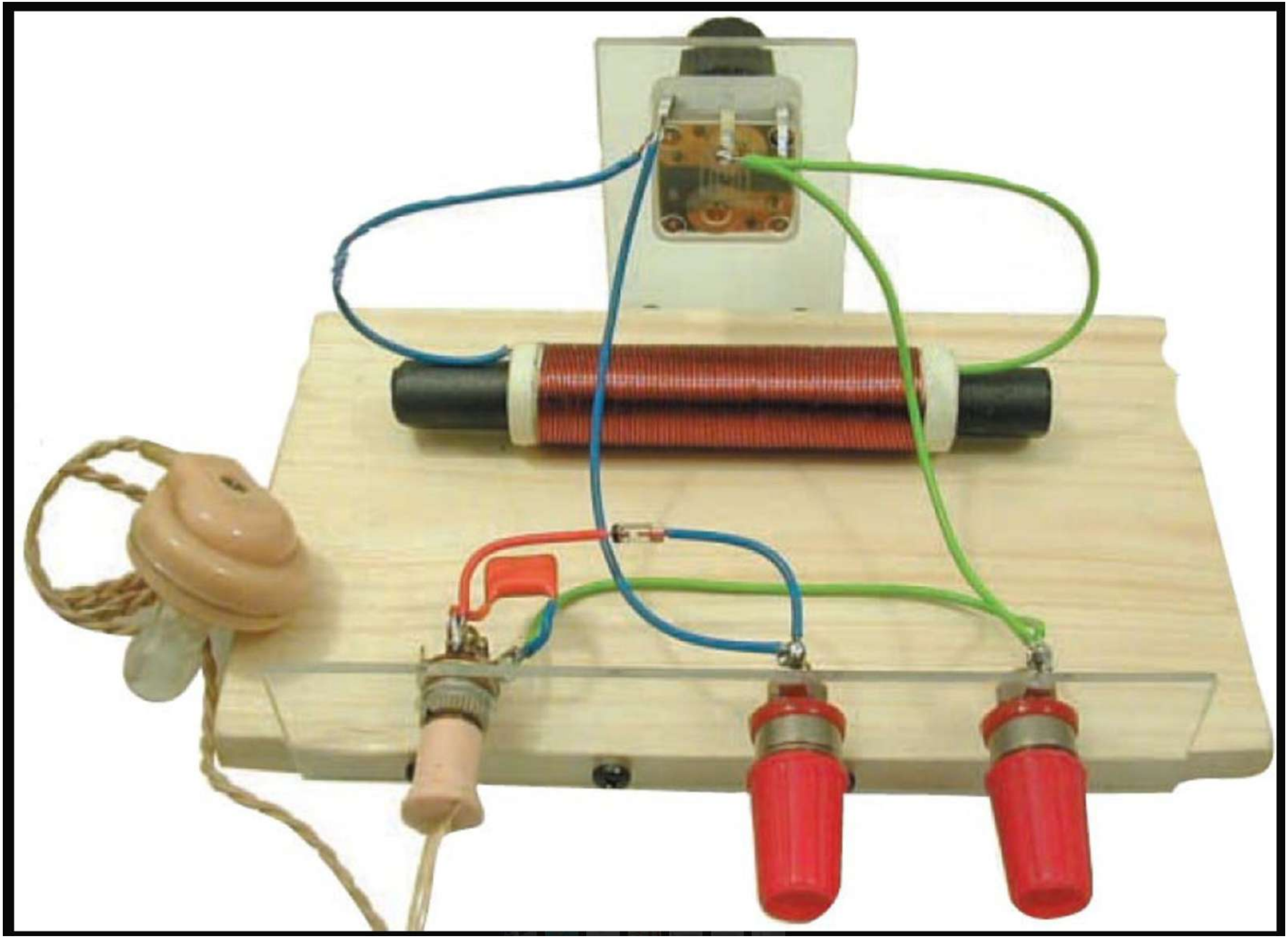

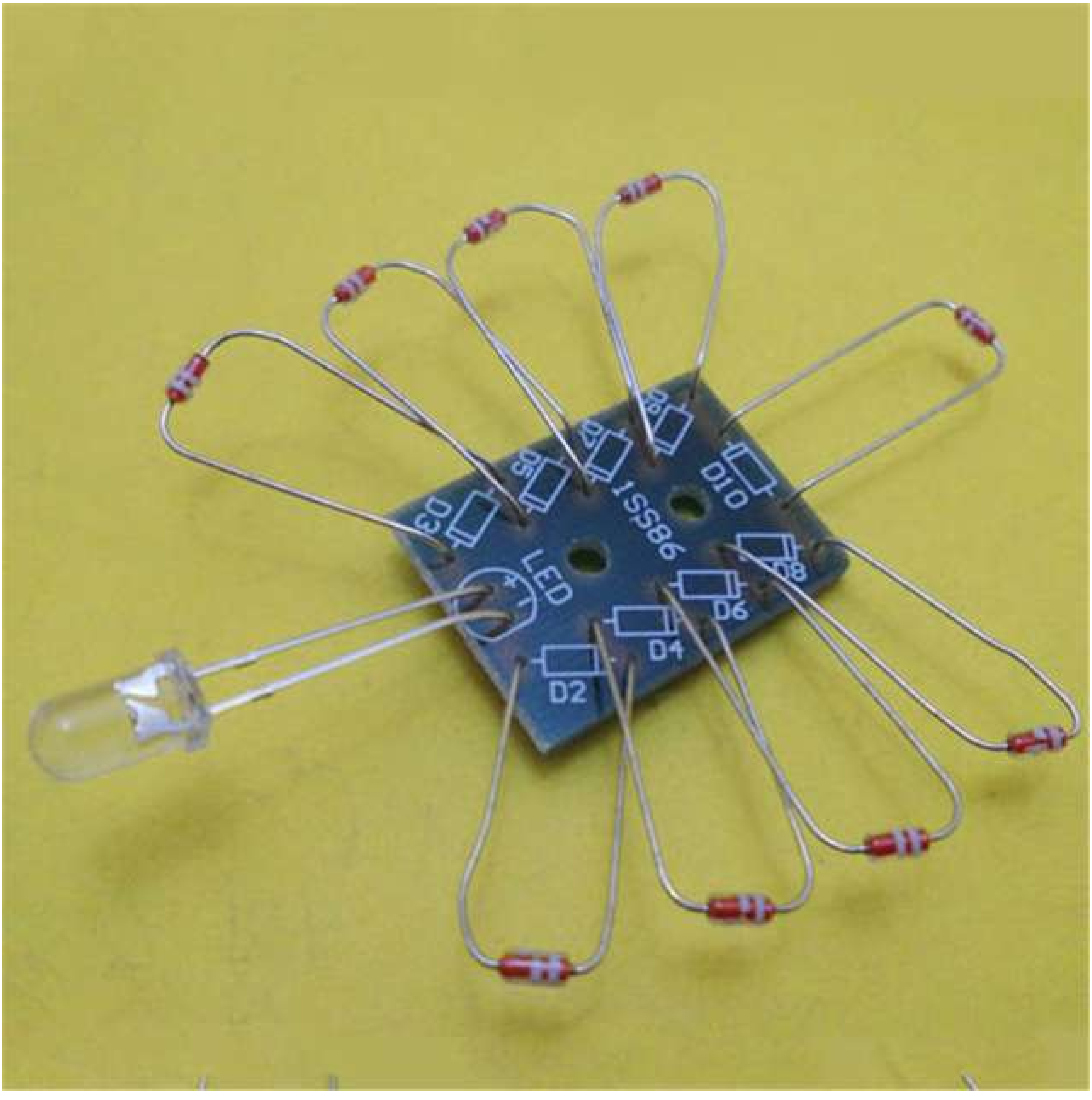
D3
D5
D7
D9
D10
1SS86
LED
D8
D6
D4
D2

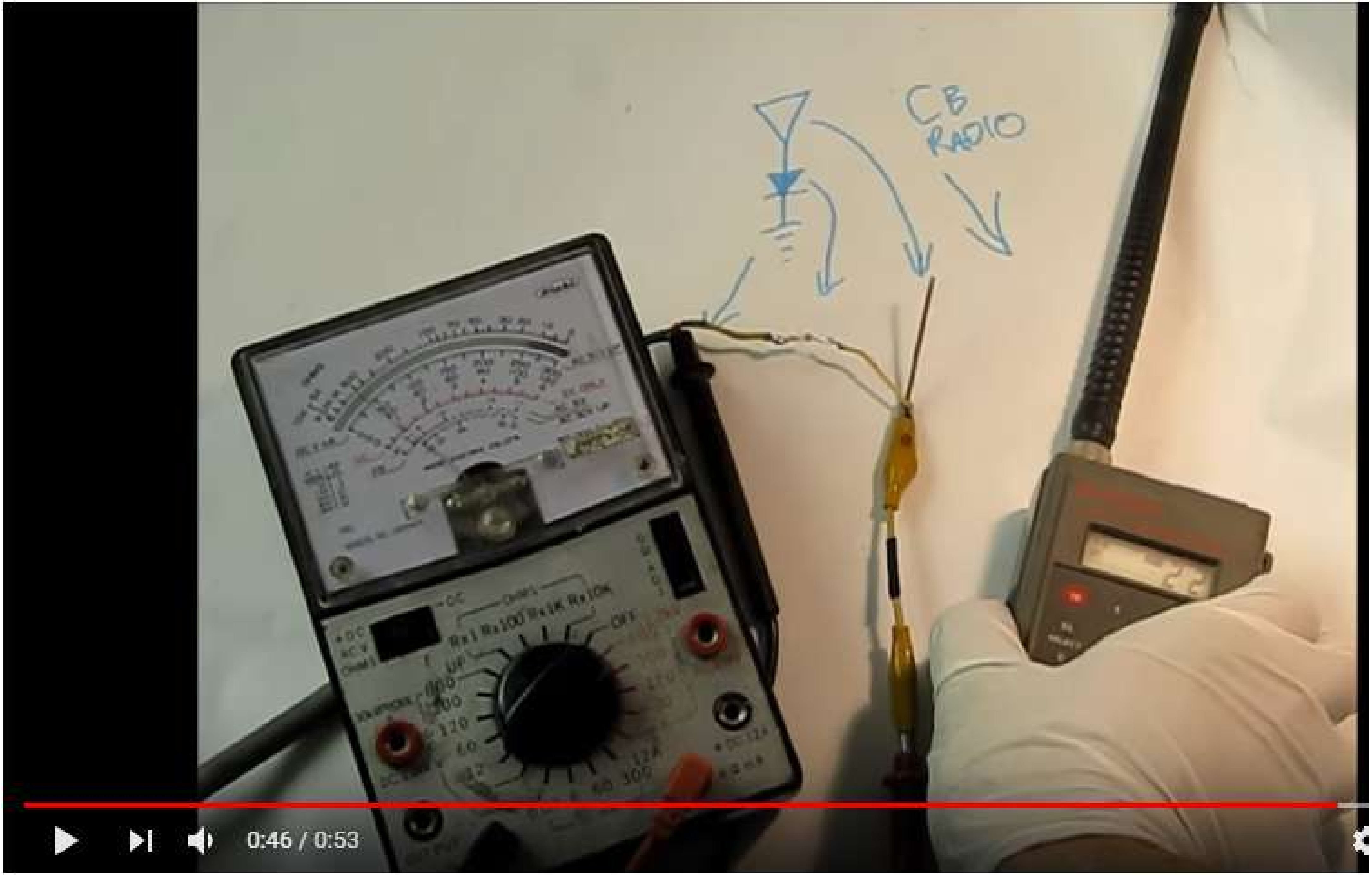
CB RADIO
0:46 / 0:53

DIY Wireless Power - Part 2: Simple Wireless Power Transmission! (CB)

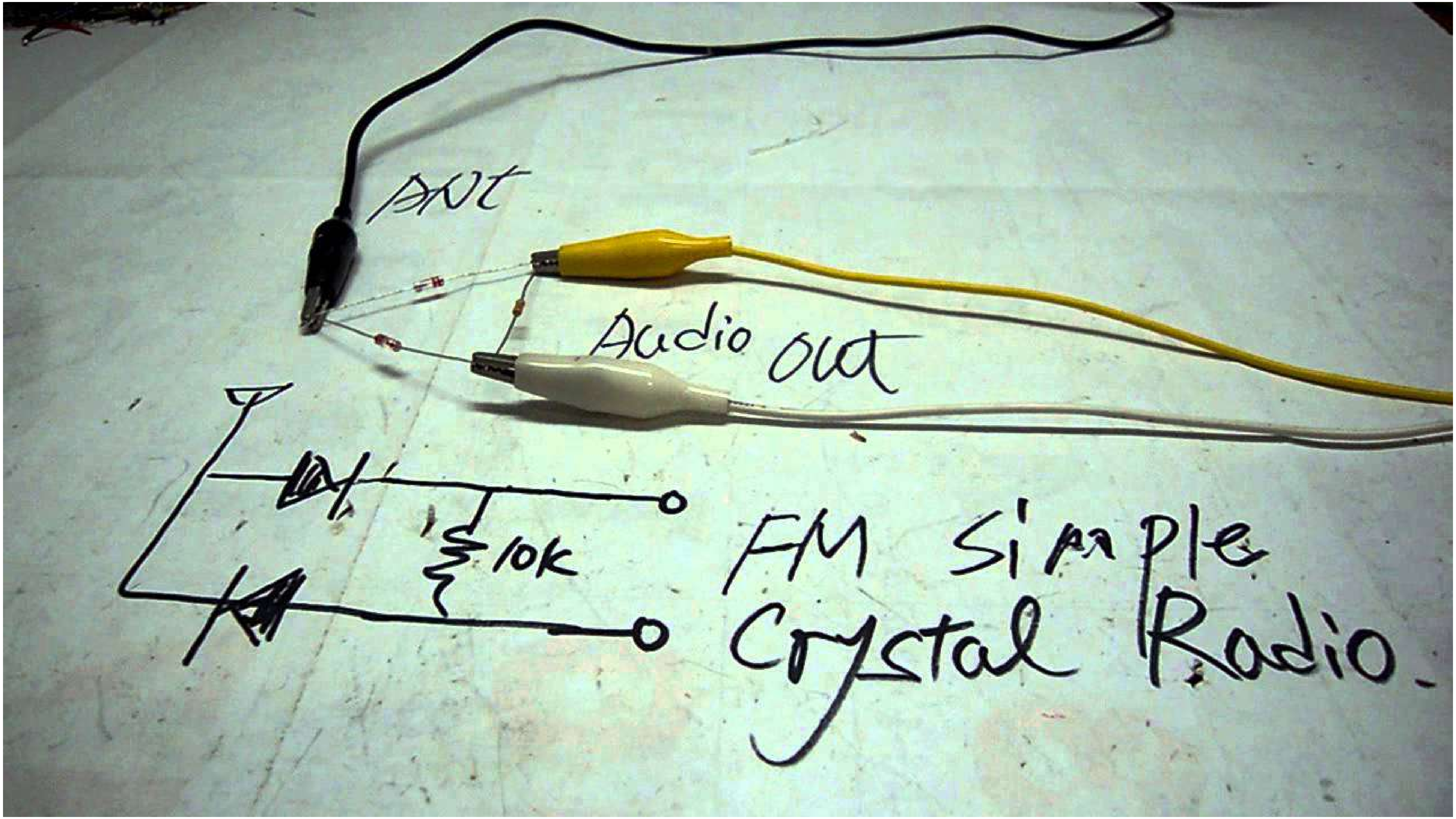
ANt
Audio out
10K
FM simple
Crystal Radio

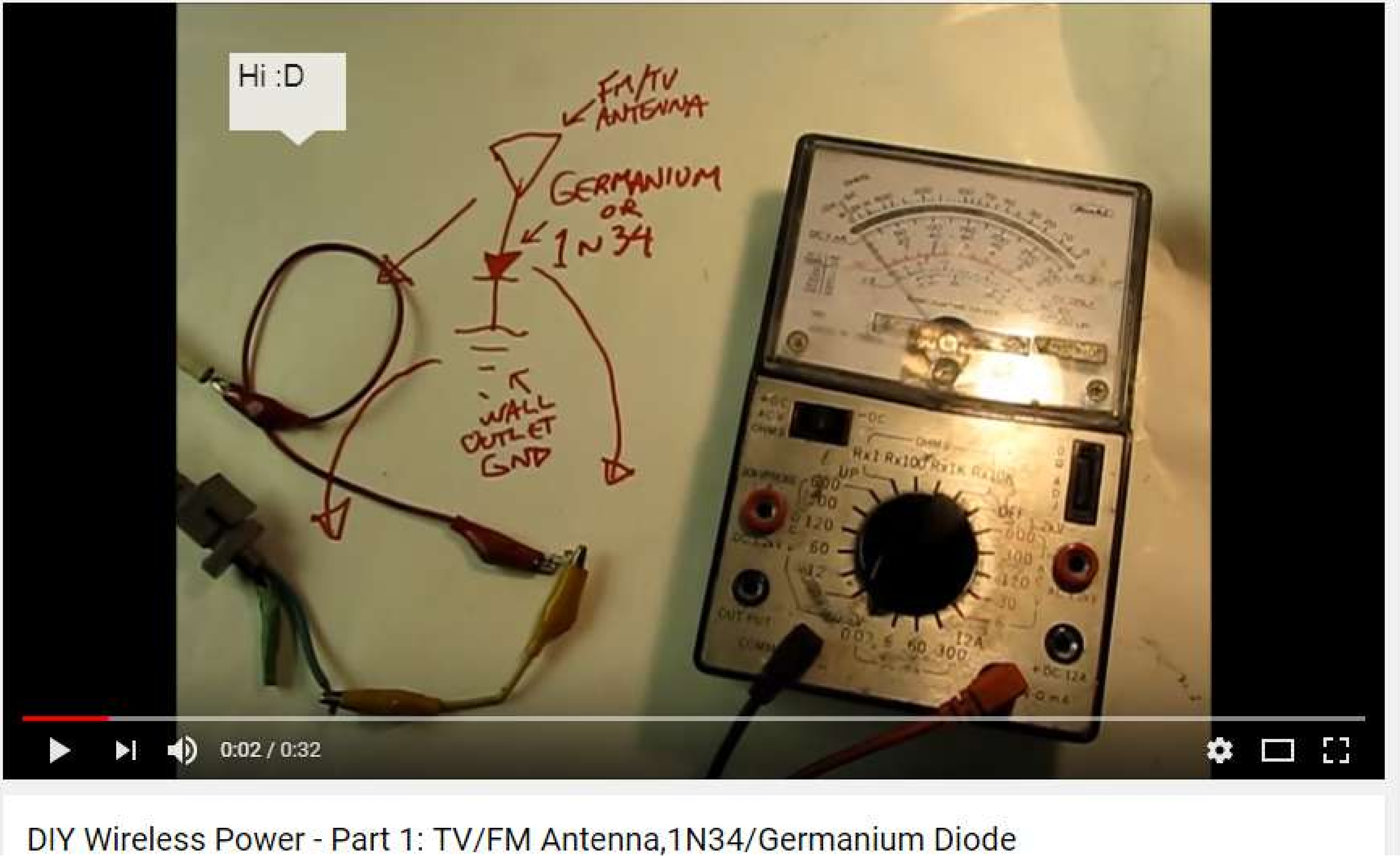
Hi :D
FM/TV ANTENNA
GERMANIUM OR 1N34
WALL OUTLET GND
0:02 / 0:32

DIY Wireless Power - Part 1: TV/FM Antenna,1N34/Germanium Diode

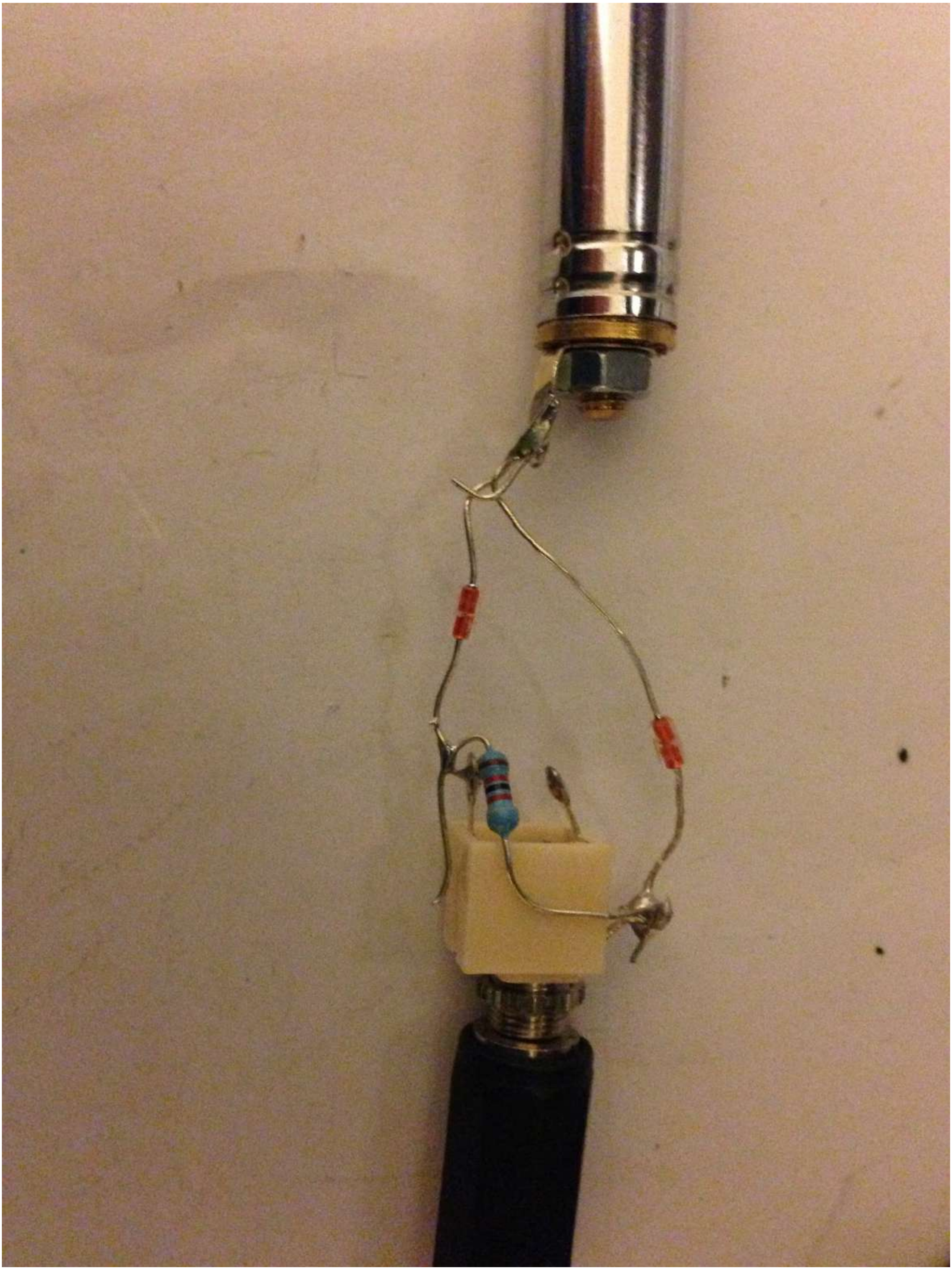

The simplest FM Crystal Radio Circuit - http://billydiy.blogspot.hk

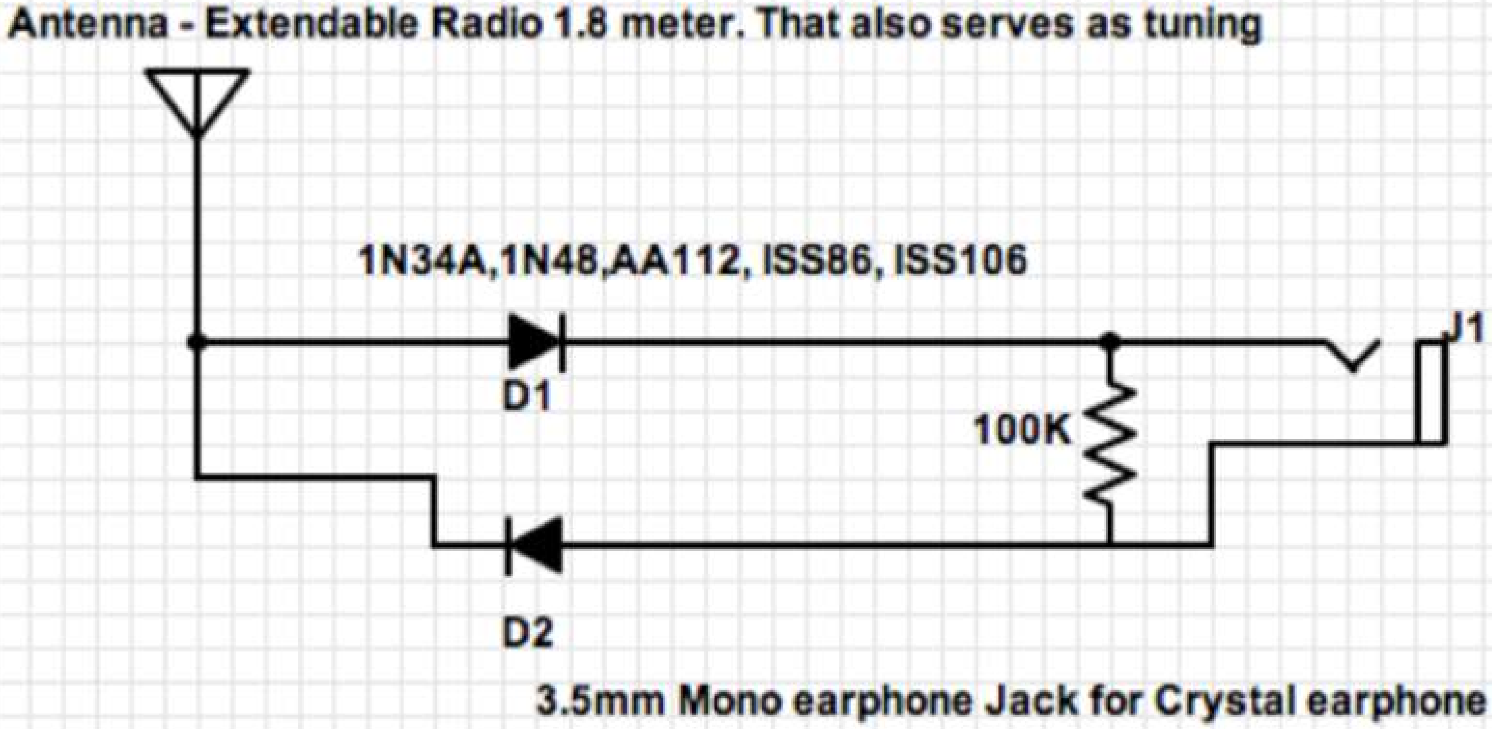
Antenna - Extendable Radio 1.8 meter. That also serves as tuning
1N34A,1N48,AA112, ISS86, ISS106
J1
D1
100K
D2
3.5mm Mono earphone Jack for Crystal earphone

Simple crystal receiver for FM

Carlo Bramanti

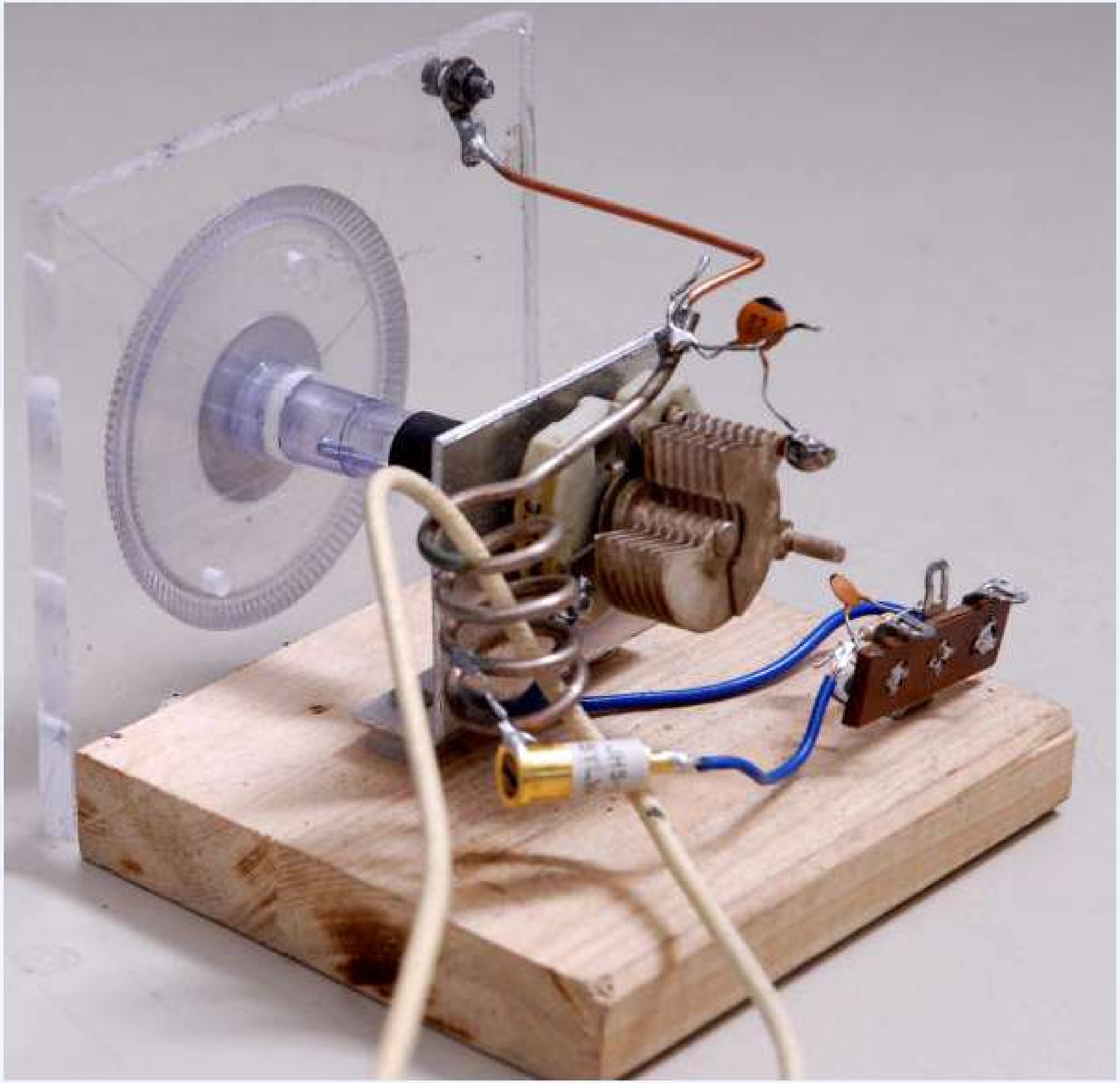

FM Crystal Radio Circuit

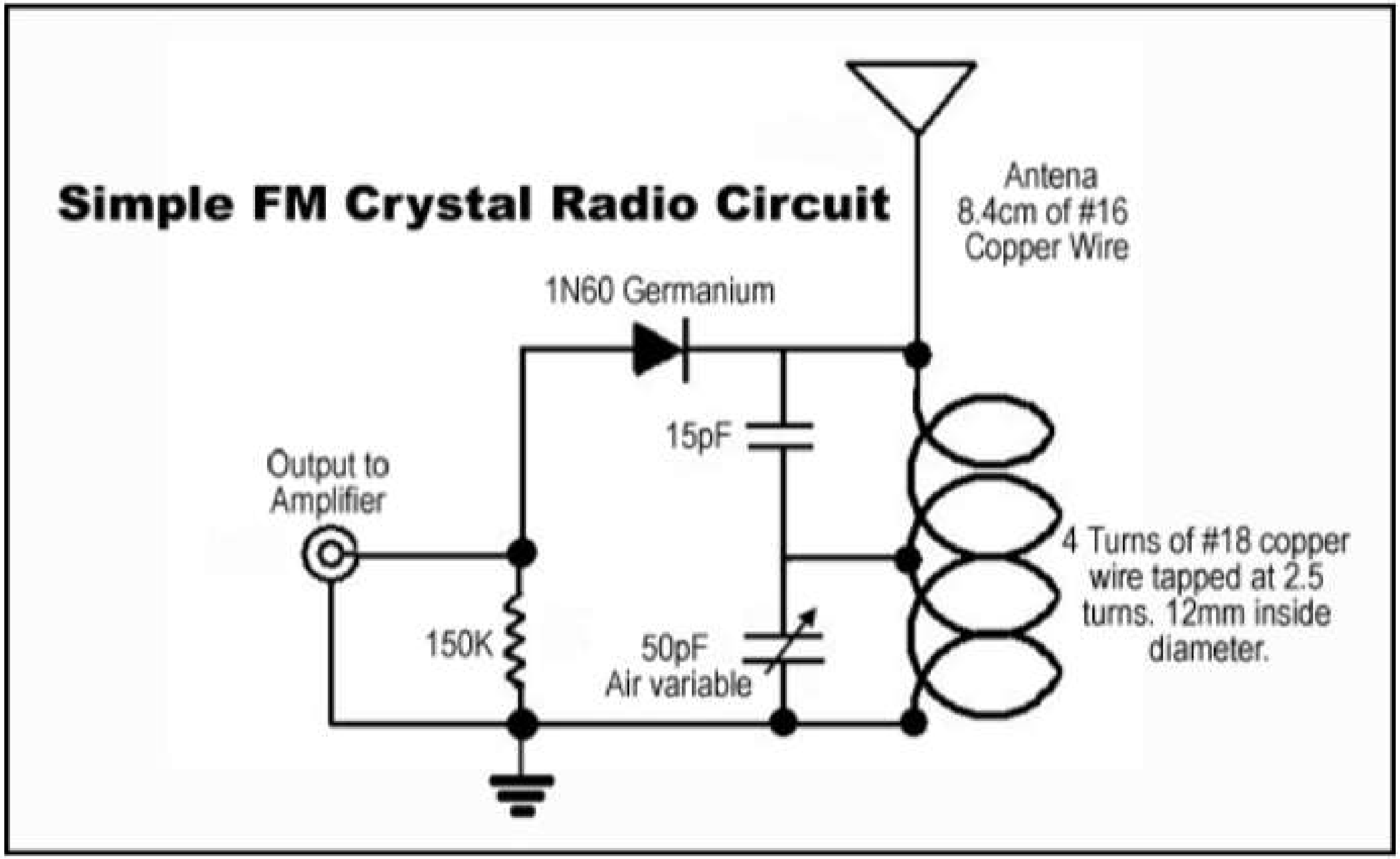
Simple FM Crystal Radio Circuit
Antena
8.4cm of #16
Copper Wire
1N60 Germanium
15pF
Output to
Amplifier
4 Turns of #18 copper
wire tapped at 2.5
turns. 12mm inside
diameter.
150K
50pF
Air variable

Parts List (some of these parts you can buy from our online store):

- 1N60 Germanium Diode
- 15pF Ceramic Capacitor
- 50pF Variable Capacitor
- 150K Ohm Resistor
- #16 & #18 Copper wires

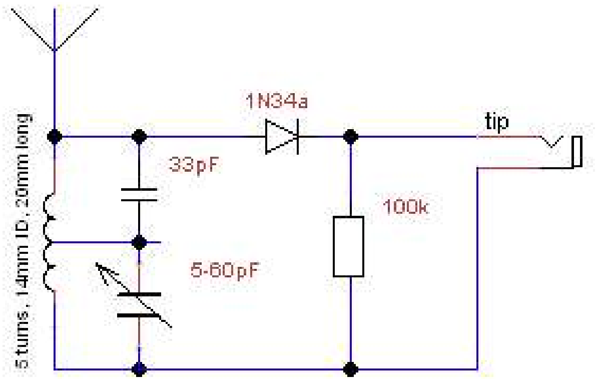
1N34a
tip
33pF
100k
5-60pF
5 turns, 14mm ID, 20mm long

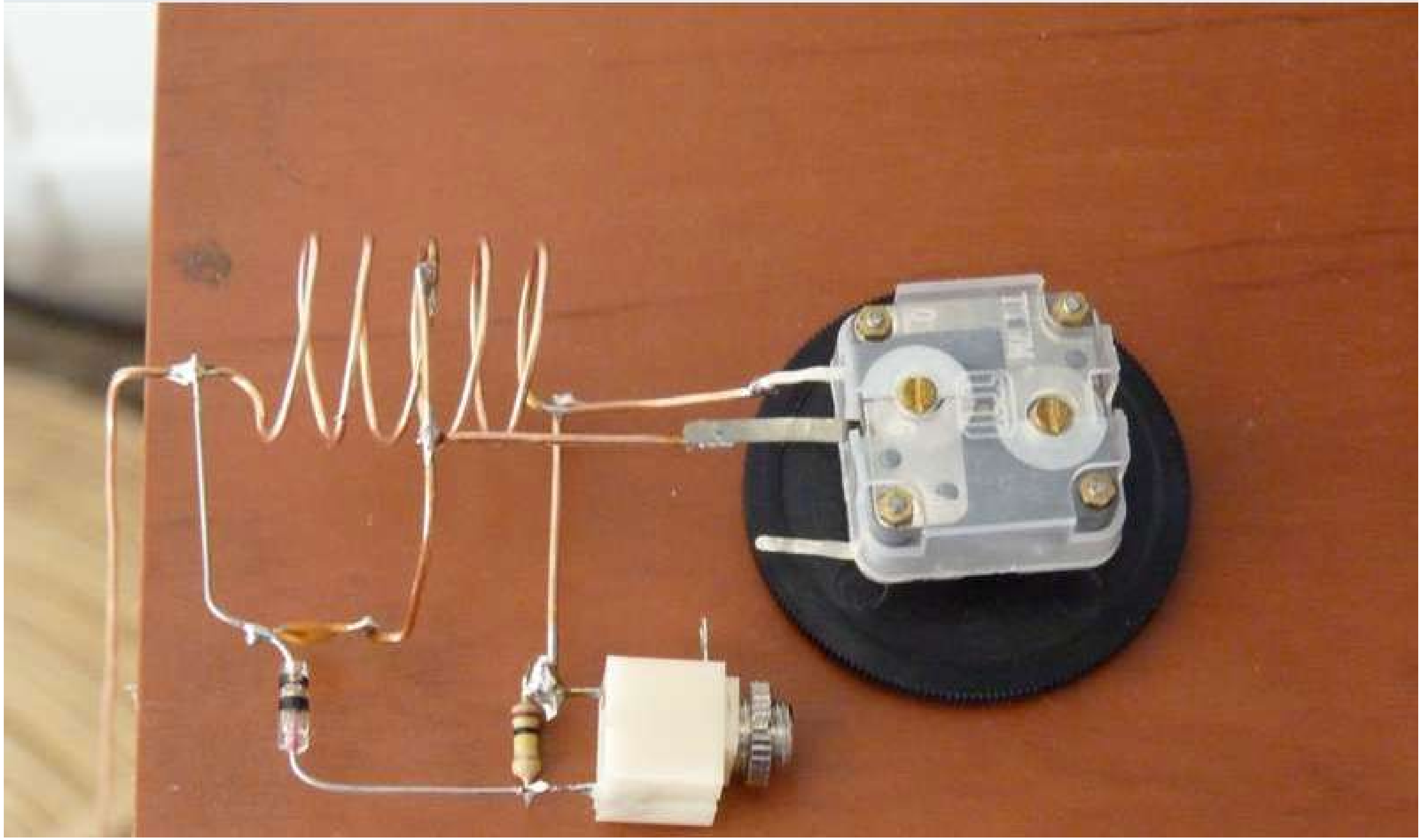

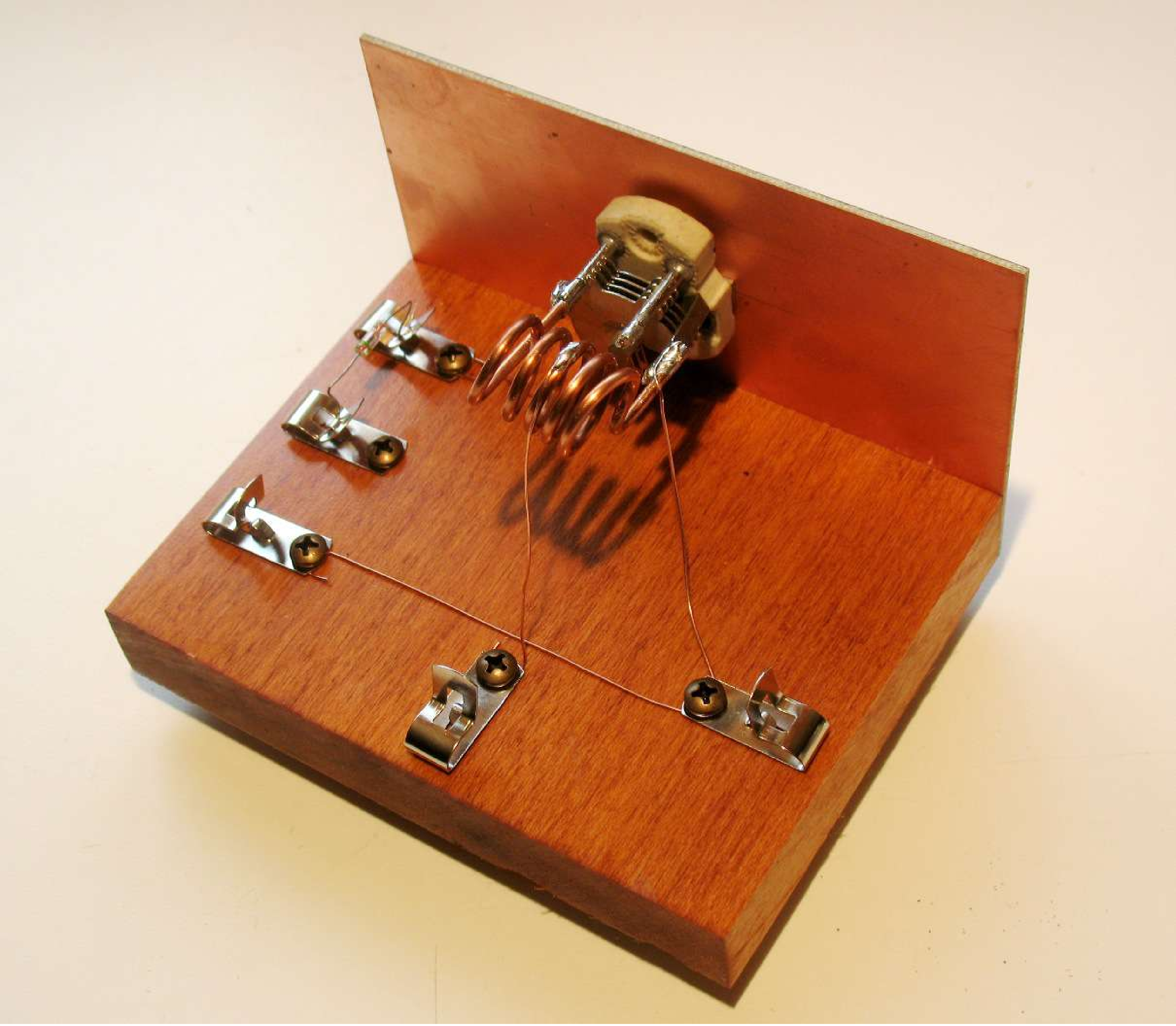

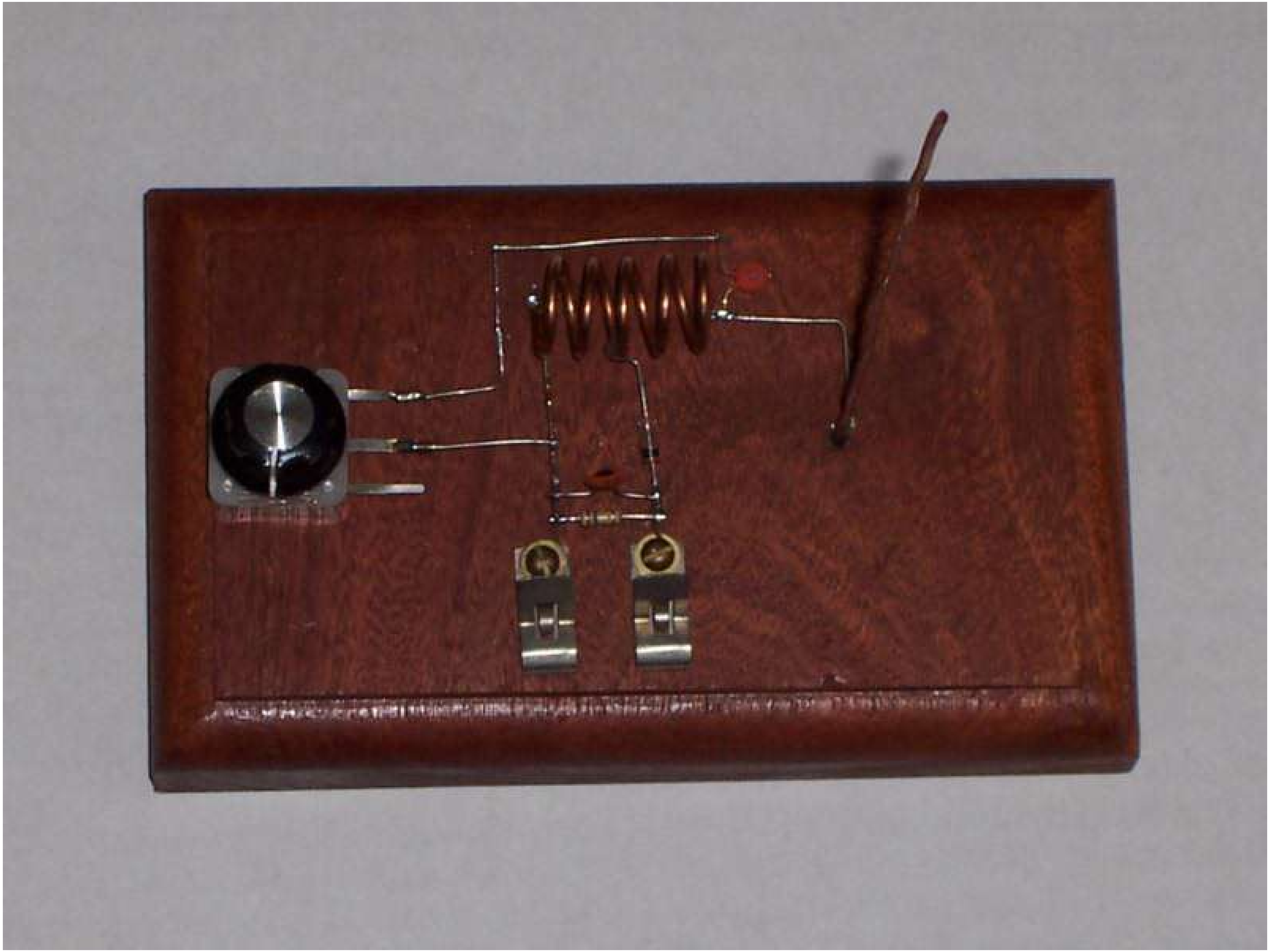

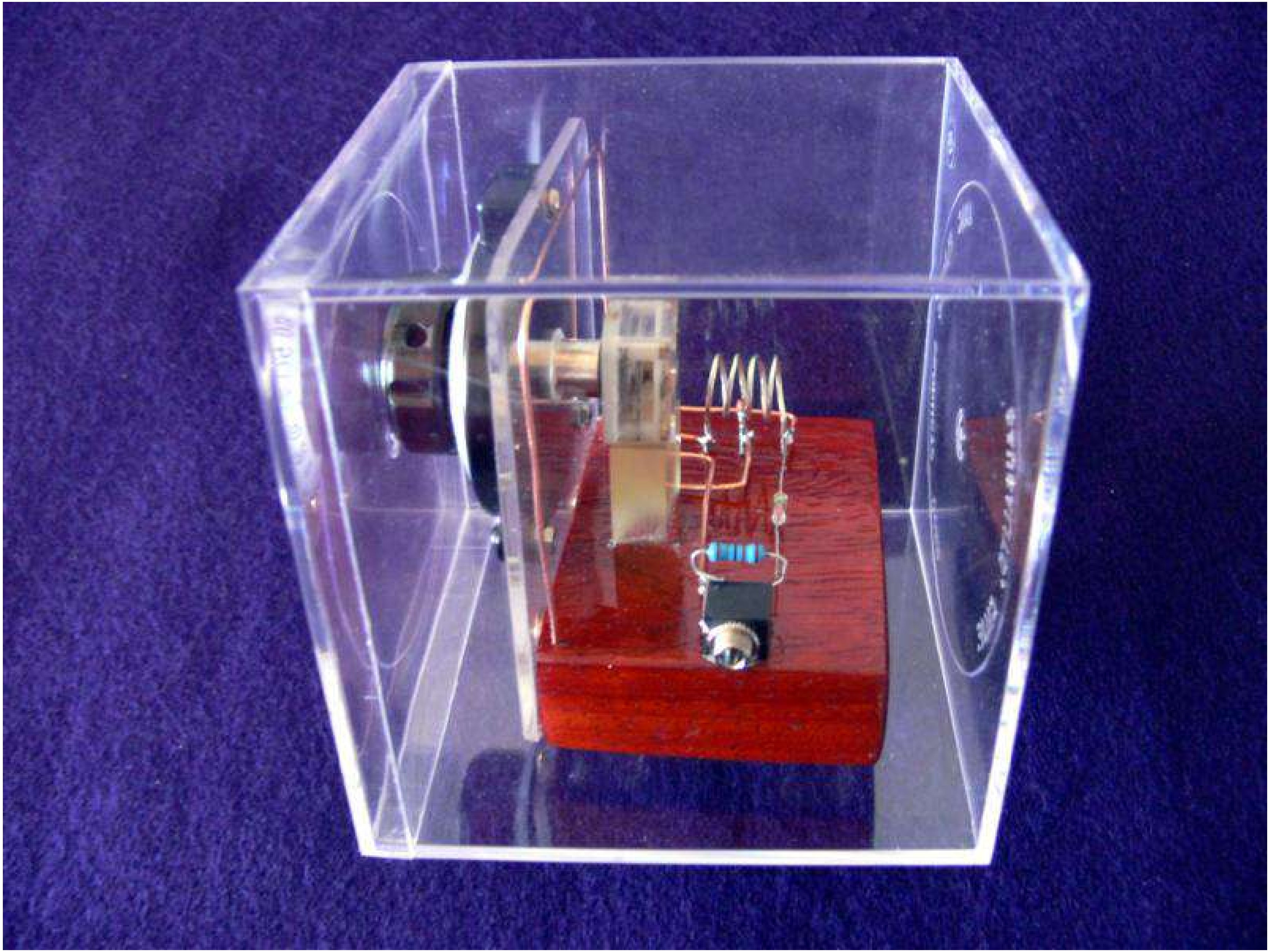

L - 4 turns #18 copper or silver wire, 12mm inside diameter, tapped at 2.5 turns
Ant - 7 inches of #18 bare copper wire
C1 - 18 pf ceramic capacitor
C2 - 50 pf air variable capacitor
D - 1N34 diode or rock crystal
R - 150K resistor

All passive components

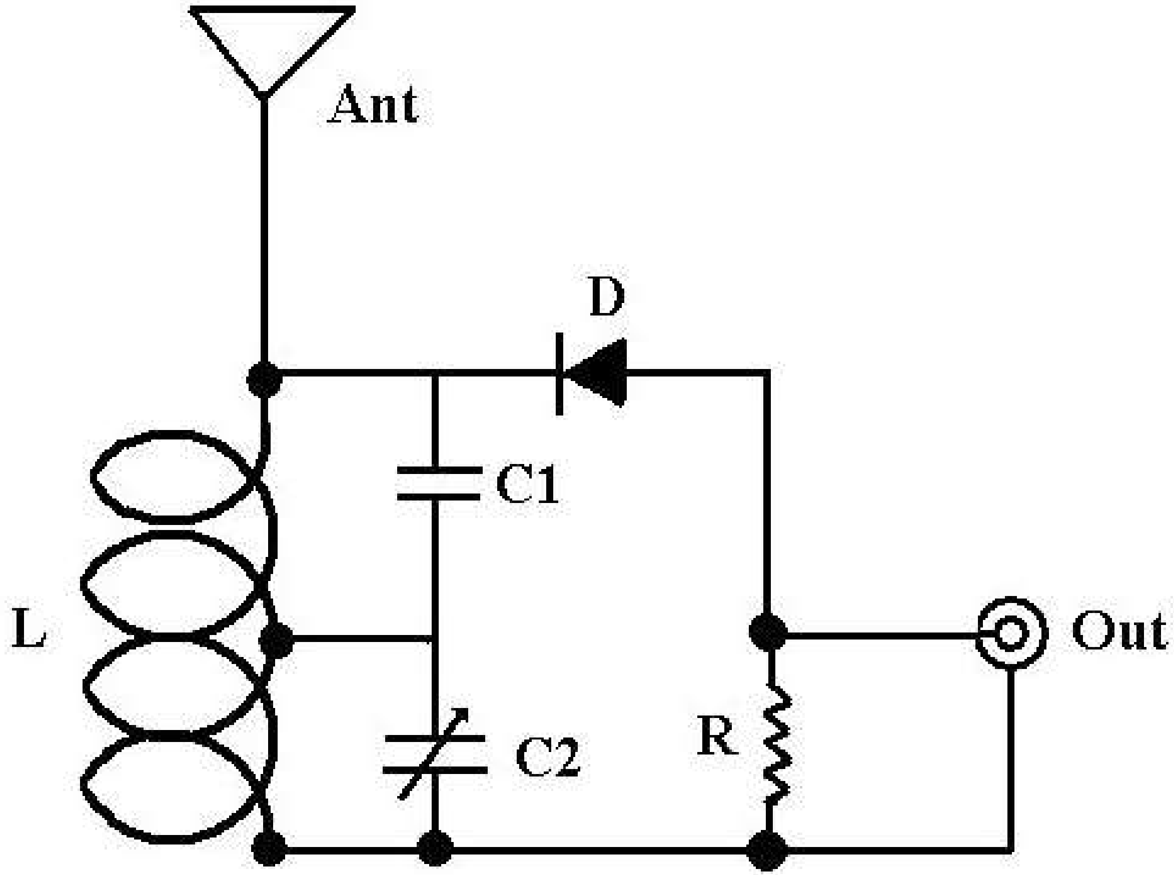

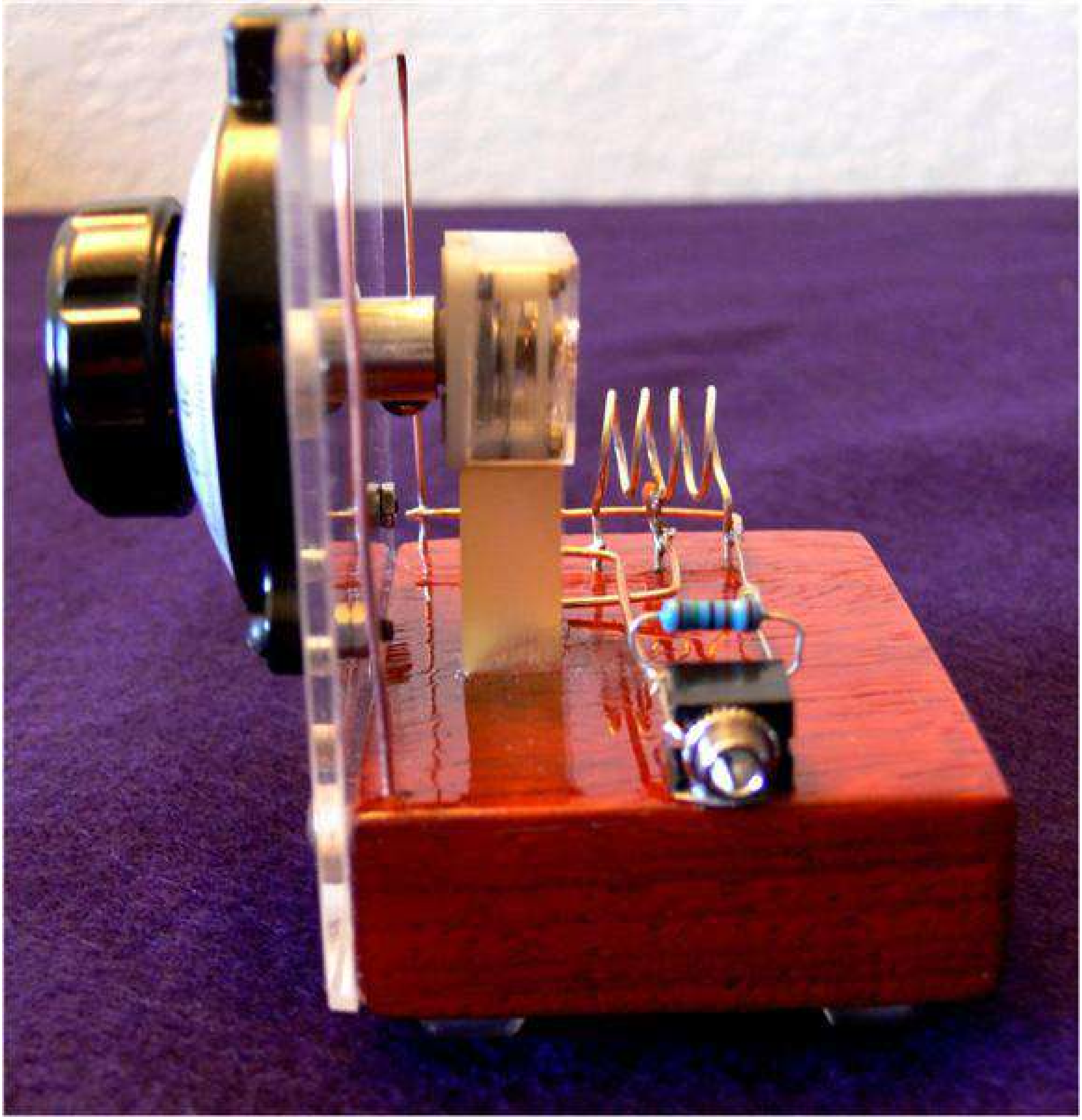

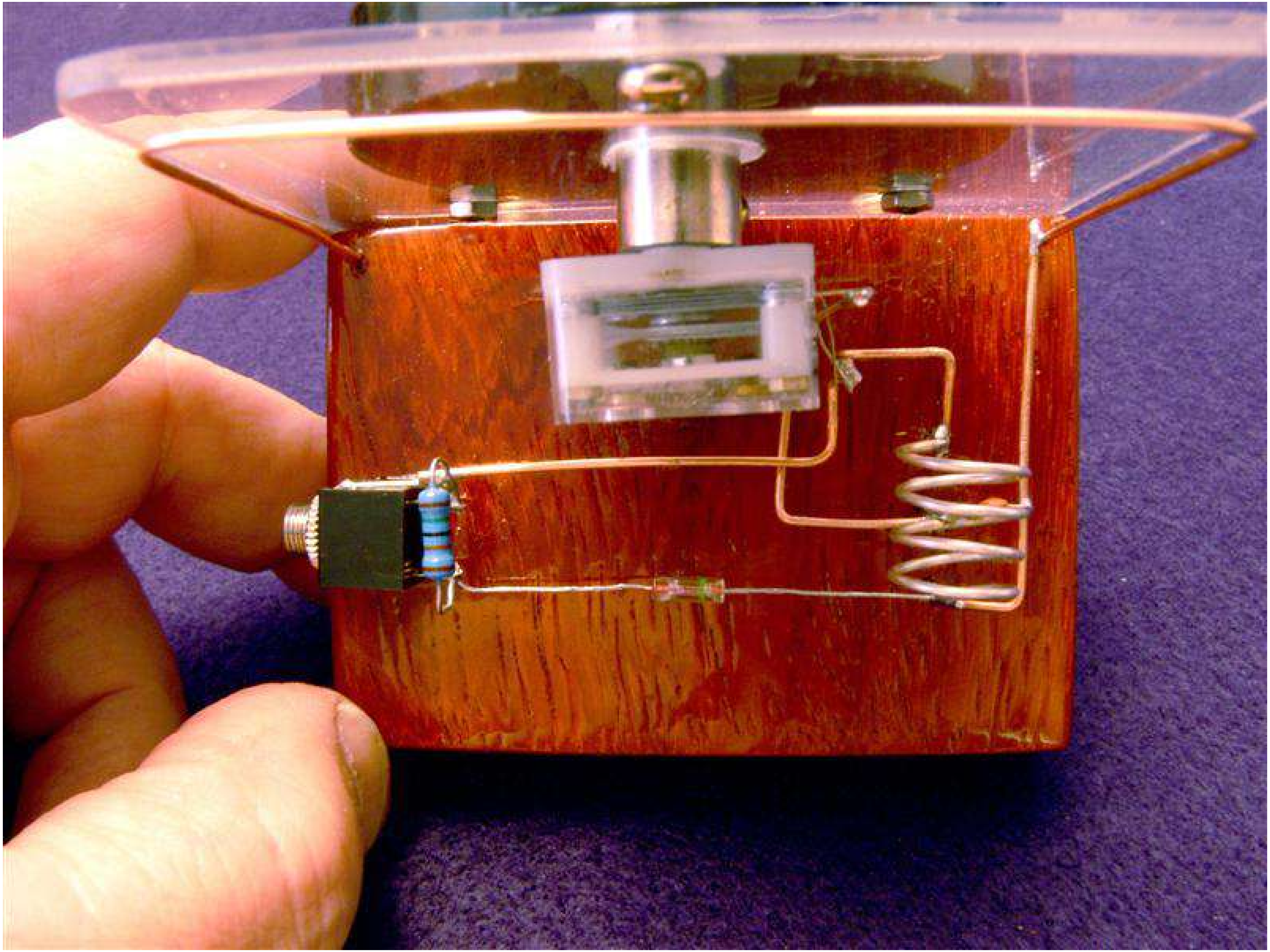

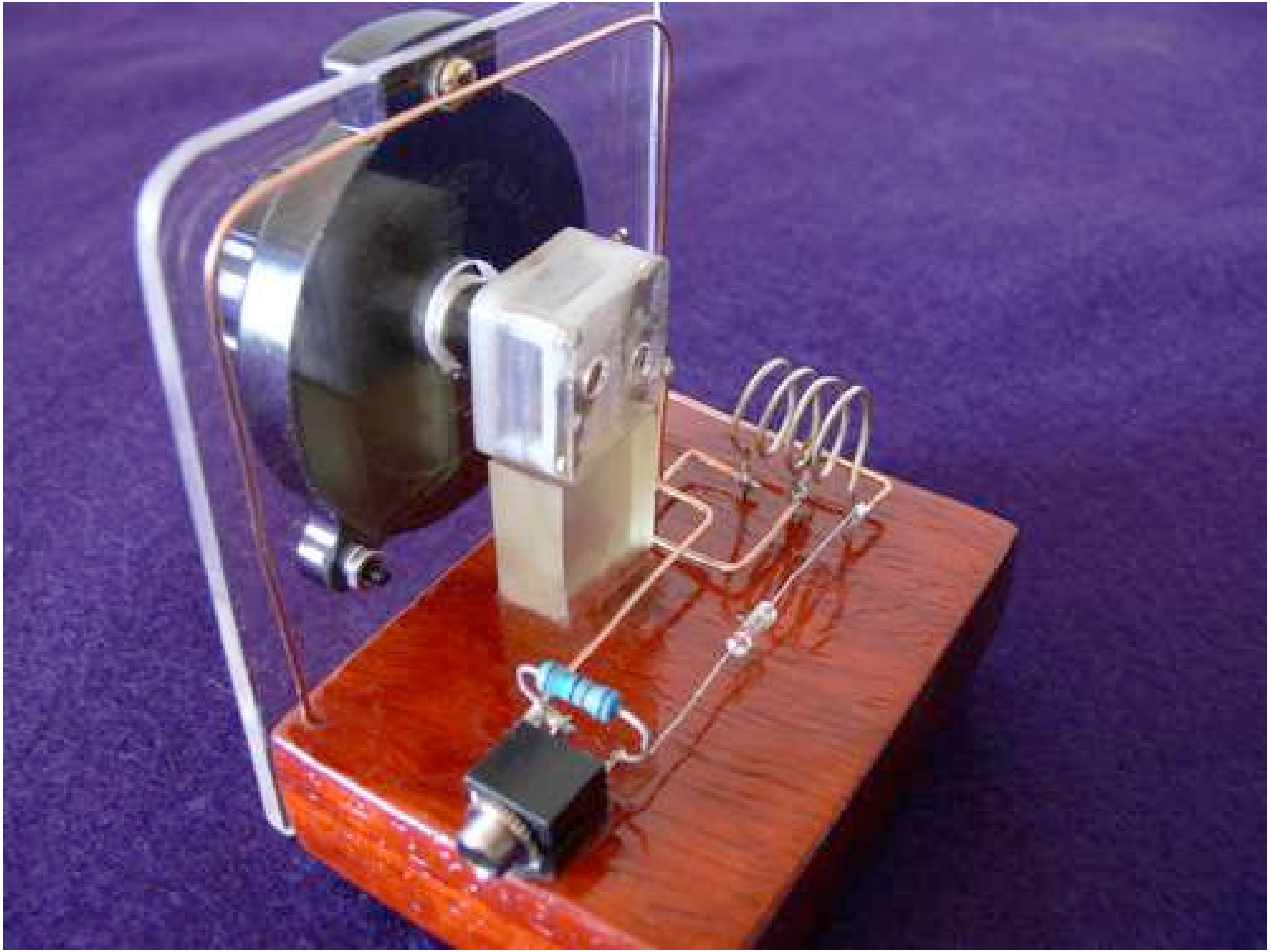

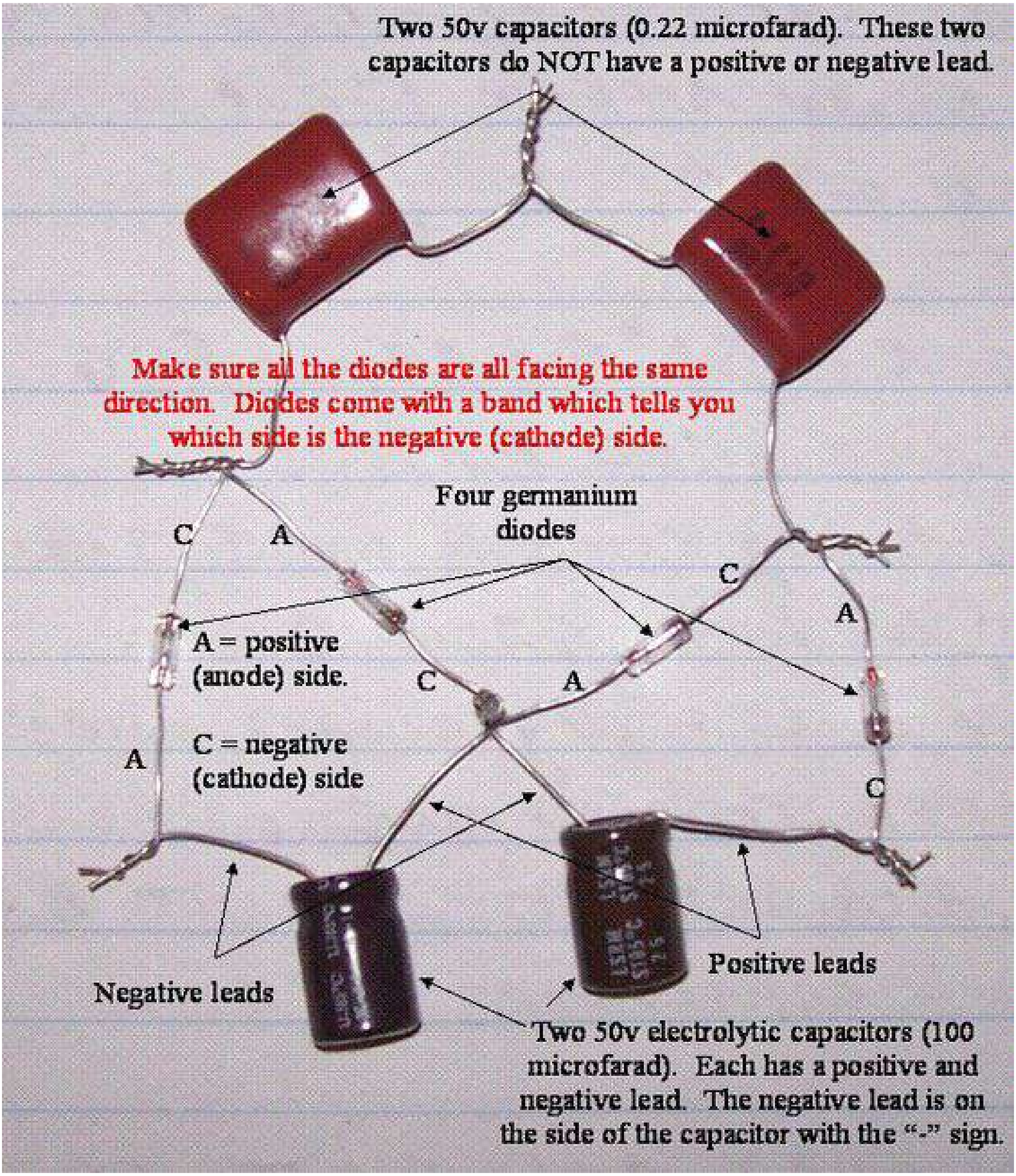

Two 50v capacitors (0.22 microfarad). These two capacitors do NOT have a positive or negative lead.

Make sure all the diodes are all facing the same direction. Diodes come with a band which tells you which side is the negative (cathode) side.

A = positive (anode) side.

C = negative (cathode) side

Two 50v electrolytic capacitors (100 microfarad). Each has a positive and negative lead. The negative lead is on the side of the capacitor with the "-" sign.

Four germanium diodes
C
A
C
A
C
A
A
C
Negative leads
Positive leads

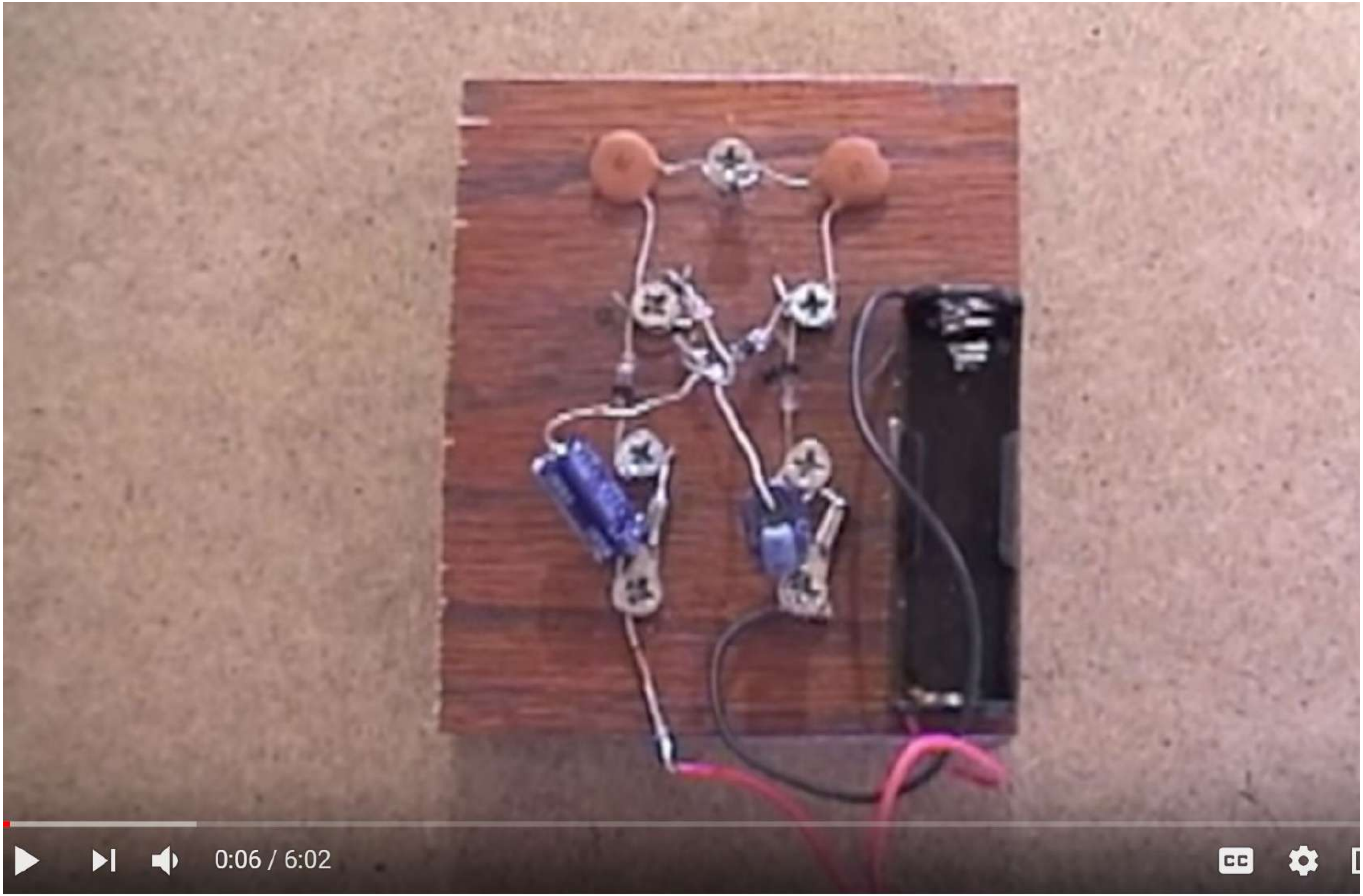
0:06 / 6:02
CC

FIG. 10 BREADBOARD

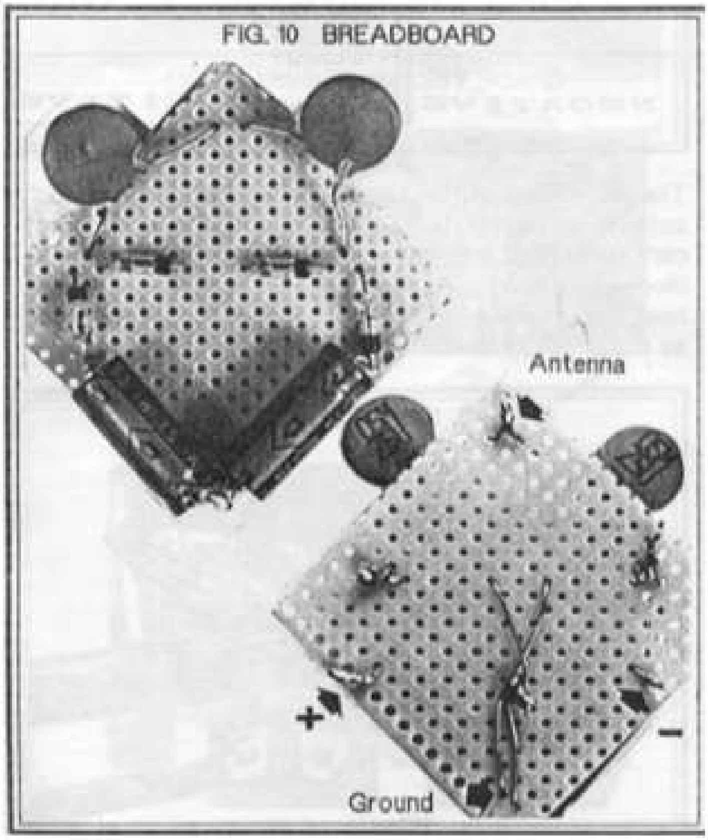
Antenna
+
−
Ground

the board as shown in Figure 11. This unit is powering a small red LED indicated by the arrow.

FIG. 11 EXPERIMENT BOARD

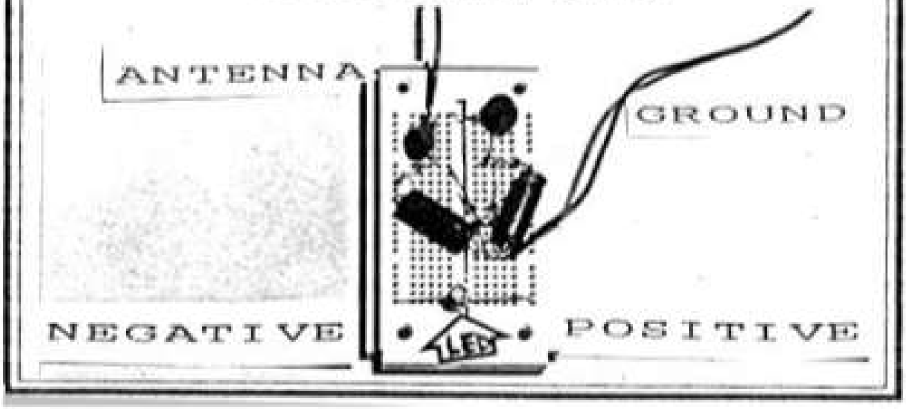
ANTENNA
GROUND
NEGATIVE
LED
POSITIVE

ıade to the twisted ends of the ceramic capacitors. When soldering the leads of the 1N34 diodes, care must be taken t

FIG. 9 TERMINAL STRIP

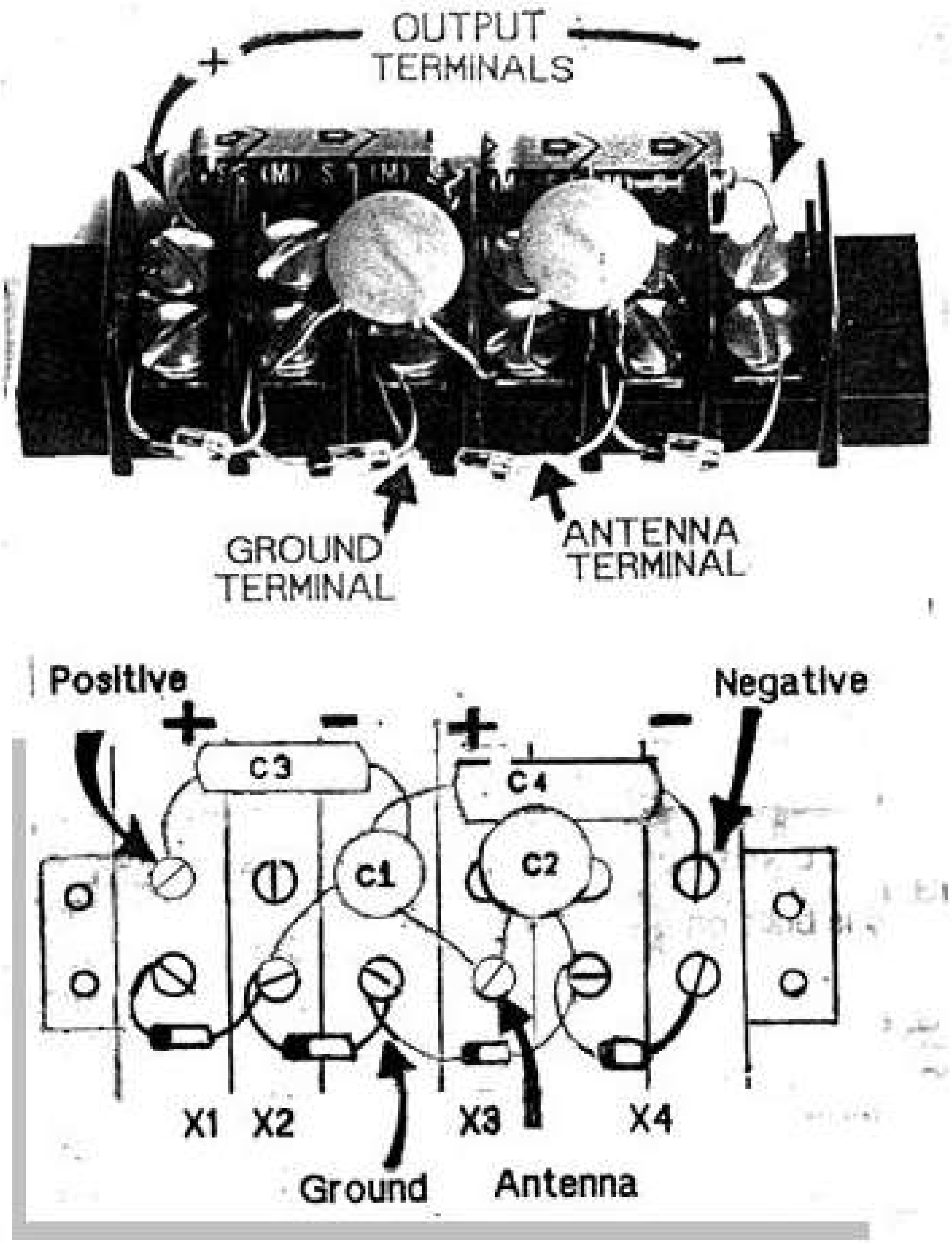
+
OUTPUT TERMINALS
−
GROUND TERMINAL
ANTENNA TERMINAL
Positive
Negative
+
−
+
−
C3
C4
C1
C2
X1
X2
X3
X4
Ground
Antenna

Amplifier circuit. The crystal radio used for example purposes is the crystal radio made from scraps here.

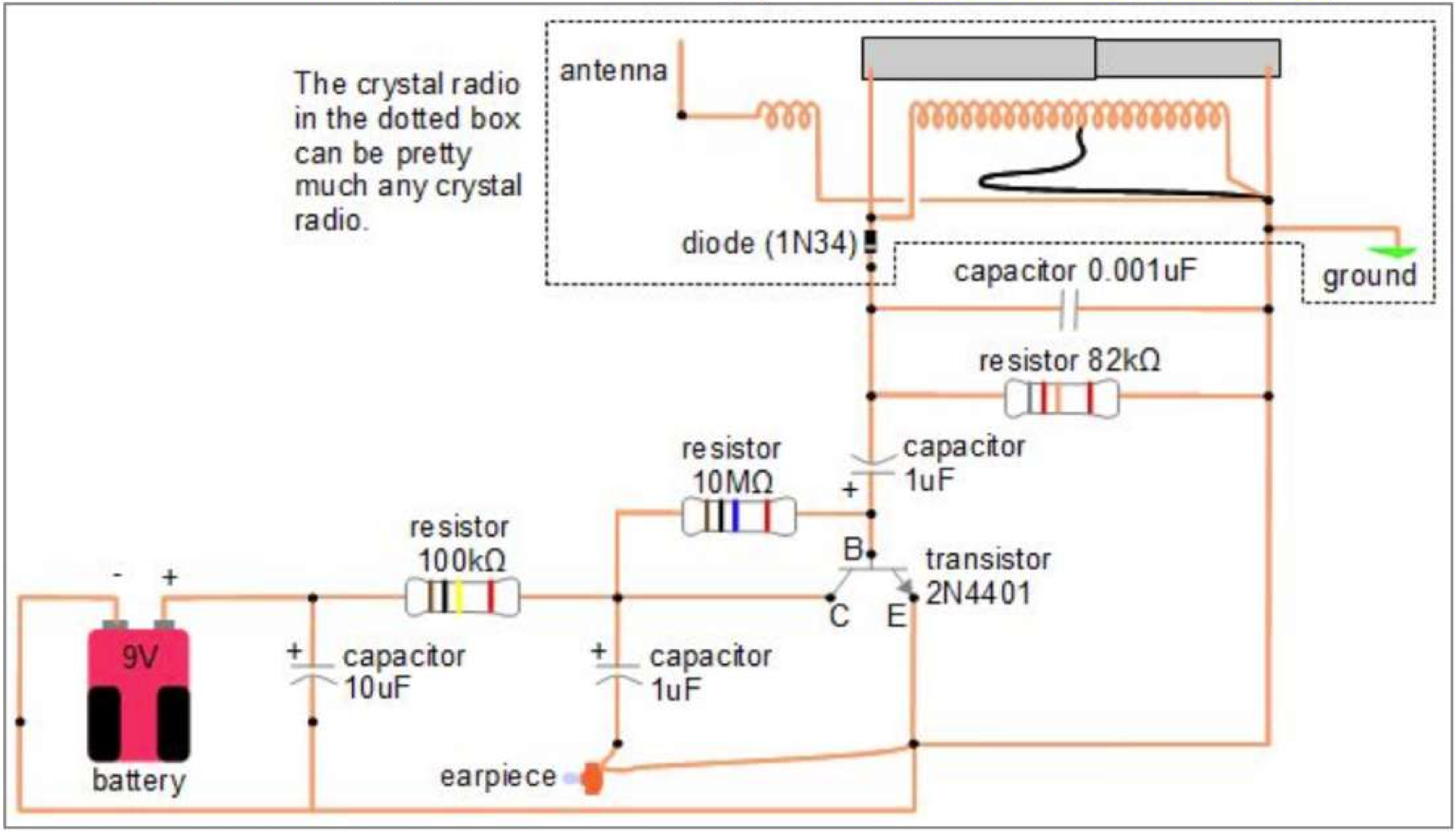
The crystal radio in the dotted box can be pretty much any crystal radio.
antenna
diode (1N34)
capacitor 0.001uF
ground
resistor 82kΩ
resistor 10MΩ
capacitor 1uF
+
B
transistor 2N4401
C
E
resistor 100kΩ
-
+
9V
battery
+
capacitor 10uF
+
capacitor 1uF
earpiece

SATURDAY, NOVEMBER 13, 2010

SIMPLE FIELD STRENGTH METER

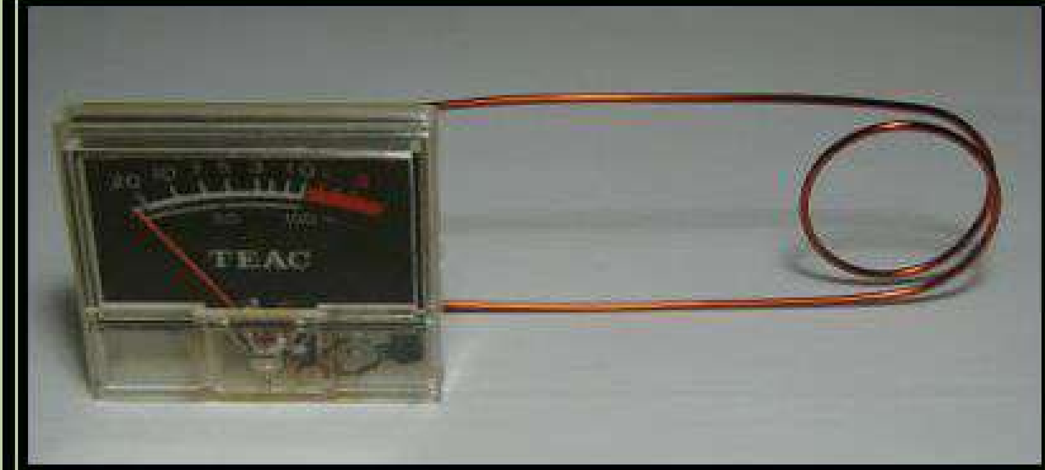

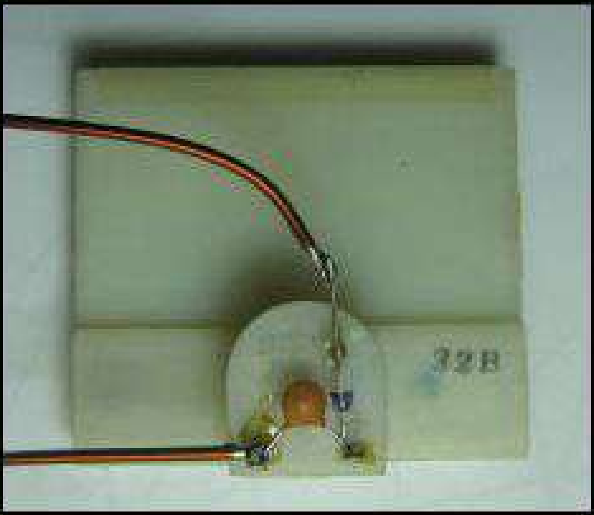

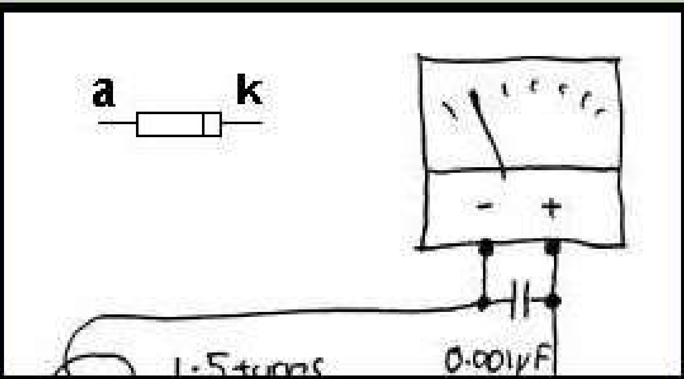
a
k
-
+
0.001μF
1.5turns

MY

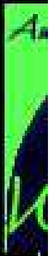

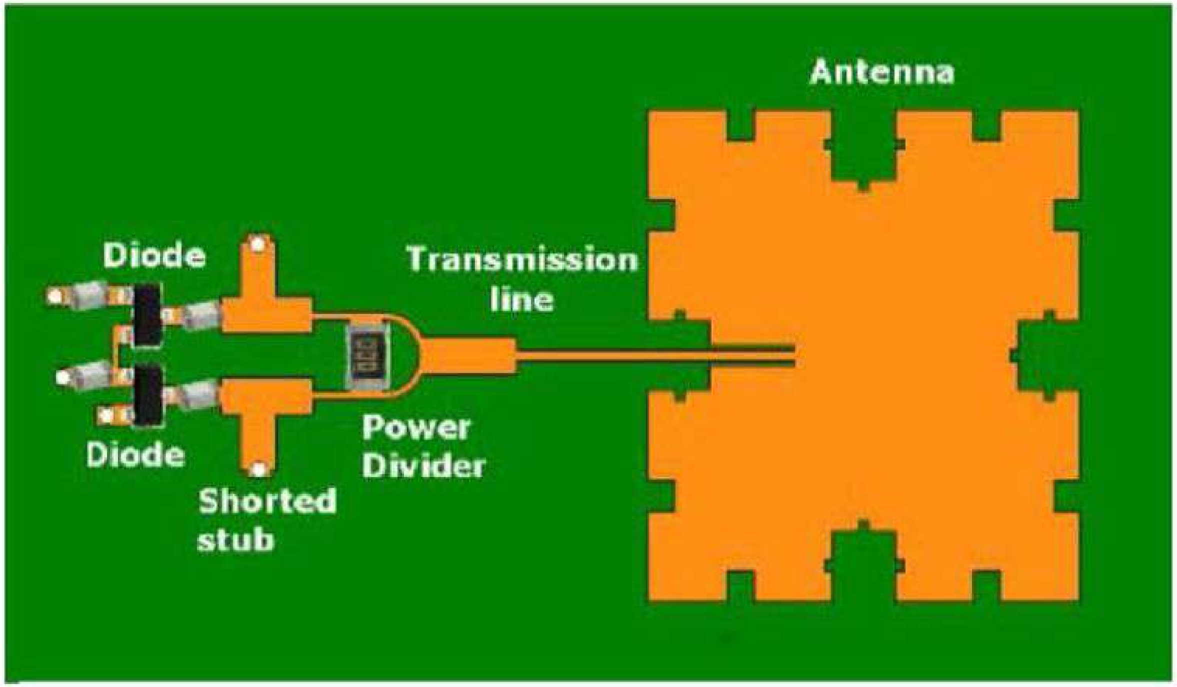
Antenna
Diode
Transmission line
Diode
Power Divider
Shorted stub

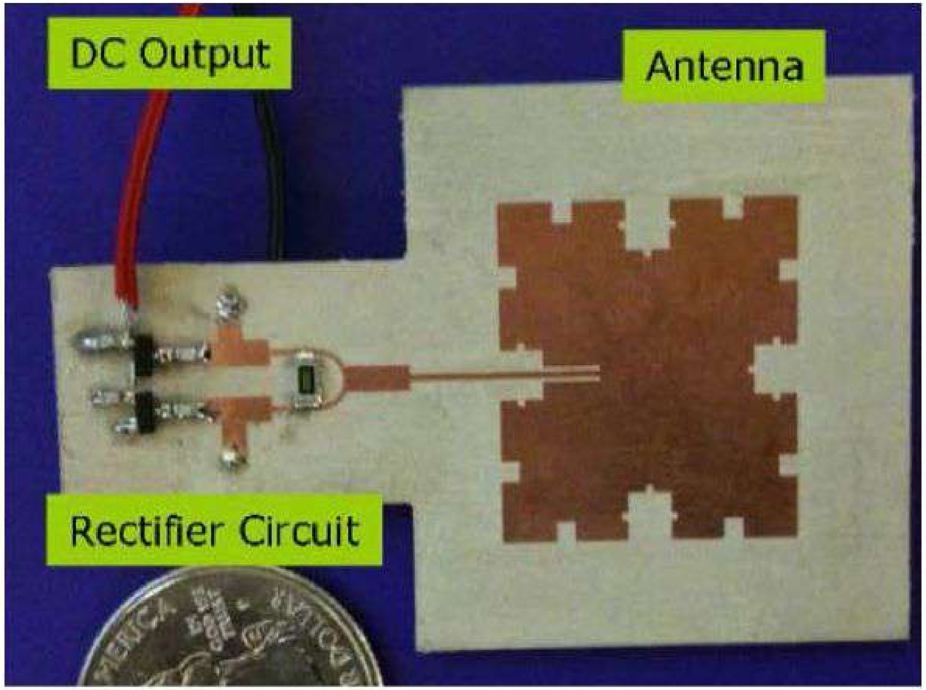
DC Output
Antenna
Rectifier Circuit

4. Results and Analysis

The simulated and measured results at the output voltage of voltage multiplier circuit are shown graphically in **Figure 8**. From the graph analysis, the simulated and the measured results agree considerably with each other. The measured results are shown to be better than the simulation results. The reason behind this may be due to the uncertainty in series resistance value of the diode obtained from SPICE parameters in modeling as explained in Equation (5). This resistance vale of diodes in practical circuit may be lower than in the model, which provides fast discharge path, in turn rise in voltage as passes through the stages and reaches to final output. In this work, the DC output voltages obtained through simulation and measurement at 0 dBm re 2.12 V and 5.0 V respectively. These results are comparatively much better than in ref. [9], where in at 0 dBm, 900 MHz they achieved 0.5 V and 0.8 V through simulation and measurement increasing to 1.4 V, 1.67 V, 1.87 V and 2.12 V for 4, 5, 6 and 7 stages respectively compared to 2 mS as shown in [10]. **Figure 12** shows that the conversion ratio of 22 is achieved at 0 dBm input power and drops to 2.5 at –40 dBm. The highest value at 0 dBm is due to the innate characteristics of the zero bias Schottky diodes which conduct fairly well at higher input voltages.

5. Conclusion

From the experimental results, it is found that the pro-

Table 2. Component used in 7 stage voltage multiplier.

Name of component	Label	Value
Stage capacitors	C_1 - C_{14}	3.3 nF
Stage diodes	D_1 - D_{14}	HSMS 2850
Filter capacitor	C_L	100 nF
Load resister	R_L	100 kΩ

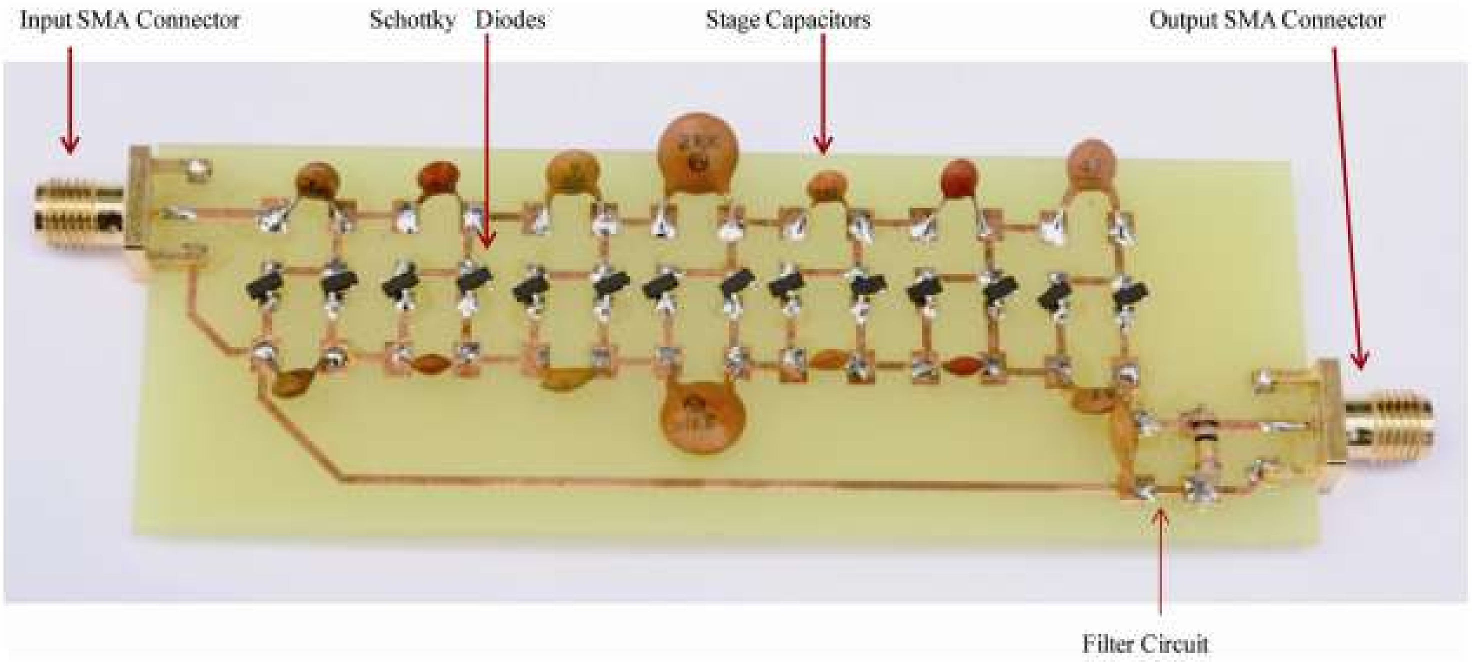

Figure 7. Photograph of assembled circuit board.

sign in this paper uses a capacitor across the load to store and provide DC leveling of the output voltage and its value only affects the speed of the transient response. Without a capacitor across the load, the output is not a good DC signal, but more of an offset AC signal.

In addition to the above, an equivalent load resistor is connected at the final node. The output voltage across the load decreases during the negative half cycle of the AC input signal. The voltage decreases is inversely proportional to the product of resistance and capacitance across the first stage was 3.3 nF, the second stage was 1.65 nF, third stage was 825 pF, fourth stage was 415 pF and so on. But keeping in view of testing, the capacitance values were chosen to have a close match with the standard available values in the market.

Simulation was carried out through 4 to 9 voltage doubler stages. Based on results obtained a 7 stage doubler is best to implemented for this application.

The design of the printed circuit board (PCB) was carried out using DipTrace software. The material used to

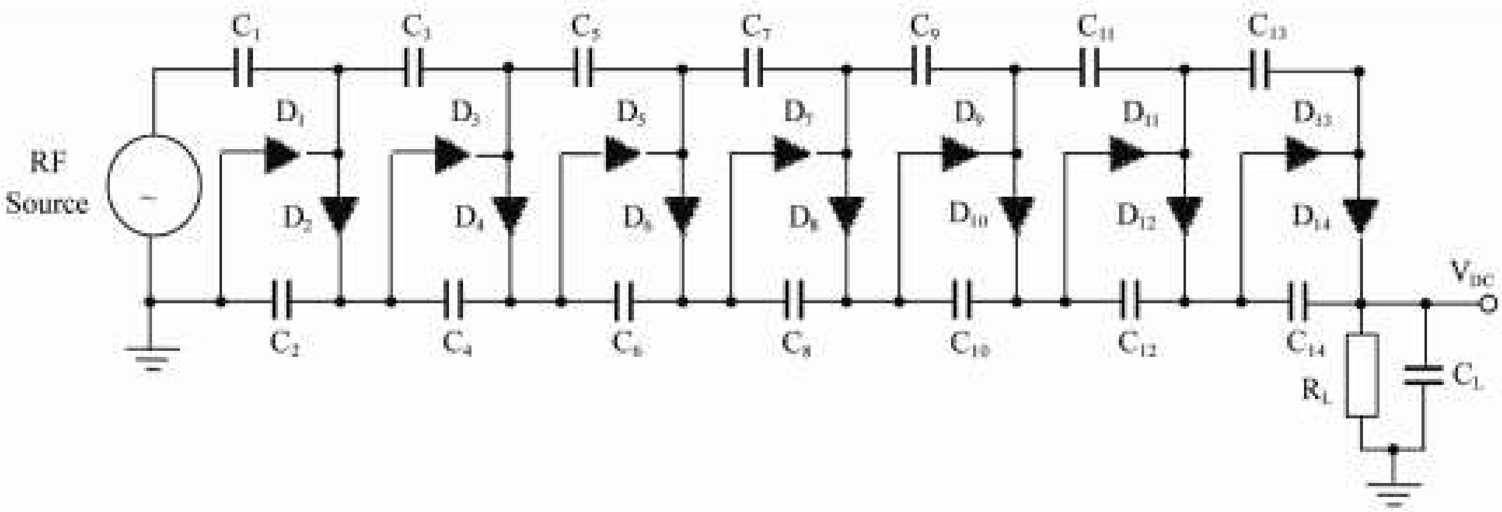

Figure 6. Schematic of 7 stage voltage multiplier.

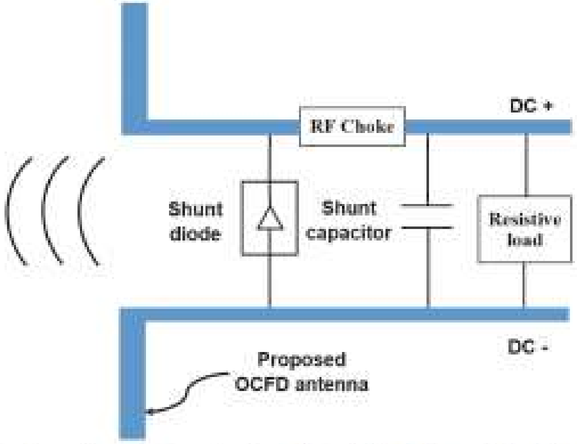

Fig. 9. Configuration of a single shunt diode (Class F) rectifier with a dipole antenna.

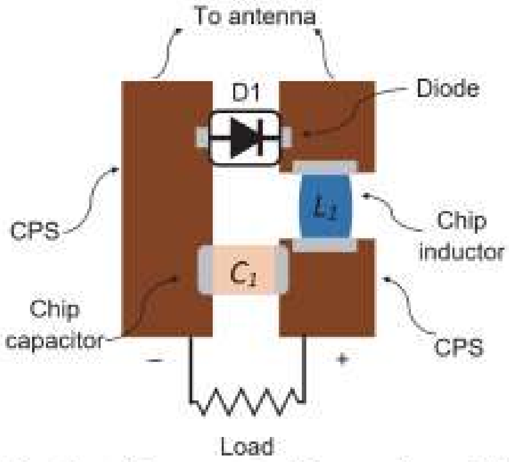

Fig. 10. Configuration of the proposed rectifier on coplanar striplines (CPS).

TABLE III
CIRCUIT COMPONENTS USED IN THE DESIGN

Component name	Nominal Value	Part number and supplier
D1	Schottky diode	SMS7630-079LF, Skyworks
L1	47 nH chip inductor	0603HP47N, Coilcraft
C1	100 nF chip capacitor	GRM188R71H104JA93D, Murata

antenna have a radius of 50 mm and a circumference angle of

respectively. While the imaginary part of the proposed OCFD is around 0 Ω at resonant frequencies 0.6 GHz, 1.2 GHz and 2.4 GHz, which are f_0, $2f_0$, and $4f_0$ respectively. These results have demonstrated that the simulated results agree with the OCFD theory as discussed in Section III–A. Furthermore, the imaginary part of the impedance of the antenna over the resonant frequency band from 1.4 to 2 GHz turns from negative values (for the reference antenna) to positive values (for the proposed antenna). As shown in Fig. 7(b), the value of the imaginary part of the proposed antenna impedance varies between 0 and 300 Ω over the desired frequency band. This feature could help the proposed antenna to produce a better conjugate matching with the rectifier, since the imaginary part of the impedance of the rectifier normally varies between -700 and 0 Ω as we discussed earlier. The simulated 3D radiation patterns of the proposed antenna at the frequencies of interest are depicted in Fig. 8. The 2D polar plots of antenna patterns in *E-plane* and *H-plane* are shown as well. Here we have only showed the directivity (maximum gain) of the antenna (without taking the mismatch loss into account). From Fig. 8, it can be seen that the antenna has symmetrical patterns about YOZ plane with a maximum directivity of 1.8 dBi at 0.9 GHz, 3.5 dBi at 1.8 GHz and 3.3 dBi at 2.4 GHz. The antenna is more directive towards the long arm direction at 1.8 GHz and 2.4 GHz with the half-power beam-widths (HPBW) of around 174° and 185° respectively. The HPBW is about 96° at 0.9 GHz.

Therefore, the proposed broadband OCFD antenna has obtained high impedance over a wide frequency range. The proposed design is just an example to illustrate the proposed new method. The details of the dipole could be modified according to the frequency of interest.

IV. RECTENNA INTEGRATION

A. *Rectifier Configuration*

The proposed high impedance OCFD antenna may directly conjugate match with the input impedance of a rectifier over a wide frequency band. The rectifier should only consist of few circuit components for rectification, DC storage and output. A single shunt diode rectifier is selected due to its very simple structure and high conversion efficiency [33]. The configuration of the single shunt diode rectifier with a dipole

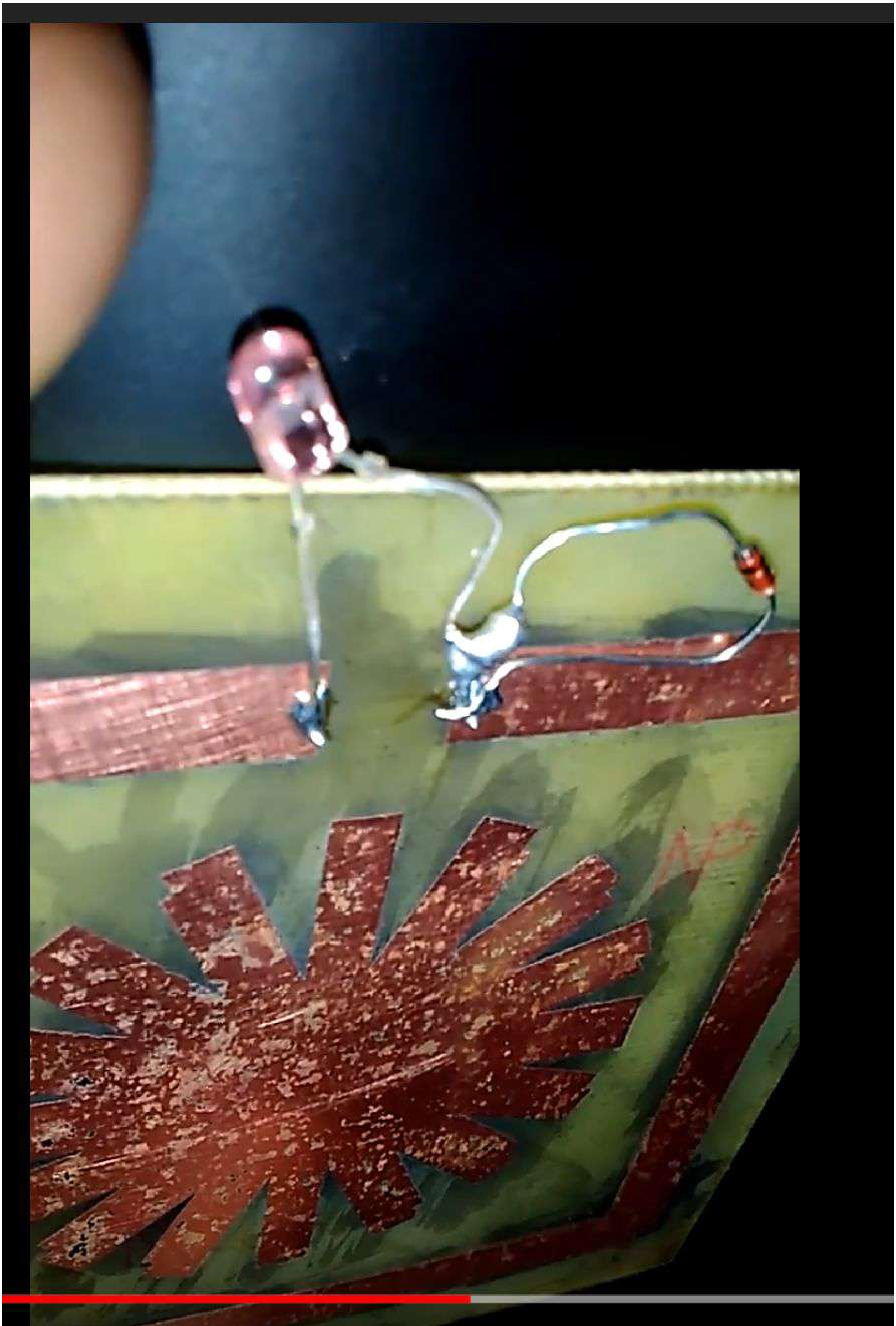

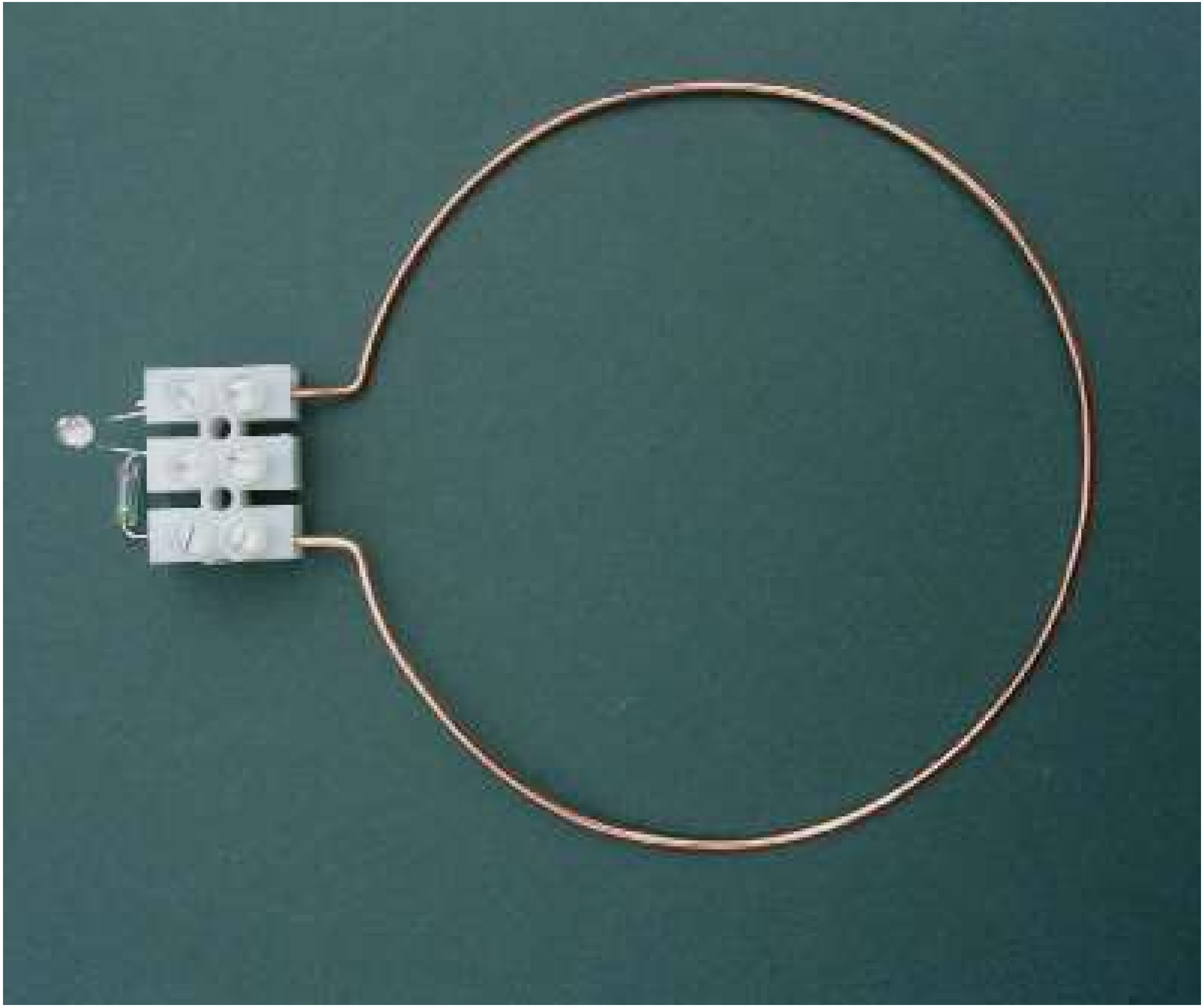

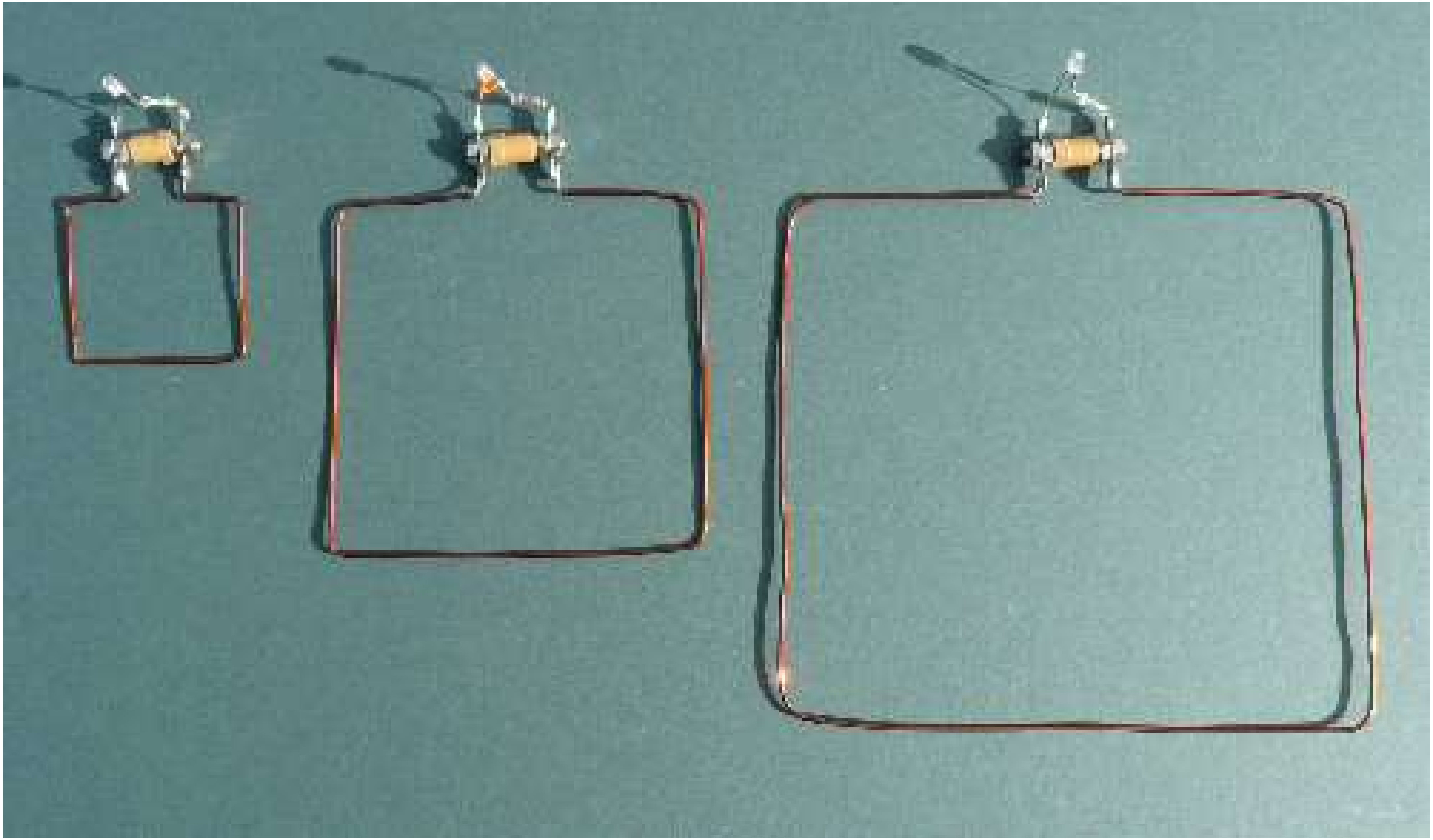

ake and it can be sent in an envelope. I made tens of them and sent them to politicians, newspapers, universities... I gave some to local people, together with a user
later version, that can be rolled up in an envelope that fits the conditions to be sent with only one stamp.

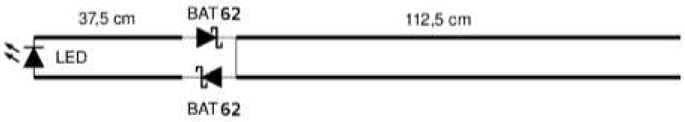

BAT62 detection diodes are no more produced. **BAT15** diodes work fine but they wear out; after a few months they stop functioning. **SMS7630** diodes are great but very little and mechanically fragile. **MMSD701T1G** diodes are sturdy and powerful; an excellent choice for a beginner. Such SMD diodes do also work for cell phone frequencies, which allows to test out a snake with a calling cell phone pushed against it. But any detection diodes that can manage 100 MHz will do.

The LED I'm currently using is the **L-7113SEC-H** . It lights up with a low tension and a very low current (the bluer a LED, the more tension it needs). Its color is red yet close to orange hence it is easily seen by the human eye (the eye is most sensitive to green, yellow and orange). The beam is quite narrow so when the LED is directed towards somebody's eyes it will appear quite bright.

For the lengths of 37.5 and 112.5 centimeters, any electric wire with two copper conductors will do. Audio signal wire is a practical solution. Use the shielding as one of the two conductors. The lengths of the two segments must not be precise. What matters is that the total length of the snake be 1.5 meters. Do not hesitate to try out if a little longer or shorter snake gives better results.

A schematic of my current probes, that I connect to a standard multimeter, measuring Volts DC. The measure displayed by the multimeter must be multiplied by 10. When using a 200.0 mV scale, just read while forgetting the dot:

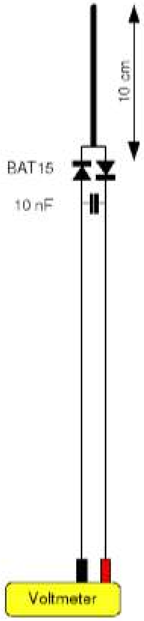

Lead Free Status	Lead Free
RoHS Status	RoHS Compliant

Features, Applications

The MMSD301T1, and MMSD701T1 devices are spin-offs of our popular MMBD301LT1, and MMBD701LT1 SOT-23 devices. They are designed for high-efficiency UHF and VHF detector applications. Readily available to many other fast switching RF and digital applications.

Extremely Low Minority Carrier Lifetime Very Low Capacitance Low Reverse Leakage AEC Qualified and PPAP Capable S Prefix for Automotive and Other Applications Requiring Unique Site and Control Change Requirements These Devices are Pb-Free, Halogen Free/BFR Free and are RoHS Compliant*

XXX G = Specific Device Code SMMSD701T1G = Date Code = Pb-Free Package
Rating Reverse Voltage MMSD701T1G, SMMSD701T1G Forward Current (DC) Continous Forward Power Dissipation = 25C Junction Temperature Storage Temperature Range Symbol VR Value to +150 Unit Vdc M G

152

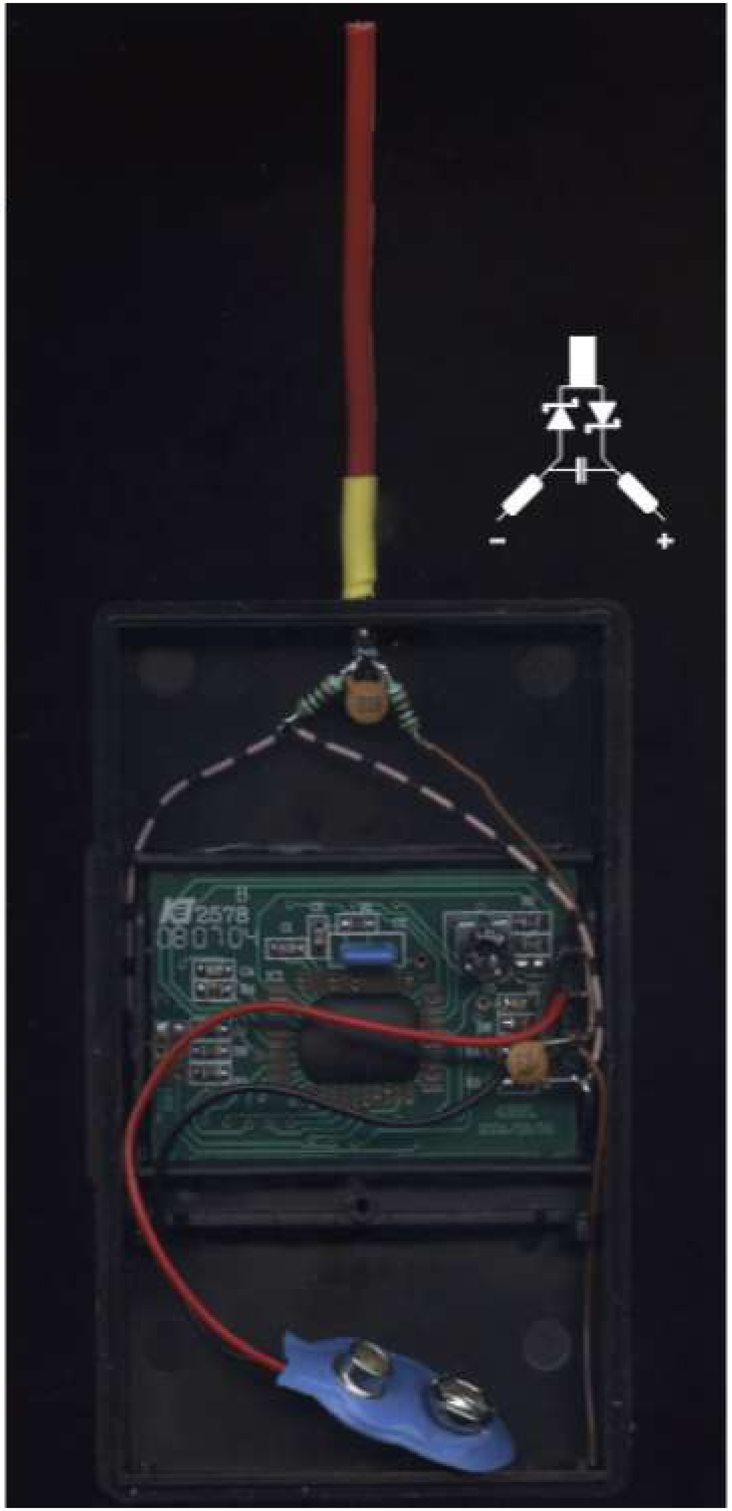
2578
080704
-
+

Parts Required:

1. 100nF & 100pF
2. n4148 Diode x(2)
3. A bright good quality LED

Circuit Diagram:

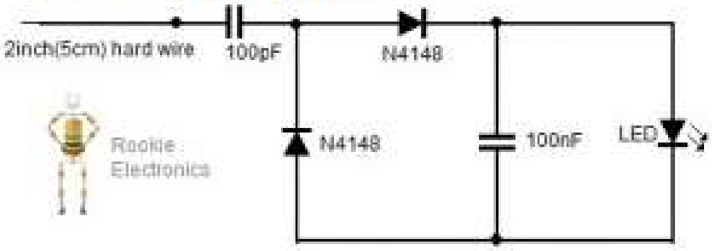
2inch(5cm) hard wire
100pF
N4148
N4148
100nF
LED
Rookie
Electronics

Bread board Arrangement:

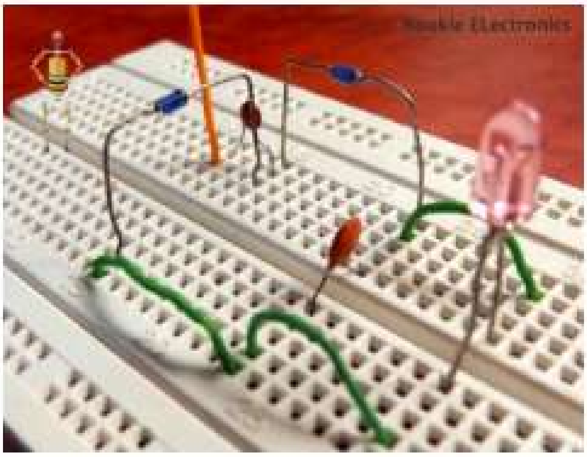
Rookie ELectronics

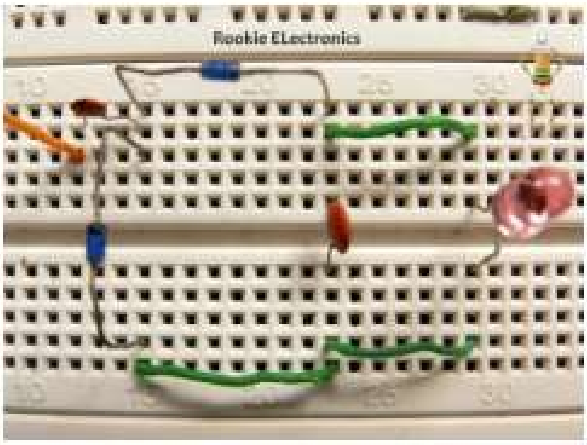
Rookie ELectronics

Strip Board:

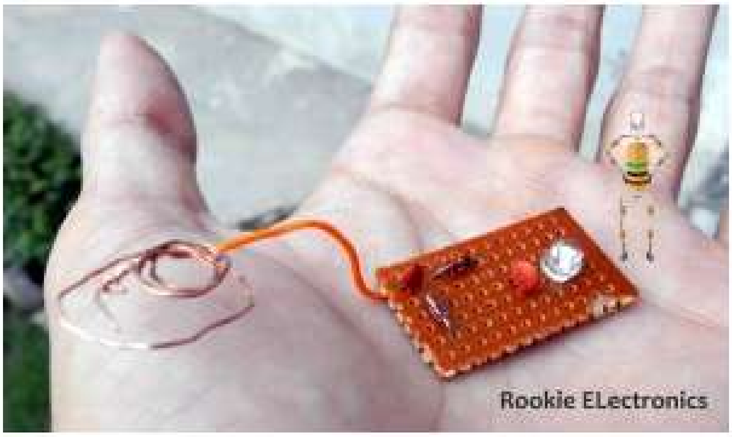
Rookie ELectronics

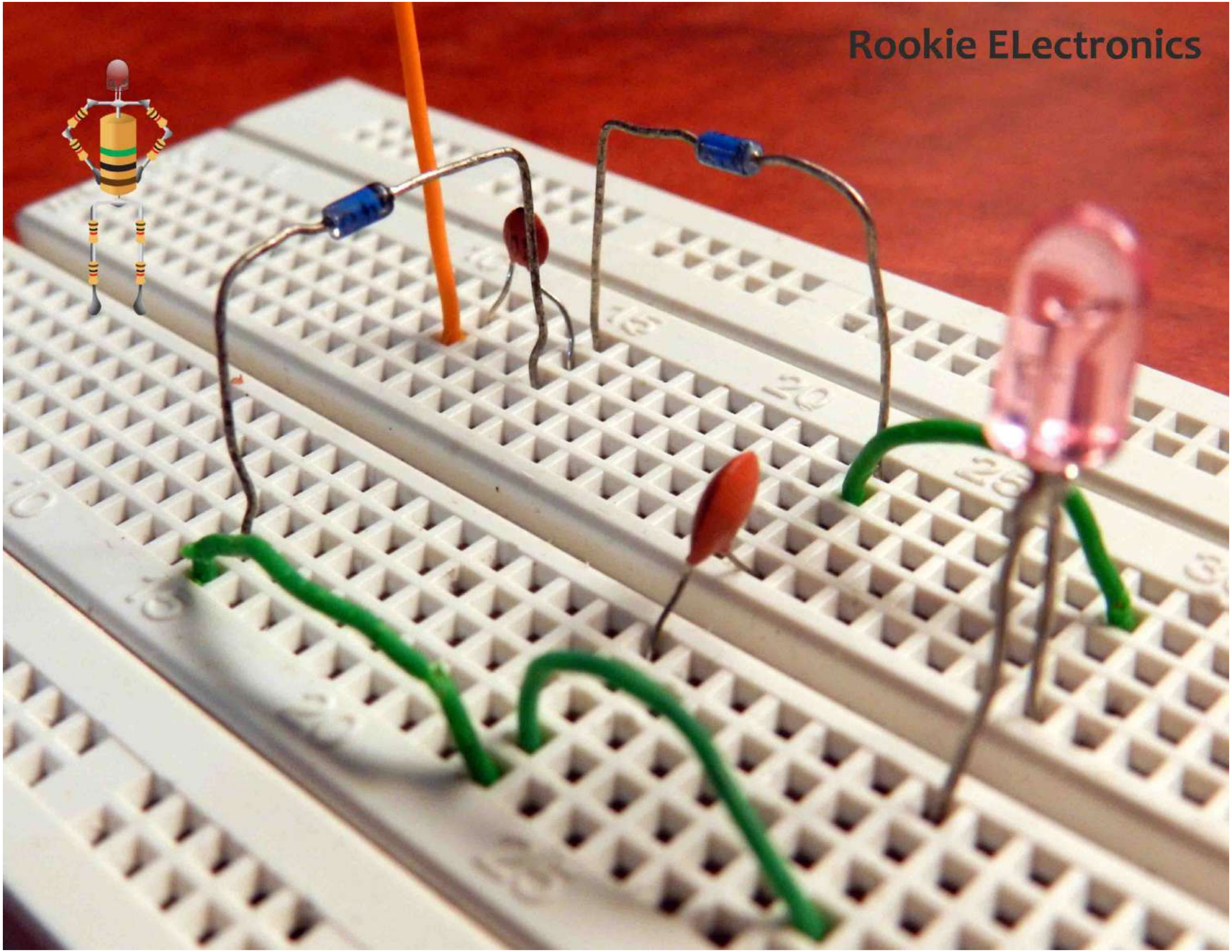
Rookie ELectronics

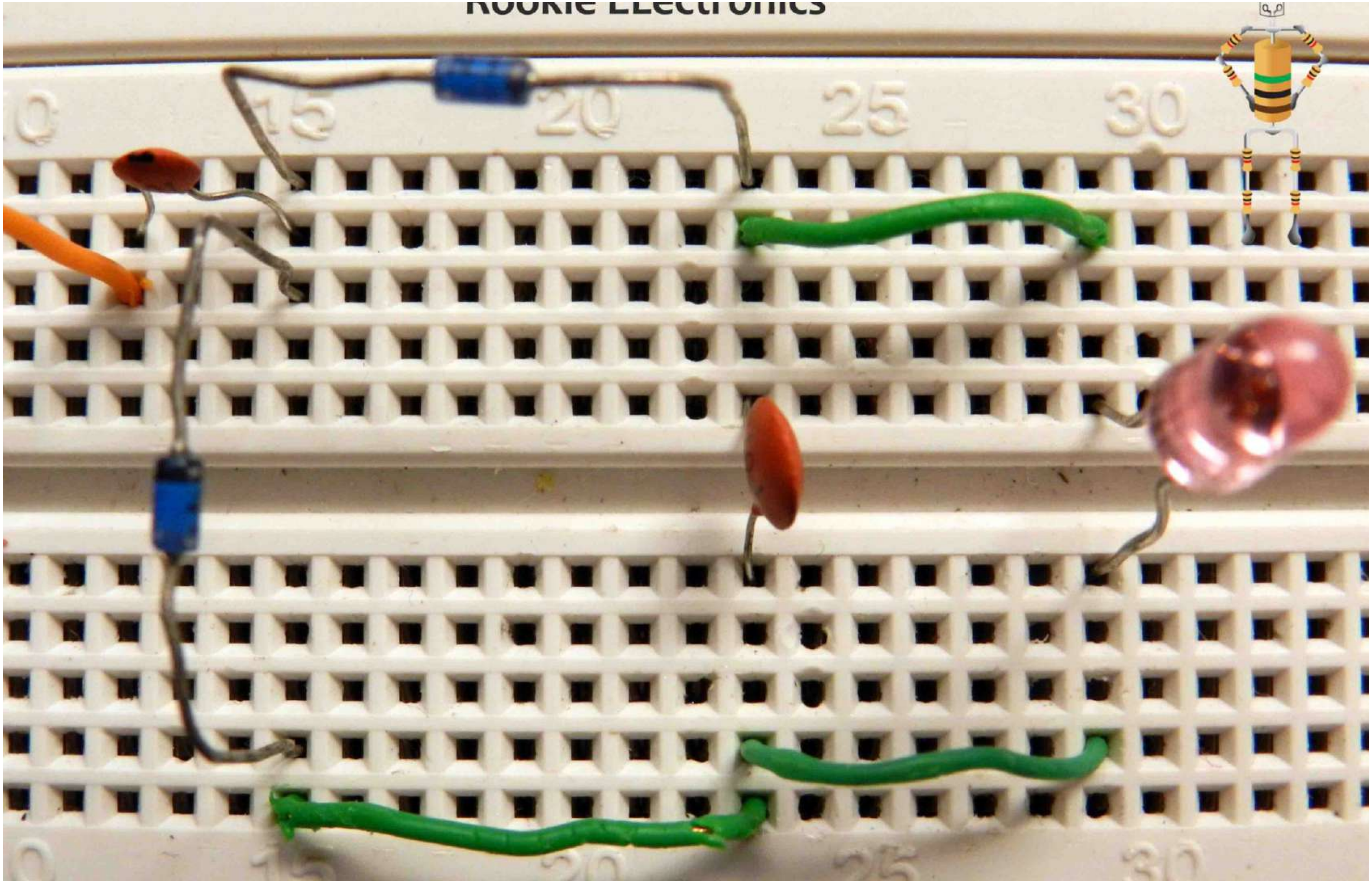
10
15
20
25
30
10
15
20
25
30

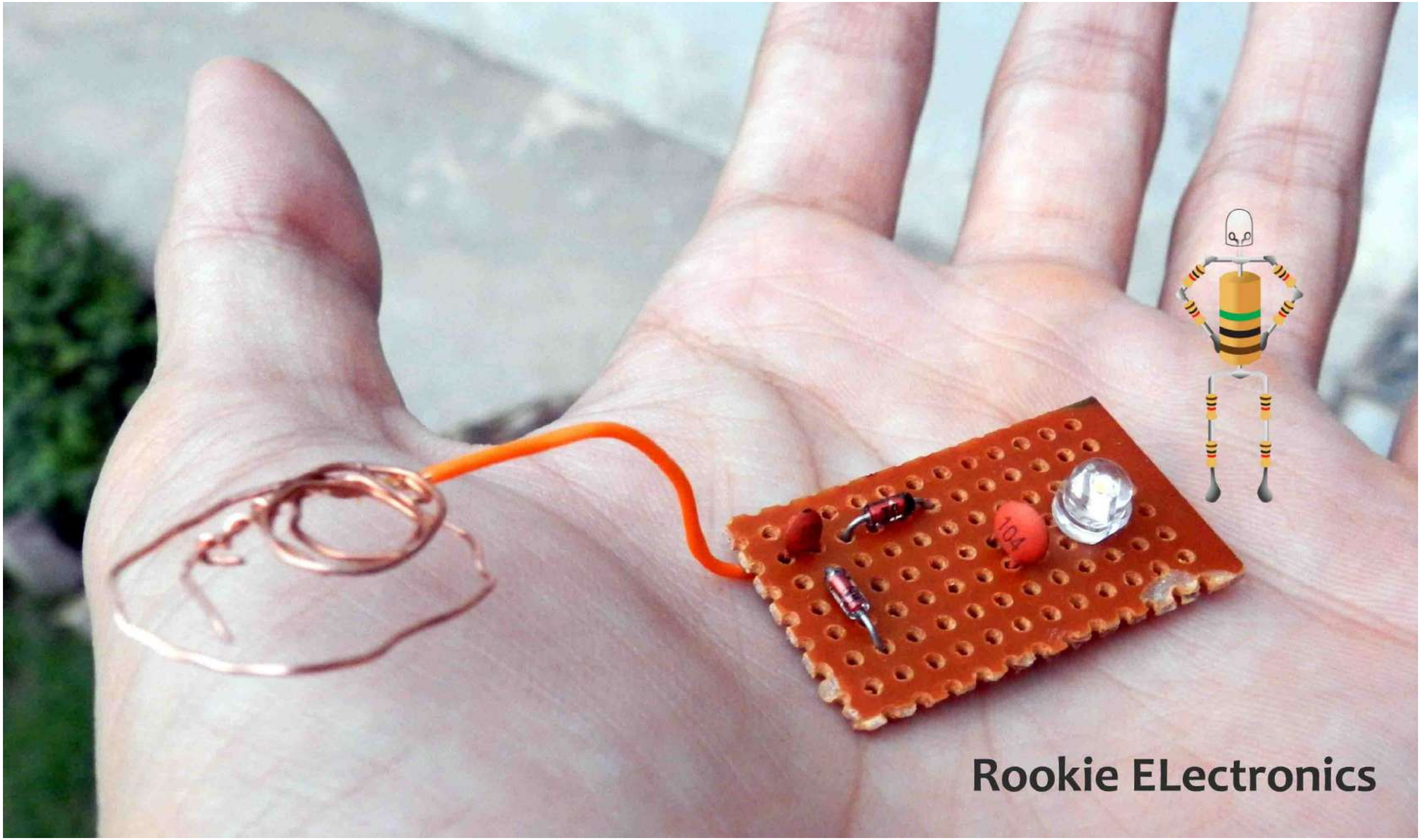

Rookie ELectronics

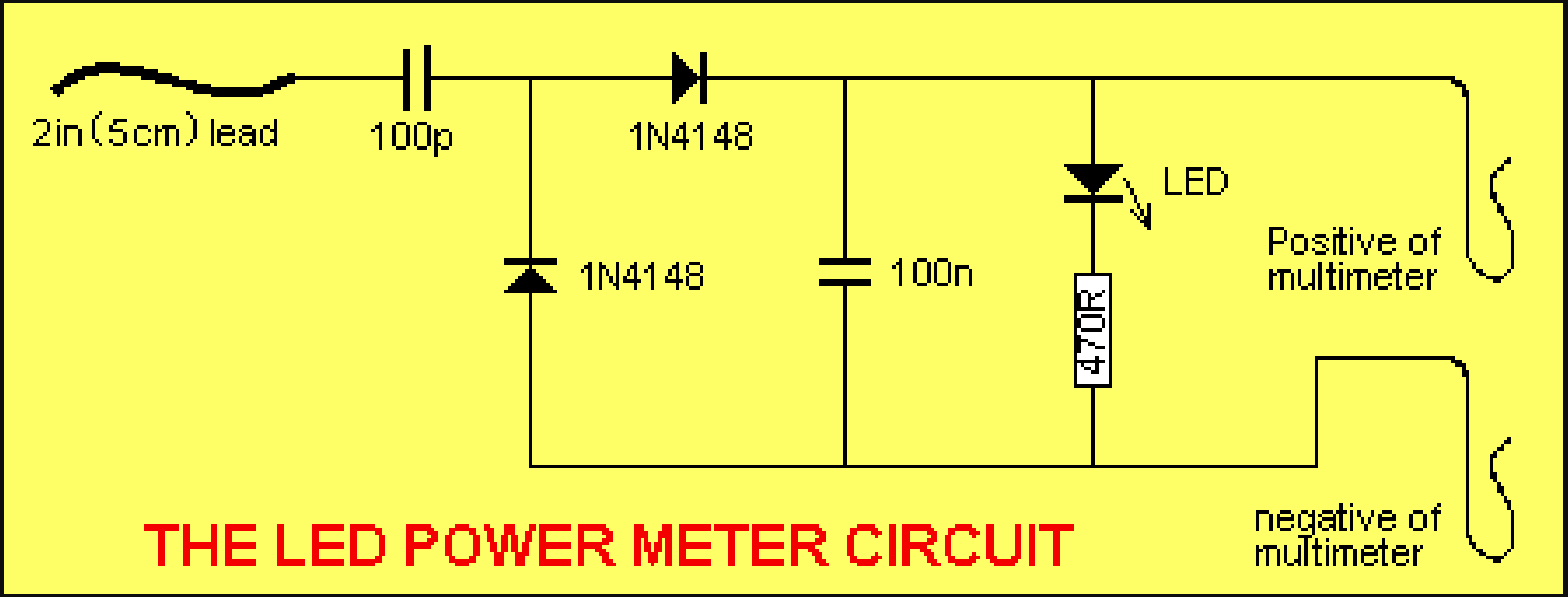
2in(5cm) lead
100p
1N4148
1N4148
100n
LED
470R
Positive of multimeter
negative of multimeter

THE LED POWER METER CIRCUIT

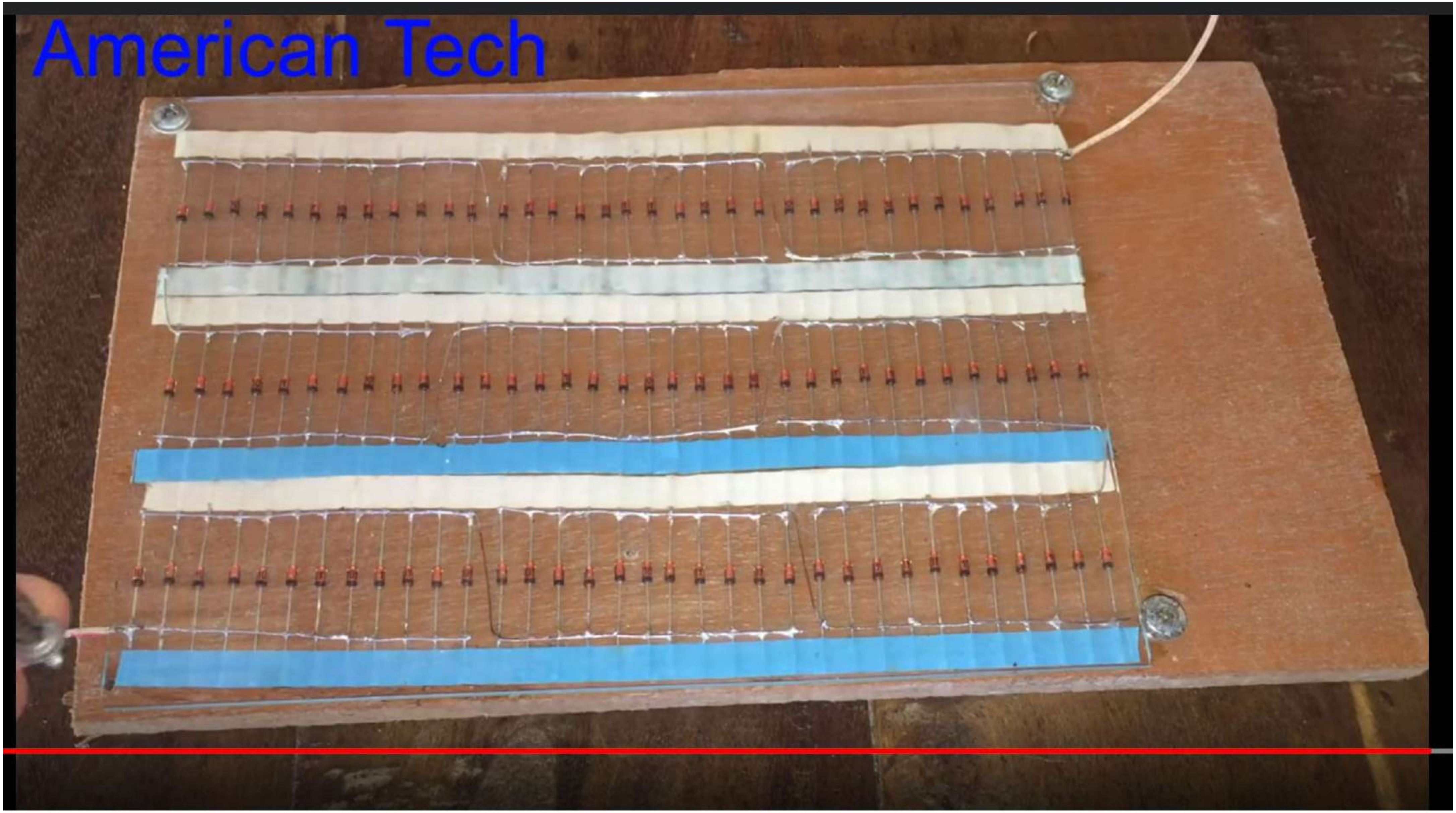
American Tech

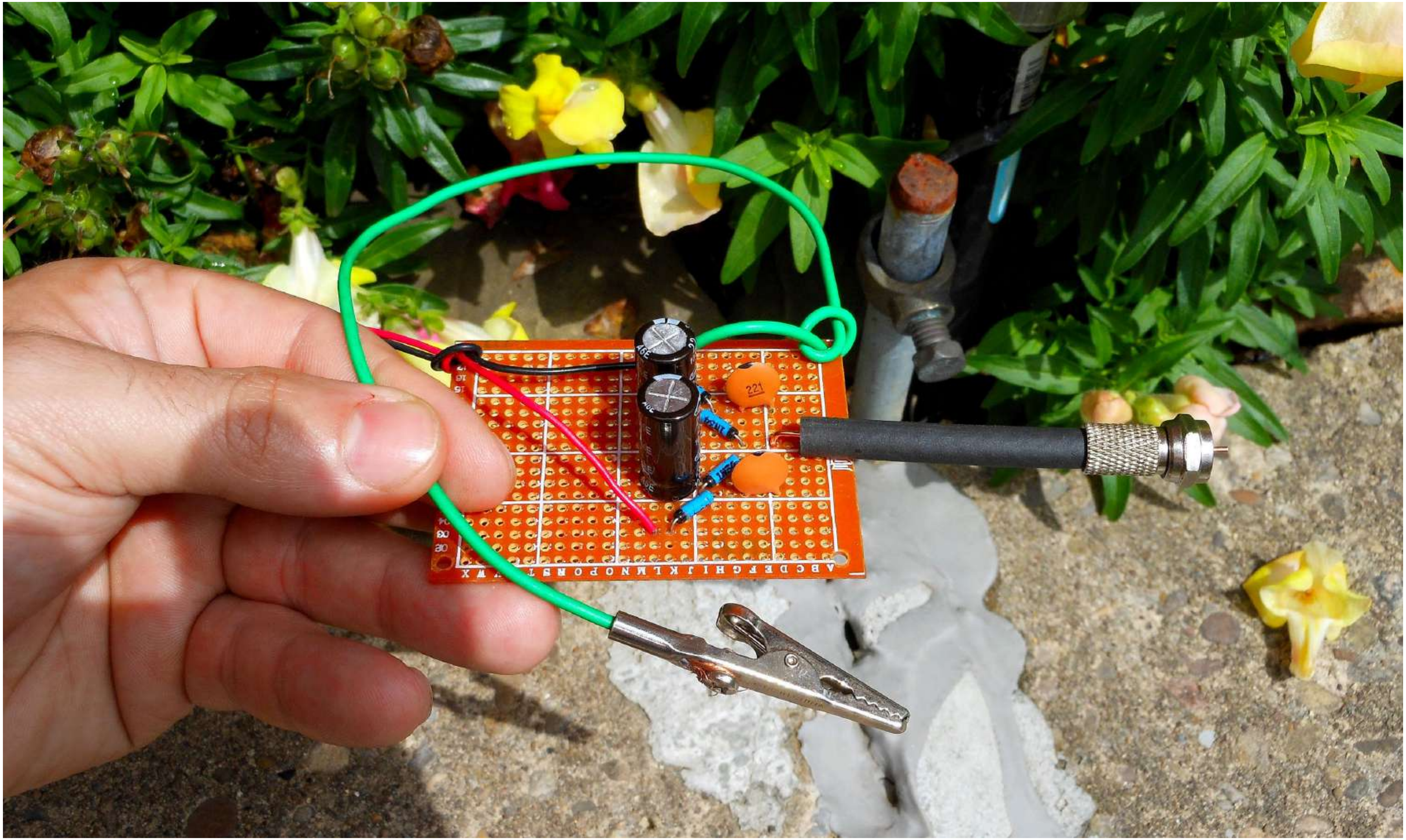
221

24 - 1N34A'S
12 - DISC CAP'S .001µF
12 - ELEC CAP'S 200V 22µF

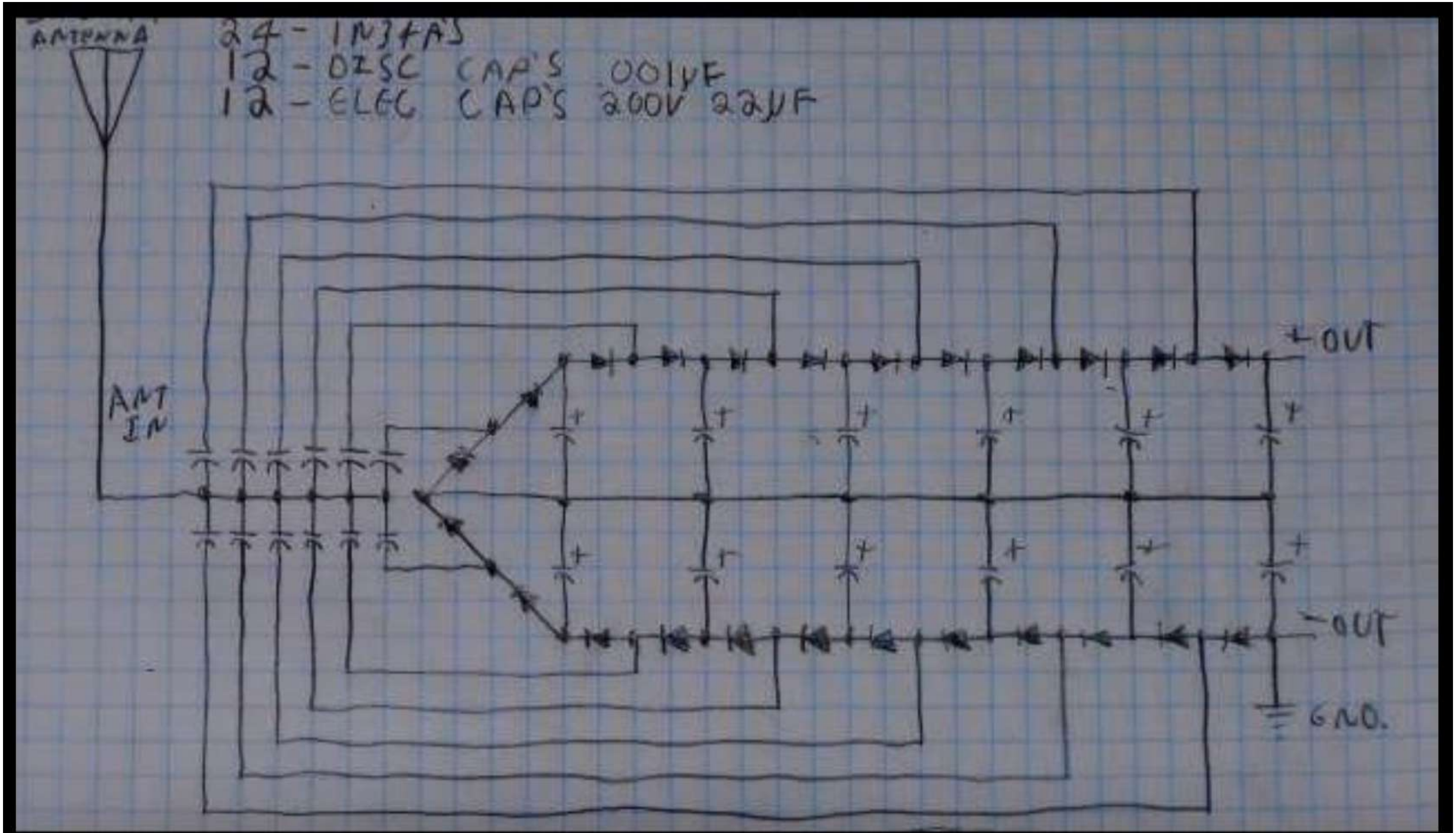
ANTENNA
ANT IN
+OUT
-OUT
GND.

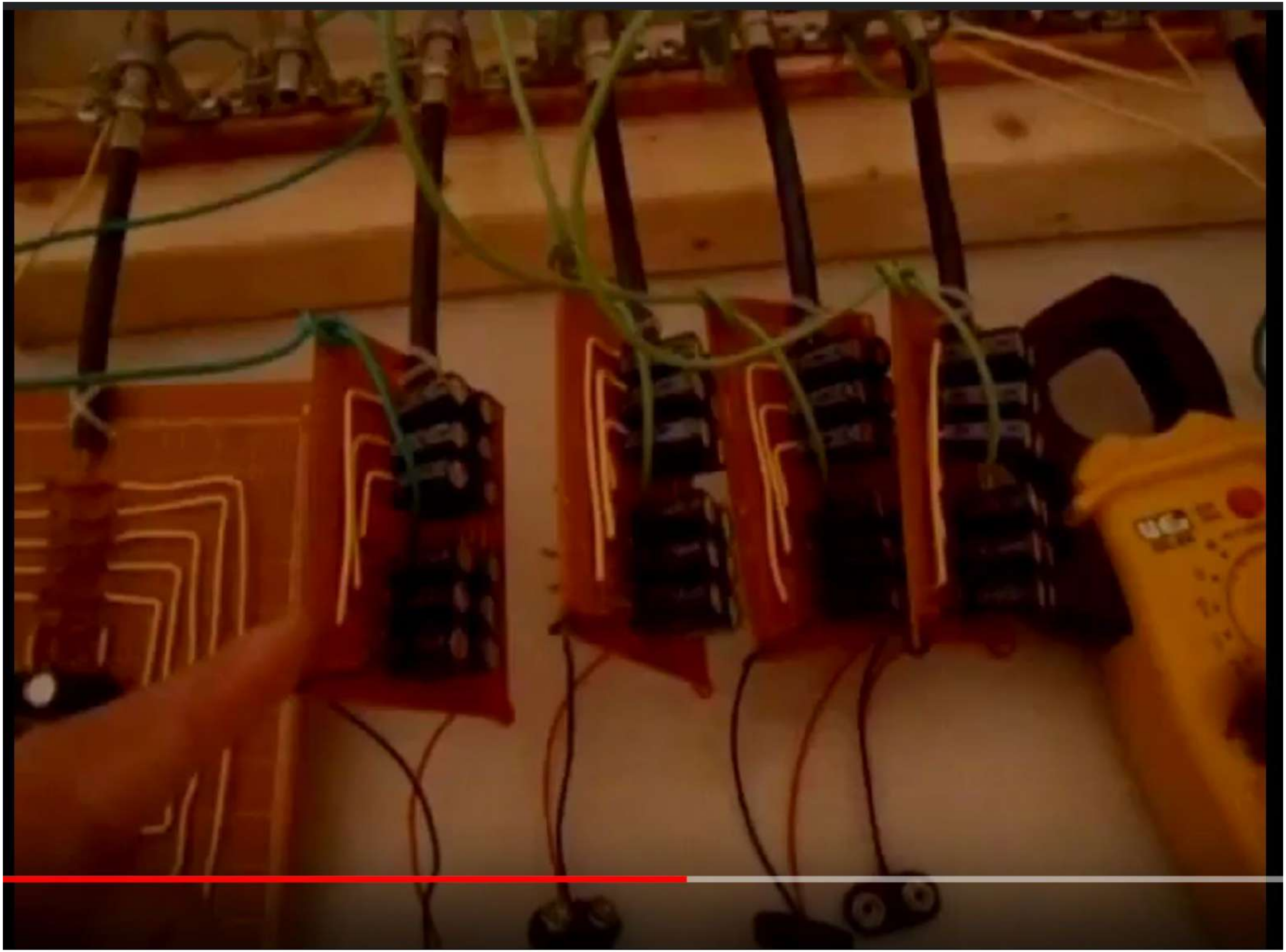

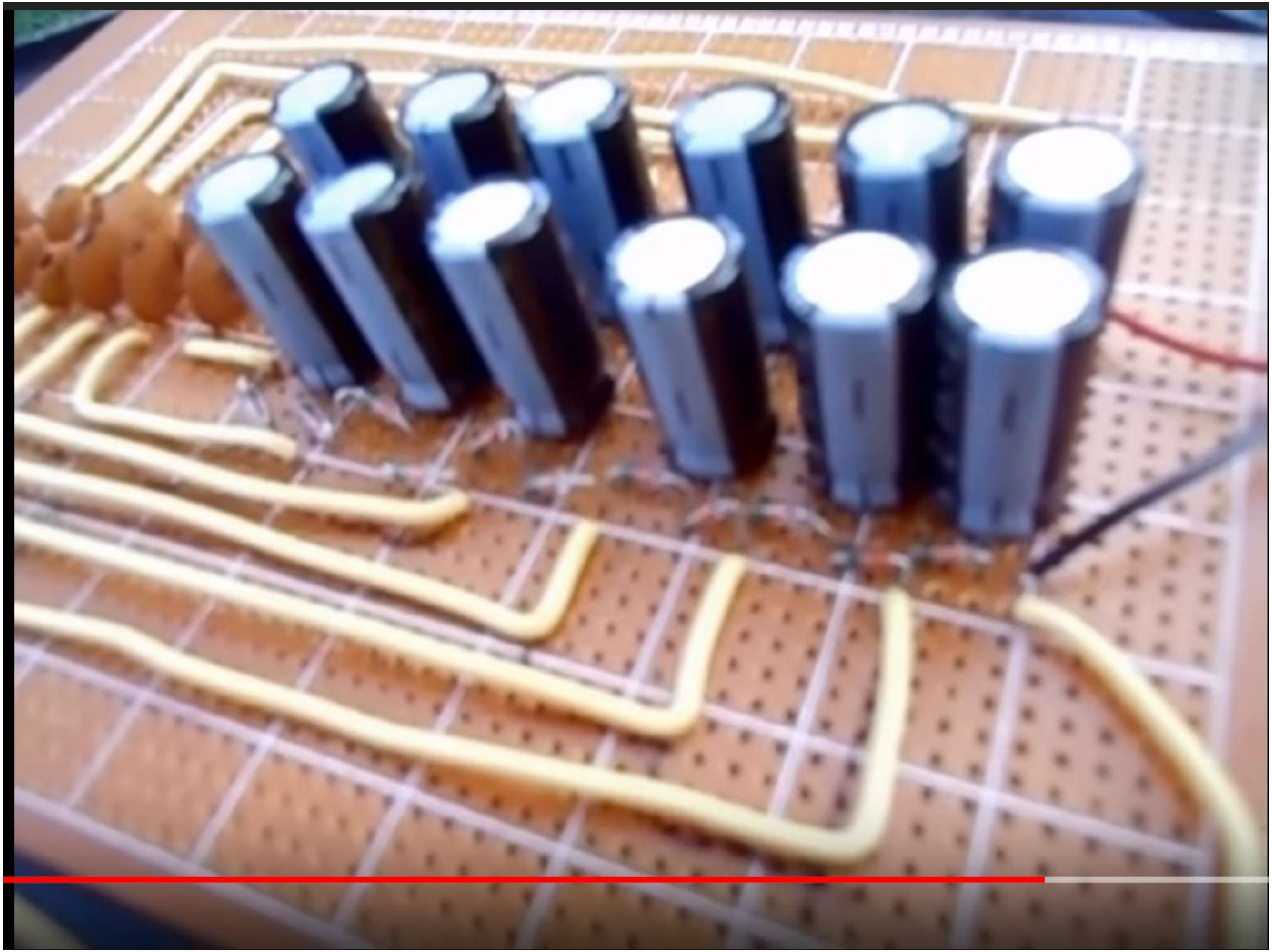

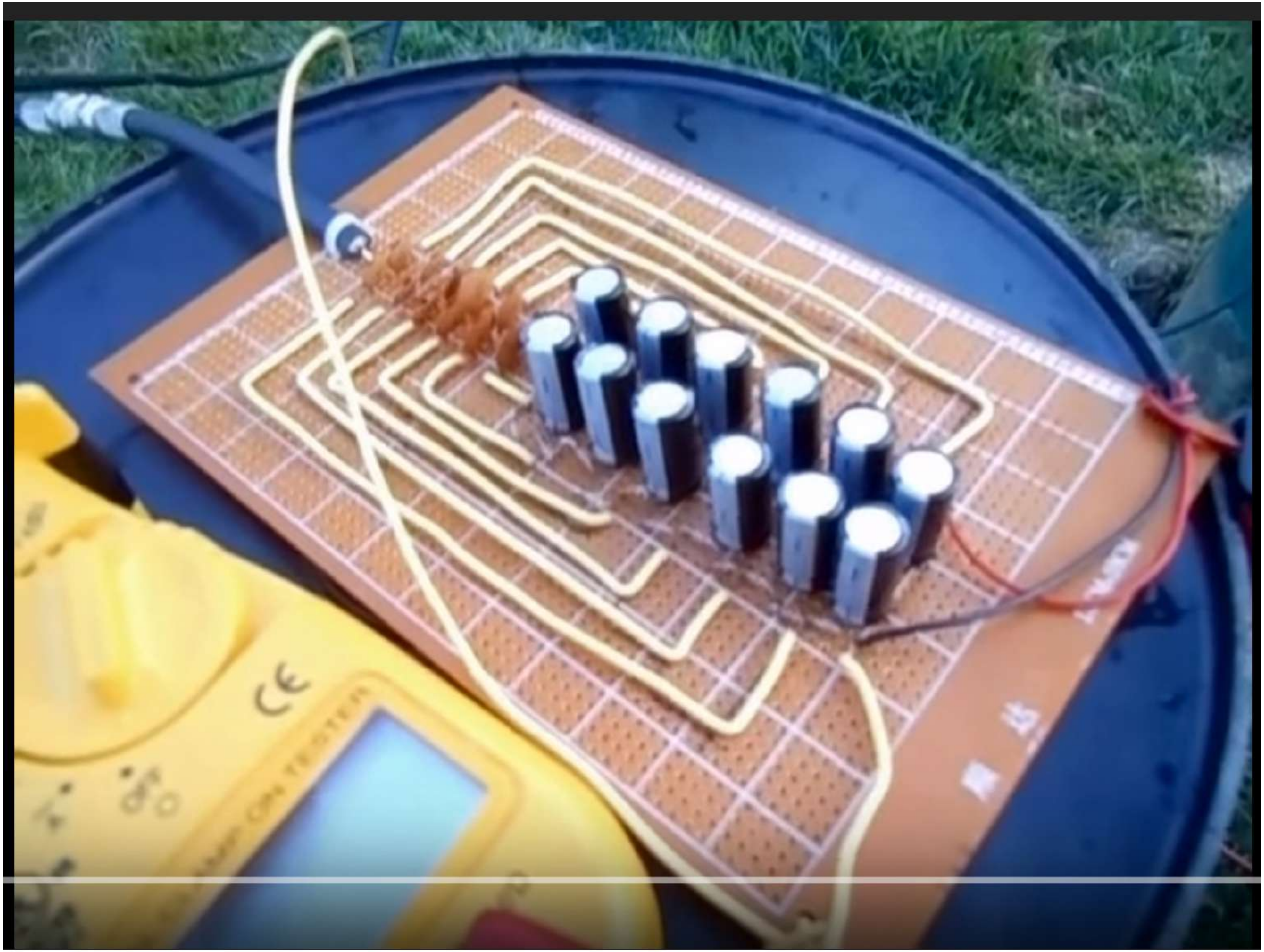
CE

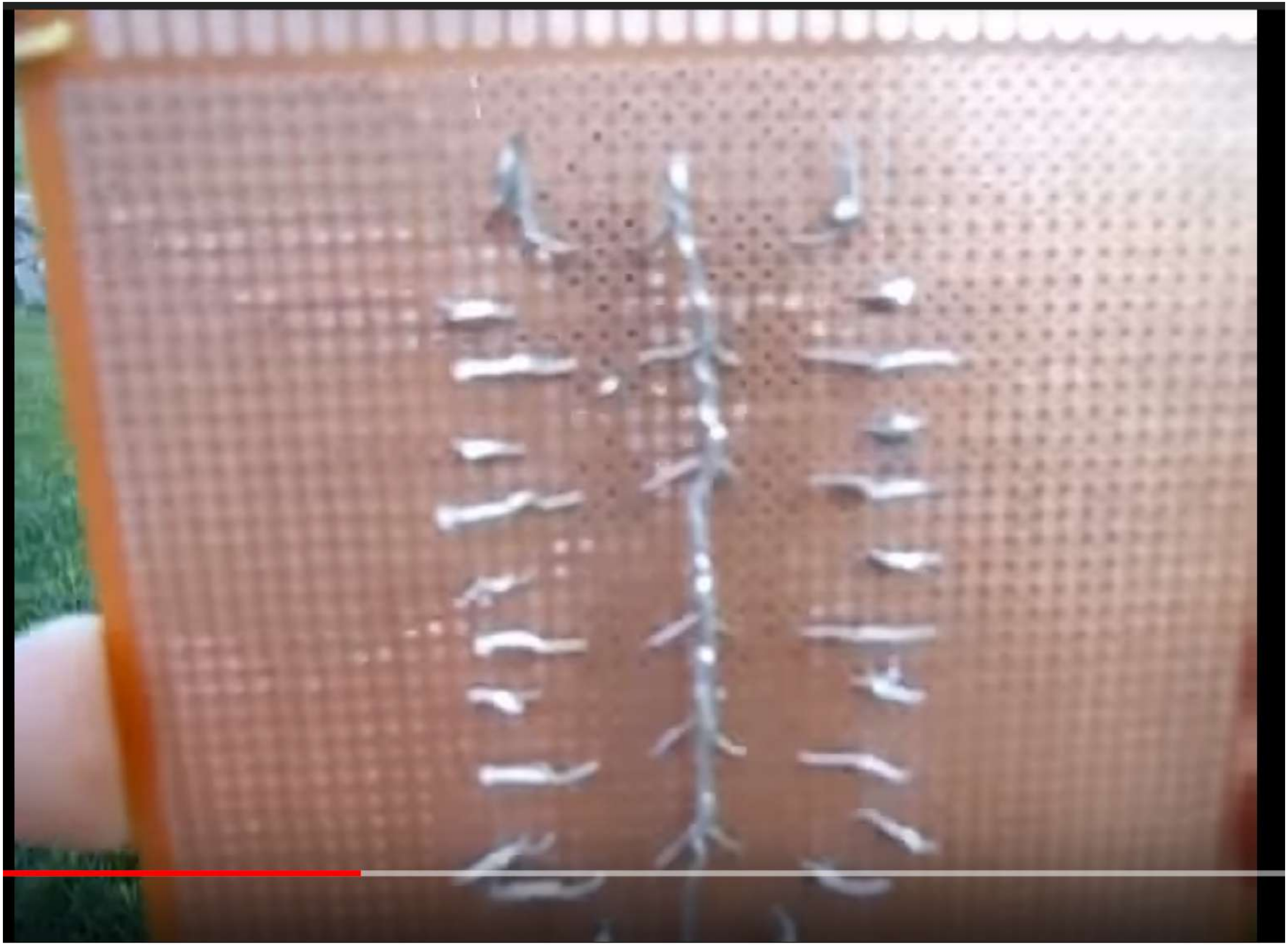

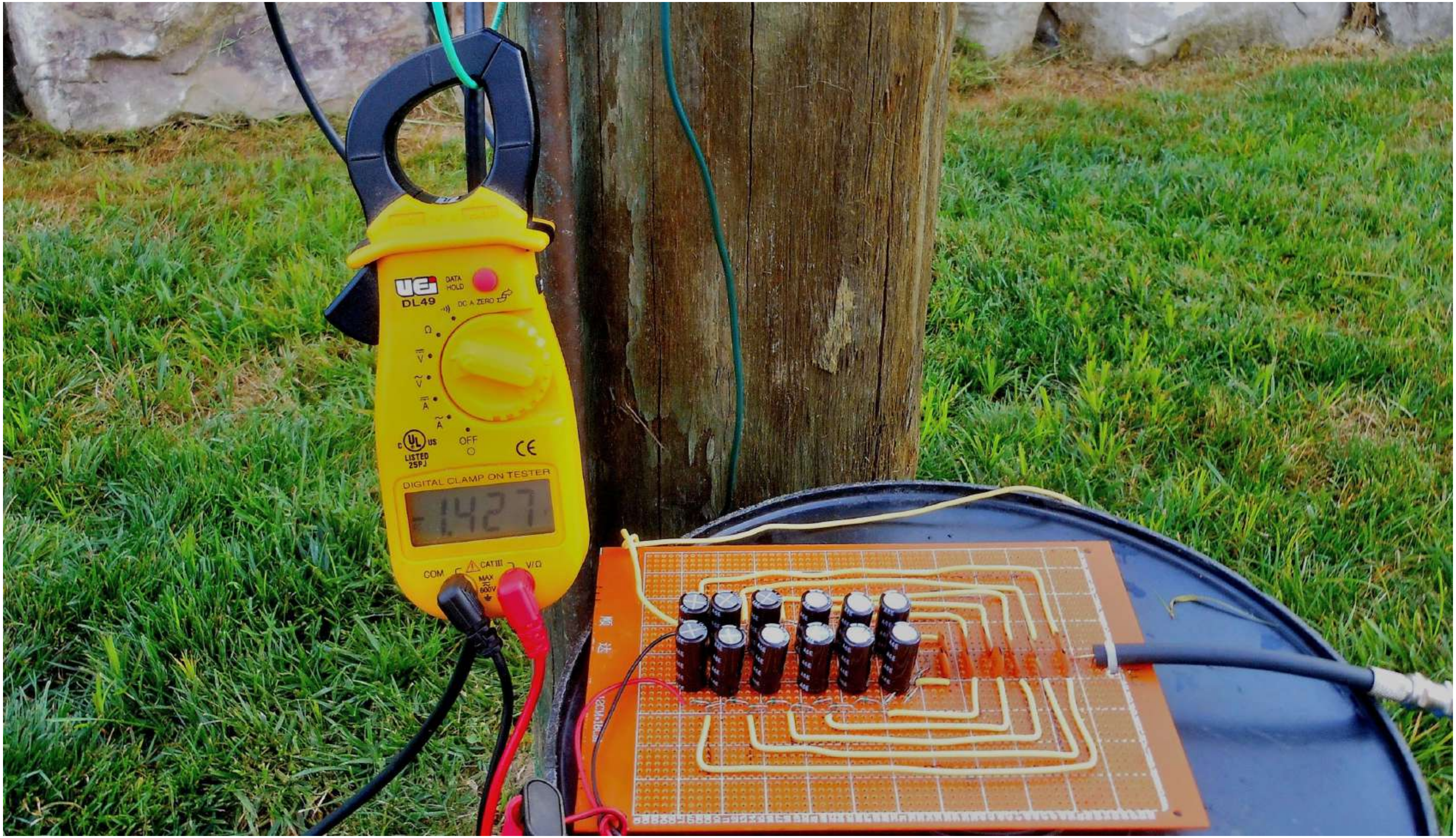
UEi
DL49
DATA HOLD
DC A ZERO
OFF
LISTED 25PJ
CE
DIGITAL CLAMP ON TESTER
-1.427
COM
CAT III
MAX 600V
V/Ω

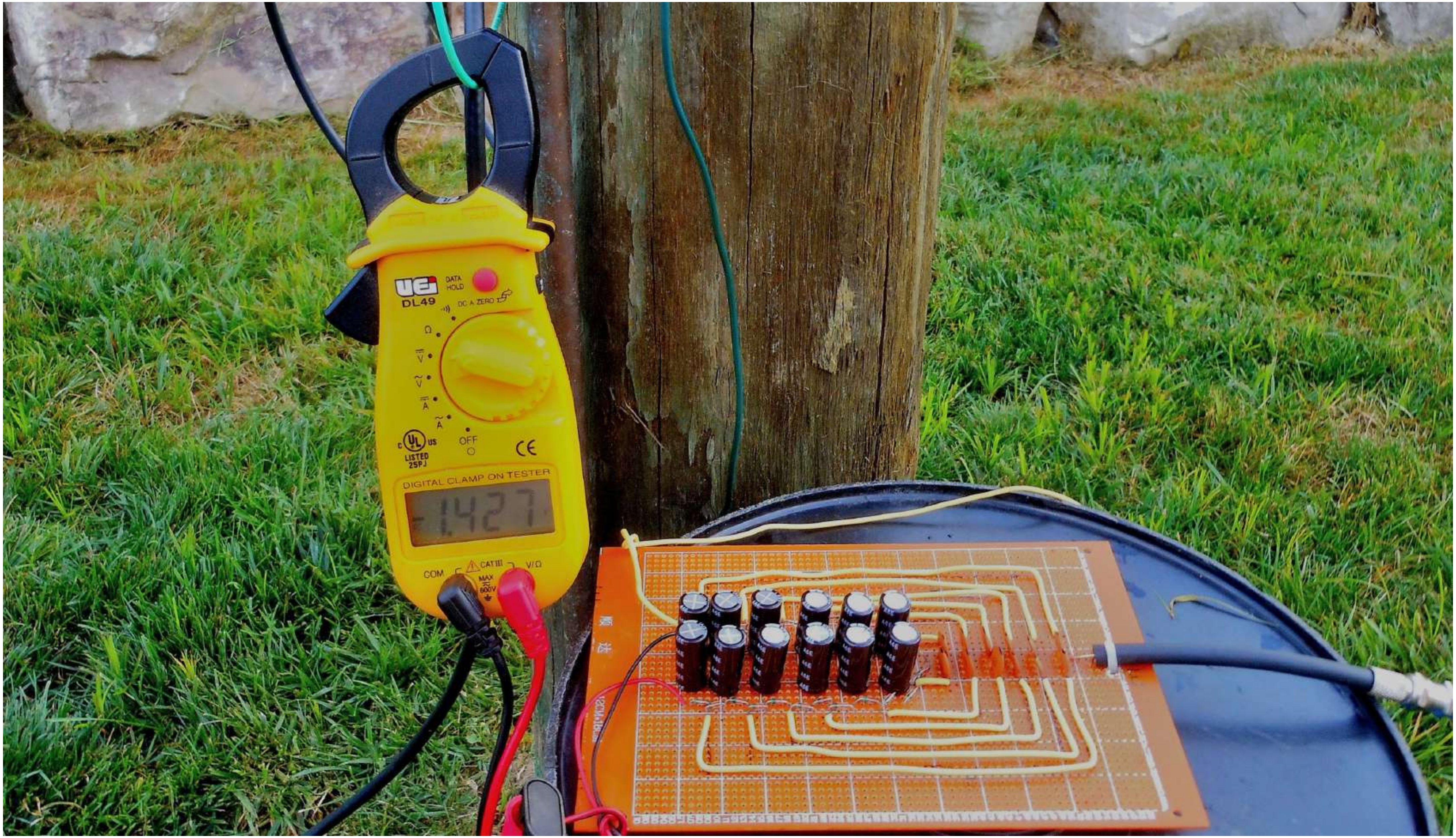
UEi
DL49
DATA HOLD
DC A ZERO
OFF
LISTED 25PJ
CE
DIGITAL CLAMP ON TESTER
-1.427
COM
CAT III
MAX 600V
V/Ω

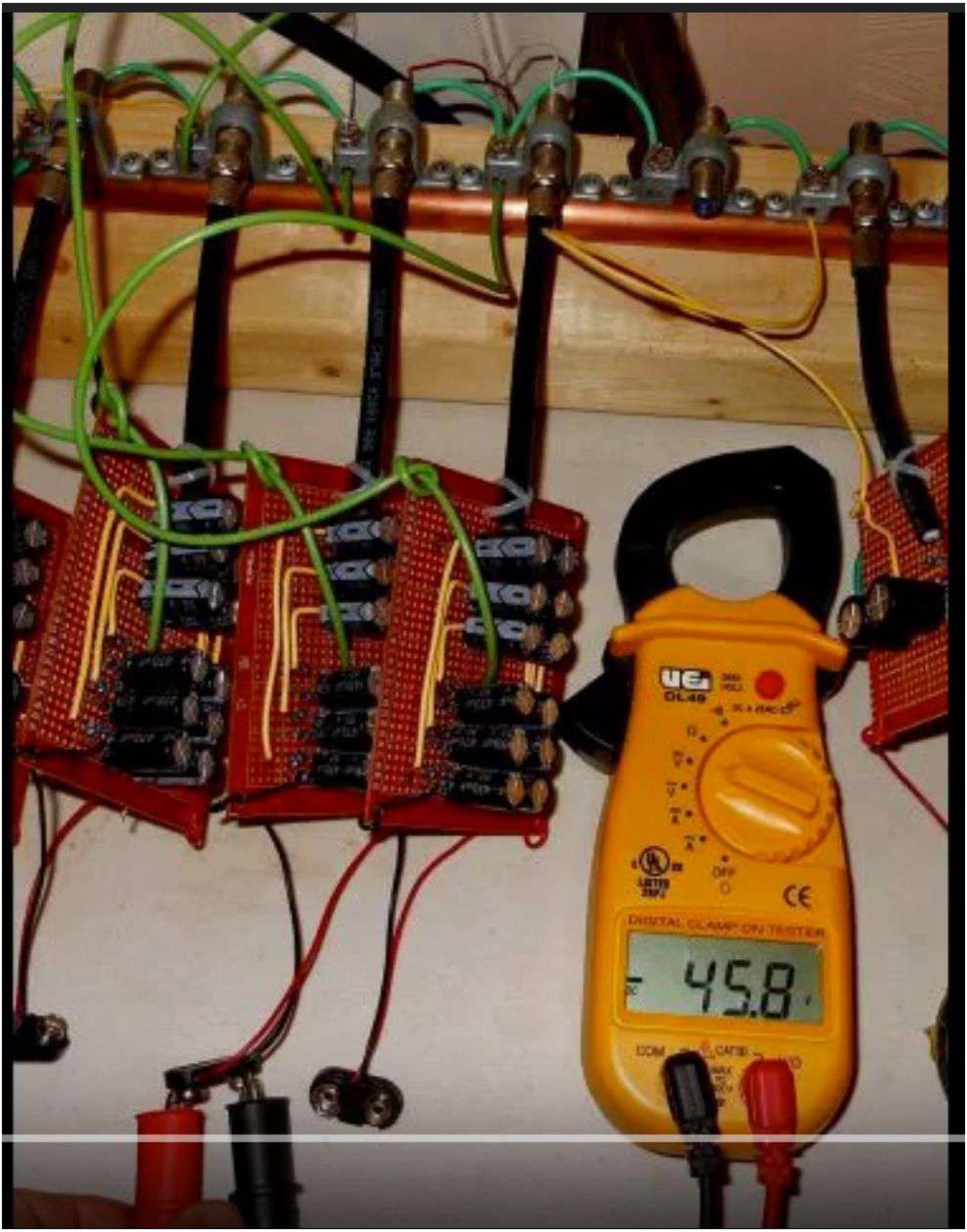
UEi
DL49
OFF
CE
DIGITAL CLAMP ON TESTER
45.8
COM

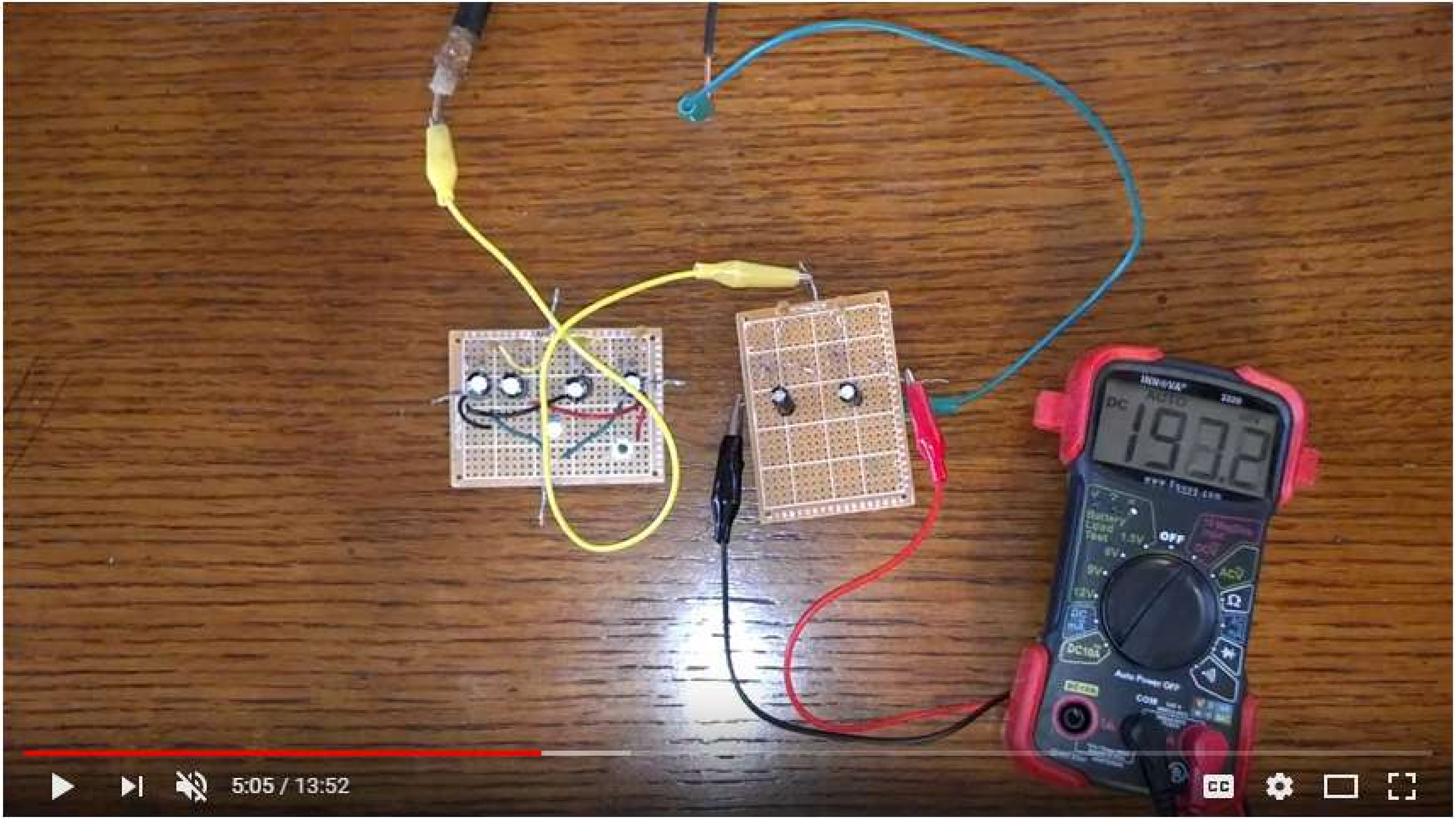
DC
193.2
OFF
9V
12V
DC10A
Auto Power OFF
COM
5:05 / 13:52
CC

Tesla Free Energy Air Circuit

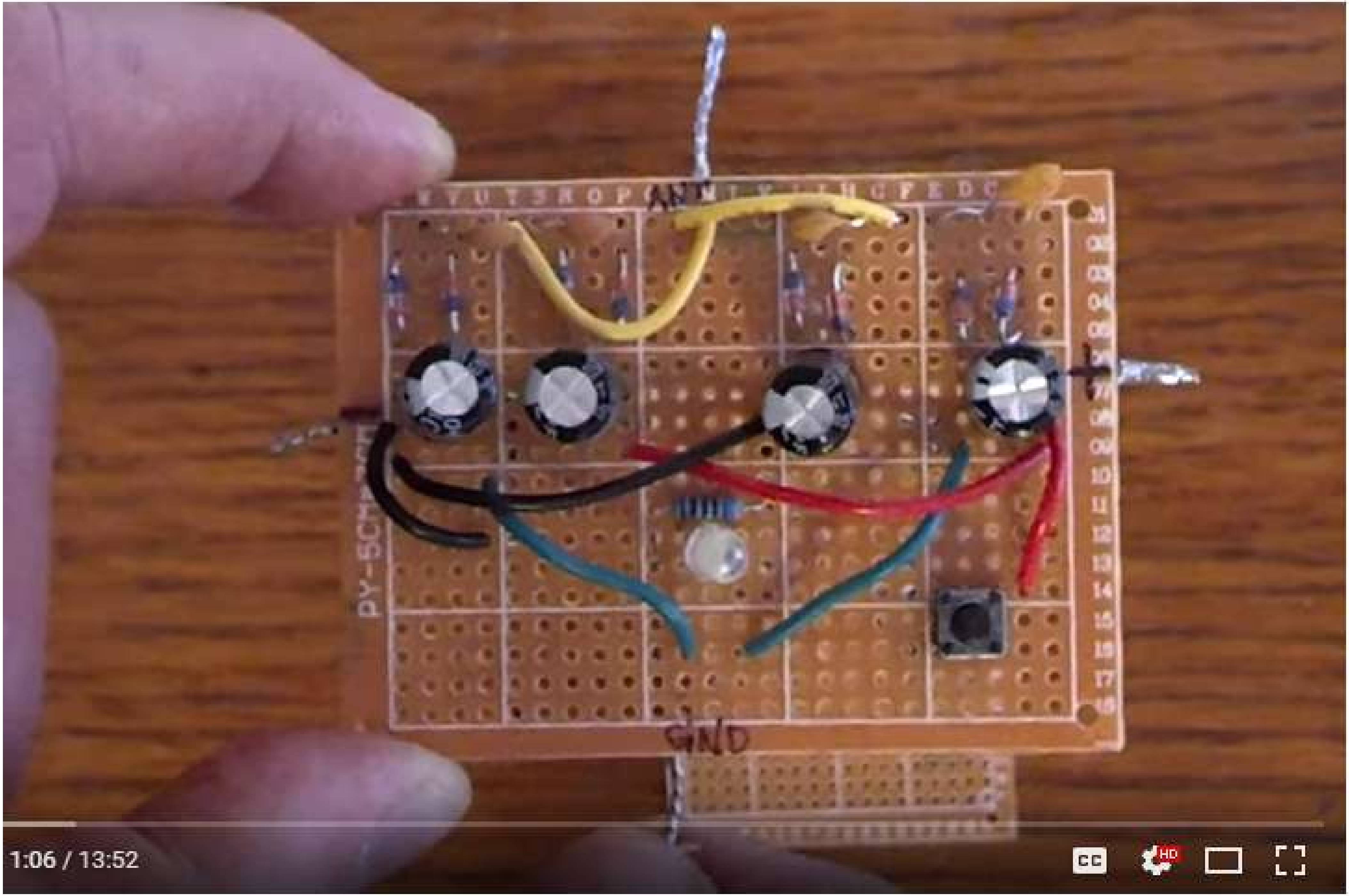
ANT
+
GND
1:06 / 13:52
HD

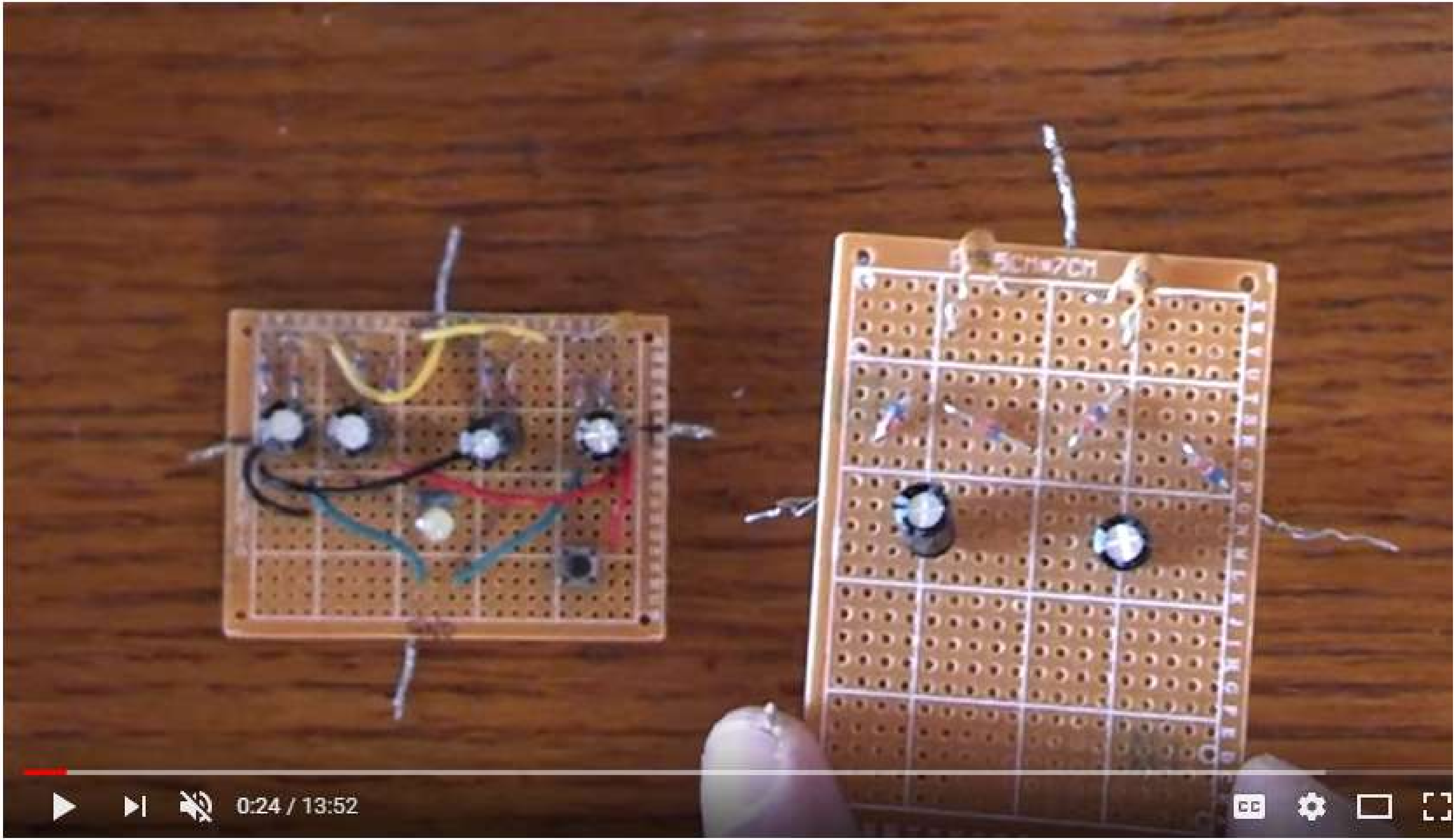
5CM*7CM
0:24 / 13:52

Tesla Free Energy Air Circuit

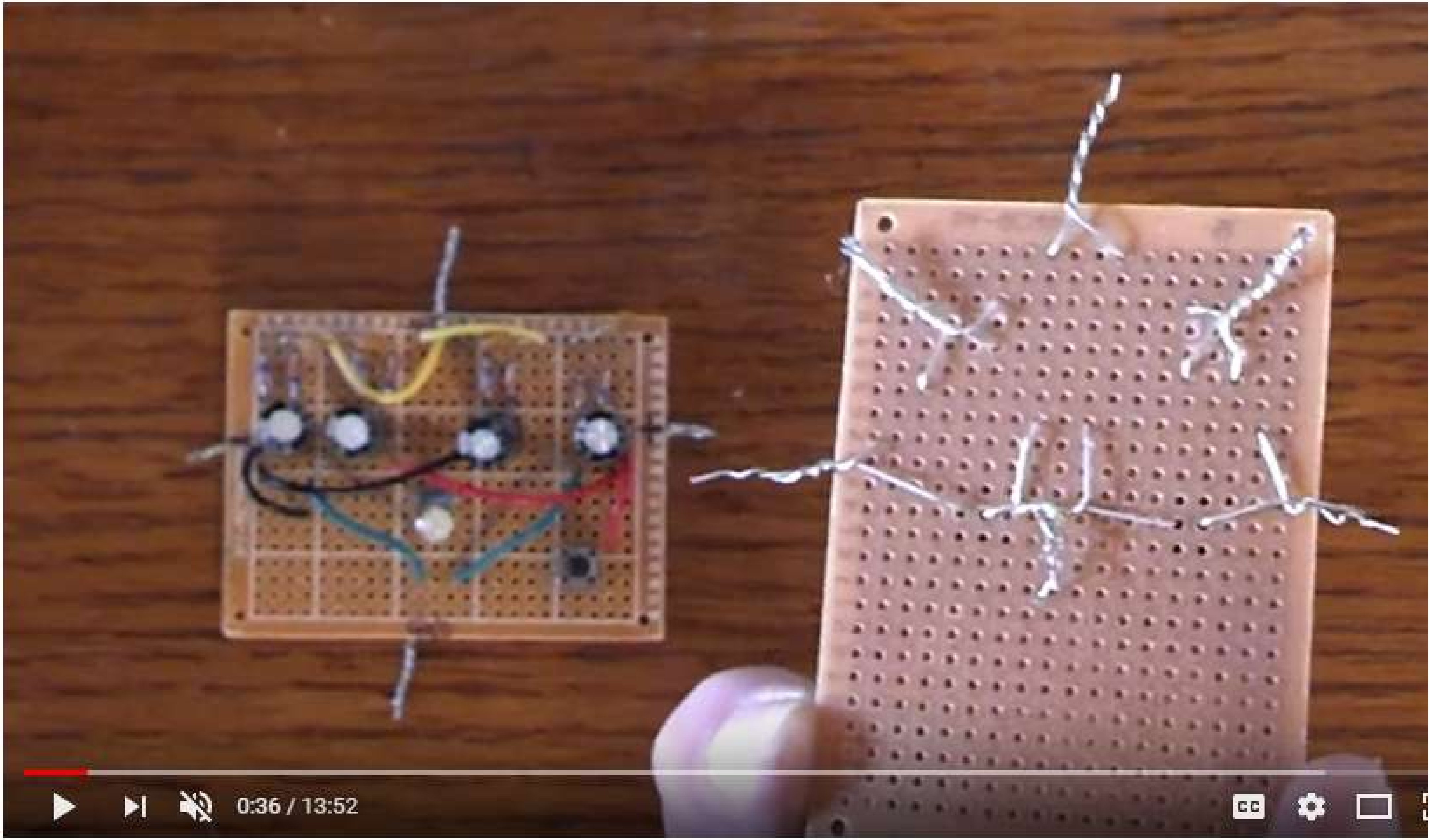

Tesla Free Energy Air Circuit

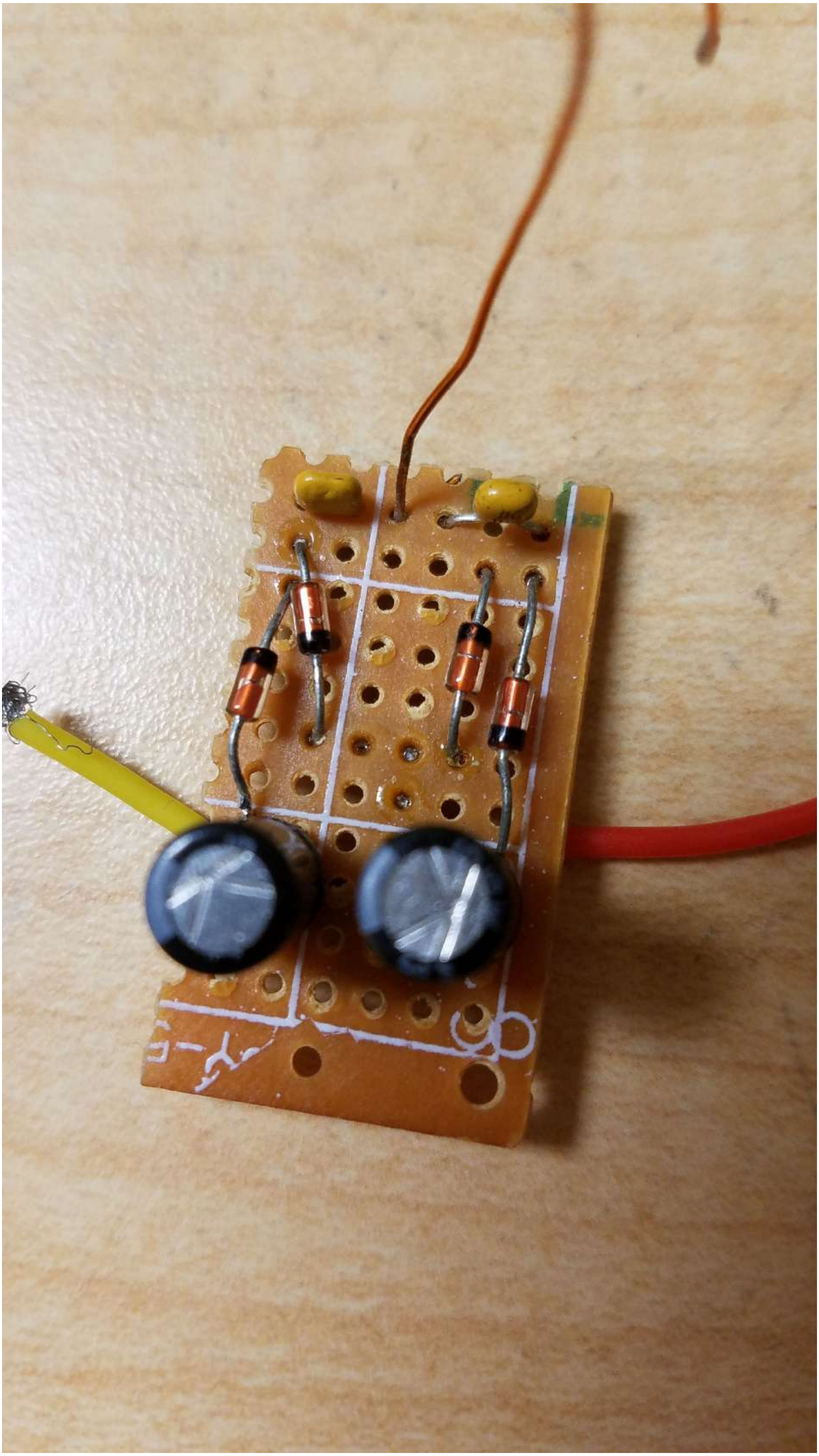

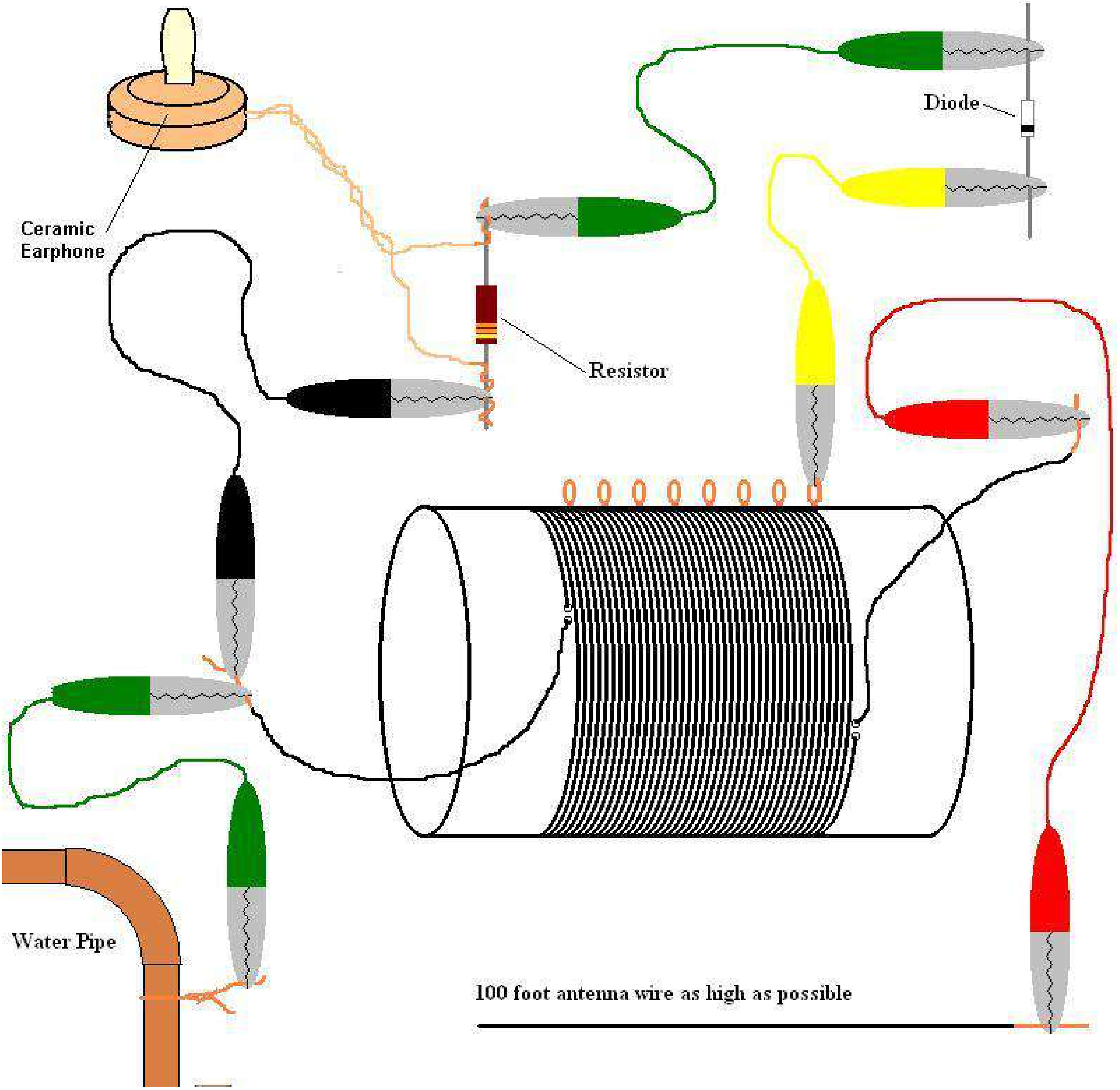
Ceramic
Earphone
Diode
Resistor
Water Pipe
100 foot antenna wire as high as possible

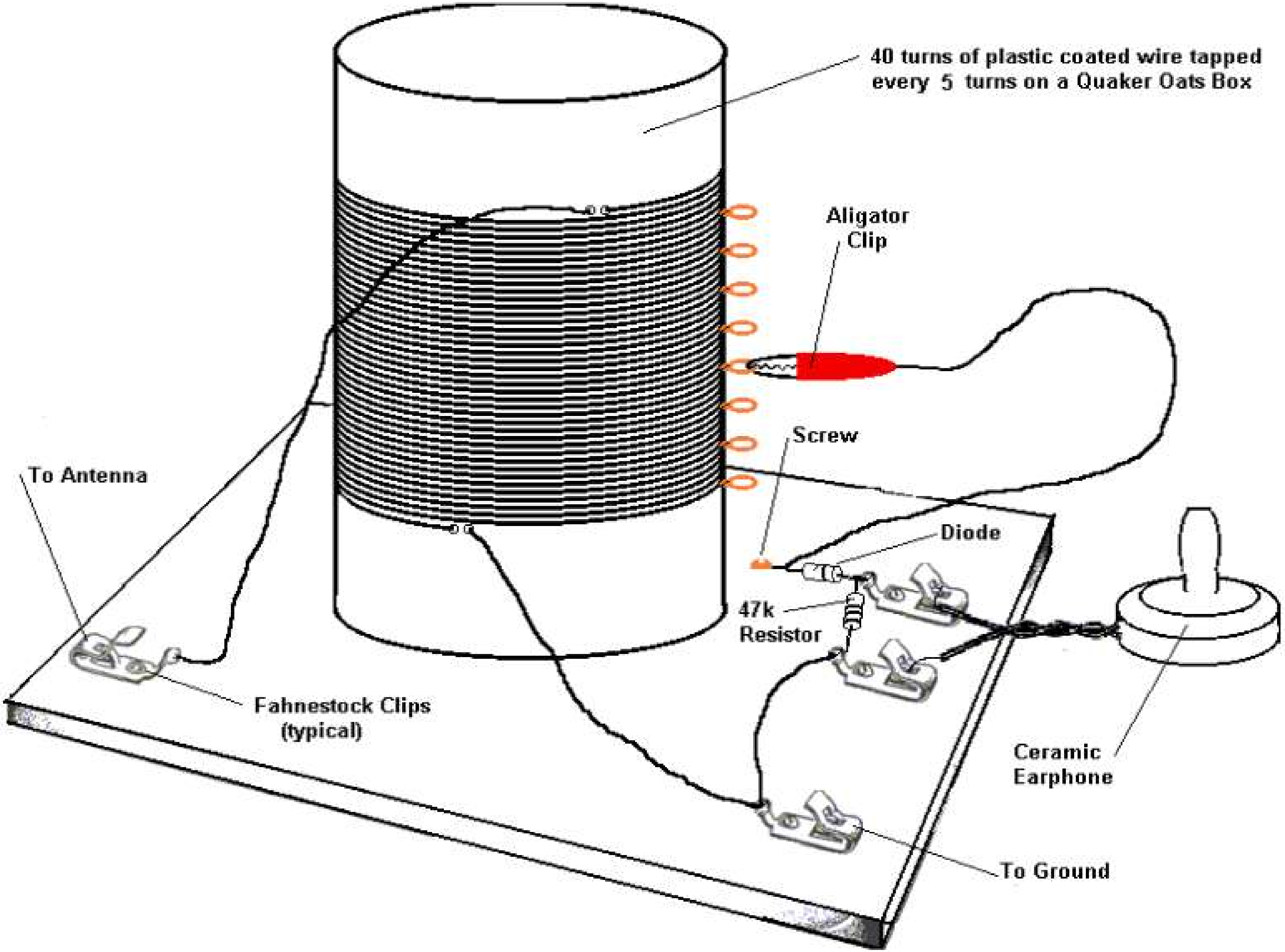
40 turns of plastic coated wire tapped every 5 turns on a Quaker Oats Box
Aligator Clip
Screw
Diode
47k Resistor
To Antenna
Fahnestock Clips (typical)
Ceramic Earphone
To Ground

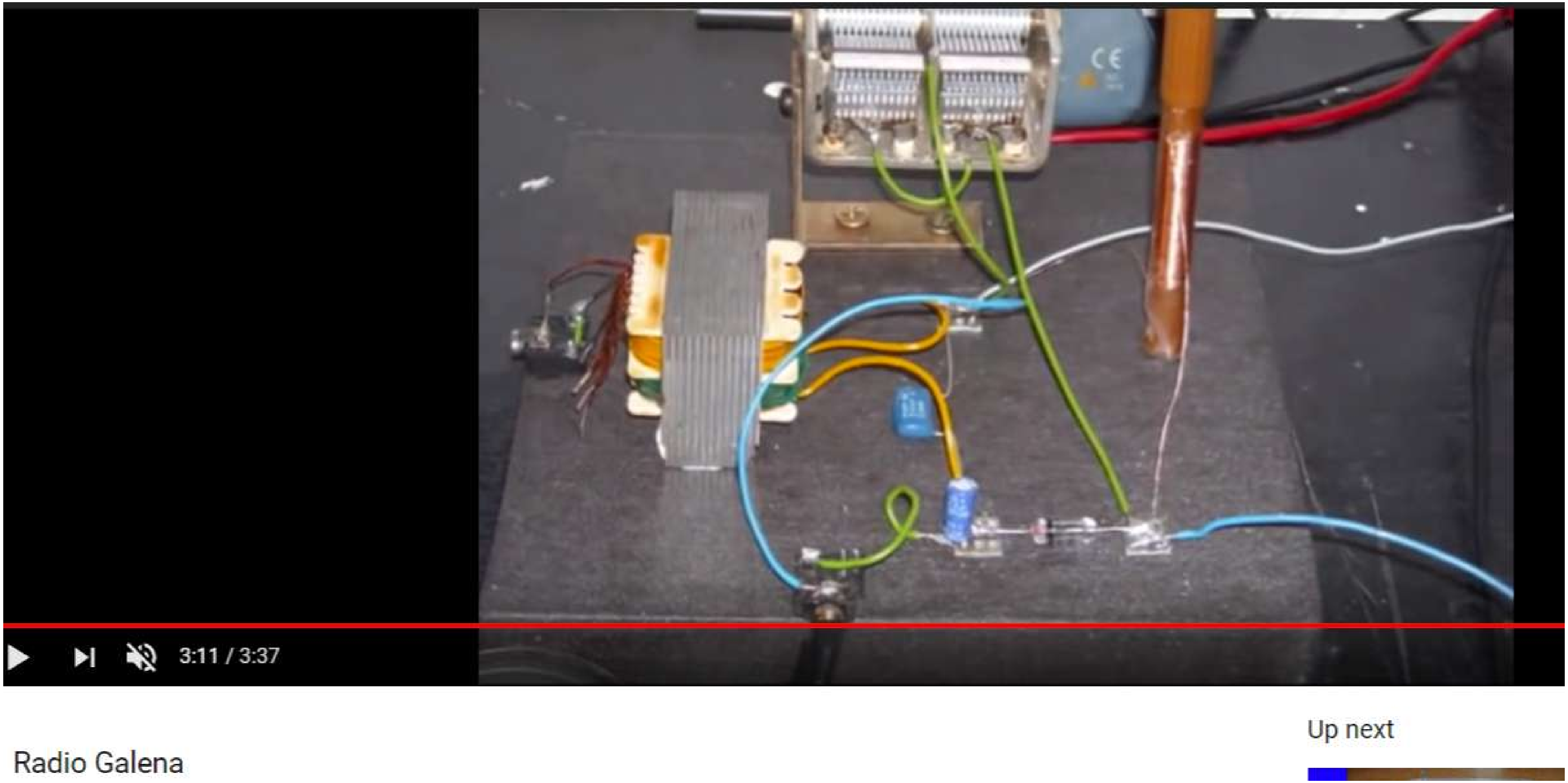
3:11 / 3:37
Up next

Radio Galena

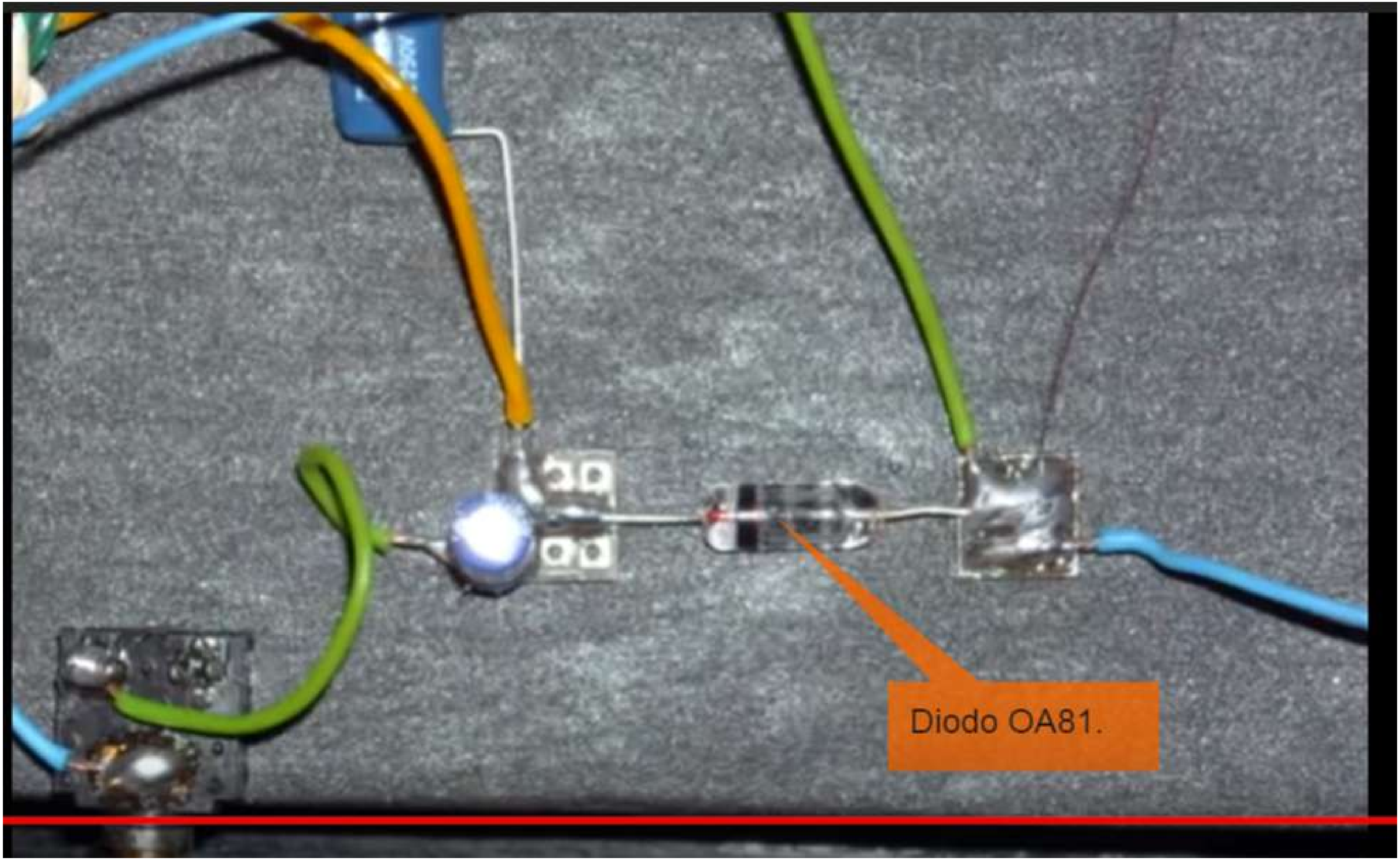
Diodo OA81.

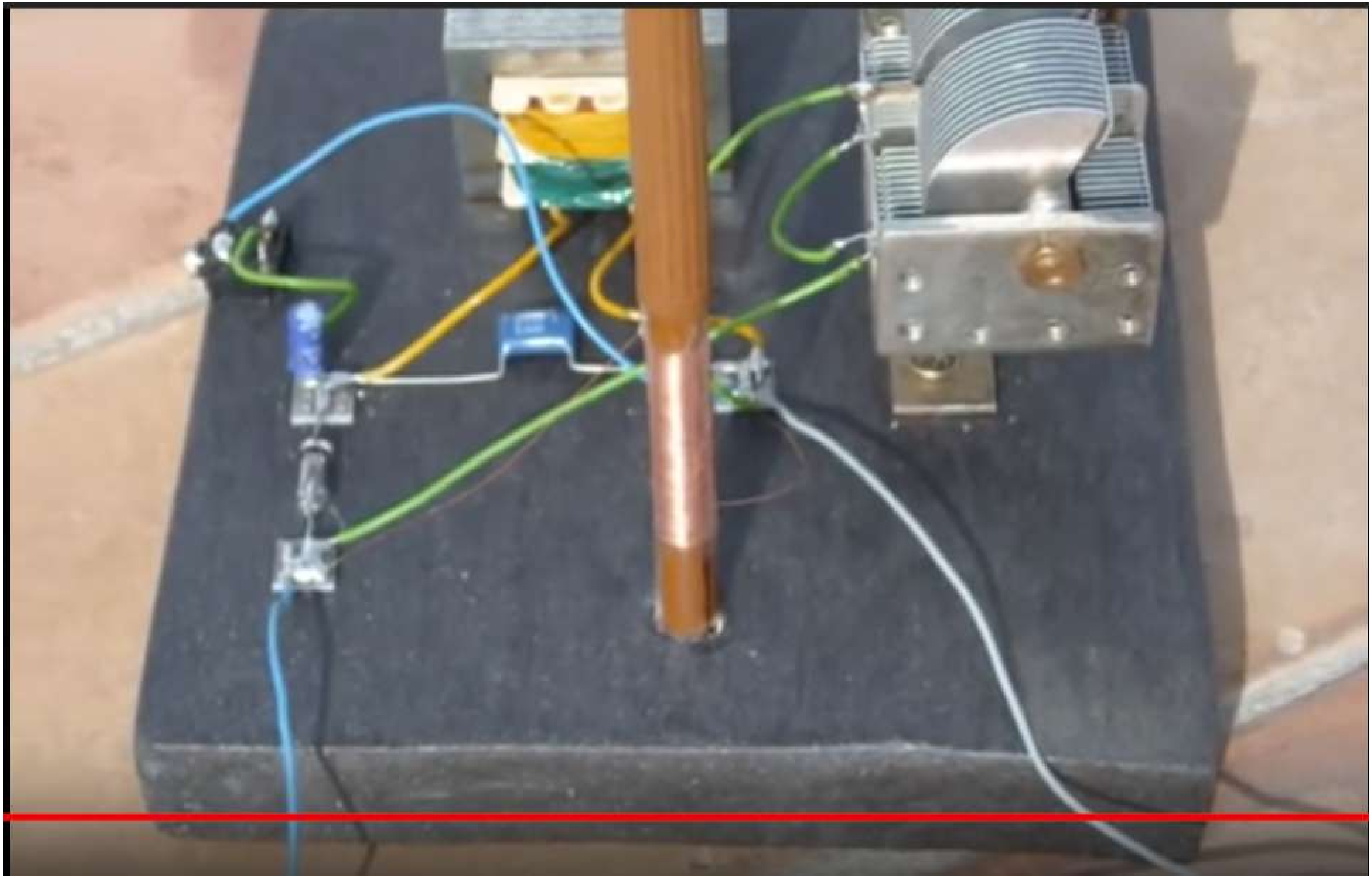

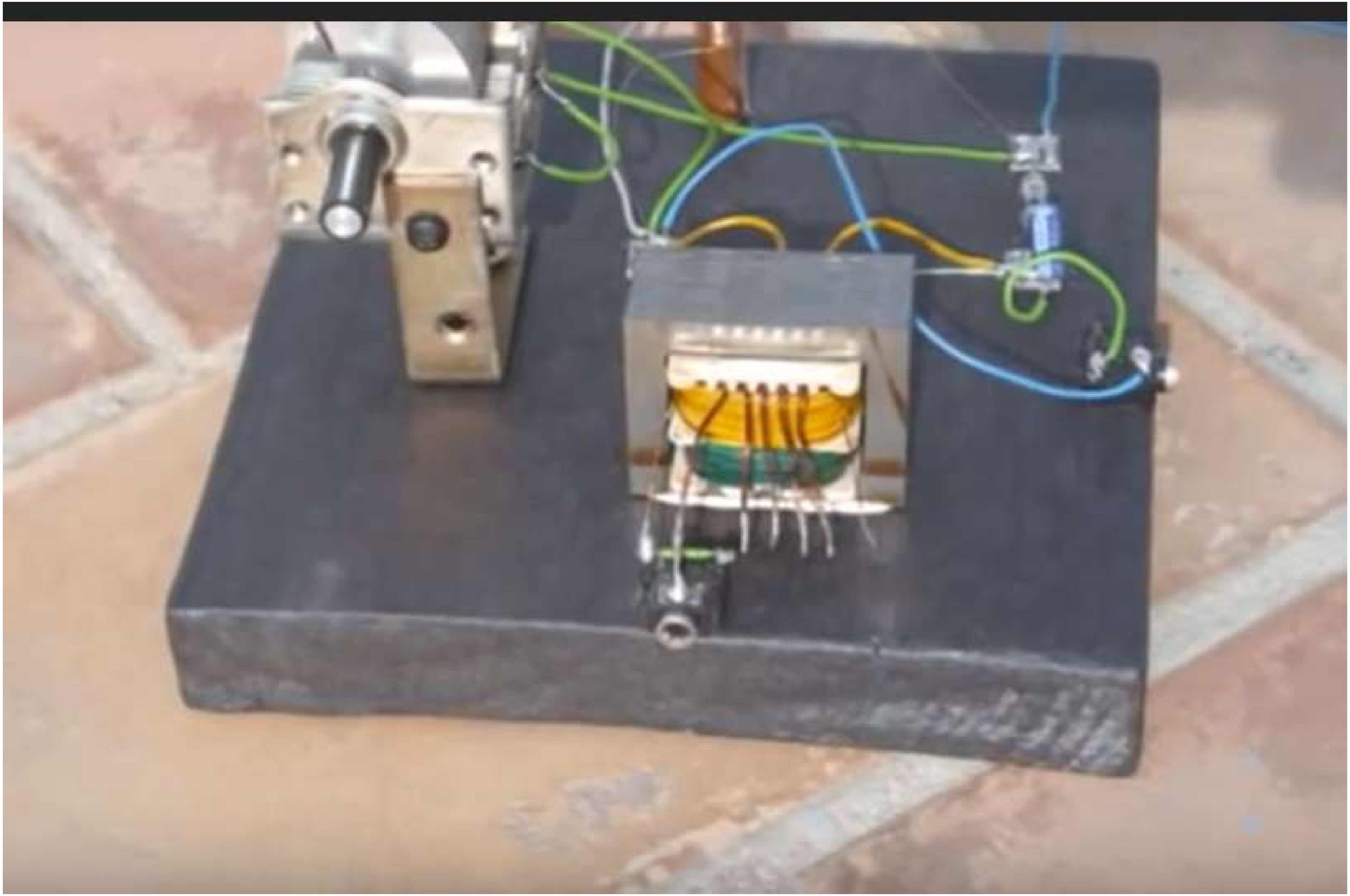

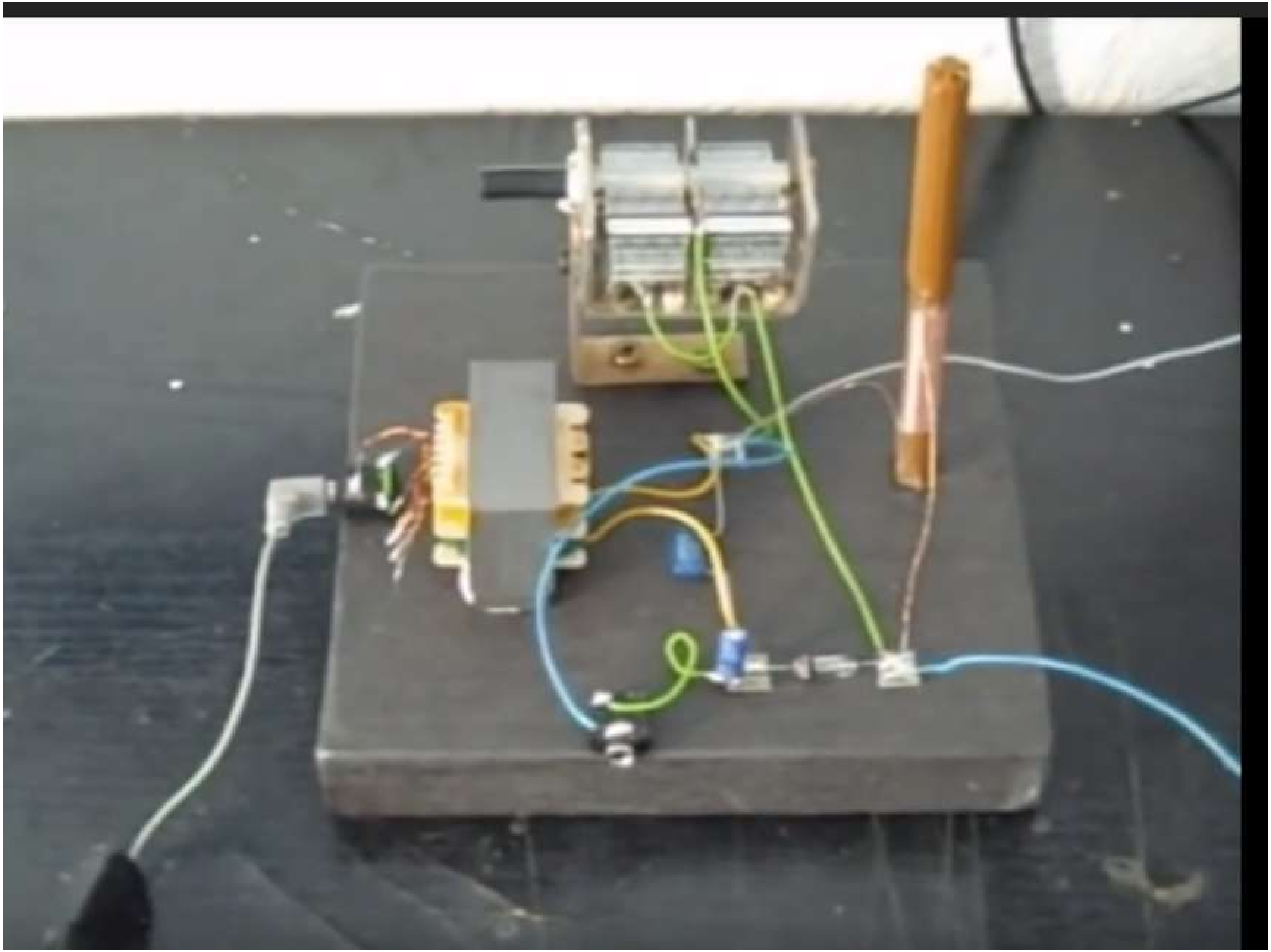

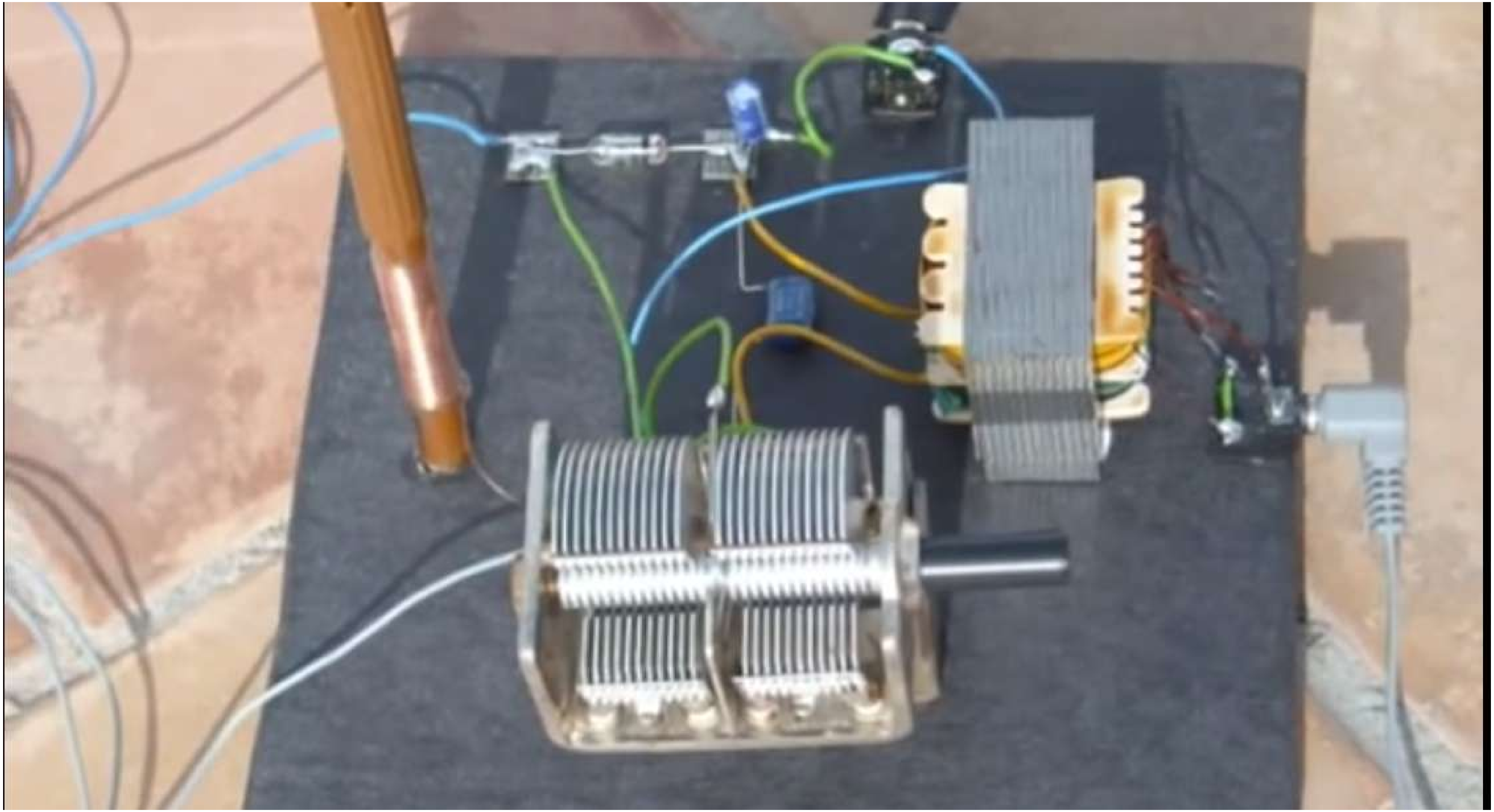

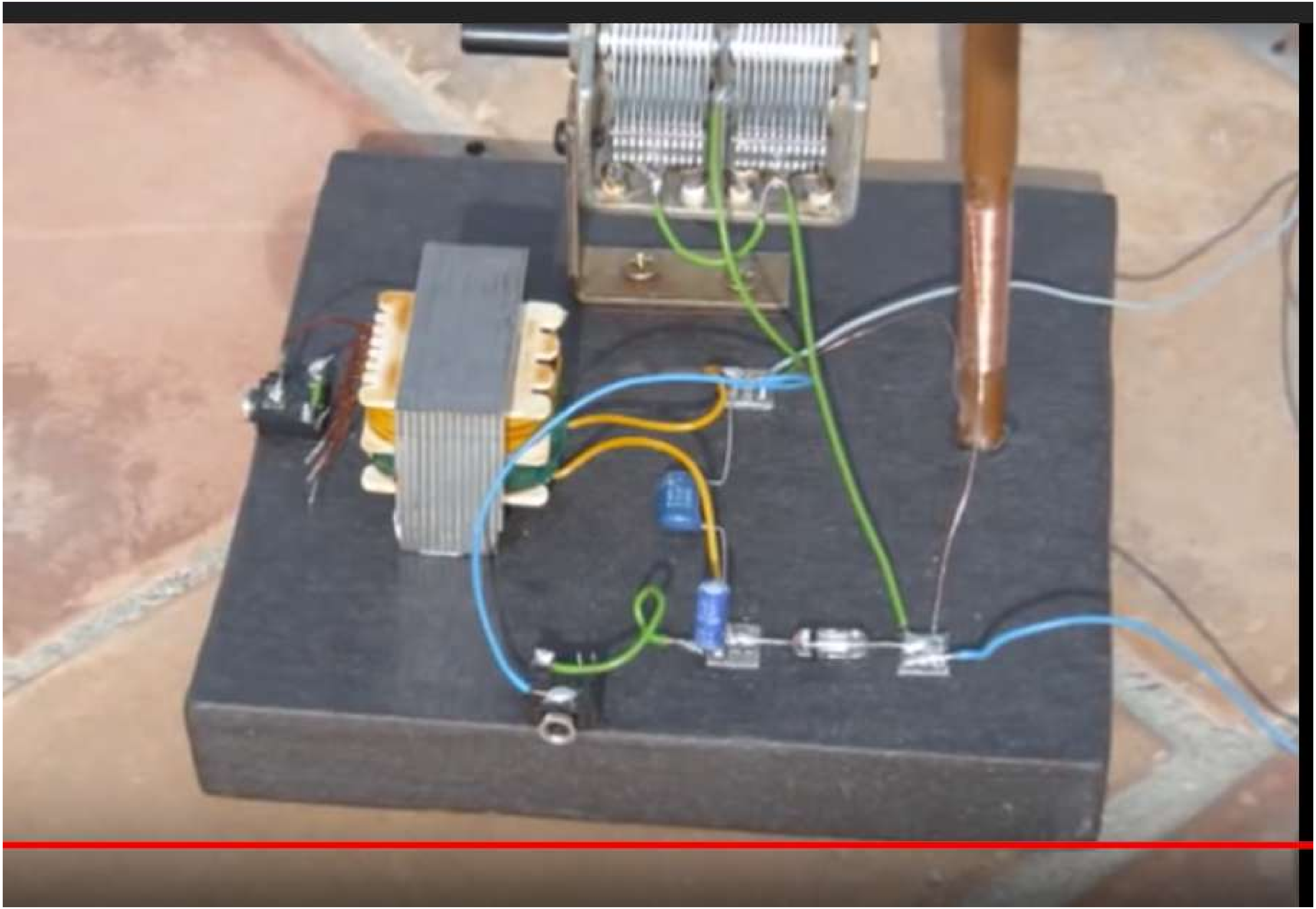

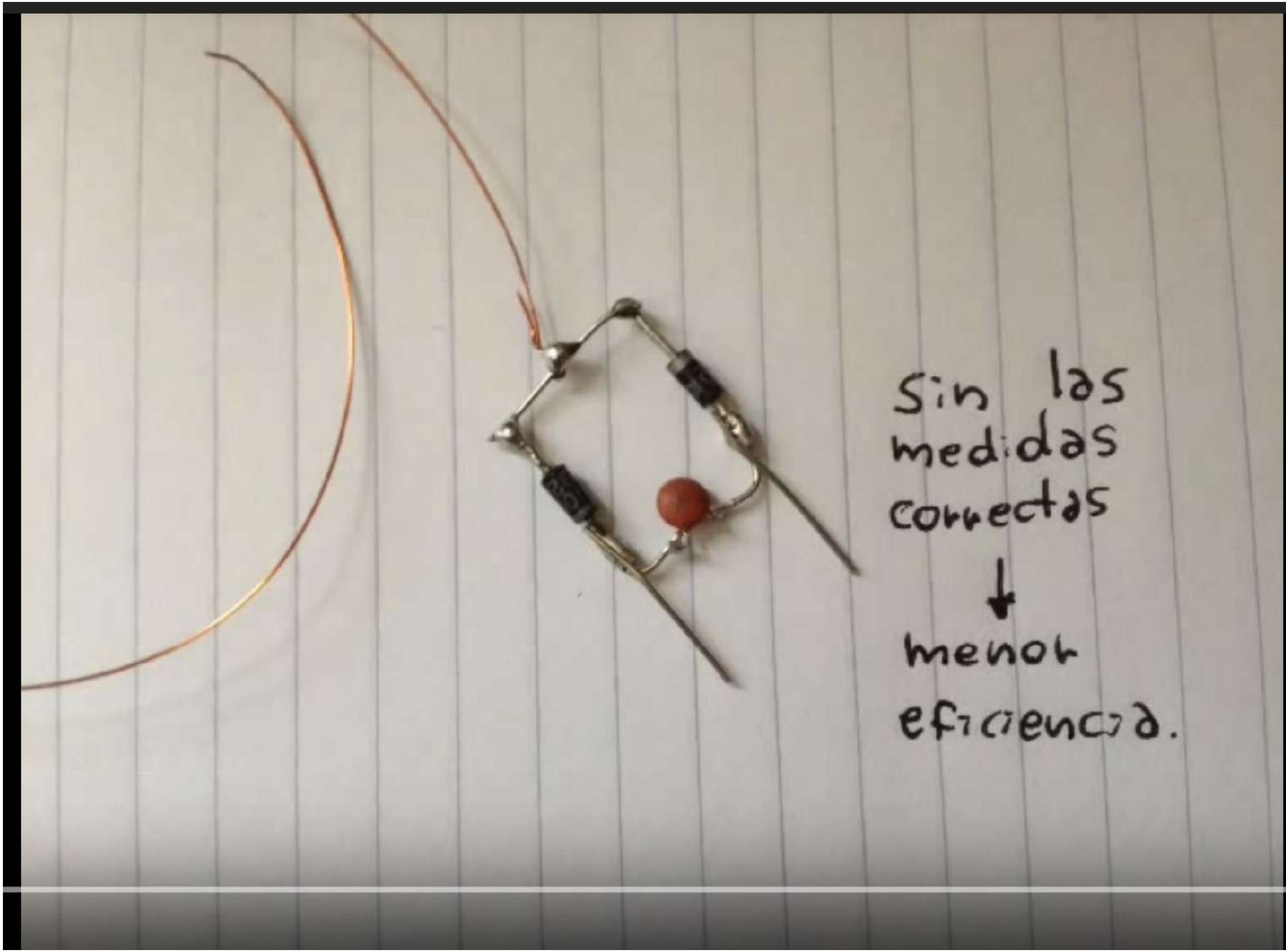

Sin las
medidas
correctas
↓
menor
eficiencia.

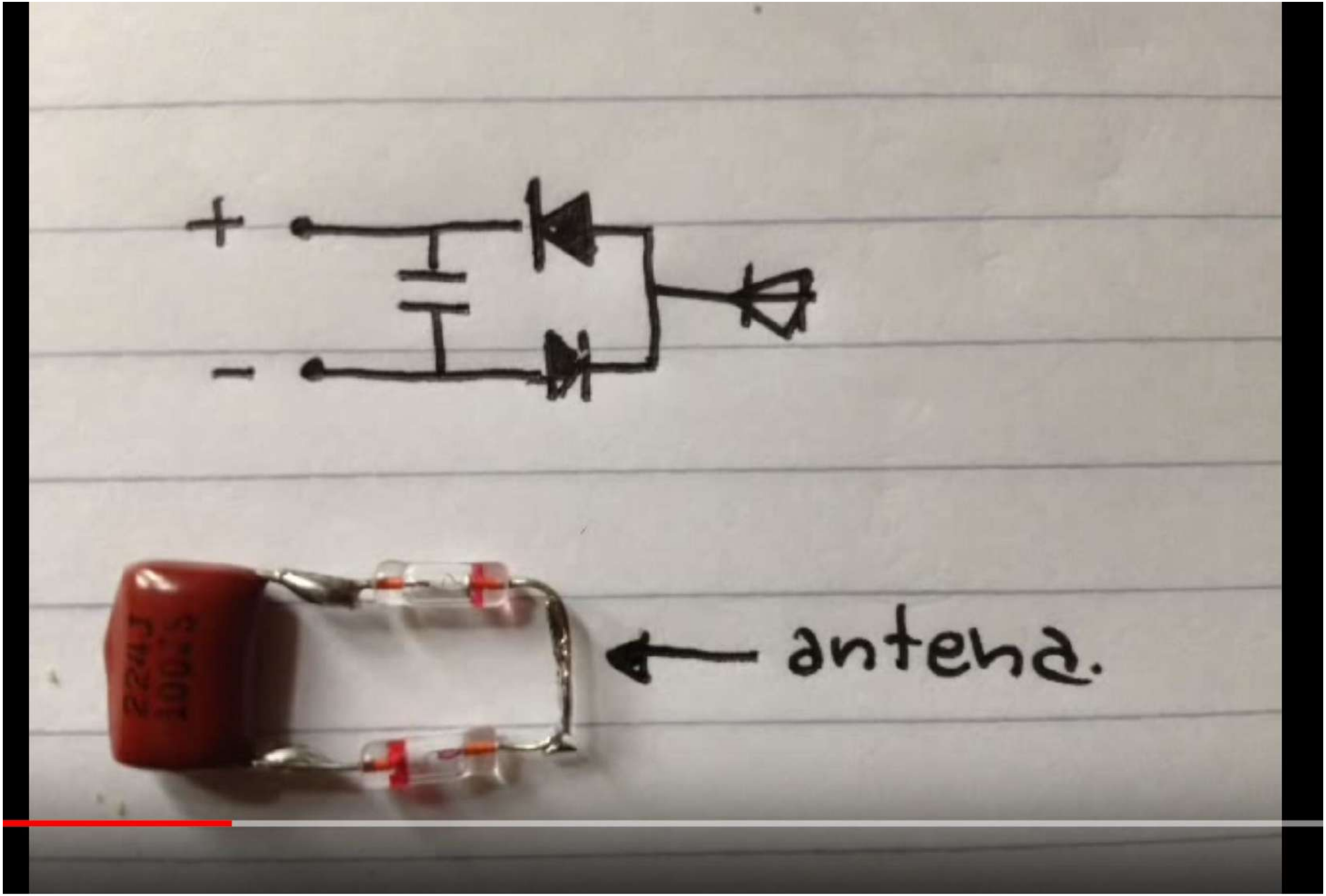

+
-
224J
100.5
antena.

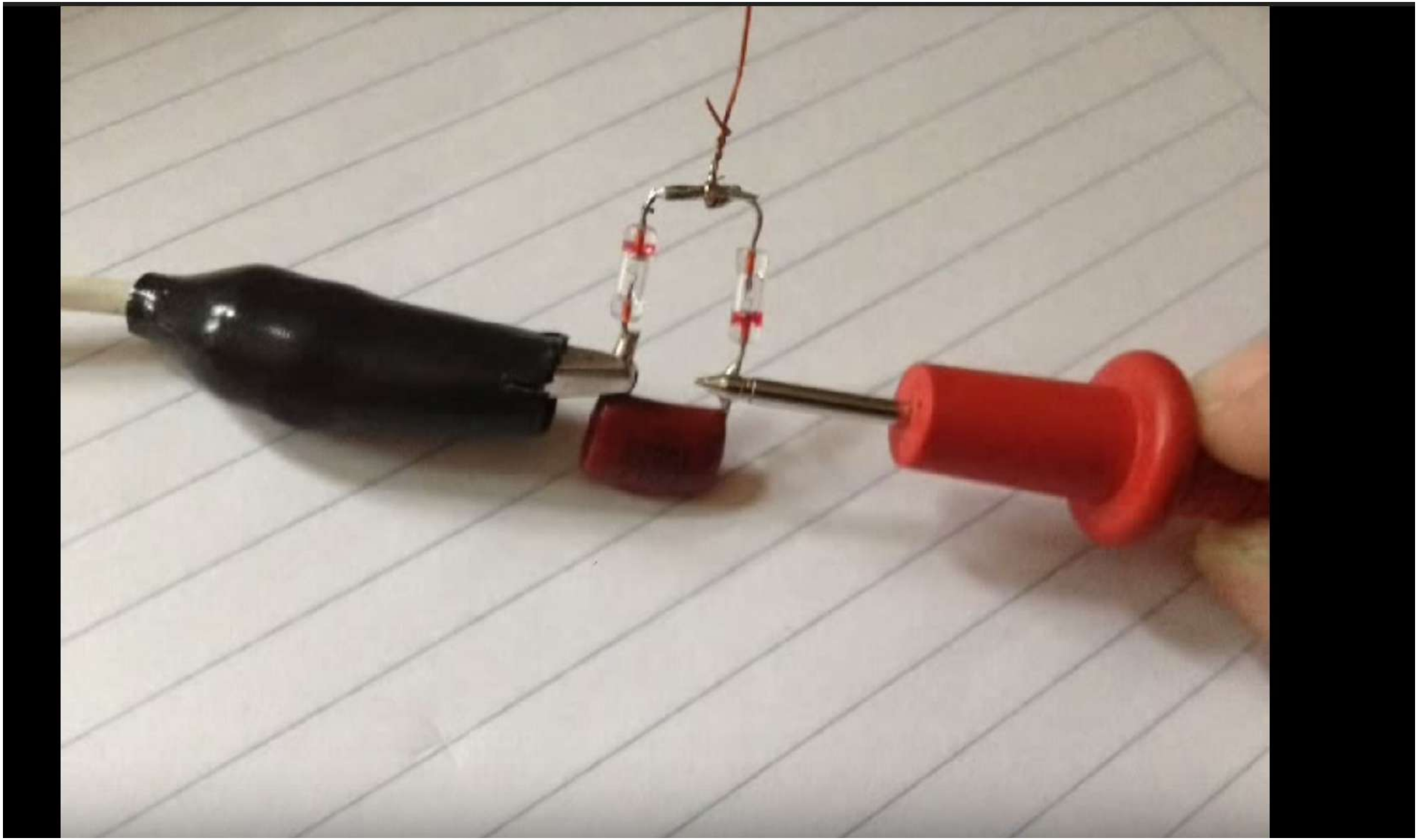

Ricevitore a cristallo per FM

Un progetto di Giacomo Cavuoti

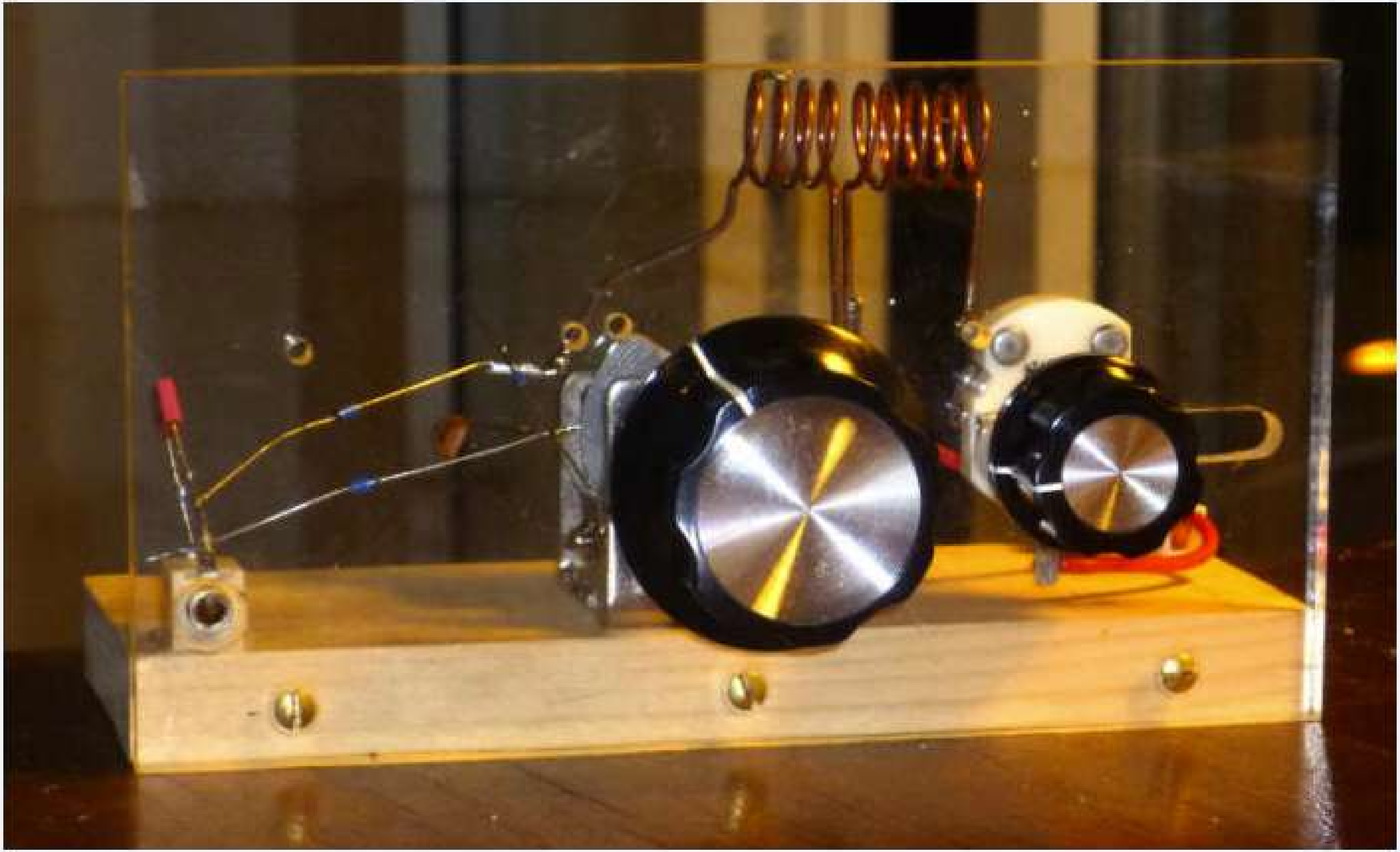

Seguendo l'invito di Leonardo ho realizzato questo semplice ricevitore a cristallo per FM (fare clic sullo schema qua sotto per vederlo ingrandito).

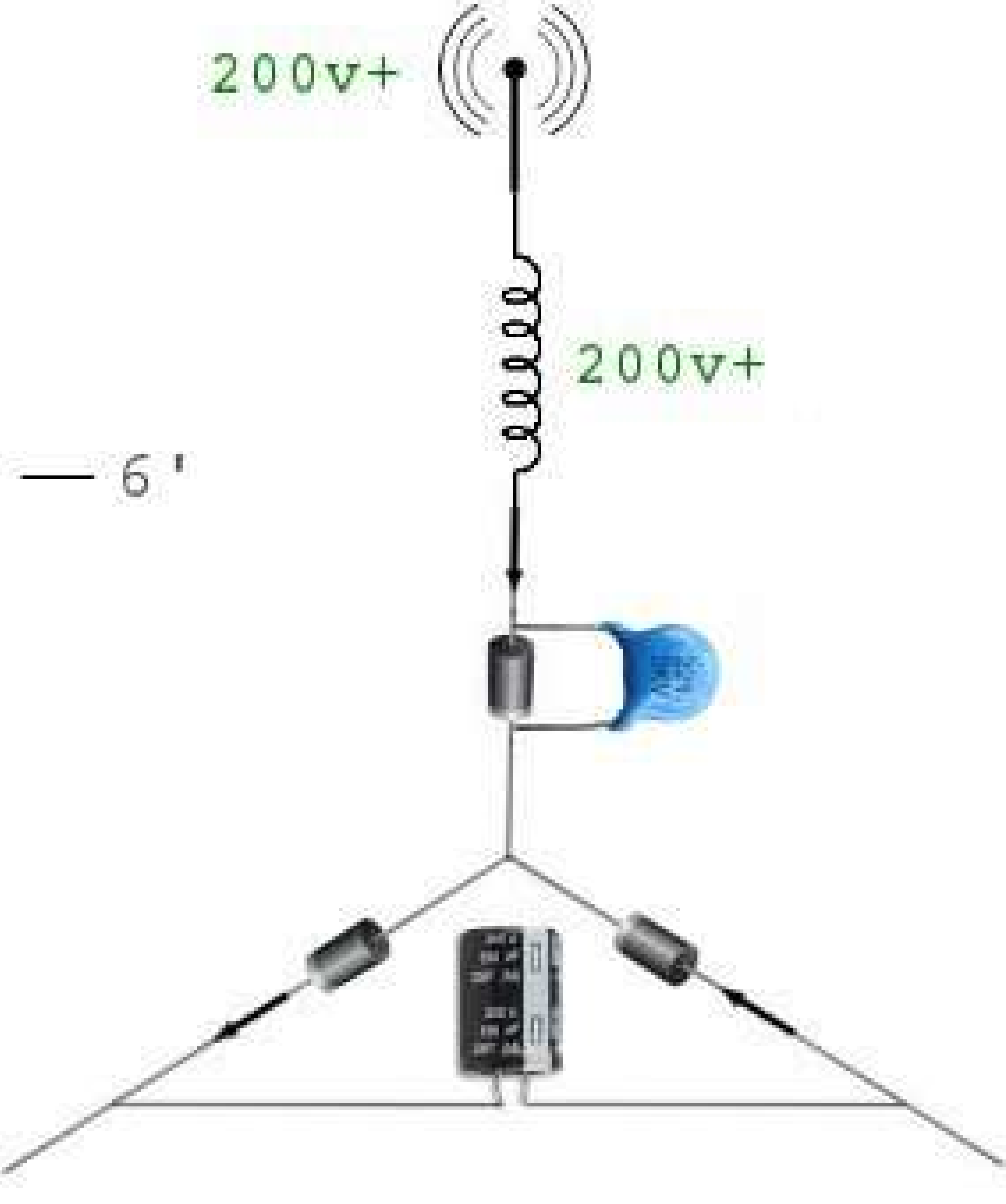
200v+
200v+
— 6'

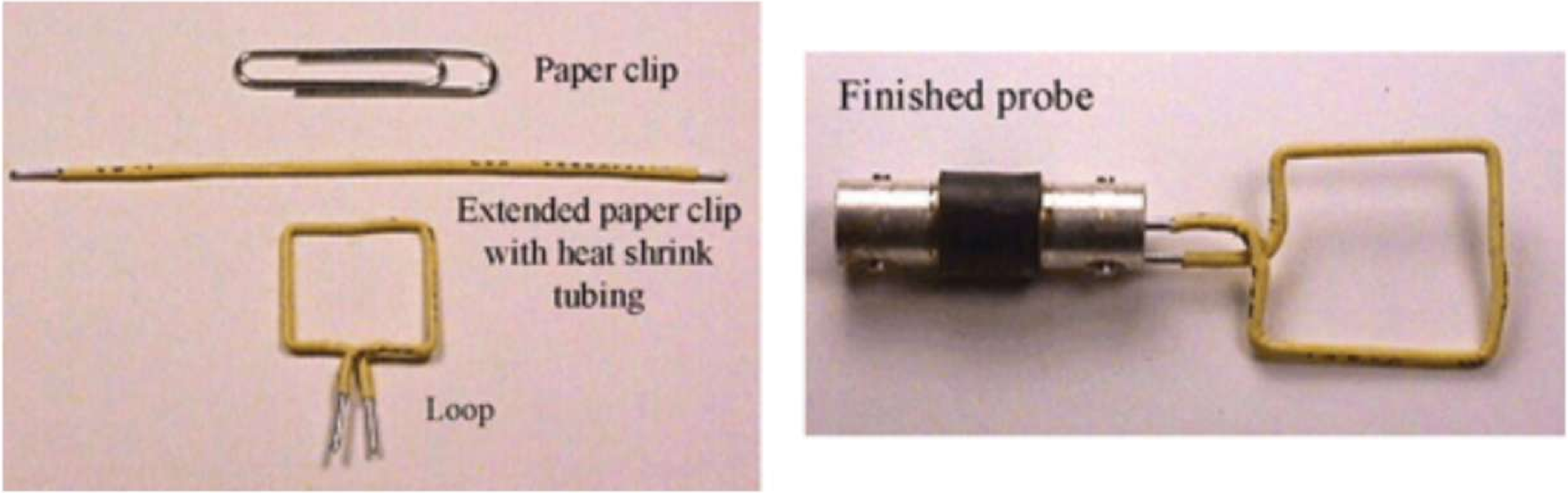
Paper clip
Extended paper clip with heat shrink tubing
Loop
Finished probe

Figure 1: Magnetic field probe build from a paper clip.

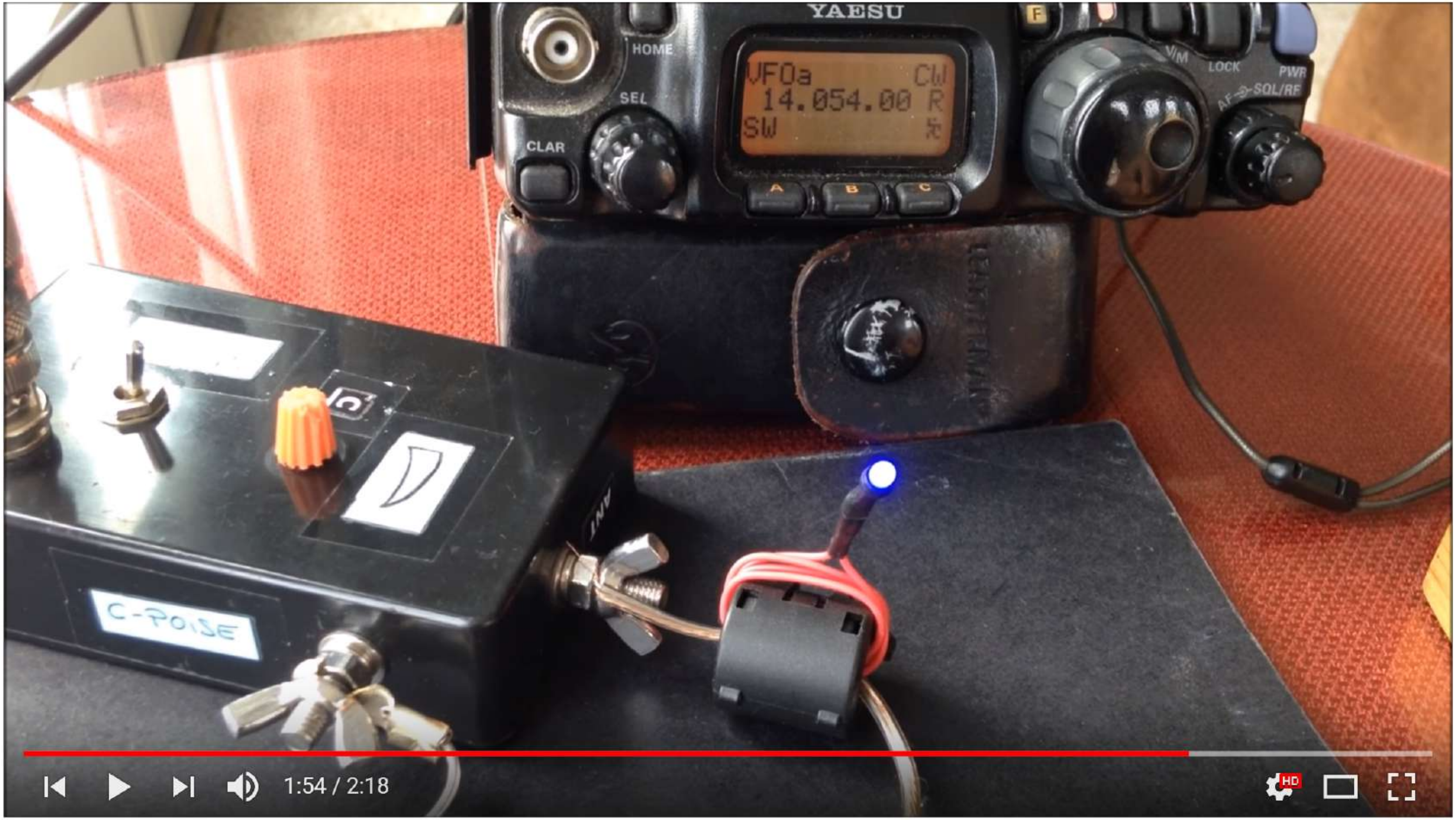
YAESU
HOME
SEL
CLAR
VFOa CW
14.054.00 R
SW
A
B
C
V/M
LOCK
PWR
AF-SQL/RF
F
C-POISE
ANT
1:54 / 2:18
HD

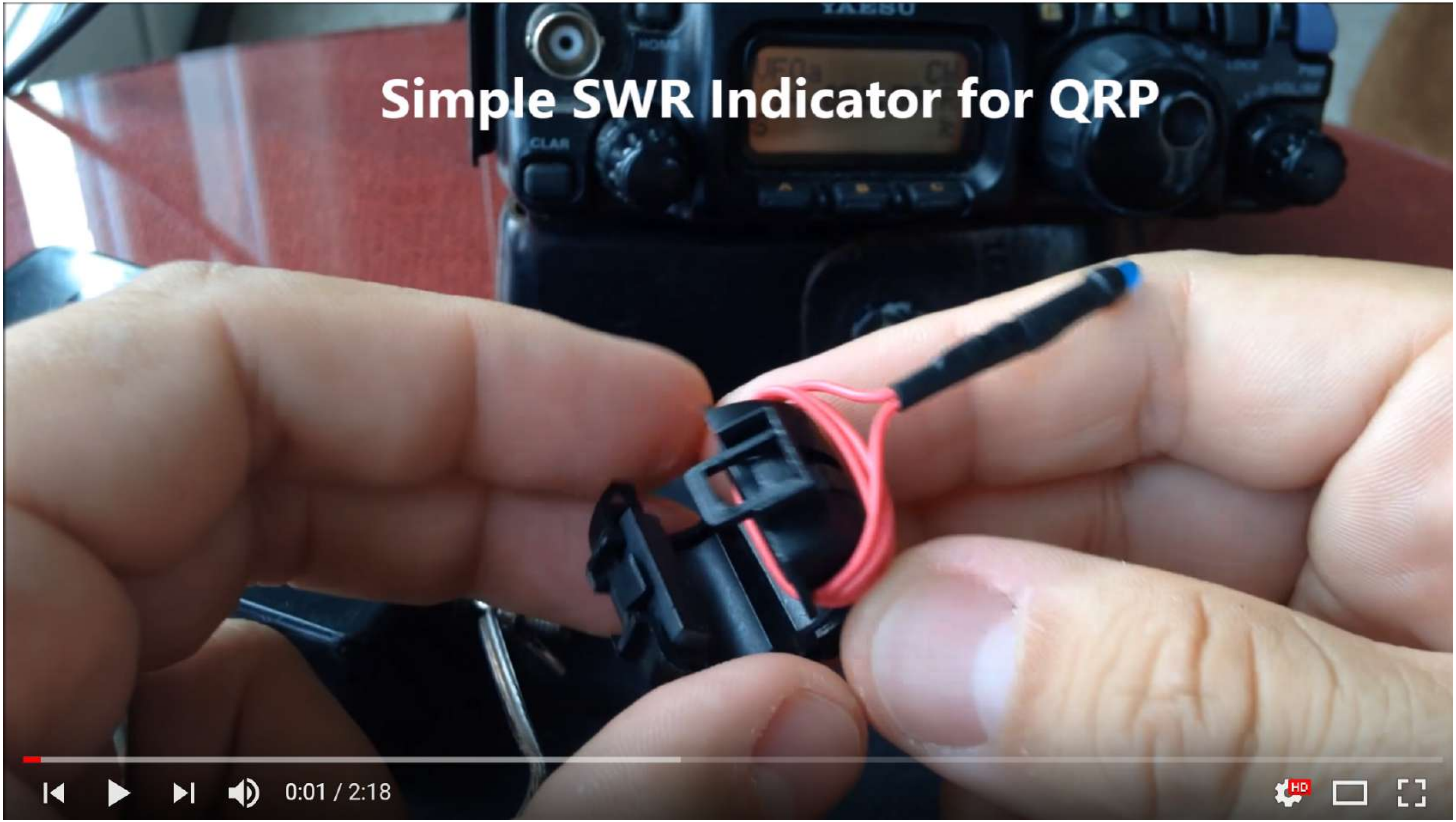

Simple SWR Indicator for QRP

YAESU
CLAR
0:01 / 2:18
HD

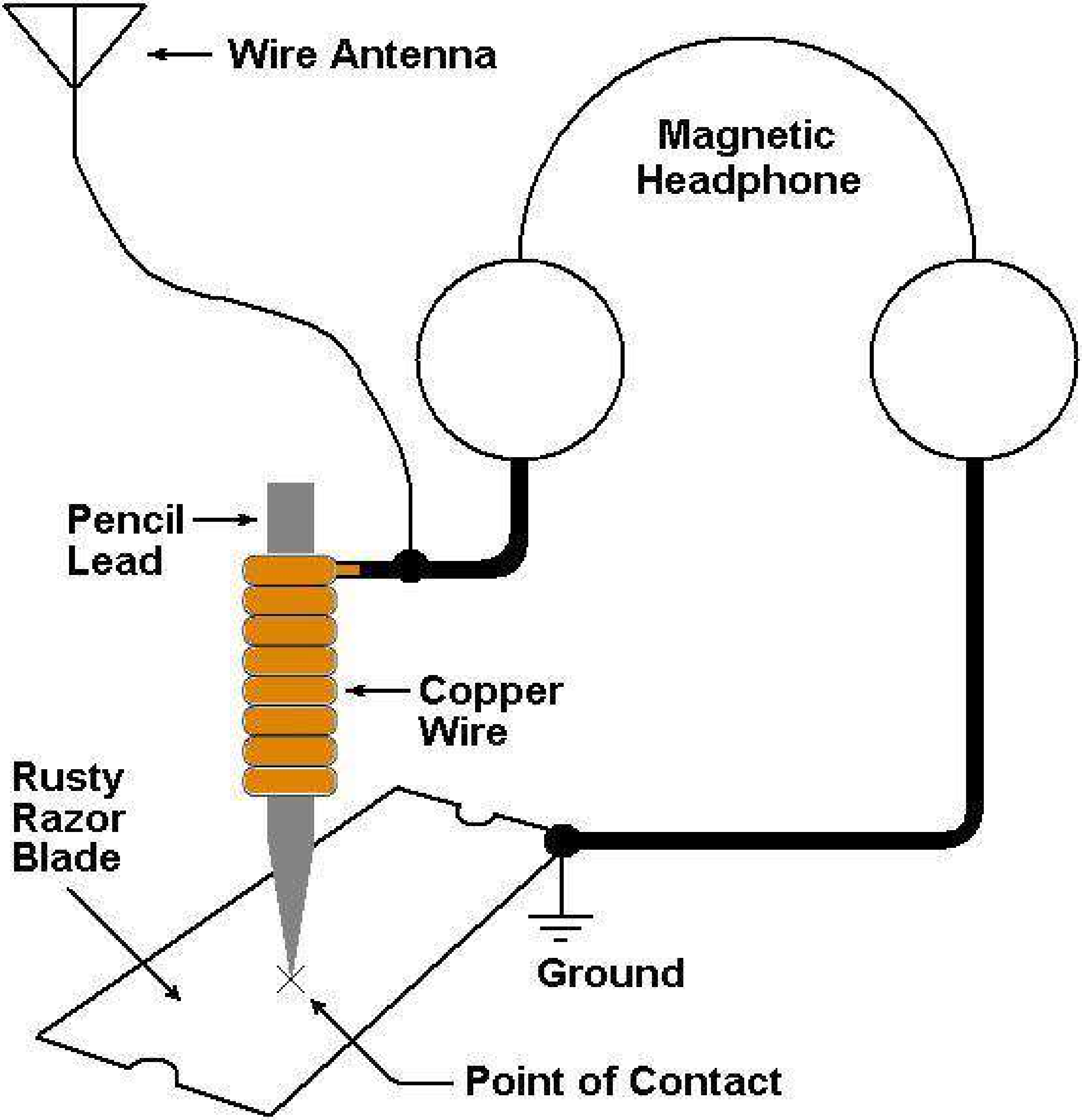
Wire Antenna
Magnetic Headphone
Pencil Lead
Copper Wire
Rusty Razor Blade
Ground
Point of Contact

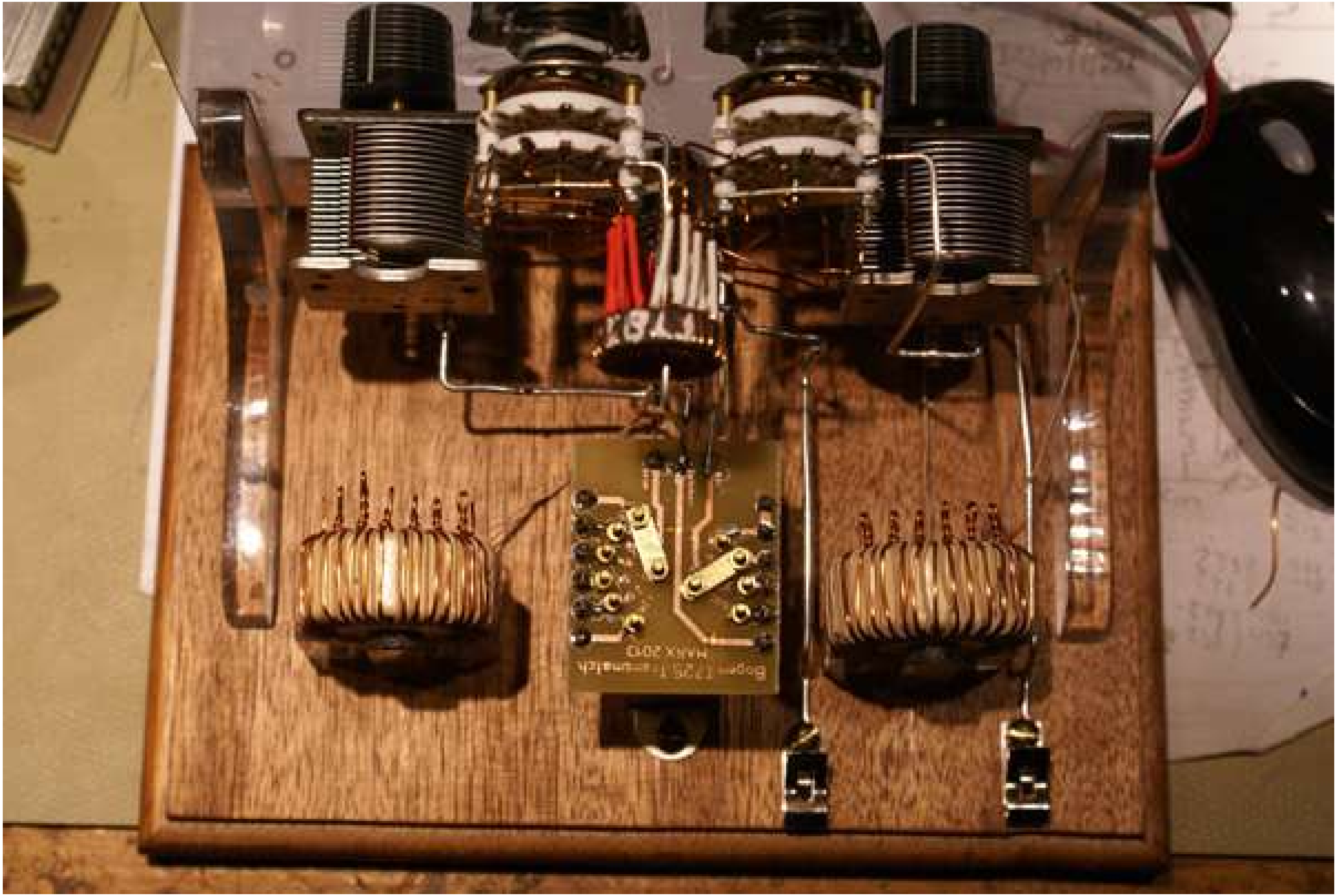
Bogen T225 Transmatch
MARX 2013

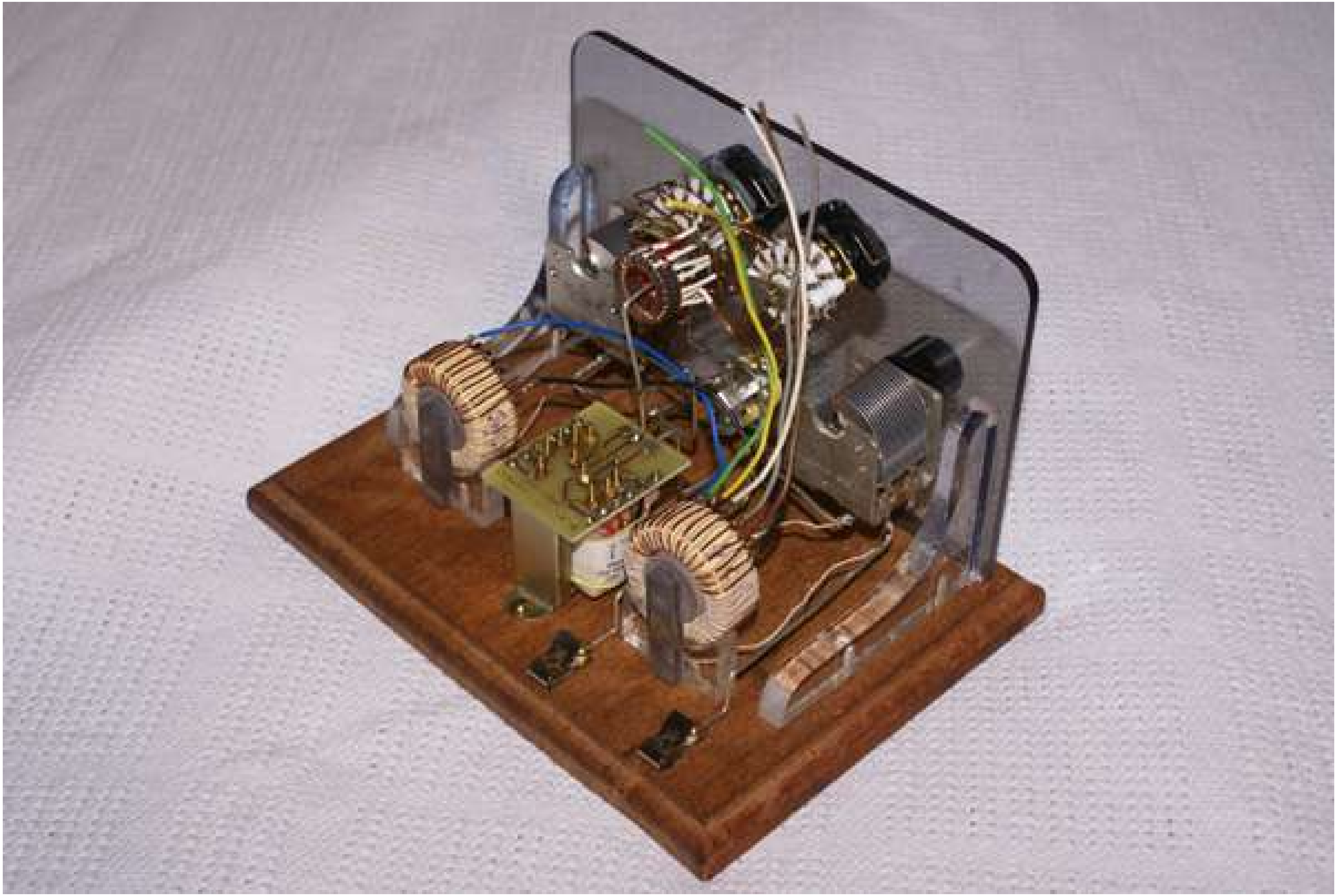

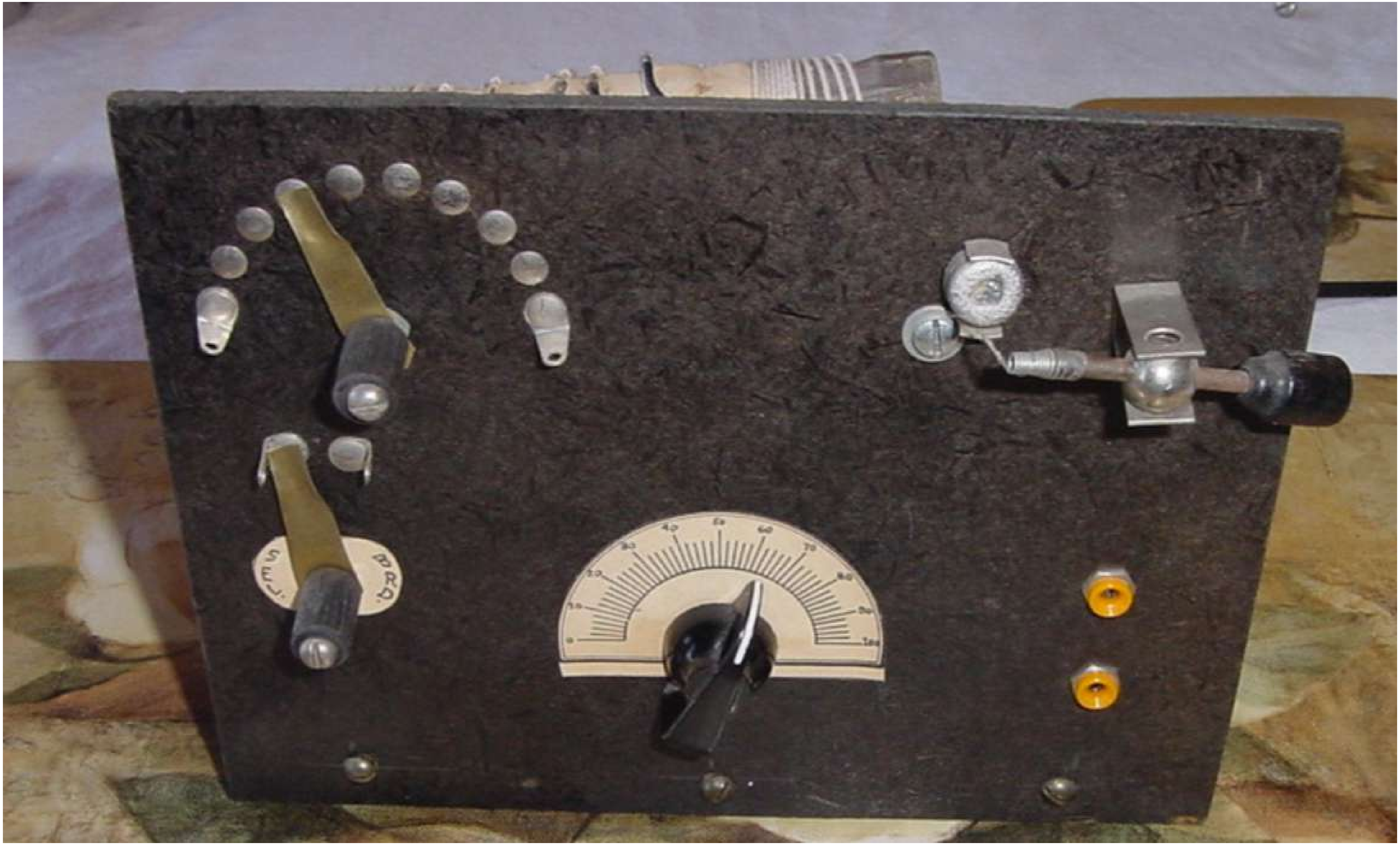

SEL.
BRD.
0
10
20
30
40
50
60
70
80
90
100

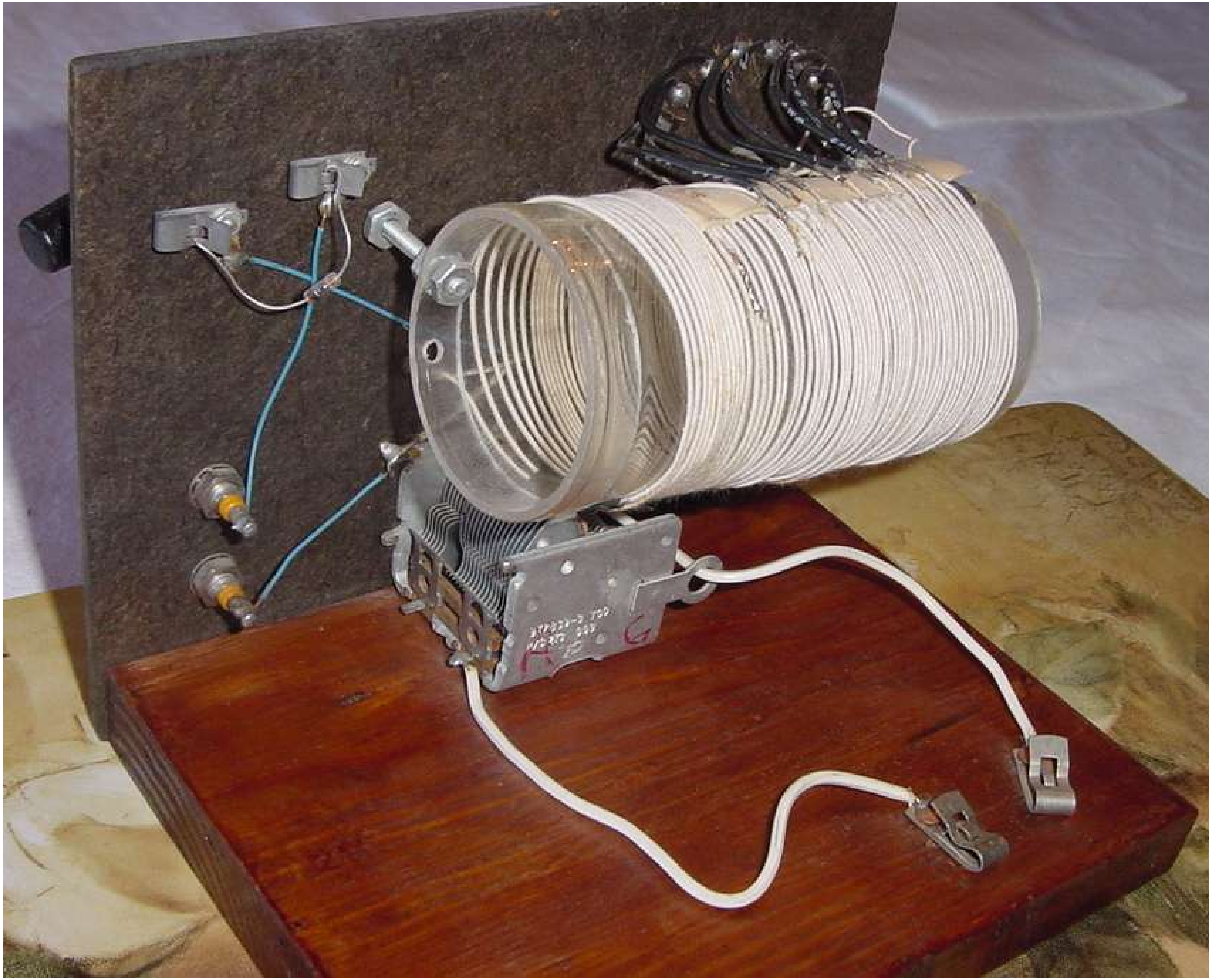
A
G

Another Prototype similar to the Heathkit using toroid coil

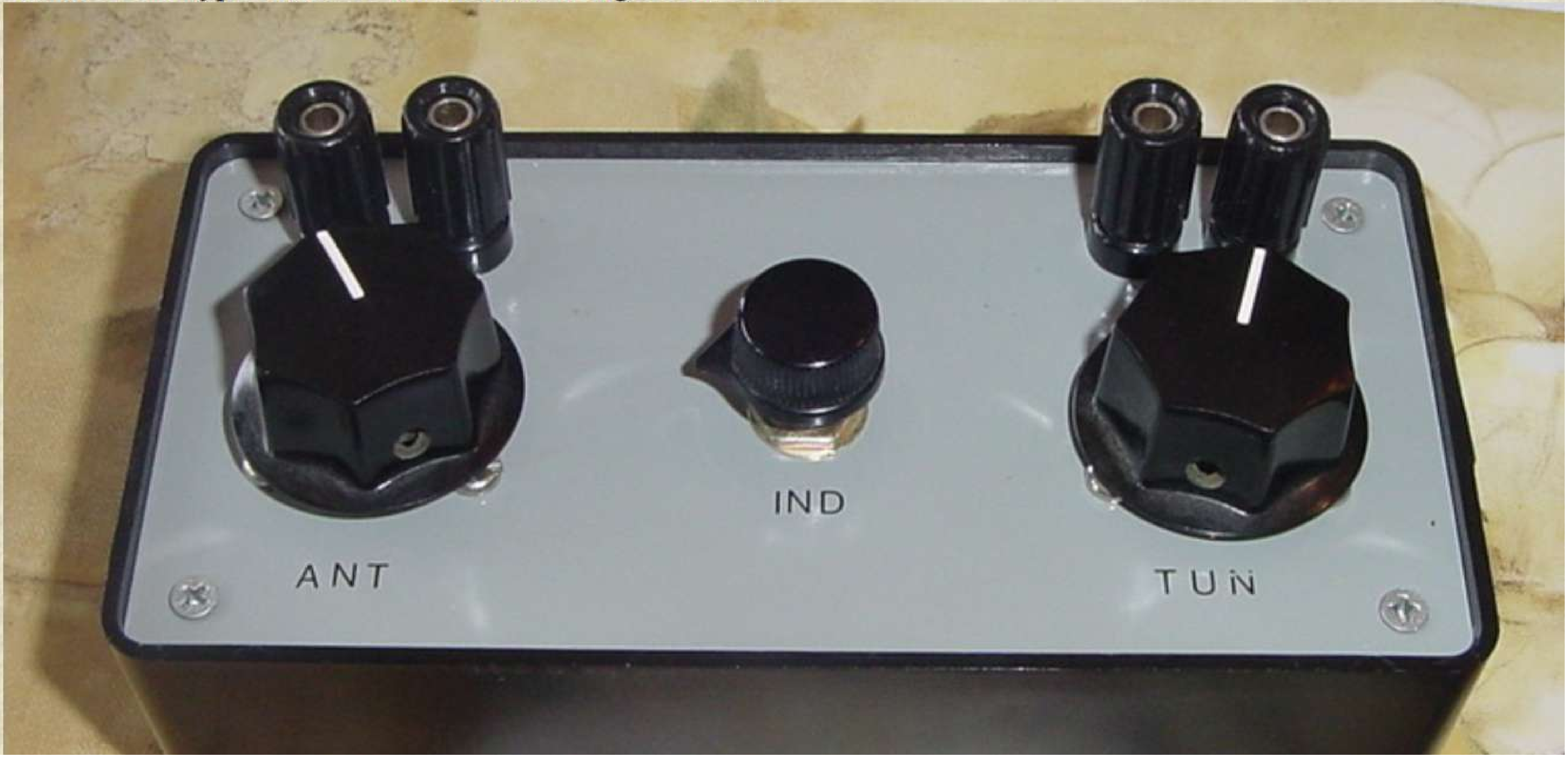

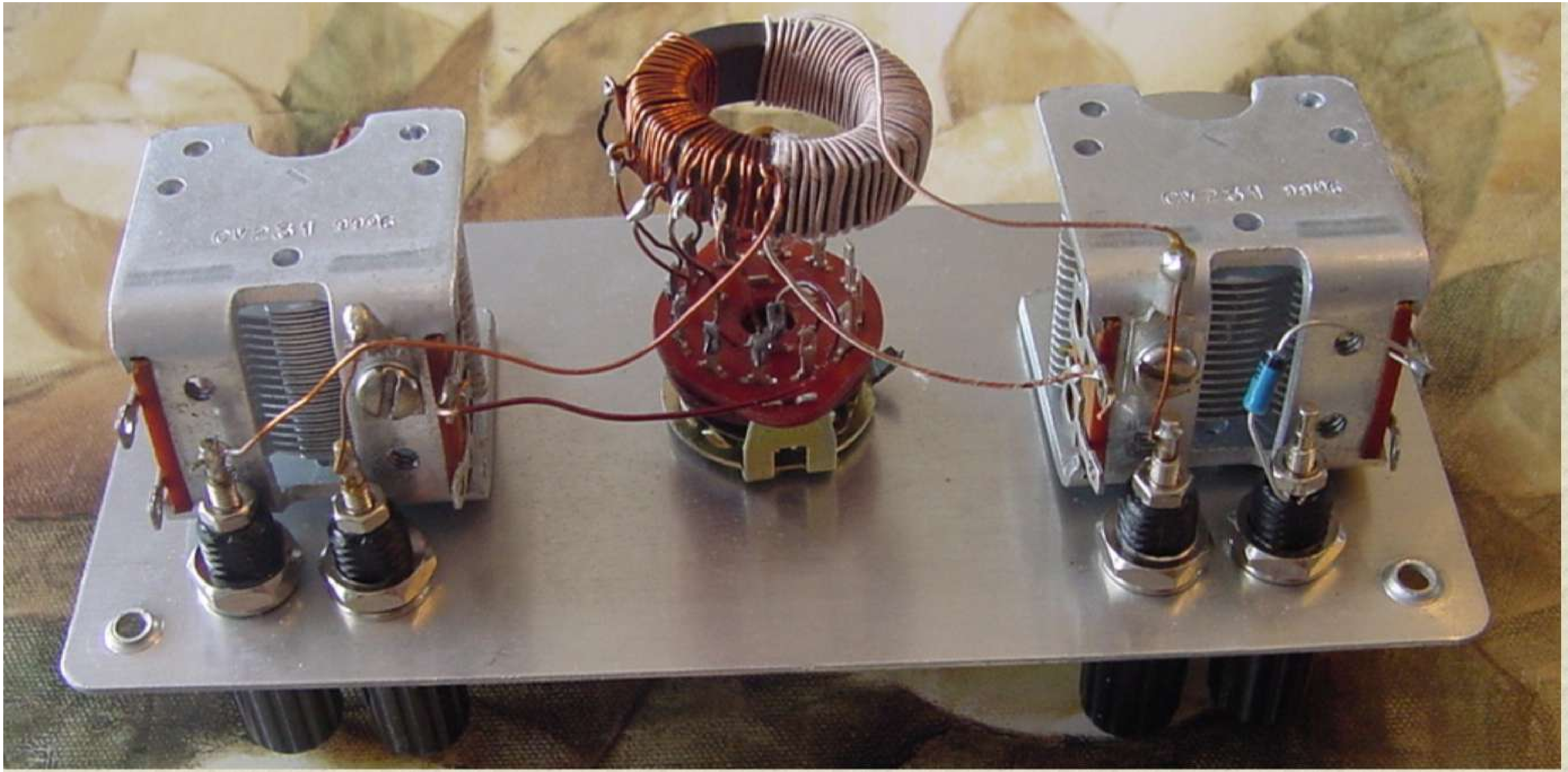

This is a dual tuned crystal set made in a box like the CR-1, just a fun prototype I made a few years back. It pulls in AM stations loud and clear.

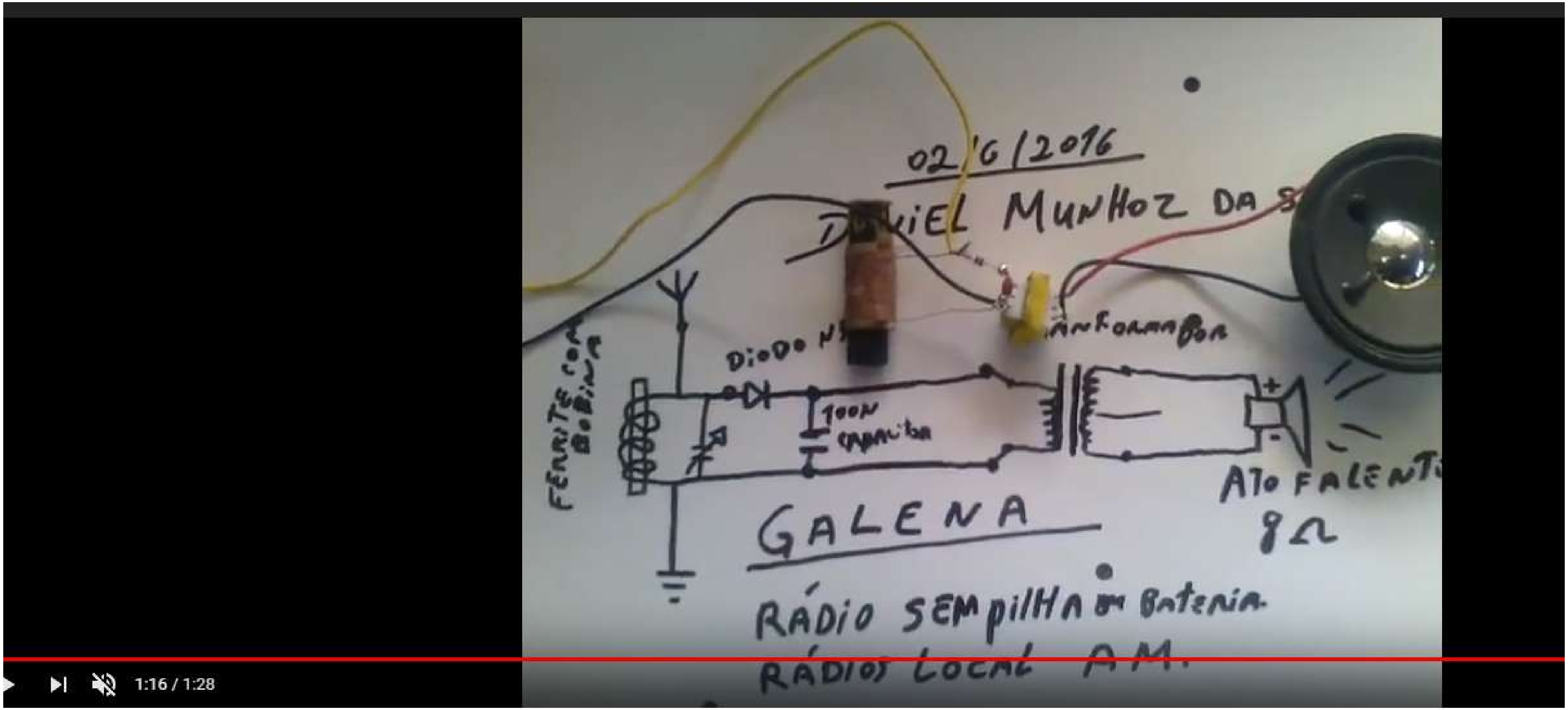
02/6/2016
DANIEL MUNHOZ DA S
DiODO N
FERRITE COM BOBINA
100N capacitor
TRANSFORMADOR
+
-
ALTO FALENTE
8Ω
GALENA
RÁDIO SEM pilHA ou BATERIA.
RÁDIOS LOCAL AM.
1:16 / 1:28

Up next

GALENA , Rádio AM ,sem pilha.

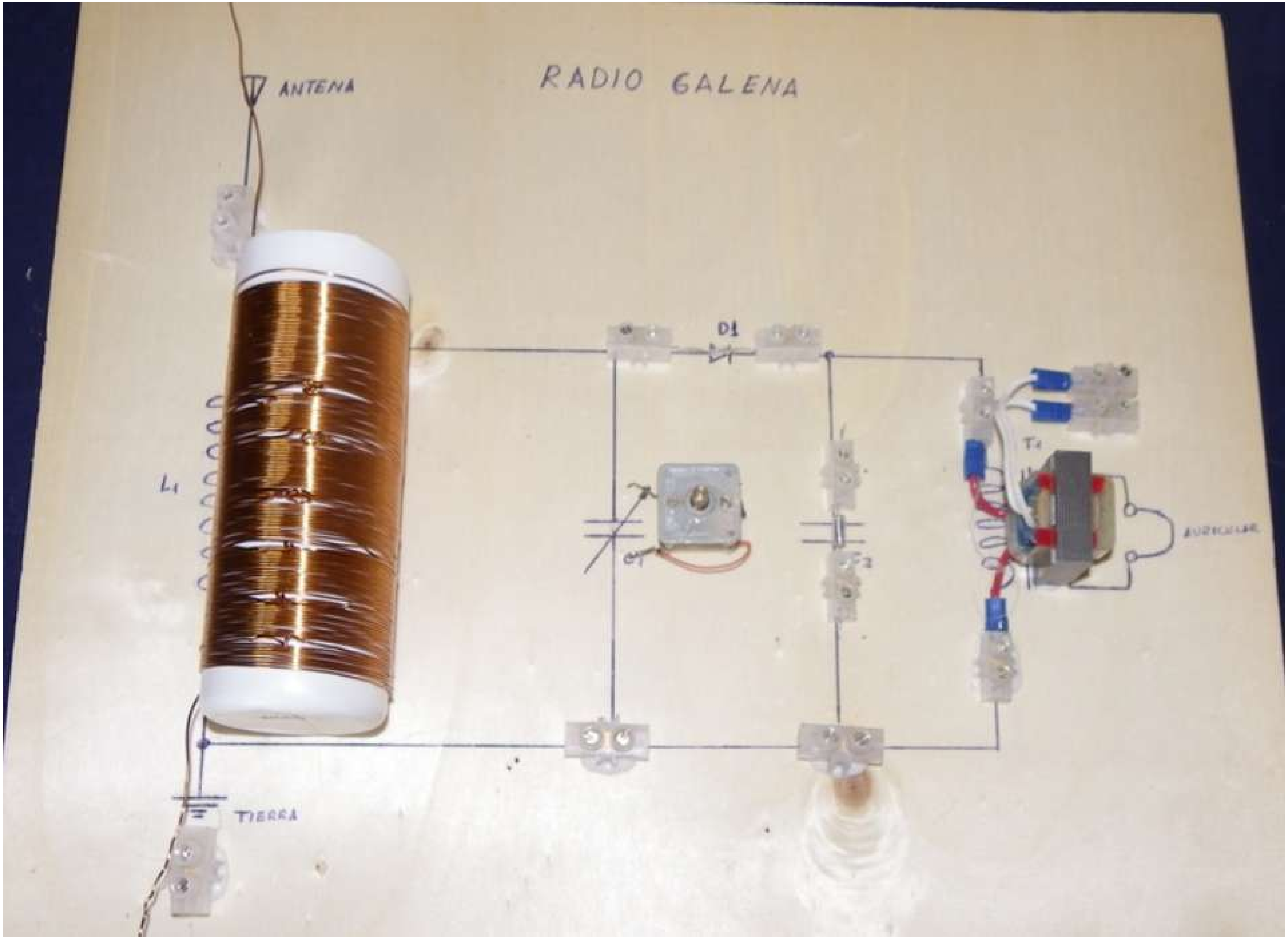
RADIO GALENA
ANTENA
L1
D1
C1
C2
T1
AURICULAR
TIERRA

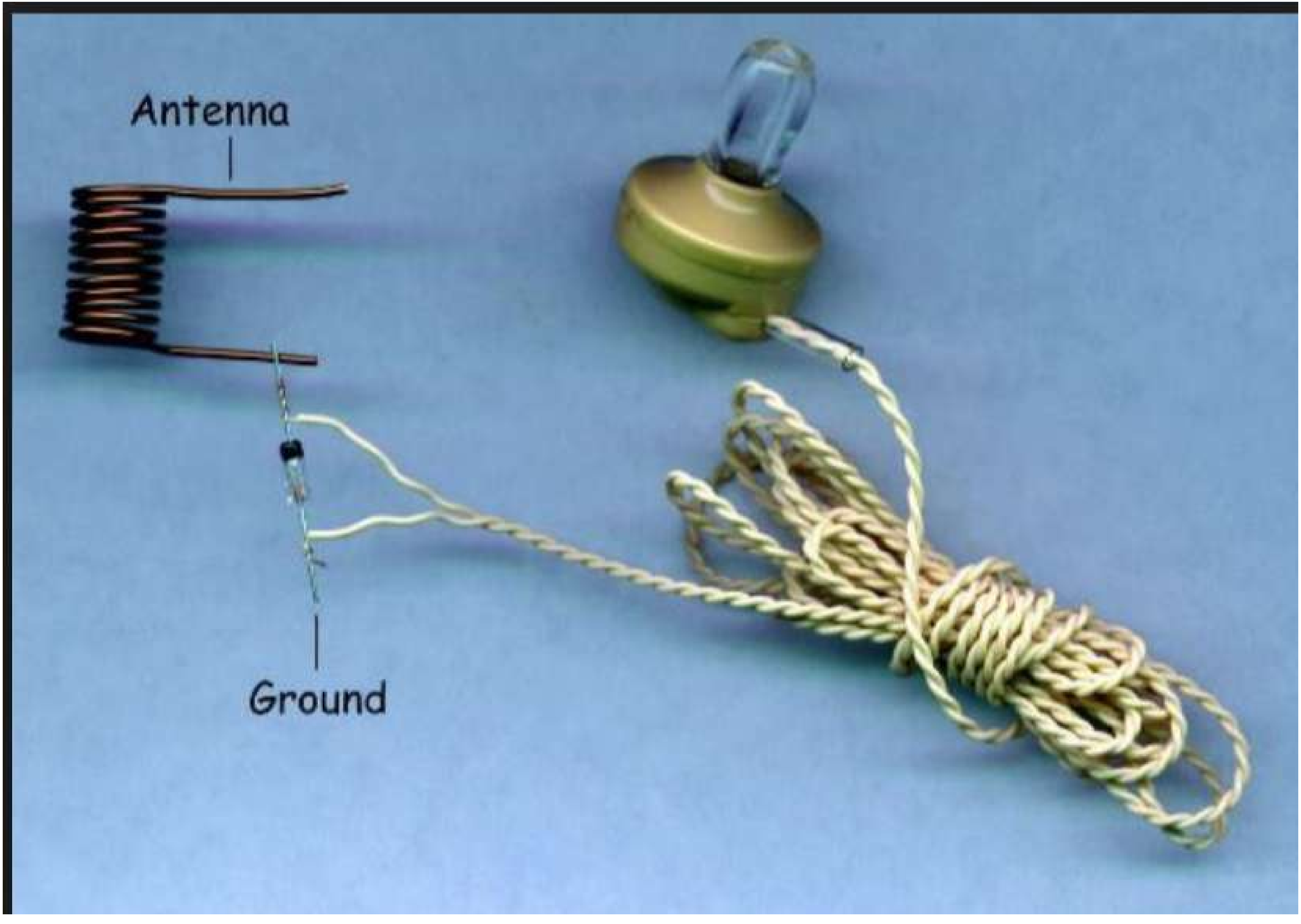
Antenna
Ground

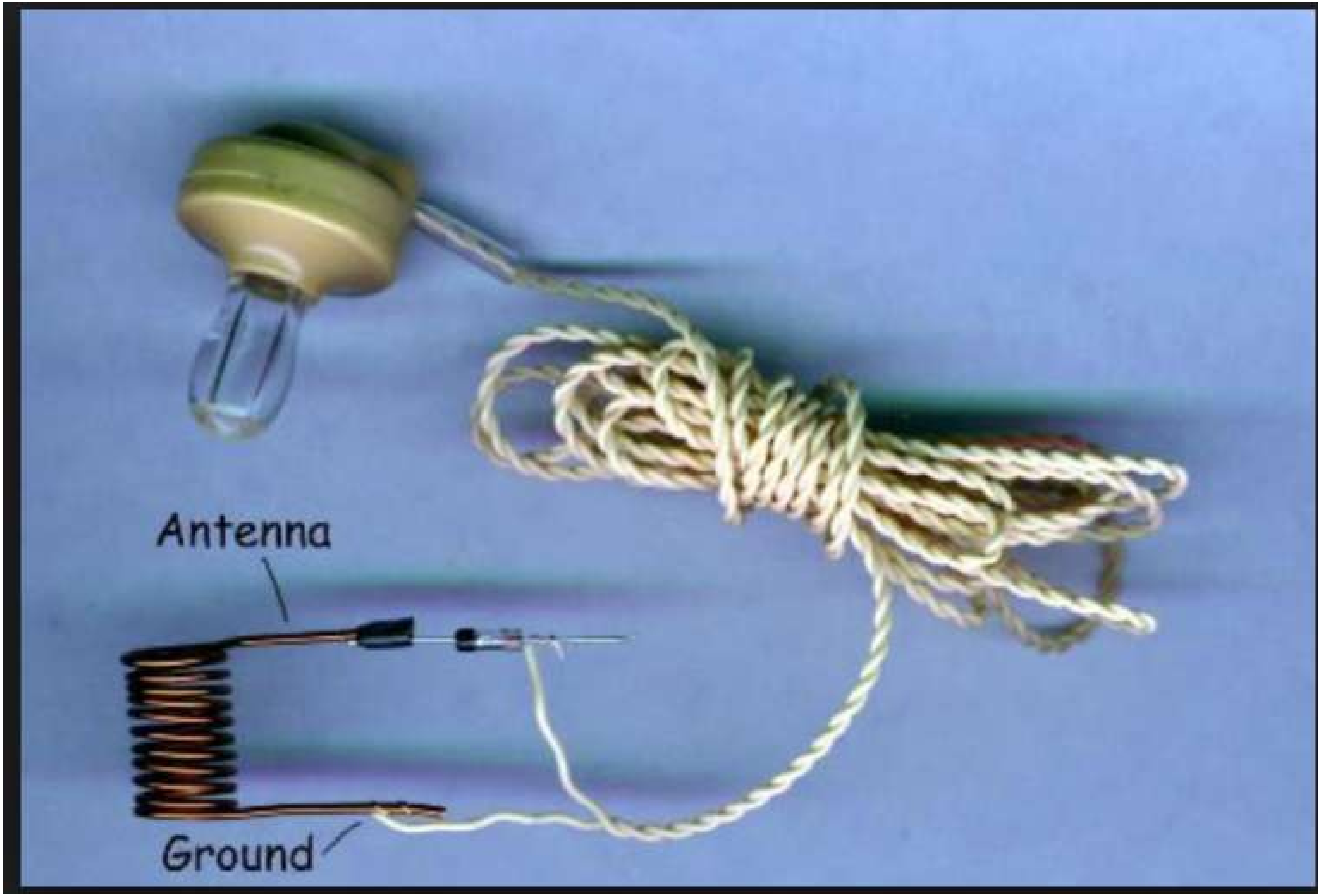
Antenna
Ground

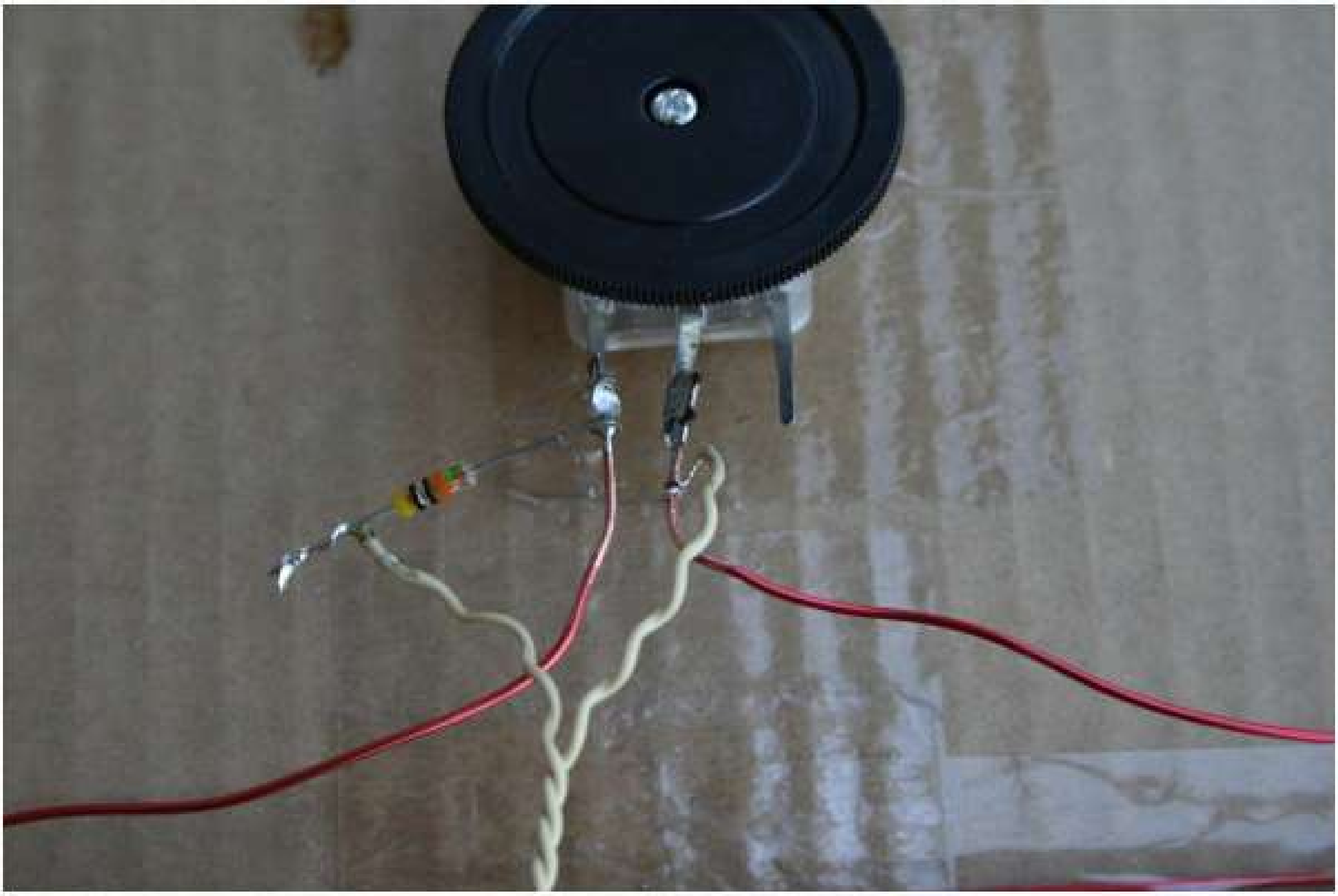

Click on photo for a larger picture

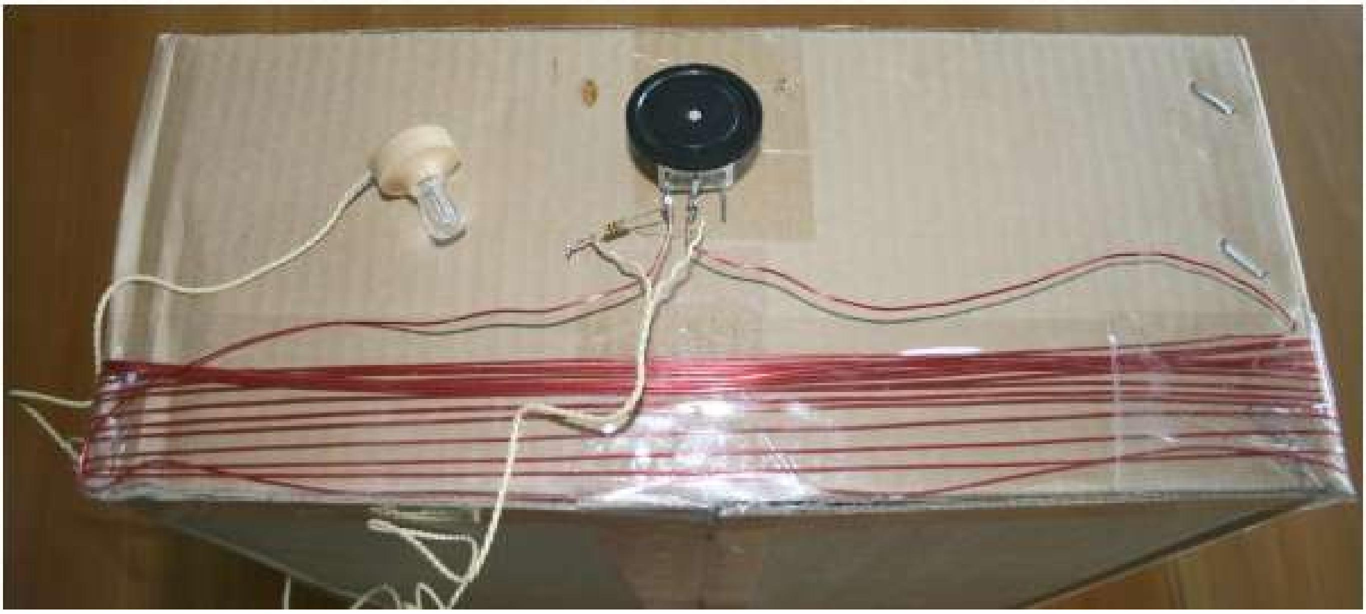

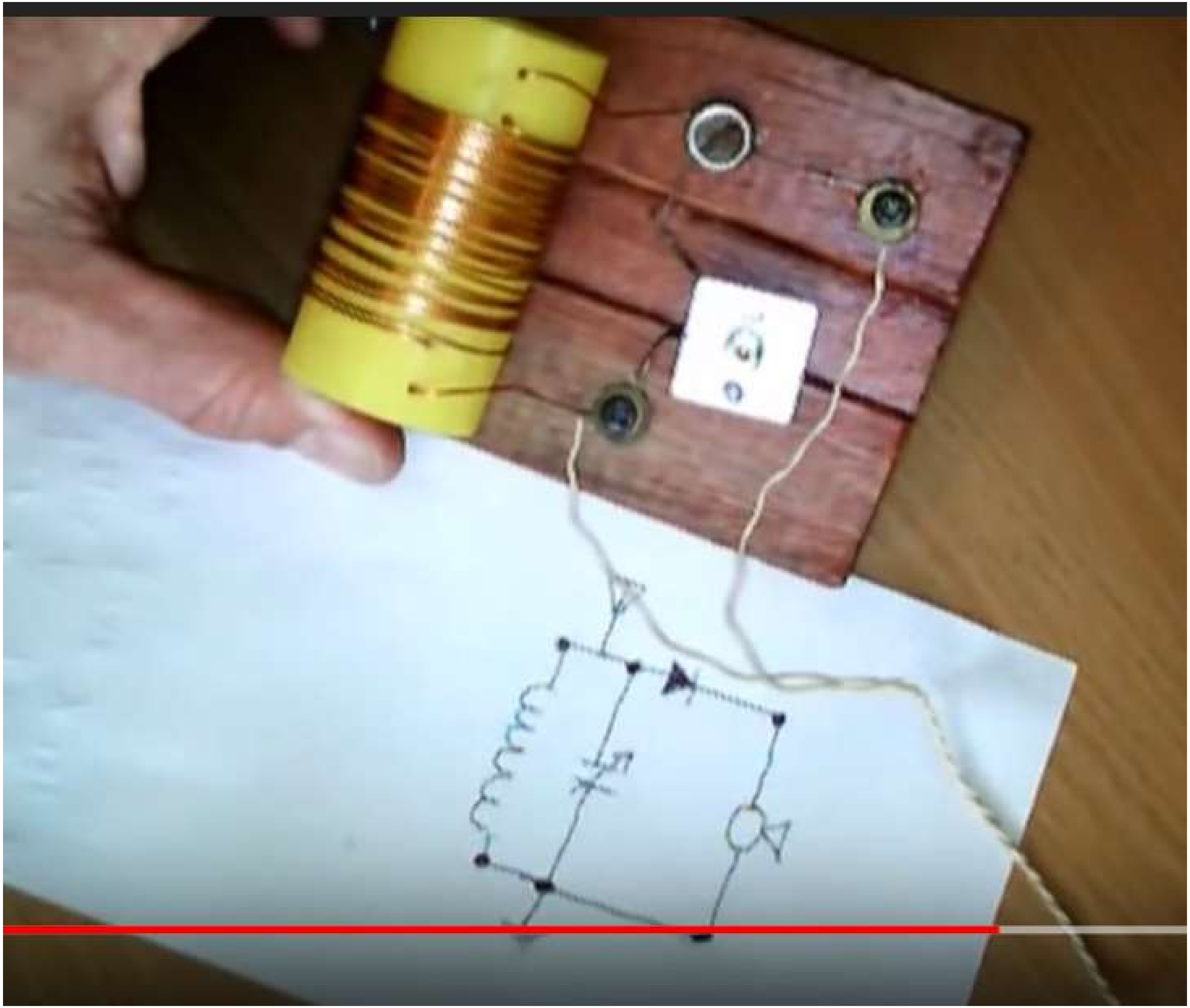

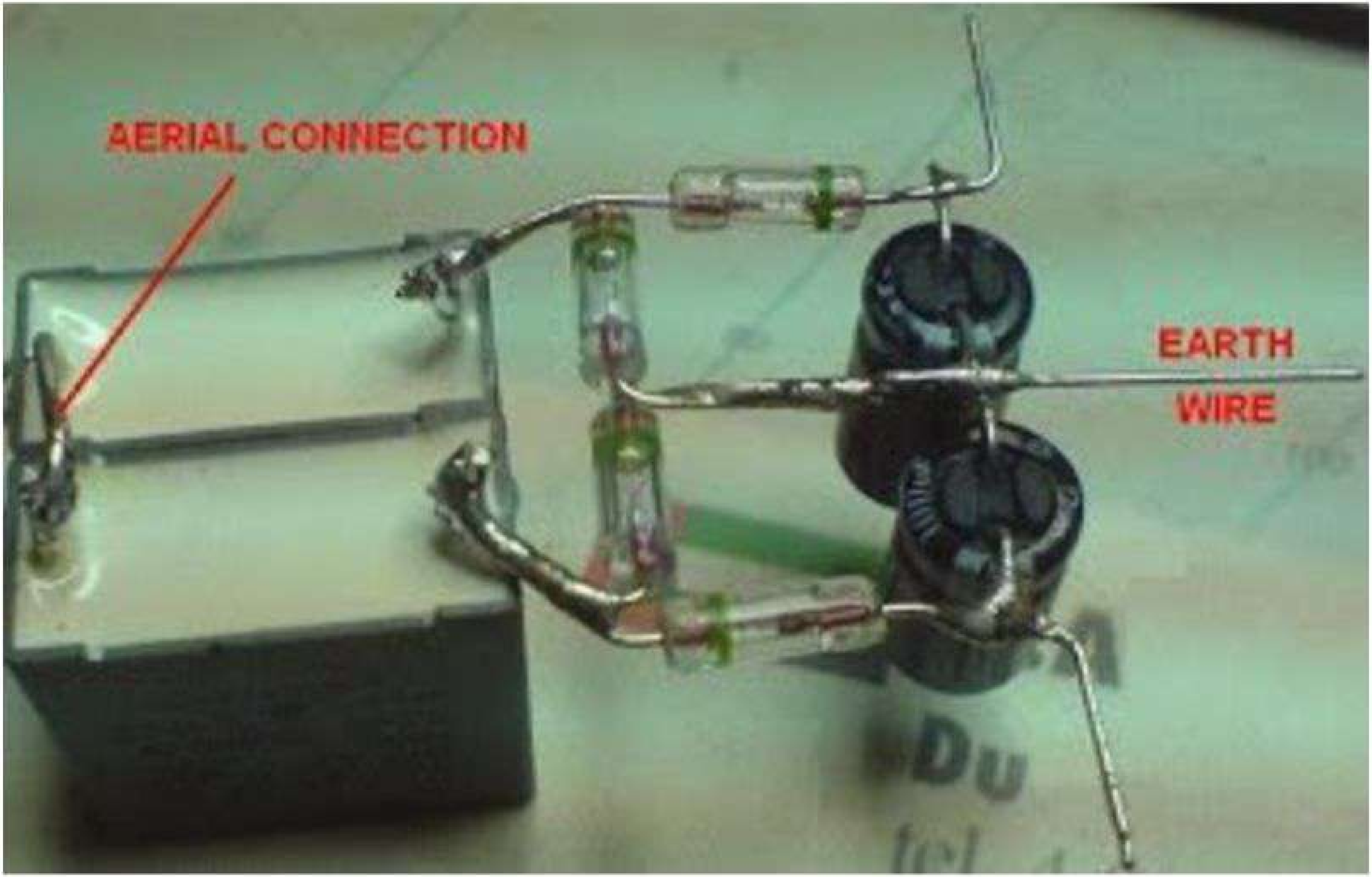
AERIAL CONNECTION
EARTH WIRE

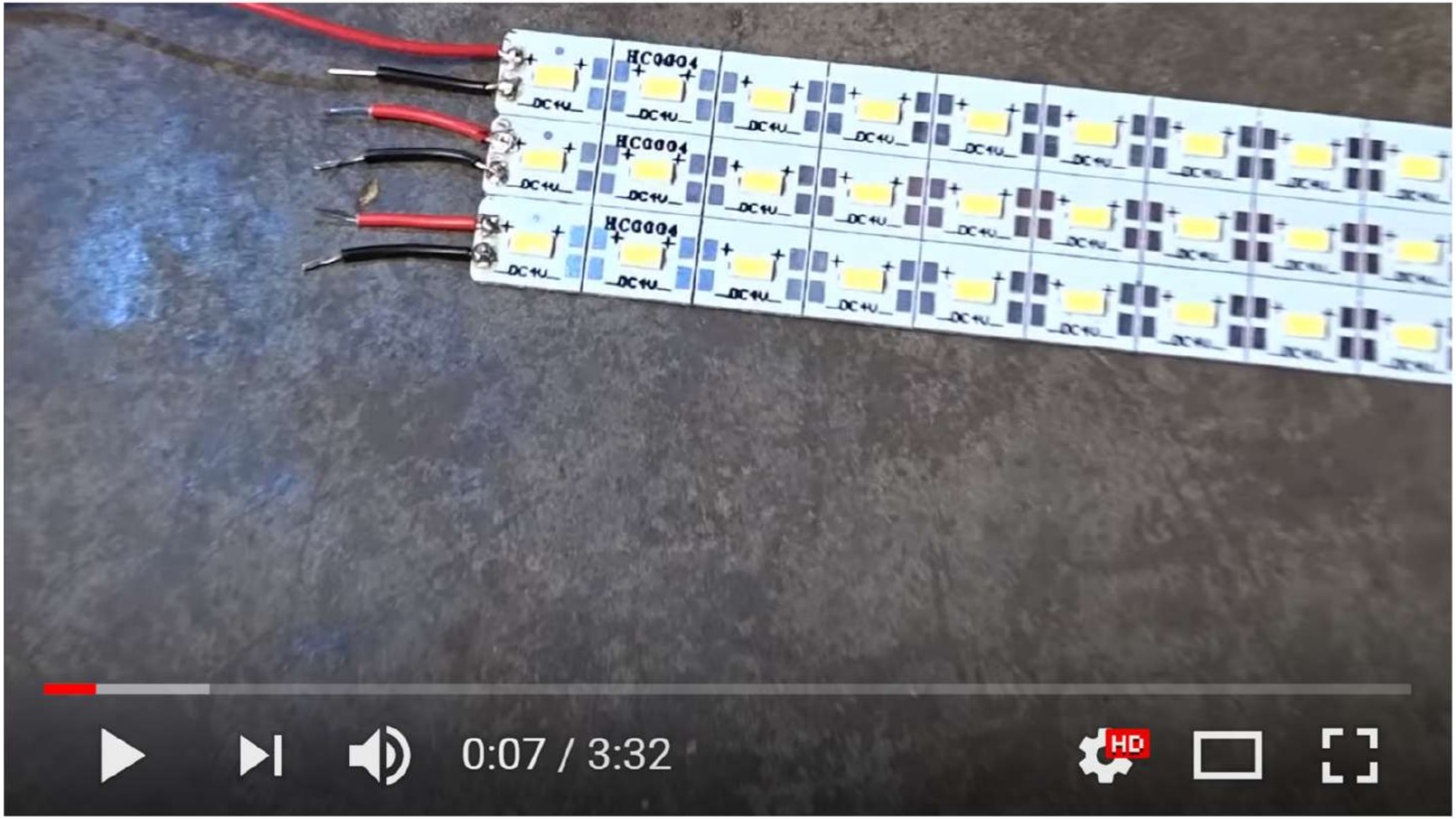
HC0004
DC4V
0:07 / 3:32
HD

New idea - You can make Solar Cell from LED (Free Energy)

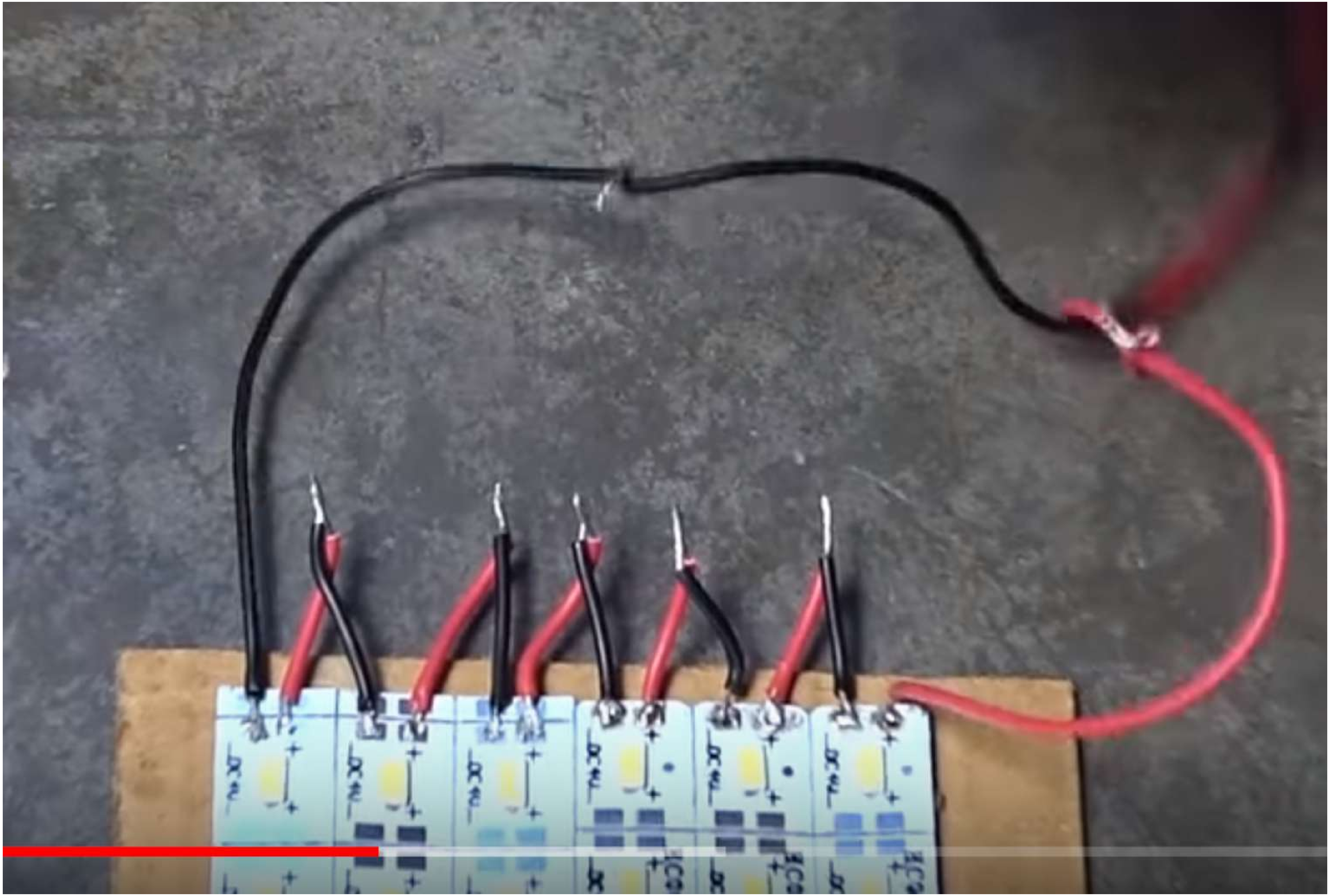

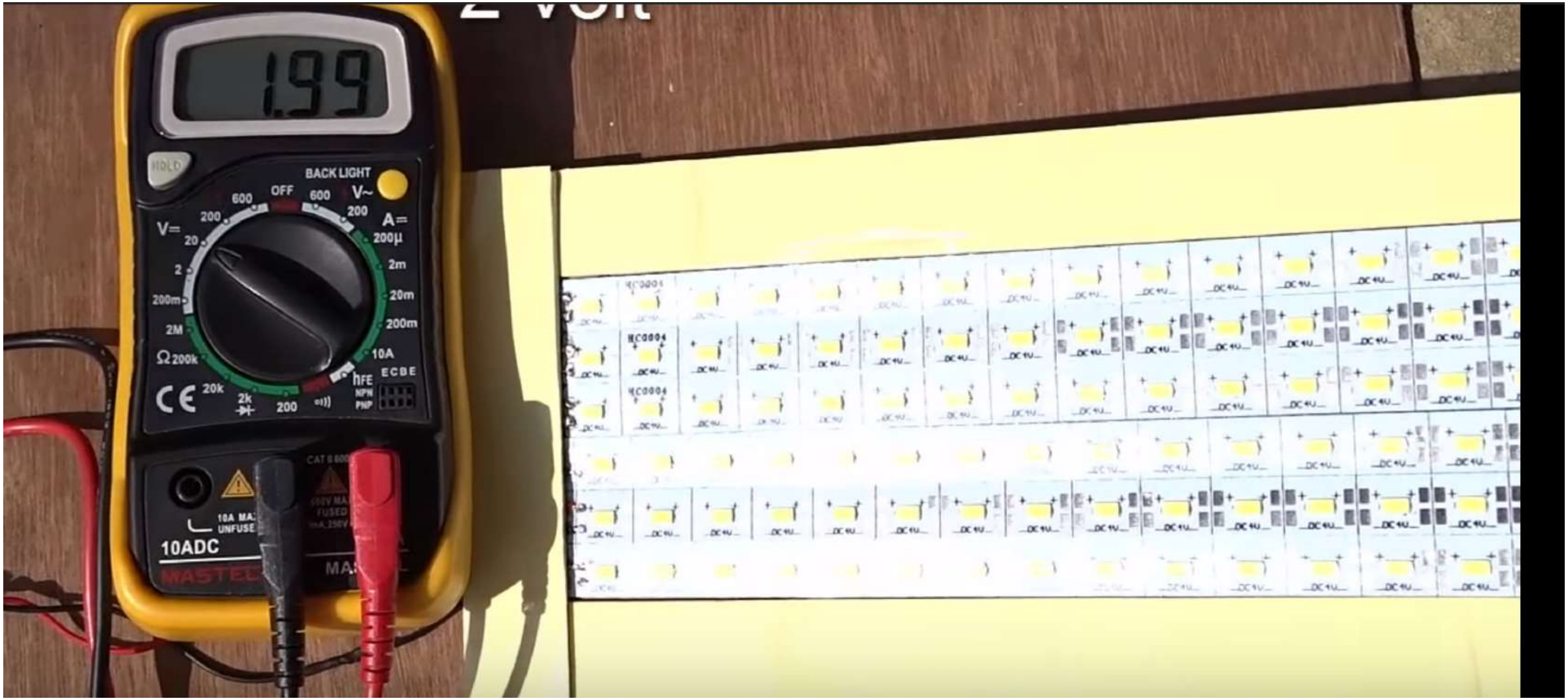
2 volt
1.99
HOLD
BACK LIGHT
OFF
V=
V~
A=
10A
hFE
ECBE
NPN
PNP
10ADC
MASTECH
DC 4V

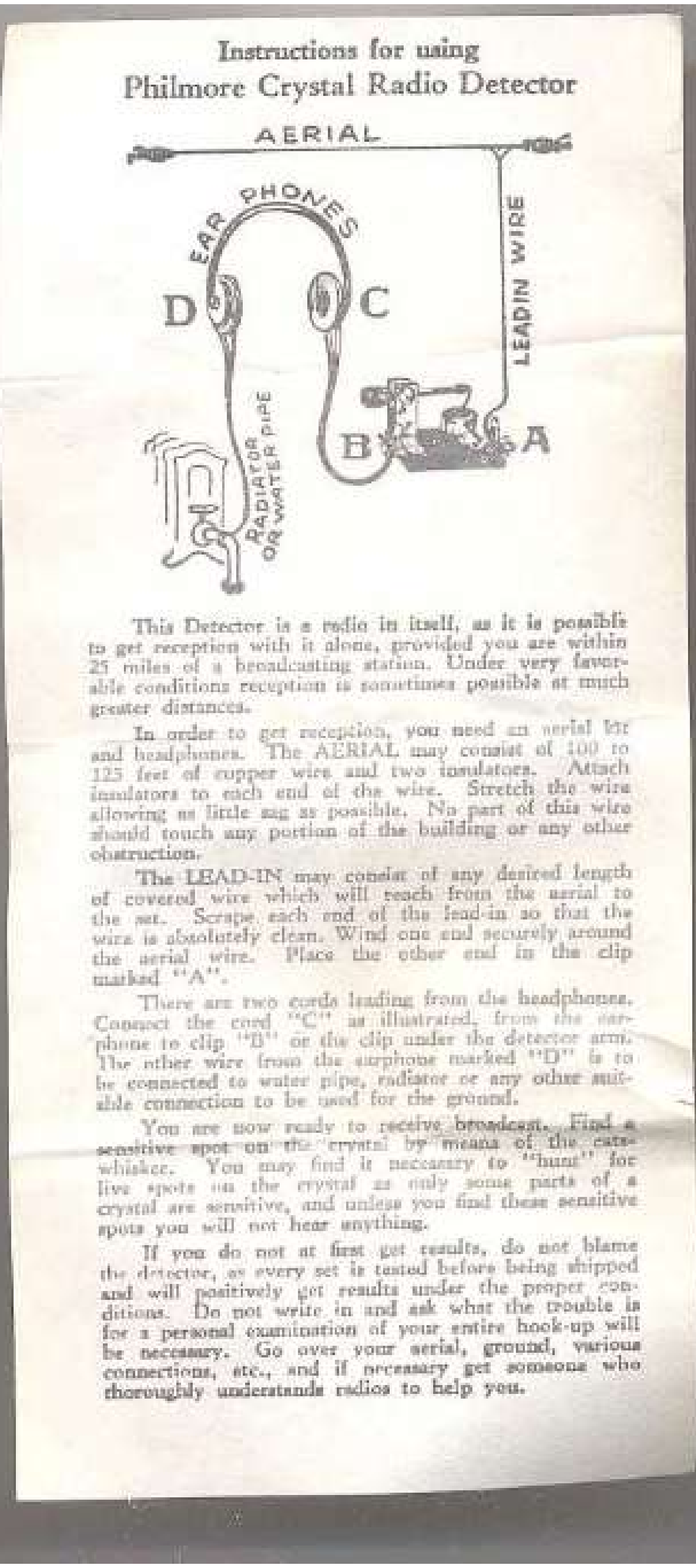

Instructions for using Philmore Crystal Radio Detector

AERIAL
EAR PHONES
LEADIN WIRE
D
C
B
A
RADIATOR OR WATER PIPE

This Detector is a radio in itself, as it is possible to get reception with it alone, provided you are within 25 miles of a broadcasting station. Under very favorable conditions reception is sometimes possible at much greater distances.

In order to get reception, you need an aerial kit and headphones. The AERIAL may consist of 100 to 125 feet of copper wire and two insulators. Attach insulators to each end of the wire. Stretch the wire allowing as little sag as possible. No part of this wire should touch any portion of the building or any other obstruction.

The LEAD-IN may consist of any desired length of covered wire which will reach from the aerial to the set. Scrape each end of the lead-in so that the wire is absolutely clean. Wind one end securely around the aerial wire. Place the other end in the clip marked "A".

There are two cords leading from the headphones. Connect the cord "C" as illustrated, from the earphone to clip "B" or the clip under the detector arm. The other wire from the earphone marked "D" is to be connected to water pipe, radiator or any other suitable connection to be used for the ground.

You are now ready to receive broadcast. Find a sensitive spot on the crystal by means of the cats-whisker. You may find it necessary to "hunt" for live spots on the crystal as only some parts of a crystal are sensitive, and unless you find these sensitive spots you will not hear anything.

If you do not at first get results, do not blame the detector, as every set is tested before being shipped and will positively get results under the proper conditions. Do not write in and ask what the trouble is for a personal examination of your entire hook-up will be necessary. Go over your aerial, ground, various connections, etc., and if necessary get someone who thoroughly understands radios to help you.

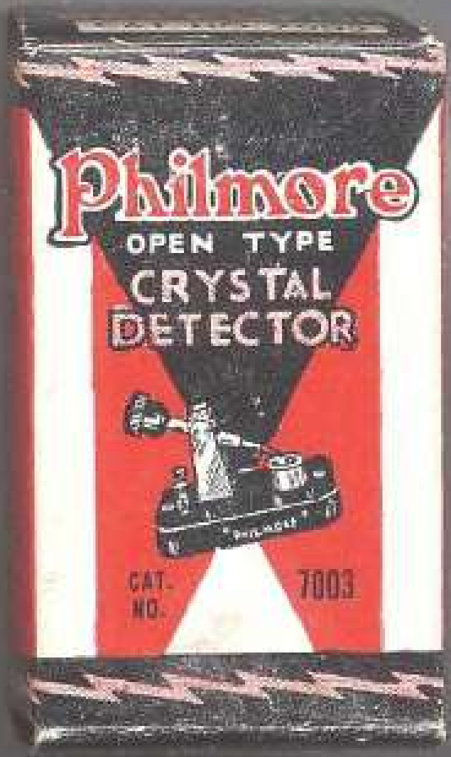

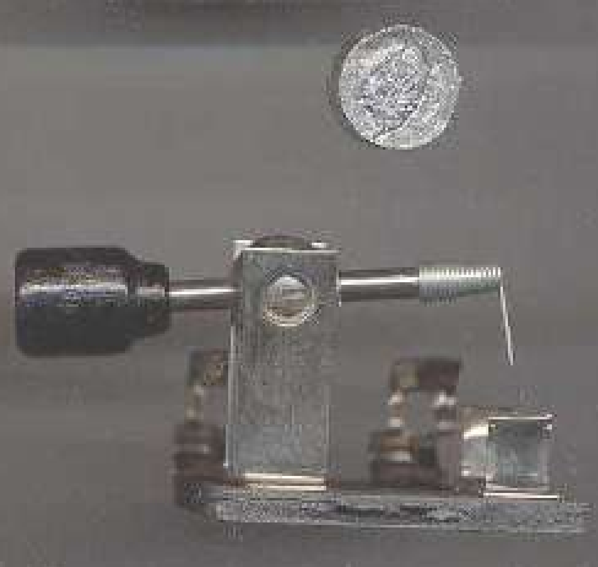

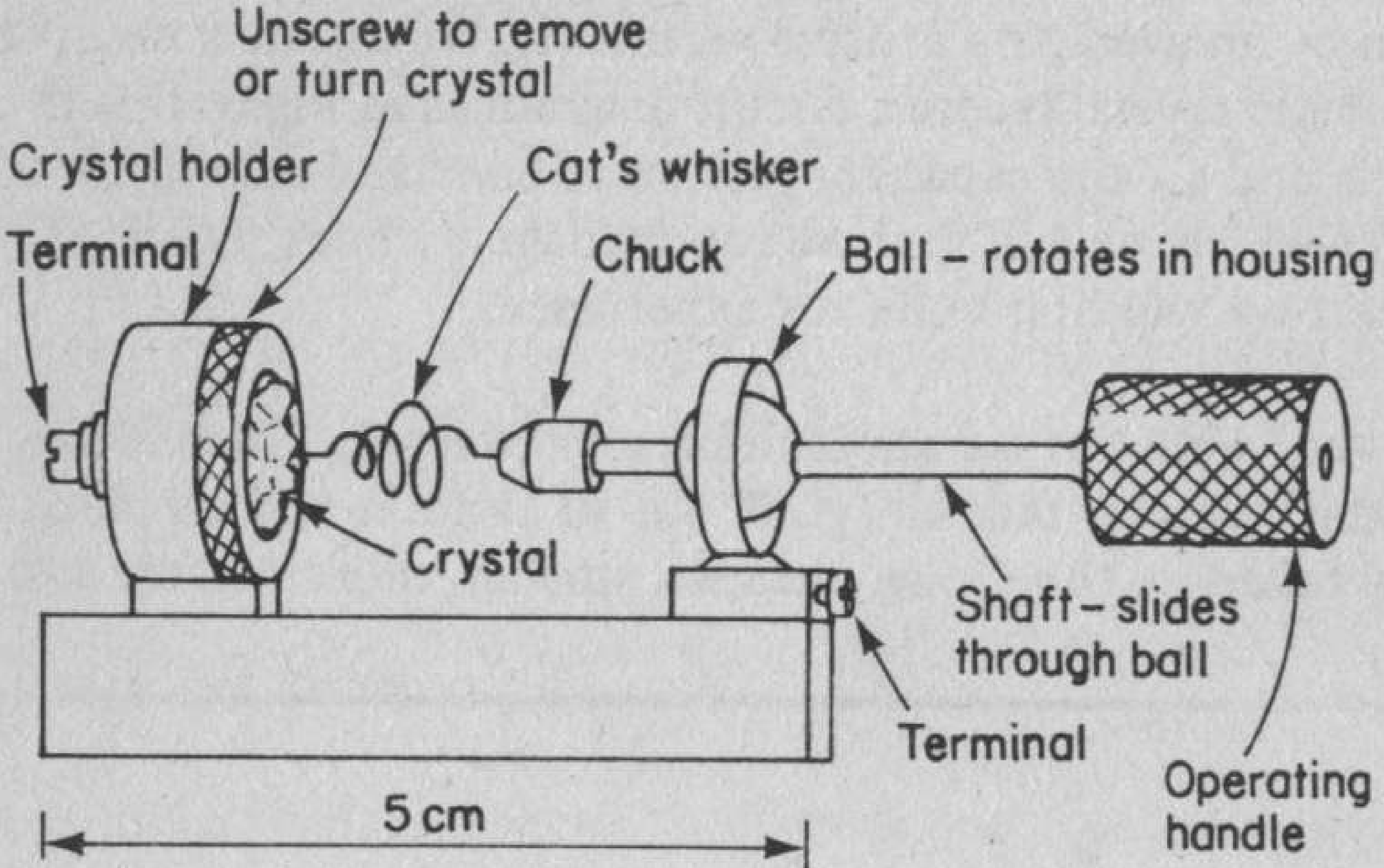
Unscrew to remove
or turn crystal
Crystal holder
Cat's whisker
Terminal
Chuck
Ball – rotates in housing
Crystal
Shaft – slides
through ball
Terminal
Operating
handle
5 cm

Modern diode (7·5 X 2·5 mm)

Majoritatea componentelor active folosite în circuitele electronice moderne sînt dispozitivele bazate pe semiconductoare.

Cel mai simplu dispozitiv este *dioda* punctiformă cu germaniu. Ea are proprietatea de bază de a se comporta ca şi cum este conectată direct la o sursă electrică de curent continuu (plusul sursei la plusul diodei) şi ca un izolator, cînd este conectată invers la aceeaşi sursă (plusul sursei la minusul diodei), ca în figura 2.9.

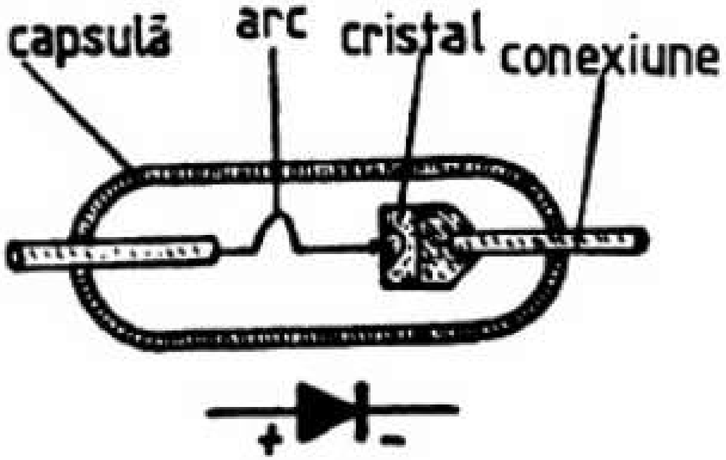
capsulă
arc
cristal
conexiune
+
-

FIG. 2.9.

Dioda punctiformă cu germaniu

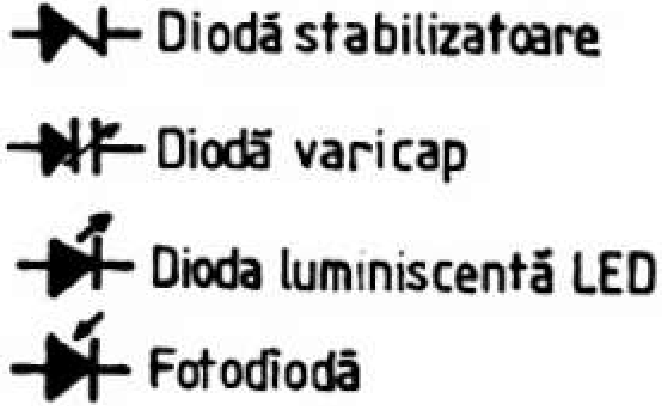
Diodă stabilizatoare
Diodă varicap
Dioda luminiscentă LED
Fotodiodă

FIG. 2.10.

Tipuri de diode semiconductoare

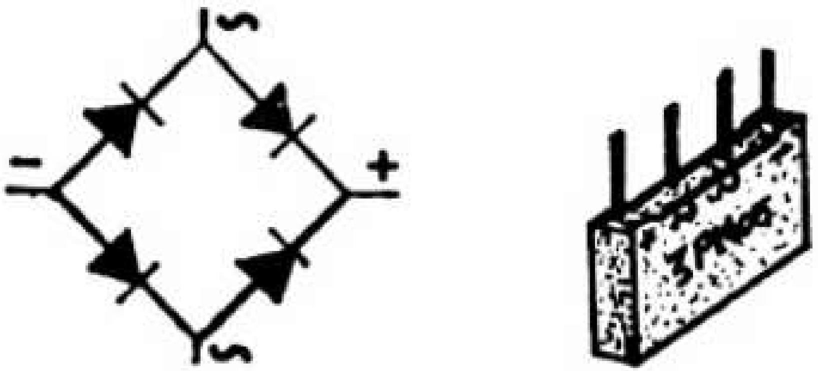

FIG. 2.11.

Puntea redresoare

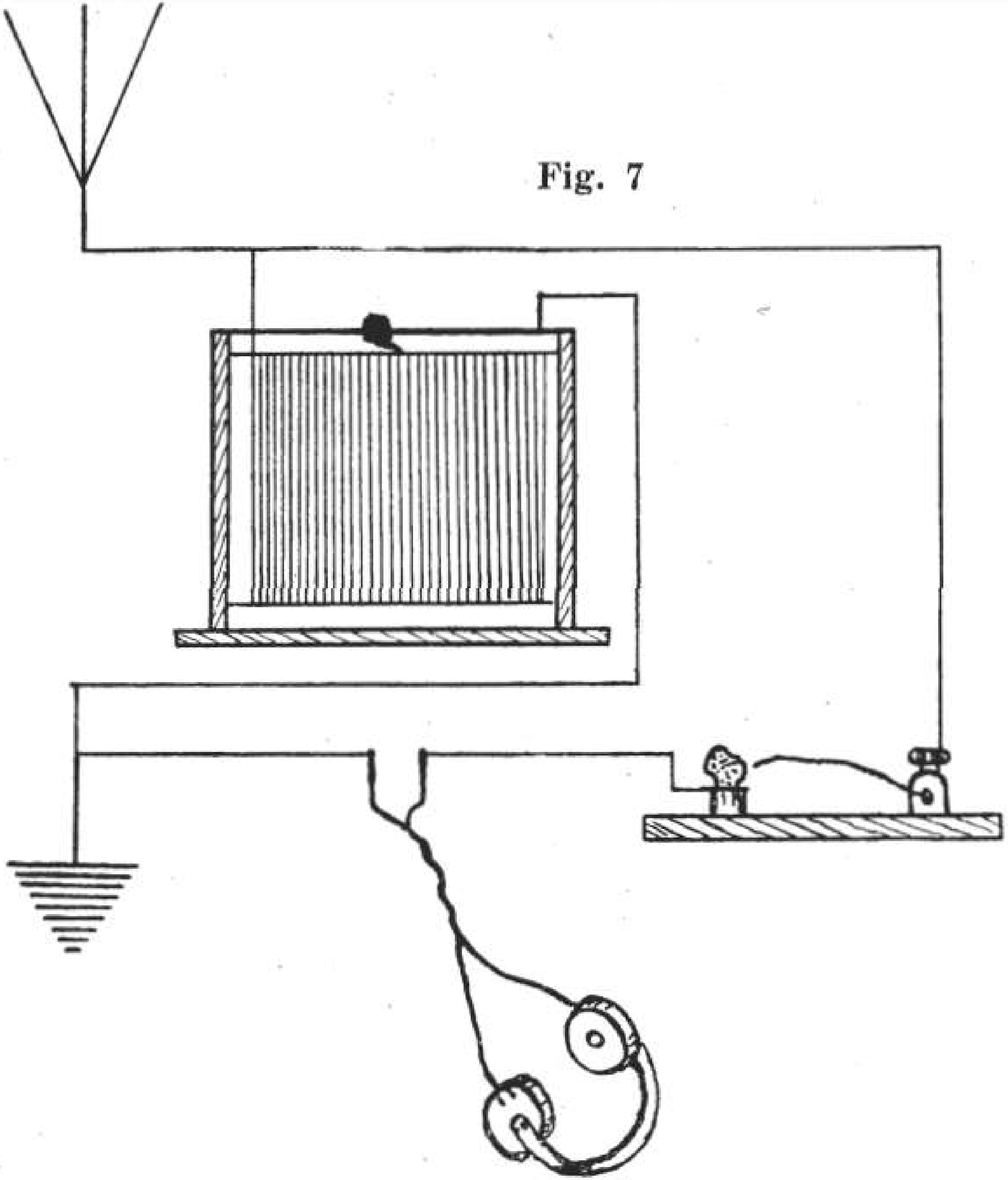

Fig. 7

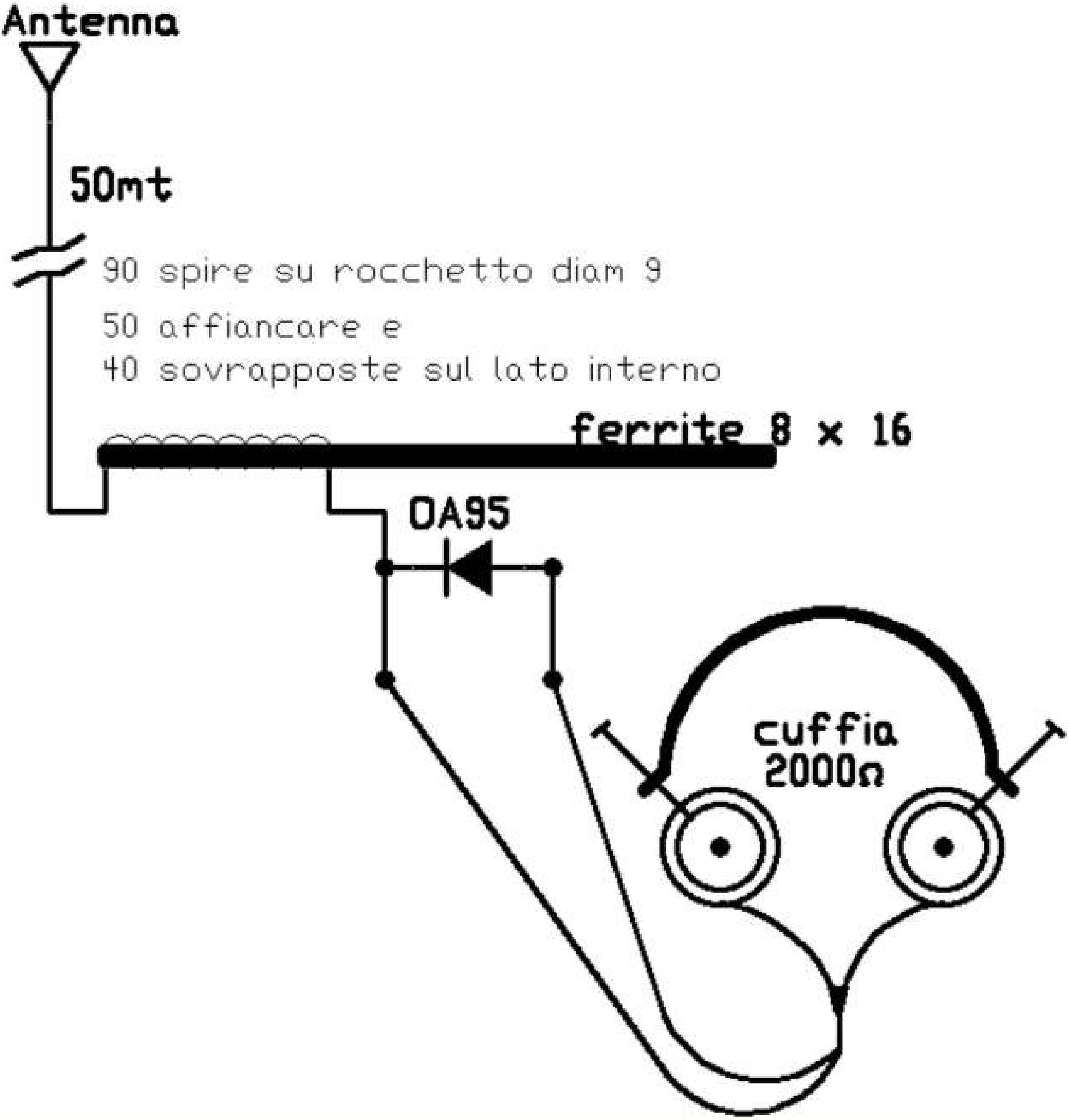
Antenna
50mt
90 spire su rocchetto diam 9
50 affiancare e
40 sovrapposte sul lato interno
ferrite 8 x 16
OA95
cuffia
2000Ω

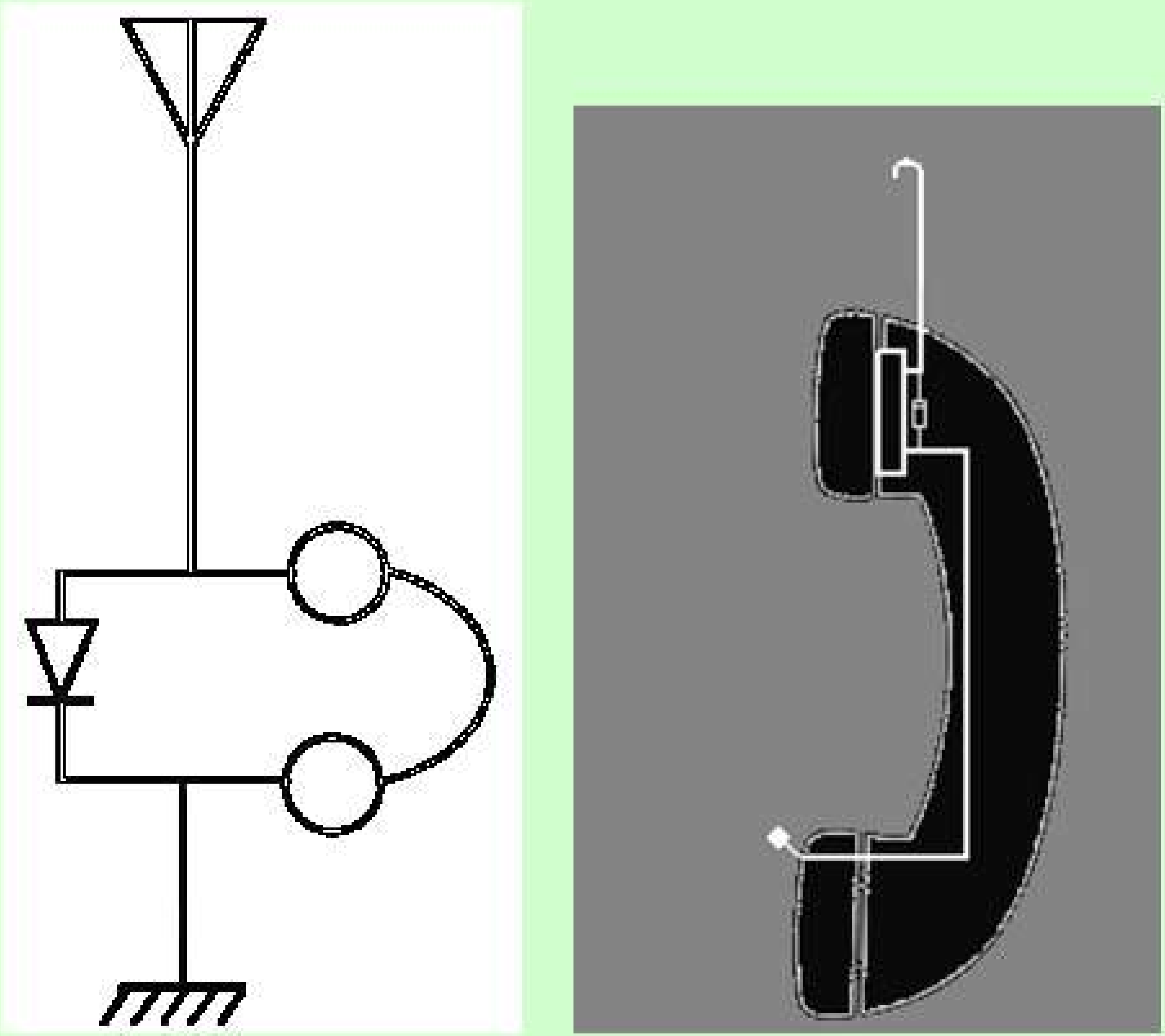

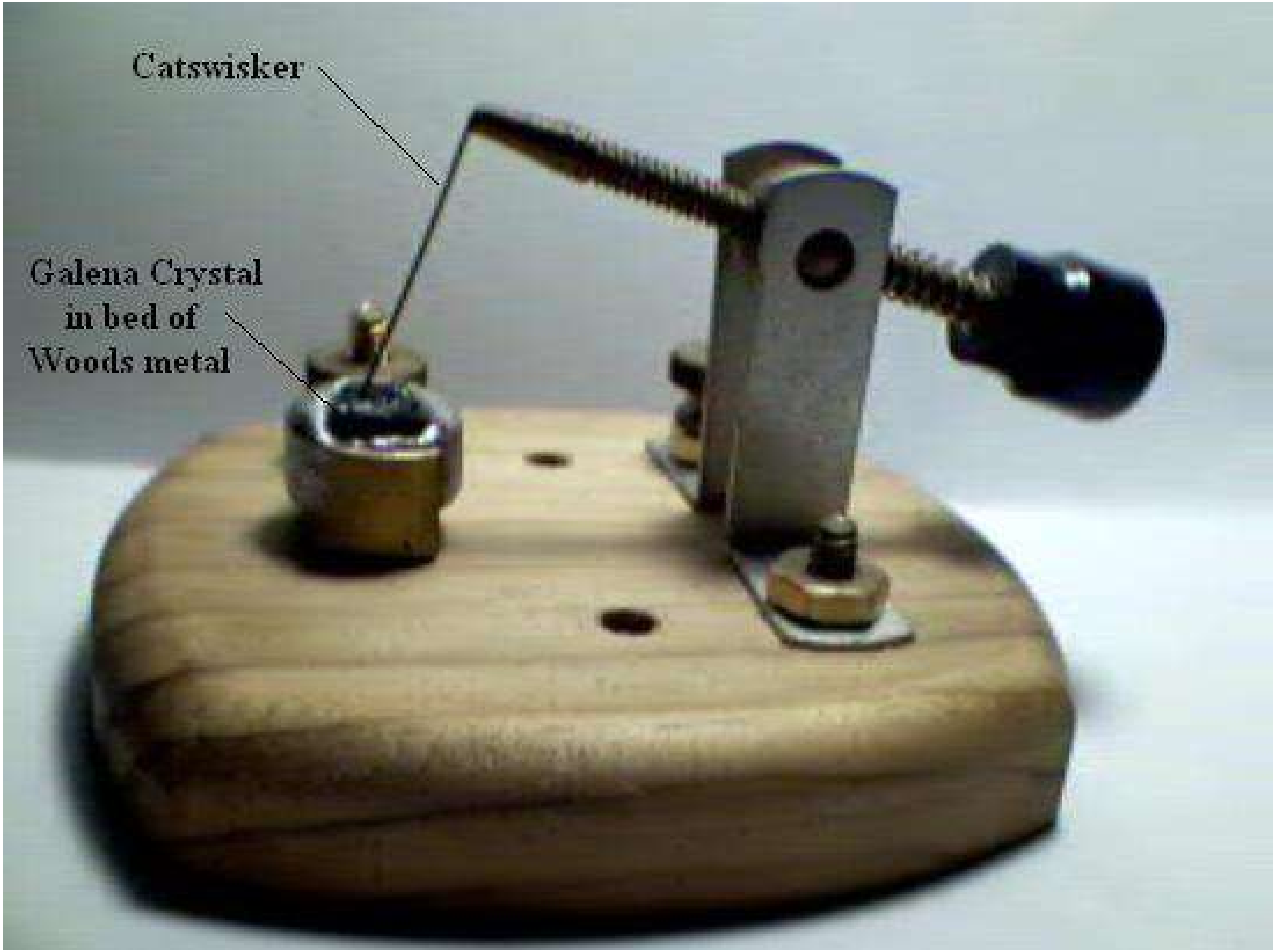
Catswisker
Galena Crystal
in bed of
Woods metal

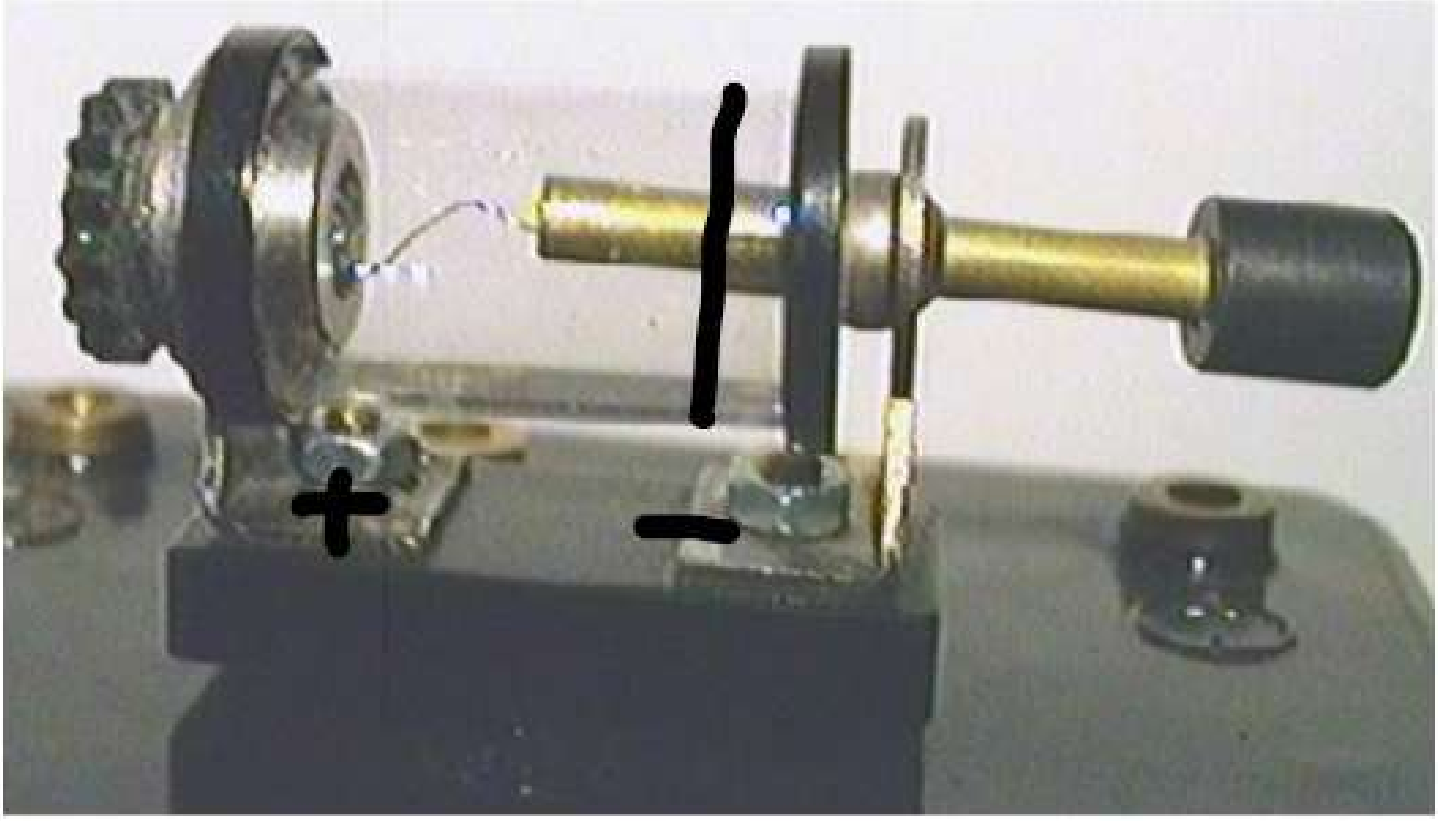
+
−

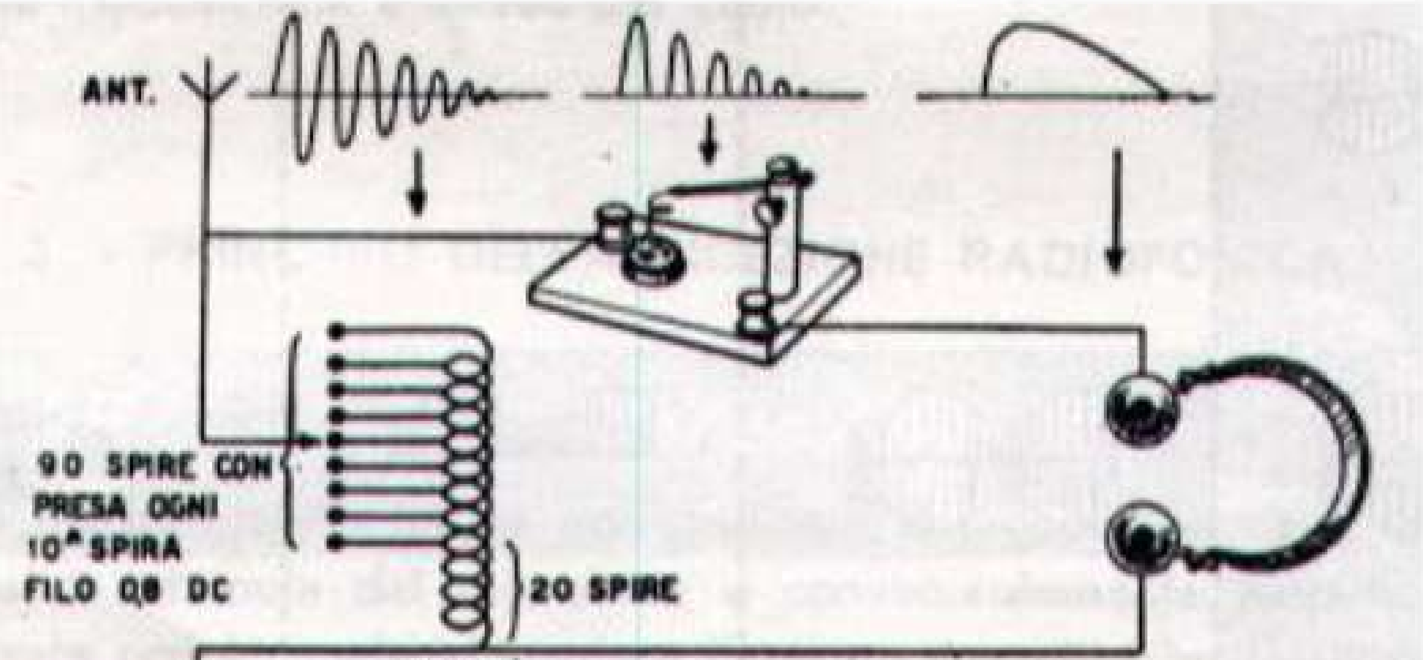
ANT.
90 SPIRE CON
PRESA OGNI
10ª SPIRA
FILO 0,8 DC
20 SPIRE
6 cm. ø
TERRA

Lead sulfide crystals, also called "galena"

Crystal detector, the lever was used to position the spring contact on a crystal face capable of performing the radio signal detection.

Yeah , but how does a galena radio work? Here is the constructive scheme of a receiver of this kind, obtained from a book of popular radio commu
to the knowledge of the radio, ed.Hepep 1943):

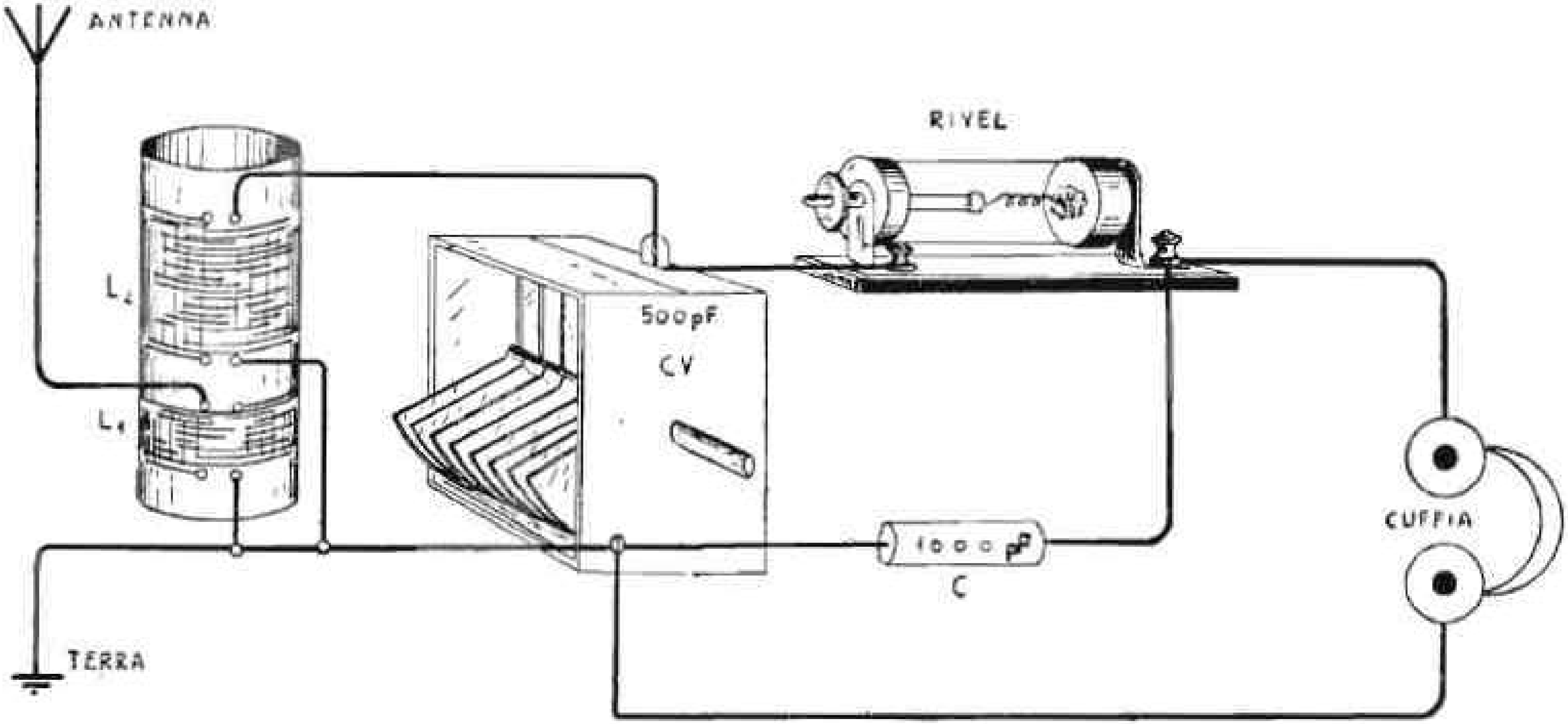

Fig. 13.5. – Realizzazione pratica dello schema di fig. 13.3 con aggiunta della bobina d'antenna.

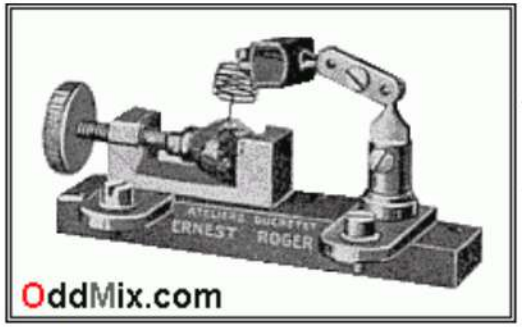

Picture 1. "Ernest Roger" Crystal detector from 1920

Many of the **younger generation** never had the pleasure to have, or to even see a real beautiful crystal detector as shown on **Picture 1**. A detector like that was the product of the early 1920's, and it was an expensive, quality, well made product. They were usually hands assembled with machine made components.

This particular "**Ernest Roger**" detector device is built on an insulated base, either wood or most likely of Bakelite, the most widely used and only available plastic material at the times.

Many of the younger generation never had the pleasure to have, or to even see a real beautiful crystal detector as shown on **Picture 1**. The new generations are much too involved with technology and playing on their computers or going to an **online University**. This excellent detector holder used copper, brass and bronze generously for the crystal holder and for all of the **electrodes**.

The most often used crystal in this and similar detectors - then and now - was **pyrite** or galena as shown in the 1925 listing in a table in the **"American Mineralogist"** publication, that has a listing of thirty three minerals catalogued with rectifying detector properties.

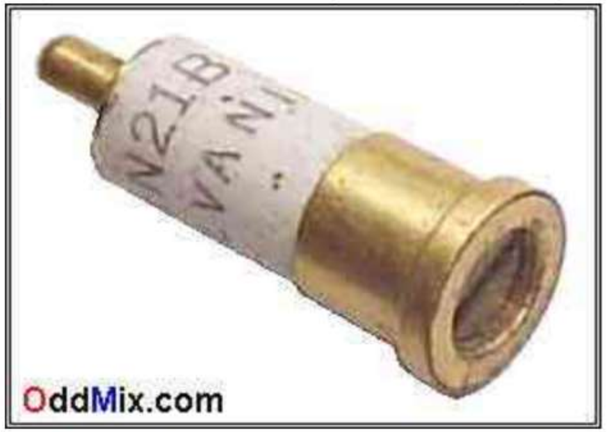

Picture 2. Early point contact diode with adjustable srew

Among the very first commercially produced diodes is the 1N21B shown on **Picture 2**. It is more than curious, that this "diode" is enclosed in a case on which there is a slotted screw-head is just visible on the wide side (lower right). That screw is connected to a fine "cats whisker" steel wire, a few turns of a spring-like device, terminating in a point that is in touch of the germanium semiconductor material connected to the top left diode terminal.

The arrangement is much smaller, and more diode like, then the detector on **Picture 1**., but it made in a very similar arrangement, which has became known as the point contact diode. From these and similar germanium diodes, evolutionary progress leads us to current PN junction Silicon, Gallium Arsenide or other more exotic semiconductor materials which are the achievement of the latest scientific age and many years of steady experimentation, research and development.

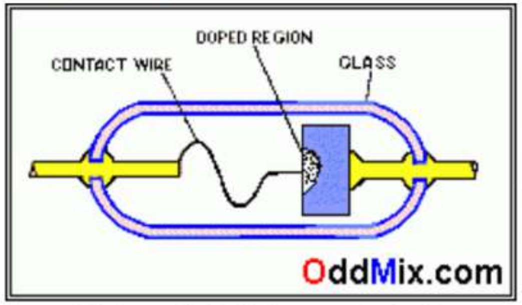

Picture 3. Point contact diode

Picture 3 shows the cross section of a hermetically glass enclosed, point contact diode. If the left side, point contact terminal would be attached to a screw and a short spring, and the glass envelope would have a threaded metal part in it, the **Picture 2** and **Picture 3** devices would be nearly identical.

These point contact "cat's whisker" devices, are still made occasionally because of their very small capacitance. As it was discovered early on, they are highly useful in high frequency electronics. All the way up to microwave frequencies they useable. The simplest radar detectors usually made with a simple horn **antenna**, a diode similar to **Picture 2** and **Picture 3** and a transistor amplifier. If the designer keeps it simple, and don't use a local oscillator, such radar detectors

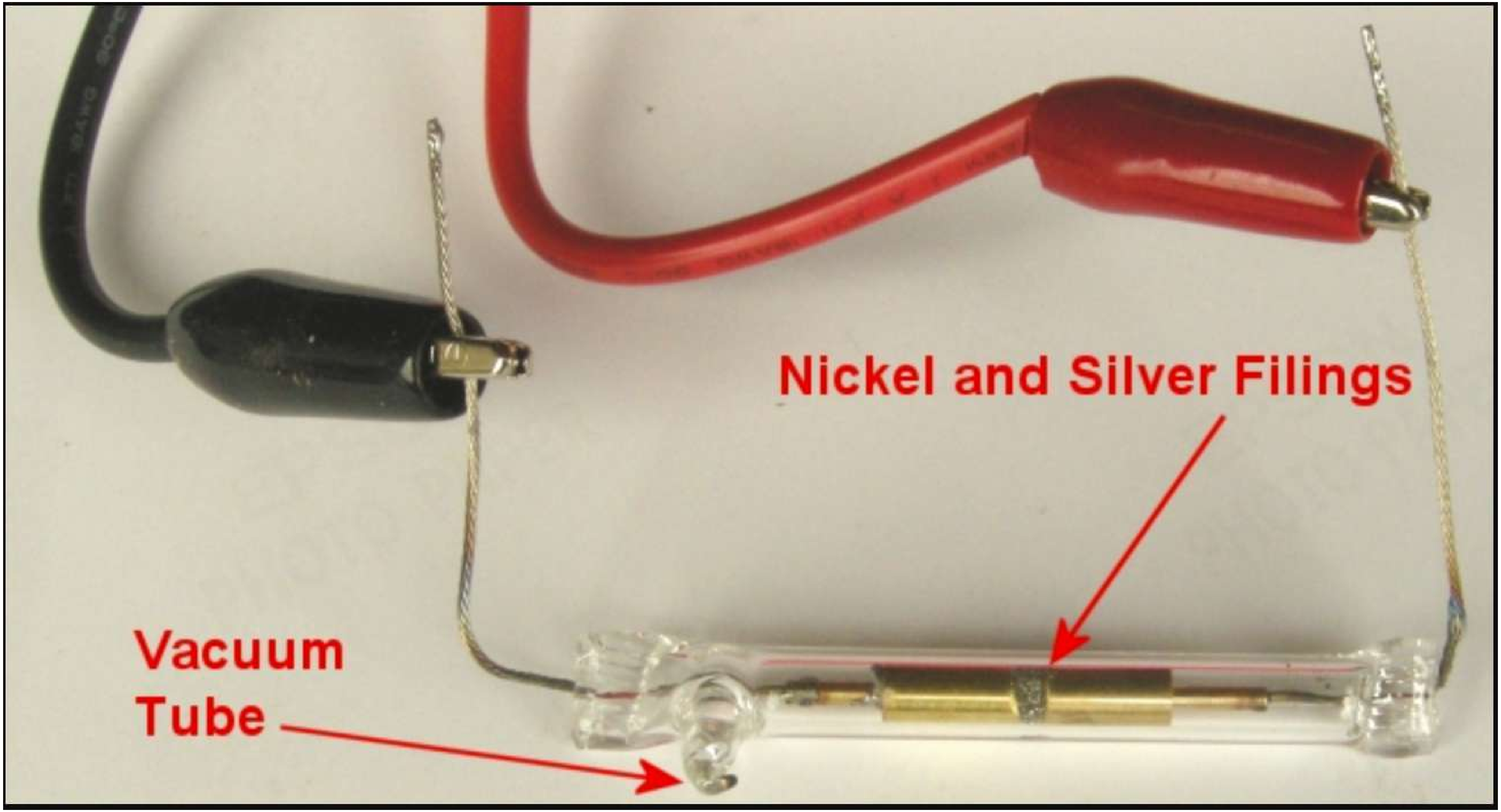
Nickel and Silver Filings
Vacuum Tube

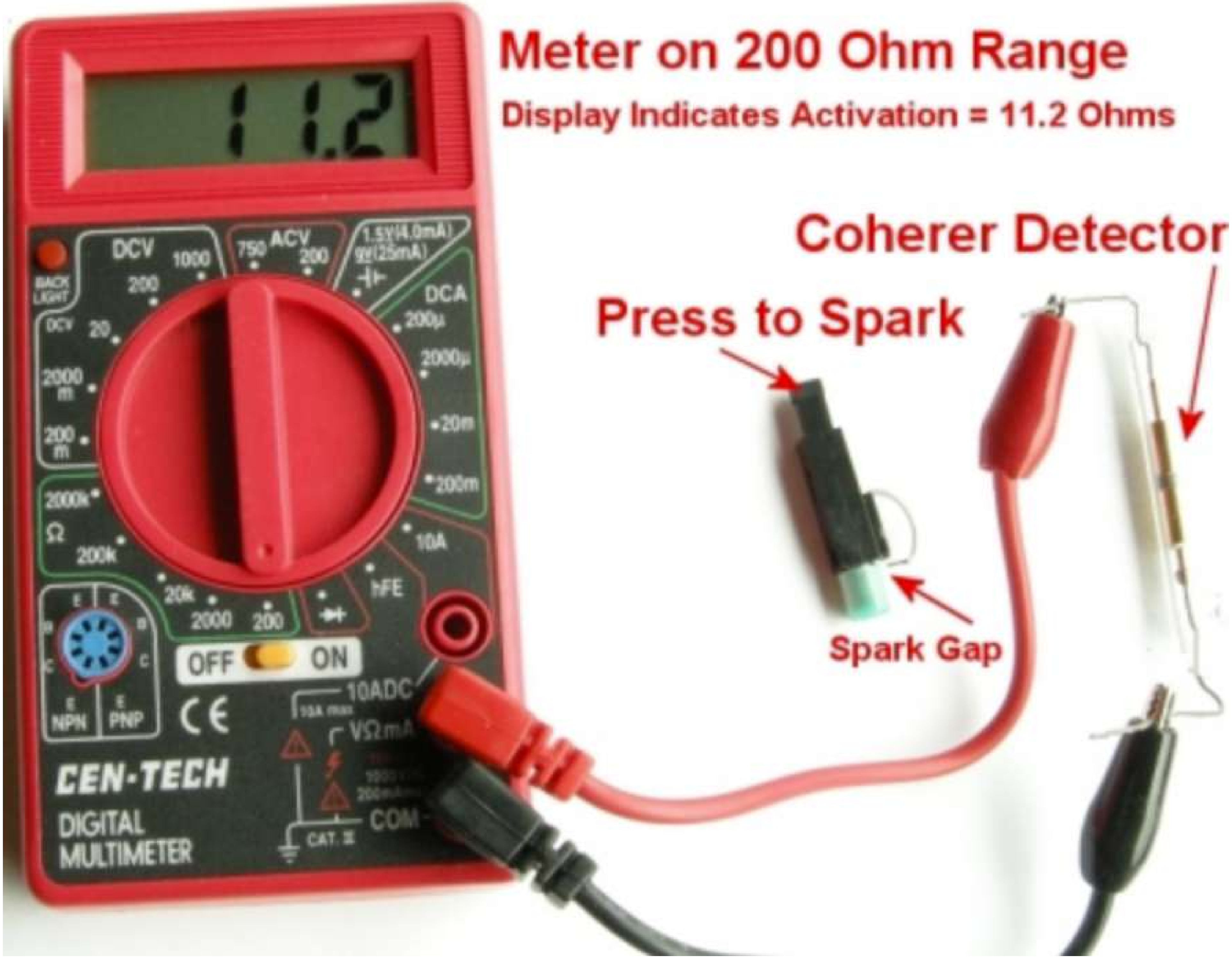

Meter on 200 Ohm Range

Display Indicates Activation = 11.2 Ohms

Coherer Detector

Press to Spark

Spark Gap

11.2
DCV
ACV
DCA
OFF
ON
10ADC
VΩmA
COM
CEN-TECH
DIGITAL
MULTIMETER

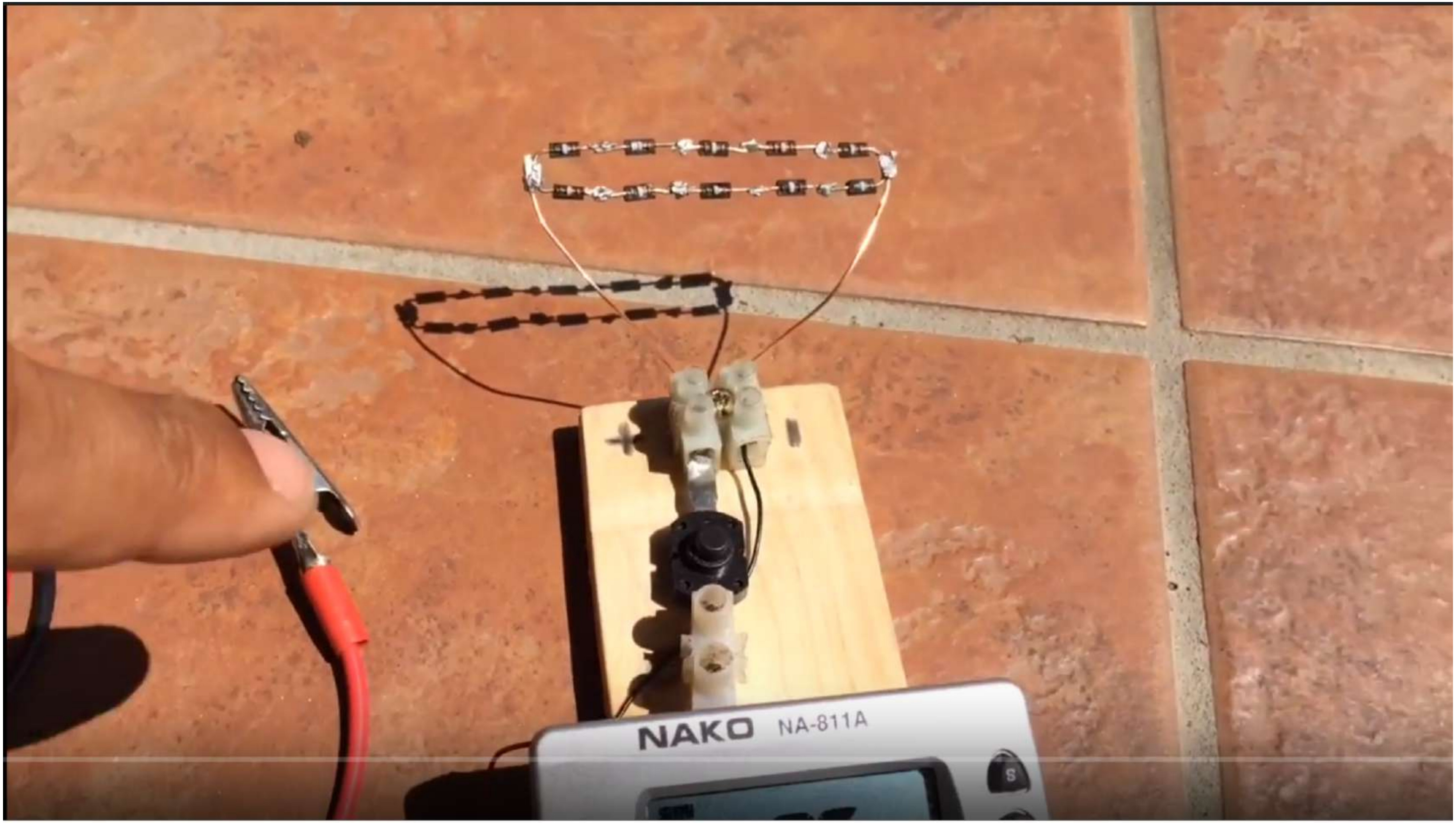
NAKO NA-811A
S

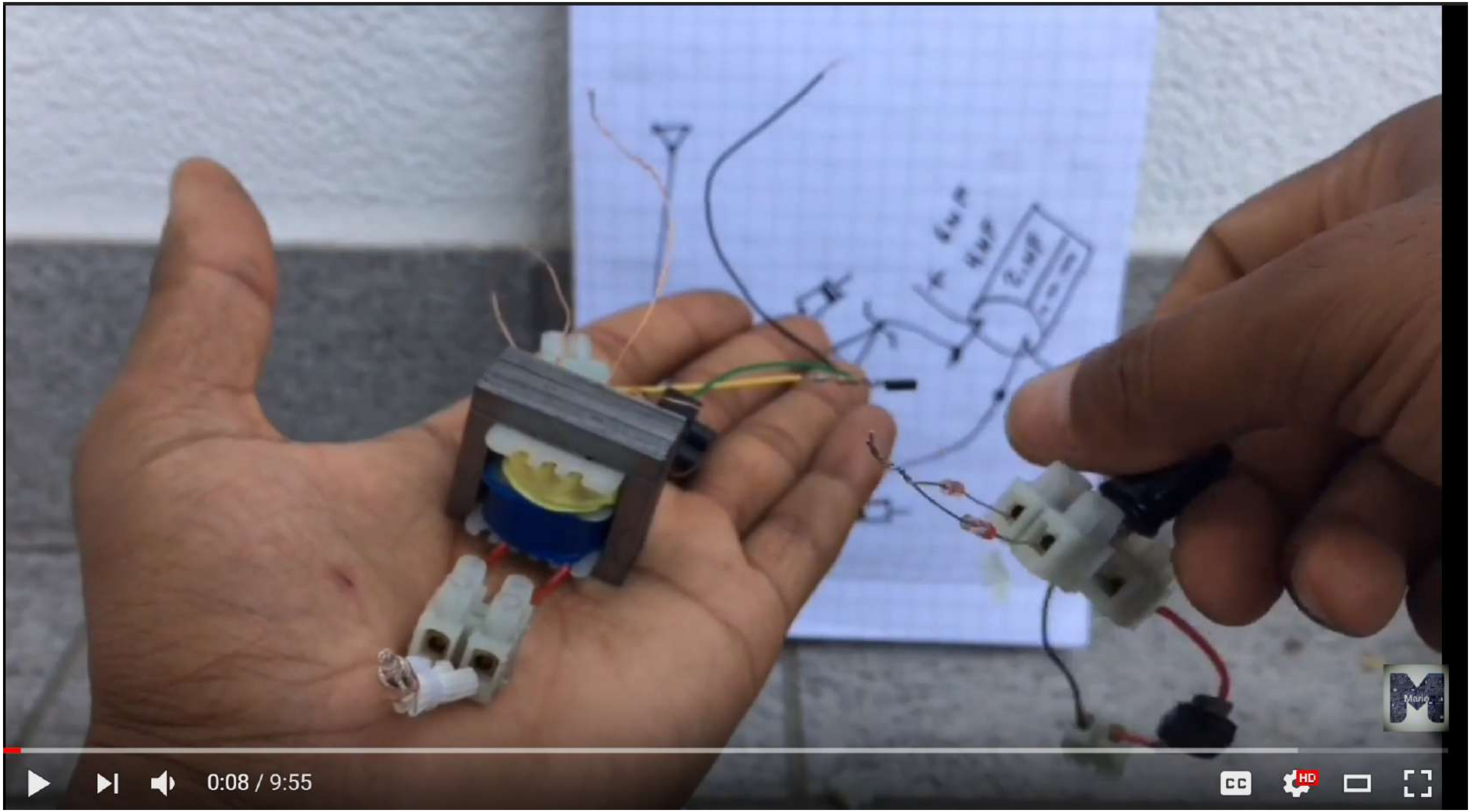
0:08 / 9:55
HD
Mario

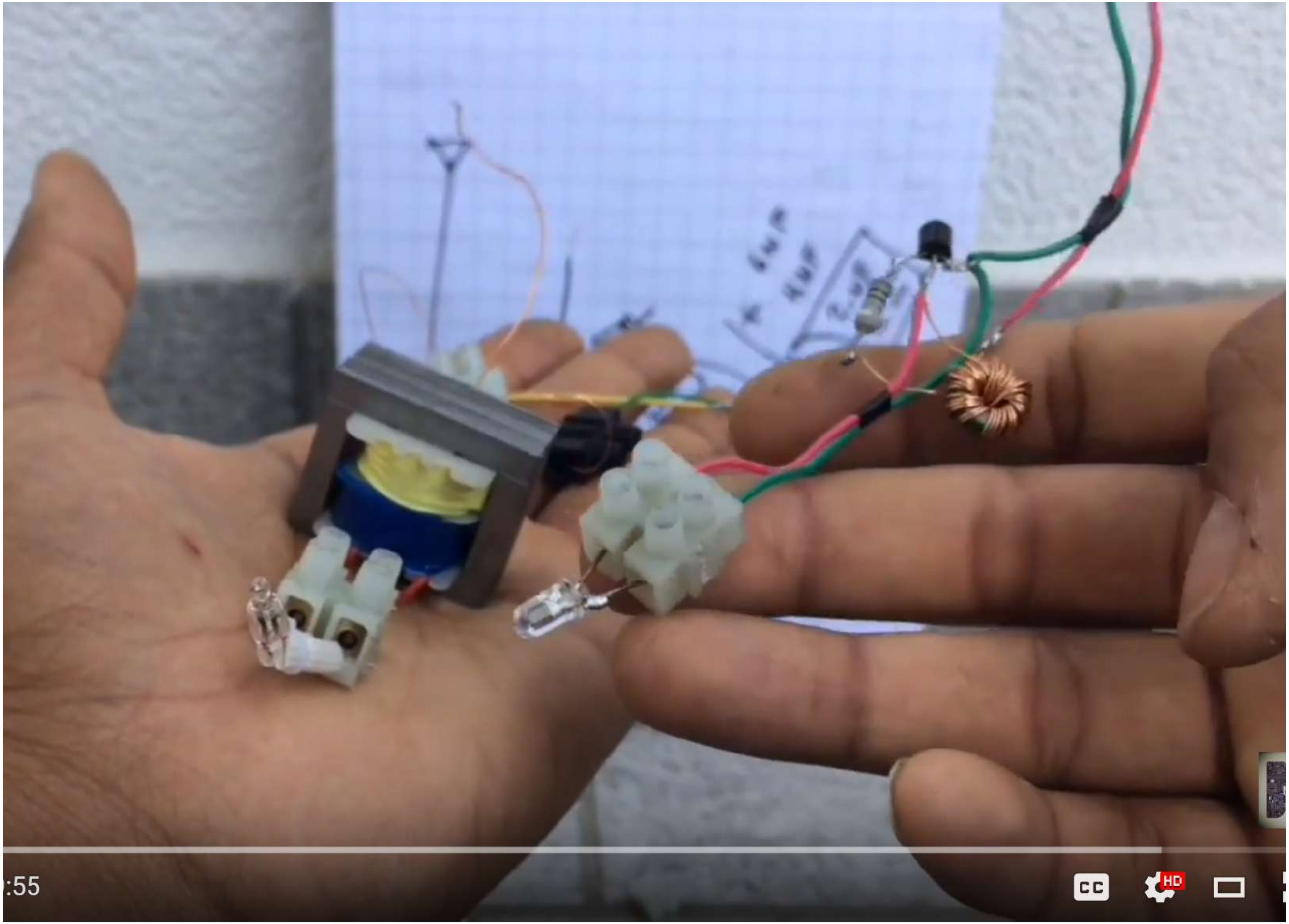
:55
CC
HD

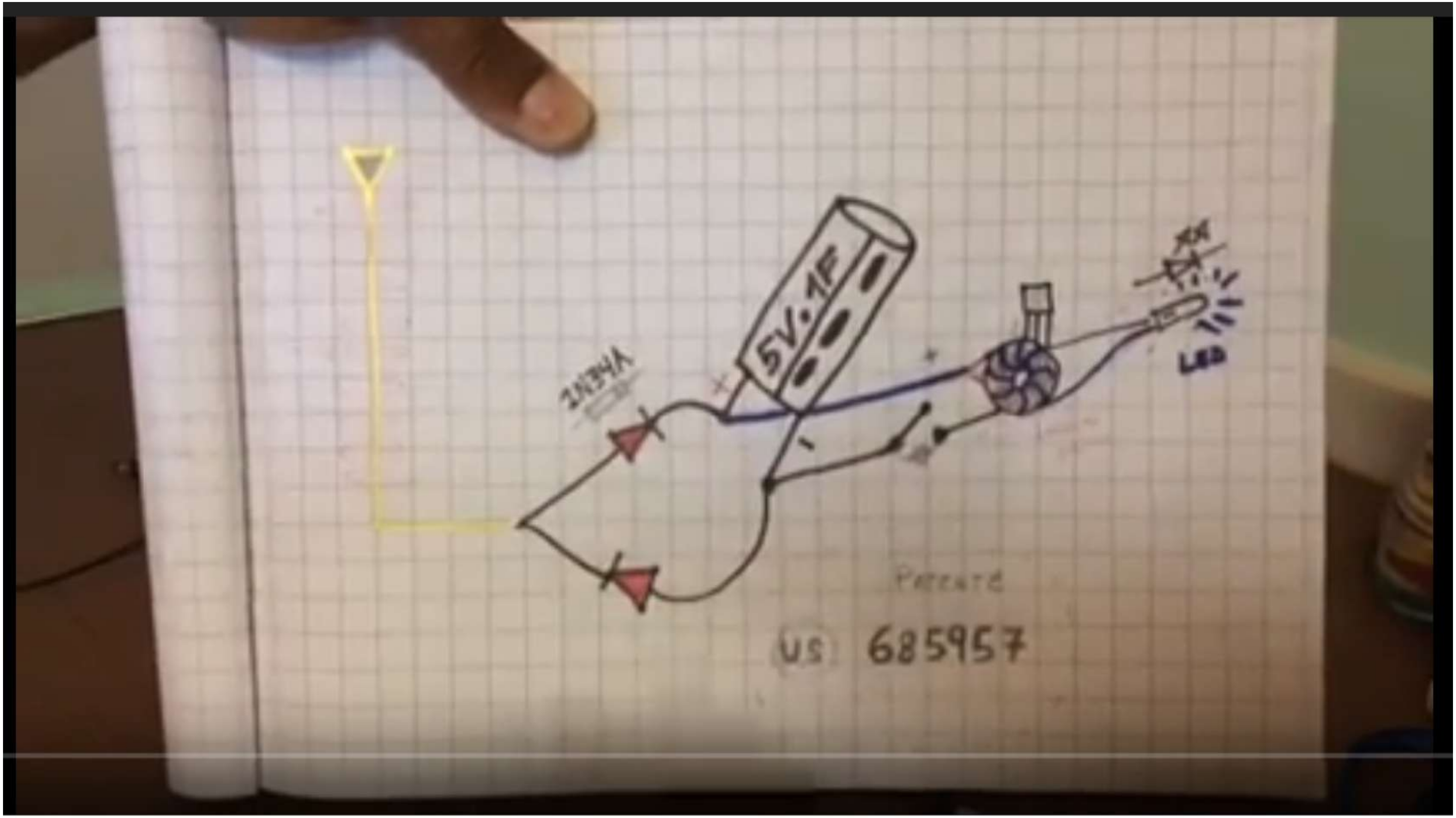
1N34A
5V·1F
+
-
+
LED
Patente
US 685957

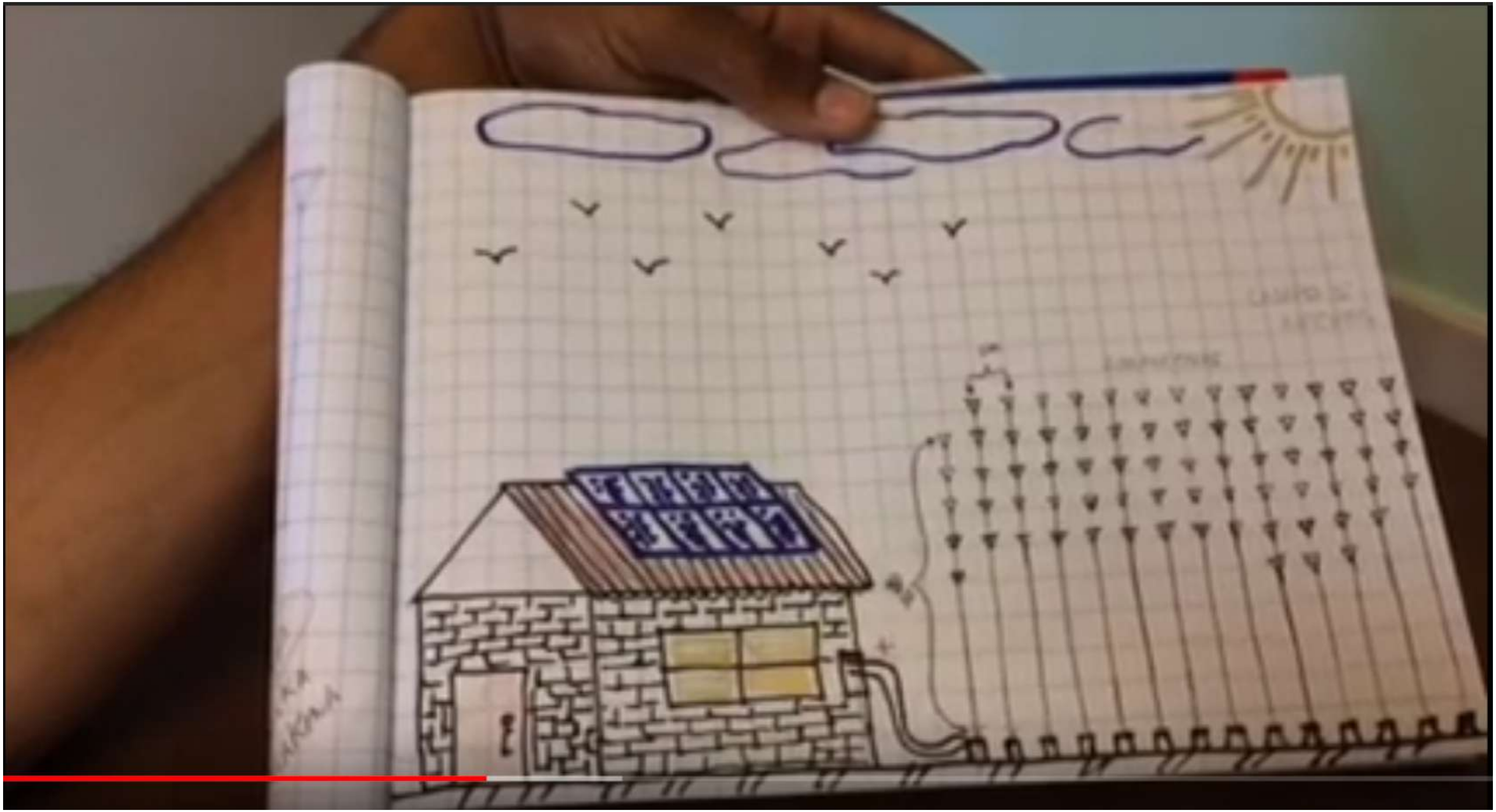

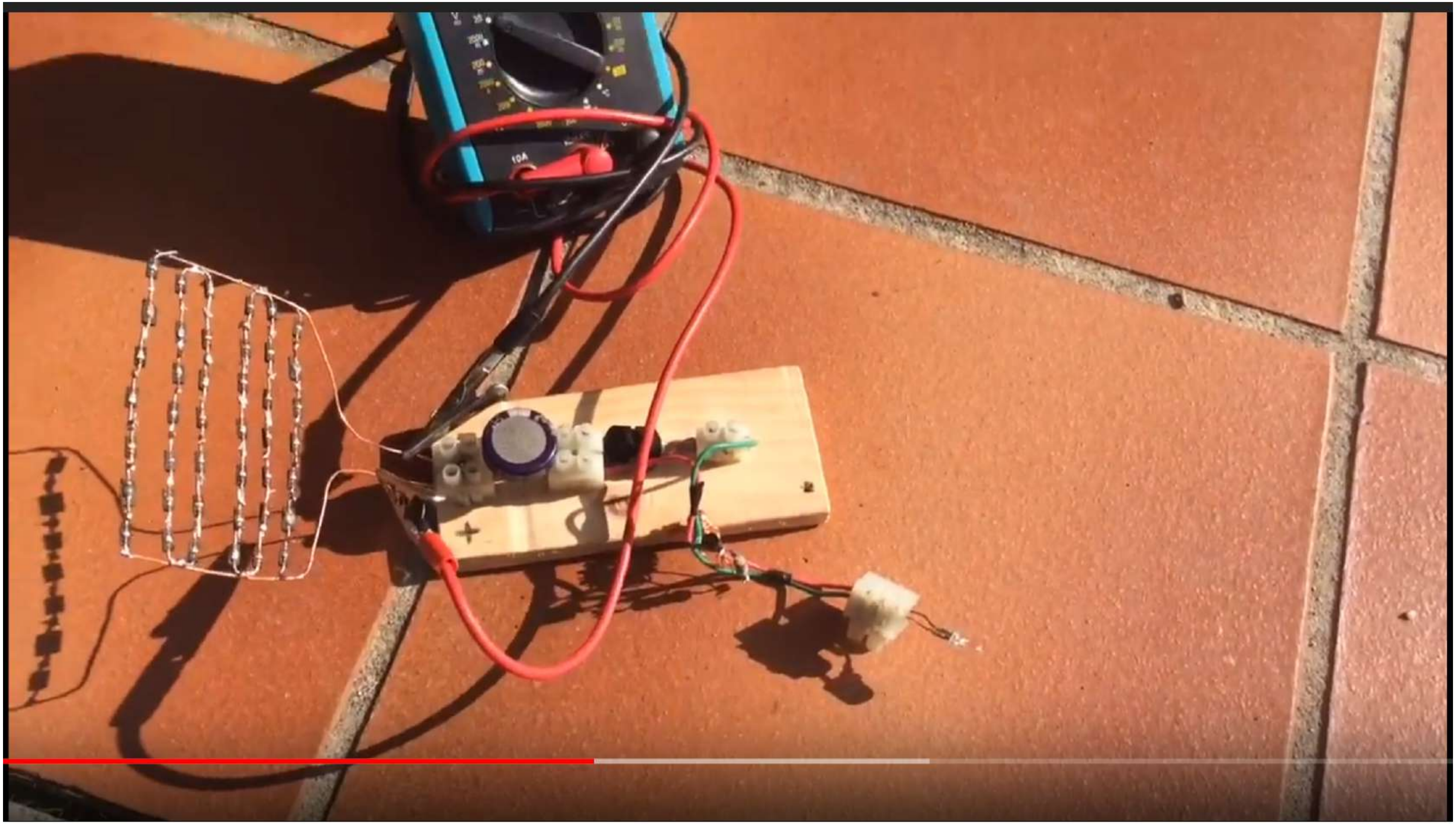
10A

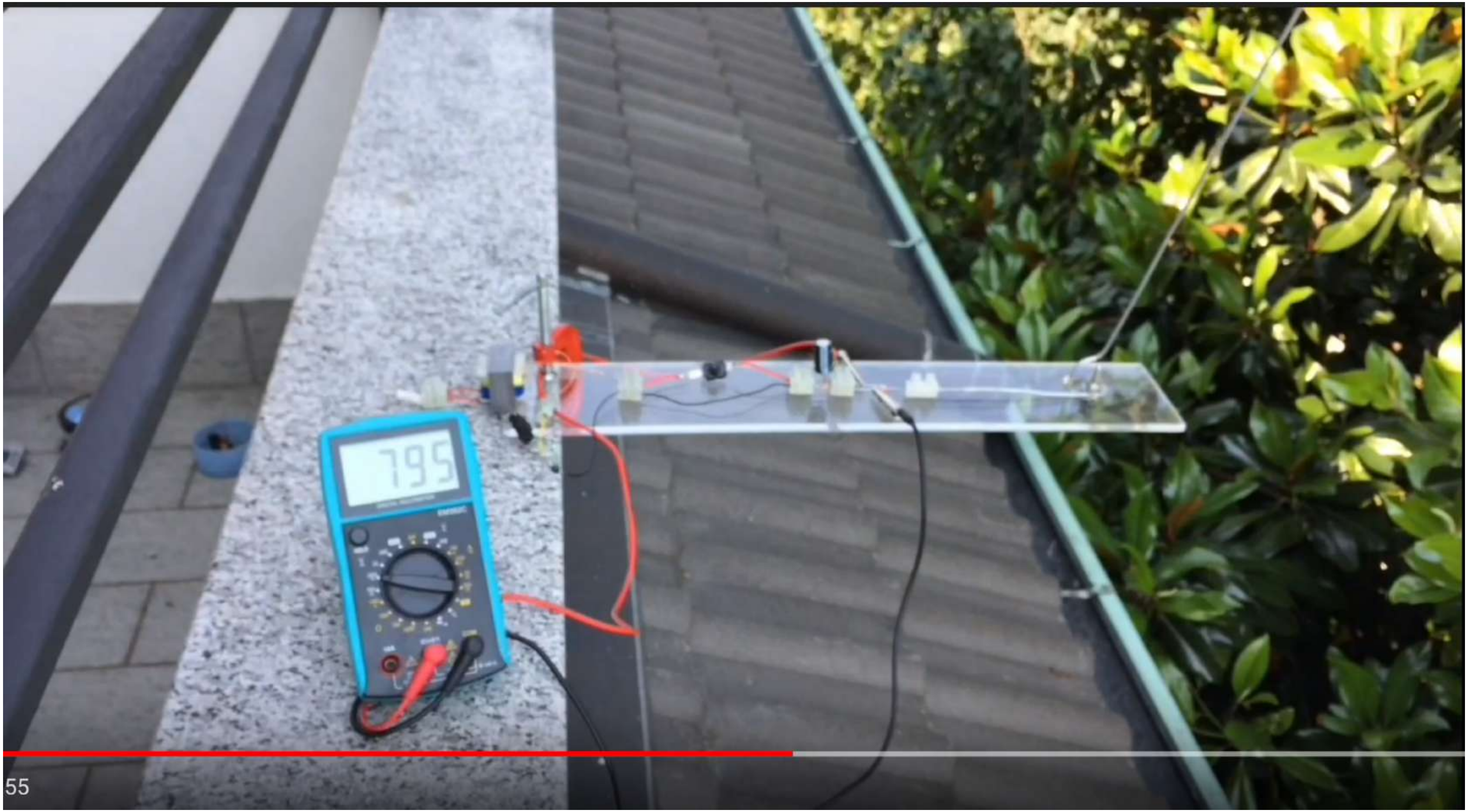
795
55

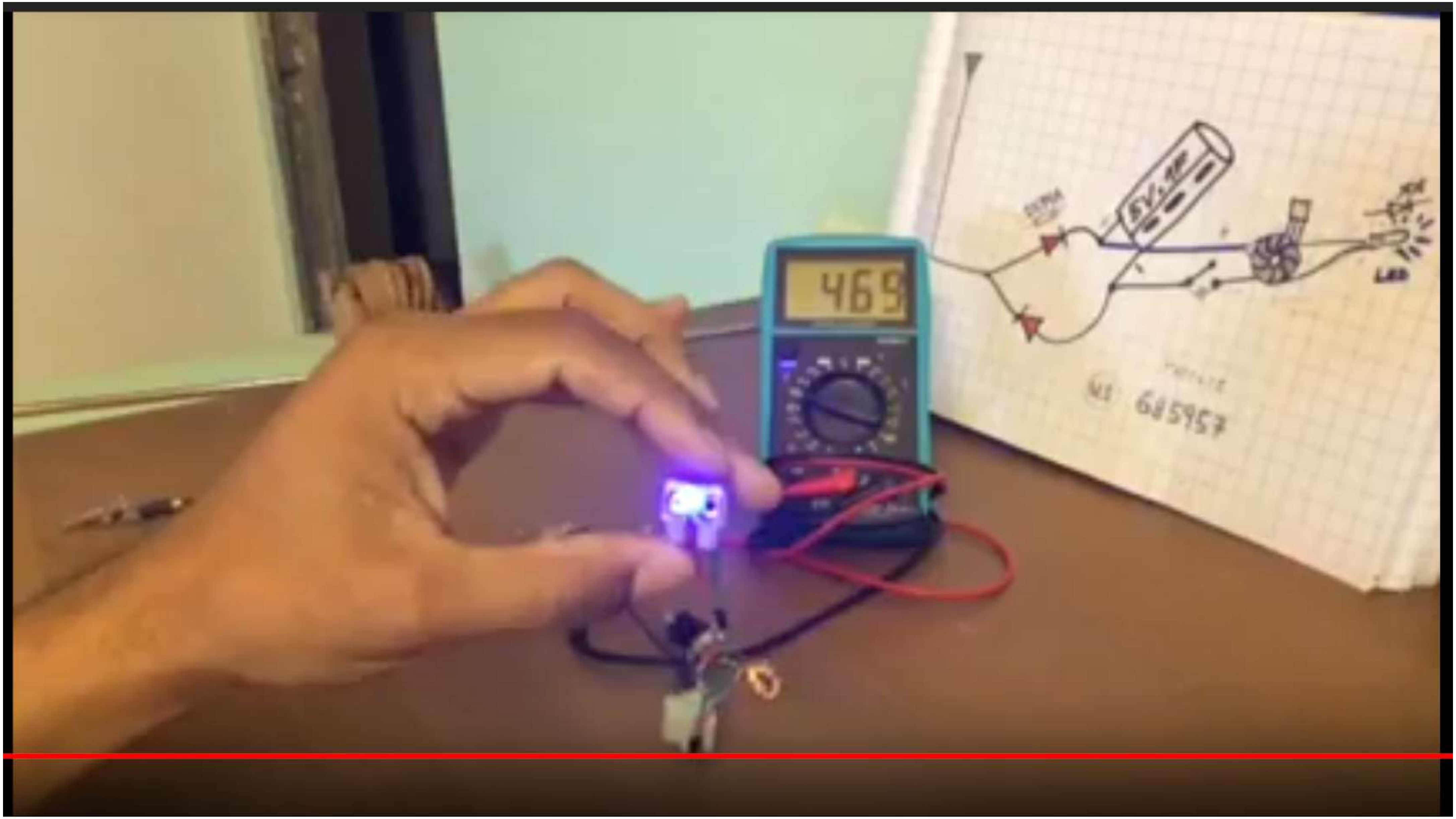
465
LED
685957

THE SUPER CRYSTAL SET
AERIAL TAP
0
1
2
3
4
5
6
7
8
9
10
TUNING
0
10
20
30
40
50
60
70
80
90
100
PHONES

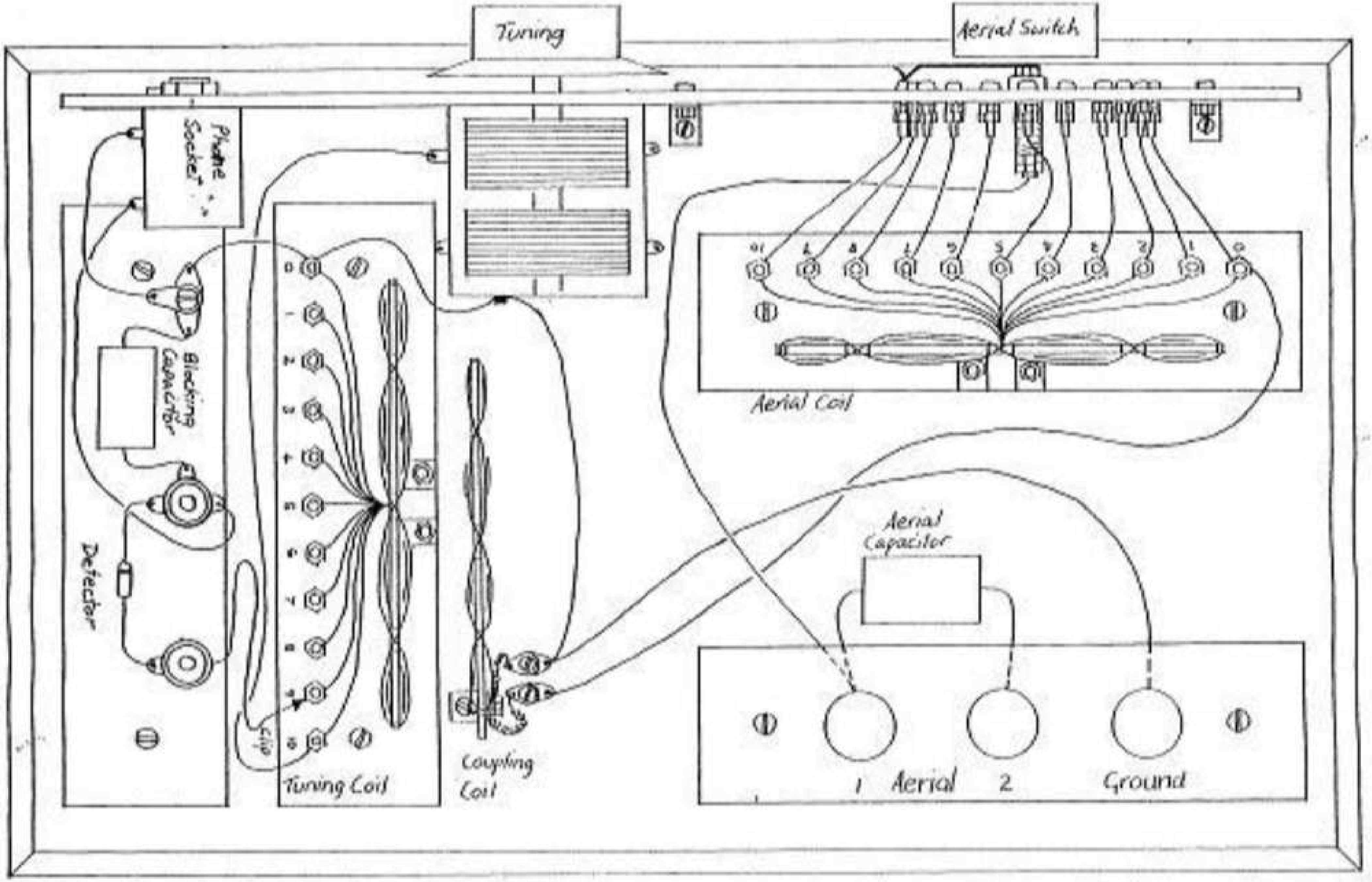
Tuning
Aerial Switch
Phone Socket
Blocking Capacitor
Detector
Clip
Tuning Coil
Coupling Coil
Aerial Coil
Aerial Capacitor
1
Aerial
2
Ground

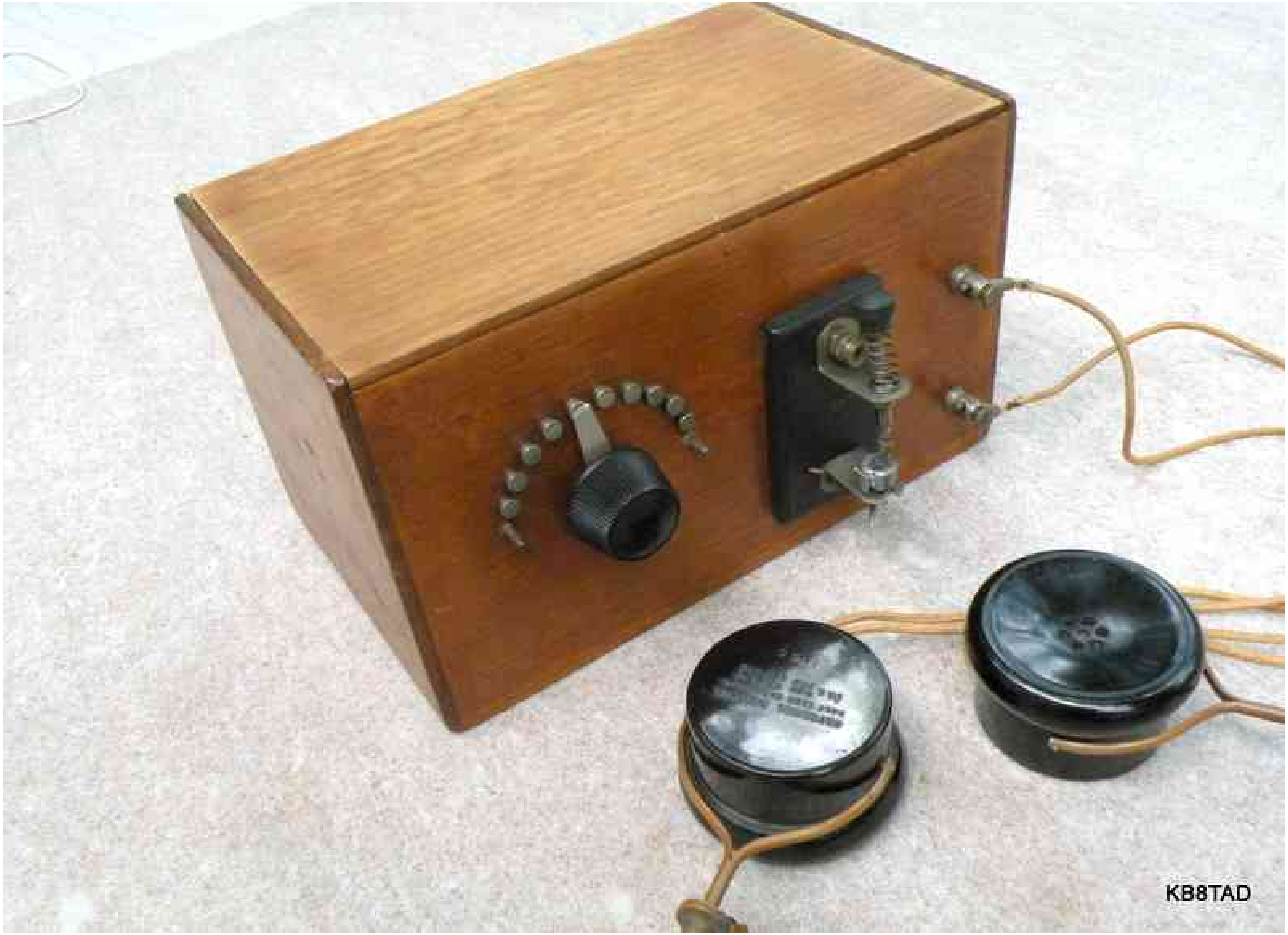

KB8TAD

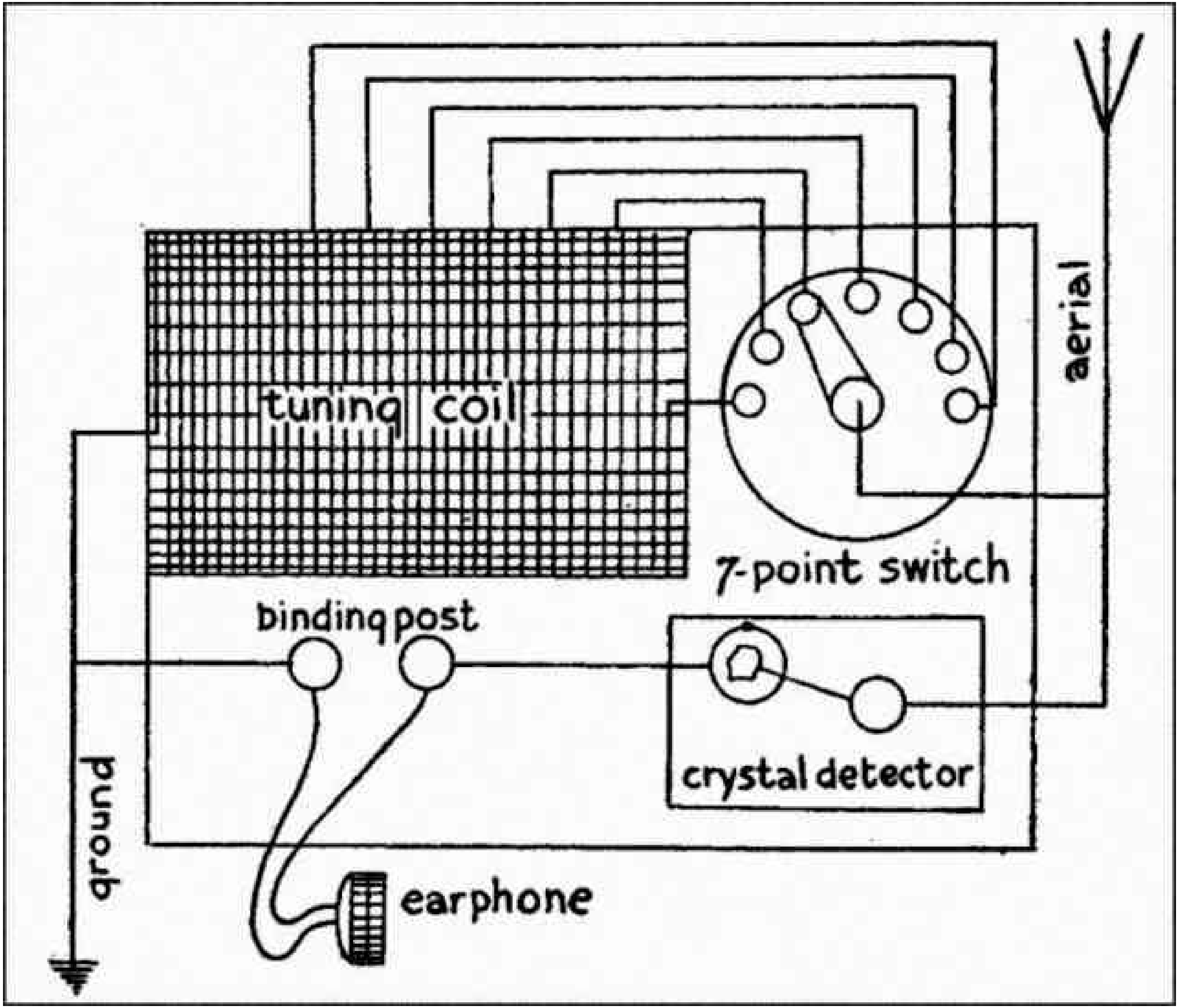
aerial
tuning coil
7-point switch
binding post
crystal detector
ground
earphone

KB8TAD

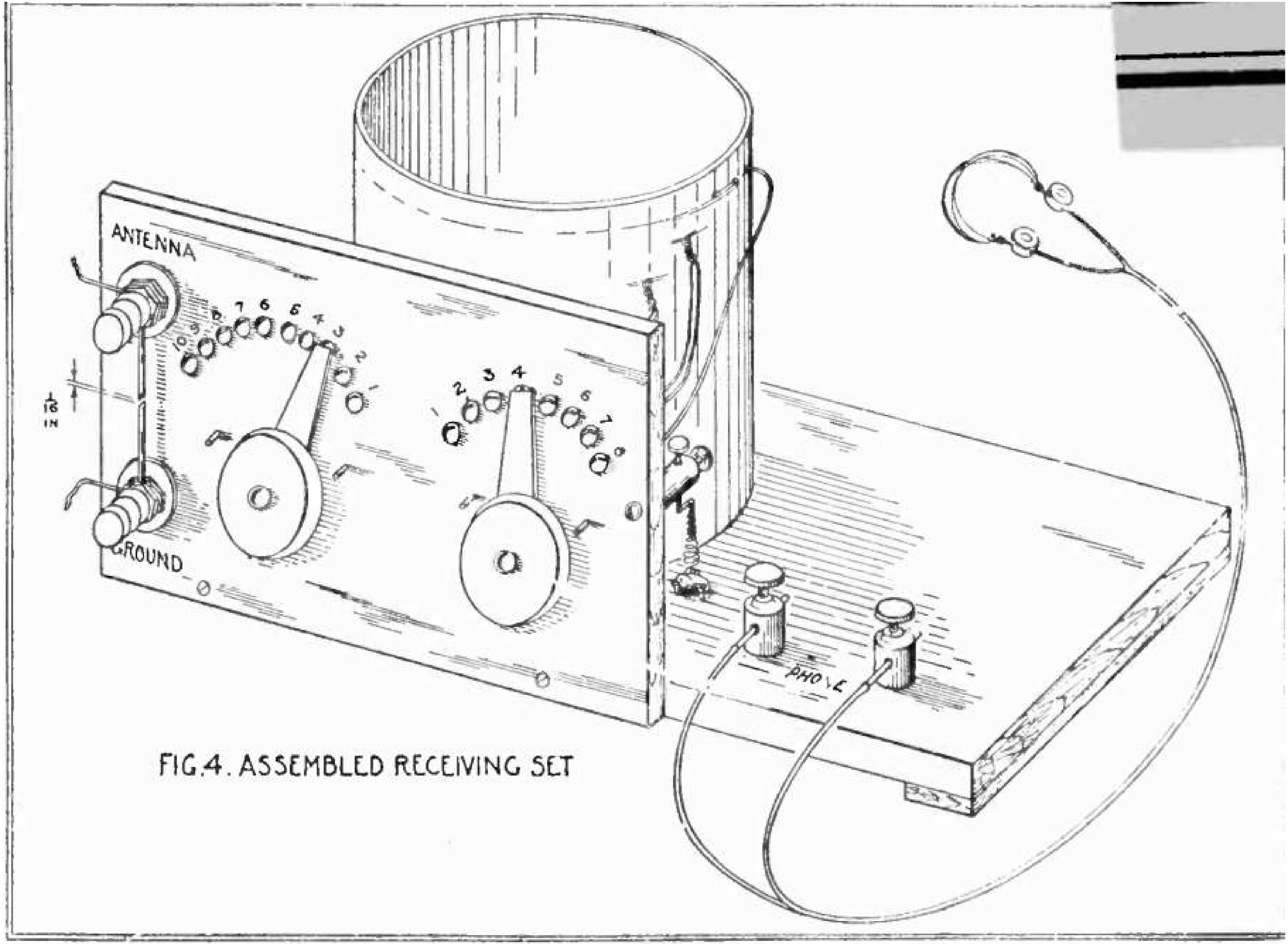
ANTENNA
GROUND
1/16 IN
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
PHONE

FIG. 4. ASSEMBLED RECEIVING SET

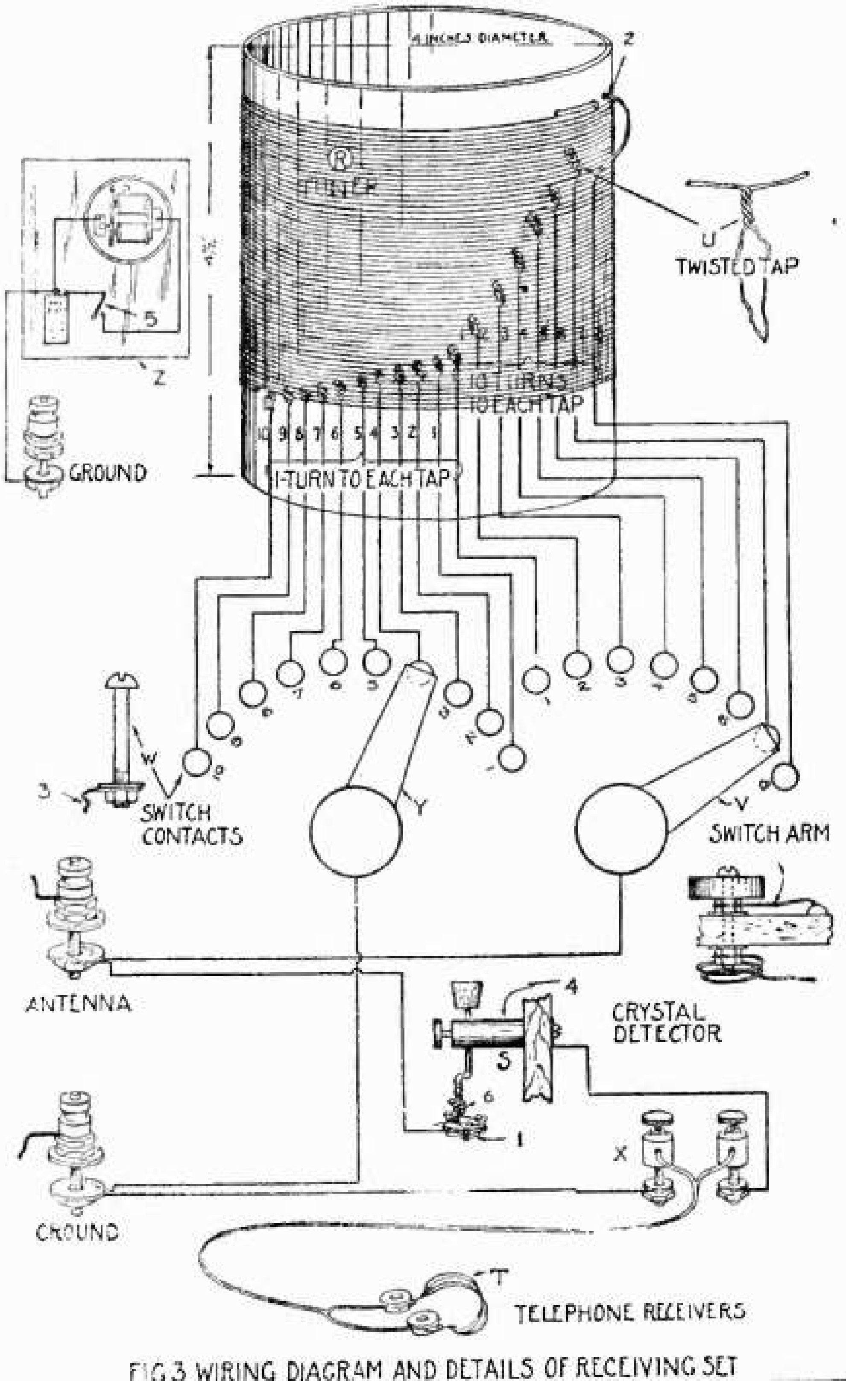
4 INCHES DIAMETER
Z
R
TUNER
U
TWISTED TAP
4½
5
Z
GROUND
10 TURNS TO EACH TAP
10 9 8 7 6 5 4 3 2 1
1-TURN TO EACH TAP
W
3
SWITCH CONTACTS
Y
V
SWITCH ARM
ANTENNA
4
CRYSTAL DETECTOR
5
6
1
X
GROUND
T
TELEPHONE RECEIVERS

FIG.3 WIRING DIAGRAM AND DETAILS OF RECEIVING SET

FIG 2. RECEIVING SET, WITH ANTENNA AND GROUND CONNECTIONS

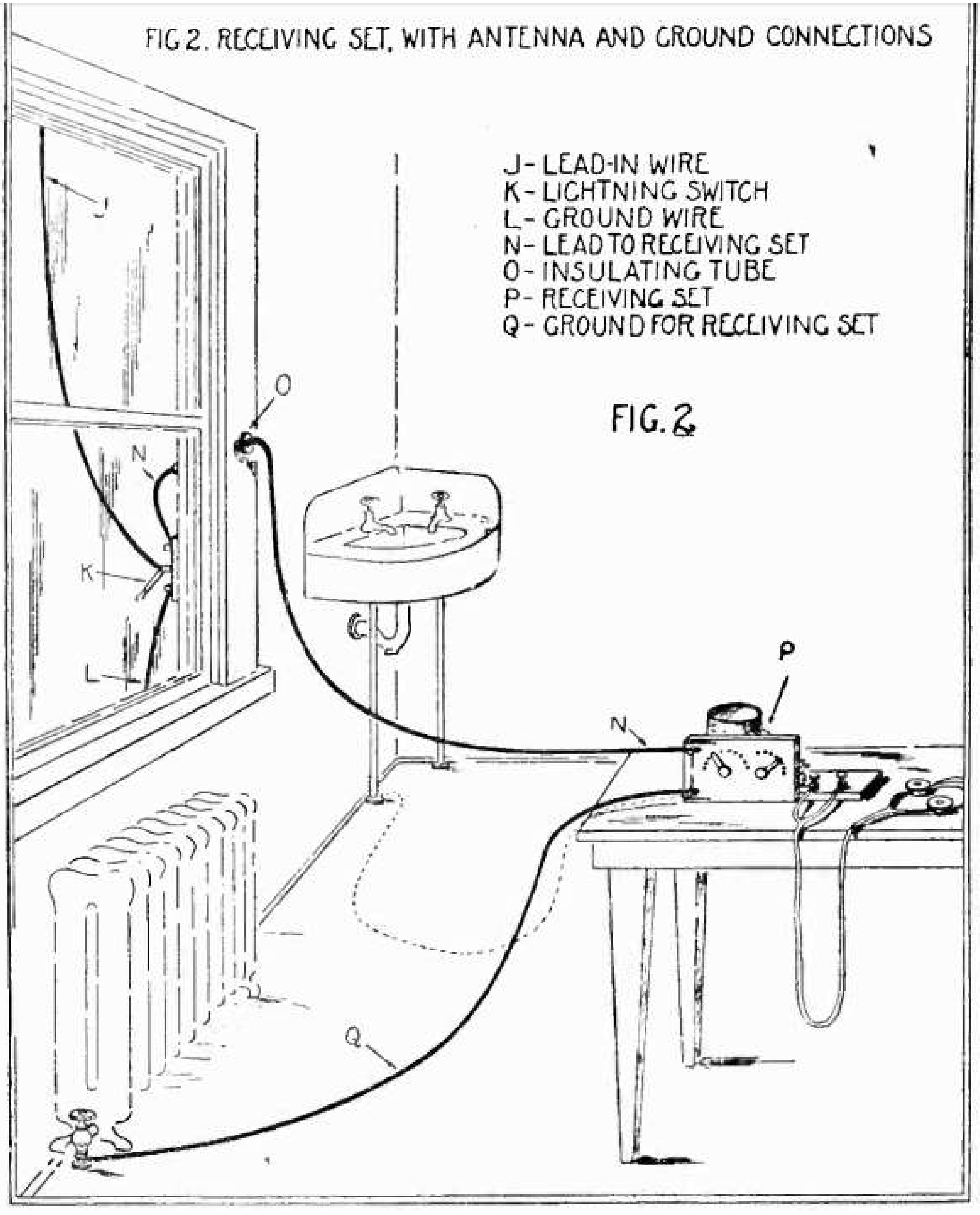

J- LEAD-IN WIRE
K- LIGHTNING SWITCH
L- GROUND WIRE
N- LEAD TO RECEIVING SET
O- INSULATING TUBE
P- RECEIVING SET
Q- GROUND FOR RECEIVING SET

FIG. 2

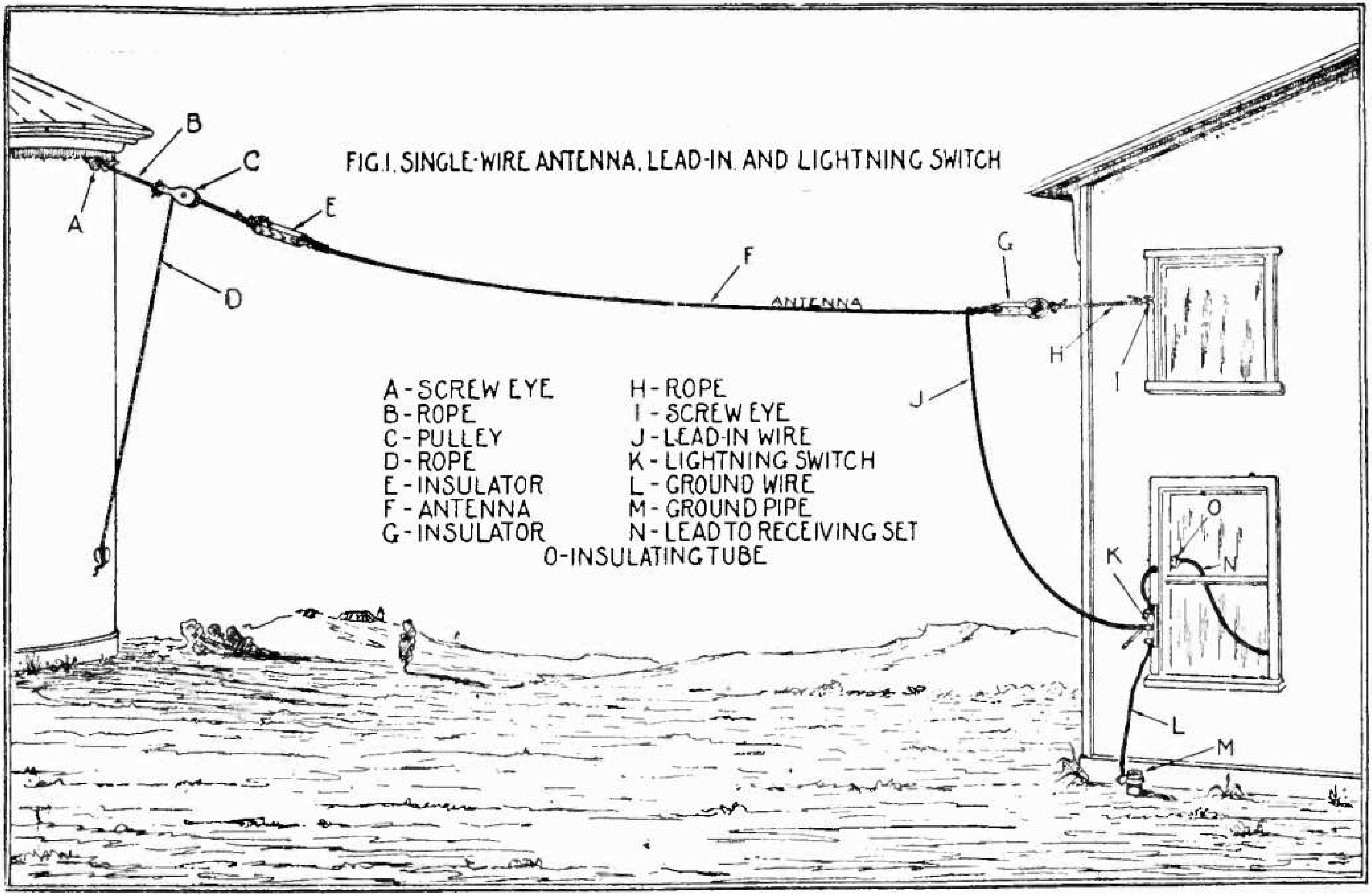

FIG. I. SINGLE-WIRE ANTENNA, LEAD-IN AND LIGHTNING SWITCH

A
B
C
D
E
F
ANTENNA
G
H
I
J
K
O
N
L
M
A - SCREW EYE
B - ROPE
C - PULLEY
D - ROPE
E - INSULATOR
F - ANTENNA
G - INSULATOR
H - ROPE
I - SCREW EYE
J - LEAD-IN WIRE
K - LIGHTNING SWITCH
L - GROUND WIRE
M - GROUND PIPE
N - LEAD TO RECEIVING SET
O - INSULATING TUBE

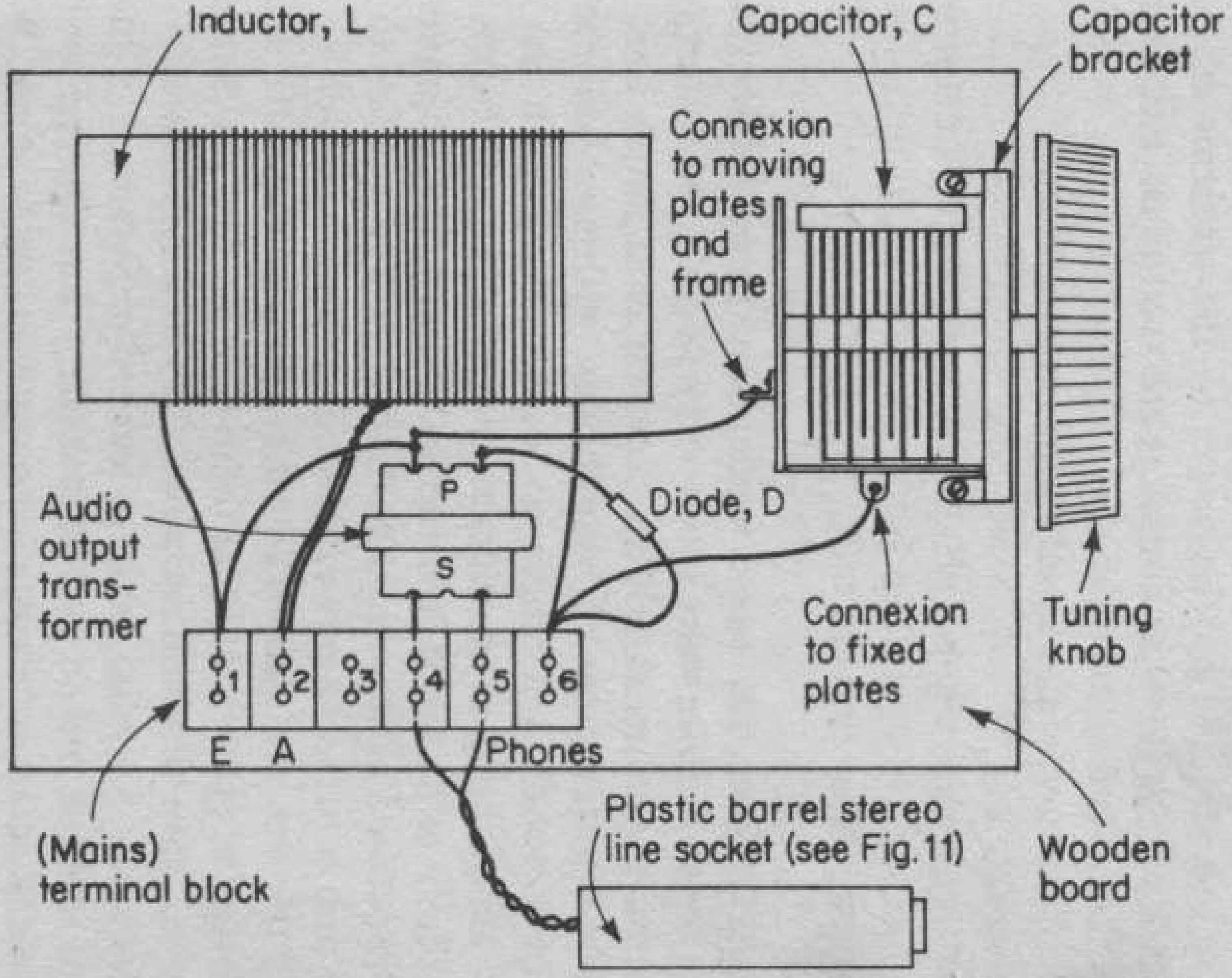
Inductor, L
Capacitor, C
Capacitor bracket
Connexion to moving plates and frame
Audio output transformer
P
S
Diode, D
Connexion to fixed plates
Tuning knob
1
2
3
4
5
6
E
A
Phones
(Mains) terminal block
Plastic barrel stereo line socket (see Fig. 11)
Wooden board

Fig. 14 Suggested component arrangement for elementary crystal set

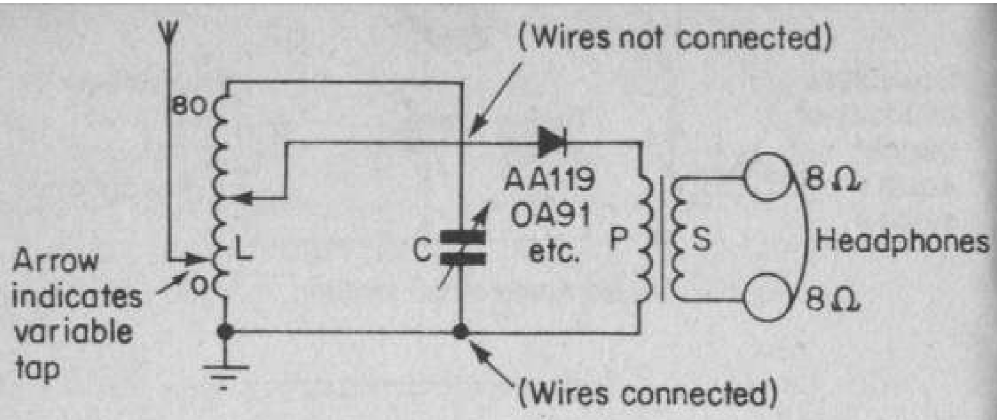
(Wires not connected)
80
AA119
OA91
etc.
C
P
S
8 Ω
Headphones
8 Ω
L
Arrow indicates variable tap
0
(Wires connected)

(i) Circuit

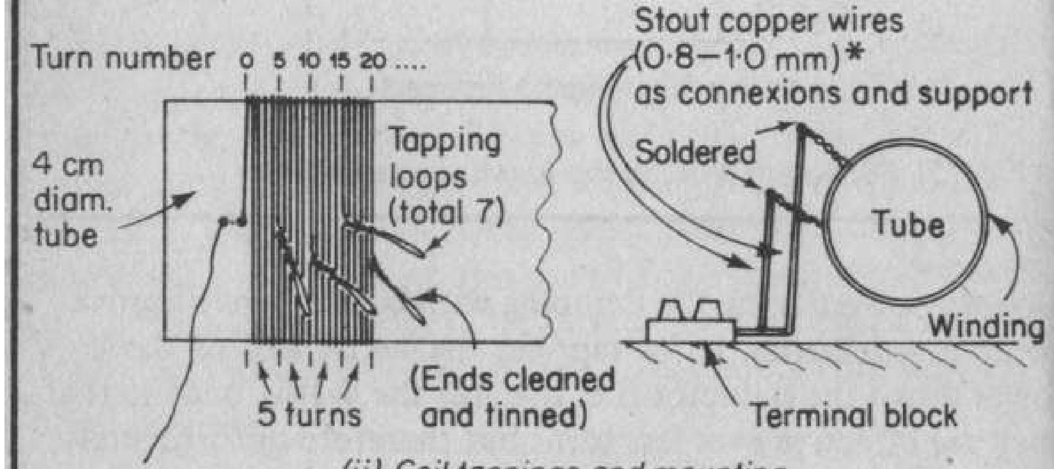
Turn number 0 5 10 15 20 ....
4 cm diam. tube
Tapping loops (total 7)
5 turns
(Ends cleaned and tinned)
Stout copper wires (0·8–1·0 mm)* as connexions and support
Soldered
Tube
Winding
Terminal block

(ii) Coil tappings and mounting

Fig. 16 Circuit with tapped inductor

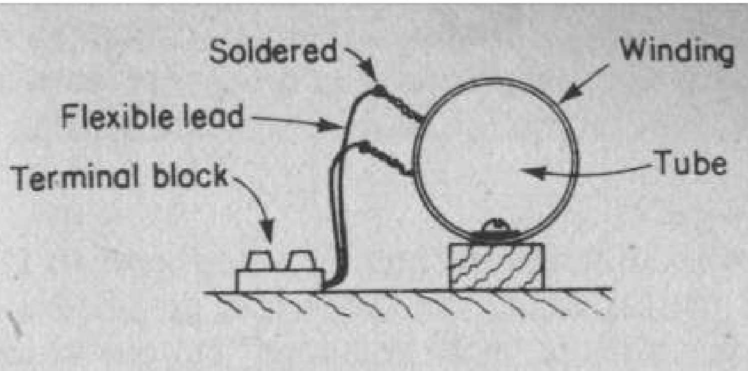
Soldered
Winding
Flexible lead
Tube
Terminal block

(iii) Alternative mounting

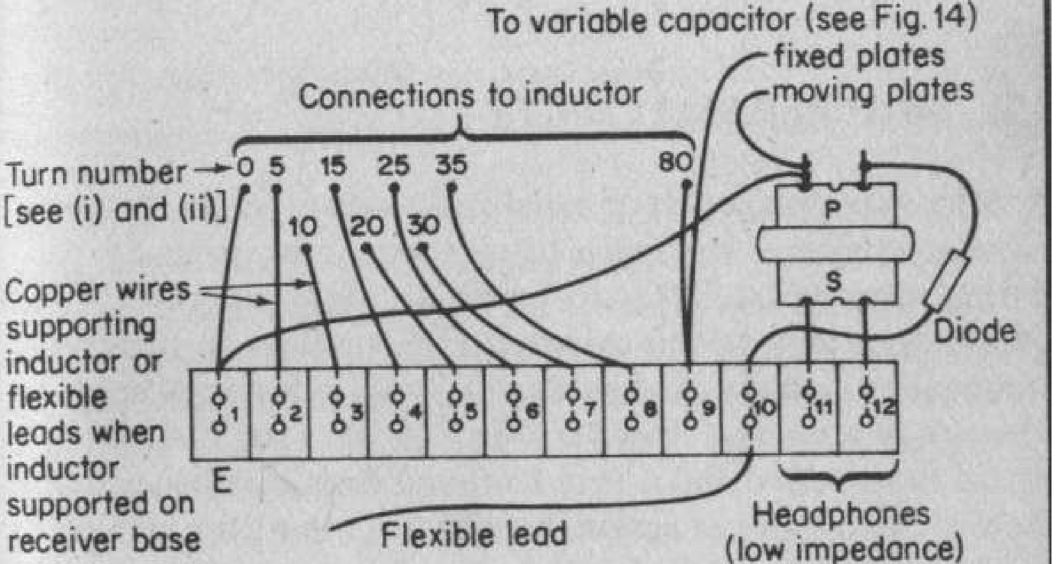
To variable capacitor (see Fig. 14)
fixed plates
moving plates
Connections to inductor
Turn number → 0 5 15 25 35 80
[see (i) and (ii)]
10 20 30
P
S
Copper wires supporting inductor or flexible leads when inductor supported on receiver base
Diode
1 2 3 4 5 6 7 8 9 10 11 12
E
Flexible lead
Headphones (low impedance)

(iv) Suggested component arrangement

---

* 21–19 SWG, 20–18 AWG

required) and a *flying lead* for the diode, these components are shown in Fig.16(iv). This is doing the job properly, there is no reason however (except for unreliability of contact) why both aerial and diode flying leads should not terminate on *crocodile clips* and be clipped onto the chosen taps as required.

Earth is connected to terminal 1 and the headphones to 11 and 12. Both the aerial and the diode lead are tried at terminals 2 – 9, a game of "poke and hope" but one which can be rewarding. Remember that generally but not always, the tappings with the lower numbers reduce loudness but increase selectivity. Aerial to terminal 8, diode to 9 brings us back to Fig.13(i) of course.

5.2 SWITCHED SELECTIVITY

A technique of selection of aerial and diode tappings by means of rotary switches has much to recommend it especially if setting changes are likely to be needed for reception of different stations. In this case two single-pole 8-way switches are required, usually obtainable as 12-way so leaving 4 spare. The drawing symbol shown in Fig.17 speaks for itself with regard to its action and it may be found from catalogues that there are two types of *action*, by which is meant the manner in which the switch changes over from one contact to the adjoining one. *Break before make* switches disconnect one circuit before the next is connected and *make before break* have a change-over period during which both circuits are connected at once for a short time. Either type is suitable for us. A knob to rotate each switch is also required, a small one, say 2 cm diameter or less is ample, if with an indicator line or pointer, so much the better. The modifications to Fig.16 for switching are given in Fig.17.

With such rotary switches a search for the optimum arrange-

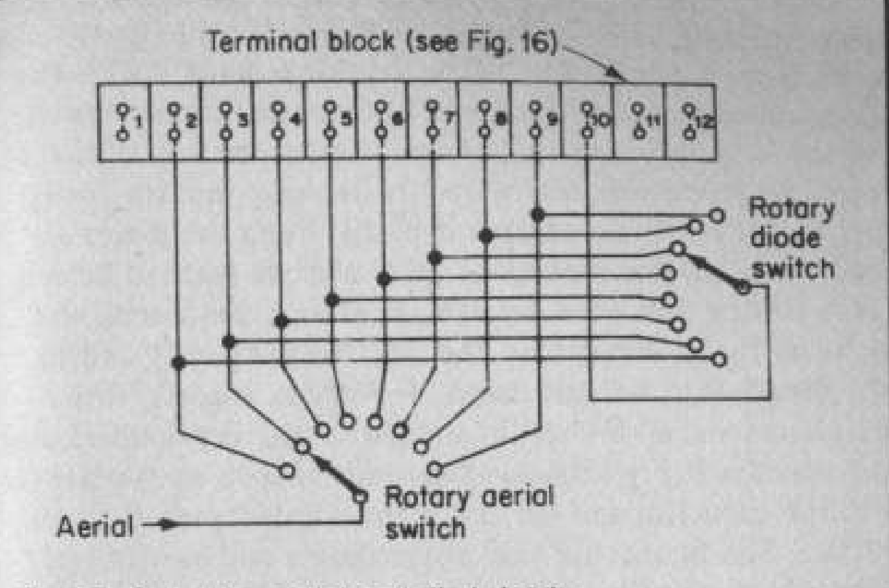

Fig. 17 · Switching aerial and diode leads

variable capacitor, the two switches and the headphone socket, the latter must then be of the chassis type. A pair of terminals for aerial and earth completes the receiver. On the panel the two rotary switches are easily labelled 1 – 8 but we are still uncertain about a tuning dial because we may not have the right variable capacitor for the inductor, this is the subject of the next section.

5.3 GETTING THE RANGE RIGHT

This is where we pause in receiver construction to remove some of the confusion which may still exist with regard to the tuning ingredients, wavelength, frequency, inductance and capacity, each of which has some bearing on the others, to fit together these four pieces of the puzzle, so to speak. For most crystal sets the inductance is fixed according to the variable capacitor available and we tune the receiver over the range by rotating the variable capacitor knob over half a

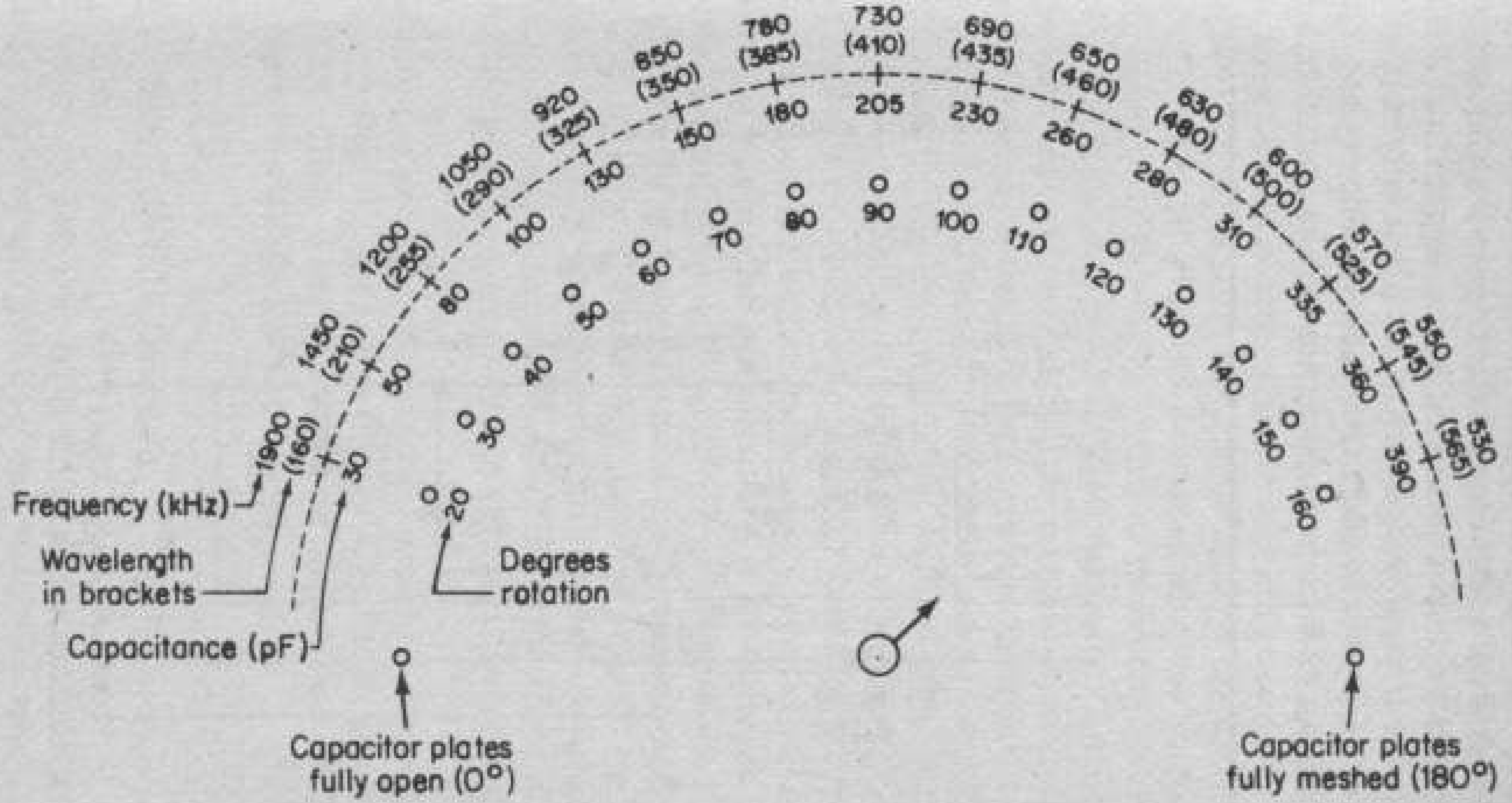
1900
(160)
30
1450
(210)
50
1200
(255)
80
1050
(290)
100
920
(325)
130
850
(350)
150
780
(385)
180
730
(410)
205
690
(435)
230
650
(460)
260
630
(480)
280
600
(500)
310
570
(525)
335
550
(545)
360
530
(565)
390
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
Frequency (kHz)
Wavelength in brackets
Capacitance (pF)
Degrees rotation
Capacitor plates fully open (0°)
Capacitor plates fully meshed (180°)

Fig. 18 Approximate capacitance and frequency values for 365 pF variable capacitor and 230 µH coil

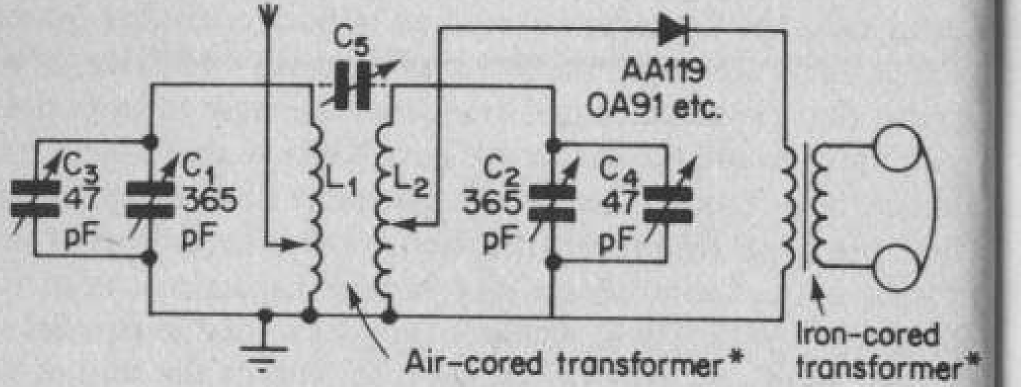
C5
AA119
OA91 etc.
C3
47
pF
C1
365
pF
L1
L2
C2
365
pF
C4
47
pF
Air-cored transformer*
Iron-cored
transformer*

(i) Circuit

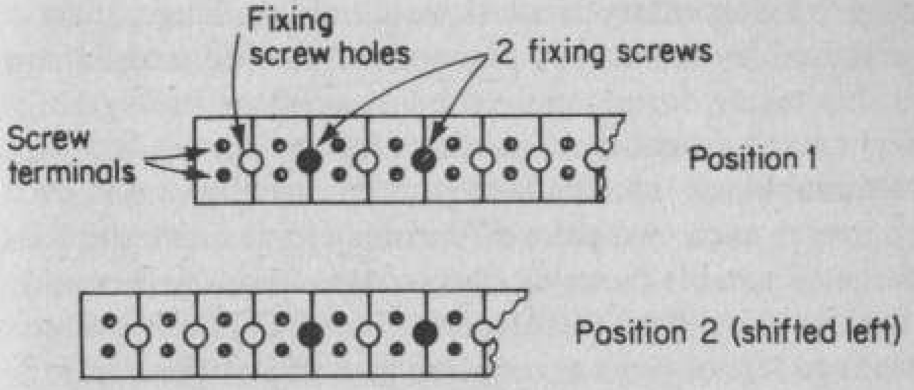
Fixing
screw holes
2 fixing screws
Screw
terminals
Position 1
Position 2 (shifted left)

(iii) Moving inductor L_1

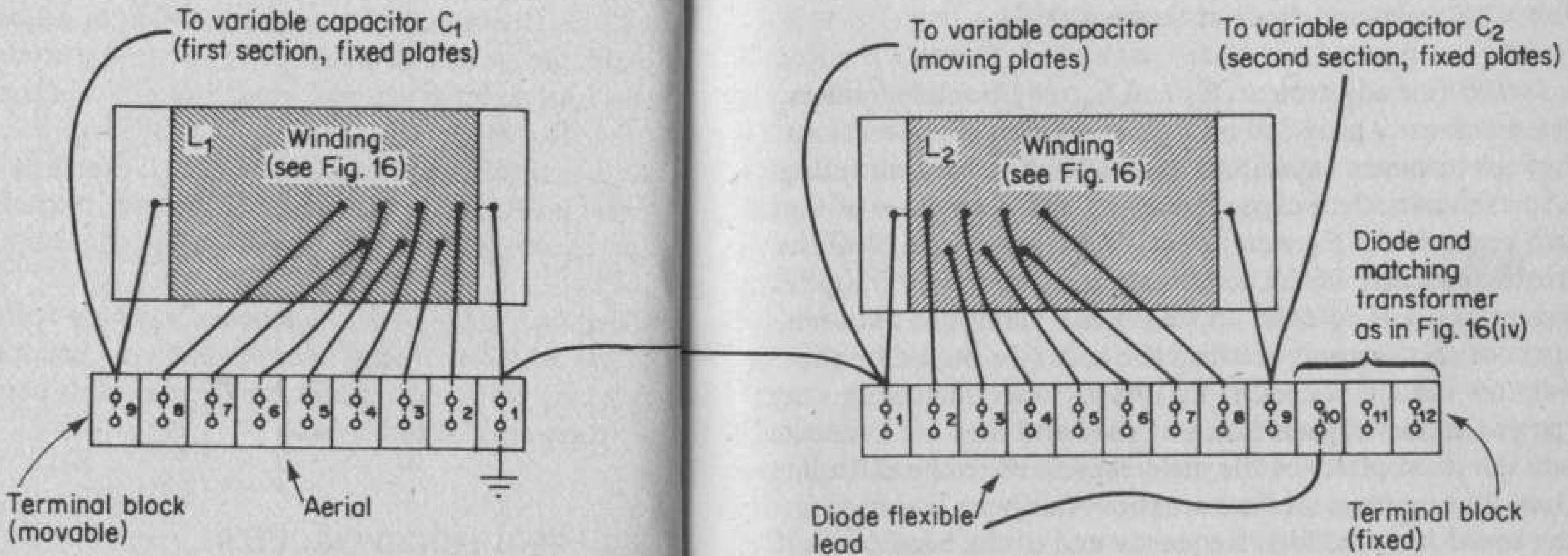
To variable capacitor C1
(first section, fixed plates)
L1
Winding
(see Fig. 16)
9
8
7
6
5
4
3
2
1
Terminal block
(movable)
Aerial
To variable capacitor
(moving plates)
To variable capacitor C2
(second section, fixed plates)
L2
Winding
(see Fig. 16)
Diode and
matching
transformer
as in Fig. 16(iv)
1
2
3
4
5
6
7
8
9
10
11
12
Diode flexible
lead
Terminal block
(fixed)

(ii) Arrangement of inductors

Fig. 21 A double-tuned crystal receiver

* See Section 2.4

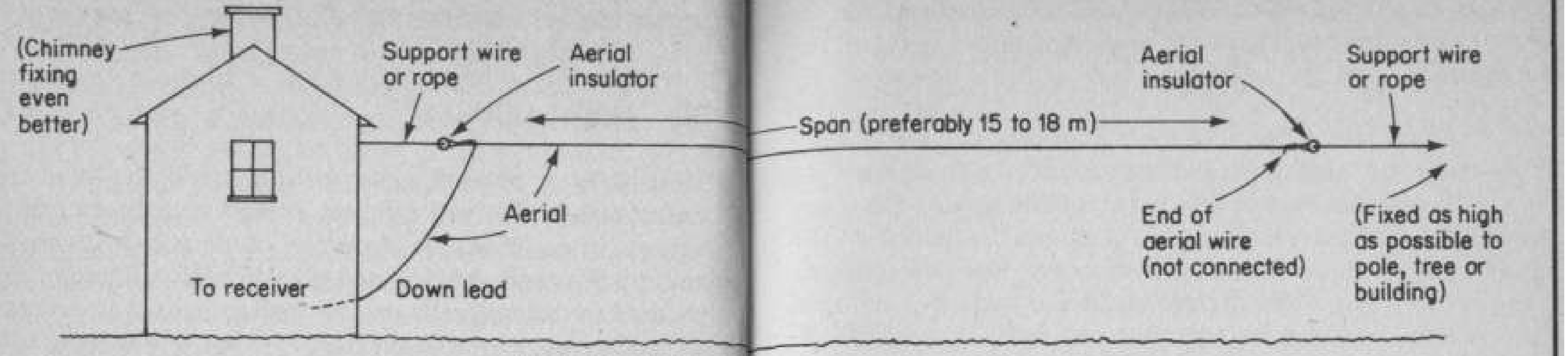
(Chimney fixing even better)
Support wire or rope
Aerial insulator
Aerial
To receiver
Down lead
Span (preferably 15 to 18 m)
Aerial insulator
Support wire or rope
End of aerial wire (not connected)
(Fixed as high as possible to pole, tree or building)

(i) Outdoor Aerial

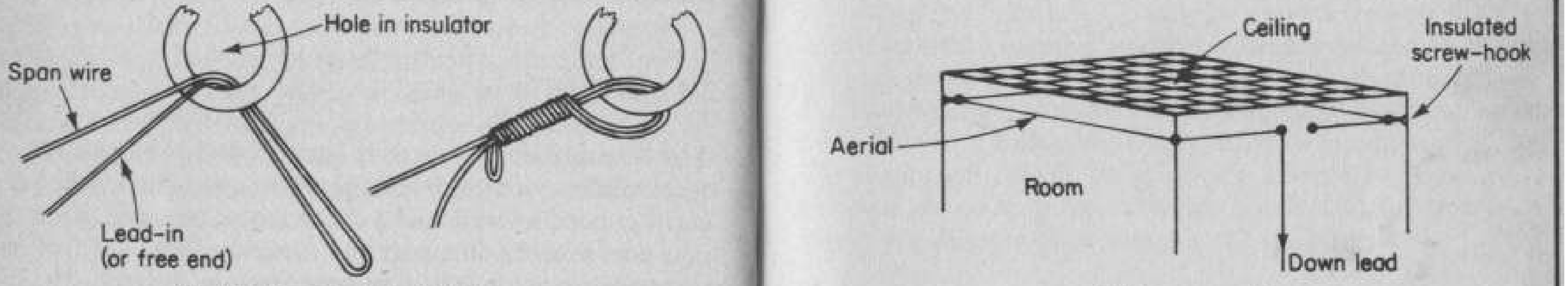
Hole in insulator
Span wire
Lead-in (or free end)
Ceiling
Insulated screw-hook
Aerial
Room
Down lead

(ii) Attaching aerial to insulator

(iii) Room aerial

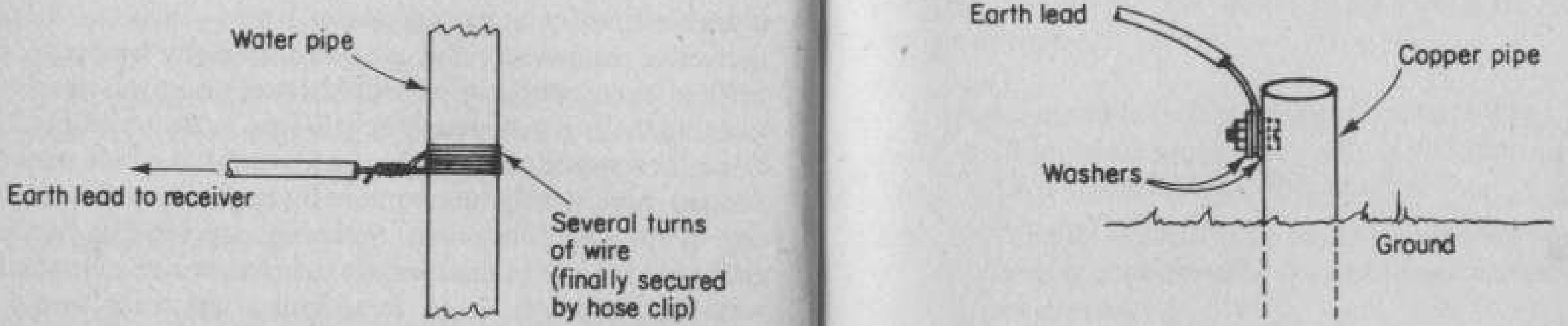
Water pipe
Earth lead to receiver
Several turns of wire (finally secured by hose clip)
Earth lead
Copper pipe
Washers
Ground

(iv) Earth connection to water pipe

(v) Lead bolted to earth spike

Fig. 12 Aerial/earth systems

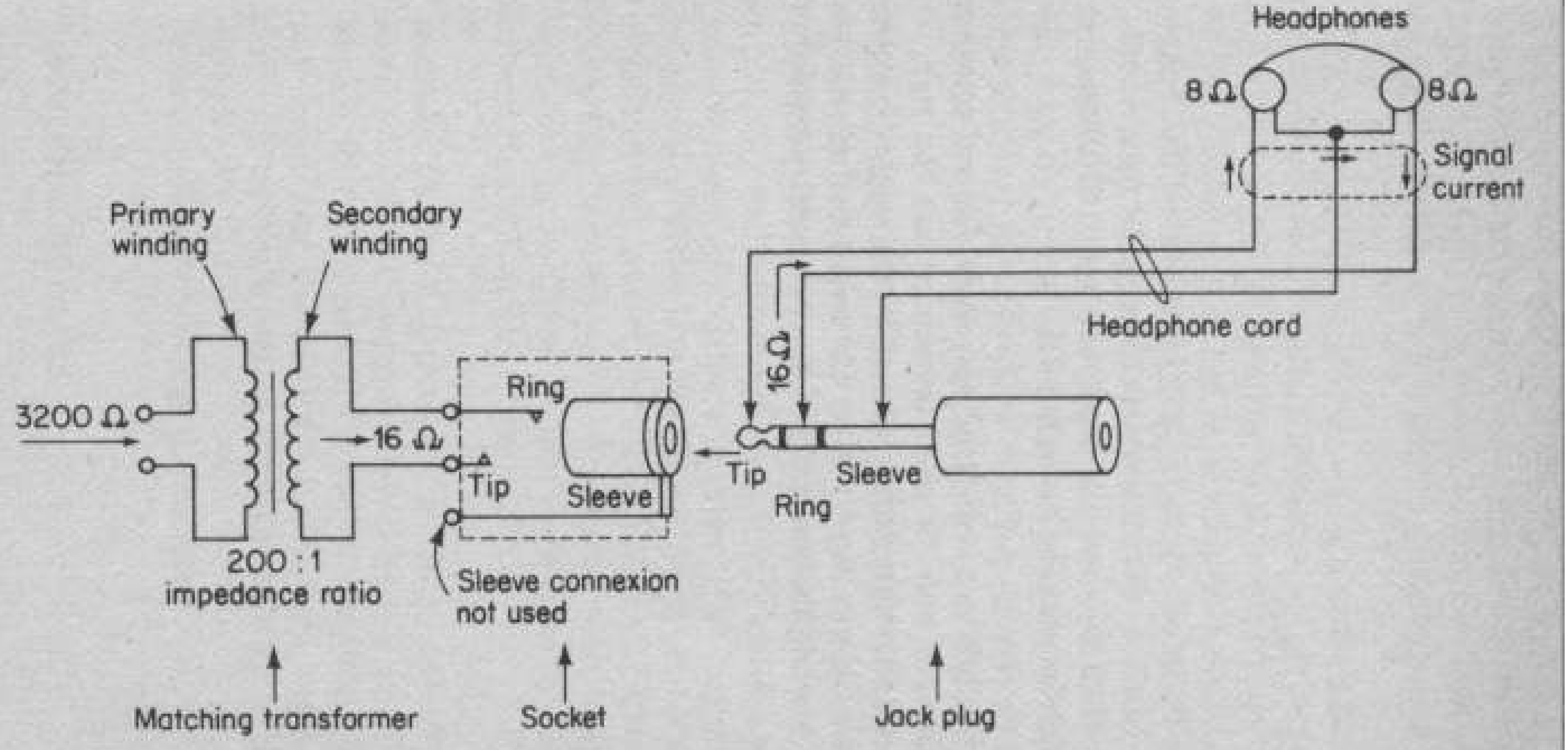
Headphones
8 Ω
8 Ω
Signal current
Primary winding
Secondary winding
Headphone cord
16 Ω
3200 Ω
16 Ω
Ring
Tip
Sleeve
Tip
Ring
Sleeve
200 : 1 impedance ratio
Sleeve connexion not used
Matching transformer
Socket
Jack plug

Fig. 11 Matching a high impedance receiver to low impedance headphones

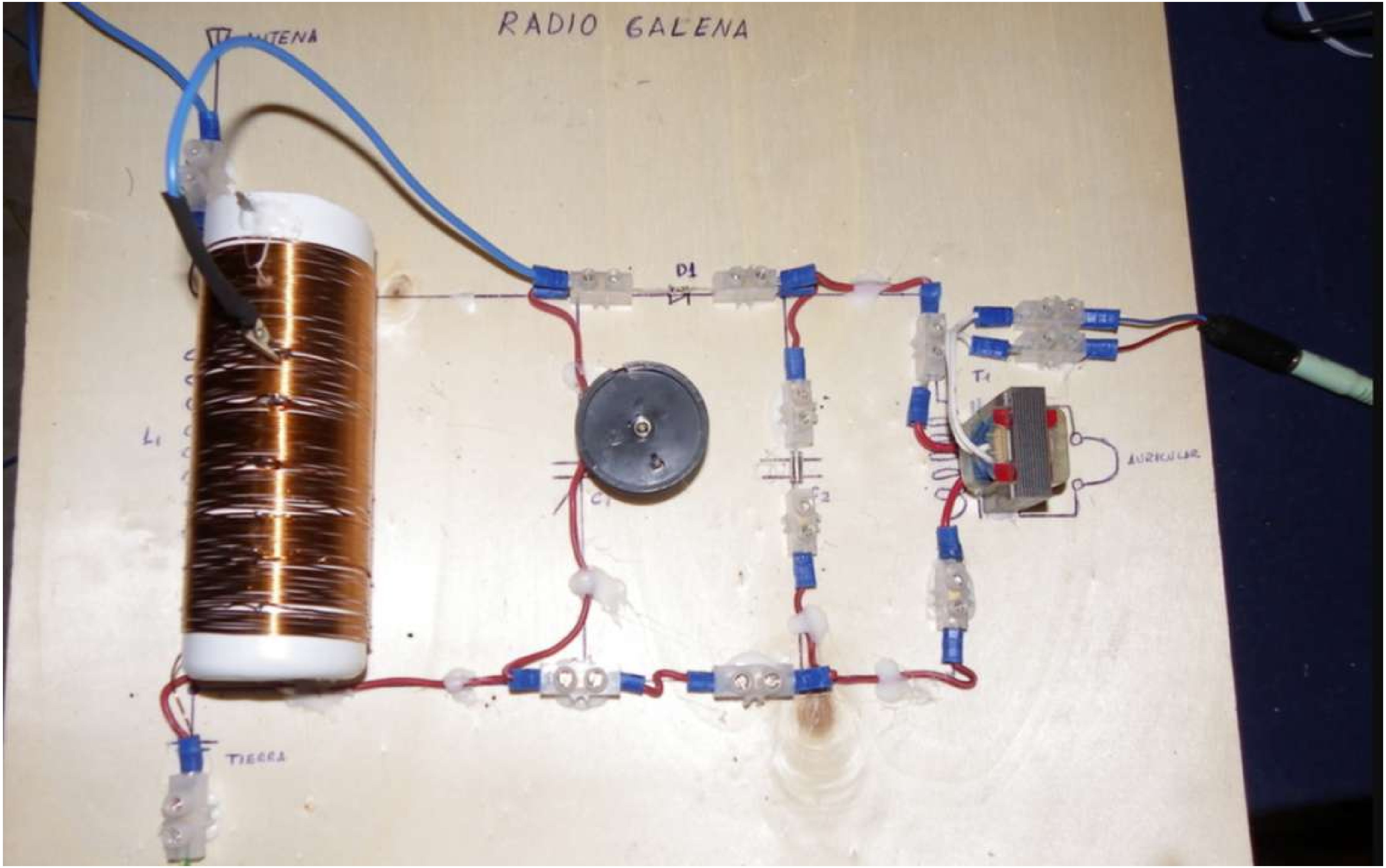
RADIO GALENA
ANTENA
D1
L1
C1
C2
T1
AURICULAR
TIERRA

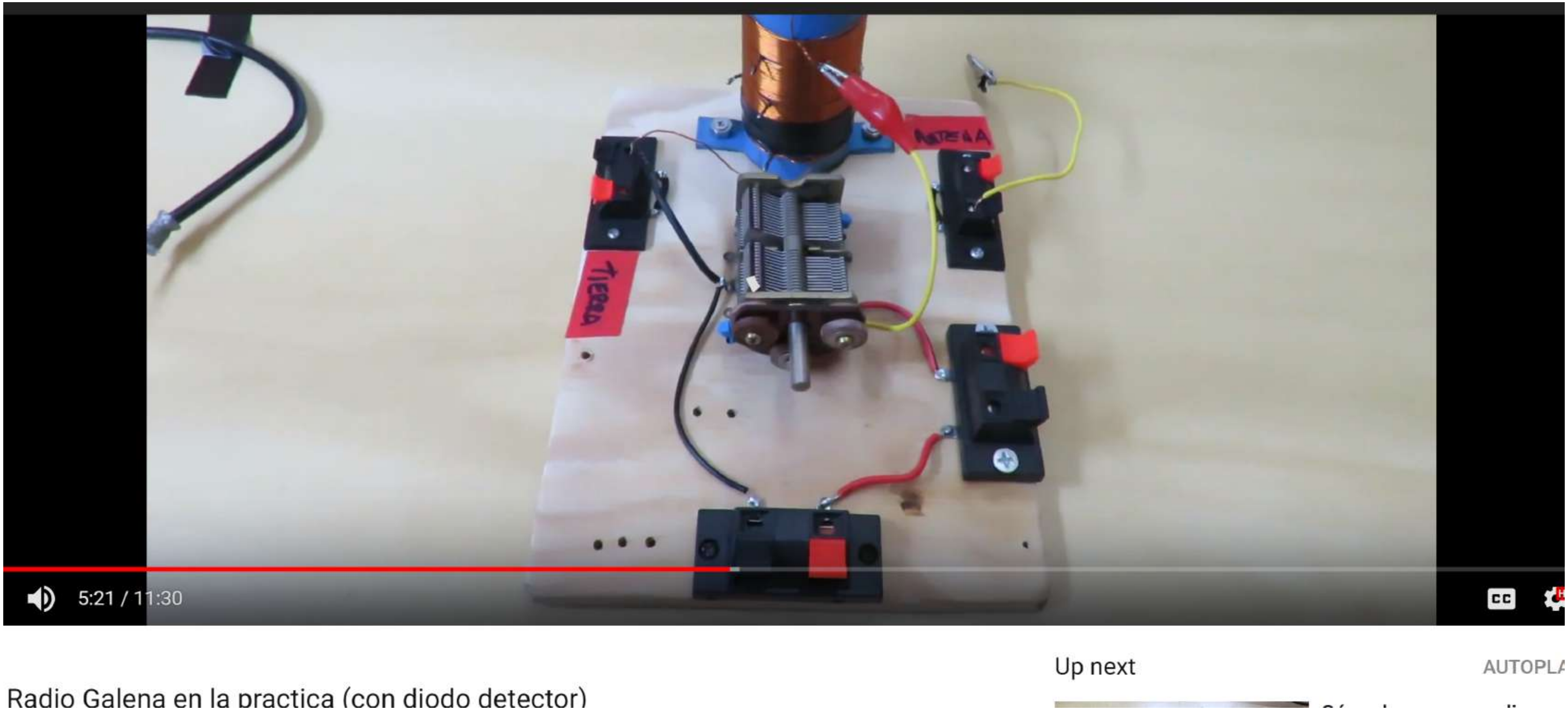
ANTENA
TIERRA
5:21 / 11:30
CC
Up next
AUTOPLA

Radio Galena en la practica (con diodo detector)

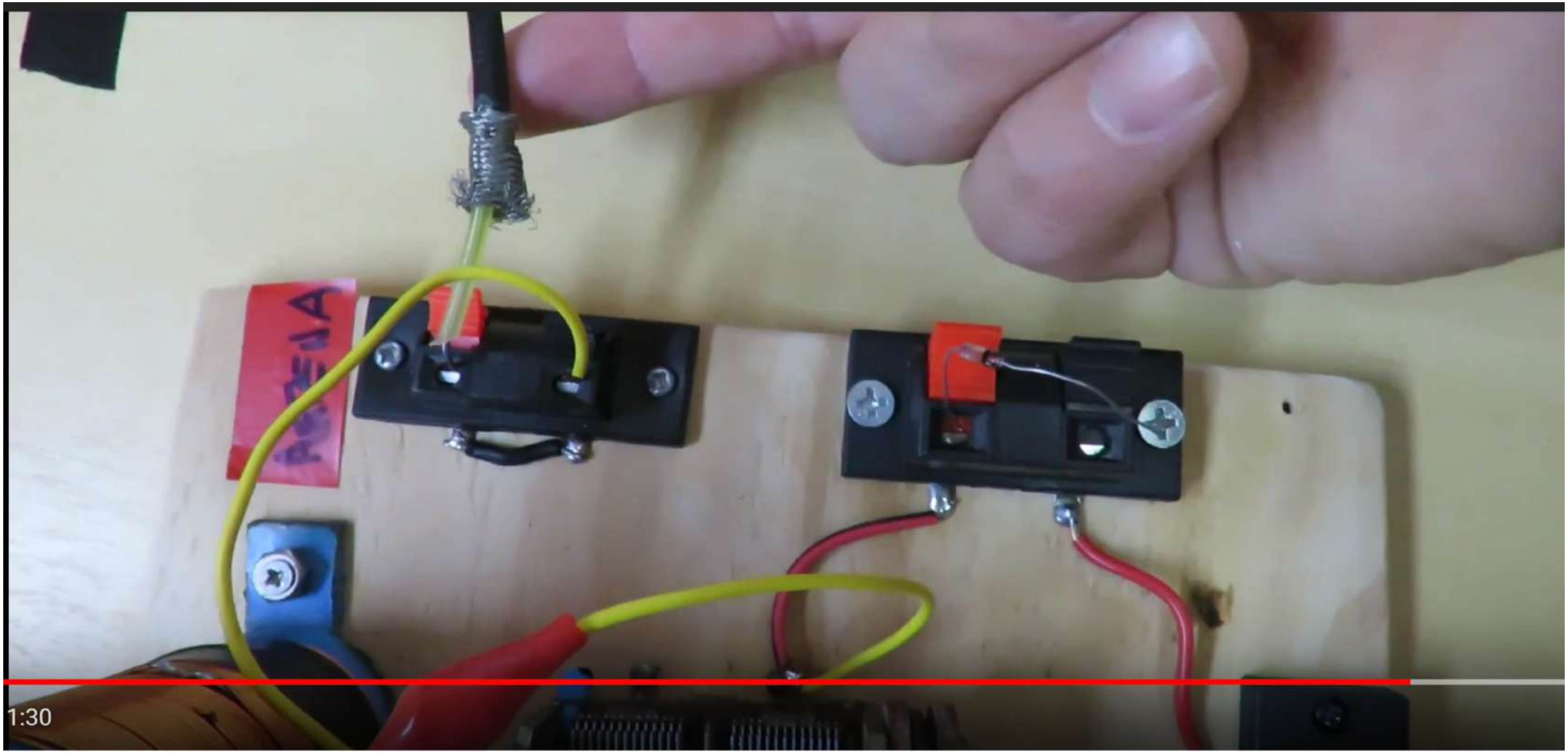
ANTENA
1:30

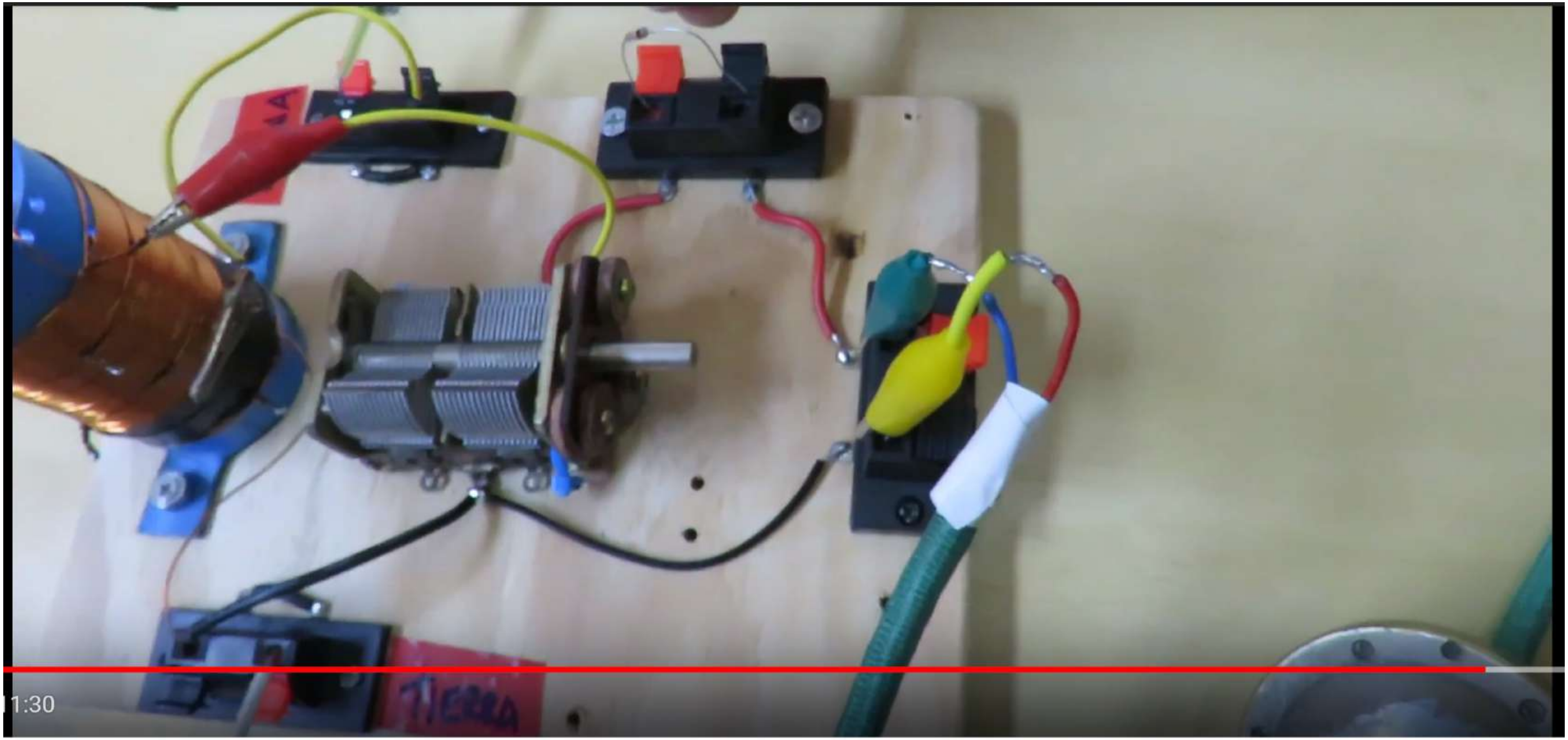
NA
TIERRA
1:30

ANTENA
11:30

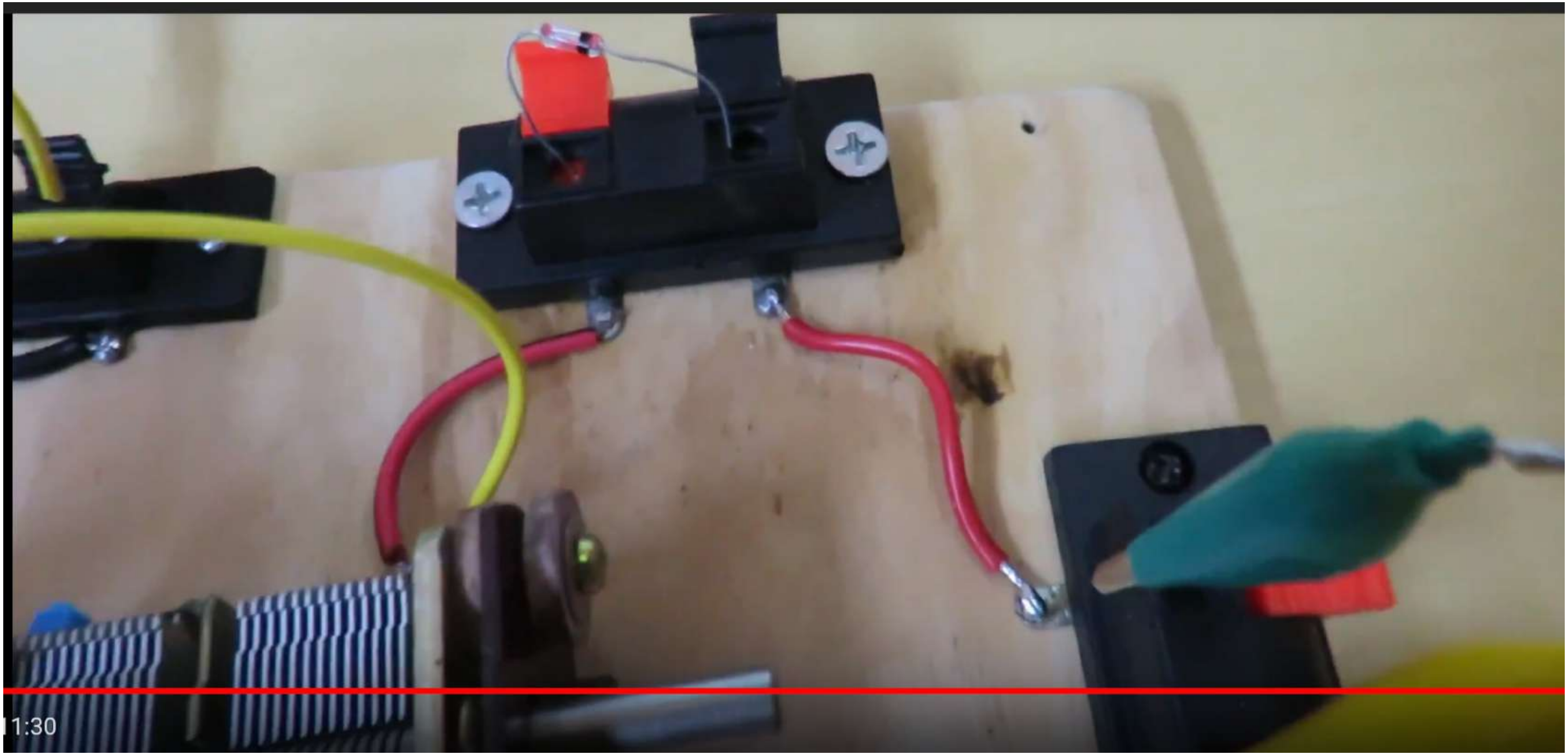
1:30

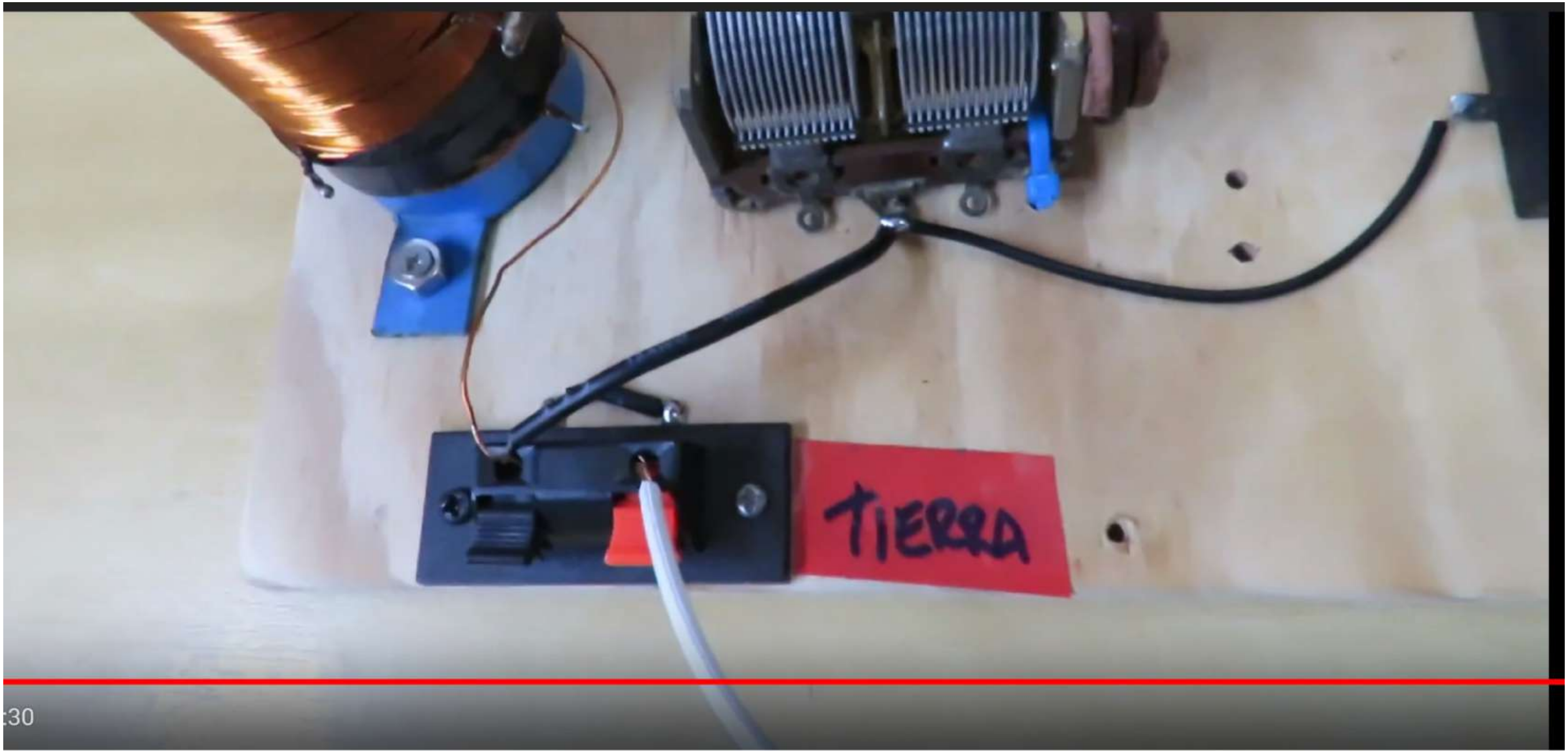
TIERRA
:30

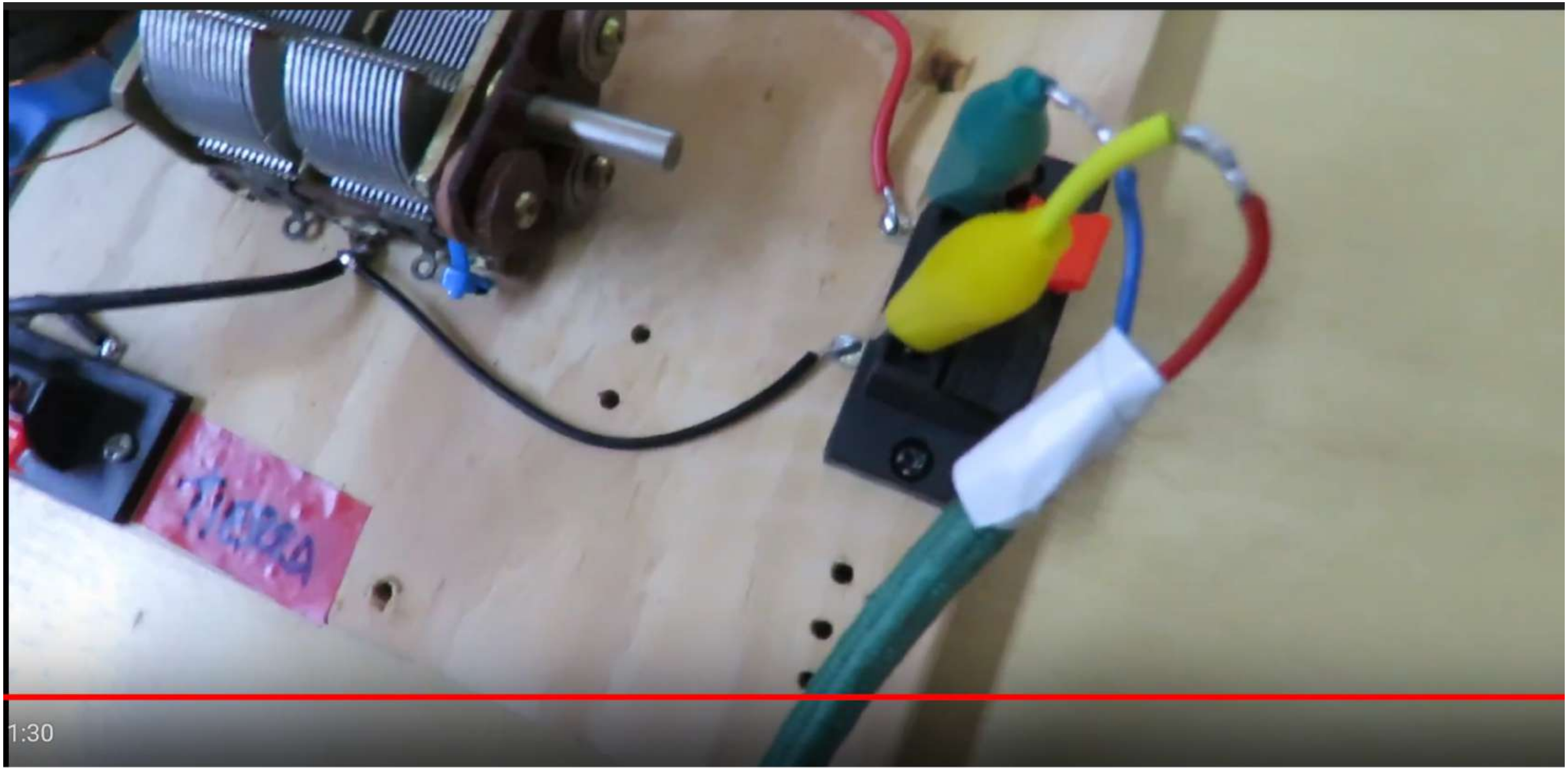
TIERRA
1:30

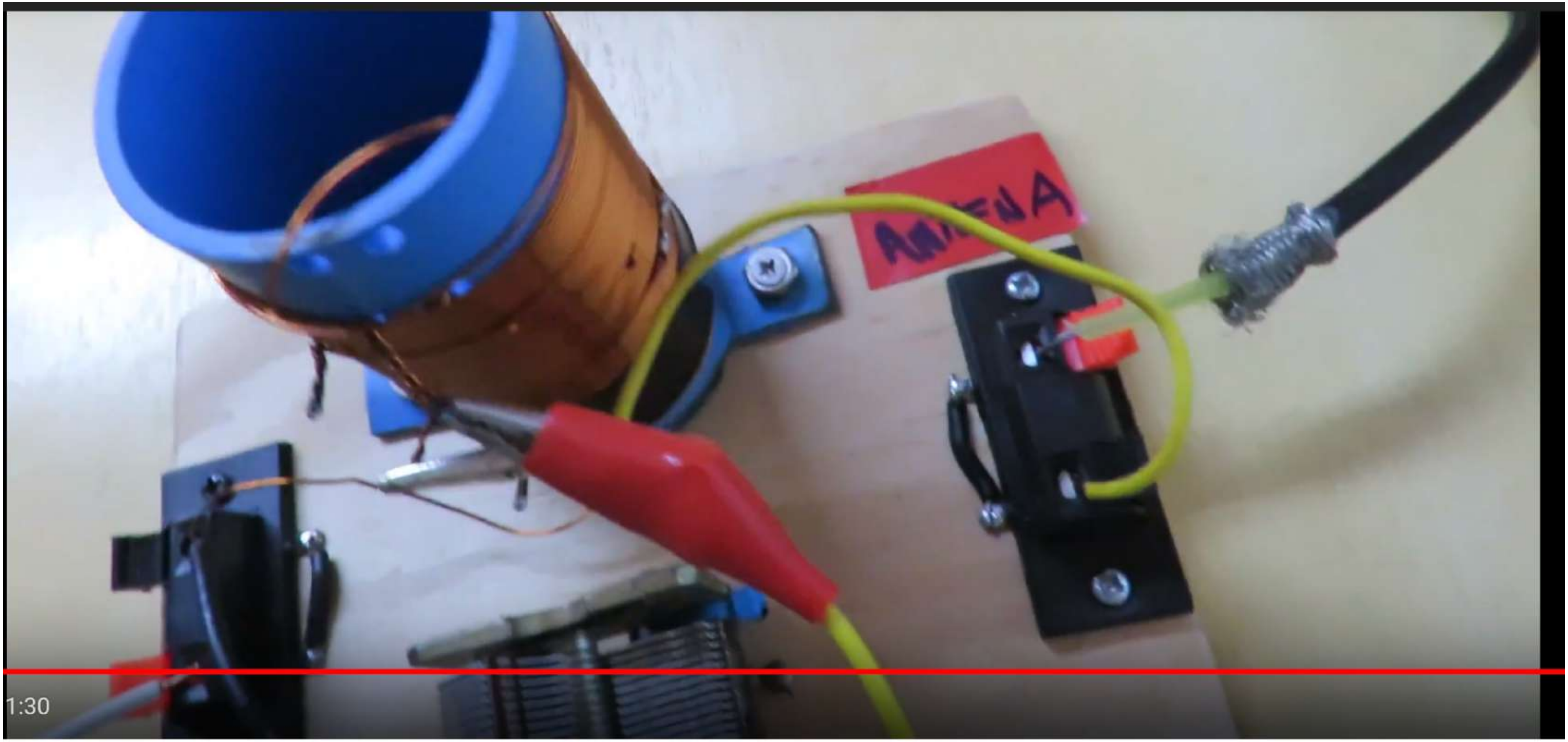
ANTENA
1:30

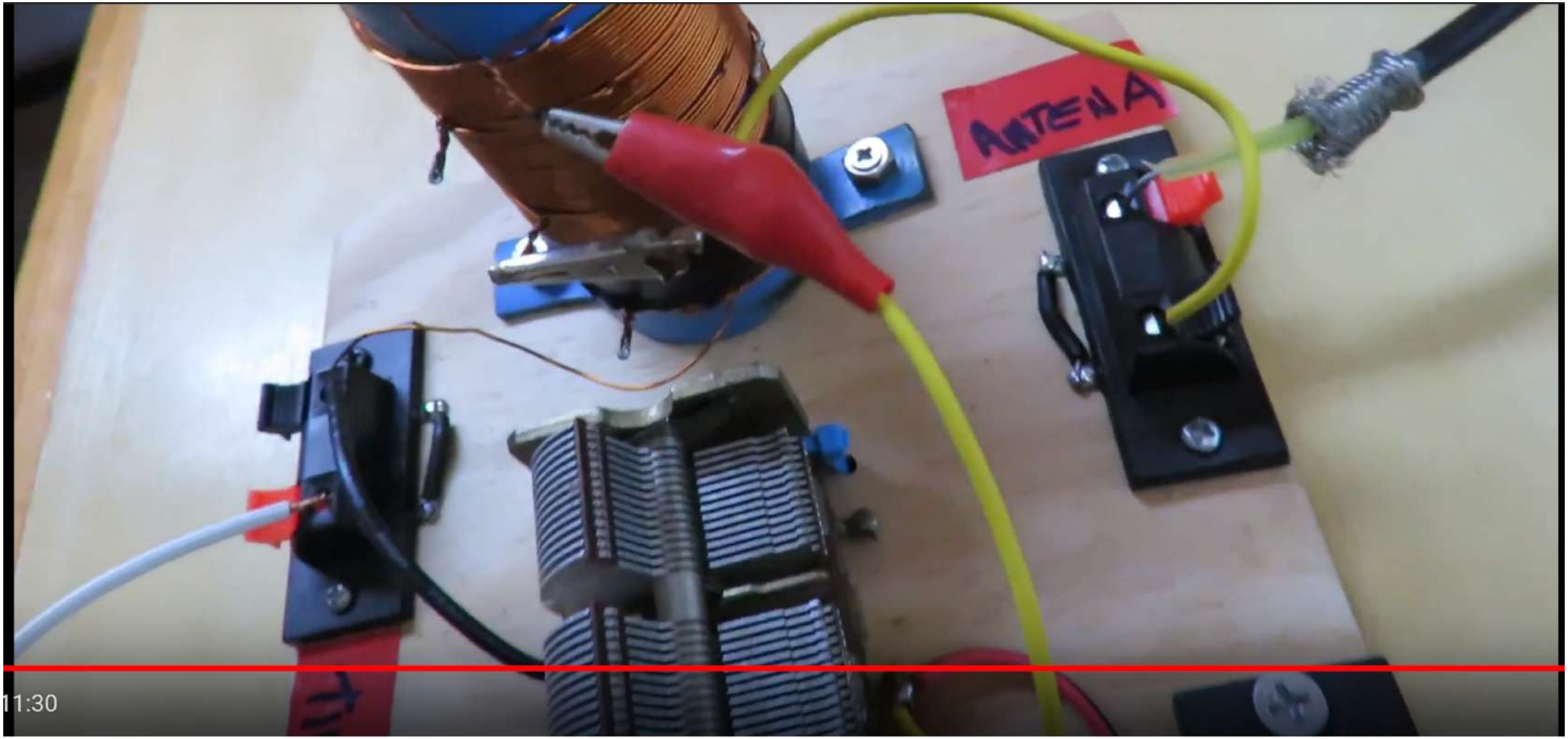
ANTENA
11:30

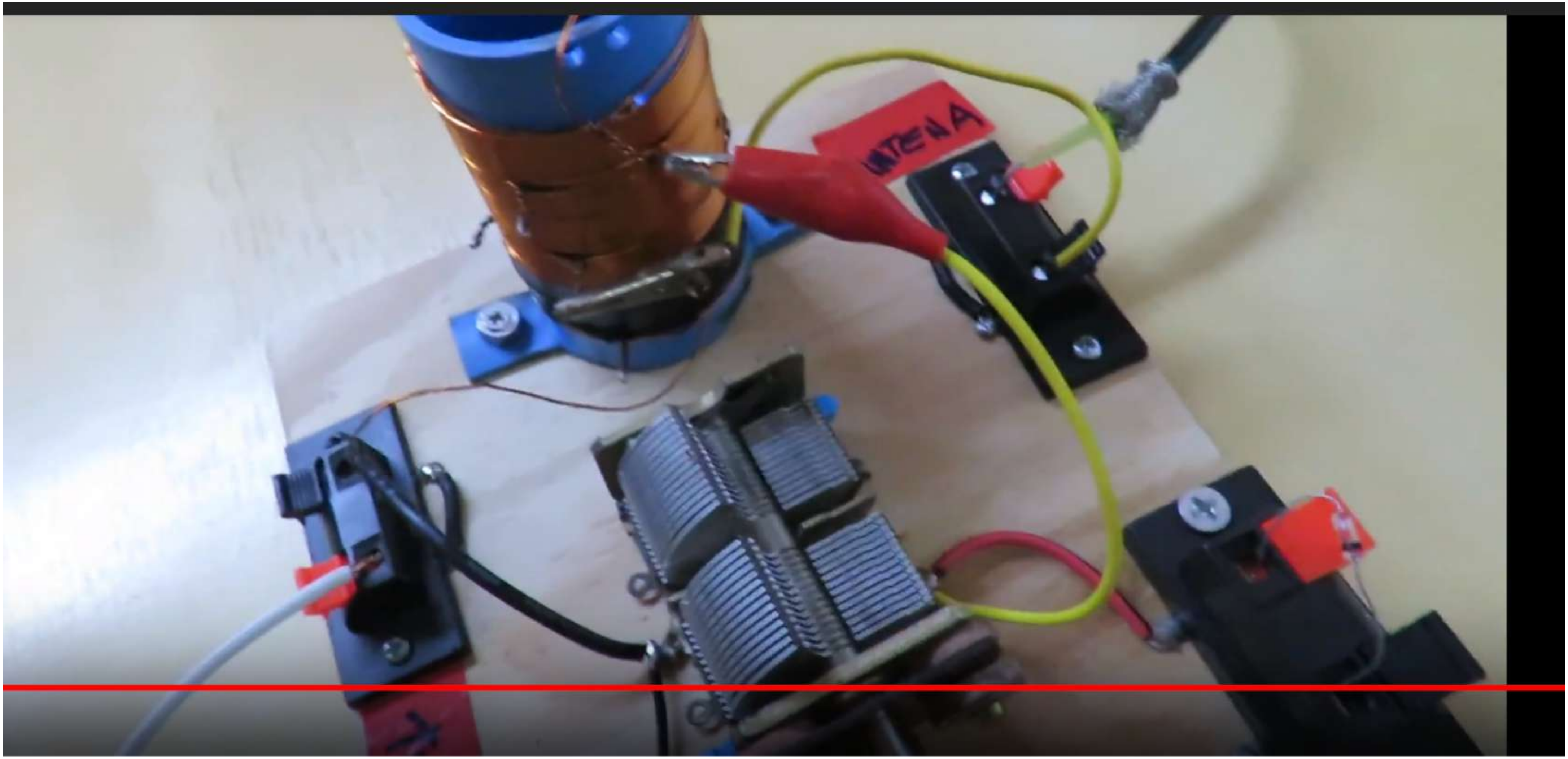
NTENA

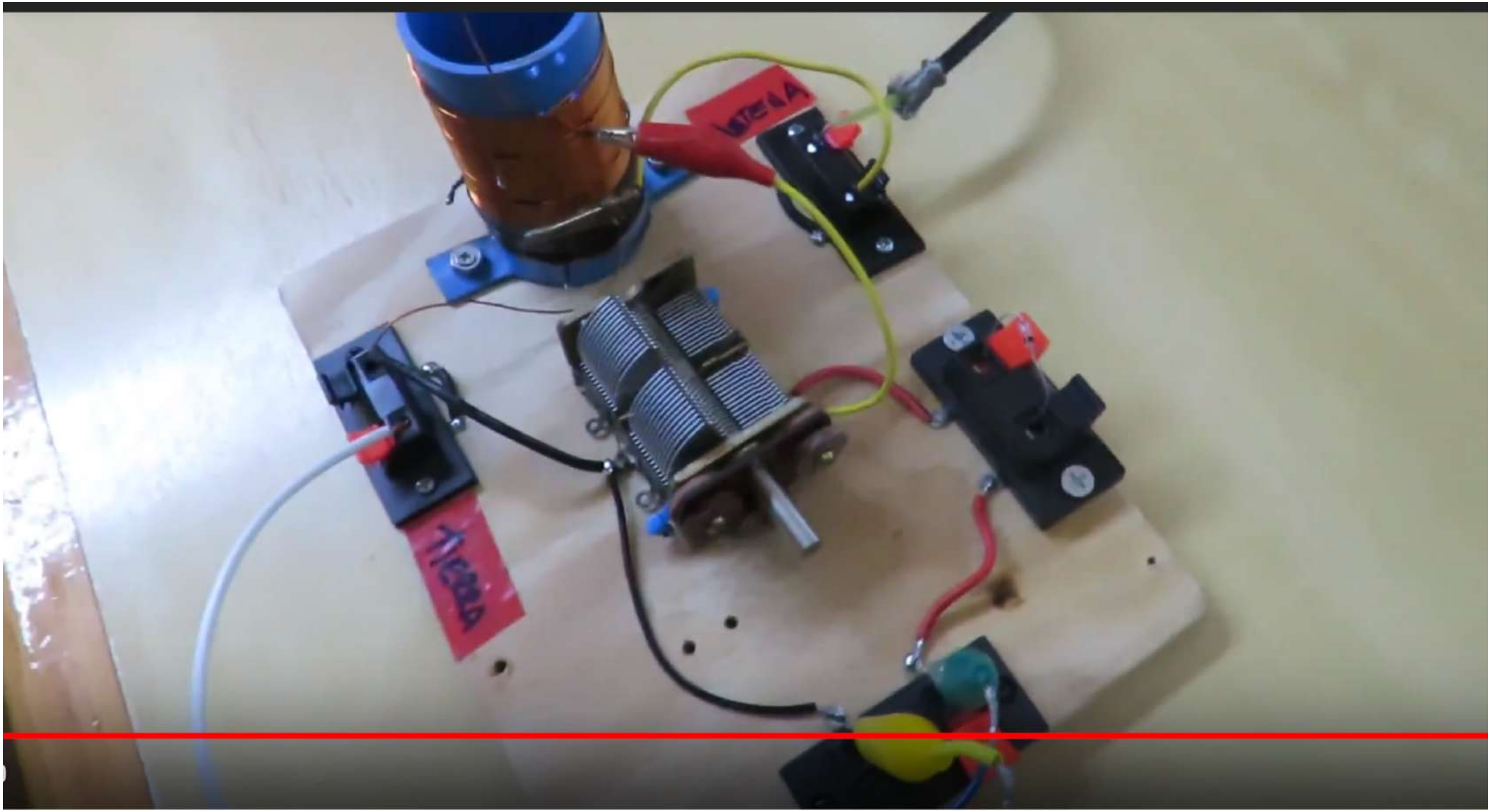
ANTENA
TIERRA

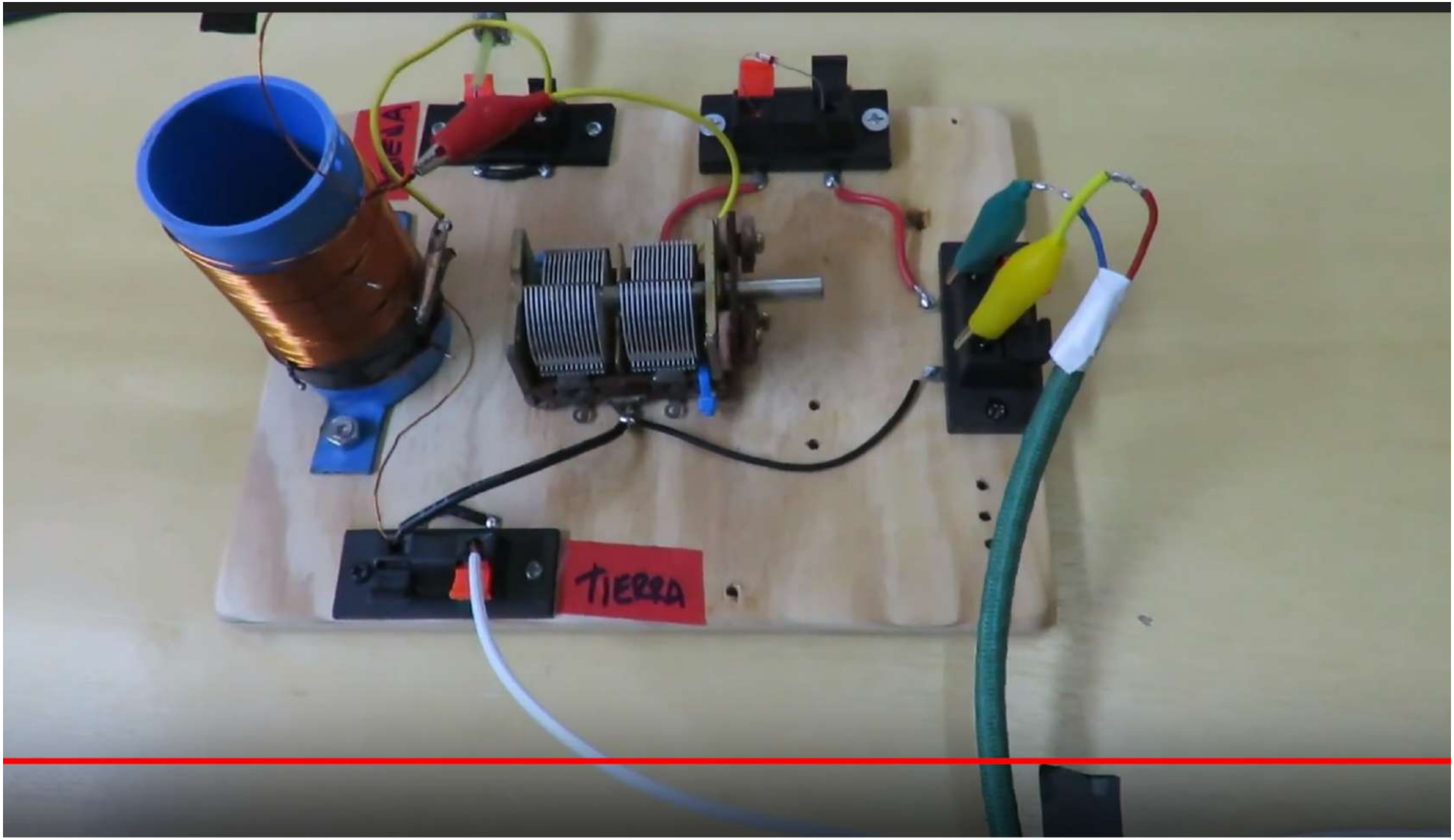
TIERRA

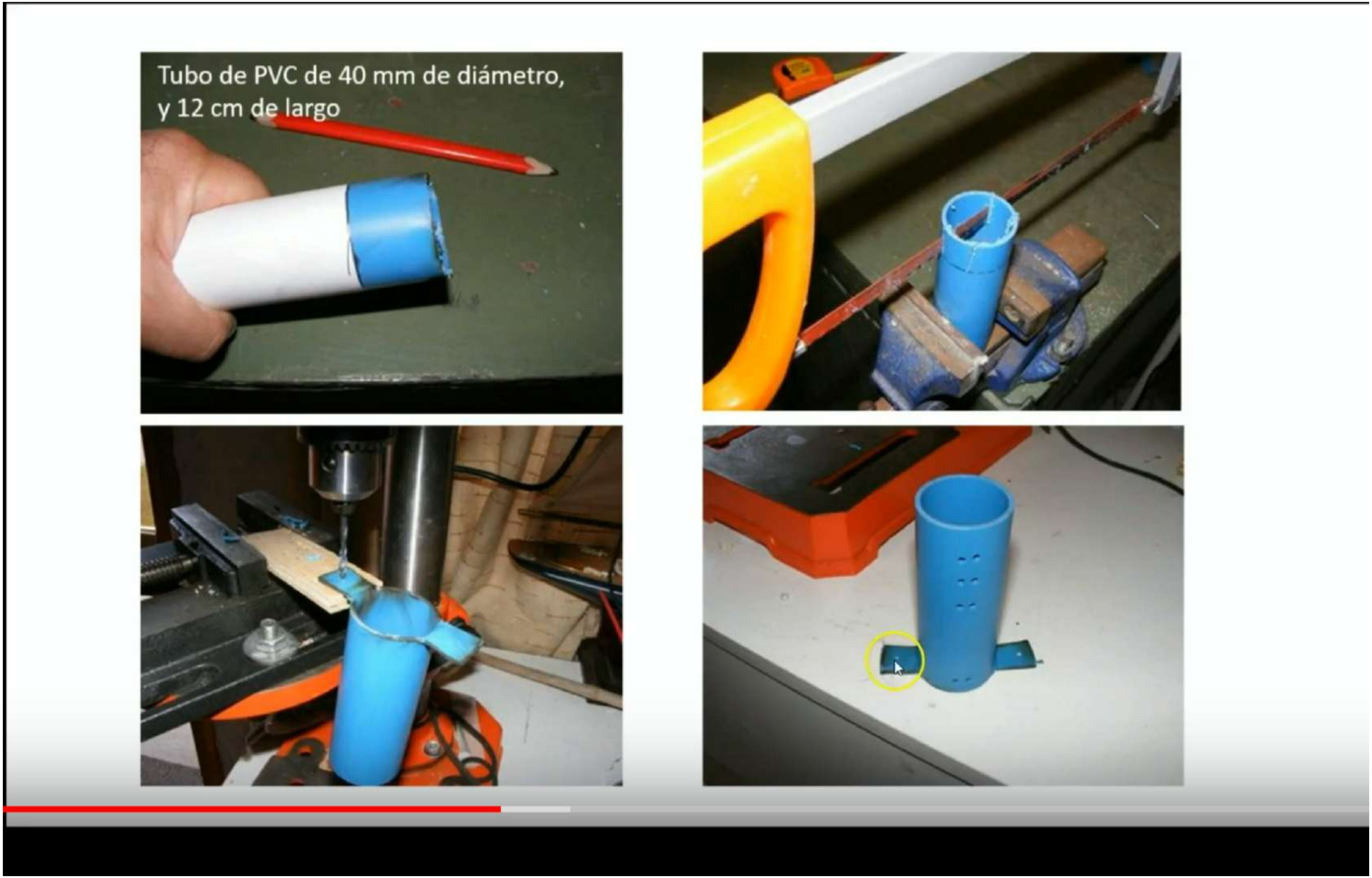

Tubo de PVC de 40 mm de diámetro, y 12 cm de largo

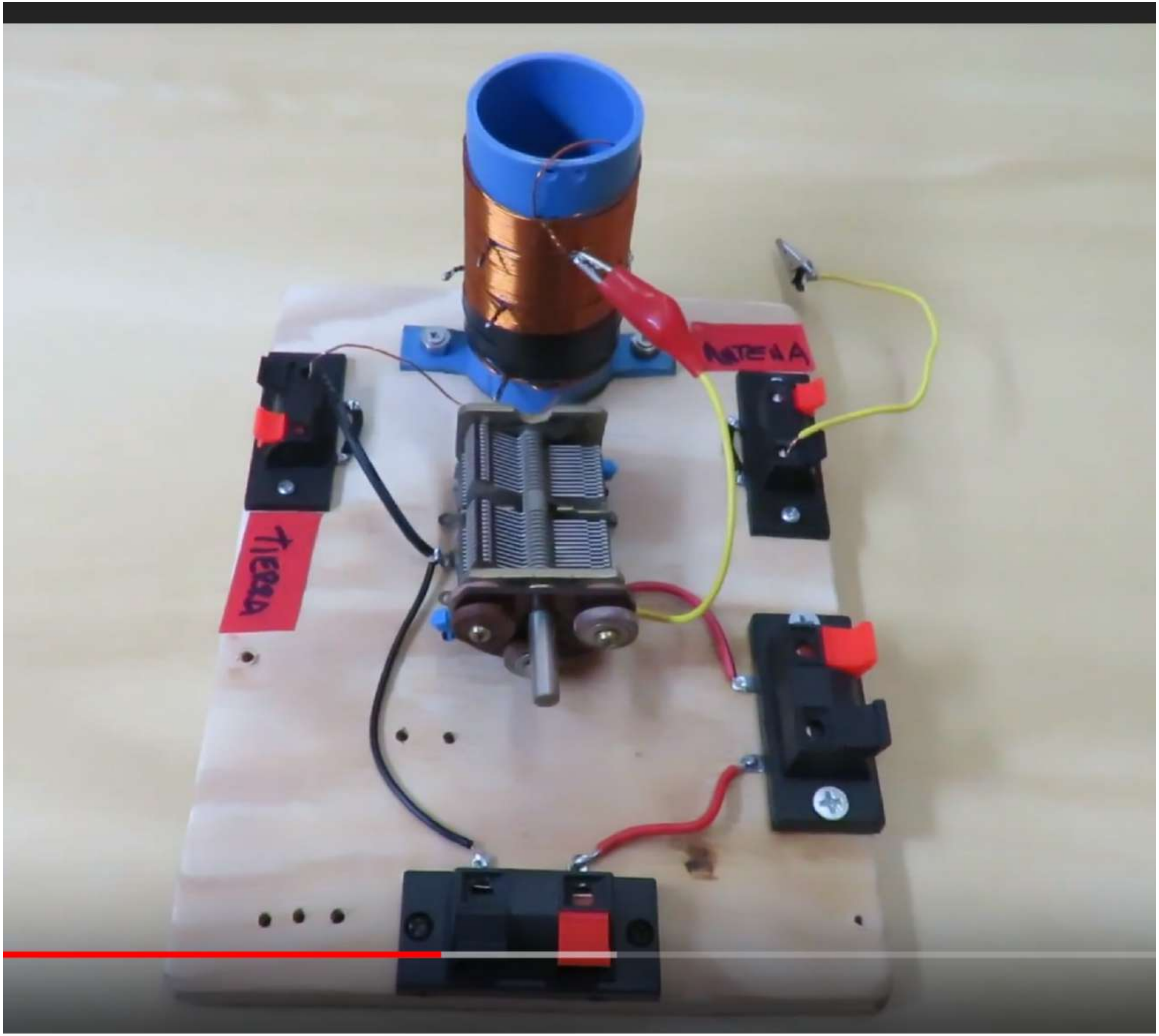
ANTENA
TIERRA

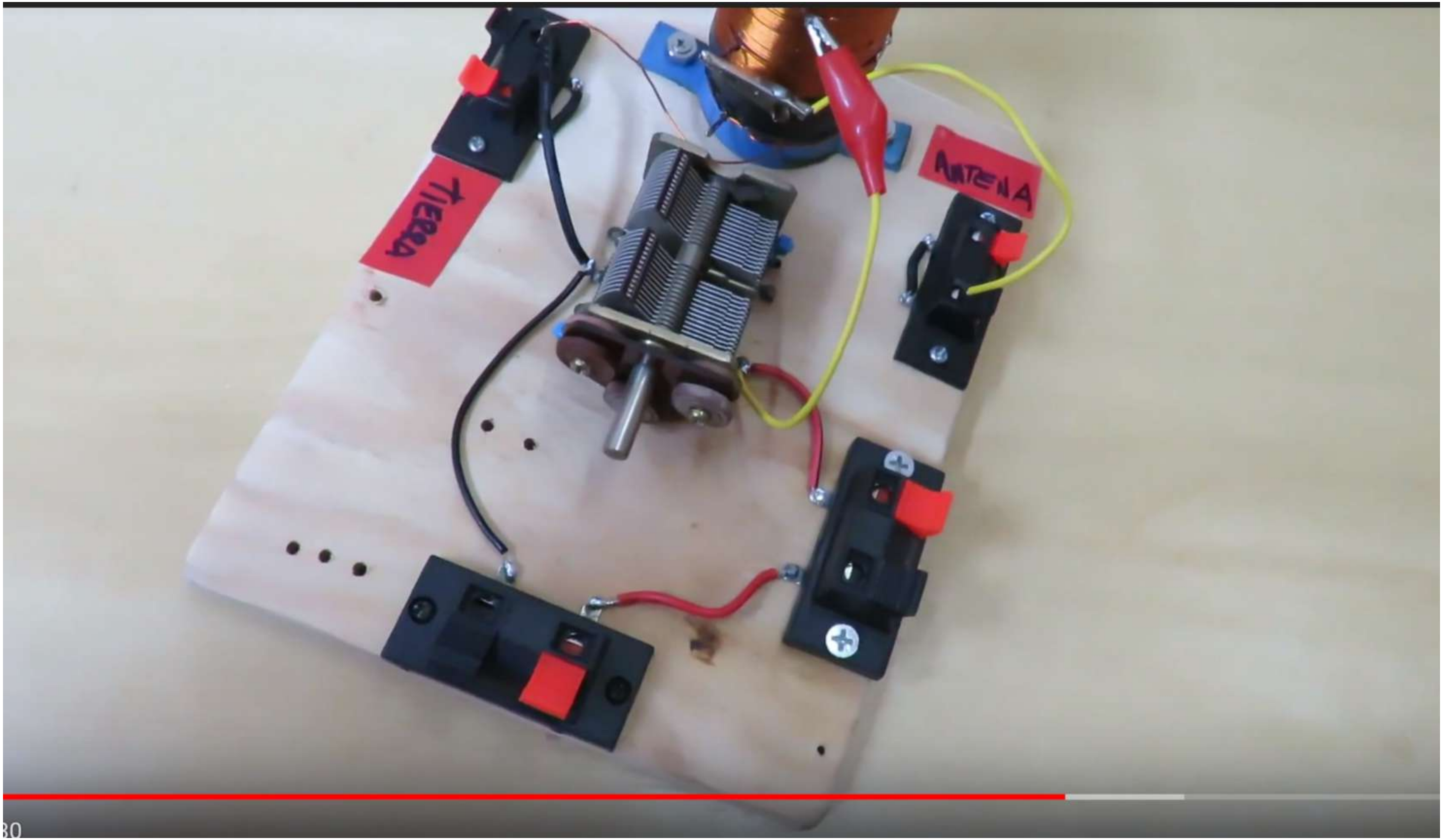
TIERRA
ANTENA

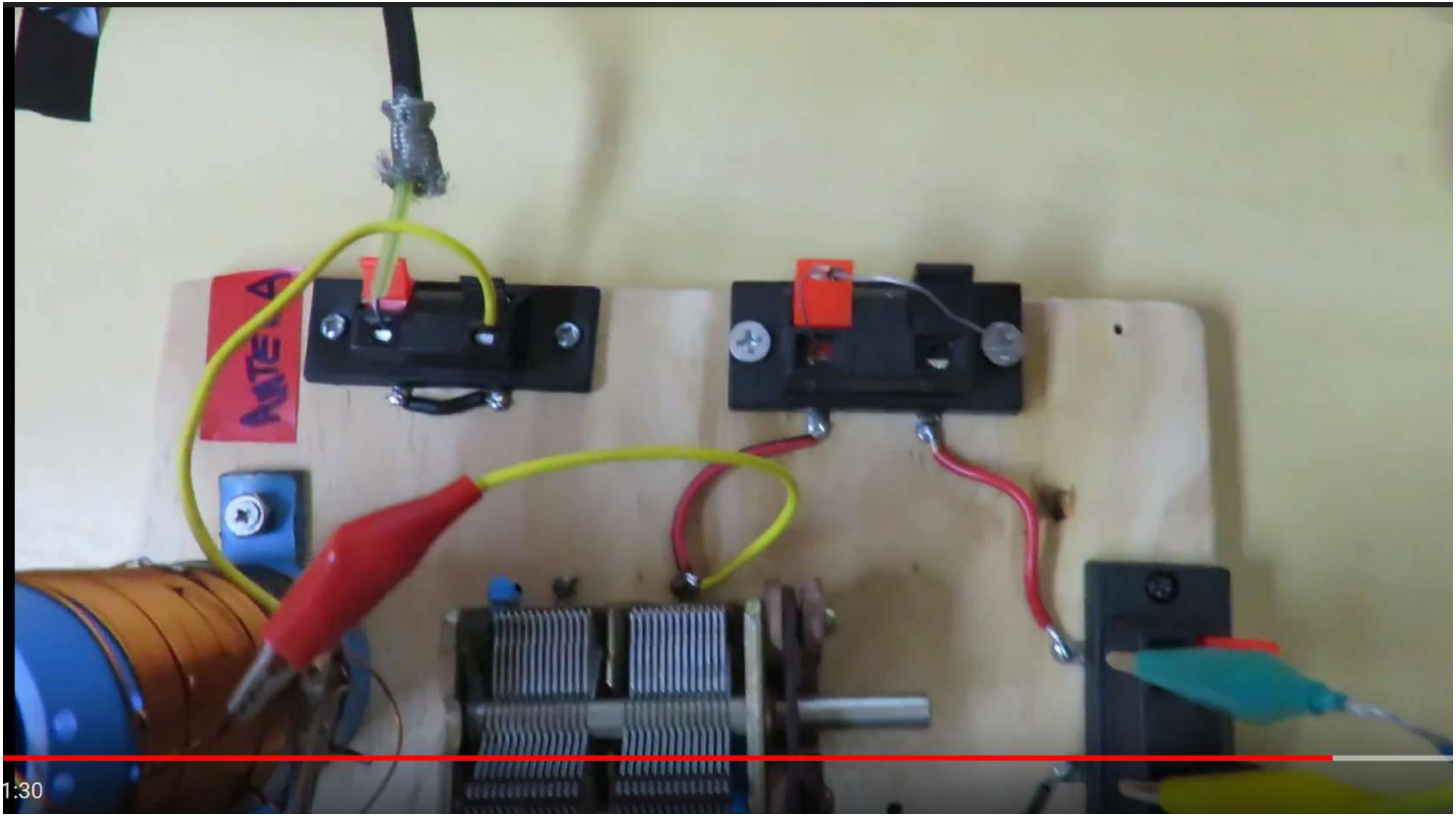
ANTE A
1:30

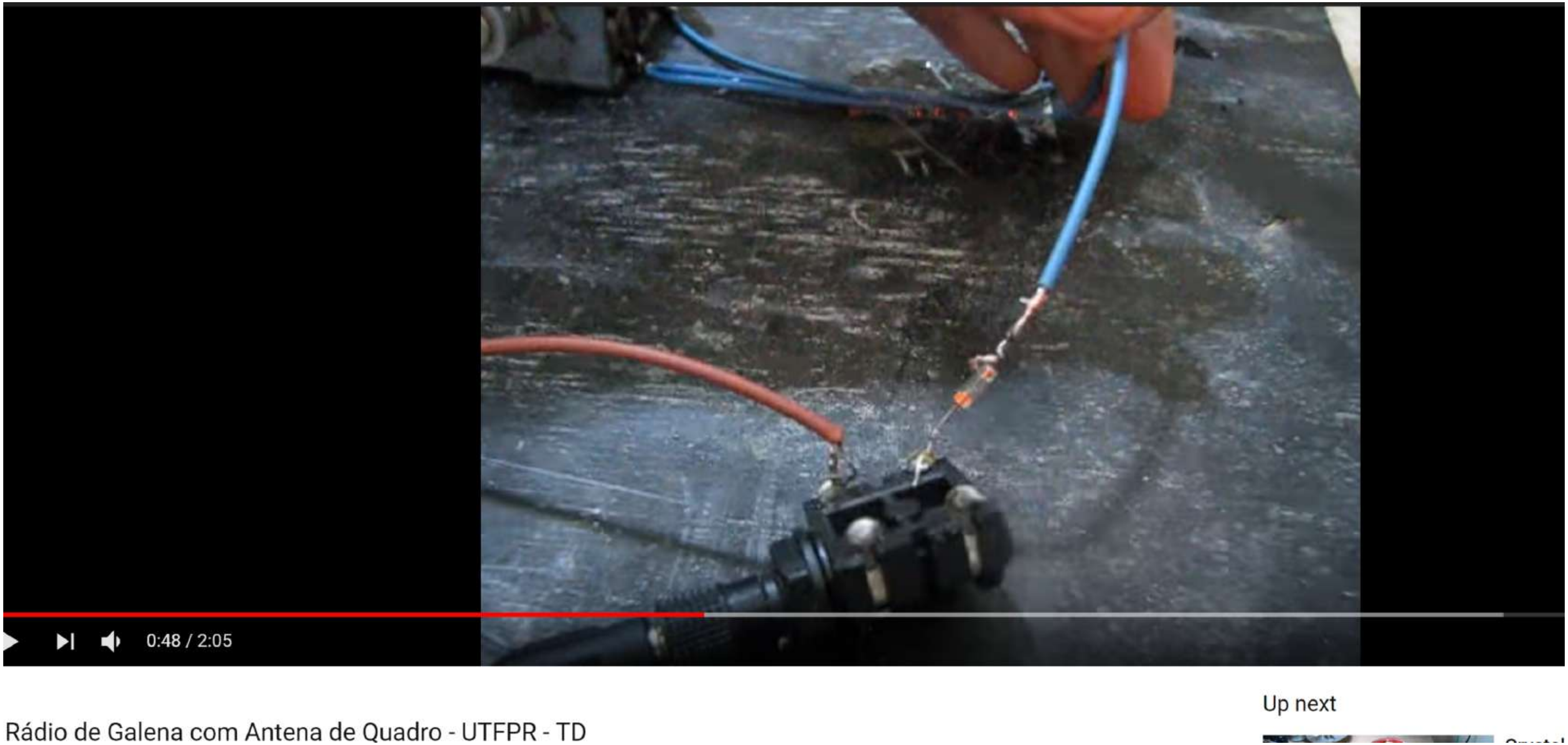
0:48 / 2:05
Up next

Rádio de Galena com Antena de Quadro - UTFPR - TD

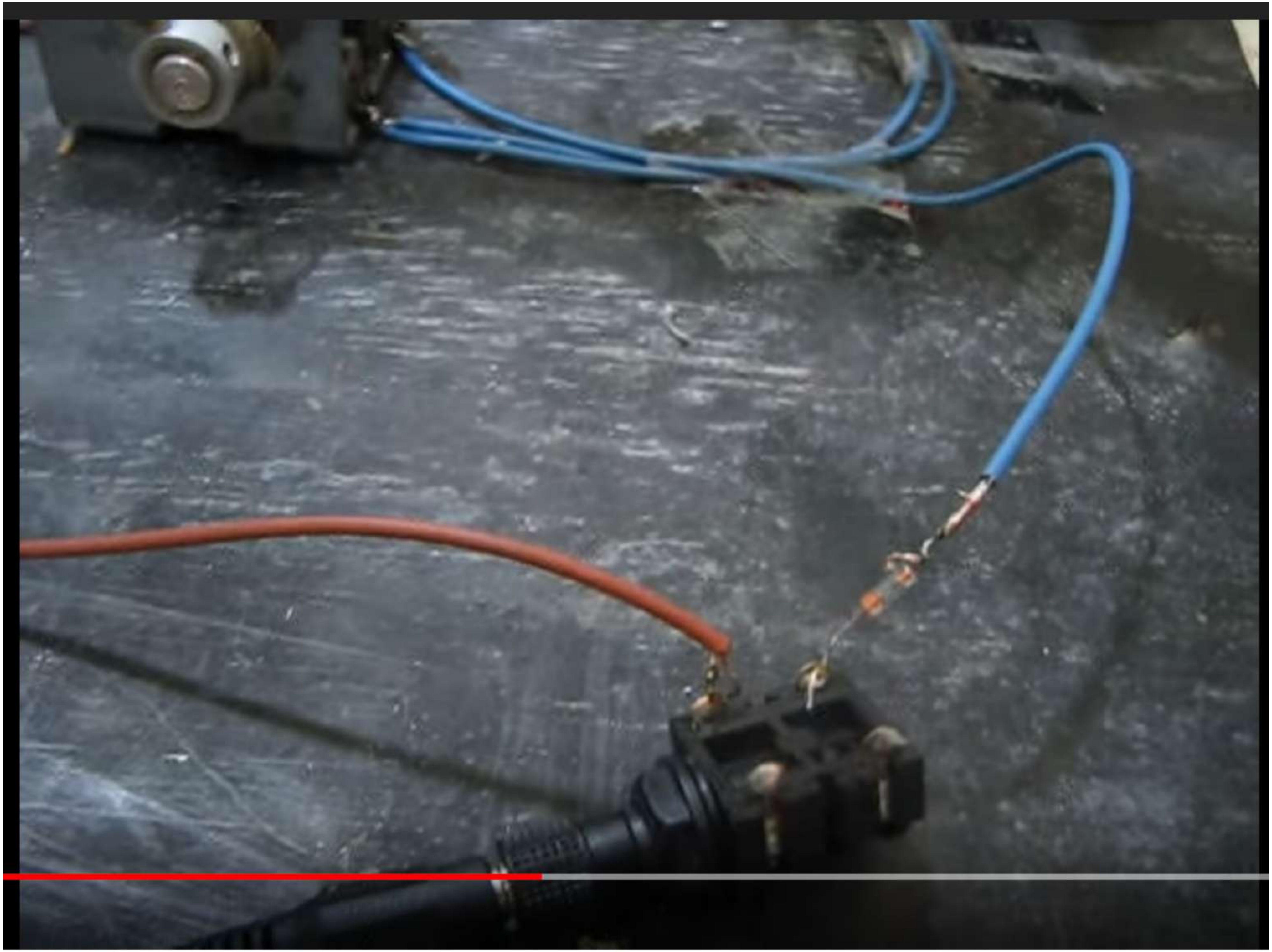

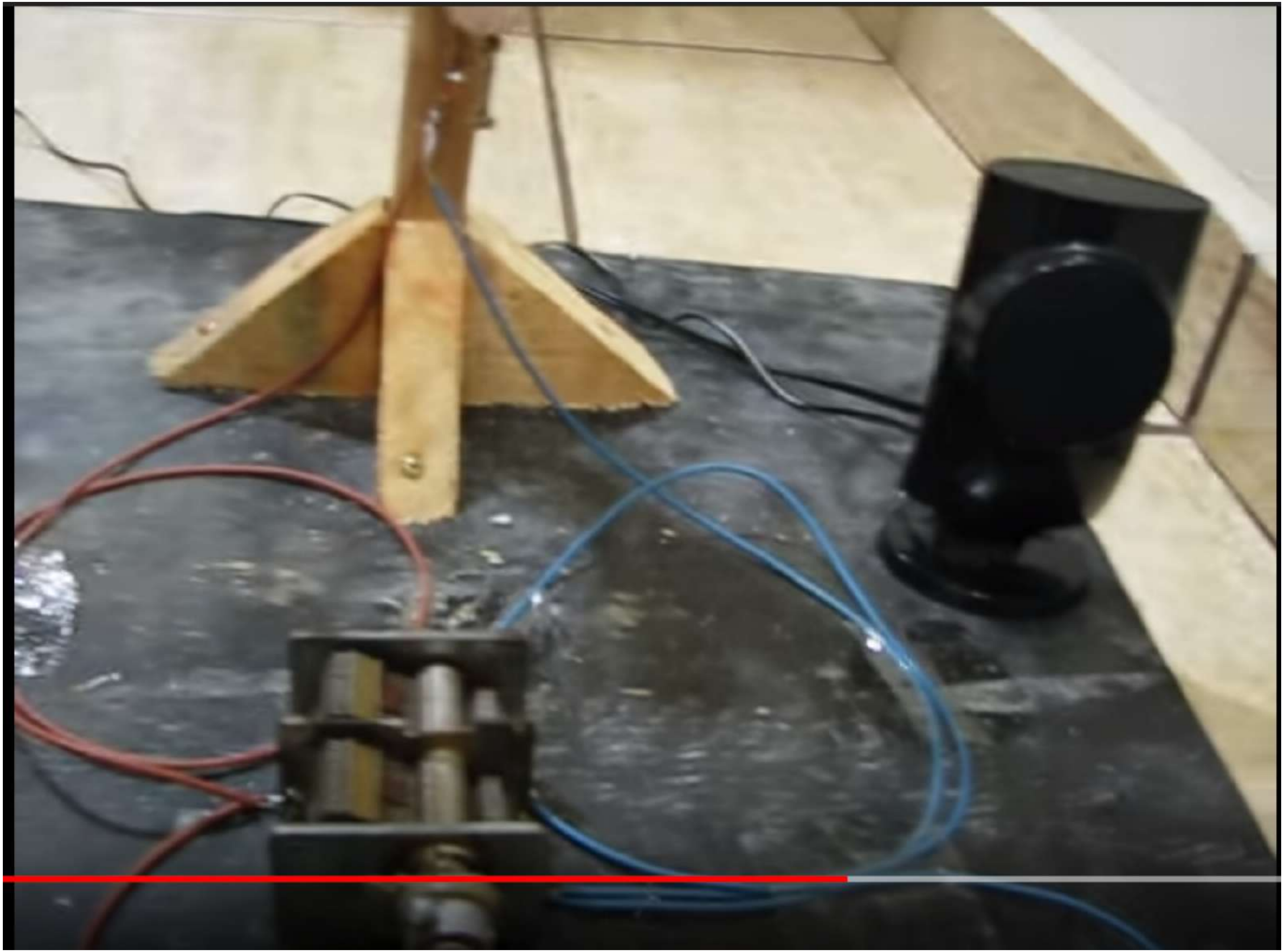

BINGR R.
BRIEL T.

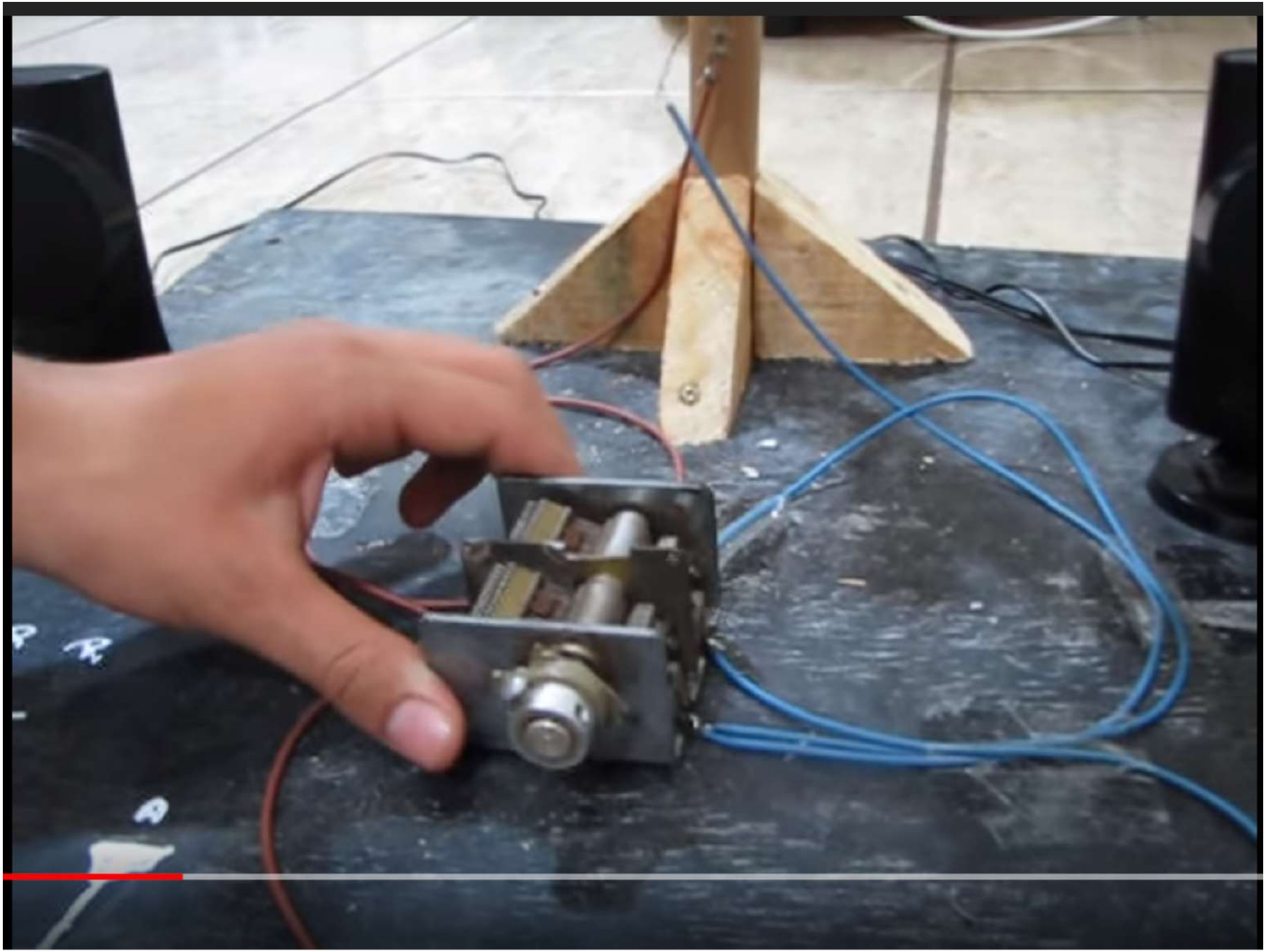
R R
A

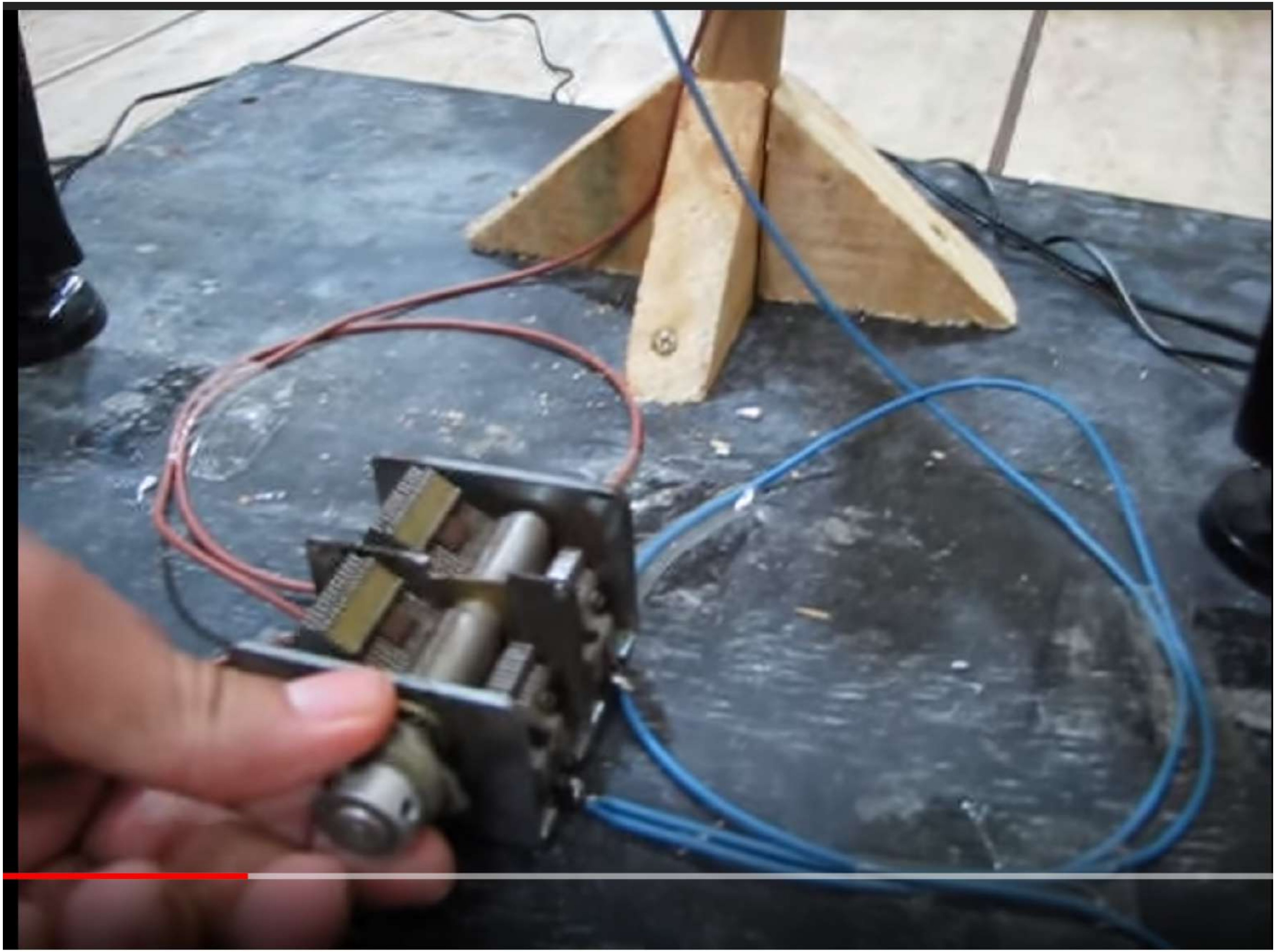

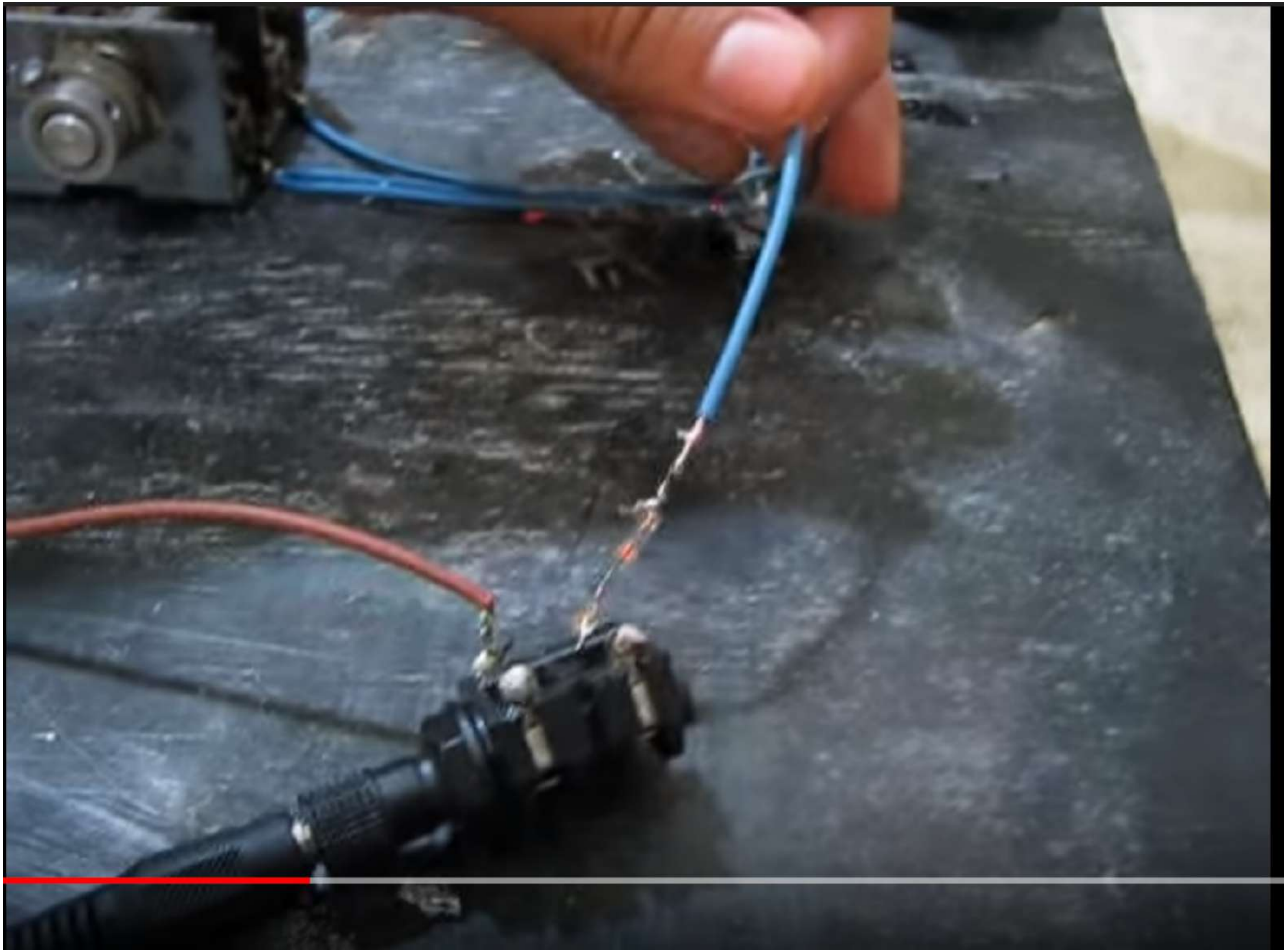

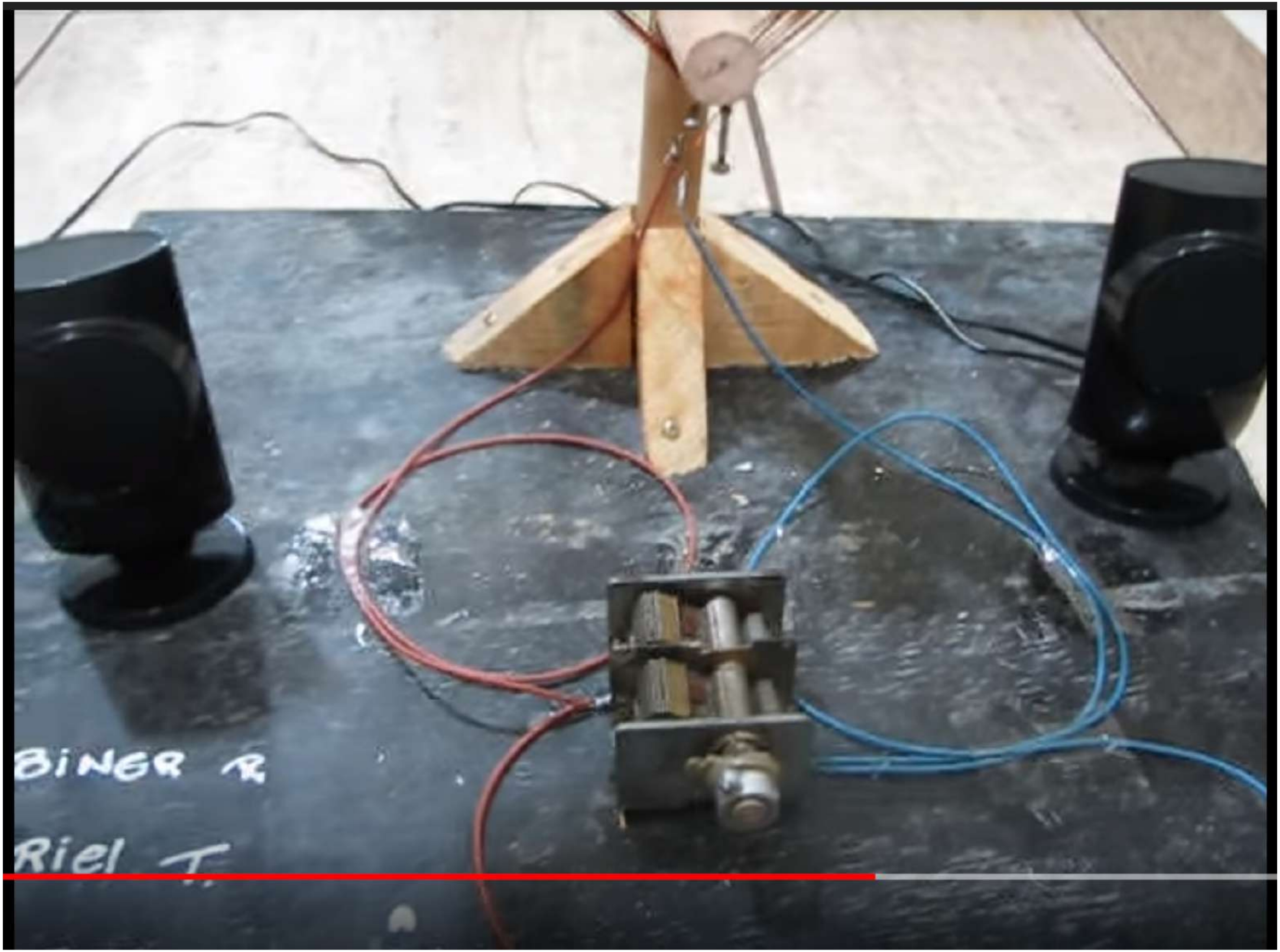
BiNGR R.
Riel T.

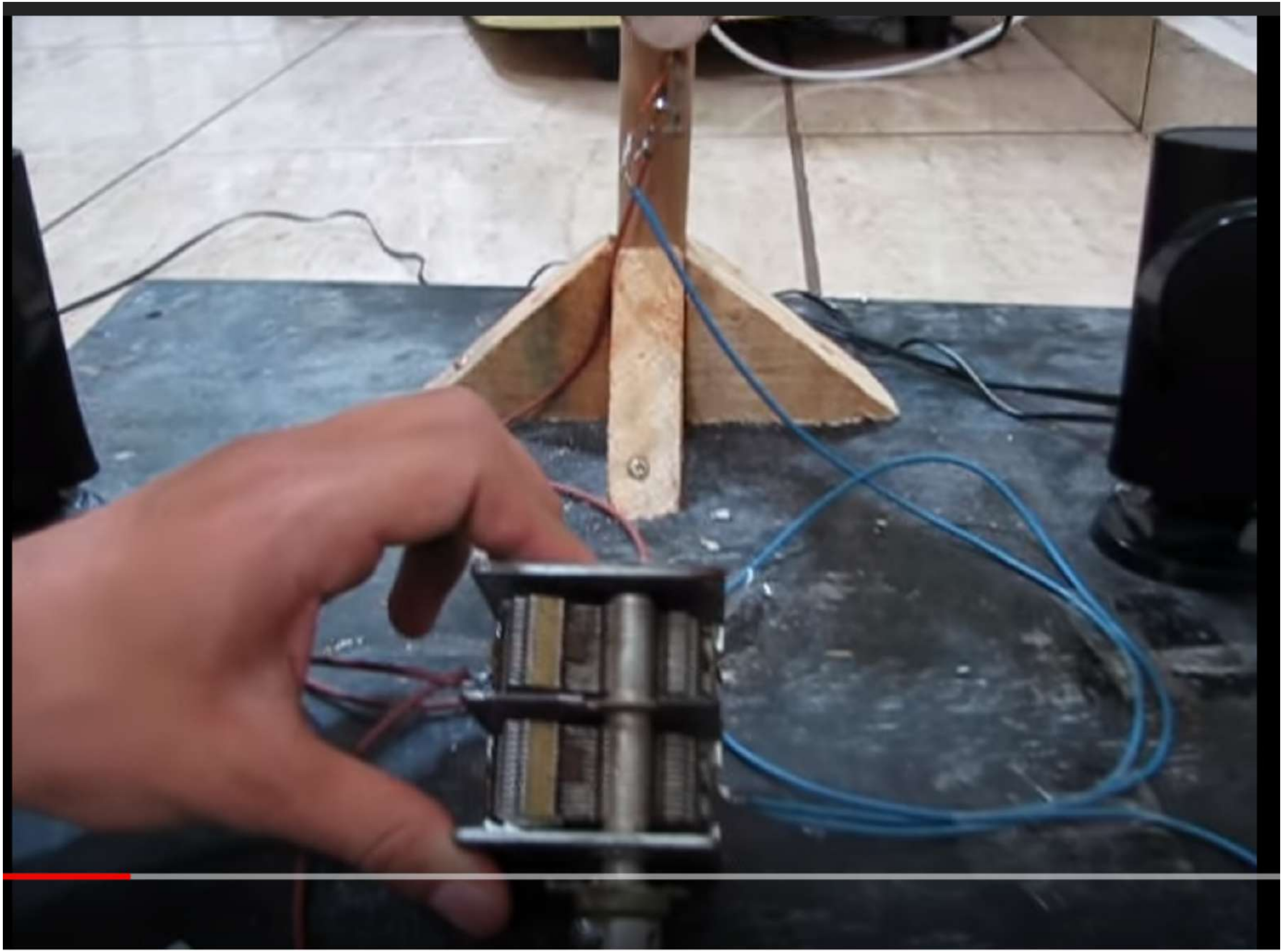

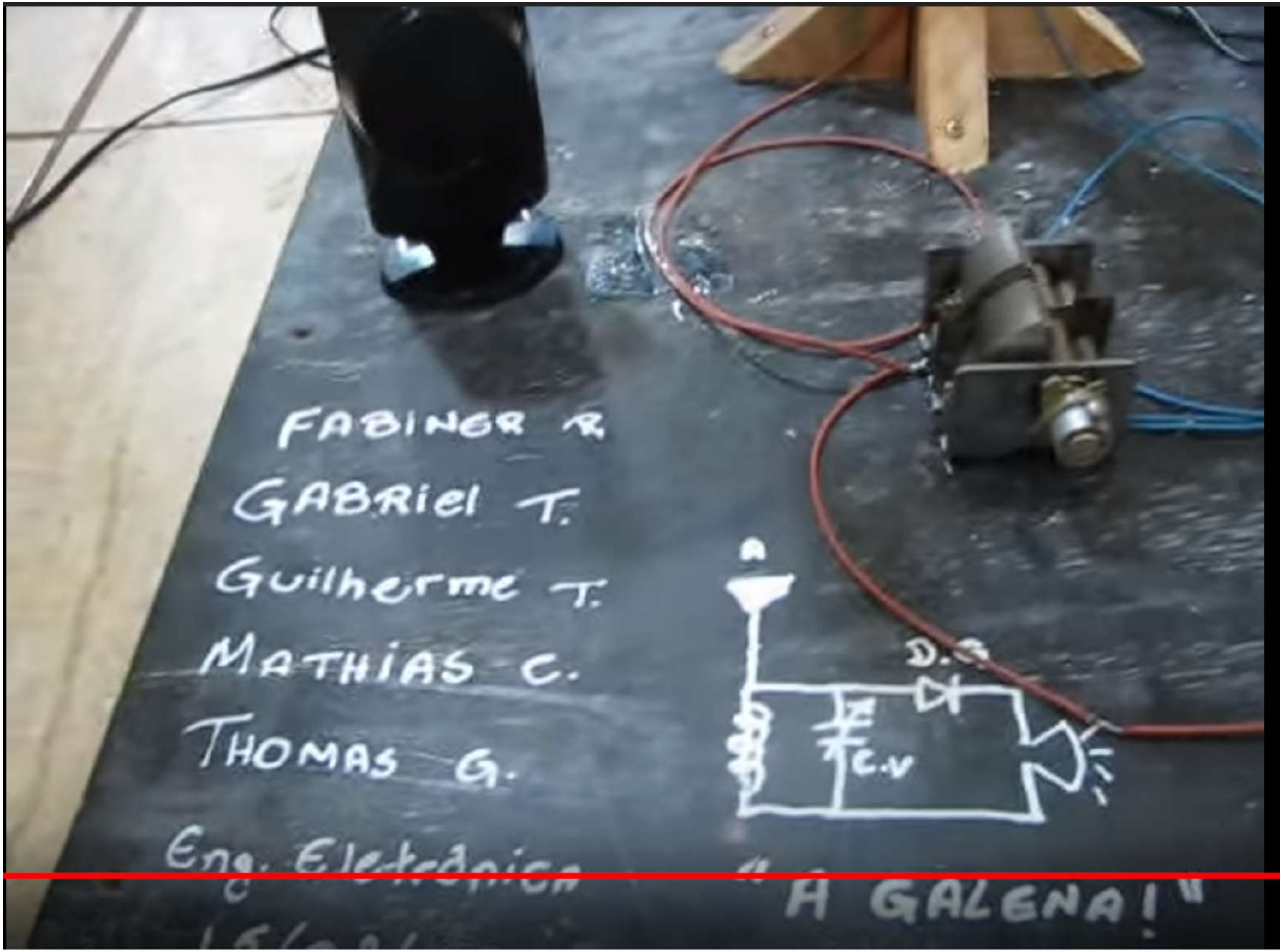
FABINER R.
GABRiel T.
Guilherme T.
MATHIAS C.
THOMAS G.
Eng. Eletrônica
D.G
C.V
"A GALENA!"

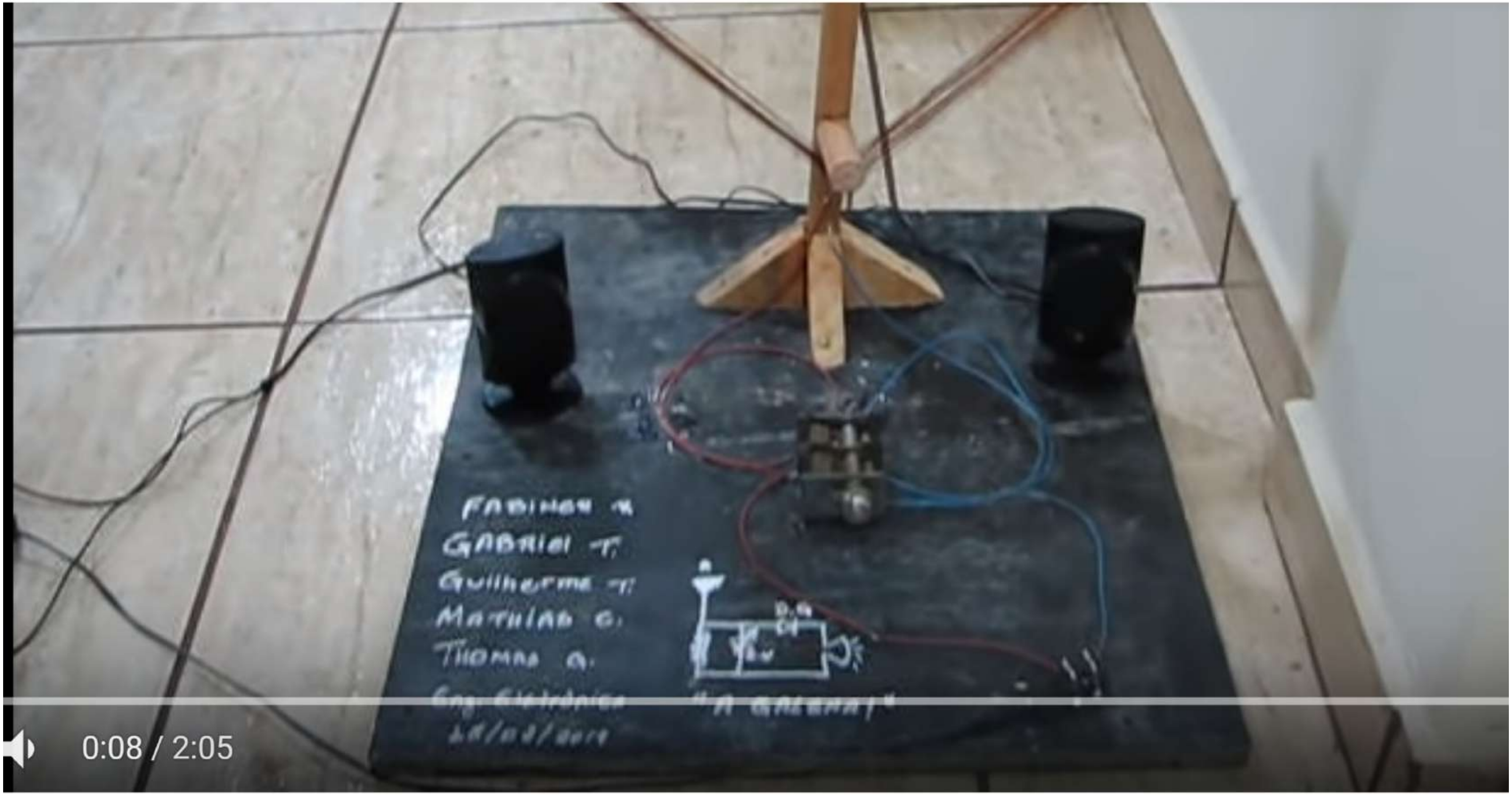
FABINHO R.
GABRIEL T.
Guilherme T.
MATHIAS C.
THOMAS A.
Eng. Eletrônica
"A GALENA!"
0:08 / 2:05

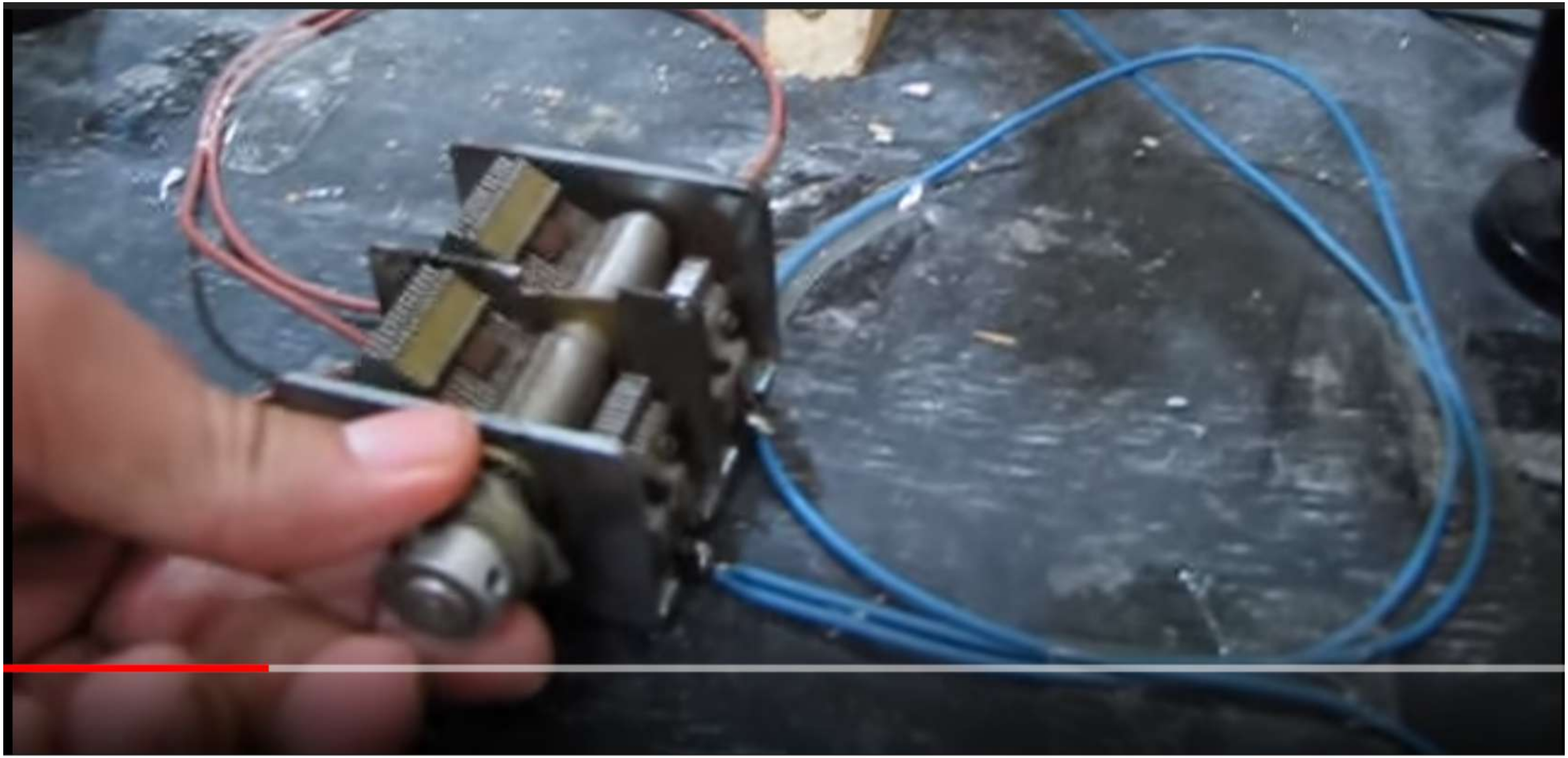

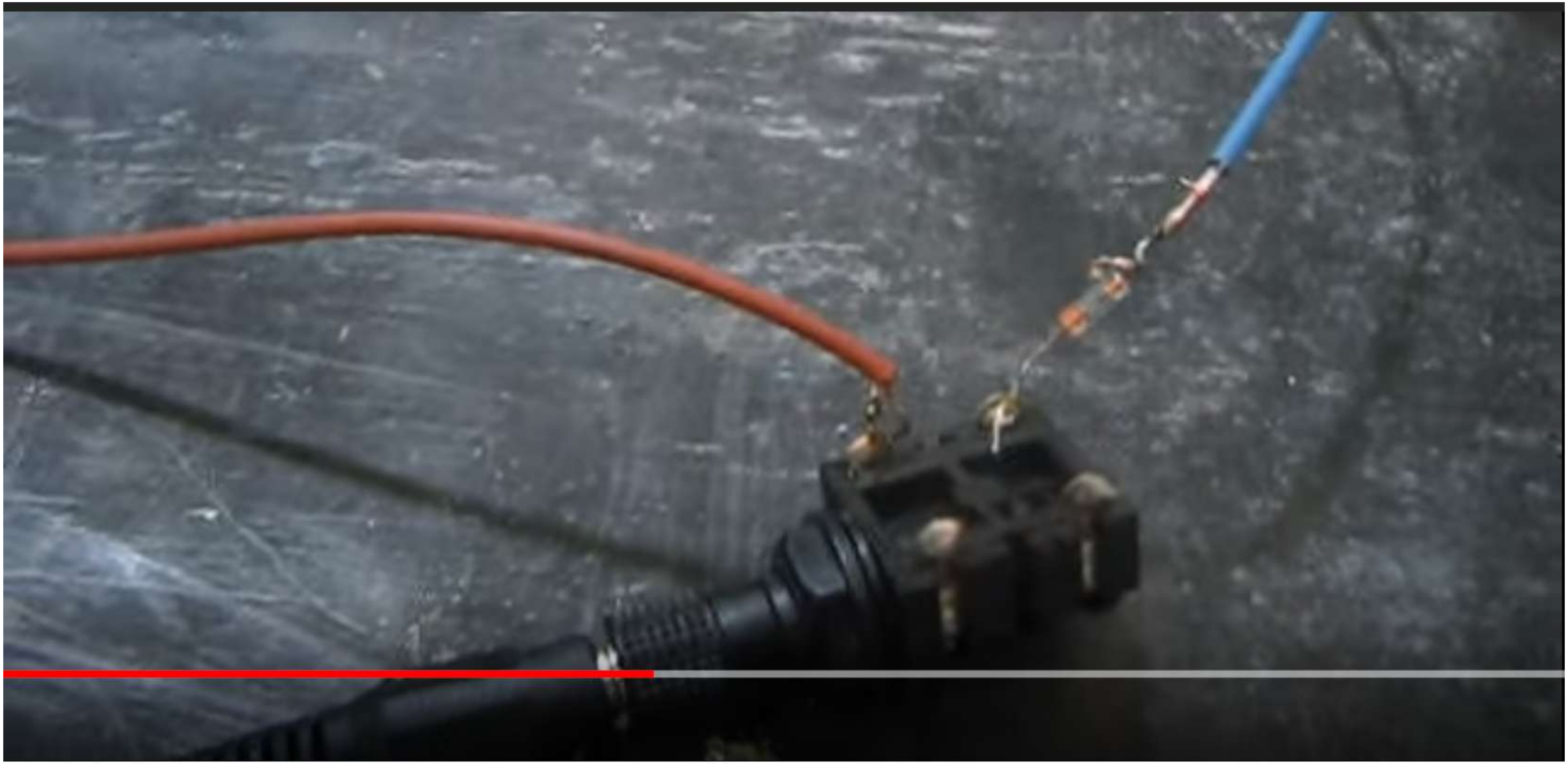

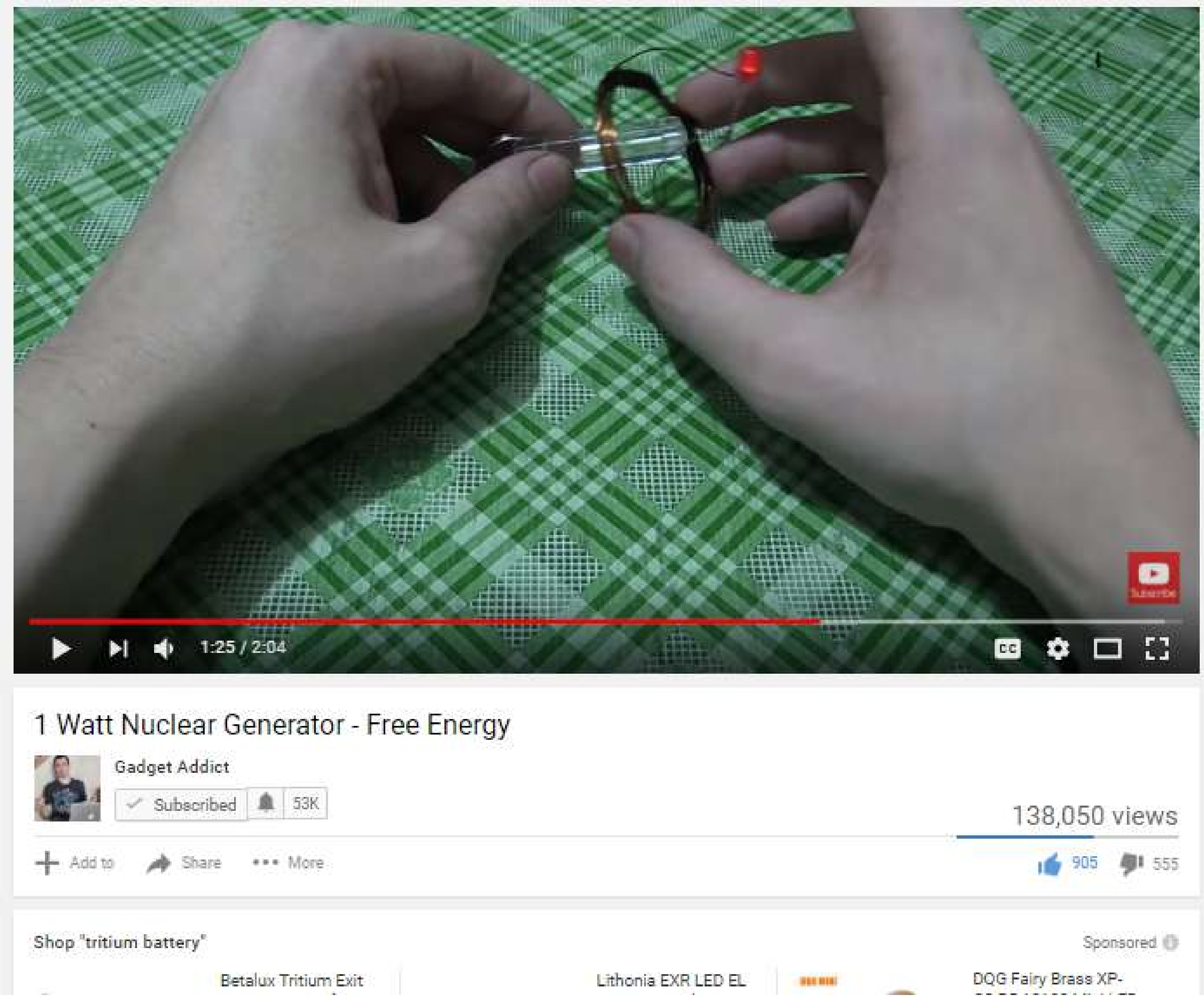
1:25 / 2:04
Subscribe
Gadget Addict
Subscribed
53K
Add to
Share
More
905
555
Sponsored
Betalux Tritium Exit
Lithonia EXR LED EL
DQG Fairy Brass XP-

1 Watt Nuclear Generator - Free Energy

138,050 views

Shop "tritium battery"

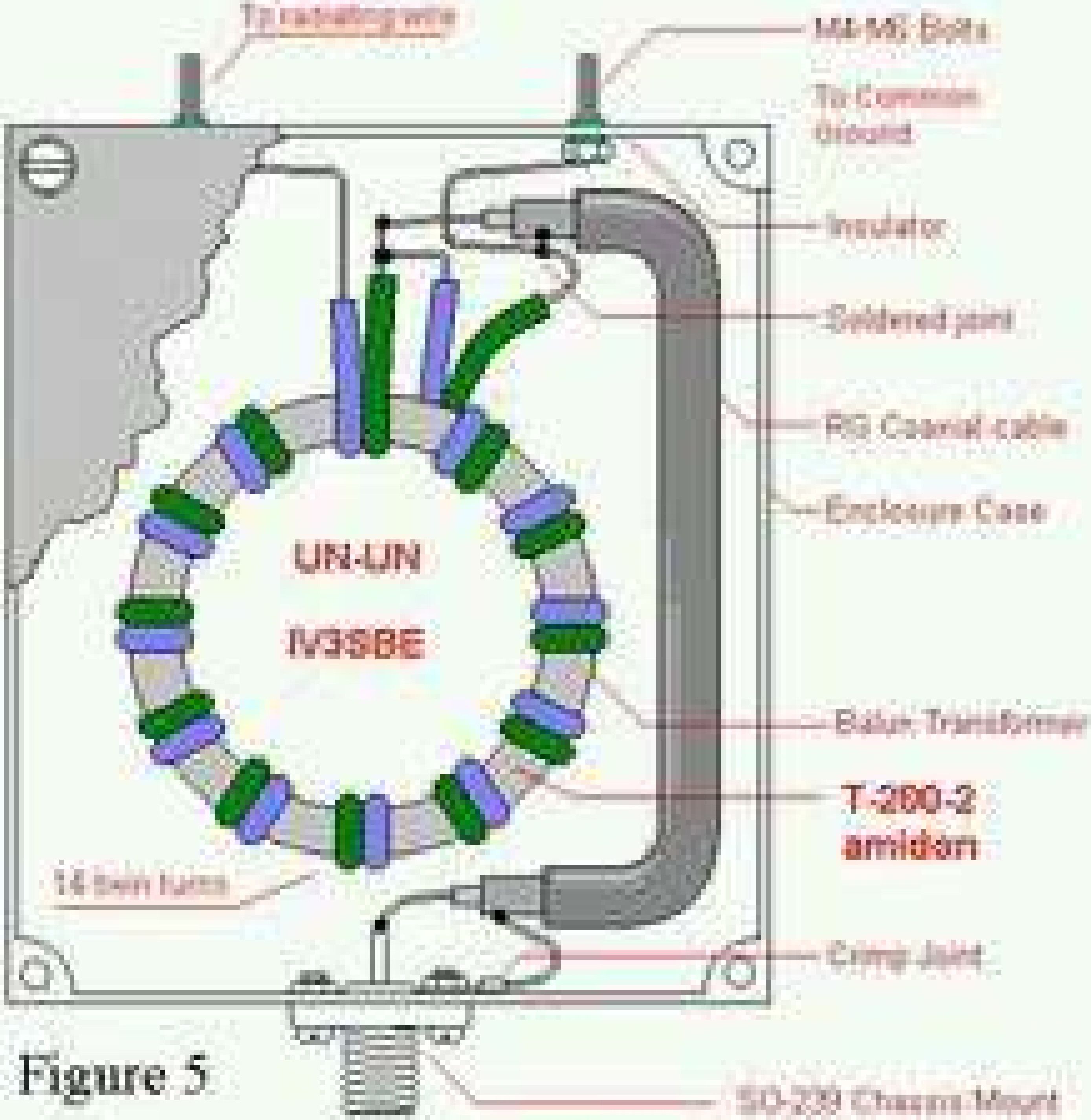
To radiating wire
M4-M6 Bolts
To Common Ground
Insulator
Soldered joint
RG Coaxial cable
Enclosure Case
UN-UN
IV3SBE
Balun Transformer
T-200-2
amidon
14 twin turns
Crimp Joint
SO-239 Chassis Mount

Figure 5

Voltage Balun / "Magnetic Balun"

From GW0VMR's plan

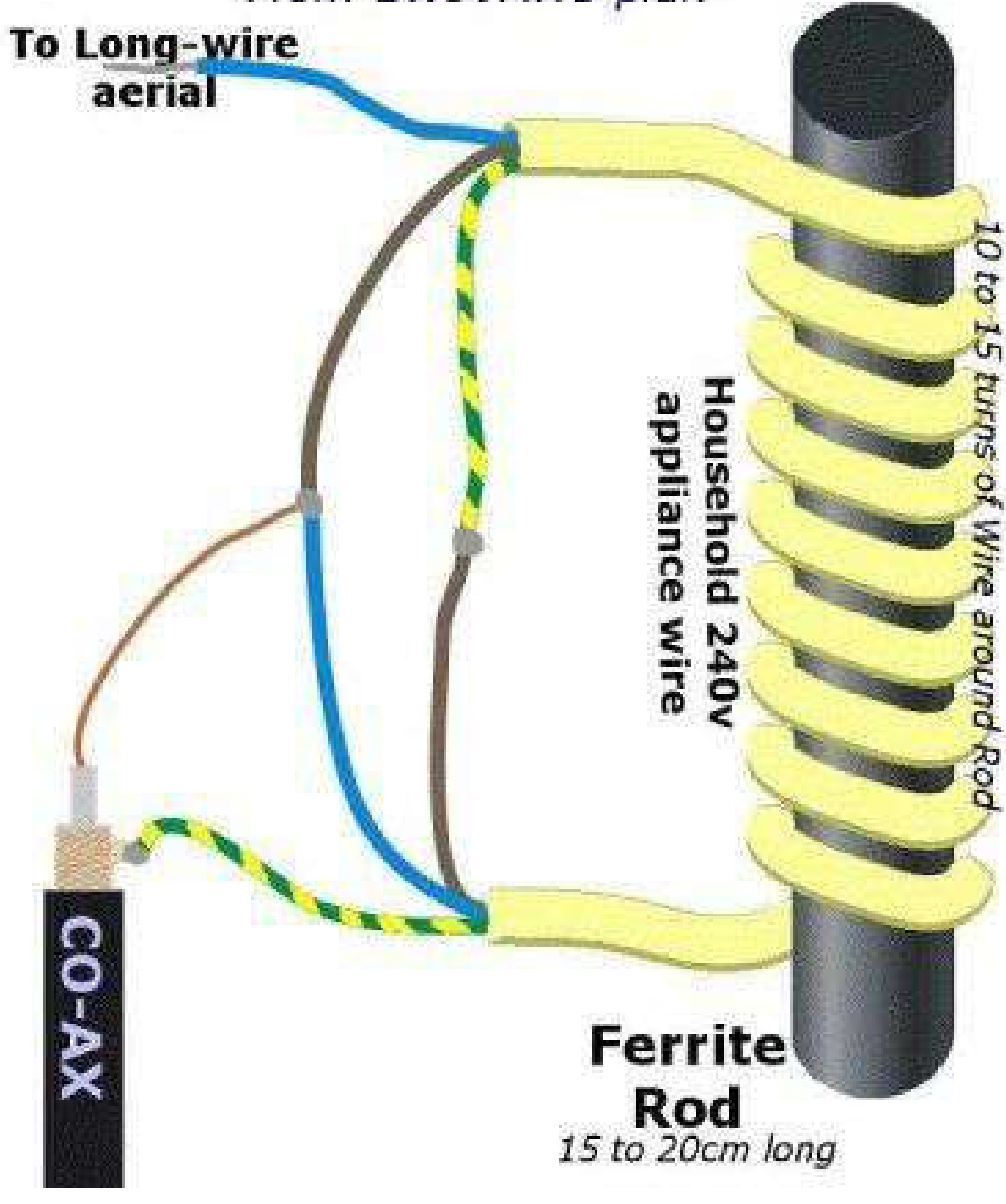
To Long-wire aerial
10 to 15 turns of Wire around Rod
Household 240v appliance wire
CO-AX
Ferrite Rod
15 to 20cm long

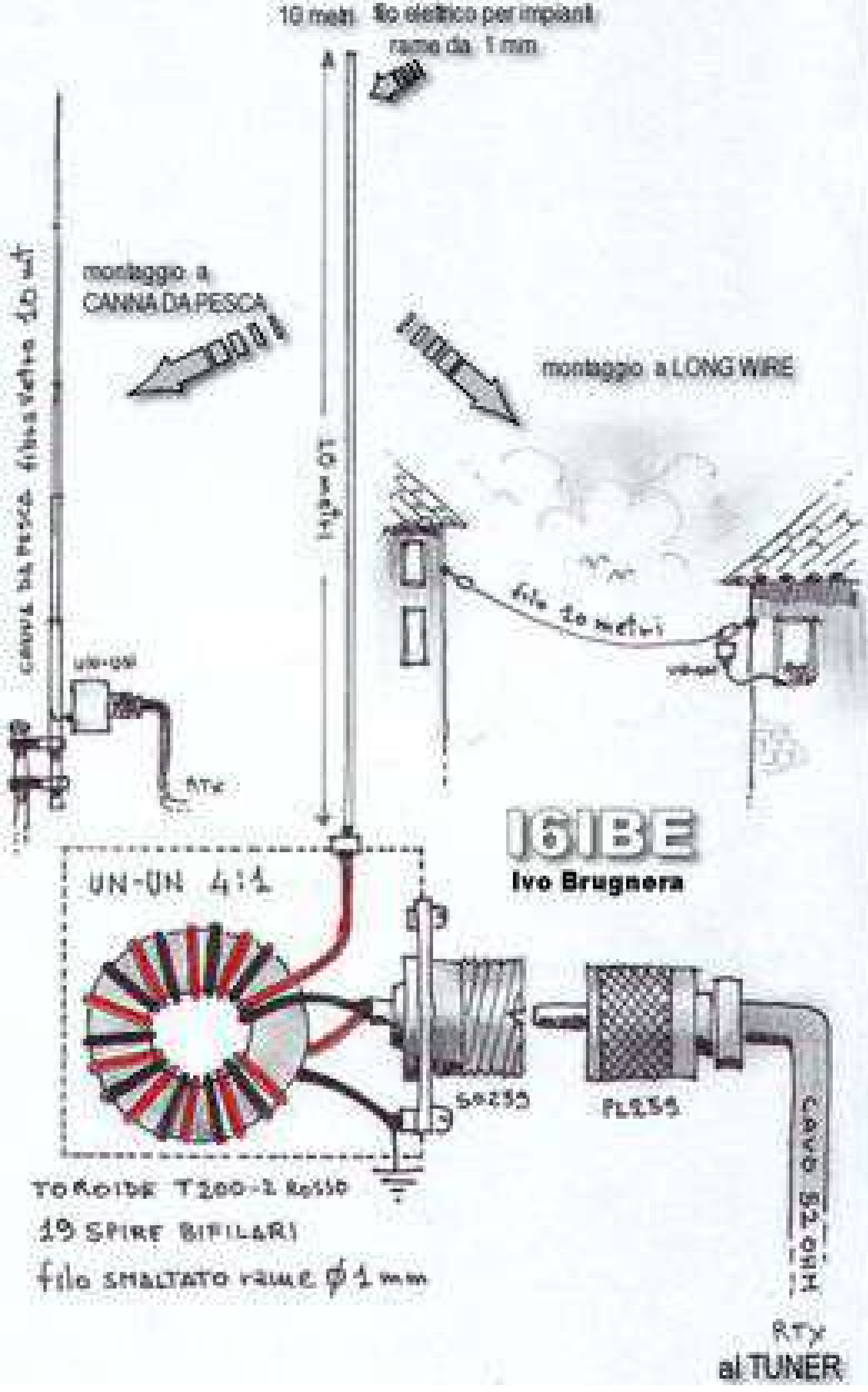
10 metri filo elettrico per impianti rame da 1 mm
montaggio a CANNA DA PESCA
CANNA DA PESCA fibra vetro 10 mt
10 metri
montaggio a LONG WIRE
filo 20 metri
un-un
un-un
RTX
I6IBE
Ivo Brugnera
UN-UN 4:1
SO239
PL259
CAVO 52 OHM
RTX
al TUNER
TOROIDE T200-2 ROSSO
19 SPIRE BIFILARI
filo SMALTATO RAME ø 1 mm

BALUN 9:1 Antenna Long Wire

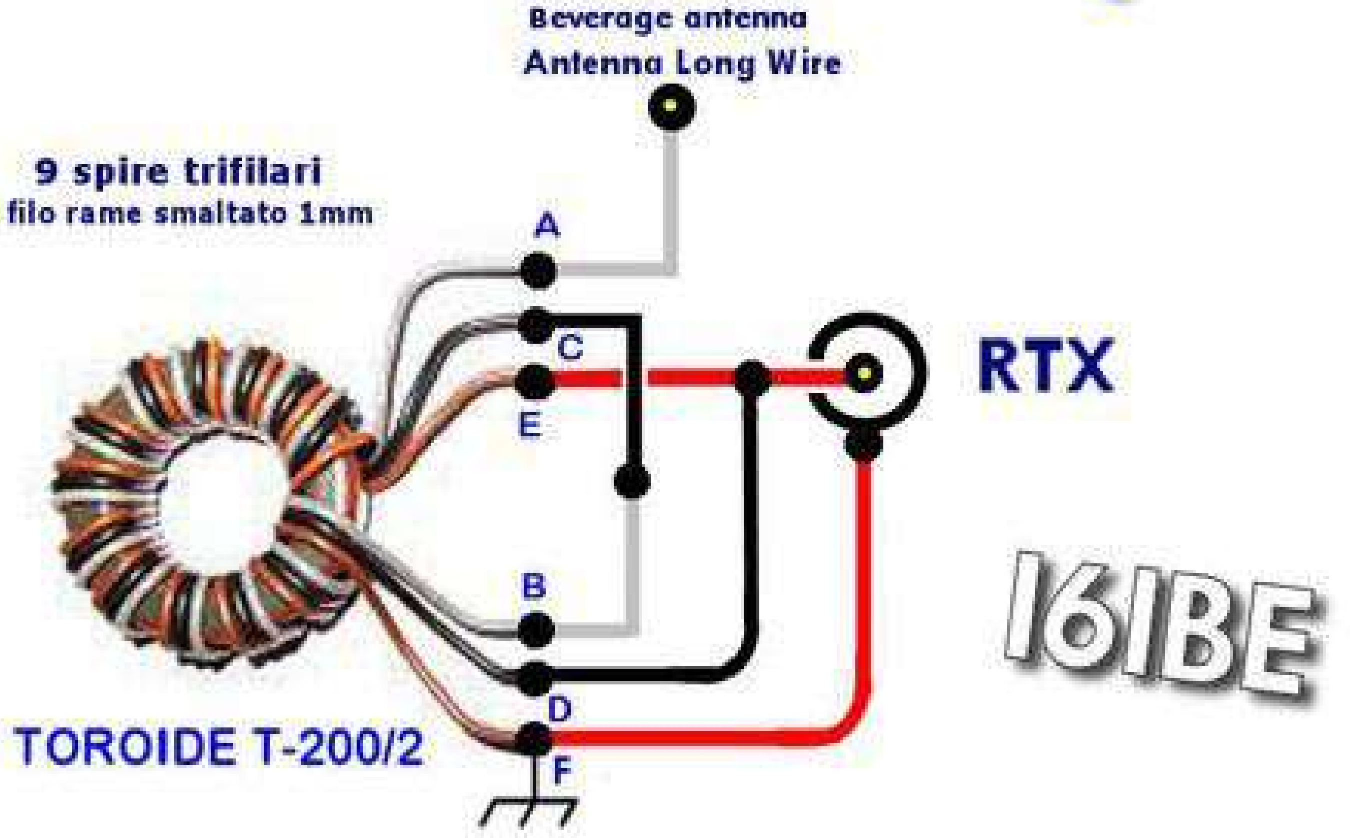
Beverage antenna
Antenna Long Wire
9 spire trifilari
filo rame smaltato 1mm
A
C
E
RTX
B
D
F
TOROIDE T-200/2
I6IBE

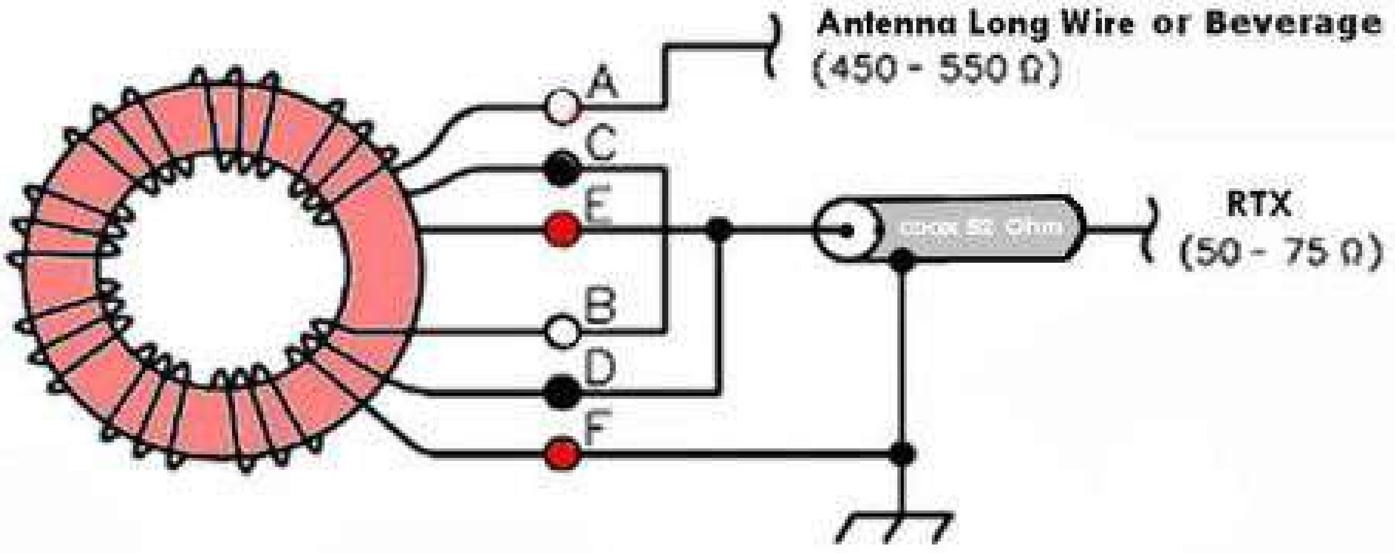
Antenna Long Wire or Beverage
(450 - 550 Ω)
A
C
E
coax 52 Ohm
RTX
(50 - 75 Ω)
B
D
F

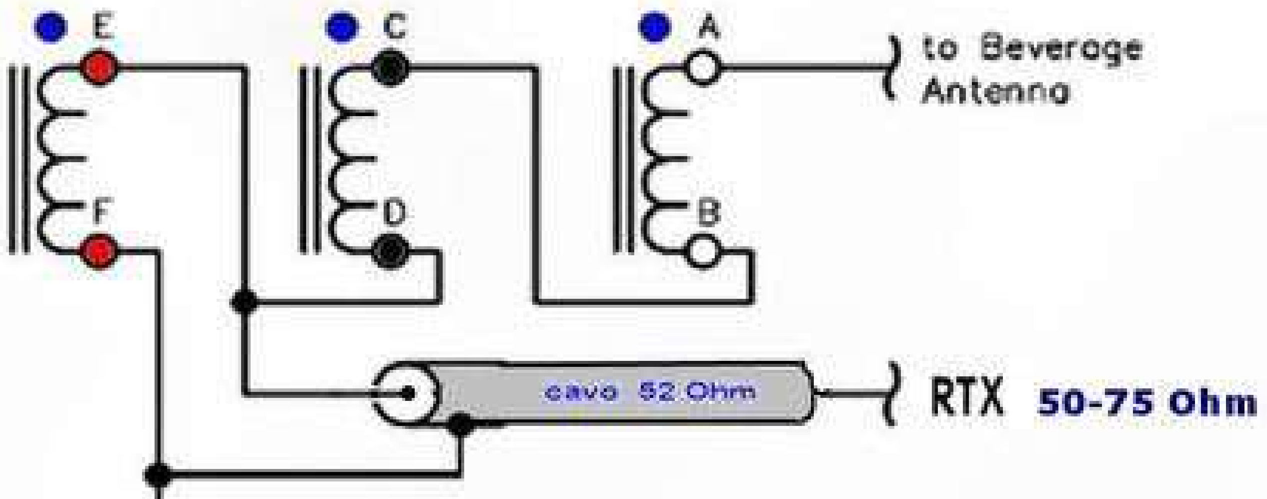
E
F
C
D
A
B
to Beveroge
Antenna
cavo 52 Ohm
RTX 50-75 Ohm

Bal-Un 2:1 per antenna DELTA-LOOP

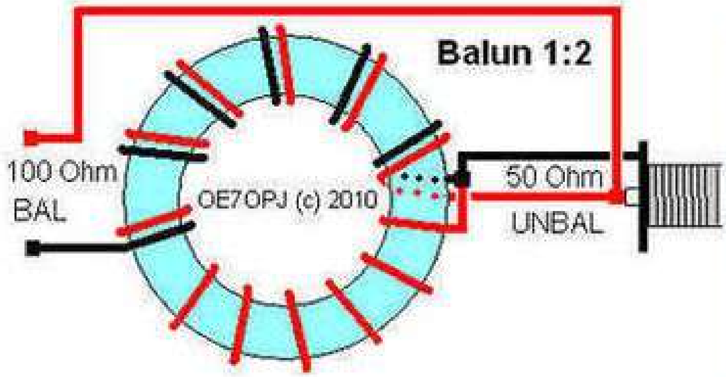
Balun 1:2
100 Ohm
BAL
OE7OPJ (c) 2010
50 Ohm
UNBAL

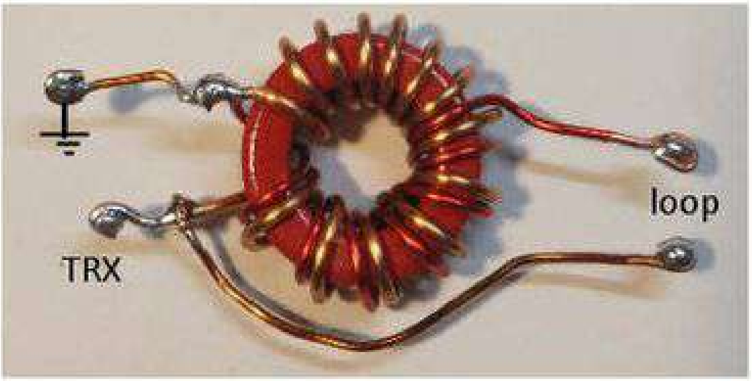

by OE7OPJ

Toroide	Numero di Spire	Potenza Supportata
T80-2	25	60 Watts
T106-2	16	100 Watts
T130-2	18	150 Watts
T157-2	16	250 Watts
T200-2	17	400 Watts
T200A-2	13	400 Watts
T400-2	14	1000 Watts

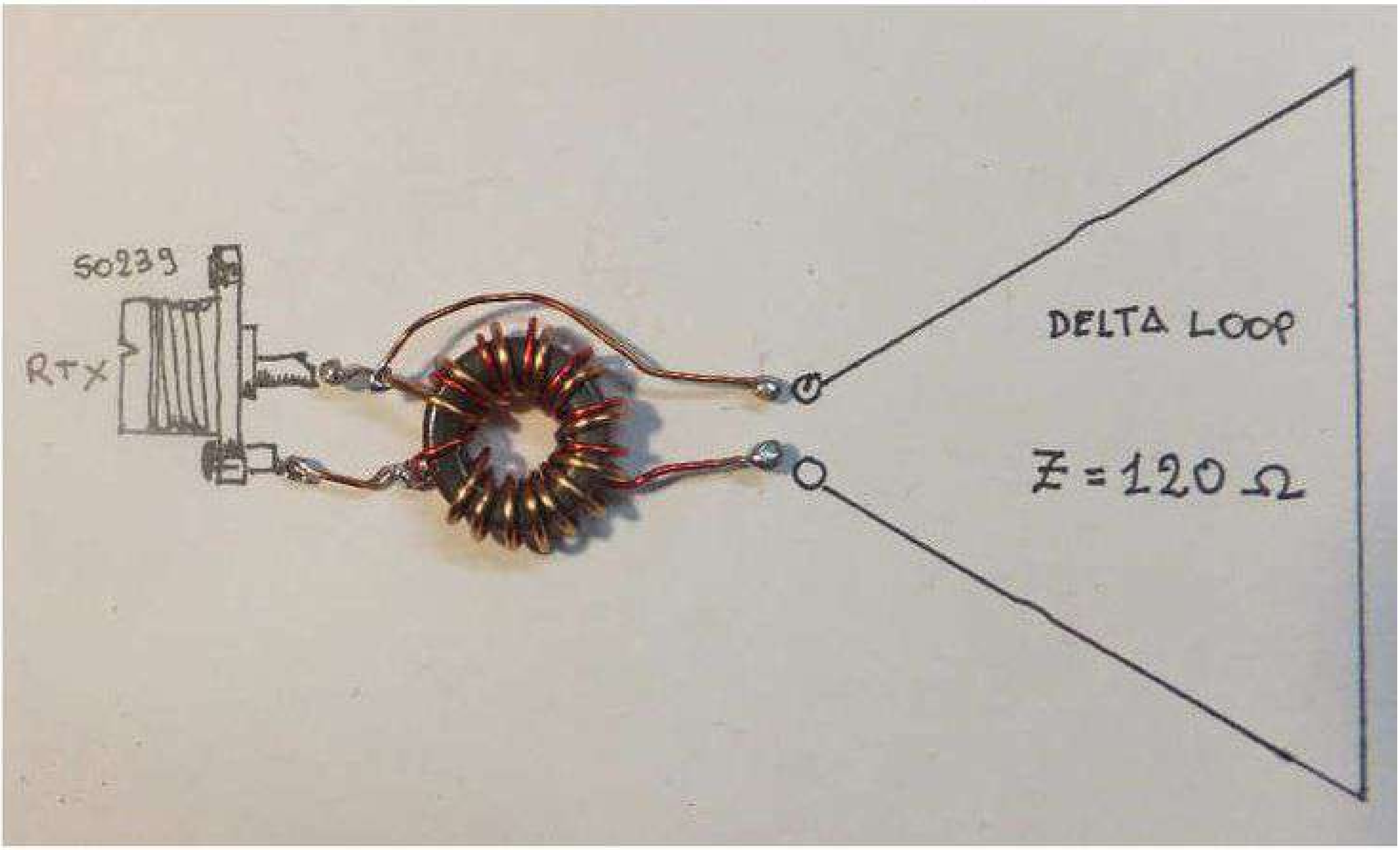
SO239
RTX
DELTA LOOP
Z = 120 Ω

Autoformer, not a parallel transmission line – Ruthroff balun (voltage)

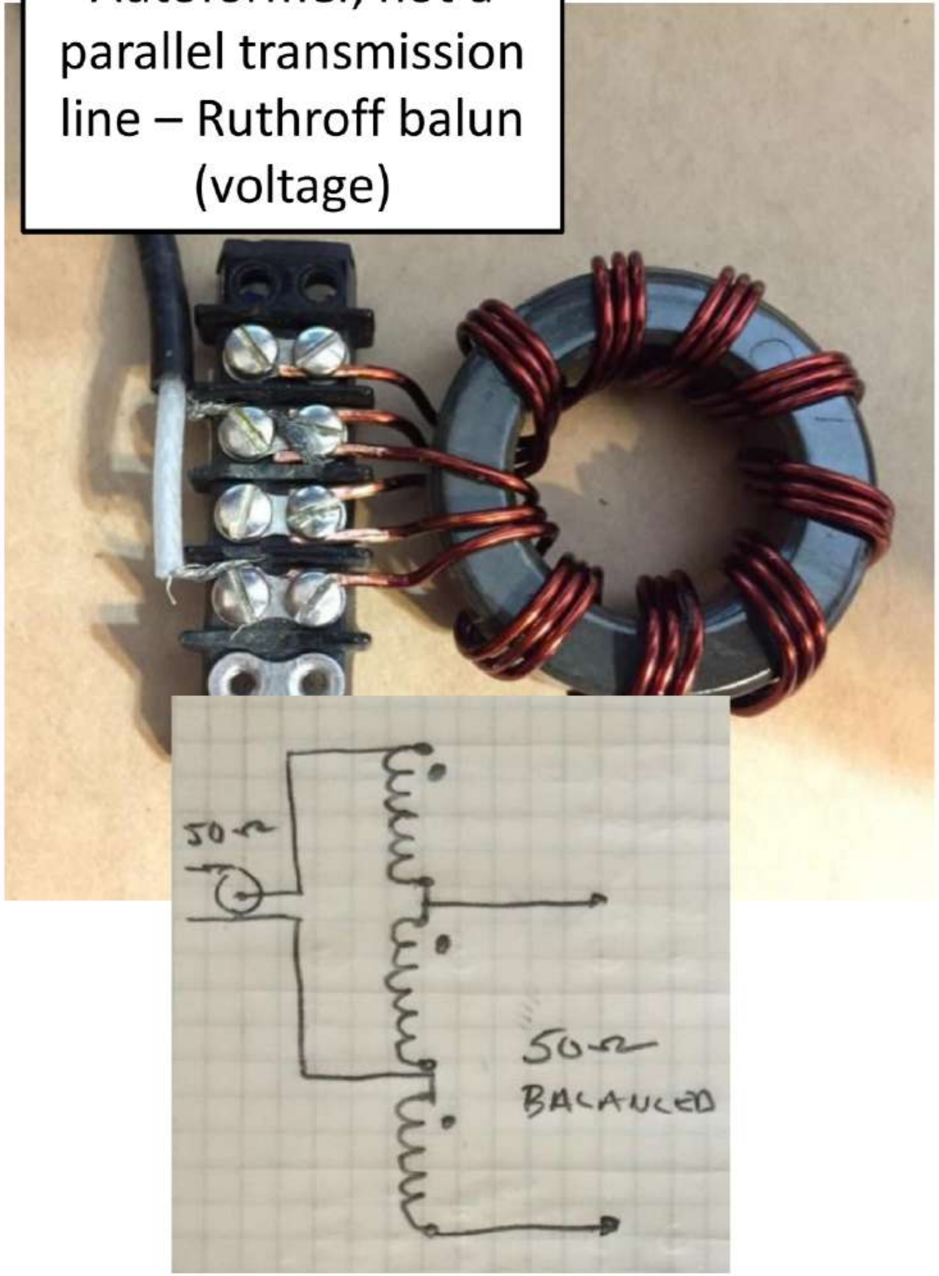

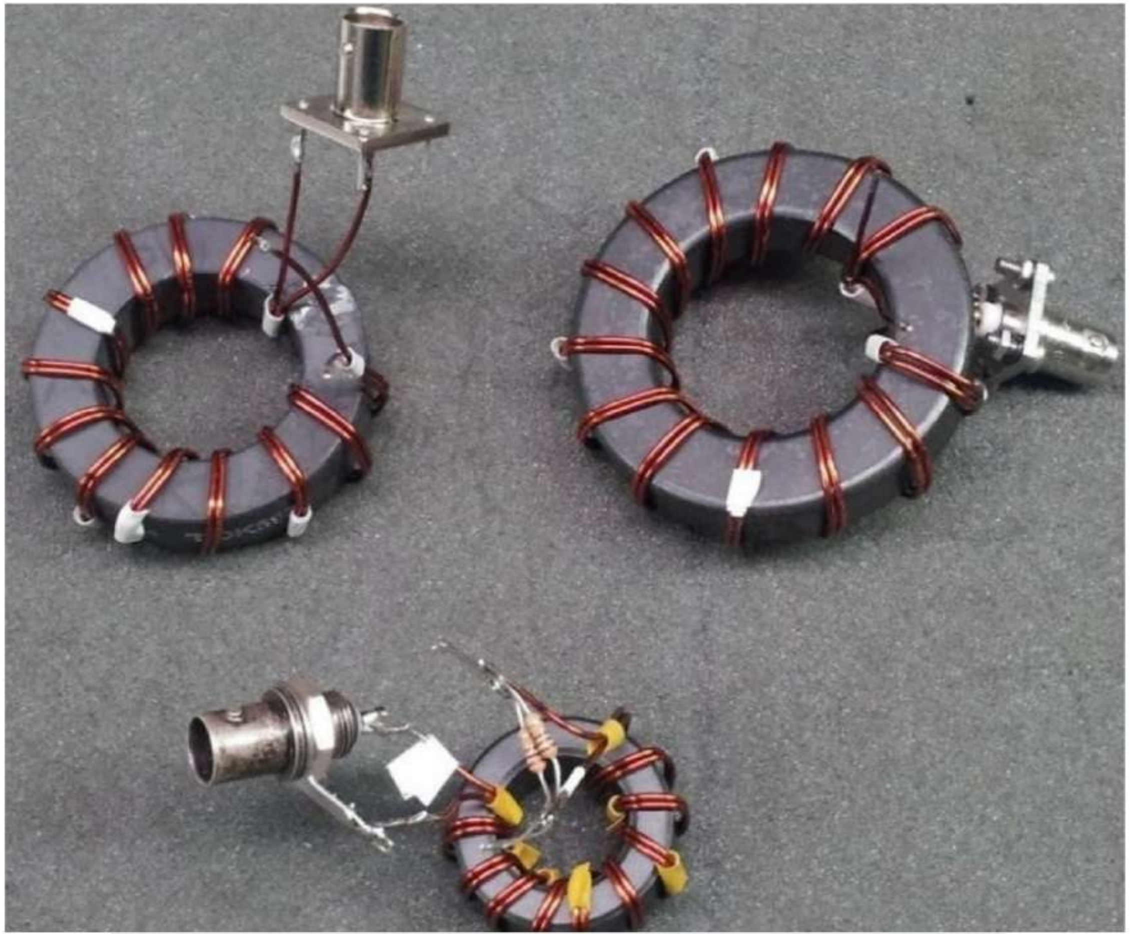

One or more parallel transmission lines, connected various ways – Guanella balun (current) – core is used solely for choking action

F/B: 0.00 dB; Rear: Azim. 120 deg, Elev. 60 deg

Coax is unbalanced …..

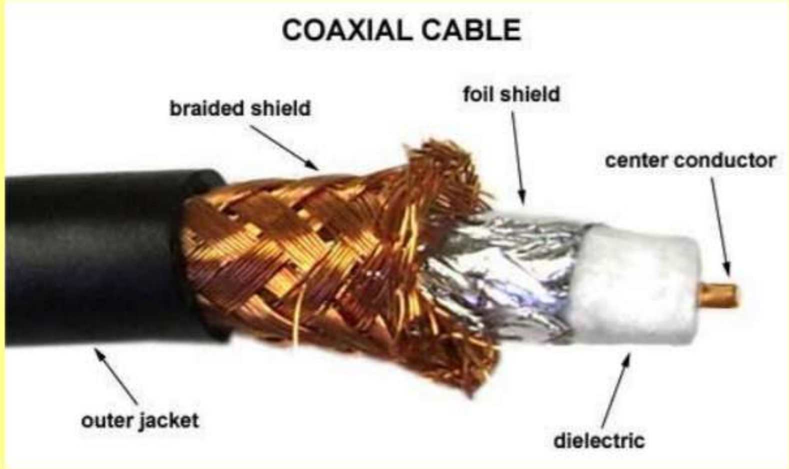
COAXIAL CABLE
braided shield
foil shield
center conductor
outer jacket
dielectric

300 Ohm
450 Ohm

… Ladder line is balanced

Coax is unbalanced, while ladder line and open wire feeder is balanced, as long as it is used properly and is kept away from metal and "earthy" materials

Convert parallel feedline to coax

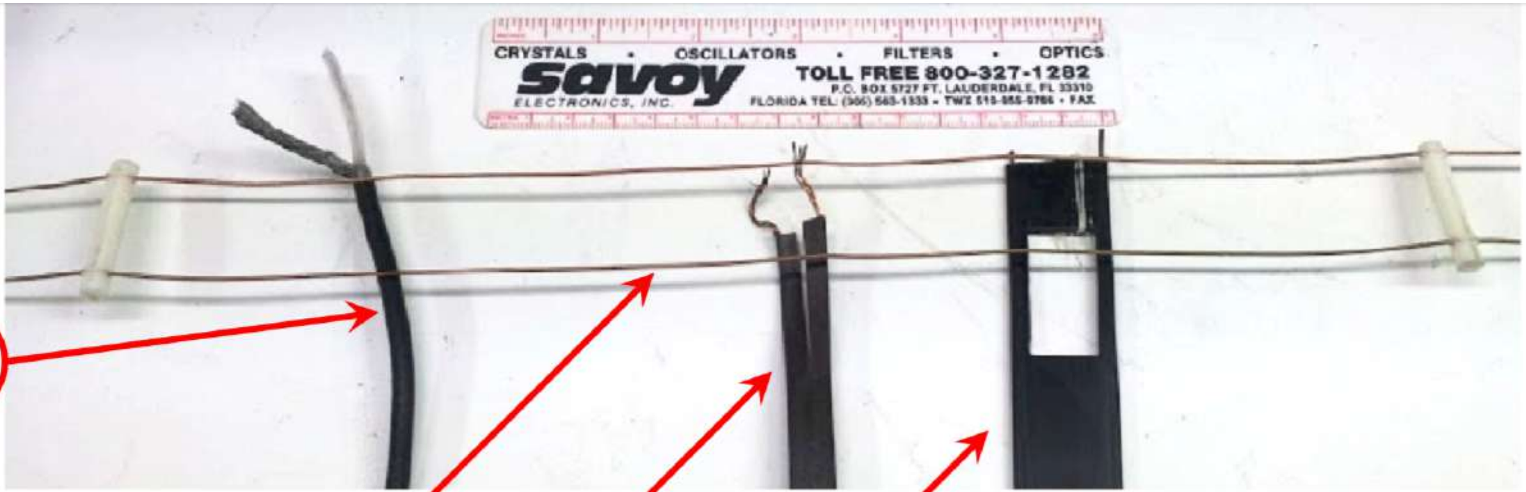
CRYSTALS • OSCILLATORS • FILTERS • OPTICS
Savoy
ELECTRONICS, INC.
TOLL FREE 800-327-1282

Types of Parallel Line

- **Open Feeder** or **Ladder line**
  - Typ 450 and 600 ohm
- **Twin Lead**
  - Common for (old) outdoor TV antennas
- **Window line**
  - Commonly available today

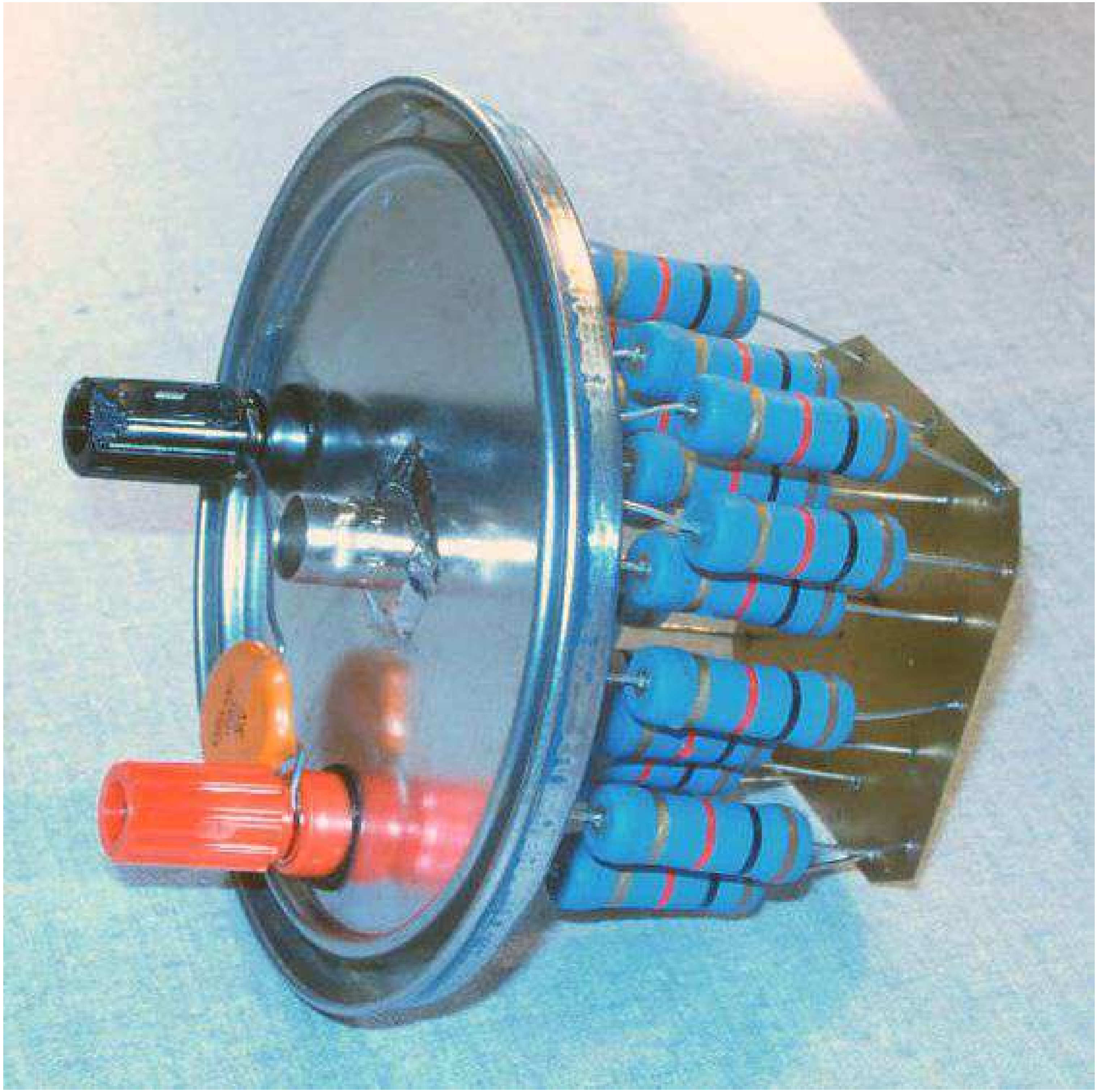

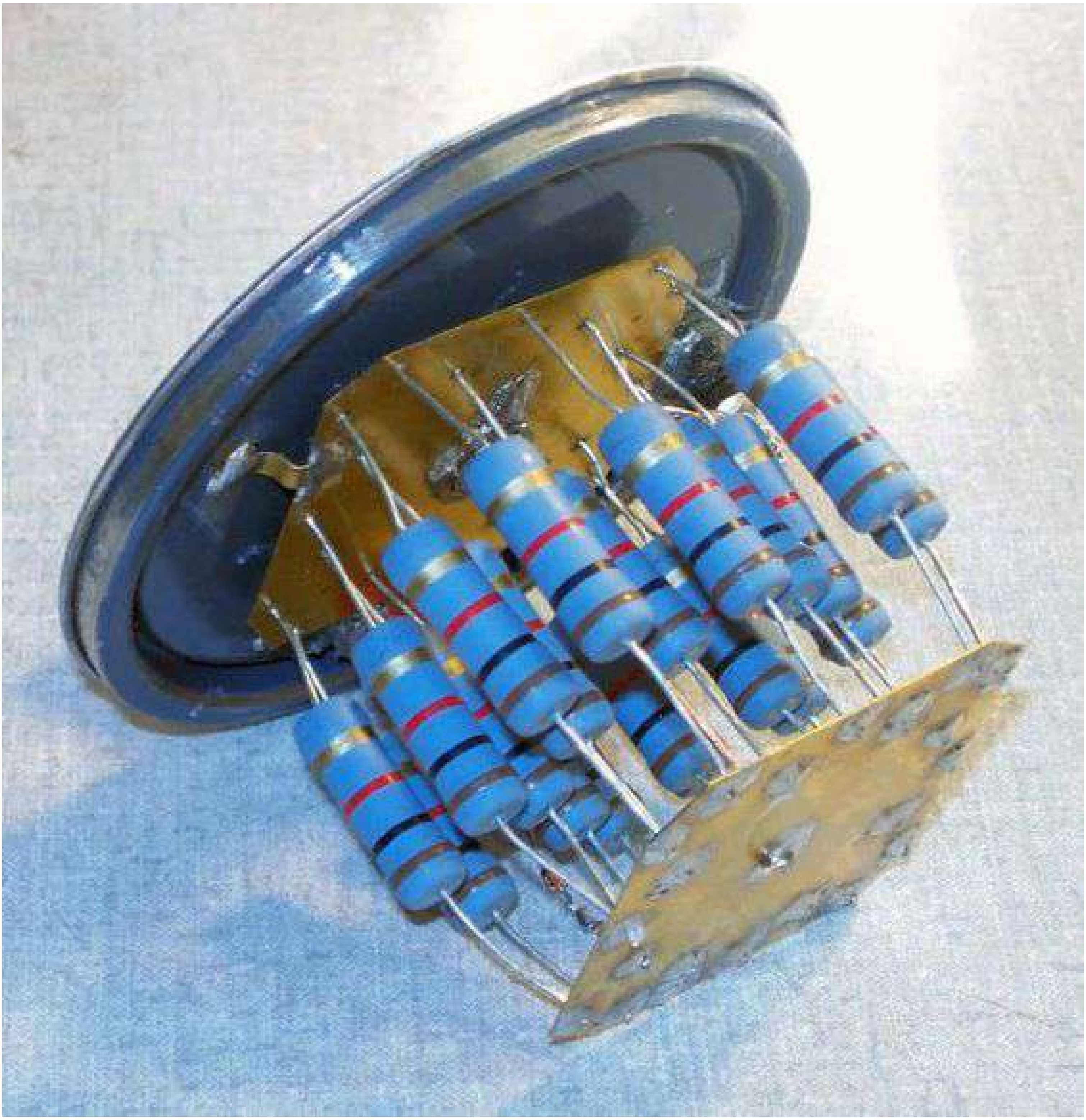

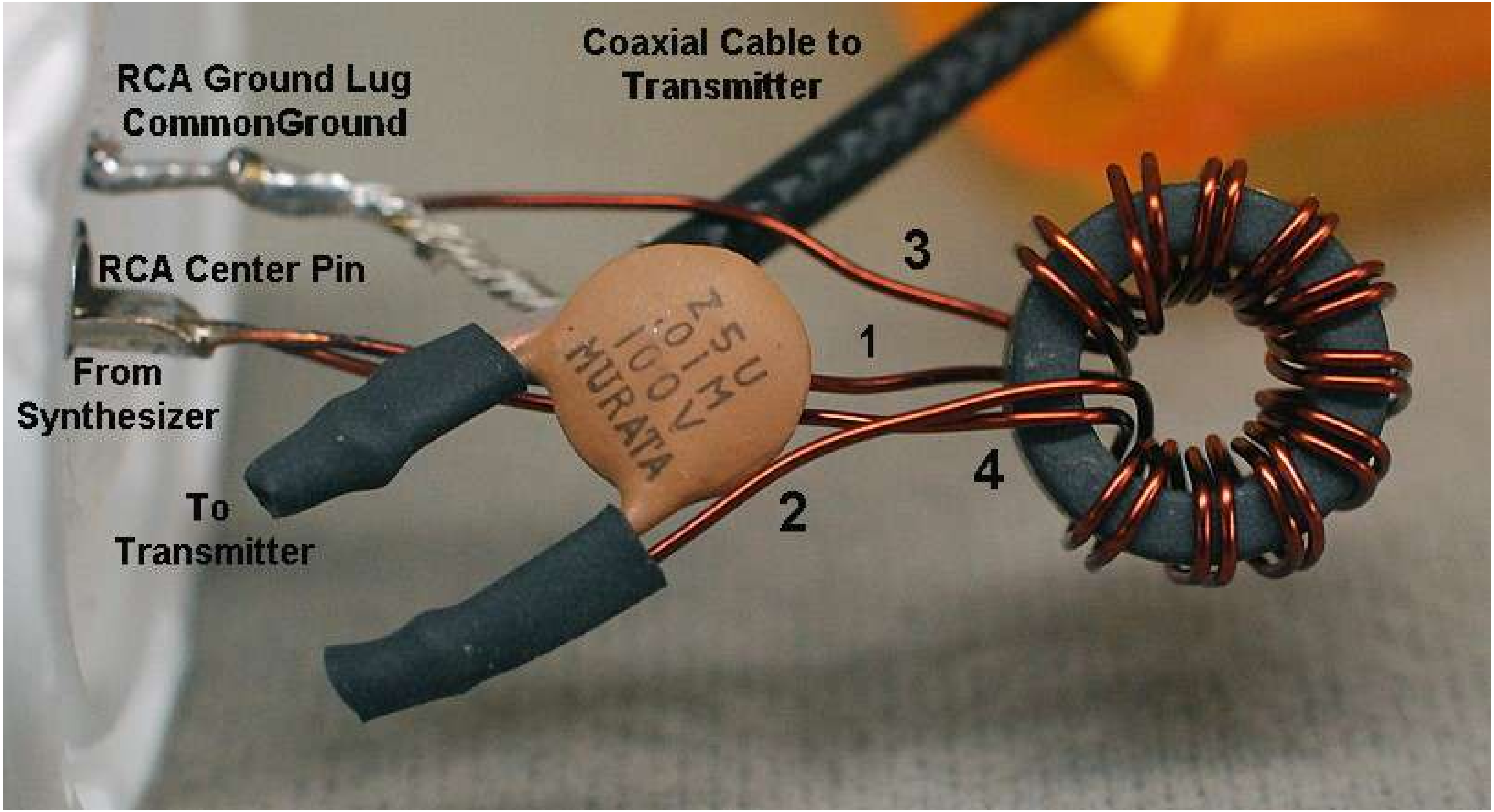
Coaxial Cable to Transmitter
RCA Ground Lug CommonGround
RCA Center Pin
From Synthesizer
To Transmitter
3
1
4
2
Z5U
.01M
100V
MURATA

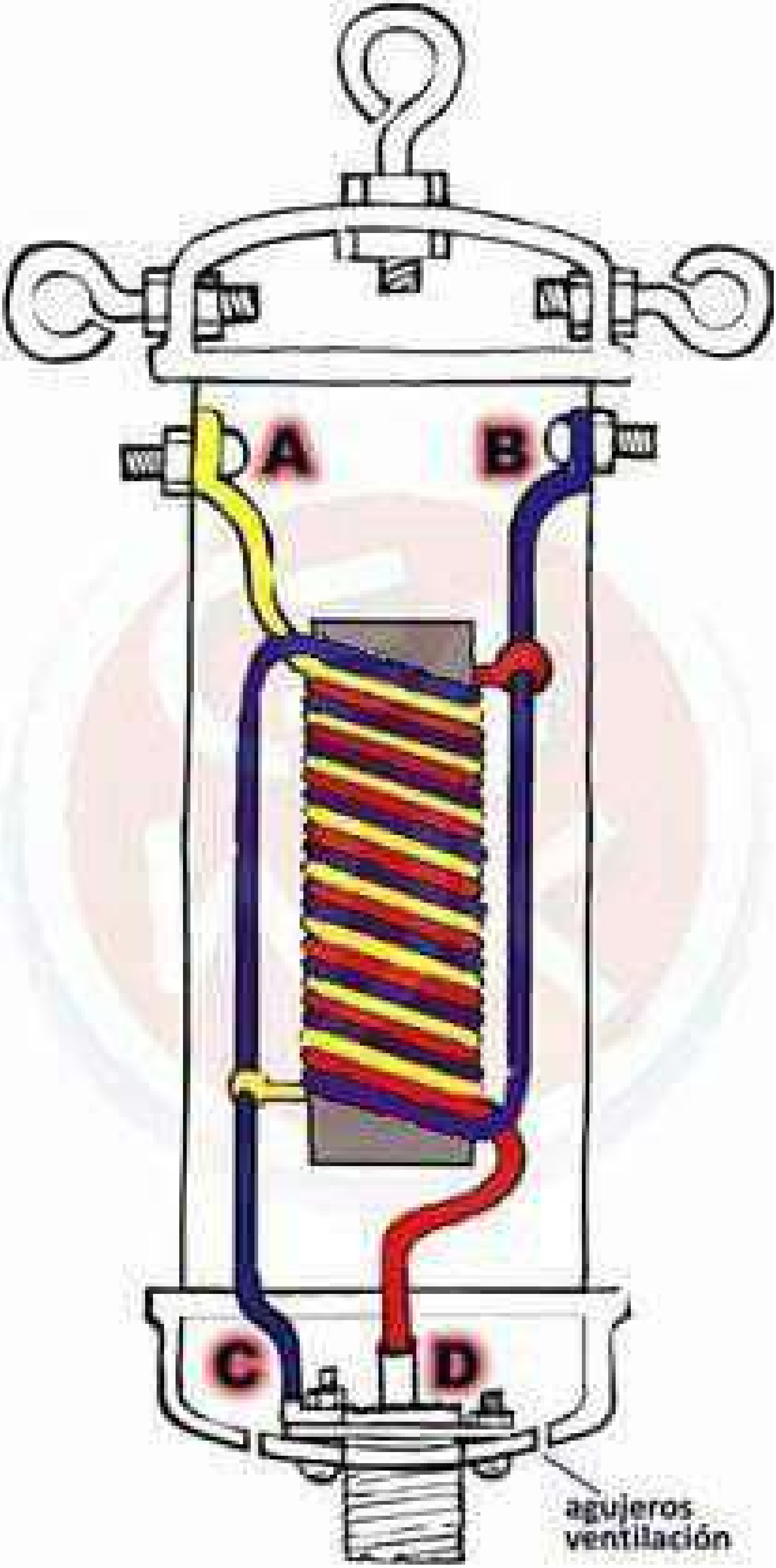
A
B
C
D
agujeros ventilación

Formula LUNGHEZZA DELTA LOOP	Esempio 14.00 MHz (20 metri)	Lunghezza TOTALE del LOOP
Velocita'Luce : Freq MHz x 0.97= L	**299.8** VL : **14.100** MHz x **0,97**	**20.624** metri totale loop
$L=\frac{300.000}{MHz} \times 0.97$	I6IBE	
Formula LUNGHEZZA Stub coax 75 Ohm	Esempio 14.00 MHz (20 metri)	Lunghezza TOTALE cavo 75 Ohm
Lunghezza LOOP x 0.66 : 4	**20.624** x **0.66** : **4**	**3.402** Metri coax RG-59

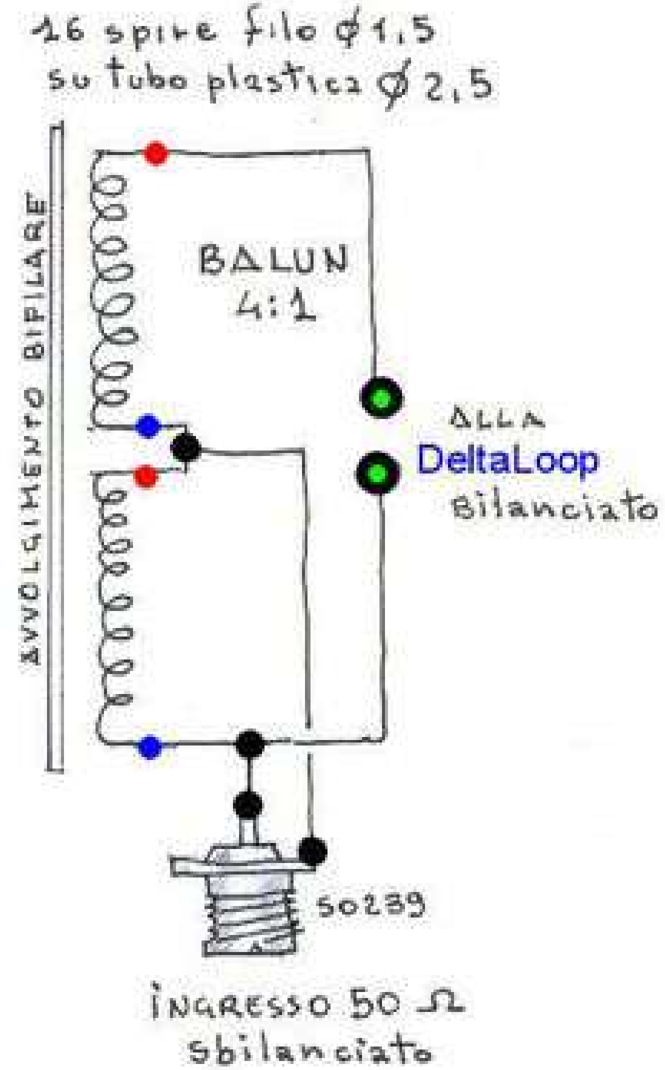
16 spire filo ∅1,5
su tubo plastica ∅2,5
AVVOLGIMENTO BIFILARE
BALUN 4:1
ALLA
DeltaLoop
Bilanciato
SO239
ingresso 50 Ω
sbilanciato

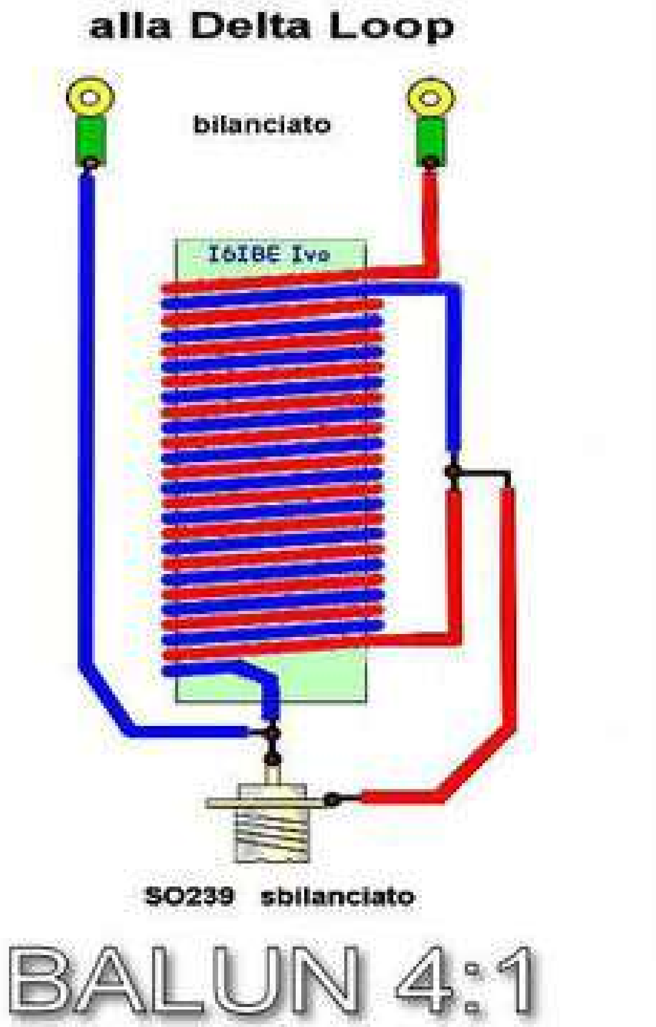
alla Delta Loop
bilanciato
I6IBE Ivo
SO239 sbilanciato
BALUN 4:1

I6IBE

Choke RFI

toroide

cavo RG-58

5+5 spire

controfase

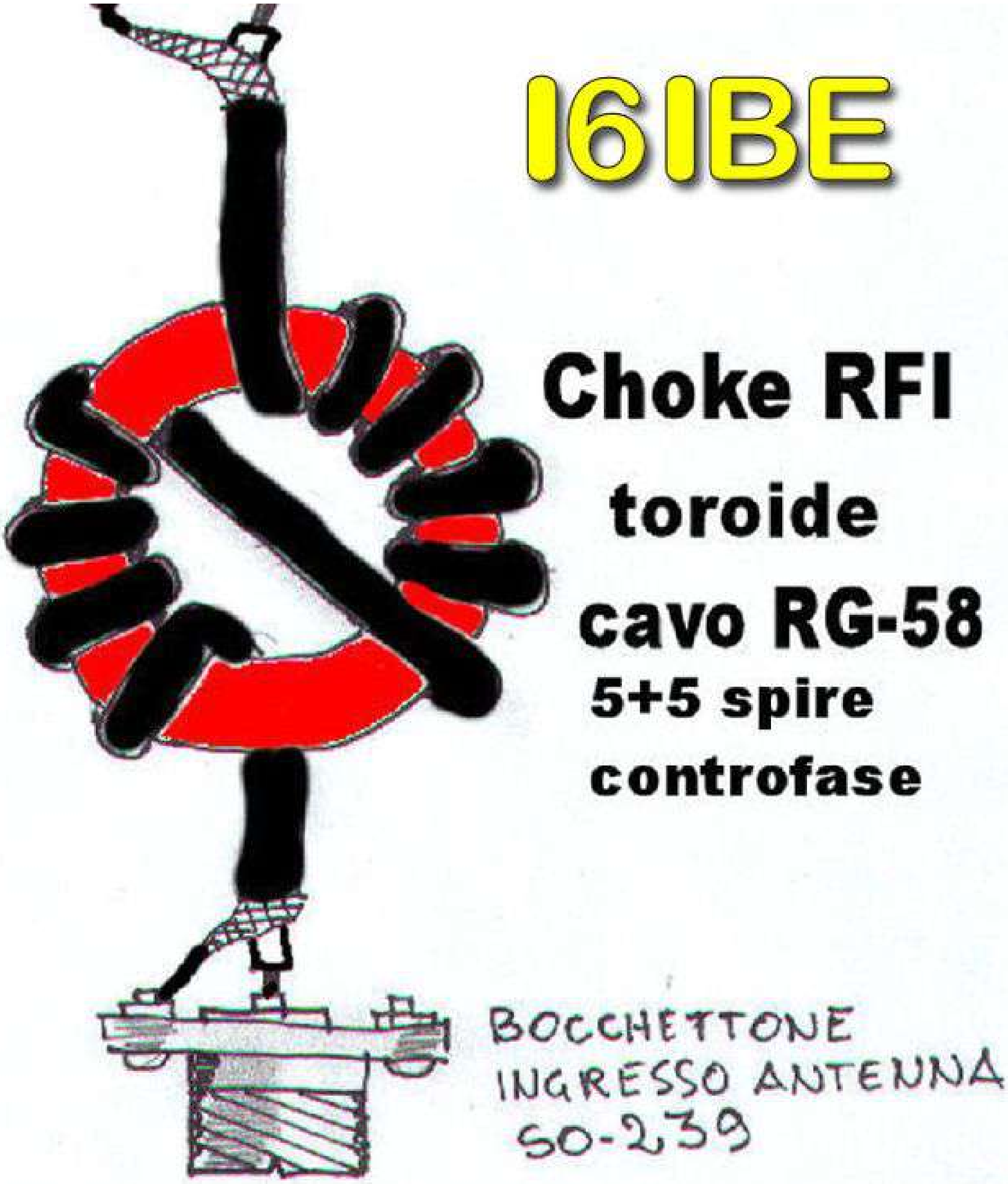
BOCCHETTONE
INGRESSO ANTENNA
SO-239

2. Insert metal electrodes and attach multimeter leads [copper (+), aluminum (-)] For measuring direct current voltage: set multimeter function switch to "DCV: 20" take a reading in volts DC. for measuring direct current: set multimeter function switch to "DCA: 20m" take a reading in milliamps (mA) DC.

Lawn battery (summer). INSET: Marsh mud battery. (Circles show position of electrodes.)

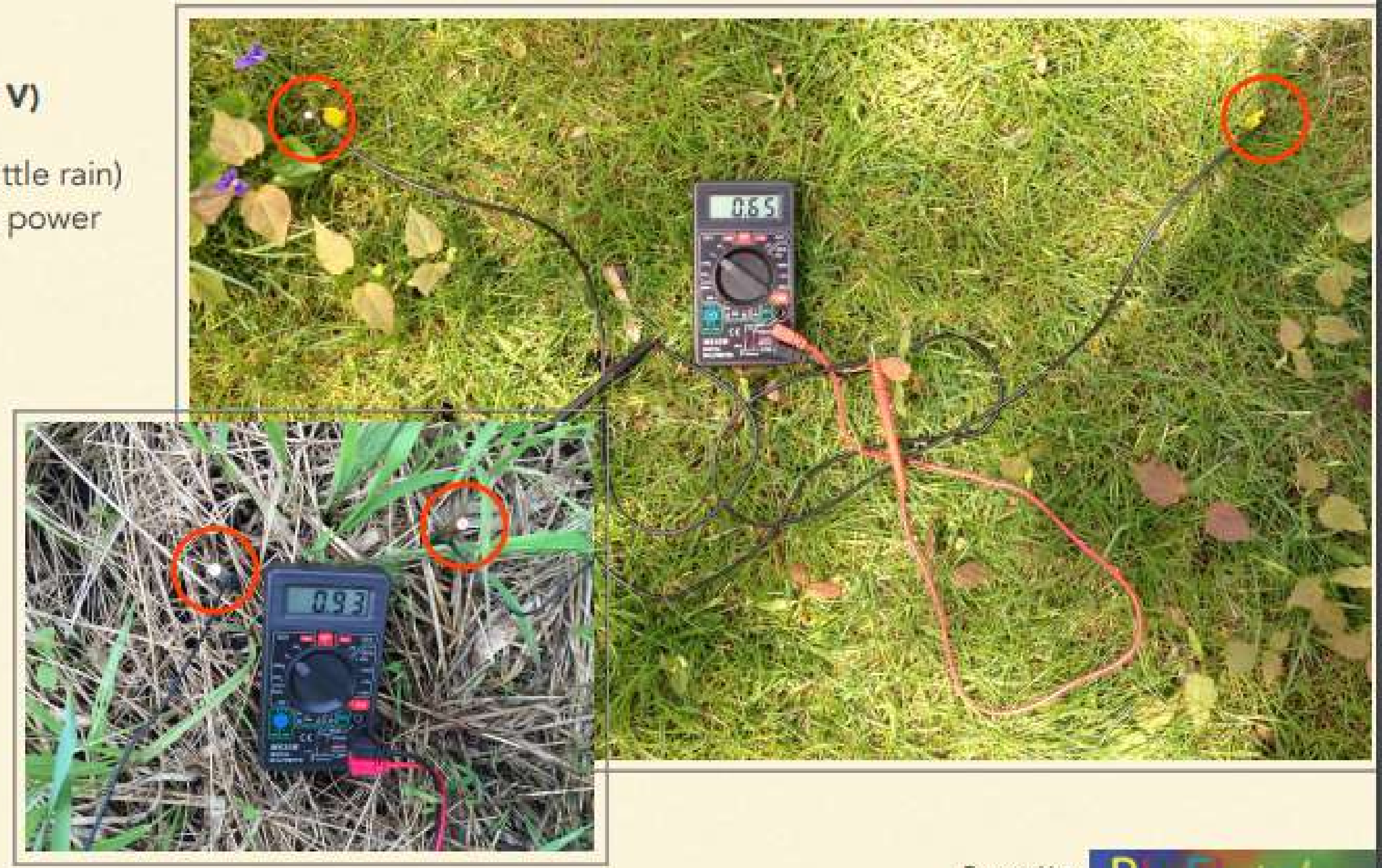

Calculating Earth Battery Power (W = I * V)

EXAMPLE: A lawn battery in late summer (little rain) produces a 0.65V, 0.2mA current. A battery power calculation of 0.00013W (0.13mW).

Powered by: PIXEL ink LLC

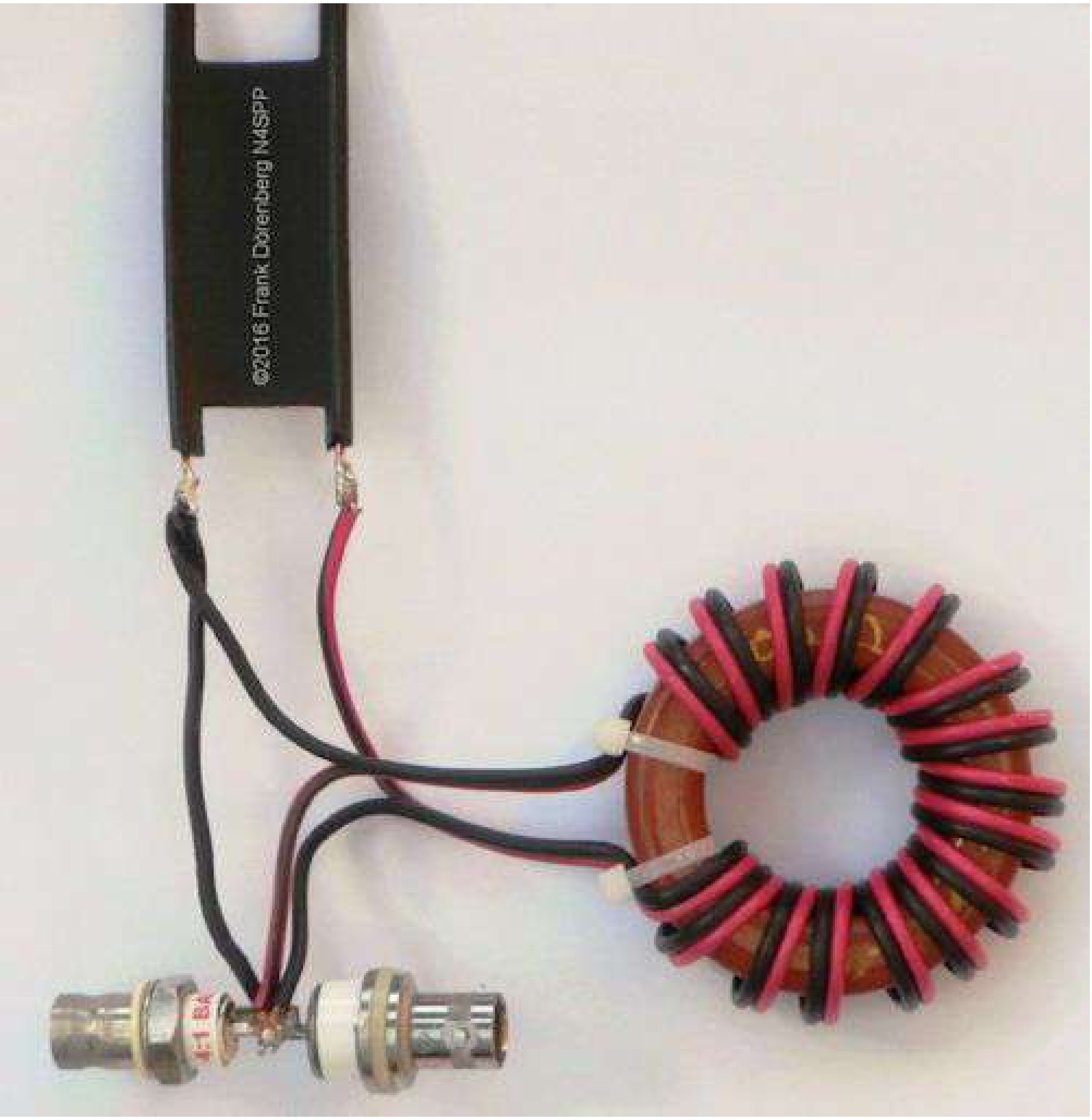
©2016 Frank Dorenberg N4SPP

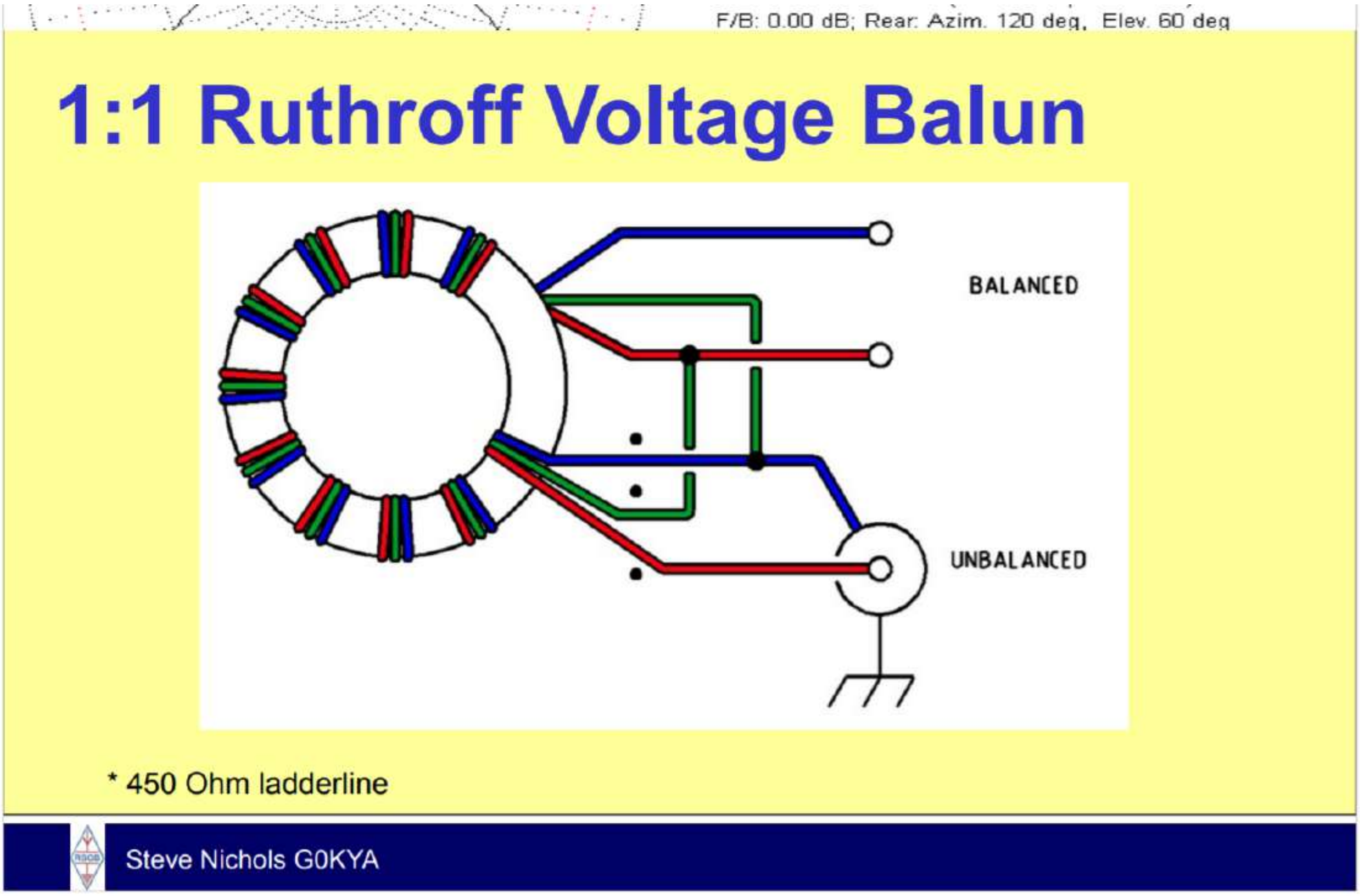

1:1 Ruthroff Voltage Balun

* 450 Ohm ladderline

Steve Nichols G0KYA

The Ruthroff or voltage balun is another type. Voltage baluns balance the voltages. Current balance is considered to be superior to voltage balance. Voltage baluns should not be used in lines with high SWR. They have the narrowest impedance and frequency range of any balun type. Properly designed voltage baluns have low common mode impedance. Properly designed current baluns have high common mode impedance, and provide better balance. If you want to stop common mode currents flowing on your coax a current balun is a better choice. But in the Carolina Windom, which uses the vertical coax as a radiator you WANT common mode currents, so a Ruthroff voltage balun is better.

F/B: 0.00 dB; Rear: Azim. 120 deg, Elev. 60 deg

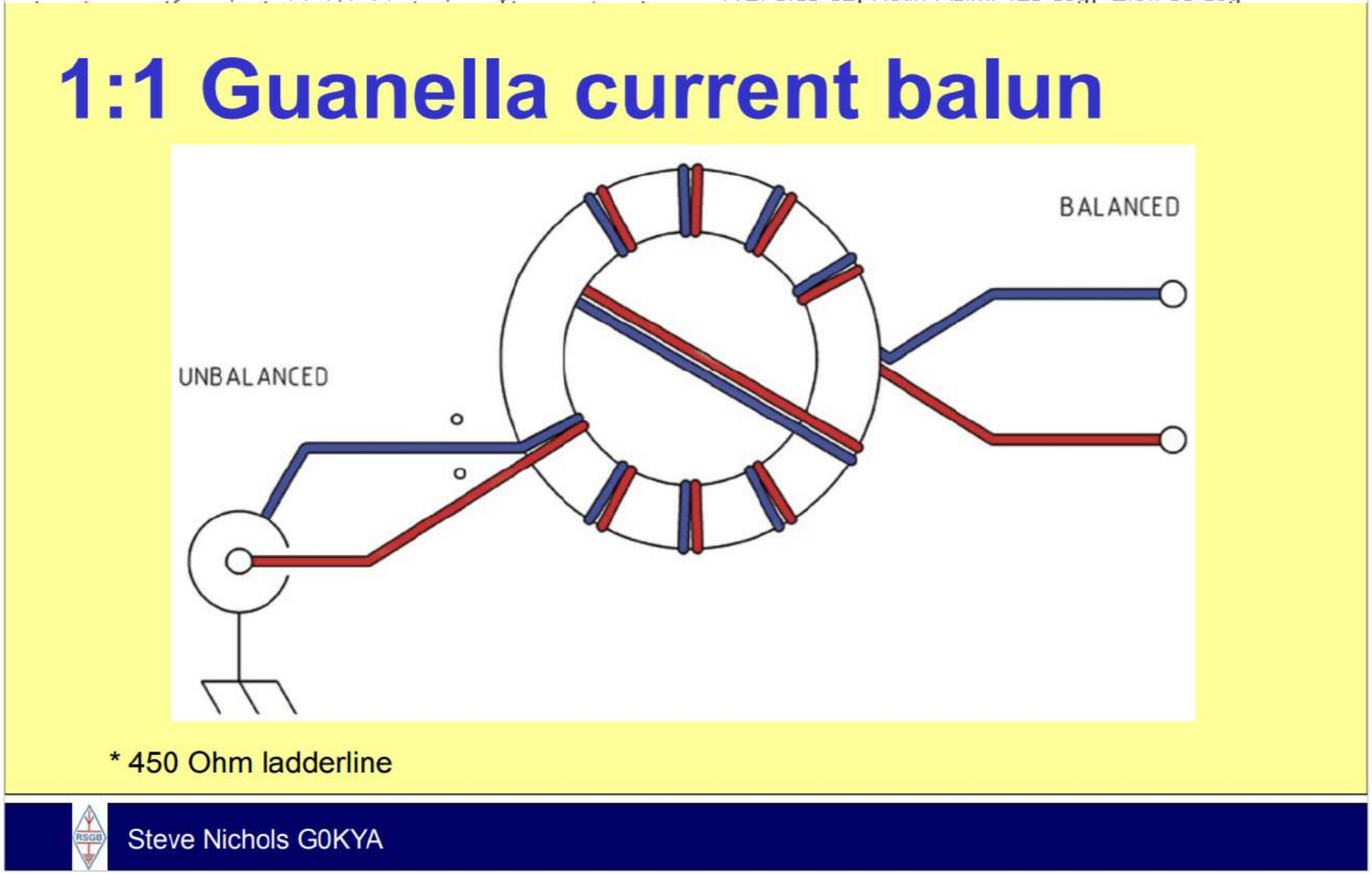

1:1 Guanella current balun

* 450 Ohm ladderline

Steve Nichols G0KYA

A Guanella or current balun is very common. Current baluns stop RF from coming back down the outside of the coax shield, and are so called that because they "force equal currents in each side of a dipole"

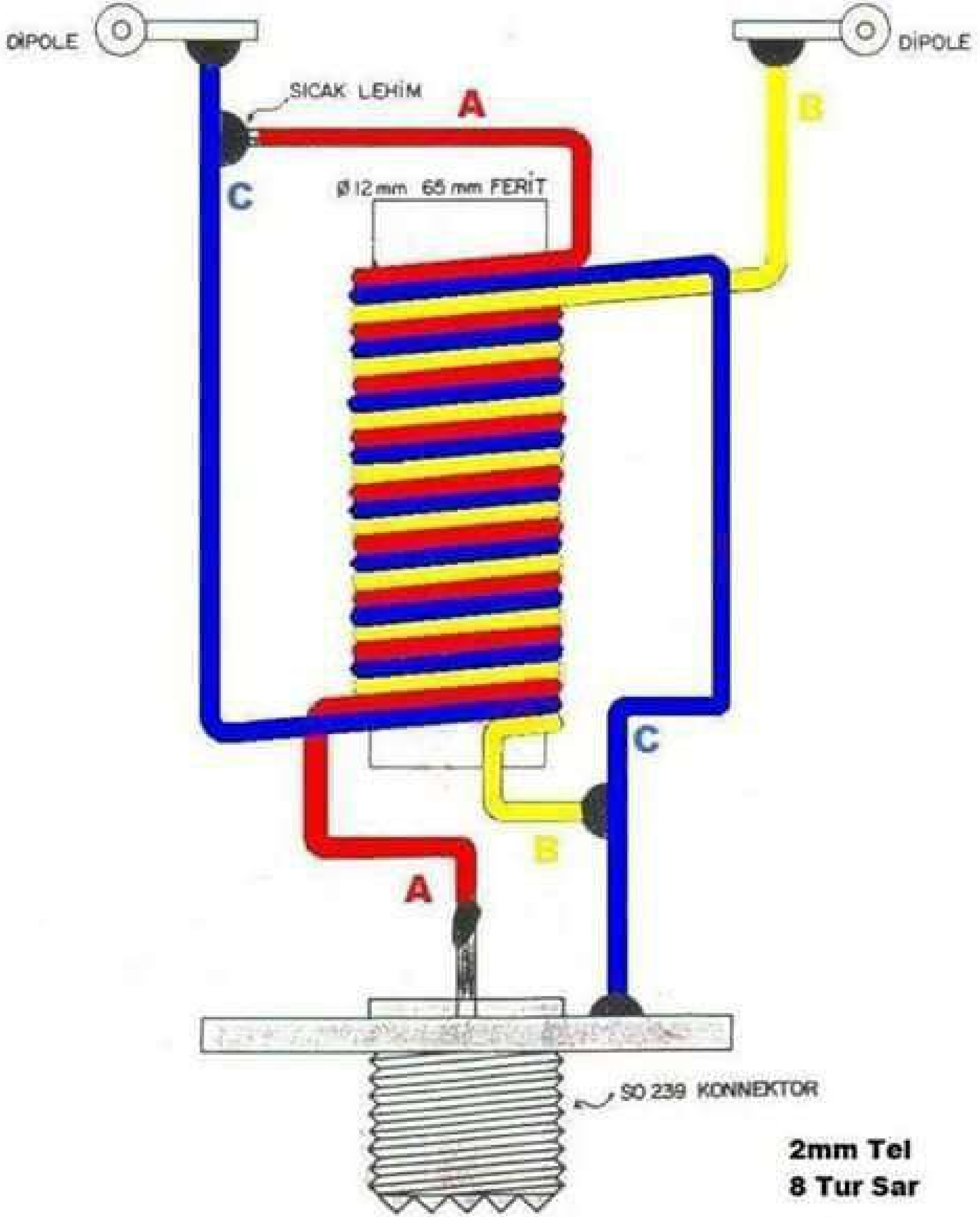
DİPOLE
DİPOLE
SICAK LEHİM
A
B
C
Ø 12 mm 65 mm FERİT
C
B
A
SO 239 KONNEKTOR

2mm Tel
8 Tur Sar

© Azega.com

Ion Harvesting Materials Test Fixture

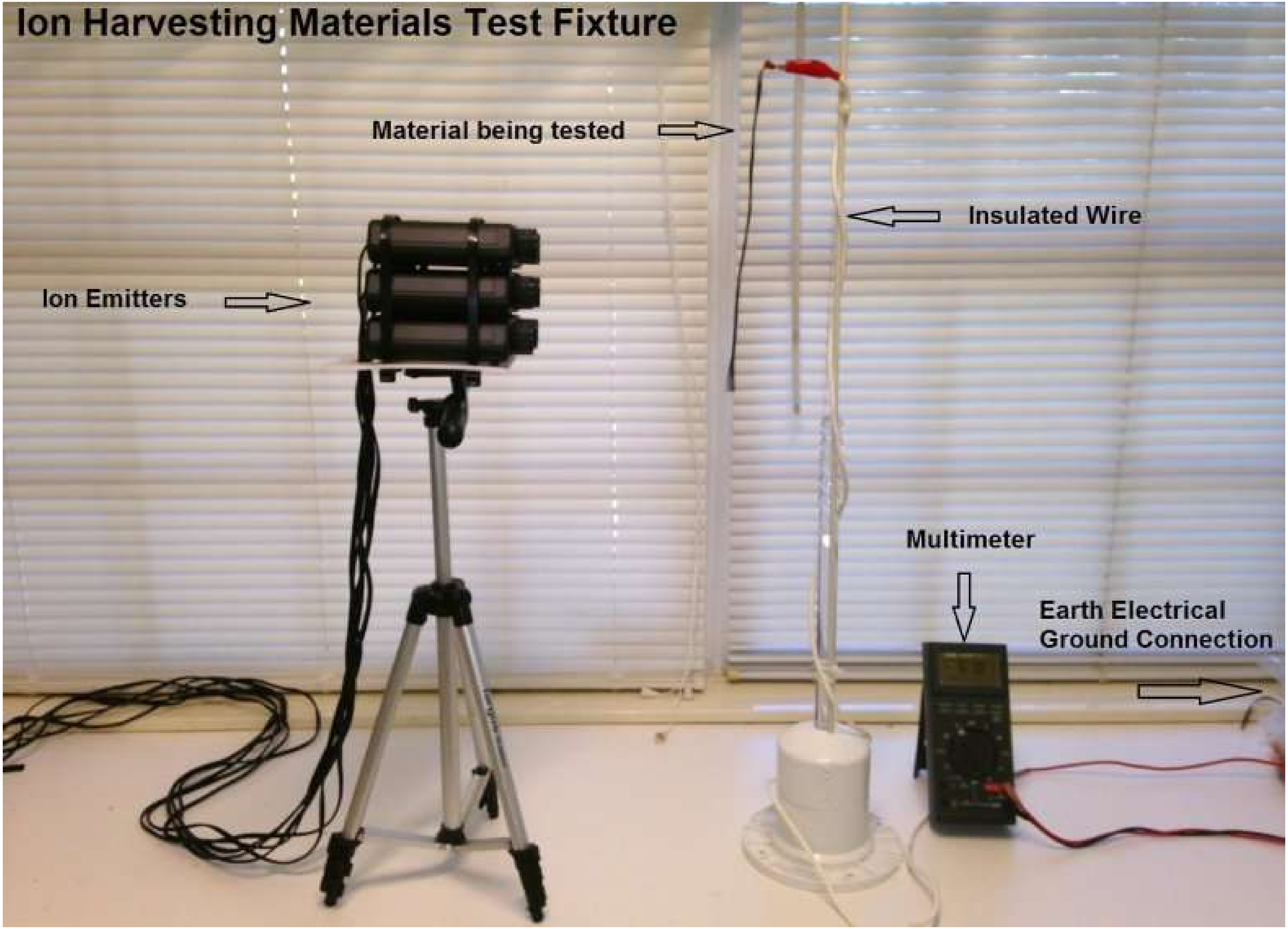
Material being tested
Insulated Wire
Ion Emitters
Multimeter
Earth Electrical
Ground Connection

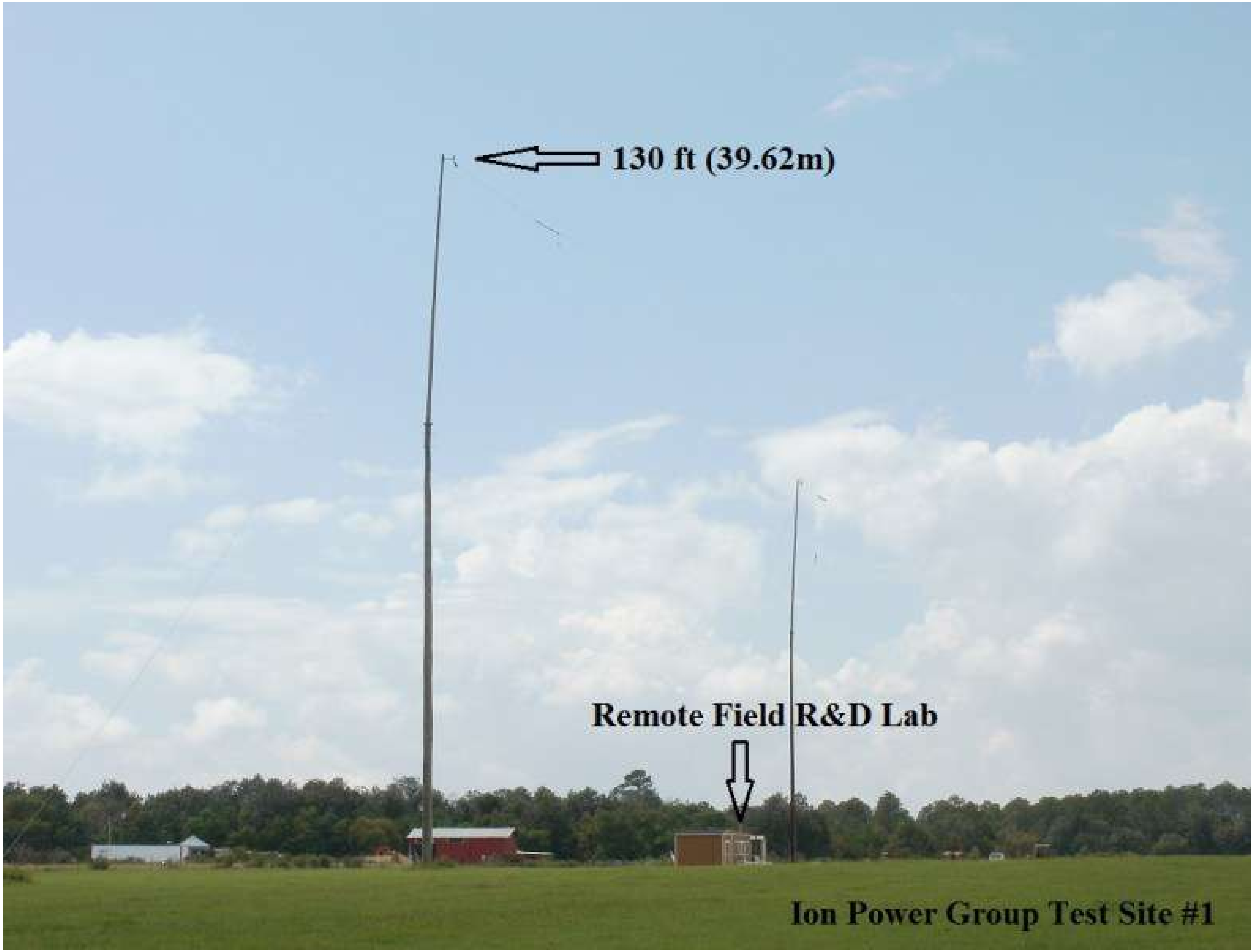
130 ft (39.62m)
Remote Field R&D Lab
Ion Power Group Test Site #1

DANGER
HIGH VOLTAGE
KEEP AWAY
1
2
3
4
5
6
7
8
9
10

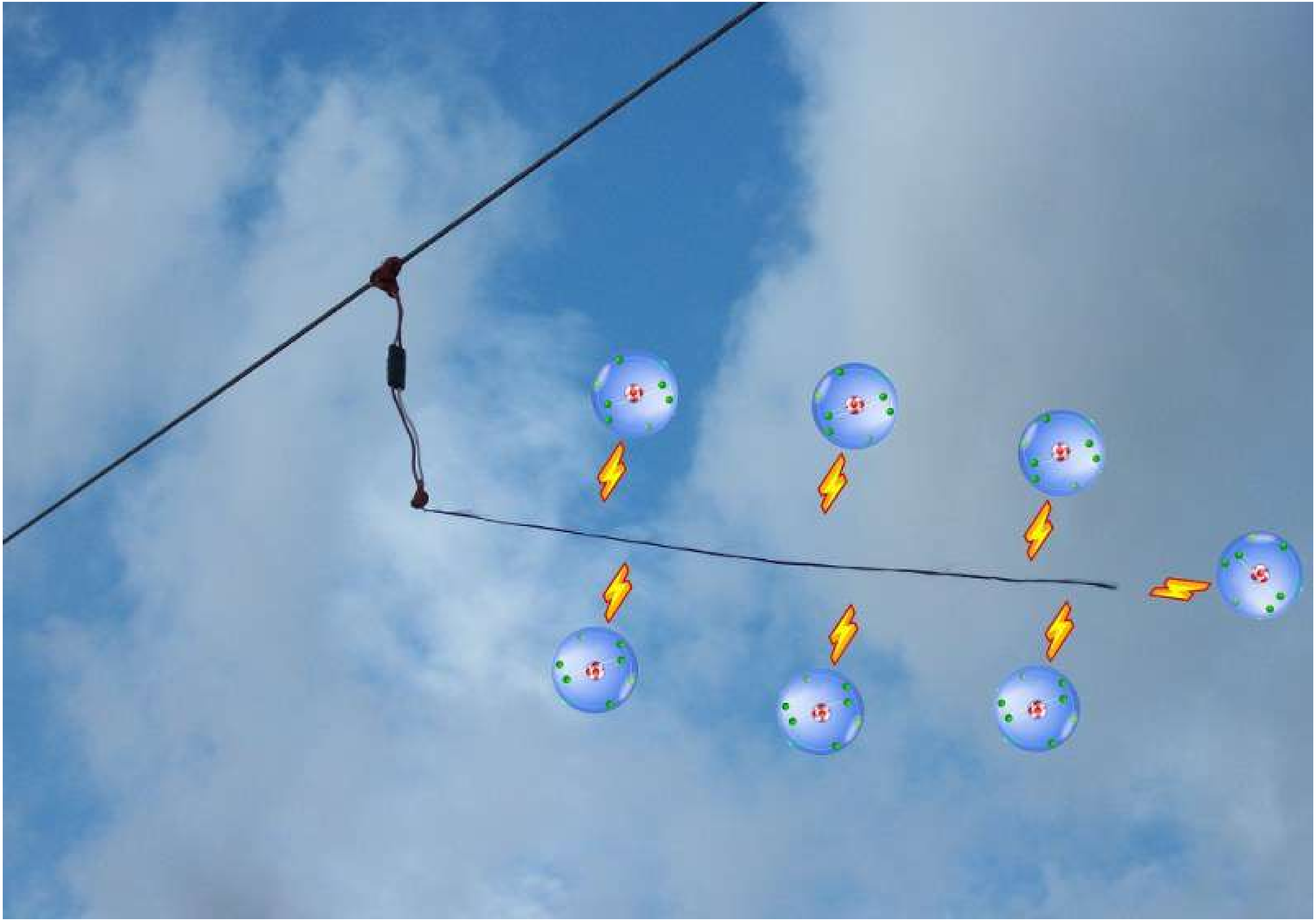

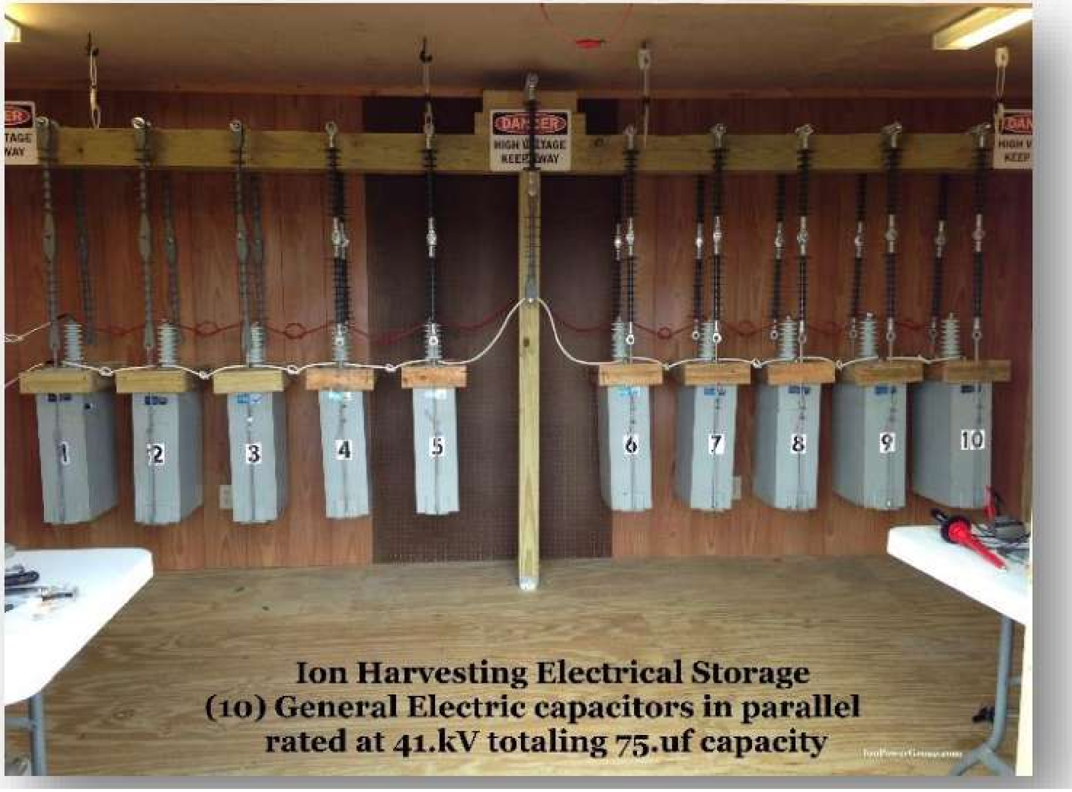
DANGER
HIGH VOLTAGE
KEEP AWAY
1
2
3
4
5
6
7
8
9
10

Ion Harvesting Electrical Storage
(10) General Electric capacitors in parallel
rated at 41.kV totaling 75.uf capacity

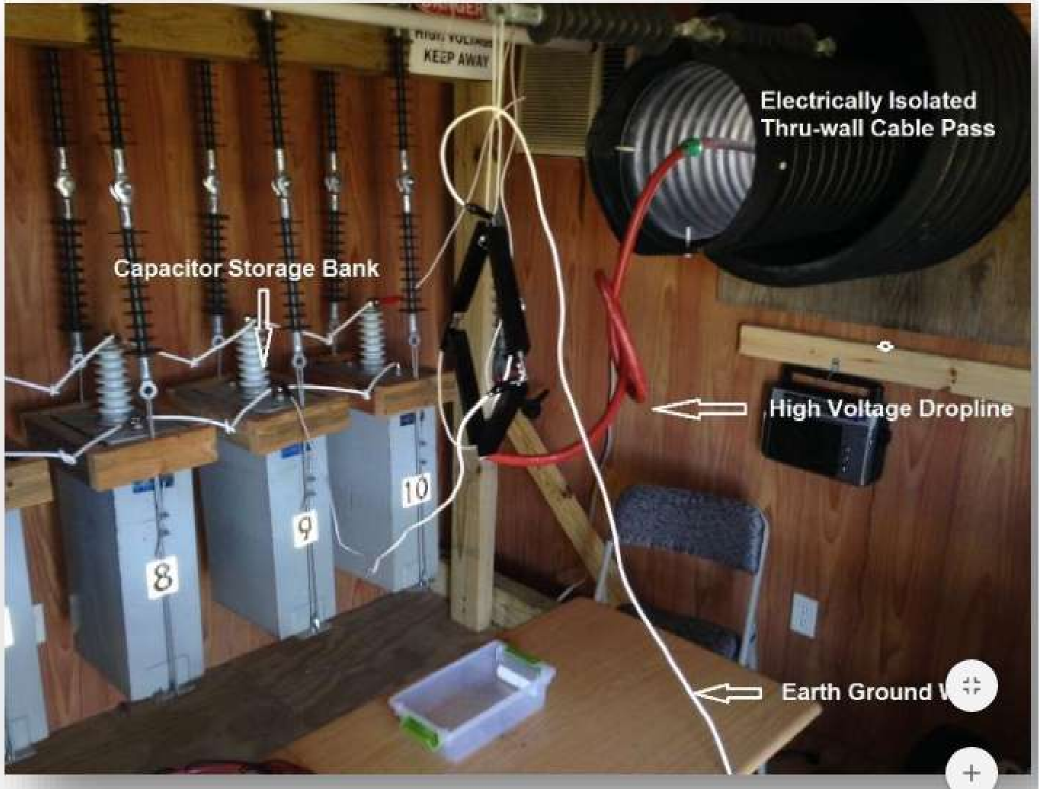
HIGH VOLTAGE
KEEP AWAY
Electrically Isolated
Thru-wall Cable Pass
Capacitor Storage Bank
High Voltage Dropline
Earth Ground W
8
9
10

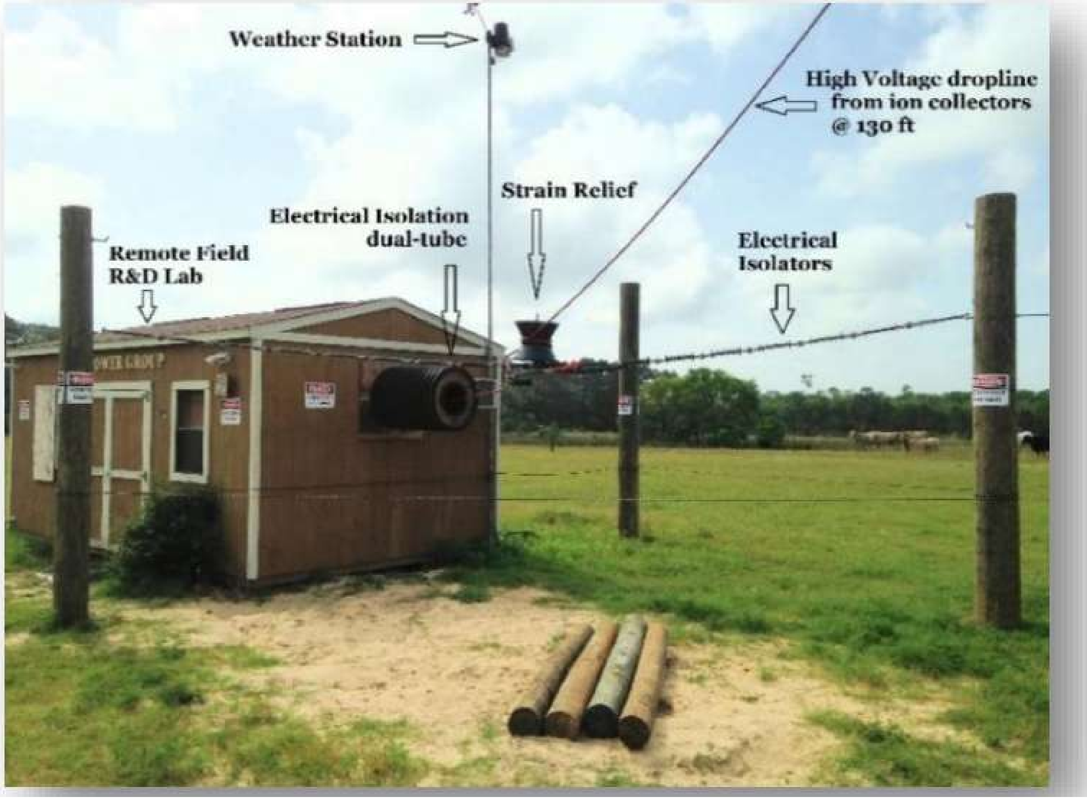
Weather Station
High Voltage dropline from ion collectors @ 130 ft
Strain Relief
Electrical Isolation dual-tube
Remote Field R&D Lab
Electrical Isolators

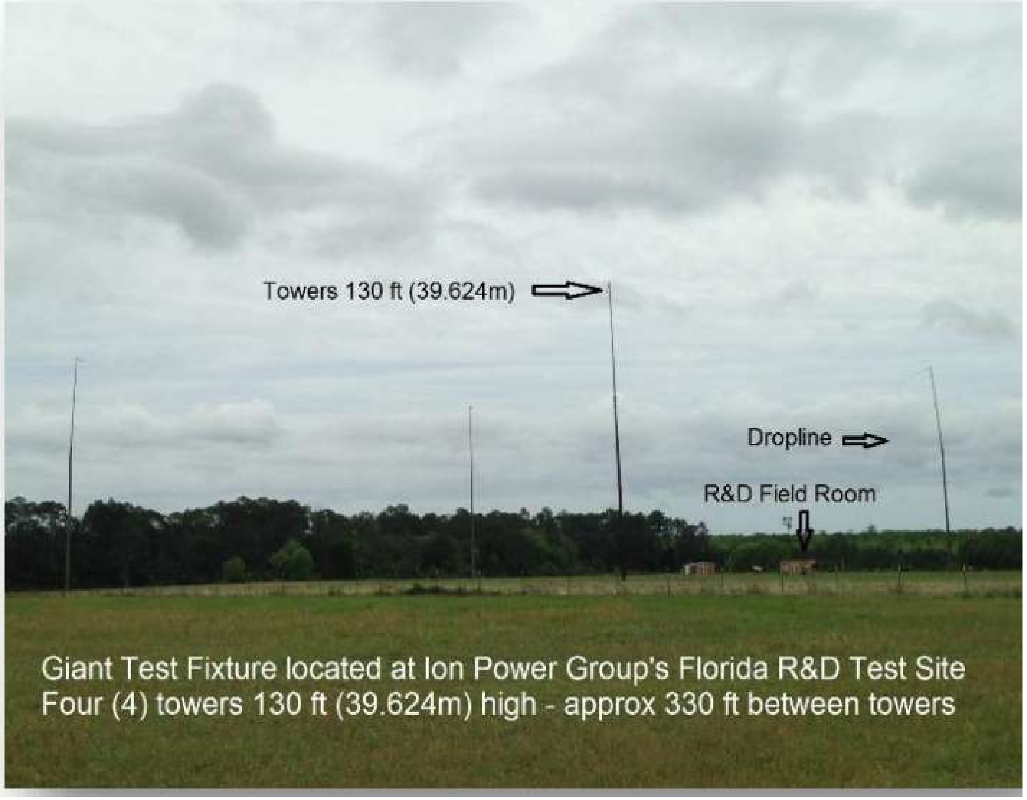
Towers 130 ft (39.624m)
Dropline
R&D Field Room

Giant Test Fixture located at Ion Power Group's Florida R&D Test Site
Four (4) towers 130 ft (39.624m) high - approx 330 ft between towers

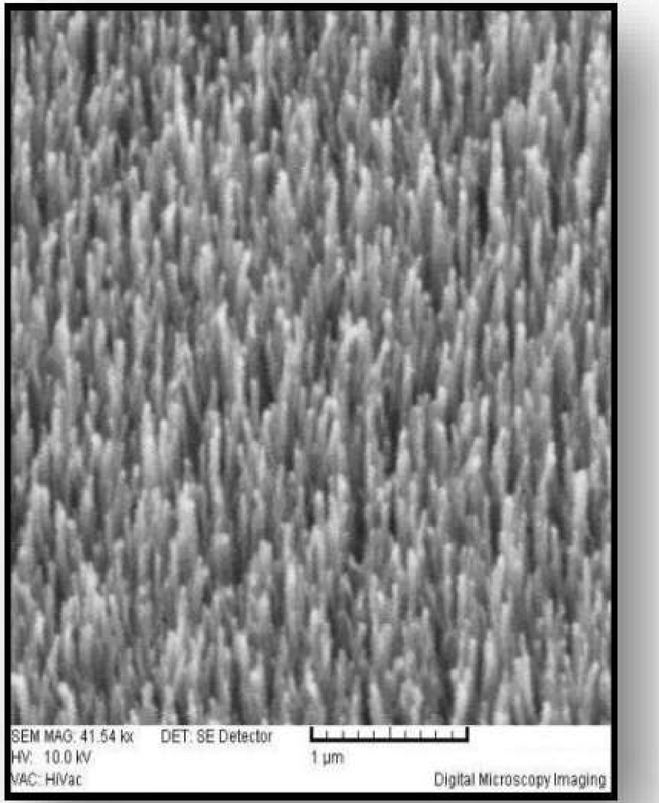
SEM MAG: 41.54 kx
DET: SE Detector
HV: 10.0 kV
VAC: HiVac
1 µm
Digital Microscopy Imaging

Ion Power Group's new method of coupling to atmospheric electricity is vastly different from all previous techniques by virtue of our patented breakthrough revealing that carbon nanomaterials such as **Graphite (and Graphene) microscopic shown at left, macroscopic shown at right are significantly more effective at coupling to airborne charge carriers (ions) than metal.** The use of carbon based nanomaterials distinguishes Ion Power Group from all other researchers.

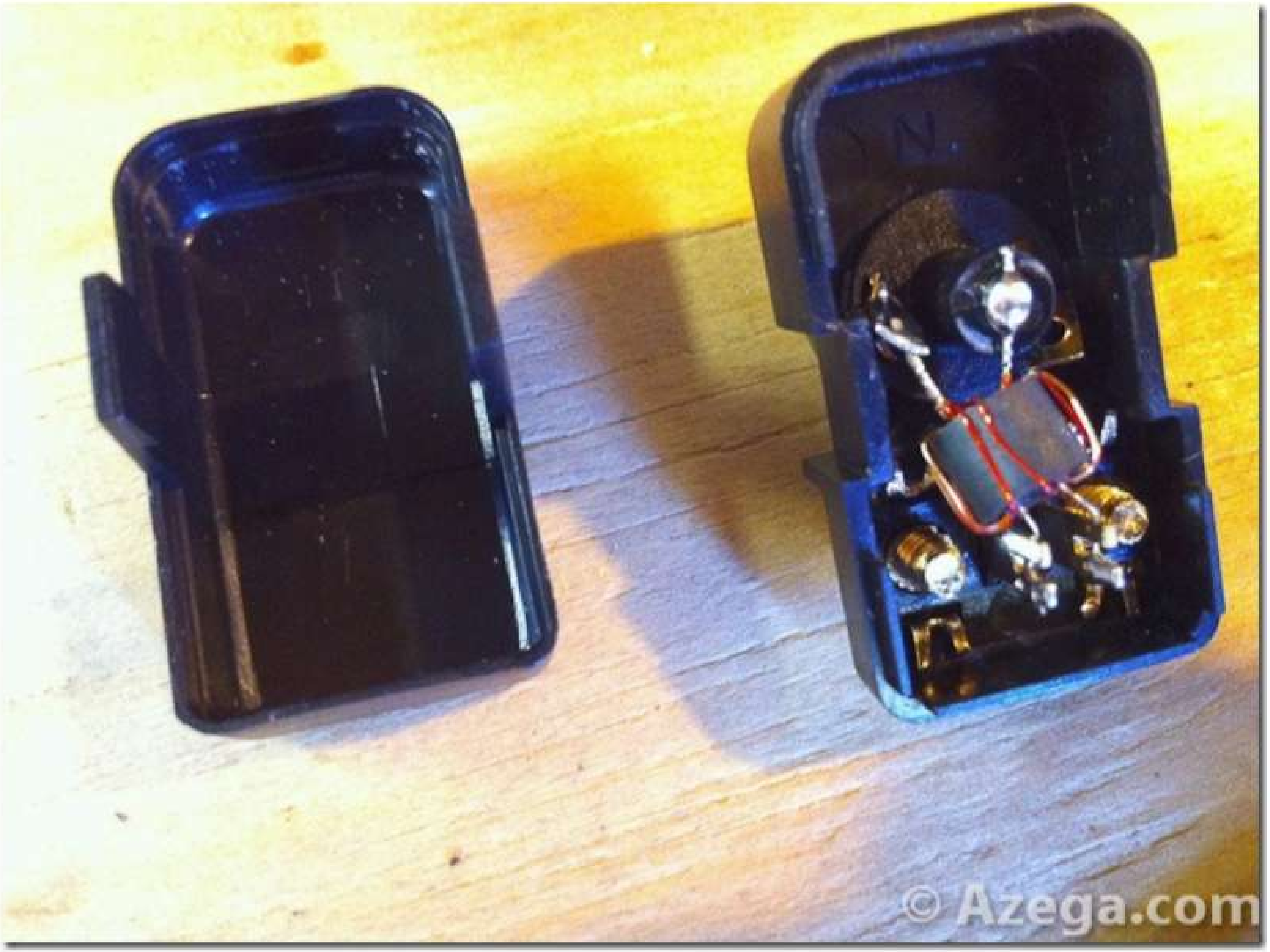
© Azega.com

I cut a hole in the side of the balun then cut the end off a piece of coax cable I had and soldered it to the contacts of the coax connection.

EXAMPLE: W1JR (Reisert) Cross-Winding Method

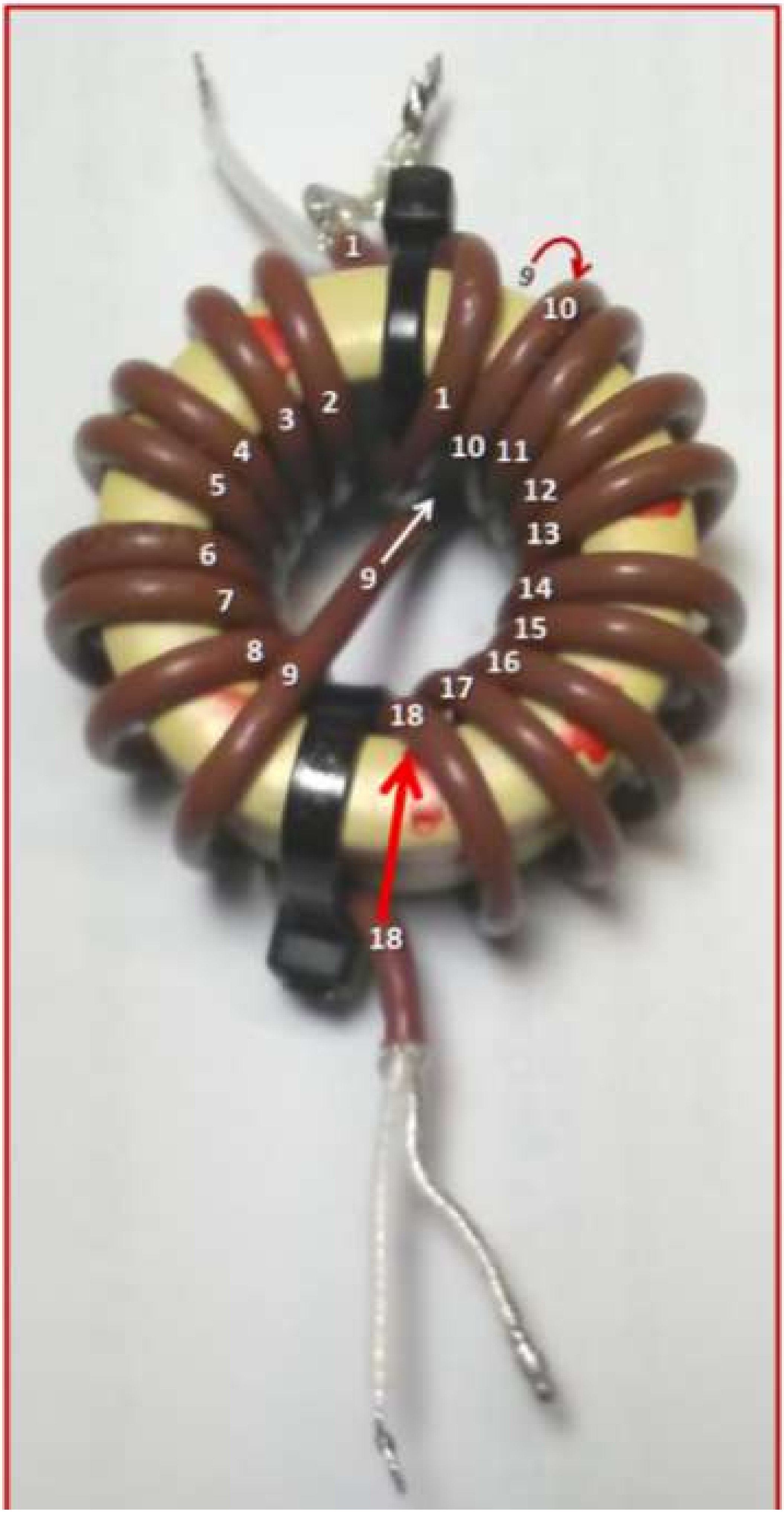
1
9
10
1
2
3
10
11
4
5
12
13
6
9
7
14
15
8
9
16
17
18
18

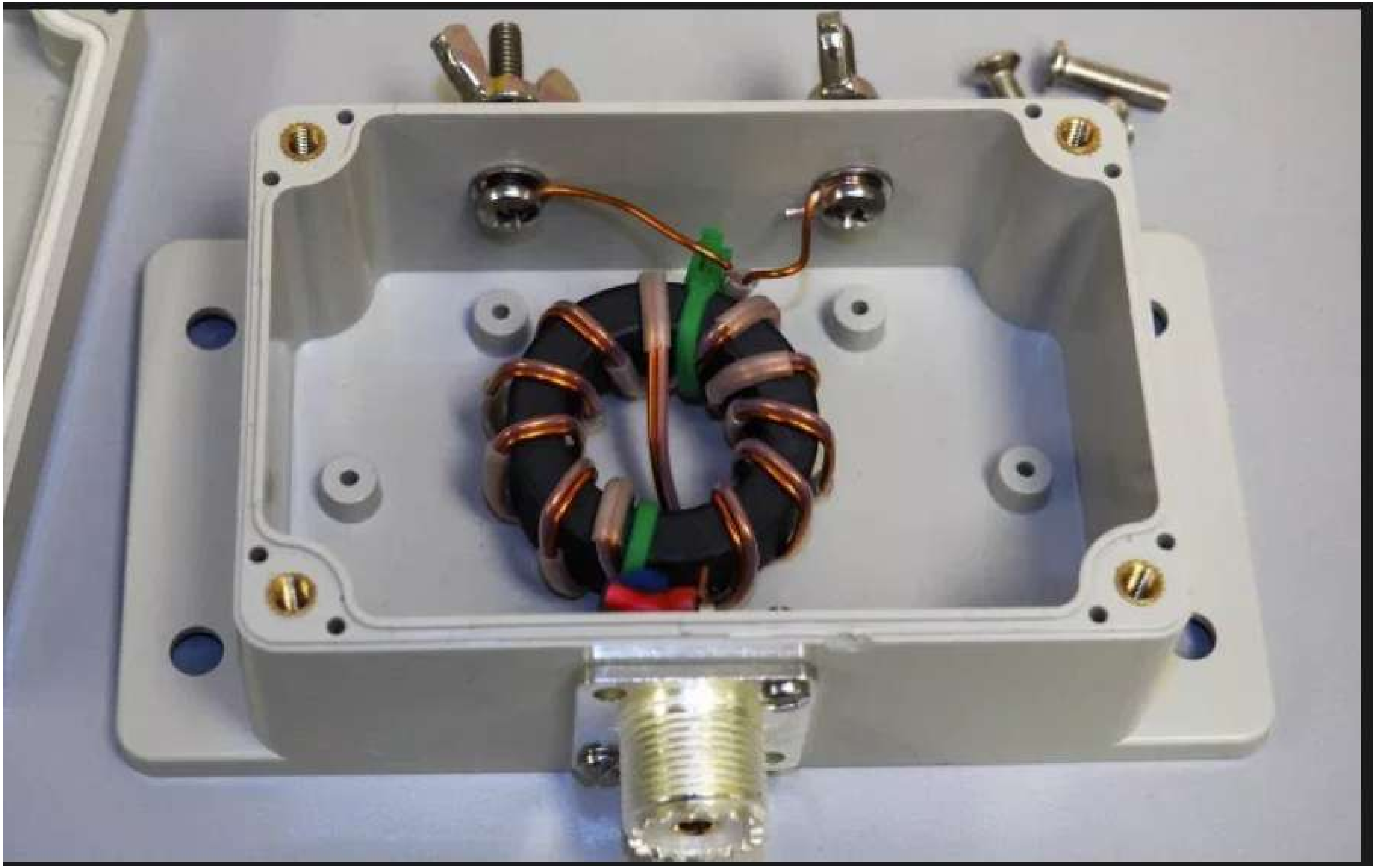

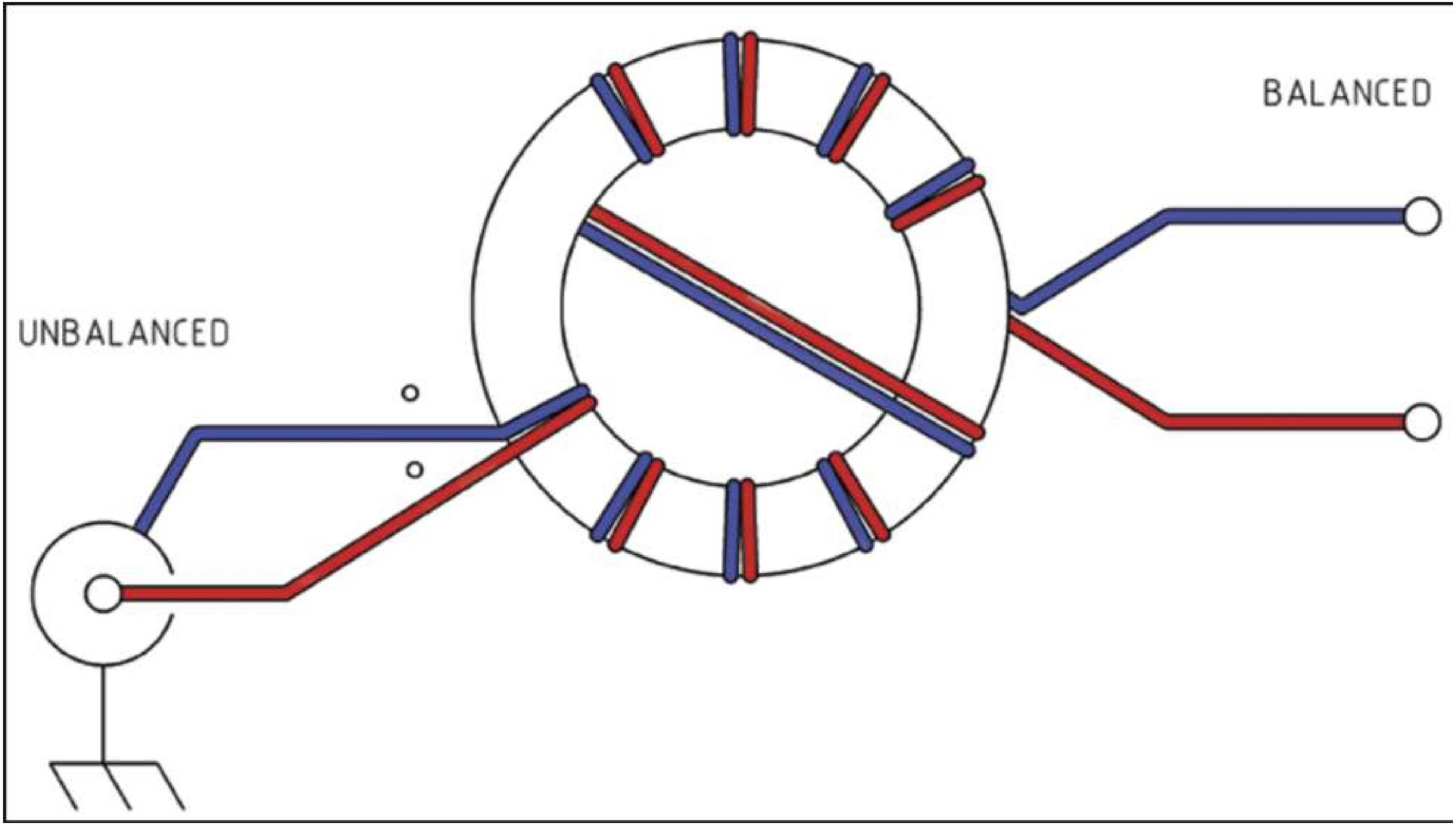
BALANCED
UNBALANCED

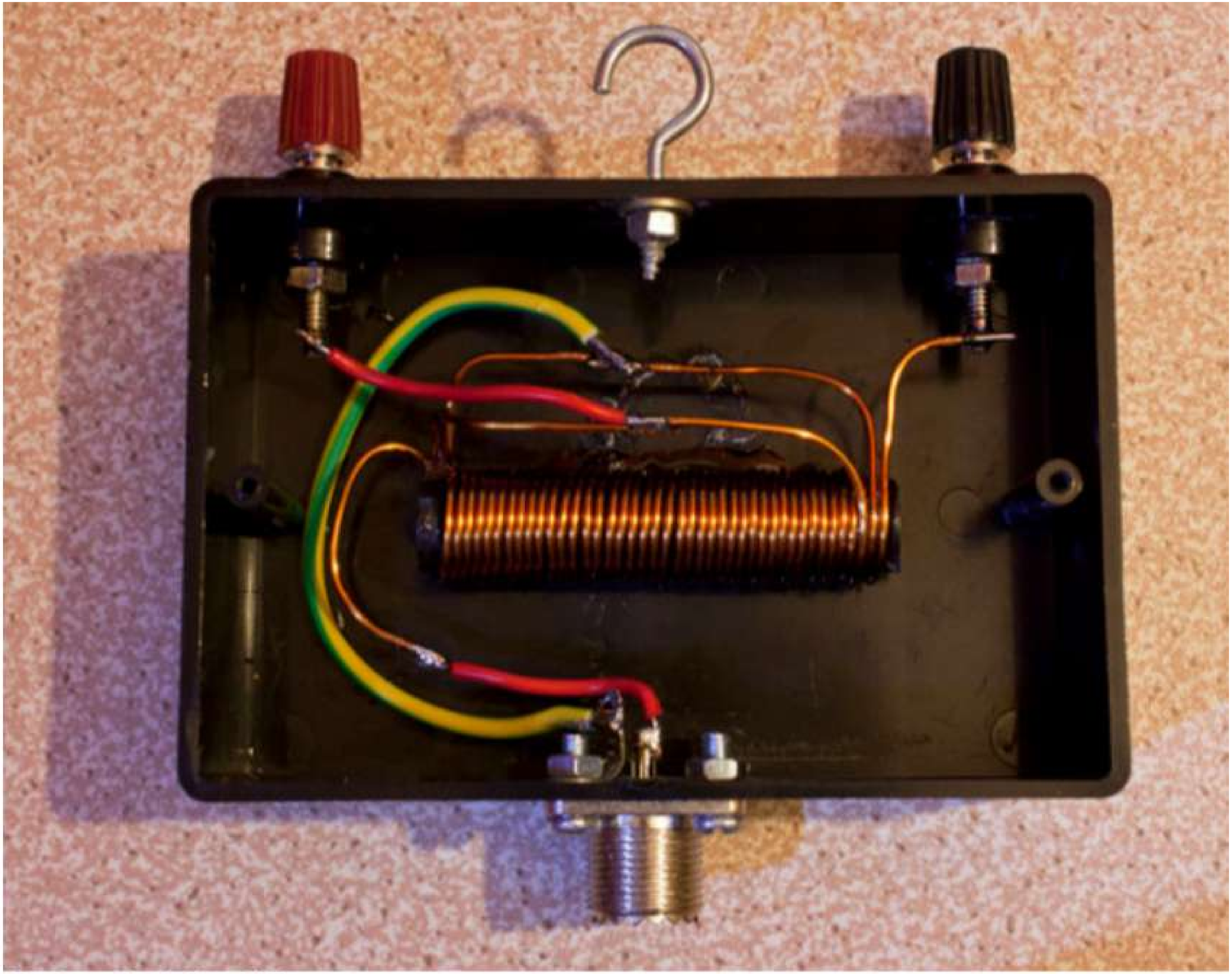

Balun 1: 1 for HF

1:1 Voltage balun ("Ruthroff")

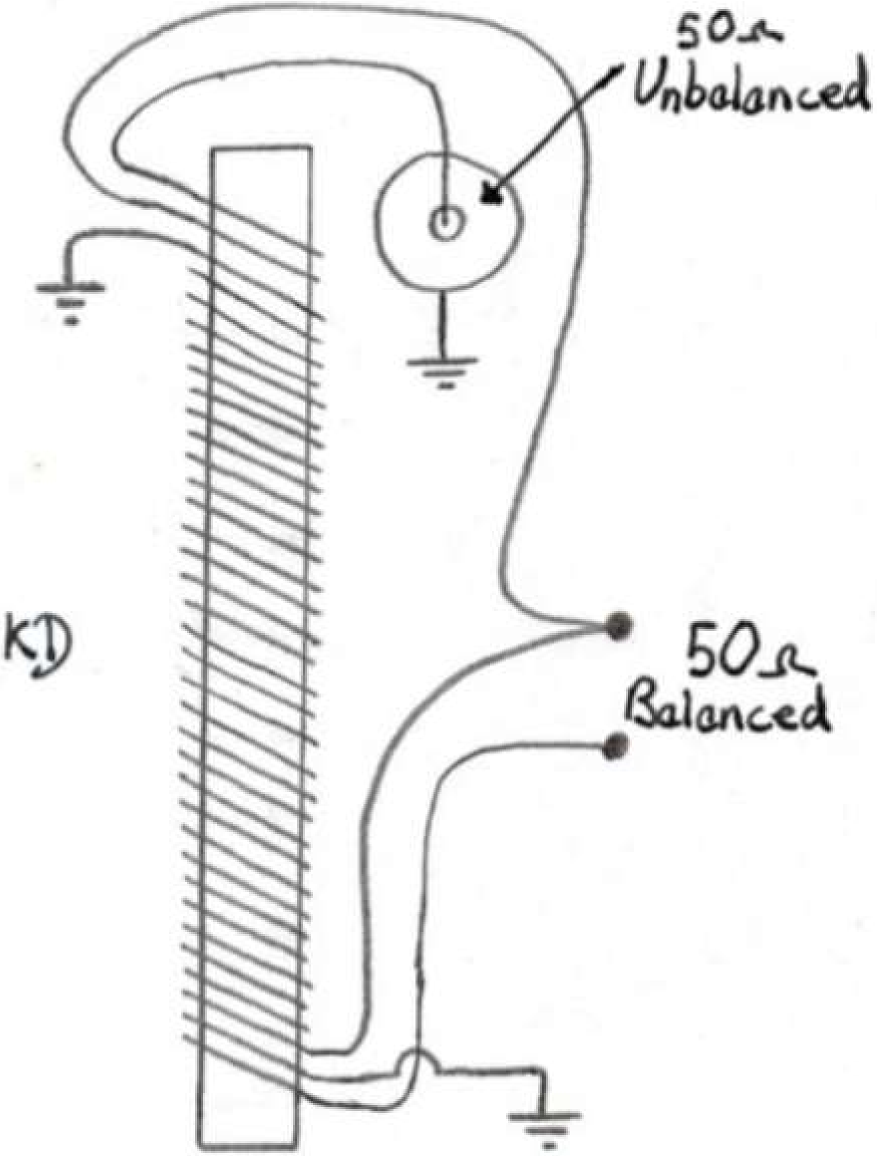
50Ω
Unbalanced
50Ω
Balanced

M0UKD

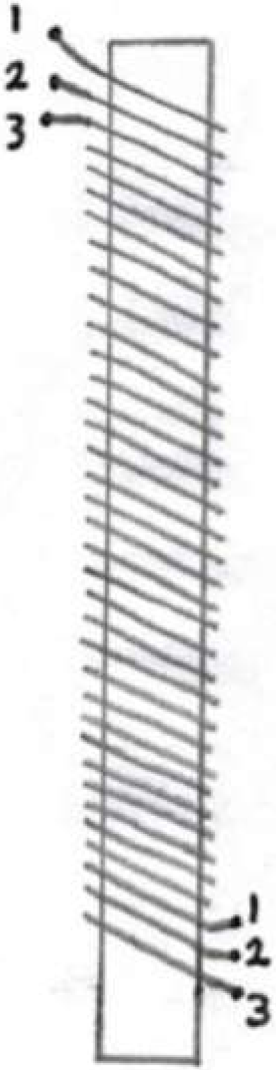

10 - 15 trifilar turns on ferrite rod.

1:1 ferrite rod voltage balun

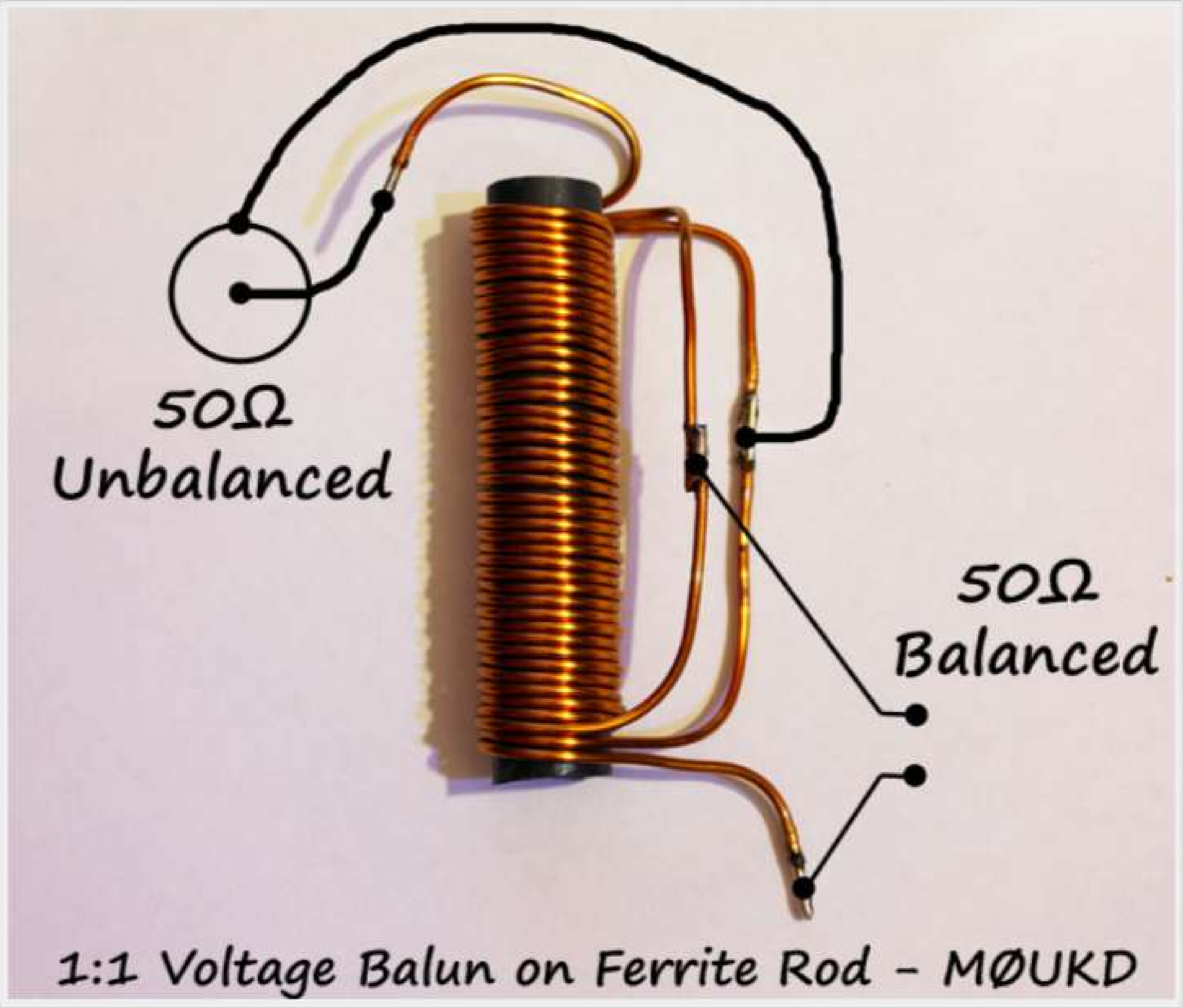
50Ω
Unbalanced
50Ω
Balanced
1:1 Voltage Balun on Ferrite Rod - MØUKD

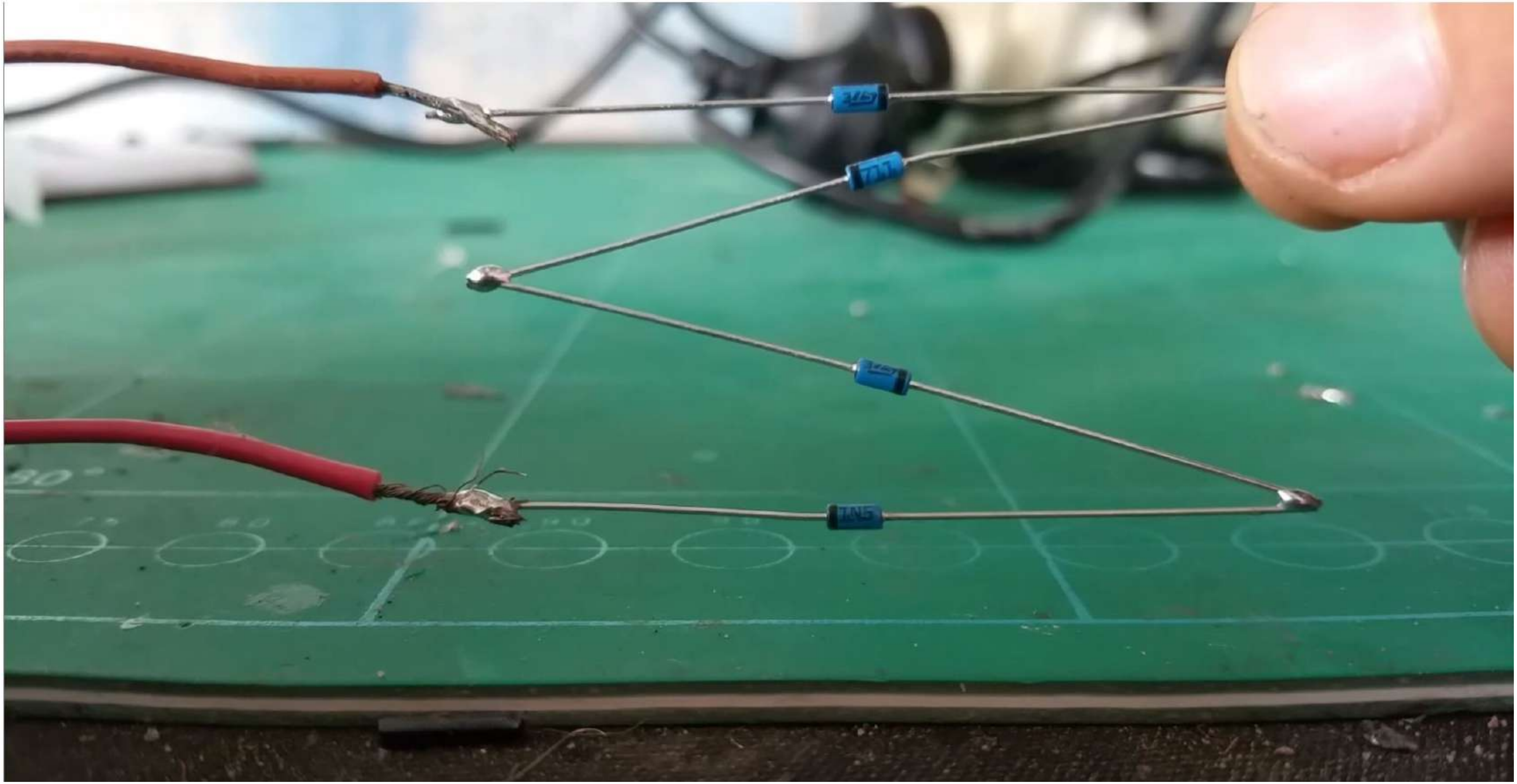

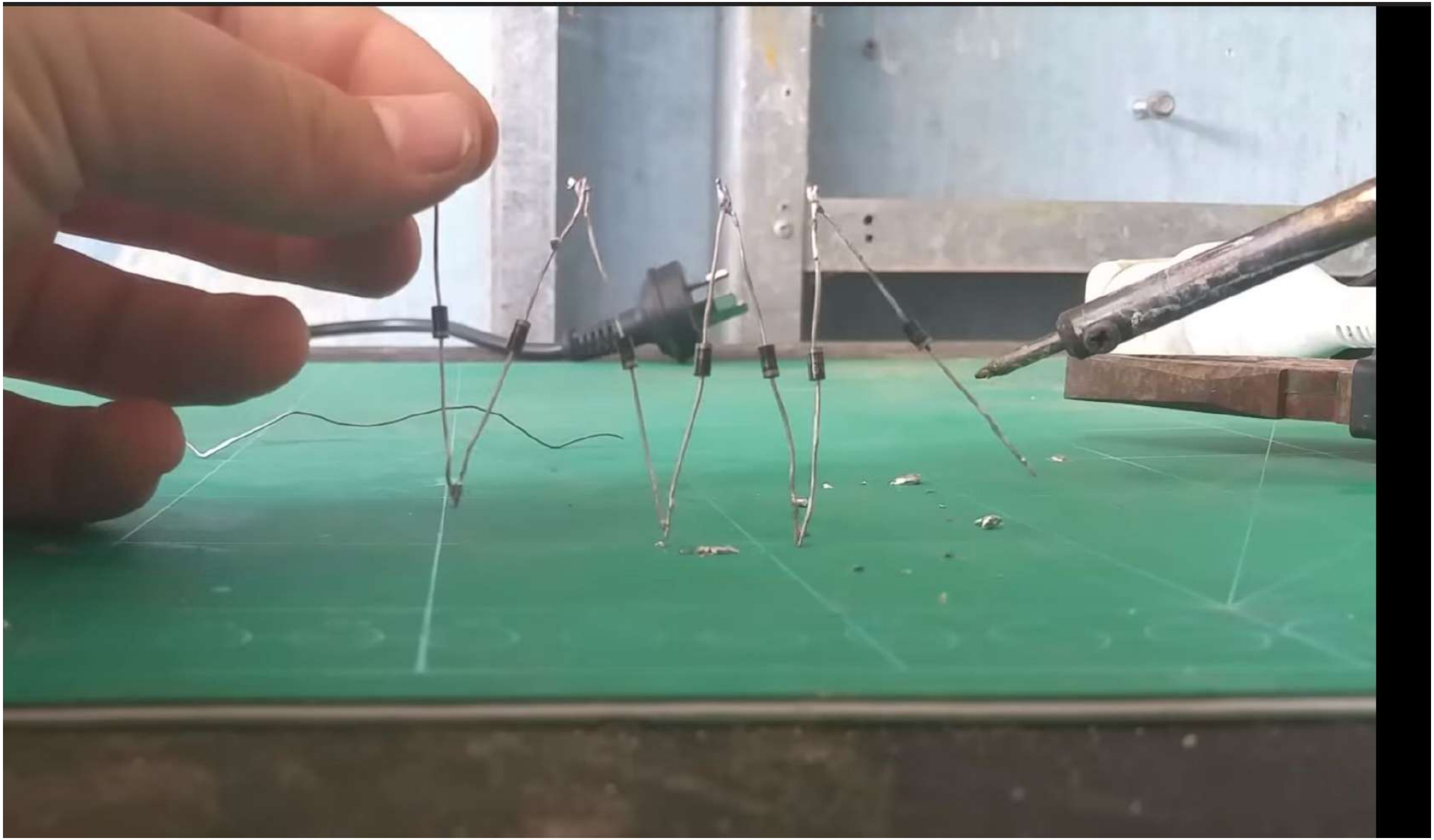

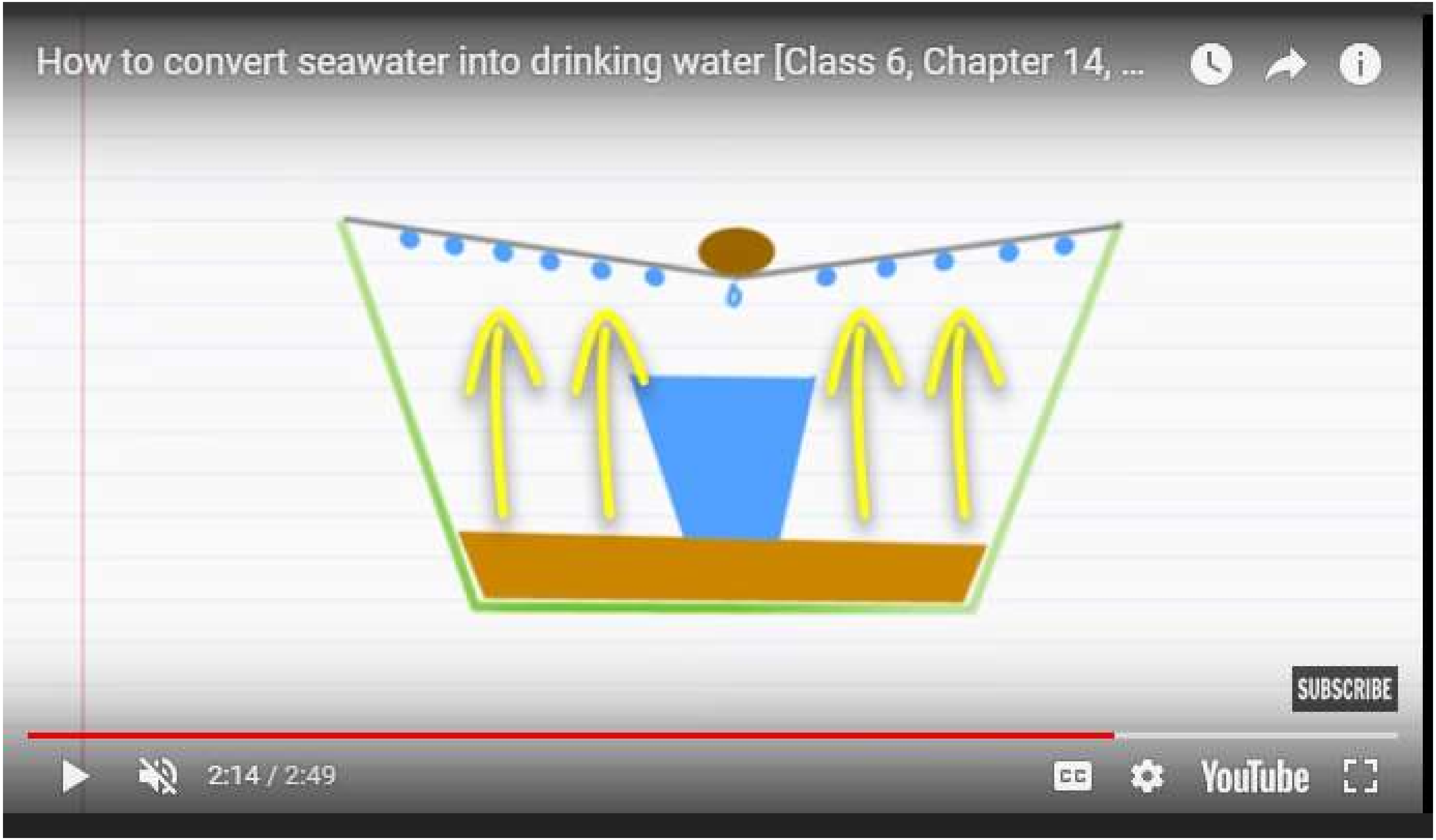

How to convert seawater into drinking water [Class 6, Chapter 14, ...

SUBSCRIBE
2:14 / 2:49
CC
YouTube

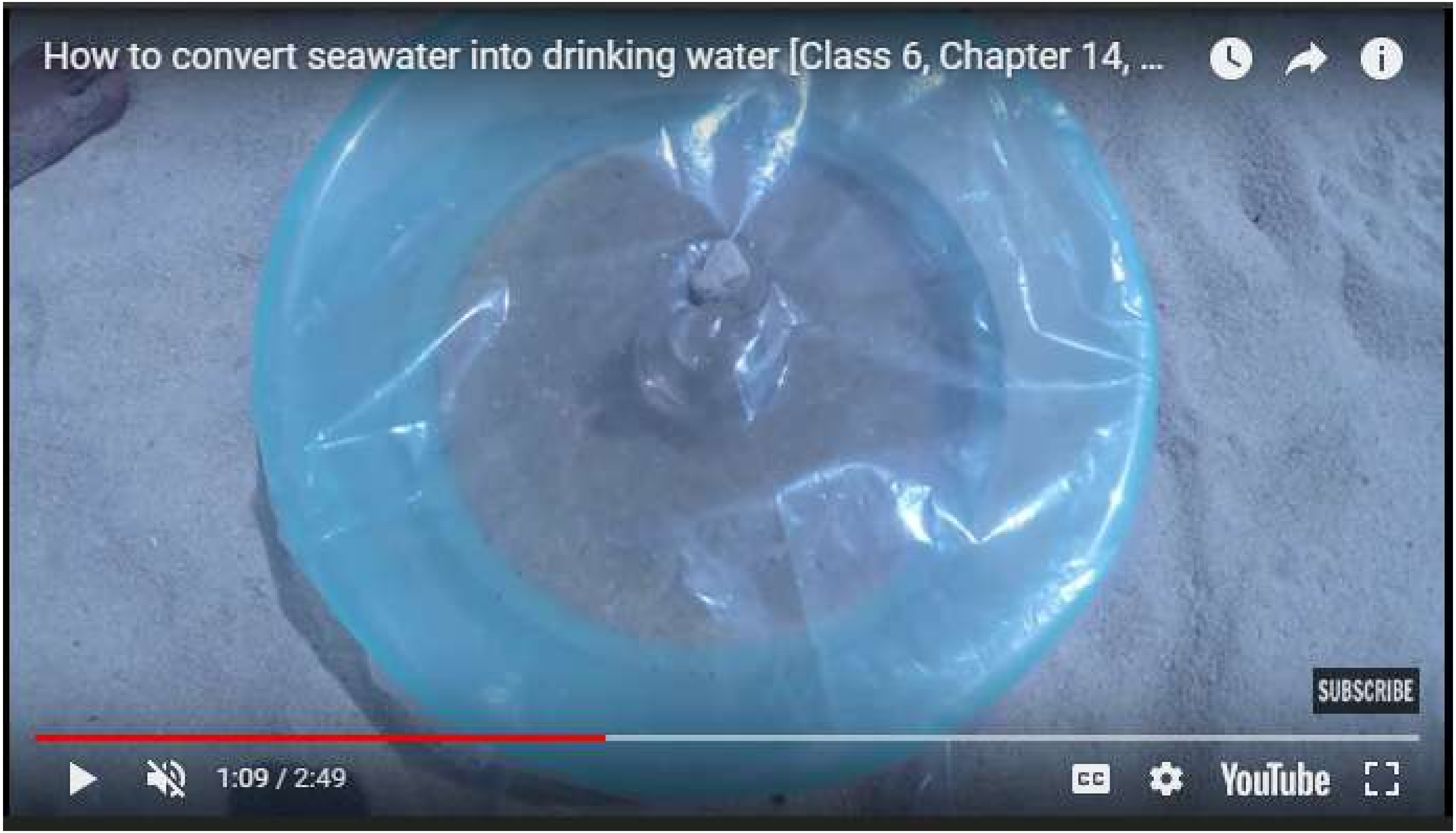
How to convert seawater into drinking water [Class 6, Chapter 14, …
SUBSCRIBE
1:09 / 2:49
YouTube

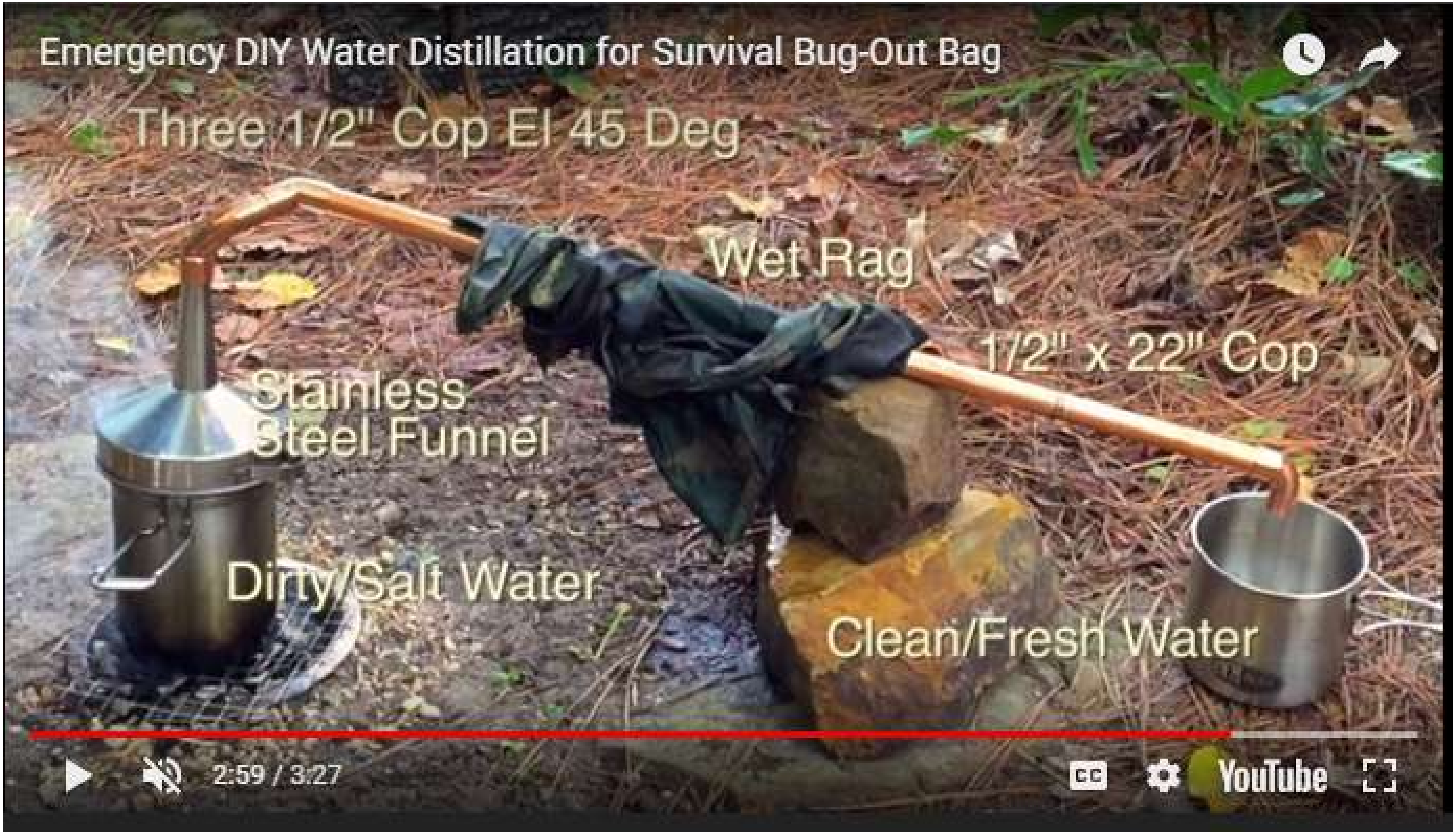
Emergency DIY Water Distillation for Survival Bug-Out Bag

Three 1/2" Cop El 45 Deg
Wet Rag
1/2" x 22" Cop
Stainless
Steel Funnel
Dirty/Salt Water
Clean/Fresh Water
2:59 / 3:27
CC
YouTube

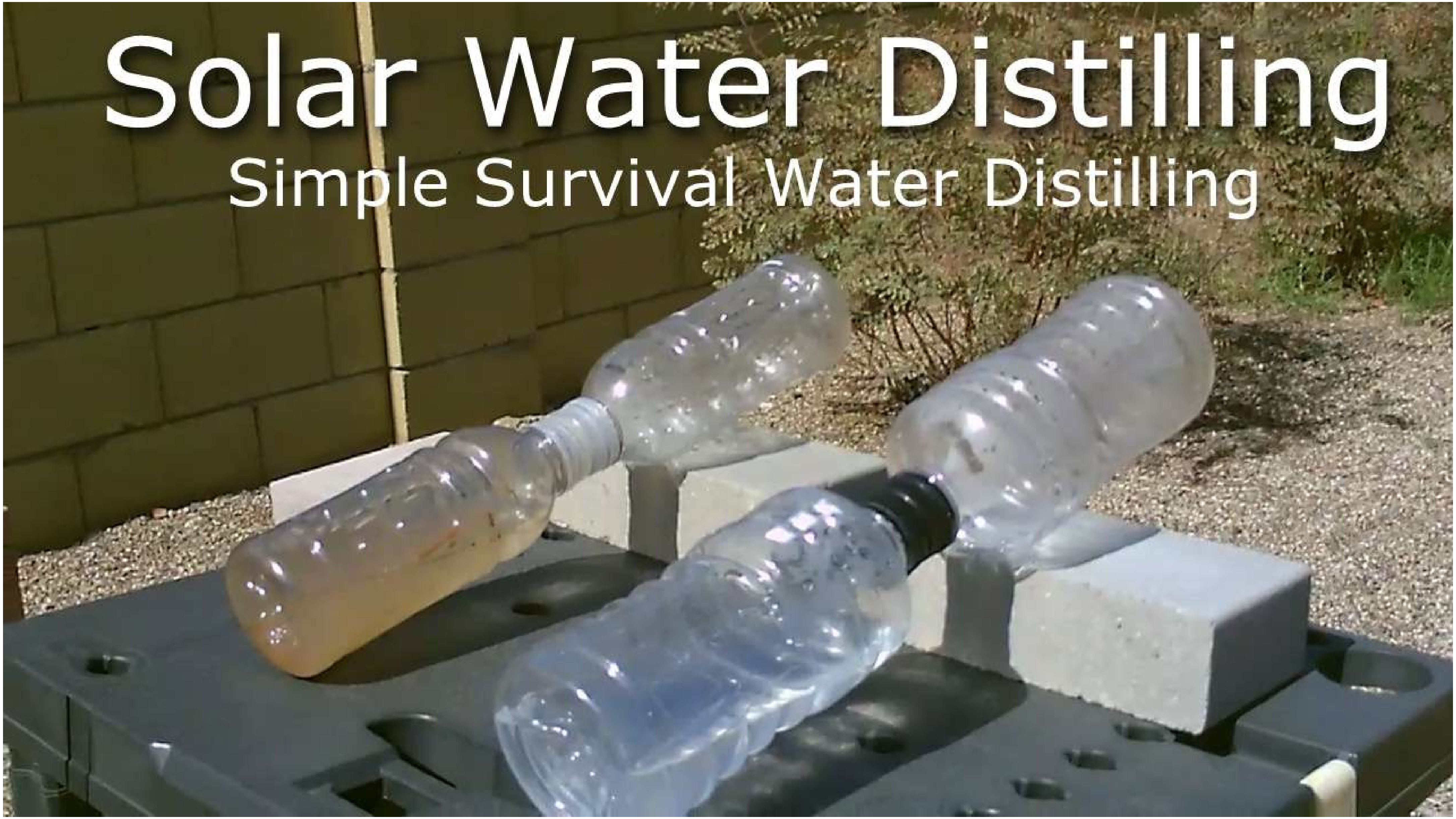

Solar Water Distilling

Simple Survival Water Distilling

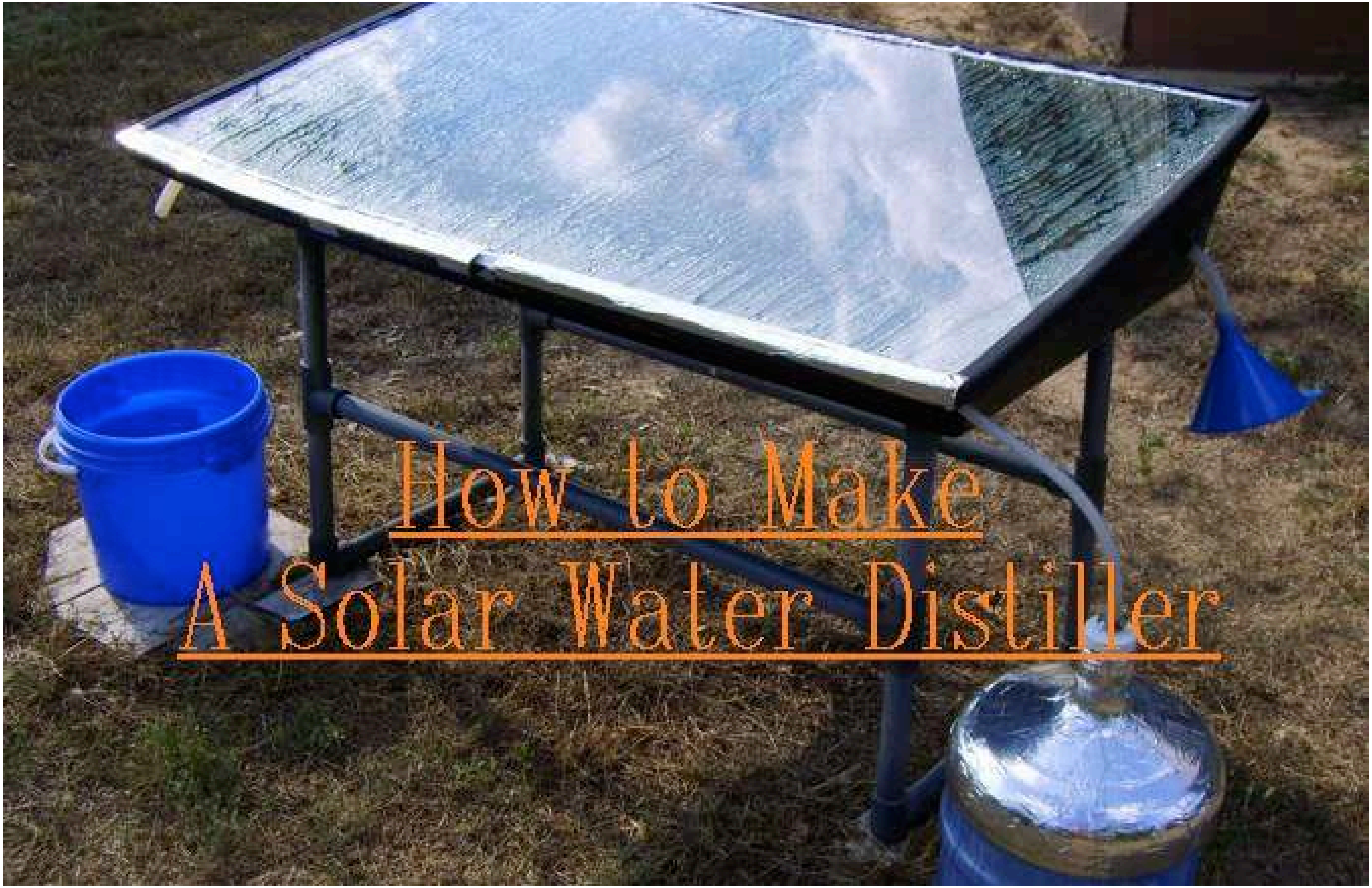

How to Make A Solar Water Distiller

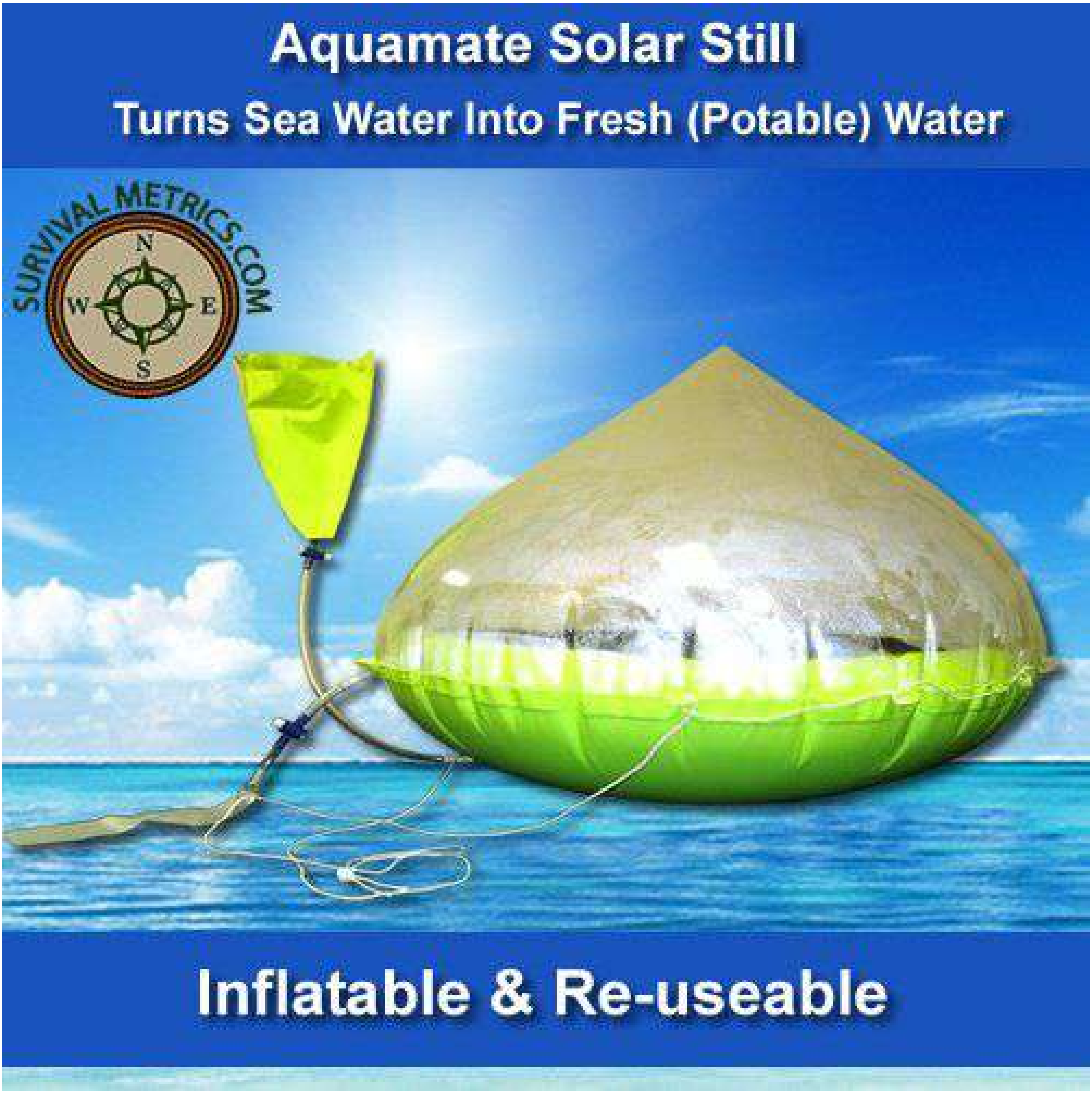
Aquamate Solar Still
Turns Sea Water Into Fresh (Potable) Water
SURVIVALMETRICS.COM
N
W
E
S
Inflatable & Re-useable

Solar Still

Turns sea water or impure water into drinking water

1 Remove from pack and retain pack to top up salt water in still.

2 Lay out Solar Still and inflate valve Ⓐ on floatation ring by mouth or by liferaft or dinghy pump.

3 Inflate clear dome to obtain shape using valve Ⓔ in base. Do not over inflate. It may be necessary to manually regulate dome pressure during high temperatures.

4 Attach tapered bag Ⓑ and attachment cord Ⓒ to boat or liferaft.

5 Pour 5 litres of sea water into tapered bag Ⓑ using pack which holds 5 litres (approx). Close valve Ⓕ after filling.

6 Ensure both valves Ⓖ are open in reservoir feed tube and drinkable water will start to collect in reservoir Ⓓ. Air may need to be squeezed out to assist collection of water.

7 Ensure the Solar Still is kept in direct sunlight whenever possible.

8 Re-fill Solar Still and inflate/deflate dome as required.

9 Daily – Agitate Solar Still to dampen fabric.

10 To drain unwanted water open valve Ⓔ in base.

11 Shelter Solar Still during bad weather.

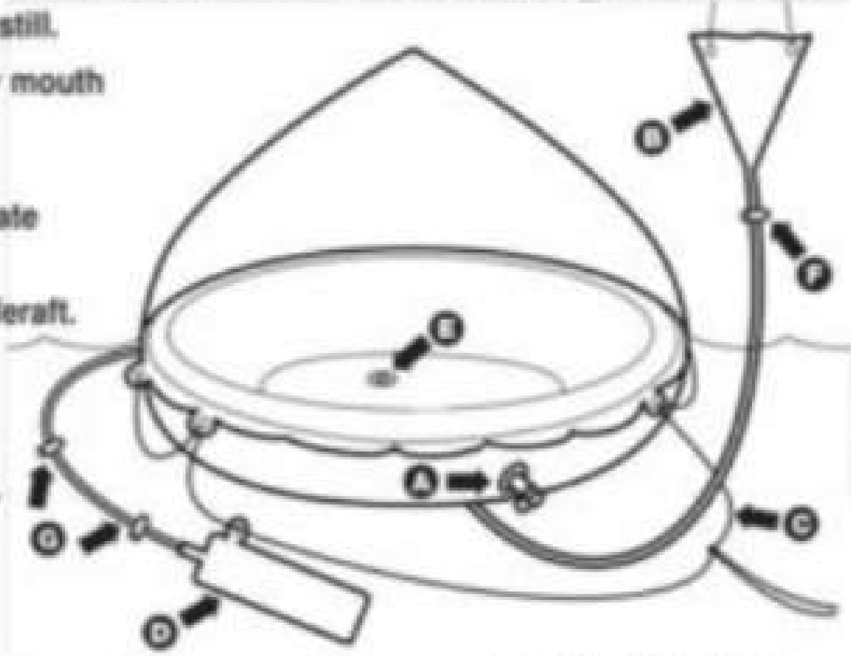

IMPORTANT
IF INNER BLACK MATERIAL SALTS UP FILL WITH SEA WATER AND AGITATE TO CLEAR SALT CRYSTALS AND DRAIN AWAY. PROCESS DRINKING WATER AS INSTRUCTIONS UNTIL SALT IS NO LONGER PRESENT. IN NO CIRCUMSTANCES DRINK WATER THAT TASTES OF SALT.

NATO Stock No. 4610-66-144-2646

AQUAMATE
PRODUCTS

P.O. Box 6032 Dunmow CM6 3AS U.K. Tel: +44 (0)1371 830 216 Fax: 831 733 e-mail: aquamatesales@aol.com

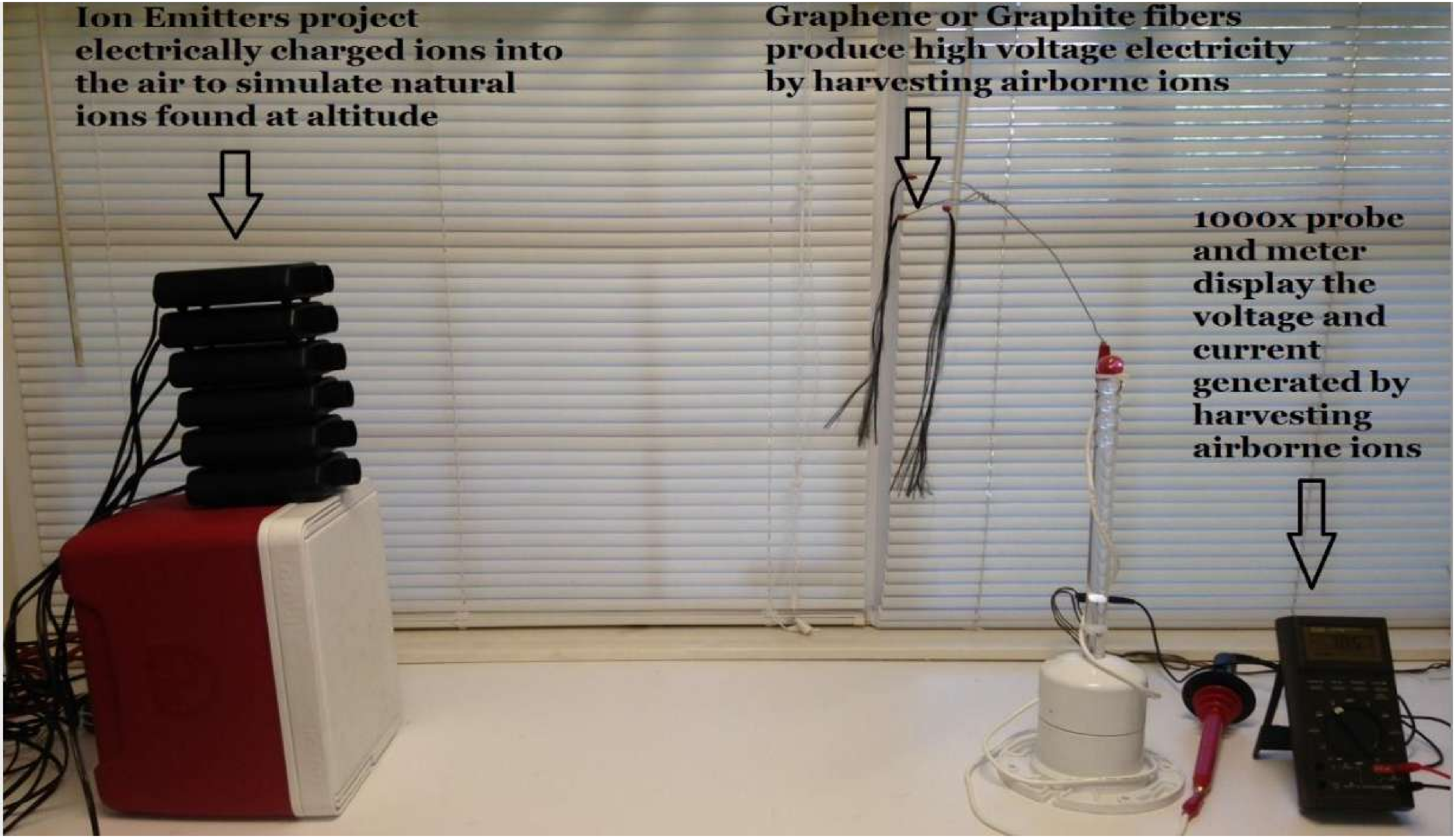

Ion Emitters project electrically charged ions into the air to simulate natural ions found at altitude

Graphene or Graphite fibers produce high voltage electricity by harvesting airborne ions

1000x probe and meter display the voltage and current generated by harvesting airborne ions

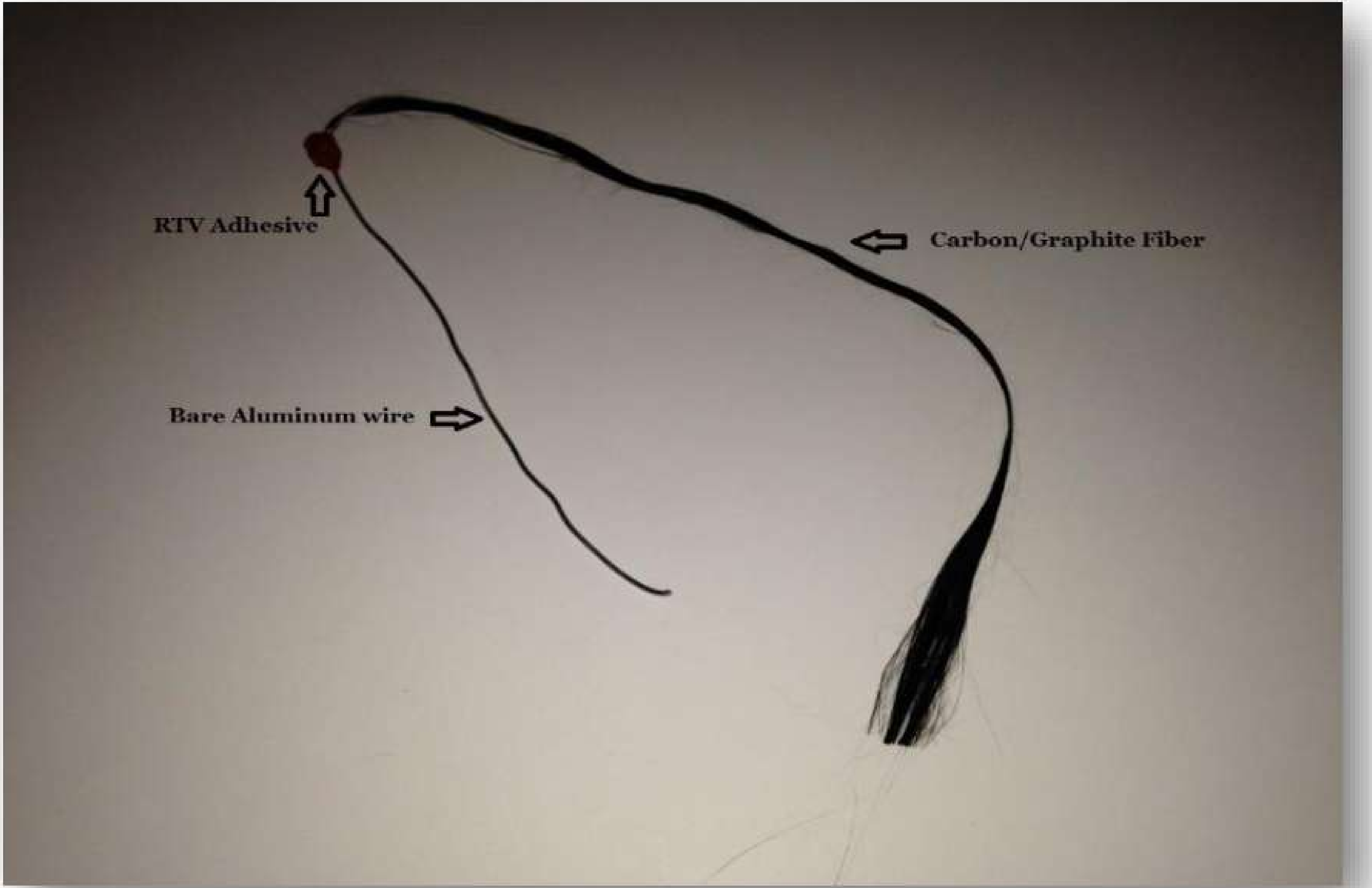
RTV Adhesive
Carbon/Graphite Fiber
Bare Aluminum wire

Image of a patented Ion Collector made of carbon/graphite weighing approximately 1 ounce.

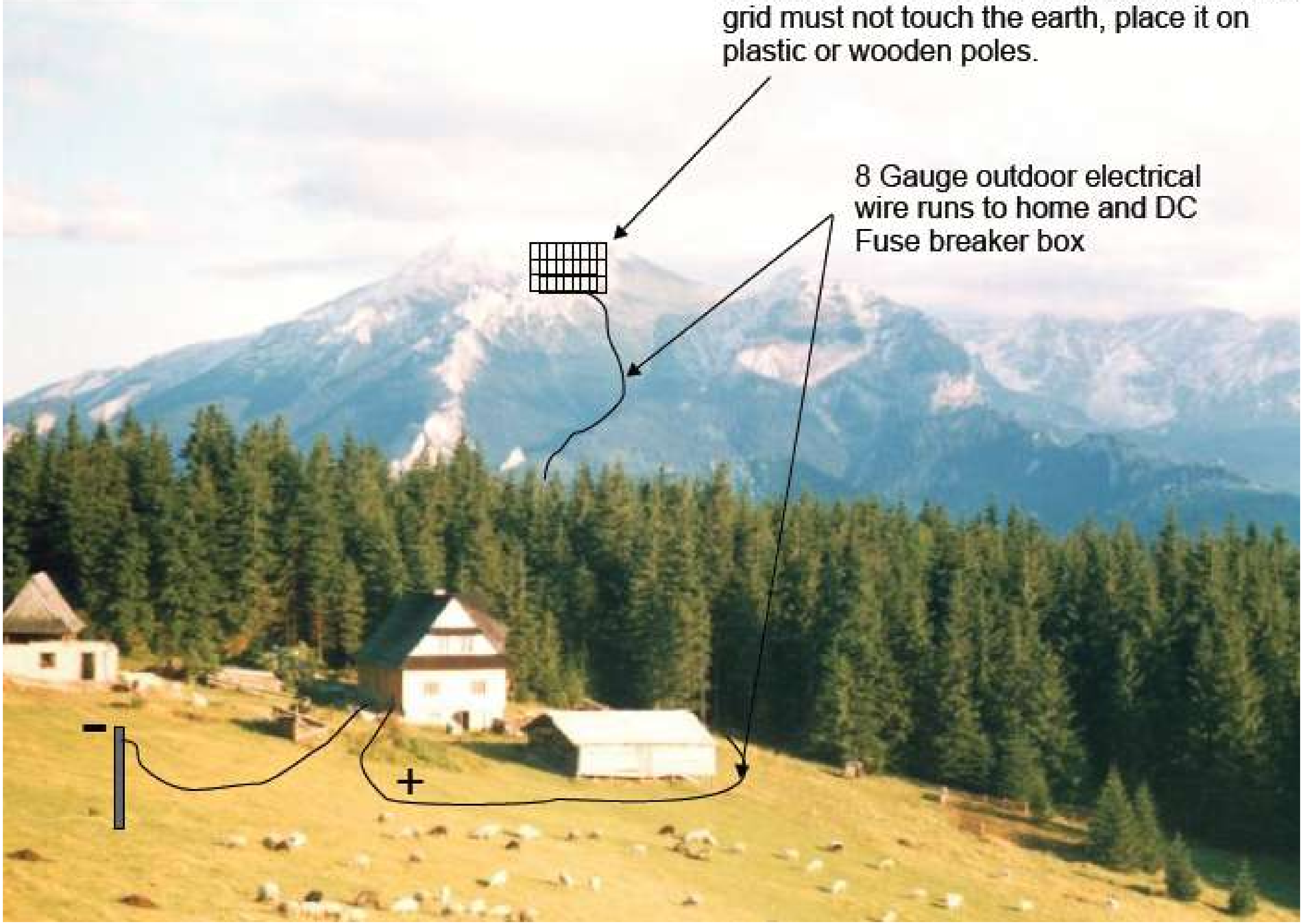

Large aluminum collection grid (Mesh type)
grid must not touch the earth, place it on
plastic or wooden poles.

8 Gauge outdoor electrical
wire runs to home and DC
Fuse breaker box

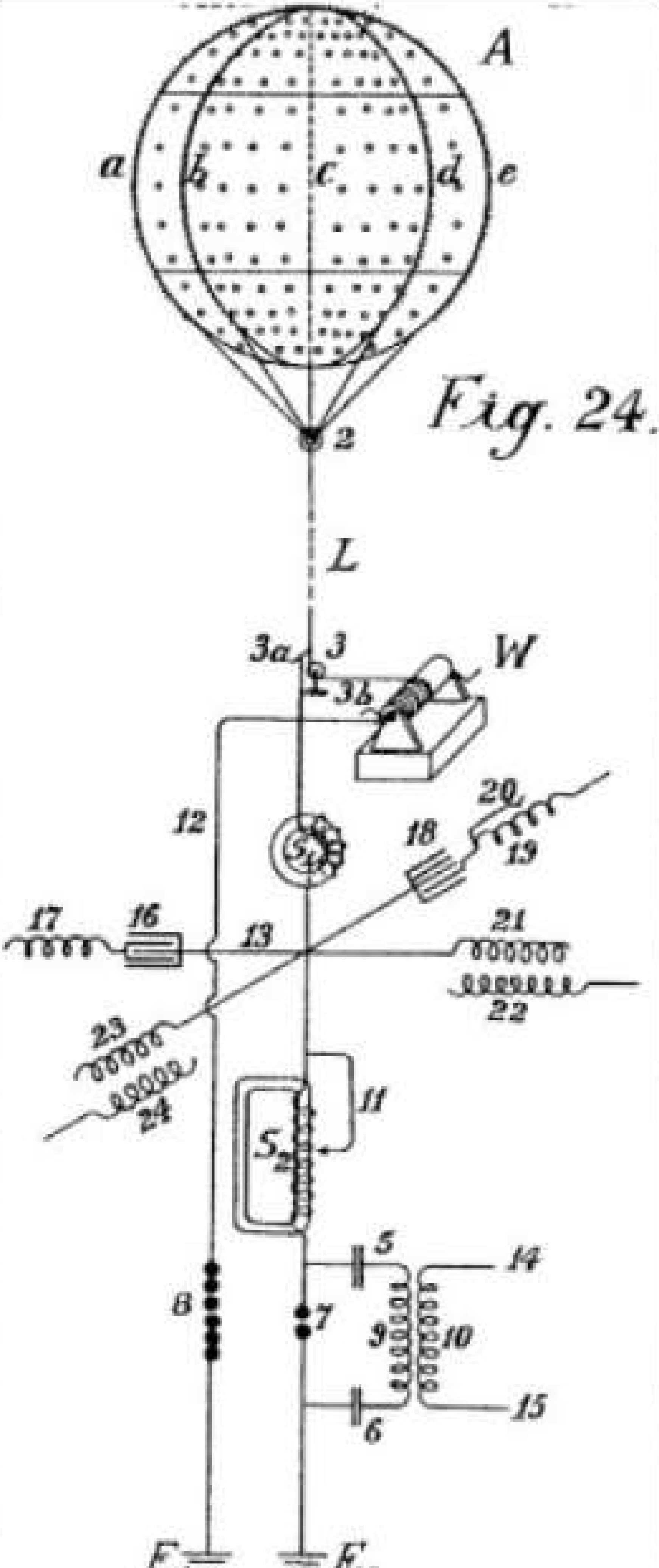
Fig. 24.
A
a
b
c
d
e
L
3a
3b
W
S4
S2
E1
E2

HARNESSING NATURE'S ELECTRICITY

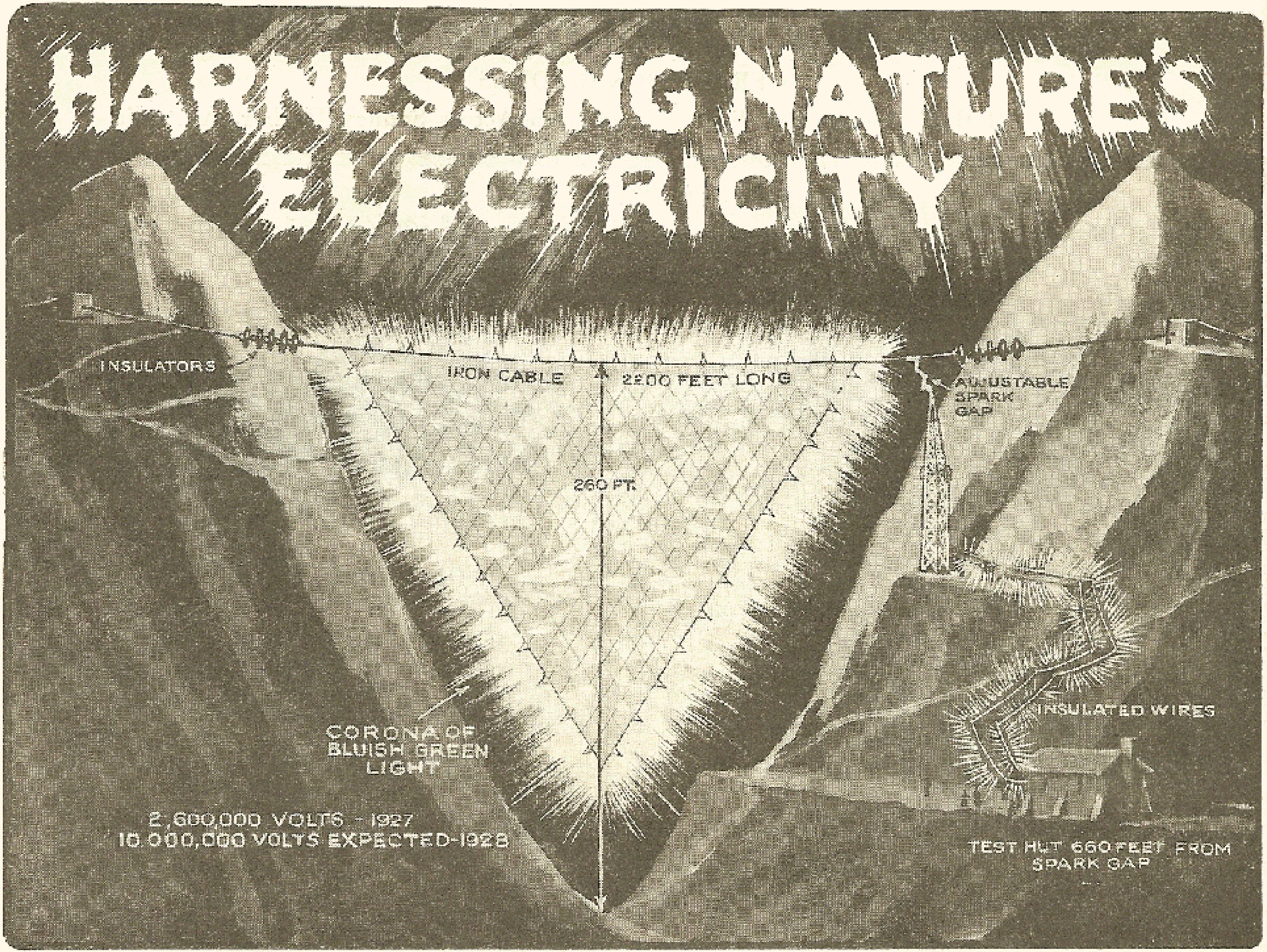
INSULATORS
IRON CABLE
2200 FEET LONG
ADJUSTABLE SPARK GAP
260 FT.
CORONA OF BLUISH GREEN LIGHT
INSULATED WIRES
2,600,000 VOLTS - 1927
10,000,000 VOLTS EXPECTED-1928
TEST HUT 660 FEET FROM SPARK GAP

Remarkable European Experiments with Atmospheric Electrical Discharges with Potentials as High as 3,000,000 Volts

By HENRY TOWNSEND

elevation of 350 ft., and these students of natural electrical phenomena have found a very desirable location in the Alps, where they can suspend between one mountain and an adjacent one, a strong iron cable having a length of about 2,000 ft. This cable is about 250 feet above the intervening valley, and from it these daring engineers have suspended a coarsely woven wire net, which serves as an electrical capacity to gather the electricity from the atmosphere. As shown in the pictures, the wire net is supplied with numerous sharp points to aid in collecting the current from the air.

As the accompanying photographs of the actual apparatus and wire cable used last year clearly show, an adjustable spark gap of considerable length is provided. By adjusting this spark gap to various lengths, it is possible to judge the voltage of the discharge which leaps the gap at any moment. Mr. F. W. Peek, Jr., the well-known American worker in the realm of high voltage measurements, together with other engineers, have provided tabulated data and curves for various lengths of both needle and sphere type spark gaps. As one of the accompanying diagrams shows, it is a simple matter to calculate the voltage when a certain length of gap is used. The engineer first checks the length of the gap on the chart; he then follows a line horizontally from the gap length, to where it intersects with the angular line on the chart; and from the point of intersection he looks in a visual line downward to a place where the voltage is given. For needle spark gap measurements, the characteristic curve on the chart is practically a straight line, while for sphere gaps the characteristic curve on the voltage versus gap length, is a curved line. Those interested in high voltage measurements by means of the spark gap method can find the voltage-gap tables and charts in the *Standardization Rules* of the American Institute of Electrical Engineers. According to Mr. Peek's researches, the voltage per foot of atmospheric electrical discharges is about 100,000, while in laboratory measurements with A.C., transformer high potential discharges, the average voltage per foot of spark was found to be about 150,000 volts. The voltage of a lightning flash may

(*Continued on page* 156)

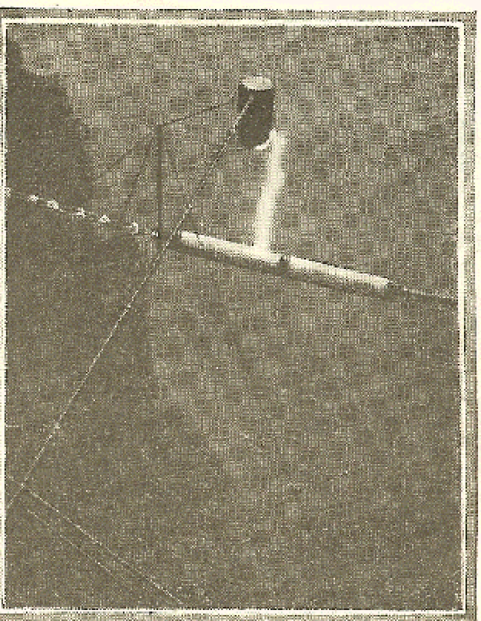

↑ **Actual photo above shows 13 ft., heavy spark obtained from the collecting net in the Alps by the German scientists. The voltage is about 2,000,000. The spark occurred once per second for 30 minutes.**

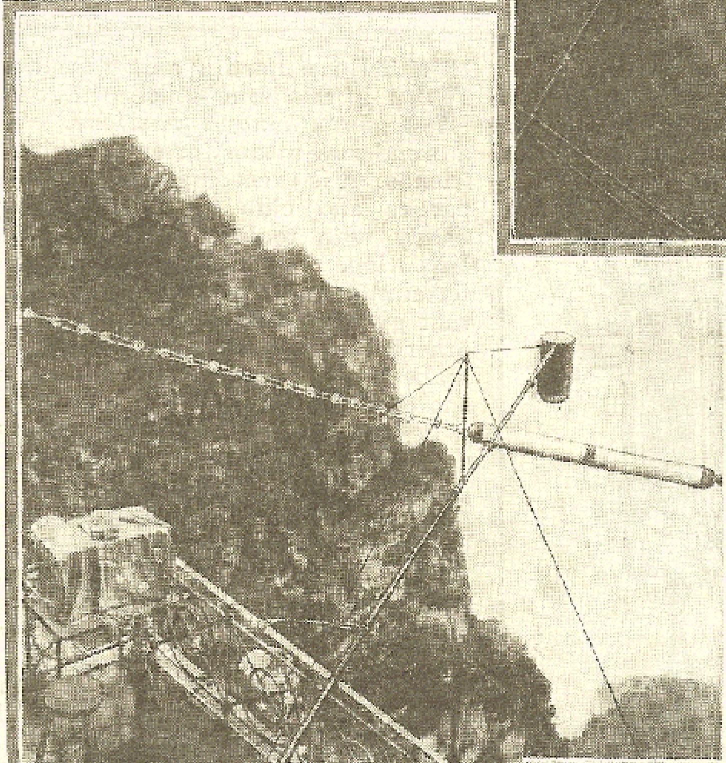

Photo, left, shows the adjustable spark gap used in the Alps. Notice the heavy electrode on the end of the adjustable arm to which the spark jumps.

←

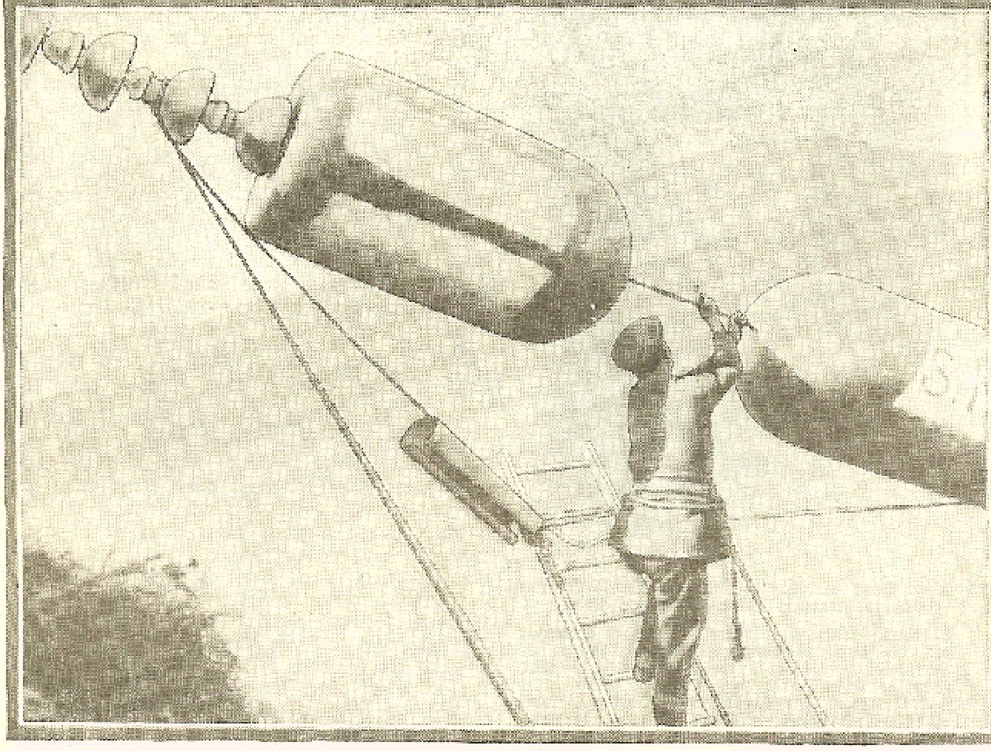

Actual photograph of the experimental "kite" used by the German experimenters in the Alps Mountains, for the purpose of accumulating high potential electrical discharges from the atmosphere. Note the size of the insulators.

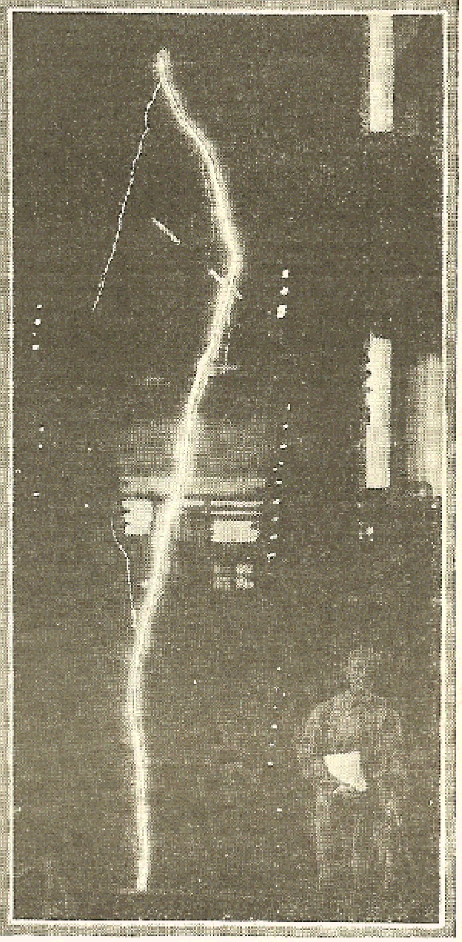

↓ **Below we see 3,000,000-volt artificial lightning stroke produced in G. E. Laboratory at Pittsfield, Mass. Note man.**

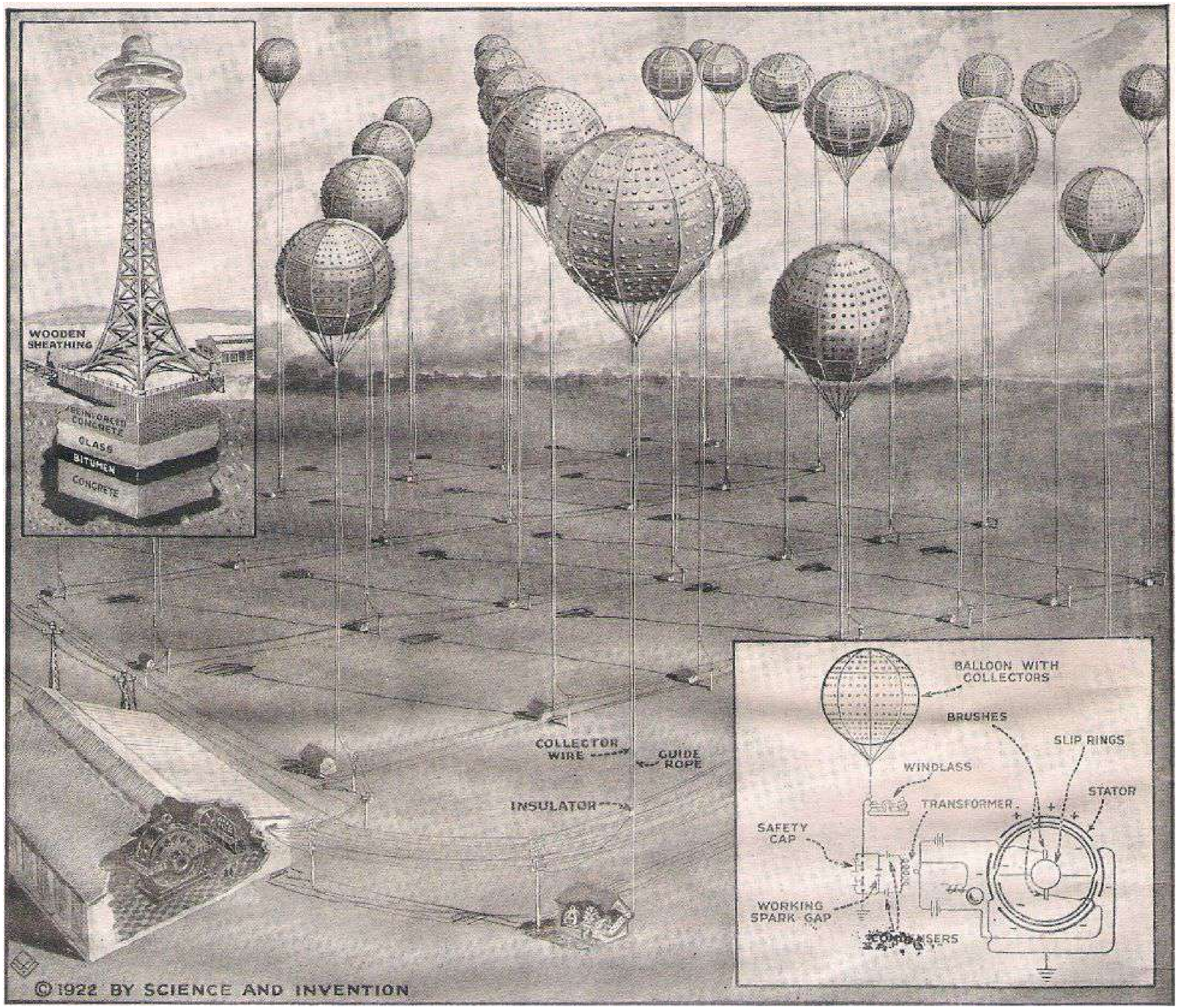
WOODEN SHEATHING
REINFORCED CONCRETE
GLASS
BITUMEN
CONCRETE
COLLECTOR WIRE
GUIDE ROPE
INSULATOR
BALLOON WITH COLLECTORS
BRUSHES
SLIP RINGS
WINDLASS
TRANSFORMER
STATOR
SAFETY CAP
WORKING SPARK GAP
CONDENSERS

© 1922 BY SCIENCE AND INVENTION

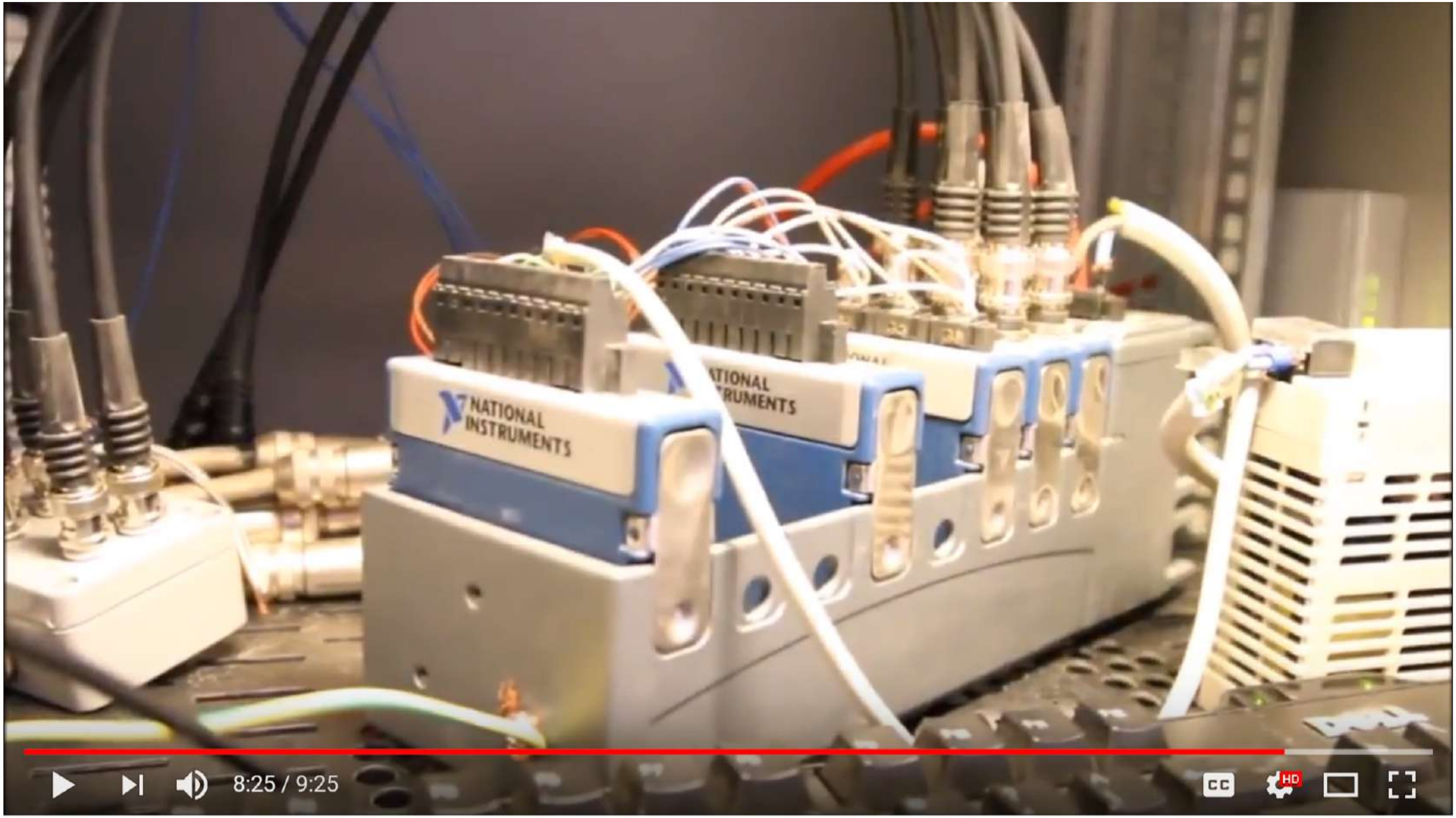
NATIONAL INSTRUMENTS
NATIONAL INSTRUMENTS
DELL
8:25 / 9:25
HD

Flat Earth CON SPIRES! Cathedral Spires Are Secretly Atmospheric Electricity Ma

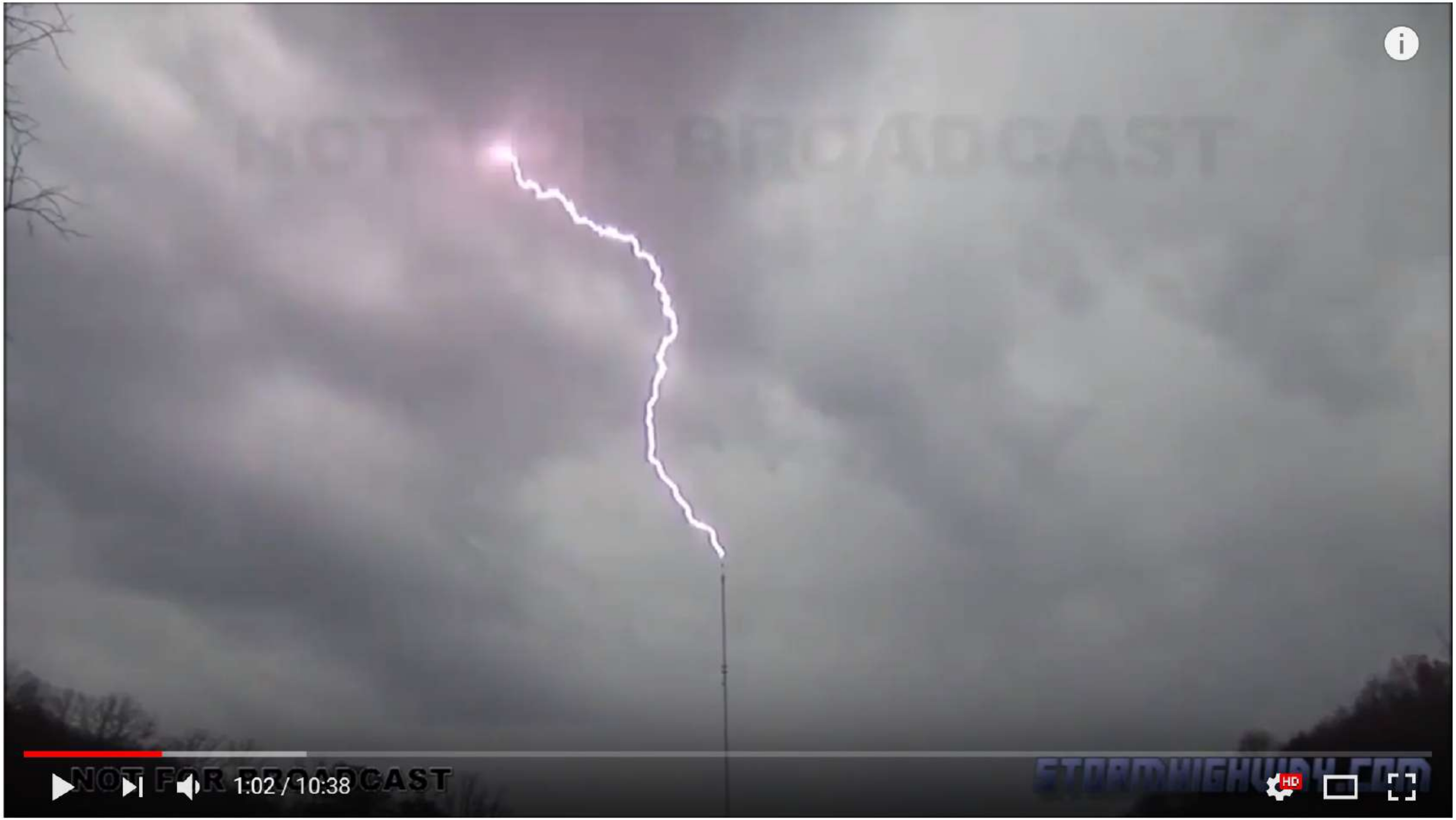
NOT FOR BROADCAST
NOT FOR BROADCAST
1:02 / 10:38
STORMHIGHWAY.COM
HD

1:45 / 9:25
CC
HD

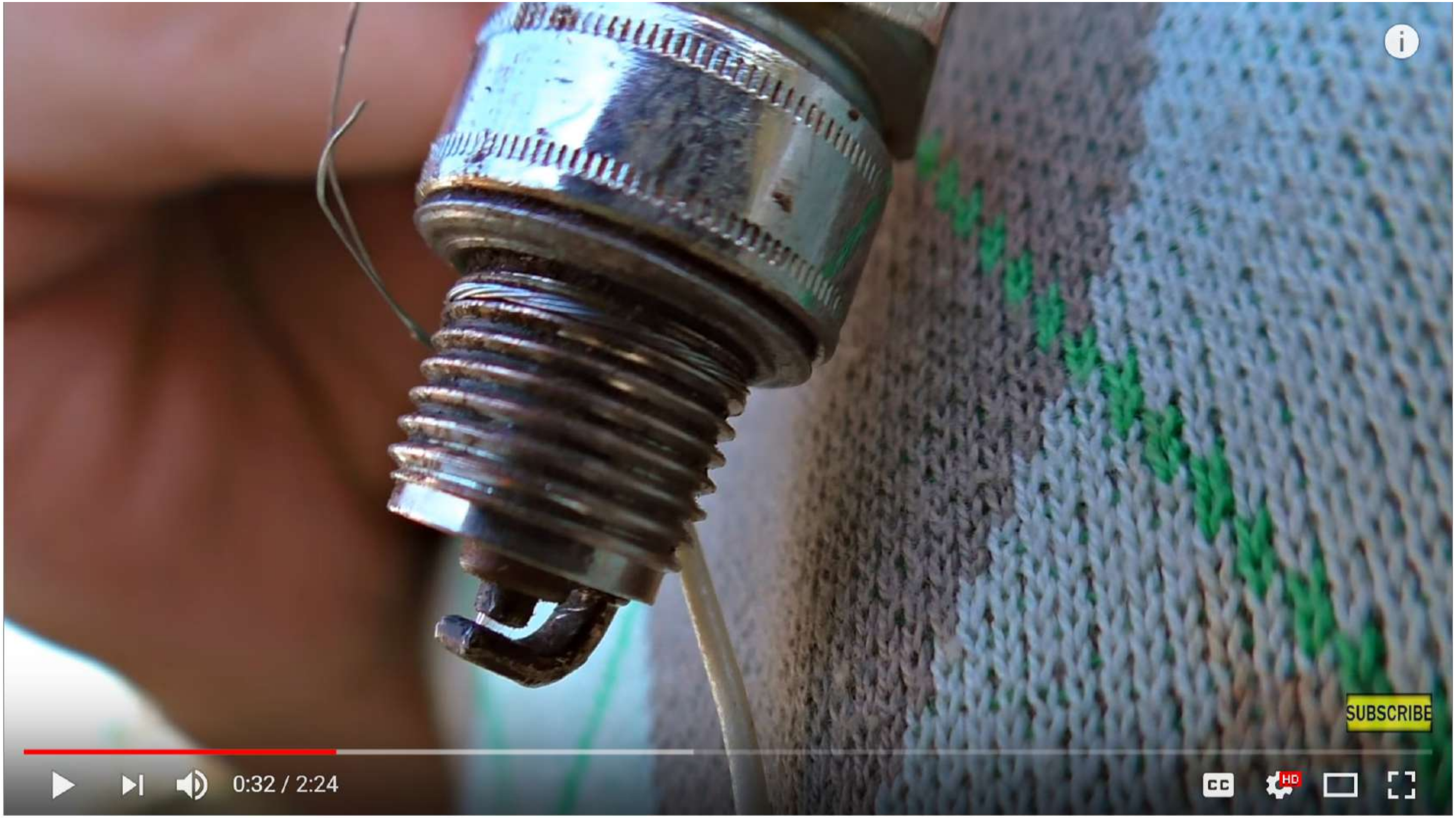
SUBSCRIBE
0:32 / 2:24
HD

Charging from the power lines.

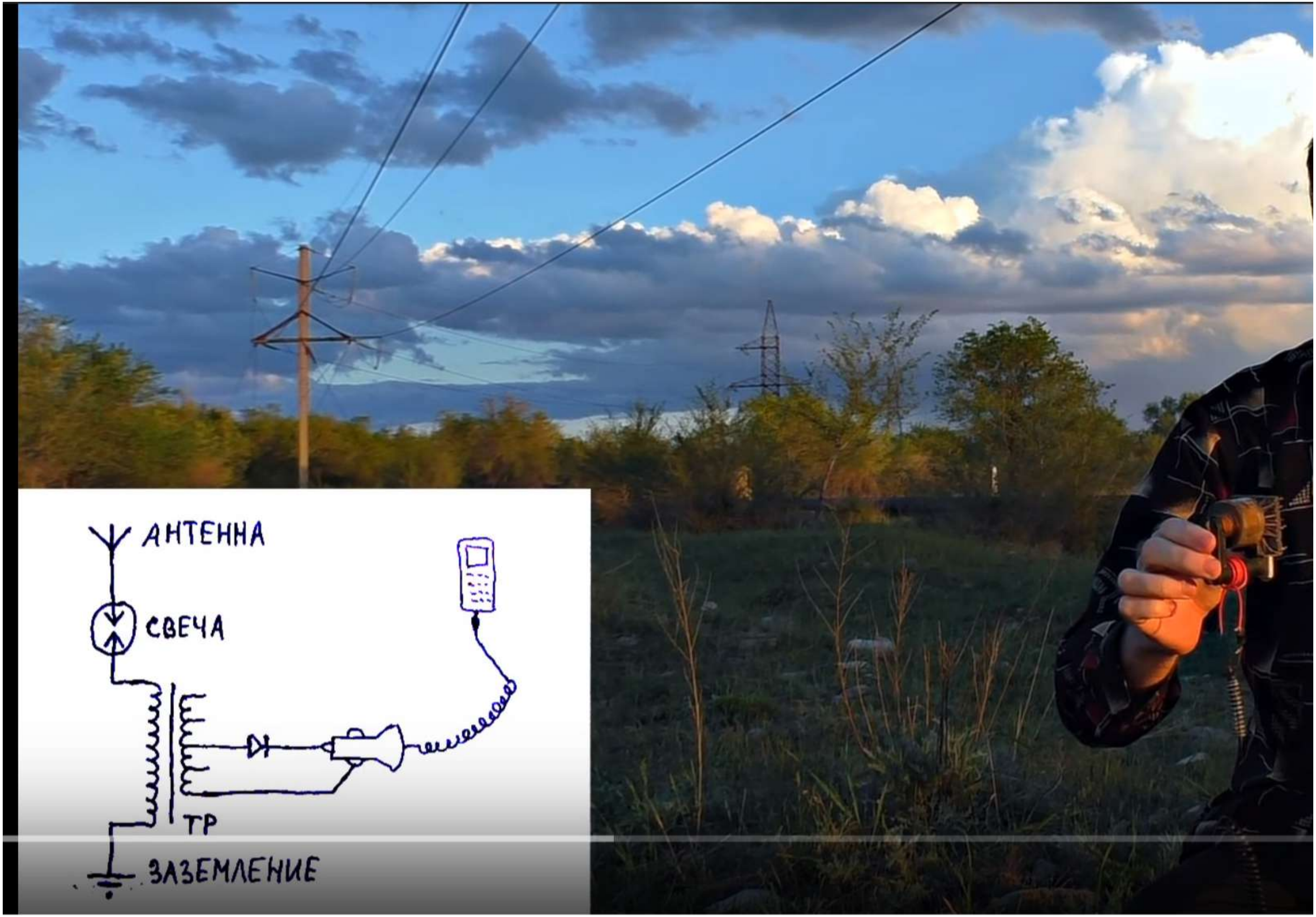
АНТЕННА
СВЕЧА
ТР
ЗАЗЕМЛЕНИЕ

E6

MULTI DISPLAY DMM
10.1
POWER
DT-920

SPARK PLUG DIODE CAR CHARGER FLYBACK TRANSFORMER TV ANTENNA
00.0
POWER
COM
DT-82

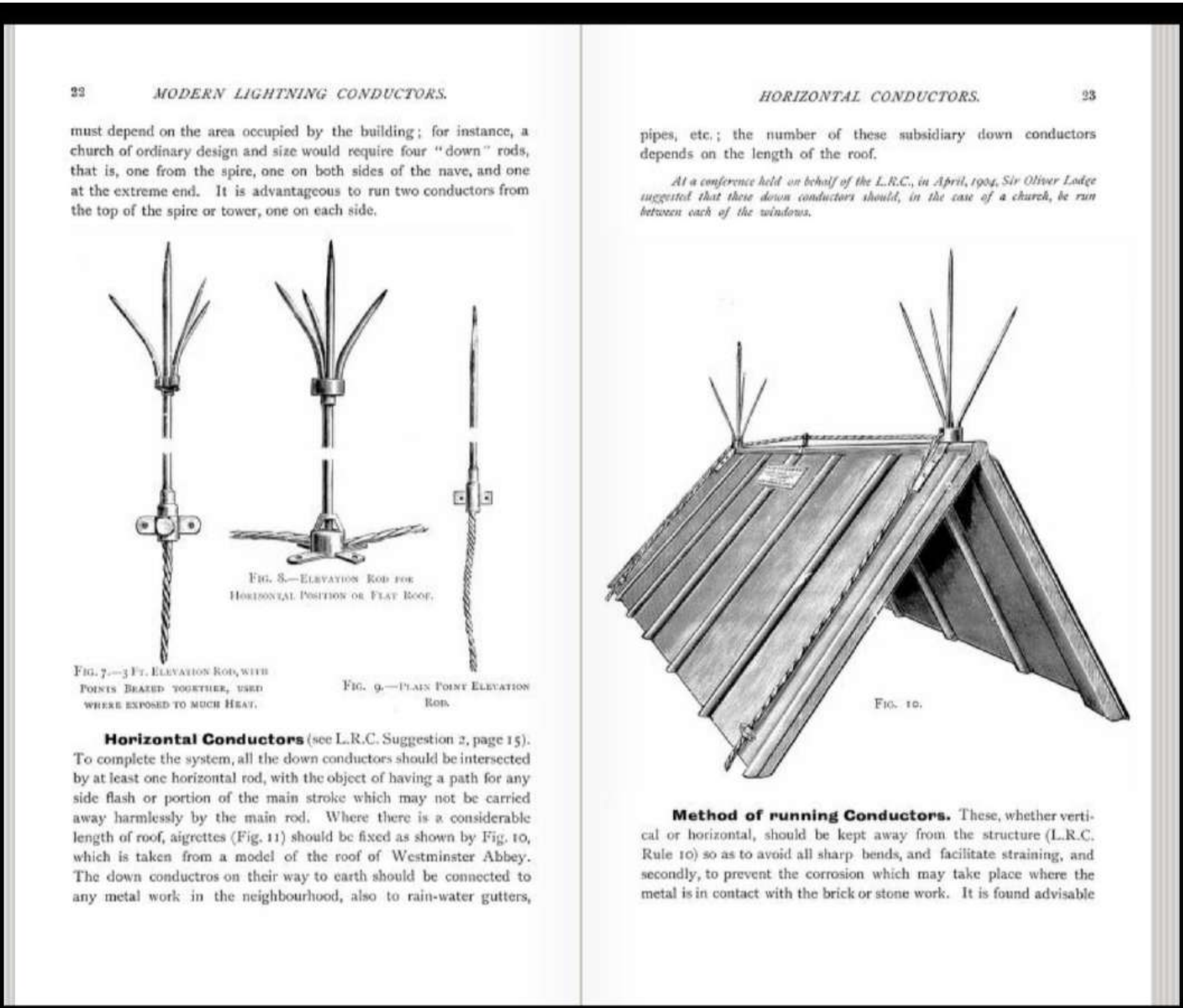

22 *MODERN LIGHTNING CONDUCTORS.*

must depend on the area occupied by the building; for instance, a church of ordinary design and size would require four "down" rods, that is, one from the spire, one on both sides of the nave, and one at the extreme end. It is advantageous to run two conductors from the top of the spire or tower, one on each side.

FIG. 7.—3 FT. ELEVATION ROD, WITH POINTS BRAZED TOGETHER, USED WHERE EXPOSED TO MUCH HEAT.

FIG. 8.—ELEVATION ROD FOR HORIZONTAL POSITION OR FLAT ROOF.

FIG. 9.—PLAIN POINT ELEVATION ROD.

Horizontal Conductors (see L.R.C. Suggestion 2, page 15). To complete the system, all the down conductors should be intersected by at least one horizontal rod, with the object of having a path for any side flash or portion of the main stroke which may not be carried away harmlessly by the main rod. Where there is a considerable length of roof, aigrettes (Fig. 11) should be fixed as shown by Fig. 10, which is taken from a model of the roof of Westminster Abbey. The down conductros on their way to earth should be connected to any metal work in the neighbourhood, also to rain-water gutters, pipes, etc.; the number of these subsidiary down conductors depends on the length of the roof.

HORIZONTAL CONDUCTORS. 23

At a conference held on behalf of the L.R.C., in April, 1904, Sir Oliver Lodge suggested that these down conductors should, in the case of a church, be run between each of the windows.

FIG. 10.

Method of running Conductors. These, whether vertical or horizontal, should be kept away from the structure (L.R.C. Rule 10) so as to avoid all sharp bends, and facilitate straining, and secondly, to prevent the corrosion which may take place where the metal is in contact with the brick or stone work. It is found advisable

Modern lightning conductors: an illustrated

surface of the joint need not necessarily exceed that of the cross section of the conductors. The joint should be put together previously by screws or rivets, and the soldered joint, especially if used in underground work, should be carefully protected from local electrical action by tarred rope. Stranded iron conductors can be connected (as previously described) by use of a box joint; the box, Fig. 28, must be of the same metal as the conductors.

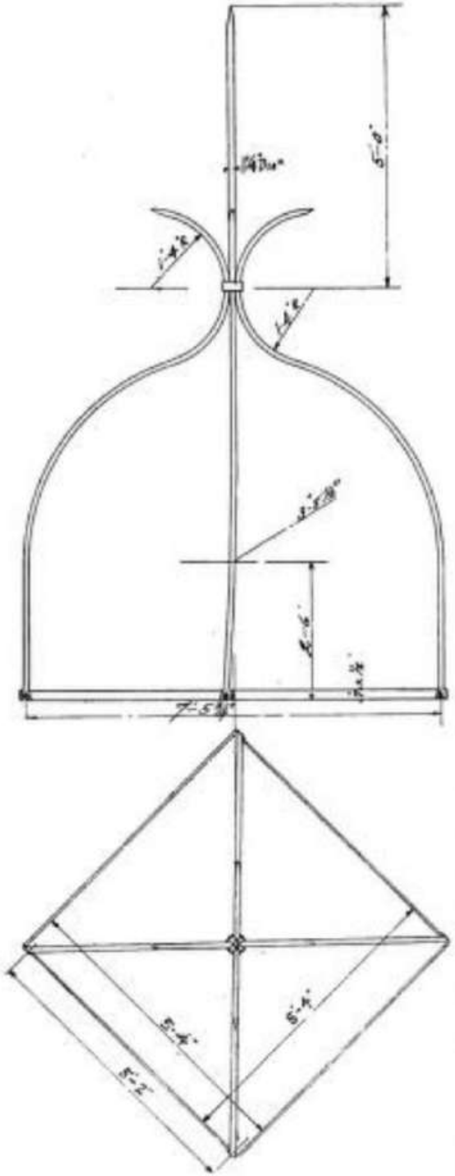

FIG. 27.—TERMINALS IN FORM OF AN ARCH FOR CHIMNEY STACK.

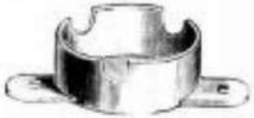

FIG. 28.

Vanes.—Particular attention must be paid to the necessity of making a permanent joint to the spindle. A clamp is prepared of the same material as the spindle, and is furnished with two bolts to tighten; if iron is used it is well to line the clamp with a piece of sheet lead. The conductor is sweated into a socket which is fitted with an eye, through which one of the tightening bolts passes. In the

case of the vanes of churches and those fixed in inaccessible positions, two separate clamps should be used.

Internal Masses of Metal.—Roof trusses fitted with longitudinal iron tie rods will, as a rule, be found to be electrically connected, but should this not be the case each truss must be joined to the conductors. All large and long masses of metal, such as beams, girders, roof trusses, tie rods, hot water systems, traveller ways, hoisting crabs, engines, boilers, large machines, and ventilators fixed in the interiors of buildings, should be connected to all conductors that pass near them, and as far as possible with one another. The discontinuous parts of traveller rails should be connected by straps, or in some cases tramway bonds might be used. If electric light wires are run in tubes, such as the "SIMPLEX," this should be earthed. Metallic contact between lead or zinc sheeting and flashings should be carefully studied, and for special work strips of sufficient size should be either burnt on to lead or soldered in such a way that the joint will stand rough usage, and allow for expansion or contraction.

Earth Connection.—"*It is essential that the lower extremity of the conductor be buried in permanently damp soil; hence proximity to rain-water pipes, and to drains, is desirable. It is a very good plan to make the conductor bifurcate close below the surface of the ground, and adopt two of the following methods for securing the escape of the lightning into the earth. A strip of copper tape may be led from the bottom of the rod to the nearest gas or water* main—*not merely to a lead pipe—and be soldered to it; or a tape may be soldered to a sheet of copper 3 feet by 3 feet and $\frac{1}{16}$ inch thick, buried in permanently wet earth, and surrounded by cinders or coke; or many yards of the tape may be laid in a trench filled with coke, taking care that the surfaces of copper are, as in the previous cases, not less than 18 square feet. Where iron is used for the rod, a galvanised iron plate of similar dimensions should be employed.*

"The use of cinders or coke appears to be questionable owing to the chemical or electrolytic effect on copper or iron. Charcoal or pulverised carbon (such as ends of arc-light rods) is better. **A tubular earth** consisting of a perforated steel spike driven tightly into moist ground and lengthened up to the surface, the conductor reaching to the bottom and being packed with granulated charcoal, gives as much **effective area** as a plate of larger surface, and can easily be kept moist by connecting it to the nearest rain-water pipe. The resistance of a tubular earth on this plan should be very low and practically constant."—*Lightning Research Committee, 1905.*

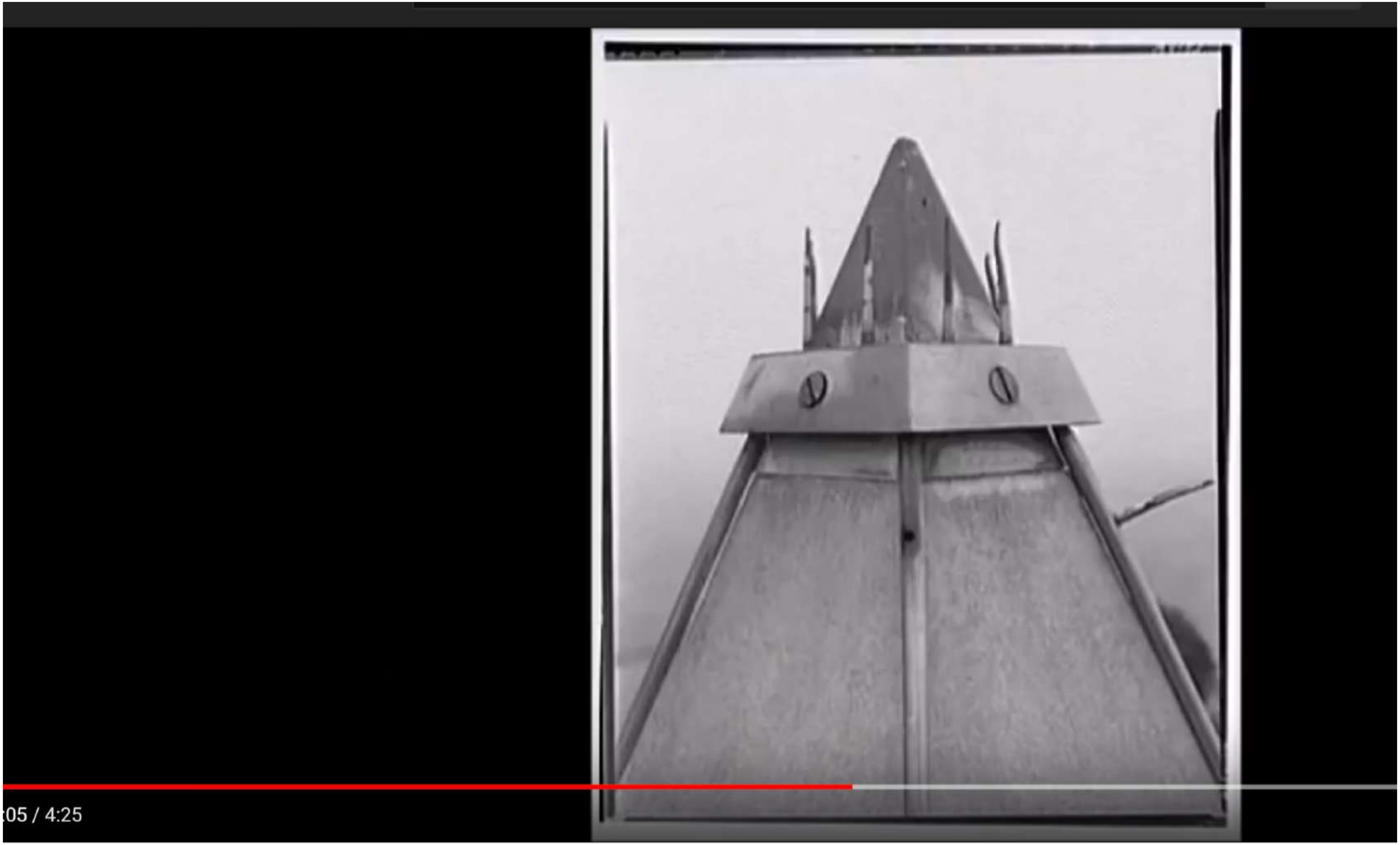

Washington Monument is an Atmospheric Electricity Mast

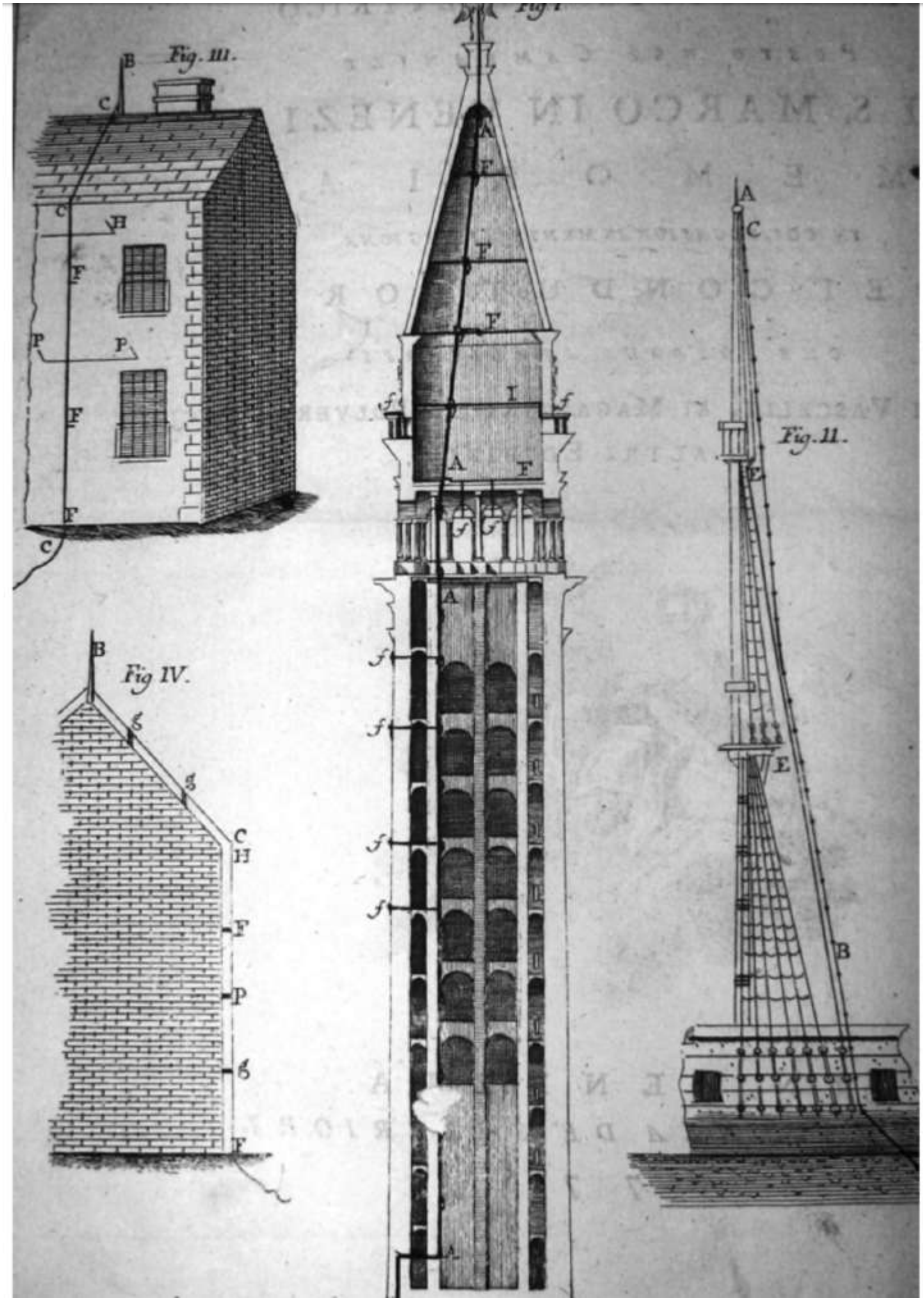
Fig. III.
B
C
C
H
F
P
P
F
F
C
Fig IV.
B
g
g
C
H
F
P
g
F
Fig. I
A
F
F
F
I
f
f
A
F
f
f
A
f
f
f
f
A
A
C
Fig. II.
E
B

Figure 2.3. The lightning rod that Toaldo designed for the church of San Marco in Venice. GiuseppeToaldo, “Del conduttore elettrico posto nel campanile” (Padua, 1776). Franklin Collec-

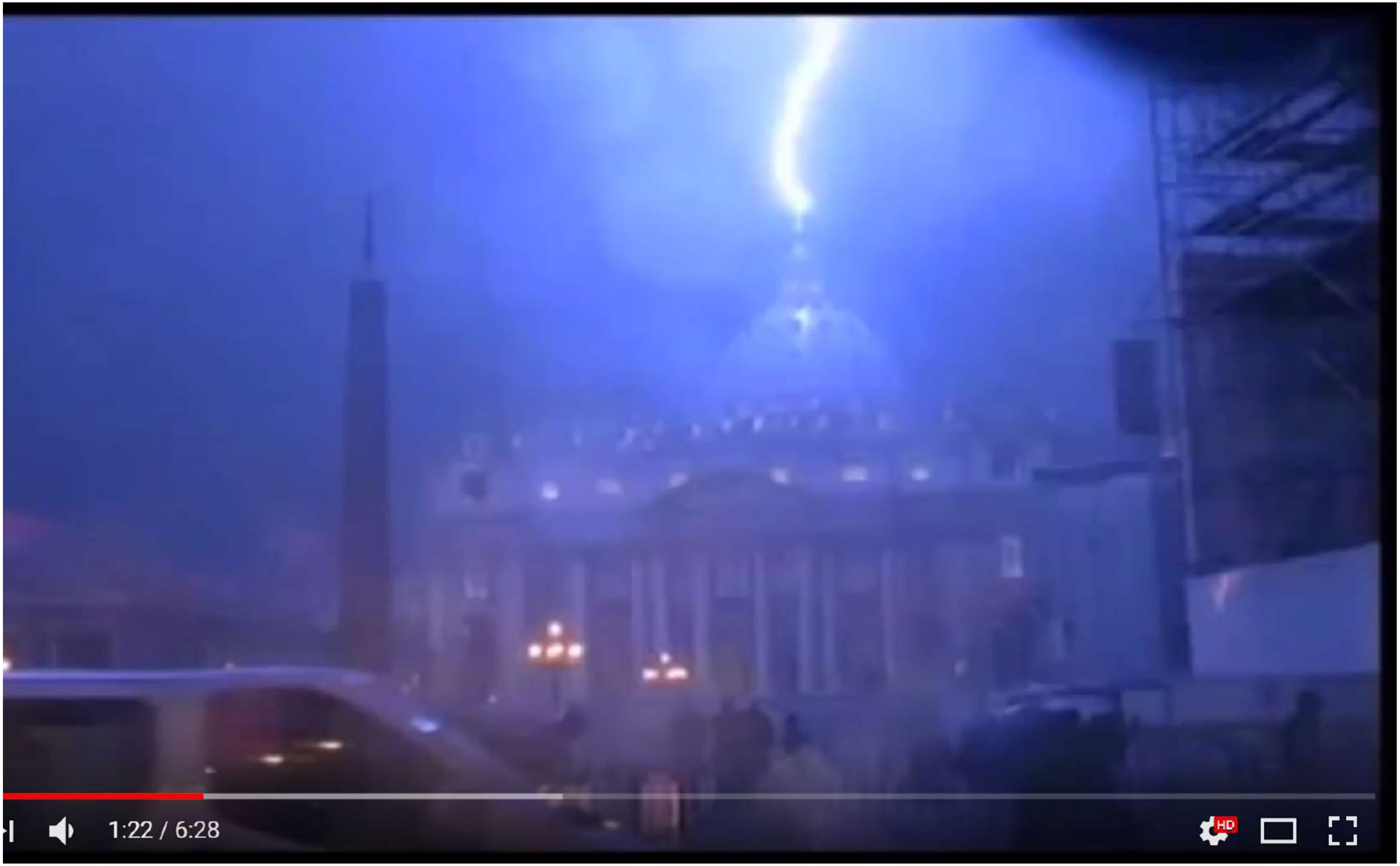

Gilded Temples Are Secretly Atmospheric Electricity Masts

How To Find S-W Stations Quickly

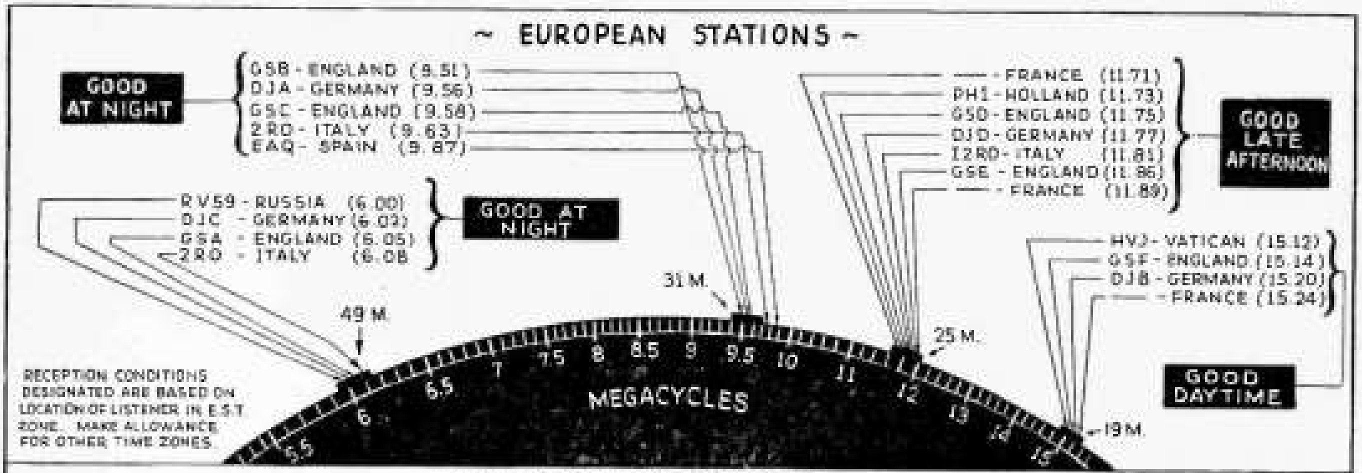

~ EUROPEAN STATIONS ~

GOOD AT NIGHT
GSB - ENGLAND (9.51)
DJA - GERMANY (9.56)
GSC - ENGLAND (9.58)
2RO - ITALY (9.63)
EAQ - SPAIN (9.87)
—— - FRANCE (11.71)
PH1 - HOLLAND (11.73)
GSD - ENGLAND (11.75)
DJD - GERMANY (11.77)
I2RO - ITALY (11.81)
GSE - ENGLAND (11.86)
—— - FRANCE (11.89)
GOOD LATE AFTERNOON
RV59 - RUSSIA (6.00)
DJC - GERMANY (6.02)
GSA - ENGLAND (6.05)
2RO - ITALY (6.08)
GOOD AT NIGHT
HVJ - VATICAN (15.12)
GSF - ENGLAND (15.14)
DJB - GERMANY (15.20)
—— - FRANCE (15.24)
GOOD DAYTIME
31 M.
49 M.
25 M.
19 M.
RECEPTION CONDITIONS DESIGNATED ARE BASED ON LOCATION OF LISTENER IN E.S.T. ZONE. MAKE ALLOWANCE FOR OTHER TIME ZONES.
MEGACYCLES
5.5
6
6.5
7
7.5
8
8.5
9
9.5
10
11
12
13
14
15

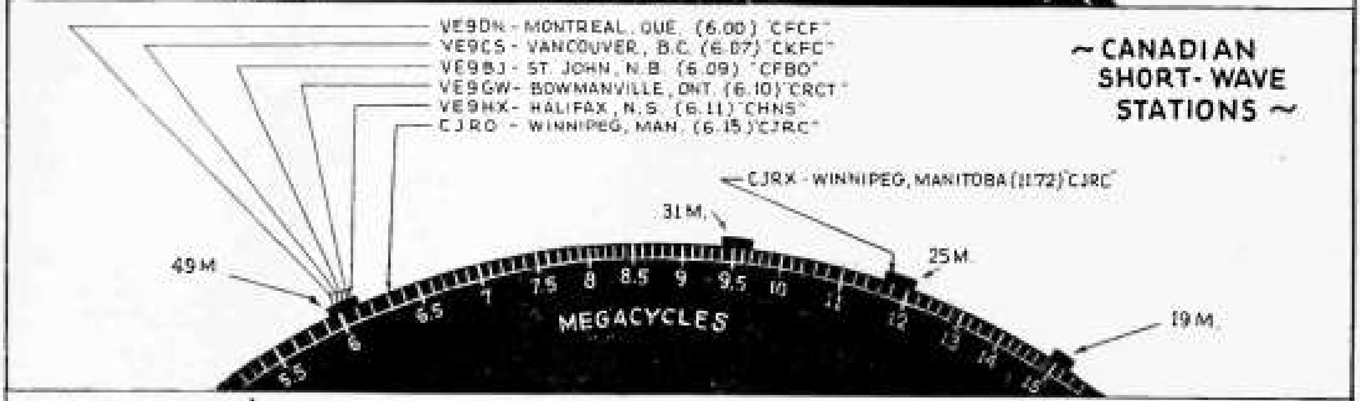

~ CANADIAN SHORT-WAVE STATIONS ~

VE9DN - MONTREAL, QUE. (6.00) "CFCF"
VE9CS - VANCOUVER, B.C. (6.07) "CKFC"
VE9BJ - ST. JOHN, N.B. (6.09) "CFBO"
VE9GW - BOWMANVILLE, ONT. (6.10) "CRCT"
VE9HX - HALIFAX, N.S. (6.11) "CHNS"
CJRO - WINNIPEG, MAN. (6.15) "CJRC"
CJRX - WINNIPEG, MANITOBA (11.72) "CJRC"
31 M.
49 M.
25 M.
19 M.
MEGACYCLES
5.5
6
6.5
7
7.5
8
8.5
9
9.5
10
11
12
13
14
15

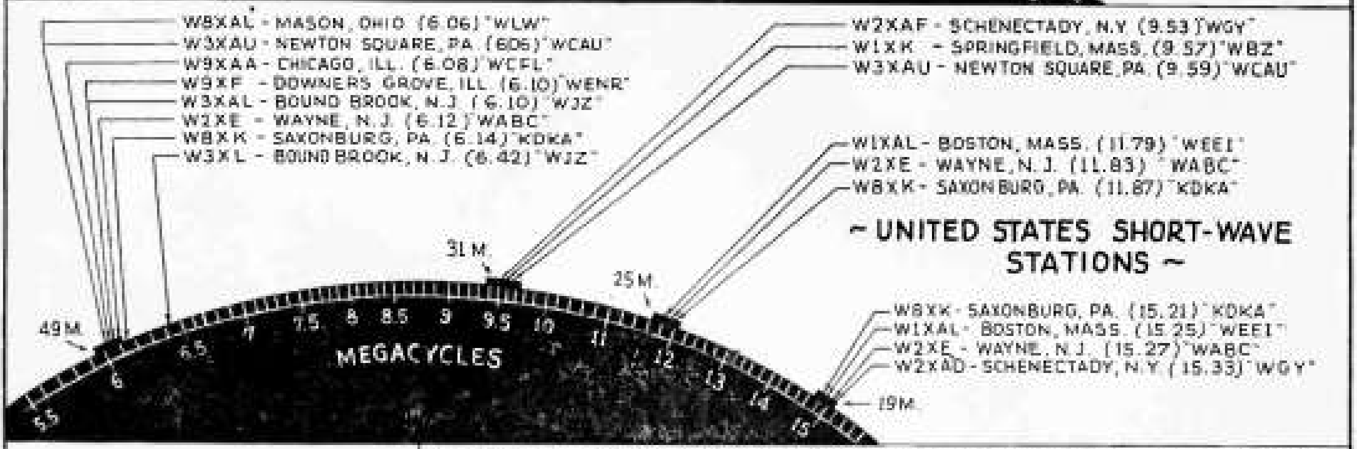

~ UNITED STATES SHORT-WAVE STATIONS ~

W8XAL - MASON, OHIO (6.06) "WLW"
W3XAU - NEWTON SQUARE, PA. (6.06) "WCAU"
W9XAA - CHICAGO, ILL. (6.08) "WCFL"
W9XF - DOWNERS GROVE, ILL. (6.10) "WENR"
W3XAL - BOUND BROOK, N.J. (6.10) "WJZ"
W2XE - WAYNE, N.J. (6.12) "WABC"
W8XK - SAXONBURG, PA. (6.14) "KDKA"
W3XL - BOUND BROOK, N.J. (6.42) "WJZ"
W2XAF - SCHENECTADY, N.Y. (9.53) "WGY"
W1XK - SPRINGFIELD, MASS. (9.57) "WBZ"
W3XAU - NEWTON SQUARE, PA. (9.59) "WCAU"
W1XAL - BOSTON, MASS. (11.79) "WEEI"
W2XE - WAYNE, N.J. (11.83) "WABC"
W8XK - SAXONBURG, PA. (11.87) "KDKA"
W8XK - SAXONBURG, PA. (15.21) "KDKA"
W1XAL - BOSTON, MASS. (15.25) "WEEI"
W2XE - WAYNE, N.J. (15.27) "WABC"
W2XAD - SCHENECTADY, N.Y. (15.33) "WGY"
31 M.
25 M.
49 M.
19 M.
MEGACYCLES
5.5
6
6.5
7
7.5
8
8.5
9
9.5
10
11
12
13
14
15

When To Tune and Where

● Until you have had considerable experience in short wave tuning, it is well to confine your efforts to the four bands indicated on the dial. These may look small, compared to the total length of the dial, but they contain the best "plums" among the foreign stations and many faint ones which, at first, you will probably pass unnoticed until you learn to tune slowly and carefully, intent upon identifying each slight "swish", which may be a station.

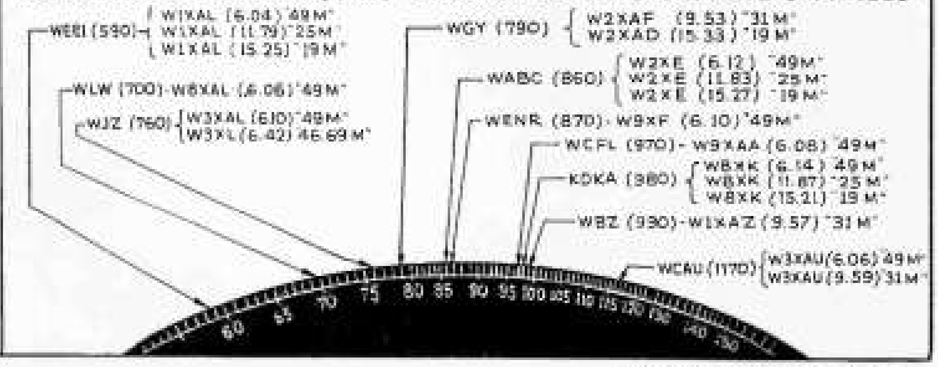

U.S. STATIONS BROADCASTING PROGRAMS ON-SHORT-WAVE CHANNELS

WEEI (590) { W1XAL (6.04) "49 M" W1XAL (11.79) "25 M" W1XAL (15.25) "19 M"
WLW (700) - W8XAL (6.06) "49 M"
WJZ (760) { W3XAL (6.10) "49 M" W3XL (6.42) 46.69 M"
WGY (790) { W2XAF (9.53) "31 M" W2XAD (15.33) "19 M"
WABC (860) { W2XE (6.12) "49 M" W2XE (11.83) "25 M" W2XE (15.27) "19 M"
WENR (870) - W9XF (6.10) "49 M"
WCFL (970) - W9XAA (6.08) "49 M"
KDKA (980) { W8XK (6.14) "49 M" W8XK (11.87) "25 M" W8XK (15.21) "19 M"
WBZ (990) - W1XAZ (9.57) "31 M"
WCAU (1170) { W3XAU (6.06) "49 M" W3XAU (9.59) "31 M"
60
65
70
75
80
85
90
95
100
105
110
115
120
130
140
150

Copyright 1935 by H. Gernsback.

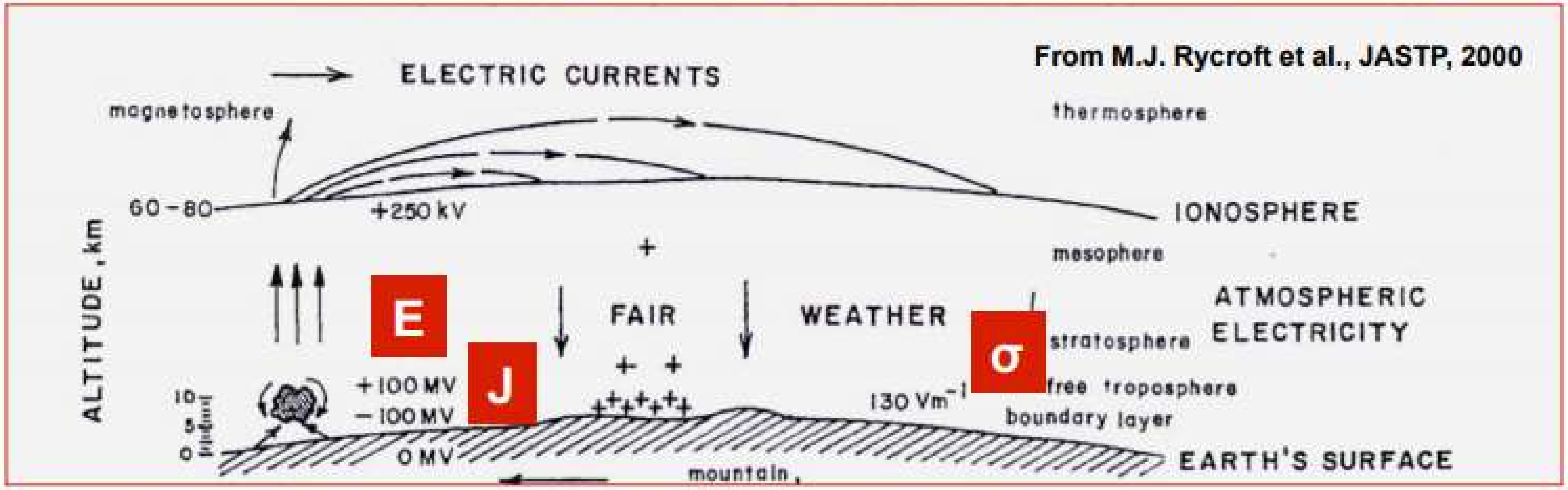
ELECTRIC CURRENTS
magnetosphere
thermosphere
60-80
+250 kV
IONOSPHERE
ALTITUDE, km
+
mesophere
E
FAIR
WEATHER
σ
ATMOSPHERIC ELECTRICITY
stratosphere
J
+ +
+100 MV
-100 MV
130 Vm^{-1}
free troposphere
boundary layer
10
5
0
0 MV
mountain
EARTH'S SURFACE

From M.J. Rycroft et al., JASTP, 2000

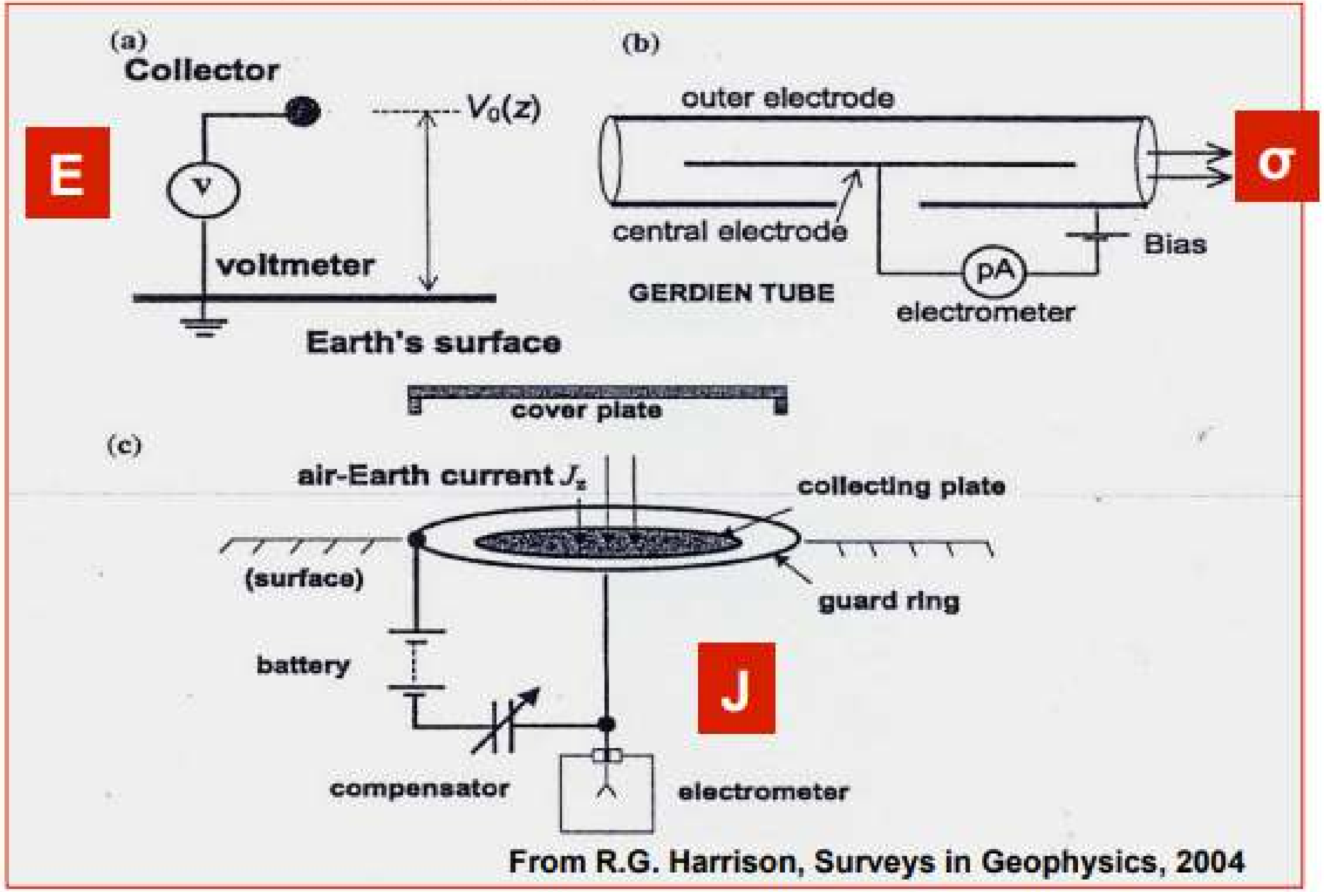
(a)
Collector
$V_0(z)$
E
V
voltmeter
Earth's surface
(b)
outer electrode
σ
central electrode
Bias
pA
GERDIEN TUBE
electrometer
cover plate
(c)
air-Earth current J_z
collecting plate
(surface)
guard ring
battery
J
compensator
electrometer

From R.G. Harrison, Surveys in Geophysics, 2004

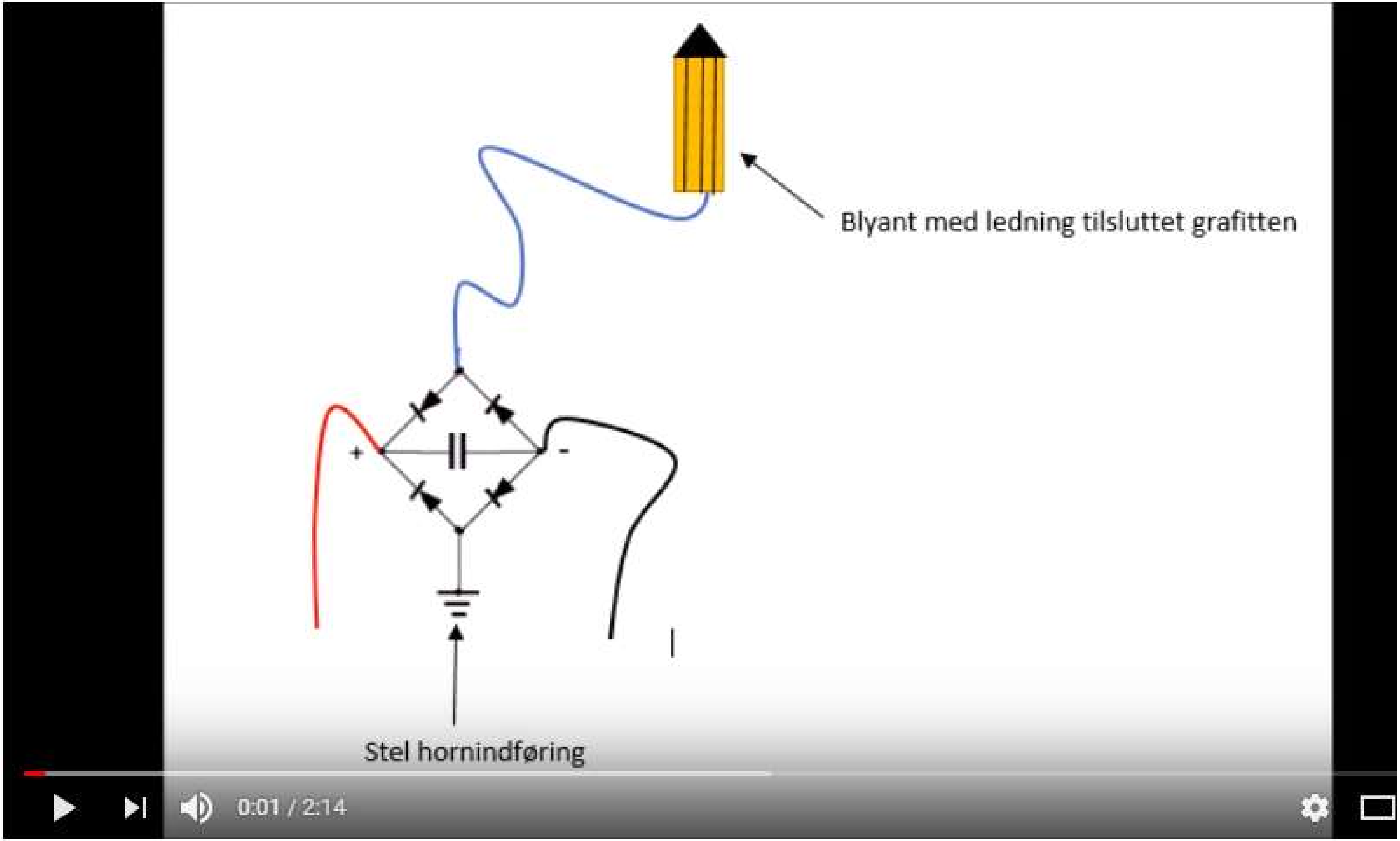
Blyant med ledning tilsluttet grafitten
+
-
Stel hornindføring
0:01 / 2:14

tesla radiant collector/ion collector with pencil

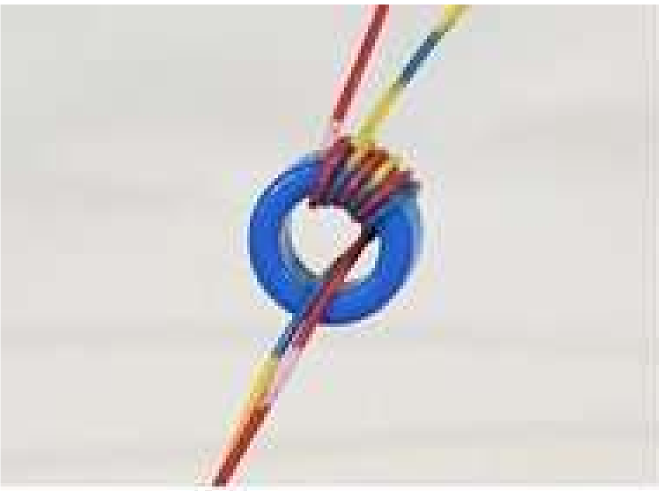

Keeping the two wires together, make a few more turns through the center.

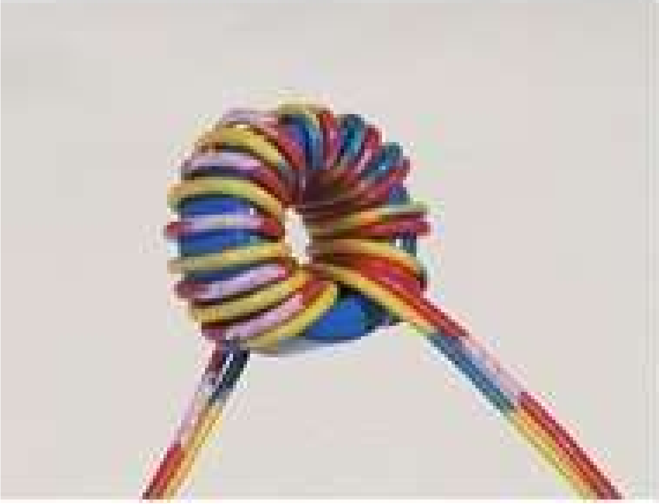

Keep winding until you fit as many turns as will fit in a single layer around the toroid, typically 7-10 turns with thin insulated wire.

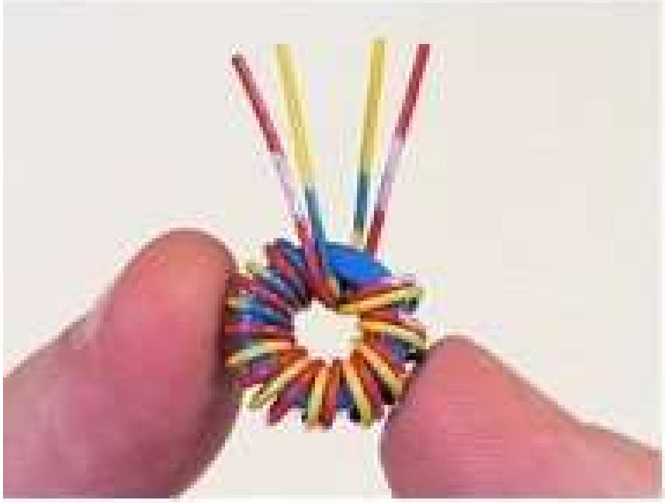

Clip the wire leads down. Note that we have two pairs of wires: one coming out the front, and one coming out the back.

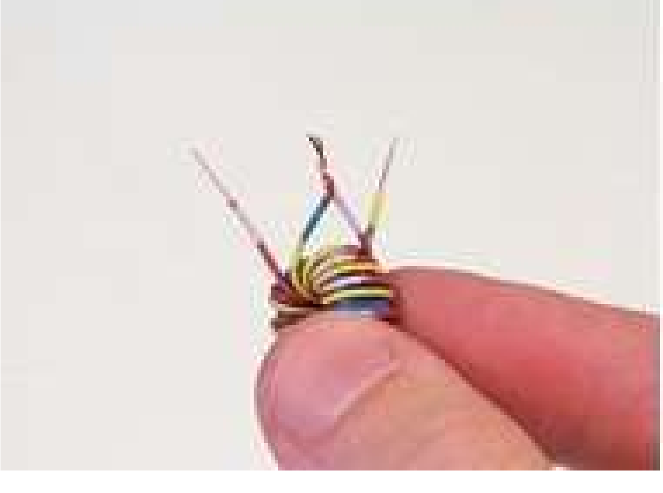

Strip the wire ends. Take one wire from each pair *of different color* and attach them together.

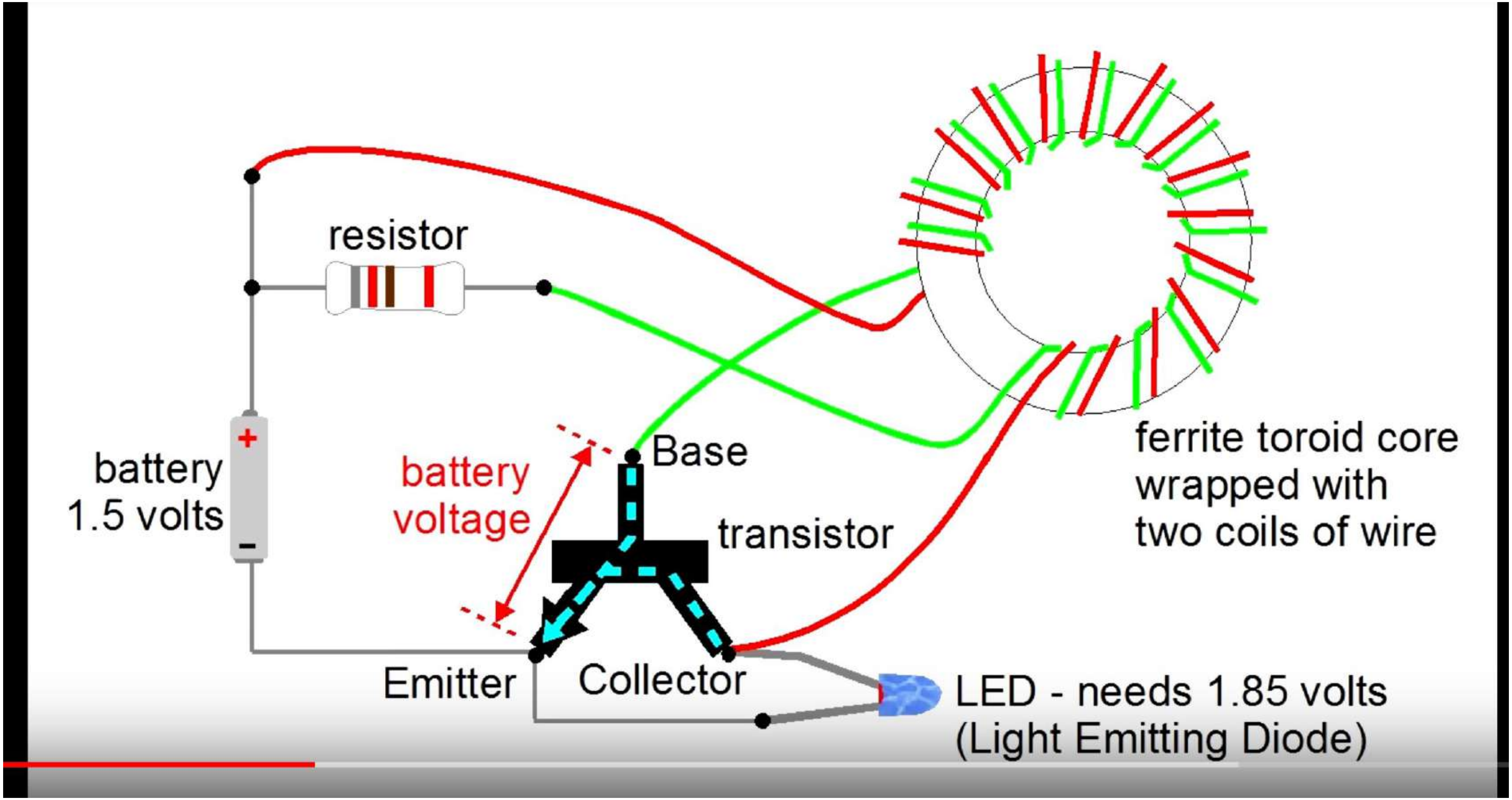
resistor
battery
1.5 volts
+
–
battery
voltage
Base
transistor
Emitter
Collector
ferrite toroid core
wrapped with
two coils of wire
LED - needs 1.85 volts
(Light Emitting Diode)

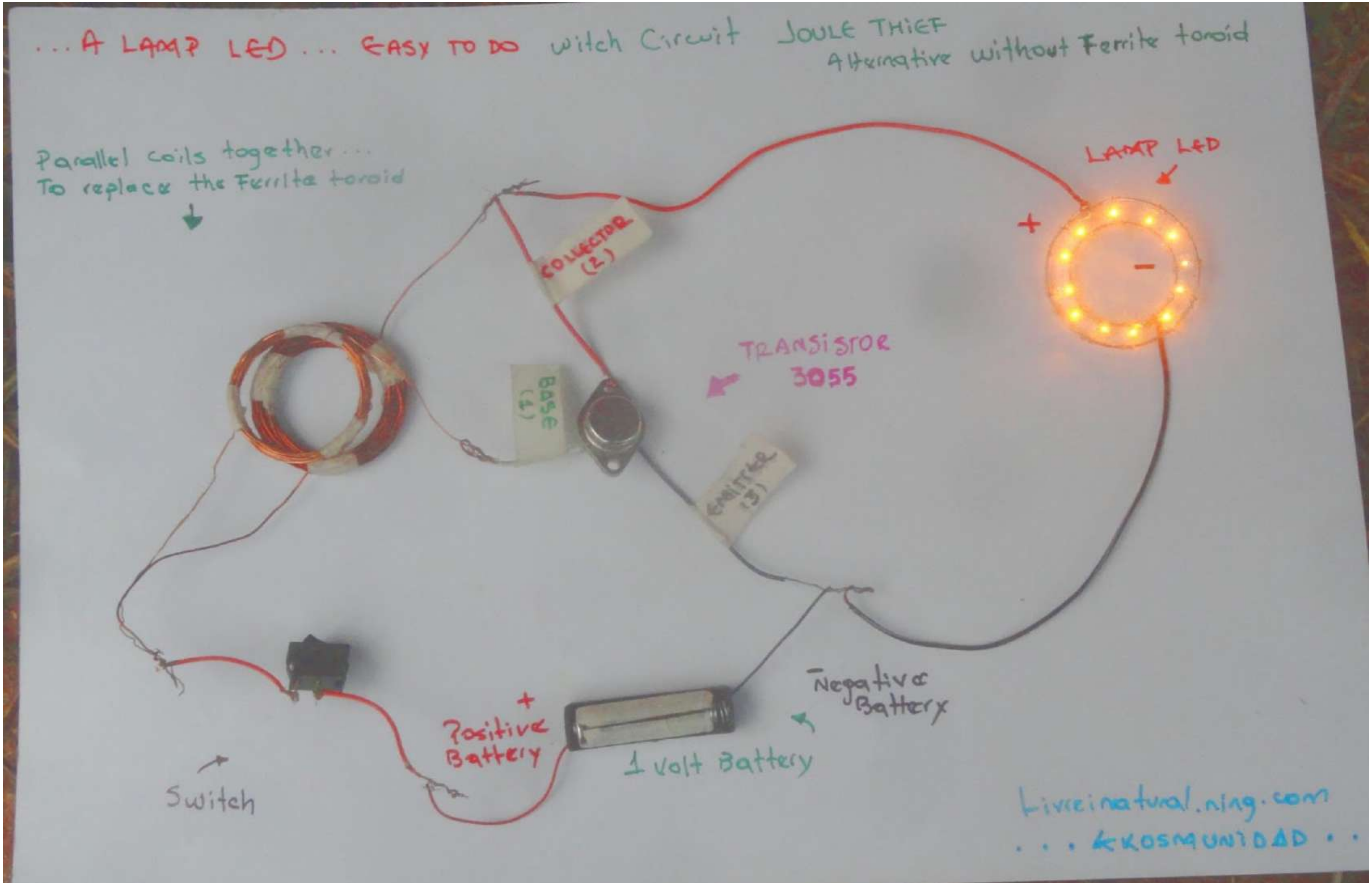... A LAMP LED ... EASY TO DO witch Circuit JOULE THIEF
Alternative without Ferrite toroid
Parallel coils together ...
To replace the Ferrite toroid
COLLECTOR (2)
BASE (1)
TRANSISTOR
3055
EMITTER (3)
LAMP LED
+
−
Switch
+
Positive Battery
1 volt Battery
−Negative Battery
Livreinatural.ning.com
... EKOSMUNIDAD ..

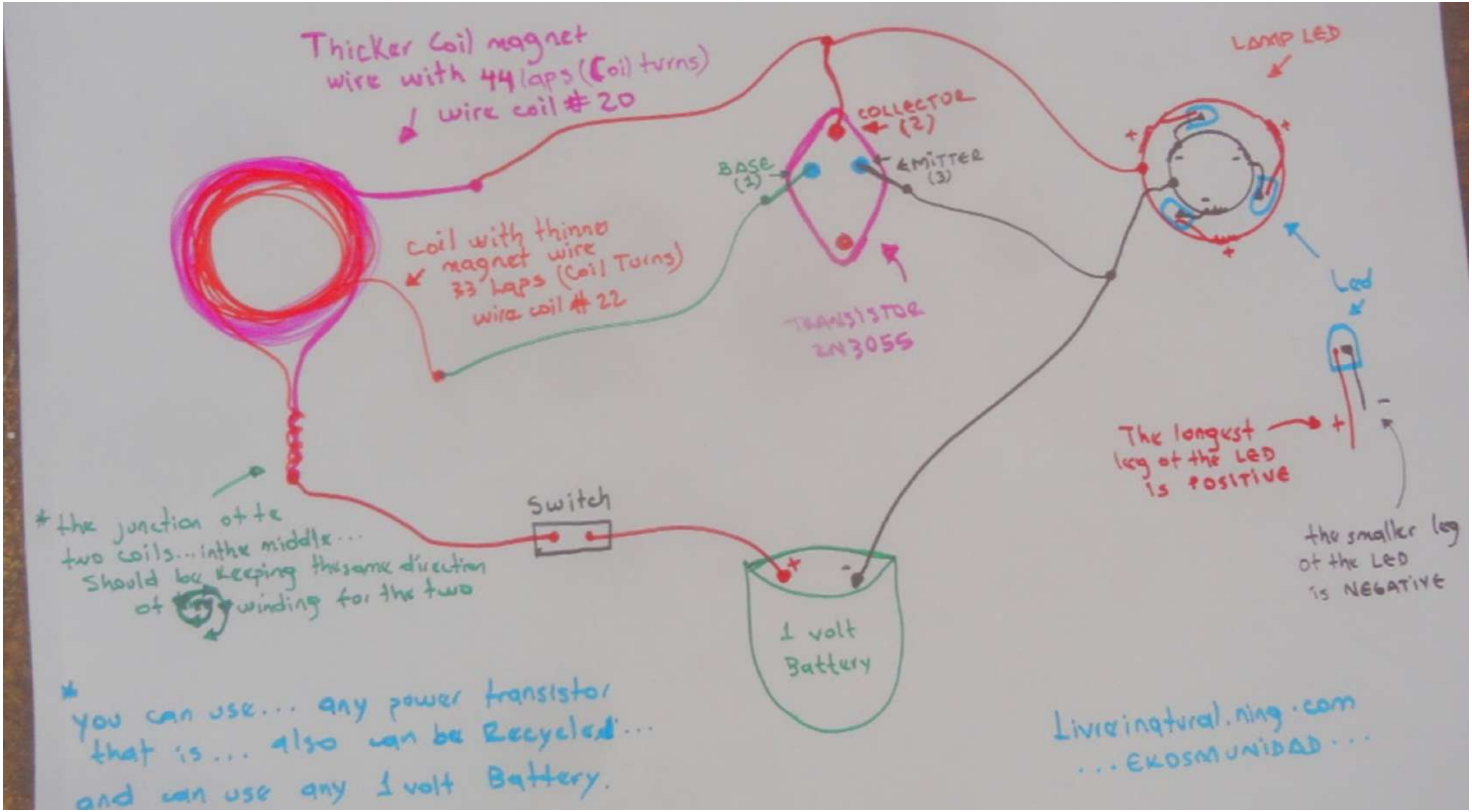

Thicker Coil magnet
wire with 44 laps (Coil turns)
wire coil # 20
coil with thinner
magnet wire
33 laps (Coil Turns)
wire coil # 22
BASE
(1)
COLLECTOR
(2)
EMITTER
(3)
TRANSISTOR
2N3055
LAMP LED
+
+
+
+
Led
+
-
Switch
+
-
1 volt
Battery

The longest leg of the LED is POSITIVE

the smaller leg of the LED is NEGATIVE

* the junction of te two coils... in the middle... Should be keeping the same direction of winding for the two

* you can use... any power transistor that is... also can be Recycled... and can use any 1 volt Battery.

Livreinatural.ning.com
...EKOSMUNIDAD...

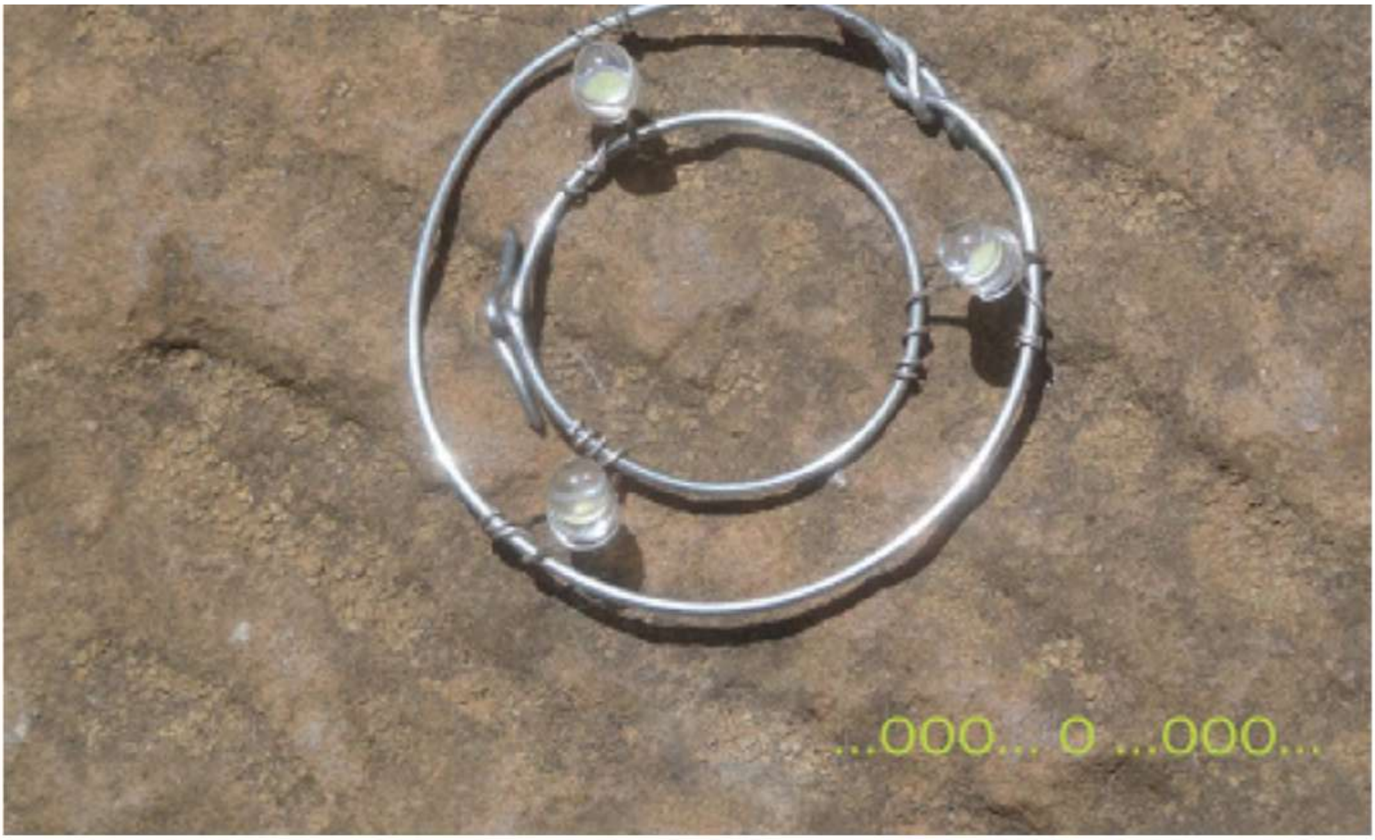
...OOO... O ...OOO...

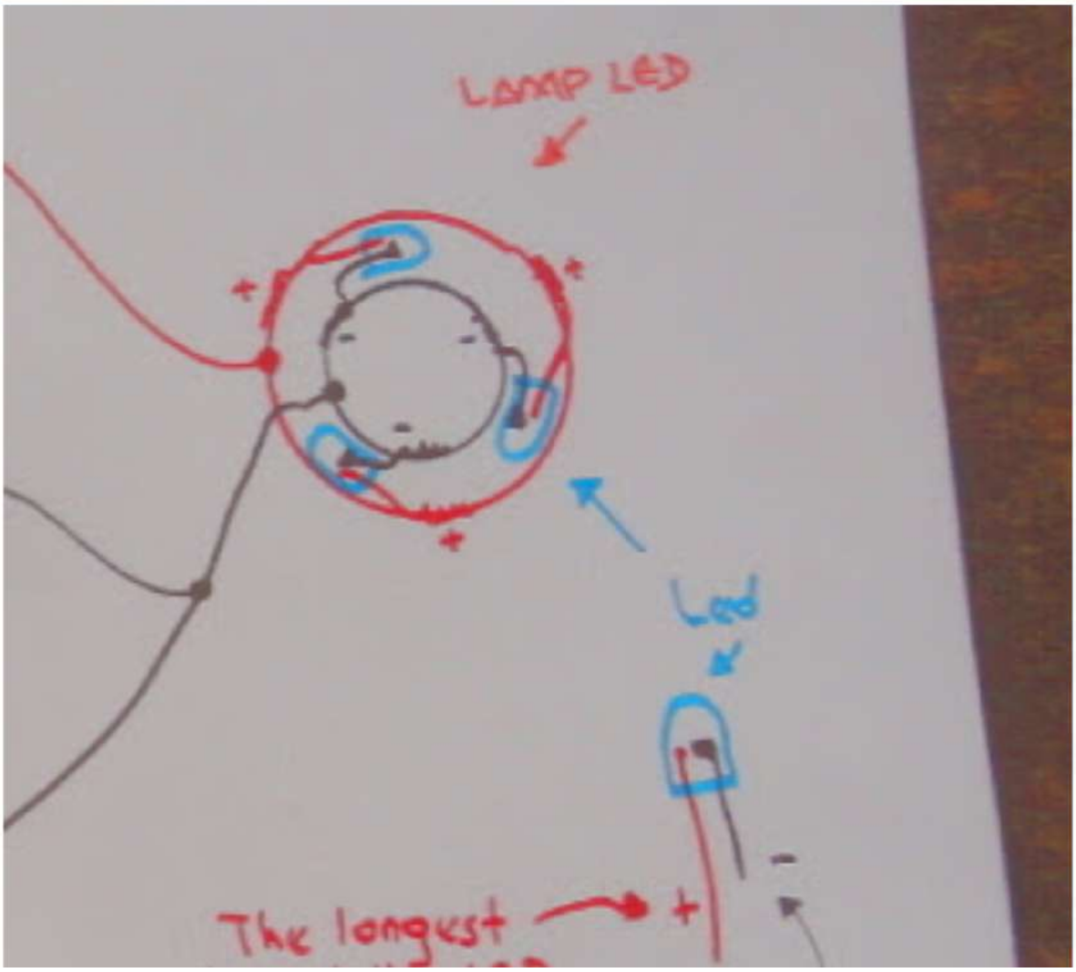
Lamp LED
+
+
+
-
-
-
Led
+
-
The longest

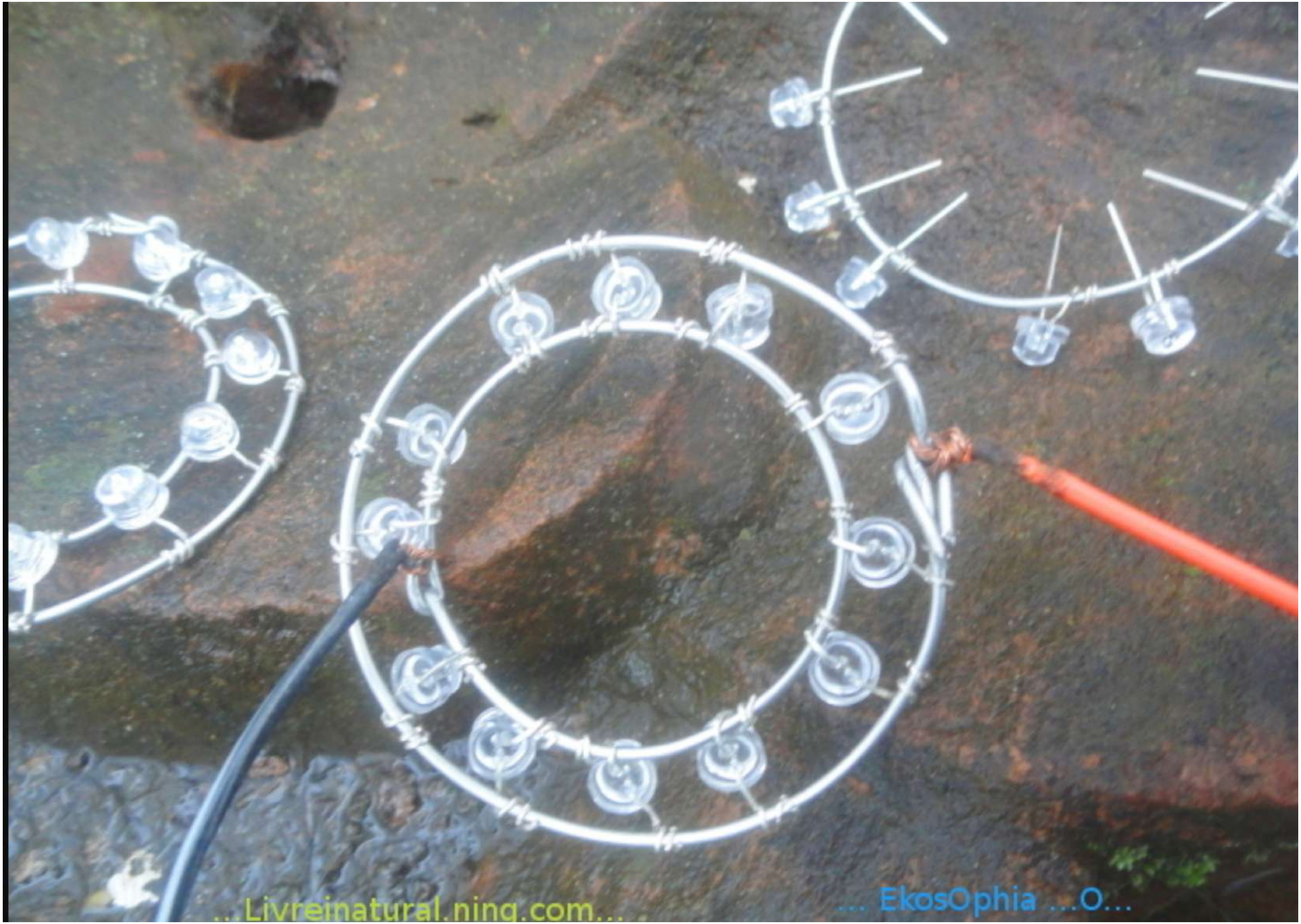
...Livreinatural.ning.com... ... EkosOphia ...O...

Livreinatural.ning.com...

... TRANSISTORES ...
3055

Vista trasera

2N3055

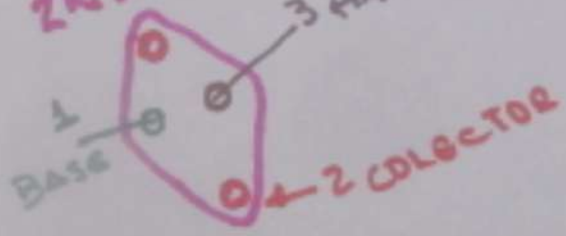
3 EMISOR
1
BASE
2 COLECTOR

Vista frontal

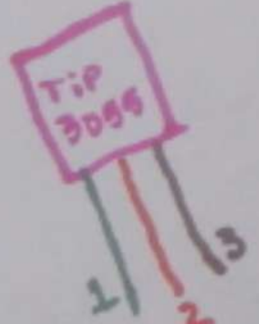
TIP
3055
1
2
3

1 → BASE
2 → COLECTOR
3 → EMISOR

... Livreinatural ...

... EKOSMUNIDAD ...

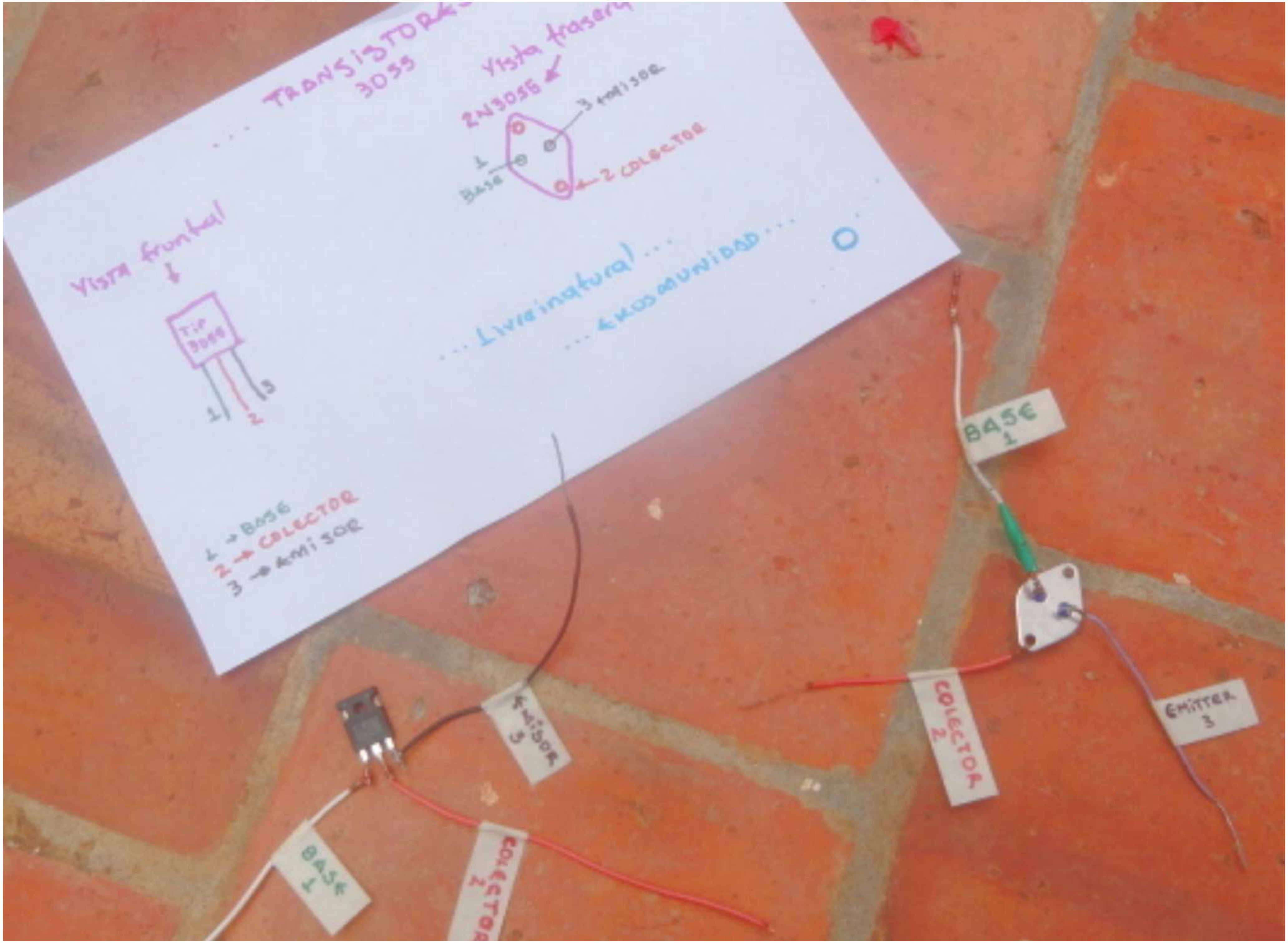

..Livrenatural.ning.com..

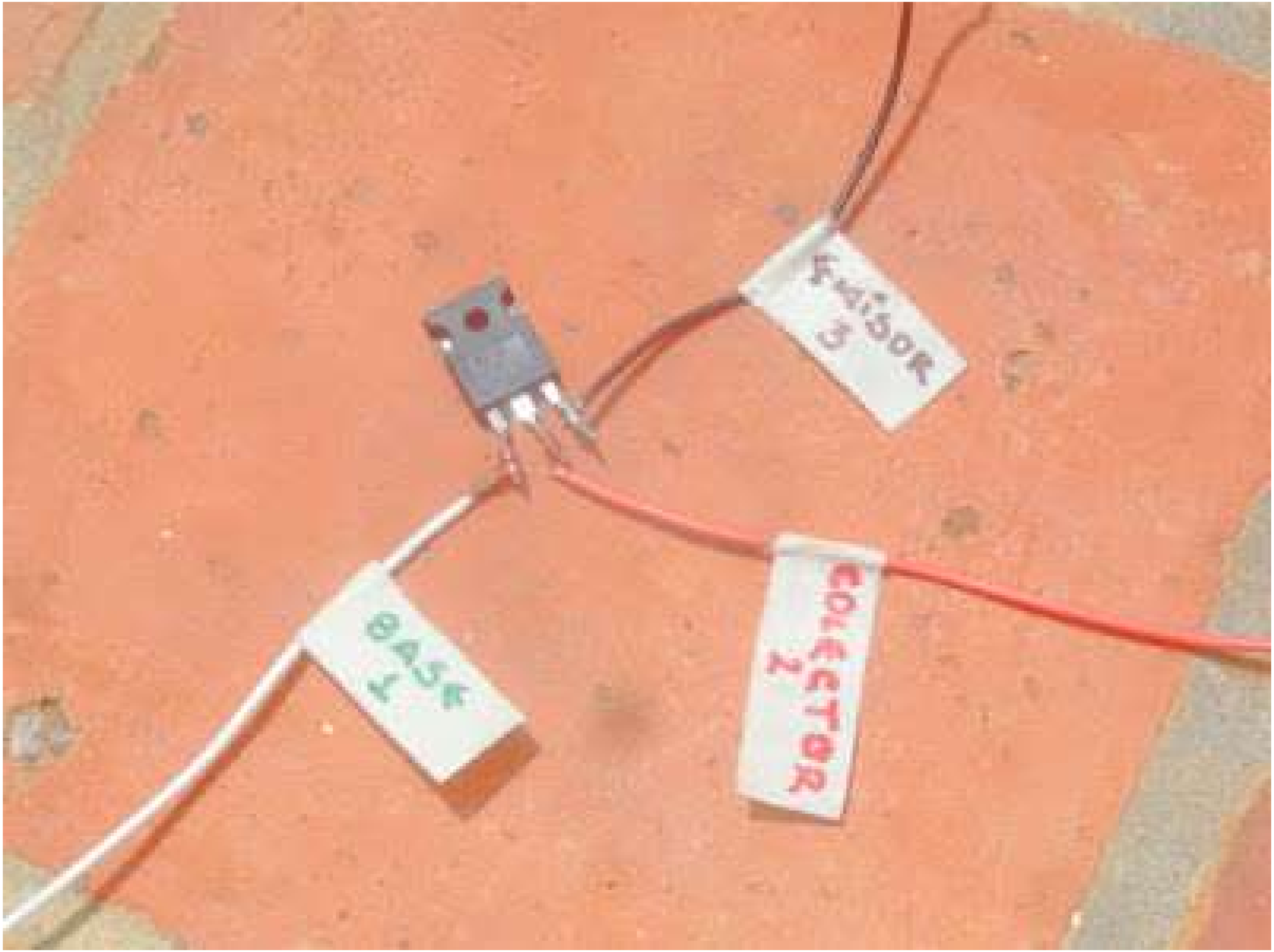
EMISOR
3
BASE
1
COLECTOR
2

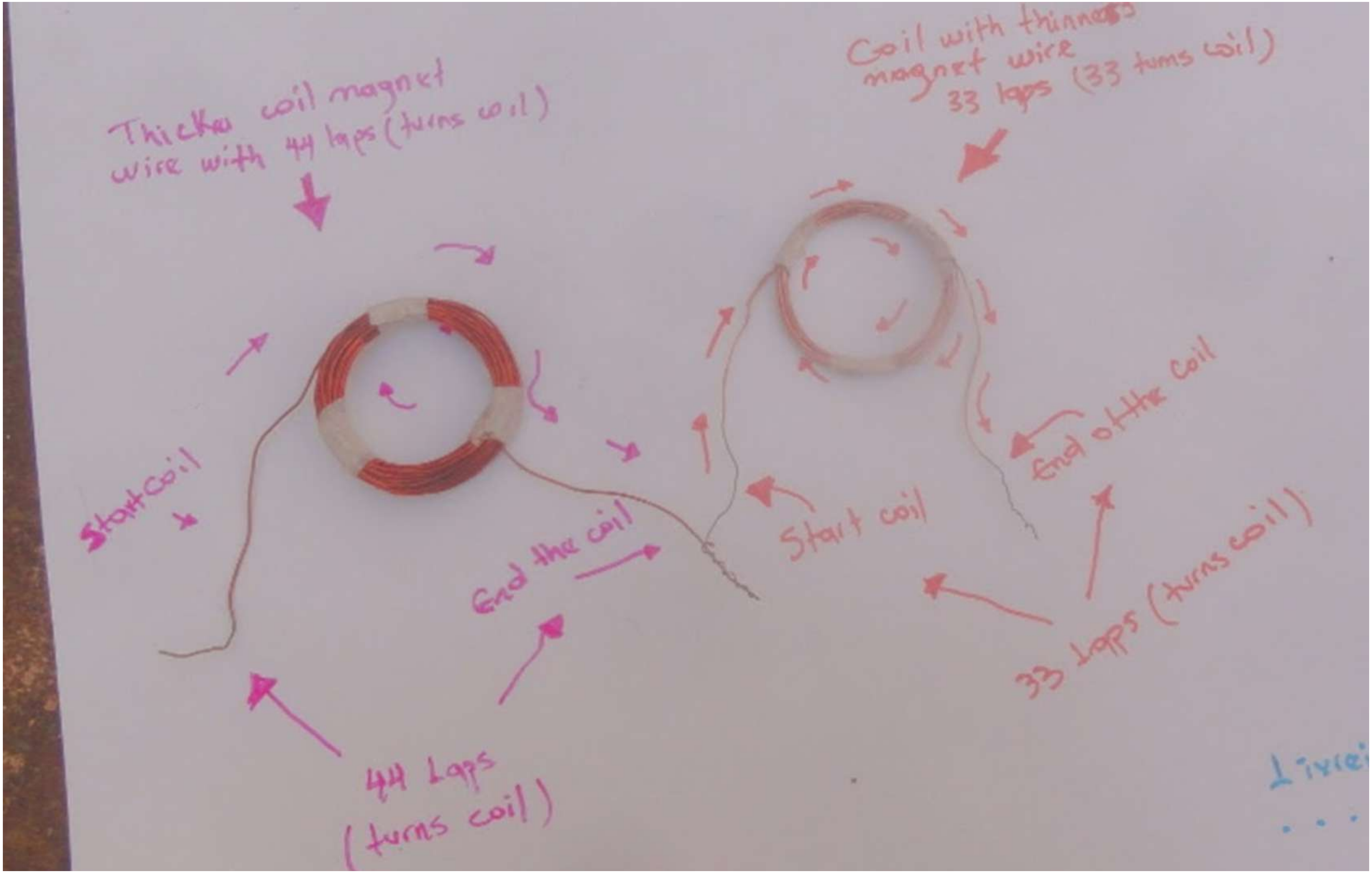
Thicker coil magnet wire with 44 laps (turns coil)
Start coil
End the coil
44 Laps (turns coil)
Coil with thinness magnet wire 33 laps (33 turns coil)
Start coil
End of the coil
33 Laps (turns coil)
Livrei

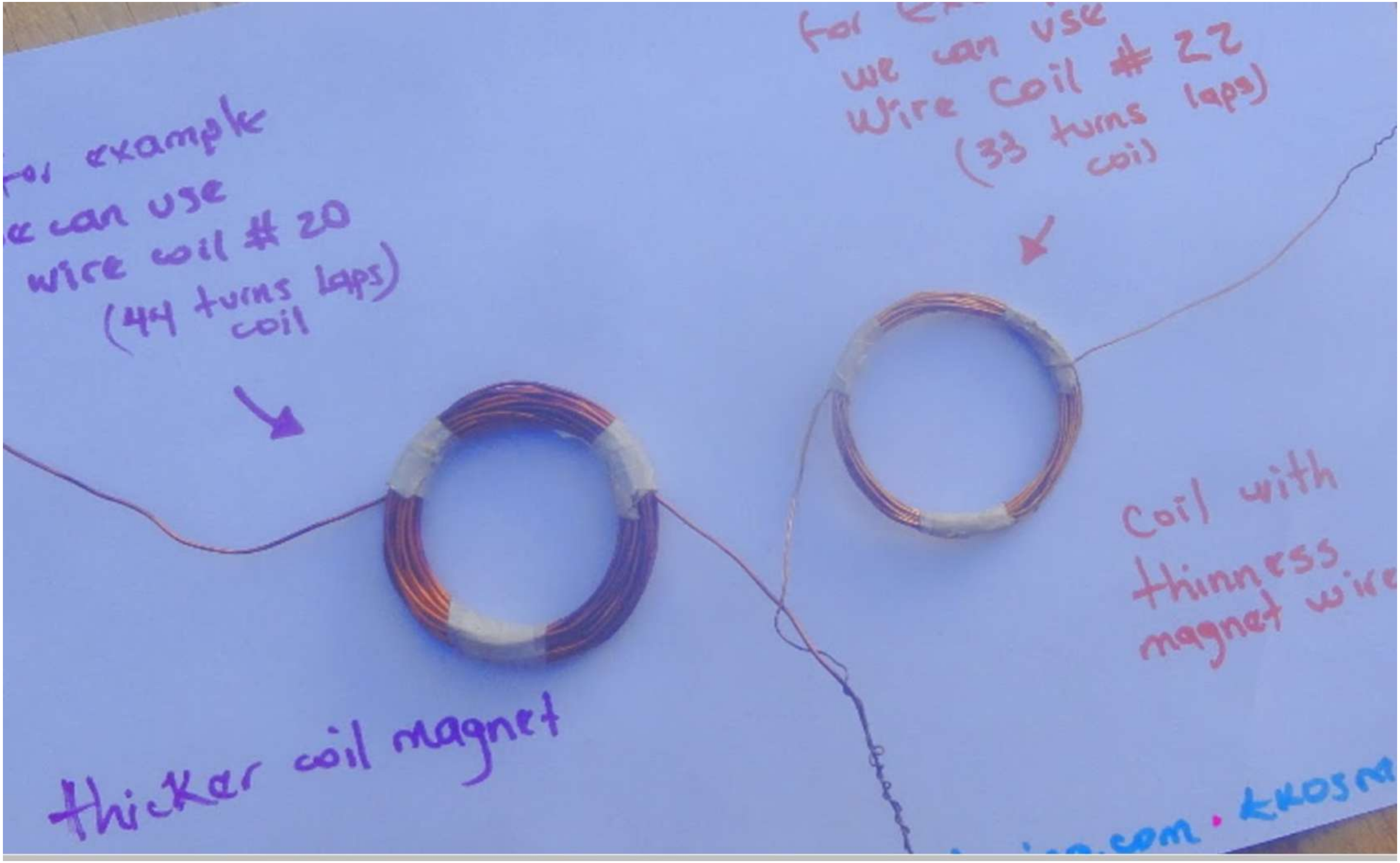
or example
e can use
wire coil # 20
(44 turns laps)
coil
for
we can use
wire coil # 22
(33 turns laps)
coil)
Coil with
thinness
magnet wire
thicker coil magnet
.com . kkosm

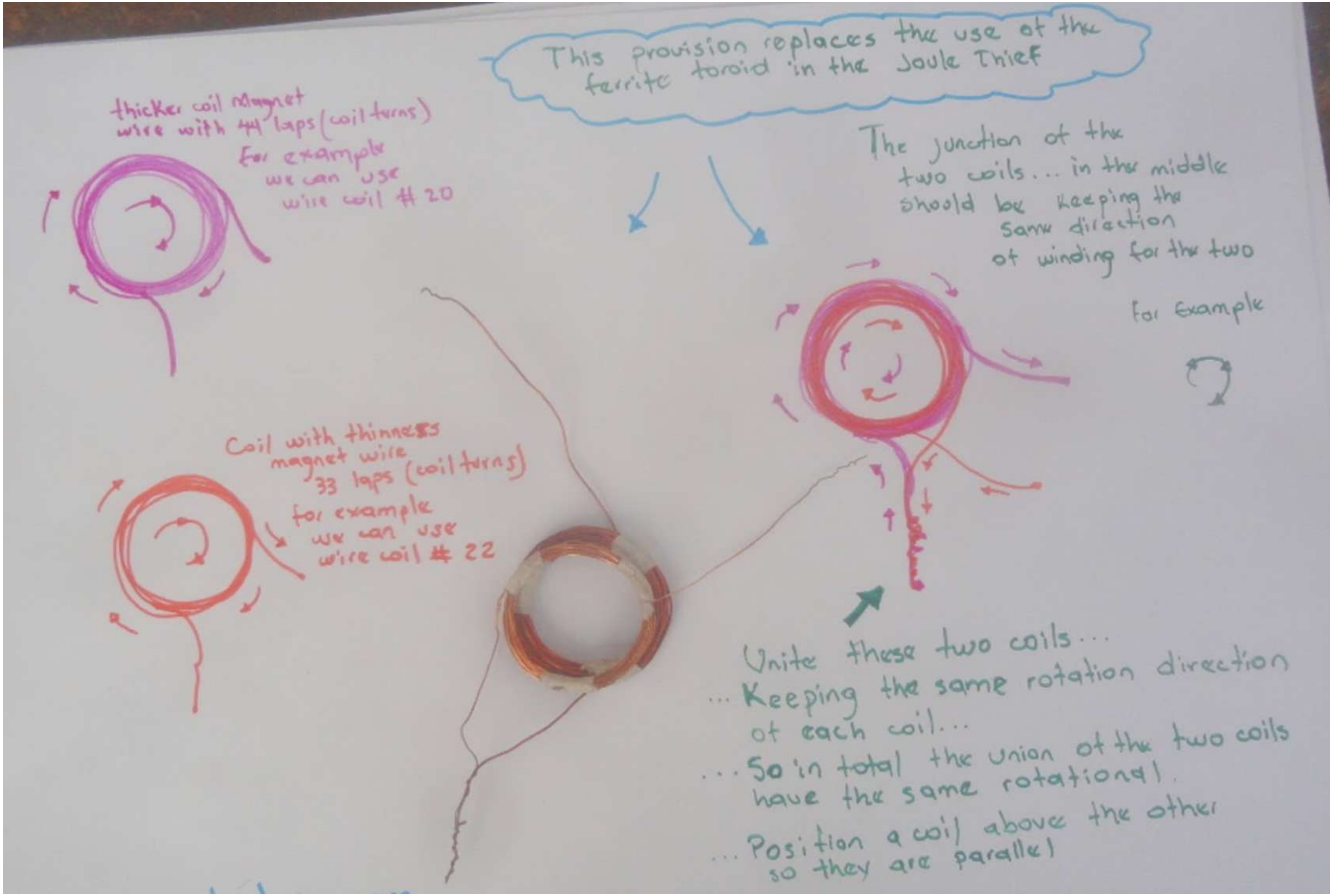

This provision replaces the use of the
ferrite toroid in the Joule Thief

thicker coil magnet
wire with 44 laps (coil turns)
for example
we can use
wire coil # 20

Coil with thinness
magnet wire
33 laps (coil turns)
for example
we can use
wire coil # 22

The junction of the
two coils... in the middle
should be keeping the
same direction
of winding for the two

for Example

Unite these two coils...
...Keeping the same rotation direction
of each coil...
...So in total the union of the two coils
have the same rotational.
...Position a coil above the other
so they are parallel

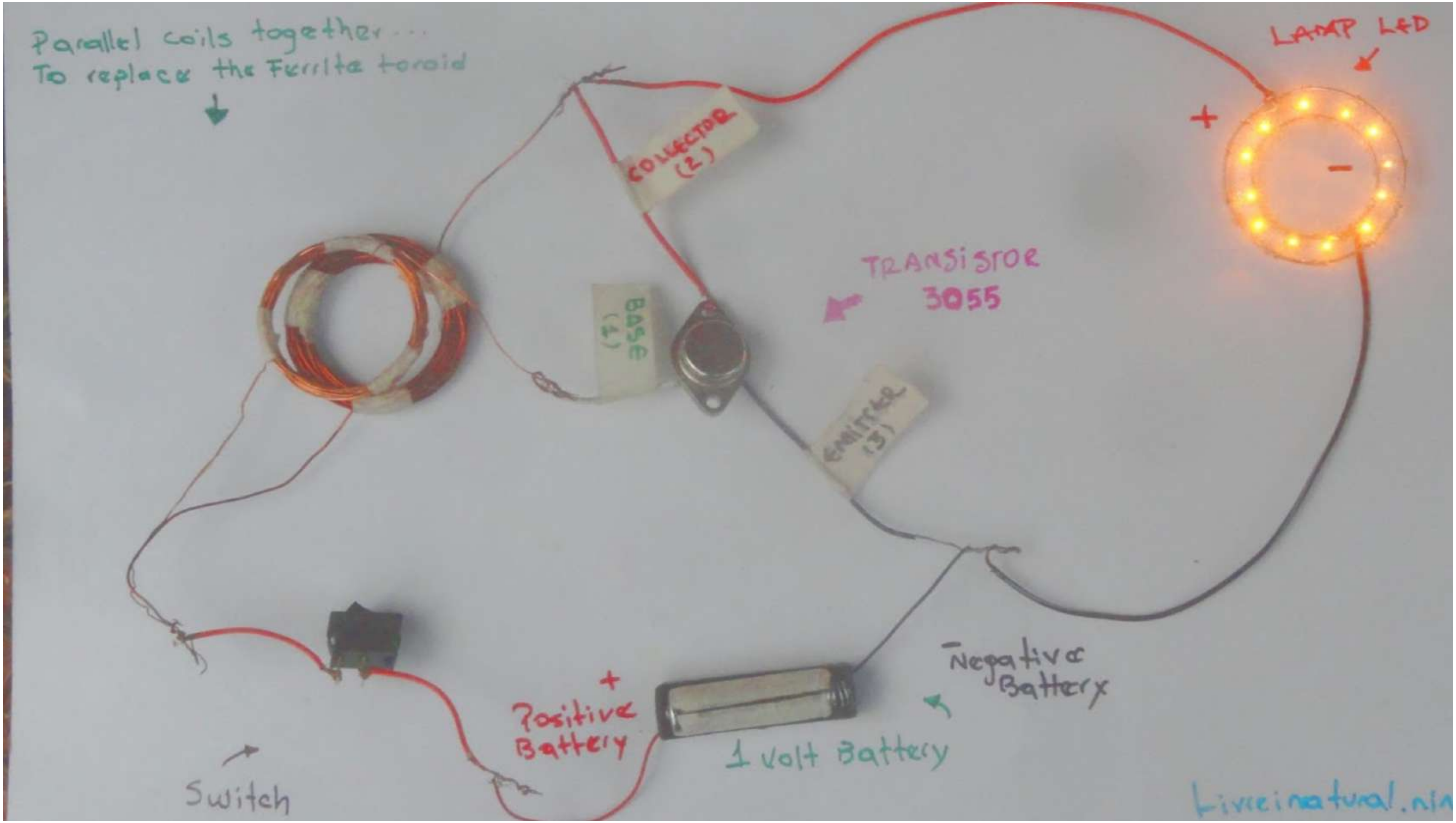
Parallel coils together...
To replace the Ferrite toroid
LAMP LED
+
-
COLLECTOR (2)
BASE (1)
TRANSISTOR 3055
EMITTER (3)
-Negative Battery
+ Positive Battery
1 volt Battery
Switch
Livreinatural.nin

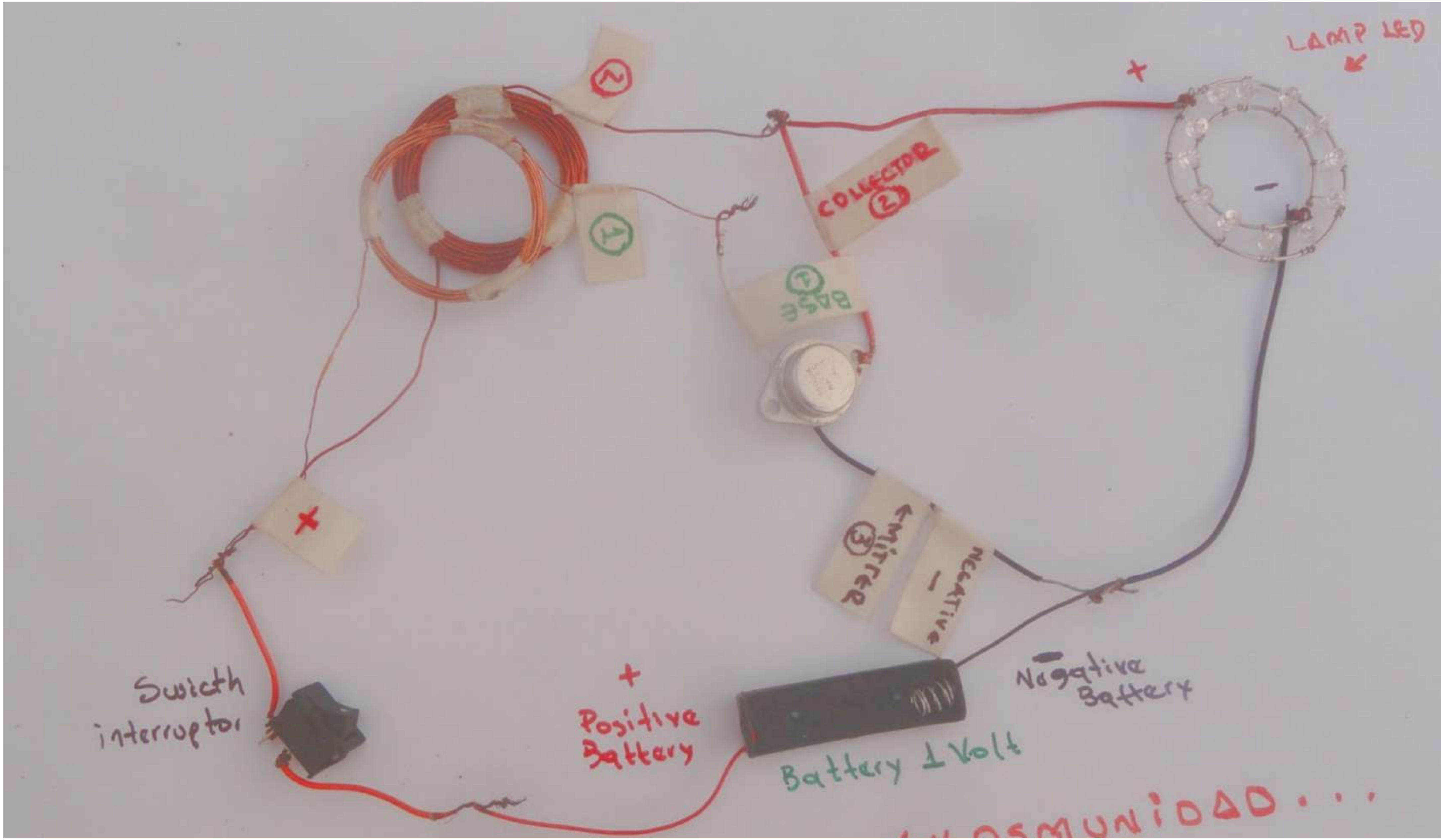
LAMP LED
+
②
①
COLLECTOR ②
BASE ①
-
+
EMITTER ③
NEGATIVE -
Swicth interruptor
+ Positive Battery
Battery 1 Volt
- Negative Battery
OSMUNIDAD...

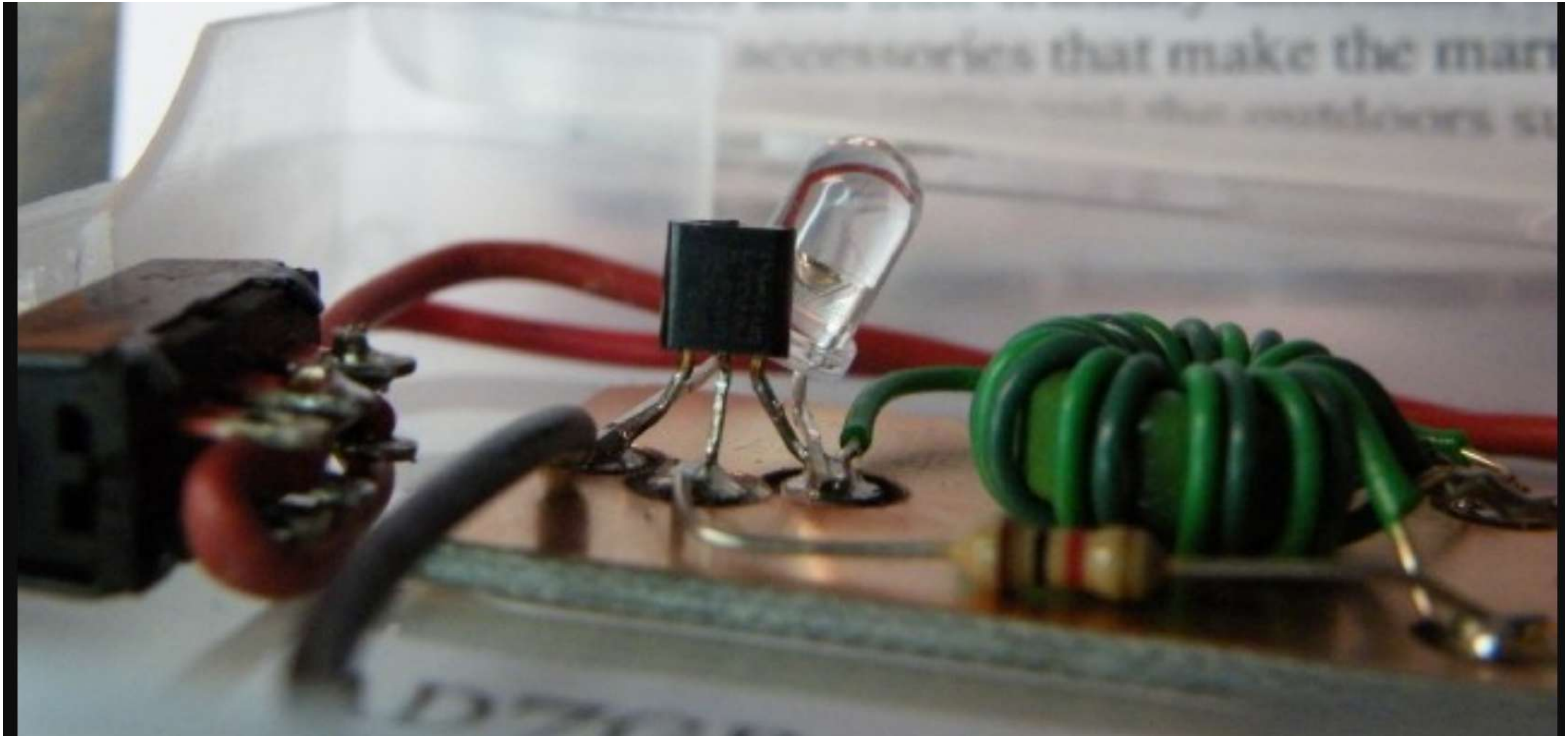

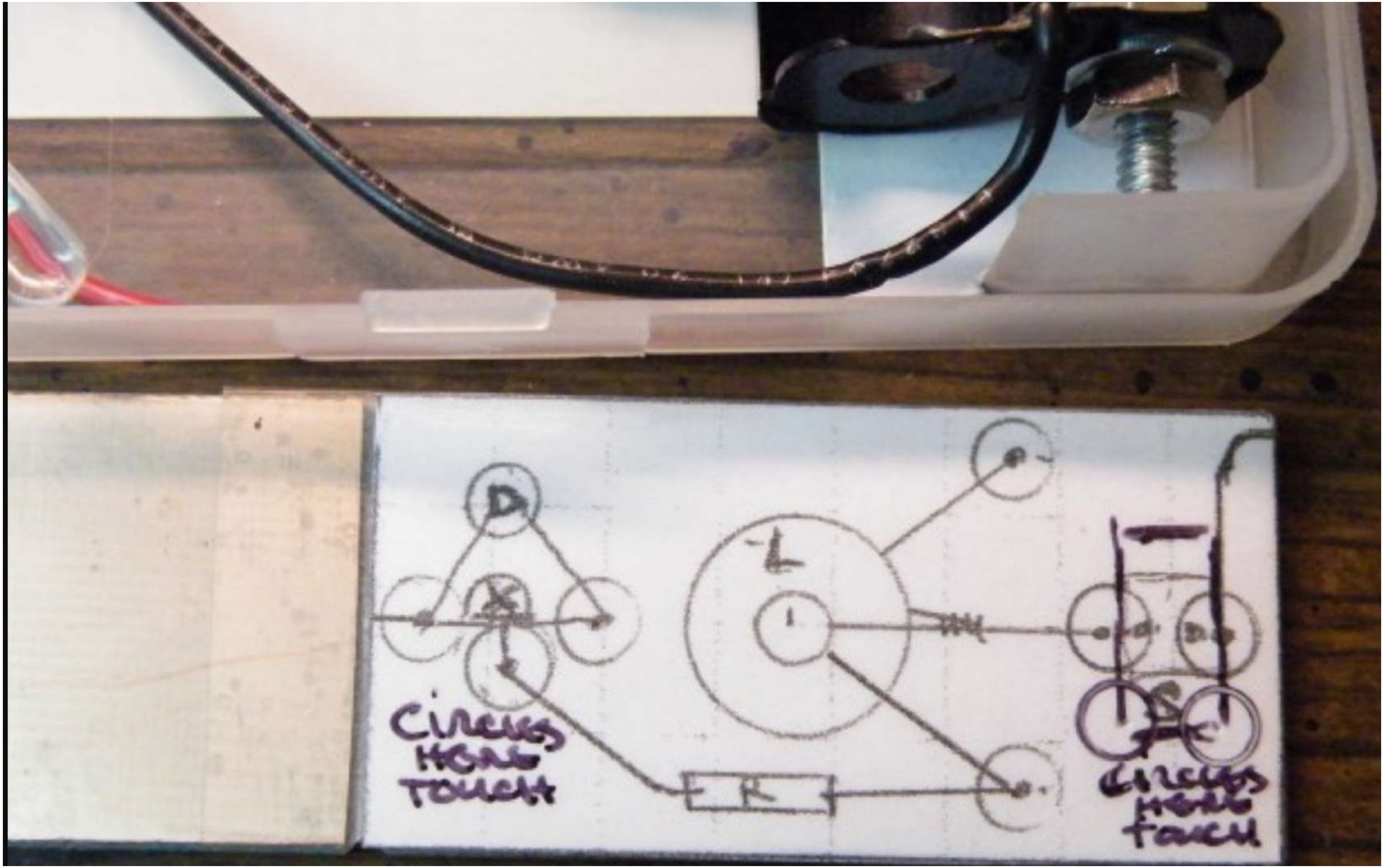
D
X
L
R
S
Circles here touch
Circles here touch

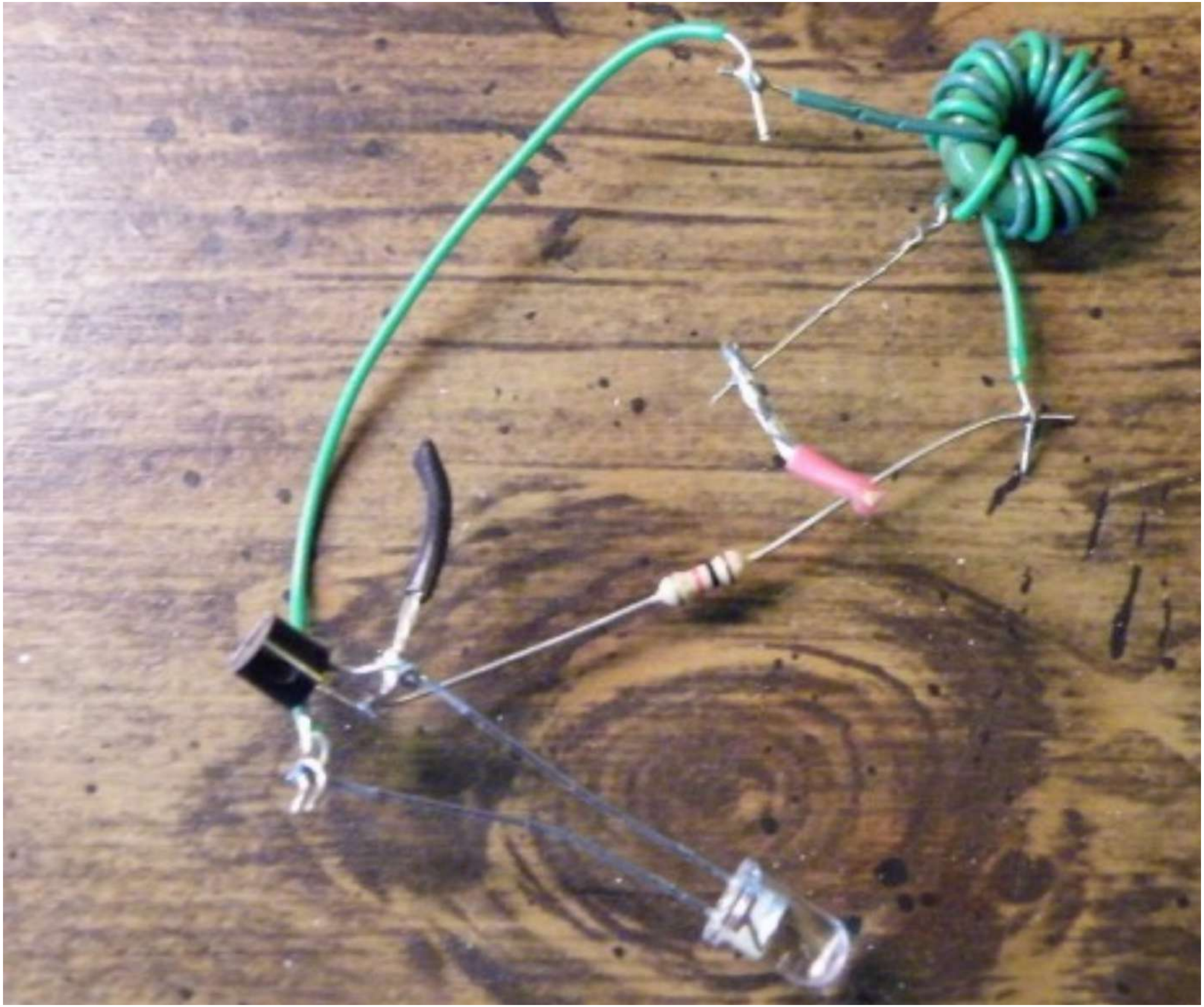

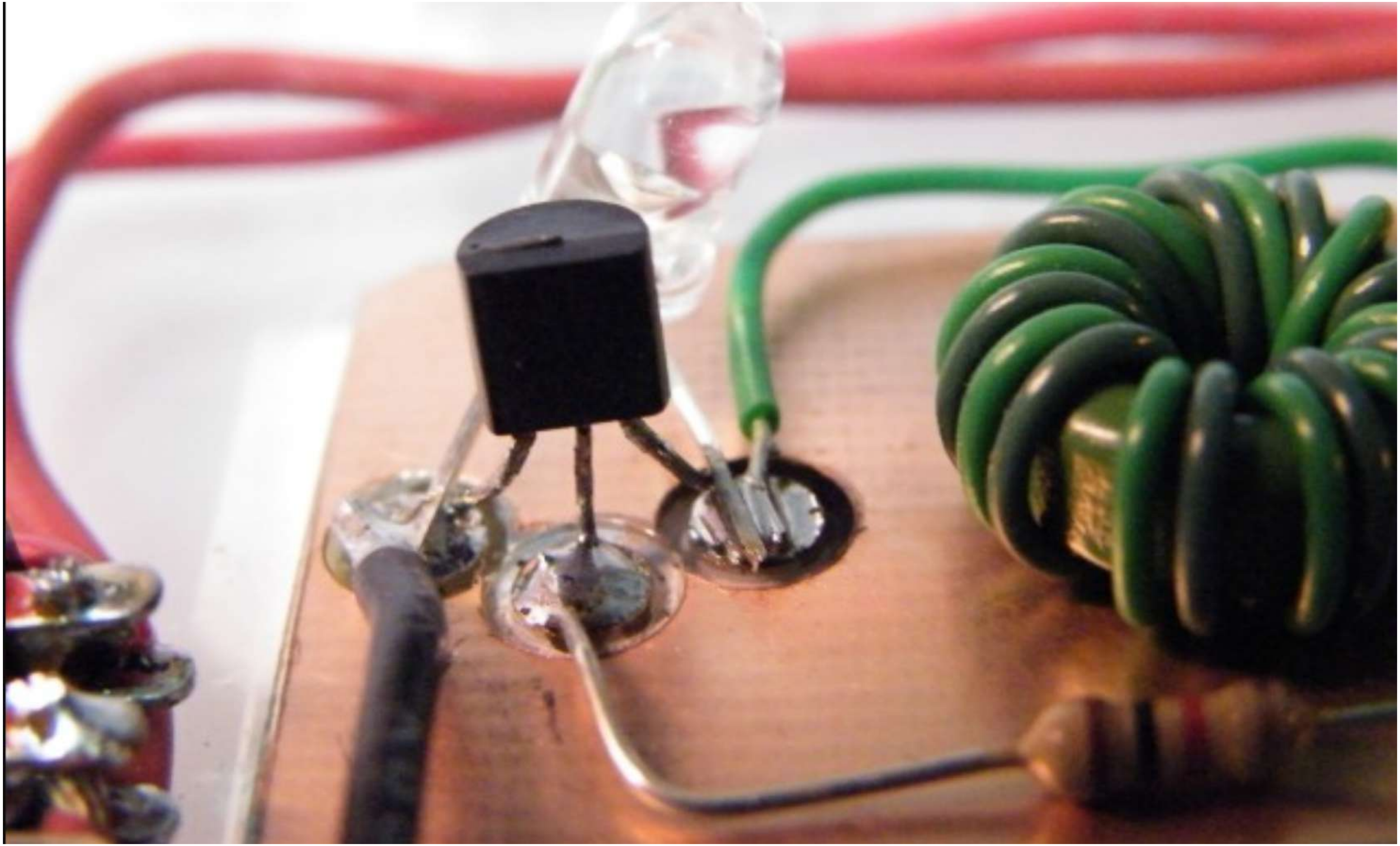

age

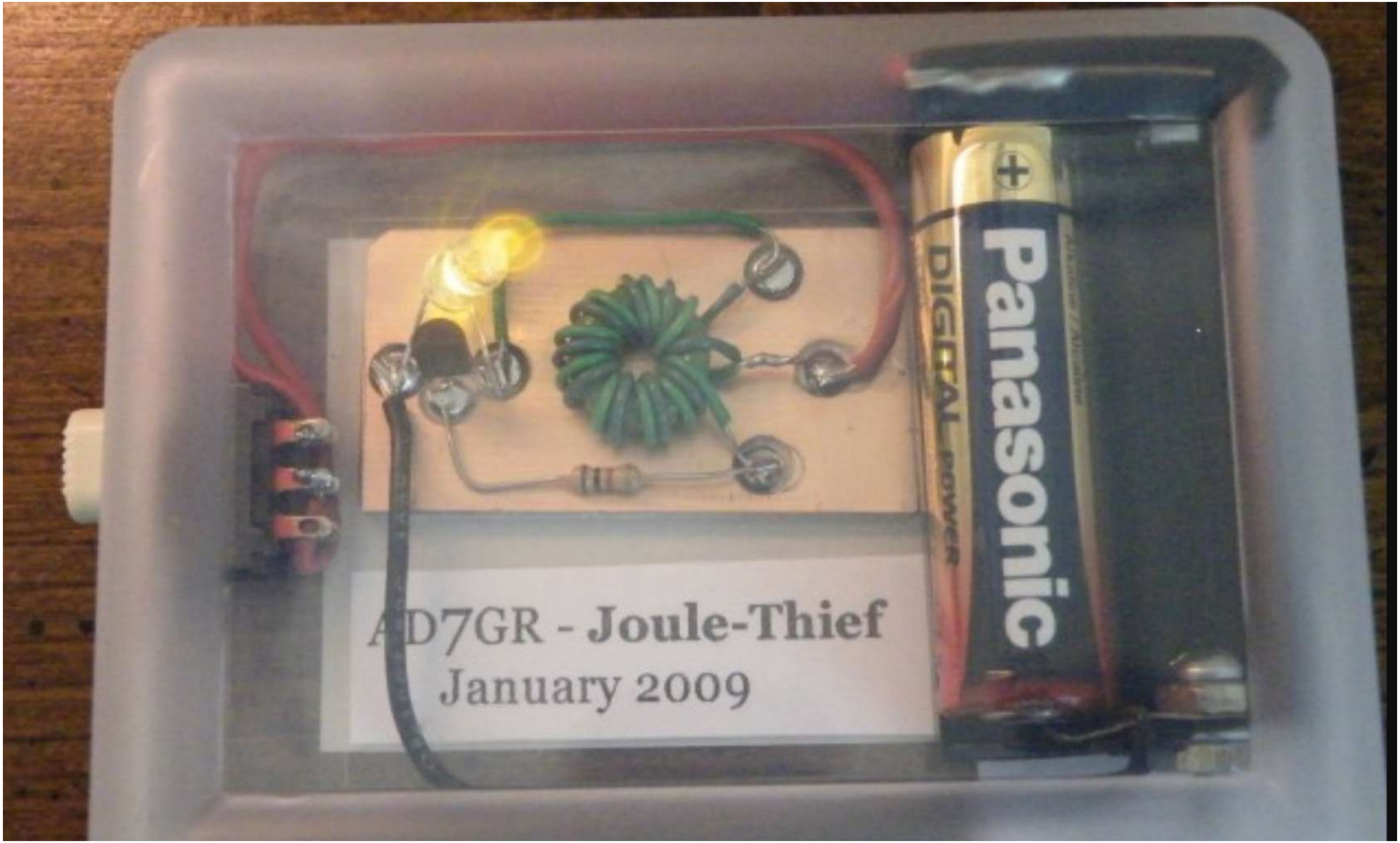
AD7GR - Joule-Thief
January 2009
Panasonic
DIGITAL POWER

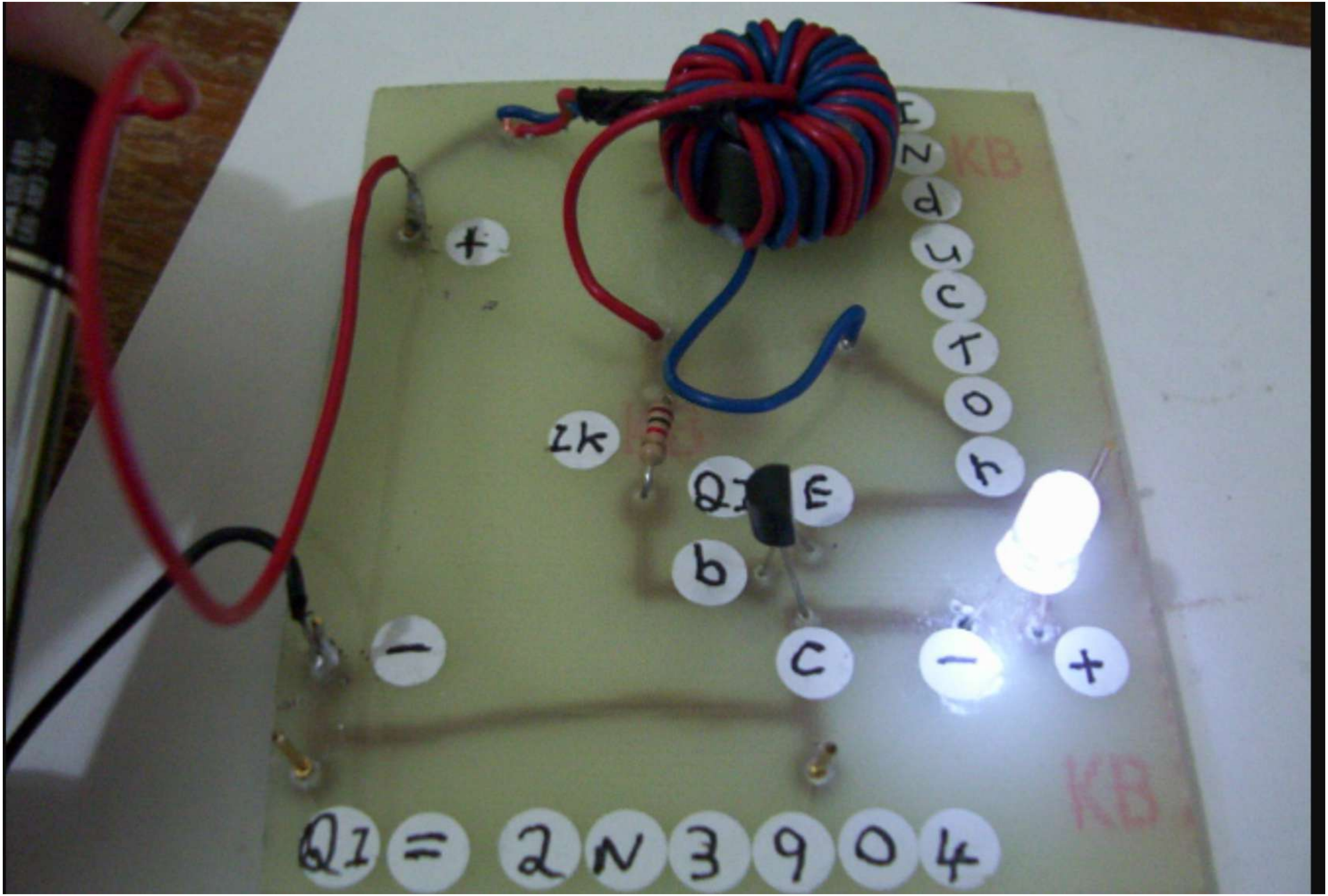

Inductor
KB
+
2k
Q1
E
b
C
−
−
+
KB
Q1 = 2N3904

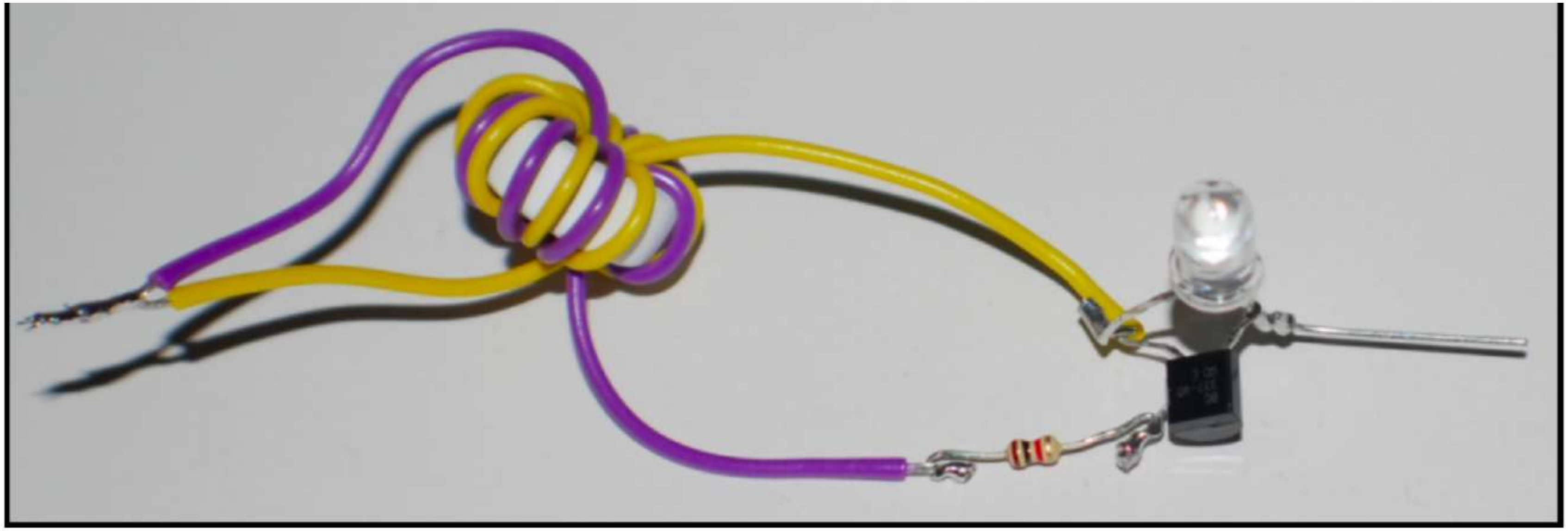

… and these are the components you need:

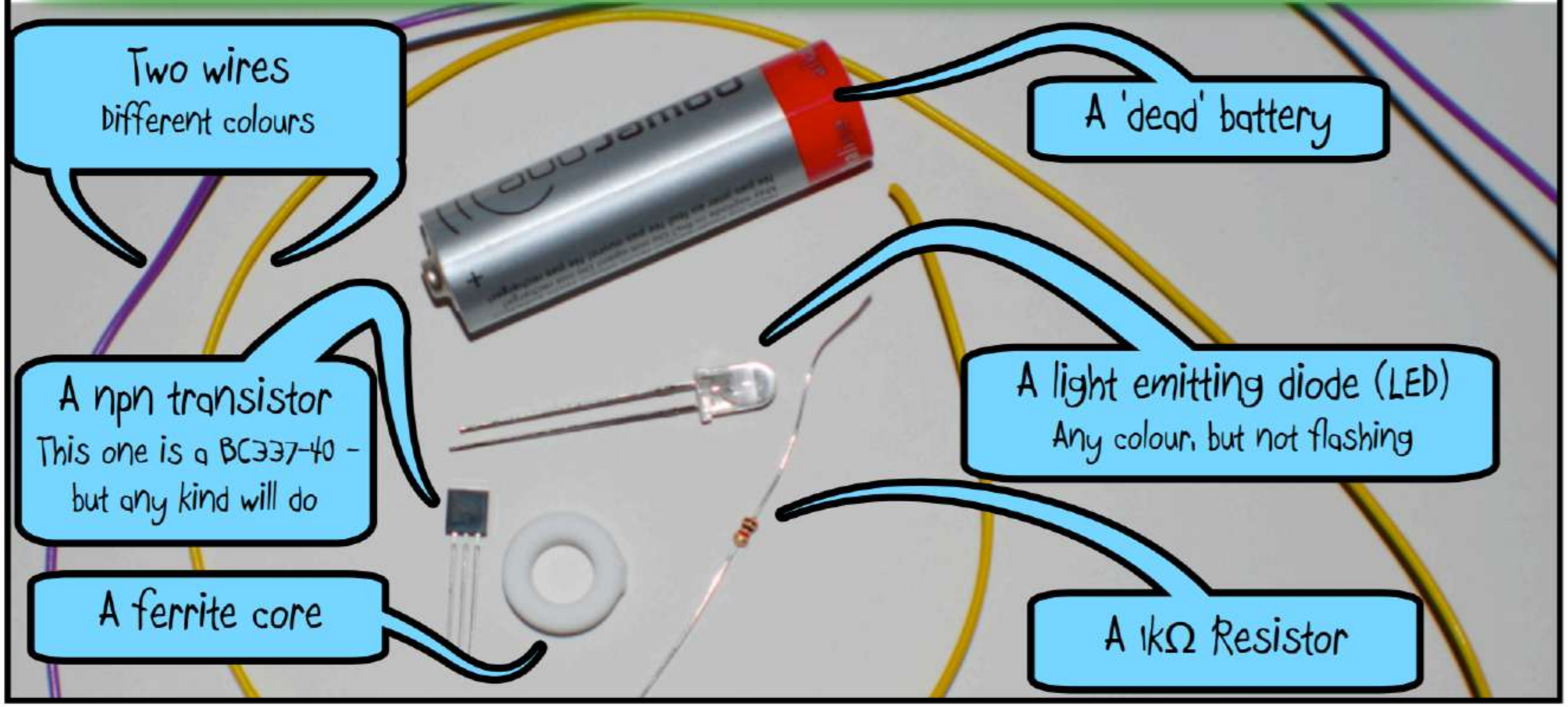
Two wires
Different colours
A 'dead' battery
A npn transistor
This one is a BC337-40 -
but any kind will do
A light emitting diode (LED)
Any colour, but not flashing
A ferrite core
A 1kΩ Resistor

Solder the 1kΩ Resistor to either of the unsoldered wires coming from the inductor.

Solder the other end of the resistor to the base lead of the transistor.

On the transistor used here, that is the middle lead.

Cut off the spare ends of leads.

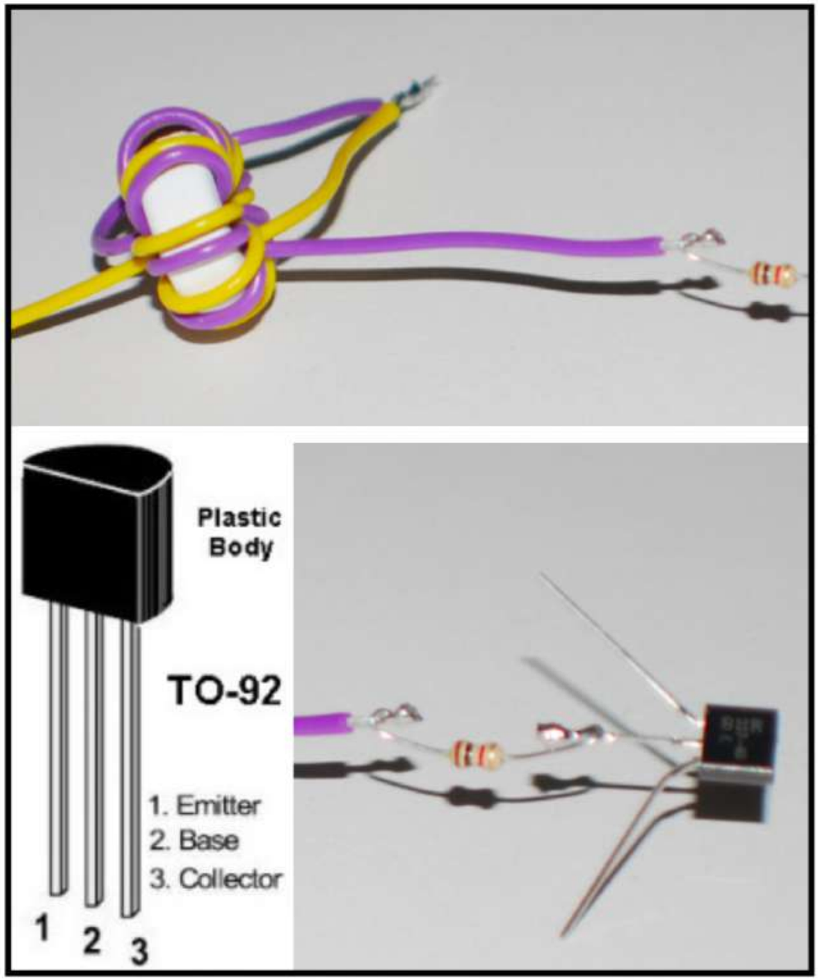
Plastic Body
TO-92
1. Emitter
2. Base
3. Collector
1
2
3

Solder the anode of the LED (the longer leg) to the collector lead of the transistor AND the remaining wire from the inductor.

Cut off the ends of the leads.

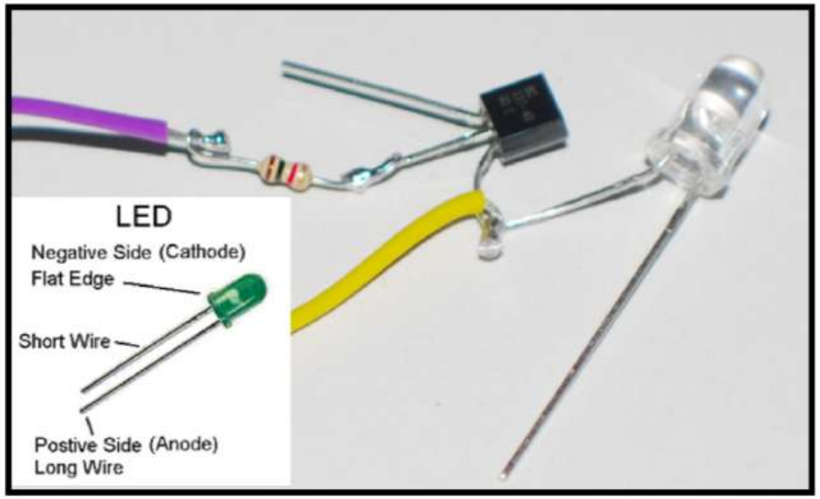
LED
Negative Side (Cathode)
Flat Edge
Short Wire
Postive Side (Anode)
Long Wire

Solder the other leg of the LED (the cathode) to the emitter lead of the transistor.
Do NOT cut off the LED leg.
(Don't wory if you have cut it off - you'll just need to solder another wire to this leg so that you can connect it to the battery.)

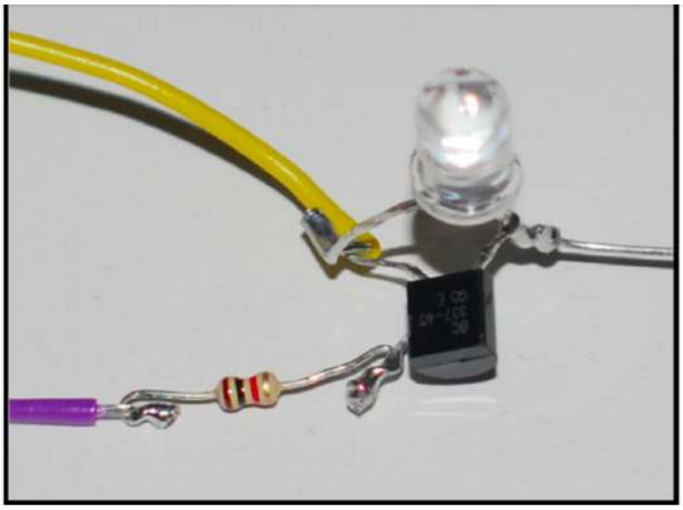

That's the circuit finished. Now use it.

First you need a 'dead' battery - that means one where the voltage is less than 1.3V.
You can see that the battery used here has a Voltage of just over 1V.

Rapid
1.037
M320

Connect the two soldered wires coming from the inductor to the positive (+) side of the battery.
Connect the uncut LED leg (the athode) to the negative (-) side of the battery.

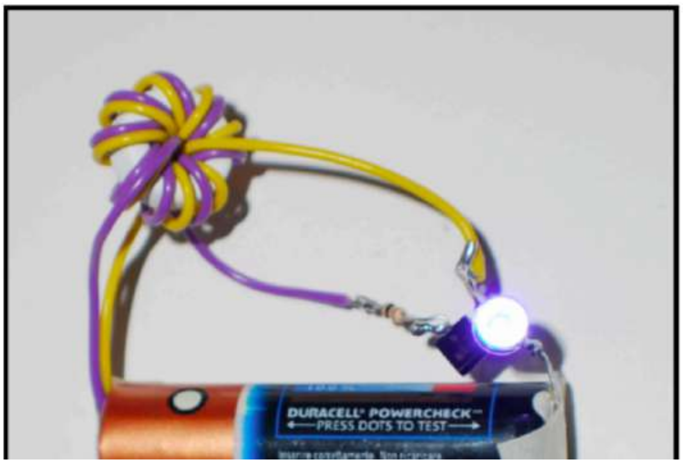

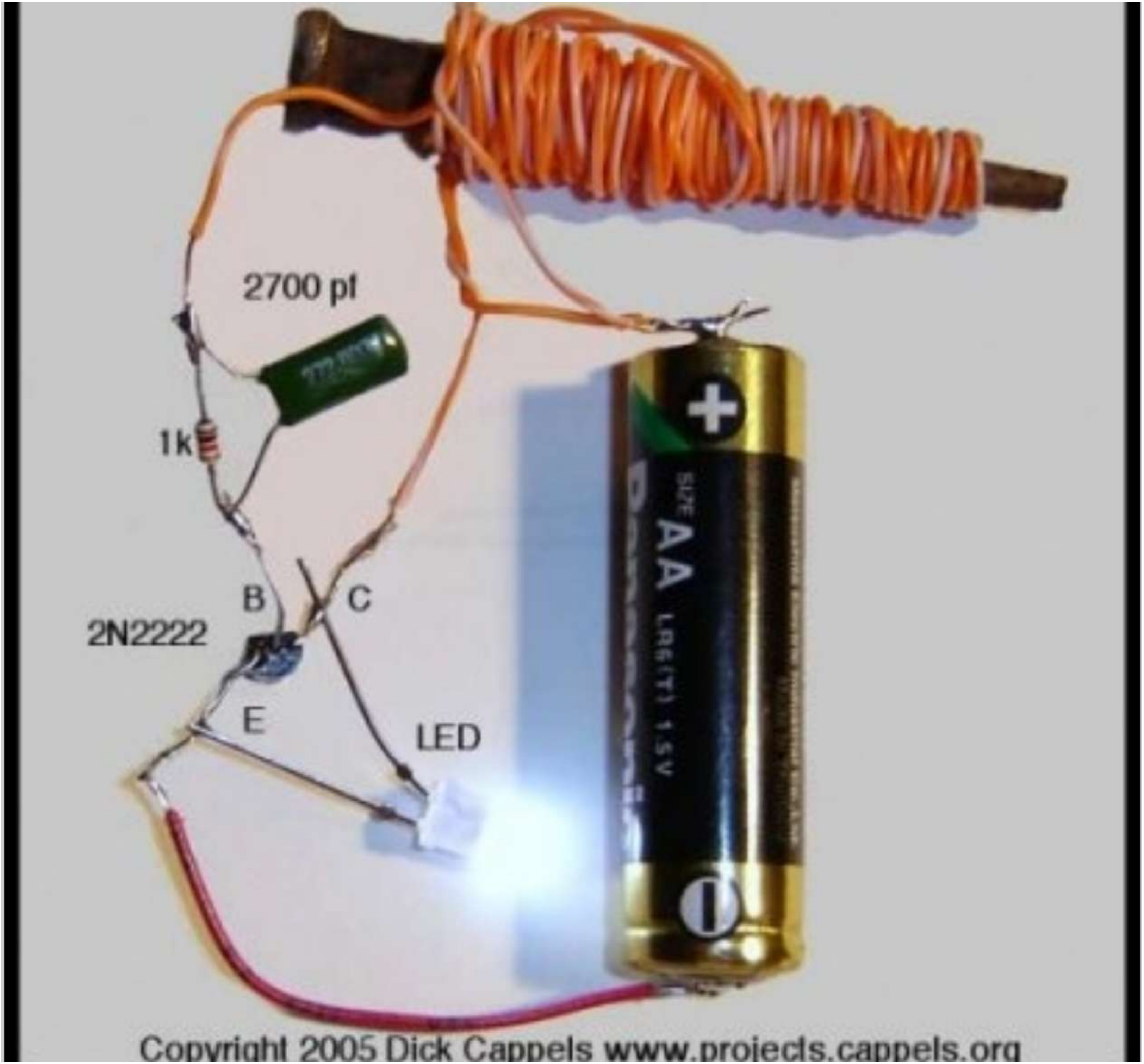
2700 pf
1k
B
C
2N2222
E
LED
+
SIZE AA LR6(T) 1.5V

Copyright 2005 Dick Cappels www.projects.cappels.org

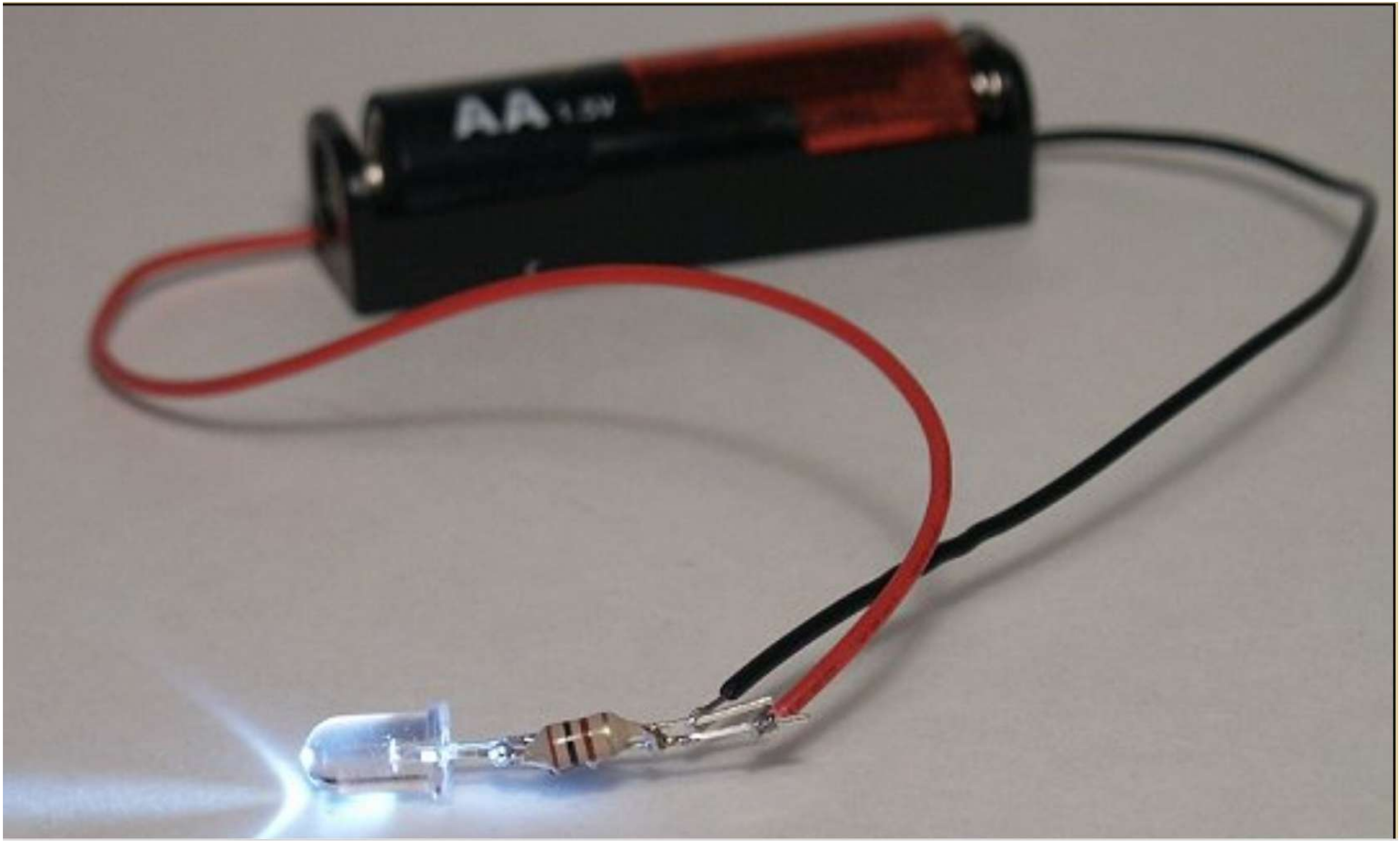
AA 1.5V

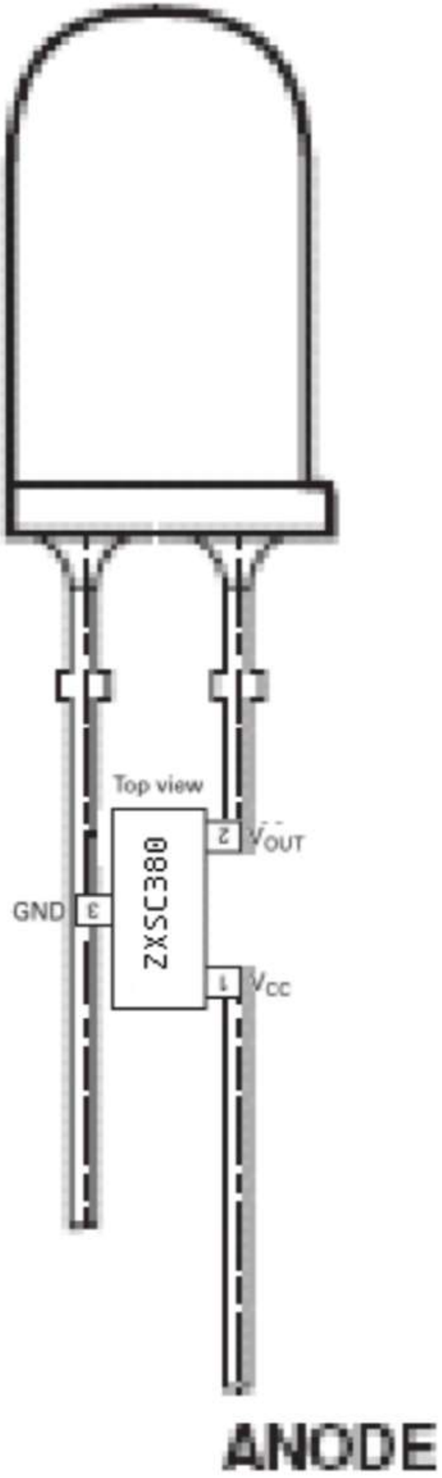

The chip is soldered onto the legs of the LED and then the wire between pins 1 and 2 is cut out so there is no connection. The inductor is then soldered so as to bridge pins 1 and 2. I soldered the inductor on before cutting out the wire but it might be easier to try some other method.

As I said previously, this is a tedious solder job, and the SOT23 IC is easily damaged (ask me how I know!) It takes patience and a steady hand to assemble these few parts.

This circuit would work for the LED indicator in the 1.5v boosters and would most likely be a better choice than the Joule Thief that I posted previously.

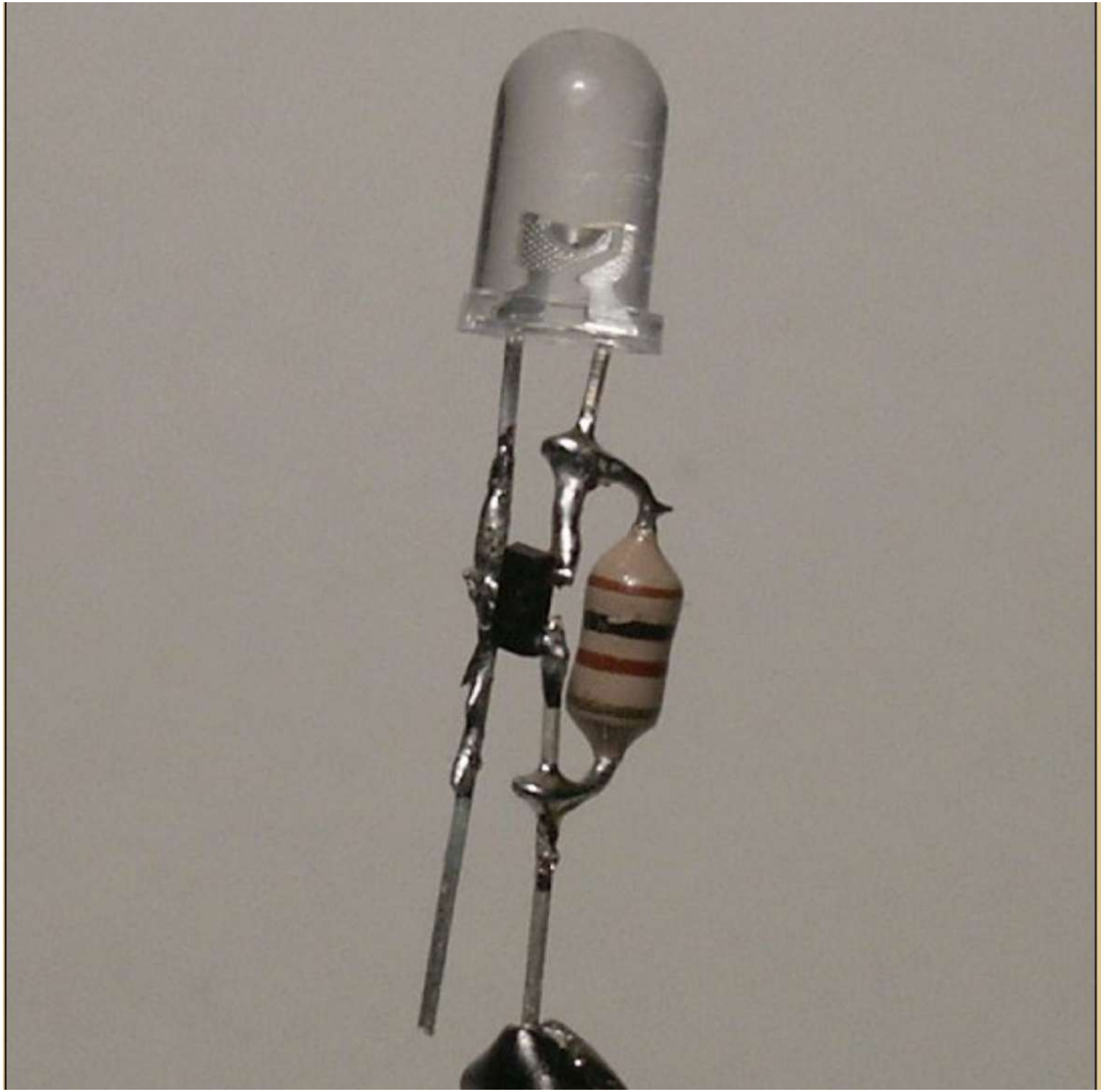

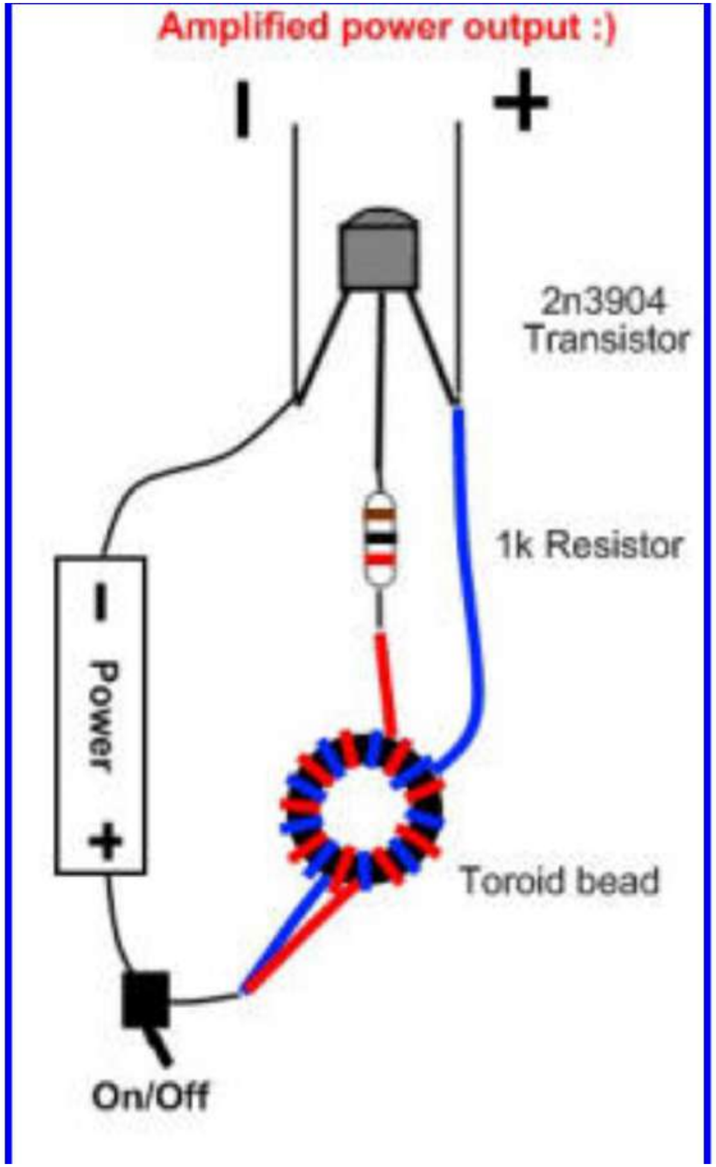
Amplified power output :)
–
+
2n3904 Transistor
1k Resistor
–
Power
+
Toroid bead
On/Off

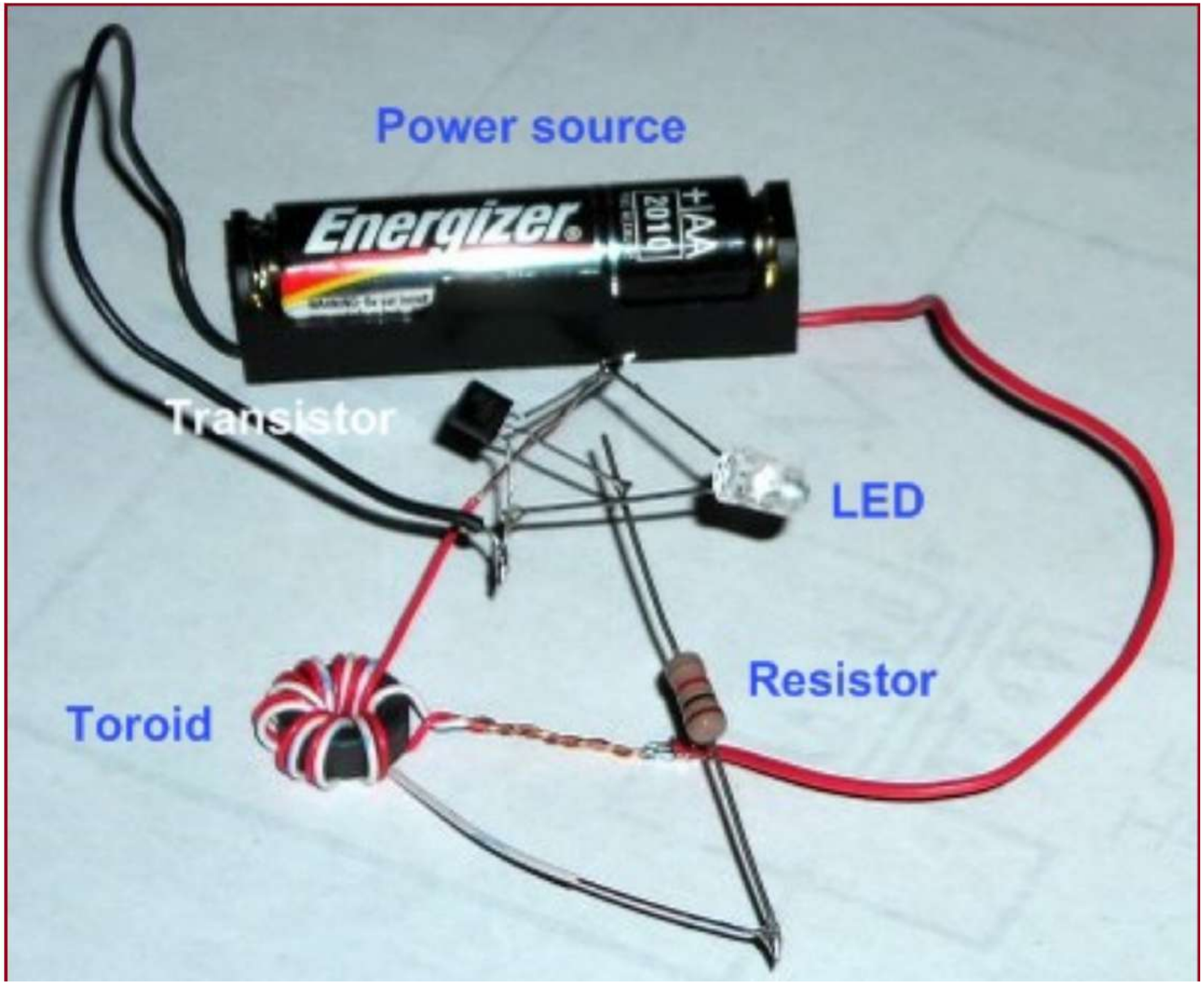
Power source
Energizer
+ AA
2010
Transistor
LED
Resistor
Toroid

The CIRCUIT

The circuit is very simple. All the work is done by the 5252F chip.
It contains an oscillator, a high speed diode and a power transistor. All these components inside the "IC" that looks like a 4-leaded transistor!
This IC is smaller and cheaper than all the parts individually and is less expensive than the competition (that costs 70 cents).
We can produce a complete project for a few dollars and it has two "test features."
You can place a LED across one of the LEDs on the board and find out the colour as many LEDs come in a "clear-as-glass" package and you cannot tell the colour until they are illuminated.
The other feature is INDUCTOR TESTING.
The current taken by the circuit changes according to the value of the inductor.
You just need a few reference values and you can work out the value of an inductor within the range of the inductors you have used as samples, or slightly higher or lower values.

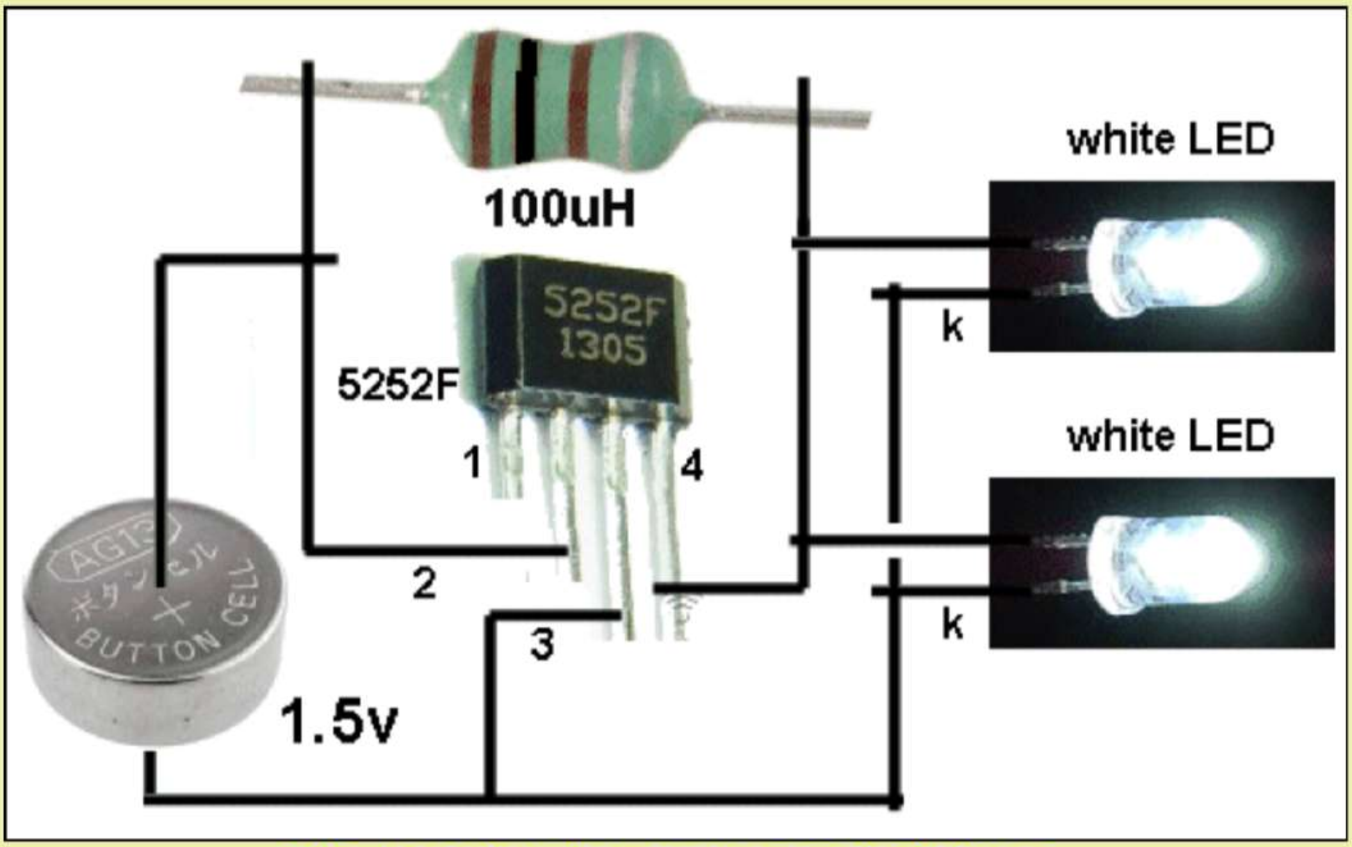

The layout using component pictures.

Using a 220uH, the circuit takes 13mA and illuminates 2 white LEDs very brightly.
Using 100uH the circuit takes 30mA and the LEDs are really the same brightness.
Using 33uH the circuit takes 80mA and the LEDs are just about the same brightness.
Obviously the 220uH creates the most efficient circuit.
The QX5252F is capable of delivering more than 100mA to the LEDs but we only need

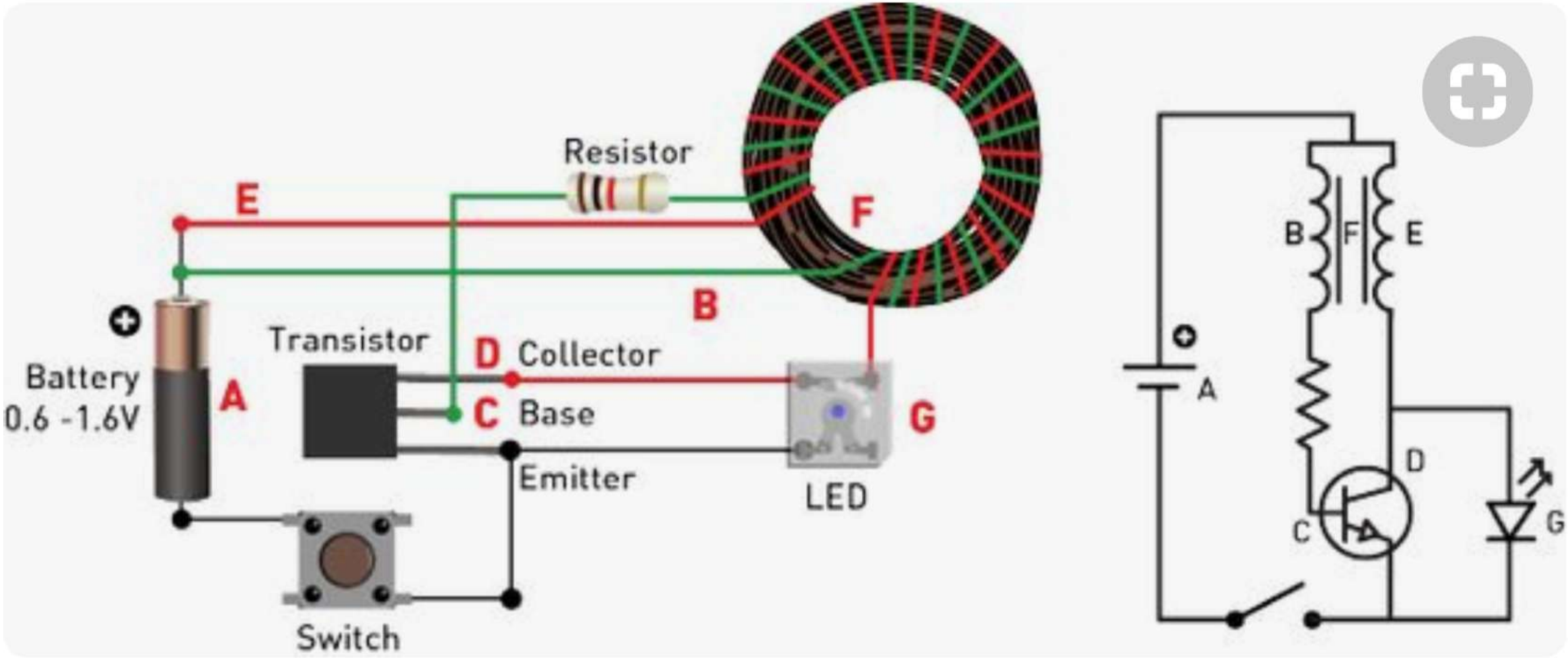
Resistor
E
F
B
Transistor
D Collector
C Base
Emitter
Battery
0.6 -1.6V
A
LED
G
Switch
B
F
E
A
C
D
G

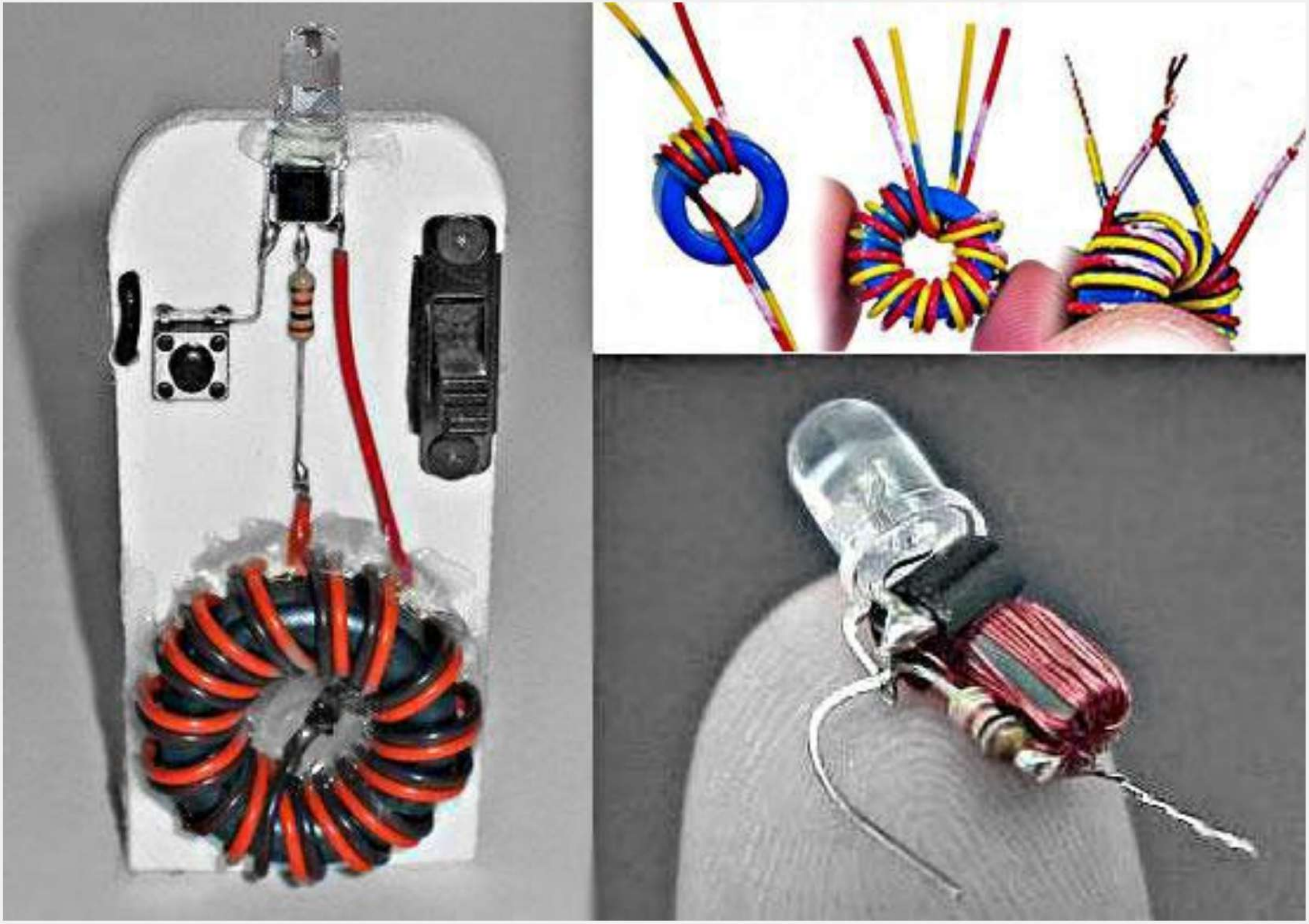

Watson's 2 Transistor "Joule Thief" V. Booster

Uses 2N3904 for output. Coil is 180 microhenry,
2.5 ohm RF choke (green blob)

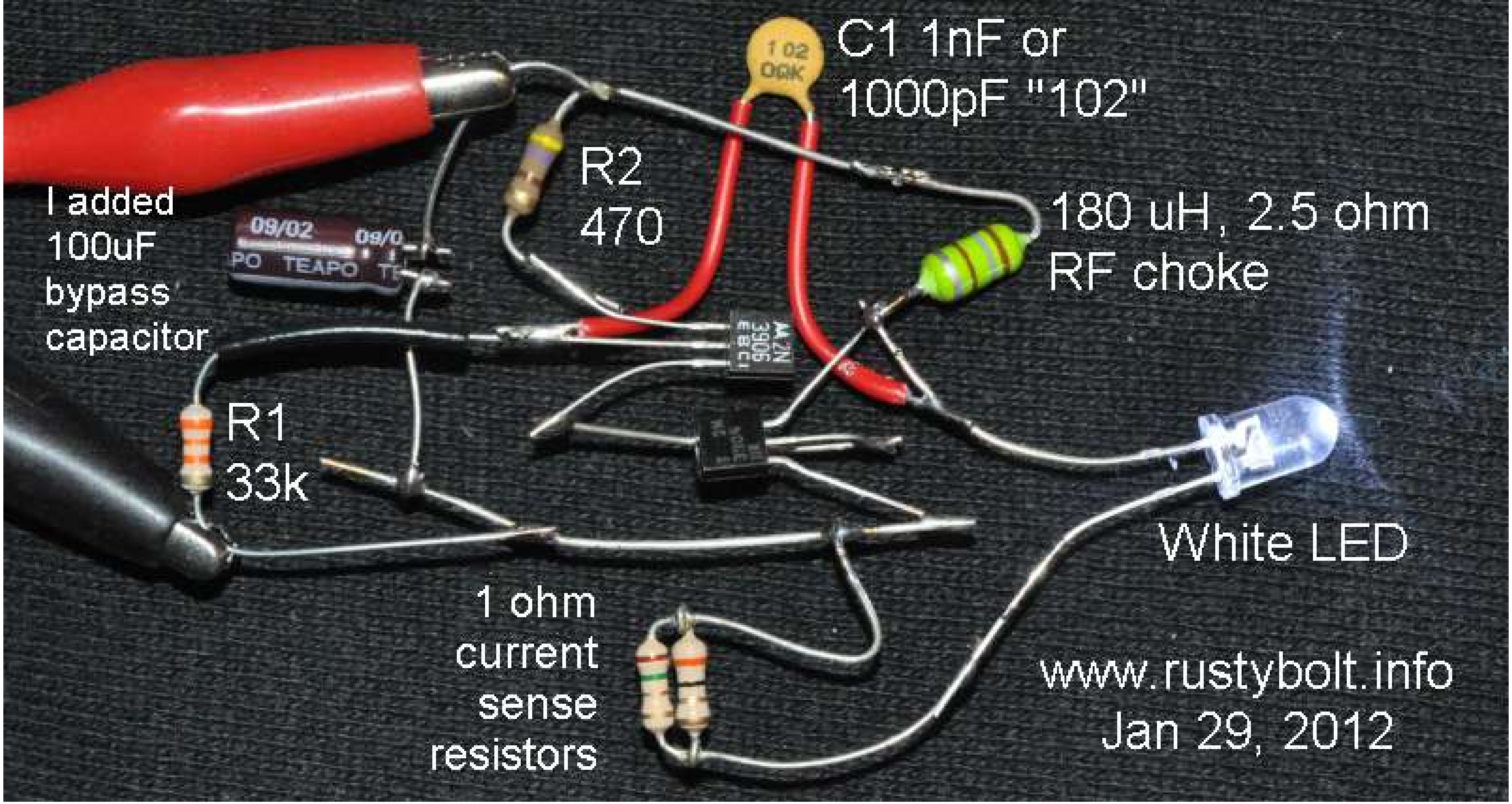
C1 1nF or
1000pF "102"
1 02
OGK
R2
470
I added
100uF
bypass
capacitor
09/02
TEAPO
2N
3906
E B C
180 uH, 2.5 ohm
RF choke
R1
33k
White LED
1 ohm
current
sense
resistors

www.rustybolt.info
Jan 29, 2012

Watson's Version of "1.5v Joule Thief -Blinks Led"

Taken from http://www.youtube.com/watch?v=GVP2QGpk5KE Nov 14, 2012

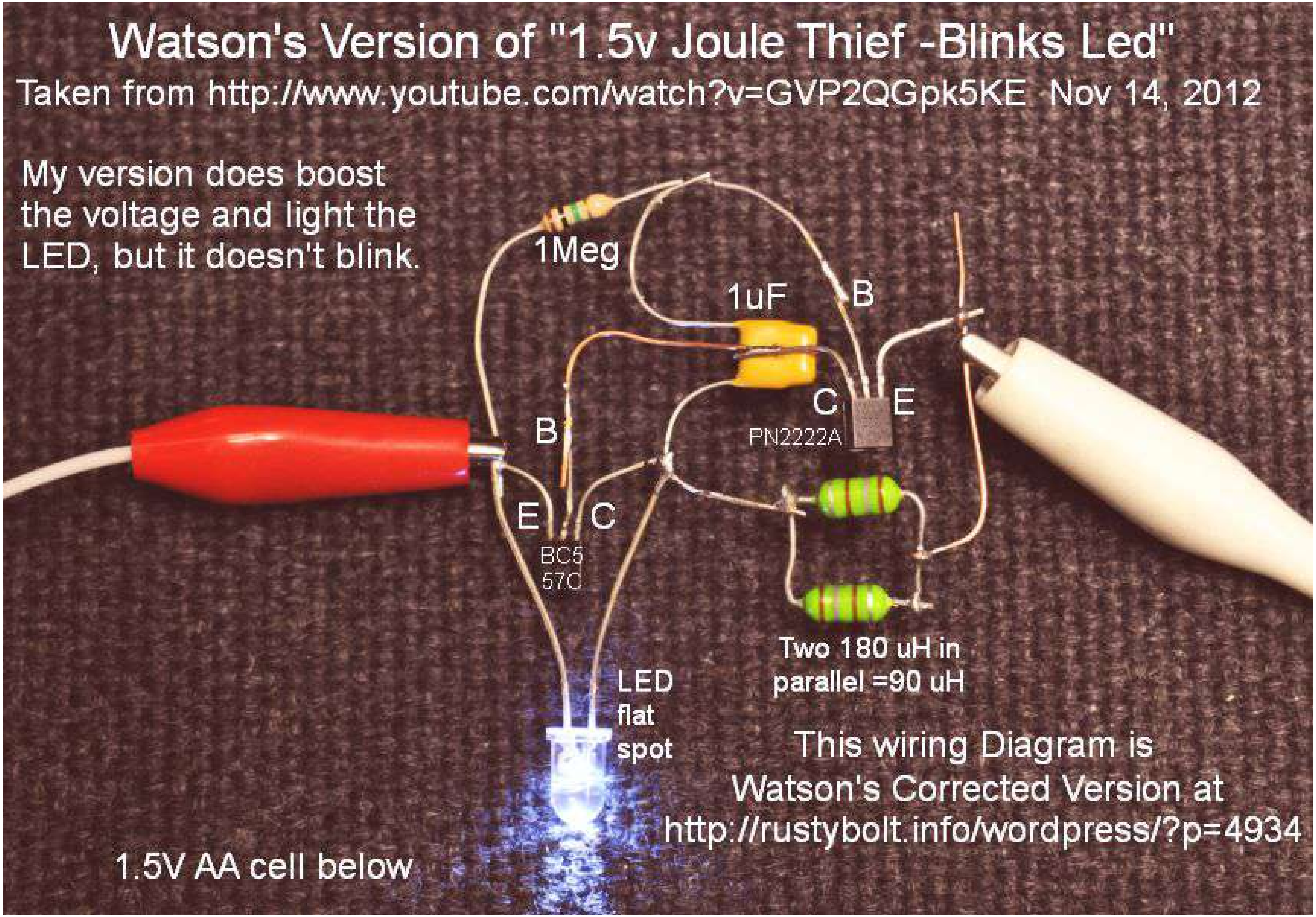

My version does boost the voltage and light the LED, but it doesn't blink.

1Meg
1uF
B
C
E
PN2222A
B
E
C
BC5
57C
LED
flat
spot
Two 180 uH in
parallel =90 uH

This wiring Diagram is Watson's Corrected Version at http://rustybolt.info/wordpress/?p=4934

1.5V AA cell below

Watson's Joule Thief Pictorial Diagram

http://rustybolt.info/wordpress/?p=128 2011 Dec 07

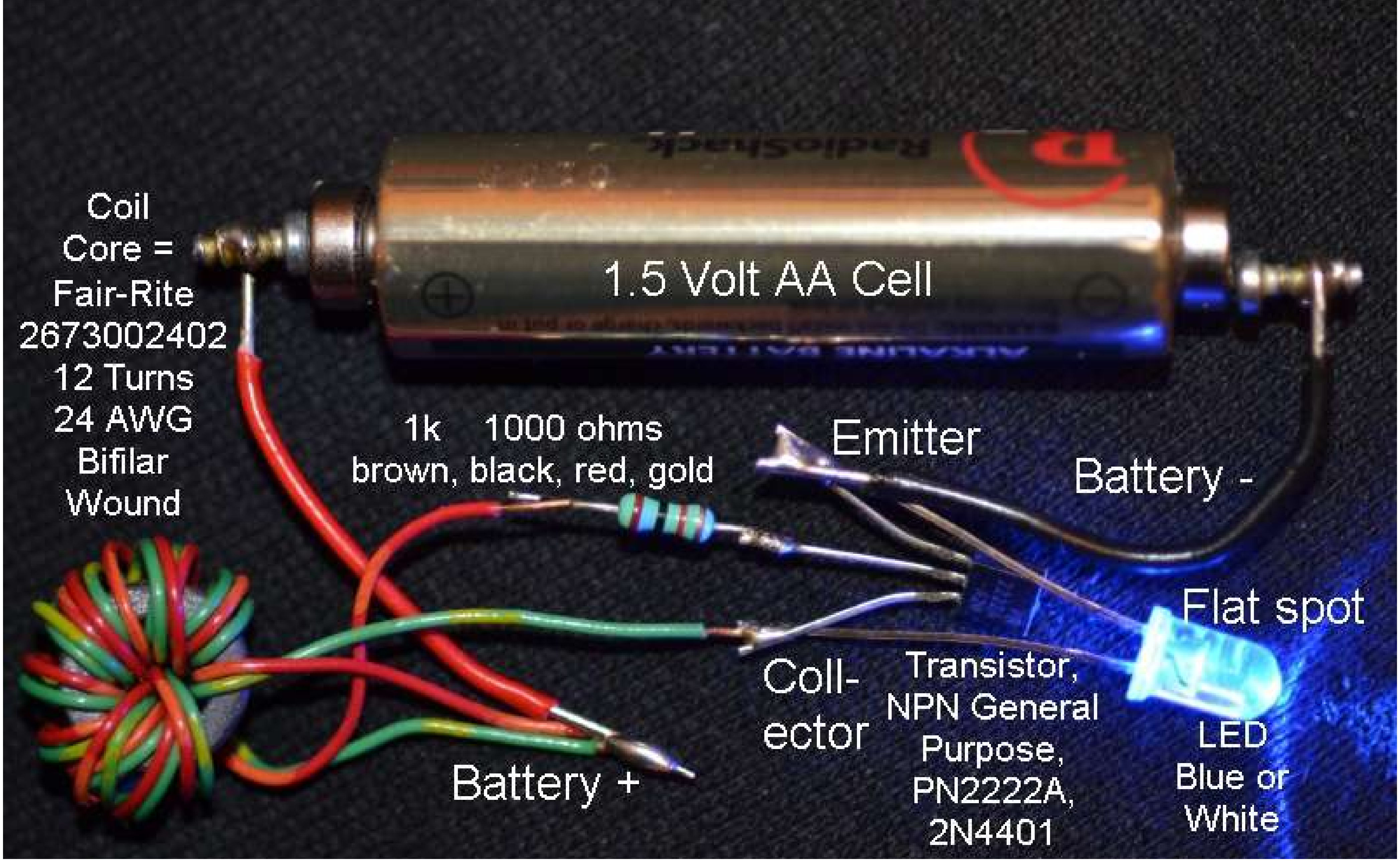
Coil
Core =
Fair-Rite
2673002402
12 Turns
24 AWG
Bifilar
Wound
1.5 Volt AA Cell
1k 1000 ohms
brown, black, red, gold
Emitter
Battery -
Flat spot
Coll-
ector
Transistor,
NPN General
Purpose,
PN2222A,
2N4401
LED
Blue or
White
Battery +

Joule Thief

Pictorial Wiring Diagram Dec 11, 2009

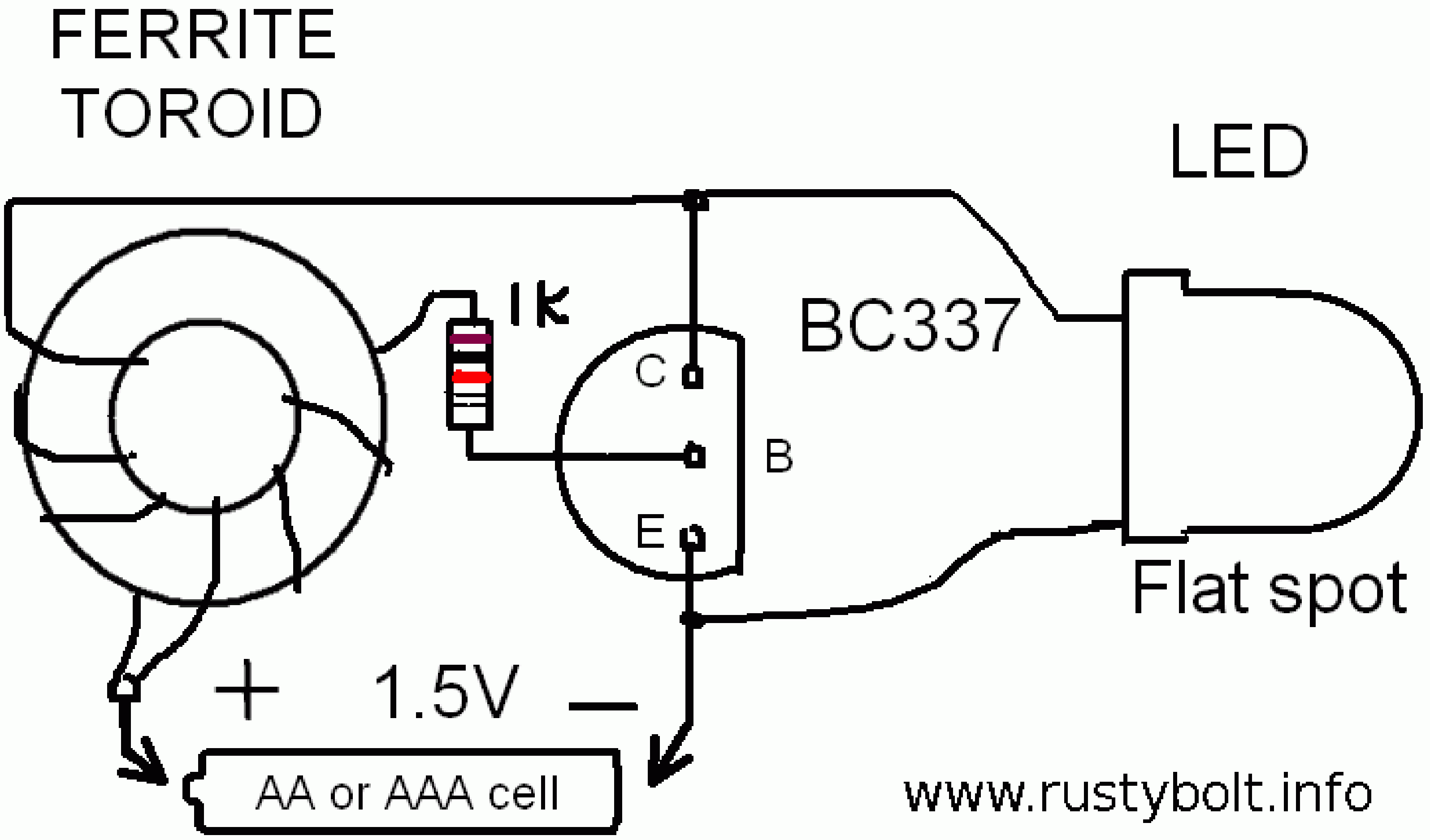
FERRITE
TOROID
LED
1k
BC337
C
B
E
Flat spot
+
1.5V
—
AA or AAA cell

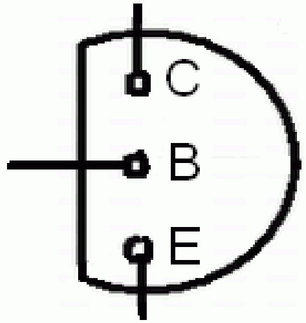

for 2N4401

www.rustybolt.info

Watson's True 2N2222 Joule Thief

Oct 31, 2010 Boo!

Not very good performance from an expensive transistor

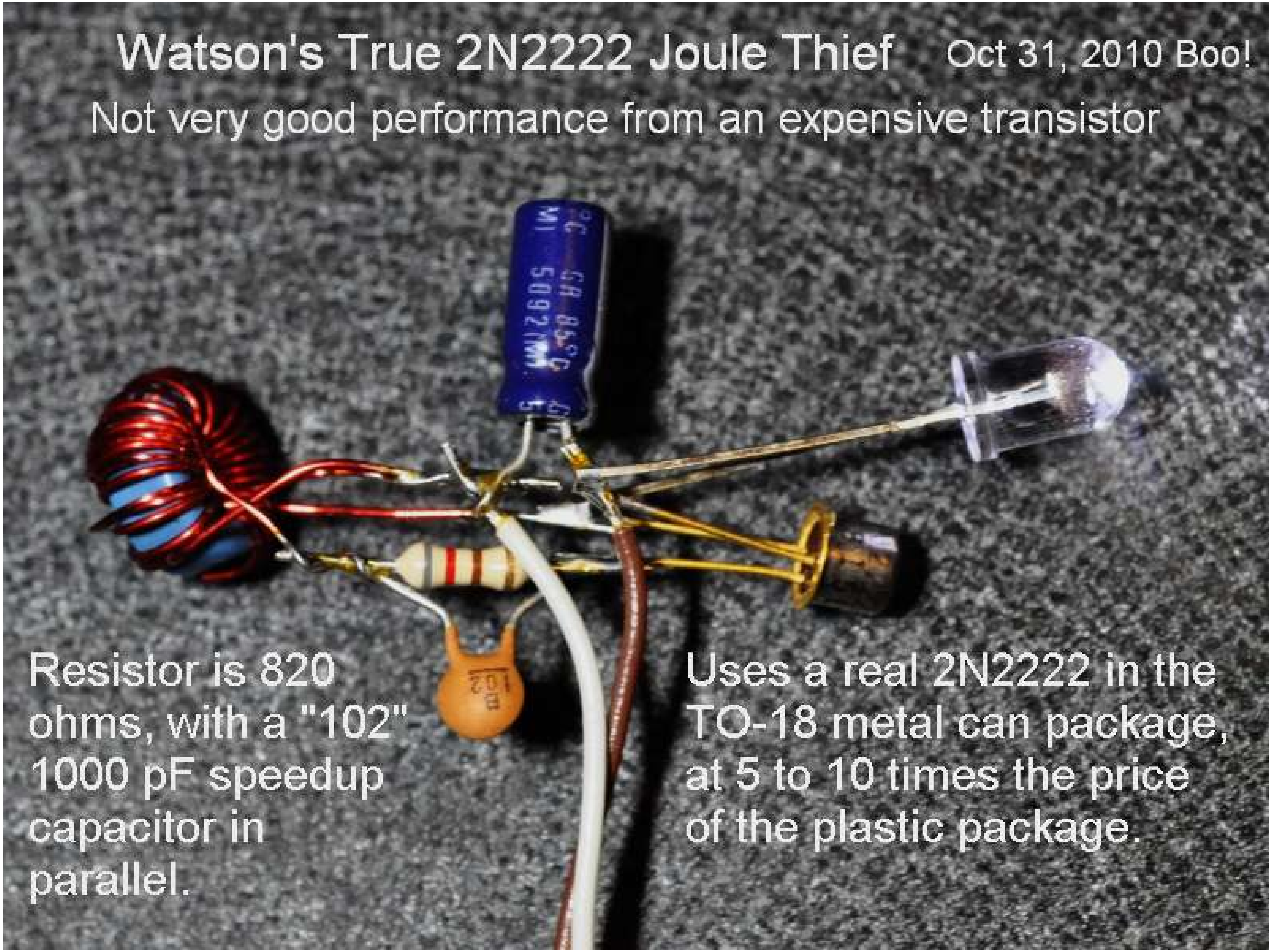

Resistor is 820 ohms, with a "102" 1000 pF speedup capacitor in parallel.

Uses a real 2N2222 in the TO-18 metal can package, at 5 to 10 times the price of the plastic package.

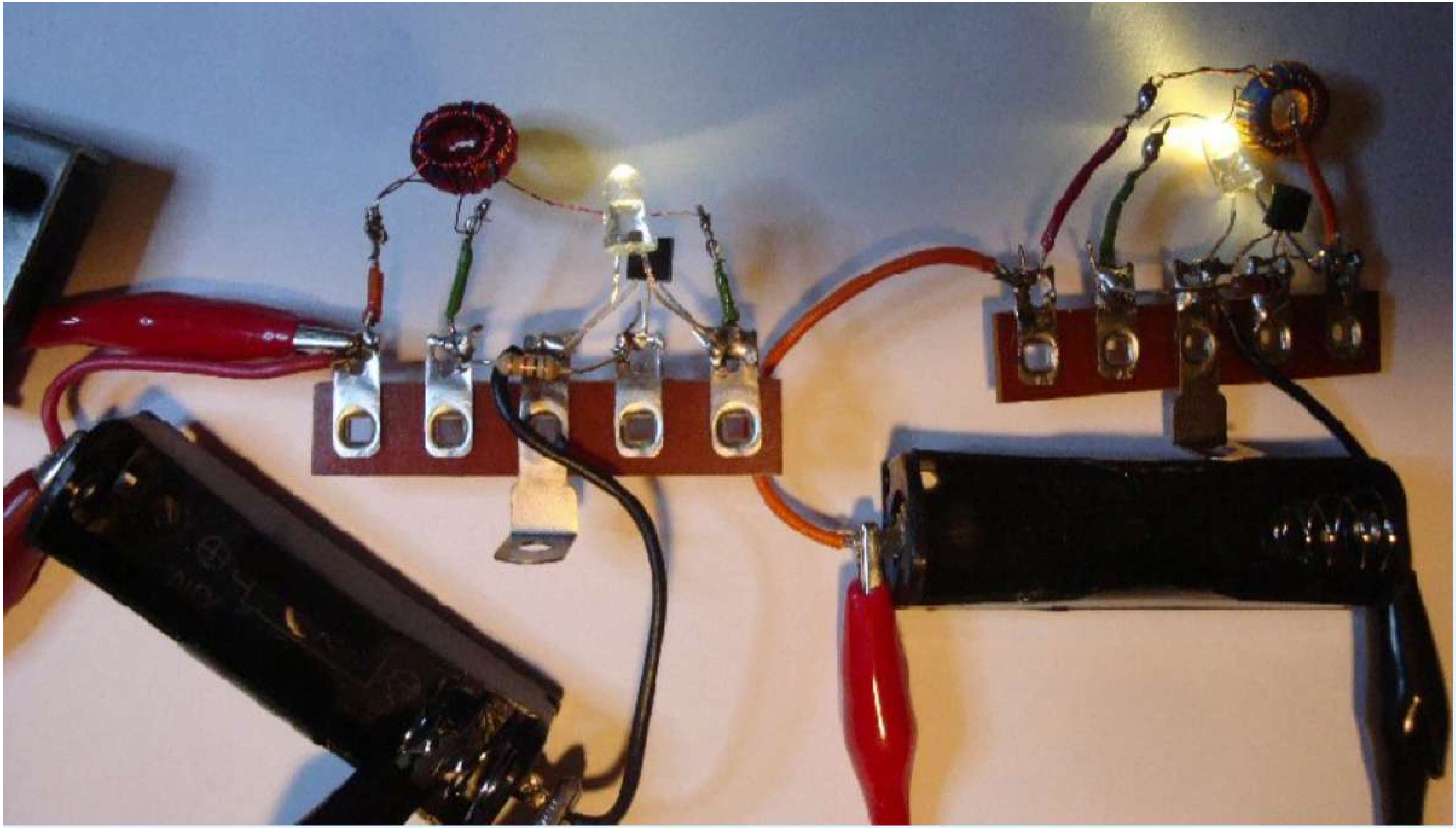

Watson's Tiny Toroid Conventional Joule Thief - Nov 18, 2009

Not much bigger than the transistor!

www.rustybolt.info

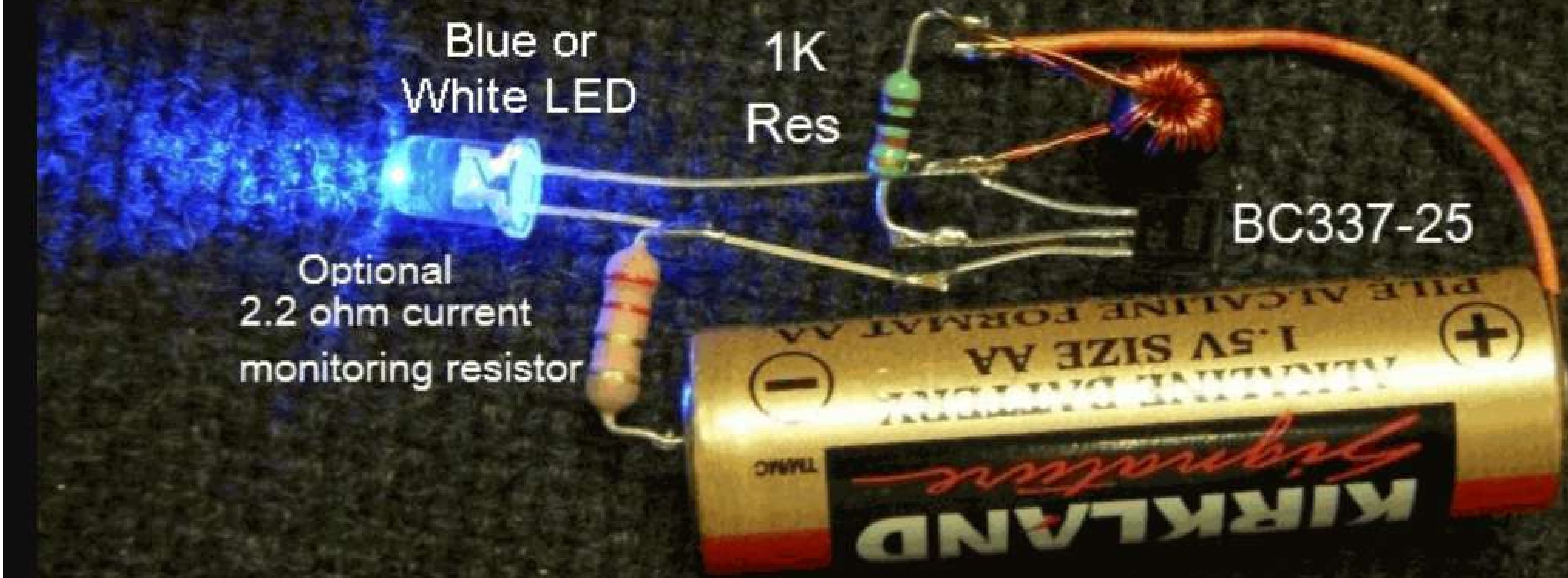
Blue or White LED
1K Res
BC337-25
Optional 2.2 ohm current monitoring resistor

Three windings 30 AWG, each about 6 inches long, trifilar wound on 0.229 inch O.D. toroid, 2 in parallel for main winding.

I took this photo just before Halloween back in 2010, after I made this Joule Thief without solder and without a toroid. It shows that you don't have to know how to solder and you don't have to have a toroid to make a working Joule Thief – all you need is a screwdriver. So do your thing and experiment a bit, and see what you can come up with.

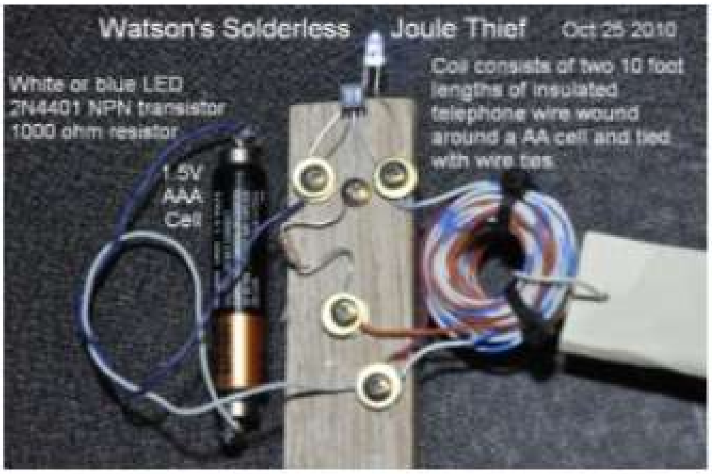
Watson's Solderless Joule Thief
Oct 25 2010
White or blue LED
2N4401 NPN transistor
1000 ohm resistor
1.5V
AAA
Cell
Coil consists of two 10 foot lengths of insulated telephone wire wound around a AA cell and tied with wire ties.

Each screw (except for one) has two washers to hold the wires. Putting two or more wires under the screw head without the washers will be difficult and the wires will try to come loose. Two or three washers makes it easier and the wires are held much firmer. The wood block is a piece of solid oak that is about a half inch (12mm) thick. The oak is very hard and the screw holes have to be drilled out before the screws are put in. This also makes it possible to use small screws with a flat end which were used for holding plastic parts together. The screw size is about #3 by 3/8 inch long, with coarse threads like a sheet metal screw.

The LED is not very bright because it's a cheap 3 mm white LED I got from an eBay seller, and it has been used a bit so it has dimmed. Use a decent LED and this circuit will be as bright as a toroid JT.

Another Joule Thief I put together is shown in the second photo. This uses the bare wires to wrap around the joints in place of solder. The joints are not as mechanically strong as solder but for a quick experiment they should work okay. If the JT is going to be used, the wire joints may become intermittent or loose. The solder joints are a much better connection.

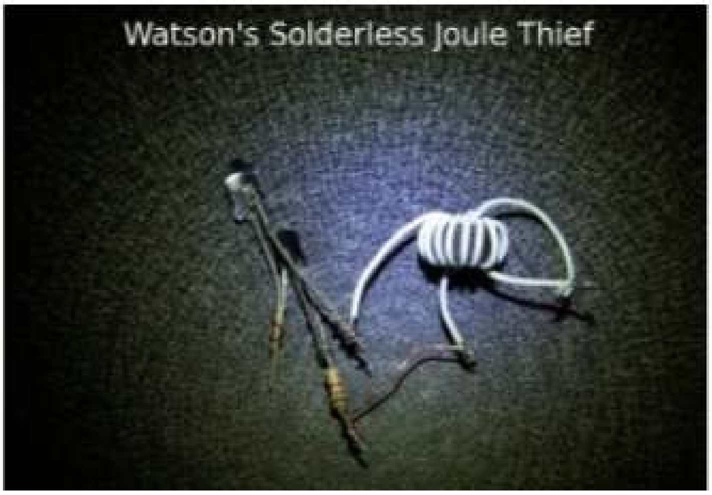
Watson's Solderless Joule Thief

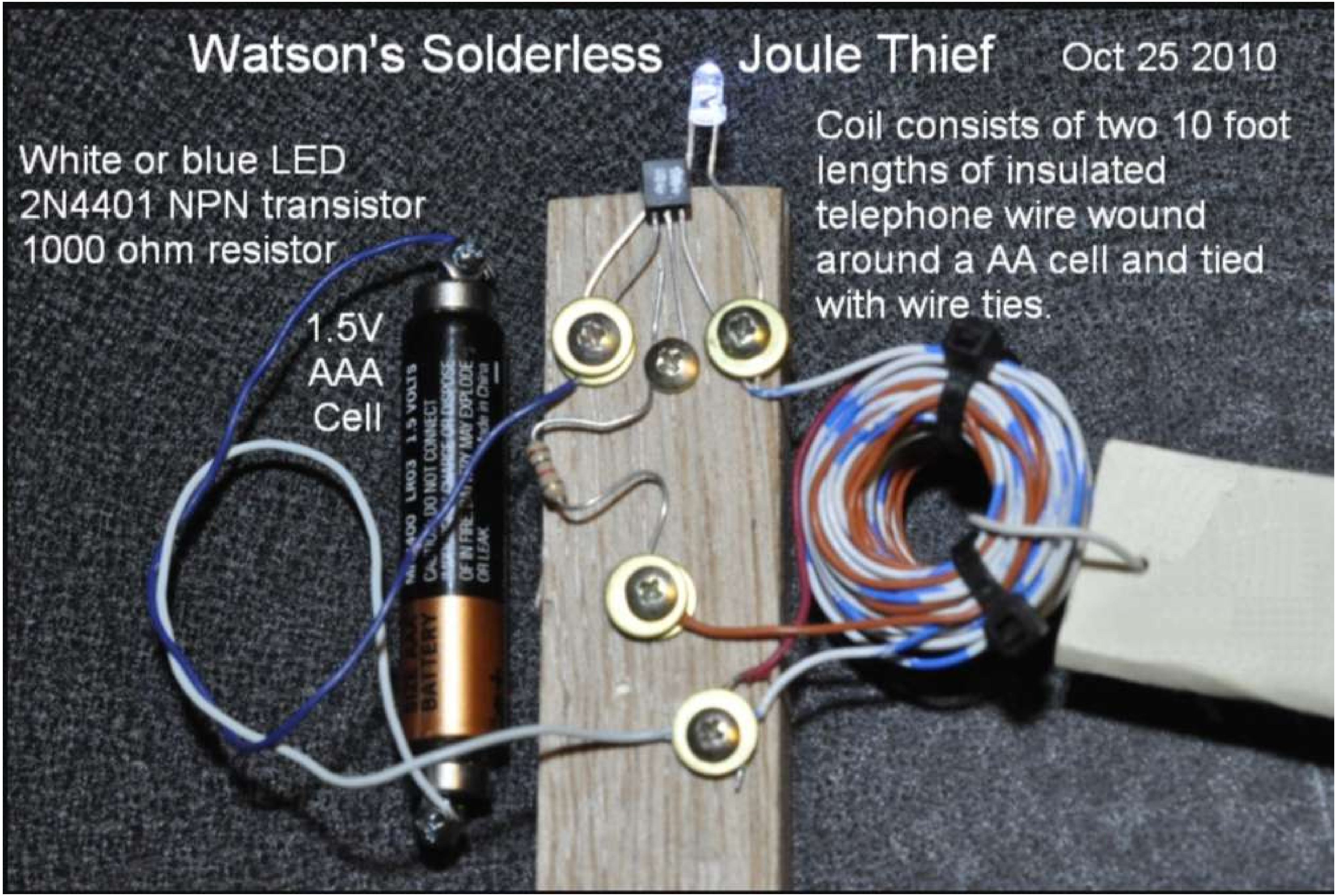

Watson's Solderless Joule Thief Oct 25 2010

White or blue LED
2N4401 NPN transistor
1000 ohm resistor

Coil consists of two 10 foot lengths of insulated telephone wire wound around a AA cell and tied with wire ties.

Watson's 'Toroid Free' Air Core Joule Thief

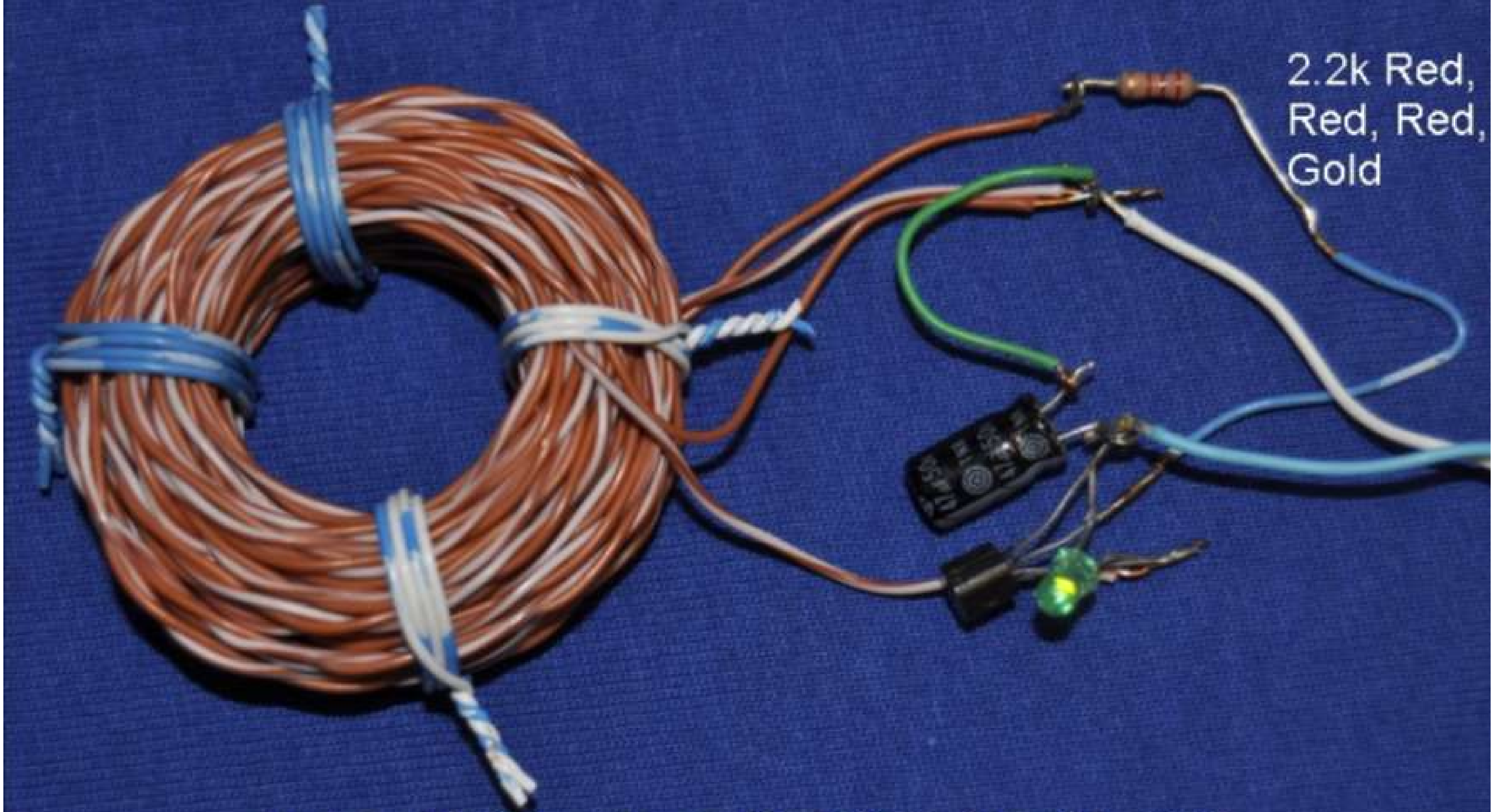

2.2k Red,
Red, Red,
Gold

rustybolt.info/wordpress/?p=2872 May 21, 2012

Watson's Nearly Disposable Joule Thief Light

rustybolt.info/wordpress/?p=2133   Fri. Apr 13 2012

Completed circuit ready for coating of silicone sealant.
Transistor is a BC337. If it was a PN2222A or 2N4401,
the flat side would be facing up.

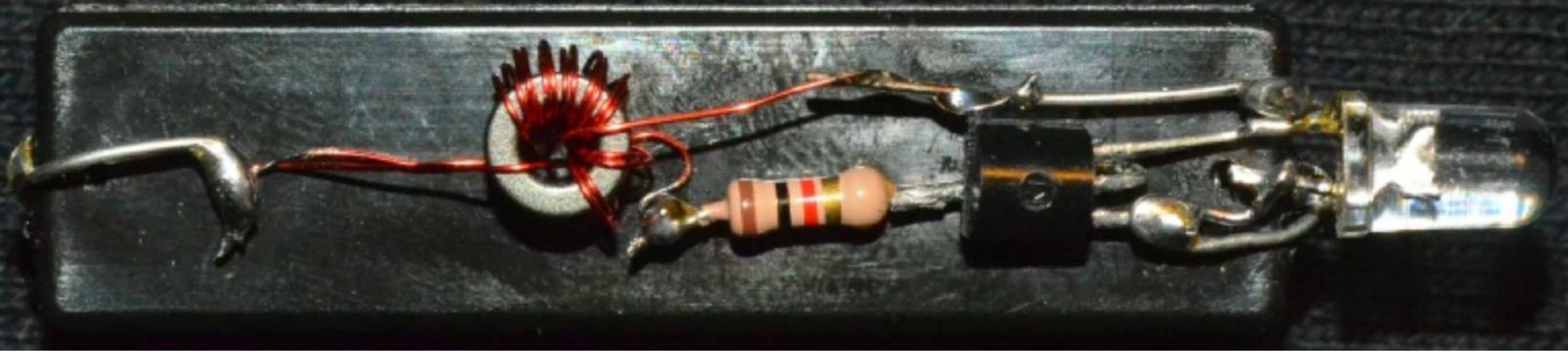

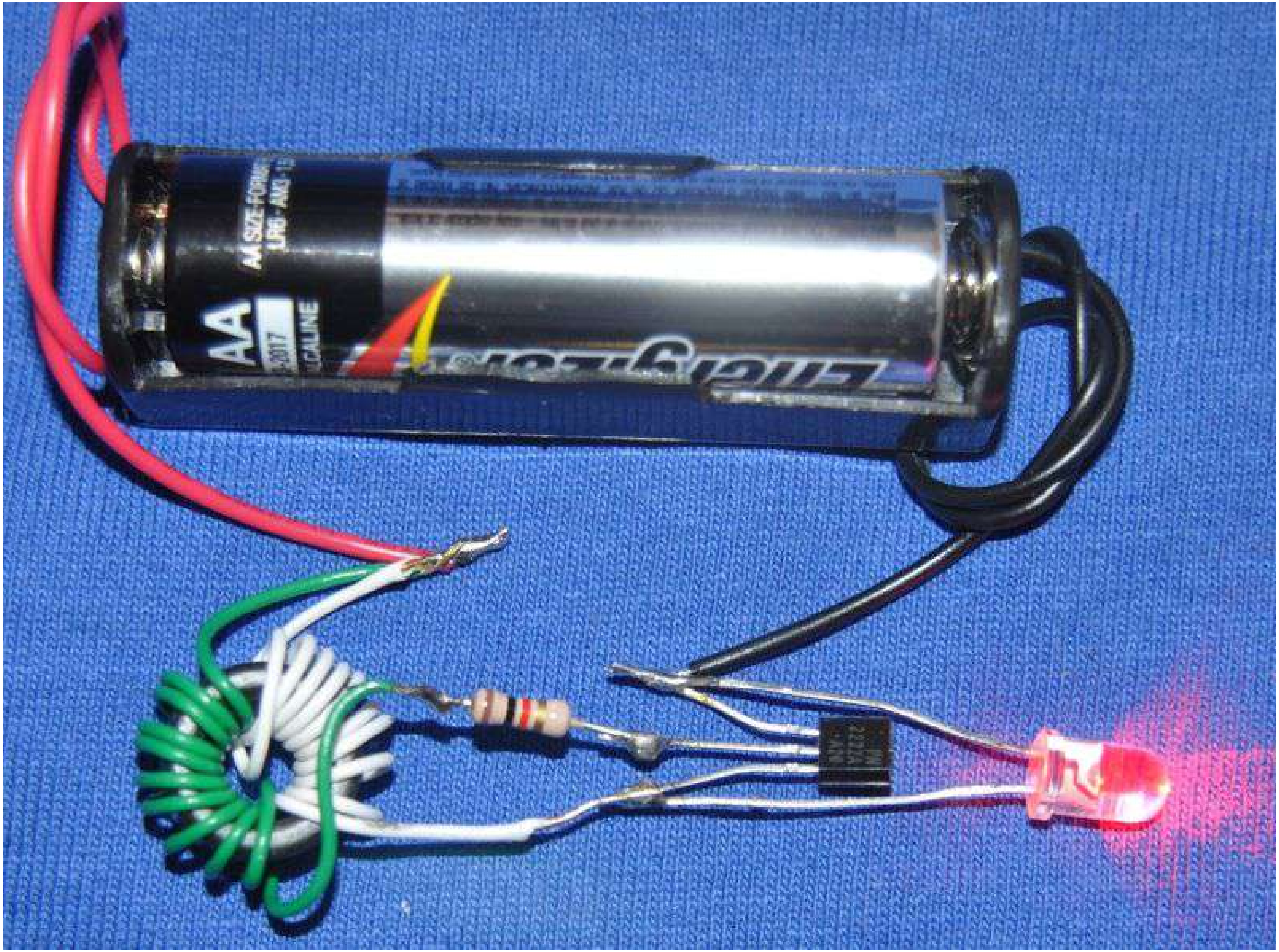
AA
AA SIZE-FORMAT
LR6 - AM3
ALKALINE
Energizer

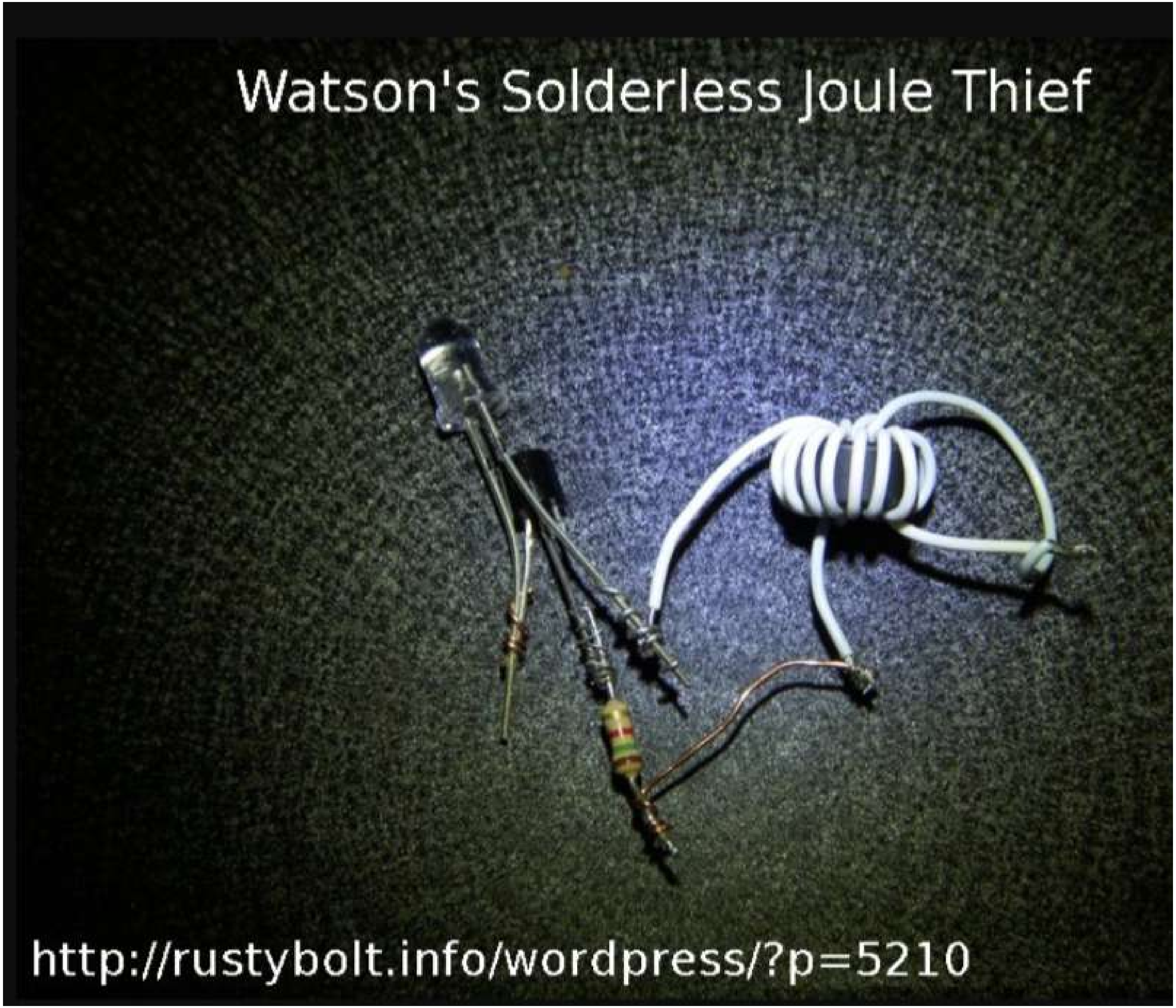

Watson's Solderless Joule Thief

http://rustybolt.info/wordpress/?p=5210

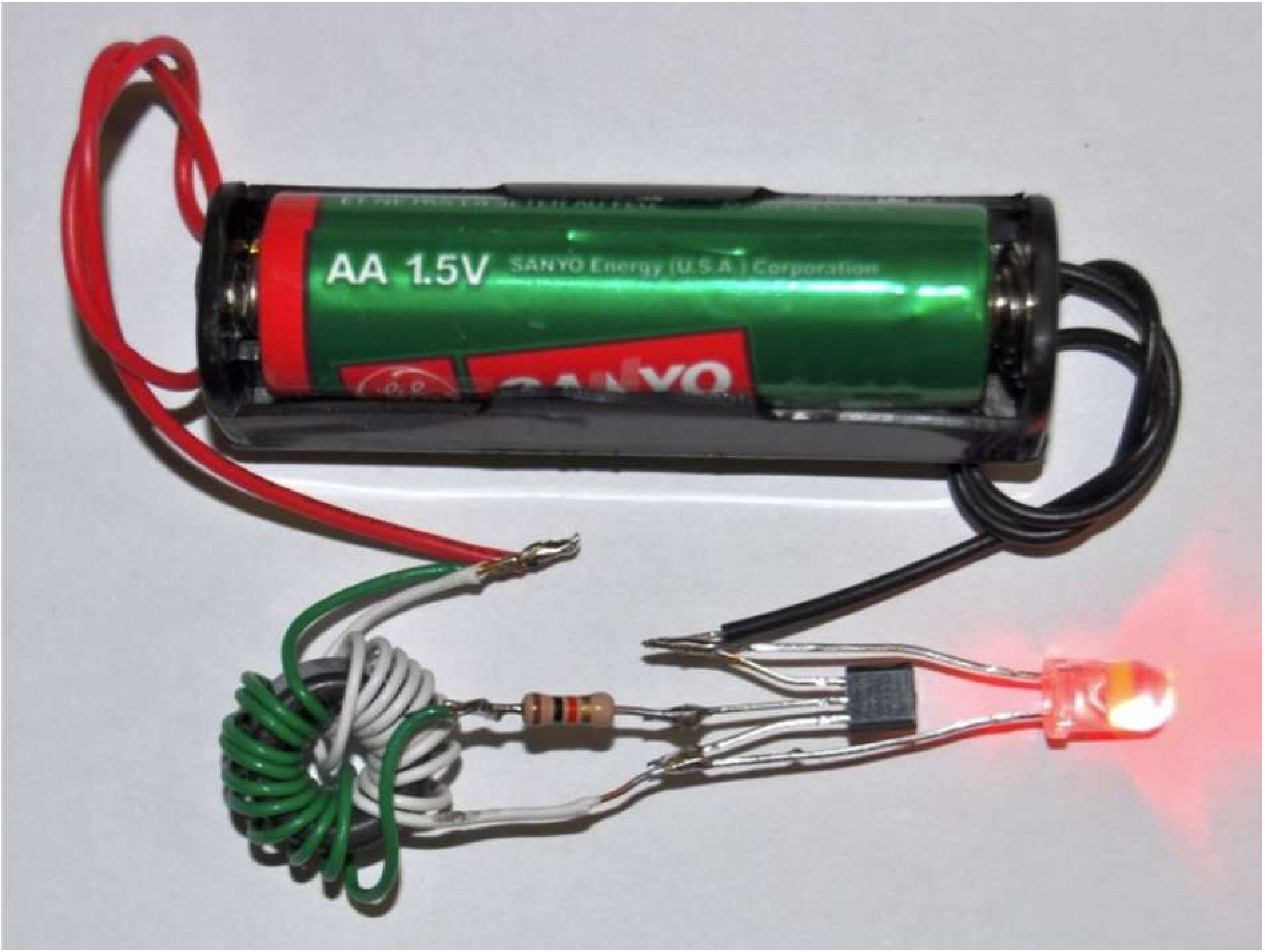
AA 1.5V
SANYO Energy (U.S.A.) Corporation
SANYO

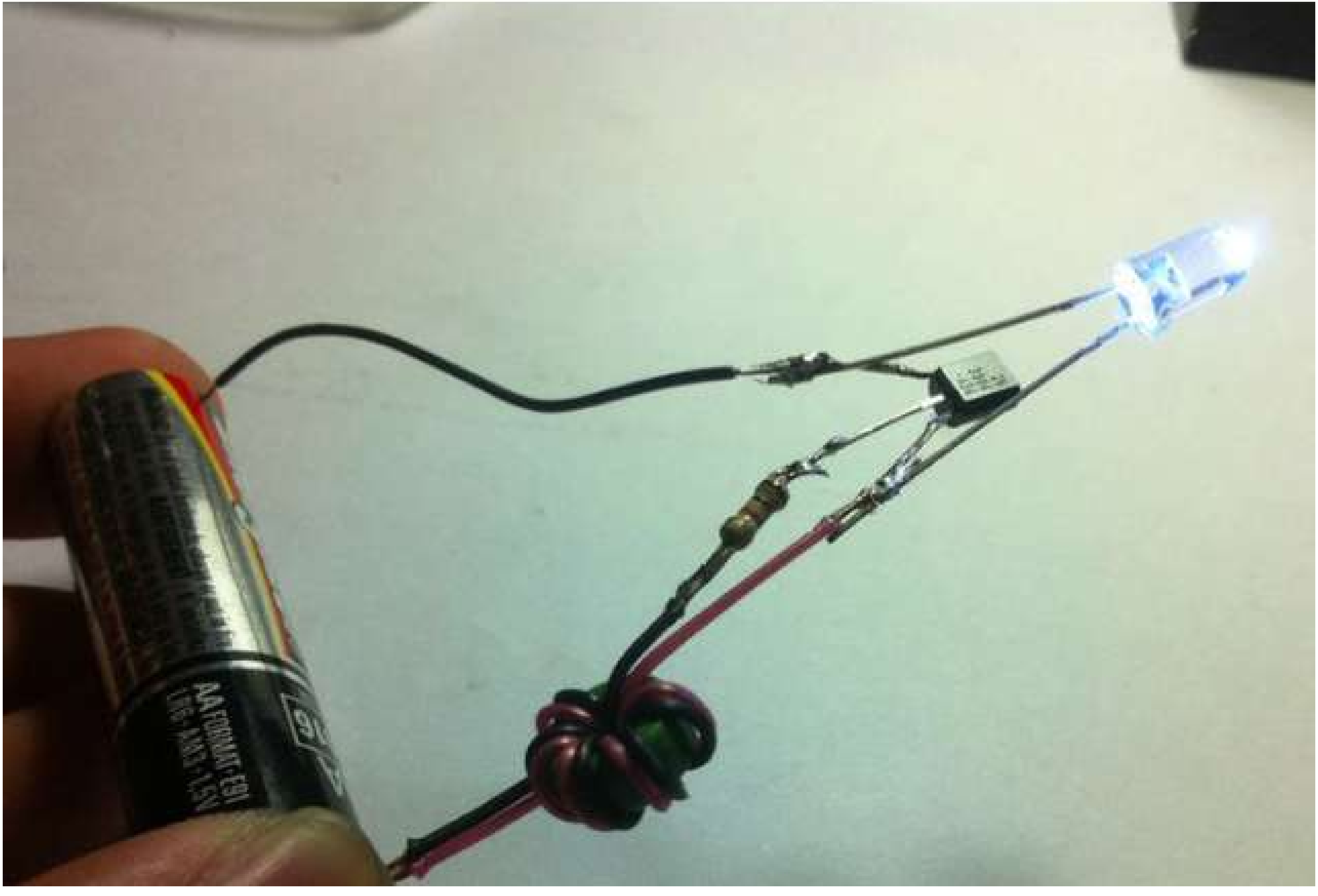
AA FORMAT · E91
LR6 · AM3 · 1.5V

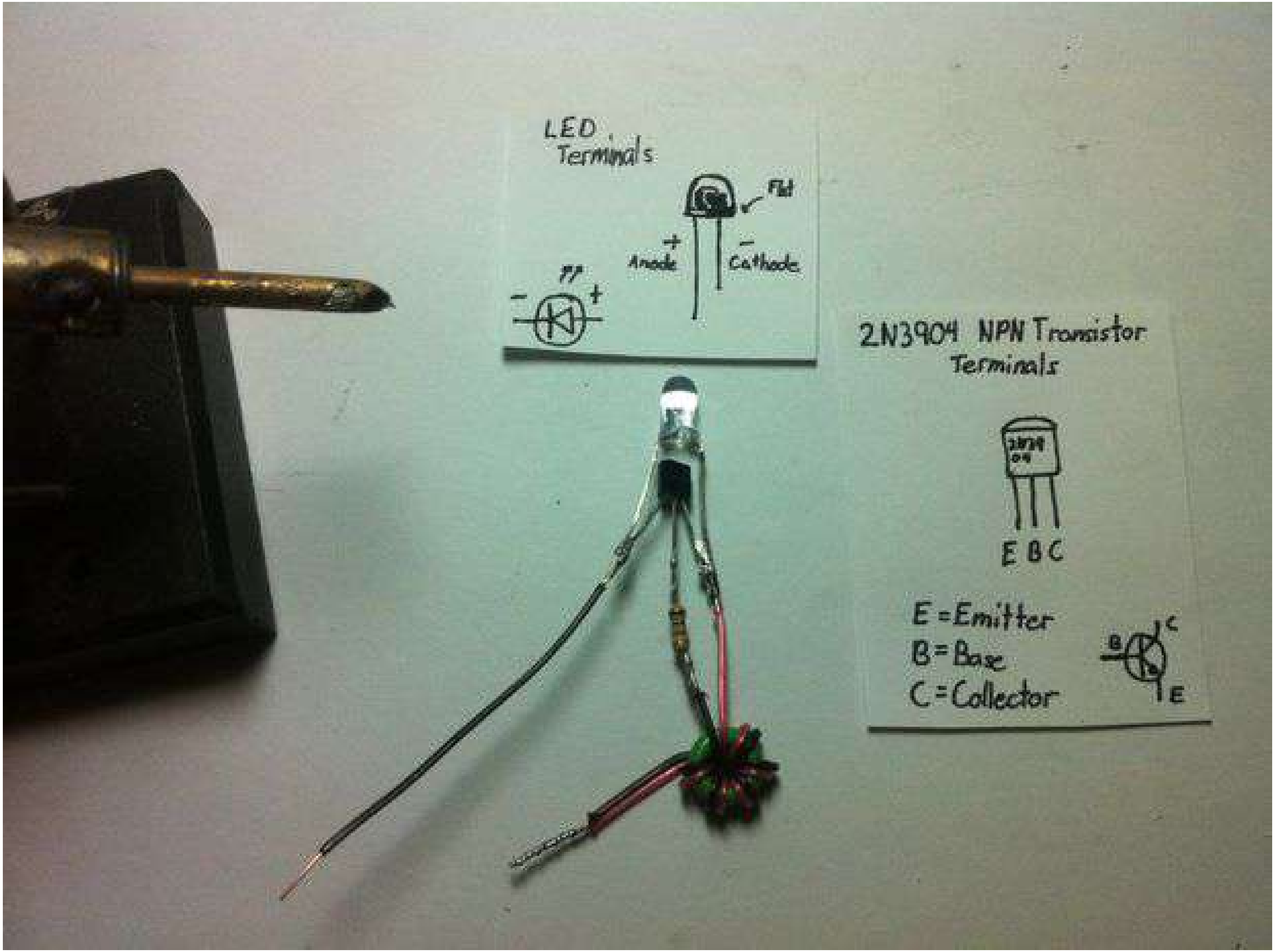
LED
Terminals
Flat
+
Anode
-
Cathode
-
+
2N3904 NPN Transistor
Terminals
2N39
04
E B C
E = Emitter
B = Base
C = Collector
B
C
E

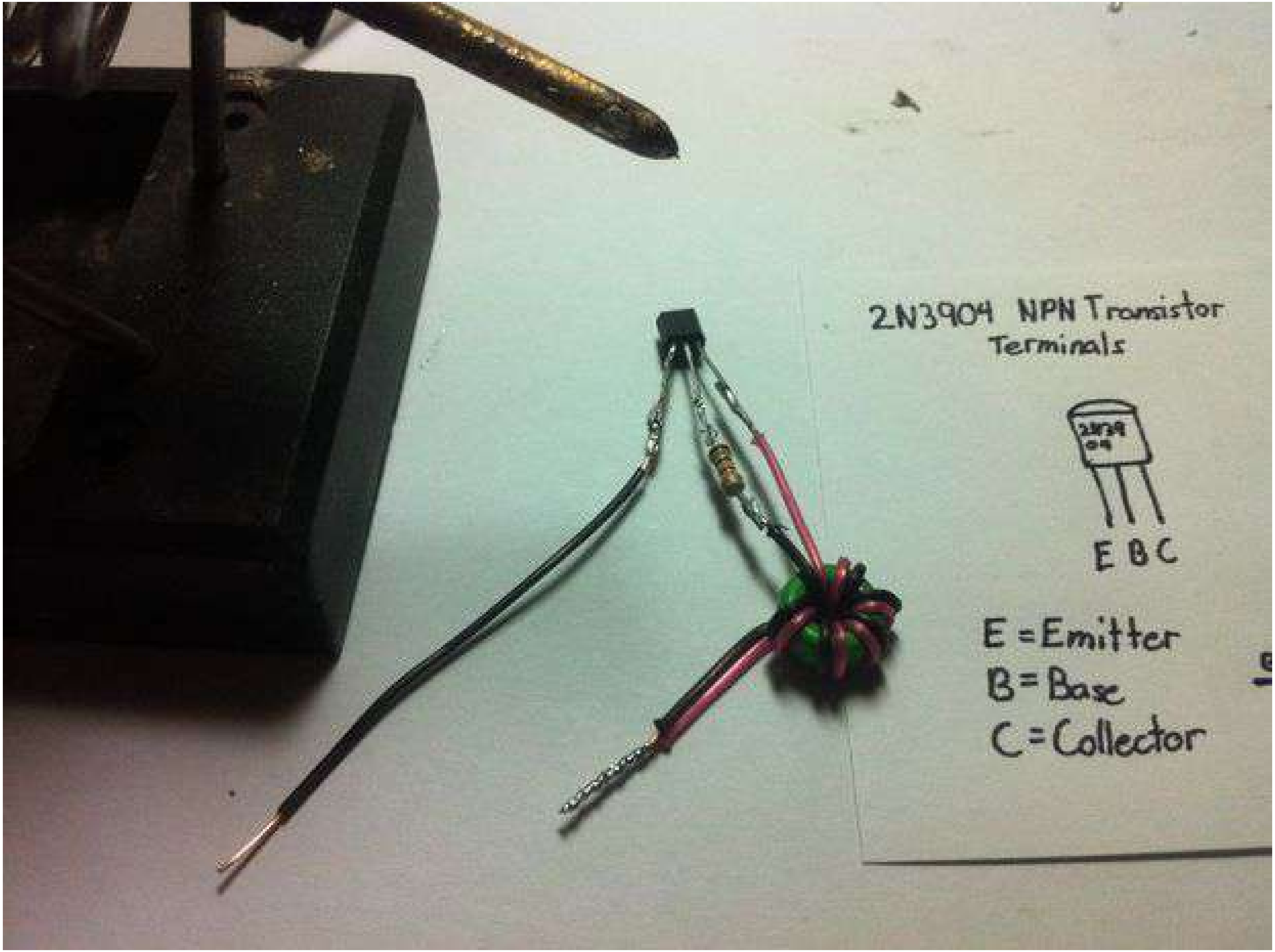
2N3904 NPN Transistor Terminals
2N39 04
E B C
E = Emitter
B = Base
C = Collector

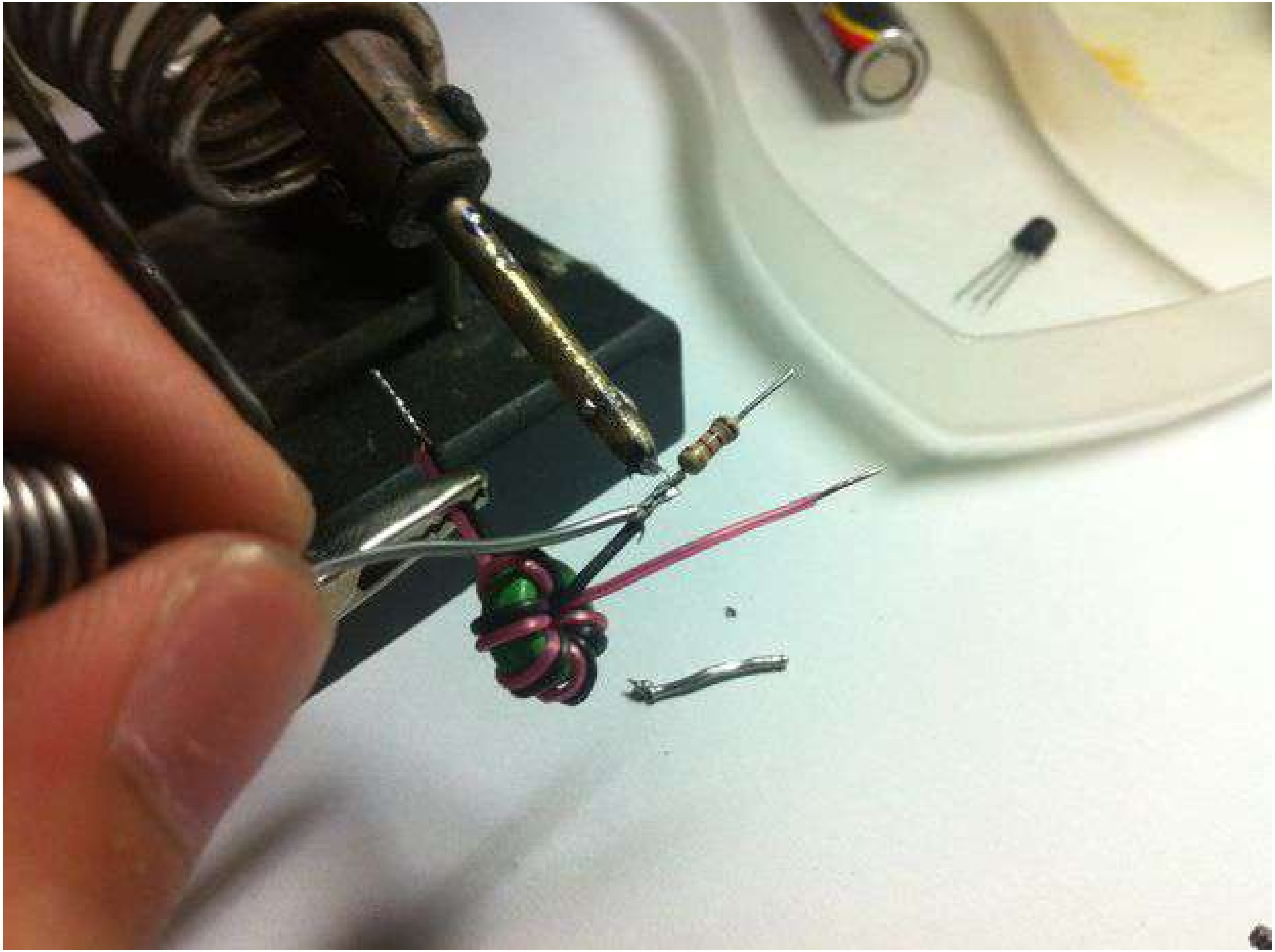

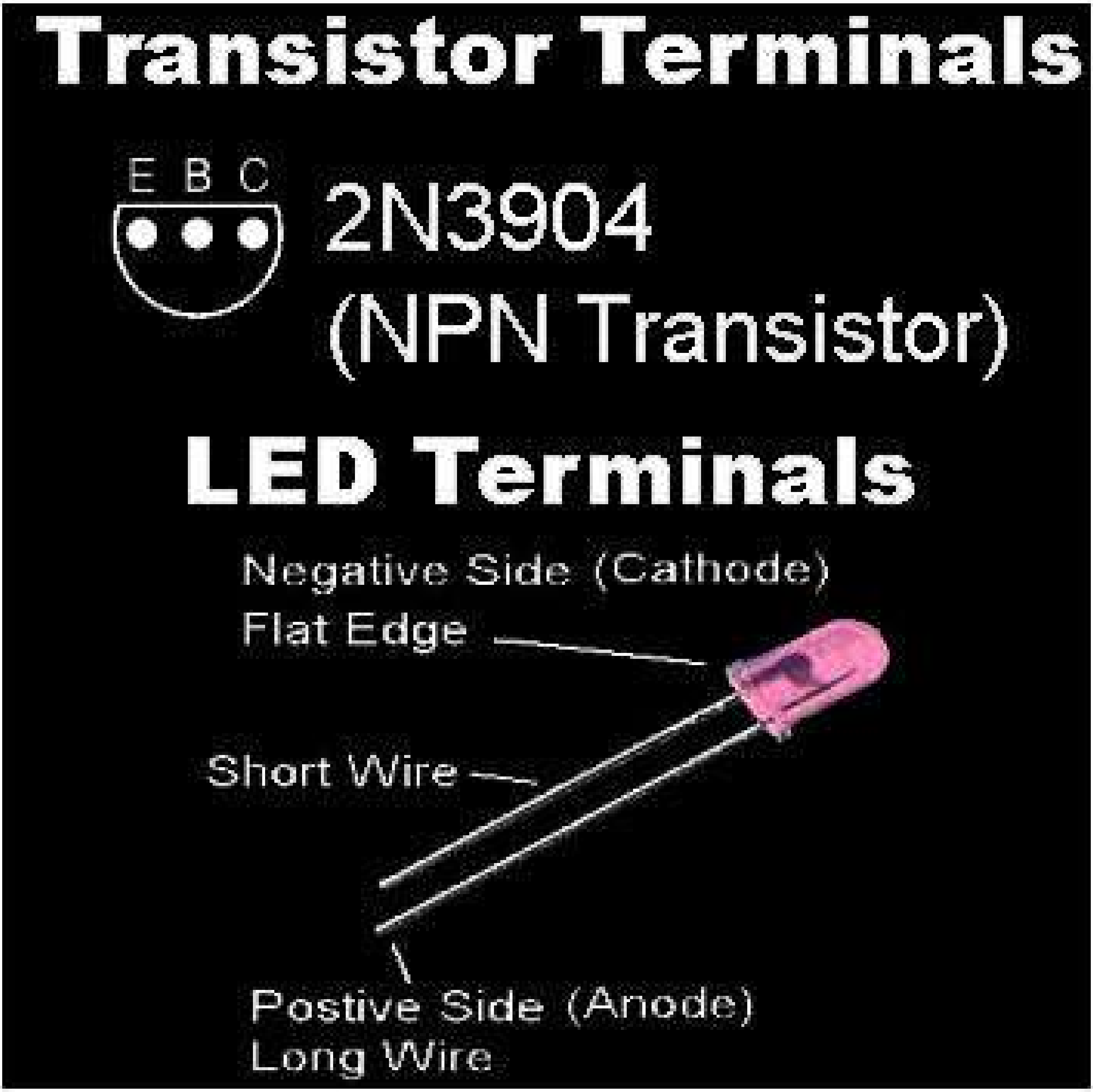

Transistor Terminals

E B C

2N3904
(NPN Transistor)

LED Terminals

Negative Side (Cathode)
Flat Edge

Short Wire

Postive Side (Anode)
Long Wire

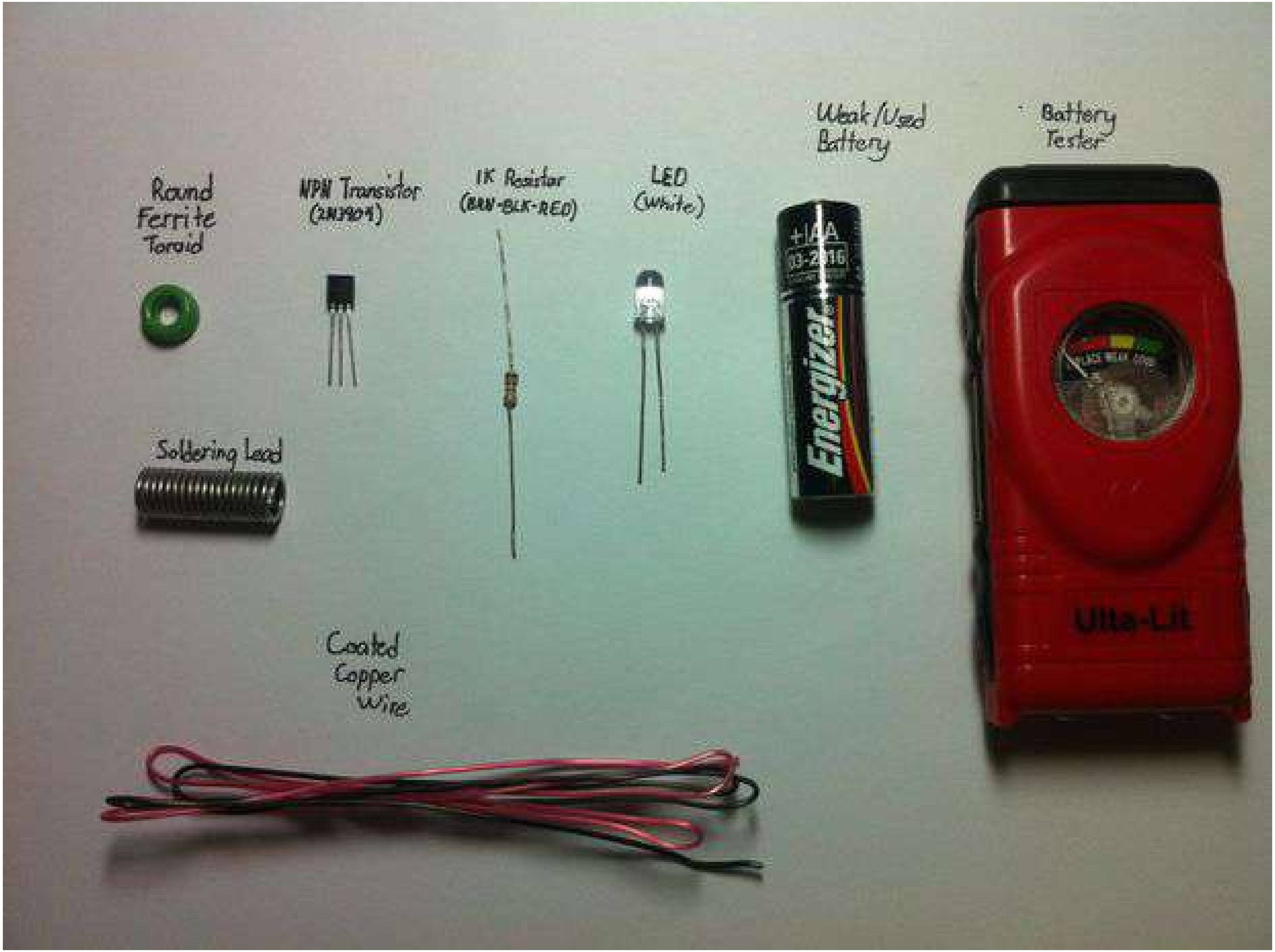
Round Ferrite Toroid
NPN Transistor (2N3904)
1K Resistor (BRN-BLK-RED)
LED (White)
Weak/Used Battery
Battery Tester
+AA
03-2016
Energizer
Soldering Lead
Ulta-Lit
Coated Copper Wire

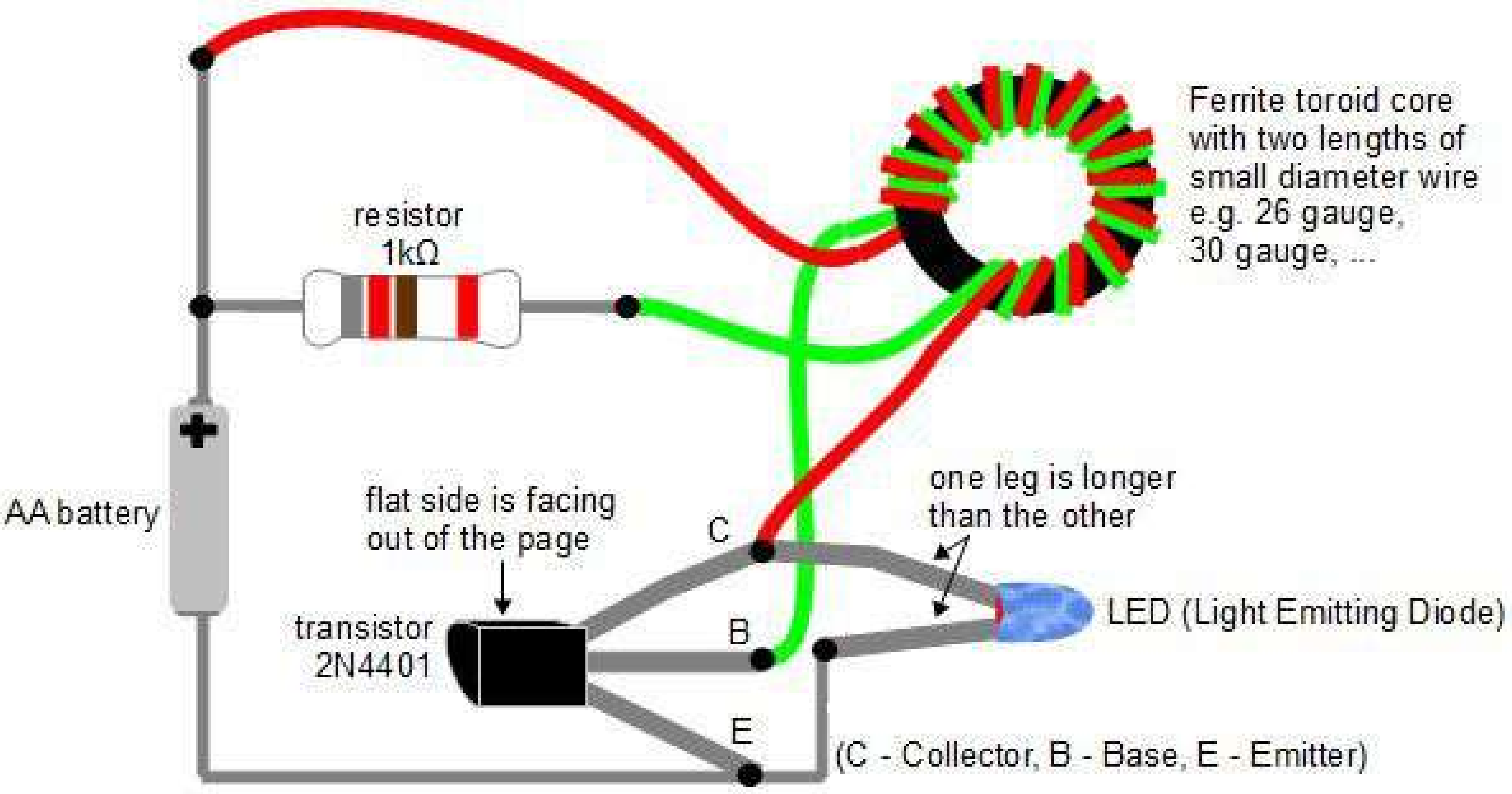
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge,
30 gauge, ...
resistor
1kΩ
AA battery
flat side is facing
out of the page
transistor
2N4401
C
B
E
one leg is longer
than the other
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

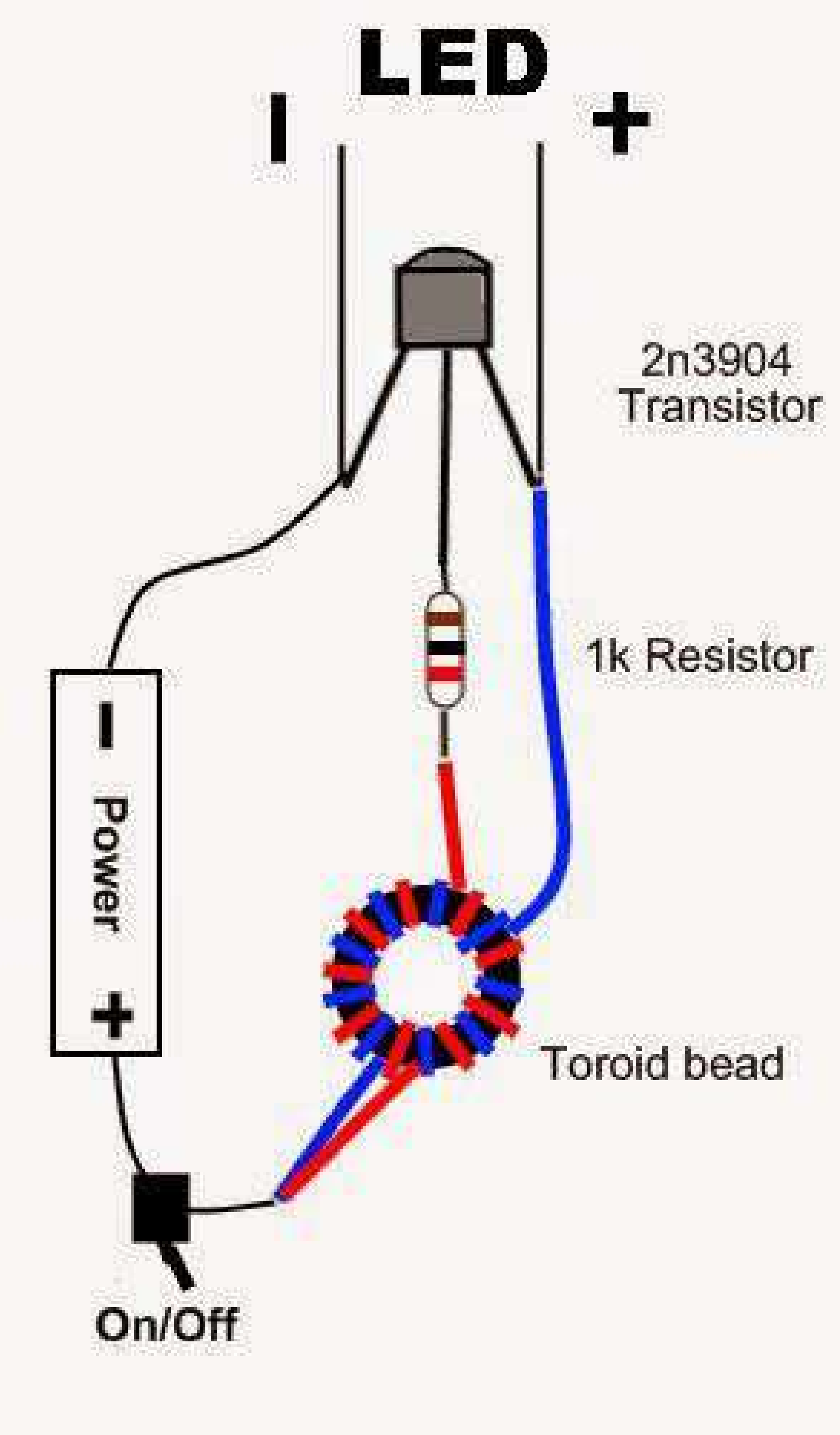
LED
I
+
2n3904
Transistor
1k Resistor
–
Power
+
Toroid bead
On/Off

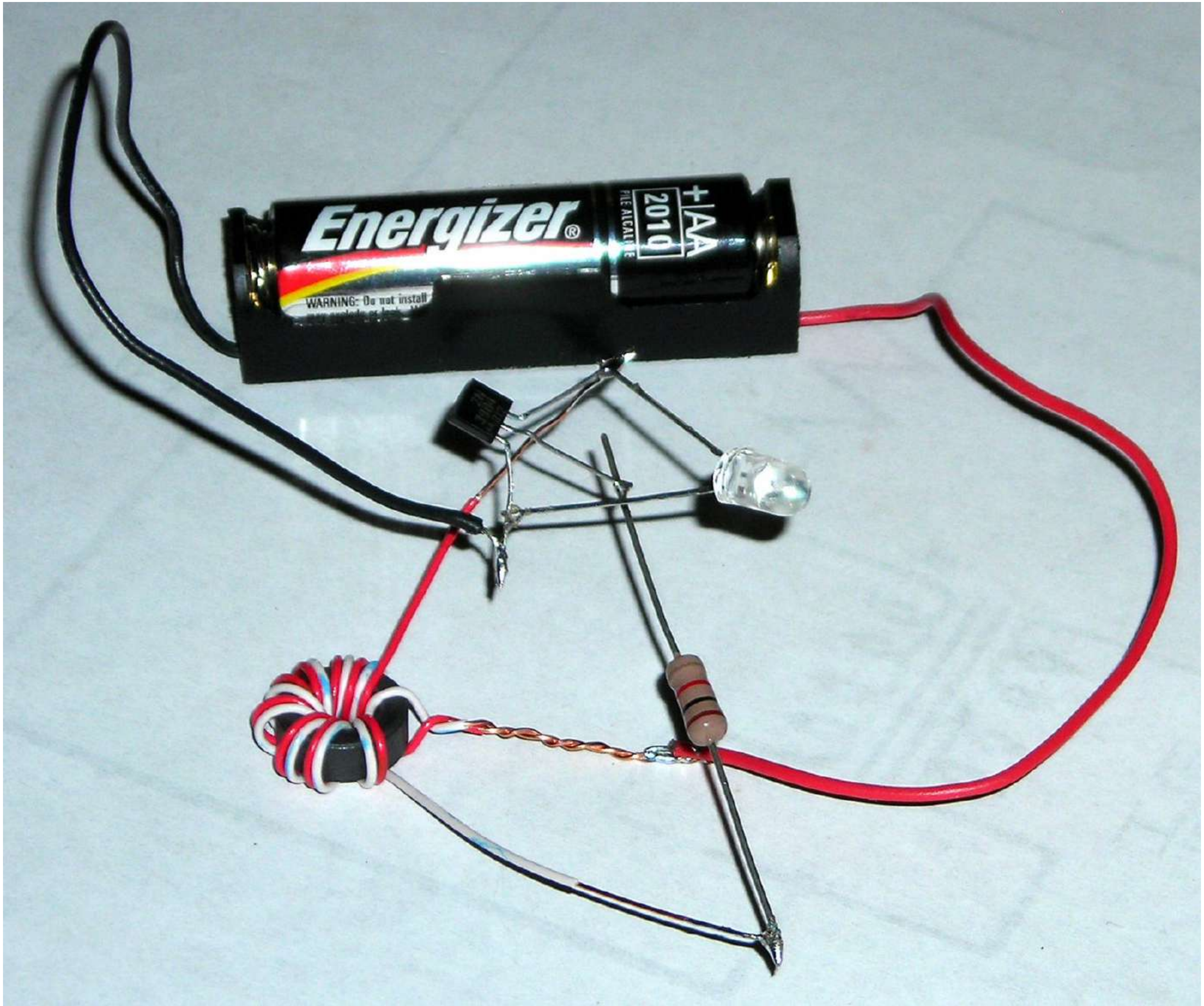
Energizer
+
AA
2010
WARNING: Do not install

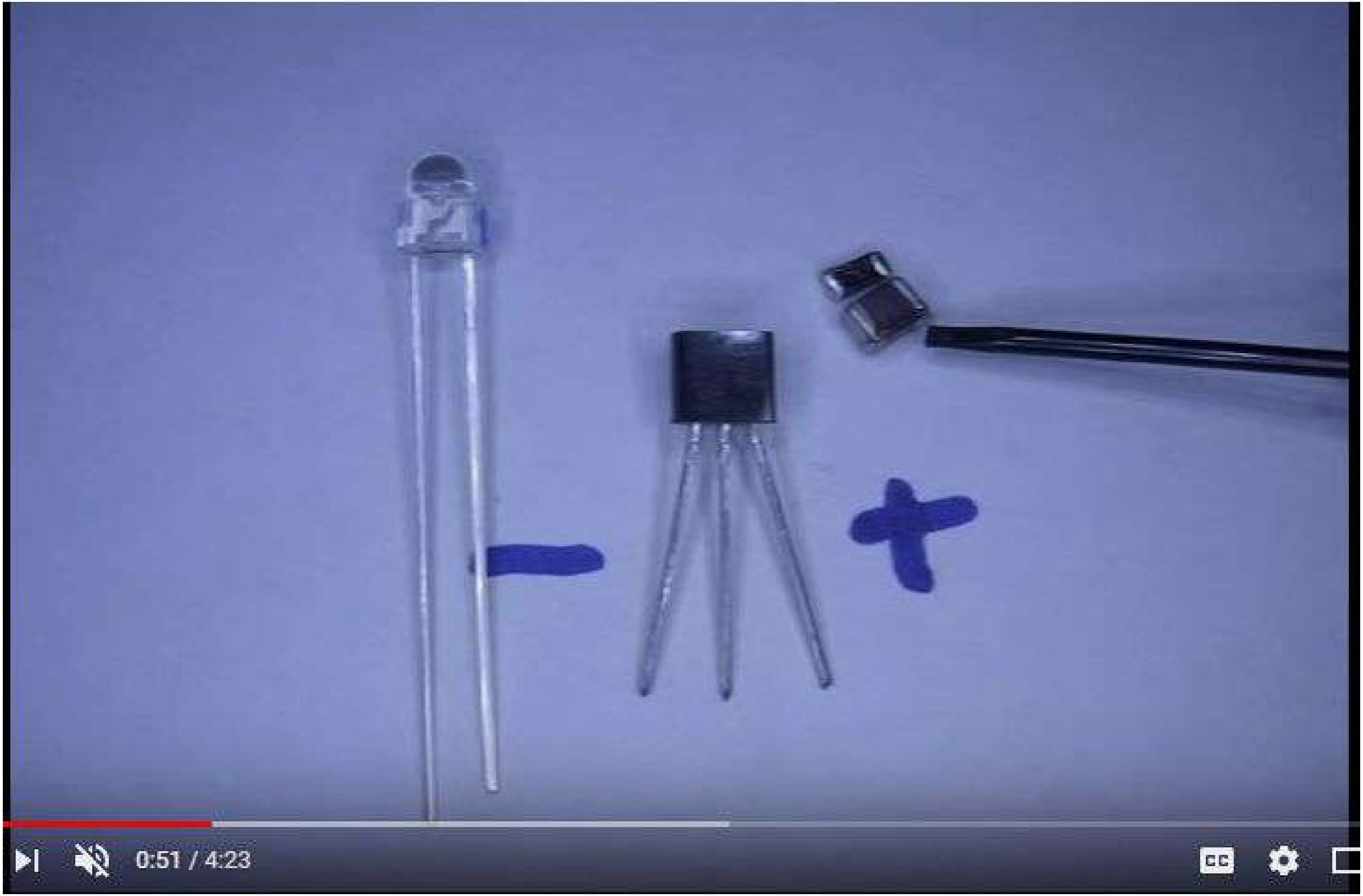
-
+
0:51 / 4:23

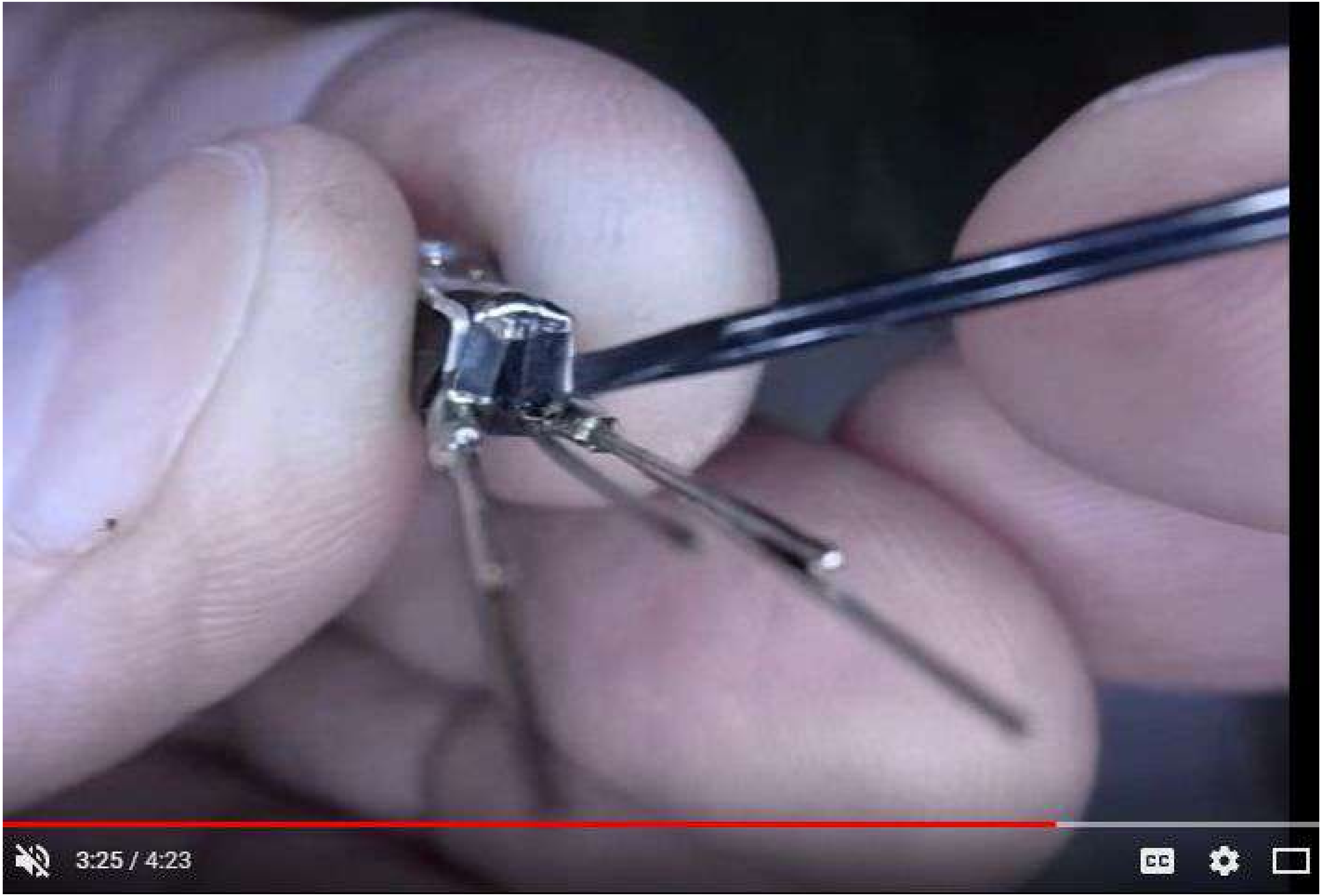
3:25 / 4:23
CC

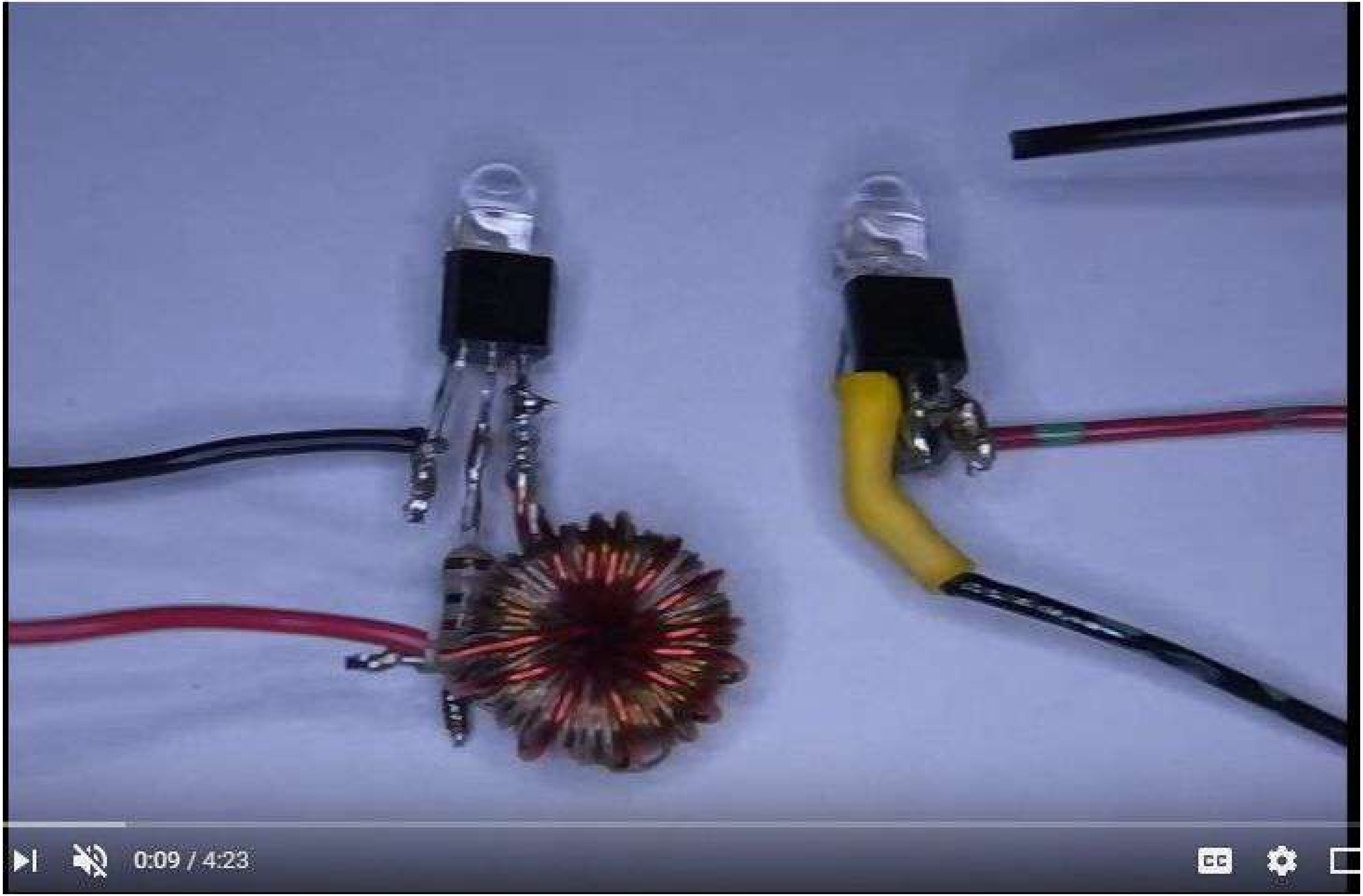
0:09 / 4:23

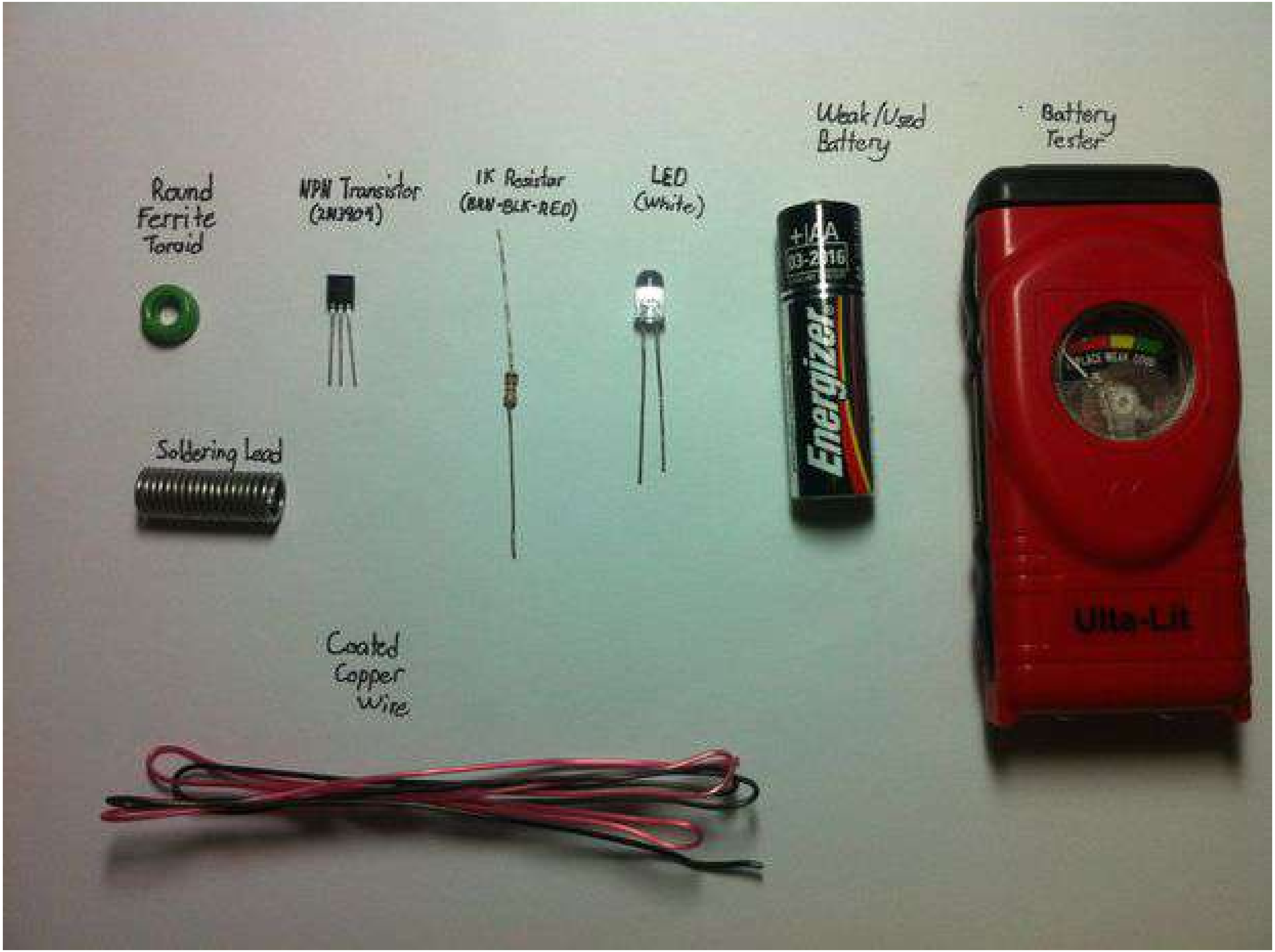
Round Ferrite Toroid
NPN Transistor (2N3904)
1K Resistor (BRN-BLK-RED)
LED (White)
Weak/Used Battery
Battery Tester
+AA
03-2016
Energizer
Ulta-Lit
Soldering Lead
Coated Copper Wire

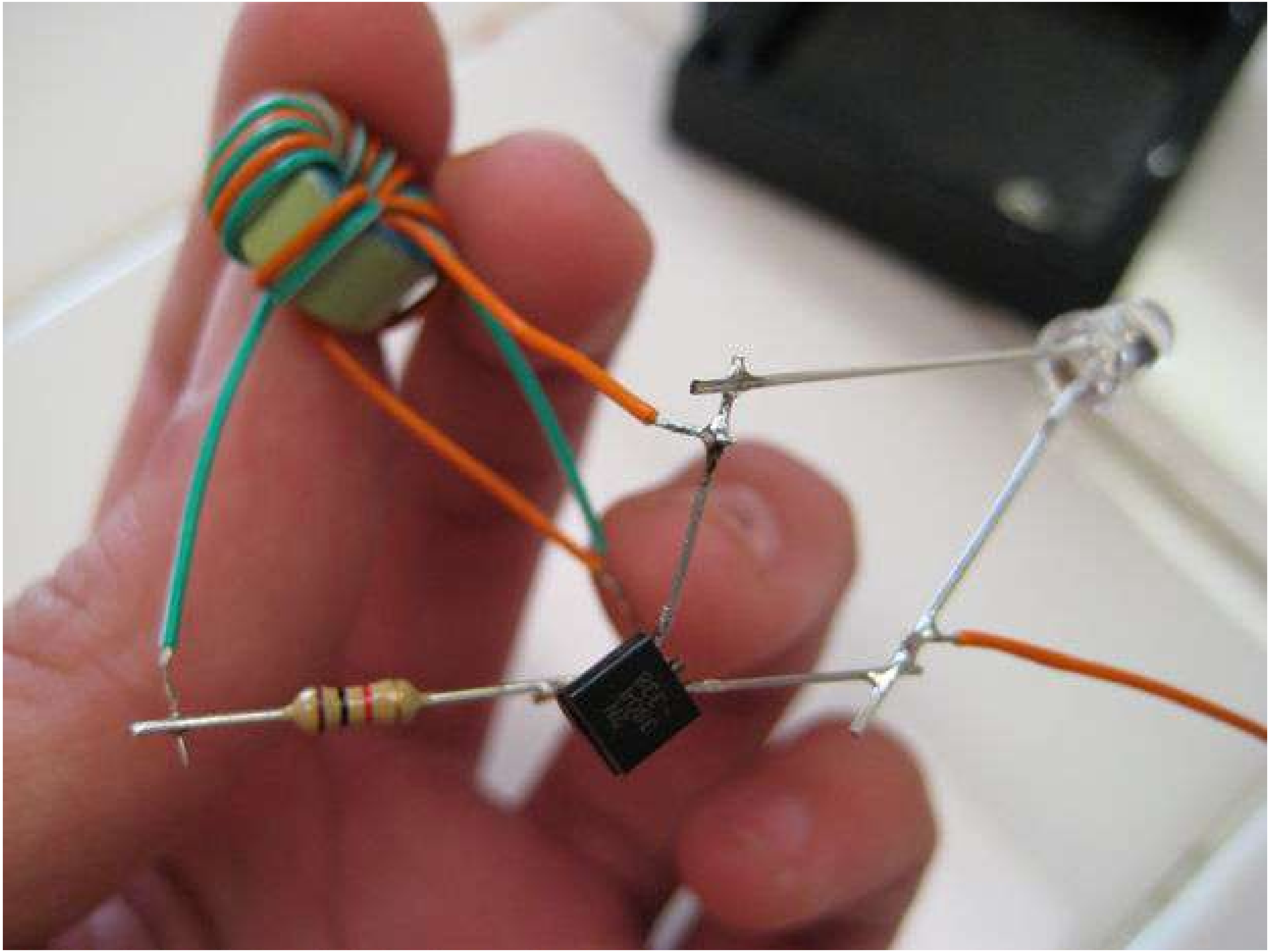

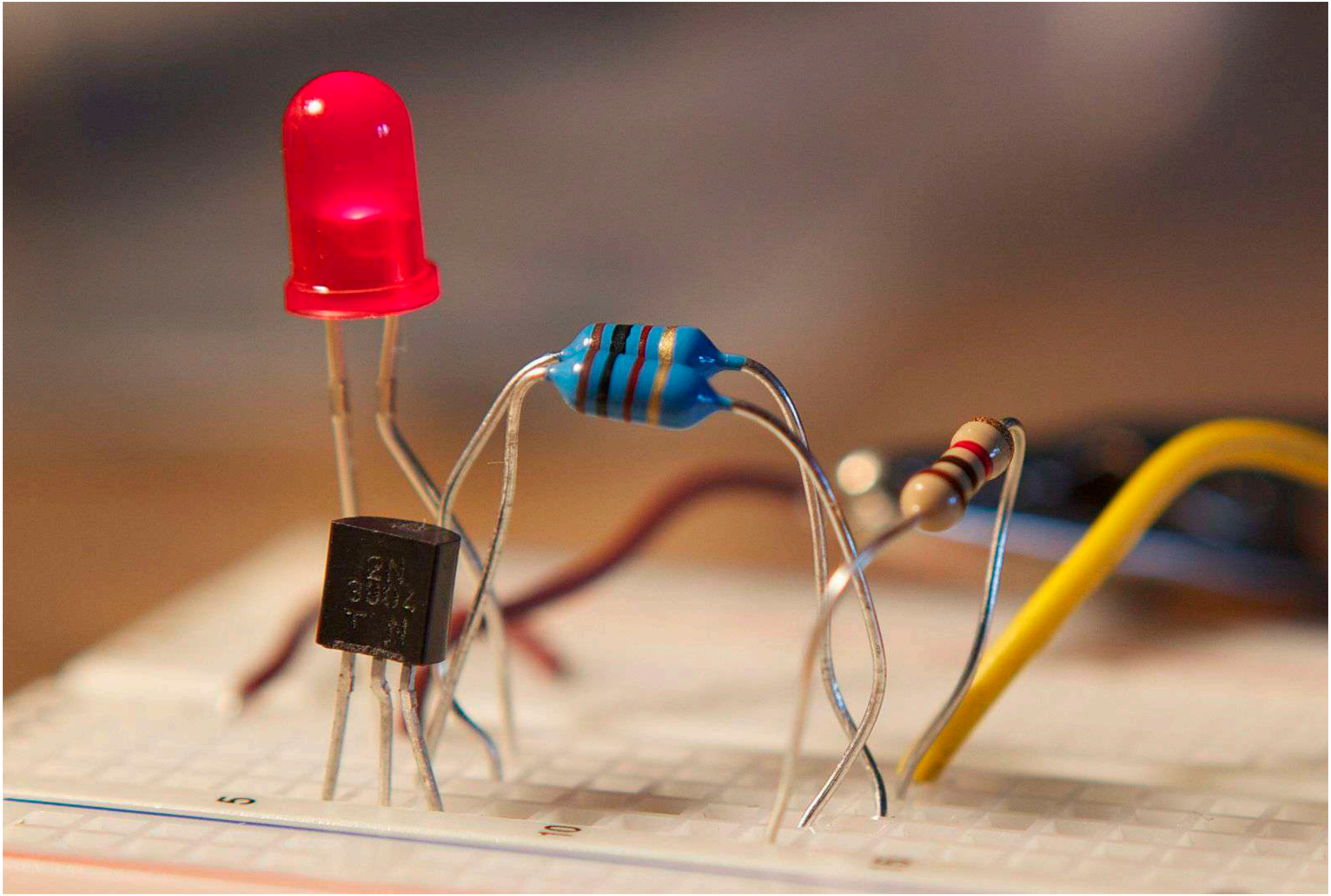
2N
3904
5
10

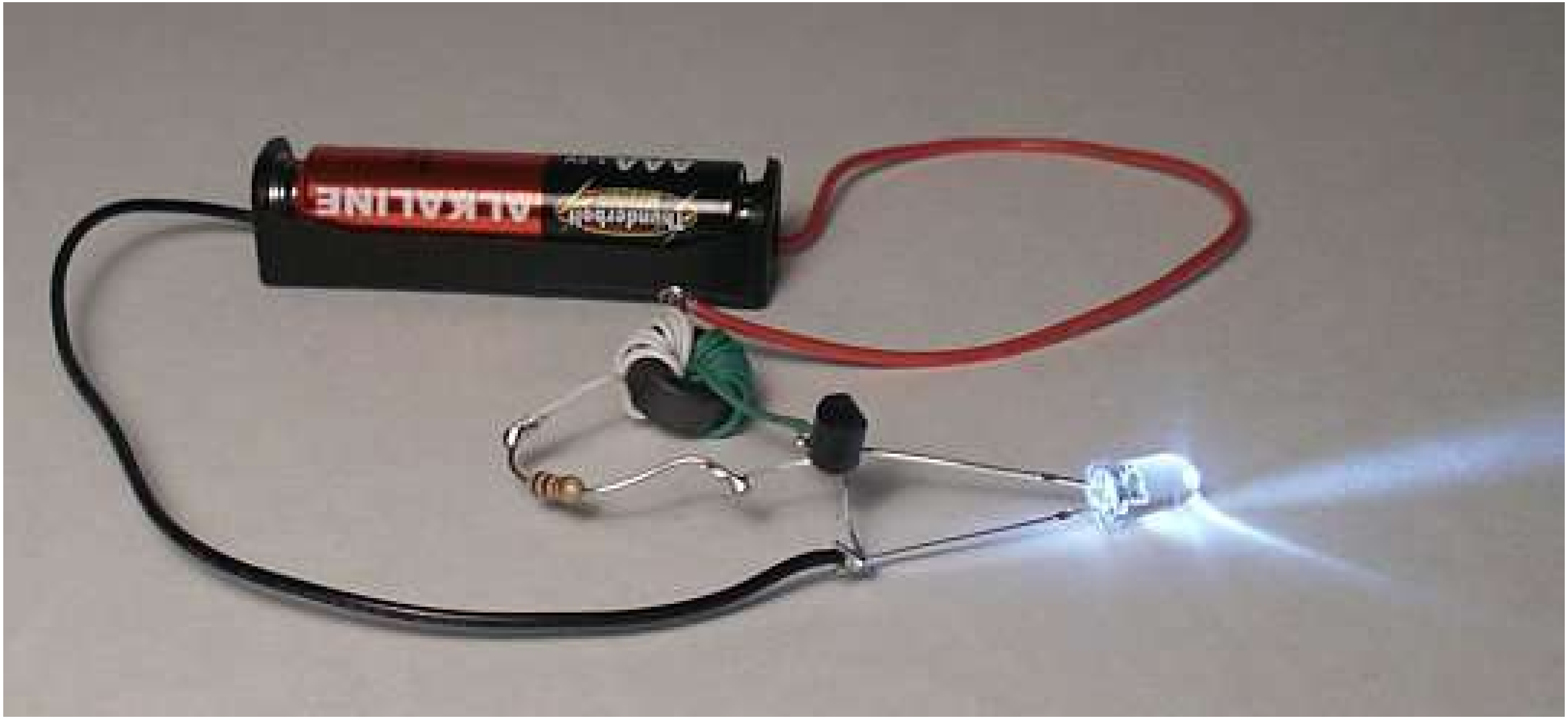
ALKALINE
Thunderbolt

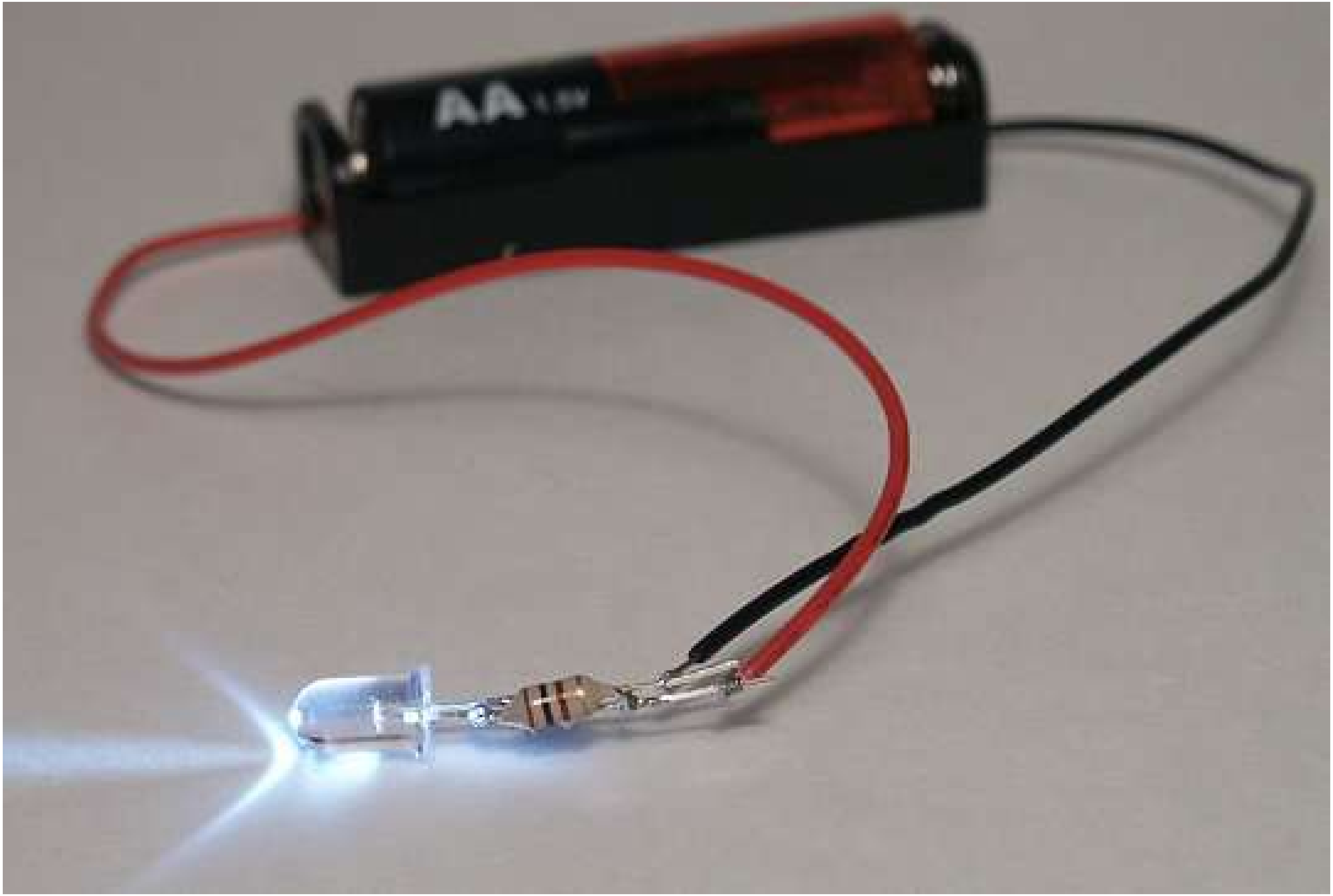
AA 1.5V

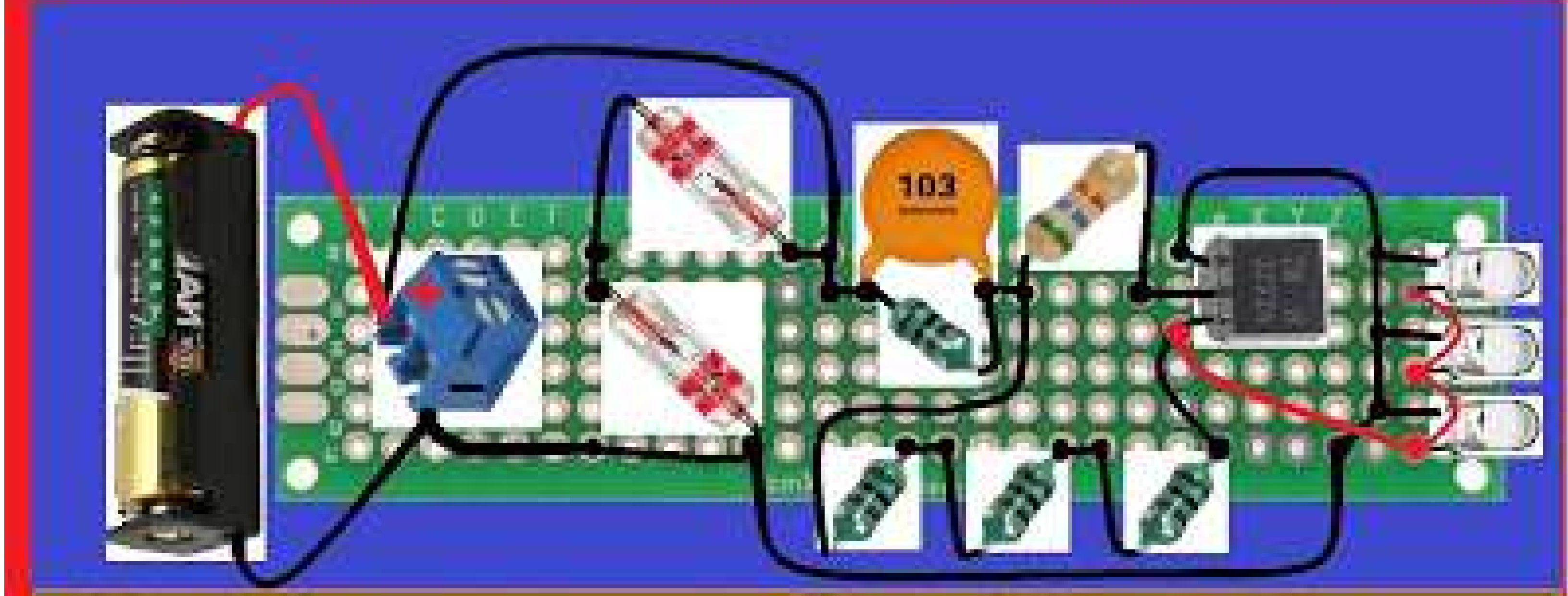

3* Axial lead inductor 82uH 1/4W

1* 10nf ceramic capacitor 103

2* AO90 Germanium signal diode

1* 2N2222 NPN Transistor bipolar 0.6A 40V

1* RESISTOR 5.6K Ohm 1/4W

3* 5mm white Emitting led Diode 20 mA 3.2 V

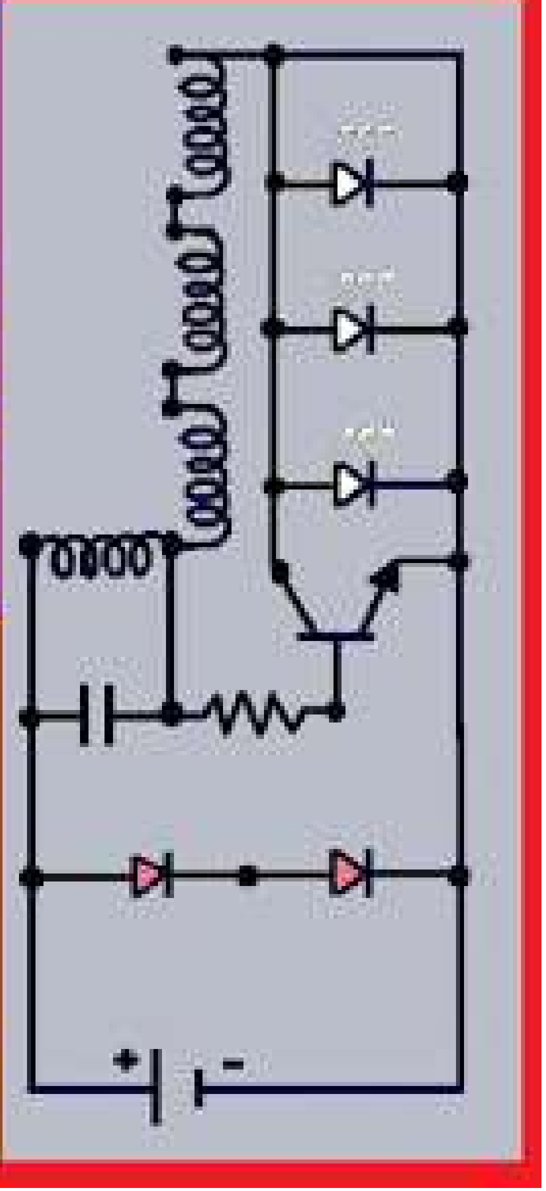

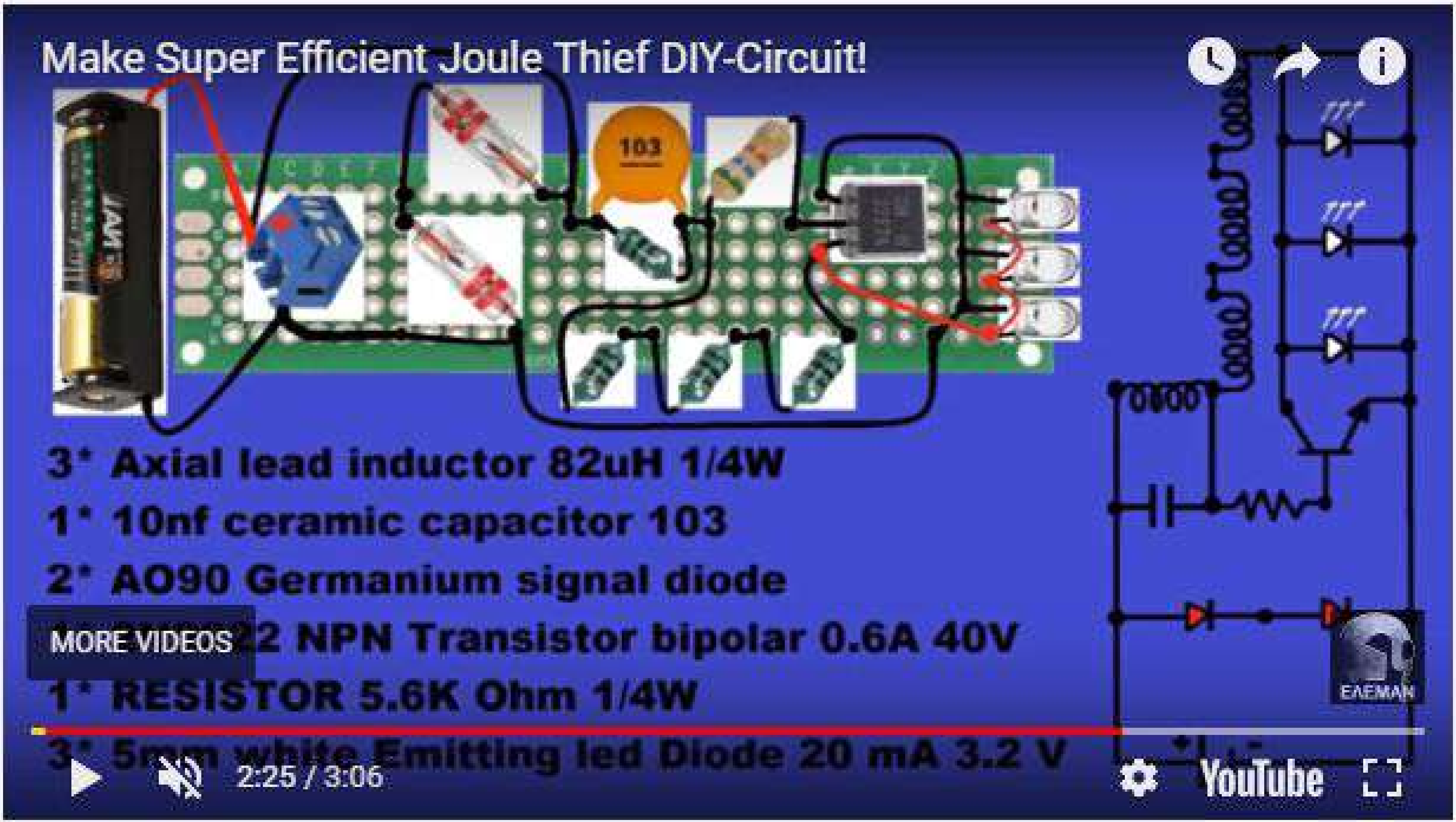
103
MORE VIDEOS
2:25 / 3:06
YouTube
ΕΛΕΜΑΝ

Make Super Efficient Joule Thief DIY-Circuit!

3* Axial lead inductor 82uH 1/4W

1* 10nf ceramic capacitor 103

2* AO90 Germanium signal diode

22 NPN Transistor bipolar 0.6A 40V

1* RESISTOR 5.6K Ohm 1/4W

3* 5mm white Emitting led Diode 20 mA 3.2 V

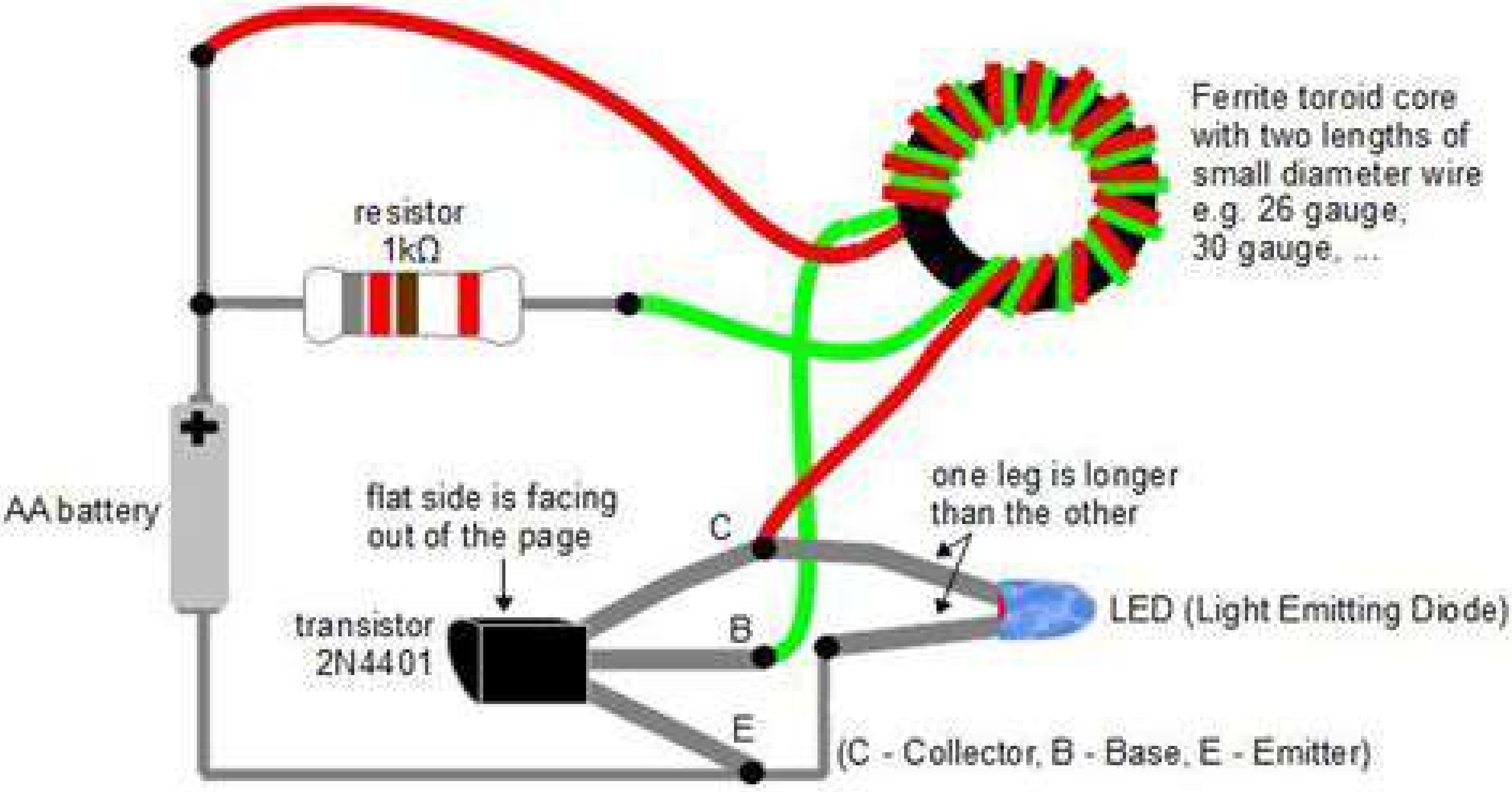
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge;
30 gauge, ...
resistor
1kΩ
AA battery
+
flat side is facing
out of the page
one leg is longer
than the other
C
transistor
2N4401
B
E
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

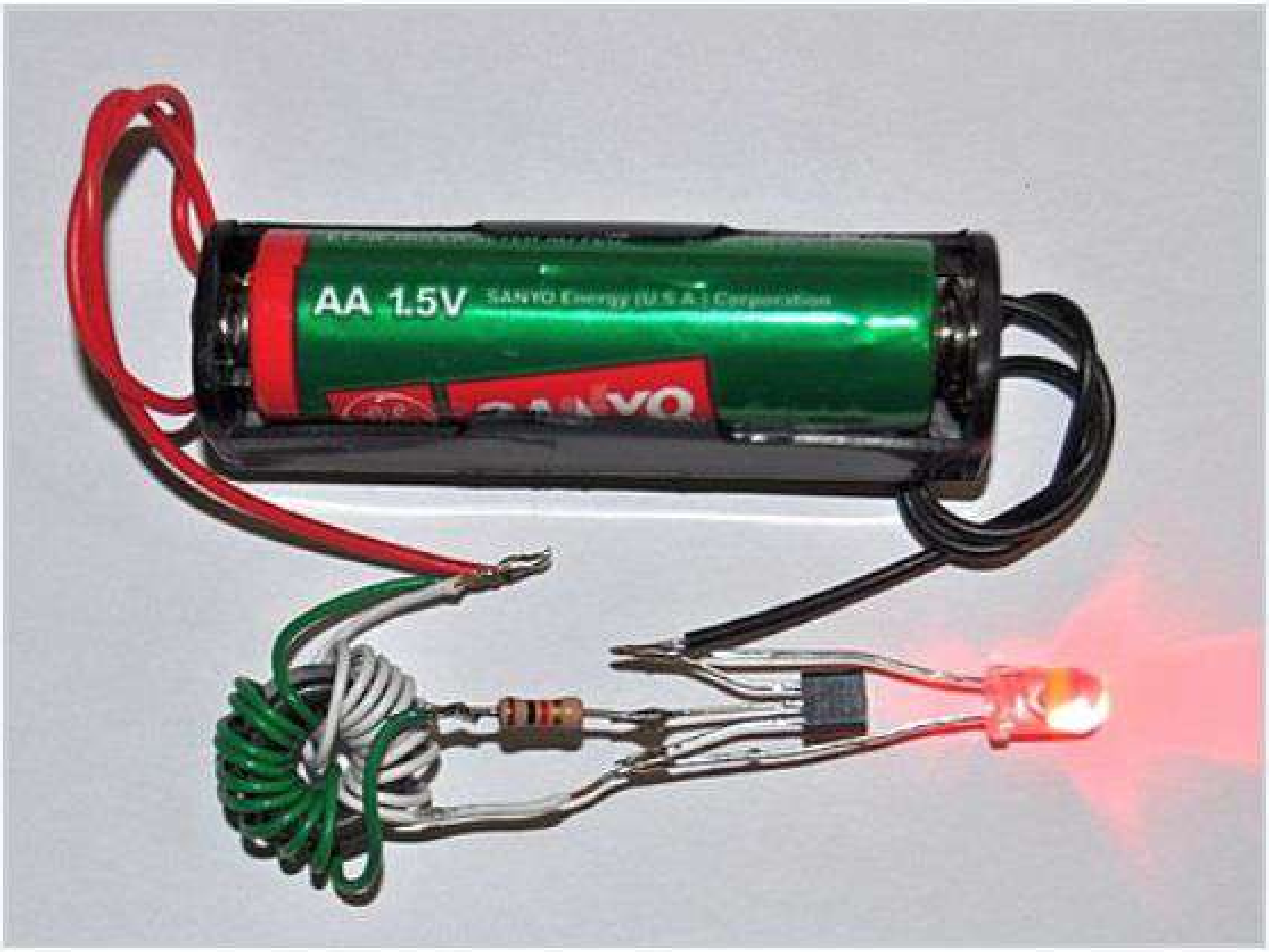
AA 1.5V
SANYO Energy (U.S.A.) Corporation
SANYO

A conventional joule thief, shows components and how they are connected. This example uses a red LED. A ferrite toroid is wound to form a coil with primary (white) and feedback (green) windings. A 2N2222A transistor and 1000 ohm resistor are used.

The Joule Thief

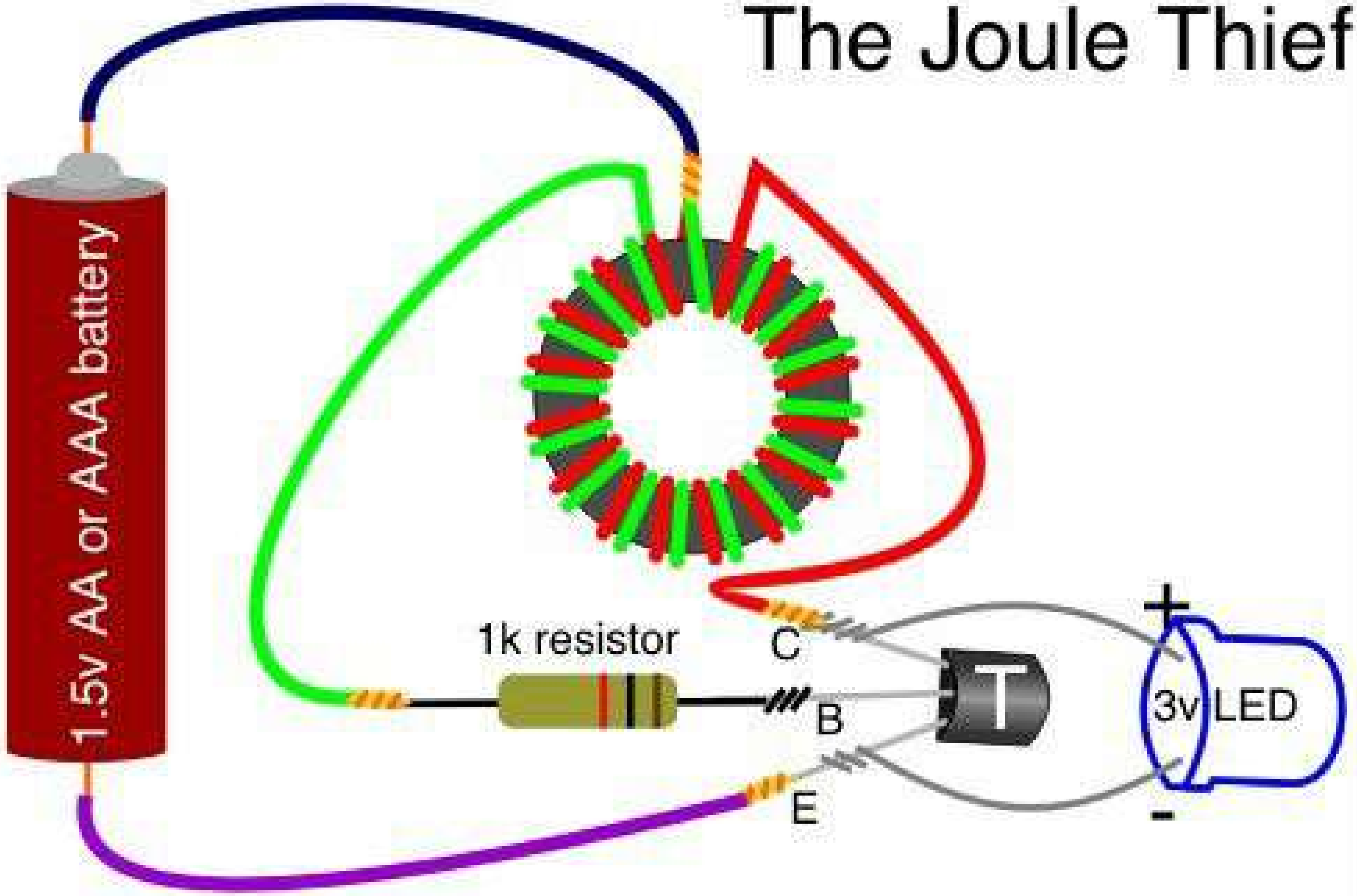
1.5v AA or AAA battery
1k resistor
C
B
E
T
+
-
3v LED

Transistor shown (T) is 2N3904 (can also use 2N2222).

Note LED pins - negative is shorter and has flat on bulb edge.

2 wires round ferrite ring are wound together, then start of one joined to end of other.

Try adding more and more LEDs...

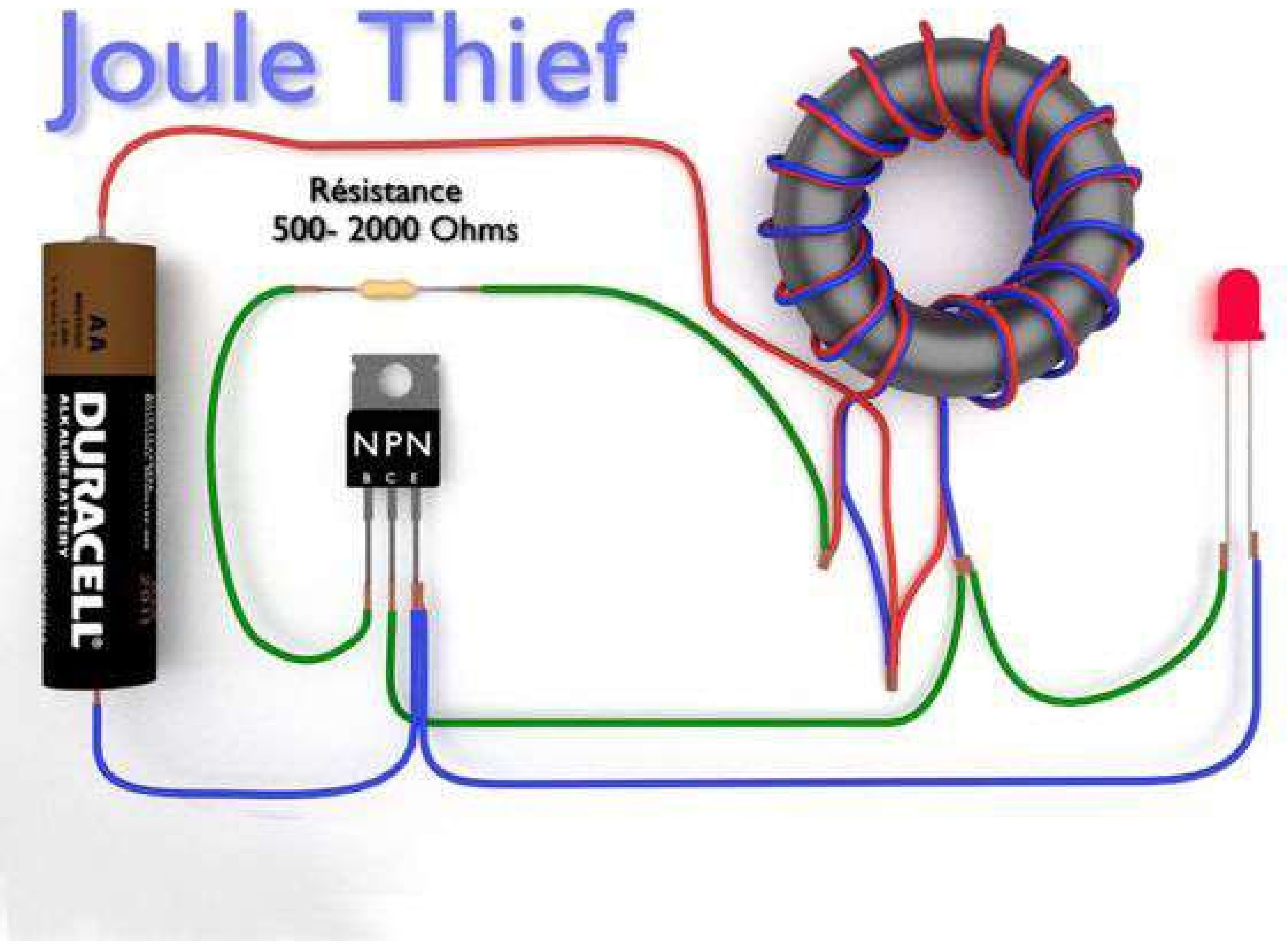
Joule Thief
Résistance
500- 2000 Ohms
NPN
B C E
AA
DURACELL
ALKALINE BATTERY

Its a Rubbish Challenge
Dog Light

Joule Thief
electronics circuit

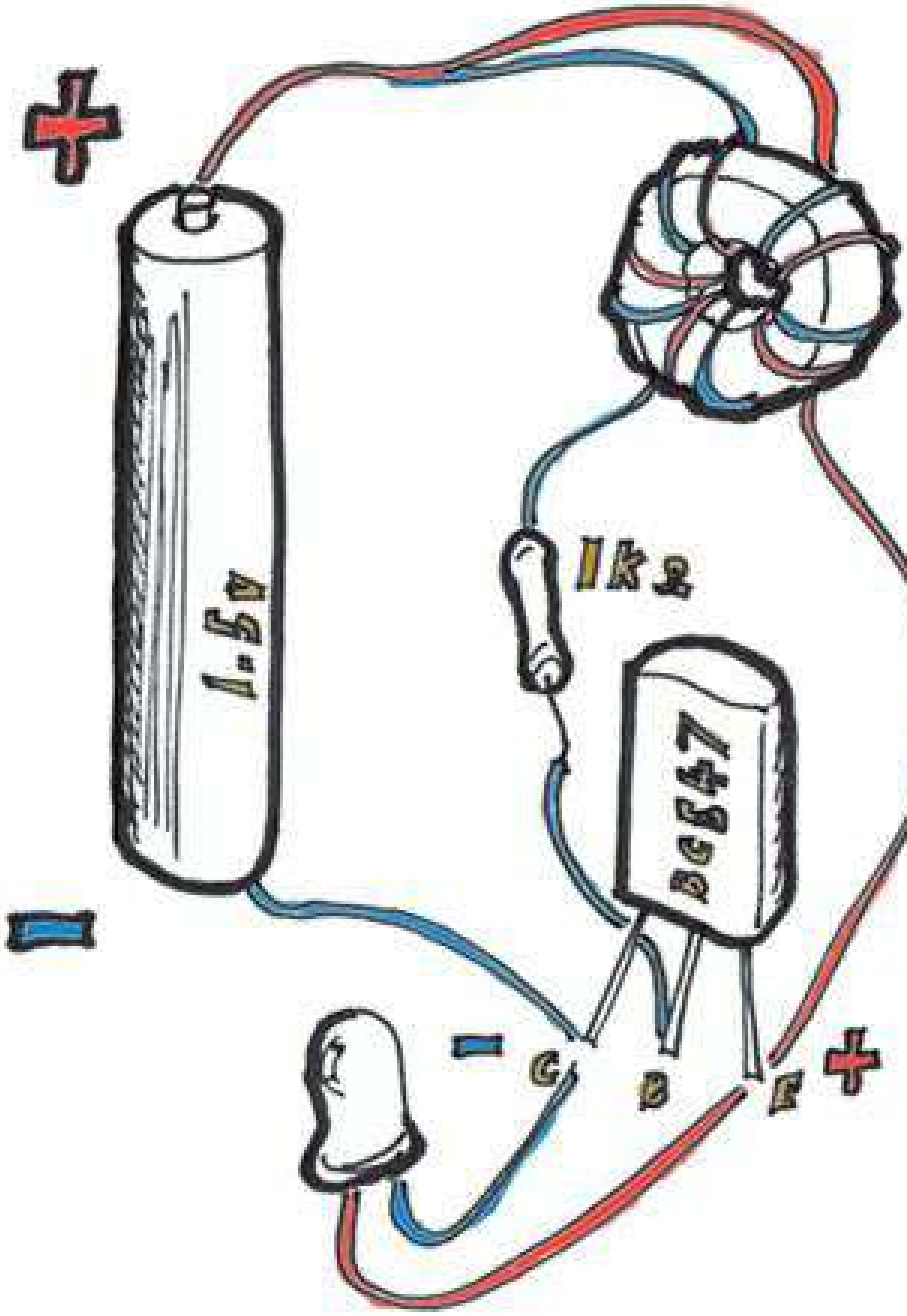

+
1.5v
-
1kΩ
BC547
-
c
e
B
+

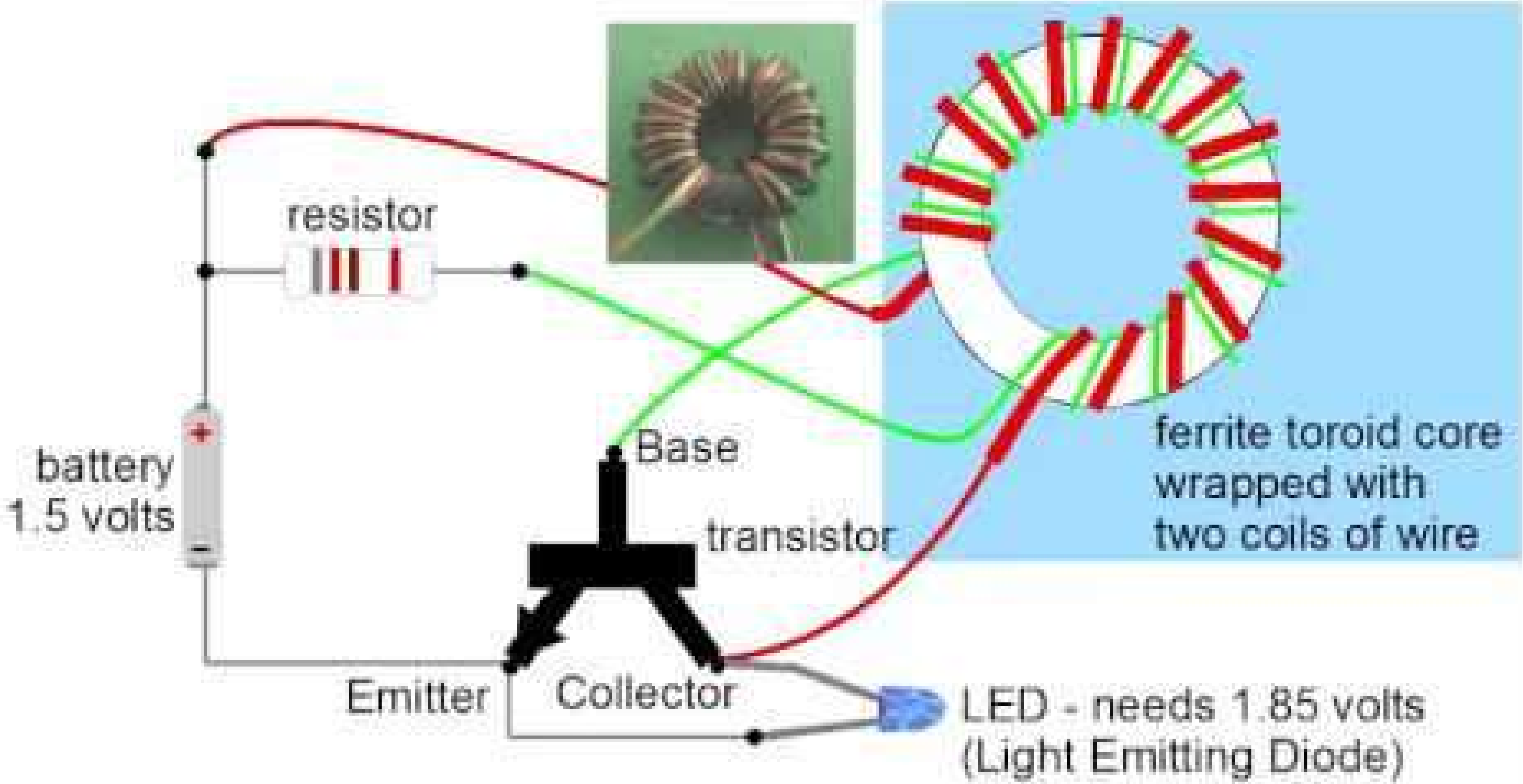
resistor
battery
1.5 volts
+
−
Base
transistor
Emitter
Collector
ferrite toroid core
wrapped with
two coils of wire
LED - needs 1.85 volts
(Light Emitting Diode)

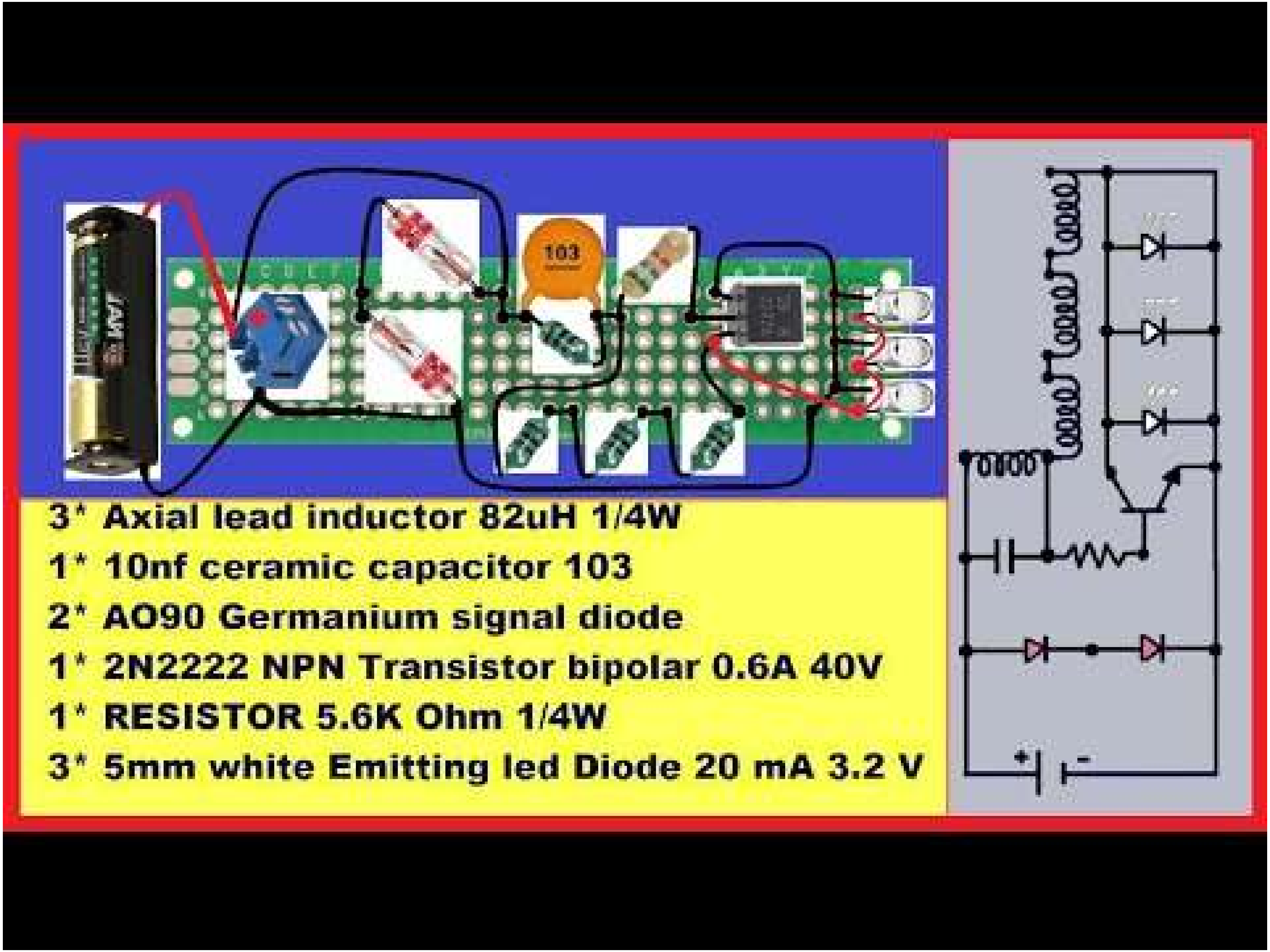

3* Axial lead inductor 82uH 1/4W

1* 10nf ceramic capacitor 103

2* AO90 Germanium signal diode

1* 2N2222 NPN Transistor bipolar 0.6A 40V

1* RESISTOR 5.6K Ohm 1/4W

3* 5mm white Emitting led Diode 20 mA 3.2 V

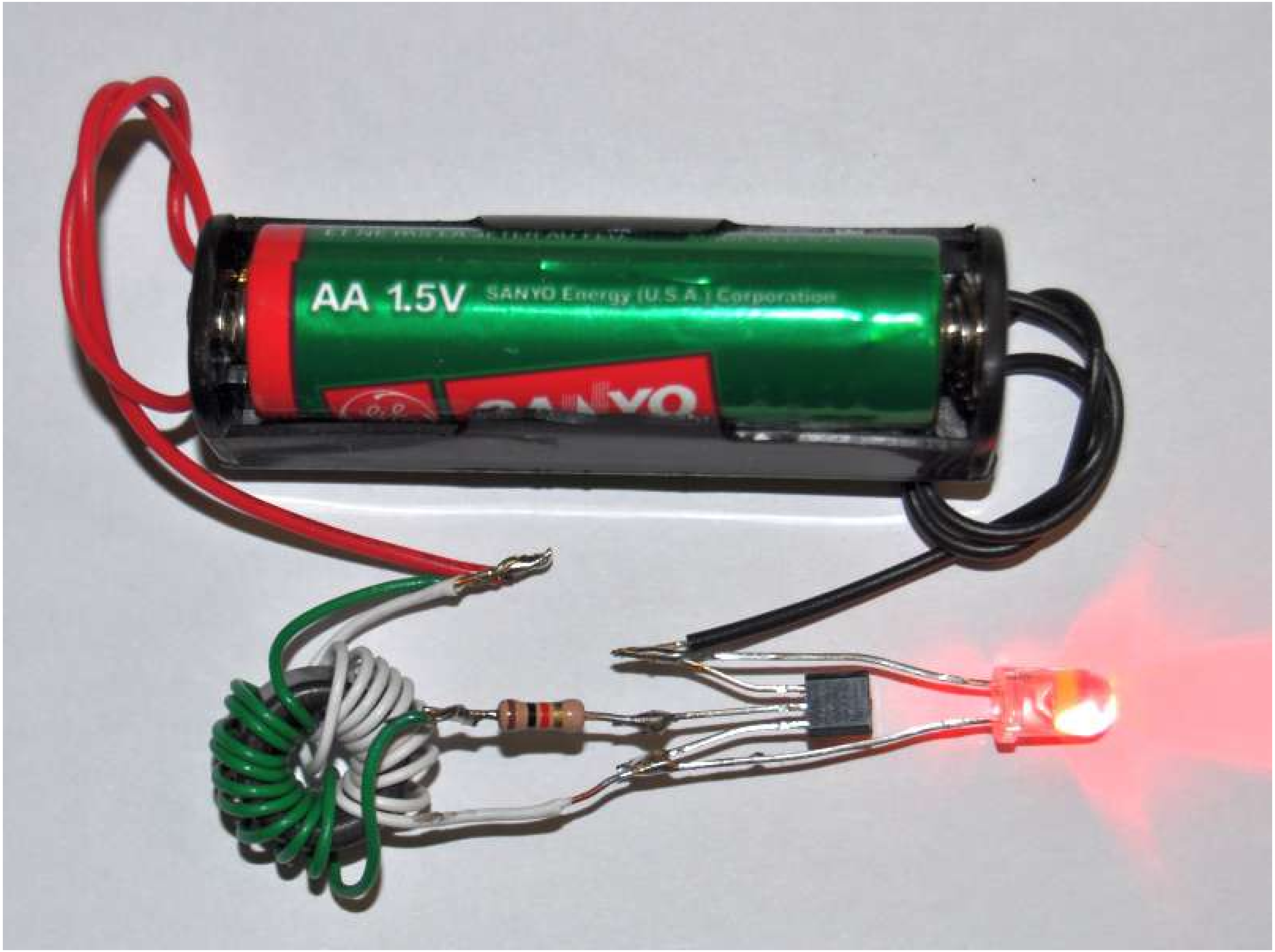
AA 1.5V
SANYO Energy (U.S.A.) Corporation
SANYO

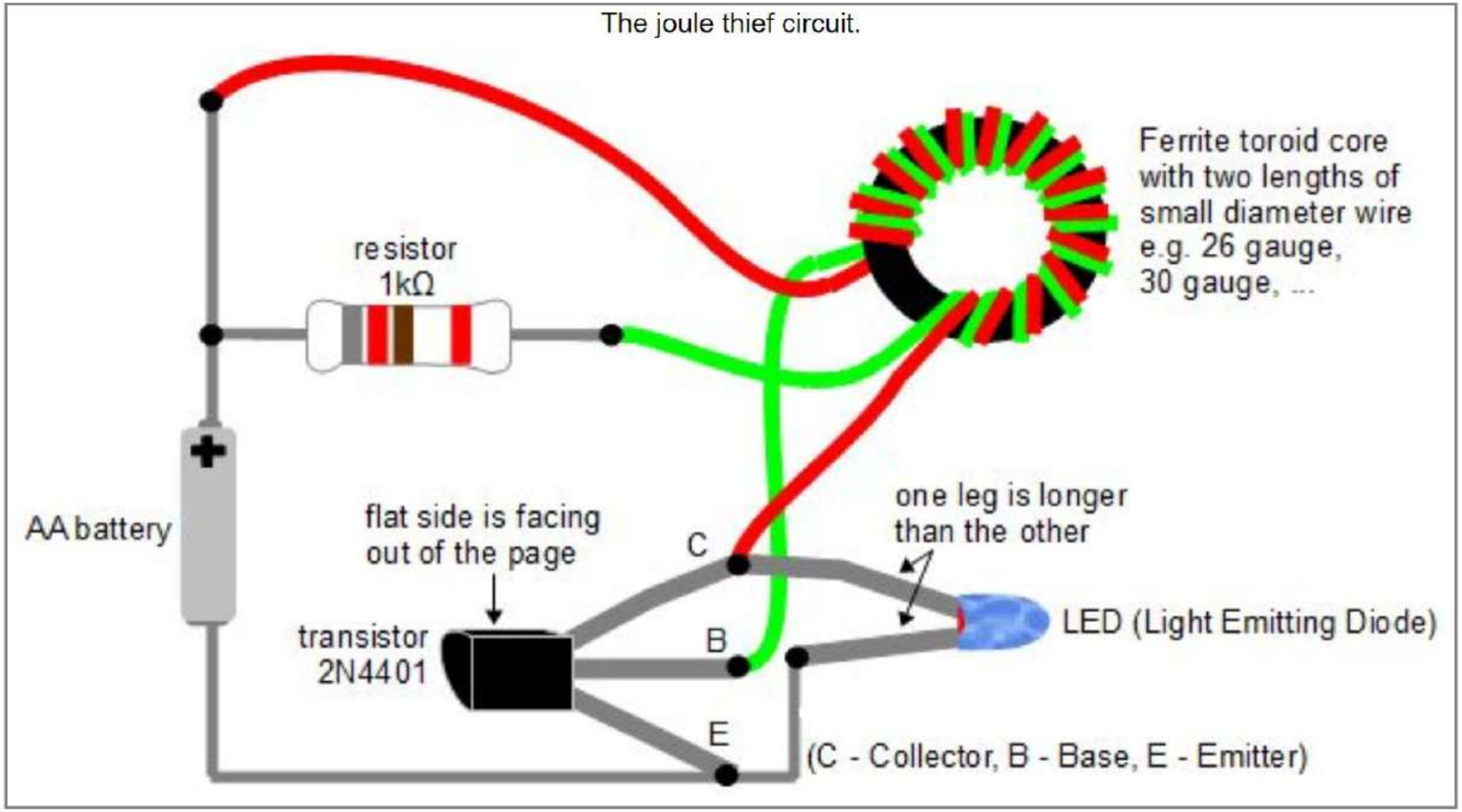
The joule thief circuit.
Ferrite toroid core with two lengths of small diameter wire e.g. 26 gauge, 30 gauge, ...
resistor 1kΩ
AA battery
flat side is facing out of the page
transistor 2N4401
C
B
E
one leg is longer than the other
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

Transistor - The legs of the transistor can be determined by noticing that there's a flat side to the transistor case. See the diagram above. A large number of transistors have been reported to work: 2N4401, NET123AP, BC547B, 2SC2500, BC337, PN2222, to name just a few.

LED - One leg of the LED is longer than the other leg. Use this to determine which one goes where. See the diagram above.

Resistor - The diagram says use a 1 kilo ohm resistor but I've used an 820 ohm one just fine. I've also seen a 2 kilo ohm one in use. Use whatever works for you. You can also use a potentiometer (a variable resistor) so that you can easily adjust it to select the resistance that gives the best light.

Toroid ferrite core - Some people have gotten these by opening up compact fluorescent lightbulbs (CFLs). I took mine out of some device whose original function I don't know. To get it working, my first one had just 13 turns for each wire and I used a 30 gauge wire and a 26 gauge wire. The wire must be insulated. A variety of number of turns will work. This is something you can play with. Look at the diagram carefully to determine where the wires connect to.

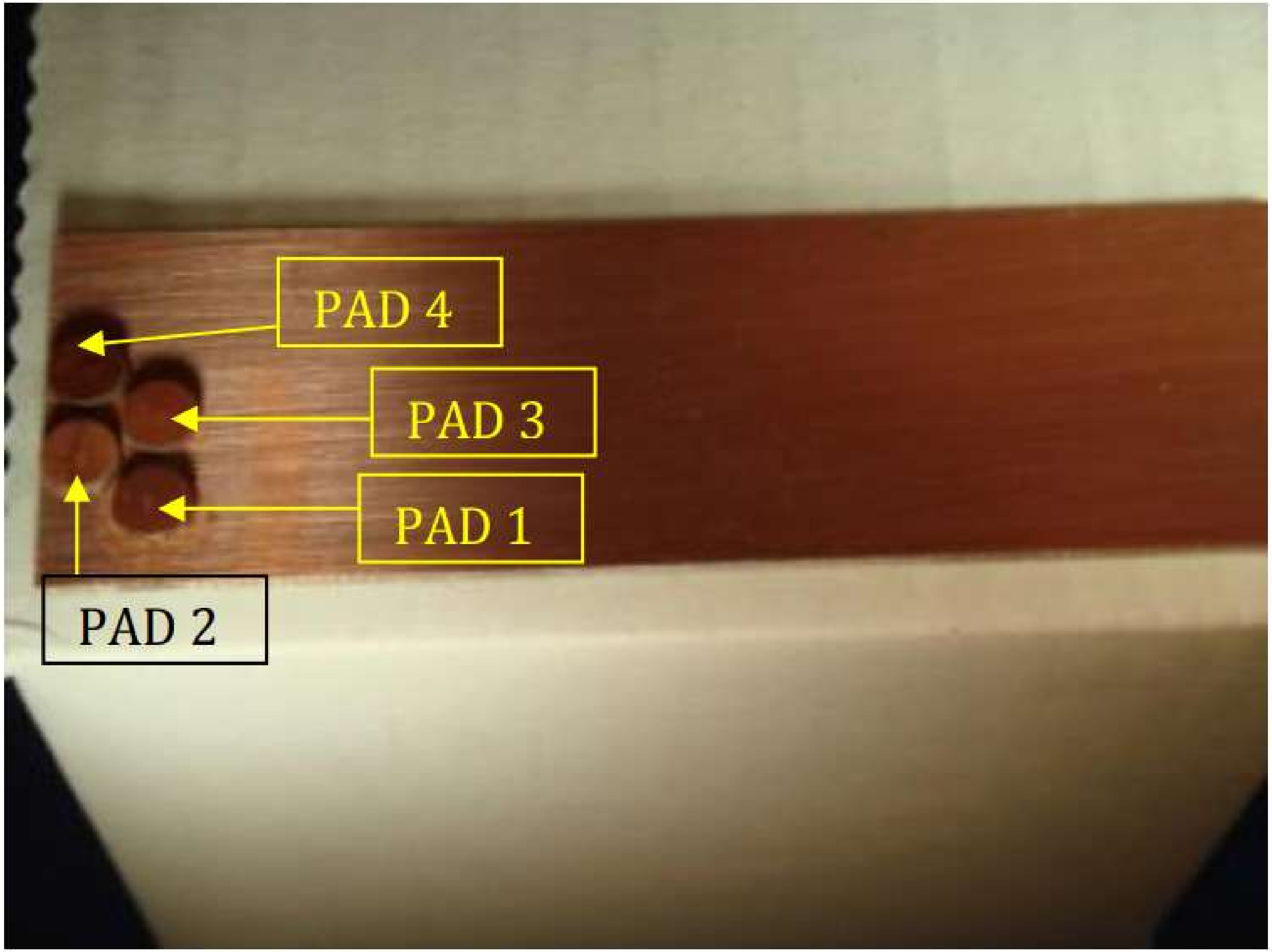

Superglue four pads to the PC board. Make sure they are not touching each other. You should have roughly 1/8 inch between the pads. Apply a small drop of superglue on the board and press the pad onto the board with the small dimple up. Use a small screwdriver to hold pressure on the pad until the glue dries (about 10 seconds).

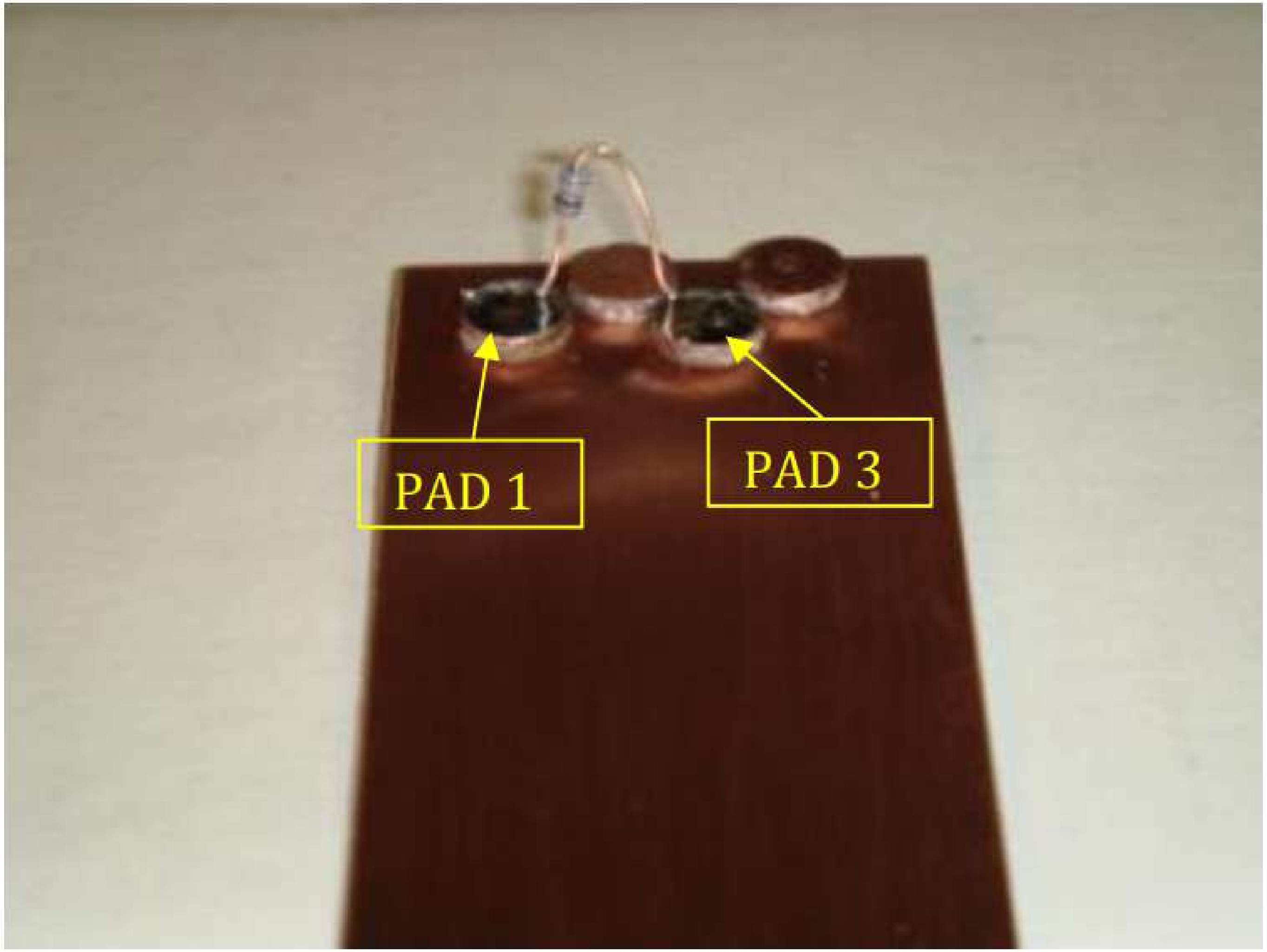

Solder the 1,000 ohm resistor onto pads 1 and 3. When soldering all components, place a 90-degree bend on the ends of the component where they touch the pad to allow better soldering. There is no polarity on a resistor.

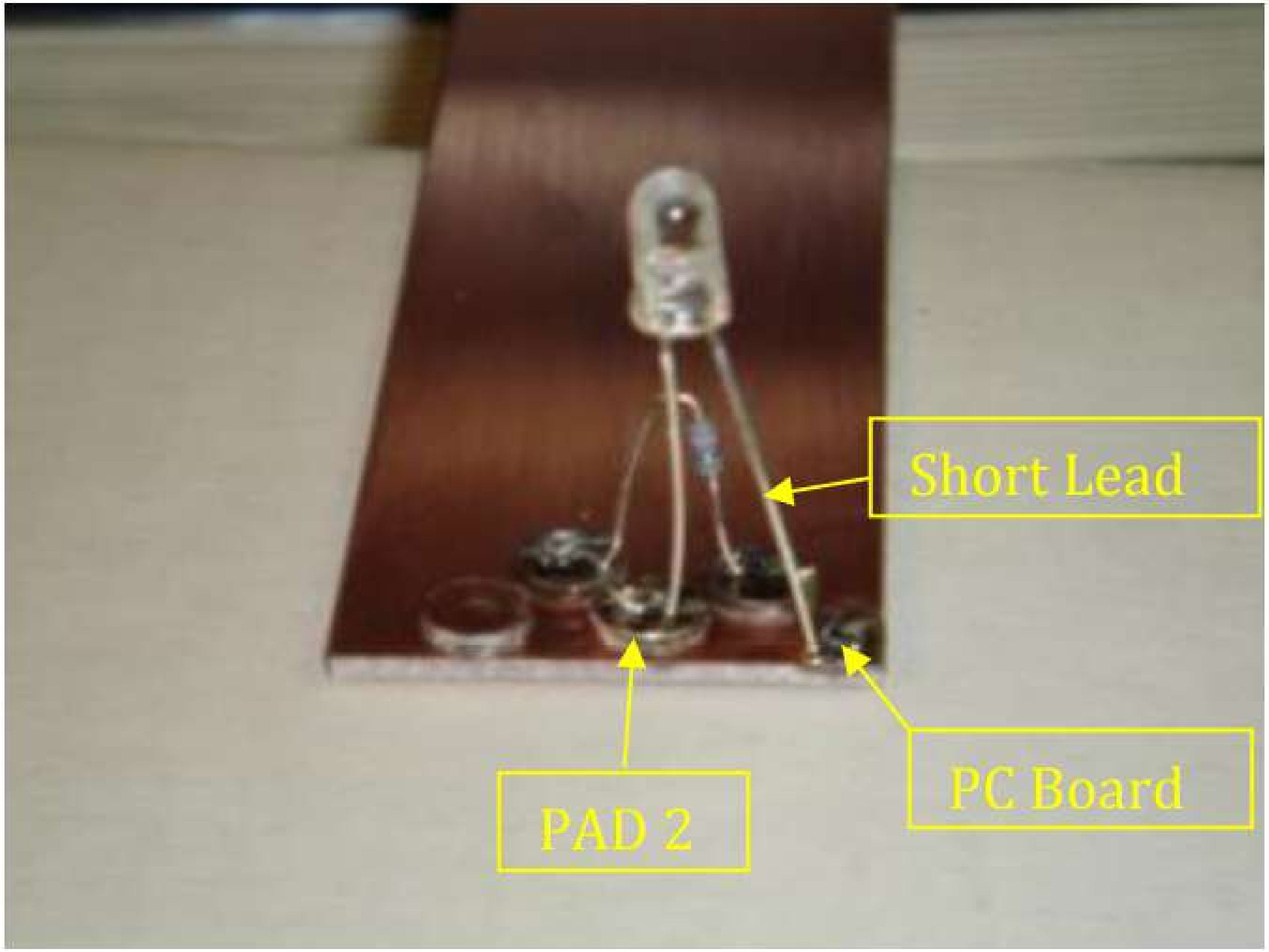

Solder the LED to the board. **VERY IMPORTANT**: Solder the **long lead** to pad 2 and the **short lead** directly to the PC board. The LED will not illuminate of you put it in backwards. If you look into the LED, the anvil shaped part is the lead that gets soldered to the PC board.

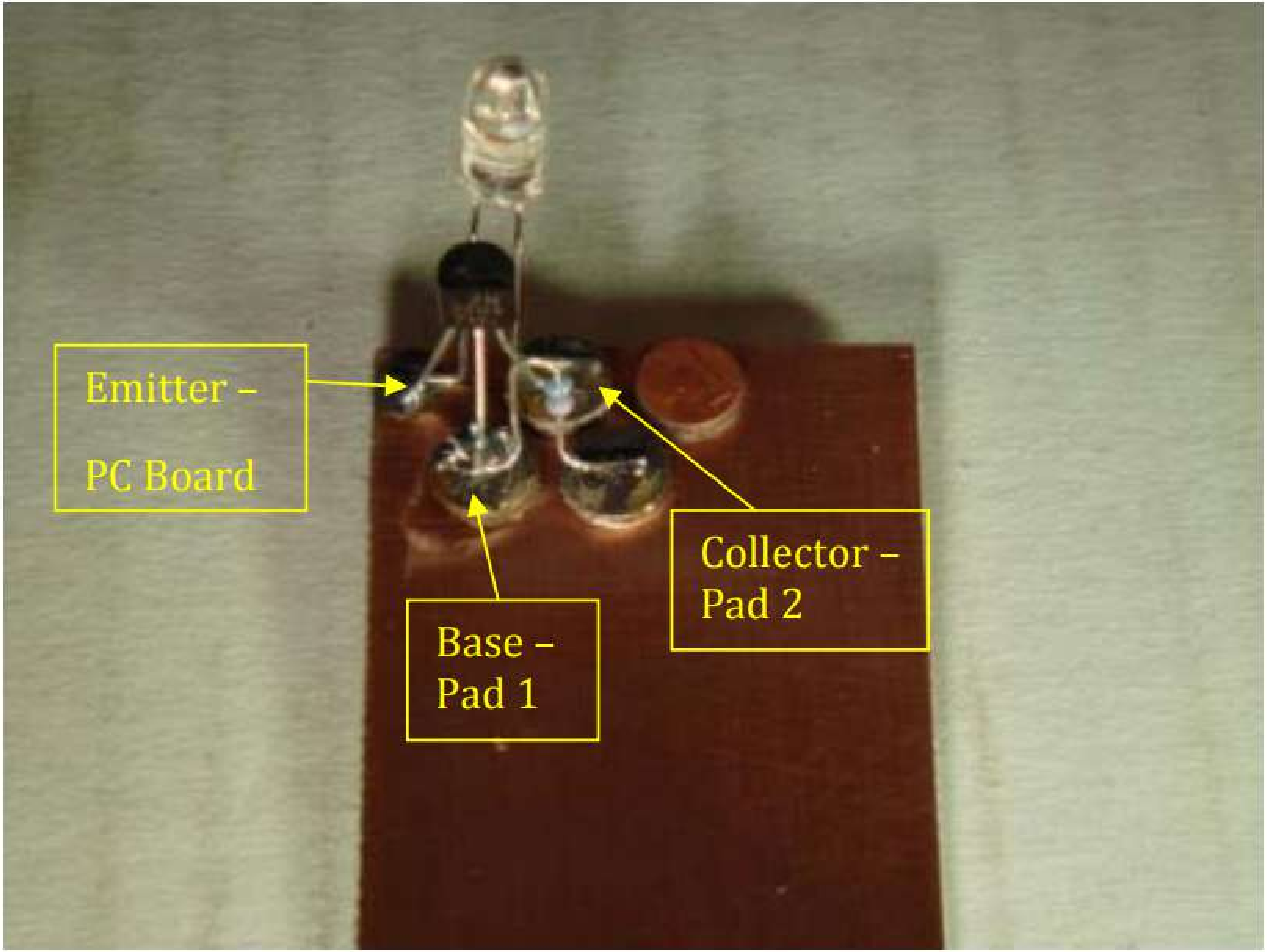
Emitter – PC Board
Base – Pad 1
Collector – Pad 2

Solder the transistor to the board. With the flat part facing you the left lead (emitter) is soldered directly to the PC board. The center lead (base) is soldered to pad 1. The right lead (collector) is soldered to pad 2. Failure to install this part correctly will prevent the Joule Thief from working.

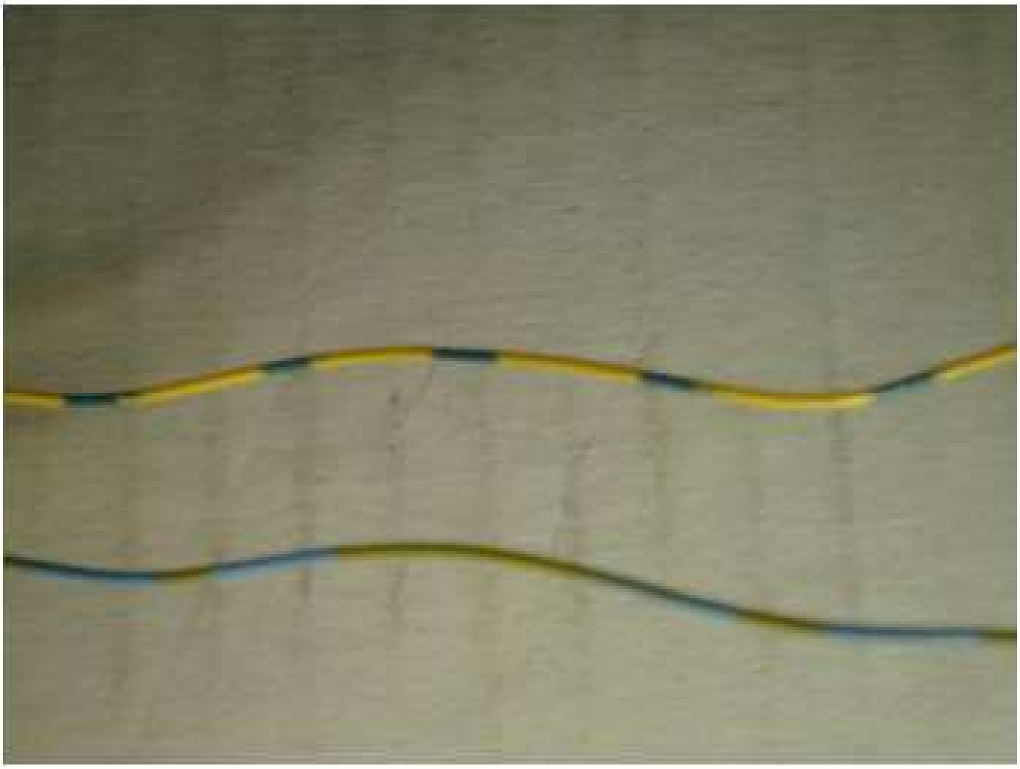

Take the two pieces of wire and **untwist** them if they are twisted. The toroid requires 12 inches of wire. You will wind the toroid with this pair of wires side by side.

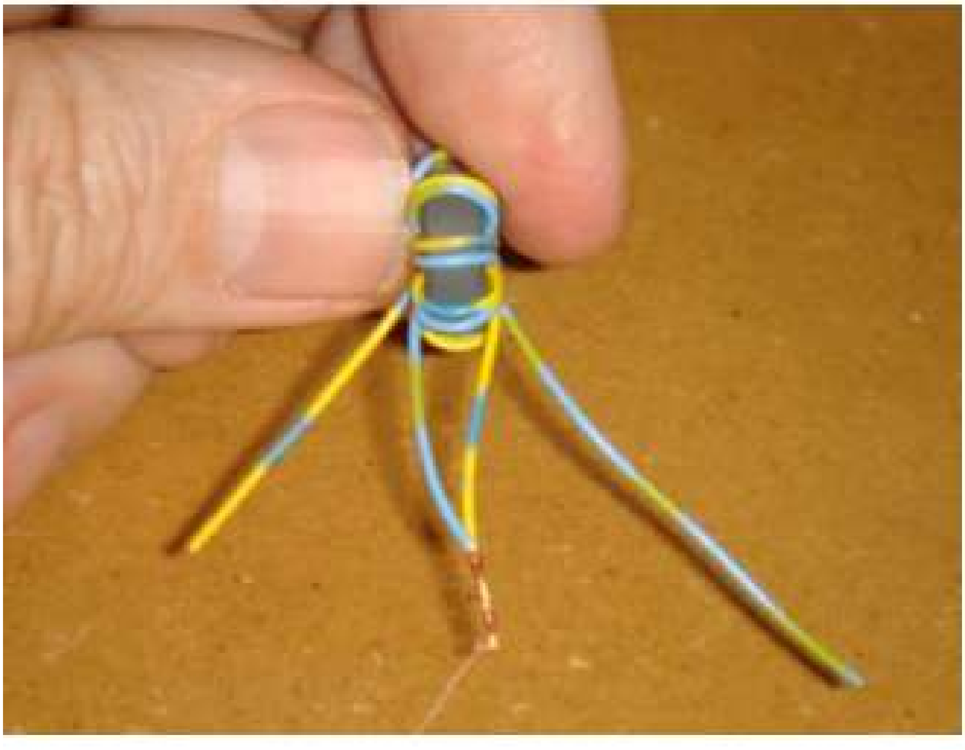

Wind the toroid with **9 turns** of the wire. Each turn through the hole in the center counts as one turn. **VERY IMPORTANT:** You must take one wire **from each side of the winding** and twist them together. This makes the transformer windings out of phase. This is critical to operation. Connect one wire of each color.

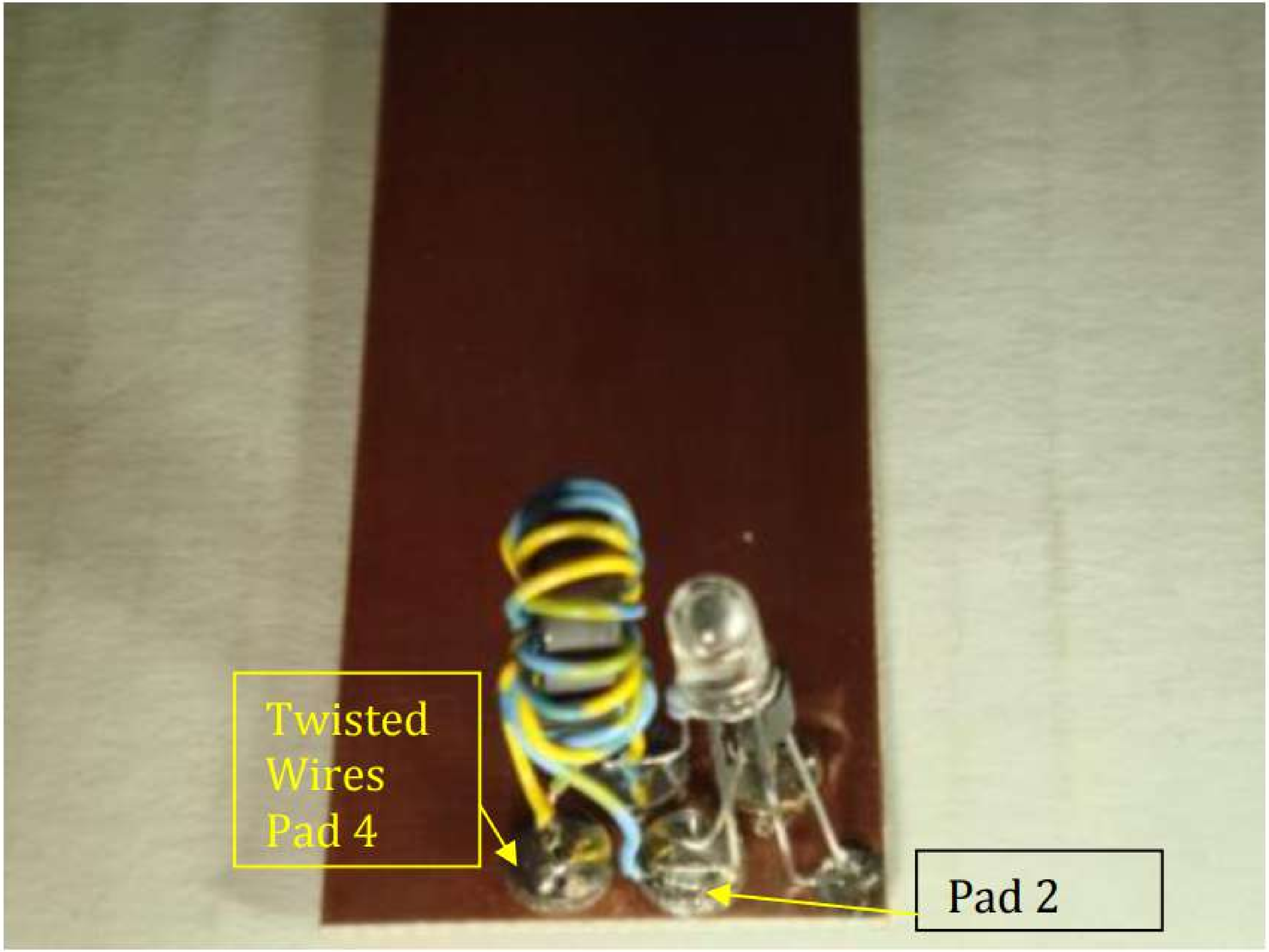
Twisted Wires Pad 4
Pad 2

Solder the twisted wires on the toroid to pad 4. Solder one of the other toroid wires (doesn't matter which one) to pad 2.

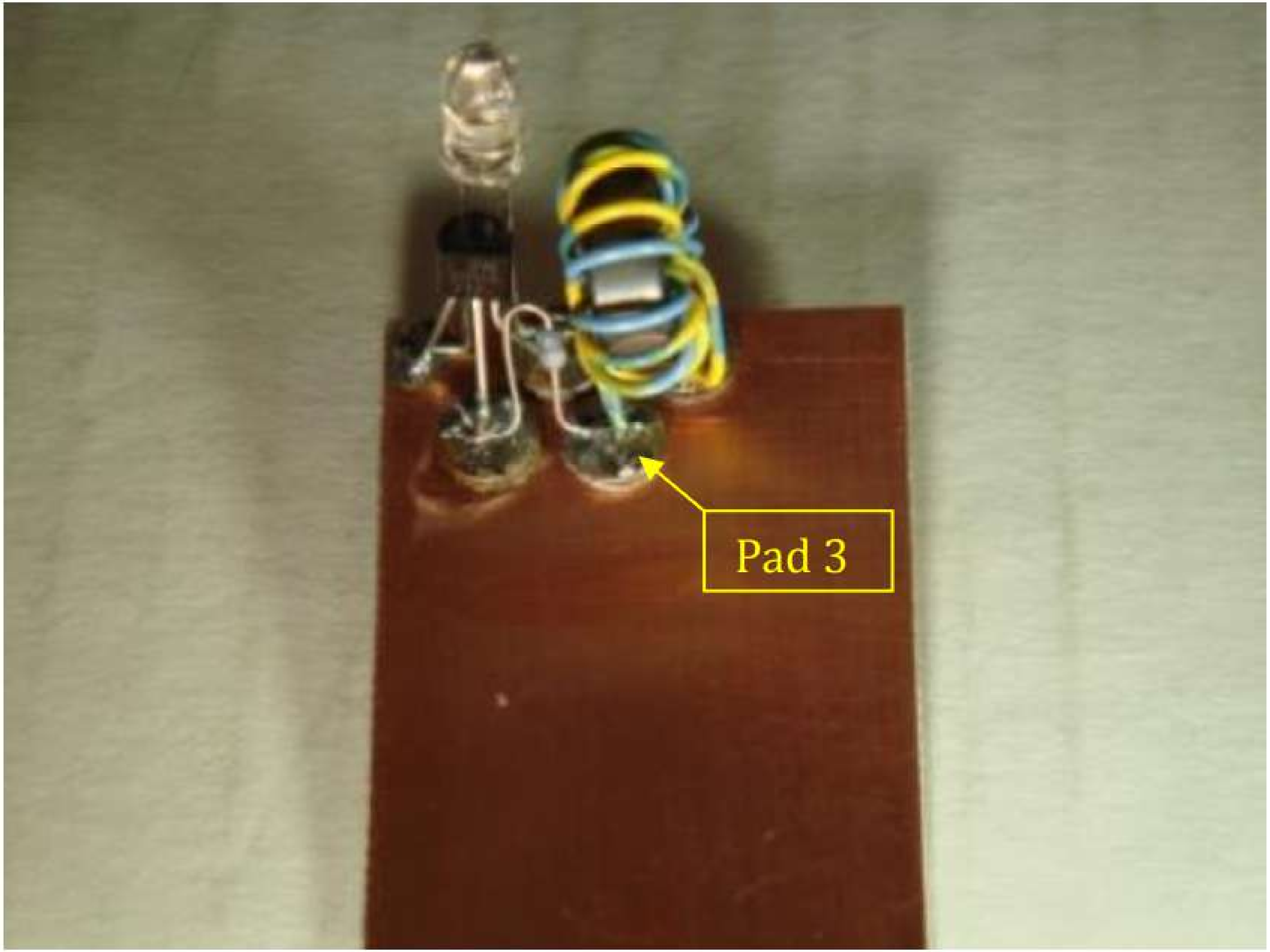

Solder the other wire on the toroid to pad 3.

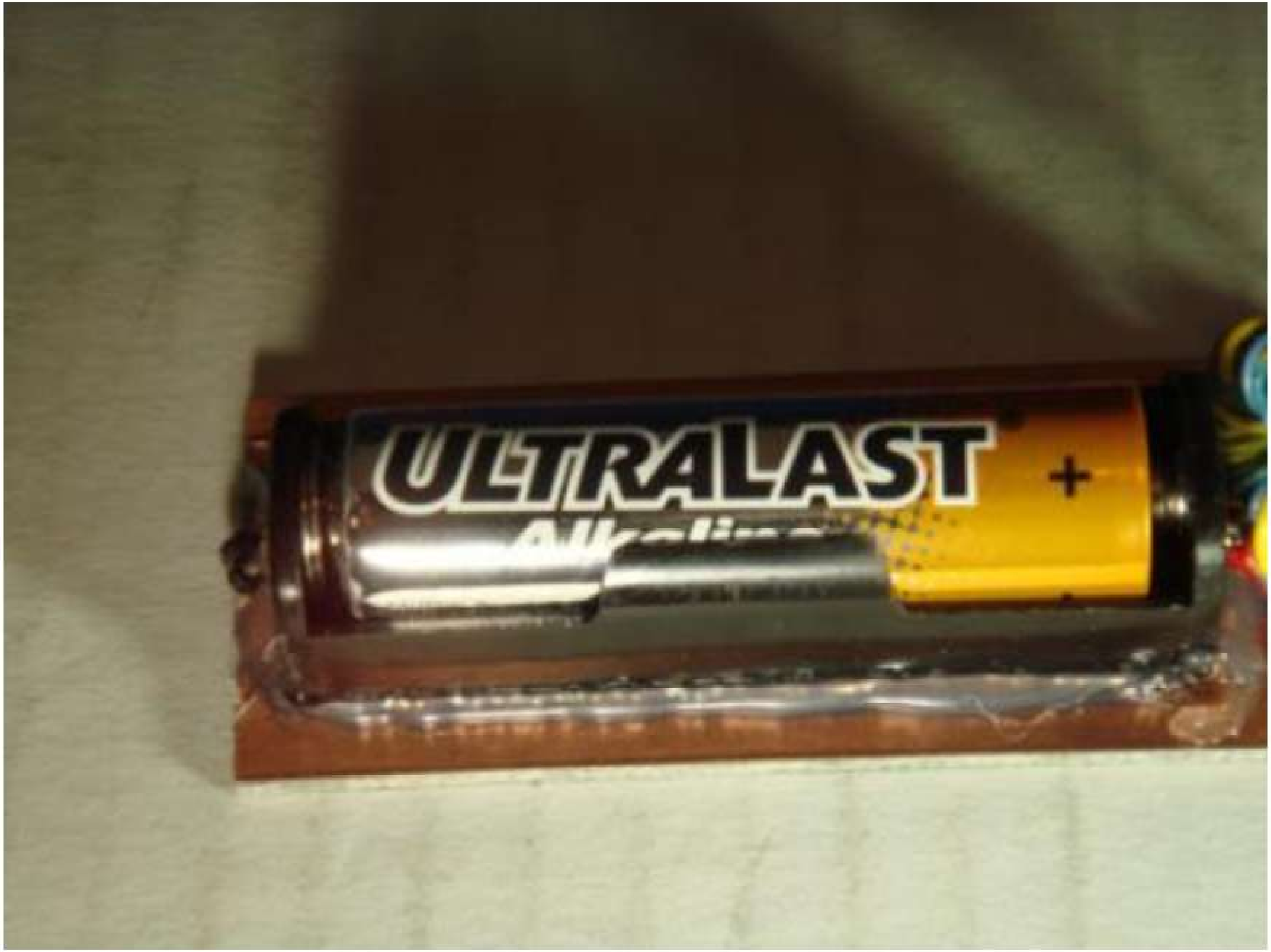

Solder the battery holder. Connect a short piece of wire from the positive terminal (the one without the spring) to pad 4. Solder the negative terminal (the one with the spring) directly to the PC board using a short piece of wire. Use hot glue or some other glue to secure the battery holder to the PC

Step 1 – Wound the Toroid

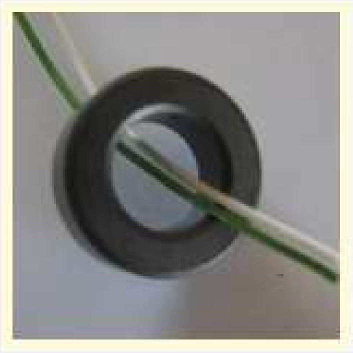
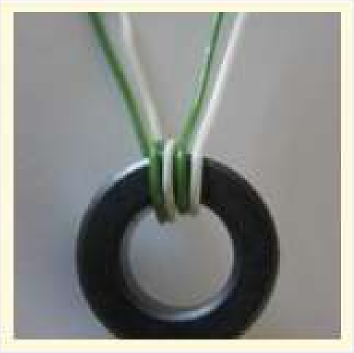
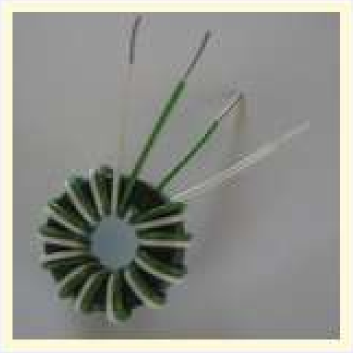

Take the two strands of plastic insulated wire and hold them together. Start off by sticking then through the middle of the toroid until you have 2 cm of wire left to make a connection with later. Keep the two wires together and wrap them around the toroid until you have covered the whole toroid. If you have wire left you can continue until you have about 2 cm of wire free on either side. Remove about a 0.5 cm of the plastic insulation on all four of the wires as in the photo. You have now wound your toroid and we are ready to continue.

Step 2 – Join the two coils of Toroid

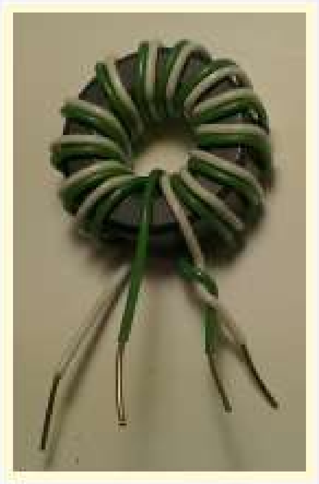

The next step is to join the two coils or windings of the toroid. You may have noticed the two dots next to the toroid windings in the circuit diagram. These two dots indicate the polarity of the transformer and as you can see the two windings are opposite to each other. To make sure that the polarity of our toroid windings are correct we must do the following:

- If we look at the toroid we noticed that we have two wires of different colours sticking out at both ends of the toroid.
- Take one wire of a specific colour (green in the example) on the one side of the toroid and one wire of the other colour (white in the example) on the other side of the toroid and connect (or solder) them together. These two wires forms the top pole of the toroid as in the circuit diagram and is connected to the positive terminal of the battery.

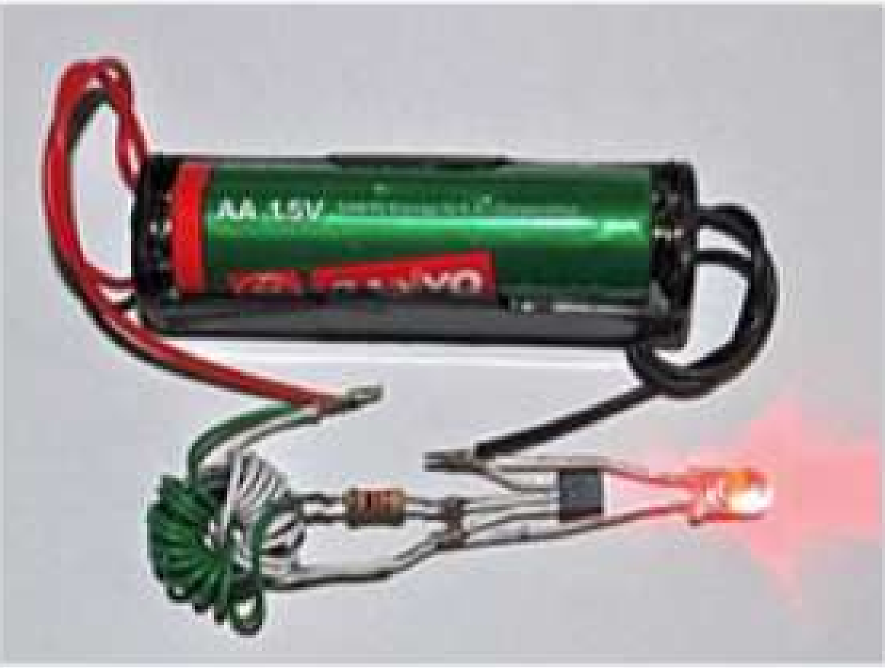

A conventional joule thief, shows components and how they are connected. This example uses a red LED. A ferrite toroid is wound to form a coil with primary (white) and feedback (green) windings. A 2N2222A transistor and 1000 ohm resistor are used.

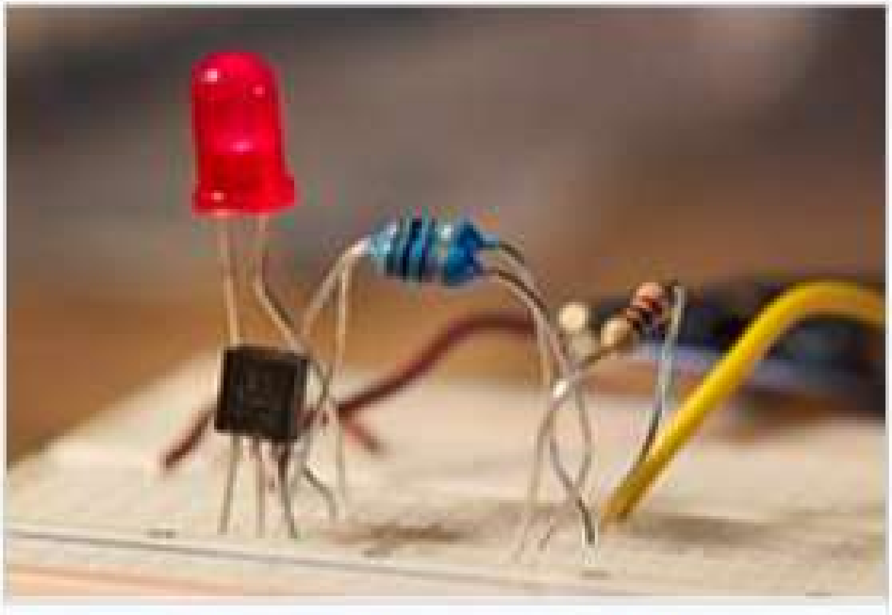

A joule thief with two axial inductors replacing the ferrite toroid, shown on a solderless breadboard

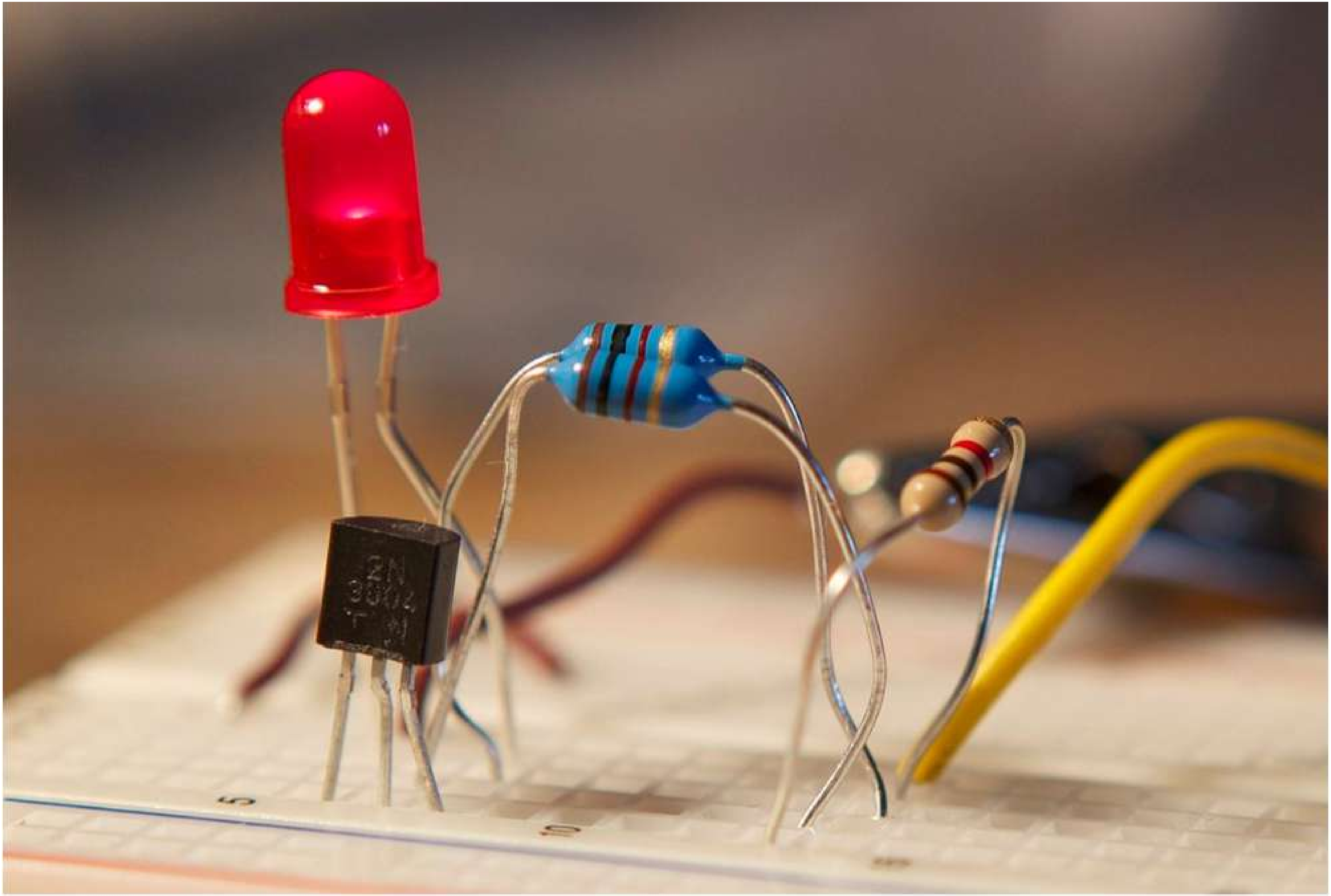
2N
3904
5
10

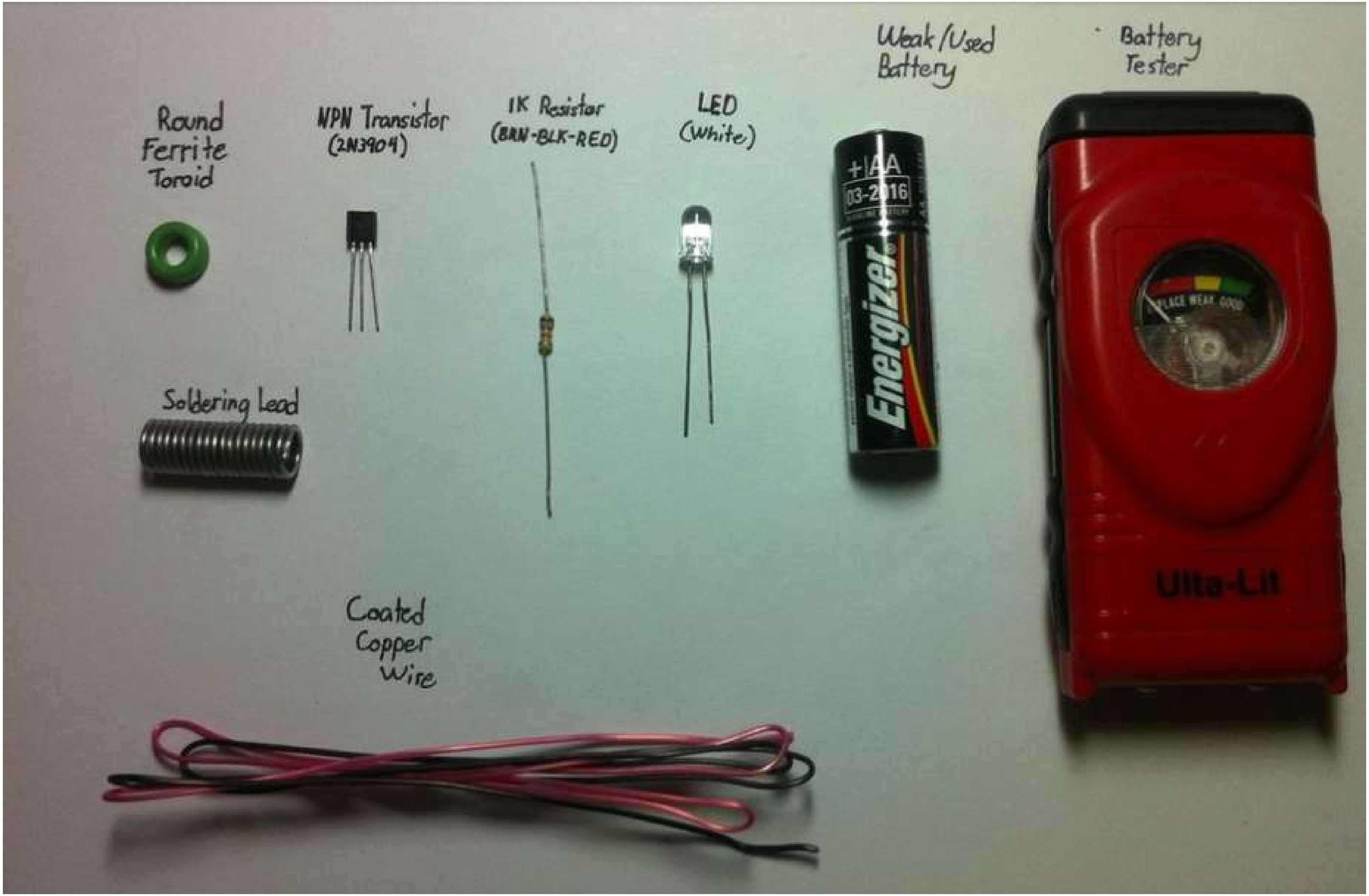
Round Ferrite Toroid
NPN Transistor (2N3904)
1K Resistor (BRN-BLK-RED)
LED (White)
Weak/Used Battery
Battery Tester
+AA
03-2016
Energizer
PLACE WEAK GOOD
Soldering Lead
Ulta-Lit
Coated Copper Wire

Step 2: Schematic Diagrams

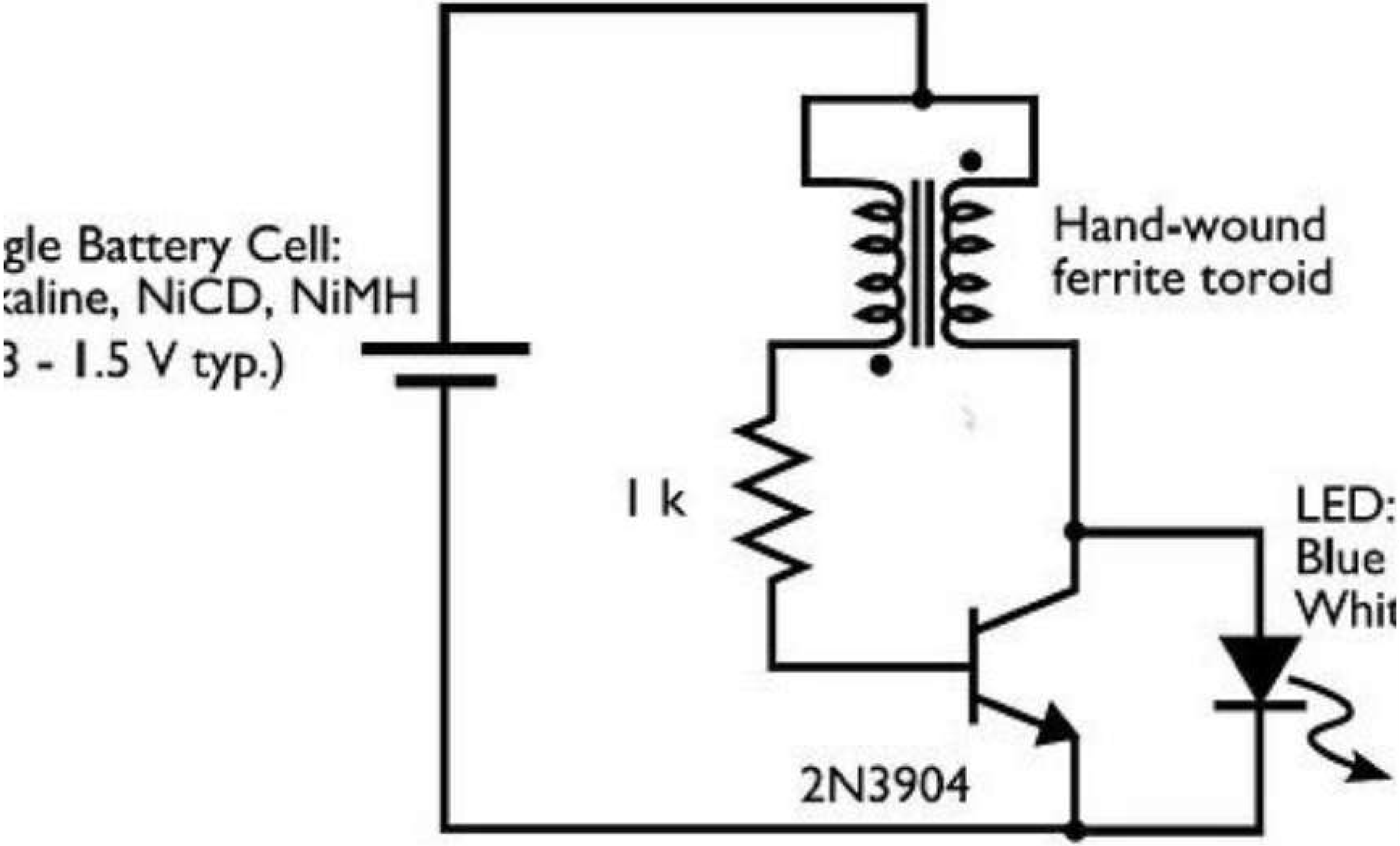
gle Battery Cell:
aline, NiCD, NiMH
3 - 1.5 V typ.)
Hand-wound
ferrite toroid
1 k
2N3904
LED:
Blue
Whit

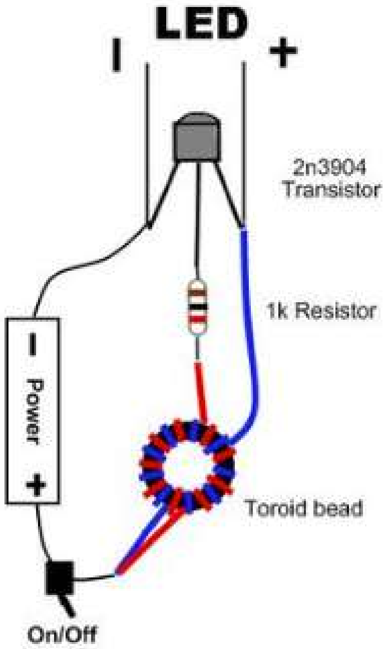
LED
I
+
2n3904
Transistor
1k Resistor
–
Power
+
Toroid bead
On/Off

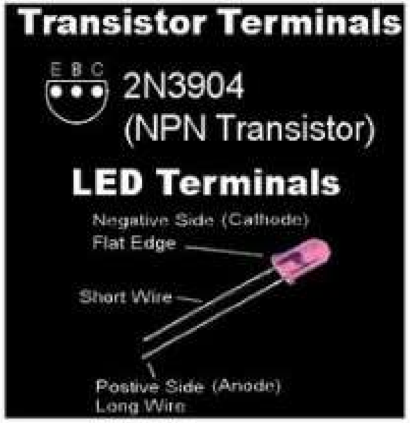
Transistor Terminals
E B C
2N3904
(NPN Transistor)
LED Terminals
Negative Side (Cathode)
Flat Edge
Short Wire
Postive Side (Anode)
Long Wire

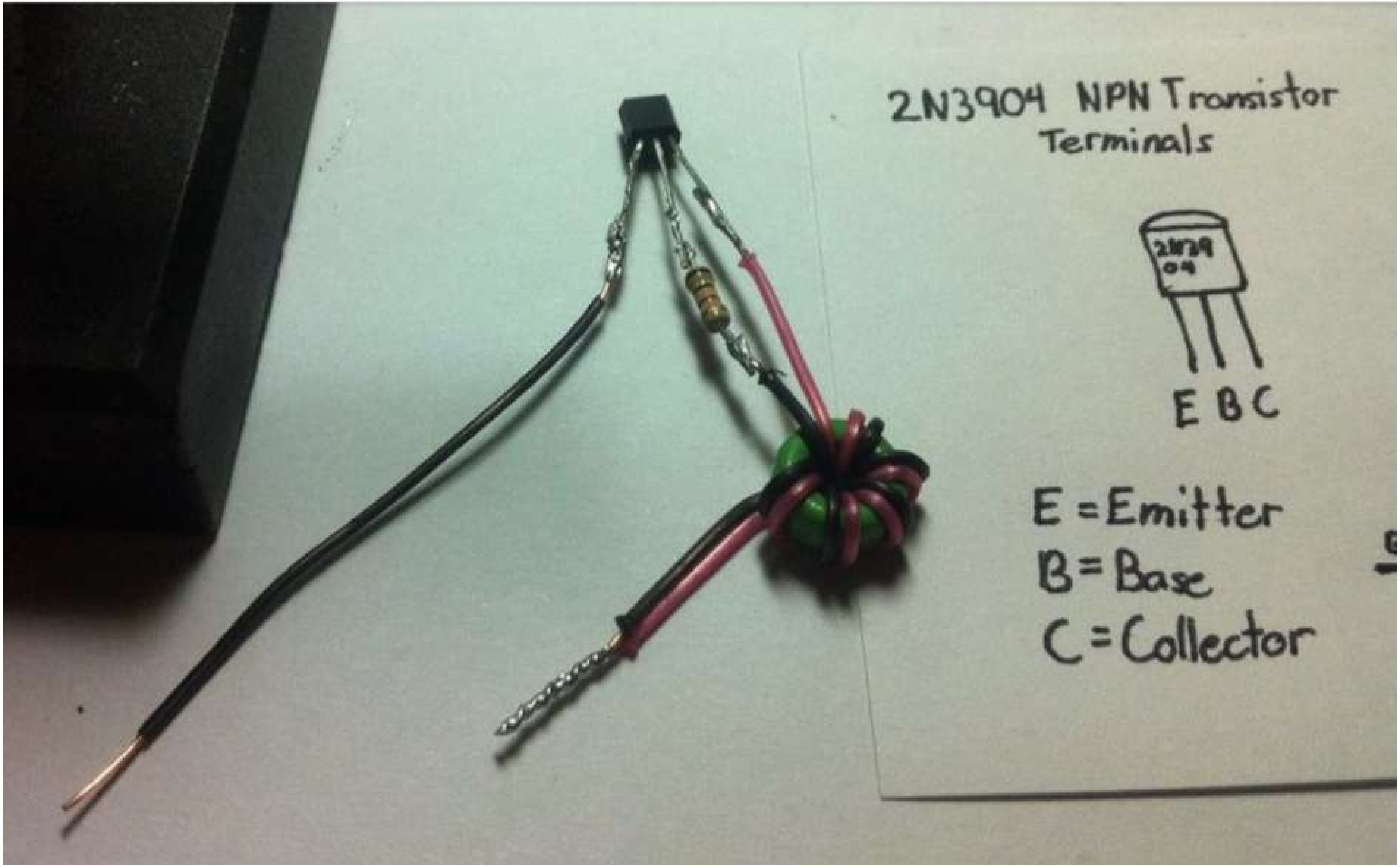
2N3904 NPN Transistor Terminals
2N39 04
E B C
E = Emitter
B = Base
C = Collector

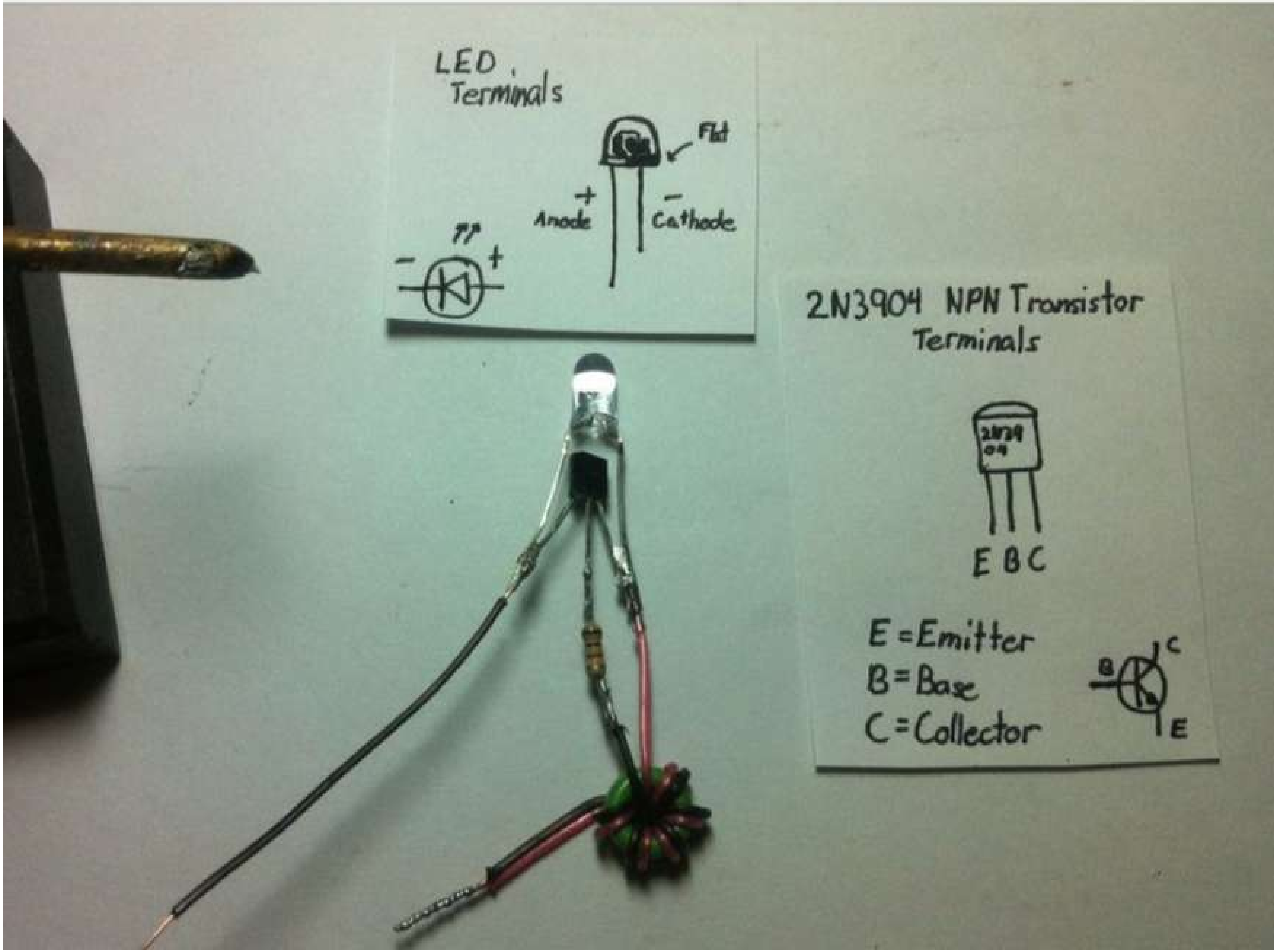
LED
Terminals
Flat
+
Anode
−
Cathode
−
+
2N3904 NPN Transistor
Terminals
2N39
04
E B C
E = Emitter
B = Base
C = Collector
B
C
E

These tiny T231212T toroids from Surplussales.com are high permeability so it doesn't take much wire to make a good Joule Thief coil. The core is less than a quarter inch diameter (here is a data sheet in .PDF.). Six inches of 30 AWG magnet wire trifilar wound, with two of the windings connected in parallel for the primary winding.

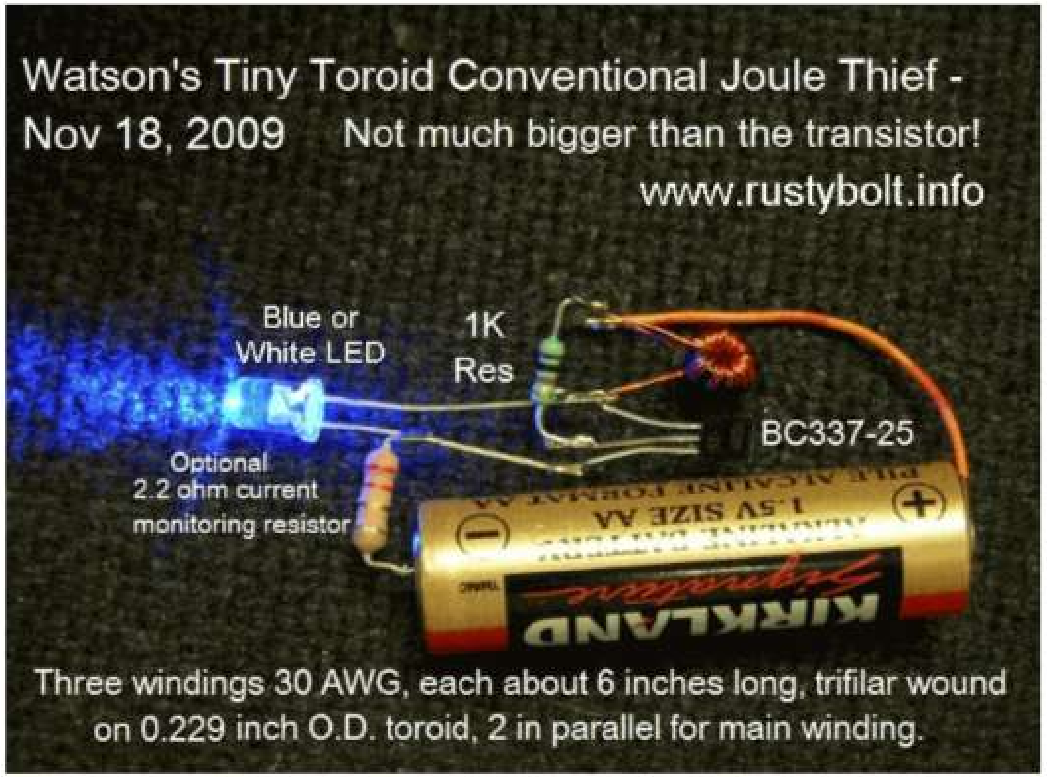

Watson's Tiny Toroid Conventional Joule Thief - Nov 18, 2009 Not much bigger than the transistor! www.rustybolt.info

Three windings 30 AWG, each about 6 inches long, trifilar wound on 0.229 inch O.D. toroid, 2 in parallel for main winding.

The transistor is a BC337-25. I put a 2.2 ohm resistor in series with the circuit to measure the battery current. It can be put there temporarily and removed when the LED current is what you want. How do you change the LED current? Change the 1k resistor to a different value. Higher values will reduce the LED current. With the BC337-25 and 1k resistor, the LED current should be close to 20 milliamps so it's probably not wise to go much below 1k. If you use another transistor, especially the pipsqueak 2N3904, you may have a hard time getting it to put out 20 milliamps.

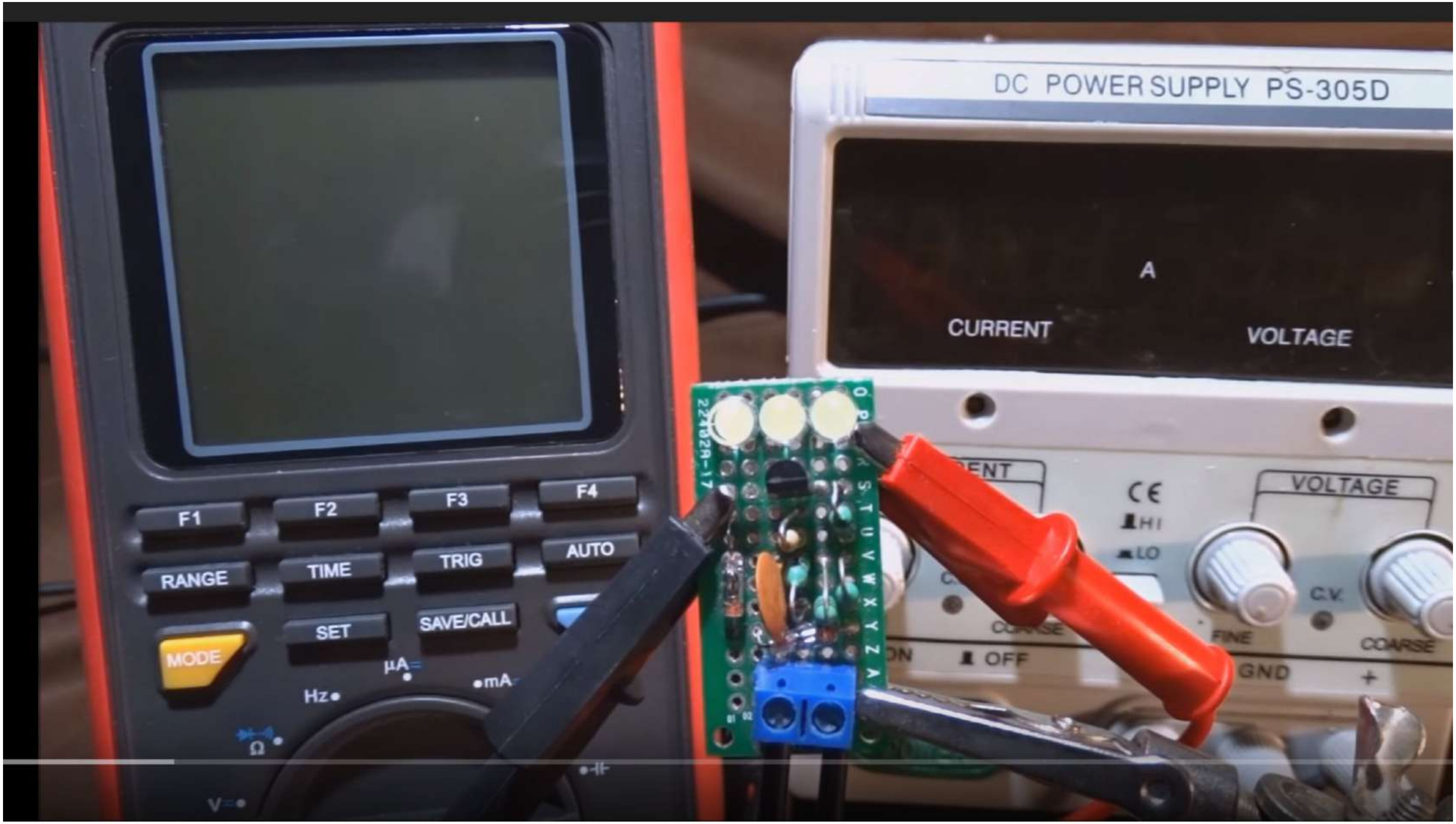
DC POWER SUPPLY PS-305D
A
CURRENT
VOLTAGE
F1
F2
F3
F4
RANGE
TIME
TRIG
AUTO
MODE
SET
SAVE/CALL
µA
Hz
mA
Ω
V
VOLTAGE
C.V.
FINE
COARSE
OFF
GND
+
CE
HI
LO
22402A-17
O P R S T U V W X Y Z A

B
103M

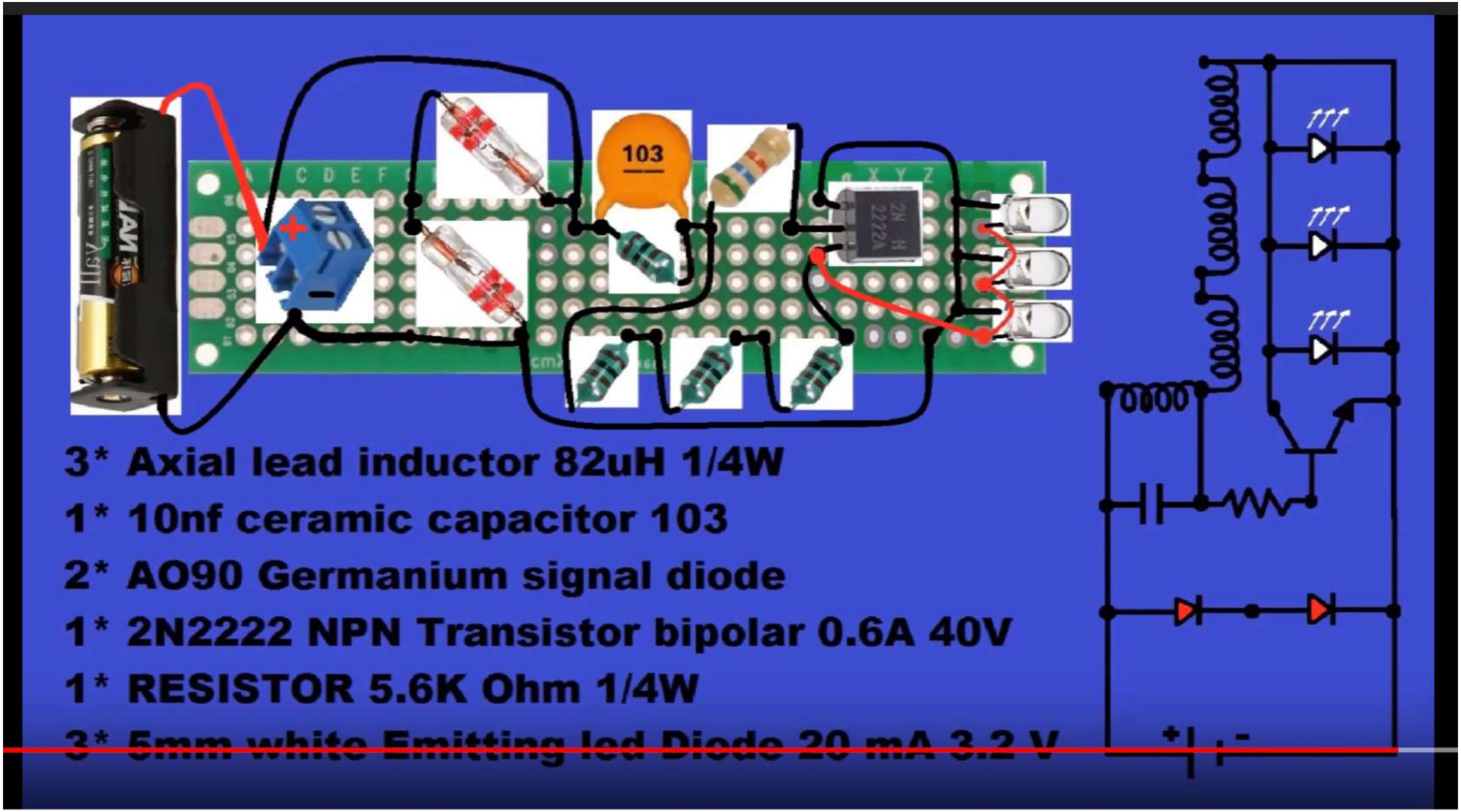
103
2N 2222A
+
-
A C D E F G
X Y Z
+
-

3* Axial lead inductor 82uH 1/4W

1* 10nf ceramic capacitor 103

2* AO90 Germanium signal diode

1* 2N2222 NPN Transistor bipolar 0.6A 40V

1* RESISTOR 5.6K Ohm 1/4W

3* 5mm white Emitting led Diode 20 mA 3.2 V

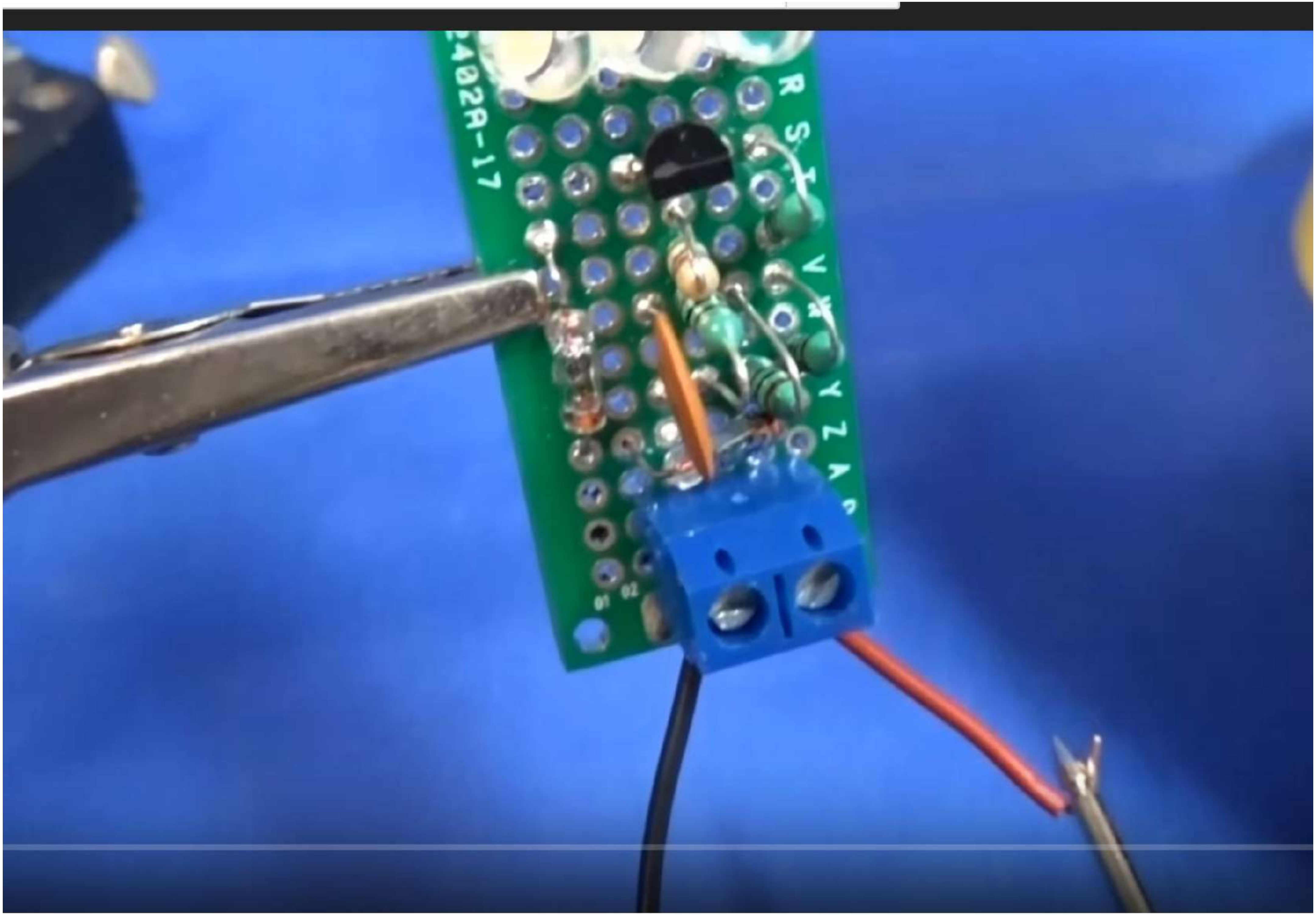
2402A-17
R
S
T
V
X
Y
Z
A
B
01
02

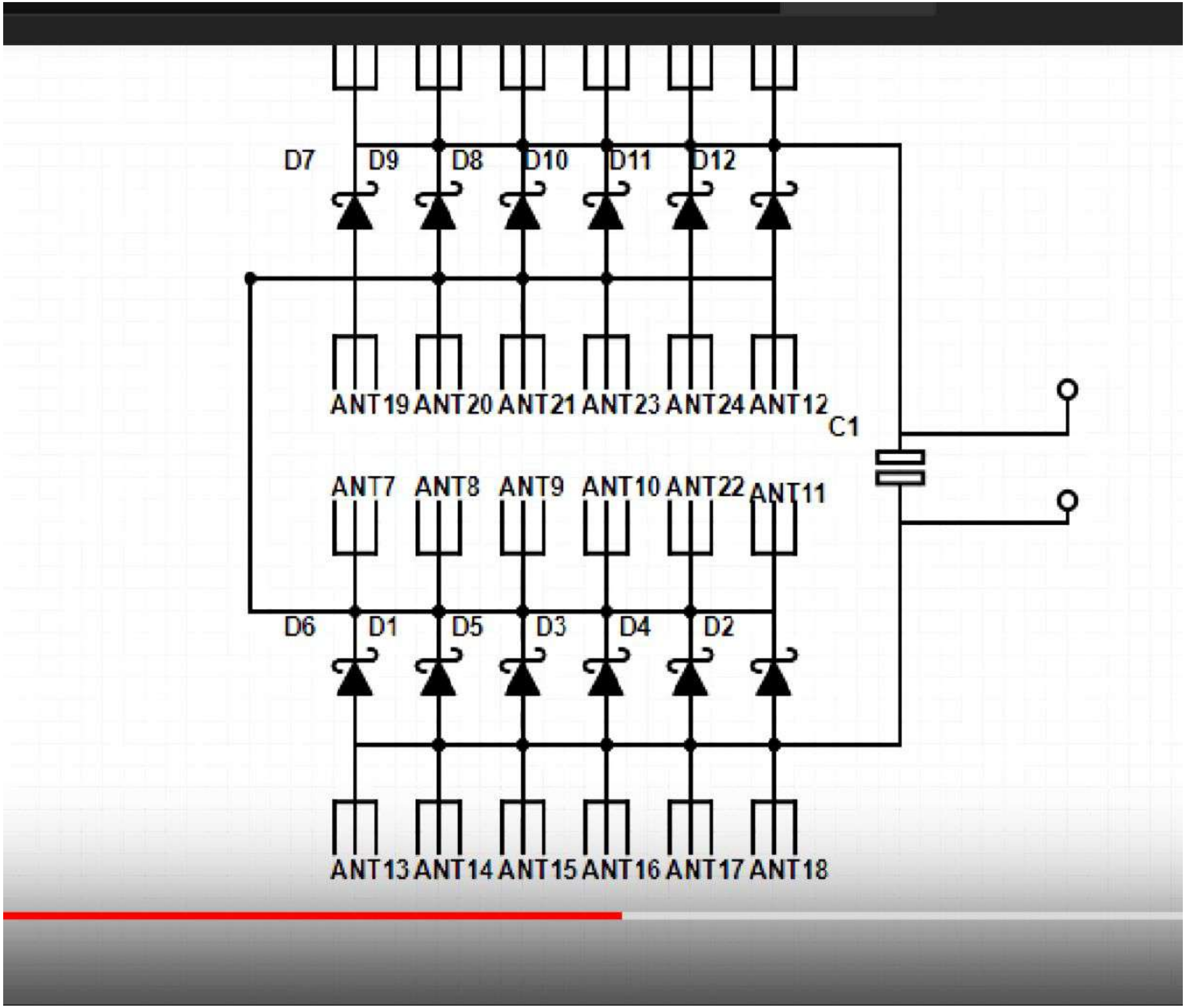
D7
D9
D8
D10
D11
D12
ANT19
ANT20
ANT21
ANT23
ANT24
ANT12
C1
ANT7
ANT8
ANT9
ANT10
ANT22
ANT11
D6
D1
D5
D3
D4
D2
ANT13
ANT14
ANT15
ANT16
ANT17
ANT18

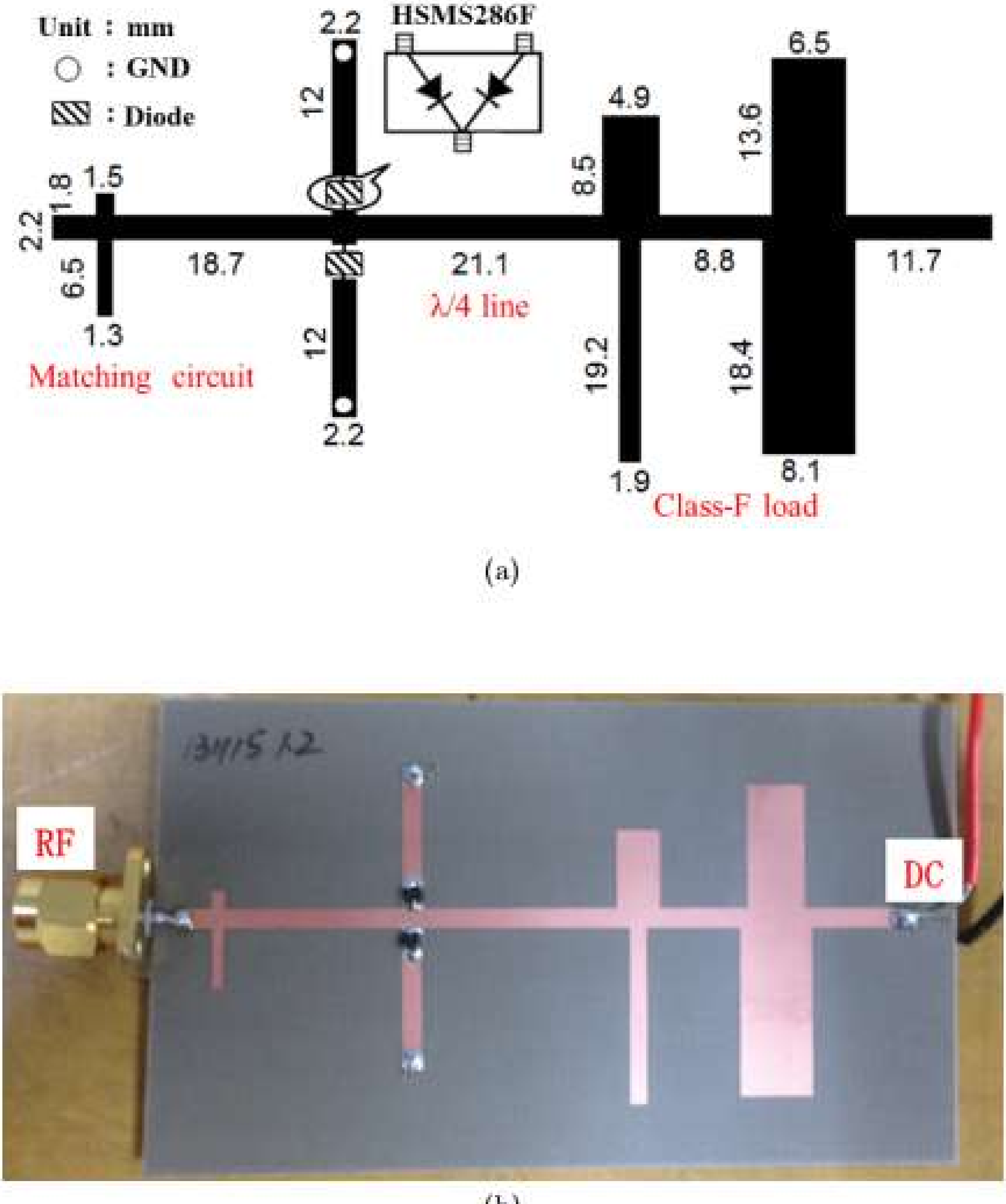
Unit : mm
○ : GND
▨ : Diode
HSMS286F
2.2
12
1.5
1.8
2.2
6.5
18.7
1.3
Matching circuit
12
2.2
21.1
λ/4 line
4.9
8.5
19.2
1.9
8.8
6.5
13.6
18.4
8.1
11.7
Class-F load
RF
DC
13115 12

(a)

(b)

Figure 4.1: Proposed positive output voltage rectifier. (a) Structure and size. (b) Photograph.

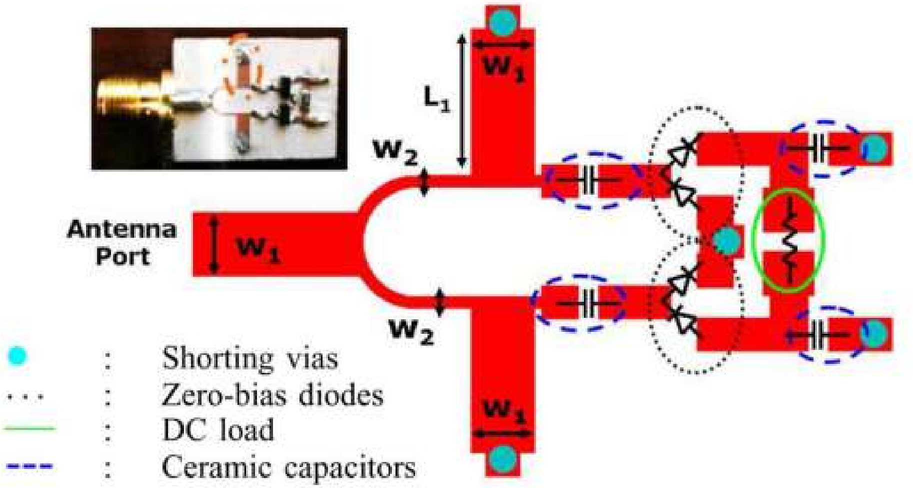
W1
L1
W2
Antenna Port
W1
W2
W1
Shorting vias
Zero-bias diodes
DC load
Ceramic capacitors

Fig. 3. Layout of the rectifier prototype, printed on RO3206. $w_1 = 72$ mil, $w_2 = 15$ mil, and $L_1 = 171$ mil. Fabricated sample is shown in top left, and the impedance matching stub is encircled.

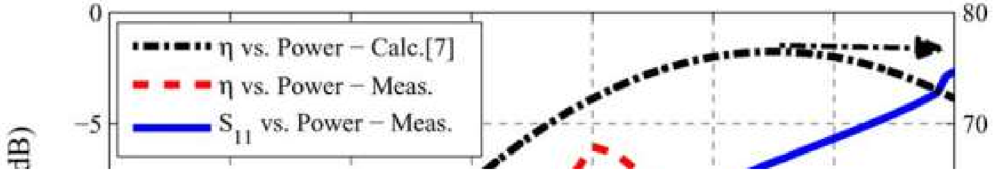

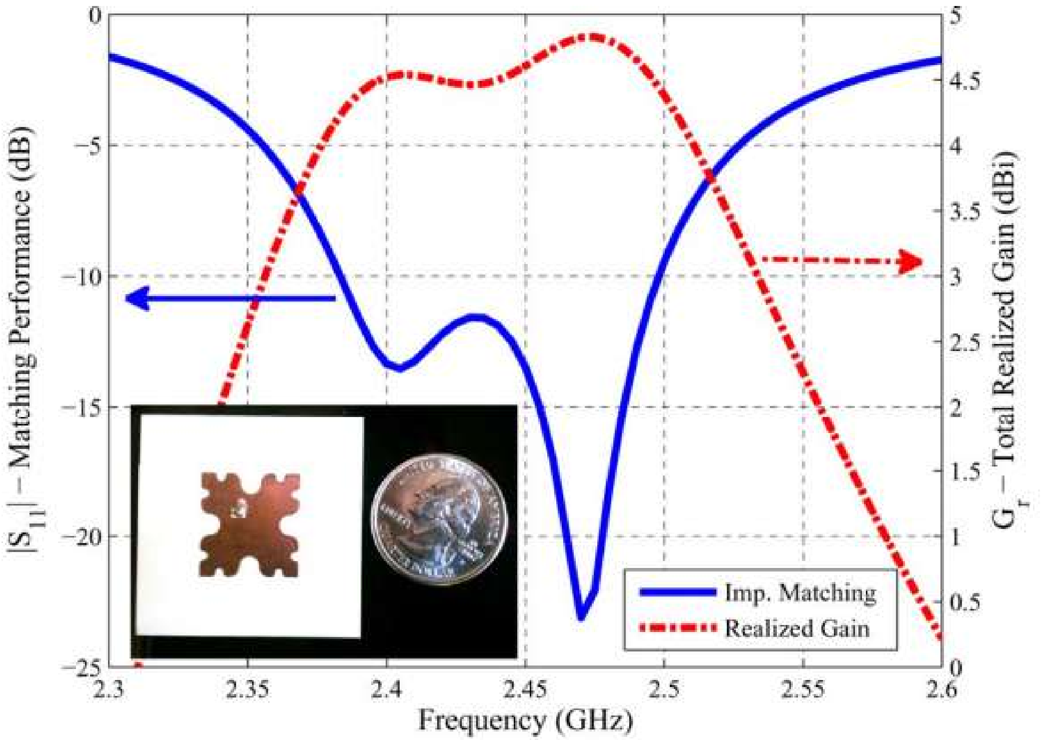

Fig. 5. Measured $|S_{11}|$ performance and total realized gain (at boresight) of the proposed antenna. Fabricated sample is shown in bottom left.

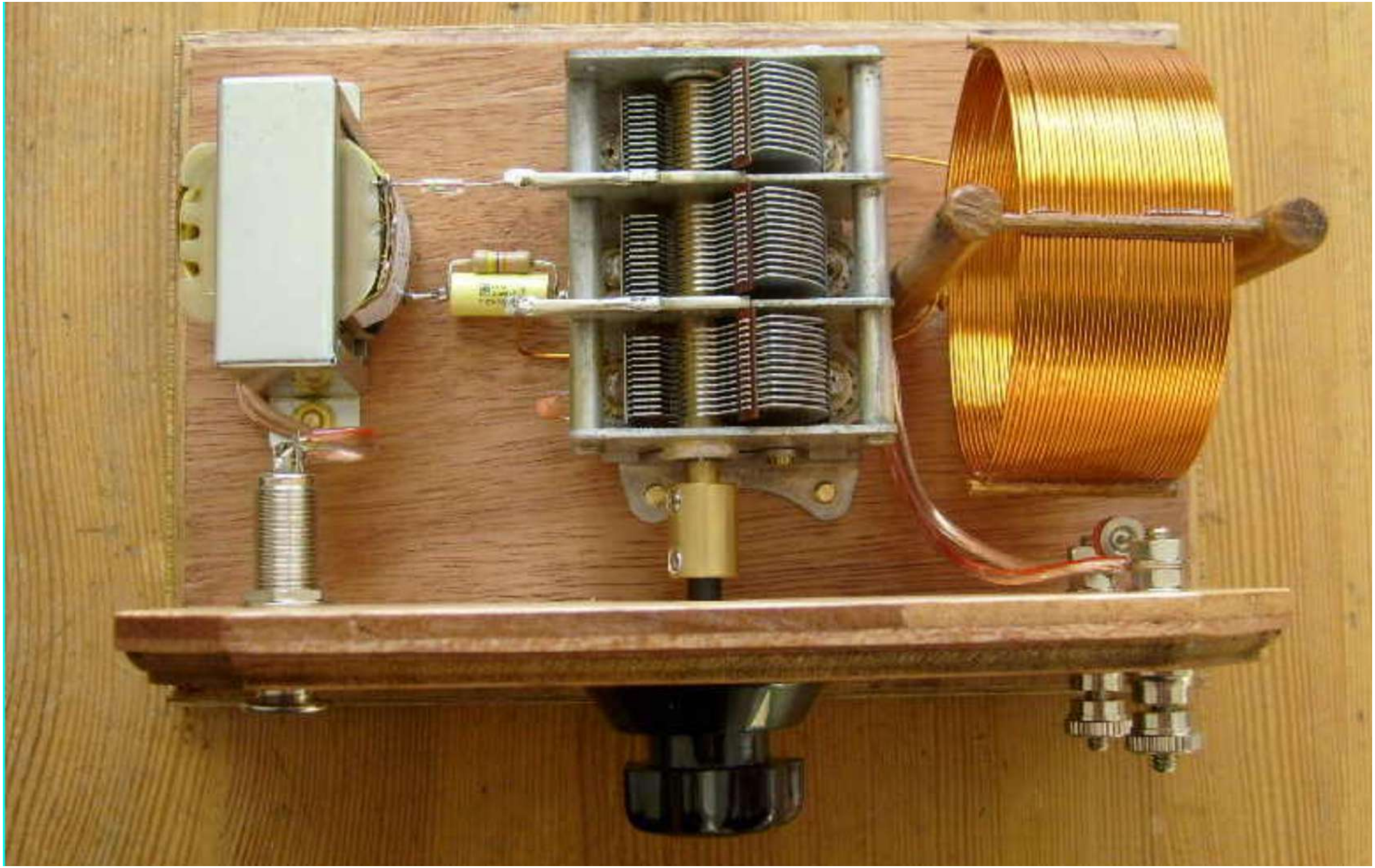

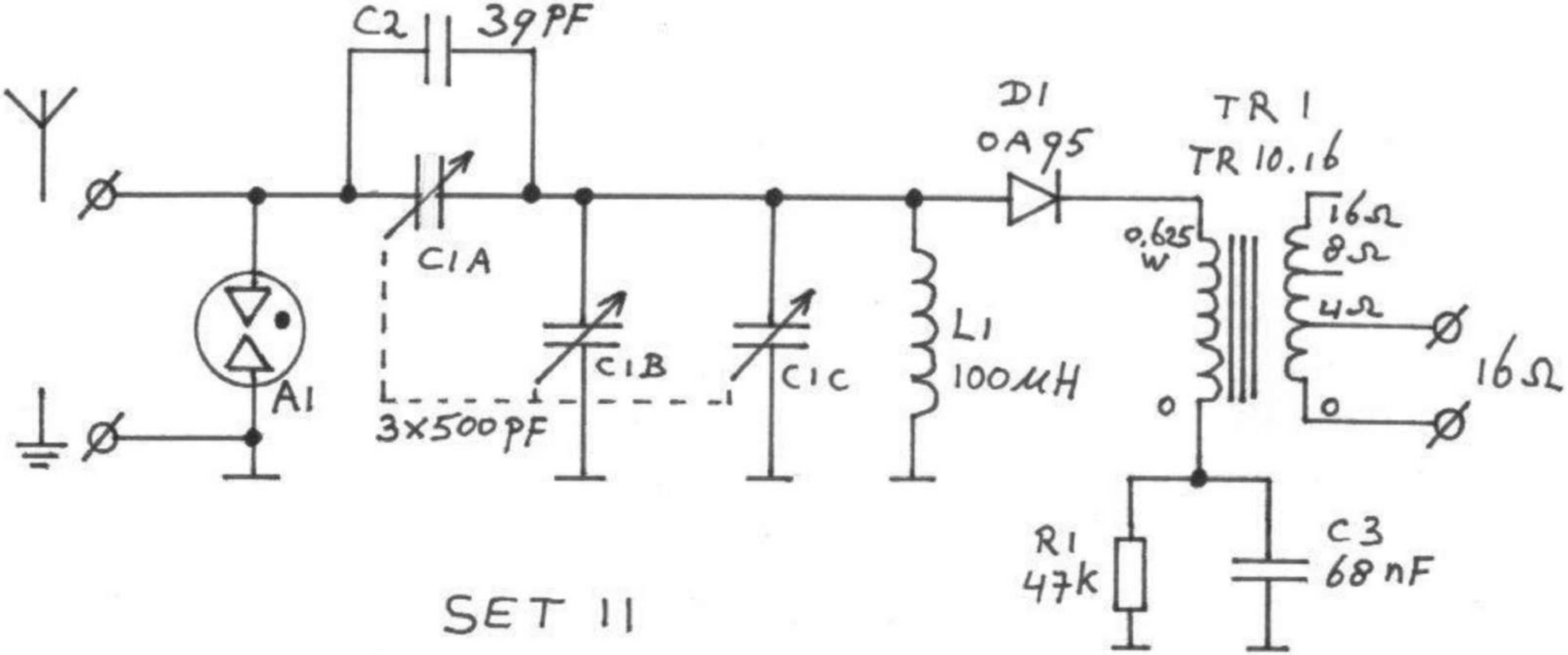
C2 39PF
D1
OA95
TR1
TR 10.16
0.625 W
16Ω
8Ω
4Ω
0
0
16Ω
C1A
C1B
C1C
L1
100µH
A1
3×500PF
R1
47k
C3
68nF

SET 11

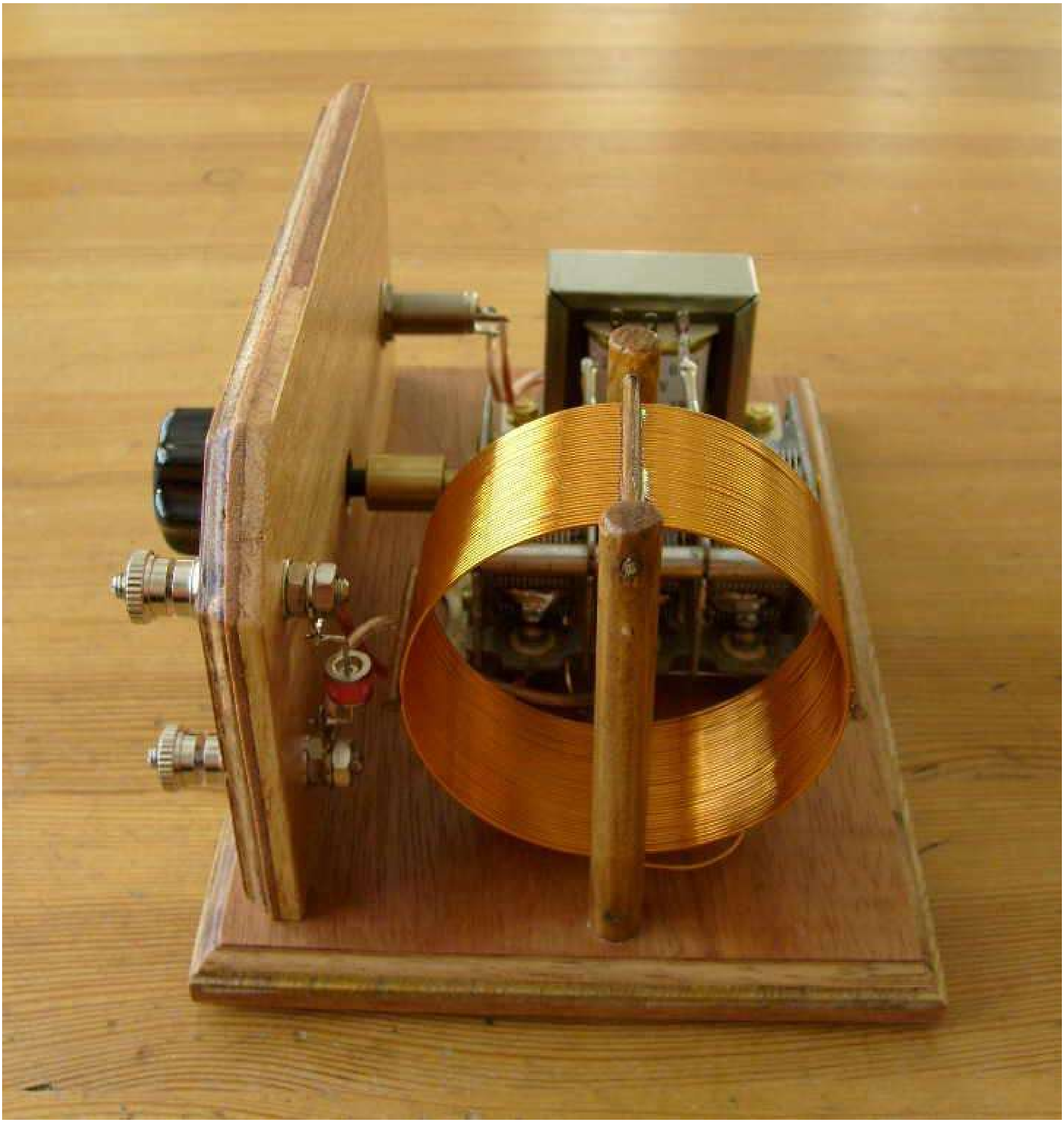

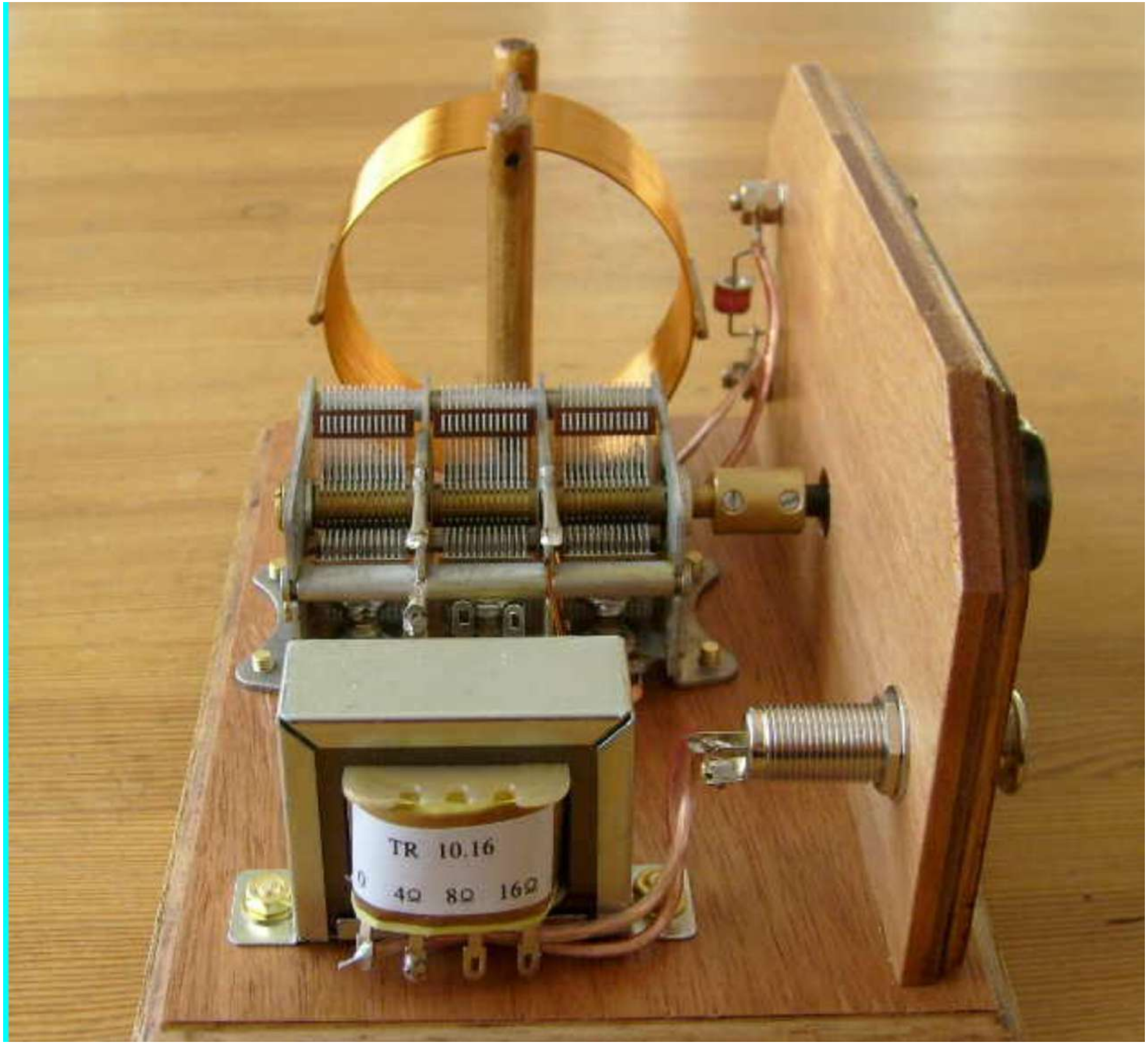
TR 10.16
0 4Ω 8Ω 16Ω

Preparation of the tuning capacitor

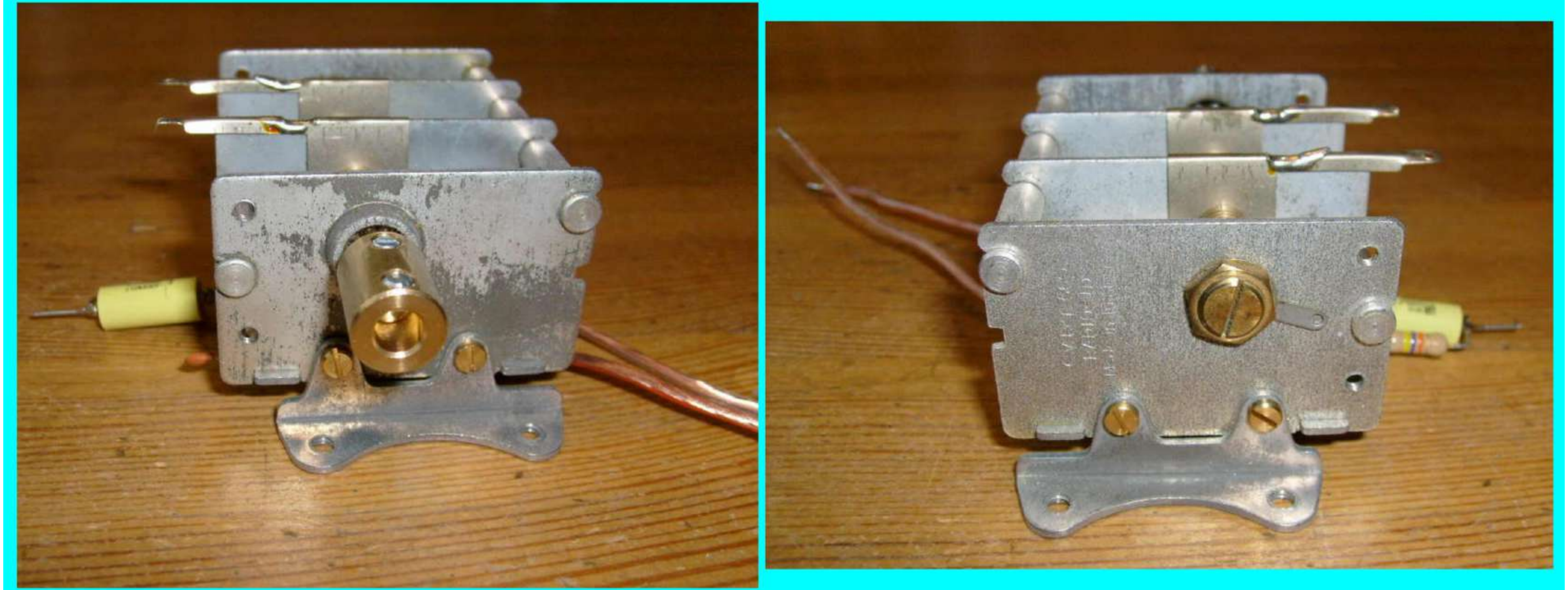

Mount the mounting supports on the tuning capacitor (they are delivered with the tuning capacitor).
Place the shaft coupler on the shaft of the tuning capacitor.

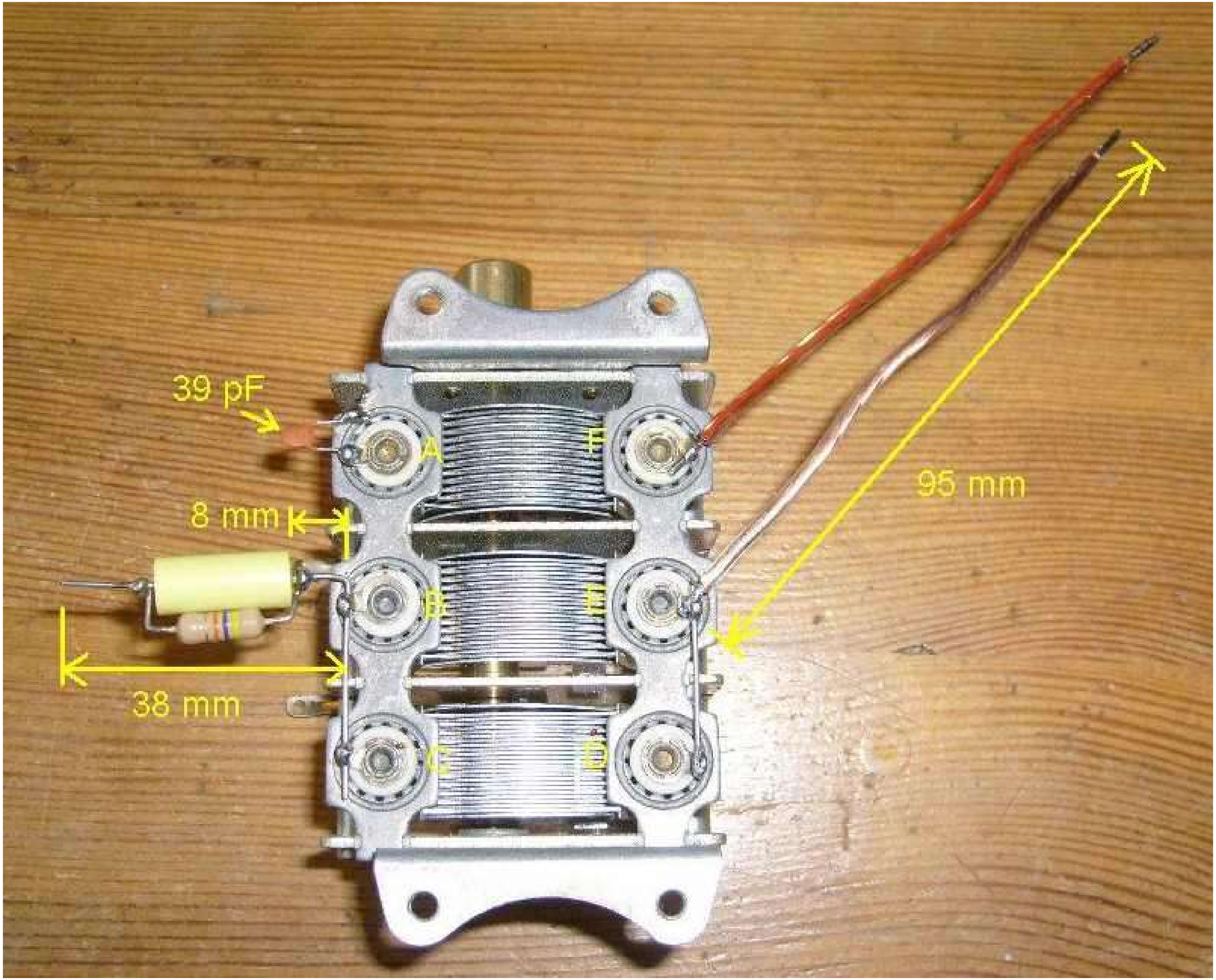
39 pF
A
F
95 mm
8 mm
B
E
38 mm
C
D

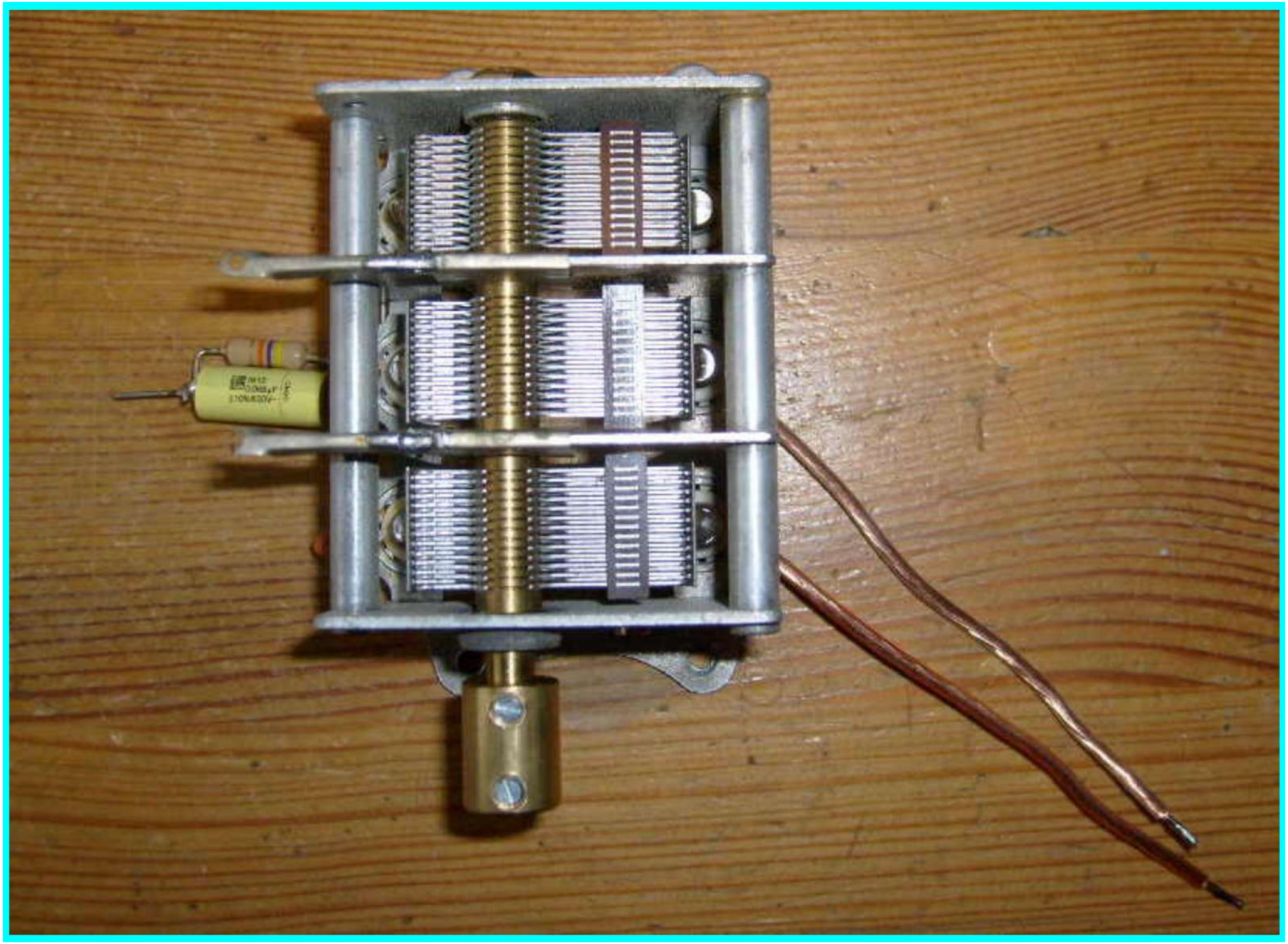

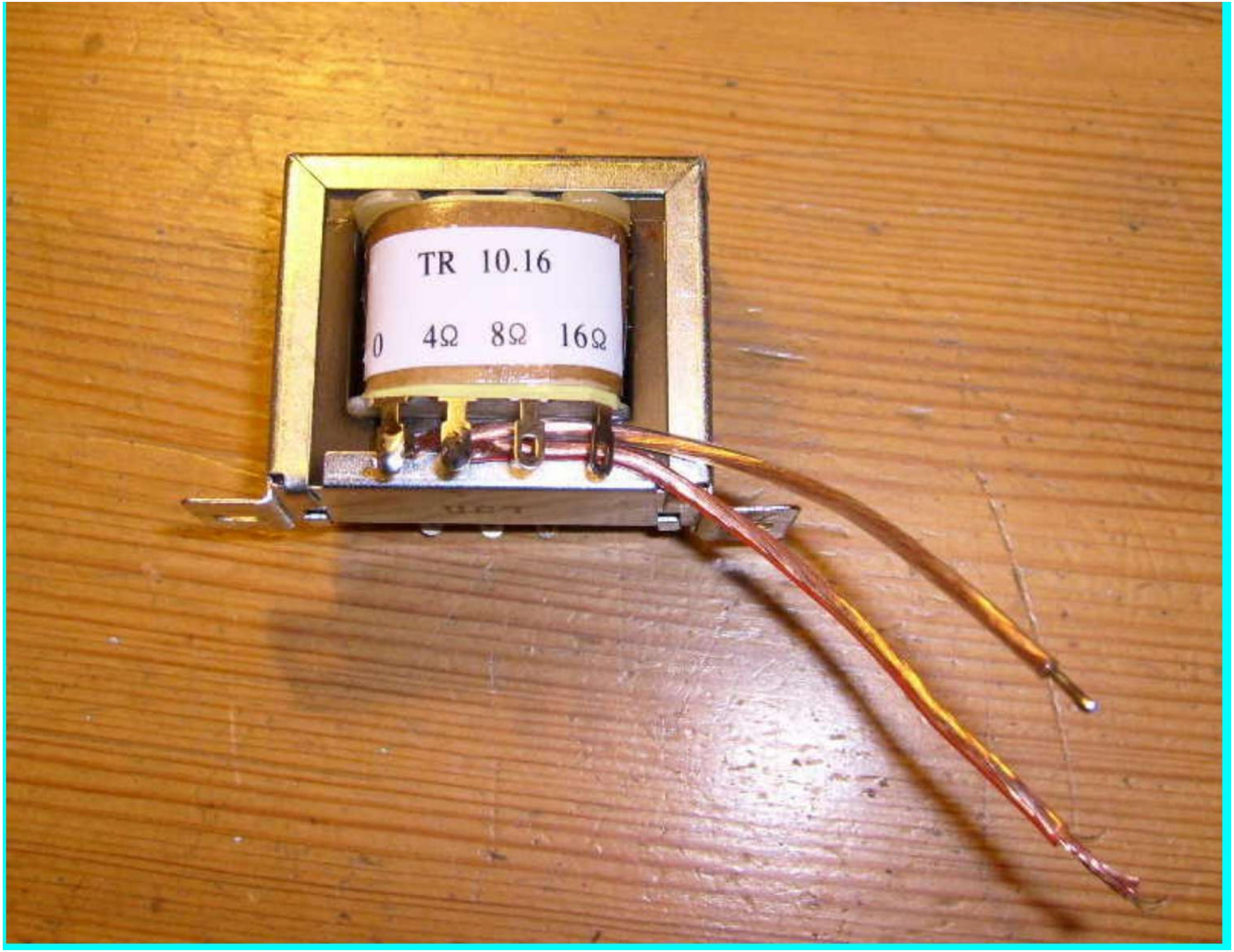
TR 10.16
0 4Ω 8Ω 16Ω

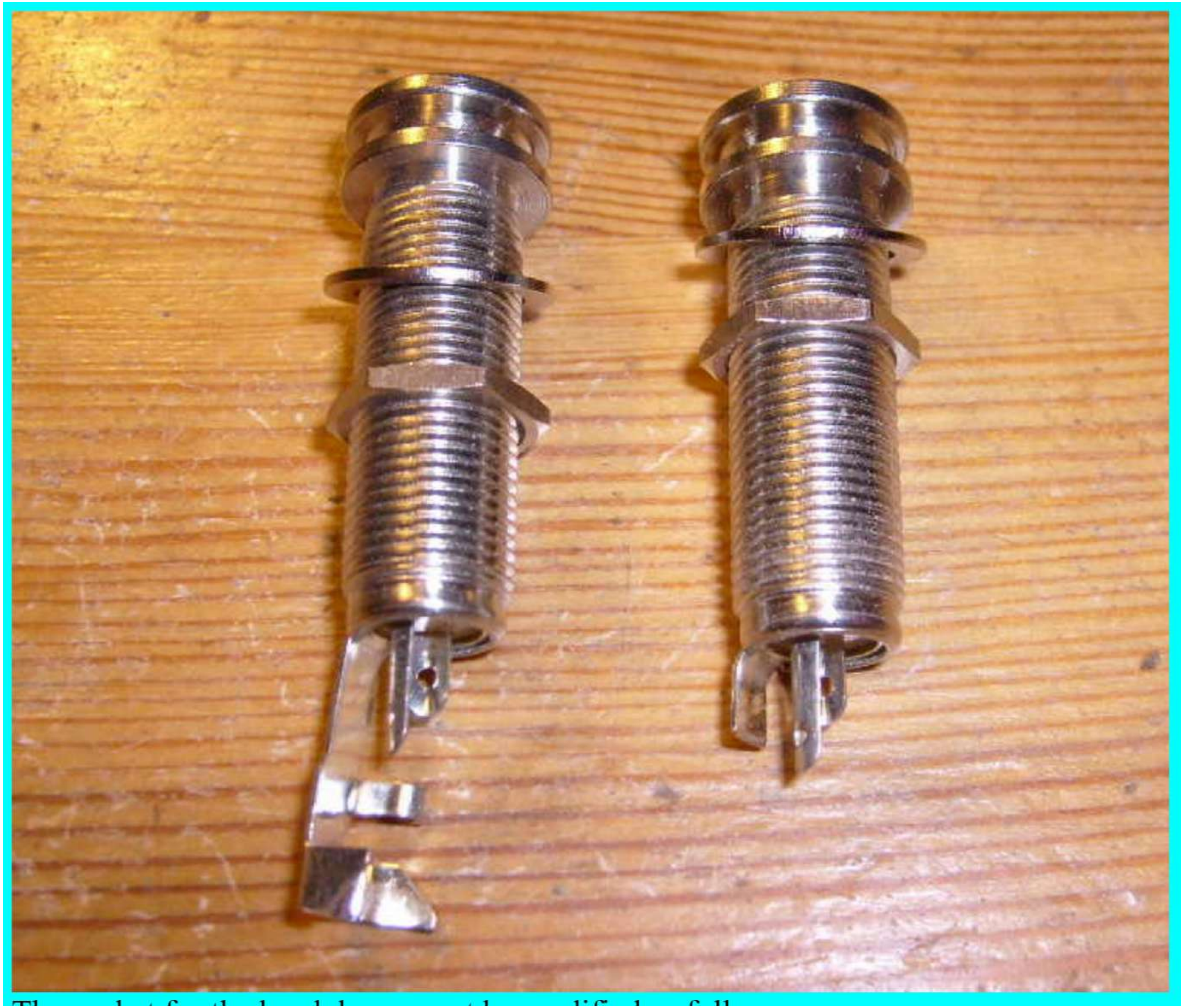

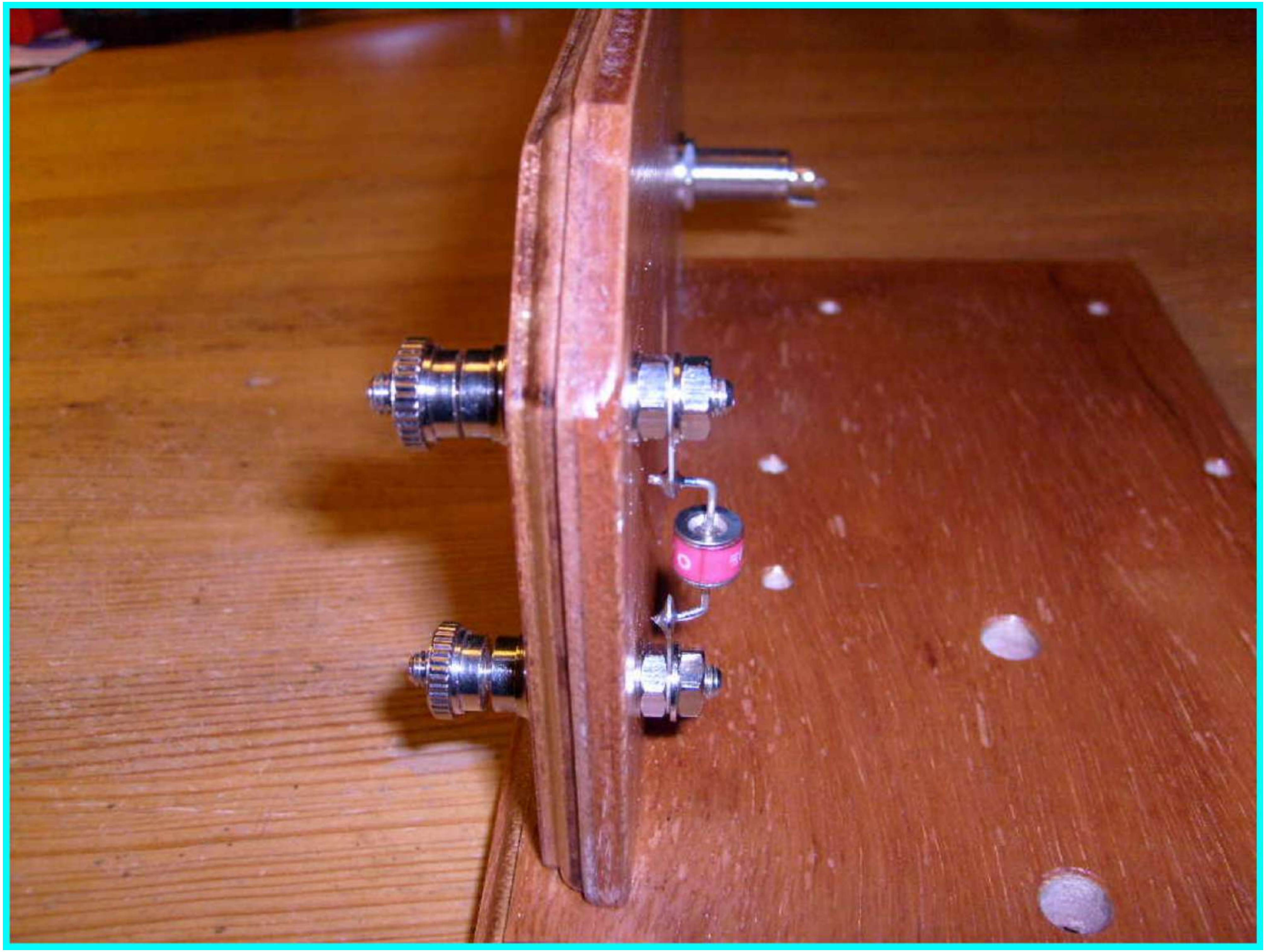

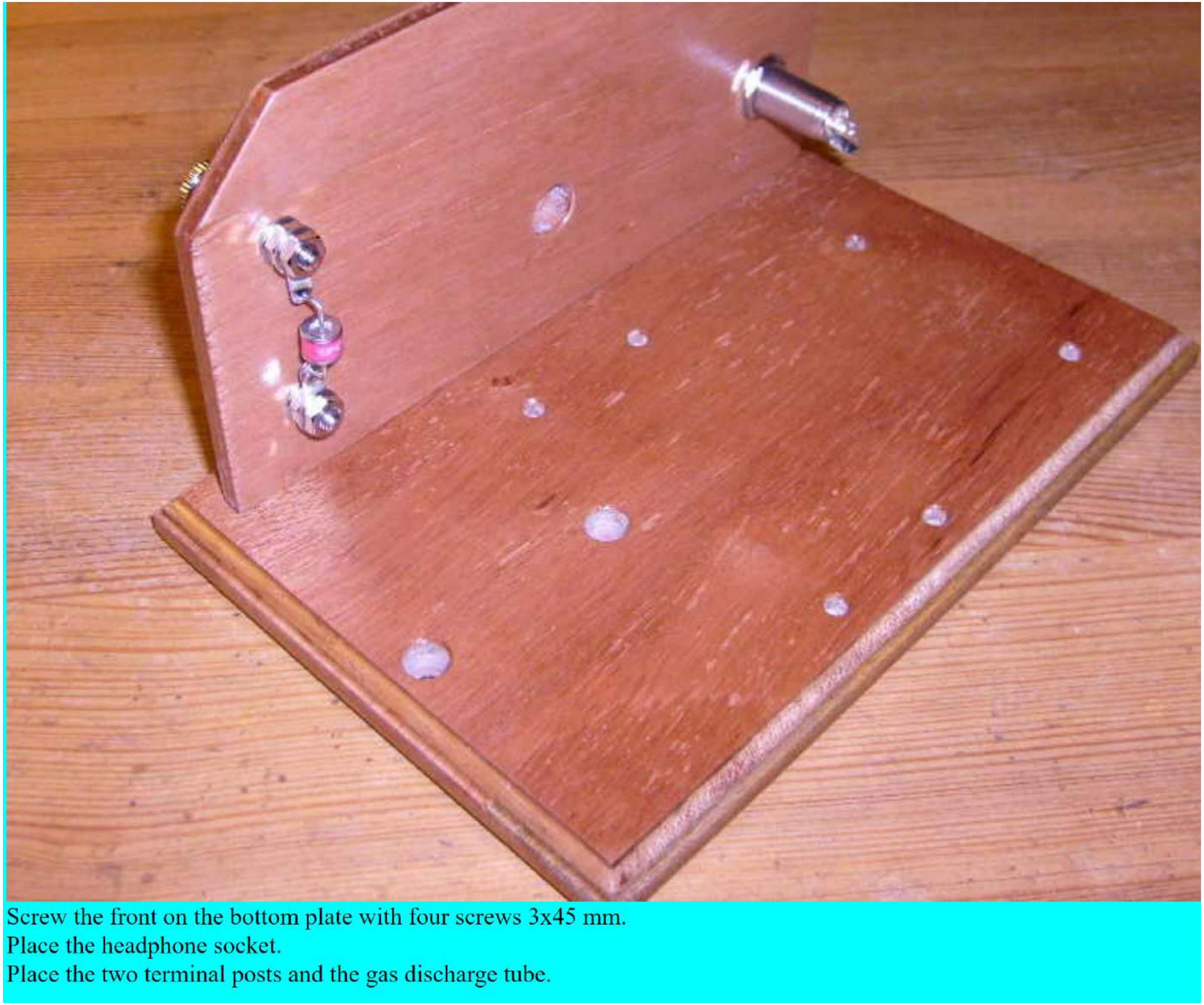

Screw the front on the bottom plate with four screws 3x45 mm.
Place the headphone socket.
Place the two terminal posts and the gas discharge tube.

Make a paper strip with 4 sections of 40x55 mm on it.
Do the strip around the tube, and fix it with a piece of tape.
Don't tape the strip onto the tube on places where later the coil is wound.

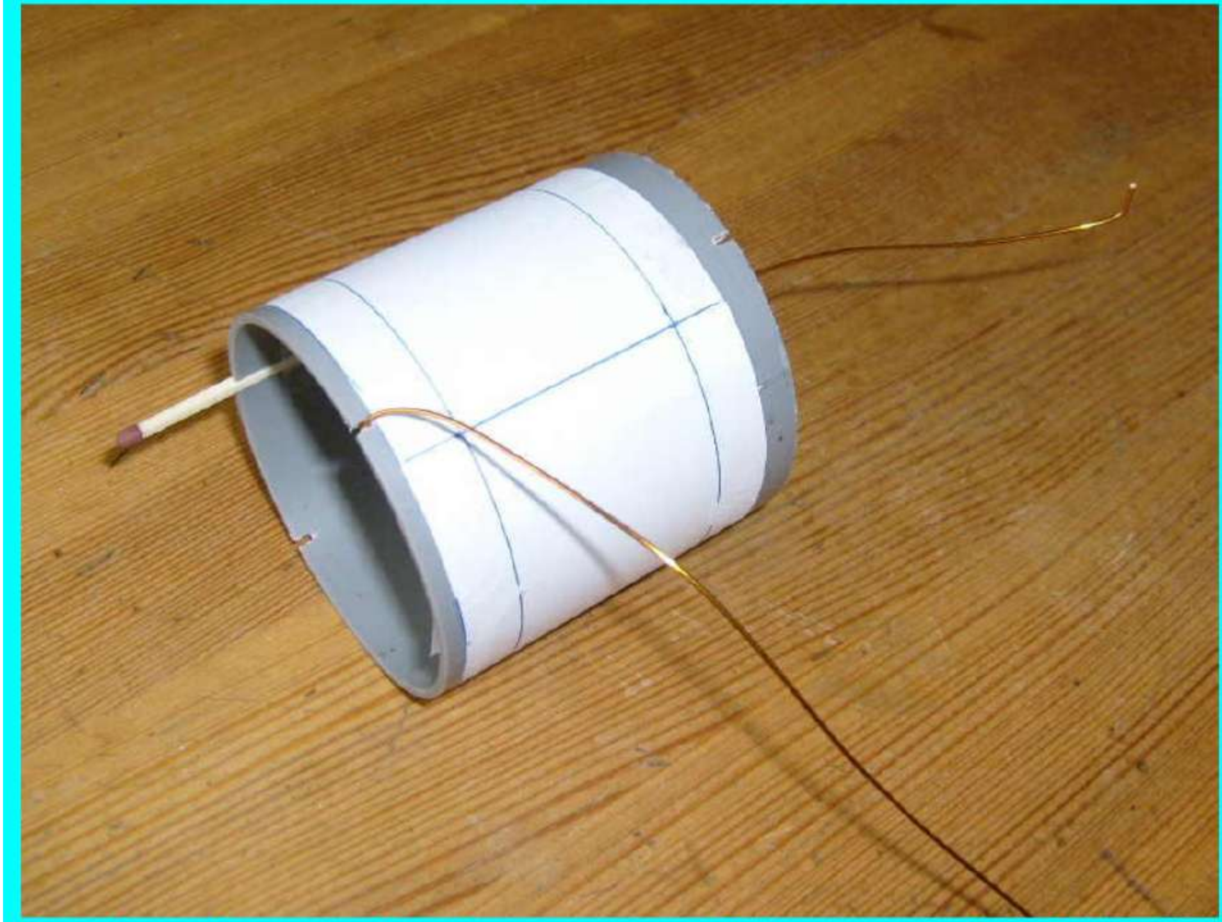

For the coil we need 8.6 meter enamelled copper wire with 0.8 mm thickness.
Hook the begin of the coilwire onto the slot in the tube.
Also fix the begin of the wire on the inside of the tube with some tape.

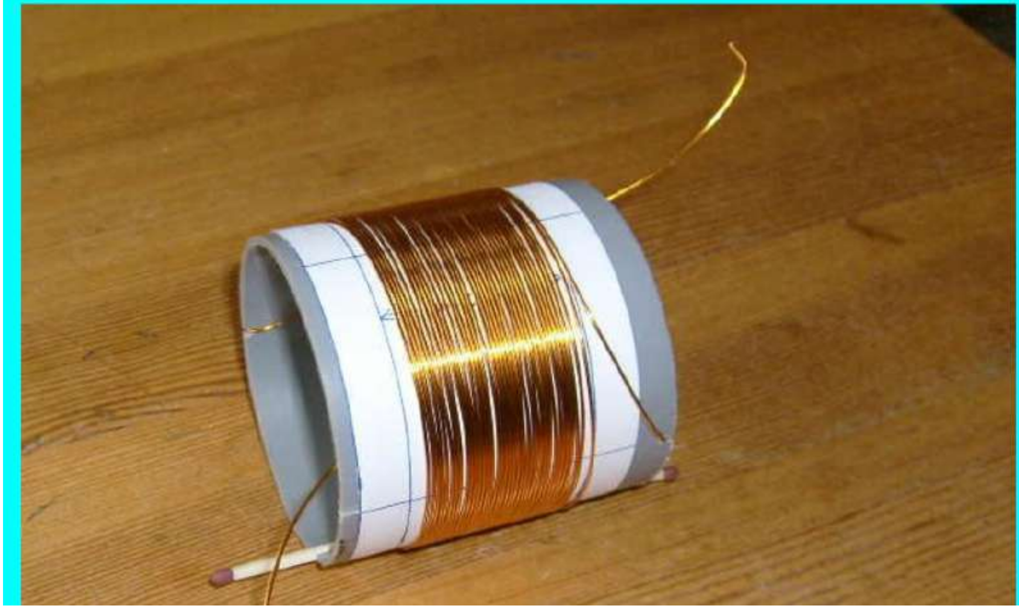

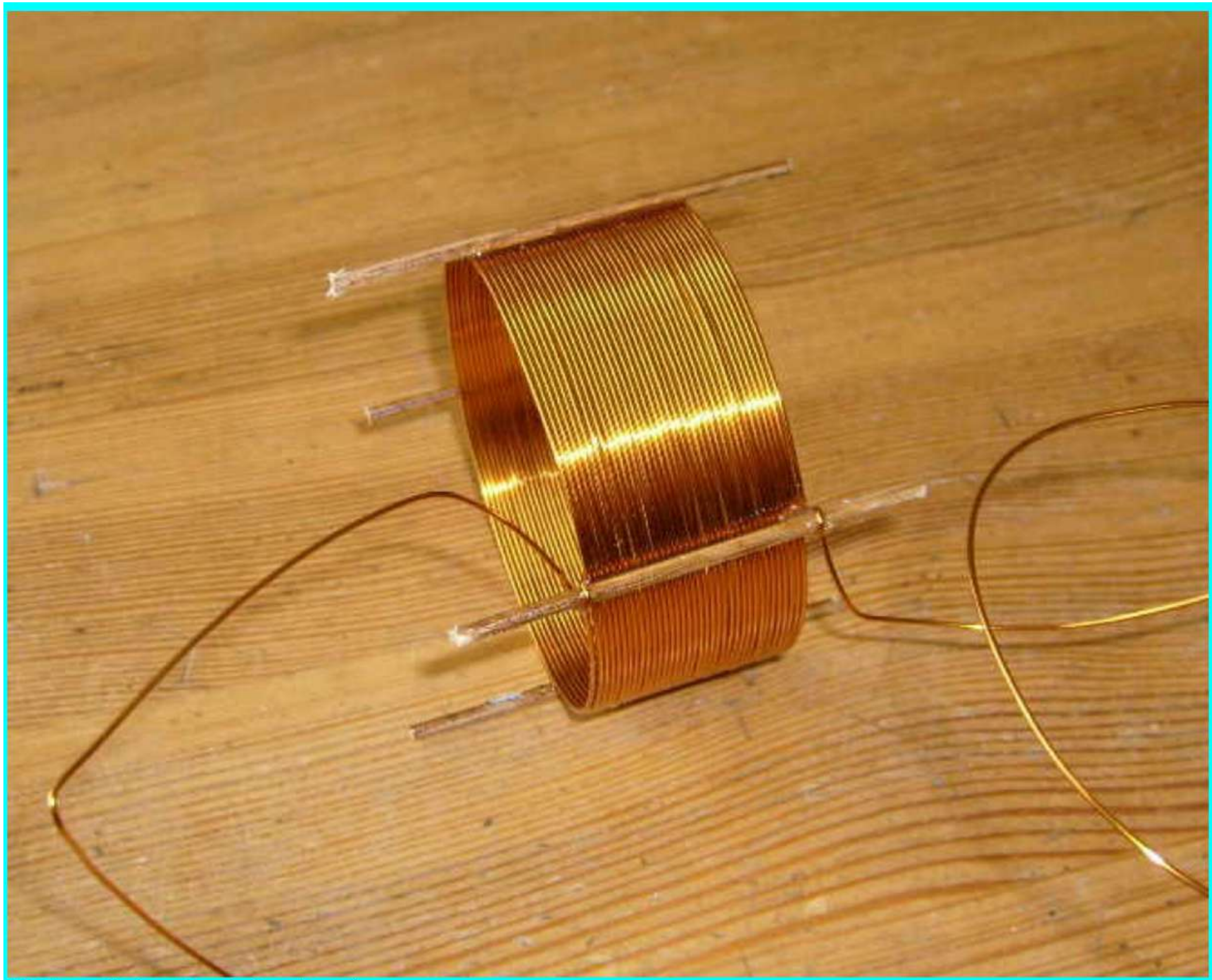

Wait until the glue is hardened.
Remove the rubber bands, the tape and the matches.
Shift the tube out of the coil, and remove the paper out of the coil.
Turn the both ends of the wire one time around the stick at the bottom side of the coil, and apply some extra glue here.

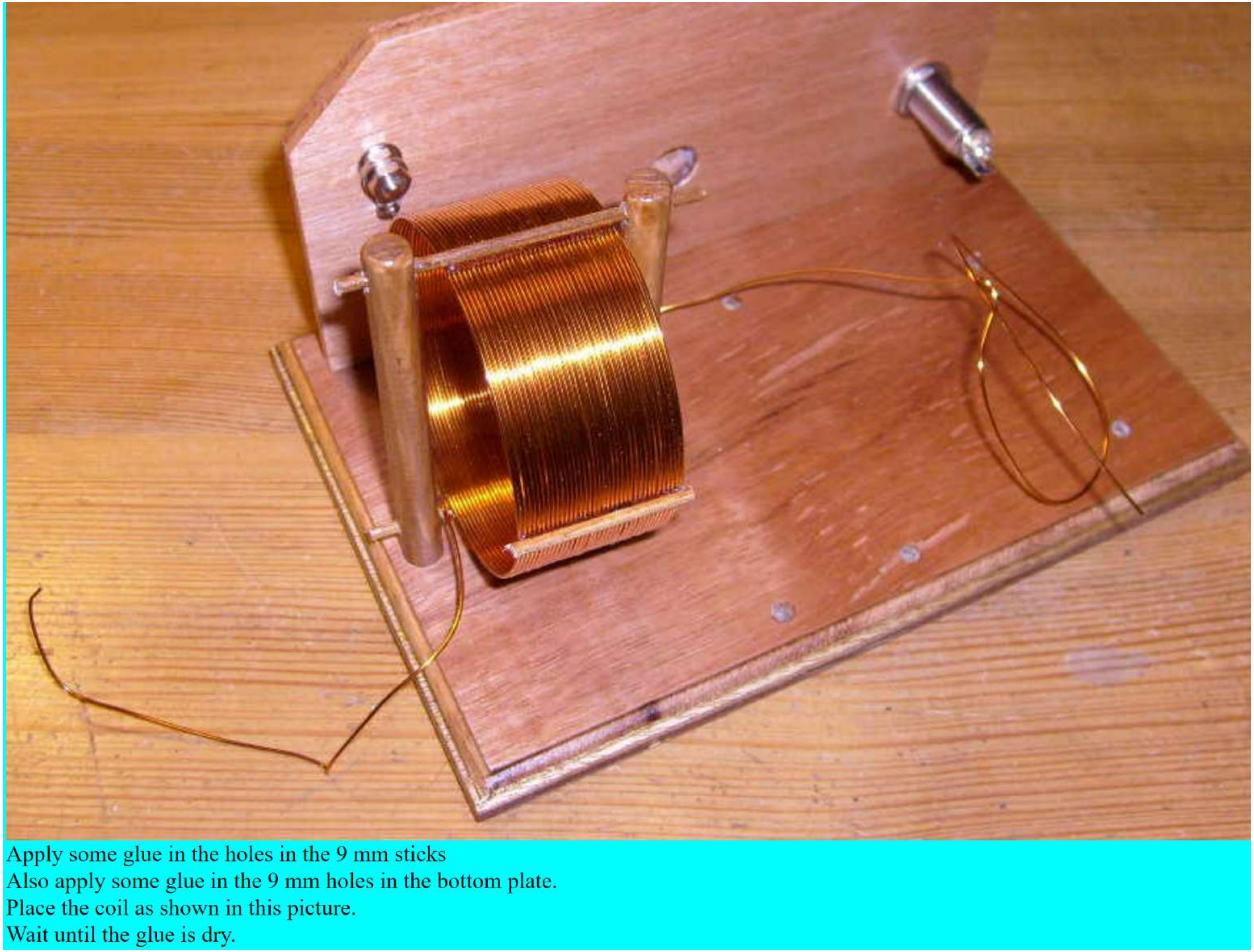

Apply some glue in the holes in the 9 mm sticks
Also apply some glue in the 9 mm holes in the bottom plate.
Place the coil as shown in this picture.
Wait until the glue is dry.

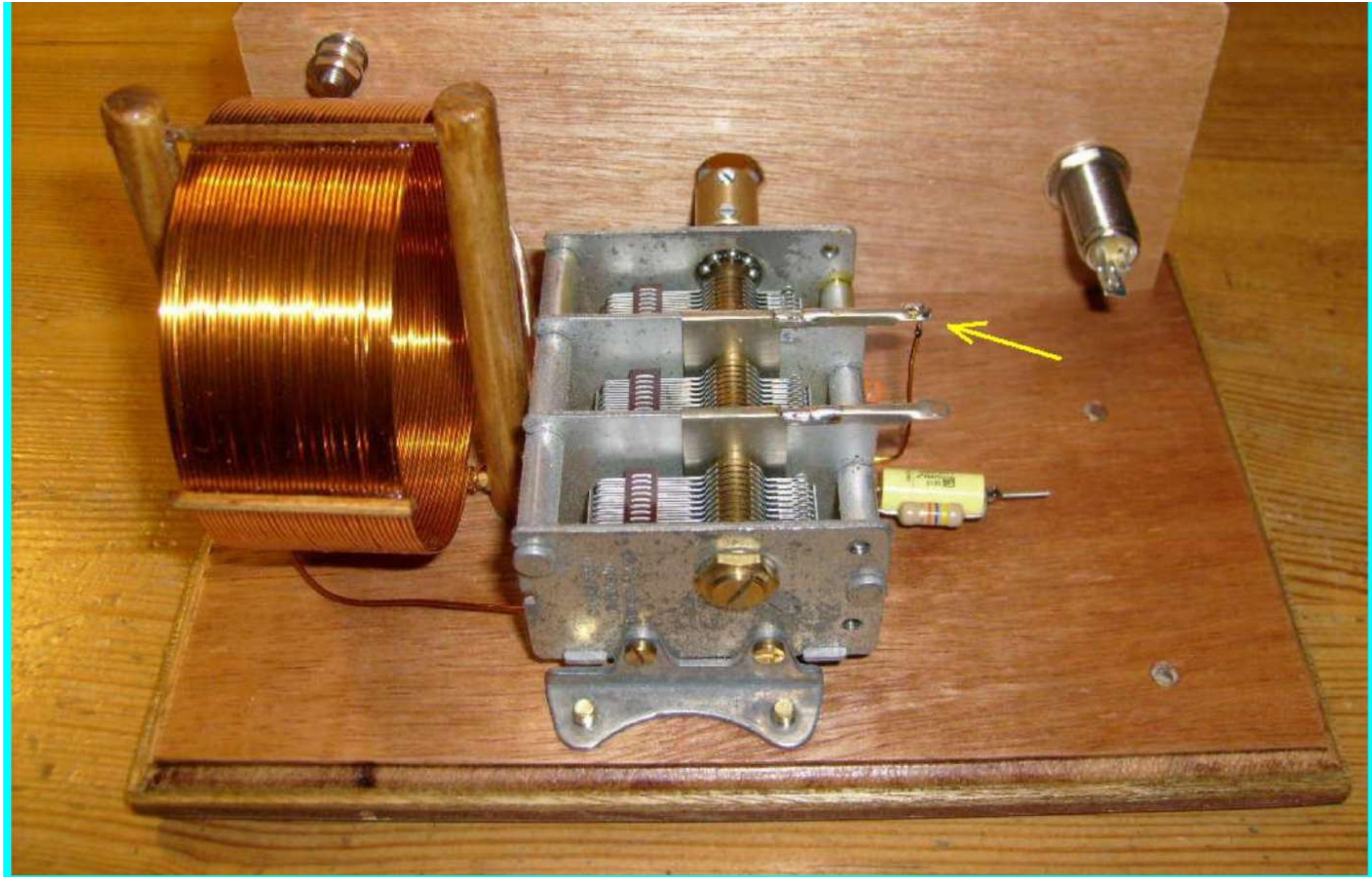

Place the tuning capacitor on the frame.
Fix it with four screws M4x16
One coilwire runs underneath the tuning capacitor, is then bend upwards and cut to length.
Solder this wire to the tuning capacitor (indicated with the arrow in the above picture).
Scrap of the lacquer from the wire, before soldering.

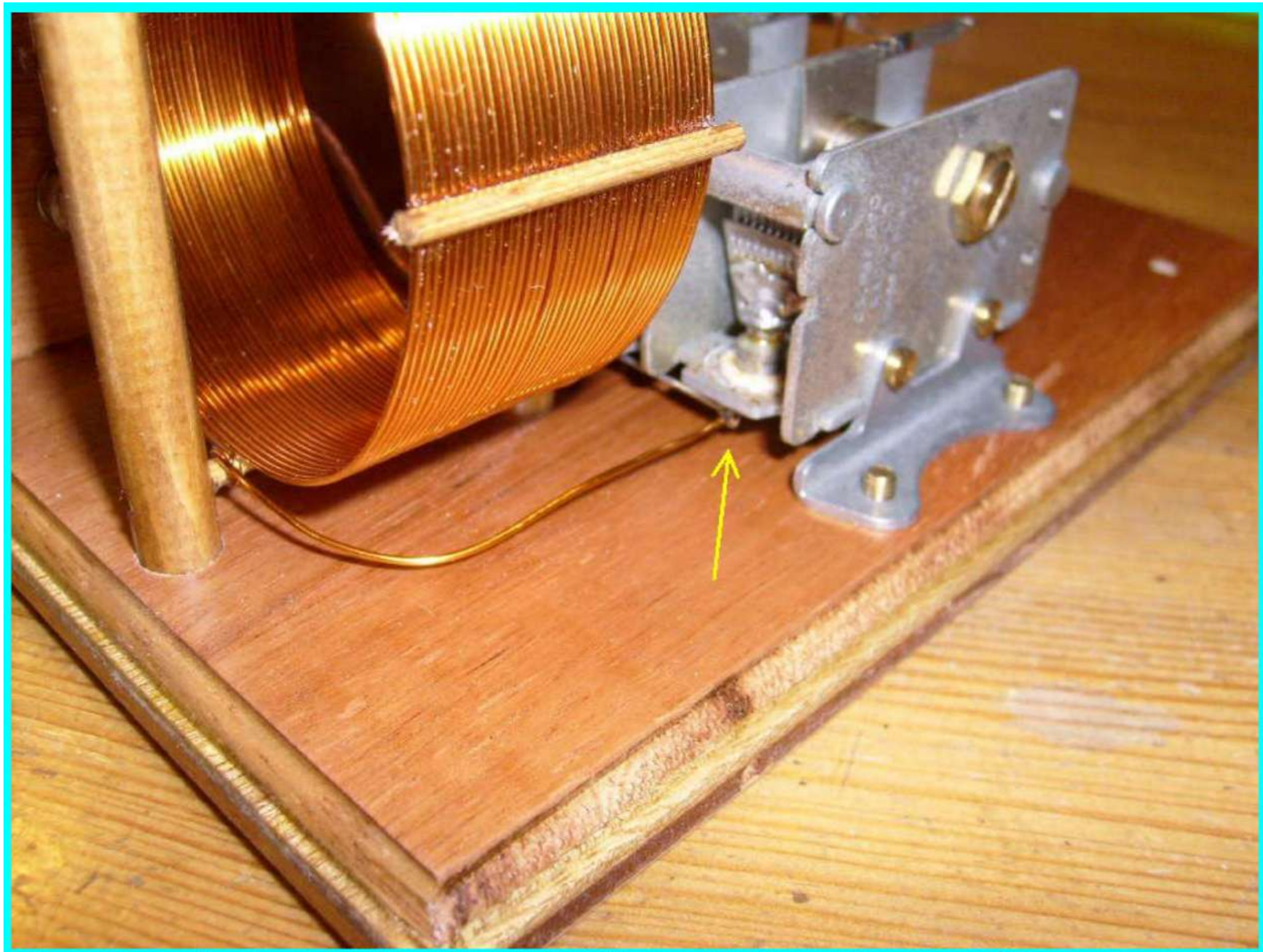

The other wire is cut to length, and soldered to the solder tag of the tuning capacitor (indicated with arrow).

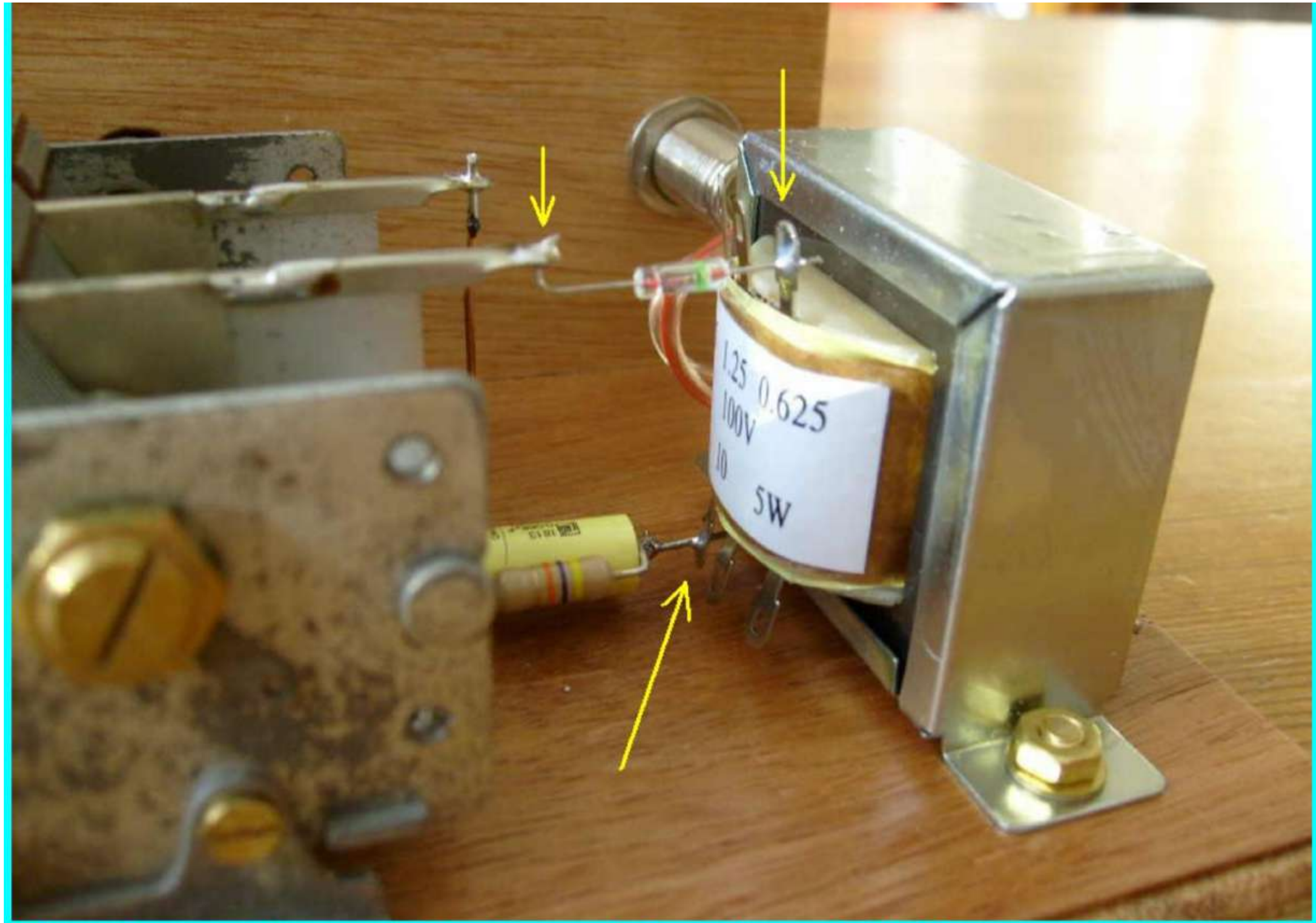

Place the transformer on the frame, fix it with two M4x16 screws, 2 rings and 2 nuts
The wire of the 68 nF capacitor must stick through the "0" connection of the transformer, and is then soldered.
Solder the OA95 diode between tuning capacitor and the "0.625W" connection of the transformer.
The green band on the diode must point towards the transformer.

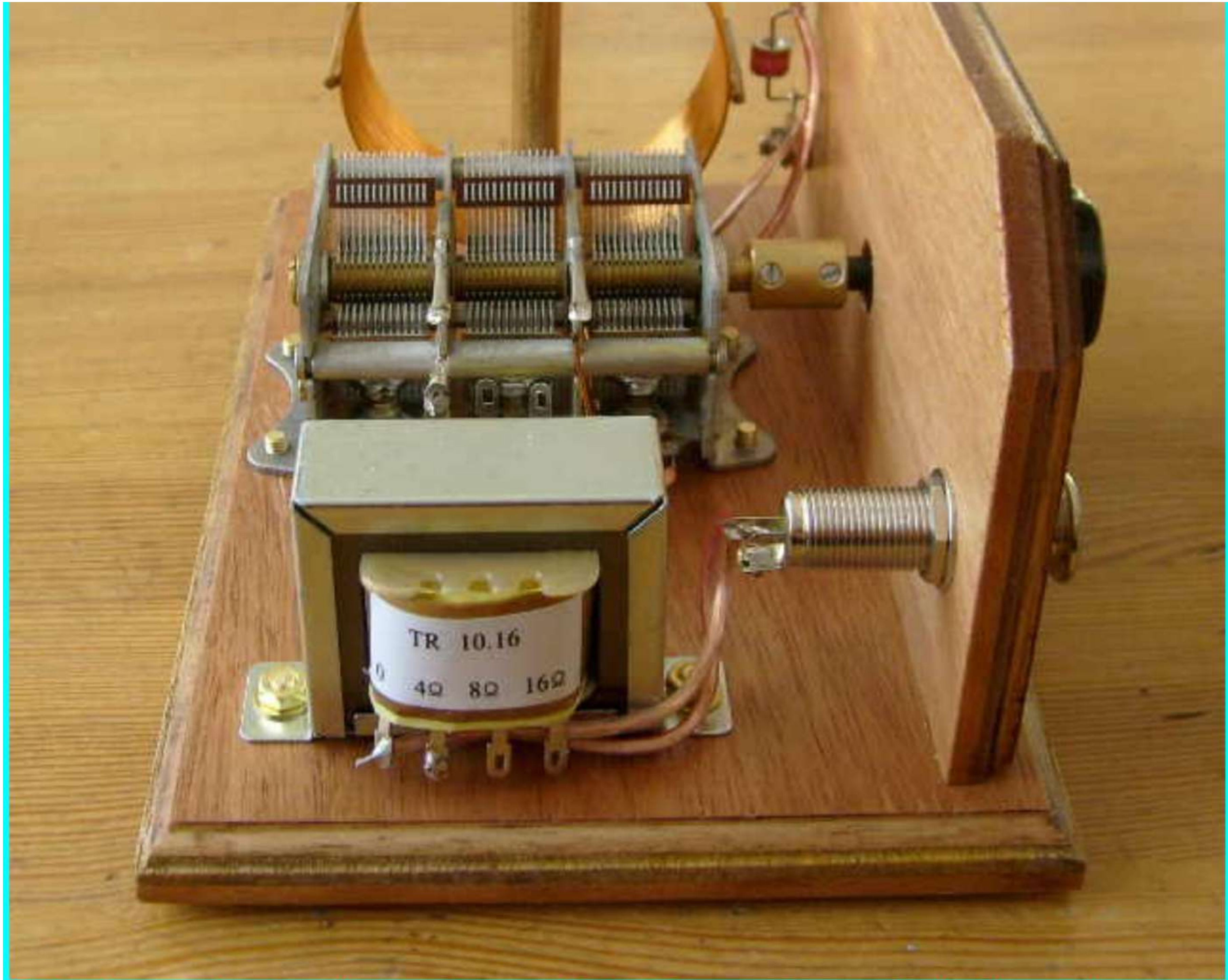
TR 10.16
0 4Ω 8Ω 16Ω

Solder the two wires from the transformer to the headphone socket (see for detail: next picture).

Solder the two wires from the tuning capacitor to the terminal posts.
The wire with the red line on it, comes on the upper (antenna) connection.
The wire without red line, comes on the lower (ground) connection.

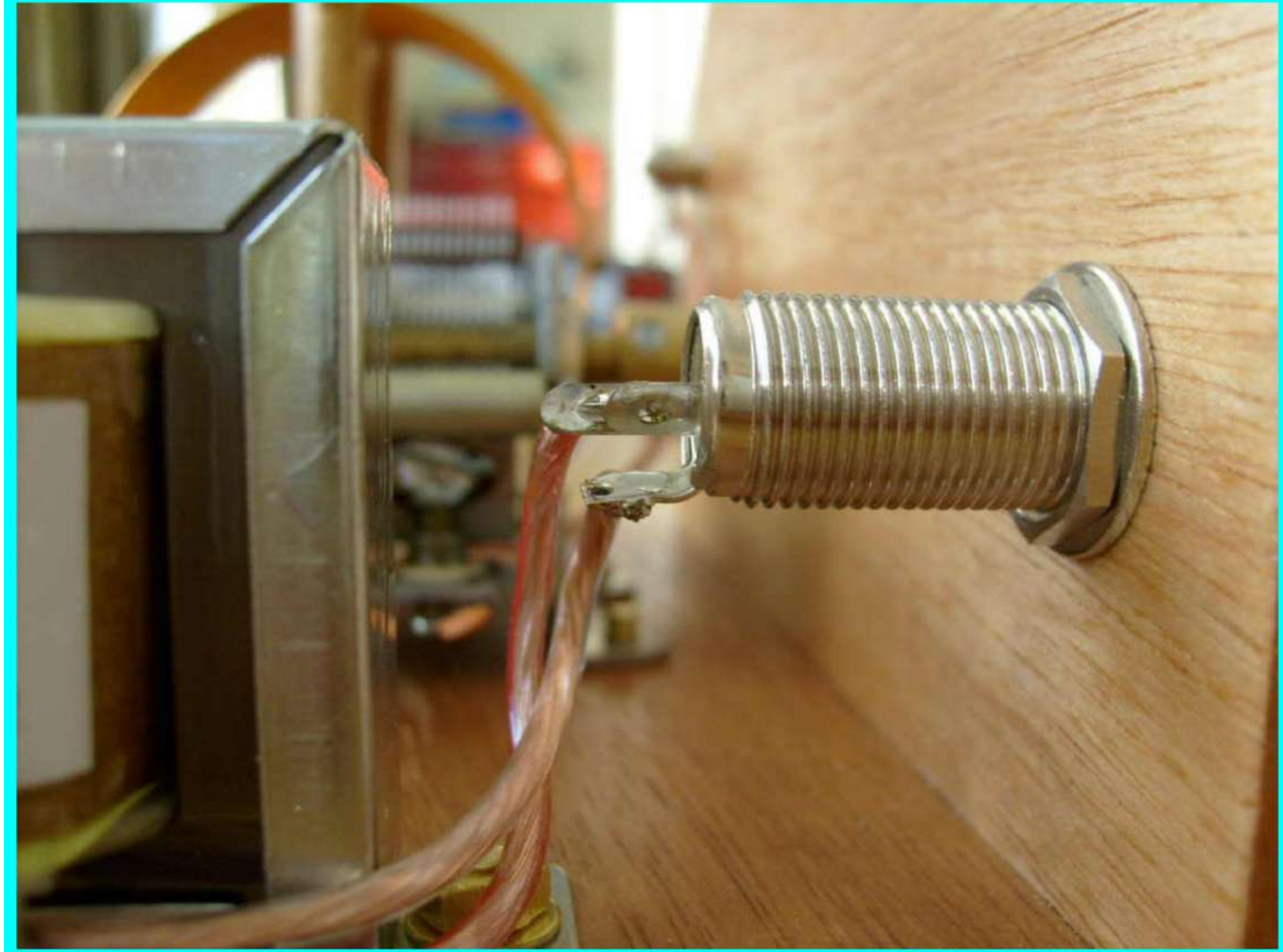

Detail of the wires on the headphone socket.
The wire without red line comes on the ground connection.
The wire with the red line comes on the two signal pins, so these two signal pins are connected together.

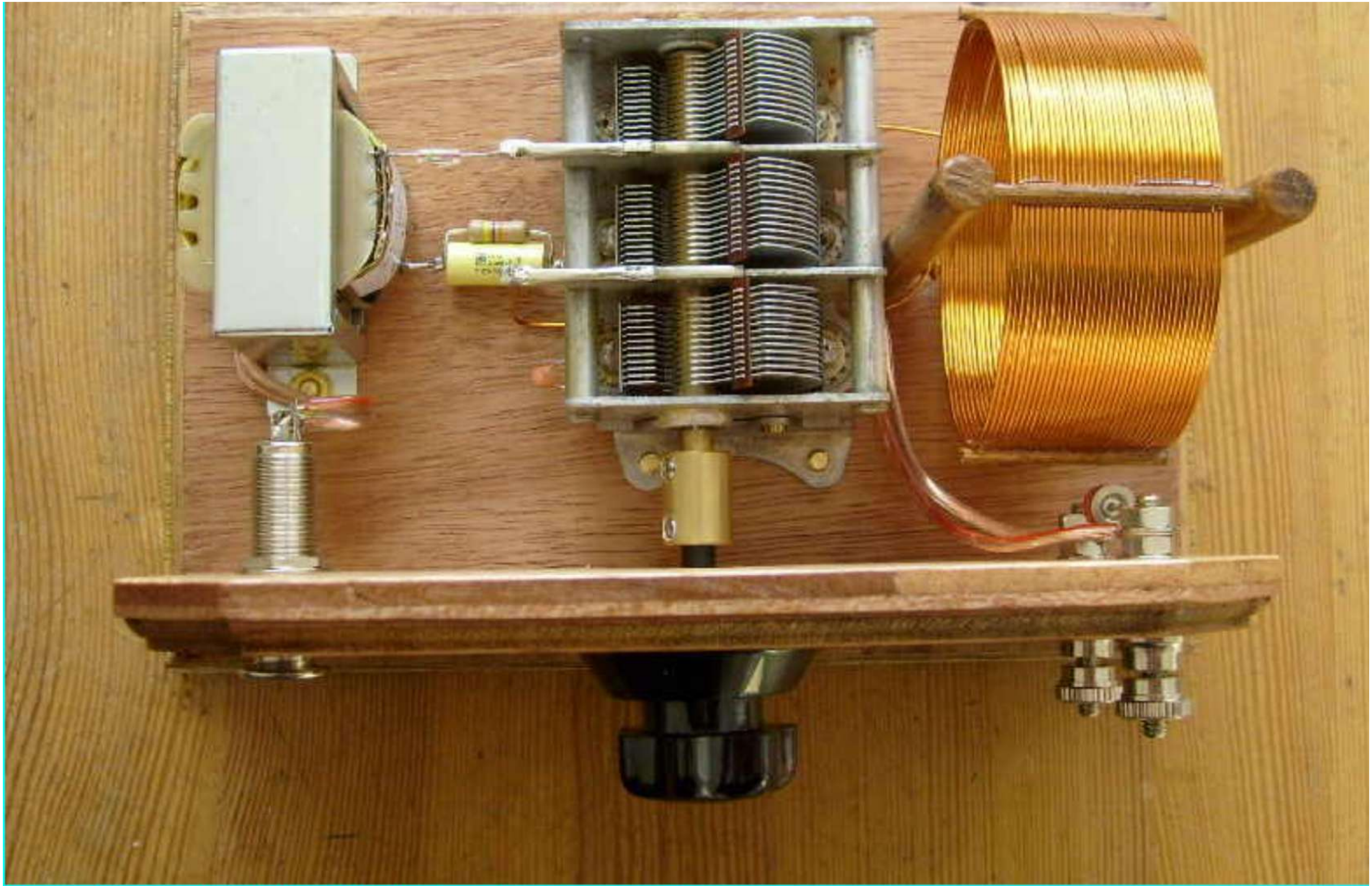

Mount the extension shaft and the knob on the tuning capacitor.
In the middle position, the pointer on the knob must point upwards.

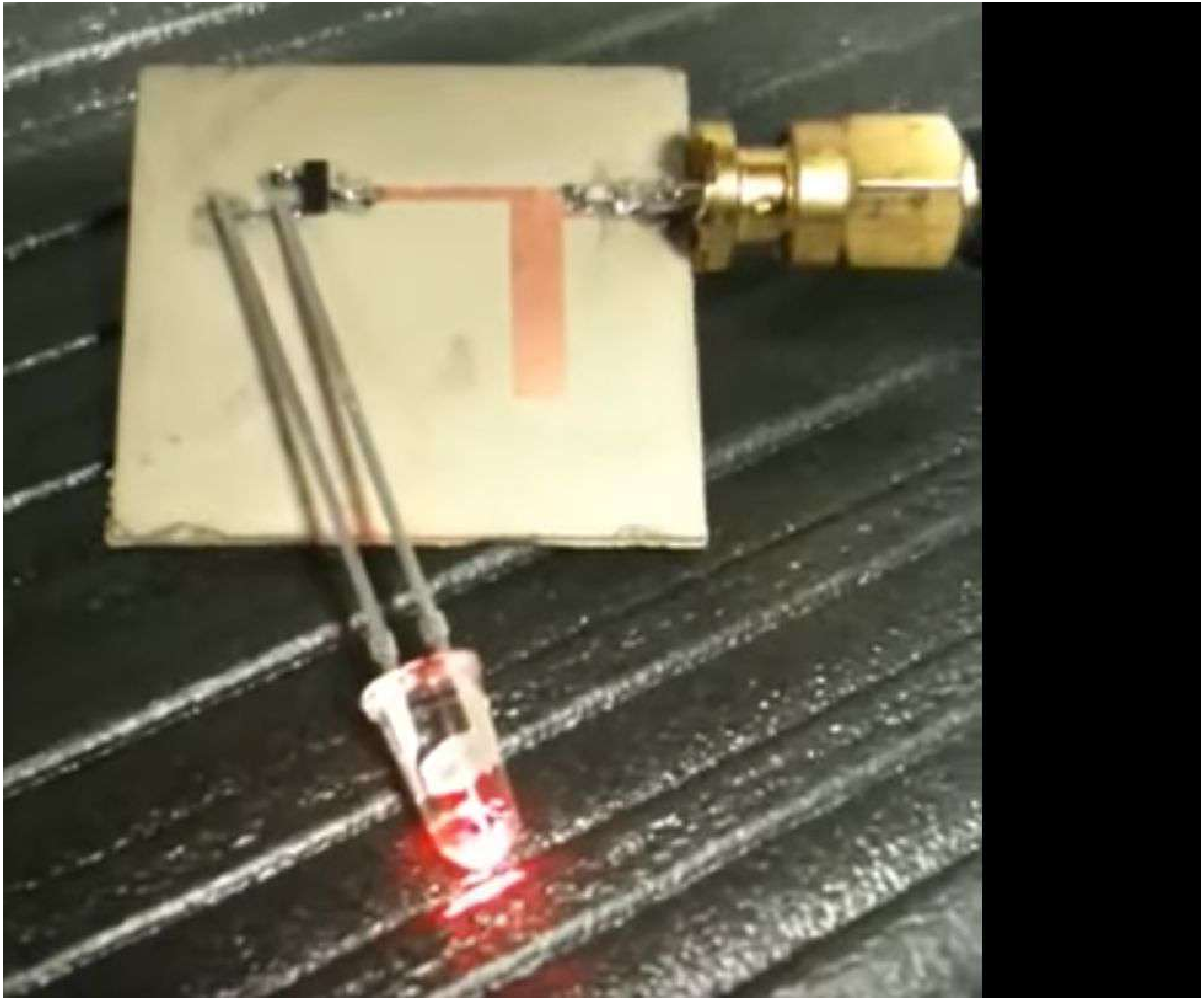

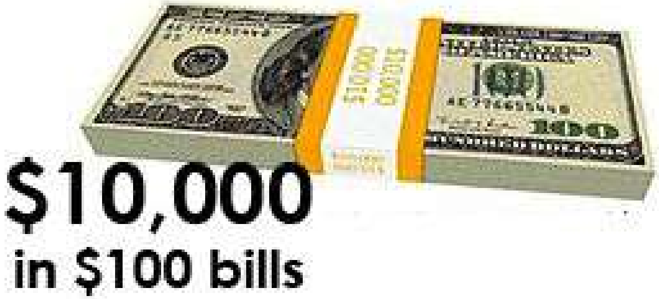
$10,000
in $100 bills

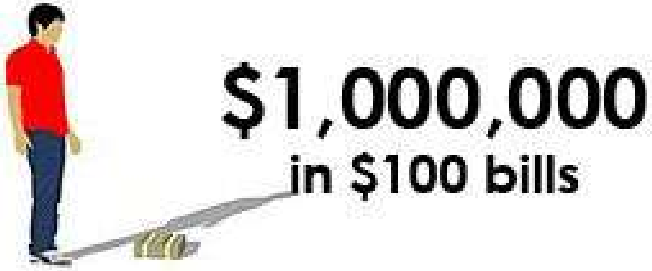
$1,000,000
in $100 bills

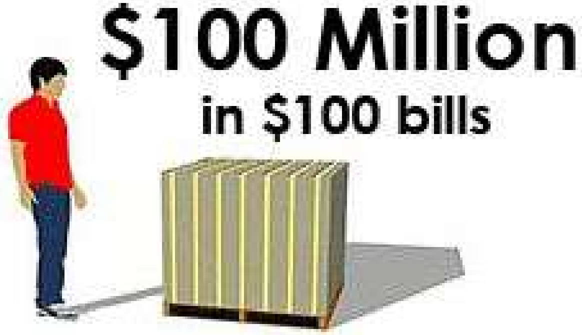
$100 Million
in $100 bills

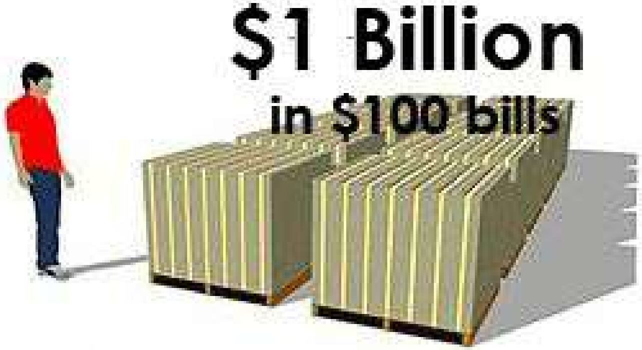
$1 Billion
in $100 bills

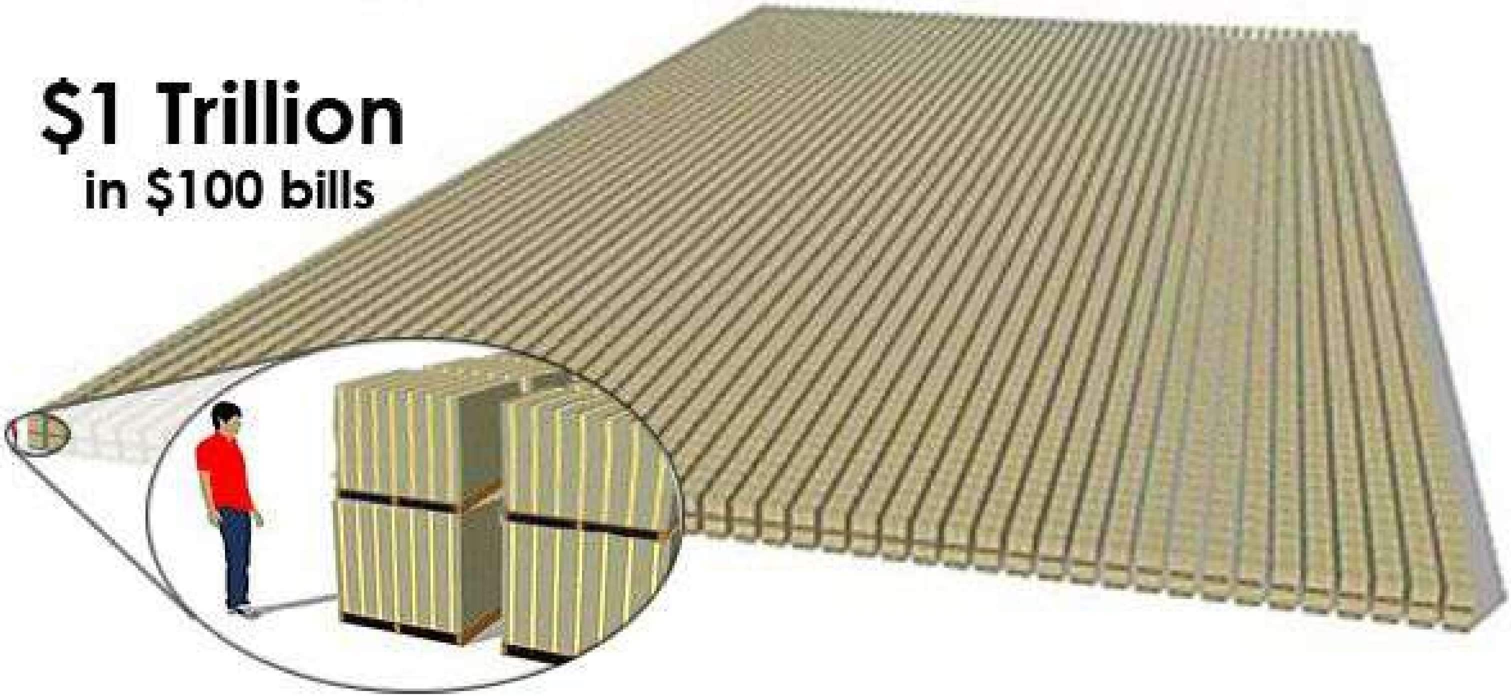
$1 Trillion
in $100 bills

USB Type A Female Field Termination Connector

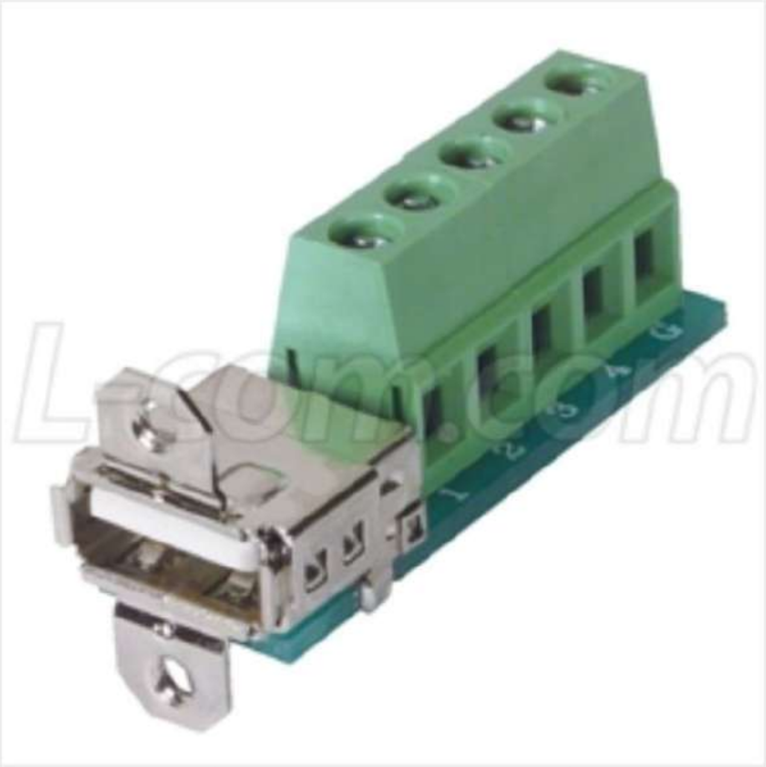

Roll over image to zoom in

L-com Item # USBAFT

List Price	$7.24
Your Price 1-9	**$7.24**
10-24	$6.81
25-99	$6.37
100 +	Call Us

Availability: In Stock

Available for Same Day Shipping

Quantity 1 ADD TO CART

Email This Page

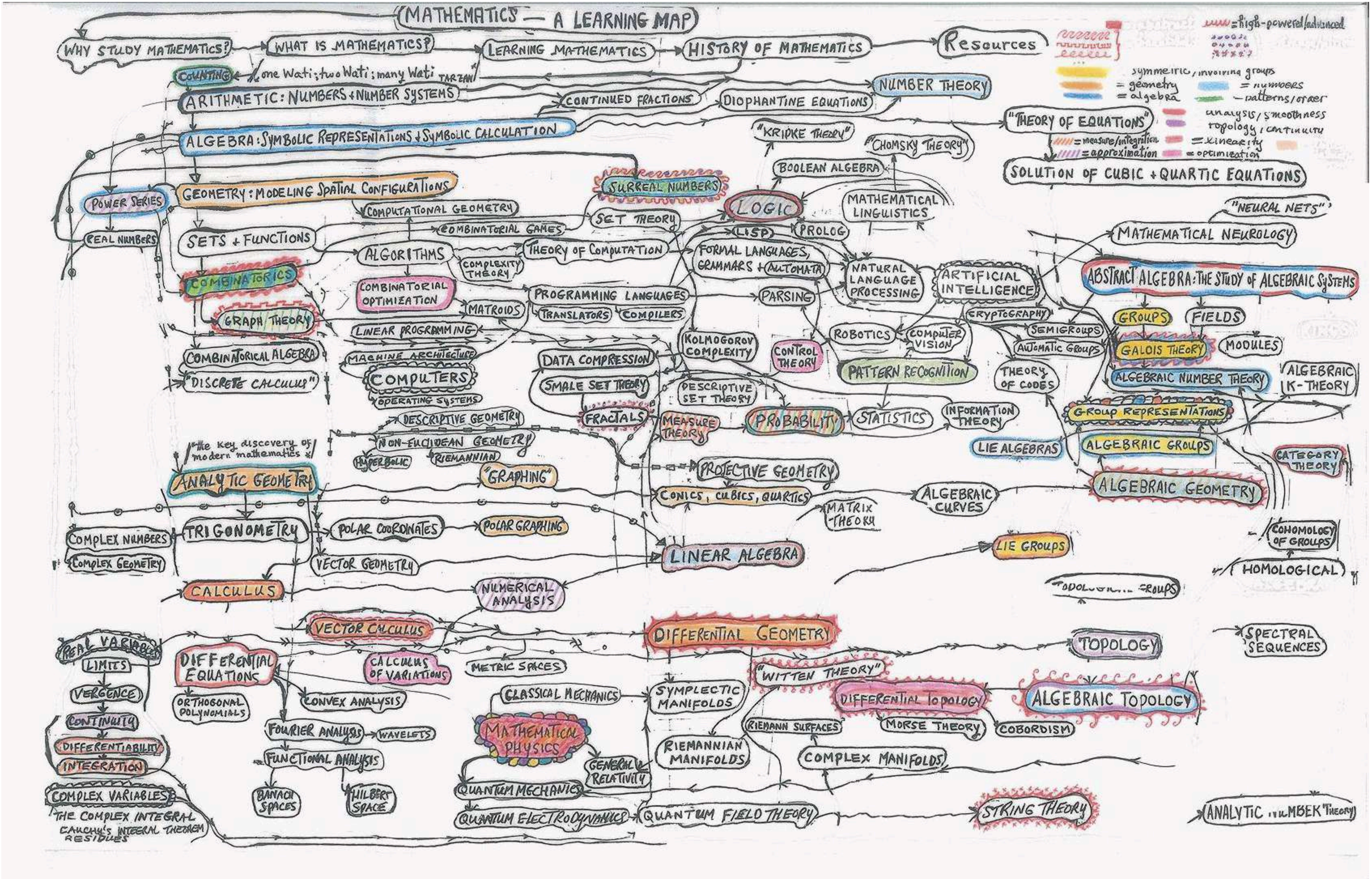
MATHEMATICS — A LEARNING MAP
WHY STUDY MATHEMATICS?
WHAT IS MATHEMATICS?
LEARNING MATHEMATICS
HISTORY OF MATHEMATICS
Resources
= high-powered/advanced
symmetric, involving groups
= geometry
= algebra
= numbers
= patterns/order
analysis/smoothness
topology, continuity
= measure/integration
= approximation
= linearity
= optimization
COUNTING
one Wati; two Wati; many Wati
TARZAN
NUMBER THEORY
ARITHMETIC: NUMBERS + NUMBER SYSTEMS
CONTINUED FRACTIONS
DIOPHANTINE EQUATIONS
ALGEBRA: SYMBOLIC REPRESENTATIONS + SYMBOLIC CALCULATION
"THEORY OF EQUATIONS"
"KRIPKE THEORY"
"CHOMSKY THEORY"
SOLUTION OF CUBIC + QUARTIC EQUATIONS
BOOLEAN ALGEBRA
SURREAL NUMBERS
POWER SERIES
GEOMETRY: MODELING SPATIAL CONFIGURATIONS
LOGIC
MATHEMATICAL LINGUISTICS
COMPUTATIONAL GEOMETRY
SET THEORY
"NEURAL NETS"
REAL NUMBERS
COMBINATORIAL GAMES
LISP
PROLOG
MATHEMATICAL NEUROLOGY
SETS + FUNCTIONS
ALGORITHMS
THEORY OF COMPUTATION
FORMAL LANGUAGES, GRAMMARS + AUTOMATA
COMPLEXITY THEORY
NATURAL LANGUAGE PROCESSING
ARTIFICIAL INTELLIGENCE
ABSTRACT ALGEBRA: THE STUDY OF ALGEBRAIC SYSTEMS
COMBINATORICS
COMBINATORIAL OPTIMIZATION
PROGRAMMING LANGUAGES
PARSING
GROUPS
FIELDS
MATROIDS
TRANSLATORS
COMPILERS
CRYPTOGRAPHY
GRAPH THEORY
LINEAR PROGRAMMING
ROBOTICS
COMPUTER VISION
SEMIGROUPS
AUTOMATIC GROUPS
GALOIS THEORY
MODULES
KOLMOGOROV COMPLEXITY
CONTROL THEORY
COMBINATORICAL ALGEBRA
MACHINE ARCHITECTURE
DATA COMPRESSION
PATTERN RECOGNITION
THEORY OF CODES
ALGEBRAIC K-THEORY
COMPUTERS
OPERATING SYSTEMS
"DISCRETE CALCULUS"
SMALE SET THEORY
DESCRIPTIVE SET THEORY
ALGEBRAIC NUMBER THEORY
DESCRIPTIVE GEOMETRY
FRACTALS
MEASURE THEORY
PROBABILITY
STATISTICS
INFORMATION THEORY
GROUP REPRESENTATIONS
the key discovery of modern mathematics
NON-EUCLIDEAN GEOMETRY
HYPERBOLIC
RIEMANNIAN
LIE ALGEBRAS
ALGEBRAIC GROUPS
CATEGORY THEORY
"GRAPHING"
PROJECTIVE GEOMETRY
ANALYTIC GEOMETRY
CONICS, CUBICS, QUARTICS
ALGEBRAIC CURVES
ALGEBRAIC GEOMETRY
MATRIX THEORY
COMPLEX NUMBERS
TRIGONOMETRY
POLAR COORDINATES
POLAR GRAPHING
COHOMOLOGY OF GROUPS
COMPLEX GEOMETRY
VECTOR GEOMETRY
LINEAR ALGEBRA
LIE GROUPS
HOMOLOGICAL
CALCULUS
NUMERICAL ANALYSIS
TOPOLOGICAL GROUPS
VECTOR CALCULUS
DIFFERENTIAL GEOMETRY
SPECTRAL SEQUENCES
REAL VARIABLES
TOPOLOGY
LIMITS
DIFFERENTIAL EQUATIONS
CALCULUS OF VARIATIONS
METRIC SPACES
"WITTEN THEORY"
VERGENCE
ORTHOGONAL POLYNOMIALS
CONVEX ANALYSIS
CLASSICAL MECHANICS
SYMPLECTIC MANIFOLDS
DIFFERENTIAL TOPOLOGY
ALGEBRAIC TOPOLOGY
CONTINUITY
FOURIER ANALYSIS
WAVELETS
MATHEMATICAL PHYSICS
RIEMANN SURFACES
MORSE THEORY
COBORDISM
DIFFERENTIABILITY
FUNCTIONAL ANALYSIS
RIEMANNIAN MANIFOLDS
COMPLEX MANIFOLDS
INTEGRATION
GENERAL RELATIVITY
COMPLEX VARIABLES
BANACH SPACES
HILBERT SPACE
QUANTUM MECHANICS
THE COMPLEX INTEGRAL
CAUCHY'S INTEGRAL THEOREM
RESIDUES
QUANTUM ELECTRODYNAMICS
QUANTUM FIELD THEORY
STRING THEORY
ANALYTIC NUMBER THEORY

Philmore

MANUFACTURING CO. Inc.
Richmond Hill 18, N. Y.

MODEL NO.
VC-1000

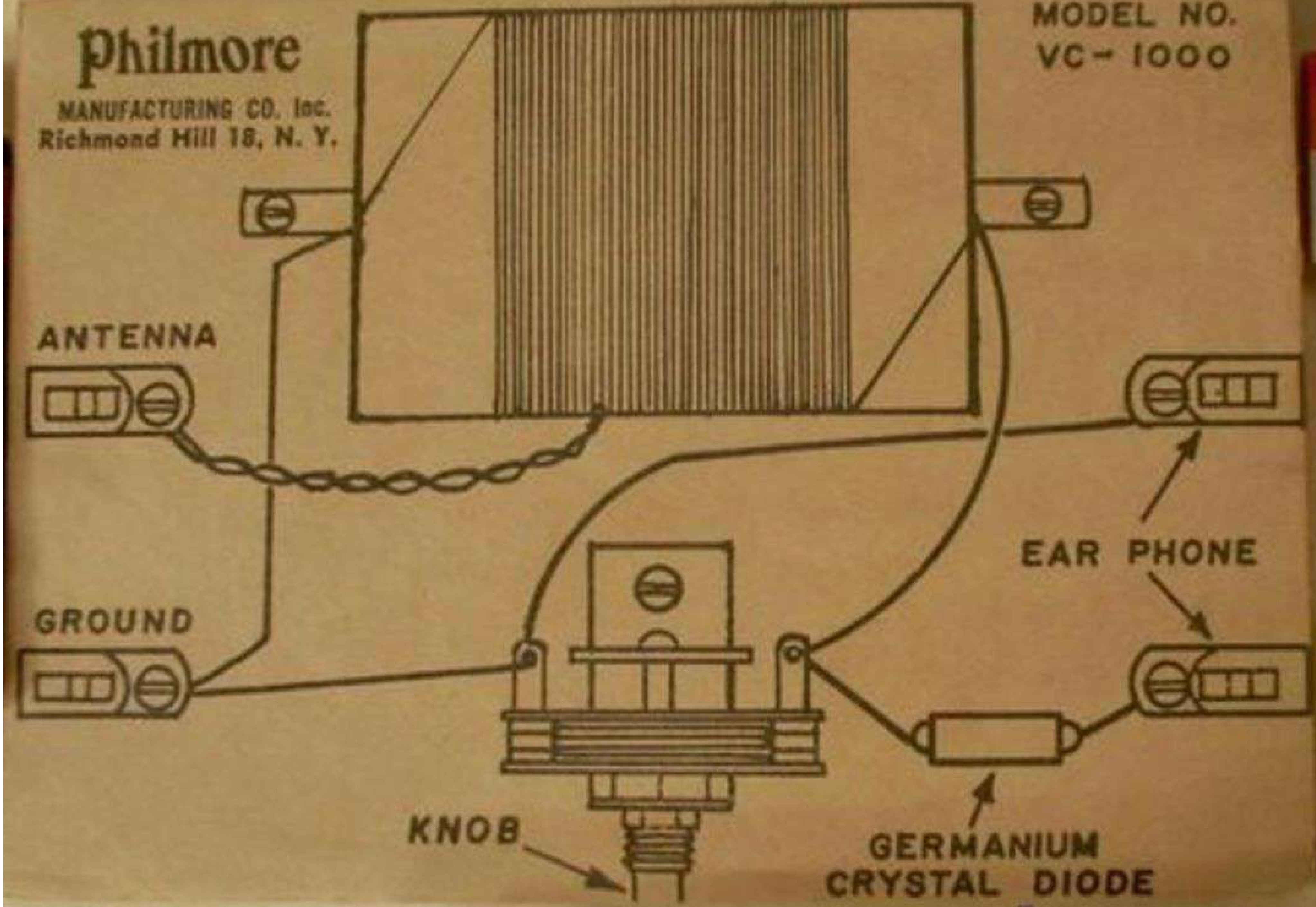
ANTENNA
GROUND
EAR PHONE
KNOB
GERMANIUM
CRYSTAL DIODE

1 LED azul
(con tensión entre 3V y 3,6V)

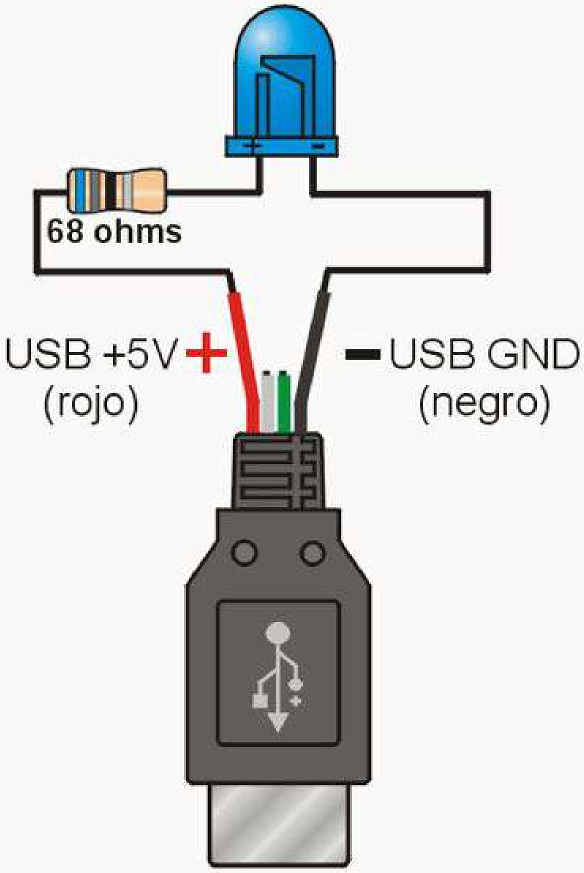
68 ohms
USB +5V +
(rojo)
- USB GND
(negro)

WWW.INVENTABLE.EU

1 LED rojo
(con tensión entre 1,8V y 2,1V)

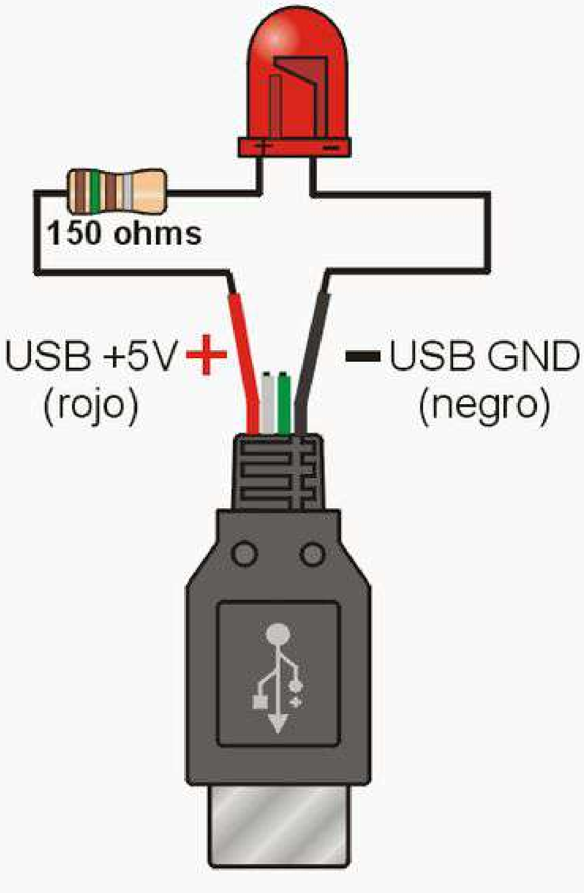
150 ohms
USB +5V +
(rojo)
- USB GND
(negro)

The Tesla Experiment

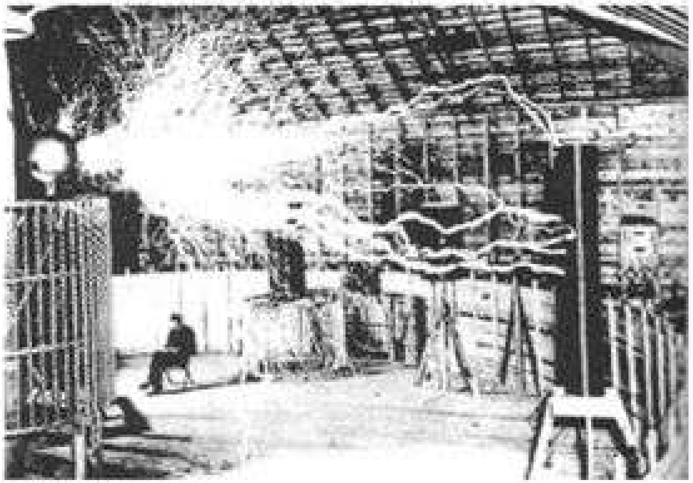

TRANSVERSE WAVE INTERFERENCE

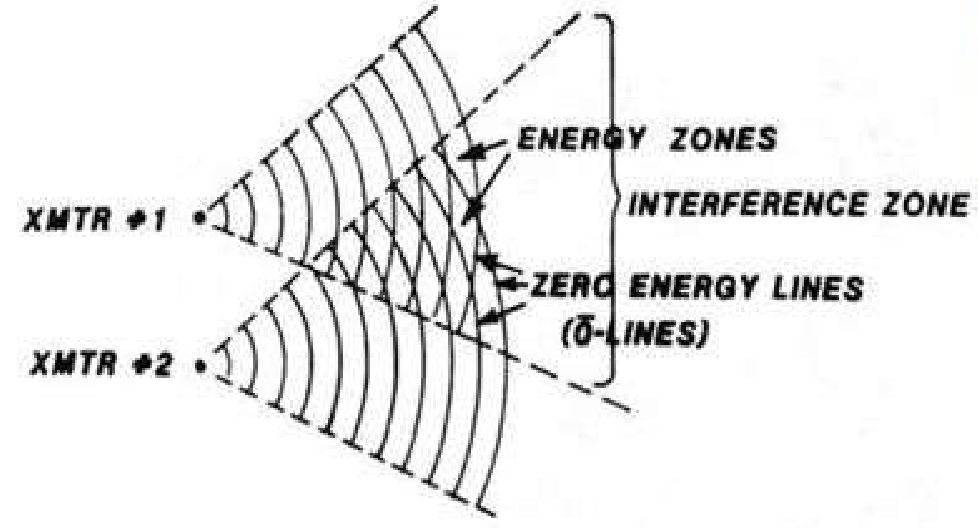
XMTR #1
XMTR #2
ENERGY ZONES
INTERFERENCE ZONE
ZERO ENERGY LINES
(0̄-LINES)

LONGITUDINAL WAVE INTERFERENCE
(SCALAR WAVES)

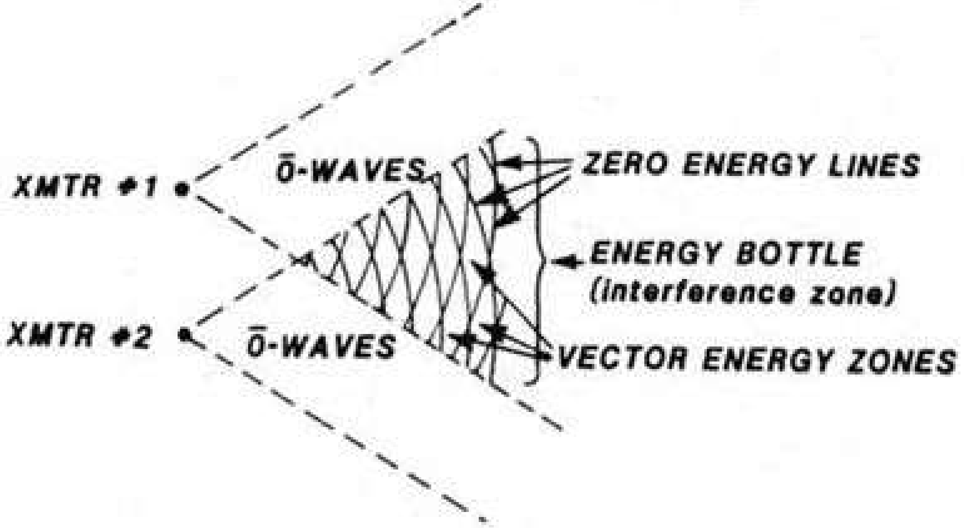
XMTR #1
XMTR #2
0̄-WAVES
0̄-WAVES
ZERO ENERGY LINES
ENERGY BOTTLE
(interference zone)
VECTOR ENERGY ZONES

© PoweredTemplates.com

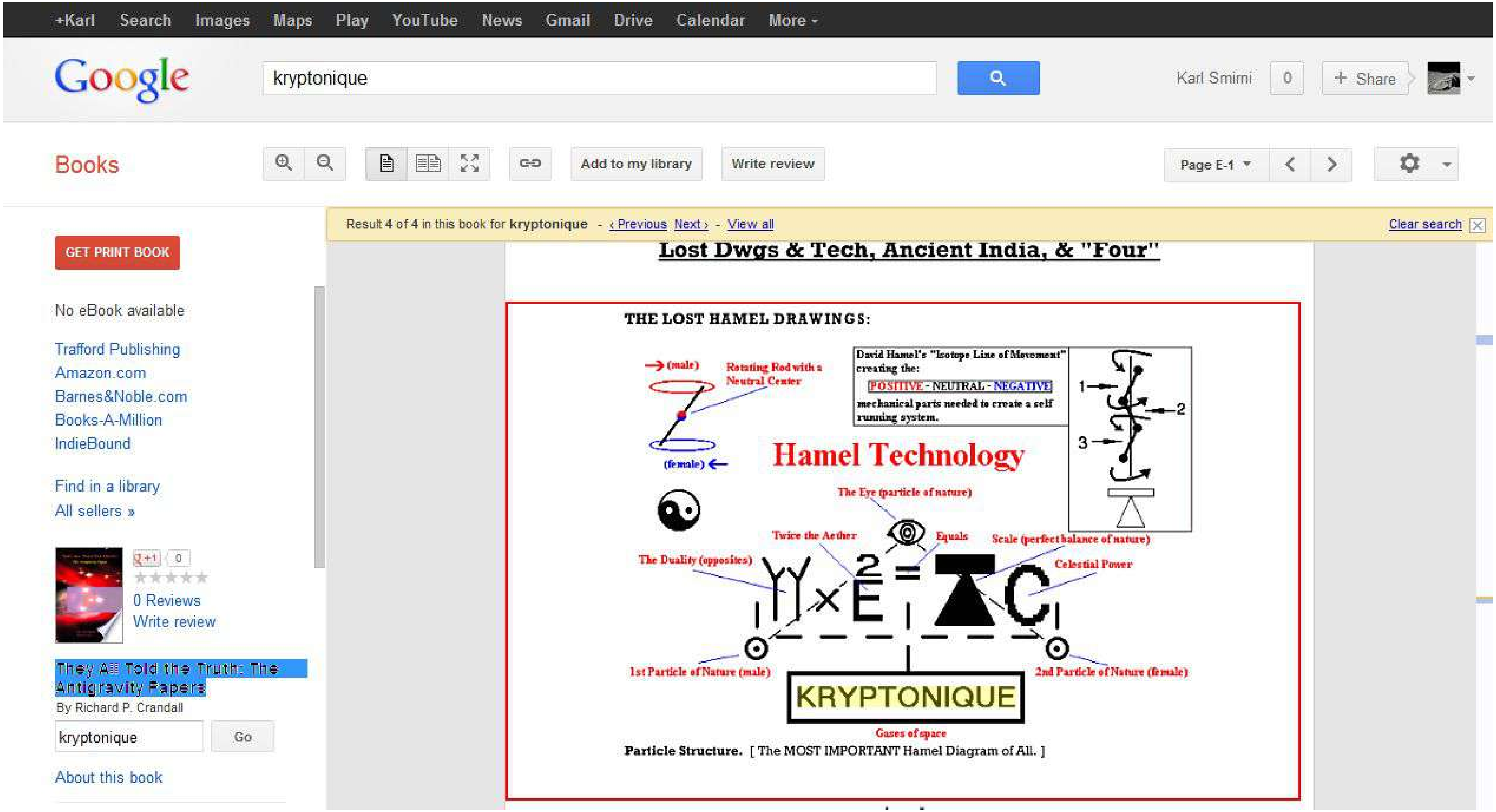
+Karl
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
kryptonique
Karl Smirni
0
Share
Books
Add to my library
Write review
Page E-1
GET PRINT BOOK
No eBook available
Trafford Publishing
Amazon.com
Barnes&Noble.com
Books-A-Million
IndieBound
Find in a library
All sellers »
+1
0
0 Reviews
Write review
kryptonique
Go
About this book
Clear search

Result 4 of 4 in this book for **kryptonique** - ‹ Previous Next › - View all

They All Told the Truth: The Antigravity Papers
By Richard P. Crandall

Lost Dwgs & Tech, Ancient India, & "Four"

THE LOST HAMEL DRAWINGS:

David Hamel's "Isotope Line of Movement" creating the:
POSITIVE - NEUTRAL - NEGATIVE
mechanical parts needed to create a self running system.

(male)
Rotating Rod with a Neutral Center
(female)
1
2
3
Hamel Technology
The Eye (particle of nature)
Twice the Aether
Equals
Scale (perfect balance of nature)
The Duality (opposites)
Celestial Power
1st Particle of Nature (male)
2nd Particle of Nature (female)
KRYPTONIQUE
Gases of space

Particle Structure. [The MOST IMPORTANT Hamel Diagram of All.]

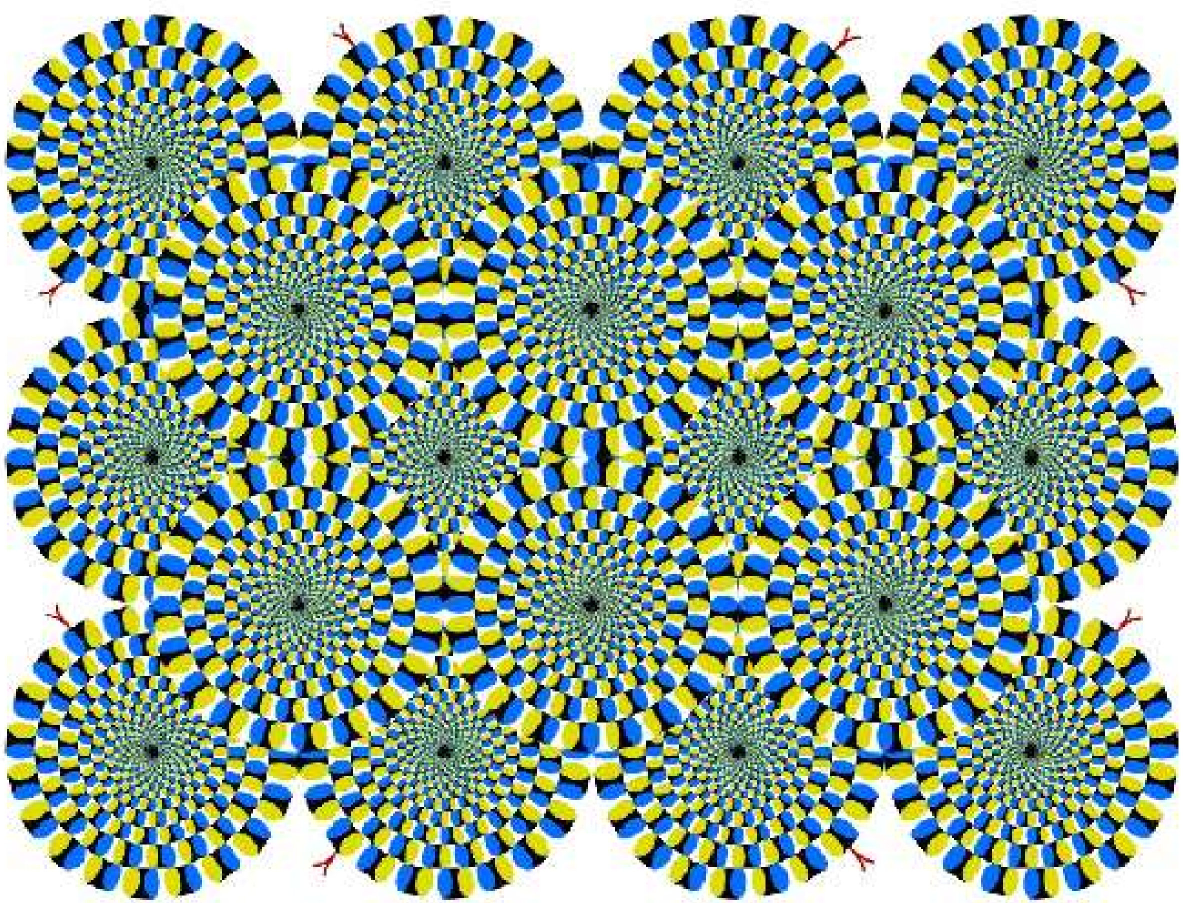

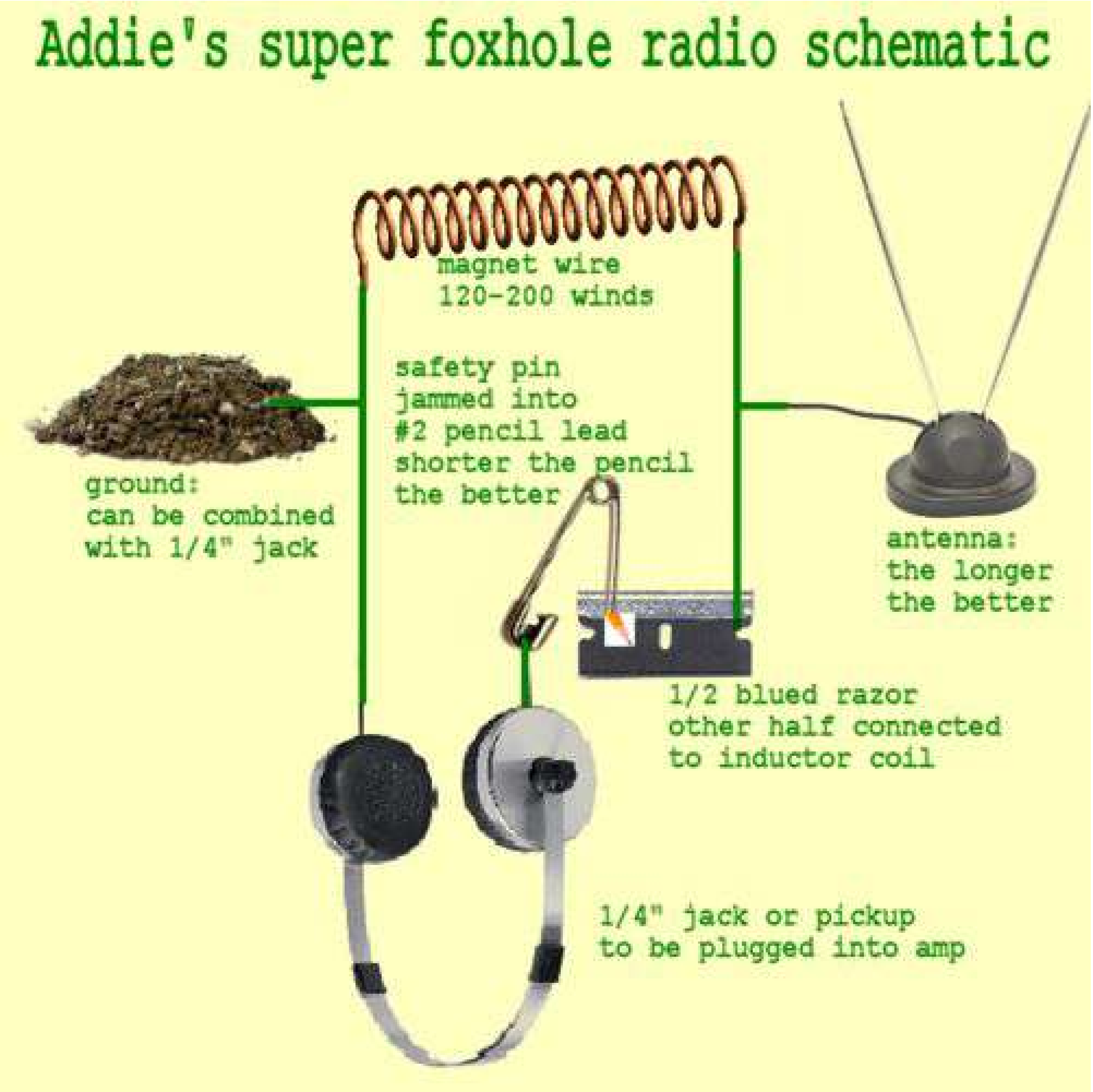
Addie's super foxhole radio schematic
magnet wire
120-200 winds
safety pin
jammed into
#2 pencil lead
shorter the pencil
the better
ground:
can be combined
with 1/4" jack
antenna:
the longer
the better
1/2 blued razor
other half connected
to inductor coil
1/4" jack or pickup
to be plugged into amp

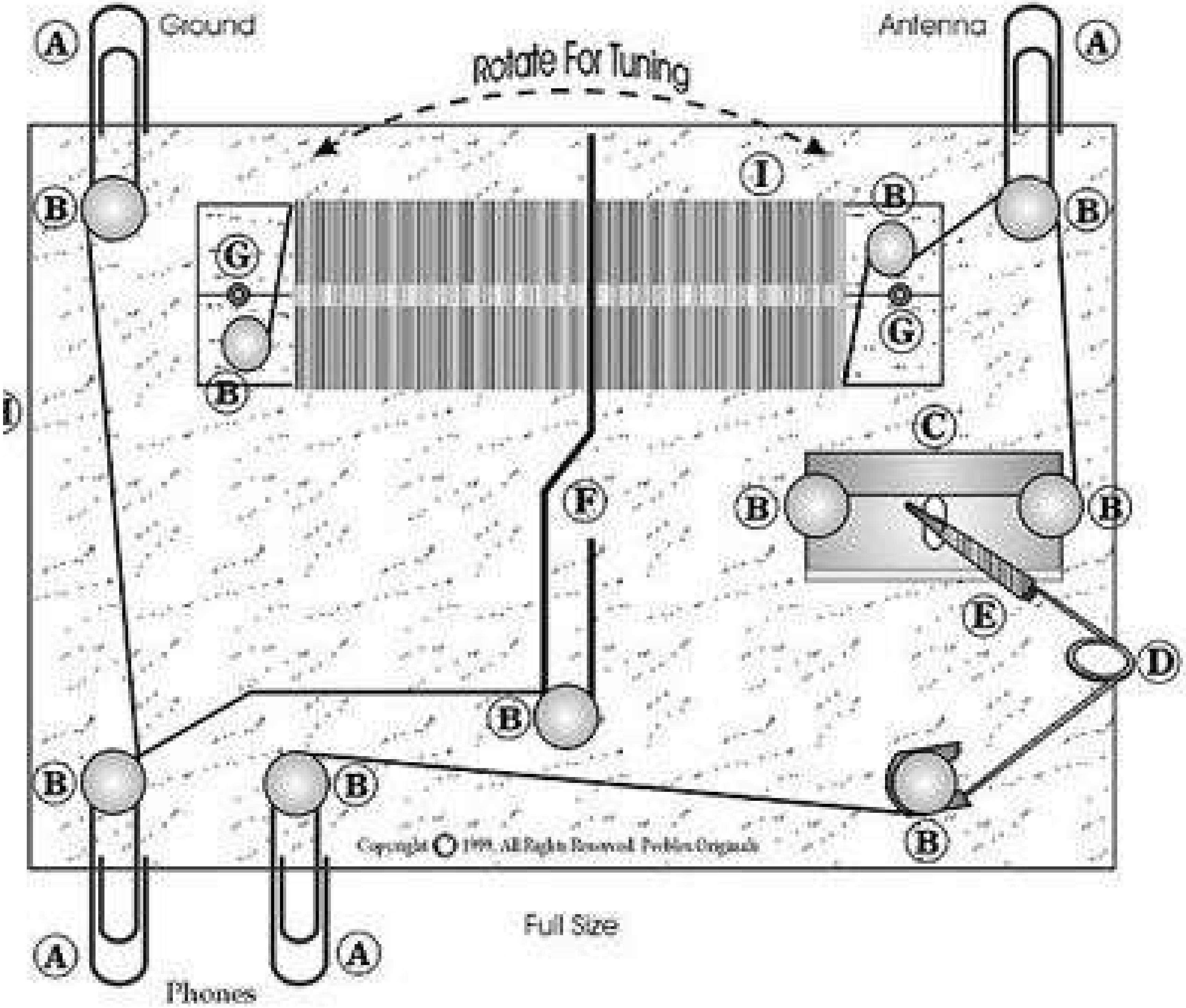
Ground
Antenna
A
A
Rotate For Tuning
I
B
B
B
G
G
B
C
B
B
F
E
D
B
B
B
B
Copyright ○ 1999, All Rights Reserved.
Full Size
A
A
Phones

The PO-102, Crystal Radio Kit, w/Slider-Tuning!

Perfect for parents, grandparents, and teachers to build with kids! Also a super starter kit for the beginning adult builder! Includes small coil form, wire, slider assembly, capacitors, and diode. Easy parts layout guide provided for kids. Excellent performance for such a simple but elegant classic slider. Earplug included.

PO-102, w/Crystal Earphone

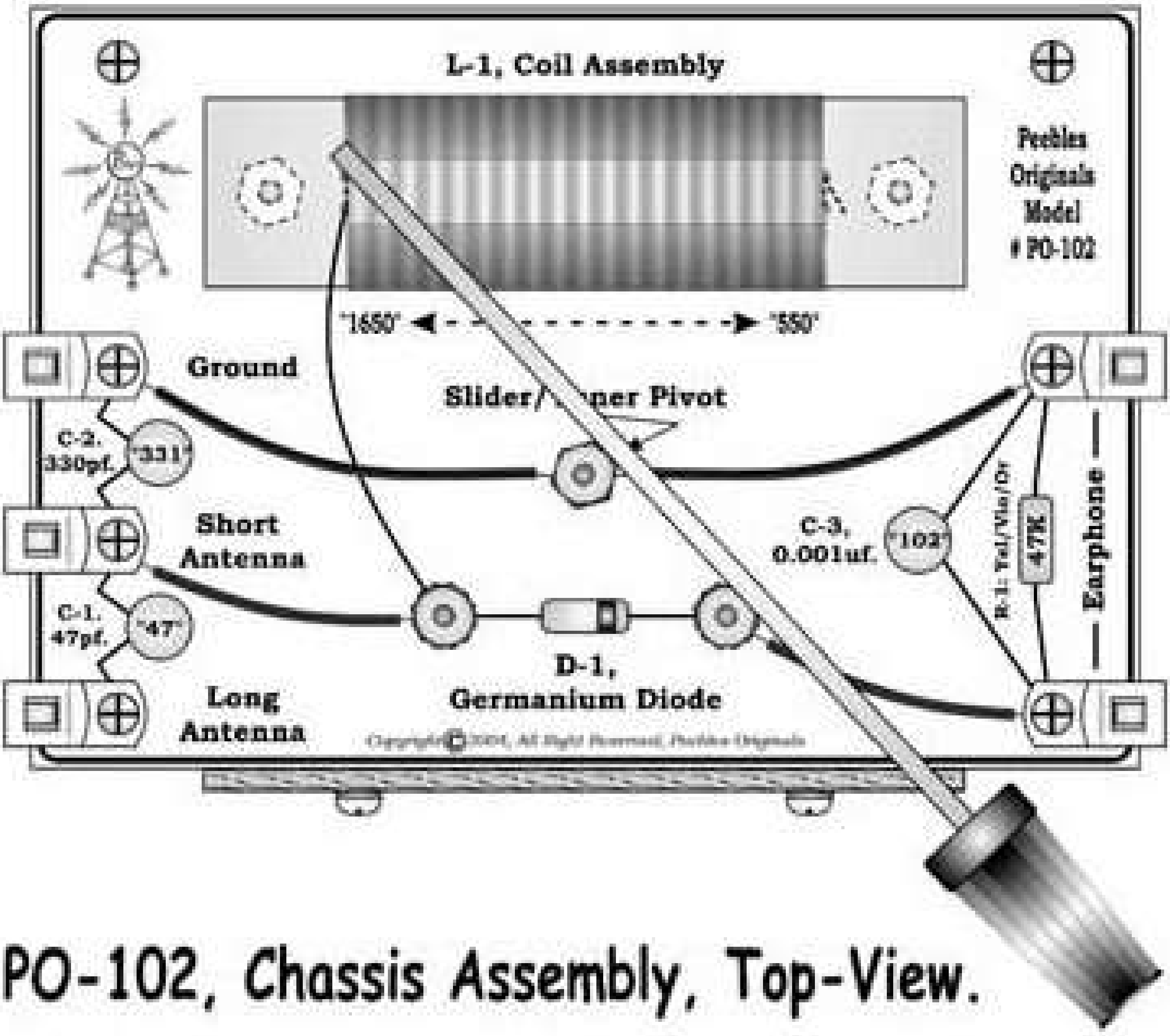
L-1, Coil Assembly
Peebles Originals Model # PO-102
"1650"
"550"
Ground
Slider/ ner Pivot
C-2. 330pf.
"331"
Short Antenna
C-1. 47pf.
"47"
Long Antenna
D-1, Germanium Diode
C-3, 0.001uf.
"102"
R-1: Yel/Vio/Or
47K
Earphone
Copyright © 2004, All Right Reserved, Peebles Originals

PO-102, Chassis Assembly, Top-View.

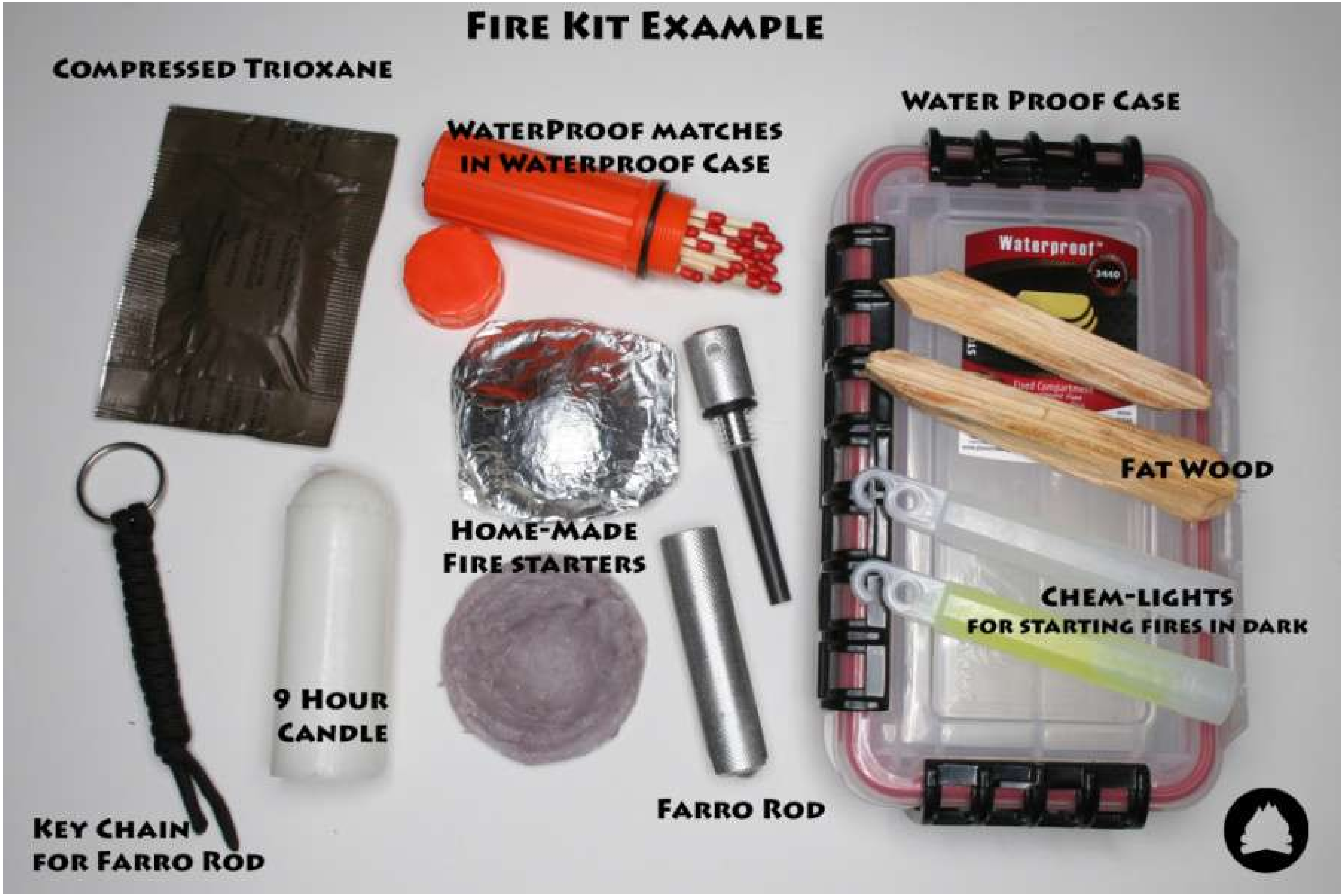

Fire Kit Example

Compressed Trioxane
WaterProof matches in Waterproof Case
Water Proof Case
Fat Wood
Home-Made Fire starters
Chem-lights for starting fires in dark
9 Hour Candle
Farro Rod
Key Chain for Farro Rod

PREPPERS SHOP
PREPARE
DEFEND
SURVIVE
FIRST
SAWYER
CONSERVE WATER
NAVIGATION
HIGHLANDER
SURVIVAL BAG

WWW.PREPPERSSHOP.CO.UK

THE ULTIMATE BUG OUT BAG

1. Anglo Arms Rambo Knife
2. Anglo Arms Machete
3. Austrian Ex-Military Water Flask
4. BCB Trekker Lifesaver Pack
5. BCB Compact Fishing Kit
6. Canadian Gas Mask w/ Filters
7. Feit Electric 500 LED Torch
8. French Military Mess Tins
9. Hexamine Solid Fuel Cooker
10. Heavy Duty Camo Tarp
11. Highlander Survival Bag
12. Highlander Solid Fuel Cooker
13. Highlander Trekker Hammock
14. Highlander Folding Saw
15. Kombat UK W/proof Matches
16. Kombat UK 40l Molle Bag
17. Kombat UK Fire Starting Kit
18. Kombat UK Compass
19. Kombat UK Throwing Axe
20. Oasis Water Purification Tablets
21. Pro Force Survival Saw
22. Sawyer Mini Water Filter
23. Sharpening Stone
24. Survival Mirror
25. Turboflame Lighter
26. UST Survival Poncho
27. Waterproof Rucksack Liner
28. Wind Up Solar Torch
29. Wooden Deluxe Knife w/ Case
30. 100ft Paracord
31. 24hr Ration Pack

£280

ERbar
BAG
DUCK
FRONTIER STRAW

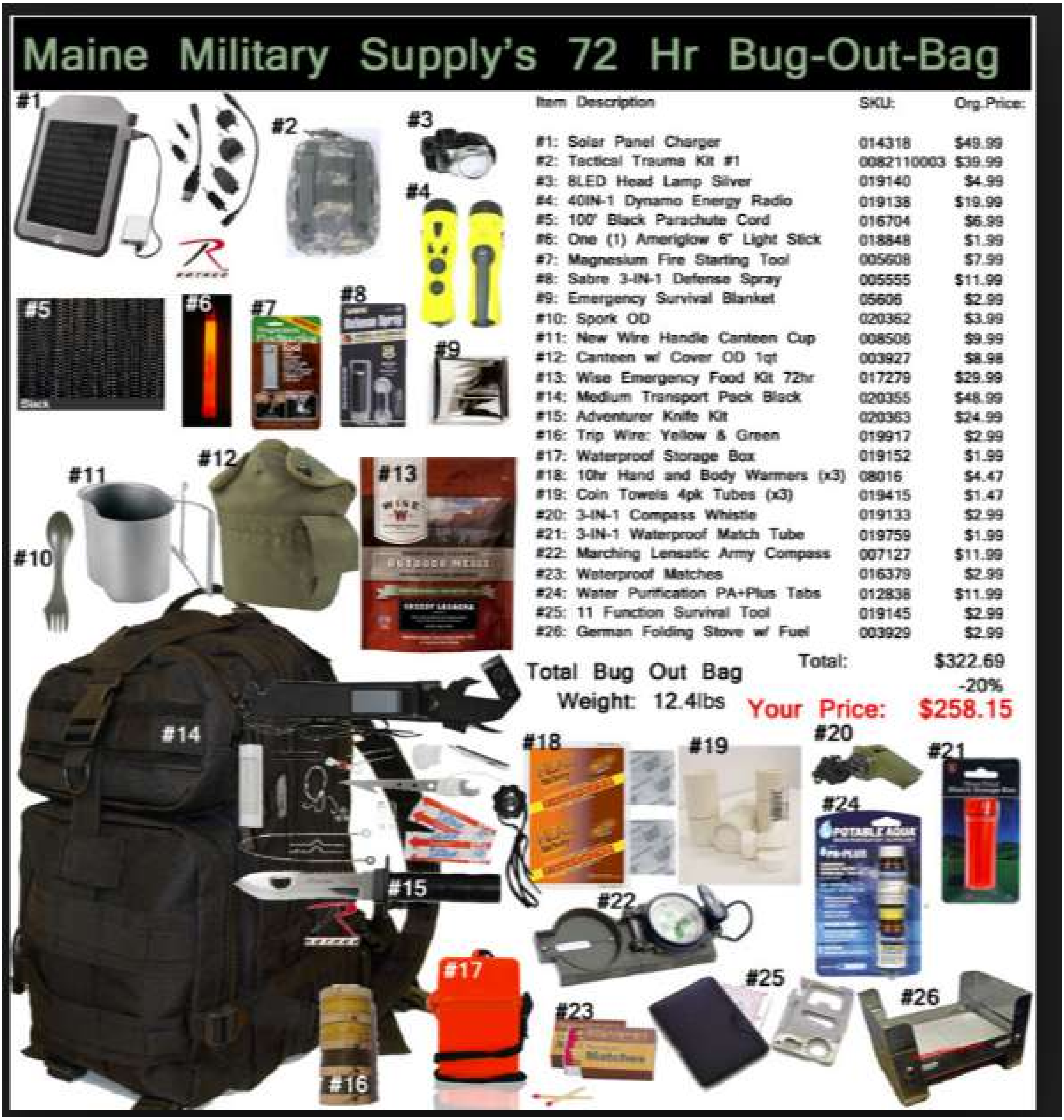

Maine Military Supply's 72 Hr Bug-Out-Bag

Item Description	SKU:	Org.Price:
#1: Solar Panel Charger	014318	$49.99
#2: Tactical Trauma Kit #1	0082110003	$39.99
#3: 8LED Head Lamp Silver	019140	$4.99
#4: 40IN-1 Dynamo Energy Radio	019138	$19.99
#5: 100' Black Parachute Cord	016704	$6.99
#6: One (1) Ameriglow 6" Light Stick	018848	$1.99
#7: Magnesium Fire Starting Tool	005608	$7.99
#8: Sabre 3-IN-1 Defense Spray	005555	$11.99
#9: Emergency Survival Blanket	05606	$2.99
#10: Spork OD	020362	$3.99
#11: New Wire Handle Canteen Cup	008506	$9.99
#12: Canteen w/ Cover OD 1qt	003927	$8.98
#13: Wise Emergency Food Kit 72hr	017279	$29.99
#14: Medium Transport Pack Black	020355	$48.99
#15: Adventurer Knife Kit	020363	$24.99
#16: Trip Wire: Yellow & Green	019917	$2.99
#17: Waterproof Storage Box	019152	$1.99
#18: 10hr Hand and Body Warmers (x3)	08016	$4.47
#19: Coin Towels 4pk Tubes (x3)	019415	$1.47
#20: 3-IN-1 Compass Whistle	019133	$2.99
#21: 3-IN-1 Waterproof Match Tube	019759	$1.99
#22: Marching Lensatic Army Compass	007127	$11.99
#23: Waterproof Matches	016379	$2.99
#24: Water Purification PA+Plus Tabs	012838	$11.99
#25: 11 Function Survival Tool	019145	$2.99
#26: German Folding Stove w/ Fuel	003929	$2.99
Total:		$322.69
		-20%
Your Price:		$258.15

Total Bug Out Bag
Weight: 12.4lbs

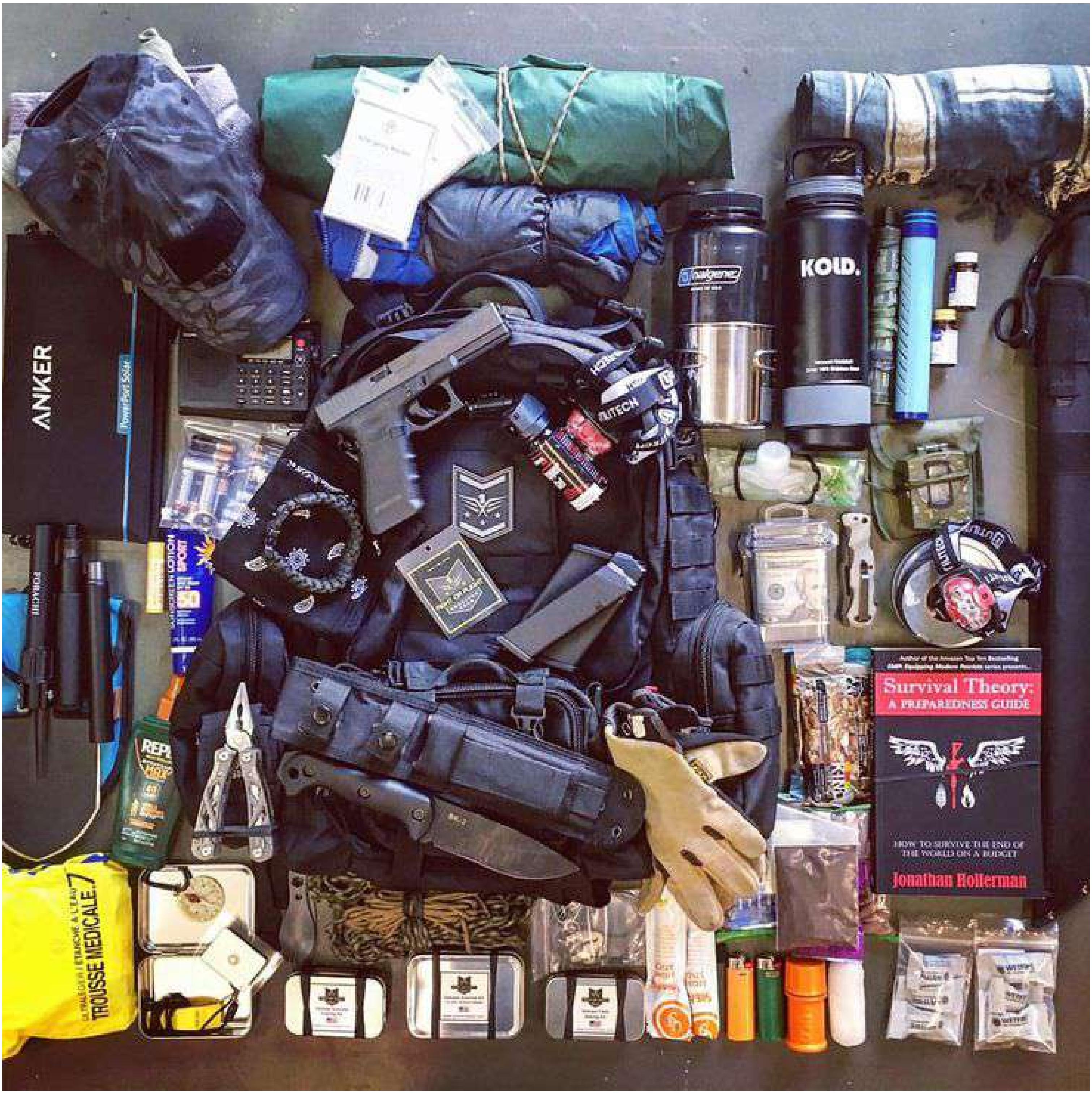
ANKER
PowerPort Solar
nalgene
KOLD.
SPORT
50
REPEL
TROUSSE MEDICALE .7
Survival Theory:
A PREPAREDNESS GUIDE
HOW TO SURVIVE THE END OF THE WORLD ON A BUDGET
Jonathan Hollerman

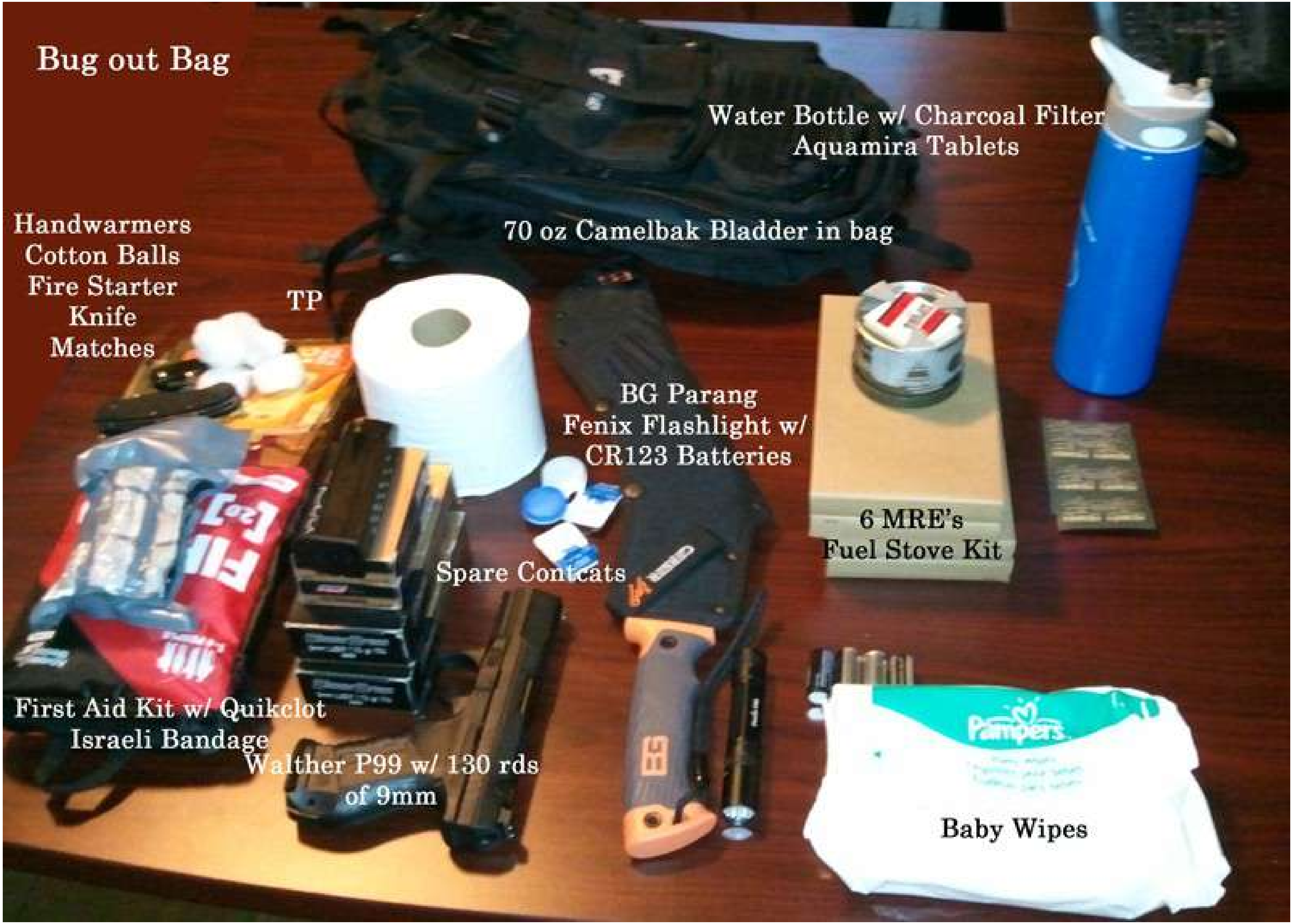

Bug out Bag

Water Bottle w/ Charcoal Filter
Aquamira Tablets

70 oz Camelbak Bladder in bag

Handwarmers
Cotton Balls
Fire Starter
Knife
Matches

TP

BG Parang
Fenix Flashlight w/
CR123 Batteries

6 MRE's
Fuel Stove Kit

Spare Contcats

First Aid Kit w/ Quikclot
Israeli Bandage

Walther P99 w/ 130 rds
of 9mm

Baby Wipes

6 EMERGENCY USES FOR PARACORD

Did you know?

A single length of paracord has been tested to handle 550 lbs of weight.

1. Snare Trap

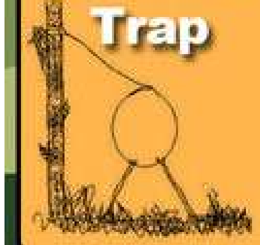

2. Fishing Line

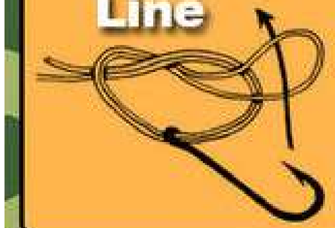

3. Shelter

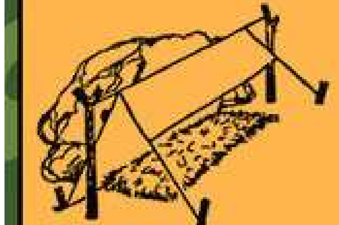

4. Fishing Net

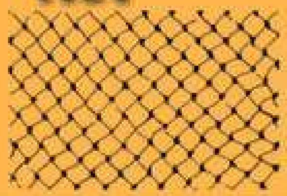

5. Fire

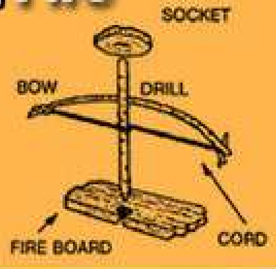
SOCKET
BOW
DRILL
FIRE BOARD
CORD

6. Bear Bag

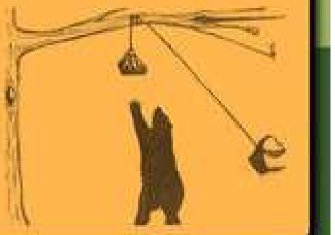

Did you know?

There are seven internal strands, each of which comes apart into two, so there's 14 thin lines.

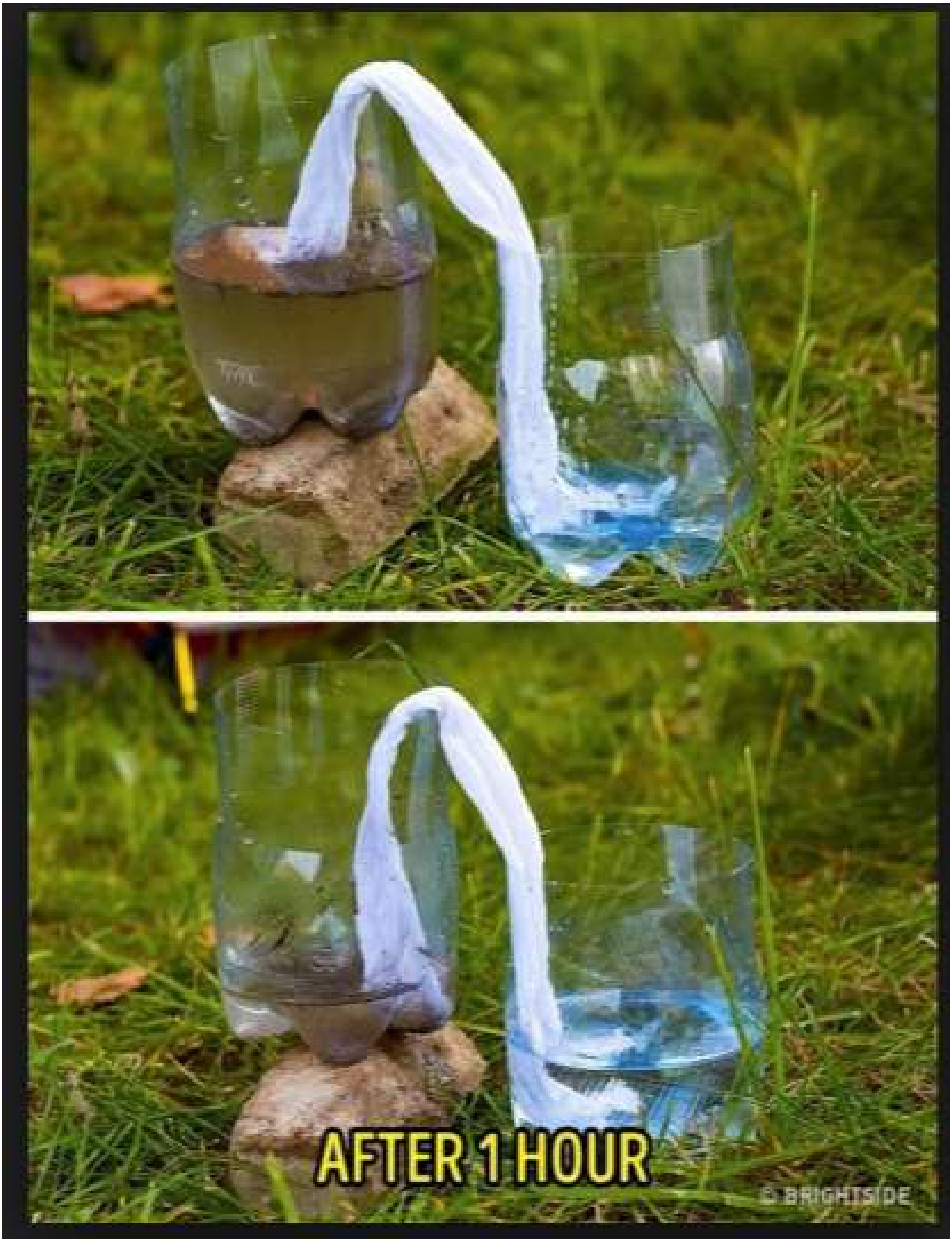
AFTER 1 HOUR
© BRIGHTSIDE

5 WAY TAG STRIP

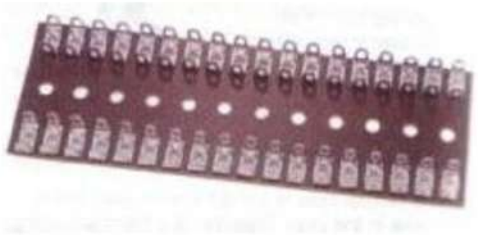

36 WAY TAG STRIP - TWO ROWS

e wired together ,with a little ingenuity, with the component wires being held together in the grip of solderless crocodile clips, whereby the conne
older.

ex circuits a plastic Terminal Block (sometimes referred to as a choc' or chocolate block) can be utilised very effectively indeed. These are used
15 Amp and 30 Amp. The 5 and 15 Amp Terminal Blocks I have found to be the most suitable. The various component wires can be trappe
makes it easy to change the components around when experiment with different circuits. See The EXPERMENTAL CRYSTAL SET for more de

'CHOCOLATE' TERMINAL BLOCK

Where to connect to the earbud jack.

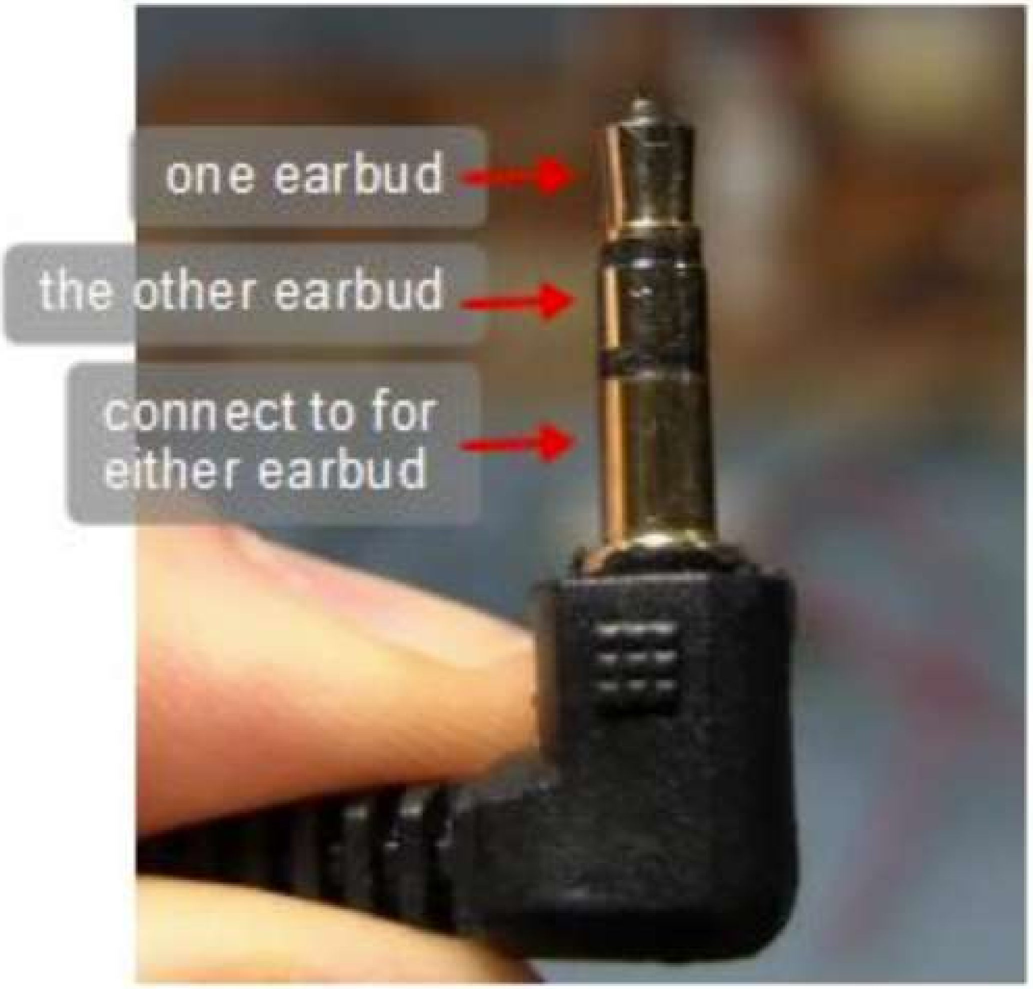
one earbud
the other earbud
connect to for either earbud

For connecting to one earbud only.

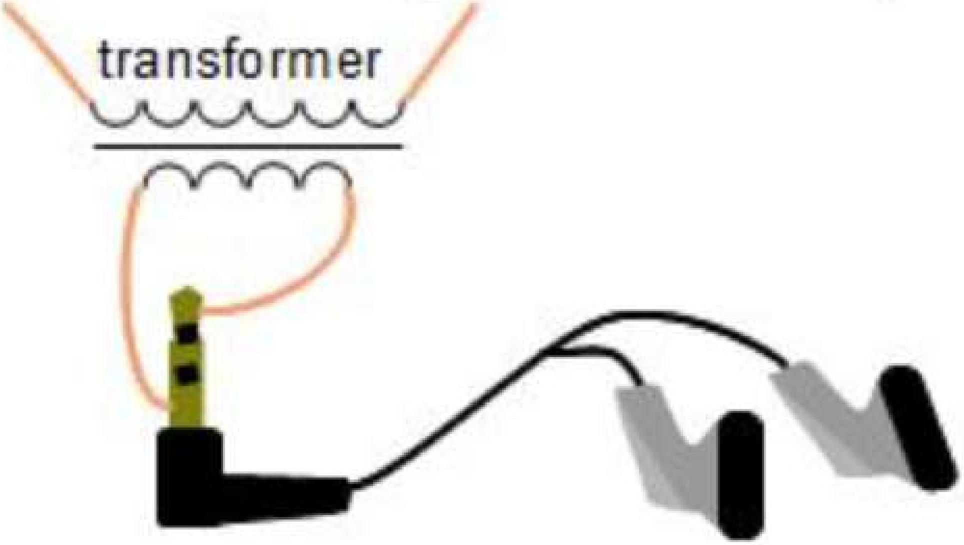
transformer

For connecting to both earbuds.

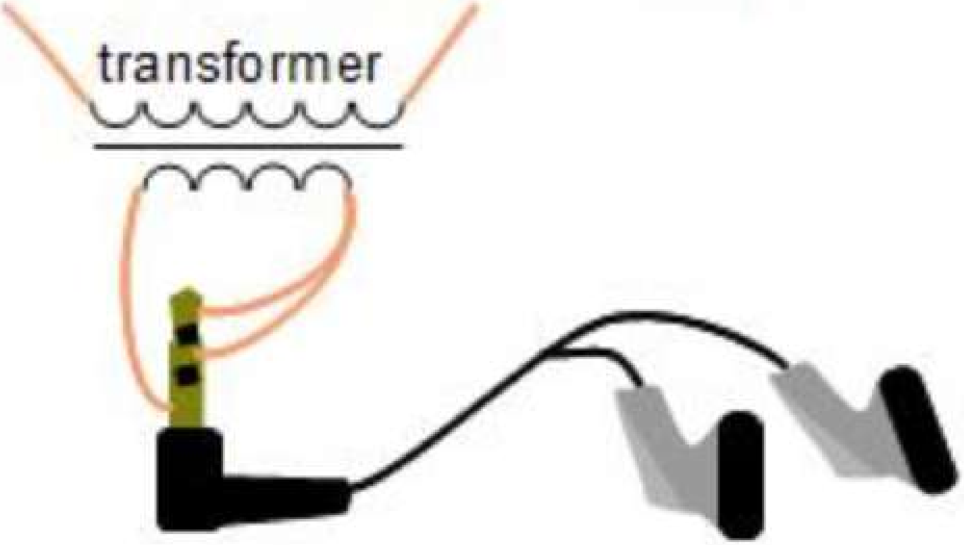
transformer

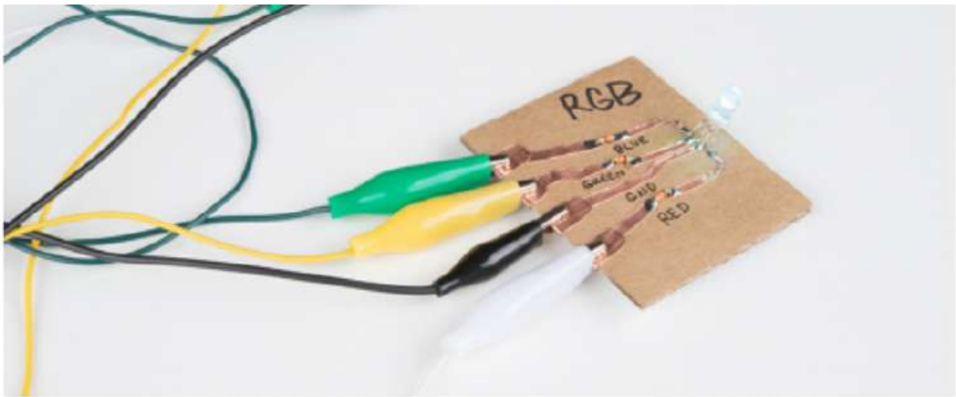
RGB
BLUE
GREEN
GND
RED

The current-limiting resistors are part of the cardboard circuit. The black spots are where I used the Bare Conductive paint

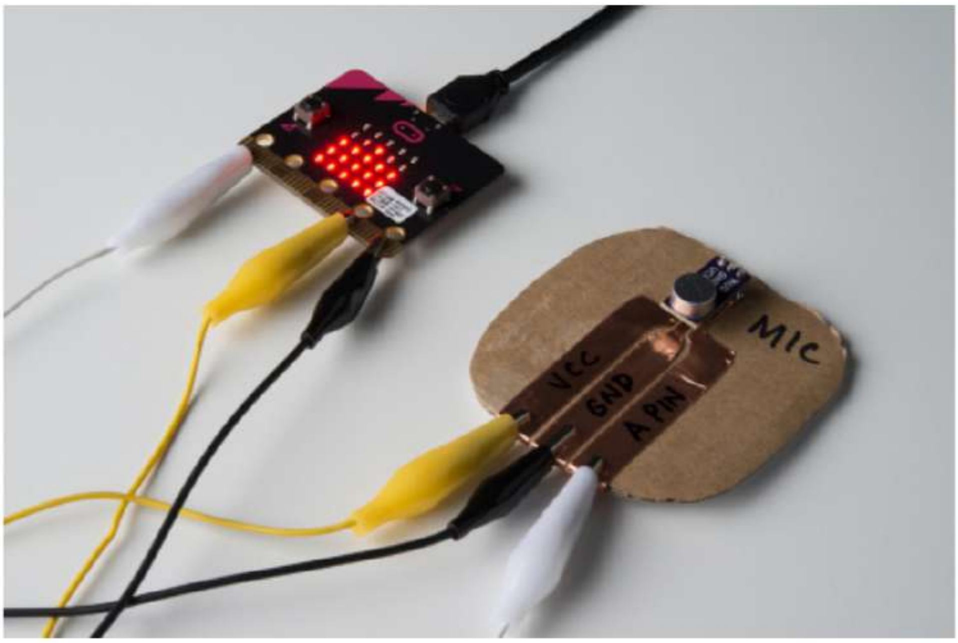
MIC
VCC
GND
A PIN

The micro:bit is plotting the analog values on the 5x5 LED matrix, which represent the noise level in the room.

D PIN
2 BUTTON BOARD
"ACTIVE LOW"
D PIN
GND
BARE ELECTRIC PAINT

All you need to get started is a roll of copper tape, a few components and conductive paint. One roll of tape and tube of paint can make about 200 circuits.

COSMIC VIBRATIONS

(THESE INTERVALS AND FIGURES ARE APPROXIMATE AND NOT EXACT)

Octave	Manifestation	Vibrations per second	Gaps
1	TOUCH	2	
2		4	
3		8	
4		16	
5	SOUND	32	
6		64	
7		128	
8		256	
9		512	
10		1,024	
11		2,048	
12		4,096	
13		8,192	
14		16,384	
15	ELECTROMAGNETIC (Longer Radio Waves)	32,768	
16		65,536	
17		131,072	
18		262,144	
19		524,288	
20		1,048,576	
21		2,097,152	
22		4,194,304	
23		8,388,608	
24		16,777,216	
25		33,554,432	
26		67,108,864	
27		134,217,728	
28		268,435,456	
29	ELECTROMAGNETIC (Microwaves)	536,870,912	
30		1,073,741,824	
31		2,147,483,648	
32		4,294,967,296	
33		8,589,934,592	
34		17,179,869,184	
35		34,359,738,368	
36		68,719,476,736	
37		137,438,953,472	
38		274,877,906,944	
39	HEAT OR INFRARED	549,755,813,888	
40		1,099,511,627,776	
41		2,199,023,255,552	
42		4,398,046,511,104	
43		8,796,093,022,208	
44		17,592,186,044,416	
45		35,184,372,088,832	
46		70,368,744,177,664	
47		140,737,488,355,328	
48		281,474,976,710,656	
49	VISIBLE LIGHT	562,949,953,421,312	
50		1,125,899,906,842,624	
51	ULTRAVIOLET	2,251,799,813,685,248	SPIRIT ELECTRONS
52		4,503,599,627,370,496	
53		9,007,199,254,740,992	
54		18,014,398,509,481,984	
55		36,028,797,018,963,968	
56		72,057,594,037,927,936	
57	X-RAY	144,115,188,075,855,872	
58		288,230,376,151,711,744	
59		576,460,752,303,423,488	
60			PSYCHIC PROJECTIONS
61			
62	GAMMA AND COSMIC RAYS	4,611,686,018,427,387,904	
63			
64		*The Vibrations in these octaves increase to an incomprehensible number per second.*	
65			
66			
67			
68			
69			
70		*The phenomena of psychic projections and soul essence occur in octaves considerably beyond those of any known material manifestations. Consequently, they must not be considered as immediately following any manifestations which we have shown on this chart.*	
71			
72			SOUL ESSENCE
73			
74			
75			
76			
77			
78			
79			
80			

NOTICE

This chart was carefully compiled by the Editorial and Research Departments of AMORC College after consulting the most eminent scientific authorities on the subjects of physics and chemistry and comparing their observations and notations with the reports from the Rosicrucian laboratories in foreign countries.

There is a slight disagreement in regard to the rates of vibrations for the various keys of the musical keyboard, due to different standards of pitch in various countries, and there is some disagreement among scientific authorities in regard to the colors associated with various musical notes. This chart represents not only the consensus of most of these authorities but also adheres strictly to what the Rosicrucians have been able to demonstrate as truth in our own laboratories and in the experiments of thousands of members in various parts of the world. Regardless, therefore, of any disagreement with authorities outside the organization, this chart is the authentic presentation of the Rosicrucian knowledge.

This chart is to be used by the members of the Rosicrucian Order in connection with their studies and experiments and is not to be used for any other purpose or in any other way. Like the lessons themselves, this chart and the facts contained on it are reserved and restricted in use under the rules and regulations of the organization.

INTERNATIONAL
SEAL OF THE
SUPREME LODGE
of the
Rosicrucian
Order

THE SOLAR SPECTRUM

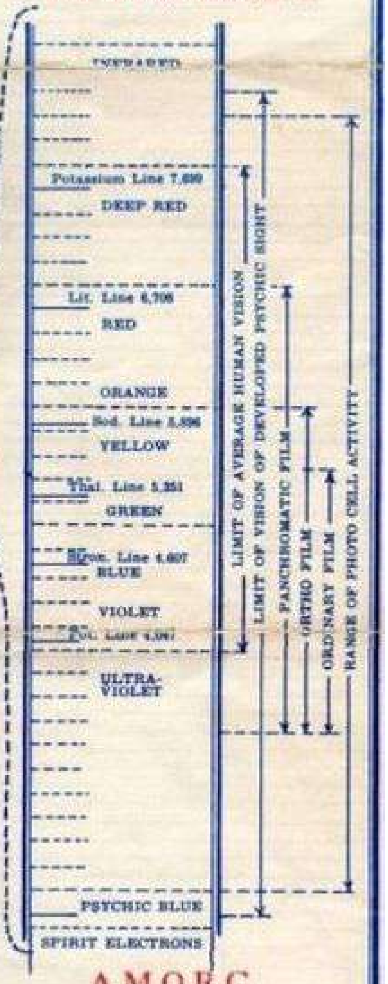

AMORC
THE
Rosicrucian Order
San Jose, California

LITHO. IN U. S. A.

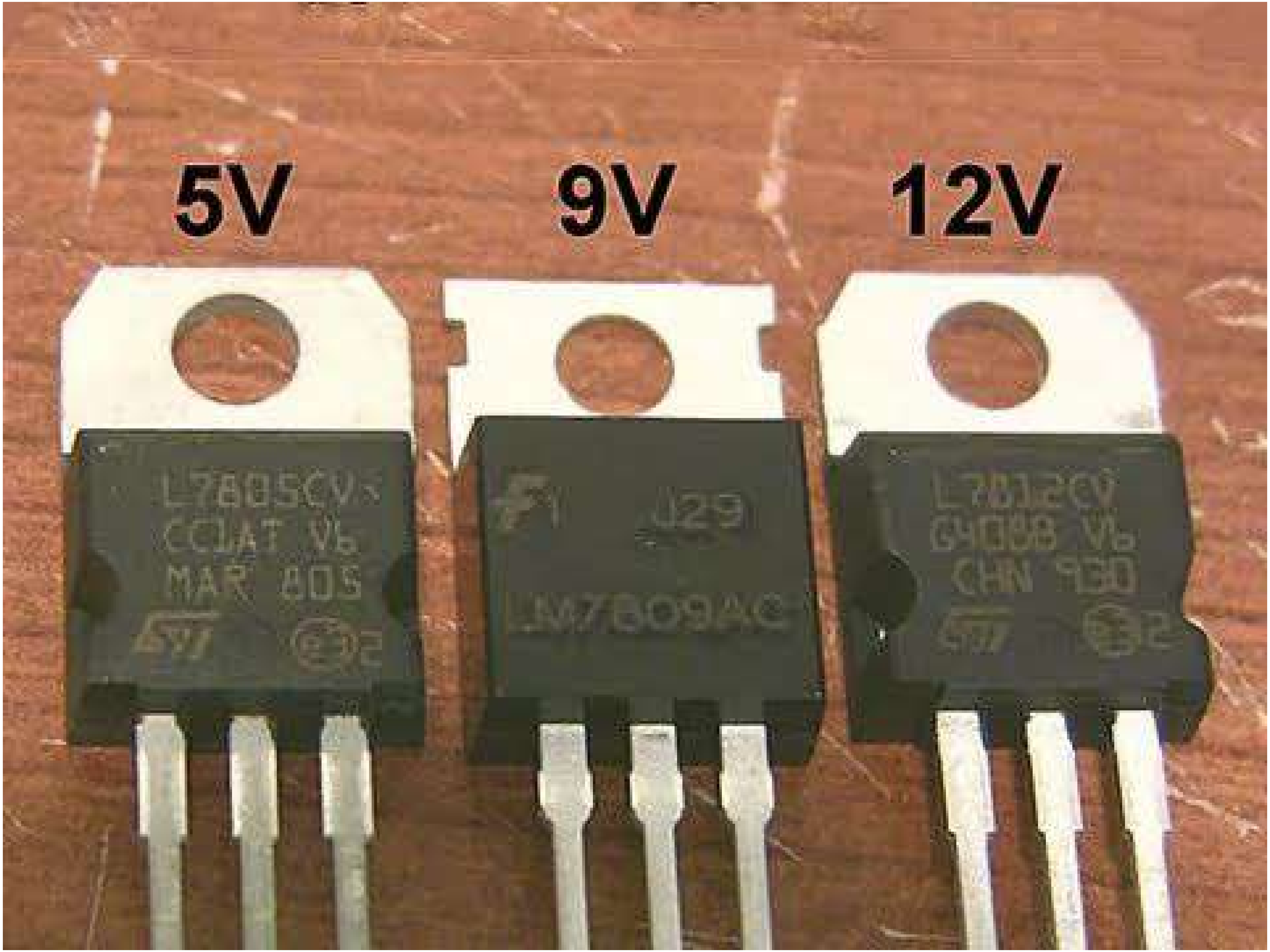
5V
9V
12V
L7805CV
CC1AT V6
MAR 80S
J29
LM7809AC
L7812CV
G4088 V6
CHN 930

7 LOGIC GATES & 4 USEFUL COMBINATIONS

Standard Logic Gates

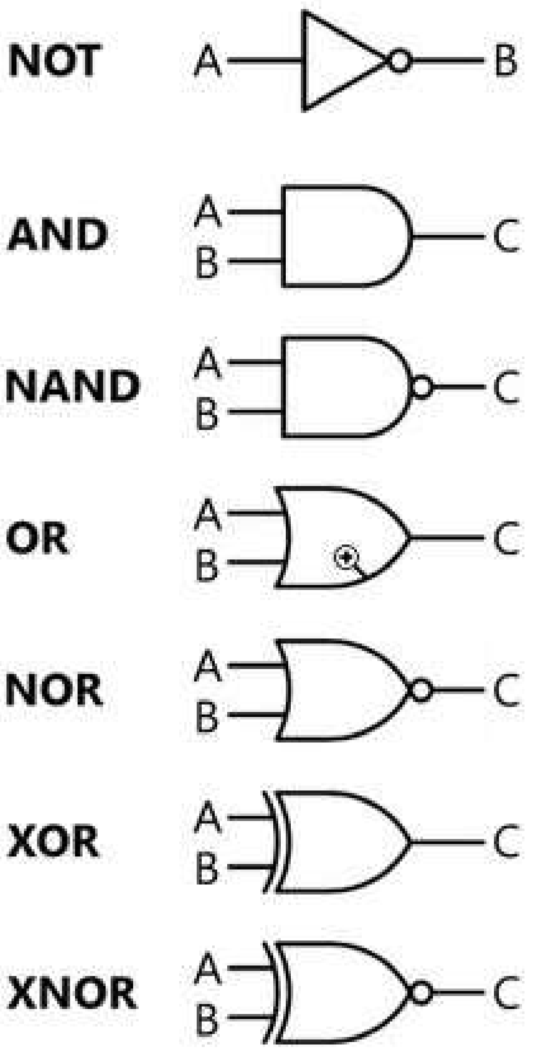
NOT
A
B
AND
A
B
C
NAND
A
B
C
OR
A
B
C
NOR
A
B
C
XOR
A
B
C
XNOR
A
B
C

NOT

A	B
1	0
0	1

AND

A	B	C
1	1	1
1	0	0
0	1	0
0	0	0

NAND

A	B	C
1	1	0
1	0	1
0	1	1
0	0	1

OR

A	B	C
1	1	1
1	0	1
0	1	1
0	0	0

NOR

A	B	C
1	1	0
1	0	0
0	1	0
0	0	1

XOR

A	B	C
1	1	0
1	0	1
0	1	1
0	0	0

XNOR

A	B	C
1	1	1
1	0	0
0	1	0
0	0	1

Preceding NOT Gate on One Input for AND, NAND, OR, NOR Gates

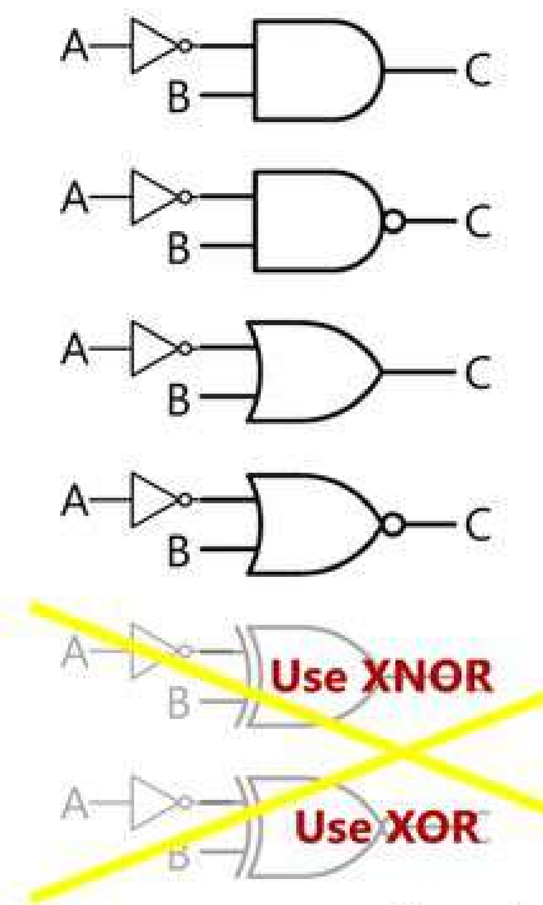
A
B
C
A
B
C
A
B
C
A
B
C
A
B
Use XNOR
A
B
Use XOR

A	B	C
1	1	0
1	0	0
0	1	1
0	0	0

A	B	C
1	1	1
1	0	1
0	1	0
0	0	1

A	B	C
1	1	1
1	0	0
0	1	1
0	0	1

A	B	C
1	1	0
1	0	1
0	1	0
0	0	0

A	B	C
~~1~~	~~1~~	~~1~~
~~1~~	~~0~~	~~0~~
~~0~~	~~1~~	~~0~~
~~0~~	~~0~~	~~1~~

A	B	C
~~1~~	~~1~~	~~0~~
~~1~~	~~0~~	~~1~~
~~0~~	~~1~~	~~1~~
~~0~~	~~0~~	~~0~~

Copyright 2012 PicoKit

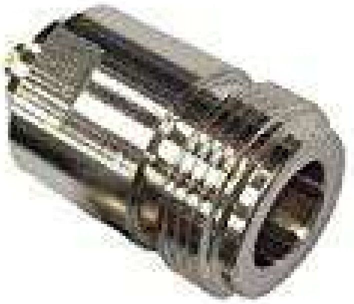

TYPE N F

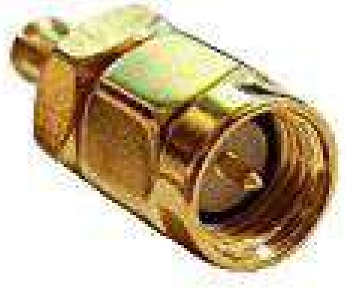

SMA M

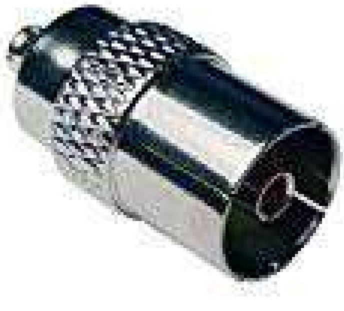

PAL/BELLING LEE F

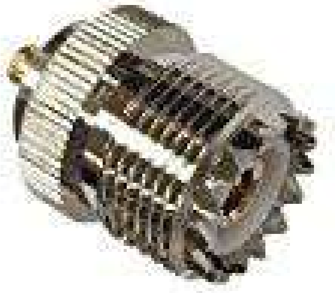

UHF F

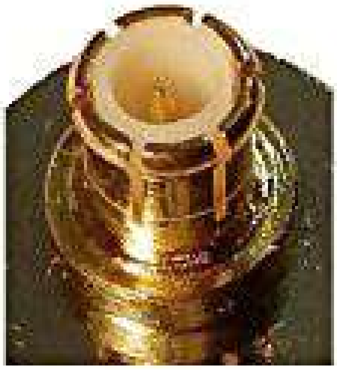

MCX MALE
(OTHER END)

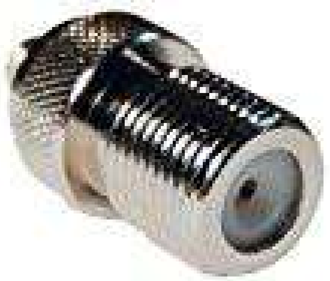

TYPE F F

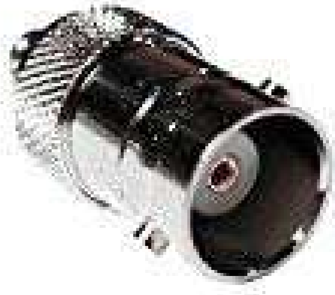

BNC F

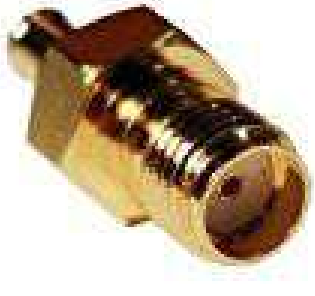

SMA F

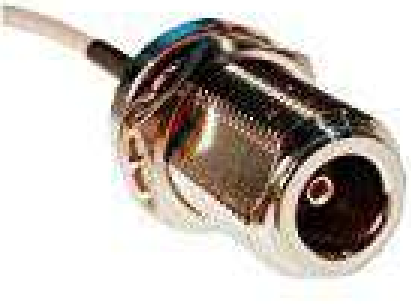

TYPE N F

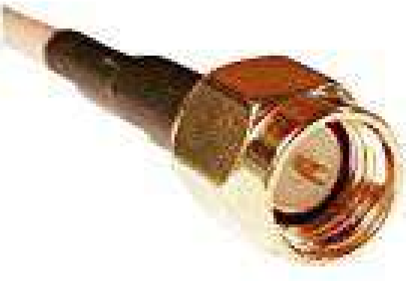

SMA M

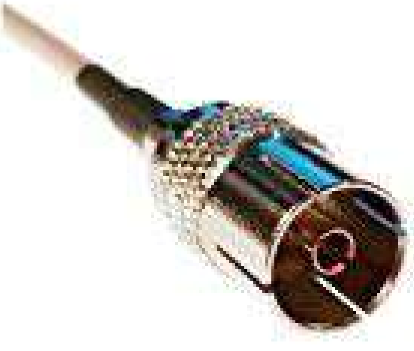

PAL/BELLING LEE F

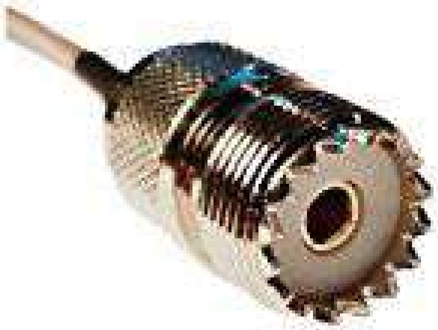

UHF F

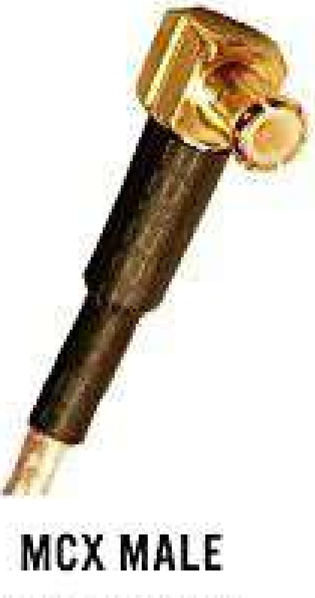

MCX MALE
(OTHER END)

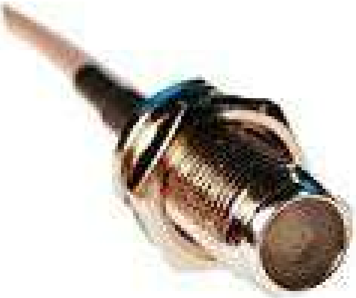

TYPE F F

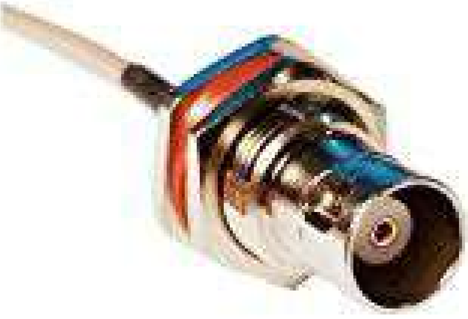

BNC F

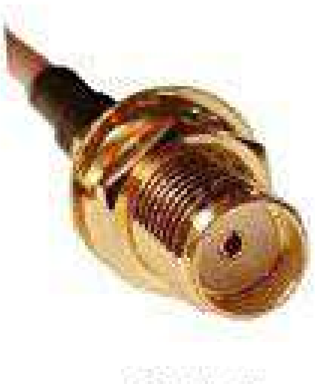

SMA F

DIODE

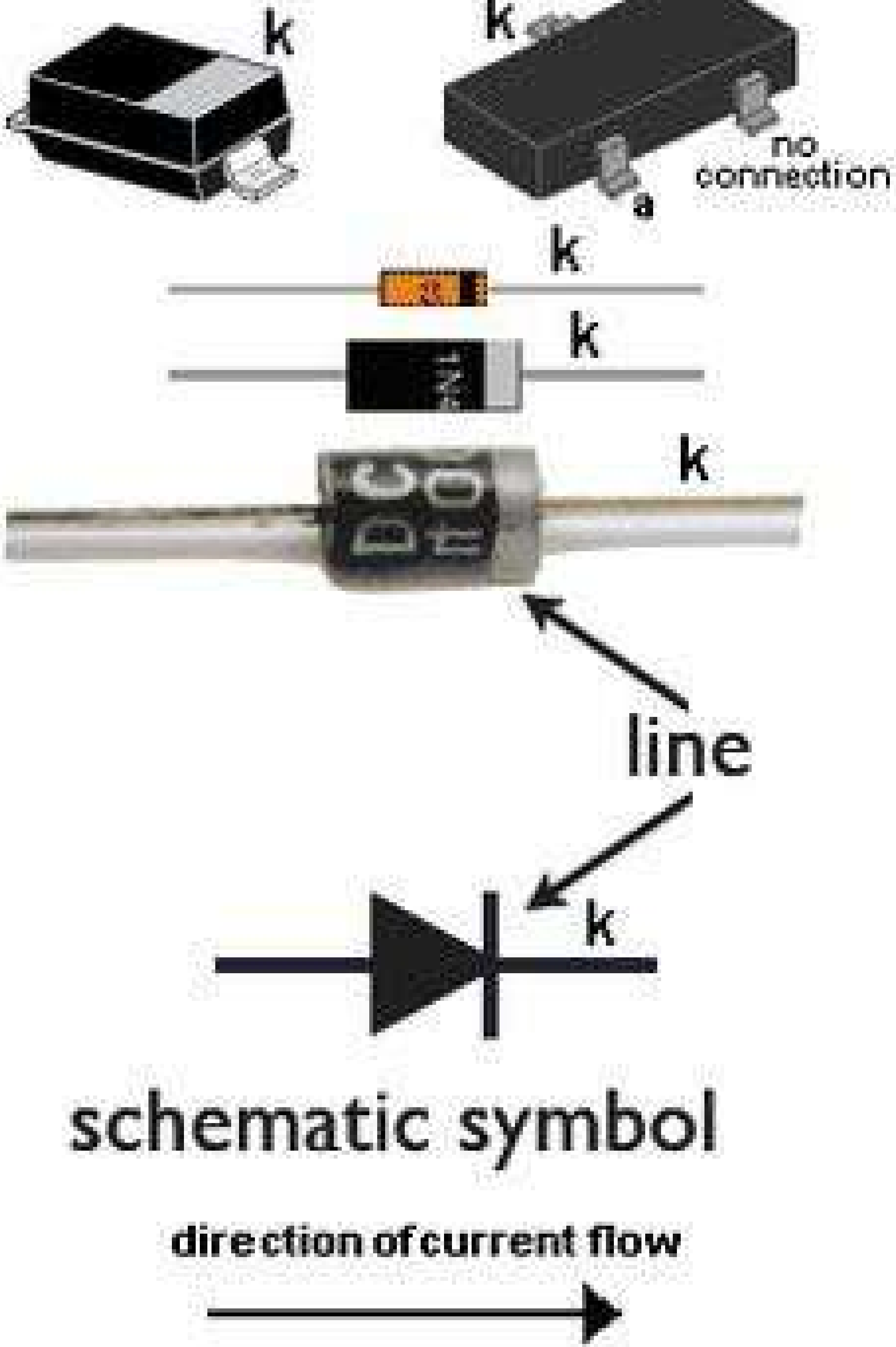
k
k
no
connection
a
k
k
k
line
k
schematic symbol
direction of current flow

7. 555 Timer Projects

In this section we will make :

- 7.1 4 Key Piano
- 7.2 Light sensitive music circuit
- 7.3 Light controlled Police Siren
- 7.4 Touch Switch
- 7.5 Timer
- 7.6 Continuity Tester
- 7.7 Knight Rider
- 7.8 Cricket Game
- 7.9 Multipurpose circuit
- 7.10 Johnson counter

Some Parts of this project are available in Video CD ROM

Internal diagram of 555

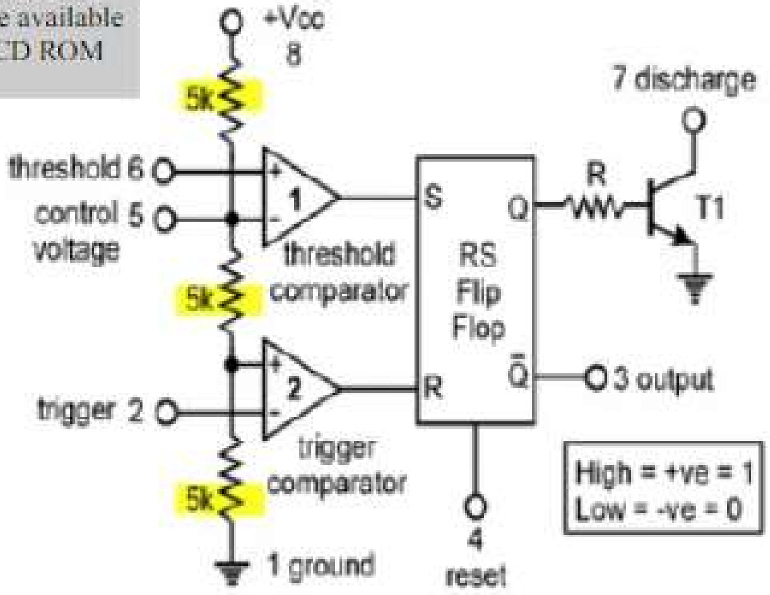
+Vcc
8
5k
7 discharge
threshold 6
control 5
voltage
1
threshold
comparator
S
Q
R
T1
RS
Flip
Flop
5k
2
trigger 2
R
Q̄
3 output
trigger
comparator
5k
High = +ve = 1
Low = -ve = 0
1 ground
4
reset

What is it about ?

555 timer integrated circuit (IC) is a very popular chip used in variety of applications like timer, pulse generation and oscillators. This is a low cost, stable and widely available chip which makes it favorite for hobbyists. The internal components of 555 as shown in figure consists of 2 comparators and a flip flop. All of these components contain 25 transistors and 15 resistors packed in the IC.

The three highlighted 5k resistors shown in figure are the reason why this IC is named as 555.

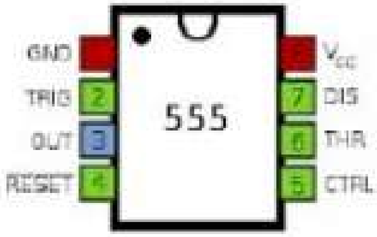
GND
TRIG
OUT
RESET
2
3
4
555
VCC
DIS
THR
CTRL
7
6
5

Please note the notch near first pin. This is made to indentify the first pin of IC.

Pin	Name	Purpose
1	GND	Ground reference voltage, low level (0 V)
2	TRIG	The OUT pin goes high and a timing interval starts when this input falls below 1/2 of CTRL voltage (which is typically 1/3 of V_{CC}, when CTRL is open).
3	OUT	This output is driven to approximately 1.7V below $+V_{CC}$ or GND.
4	RESET	A timing interval may be reset by driving this input to GND, but the timing does not begin again until RESET rises above approximately 0.7 volts. Overrides TRIG which overrides THR.
5	CTRL	Provides "control" access to the internal voltage divider (by default, 2/3 V_{CC}).
6	THR	The timing (OUT high) interval ends when the voltage at THR is greater than that at CTRL.
7	DIS	Open collector output which may discharge a capacitor between intervals. In phase with output.
8	V_{cc}	Positive supply voltage, which is usually between 3 and 15 V depending on the variation.

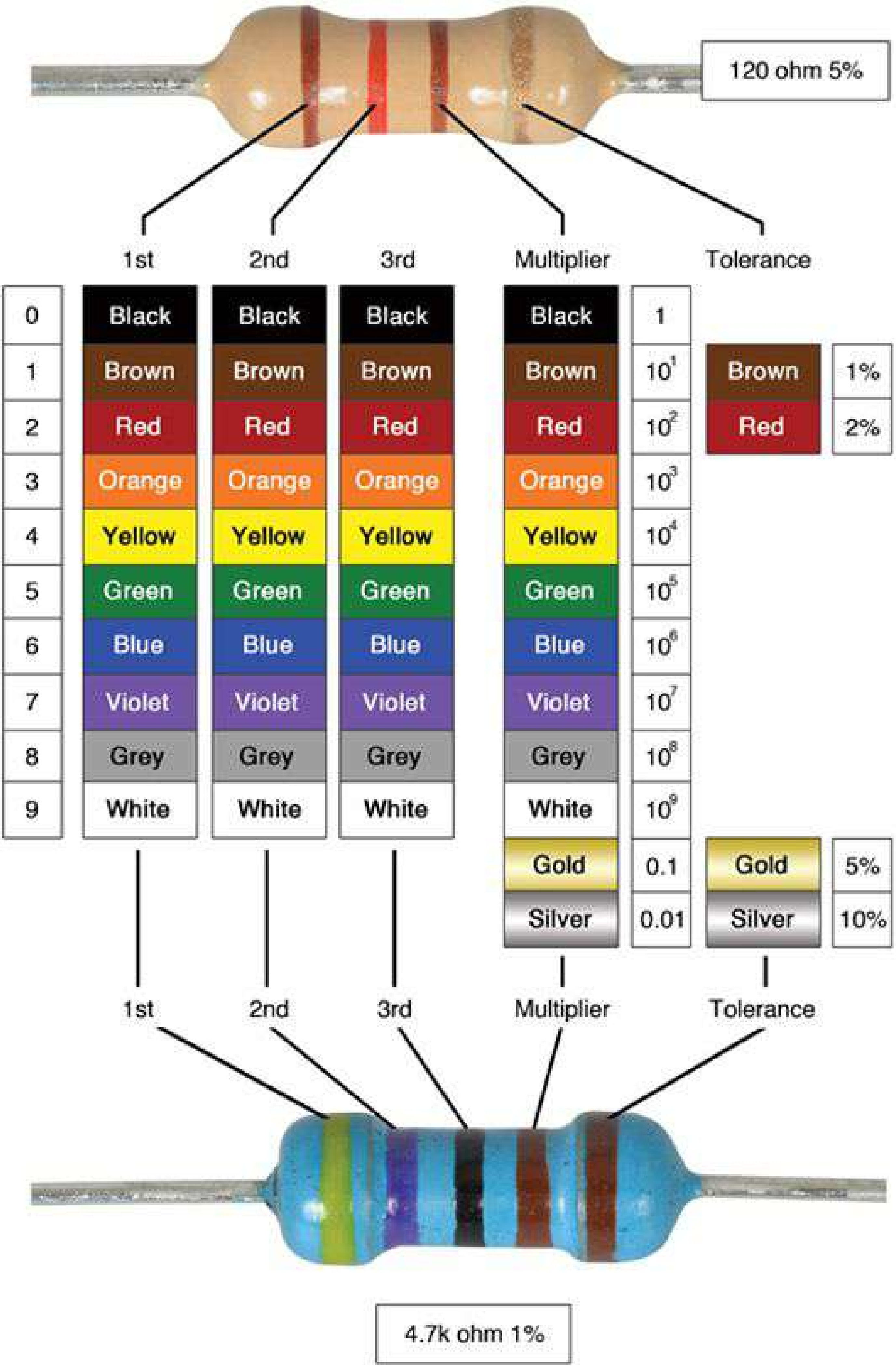

	1st	2nd	3rd	Multiplier		Tolerance	
0	Black	Black	Black	Black	1		
1	Brown	Brown	Brown	Brown	10^1	Brown	1%
2	Red	Red	Red	Red	10^2	Red	2%
3	Orange	Orange	Orange	Orange	10^3		
4	Yellow	Yellow	Yellow	Yellow	10^4		
5	Green	Green	Green	Green	10^5		
6	Blue	Blue	Blue	Blue	10^6		
7	Violet	Violet	Violet	Violet	10^7		
8	Grey	Grey	Grey	Grey	10^8		
9	White	White	White	White	10^9		
				Gold	0.1	Gold	5%
				Silver	0.01	Silver	10%

120 ohm 5%
1st
2nd
3rd
Multiplier
Tolerance
4.7k ohm 1%

ELECTROMAGNETIC RADIATION SPECTRUM

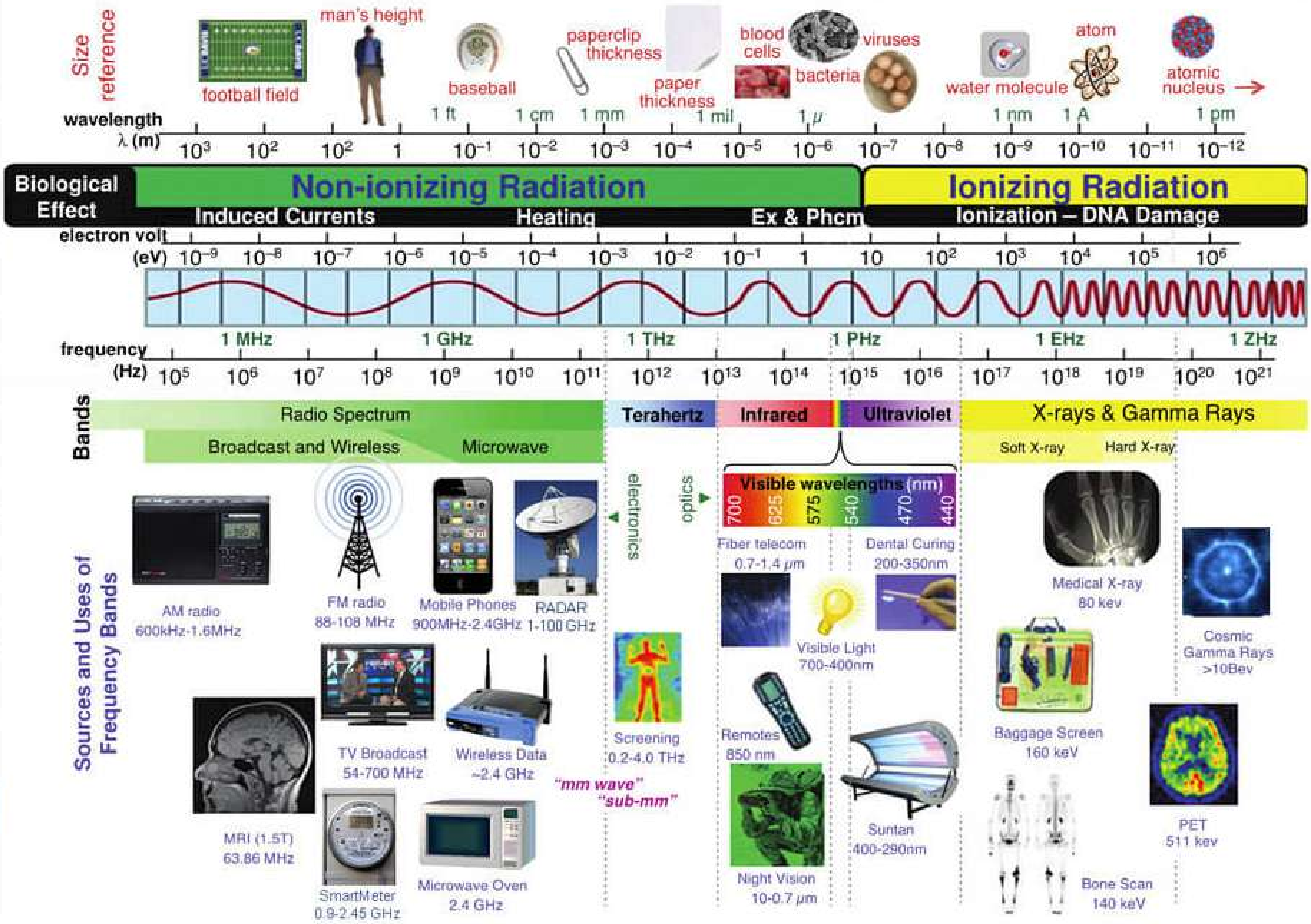
Size reference
football field
man's height
baseball
paperclip thickness
paper thickness
blood cells
viruses
bacteria
water molecule
atom
atomic nucleus →
wavelength λ (m)
1 ft
1 cm
1 mm
1 mil
1 μ
1 nm
1 A
1 pm
10³ 10² 10² 1 10⁻¹ 10⁻² 10⁻³ 10⁻⁴ 10⁻⁵ 10⁻⁶ 10⁻⁷ 10⁻⁸ 10⁻⁹ 10⁻¹⁰ 10⁻¹¹ 10⁻¹²
Biological Effect
Non-ionizing Radiation
Ionizing Radiation
Induced Currents
Heating
Ex & Phcm
Ionization – DNA Damage
electron volt (eV)
10⁻⁹ 10⁻⁸ 10⁻⁷ 10⁻⁶ 10⁻⁵ 10⁻⁴ 10⁻³ 10⁻² 10⁻¹ 1 10 10² 10³ 10⁴ 10⁵ 10⁶
frequency (Hz)
1 MHz
1 GHz
1 THz
1 PHz
1 EHz
1 ZHz
10⁵ 10⁶ 10⁷ 10⁸ 10⁹ 10¹⁰ 10¹¹ 10¹² 10¹³ 10¹⁴ 10¹⁵ 10¹⁶ 10¹⁷ 10¹⁸ 10¹⁹ 10²⁰ 10²¹
Bands
Radio Spectrum
Broadcast and Wireless
Microwave
Terahertz
Infrared
Ultraviolet
X-rays & Gamma Rays
Soft X-ray
Hard X-ray
electronics
optics
Visible wavelengths (nm)
700 625 575 540 470 440
Sources and Uses of Frequency Bands
AM radio
600kHz-1.6MHz
FM radio
88-108 MHz
Mobile Phones
900MHz-2.4GHz
RADAR
1-100 GHz
MRI (1.5T)
63.86 MHz
TV Broadcast
54-700 MHz
Wireless Data
~2.4 GHz
SmartMeter
0.9-2.45 GHz
Microwave Oven
2.4 GHz
Screening
0.2-4.0 THz
"mm wave"
"sub-mm"
Fiber telecom
0.7-1.4 μm
Dental Curing
200-350nm
Visible Light
700-400nm
Remotes
850 nm
Night Vision
10-0.7 μm
Suntan
400-290nm
Medical X-ray
80 kev
Cosmic Gamma Rays
>10Bev
Baggage Screen
160 keV
Bone Scan
140 keV
PET
511 kev

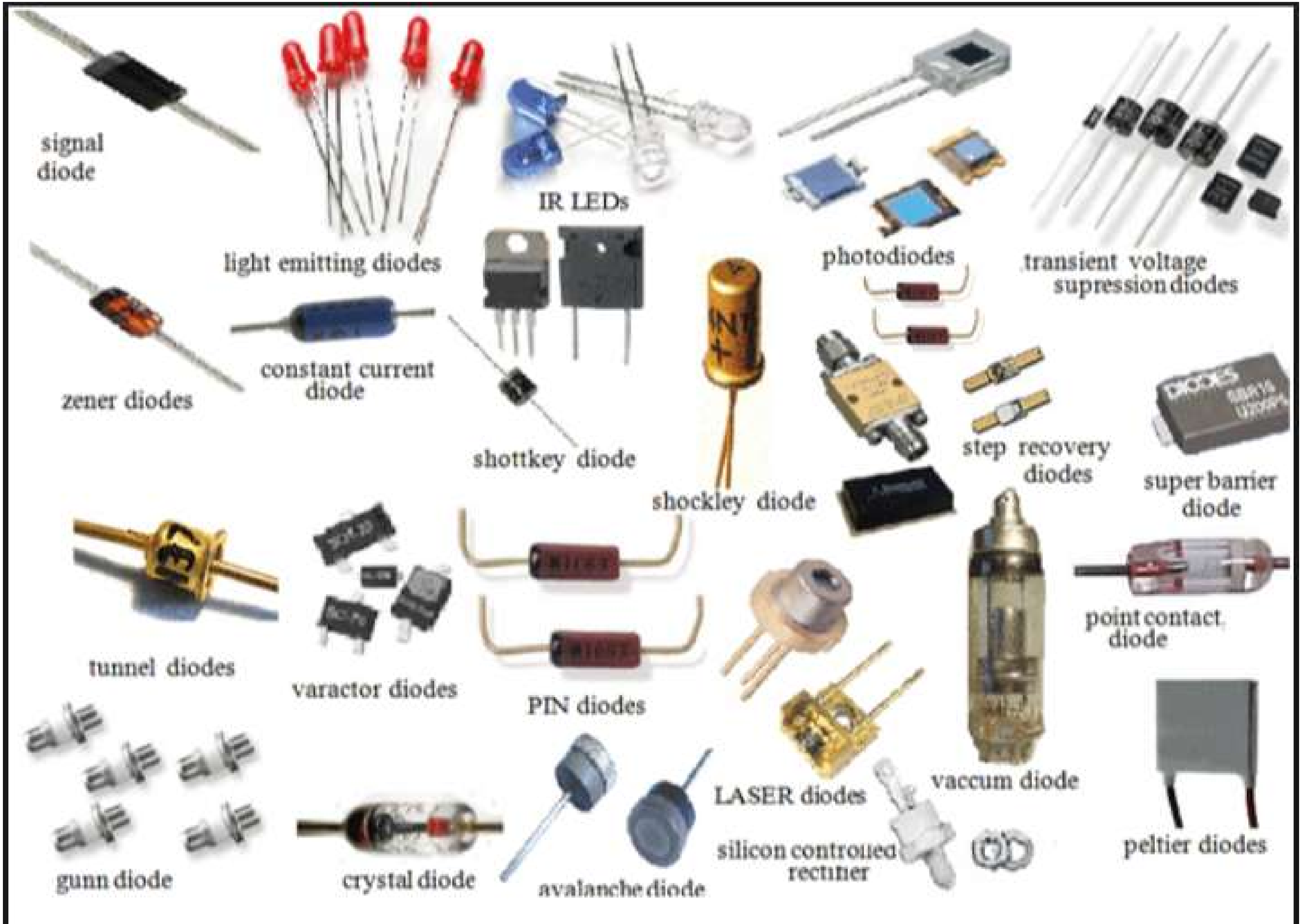
signal diode
light emitting diodes
IR LEDs
photodiodes
transient voltage supression diodes
zener diodes
constant current diode
shottkey diode
shockley diode
step recovery diodes
super barrier diode
tunnel diodes
varactor diodes
PIN diodes
point contact diode
vaccum diode
LASER diodes
gunn diode
crystal diode
avalanche diode
silicon controued rectifier
peltier diodes

Types of Diode

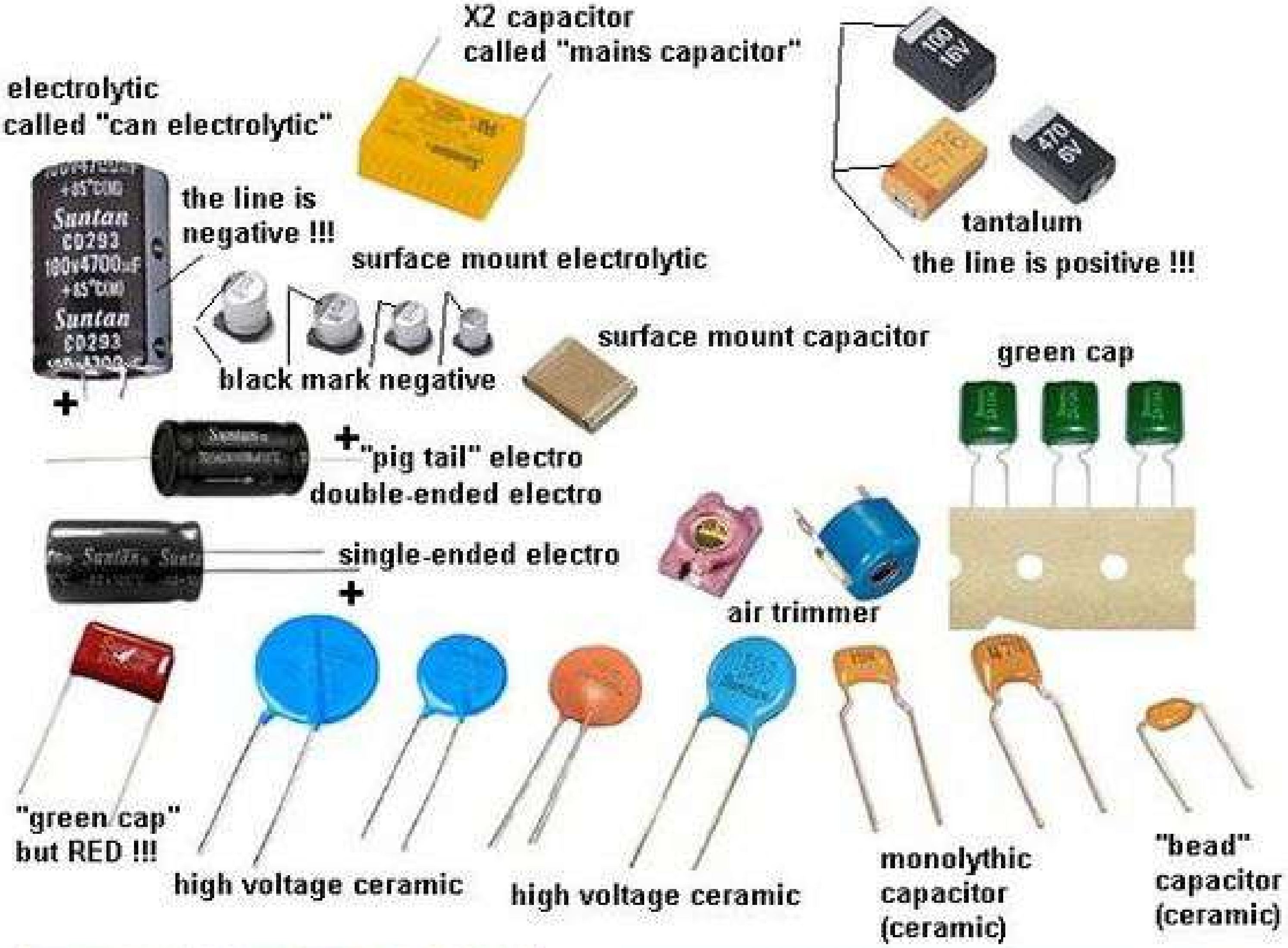
X2 capacitor
called "mains capacitor"
electrolytic
called "can electrolytic"
the line is
negative !!!
surface mount electrolytic
tantalum
the line is positive !!!
black mark negative
surface mount capacitor
green cap
+
+"pig tail" electro
double-ended electro
single-ended electro
+
air trimmer
"green cap"
but RED !!!
high voltage ceramic
high voltage ceramic
monolythic
capacitor
(ceramic)
"bead"
capacitor
(ceramic)

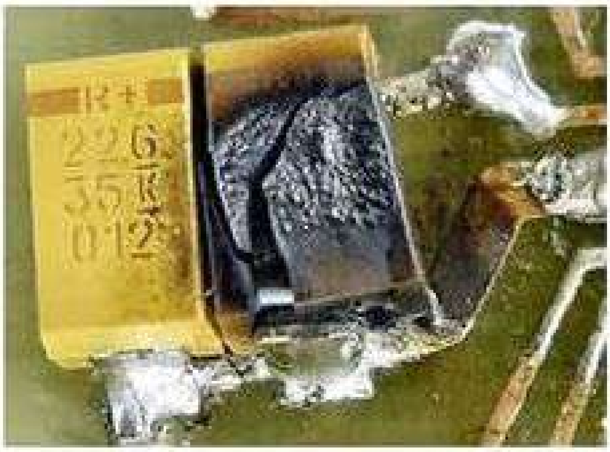

tantalum capacitors can
"go up in smoke"

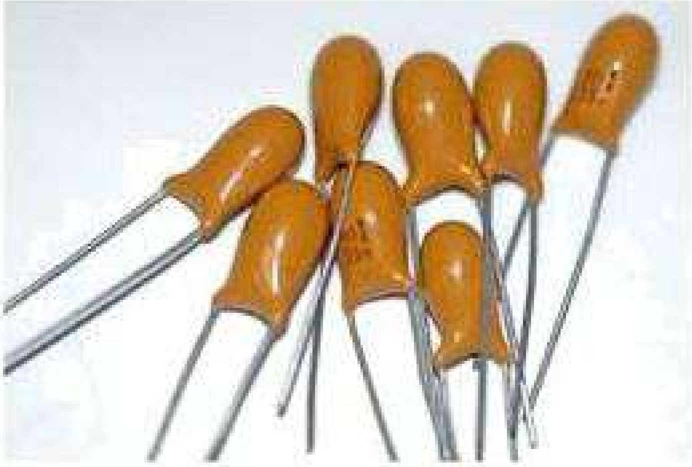

"bead" tantalum

Electronics
Reference Sheet v1.1b

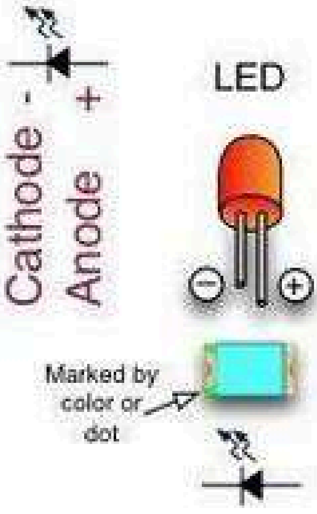

Cathode
-
Anode
+
LED
Marked by color or dot

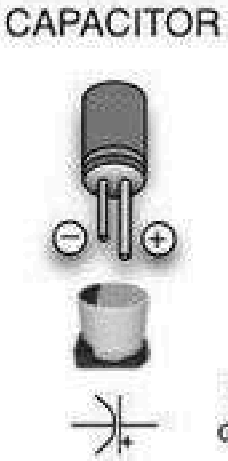

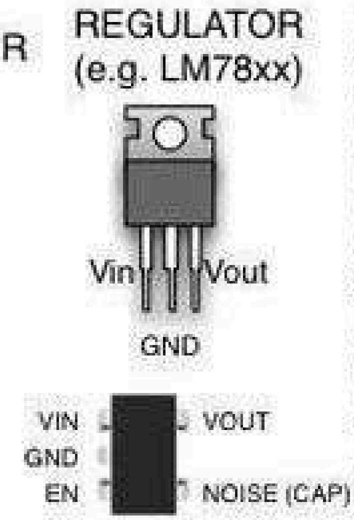

REGULATOR
(e.g. LM78xx)
Vin
Vout
GND
VIN
GND
EN
VOUT
NOISE (CAP)

Resistor

Resistor	Digit	Multiplier	Tolerance
Silver	-	0.01	±10%
Gold	-	0.1	±5%
Black	0	1	-
Brown	1	10	±1%
Red	2	100	±2%
Orange	3	1k	-
Yellow	4	10k	-
Green	5	100k	±0.5%
Blue	6	1M	±0.25%
Violet	7	10M	±0.1%
Gray	8	-	-
White	9	-	-

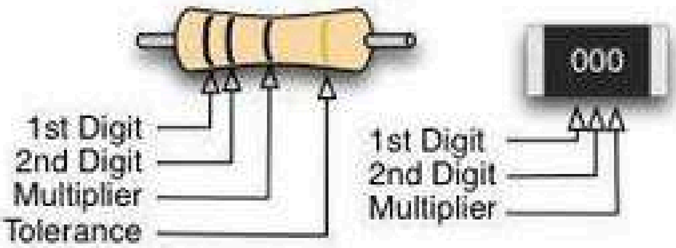

1st Digit
2nd Digit
Multiplier
Tolerance
000
1st Digit
2nd Digit
Multiplier

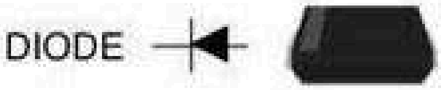

NPN transistor (Current sink)
(e.g. PN2222)

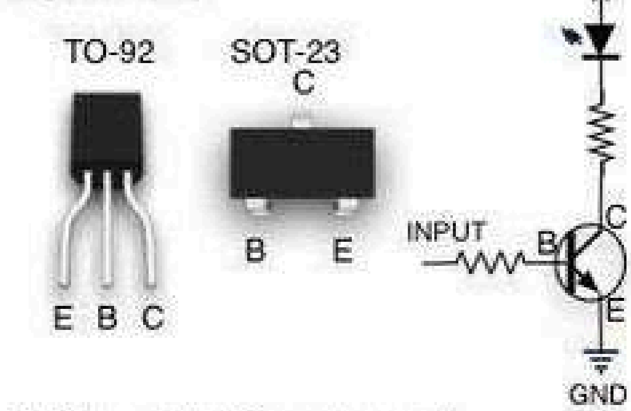

TO-92
E B C
SOT-23
C
B
E
VCC
INPUT
B
C
E
GND

N-channel MOSFET

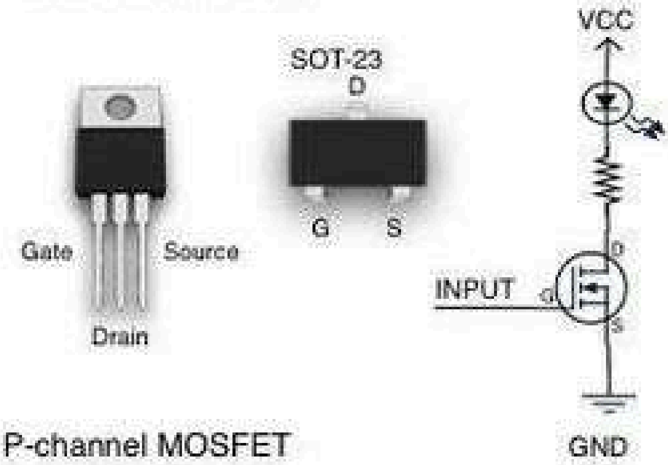

Gate
Source
Drain
SOT-23
D
G
S
VCC
D
INPUT
G
S
GND

PNP transistor (Current source)
(e.g. PN2907)

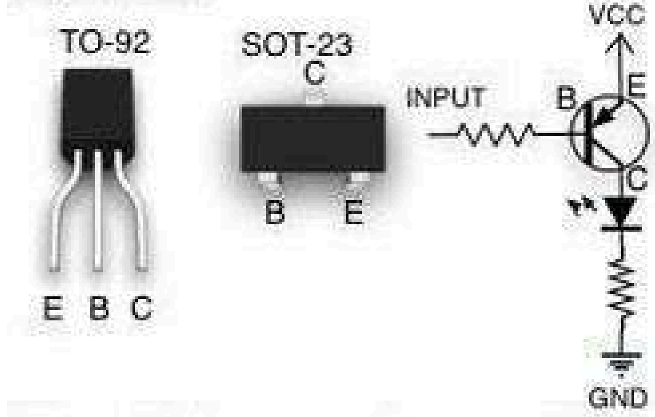

TO-92
E B C
SOT-23
C
B
E
VCC
INPUT
B
E
C
GND

P-channel MOSFET

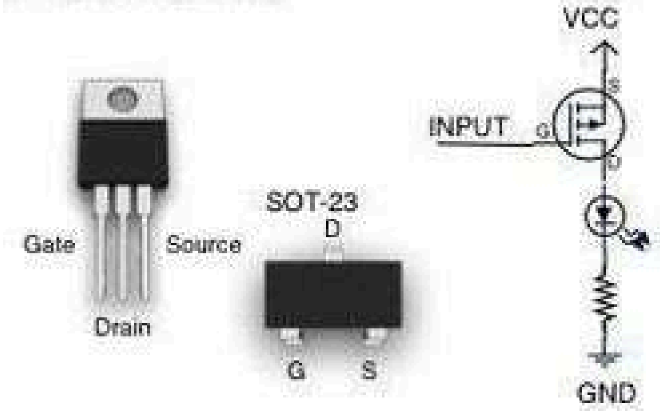

Gate
Source
Drain
SOT-23
D
G
S
VCC
S
INPUT
G
D
GND

* Please note that some components may have a different pinout than the one showed above, you should always check the data sheet before using a new component.

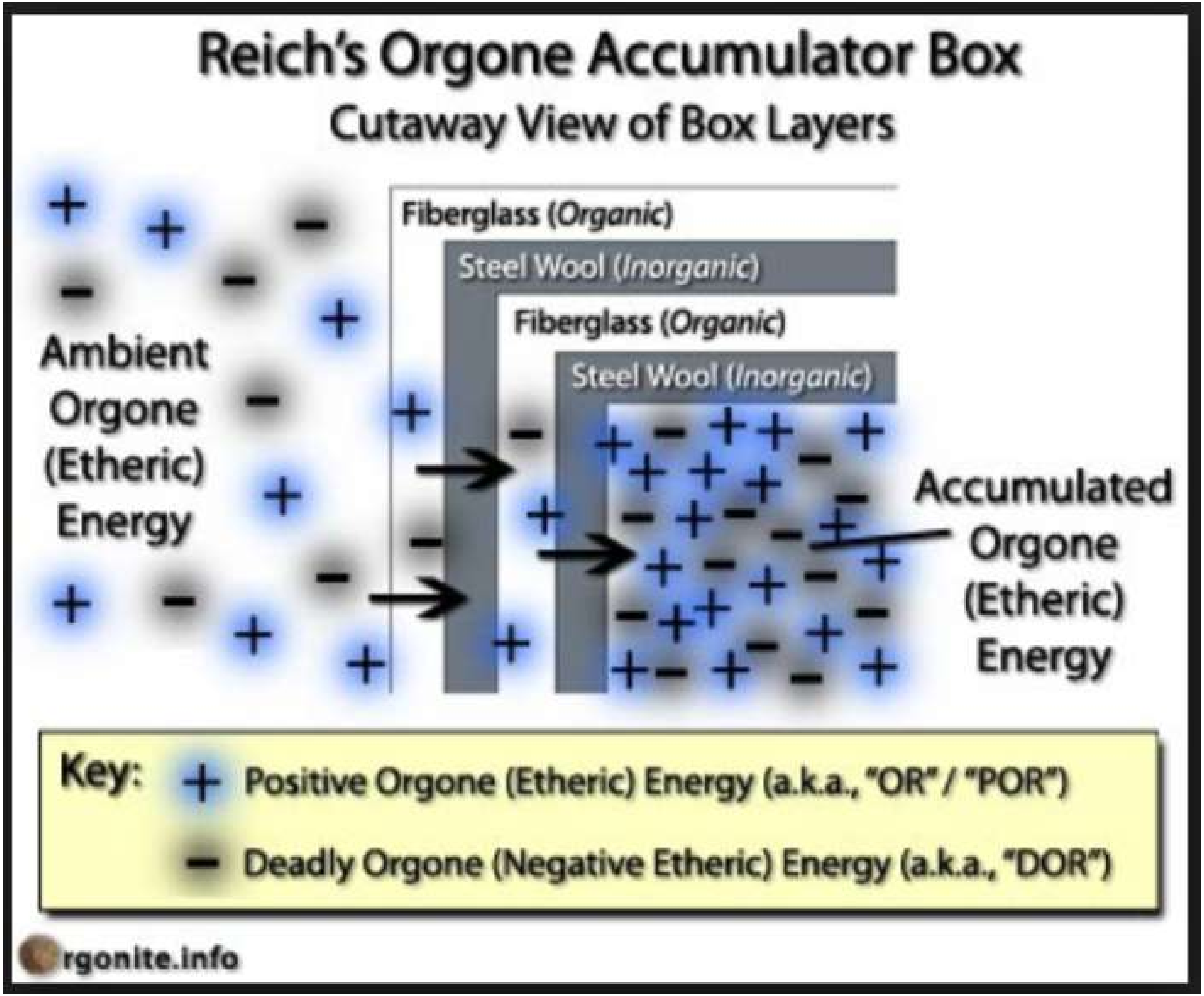
Reich's Orgone Accumulator Box
Cutaway View of Box Layers
Fiberglass (Organic)
Steel Wool (Inorganic)
Fiberglass (Organic)
Steel Wool (Inorganic)
Ambient Orgone (Etheric) Energy
Accumulated Orgone (Etheric) Energy
Key:
+ Positive Orgone (Etheric) Energy (a.k.a., "OR" / "POR")
− Deadly Orgone (Negative Etheric) Energy (a.k.a., "DOR")
rgonite.info

How to build a Don Croft ORGONITE CLOUD BUSTER

Refer to (Part 1 - 2) How to Make an Orgonite CloudBuster:
http://www.youtube.com/watch?v=cCn1DO2Rj6k

Materials List:
Wear safety gear as necessary
- respirator, gloves when pouring resin.
- gloves & goggles when cutting wood - with a respirator if cutting MDF etc.

4 x 10' lengths of copper (or other) pipe (cut 3 lengths in half, and one length into six 1' pieces)

6 pipe endcaps and 6 pipe connectors

2 gallon bucket (polypropylene if possible)

10 double terminated quartz (6 of these must fit into garden hose for insertion into base of pipes)

6 pieces of garden hose (3/4" - 1" long)

Resin and metal suitable for orgonite

A roll of tape (duct tape etc)

Glue (silicone glue will work)

Stick for mixing metal into resin when pouring layers

Top, middle (temporary) and bottom spacers cut and drilled out of wood. If not buying readymade, use templates or measure with copmass, ruler, pencil.
Drill the holes with 1 1/8" and 1 1/4" wood drill bits, and cut into shape with a Jigsaw.

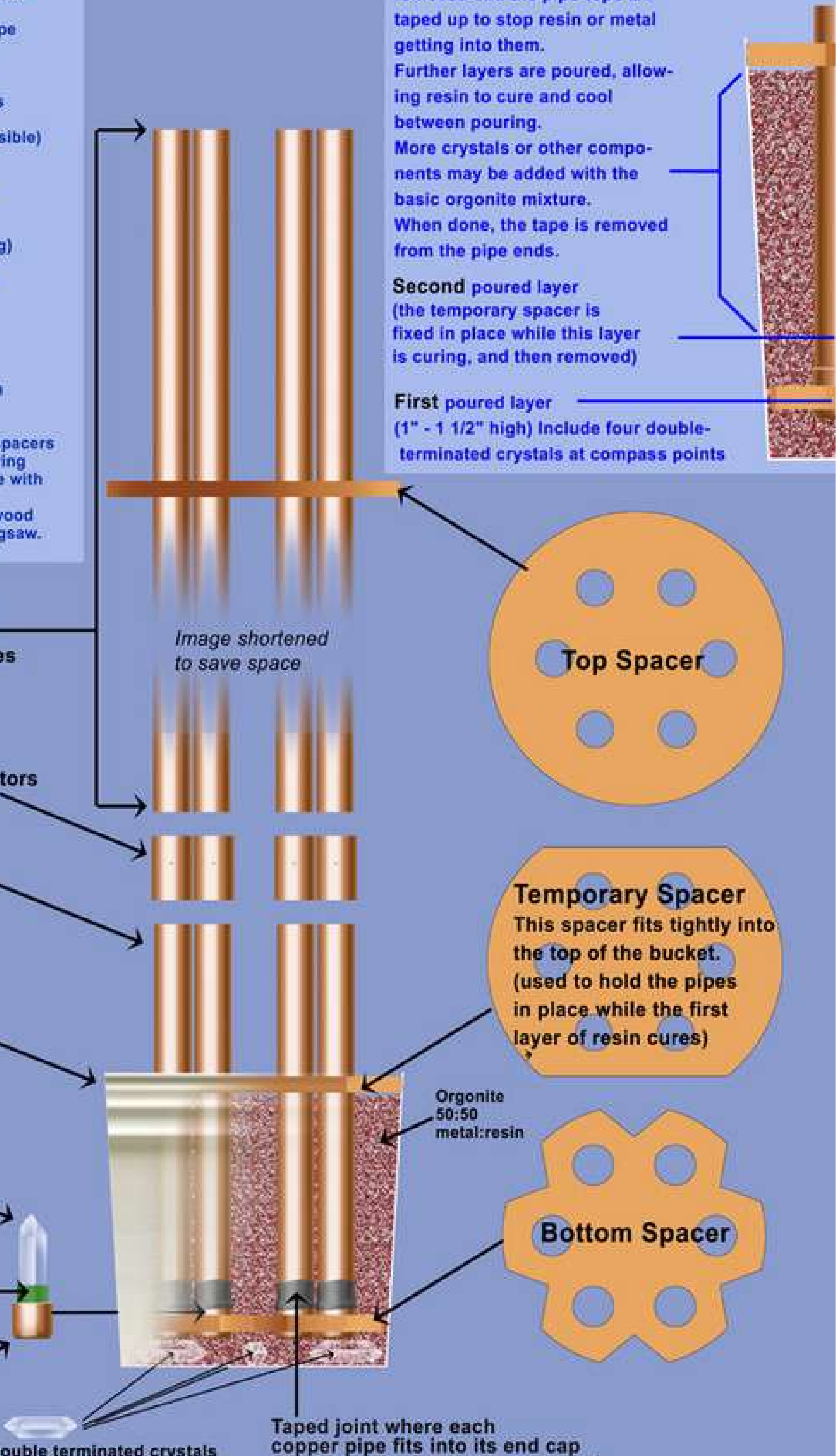

Third - The temporary spacer is removed and the pipe-tops are taped up to stop resin or metal getting into them.
Further layers are poured, allowing resin to cure and cool between pouring.
More crystals or other components may be added with the basic orgonite mixture.
When done, the tape is removed from the pipe ends.

Second poured layer
(the temporary spacer is fixed in place while this layer is curing, and then removed)

First poured layer
(1" - 1 1/2" high) Include four double-terminated crystals at compass points

Six 5' lengths of copper pipe
When the orgonite base of the cloud buster is complete, fix the top pipes to the bottom ones using pipe connectors, then fix the top spacer in place to provide stability.

Pipe connectors

Six 1' lengths of copper pipe
(1" diameter pipe)

Temporary Spacer
This spacer fits tightly into the top of the bucket. (used to hold the pipes in place while the first layer of resin cures)

2 Gallon ......plastic bucket
(8" bottom-interior diameter, 9" top interior diameter)

Orgonite
50:50
metal:resin

Double Treminated Crystal
(secured into hose with glue)

Section of garden hose 3/4" - 1" piece
(glued into end cap with silicone glue)

Copper end cap

Four double terminated crystals in the base are aligned with the cardinal directions

Taped joint where each copper pipe fits into its end cap (duct tape or any kind to stop resin seeping into endcap)

BOAT
BOAT
GOOP
HOUSEHOLD

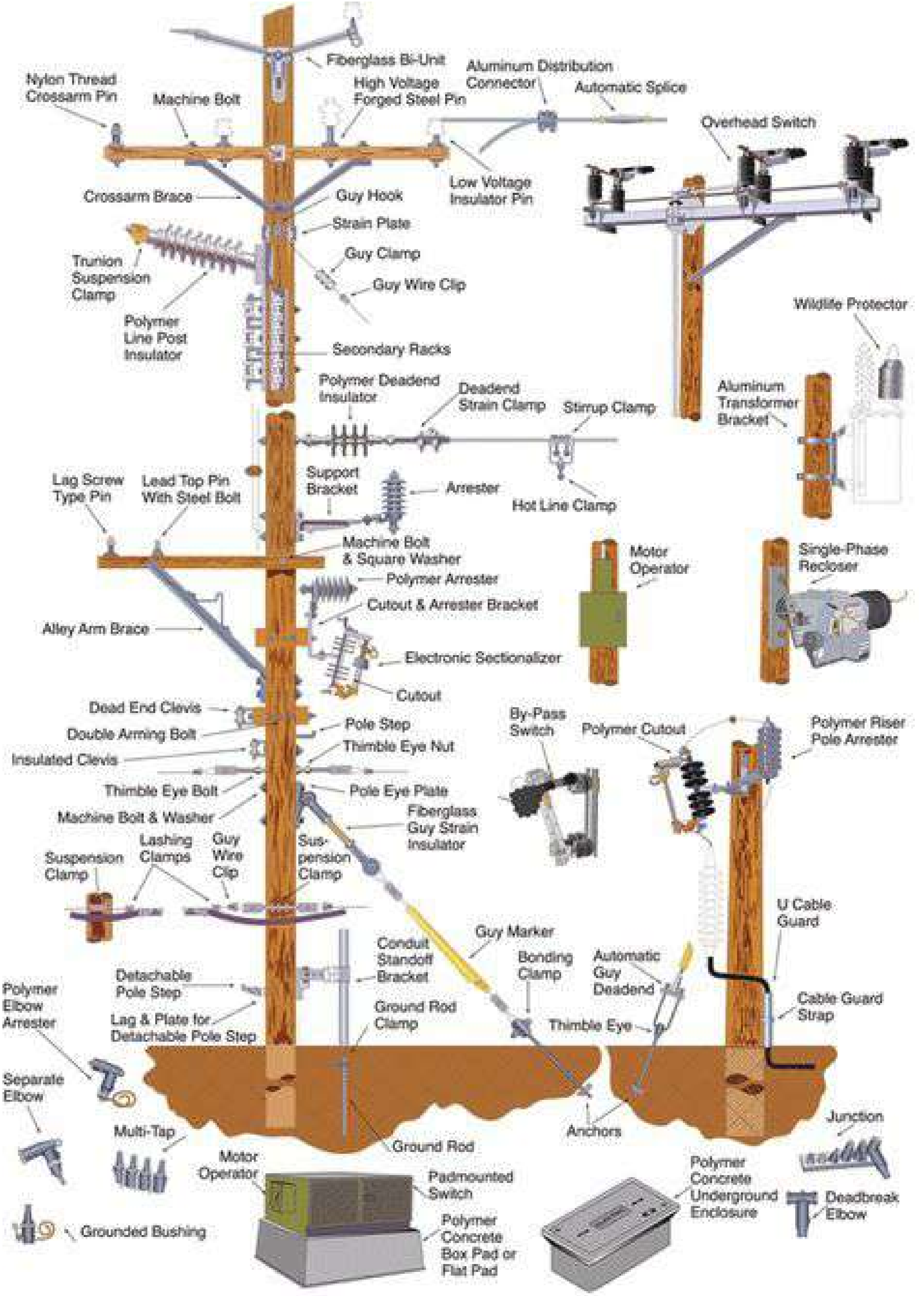
Fiberglass Bi-Unit
Nylon Thread Crossarm Pin
Machine Bolt
High Voltage Forged Steel Pin
Aluminum Distribution Connector
Automatic Splice
Overhead Switch
Low Voltage Insulator Pin
Crossarm Brace
Guy Hook
Strain Plate
Trunion Suspension Clamp
Guy Clamp
Guy Wire Clip
Polymer Line Post Insulator
Secondary Racks
Wildlife Protector
Polymer Deadend Insulator
Deadend Strain Clamp
Stirrup Clamp
Aluminum Transformer Bracket
Support Bracket
Arrester
Hot Line Clamp
Lag Screw Type Pin
Lead Top Pin With Steel Bolt
Machine Bolt & Square Washer
Motor Operator
Single-Phase Recloser
Polymer Arrester
Cutout & Arrester Bracket
Alley Arm Brace
Electronic Sectionalizer
Cutout
Dead End Clevis
Double Arming Bolt
Pole Step
By-Pass Switch
Polymer Cutout
Polymer Riser Pole Arrester
Insulated Clevis
Thimble Eye Nut
Thimble Eye Bolt
Pole Eye Plate
Machine Bolt & Washer
Fiberglass Guy Strain Insulator
Suspension Clamp
Lashing Clamps
Guy Wire Clip
Sus-pension Clamp
U Cable Guard
Guy Marker
Conduit Standoff Bracket
Bonding Clamp
Automatic Guy Deadend
Detachable Pole Step
Polymer Elbow Arrester
Ground Rod Clamp
Cable Guard Strap
Lag & Plate for Detachable Pole Step
Thimble Eye
Separate Elbow
Anchors
Junction
Multi-Tap
Ground Rod
Motor Operator
Padmounted Switch
Polymer Concrete Underground Enclosure
Deadbreak Elbow
Grounded Bushing
Polymer Concrete Box Pad or Flat Pad

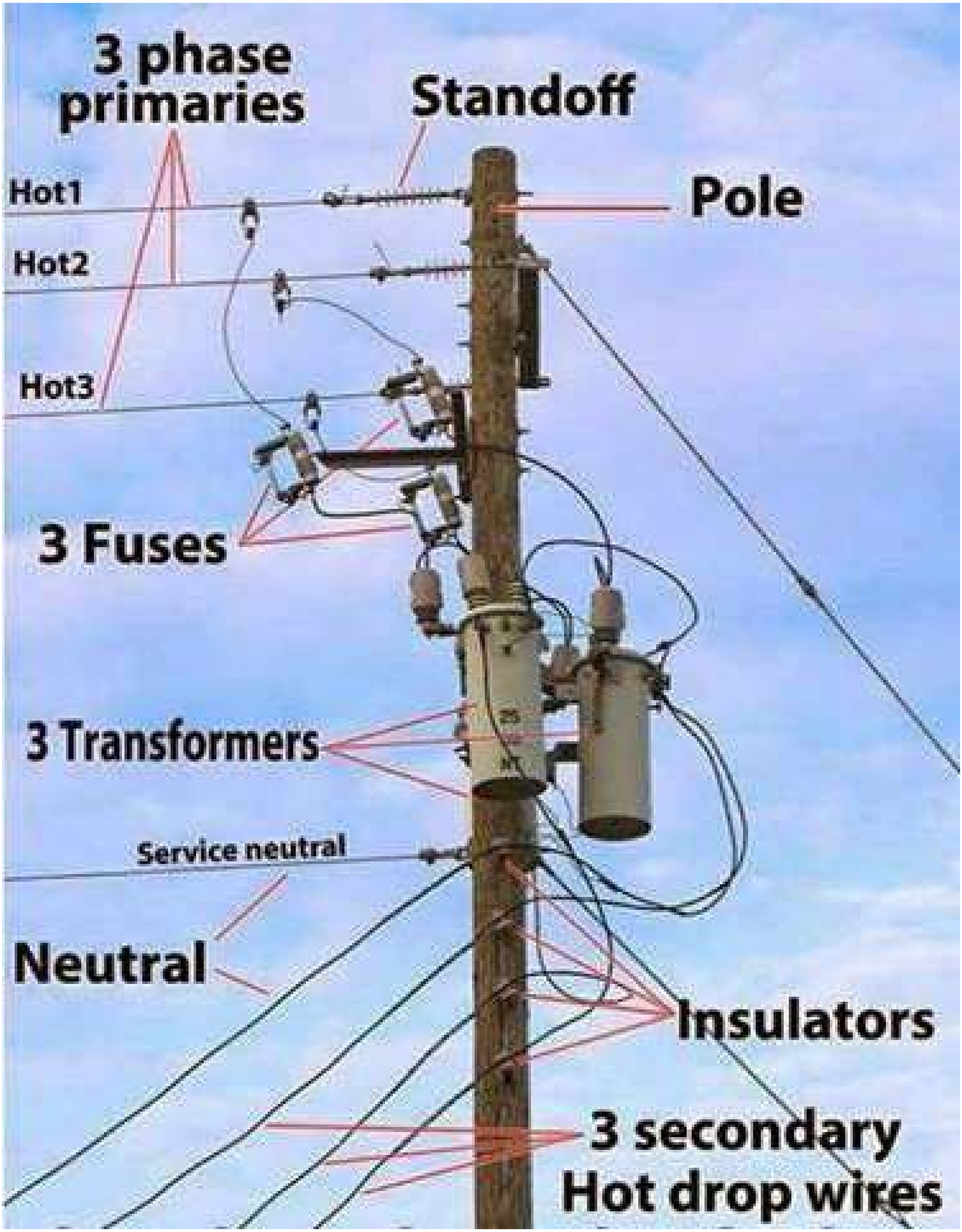
3 phase primaries
Standoff
Hot1
Pole
Hot2
Hot3
3 Fuses
3 Transformers
Service neutral
Neutral
Insulators
3 secondary Hot drop wires

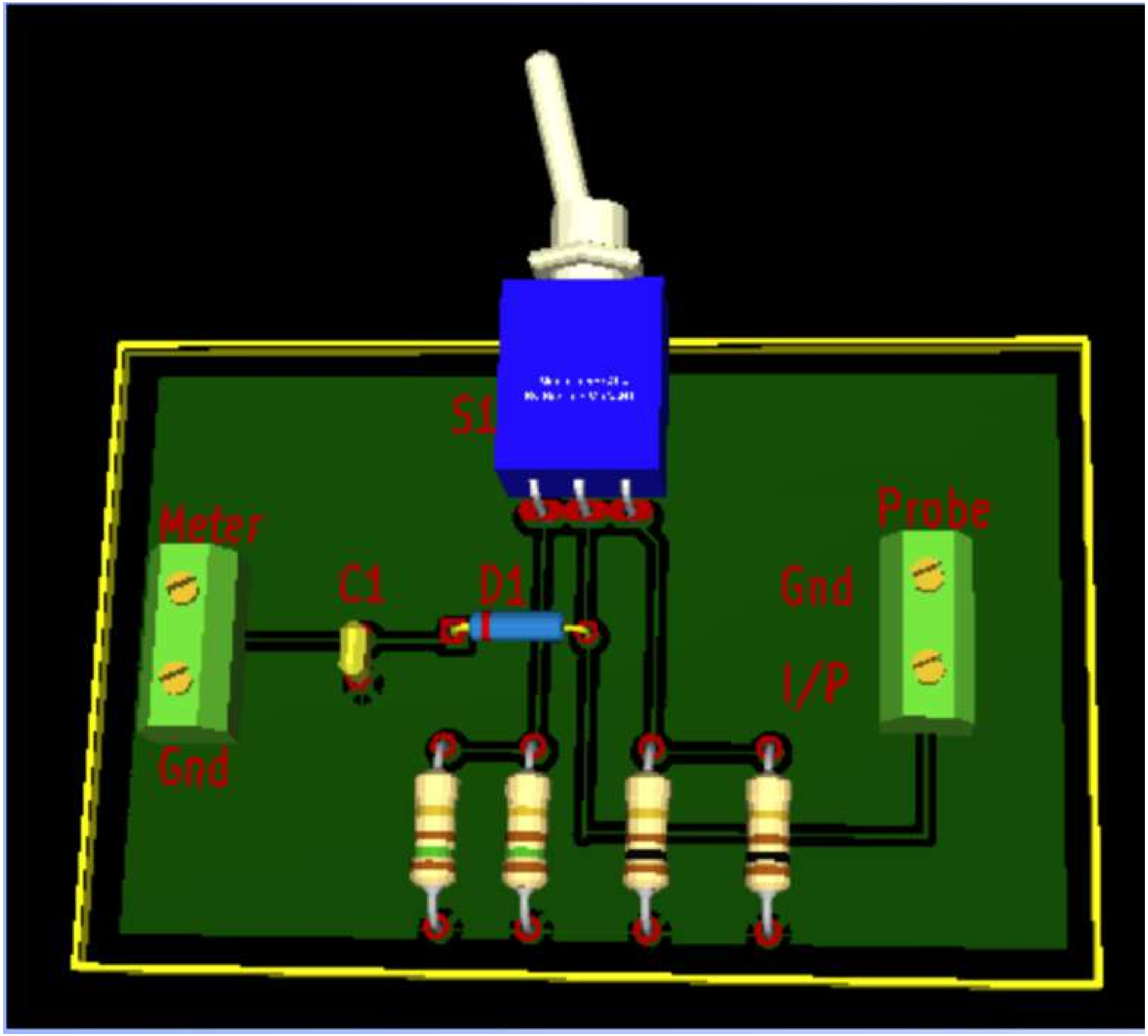
S1
Meter
C1
D1
Probe
Gnd
I/P
Gnd

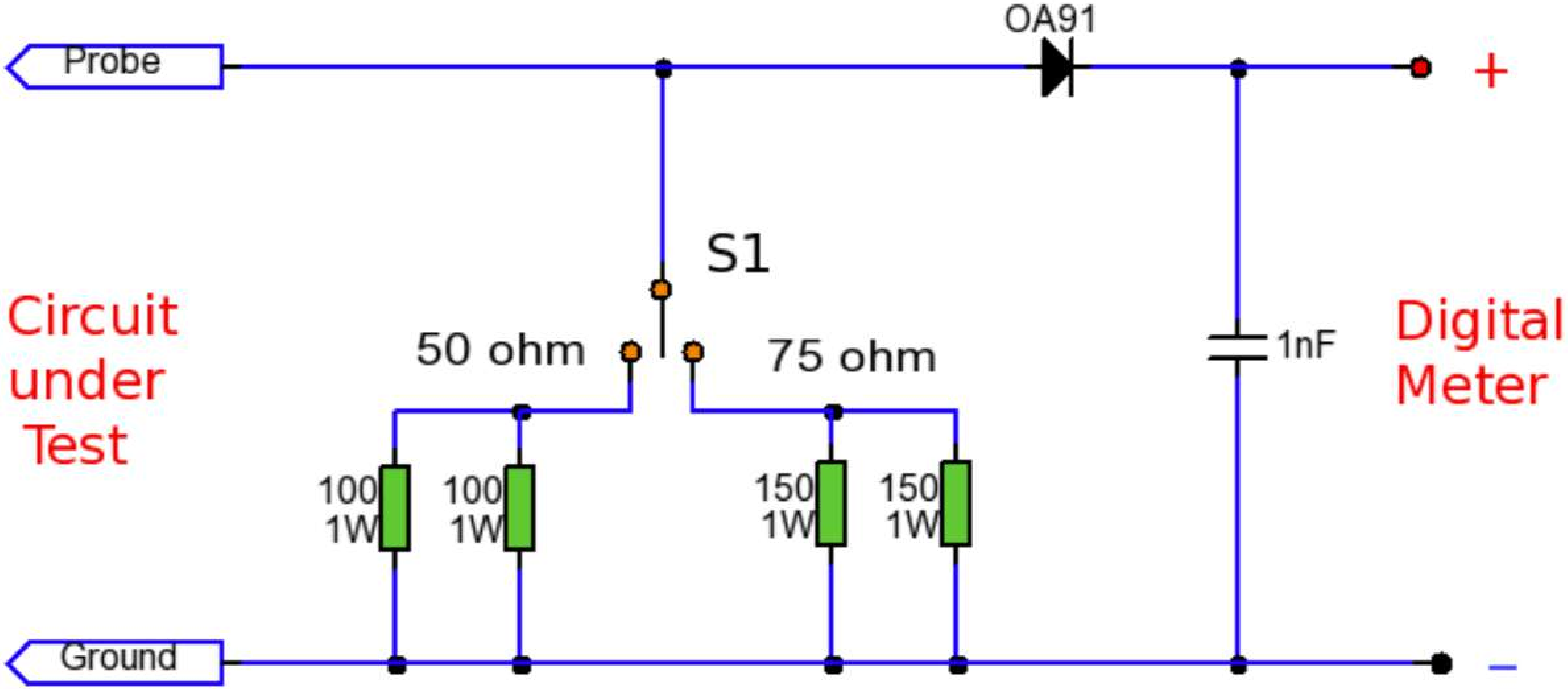
OA91
Probe
+
S1
Circuit under Test
50 ohm
75 ohm
1nF
Digital Meter
100 1W
100 1W
150 1W
150 1W
Ground
–

N. TESLA.
APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.
APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.

1,119,732. Patented Dec. 1, 1914.

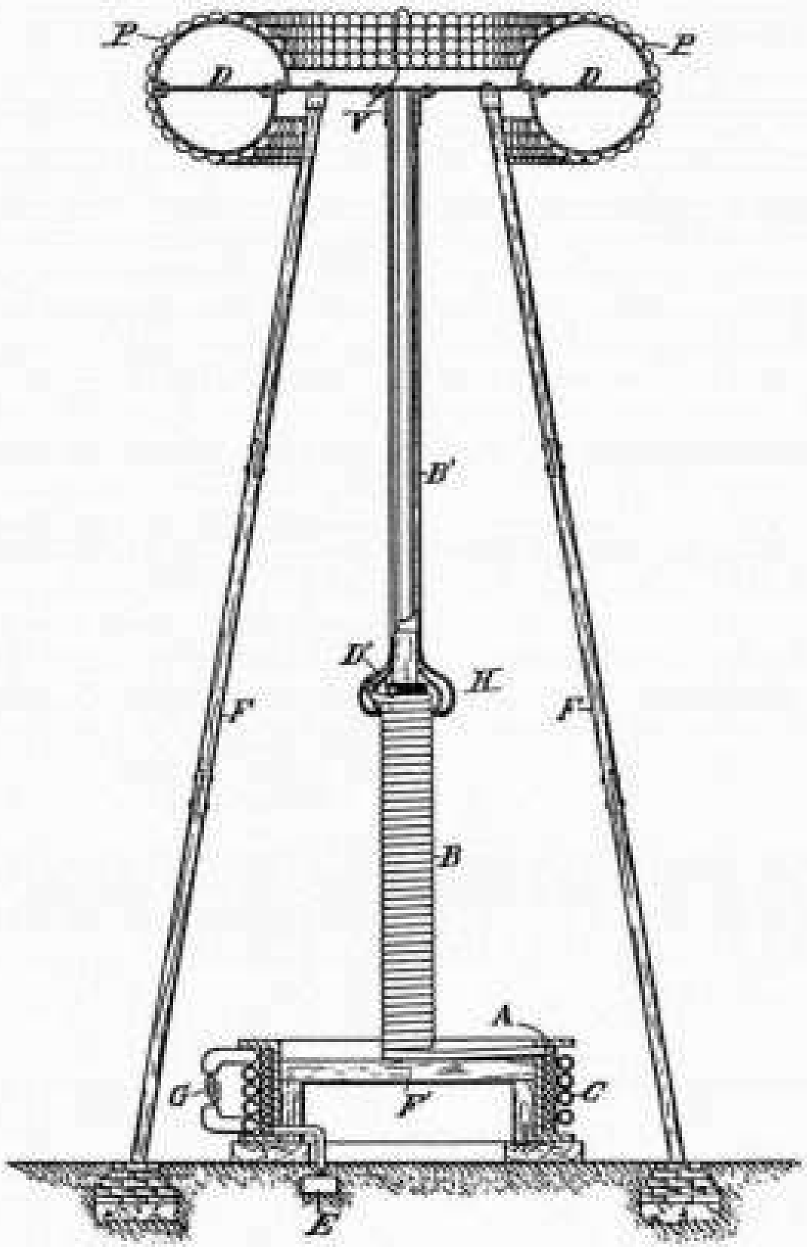
P
P
D
D
V
B'
H
H
F
F
B
A
G
C
F'
E

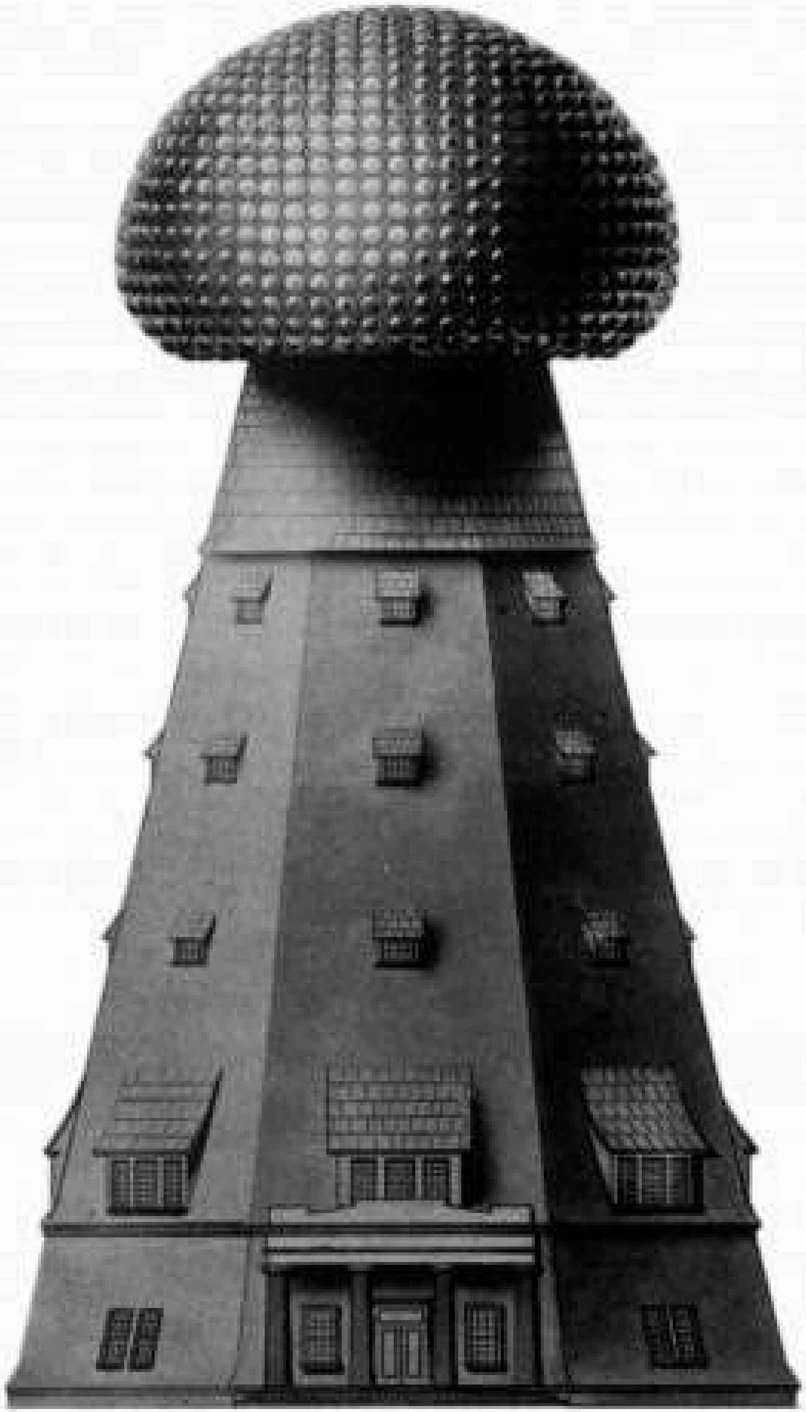

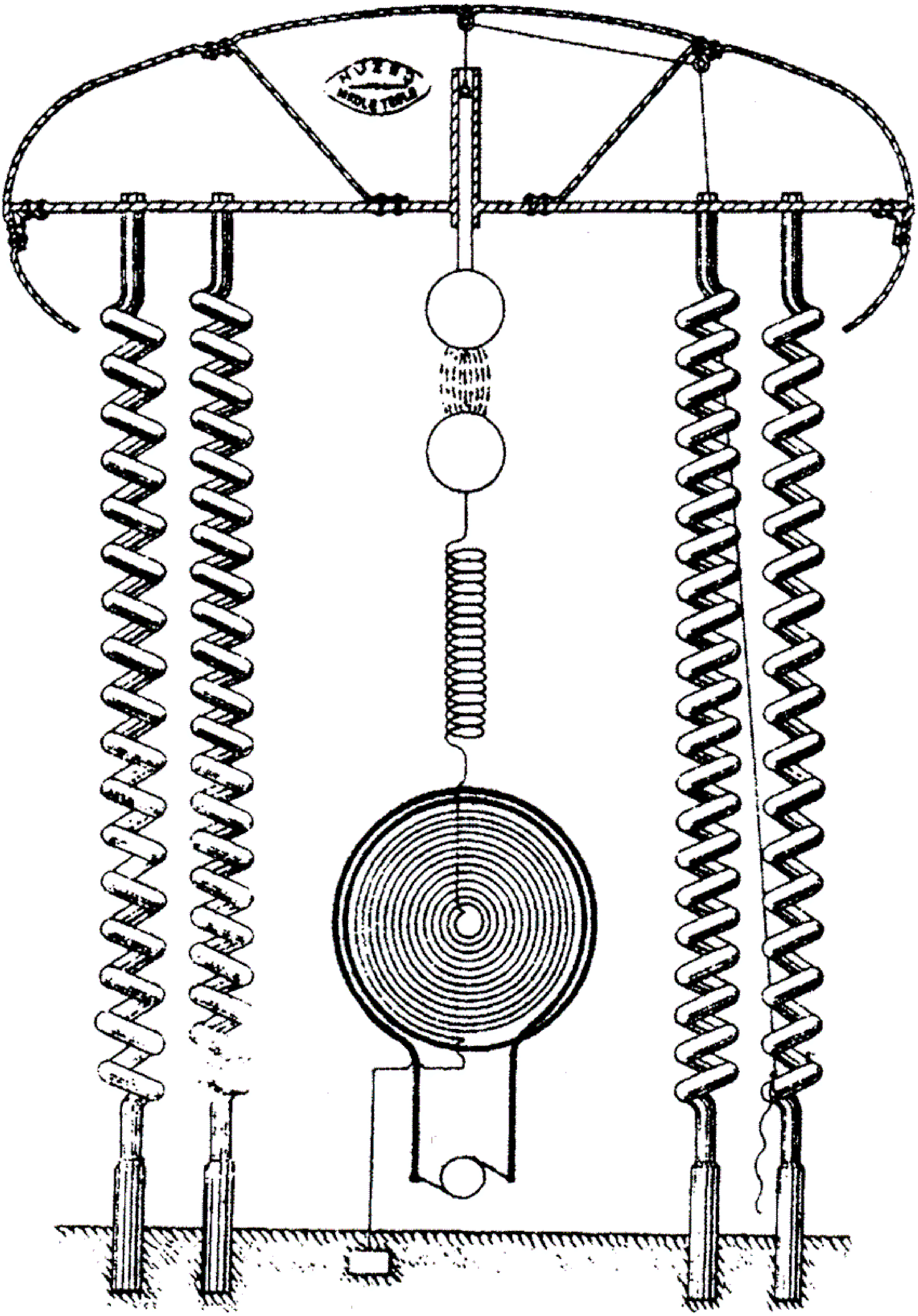

New York, Aug. 30th, 1901
46 & 48 East Houston Str.

Mr. Stanford White
160 Fifth Ave.
New York City

My Dear Stanford:

Many thanks for your suggestions. I am writing to Mr. Powell today. Perhaps he will be able to clear the land altogether.

I want you to understand that I went to the American Bridge Company simply because of my anxiety to have the work pushed through as fast as practicable. I am only too glad to follow your advice and beg you to consider yourself absolutely free in your choice and arrangements regarding this work.

Yours very sincerely,
N. Tesla

New York, Sep. 12th, 1901
46 & 48 East Houston Str.

Babcock & Wilcox Co.
85 Liberty Street
New York City

Gentlemen:

Under enclosure I forward sketch showing your two boilers as they will be placed in my building and their position relative to and exact distance from the chimney. The scale is ½ inch to a foot.

You will greatly oblige me by furnishing the drawings of the flues leading to the chimney and the position of the breech, as the builder cannot proceed without this information.

Yours very truly,
Encl.

Anyone familiar with the Wardenclyffe Tower knows it to have been a colossal structure. Yet, few realize that it was supposed to have been even larger. Although the exact figures are not revealed, Tesla must have drastically underestimated the cost of building his structure as is evidenced by the following response to White.

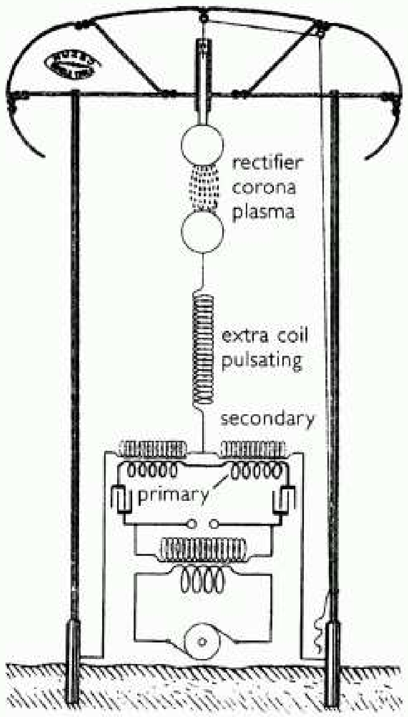

Fig. 5 Oscillating statically charged terminal.

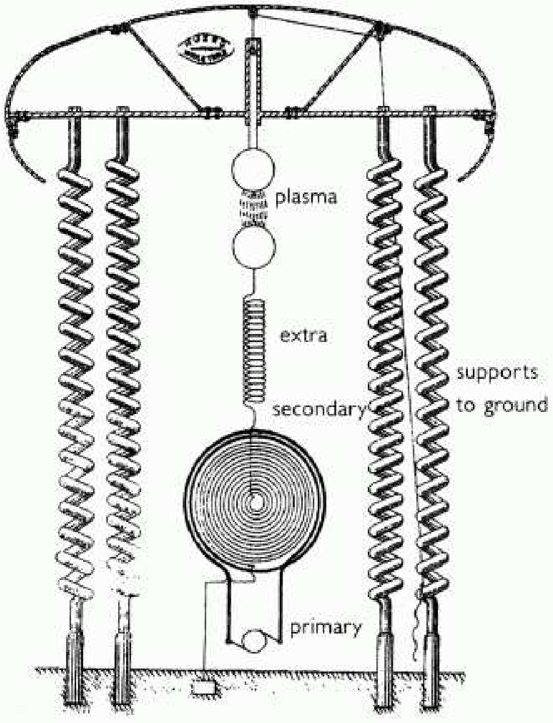

Fig. 6 Oscillating electrostatically charged dome.

N. TESLA
APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.
APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.

1,119,732. Patented Dec. 1, 1914.

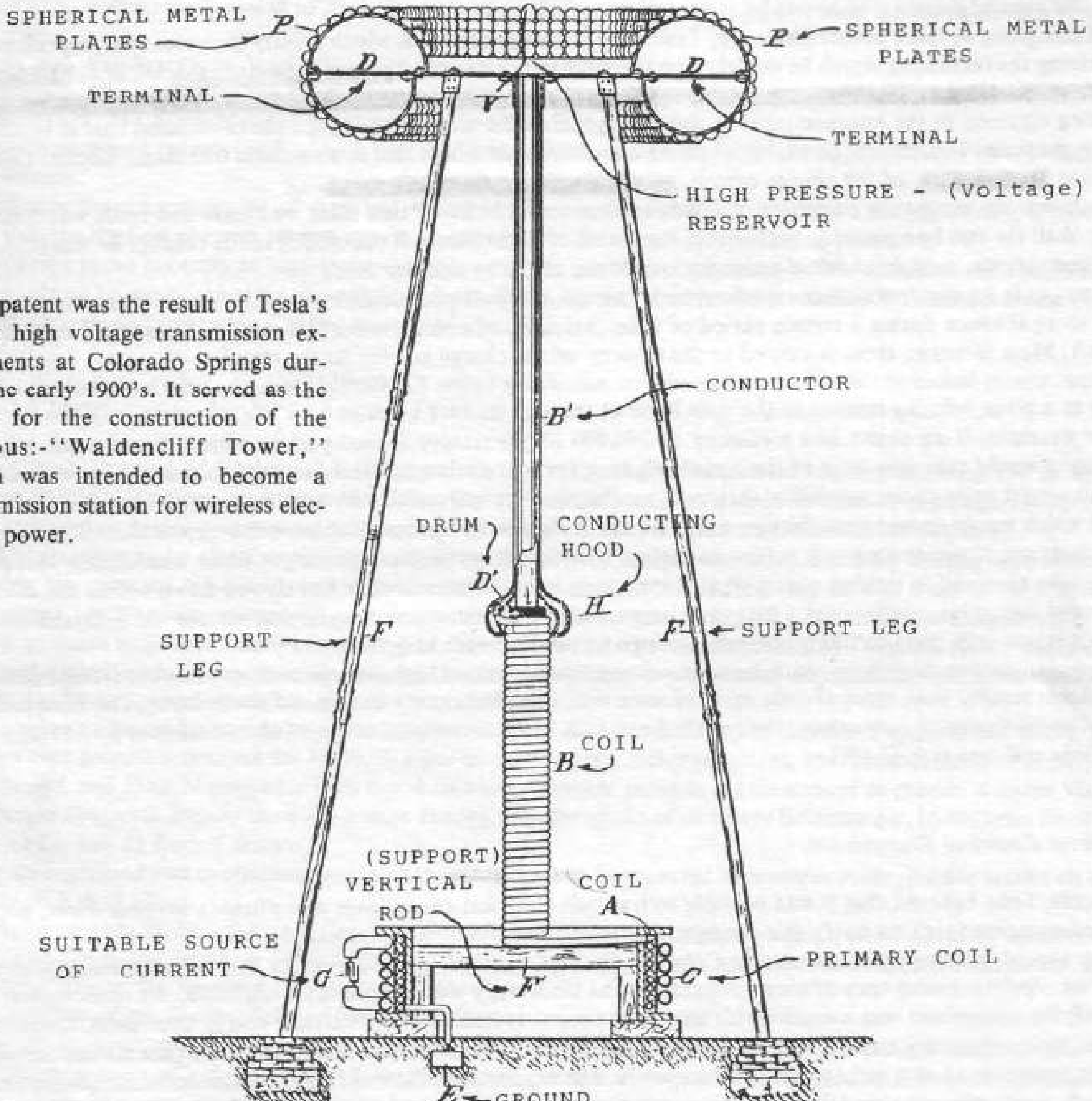
SPHERICAL METAL PLATES
P
P
SPHERICAL METAL PLATES
D
D
TERMINAL
TERMINAL
V
HIGH PRESSURE - (Voltage) RESERVOIR
CONDUCTOR
B'
DRUM
CONDUCTING HOOD
D'
H
SUPPORT LEG
F
F
SUPPORT LEG
COIL
B
(SUPPORT) VERTICAL ROD
COIL
A
SUITABLE SOURCE OF CURRENT
G
F'
C
PRIMARY COIL
E
GROUND PLATE

This patent was the result of Tesla's many high voltage transmission experiments at Colorado Springs during the early 1900's. It served as the basis for the construction of the famous:-''Waldencliff Tower,'' which was intended to become a transmission station for wireless electrical power.

WITNESSES:

M. Lawson Dyer

Benjamin Miller.

Nikola Tesla, INVENTOR.

BY Kerr, Page & Cooper, his ATTORNEYS.

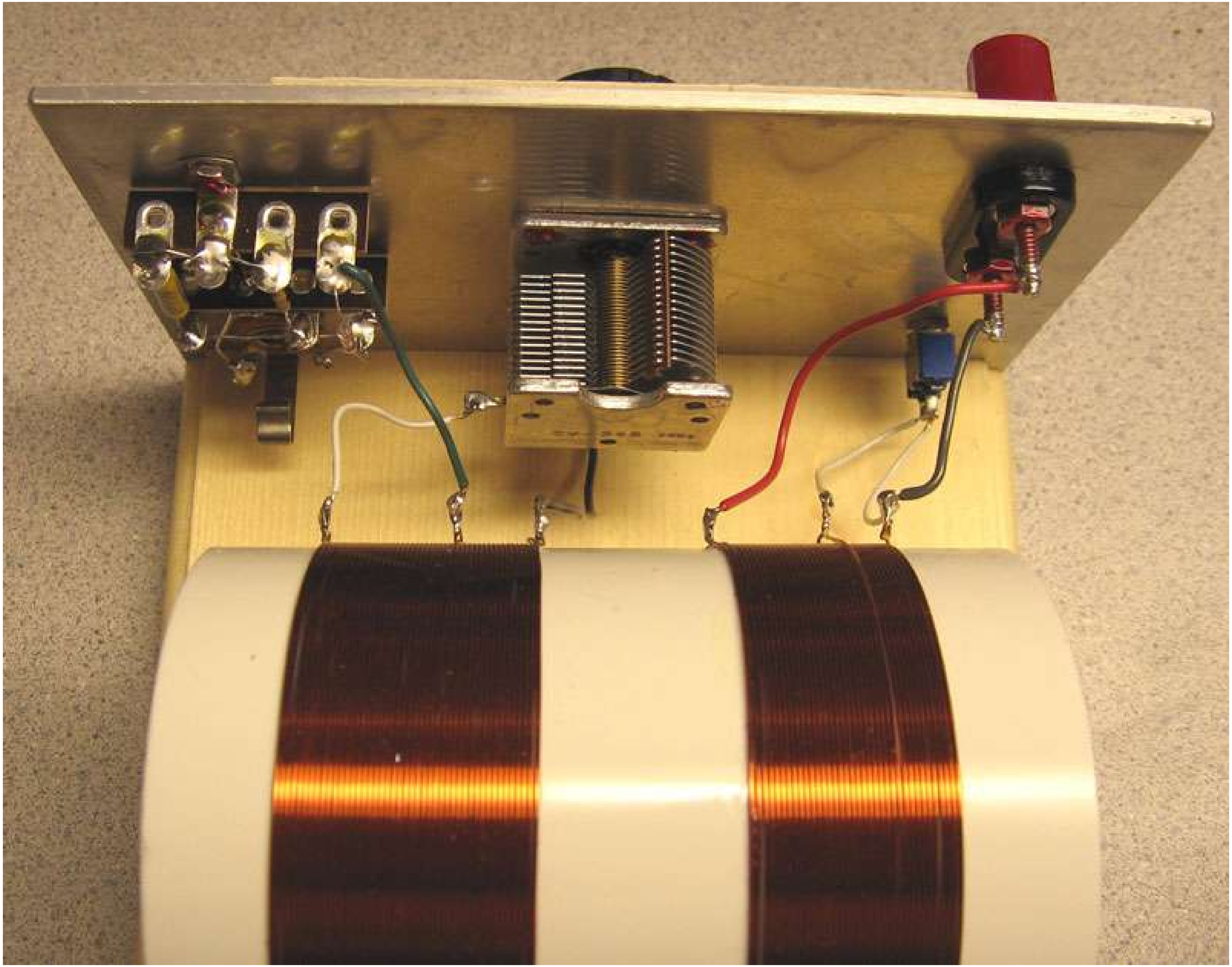

"HIGH SELECTIVITY CRYSTAL RADIO"
[WITH SHORT WAVE]

RGK 2007

570KHZ → 2.5MHZ

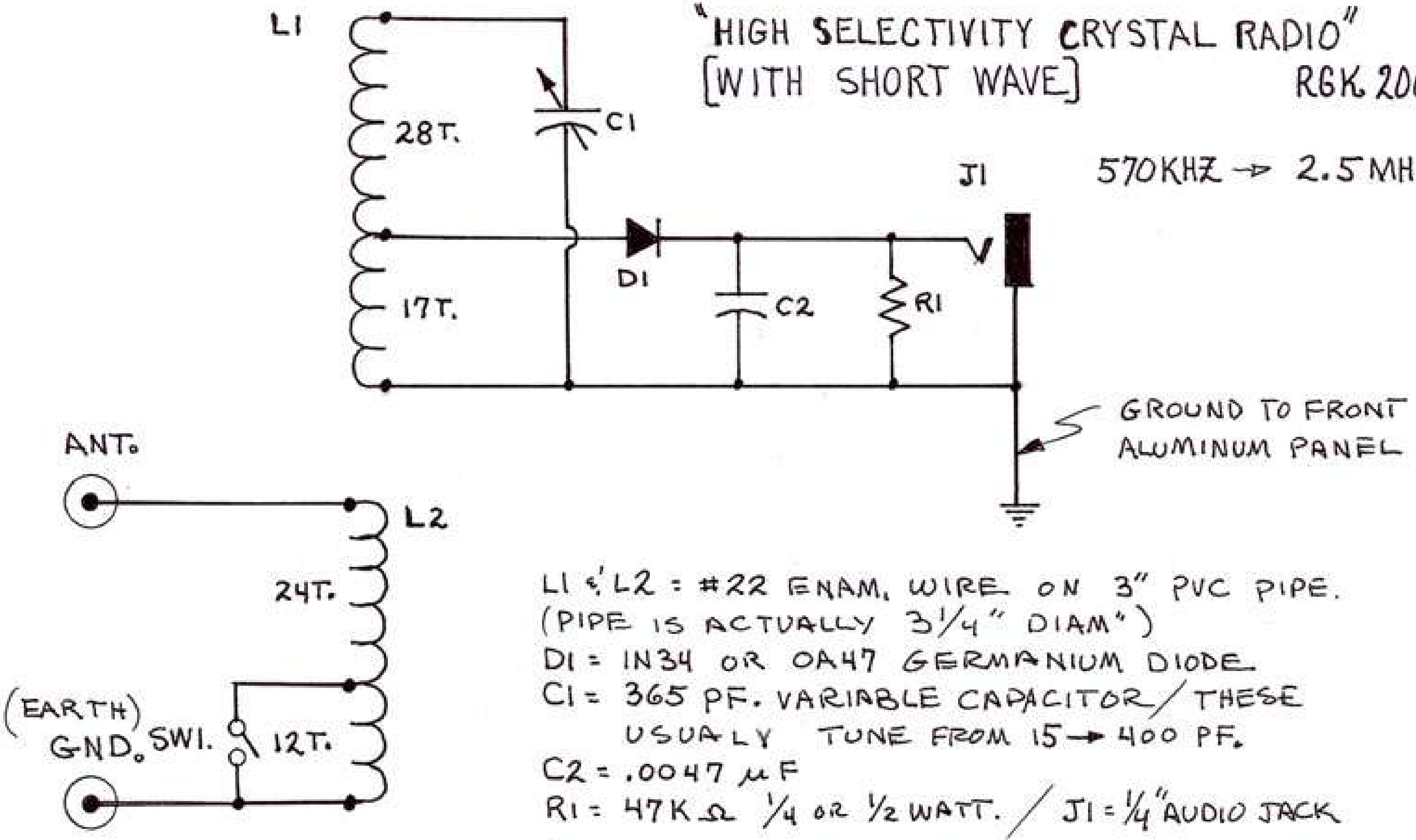
L1
28T.
C1
17T.
D1
C2
R1
J1
GROUND TO FRONT ALUMINUM PANEL
ANT.
L2
24T.
(EARTH) GND.
SW1.
12T.

L1 & L2 = #22 ENAM. WIRE ON 3" PVC PIPE.
(PIPE IS ACTUALLY 3 1/4" DIAM")
D1 = 1N34 OR OA47 GERMANIUM DIODE
C1 = 365 PF. VARIABLE CAPACITOR / THESE
USUALY TUNE FROM 15 → 400 PF.
C2 = .0047 µF
R1 = 47K Ω 1/4 OR 1/2 WATT. / J1 = 1/4" AUDIO JACK

[L1/L2 NOTES: L1 = 45 TURNS TOTAL & L2 = 36 TURNS TOTAL. THE SPACE BETWEEN L1 & L2 IS 1 1/16".]

SW1 = SP.S.T. SWITCH / FOR ANTENNA MATCHING

630V 0245
150M

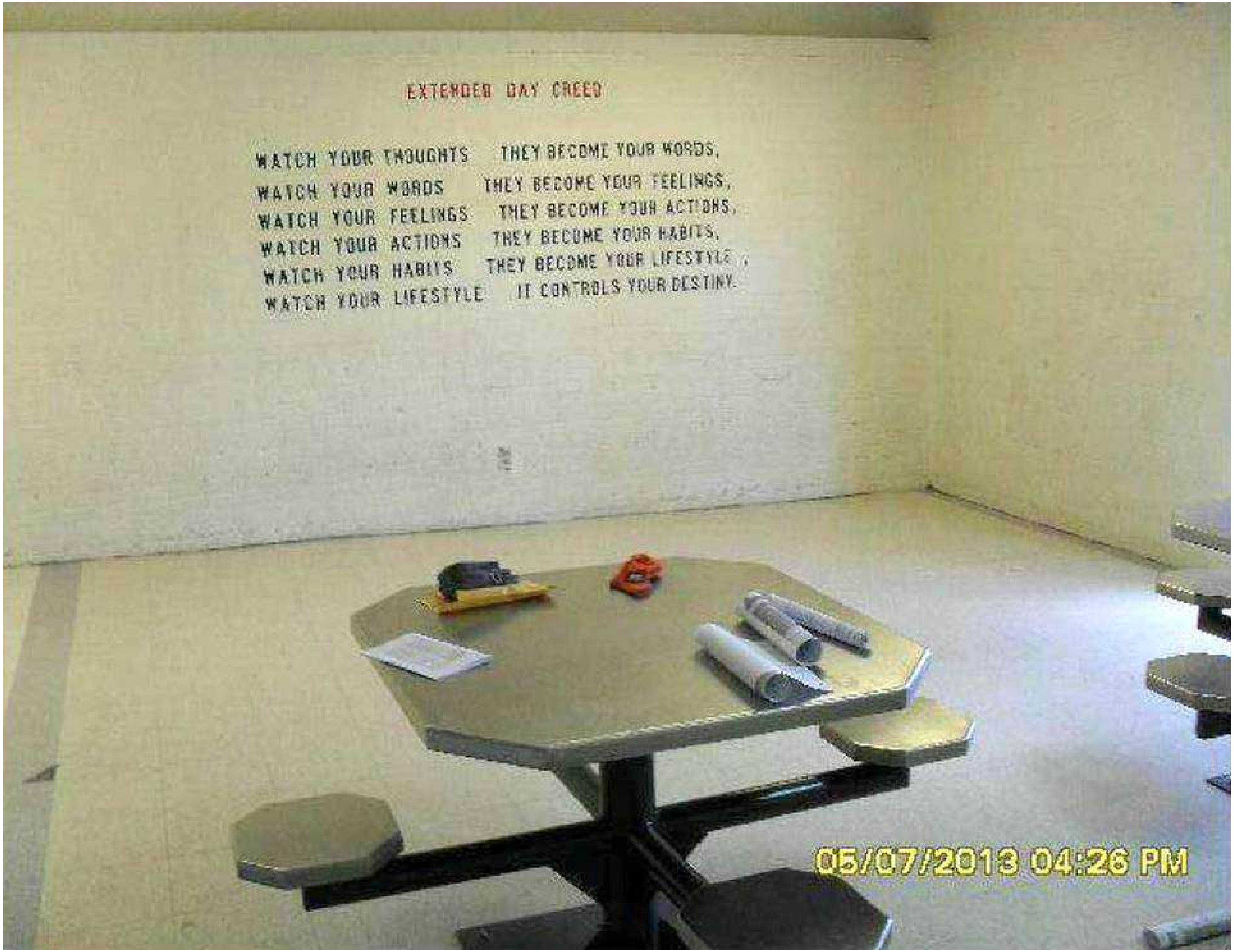
EXTENDED DAY CREED
WATCH YOUR THOUGHTS THEY BECOME YOUR WORDS,
WATCH YOUR WORDS THEY BECOME YOUR FEELINGS,
WATCH YOUR FEELINGS THEY BECOME YOUR ACTIONS,
WATCH YOUR ACTIONS THEY BECOME YOUR HABITS,
WATCH YOUR HABITS THEY BECOME YOUR LIFESTYLE,
WATCH YOUR LIFESTYLE IT CONTROLS YOUR DESTINY.
05/07/2013 04:26 PM

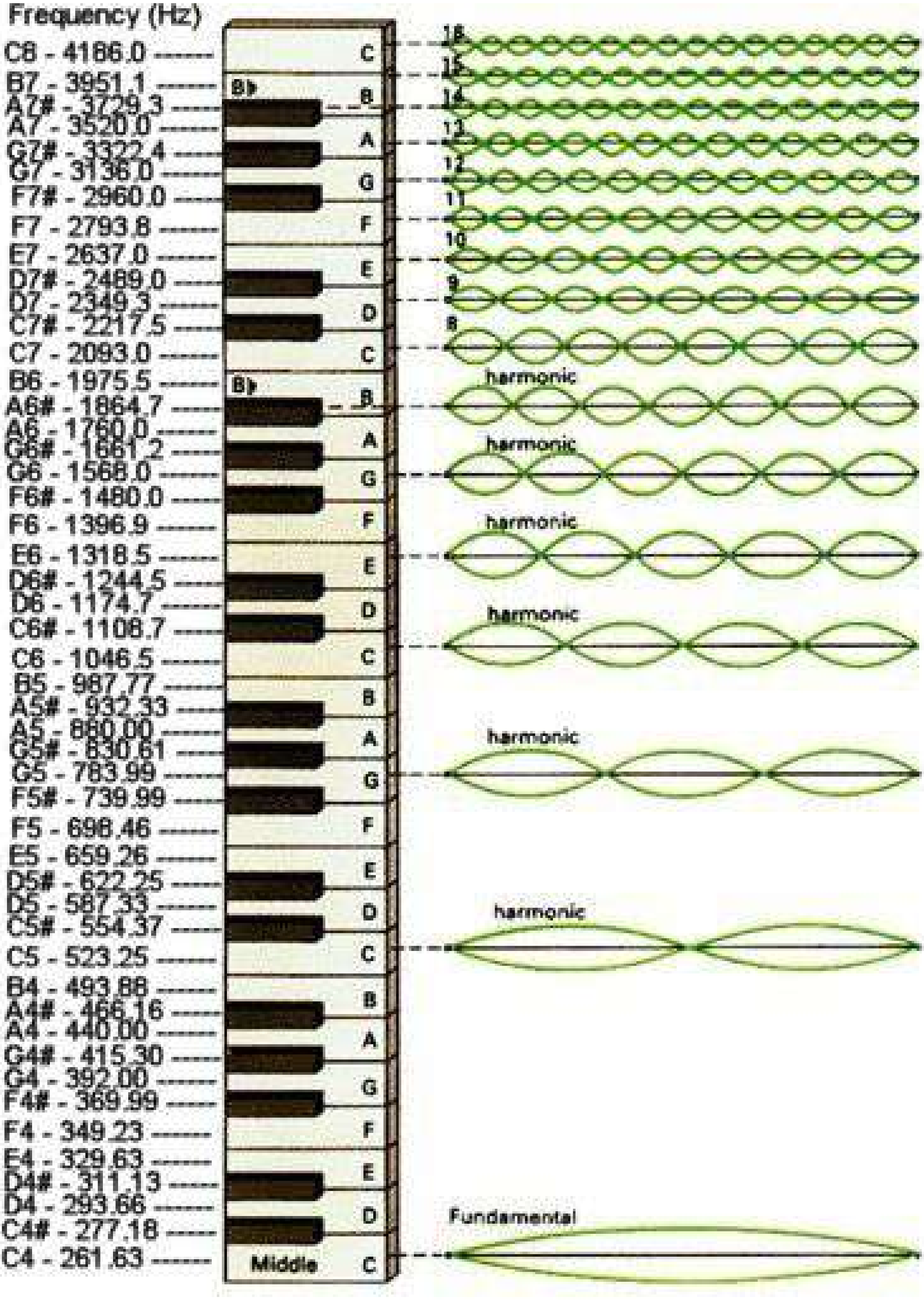
Frequency (Hz)
C8 - 4186.0
B7 - 3951.1
A7# - 3729.3
A7 - 3520.0
G7# - 3322.4
G7 - 3136.0
F7# - 2960.0
F7 - 2793.8
E7 - 2637.0
D7# - 2489.0
D7 - 2349.3
C7# - 2217.5
C7 - 2093.0
B6 - 1975.5
A6# - 1864.7
A6 - 1760.0
G6# - 1661.2
G6 - 1568.0
F6# - 1480.0
F6 - 1396.9
E6 - 1318.5
D6# - 1244.5
D6 - 1174.7
C6# - 1108.7
C6 - 1046.5
B5 - 987.77
A5# - 932.33
A5 - 880.00
G5# - 830.61
G5 - 783.99
F5# - 739.99
F5 - 698.46
E5 - 659.26
D5# - 622.25
D5 - 587.33
C5# - 554.37
C5 - 523.25
B4 - 493.88
A4# - 466.16
A4 - 440.00
G4# - 415.30
G4 - 392.00
F4# - 369.99
F4 - 349.23
E4 - 329.63
D4# - 311.13
D4 - 293.66
C4# - 277.18
C4 - 261.63
C
B♭
B
A
G
F
E
D
C
B♭
B
A
G
F
E
D
C
B
A
G
F
E
D
C
B
A
G
F
E
D
Middle
C
16
15
14
13
12
11
10
9
8
harmonic
harmonic
harmonic
harmonic
harmonic
harmonic
Fundamental

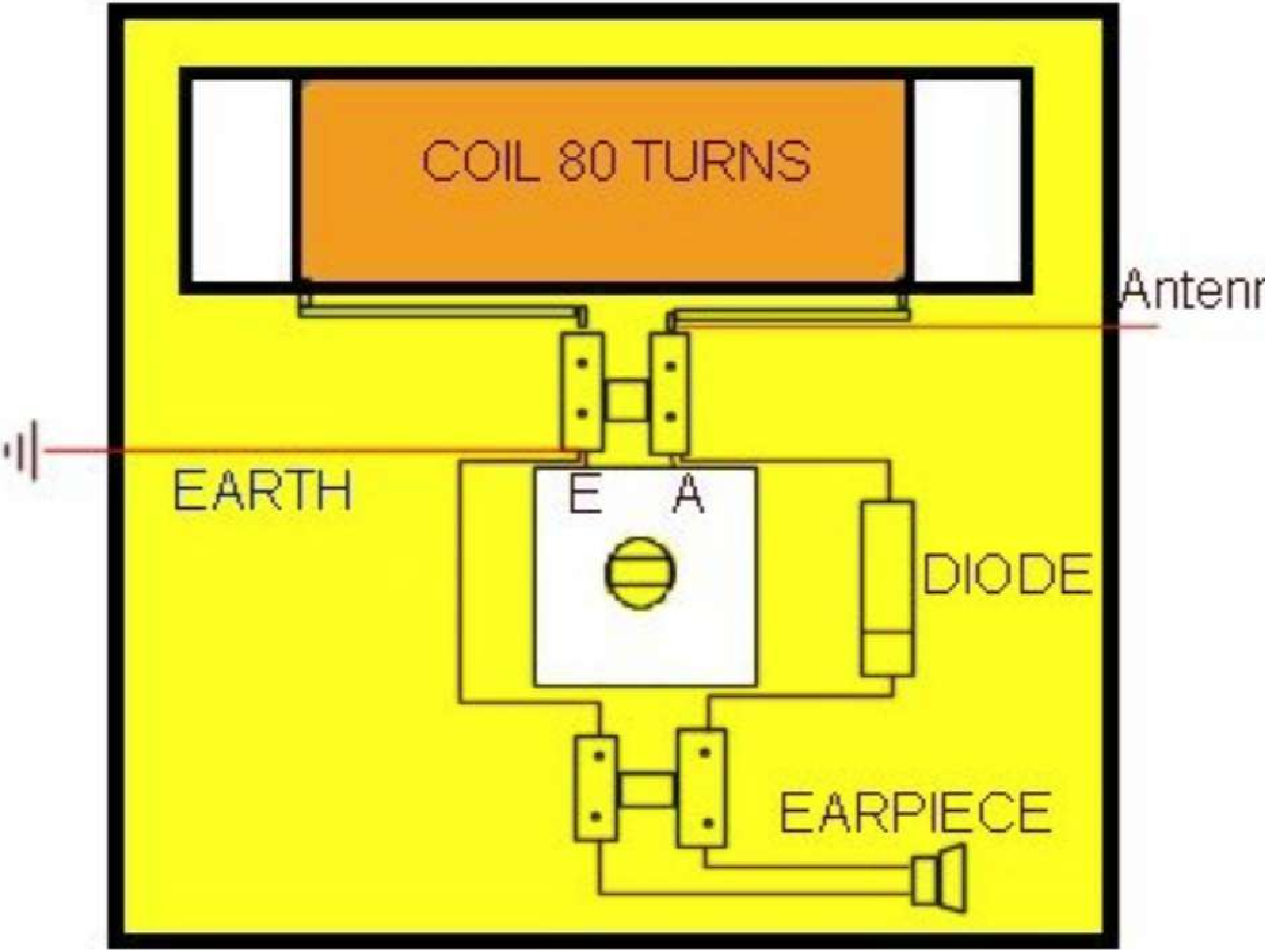
COIL 80 TURNS
Antenn
EARTH
E
A
DIODE
EARPIECE

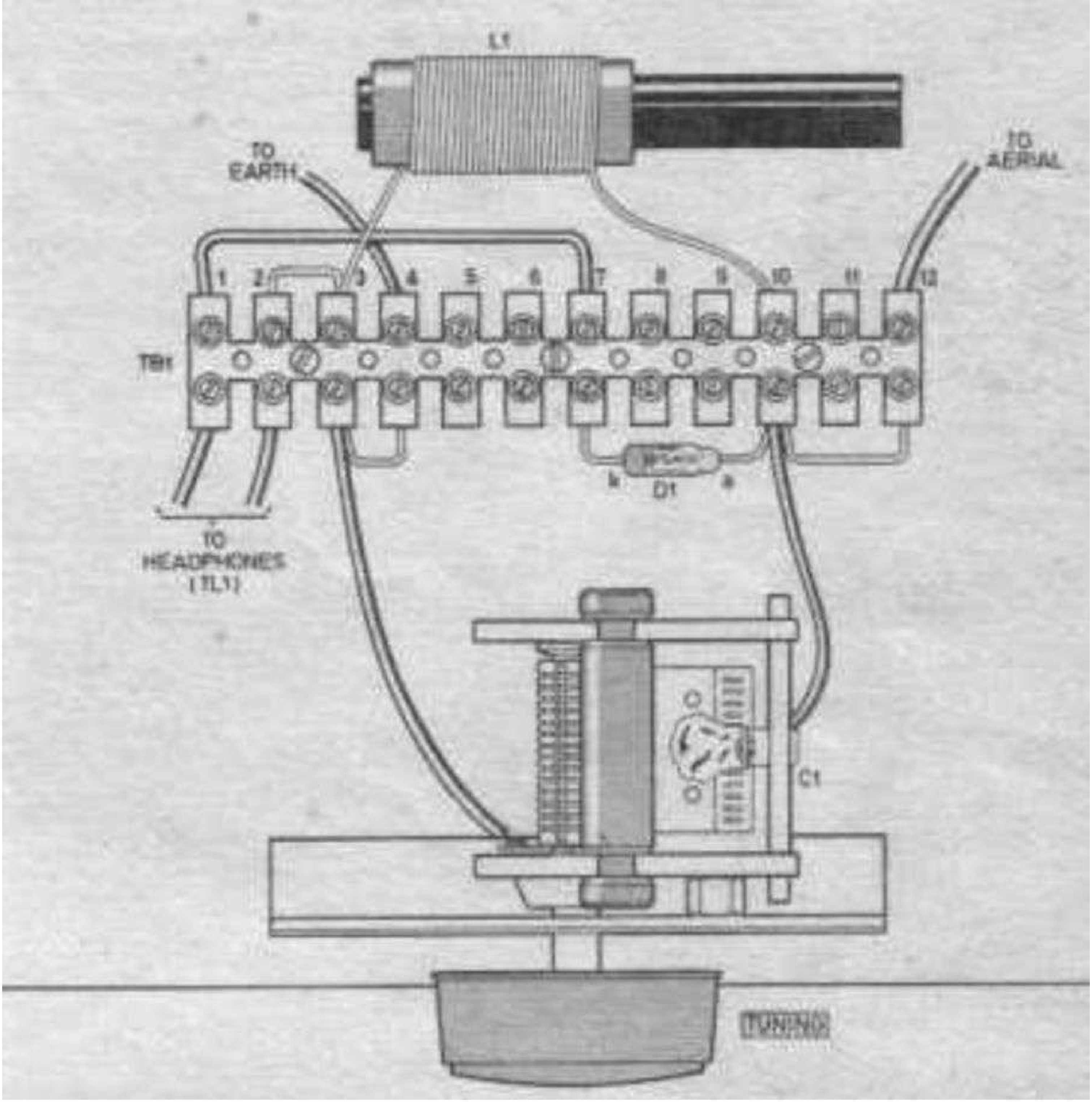
L1
TO EARTH
TO AERIAL
1
2
3
4
5
6
7
8
9
10
11
12
TB1
k
D1
a
TO HEADPHONES (TL1)
C1
TUNING

4

The **immortal jellyfish** (*Turritopsis dohrnii*) is capable of biological immortality.

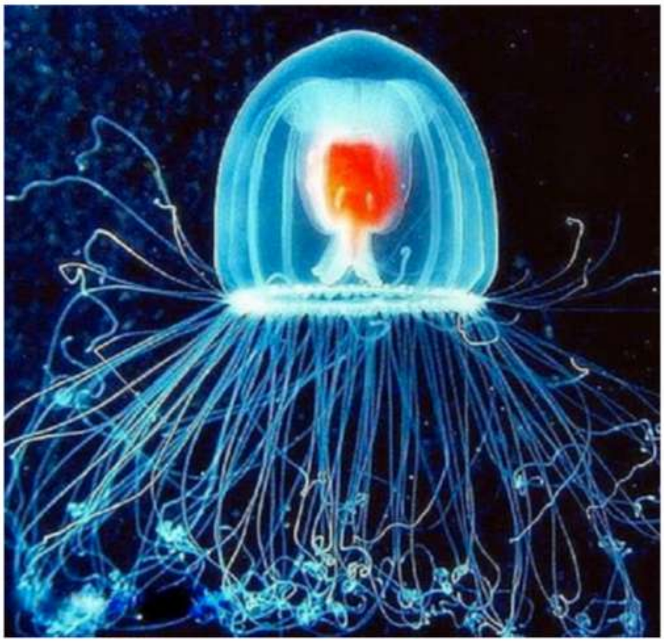

It's one of few known species capable of reverting completely to a sexually immature, colonial polyp stage after having reached sexual maturity as a solitary (free-floating) individual (called a medusa).

> Theoretically, this process can go on indefinitely, effectively rendering the jellyfish biologically immortal

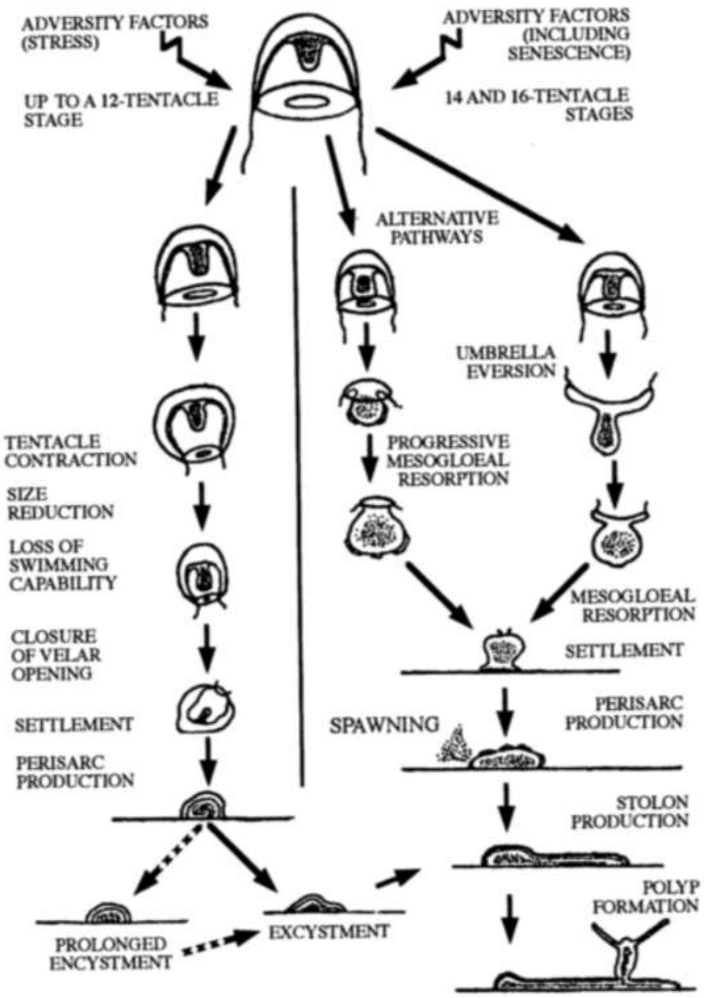
ADVERSITY FACTORS (STRESS)
UP TO A 12-TENTACLE STAGE
ADVERSITY FACTORS (INCLUDING SENESCENCE)
14 AND 16-TENTACLE STAGES
ALTERNATIVE PATHWAYS
UMBRELLA EVERSION
TENTACLE CONTRACTION
SIZE REDUCTION
LOSS OF SWIMMING CAPABILITY
PROGRESSIVE MESOGLOEAL RESORPTION
CLOSURE OF VELAR OPENING
MESOGLOEAL RESORPTION
SETTLEMENT
SETTLEMENT
SPAWNING
PERISARC PRODUCTION
PERISARC PRODUCTION
STOLON PRODUCTION
POLYP FORMATION
PROLONGED ENCYSTMENT
EXCYSTMENT

Figure 3. Pathways of transformation from medusa into polyp. Fate of stressed medusae up to 12-tentacle stage (left side), and alternative transformations of stressed or spawning medusae from a 14-tentacle or 16-tentacle stage (right side). The final product is always the polyp colony (bottom), directly or through a resting stage.

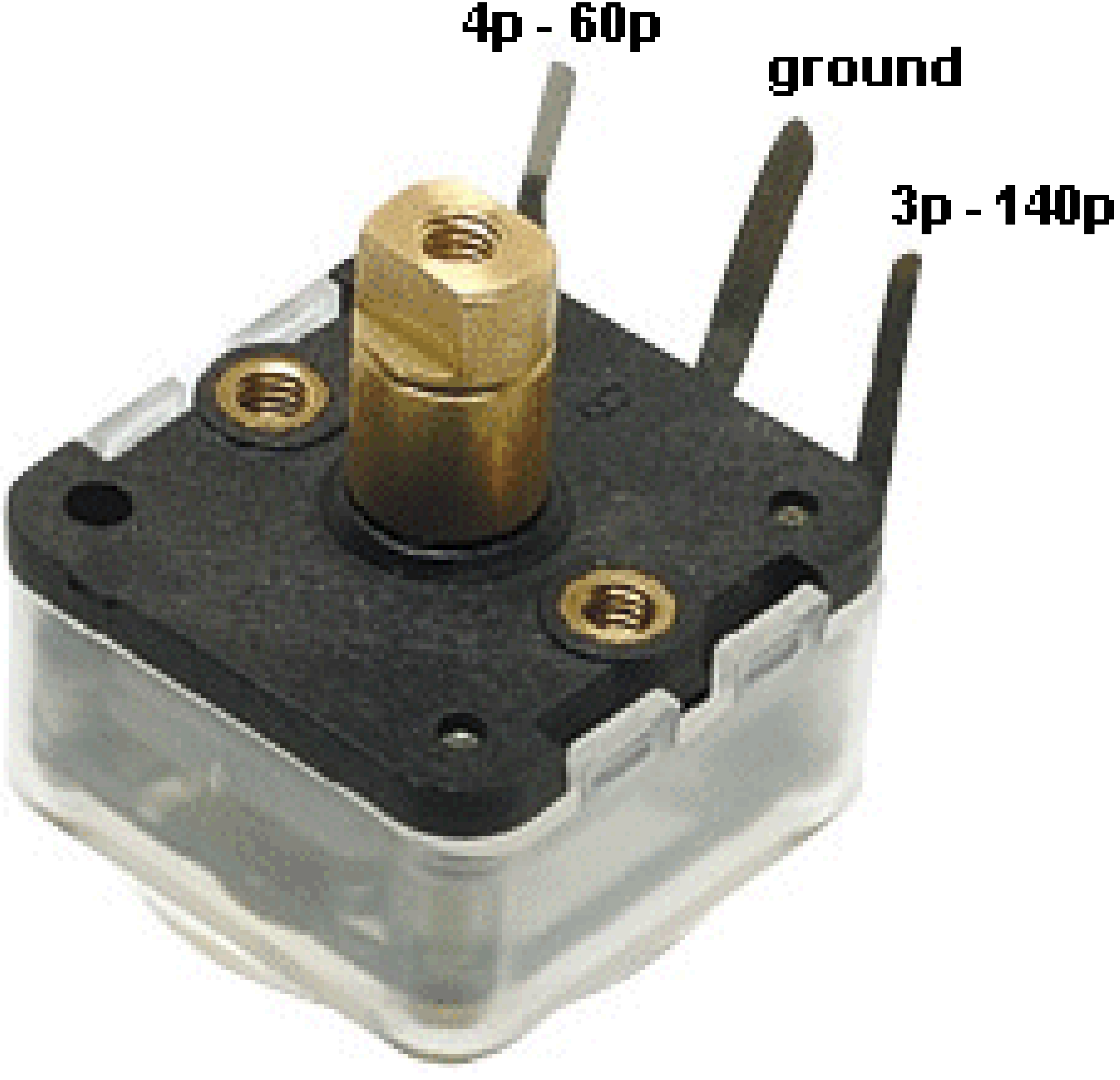
4p - 60p
ground
3p - 140p

© www.petervis.com

© Copyright Protec

eisco
Germanium Diode
A
K
1N60
Max. 50 V
CE

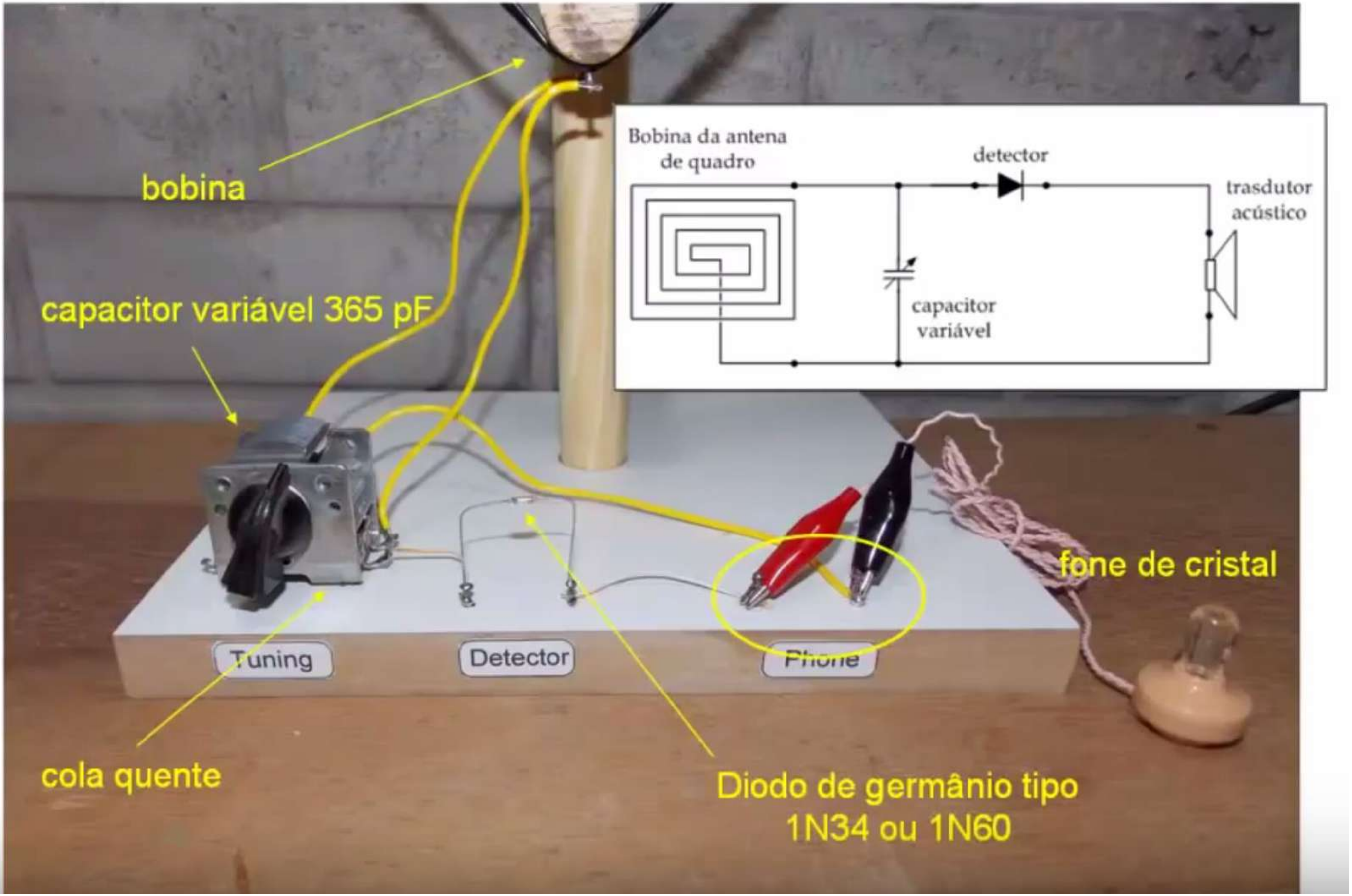
bobina
capacitor variável 365 pF
cola quente
Tuning
Detector
Phone
Diodo de germânio tipo
1N34 ou 1N60
fone de cristal
Bobina da antena
de quadro
detector
trasdutor
acústico
capacitor
variável

Figure 12. (**a**) Output power for a half-wave and full-wave CWVM, and (**b**) portable electronic calculator running with the AM energy harvester. Detected input signal in LC resonator is 1.5 V at 1 MHz.

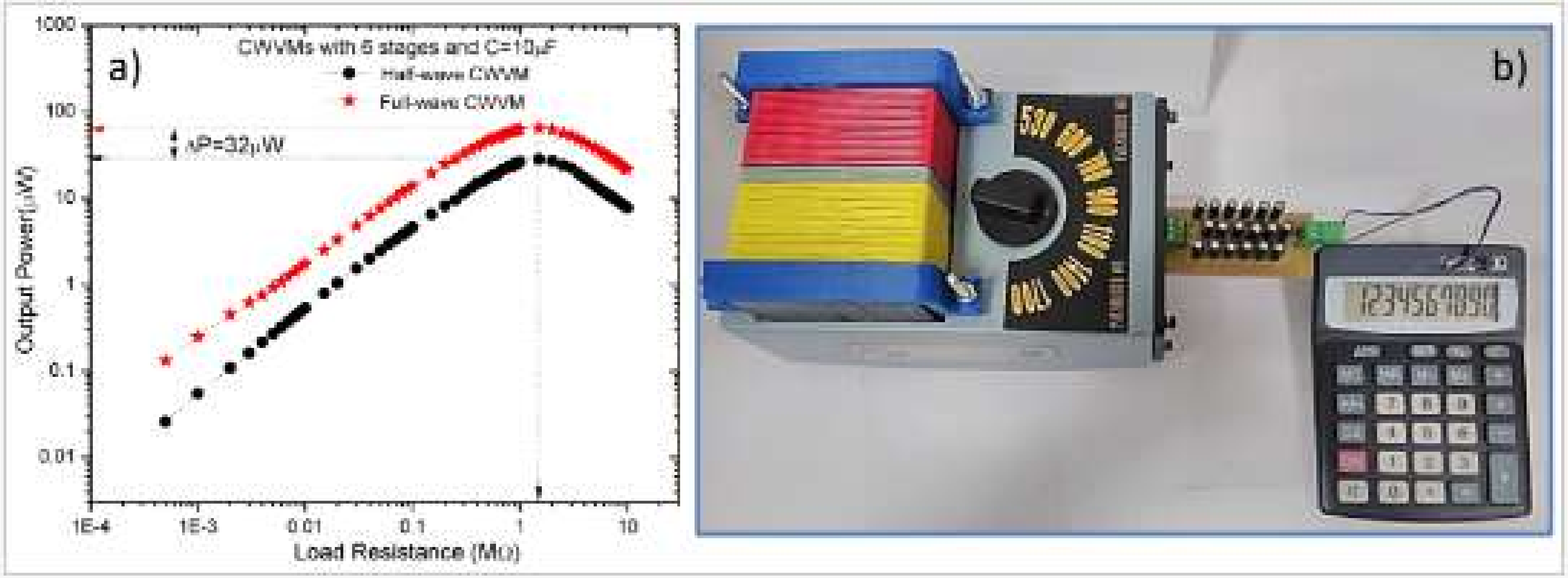

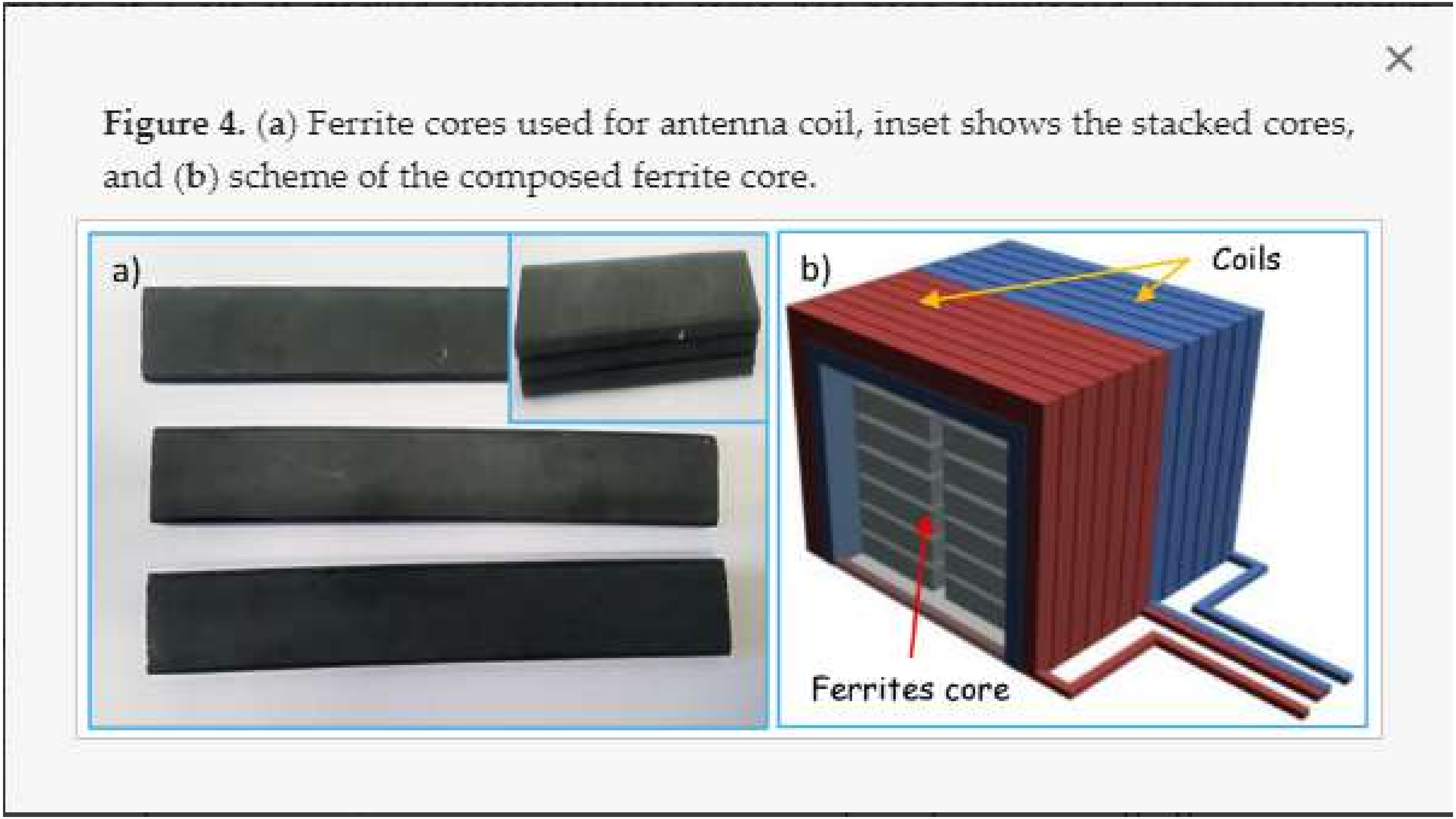

Figure 4. (**a**) Ferrite cores used for antenna coil, inset shows the stacked cores, and (**b**) scheme of the composed ferrite core.

×
a)
b)
Coils
Ferrites core

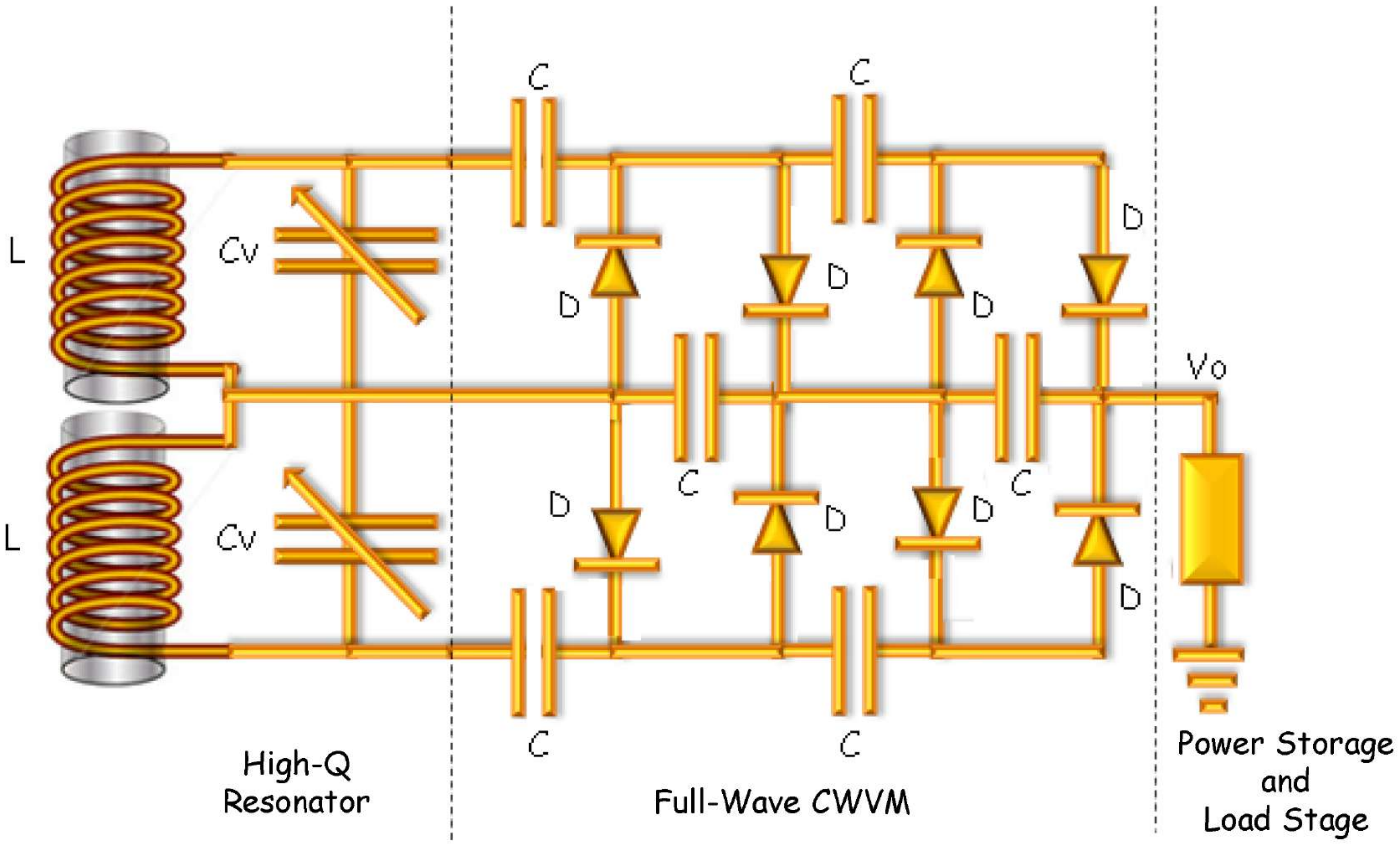
L
Cv
L
Cv
C
C
D
D
D
D
C
C
Vo
D
D
D
D
C
C
D
High-Q
Resonator
Full-Wave CWVM
Power Storage
and
Load Stage

Figure 5. (a) circuit for a six-stage conventional CWVM, and (b) circuit for a six-stage full wave-CWVM. Here, Cs stands for the series capacitances.

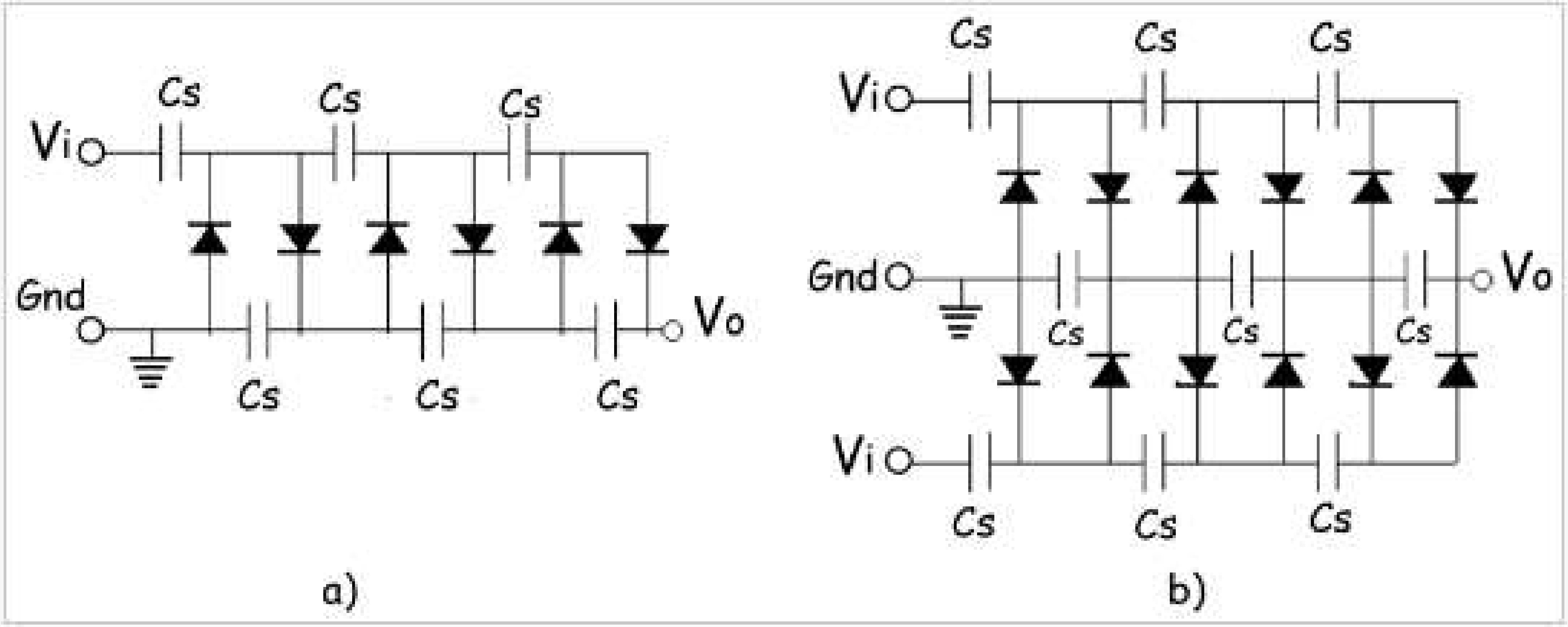
Vi
Cs
Cs
Cs
Gnd
Vo
Cs
Cs
Cs
a)
Cs
Cs
Cs
Vi
Gnd
Vo
Cs
Cs
Cs
Vi
Cs
Cs
Cs
b)

Figure 6. Pictorial image of the AM-RF energy harvesting system.

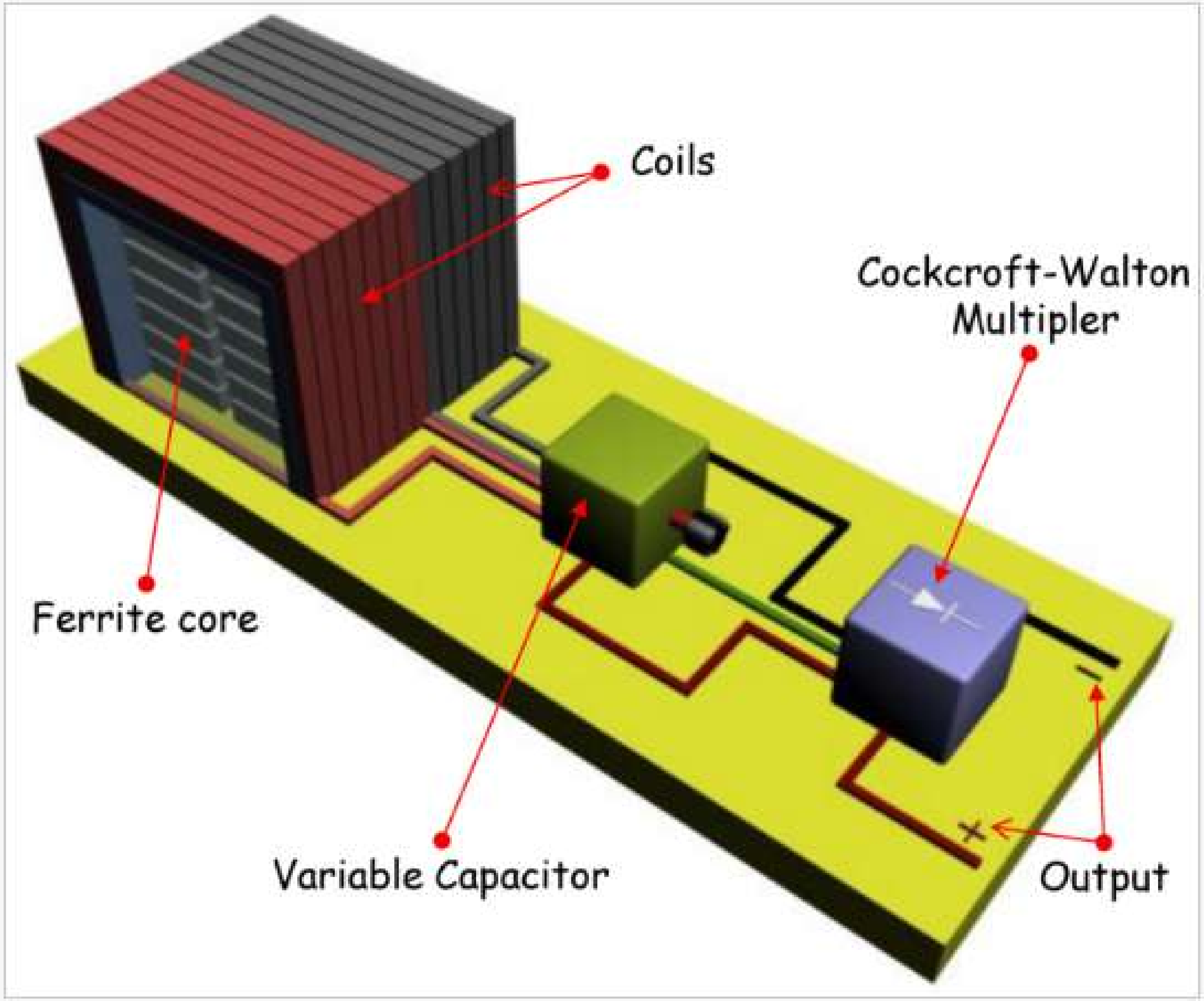
Coils
Cockcroft-Walton Multipler
Ferrite core
Variable Capacitor
−
+
Output

Figure 11. (a) AM resonator implemented, (b) Full-wave CWVM implemented, (c) top view of the AM resonator and full-wave CWVM, and (d) lateral view of the AM resonator and full-wave CWVM.

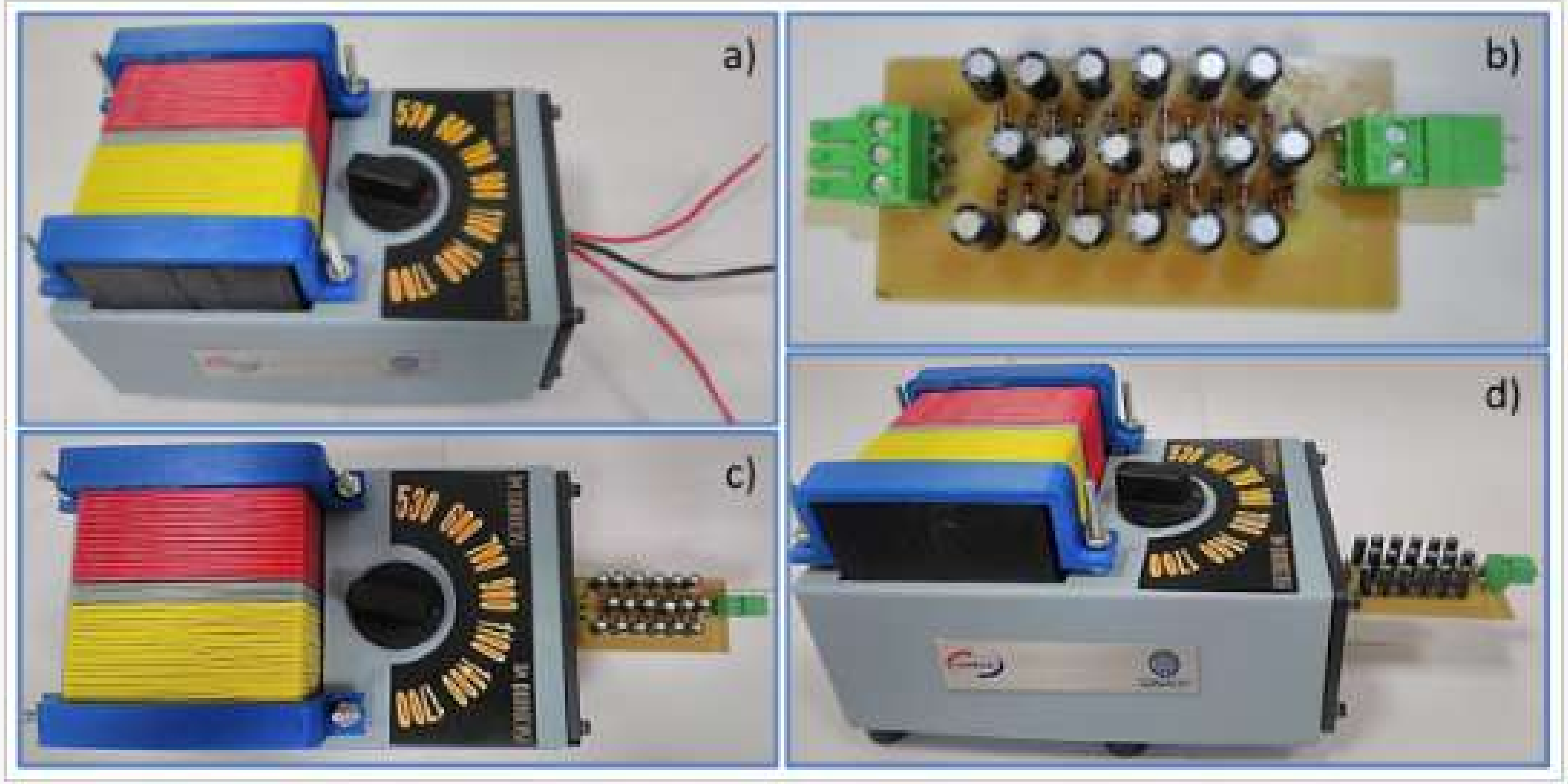

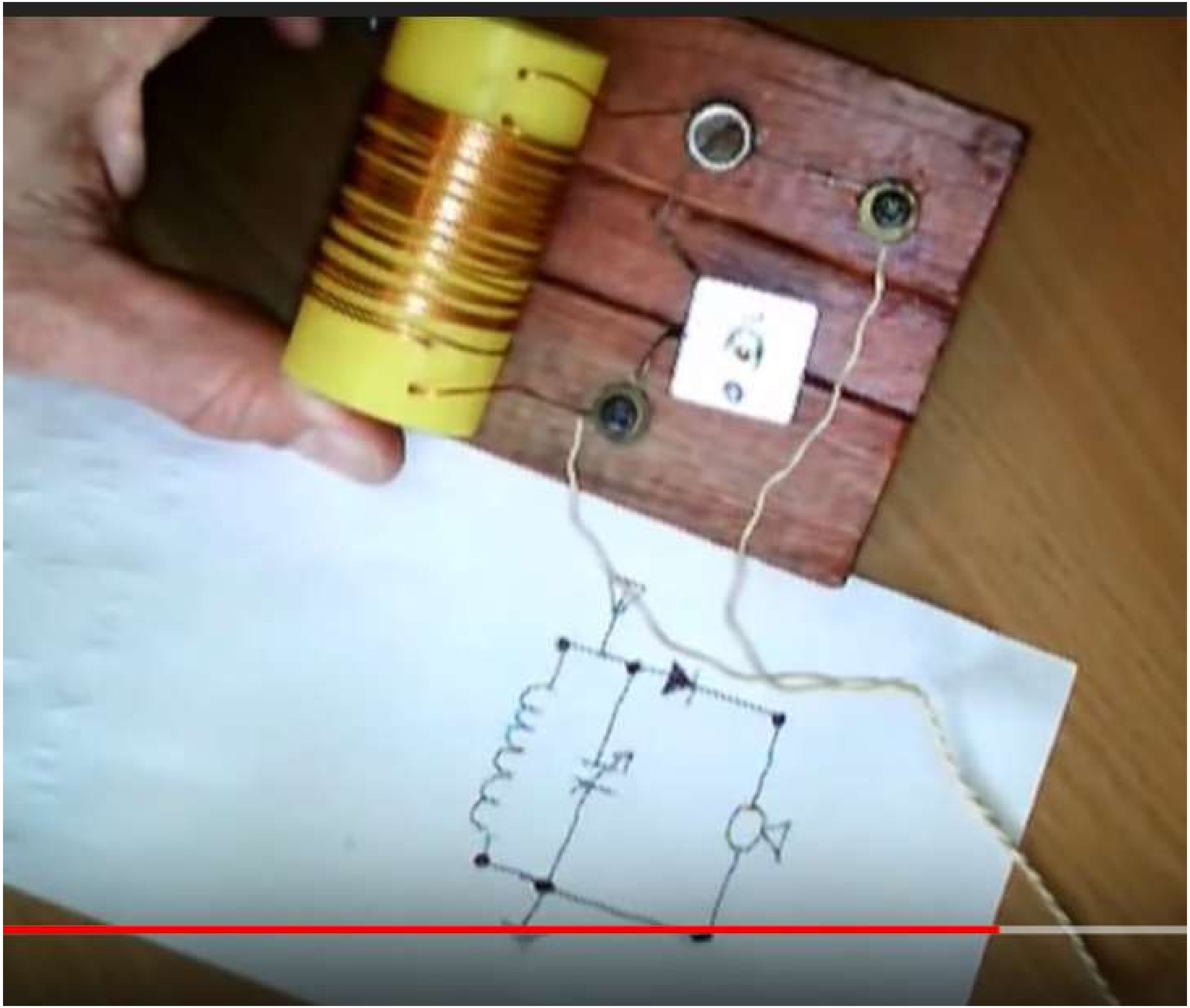

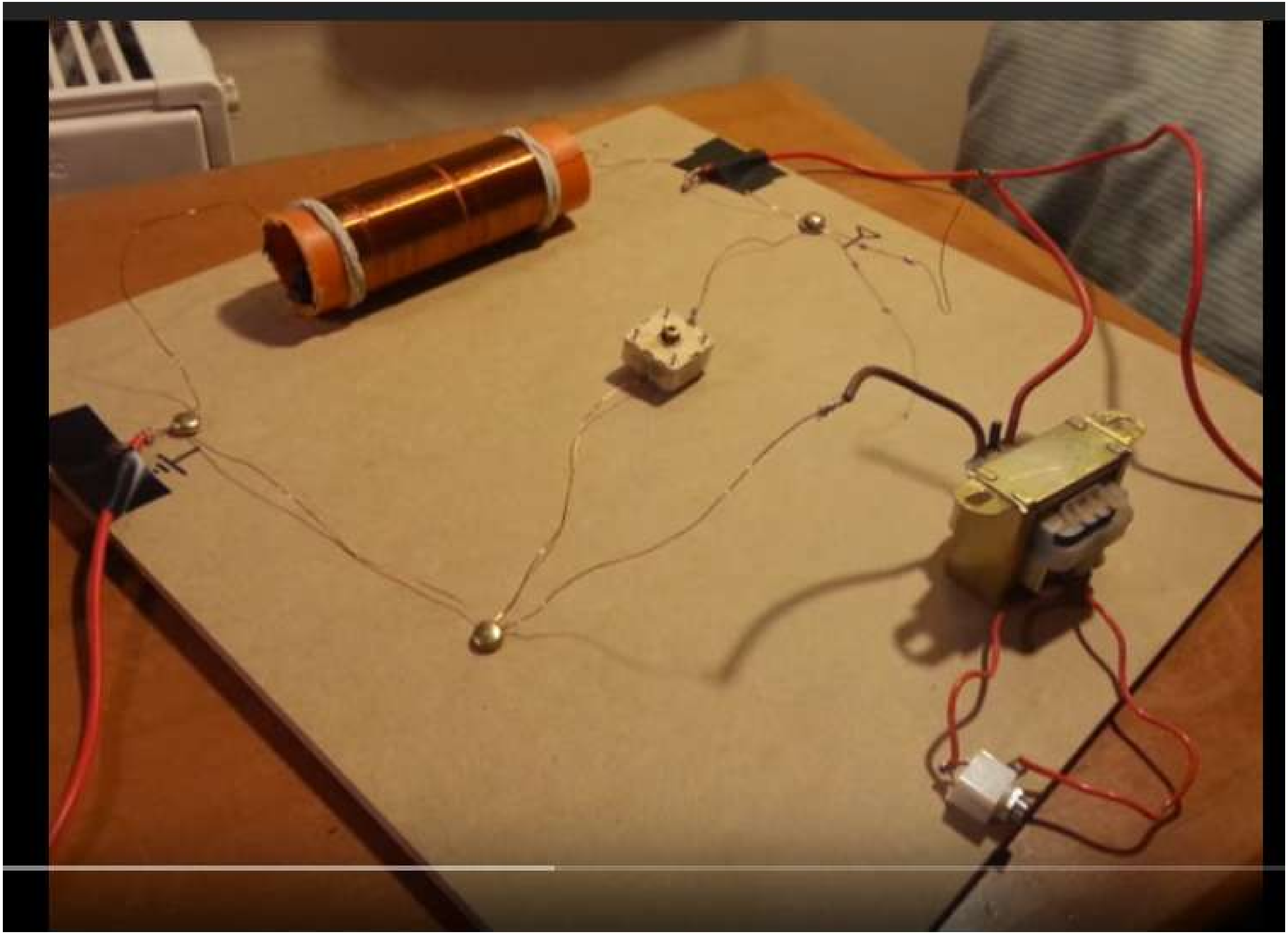

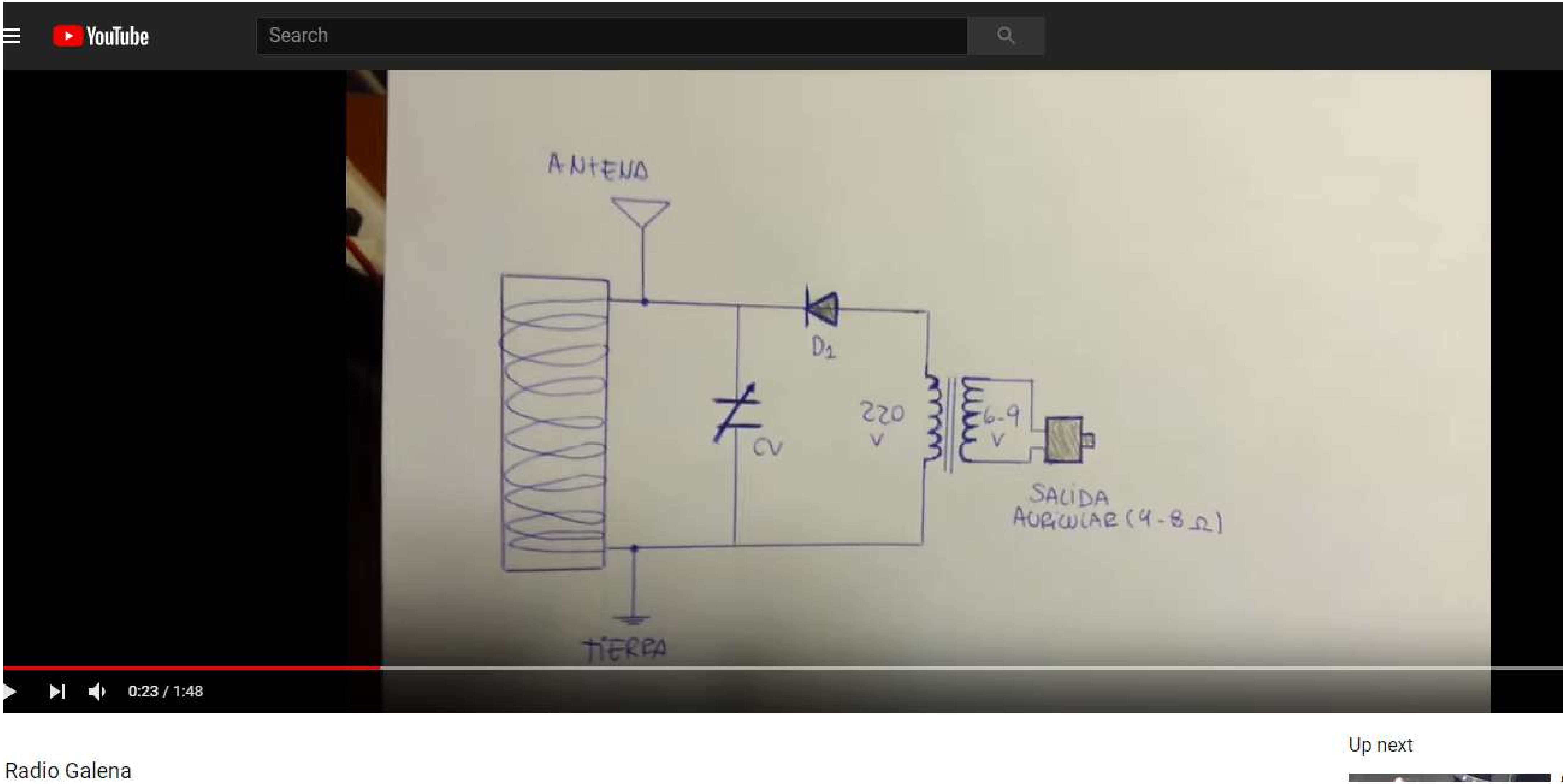
YouTube
Search
ANTENA
D1
CV
220 V
6-9 V
SALIDA AURICULAR (4-8 Ω)
TIERRA
0:23 / 1:48
Up next

Radio Galena

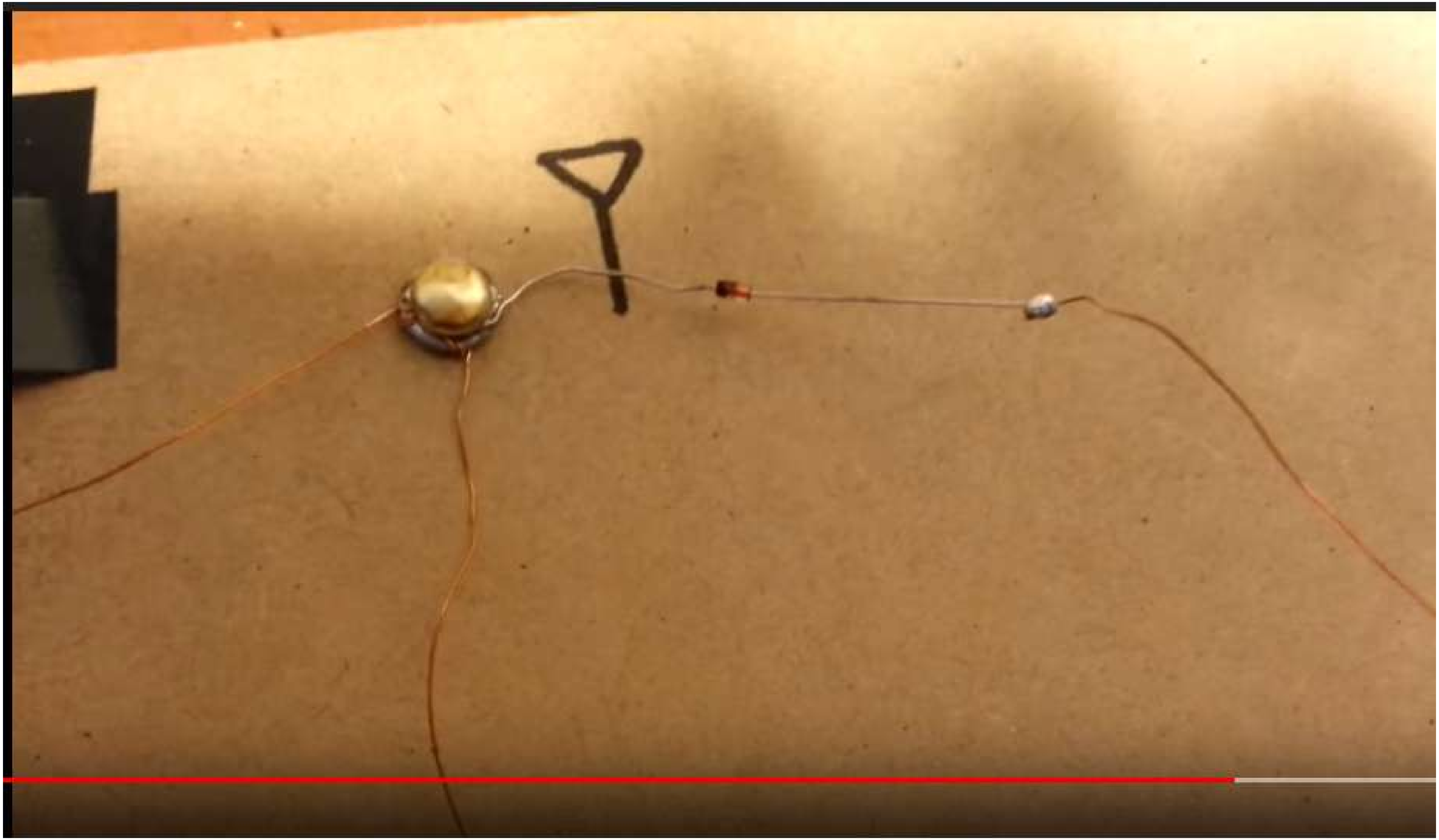

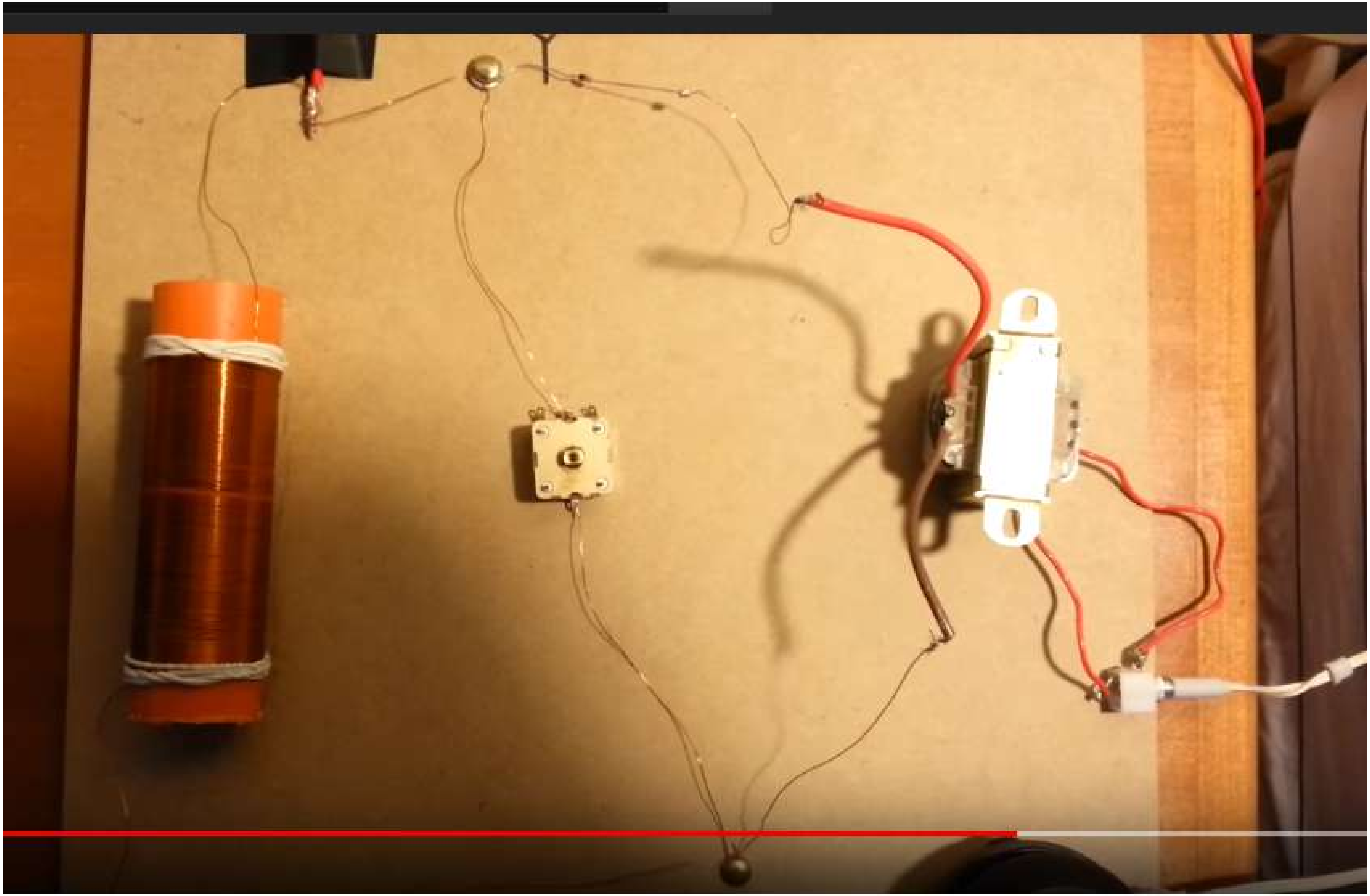

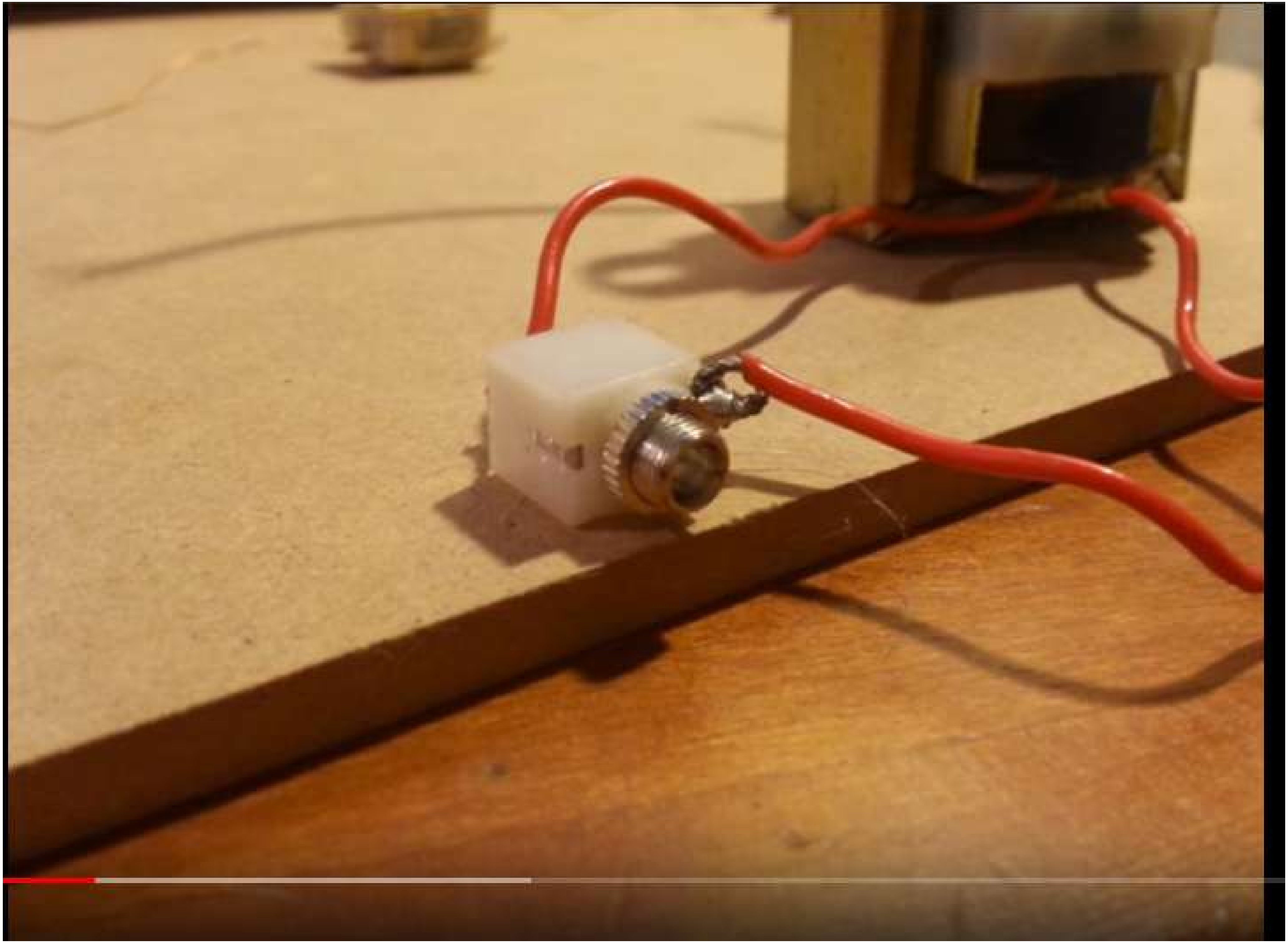

Step 1: LED+RF Diode

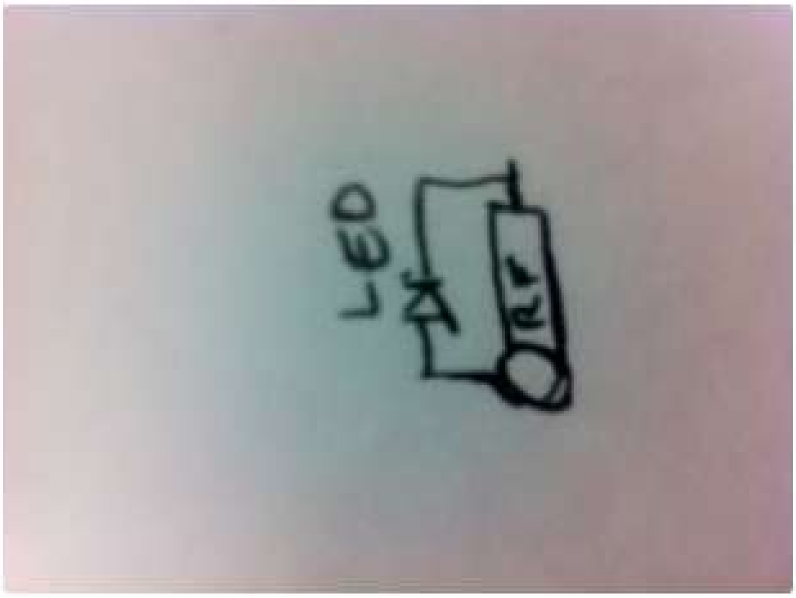

First Solder the led parallel to the Rf diode

Add Tip Ask Question Comment Download

Step 2: RF Diode+ LED+ Wires

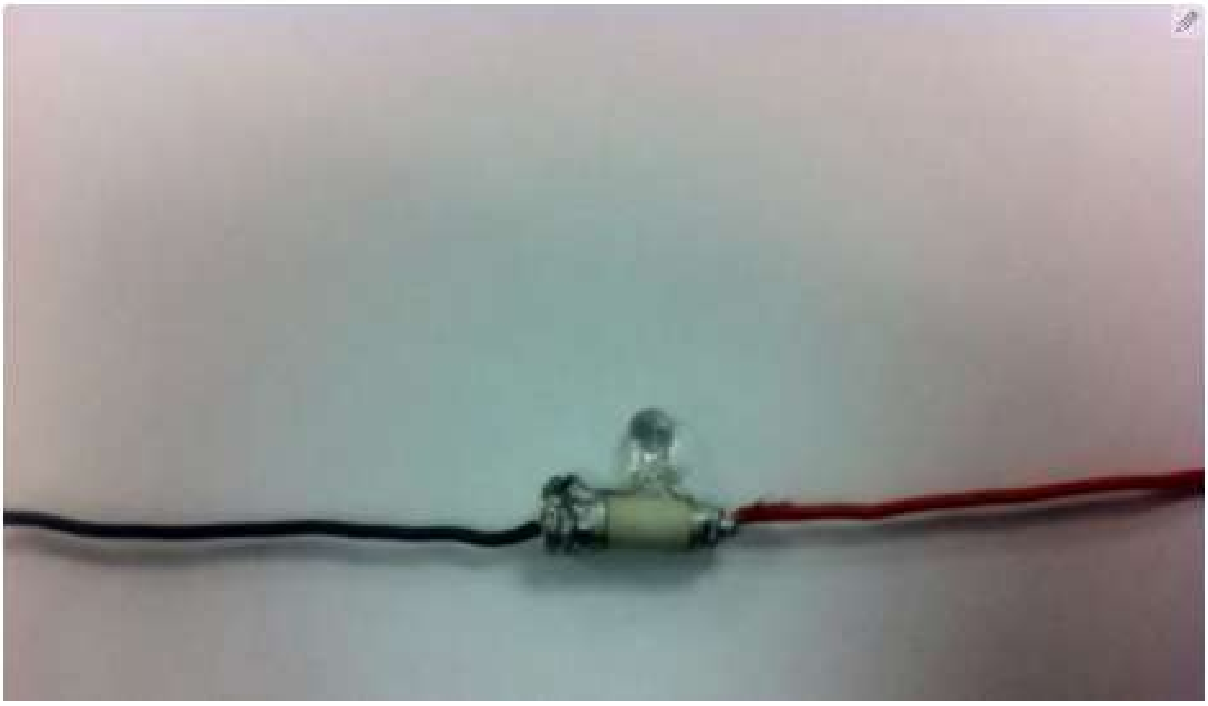

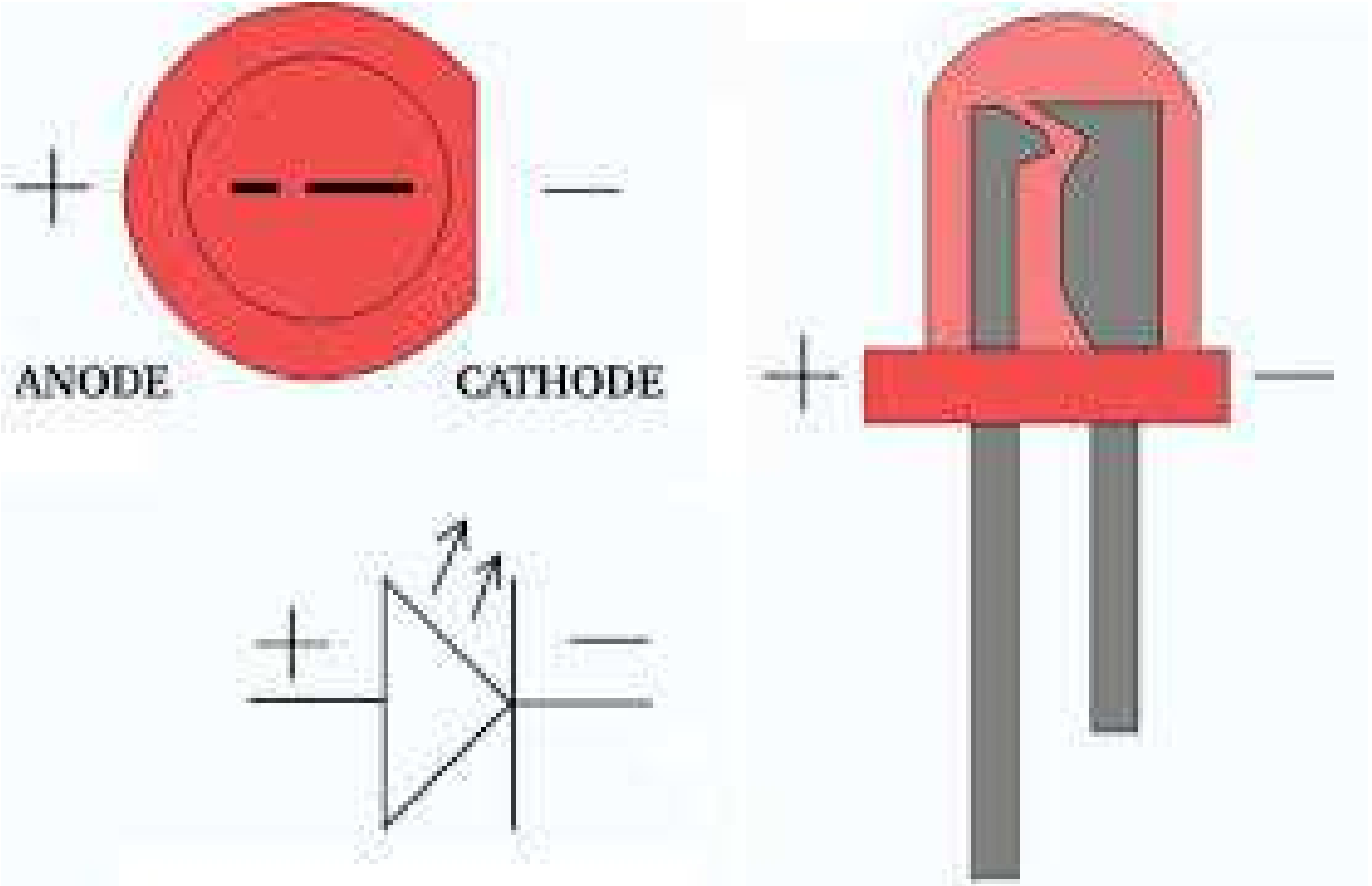
+
—
ANODE
CATHODE
+
—
+
—

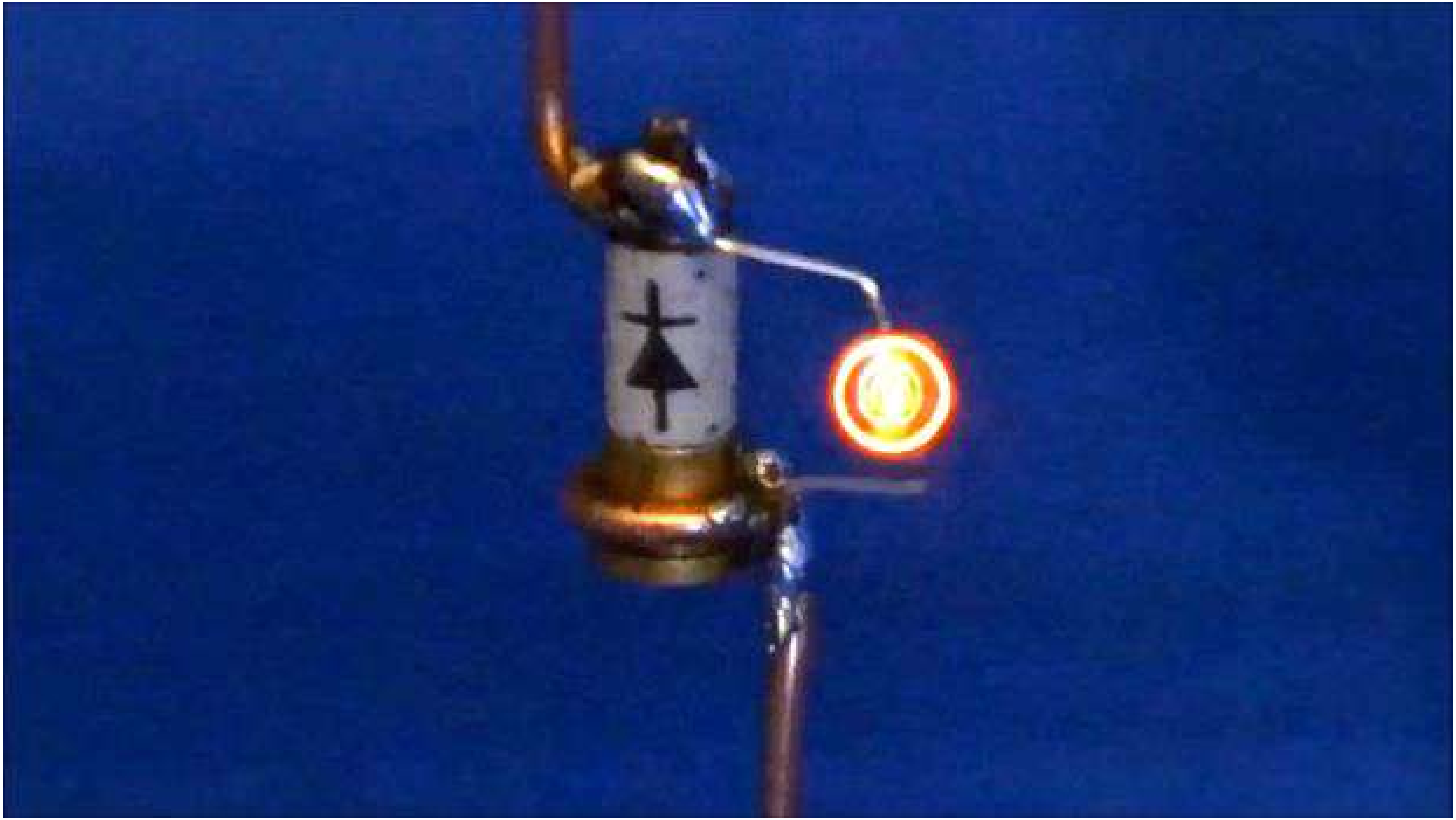

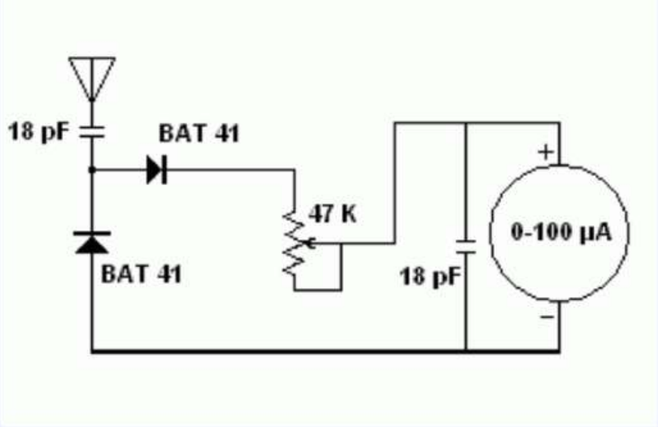
18 pF
BAT 41
47 K
BAT 41
18 pF
+
0-100 μA
−

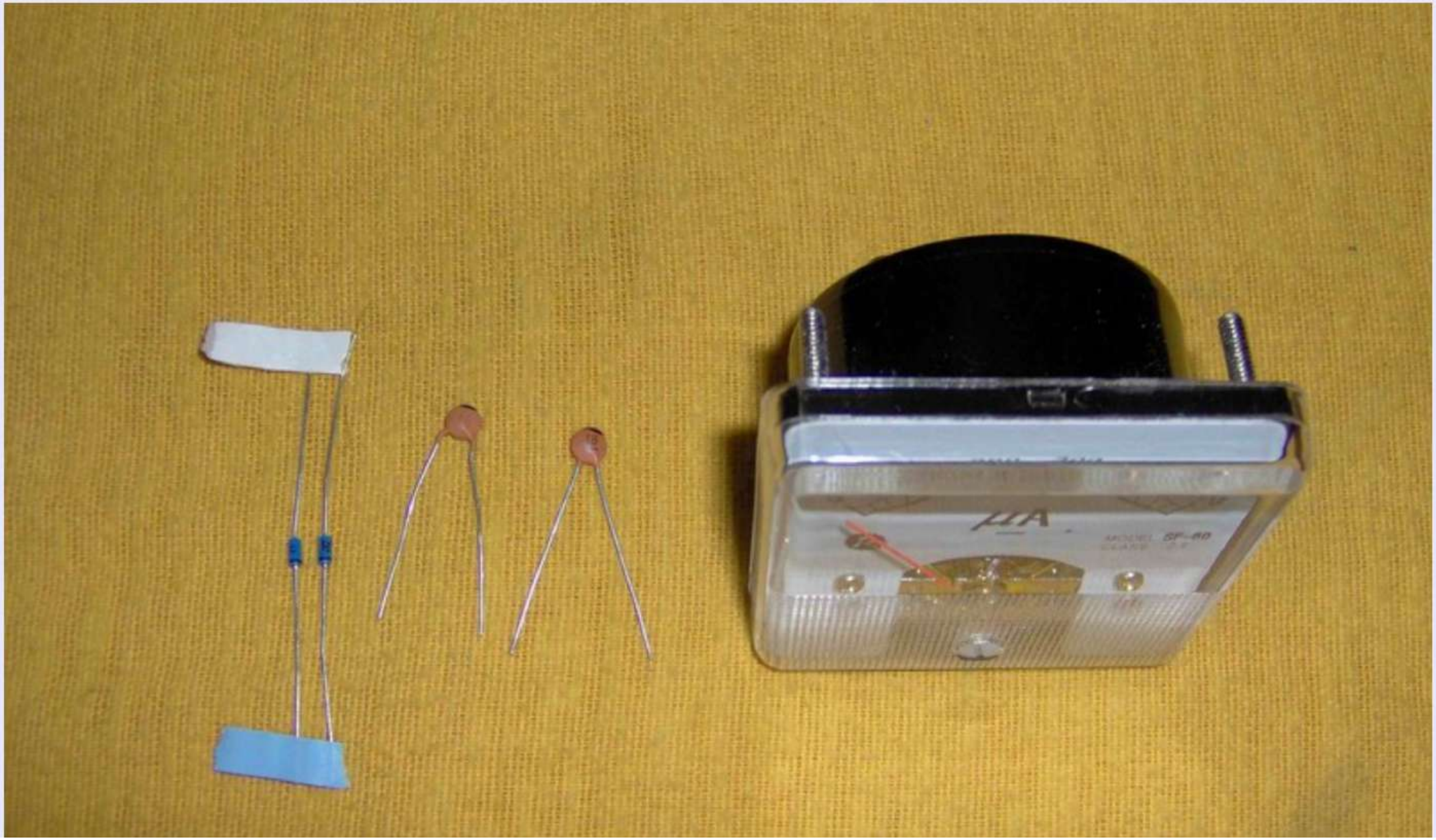
μA

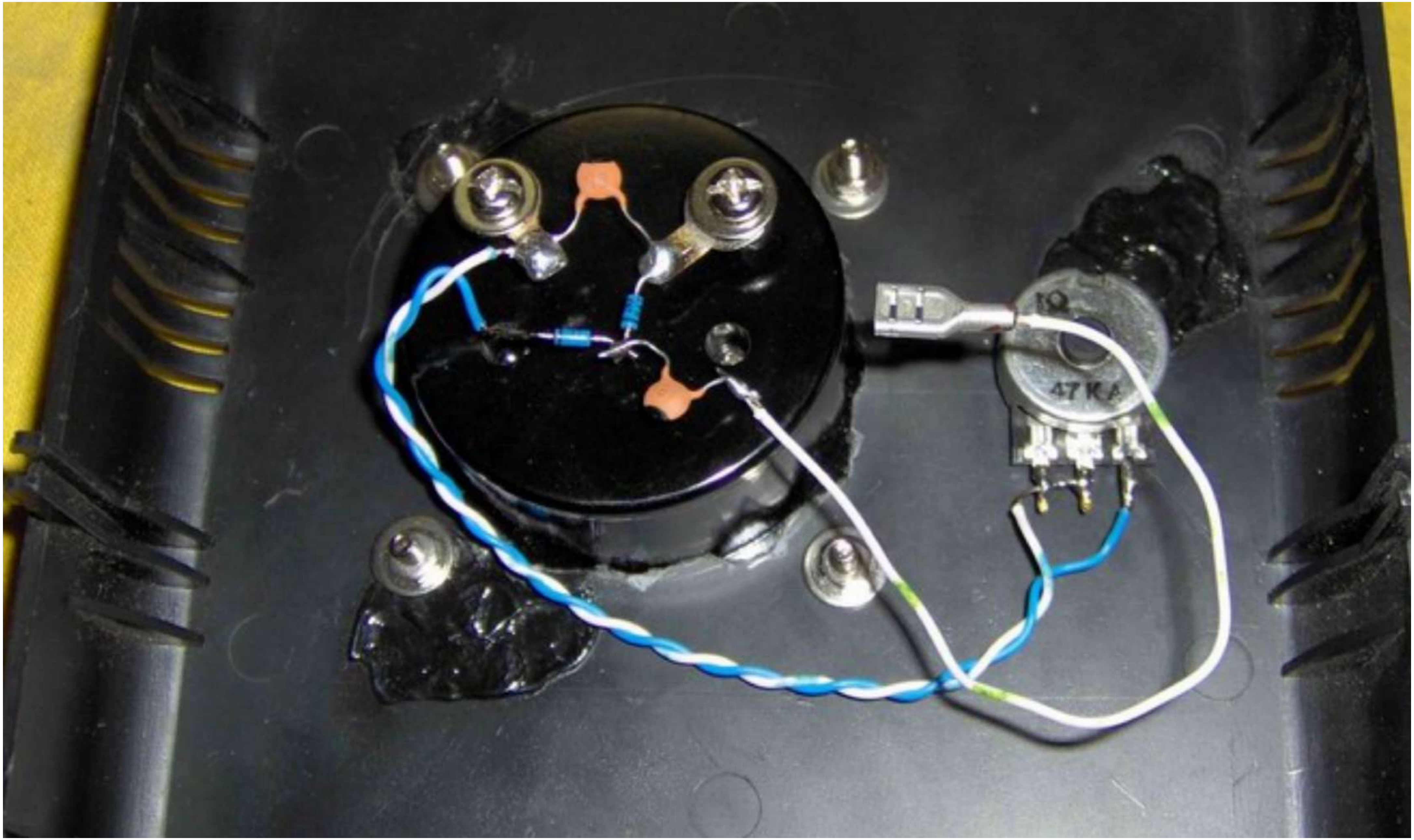
47 KA

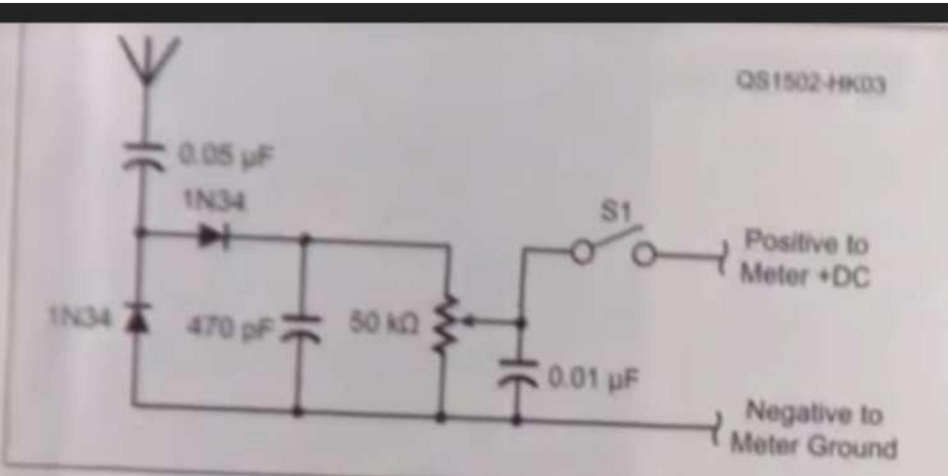
QS1502-HK03
0.05 µF
1N34
S1
Positive to Meter +DC
1N34
470 pF
50 kΩ
0.01 µF
Negative to Meter Ground

Figure 3 — The schematic for the field strength detector board.

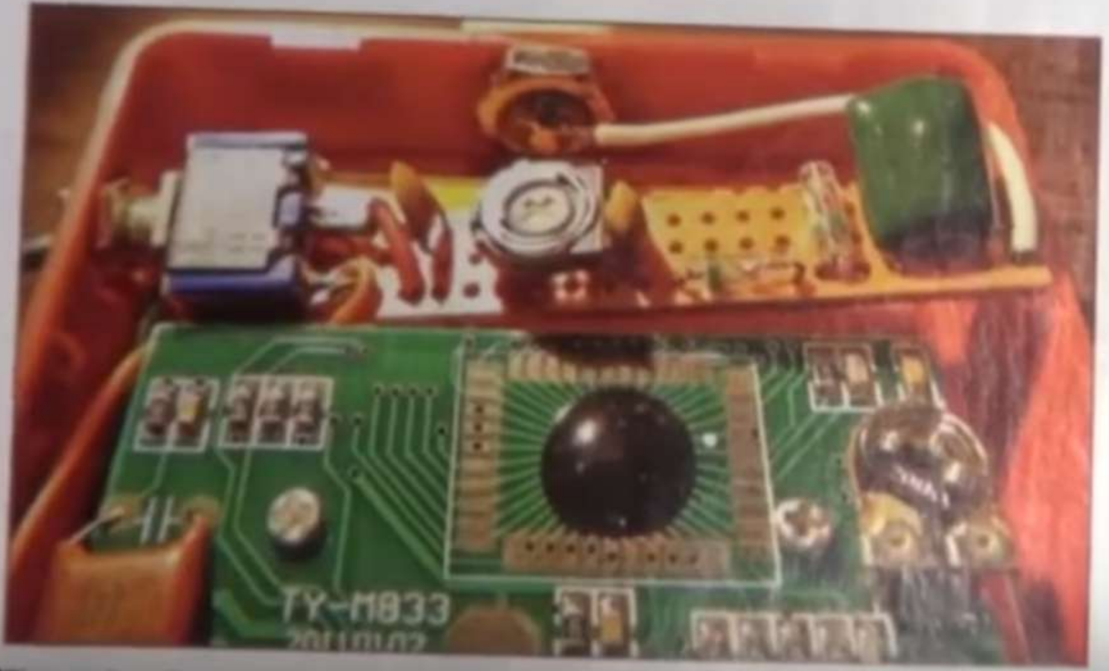

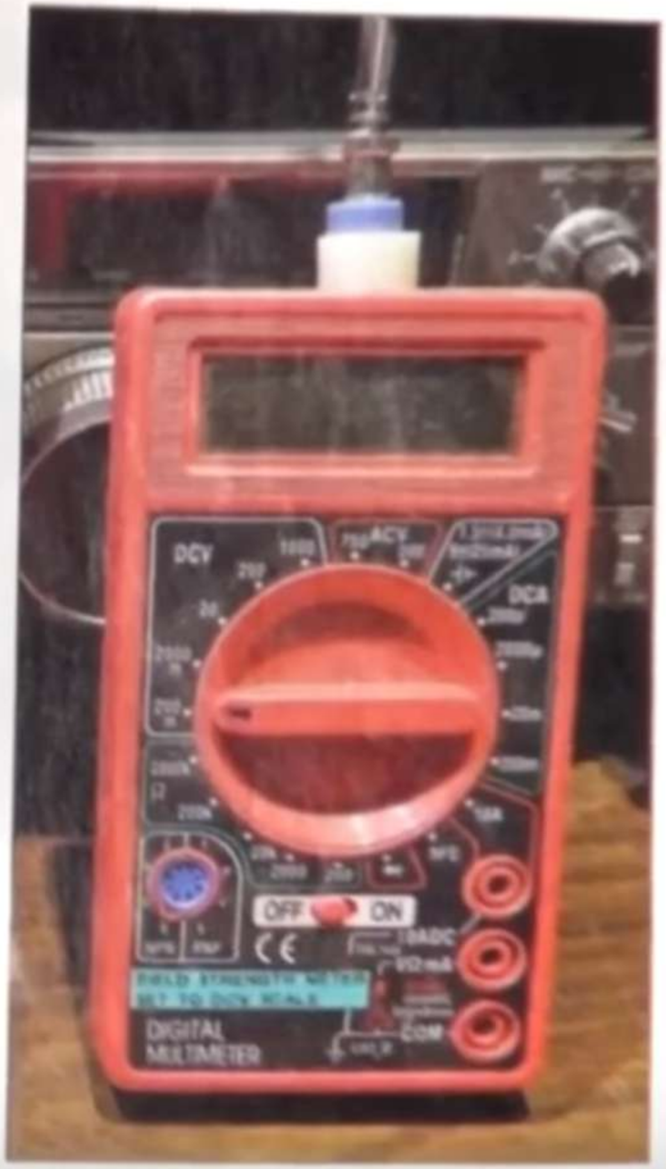
DCV
OFF
ON
FIELD STRENGTH METER
SET TO DCV SCALE
DIGITAL
MULTIMETER
COM

Figure 4 — The detector board mounted inside the multimeter case. The bottom

Figure 5 — The original multimeter and its upgraded companion. [Richard Purse

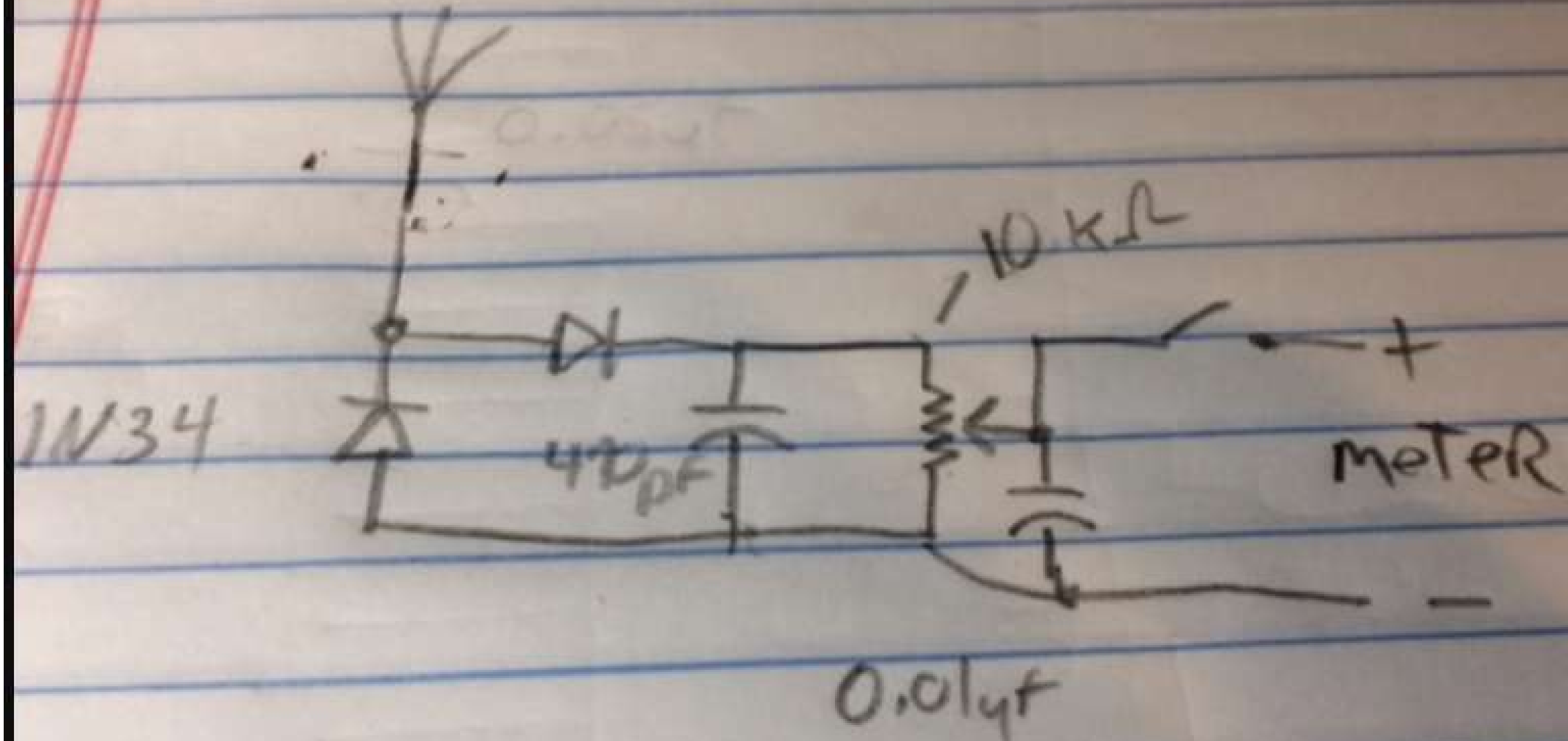
10 kΩ
1N34
470pF
meteR
+
−
0.01uf

1 000 000 000 000

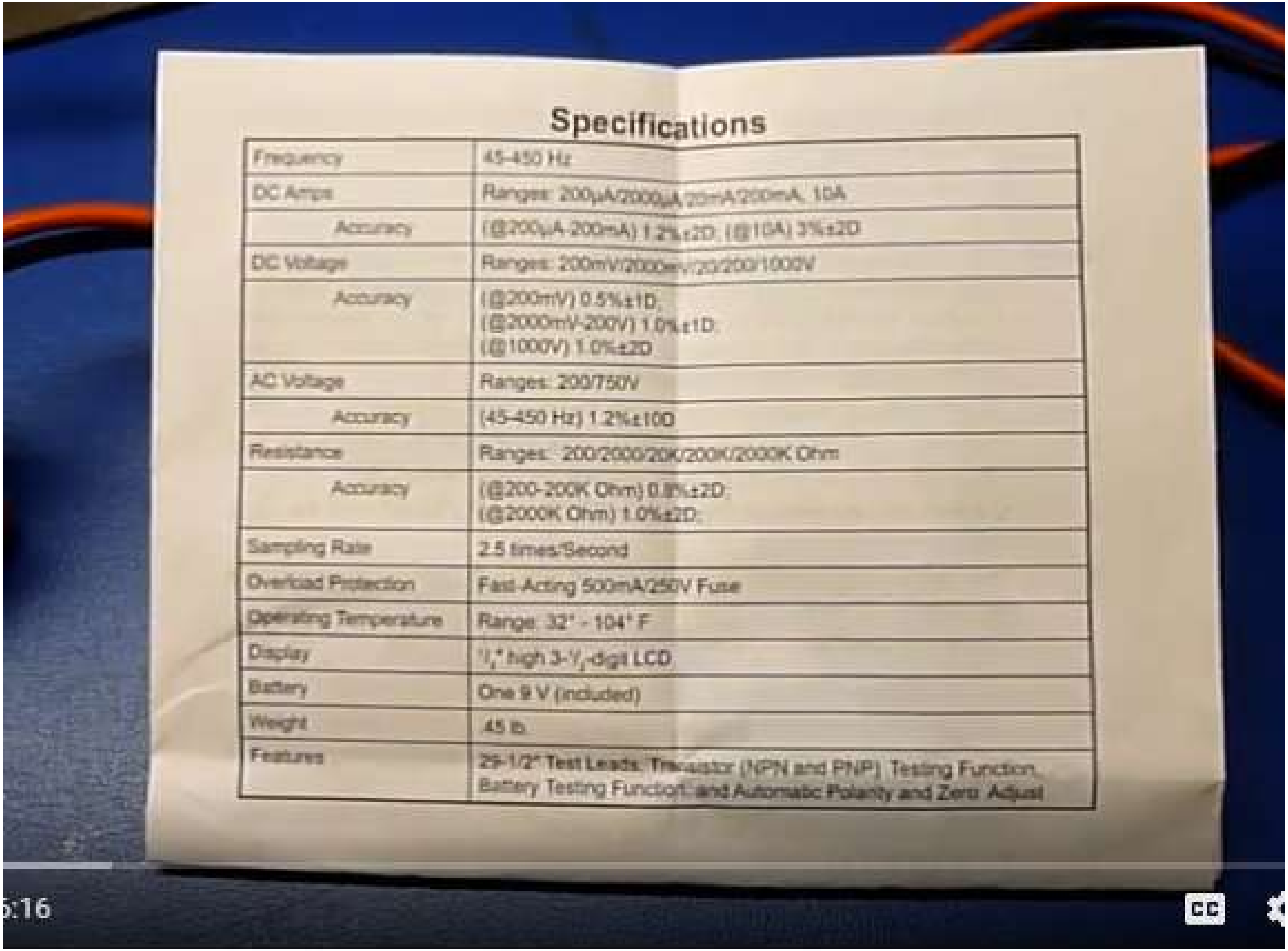

Specifications

Frequency	45-450 Hz
DC Amps	Ranges: 200µA/2000µA/20mA/200mA, 10A
Accuracy	(@200µA-200mA) 1.2%±2D; (@10A) 3%±2D
DC Voltage	Ranges: 200mV/2000mV/20/200/1000V
Accuracy	(@200mV) 0.5%±1D; (@2000mV-200V) 1.0%±1D; (@1000V) 1.0%±2D
AC Voltage	Ranges: 200/750V
Accuracy	(45-450 Hz) 1.2%±10D
Resistance	Ranges: 200/2000/20K/200K/2000K Ohm
Accuracy	(@200-200K Ohm) 0.8%±2D; (@2000K Ohm) 1.0%±2D;
Sampling Rate	2.5 times/Second
Overload Protection	Fast-Acting 500mA/250V Fuse
Operating Temperature	Range: 32° - 104° F
Display	½" high 3-½-digit LCD
Battery	One 9 V (included)
Weight	.45 lb.
Features	29-1/2" Test Leads; Transistor (NPN and PNP) Testing Function, Battery Testing Function and Automatic Polarity and Zero Adjust

CEN-TECH
7 FUNCTION
DIGITAL
MULTIMETER
OFF
ON
CE
CEN-TECH
DIGITAL
MULTIMETER

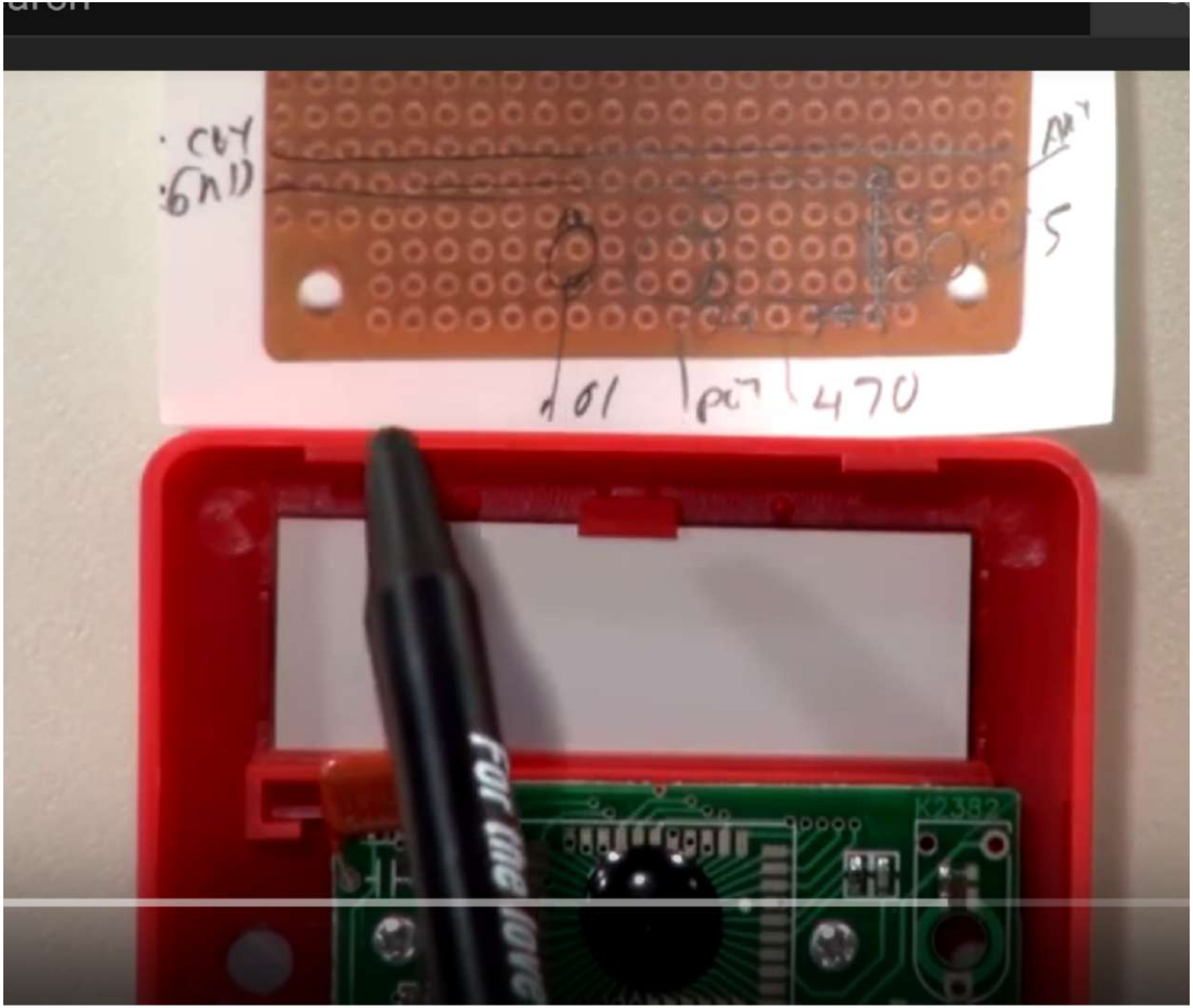
CBY
GND
ANT
5
.01
pot
470
For the love
K2382

Up next

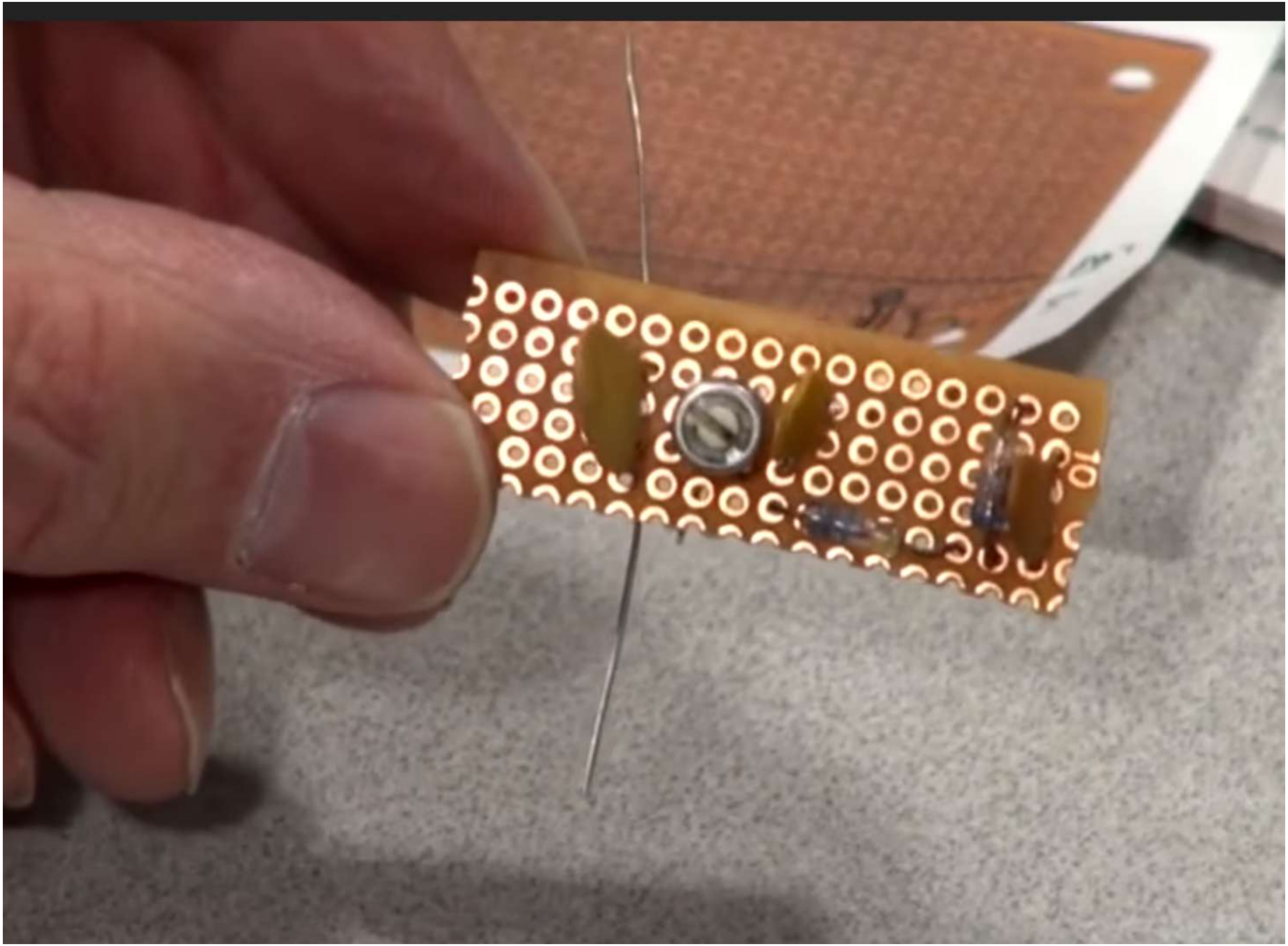

10

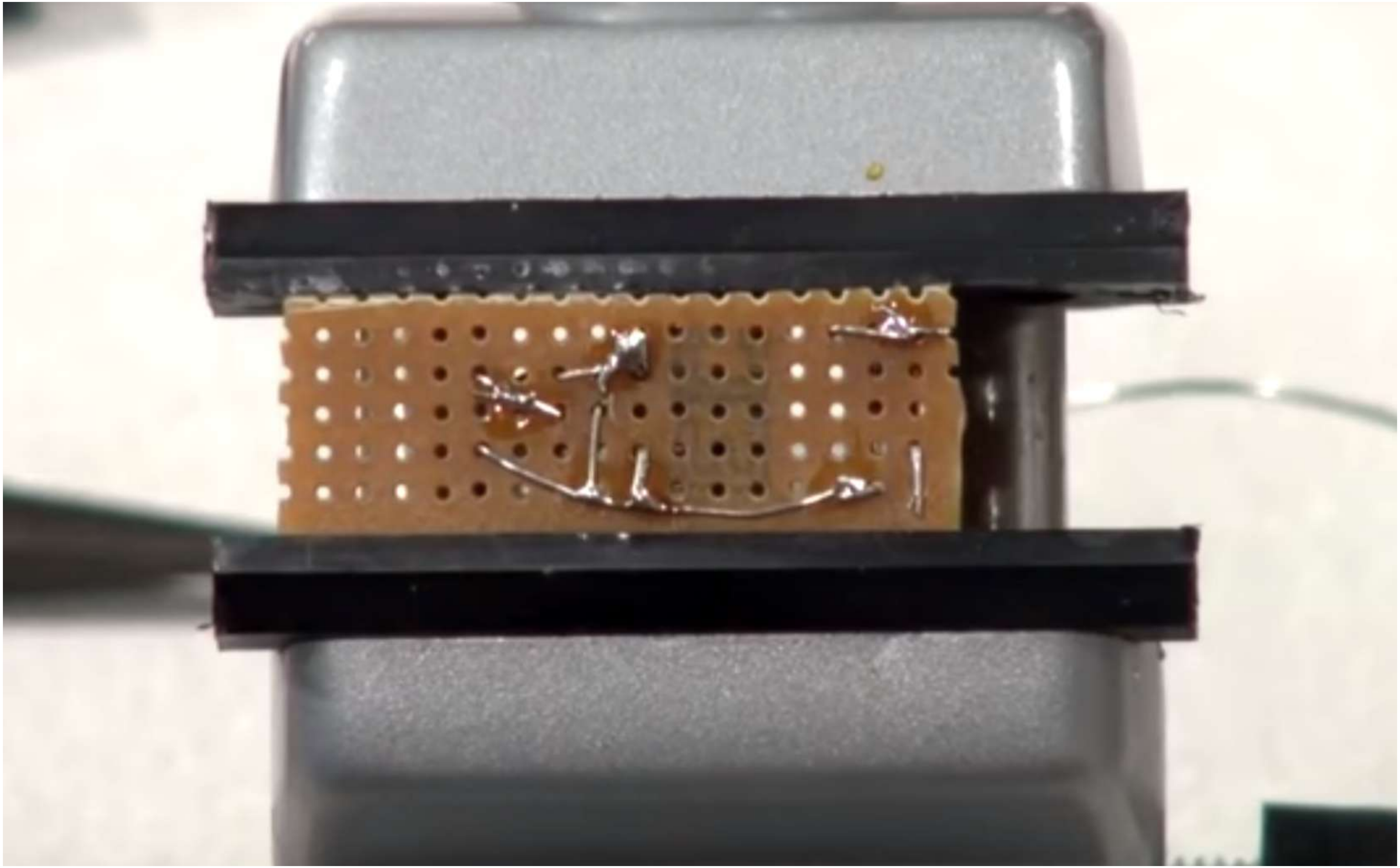

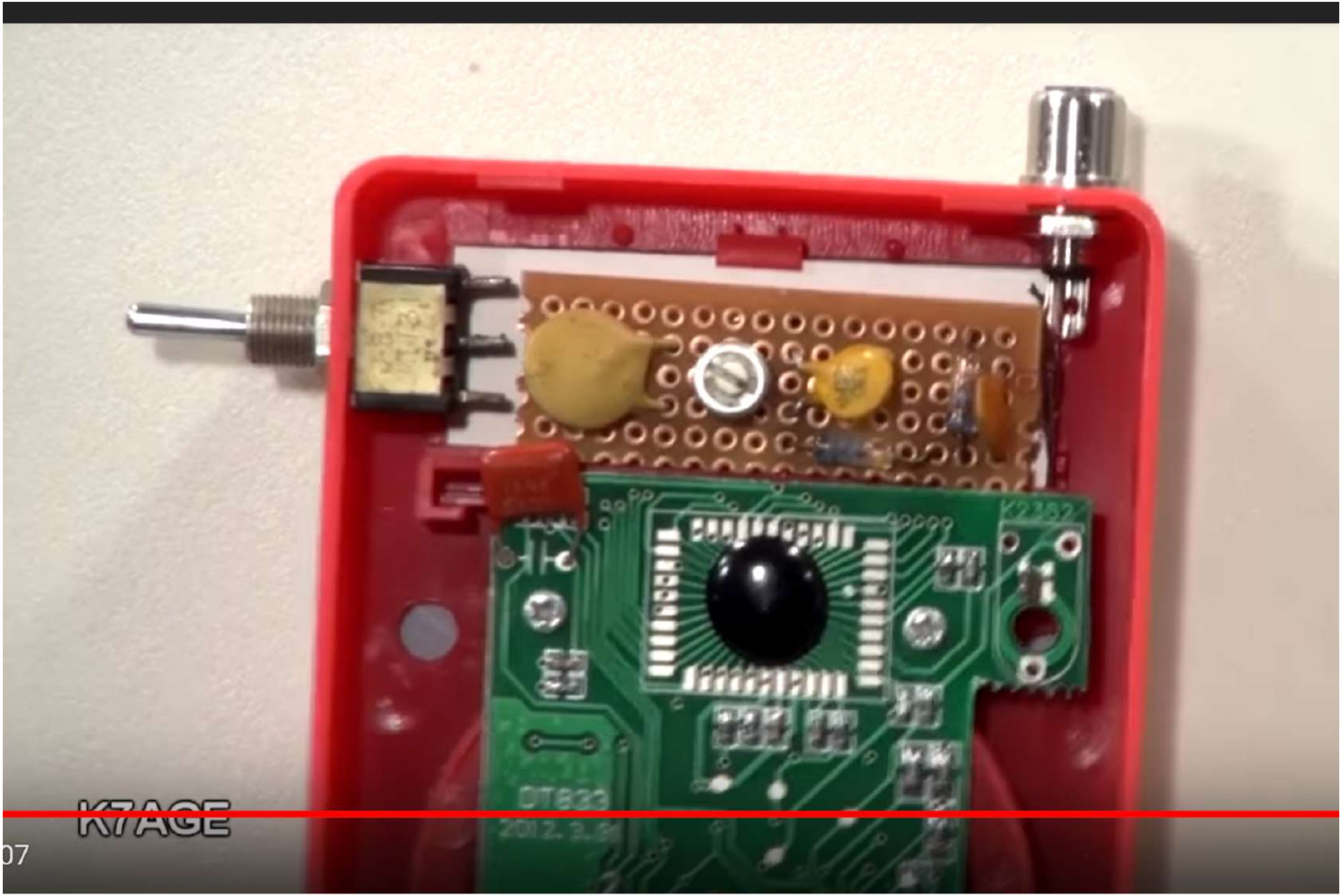
K7AGE
DT833
2012. 3. 8
K2382
07

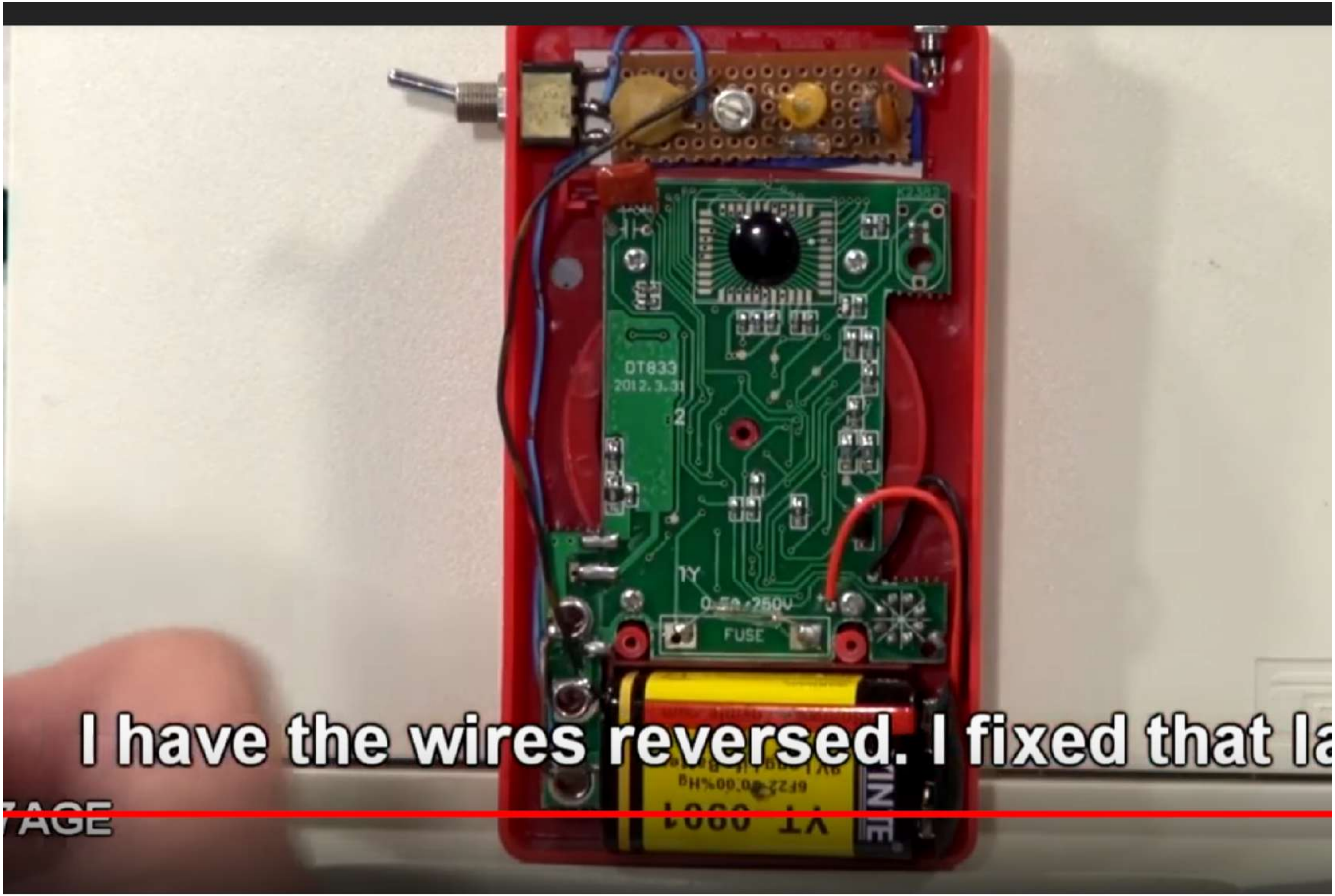
DT833
2012.3.31
0.5A/250V
FUSE
AGE

I have the wires reversed. I fixed that la

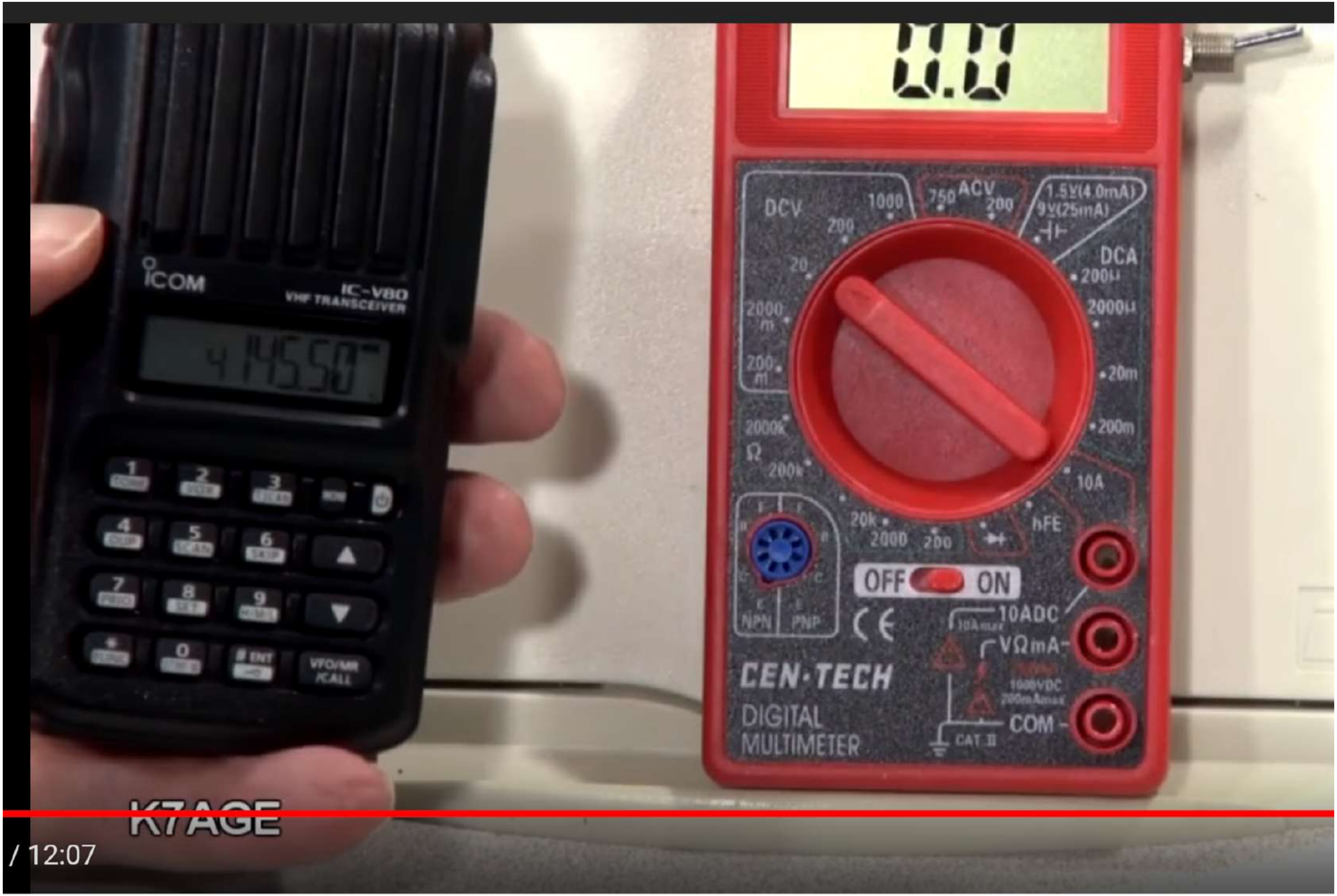
ICOM
IC-V80
VHF TRANSCEIVER
145.50
0.0
DCV
ACV
DCA
OFF ON
10ADC
VΩmA
COM
CEN·TECH
DIGITAL MULTIMETER
K7AGE
/ 12:07

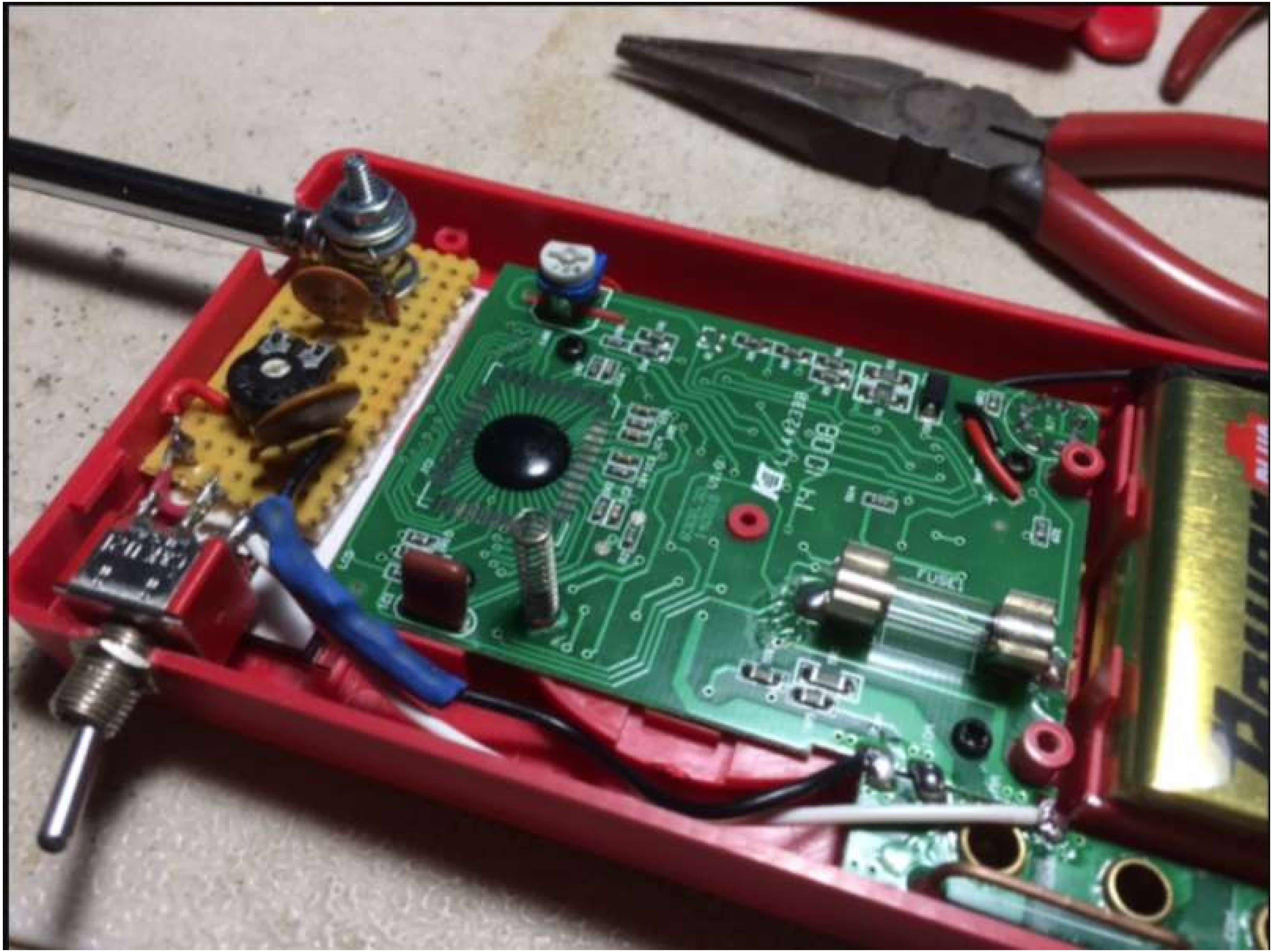
FUSE1

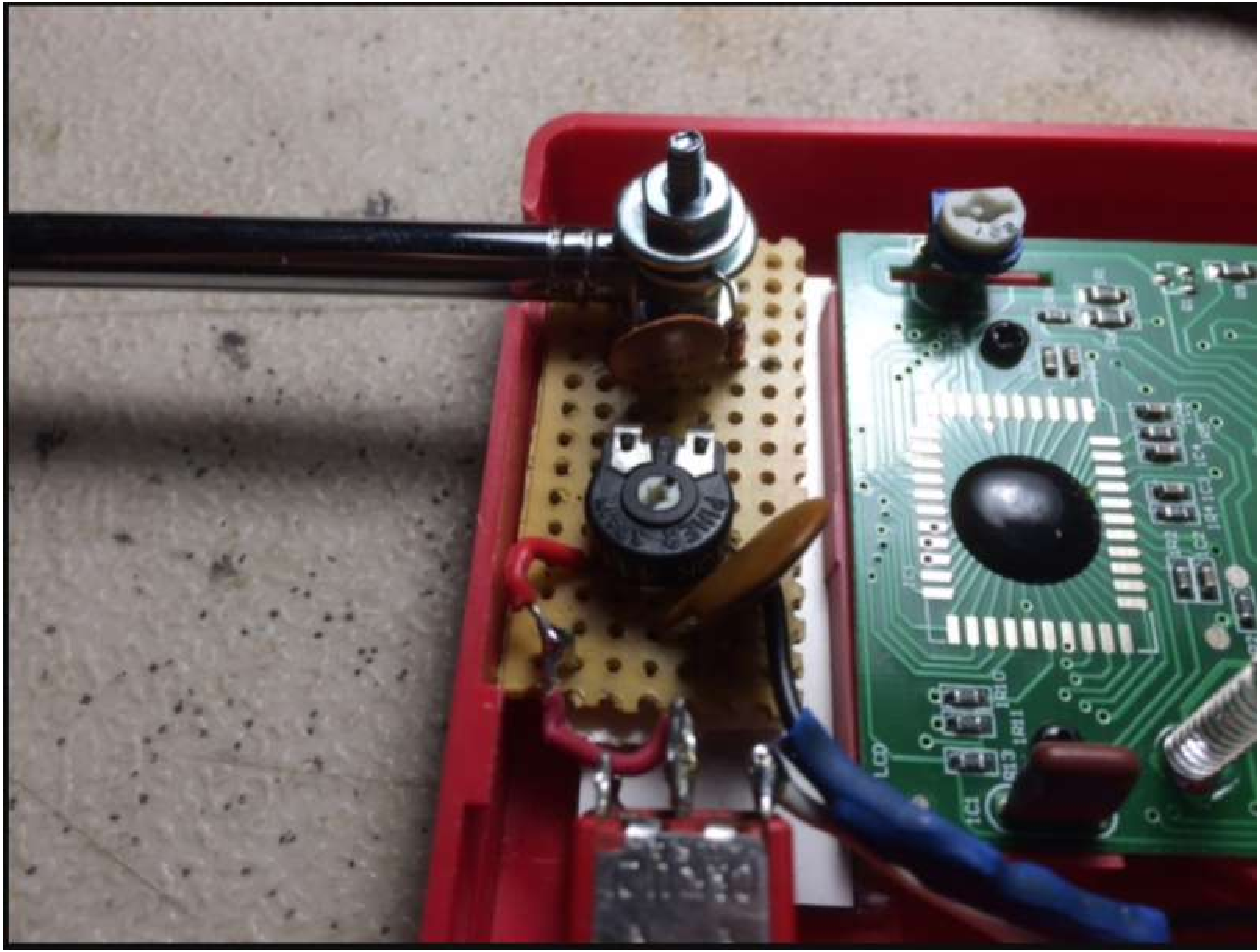

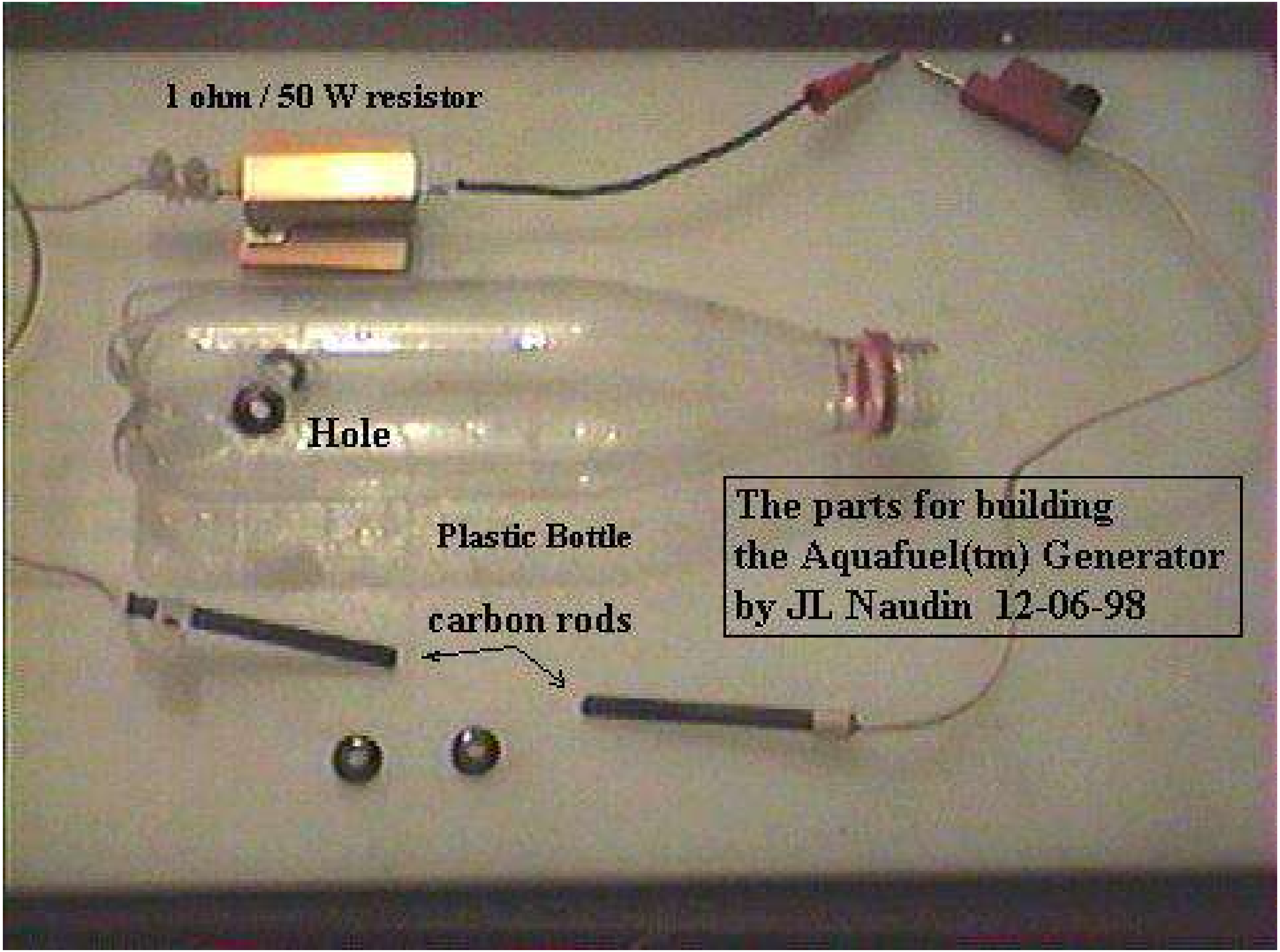
1 ohm / 50 W resistor
Hole
Plastic Bottle
carbon rods

The parts for building
the Aquafuel(tm) Generator
by JL Naudin 12-06-98

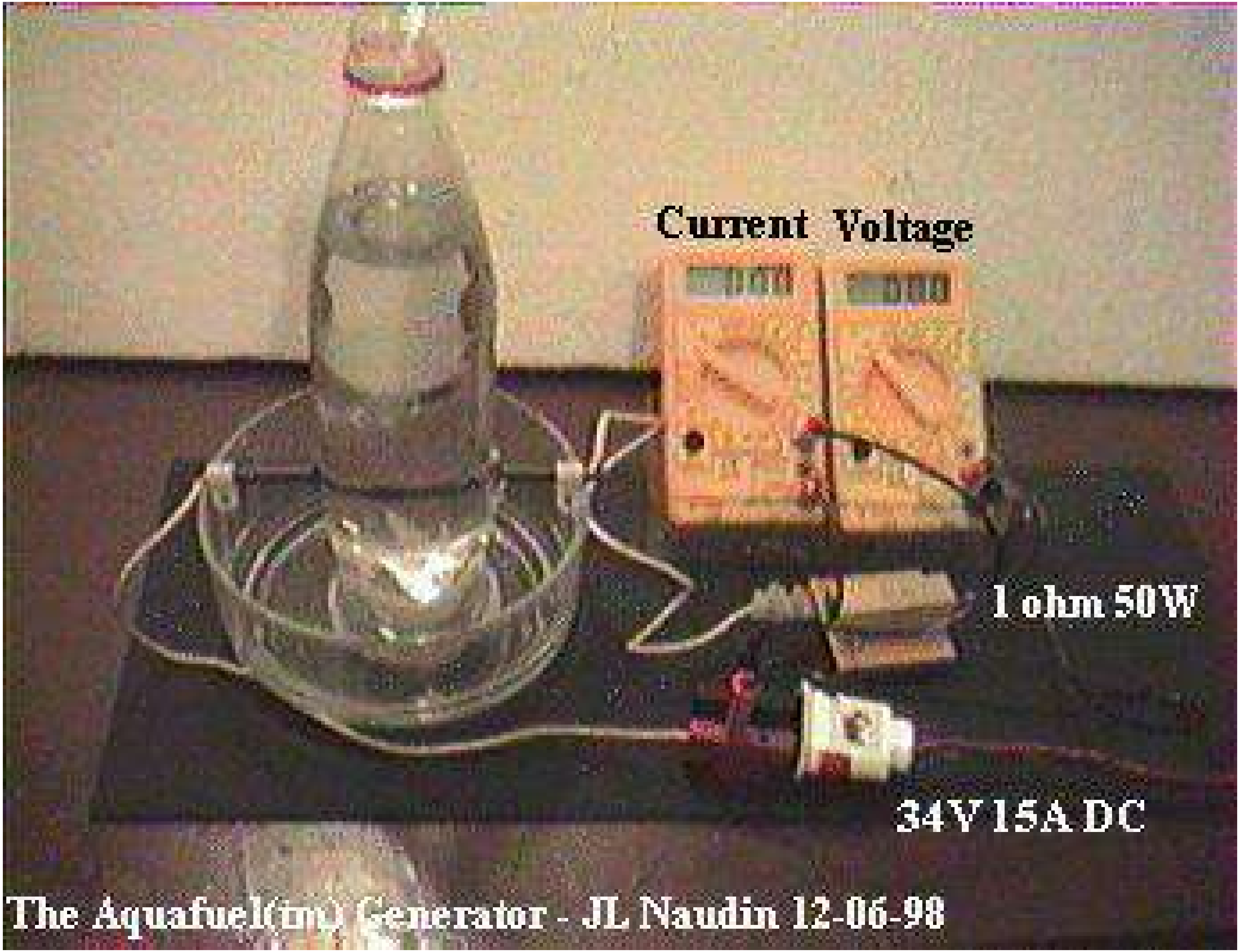
Current
Voltage
1 ohm 50W
34V 15A DC

The Aquafuel(tm) Generator - JL Naudin 12-06-98

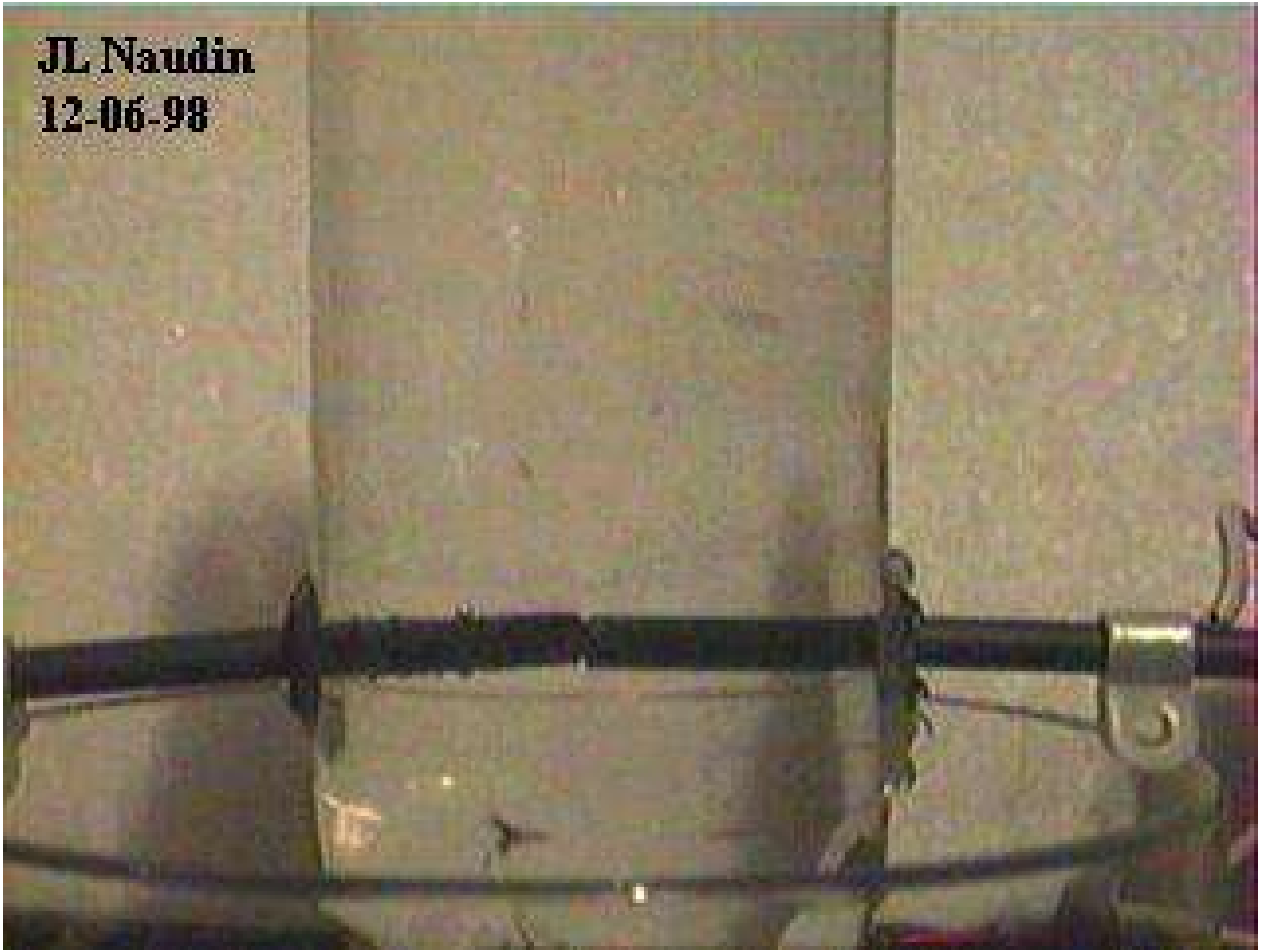
JL Naudin
12-06-98

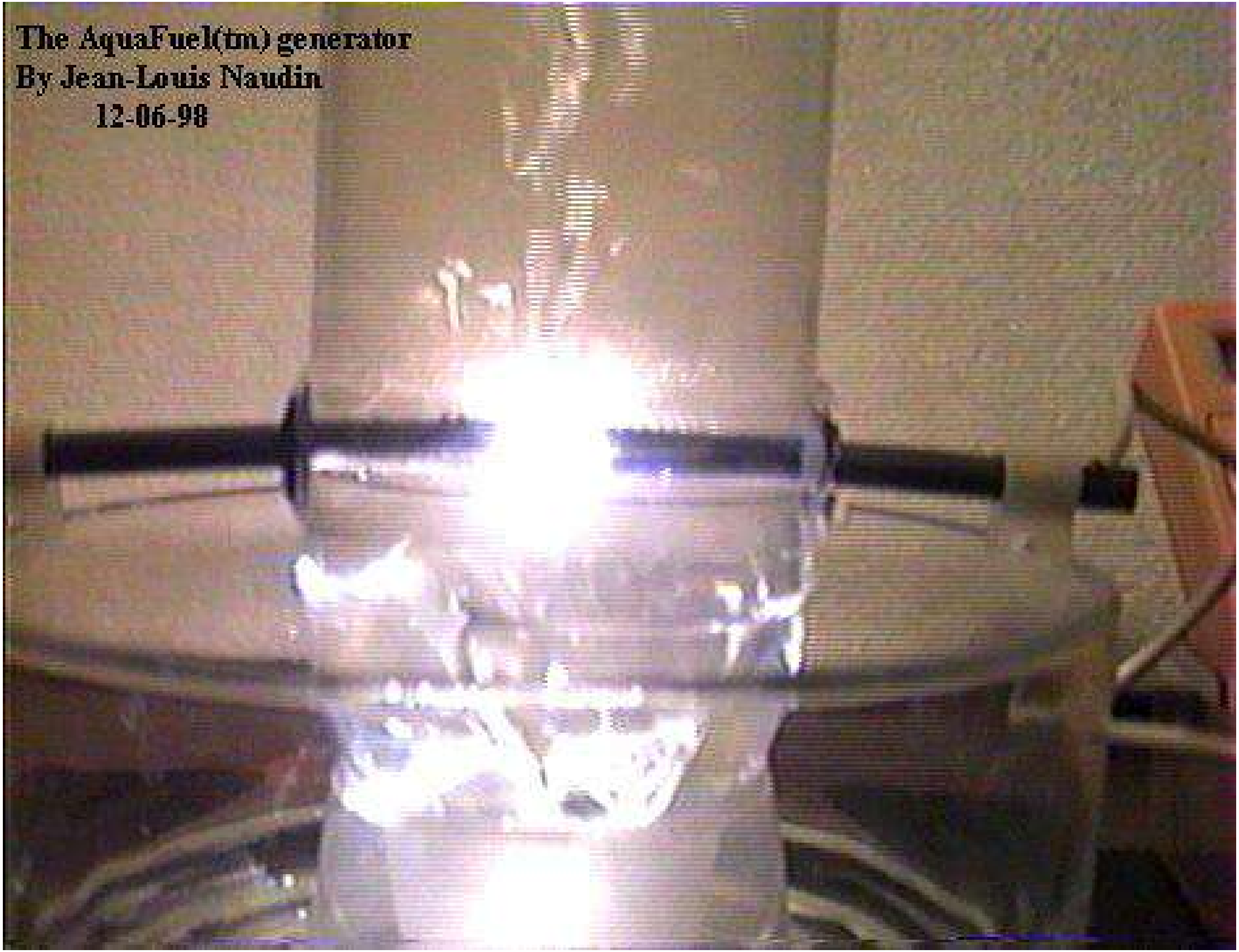
The AquaFuel(tm) generator
By Jean-Louis Naudin
12-06-98

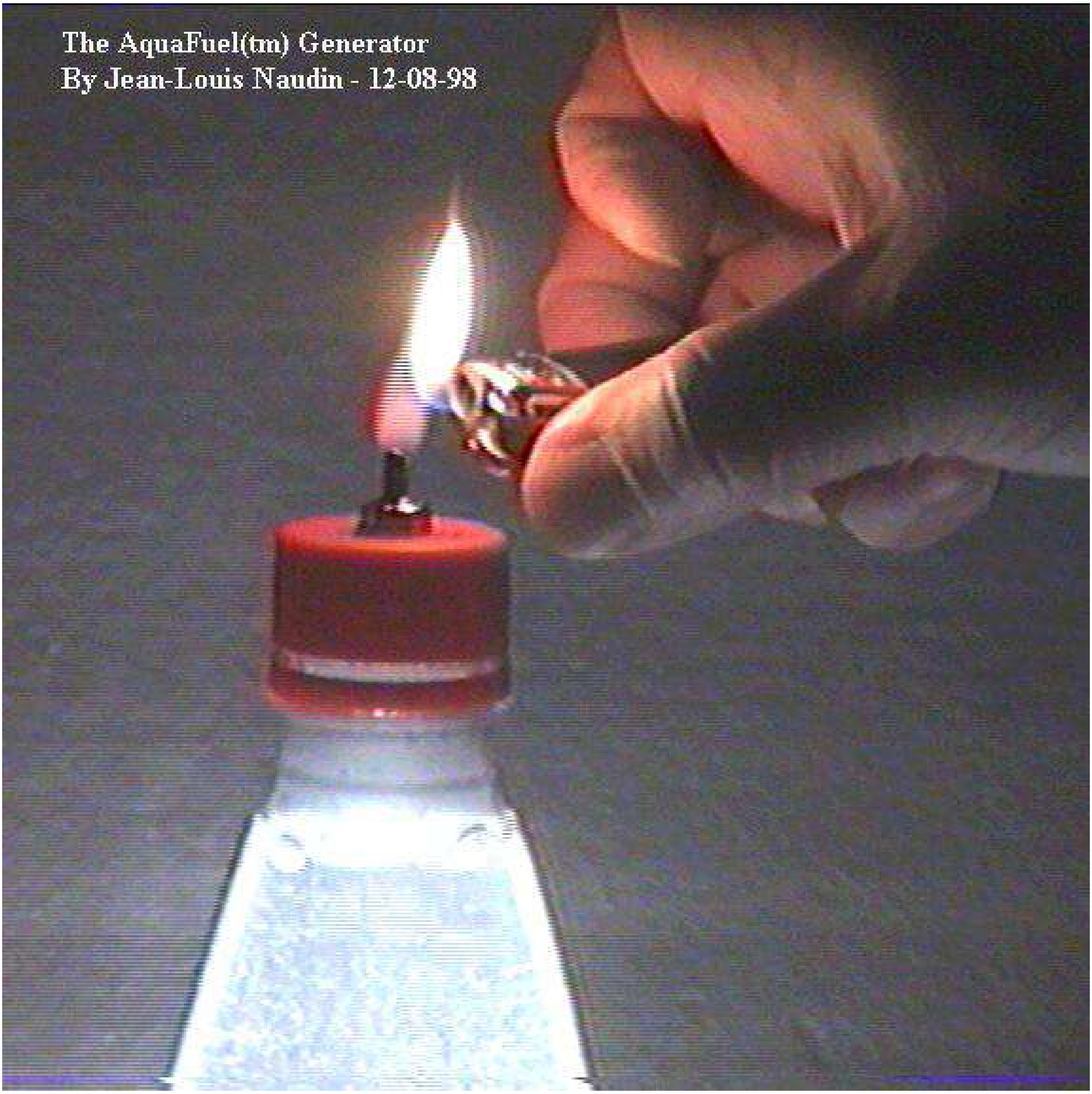
The AquaFuel(tm) Generator
By Jean-Louis Naudin - 12-08-98

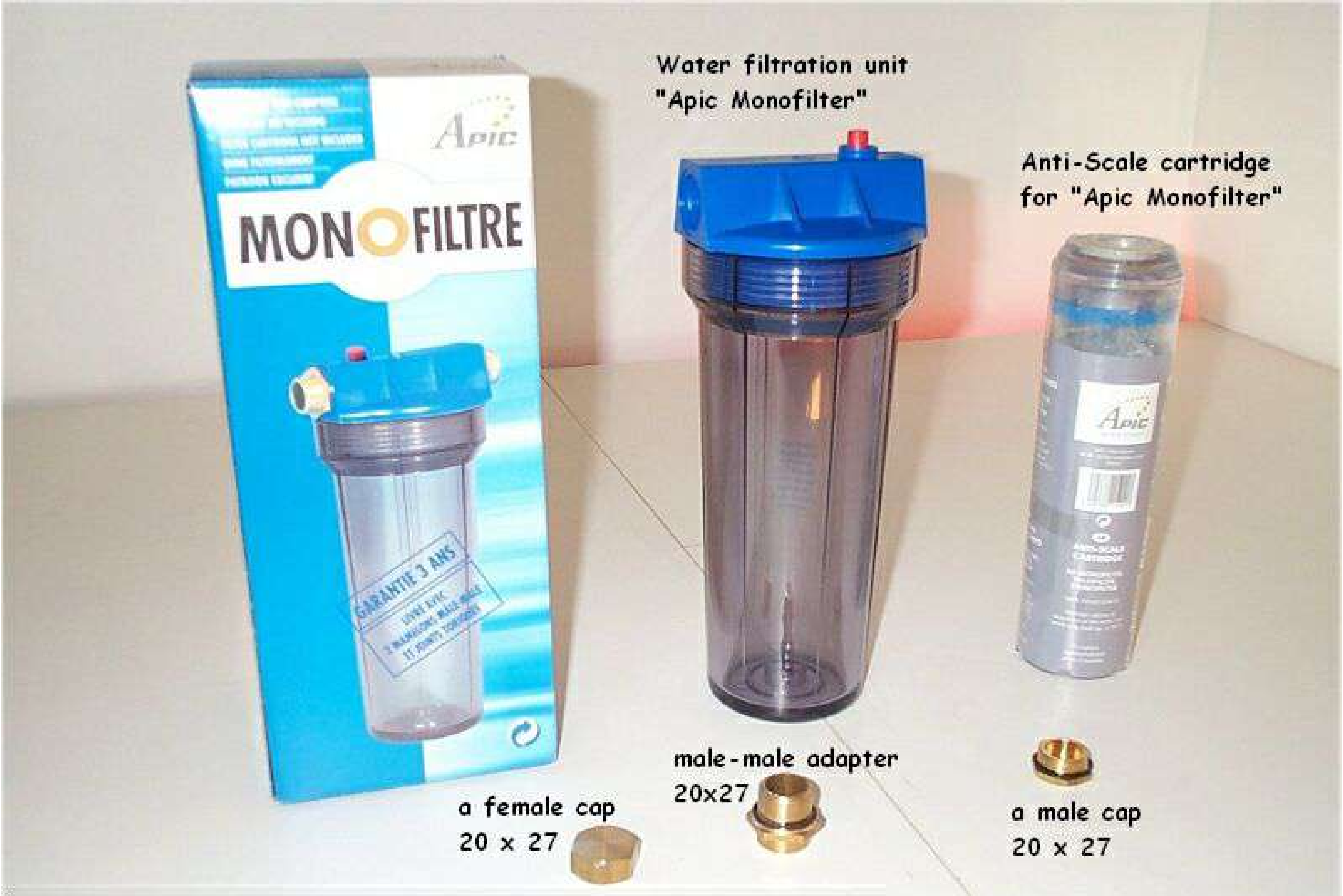
Water filtration unit
"Apic Monofilter"
Anti-Scale cartridge
for "Apic Monofilter"
Apic
MONOFILTRE
GARANTIE 3 ANS
male-male adapter
20x27
a female cap
20 x 27
a male cap
20 x 27

BingoFuel Reactor v1.1 - The components for the reactor and and reaction chamber
(c) 2003 by Jean-Louis Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

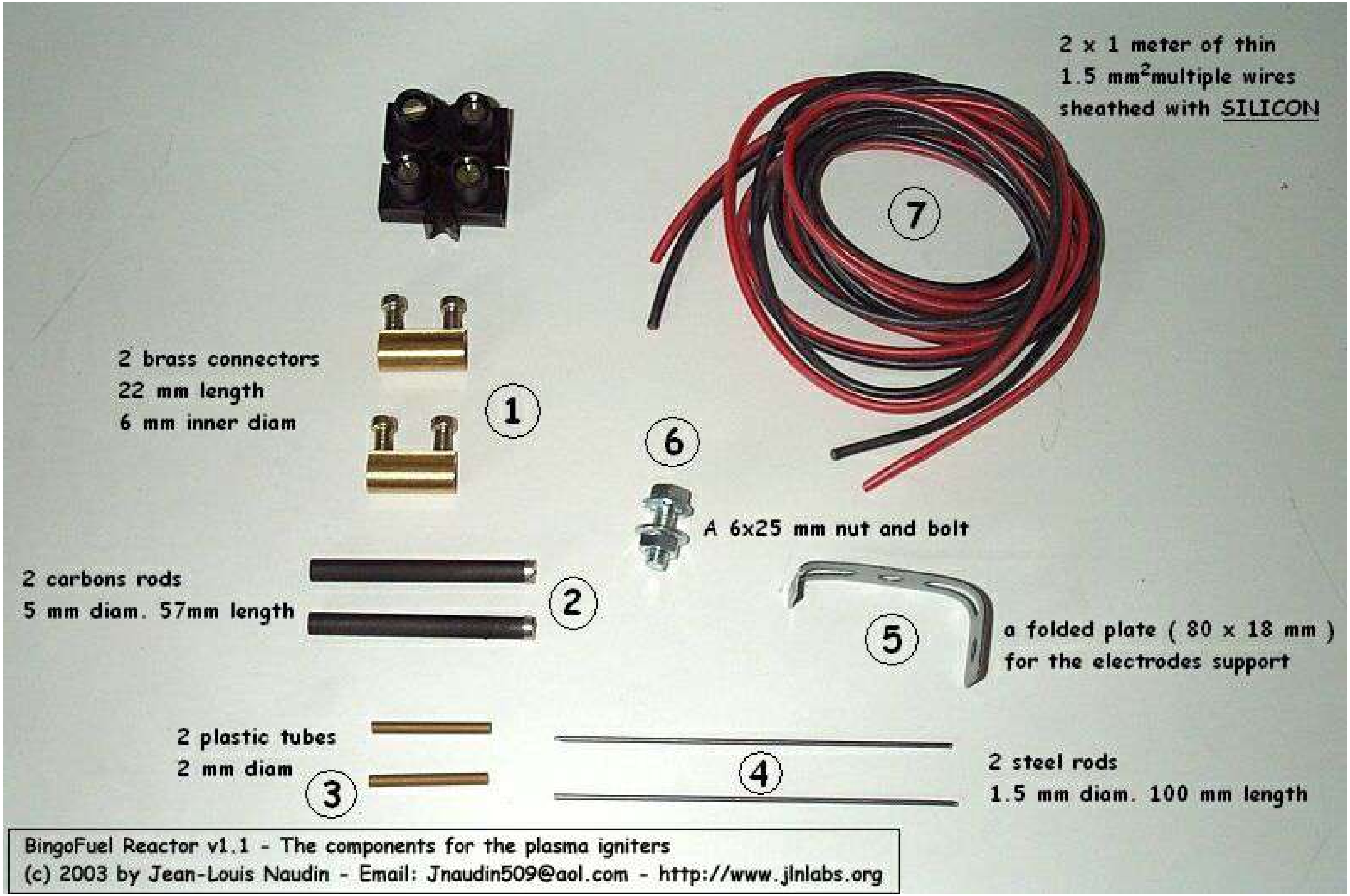
2 x 1 meter of thin
1.5 mm² multiple wires
sheathed with SILICON
7
2 brass connectors
22 mm length
6 mm inner diam
1
6
A 6x25 mm nut and bolt
2 carbons rods
5 mm diam. 57mm length
2
5
a folded plate (80 x 18 mm)
for the electrodes support
2 plastic tubes
2 mm diam
3
4
2 steel rods
1.5 mm diam. 100 mm length

BingoFuel Reactor v1.1 - The components for the plasma igniters
(c) 2003 by Jean-Louis Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

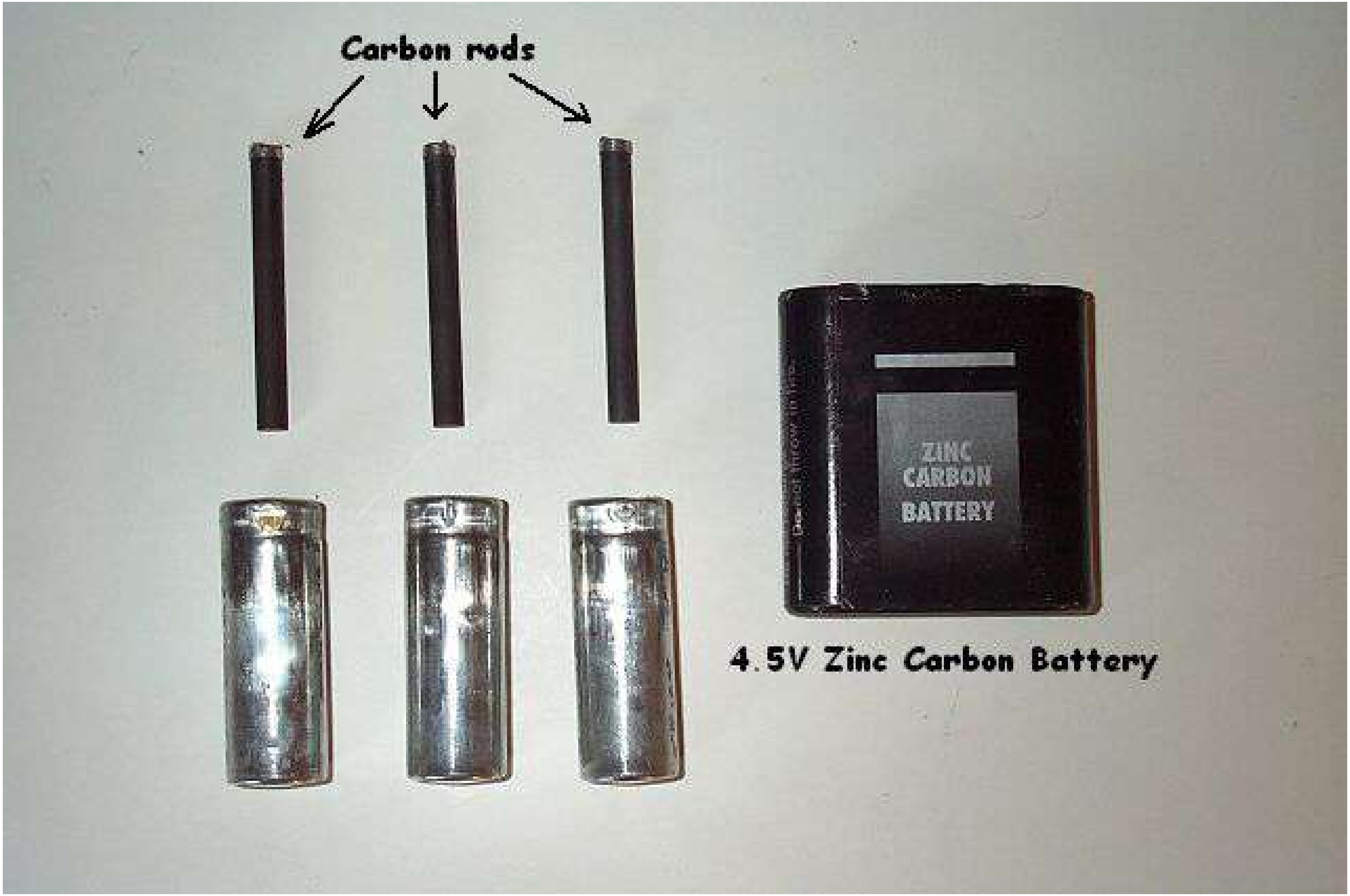
Carbon rods
ZINC CARBON BATTERY
4.5V Zinc Carbon Battery

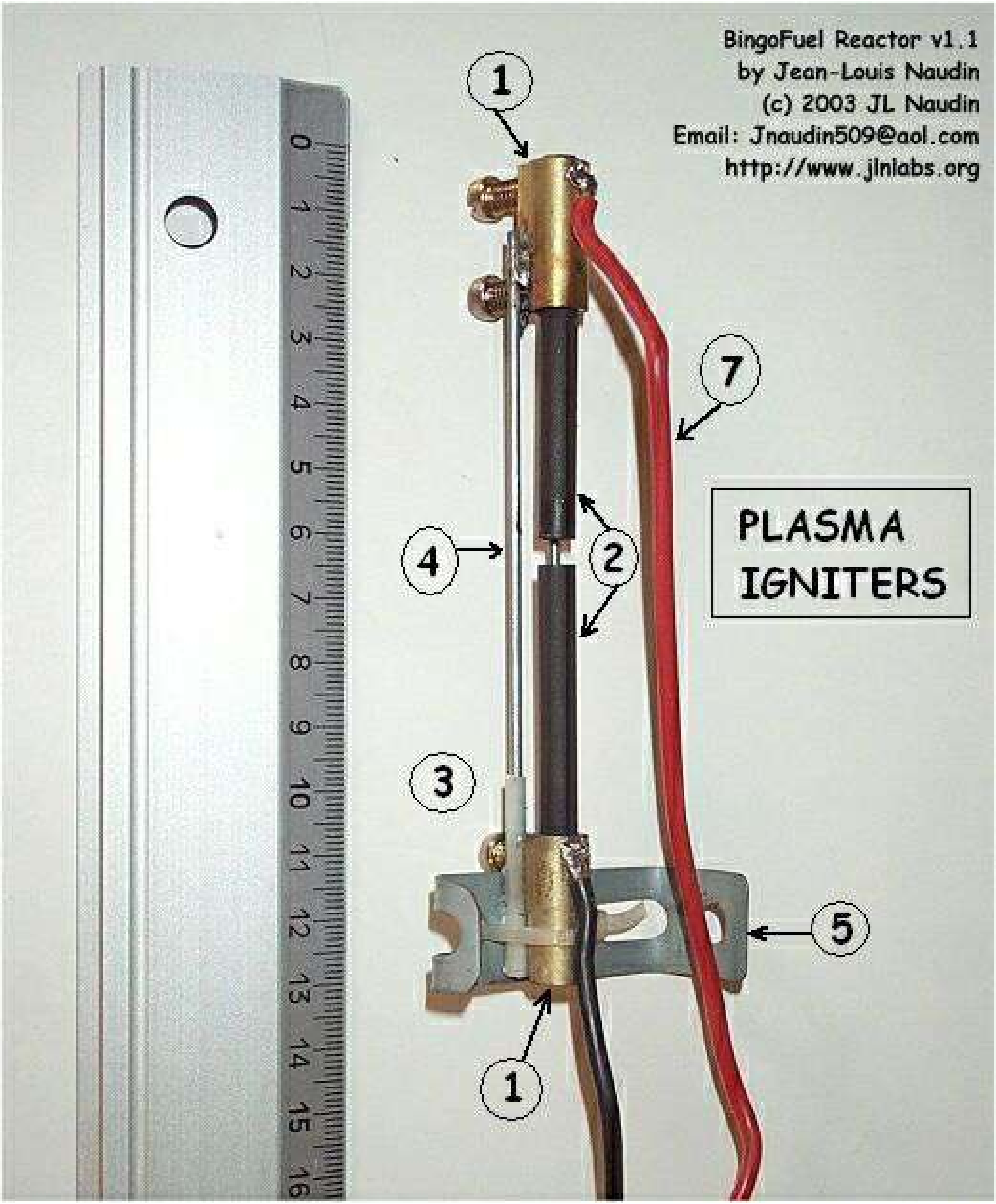
BingoFuel Reactor v1.1
by Jean-Louis Naudin
(c) 2003 JL Naudin
Email: Jnaudin509@aol.com
http://www.jnlabs.org
PLASMA
IGNITERS
1
2
3
4
5
7
1

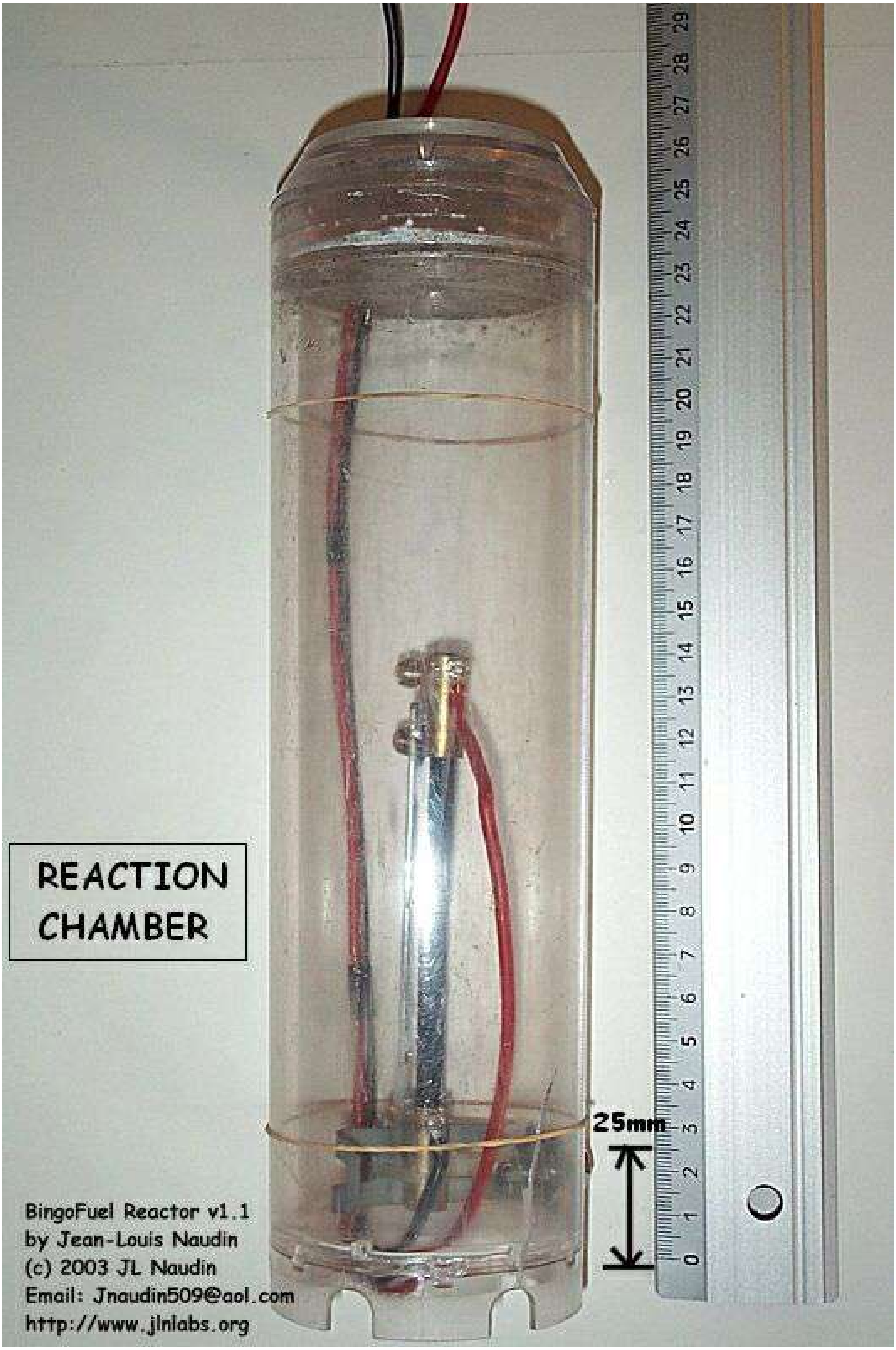
REACTION CHAMBER
25mm

BingoFuel Reactor v1.1
by Jean-Louis Naudin
(c) 2003 JL Naudin
Email: Jnaudin509@aol.com
http://www.jlnlabs.org

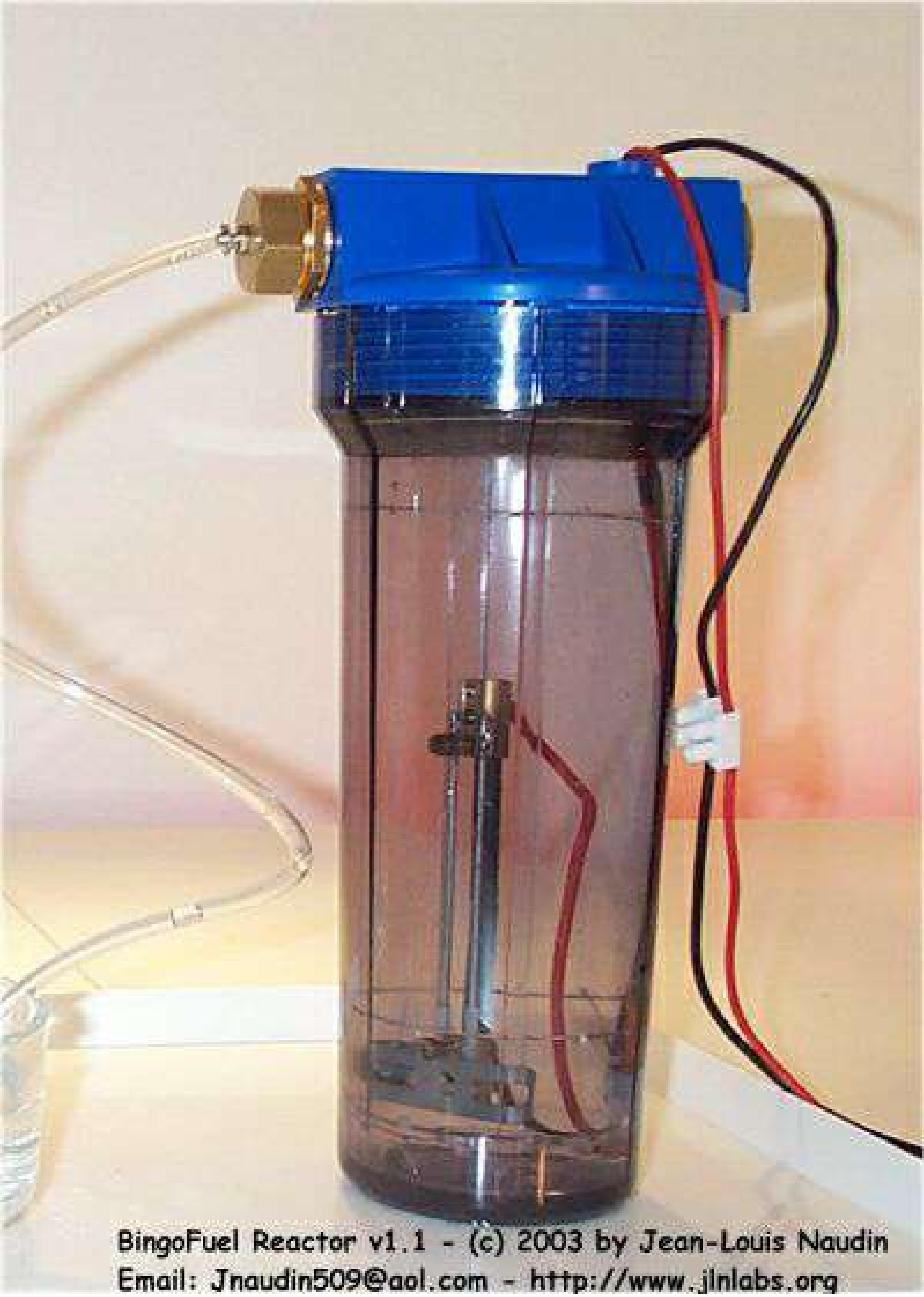

BingoFuel Reactor v1.1 - (c) 2003 by Jean-Louis Naudin
Email: Jnaudin509@aol.com - http://www.jlnlabs.org

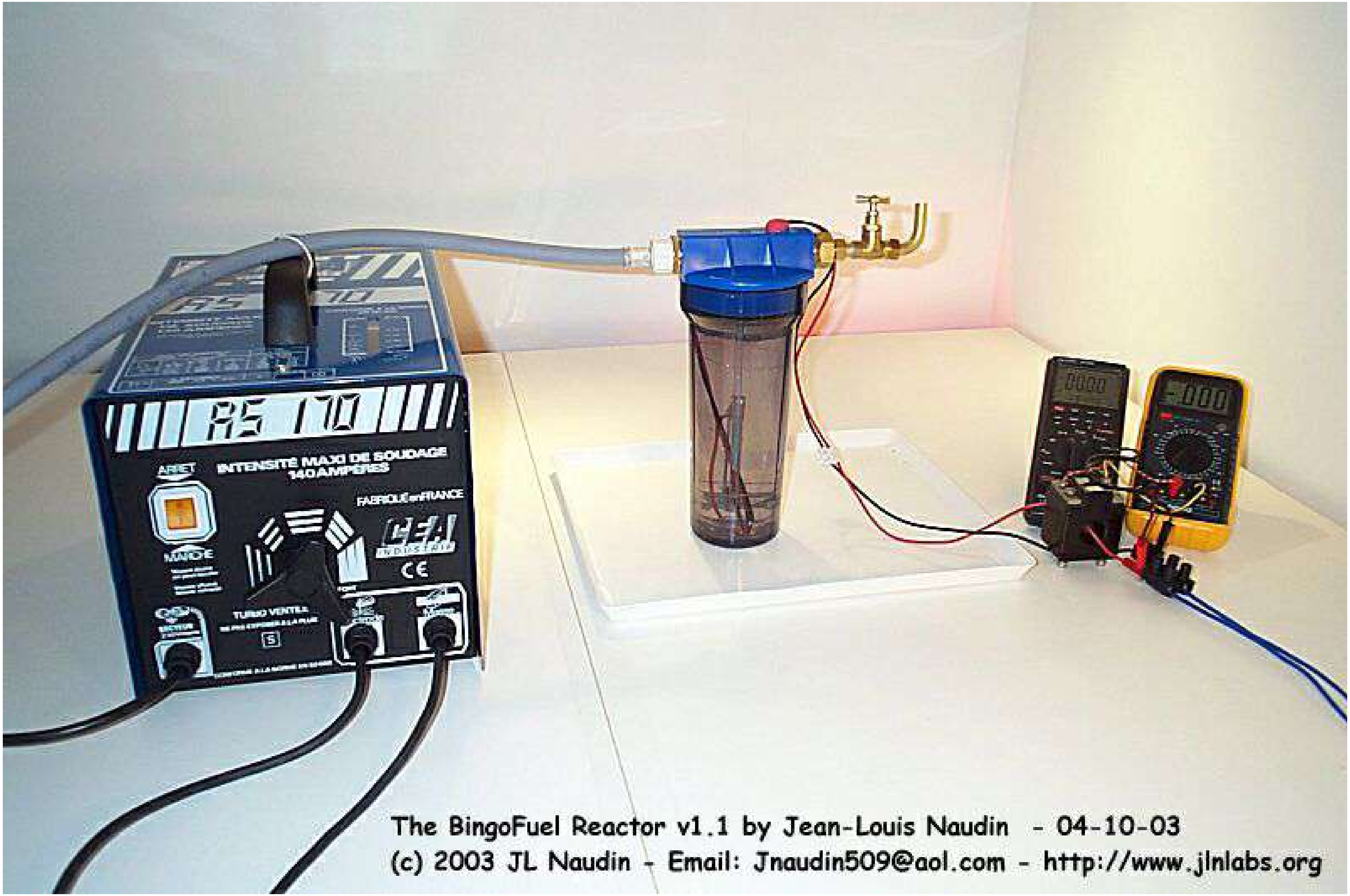
RS 170
ARRET
INTENSITE MAXI DE SOUDAGE
140 AMPERES
FABRIQUE en FRANCE
MARCHE
CEA
CE
TURBO VENTILE

The BingoFuel Reactor v1.1 by Jean-Louis Naudin - 04-10-03
(c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

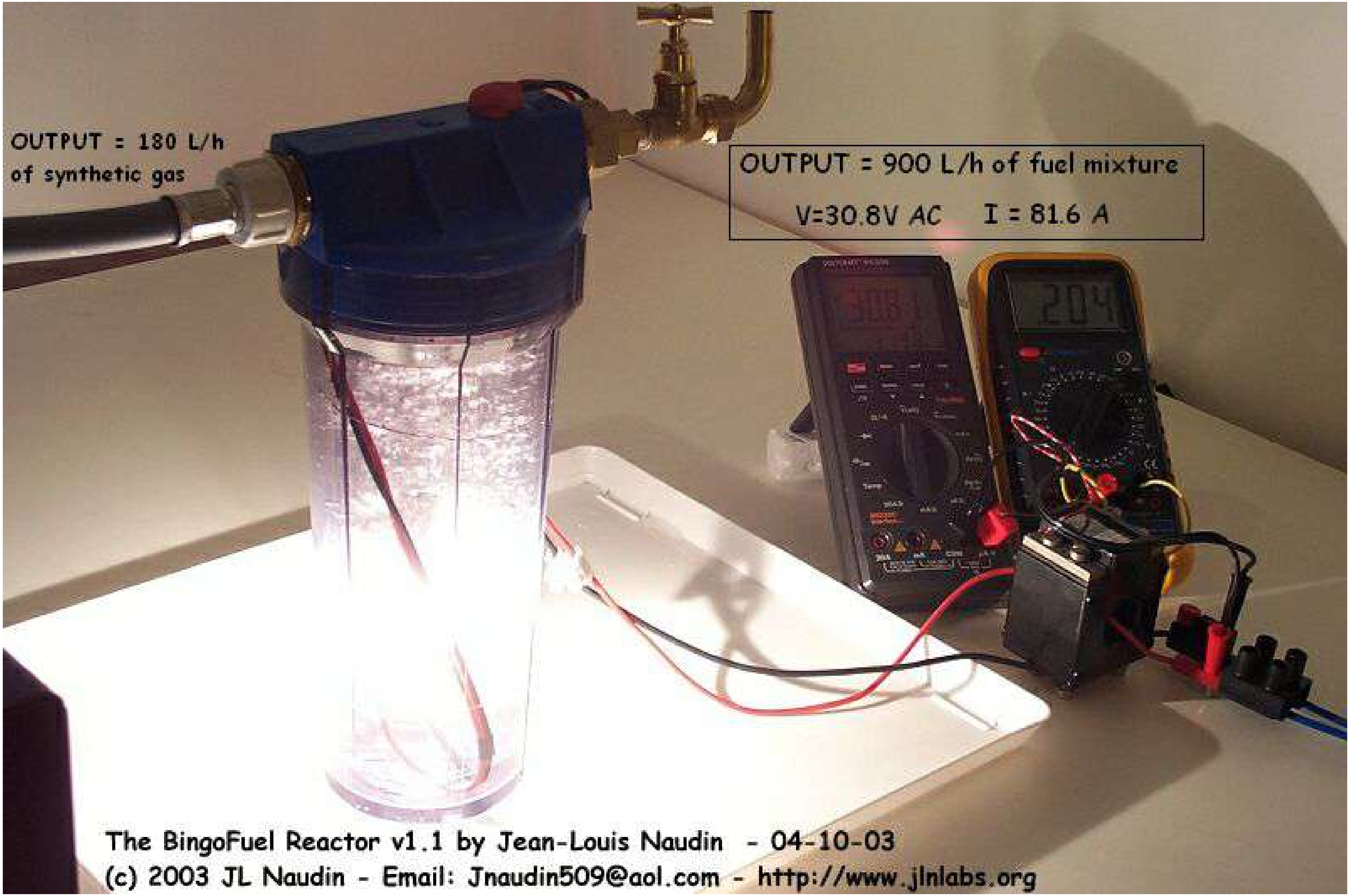
OUTPUT = 180 L/h
of synthetic gas
OUTPUT = 900 L/h of fuel mixture
V=30.8V AC I = 81.6 A
2.04

The BingoFuel Reactor v1.1 by Jean-Louis Naudin - 04-10-03
(c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

HONDA
GC160
5.0
SDMO
RANGER 2500
www.sdmo.com

GROUPES ÉLECTROGÈNES MONOPHASÉS

50 hz			Moteur							Alternateur					Options				
Type	Puissance max 230V		Marque	Type	Sécurité huile	Démarrage électrique	HP 3600 tr/mn	h Autonomie	L Réservoir	230V Disjoncteur	Niveau sonore CEE Lwa	dB(A) @ 7 m	Dimensions L x l x h cm	Kg Poids	Kit brouette	Disjoncteur	Quick'lock	Com. à distance	Coffret dém. auto
	kW ISO 8528	kVA(1) Continu																	
RANGER™ 2500	2,1	2,6	Honda OHC	GC 160	•	x	5	2,2	2	•	**98**	75	58x46x44	30	x	x	x	x	x

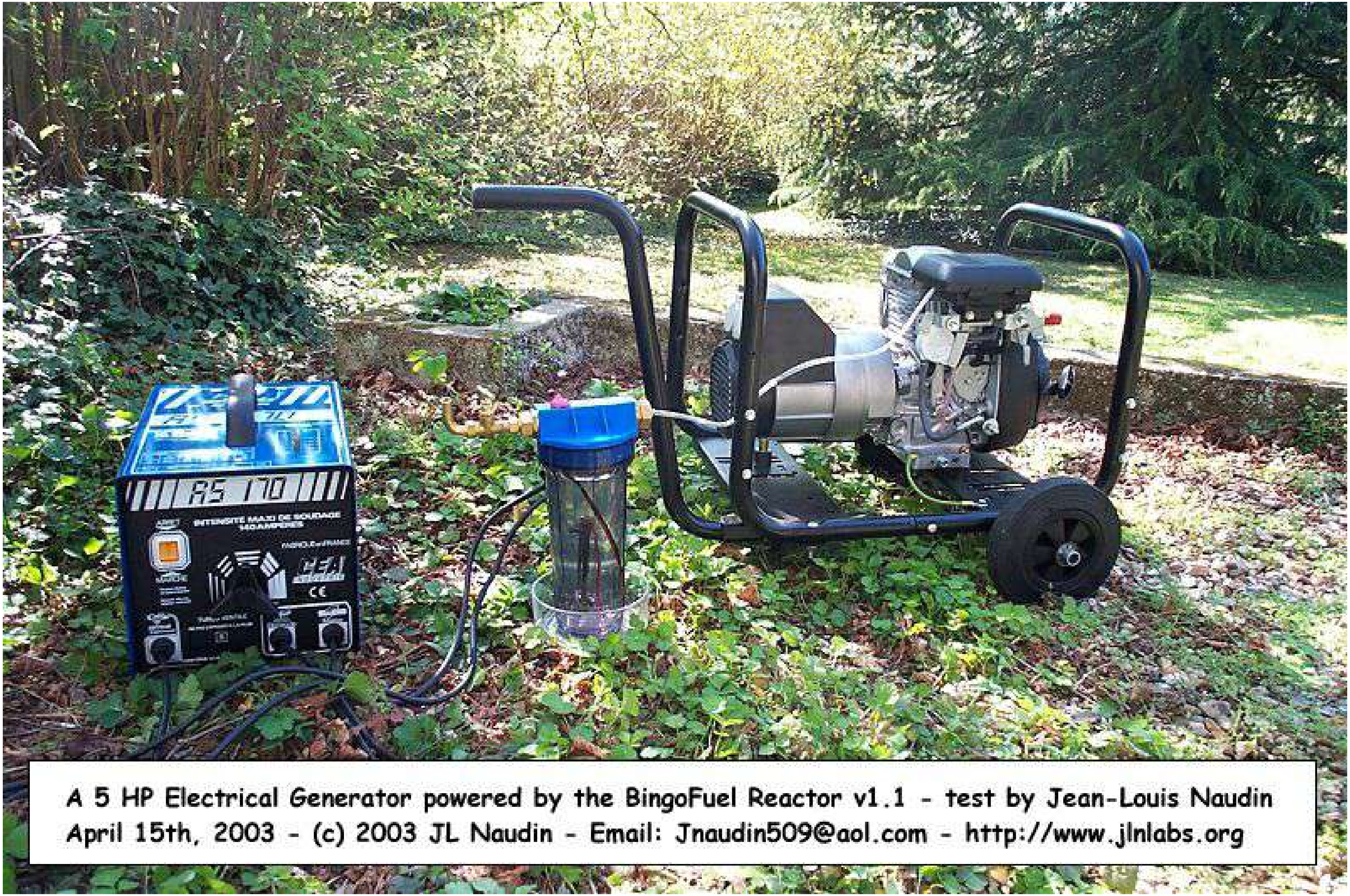
RS 170
INTENSITE MAXI DE SOUDAGE 140 AMPERES
CE

A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

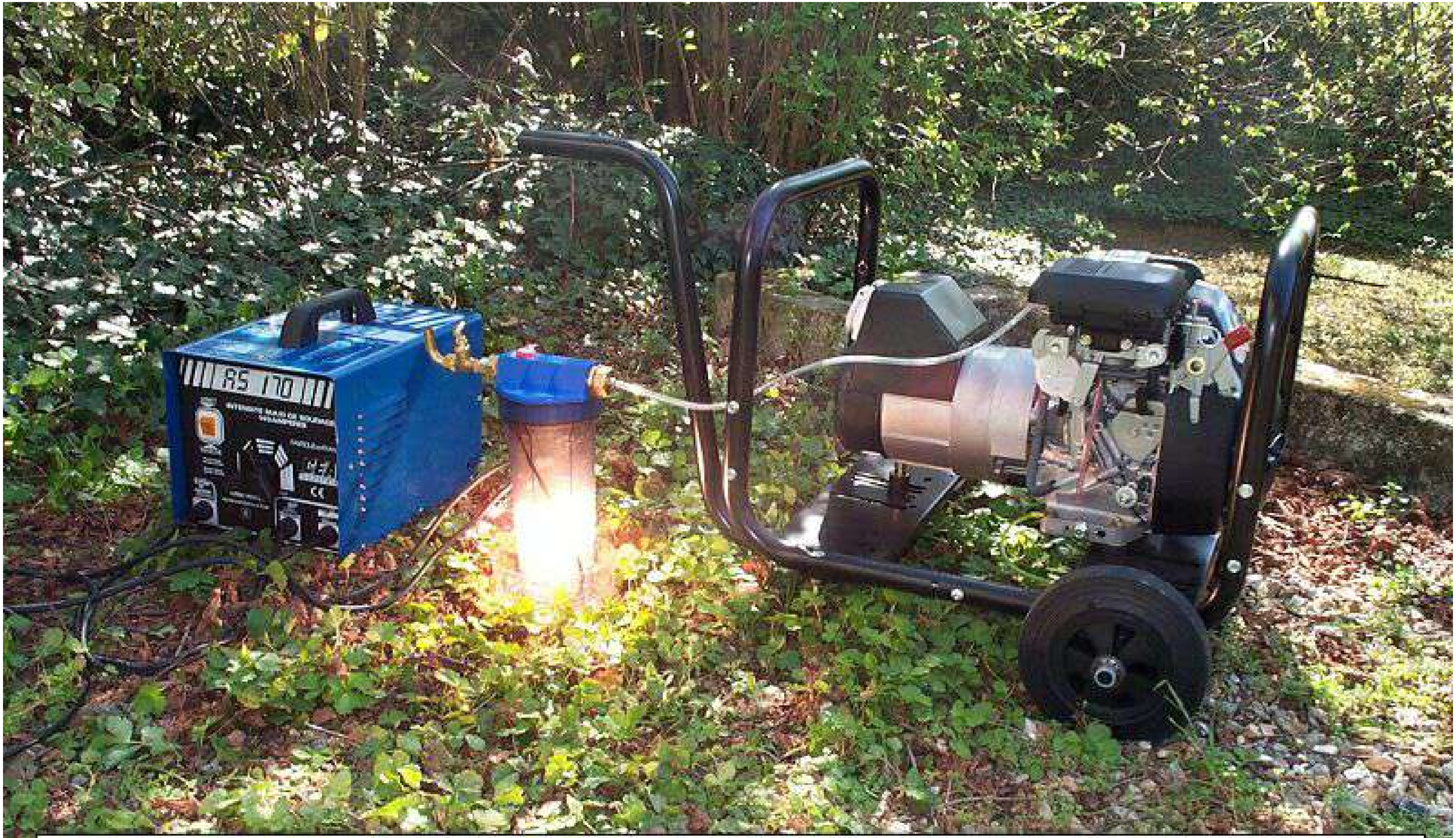

A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

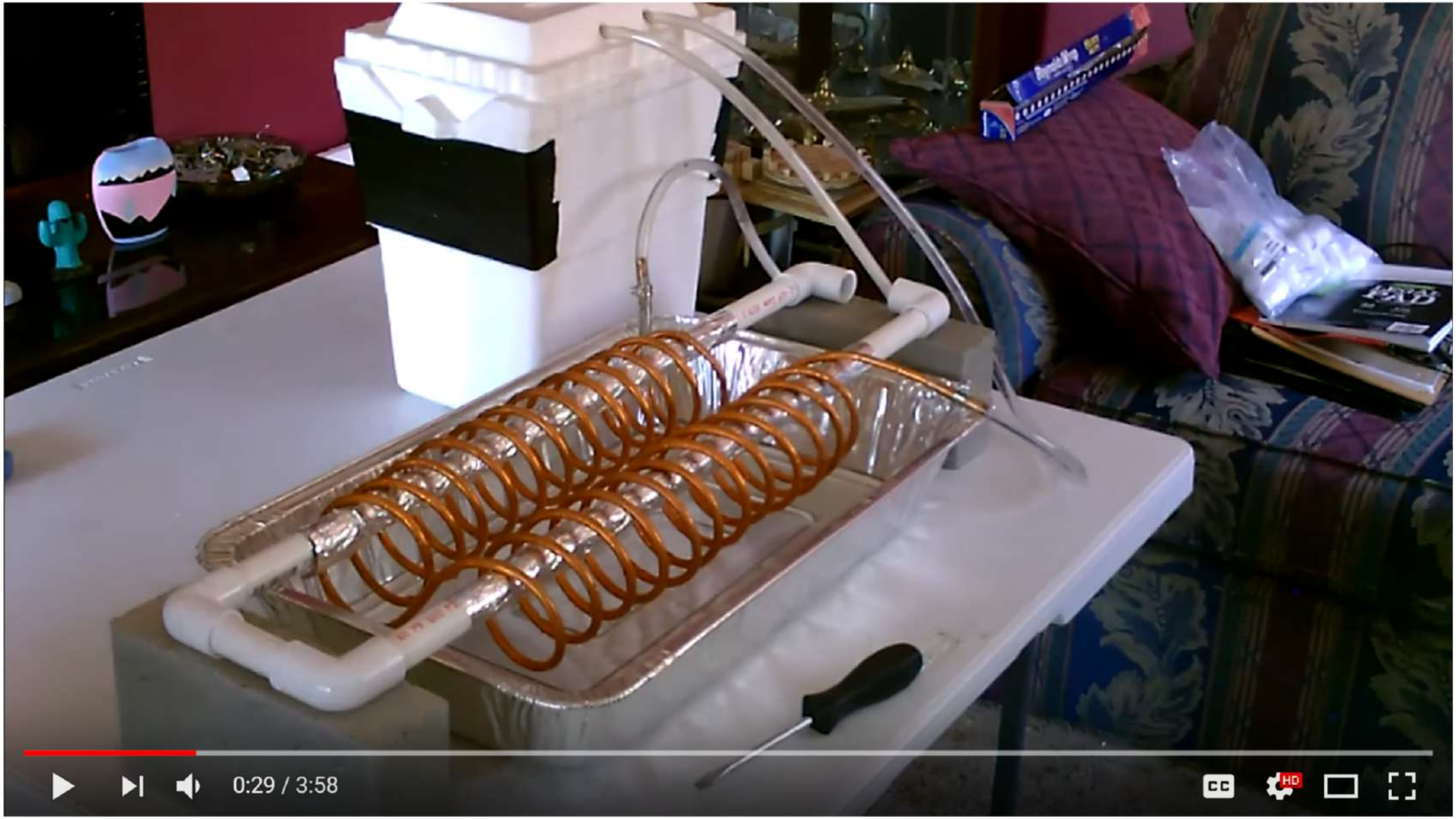

DIY Atmospheric Water Generator! - Produces/Extracts Distilled Water from the air! - DIY distiller

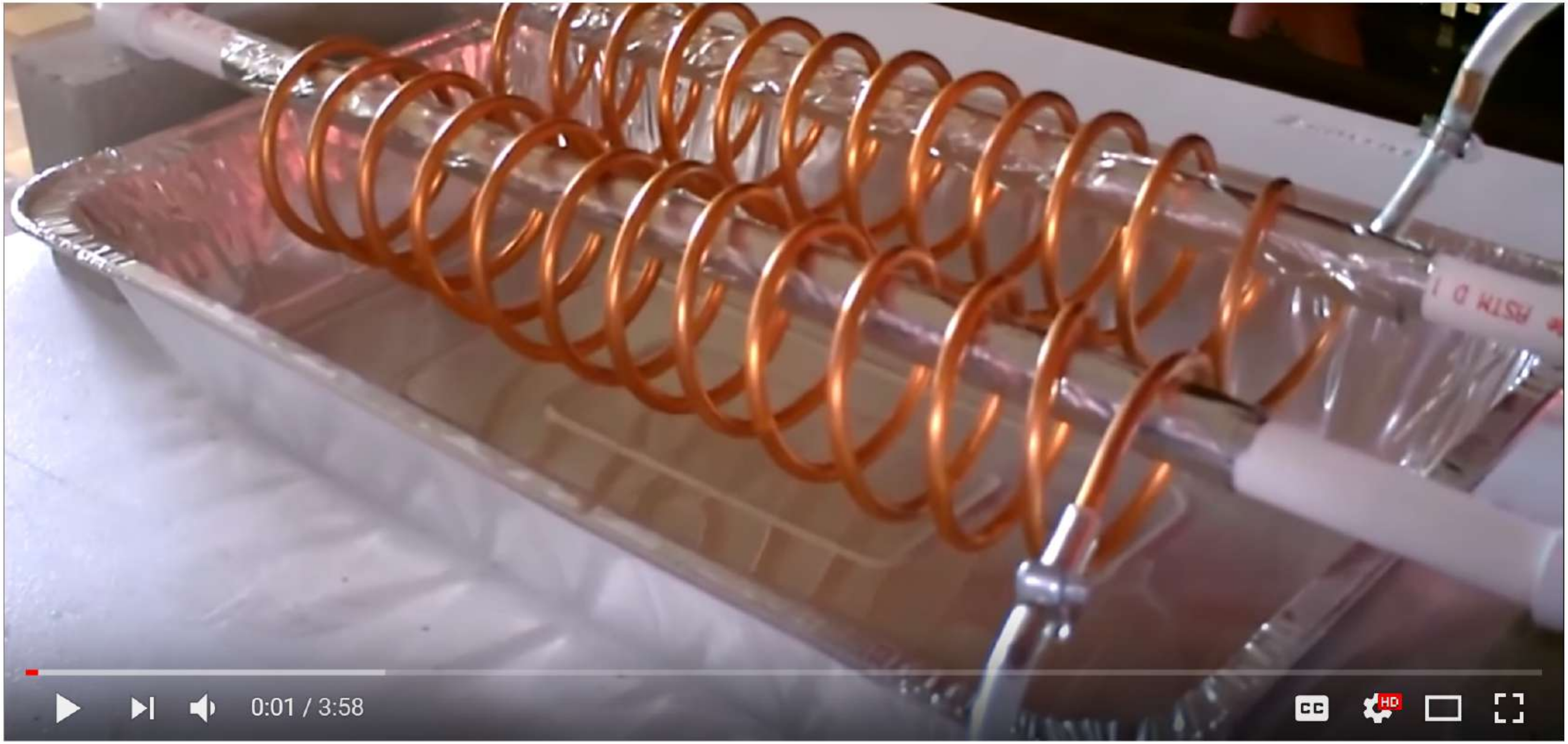
ASTM D
0:01 / 3:58
HD

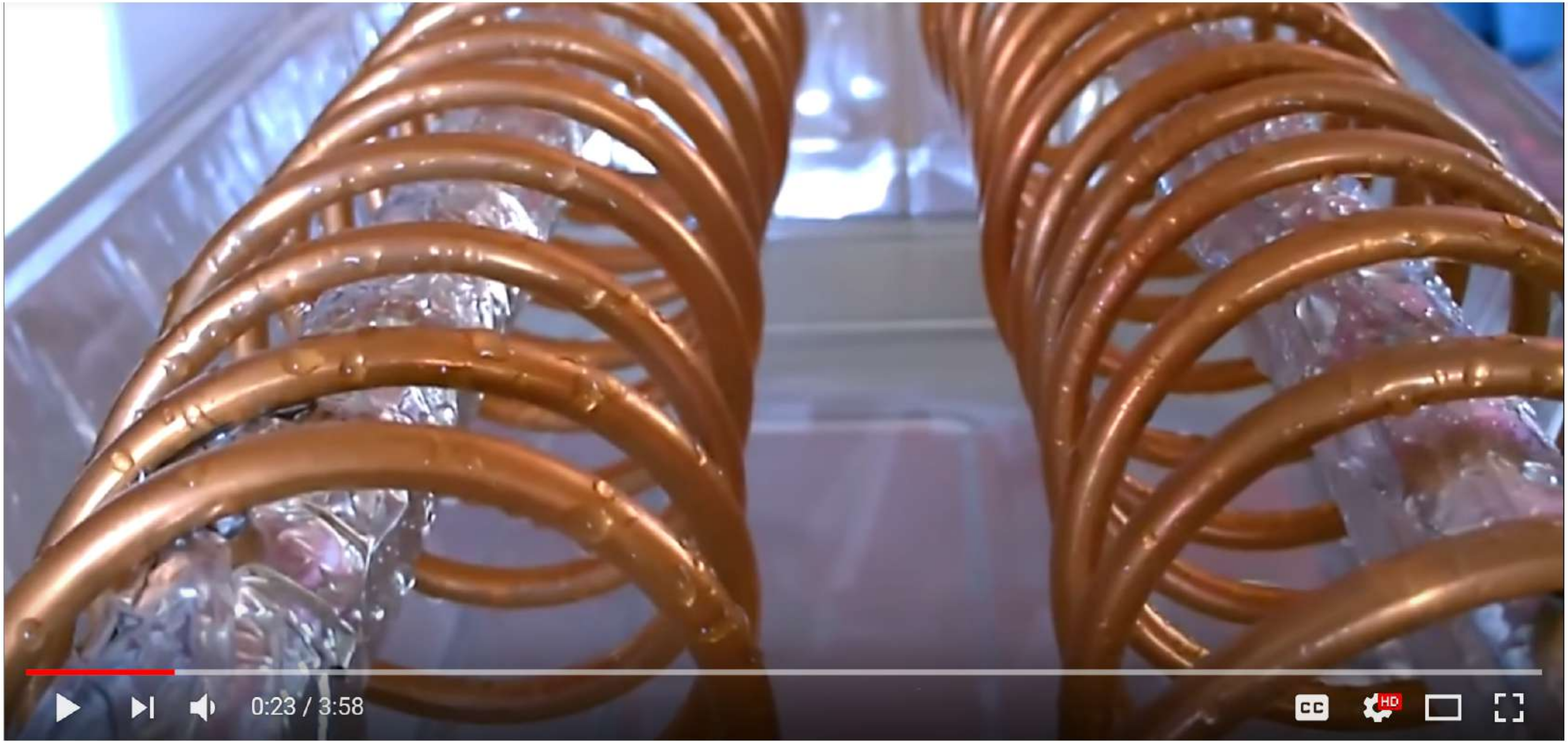
0:23 / 3:58
HD

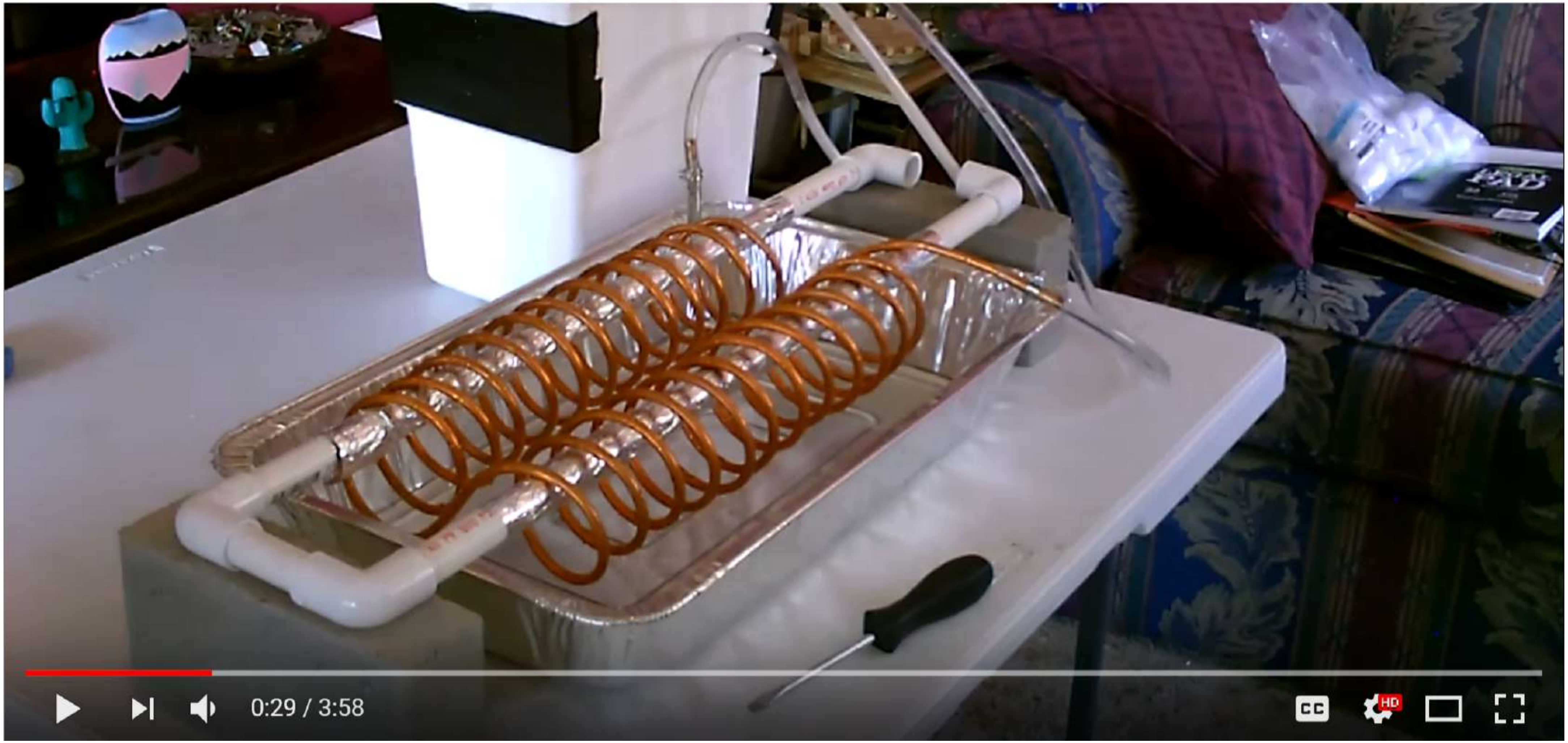
0:29 / 3:58
CC
HD

www.electronicecircuits.com

previous next

image 2 of 2 Transmitter RF Output LED Indicator Circuit

close

2N3904
NPN General Purpose Amplifier

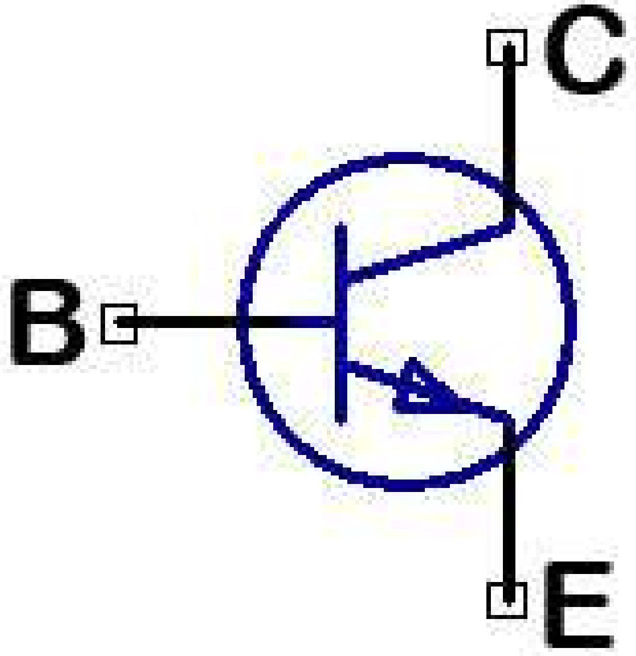

E = EMITTER
C = COLLECTOR
B = BASE

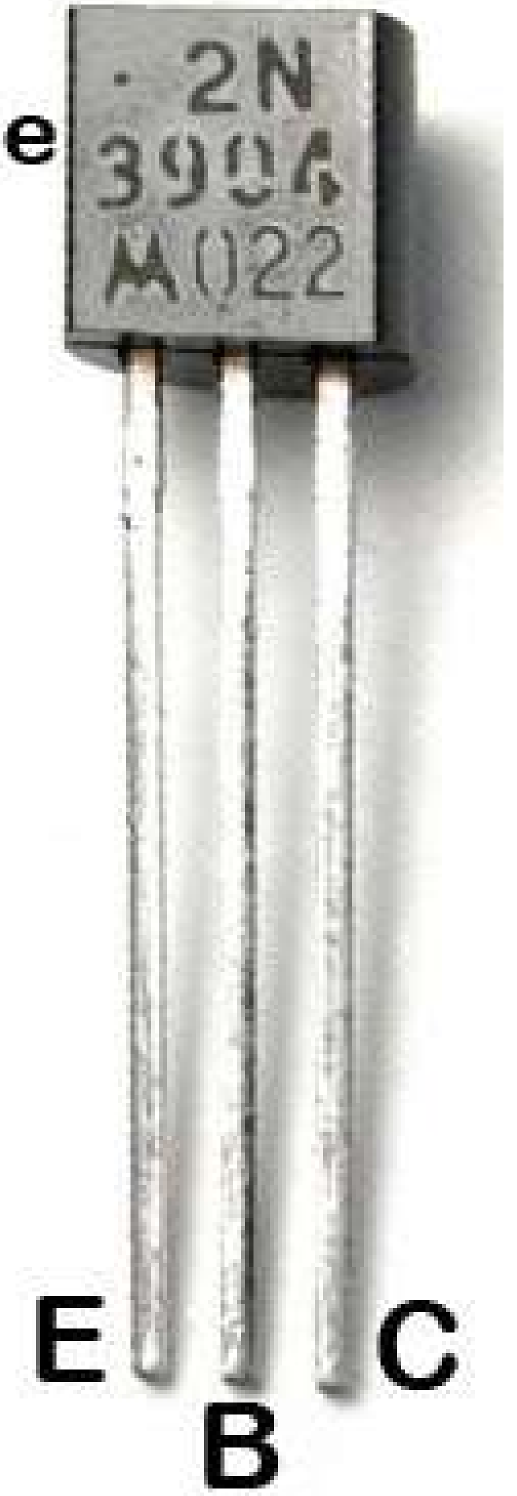
2N
3904
M022
E
B
C

www.electronicecircuits.com

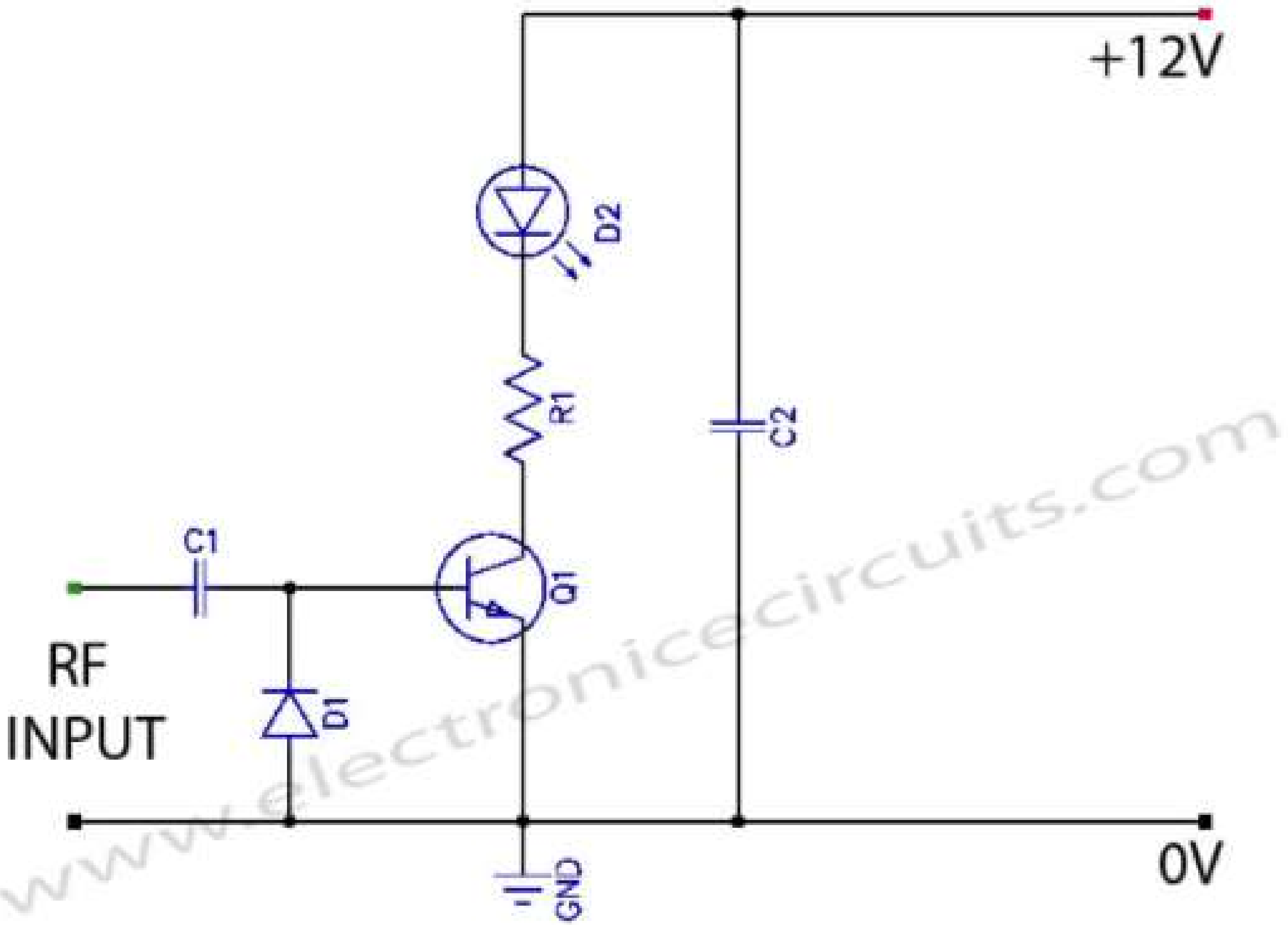
+12V
D2
R1
C2
C1
Q1
RF
INPUT
D1
GND
0V

www.electronicecircuits.com

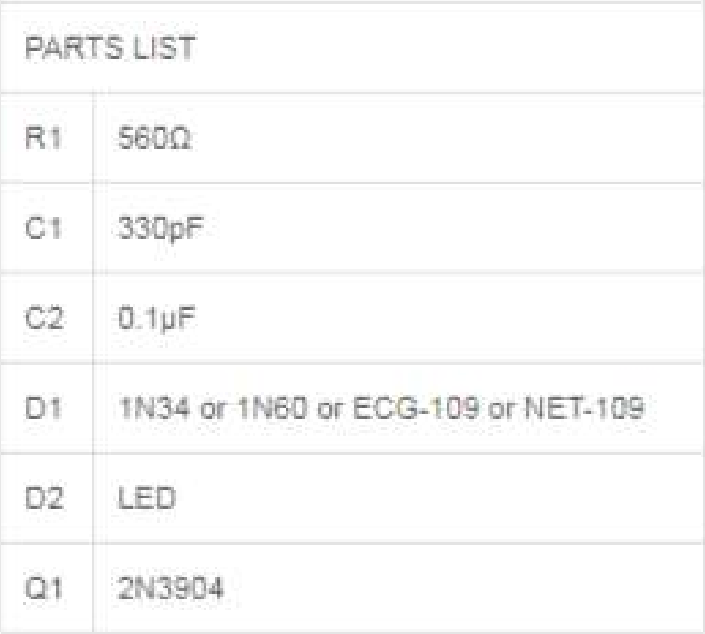

PARTS LIST	
R1	560Ω
C1	330pF
C2	0.1µF
D1	1N34 or 1N60 or ECG-109 or NET-109
D2	LED
Q1	2N3904

www.electronicecircuits.com
previous
next
close

image 1 of 2 Radio frequency LED Output Indicator Circuit

Step 1: Assembly Instructions

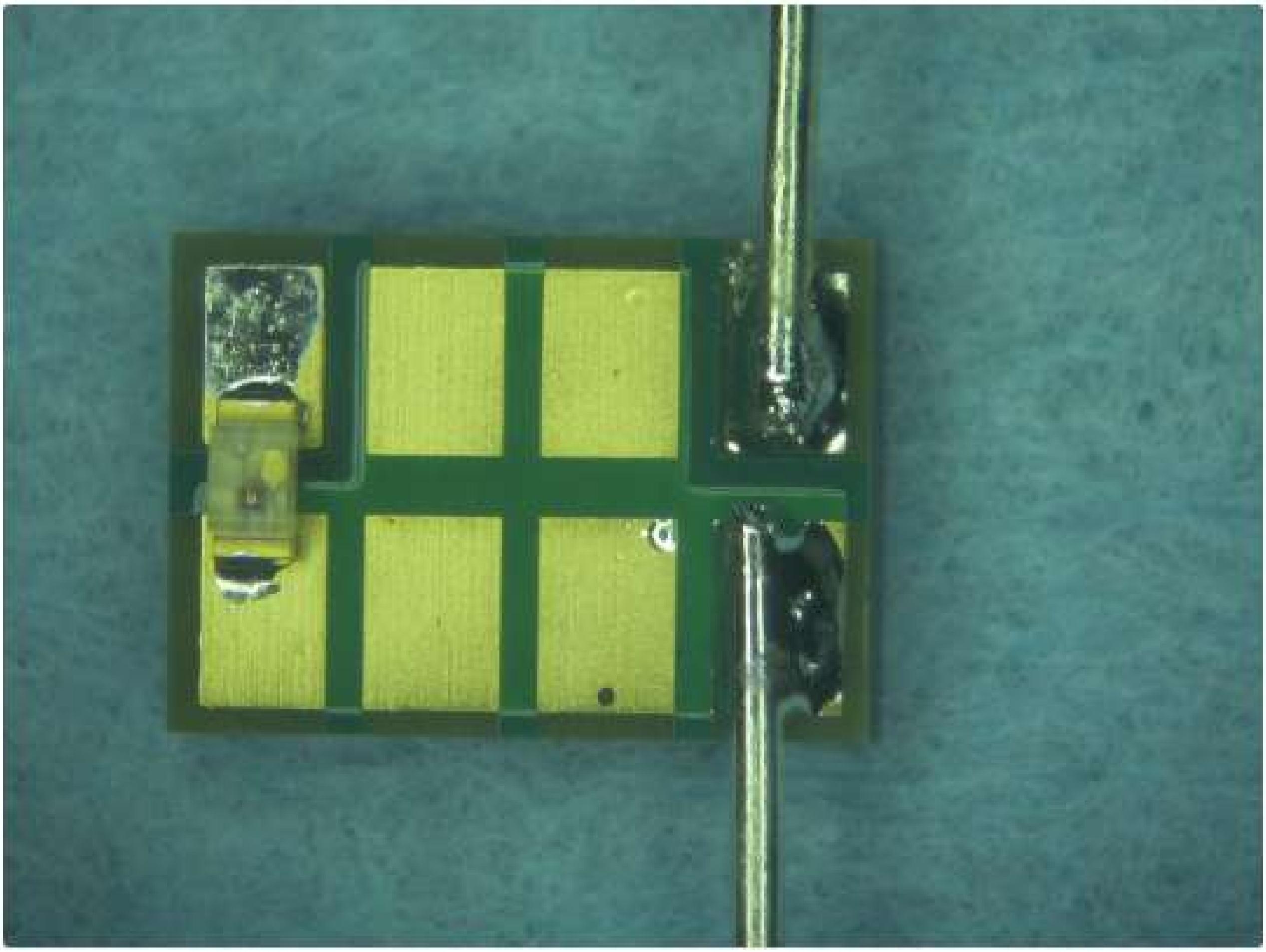

Cut the resistor wires off next to the resistor. These are just the right size at 1 1/8" long for a 2.5GHz dipole. Throw away the resistor and keep the wires.

Put solder paste on the module at pins 1 & 8 and at pins 4 and 5.Place the wires on pins 4 and 5 and solder carefully using tweezers to hold the wires (it will burn you otherwise). Solder at the lowest soldering temperature possible to avoid damaging the module. If the iron is too hot then you may damage the internal connections inside the module. Use a minimum of time for soldering (<10secs). The wires work as a dipole antenna to collect the 2.5GHz energy into the RF (Radio Frequency) Input of the module.

Place the LED with the anode (positive side) onto pin 1 and the cathode (negative side) on pin 8 and solder carefully. For those not familiar with LEDs, the triangle symbol of the diode should point to the ground pin of the module (pin 8). Your final microwave harvester should look like figure 2

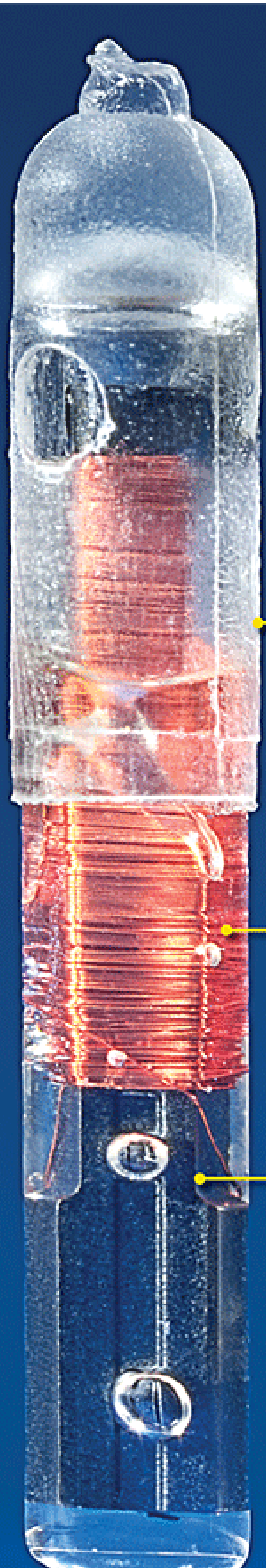

LIBERTY
IN GOD WE TRUST
200

SIZE The device is 11 millimeters long and about 1 mm in diameter, comparable to a grain of rice.

TISSUE-BONDING CAP A cap made from a special plastic covers a hermetically sealed glass capsule containing the RFID circuitry. The plastic is designed to bond with human tissue and prevent the capsule from moving around once it has been implanted.

ANTENNA The coils of the antenna turn the reader's varying magnetic field into current to power the chip. The coil is coupled to a capacitor to form a circuit that resonates at 134 kilohertz.

ID CHIP The chip modulates the amplitude of the current going through the antenna to continuously repeat a 128-bit signal. The bits are represented by a change in amplitude—low to high or high to low. An analysis by Jonathan Westhues, of Cambridge, Mass., indicated that only 32 of the bits varied between any two VeriChips. The rest of the bits probably tell the reader when the loop starts and may also contain some error-checking or correction data.

Materials:

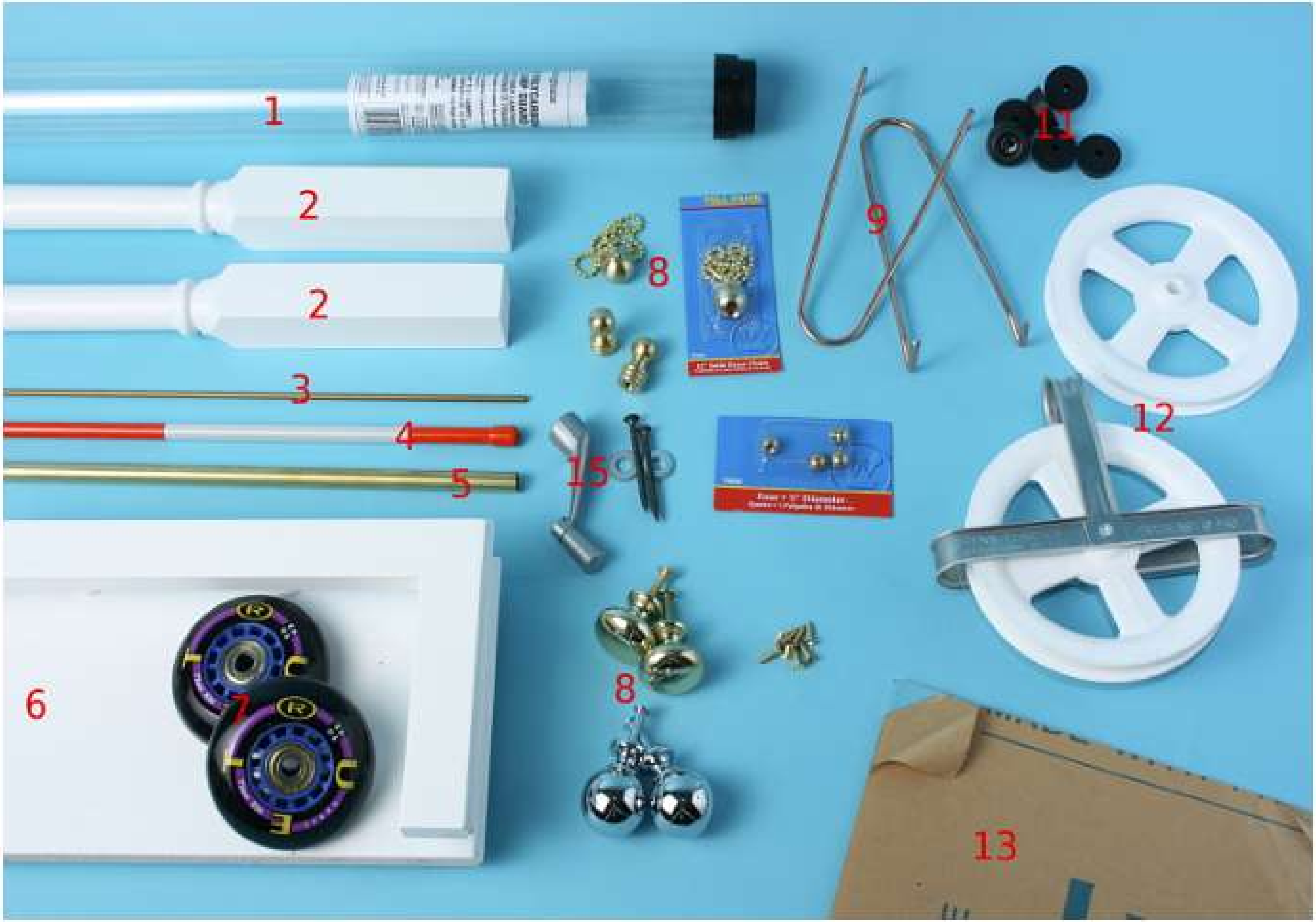
1
2
2
3
4
5
6
7
8
8
8
9
11
12
13
15

1. Fluorescent Lamp Protector Sleeve – used to make the two Leyden jars.
2. Staircase Balusters – these will be the supports for the rotating disks.
3. 1/8" Bronze Brazing Rod – will be used to fabricate all of the conductors. If you can't find this at your local hardware store look for a welding supply shop, they are sold by the pound and are incredibly useful for many things even if you don't own an oxyacetylene torch.
4. Fiberglass Driveway Marker Rod – Make sure it's round and 5/16" in diameter; these will be the shafts and insulated supports.
5. 3/8" OD Thin Wall Brass Tubing – one 3' section.
6. Knick-Knack Shelf Kit – approximately 24" by 6". You can use any ¾" board you desire, the shelf included has a nice rail that will add to the overall look of the project.
7. Inline Skate Replacement Wheels – Quantity 2.
8. Lamp Parts – You will need a selection of lamp parts which may vary depending on what is available at your particular store. Pictured here are pull chains, finials, and ball nuts used to make parts of the charge collector combs and discharge electrodes. Also pictured are cabinet knobs which were not used in this project but would make good alternatives. See the charge collector construction step for details.
9. 1" Copper Pipe Hangers – These you'll find in the plumbing section, they are copper plated steel.
10. Solder wick (not pictured) – for the neutralizing brushes, you might have to visit Radio Shack for this.
11. Rubber feet – Quantity 6.
12. Clothes Line Pulleys – must be plastic.
13. 3/16" Acrylic Glazing – enough to cut (2) 14" circles from. Polycarbonate will work too and is easier to work with but costs more than twice as much.
14. Aluminum tape (not pictured) – found with the duct tape and HVAC supplies, get the kind with the peel off paper backing.
15. Rubber O-ring belts (not pictured) – available from McMaster-Carr, part number: 9411SK259 about $15 for a package of eight.

The total cost of purchasing the materials new is about $100. However, these are all relatively common items

Disks and Drive Components:

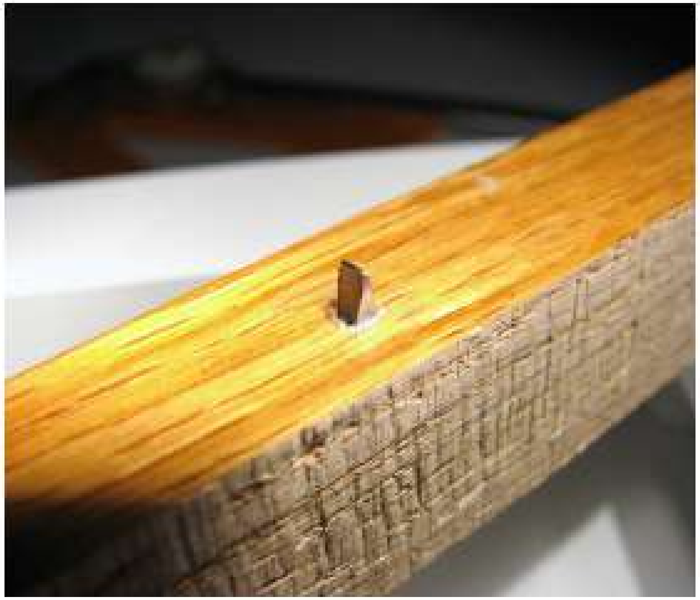

wimshurst-circle-cutter-inset.jpg

Make the cutting tool:

1. To cut the two 14" acrylic circles we will first need to make a tool. Cut a 12" length of wood ¾" square. Pine will work but hardwood is preferable.
2. Drill a pilot hole near one end and press or drive a #6 penny nail through the stick so the point sticks out about ¼".
3. Drill a second hole exactly 7" from the first and insert another #6 penny nail into it.
4. Use a fine metalworking file to shape the point of the second nail as shown. You want to make a chisel point with a slight undercut on the leading face.

Cut the acrylic disks:

1. Lay out your circles with a compass to be sure they will both fit on your sheet of acrylic.
2. Drill a 1/8″ hole in the center of your circle. Be gentle when drilling acrylic, it cracks easily. Polycarbonate is quite a bit tougher.
3. Working on a carpeted floor, insert the unmodified nail in the center and begin scoring your circle. Cut about a quarter of the way with each stroke and work your way around the circumference.
4. If the cutter sticks, lift it out and move to a different spot.
5. When you think you've gone about halfway through, flip the acrylic sheet over and cut from the other side. You may end up flipping the sheet several times before the circle pops free.
6. Clean up the edge of the circle with some 400 grit sand paper and set them aside.

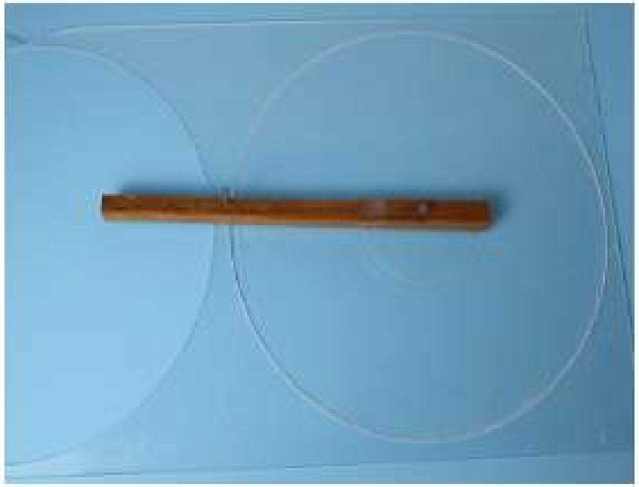

wimshurst-circle-cutter.jpg

Cut belt grooves in the skate wheels:

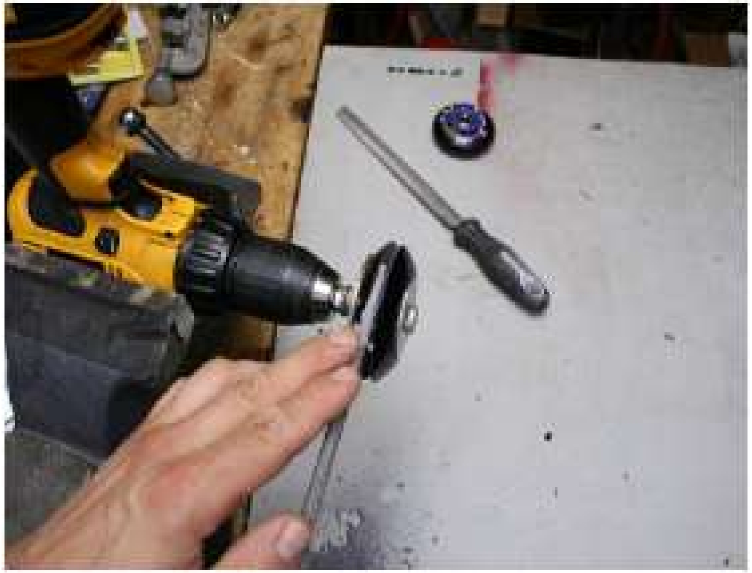

wimshurst-skate-wheel-groove.jpg

1. Gently clamp or strap your drill to a workbench as pictured.
2. Assemble a mandrel from a 5/16″ bolt and some large (fender) washers, when assembled the entire wheel must spin, not just the bearings.
3. Chuck the assembly into the drill. The wheel should turn toward you and the speed should be fairly fast.
4. With a crosscut bastard file make a ¼" wide flat on the wheel and then switch to a rat-tail file to cut the grove. Apply light and even pressure to the file.

Attach the skate wheels to the disks:

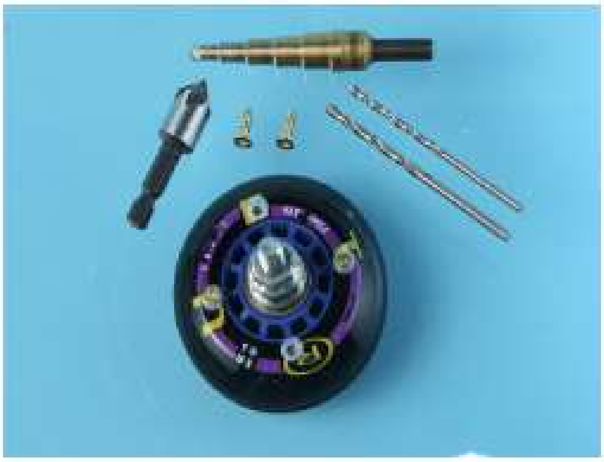

wimshurst-skate-wheel-mount.jpg

1. Use a step drill bit like the one pictured to increase the size of the hole in the acrylic disk to 5/16". Remember, be gentle and go slowly because acrylic is easily cracked.
2. Remove the washers from the wheel and use the 5/16" bolt to center the wheel against the disk.
3. Drill (4) 1/8" holes through the disk, don't drill into the wheel.
4. Switch to a 3/32" bit and drill partway into the wheel in 4 places.
5. Finish the holes with a counter sink.
6. Now remove the 5/16" bolt and drill the center hole out to ½" or 5/8" using a step drill, you want the edges of the hole completely clear of the rotating parts of the wheel bearing.
7. Install (4) small counter sunken screws, tighten these so they *just* touch the disk, the disk must remain as flat as possible.

Cut the sectors:

wimshurst-sector-cutting.jpg

1. Decide how many sectors you are willing to cut. I'm rather lazy and opted for fewer sectors, 16 per disk. If you decide to make 24 or even 32 sectors you'll have to make them smaller but you will be rewarded with longer sparks.
2. The sectors are cut from aluminum tape. Make a template from a piece of plastic milk jug and trace each sector. Cut them individually, don't be tempted to stack multiple layers of tape; the cut will end up ragged and will bleed charge away into the air.
3. Tip: I found it easiest to use an X-acto knife and straight edge to cut the long sides and then switch to scissors for the curved ends.

Attach the sectors:

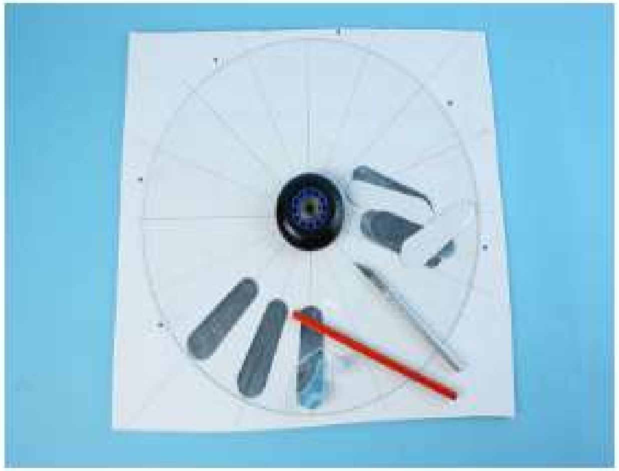

wimshurst-affix-sectors.jpg

1. Lay out a circle on a piece of foam board.
2. Draw radial lines to correspond with the number of sectors you've chosen
3. Place your template centered at 6 o'clock and trace it. The large end should face out and be about ¼" from the edge of the disk.
4. Set the disk on the foam board and insert push pins around the circumference so it turns in place.
5. Carefully peel and stick the sector in place. It's a good idea to make some extra sectors and practice this operation first. A length of fiberglass rod makes an excellent burnishing tool.
6. Turn the disk one line to the left and repeat. Always index the line to the first sector you stuck down, this will help make the spacing as even as possible.

wimshurst-drill-pulley.jpg

Prepare the drive pulleys:

1. Remove the pulleys from their cages by drilling out the rivets.
2. Use the step drill to enlarge the holes to 5/16". Drill from one side, then the other to enlarge the full depth of the hole in the pulley. Note: The use of the step drill is especially important here because of its self-centering characteristics.
3. Cut (2) 7" lengths of fiberglass rod, slightly bevel the ends with a file to prevent chip out. Be careful of the glass fibers, they can be really irritating!
4. Drill the splines out of the window crack bore with a regular 5/16" drill bit. Clamp the crank in a vise and go slowly; making sure the bit is in line with the axis.

Cut and drill the supports:

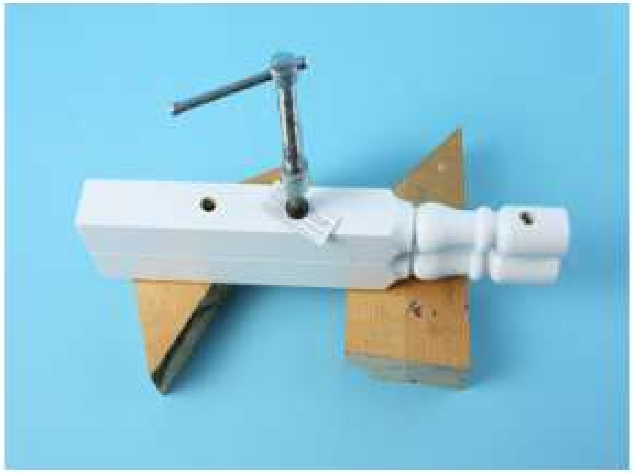

wimshurst-drill-upright.jpg

1. Cut 12" off of each of the staircase balusters. Choose the end that you think looks best. On my prototype machine I used both ends of the same baluster and thus had two different style supports.

2. Clamp the two supports together as shown and drill 5/16" holes 3 ¼" inches from the bottom (square end) and 11" inches from the bottom.

3. The lower hole will need to be reamed out so that the fiberglass axle turns freely in it. Use a slightly larger drill or rat-tail file for that. You can also drill it larger and insert plastic bushings for smoother operation. Alternatively you can bore it out with a step drill to match the diameter of a pair of skate bearings – this works exceptionally smoothly and is what I ultimately did to my own machine.

Attach the supports to the base:

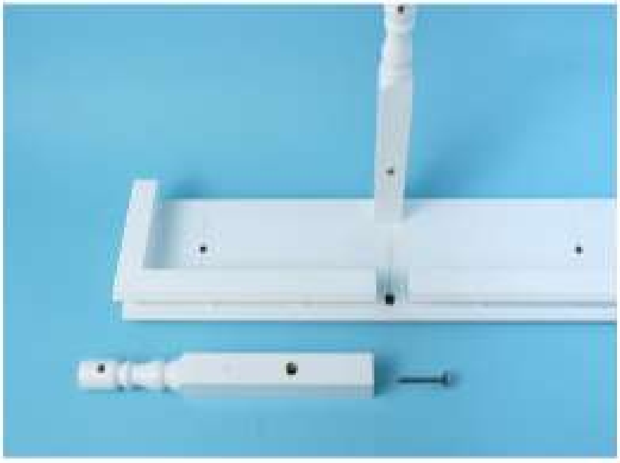

wimshurst-screw-uprights.jpg

1. Draw a line parallel to the back of the base 2 ½" in, this is not quite to the center. Draw a second line perpendicular to the first on the center of the base.
2. Cut a 1 ¼" gap in the rail on the center line, as pictured.
3. Drill (2) 3/8" inch holes through the base on the center line 5/8" from the front and back edges.
4. Use 2" drywall screws and large washers to attach the supports to the base. The combination of the large washer and 3/8" hole will allow you to adjust and align the position of the rotating disk precisely.
5. Drill (2) 5/16" holes on the line parallel to the long dimension and 7 5/8" from the centerline on each side – these holes need to be straight up and down so drill carefully, use a small carpenter's square to line up the drill.

Charge Combs and Neutralizing Bars:

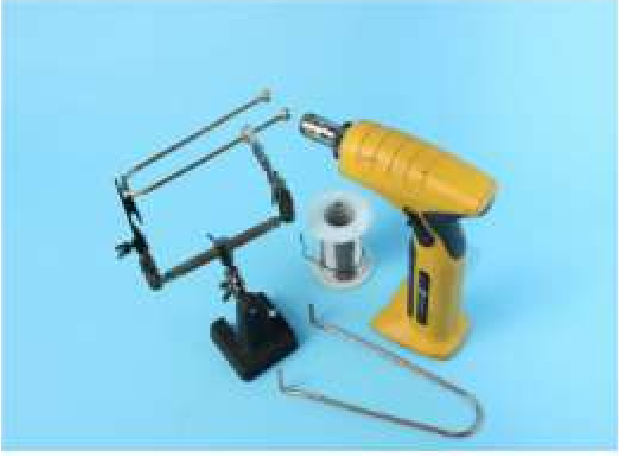

wimshurst-solder-balls.jpg

Prepare the charge collectors:

1. Use a hacksaw to cut off the nail ends of the pipe hanger. The overall length should be 5".
2. You'll find small brass ball cap nuts in the electrical section at the hardware store; they are most commonly used to secure the top of brass outdoor lighting fixtures.
3. Place the small brass ball nuts on the ends of the hanger, heat them with a small torch and apply just enough solder to fill the joint. Note: Be careful not to overheat the pipe hanger, it is copper plated steel and it you heat it too much the solder may not adhere.

The torch pictured is a Lenk LSP-180 butane torch/soldering iron and it is a marvelous tool.

Attach the collector comb prongs:

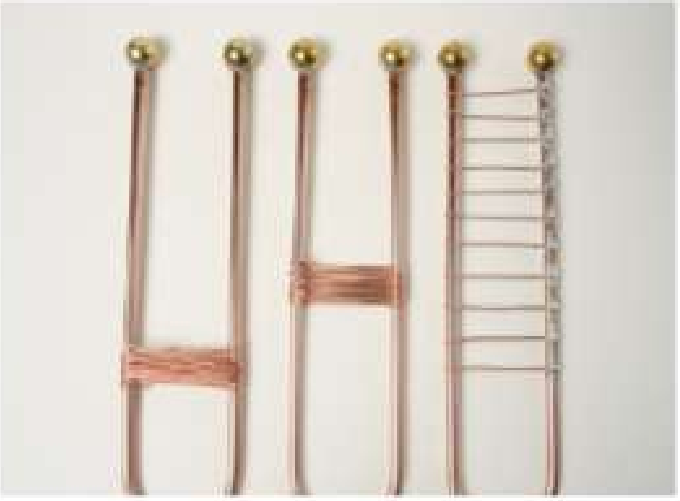

wimshurst-prongs-progression2.jpg

1. You need to make 8–12 pointy prongs down each side of the collector comb. I stripped the conductors out of a 3' section of telephone wire to make them.

2. Wrap the copper wire around the pipe hanger as shown in the left-most example. I made 11 turns.

3. Cut away the center portion of the wire on one side only and bend the cut ends around the pipe hangers.

4. Spread the prongs out evenly along the portion of the charge collector that will be opposite the sector.

Solder the prongs:

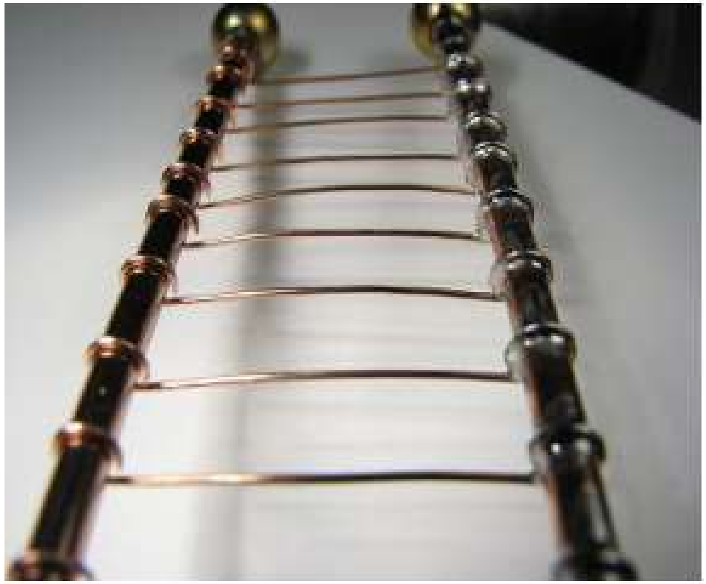

wimshurst-solder-prongs.jpg

1. Crimp the ends tightly around the pipe hanger.
2. Use a large soldering iron to solder each joint. Apply sufficient solder so that when you take the soldering iron away solder flows down to fill the gap at the end of each length of wire. We want to avoid any points other then the prongs themselves.
3. Once you've soldered all of the joints cut down the center of the wires but don't trim them to length until it's time to install the combs.

Charge collector mount:

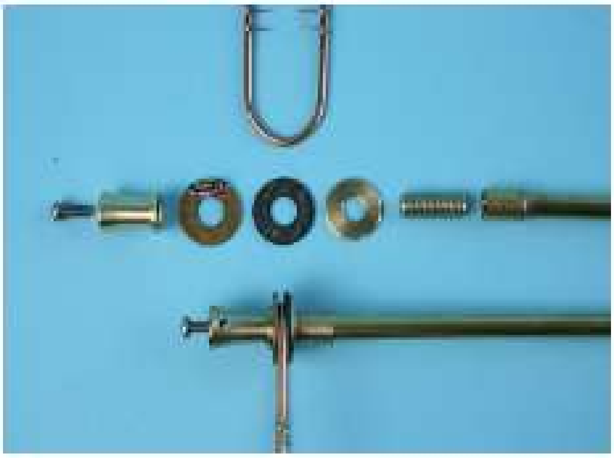

wimshurst-collector-assembly.jpg

I made a couple of different collector mounts using various lamp parts and cabinet knobs. This was the simplest, but you may have to improvise if you can't find these particular lamp parts at your local hardware store.

Pictured here right to left:

- 3/8" OD thin wall brass tubing 6" long
- 3/8" threaded collar
- 3/8" lamp "nipple" 1" long
- Lamp washer nut (threaded)
- Rubber flat washer
- 3/8" brass washer
- 3/8" threaded lamp finial
- #8-32 screw

Prepare the collector mount:

1. Using the step drill, bore out one half of the threaded collar.
2. Screw the nipple halfway into the collar and insert the brass tubing into the opposite end and solder it in place.
3. Drill one hole straight down into the top of the finial and thread with a #6-32 tap. Use the drill size written on the tap.
4. Drill a 1/8" hole through the body of the finial as pictured, this is for the discharge electrode.
5. Cut a ½" length from the extra you trimmed off of the pipe hanger earlier and solder it to the brass washer, this will allow the assembly to clamp and hold the charge collector perpendicular to the support.
6. Test assemble the mount and then disassemble and set aside.

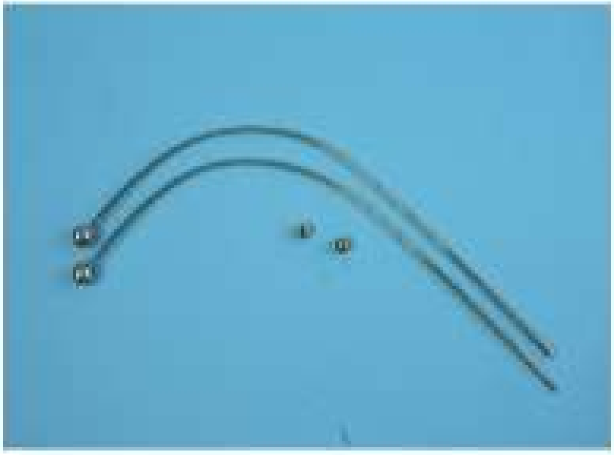

wimshurst-discharge-assembly.jpg

Prepare the discharge electrodes:

1. Cut two 15" lengths of brazing rod and bend them as shown. I bent mine by hand but you could bend a 30" length around a five gallon pail and then cut it in the center for a neater appearance.

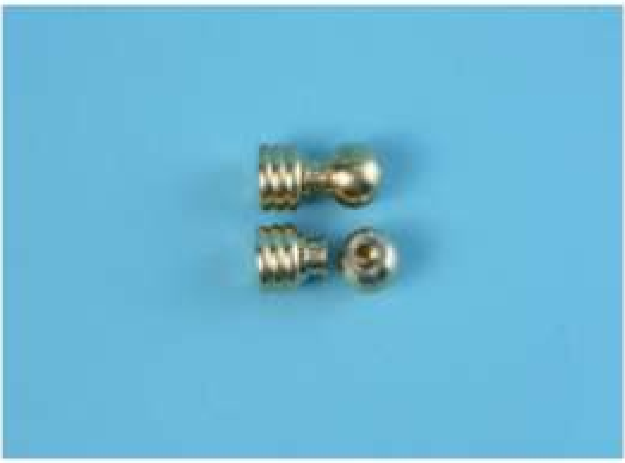

wimshurst-discharge-ball-cut.jpg

2. The balls for the discharge electrodes come from some more lamp finials, cut them off just below the ball with a hacksaw. These balls are about ½" in diameter.

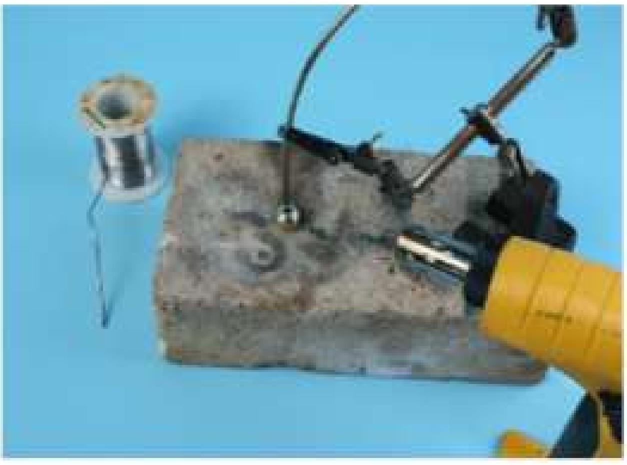

wimshurst-discharge-ball-solder-prep.jpg

3. Solder the discharge balls to the electrodes; fill the hole with solder so it makes a smooth transition to the rod.

Note: do not solder the small ball nuts in place!

Fabricate the neutralizing brushes:

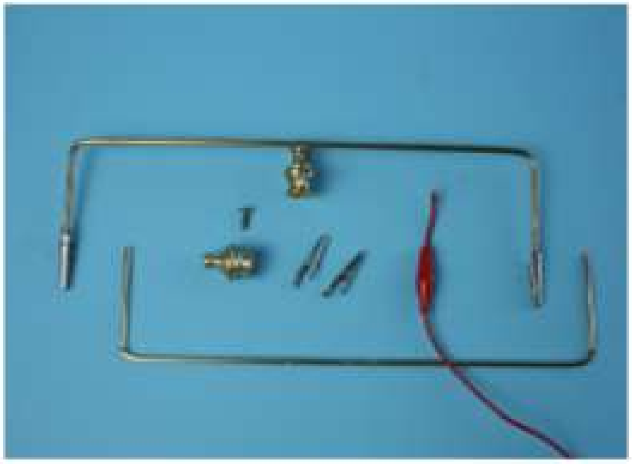

wimshurst-neutralizer-bar-parts.jpg

The neutralizing brushes are made with more brazing rod, alligator clips salvaged from a pair of clip-leads, and yet another type of lamp finial.

Bend the brush support:

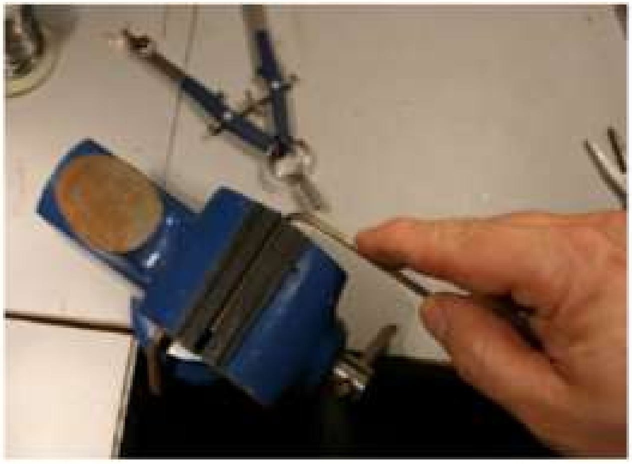

wimshurst-neutralizer-bend.jpg

1. Cut a length of brazing rod 14" long and mark it 2" from either end.
2. Make (2) 90 degree bends in the rod at the 2" marks.

Solder the brush support to the brush boss:

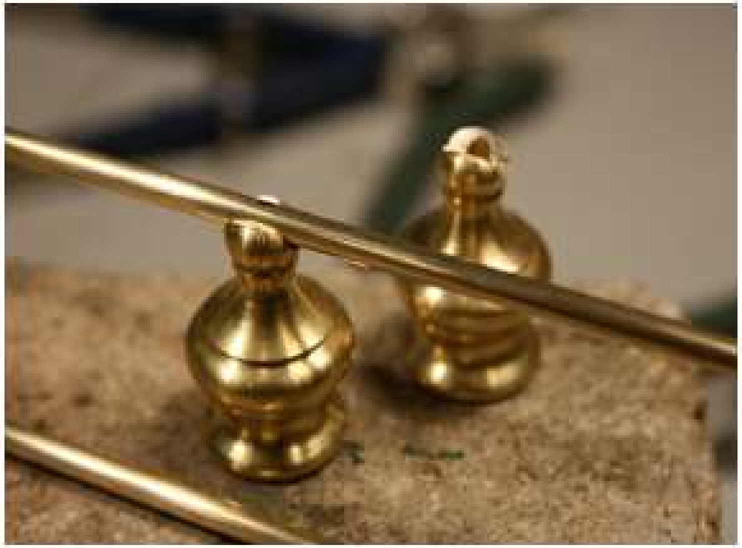

wimshurst-neutralizer-hub.jpg

1. Drill a hole for a set screw in the base of the finial and tap with the #6-32 tap.
2. File a groove in the top of the lamp finials, these particular finals have a 3/8" threaded hole in the bottom and a small hole in the top. I think they are made for ceiling fixtures that have a center pull string.
3. Center the neutralizer bar on the finial and prop it so its parallel to the workbench top and solder it in place.

Attach the brush clips to the support:

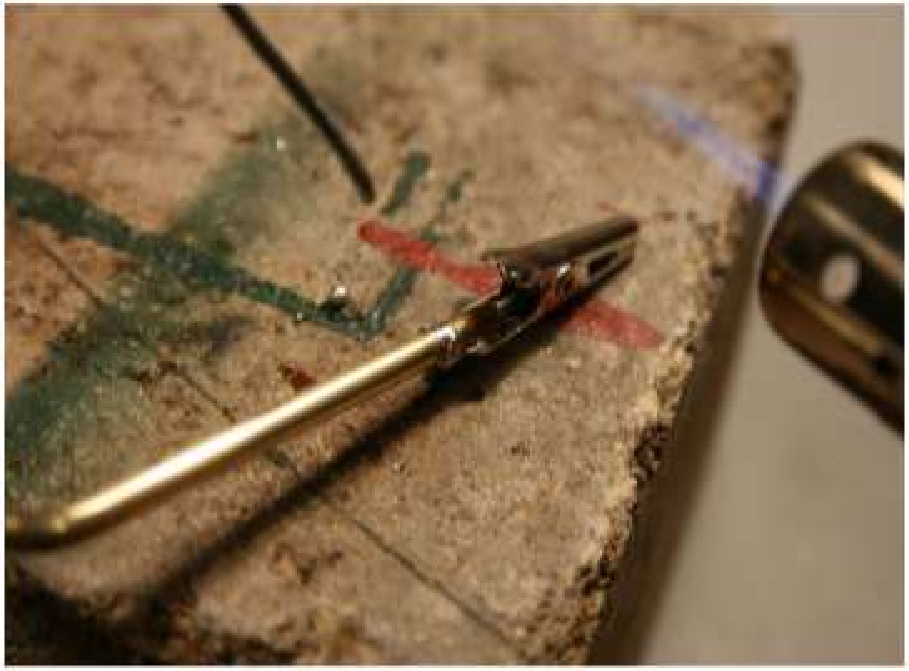

wimshurst-neutralizer-clip.jpg

1. Crimp the alligator clips on to the ends of the neutralizer bar and solder.

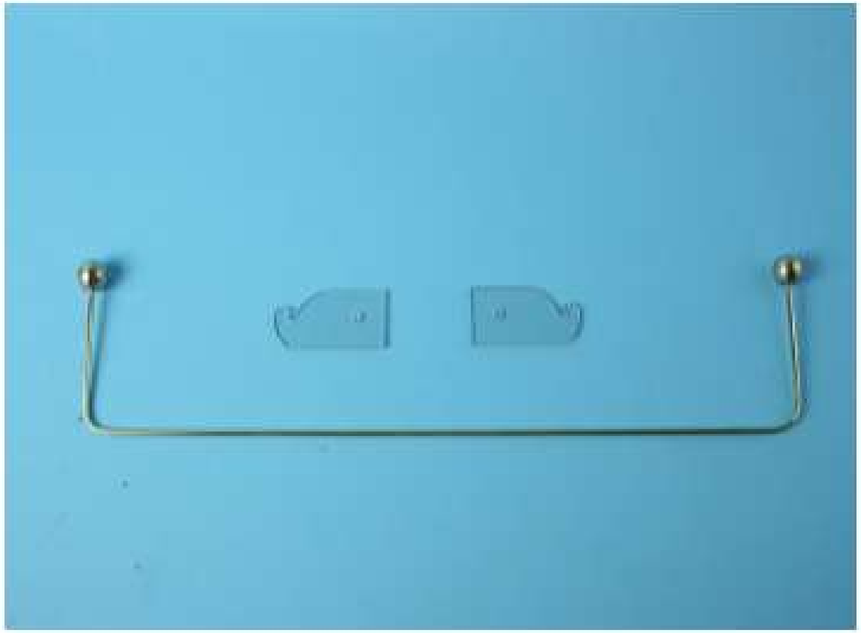

wimshurst-layden-shunt.jpg

Fabricate the Leyden jar shunt:

1. Cut a 22" length of brazing rod.
2. Make 90-degree bends, 3 ½" in from each end.
3. Solder two brass balls to the end. These are the large brass lamp chain pull balls, smaller finial balls or cabinet knobs would work here, too. If you use knobs be sure to remove any lacquer finish.

Cut the Leyden jar body:

wimshurst-layden-cut-tube.jpg

1. Using the miter box and fine tooth hacksaw, cut two 7 1/2" lengths from the fluorescent lamp protector sleeve.

Cut and affix the inner plate:

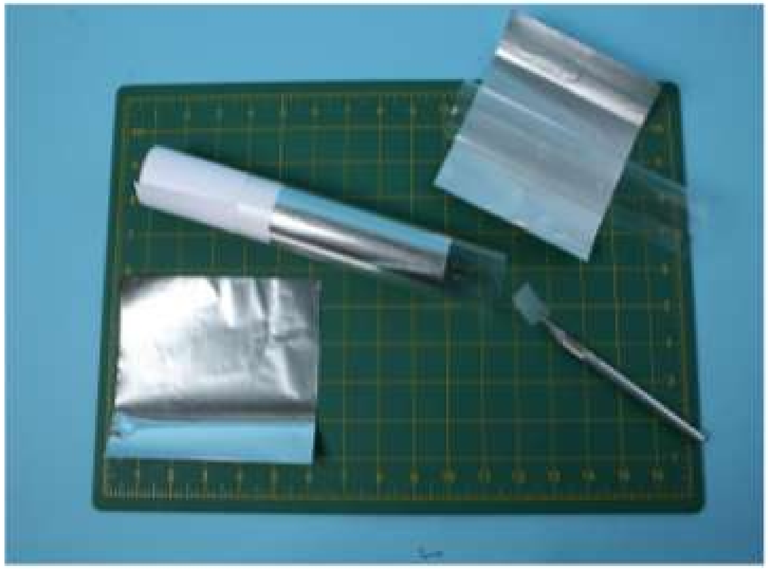

wimshurst-layden-plate-inner.jpg

1. Cut (4) 5" by 6" sheets of heavy duty aluminum foil.
2. Form one sheet by wrapping it around the tube and then rolling it so it can be inserted. Roll along the 6" axis so the foil cylinder ends up being 5" high.
3. Insert the foil into the tube so that it is 1" from one end. Use a couple of rolled up sheets of paper to hold the foil firmly against the inside of the tube while you tape it in place. The tighter you can make it to the inside of the tube the better.

Affix the outer plate:

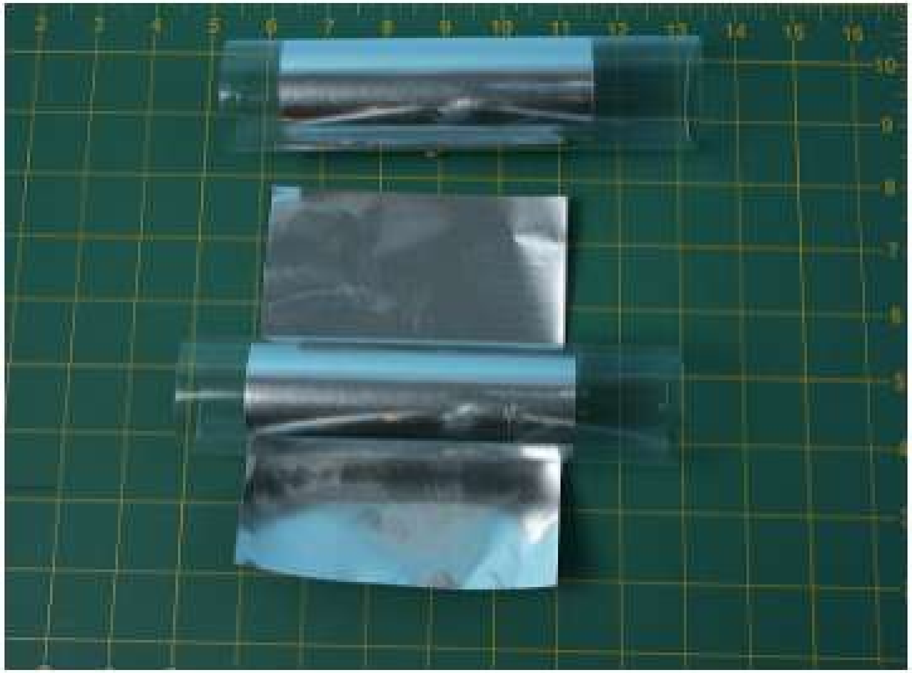

wimshurst-layden-plate-outer.jpg

1. Wrap another piece of aluminum foil around the outside and tape it in place. Again, the tighter the better, but don't wrinkle the foil.

wimshurst-layden-plate-bottom.jpg

Make the bases:

1. Snap the tube ends onto the opening that is 1" from the foil
2. Make the Leyden jar bases from a pair of plastic closet pole mounts. Drill out the center hole to 5/16".

Note: These are Stanley brand and I had to trim some reinforcing ribs off with an X-acto knife to make them slide into the tubes.

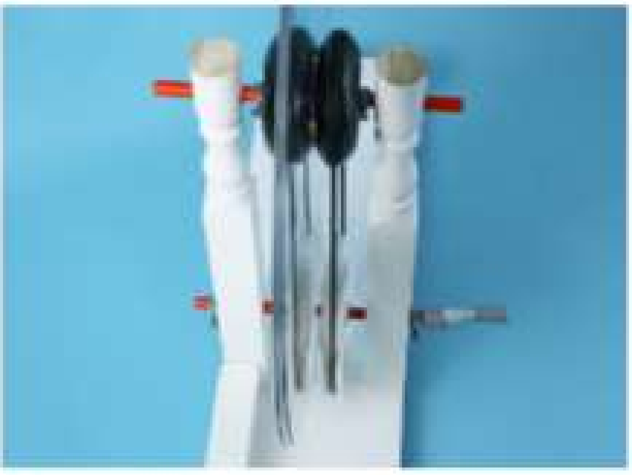

wimshurst-axle-collars.jpg

Mount the disks and drive line:

1. Slide the disk axle into a support and put on a 5/16" set screw collar, an O-ring belt, the two disks, the other belt, and another collar.
2. Attach the casement window crank to the drive shaft, insert the bushings in the supports if you are using them and slide the shaft through the pulleys. The pulleys should be a tight fit and you will have to twist the shaft back and forth to get it through. Don't forget about the belts hanging from the top shaft, one will need a twist so that the disks rotate in opposite directions. A collar goes on either end of the drive shaft.
3. Once both shafts are in place, stretch the belt around the pulleys. (In the picture, the belt with the twist is hidden behind the disk. What you are seeing is a reflection of the untwisted belt.)

Note: 5/16" set screw collars can be found at the hardware store but I made my own by drilling out a 5/16" nut and threading a #6-32 screw into the side.

Note: I found that my machine became difficult to turn once it was fully charged due to the electrostatic attraction of the disks. I cut a 2 ½" washer from a plastic milk jug and placed it on the shaft between the disks to remedy this problem.

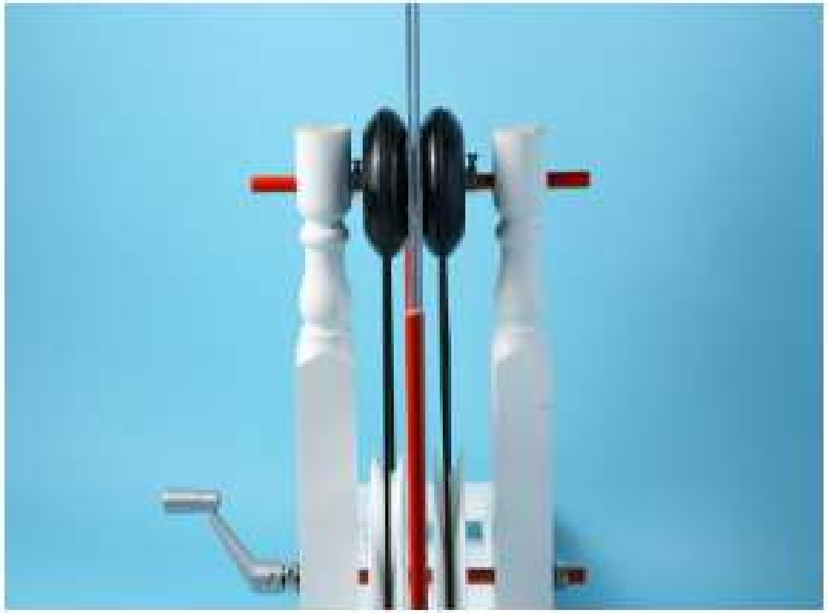

wimshurst-lineup-collectors.jpg

Align the disk and collector supports:

1. Cut two 11" lengths of fiberglass rod and press them into the holes made earlier in the base.
2. Loosen the screws that hold the two supports to the base and slide them around to adjust the disks so they line up with the charge collector supports.
3. Re-tighten the supports.

Install the Leyden jar base and inner plate contact:

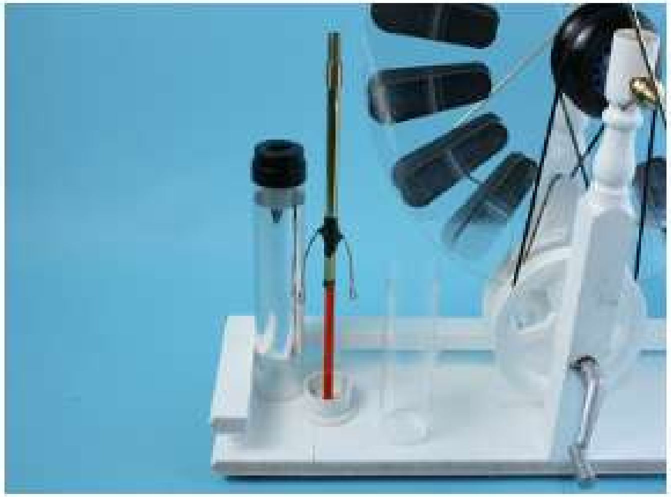

wimshurst-layden-contact-inner.jpg

1. Slide the Leyden jar bases onto the fiberglass charge collector supports.

2. Slide the charge collector assembly over the fiberglass supports.

3. Using about 6" of 14 AWG solid copper wire, form the inner plate contact. Wrap it once around the brass tube and form two loops in the ends.

4. Using a scrap of the plastic tube as a guide, adjust the inner plate contacts so they apply even and gentle pressure. You want good contact with the foil but you don't want to rip the foil when you install the Leyden jars.

Epoxy the charge collector assembly in place:

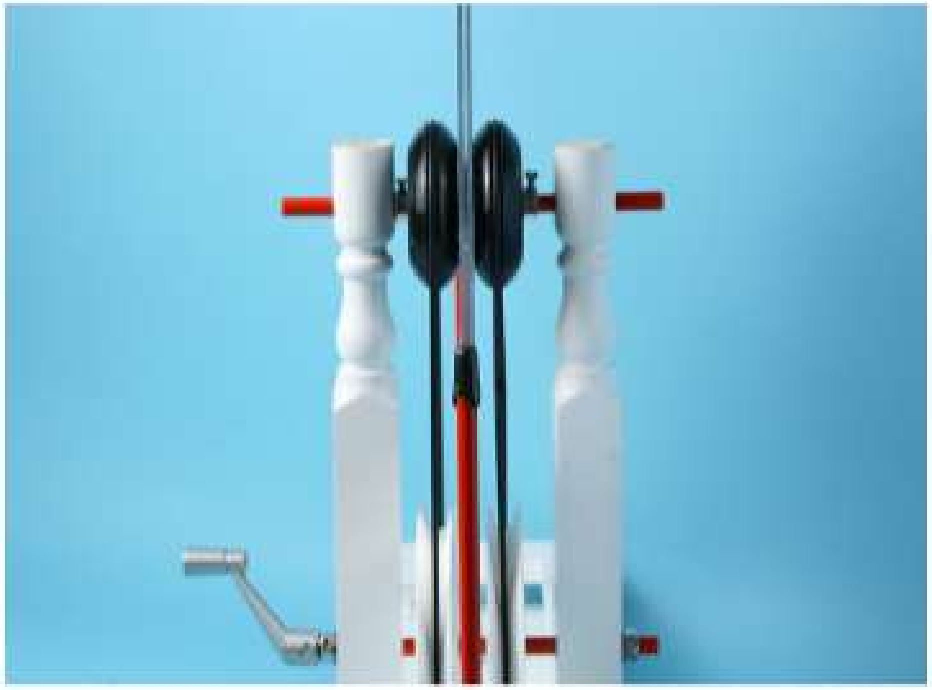

wimshurst-collector-epoxy.jpg

1. Apply epoxy to the end of the rod and slide the brass charge collector assembly down onto the fiber glass support rod.

2. Set aside while the epoxy cures.

Install Leyden jar and assemble collector:

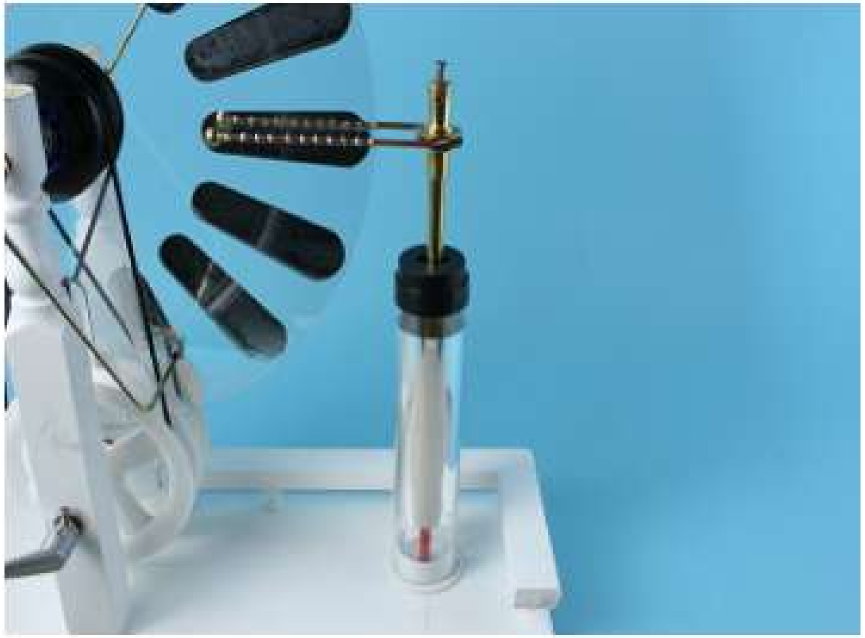

wimshurst-layden-complete.jpg

1. Slide the Leyden jar onto its base, being careful not to tear the foil as makes contact.

2. Line up the charge collector comb and trim the prongs. Test spin the disks to see if there is any wobble and trim the prongs to come as close as possible to the disks without touching.

3. Assemble the charge collectors.

Install discharge electrode:

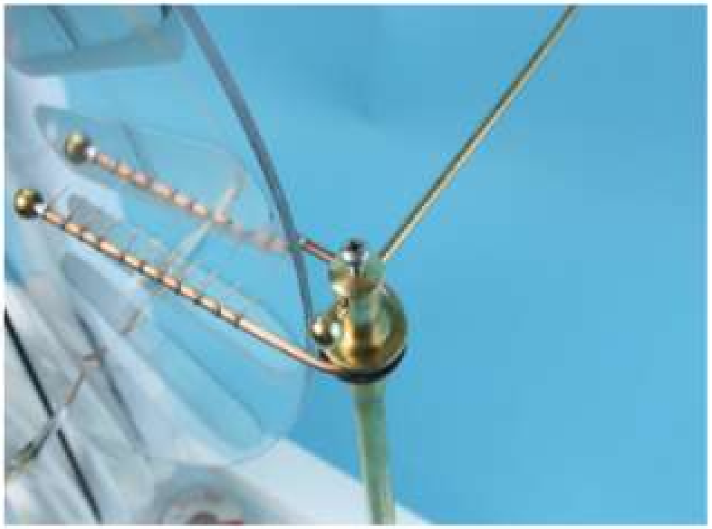

wimshurst-collector-inplace-2.jpg

1. Insert the discharge electrodes into the lamp finial on the charge collector and tighten the screw to hold it in place.

2. The finial should be tight enough to hold the collector comb but allow the discharge electrode to move back and forth. If it's too tight, or not tight enough, the support rod can be twisted in the base to accommodate.

3. Wrap a small bit of tape around the end of the electrode and screw on one of the small ball nuts; this will prevent charge from bleeding off the sharp end.

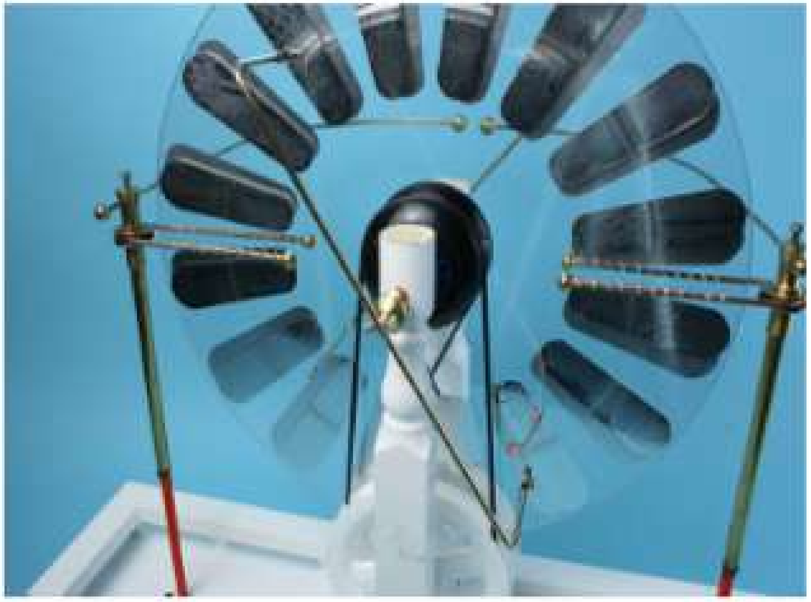

wimshurst-neutralizer.jpg

Install neutralizing brushes:

1. Slide the neutralizing bars onto the upper shaft and adjust them to be about 45 degrees from the collector combs.
2. Sectors should pass through a charge collector, encounter a neutralizing bar after about 1/6 of a rotation, and then encounter the other charge collector after a further 1/3 of a rotation.
3. Tighten the set screw to secure.

Position brushes:

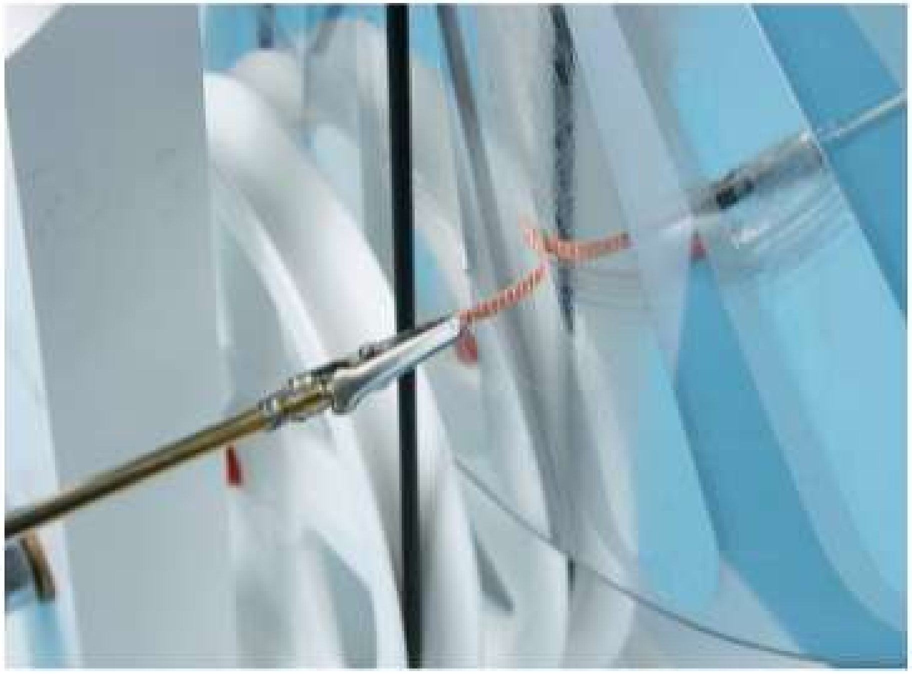

wimshurst-neutralizer-brush.jpg

1. Clip (2) 1 ½" lengths of Solder Wick™ to the ends of the neutralizing rods so they make good contact with the disk.

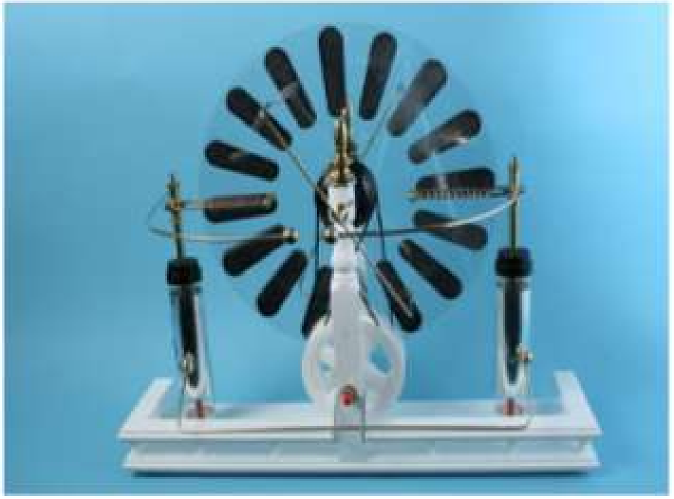

wimshurst-complete-front.jpg

Mount the Leyden jar shunt and add optional finials:

1. Use small brass wood screws to attach the (2) acrylic brackets to the front disk support, leave them a little loose at first.
2. Place the Leyden jar shunt in the brackets and line them up so the balls on the shunt lean comfortable against the Leyden jars.
3. Tighten the brackets.

The two tops of the disk supports looked a little bare to me so I raided my junk box for more lamp parts and came up with these decorative finials. The wealth of finial and cabinet knobs at the typical home center means that there are infinite

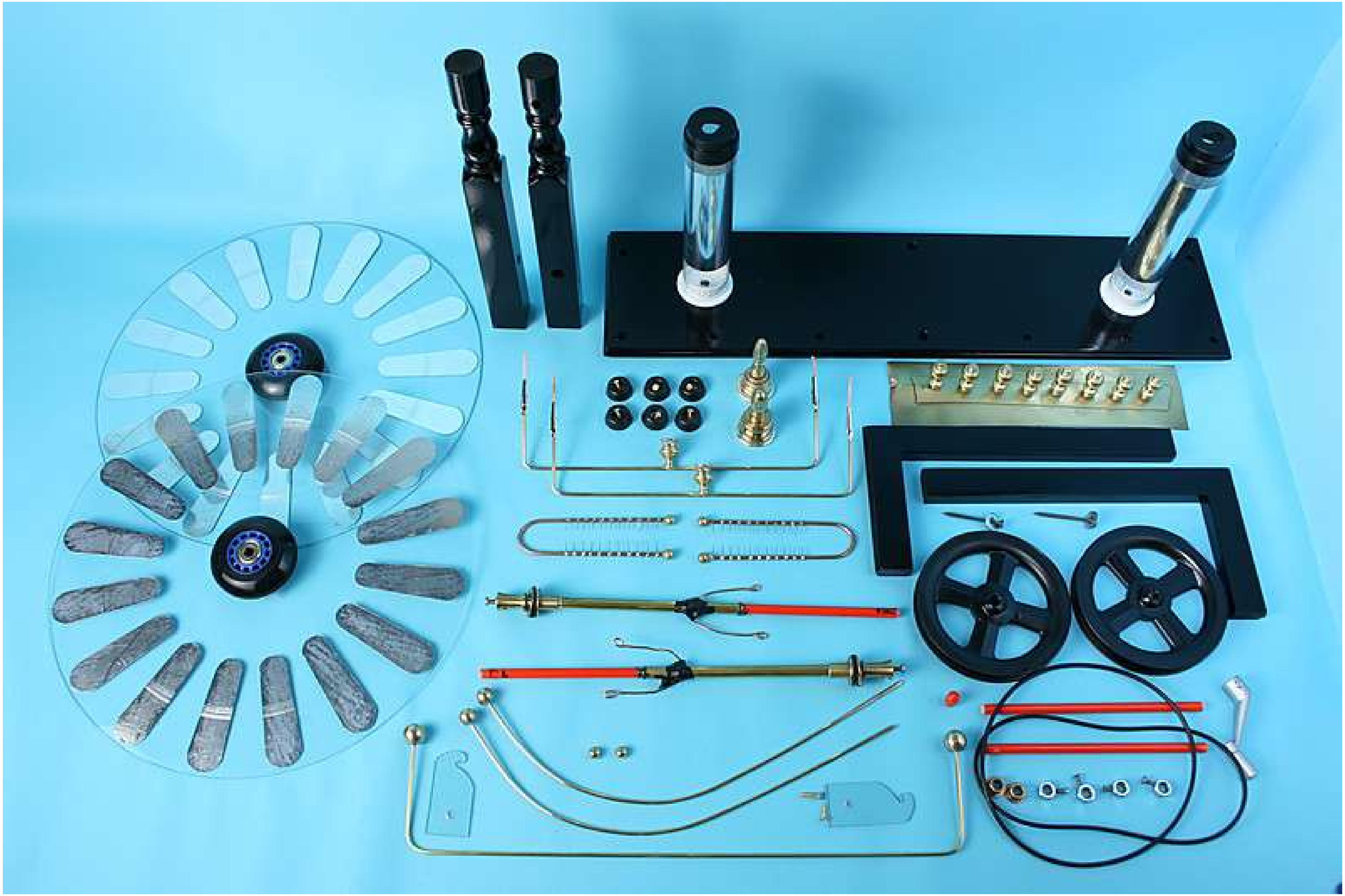

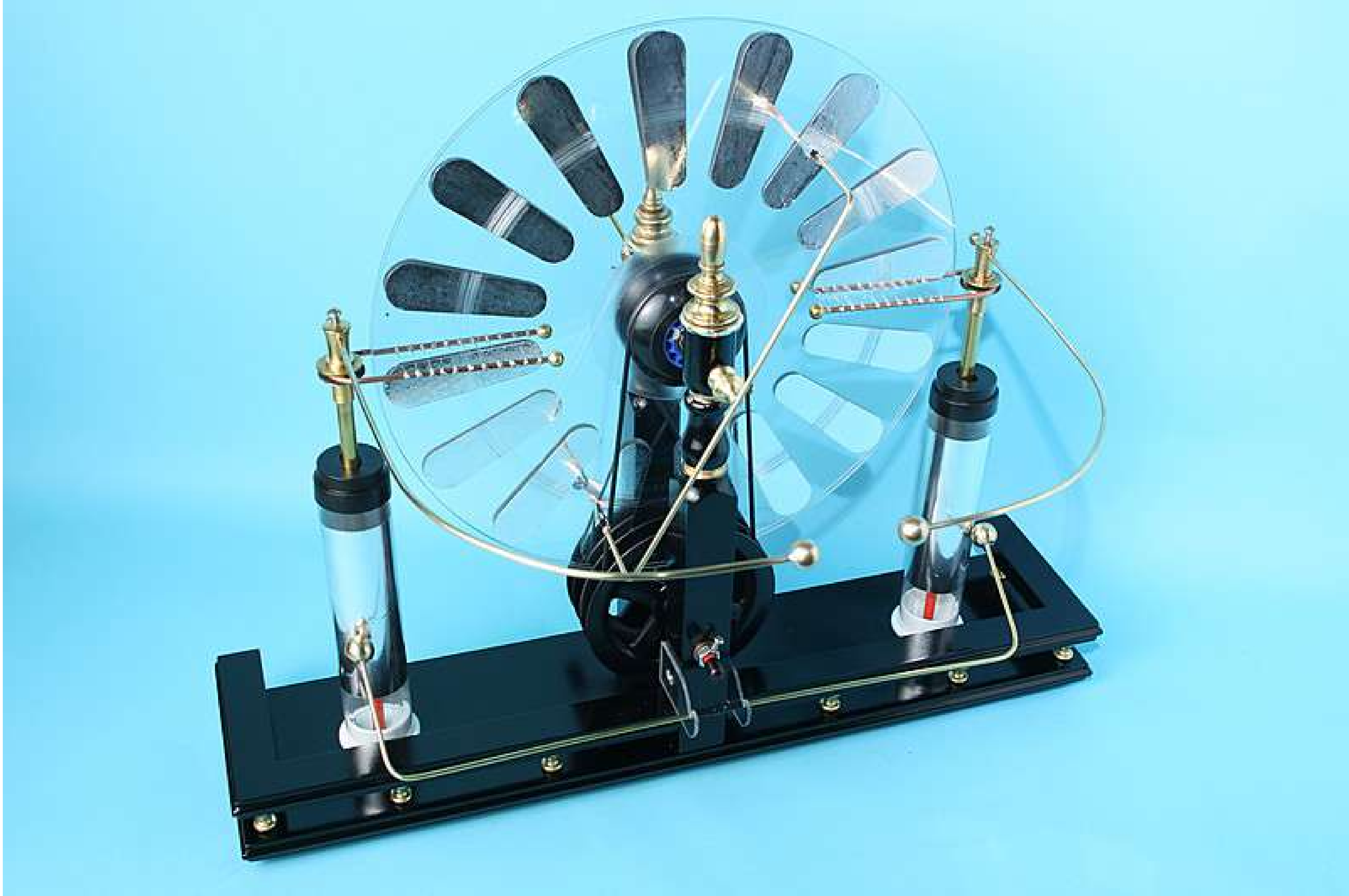

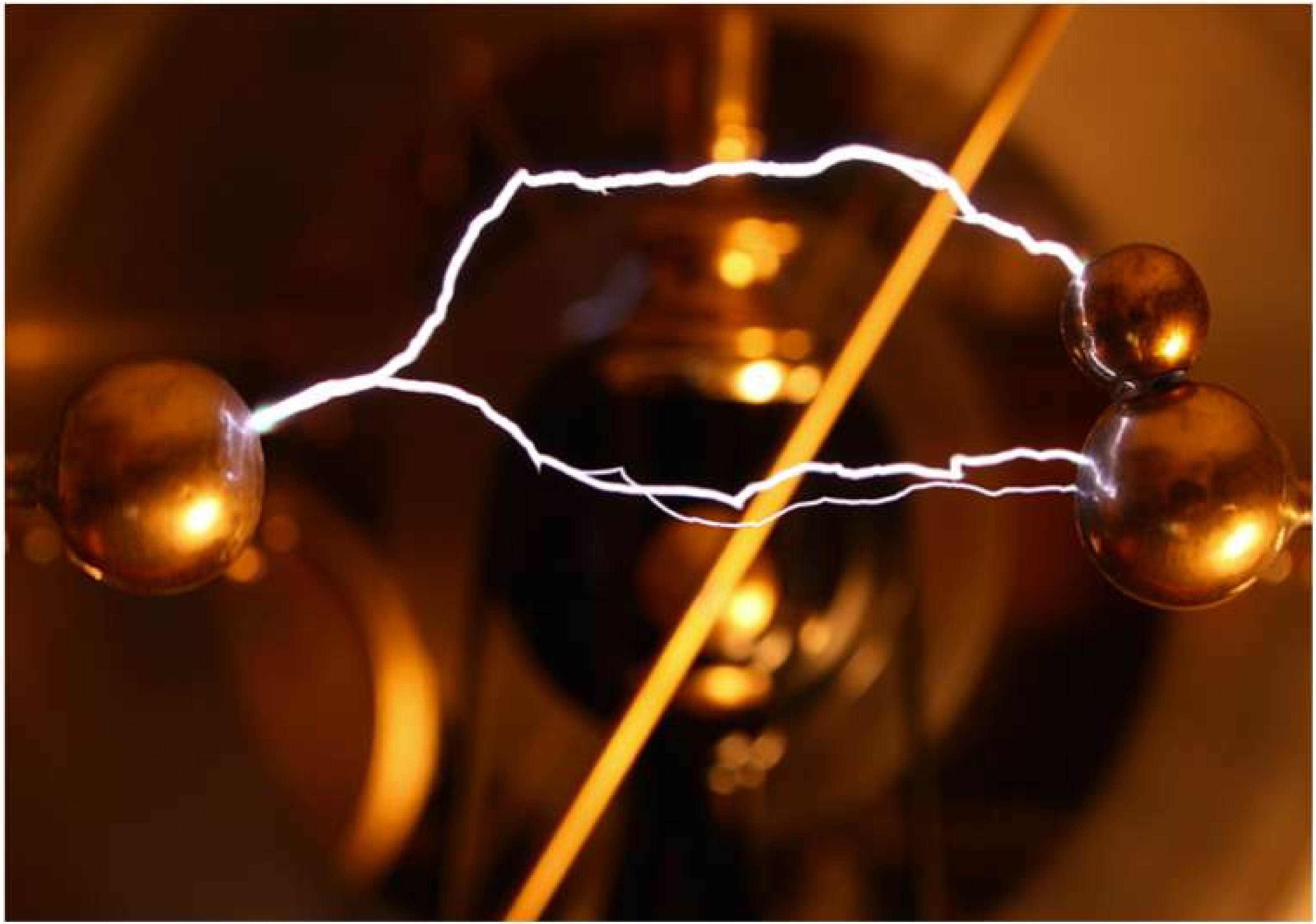

Attaching a small ball to the positive electrode will result in larger and more interesting sparks. The small ball creates a plume of ionized air that helps the spark jump the gap.

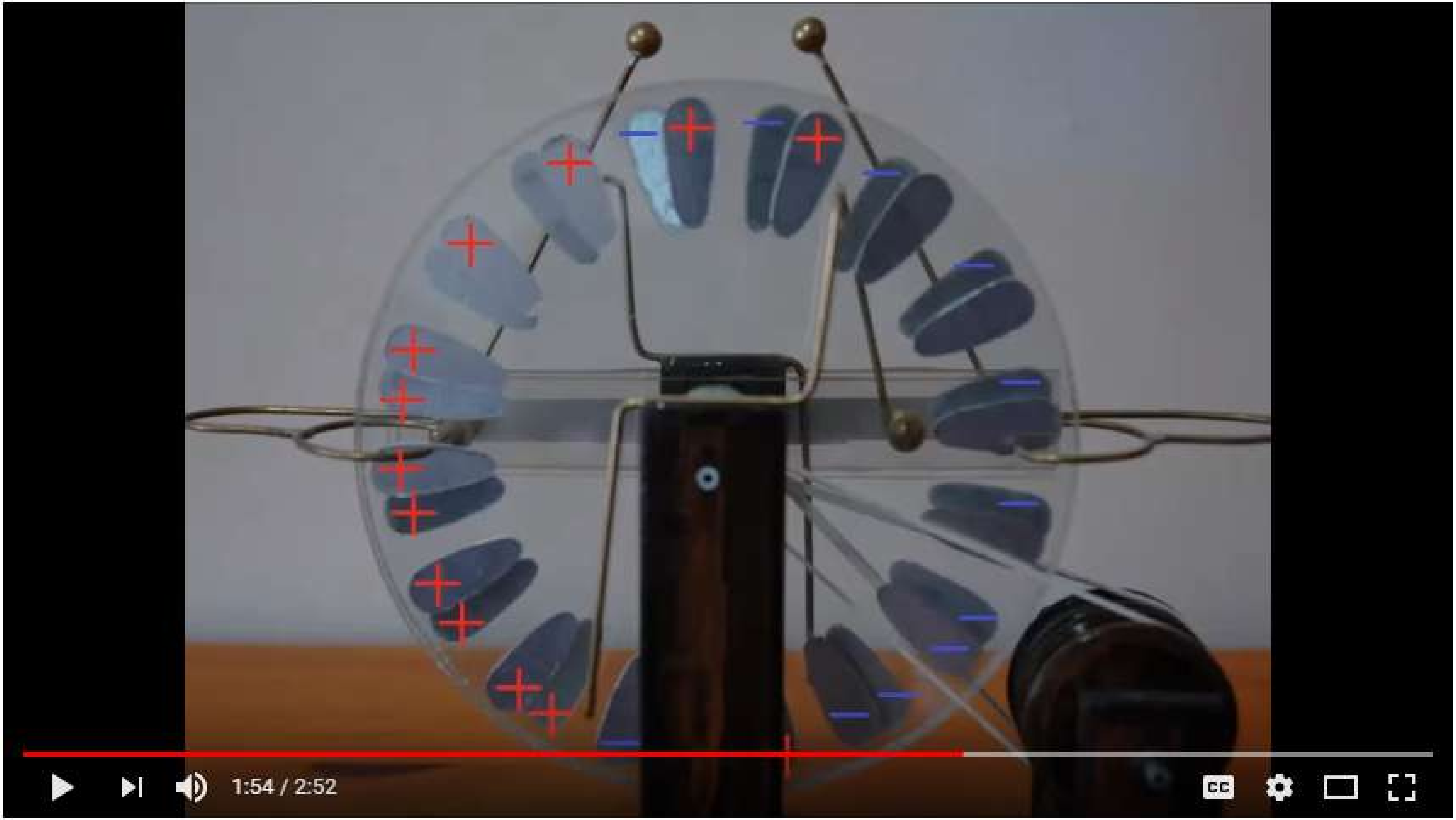

How the Wimshurst Works

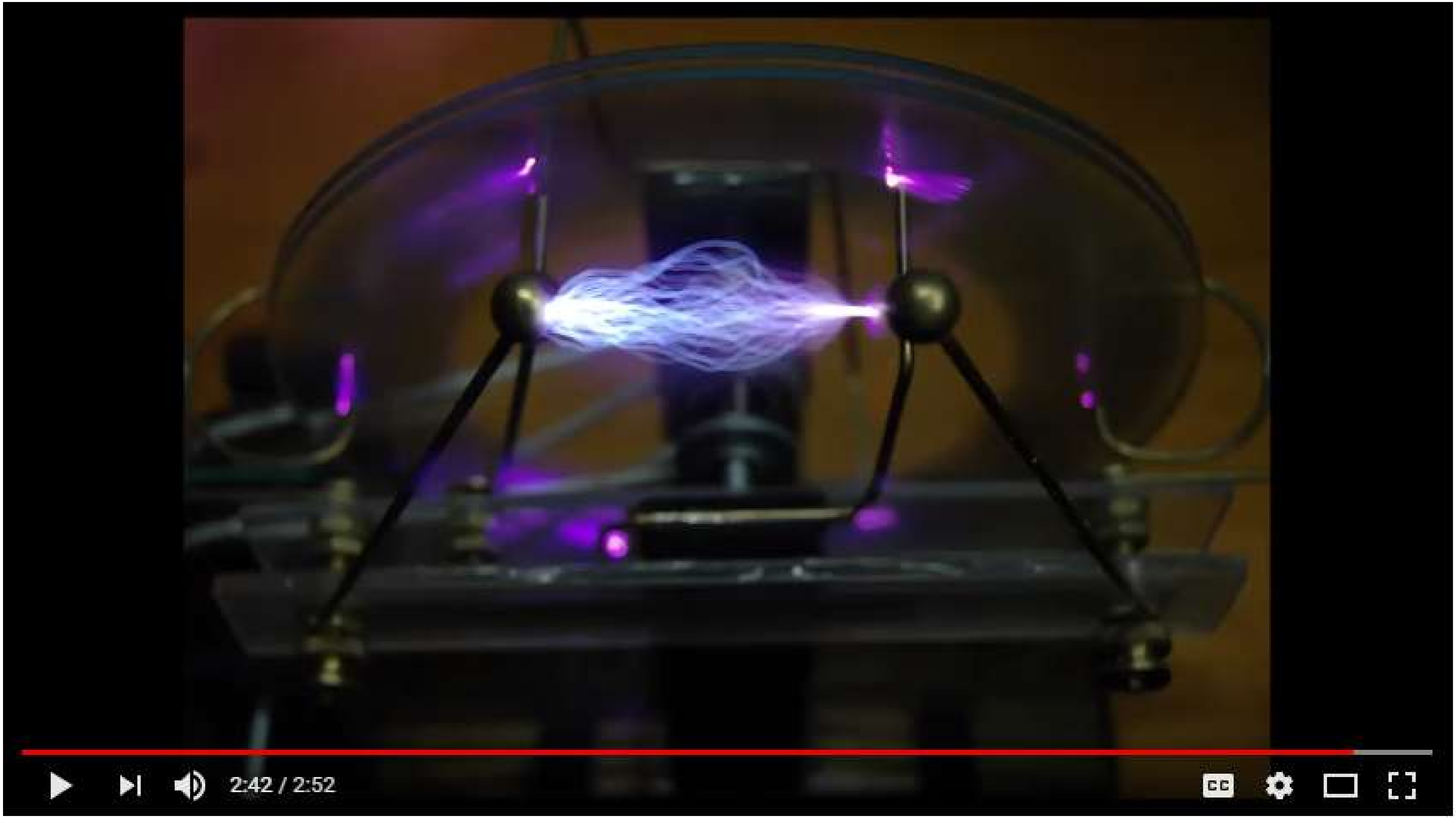

How the Wimshurst Works

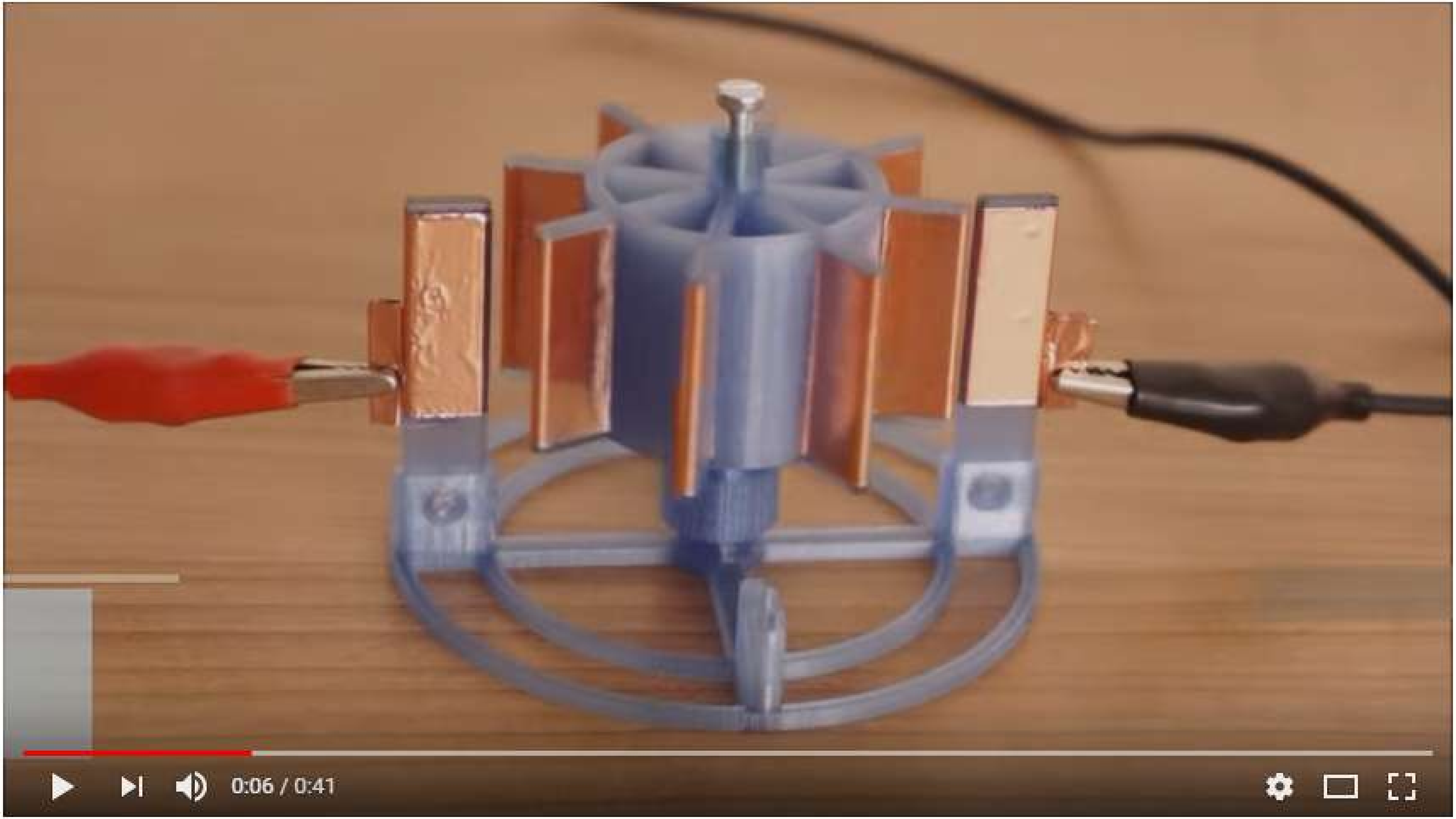

Electrostatic Motor

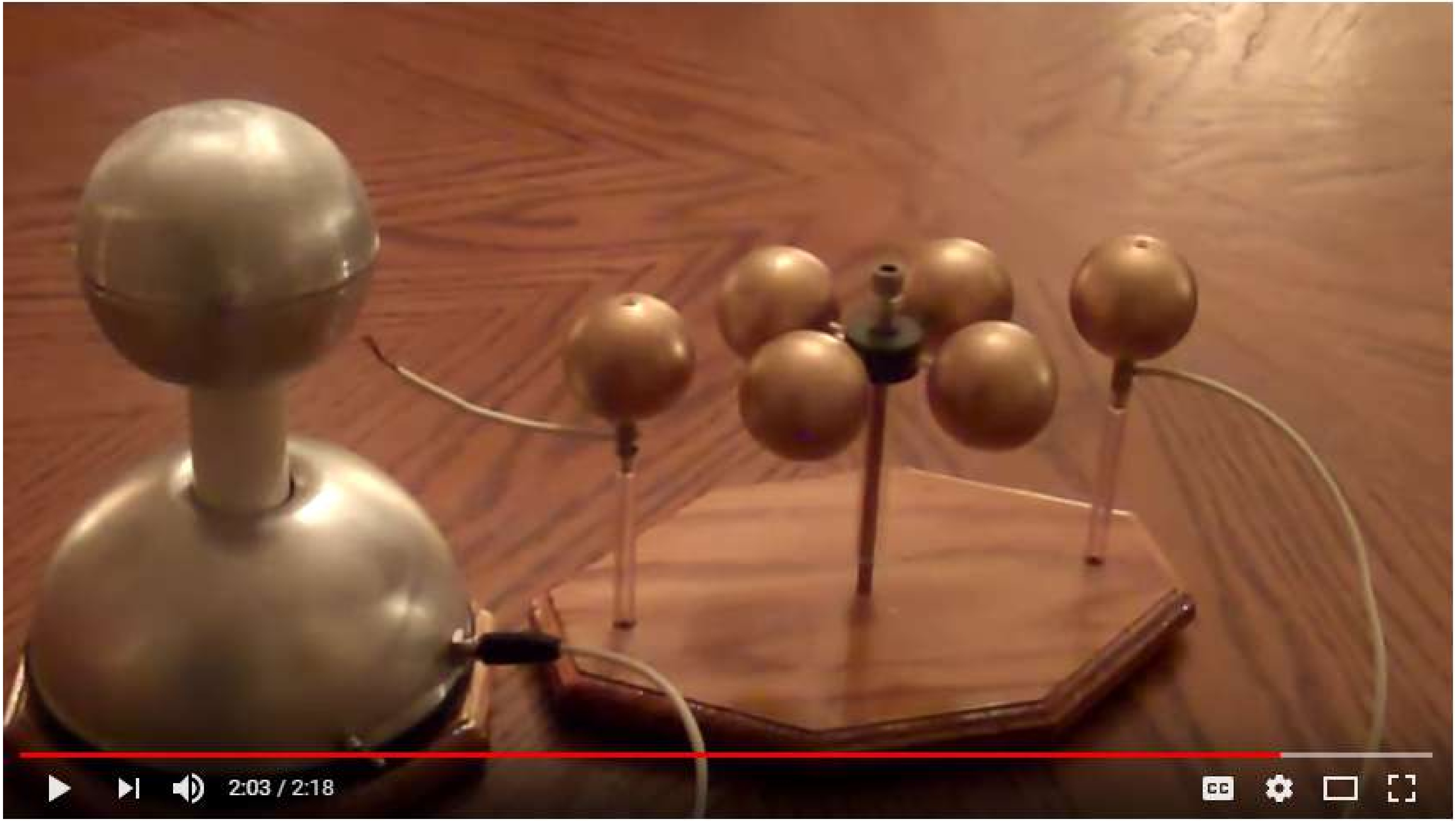

Home Made Electrostatic Motor

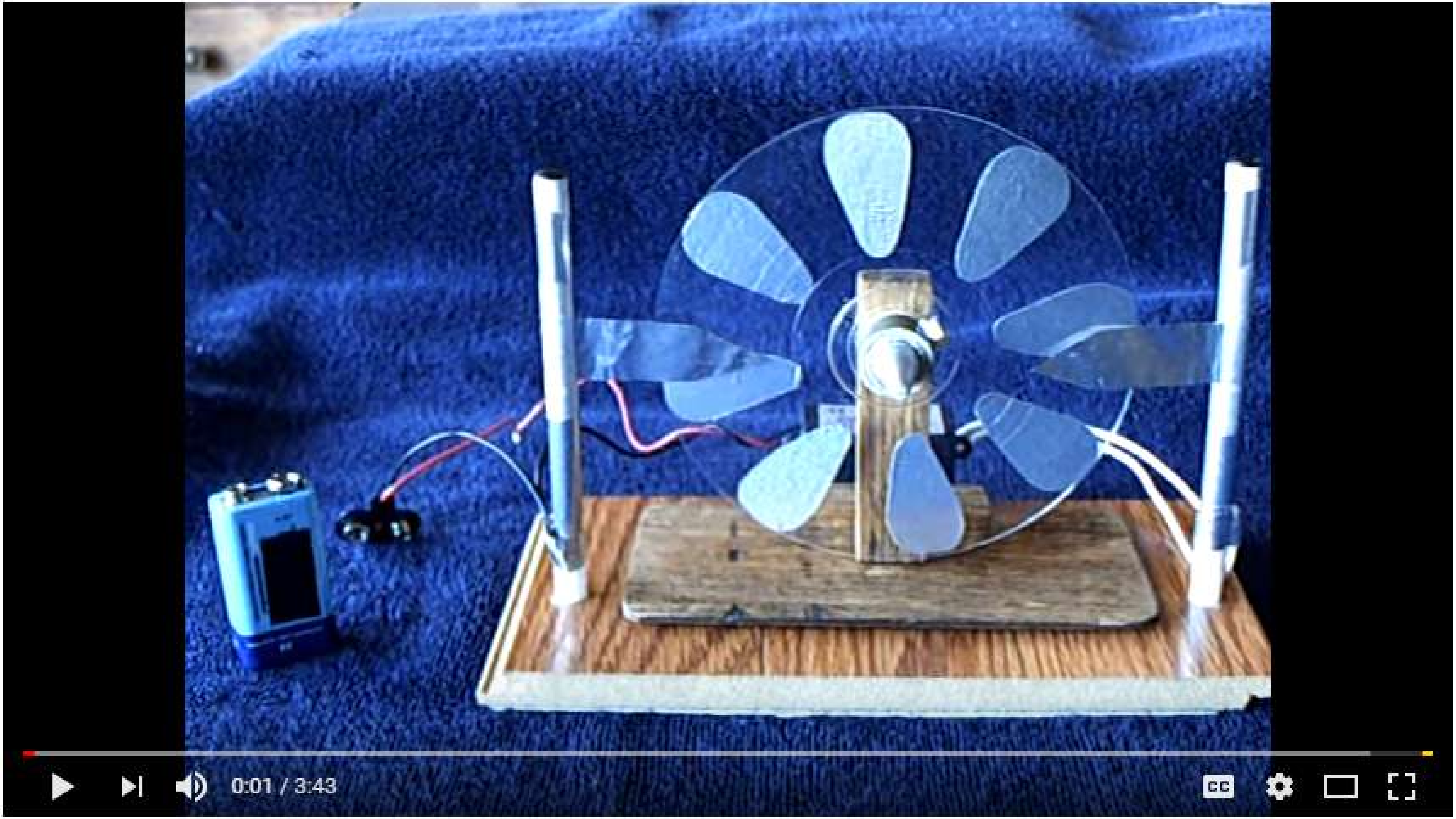

Clear Disk Electrostatic Motor

CE
2:57 / 3:43
CC

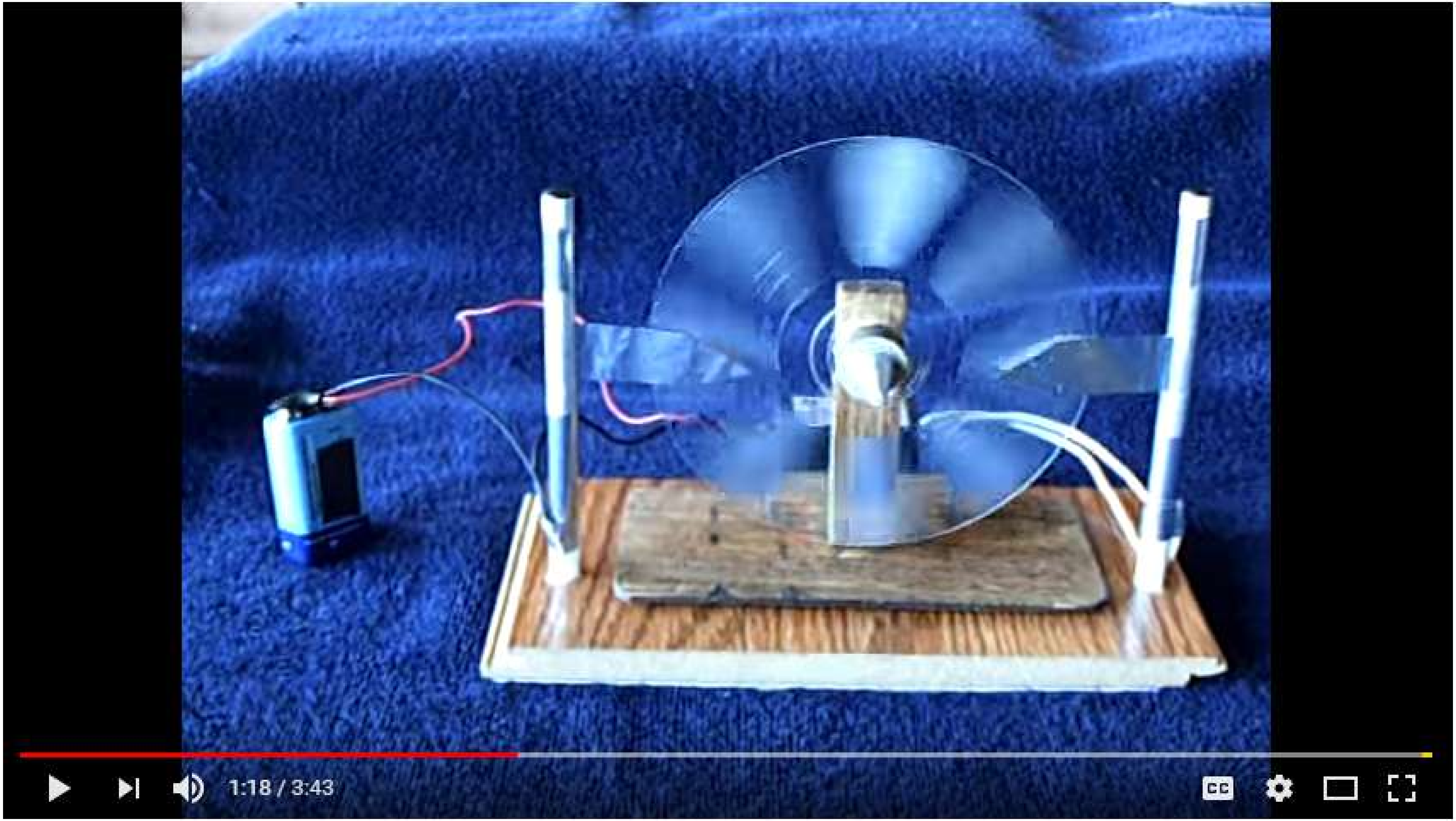
1:18 / 3:43
CC

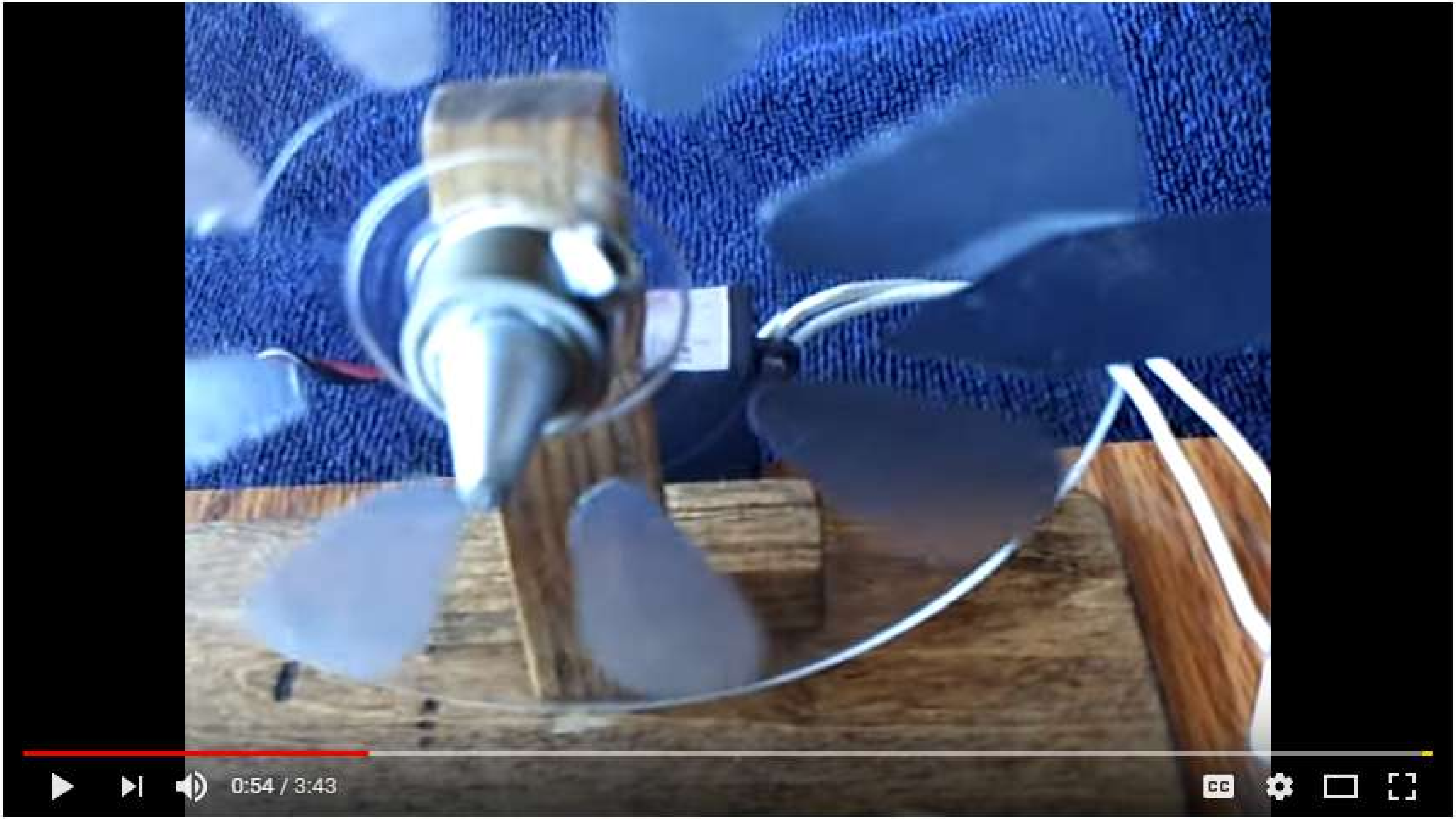
0:54 / 3:43

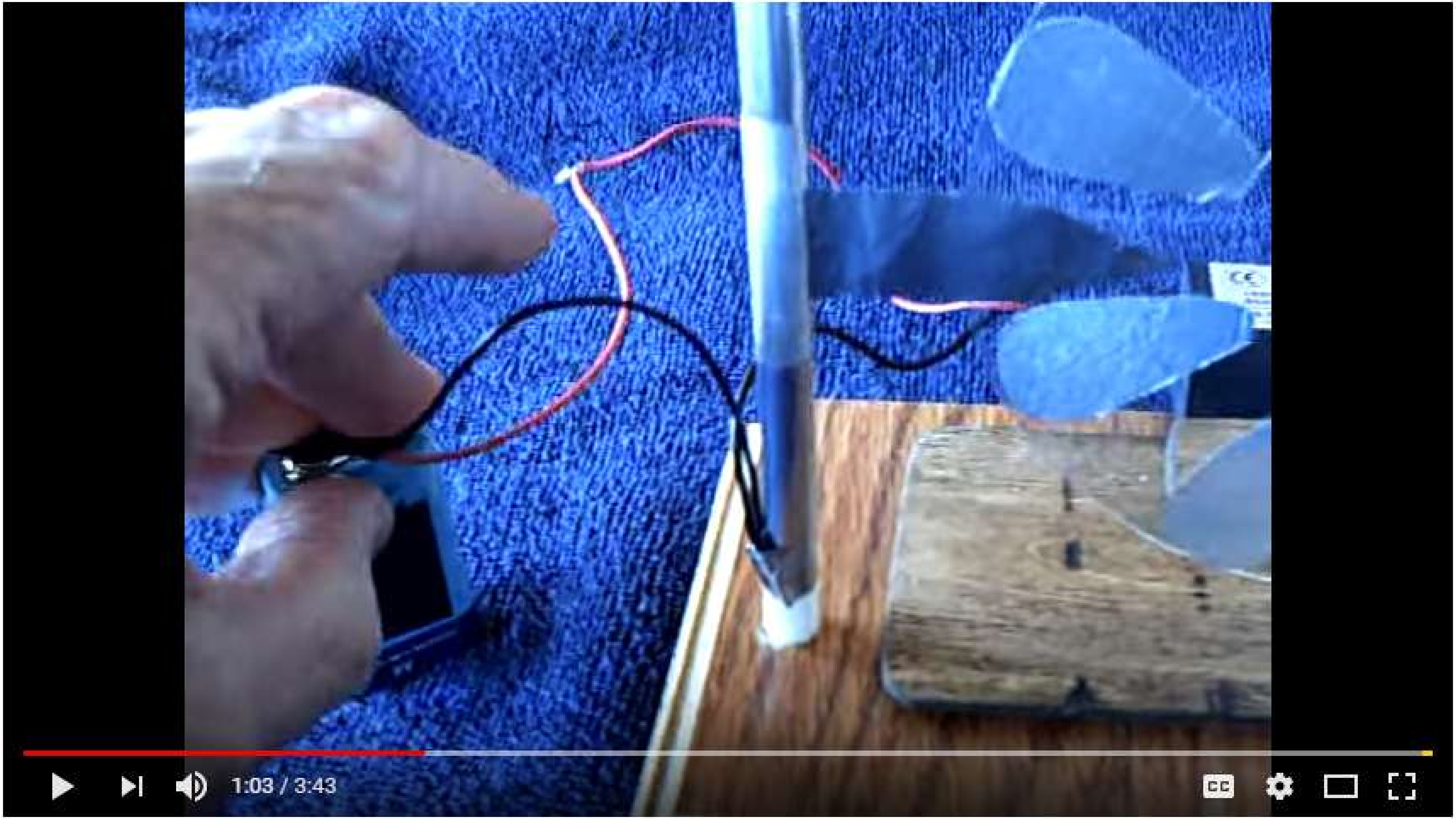
1:03 / 3:43
CC

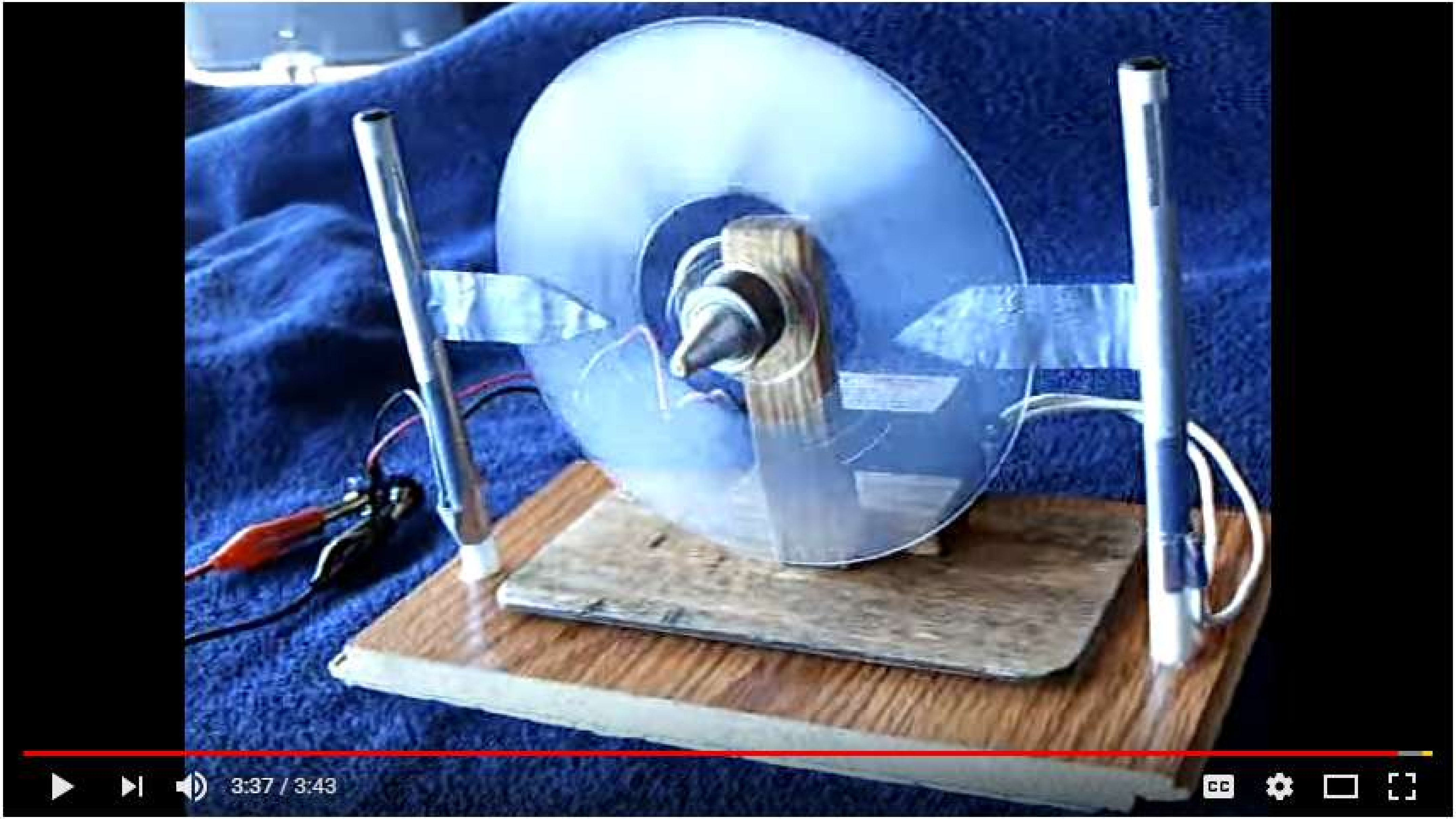
3:37 / 3:43
CC

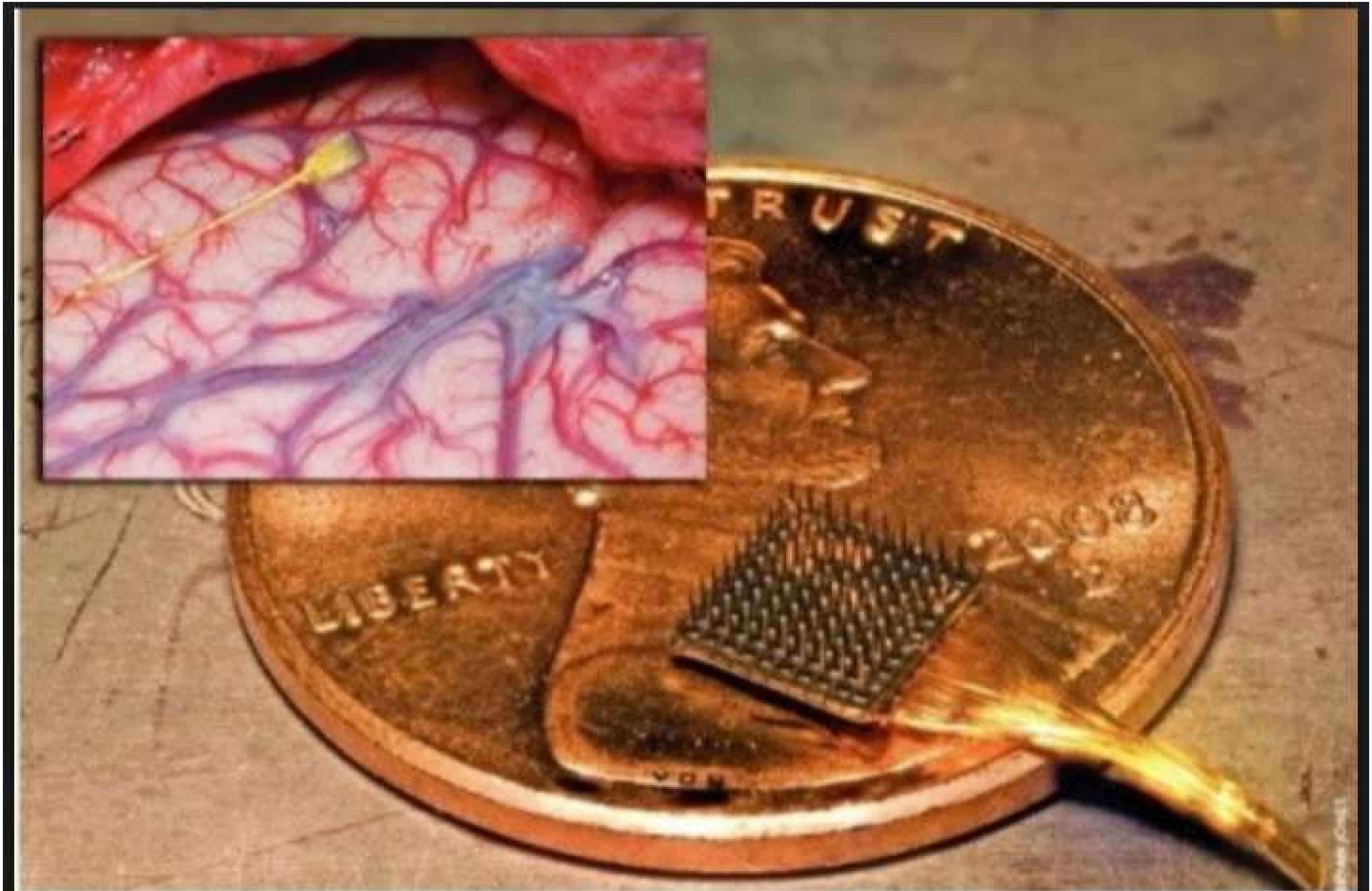
TRUST
LIBERTY
2008

Thought control: The Utah Electrode Array can be implanted on a human brain. For a podcast and more photos, go to CityWeekly.net.

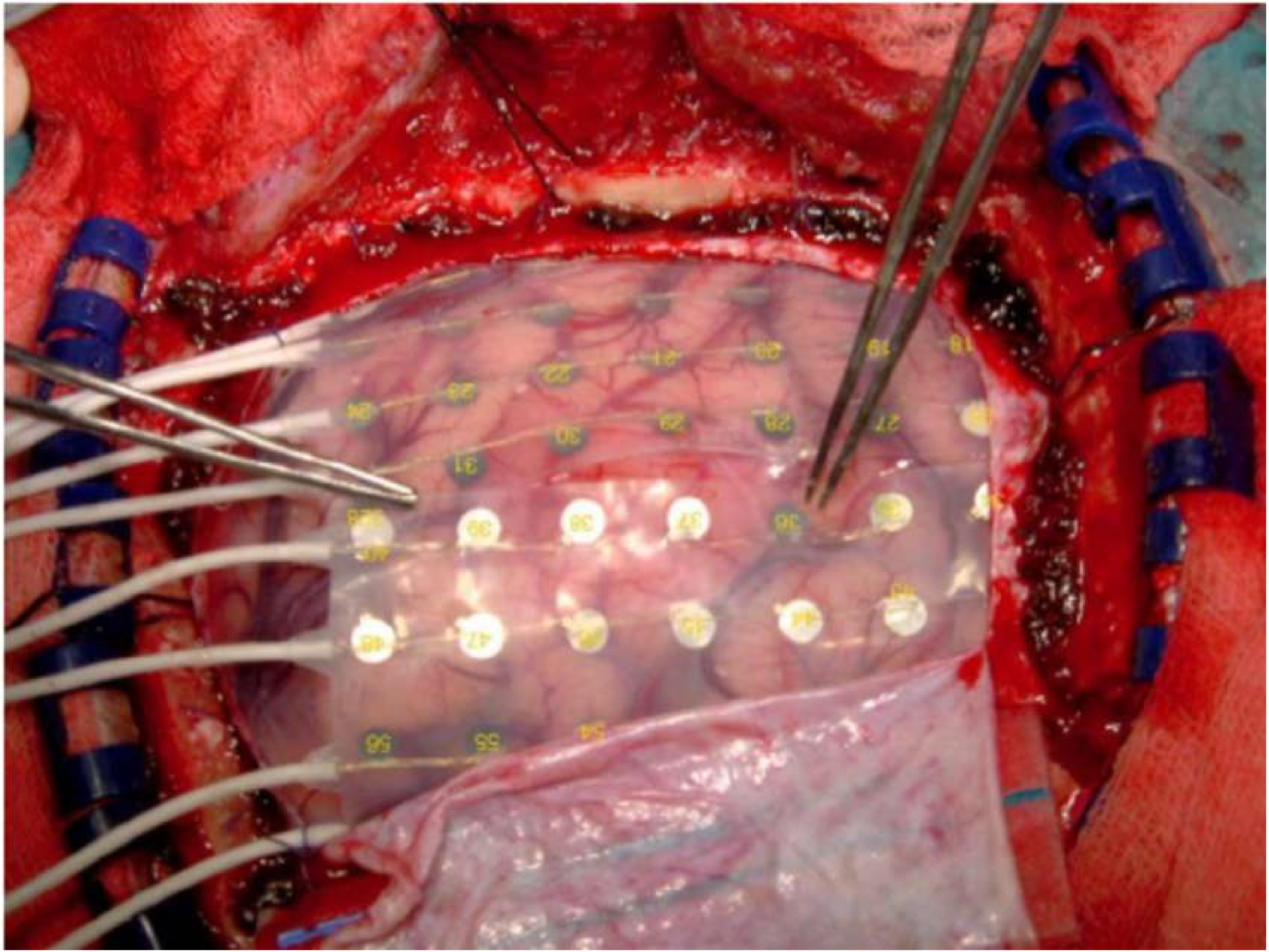

19
18
20
21
22
23
24
30
29
28
27
31
36
37
38
39
44
45
47
48
56
55
54

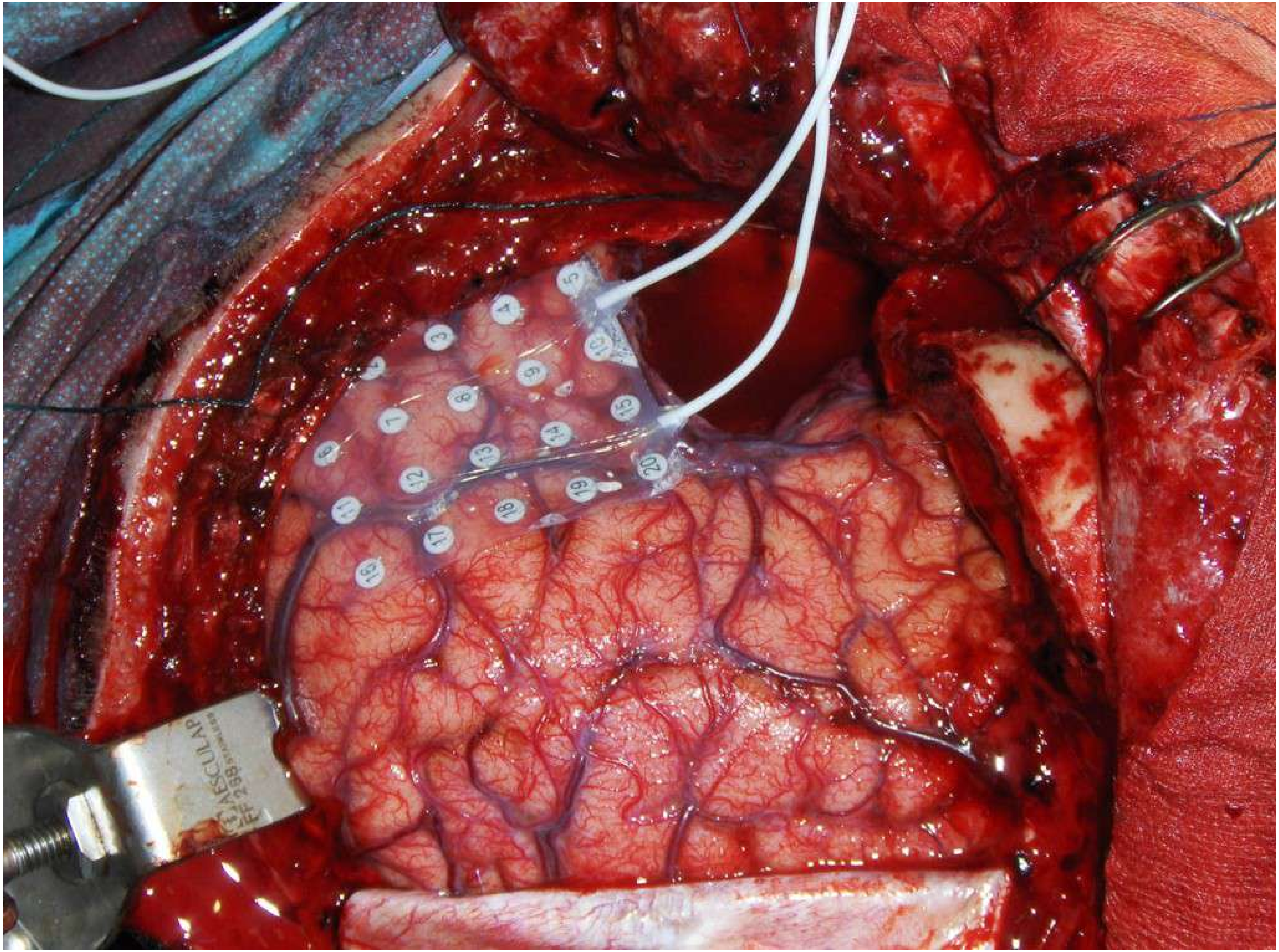
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
AESCULAP
FF 268

WWW.SADOUN.COM
1-888-519-9595

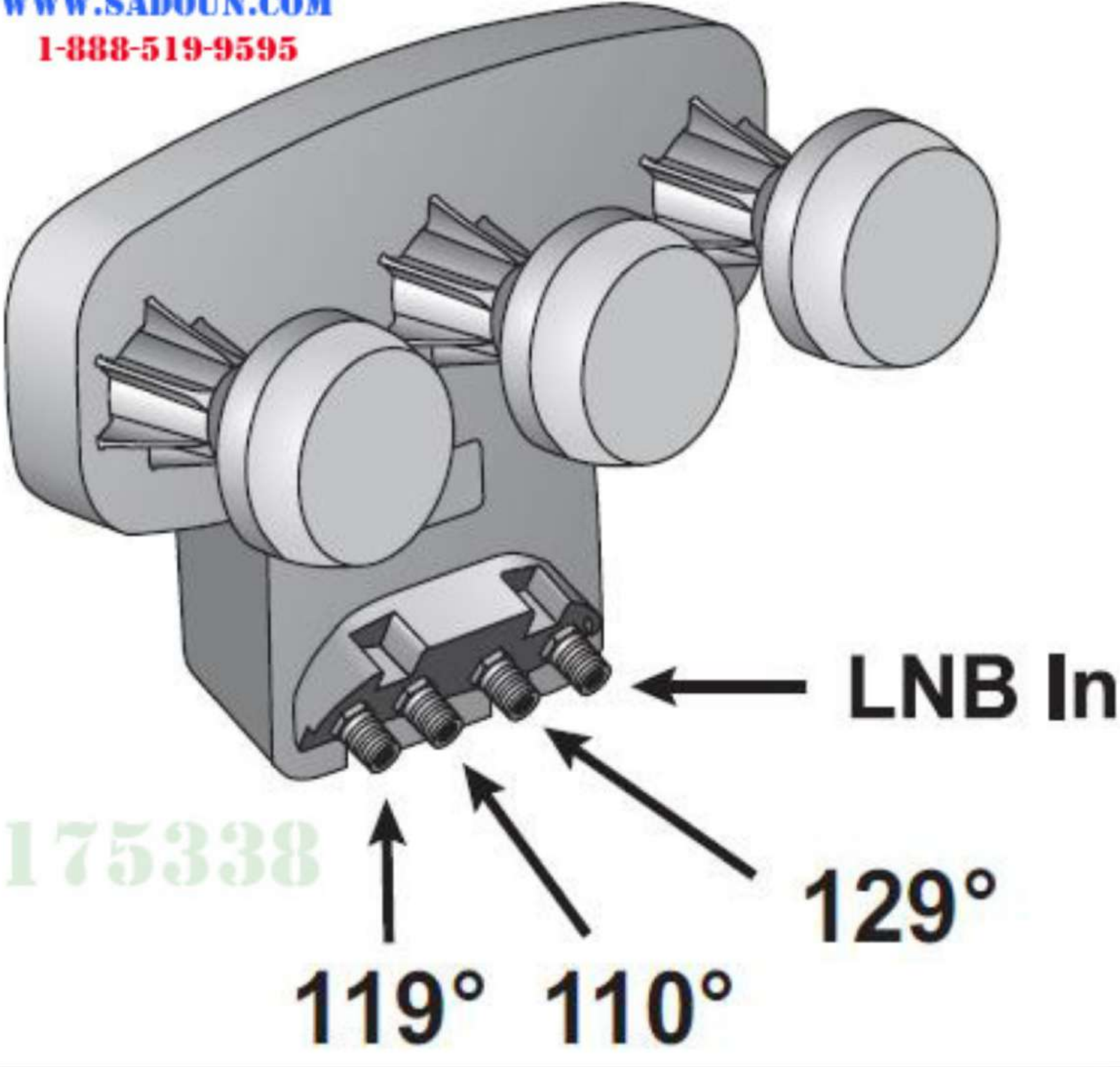
LNB In
129°
119°
110°

175338

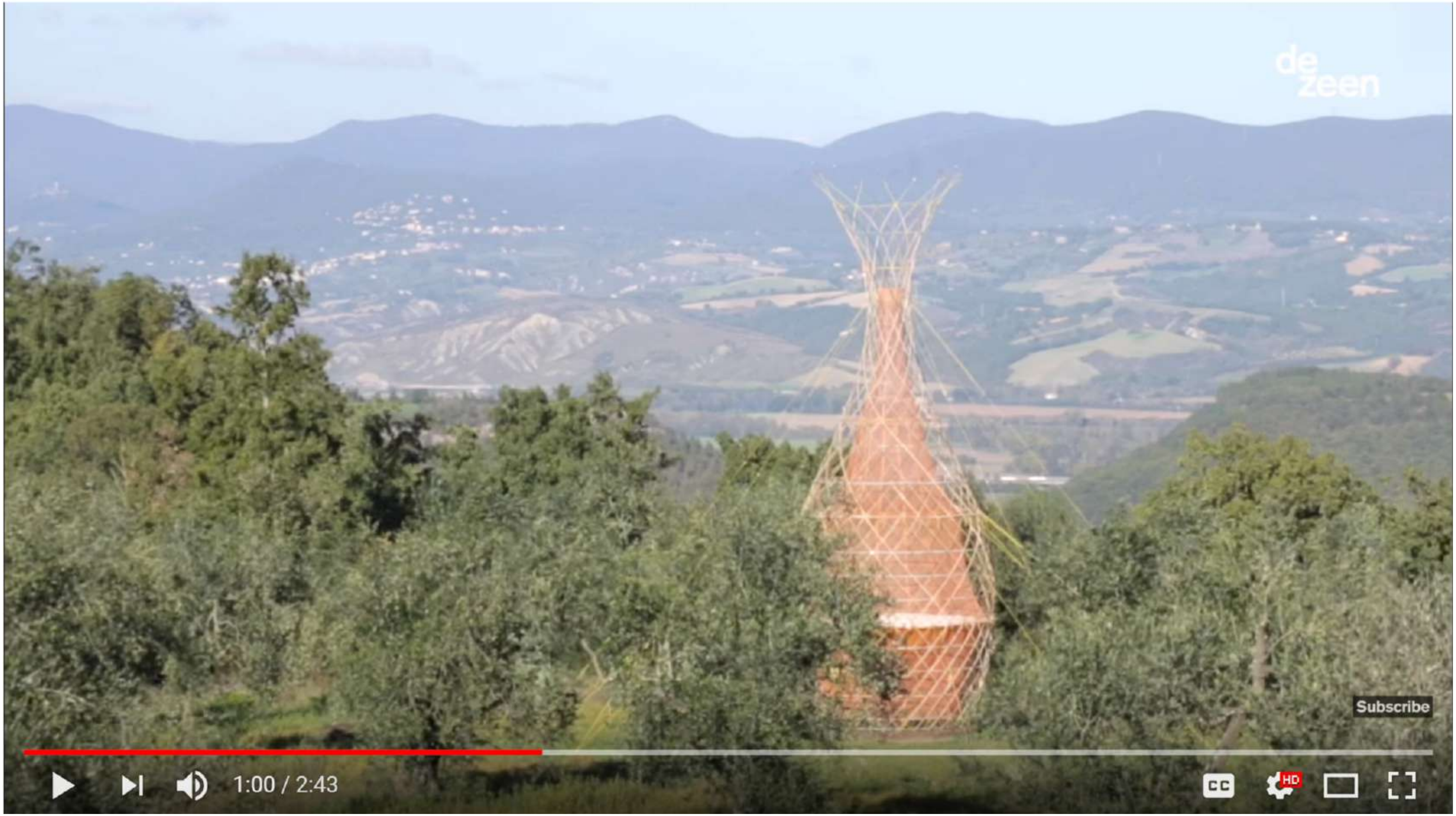
de
zeen
Subscribe
1:00 / 2:43
CC
HD

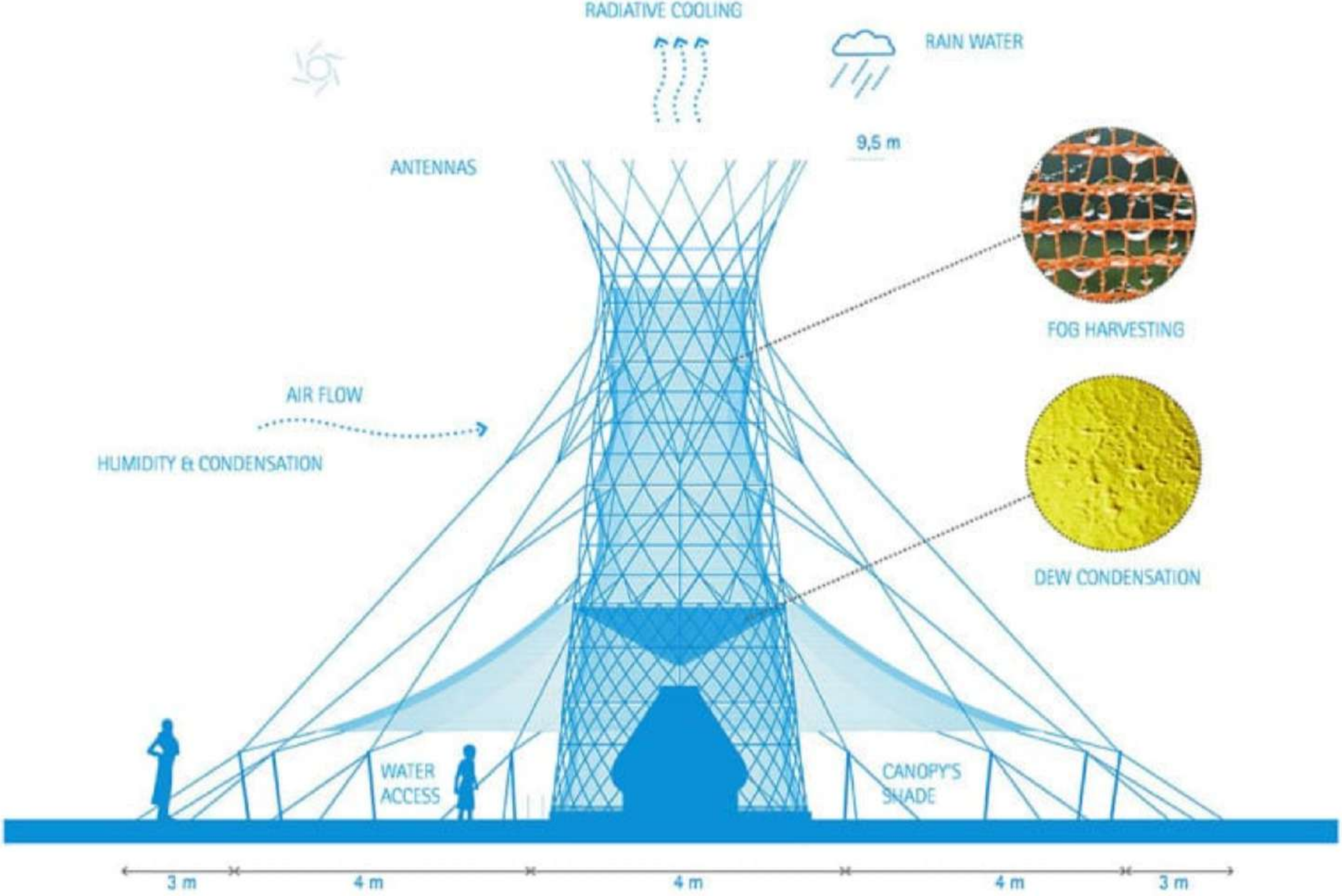
RADIATIVE COOLING
RAIN WATER
9,5 m
ANTENNAS
FOG HARVESTING
AIR FLOW
HUMIDITY & CONDENSATION
DEW CONDENSATION
WATER ACCESS
CANOPY'S SHADE
3 m
4 m
4 m
4 m
3 m

Warka Water

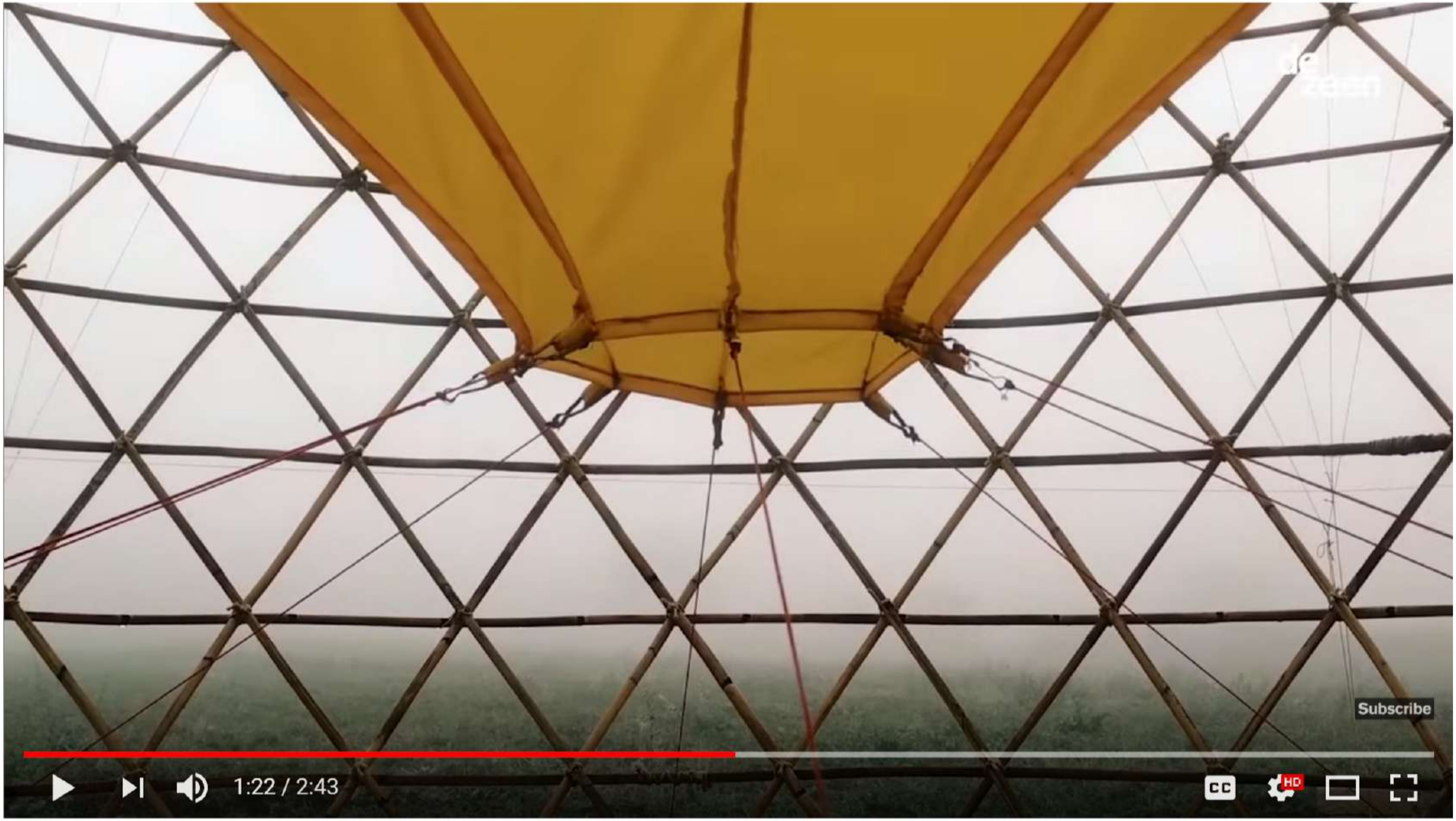
dezeen
Subscribe
1:22 / 2:43
CC
HD

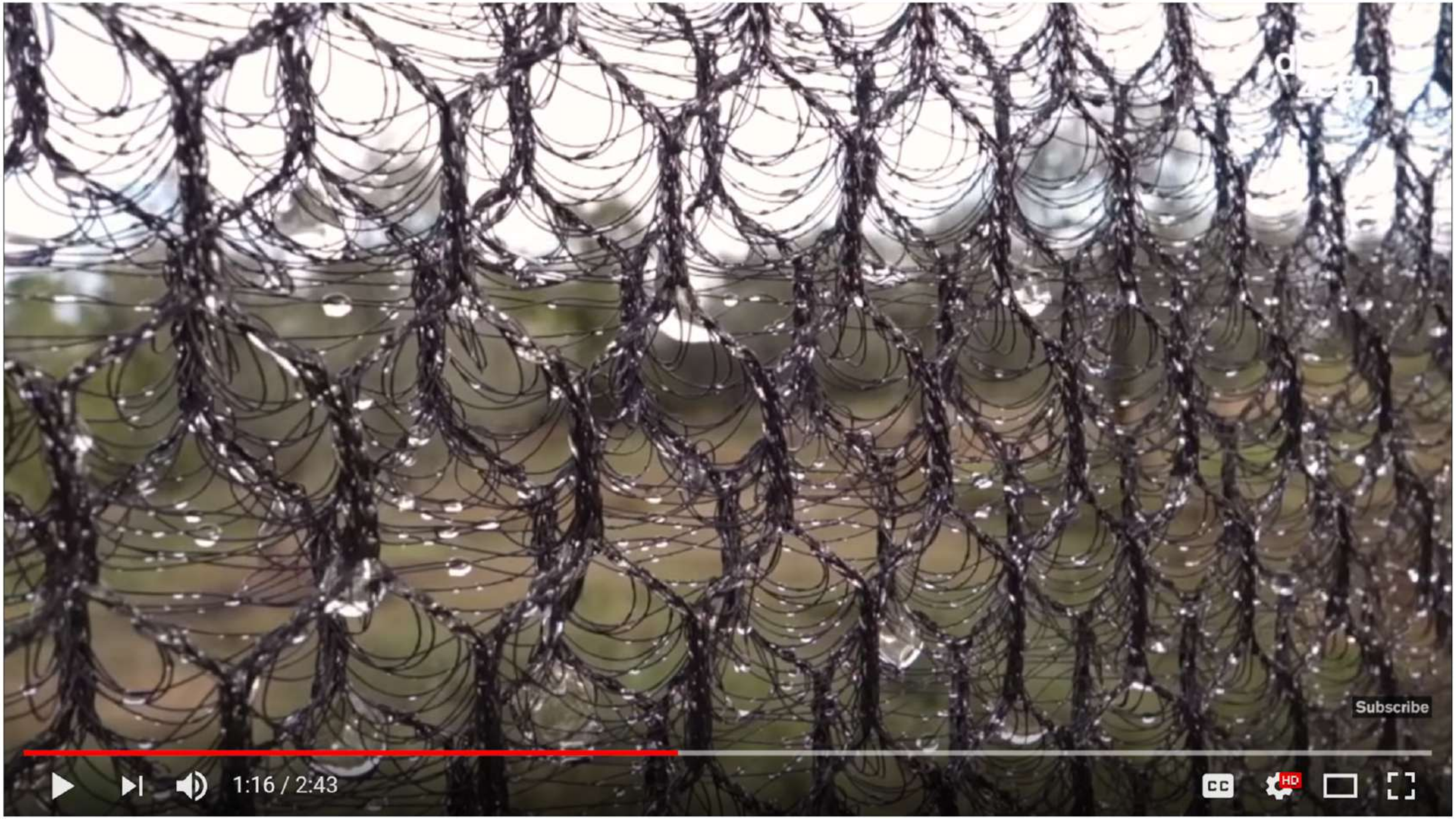

Subscribe
1:16 / 2:43
HD

Warka Water towers harvest drinkable water from the air

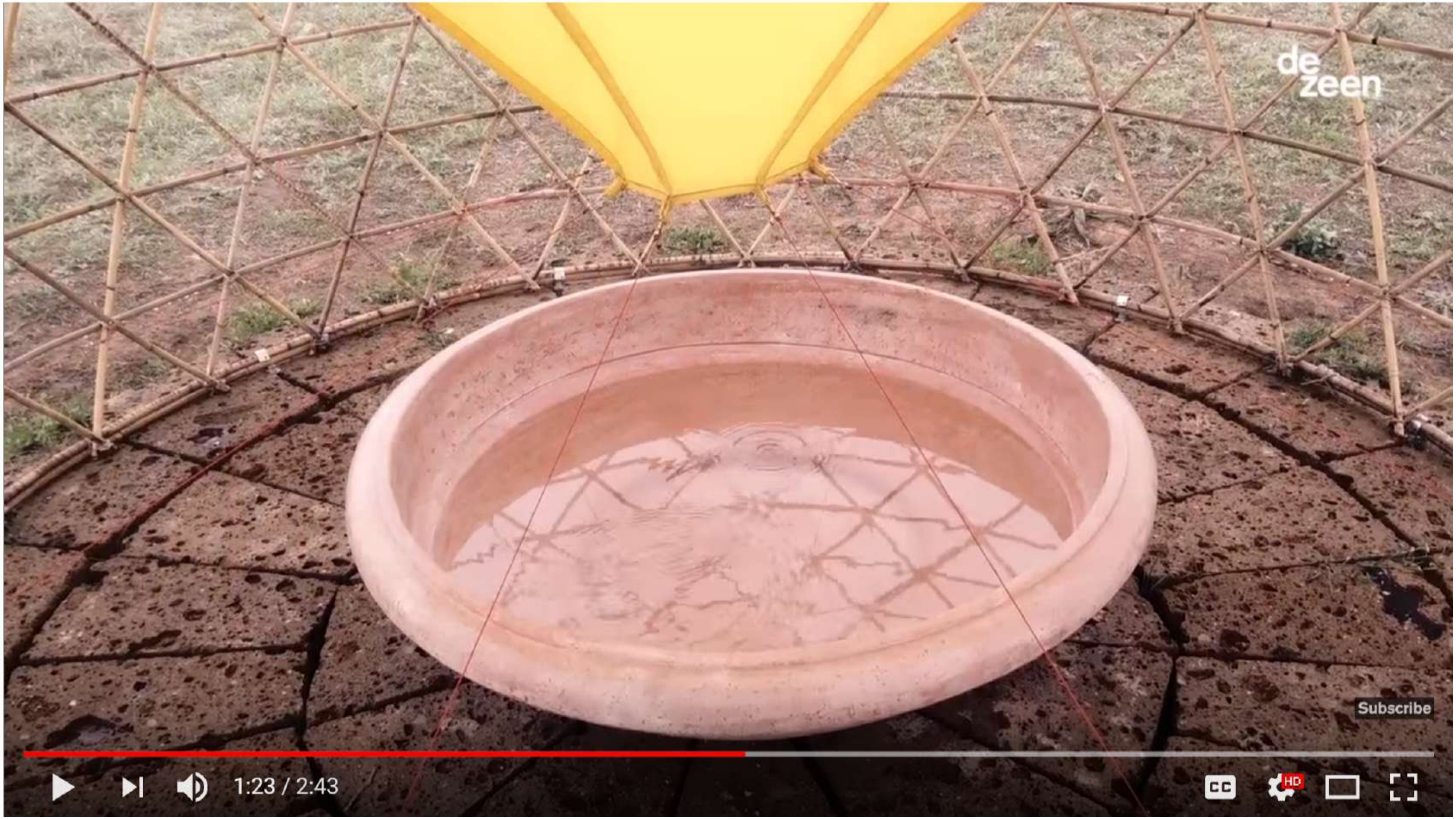
dezeen
Subscribe
1:23 / 2:43
HD

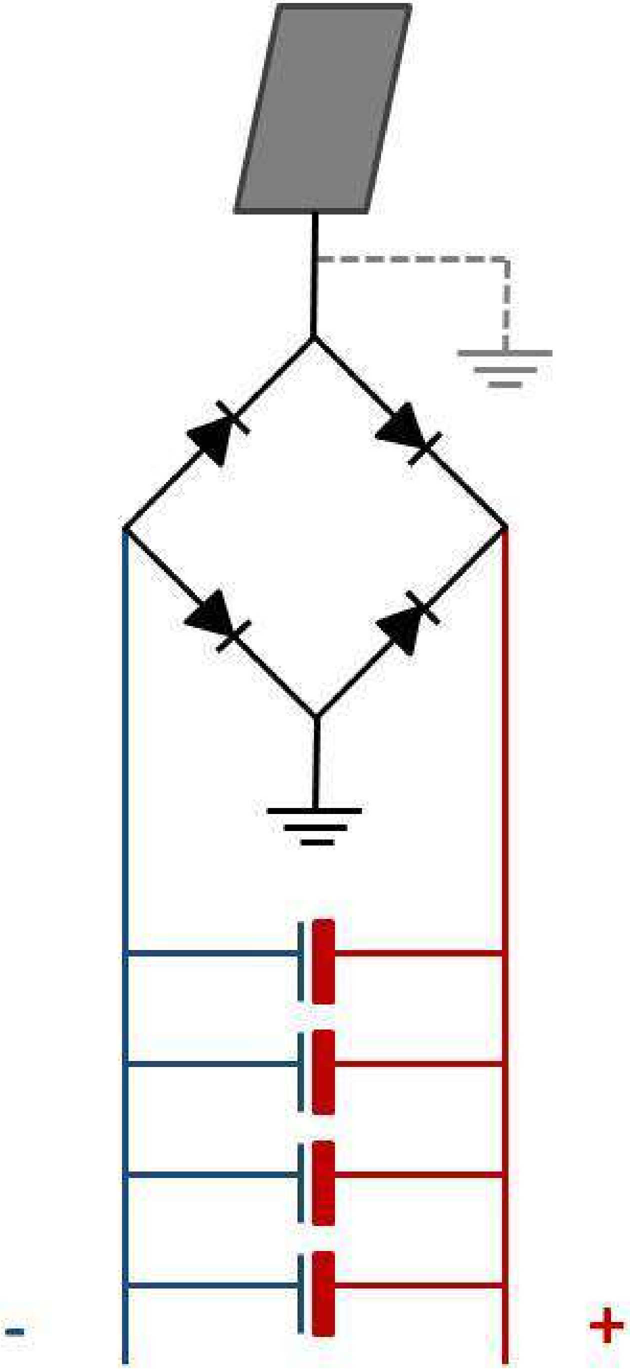

Insulated, polished aluminium plate high up in air

An extra direct earth connection might help

1N34a germanium diodes as full-wave bridge rectifier

4mm high load single core copper wire

Earth is copper pipe 2 metres deep in moist soil

100uF 50V electrolytic capacitors in parallel

IMPROVEMENT OF RECEIVER:

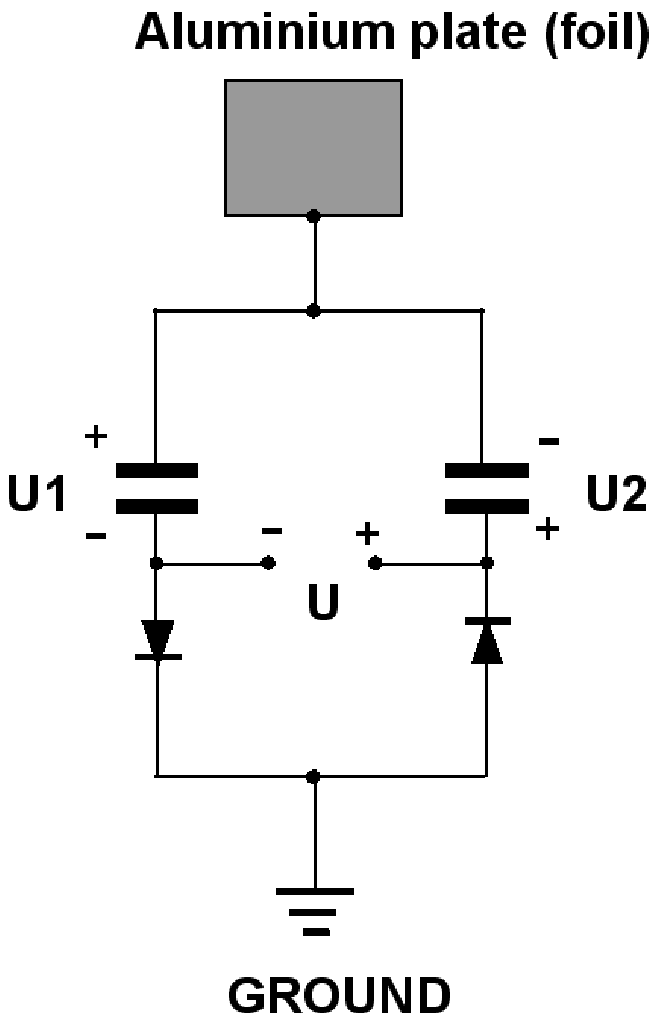
Aluminium plate (foil)
+
U1
-
-
U
+
-
U2
+
GROUND

$$U = U1 + U2$$

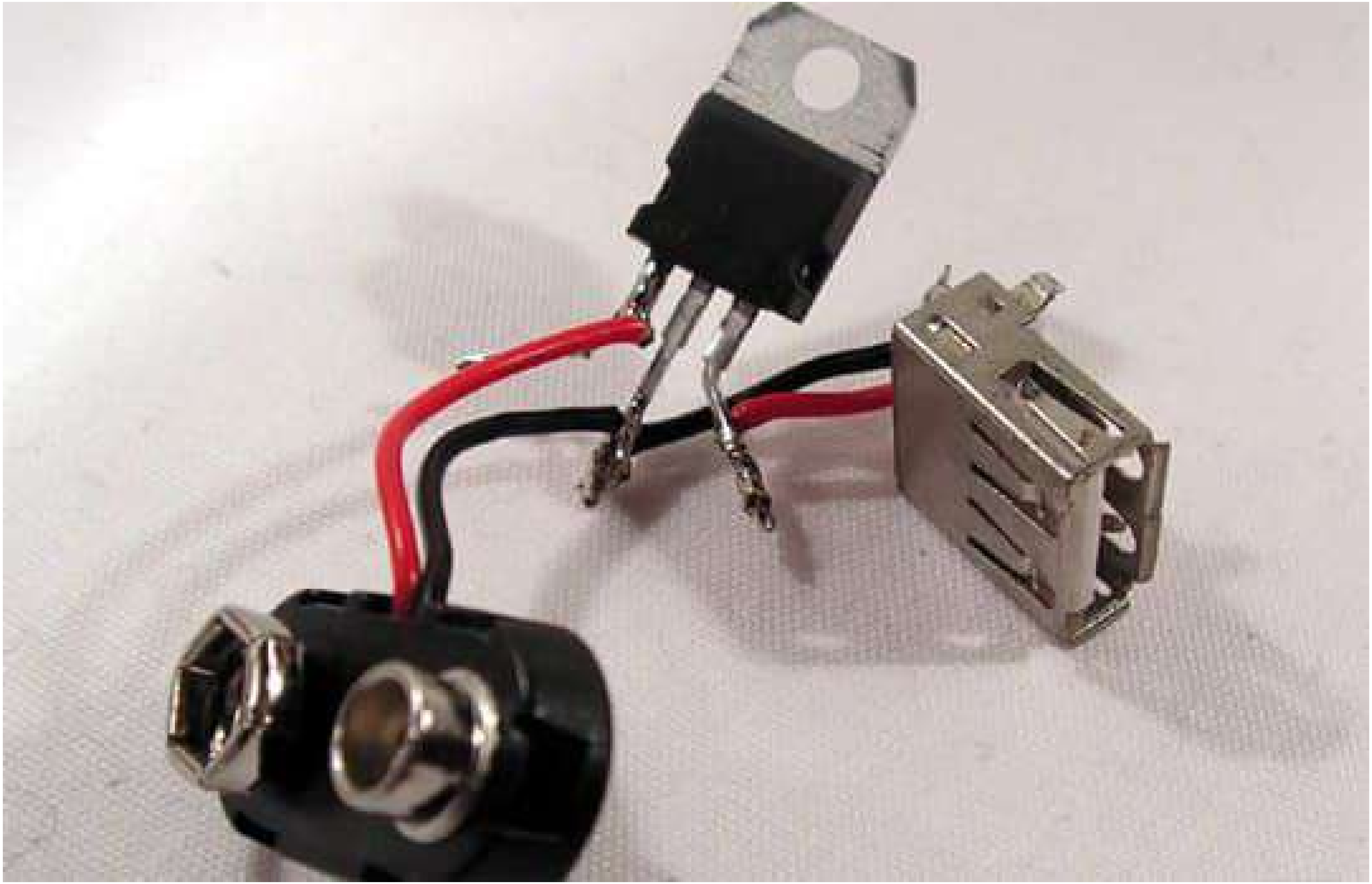

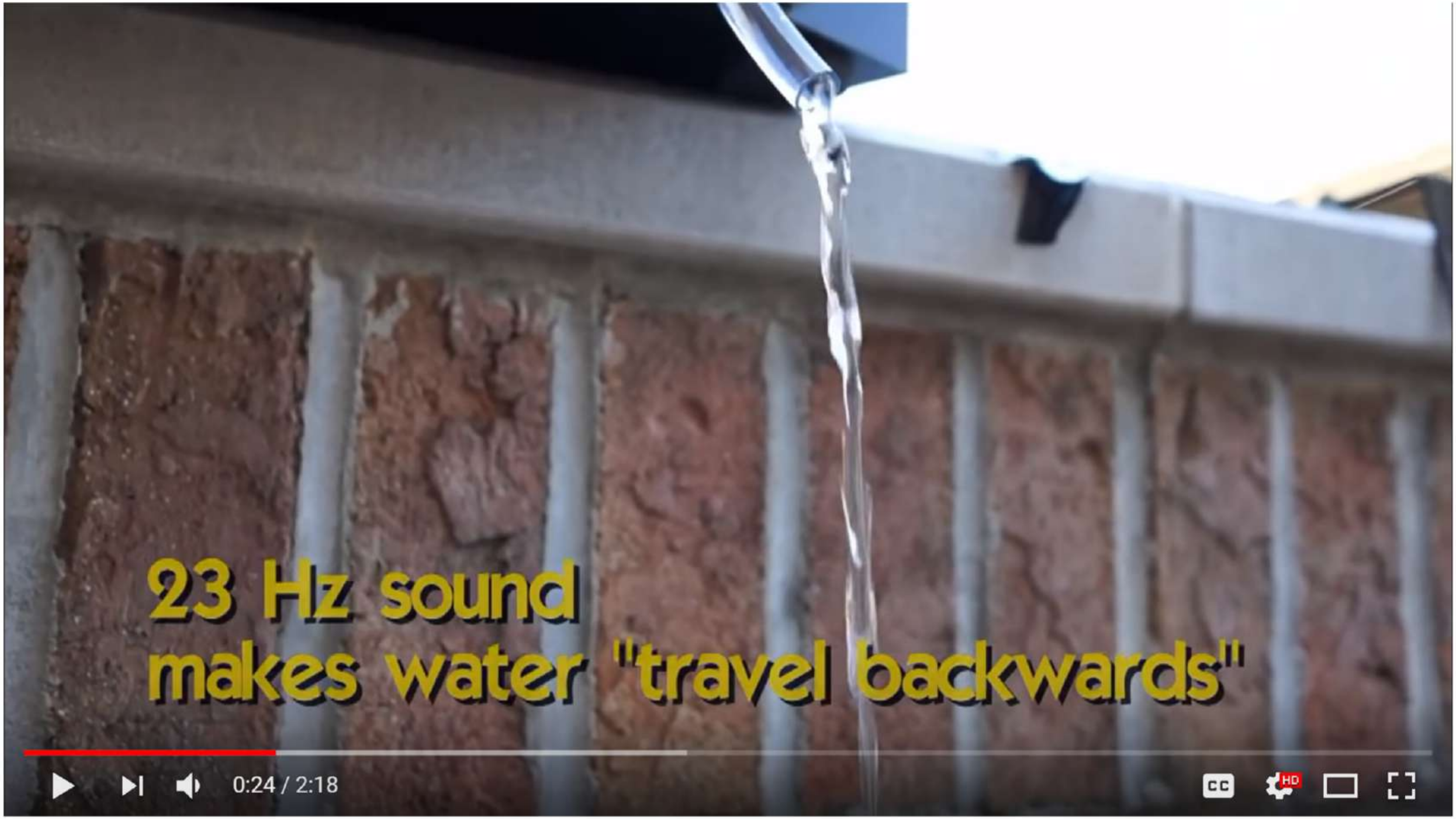
23 Hz sound
makes water "travel backwards"
0:24 / 2:18
HD

Make Water Appear Frozen In Time Using Sound

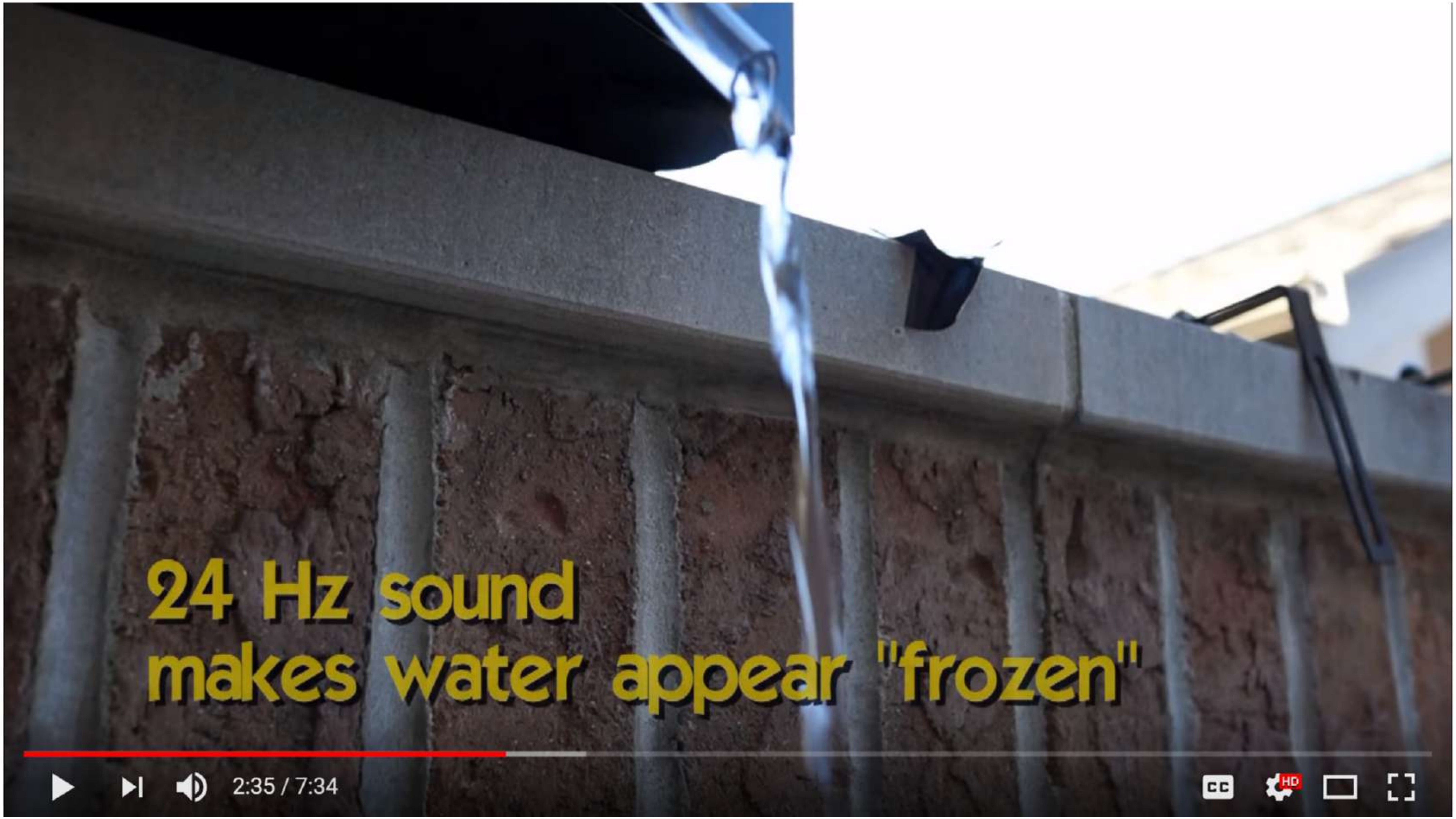

24 Hz sound
makes water appear "frozen"

2:35 / 7:34
CC
HD

f1=24.0 Hz
f2=23.0 Hz

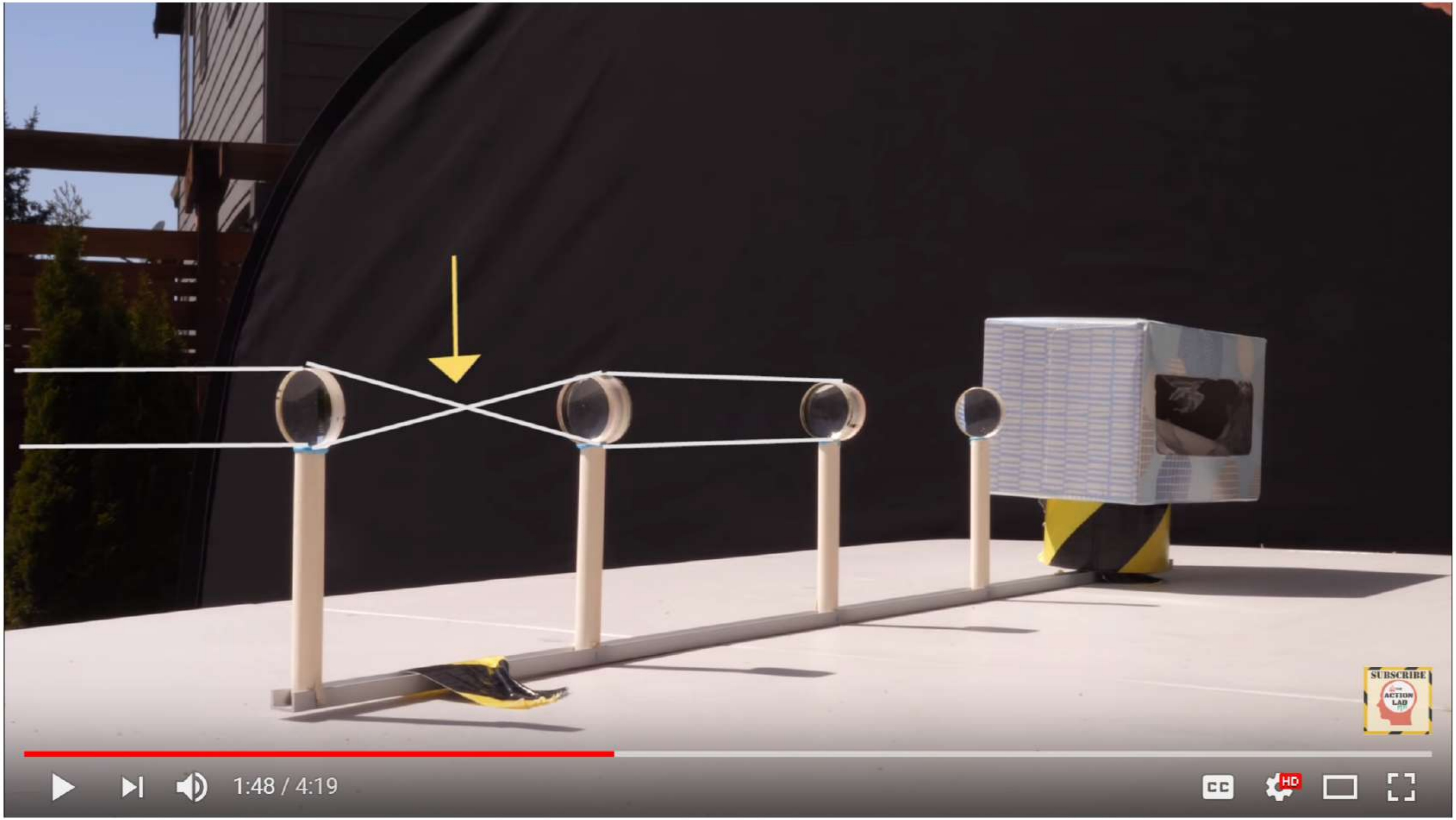
SUBSCRIBE
ACTION LAB
1:48 / 4:19
CC
HD

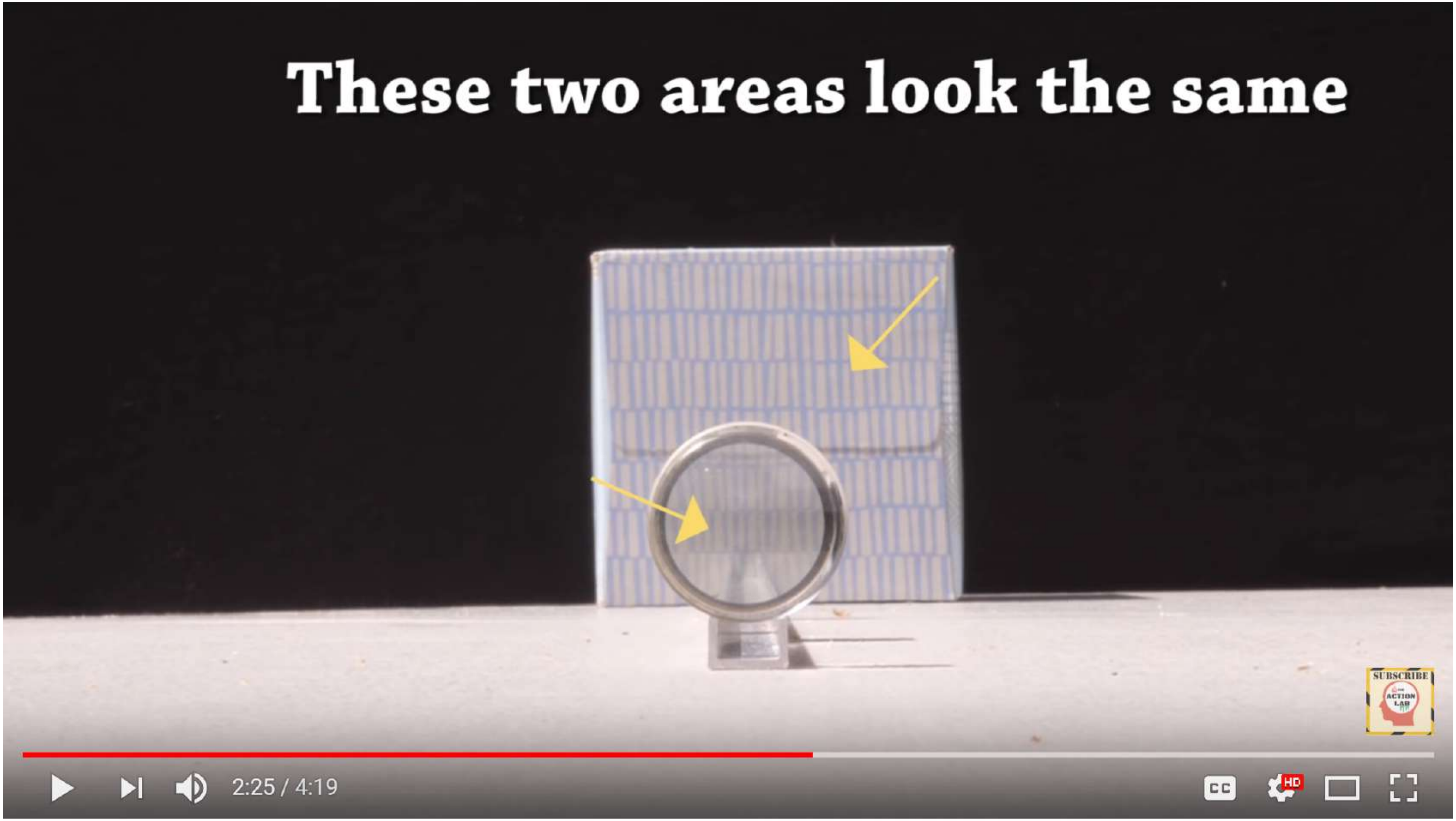

These two areas look the same

SUBSCRIBE
THE ACTION LAB
2:25 / 4:19
CC
HD

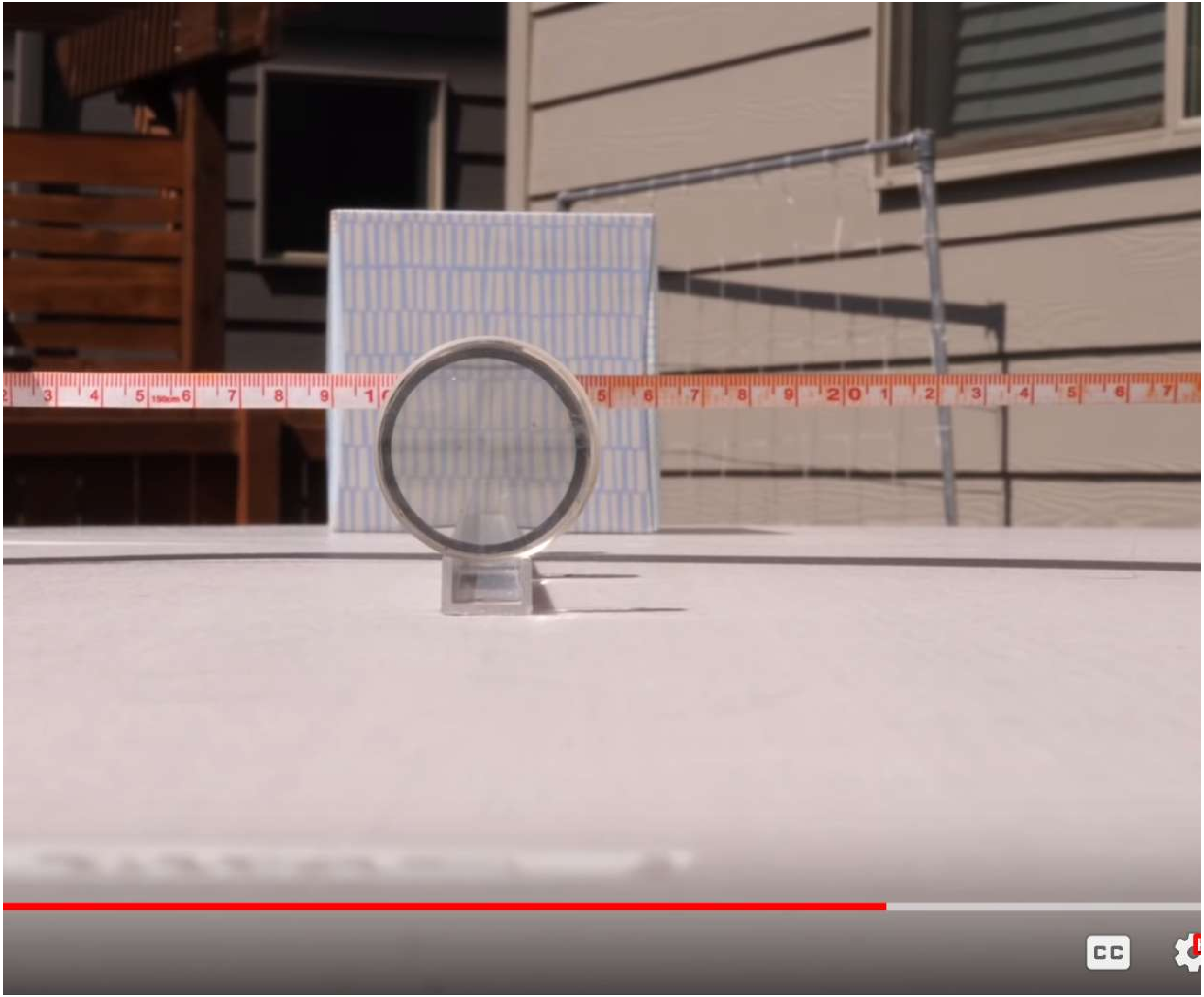
3 4 5 150cm 6 7 8 9 1
5 6 7 8 9 20 1 2 3 4 5 6 7
CC

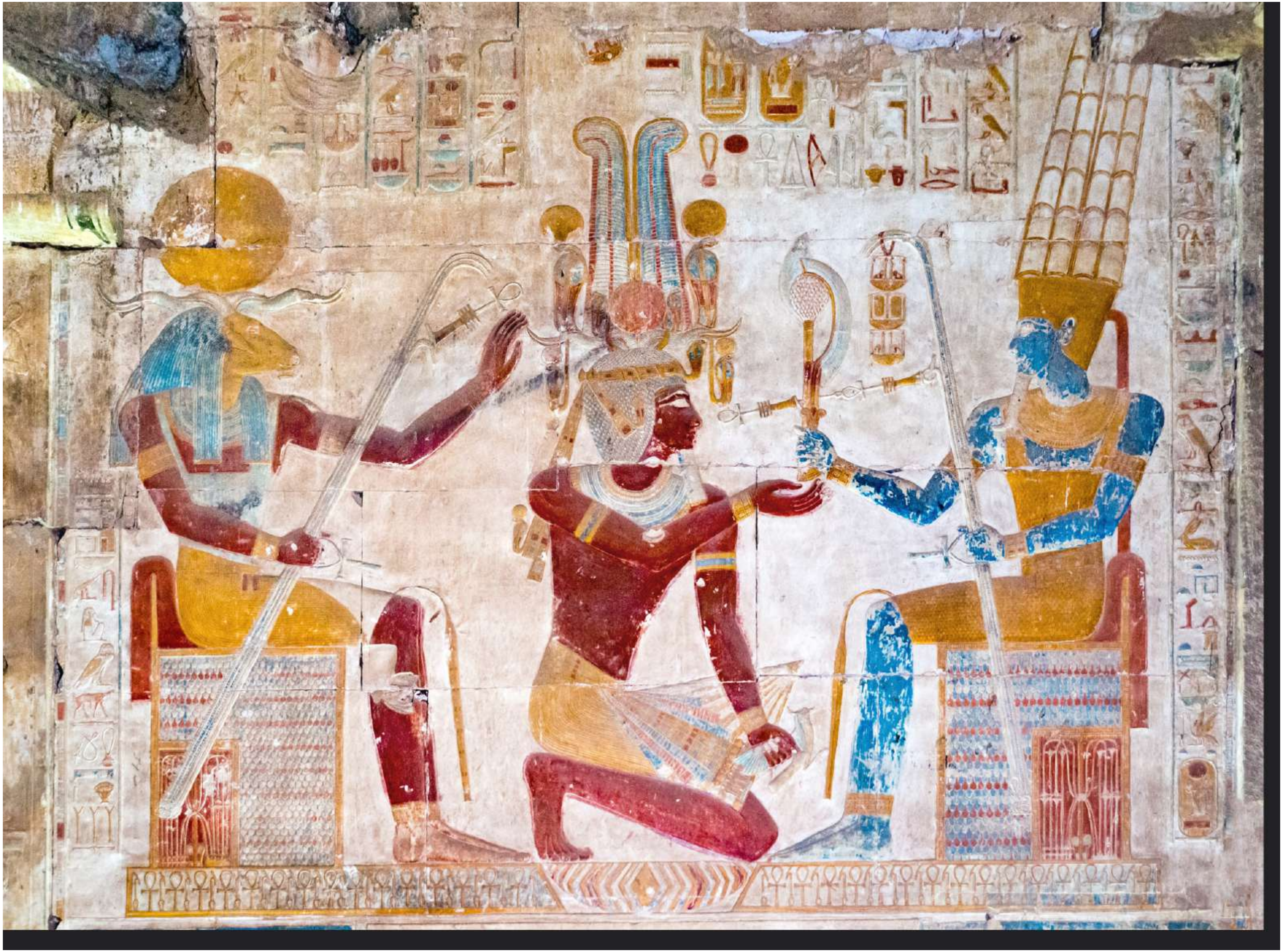

1. *Priest*
2. *ionized fumes*
3. *electric discharge (snake)*
4. *Lamp socket (Lotos)*
5. *Cable (Lotos stem)*
6. *Air god*
7. *Isolator (Djed-Pillar)*
8. *Light bringer Thot with knifes*
9. *Symbol for "current"*
10. *Inverse polarity (Haarpolarität +)*
11. *Energy storage (electrostatic Generator?)*

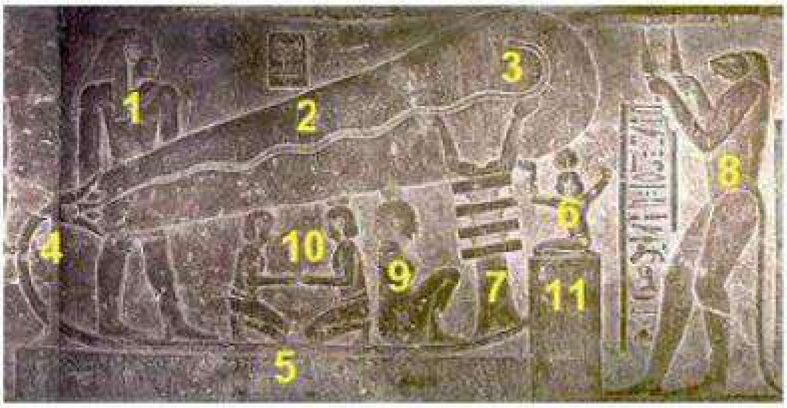

http://www.facebook.com/[illegible]

http://www.facebook.com/UfoKruhyV0bili

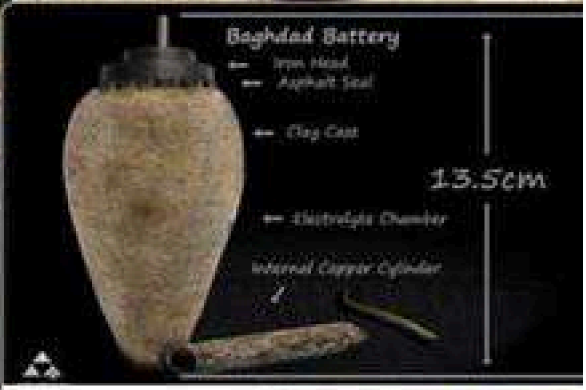
Baghdad Battery
Iron Head
Asphalt Seal
Clay Case
13.5cm
Electrolyte Chamber
Internal Copper Cylinder

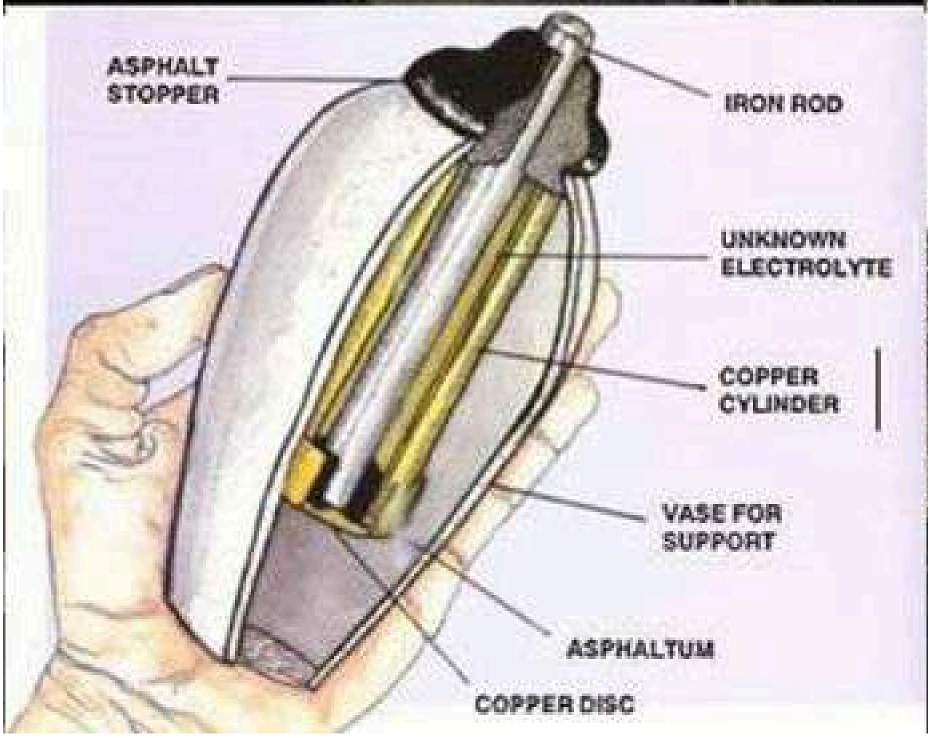
ASPHALT STOPPER
IRON ROD
UNKNOWN ELECTROLYTE
COPPER CYLINDER
VASE FOR SUPPORT
ASPHALTUM
COPPER DISC

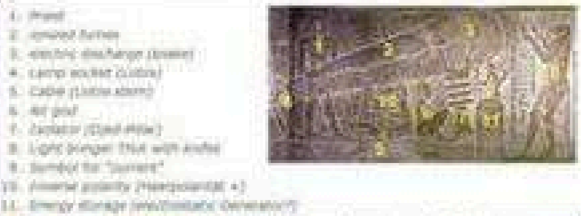
1. Priest
2. ionized fumes
3. electric discharge (snake)
4. Lamp socket (Lotos)
5. Cable (Lotos stem)
6. Air god
7. Isolator (Djed-Pillar)
8. Light bringer Thot with knifes
9. Symbol for "current"
10. inverse polarity (Heteropolarität +)
11. Energy storage (electrostatic Generator?)

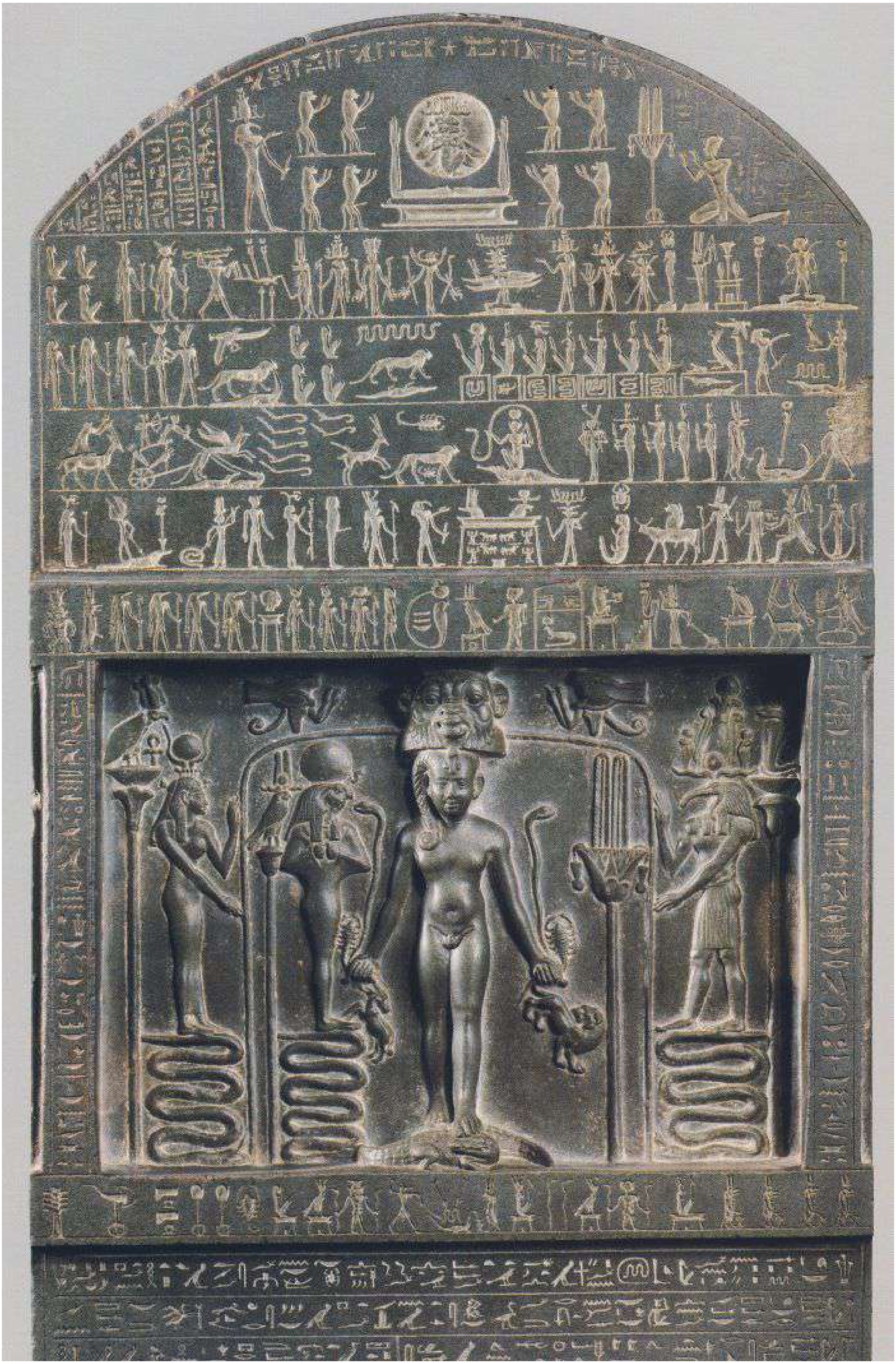

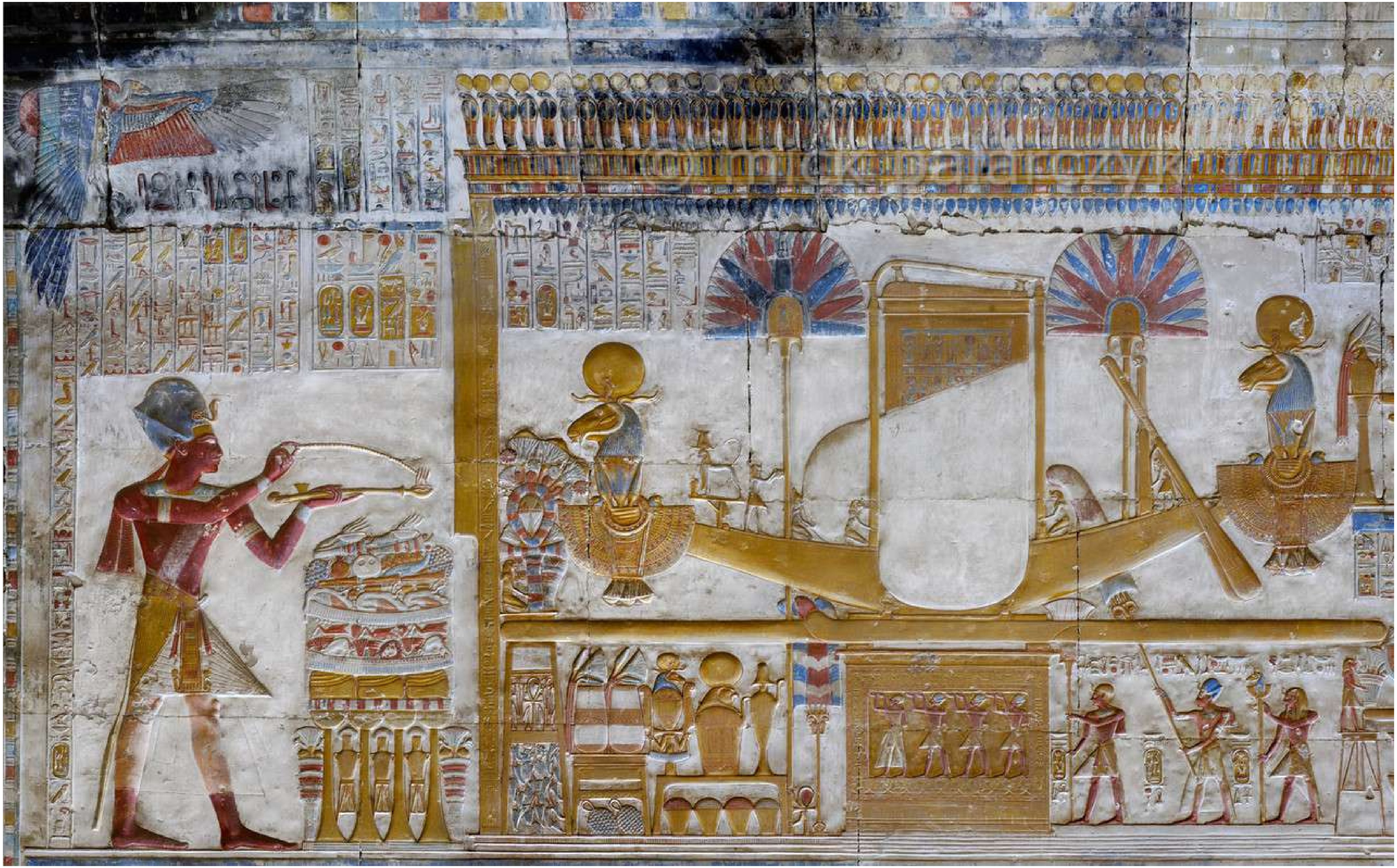

© mickypalarczyk

© michal pelarczyk

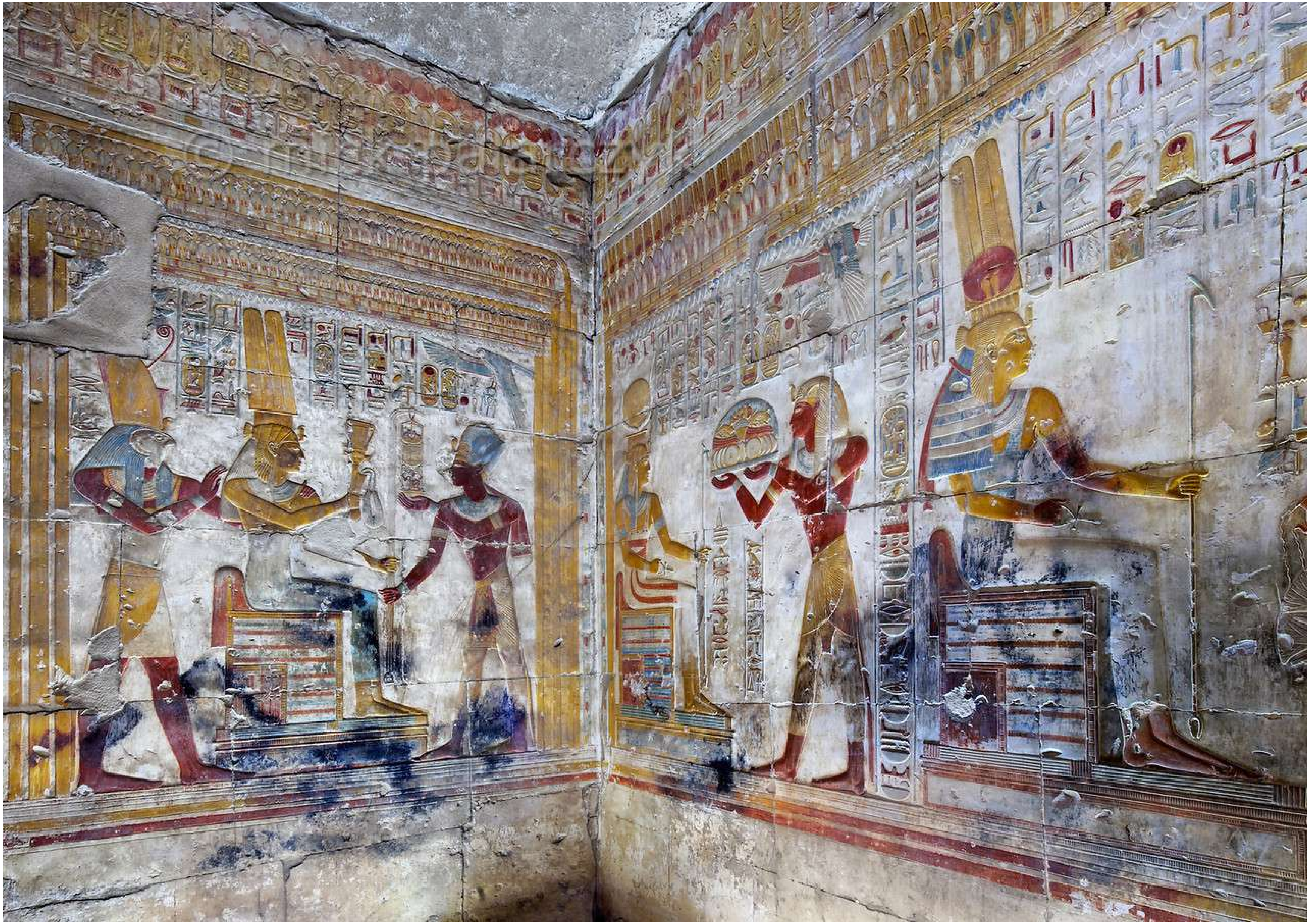

PETRIFIED FOREST
NATIONAL PARK
NATIONAL PARK SERVICE

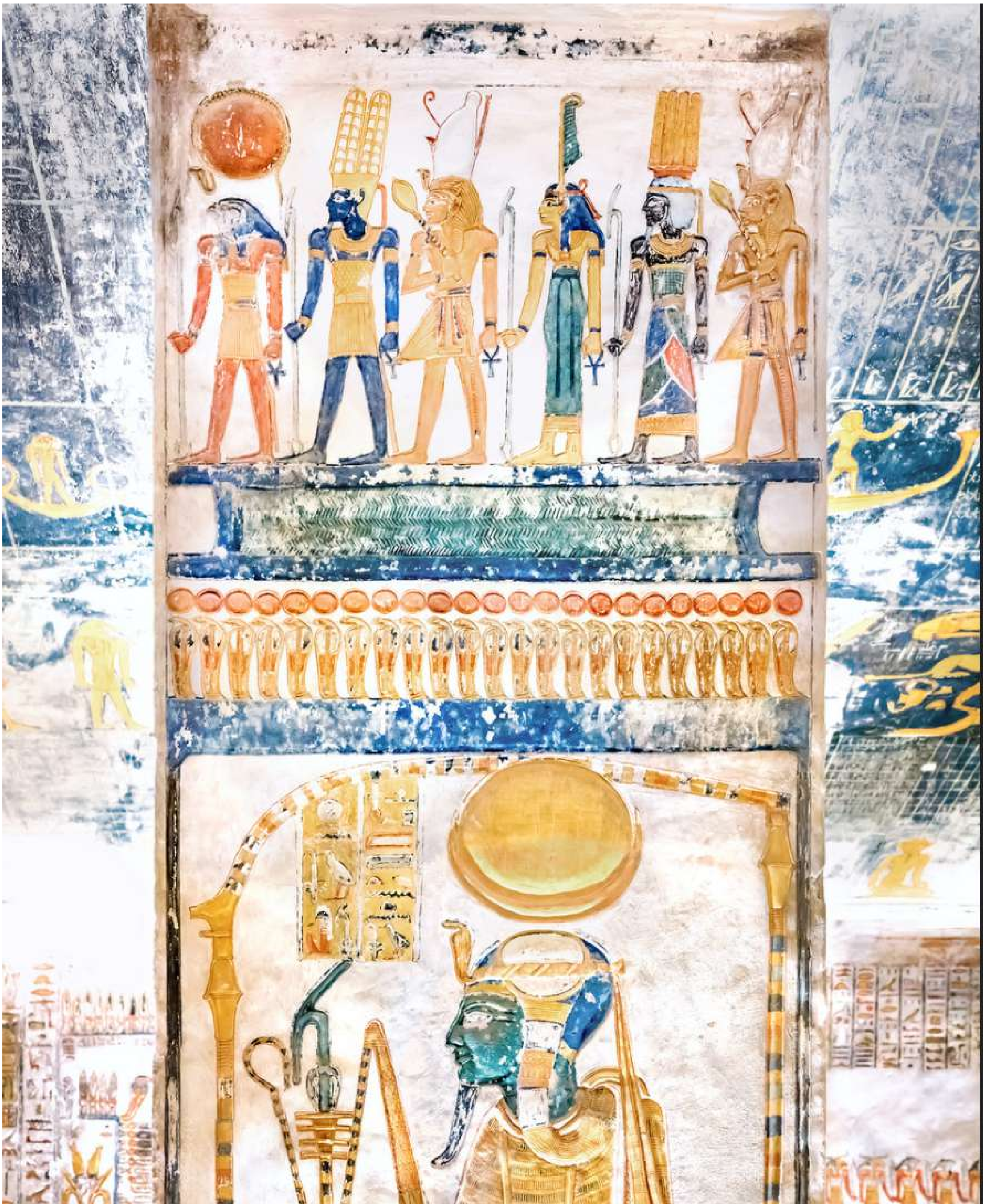

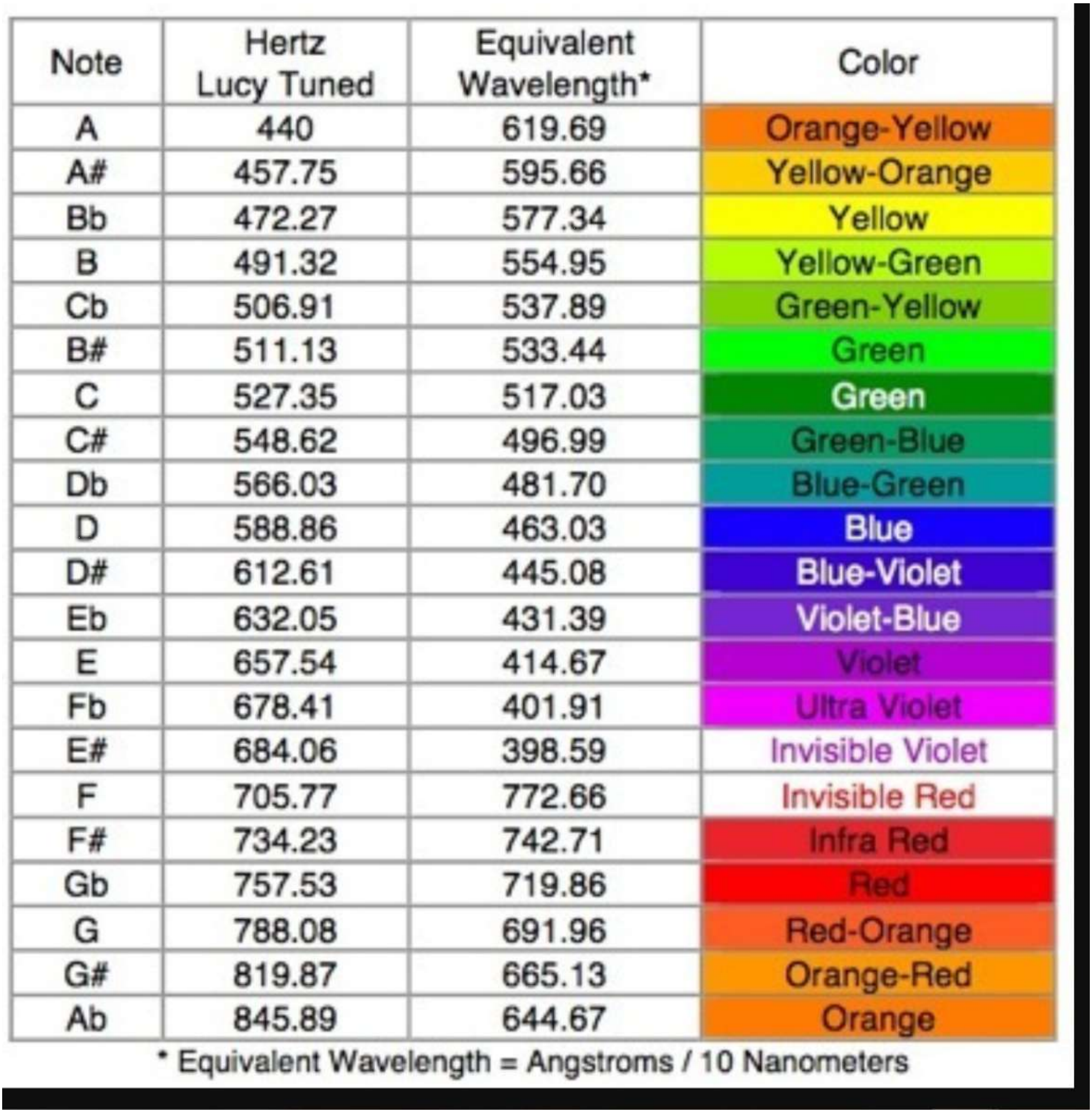

Note	Hertz Lucy Tuned	Equivalent Wavelength*	Color
A	440	619.69	Orange-Yellow
A#	457.75	595.66	Yellow-Orange
Bb	472.27	577.34	Yellow
B	491.32	554.95	Yellow-Green
Cb	506.91	537.89	Green-Yellow
B#	511.13	533.44	Green
C	527.35	517.03	Green
C#	548.62	496.99	Green-Blue
Db	566.03	481.70	Blue-Green
D	588.86	463.03	Blue
D#	612.61	445.08	Blue-Violet
Eb	632.05	431.39	Violet-Blue
E	657.54	414.67	Violet
Fb	678.41	401.91	Ultra Violet
E#	684.06	398.59	Invisible Violet
F	705.77	772.66	Invisible Red
F#	734.23	742.71	Infra Red
Gb	757.53	719.86	Red
G	788.08	691.96	Red-Orange
G#	819.87	665.13	Orange-Red
Ab	845.89	644.67	Orange

* Equivalent Wavelength = Angstroms / 10 Nanometers

Musical Notes-Color Correspondences in the Visible Spectrum of Light

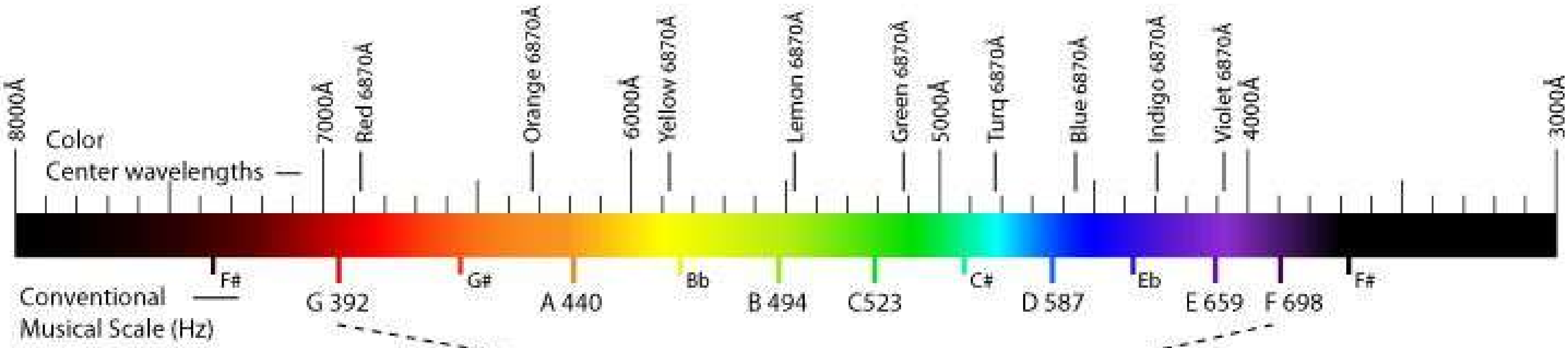

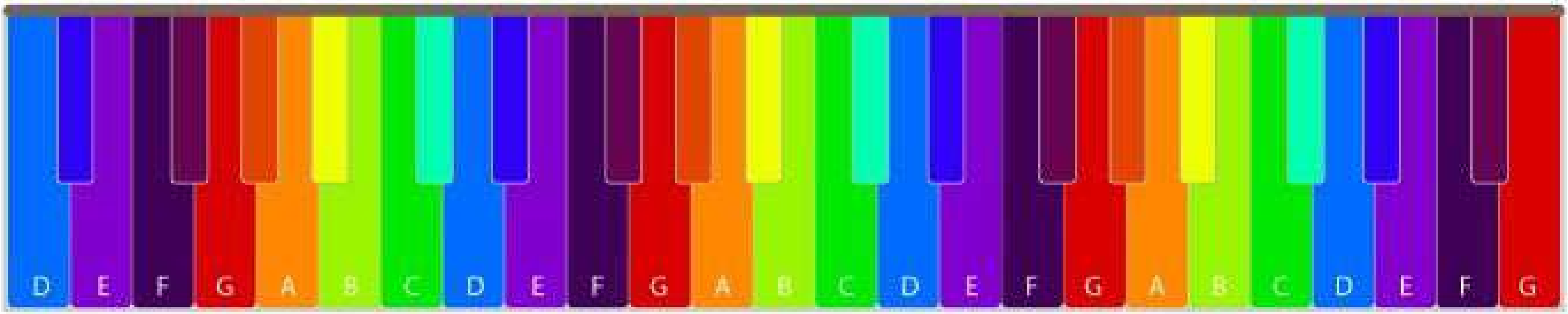
D
E
F
G
A
B
C
D
E
F
G
A
B
C
D
E
F
G
A
B
C
D
E
F
G

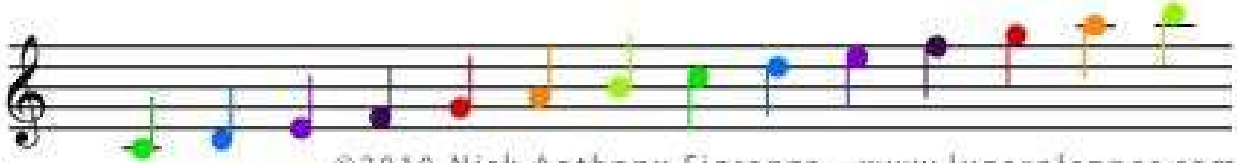

©2010 Nick Anthony Fiorenza - www.lunarplanner.com

Guitar Fret Board (Standard Tuning)

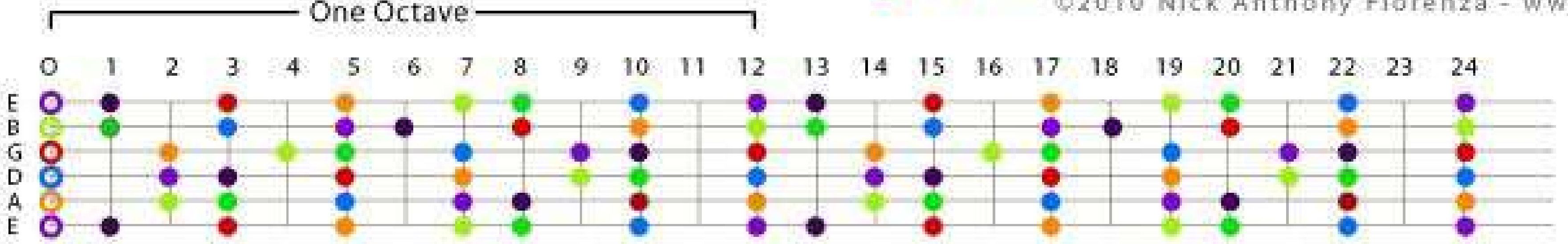
One Octave
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
E
B
G
D
A
E

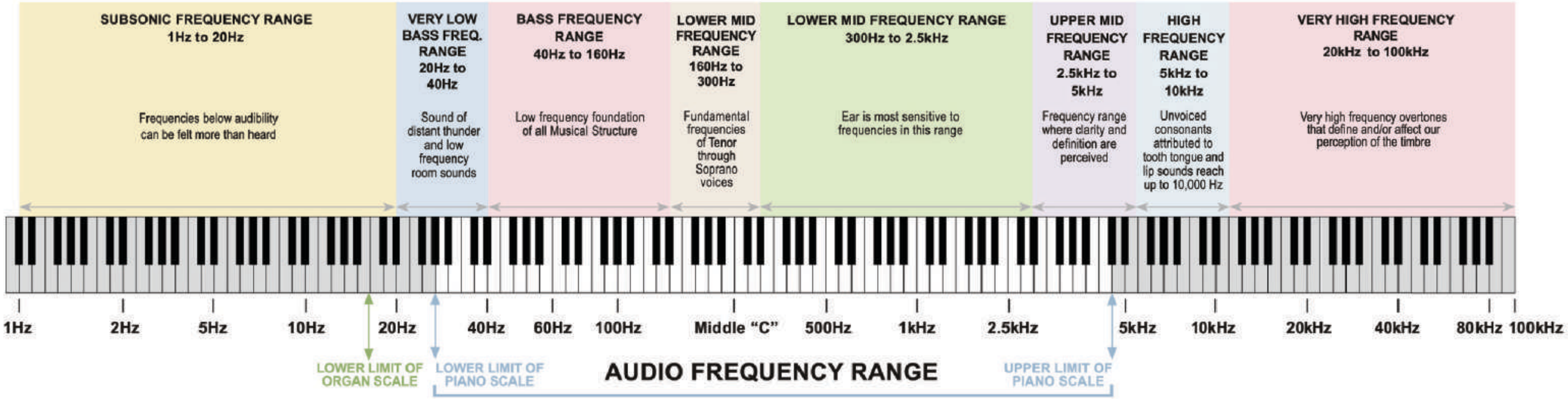

SUBSONIC FREQUENCY RANGE
1Hz to 20Hz

Frequencies below audibility can be felt more than heard

VERY LOW BASS FREQ. RANGE
20Hz to 40Hz

Sound of distant thunder and low frequency room sounds

BASS FREQUENCY RANGE
40Hz to 160Hz

Low frequency foundation of all Musical Structure

LOWER MID FREQUENCY RANGE
160Hz to 300Hz

Fundamental frequencies of Tenor through Soprano voices

LOWER MID FREQUENCY RANGE
300Hz to 2.5kHz

Ear is most sensitive to frequencies in this range

UPPER MID FREQUENCY RANGE
2.5kHz to 5kHz

Frequency range where clarity and definition are perceived

HIGH FREQUENCY RANGE
5kHz to 10kHz

Unvoiced consonants attributed to tooth tongue and lip sounds reach up to 10,000 Hz

VERY HIGH FREQUENCY RANGE
20kHz to 100kHz

Very high frequency overtones that define and/or affect our perception of the timbre

1Hz
2Hz
5Hz
10Hz
20Hz
40Hz
60Hz
100Hz
Middle "C"
500Hz
1kHz
2.5kHz
5kHz
10kHz
20kHz
40kHz
80kHz
100kHz
LOWER LIMIT OF ORGAN SCALE
LOWER LIMIT OF PIANO SCALE
AUDIO FREQUENCY RANGE
UPPER LIMIT OF PIANO SCALE

THE COAST OF ANTARCTICA, AND THE ANTARCTIC SHELF.
THIS IS THE ICE WALL, THE CONTAINER OF THE EARTH,
A GIANT LAND MASS SURROUNDING THE EARTH.
WATER ALWAYS LEVELS OUT AND ALWAYS TAKES
THE SHAPE OF ITS CONTAINER. THERE IS NO EDGE.

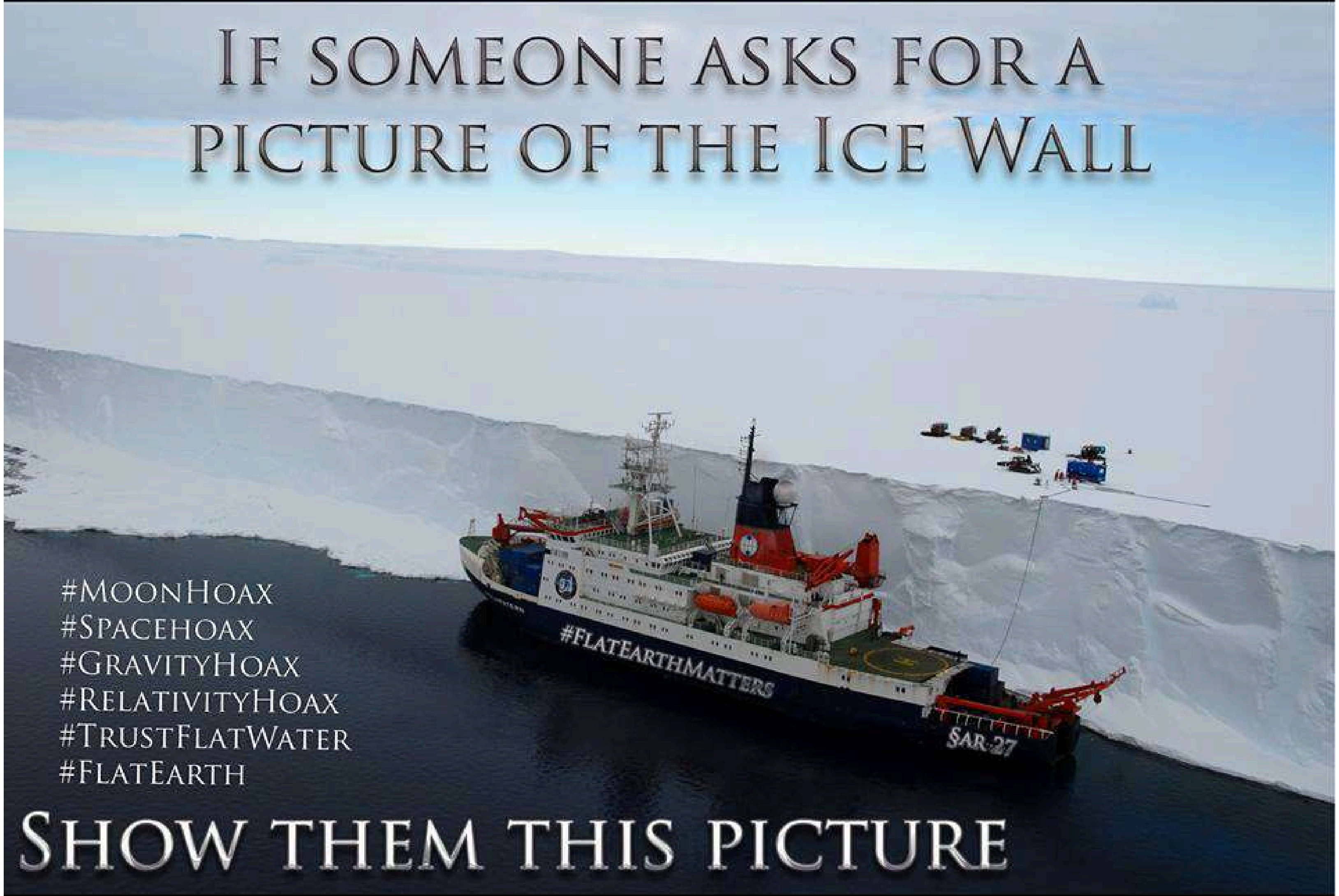
IF SOMEONE ASKS FOR A
PICTURE OF THE ICE WALL
#MOONHOAX
#SPACEHOAX
#GRAVITYHOAX
#RELATIVITYHOAX
#TRUSTFLATWATER
#FLATEARTH
#FLATEARTHMATTERS
SAR-27
SHOW THEM THIS PICTURE

FISH TANKS, TUBS, POOLS & LAKES OF WATER ARE FLAT AND LEVEL.

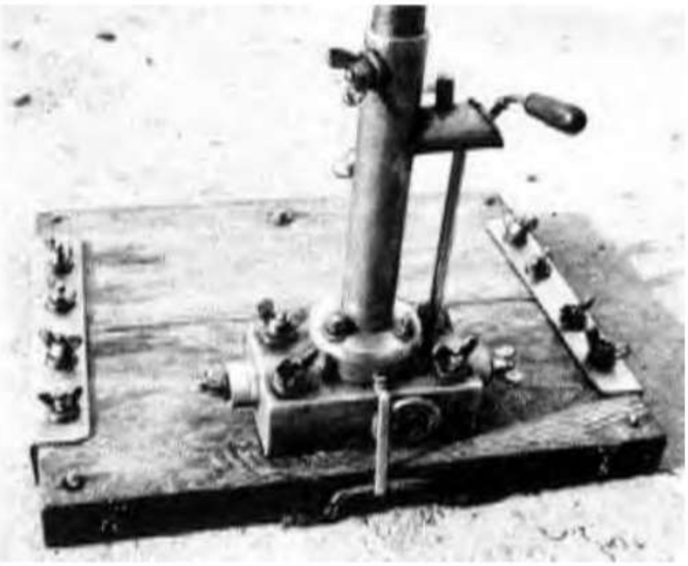

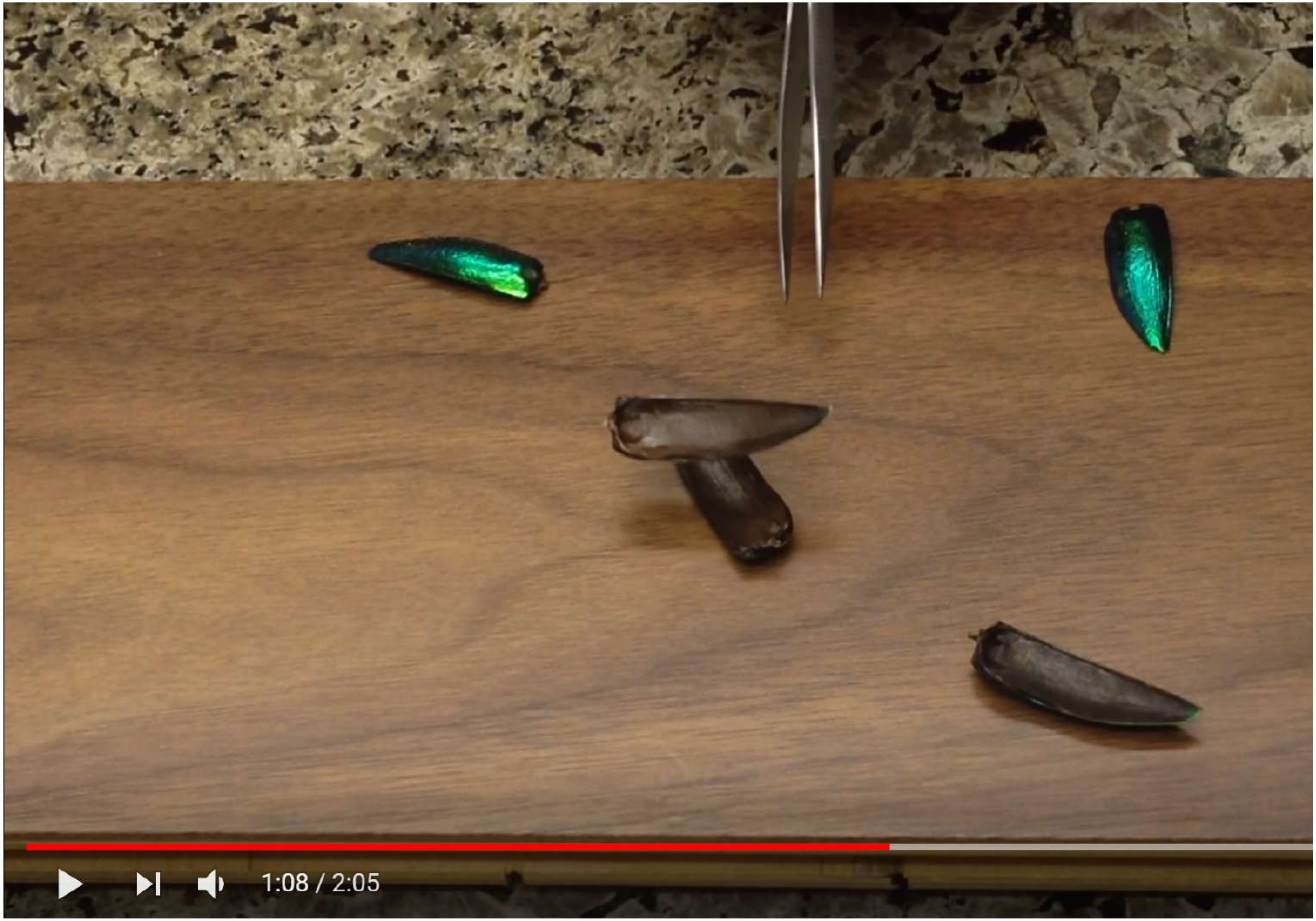

Hovering Elytra - Grebennikov Video Replication Series

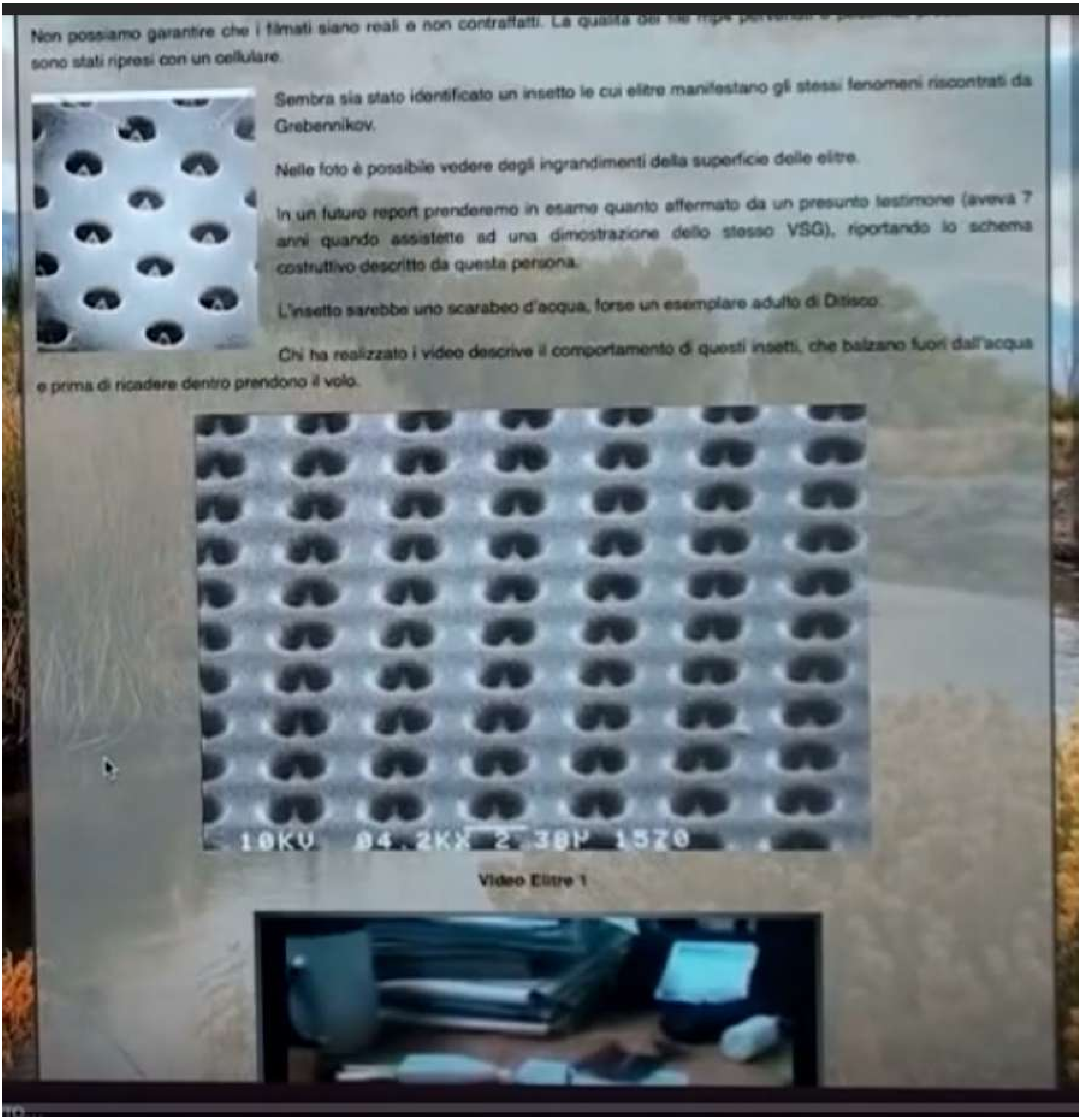

Non possiamo garantire che i filmati siano reali e non contraffatti. La qualità dei file

sono stati ripresi con un cellulare.

Sembra sia stato identificato un insetto le cui elitre manifestano gli stessi fenomeni riscontrati da Grebennikov.

Nelle foto è possibile vedere degli ingrandimenti della superficie delle elitre.

In un futuro report prenderemo in esame quanto affermato da un presunto testimone (aveva 7 anni quando assistette ad una dimostrazione dello stesso VSG), riportando lo schema costruttivo descritto da questa persona.

L'insetto sarebbe uno scarabeo d'acqua, forse un esemplare adulto di Ditisco.

Chi ha realizzato i video descrive il comportamento di questi insetti, che balzano fuori dall'acqua e prima di ricadere dentro prendono il volo.

Video Elitre 1

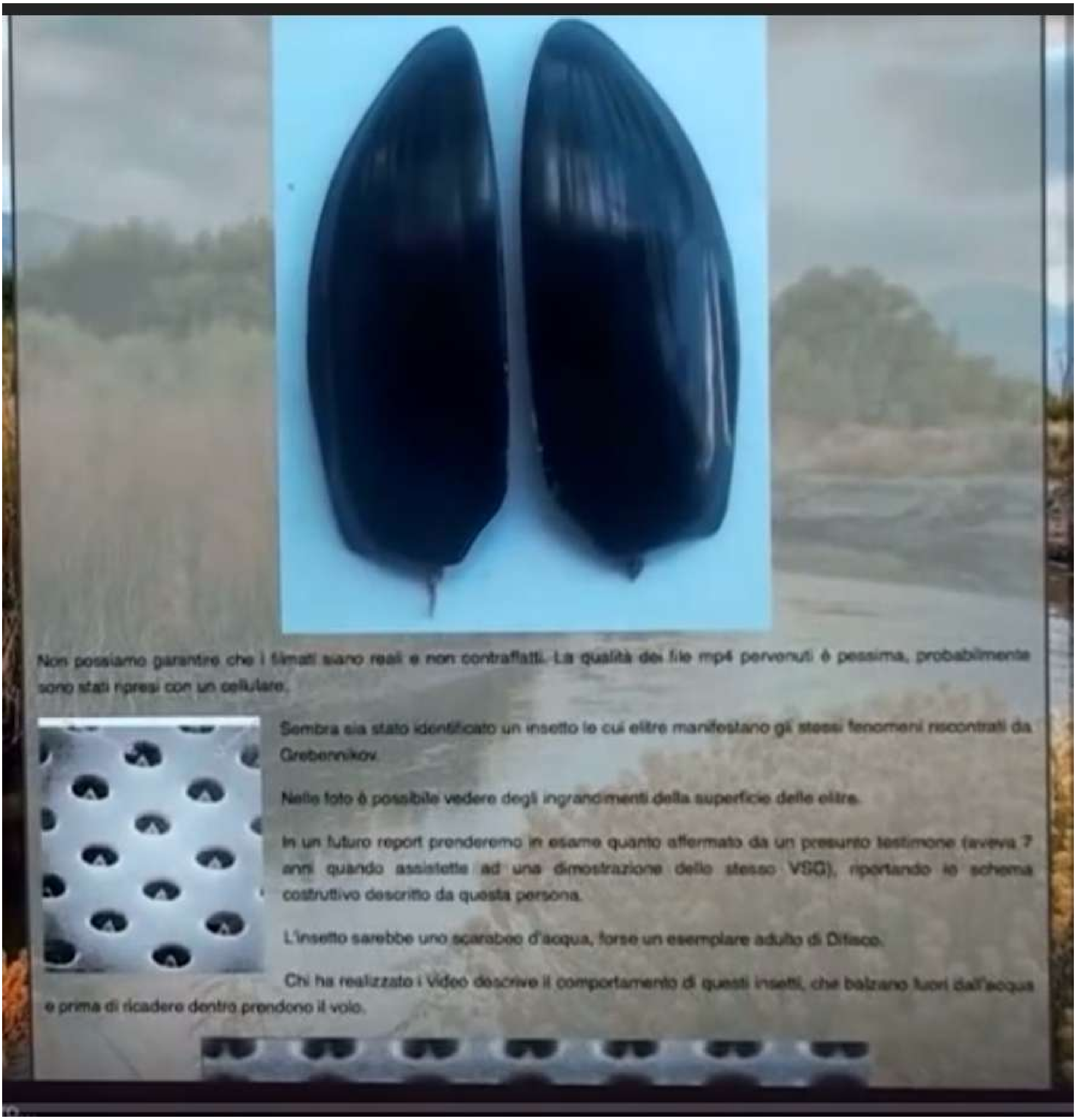

Non possiamo garantire che i filmati siano reali e non contraffatti. La qualità dei file mp4 pervenuti è pessima, probabilmente sono stati ripresi con un cellulare.

Sembra sia stato identificato un insetto le cui elitre manifestano gli stessi fenomeni riscontrati da Grebennikov.

Nelle foto è possibile vedere degli ingrandimenti della superficie delle elitre.

In un futuro report prenderemo in esame quanto affermato da un presunto testimone (aveva 7 anni quando assistette ad una dimostrazione dello stesso VSG), riportando lo schema costruttivo descritto da questa persona.

L'insetto sarebbe uno scarabeo d'acqua, forse un esemplare adulto di Ditisco.

Chi ha realizzato i Video descrive il comportamento di questi insetti, che balzano fuori dall'acqua e prima di ricadere dentro prendono il volo.

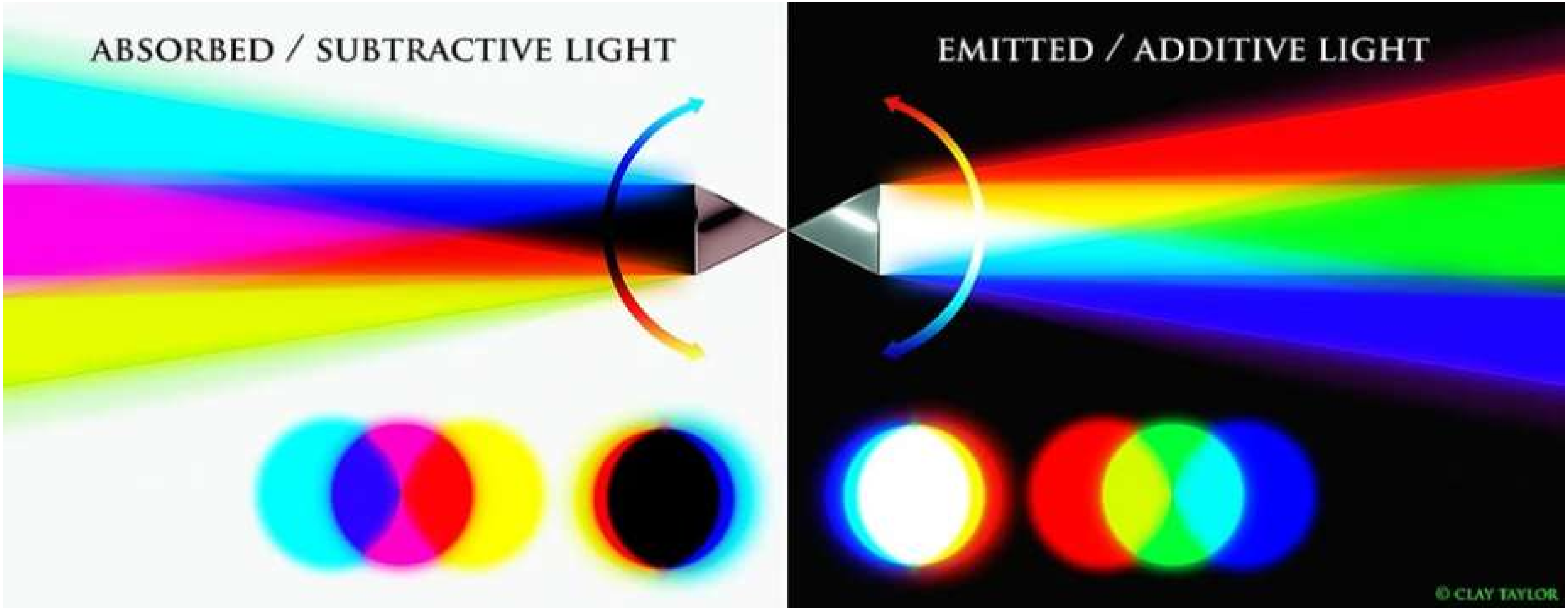
ABSORBED / SUBTRACTIVE LIGHT
EMITTED / ADDITIVE LIGHT
© CLAY TAYLOR

Free Energy From Air Circuit

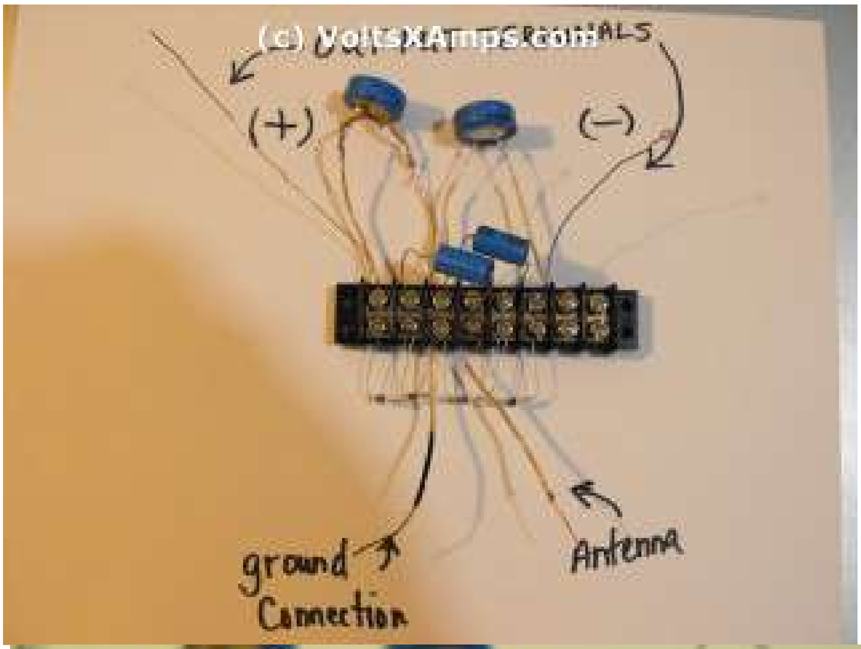
(c) VoltsXAmps.com
(+)
(−)
ground Connection
Antenna

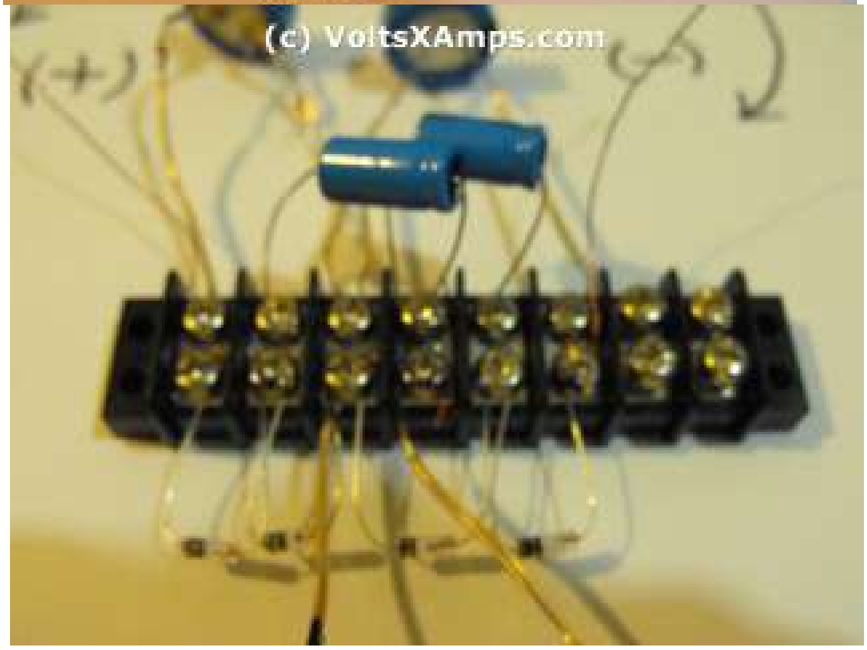
(c) VoltsXAmps.com
(+)
(−)

Last year I found a United States Patent that showed how to collect ambient energy right from the air. I finally decided to build this curcuit just to see what it could do.

[protected]

The above images are from the circuit I built. Most of the text and info comes direct from the US4628299filed by Joseph Tate. US4628299

The Amazing Ambient Power Module

Parts List for the APM-2

Four 1N34 germanium diodes (Radio shack #276-1123) ~ Figure 1, X1, X2, X3, & X4
Two 0.2 mfd 50 V ceramic capacitors ~ Figure 1, C1 & C2
Two 100 mfd 50V electrolytic capacitors (Radio Shack #272-1016) ~ Figure 1, C3 & C4
Copper wire for antenna & ground connections

Introduction

The Ambient Power Module (APM) is a simple electronic circuit which, when connected to antenna and earth ground, will deliver low voltage up to several milliwatts. The amount of voltage and power will be determined by local radio noise levels and antenna dimensions

Generally a long wire antenna about 100' long and elevated in a horizontal position about 30' above ground works best. A longer antenna may be required in some locations. Any type copper wire, insulated or not, may be used for the antenna. More details about the antenna and ground will be discussed further on.

The actual circuit consists of two oppositely polarized voltage doublers (Figure 1). The DC output of each doubler is connected in series with the other to maximize voltage without using transformers. Single voltage doublers were often found in older TV sets for converting 120 VAC to 240 VDC. In the TV circuit the operating frequency is 60 Hz.

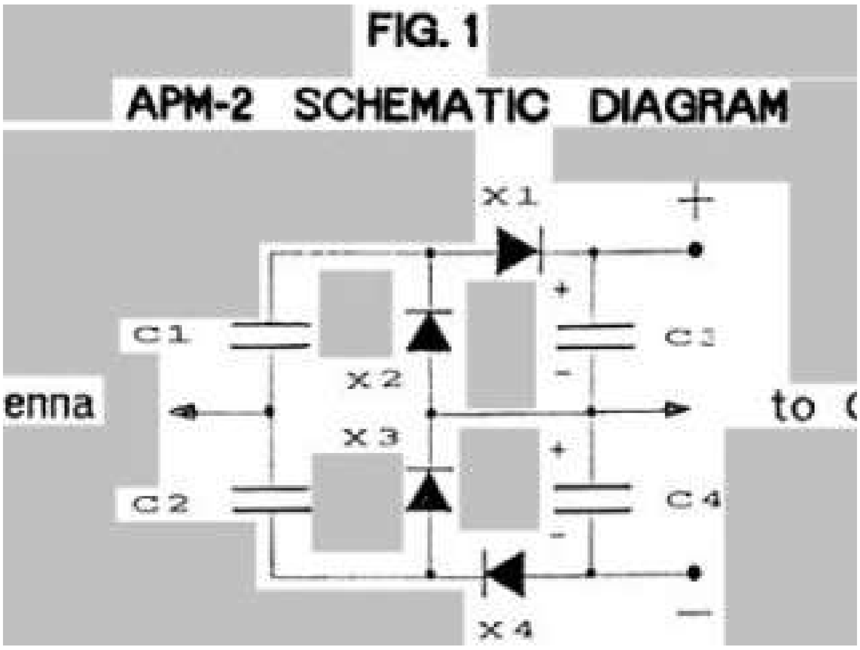

The APM operates at radio frequencies, receiving most of its power from below 1 MHz. The basic circuit may be combined with a variety of voltage regulation schemes, some of which are shown in Figure 2. Using the APM-2 to charge small NiCad batteries provides effective voltage regulation as well as convenient electrical storage. This is accomplished by connecting the APM-2 as shown in Figure 2B.

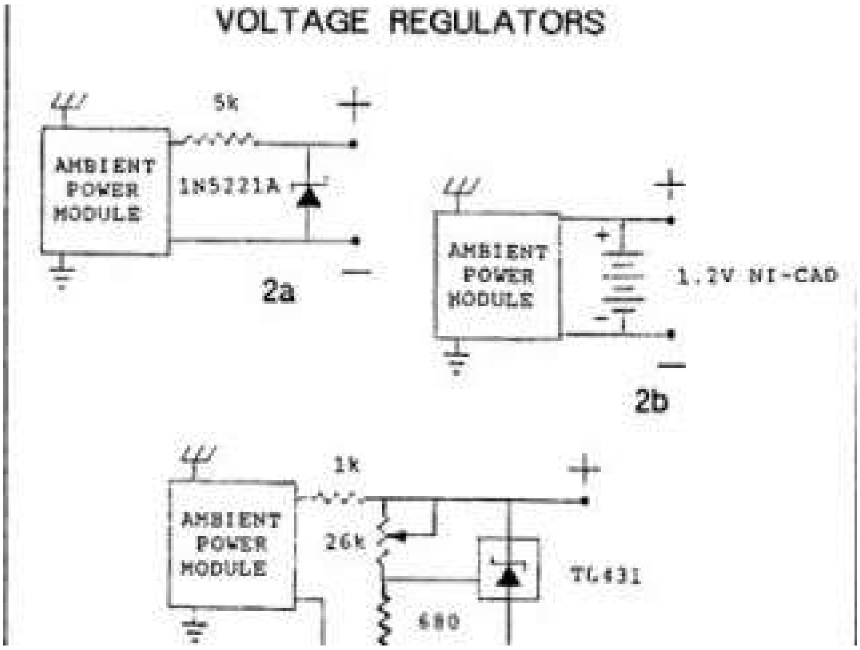

Charging lead acid batteries is not practical because their internal leakage is too high for the APM to keep up with. Similarly, this system will not provide enough power for incandescent lights except in areas of very high radio noise.

It can be used to power small electronic devices with CMOS circuitry, like clocks and calculators. Smoke alarms and low voltage LEDs also can be powered by the APM.

Figure 3 is a characteristic APM power curve measured using various loads from 0-19 kOhm. This unit was operating from a 100' horizontal wire about 25' high in Sausalito CA. As can be seen from the plot, power drops rapidly as the load resistance decrease from 2 kOhm. This means that low voltage, high impedance devices, like digital clocks, calculators and smoke alarms are the most likely applications for this power source. Some applications are shown in Figures 4 through 7.

FIG. 3 AMP Power Envelope

Volts vs. Resistance

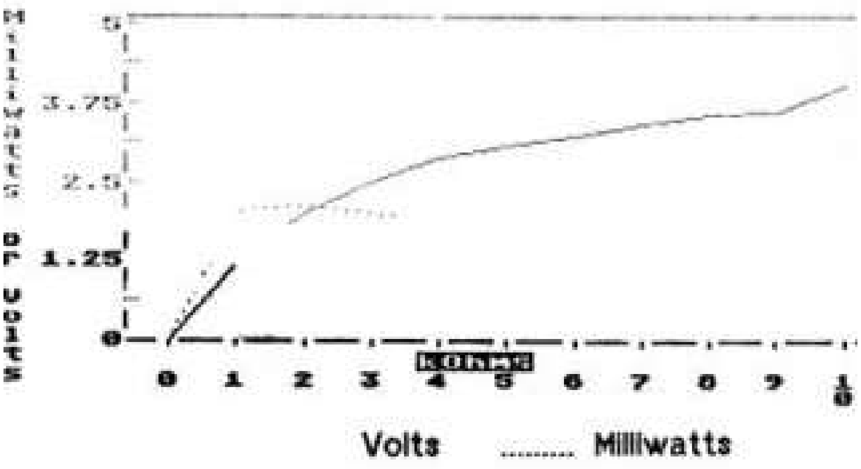

Figure 4 ~ A digital clock is shown powered by the APM-2. The 1.5 volt clock draws 28 microamps. Its position on the power envelope curve would be off the scale to the right and almost on the bottom line, dissipating only 42 microwatts.

FIG. 4 Clock powered by APM

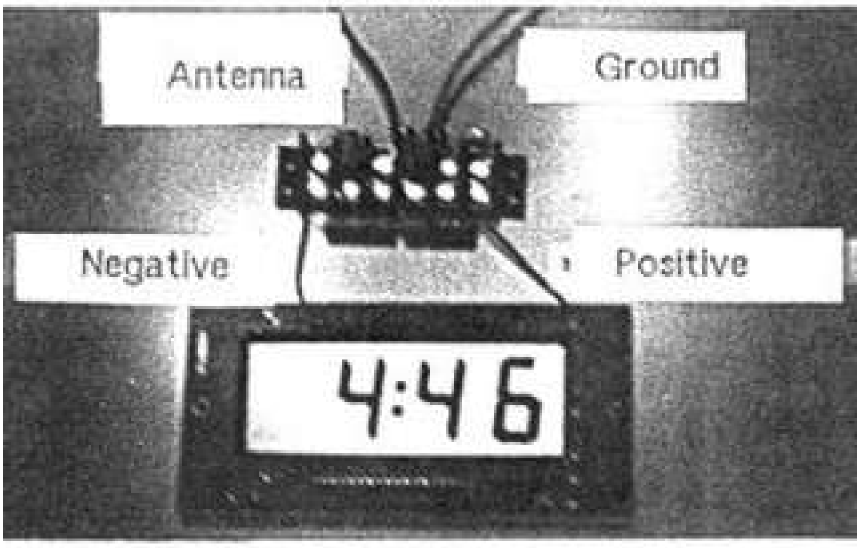

FIG. 5 APM charging a battery

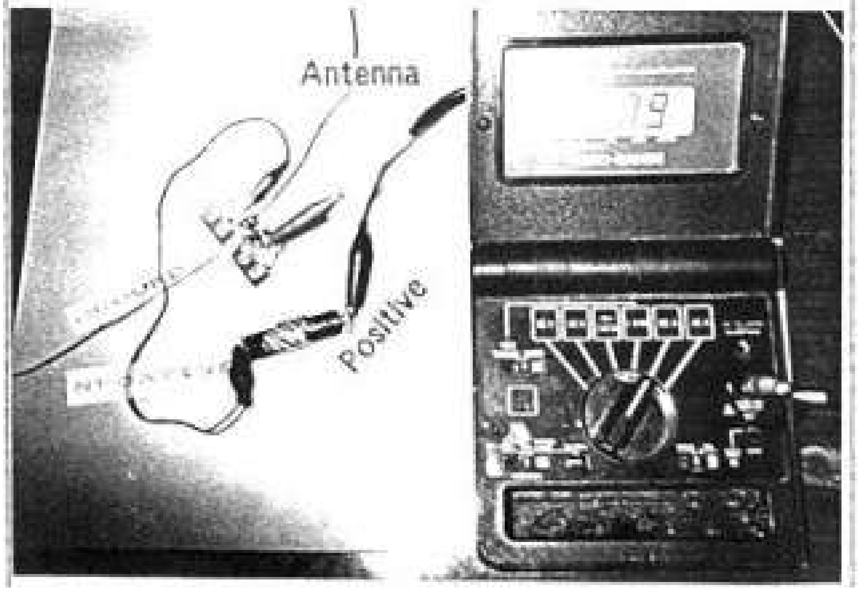

Figure 6 shows a clock which has the APM-2 built into it so it is only necessary to connect the antenna and ground wires directly to the clock. The antenna for this clock, which is a low frequency marine type, is shown in Figure 7.These antenna are expensive, not generally available, and usually don't work any better than the long wire mentioned above. But it may be necessary to use them in urban areas where space is limited and radio noise is high.

FIG. 6

Building the Module

The builder has a choice of wiring techniques which may be used to construct the module. It may be hand wired onto a terminal strip, laid out on a bread board, experiment board, or printed circuit. Figure 8 shows some of the different ways of constructing the APM-2.

Figure 8A is constructed on a screw strip terminal; Figure 8B is constructed on a perforated breadboard; Figure 8C is built on a standard experiment board; Figures 8D, 8E, and 8F are all printed circuits; Figure 8F is made up on a solder strip terminal.

FIG. 8 a, b, c, d, e, f & g

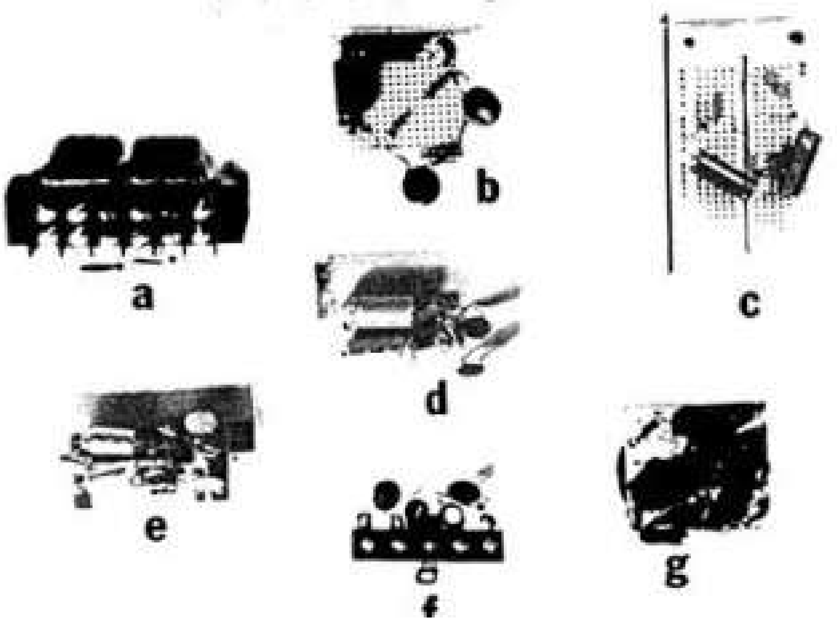

If you wish to make only one or two units, hand wiring will be most practical, either on a terminal strip or breadboard. Assembly on the terminal strip (Figure 8A) can be done easily and without soldering. It is important to get the polarity correct on the electrolytic capacitor. The arrow printed on the side of the capacitor points to negative.

Figure 9 is a closer view of the terminal strip with an illustration of the components and how they are connected.

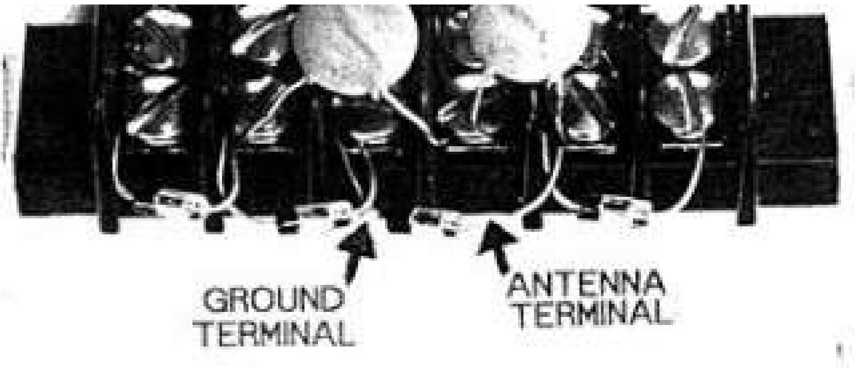

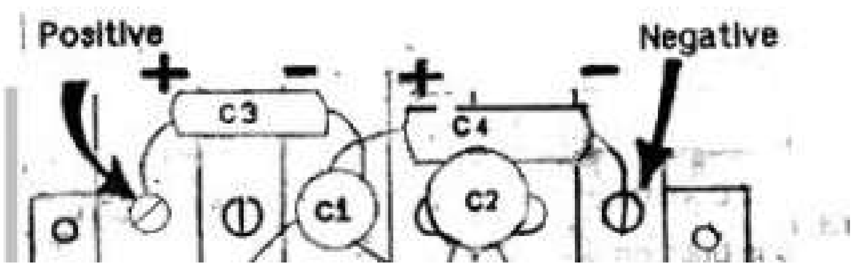

The breadboard unit is shown in Figure 10 with all components on one side and all connections on the other. All you need is a 2" x 2" piece of perforated breadboard (Radio Shack #276-1395) and the components on the parts list. Push component wires through the holes and twist them together on the other side. Just follow the pattern in the photo, making sure to observe the correct polarity on the electrolytic capacitors and the diodes. The ceramic capacitors may be inserted in either direction.

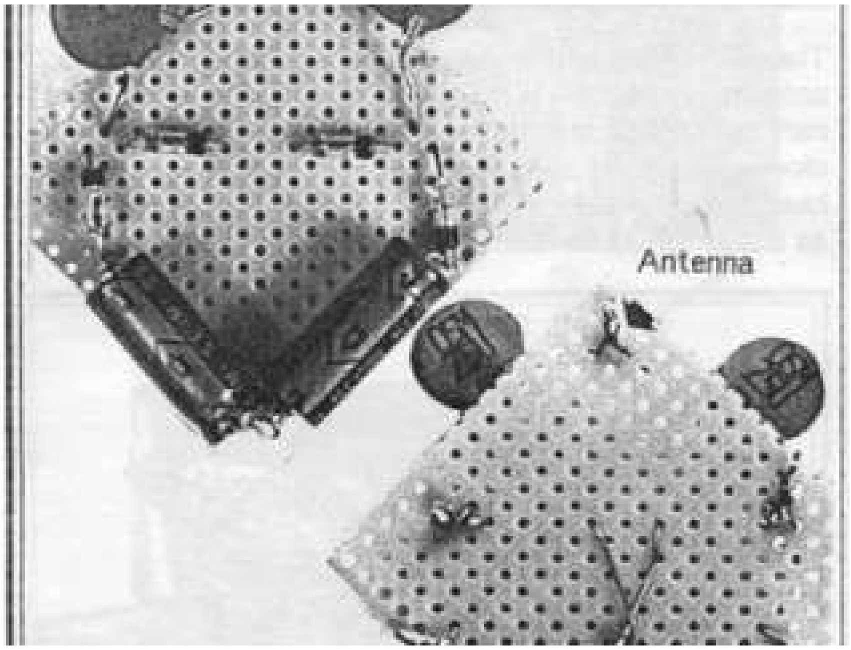

The experiment board unit is assembled by simply pushing the component leads into the board as shown in Figure 11. This unit is powering a small red LED indicated by the arrow.

FIG. 11 EXPERIMENT BOARD

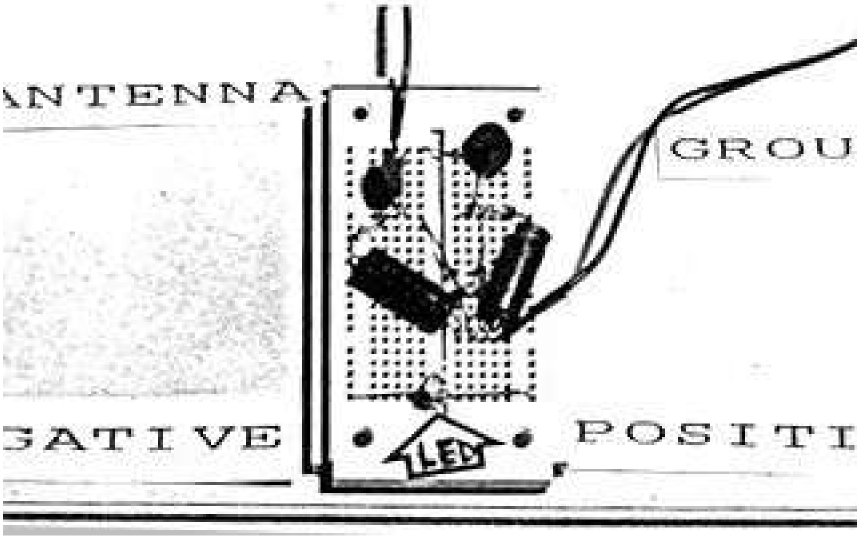

The solder strip unit is made up on a five terminal strip. The antenna connection is made to the twisted ends of the ceramic capacitors. When soldering the leads of the 1N34 diodes, care must be taken to avoid overheating. Clip a heat sink onto the lead between the diode and the terminal as shown in Figure 12.

FIG. 12 SOLDER STRIP UNIT

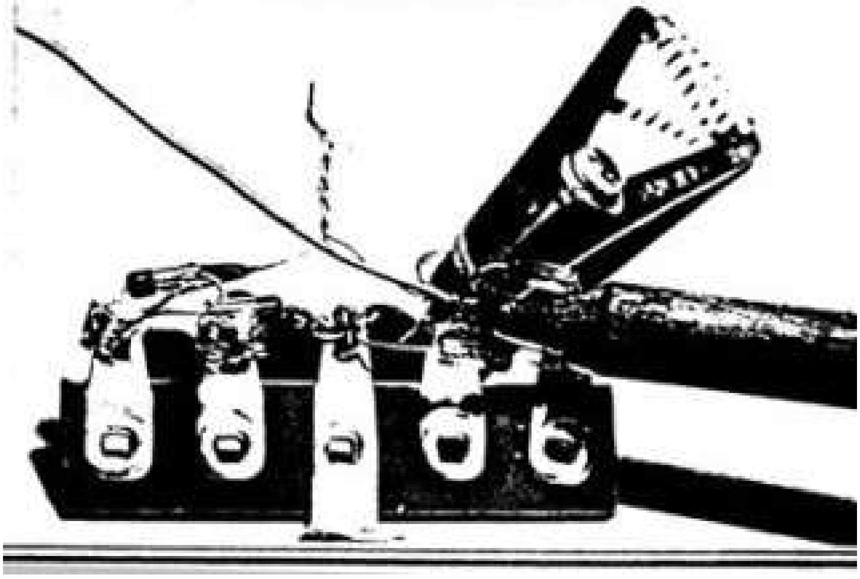

FIG. 13

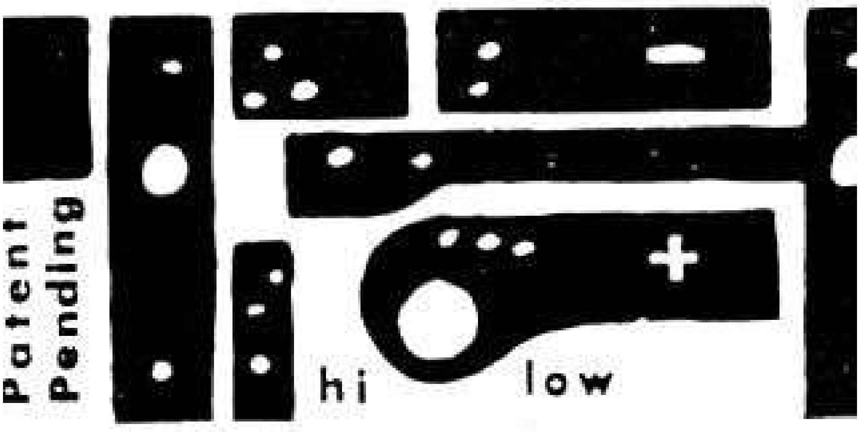
Patent Pending
+
-
hi
low

Figure 14 shows the front and back view of the completed printed circuit.

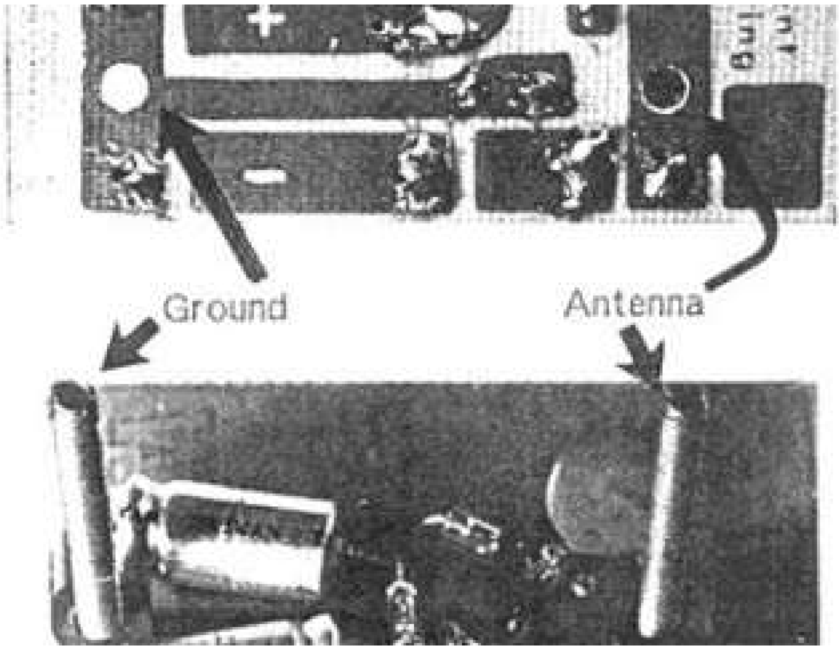
Ground
Antenna

A small switch may be installed on the board to activate the zener regulator (Figure 15). This board was designed for use in clocks.

FIG. 15

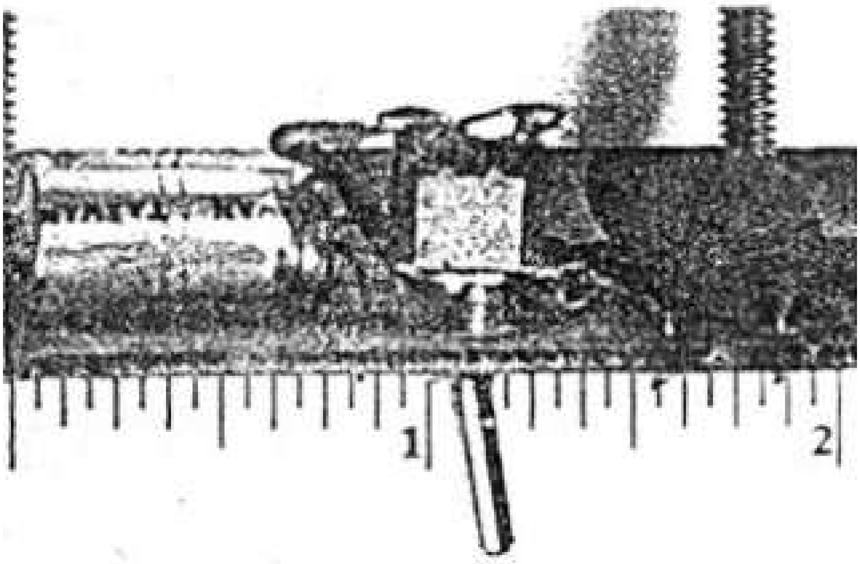

Antenna Requirements

The antenna needs to be of sufficient size to supply the APM with enough RF current to cause conduction in the germanium diodes and charge the ground coupling capacitors. It has been found that a long horizontal wire works best. It will work better when raised higher. Usually 20-30 feet is required. Lower elevations will work, but a longer wire may be necessary.

In most location, possible supporting structures already exist. The wire may be stretched between the top of a building and some nearby tree or telephone pole. If live wires are present on the building or pole, care should be taken to keep your antenna and body well clear of these hazards.

To mount the wire, standard commercial insulators may be sued as well as homemade devices. Plastic pipe makes an excellent antenna insulator. Synthetic rope also works very well, and has the advantage of being secured simply by tying a knot. It is convenient to mount a pulley at some elevated point so the antenna wire may be pulled up to it using the rope which doubles as an insulator (Figure 16).

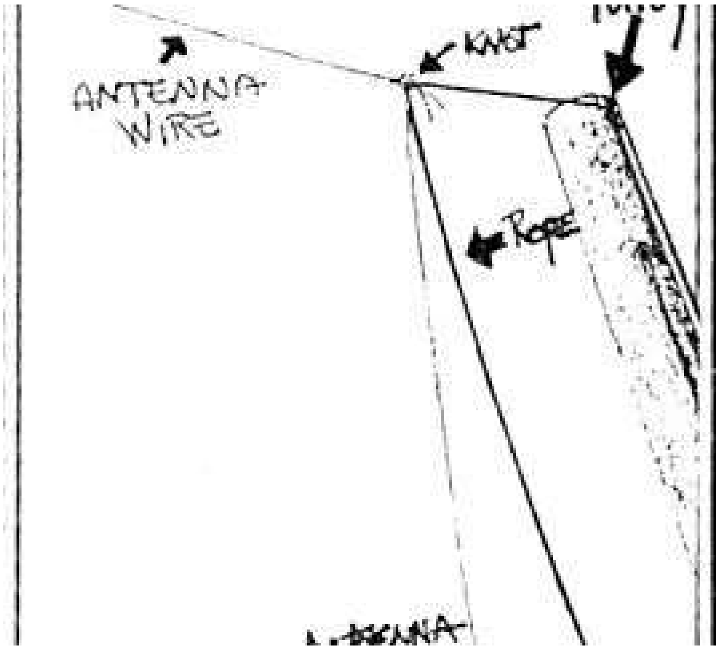

Figure 17 is an illustration of a horizontal wire antenna using a building and tree for supports.

FIG. 17

Horizontal Wire Installation

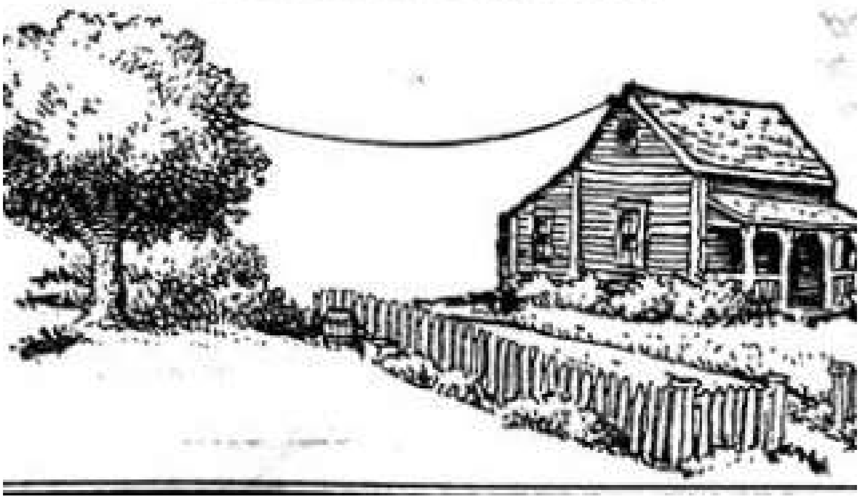

Grounding

Usually a good ground can be established by connecting a wire to the water or gas pipes of a building. Solder or screw the wire to the APM-2 ground terminal. In buildings with plastic pipes or joints, some other hookup must be used. A metal rod or pipe may be driven into the ground in a shady location where the earth usually is damper. Special copper coated steel rods are made for grounds which have the advantage of good bonding to copper wire. A ground of this type usually is found within the electrical system of most buildings.

Conduit is a convenient ground provided that the conduit is properly grounded. This may be checked with an ohmmeter by testing continuity between the conduit and system ground (ground rod). Just as with the antenna, keep the ground wire away form the hot wires. The APM's ground wire may pass through conduit with other wires but should only be installed by qualified personnel.

Grounding in extremely dry ground can be enhanced by burying some salts around the rod. The slats will increase the conductivity of the ground and also help retain water. More information on this subject may be found in an antenna handbook.

Good luck getting your Ambient Power Module working. It is our hope that experimenters will find new applications and improve the power capabilities of the APM.

CopyRight Joseph Tate

Inventors: Tate, Joseph B. (Sausalito, CA)
Brown, David E. (Mill Valley, CA)

Application Number: 06/695632

Publication Date: 12/09/1986

Filing Date: 01/28/1985 NOTE: The images maybe hard to see as they were originally scanned and uploaded in black and white. If it helps you to replicate this device you may want to check the images of the one I built at the very top (in color) Here are some more photos of my completed unit for your review.

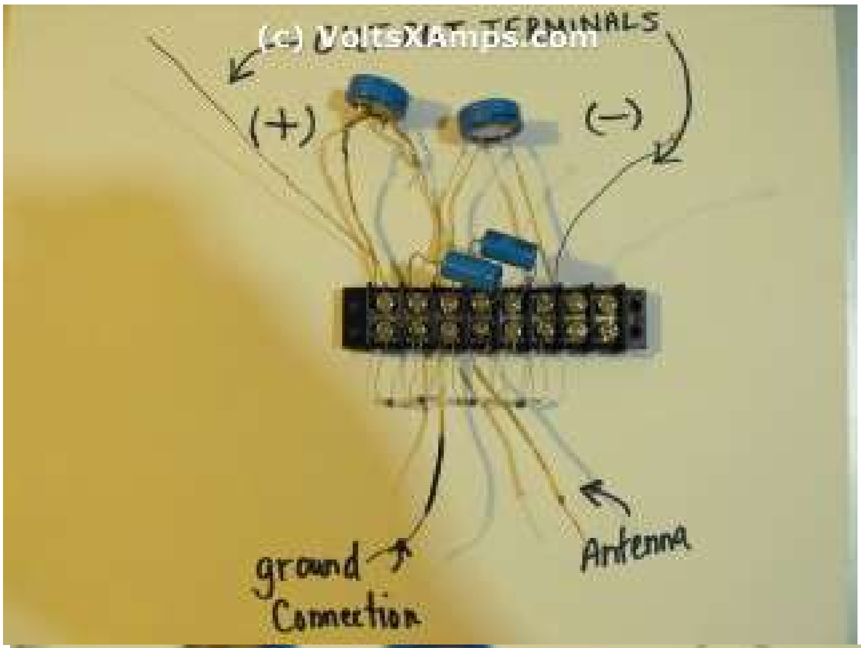
(c) VoltsXAmps.com
TERMINALS
(+)
(−)
ground Connection
Antenna

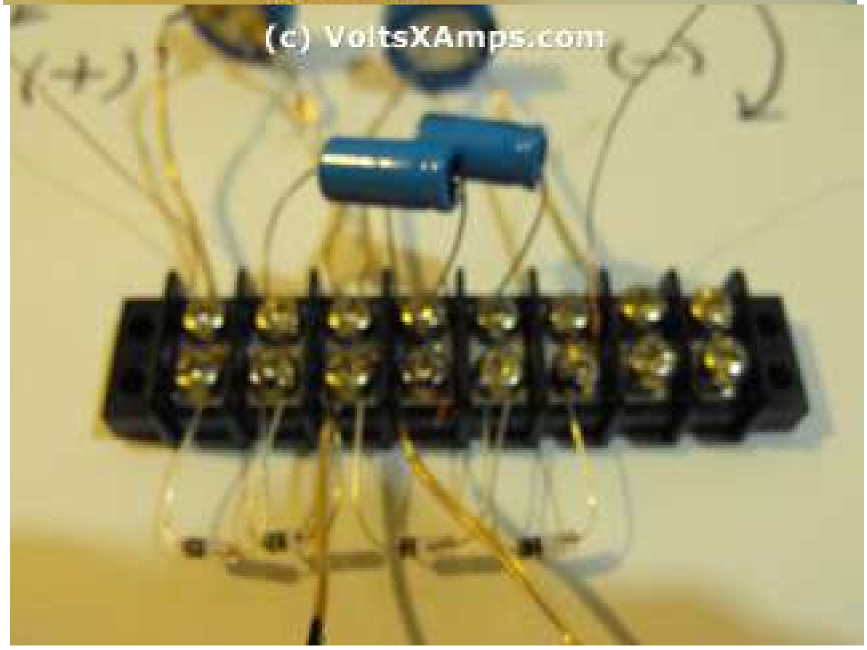
(c) VoltsXAmps.com
(+)
(−)

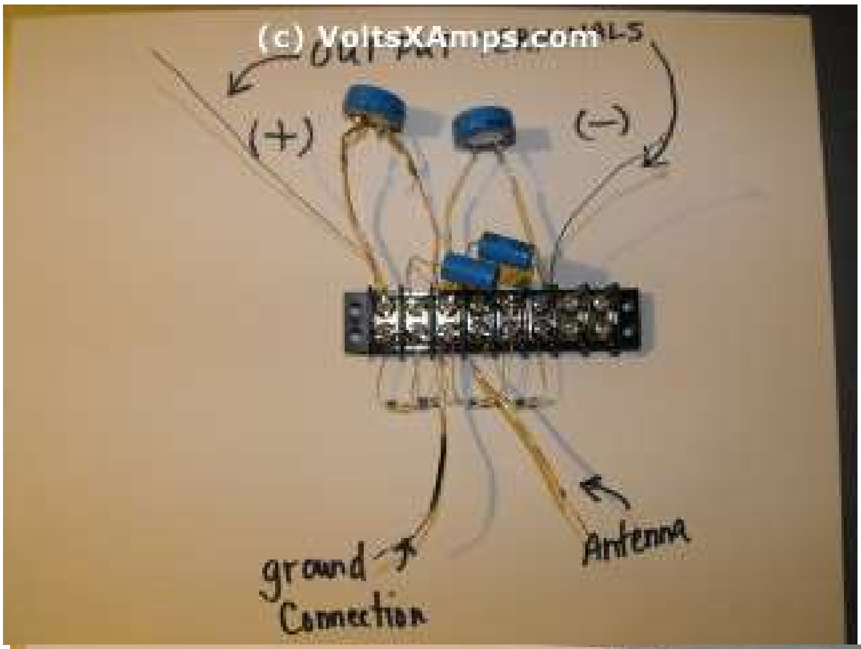
(c) VoltsXAmps.com
(+)
(–)
ground Connection
Antenna

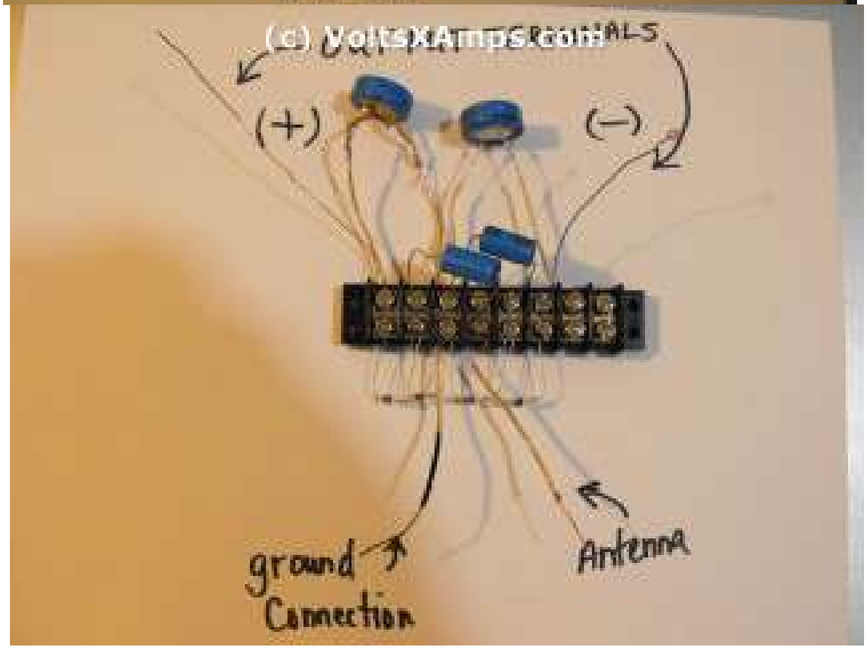
(c) VoltsXAmps.com
(+)
(–)
ground Connection
Antenna

[/protected]

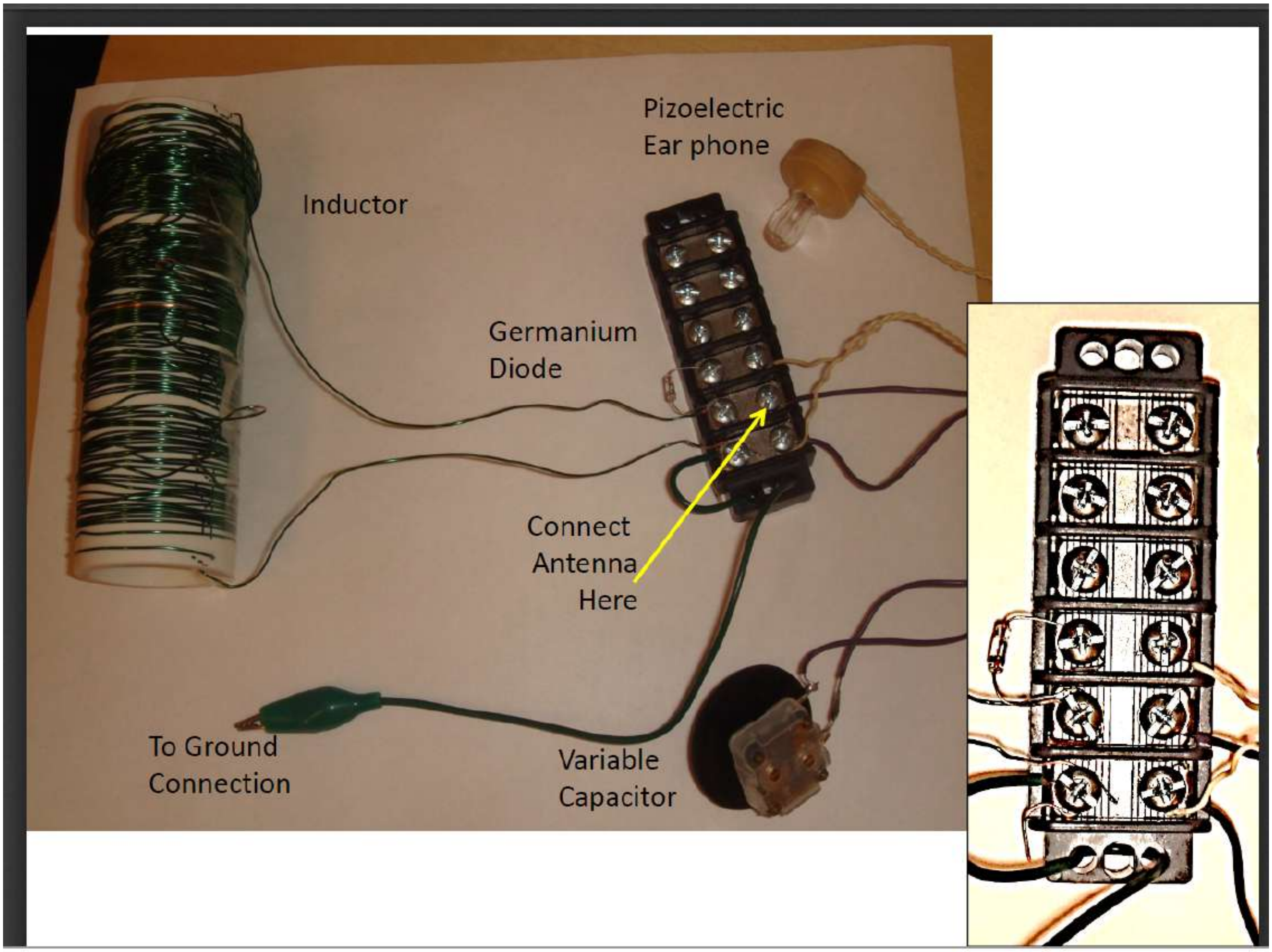
Pizoelectric
Ear phone
Inductor
Germanium
Diode
Connect
Antenna
Here
To Ground
Connection
Variable
Capacitor

FM Crystal Radios?

copyright © 2006 by Larry J Solomon

I have heard, even from a physicist, that it is impossible to build FM crystal radios. On the other hand some experimenters claim that they have built them. This argument intrigued me to try and build an FM crystal radio, which I have done successfully. To my surprise, the result is an astounding performer, pulling in four local stations in Tucson. When connected as a receiver to a good sound system the sound fidelity is as good or better than more expensive AM radios. In fact, it sounds "high-fidelity".

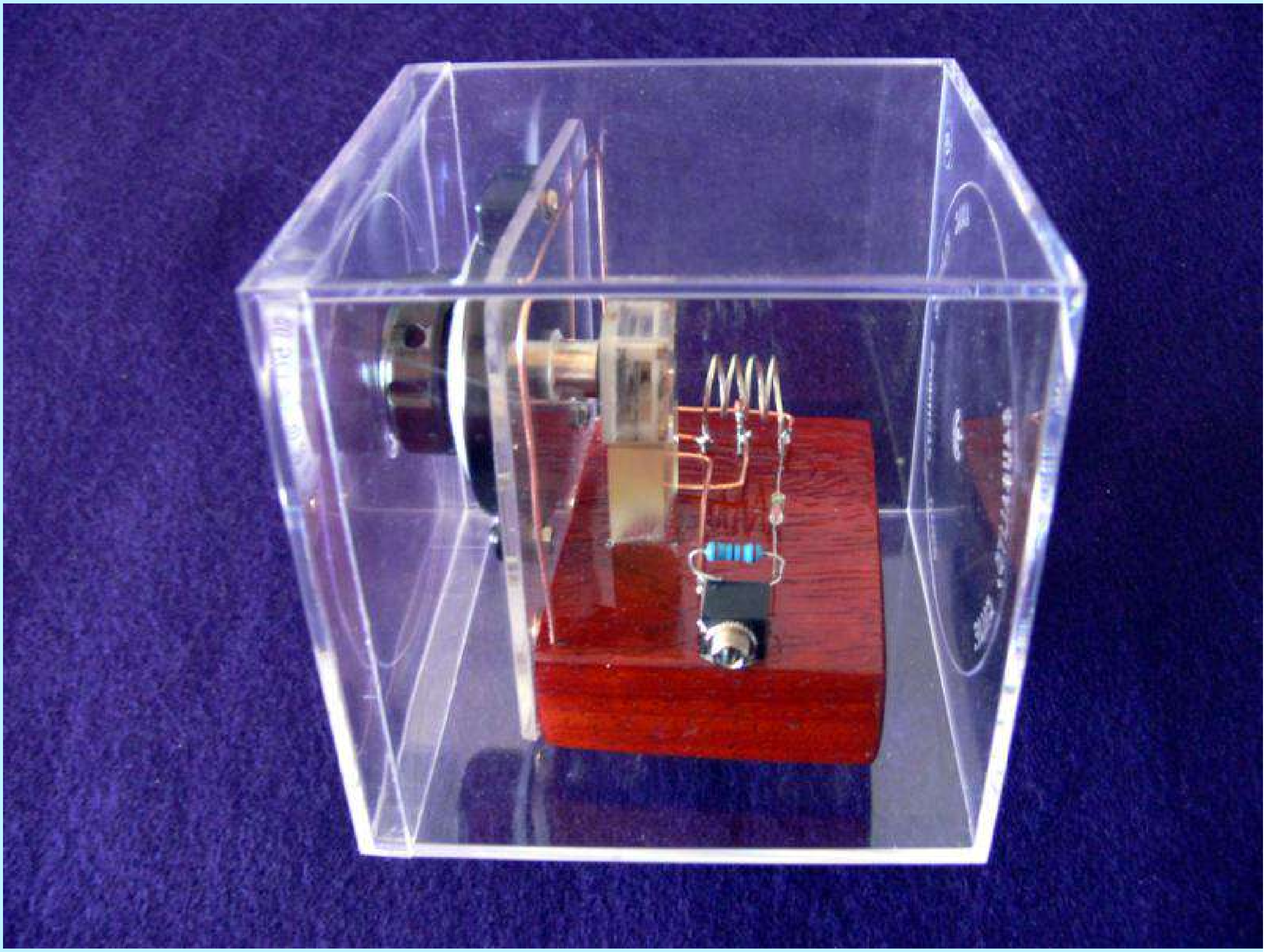

This picture shows the Solomon FM Crystal Set in an acrylic display case. I made the set specifically to fit inside this case (the case came first).

My definition of a crystal radio is one that is not powered, except by the radio transmission itself and employs a crystal detector. So, it should work without any batteries or AC power. An FM crystal receiver must be able to detect and receive FM signals well enough to be heard in earphones without any such extra power.

This FM receiver is an amazing performer. It has crystal clear reception (pun intended), good sensitivity, but only fair selectivity. This set was a discovery for me. I started out by designing and building the normal AM sets. Then one day while testing the "Mystery" set (see my other web links), to my surprise, in addition to the expected panoply of AM stations, I heard a very faint signal that I could not tune out. At first, it seemed too weak to identify. When I tuned out all the AM stations, I was astonished to hear the announcement "KiiM FM, 99.5"! This is a

country music FM station here in Tucson. It was all over the dial, untunable, but the much louder AM signals masked it when they were tuned in.

I set myself the task of trying to improve the FM reception. I tried some simple circuit modifications that did not seem to improve anything. Then I connected a dipole antenna instead of the AM antenna I normally use. Suddenly, the FM signal was much clearer, although still weak. By using the audio output and sound system amplifier, I was even more amazed that four different FM stations came in loud (or rather medium) and clear. I found that changing the telescoping antenna length and position I could tune the stations in and out. They were KRQ, KLPX, KiiM, and KHYT all local FM stations with transmitters nearby. Their reception was also affected by the length and position of the audio output cable.

After doing some research, I discovered that there was a physical theory that claimed that FM reception was possible and even probable using the same circuit as an AM receiver. The theory is called "slope detection". So, I set out to find circuit improvements. A web search yielded little, mostly theory. But there was enough information that I thought I could make some modifications to the AM circuits to make them more tunable to FM signals and less tunable to AM. Since FM operates at higher frequencies, all I had to do, I thought, was make the coil and caps smaller. After much "tinkering" I arrived at the current circuit.

The circuit looks identical to a classic AM crystal circuit but is even simpler to build. The components were reduced in dimension to resonate at higher frequencies. This was done by experimenting with smaller and smaller coils and capacitors. The antenna is also much reduced in size (from that of AM) to resonate at higher frequencies (the antenna is crucial). The air variable capacitor I used has two trimmers in it which should be adjusted for best reception.I have found that a commonly available vernier dial and knob will fit the capacitor nicely. See end of article for a picture of the variable. C3 is a ceramic capacitor of 18 pf, but may be anywhere from 10 to 50pf. A detected FM signal is converted to AM due to an effect called slope detection that modulates amplitude.

This FM Crystal Set works best near the transmitter (I have not tested it beyond about 10 miles). Secondly, the sound level is quiet, especially without an amplifier. A quiet room is needed for listening with earphones. One must be willing to move the set around to find a location for the best reception of signals. However, in addition to listening with high impedance earphones (crystal or otherwise), the set can be connected directly to an audio amplifier's low level magnetic input which can then play amplified through a sound system at any volume -- sounds GREAT. In fact, I recommend starting tests with the FM crystal set by connecting it to the low-level phono inputs of a receiver or preamplifier. (Nowadays, many receivers don't even have a phono input!) That way you can crank up the volume, which makes it more likely to find the FM stations. If no signals are detected, I also recommend connecting an external "rabbit ear" antenna or hanging a short wire (12 inches or so) in various positions next to the internal antenna. The variable length of rabbitt ears can help to tune in stations.

No additional wiring or antenna is necessary (the antenna is optimized in length for FM.)

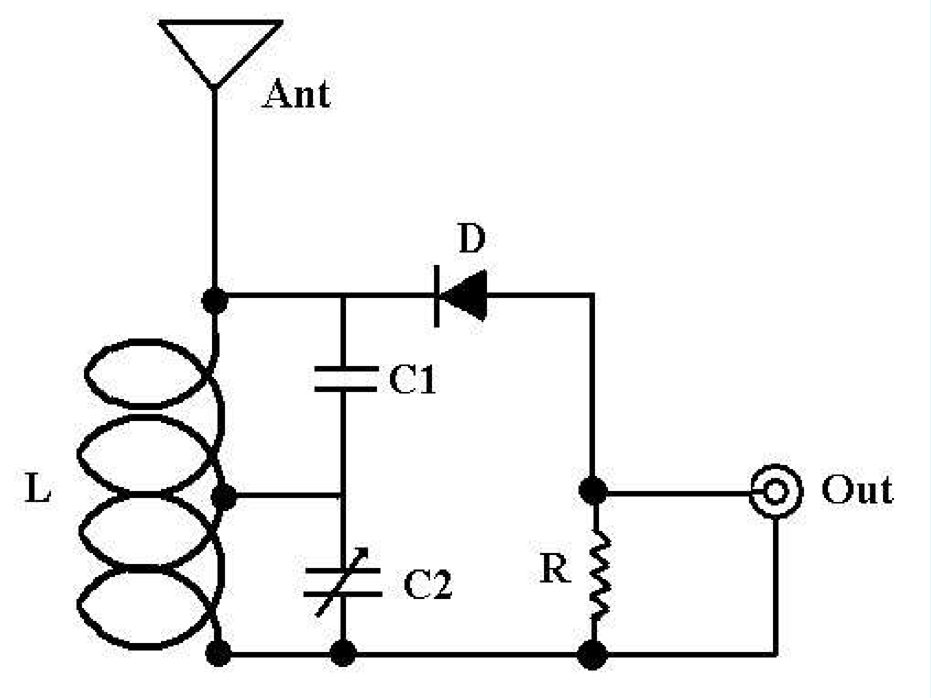

L - 4 turns #18 copper or silver wire, 12mm inside diameter, tapped at 2.5 turns
Ant - 7 inches of #18 bare copper wire
C1 - 18 pf ceramic capacitor
C2 - 50 pf air variable capacitor
D - 1N34 diode or rock crystal
R - 150K resistor

The diode is tapped directly to the antenna. The vernier dial fits directly on the tuning capacitor. The antenna parallels the perimeter of the acrylic face plate. "Military style" #18 AWG wiring is used without any insulation. It is important to keep the components physically close together. The component specifications are the same as in circuit #2. The coil is silver rather than copper, but copper does just as well. I think that the contrast of the silver and copper is beautiful. The coil was wrapped around a Sharpie Permanent Marker, then slipped off and expanded slightly. The wooden base is made from lacquered, polyurethane padouk.

I consider this set a work of art as well as science and think it is the most elegant crystal receiver I have created. I love the contrast of the silver coil, the copper antenna, the clear acrylic faceplate, the black vernier dial, the white and transparent variable capacitor, and the subtle colorings on the resistor, the diode, and the lucite base. Yet the circuit is so ridiculously simple that some will not believe it is possible without building it themselves. No shielding is necessary, and there is no problem with hand capacitance. However, the output cable position may affect reception sensitivity.

Photos of wired circuit

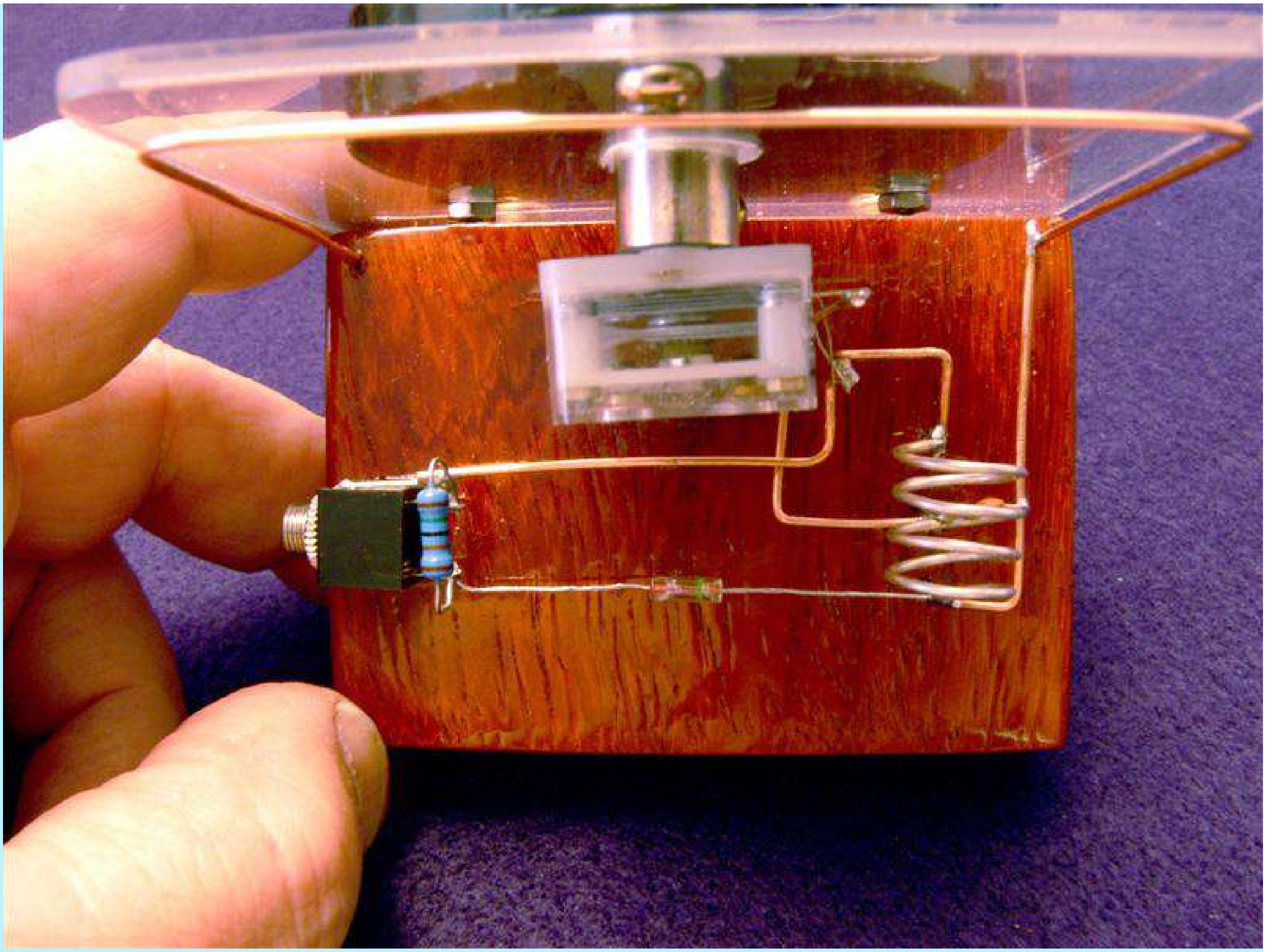

A hand is included in this photograph to show scale. Note the military style wiring, diode, and antenna. I wanted the wiring to create a modern design similar to a Mondrian painting. Not only is this set beautiful, it works! No power and no long antenna! It looks like a work of fiction.

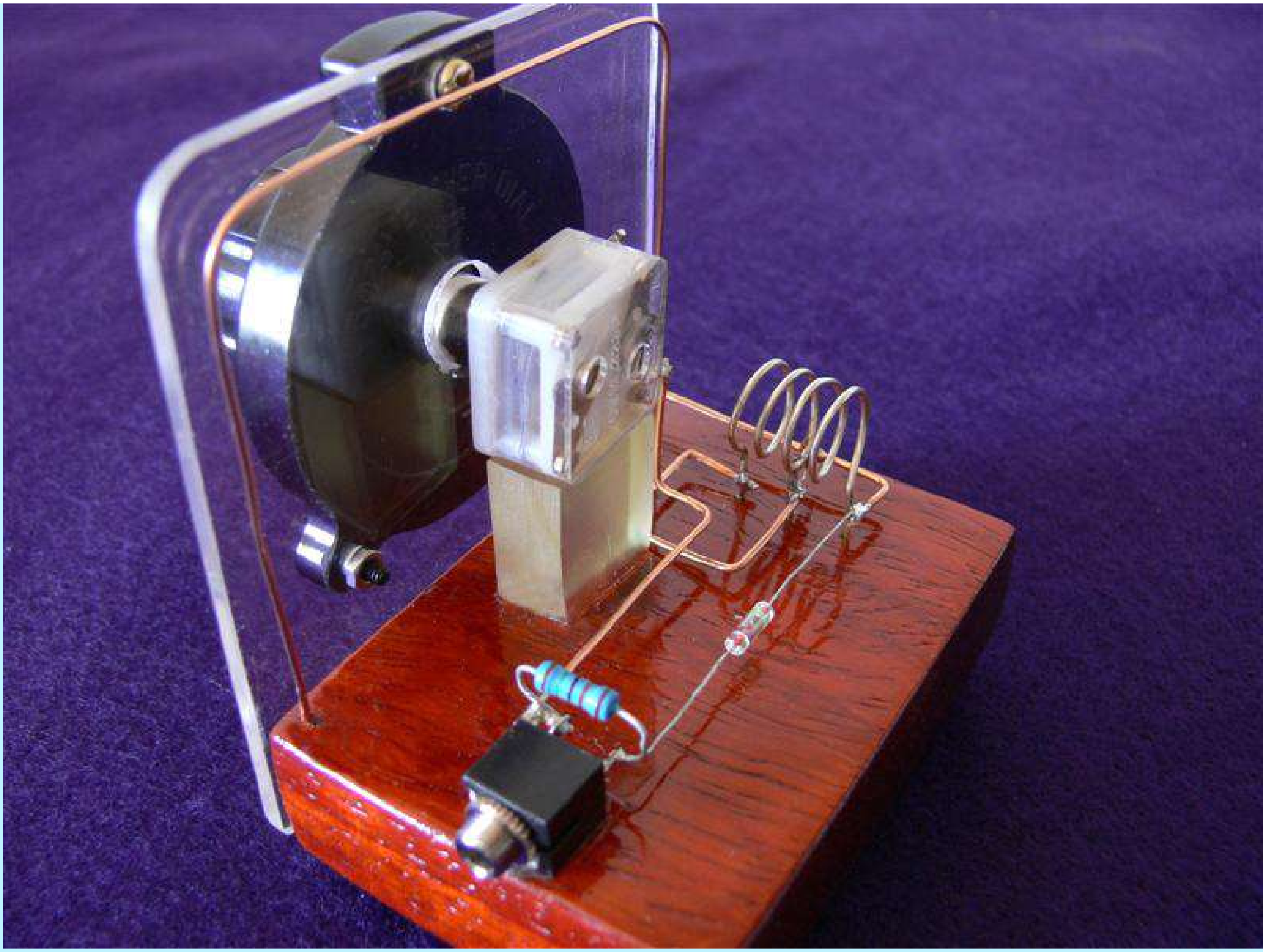

Is this thing imaginary -- science fiction? Well, imagination did play a part, but it is definitely not science fiction. This shot shows the elegance of the FM set best, I think. There is only one resistor and one fixed capacitor.

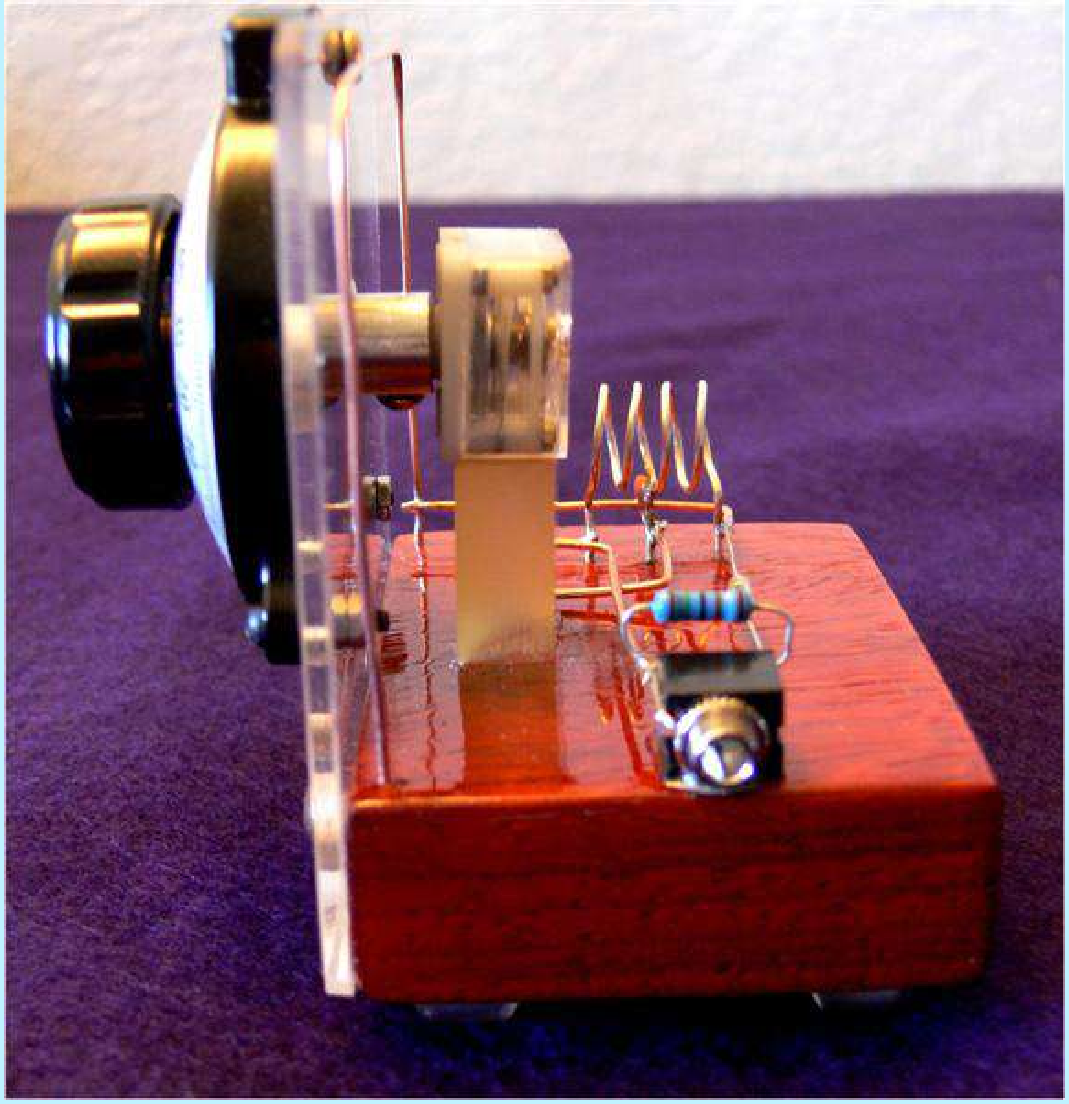

The inside of the tuning capacitor and the phono jack/output can be seen here. Can you spot the fixed ceramic capacitor? Note the polished edge of the face plate and the reflection in the wooden base.

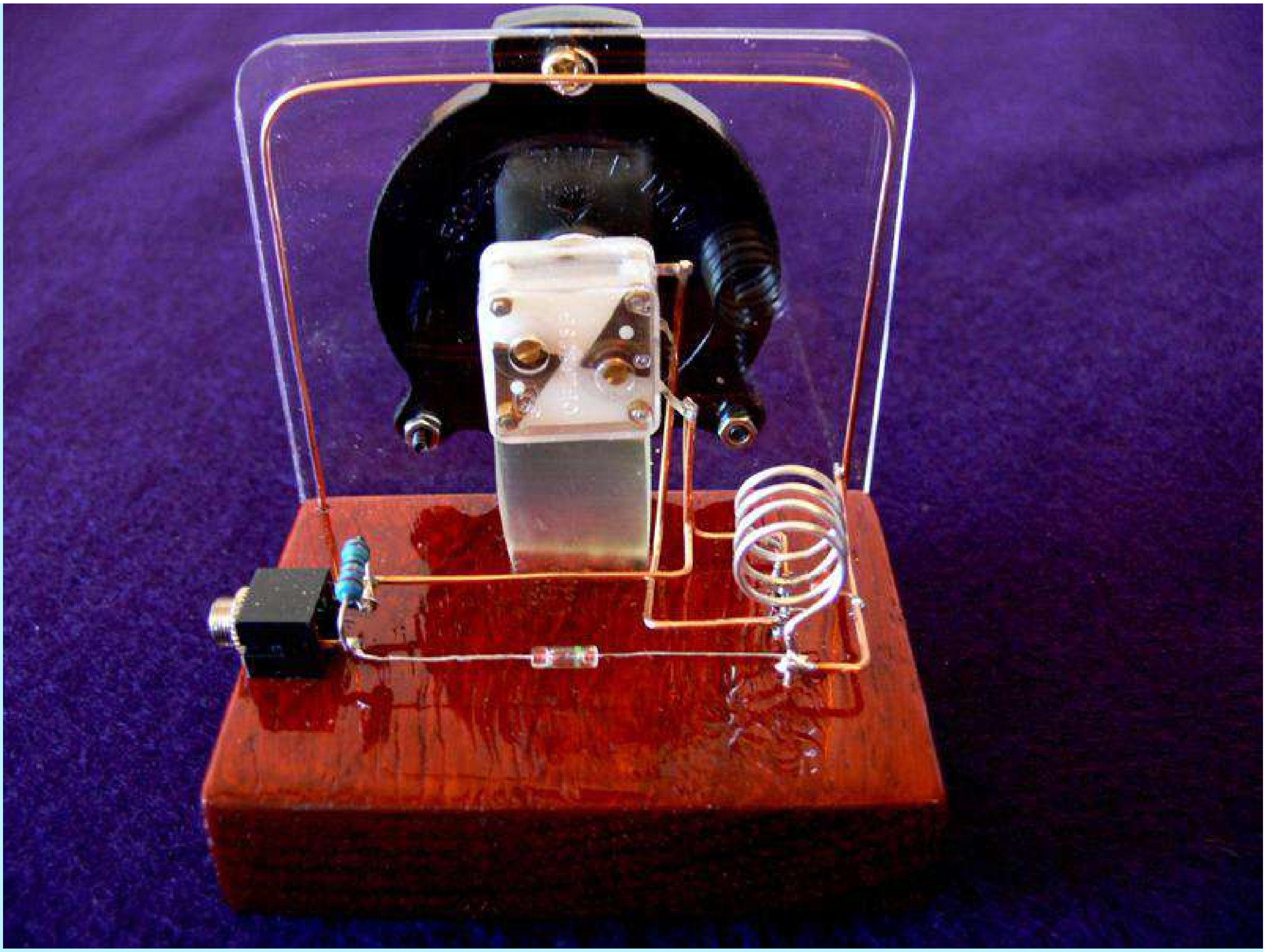

A quarter-inch piece of lucite was fitted under the tuning capacitor to anchor it. Note the two tiny trimmers on the back of the tuning capacitor. Brass screws were used to enhance appearance.

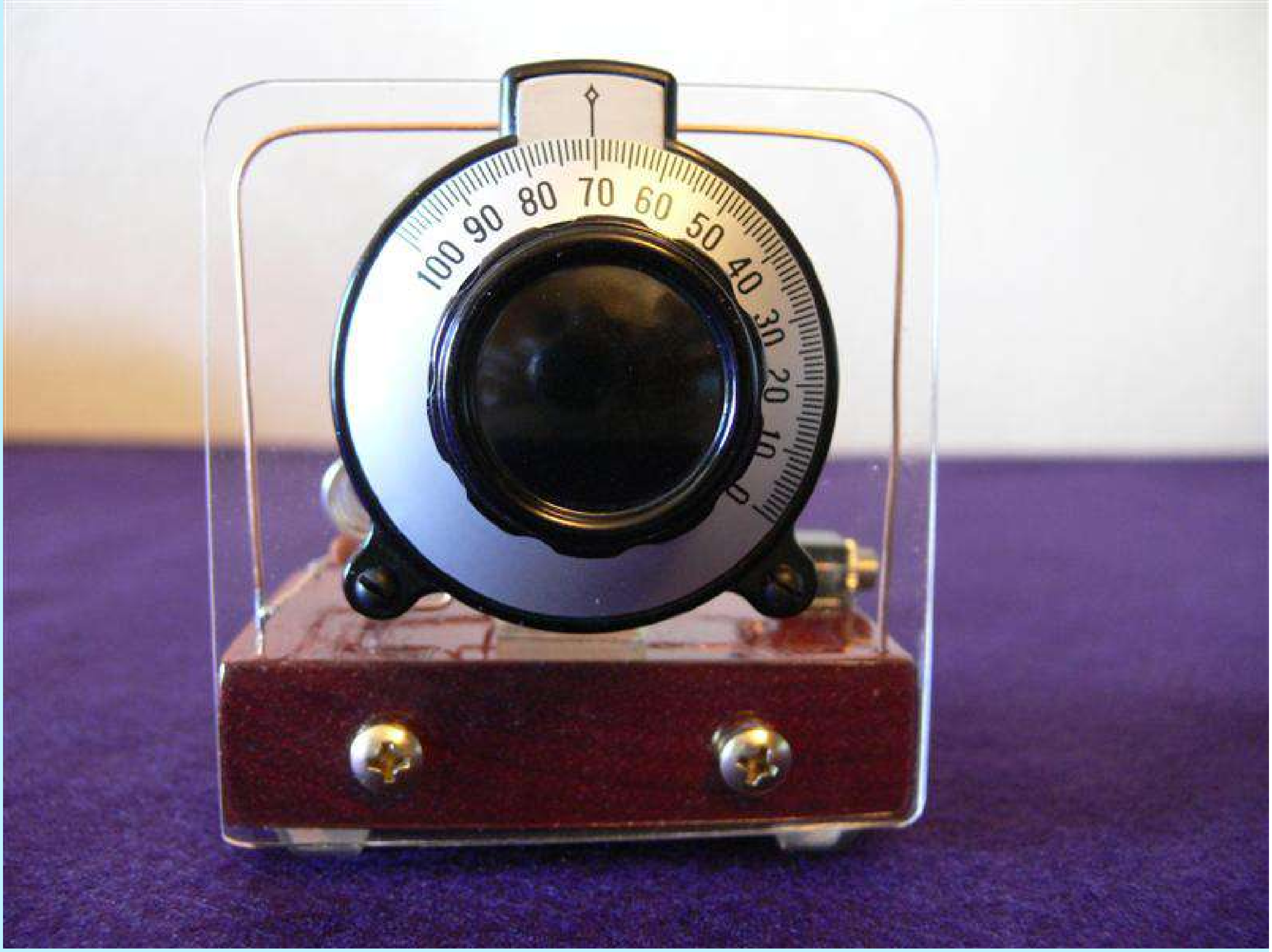

The vernier dial is large to accomodate ease of tuning, and the vernier makes it easy to separate stations. Two golden (brass) wood screws fix the face plate to the base. Holes for the face plate were made with special plastic drills, but ordinary drills may be used if drilled very SLOWLY. The knob is removable.

FM Crystal Circuit #2

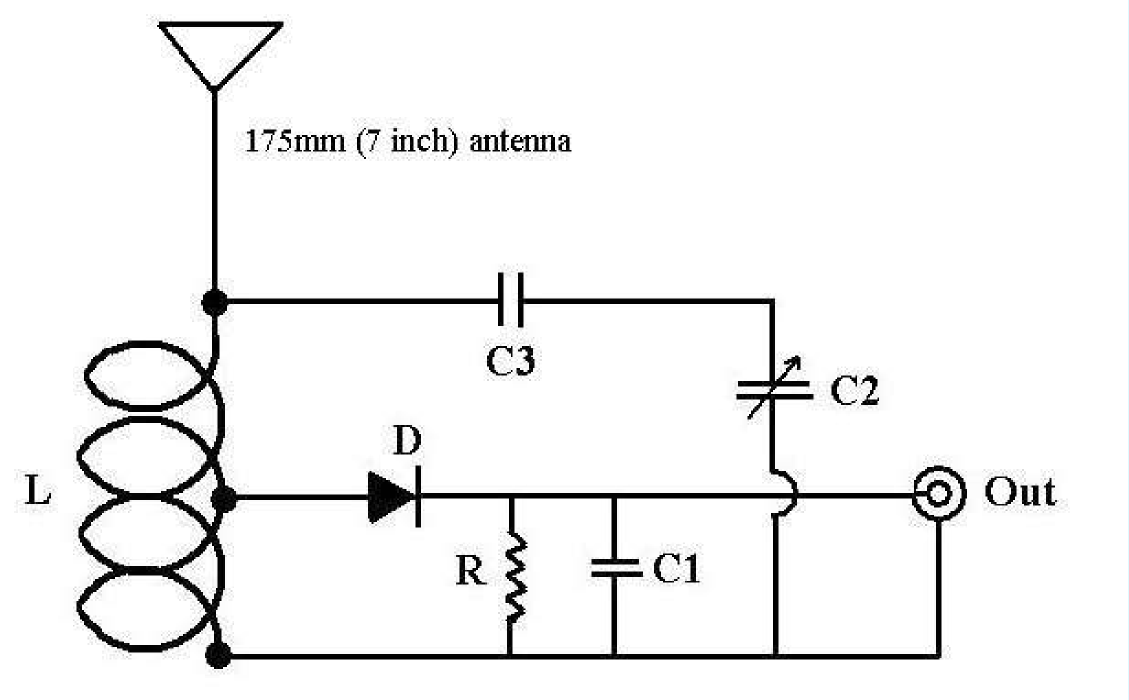

L - 5 turns AWG#18 bare copper or silver wire, 12mm inside diameter, tapped at 2.5 turns
D - 1N34 or rock crystal diode
C1 - 82 pf capacitor
C2 - 80 pf air variable capacitor
C3 - 18 pf capacitor
R - 150K resistor

The following photographs show the circuit wired with the handmade Saturn Dial. and knob. It is perhaps not as visually striking as set No. 1, but it works just as well. In fact, this set was the original version. Notice that all the wiring and coil are copper.

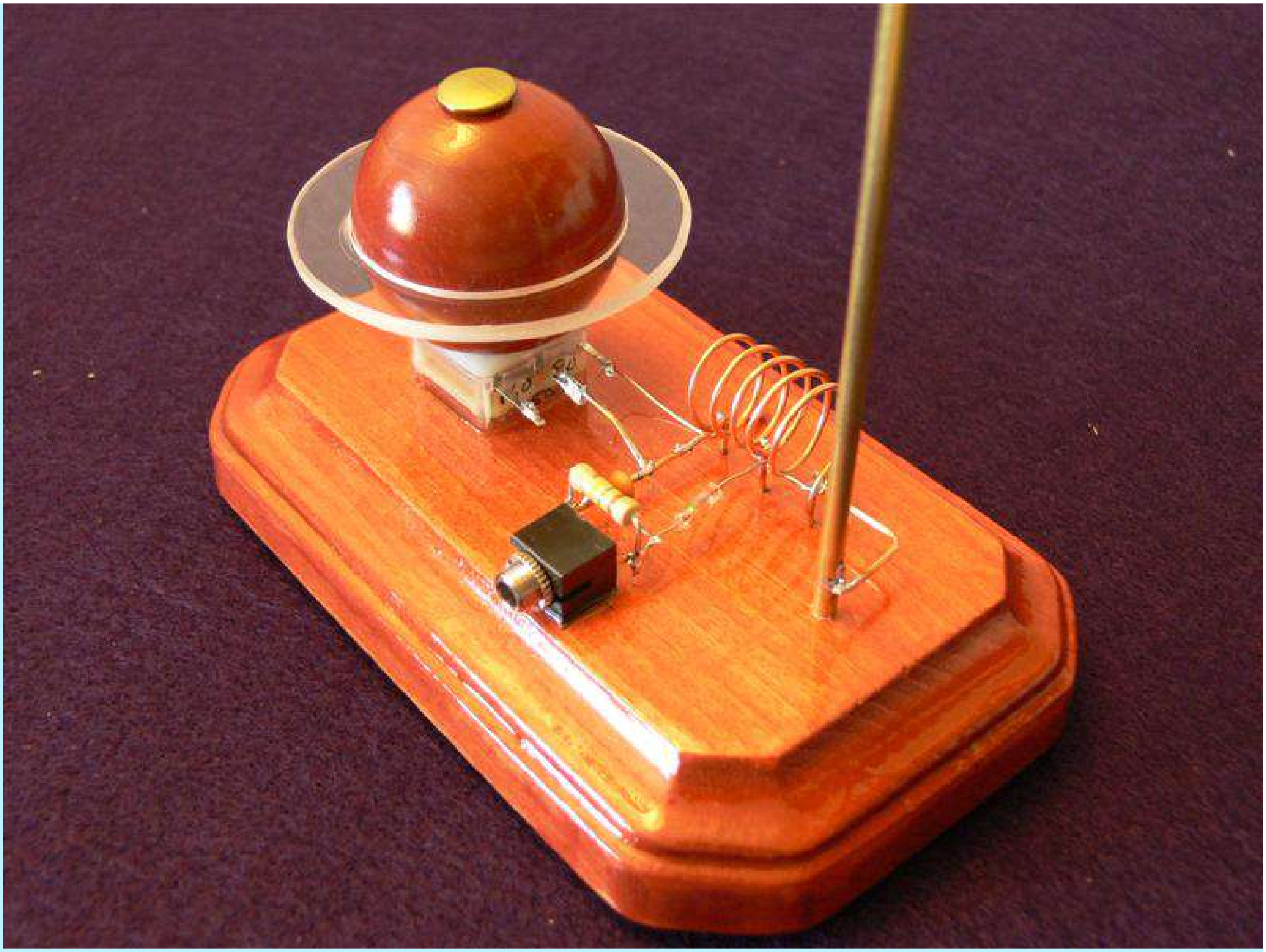

The Saturn dial and knob were fashioned from a "doll's head" from Michael's Arts and Crafts, a piece of lucite cut with two circle cutters, and a brass paper fastener. The knob is fixed to the tuning capacitor with a small machine screw that fits in the hole below the brass fastener. The most difficult part of this was fashioning "Saturn's rings". This must be done very carefully and slowly. The inside edge should be cut slightly undersized and then sanded with a drum sander to fit snugly. The outside edges can be sanded with fine sandpaper and polished with a plastic polisher.

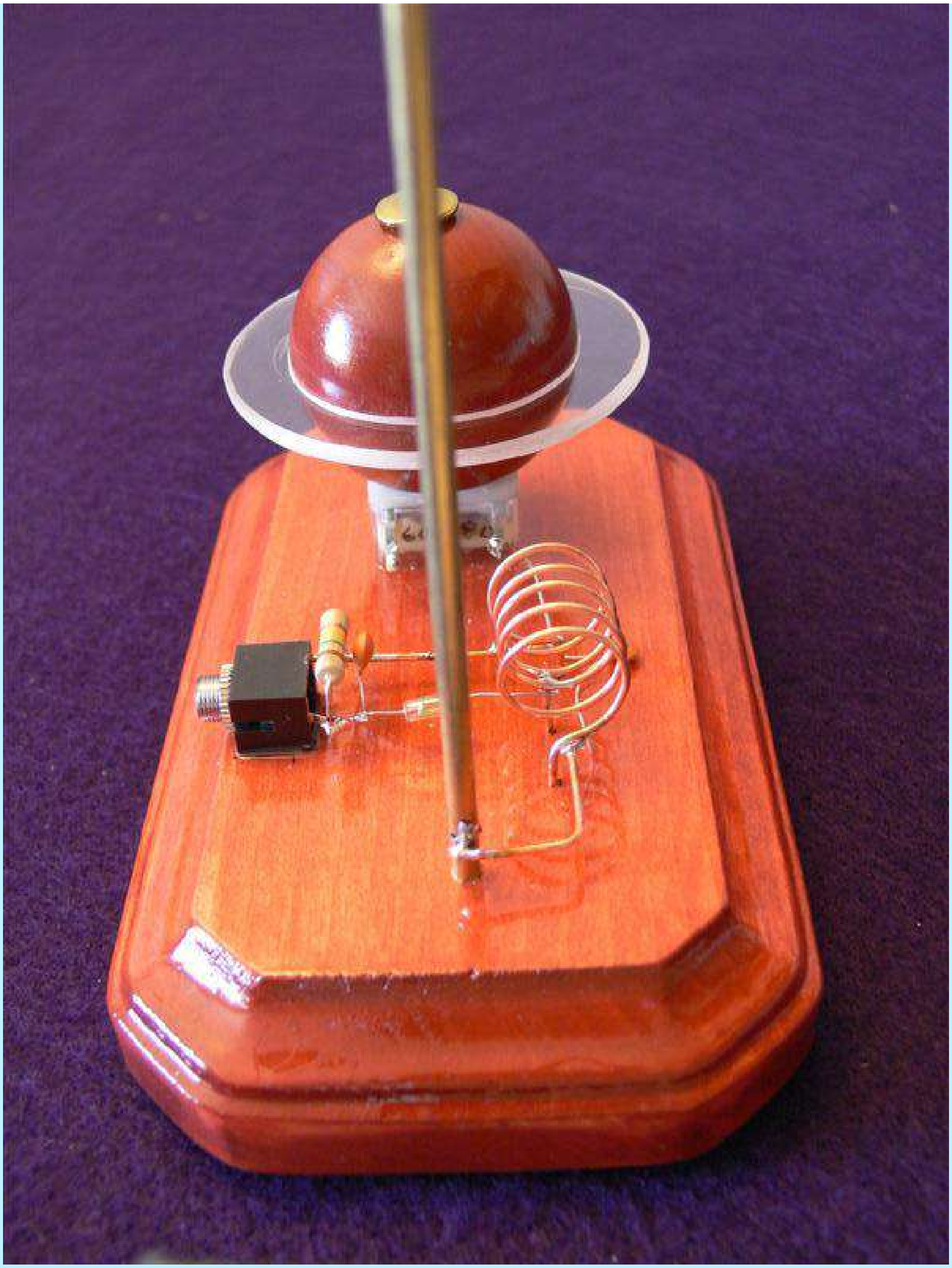

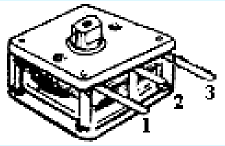

The air variable capacitor may be obtained from Electronix Express at http://www.elexp.com/. Part number 14VCRF10-280P. The 80 pf side is recommended for the second circuit, contacts 2-3. Contacts 1 and 3 were used for the first circuit (50pf).

- OSC: 5-59 pf
- ANT: 5-142 pf
- OSC and ANT Trimmer 10pf range

Home -- If you want to contact me go to the home page and go through the Music Clinic link.

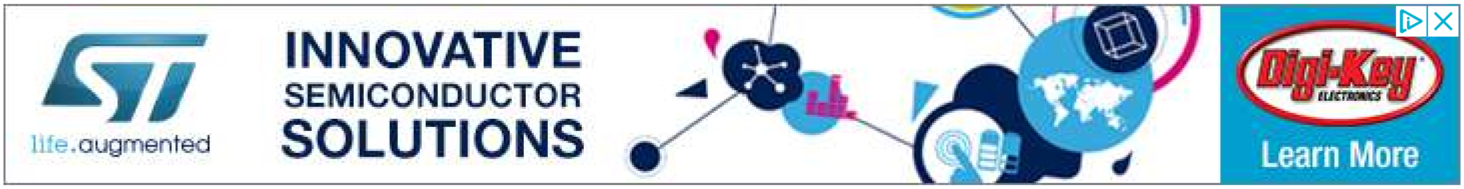

life.augmented
INNOVATIVE
SEMICONDUCTOR
SOLUTIONS
Digi-Key
ELECTRONICS
Learn More

advertisement

FREE ENERGY FROM THIN AIR!

Technology (/technology/) > Science | by DrewPaulDesigns (/member/DrewPaulDesigns/) Follow

79,181 1,078 230

Posted May. 24, 2016 | BY-NC-SA

FREE ENERGY
FROM THIN AIR!
*INVOLVES CRYSTALS AND
DANGEROUS VOLTAGES!
Copyright 2016. Drew Paul Designs, All Rights Reserved.

(https://cdn.instructables.com/FAT/Z4RP/IOKBVYWL/FATZ4RPIOKBVYWL.LARGE.jpg)

DrawDayDesigns.com

(https://cdn.instructables.com/F5S/A5K4/IOKBUDZE/F5SA5K4IO

"Dendera Light"
circa1500 BC
Image: creative commons, Wikipedia

(https://cdn.instructables.com/FQH/L0OE/IOKBVTOS/

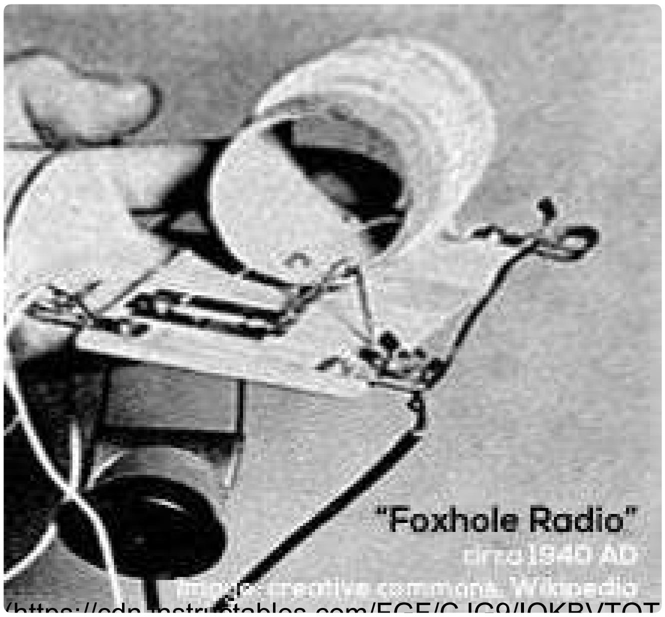
"Foxhole Radio"
circa 1940 AD
Image: creative commons, Wikipedia

(https://cdn.instructables.com/FGE/CJG9/IOKBVTOT

3.74
CEN-TECH

(https://cdn.instructables.com/FHS/HM1V/IOKBT6CY/FHSHM1V

"Free energy from the air?, Yea, right!" Sardonic skepticism was my first reaction to this unusual concept, as well.

Though, its not so far out there, in fact. Light can be converted to DC current with solar panels, electricity can be converted to magnetism as I did in my last article (https://www.instructables.com/id/DIY-Electro-Magnetic-Levitation/), in a microphone sound waves are converted to an electrical signal (by vibrating a magnet near a coil (http://hyperphysics.phy-astr.gsu.edu/hbase/audio/mic.html)), solar rays can even be focused and converted to heat in awesome devices like this! (http://www.gosunstove.com/) When we think about it, energy is all around us and can be harvested in an enumerable many of ways.

Download

Today, we are going to take a rather novel approach. We are going to build a device specifically designed to sense and capture a particular band of energy which is all around us.

The earth is magnetic and anyone who has ever used a compass knows this. Magnetic bodies in motion produce electricity, we can see this in any alternator, like the one in your car. So, therefore the earth is electric as well as magnetic, by definition.

Can we detect this energy? Yes, we sure can! Ever turn on a radio in the middle of nowhere and heard static? That is your radio picking up naturally occurring energy in the RF spectrum!

Can we use this energy to do work? Absolutely! This has been known for a long time. Crystal radios (https://en.wikipedia.org/wiki/Crystal_radio) have been around since before the 1930's and can run with no input energy other than the radio signal. Even when completely isolated, but from the atmosphere, a crystal radio will produce a voltage in the earpiece resulting in a sound (albeit and undesirable one).

Well, this is where it gets interesting...

Can we replicate this effect? Yea, and with modern components like the high quality crystals found in germanium diodes, we can even increase efficiency. By applying this concept as a Crystal Energy Receiver we can take advantage of a wide range of energetic frequencies rather than tuning in to just one.

Can we scale it up? Definitely. Things like micro germanium diodes, high efficiency antennas and compact contemporary capacitors make the components that are required to build a crystal receiver fit in the palm of your hand. While there may or may not be a more efficient way, this renewable energy solution is simple to employ and can be scaled up, or down, indefinitely.

Download

It sounds like we can build a Crystal Energy Receiver. Let's give it a shot...

Add Tip Ask Question

Step 1: What You'll Need

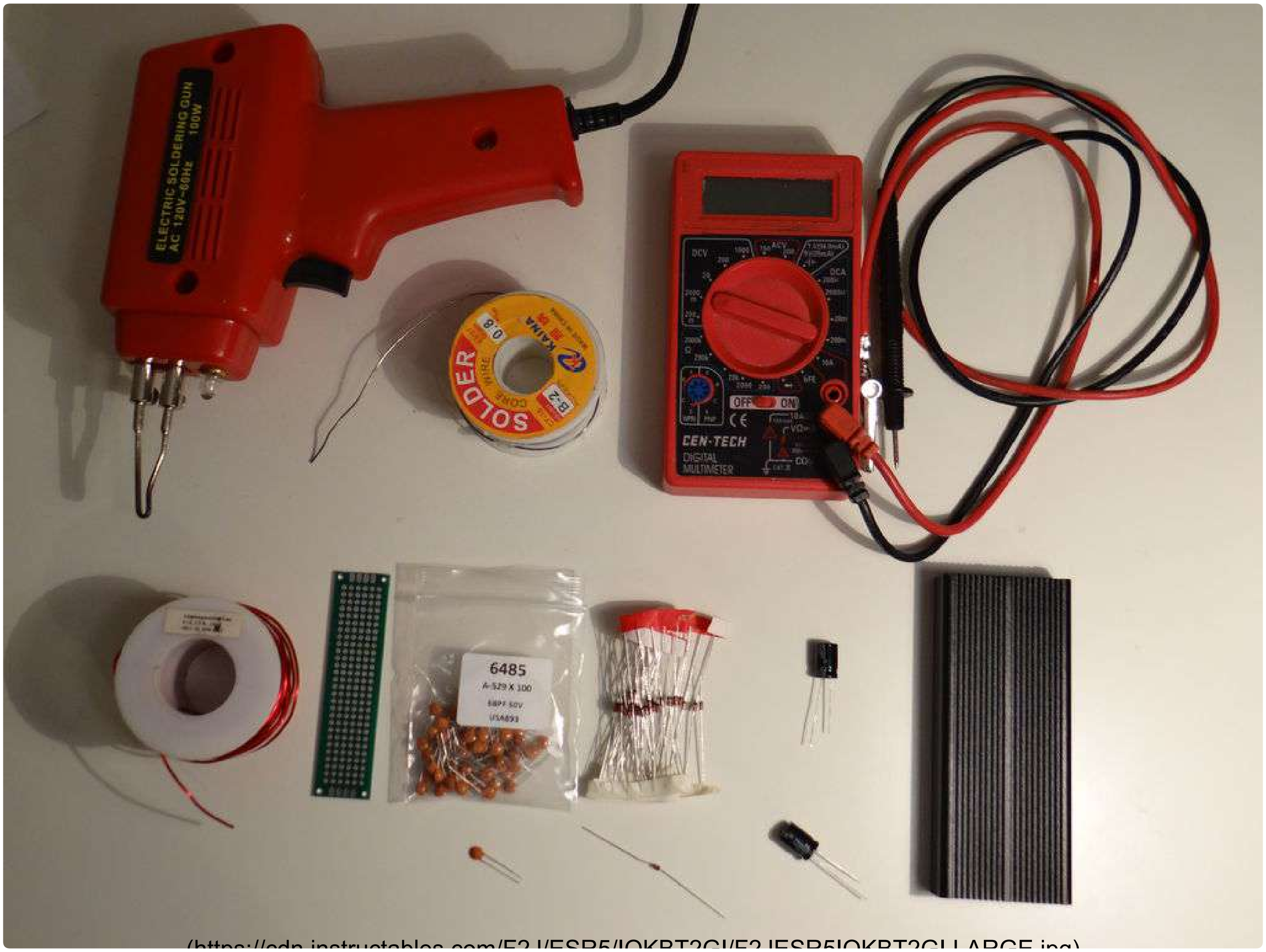
ELECTRIC SOLDERING GUN
AC 120V~60Hz 100W
SOLDER
WIRE
KAINA
0.8
CEN-TECH
DIGITAL MULTIMETER
OFF
ON
6485

(https://cdn.instructables.com/F2J/ESR5/IQKBT2GI/F2JESR5IQKBT2GI.LARGE.jpg)

Download

(https://cdn.instructables.com/F53/FHGM/JOKBT2GV/F53F

(https://cdn.instructables.com/FP2/G266/JOKBT2GG/FP2G

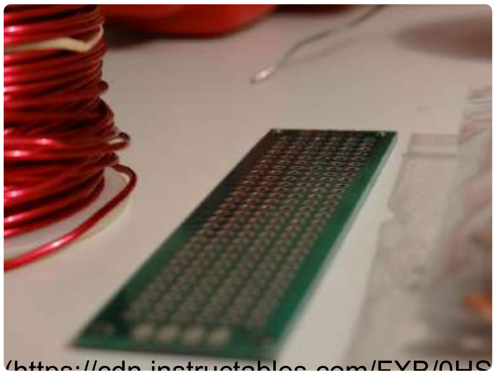

(https://cdn.instructables.com/FXB/0HS

(https://cdn.instructables.com/FXS/DZ

(https://cdn.instructables.com/FIH/FSH

Download

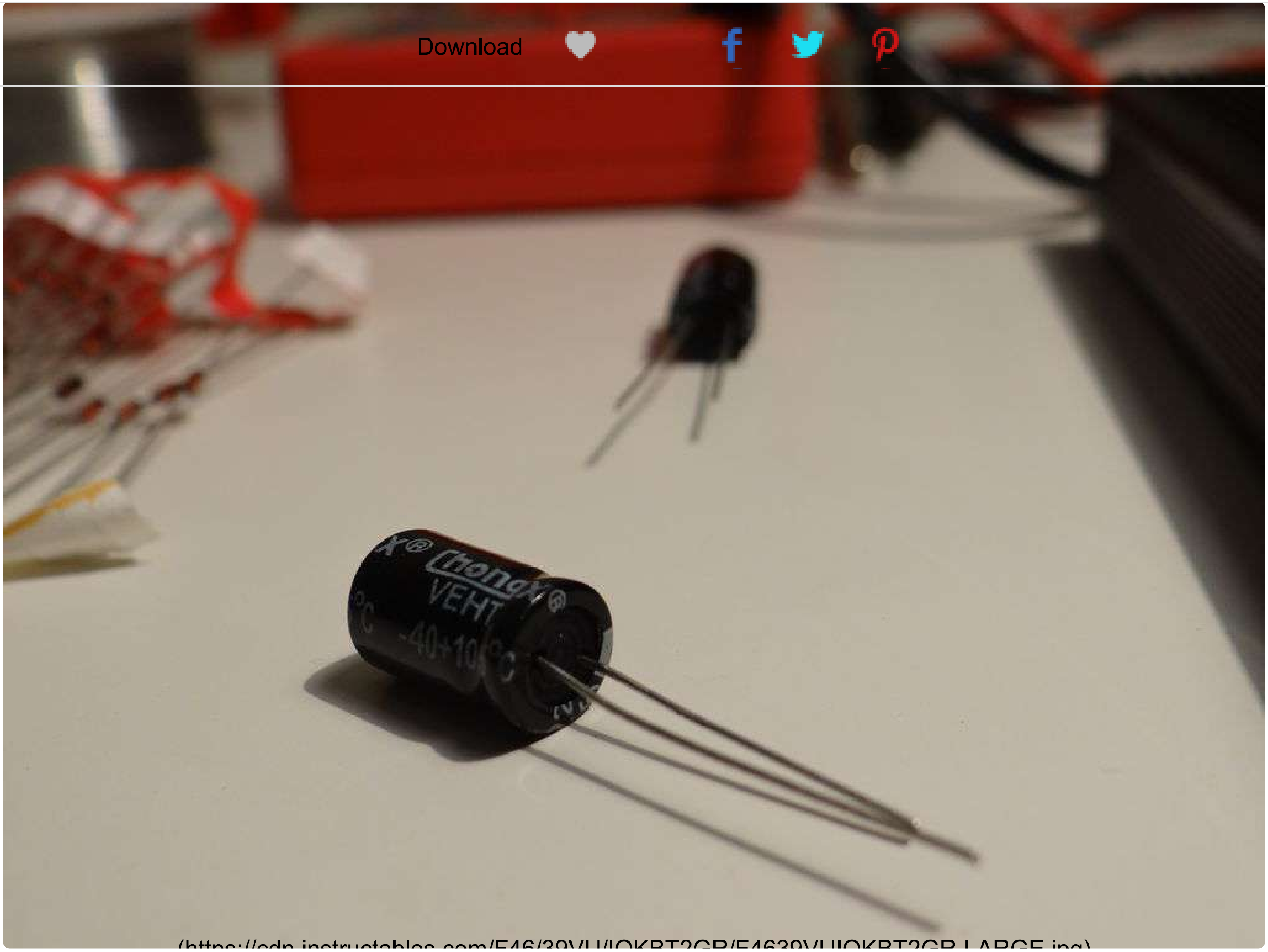
X®
Chongx®
VEHT
-40+10
C

(https://cdn.instructables.com/F46/39VU/IOKBT2GR/F4639VUIOKBT2GR.LARGE.jpg)

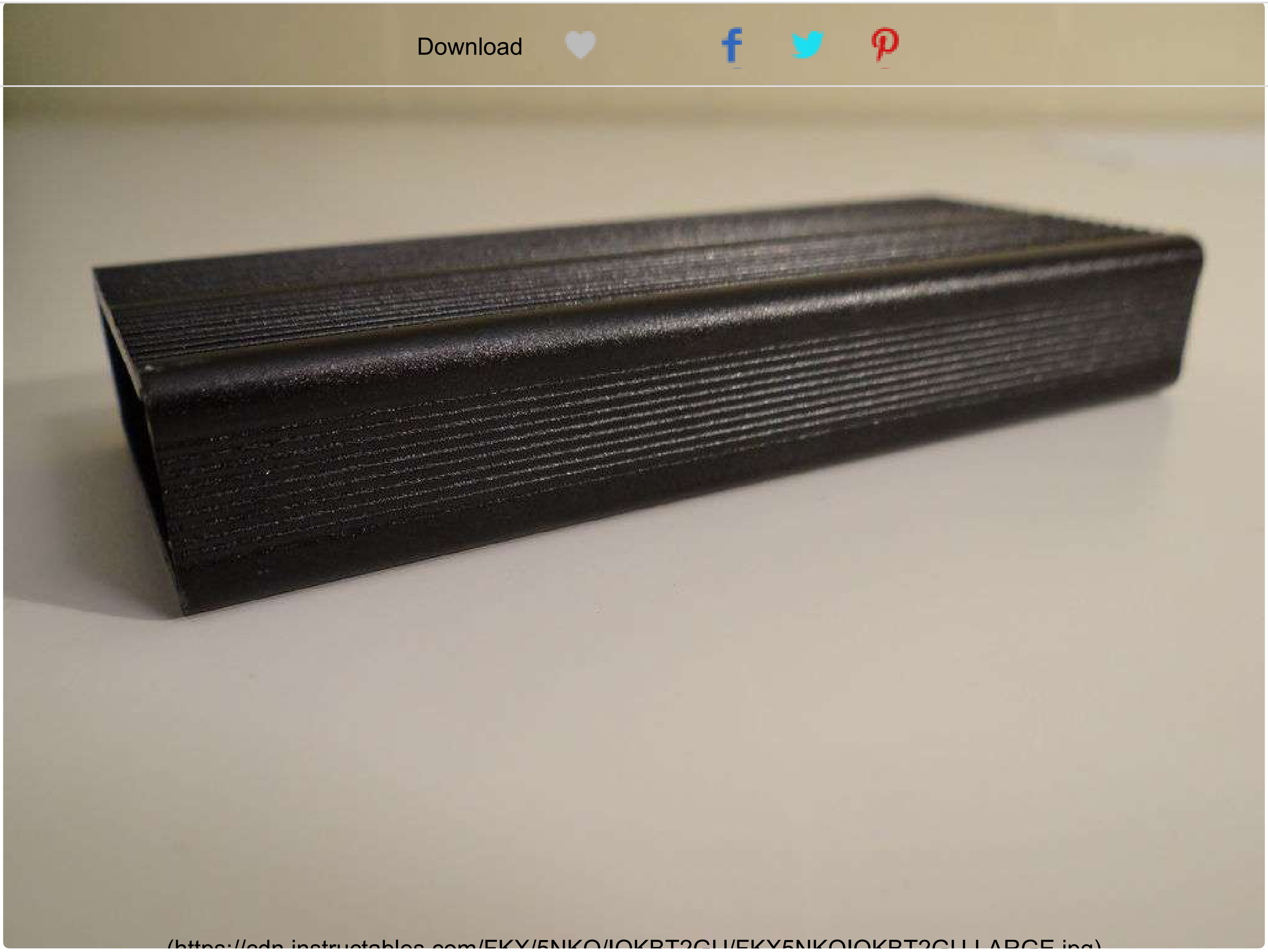

(https://cdn.instructables.com/FKY/5NKQ/IQKBT2GU/FKY5NKQIQKBT2GU.LARGE.jpg)

One of the reasons this particular renewable energy harvesting method is so viable is the relatively few and easy to obtain materials required.

The simplest crystal receiver design needs no power and can be built with only three parts: a coil, a crystal and a resistor. We're going to optimize that design in order to produce a cleaner and more reliable output signal by first polarizing the input amplitude, then rectifying and filtering the signal. Then we'll add an antenna, case and connections.

Get the circuit diagram here (http://www.drewpauldesigns.com/crystal-energy-receiver-kit.html)
Get the kit here (http://www.drewpauldesigns.com/crystal-energy-receiver-kit.html)

The parts for the circuit include:
(1) Circuit Board (http://www.drewpauldesigns.com/crystal-energy-receiver-kit.html)
(1) 10-18 gauge Copper Wire
(2-12+) Ceramic Capacitors (matched)
(2-6+) Electrolytic Capacitors (matched)
*note various types of capacitors can be used
(4) Germanium Crystal Diodes (1A+)

Total Unit Cost: +/- $0.40 (USD, scaled for volume of 1,000+ units)

In addition, you'll probably want to get:
(1) Project box (optional)
(1) Antenna (a loop antenna or elevated antenna is recommended and can be made with copper wire)

The tools you'll need are:
Soldering Iron/ Solder (optional)
Multimeter
Oscilloscope (http://www.seeedstudio.com/depot/DSO-Nano-v3-p-1358.html?cPath=63_65) (optional)

That's it. Yup, that's all. Once we've got it all, let's begin.

Add Tip Ask Question

Step 2: Build the Circuit

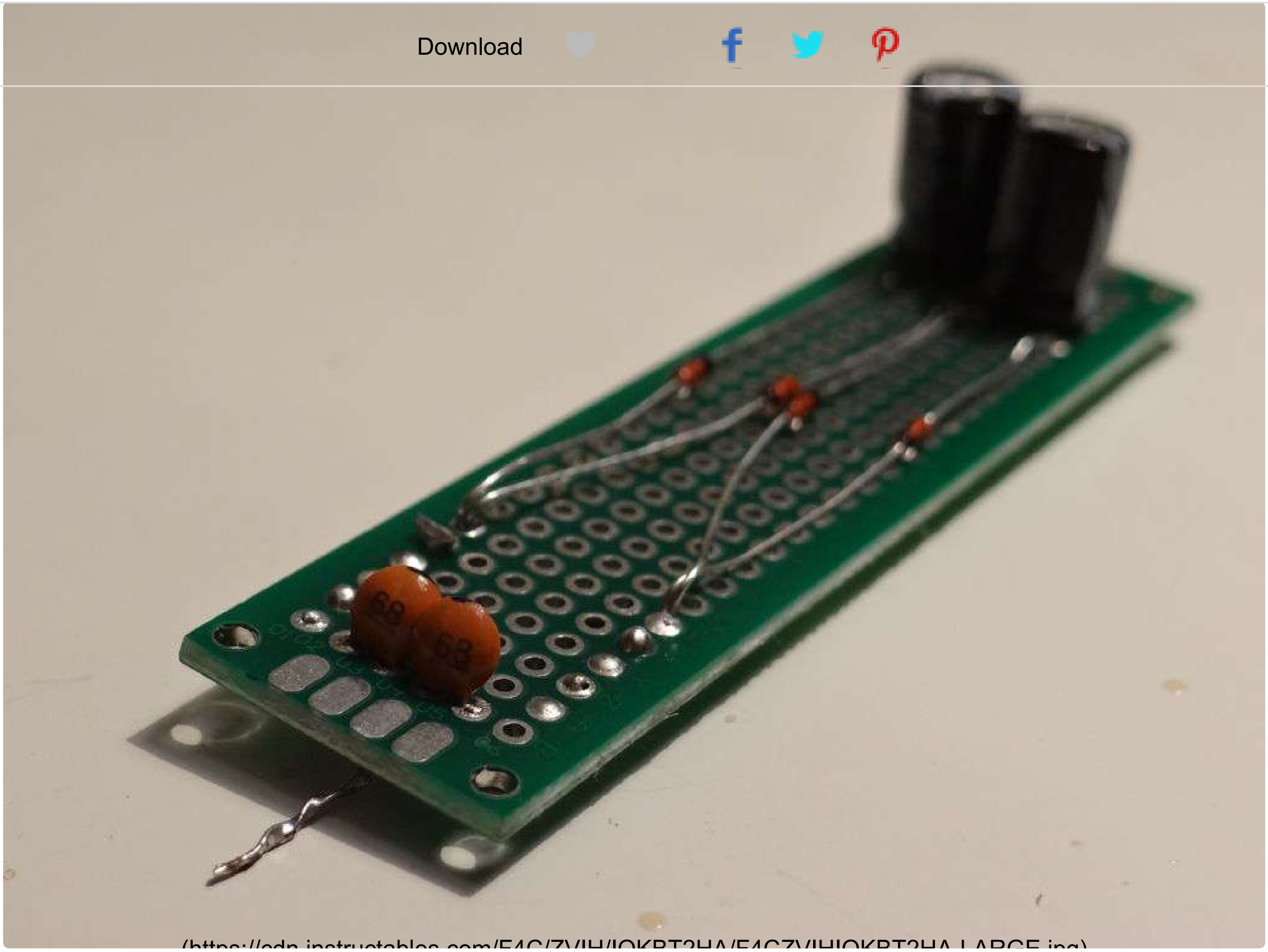

(https://cdn.instructables.com/F4G/ZVIH/IQKBT2HA/F4GZVIHIQKBT2HA.LARGE.jpg)

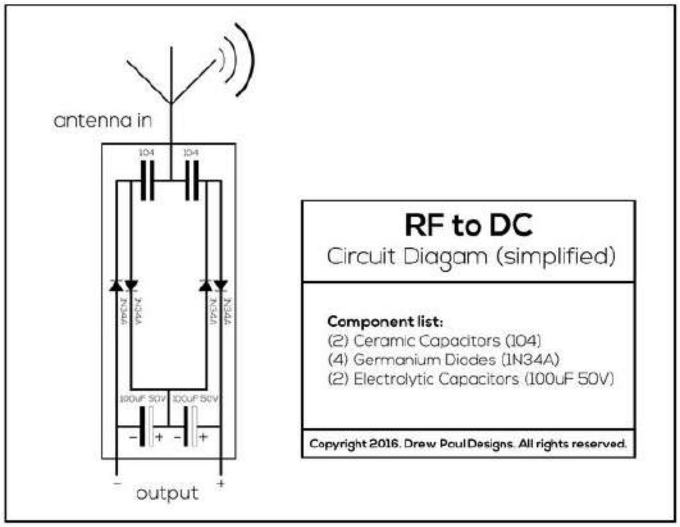
antenna in
104
104
1N34A
1N34A
1N34A
1N34A
100uF 50V
100uF 50V
-
+
output
RF to DC
Circuit Diagam (simplified)
Component list:
(2) Ceramic Capacitors (104)
(4) Germanium Diodes (1N34A)
(2) Electrolytic Capacitors (100uF 50V)
Copyright 2016. Drew Paul Designs. All rights reserved.

(https://cdn.instructables.com/FAW/ABS2/IQKBUKAD/FAV

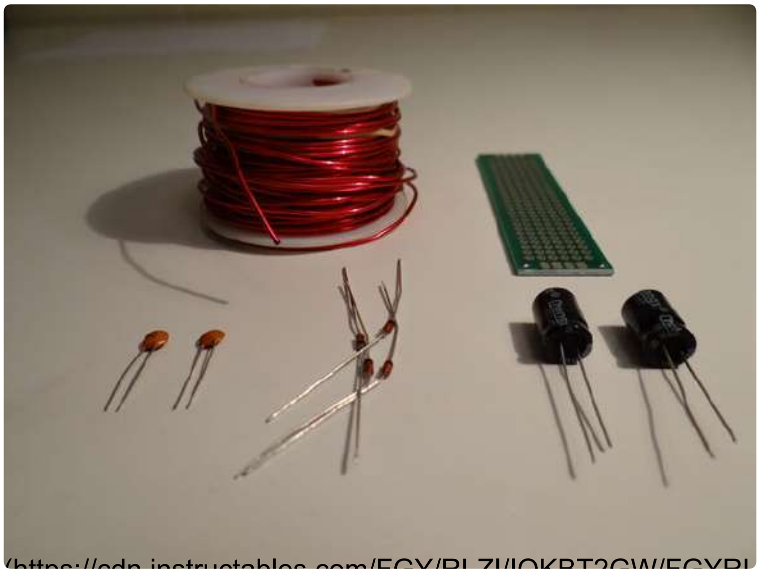

(https://cdn.instructables.com/FGY/PLZI/IQKBT2GW/FGYPL

Download

(https://cdn.instructables.com/FRH/JS

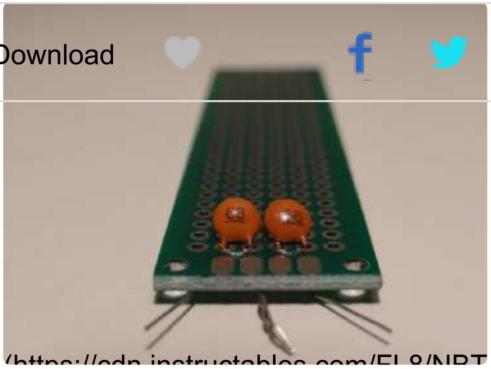

(https://cdn.instructables.com/FL8/NBT

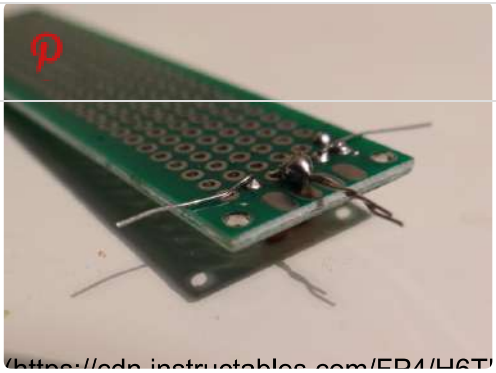

(https://cdn.instructables.com/FR4/H6T

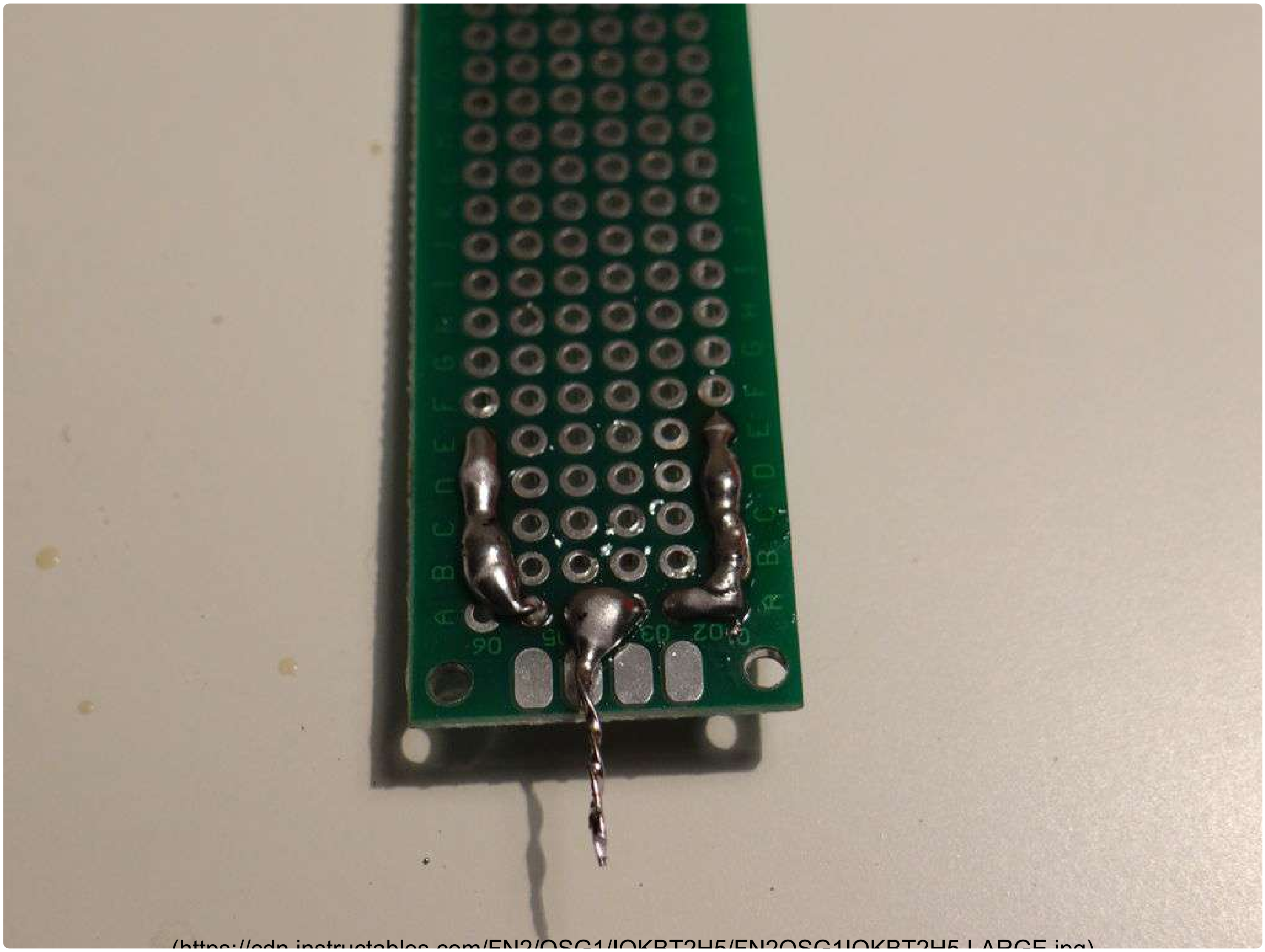

(https://cdn.instructables.com/FN2/QSC1/JOKBT2H5/FN2QSC1JOKBT2H5.LARGE.jpg)

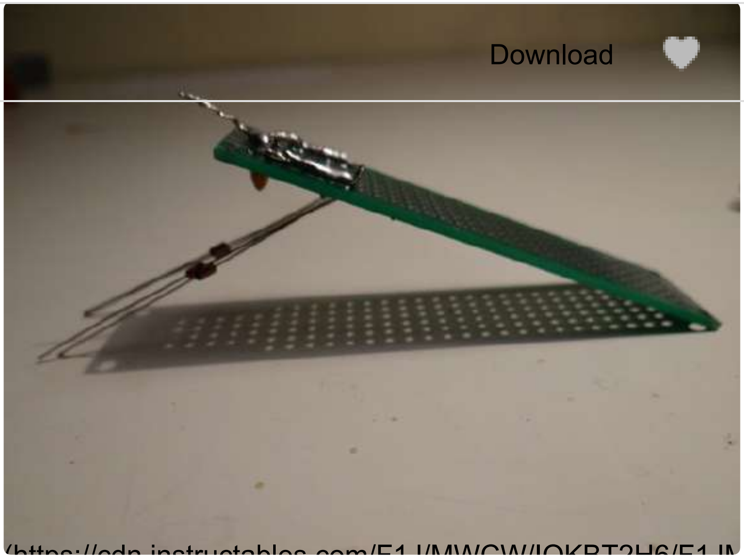

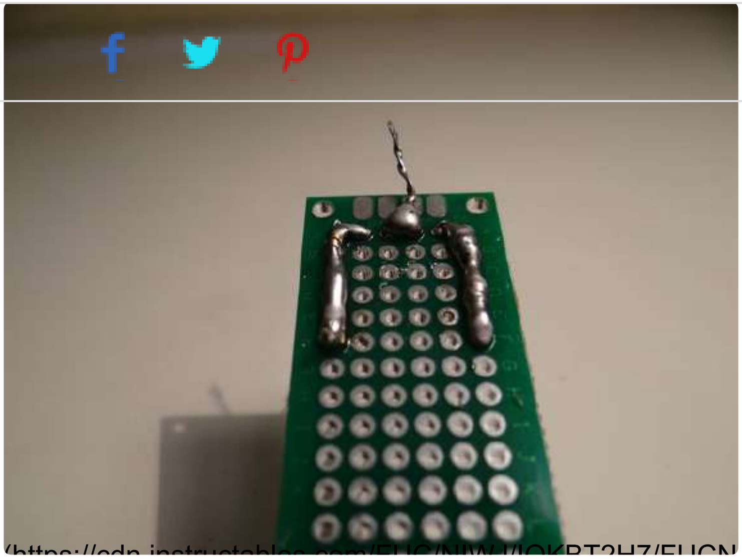

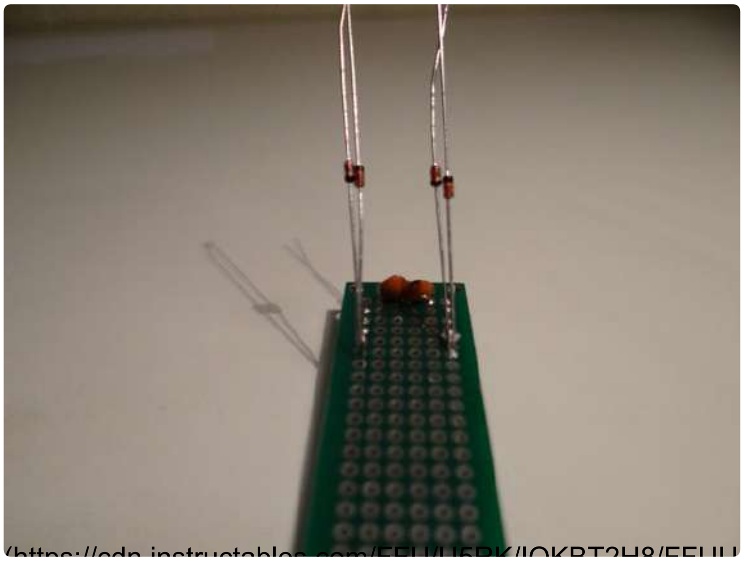

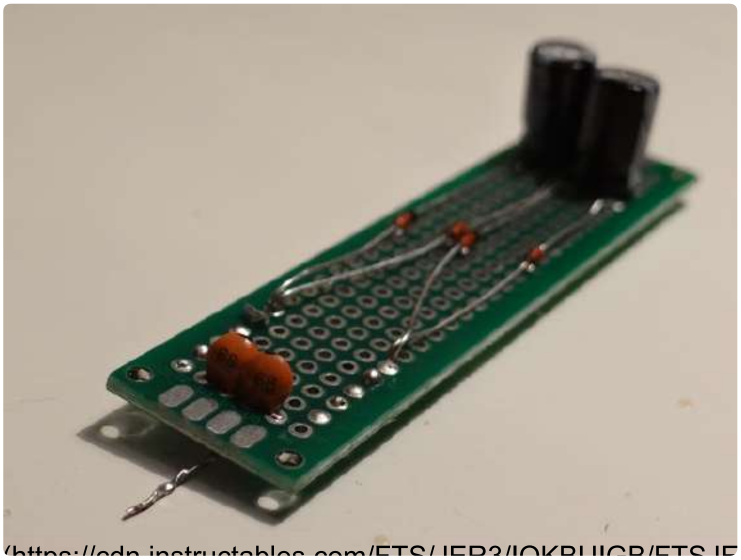

We're going to build the simplest version of this circuit variation in order to understand how each component interacts and as a proof of concept.

There are three simple systems at work in the circuit that are composed of capacitors, which store energy, and diodes that direct it.

Energy in the band of radio waves, among others, will vibrate a wire antenna on an atomic level, sending a discernible signal to its lead. This signal will then meet the junction between two ceramic capacitors wired in series. This junction will force positive charge from the wave to travel in one direction and negative charge in the other direction which, when collected again, makes the signal uniform and polar. Connecting the two capacitors in series creates leads on each end; the now positively charged side of one and the now negatively charged side of the other creates a two cell battery.

Download

This next stage of the circuit takes a signal with a net value of zero, adds the absolute values of the positive and negative amplitudes with respect to the origin and produces a positive integer. This concept can be thought of as taking:

(+1) + (-1) + (+1) + (-1) = 0

and converting it to:

| [(+1) + (+1)] | = 2
+
| [(-1) + (-1)] | = 2
= 4

Isn't math fun?

To each of these leads from our two capacitors in series, we will connect two crystal diodes, one facing each direction, to form what is called a bridge rectifier. A bridge rectifier is a configuration which will convert an alternating current to a direct one by cleverly rerouting the signal.

By connecting the bridge rectifier as shown in the circuit diagram, this direct current from the diodes then charges the electrolytic capacitors. This stage normalizes the amplitude, making the current constant and usable.

Components can easily be twisted together for testing and then soldered to a circuit board to secure.

Add Tip Ask Question

Step 3: Test and Optimize Your Circuit

Download

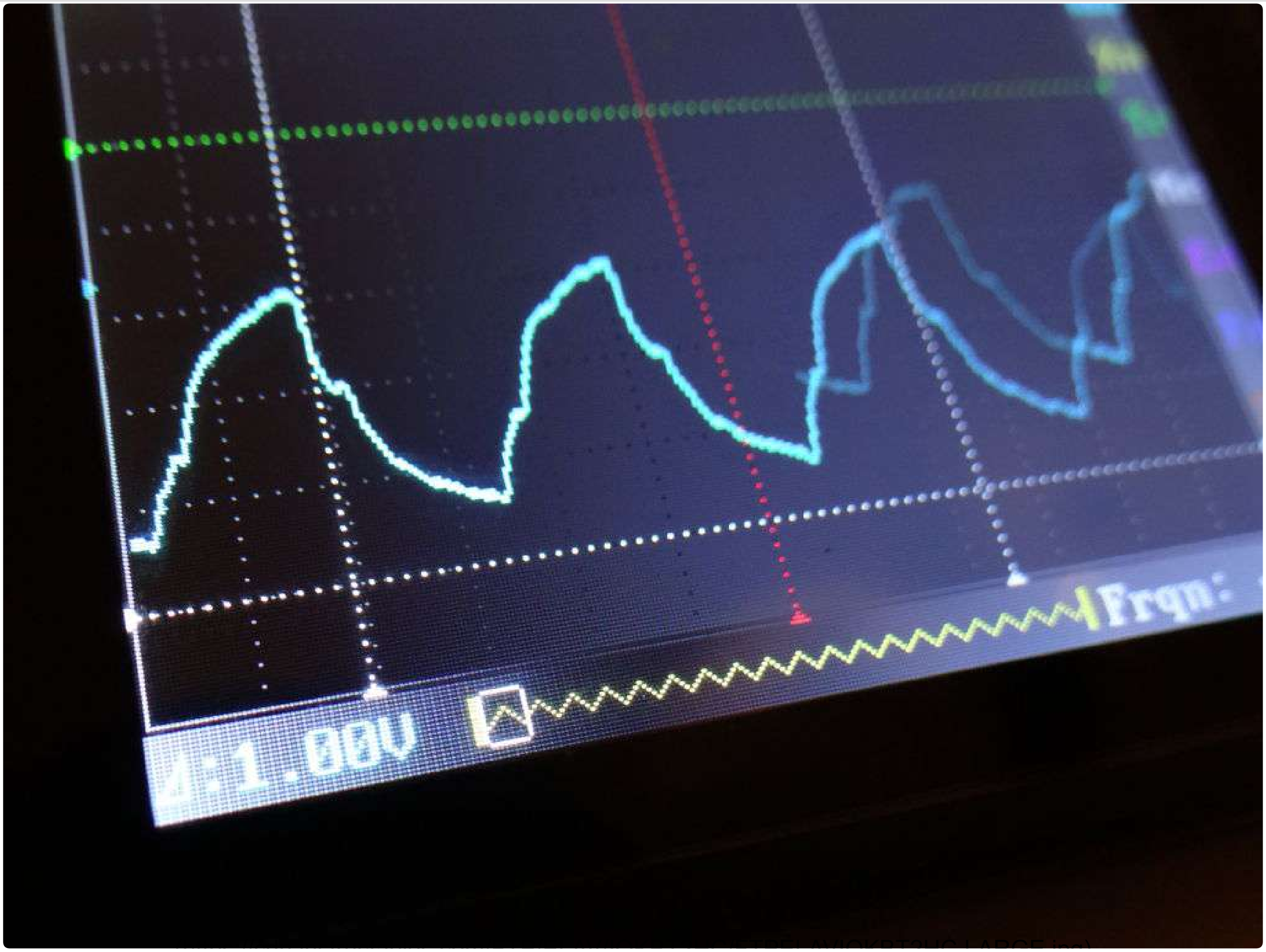

ETPBL4V/IOKBT3HC.LARGE.jpg)

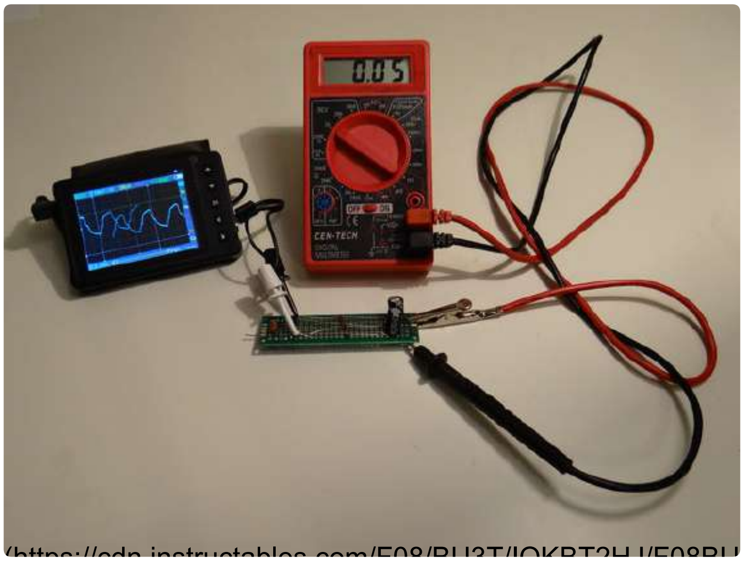
0.05
CEN-TECH

(https://cdn.instructables.com/F08/RU3T/IOKBT2HJ/F08RU

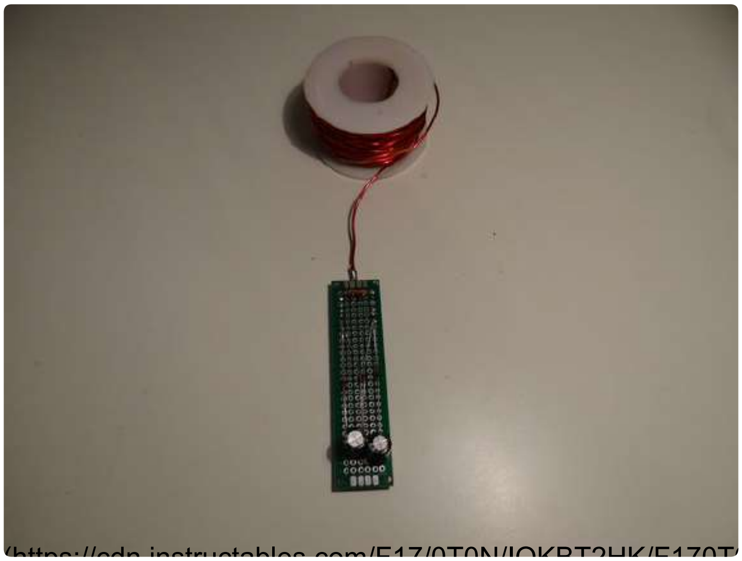

(https://cdn.instructables.com/F17/0T0N/IOKBT2HK/F170T

Download

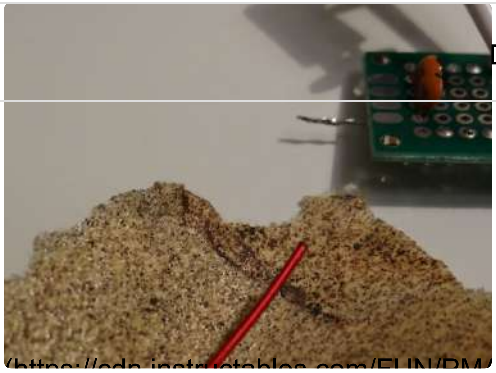

(https://cdn.instructables.com/FUN/RM

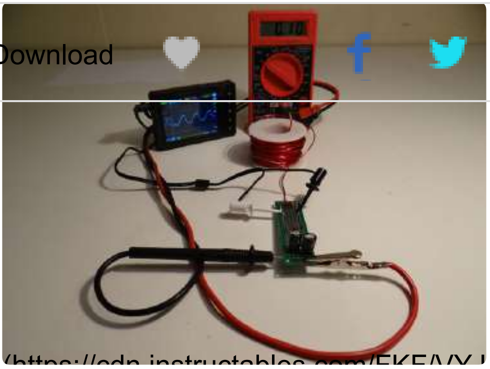

(https://cdn.instructables.com/FKE/VX

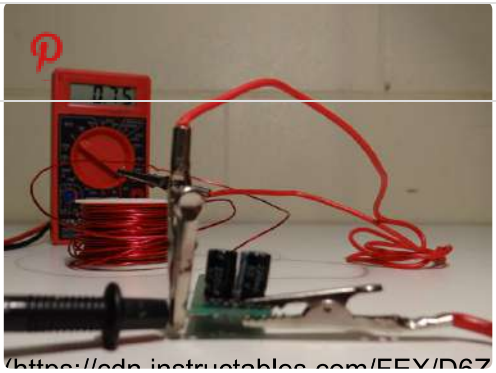

(https://cdn.instructables.com/FEX/D6Z

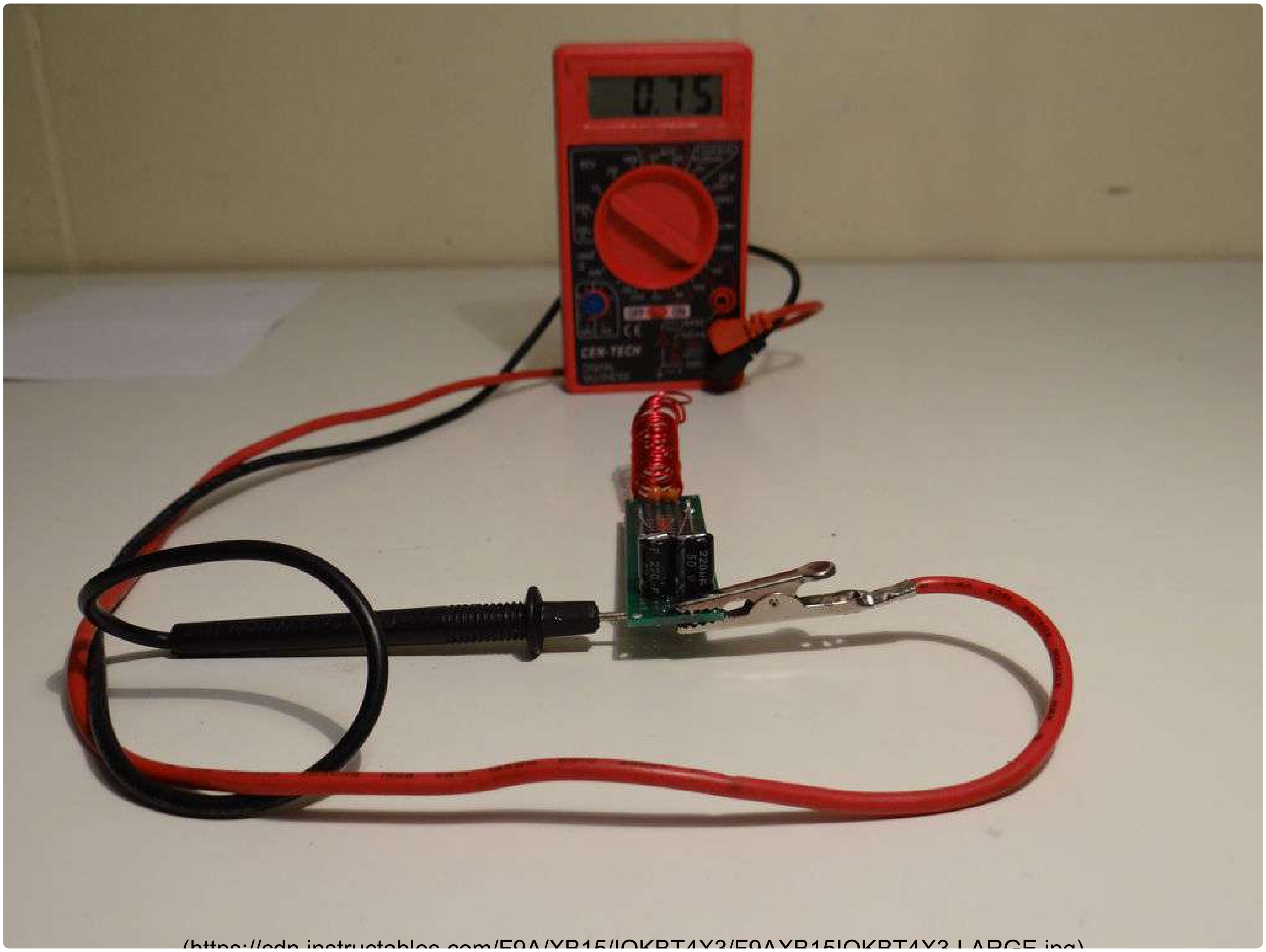

0.75
CEN-TECH
220µF
50V

(https://cdn.instructables.com/F9A/XB15/IOKBT4X3/F9AXB15IOKBT4X3.LARGE.jpg)

Download

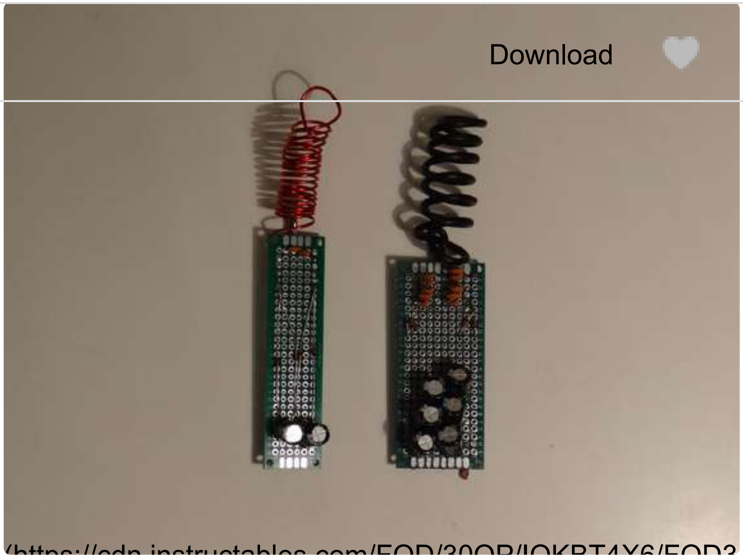
(https://cdn.instructables.com/FOD/30OR/IQKBT4X6/FOD3

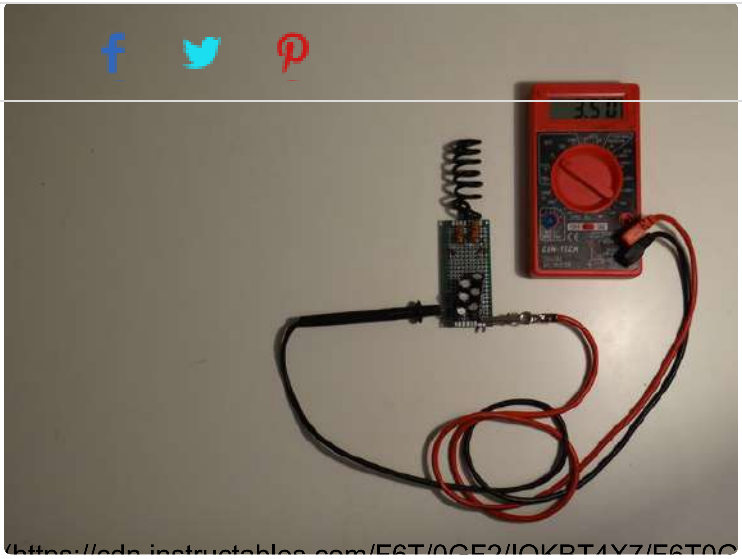

(https://cdn.instructables.com/F6T/0GF2/IQKBT4X7/F6T0G

To test and analyze our circuit, we'll be using a digital voltmeter and oscilloscope.

By connecting a voltmeter to the output, we'll immediately begin to see a small voltage climbing in the 10-100mV range. If not, we'll want to check our connections and make sure the circuit is not isolated from the environment by taking it outside to a clear area.

Then, by connecting an oscilloscope to the outside leads of our two ceramic capacitor bank, we will see the the polarized signal being captured from the air around us. We can then connect after the diodes to see our varying direct current and then to after the electrolytic capacitors to see a normalized, usable direct current at our output.

We can then optimize the input resistance in two ways. Firstly, we can add additional ceramic capacitors in parallel to our original two and make sure our soldered connections are consistent and thick in this area.

We can optimize the circuit's capacity by adding electrolytic capacitors in parallel to our original two which will allow this circuit to charge slightly when not in use. For this purpose, a charging circuit can also be added here in order to incorporate an optional battery bank.

We can optimize the antenna by attaching loops and coils of copper wire in various positions, store-bought antennas or by stringing some wire up to the highest point you can reach.

We don't have to stop there, either. We can now connect multiple circuits in series to increase voltage or in parallel to increase current. This can be done indefinitely.

Add Tip Ask Question

Step 4: Add a Case and Antenna

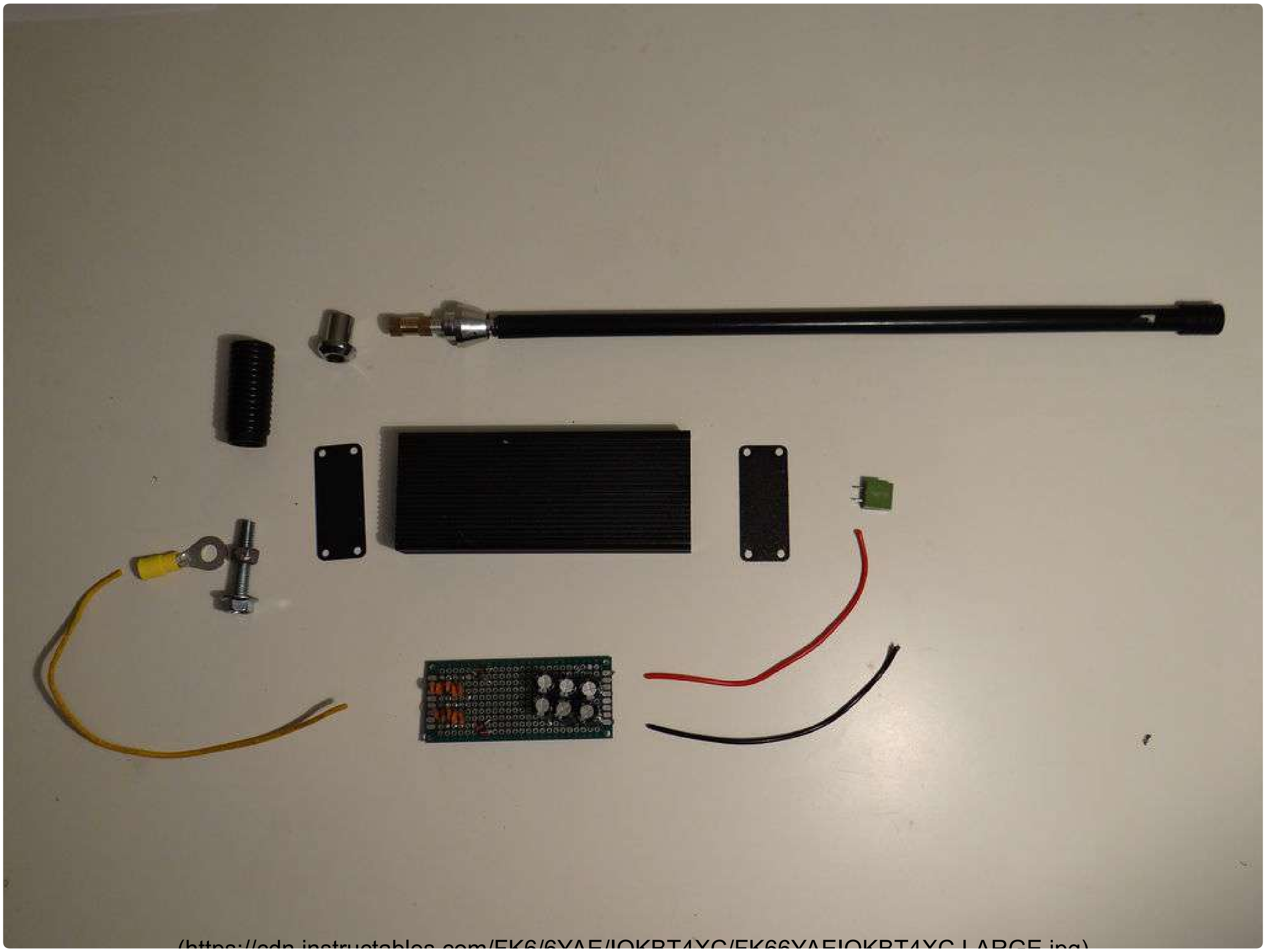

(https://cdn.instructables.com/FK6/6YAE/IOKBT4XC/FK66YAEIOKBT4XC.LARGE.jpg)

Download

(https://cdn.instructables.com/F62/QBPB/IQKBT4XA/F62Q

(https://cdn.instructables.com/F9J/W9B4/IQKBT4XD/F9JW

(https://cdn.instructables.com/F09/ES7

(https://cdn.instructables.com/F5Q/5N

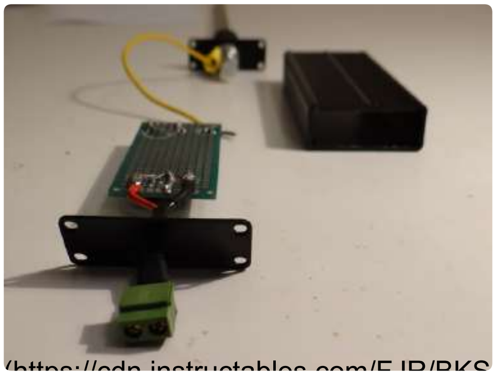
(https://cdn.instructables.com/FJR/BKS

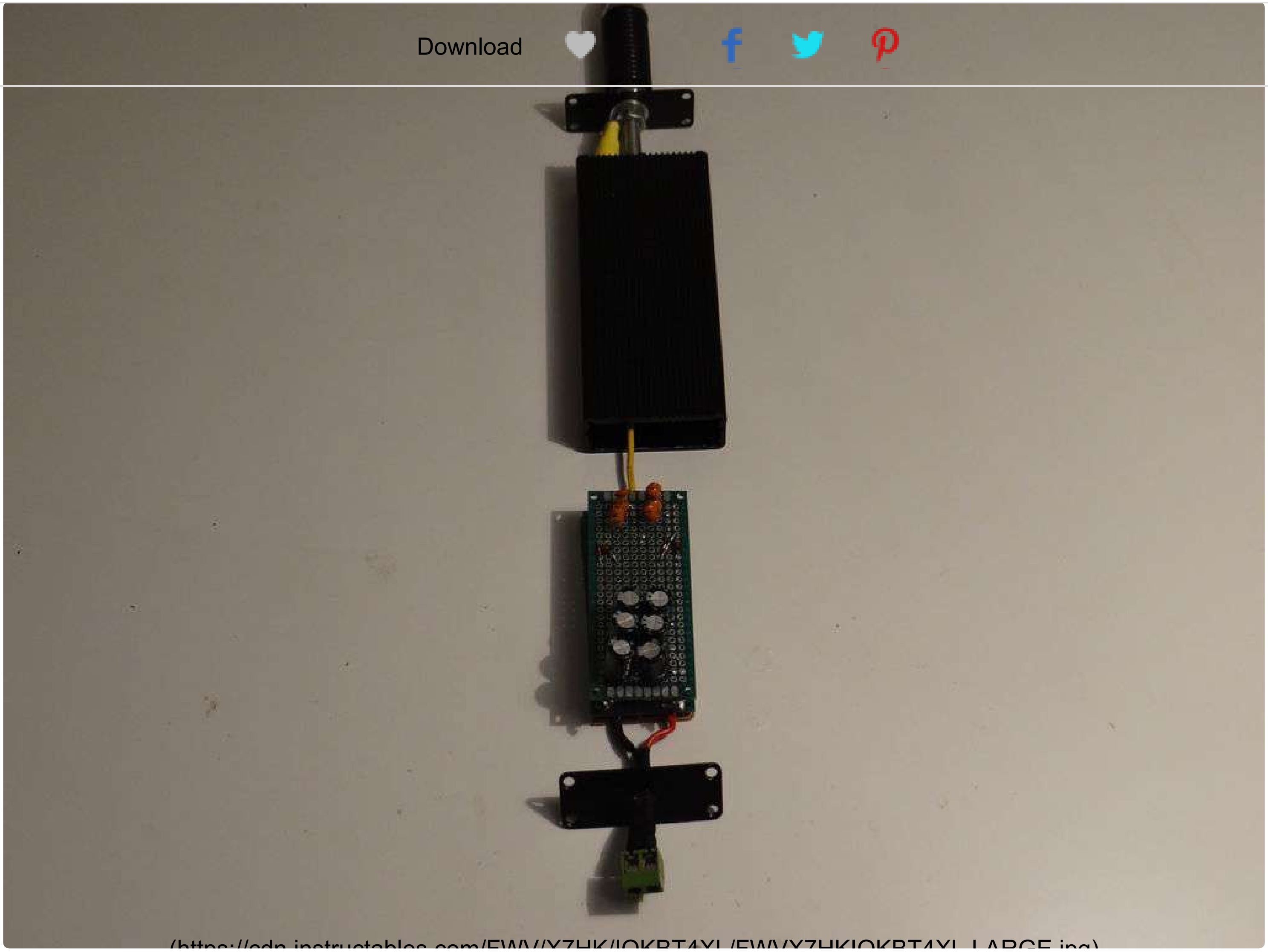

(https://cdn.instructables.com/FWV/YZHK/IQKBT4XL/FWVYZHKIQKBT4XL.LARGE.jpg)

(https://cdn.instructables.com/FTN/GNK

(https://cdn.instructables.com/FM8/8IF

(https://cdn.instructables.com/FD9/OGS

After choosing an antenna in the last step we'll now want to permanently wire it. Whether you choose a compact antenna for portability or a tall fixed antenna for power and range, we will wire it in the same manner according to the diagram in the previous step. Note that the input on the configuration here is grounded to the metallic case, and thus the users hand, and incorporation of longer antennas will require proportionally more substantial grounding.

We will then attach a terminal to the output to allow us to connect this circuit to an electrical device or charging circuit and battery bank.

Next, we will add a case, making sure to isolate exposed leads with non-conductive material especially if mounting in a conductive case. A piece of cardboard secured with glue is sufficient for the circuit's bottom and shrink wrap or electrical tape can be used in the case of any additional exposed leads. Drill two holes in your enclosure, one for the antenna or antenna lead and another for your output terminals. You can then insert your components, fasten the enclosure and your device is ready to use!

Add Tip Ask Question

Step 5: Your Crystal Energy Receiver Is Complete!

Download

(https://cdn.instructables.com/FEC/D3WF/IOKBT69X/FECD3WFIOKBT69X.LARGE.jpg)

(https://cdn.instructables.com/FX4/8V5I/IOKBT4Y6/FX48V5

(https://cdn.instructables.com/FLD/DTVI/IOKBT4YC/FLDDT

Download

Your Crystal Energy Receiver is now complete and ready to use!

I built a portable version, for proof of concept and demonstration purposes. However, you can go as big as you want- to passively charge batteries or run equipment remotely; or go as small as you want- to power sensors, RFID devices, small electronics and more.

I used this harvested energy to easily power a low-consumption quartz clock, a digital chronograph with integrated circuits and LCD and was even able to momentarily rotate a small dc motor.

Because of its simplicity this device is a durable, efficient and reasonably effective method of harvesting radiant energy in a simple, replicable and sustainable way. I humbly hope that the contributions made here, and by those reading, can be one day used by people worldwide to conveniently capture free energy.

Thanks for checking out my project and I look forward to seeing everyone's variations, suggestions and improvements!

Add Tip Ask Question

advertisement

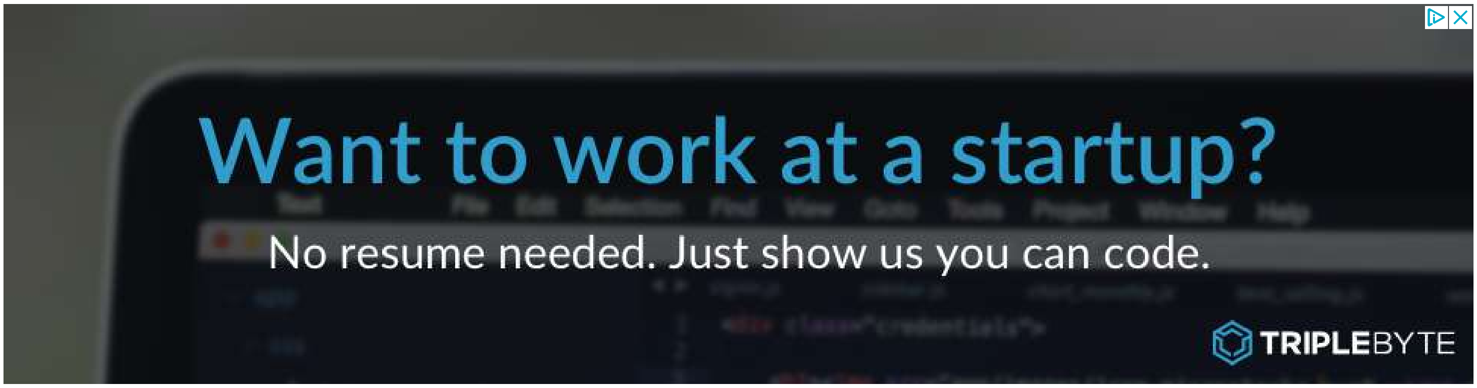

Want to work at a startup?

No resume needed. Just show us you can code.

TRIPLEBYTE

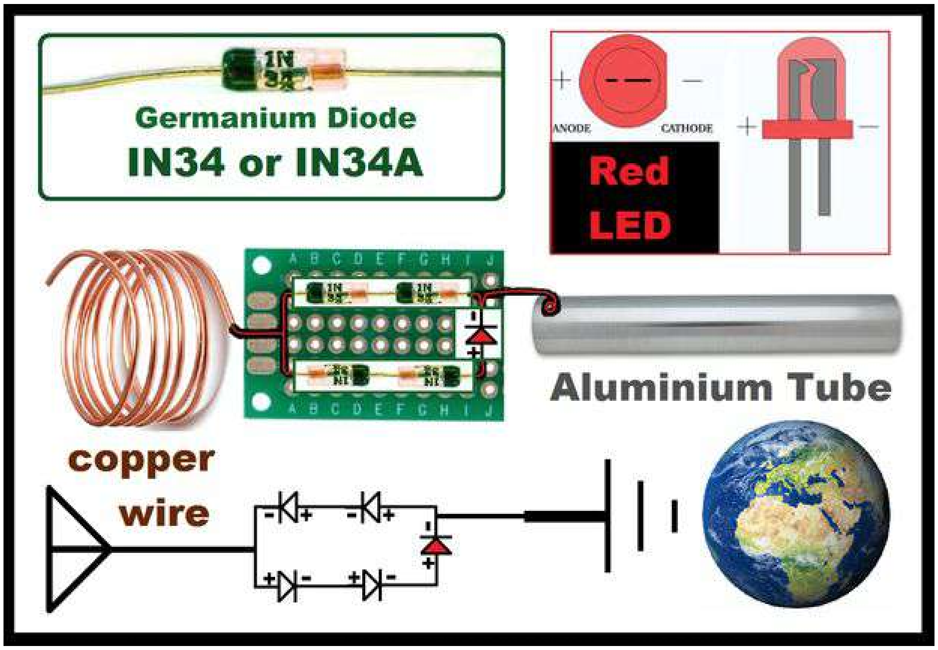
1N 34
Germanium Diode
IN34 or IN34A
ANODE
CATHODE
Red
LED
A B C D E F G H I J
Aluminium Tube
copper
wire

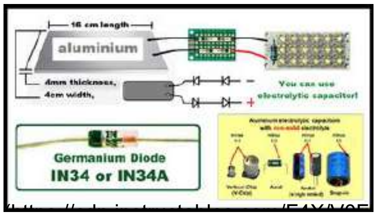
16 cm length
aluminium
4mm thickness, 4cm width,
You can use electrolytic capacitor!
Germanium Diode
IN34 or IN34A

(https://cdn.instructables.com/F4X/Y0F

1N 34
Germanium Diode
IN34 or IN34A

(https://cdn.instructables.com/FUD/RDY8/IIRRGZ3R/FUDR

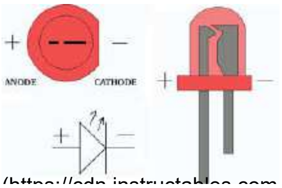
ANODE
CATHODE

(https://cdn.instructables.com

(https://cdn.instructables.com/F9F

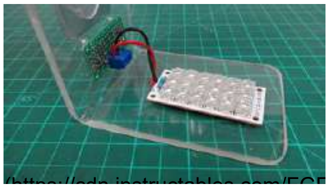

(https://cdn.instructables.com/FCF

advertisement

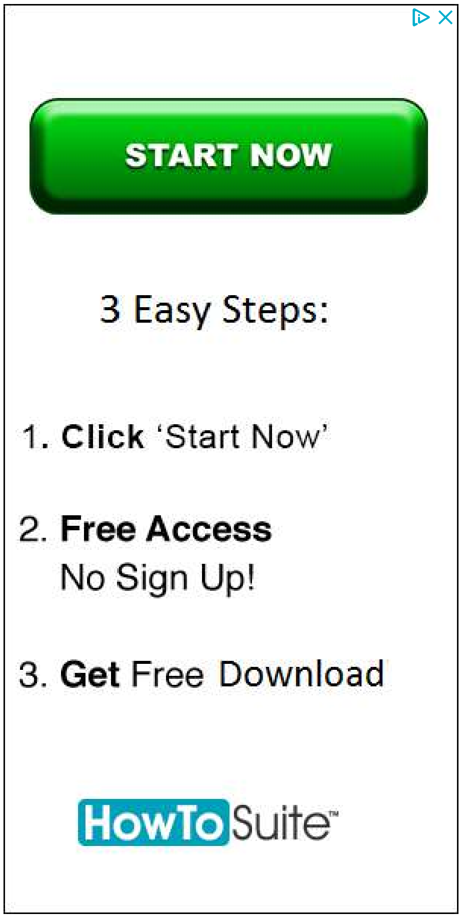

START NOW

3 Easy Steps:

1. Click 'Start Now'

2. Free Access
No Sign Up!

3. Get Free Download

HowToSuite

About This Instructable

2,351 views

46 favorites

License:

(cc) BY-NC-SA

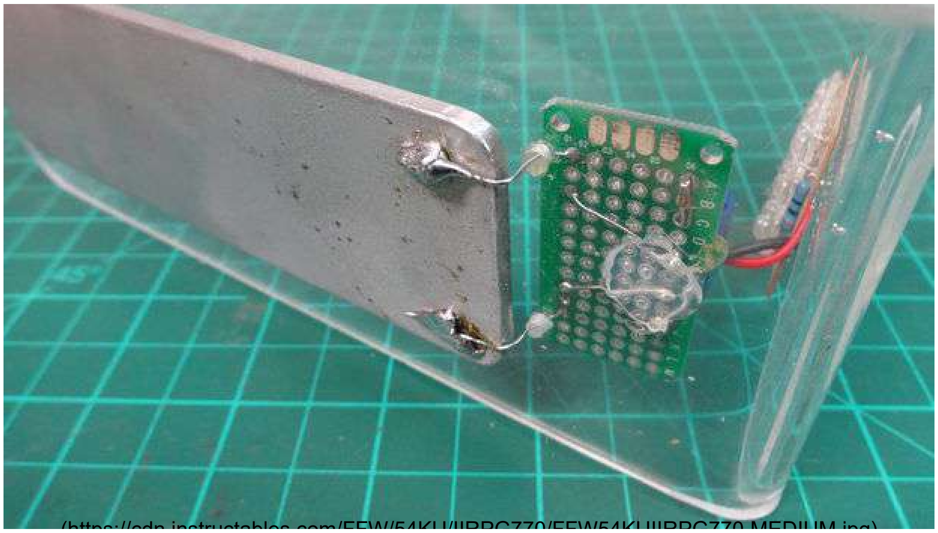

(https://cdn.instructables.com/FEW/54KU/IIRBC770/FEW54KUIIRBC770.MEDIUM.jpg)

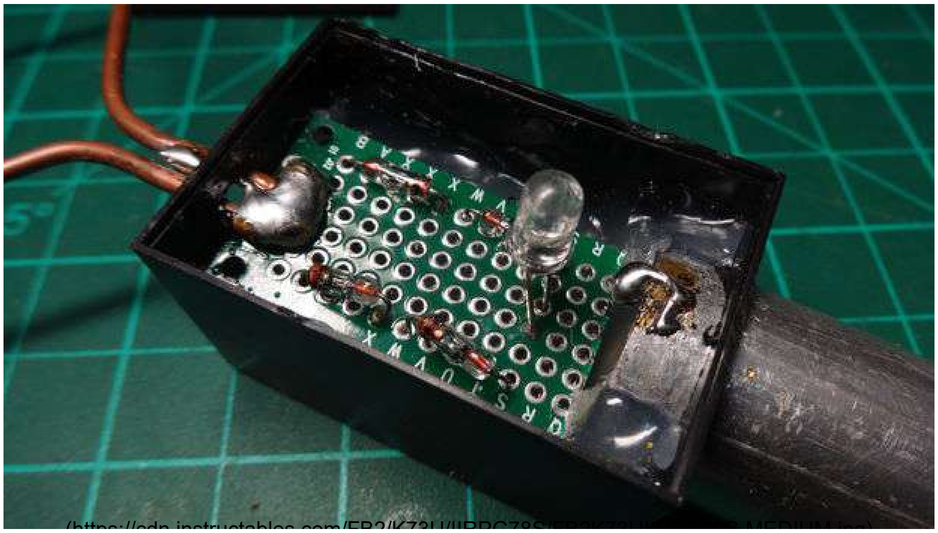

(https://cdn.instructables.com/FB2/K73U/IIRBC78S/FB2K73U MEDIUM.jpg)

How to Make a Electrostatic RF Detector or Ghost Detector, Hunting Solution. Much cheaper and easier than you may think For Under $5 ! A simple description is that IN34 or IN34A RF Germanium Diodes for Crystal Radio can take "Ambient" or "Radio Waves" to turns them into DC energy. I would like to show the Tesla Coil Transmitter, where I made an homemade "RF Detector" this is the simplest detector! without any batteries or hardships circuit, only with IN34 diodes and Ether to Earth!.. Stay Tuned and Enjoy! Please click the "like" button, subscribe my channel, and share this with your friends... To support my "work" every little bit helps! Watch the "ADS" THANKS

- See the original video https://www.youtube.com/watch?v=cSxqeGVC9t0

Step 1:

GEORGE CHANIOTAKIS (/member/GEORGE+CHA
(http://https://www.youtube.com/
Follow 84
(/member/GEORGE+CHANIOTAKIS/)

Bio: A big hello to all mankind! This is my channel I hope you can find here various interesting things on in Renewable Energy Technologies and ... More » (/member/GEORGE%20CHANIOTAKIS/)

More by GEORGE CHANIOTAKIS:

Synchro
Diy
Make

(/id/MAKE-PERMANENT-MAGNET-SYNCHRONOUS-MOTOR-GENERATOR/)

(/id/Make-Wind-Turbine/)

(/id/High-Voltage-Die/)

advertisement

You Should Never Shop On Amazon Without Using This Trick – Here's Why

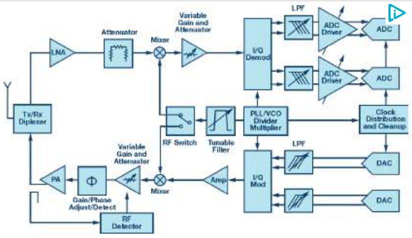

LNA
Attenuator
Mixer
Variable Gain and Attenuator
I/Q Demod
LPF
ADC Driver
ADC
Tx/Rx Diplexer
RF Switch
Tunable Filter
PLL/VCO Divider Multiplier
Clock Distribution and Cleanup
PA
Gain/Phase Adjust/Detect
RF Detector
Amp
I/Q Mod
DAC

Get Complete RFIC Guide

ANALOG DEVICES

Download all Analog Devices RFIC Parts for Each Component in RF Signal Chain.

Comments

We have a be nice comment policy.
Please be positive and constructive.

I Made it! Add Images Post Comment

LazaroG4 (/member/LazaroG4/) 2017-11-14 Reply

I can't download and I was doing it before. I now need to go premium. This why i do not give any of my own files for the idea i have posted for free on You Tube. sorry for the not using of file and I have integrity and will not copy and paste as many do for free.

More Comments

How to Make a Electrostatic RF Detector or Ghost Detector by

GEORGE CHANIOTAKIS (/member/GEORGE+CHANIOTAKIS/) in leds (/technology/leds/)

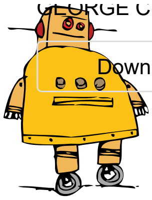

Download

(/id/How-to-Make-a-Electrostatic-RF-Detector-or-Ghost-D/) 1 Steps

Newsletter

Let your inbox help you discover our best projects, classes, and contests. Instructables will help you learn how to make anything!

enter email | I'm in!

+ Collection | I Made it! | Favorite | Share

About Us

Who We Are (/about/)

Advertise (/advertise/)

Contact (/about/contact.jsp)

Jobs (/community/Positions-available-at-Instructables/)

Help (/id/how-to-write-a-great-instructable/)

Find Us

- Facebook (http://www.facebook.com/instructables)
- Youtube (http://www.youtube.com/user/instructablestv)
- Twitter (http://www.twitter.com/instructables)
- Pinterest (http://www.pinterest.com/instructables)
- Google+ (https://plus.google.com/+instructables)

Resources

For Teachers (/teachers/)

Residency Program (/pier9residency)

Gift Premium Account (/account/give?sourcea=footer)

Forums (/community/?categoryGroup=all&category=all)

Answers (/tag/type-question/?sort=RECENT)

Sitemap (/sitemap/)

advertisement

AdChoices

JOHN LEWIS CLEARANCE.
MORE OF WHAT YOU LOVE FOR LESS.
NOW ON.
Shop now >
John Lewis

Terms of Service (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21959721) |
Privacy Statement (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21292079) |
Legal Notices & Trademarks (http://usa.autodesk.com/legal-notices-trademarks/) | Mobile Site (https://www.instructables.com)
(http://usa.autodesk.com/adsk/servlet/pc/index?id=20781545&siteID=123112)

© 2017 Autodesk, Inc.

Build a Homebrew Radio Telescope

Explore the basics of radio astronomy with this easy to construct telescope.

Mark Spencer, WA8SME

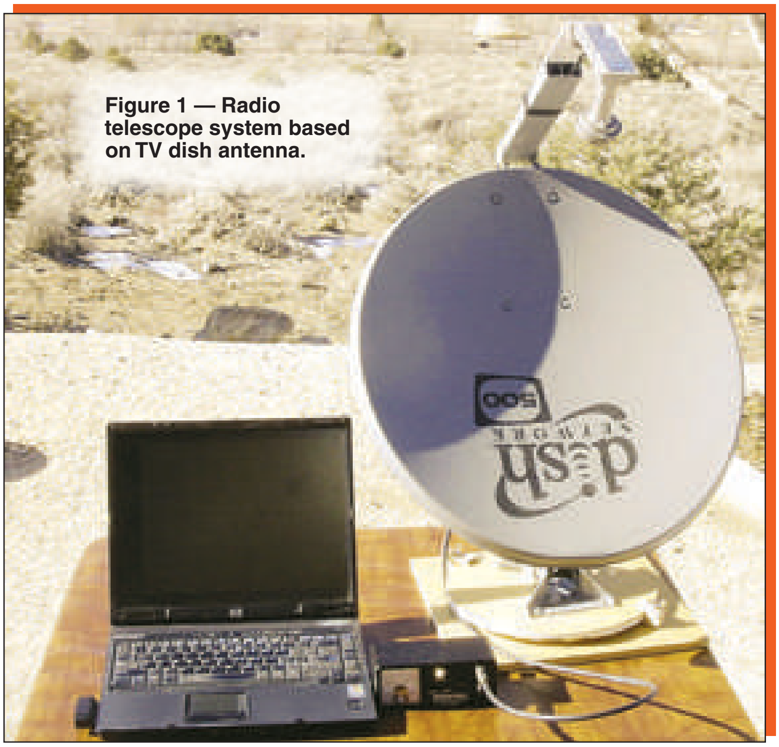

Figure 1 — Radio telescope system based on TV dish antenna.

There are many ham radio related activities that provide a rich opportunity to explore and learn more about the science of radio. One of those opportunities is radio astronomy.

All matter emits radio frequency (RF) energy dependent on the temperature and makeup of the matter, including the matter in space. The foundation of radio astronomy is to study the heavens by collecting and analyzing the RF energy that is emitted by bodies in space, very much as optical astronomers use light energy collected by telescopes. It sounds complicated. While professionals use very sophisticated and expensive equipment, you can, with some simple equipment and a little investment, build a radio telescope that will allow you to learn and explore the fundamentals of radio astronomy.

A Homemade Radio Telescope

In this article, I will build on an existing design of a radio telescope made from one of those ubiquitous TV dish antennas that you see around your neighborhood. The radio telescope (RT) project described here can easily be reproduced. Although this is not a fully capable RT, it can provide a wonderful learning opportunity for you, or perhaps students in your local school.

Figure 1 shows the radio telescope set up. The major components include a modified TV dish antenna mounted on a wooden support structure to allow pointing the antenna, a commercial satellite signal strength detector that displays the signal strength of signals collected by the dish on a meter and an interface that converts the signal strength into a amplitude modulated tone. The tone is fed into a computer sound card and finally a computer and software graphically displays the signal strength as a function of time.

The TV dish modifications are structural, and any available TV dish system can be used. The signal strength detector costs between $40 and $65 and is widely available from Web retailers. The interface circuit, which will be described shortly, is easily duplicated and costs approximately $20. Finally, the display software is free.

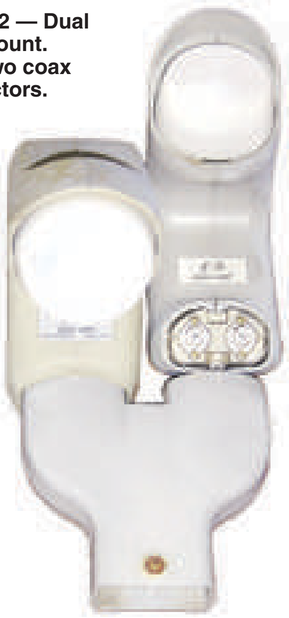

Figure 2 — Dual LNB mount. Note two coax connectors.

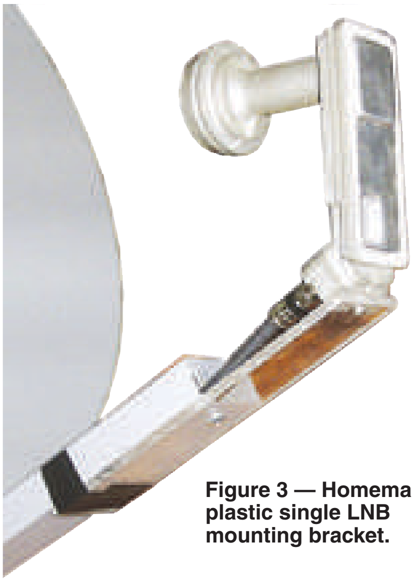

Figure 3 — Homemade plastic single LNB mounting bracket.

What it Can Do

The following is just a sample of what you can do with this simple RT:

- Use the sun to study and determine the beamwidth of the dish and verify the mathematic formula that is used to predict dish antenna performance.
- Measure the radiation intensity of the Sun and perhaps detect changes in solar activity.
- Measure the relative changes in the surface temperature of the moon.
- Learn about and explore a common radio astronomy collection technique called the *drift scan*.

From June 2009 QST © ARRL

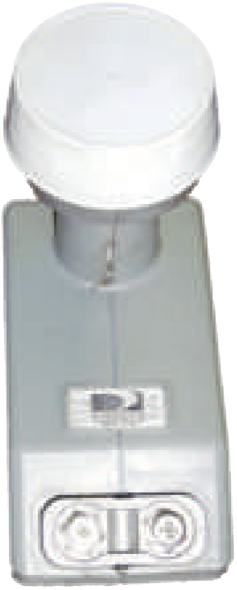

Figure 4 — Dual coax connector configured LNB. Terminate one connector with a dummy load.

- Explore the fundamental principle of energy emission as a function of temperature by detecting the relative differences between the temperatures of emitting bodies.
- Detect satellites parked along the Clarke Belt in geosynchronous orbit and illustrate how crowded space has become.

Figure 5 — CM satellite signal strength meter.

- Detect the Earth's rotation around the Sun and the Earth's spin on its axis by comparing daily drift scans of the horizon.

Antenna Subsystem

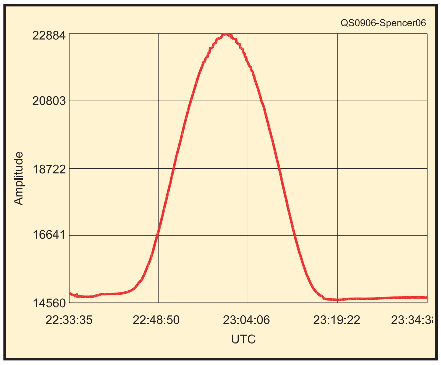

Figure 6 — *SkyPipe* screen showing antenna response.

The basic RT system is based on the "Itty-Bitty" design that is described in two Web pages.[1,2] The TV dish is an offset 18 inch dish that has down converter(s) mounted at the focal point of the dish. The down converter is called a low noise block (LNB). The LNB is a preamplifier/down converter that converts the satellite signals from around 12 GHz down to around 2.4 GHz. Most modern dishes have two or more LNBs to access more than one TV satellite at a time without changing the pointing of the dish (Figure 2). The LNBs are mounted to share the focal point of the dish. Since only one LNB is required for the RT, I made a minor adjustment to the published Itty-Bitty design to position the single LNB at the dish focal point. Mounting the single LNB at the focal point really helps in pointing the antenna.

I used the existing LNB housing and mounting bracket as a template to determine the distance between the edge of the mounting arm to the mounting hole of the LNB. I then used a piece of plastic to fabricate a new mounting bracket for the LNB as shown

---

[1] Notes appear on page 45.

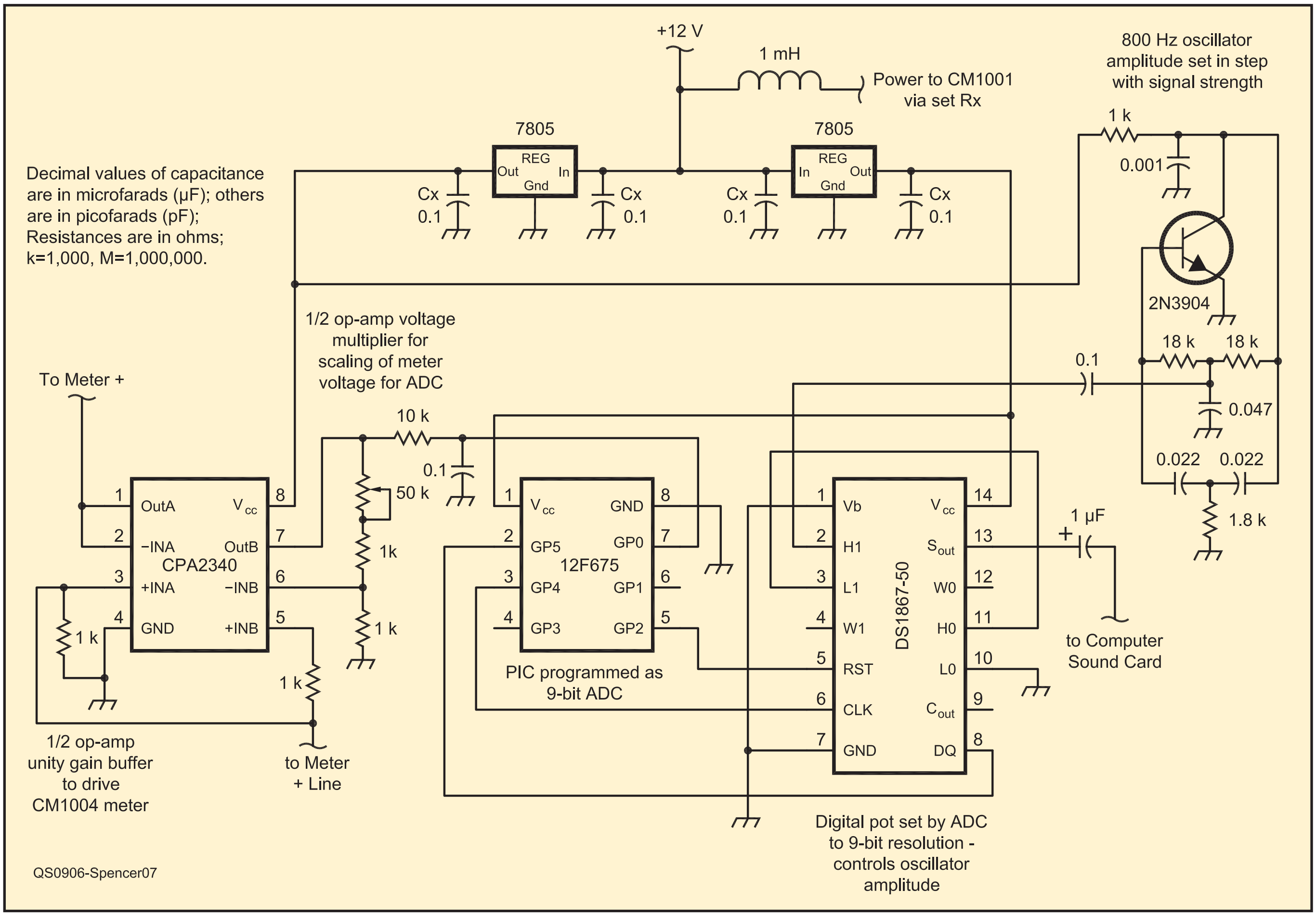
+12 V
1 mH
Power to CM1001 via set Rx
800 Hz oscillator amplitude set in step with signal strength
7805
REG
Out
In
Gnd
Cx 0.1
Decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); Resistances are in ohms; k=1,000, M=1,000,000.
1 k
0.001
2N3904
18 k
0.1
0.047
0.022
1.8 k
1/2 op-amp voltage multiplier for scaling of meter voltage for ADC
To Meter +
10 k
50 k
1k
1 k
OutA
Vcc
−INA
OutB
CPA2340
+INA
−INB
GND
+INB
Vcc
GND
GP5
GP0
12F675
GP4
GP1
GP3
GP2
PIC programmed as 9-bit ADC
Vb
H1
L1
W1
RST
CLK
GND
Vcc
Sout
W0
H0
L0
Cout
DQ
DS1867-50
1 µF
to Computer Sound Card
1/2 op-amp unity gain buffer to drive CM1004 meter
to Meter + Line
Digital pot set by ADC to 9-bit resolution - controls oscillator amplitude
QS0906-Spencer07

Figure 7 — RT Interface circuit diagram.

From June 2009 QST © ARRL

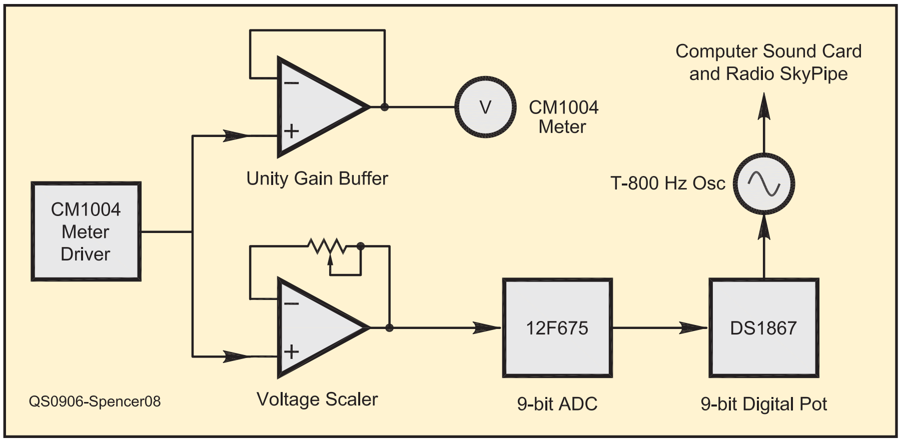

Figure 8 — RT Interface block diagram.

in Figure 3. The dimensions are not super critical, but careful placement certainly will improve the RT performance.

Some LNBs have two coax connectors. Only one will be used in the RT (Figure 4). It is a good idea to terminate the extra coax connector with a 75 Ω dummy load plug to balance the load on the LNB. The dummy loads for F type TV coax connectors are readily available from electronic parts retailers.

Note that the dish is mounted upside down. Though this orientation is not ideal for receiving satellite signals, this arrangement helps with pointing the dish in its radio telescope role.

Satellite Detector

The detector used in this project is the Channel Master (CM) satellite signal level meter model 1004IFD (Figure 5).[3] The CM is connected to the LNB. Power is supplied to the LNB through the coax connection from the CM. The CM detects the signal coming from the LNB and gives a meter indication of the signal strength and also varies the frequency of an audio tone to help technicians point the dish at the desired satellite. As you move the dish through the beam coming from the satellite, the meter indication will increase and then decrease coincident with the pitch of the audio tone.

The Itty-Bitty plans detail how to connect power to the CM and in turn connect power to the LNB (this power connection is handled by the interface in this project). Though somewhat effective, the CM meter and variable frequency tone indications provide limited utility in detecting changes in signal strengths required for radio astronomy.

Display

To really study the signals received by the RT, you will need to see them displayed graphically on a strip chart. There is an excellent software package called *Radio-SkyPipe* that is posted on radio astronomy Web sites.[4] The free version of this software is a good place to start. *SkyPipe* uses the computer sound card to measure the incoming signal strength and graphically displays the signal strength as a function of time. Figure 6 is illustrative of a signals detected by the RT. *SkyPipe* is very easy to use but some study of the HELP files will make it easier for you to fully tap into the capabilities of this software.

SkyPipe requires audio signals to be fed into the sound card MICROPHONE jack. The output of the CM detector is either an analog meter reading or a frequency modulated (constant amplitude) tone that is not really compatible with *SkyPipe*. An interface is required.

Interface

What is required to make the CM output work with *SkyPipe* and a sound card is to convert the signal level into an amplitude varying audio tone. The interface designed to do this is shown in Figure 7 and as a block diagram in Figure 8. Refer to the block diagram during the description of the interface function.

The unity-gain op-amp is used as a buffer between the CM meter driver circuit and the analog meter. The other op-amp is used as a voltage multiplier to scale the CM meter driver output voltage to match the 5 V reference voltage of the following analog to digital converter (ADC). The variable resistor in this voltage multiplier circuit is used to calibrate the CM to *SkyPipe*. The voltage from the multiplier is fed to a programmable interface controller (PIC) that is programmed as a 9-bit ADC to covert the analog voltage that is a function of received signal strength to a 9-bit digital word that is used to control a digitally controlled variable resistor. The interface includes a simple Twin-T audio oscillator circuit that provides a tone of approximately 800 Hz that is fed to the computer sound card. The amplitude of this audio oscillator is varied by the digital pot that is being controlled by the PIC. The result is the audio amplitude being varied in step with the signal strength detected by the CM.

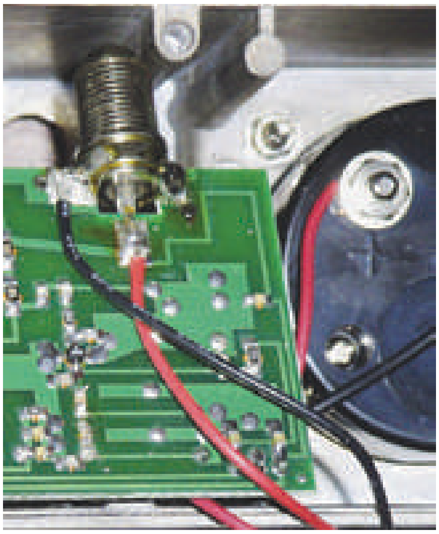

Figure 9 — Power and ground connection to CM board.

The circuit provides power to the CM and the LNB. A 12 V source in the CM is tapped through an RF choke and this is connected to the LNB coax connector inside the CM (Figure 9). The 12 V is also regulated to 5 V to provide power to the interface. Though probably not required, there are two 5 V sources, one for the digital components of the interface, and the other for the analog components with one common ground point. This arrangement is used to isolate potential digital and analog noise sources within the circuit.

The interface is built on a circuit board and mounted right inside the CM box

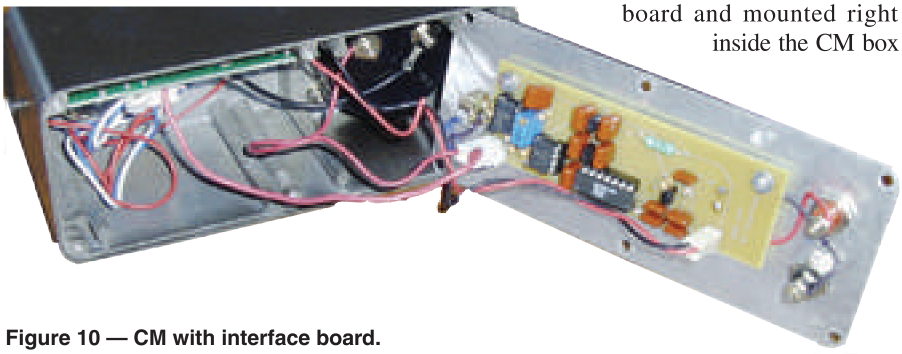

Figure 10 — CM with interface board.

From June 2009 QST © ARRL

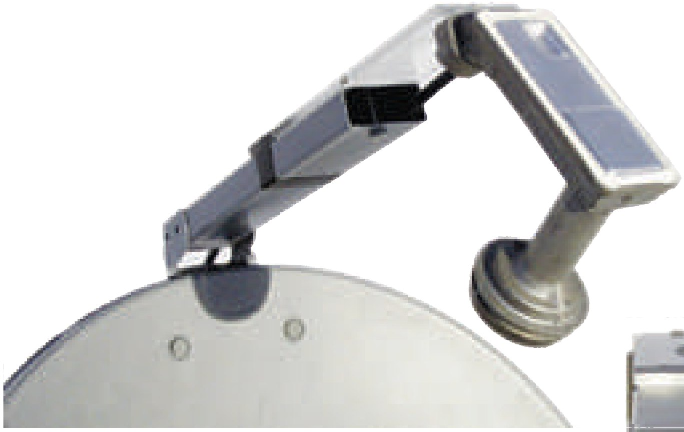

Figure 11 — Aiming the RT at the Sun, note LNB shadow location.

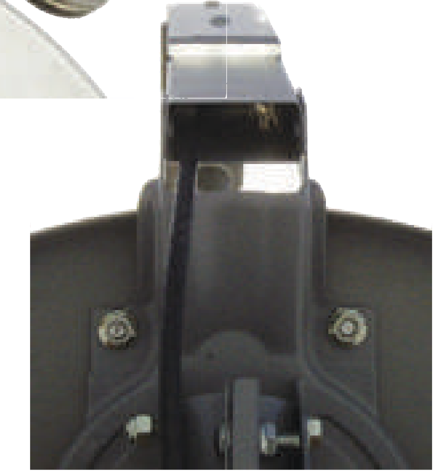

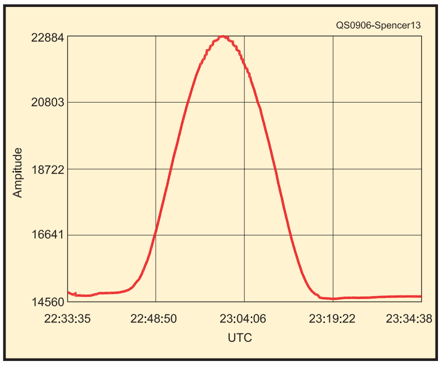

Figure 13 — Drift scan of the Sun indicating antenna's azimuth pattern.

(Figure 10). Though I made an etched circuit board for the circuit, the hand wired prototype worked equally well for those who would rather roll their own. The PIC firmware is available on the *QST* Web site.[6]

RT in Action

The first thing you need to do is learn how to point the RT antenna. The best place to start is to connect the CM to the antenna and point the antenna at the Sun. *Caution*: Do not look into the Sun as you do this, or at any time. Adjust the pointing angle and elevation until you get peak signal strength as indicated on the CM meter or hear the highest pitch audio tone. With the antenna pointed directly at the Sun, take note of the position of the shadow of the LNB on the surface of the dish (left in Figure 11). If you look from behind the dish, along the LNB

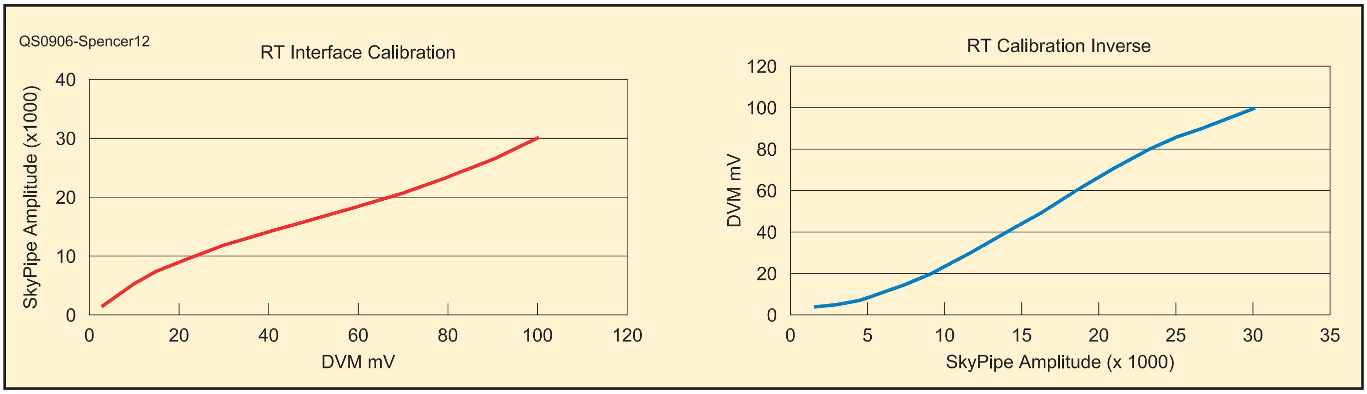

Figure 12 — Example calibration curves.

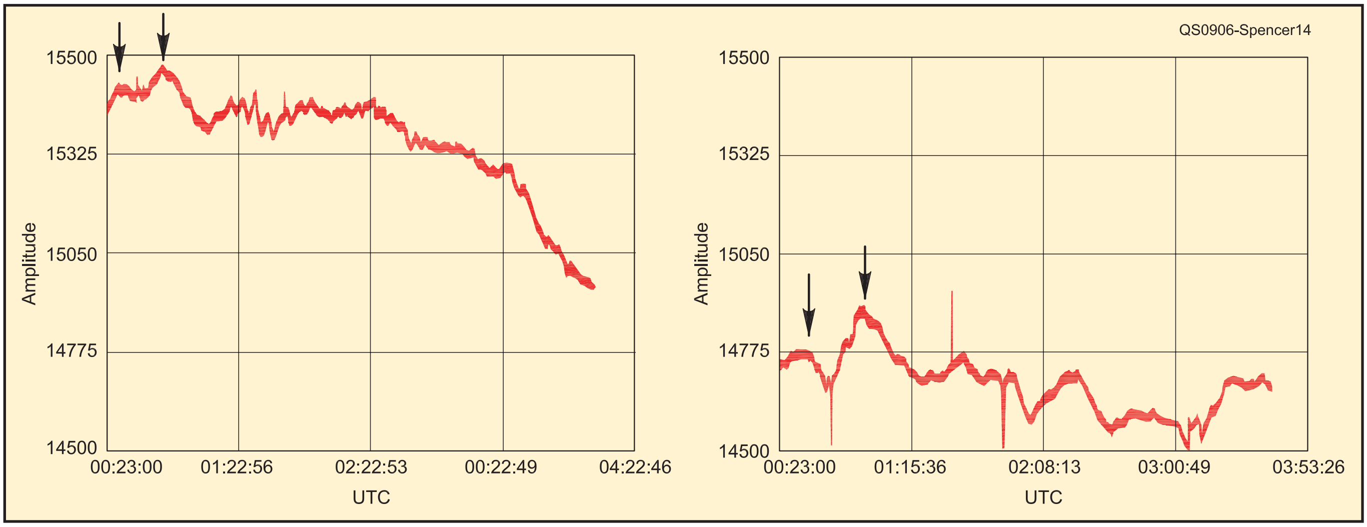

Figure 14 — Sequential drift scans. Note the time offsets between the peaks.

From June 2009 QST © ARRL

supporting arm (between the arm and the rim of the dish), you will see the Sun being blocked by the LNB.

Once you have the RT set up, it needs to be calibrated to match the output of the CM to *SkyPipe*. I have developed an *Excel* spreadsheet template to help with the calibration and a few of the other activities that you can accomplish with the RT (also available from the *QST* Web site). Turn the RT to a signal source, the Sun, or the side of a building would work. Turn the gain control of the CM to set the meter to maximum. Run *SkyPipe* and adjust the variable resistor on the interface board until you get a reading on the *SkyPipe* graph vertical (y) axis of approximately 32,000. With the maximum value set, adjust the CM gain control through the voltage range (0 to 100 mV) in 10 mV steps and record the corresponding y axis value on *SkyPipe*. This data is entered into the *Excel* spreadsheet to compute the calibration curve between voltage and y axis value. Both voltage and y axis values are used in analyzing recorded signal strength data (Figure 12).

A good first activity is to do a drift scan of the Sun. A drift scan means that you set the antenna azimuth (AZ) and elevation (EL) to some fixed pointing angle and allow the Earth to serve as the rotator to drag the antenna across the sky. To do a drift scan of the Sun, first set the elevation and azimuth to point directly at the Sun (maximum signal) and then move the azimuth toward the west (leave the elevation set) until you are off the peak signal. Now start *SkyPipe*. In about 15 minutes, the Sun will pass through the antenna pattern beam width and the result will be as illustrated in Figure 13. You can also use this collection technique to explore the antenna performance parameters.

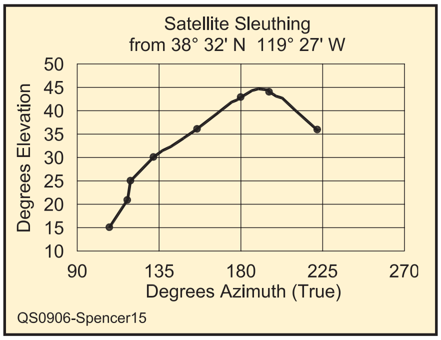

Figure 15 — Clarke Belt plot — tracking down satellites.

A good second activity is to do two drift scans of the night sky on two consecutive nights (beginning the scans at the same time each night) using the same fixed antenna azimuth (AZ) and elevation (EL). Figure 14 shows two such drift scans. Although at first glance they may not seem similar, there are some interesting features that are pointed to by arrows. If you compare the time that these two peaks occurred, the time difference is about 4.5 minutes. This shift is the result of the distance the Earth had traveled during the 24 hours between collections. This illustrates that the Earth's rotation as well as its travel in orbit needs to be considered when comparing drift scans. Enough to make your head spin (pun intended)?

A final good starting activity is to point the antenna toward the Clarke Belt and find all the satellites in geosynchronous orbit transmitting on 12 GHz. If you record signal strength peaks and AZ and EL for each peak, you will develop a graph of the Clarke belt as illustrated in Figure 15.

I have only scratched the surface, and the sky is the limit of this little project. The RT project can certainly broaden your horizons and expand your understanding of our universe. If you would like more detail than can be presented here, please contact the author.

Notes

[1] **www.setileague.org/articles/lbt.pdf**.

[2] **www.aoc.nrao.edu/epo/teachers/ittybitty/procedure.html**.

[3] **www.pctinternational.com/channelmaster/0612/satellite.html**.

[4] **radiosky.com/skypipeishere.html**.

[5] **en.wikipedia.org/wiki/Geostationary**.

[6] **www.arrl.org/files/qst-binaries/**.

Mark Spencer, WA8SME, is ARRL Education and Technology Program Coordinator. He is an ARRL member and holds an Amateur Extra class license. You can reach Mark at 774 Eastside Rd, Coleville, CA 96107 or at **wa8sme@arrl.org**.

Did you enjoy this article?
Cast your vote at:
VOTE www.arrl.org/members-only/qstvote.html

From June 2009 QST © ARRL

Simple demonstration to explore the radio waves generated by a mobile phone.

Dr Jonathan Hare, Sussex University, Department of Physics, Falmer, Brighton. BN1 9QH

Note: this article has been published: Simple demonstration to explore the radio waves generated by a mobile phone
J P Hare, 2010, Journal of Physics Education, Institute of Physics, 45, p. 481 45 481

Also see the brief full article at: mobile phone detector

IMPORTANT NOTE: this device works very well on the old style mobile phones (as shown in the photo above). However, it does not always work well with modern smart phones. This may be because modern phones use higher frequencies, less power and use the power in a slightly different way (e.g. spread spectrum). Some smart phones do work and success may be due to the signal strength of the local mobile phone mast nearby. If you are in a low signal area the phone will create more power to ensure reliable communications. If you are in a very strong signal area (very near the local network) your phone will drop its output power and consiquently there will be less power to pick-up and to convert to a voltage to light the LED.

Described is a simple low cost home-made device that converts the radio wave energy from a mobile phone signal into electricity to light an LED. No battery or complex circuitry is required. The device can form the basis of a range of interesting experiments on the physics and technology of our mobile phones.

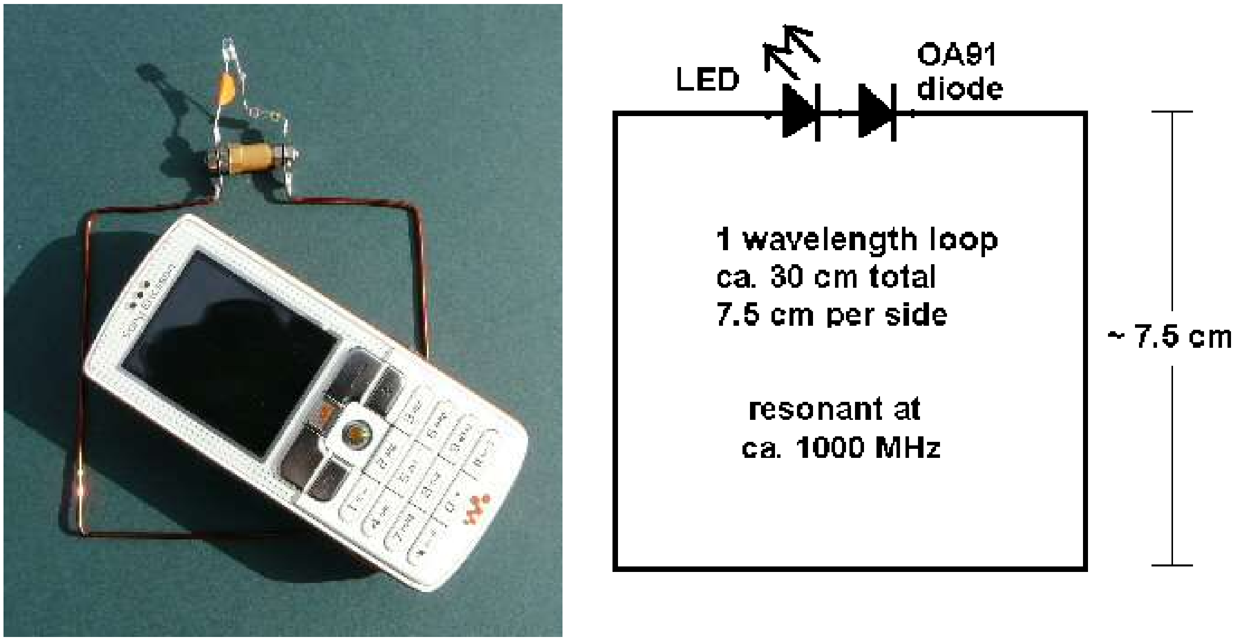

Fig. 1: left: mobile phone radio wave detector and right: the simple schematic

Introduction

Electromagnetic radiation (EMR) is at the heart of modern mobile phone data communications networks. The way a mobile phone and local base stations (the antenna covered masts you see dotted all around the place) communicate between each other is by using EMR in the radio wave part of the spectrum [1,2,3]. On switch-on your mobile sends digital information pulses by rapidly switching on and off the radio waves rather like a fast Morse code signal. Your text or voice is also converted into a series of digital pulses and sent across the network to be decoded (reassembled) by another mobile phone you dialled.

EM radiation and radio waves

Mobiles make use of various bands of radio frequencies to communicate between the mobile to base and the base to mobile: in Europe these include 900 and 1800 MHz (850 and 1900 MHz in the USA and Canada) [2, 3].

The relationship between wavelength, speed of light and the frequency follows the well known formula:
Wavelength λ (m) = speed / frequency = c (ms^{-1}) / ν (Hz)
λ (m) = 300,000,000 / ν (Hz) or approximately:
λ (m) = 300 / ν (MHz) ..... Equation 1.

So for a mid-range of about 1000 MHz (1 GHz) we get a typical mobile phone wavelength of about:
λ = 300/ 1000 = 0.3 m = 30 cm.

Simple radio wave detector

The loop consists of about a wavelength of wire, ca. 30 cm so each side is about 30/4 = 7.5 cm. The dimensions are not critical. The two ends are connected directly to a simple series circuit consisting of a high brightness LED and a germanium diode. They need to be connected correctly. All these components are cheap and readily available from electronic stores [4]. The loop can be made from a piece of copper wire roughly bent into a square (although a circular loop or rectangle will also work). If the wire is insulated remember to scrap off the insulation and solder-tin the ends. Simply solder the germanium diode and LED into circuit as shown in the diagram.
On a new LED the long lead is the positive (anode) while the short lead is the negative (cathode). The germanium diode has a line (band) around the end which is the cathode. When correctly wired the LED and the germanium diodes are connected so they both allow current to pass in the same direction, i.e. in the circuit diagram the arrows point in the same direction. In practice this means the LED and germanium diode are joined at the cathode of one and the anode of the other.
In my prototypes I used an insulator between the loop ends (light coloured cylinder in the photo) to make the whole thing more sturdy but this was purely for mechanical reasons and is not needed for the circuit to function properly. Note: a much more sensitive version using a x10 and x100 DC amplifier is described on my web site [6].

How it works and how to use it

When a radio wave passes across a metal object the EM fields cause the charged electrons in the metal to oscillate and this causes small AC currents at the same frequency to be induced into the metal. If a mobile is brought near to the loop and a call or text is made [5] the radio waves emitted from the phone pass across the loop. This induces a voltage into the antenna (the loop) and if it is close enough will be large enough to light the LED. As the loop is about one wavelength in size it is resonant and so there is a good transfer of power (low reactance) between the radio wave and LED.

The mobile phone automatically tests the network and adjusts its transmission power to maximise the battery life and minimise network interference. As a result the brightness of the LED will depend on the data being sent (the average signal), the local signal strength and how close the loop is to the phone. Why the second diode? - It's curious why the germanium diode is needed at all. The LED is a Light Emitting Diode after all and one would not think that another diode would help. However my initial experiments failed because I had not included it. The LED will have a relatively high capacitance which at these frequencies will tend to de-tune the loop and short out the LED. The germanium diode however is made up of a tiny wire which only makes a point-of-contact onto a piece of semiconducting germanium so it's 'self' capacitance is very low keeping the loop resonant.

The germanium diode will rectify the AC signal from the loop forming a series of DC pulses that will be nicely smoothed by the LED's capacitance. Without the diode however the raw AC signal from the loop will tend to be averaged to zero by the LED's capacitance.

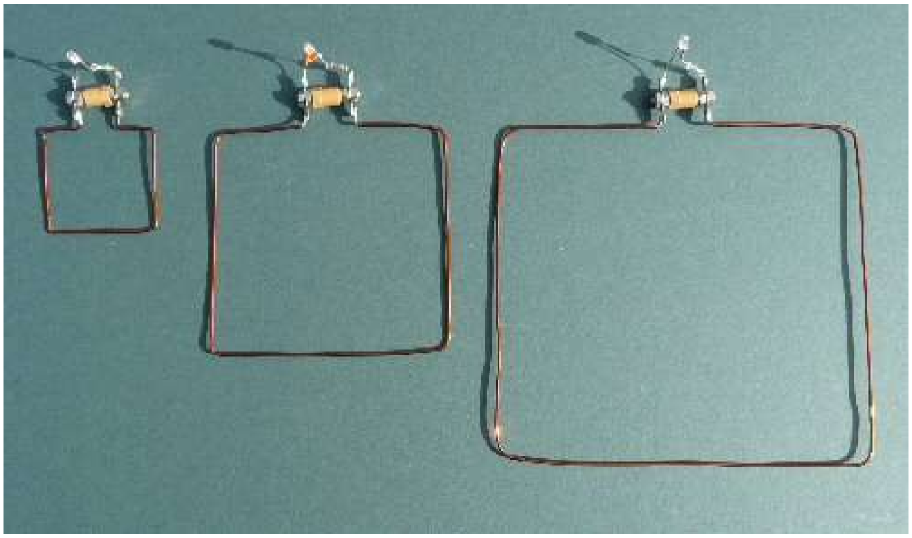

Fig. 2: Three loops in a row of varying size. The one described here and shown in Fig. 1 is shown in the middle. Smaller ones may well work better for higher frequencies such as the 3G networks (see below).

Other size loops

Fig. 2 shows a set of three loop devices with edge lengths of roughly 3.7, 7.5 and 15 cm. You can find out for yourself that the best match to the mobile signal is with the full wave loop of ca. 7.5cm per side. The other loops do work to varying degrees however (smaller ones may work better for the 3G network). The larger loop works well for the '70 cm' amateur radio bands.

Polarisation

The electric and magnetic fields making up the EM wave are orthogonal (they are at right angles to each other as they pass through space) to each other but depending how they are generated by the transmitting antenna can arrange themselves in any orientation with respect to the ground. If the electric field is parallel with the ground we say the wave is 'horizontally polarised' while if its normal to the ground we say its 'vertically polarised'. The loop antenna will respond best to one type of polarisation (depending on its orientation) so it's worth experimenting with the orientation of the mobile (or the loop) to get the strongest signal - brightest LED.

Mobile antenna

Inside your mobile phone is a transmitter / receiver and antenna. Many mobiles have this antenna at the top of the phone but some of the PDA type phones have it at the bottom. As a result you can locate the position of the antenna by moving it around the center of the loop till you get maximum LED brightness.

Networks

There are various different networks that a mobile may use both in the UK and abroad. It may be that you need to adjust the network phone settings on your mobile i.e. change from "automatic select" to set for "GSM" so as it get the strongest signal to light the LED. Note: the 3G network might not be powerful enough to light the LED. As the GMS network is currently the main network over the UK the device should work anywhere where you can get a signal as long as you check the correct selection on your mobile menus [5, 7]. The 3G network operates on a higher frequency (smaller wavelength) so you might find a smaller loop will work better than the main one described here. See 'other experiments' section below.

Test signals

In order to pick up the radio wave energy from the phone it obviously needs to be transmitting a signal. There are a few ways to do this:

1) On switch 'on' (or change of network) you can see that the mobile initially transmits for a few seconds to the network to tell it it's there (especially if you have moved since turning it off). You don't actually need to dial a number to detect these signals.
2) Even if don't text or call, throughout the day the mobile will send out data to 'keep up' with the network, especially if you are moving around (going through train tunnels etc. see below).

3) When you make a phone call you will transmit. Initially there is quite a lot of data being sent but in a few seconds data / power only gets transmitted when you speak. So to light the LED continuously you need to talk or provide some background sound continuously. Your service provider voicemail might be a good free phone number to try for these experiments [5].

4) Texting is the easiest way to show the radio wave power being transmitted. Long texts will light the LED for longer than short texts.

5) Finally set up the mobile on the loop and use another phone to text or phone the mobile. Even though you are not directly using the phone you will see that even on 'receive' the mobile phone transmits data to and fro. Ring off before you get charged.

Note: If you can use a free phone number it will save you money [5].

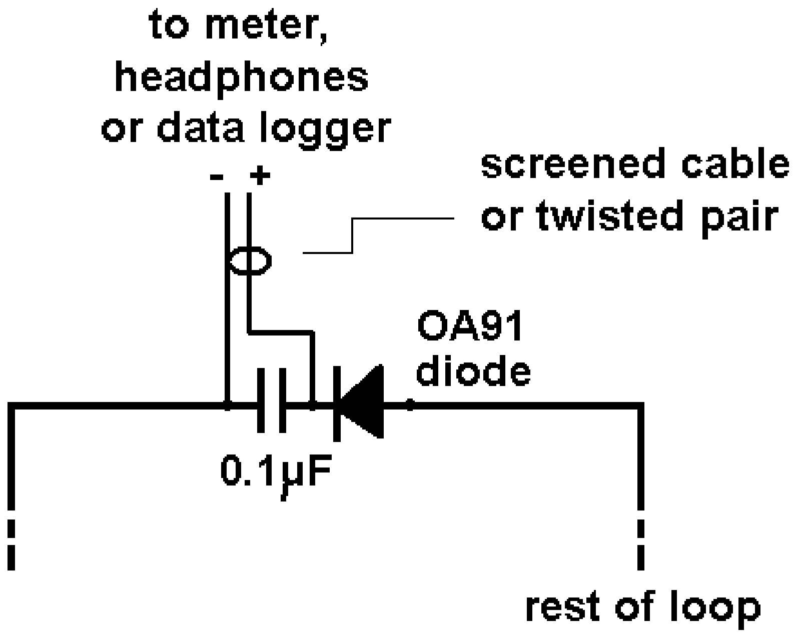

Fig. 3: Adding a capacitor and coax (or twin) lead so that headphones, a meter or a data logger can be connected (Note: diode is reverse wired compared to Fig. 1).

Other experiments:
Hearing data - if headphones are wired across the LED they will convert the voltages into sound and you can 'hear' the clicks of the digital data being transmitted. These are the same clicks that so easily get picked up by sensitive electronics such as a stereo amp or recording equipment when making a video for example. Hence - 'no phones on' when filming.

Logging data - if a meter, or better still a stand-alone data logger, is attached across the LED then one can monitor the EMR from the phone. For example even if you are not making a call your mobile will send signals too (and receive signals from) the network while travelling around. Fig. 3 shows a simple modification using a de-coupling capacitor so that a coax cable (or twisted pair) can be used to go to headphones, meter or data logger. Note the diode has been reversed so that the logger has the correct + and – connections for a unipolar input logger. The capacitor should help average the signal and stop radio frequencies going down to the logger. If one is available a few turns of the wire can be wound within a ferrite ring near to the logger so that maximum immunity to the mobile phone signal can be obtained for the logger electronics.

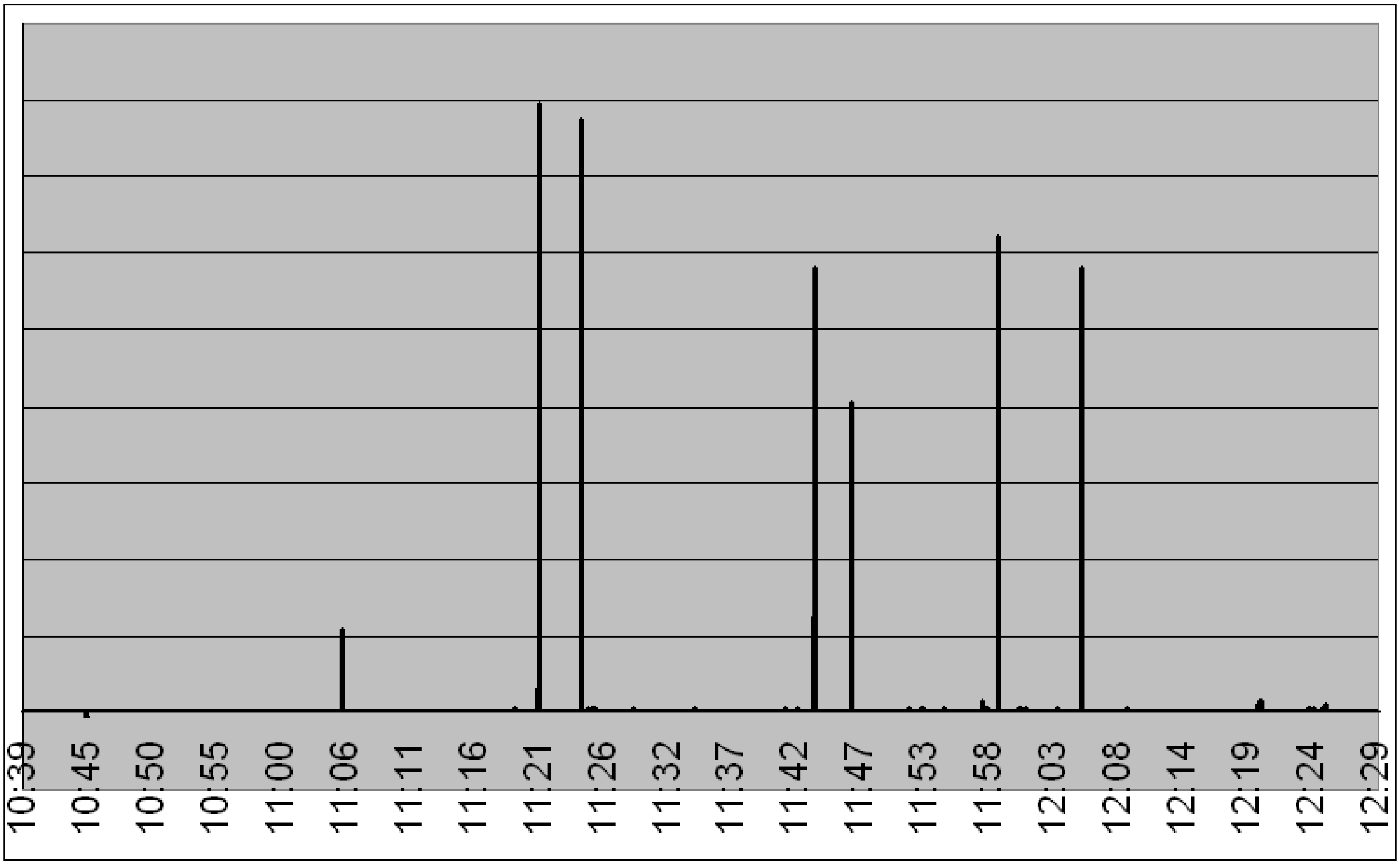

Fig. 4: Typical mobile phone data signals sent out onto the network while travelling around. These were recorded by a data logger from a mobile using the loop (no calls or text were made) while travelling on the train from Brighton to London Victoria (and then around London and return). Many of these peaks were the phone sending out 'I am here' data after coming out of one of the many long tunnels under the South Downs during the journey.

Out and about - Once you can log data you can discover all sorts of interesting things your mobile phone is doing without you realising it. Fig. 4 shows the plot over a few hours of travelling between Brighton and London (and within London) on the train. The detector was simply placed near to a phone that was not making or receiving a phone call or text, but was turned on.

The graph shows that the mobile sends out signals to tell the network where it is as it travels along and in particular goes in and out of long train tunnels. The peak heights vary because of the different powers the mobile transmits at depending on the signal strength of the local network and also because of the way the data logger 'snatches' a reading from the circuit every few seconds. As your phone sends out data onto the network to ensure the very best communications as you move around, so your mobile and the network obviously knows where you are and where you have been. Thieves and criminals beware the police can track you!

The inverse square law - If the transmitting mobile phone is moved away from the loop one would expect the signal to drop off. Unfortunately because both diodes need a certain threshold before they conduct the detector is not sensitive to small signals and not very linear. Therefore it's not very easy to use the device to measure the inverse square law (drop in signal v distance away) but of course you can see the signal go down. You could perhaps use the device to plot isobars - i.e. plot the equal intensity signals around the phone / nearby objects.

Changing the resonant frequency of the loop - you might be able to make some simple sliding mechanism (e.g. a small trombone-like mechanism) out of metal tube for example to tune the loop device for different frequencies. Then you can use it to find the average wavelength and so determine the center frequency by adjusting the size for maximum brightness of the LED. The wavelength can be determined by measuring the total distance around the loop. If we assume the antenna is one wavelength in total length then the frequency can be established by rearranging Equation 1, i.e. ν (MHz) = 30,000 / L (cm), where L is the length around the loop (cm).

Note: You will need to allow the transmitted digital signal to 'settle down' i.e. make measurements only after a few seconds after dialling / pick up so that only the sound data is being transmitted rather than the initial connection data. A constant sound will also need to be made so that the mobile phone continuously transmits data. It's worth playing music near to the phone or constantly whistling to keep sound coming into the phones microphone.

Mobile phone detector - teachers who want to know if the students / pupils really have turned-off their mobile phones (rather than just put on 'silent') can wire the loop device into the class room white-board speakers. Any mobile that is on in the class will send out signals which (if you are close enough) you will hear the data going to and fro - you will have your very own 'who's got their mobiles on' device which might be useful for exams etc.

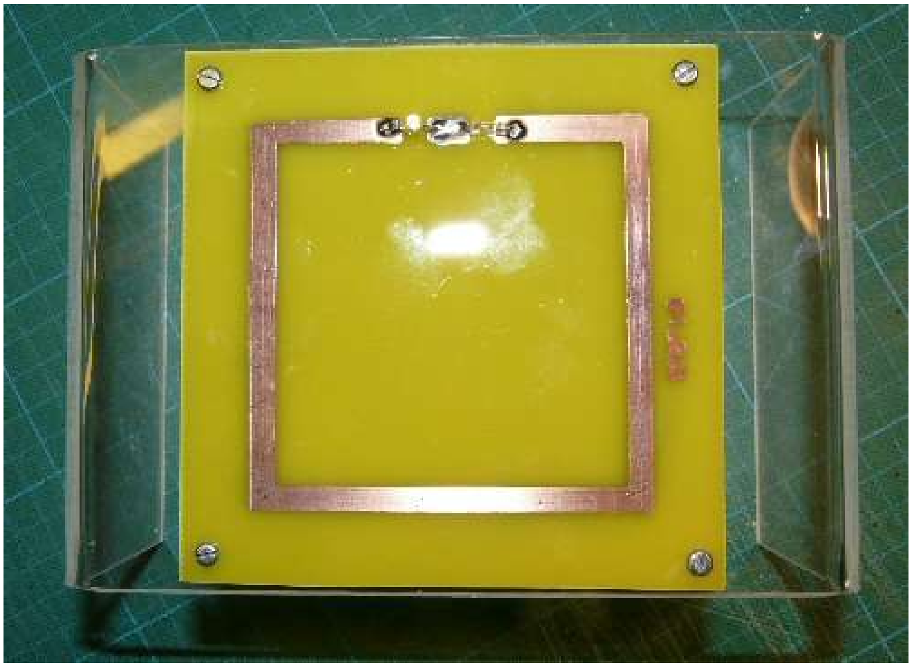

Fig. 5: The SEPNet 'deluxe' printed circuit board version (pcb) on a perspex stand where the loop is composed of a pcb copper track and the diode and LED soldered onto the board (top) [8,9].

Summary

All in all then, for such a simple easy to make device I hope you agree that there is a lot of scope for interesting science / technology investigations with your mobile phone. The device would make a good science week project (for radio amateur clubs etc.) A 'deluxe' pcb version (Fig. 5) on a perspex display case (Fig. 5) is currently going around the southern UK as part of the SEPnet outreach work, see the 'Radiation Exhibition' [8] and also as part of my on-going lecture series [9].

Post publication additions

(What follows was not included in the published article as this calculation was worked out later). A full wave loop is resonant and so looks purely resistive to the radio waves. Such a loop will have a resistance of about 100 ohms (Note: this is the AC resistance and not the DC resistance which will be very low). Now power P = V x I (V = voltage and I = current) and resistance = R = V/ I therefore $P = V^2 / R$ or rearranging $V = \sqrt{(P \times R)}$ which means that the voltage created by a power level of say 50mW (say for argument that roughly half the mobile phone power) arriving at the antenna will be about $V = \sqrt{(100 \times .05)}$ which is aprox. V = 2V, enough to light an LED.

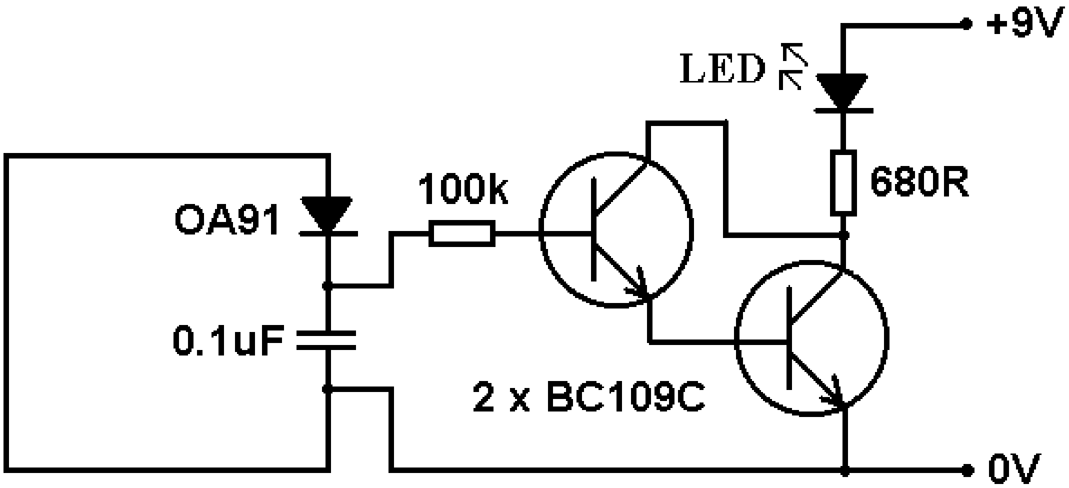

This circuit uses a two transistor darlington driver to amplify the signal from the loop and diode making the detector much more sensitive. The LED will be much brighter using this circuit. Note: the circuit needs a battery to power it (e.g. a PP3 9V)

References
[1] These ultra high frequencies (UHF, > 1000 MHz) are also often called microwaves.
[2] wiki pages
[3] Elektor Electronics magazine, June 2005
[4] order codes for the germanium diode and LED are:
e.g. Germanium diode: Maplin Electronics: QH71N, Rapid Electronics: 47-3114
e.g. LED: Maplin Electronics: UF72P , Rapid Electronics: 55-0085
[5] to save money use your voice mail service (often you simply dial 121).
[6] for details of an amplified detector see: wavemeter
[7] select 'network setting' from the mobile phone 'settings' menu and then go to 'network mode' and select 'GSM 900/1800' rather than 'automatic'.
[8] SEPnet mobile phone device on display throughout southern UK. [9] for details of my talks see: talks and workshops

THE CREATIVE SCIENCE CENTRE

Dr Jonathan Hare, Brighton, East Sussex. BN1 9QJ.

home | diary | whats on | CSC summary | latest news

Simple demonstration to show mobile phones emit radio waves

Dr Jonathan Hare, Sussex University, Department of Physics, Falmer, Brighton. BN1 9QH

Note: this article is in press: Elektor Magazine, July-August 2010, p. 56-57

For other experiments with this device please see my full article at: mobile phone detector

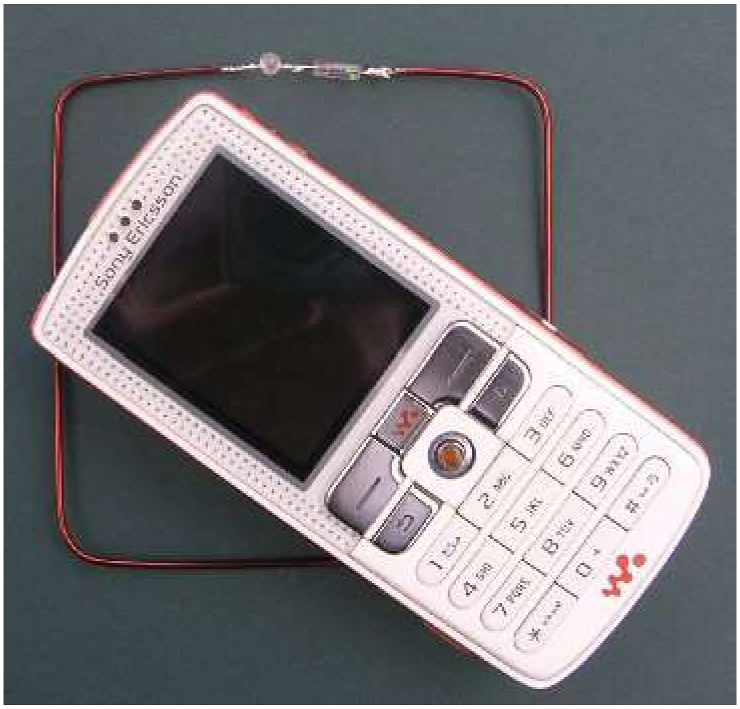

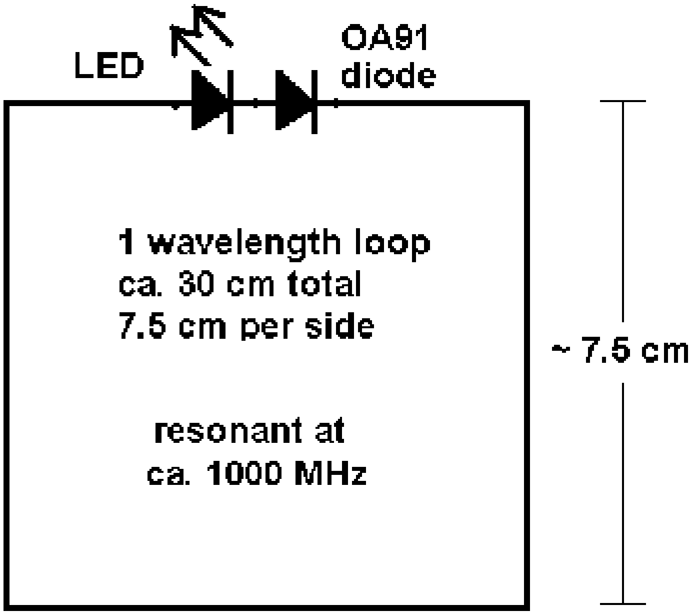
LED
OA91
diode
1 wavelength loop
ca. 30 cm total
7.5 cm per side
resonant at
ca. 1000 MHz
~ 7.5 cm

left: mobile phone radio wave detector and right: the simple schematic. Below: detail of the LED and germanium diode.

IMPORTANT NOTE: this device works very well on the old style mobile phones (as shown in the photo above). However, it does not always work well with modern smart phones. This may be because modern phones use higher frequencies, less power and use the power in a slightly different way (e.g. spread

spectrum). Some smart phones do work and success may be due to the signal strength of the local mobile phone mast nearby. If you are in a low signal area the phone will create more power to ensure reliable communications. If you are in a very strong signal area (very near the local network) your phone will drop its output power and consiquently there will be less power to pick-up and to convert to a voltage to light the LED.

This is a very simple and cheap device that demonstrates mobile phones ('cell phones' or 'handies') generate radio waves. We have a 30 cm (7.5 cm per side) full-wavelength loop antenna (a 'Quad' to radio amateurs) connected to a germanium diode and a hyper-bright LED. The loop can be made of copper wire, thin sheet metal or a track on a pcb. The diodes need to be wired correctly. I think the germanium diode is needed as the LED probably has too great a self-capacitance to perform at the very high AC frequencies generated by the phone (ca. 900 or 1800 Hz) but will work well with the DC pulses from the germanium diode (which has a very small capacitance).

To show the mobile generates radio waves put the mobile near to the loop and dial a number (use a free phone number, e.g. your voice mail) or text. The radio waves will induce a voltage into the loop, large enough to light the LED. The LED will flash indicating the digital data being sent by the mobile phone transmitter. You may need to set your phone to 'GSM 900/1800' rather than the '3G' network in the settings menu.

parts:
germanium diode: Maplin Electronics: QH71N or Rapid Electronics: 47-3114
LED: Maplin Electronics: UF72P or Rapid Electronics: 55-0085

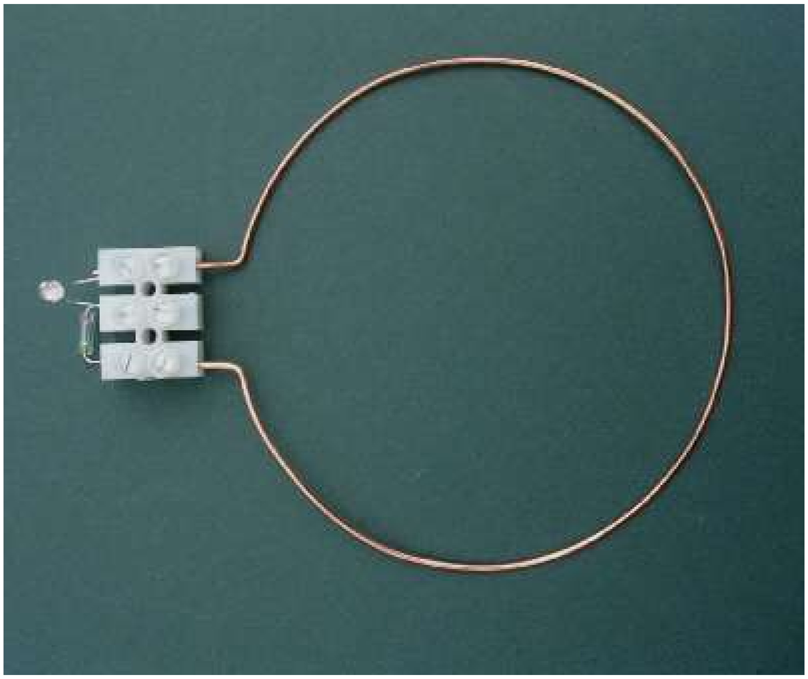

A very simple connector block version and a circular 1 wavelength loop

THE CREATIVE SCIENCE CENTRE

Dr Jonathan Hare, Brighton, East Sussex. BN1 9QJ.

home | diary | whats on | CSC summary | latest news

BACK TO G1EXG RADIO PAGE

EPAD MOSFET - 'near zero' and 'zero' threshold devices (e.g. the ALD110900)

I will add more details to this page later on but for the moment it gives links to important and inspiring sites.

ARTICLES
1) High Sensitivity Crystal sets, Technical Topics, Pat Hawker, Rad Com, p. 77-78, May 2007.
2) A novel kind of 'crystal set' radio, Giles Read, Rad Com, p. 60-61, June 2007.
3) also see Pat Hawker, Technical Topics, Rad. Com. January 200, page 53-57.
4) High Sensitivity Crystal Set by By Bob Culter (N7FKI)
5) Next-generation Zero-Threshold Voltage EPAD™ design enables circuits with greater operating range in low voltage supply environments

LINKS ABOUT THE CHIPS / DEVICES
1) Advanced Linear Devices web site
2) EPAD MOSFET (ALD web site)
3) ALD110800 and ALD110900 page (ALD web site)
4) ALD110802 and ALD110802 page (ALD web site)

ORDERING THE CHIPS / DEVICES
1) mouser.com (USA)

'CRYSTAL SET' RADIO'S - DIFFERENT DETECTORS

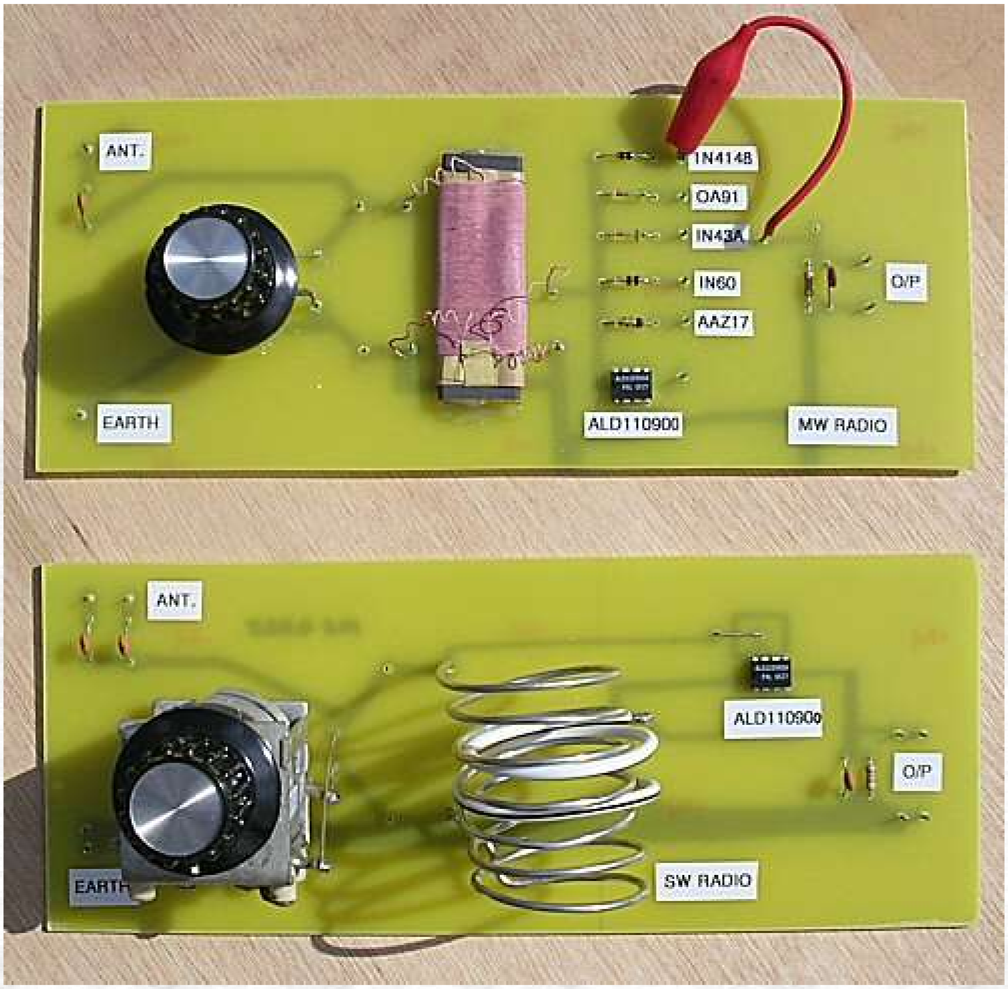

Top: the medium wave (MW) test radio with a selection of 'detectors' (selected by the croc clip lead) Bottom: the short wave (SW) version with a ALD110900 detector (on each circuit board the connections are: antenna - top left, Earth - bottom left, audio output (heaphones / amp) - far right.)

RADIO AND RADIO WAVES

The simplest 'radio' can be a piece of wire attached to the input of an amplifier. Why does this pick up radio signals? Well in principle it should not be able to pick up anything but in reality poor solder joints in the amplifier circuitry, as well as point contact effects between the ends of the anenna wire going into the input socket of the amp as well as other effects mean that you do often hear radio signals.

Radio waves are electromagnetic waves as they pass through a metal they induce small voltages into it. Any piece of metal e.g. metal fram specs, tape measures, metal window frames, a piece of wire etc. will act as an antenna and have tiny voltages induced into them. These voltages will be due to natural radio waves (from Space as well as the Sun and Earth), radio stations, satellites, mobile phones, garage door remote controls, microwave ovens .... the list goes on and on.

In the case of long, medium and short wave radio stations the amplitude of the radio wave signal is modified by the music, voice or program - we call it Amplitude modulation AM. Here the strength of the radio waves varies as the tones, loudness and pitch of the program vary. As a result the voltages induced in the metal object antennas distant from the radio wave stations also vary in accord this program information or modulation.

To actually hear the programs on the radio waves you cant actually take this tiny signal and listen to it directly (say with headphones or an amplifier). This is because the signal is a very high frequency signal out of the range of human hearing. To get the audible information - the program - 'off' the radio signal you need a device known as a 'detector'. This is usually a diode but lots of other things can act as inefficent detectors for example a mineral called galena, coke (burnt coal) ... as well as rusty screws and bad solder joints which is how the ampilfier mentioned above apparently picks up radio signals.

In the early days of radio crystals of galina were used as diodes and so these radios became known as 'crystal set radios'. They did not have any amplifiers and did not require any battery to work. They got all their power from that induced into the (of very long) antenna by the radio signal(s). The radio detector is an extremely important part of the radio and its proper function determines how well the radio works.

The radio signal is an AC signal. A speaker or headphones can not resonate at such high frequencies as radio waves and if you try to wire in the RF to the headphones the diaphram just averages the power. As half the time the RF is positive and half the time negative, in an AC signal, the average is zero and so you dont hear anything. The diode or detector is a device called a diode which only conducts electricity in one direction. So it only allows the positive, or negative half (one half but but not both) of the AC signal to go through (depending on which way it is wired). After the diode the average is now no longer zero its actually a changing signal dependant on the modulation of the radio wave - which is the information we want to hear. As this is at audio frequencies, and as the diaphram can move at audio frequencies, you hear the information on the detected RF signal - the music, voice or what ever.

Why should some detectors (diodes) be 'better' than others? The answer is not that some are better at 'magnifying' the signals but rather that each diode requires a certain threshold voltage in order to start conducting (one way). Signals below this level will not get passed through and a signal only slightly larger will therefore only pass through weekly (the diode unfortunatly absorbing the majority of the signal). As a result for large signals you wont expect to see much difference between different detectors but for weak signals there can be a great deal of difference. In an ideal world the diode would start to (forward) conduct at just above zero volts and this ideal detector (which does not conduct at all in reverse) would produce the greatest signal possible for the particular strength signal applied to it.

Recently I was playing around with OA91 germanium diodes for a mobile phone detector. I made a simple 10cm per side square loop and wired it directly to an LED. I was hoping that if I brought the mobile near to the loop and texted, or made a phone call, the RF should be picked up by the loop (which is almost resonant) and the LED should light - it didnt work! But putting an OA91 in series I did get it to work. So why did I need the OA91 diode when the LED is itself a diode?

It can't be due to the threshold voltage in this case as I am still using the LED and I did get it to work with the same signal! So I guess it because the capacitance of the LED is very high compared to the point-contact germanium diode. If we extend this thinking to the crystal set radio I guess the germanium diodes have very little capacitance and so dont allow any of the 'wrong' half of the AC signal through and so act as a more perfect detector. An LED would have a high capacitance and so let some of the wrong half of the AC signal through which on averaging by headphones would be nearer zero (see above) - hence less sound.

A low threshold (turn-on) voltage and a low self capacitance are crucial things in a good detector. Point contact germanium diodes having a very small area (a point of contact) have low capacitance and the germanium semiconductor-metal junction of the contact has a low turn on voltage. EPAD Mosfets also have low capacitance, high input resistance (low loading of the signal) and low turn on voltage.

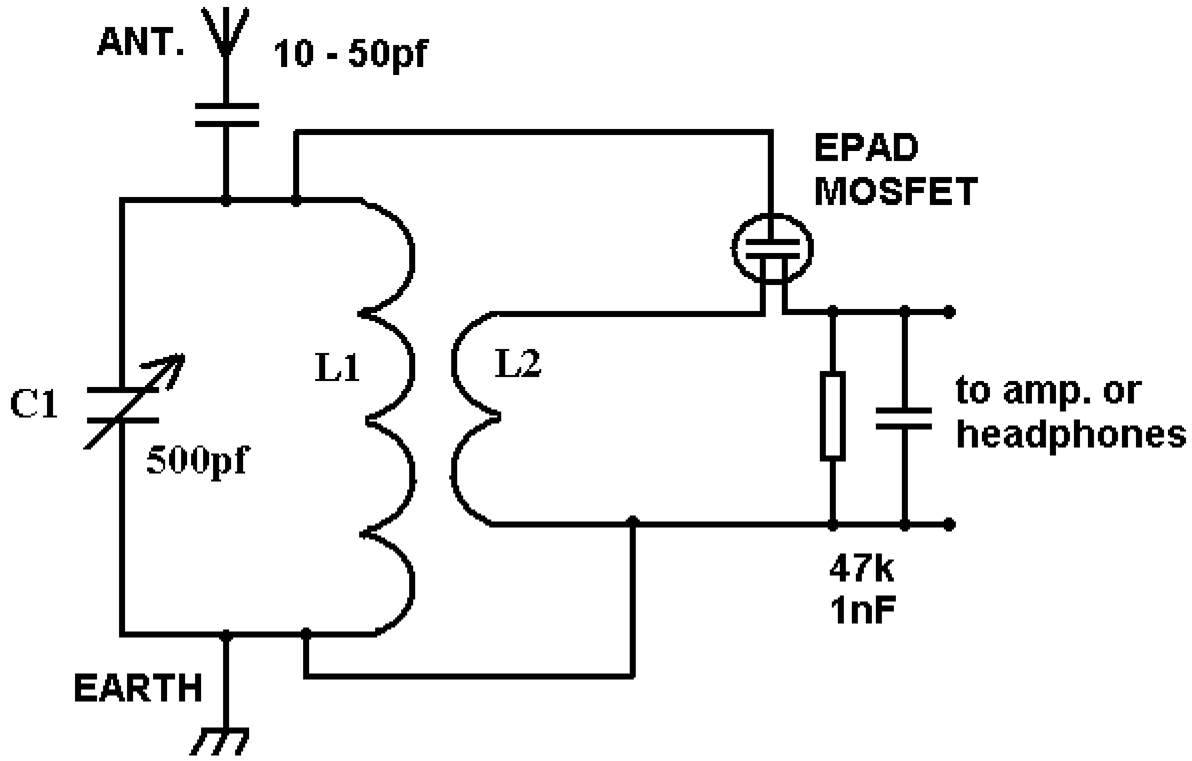

MEDIUM WAVE TEST CIRCUIT (top in photo)
As a simple test circuit to compare detectors I made up a simple 'crystal set' radio from a ferrite rod coil from an old radio, tuning capacitor an array of diodes (crystals) and a simple RF filter composed of a 47k resistor and 1nF capacitor. The coil and capacitor formed the resonant circuit which was fed via a 10 - 50pf capacitor by the antenna (long wire). This capacitor helps to reduce the effect of the antennas own capacitance and inductance from modifying the resonant frequency of the tuned circuit. Ideally this should be as small a capacitance as you can get away with but too small a value will reduce the radios sensitivity ....

An earth was connected to the ground connection of the circuit. Instead of taking the antenna end of the resonant circuit to a diode detector (as many simple crystal sets do) I used a ca. 10 turn coupling coil around the coil. This was done to limit the loading on the resonant circuit by the detector. It should help keep the tuning sharp and also help to keep the overall (band) spread of tuning as large as possible.

Because of the reduction in turns (ca. 100 turns : 10 turns, i.e. 10:1) the coupling coil reduces the signal that is available to drive the detector. For example if I use a standard silicon diode (e.g. a 1N4148) which requires about 0.6V to conduct then we might need ca. 10 x 0.6 = 6V of RF to be developed in the resonant circuit! A high Q resonant circuit might do this with a long antenna coupled to it but without such a generious antenna such a set-up wont be very sensitive.

RESULTS
I tried an array of 'crystal set' type diodes to campare them. Typical rough results are shown in the table below. This was with an Earth and about 3-4m of antenna wire randomly strung near to ground level (so its a pretty bad set up really - a good test).

The EPAD MOSFET data sheets described how the gate and drain of one of the MOSFETS can be joined together to form the anode of a diode the source then forms the cathode. I tried this arangement with the ALD110900 in order to make up a 'crystal set' diode from this MOSFET but it didnt perform well.

Then I tried the alternative arrangement as described in the references above. In these articles they dont use the MOSFETS as diodes but connect the gate directly to the resonant circuit. As the input resistance is extremely high (c.a. 1E12 ohms) and the input capacitance is very low (c.a 3pf) it does not load the tuned circuit by any appreciable amount. The source and drain are then wired between the coupling coil and filter just as the usual diode detector is in a crystal set radio. In this configuration the MOSFET is working more as a gate voltage controlled resistor rather than a diode i.e 'off' for half an RF cycle and 'on' for the rest - this lets the envelope through and you hear the music or sound. One article describes the arrangement as a synronious detector. As there are two MOSFETS in one package the two are wired in parallel i.e. gate to gate, drain to drain (there is a common source) etc.

DETECTOR TYPE	**AF voltage produced after filtering**
1N4148 silicon	no signal
OA91 germanium	1mV
1N43A germanium	1mV
1N60 germanium	1mV
AAZ17 germanium	1.5mV
ALD110900 EPAD MOSFET wired as diode	<1mV
ALD110900 EPAD MOSFET wired as sync. detc.	10mV

Typical results of the AF voltages produced by the detector (for use by headphones or amplifier) for an Earth connection, small (ca. 3-4m antenna wire near ground level) and the prototype 'crystal set' radio.

For the simple set-up described above there was not enough signal to get the silicon diode to work at all (although a much longer wire did provide a very small signal). A simple listening test and measurement of the AF prodiced by each diode showed that their was very little difference between the various germanium diodes. The MOSFET simply wired as a diode (gate wired directly the drain) was a poor detector - better than a silicon but not nearly so good as any of the germanium diodes. Finally the EPAD MOSFET wired as a syncronious detector worked really very well apparently producing ca. 10 times the signal (in other words it has much less loss than a germanium diode) for this particular set up and station than the germanium diodes.

As the syncronious arrangement was so succesful it got me thinking about this circuit. As the gate goes straight in on the main resonant circuit where the voltages, in a good Q circuit, might well be relatively high I thought I would try a standard FET (2N3819) and MOSFET (VN10KM) instead - and they worked! (they are a lot cheaper than an EPAD MOSFET chip and easier to get). Some FETS have a larger input capacitance so it helped to put a 10pf capacitor in series to the gate (a 10M ohm from gate to earth might reduce 50Hz pick up).

They were not as good as the EPAD MOSFET (although I only tried a single device) but the arrangment was still better than a simple 'crystal set' diode. For weak signals the EPAD MOSFET should perform better as it is able to keep going while the FET and the standard MOSFET will start to trail off as the resonant circuit voltage drops below the relatively higher gate voltage required for these devices. However it was interesting to see that it seems to be possible to get a more efficent set-up for a 'crystal' radio using a standard FET or MOSFET than with the standard germanium diode detector.

By the way I recently re-tried the 1N4148 diode but wired a 1.5V AA battery with a 100k resistor across it to forward bias it - and it worked ! It seemed to work better than a OA91!

SHORT WAVE 'CRYSTAL SET RADIO (bottom in photo)

The bottom circuit shown in the photo above is my own short wave version of the 'crystal set' radio but based on the references above. It's the same circuit as the MW radio but has a reduction drive tuning capacitor and different coil. I used about 7 turns (ca. 4cm diameter) for the resonant circuit and a couple of turns of insulated wire (white) for the coupling coil. I guess the radio tunes from about 3-8 MHz - I havent tried a sig gen on it yet.

I made this in the summer months (not the best time for listening on this part of the spectrum). Using an Earth and few meters of wire for antenna I heard 2 or 3 stations from around the world during the day while at night there were at least 7 including the voice of Russia world service, Radio Thailand as well as German and French speaking stations.

BACK TO G1EXG RADIO PAGE

THE CREATIVE SCIENCE CENTRE

Dr Jonathan Hare, The University of Sussex
Brighton, East Sussex. BN1 9QJ

home | diary | whats on | CSC summary | latest news

PUBLISHED BY

INTECH
open science | open minds

World's largest Science,
Technology & Medicine
Open Access book publisher

3,200+
OPEN ACCESS BOOKS

105,000+
INTERNATIONAL
AUTHORS AND EDITORS

111+ MILLION
DOWNLOADS

BOOKS
DELIVERED TO
151 COUNTRIES

AUTHORS AMONG
TOP 1%
MOST CITED SCIENTIST

12.2%
AUTHORS AND EDITORS
FROM TOP 500 UNIVERSITIES

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the
Book Citation Index in Web of Science™
Core Collection (BKCI)

Chapter from the book *Sustainable Energy Harvesting Technologies - Past, Present and Future*

Downloaded from: http://www.intechopen.com/books/sustainable-energy-harvesting-technologies-past-present-and-future

Interested in publishing with InTechOpen?
Contact us at book.department@intechopen.com

10

Design Issues in Radio Frequency Energy Harvesting System

Chomora Mikeka and Hiroyuki Arai
Yokohama National University
Japan

1. Introduction

Emerging self powered systems challenge and dictate the direction of research in energy harvesting (EH). State of the art in energy harvesting is being applied in various fields using different single energy sources or a combination of two or more sources. In certain applications like smart packaging, radio frequency (RF) is the preferred method to power the electronics while for smart building applications, the main type of energy source used is solar, with vibration & thermal being used increasingly. The main differences in these power sources is the power density; for example RF (0.01 ~ 0.1 μW/cm^2), Vibration (4 ~ 100 μW/cm^2), Photovoltaic (10 μW/cm^2 ~ 10mW/cm^2) and Thermal (20 μW/cm^2 ~ 10mW/cm^2). Obviously RF energy though principally abundant, is the most limited source on account of the incident power density metric, except when near the base stations. Therefore, in general, RF harvesting circuits must be designed to operate at the most optimal efficiencies.

This Chapter focuses on RF energy harvesting (EH) and discusses the techniques to optimize the conversion efficiency of the RF energy harvesting circuit under stringent conditions like arbitrary polarization, ultra low power (micro or nanopower) incidences and varying incident power densities. Harvested power management and application scenarios are also presented in this Chapter. Most of the design examples described are taken from the authors' recent publications.

The Chapter is organised as follows. Section 2.1 is the introduction on RF energy sources. Section 2.2 presents the antenna design for RF EH in the cellular band as well as DTV band. The key issue in RF energy harvesting is the RF-to-DC conversion efficiency and is discussed in Section 2.3, whereas Section 2.4 and 2.5 present the design of DTV and cellular energy harvesting rectifiers, respectively. The management of micropower levels of harvested energy is explained in Section 2.6. Performance analysis of the complete RF EH system is presented in Section 3.0. Finally, conclusions are drawn in Section 4.0.

1.1 RF energy sources

These include FM radio, Analogue TV (ATV), Digital TV (DTV), Cellular and Wi-Fi. We will present a survey of the measured E-field intensity (V/m) for some of these RF sources as shown in Table 1, [1]-[2]. Additionally, measured RF spectrums for DTV and Cellular signals are presented as shown in Fig. 1 to show on the potential for energy-harvesting in these frequency bands. In general, many published papers on RF-to-DC conversion, have presented circuits capable of converting input or incident power as low as -20dBm. This means that, if an RF survey or scan finds signals in space, with power spectrum levels around -20dBm, then, it is potentially viable to harvest such signal power. In Fig. 1 (left side), the spectrum level is well above -20dBm and hence, a higher potential for energy harvesting. In Fig. 1 (right side), while the spectrum level is below -20dBm, what we observe is that the level increases with decrease in the distance toward the base station (BTS). Using free space propagation equation with this data, it was calculated that at a distance 1.4 m from the BTS, the spectrum level could measure 0dBm. An example calculation and plot for the estimated received power level, assuming 0dBi transmitter (BTS) and receiver antenna gains and free space propagation loss (FSPL) for FM and DTV is presented in Section 2.1.1. For the example estimation in Section 2.1.1, we select FM and DTV because they measured with a higher level than cellular and Wi-Fi for example.

Source	V/m	dBm	Reference	
FM radio	0.15~3		Asami et al.	
Analogue TV	0.3~2			
Digital TV	0.2~2.4	-40~0.0	Asami et al.	Arai et al.
Cellular		-65~0.0	Mikeka and Arai	
Wi-Fi		≅ -30		

Table 1. RF energy sources, measured data.

In Table 1, FM radio has the highest E-field intensity implying the highest potential for energy harvesting. However, due to the requirements for a large antenna size and the challenges for simulations and measurements at the FM frequency i.e. 100 MHz or less (See Section 2.2.3, example FM antenna at 80 MHz), this Chapter will focus on DTV (470~770 MHz band) and Cellular (2100 MHz band) energy harvesting.

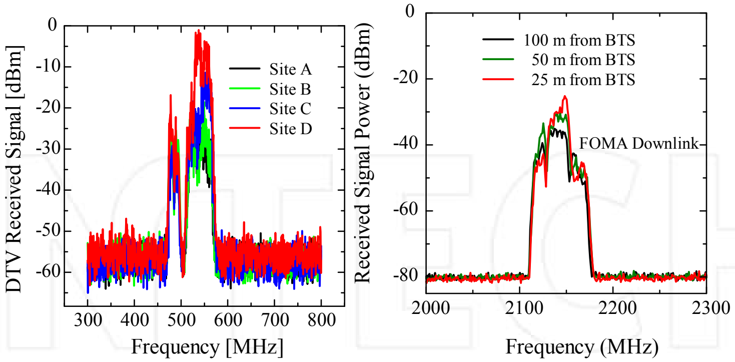

Fig. 1. DTV signal spectrum measured in Tokyo City (left side graph) and Cellular signal spectrum measured in Yokohama City (right side graph).

The received DTV signal power is high and also wide band, presenting high potential for increased energy harvesting unlike in cellular signals. We demonstrated in [2] that the total RF-to-DC converted power is roughly the integral over the DTV band (1), and is significantly larger than in the case of narrow band cellular energy harvesting.

$$P_{DC(DTV)} = \alpha \int_{470}^{770} \delta P_{DC}(f)\, df \,, \tag{1}$$

where α is the attenuation factor on the rectifying antenna's RF-to-DC conversion efficiency due to multiple incident signal excitation. δP_{DC} is the small converted DC power from each of the single DTV signals in the 470 MHz to 770 MHz band.

In detail, we derive (1) from fundamentals as follows.

The incident power density on the rectifying antenna (rectenna), $S(\theta,\phi,f,t)$, is a function of incident angles, and can vary over the DTV spectrum and in time. The effective area of the antenna, $A_{eff}(\theta,\phi,f)$, will be different at different frequencies, for different incident polarizations and incidence angles. The average RF power over a range of frequencies at any instant in time is given by:

$$P_{RF}(t) = \frac{1}{f_{high} - f_{low}} \int_{f_{low}}^{f_{high}} \int_{0}^{4\pi} S(\theta,\phi,f,t) A_{eff(\theta,\phi,f)}\, d\Omega\, df \tag{2}$$

The DC power for a single frequency (f_i) input RF power, is given by

$$P_{DC}(f_i) = P_{RF}(f_{i,}t).\eta\left(P_{RF}(f_i,t),\rho,Z_{DC}\right), \tag{3}$$

where η is the conversion efficiency, and depends on the impedance match $\rho(P_{RF},f)$ between the antenna and the rectifier circuit, as well as the DC load impedance. The reflection coefficient in turn is a nonlinear function of power and frequency.

The estimated conversion efficiency is calculated by P_{RF}/P_{DC} . This process should be done at each frequency in the range of interest. However, DC powers obtained in that way cannot be simply added in order to find multi-frequency efficiency, since the process is nonlinear. Thus, if simultaneous multi-frequency or broadband operation like in DTV band is required, the above characterization needs to be performed with the actual incident power levels and spectral power density. In this Chapter, we shall demonstrate DTV spectrum power harvest, given a rectenna than has been characterised in house at each single frequency in the DTV band.

1.1.1 An example calculation and plot for the estimated received power level

In this example we consider Tokyo's DTV and FM base stations (BS) as the RF sources. Both DTV and FM BS transmitter power (P_t) equals 10 kW (70dBm). The antenna gains are assumed 0dBi in both cases but also at the points of reception for easiness of calculation but with implications as follows. Assuming 0dBi antenna at each reception point, demands that we specify the frequency of the transmitted signal. For this reason we specify DTV signal frequency to be equal to 550 MHz while the FM signal frequency equals 80 MHz (Tokyo FM).

The received power, P_r is calculated using the simplest form of Friis transmission equation given by

$$P_r = P_t + G_t + G_r + FSPL\,, \tag{4}$$

where P_t = 70dBm, $G_t = G_r$ = 0dBi. G_r is the receiving antenna gain while FSPL is the free-space path loss given by

$$FSPL(dB) = 20\log(d) + 20\log(f) + 32.45\,, \tag{5}$$

where ***d*** is in (km) and ***f*** is in (MHz). The plot for the received power as a function of distance from the DTV and FM base stations is shown in Fig. 2.

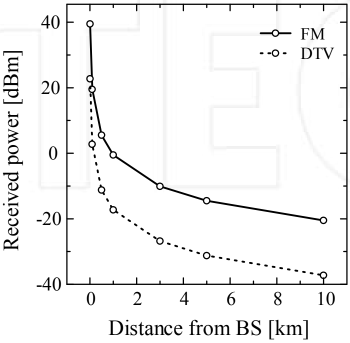

Fig. 2. DTV and FM received signal power level against distance.

With respect to Fig. 2, FM registers higher received power level than DTV at every reception point due to its lower transmit frequency and hence lower free-space path loss. For example at 1 km distance, FM received power level is -0.51dBm while for DTV, the received power is -17.26dBm. The important thing however, is that the received power level is frequency independent. It means that P_t is the transmitter power and the received power level at the position of distance ***d*** is $\frac{P_t}{4\pi d^2}$. However, if we assume 0dBi antenna at each reception point as in the above example, the power level is different because the antenna size of 0dBi is frequency dependent. As a result, high transmit power level is favorable for RF energy harvesting. Also near the base station is favorable.

1.2 Antenna design for the proposed RF energy harvesting (EH) system

It is well known that RF EH system requires the use of antenna as an efficient RF signal power receiving circuit, connected to an efficient rectifier for RF-to-DC power conversion. Depending on whether we want to harvest from cellular or DTV signals, the antenna design requirements are different. We will discuss the specific designs in the following sub sections.

1.2.1 Cellular energy harvesting antenna design

We propose a circular microstrip patch antenna (CMPA) for easy integration with the proposed rectifier (Section 2.5.1). However, the use of circular microstrip patch antennas (CMPA) is often challenged by the need for impedance matching, circular polarization (CP) and higher order harmonic suppression.

To address the above concerns, we create notches on the circular microstrip patch antenna. In our approach, we use only two, thin, fully parameterized triangular notches to achieve higher order harmonic suppression, impedance matching and circular polarization, all at once. This is the novelty in our proposed antenna. Our proposed CMPA is shown in Fig. 3. We study the behaviour of CMPA surface current vectors when notches (triangles ABC) are created on the structure at $\alpha = 45^0$ and $\alpha = 225^0$.

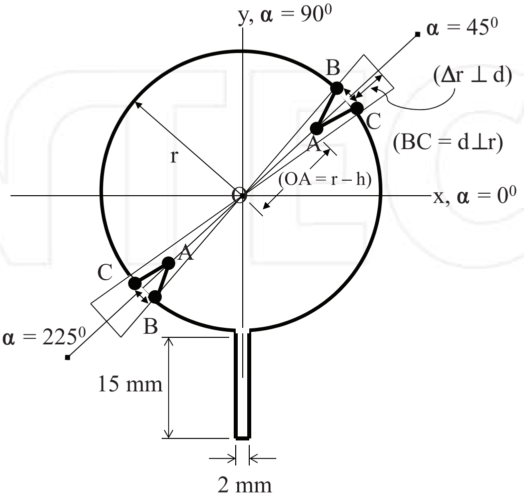

Fig. 3. Cellular energy harvesting antenna structure.

Notch parameters ***d*** and ***h*** in Fig. 3 were investigated by calculation using CST microwave Studio.

Without notches, the CMPA's input is not matched at f_c= 2.15 GHz as shown in Fig. 4 (left side). However, with notches, matching is achieved. The parameter combination ***d*** = 7 and ***h*** = 6 offers a matched and widest band input response and hence we adopt it for cellular energy harvesting applications.

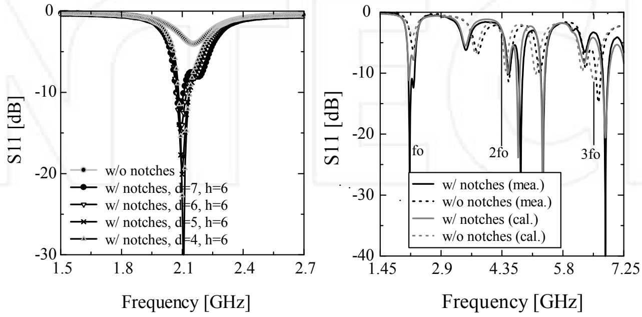

Fig. 4. *Left side:* ***d*** and ***h*** parameter investigation. *Right side:* Comparison between (cal.) and (mea.) S_{11} at f_0 = 2.175 GHz, $2f_0$ = 4.35 GHz and $3f_0$ = 6.53 GHz. The adopted notch parameters are ***d***=7 mm while ***h***= 6 mm.

The comparison between calculated and measured S_{11} is shown in Fig. 4 (right side). The 2nd and 3rd harmonics are suppressed as required by design. The comparison between calculated and measured radiation patterns is shown in Fig. 5, where $E_\vartheta \cong E_\varphi$ due to the 45^0 tilted surface current vector. Ordinarily, without notches, the surface current vector is parallel to the microstrip feeder axis. In conclusion, our proposed CMPA is sufficiently able to suppress higher order harmonics while simultaneously radiating a circularly polarized (CP) wave. The CP is required to efficiently receive the arbitrary polarization of the incident cellular signals at the rectenna.

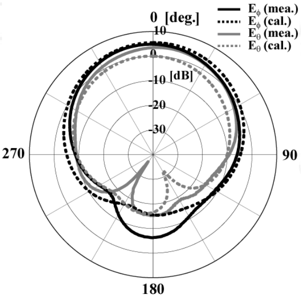

Fig. 5. Cellular energy harvesting antenna pattern at f_c= 2.15 GHz.

1.2.2 DTV energy harvesting antenna design

Unlike in the cellular energy harvesting antenna, the DTV energy harvesting antenna must be wideband (covering 470 MHz to 770 MHz), horizontally polarized and omni-directional.

The proposed antenna is typically a square patch (57 mm x 76 mm) with a partial ground plane (9 mm x 100 mm). The patch is indirectly fed by a strip line (9 mm x 3 mm). The proposed antenna geometry is shown in Fig. 6. The partial ground plane is used to achieve omni-directivity and a certain level of wide bandwidth. To tune the impedance of this antenna as well as to adjust the bandwidth within the target band, a "throttle" with stepped or graded structures is used between the microstrip feed line and the square patch, as shown in Fig. 6 (left side).

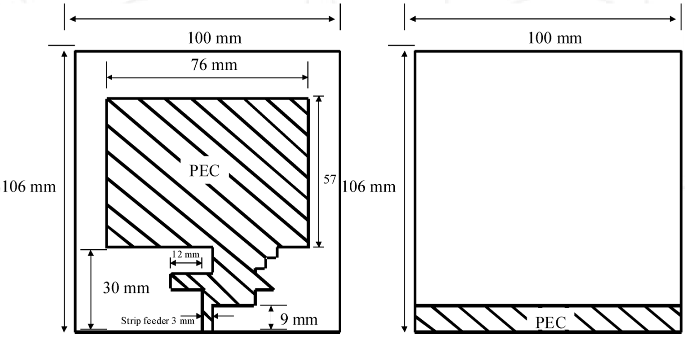

Fig. 6. Proposed DTV antenna geometry. *Left side:* Front view. *Right side:* Back view. The antenna is printed on FR4 substrate; t = 1.6 mm, ε_r = 4.4.

The input response for the proposed antenna is shown in Fig. 7 (left side). The omni directivity is confirmed by measurement at 500 MHz, 503 MHz, and 570 MHz as shown in Fig. 7 (right side). The radiation patterns shown in Fig. 7 are for the xz plane, which happens to be the vertical polarization for the antenna. DTV signals are horizontally polarized and therefore, when using this antenna, the orientation must be in such a way as to efficiently receive the DTV signal. Simply a 90 degree rotation of the antenna along the z axis achieves this requirement.

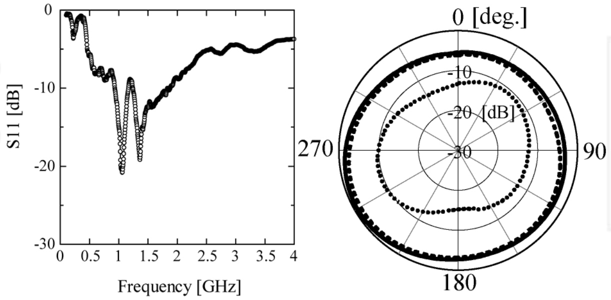

Fig. 7. Proposed DTV antenna performance. *Left side:* The antenna's measured input performance. *Right side:* The omni directivity in the vertical plane is confirmed at 500 MHz, 503 MHz, and 570 MHz. The outermost, black solid and dotted line patterns represent 503 MHz and 500 MHz directivity, respectively. The innermost dotted line pattern is the directivity at 570 MHz.

1.2.3 Example design for an 80 MHz FM half-wave dipole antenna

A half-wave dipole is the simplest practical antenna designed for picking up electromagnetic radiation signals, see Fig. 8 (courtesy of Highfields Amateur Radio Club). Calculating the optimal antenna length to pick up a certain frequency signal is fairly straightforward because antenna physics demand that the total length of wire used in the antenna be equal to one wavelength of the type of electromagnetic radiation it will be picking up. This means that the total length of the antenna should be equal to half the desired wavelength. By converting the 80 MHz frequency into a wavelength, you can thus obtain your antenna length as 1.875m by using the magic equation, $\lambda = \frac{c}{f}$. However, the actual length is typically about 95% of a half wavelength in free space, hence a half-wave dipole for this frequency should be 1.788m long, which would make each leg of the dipole 0.894m in length.

1.3 RF-to-DC conversion efficiency improvement techniques

A Schottky diode circuit connected to an antenna is used for RF-to-DC power conversion. To convert more of the antenna surface incident RF power to DC power, high RF-to-DC conversion efficiency is required of the rectifying circuit. Many authors have shown that the efficiency depends on several factors like Schottky diode type, harmonics suppression capability, load resistance selection, and the capability to handle arbitrary polarized incident waves. What is missing in most of these published works is the efficiency optimization for

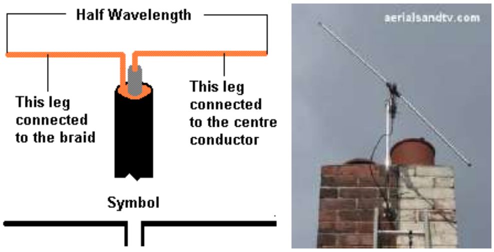
Half Wavelength
This leg connected to the braid
This leg connected to the centre conductor
Symbol
aerialsandtv.com

Fig. 8. Half-wave dipole. *Left side:* Antenna structure. *Right side:* Typical deployment.

ultra low power incident waves and the explanation of the physical phenomena behind most of the recommended efficiency optimization approaches.

This Chapter will show for example that a Schottky diode that delivers the highest efficiency at 0dBm incidence may not necessarily deliver the highest efficiency at lower power incidence e.g. –20dBm. We will therefore classify which diodes perform better at given power incidences; of course, this will also be compared to the diode manufacturers' application notes. Simulations in Agilent's ADS using SPICE and equivalent circuit models will compare the performance of few selected Schottky diodes namely; HSMS-2820, HSMS-2850, HSMS-2860, HSC-276A, and SMS7630. Moreover, the effect of the Schottky diode's junction capacitance (C_j) and junction bias resistance (R_j) on the conversion efficiency will be shown from which, special techniques for Schottky diode harmonic suppression and rectifying circuit loading for maximum efficiency point tracking will be presented.

1.3.1 The schottky diode

The classical *pn* junction diode commonly used at low frequencies has a relatively large junction capacitance that makes it unsuitable for high frequency application [3]. The Schottky barrier diode, however, relies on a semiconductor-metal junction that results in a much lower junction capacitance. This makes Schottky diodes suitable for higher frequency conversion applications like rectification (RF-to-DC conversion) [3]. We will demonstrate the effects of junction capacitance and resistance in the following sub section.

1.3.2 The effect of Schottky diode's C_j and R_j on the conversion efficiency

We have studied Schottky diode's C_j and R_j and published our results in [4]. In this work, we designed a rectifying antenna tuned for use at 2 GHz. The circuit proposed in [4] is a voltage doubler by configuration, but we replaced the amplitude detection diode (series diode) with its equivalent circuit adapted from [5]. The results of this investigation show that variation of C_j shifts the tuned frequency position and also introduces a mismatch in the resonant frequency, see Fig. 9 (left side graph). Therefore for this circuit at 2 GHz, we recommend using a Schottky diode having $C_j = 0.2$pF. In general, a smaller value of C_j is desirable at higher frequencies. Similarly, for R_j investigation, a smaller value is desirable for better matching at 2 GHz for example. If the R_j is increased towards 10kΩ, there is a mismatch in the resonant frequency but no shift in the frequency, see Fig. 9 (right side graph). Another approach to the study of Schottky diodes for higher frequency and efficiency rectenna design is presented in [6].

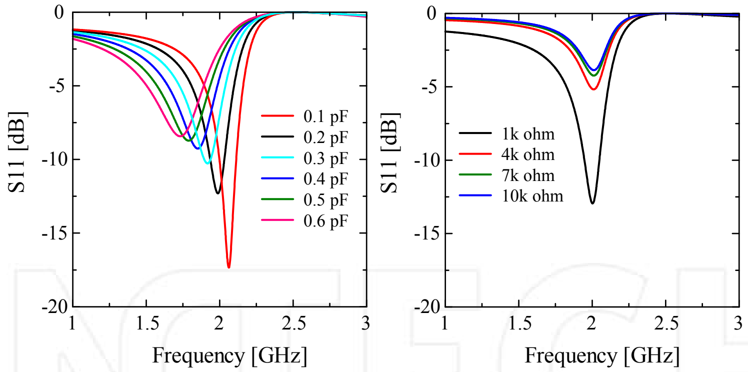

Fig. 9. Schottky diode's C_j effect (Left) and R_j effect (Right) on 2 GHz rectenna's input response.

1.4 Rectifying circuit for DTV energy harvesting

In the design of a DTV energy harvesting circuit, several basic design considerations must be paid attention to. First is the antenna; it must be wideband (covering 470 MHz to 770 MHz), horizontally polarized and omni-directional. Secondly is the rectifier; it must also be wideband, and optimized for RF-to-DC conversion for incident signal power at least -40dBm. Until recently, very few authors have published on DTV energy harvesting circuit. For the few publications, the antenna could not meet all those three requirements and a discussion on the performance of the harvesting circuit for ultra low power incidences has been neglected. In this Chapter we will present such a rectenna with conversion efficiencies above 0.4% at -40dBm, above 18.2% at -20dBm and over 50% at -5dBm signal power incidence. We will closely compare simulated and measured performance of the rectenna and discuss any observed disparities.

Agilent's ADS will be used to simulate the nonlinear behaviour of the rectifying circuit based on harmonic balance tuning methods. To simulate the multiple incident waves, a multi-tone excitation in the DTV band will be invoked. The wideband input characteristic will be achieved by the input matching inductors and capacitors.

The generic version of our proposed DTV energy harvesting circuit is shown below in Fig. 10. The implementation, however, is in two phases or scenarios as follows. First, we investigate the class called "ultra low power" DTV band rectenna. Secondly, we introduce the "medium power" DTV band rectenna.

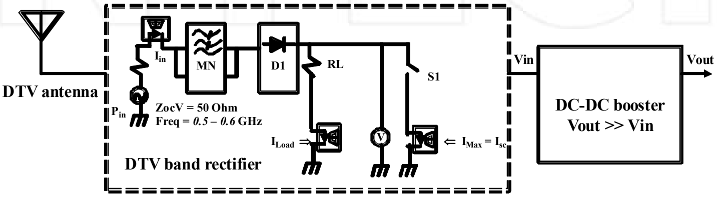

Fig. 10. Generic version of our proposed DTV energy harvesting circuit.

1.4.1 Ultra low power DTV rectenna

We define an ultra low power rectenna as one impinged by RF power incidence in the range between – 40dBm and -15dBm. Below in Fig. 11 is the circuit we designed; optimized for -20dBm input. The matching network is complex so as to achieve a wide band input characteristic. The fabricated circuit was well matched for the frequency range between 470 MHz and 600 MHz. More details about the circuit design can be found in [7].

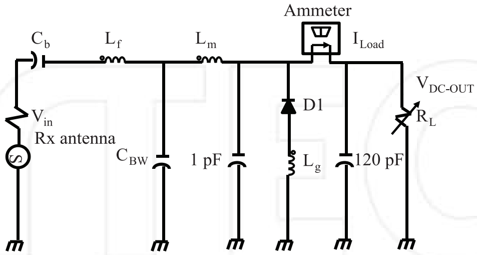

Fig. 11. Ultra low power DTV band rectenna circuit. SMS7630 Schottky diode by SKYWORKS offered the best performance.

The RF-to-DC conversion efficiency for this circuit is shown in Fig. 12 where at input power equal to -40dBm, efficiency is at least 0.4% and rectified voltage equals 1mV; at -20dBm, we have at least 18.2% by measurement and a rectified voltage of 61.7mV. The level of rectified voltage is too low and disqualifies this circuit for purposes of charging capacitors or batteries to accumulate such micropower over time. Instead, boosting the low voltage to usable levels is the option available and we shall discuss this at a later stage, (in Section 2.6).

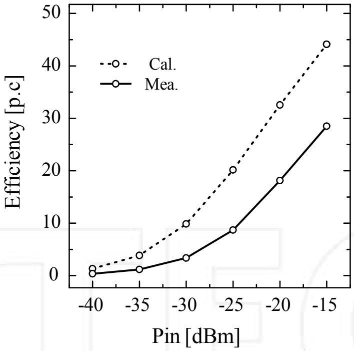

Fig. 12. Ultra low power DTV band rectenna efficiency.

1.4.2 Medium power DTV rectenna

We define a medium power rectenna as one impinged by RF power incidence in the range between – 5dBm and 0dBm. Below in Fig. 13 is the circuit we designed, optimized for -5dBm input. The matching network is simpler than as shown in section 2.4.1 since we require a narrow band around 550 MHz, with received peak power spectrum levels at least -5dBm. The circuit in Fig. 13 is a modification of Greinacher's doubler rectifier. In the circuit, C_b equals 1 pF and is used to block DC current against flowing towards the source. The shunt

INTECH

capacitance, C_{BW} equals 3300 pF and is used to set the input bandwidth. The grounding inductance, L_g equals 56nH (optimal) and is used to improve the RF-to-DC conversion efficiency by cancelling the Schottky diodes (D_b and D_D) capacitive influence; thereby minimizing the harmonic levels (harmonic suppression). We used HSMS2850 diodes in these circuits for their better performance at this level of incident power.

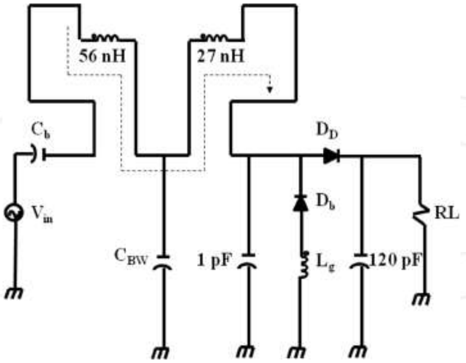

Fig. 13. Medium power DTV rectenna circuit. HSMS 2850 or 2820 from Hewlett-Packard offered the best performance.

The RF-to-DC conversion efficiency for this circuit is shown in Fig. 14 where at input power equal to -5dBm, we achieve at least 50% conversion efficiency by measurement, equivalent to 1.2 V DC rectified at 8.2kΩ optimal load. If we change the load to 47kΩ, over 2 V DC is rectified. This rectenna circuit is ideal for powering small sensors that run on 1.5 V or 2.2 V and draw around 6μA nominal current. If we need to power sensors demanding more power, say at least 2.2 V and 0.3mA to 1.47mA current consumption, we have to accumulate the power in a capacitor over time.

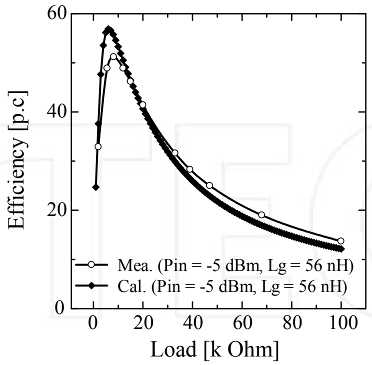

Fig. 14. Medium power DTV band rectenna efficiency.

1.4.3 DTV energy harvesting scenario and application demo

Using the medium power DTV band rectenna, connected to a gold capacitor as an accumulator, energy harvesting was initiated as shown in Fig. 15. Details about the gold capacitor, which include its charge function, backup time and leakage losses are presented in [8]. For the scenario shown in Fig. 15, the accumulated voltage by measurement i.e. capacitor charge function follows the path;

$$V_{acc} = 0.5388\ln(t) + 1.4681 \tag{6}$$

where V_{acc} is the accumulated voltage in volts and t the time in hours. It takes 4.5 hours to accumulate 2.25 V, given a rectified charging voltage and current of 2.4 V and 51µA, respectively, supplied by the DTV band rectenna instantaneously.

With this rectenna, it was possible to power up many different kinds of sensors. Sensors with ultra low power consumption were powered directly, without need to accumulate the power in a capacitor, as shown in Fig. 16.

INTECH

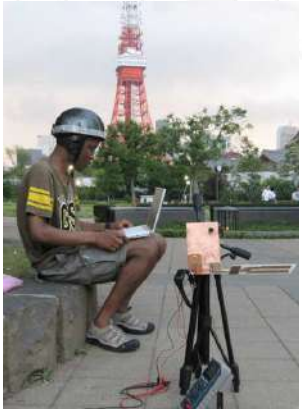

Fig. 15. DTV energy harvesting in a park at some line of sight from the base station.

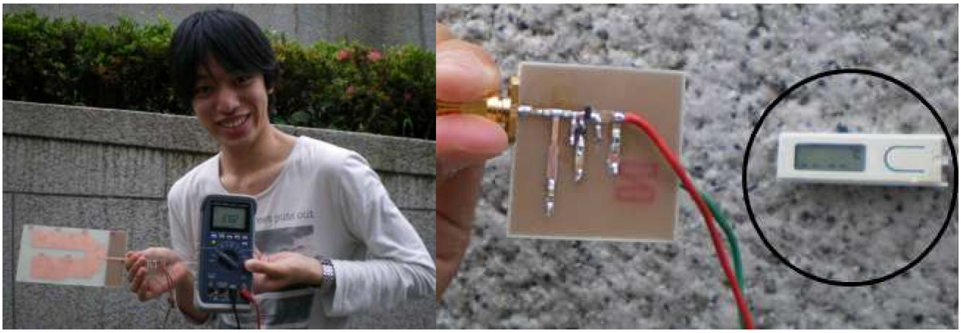

Fig. 16. Directly powering a thermometer mounted on a car park wall (right picture). The maximum instant voltage rectification on record equals 3.7 V (left picture).

1.5 Rectifying circuit for cellular energy harvesting

Unlike in the DTV energy harvesting circuit, for cellular energy harvesting, the antenna must be narrowband (50 MHz bandwidth is acceptable), and circularly polarized even though cellular signals are vertically polarized. The circular polarization is desired to maximize the RF-to-DC conversion efficiency of the arbitrary polarization incident signals in the multipath environment. Similarly, the rectifier must be narrowband, and optimized for RF-to-DC conversion over a wide range of incident signal power.

Thinking about the potential applications for cellular energy harvesting is useful. Other authors have reported on powering a scientific calculator or a temperature sensor from GSM energy harvesting. In this Chapter we will present a special application for energy harvesting in the vicinity of the W-CDMA cellular base station and analyze the system performance by calculation from experimental data. A cellular energy harvesting circuit optimized for over 50% RF-to-DC conversion efficiency given approximately 0dBm incidence will be presented.

1.5.1 Cellular band rectenna

Below in Fig. 17 is the circuit we designed, optimized for 0dBm input. Simple input matching network is ideal since we require a narrow band response around 2.1 GHz. The optimum value for L_g equals 5.6nH, where L_g is used to improve the RF-to-DC conversion efficiency as earlier discussed. HSMS2850 diode was used.

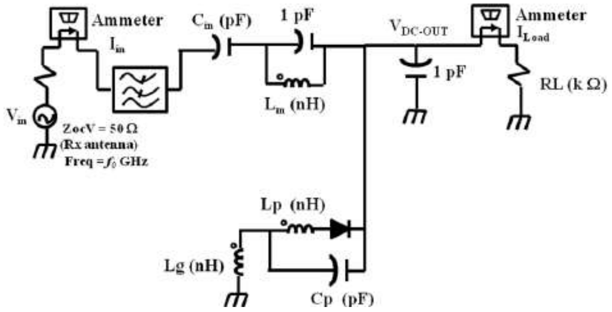

Fig. 17. Shunt rectifier configuration for the cellular band. The matching elements L_m = 3.2nH, while C_{in} =2.5pF. The load resistance is fixed at R_L = 2.1kΩ.

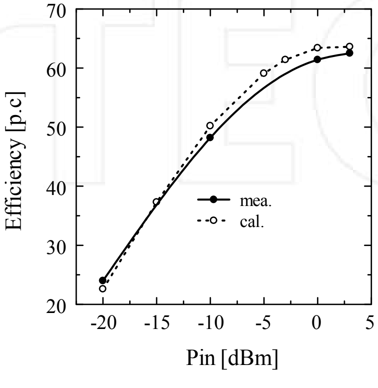

Fig. 18. Conversion efficiency as a function of input power (P_{in}) in the cellular band.

The RF-to-DC conversion efficiency for this circuit is shown in Fig. 18 where at input power equal to 0dBm, we achieve at least 60% conversion efficiency by measurement, given a 2.1kΩ optimal load. This rectenna circuit is ideal for powering small sensors that run on 1.5 V or 2.2 V and 6μA nominal current consumption. If we need to power sensors demanding more power, say at least 2.2 V and 0.3mA to 1.47mA, we have to accumulate the power in a capacitor over time as discussed in section 2.4.3 above.

1.5.2 Cellular energy harvesting application example

Environmental power generation in the neighbourhood of a cellular base station to power a temperature sensor is proposed as shown in Fig. 19 below. Electric field strength measurements in the base station neighbourhood have demonstrated the potential for environmental power generation, and the proposed temperature sensor system is designed based on these values. The rectenna described in Section 2.5.1 is used as the RF-to-DC rectifying circuit with the notched circular microstrip patch antenna (CMPA) proposed in Section 2.2.1. RF-to-DC conversion efficiency equal to 53.8% is obtained by measurement. The temperature sensor made for trial purposes clarifies the capability for temperature data wireless transmission for 20 seconds per every four hours in the base station neighbourhood.

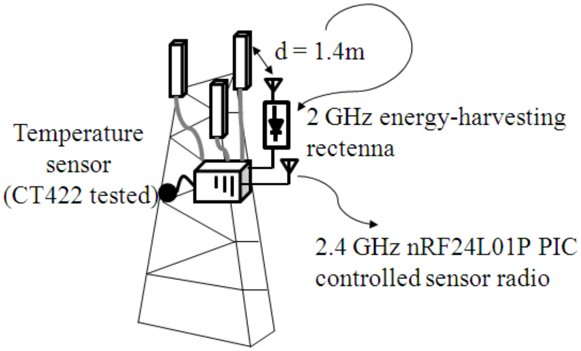
d = 1.4m
2 GHz energy-harvesting rectenna
Temperature sensor (CT422 tested)
2.4 GHz nRF24L01P PIC controlled sensor radio

Fig. 19. Application example in the vicinity of the cellular base station.

1.6 Micropower energy harvesting management

A rectifying antenna circuit for -40dBm incident power harvesting generates 1mV at 2kΩ load, given 0.4% efficiency as presented in Section 2.4.1. At -20dBm incidence and at least 18.2% efficiency, 61.7mV is generated given a 2kΩ load [7]. The generated DC power in both of these two cases is in the μW range, hence the micropower definition. To manage such micropower, power accumulation or energy storage is required. Storage devices may either be a gold capacitor, super capacitor, thin film battery or the next generation flexible paper batteries. These storage devices have specific or standard maximum voltage and trickle charging current minimum requirements. Typically, gold capacitors have voltage ratings like 2.7 V, 5.5 V for 100 μA, 10mA or 100mA maximum discharge current. On the other hand, standard ratings for batteries are 1.8 V, 2 V, 3.3 V and 4.1 V. Therefore, to directly charge any of these storage devices from 1mV, or 61.7mV DC is impractical.

Published works have demonstrated the need for a DC-to-DC boost converter placed between the rectifying antenna circuit (rectenna) and the storage device. Recent efforts have demonstrated that a 40mV rectenna output DC voltage could be boosted to 4.1 V to trickle charge some battery. A Coilcraft transformer with turns ratio ($N_s : N_p$) equal to 100 was used in the boost converter circuit. An IC chip leading manufacturer (Linear Technology Corp., LT Journal, 2010) has released a linear DC-to-DC boost regulator IC chip capable of boosting an input DC voltage as low as 20 mV and supplying a number of possible outputs, specifically suited for energy harvesting applications. While this IC is a great milestone, readers and researchers need to understand the techniques to achieve such ICs and also the limitations that apply. In the following sub section, we will describe the methods toward designing a DC-DC boost converter, suitable for micropower RF energy harvesting.

In the design, we will attempt to clarify the parameters that affect the DC-DC conversion efficiency. For this design, Envelope simulation in Agilents's ADS is used. This simulation technique is the most efficient for the integrated rectenna and DC-DC boost converter circuits.

1.6.1 DC-DC boost converter design theory and operation

The DC-DC boost converter design theory and actual implementation are presented in this section. The inequality $V_{in} \ll V_{out}$ defines the boost operation. In this Chapter, our boost converter concept is illustrated in Fig. 20. A small voltage, V_{in} is presented at the input of the boost converter inductive pump which as a result, generates some output voltage, V_{out}. The output voltage is feedback to provide power for the oscillator. The oscillator generates a square wave, F_{OSC} that is used for gate signalling at the N-MOSFET switch.

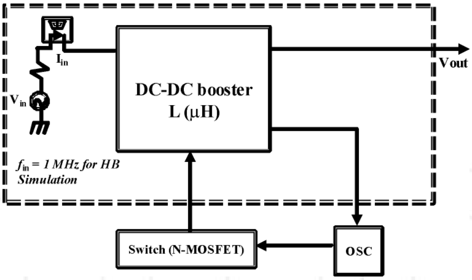

Fig. 20. Boost converter concept.

The drain signal of the N-MOSFET is used as the switch node voltage, V_{sn} at the anode of the diode inside the boost converter circuit block. From the concept presented in Fig. 20, the actual implemented circuit is shown in Fig. 21. The circuit was designed in Agilent's ADS and fabricated for investigation by measurement.

The circuit in Fig. 21 is proposed for investigation. Since a DC-DC boost converter is supposed to connect to the rectenna's output, it therefore, becomes the load to the rectenna circuit. This condition demands that the input impedance of the boost converter circuit emulates the known optimum load of the rectenna circuit. This has the benefit of ensuring maximum power transfer and hence higher overall conversion efficiency from the rectenna input (RF power) to the boost converter output (DC power). In this investigation, as shown in [7], the optimum load for the rectenna is around 2kΩ. In general, emulation resistance R_{em} is given by

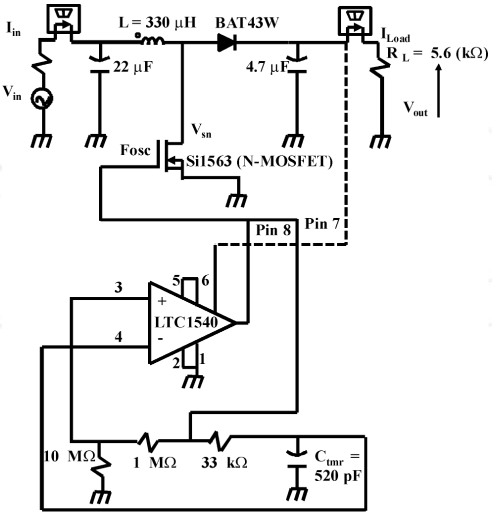

Fig. 21. The proposed boost converter circuit diagram. Designed in Agilent's ADS and fabricated for investigation by measurement.

$$R_{em} = \frac{2LT}{t_1^2 k}\left(\frac{M-1}{M}\right) \tag{7}$$

where L is the inductance equal to 330µH as shown in Fig. 20, $M = \frac{V_{out}}{V_{in}}$, T is the period of F_{OSC}, t_1 is the switch"ON" time for the N-MOSFET, and k is a constant that according to [3] is a low frequency pulse duty cycle if the boost converter is run in a pulsed mode and typically, k may assume values like 0.06 or 0.0483. With reference to (7), we select L as the key parameter for higher conversion efficiency while V_{in} = 0.4 V DC is selected as the lowest start up voltage to achieve oscillations and boost operation. Computing the DC-DC boost conversion efficiency against different values of L, we have results as shown in Fig. 22.

From the results above, L = 100µH is the optimum boost inductance that ensures at least 16.5% DC-DC conversion efficiency, given R_L = 5.6kΩ.

Now having selected the optimum boost inductance given some load resistance, the emulation resistance shown in Fig. 23 is evaluated from the ratio of voltage versus current at the boost converter circuit's input.

The results show a constant resistance value against varying inductance. In general, we can say that this boost converter circuit has a constant low input impedance around 82.5Ω. This impedance is too small to match with the optimum rectenna load at 2kΩ. This directly affects the overall RF-to-DC conversion efficiency.

INTECH

The results show a constant resistance value against varying inductance. In general, we can say that this boost converter circuit has a constant low input impedance around 82.5Ω. This impedance is too small to match with the optimum rectenna load at 2kΩ. This directly affects the overall RF-to-DC conversion efficiency.

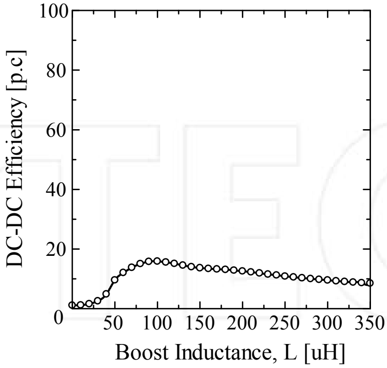

Fig. 22. Boost inductance variation with DC-DC conversion efficiency for a 5.6 kΩ load.

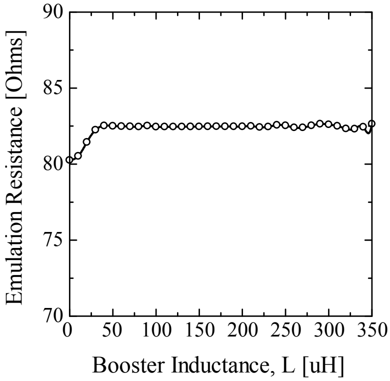

Fig. 23. Boost converter’s input impedance: the emulation resistance.

Another factor, which affects the overall conversion efficiency is the power lost in the oscillator circuit. Unlike the circuit proposed in [9], which uses two oscillators; a low frequency (LF) and high frequency (HF) oscillator; in Fig. 21, we have attempted to use a single oscillator based on the LTC1540 comparator, externally biased as an astable multivibrator.

The power loss in this oscillator is the difference in the DC power measured at Pin 7 (supply) to the power measured at pin 8 (output). We term this loss, L_{OSC}; converted to heat or sinks through the 10MΩ load. A comparison of the oscillator power loss to the power available at the boost converter output is shown in Fig. 24.

Looking at Fig. 24; we notice that the power loss depends on whether the oscillator output is high or low. The low loss corresponds to the quiescent period where the power lost is almost negligible. However, during the active state, the lost power (power consumed by the oscillator) nearly approaches the DC power available at the boost converter output. This results in low operational efficiency.

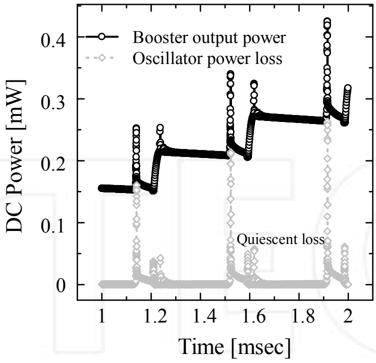

Fig. 24. The power loss in the oscillator.

To confirm whether or not the circuit of Fig. 21 works well, we did some measurements and compared them with the calculated results. Unlike in calculation (simulation), during measurement, L = 330μH was used due to availability. All the other component values remain the same both in calculation and measurement. In Fig. 25 (left side graph) and (right side graph), we see in general that the input voltage is boosted and also that the patterns of F_{osc} and V_{sn} are comparable both by simulation and measurement. To control the duty cycle of the oscillator output (F_{osc}), and the level of ripples in the boost converter output voltage (V_{out}), we change the value of the timing capacitance, C_{tmr} in the circuit of Fig. 21. Simulations in Fig. 25 (left side graph) show that C_{tmr} = 520pF realizes a better performance i.e. nearly constant V_{out} level (very low ripple).

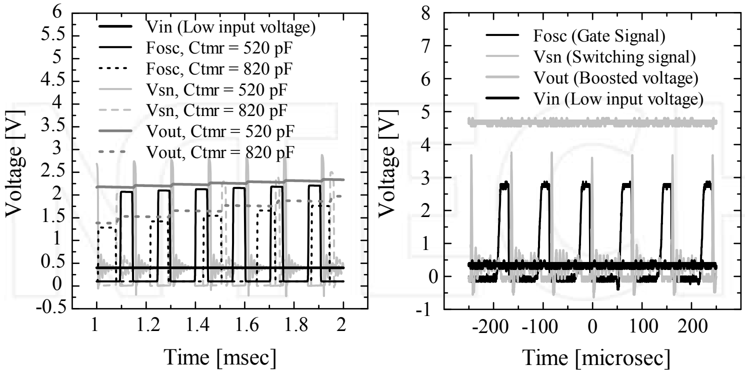

Fig. 25. Voltage characteristics of the developed boost converter circuit. The left side graph represents simulation while the right side graph is for measurements.

Generally, we observe that with this kind of boost converter circuit topology, it is difficult to start up for voltages as low as 61.7mV DC generated by the rectenna at -20dBm power incidence and at least 18.2% rectenna RF-to-DC conversion efficiency. Self starting is the issue for this topology at very low voltages.

At least 11.3% DC-DC conversion efficiency was recorded by measurement and is comparable to the calculation in Fig. 22. During measurement it was clearly revealed that the boost converter efficiency does depend on the value of **L** and the duty cycle derived from $\mathbf{t_1}$. To efficiently simulate the complete circuit, from the RF input to the DC output, envelope transient simulation (ENV) in Agilent's ADS was used. The (ENV) tool is much more computationally efficient than transient simulation (Tran). This simulation is appropriate for the boost converter circuit's resistor emulation task. Moreover, the boost converter's DC-DC conversion efficiency, and the overall RF-to-DC conversion efficiency can be calculated at once with a single envelope transient simulation.

In summary, though not capable to operate for voltages as low as 61.7mV DC, the proposed boost converter has by simulation and measurement demonstrated the capability to boost voltages as low as 400mV DC, sufficient for battery or capacitor recharging, assuming that the battery or the capacitor has some initial charge or energy enough to provide start-up to the boost converter circuit.

The limitations of our proposed boost converter circuit include; low efficiency, lack of self starting at ultra low input voltages, and unregulated output. To address these limitations, circuit optimization is required. Moreover, alternative approaches which employ a flyback transformer to replace the boost converter inductance must be investigated. A regulator circuit with Low Drop Out (LDO) is necessary to fix the boost converter output voltage commensurate with standard values like 2.2 V DC for example. For further reading, see [7]

2. Performance analysis of the complete RF energy harvesting sensor system

To demonstrate how one may analyze the performance of an RF energy harvesting system including its application, we extend the discussion of Section 2.5.2 to this Section. We propose a transmitter assembled as in Fig. 26 for temperature sensor wireless data transmission.

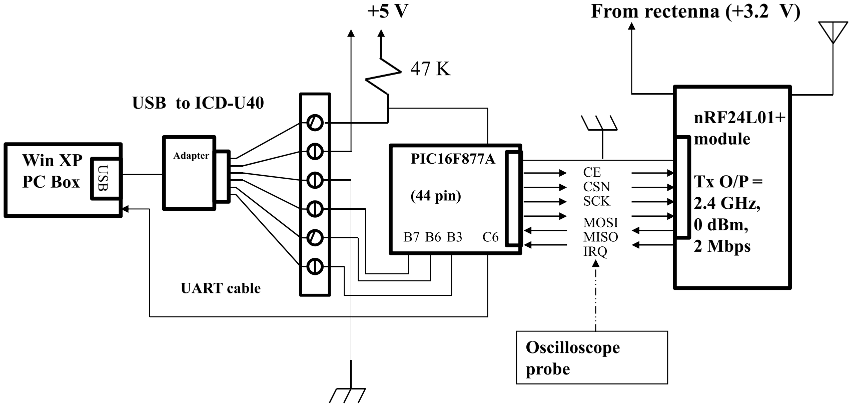

Fig. 26. The assembly and test platform for the proposed battery-free sensor transmitter.

The transmitter consists of one-chip microcomputer (MCU) PIC16F877A and wireless module nRF24L01P for the control, and MCU can be connected with an outside personal computer using ICD-U40 or RS232 cable. The wireless module operates in transmission and reception mode, and controls power supply on-off, transmitting power level, the receiving mode status, and transmission data rate via Serial Peripheral Interface (SPI). Figure 27 shows the operation flow when transmitting.

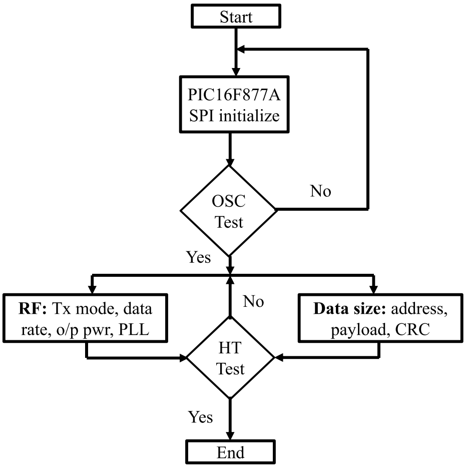
Start
PIC16F877A SPI initialize
OSC Test
No
Yes
RF: Tx mode, data rate, o/p pwr, PLL
No
Data size: address, payload, CRC
HT Test
Yes
End

Fig. 27. Operation flow during transmission.

The experimental system composition is shown in Fig. 28 to transmit acquired data by the temperature sensor with WLAN at 2.4 GHz (ISM band). An ISM band sleeve antenna is used for the transmission. Using the cellular band rectenna shown and discussed in Section 2.5.1, at least 3.14 V is stored in the electric double layer capacitor over a period of four hours. To harvest a maximum usable power for the overall system, we charge the capacitor up to 5V. The operation voltage for the wireless module presented in Fig. 26 above is between 1.9V and 3.6V.

The signal was transmitted from the wireless module while a sleeve antenna, same like the one for transmission was used with the spectrum analyzer and the reception experiment was performed. Received signal level equal to -43.4dBm was obtained at a distance 3.5m between transmitter and reception point. The capacitor's stored voltage was used to supply the wireless module in the above-mentioned experiment. Successful transmission was possible for 5.5 minutes after which, the capacitor terminal voltage decreased from 3.16V to 1.47V, and the transmission ended. The sending and receiving distance of data can be estimated to be about 10m when the sensitivity of the receiver is assumed to be -60dBm, given 0dBm maximum transmit power.

Hereafter, the overall system examination is done by environmental power generation using the transmitted electric waves from the cellular phone base station, proposed based on the above-mentioned results. First of all, the power consumption shown in Fig. 29 is based on the fact that 120mW (5V, 24mA) is saved in the electric double layer capacitor by environmental power generation, achieved by calculation as discussed earlier.

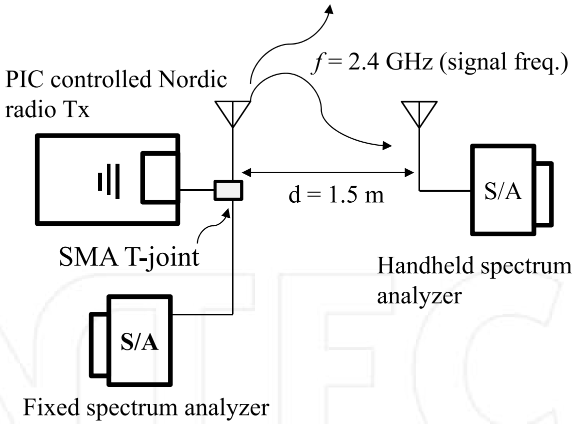
f = 2.4 GHz (signal freq.)
PIC controlled Nordic radio Tx
d = 1.5 m
S/A
SMA T-joint
Handheld spectrum analyzer
S/A
Fixed spectrum analyzer

Fig. 28. Indoor measurement setup for received traffic from the sensor radio transmitter.

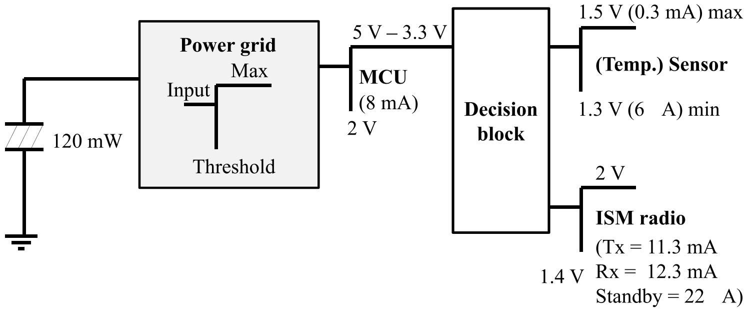
Power grid
Max
Input
Threshold
120 mW
5 V – 3.3 V
MCU
(8 mA)
2 V
Decision block
1.5 V (0.3 mA) max
(Temp.) Sensor
1.3 V (6   A) min
2 V
ISM radio
(Tx = 11.3 mA
Rx = 12.3 mA
Standby = 22   A)
1.4 V

Fig. 29. Power management scheme for the cellular energy-harvesting sensor node.

The sensor data packet is transmitted wirelessly in ShockBurst mode for energy efficient communication. The data packet format includes a pre-amble (1 byte), address (3 bytes), and the payload i.e. temperature data (1 byte). The flag bit is disregarded for easiness, and cyclic redundancy check (CRC) is not used.

The operation of the proposed system is provisionally calculated. When the rectenna is set up in the place where power incidence of 0dBm is obtained in the base station neighbourhood (as depicted in Section 2.5.2), an initially discharged capacitor accumulates up to 3.3V by a rectenna with 53.8% conversion efficiency (presented in Section 2.5.1). At this point, it takes 1.5 minutes to start and to initialize a wireless module, and the voltage of the capacitor decreases to 2V. This trial calculation method depends on the capacitor's back up time discussed in [8]. After this, when the wireless module is assumed to be in sleep mode, the capacitor is charged by a 0.28mA charging current for four hours whereby the capacitor's stored voltage increases up to 5V. The power consumption in the sleep mode or standby is 33µW (1.5V, 22µA).

When the wireless module starts, after data transmission and the confirmation signal is sent, the voltage of the capacitor decreases by 0.6V, and consumes the electric power of 7.4mW.

The voltage of the capacitor decreases to 2V when 3.2mW is consumed to the acquisition of the sensor data, and the operation time of MCU is assumed to be one minute to the data storage in the wireless module etc. As for the capacitor voltage, when the wireless module continuously transmits data for 20 seconds, it decreases from 2V to 1.4V and even the following operation saves the electric power. Therefore, a temperature sensing system capable of transmitting wireless data in every four hours becomes feasible by environmental power generation from the cellular phone base station if we consider intermittent operation by sleep mode.

3. Conclusion

This Chapter has given an overview of the present energy harvesting sources, but the focus has stayed on RF energy sources and future directions for research. Design issues in RF energy harvesting have been discussed, which include low conversion efficiency and sometimes low rectified power. Solutions have been suggested by calculation and validated by measurement where possible, while highlighting the limitations of the proposed solutions. Potential applications for both DTV and cellular RF energy harvesting have been proposed and demonstrated with simple examples. A discussion is also presented on the typical performance analysis for the proposed RF energy harvesting system with sensor application.

4. Acknowledgment

The authors would like to thank Prof. Apostolos Georgiadis of Centre Tecnològic de Telecomunicacions de Catalunya (CTTC, Spain) for the collaboration on the design and development of the DC-DC boost converter circuit. Further thanks go to all those readers who will find this Chapter useful in one way or the other.

5. References

[1] Keisuke, T.; Kawahara, Y. & Asami, T. (2009). *RF Energy Intensity Survey in Tokyo..*, (c)2009 IEICE, B-20-3, Matsuyama-shi, Japan

[2] Mikeka, C.; Arai, H. (2011). *Dual-Band RF Energy-Harvesting Circuit for Range Enhancement in Passive Tags*, (c)2011 EuCAP, Rome, Italy

[3] Pozar, D. (2005). *Microwave Engineering*, Wiley, ISBN 978-0-471-44878-5, Amherst, MA, USA

[4] Mikeka, C.; Arai, H. (2010). Techniques for the Development of a Highly Efficient Rectenna for the Next Generation Batteryless System Applications, *IEICE Tech. Rep.*, pp. 101-106, Kyoto, Japan, March, 2010

[5] http://www.secomtel.com/UpFilesPDF/PDF/Agilent/PDF_DOCS/SKYDIODE/ 03_SKYDI/HSMS2850.PDF (Last accessed on 13 July, 2011)

[6] McSpadden, J. et al., H. (1992). Theoretical and Experimental Investigation of a Rectenna Element for Microwave Power Transmission, *IEEE Trans., on Microwave Theory and Tech., Vol. 40, No. 12.*, pp. 2359-2366, Dec., 1992

[7] Mikeka, C.; Arai, H. ; Georgiadis A. ; and Collado A. (2011). DTV Band Micropower RF Energy-Harvesting Circuit Architecture and Performance Analysis, *RFID-TA Digest*, Sitges, Spain, Sept., 2011

[8] Mikeka, C.; Arai, H. (2009). Design of a Cellular Energy-Harvesting Radio, *Proc. 2 nd European Wireless Technology Conf.*, pp. 73-76, Rome, Italy, Sept., 2009

[9] Popovic Z., et al., (2008). Resistor Emulation Approach to Low-Power RF Energy Harvesting, *IEEE Trans. Power Electronics, Vol. 23, No. 3*, 2008

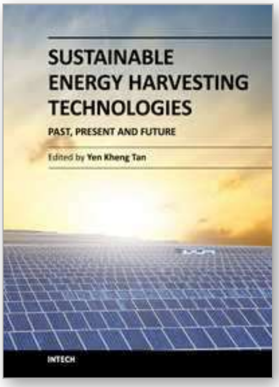

Sustainable Energy Harvesting Technologies - Past, Present and Future
Edited by Dr. Yen Kheng Tan

ISBN 978-953-307-438-2
Hard cover, 256 pages
Publisher InTech
Published online 22, December, 2011
Published in print edition December, 2011

In the early 21st century, research and development of sustainable energy harvesting (EH) technologies have started. Since then, many EH technologies have evolved, advanced and even been successfully developed into hardware prototypes for sustaining the operational lifetime of low?power electronic devices like mobile gadgets, smart wireless sensor networks, etc. Energy harvesting is a technology that harvests freely available renewable energy from the ambient environment to recharge or put used energy back into the energy storage devices without the hassle of disrupting or even discontinuing the normal operation of the specific application. With the prior knowledge and experience developed over a decade ago, progress of sustainable EH technologies research is still intact and ongoing. EH technologies are starting to mature and strong synergies are formulating with dedicate application areas. To move forward, now would be a good time to setup a review and brainstorm session to evaluate the past, investigate and think through the present and understand and plan for the future sustainable energy harvesting technologies.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Chomora Mikeka and Hiroyuki Arai (2011). Design Issues in Radio Frequency Energy Harvesting System, Sustainable Energy Harvesting Technologies - Past, Present and Future, Dr. Yen Kheng Tan (Ed.), ISBN: 978-953-307-438-2, InTech, Available from: http://www.intechopen.com/books/sustainable-energy-harvesting-technologies-past-present-and-future/design-issues-in-radio-frequency-energy-harvesting-system

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:

ELECTRICAL AND ELECTRONIC ENGINEERING

ELECTRONIC AND SPINTRONIC DEVICES

Received
4 September 2014

Accepted
19 February 2015

Published
7 May 2015

Correspondence and requests for materials should be addressed to N.S. (negin.shariati@rmit.edu.au)

Multi-Service Highly Sensitive Rectifier for Enhanced RF Energy Scavenging

Negin Shariati, Wayne S. T. Rowe, James R. Scott & Kamran Ghorbani

School of Electrical and Computer Engineering, RMIT University, Melbourne, VIC 3001, Australia.

Due to the growing implications of energy costs and carbon footprints, the need to adopt inexpensive, green energy harvesting strategies are of paramount importance for the long-term conservation of the environment and the global economy. To address this, the feasibility of harvesting low power density ambient RF energy simultaneously from multiple sources is examined. A high efficiency multi-resonant rectifier is proposed, which operates at two frequency bands (478–496 and 852–869 MHz) and exhibits favorable impedance matching over a broad input power range (−40 to −10 dBm). Simulation and experimental results of input reflection coefficient and rectified output power are in excellent agreement, demonstrating the usefulness of this innovative low-power rectification technique. Measurement results indicate an effective efficiency of 54.3%, and an output DC voltage of 772.8 mV is achieved for a multi-tone input power of −10 dBm. Furthermore, the measured output DC power from harvesting RF energy from multiple services concurrently exhibits a 3.14 and 7.24 fold increase over single frequency rectification at 490 and 860 MHz respectively. Therefore, the proposed multi-service highly sensitive rectifier is a promising technique for providing a sustainable energy source for low power applications in urban environments.

AMBIENT energy harvesting is attracting widespread interest as it has the potential to provide a sustainable energy source for future growth and protection of the environment. Considerable research effort has been directed toward low-profile, low-power, energy efficient and self-sustainable devices aiming to harvest energy from inexhaustible sources such as solar energy, thermal, biomass, mechanical sources (e.g. wind, kinetic, vibration, and ocean waves) wastewater, and microwave energy. A thorough set of reviews is given in the literature[1–4]. Among these green energy sources, there has been a growing interest for radio frequency (RF) energy scavenging, as the availability of ambient RF energy has increased due to advancements in broadcasting and wireless communication systems. Furthermore, the development of wireless power transmission (WPT) technologies[5] that allow micro sensors[6], mobile electronic devices[7], wireless implantable neural interfaces[8] and far-field passive RFID (Radio-Frequency Identification) systems[9–11] to operate without batteries has triggered impetus for RF energy harvesting.

Efficient RF energy harvesting is a very challenging issue, as it deals with the very low RF power levels available in the environment. Furthermore, the scavengeable power level can vary unpredictably, depending on several factors such as the distance from the power source, the transmission media, the telecommunication traffic density and the antenna orientation. The majority of available literature on RF rectification has been dedicated to narrowband rectennas, which essentially operate at a single frequency and hence provide low DC output power[12,13]. Various topologies, such as voltage doublers or multipliers have been employed in order to increase the RF to DC conversion efficiency and the output DC voltage for specific applications[14–16]. However, from an ambient RF scavenging perspective, harvesting energy from various available frequencies could maximize power collection and hence increase the output DC power. Ultra-wideband and broadband rectenna arrays have been proposed as a potential solution[17,18]. However in some cases, simulation and experimental results were not provided to demonstrate the findings[17]. A broadband rectenna consisting of a dual-circularly polarized spiral rectenna array operating over a frequency range of 2–18 GHz was demonstrated[18]. The rectified DC power was characterized as a function of DC load, RF frequency and polarization for power densities between 10^{-5} and 10^{-1} mW/cm^2. However, the proposed rectenna was matched at a single input RF power level for a specified load resistance for the characterization. Also, due to the low Q value of the rectifier circuit, the conversion efficiency was a fraction of 1% at −15.5 dBm. From a design point of view, while it is relatively easy to achieve a broadband antenna, it is very challenging to realize a broadband rectenna due to the non-linearity of the rectifier impedance with input power across the frequency band[19].

To address this, a promising approach is to use a dual-band or multi-band configuration. This can maximize the power conversion efficiency (PCE) at the specific frequencies where the maximum ambient signal level is available. Various dual-band RF energy harvesting systems has been demonstrated[20–24], however a large signal analysis of the rectifier was commonly not provided over a broad input power range. A dual-band RF energy harvesting using frequency limited dual-band impedance matching has been proposed[20] and the PCE was shown over a high power range of 0 to 160 mW, however it was only matched at a single input power level (10 dBm). A CMOS dual-narrowband energy harvester circuit was modeled at environmental power levels[21]. Again, the rectifier efficiency was demonstrated with only single input power levels of −19 and −19.3 dBm at 2 GHz and 900 MHz respectively, and a large signal analysis was not presented. A compact dual-band rectenna operating at 915 MHz and 2.45 GHz has been demonstrated[22] and the PCE was shown for input power levels of −15, −9 and −3 dBm. However, the reflection coefficient was evaluated at a single incident power level. Furthermore, the efficiency results with dual-tone excitation simultaneously and single-tone excitation (at 915 MHz) are very similar, hence the impact of applying a dual-band technique does not demonstrate a clear advantage over a single band. A dual-frequency rectenna for WPT has been proposed[23] which achieved a conversion efficiency of 84.4% and 82.7% at 2.45 and 5.8 GHz with a high input power level of 89.84 and 49.09 mW respectively. These power levels far exceed ambient levels in the environment[19]. A conformal hybrid solar and electromagnetic (EM) energy harvesting rectenna has been presented[24] and the PCE was provided with −30 to 5 dBm input power, achieving an efficiency up to 40% at 1.85 GHz for higher input power levels (above −5 dBm). However the reflection coefficient was not provided at low input power range.

A multi-resonant rectenna that uses a multi-layer antenna and rectifier has been evaluated for a −16 dBm to +8 dBm RF received power level, but the rectifier circuit layout and large signal analysis were not provided to clarify the findings[25]. Furthermore, a rectenna for triple-band biotelemetry communications has been proposed using a triple-band antenna and single frequency rectifier[26]. However, this rectenna is not suitable for RF energy scavenging due to the low efficiency at lower input power levels. Another triple band rectenna presented an RF-DC efficiency over the input power range of −14 to +20 dBm[27], however the reflection coefficient results were only evaluated at a single input power level. This rectenna was shown to harvest 7.06 μW of DC power from three sources simultaneously at a high input power level of +10 dBm. A multi-band harvesting system has also been proposed where four individual harvesters are designed to cover four frequency bands[28]. However, a large signal analysis was not provided over a broad input power range. Furthermore, the proposed harvesting system has a minimum sensitivity of −25 dBm, whilst in a real environment more sensitive systems are required as the available RF power levels are very low[19].

Tunable impedance matching networks have been demonstrated in order to collect RF signals from various sources and convert them to DC power[29]. However from an application point of view, this is still single frequency rectification and it is not widely applicable to environmental RF energy scavenging where the available power is very low.

In order to increase the amount of RF energy scavenged by a rectenna, it is crucial to identify and harvest multiple ambient frequency sources over their realistic available energy range. Our previous research has demonstrated the feasibility of RF energy harvesting through RF field investigations and maximum available power analysis in metropolitan areas of Melbourne, Australia[19]. The maximum available power for different frequency bands based on antenna aperture and number of antennas in a given collection area was analyzed. Measured results and analysis indicated that cellular systems and broadcast sources are well suited to harvesting, with scavengeable RF power ranging from −40 to −10 dBm. This identifies two important considerations in the design of efficient rectenna for RF energy harvesting: the scavengeable ambient RF power sources available, and the significant variance of this power.

The RF to DC rectifier solutions proposed in recent literature have focused on maximizing the system efficiency at a given, and often quite high, input power level. This neglects the issues related to input power variation which can lead to unexpected variations in the matching network due to diode non-linearity. Also, the scavangeable levels of ambient RF power have been shown to be orders of magnitude lower. Therefore, based on our previous research outcomes and recommendations[19], an efficient power harvesting solution could encompass a multi-band matching circuit at the specific frequencies where maximum signal power is available, enabling greater power harvesting due to the combination of RF signals. This also results in a higher power being fed to a single rectifier, utilizing the diode function more efficiently.

This paper presents an RF energy harvesting method that can scavenge a wide range of ambient power levels which are orders of magnitude lower than previous reported techniques in the literature. An efficient dual resonant rectifier circuit is proposed, matched to a 50 Ω input port at 490 and 860 MHz over a broad low input RF power range from −40 to −10 dBm. The proposed dual resonant matching network operates efficiently at two identified harvesting frequency bands over a wide input power range, maximizing DC power by scavenging two sources simultaneously.

The remainder of this paper is organized as follows. First, the key results for the reflection coefficient and output DC power are presented. Subsequently, the Discussion section summarizes the results and demonstrates their potential implications, the limitations of this study, open questions and future research. Finally, the Method section describes the proposed rectifier design.

Results

A dual resonant rectifier was fabricated on a 1.58 mm FR-4 substrate with a dielectric constant $\varepsilon_r \approx 4.5$ and a loss tangent $\delta \approx 0.025$. These substrate parameters were measured using the Nicolson-Ross method[30] so accurate values could be used in the rectifier design. A photograph of the fabricated dual resonant rectifier is shown in Fig. 1 which depicts input RF port, dual-band matching network lumped components, Schottky diodes and the output terminal. The performance of the rectifier was verified by measuring the input reflection properties, and the output power was calculated from the measured output DC voltage for the input powers from −40 to −10 dBm.

Reflection Coefficient. The $|S_{11}|$ of the rectifier was evaluated using a vector network analyzer (VNA). The VNA was re-calibrated for each input power level. Figure 2 compares the simulated and measured $|S_{11}|$ versus frequency for the dual resonant rectifier circuit at four different input power levels from −40 to −10 dBm. The measured results show very good agreement to the simulations. Slightly higher reflection was observed for the resonant frequencies at the lower part of the input power range (due to the diode characteristics). However, the proposed rectifier circuit is well-matched ($|S_{11}| < -10$ dB) at the desired frequency bands of

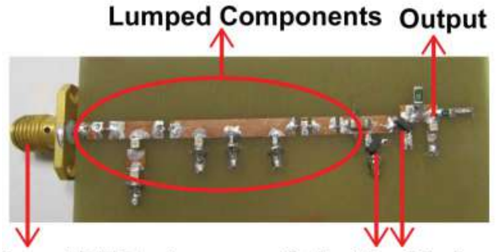

Figure 1 | Fabricated rectifier prototype.

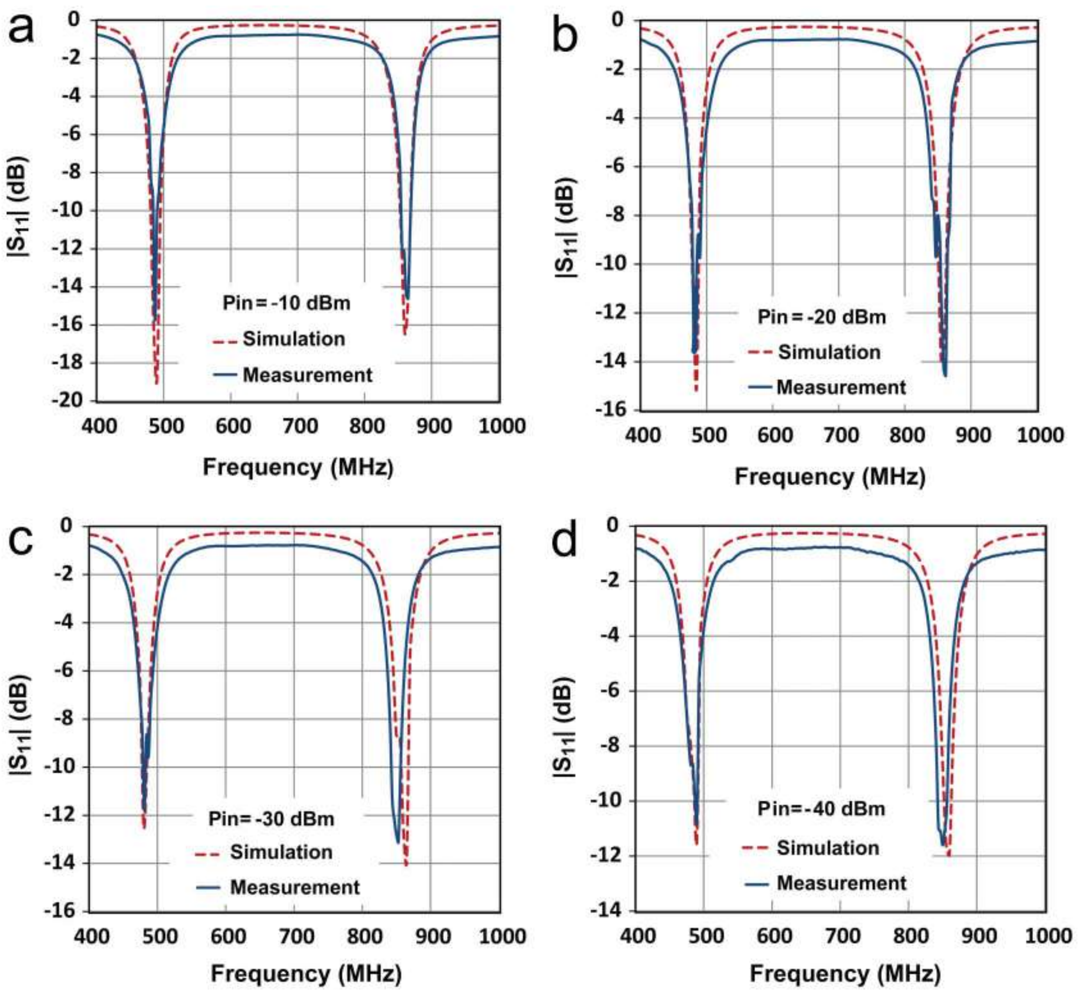

Figure 2 | Simulated and measured $|S_{11}|$ as a function of frequency and input RF power for the proposed dual resonant rectifier circuit. (a) −10 dBm. (b) −20 dBm. (c) −30 dBm. (d) −40 dBm.

478–496 MHz and 852–869 MHz over the broad range of input powers from −40 to −10 dBm. The small difference between simulation and measurement is due to the parasitic extraction accuracy.

Output DC Power. In the frequency domain, the Harmonic Balance method of analysis provides a comprehensive treatment of a multispectral problem[18]. The method intrinsically takes into account the DC component and a specified number of harmonics, while allowing the ability to specify the source impedance and harmonic terminations. A Harmonic Balance simulation was used to numerically evaluate the output DC voltage of the dual resonant rectifier for both a single and two tone input. The output DC voltage across the load resistor was also measured and used to calculate output DC power. Measurements were performed using a Wiltron 68247B synthesized signal generator as a RF power source for the rectifier circuit. Recording of the output DC voltage across the load resistance was achieved with a Fluke 79III digital voltage meter. The RF source power was initially set at −10 dBm, and decreased in 2 dB steps. In the dual-band measurement case, two RF signal generators were fed to the rectifier circuit simultaneously via a power combiner.

The simulation and measurement results for single and dual input tones are summarized in Fig. 3(a) and (b). A measured DC voltage of 772.8 mV is achieved with two simultaneous input tones at an input power of −10 dBm. For single tone measurements, DC voltages of 436 mV and 286 mV at 490 MHz and 860 MHz respectively are produced. The comparison between the 490 and 860 MHz single rectifiers highlights the impact of the input frequency on the PCE. A higher amount of DC voltage can be generated at the lower frequency. This difference comes from decreasing diode performance at the higher frequency due to the higher junction capacitance of the diode[31].

Importantly, a slightly higher DC voltage can be generated with the dual resonant rectifier as compared to the sum of output voltage from the two single bands, particularly at the lower input power levels as can be seen in Fig. 3(b) which shows the lower power section of Fig. 3(a) in more detail. By maximizing power collection from various sources of different frequencies and delivering the combined power to the rectification circuit, the diode conversion efficiency is enhanced which results in a higher level of rectified voltage. Figure 4 compares the simulated and measured output DC power for the dual resonant rectifier circuit with both single and dual input tones. A measured DC power of 17.3 μW and 7.5 μW can be generated at 490 MHz and 860 MHz respectively with a single tone input of −10 dBm (100 μW). This represents true efficiencies of 17.3% and 7.5% for the individual single band rectification (see Fig. 5). However, the measured DC output power with two concurrent input tones of −10 dBm is 54.3 μW which corresponds to an effective efficiency of 54.3% for the dual-band rectifier (see Fig. 6). This represents a 3.14 and 7.24 times increase in output DC power over the single tone excitation at 490 MHz or 860 MHz respectively. This trend is evident down to low input power levels (around 40 μW). Furthermore, there is a significant increase in the PCE of the dual resonant rectifier for lower input power levels (<40 μW).

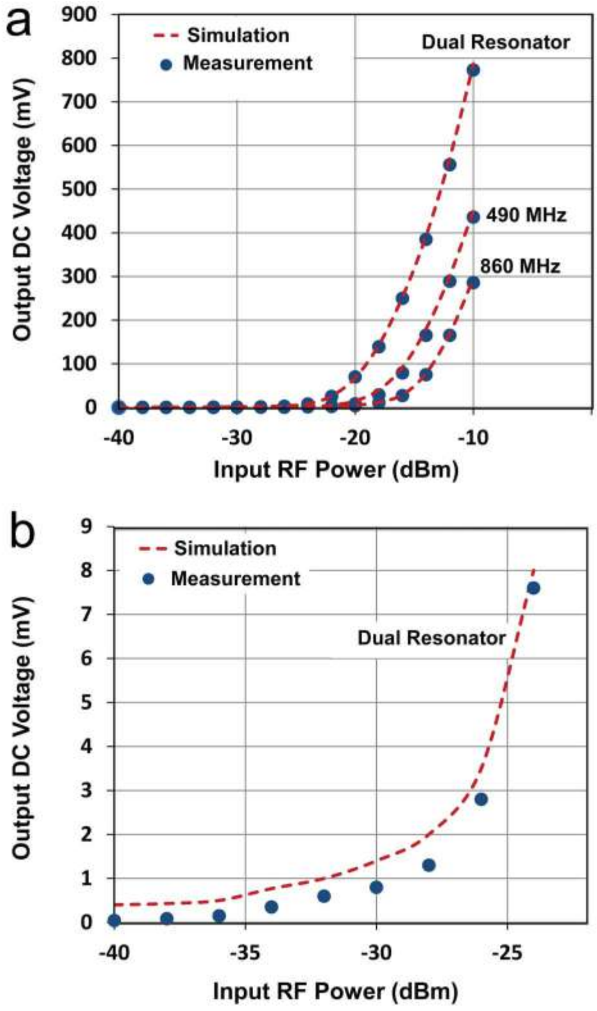

Figure 3 | Output DC voltage as a function of input RF power for single input tone at both 490 MHz and 860 MHz and for dual input tones (a) with −40 to −10 dBm input RF power (b) with −40 to −25 dBm input RF power. (This power range is associated with the signal source).

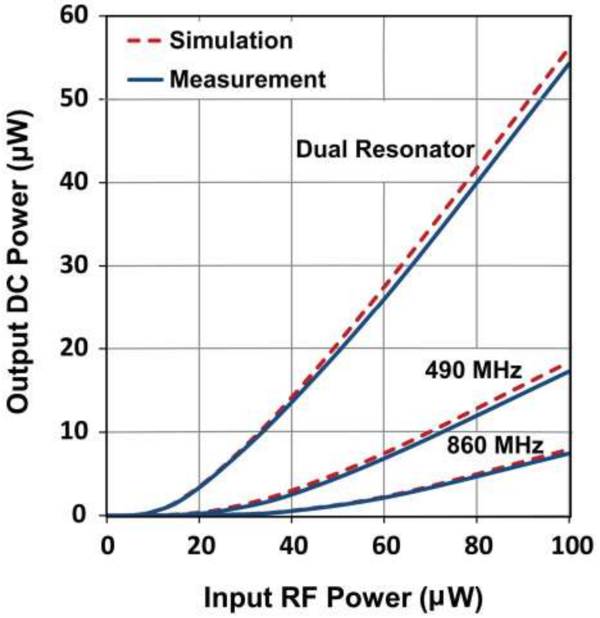

Figure 4 | Output DC power as a function of input RF power for single and dual input tones.

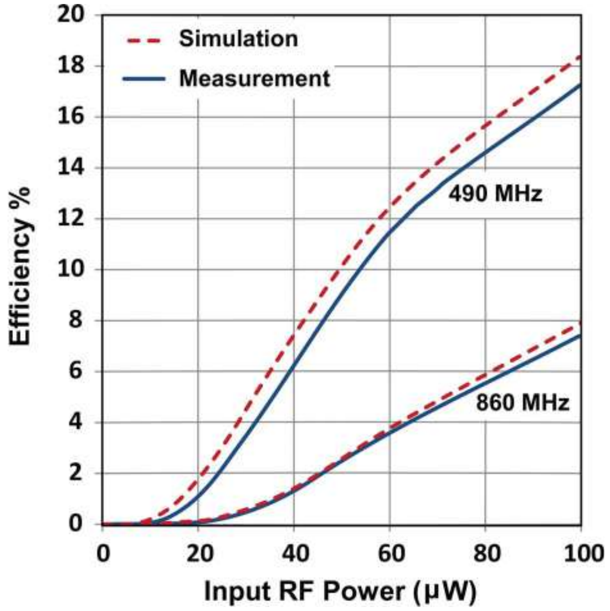

Figure 5 | RF to DC conversion efficiency as a function of input RF power for single band rectification.

Here, the effective efficiency is defined as the ratio of output DC power to the available input RF power rather than the power delivered to the diodes (equation (1)). The available power level is associated with the signal source. For the single resonator, the available input power is −10 dBm and the delivered power is also −10 dBm (assuming no loss). However, by creating a dual resonant matching network the power delivered to the diodes is −7 dBm (combined total input power from two signal generators) but the available power is still −10 dBm.

$$\eta = \frac{p_o}{p_i} = \frac{\left(\frac{{V_o}^2}{R_L}\right)}{p_i} \qquad (1)$$

Therefore, combining input RF signals into a single rectification stage results in high sensitivity rectifier, which is widely applicable to

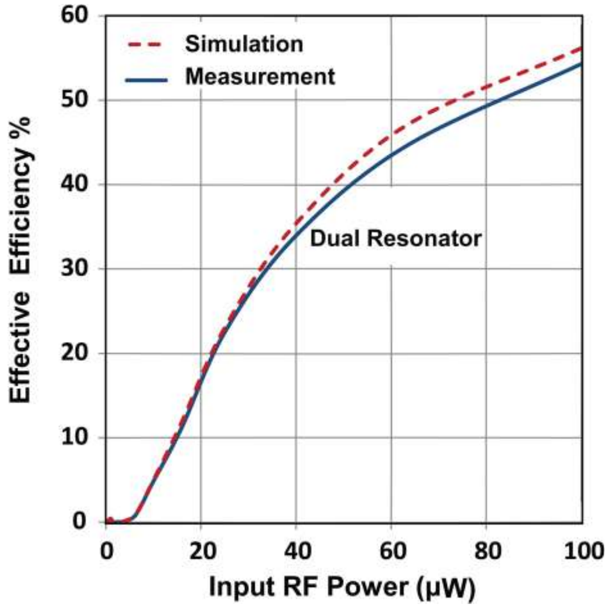

Figure 6 | Effective RF to DC conversion efficiency as a function of input RF power for dual resonant rectification.

Table 1 | Rectifier performance comparison

Ref.	Technology	Measured Efficiency (%)	RF power variation (in PCE evaluation)	Rectification Technique
13	Schottky diode	82@50 mW	N/A	Single resonator
16	Schottky diode	44@−10 dBm	N/A	Single resonator
18	Schottky diode	20@0.07 mW/cm^2 0.1@5×10^{-5} mW/cm^2	10^{-5} to10^{-1} mW/cm^2	Broad band
20	Schottky diode	77.13@22 dBm (158.49 mW)	0 to 160 mW	Dual resonator
21	CMOS	9.1@−19.3 dBm (900 MHz) 8.9@−19 dBm (2 GHz)	N/A	Dual resonator
22	Schottky diode	37 (915 MHz)@−9 dBm 30 (2.45 MHz)@−9 dBm	−40 to 0 dBm	Dual resonator
23	Schottky diode	84.4@89.84 mW (2.45 GHz) 82.7@49.09 mW (5.8 GHz)	0 to 100 mW	Dual resonator
27	Schottky diode	80@10 dBm (940 MHz) 47@8 dBm (1.95 GHz) 43@16 dBm (2.44 GHz)	−14 to 20 dBm	Triple resonator
29	Schottky diode	50@−5 dBm	−25 to 0 dBm	Dual resonator with tunable input response
This work	Schottky diode	54.3@−10 dBm 11.25@−18 dBm (490 and 860 MHz)	−40 to −10 dBm	Dual resonator

Table 2 | Environmental measurement results

Suburb	Available frequencies (MHz)	Respective available RF power (dBm) [μW]	Measured DC power (μW)
Bayswater	486, 488, 489, 490, 491, 867, 868, 869, 870, 871, 872, 873, 874	−19[12.5], −20[10], −17[19.95], −15[31.62], −22[6.3], −37[0.199], −37[0.199], −30[1], −24[3.98], −20[10], −30[1], −37[0.199], −40[0.1]	39.38
Bentleigh	491, 492, 494, 495, 865, 866, 867, 868, 869, 870, 871	−12[63.09], −46[0.02], −42[0.063], −57[0.001], −27[1.99], −27[1.99], −30[1], −37[0.199], −40[0.1], −40[0.1], −41[0.07]	30.9
RMIT University (Melbourne CBD)	487, 488, 489, 490, 491, 851, 861, 862, 866, 867, 868, 869	−30[1], −22[6.3], −29[1.25], −22[6.3], −20[10], −23[5.01], −21[7.94], −21[7.94], −30[1], −35[0.31], −40[0.1], −40[0.1]	14.5

real environmental RF energy scavenging. This multi-band technique can provide higher DC power than combining two separate single frequency rectifier circuits operating at the same frequencies. This is due to the fact that harvesting RF energy from various available sources simultaneously increases the delivered power to the rectifier, which improves the diode conversion efficiency and consequently enhances the output DC power. Table 1 summarizes this work as compared to previous published work.

In order to provide a realistic scenario for the proposed dual-band rectifier, measurement results were taken in three suburbs of Melbourne, Australia, congruent with our previous research outcomes[19]. Table 2 summarizes these environmental measurement results. It should be noted that the lower band (478–496 MHz) has a 3.67% fractional bandwidth and the higher band (852–869 MHz) has around 2% fractional bandwidth. Hence, various RF frequencies from different sources can be harvested within these two bands. The environmental measurement results demonstrate the feasibility of harvesting ambient EM energy from multiple sources simultaneously.

Discussion

The feasibility of harvesting ambient EM energy from multiple sources simultaneously is investigated in this paper. The proposed dual resonant rectifier operates at two frequency bands (478–496 and 852–869 MHz), which are used for broadcasting and cellular systems respectively. The dual resonant rectifier exhibits favorable impedance matching over a broad input power range (−40 to −10 dBm) at these two bands. The achieved sensitivity and dynamic range demonstrate the usefulness of this innovative low input power rectification technique. Simulation and experimental results of input reflection coefficient and rectified output power are in excellent agreement. The measurement results demonstrate that a two tone input to the proposed dual-band RF energy harvesting system can generate 3.14 and 7.24 times more power than a single tone at 490 or 860 MHz respectively, resulting in a measured effective efficiency of 54.3% for a dual-tone input power of −10 dBm. It is evident that this dual resonant rectification technique increases the RF to DC effective conversion efficiency, and hence the recoverable DC power for low power applications. Furthermore from a design and economic perspective, utilizing a large number of components (e.g. antennas, diodes) to realize individual rectifier circuits for each frequency band creates additional expense. In order to provide more realistic measurement results, the proposed dual-band rectifier was tested in three suburbs of Melbourne, Australia. Therefore, this dual-band technique offers a simple and cost-effective solution which is of paramount importance for environmental power harvesting systems. This innovative technique has the potential to generate a viable perpetual energy source for low power applications in urban environments.

Limitation of the study, open questions and future work

Utilizing diodes which are more suitable to low power applications ($P_i < -20$ dBm) could increase the voltage sensitivity, resulting in a higher RF-DC conversion efficiency[32]. Applying a power optimized waveform excitation to the rectifier circuit in these frequency bands, a higher amount of DC power can be generated when compared with a single and dual tone excitations with the same input power[33,34]. However this technique is not applicable to energy harvesting where the input waveform is arbitrary.

Utilizing our proposed dual resonant rectification technique to combine resonant circuits for any other arbitrary frequency bands could lead to PCE improvement, provided that suitable diodes for the desired frequency bands are selected. Note that by increasing the operating frequency the rectification performance degrades due to

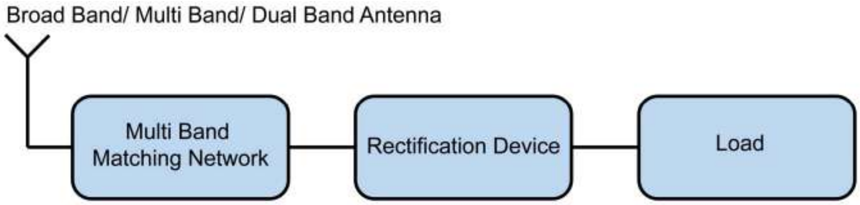
Broad Band/ Multi Band/ Dual Band Antenna
Multi Band Matching Network
Rectification Device
Load

Figure 7 | **General block diagram of the RF energy harvesting system.**

the higher junction capacitance of the diode. Hence, lower output voltage is expected at higher frequency bands.

It is the object of our future work to design a multi-band rectenna array for enhanced RF energy harvesting. Furthermore, increasing the bandwidth, sensitivity and efficiency will also be investigated.

Methods

The major goal in designing an efficient RF harvesting system is to produce high DC output power. Toward this goal, a high sensitivity rectifier is crucial for optimum RF scavenging. A significant factor governing the sensitivity of a rectifier is the threshold voltage of the diode used for rectification. The diode must be able to "switch on" for very low ambient energy levels.

To address this sensitivity issue, a system that scavenges power from multiple frequency bands and combines them to activate a rectification circuit is proposed. The general block diagram of the proposed system is depicted in Fig. 7. Various environmental RF energy sources of different frequencies are collected by an appropriately designed antenna, and delivered to the rectification circuit via a multi-band matching network. The rectification circuit converts the combination of RF signals into DC power for low-power applications. The embodiment in this paper realizes a dual resonant matching circuit as a transition between a 50 Ω nominal antenna output and the non-linear rectification device at 490 and 860 MHz. Based on the Australian Radiofrequency Spectrum Plan[35], these bands are allocated to broadcasting services and cellular systems.

Device Selection. Due to the very low ambient power available in a real environment[19], a very low threshold voltage rectification device is required in order to increase sensitivity. For this reason, Schottky diodes (GaAs or Si) are commonly employed for RF energy harvesting. In this work, a microwave Schottky detector HSMS2820 ($C_{j0} = 0.7$ pF, $R_s = 6\ \Omega$, $I_s = 2.2e^{-8}$ A) is chosen due to its excellent high frequency performance, low series resistance (R_s) and junction capacitance (C_j), and low threshold voltage with high-saturation current[31]. This low threshold voltage (0.15–0.3 V) supports rectification at low input power levels.

Proposed Rectifier Design. In order to design an efficient RF harvesting system, the non-linearity of the rectifier impedance with frequency and input power should be matched to the 50 Ω output of the antenna at the desired frequency bands. Therefore, the diode input impedance as a function of frequency and different power levels were calculated and analyzed[36]. In order to match the input impedance of the rectifier to the 50 Ω output of the antenna, the total load impedance for different input power and frequencies should be determined. A circuit consisting of a pair of Schottky Barrier Diodes (SBD) terminated with a load resistor ($R_{Load} = 11$ kΩ) and an output bypass capacitor ($C2 = 6.8$ pF) was simulated using Agilent ADS software. Figure 8 shows the proposed geometry of the voltage-doubler topology[31,37]. The voltage doubler rectifier structure is employed for the design of the RF-DC power conversion system as this topology is well suited to low power rectification. The resistor and capacitor at the output will filter high frequencies. The high load resistor (11 kΩ) was chosen to observe a reasonable output voltage at very low currents. Using Large Signal S-Parameters analysis in Agilent ADS software, the load impedance and bypass capacitor were determined and optimized.

The voltage doubler rectifier in Fig. 8 consists of a peak rectifier formed by *D2* and bypass capacitor *C2* (6.8 pF) and a voltage clamp formed by *D1* and *C1* (total capacitance of the transmission lines and diode's parasitic capacitance (C_p)). In the negative phase of the input, current flows through *D1* while *D2* is cutoff. The voltage across *D1* stays constant around its threshold voltage and the voltage at node 1 is charged to $-V_{th1}$ (where $-V_{th1}$ is the threshold voltage of the *D1*). At the negative peak, the voltage across *C1* is $V_{amp} - V_{th1}$ (where V_{amp} is the amplitude of the input signal). In the positive phase of the input, current flows through *D2* while *D1* is cutoff. The voltage across *C1* remains the same as the previous phase because it has no way to discharge. At the positive peak, the voltage across *D1* is $2V_{amp} - V_{th1}$. Since *D2* is conducting current to charge *C2*, the voltage at the output is $V_{out} = 2V_{amp} - V_{th1} - V_{th2}$.

The DC equivalent circuit of the SBD is a voltage source in series with the junction resistor R_j which is obtained by differentiating the diode voltage–current characteristic and is given by equation (2)[31,38]:

$$R_j = \frac{nKT}{q(I_s + I_b)} \tag{2}$$

Where n is the diode ideality factor, *K* is the Boltzmann's constant, *T* is the temperature in degrees Kelvin, *q* is the electronic charge, I_s is the diode saturation current and I_b is the external bias current. At low power levels, the saturation current is very small ($I_s = 2.2\ e^{-8}$ A) and for a zero-biased diode, $I_b = 0$. Therefore, the resulting value of junction resistance at room temperature is approximately 1.7 MΩ. Since, the saturation current is highly temperature dependent, R_j will be even higher at lower temperatures which tends to decrease the output voltage. As the input power increases, some circulating rectified current will cause a drop in the value of R_j and this phenomenon will increase the value of the DC output voltage. Furthermore, it is worth to highlight that the rectified current produced by the first diode (*D1*) in Fig. 8 constitutes the external bias current of the second diode (*D2*) which will help to reduce the R_j and hence the detection sensitivity is improved. Therefore, depending on the amount of available bias current, R_j is varying (equation (2)), hence the matching network is changing which impacts the amount of delivered power to the diode and results in different values of PCE.

A Schottky barrier diode can be modeled by the linear equivalent circuit shown in Fig. 9, where L_p and C_p are the diode's parasitic inductance and capacitance respectively due to packaging ($L_p = 2$ nH and $C_p = 0.08$ pF) which are generally unwanted[39]. This linear model is used for determining the diode impedance at a given input power.

The diode impedance analyzed using a Harmonic Balance simulator and a non-linear model of the diodes over the frequency range of 400 to 900 MHz at various input power levels (Fig. 10). Due to the large junction resistor at low input RF power levels, the rectification device is turned off in absence of an appropriate matching network. Large Signal S-parameter analysis was conducted and higher input power (associated with the signal source) is applied directly to the Schottky diodes configuration of Fig. 8 which does not include a matching network in order to turn on the diodes (reduce the value of R_j) and extract approximate input impedance value as our starting point in design of a matching network. As it can be seen in Fig. 10, with increasing the source power, the diode impedance is varying and it is beginning to switch on. Hence, the input impedance needs to be determined when the diode is

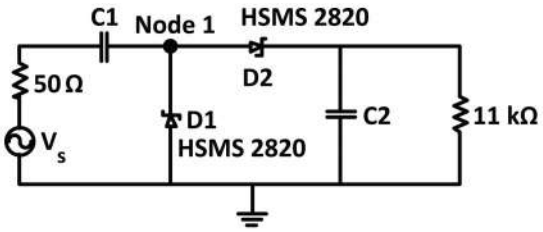

Figure 8 | **Schematic of a voltage-double rectifier without matching network.**

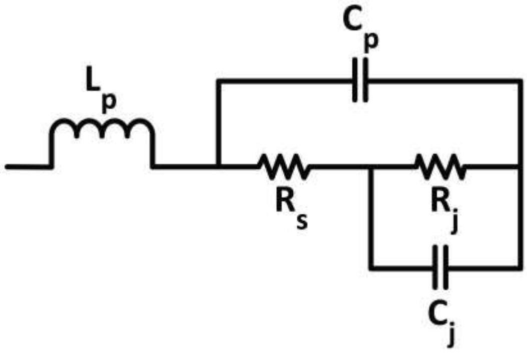

Figure 9 | **HSMS 2820 Schottky diode equivalent circuit.**

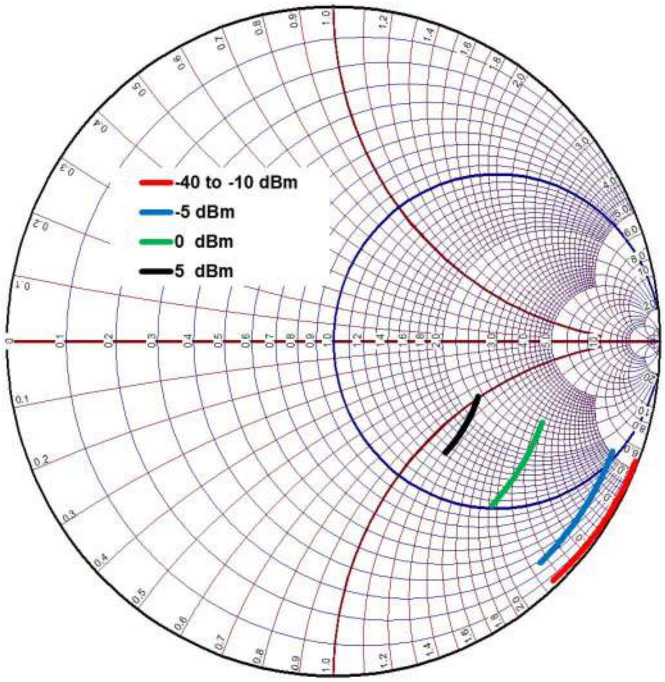

Figure 10 | Diode input impedance calculated with Large Signal S-parameter analysis over the frequency range of 400 to 900 MHz with various unmatched input power levels.

turned on to realize the matching network for a rectifier circuit. Obviously, in the presence of an appropriate matching network the rectification device can be turned on at lower power levels, whilst in the absence of a matching network a higher input power should be applied to switch on the diode. (Note that, with an unmatched rectifier the total applied input power from the signal source cannot be delivered to the diode due to the high reflection in the circuit).

The aim is to match the input impedance of the device to 50 Ω at 478–496 MHz and 852–869 MHz bands over a broad range of input RF powers. The procedure commences by matching the diode input impedance at high unmatched source power and shifting the diode input impedance at various power levels to within the voltage standing wave ratio (VSWR) <2 circle on the Smith chart. This procedure assumes that diode input impedance does not drastically change in this low power range. The simulation results of Fig. 10 prove that this is the case.

In order to provide maximum power transfer from the antenna to the rectifier circuit, a dual resonant rectifier network is designed as a transition between a 50 Ω nominal antenna output and the non-linear rectification device over the power range of −40 to −10 dBm (see Fig. 11). Hence, a coupled-resonator structure with both series and shunt resonators is designed to achieve a dual-band network[40]. The linear equivalent circuit model of the SBD chip[39] has been taken into consideration to design the dual band match at the desired frequency bands. In Fig. 11, $C_{equivalent}$ represents the total capacitance of the diodes and bypass capacitor and $L_{equivalent}$ is the overall parasitic inductance of the diodes. The series *L-C* resonator ($L4 + L_{equivalent}$ and $C_{equivalent}$) and the parallel L-C resonator (*C3* and *L3*) define the dual resonant circuit. The series resonator corresponds closely to the higher band specification of 852–869 MHz, whilst the parallel resonator approximates the lower 478–496 MHz band. A minimum number of components were used in order to reduce the ohmic and parasitic losses.

The resonant frequency of each sub-circuit was determined in isolation using the following equation:

$$f = \frac{1}{2\pi\sqrt{LC}} \tag{3}$$

The 852–869 MHz band resonator circuit components were calculated. Here, $C = C_{equivalent} \cong 1.3$ pF consists of the combination of the bypass capacitor (6.8 pF) and

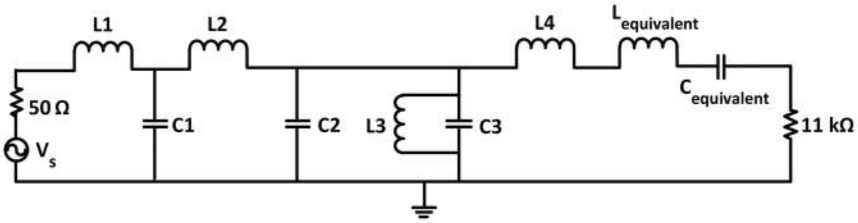

Figure 11 | Schematic of a dual resonant rectifier (optimized parameters of the chip components are: $L1 = 3.9$ nH, $C1 = 0.2$ pF, $L2 = 12$ nH, $C2 = 1.8$ pF, $L3' = 3.9$ nH, $C3' = 7.5$ pF, $L4' = 11.6$ nH, $L_{equivalent} \cong 1$ nH, $C_{equivalent} \cong 1.3$ pF).

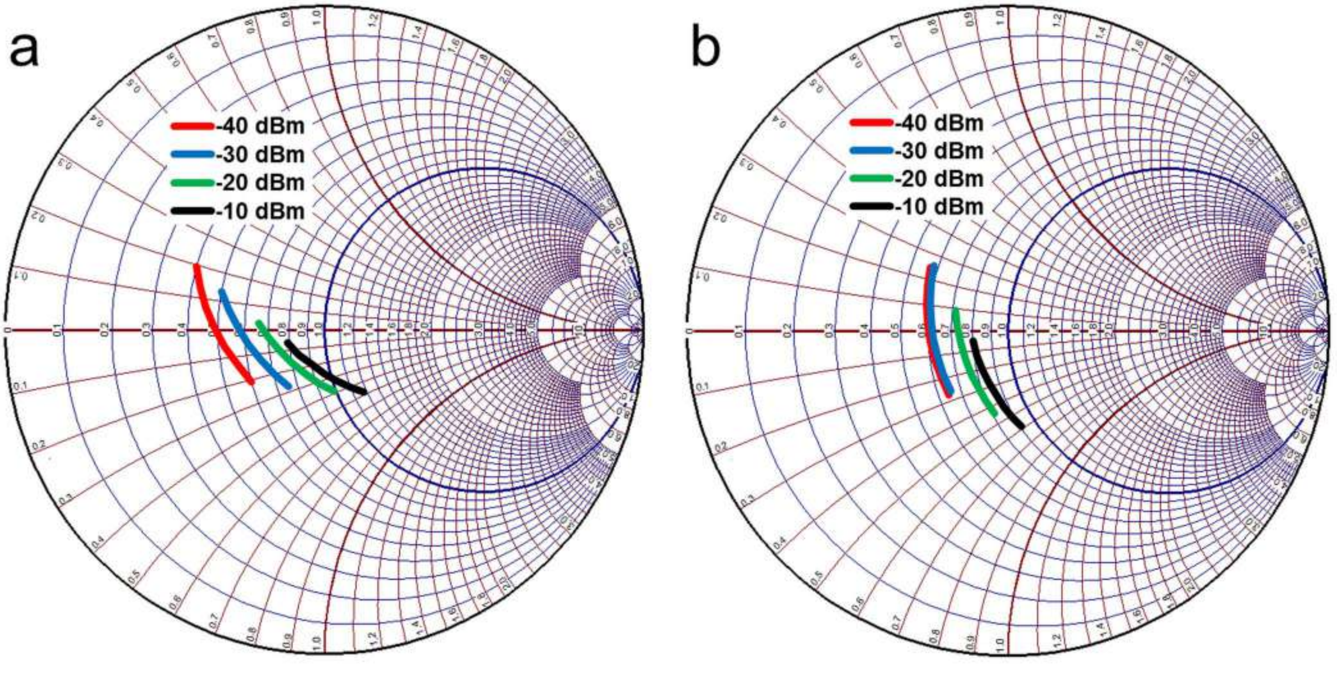

Figure 12 | Dual resonant impedance matching with −40 to −10 dBm input RF power. (a) 478–496 MHz. (b) 852–869 MHz.

the overall junction ($C_{j0} = 0.7$ pF) and parasitic capacitance ($C_p = 0.08$ pF) of *D1* and *D2*. Thus, *L* is calculated to be 26.5 nH in order to achieve an appropriate resonant frequency. Note that, *L* consists of L_4 and the overall parasitic inductance ($L_{equivalent} \cong 1$ nH) of *D1* and *D2*. The 478–496 MHz band resonator circuit components were calculated as $C3 \cong 15$ pF and $L3 \cong 7.2$ nH. Hence, the initial component values are determined for the two resonant circuits.

Initially these resonators were combined to achieve a dual-band structure. Then standard *LC* matching technique[40] is utilized to determine *C1*, *C2*, *L1*, and *L2* to achieve minimum reflection at the resonant frequencies. The substitution of realistic chip component values with their associated parasitics, and addition of 50 Ω microstrip lines and T-junctions introduce delay and shift the imaginary part of the input impedance. The via-holes also contribute to extra inductance in the circuit. Hence minor circuit adjustments are made in order to fine tune the resonant frequencies to the desired values. The final optimized values of the standard chip components are: $L3' = 3.9$ nH, $C3' = 7.5$ pF and $L4' = 11.6$ nH. Large Signal S-parameter analysis is also performed to demonstrate the matching network performance as the input power is varied. Simulation results for the input impedance of the circuit depicted in Fig. 11 are illustrated in Fig. 12. The proposed dual-resonant matching circuit achieves a VSWR <2 at 478–496 MHz and 852–869 MHz for input power ranging from -40 to -10 dBm. It should be noted that the matching circuit was designed based on the input impedance of two diodes and the output resistor and capacitor (Fig. 10). Therefore, selecting a different value for the load resistor requires a new matching circuit to be designed.

1. Paradiso, J. A. & Starner, T. Energy scavenging for mobile and wireless electronics. *IEEE Pervasive Comput.* **4**, 18–27 (2005).
2. Singh, R., Gupta, N. & Poole, K. F. Global green energy conversion revolution in 21st century through solid state devices. 26th International Conference on Microelectronics, Nis, Serbia. *IEEE.* DOI:10.1109/ICMEL.2008.4559221. 45–54 (11–14 May, 2008).
3. Wang, Y. K., Sheng, G. P., Shi, B. J., Li, W. W. & Yu, H. Q. A Novel electrochemical membrane bioreactor as a potential net energy producer for sustainable wastewater treatment. *Sci. Rep.* **3**, DOI:10.1038/srep01864 (2013).
4. Inayat, S. B., Rader, K. R. & Hussain, M. M. Nano-materials enabled thermoelectricity from window glasses. *Sci. Rep.* **2**, DOI:10.1038/srep00841 (2012).
5. Brown, W. C. The history of power transmission by radiowaves. *IEEE Trans. Microw. Theory Tech.* **32**, 1230–1242 (1984).
6. Le, T., Mayaram, K. & Fiez, T. Efficient far-field radio frequency energy harvesting for passively powered sensor networks. *IEEE J. Solid-State Circuits.* **43**, 1287–1302 (2008).
7. Paolo, J. & Gaspar, P. D. Review and future trend of energy harvesting methods for portable medical devices. Proceedings of the World Congress on Engineering (WCE). London, U. K. 909–914 (30 June–2 July, 2010).
8. Matsuki, H., Yamakata, Y., Chubachi, N., Nitta, S. & Hashimoto, H. Transcutaneous DC-DC converter for totally implantable artificial heart using synchronous rectifier. *IEEE Trans. Magn.* **32**, 5118–5120 (1996).
9. Scorcioni, S. *et al.* RF to DC CMOS rectifier with high efficiency over a wide input power range for RFID applications. Microwave Symposium Digest (IMS), *IEEE MTT-S.* Montreal, QC, Cabada. DOI:10.1109/MWSYM.2012.6259760. 1–3 (17–22 June, 2012).
10. Shameli, A., Safarian, A., Rofougaran, A., Rofougaran, M. & De Flaviis, F. Power harvester design for passive UHF RFID tag using a voltage boosting technique. *IEEE Trans. Microw. Theory Tech.* **55**, 1089–1097 (2007).
11. Mandal, S. & Sarpeshkar, R. Low-power CMOS rectifier design for RFID applications. *IEEE Trans. Circuits Syst. I, Reg. Papers.* **54**, 1177–1188 (2007).
12. Sun, H., Guo, Y. X., He, M. & Zhong, Z. Design of a high-efficiency 2.45-GHz rectenna for low-input-power energy harvesting. *IEEE Antennas Wireless Propag. Lett.* **11**, 929–932 (2012).
13. McSpadden, J. O., Fan, L. & Chang, K. Design and experiments of a high-conversion-efficiency 5.8-GHz rectenna. *IEEE Trans. Microw. Theory Tech.* **46**, 2053–2060 (1998).
14. Din, N. M., Chakrabarty, C. K., Bin Ismail, A., Devi, K. K. A. & Chen, W. Y. Design of RF energy harvesting system for energizing low power devices. *PIER.* **132**, 49–69 (2012).
15. Barnett, R. E., Liu, J. & Lazar, S. A RF to DC voltage conversion model for multi-stage rectifiers in UHF RFID transponders. *IEEE J. Solid-State Circuits.* **44**, 354–370 (2009).
16. Kanaya, H. *et al.* Energy harvesting circuit on a one-sided directional flexible antenna. *IEEE Microw. Wireless Compon. Lett.* **23**, 164–166 (2013).
17. Buonanno, A., D'Urso, M. & Pavone, D. An ultra wide-band system for RF energy harvesting. 5th European Conference on Antennas and Propagation (EUCAP). Rome, Italy. *IEEE.* 388–389 (11–15 April, 2011).
18. Hagerty, J. A., Helmbrecht, F. B., McCalpin, W. H., Zane, R. & Popovic, Z. B. Recycling ambient microwave energy with broad-band rectenna arrays. *IEEE Trans. Microw. Theory Tech.* **52**, 1014–1024 (2004).
19. Shariati, N., Rowe, W. S. T. & Ghorbani, K. RF field investigation and maximum available power analysis for enhanced RF energy scavenging. 42nd European Microwave Conference (EuMC). Amsterdam RAI, The Netherlands. *IEEE.* 329–332 (29 October–1 November, 2012).
20. Kim, P., Chaudhary, G. & Jeong, Y. A dual-band energy harvesting using frequency limited dual-band impedance matching. *PIER.* **141**, 443–461 (2013).
21. Li, B. *et al.* An antenna co-design dual band RF energy harvester. *IEEE Trans. Circuits Syst. I, Reg. Papers.* **60**, 3256–3266 (2013).
22. Niotaki, K. *et al.* A compact dual-band rectenna using slot-loaded dual band folded dipole antenna. *IEEE Antennas Wireless Propag. Lett.* **12**, 1634–1637 (2013).
23. Suh, Y. H. & Chang, K. A high-efficiency dual-frequency rectenna for 2.45- and 5.8-GHz wireless power transmission. *IEEE Trans. Microw. Theory Tech.* **50**, 1784–1789 (2002).
24. Collado, A. & Georgiadis, A. Conformal hybrid solar and electromagnetic (EM) energy harvesting rectenna. *IEEE Trans. Circuits Syst. I, Reg. Papers.* **60**, 2225–2234 (20013).
25. Rizzoli, V., Bichicchi, G., Costanzo, A., Donzelli, F. & Masotti, D. CAD of multi-resonator rectenna for micro-power generation. European Microwave Integrated Circuits Conference (EuMIC). Rome, Italy. *IEEE.* 331–334 (228–29 September, 2009).
26. Huang, F. J. *et al.* Rectenna application of miniaturized implantable antenna design for triple-band biotelemetry communication. *IEEE Trans. Antennas Propag.* **59**, 2646–2653 (2011).
27. Pham, B. L. & Pham, A. V. Triple bands antenna and high efficiency rectifier design for RF energy harvesting at 900, 1900 and 2400 MHz. Microwave Symposium Digest (IMS), *IEEE MTT-S.* Seattle, WA. DOI:10.1109/MWSYM.2013.6697364. 1–3 (2013).
28. Pinuela, M., Mitcheson, P. D. & Lucyszyn, S. Ambient RF energy harvesting in urban and semi-urban environments. *IEEE Trans. Microw. Theory Tech.* **61**, 2715–2726 (2013).
29. Mikeka, C. & Arai, H. Microwave Tooth for sensor power supply in battery-free applications. Proceedings of the Asia-Pacific Microwave Conference (APMC). Melbourne, VIC. *IEEE.* 1802–1805 (5–8 December, 2011).
30. Baum, T., Thompson, L. & Ghorbani, K. Complex dielectric measurements of forest fire ash at X-band frequencies. *IEEE Geosci. Remote Sens. lett.* **8**, 859–863 (2011).
31. Avago Technologies, *Surface Mount RF Schottky Barrier Diodes HSMS-282X Series. Data sheet.* (2005). Available at: http://www.avagotech.com/docs/AV02-1320EN. (Accessed: 22/11/2014).
32. Avago Technologies, *Surface Mount Zero Bias Schottky Detector Diodes HSMS-285X Series. Data sheet.* (2005). Available at: http://www.avagotech.com/docs/AV02-1377EN. (Accessed: 22/11/2014).
33. Valenta, C. R. & Durgin, G. D. Rectenna performance under power-optimized waveform excitation. IEEE International Conference on RFID. Penang. DOI:10.1109/RFID.2013.6548160. 237–244 (30 April–2 May, 2013).
34. Trotter, M. S. & Durgin, G. D. Survey of range improvement of commercial RFID tags with power optimized waveforms. IEEE International Conference on RFID. Orlando. DOI:10.1109/RFID.2010.5467265. 195–202 (14–16 April, 2010).
35. Australian Government, *Australia Radio Frequency Spectrum Plan (ACMA).* (January 2013). Available at: http://acma.gov.au/~/media/Spectrum%20Transformation%20and%20Government/Information/pdf/Australian%20Radiofrequency%20Spectrum%20Plan%202013.pdf. (Accessed: 22/09/2014).
36. Keyrouz, S., Visser, H. J. & Tijhuis, A. G. Rectifier analysis for radio frequency energy harvesting and power transport. 42nd European Microwave Conference (EuMC). Amsterdam RAI, The Netherlands. *IEEE.* 428–431 (29 October–1 November, 2012).
37. Avago Technologies, *Schottky Diode Voltage Doubler. Application Note 956-4.* (2005). Available at: http://c1233384.r84.cf3.rackcdn.com/UK_HPA_HSCH-5331_7AN.pdf. (Accessed: 22/11/2014).
38. Eriksson, H. & Waugh, R. W. *A temperature compensated linear diode detector. Design Tip*. Agilent Technologies. (2000). Available at: http://electronix.ru/forum/index.php?act=Attach&type=post&id=13726. (Accessed: 22/11/2014).
39. Avago Technologies, *The Zero Bias Schottky Diode Detector at Temperature Extremes-Problems and Solutions. Application Note AN 1090.* (2005). Available at: http://www.efo.ru/download/Zero_bias_shottky_over_temperature.pdf. (Accessed: 22/11/2014).
40. Mattaei, G., Young, L. & Jones, E. M. T. *Microwave Filters, Impedance-Matching Networks and Coupling Structures.* Artech house books, Dedham, MA. (1980).

Acknowledgments

The authors would like to acknowledge Dr. Khashayar Khoshmanesh for valuable advice. The authors also acknowledge Mr. David Welch for assistance in fabrication of the rectifier circuit.

Author contributions

N.S. designed and simulated the rectifier, carried out the measurements, interpreted results and wrote the paper. K.G. directed the research, contributed to perform the simulation and measurement and validation of design and results. W.S.T.R. supervized the research, analyzed the data and contributed to the general concept, validation of design and results. J.R.S. analyzed the data and contributed to the general concept, validation of design and results. All authors reviewed the manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Shariati, N., Rowe, W.S.T., Scott, J.R. & Ghorbani, K. Multi-Service Highly Sensitive Rectifier for Enhanced RF Energy Scavenging. *Sci. Rep.* **5**, 9655; DOI:10.1038/srep09655 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Sensors **2012**, *12*, 13636-13663; doi:10.3390/s121013636

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Optimization of Passive Low Power Wireless Electromagnetic Energy Harvesters

Antwi Nimo *, Dario Grgić and Leonhard M. Reindl

Department of Microsystems Engineering, Laboratory for Electrical Instrumentation,
University of Freiburg, IMTEK, Georges-Köhler-Allee 106, 79110 Freiburg, Germany;
E-Mails: grgic@imtek.de (D.G.); leonhard.reindl@imtek.uni-freiburg.de (L.M.R.)

* Author to whom correspondence should be addressed; E-Mail: antwi.nimo@imtek.de;
Tel.: +49-761-203-7223; Fax: +49-761-203-7222.

Received: 13 August 2012; in revised form: 28 September 2012 / Accepted: 28 September 2012 / Published: 11 October 2012

Abstract: This work presents the optimization of antenna captured low power radio frequency (RF) to direct current (DC) power converters using Schottky diodes for powering remote wireless sensors. Linearized models using scattering parameters show that an antenna and a matched diode rectifier can be described as a form of coupled resonator with different individual resonator properties. The analytical models show that the maximum voltage gain of the coupled resonators is mainly related to the antenna, diode and load (*remote sensor*) resistances at matched conditions or resonance. The analytical models were verified with experimental results. Different passive wireless RF power harvesters offering high selectivity, broadband response and high voltage sensitivity are presented. Measured results show that with an optimal resistance of antenna and diode, it is possible to achieve high RF to DC voltage sensitivity of 0.5 V and efficiency of 20% at −30 dBm antenna input power. Additionally, a wireless harvester (*rectenna*) is built and tested for receiving range performance.

Keywords: RF energy harvesting; wireless power transmission; coupled resonators; Schottky diode; RF to DC power converter; impedance matching; PI-matching; L-matching; rectenna

---

1. Introduction

For autonomous powering of sensor nodes in remote or inaccessible areas, wireless power transfer provides the only viable option to power them from an energy source. Due to the low power density of ambient RF at far-field from transmitters, there is a need to optimize each aspect of a wireless RF energy harvester for possible realistic applications. Today remote autonomous sensors are mostly powered by batteries, which have limited lifespan. Renewable powering has the potential to power autonomous sensors perpetually. Due to the expansion of telecommunications technology ambient electromagnetic (EM) power is among the most common sources of ambient energy. There are power transmitters/receivers scattered in practically any society, ranging from television transmission stations to cell phone transmitters and even wireless routers in our homes/offices or mobile phones. These transmitters in our environment and others which are on special dedicated frequencies produce ambient RF power (on the order of microwatts) which can be used as a source for powering remote microwatt budget sensors through wireless energy harvesting. This work presents different matching techniques based on different application requirements using Schottky diode-based RF to DC power converting circuits for wireless remote EM energy harvesting around 434 MHz and 13.6 MHz. Generalized analytical models and limitations of the matched RF to DC power converters are discussed. A wireless RF energy harvester consisting of an antenna and a matched diode rectifier is then realized and its performance tested. Passive wireless energy harvesting also finds applications in near field communications (NFC) [1], RFID tags [2–5], implantable electronics [6,7], and environmental monitoring [8], among others.

1.1. State of the Art

Hertz was the first to demonstrate the propagation of EM waves in free space and to demonstrate other properties of EM waves such as reflection using parabolic reflectors [9]. Wireless power transmission was then investigated and demonstrated for possible wireless remote powering by Tesla. Electromagnetic power beaming for far field wireless power transfer using collimated EM waves was proposed in the 1950s [9]. Recent advances in ultralow power sensors means ambient omni-directional EM power can be used as a source for powering remote sensors without the need to collimate the EM power through the wireless space. Mickle [10] and McSpadden [11] have presented earlier work on wireless energy harvesting systems using Schottky diodes and rectennas where the usability of ambient RF power into DC power was shown. Sample [12] presented a wireless harvester which can harvest EM power from TV and radio base stations transmitting 960 kW of effective radiated power; 60 μW was harvested at a range of about 4 km. Umeda [13] and Le [14] have presented more integrated wireless energy harvesters based on CMOS RF to DC rectifying circuits. CMOS-based rectifying power converters provide full compatibility with standard CMOS technologies and have advantages in batch processes for mass production. The drawback of CMOS-based diode connected transistors is the need to bias the gate of the transistors for the rectifying circuits to effectively function. This gate bias is provided externally, which makes the system not passive. Without the injection of external charges or a biasing of the transistor gate, the circuit has low efficiency, especially when the amplitude of the input voltage is low [15]. Shameli [2] presented a passive CMOS RF to DC power converter with a voltage sensitivity of 1 V at −14.1 dBm input, but the circuit efficiency was only 5 %. Zbitou [16] presented an RF to DC converter based on Schottky diodes and achieved 68 % efficiency at 20 dBm RF input power. Ungan [17,18] presented antennas and high quality factor RF to DC power converters at 24 MHz and 300 MHz for RF wireless energy harvesting at -30 dBm input power. The power converter used high quality factor resonators for impedance matching the EM source and the diodes and achieved high open circuit voltage sensitivity of 1 V/μW. Boquete [19] presented a risk assessment system for calculating insurance premiums by monitoring mobile phone usage while driving. This was done by harvesting EM power from detected mobile phone usage during driving for risk assessment. Heikkinen [20] presented rectennas on different substrates at 2.4 GHz using transmisson lines to match the antennas output resistance (*at resonance*) to the rectifying diodes. Akkermans [21] presented a rectenna design by complex conjugating impedance provided by a microstrip structure to a diode so that resonance may be achieved for a working frequency. This design approach may need sophisticated tools to realize and the dominant resonance frequency of the rectenna can be unpredictable in practice. Hagerty [22] presented rectenna arrays for broadband ambient EM harvesting and characterized the harvesters from 2 GHz to 18 GHz; rectennas combine impedance matching the RF rectifying circuit and the antenna into one compact device, but an array of rectennas may increase the overall size of an EM harvester. Herb [23] and Vullers [24] have provided a comprehensive state of the art for micro energy harvesting and have explored the various techniques used for harvesting ambient renewable energy.

2. RF to DC Power Converter

2.1. Diode Rectifier

A junction diode equivalent circuit and simple Schottky diode rectifier are shown in Figure 1. R_{DS} is the diode resultant series resistance, C_{DS} is the diode resultant series capacitance, R_{DP} is the diode resultant parallel resistance, C_{DP} is the diode resultant parallel capacitance, Vs is the sinusoidal source voltage and Vc is the voltage across the capacitor.

Figure 1. (**a**) Diode series equivalent model, (**b**) Diode parallel equivalent model, (**c**) Simple diode detector.

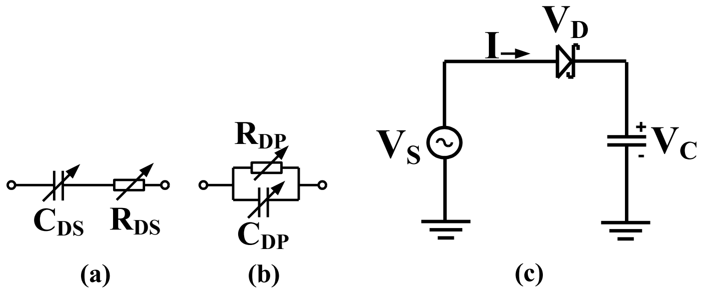

The diode capacitive impedance is mainly due to the junction capacitances provided by the metal, its passivation and the semiconductor forming the diode. AC power incident on a forward biased diode input is converted to DC power at the output. The current-voltage behavior of a single metal/semiconductor diode is described by the Richardson equation [25] as in Equation (1):

$$I = I_S\left(e^{\left(qV_D/nKT\right)} - 1\right) \tag{1}$$

where I is the current through the diode, I_S is the saturation current, q is the charge of an electron, V_D is the voltage across the diode, T is the temperature in degrees Kelvin and K is Boltzmann constant. The voltage equation around the loop can be derived from Figure 1(c) and is given in Equation (2):

$$V_D = V_S - V_C \tag{2}$$

Since the same current flows through the diode and the capacitor, one can find the average current through the circuit by integrating Equation (1) over a time period. By substituting Equation (2) into Equation (1), V_C can be expressed in terms of V_S by averaging the diode current to zero. This is given in Equation (3) [26]:

$$V_C = \frac{KT}{q}\ln\left[\vartheta_0\left(\frac{qV_S}{KT}\right)\right], \tag{3}$$

where ϑ_0 is the series expansion of the sinusoidal source voltage. Equation (3) can further be simplified for very small amplitude V_S as Equation (4):

$$V_C \approx \frac{qV_S^2}{4KT} \tag{4}$$

Equation (4) shows that for a small voltage source, the circuit output voltage is proportional to the square of the input sinusoidal voltage; hence it's so-called square law operation. Extensions of this model for voltage multipliers and other input signals are presented in [27] and [28]. Equation (4) further confirms that for low input voltage (power ≤ 10 dBm), an impedance matching network between the source and the diode is necessary to improve the detected output voltage and efficiency.

2.2. Impedance Matching

The maximum power transfer theorem states that the highest power is transferred to the load when the source resistance is the same as the load resistance. For systems with both resistive and reactive impedances from source and load, the source and the load impedance should be adjusted in a way that they are the complex conjugate of each other through impedance matching. For the purposes of this work, a 50 Ω resistive source is chosen as reference for load impedance matching. The antenna which captures the ambient RF signals is tuned to provide this source resistance at resonance for the rectifying circuit in a complete EM wireless remote harvester. The load is the resistance of the Schottky diodes and the actual connected resistance (*remote sensor*). The specific type of matching network which can be used for complex conjugation depends on the nature of load or source impedance, the desired RF to DC converter functionality and other factors like circuit size, cost, *etc.* The response of a matched RF to DC power converter depends on the matching network used as well as the source or load component quality factors and impedances.

2.3. Diode Impedance

Schottky diodes HSMS-285C and HSMS-286C from Avago [29,30] are used to build the RF to DC power converters. The HSMS-285× or 286× series diodes can be operated as zero biased with relatively low forward junction potential. This allows for the realization of completely passive RF to DC power converters for wireless energy harvesting. The HSMS-285C or 286C is a pair of series connected Schottky diodes in a SOT-323 package. The impedance of the HSMS-285C and HSMS-286C diodes was first measured so it can be matched to the resistance (50 Ω) of the antenna source. This is done by connecting the input of the diodes to a network analyzer and measuring the scattering parameters. These scattering parameters are then converted to the corresponding impedances. The input impedance of a diode depends mainly on the resistive and capacitive impedance provided by the junction of the diode and its connected load. For a couple of diodes arranged in a package such as the HSMS-285C or 286C, the input impedance is the vector sum of the impedances provided by each diode in the package arrangement, the extra impedance associated with the packaging and the connected load. The diode measuring board is as shown in Figure 2. The diodes were measured at room temperature for an input power of -30 dBm at a diode connected load of 1 MΩ with a 100 pF filter capacitor. For the sake of this work, the input impedance of the diodes will always be referred to at these connected load conditions.

Figure 2. (**left**) Reference circuit layout for measuring diodes input impedance, (**right**) measuring printed circuit board (PCB) for diodes input impedance on 1 mm FR4 substrate.

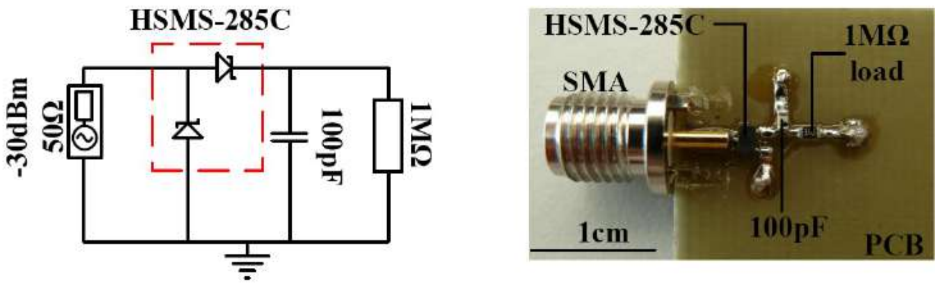

Figure 3. Measured input impedance (Δ resistive, □ capacitive) of HSMS-285C (**left**) and HSMS-286C (**right**) diodes at −30 dBm input with 1 MΩ load and 100 pF filter.

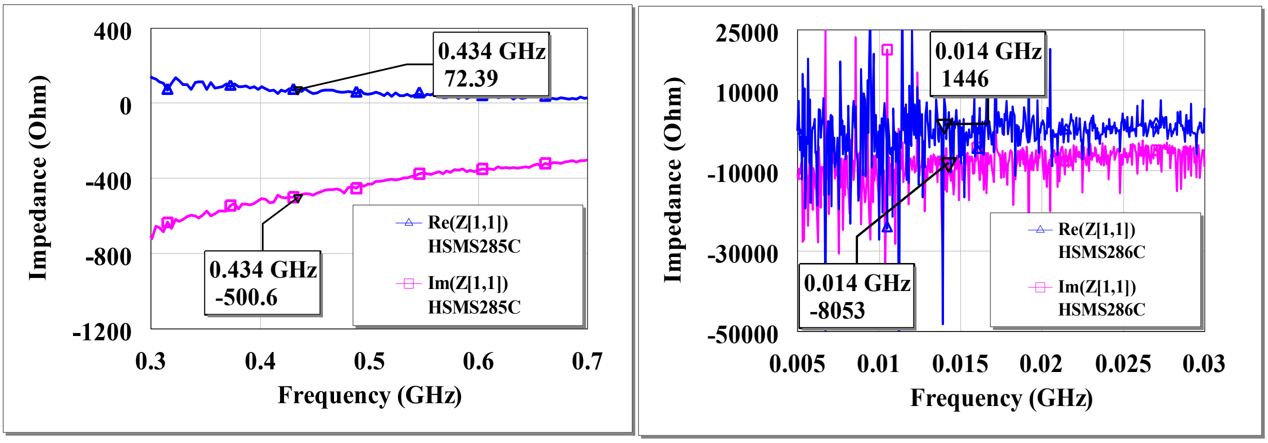

The board is fabricated such that components are soldered directly one into another to prevent additional impedances introduced by copper route. The PCB backside had the ground layer. An example of measured input impedance for HSMS-285C and HSMS-286C is shown in Figure 3.

The diodes quality factor is given by $X_{DS}R_{DS}^{-1}$, where X_{DS} is the resultant series capacitive impedance of the diodes. At an input power of −30 dBm, the measured input impedance of the HSMS-285C diodes is 72–j501 Ω at 434 MHz and 587–j1239 Ω at 13.6 MHz. For HSMS-286C diodes, it is 10–j503 Ω at 434 MHz and ~1.5–j8.1 kΩ at 13.6 MHz for −30 dBm input. The measured impedance of the HSMS-286C diodes at low frequencies (< 60 MHz) shows pronounced fluctuations. The low-frequency excess flicker noise and the shot noise observed in the HSMS-286C have been studied by several authors [31–33]. The pronounced presence of trap states in the depletion region of the semiconductor, mobility fluctuations in carriers, edge effects among other reasons is reported to cause deviations from the ideal Schottky diode behavior and hence generation-recombination noise for some diodes such as the HSMS-286C [34]. When a diode rectifier is matched at a reference operating condition, the matching network may function less effectively at other input power levels, connected load and other operating frequencies. This is due to possible changes in the diode input impedance. Throughout this work the imperfections of the matching circuit at other operating conditions away from the matched reference conditions are accepted without changes to the matching network.

2.4. Voltage Doubler

The Delon voltage doubler and Greinacher doubler are both used to realize the RF to DC power converters presented in this work. The Delon voltage doubler and Greinacher doubler are shown in Figure 4. The diodes output voltage (V_{out}) is doubled what is detected by a simple detector circuit shown in Figure 1. Both doublers produce the same output performance, the only difference is that the Delon doubler has an instantaneous input ground which is not shared with the output.

Figure 4. Circuit diagram of voltage doubler, (**a**) Delon doubler and (**b**) Greinacher doubler.

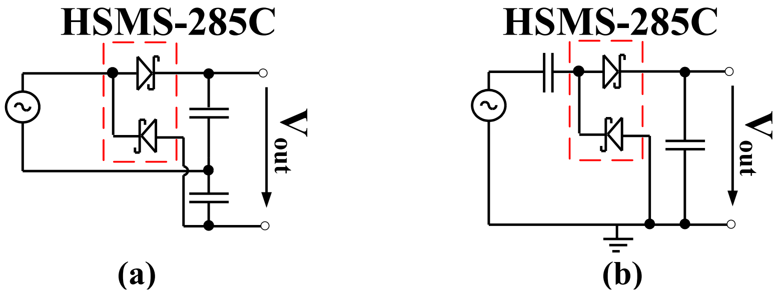

(a) **(b)**

2.5. Matching Techniques for Antenna Source and RF to DC Power Converter

2.5.1. L-match RF to DC Power Converter

An L-match network converts a source series impedance to its equivalent load parallel impedance or *vice-versa* and tunes out by subtracting or adding any surplus reactance from the load or source with the counter impedance. Series impedance is converted to its parallel equivalent impedance using Equations (5–7):

$$Q_S = \frac{X_S}{R_S} \tag{5}$$

$$Q_P = \frac{R_P}{X_P} \tag{6}$$

where Xs is the total series reactive impedance, Rs is the total series resistance, R_P is the total parallel resistance, Xp is the total parallel reactive impedance, Q_S and Qp are the series and parallel quality factors respectively:

$$R_S + jX_S = \frac{R_P \times jX_P}{R_P + jX_P} \tag{7}$$

Equation (7) is the equation of a series sum of impedances and a parallel sum of impedances. It is interesting to note that Q_S and Q_P from an L-matched network may be different from the individual component quality factors as a result of the inherent resistive and reactive impedances in that component. By virtue of Equation (7), Q_S and Q_P must be equal in an L-matched network. Using Equations (5,6) and (7), the ratio of the parallel resistance (*or reactance*) to the series resistance (*or reactance*) can be derived in terms of the quality factors Q_P or Q_S [35]. Since at match conditions, only the resistive impedances dissipate power, the loaded quality factor Q, of the L-matched network can be expressed as in Equation (8):

$$R_P = (Q^2 + 1)R_S \tag{8}$$

Using Equations (5,6) and (8), series impedance can be converted to its parallel equivalent for a fixed frequency and power level. As an example; a series impedance 72–j501 Ω (HSMS-285C at 434 MHz for −30 dBm input power) is easily converted to $-j510(3519)/(-j510 + 3519)$ Ω as its parallel equivalent with a component quality factor of 6.96. The source resistance is taken as part of the parallel matching network in an L-match circuit if the source series equivalent resistance is greater than the load series equivalent resistance. On the other hand, the load resistance is taken as part of the parallel matching network if the load series equivalent resistance is greater than the source series equivalent resistance. For the purpose of this work, inductors were only used for series impedance matching and capacitors as shunts. This prevents power seeping through any shunt inductor used for impedance matching due the short circuit provided by a shunt inductor to ground and resulting in less output efficiency. Resistors were not used for impedance matching.

2.5.2. L-match RF to DC Converter Generalized Analytical Model

The classical matching technique using Equations (5,6) and (8) is first used to L-match the 50 Ω resistance of the antenna to the resistance of the HSMS-286C diodes (and load) at 434 MHz for −30 dBm input and then the generalized model is discussed. The antenna source resistance was L-matched to the resistance of the diodes (and load). The 50 Ω resistance of the antenna is taken as the parallel matching component and the diodes 10 Ω resistance is the series matching component. The loaded Q is found as 2 between the 50 Ω antenna source resistance and the 10 Ω diode series resistance using Equation (8). From this loaded Q, a shunt capacitive impedance of 25 Ω (14.6 pF at 434 MHz) using Equation (6) and a series inductive impedance of 20 Ω (7.3 nH at 434 MHz) using Equation (5) will match the 50 Ω source to the 10 Ω HSMS-286C diodes (and load) series resistance at −30 dBm input. Since the HSMS-286C diodes inherently provides 503 Ω series capacitive impedance at −30 dBm, a resultant series inductive impedance of 523 Ω (192 nH at 434 MHz) is needed to tune the 50 Ω resistive source to the complete HSMS-286C diodes impedance at 434 MHz for −30 dBm input. The L-matched HSMS-286C diodes rectifier is as shown in Figure 5(a).

Figure 5. (**a**) L-match RF to DC harvester using the HSMS-286C diodes at 434 MHz for −30 dBm input. (**b**) Small signal impedance model of a generalized L-matched RF to DC power converter as capacitive coupled series RLC resonator with different resonator elements.

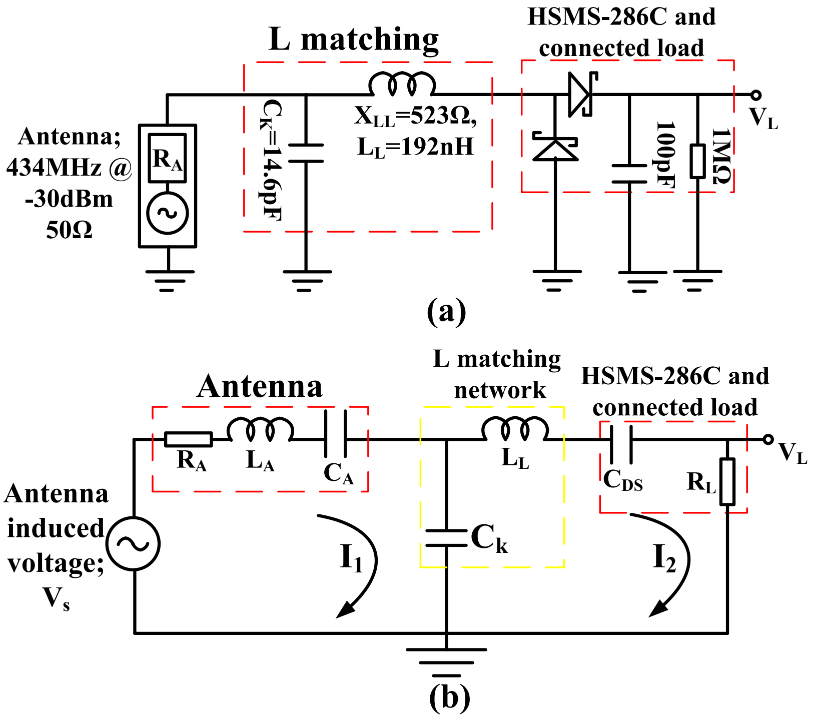
L matching
HSMS-286C and connected load
Antenna; 434MHz @ -30dBm 50Ω
R_A
C_K=14.6pF
X_{LL}=523Ω, L_L=192nH
100pF
1MΩ
V_L
(a)
Antenna
L matching network
HSMS-286C and connected load
R_A
L_A
C_A
L_L
C_{DS}
R_L
V_L
Antenna induced voltage; V_s
I_1
C_k
I_2
(b)

C_K is the tuning capacitance, L_L is the tuning inductance, X_{LL} is the tuning inductive impedance, C_{DS} is the diodes series capacitance, X_{DS} is the diodes series capacitive impedance, V_S is the antenna captured ambient EM voltage, R_A is the resistance of antenna, L_A is the inductance of antenna, C_A is the capacitance of antenna, R_L is the resultant series resistance from the diodes and the connected load resistance, V_L is the resistive load voltage. From Figure 5(a) the power dissipated in the resistance of the diodes (and connected load); P_L is given by Equation (9), where R_L is the series resistance of the diodes and load:

$$P_L = \frac{V_L^{\ 2}}{R_L} \tag{9}$$

The source power; P_S is given by Equation (10), where V_S* is the root mean squared (RMS) antenna captured source voltage. Half of the source power is transferred to the resistance of the diodes (and connected load) at match conditions as described by the maximum power transfer theorem:

$$P_S = \frac{V_S^{\ 2}}{R_A} \quad \text{or} \quad P_S = \frac{V_{S*}^{\ 2}}{2R_A} \tag{10}$$

Equating P_L and half RMS P_S gives a condition of maximum voltage gain for the matched RF to DC power converter shown in Figure 5(a):

$$\frac{V_L}{V_{S*}} = \frac{1}{2}\sqrt{\frac{R_L}{R_A}} \tag{11}$$

From Equation (8), substituting the series and parallel resistance ratio into Equation (11) the voltage gain can be expressed in terms of the loaded quality factor as in Equations (12) and (13), where Q is the loaded quality factor of the RF to DC power converter:

$$\frac{V_L}{V_{S*}} = \frac{1}{2}\sqrt{\frac{1}{1 + Q^2}} \tag{12}$$

Equation (12) is the voltage gain in-terms of the loaded Q if the resistance of the diodes (and connected load) is part of the series matching network and the resistance of the antenna source is part of the parallel matching network as in Figure 5(a). If the resistance of the diodes is part of the parallel matching network, then Equation (13) may be written as the voltage gain in-terms of the loaded Q in an L-matched circuit:

$$\frac{V_L}{V_{S*}} = \frac{1}{2}\sqrt{1 + Q^2} \tag{13}$$

Equations (12) and (13) shows that the maximum voltage gain is directly related to the relative differences between the diodes (and connected load) resistance and source resistance at matched conditions or the circuit loaded quality factor. It is interesting to note that the circuit shown in Figure 5(a) has a loaded Q of 2, but an HSMS-286C unloaded quality factor of 50 (at 434 MHz for −30 dBm).

Figure 5(a) is generally modeled as capacitive coupling of two series RLC resonators with a voltage source. This linearized model can be made at any defined frequency and power level. The model however neglects the metal/semiconductor physics of the diode's junction potentials which results in a Schottky barrier. The first series RLC resonator is modeled as impedance from the antenna with or without some passive matching components. The voltage source V_S, is the antenna captured electromagnetic voltage. The second series RLC resonator is the impedance from the diodes (at a defined condition), connected resistance and some passive matching components. Ck is modeled as the coupling element between the two series RLC resonators. Figure 5(b) gives a more general look at the special scenario shown in Figure 5(a). The voltage equations in the two loops are given by Equations (14,15) according to Kirchhoff's voltage loop laws, where ω is the angular frequency and I_1, I_2 are the currents in the first loop and second loop, respectively:

$$V_S = I_1\left[R_A + j\omega L_A - \frac{j}{\omega C_A} - \frac{j}{\omega C_K}\right] + \frac{jI_2}{\omega C_K} \tag{14}$$

$$0 = \frac{jI_1}{\omega C_K} + I_2\left[R_L + j\omega L_L - \frac{j}{\omega C_{DS}} - \frac{j}{\omega C_K}\right] \tag{15}$$

Using Cramers rule, I_2 can be expressed as:

$$I_2 = \frac{\dfrac{-jV_S}{\omega C_K}}{\left[R_A + j\omega L_A - \dfrac{j}{\omega C_A} - \dfrac{j}{\omega C_K}\right]\left[R_L + j\omega L_L - \dfrac{j}{\omega C_{DS}} - \dfrac{j}{\omega C_K}\right] + \dfrac{1}{\omega^2 C_K^2}}. \tag{16}$$

The voltage across R_L is V_L; given by I_2R_L:

$$V_L = \frac{\dfrac{-jV_S}{\omega C_K} R_L}{\left[R_A + j\omega L_A - \dfrac{j}{\omega C_A} - \dfrac{j}{\omega C_K}\right]\left[R_L + j\omega L_L - \dfrac{j}{\omega C_{DS}} - \dfrac{j}{\omega C_K}\right] + \dfrac{1}{\omega^2 C_K^2}} \tag{17}$$

The voltage gain of the coupled resonator can be expressed as in Equation (18):

$$\frac{V_L}{V_S} = \frac{\dfrac{-jR_L}{\omega C_K}}{\left[R_A + j\omega L_A - \dfrac{j}{\omega C_A} - \dfrac{j}{\omega C_K}\right]\left[R_L + j\omega L_L - \dfrac{j}{\omega C_{DS}} - \dfrac{j}{\omega C_K}\right] + \dfrac{1}{\omega^2 C_K^2}} \tag{18}$$

At resonance, there is no resultant reactance in the RLC resonators or the capacitive and inductive impedances become equal; hence Equation (19) can be written:

$$\omega L_A - \frac{1}{\omega}\left\{\frac{1}{C_A} + \frac{1}{C_K}\right\} = 0 \text{ and } \omega L_L - \frac{1}{\omega}\left\{\frac{1}{C_{DS}} + \frac{1}{C_K}\right\} = 0 \tag{19}$$

Equations in Equation (19) can be used to find the resonant frequencies of the series coupled resonator. The voltage gain of the coupled resonator at resonance can then be expressed as in Equation (20):

$$\frac{V_L}{V_S} = V_{gain} = \frac{\dfrac{-jR_L}{\omega C_K}}{R_A R_L + \dfrac{1}{\omega^2 C_K^2}} \tag{20}$$

where V_{gain} is the voltage gain. V_{gain} at resonance is a function of the resistance of the source and load and the coupling element. The maximum of Equation (20) is obtained when:

$$\frac{dV_{gain}}{dC_K} = 0. \tag{21}$$

This gives the results as in Equation (22):

$$\frac{dV_{gain}}{dC_K} = \frac{j2R_L}{\omega^3 C_K^4} - j\left\{R_A R_L + \frac{1}{\omega^2 C_K^2}\right\}\frac{R_L}{\omega C_K^2} = 0 \text{ or } R_A R_L^2 = \frac{R_L}{\omega^2 C_K^2} \tag{22}$$

Equation (22) can be simplified to find $C_{K(max)}$:

$$C_{K(\max)} = \pm\frac{1}{\omega}\sqrt{\frac{1}{R_A R_L}} \tag{23}$$

where C_{Kmax} is the value of the coupling element where maximum power transfer from the first resonator to the second resonator occurs. Using Equations (19) and (23) the unknown optimal matching impedances can be found from the known impedances just like the classical L-matched procedure using Equations (5,6) and (8). By substituting $C_{K(max)}$ into Equation (20) and taking the magnitude of V_{gain}, gives the maximum voltage gain of the coupled series resonator at resonance:

$$\left|\frac{V_L}{V_S}\right| = \frac{1}{2}\sqrt{\frac{R_L}{R_A}} \text{ or simply } \frac{V_L}{V_{S^*}} = \frac{1}{2}\sqrt{\frac{R_L}{R_A}} \tag{24}$$

For wireless harvesters consisting of an antenna and a diode rectifying circuit, the diode resistive impedance at any condition is dependent on the diode realized parameters, signal frequency, connected load and the input power level. The source impedance is determined by the impedance of the antenna. For maximum efficiency, the ratio of the source resistance to the load resistance must tend to zero at matched conditions. The efficiency η of the circuit is given by Equation (25):

$$\eta = \frac{P_L}{P_S}\,;\, \eta \to 1 \;\; when \; \frac{R_A}{R_L} \to 0 \tag{25}$$

2.5.3. L-Match RF to DC Converter Experimental Results and Discussion

The presented circuit was L-matched between the 50 Ω resistance of the antenna source and the resistance of the HSMS-285C diodes (and load) at 434 MHz for −30 dBm input as shown in Figure 6. Since the series equivalent resistance of the HSMS-285C diodes and load (72 Ω) is greater than the 50 Ω series resistive antenna source, the diode is taken as parallel matching network with a parallel equivalent impedance of $-j510(3519)/(-j510 + 3519)$ Ω. The analysis follows the same procedure as in Section 2.5.2 after this step. Figure 6(b) shows the resultant L-matched RF to DC power converter. C_{DP}* is the resultant shunt matching capacitance.

Figure 6. (**a**) L-matched impedance circuit for matching the HSMS-285C diodes at 434 MHz for −30 dBm input. (**b**) Resultant network, (**c**) PCB layout of the L-matched Delon doubler with adjusted values on FR4 substrate (**d**) Fabricated PCB of the L-network matched Delon voltage doubler.

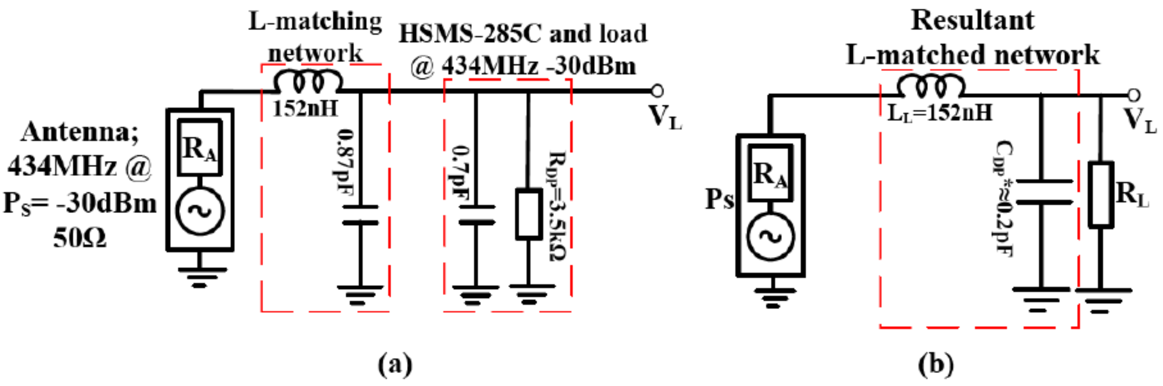
L-matching network
HSMS-285C and load @ 434MHz -30dBm
152nH
0.87pF
0.7pF
R_DP=3.5kΩ
V_L
Antenna; 434MHz @ P_S= -30dBm 50Ω
R_A
(a)
Resultant L-matched network
L_L=152nH
C_DP*≈0.2pF
R_L
Ps
R_A
V_L
(b)

Figure 6. *Cont.*

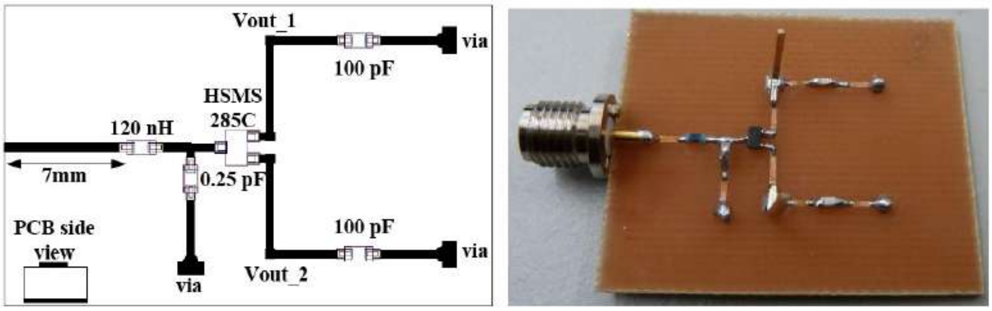

(c) **(d)**

Figure 6(a,b) assume perfect characteristic impedance between the various components in the matched circuit. When a copper route is introduced between components and on a material substrate, it must be accounted for in the total impedance as seen by the source or load. This PCB impedance compensation is carried out in Advance Design Systems (ADS) from Agilent [36]. ADS has extensive models for microstrip substrates to account for its impedances. The optimized layout using ADS microstrip models and its compensated values in the passive tuning components for a Delon doubler is shown in Figure 6(c).

The circuit reflection coefficient (S_{11}) and input impedance at open circuit are shown in Figure 7. There is high return loss and resonance around 434 MHz. The circuit input impedance at open circuit conditions is ~38 Ω at resonance for −40 dBm and ~17 Ω at −10 dBm input.

The measured L-matched circuit efficiency and voltage sensitivity is shown in Figure 8. The maximum measured L-matched efficiency at −30 dBm is 22% at ~20 kΩ load and an open circuit voltage of 124 mV. At −10 dBm, the maximum efficiency and open circuit voltage is 47% and 2 V respectively. At the optimal load of ~20 kΩ, the detected voltage is 58 mV and 1 V at −30 dBm and −10 dBm respectively.

Figure 7. Measured open circuit S_{11} of the L-matched Delon circuit at different input power levels from a 50 Ω source (**left**), measured open circuit input impedance at −10 dBm and −40 dBm of the L-matched circuit (**right**).

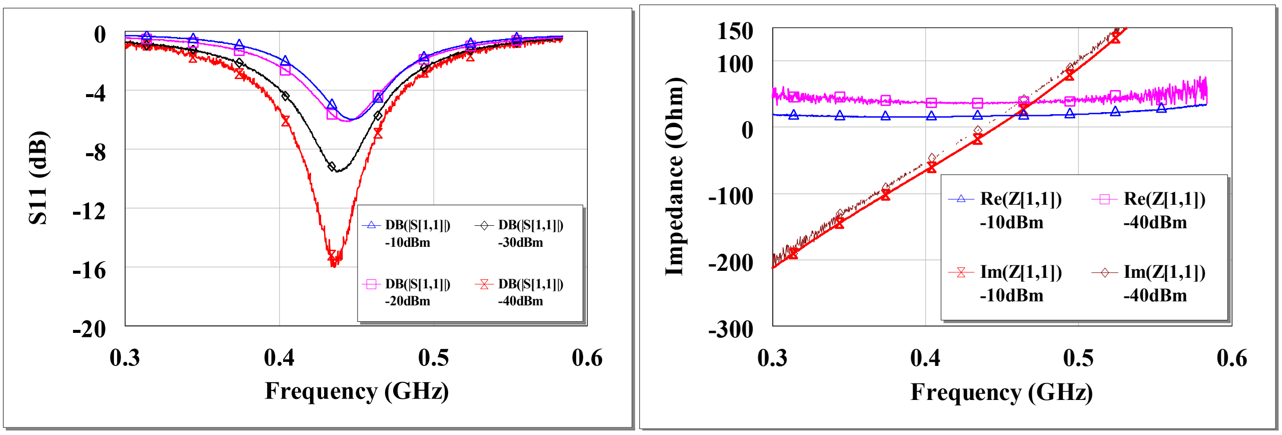

The open circuit voltage gain is 25 at −30 dBm and 40 at −10 dBm. The maximum measured efficiency at −35 dBm is 27%. This is higher than that of −30 dBm due to the better matched circuit impedance at −35 dBm (35 Ω) than at −30 dBm (27 Ω) input. The L-matched RF to DC power converter has a loaded *Q*, sensitivity and efficiency determined mainly by the diodes resistance, diodes junction potential, connected resistance and antenna source resistance at matched conditions.

Figure 8. Measured L-matched circuit efficiency *versus* resistive load at various input power levels at 434 MHz (**left**), measured open circuit voltage and at 17 kΩ load *versus* input power at 434 MHz (**right**).

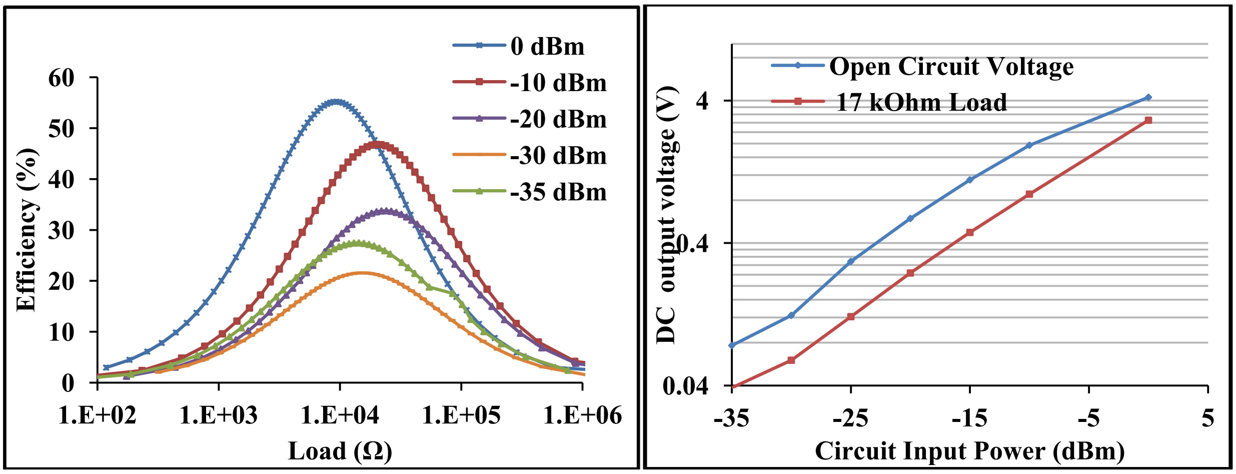

2.5.4. PI-match RF to DC Power Converter

A highly selective or small frequency bandwidth RF power converter is realized with a PI-network in-between the source impedance from the antenna and the diode rectifier. A PI-network is a ‘back to back’ L-network that are both configured to match the load and source impedance to an invisible resistance located at the junction between the two L-networks [37]. The quality factor of the L-network with the parallel resistance is given by Equation (26):

$$Q_P^* = \sqrt{\frac{R_P}{R^*} - 1}, \tag{26}$$

where R_P is the parallel resistance, R^* is a virtual resistance and Q_P* is the quality factor of the L-network with the parallel resistance. The quality factor of the L-network with the series resistance is given by Equation (27):

$$Q_S^* = \sqrt{\frac{R_S}{R^*} - 1}, \tag{27}$$

where Q_S* is the quality factor of the L-network with the series resistance. The unloaded quality factor; Q_S* or Q_P* is set higher than what is normally achieved with a single L-network [37] to realize the small frequency bandwidth circuit. The resistance of the load is assigned the parallel network in a PI-matched circuit if its series equivalent resistance is higher than the source series equivalent resistance; the opposite is true if the source is higher than the load. Equation (26) and Equation (27) are synonymous to Equation (8), except the lowest resistive impedance in Equation (8) is substituted with the virtual resistance which is dependent on the newly desired circuit selectivity. From Equations (26) and (27) the loaded quality factor of the PI-matched circuit can be written as Equation (34) in terms of Q_S* and Q_P*:

$$Q^2 = \left[\left(\frac{Q_P^{*2} + 1}{Q_S^{*2} + 1} \right) - 1 \right], \tag{28}$$

where Q is the loaded quality factor of the PI-network. Q_S* or Q_P* are the unloaded quality factors of the PI-matched network. The larger value among the unloaded quality factors result in small frequency bandwidth response which is desired when matching a source and load impedance with a PI-network. Some authors approximate the highest value of Q_S* or Q_P* or their algebraic sum as the loaded quality factor of the PI-network as in [35] and [37], but Equation (28) gives the exact loaded Q of the PI-matched circuit in terms Q_S* and Q_P*. This allows for the correct estimation of the maximum voltage gain from the loaded quality factor.

2.5.5. Selectivity RF to DC Converter Generalized Analytical Model

An example of a PI-matched RF to DC converter using the HSMS-285C diodes operating at 434 MHz for −30 dBm input is presented first and then the generalized model is discussed. The circuit is matched for Q_P* of 60 between the antenna and the resistance of the diodes as shown in Figure 9.

Figure 9. Impedance diagram of PI-matched RF power converter; (**a**) Impedance diagram of 50 Ω source and the HSMS-285C diodes at 434 MHz, (**b**) Resultant PI matched network between the antenna source and load resistance.

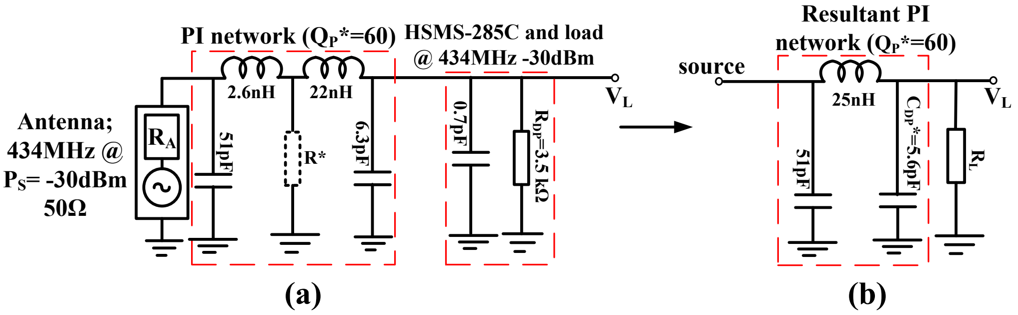
PI network (QP*=60)
HSMS-285C and load
@ 434MHz -30dBm
Antenna;
434MHz @
PS= -30dBm
50Ω
RA
2.6nH
22nH
51pF
R*
6.3pF
0.7pF
RDP=3.5 kΩ
VL
(a)
Resultant PI
network (QP*=60)
source
25nH
51pF
CDP*=5.6pF
RL
VL
(b)

Figure 9(b) can also be modeled as an inductive coupling of two parallel RC circuits. A more general look at such a circuit is shown in Figure 10, as an inductive coupling of two parallel RLC resonators with a current source.

Figure 10. Inductive coupled parallel RLC small signal model of a generalized PI-matched antenna and diode rectifier.

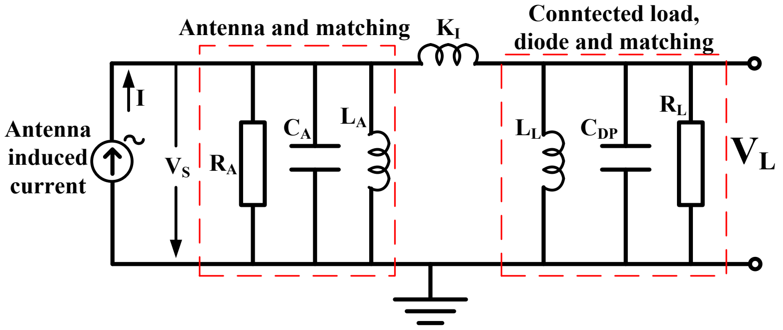
Antenna and matching
KI
Conntected load,
diode and matching
I
Antenna
induced
current
VS
RA
CA
LA
LL
CDP
RL
VL

The first parallel RLC resonator is modeled as impedance from the antenna and some passive matching components. The second parallel RLC resonator is modeled as impedance from the linearized diodes, its connected load and some passive matching components. I is the antenna induced current, V_S this time is the voltage across the parallel R_A and K_1 is the coupling element between the two parallel resonators. Using Kirchoff's current laws, the node equations can be expressed as Equations (29) and (30):

$$I = V_S\left[\frac{1}{R_A} + j\omega C_A - \frac{j}{\omega L_A} - \frac{j}{\omega K_1}\right] + \frac{jV_L}{\omega K_1} \tag{29}$$

$$0 = \frac{jV_S}{\omega K_1} + V_L\left[\frac{1}{R_L} + j\omega C_{DP} - \frac{j}{\omega L_L} - \frac{j}{\omega K_1}\right] \tag{30}$$

Load voltage (V_L) and the source voltage (V_S) at resonance are given by the equations in Equation (31). The resonance frequencies are given by Equation (32):

$$V_L = \frac{\frac{-jI}{\omega K_1}}{\left[\frac{1}{R_A R_L} + \frac{1}{\omega^2 K_1^2}\right]} \quad \text{and} \quad V_S = \frac{\frac{I}{R_L}}{\left[\frac{1}{R_A R_L} + \frac{1}{\omega^2 K_1^2}\right]} \tag{31}$$

$$\omega C_A - \frac{1}{\omega}\left\{\frac{1}{L_A} + \frac{1}{K_1}\right\} = 0 \text{ and } \omega C_{DP} - \frac{1}{\omega}\left\{\frac{1}{L_L} + \frac{1}{K_1}\right\} = 0 \tag{32}$$

From V_L and Vs in Equation (31), the voltage gain at resonance can be expressed as:

$$\frac{V_L}{V_S} = \frac{R_L}{j\omega K_1} \tag{33}$$

The maximum of Equation (33) is obtained when:

$$j\omega K_1 \to 0 \;\; or \;\; R_L \to \infty \tag{34}$$

Since $j\omega K_1$ is restricted by the conditions in Equation (32) to attain resonance, one cannot manipulate $j\omega K_1$ alone without changing the resonance conditions. What can drive the voltage gain is if R_L is very large at resonance conditions. If the input impedance (V_S/I) of the coupled resonator is maximum at resonance, conditions in Equation (35) hold:

$$\left(\frac{Vs}{I}\right) \to maximum \;\; when \;\; \frac{R_L}{\omega^2 K_1^2} \to 0 \tag{35}$$

Equation (36) may be assumed when $\frac{R_L}{\omega^2 K_1^2} \to 0$:

$$\frac{V_S}{I} = R_A \tag{36}$$

Under these conditions and an optimal coupling coefficient K_{1max}, the maximum voltage gain of the parallel coupled resonator can be written as in Equation (37), where K_{1max} is given by Equation (38):

$$\left|\frac{V_L}{V_S}\right| = \left|V_{gain}\right| = \left|\frac{1}{2}\sqrt{\frac{R_L}{R_A}}\right| \tag{37}$$

$$K_{1(\max)} = \mp\frac{1}{\omega}\sqrt{R_A R_L} \tag{38}$$

The analysis of Section 2.5.2 and parallel coupled RLC resonators show that any antenna and matched rectifying diode can be described as an equivalent circuit of a coupled resonator at a defined operating point. This general model can be applied to optimize other harvesters with complex output impedance such as piezo-harvesters or vibration harvesters for maximum transfer of power or voltage to its connected load. The model can also be applied to near field magnetically coupled antennas/coils for optimization.

2.5.6. Broadband RF to DC power converter

A broadband network is preferred when an RF to DC power converter is to be operated for a wide range of frequencies. A broadband converter is realized by connecting successive L-networks together in a multi-network between the antenna source and the rectifying diodes. The result is broadband or multiband RF power converter around certain frequencies. This can be deduced from the general model of a coupled resonators that by choosing certain passive components between a source and the load, it is possible to have more frequencies (ω) fulfilling Equation (32) and hence a result of multiple resonant frequencies or broader bandwidth at match conditions. For a two stage L-connected match, the quality factor of the L-network with the parallel resistance is given by Equation (39):

$$Q_P^{\,*} = \sqrt{\frac{R_P}{R^*} - 1} \tag{39}$$

The quality factor of the L-network with the series resistance is given by Equation (40):

$$Q_S^{\,*} = \sqrt{\frac{R_*}{R_S} - 1} \tag{40}$$

From Equations (39) and (40) the loaded quality factor of the two stage L-connected broadband network may be written as Equation (41) in terms of the unloaded quality factors; Q_S* and Q_P*:

$$Q^2 = \{(Q_P^{*2} + 1)(Q_S^{*2} + 1)\} - 1 \tag{41}$$

R* in this case may be chosen if it is larger than R_S and lower than the R_P. The highest possible bandwidth between a resistive source and resistive load is found for a virtual resistance (R*) given in Equation (42) [37]:

$$R^* = \sqrt{R_S R_P} \tag{42}$$

For complex loads such as rectifying diodes or transistors, the largest achievable bandwidth prescribed by Equation (42) is limited by the load or source component quality factor, since Equation (42) does not take into account reactive impedance associated with the source or load.

2.5.7. Broadband-Match RF to DC Converter Results and Discussion

The antenna source resistance was broadband matched to the HSMS-285C diodes (and load) resistance at −30 dBm input around 434 MHz. For a desired Q_P* and Q_S* of 2.7 there is ~0.4 pF inherent diode capacitance which is un-tuned using a two stage L-matching network [Figure 11(b)]. This is because the HSMS-285C diodes provides an inherent component quality factor of 6.96 at 434 MHz for −30 dBm input, hence a broadband circuit with Q_P* lower than this inherent component quality factor of the diodes (and load) is difficult to achieve without trade-offs. However, connected L-networks with Q_P* as high as the diode component quality factor may perform worse than a single L-matched network with similar loaded quality factor. This is due to redundant components of the connected L-networks which have inherent losses.

Figure 11. Impedance diagram of broadband RF power converter; (**a**) Broadband match around 434 MHz with loaded Q of 2.7, (**b**) Resultant impedance matching network with un-turned capacitance of 0.4 pF.

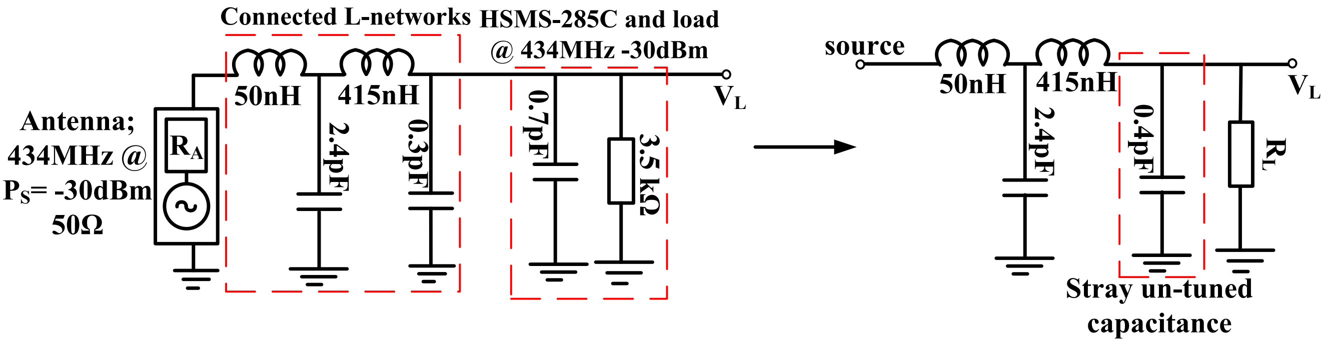
Connected L-networks
HSMS-285C and load
@ 434MHz -30dBm
Antenna;
434MHz @
P_S= -30dBm
50Ω
R_A
50nH
415nH
2.4pF
0.3pF
0.7pF
3.5 kΩ
V_L
source
50nH
415nH
2.4pF
0.4pF
R_L
V_L
Stray un-tuned
capacitance

(a) **(b)**

Figure 12. Measured open circuit S_{11} of the broadband circuit around 434 MHz at different input power levels from a 50 Ω source (**left**), measured open circuit input impedance at −10 dBm and −40 dBm of the broadband circuit (**right**).

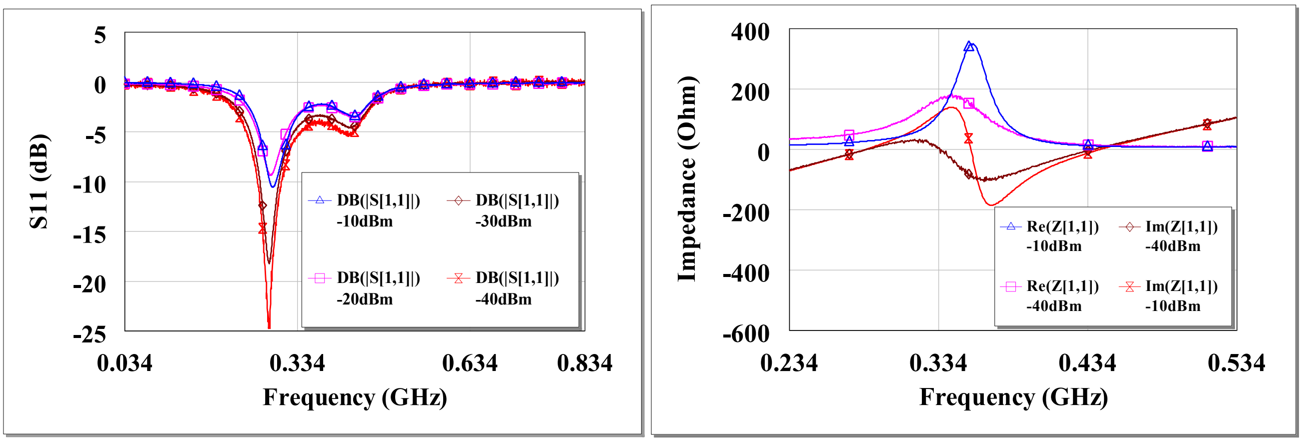

Therefore the broadband circuit is matched for Q_P* of 2.7, notwithstanding the un-tuned shunt capacitance as can be seen in Figure 11(b). Figure 12 shows the circuit S_{11} at various input power levels and input impedance at open circuit conditions. From Figure 12 (left) there is ~−5 dB return loss from 200 MHz to 500 MHz providing an operating band of ~300 MHz. The impedance of the circuit shows resonances at ~290 MHz and ~450 MHz as shown in Figure 12(right). A third resonance occurs around 356 MHz at −10 dBm as the frequency of highest harvester input resistance (~350 Ω) and where the reactive impedances approach their extremes. Figure 12 show that a wireless EM harvester can exhibit different resonance scenarios depending on the dominant instantaneous conditions. The efficiency and voltage sensitivity of the broadband matched wireless EM harvester are shown in Figure 13. The average open circuit voltage is 47 mV and 1.1 V at −30 dBm and −10 dBm, respectively, when operating from 200 MHz to 500 MHz.

Figure 13. Measured open circuit voltage *versus* frequency sweep from 200 MHz to 500 MHz for −10 dBm and −30 dBm (**left**), measured efficiency at 17 kΩ load *versus* frequency sweep for −10 dBm and −30 dBm (**right**).

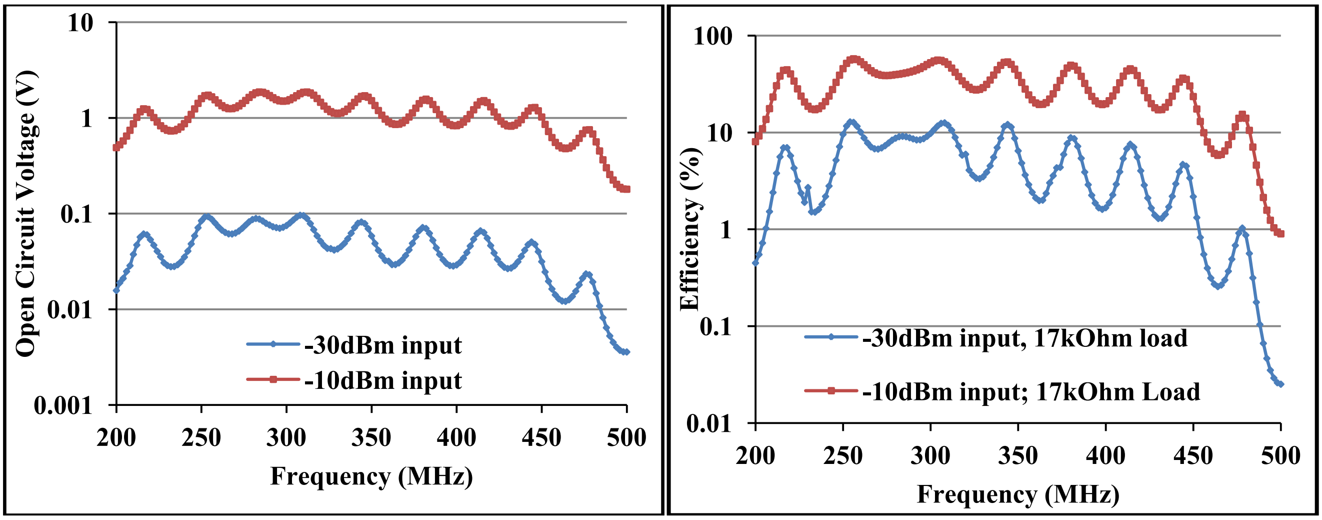

The broadband circuit achieves average efficiency of 5% at 17 kΩ load for −30 dBm and 30% at 17 kΩ load for −10 dBm input power from 200 MHz to 500 MHz. Figure 13 further confirm a direct link between frequency response and the unloaded quality factors. For Q_S* and Q_P* of ~2.7, the circuit response is broadband around 434 MHz.

2.6. High Voltage Sensitive RF to DC Converter

The current state of the art low power remote sensors would require a DC voltage supply of about 1 V and DC current of about 30 μA for operation. Therefore, the issue is not only how efficient a wireless EM harvester is in converting RF to DC power, but also what the output DC voltage and current of the EM harvester are at the RF input power level [38]. Equations (11,24) and (33) show that the maximum voltage sensitivity of a coupled resonator system or an RF to DC power converter is mostly related to the load and the source resistances at resonance. Therefore high voltage sensitive wireless EM harvester can be achieved with a diode voltage doubler with a very high input resistance relative to the antenna source without the need to cascade the diodes as in voltage multipliers. If the diodes been used for the RF to DC power conversion cannot provide high resistive impedance at the working frequency relative to the antenna source, then a DC-DC converter can be applied after the EM harvester as presented in [39] or the diodes may be cascaded by way of multipliers as presented in our earlier work [40] and by several other authors [3,5,14]. In case of multipliers, the input voltage ought to be high enough to overcome the junction potential of the several diodes in the multiplier network. If frequency is not a constraint, then a frequency sweep *versus* impedance for the diodes can be made and the frequency where the diodes exhibits high resistive impedance can be used to realize high voltage sensitive wireless RF harvester. For Schottky diodes, high resistive impedance occurs mostly at lower frequencies (see Figure 3). The measured voltage gain of a high resistive diode pair (voltage doubler) is presented in the next results.

2.6.1. High Voltage Sensitive RF to DC Converter Results and Discussion

The presented result was L-matched using 50 Ω resistance of the antenna source and the resistance of the HSMS-286C diodes (and load). The HSMS-286C diodes do provide high resistive impedance at low frequencies; notwithstanding the flicker noise which causes its resistive (and reactive) impedance to fluctuate. The HSMS-286C has low forward junction potential (~350 mV at 1 mA) per diode and series impedance of ~1.5–j8.1 kΩ or parallel impedance of ~−*j*8.3(46.3)/(−*j*8.3 + 46.3) kΩ at 13.6 MHz for −30 dBm input. Even though the HSMS-286C diodes unloaded component quality factor at 13.6 MHz is similar to that of the HSMS-285C diodes at 434 MHz, the elevated resistive impedance at 13.6 MHz fulfills the condition for high voltage sensitivity relative to a 50 Ω antenna source at resonance conditions.

Figure 14. (**a**) L-matched impedance diagram for matching the HSMS-286C diodes at 13.6 MHz at -30 dBm input. (**b**) Resultant network, (**c**) PCB layout of the L-matched Greinacher doubler with adjusted values due to impedances provided by copper route on FR4 substrate with thickness of 1 mm. (**d**) Fabricated PCB of the L-matched RF to DC power converter.

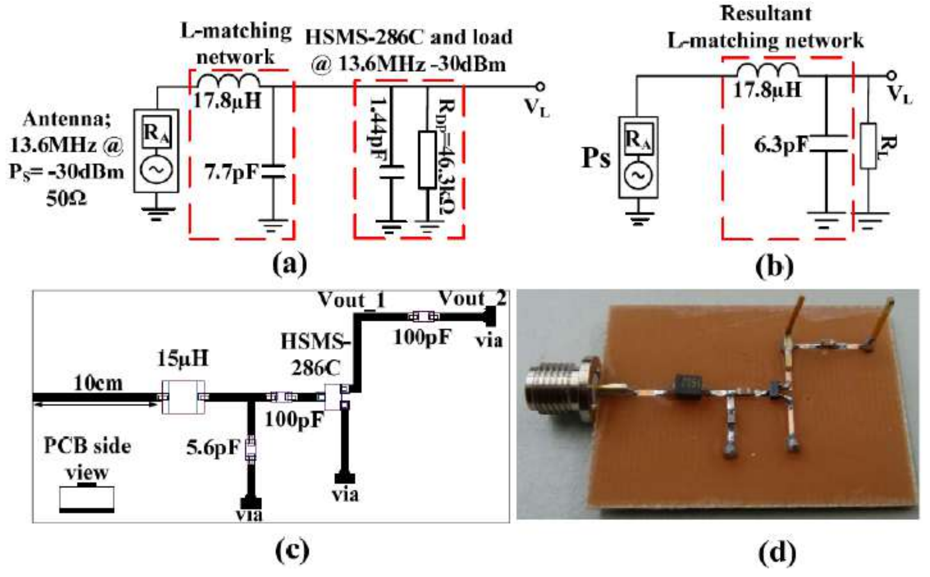
L-matching network
HSMS-286C and load @ 13.6MHz -30dBm
Antenna; 13.6MHz @ P$_S$= -30dBm 50Ω
R$_A$
17.8μH
7.7pF
1.44pF
R$_{DP}$=46.3kΩ
V$_L$
(a)
Resultant L-matching network
Ps
R$_A$
17.8μH
6.3pF
R$_L$
V$_L$
(b)
Vout_1
Vout_2
100pF
via
HSMS-286C
15μH
10cm
100pF
PCB side view
5.6pF
via
via
(c)
(d)

The high voltage sensitive EM harvester operating at 13.6 MHz is as shown in Figure 14. On the realized PCB is a Greinacher doubler. An inductance of 15 μH and a shunt capacitance of 5.6 pF were the adjusted values after the microstrip contributions.

The measured S_{11} and input impedance at open circuit are shown in Figure 15. There is high return loss and resonance around 13.6 MHz. The circuit input impedance at open circuit conditions is 58 Ω at resonance for both −40 dBm and −10 dBm.

Figure 15. Measured open circuit S_{11} of the L-matched HSMS-286C diodes at 13.6 MHz for different input power levels from a 50 Ω source (**left**), measured open circuit input impedance at −10 dBm and −40 dBm of the L-matched HSMS-286C diode at 13.6 MHz (**right**).

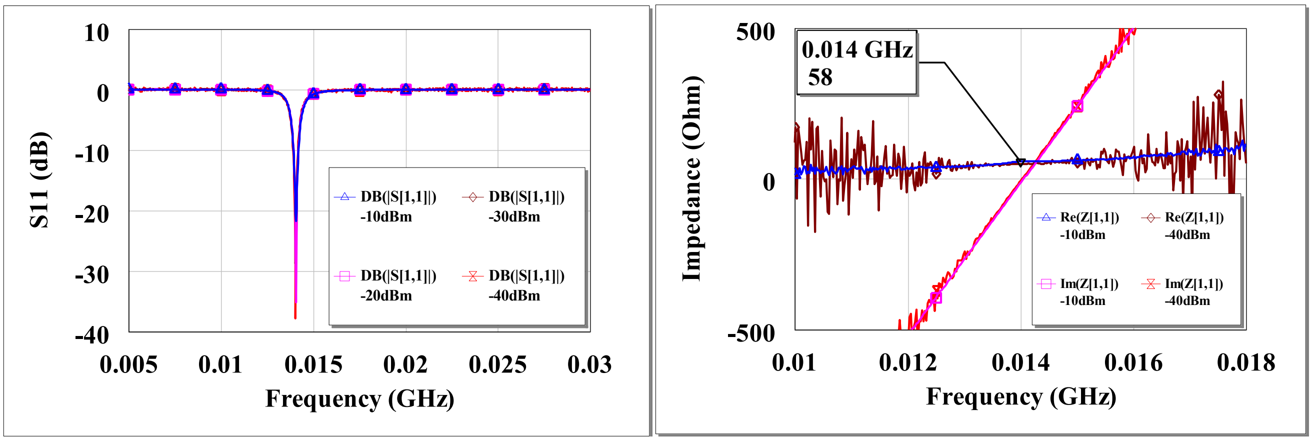

The efficiency and voltage sensitivity of the high voltage sensitive wireless EM harvester are shown in Figure 16.

Figure 16. Measured circuit efficiency *versus* load at various input power levels at 13.6 MHz (**left**), measured open circuit voltage and at 200 kΩ load *versus* input power at 13.6 MHz (**right**).

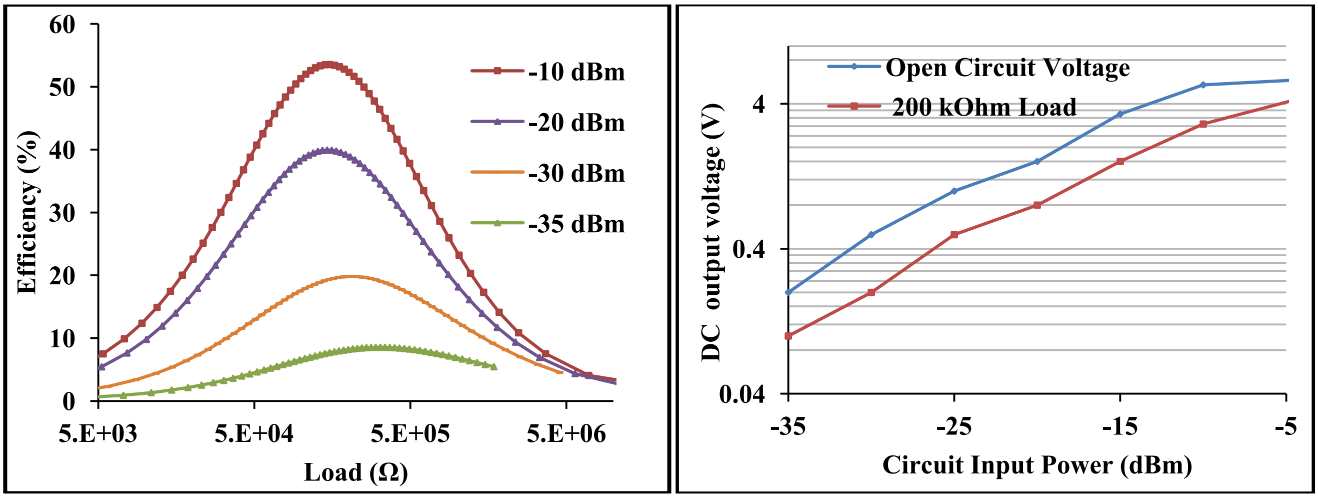

The maximum measured efficiency at −30 dBm is 20% for ~200 kΩ load and an open circuit voltage of 0.5 V. At −10 dBm, the maximum efficiency and open circuit voltage are 54% and 5.4 V respectively. At the optimal load of ~200 kΩ, the detected voltage is 0.2 V and 2.9 V at −30 dBm and −10 dBm respectively. The open circuit voltage gain is 100 at −30 dBm and 108 at −10 dBm.

Even though the RF to DC converter presented in Section 2.5.3 is the same as the L-match circuit realized with the HMSM-286C diodes at 13.6 MHz, the voltage gain is increased by a factor of 4 due to the large difference between the diodes (and load) resistance and source resistance so that at matched conditions high voltage gain occurs. The loaded *Q* of the L-matched circuit is 30 which results in small frequency bandwidth just like a PI-matched diode rectifier presented in our earlier work [40]. From this result and the results from our earlier presented PI-matched EM harvester, it can be inferred that all high loaded *Q* RF to DC circuits have high selectivity but not all highly selective RF to DC circuits have high loaded *Q*. The voltage sensitivity of the matched HSMS-286C diode at 13.6 MHz can be improved if its resistive impedance is not lowered by the flicker noise.

3. Wireless EM Power Harvester

A wireless EM harvester, consisting of a rectifying antenna (*rectenna*) was designed to find a compromise between size and performance of its antenna. The rectenna is shown in Figure 17.

Figure 17. Rectenna realized on a Duroid 5880, 1.57 mm substrate. (**a**) Backside of the rectenna, (**b**) cross-section of antenna output coplanar stripline dimensions (**c**) frontside of the rectenna.

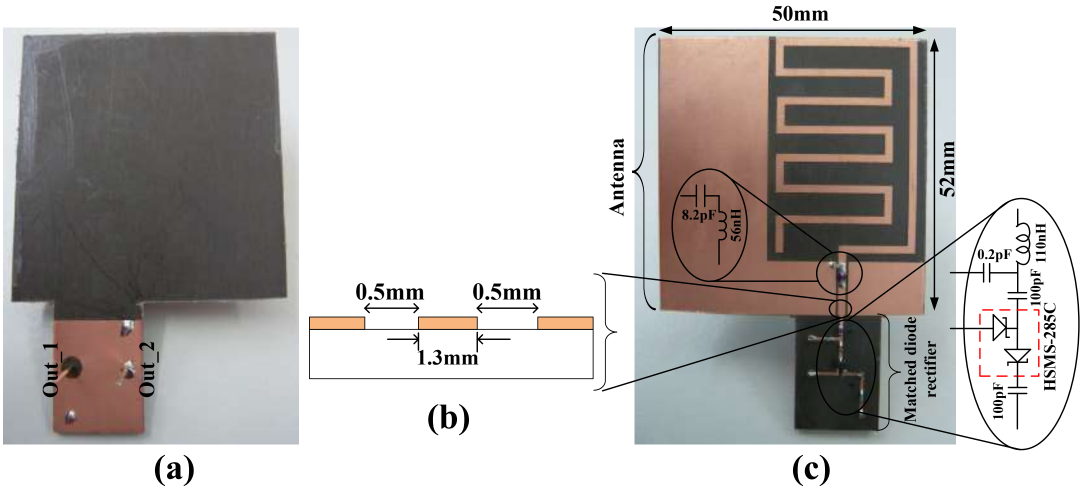

The antenna (planar) part of the rectenna is based on our earlier work [41]. In contrast to the earlier presented antenna, this rectenna is realized on a Duroid [42] substrate of thickness 1.57 mm. Duroid 5880 has lower loss tangent of 0.0004 at 1 MHz compared to 0.025 at 1 MHz for FR4. This means there is less loss in the transmission of signals on a Duroid PCB at this frequency range. The antenna part is fabricated to resonate around 434 MHz; hence its dimensions of 5 × 5.2 cm make it electrically small. The antenna is tuned with a chip inductor and a capacitor to achieve the resonance frequency around 434 MHz [Figure 17(c)]. This is done at a cost of reduced antenna radiation efficiency. An antenna is one of the few components the size of which is related to the operating frequency. Thus, if the size of an antenna is fixed, resonance frequency reduction of the antenna can only be achieved with penalty factors [10]. The antenna's output impedance is tuned with the dimensions of the coplanar stripline as shown in Figure 17(b).

HFSS [43] was used to simulate the presented antenna and to find the correct capacitive and inductive components for frequency tuning before the optimized design was fabricated. The simulated antenna resonances occur at 438 MHz and 445 MHz. At these frequencies, the radiation efficiency is 20% and a peak gain of −6 dBi. The rectifying part of the rectenna consists of L-matched HSMS-285C diodes (Figure 17(c)). The L-matched HSMS-285C part of the rectenna can be engineered to be as small as possible if required. The separate parts of the rectenna were characterized by terminating their ends and measuring the individual reflection coefficients just like the power converters presented in Section 2. Figure 18 shows the measured antenna and matched rectifier individual S_{11} and impedance. Figure 18 (left) also show the HFSS simulated S_{11} results. From Figure 18 (right), the measured antenna resonance where the input impedance is at maximum is ~434 MHz. At ~434 MHz, the antenna input resistance is 376 Ω and the reactive impedances approach their extreme (*so called anti-resonance*). The other resonance occurs when the input resistance is 'finite' and the reactive impedance is zero; at ~441 MHz. The input resistance is 57 Ω at ~441 MHz. The rectifier circuit is matched for the antenna's resistance at ~441 MHz.

Figure 18. Antenna HFSS simulated, antenna measured, and measured L-matched diode rectifier S_{11} on a Duroid 5880 PCB for −30 dBm input (**left**), Measured open circuit input impedance of antenna and rectifier at −30 dBm input (**right**).

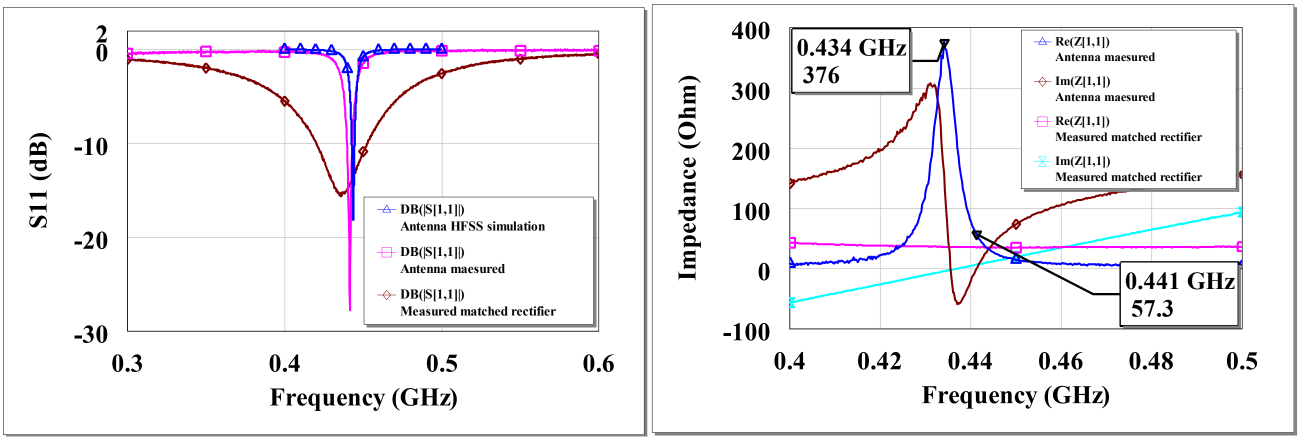

3.1. EM Range Results and Discussion

At far field between wireless EM transmitting and receiving antenna, the coupling mechanism between the transmitting and receiving antenna is neither capacitive nor inductive as is the case for the RF to DC converters. The coupling is radiative which can be described by the Friis equation of transmission on the assumption that the transmitting and receiving antenna are in free space [44]. A modified Friis equation for a transmitting and receiving antenna at far-field (R >> λ and R >> transmitting antenna largest dimension) to each other at a specified direction is given by Equation (43) [45]. Equation (43) assumes real world open space conditions:

$$\frac{P_r}{P_t} = F_{envt} G_t G_r \left(\frac{\lambda}{4\pi R}\right)^2, \tag{43}$$

where Pr is the power at the receiving antenna port, Pt is the power supplied at the transmitting antenna port, F_{envt} is a factor accounting for environmental effects as such ground reflections among others, Gt *and* Gr are the transmitting and receiving antenna gain (*at specified direction*) respectively. R is the distance between the transmitting and receiving antenna and λ is the wavelength of the transmitting EM wave. The rectenna receiving range measurements were carried out in an open space (*hall*) with the antennas 2 m above ground level. The transmitting and receiving antennas were arranged in the direction of their peak gain. The rectenna range performance is shown in Figure 19. According to Equation (43), the efficiency of RF power transferred between a sending and receiving antenna depends on controllable factors like the gain of the antennas in the arranged direction and the radiation efficiency of the antennas. Since the receiving/transmitting antenna's incorporated in remote harvesters for sensor powering are normally small in relation to their operating frequencies, they tend to be less efficient.

Figure 19. Rectenna receiving range performance by sending 17 dBm (50 mW) at a gain of −6 dBi at 437 MHz. Output DC voltage *versus* receiving distance for different loads (**left**), loads output power *versus* receiving distance (**right**).

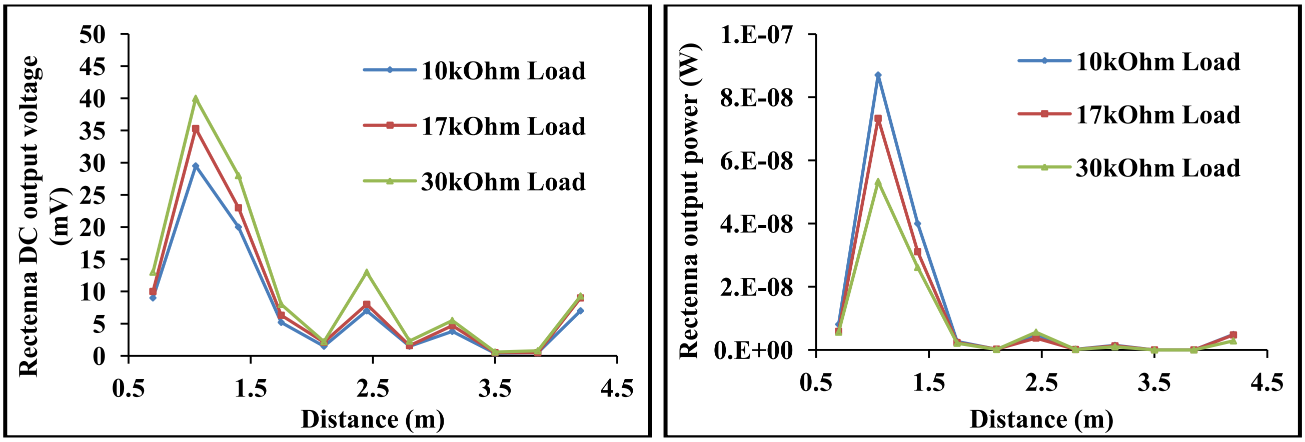

The efficiency of the rectenna's antenna is ~20% at resonance. A 'perfectly' matched RF to DC power converter operating in its square law region has efficiencies in the region of 20% as depicted in Section 2. The transmitting antenna was the same as the antenna incorporated in the rectenna. By transmitting the EM power with a small antenna (5 cm × 5.2 cm) at 437 MHz with efficiency of ~20% and at a gain of −6 dBi, the power delivered by the rectenna is generally low at far-field from the transmitter as can be seen in Figure 19. A mediocre transmitting antenna was used to transmit the EM waves due to limitations in the European Union about transmitting EM power at certain frequencies; so the goal in the rectenna range experiment is to show the lowest limit functionality of such a harvester. At 4.2 m from the electrically small transmitting antenna transmitting at 17 dBm, the rectenna harvested DC voltage and power are 9 mV and 5 nW respectively for 10 kΩ load. It can be seen from Figure 19 that the harvested voltage/power generally degrades as an inverse square of distance from transmitter as described by Friis equation. The measured received power however alternate along this R^{-2} fit as shown in Figure 19. This anomaly is accounted for by F_{envt} [Equation (43)] as influence of ground reflections and polarization in real world open field measurements [45]. For any particular distance R, the signals reflected from ground can be constructive with the direct signal to the rectenna, in which case the measured power may be higher than that predicted by the original Friss equation as in [44]. The ground effect can also be destructive, in which case the measured power will be lower than what is predicted by the original Friis equation.

4. Conclusions

Optimization of Schottky diode-based RF to DC power converters using different matching techniques for wireless EM energy harvesting applications is presented. Using scattering parameters for small signal modeling, it is shown that wireless EM harvesters can be generally described as coupled resonators with efficiencies and maximum voltage sensitivity depending mostly on the source and load resistances under matched conditions. The analytical models allow systematic control in the design of passive wireless EM harvesters. Based on these analyses, a rectenna is built and tested for lower limit functionality from harvesting ambient EM waves. The analysis presented in this work may also be applied to optimize derivatives of wireless EM harvesters like RFID tags, NFC, wireless chargers *etc.*, for efficient powering of their sensors or integrated circuits. Generally, most energy harvesters and their matched loads can be described as coupled resonators and thus may be optimized with the methods presented in this work.

Acknowledgments

This work is part of the graduate program GRK 1322 Micro Energy Harvesting at IMTEK, University of Freiburg, funded by the German Research Foundation (DFG). Special thanks to Daniela Ohnemus for the PCB preparations and Uwe Burzlaff for antenna range measurements.

References

1. Fischer, J. NFC in cell phones: The new paradigm for an interactive world [Near-Field Communications]. *IEEE Commun. Mag.* **2009**, *47*, 22–28.
2. Shameli, A.; Safarian, A.; Rofougaran, A.; Rofougaran, M.; de Flaviis, F. Power harvester design for passive UHF RFID tag using a voltage boosting technique. *IEEE Trans. Microw. Theor. Tech.* **2007**, *55*, 1089–1097.
3. Karthaus, U.; Fischer, M. Fully integrated passive UHF RFID transponder IC with 16.7-µW minimum RF input power. *IEEE J. Solid-State Circuits* **2003**, *38*, 1602–1608.
4. Lee, S.H.; Jin, I.S. Interoperation of an UHF RFID reader and a TCP/IP device via wired and wireless links. *Sensors* **2011**, *11*, 10664–10674.
5. Mandal, S.; Sarpeshkar, R. Low-power CMOS rectifier design for RFID applications. *IEEE Trans. Circuits Syst. I: Regul. Papers* **2007**, *54*, 1177–1188.
6. Yu, H.; Bashirullah, R. A Low Power ASK Clock and Data Recovery Circuit for Wireless Implantable Electronics. In *Proceedings of IEEE Custom Integrated Circuits Conference, CICC '06*, San Jose, CA, USA, September 2006; pp. 249–252.
7. Hannan, M.A.; Abbas, S.M.; Samad, S.A.; Hussain, A. Modulation techniques for biomedical implanted devices and their challenges. *Sensors* **2011**, *11*, 297–319.

8. Cho, N.; Song, S.-J.; Lee, J.Y.; Kim, S.; Kim, S.; Yoo, H.-J. A 8-μW, 0.3-mm2 RF-powered Transponder with Temperature Sensor for Wireless Environmental Monitoring. In *Proceedings of IEEE International Symposium on Circuits and Systems, ISCAS*, Kobe, Japan, May 2005; Volume 5, pp. 4763–4766.
9. Brown, W.C. The history of power transmission by radio waves. *IEEE Trans. Microw. Theor. Tech.* **1984**, *32*, 1230–1242.
10. Mickle, M.H.; Mi, M.; Mats, L.; Capelli, C.; Swift, H. Powering autonomous cubic-millimeter devices. *IEEE Antennas Propag. Mag.* **2006**, *48*, 11–21.
11. McSpadden, J.O.; Chang, K. A Dual Polarized Circular Patch Rectifying Antenna at 2.45 GHz for Microwave Power Conversion and Detection. In *Proceedings of IEEE MTT-S International Microwave Symposium Digest*, San Diego, CA, USA, May 1994; Volume 3, pp. 1749–1752.
12. Sample, A.; Smith, J.R. Experimental Results with Two Wireless Power Transfer Systems. In *Proceedings of the 4th International Conference on Radio and Wireless Symposium*, San Diego, CA, USA, January 2009; pp. 16–18.
13. Umeda, T.; Yoshida, H.; Sekine, S.; Fujita, Y.; Suzuki, T.; Otaka, S. A 950-MHz rectifier circuit for sensor network tags with 10-m distance. *IEEE J. Solid-State Circuits* **2006**, *41*, 35–41.
14. Le, T.; Mayaram, K.; Fiez, T. Efficient far-field radio frequency energy harvesting for passively powered sensor networks. *IEEE J. Solid-State Circuits* **2008**, *43*, 1287–1302.
15. Yuan, F. *CMOS Circuits for Passive Wireless Microsystems*; Springer: New York, NY, USA, 2011.
16. Zbitou, J.; Latrach, M.; Toutain, S. Hybrid rectenna and monolithic integrated zero-bias microwave rectifier. *IEEE Trans. Microw. Theor. Tech.* **2006**, *54*, 147–152.
17. Ungan, T.; Le Polozec, X.; Walker, W.; Reindl, L. RF Energy Harvesting Design Using High Q Resonators. In *Proceedings of IEEE MTT-S International Microwave Workshop on Wireless Sensing, Local Positioning and RFID, IMWS 2009*, Cavtat, Croatia, September 2009; pp. 1–4.
18. Ungan, T.; Freunek, M.; Muller, M.; Walker, W.D.; Reindl, L.M. Wireless Energy Transmission Using Electrically Small Antennas. In *Proceedings of IEEE Radio and Wireless Symposium, RWS '09*, San Diego, CA, USA, January 2009; pp. 526–529.
19. Boquete, L.; Rodríguez-Ascariz, J.M.; Barea, R.; Cantos, J.; Miguel-Jiménez, J.M.; Ortega, S.; Data acquisition, analysis and transmission platform for a pay-as-you-drive system. *Sensors* **2010**, *10*, 5395–5408.
20. Heikkinen, J.; Salonen, P.; Kivikoski, M. Planar Rectennas for 2.45 GHz Wireless Power Transfer. In *Proceedings of Radio and Wireless Conference, RAWCON 2000*, Denver, CO, USA, 10–13 September 2000; pp. 63–66.
21. Akkermans, J.A.G.; van Beurden, M.C.; Doodeman, G.J.N.; Visser, H.J. Analytical models for low-power rectenna design. *IEEE Antenn. Wireless Propag. Lett.* **2005**, *4*, 187–190.
22. Hagerty, J.A.; Helmbrecht, F.B.; McCalpin, W.H.; Zane, R.; Popovic, Z.B. Recycling ambient microwave energy with broad-band rectenna arrays. *IEEE Trans. Microw. Theor. Tech.* **2004**, *52*, 1014–1024.
23. Harb, A. Energy harvesting: State-of-the-art. *Renew. Energ.* **2011**, *36*, 2641–2654.
24. Vullers, R.J.M.; van Schaijk, R.; Doms, I.; van Hoof, C.; Mertens, R. Micropower energy harvesting. *Solid-State Electron.* **2009**, *53*, 684–693.

25. Sah, C.-T. *Fundamentals of Solid-State Electronics*; World Scientific: Singapore, Singapore, 1991.
26. Wetenkamp, S. Comparison of Single Diode *vs.* Dual Diode Detectors for Microwave Power Detection. In *Proceedings of IEEE MTT-S International Microwave Symposium Digest*, Boston, MA, USA, June 1983; pp. 361–363.
27. Cardoso, A.J.; Schneider, M.C.; Montoro, C.G. Design of Very Low Voltage CMOS Rectifier Circuits. In *Proceedings of Circuits and Systems for Medical and Environmental Applications Workshop (CASME)*, Merida, Mexico, December 2010; pp. 1–4.
28. Cardoso, A.J.; de Carli, L.G.; Galup-Montoro, C.; Schneider, M.C. Analysis of the Rectifier Circuit Valid Down to Its Low-Voltage Limit. *IEEE Trans. Circuits Syst. I: Regul. Papers* **2012**, *59*, 106–112.
29. Data sheet HSMS-285x. Avago Technologies. Available online: http://www.avagotech.com (accessed on 28 September 2012).
30. Data sheet HSMS-286x. Avago Technologies. Available online: http://www.avagotech.com (accessed on 28 September 2012).
31. Hsu, T.S. Low-frequency excess noise in metal—Silicon Schottky barrier diodes. *IEEE Trans. Electron Devices* **1970**, *17*, 496–506.
32. Hastas, N.A.; Dimitriadis, C.A.; Dozsa, L.; Gombia, E.; Amighetti, S.; Frigeri, P. Low frequency noise of GaAs Schottky diodes with embedded InAs quantum layer and self-assembled quantum dots. *J. Appl. Phys.* **2003**, *93*, 3990.
33. Gomila, G.; Reggiani, L.; Rubı́, J.M. Shot-noise suppression in Schottky barrier diodes. *J. Appl. Phys.* **2000**, *88*, 3079.
34. Kleinpenning, T.G.M. Low-frequency noise in Schottky barrier diodes. *Solid-State Electron.* **1979**, *22*, 121–128.
35. Lee, T.H. *The Design of CMOS Radio-Frequency Integrated Circuits*; Cambridge University Press: Cambridge, UK, 2004.
36. Advanced Design System. Agilent Technologies: Santa Clara, CA, USA.
37. Bowick, C.; Blyler, J.; Ajluni, C.J. *RF Circuit Design*; Newnes/Elsevier: Amsterdam, The Netherlands, 2008.
38. Joe, J.; Chia, M.; Marath, A.; Ang, C. Zero Bias Schottky Diode Model for Low Power, Moderate Current Rectenna. In *Proceedings of Defence Electronics Technology Seminar (DETS 97)*, Singapore, Singapore, November 1997; pp. 1–5.
39. Visser, H.J.; Vullers, R.J.M.; Veld, B.O.h.; Pop, V. Remote RF Battery Charging. In *Proceedings of Power MEMS 2010*, Leuven, Belgium, December 2010; pp. 37–40.
40. Nimo, A.; Grgic, D.; Reindl, L.M. Impedance optimization of wireless electromagnetic energy harvester for maximum output efficiency at μW input power. *Proc. SPIE* **2012**, *8341*, 83410W:1–83410W:14.
41. Nimo, A.; Grgic, D.; Reindl, L.M. Electrically small Planner Antenna for Compact Electromagnetic (EM) Wireless Energy Harvester. In *Proceedings of Power MEMS 2011*, Seoul, Korea, November 2011; pp. 310–313.
42. RT-duroid-5870–5880-Data-Sheet; Rogers Corporation: Rogers, CT, USA, 2011.

43. *HFSS*; Ansys Corporation. Ansys, Inc: Canonsburg, PA, USA, 2005.
44. Friis, H.T. A Note on a Simple Transmission Formula. *Proc. IRE* **1946**, *34*, 254–256.
45. Kvaksrud, T.I. *Range Measurements in an Open Field Environment*, *Design Note DN018*; Texas Instrum. Incorporated: Dallas, TX, USA, 2008.
46. *LabView*; National Instruments: Austin, TX, USA, 2009.

Appendix A: Measuring Setup for RF Rectifier Efficiency and Voltage Sensitivity

The measuring setup is as shown in Figure A1.

Figure A1. RF to DC Power converter characterization setup.

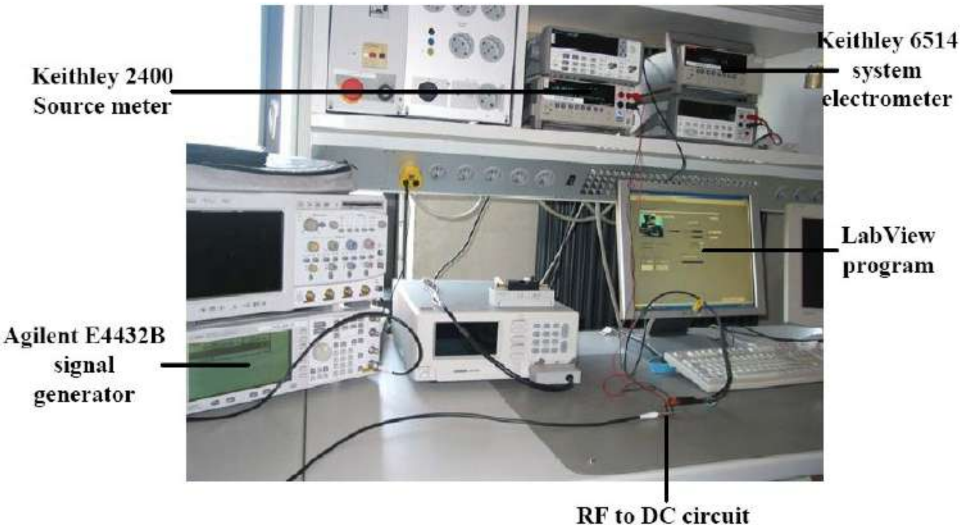

The RF to DC circuit efficiency and voltage sensitivity measurements were made with a Keithley 2400 source meter and Keithley 6514 system electrometer with an Agilent E4432B signal generator providing 50 Ω RF signal into the circuit board.

The closed circuit current drawn by the RF to DC power converter (*without load*) from the generator is first determined by the Keithley 2400 source meter; then starting from this current, the value of the current is decreased at set intervals to creates virtual load resistances to the circuit for up to a lowest current of 0.1 µA. The 6514 system electrometer is used to measure the output voltage. The number of data point is set through LabView [46] as well as the measurements. Additionally open circuit voltage or at specific loads and frequency sweep can be made through the LabView program. At −40 dBm input power and below, the detected voltages and currents were difficult to measure accurately with the measuring setup; hence measurements were made up to a minimum of -35 dBm input power. The circuit layout for the efficiency and voltage sensitivity measurements is schematically shown in Figure A2.

Figure A2. RF to DC power converter characterization circuit.

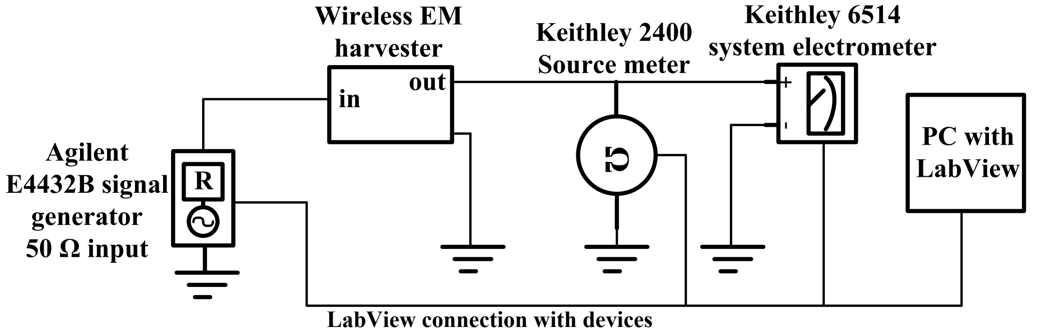
Wireless EM harvester
in
out
Keithley 2400 Source meter
Keithley 6514 system electrometer
+
-
Agilent E4432B signal generator 50 Ω input
R
PC with LabView
LabView connection with devices

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).

STANDARD ANTENNA TERMS AND RELATED FORMULAS

The "IEEE Standard Definitions of Terms for Antennas" (IEEE STD-145) represents a consistent and comprehensive vocabulary suited for the effective communication and understanding of antenna theory. General use of these definitions of terms would eliminate much of the wide-spread inconsistency concerning antenna characteristics, particularly with regard to the basic parameters of gain, beamwidth, polarization and efficiency. For convenience, IEEE antenna terms of general interest are listed here. Wherever these terms appear in this catalog, the definitions given below apply. Other commonly used terms, not covered by the IEEE standard, are shown with an "*."

ANTENNA APERTURE. A surface, near or on an antenna, on which it is convenient to make assumptions regarding the field values for the purpose of computing fields at external points. **Note:** The aperture is often taken as that portion of a plane surface near the antenna, perpendicular to the direction of maximum radiation, through which the major part of the radiation passes.

ANTENNA EFFICIENCY OF APERTURE - TYPE ANTENNA. For an antenna with a specified planar aperture, the ratio of the maximum effective area of the antenna to the aperture area.

*** ANTENNA FACTOR.** That quantity by which the voltage developed across the output of an antenna is related to the incident field strength in which the antenna is immersed. **Note:** Applicable to low frequency antennas and usually refers to a 50 ohm output.

$$\mathrm{AFE\ (dB\ m^{-1}) = E\ (dB\ V/m)\ -\ V\ (dB\ V)}$$

AFE = Electric Field Antenna Factor
E = Electric Field Strength at antenna
V = Voltage at terminals of antenna

$$\mathrm{AFH\ (dB\ AV^{-1}\ m^{-1}) = H\ (dB\ A/m) - V(db\ V)}$$

AFH = Magnetic Field Antenna Factor
H = Magnetic Field Strength at antenna
V = Voltage at terminals of antenna

$$\mathrm{AFE\ (dB\ m^{-1})\ =\ AFH\ (dB\ AV^{-1}\ m^{-1}) + 51.53}$$

for a plane wave in free space.

$$\mathrm{AFB\ (dB\frac{pT}{\mu V})\ =\ B\ (dBpT)\ -\ V_o\ (dB\mu V)}$$

AFB = Magnetic flux Antenna Factor
B = Magnetic flux at the antenna
pT: picoTesla units
V = Voltage at the terminals of the antenna

$$\mathrm{AFB\ (dB\frac{pT}{\mu V}) = AFH\ (dB\frac{A}{Vm})\ +2}$$

APERTURE ILLUMINATION. The field over the aperture as described by amplitude, phase, and polarization distributions.

APERTURE ILLUMINATION EFFICIENCY. For a planar antenna aperture, the ratio of its directivity to the directivity obtained when the aperture illumination is uniform.

BEAM. The major lobe of the radiation pattern.

CIRCULAR POLARIZATION. It may be either right hand circular polarization (RHCP) or left hand circular polarization (LHCP). The sense of polarization is determined by observation of the direction of rotation of the electric field vector from a point behind the source, RHCP and LHCP correspond to clockwise and counter-clockwise respectively. **Note:** RHCP transmit requires a like polarization to receive.

CO-POLARIZED. The polarization which the antenna is intended to radiate or receive. Also "like polarization".

*** CROSS POLARIZATION DISCRIMINATION (XPD).** Cross polarization discrimination is the measure of the antenna's ability to differentiate between the vertical and the horizontal polarization of an antenna. This difference, shown in relative signal level, is indicated on directional pattern envelopes (DPE's).

STANDARD ANTENNA TERMS AND RELATED FORMULAS

* **DIRECTIONAL PATTERN ENVELOPES (DPE'S).** In accordance with standard practice, radiation characteristics in any given plane of polarization are measured and plotted using 360-degree polar coordinate systems. The resultant Directional Pattern Envelope is the smoothed composite of all these measurements. The purpose of these DPE's is to emphasize the worst composite condition.

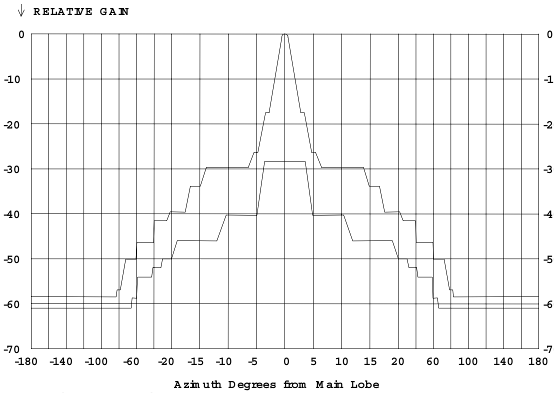

DIRECTIVE GAIN. In a given direction, 4 times the ratio of the radiation intensity in that direction to the total power radiated by the antenna.

DIRECTIVITY. The value of the directive gain in the direction of its maximum value.

EFFECTIVE AREA OF AN ANTENNA. In a given direction, the ratio of power available at the terminals of a receiving antenna to the power per unit area of a plane wave incident on the antenna from that direction, polarized coincident with the polarization that the antenna would radiate.

FAR FIELD REGION. That region of the field of an antenna where the angular field distribution is essentially independent of the distance from a specified point in the antenna region.

* **FRONT-TO-BACK RATIO**. The ratio of the maximum directivity of an antenna to its directivity in a specified rearward direction.

* **Gain, dBi**. The gain expressed in decibels relative to an isotropic radiator that is linearly polarized.

$$G(\mathrm{dBi}) = 10\log(G) \qquad G = 10^{\frac{G(\mathrm{dBi})}{10}}$$

* **GAIN, dBic**. The gain expressed in decibels relative to an isotropic radiator that is circularly polarized.

HALF-POWER BEAMWIDTH. In plane containing the direction of the maximum of a beam, the angle between the directions in which the radiation intensity is one half the maximum value of the beam.

HALF-WAVE DIPOLE. A half wavelength antenna, center fed so as to have equal current distribution in both halves. Mounted vertically, it has a doughnut shaped pattern, circular in the horizontal plane. It is an antenna that can be constructed. It has some inherent losses. When used as a gain reference, the half-wave dipole has a power gain of about 1.7 dBi.

* **ISOLATION.** Refers to the ability of one port of a dual polarized feed to discriminate against a signal fed into the other port.

ISOTROPIC RADIATOR. A hypothetical antenna having equal radiation intensity in all directions. Note: An isotropic radiator represents a convenient reference for expressing the directive properties of actual antennas.

NEAR-FIELD REGION. The spherical region of space between the antenna and the far field region.

NULL. The region of a radiation pattern, either computed or measured, where the amplitude goes through a minimum value. Note: (1) It represents the angular position where the phase or the far field pattern crosses the zero axis if the pattern is plotted as a phasor instead of a scalar value. Note: (2) The region outside the main beam of a directive antenna pattern consists of a series of minor lobes separated by nulls.

PARALLEL POLARIZATION. The condition where the electric vector is parallel to the local conducting surface. Note: Over the earth, this is usually referred to as being horizontal polarization.

PHASE CENTER. The location of a point associated with an antenna such that, if it is taken as the center of a sphere whose radius extends into the far-field, the phase of a given component over the surface of that radiation sphere is essentially constant, at least over the portion of the sphere where the radiation is significant.

STANDARD ANTENNA TERMS AND RELATED FORMULAS

POLARIZATION. The polarization of an antenna is defined as the polarization of the electromagnetic wave as described by the shape and orientation of an ellipse, which is the locus of the extremity of the field vector, and the sense in which the ellipse is traversed with time. The elliptical locus is called the polarization ellipse and the wave is said to elliptically polarized. Circular polarization and linear polarization are degenerate cases of elliptical polarization.

POWER DENSITY AT A POINT

$$S_{av} = \frac{GP_t}{4\pi r^2}$$

S_{av} = Time average power density in W/m^2
P_t = Power transmitted in watts
G = Antenna gain relative to an isotrope
r = Distance from antenna to point in meters

POWER DENSITY TO VOLTS/METER IN FREE SPACE

$$E^2 \ (V/m) = 377 \ S_{av} \ (W/m^2)$$

$$S_{av} \ (W/m^2) = E^2 \ (V/m) / 377$$

$$1 \ V/m = 2.65 \ mW/m^2$$

POWER GAIN. In a given direction, 4 times the ratio of the radiation intensity in that direction to the net power accepted by the antenna from the connected transmitter. **Note:** (1) When the direction is not stated, the power gain is usually taken to be the power gain in the direction of its maximum value. (2) Power gain does not include reflection losses arising from mismatch of impedance.

POWER GAIN IN PHYSICAL MEDIA. In a given direction and at a given point in the far field the ratio of the power flux per unit area from an antenna to the power flux per unit area from an isotropic radiator at a specified location with the same power input as the subject antenna.

Note: The isotropic radiator must be within the smallest sphere containing the antenna. Suggested locations are antenna terminals and points of symmetry, if such exist.

POWER GAIN REFERRED TO A SPECIFIED POLARIZATION. The power gain of an antenna, reduced by the ratio of that portion of the radiation intensity corresponding to the specified polarization to the radiation intensity.

POWER TRANSMISSION FORMULAS

$$P_r = P_t \frac{G_t G_r \lambda^2}{(4\pi r)^2}$$

$$P_r \text{ (dB W)} = P_t \text{ (dB W)} + G_t \text{ (dBi)} + G_r \text{ (dBi)} - 20 \log r - 20 \log f + 27.56$$

$$P_r \text{ (dB W)} = P_t \text{ (dB W)} - AFE_t \text{ (dB m}^{-1}) - AFE \text{ (dB m}^{-1}) - 20 \log r + 20 \log f - 32$$

P_r = Power received
P_t = Power transmitted
G_r = Gain of receiving antenna
G_t = Gain of transmitting antenna
f = Frequency in MHz, λ = Wavelength
r = distance between antennas in meters
AFE_r = AFE of receiving antenna
AFE_t = AFE of transmitting antenna

RADIATOR. Any antenna or radiation element that is a discrete physical and functional entity.

RADIATION, ELECTROMAGNETIC. The emission of energy in the form of electromagnetic waves.

RADIATION INTENSITY. In a given direction, the power radiated from an antenna per unit solid angle.

RADIATION LOBE. A portion of the radiation pattern bounded by regions of relatively weak radiation intensity

RADIATION PATTERN (ANTENNA PATTERN). A graphical representation of radiation properties of the antenna as a function of space coordinates. **Note:** (1) In the usual case the radiation pattern is determined in the far-field region and is represented as a function of directional coordinates. (2) Radiation properties include power flux density, field strength, phase, and polarization.

* **RADIATION RESISTANCE OF AN ELECTRICALLY SMALL LOOP ANTENNA.** The resistive component of an antenna's input impedance that results from the coupling of the antenna to its environment. This resistance dissipates the power that is actually radiated from the antenna.

$$R_r = 20\,(2\pi/\lambda)^4\, n^2 A^2 \text{ ohms}$$

n = number of turns
A = area of the loop

REALIZED GAIN. The power gain of an antenna in its environment, reduced by the losses due to the mismatch of the antenna input impedance to a specified impedance.

* **REALIZED RADIATOR EFFICIENCY.** The efficiency of an antenna in its environment reduced by all losses suffered by it, including: ohmic losses, mismatch losses, feedline transmission losses, and radome losses.

RELATIVE POWER GAIN. The ratio of the power gain in a given direction to the power gain of a reference antenna in its reference direction. **Note:** Common reference antennas are half-wave dipoles, electric dipoles, magnetic dipoles, monopoles, and calibrated horn antennas.

RETURN LOSS. The reflection coefficient of a mismatch expressed in decibels. **Note:** Modern swept VSWR techniques actually sense the reflected component which is normalized to the forward component to yield return loss. A 2:1 VSWR is equivalent to 9.5 dB return loss.

VSWR. The voltage standing wave ratio of a component such as an antenna. It is referred to the characteristic impedance of the transmission line being used. **Note:** The most common characteristic impedance is 50 ohms, but 75 and 300 ohms are frequently used in coaxial or twin lines for VHF, UHF applications.

NOTES:

THE HANDYMAN'S GUIDE TO OSCILLOSCOPES (Part 2 of 2)

by Paul Harden, NA5N

Making some advanced measurements with your Oscilloscope

Print as .pdf file
4 pages
8½ x 11 or A4

In Part 1, oscilloscope operation was covered for making basic voltage, time and frequency measurements. In this part, we'll continue with some more advanced uses of a scope, and in particular, how to use a scope for testing and troubleshooting ham radio QRP transceivers in the homebrewer's workshop.

Receiver Filter Bandwidth.

This procedure uses a scope (a DVM can be used with less accuracy) for determining the overall filter bandwidth (or selectivity) of a receiver. It is basically measured by plotting output voltage vs. audio frequency to construct a picture of the filter response.

Connect scope to the receiver audio output (speaker or phone jack); measurement will be based on peak-to-peak voltages (Vpp) on a scope, or rms voltage (Vrms) on a DVM.

Using a signal generator, set the frequency for the band of interest on your radio. For example, on a general coverage shortwave receiver, you might set it for 10 MHz (top end of the 31M band), or perhaps to 7.040 MHz on a 40M ham radio receiver/transceiver. Tune the receiver to the signal generator signal. If you don't have a signal generator, you can also tune to a steady carrier or station to produce a hetrodyne audio "pitch." Tune in the signal to the pitch that causes the maximum peak-to-peak display. Adjust the scope and volume control to produce a 2Vpp display (4 divisions). This is the peak response of the overall filtering stages as shown in **Fig. 13**.

Now determine the audio frequency at this peak response by measuring the time period between cycles and covert to frequency. In the example to the right, the period of one cycle is 1.7mS, which is an audio tone of 750Hz (1/.0017sec). A frequency counter on the output can also be used.

Next, tune the receiver such that the sidetone pitch goes UP in frequency and the peak-to-peak signal will decrease in magnitude. Tune to the point where the signal is exactly 1Vpp on the scope. See **Fig. 12**.

This is the –6dB point of the high end of the filter (20log 1v/2v =–6dB). Determine the frequency of the audio pitch as before. In the example, this is 1100 Hz. Record the data.

From these two data points, the –6dB bandwidth can be estimated. The bandwidth from the filter peak (750 Hz) to the –6dB point (1100 Hz) is 350 Hz. The bandwidth (BW) between the two –6dB points is usually twice this value, or 700 Hz. A filter with a –6dB BW of 700Hz is a mediocre filter for CW reception, and way too narrow for SSB or AM.

Of course you can determine the exact –6dB BW by tuning the receiver back to the 2Vpp peak response, and continue tuning DOWNward in frequency until the audio is again exactly 1Vpp. Determine this frequency and record. In this example, it should occur around 400 Hz if the filter shape is symmetrical.

Plot these three data points on a sheet of graph paper as shown in **Fig. 15** to construct the filter shape. Return to the upper –6dB point (1100Hz in the example) and continue tuning upwards in audio pitch, recording the frequency at 0.5v (–12 dB), 0.25v (–18 dB), 125mV (–24 dB), etc. Everytime you "halve the voltage," it is a 6 dB change. The more points you collect, the more accurate your filter response plot will be.

Fig. 13 – Displaying Filter Peak Response

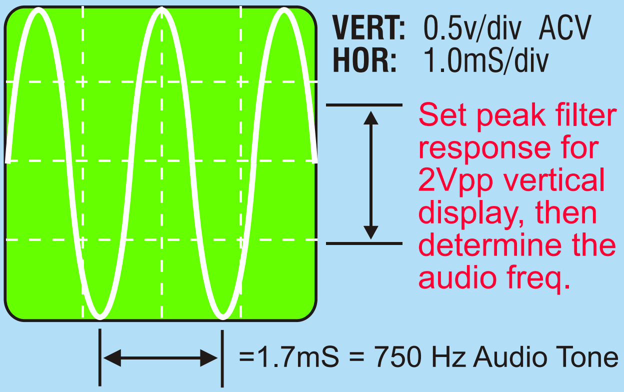
VERT: 0.5v/div ACV
HOR: 1.0mS/div
Set peak filter response for 2Vpp vertical display, then determine the audio freq.
=1.7mS = 750 Hz Audio Tone

Fig. 14 – Displaying –6dB Frequency

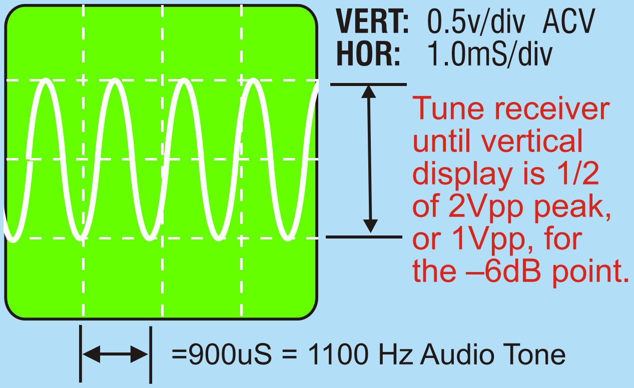
VERT: 0.5v/div ACV
HOR: 1.0mS/div
Tune receiver until vertical display is 1/2 of 2Vpp peak, or 1Vpp, for the –6dB point.
=900uS = 1100 Hz Audio Tone

Of interest to proper rig alignment is to repeat the above using the output of the product detector. To maximize the effectiveness of the filtering, the receive offset frequency set by the BFO should be adjusted to the same frequency as the peak frequency response of the audio. In this example, with the peak audio response occuring at 750Hz, if your BFO is set for a sidetone frequency of 700 Hz, you are loosing 2–3 dB, since this is in your filter skirt. This is shown on the response plot in **Fig. 15** by the dashed lines. By adjusting your BFO for a sidetone frequency of 750 Hz, you will pick up 3-4dB of overall gain in your receiver, plus increase the selectivity a bit as well. Why? Because nearby stations, such as one at 800–900Hz tone, could actually be louder than the 700Hz tone signal you are trying to copy, since the gain of the receiver is greater at those tones than at 700Hz, as shown in the plot.

Fig. 15 – Plotting Audio Response

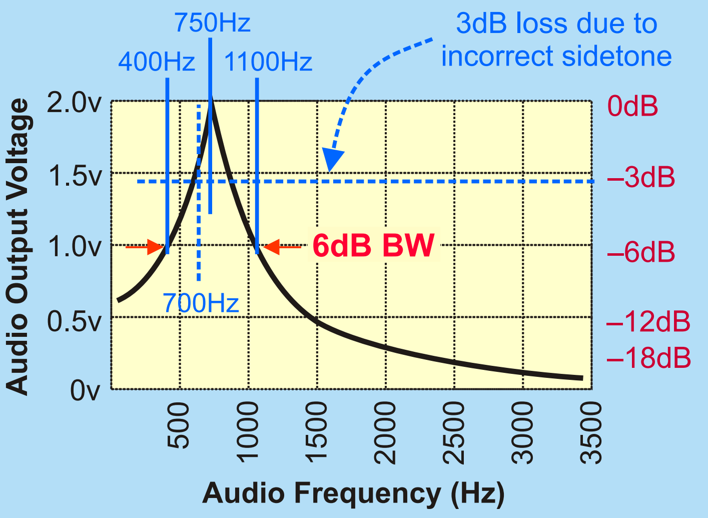

Opposite Sideband Rejection

A superhet receiver is supposed to pass just one sideband and reject the other. Poor opposite sideband rejection could indicate the crystals in your IF filter are not well matched or other problems. It is measured almost identical to plotting the filter response just described. First, you tune the receiver to the test signal to find the peak response frequency, or 750Hz in this example. Set the scope display for 2Vpp. Now tune downward in audio pitch, passing zero-beat, and continue tuning and you should hear the test signal, much weaker, now rising again in tone. This is the *opposite sideband*. Measure the peak-to-peak voltage, if you can. For example, say it is 80mVpp, as shown in **Fig. 16**. Calculate the opposite sideband rejection by:

$$\text{rejection} = 20\log\frac{80\text{mV}}{2.0\text{v}} = \frac{80\text{mV}}{2000\text{mV}} = -28\text{dB}$$

If you can't hear the opposite sideband, then obviously you have excellent filter rejection. If you can just barely hear it, you may have to increase the sensitivity of your scope (set vertical gain to 20 or 50 mV/div). In this example, –28dB rejection of the opposite sideband is quite good. A –30dB rejection means the opposite sideband is only 1/1000th of the desired sideband, a very suitable attenuation of the opposite sideband.

Fig. 16
Calculating Opposite Sideband Rejection

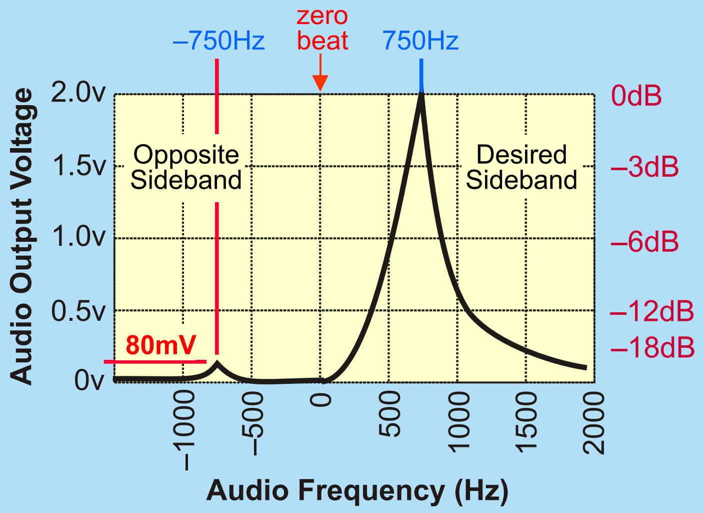

These tests are important to perform on your rig for documenting it's current performance, and repeated periodically to detect unfavorable changes or for troubleshooting when a problem is evident. For homebrewing, these tests can allow you to evaluate different circuits or when experimenting with different components or part values.

Oscillator Phase Noise

When homebrewing a basic oscillator circuit, such as a VFO, looking at the oscillator output on a scope can reveal several problems. One is to check for excessive phase noise. Phase noise is small variations of the oscillator frequency that causes power in the close-in sidebands, usually measured on laboratory equipment within 100KHz, or even within 10 KHz.

To check for oscillator phase noise, connect the oscilloscope to the oscillator output, loading the output of the oscillator as little possible. Most scopes have sufficiently high input impedances where this shouldn't be a problem, but some cheaper scopes can load an oscillator circuit. If you suspect your scope is loading the oscillator, couple the scope to the circuit with a small value capacitor, less than 20pF.

Display 2–3 cycles of the oscillator output as shown on the scope display shown in **Fig. 17**. Properly focus the scope and carefully observe if the waveform appears in focus at the peaks, but slightly out-of-focus at the zero-crossing points, that is, on the rising and falling edges of the sine wave.

If it appears out-of-focus, this is excessive phase noise jittering the signal and "smearing" the waveform along the time (horizontal) axis. Extreme phase noise may show 2-3 sine waves very close to each other, as shown in the exagerated waveform to the right — assuming you have your scope properly triggered.

Fig. 17 – Checking Oscillator Phase Noise

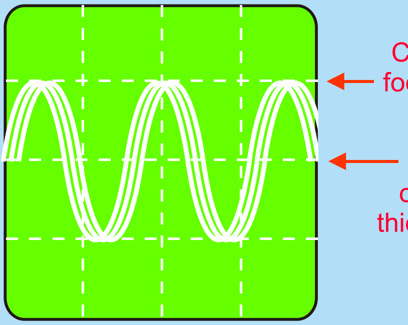

Check for a clean, focused signal here

Check for an out-of-focus or thicker signal here

Phase noise is random, instantaneous changes in the oscillator frequency that smears the display. If you can see this on a scope, the phase noise is excessive! If you can't see it, it doesn't mean the oscillator has no phase noise (all oscillators have some phase noise), it just means it is not excessive enough to see on a scope. A scope is not a good instrument for checking phase noise, but for homebrew circuits, it is a check to ensure you do not have a serious oscillator problem.

Excessive oscillator phase noise in receivers can cause IMD products and noise in the audio range at the output of the mixer(s), including the product detector. In a transmitter, excessive oscillator phase noise will put power in the close-in sidebands of your carrier, not only wasted power for lower transmitter efficiency, but may produce strange sounds (buzzing or chirping) to the receiving station. A few causes of phase noise are excessive current in the oscillator transistor, low-Q coil(s), high dissipation in the tuning caps or poor power supply filtering at the oscillator frequency.

AM (Amplitude) Noise

Another oscillator problem may be AM noise, or amplitude modulated noise. It is an opposite effect on a scope when displaying the oscillator output — the sine wave appears out-of-focus or thicker at the peaks, and in-focus elsewhere as shown in **Fig. 18**.

Fig. 18 – Checking for AM Oscillator Noise

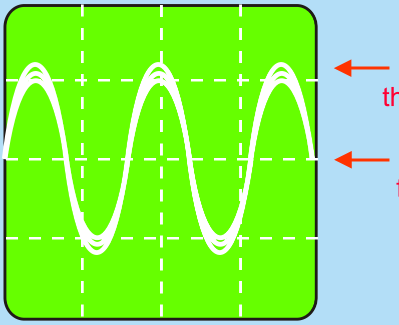

Check for an out-of-focus or thicker signal here

Check for a clean, focused signal here

If you detect AM noise, slow down the scope's sweep rate to the audio frequencies or slower to see if you can notice a lower frequency component. A common cause of AM oscillator noise is 60 Hz from the power mains leaking into the circuits. This is particularly true if using a power supply off of 120v 60 Hz, or sometimes it can be due to the AC lighting above your head! If the AM noise seems to be at the same frequency as the audio output tone, it means audio is getting into the Vcc bus, likely due to poor bypass filtering at the audio amplifier stages (particular if using an LM386 or similar).

If you can't find a low frequency component, the AM noise may be random, which may indicate poor voltage regulation, a noisy voltage regulator, or perhaps a circuit in VHF oscillation. If the AM noise seems to occur on key-down only on a transmitter, the transmit current may be loading the power supply, the voltage regulator is under-rated, or just simply loading the oscillator. In the case of loading the power supply or regulator(s), perhaps a separate voltage regulator or zener circuit should be used, dedicated for the oscillator(s). In the case of transmit loading, adding a buffer amplifier or emitter follower to isolate the load from the oscillator may help.

Much of this can be diagnosed also with the scope, by looking at the AC ripple on the DC power lines. You should have less than 50mV of any AC component on the 8-12v DC wiring, whether 60Hz, audio or RF. If >50mV, then additional low or high frequency filtering on the DC power is needed.

Monitoring Transmitter RF Output

RF Power (in watts) is E^2/R, where E is in rms and R=50 . The voltage displayed on the scope (peak-to-peak) must be converted to rms by Erms=.707(Epp/2). To measure properly, the transmitter should be on a 50 dummy load using the normal hi-Z scope input. If your scope has a selectable "50 "input, it can be the transmitter dummy load directly, providing it can tolerate the 50Vpp input. (Always know what the maximum p-p input voltage your oscope will tolerate. It is often stated on your scope at the vertical channel inputs. 50Vpp to 100Vpp are typical).

Measuring Transmitter RF Output Power

Figure 19 shows how you can use an oscilloscope to fairly accurately measure output power from a low power transmitter (generally 10W or less with a 1X probe and 20-30W with a 10X probe). Connect transmitter output to a 50 dummy load. Connect the scope lead to the transmitter output (after the low pass filters) or directly to the dummy load. Do not connect the scope leads to the collector or drain of the final PA transistor. The displayed voltage in this case will be erroneous.

In the example, the output transmit voltage across the dummy load measures 24Vpp. Convert this voltage to Vrms, then use the equation (Vrms squared divided by the load resistance) to calculate the power in watts. In the example, 24Vpp, the output power is 1.4 watts. A 5W QRP transmitter should produce about 45Vpp.

The accuracy of your power measurements depend upon the condition of your oscilloscope:

1) Ensure your scope's vertical sensitivity (volts/division) is properly calibrated (see Part I)
2) Ensure your vertical sensitivity is properly calibrated with the 1X or 10X probe you are using.
3) Ensure that the transmitter frequency is well within the bandwidth of your scope. A 100 MHz scope should give reliable readings on all HF bands. A 50 MHz scope should be reliable on all HF bands except slight errors at 28MHz.

Fig. 19 –Oscilloscope as a Power Meter

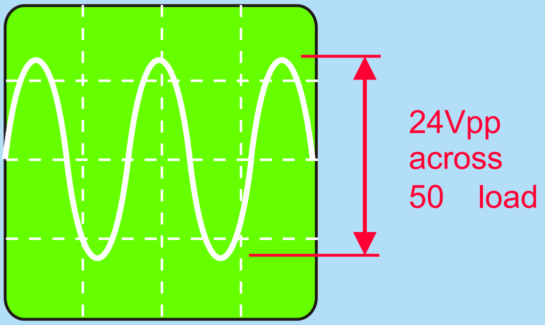

1. Measure output voltage. 24Vpp
2. Convert Vpp to Vrms

Vrms=.707(Vpp/2) 8.5Vrms

3. Calculate power in watts

$$\frac{Vrms^2}{R} = \frac{8.5v^2}{50} = 1.4W$$

Checking Transmitter Purity

Phase Noise of the transmitter is measured identically to the phase noise checks on page 2–3. With the scope properly triggered and focused, you are looking for an out-of-focus "fuzziness" around zero crossing. If it appears phase noise exists, the fault is generally at the beginning of the transmit chain. That is, seldom does the PA transistor add phase noise; it's merely amplifying what it is given. Phase noise on the transmitter output is more likely due to the transmit oscillator, or the transmit mixer if used.

Harmonic Power can also be detected on a scope while looking at the transmitter RF output power. In **Fig. 20**, notice the "dips" or the two "peaks" on the top and bottom of each sine wave. This is caused by excessive 2nd harmonic output power. The rule-of-thumb is – if you can see any 2nd harmonic power (a dip or flattening at the peaks), then you are right at or exceeding the –30dBc FCC harmonic specification. A clean sine wave implies FCC compliance, providing the 2nd harmonic is within your scope's bandwidth. Obviously, a spectrum analyzer should be used for accurate harmonic evaluation, but a scope can be used to indicate if you have a problem. If you build your own transmitter, it is your responsibility to ensure it is compliant. An approximate method of estimating 2nd harmonic attenuation is shown.

–30dBc means the 2nd harmonic power is 1/1000th of the fundamental power. For a 5W transmitter, this means the 2nd harmonic power should be 5mW or less, or about 1Vpp. In displaying a 5W signal on a scope (45Vpp), it would be difficult to resolve much less than a 1Vpp dip or ripple.

Fig. 20 –Checking 2nd Harmonic Power

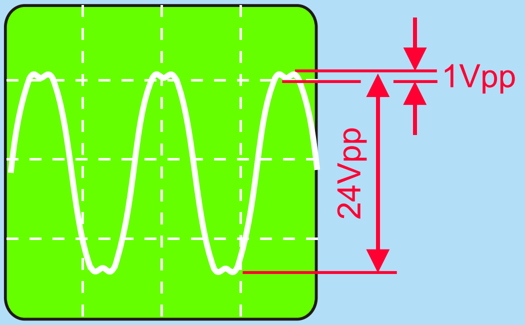

1. Measure output voltage. 24Vpp
2. Measure ripple voltage. 1Vpp
3. Calculate dBc of 2nd harmonic ripple

$$20\log\frac{V2}{V1} = 20\log\frac{1v}{24v} = -27.6\text{ dBc}$$

or, not quite FCC compliant!

Conclusion. This concludes the NA5N "Handiman's Guide to Oscilloscopes." Hopefully, it has helped you to understand your oscilloscope to make simple measurements, and with some practice, perform the more advanced measurements discussed. An oscilloscope is a very powerful tool that is invaluable on the homebrewer's workbench. If there is a particular measurement you feel I omitted, please let me know and I will add it when time permits.

73, Paul Harden, NA5N (na5n@zianet.com or pharden@nrao.edu)

THE HANDYMAN'S GUIDE TO OSCILLOSCOPES (Part 1 of 2)

by Paul Harden, NA5N

Getting Acquainted with your Scope and making some measurements

Updated May 2004 with actual oscope waveforms

Print as .pdf file
5 pages
8½ x 11 or A4

When in the course of human events, it becomes necessary to look at neat signals floating around your radio, you go to a hamfest and buy that $50 o-scope. Now what? This two part article will attempt to explain basically how an oscilloscope works, operator functions, basic measurements, and some advanced applications. An o-scope is a powerful tool in any shack – even a real "cheapie" with limited bandwidth.

HOW AN OSCILLOSCOPE WORKS

A block diagram of a typical o-scope is shown in **Fig. 1**. The test probe usually plugs into the scope via a BNC connector, then passes through a switch to determine whether the input signal will be dc or ac coupled (to remove any dc component). Often this switch will have a "ground" position for setting the zero-volts reference. Next is the input attenuators. The vertical input amplifier is quite sensitive, designed for 20-50mV of input. For larger input voltages, the signal is applied to attenuators comprised of simple voltage dividers. This is the first area of concern for cheap o-scopes, as the input attenuators may not be very linear or accurate. For example, if you appy a 10Vpp signal on the 10v/division setting, the signal should be 1 division high. Switching to 1v/div., the signal should be 10 divisions (usually full-scale) high. If it is not exactly 10 divisions, the attenuator for that setting needs adjusting. Some scopes have internal adjustments for fine-tuning each attenuator setting.

Following the attenuators, the signal is applied to the first vertical amplifier, which converts the input to a differential signal. This differential signal is amplified up to high voltages for the oscilloscope deflecting plates – moving the beam up and down (in the vertical axis).

The sweep generator is usually a constant current source charging a capacitor to make a sawtooth waveform that eventually deflects the beam in the horizontal axis. The frequency of the sawtooth determines how fast the beam travels from the left to the right side of the tube, and is controlled by the sweep control, usually calibrated in seconds, milli-seconds or micro-seconds per division. This is the second area of concern for an oscilloscope – how linear the sawtooth waveform is generated. For example, a sawtooth with a nonlinear ramp will cause the signal displayed in the central portion of the tube to be expanded or compressed compared to the signal at the ends.

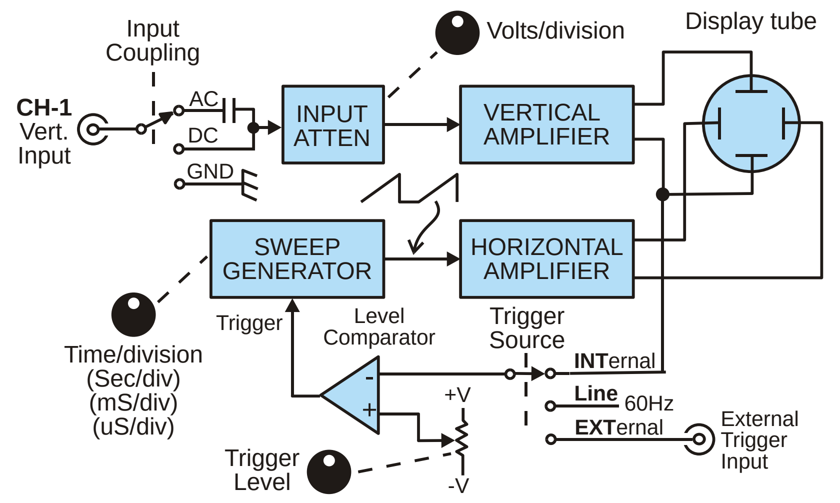

FIG. 1 - Basic block diagram of an Oscilloscpe

The sawtooth ramp is amplified to high voltages, applied to the oscope tube, to deflect the beam from left to right. An important task of an oscope is when the horizontal deflection begins. Normally a switch labeled "Trigger Source" determines what initiates the sawtooth ramp. In the "Internal" position, a sample of the input signal (in the vertical amplifiers) is sampled, with a variable resistor setting the level. When the input signal exceeds the "Trigger Level," a pulse is generated to start the sawtooth ramp and hence the horizontal sweep. The purpose of triggering is to keep the input waveform synchronized to the sweep so it appears stationary on each sweep. The trigger source usually has a "Line" position, which simply triggers the sweep off of 60Hz from the power supply. This synchronizes the sweep to the AC power frequency and is useful for checking television signals, which are synchronized to the power mains. Also, an "External" position may be present, which connects an external input signal (via a BNC connector) to trigger the sweep generator.

Other features your oscope may have are two vertical channels for dual trace operation, various modes to display both waveforms (alternate, chopped, A+B added, etc.), delayed sweep features, dual sweep time bases, built in calibrators, etc.

CALIBRATING YOUR OSCILLOSCOPE

The first thing you should do upon acquiring an o-scope is to check its calibration.

The vertical amplfiers can be checked with a known voltage source or 9v transistor radio battery. Measure the output voltage of the battery with an accurate voltmeter. Let's say it just happens to be +9v exactly. Set the input coupling to ground (0v) and move the trace to the bottom division. Switch the input coupling to DC and set the attenuators to 1v/div. The deflection should be 9 divisions. Switching to 10v/div., deflection should be 0.9 divisions. Internal to the oscope (or perhaps accessible from the outside) are adjustments for the vertical amplifier gain. Adjust this for 9 divisions of deflection in the 1v/div. range. Procedure can be repeated with a 1.5v flashlight battery (assuming you know the exact voltage from a DVM).

The horizontal amplifiers should be checked/calibrated using a signal generator. For example, a 1MHz signal has a period of 1uS. Setting the sweep rate to 1.0uS/div., a 1MHz signal should take exactly 1 division per cycle. Set the horizontal width control properly to ensure the beam starts at the first division and ends at the last division. If the sweep rate appears incorrect, an internal adjustment (Sweep gain or similar) can be set for proper display of the test signal.

The main operator controls are:

- **Intensity** - controls the brightness of the beam. **NOTE: Too bright a beam can damage to the CRT tube!**
- **Focus** - adjusts the beam for the thinnest and sharpest display.
- **VERT & HOR Position** - controls the vertical and horizontal position of the display respectively
- **VERT V/div** - controls the vertical sensitivity of the display, i.e., how many volts (or mV) per division.
- **HOR Sweep Speed** - sets the horizontal sensitivity, i.e., how many mS or uS per division.
- **VERT & HOR vernier** - allows the vertical and horizontal sensitivity settings to be varied in small steps.

Other adjustments you may find on your scope are:

Astigmatism - With the scope intensity and focus properly set, this adjustment compensates for the curvature of the CRT tube by making it in-focus across the sweep. If your trace is out-of-focus in certain areas, but in-focus elsewhere, the astigmatism needs to be adjusted. See ***Fig. 2***.

Trace Rotation - is a small coil around the CRT that skews the trace to ensure it is perfectly horizontal. On scopes without this adjustment, the trace is leveled by physically rotating the CRT to align the trace to the graticle grid. See ***Fig. 2***.

Fig. 2 – Effects of Astigmatism & Trace Rotation

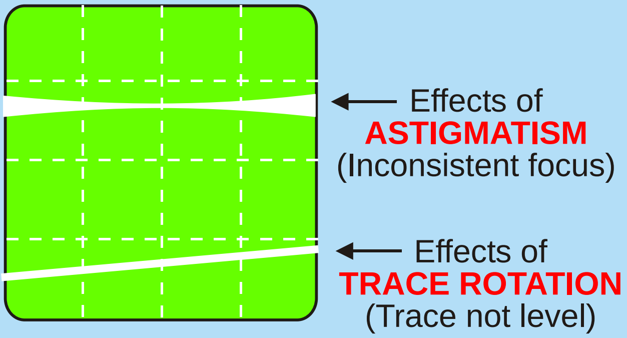

DC BAL (DC Balance) - is a dc offset in the vertical amplifiers that causes a shift in the trace baseline when changing vertical scales. It is most obvious when measuring ac voltages. For example, you are displaying a 10Vpp sine wave, centered on the center graticle, at 2v/div. Changing to 5v/div, the sine wave shifts off the center graticle ... that is, it assumes a dc bias error. The DC BAL is adjusted until the shift no longer occurs when changing vertical scales.

HV ADJ. - is the high voltage that controls the intensity of the trace. Turn up the **Intensity** control to its brightest position, then adjust the HV ADJ for a trace slightly brighter than normal intensity. The **Intensity** control now has the proper range. The HV ADJ might have to be re-adjusted to acquire proper focus.

NOTE: Very bright trace displays can cause permanent damage to the CRT, particularly on a well-used scope.

LET'S MAKE SOME MEASUREMENTS

It is assumed you have your scope relatively calibrated and familiar with the front panel controls. The sample o-scope displays are based on eight vertical and ten horizontal divisions on the CRT screen, typical to most oscilloscopes. Most waveforms are actual displays of the signal cited, photographed from my trusty Tektronix 475 oscilloscope.

First ... a word on TRIGGERING.

Most oscilloscopes have a knob or two for "Triggering." This tells the oscilloscope when to start the sweep. When the **Triggering Slope** is placed in the **(+)** position, the scope will begin its trace when the *input* signal goes positive. Likewise, when **(–)** triggering is selected, the trace will begin when the input signal goes negative, as shown in **Fig. 3**. Often there will be the option to chose the **Triggering Source**, such as "CH.1" or "LINE." Line means the scope is triggered off the 60Hz line voltage, and is useful when synchronizing on television signals or looking at 60Hz power supply noise. CH.1 or CH.2 means the scope will trigger off the signal on channel 1 or 2 respectively. **Trigger Level** is at what voltage of the input signal triggering begins. For example, if set high, triggering may not begin until the input signal reaches several volts. When set around zero, it will trigger the moment the signal goes positive (if set for (+) triggering). This setting can be troublesome if noise exists on the signal. Adjust for stable triggering.

Fig. 3 – Triggering

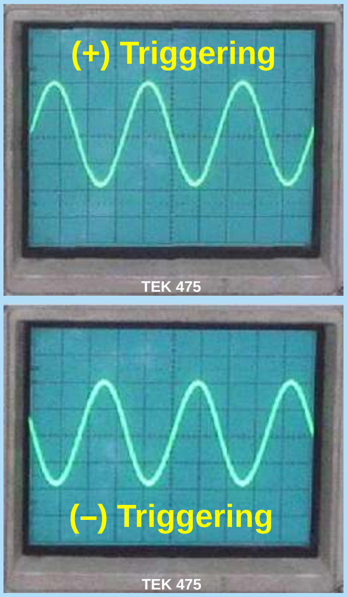

DC Voltages.

Say you want to check the transmit-receiver (T-R) switch in your QRP rig, or other digital signal. See ***Fig. 4.*** The key line is the input to the HCT240 inverter to form the 0v TX– on key- down and the 0v RX– on key- up. This switches the rig between transmit and receive (T-R Switch). It is a logic function, that is, a voltage to represent ON or OFF.

Place the scope lead on pin 13 at 10v/div. and you should see the waveform like the top trace in ***Fig. 4*** ... about +6v on key-up and 0v on key down. Move the scope lead to pin 7 and you should see 0v on key-up and about +8v on key-down (bottom trace). If the output does not go "HI" (+8v) on key-down, or does not go to a solid "LO" (<1v) on key-up, the inverter is not working properly. (It's busticated). Many shortwave receivers use similar schemes for switching filters or attenuators.

While this test could be done with a DVM, the integration time is slow, requiring long key-downs to get the voltages. A scope will also show you how clean the switching is, or if there is an ac voltage (or RF noise) riding on the T-R voltage.

Scopes are thus good dc voltmeters, with about a 5% reading accuracy.

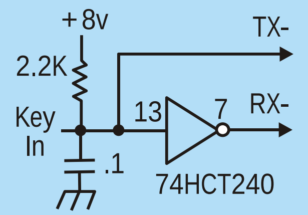

Fig. 4 - 38-Special T-R Switch

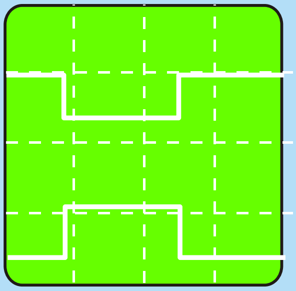

VERT: 10v/div DC
HOR: 500mS/div

AC Voltages.

Here is where an oscilloscope pays for itself by making AC voltage (and frequency) measurements. You must remember, AC voltages are displayed on a scope as ***peak-to-peak voltages***, while a voltmeter measures in ***rms***. RMS voltages are about 1/3 the p-p voltage read on a scope, or specifically:

$$V_{rms} = \tfrac{1}{2}(.707 \times V_{pp}) = 0.354 \times V_{pp}$$

For example, let's measure the output voltage and frequency from the sidetone oscillator in your QRP rig. Place the scope lead on the audio amplifier output. On key-down, you get the waveform shown in ***Fig. 5***. The transmit sidetone audio is 1.9Vpp.

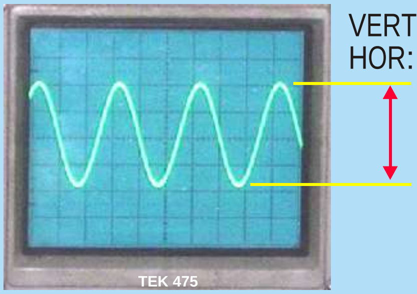

VERT: 0.5v/div ACV
HOR: 0.5mS/div

3.8 divisions peak-to-peak times 0.5v/div
= 1.9 Vpp
= 0.67 Vrms

Fig. 5

AC Frequency Measurement.

With this waveform, we might as well see what frequency our sidetone or transmit-offset frequency is. Most operators prefer the sidetone to be about 700–750Hz. Trigger the scope for a stable waveform and set the time-base (sweep) to display 2 or 3 cycles, as shown in ***Fig. 6***. Center the waveform between two horizontal divisions so zero volts on the waveform is on a graticule line, then move the horizontal position so the first "zero– crossing" is also on a division line.

Fig. 6

VERT: 0.5v/div ACV
HOR: 0.5mS/div

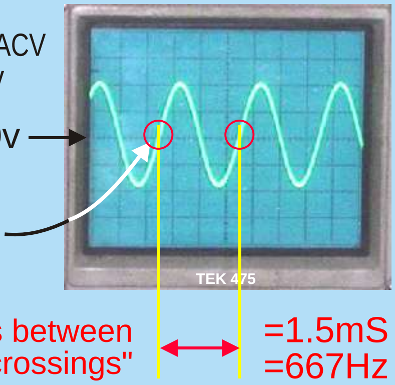

0v

First "Zero Crossing"

3 divisions between "zero-crossings"

=1.5mS
=667Hz

Measure the time it takes to make one complete sine wave from one zero-crossing to the next. In this example, it is 1.5 divisions, at 1mS per division, or 1.5mS. Frequency is simply the reciprocal of time, such that the sidetone frequency is:

$$f = \frac{1}{t} = \frac{1}{1.5\text{mS}} = 667\text{ Hz}$$

For some, this may be about right. For others, this may be a little low to your liking. To raise it to 700Hz, calculate the time period of 700Hz (1/700 = 1.4mS). At 1.0mS/div, you can adjust your sidetone or transmit offset until zero-crossings for a single sinewave is 1.4 divisions. This will be about 700 Hz. (Sidetone may not be adjustable on some rigs).

All frequency measurements are made in this fashion, by measuring the distance between zero-crossings (or from one peak to the next) and converting the time period to frequency. This should emphasize the importance of ensuring your sweep speed is calibrated; as any error in the time base will cause a corresponding error in the accuracy of your time or frequency measurements.

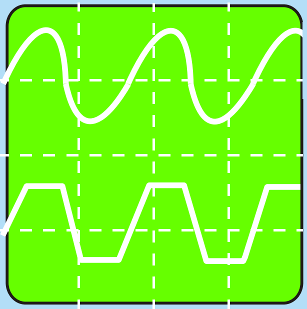

Fig. 7

Example of Waveform Quality

VERT: 2v/div ACV
HOR: 1.0mS/div

Quality of the waveform is another feature of a scope that is unsurpassed since you are "seeing" the waveform in real time. Two examples of waveform quality are shown in ***Fig. 7***.

The top trace shows the sidetone frequency with distortion, perhaps due to improper time-constant on the coupling capacitors or improperly biased audio amplifiers. The bottom trace would be a raspy sounding side tone, due to the amplifier being over-driven and in compression (clipping). The o-scope is an invaluable tool for detecting and diagnosing such impurities in the signal quality.

FIG. 8 – Full Scale Signal Display

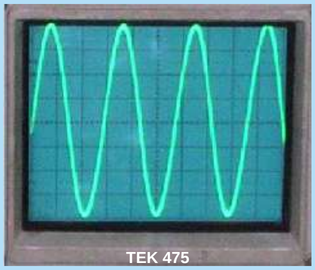

Actual display
Tektronix 475

MORE NIFTY MEASUREMENTS

Amplifier Gain.

The gain of an amplifier can be measured in terms of voltage or decibels (dB). For voltage gain, it is simply Vout/Vin of the amplifier. For example, if the input is 1Vpp and the output is 4Vpp, then the amplifier has a voltage gain of 4.

Gain in dB is often more useful and is how the gains of amplifiers are usually expressed. With dB's, every-time you double the AC voltage, you add 6dB of gain. It is the **ratio** of output to the input, and this **ratio** is easy to measure on a scope.

It is often easier to start with the output. Set the vertical amplifier gain to display the amplifier **output** as a full-scale signal as shown in ***Fig. 8***. Now move the scope probe to the amplifier **input** without disturbing the scope gain. You will of course have a much smaller signal, and the ratio of the input to the output will be the gain in dB. In our example of using eight divisions for full-scale, then four divisions would be 6db, 2 divisions 12dB, etc. as shown in **Fig. 9.** You may want to add your own dB scale along your scope display to remind you of this relationship. Note: this is ***voltage gain*** (Av=20log x Vout/Vin). In this example, with 4Vpp output and 1Vpp input (Av=4), then the gain is dB=20log(4) = 20(0.602) = 12dB, or as shown directly on the CRT tube. Since this is a relative measurement, the absolute Vin or Vout voltage does not need to be determined.

Fig. 9 – Volt. vs. dB relationship

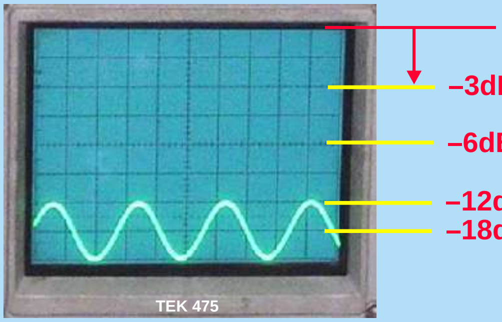

Fig. 10 Measuring Phase Shifts

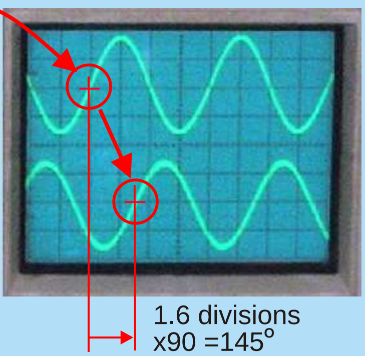

Insertion Loss.

In some circuits, such as filters or attenuators, the ***loss*** in the circuit needs to be measured, and like circuit gain – expressed in dB. The loss through a circuit is called the ***insertion loss.*** It is determined in the same way as amplifier gain just presented, except start with the input (the highest AC voltage) as the full-scale or reference display, then measure the output AC voltage (the lowest level). **The ratio is the insertion loss in dB.**

For example, with a signal generator connected to your receiver, you want to measure the insertion loss through the IF crystal filter. At the filter input, you can just barely squeek out 2 divisions of input signal on your scope at its most sensitive setting. The output from the crystal filter is 1.5 divisions. The insertion loss would be 20log(1.5/2.0 div.) = –2.5 dB. If the output were only 1.0 division (50% reduction), the insertion loss would be 6dB.

Fig. 11 –Phase Shift by imposition

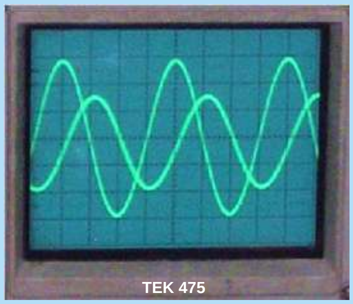

Measuring Phase Shifts.

Phase relationships between two signals at the same frequency can be measured with 2-5° accuracy with a scope, although more suited for a dual-trace scope. The reference signal is applied to CH. 1 and the signal to be measured to CH. 2. For proper phase measurements, ensure your dual trace display is in the **chopped** mode, not **alternate** mode for proper phased referenced triggering.

There are many methods to do this. One is to stretch out the signal so it takes 4 horizontal divisions, such that each division is 90° of phase, as shown in **Fig. 10**. By measuring from a common point on one signal (zero-crossing or from peak-to-peak) to the next, the phase can be measured. For example, say you are making a phased-array antenna in which one feedline must cause a 90° delay. You calculate the electrical length for a ¼λ [L=(246/f) x Velocity Factor] and cut the coax to that length. You are now working on blind faith that you have exactly 90°. With a scope, you can measure it fairly accurately by injecting a signal into one end with a signal generator (at the frequency of interest) and a 50Ω load on the other. Connect the scope CH.1 to the coax (signal) input and CH.2 to the load end and measure the phase. In the **Fig. 10** example, the CH.2 signal is delayed by 1.6 divisions, at 90° /div is 145. ***Your delay line is too long!*** Cut off an inch or two at a time until the CH.2 signal is 90° from CH.1 for precise tuning of the delay line. (While departing from o-scopes for a moment, the sharp null of a phased array is astounding when exactly 90° delay is achieved. More than 10-15° in error causes a very “mushy” null with little difference over a single vertical antenna. Most errors in achieving exactly 90° by the "measure-and-cut" method are due to uncertainties in the stated velocity factor of the coax).

Another method is to superimpose the two signals on top of eachother. Make one signal larger than the other so you know which one is what, as shown in **Fig. 11**. In this example, the smaller signal lags the larger signal by about 100°, estimated by where they cross. For more accurate determination, use the time base to measure the time period of one cycle (T1), then the time period one signal lags (or leads) the other (T2). The phase shift is then $\theta=(T2/T1)x360^{o}$.

Phase measurements can be made on a single trace scope as well. First, connect the reference signal, uisng a BNC "T," to both the *external trigger* and the normal *vertical input*. Adjust the trigger level so the zero-crossing occurs at the beginning of the trace (left-hand graticle). Remove the reference from the vertical input, but not the e*xternal trigger,* and apply the signal to be tested to the vertical input – without altering the time base or trigger level. The distance of zero-crossing of the test signal is from the left-hand graticle can now be measured to determine the phase, though with slightly less accuracy than using a dual-trace scope.

An interesting experiment is to measure the phase shift of the audio signal at different frequencies as it travels through the stages in a CW, SSB or AM active filter. What is the phase shift of the wanted vs unwanted frequencies?

Measuring Rise and Fall Times.

In digital circuits, it is sometimes important to know the rise and fall times of a signal through a gate. In amateur radio transceivers, this same interest could be applied to how fast the T-R switch switches. On key-down, if the transmitter turns on slightly before the receiver is turned off, it can produce an annoying "thump" in the receiver. Rise and fall times are measured by triggering on the edge of the signal of interest, then increase to a faster sweep speed to measure the time it takes the signal to reach 90% of its final level. The signal to be measured is shown in **Fig. 12** on the top trace, and the expanded version on the bottom. For proper rise times, the signal being measured should be well within the bandwidth of your scope and using a low capacity probe.

For example, in **Fig. 12** (bottom trace), the rise time is about 1/4th of a division. If the sweep speed is 100nS/division, the rise time would then be about 25nS.

Fig. 12 –
Rise and Fall Times

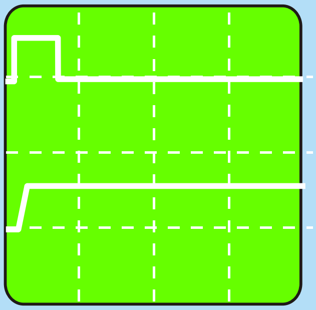

USING LIMITED BANDWIDTH SCOPES

Today's scopes have 200–500MHz bandwidths. Likely your scope is much less than that. A limited bandwidth scope is still very useful to the amateur or homebrewer. Say the bandwidth of your scope is 5MHz. This does not mean you can't see 7MHz signals. It just means the peak-to-peak value has lost meaning, and will likely be very weak, since it is beyond the bandwidth of the scope. (Like other bandwidth measurements in electronics, the “bandwidth” of a scope is usually based on the “3dB bandwidth.” That is, at the maximum bandwidth, you are already at the –3dB point, or a 25% reduction in the peak-to-peak voltage display). You can still resolve individual cycles higher than the cited bandwidth to a certain degree and make gain and phase measurements, since they are based on ***ratios.***

Most of the examples in this article explore many regions of a communications receiver or ham transceiver without the benefit of any great bandwidth. Experiment with your scope to learn its limitations. ***Use a good scope probe and make measurements with a good ground to get the most out of the bandwidth you have.***

For the homebrewer building circuits in the HF bands, a 50 MHz scope **with good calibration** will yield fairly accurate measurements up to 30 MHz with little concern for accuracy. The old 465 or 475 series of Tektronix scopes, with 100/200 MHz bandwidths, make an excellent oscilloscope for the amateur or experimenter. They can often be found at hamfests today for $100–150, and tend to maintain a fairly good calibration almost regardless of how much use they have seen.

In Part 2 - we'll probe (bad pun) into some advanced measurement techniques, even with a simple scope ... such as measuring sideband rejection, filter responses, VCO phase noise, etc. (and what it all means).

72, Paul NA5N

na5n@zianet.com
pharden@nrao.edu

Colophon:
Article prepared by NA5N using CorelDraw 11

START NOW

3 Easy Steps:
1) Click 'Start Now'
2) Download on our website!
3) Get Free File Converter

fromdoctopdf.com

advertisement

NIC NAC TIC TAC CRYSTAL RADIO SET - AN IDEAL PRIMARY SCHOOL PROJECT...

Technology (/technology/) > Science | by mk484 (/member/mk484/) Follow

10,750 83 16

Posted Jul. 4, 2014 | PUBLICDOMAIN

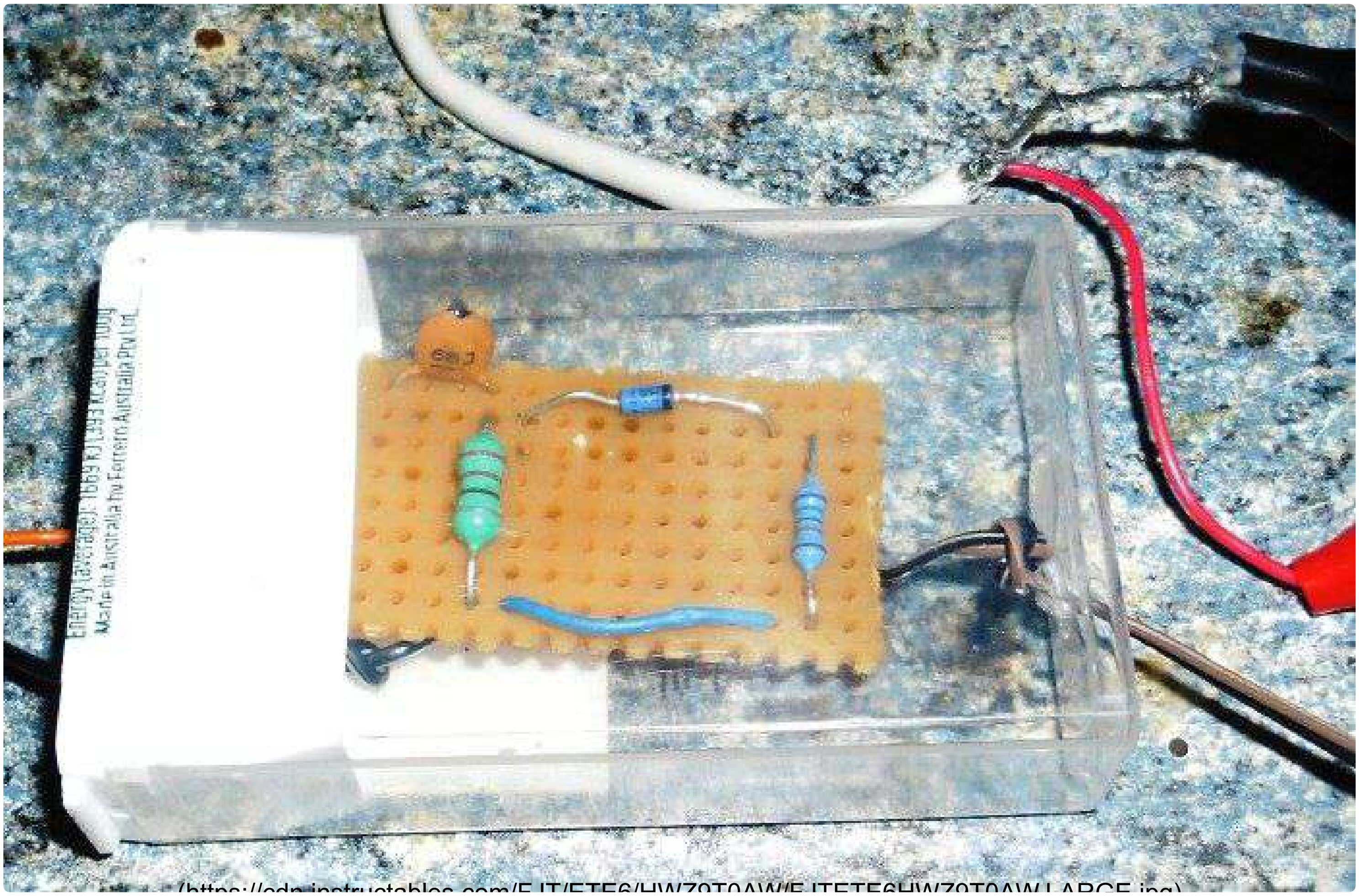
Energy (average): 1669 kJ (393 kcal) per 100g
Made in Australia by Ferrero Australia Pty Ltd

(https://cdn.instructables.com/FJT/FTE6/HWZ9T0AW/FJTFTE6HWZ9T0AW.LARGE.jpg)

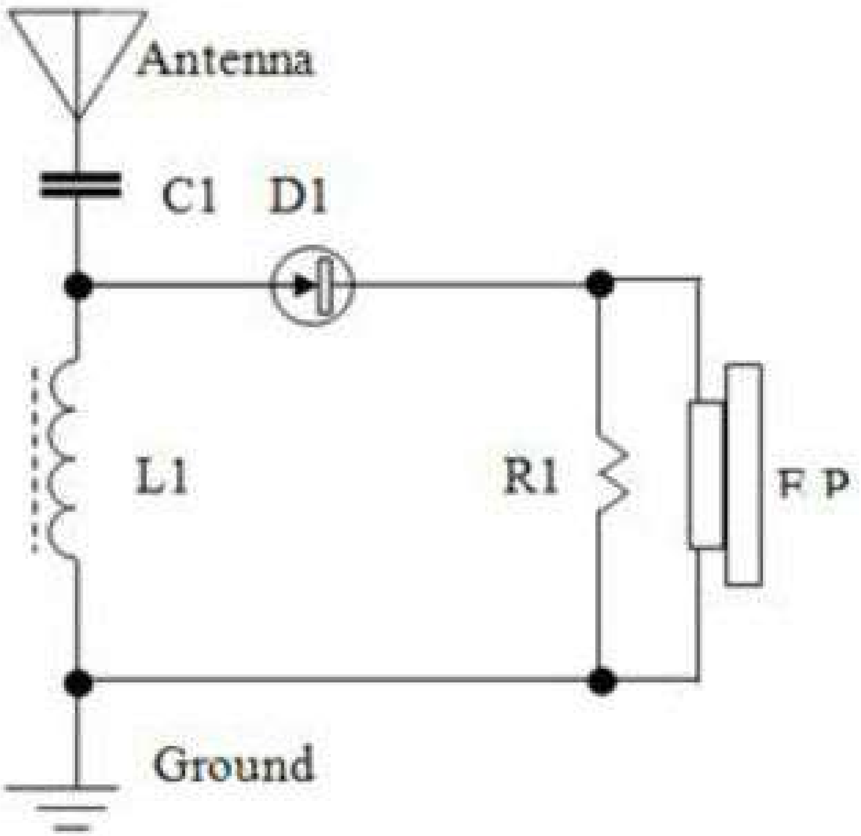

Queensland Historical Atlas

(https://cdn.instructables.com/F8Q/OFKG/HX8HK0L5/F8QOFKGHX8H (https://cdn.instructables.com/F7L/AJMS/

Hi there! This Instructable is all about building a basic crystal radio set that is so simple to build and understand, that a child could do it - with help from mum or dad - or even at school, as a class project. Parts can be bought "off the shelf" at Jaycar Electronics and other suppliers, or purchased online via Ebay and Paypal.

I built my first ever crystal set when I was 9 years old, but that was a different world. We're going to build a 21st century style radio with all modern components which are relatively easy to obtain at your local electronics store or on the Internet.

The radio is contained inside a Tic Tac box, and is called "The Nic Nac Tic Tac Radio". It is built on a square of matrix board, with the component leads poked through the holes in the board, and joined underneath. It is a very simple crystal set, designed to initially receive only one station, but you can expand the tuning range with the addition of only one extra component.

The Tic Tac radio can be constructed by simply twisting the component leads together, underneath the matrix board and attaching antenna/ground wires as well as the earphone wires to the board, using that same technique. You don't have to solder any wires, but there is another way to join them all together - more on that later.

The Tic Tac Radio is a very easy set to put together - you don't need to wind bulky coils or use heavy tools and screw strange looking parts onto a large breadboard. See step 1 for the parts list you will need.

Add Tip
Ask Question

Step 1: Parts and Tools You Will Need

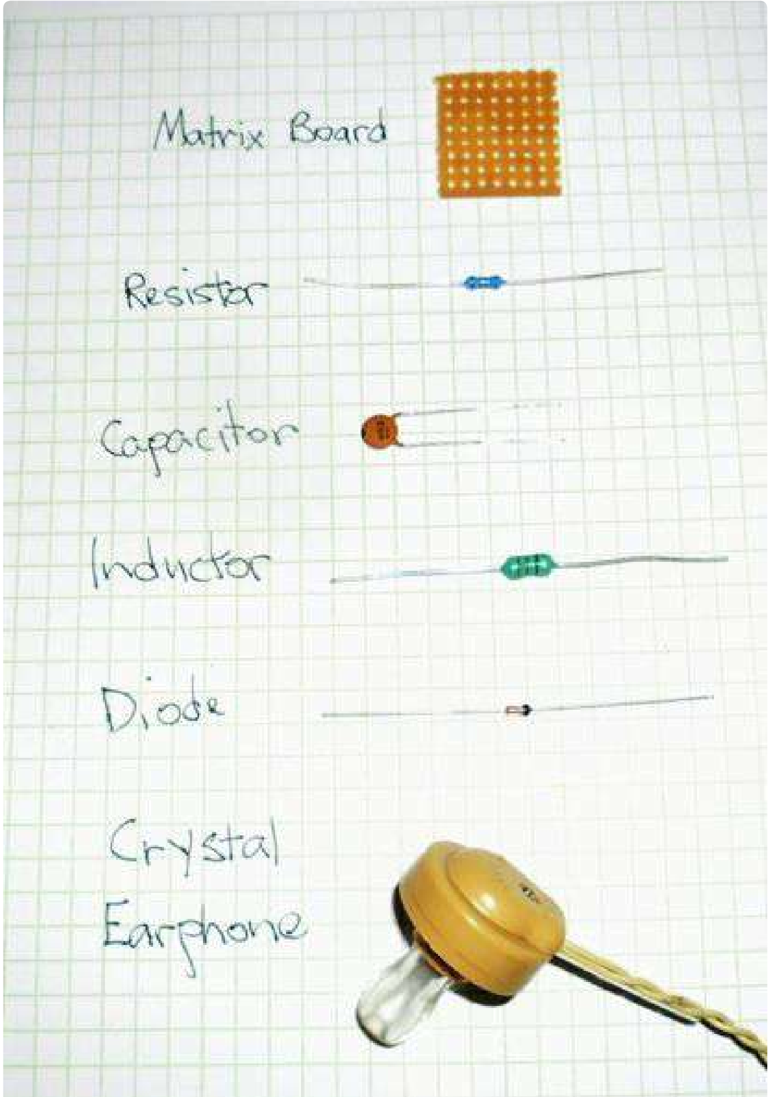
Matrix Board
Resistor
Capacitor
Inductor
Diode
Crystal Earphone

(https://cdn.instructables.com/F9B/WLEJ/HX5H54TY/F9BWL

(https://cdn.instructables.com/F99/H0BC/H

(https://cdn.ins com/FF9/DK

(https://cdn.instructables.com/FXV/5

You will need to acquire the following components from a local electronics shop or an online electronics business.

I am adding the catalogue numbers for Jaycar Electronics, so if you live in Australia or North America, you may be able to visit a Jaycar store and buy the parts over the counter, if not then you may be able to do an online mail order via Paypal:

Parts List

Resistor - 47k - yellow-purple-black-red and brown - RR 0612 (pkt of 8)

Capacitor - 68pF - ceramic x 2 - RC 5322 (pkt of 2) and a 100 or 120 pF value as well for experiments.

Inductor - 220 uH - red-red-brown silver - LF 1538 (resistive type)

Polyvaricon tuning capacitor - 220 pF - RV 5728 - with knob and mounting screws

Diode - BAT46 - ZR 1141 (You can also use a 1N34A Germanium Diode too if you have one at home)

Ceramic Earphone - AS 3305 *

A 25 meter roll of yellow hook up wire for the Antenna wire and a 3 meter length of wire for the Ground wire.

Please note that some Jaycar parts come in multiples of 2 or more per packet. And please note the following:

PLEASE NOTE: THIS PROJECT/KIT CONTAINS SMALL PARTS THAT MAY FORM A CHOKING HAZARD FOR SMALL CHILDREN OR PETS. NOT SUITABLE FOR CHILDREN UNDER FIVE (5) YEARS OLD.

*A normal crystal radio earphone is OK, but if you can't get one of these, or if the one you bought goes dead (as they sometimes do,) you can use a substitute, such as the Murata PKM44EW passive transducer (see picture above) which is available from an old Telstra TF200 touchphone, (the one on the left in the diagram above,) or an equivalent, such as the ARIO transducer, from an old Telstra T1000 pushbutton phone.

The ARIO unit is soldered to the phone's pc board so you'll need to be able to unsolder the three mounting pins underneath the board, or find someone in the neighbourhood who is able.

Take the back off the TF200 (if you've obtained one of these phones,) and you'll see a black disc shaped object 2" round by 1/2' thick - with a red and black wire. Unplug the wires from the circuit board, and unscrew any retaining screws and remove the transducer. Cut the mini plug off, carefully strip the insulation from the ends of the wires and extend them by about 18" with 2 thin lengths of hookup wire. These pietzo devices make good earphones for crystal sets and can be housed in an old pair of ear muffs.

Miscellaneous Materials:

A Tic Tac box (smaller size)

A piece of matrix board at least 7 holes long by 8 holes across. Cut the board to fit neatly inside the Tic Tac box.

A short length of 2 differently coloured wires 60 mm in length and 2 crocodile clips with red and black plastic covers

A length of antenna wire at least 25 metres long and a 3 metre length of a different colour for the Ground wire.

A metal rod or cold water pipe for the ground stake. Be careful which pipes you connect your Ground wire to.

You will need a small sharp object for punching holes in the Tic Tac box.

A small pair of wire cutters for cutting and stripping wires.

Add Tip Ask Question

Step 2: How a Crystal Radio Works

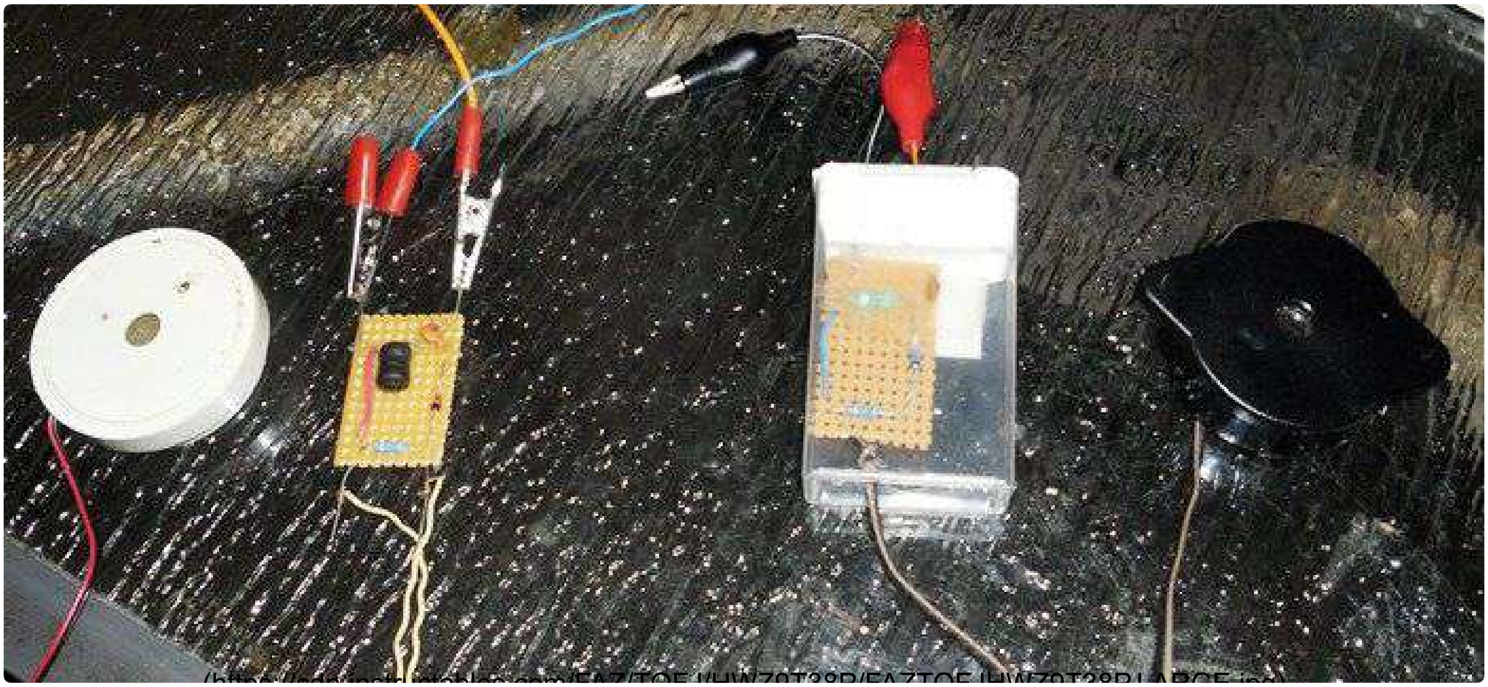

(https://cdn.instructables.com/FAZ/TOEJ/HWZ9T38P/FAZTOEJHWZ9T38P.LARGE.jpg)

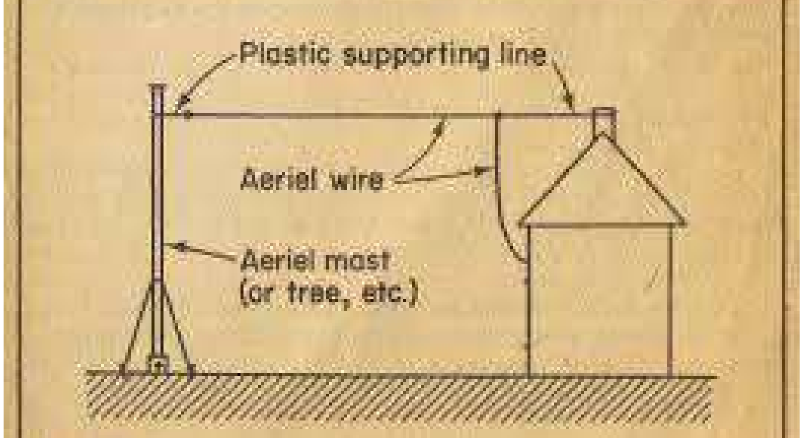

FIG. 2

(https://cdn.instructables.com/FXB/26RU/HX1W6FCZ/FXB26RI

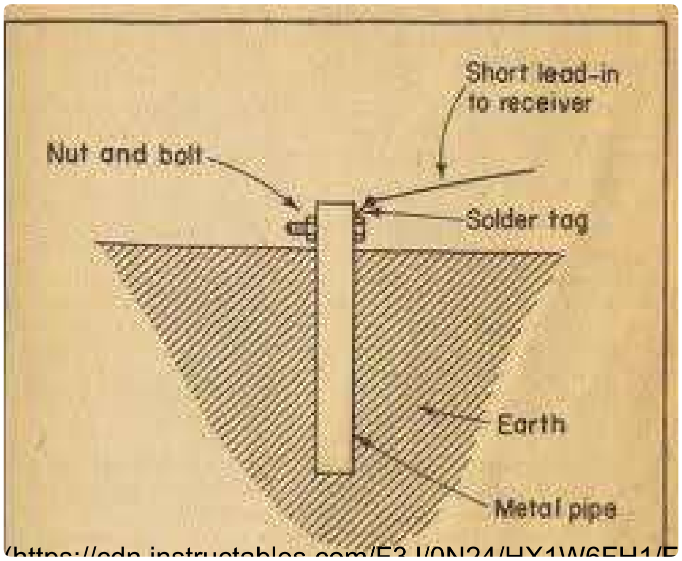

(https://cdn.instructables.com/F3J/0N24/HX1W6FH1/F

Radio signals consists of two parts - the 'carrier wave' which is the AM radio station's frequency of operation, and the 'program signal' which is mixed with the carrier wave for transmission.

Radio waves travel out from the AM transmitter tower through the atmosphere. We want to capture one specific frequency so we can listen to it, so we need the antenna/ground wire system to capture that signal. We also need a 'tuned circuit' that will filter out the desired AM signal, and discard the rest, so that all the other unwanted radio signals pass out through the ground wire to earth.

Two components in our circuit will perform that task for us. The capacitor C1, together with L1 inductor, form a basic 'series tuned' circuit. Their respective values will determine just which local AM radio station we will capture.

We also need a diode to 'detect' the voice and music, so we can hear them in our earphone, which transduces electrical signals into sound waves that we can hear.

In the photo above you can see a completed Tic Tac crystal radio. It is already inside the box. The other set is connected to the antenna/ground wire circuit, undergoing a 'soak test'. It is necessary to do this to ensure that the radio will work once inside the box!

The first diagram shows a typical antenna wire installation. Coming out of a window, the wire is anchored to the building and then over some distance (10 metres +) to a nearby tree or other building.

You must take great care not to erect antenna wires near to power or telephone cables, near your home!

The ground wire comes out the same window and is anchored to a metal pipe/water pipe or metal ground stake, embedded in soft, moist soil.

You must not connect ground wires to electrical mains wiring installations, including wall power outlets!

Another electrical hazard to consider is lightning strikes! Although it is very rare for anyone to be seriously injured or die from a lightning strike it is not impossible. So, if you hear a thunderstorm coming your way (you may hear the lightning 'crashes' in your earphone first,) then disconnect your antenna wire immediately, connect it to the ground wire and put it well up out of the way. Stay well clear of this temporary antenna/ground connection until the storm has completely passed away from your general area - miles away!

Add Tip Ask Question

Step 3: How to Build the Tic Tac Crystal Radio Set

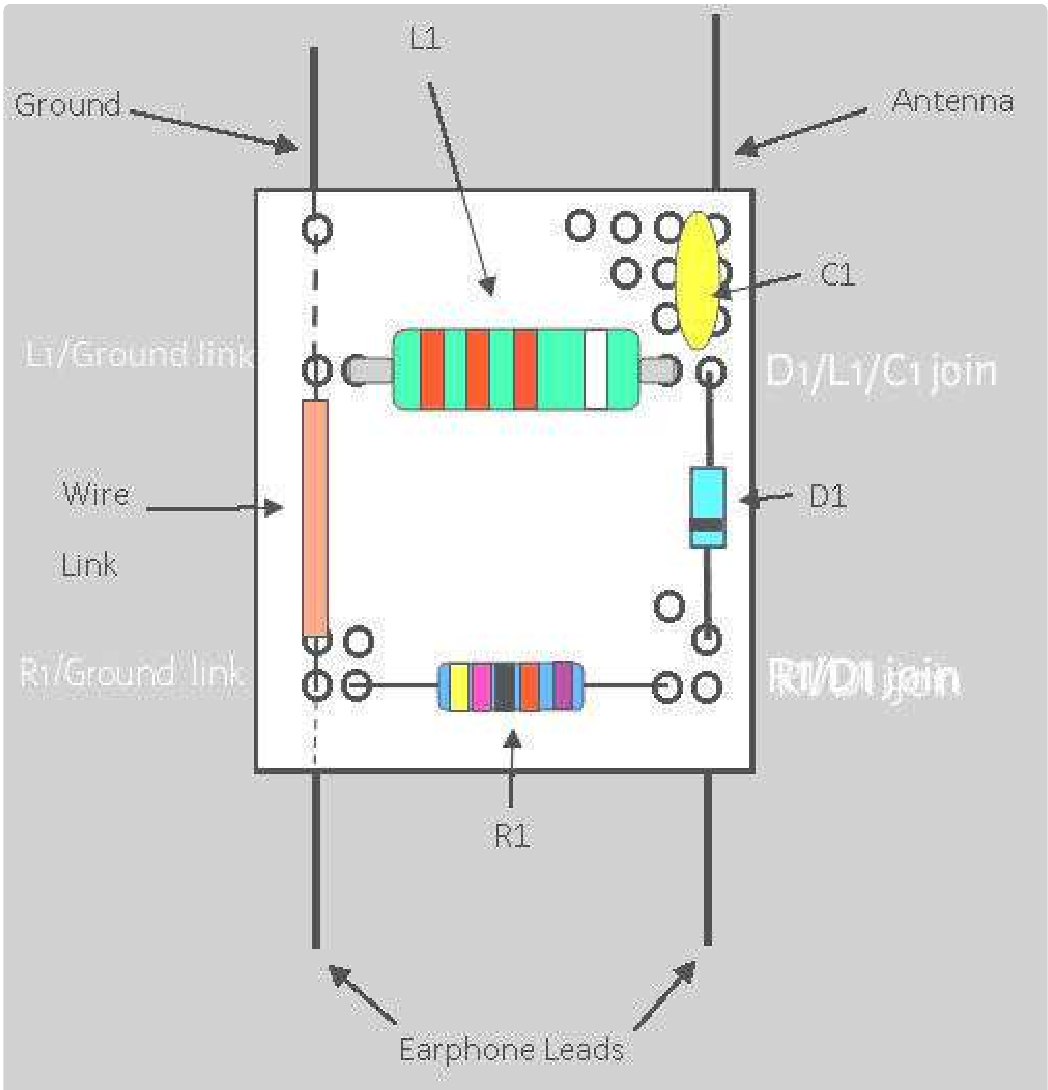
L1
Ground
Antenna
C1
L1/Ground link
D1/L1/C1 join
Wire
Link
D1
R1/Ground link
R1/D1 join
R1
Earphone Leads

(https://cdn.instructables.com/FA5/FW4E/HXA9BXF0/FA5FW4EHXA9BXF0.LARGE

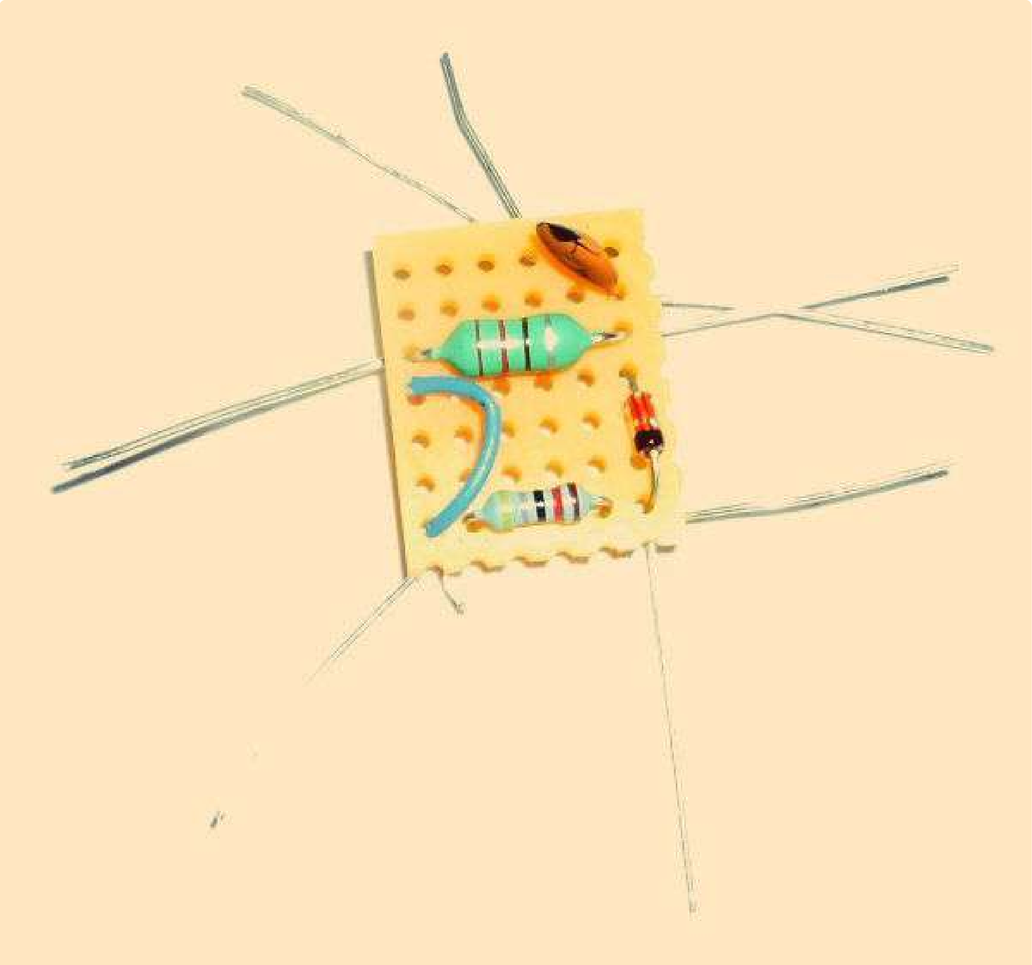
(https://cdn.instructables.com/FS8/N068/HXA9RY3Q/FS8N068HXA9RY3Q.LARGE

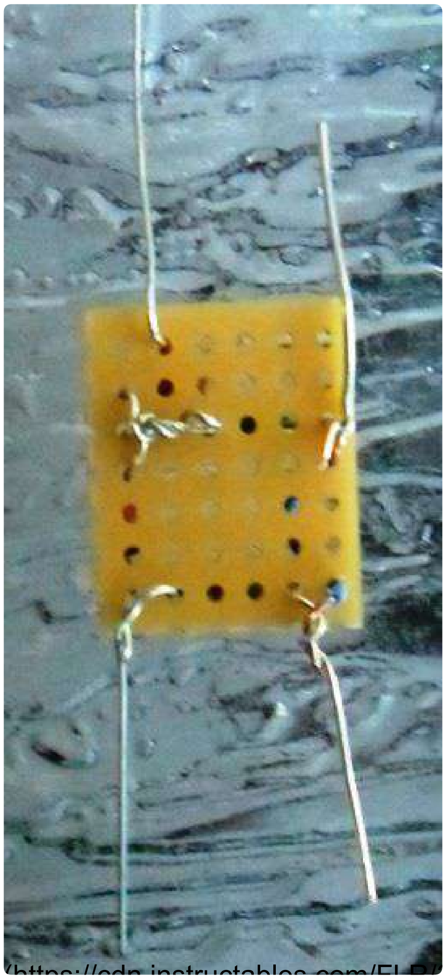
(https://cdn.instructables.com/FLR/

1. Take the components and lay them out on a clean surface.

2. Take the square of matrix board (this is also called 'perf board' because of all the holes or perforations in the material) and beginning with the 220uH (uH = micro-Henry) and place it as shown in the diagram.

3. Then take the diode, capacitor and resistor and place them in the places shown for them, making sure that the coloured band at one end of the diode joins with the resistor wire, as shown

4. Then take each junction where wires come together through their respective holes and gently twist them together, until they form a neat, tight bundle. Take your side cutters and cut off any excess length, taking care not to cut any one wire too short, so that it comes undone from the join.

.5. Take the link wire, strip 2 centimetres of insulation from each end of the wire and install that wire between the free ends of the inductor and resistor, and the matrix board construction is complete.

6. Then take the ceramic earphone, cut the plug off the end, and strip the 2 wire ends about 1.5 centimetres in length. Wrap each earphone wire around each end of the resistor component, underneath the matrix board.

7. Finally, strip the insulation off both ends of the 60mm long differently coloured wires, and attach them to the matrix board - one goes to the inductor/link wire junction and this will be the Ground connection wire. Attach the other one to the free end of capacitor C1 and this will be the antenna wire connection. Both connections are made underneath the matrix board. The Nic Nac Tic Tac Crystal Radio is now ready for testing.

Add Tip Ask Question

Step 4: Final Checks and Installation of Wires

24g NET
FRESHNESS IN 2 CALORIES
tic tac
Strawberry Fields

(https://cdn.instructables.com/FX9/8BCY/HX8HJEU3/FX98BCYHX8HJEU3.LAR

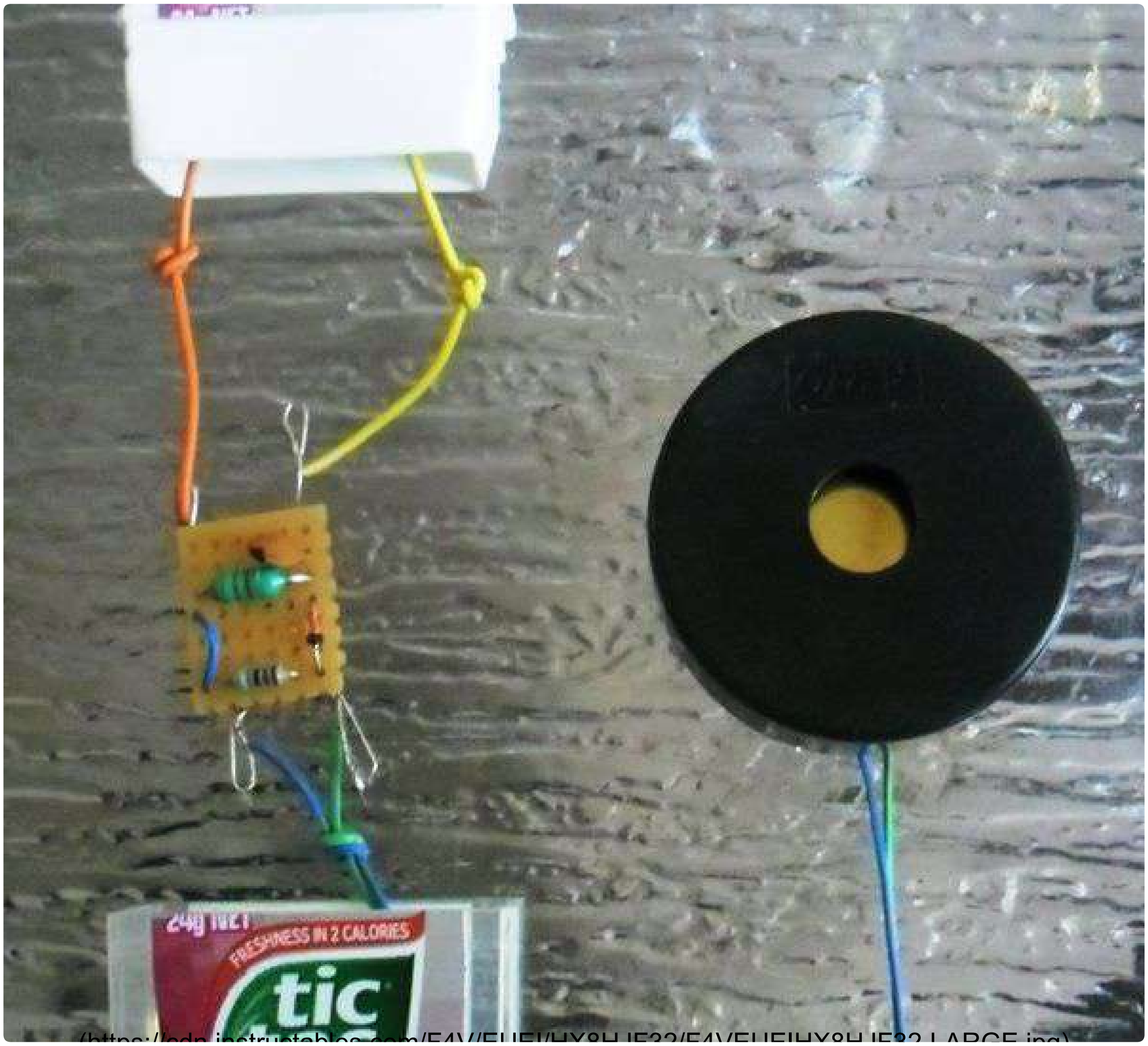

(https://cdn.instructables.com/F4V/FUFI/HX8HJF32/F4VFUFIHX8HJF32.LARGE.jpg)

When you've finished constructing the matrix board circuit and have clipped all the excess component leads off, puncture three (3) holes in the Tic Tac box - 2 small holes on the top lid of the box, about an inch (24mm) apart, so that your antenna and ground wire leads can come through the box lid, and be connected to the matrix board at those 2 points.

Feed the wires through the holes and then tie small knots in each one, near the underside of the box lid, so that they won't pull back out if strained, and disconnect themselves from the matrix board.

Then make one larger hole in the centre of the bottom of the box (clear part) so that your earphone wires can be fed through to the connecting points on that part of the matrix board. Tie a larger knot in the earphone lead so as to prevent it from pulling out of the box if it is strained. Strip the insulation off the earphone leads, and wrap them around the matrix board leads at those 2 points. You are now ready to test out the Nic Nac Tic Tac Crystal Radio.

Add Tip Ask Question

Step 5: Testing Out the Tic Tac Radio

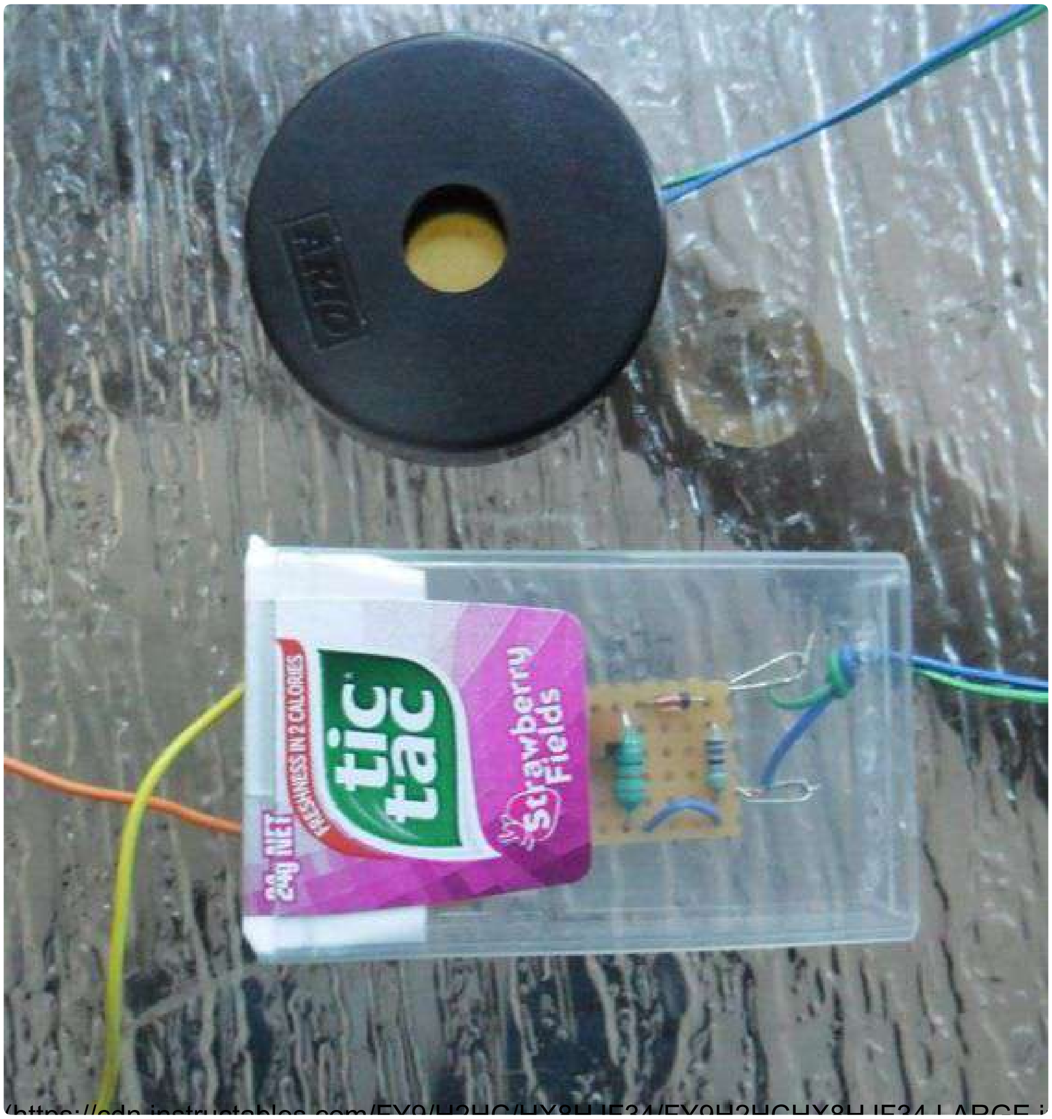
tic tac
Strawberry Fields
FRESHNESS IN 2 CALORIES
24g NET

(https://cdn.instructables.com/FX9/H2HC/HX8HJF34/FX9H2HCHX8HJF34.LARGE.

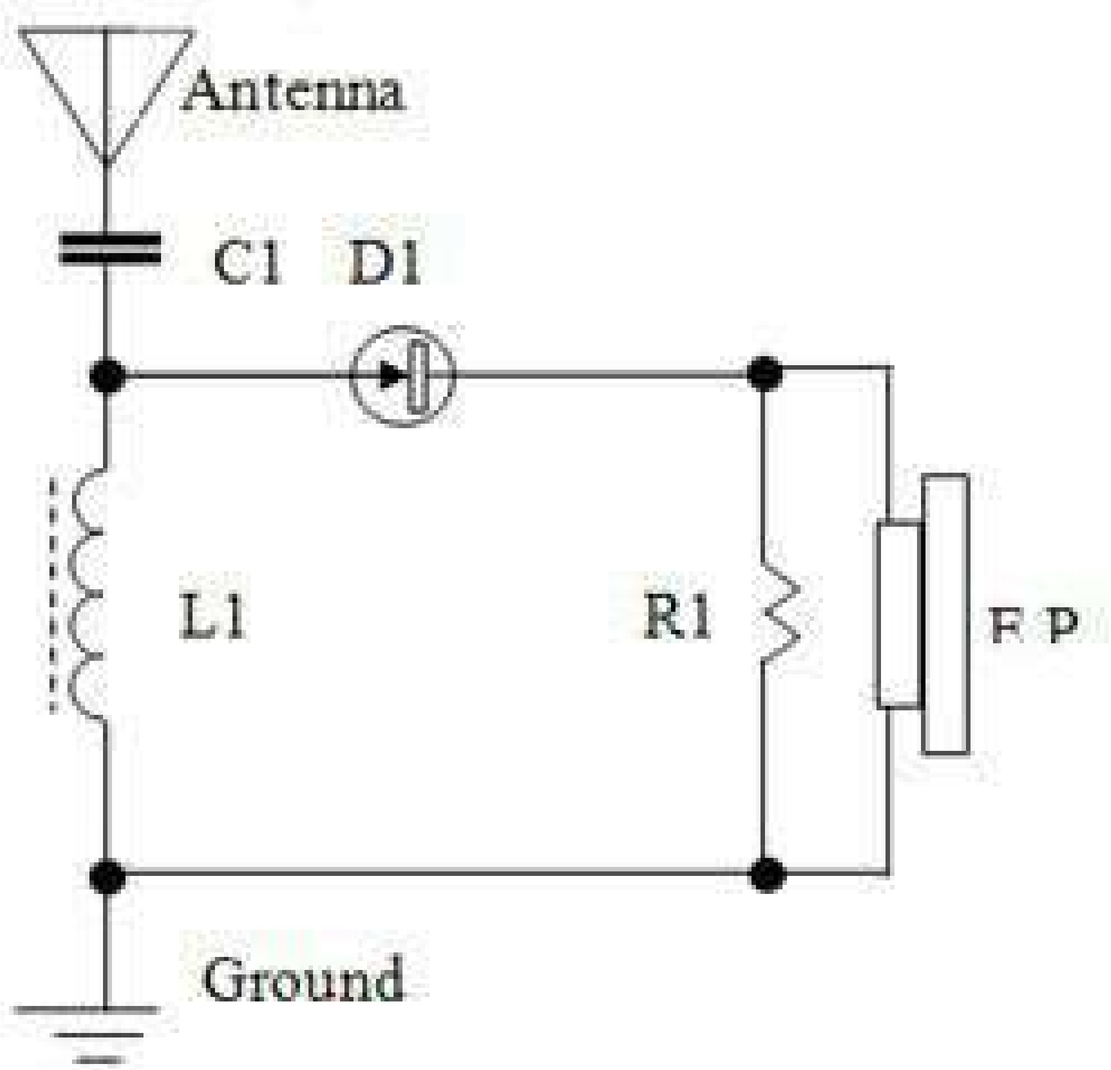

(https://cdn.instructables.com/FL6/41GU/HX5H55A2/FL641GUHX5H55A2.LARGE.jpg)

Take your finished Tic Tac Radio (with the lid part gently pushed back inside the top of the box) and connect your Antenna and Ground wires to the lead outs from the box lid. Place the earphone in/over your ear,and listen carefully for a local AM radio station. This crystal radio is a simple one, and you may have to make one or two adjustments to the components, before you succeed in receiving one or more local AM radio stations, in your area.

If you can't hear anything in the earphone, don't panic. It might just be a simple wiring mistake, which is easily fixed. Go back over all of the steps, making sure that you have the right value components from the electronics store. Make sure that each component is in the right place on the matrix board, (don't confuse the L1 inductor with the R1 resistor - they look a lot like each other!) and that no component wires have come undone from the twisted joins you have just made. I'll be writing up a troubleshooting step soon, so if you run into any problems, post your questions in the 'comments' section down at the bottom of the page, and I'll try and answer them as soon as possible.

Make sure that your antenna and ground wires aren't snagged on anything metal or anything dangerous!

If you have ANY doubts about the electrical safety of your antenna or ground wires, then consult a licensed electrical trades person, who will be able to advise you on electrical safety principles and procedures!

Always remember that electrical safety is your responsibility! If you don't think it is safe to proceed, then don't!

You will need at least 10 to 15 metres of antenna wire, strung between 2 insulating points (not connected to anything metallic, or that gets wet,) at least 2 to 3 metres in height - anything less than this minimum arrangement may mean that you cannot receive any signals at all.

Some places are regarded as 'radio dead spots', so you may need to try an open space, such as a park or a remote corner of a beach. If you do erect antenna and ground wires in public places, hang some streamers or balloons off of them, to alert people to their presence, otherwise people going past may become entangled with them - and get cranky with you!!!

As a final word for now, you'll be happy to know that the completed "Nic Nac Tic Tac Radio" shown in the picture above, picked up local AM radio station 1116 khz 4BC here in Brisbane, with a very clear signal and quite good volume. It works! So be safe kids, have some fun and look forward to more... mk484

Add Tip Ask Question

Step 6: Nic Nac Extras

(https://cdn.instructables.com/FCL/RHOH/HX8HJG9V/FCLRHOHHX8HJG9V.LARGE.jpg)

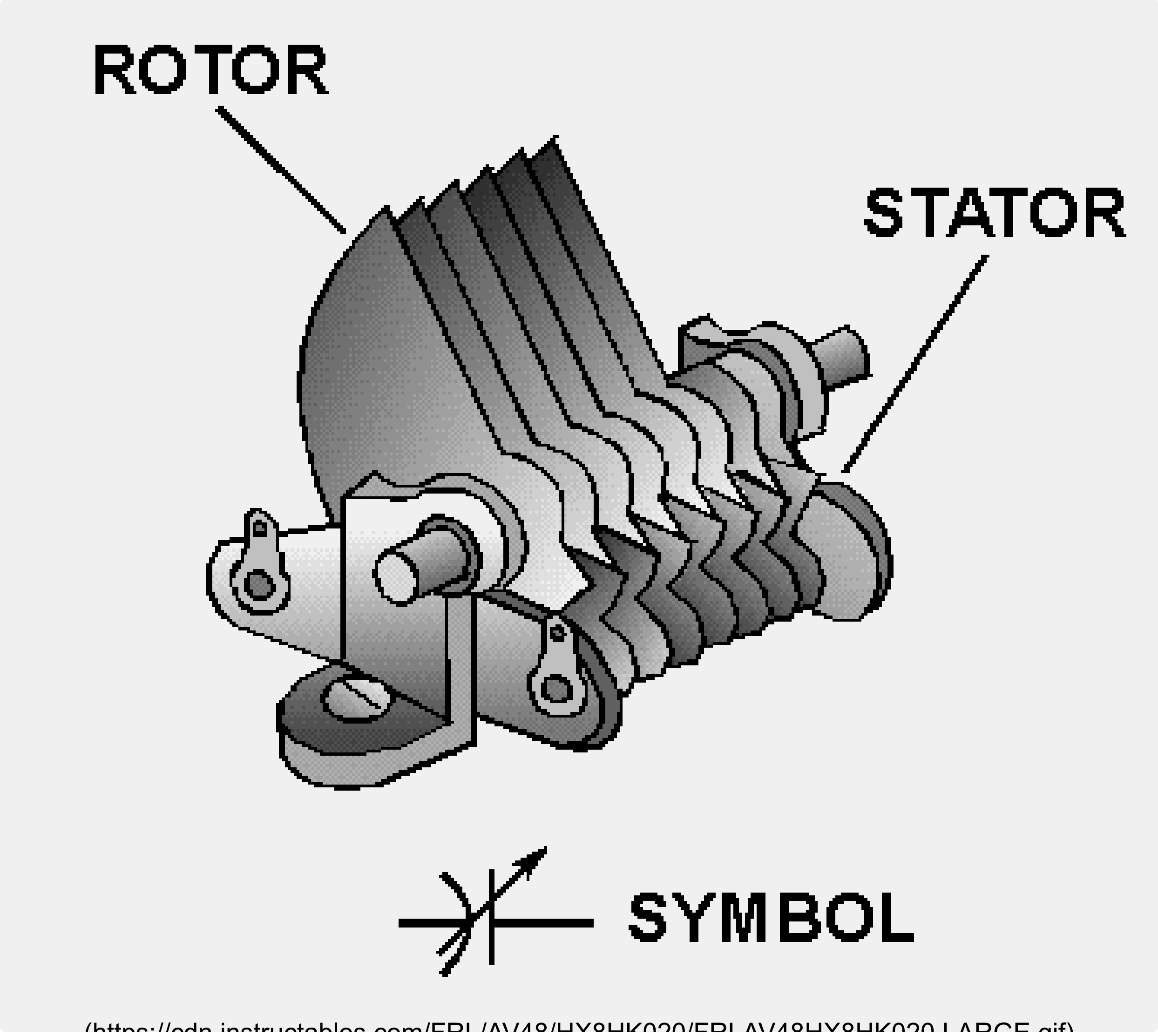
ROTOR
STATOR
SYMBOL

(https://cdn.instructables.com/FBL/AV48/HX8HK020/FBLAV48HX8HK020.LARGE.gif)

If you've built your Tic Tac crystal set and found that you can't receive a local AM radio station yet - don't panic - help is in on the way. You have just built the simplest version of the Tic Tac crystal radio and you may need to add one more component part for it to work properly. This is called a 'polyvaricon' - a miniature tuning capacitor, which can vary the frequency that your radio will receive at. You can see a picture of one up above - the small white box with the black knob next to it. It has 3 connecting tags - the one in the middle ('G') goes to the moving plates and the shaft, while the 2 outer tags ('O' and 'A') go to 2 sets of fixed plates. the smaller set of plates has a value of 60pF - pF is short for 'picoFarad' - a unit of measurement for capacitance. The larger set of plates is valued at 160 pF so that the combined value of the polyvaricon is 220 pF - or 220 picoFarads.

The other picture shows you what happens inside a basic tuning capacitor. there are 2 sets of metal plates - one set is fixed and the other set moves on a rotating shaft, connected to the tuning knob of your radio set. Both sets of plates are mounted on an insulating frame so that they won't 'short out' by touch each other.

The fixed capacitor C1, was chosen to tune somewhere close to the middle of the AM radio band. This band of frequencies starts at 531 kilohertz (Khz) and goes as far as 1701 Khz here in Australia. So we need a combination of coil and capacitor which will tune across all of those frequencies. Our simple Tic Tac radio is known as a 'series tuned' set. If you look at the circuit diagram, you can see electronic symbols for all components in the radio set.If you start at the top with the antenna symbol, you can see the capacitor C1 underneath that, the inductor L1 underneath C1 and then finally, the ground symbol - all wired in series with each other.

The diode D1 detector, the resistor and the earphone plus the link wire, can be considered as one block - the "detector unit". The Antenna wire, ground wire, inductor/coil and capacitor/polyvaricon, can be also be considered as another block - the "tuned circuit". So joining both blocks together, we have the tuned circuit that tunes in only one frequency, passing all other unwanted signals out through the ground wire to earth. This one 'tuned frequency' passes through the diode detector, which strips away the 'carrier wave' and leaves only the 'program signal' (voice, music etc,) behind, which is then fed via the resistor into the earphone. The earphone changes electrical impulses from the diode detector, into sound waves that we can hear. You need the R1 resistor to provide a pathway for the signals going through the diode, out to the ground wire connection. Without this resistor, the signals would sound very distorted and you couldn't hear the program signal very clearly.

Add Tip Ask Question

Step 7: Adding the Polyvaricon Tuning Capacitor

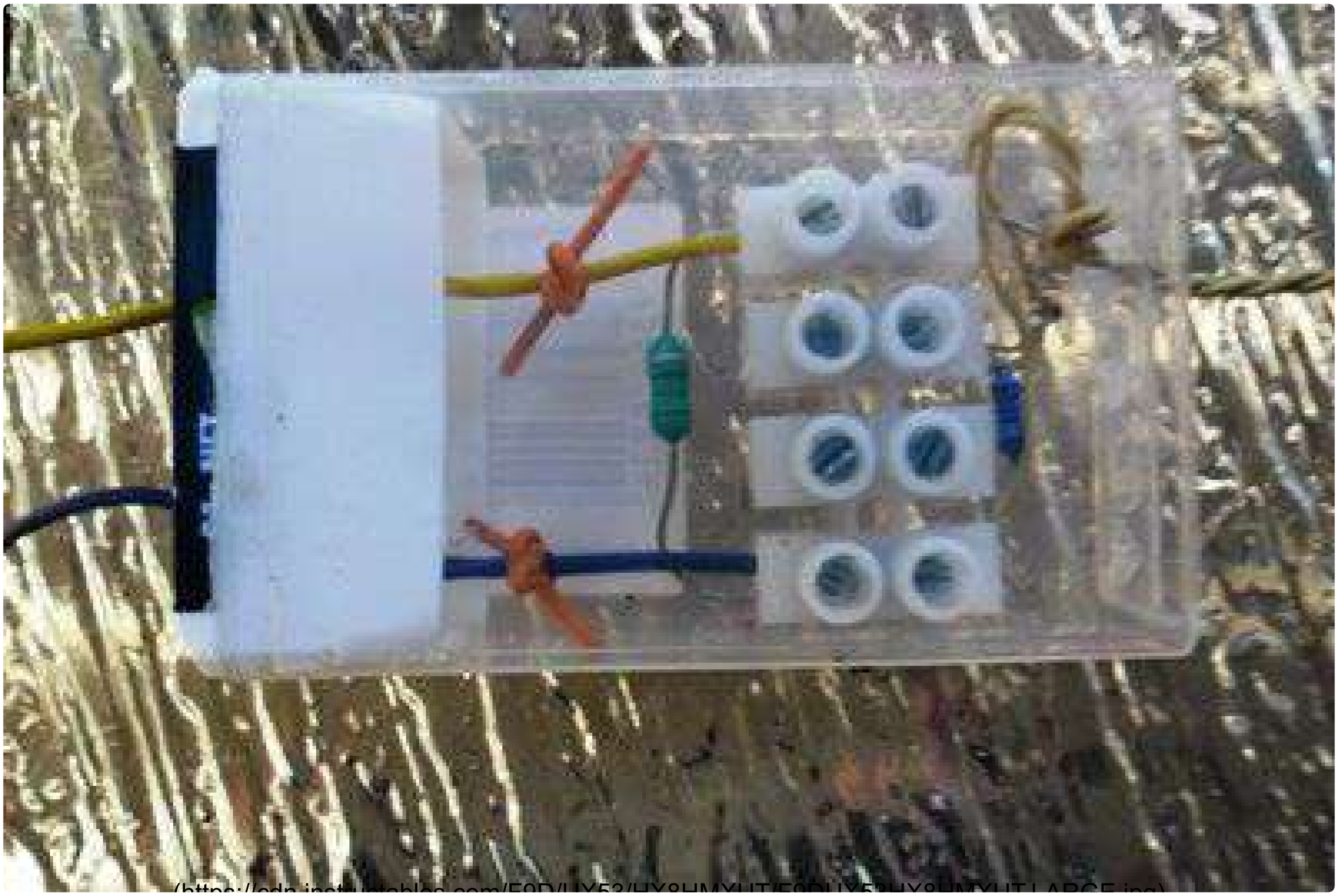

(https://cdn.instructables.com/F9D/UY53/HY8HMYUT/F9DUY53HY8HMYUT.LARGE.jpg)

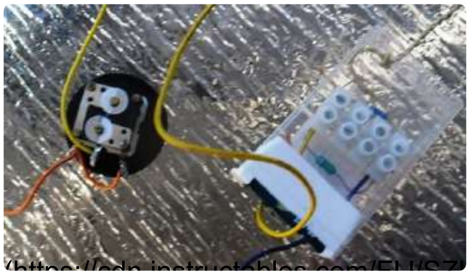

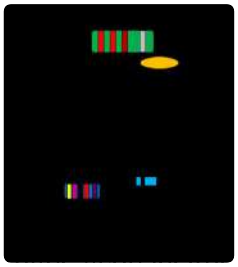

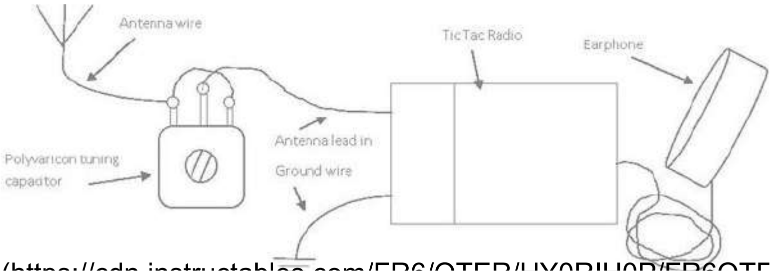

If you want to expand the tuning range of your Tic Tac Radio, than all you have to do is a simple modification (change) to the circuit of your radio set. You can see from the 2 pictures above, that there's an alternative way of building the crystal radio - you can use the matrix board method or you can use a 4 way screw terminal strip.

To use the screw terminal version, cut yourself a 4 way strip of terminals as shown in the diagram and pictures. Undo the screws right out as far as they will go without falling out of the strip. Connect the component leads and wires from the antenna/ground system as well as the earphone wires. Cut off any excess from component leads that you don't need. Wrap thin wires from the antenna/ground system and the earphones, around the thicker component leads before screwing the screws in tight.

Regardless of which method you have used, all we will now wire up the tuning capacitor between the actual Antenna wire and the junction (join) where the diode D1, the fixed capacitor C1 and one end of the inductor meet. Remove the capacitor from the matrix board, and connect the Antenna lead out wire straight to the join of the diode and inductor. Then take the polyvaricon and another piece of wire. Strip both ends of that wire and join the two outer tags (tagged 'O' and 'A' - the centre on is tagged 'G') and then connect you actual Antenna wire to one of the outer tags. Connect the antenna lead out wire, coming out of the box, to the 'G' (middle) tag of the tuning capacitor,so that the antenna wiring now looks like the picture up above.

If you're having problems following the pictures, then refer to the diagrams, which clearly shows all of the connections Make sure your ground wire is connected to the set, and then, listening with your earphone in/over your ear, slowly turn the tuning cap's flat knob, until you hear one or more stations. Congratulations - you now have a "tuneable" Tic Tac crystal set! Happy listening! And don't forget to post in your results, questions, problems etc... mk484 :)

Add Tip

Ask Question

Step 8: Tic Tac One Transistor Amplifier - Use Your Earbuds and Hear Great Sound...

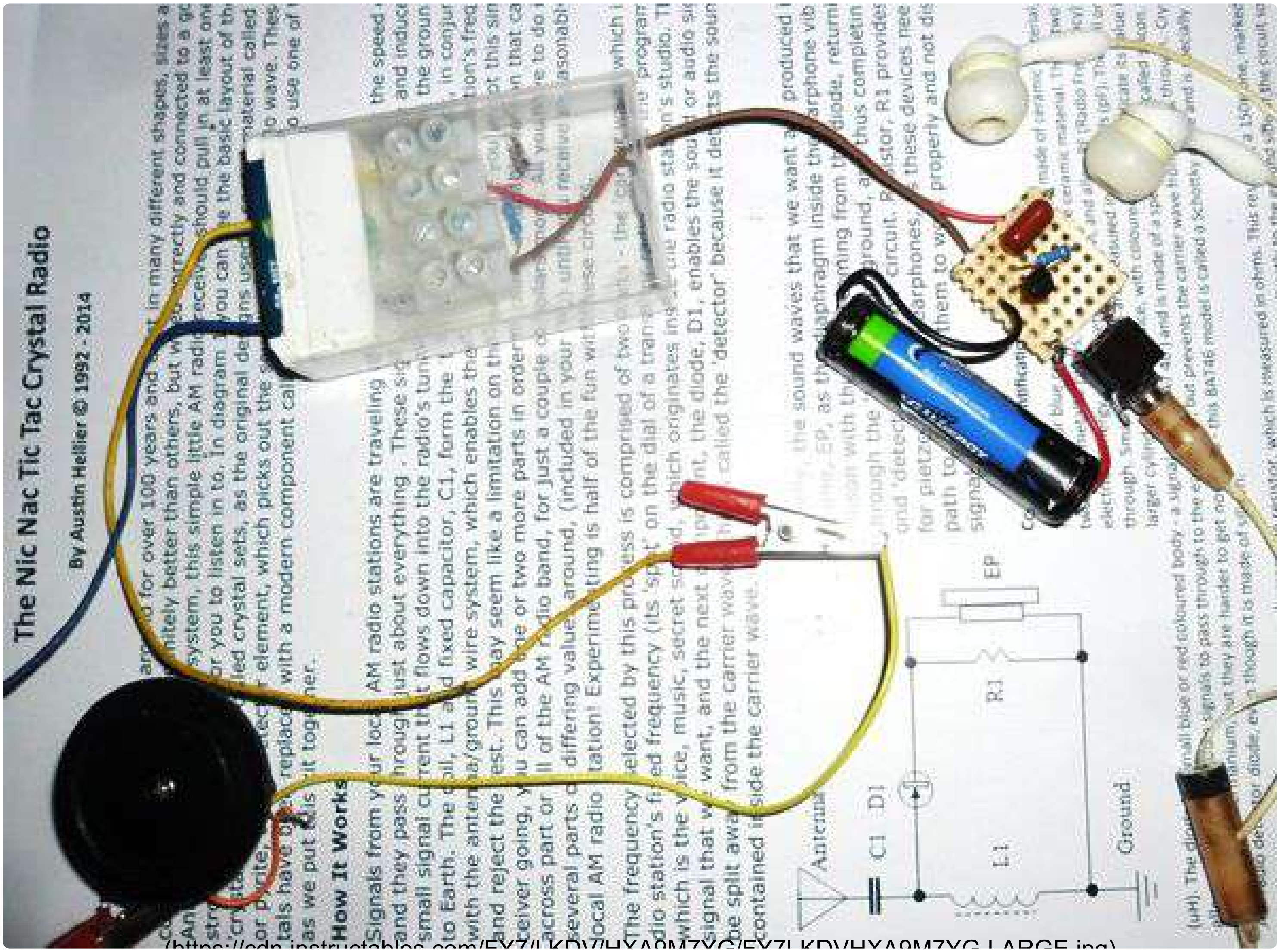
The Nic Nac Tic Tac Crystal Radio
By Austin Hellier © 1992 - 2014
How It Works
EP
R1
D1
C1
L1
Antenna
Ground

(https://cdn.instructables.com/FYZ/LKDV/HYA9MZYC/FYZLKDVHYA9MZYC.LARGE.jpg)

Hi there folks - this isn't really an extra 'Step' as such - it's a sneak preview at a picture of my upcoming Instructable for a one transistor amplifier, which will connect to the Tic Tac Radio - and give you some really good volume - in your Iphone earbuds!

Yes, it's entirely possible nowadays, with modern circuit design, to fit a one transistor amp in such a small space (yes kiddies - it WILL fit inside the smaller Tic Tac box...) and at the same time, get that great sound that comes from those "inside your ear" type earbuds.

This circuit took me about a half hour to complete, uses only 3 electronic components, costing about $1.00, a stereo earphone jack and a 1.5 volt AAA battery and plastic holder costing about another $3 to $4, so you can build the amplifier for about $5 all up - don't forget to shop around and - mums and dads - cheap batteries will be OK for this project, and the battery can simply be replaced without soldering or undoing half a dozen screws....

I can hear my Tic Tac Radio with great volume and clarity - and so will you, so "stay tuned" (ha ha ha) to this series of Instructables, and you will hear great sound too...

mk484

Add Tip Ask Question

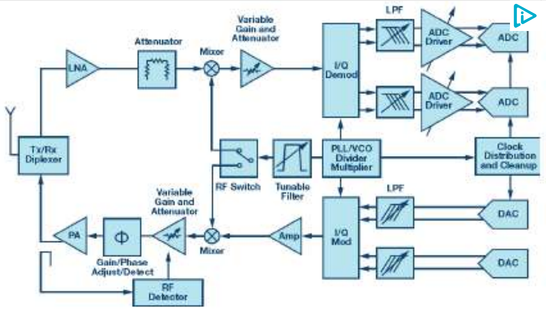
LNA
Attenuator
Mixer
Variable Gain and Attenuator
I/Q Demod
LPF
ADC Driver
ADC
Tx/Rx Diplexer
RF Switch
Tunable Filter
PLL/VCO Divider Multiplier
Clock Distribution and Cleanup
PA
Gain/Phase Adjust/Detect
RF Detector
Amp
I/Q Mod
DAC

Get Complete RFIC Guide

ANALOG DEVICES

Download all Analog Devices RFIC Parts for Each Component in RF Signal Chain.

advertisement

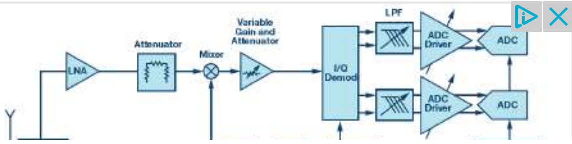

go to: Talking Electronics Website

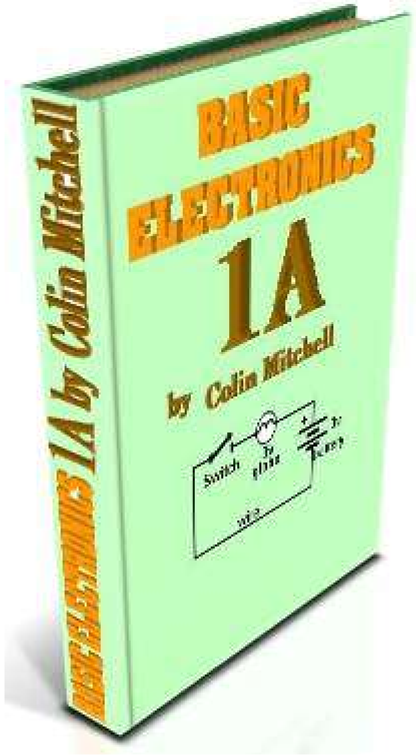
BASIC ELECTRONICS 1A
by Colin Mitchell
BASIC ELECTRONICS 1A by Colin Mitchell
Switch
wire

For any enquiries email Colin Mitchell

THE MULTIMETER

Page 1: Basic Electronics
The capacitor - how it works
The Diode - how the diode works
Circuit Symbols - EVERY Circuit Symbol
Soldering - videos
Page 2: The Transistor
- PNP or NPN Transistor TEST
Page 2a: The 555 IC
The 555 - 1
The 555 - 2
The 555 - 3
The 555 TEST
Page 3: The Power Supply download as .pdf (900kB)
3a: - Constant Current
3b: - Voltage Regulator
3c: - Capacitor-fed Power Supply
Page 4: Digital Electronics
4a: - Gates Touch Switch Gating
Page 5: Oscillators
Page 6: Test - Basic Electronics (50 Questions)
Page 7: The Multimeter - this page

INDEX

Analogue Multimeter
Buying A Multimeter
Current - measuring
Digital Multimeter
Diode - measuring
Pressure Cell
Resistance - measuring
TEST
Using A Multimeter
Voltage - measuring

to Index

TWO MULTIMETERS

There are basically two different types of MULTIMETER. ANALOGUE and DIGITAL
Analogue Multimeters have a NEEDLE or POINTER that moves across a scale.
Digital Multimeters have a numeric display of 3 or more digits. A Digital Multimeter with 3½ digits means the first digit shows only "1."

You really need both types to cover the number of tests needed for designing and repair-work. We will discuss how they work, how to use them and some of the differences between them.

The black (negative lead) ALWAYS stays in the

black hole and the red lead changes to the other red hole to measure 10 amps

to Index

BUYING A MULTIMETER

There are many different types on the market.
The cost is determined by the number of ranges and also the extra features such as diode tester, buzzer (continuity), transistor tester, high DC current and others.
Since most multimeters are reliable and accurate, buy one with the greatest number of ranges at the lowest cost. The cheapest multimeters are on **eBay**.
This article explains the difference between an analogue meter and a digital meter.

Multimeters are sometimes called a "meter", a "VOM" (Volts-Ohms-Milliamps or Volt Ohm Meter) or "multi-tester" or even "a tester" - they are all the same.

One term used to describe a **DIGITAL MULTIMETER** is **3½ digits**.
This is the number of digits on the display. The first digit is usually made from two pixels and can only produce "1." This is called a half-digit. The other digits are full digits. The cheapest digital multimeters have 3½ digits. This will produce a reading of 1999 and the decimal point can produce values from 1.999 to 19.99 to 199.9 to 1999.
Another term is **DISPLAY COUNTS**. This is connected with the accuracy of the display, but since digital meters are accurate to 1% or less and we are using resistors with an accuracy of 5%, even a $10.00 digital meter will be perfect.

to Index

USING A MULTIMETER

Analogue and **digital** multimeters have either a rotary selector switch or push buttons to select the appropriate function and range. Some Digital Multimeters (DMMs) are auto ranging; they automatically select the correct range of voltage, resistance, or current when doing a test. However you need to select the function.

Before making any measurement you need to know what you are checking. If you are measuring voltage, select the AC range (10v, 50v, 250v, or 1000v) or DC range (0.5v, 2.5v, 10v, 50v, 250v, or 1000v). If you are measuring resistance, select the Ohms range (x1, x10, x100, x1k, x10k). If you are measuring current, select the appropriate current range DCmA 0.5mA, 50mA, 500mA, 10A. Every multimeter is different however the photo below shows a low cost Analogue multimeter with the basic ranges.

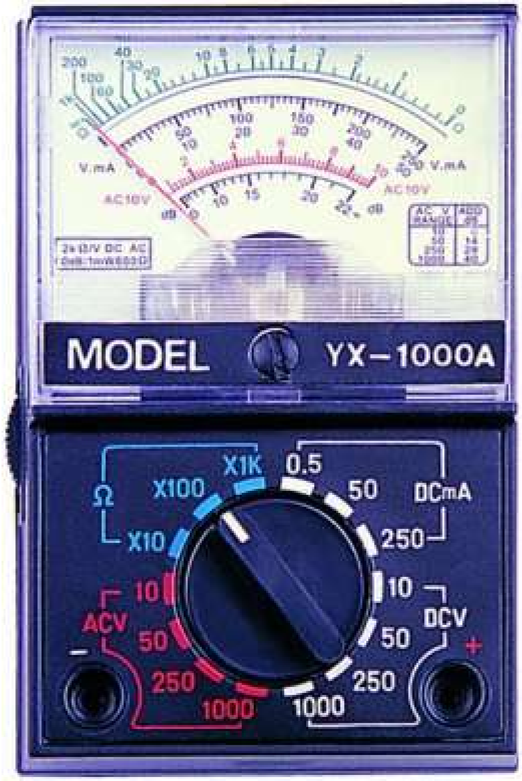
MODEL YX-1000A
Ω
X10
X100
X1K
0.5
50
DCmA
250
10
DCV
50
250
1000
ACV
10
50
250
1000
V.mA
AC10V
dB

An ANALOGUE MULTIMETER

The most important point to remember is this:
You must select a voltage or current range that is bigger or HIGHER than the maximum expected value, so the needle does not swing across the scale and hit the "end stop."
If you are using a DMM (Digital Multi Meter), the meter will indicate if the voltage or current is higher than the selected scale, by showing "OL" - this means "Overload." If you are measuring resistance such as 1M on the x10 range the "OL" means "Open Loop" and you will need to change the range. Some meters show "1' on the display when the measurement is higher than the display will indicate and some flash a set of digits to show over-voltage or over-current. A "-1" indicates the leads should be reversed for a "positive reading."
If it is an AUTO RANGING meter, it will automatically produce a reading, otherwise the selector switch must be changed to another range.

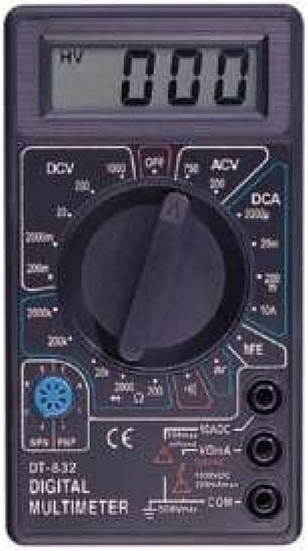
HV
000
DCV
OFF
ACV
DCA
hFE
10A
NPN PNP
CE
DT-832
DIGITAL
MULTIMETER
10ADC
VΩmA
COM

A typical DIGITAL Multimeter

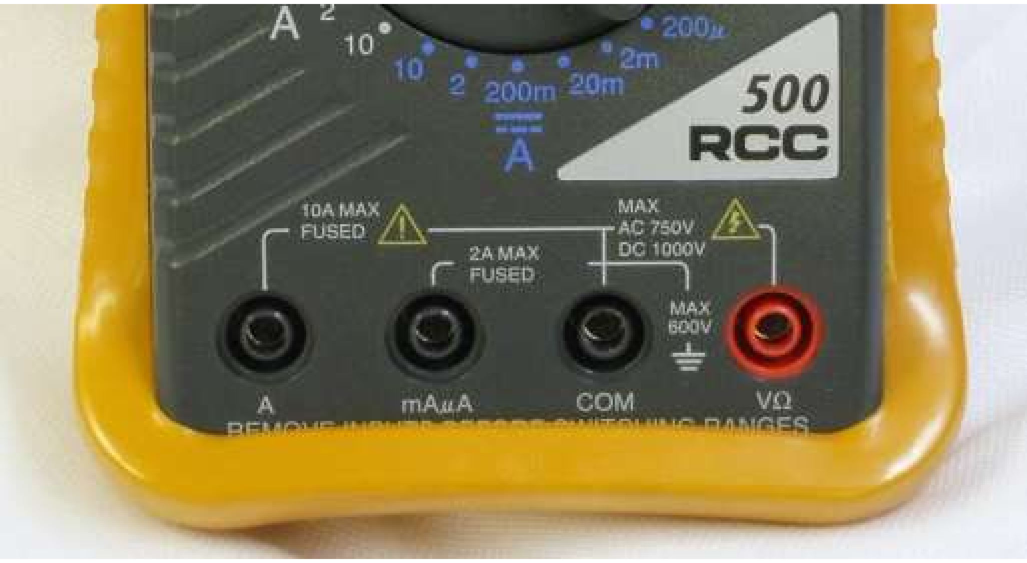
A
10
10
2
200m
20m
2m
200µ
A
500
RCC
10A MAX
FUSED
2A MAX
FUSED
MAX
AC 750V
DC 1000V
MAX
600V
A
mAµA
COM
VΩ

The Common (negative) lead ALWAYS fits into the "COM" socket. The red lead fits into the red socket for Voltage and Resistance. Place the red lead (red banana plug) into "A" (for HIGH CURRENT "Amps") or mA,uA for LOW CURRENT.

The black "test lead" plugs into the socket marked "-" "Common", or "Com," and the red "test lead" plugs into the meter socket marked "+" or "V-W-mA."   The third banana socket measures HIGH CURRENT and the positive (red lead) plugs into this. You DO NOT move the negative "-" lead at any time.
The following two photos show the test leads fitted to a digital meter. The probes and plugs have "guards" surrounding the probe tips and also the plugs so you can measure high voltages without getting near the voltage-source.

FLUKE
600.0
OFF
COM
A
10ADC

Analogue meters have an "Ohms Adjustment" to allow for the change in voltage of the battery inside the meter (as it gets old).

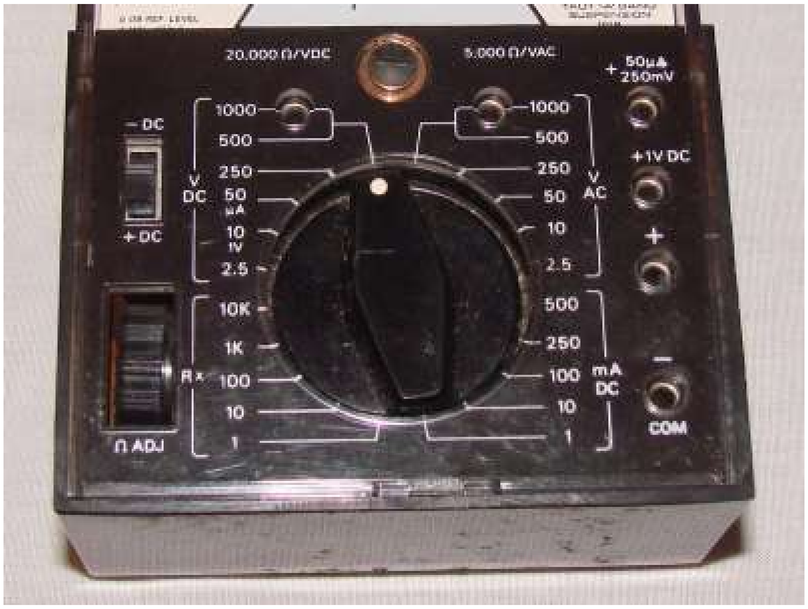

"Ohms Adjust" is also called "ZERO SET"
The sensitivity of this meter is 20,000ohms/volt on the DC ranges and 5k/v on the AC ranges

Before taking a resistance reading (each time, for any of the Ohms scales) you need to "ZERO SET" the scale, by touching the two probes together and adjust the pot until the needle reads "0" (swings FULL SCALE). If the pointer does not reach full scale, the batteries need replacing. Digital multimeters do not need "zero adjustment."

to Index

ANALOGUE Vs DIGITAL

You cannot say one meter is better than the other because BOTH have advantages and disadvantages.
An analogue multimeter is the "old style" and it puts a load on a circuit and this may change the reading to give an incorrect readout, but it has the advantage of the needle moving across the scale fairly quickly so you can sometimes see if the voltage is fluctuating.
It also gives a more-accurate result in some high frequency circuits as it does not pick up stray fields and produce a false reading.
Digital meters put almost no load on a circuit and produce accurate readings from both low-impedance and high-impedance circuits.
Digital meters can display very low resistances.
You must remember to turn a Digital meter OFF to prevent the battery going flat.

If you are testing a circuit containing a high-frequency oscillator, use BOTH an ANALOGUE and DIGITAL meter to check the reading. Sometimes the leads of a Digital multimeter will pick up signals and create a false reading.
Sometimes you will get a voltage reading with a Digital multimeter due to a high resistance leak and a zero reading with an Analogue meter. This is why you need **BOTH** meters.

to Index

MEASURING VOLTAGE

Most of the readings taken with a multimeter will be VOLTAGE readings.
Before taking a reading, you should select the highest range and if the needle does not move up scale (to the right), you can select another range.
Always switch to the highest range before probing a circuit and keep your fingers away from the component being tested.
If the meter is Digital, select the highest range or use the auto-ranging feature, by selecting "V." The meter will automatically produce a result, even if the voltage is

AC or DC.
If the meter is not auto-ranging, you will have to select V⎓ if the voltage is from a DC source or V~ if the voltage is from an AC source. DC means Direct Current (but this does not mean you select the CURRENT range - you are taking a voltage reading that is not rising and falling. That's why we say it is **DC** and do not say the words "direct current"). The voltage is coming from a battery or supply where it is steady and not "rising and falling."
You can measure the voltage at different points in a circuit by connecting the black probe to chassis. This is the 0v reference and is commonly called "Chassis" or "Earth" or "Ground" or "0v."
The red lead is called the "measuring lead" or "measuring probe" and it can measure voltages at any point in a circuit. Sometimes there are "test points" on a circuit and these are wires or loops designed to hold the tip of the red probe (or a red probe fitted with a mini clip).
You can also measure voltages ACROSS A COMPONENT. In other words, the reading is taken in PARALLEL with the component. It may be the voltage across a transistor, resistor, capacitor, diode or coil. In most cases this voltage will be less than the supply voltage.
If you are measuring the voltage in a circuit that has a HIGH IMPEDANCE, the reading will be inaccurate, up to 90% !!!, if you use a cheap analogue meter.

Here's a simple case.

The circuit below consists of two 1M resistors in series. The voltage at the mid point will be 5v when nothing is connected to the mid point. But if we use a cheap analogue multimeter set to 10v, the resistance of the meter will be about 100k, if the meter has a sensitivity of 10k/v and the reading will be incorrect.
Here how it works:
Every meter has a sensitivity. The sensitivity of the meter is the sensitivity of the movement and is the amount of current required to deflect the needle FULL SCALE. This current is very small, normally 1/10th of a milliamp and corresponds to a sensitivity of 10k/volt (or 1/30th mA, for a sensitivity of 30k/v).
If an analogue meter is set to 10v, the internal resistance of the meter will be 100k for a 10k/v movement.
If this multimeter is used to test the following circuit, the reading will be inaccurate.
The reading should be 5v as show in diagram **A**.
But the analogue multimeter has an internal resistance of 100k and it creates a circuit shown in **C**.
The top 1M and 100k from the meter create a combined PARALLEL resistance of 90k. This forms a series circuit with the lower 1M and the meter will read less than 1v
If we measure the voltage across the lower 1M, the 100k meter will form a value of resistance with the lower 1M and it will read less than 1v
If the multimeter is 30k/v, the readings will be 2v. See how easy it is to get a totally inaccurate reading.

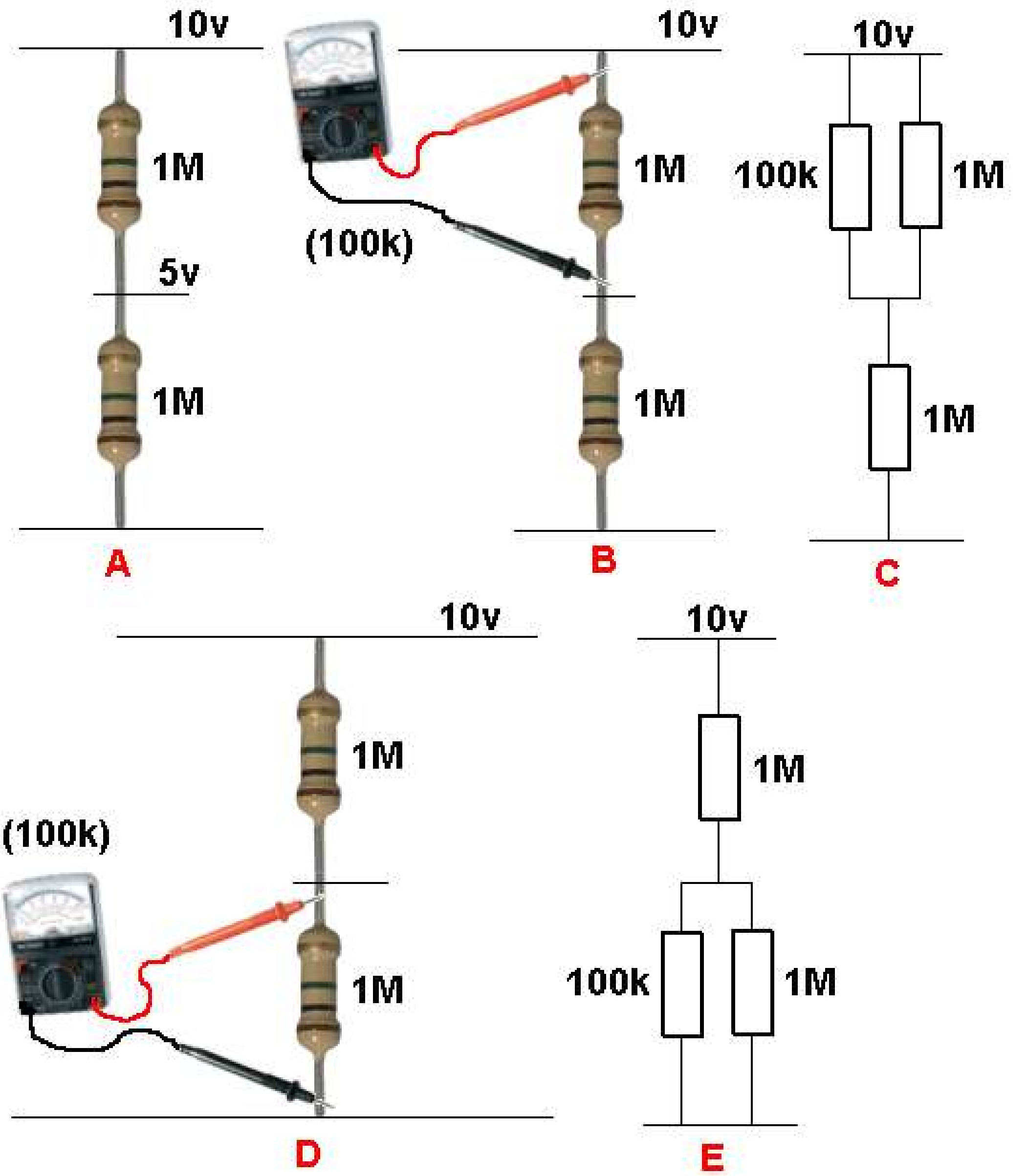

If the reading is taken with a Digital Meter, it will be more accurate as a DMM does not take any current from the circuit (to activate the meter). In other words it has a very HIGH input impedance. Most Digital Multimeters have a fixed input resistance (impedance) of 10M - no matter what scale is selected. That's the reason for choosing a DMM for high impedance circuits. It also gives a reading that is accurate to about 1%.

to Index

MEASURING VOLTAGES in a CIRCUIT

You can take many voltage-measurements in a circuit. You can measure "across" a component, or between any point in a circuit and either the positive rail or earth rail (0v rail). In the following circuit, the 5 most important voltage-measurements are shown. Voltage "A" is across the electret microphone. It should be between 20mV and 500mV. Voltage "B" should be about 0.6v. Voltage "C" should be about half-rail voltage. This allows the transistor to amplify both the positive and negative parts of the waveform. Voltage "D" should be about 1-3v. Voltage "E" should be the battery voltage of 12v.

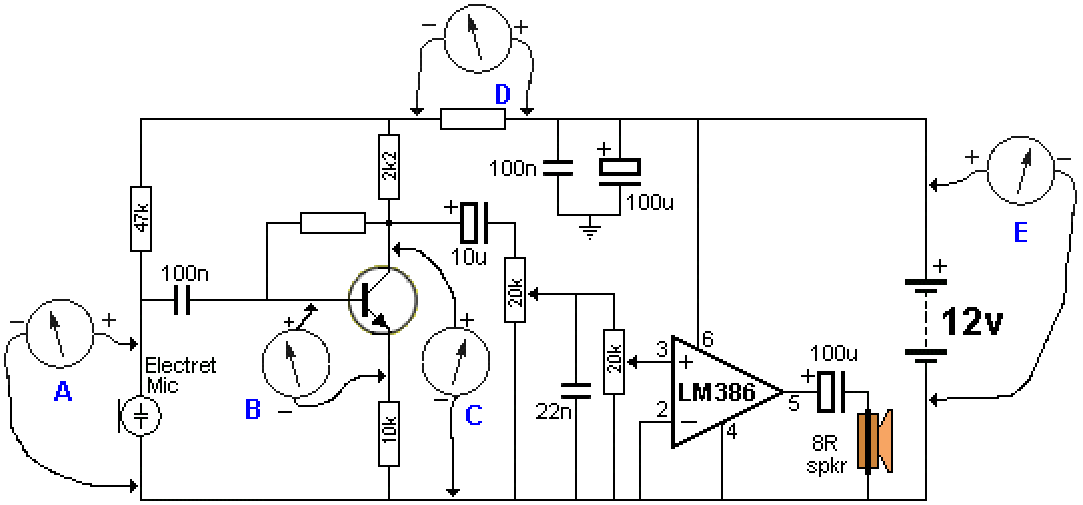

to Index

MEASURING CURRENT

You will rarely need to take current measurements, however most multimeters have DC current ranges such as 0.5mA, 50mA, 500mA and 10Amp (via the extra banana socket) and some meters have AC current ranges. Measuring the current of a circuit will tell you a lot of things. If you know the normal current, a high or low current can let you know if the circuit is overloaded or not fully operational.

Current is always measured when the circuit is working (i.e: with power applied).
It is measured IN SERIES with the circuit or component under test.
The easiest way to measure current is to remove the fuse and take a reading across the fuse-holder. Or remove one lead of the battery or turn the project off, and measure across the switch.
If this is not possible, you will need to remove one end of a component and measure with the two probes in the "opening."
Resistors are the easiest things to desolder, but you may have to cut a track in some circuits. You have to get an "opening" so that a current reading can be taken.
The following diagrams show how to connect the probes to take a CURRENT reading.
Do not measure the current ACROSS a component as this will create a "short-circuit."
The component is designed to drop a certain voltage and when you place the probes across this component, you are effectively adding a "link" or "jumper" and the voltage at the left-side of the component will appear on the right-side. This voltage may be too high for the circuit being supplied and the result will be damage.

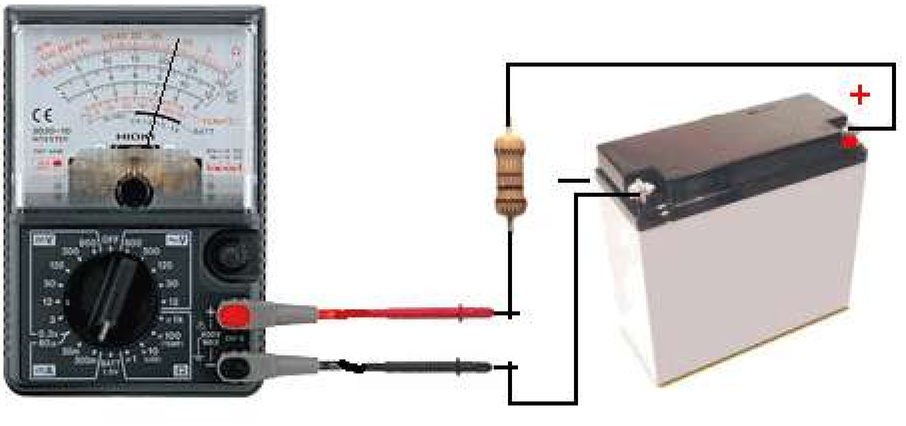

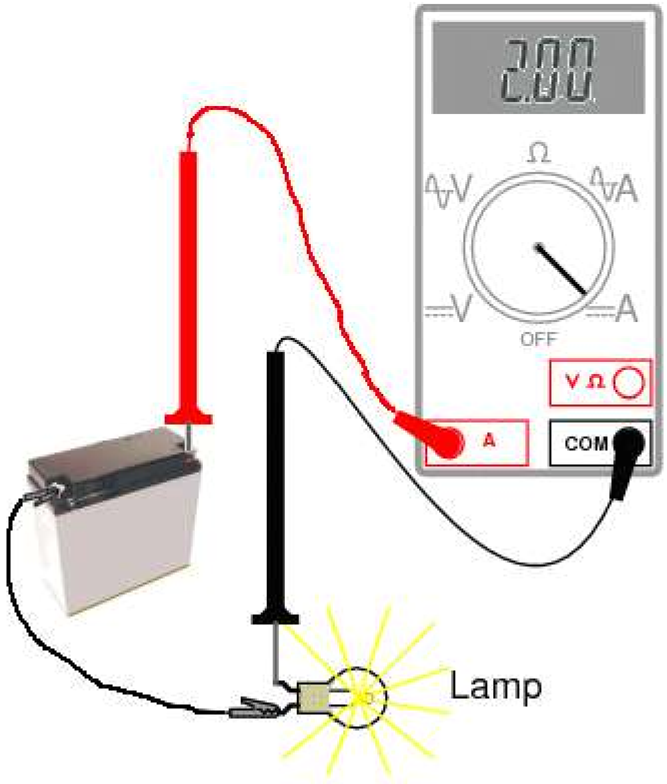

Measuring the current of a globe

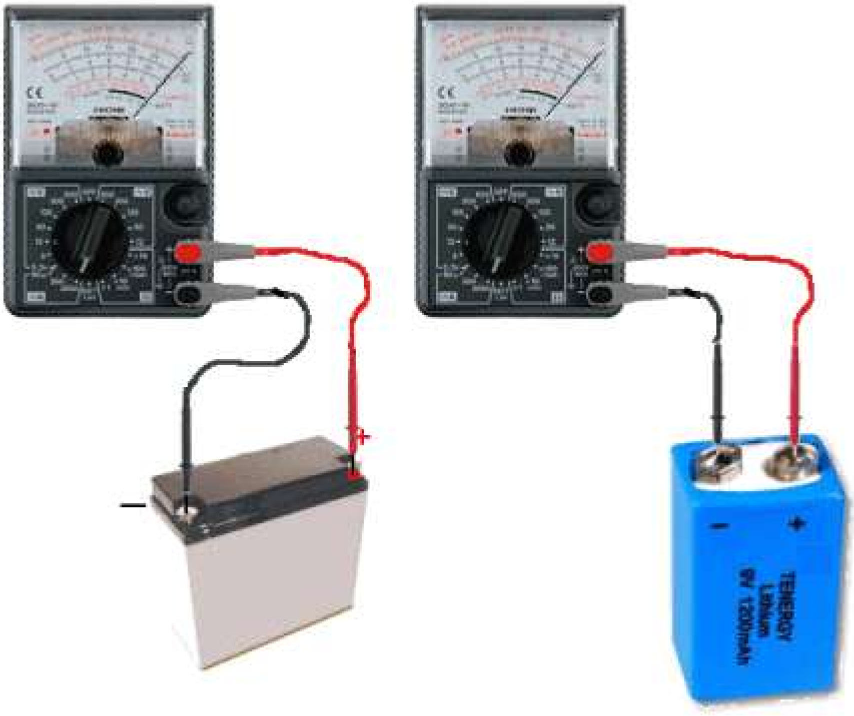

Do NOT measure the CURRENT of a battery
(by placing the meter directly across the terminals)
A battery will deliver a very HIGH current
and damage the meter

Do not measure the "current a battery will deliver" by placing the probes across the terminals. It will deliver a very high current and damage the meter instantly. There are special battery testing instruments for this purpose.

When measuring across an "opening" or "cut," place the red probe on the wire that supplies the voltage (and current) and the black probe on the other wire. This will produce a "POSITIVE" reading.
A positive reading is an UPSCALE READING and the pointer will move across the scale - to the right. A "NEGATIVE READING" will make the pointer hit the "STOP" at the left of the scale and you will not get a reading. If you are using a Digital Meter, a negative sign "-" will appear on the screen to indicate the probes are around the wrong way. No damage will be caused. It just indicates the probes are connected incorrectly.
If you want an accurate CURRENT MEASUREMENT, use a digital meter.

to Index

MEASURING RESISTANCE

Turn a circuit off before measuring resistance.
If any voltage is present, the value of resistance will be incorrect.
In most cases you cannot measure a component while it is in-circuit. This is because the meter is actually measuring a voltage across a component and calling it a "resistance."
The voltage comes from the battery inside the meter. If any other voltage is present, the meter will produce a false reading.
If you are measuring the resistance of a component while still "in circuit," (with the power off) the reading will be lower than the true reading.

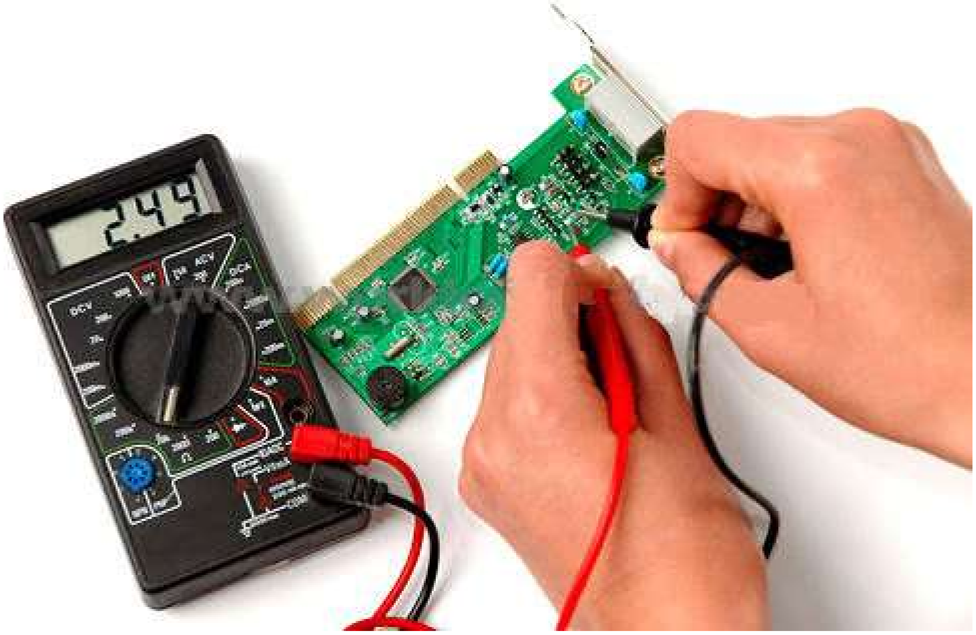

Measuring resistance

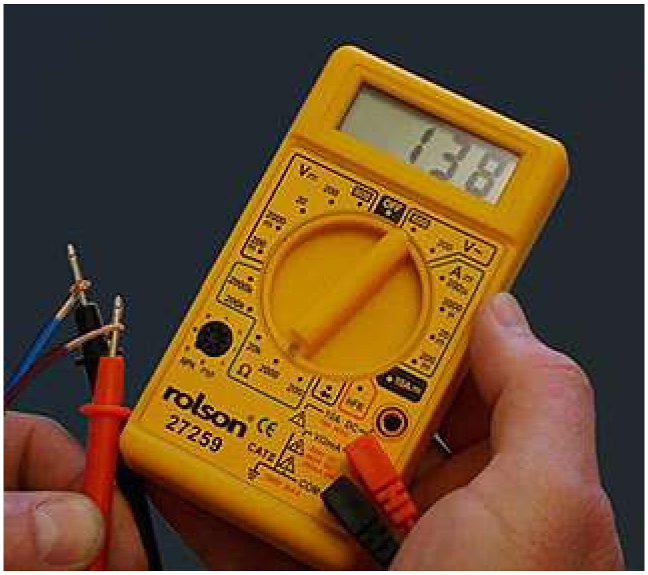
138
rolson
27259

Measuring resistance of a heater
(via the leads)

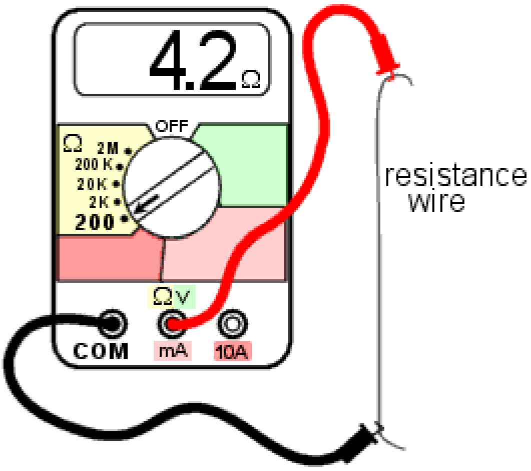
4.2Ω
OFF
Ω 2M
200 K
20K
2K
200
ΩV
COM
mA
10A
resistance wire

Measuring the resistance of a piece of resistance-wire

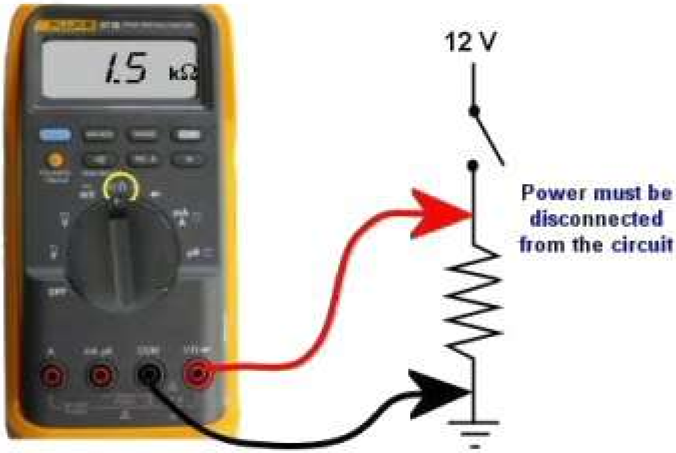
1.5 kΩ
12 V
Power must be disconnected from the circuit

Measuring the resistance of a resistor

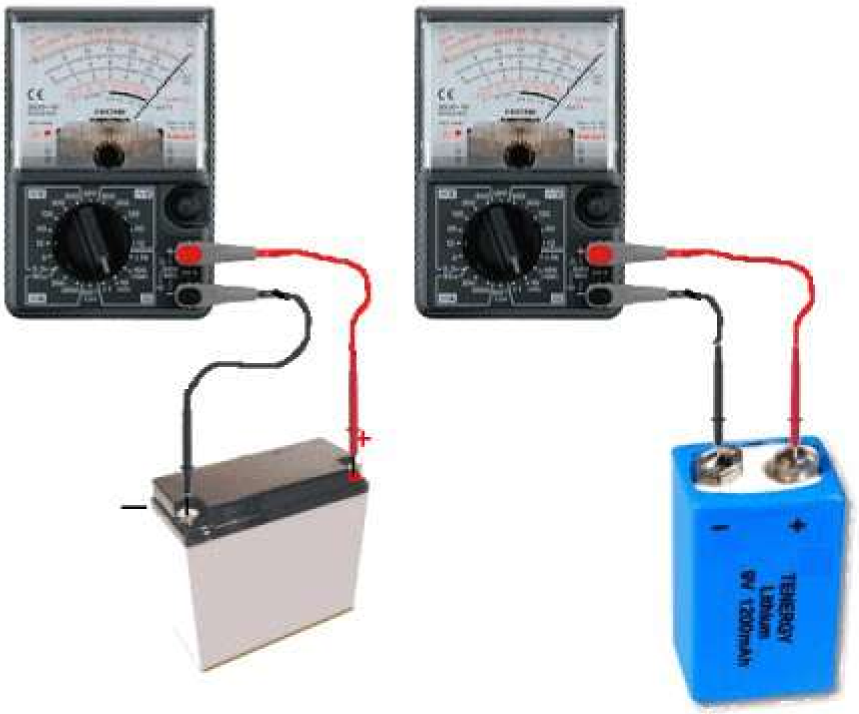

Do not measure the "Resistance of a Battery"

1. Do not measure the "resistance of a battery." The resistance of a battery (called the Internal impedance) is not measured as shown in the diagrams above. It is measured by creating a current-flow and measuring the voltage across the battery. Placing a multimeter set to **resistance** (across a battery) will destroy the meter.
2. Do not try to measure the resistance of any voltage or any "supply."

Resistance is measured in OHMs.
The resistance of a 1cm x 1cm bar, one metre long is 1 ohm.
If the bar is thinner, the resistance is higher. If the bar is longer, the resistance is higher. If the material of the bar is changed, the resistance is higher.
When carbon is mixed with other elements, its resistance increases and this knowledge is used to make RESISTORS.
Resistors have RESISTANCE and the main purpose of a resistor is to reduce the CURRENT FLOW.
It's a bit like standing on a hose. The flow reduces.
When current flow is reduced, the output voltage is also reduced and that why the water does not spray up so high. Resistors are simple devices but they produce many different effects in a circuit.
A resistor of nearly pure carbon may be 1 ohm, but when non-conducting "impurities" are added, the same-size resistor may be 100 ohms, 1,000 ohms or 1 million ohms.
Circuits use values of less than 1 ohm to more than 22 million ohms.

Resistors are identified on a circuit with numbers and letters to show the exact value of resistance - such as 1k 2k2 4M7
The letter Ω (omega - a Greek symbol) is used to identify (or express) (or represent) the word "Ohm."
But this symbol is not available on some word-processors, so the letter "R" is used. The letter "E" is also sometimes used and both mean "Ohms."
A one-ohm resistor is written "1R" or "1E." It can also be written "1R0" or "1E0."
A resistor of one-tenth of an ohm is written "0R1" or "0E1." The letter takes the place of the decimal point.
10 ohms = 10R
100 ohms = 100R

1,000 ohms = 1k (k= kilo = one thousand)
10,000 ohms = 10k
100,000 ohms = 100k
1,000,000 ohms = 1M   (M = MEG = one million)

The size of a resistor has nothing to do with its resistance. The size determines the wattage of the resistor - how much heat it can dissipate without getting too hot.
Every resistor is identified by colour bands on the body, but when the resistor is a surface-mount device, numbers are used and sometimes letters.
You MUST learn the colour code for resistors and the following table shows all the colours for the most common resistors from 1/10th of an ohm to 22 Meg ohms for resistors with 5% and 10% tolerance.

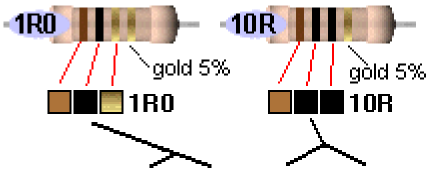

If 3rd band is gold, Divide by 10
If 3rd band is silver, Divide by 100
(to get 0.22ohms etc)

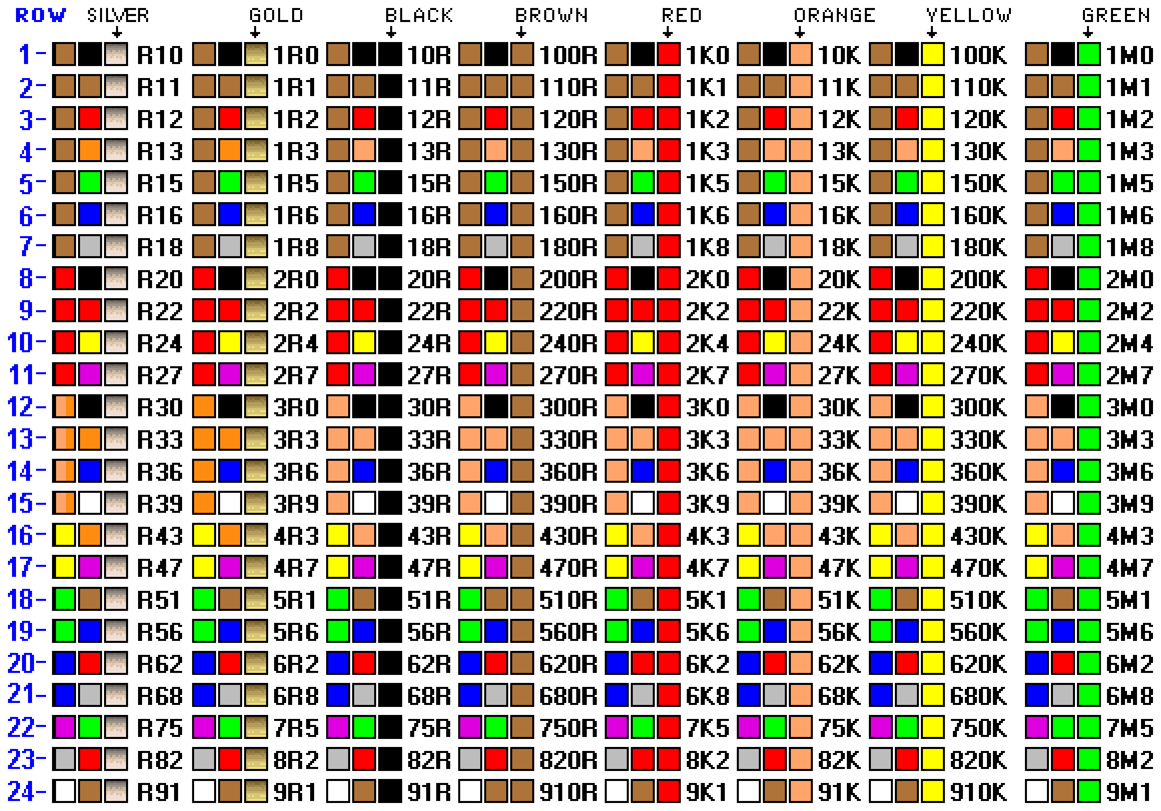

ROW	SILVER	GOLD	BLACK	BROWN	RED	ORANGE	YELLOW	GREEN
1-	R10	1R0	10R	100R	1K0	10K	100K	1M0
2-	R11	1R1	11R	110R	1K1	11K	110K	1M1
3-	R12	1R2	12R	120R	1K2	12K	120K	1M2
4-	R13	1R3	13R	130R	1K3	13K	130K	1M3
5-	R15	1R5	15R	150R	1K5	15K	150K	1M5
6-	R16	1R6	16R	160R	1K6	16K	160K	1M6
7-	R18	1R8	18R	180R	1K8	18K	180K	1M8
8-	R20	2R0	20R	200R	2K0	20K	200K	2M0
9-	R22	2R2	22R	220R	2K2	22K	220K	2M2
10-	R24	2R4	24R	240R	2K4	24K	240K	2M4
11-	R27	2R7	27R	270R	2K7	27K	270K	2M7
12-	R30	3R0	30R	300R	3K0	30K	300K	3M0
13-	R33	3R3	33R	330R	3K3	33K	330K	3M3
14-	R36	3R6	36R	360R	3K6	36K	360K	3M6
15-	R39	3R9	39R	390R	3K9	39K	390K	3M9
16-	R43	4R3	43R	430R	4K3	43K	430K	4M3
17-	R47	4R7	47R	470R	4K7	47K	470K	4M7
18-	R51	5R1	51R	510R	5K1	51K	510K	5M1
19-	R56	5R6	56R	560R	5K6	56K	560K	5M6
20-	R62	6R2	62R	620R	6K2	62K	620K	6M2
21-	R68	6R8	68R	680R	6K8	68K	680K	6M8
22-	R75	7R5	75R	750R	7K5	75K	750K	7M5
23-	R82	8R2	82R	820R	8K2	82K	820K	8M2
24-	R91	9R1	91R	910R	9K1	91K	910K	9M1

10M
BLUE

COLOR CODES FOR THE WHOLE E12/E24 RANGE OF RESISTORS
The twelve odd rows - 1, 3, 5...- represent values available in the E12 range only, plus 10M

to Index

MAKE YOUR OWN RESISTOR

Make your own variable resistor that changes resistance according to the pressure.
Use a piece of conductive foam used to package Integrated Circuits. You can ask at an electronics shop.
Use two coins or pieces of printed circuit board or aluminium foil for the top and bottom conductors.
You can solder wires to the PC board or fold the aluminium foil over a few times to hold the wires.
The resistance of the foam will reduce as you press on the "cell."
The actual resistance-values will depend on the size of the foam, the thickness and

pressure.
This cell is a very simple cell called a **LOAD CELL**.

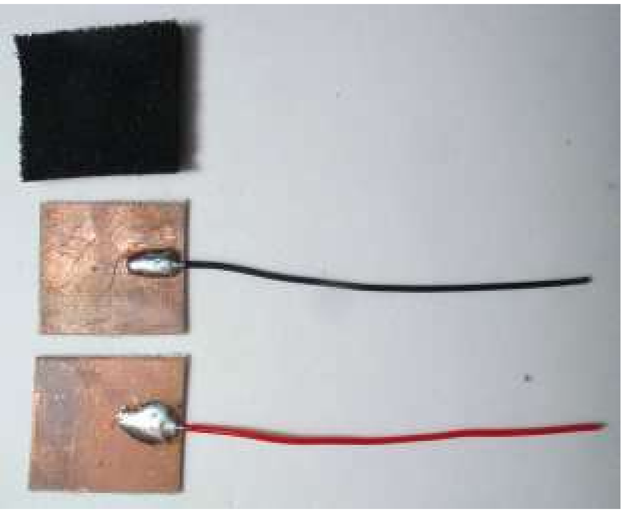

The top and bottom "plates"

The foam is placed between the plates.

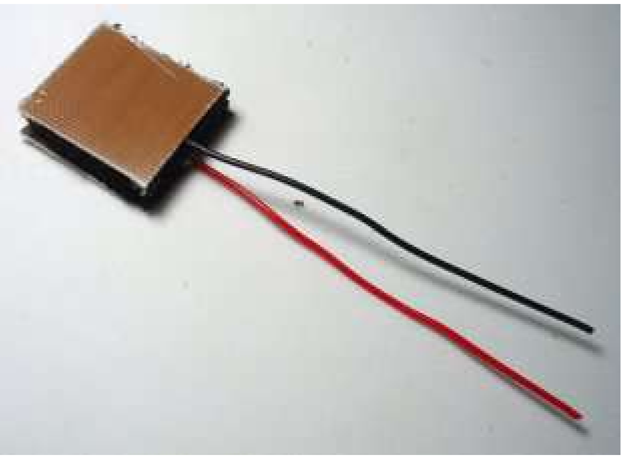

The complete LOAD CELL

The resistance of the unloaded LOAD CELL

The fully loaded resistance can be as low as 9,330 ohms

to Index

MEASURING CONTINUITY

CONTINUITY is the same as **ZERO OHMS** or the resistance of a short length of wire. It can also mean the resistance through a switch or globe or a low-value resistor.
It basically means a "PATH" and sometimes refers to a whole circuit when the switch is closed. In other words CONTINUITY means we have a "circuit." We have "current flowing" and generally refers to a low-resistance circuit.
Both ANALOGUE and DIGITAL multimeters can measure CONTINUITY and you have to work out the approximate value of resistance for the circuit you are testing, - BEFORE TAKING A READING.
If the reading is above 300 ohms or contains a diode, you cannot use a DIGITAL MULTIMETER as the buzzer on the continuity setting will not respond.

The project being tested must not have the power applied as the resistance ranges on a multimeter are actually measuring a voltage across the leads and any voltage on the circuit or contained in any electrolytics, will upset the reading.

To take a reading with an **ANALOGUE** multimeter, select the x1 setting and the pointer will move across the scale to the actual value of resistance.
It it move full scale, you have ZERO OHMS resistance and this can mean a short-circuit or continuity via a wire.
If a diode is in the circuit you must also reverse the leads to get a reading.
The resistance of a globe will be very low when it is not illuminated, so don't think a fault is present.
Measuring CONTINUITY is the same as measuring LOW RESISTANCE.

To take a reading with a **DIGITAL** multimeter, select the buzzer setting. It will respond if the resistance is less than 300 ohms. It will not respond if a diode is in the circuit.

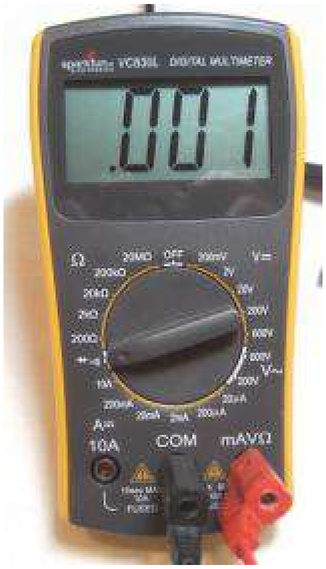

Meter set to BUZZER - CONTINUITY

You can also use the x1 resistance setting to get an accurate value of resistance. Touch the probes together to get the initial reading and subtract this value from the final reading.
When probing a circuit containing electrolytics, you may get a beep from the buzzer. This indicates the resistance is low because the multimeter is charging the electrolytic and it will beep until the electrolytic is charged to about 0.7v.
The same applies when probing across the power rails of a circuit. The circuit may contain electrolytics that will charge when probing and the buzzer will beep.
The Digital multimeter is actually detecting a voltage less than 0.7v across the probes and is created by a voltage-divider network inside the meter.
The voltage divider put 2v across the probes and when this drops to less than 0.5v, the buzzer is activated. That why it odes not buzz when testing a diode as the diode drops the voltage to 0.6v.

to Index

MEASURING A DIODE

A diode can be measured to see if it is "open" or "damaged" or "working" by placing the probes across the component.
If the diode is "open" (it will not work), the needle will NOT swing across the scale when touching the component with the probes in one direction or when the probes are reversed.
If the diode is "damaged" (does not work), the needle will swing fully across the scale when touching the component with the probes in one direction or when the probes are reversed.
If the diode is FUNCTIONAL, (works) the needle will swing about mid-way when touching the leads of the diode in one direction and it will not move when the probes are reversed.

WHY?

The positive of the battery inside an analogue multimeter comes out the black probe and that is why you will get a reading when the probers are "around the wrong way." The needle will swing a different amount for each resistance setting on the dial as the needle represents 0.6v drop and NOT an actual resistance.

There are two things you must remember.
1. When the diode is measured in one direction, the needle will will **not move at**

all. The technical term for this is the diode is **reverse biased**. It will not allow any current to flow. Thus the needle will not move.
When the diode is connected around the other way, the needle will swing to the right (move up scale) to about 80% of the scale. This position represents the voltage drop across the junction of the diode and is NOT a resistance value. If you change the resistance range, the needle will move to a slightly different position due to the resistances inside the meter. The technical term for this is the diode is **forward biased**. This indicates the diode is not faulty.
The needle will swing to a slightly different position for a "normal diode" compared to a Schottky diode. This is due to the different junction voltage drops.
However we are only testing the diode at very low voltage and it may break-down when fitted to a circuit due to a higher voltage being present or due to a high current flowing.

2. The leads of an **Analogue Multimeter** have the positive of the battery connected to the black probe and the readings of a "good diode" are shown in the following two diagrams:

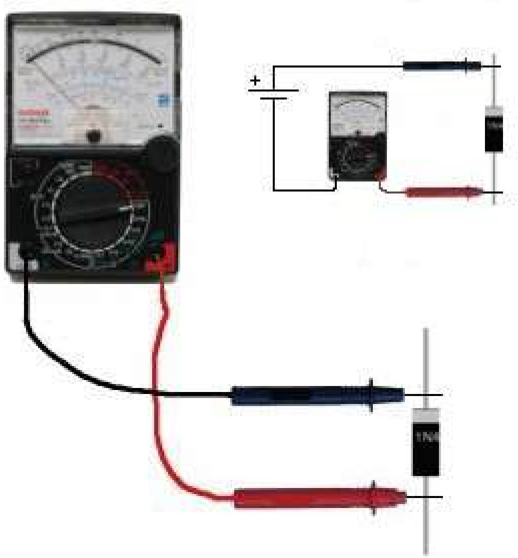

The diode is REVERSE BIASED in the diagram above and diodes not conduct.

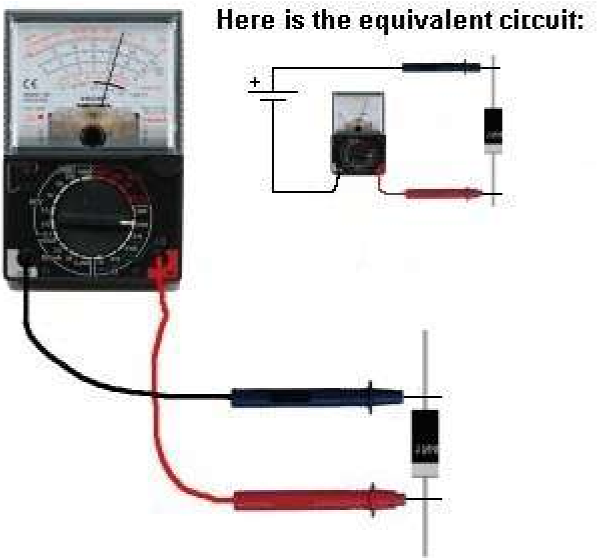

The diode is FORWARD BIASED in the diagram above and it conducts

TESTING A DIODE ON A DIGITAL METER

A Digital multimeter will measure the voltage-drop across the diode when the probes are connected in one direction (approx 0.640 on the scale) and a high reading (1) in the other direction. You need to select the "DIODE" setting on the dial as the other settings will produce a meaningless reading.

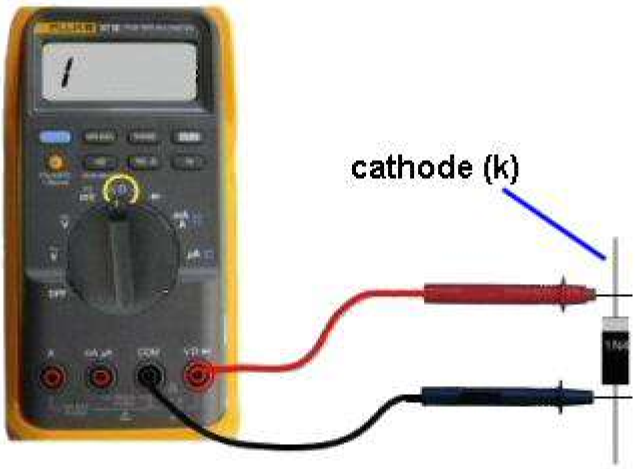

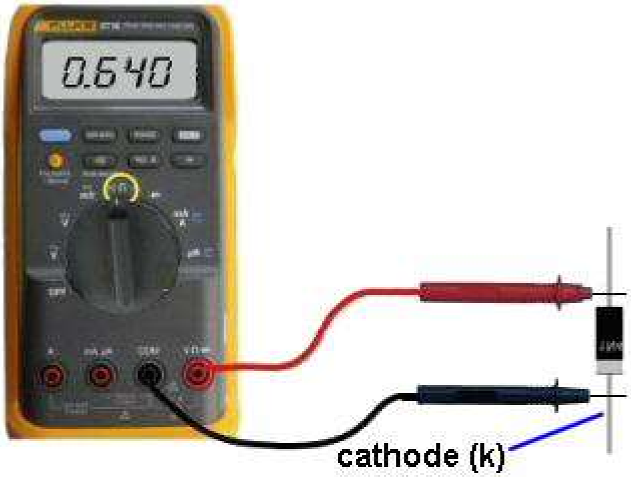

Some DIGITAL MULTIMETERS will show mV drop across the diode when the setting on the meter is "diode" or the "x1" or "x10" resistance range.

to Index

TESTING A LED

Some multimeters will test LEDs.
It depends on the voltage of the battery inside the case of the multimeter.
Many analogue multimeters have a single 1.5v cell and these cannot test LEDs.
Analogue Multimeters with 3v (for the resistance ranges) can test some LEDs.
White LEDs need about 3.6v and they may not illuminate on 3v.

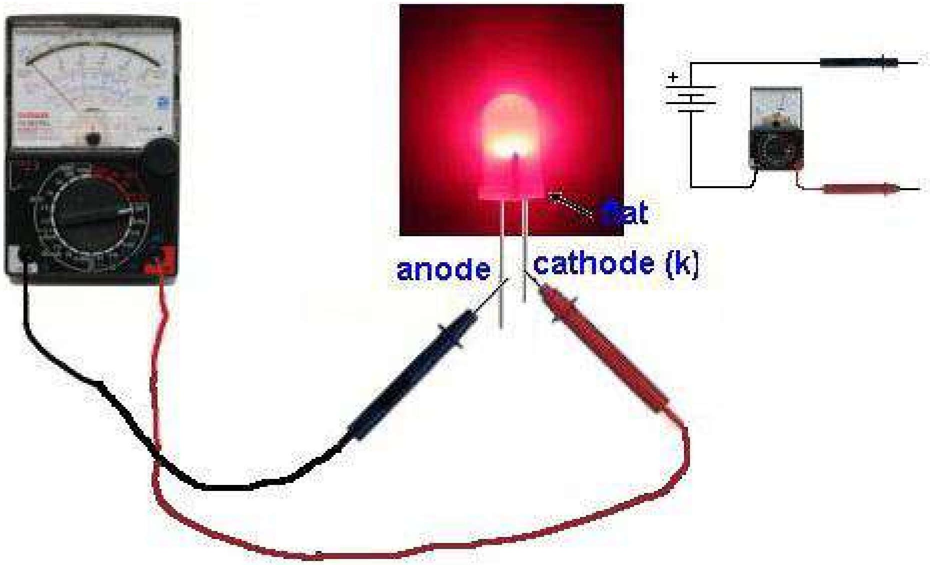

The negative lead of an ANALOGUE meter is POSITIVE!
The multimeter must have 3v (2 cells)

Digital multimeters have a 9v battery and they will illuminate all colour LEDs when the leads are placed as shown in the diagram:

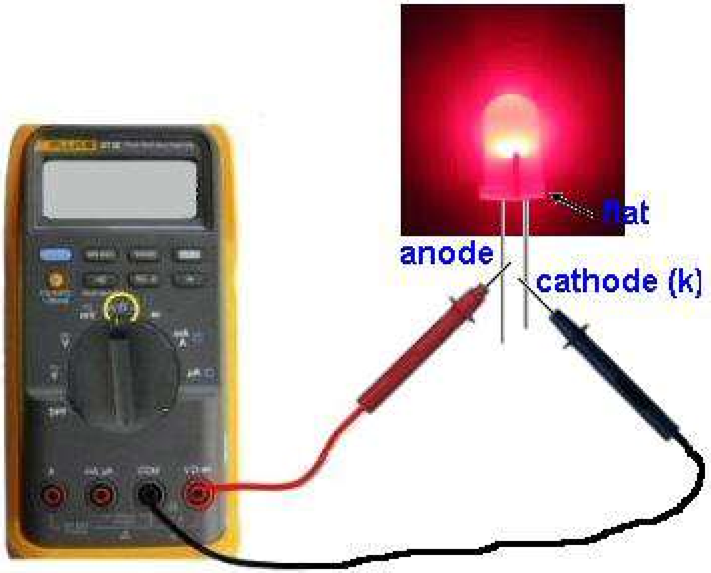

A Digital meter will illuminate all LEDs and the black probe touches the cathode.

to Index

TESTING A TRANSISTOR WITH A DIGITAL METER

Testing a transistor with a **Digital Meter** must be done on the "DIODE" setting as a digital meter does not deliver a current through the probes on some of the resistance settings and will not produce an accurate reading.
The "DIODE" setting must be used for diodes and transistors. It should also be called a "TRANSISTOR" setting.

TESTING A TRANSISTOR WITH AN ANALOGUE METER

The first thing you may want to do is test an unknown transistor for COLLECTOR, BASE AND EMITTER. You also want to perform a test to find out if it is NPN or PNP. That's what this test will provide.
You need a cheap multimeter called an ANALOGUE METER - a multimeter with a

scale and pointer (needle).
It will measure resistance values (normally used to test resistors) - (you can also test other components) and Voltage and Current. We use the resistance settings. It may have ranges such as "x10" "x100" "x1k" "x10"
Look at the resistance scale on the meter. It will be the top scale.
The scale starts at zero on the right and the high values are on the left. This is opposite to all the other scales.
When the two probes are touched together, the needle swings FULL SCALE and reads "ZERO." Adjust the pot on the side of the meter to make the pointer read exactly zero.

How to read: "x10" "x100" "x1k" "x10"
Up-scale from the zero mark is "1"
When the needle swings to this position on the "x10" setting, the value is 10 ohms.
When the needle swings to "1" on the "x100" setting, the value is 100 ohms.
When the needle swings to "1" on the "x1k" setting, the value is 1,000 ohms = 1k.
When the needle swings to "1" on the "x10k" setting, the value is 10,000 ohms = 10k.
Use this to work out all the other values on the scale.
Resistance values get very close-together (and very inaccurate) at the high end of the scale. [This is just a point to note and does not affect testing a transistor.]

Step 1 - FINDING THE BASE and determining NPN or PNP
Get an unknown transistor and test it with a multimeter set to "x10"
Try the 6 combinations and when you have the black probe on a pin and the red probe touches the other pins and the meter swings nearly full scale, you have an NPN transistor. The black probe is BASE
If the red probe touches a pin and the black probe produces a swing on the other two pins, you have a PNP transistor. The red probe is BASE
If the needle swings FULL SCALE or if it swings for more than 2 readings, the transistor is **FAULTY**.

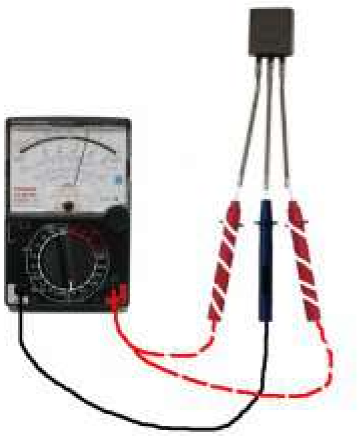

This is an NPN transistor
The black probe is the BASE

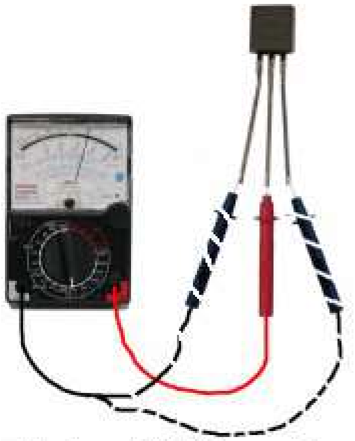

This is a PNP transistor
The red probe is the BASE

Step 2 - FINDING THE COLLECTOR and EMITTER
Set the meter to "x10k."
For an NPN transistor, place the leads on the transistor and when you press hard on the two leads shown in the diagram below, the needle will swing almost full scale.

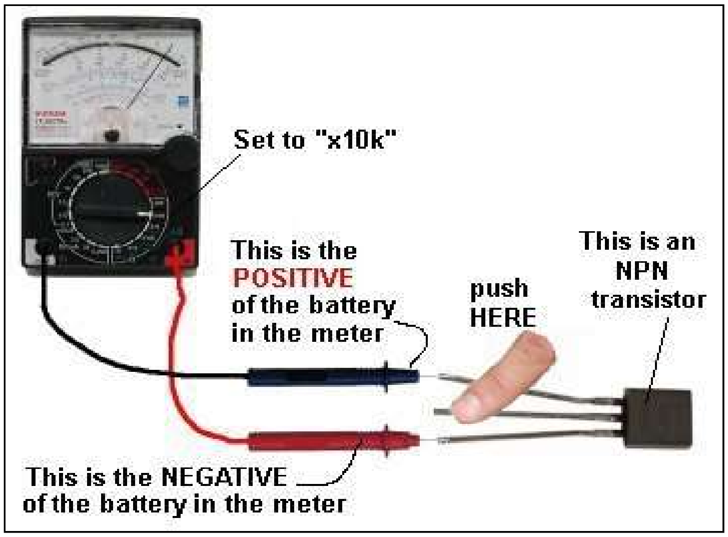
Set to "x10k"
This is the POSITIVE of the battery in the meter
push HERE
This is an NPN transistor
This is the NEGATIVE of the battery in the meter

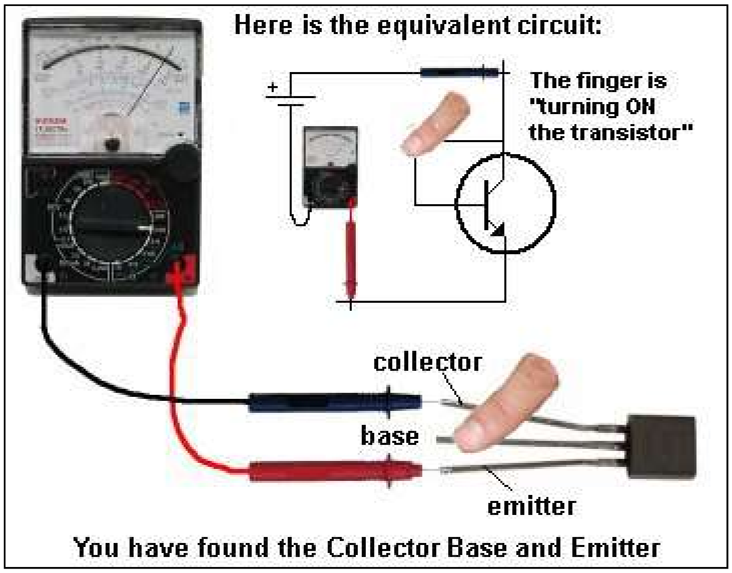
Here is the equivalent circuit:
+
The finger is "turning ON the transistor"
collector
base
emitter
You have found the Collector Base and Emitter

For a PNP transistor, set the meter to "x10k" place the leads on the transistor and when you press hard on the two leads shown in the diagram below, the needle will swing almost full scale.

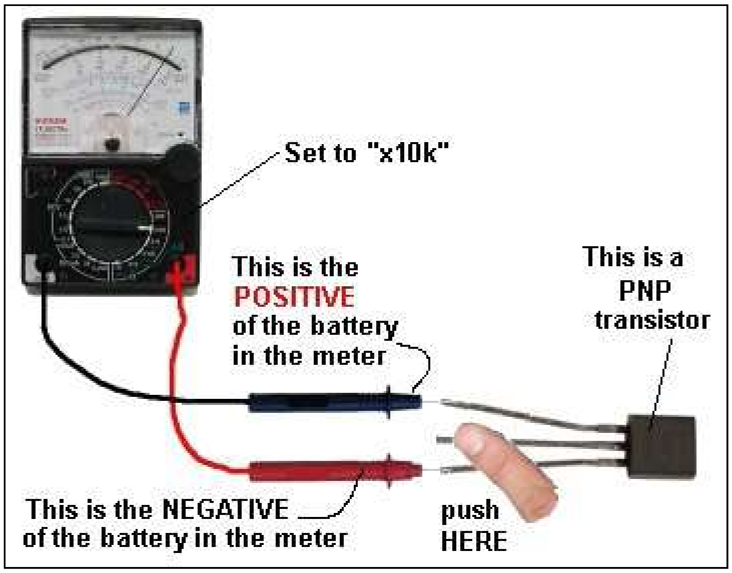
Set to "x10k"
This is the POSITIVE of the battery in the meter
This is a PNP transistor
This is the NEGATIVE of the battery in the meter
push HERE

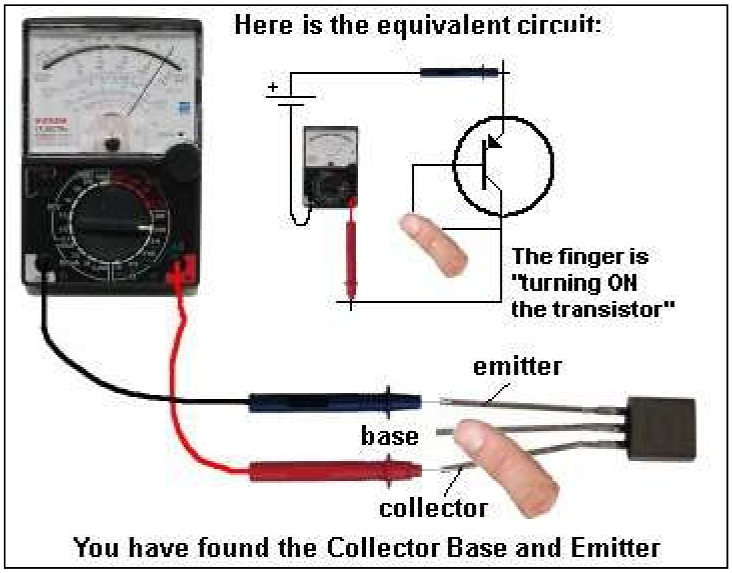
Here is the equivalent circuit:
+
The finger is "turning ON the transistor"
emitter
base
collector
You have found the Collector Base and Emitter

to Index

For more details on testing components with a multimeter, see: Testing Electronic Components.

to Index

MULTIMETER TEST

1. What is the reading on the multimeter:

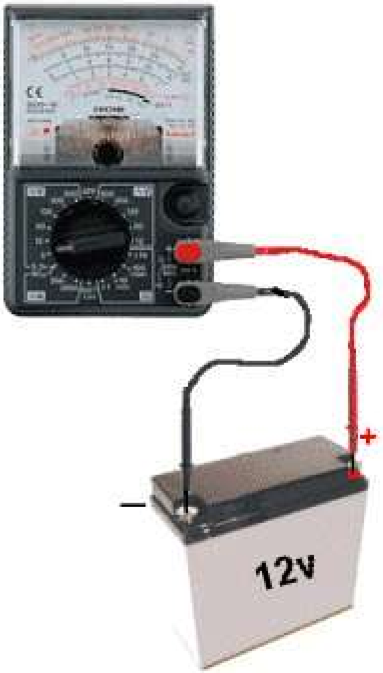

Answer

to Index

2. What is the reading on the multimeter:

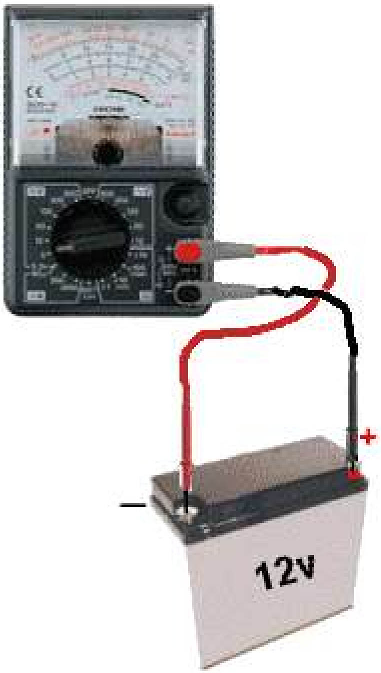

Answer

to Index

3. What is the reading on the multimeter:

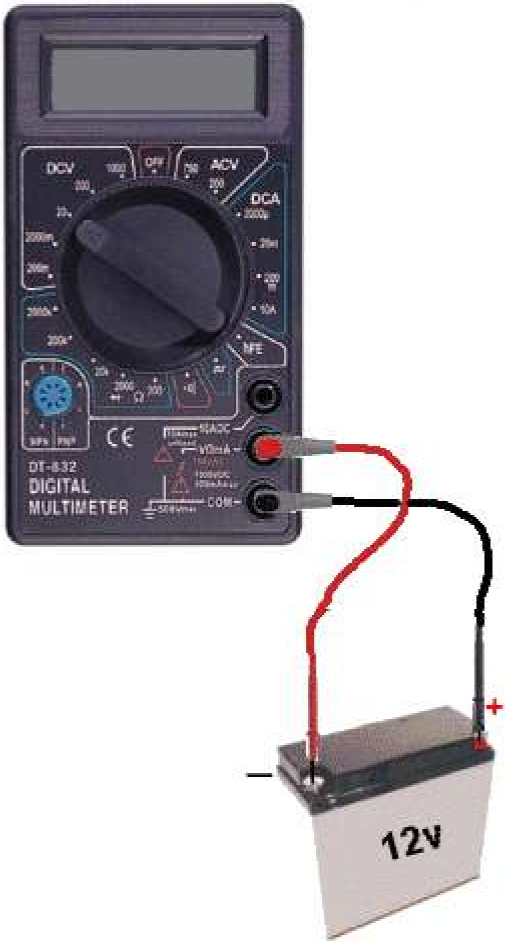
DCV
1000
200
20
2000m
200m
OFF
200
ACV
DCA
2000μ
20m
200m
10A
hFE
2000k
200k
20k
2000
200
Ω
NPN
PNP
CE
10ADC
VΩmA
COM
DT-832
DIGITAL
MULTIMETER
+
–
12v

Answer

to Index

4. What is the voltage between the following points:

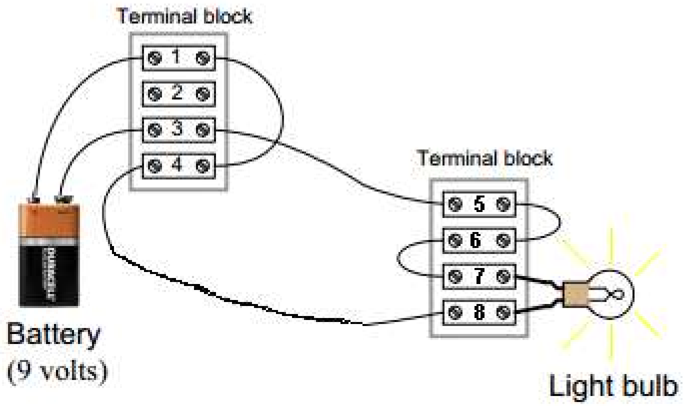
Terminal block
1
2
3
4
Terminal block
5
6
7
8
Battery
(9 volts)
Light bulb

Between 1 & 3	Between 5 & 7	Between 3 & 8	Between 1 & 7
Answer	**Answer**	**Answer**	**Answer**

to Index

5. What is the voltage on each multimeter:

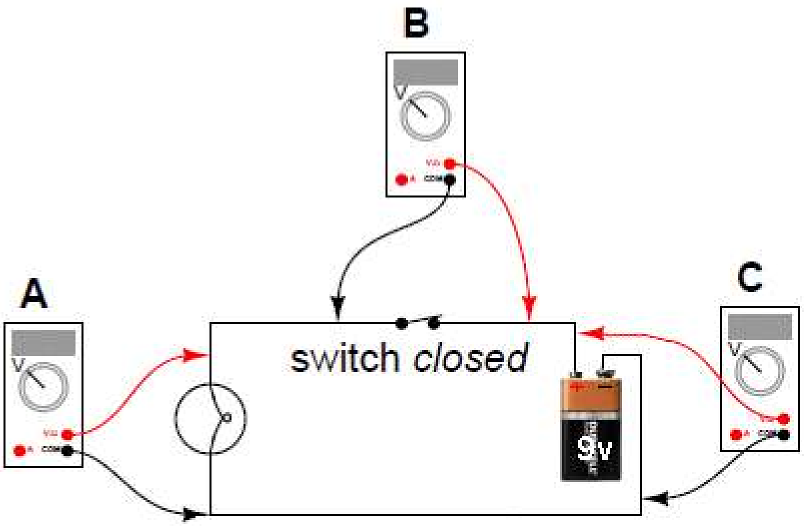
B
A
C
switch closed
9v

Meter A	**Meter B**	**Meter C**
Answer	**Answer**	**Answer**

to Index

6. What is the voltage on each multimeter:

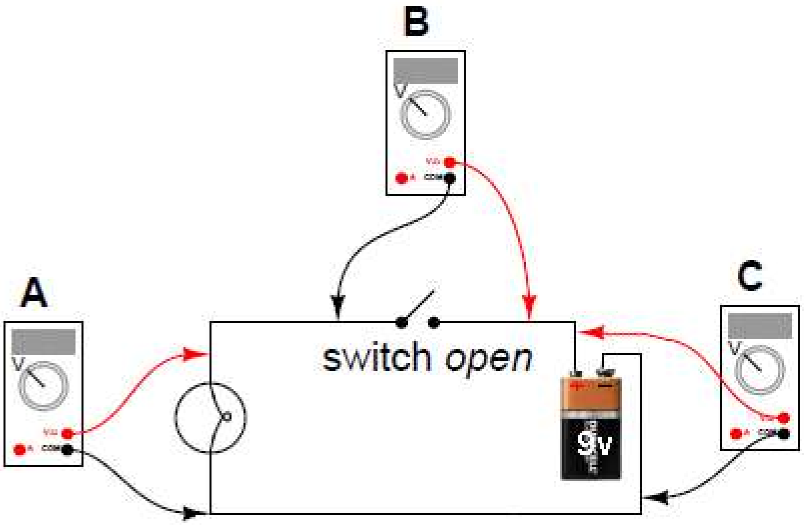
B
A
C
switch *open*
9v

Meter A **Meter B** **Meter C**

Answer Answer Answer

to Index

7. What is the reading on the multimeter:

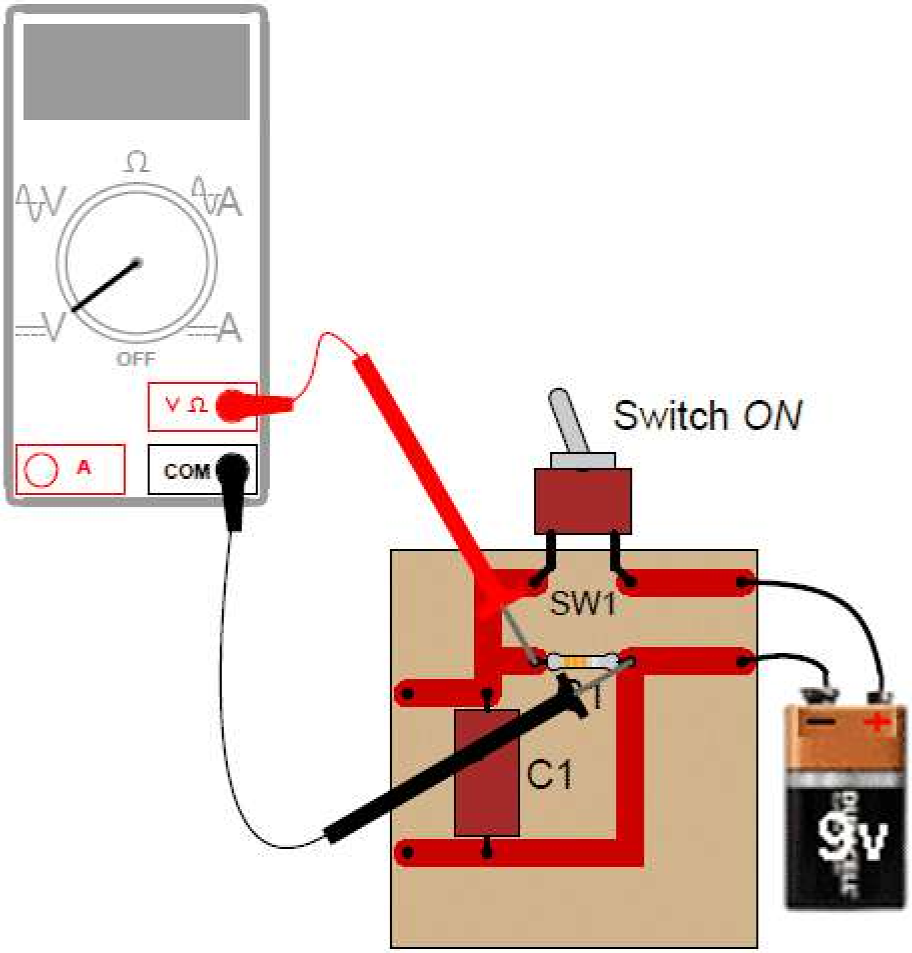
Ω
V
A
V
A
OFF
V Ω
A
COM
Switch *ON*
SW1
R1
C1
9v

Answer

to Index

8. The multimeter has three resistance ranges: **x1 ohms** **x100 ohms** **x10k ohms**. How far across the scale will the needle move on each range for a 4k7 resistor:

x1	**x100**	**x10k**
Answer	**Answer**	**Answer**

to Index

9. The resistance ranges on an ANALOGUE multimeter use the battery inside the case to move the pointer. If the multimeter is left on "**ohms range**" with the probes apart, will the battery go flat?

Answer

to Index

to Index

PCB CAD, Fabrication & Assembly
Pad2Pad
LEARN MORE »

1 | Download PDF Manual - Free
Download Free Manuals With AtoZManuals free.atozmanuals.com/PDF/Manual

2 | Splunk 2018 Predictions
Learn from Splunk experts about the tech & trends poised to transform business in 2018 splunk.com

Download PDF Manual - Free
Download Free Manuals With AtoZManuals free.atozmanuals.com/PDF/Manual

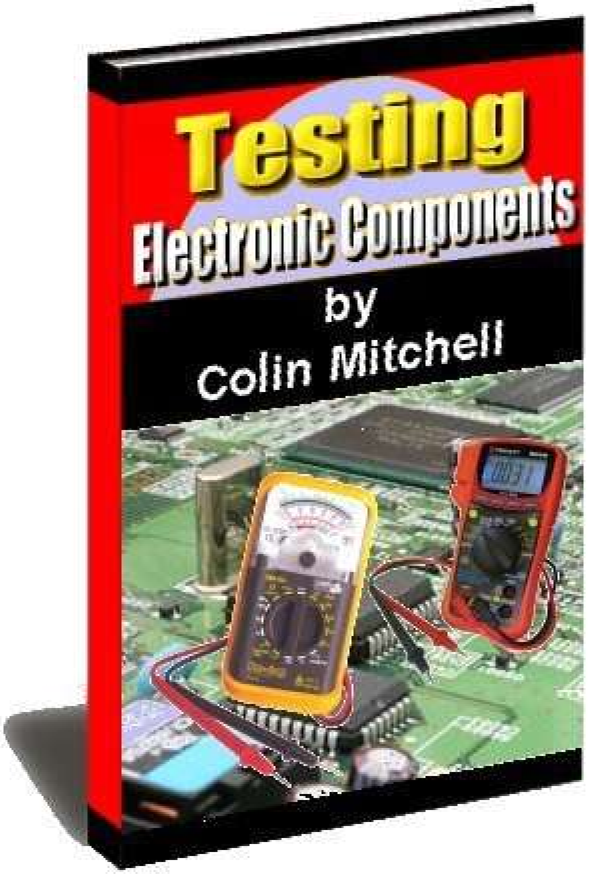

This eBook shows you how to TEST COMPONENTS.
To do this you need "TEST GEAR." The best item of Test Gear is a MULTIMETER. It can test almost 90% of all components. And that's what we will do in this eBook:

CONTENTS

Analogue Multimeter
Audio Stages
Batteries - testing
Burnt Resistor
Buying A Multimeter
Capacitors
Capacitors - decoupling caps
Capacitors in Parallel
Capacitor Substitution Box
Cells - batteries
Circuit Symbols
Co-Ax Cables
Colour Code (Resistor)
Coils
Continuity
Creating any value of R
Multimeters
Non-polar Capacitor (electrolytic)
Open Circuit
"Open" Resistor - damaged
Opto-couplers
Parallel - resistors
Parallel and Series Capacitors
Piezo Diaphragms
Piezo Buzzers
Potentiometers
Pots - testing
Power Diodes
Power Supply -see Basic Electronics 1A
Reed Switch
Relays
Remote Controls

Current - measuring
Current Sensing Resistors
Current Transformer
Damper Diodes
Darlington Transistors
Digital Chips
Digital Multimeter
Digital Transistors
Diodes replacing a diode
Discharge a Capacitor
Earth Leakage Detectors
Electrolytics
FETs
Fingers - to feel temperature
Focus pots
Fuses and 100mA fuses
Germanium Diodes
Heatsinks
Impedance - of a stage
IC's - also called Digital Chips
IC's - Analogue Chips
"In-Circuit" testing
Inductors
Inductors - measuring
Integrated Circuits
Isolation Transformer
Latching Reed Switch
LEDs
Logic Probe MkIIB
Logic Probe - Simple
Logic Probe - using CD4001
Logic Probe - using CD4011
Making your own components
Measuring Resistance
Measuring Voltage
Mica Washers and Insulators
Motor - testing
MOSFETs
Resistor Colour Code
Resistor Networks
Resistors - series
Resistor Substitution Box
Ripple Factor
Schottky Diodes
SCRs
Short Circuit
Signal Injector
Silicon diodes
Soldering
Spark Gaps
Speaker
Substitution Box
Super Probe MkII
Surface Mount - Packs
Surface-Mount Resistors
Surface-Mount Resistor Markings
Switches
Symbols
Test Equipment
Testing A Circuit
Testing A Resistor
Testing Components "In-Circuit"
Transformers - current transformer
Transformer Ratings 28v-0-28v
Transistor Outlines
Transistors
Transistors Digital
Triacs
Unknown resistors - testing
Using A Multimeter
Voltage Divider
Voltage Regulators
Voltages on a circuit
Wire Wound Resistor
Yokes
Zener Diodes
4-Band Resistors
5-Band Resistors
10-Turn Pots

Use your "brain, knowledge and your fingers."

Before we start, fixing anything is a combination of skill, luck and good diagnosis.
Sometimes you can fix something by letting it run until it finally fails.
Some things start to work as soon as you touch them.
Some things can never be fixed.
But some things can be fixed by feeling the temperature rise and deciding if it is getting too hot.
Sometimes you can smell something getting too hot.
Sometimes you can see SMOKE.
All these things make you a very clever technician and about 50% of faults will be fixed by looking for dry joints, burnt parts, overheating and carefully inspecting an item before you disturb it.

By simply touching different items you can quite often feel a hot item and home-in on the fault - at a saving of hours of work.
Servicing is not "A bull at a gate" approach.
You may be able to service something by turning it on and leaving it for hours - and start thinking.
It may take you a day to come up with the answer.
Believe me. That's how it worked for me - while fixing over 35,000 TV's.

TEST EQUIPMENT

Everyone thinks TEST EQUIPMENT will "solve the problem."
This is a big big MISTAKE.
Test equipment can help solve a problem and it can "lead to frustration," "give an incorrect answer," "mess you up," and make things worse.
You have to be very careful with test equipment and especially EXPENSIVE equipment because it is very sensitive and can detect pulses and glitches and voltages that are not affecting the operation of the circuit.
You will learn a lot of tricks when reading through this article, but let me say two things.
There are lots of faults and components that you cannot test with "test equipment" because they are either intermittent or the equipment does not load the device to the same extent as the circuit.
And secondly you need both an ANALOGUE multimeter and a DIGITAL meter to cover all the situations.
And if you are working on a car, you only need a $5.00 analogue meter because it will be dropped or fall into a crack, and you will only lose $5.00
You will learn that a digital meter will pick up spikes and signals on a line and show an incorrect reading.
That's why you need to back-up your readings with an analogue meter.
When you charge a battery it gets a "floating voltage" and this will be higher than the actual voltage, when the battery is fitted to a project. An analogue meter will draw a slight current and remove the "floating voltage."
Component testers can also give you a false reading, either because the component is out of range of the tester or intermittent and you need to be aware of this.
Oscilloscopes can also display waveforms that are parts of glitches or noise from other chips and these do not affect the operation of the part of the circuit you are investigating.
Sometimes you cannot pickup a pulse because it is not regular and the trigger on the oscilloscope does not show it on the screen. You may think it is missing.
It all depends on the "speed of the oscilloscope" - it's maximum frequency of operation.
Lastly- Power Supplies. You cannot test globes and motors on a power supply because the starting current can be 5 times more than the operating current. The power supply may not be able to deliver this high current and thus you will think the motor or globe is faulty.

MULTIMETERS

There are two types:

DIGITAL and ANALOGUE

A **Digital Multimeter** has a set of digits on the display and an Analogue Multimeter has a scale with a pointer (or needle).
You really need both types to cover the number of tests needed for designing and repair-work. We will discuss how they work, how to use them and some of the differences between them.

DIGITAL AND ANALOGUE MULTIMETERS

BUYING A MULTIMETER

There are many different types on the market.
The cost is determined by the number of ranges and also the extra features such as diode tester, buzzer (continuity), transistor tester, high DC current and others.
Since most multimeters are reliable and accurate, buy one with the greatest number of ranges at the lowest cost.
This article explains the difference between a cheap analogue meter, an expensive analogue meter and a digital meter. You will then be able to work out which two meters you should buy.

Multimeters are sometimes called a "meter", a "VOM" (Volts-Ohms-Milliamps or Volt Ohm Meter) or "multi-tester" or even "a tester" - they are all the same.

USING A MULTIMETER

Analogue and digital multimeters have either a rotary selector switch or push buttons to select the appropriate function and range. Some Digital Multimeters (DMMs) are auto ranging; they automatically select the correct range of voltage, resistance, or current when doing a test. However you need to select the function.

Before making any measurement you need to know what you are checking. If you are measuring voltage, select the AC range (10v, 50v, 250v, or 1000v) or DC range (0.5v, 2.5v, 10v, 50v, 250v, or 1000v). If you are measuring resistance, select the Ohms range (x1, x10, x100, x1k, x10k). If you are measuring current, select the appropriate current range DCmA 0.5mA, 50mA, 500mA. Every multimeter is different however the photo below shows a low cost meter with the basic ranges.

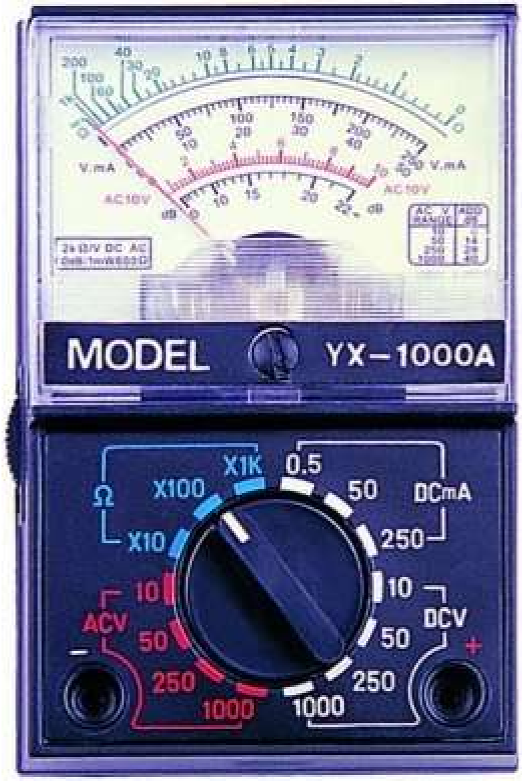
MODEL
YX-1000A
Ω
X10
X100
X1K
0.5
50
250
DCmA
10
50
250
1000
DCV
ACV
10
50
250
1000
−
+

The most important point to remember is this:
You must select a voltage or current range that is bigger or HIGHER than the maximum expected value, so the needle does not swing across the scale and hit the "end stop."
If you are using a DMM (Digital Multi Meter), the meter will indicate if the voltage or current is higher than the selected scale, by showing "OL" - this means "Overload." If you are measuring resistance such as 1M on the x10 range the "OL" means "Open Loop" and you will need to change the range. Some meters show "1' on the display when the measurement is higher than the display will indicate and some flash a set of digits to show over-voltage or over-current. A "-1" indicates the leads should be reversed for a "positive reading."
If it is an AUTO RANGING meter, it will automatically produce a reading, otherwise the selector switch must be changed to another range.

HV
000
DCV
OFF
ACV
DCA
hFE
10ADC
VΩmA
COM
DT-832
DIGITAL
MULTIMETER

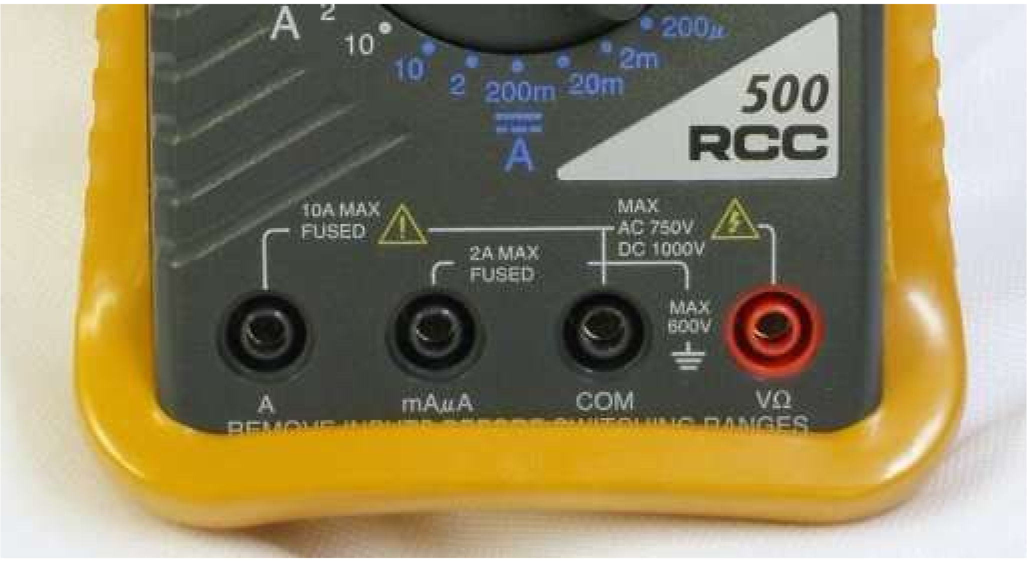
A
10
10
2
200m
20m
2m
200µ
A
500
RCC
10A MAX
FUSED
2A MAX
FUSED
MAX
AC 750V
DC 1000V
MAX
600V
A
mAµA
COM
VΩ

The Common (negative) lead ALWAYS fits into the "COM" socket. The red lead fits into the red socket for Voltage and Resistance. Place the red lead (red banana plug) into "A" (for HIGH CURRENT "Amps") or mA,uA for LOW CURRENT.

The black "test lead" plugs into the socket marked "-" "Common", or "Com," and the red "test lead" plugs into meter socket marked "+" or "V-W-mA."   The third banana socket measures HIGH CURRENT and the positive (red lead) plugs into this. You DO NOT move the negative "-" lead at any time.
The following two photos show the test leads fitted to a digital meter. The probes and plugs have "guards" surrounding the probe tips and also the plugs so you can measure high voltages without getting near the voltage-source.

FLUKE
600.0
OFF
COM
A

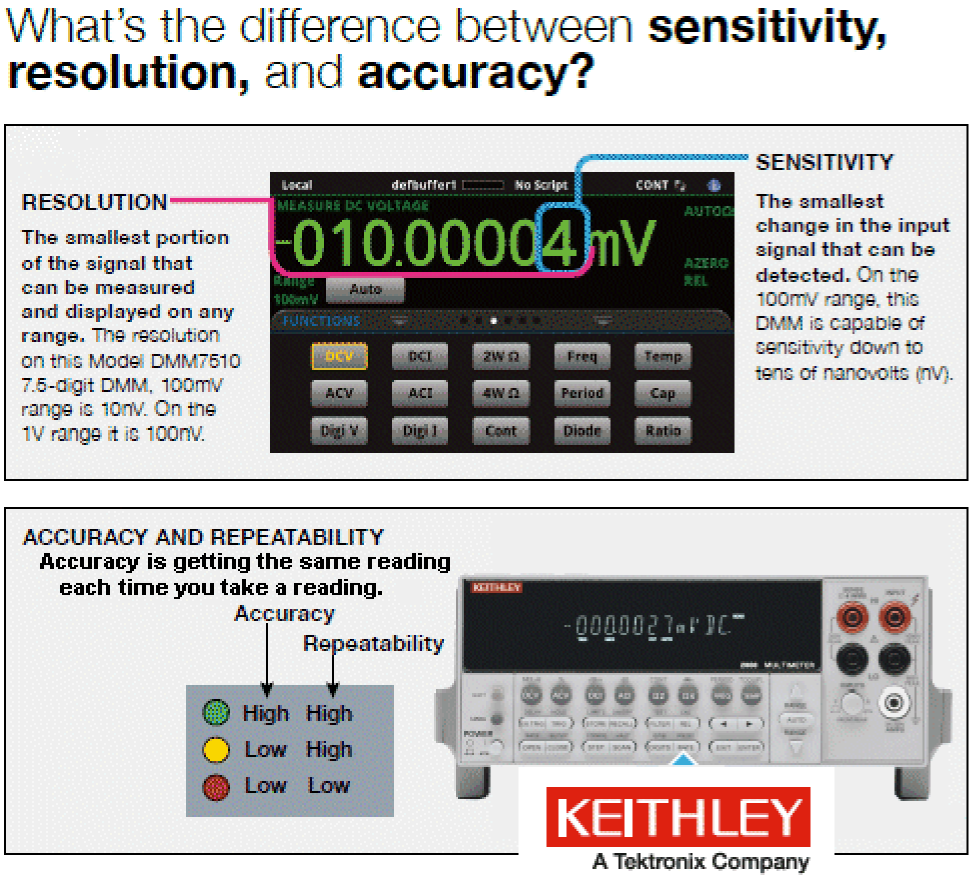
What's the difference between **sensitivity, resolution,** and **accuracy?**

RESOLUTION

The smallest portion of the signal that can be measured and displayed on any range. The resolution on this Model DMM7510 7.5-digit DMM, 100mV range is 10nV. On the 1V range it is 100nV.

SENSITIVITY

The smallest change in the input signal that can be detected. On the 100mV range, this DMM is capable of sensitivity down to tens of nanovolts (nV).

Local
defbuffer1
No Script
CONT
MEASURE DC VOLTAGE
-010.000004mV
AUTOZ
AZERO
REL
Range 100mV
Auto
FUNCTIONS
DCV
DCI
2W Ω
Freq
Temp
ACV
ACI
4W Ω
Period
Cap
Digi V
Digi I
Cont
Diode
Ratio

ACCURACY AND REPEATABILITY

Accuracy is getting the same reading each time you take a reading.

	Accuracy	Repeatability
🟢	High	High
🟡	Low	High
🔴	Low	Low

KEITHLEY
-0.00002 mV DC
2000 MULTIMETER

KEITHLEY
A Tektronix Company

The question above applies to both (every) type of multimeter and the type of meter you use depends on the accuracy you need. Sometimes you are looking for 1mV change on a 20v rail. Only a DMM will (or a CRO) will produce a result.

Analogue meters have an "Ohms Adjustment" to allow for the change in voltage of the battery inside the meter (as it gets old).

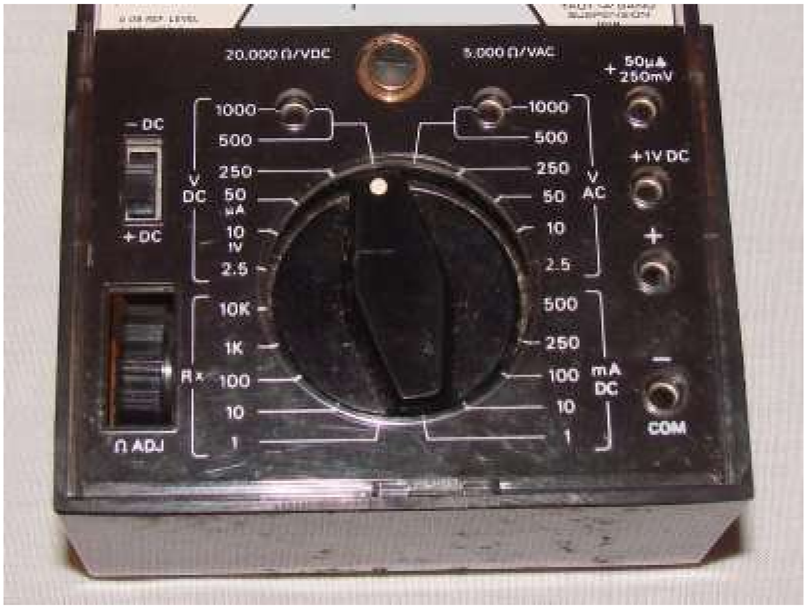

20,000 Ω/VDC
5,000 Ω/VAC
+ 50μA 250mV
+1V DC
+
−
COM
− DC
+ DC
Ω ADJ
V DC
1000
500
250
50 μA
10 1V
2.5
10K
1K
R×
100
10
1
V AC
mA DC

"Ohms Adjust" is also called "ZERO SET"

The sensitivity of this meter is 20,000ohms/volt on the DC ranges and 5k/v on the AC ranges

Before taking a resistance reading (each time on any of the Ohms scales) you need to "ZERO SET" the scale, by touching the two probes together and adjust the pot until the needle reads "0" (swings FULL SCALE). If the pointer does not reach full scale, the batteries need replacing. Digital multimeters do not need "zero adjustment."

FIXING A MULTIMETER

A multimeter can get "broken" "damaged" and go "faulty."
I don't know why, but eventually they stop working.
It can be something simple like a flat battery, corroded battery contacts, broken switch or something complex, like the circuitry failing.
Multimeters are so cheap, you can buy a new one for less than $10.00
These meters can have a 10 amp range, transistor tester and measure up to 2 meg ohms.
That's why I suggest buying a $10.00 meter. They are just as good as a $60.00 meter and the cheapest meters last the longest.
Dropping an analogue meter can cause the hair spring to loop over one of the supports and the needle will not zero correctly. You will need to open the cover on the movement and lift the spring off the support with a needle.
A faulty meter can be used in a battery-charger circuit to measure the current or voltage if that scale is still reading-correctly.
Otherwise keep the leads and throw the meter out. It is too dangerous keeping a meter that shows an incorrect reading.

MEASURING FREQUENCY

VICHY
VC99
APO
DC
-000.0
AUTO
mV
HOLD
RANGE
REL
DC/AC
Hz
hFE
V~
V⎓
OFF
AUTO OFF POWER
A
mA
COM
VΩHz
TEMP
CAT II 1000V

Before we cover the normal uses for a multimeter, it is interesting to note that some **Digital Multimeters (DMM)** have features such as Capacitance, Frequency and measuring the gain of a transistor as well as a number of other features using probes such as a temperature probe. The VICHY VC99 meter above is an example and costs about $40.00.

Basic function	Range
DCV	600mV/6V/60V/600V/1000V
ACV	6V/60/600/1000V
DCA	600uA/6000uA/60mA/600mA/6A/20A
ACA	600uA/6000uA/60mA/600mA/6A/20A
Capacitance	40nF/ 400nF/4uF/40uF/400uF/2000uF
Resistance	600Ω/6kΩ/60kΩ/600kΩ/6MΩ/60MΩ
Frequency	100Hz/1kHz/10kHz/100kHz/1MHz/60MHz

Temperature	-40°C~1000°C
	0°F~1832°F

MEASURING VOLTAGE

Most of the readings you will take with a multimeter will be VOLTAGE readings.
Before taking a reading, you should select the highest range and if the needle does not move up scale (to the right), you can select another range.
Always switch to the highest range before probing a circuit and keep your fingers away from the component being tested.
If the meter is Digital, select the highest range or use the auto-ranging feature, by selecting "V." The meter will automatically produce a result, even if the voltage is AC or DC.
If the meter is not auto-ranging, you will have to select V⎓ if the voltage is from a DC source or V~ if the voltage is from an AC source. DC means Direct Current and the voltage is coming from a battery or supply where the voltage is steady and not changing and AC means Alternating Current where the voltage is coming from a voltage that is rising and falling.
You can measure the voltage at different points in a circuit by connecting the black probe to chassis. This is the 0v reference and is commonly called "Chassis" or "Earth" or "Ground" or "0v."
The red lead is called the "measuring lead" or "measuring probe" and it can measure voltages at any point in a circuit. Sometimes there are "test points" on a circuit and these are wires or loops designed to hold the tip of the red probe (or a red probe fitted with a mini clip or mini alligator clip).
You can also measure voltages ACROSS A COMPONENT. In other words, the reading is taken in PARALLEL with the component. It may be the voltage across a transistor, resistor, capacitor, diode or coil. In most cases this voltage will be less than the supply voltage.
If you are measuring the voltage in a circuit that has a HIGH IMPEDANCE, the reading will be inaccurate, up to 90% !!!, if you use a cheap analogue meter.

Here's a simple case.

The circuit below consists of two 1M resistors in series. The voltage at the mid point will be 5v when nothing is connected to the mid point. But if we use a cheap analogue multimeter set to 10v, the resistance of the meter will be about 100k, if the meter has a sensitivity of 10k/v and the reading will be incorrect.
Here how it works:
Every meter has a sensitivity. The sensitivity of the meter is the sensitivity of the movement and is the amount of current required to deflect the needle FULL SCALE. This current is very small, normally 1/10th of a milliamp and corresponds to a sensitivity of 10k/volt (or 1/30th mA, for a sensitivity of 30k/v).
If an analogue meter is set to 10v, the internal resistance of the meter will be 100k for a 10k/v movement.
If this multimeter is used to test the following circuit, the reading will be inaccurate.
The reading should be 5v as show in diagram **A**.
But the analogue multimeter has an internal resistance of 100k and it creates a circuit shown in **C**.
The top 1M and 100k from the meter create a combined PARALLEL resistance of 90k.
This forms a series circuit with the lower 1M and the meter will read less than 1v
If we measure the voltage across the lower 1M, the 100k meter will form a value of resistance with the lower 1M and it will read less than 1v
If the multimeter is 30k/v, the readings will be 2v. See how easy it is to get a totally inaccurate reading.

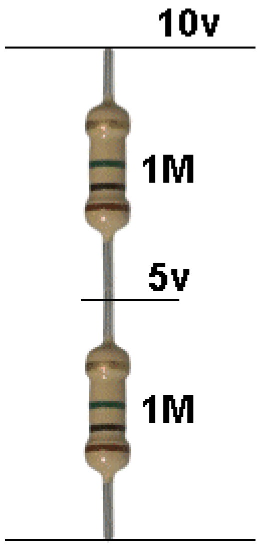

A

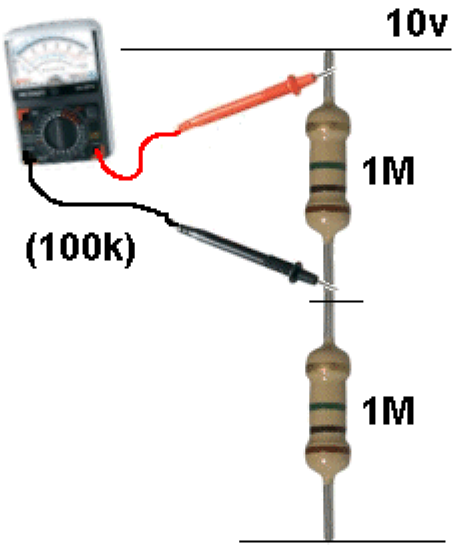

B

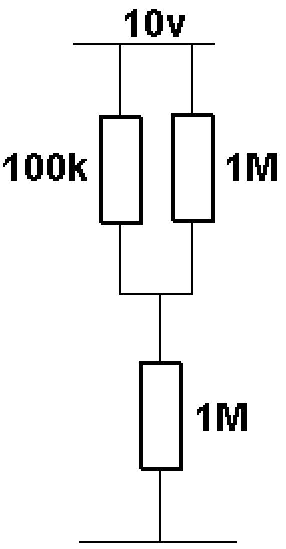

C

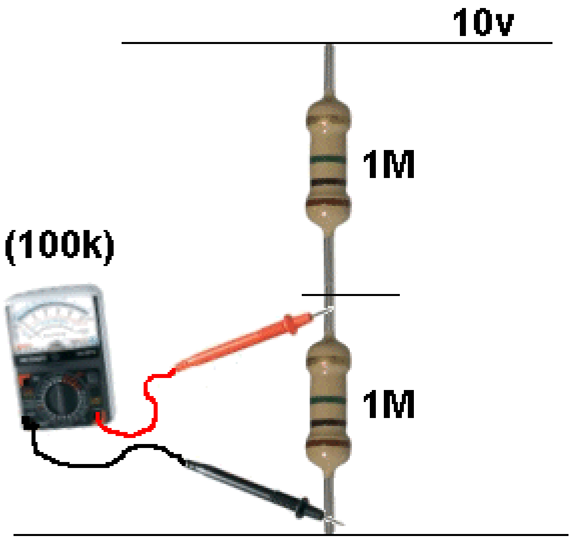

D

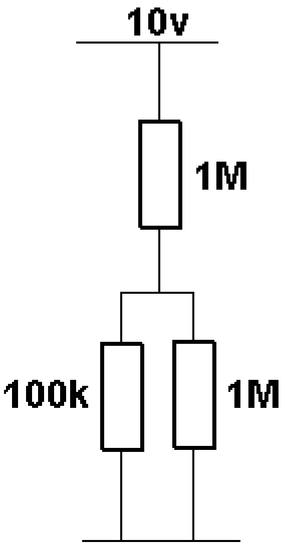

E

This introduces two new terms:
HIGH IMPEDANCE CIRCUIT and "RESISTORS in SERIES and PARALLEL."

If the reading is taken with a Digital Meter, it will be more accurate as a DMM does not take any current from the circuit (to activate the meter). In other words it has a very HIGH input impedance. Most Digital Multimeters have a fixed input resistance (impedance) of 10M - no matter what scale is selected. That's the reason for choosing a DMM for high impedance circuits. It also gives a reading that is accurate to about 1%.

MEASURING VOLTAGES IN A CIRCUIT

You can take many voltage-measurements in a circuit. You can measure "across" a component, or between any point in a circuit and either the positive rail or earth rail (0v rail). In the following circuit, the 5 most important voltage-measurements are shown. Voltage "A" is across the electret microphone. It should be between 20mV and 500mV. Voltage "B" should be about 0.6v. Voltage "C" should be about half-rail voltage. This allows the transistor to amplify both the positive and negative parts of the waveform. Voltage "D" should be about 1-3v. Voltage "E" should be the battery voltage of 12v.

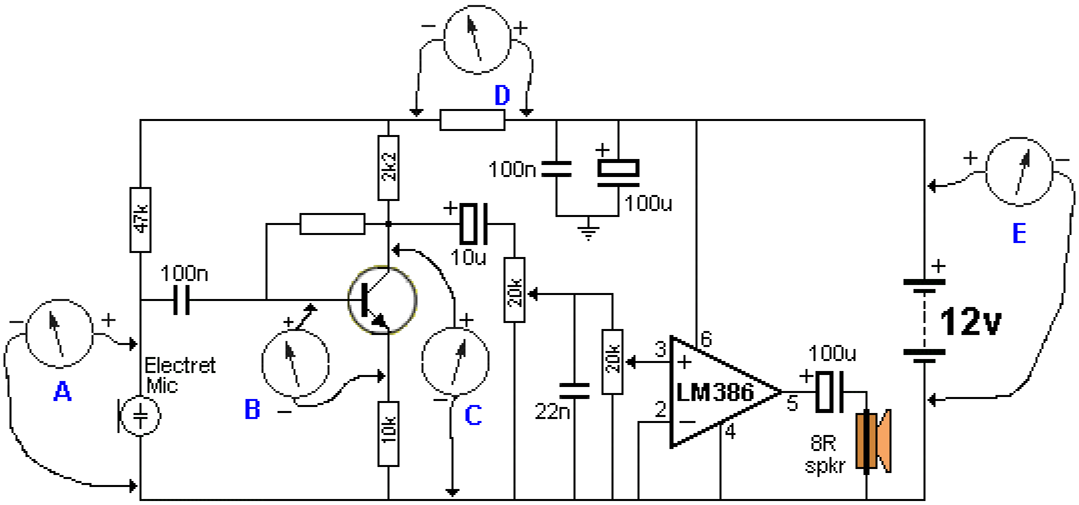

MEASURING VOLTAGES IN A CIRCUIT

MEASURING CURRENT

You will rarely need to take current measurements, however most multimeters have DC current ranges such as 0.5mA, 50mA, 500mA and 10Amp (via the extra banana socket) and some meters have AC current ranges. Measuring the current of a circuit will tell you a lot of things. If you know the normal current, a high or low current can let you know if the circuit is overloaded or not fully operational.

Current is always measured when the circuit is working (i.e: with power applied).
It is measured IN SERIES with the circuit or component under test.
The easiest way to measure current is to remove the fuse and take a reading across the fuse-holder. Or remove one lead of the battery or turn the project off, and measure across the switch.
If this is not possible, you will need to remove one end of a component and measure with the two probes in the "opening."
Resistors are the easiest things to desolder, but you may have to cut a track in some circuits. You have to get an "opening" so that a current reading can be taken.
The following diagrams show how to connect the probes to take a CURRENT reading.
Do not measure the current ACROSS a component as this will create a "short-circuit."
The component is designed to drop a certain voltage and when you place the probes across this component, you are effectively adding a "link" or "jumper" and the voltage at the left-side of the component will appear on the right-side. This voltage may be too high for the circuit being supplied and the result will be damage.

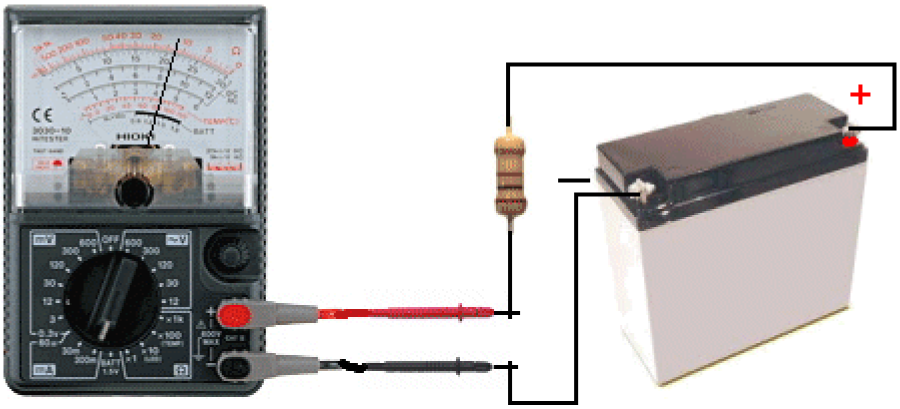

Measuring current through a resistor

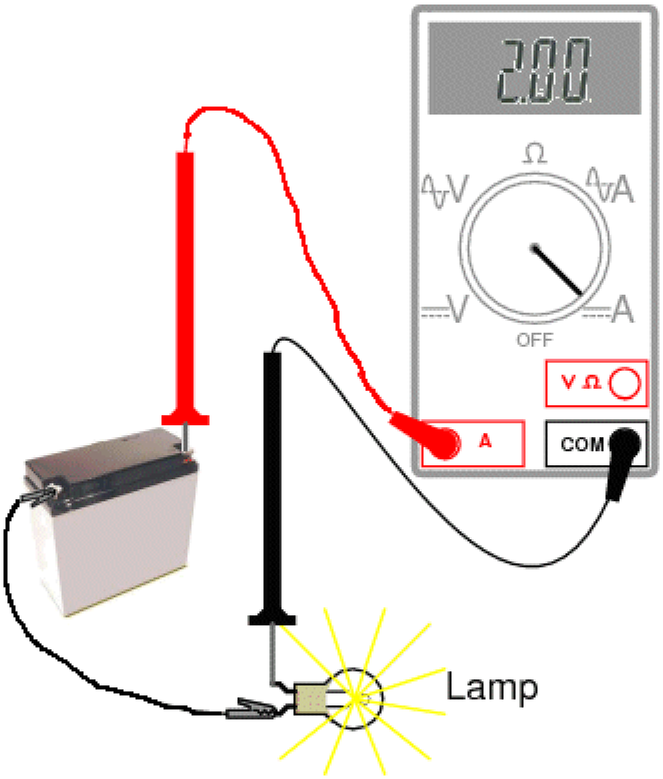

Measuring the current of a globe

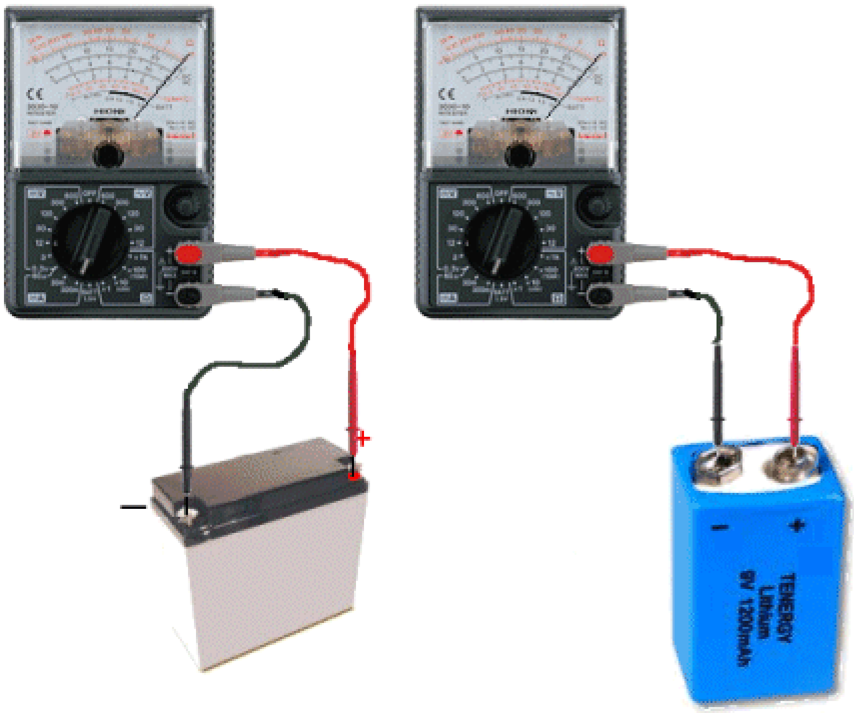

Do NOT measure the CURRENT of a battery
(by placing the meter directly across the terminals)
A battery will deliver a very HIGH current
and damage the meter

Do not measure the "current a battery will deliver" by placing the probes across the terminals. It will deliver a very high current and damage the meter instantly. There are special battery testing instruments for this purpose.
When measuring across an "opening" or "cut," place the red probe on the wire that supplies the voltage (and current) and the black probe on the other wire. This will produce a "POSITIVE" reading.

A positive reading is an UPSCALE READING and the pointer will move across the scale - to the right. A "NEGATIVE READING" will make the pointer hit the "STOP" at the left of the scale and you will not get a reading. If you are using a Digital Meter, a negative sign "-" will appear on the screen to indicate the probes are around the wrong way. No damage will be caused. It just indicates the probes are connected incorrectly.

If you want an accurate CURRENT MEASUREMENT, use a digital meter.

MEASURING RESISTANCE

Turn a circuit off before measuring resistance.

If any voltage is present, the value of resistance will be incorrect.

In most cases you cannot measure a component while it is in-circuit. This is because the meter is actually measuring a voltage across a component and calling it a "resistance." The voltage comes from the battery inside the meter. If any other voltage is present, the meter will produce a false reading.

If you are measuring the resistance of a component while still "in circuit," (with the power off) the reading will be lower than the true reading.

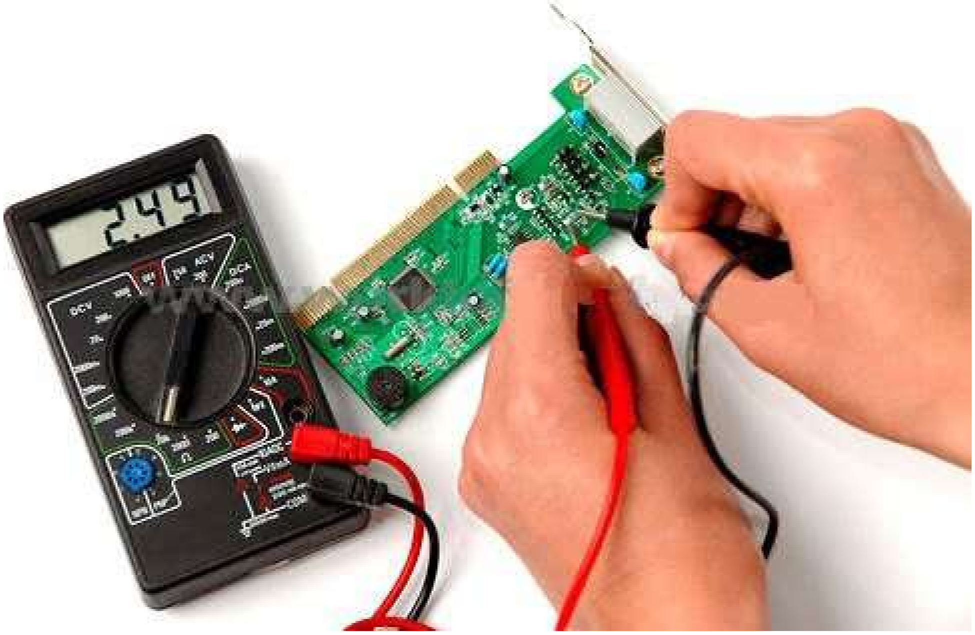

Measuring resistance

138
rolson
27259

Measuring resistance of a heater
(via the leads)

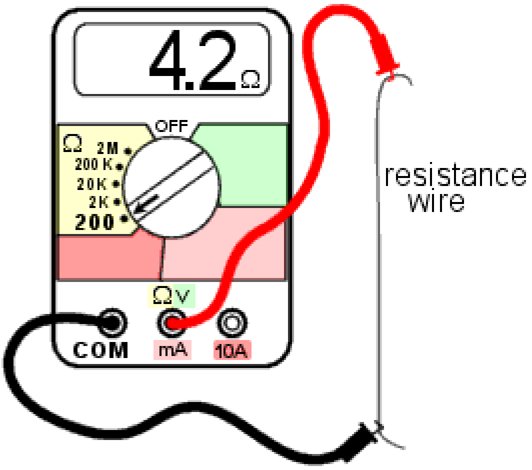
4.2 Ω
OFF
Ω 2M
200 K
20 K
2 K
200
ΩV
COM
mA
10A
resistance wire

Measuring the resistance of a piece of resistance-wire

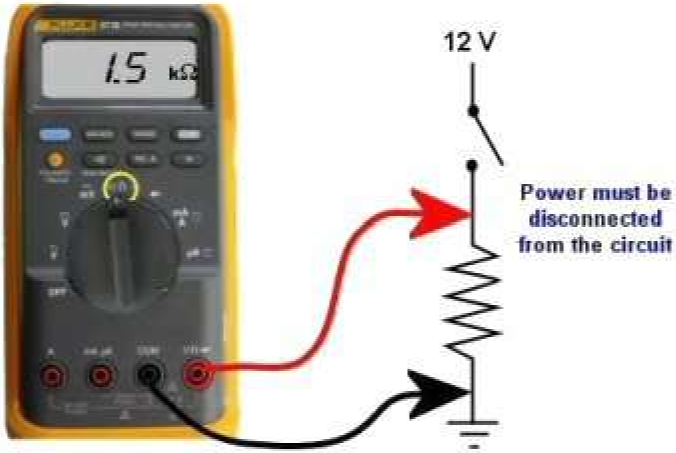

Power must be disconnected from the circuit

Measuring the resistance of a resistor

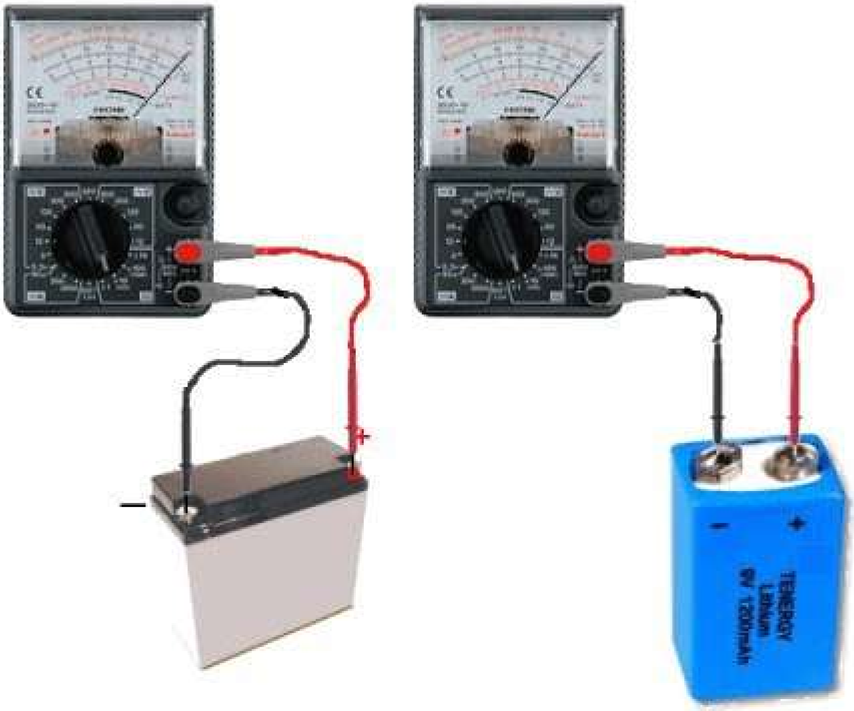
+
−
-
+
TENERGY
Lithium
9V 1200mAh

Do not measure the "Resistance of a Battery"

1. Do not measure the "resistance of a battery." The resistance of a battery (called the Internal impedance) is not measured as shown in the diagrams above. It is measured by creating a current-flow and measuring the voltage across the battery.

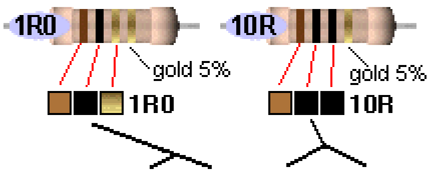

Placing a multimeter set to **resistance** (across a battery) will destroy the meter.
2. Do not try to measure the resistance of any voltage or any "supply."

Resistance is measured in OHMs.
The resistance of a 1cm x 1cm bar, one metre long is 1 ohm.
If the bar is thinner, the resistance is higher. If the bar is longer, the resistance is higher.
If the material of the bar is changed, the resistance is higher.
When carbon is mixed with other elements, its resistance increases and this knowledge is used to make RESISTORS.
Resistors have RESISTANCE and the main purpose of a resistor is to reduce the CURRENT FLOW.
It's a bit like standing on a hose. The flow reduces.
When current flow is reduced, the output voltage is also reduced and that why the water does not spray up so high. Resistors are simple devices but they produce many different effects in a circuit.
A resistor of nearly pure carbon may be 1 ohm, but when non-conducting "impurities" are added, the same-size resistor may be 100 ohms, 1,000 ohms or 1 million ohms.
Circuits use values of less than 1 ohm to more than 22 million ohms.

Resistors are identified on a circuit with numbers and letters to show the exact value of resistance - such as 1k 2k2 4M7
The letter Ω (omega - a Greek symbol) is used to identify the word "Ohm." but this symbol is not available on some word-processors, so the letter "R" is used.
The letter "E" is also sometimes used and both mean "Ohms."
A one-ohm resistor is written "1R" or "1E." It can also be written "1R0" or "1E0."
A resistor of one-tenth of an ohm is written "0R1" or "0E1." The letter takes the place of the decimal point.
10 ohms = 10R
100 ohms = 100R
1,000 ohms = 1k (k= kilo = one thousand)
10,000 ohms = 10k
100,000 ohms = 100k
1,000,000 ohms = 1M (M = MEG = one million)

The size of a resistor has nothing to do with its resistance. The size determines the wattage of the resistor - how much heat it can dissipate without getting too hot.
Every resistor is identified by colour bands on the body, but when the resistor is a surface-mount device, numbers are used and sometimes letters.
You MUST learn the colour code for resistors and the following table shows all the colours for the most common resistors from 1/10th of an ohm to 22 Meg ohms for resistors with 5% and 10% tolerance.

If 3rd band is gold, Divide by 10
If 3rd band is silver, Divide by 100
(to get 0.22ohms etc)

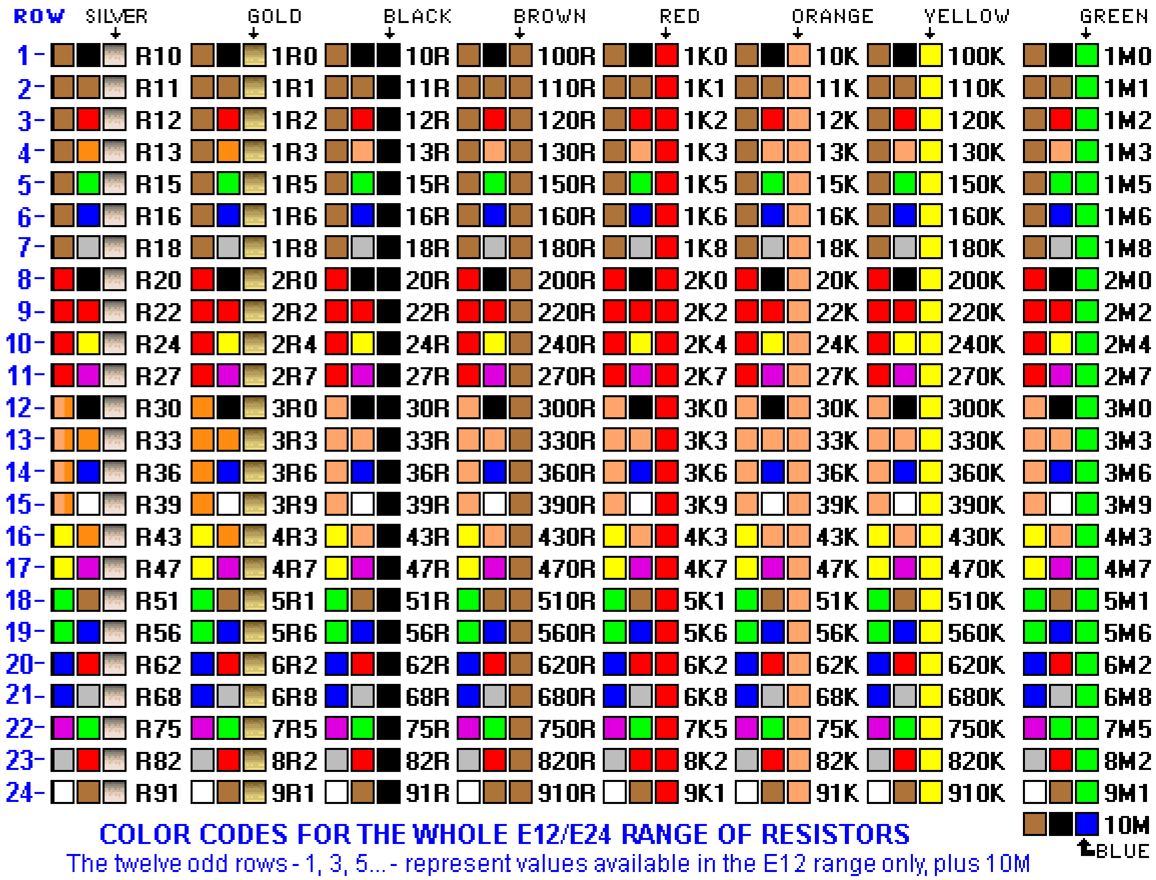

ROW	SILVER	GOLD	BLACK	BROWN	RED	ORANGE	YELLOW	GREEN
1-	R10	1R0	10R	100R	1K0	10K	100K	1M0
2-	R11	1R1	11R	110R	1K1	11K	110K	1M1
3-	R12	1R2	12R	120R	1K2	12K	120K	1M2
4-	R13	1R3	13R	130R	1K3	13K	130K	1M3
5-	R15	1R5	15R	150R	1K5	15K	150K	1M5
6-	R16	1R6	16R	160R	1K6	16K	160K	1M6
7-	R18	1R8	18R	180R	1K8	18K	180K	1M8
8-	R20	2R0	20R	200R	2K0	20K	200K	2M0
9-	R22	2R2	22R	220R	2K2	22K	220K	2M2
10-	R24	2R4	24R	240R	2K4	24K	240K	2M4
11-	R27	2R7	27R	270R	2K7	27K	270K	2M7
12-	R30	3R0	30R	300R	3K0	30K	300K	3M0
13-	R33	3R3	33R	330R	3K3	33K	330K	3M3
14-	R36	3R6	36R	360R	3K6	36K	360K	3M6
15-	R39	3R9	39R	390R	3K9	39K	390K	3M9
16-	R43	4R3	43R	430R	4K3	43K	430K	4M3
17-	R47	4R7	47R	470R	4K7	47K	470K	4M7
18-	R51	5R1	51R	510R	5K1	51K	510K	5M1
19-	R56	5R6	56R	560R	5K6	56K	560K	5M6
20-	R62	6R2	62R	620R	6K2	62K	620K	6M2
21-	R68	6R8	68R	680R	6K8	68K	680K	6M8
22-	R75	7R5	75R	750R	7K5	75K	750K	7M5
23-	R82	8R2	82R	820R	8K2	82K	820K	8M2
24-	R91	9R1	91R	910R	9K1	91K	910K	9M1

10M
BLUE

COLOR CODES FOR THE WHOLE E12/E24 RANGE OF RESISTORS

The twelve odd rows - 1, 3, 5... - represent values available in the E12 range only, plus 10M

Reading 4-band resistors

The most "common" type of resistor has 4 bands and is called the 10% resistor. It now has a tolerance of 5% but is still called the "10% type" as the colours increase by 20% so that a resistor can be 10% higher or 10% lower than a particular value and all the resistors produced in a batch can be used.

The first 3 bands produce the resistance and the fourth band is the "tolerance" band.

Gold = 5%

(Silver =10% but no modern resistors are 10%!! - they are 5% 2% or 1%)

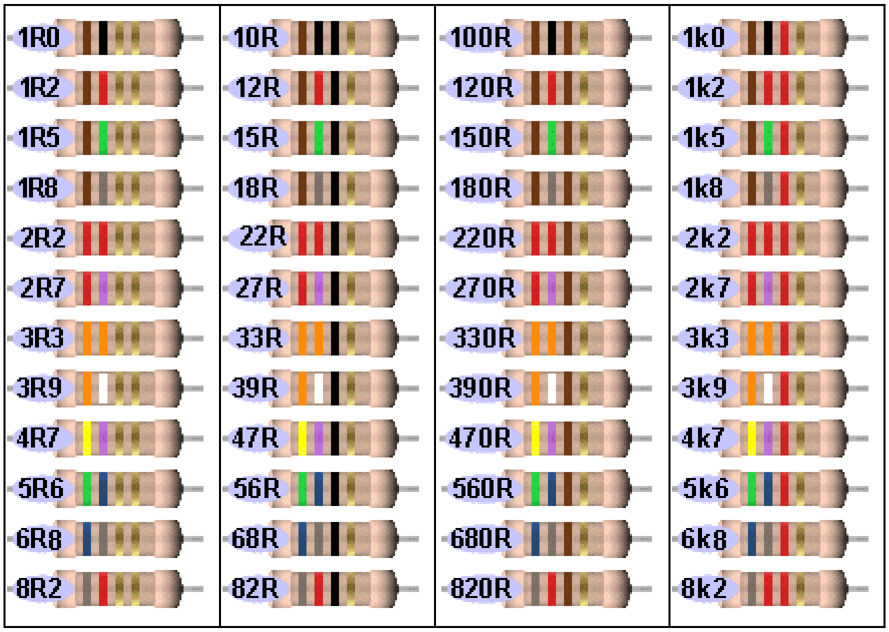
1R0
10R
100R
1k0
1R2
12R
120R
1k2
1R5
15R
150R
1k5
1R8
18R
180R
1k8
2R2
22R
220R
2k2
2R7
27R
270R
2k7
3R3
33R
330R
3k3
3R9
39R
390R
3k9
4R7
47R
470R
4k7
5R6
56R
560R
5k6
6R8
68R
680R
6k8
8R2
82R
820R
8k2

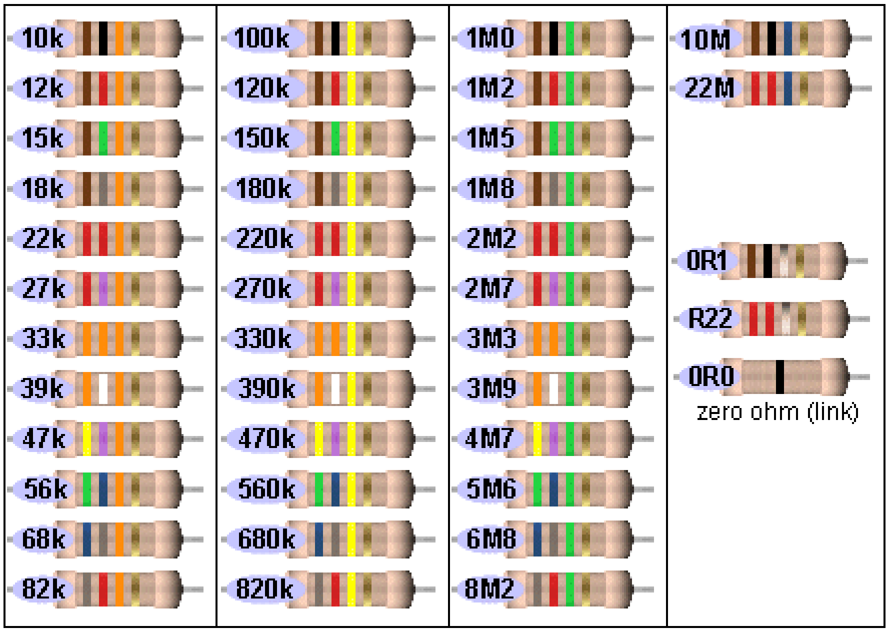
10k
12k
15k
18k
22k
27k
33k
39k
47k
56k
68k
82k
100k
120k
150k
180k
220k
270k
330k
390k
470k
560k
680k
820k
1M0
1M2
1M5
1M8
2M2
2M7
3M3
3M9
4M7
5M6
6M8
8M2
10M
22M
0R1
R22
0R0
zero ohm (link)

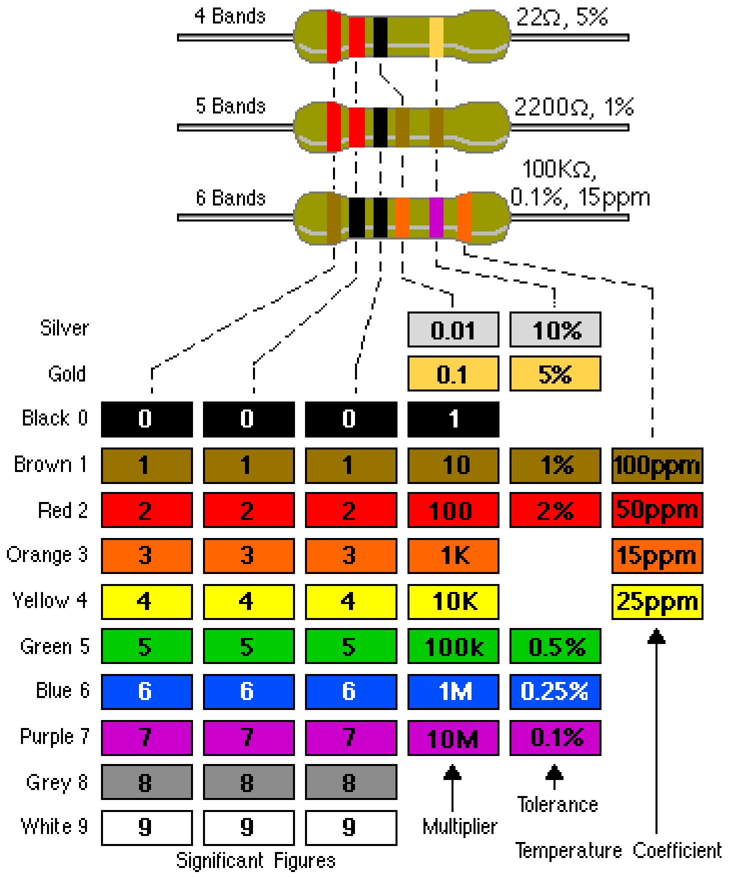
4 Bands
22Ω, 5%
5 Bands
2200Ω, 1%
6 Bands
100KΩ,
0.1%, 15ppm

	Significant Figures			Multiplier	Tolerance	Temperature Coefficient
Silver				0.01	10%	
Gold				0.1	5%	
Black 0	0	0	0	1		
Brown 1	1	1	1	10	1%	100ppm
Red 2	2	2	2	100	2%	50ppm
Orange 3	3	3	3	1K		15ppm
Yellow 4	4	4	4	10K		25ppm
Green 5	5	5	5	100k	0.5%	
Blue 6	6	6	6	1M	0.25%	
Purple 7	7	7	7	10M	0.1%	
Grey 8	8	8	8			
White 9	9	9	9			

Resistor Color Code System

Here is another well-designed resistor colour code chart:

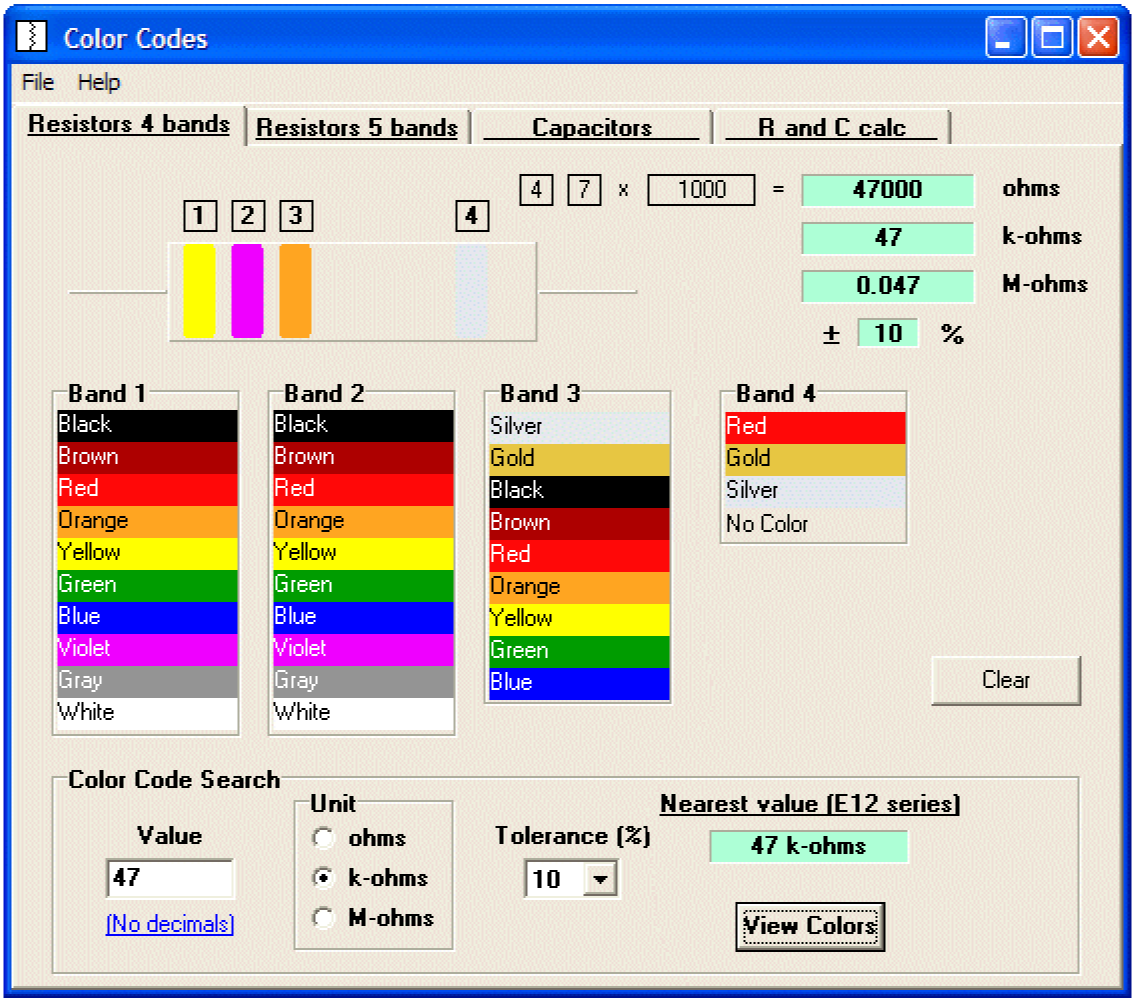
Color Codes
File
Help
Resistors 4 bands
Resistors 5 bands
Capacitors
R and C calc
1
2
3
4
Band 1
Black
Brown
Red
Orange
Yellow
Green
Blue
Violet
Gray
White
Band 2
Black
Brown
Red
Orange
Yellow
Green
Blue
Violet
Gray
White
Band 3
Silver
Gold
Black
Brown
Red
Orange
Yellow
Green
Blue
Band 4
Red
Gold
Silver
No Color
Clear
Color Code Search
View Colors

4 7 × 1000 = 47000 ohms
47 k-ohms
0.047 M-ohms
± 10 %

Value: 47 (No decimals)
Unit: ○ ohms ● k-ohms ○ M-ohms
Tolerance (%): 10
Nearest value (E12 series): 47 k-ohms

Download the program and save it on your desk-top for future reference:
ColourCode.exe (520KB)
ColourCode.zip (230KB)
ColourCode.rar (180KB)

RESISTORS LESS THAN 10 OHMS

When the **third** band is gold, it indicates the value of the "colors" must be divided by 10.
Gold = "divide by 10" to get values 1R0 to 8R2
When the **third** band is silver, it indicates the value of the "colors" must be divided by 100. (Remember: more letters in the word "silver" thus the divisor is "a larger division.")
Silver = "divide by 100" to get values R1 to R82
e.g: 0R1 = 0.1 ohm   0R22 = point 22 ohms
See 4th Column above for examples.

The letters "R, k and M" take the place of a decimal point.
e.g: 1**R**0 = 1 ohm   2**R**2 = 2 point 2 ohms   22**R** = 22 ohms
2**k**2 = 2,200 ohms   100**k** = 100,000 ohms
2**M**2 = 2,200,000 ohms

HOW TO REMEMBER THE COLOUR CODE:

Each colour has a "number" (or divisor) corresponding to it.
Most of the colours are in the same order as in the spectrum. You can see the spectrum in a rainbow. It is: ROY G BIV and the colours for resistors are in the same sequence.

black
brown - colour of increasing temperature

red
orange
yellow
green
blue
(indigo - that part of the spectrum between blue and violet)
violet
gray
white

colour	value	No of zero's
silver	-2	divide by 100
gold	-1	divide by 10
black	0	No zeros
brown	1	0
red	2	00
orange	3	,000 or k
yellow	4	0,000
green	5	00,000
blue	6	M
violet	7	
gray	8	
white	9	

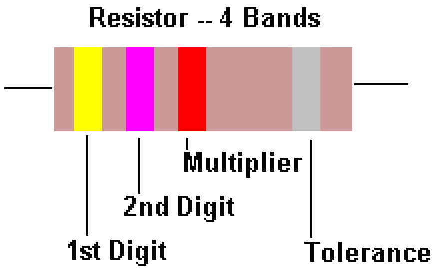
Resistor -- 4 Bands
Multiplier
2nd Digit
1st Digit
Tolerance

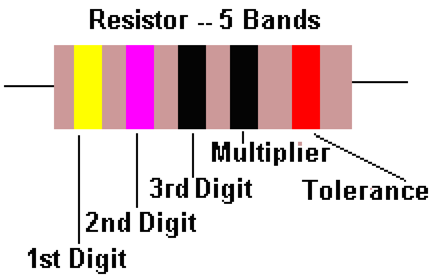
Resistor -- 5 Bands
Multiplier
3rd Digit
Tolerance
2nd Digit
1st Digit

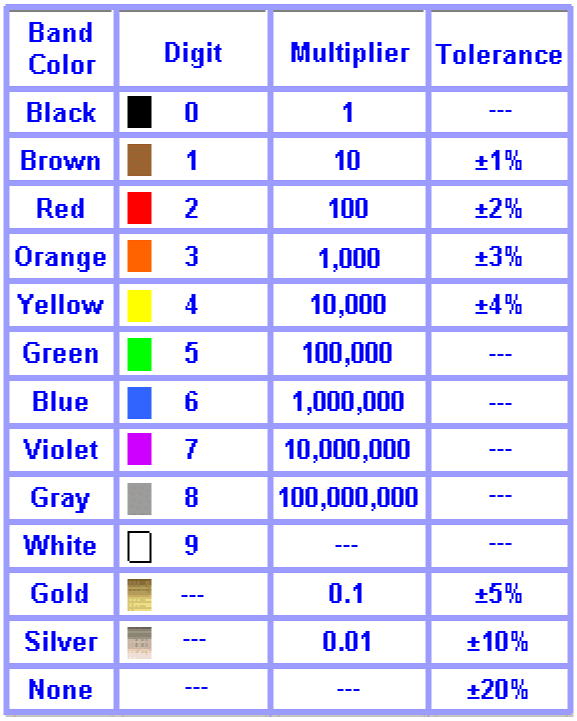

Band Color	Digit	Multiplier	Tolerance
Black	⬛ 0	1	---
Brown	🟫 1	10	±1%
Red	🟥 2	100	±2%
Orange	🟧 3	1,000	±3%
Yellow	🟨 4	10,000	±4%
Green	🟩 5	100,000	---
Blue	🟦 6	1,000,000	---
Violet	🟪 7	10,000,000	---
Gray	▨ 8	100,000,000	---
White	⬜ 9	---	---
Gold	▨ ---	0.1	±5%
Silver	▨ ---	0.01	±10%
None	---	---	±20%

Here are some common ways to remember the colour code:

Bad Beer Rots Our Young Guts, But Vodka Goes Well
Bright Boys Rave Over Young Girls But Violet Gets Wed
Bad Boys Rave Over Young Girls But Violet Gets Wed with Gold and Silver.

Reading 5-band resistors:

5-band resistors are easy to read if you remember two simple points. The first three bands provide the digits in the answer and the 4th band supplies the number of zero's.

Reading "STANDARD VALUES" (**on 5-band resistors**)

5-band resistors are also made in "Standard Values" but will have different colours to 4-band "common" resistors - and will be confusing if you are just starting out. For instance, a 47k 5% resistor with 4-bands will be: yellow-purple-orange-gold. For a 47k 1% resistor the colours will be yellow-purple-black-red-brown. The brown colour-band represents 1%.
The first two colour-bands for a STANDARD VALUE or "common value" in 1% or 5% will be the SAME. These two bands provide the digits in the answer.
It's the 3rd band for a 5% resistor that is expanded into two bands in a 1% resistor. But it's easy to follow.
For a standard value, the 3rd band in a 1% resistor is BLACK. This represents a ZERO in the answer. (For 5-band resistors BLACK represents a ZERO when in the third band. This is different to 4-band resistors where black represents the word OHMS! If the third band is BROWN, the answer will be 1).
So the 4th band has to represent one-less ZERO and is one colour UP THE COLOUR CHART! In other words the 3rd and 4th bands (combined) on a 1% resistor produces the same number of zero's as the 3rd band on a 5% resistor!

Resistors come in a range of values and the two most common are the E12 and E24 series. The E12 series comes in twelve values for each decade. The E24 series comes in twenty-four values per decade.

E12 series - 10, 12, 15, 18, 22, 27, 33, 39, 47, 56, 68, 82

E24 series - 10, 11, 12, 13, 15, 16, 18, 20, 22, 24, 27, 30, 33, 36, 39, 43, 47, 51, 56, 62, 68, 75, 82, 91

Here is the complete list of 1% 1/4watt resistors from:
CIRCUIT SPECIALISTS. The following list covers 10 ohms (10R) to 1M.
To buy 1% resistors from Circuit Specialists, click: HERE.

10R
12R1
15R
18R2
22R1
27R4
30R1
33R2
36R5
39R2
47R5
49R9
51R1
56R2
68R1
75R
82R5
90R9
100R
121R
150R
182R
200R
221R
240R
249R
274R
301R
332R
348R
392R
402R
475R
499R
565R
604R
681R
750R
806R
825R
909R
1k0
1k21
1k5
1k82
2k
2k21
2k2
2k43
2k49
2k67
2k74
3k01
3k32
3k48
3k57
3k74
3k83
3k92
4k02
4k22
4k64
4k75
4k7
4k87
4k99
5k11
5k23
5k36
5k49
5k62
5k76
5k9
6k04
6k19
6k81
7k15
7k5
7k87
71k5
8k06
8k25
8k45
8k66
8k87
9k09
9k31
9k53
9k76
10k
11k
12k
12k1
12k4
13k
14k7
15k
15k8
16k9
17k4
17k8
18k2
20k
22k1
22k6
23k7
24k9
27k4
29k4
30k1
33k2
34k8
36k5
38k3
39k2
40k2
44k2
46k4
47k
47k5
49k9
51k1
53k6
56k2
61k9
68k1
69k8
75k0
82k5
90k
90k9
95k3
100k
121k
147k
150k
182k
200k
212k
221k
226k
249k
274k
301k
332k
357k
392k
475k
487k
499k
562k
604k
1M

Here is the list of 1% resistors from suppliers (such as Farnell):

1R0	13R	68R	360R	1k8	9k1	47k	240k
1R2	15R	75R	390R	2k0	10k	51k	270k
1R5	16R	82R	430R	2k2	11k	56k	300k
2R2	18R	91R	470R	2k4	12k	62k	330k
2R7	20R	100R	510R	2k7	13k	68k	360k
3R3	22R	110R	560R	3k	15k	75k	390k
3R9	24R	120R	620R	3k3	16k	82k	430k
4R7	27R	130R	680R	3k6	18k	91k	470k
5R6	30R	150R	750R	3k9	20k	100k	510k
6R2	33R	160R	820R	4k3	22k	110k	560k
6R8	36R	180R	910R	4k7	24k	120k	620k
7R5	39R	200R	1k	5k1	27k	130k	680k
8R2	43R	220R	1k1	5k6	30k	150k	750k
9R1	47R	240R	1k2	6k2	33k	160k	820k
10R	51R	270R	1k3	6k8	36k	180k	910k
11R	56R	300R	1k5	7k5	39k	200k	1M
12R	62R	330R	1k6	8k2	43k	220k	

Surface Mount Resistors

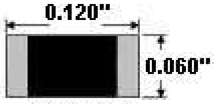

"120 thou"
x
"60 thou"
"one hundred and
twenty thousands
of an inch"

10R
10 ohm

0.047R
0.025R
sense resistor

121k

499R

240R

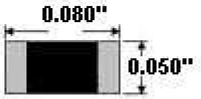

11.8 ohm

10 ohm

474	M47
470 kΩ	470 kΩ

R47	4R7	47R	K47	4K7	47K
0.47 Ω	4.7 Ω	47 Ω	470 Ω	4.7 kΩ	47 kΩ

R464	464R	4K64	471	472	473
0.464 Ω	464 Ω	4.64 kΩ	470 Ω	4.7 kΩ	47 kΩ

0	00	000	0000

SHORT-CIRCUITING "ZERO-OHM LINKS" OR "JUMPERS"

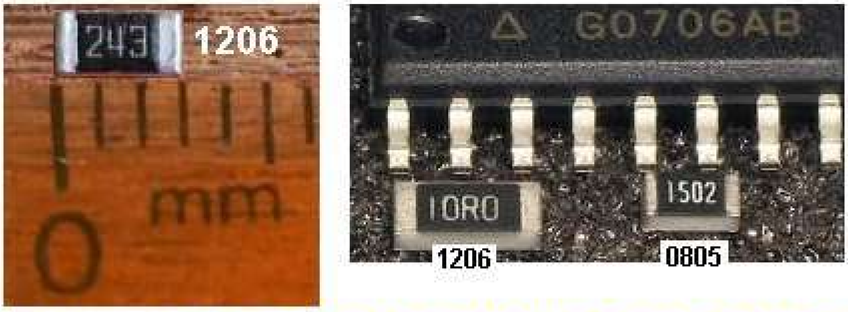

1206 and 0805 SURFACE MOUNT RESISTORS

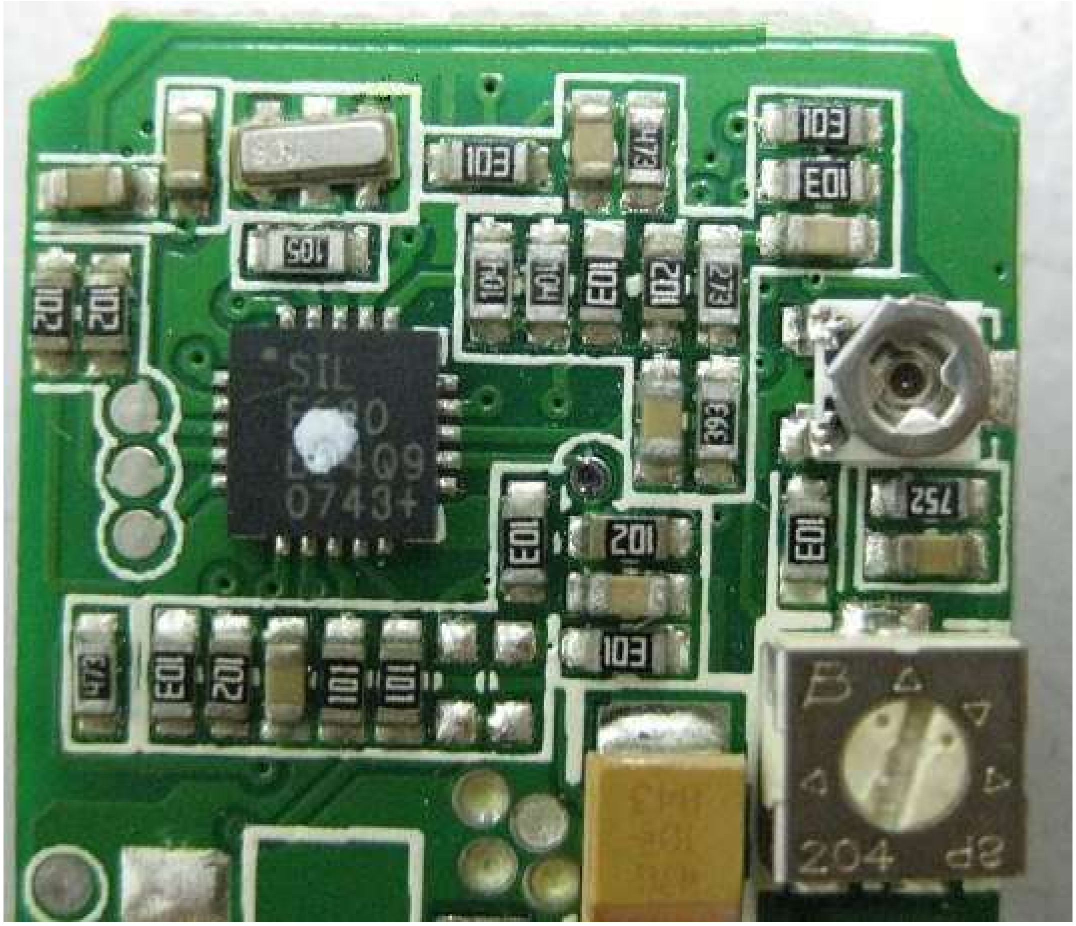
103
105
102
473
393
752
SIL
0743+
B
204

3-digit Surface Mount resistors on a PC board

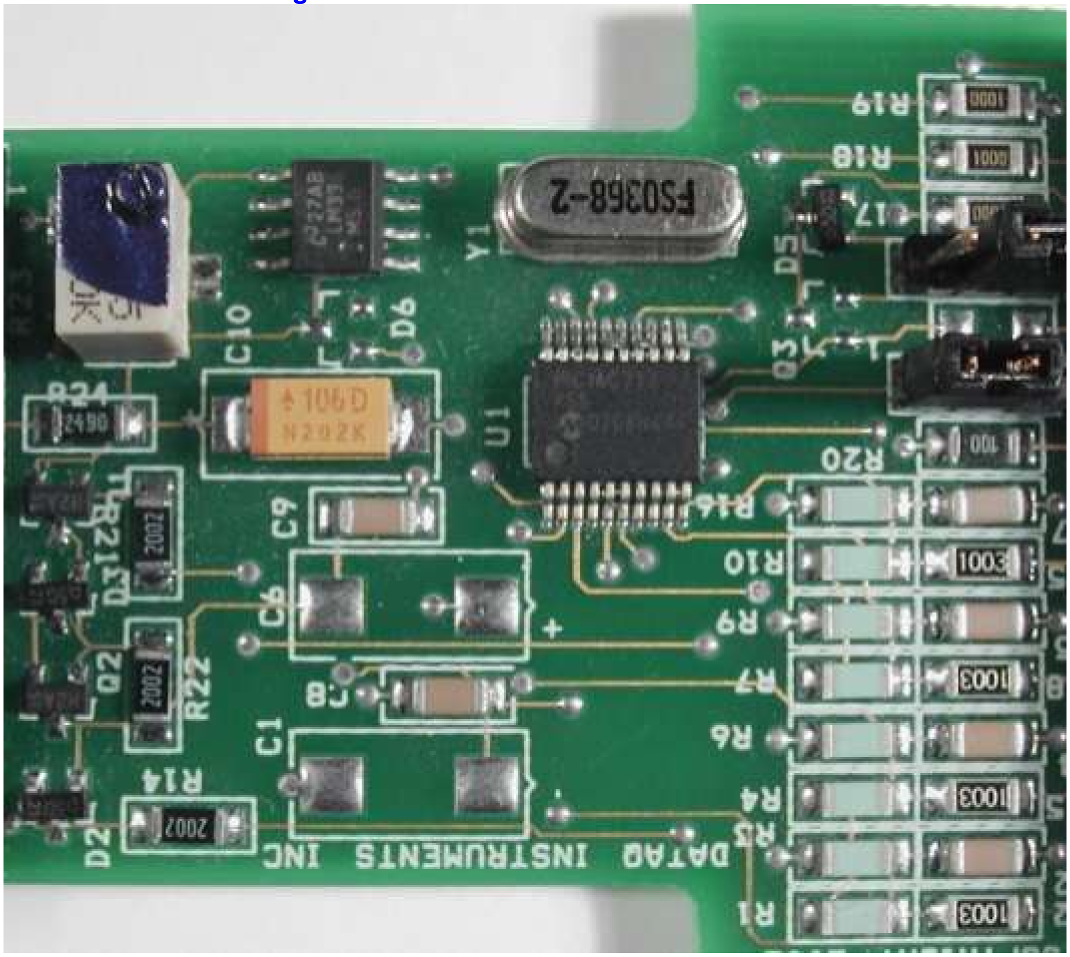
R19
R18
R17
1000
FS0368-2
Y1
U1
D5
D6
Q3
C10
R24
2490
106D
R20
100
R16
R10
1003
R9
R7
R6
R4
R3
R1
C9
C6
C8
C1
R21
R22
Q2
D2
R14
2002
DATAQ INSTRUMENTS INC

4-digit Surface Mount resistors on a PC board

The photo above shows surface mount resistors on a circuit board. The components

that are not marked are capacitors (capacitors are NEVER marked).
All the SM resistors in the above photos conform to a 3-digit or 4-digit code. But there are a number of codes, and the 4-digit code caters for high tolerance resistors, so it's getting very complicated.
Here is a basic 3-digit SM resistor:

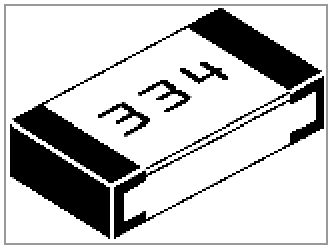

A 330k SM resistor

The first two digits represent the two digits in the answer. The third digit represents the number of zero's you must place after the two digits. The answer will be OHMS. For example: 334 is written 33 0 000. This is written 330,000 ohms. The comma can be replaced by the letter "k". The final answer is: 330k.
222 = 22 00 = 2,200 = 2k2
473 = 47 000 = 47,000 = 47k
474 = 47 0000 = 470,000 = 470k
105 = 10 00000 = 1,000,000 = 1M = one million ohms
There is one trick you have to remember. Resistances less than 100 ohms are written: 100, 220, 470. These are 10 and NO zero's = 10 ohms = 10R or 22 and no zero's = 22R or 47 and no zero's = 47R. Sometimes the resistor is marked: 10, 22 and 47 to prevent a mistake.

Remember:
R = ohms
k = kilo ohms = 1,000 ohms
M = Meg = 1,000,000 ohms
The 3 letters (R, k and M) are put in place of the decimal point. This way you cannot make a mistake when reading a value of resistance.

Surface Mount **CURRENT SENSING** Resistors

Many new types of CURRENT SENSING surface-mount resistors are appearing on the market and these are creating lots of new problems.
Fortunately all resistors are marked with the value of resistance and these resistors are identified in MILLIOHMS. A miili ohm is one thousandth or an ohm and is written 0.001 when writing a normal mathematical number.
When written on a surface mount resistor, the letter R indicates the decimal point and it also signifies the word "OHM" or "OHMS" and one milli-ohm is written R001
Five miiliohms is R005 and one hundred milliohms is R100
Some surface mount resistors have the letter "M" after the value to indicate the resistor has a rating of 1 watt. e.g: R100M These surface-mount resistors are specially-made to withstand a high temperature and a surface-mount resistor of the same size is normally 250mW or less.
These current-sensing resistors can get extremely hot and the PC board can become burnt or damaged.
When designing a PC board, make the lands very large to dissipate the heat.
Normally a current sensing resistor is below one ohm (1R0) and it is easy to identify them as R100 etc.
You cannot measure the value of a current sensing resistor as the leads of a multimeter have a higher resistance than the resistor and few multimeters can read values below one ohm.
If the value is not visible, you will have to refer to the circuit.
Before replacing it, work out why it failed.
Generally it gets too hot. Use a larger size and add tiny heatsinks on each end.
Here are some surface=mount current-sense resistors:

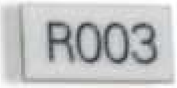

3 milli-ohm

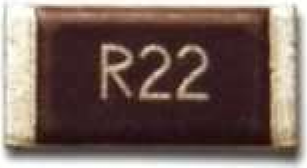

220 milli-ohm

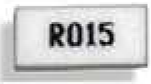

15 milli-ohm

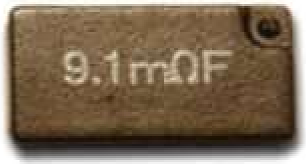

9 point 1 milli-ohm

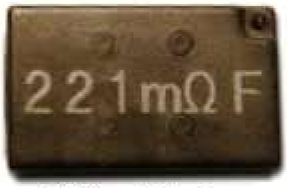

220 milli-ohm

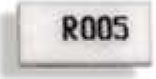

5 milli-ohm

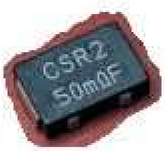

50 milli-ohm 1%

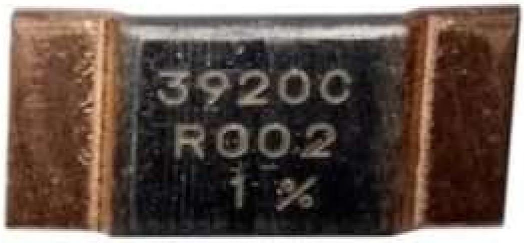
3920C
R002
1%

2 milli-ohm

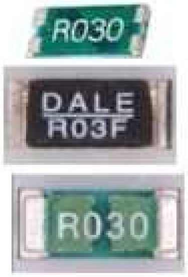
R030
DALE
R03F
R030

30 milli-ohm

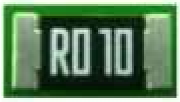

10 milli-ohm

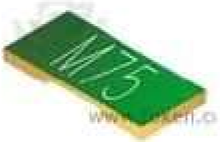

1-watt ????

THE COMPLETE RANGE OF SM RESISTOR MARKINGS

Click to see the complete range of SM resistor markings for 3-digit code:

Click to see the complete range of SM resistor markings for 4-digit code:

0000 is a value on a surface-mount resistor. It is a zero-ohm **LINK**!
Resistances less than 10 ohms have 'R' to indicate the position of the decimal point.
Here are some examples:

Three Digit Examples	Four Digit Examples
330 is 33 ohms *- not 330 ohms*	**1000 is 100 ohms *- not 1000 ohms***
221 is 220 ohms	**4992 is 49 900 ohms, or 49k9**
683 is 68 000 ohms, or 68k	**1623 is 162 000 ohms, or 162k**
105 is 1 000 000 ohms, or 1M	**0R56 or R56 is 0.56 ohms**
8R2 is 8.2 ohms	

A new coding system has appeared on **1% types**. This is known as the EIA-96 marking method. It consists of a three-character code. The first two digits signify the 3 significant digits of the resistor value, using the lookup table below. The third character - a letter - signifies the multiplier.

code	value	code	value	code	value	code	value	code	value	code	value
01	100	**17**	147	**33**	215	**49**	316	**65**	464	**81**	681
02	102	**18**	150	**34**	221	**50**	324	**66**	475	**82**	698
03	105	**19**	154	**35**	226	**51**	332	**67**	487	**83**	715
04	107	**20**	158	**36**	232	**52**	340	**68**	499	**84**	732
05	110	**21**	162	**37**	237	**53**	348	**69**	511	**85**	750
06	113	**22**	165	**38**	243	**54**	357	**70**	523	**86**	768
07	115	**23**	169	**39**	249	**55**	365	**71**	536	**87**	787
08	118	**24**	174	**40**	255	**56**	374	**72**	549	**88**	806
09	121	**25**	178	**41**	261	**57**	383	**73**	562	**89**	825
10	124	**26**	182	**42**	267	**58**	392	**74**	576	**90**	845
11	127	**27**	187	**43**	274	**59**	402	**75**	590	**91**	866
12	130	**28**	191	**44**	280	**60**	412	**76**	604	**92**	887
13	133	**29**	196	**45**	287	**61**	422	**77**	619	**93**	909
14	137	**30**	200	**46**	294	**62**	432	**78**	634	**94**	931
15	140	**31**	205	**47**	301	**63**	442	**79**	649	**95**	953
16	143	**32**	210	**48**	309	**64**	453	**80**	665	**96**	976

The **multiplier** letters are as follows:

letter	mult	letter	mult
F	**100000**	**B**	**10**
E	**10000**	**A**	**1**
D	**1000**	**X or S**	**0.1**
C	**100**	**Y or R**	**0.01**

22A is a 165 ohm resistor, **68C** is a 49900 ohm (49k9) and **43E** a 2740000 (2M74). This marking scheme applies to 1% resistors only.

A similar arrangement can be used for **2% and 5%** tolerance types. The multiplier letters are identical to 1% ones, but occur **before** the number code and the following **code** is used:

2%				**5%**			
code	**value**	**code**	**value**	**code**	**value**	**code**	**value**
01	100	**13**	330	**25**	100	**37**	330
02	110	**14**	360	**26**	110	**38**	360
03	120	**15**	390	**27**	120	**39**	390
04	130	**16**	430	**28**	130	**40**	430
05	150	**17**	470	**29**	150	**41**	470
06	160	**18**	510	**30**	160	**42**	510
07	180	**19**	560	**31**	180	**43**	560

08	200	**20**	620	**32**	200	**44**	620
09	220	**21**	680	**33**	220	**45**	680
10	240	**22**	750	**34**	240	**46**	750
11	270	**23**	820	**35**	270	**47**	820
12	300	**24**	910	**36**	300	**48**	910

With this arrangement, **C31** is 5%, 18000 ohm (18k), and **D18** is 510000 ohms (510k) 2% tolerance.
Always check with an ohm-meter (a multimeter) to make sure.

Chip resistors come in the following styles and ratings:
Style: 0402, 0603, 0805, 1206, 1210, 2010, 2512, 3616, 4022
Power Rating: 0402(1/16W), 0603(1/10W), 0805(1/8W), 1206(1/4W), 1210(1/3W), 2010(3/4W), 2512(1W), 3616(2W), 4022(3W)
Tolerance: 0.1%, 0.5%, 1%, 5%
Temperature Coefficient: 25ppm 50ppm 100ppm

EIA marking code for surface mount (SMD) resistors

01S = 1R	01R = 10R	01A = 100R	01B = 1k	01C = 10k	01D = 100k	01E = 1M	01F = 10M
02S = 1R02	02R = 10R2	02A = 102R	02B = 1k02	02C = 10k2	02D = 102k	02E = 1M02	
03S = 1R05	03R = 10R5	03A = 105R	03B = 1k05	03C = 10k5	03D = 105k	03E = 1M05	18F = 15M
04S = 1R07	04R = 10R7	04A = 107R	04B = 1k07	04C = 10k7	04D = 107k	04E = 1M07	
05S = 1R1	05R = 11R	05A = 110R	05B = 1k1	05C = 11k	05D = 110k	05E = 1M1	30F = 20M
06S = 1R13	06R = 11R3	06A = 113R	06B = 1k13	06C = 11k3	06D = 113k	06E = 1M13	
07S = 1R15	07R = 11R5	07A = 115R	07B = 1k15	07C = 11k5	07D = 115k	07E = 1M15	
08S = 1R18	08R = 11R8	08A = 118R	08B = 1k18	08C = 11k8	08D = 118k	08E = 1M18	
09S = 1R21	09R = 12R1	09A = 121R	09B = 1k21	09C = 12k1	09D = 121k	09E = 1M21	
10S = 1R24	10R = 12R4	10A = 124R	10B = 1k24	10C = 12k4	10D = 124k	10E = 1M24	
11S = 1R27	11R = 12R7	11A = 127R	11B = 1k27	11C = 12k7	11D = 127k	11E = 1M27	
12S = 1R3	12R = 13R	12A = 130R	12B = 1k3	12C = 13k	12D = 130k	12E = 1M3	
13S = 1R33	13R = 13R3	13A = 133R	13B = 1k33	13C = 13k3	13D = 133k	13E = 1M33	
14S = 1R37	14R = 13R7	14A = 137R	14B = 1k37	14C = 13k7	14D = 137k	14E = 1M37	
15S = 1R4	15R = 14R	15A = 140R	15B = 1k4	15C = 14k	15D = 140k	15E = 1M4	
16S = 1R43	16R = 14R3	16A = 143R	16B = 1k43	16C = 14k3	16D = 143k	16E = 1M43	
17S = 1R47	17R = 14R7	17A = 147R	17B = 1k47	17C = 14k7	17D = 147k	17E = 1M47	
18S = 1R5	18R = 15R	18A = 150R	18B = 1k5	18C = 15k	18D = 15k	18E = 1M5	
19S = 1R54	19R = 15R4	19A = 154R	19B = 1k54	19C = 15k4	19D = 154k	19E = 1M54	
20S = 1R58	20R = 15R8	20A = 158R	20B = 1k58	20C = 15k8	20D = 158k	20E = 1M58	
21S = 1R62	21R = 16R2	21A = 162R	21B = 1k62	21C = 16k2	21D = 162k	21E = 1M62	
22S = 1R65	22R = 16R5	22A = 165R	22B = 1k65	22C = 16k5	22D = 165k	22E = 1M65	
23S = 1R69	23R = 16R9	23A = 169R	23B = 1k69	23C = 16k9	23D = 169k	23E = 1M69	
24S = 1R74	24R = 17R4	24A = 174R	24B = 1k74	24C = 17k4	24D = 174k	24E = 1M74	
25S = 1R78	25R = 17R8	25A = 178R	25B = 1k78	25C = 17k8	25D = 178k	25E = 1M78	
26S = 1R82	26R = 18R2	26A = 182R	26B = 1k82	26C = 18k2	26D = 182k	26E = 1M82	
27S = 1R87	27R = 18R7	27A = 187R	27B = 1k87	27C = 18k7	27D = 187k	27E = 1M87	
28S = 1R91	28R = 19R1	28A = 191R	28B = 1k91	28C = 19k1	28D = 191k	28E = 1M91	
29S = 1R96	29R = 19R6	29A = 196R	29B = 1k96	29C = 19k6	29D = 196k	29E = 1M96	
30S = 2R0	30R = 20R0	30A = 200R	30B = 2k0	30C = 20k0	30D = 200k	30E = 2M0	
31S = 2R05	31R = 20R5	31A = 205R	31B = 2k05	31C = 20k5	31D = 205k	31E = 2M05	
32S = 2R10	32R = 21R0	32A = 210R	32B = 2k10	32C = 21k0	32D = 210k	32E = 2M10	
33S = 2R15	33R = 21R5	33A = 215R	33B = 2k15	33C = 21k5	33D = 215k	33E = 2M15	
34S = 2R21	34R = 22R1	34A = 221R	34B = 2k21	34C = 22k1	34D = 221k	34E = 2M21	
35S = 2R26	35R = 22R6	35A = 226R	35B = 2k26	35C = 22k6	35D = 226k	35E = 2M26	
36S = 2R32	36R = 23R2	36A = 232R	36B = 2k32	36C = 23k2	36D = 232k	36E = 2M32	
37S = 2R37	37R = 23R7	37A = 237R	37B = 2k37	37C = 23k7	37D = 237k	37E = 2M37	
38S = 2R43	38R = 24R3	38A = 243R	38B = 2k43	38C = 24k3	38D = 243k	38E = 2M43	
39S = 2R49	39R = 24R9	39A = 249R	39B = 2k49	39C = 24k9	39D = 249k	39E = 2M49	
40S = 2R55	40R = 25R5	40A = 255R	40B = 2k55	40C = 25k5	40D = 255k	40E = 2M55	
41S = 2R61	41R = 26R1	41A = 261R	41B = 2k61	41C = 26k1	41D = 261k	41E = 2M61	
42S = 2R67	42R = 26R7	42A = 267R	42B = 2k67	42C = 26k7	42D = 267k	42E = 2M67	
43S = 2R74	43R = 27R4	43A = 274R	43B = 2k74	43C = 27k4	43D = 274k	43E = 2M74	
44S = 2R80	44R = 28R0	44A = 280R	44B = 2k80	44C = 28k0	44D = 280k	44E = 2M80	
45S = 2R87	45R = 28R7	45A = 287R	45B = 2k87	45C = 28k7	45D = 287k	45E = 2M87	
46S = 2R94	46R = 29R4	46A = 294R	46B = 2k94	46C = 29k4	46D = 294k	46E = 2M94	
47S = 3R01	47R = 30R1	47A = 301R	47B = 3k01	47C = 30k1	47D = 301k	47E = 3M01	
48S = 3R09	48R = 30R9	48A = 309R	48B = 3k09	48C = 30k9	48D = 309k	48E = 3M09	
49S = 3R16	49R = 31R6	49A = 316R	49B = 3k16	49C = 31k6	49D = 316k	49E = 3M16	
50S = 3R24	50R = 32R4	50A = 324R	50B = 3k24	50C = 32k4	50D = 324k	50E = 3M24	
51S = 3R32	51R = 33R2	51A = 332R	51B = 3k32	51C = 33k2	51D = 332k	51E = 3M32	
52S = 3R4	52R = 34R0	52A = 340R	52B = 3k4	52C = 34k0	52D = 340k	52E = 3M4	
53S = 3R48	53R = 34R8	53A = 348R	53B = 3k48	53C = 34k8	53D = 348k	53E = 3M48	

54S = 3R57	54R = 35R7	54A = 357R	54B = 3k57	54C = 35k7	54D = 357k	54E = 3M57
55S = 3R65	55R = 36R5	55A = 365R	55B = 3k65	55C = 36k5	55D = 365k	55E = 3M65
56S = 3R74	56R = 37R4	56A = 374R	56B = 3k74	56C = 37k4	56D = 374k	56E = 3M74
57S = 3R83	57R = 38R3	57A = 383R	57B = 3k83	57C = 38k3	57D = 383k	57E = 3M83
58S = 3R92	58R = 39R2	58A = 392R	58B = 3k92	58C = 39k2	58D = 392k	58E = 3M92
59S = 4R02	59R = 40R2	59A = 402R	59B = 4k02	59C = 40k2	59D = 402k	59E = 4M02
60S = 4R12	60R = 41R2	60A = 412R	60B = 4k12	60C = 41k2	60D = 412k	60E = 4M12
61S = 4R22	61R = 42R2	61A = 422R	61B = 4k22	61C = 42k2	61D = 422k	61E = 4M22
62S = 4R32	62R = 43R2	62A = 432R	62B = 4k32	62C = 43k2	62D = 432k	62E = 4M32
63S = 4R42	63R = 44R2	63A = 442R	63B = 4k42	63C = 44k2	63D = 442k	63E = 4M42
64S = 4R53	64R = 45R3	64A = 453R	64B = 4k53	64C = 45k3	64D = 453k	64E = 4M53
65S = 4R64	65R = 46R4	65A = 464R	65B = 4k64	65C = 46k4	65D = 464k	65E = 4M64
66S = 4R75	66R = 47R5	66A = 475R	66B = 4k75	66C = 47k5	66D = 475k	66E = 4M75
67S = 4R87	67R = 48R7	67A = 487R	67B = 4k87	67C = 48k7	67D = 487k	67E = 4M87
68S = 4R99	68R = 49R9	68A = 499R	68B = 4k99	68C = 49k9	68D = 499k	68E = 4M99
69S = 5R11	69R = 51R1	69A = 511R	69B = 5k11	69C = 51k1	69D = 511k	69E = 5M11
70S = 5R23	70R = 52R3	70A = 523R	70B = 5k23	70C = 52k3	70D = 523k	70E = 5M23
71S = 5R36	71R = 53R6	71A = 536R	71B = 5k36	71C = 53k6	71D = 536k	71E = 5M36
72S = 5R49	72R = 54R9	72A = 549R	72B = 5k49	72C = 54k9	72D = 549k	72E = 5M49
73S = 5R62	73R = 56R2	73A = 562R	73B = 5k62	73C = 56k2	73D = 562k	73E = 5M62
74S = 5R76	74R = 57R6	74A = 576R	74B = 5k76	74C = 57k6	74D = 576k	74E = 5M76
75S = 5R9	75R = 59R0	75A = 590R	75B = 5k9	75C = 59k0	75D = 590k	75E = 5M9
76S = 6R04	76R = 60R4	76A = 604R	76B = 6k04	76C = 60k4	76D = 604k	76E = 6M04
77S = 6R19	77R = 61R9	77A = 619R	77B = 6k19	77C = 61k9	77D = 619k	77E = 6M19
78S = 6R34	78R = 63R4	78A = 634R	78B = 6k34	78C = 63k4	78D = 634k	78E = 6M34
79S = 6R49	79R = 64R9	79A = 649R	79B = 6k49	79C = 64k9	79D = 649k	79E = 6M49
80S = 6R65	80R = 66R5	80A = 665R	80B = 6k65	80C = 66k5	80D = 665k	80E = 6M65
81S = 6R81	81R = 68R1	81A = 681R	81B = 6k81	81C = 68k1	81D = 681k	81E = 6M81
82S = 6R98	82R = 69R8	82A = 698R	82B = 6k98	82C = 69k8	82D = 698k	82E = 6M98
83S = 7R15	83R = 71R5	83A = 715R	83B = 7k15	83C = 71k5	83D = 715k	83E = 7M15
84S = 7R32	84R = 73R2	84A = 732R	84B = 7k32	84C = 73k2	84D = 732k	84E = 7M32
85S = 7R5	85R = 75R0	85A = 750R	85B = 7k5	85C = 75k0	85D = 750k	85E = 7M5
86S = 7R68	86R = 76R8	86A = 768R	86B = 7k68	86C = 76k8	86D = 768k	86E = 7M68
87S = 7R87	87R = 78R7	87A = 787R	87B = 7k87	87C = 78k7	87D = 787k	87E = 7M87
88S = 8R06	88R = 80R6	88A = 806R	88B = 8k06	88C = 80k6	88D = 806k	88E = 8M06
89S = 8R25	89R = 82R5	89A = 825R	89B = 8k25	89C = 82k5	89D = 825k	89E = 8M25
90S = 8R45	90R = 84R5	90A = 845R	90B = 8k45	90C = 84k5	90D = 845k	90E = 8M45
91S = 8R66	91R = 86R6	91A = 866R	91B = 8k66	91C = 86k6	91D = 866k	91E = 8M66
92S = 8R87	92R = 88R7	92A = 887R	92B = 8k87	92C = 88k7	92D = 887k	92E = 8M87
93S = 9R09	93R = 90R9	93A = 909R	93B = 9k09	93C = 90k9	93D = 909k	93E = 9M09
94S = 9R31	94R = 93R1	94A = 931R	94B = 9k31	94C = 93k1	94D = 931k	94E = 9M31
95S = 9R53	95R = 95R3	95A = 953R	95B = 9k53	95C = 95k3	95D = 953k	95E = 9M53
96S = 9R76	96R = 97R6	96A = 976R	96B = 9k76	96C = 97k6	96D = 976k	96E = 9M76

If you want an accurate RESISTANCE measurement, remove the resistor from the circuit and use a Digital meter.

SURFACE MOUNT COMPONENTS - PACKS

Talking Electronics has packs of components for the repairman. The following packs are available:

SURFACE MOUNT RESISTOR PACK consists of 1 off each standard value
10 ohms to 1M & 2M2 (60 resistors)
$14.20   including pack and post

SURFACE MOUNT CAPACITOR PACK consists of:
2 - 10p   5 - 47p 5 - 100p 5 - 470p 5 - 1n 5 - 10n   5 - 22n   5 - 100n
5 - 1u 16v electrolytic   5 - 10u 16v electrolytic
(40 components)
$23.80 including pack and post

SURFACE MOUNT DIODE PACK consists of:
5 - 1N 4148 (marked as "A6")
$10.00   including pack and post

SURFACE MOUNT TRANSISTOR PACK consists of:
5 - BC 848 (marked as "1K") NPN
5 - BC858 PNP
$10.00   including pack and post

email **Colin Mitchell** for details on how to pay by credit card or PayPal.

CREATING ANY VALUE OF RESISTANCE

Any value of resistance can be created by connecting two resistors in PARALLEL or SERIES.
You can also create a higher wattage resistor by connecting them in SERIES OR PARALLEL.
We are only going to cover two EQUAL VALUE resistors in SERIES or in PARALLEL.
If you want to create a "Special Value," simply connect two resistors and read the value with a Digital Meter. Keep changing the values until you get the required value.
We are not going into series or Parallel formulae. You can easily find a value with a multimeter.

TWO EQUAL-VALUE RESISTORS IN SERIES

Two equal-value resistors IN SERIES creates a value of DOUBLE. You simply ADD the values.
This can be done with any to two values as shown. Three equal-value resistors in series is three times the value.

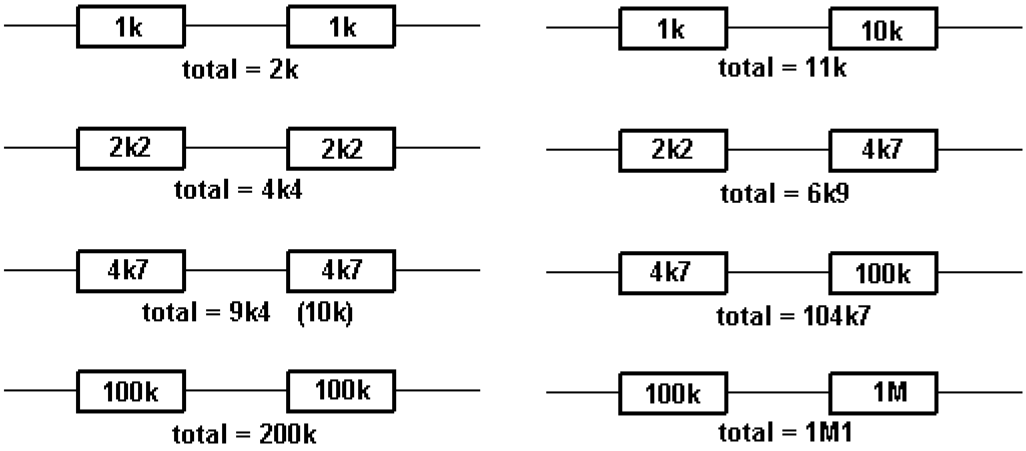
1k
1k
total = 2k
1k
10k
total = 11k
2k2
2k2
total = 4k4
2k2
4k7
total = 6k9
4k7
4k7
total = 9k4 (10k)
4k7
100k
total = 104k7
100k
100k
total = 200k
100k
1M
total = 1M1

TWO EQUAL-VALUE RESISTORS IN PARALLEL

Two equal-value resistors IN PARALLEL creates a value of HALF. Three equal-value resistors in parallel is equal to one-third the value.

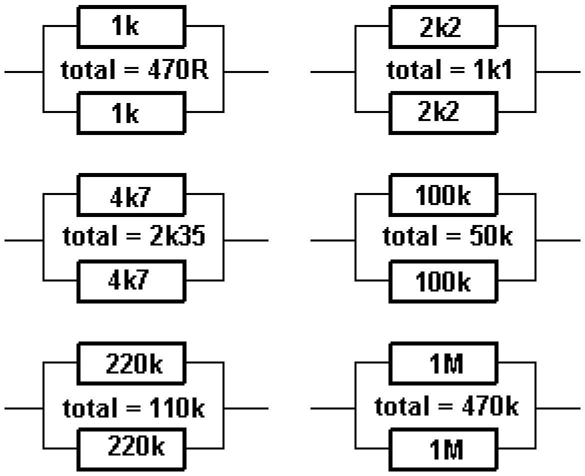
1k
total = 470R
1k
2k2
total = 1k1
2k2
4k7
total = 2k35
4k7
100k
total = 50k
100k
220k
total = 110k
220k
1M
total = 470k
1M

If you want a particular value and it is not available, here is a chart.
Use 2 resistors in series or parallel as shown:

Required Value	R1	Series/ Parallel	R2	Actual value:

10	4R7	S	4R7	9R4
12	10	S	2R2	12R2
15	22	P	47	14R9
18	22	P	100	18R
22	10	S	12	22
27	22	S	4R7	26R7
33	22	S	10	32R
39	220	P	47	38R7
47	22	S	27	49
56	47	S	10	57
68	33	S	33	66
82	27	S	56	83

There are other ways to combine 2 resistors in parallel or series to get a particular value. The examples above are just one way. 4R7 = 4.7 ohms

TESTING A RESISTOR

To check the value of a resistor, it should be removed from the circuit. The surrounding components can affect the reading and make it lower.

Resistors **VERY RARELY** change value, but if it is overheated or damaged, the resistance can increase. You can take the reading of a resistor "in-circuit" in one direction then the other, as the surrounding components may have diodes and this will alter the reading.

You can also test a resistor by feeling its temperature-rise. It is getting too hot if you cannot hold your finger on it (some "metal film" resistors are designed to tolerate quite high temperatures).

TESTING AN "AC" RESISTOR

There is no such thing as an "AC" resistor. Resistors are just "resistors" and they can be in AC circuits or DC circuits. Resistors can be given names such as "Safety Resistor" "Ballast Resistor" "LOAD Resistor" "Feed Resistor" "Dropper Resistor" or "Supply Resistor." These are just normal resistors with a normal resistance - except a "Safety Resistor."

A safety resistor is made of a flame-proof material such as metal-oxide-film and not carbon-composition. It is designed to "burn out" when too much current flows BUT NOT CATCH FIRE.

It is a low-value resistor and has a voltage-drop across it but this is not intentional. The voltage-drop is to create a "heating-effect" to burn out the resistor. In all the other types of resistor, the voltage-drop is intentional.

A Ballast resistor is a normal resistor and can be called a Power resistor, Dropper resistor, Supply resistor or Feed resistor. It is designed to reduce the voltage from one source and deliver a lower voltage. It is a form of: "in-line" resistor.

A Load Resistor is generally connected across the output of a circuit and turns the energy it receives, into heat.

RESISTOR NETWORKS

To reduce the number of resistors in a circuit, some engineers use a set of identical resistors in a package called a Single-In-Line (SIL) resistor network. It is made with many resistors of the same value, all in one package. One end of each resistor is connected all the other resistors and this is the common pin, identified as pin 1 and has a dot on the package.

These packages are very reliable but to make sure all the resistors are as stated, you need to locate pin 1. All values will be identical when referenced to this pin.

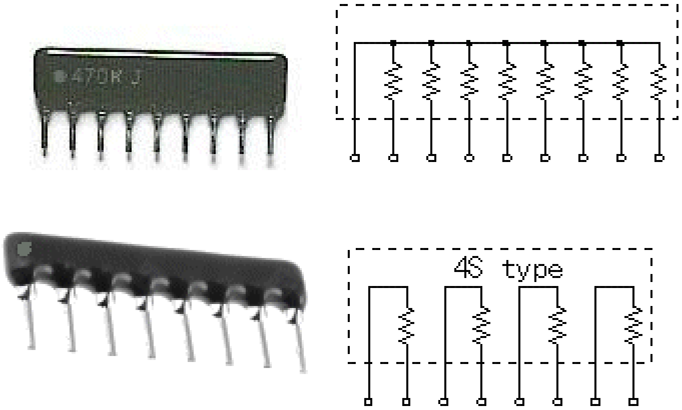

RESISTOR NETWORKS

Some resistor networks have a "4S" printed on the component. The 4S indicates the package contains 4 independent resistors that are not wired together inside. The housing has eight leads as shown in the second image.
Independent resistors have an even number of pins and measuring between each pair will produce identical values. Resistance between any pair will indicate leakage and may be a fault.

WIRE WOUND RESISTOR

A wire wound resistor is also called a POWER RESISTOR. This type of resistor can have a resistance as low as 0.1 ohms (one-tenth of an ohm) or as high as about 10k.
The image shows a 0.68 ohm resistor as the letter "R" represents the DECIMAL POINT and R68 is the same a .68 and this is 0.68 ohms. The wattage is 9 watts.
This resistor will allow xxx amps to flow. To work out the current, use the formula:
Power = Current x Current x resistance
9 = Current x Current x .68
Divide both sides by 0.68
13.2 = Current x Current
Find the square root of 13.2
Current = 3.6 amps

When 3.6 amps flow through the resistor, the voltage appearing across it will be:
V = current x resistance
= 3.6 x 0.68
= 2.5v and the wattage (heat) loss will be 9 watts.

The purpose of a resistor like this is to stop or reduce "ripple." Ripple is the noise or hum in an amplifier when the sound is turned up.
There are many reasons why you need to reduce the level of hum and this resistor will remove ripple as large as 2.5v when 3.6 amps is flowing, provided you have filter electrolytics on both side of the resistor to assist in removing the ripple.

If the letter "R" is in a different position, the value of resistance would be:

68R = 68Ω
6R8 = 6.8Ω
R68 = 0.68Ω

If you replace the R68 resistor a 6R8 resistor by mistake, the voltage across it will rise to 25v and if 3.6 amps flows, the wattage will be: 90 watts!!!
The resistor will glow red and burn out.

TESTING A POSISTOR

A Posistor is a resistor that connects in series with the degaussing coil around the picture tube or **Monitor.** When cold, it has a very low resistance and a large current flows when the monitor or TV is switched on. This current heats up the Posistor and the resistance increases. This causes the current to decrease and any magnetism in the shadow mask is removed. The posistor can one or two elements and it is kept warm so the resistance remains high. Many Posistors have a second element inside the case that connects directly to the supply to keep the Positive Temperature Coefficient resistor high so that the current through the degaussing coil falls to almost zero. This constant heat eventually destroys the package.
The heavy current that flows when a set is turned ON also causes the posistor to crack and break and this results in poor purity on the screen - as the shadow mask gradually becomes magnetic..
Posistors have different resistance values from different manufacturers and must be replaced with an identical type.
They can be checked for very low resistance when cold but any loose pieces inside the case will indicate a damaged component.

A "BURNT" RESISTOR **- normally and technically called a "burnt-out" resistor.**

The resistance of a "burnt" resistor can sometimes be determined by scraping away the outer coating - if the resistor has a spiral of resistance-material. You may be able to find a spot where the spiral has been damaged.

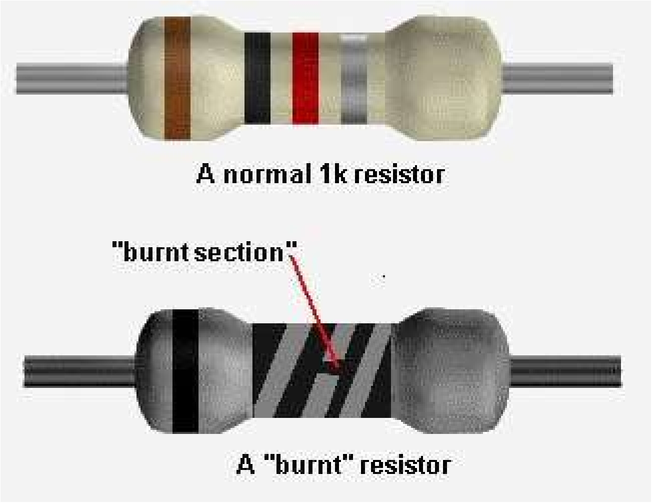

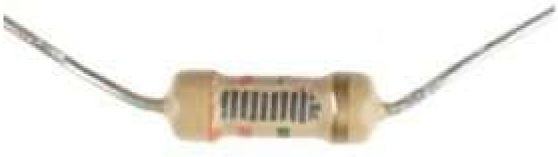

**Note the spirals of conductive carbon.
The number of spirals has nothing to with the resistance.
It is the amount of carbon particles in the "track" that
determines the resistance. It is also the thickness and width
of the track that determines the resistance.
And then it is the overall size of the resistor that determines the wattage.
And then the size of the leads, the closeness to the PCB and
the size of the lands that eventually determines how hot the resistor
will get.**

Clean the "spot" (burnt section of the spiral) very carefully and make sure you can get a good contact with the spiral and the tip of your probe. Measure from one lead of the resistor to the end of the damaged spiral. Then measure from the other lead to the other end of the spiral.
Add the two values and you have an approximate value for the resistor. You can add a small amount for the damaged section.
This process works very well for damaged wire-wound resistors. They can be pulled apart and each section of the resistance-wire (nichrome wire) measured and added to get the full resistance.

There is another way to determine the value of a damaged resistor.
Get a set of resistors of the same wattage as the damaged component and start with a high value. It's handy to know if the resistor is in the range: 10ohm to 100ohms or 1k to 10k etc, but this is not essential.
Start with a very high value and turn the circuit ON. You can perform voltage tests and if you know the expected output voltage, decrease the resistance until this voltage is obtained.
If you do not know the expected voltage, keep reducing the value of resistance until the circuit works as designed.
This is the best advice in a situation where you do not know the value of a resistor.

There is a third way to determine the value and this requires measuring the voltage drop across the resistor and the current-flow. By multiplying the two you will get a wattage and this must be less than the wattage of the resistor being replaced.

TESTING POTENTIOMETERS (variable resistors)

To check the value of a variable resistor, it should be removed from circuit or at least 2 legs should be removed. A Rheostat is a variable resistor using only one end and the middle connected to a circuit.
The resistance between the two outside pins is the value marked on the component and the centre leg will change from nearly zero to the full resistance as the shaft is rotated.
"Pots" generally suffer from "crackle" when turned and this can be fixed by spraying up the shaft and into the pot via the shaft with a tube fixed to a can of "spray-lubricant" (contact cleaner).
"Pre-set pots" and "trim pots" are miniature versions of a potentiometer and they are all tested the same. The photo shows a pot, two mini pots and 3 mini trim pots.

102
103
P102
N911

10-Turn POTs

A 10-turn pot is one of the worst items to be designed. I remove them immediately from any design.
You don't know the position of the wiper. You don't know which way you are turning the wiper and you can't remember which way you turned the post "the last time."
The screwdriver always falls out of the slot.
If you need fine adjustment, place fixed resistors on each side of the pot and use a normal mini trim pot with much less resistance.

FOCUS POTS

Focus pots quite often get a spot of dirt where the wiper touches the track. Cleaning with spray fixes the bad focus but if the pot is leaking to chassis from inside the pot (due to the high voltage on the terminals) simply remove it from the chassis and leave it floating (this will restore the high voltage to the picture tube) or you can use one from an old chassis.

MAKING YOUR OWN RESISTOR, CAPACITOR, INDUCTOR or DIODE

Quite often you will not have the exact value of resistance or capacitance for a repair.
We have already covered placing resistors and capacitors in parallel and series:

Resistors in Parallel and/or Series
Capacitors in Parallel and/or Series

Here are some extras:

RESISTORS
Two 1k 0.5watt resistors in parallel produces a 470R 1watt resistor.
Two 1k 0.5watt resistors in series produces a 2k 1watt resistor.

CAPACITORS
Two 100n 100v capacitors in series produces a 50n capacitor @200v

INDUCTORS: Two inductors in series - **ADD THE VALUES**

DIODES: Two 1Amp 400v diodes in series produces a 1Amp 800v diode
Two 1Amp 400v diodes in parallel produces a 2Amp 400v diode

ZENER DIODES: Zener diodes can be connected in series to get a higher voltage.
Two 12v zener diodes in series produces a 24v zener.

CONTINUITY

Some multimeters have a "buzzer" that detects when the probes are touching each other or the resistance between the probes is very LOW. This is called a CONTINUITY TESTER.
You can use the resistance scale "x1" or "x10" to detect low values of resistance.
Set the pointer to "0" (right end of the scale) by touching the probes together and adjusting the "zero ohms" control.
When taking a reading, you will have to decide if a low value of resistance is a short-circuit or an "operating value."
For instance, the cold resistance of a 12v car globe is very low (about 2 ohms) and it increases (about 6 times) to 12 ohms when hot.
The "resistance of a circuit" may be very low as the electrolytics in the circuit are uncharged. This may not indicate a true "short-circuit."
The measurement across a diode is not a resistance-value but a "voltage-drop" and that is why the needle swings nearly full-scale.
Leads and wires and cords have a small resistance and depending on the length of the lead, this small resistance may be affecting a circuit.
Remember this:
When a circuit takes 1 amp, and the resistance of the leads is 1 ohm, the voltage drop across the leads will be 1v.
That's why a 12v battery supplying a circuit with these leads will have 11v at the circuit.
Note:
Turn off the equipment before making any continuity tests. The presence of even a small voltage (from an electrolytic) can give a false reading.
You can determine the resistance of a lead very accurately by taking the example above and applying it to your circuit.
If the battery is 12.6v and the voltage across the circuit is 10v, when the current is 2.6 amps, the resistance of the "leads" is 12.6 - 10 = 2.6 R=V/I = 2.6/2.6 = 1ohm. By making the lead shorter or using thicker wire, the resistance will be less and the voltage on the project will increase.
When taking readings in a circuit that has a number of diodes built-into IC's (Integrated Circuits) and transistors, some Continuity Testers will beep and give a false reading.
The following circuit has the advantage of providing a beep when a short-circuit is detected but does not detect the small voltage drop across a diode. This is ideal when testing logic circuits as it is quick and you can listen for the beep while

concentrating on the probe. Using a multimeter is much slower.

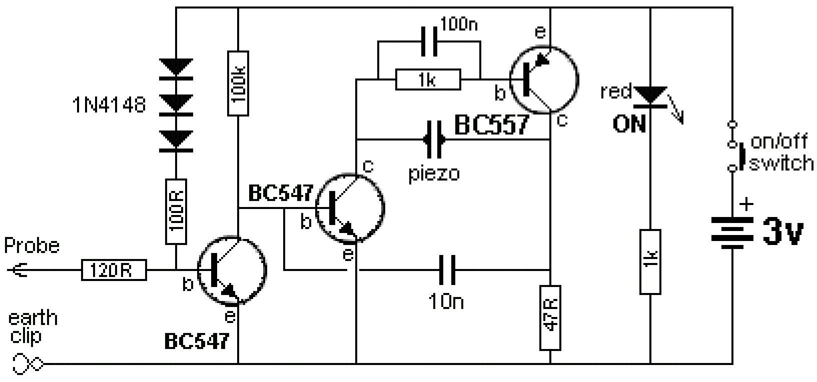

CONTINUITY TESTER

You can build the circuit on Matrix Board and add it to your Test Equipment.
You will need lots of "Test Equipment" and they can be built from circuits in this eBook.

TESTING FUSES, LEADS AND WIRES

All these components come under the heading TESTING for CONTINUITY. Turn off all power to the equipment before testing for shorts and continuity. Use the low resistance "Ohms Scale" or CONTINUITY range on your multimeter. All fuses, leads and wires should have a low, very low or zero resistance. This proves they are working.

A BLOWN FUSE

The appearance of a fuse after it has "blown" can tell you a lot about the fault in the circuit.
If the inside of the glass tube (of the fuse) is totally blackened, the fuse has been damaged very quickly. This indicates a very high current has passed through the fuse.
Depending on the rating of the fuse, (current rating) you will be able to look for components that can pass a high current when damaged - such as high power transistors, FETs, coils, electrolytics. Before re-connecting the supply, you should test the "SUPPLY RAILS" for resistance. This is done by measuring them on a low OHMs range in one direction then reverse the leads to see if the resistance is low in the other direction.
A reading can be very low at the start because electrolytics need time to charge-up and if the reading gradually increases, the power rail does not have a short. An overload can occur when the supply voltage rises to nearly full voltage, so you sometimes have to fit a fuse and see how long it takes to "blow."
If the fuse is just slightly damaged, you will need to read the next part of this eBook, to see how and why this happens:

FAST AND SLOW BLOW FUSES

There are many different sizes, shapes and ratings of a fuse. They are all current ratings as a fuse does not have a voltage rating. Some fuses are designed for cars as they fit into the special fuse holders. A fuse can be designed for 50mA, 100mA, 250mA, 315mA, 500mA, 1Amp, 1.5amp, 2amp, 3amp, 3.15amp 5amp, 10amp, 15amp, 20amp, 25amp, 30amp, 35amp, 50amp and higher.
Some fuses are fast-blow and some are slow-blow.
A "normal" fuse consists of a length of thin wire. Or it may be a loop of wire that is thin near the middle of the fuse. This is the section that will "burn-out."
A "normal" fuse is a fast-blow fuse. For instance, a 1amp fuse will remain intact when up to 1.25 amp flows. When a circuit is turned on, it may take 2-3 amps for a very short period of time and a normal 1 amp fuse will get very hot and the wire will stretch but not "burn-out." You can see the wire move when the supply turns on.
If the current increases to 2amps, the fuse will still remain intact. It needs about 3 amp to heat up the wire to red-hot and burn out.

If the current increases to 5 amp, the wire VOLATILISES (burns-out) and deposits carbon-black on the inside of the glass tube.

A slow-blow fuse uses a slightly thicker piece of wire and the fuse is made of two pieces of wire joined in the middle with a dob of low-temperature solder. Sometimes one of the pieces of wire is a spring and when the current rises to 2.5 amp, the heat generated in the wire melts the solder and the two pieces of wire "spring apart."

A slow-blow fuse will allow a higher current-surge to pass through the fuse and the wire will not heat up and sag.

Thus the fuse is not gradually being damaged and it will remain in a perfect state for a long period of time.

A fuse does not protect electronic equipment from failing. It acts AFTER the equipment has failed.

It will then protect a power supply from delivering a high current to a circuit that has failed.

If a slow-blow fuse has melted the solder, it could be due to a slight overload, slight weakening of the fuse over a period of time or the current-rating may be too low. You can try another fuse to see what happens.

You can replace a fast-acting fuse (normal fuse) with a slow blow if the fast-acting fuse has been replaced a few times due to deterioration when the equipment is turned on.

But you cannot replace a slow-blow fuse with a fast acting fuse as it will be damaged slightly each time the equipment is turned on and eventually fail.

100mA FUSES

Fuses below about 100mA are very hard to make and very unreliable.

Many circuits take a high current when turned to charge the electrolytics and a 100mA (or 50mA or 63mA fuse) will bow and stretch and change shape, every time the equipment is turned ON. Eventually it will break, due to it heating-up and stretching.

To produce a reliable fuse below 100mA, some manufacturers have placed a resistor inside the fuse and connected it to a spring. One end of the resistor is soldered to a wire with low-temperature metal and when the resistor gets hot, the metal softens and the spring pulls the resistor away from the wire.

Quite often you can heat up the metal and connect the wire and the fuse is perfect.

This type of fuse is called a DELAY fuse and the current rating is shown on the end-cap.

The value of the resistor determines the current rating.

There is a small voltage across this type of fuse and it means the circuit sees a slightly lower voltage than the supply voltage.

The third photo shows the pot of solder or low-temp metal and a wire connected to a spring. The heat generated in the wire is passed to the solder and it softens. The spring pulls the two components apart. You can smash the glass and set up the fuse in the two fuse-holders and repair the fuse while you wait for a new fuse to arrive.

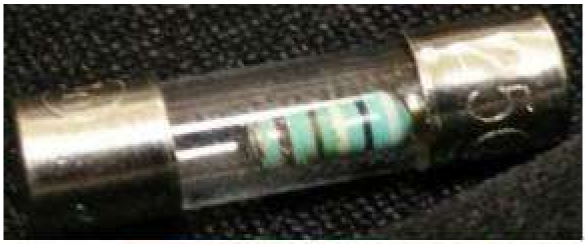

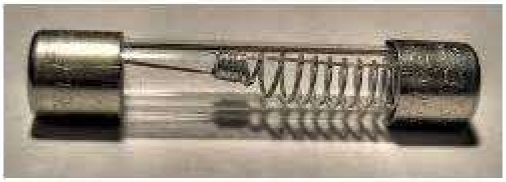

TESTING COILS, INDUCTORS and YOKES

Coils, inductors, chokes and yokes are just coils (turns) of wire. The wire may be wrapped around a core made of iron or ferrite.
It is labeled "L" on a circuit board.
You can test this component for continuity between the ends of the winding and also make sure there is no continuity between the winding and the core.
The winding can be less than one ohm, or greater than 100 ohms. A coil of wire is also called an INDUCTOR and it might look like a very simple component, but it can operate in a very complex way.
The way it works is a discussion for another eBook. It is important to understand the turns are insulated but a slight fracture in the insulation can cause two turns to touch each other and this is called a "SHORTED TURN" or you can say the inductor has "SHORTED TURNS."
When this happens, the inductor allows the circuit to draw MORE CURRENT. This causes the fuse to "blow."
The quickest way to check an inductor is to replace it, but if you want to measure the inductance, you can use an INDUCTANCE METER. You can then compare the inductance with a known good component.
An inductor with a shorted turn will have a very low or zero inductance, however you may not be able to detect the fault when it is not working in a circuit as the fault may be created by a high voltage generated between two of the turns.
Faulty yokes (both horizontal and vertical windings) can cause the picture to reduce in size and/or bend or produce a single horizontal line.
A TV or monitor screen is the best piece of Test Equipment as it has identified the fault. It is pointless trying to test the windings further as you will not be able to test them under full operating conditions. The fault may not show up when a low voltage (test voltage) is applied.

MEASURING AND TESTING INDUCTORS

Inductors are measured with an INDUCTANCE METER but the value of some inductors is very small and some Inductance Meters do not give an accurate reading. The solution is to measure a larger inductor and note the reading. Now put the two inductors in SERIES and the values ADD UP - just like resistors in SERIES. This way you can measure very small inductors. VERY CLEVER!

Question from a reader: *Can I add an inductor to stop a fuse blowing?*

Basically, an inductor NEVER prevents a fuse blowing because an inductor prevents spikes on one lead (we will call the INPUT lead), appearing on its other lead.
This is the detection and prevention of current that exists for a very short period of time.
A fuse detects an excess of current that occurs over a very long period of time and they are entirely two different "detectors."
One cannot assist the other in any way.

An inductor is basically a coil of wire. It may be thick or thin wire. The value of the inductor is a combination of the number of turns and the material on which the wire is wound.
The value of an inductor does not change over say a period of 20 years but it can go faulty by the enamel cracking and two turns touching. This can also be due to the difference in voltage between the two turns creating a spark between the turns and creating a "short."
When you test it, the high voltage is not present and it will test ok.
You may not think a few turns of wire will have any effect on improving a circuit, but spikes are very high frequency and the inductor will have a very big effect on reducing them.
An inductor (say 100uH) can be produced in many different sizes and the thickness of the wire will be important as it determines the current that can flow through the inductor.
The term "inductor" also includes those with two or more windings and these components are called TRANSFORMERS. These devices can get "shorts" and "leaks" between the windings and sparks can be seen between the windings. These sparks do not occur when you are testing them on test-equipment so the only way to

guarantee success is to replace it with an identical replacement.

TESTING SWITCHES and RELAYS

Switches and relays have contacts that open and close mechanically and you can test them for CONTINUITY. However these components can become intermittent due to dirt or pitting of the surface of the contacts due to arcing as the switch is opened.
It is best to test these items when the operating voltage and current is present as they quite often fail due to the arcing. A switch can work 49 times then fail on each 50th operation. The same with a relay. It can fail one time in 50 due to CONTACT WEAR.
If the contacts do not touch each other with a large amount of force and with a large amount of the metal touching, the current flowing through the contacts will create HEAT and this will damage the metal and sometimes reduce the pressure holding the contact together.
This causes more arcing and eventually the switch heats up and starts to burn. Switches are the biggest causes of fire in electrical equipment and households.

A relay also has a set of contacts that can cause problems.
There are many different types of relays and basically they can be put into two groups.
1. An electromagnetic relay is a switch operated by magnetic force. This force is generated by current through a coil. The relay opens and closes a set of contacts. The contacts allow a current to flow and this current can damage the contacts.
Connect 5v or 12v to the coil (or 24v) and listen for the "click" of the points closing. Measure the resistance across the points to see if they are closing.
You really need to put a load on the points to see if they are clean and can carry a current.
The coil will work in either direction.
If not, the relay is possibly a CMOS relay or Solid State relay.

2. An electronic relay (Solid State Relay) does not have a winding. It works on the principle of an opto-coupler and uses a LED and Light Activated SCR or Opto-TRIAC to produce a low resistance on the output. The two pins that energise the relay (the two input pins) must be connected to 5v (or 12v) around the correct way as the voltage is driving a LED (with series resistor). The LED illuminates and activates a light-sensitive device.

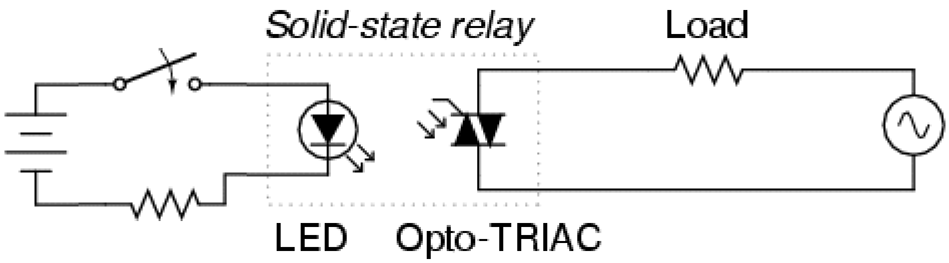

TESTING REED SWITCHES

A reed switch is generally contained in a long glass tube:

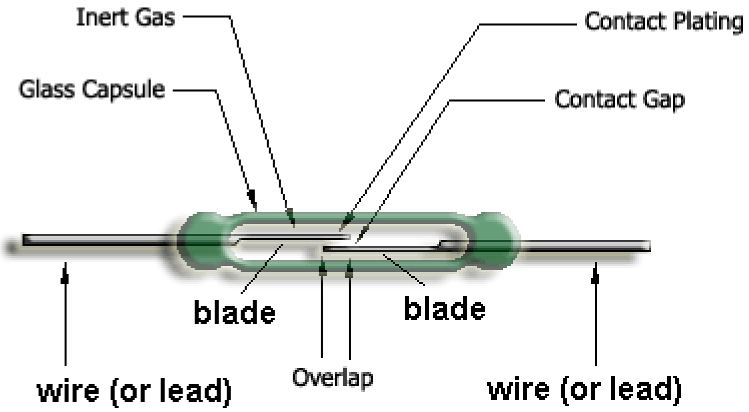

A wire or lead comes out each end for soldering to the reed switch to the project. The two "blades" inside the switch are made from a material that can be magnetised but does not retain its magnetism. This effect is called "temporally magnetised" (not permanently magnetised) and really only "passes" magnetism from one end to the other when in the presence of a magnet. One of the blades is made of a soft material and it will bend very easily. The other one is much stiffer.

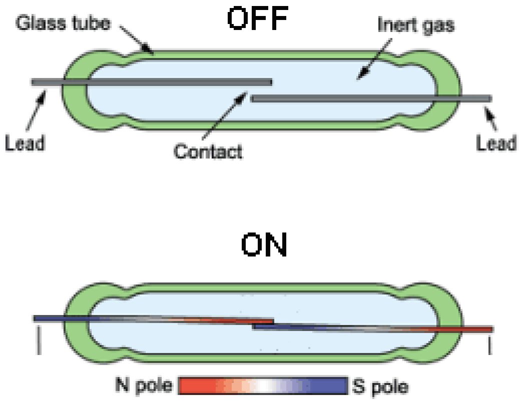

When a magnet is placed under the two blades (or on top), the magnetism from the magnet is passed to the two blades (INDUCTION or MUTUAL INDUCTION - commonly called INDUCED) and it produces a very weak magnet (in the blade) that is identical to the powerful magnet as far as the position of the north and south poles are concerned). Initially it produces a N-S and N-S set of poles and this makes the two blades click together because the top blade will be South at the contact and the bottom blade will be North.

When the two blades click together the magnetism runs through the two blades and keeps them together. The two blades attract and the switch is closed. When the magnet is removed, the magnetism in the two blades ceases and the two blades move apart.

Since there is a very small amount of movement in the top blade, this switch has a limited number of operations. Eventually it will fail. It is a mechanical device and is not suited for detecting a spinning shaft as 100,000 revolutions will very quickly weaken the switch.

If the switch does not make contact or remains closed, the moveable blade can be cracked or broken. This can be very hard to see. So replace the switch.

LATCHING REED SWITCH

A "normal" reed switch can be converted into a **LATCHING REED SWITCH** by carefully placing a magnet below the switch and moving it away so the two blades open. Now move it slightly closer but do not allow the blades to close.

This is called putting a "SET" on the switch and the two blades will have a small magnetic effect "induced" in them but not enough to close the contacts:

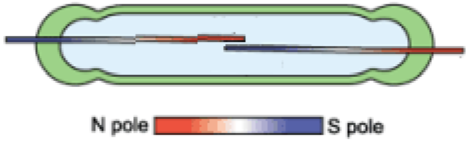

Now bring a strong magnet up to the reed switch on the other side of the glass tube with the north pole above the north of the lower magnet. This effect will increase the INDUCED MAGNETISM in the blades and close the contacts:

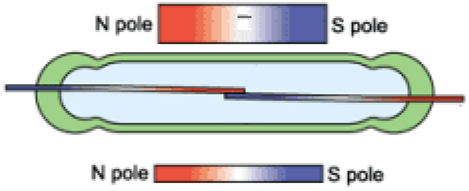

Remove the top magnet and the lower magnet will induce enough magnetism into the blades to keep them closed:

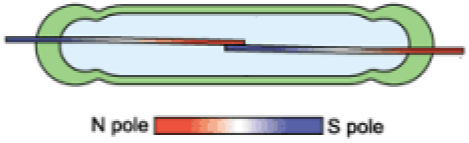

Now bring the upper magnet near the reed switch with the south pole above the north pole of the lower magnet. (In other words: AROUND THE OTHER WAY) This will have the effect of reducing the induced magnetism in the blades and a point will be reached when the two contacts will separate:

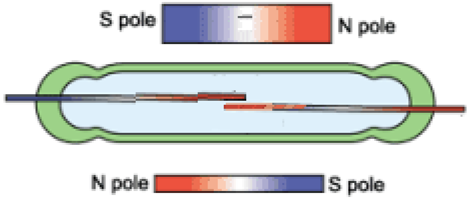

Remove the top magnet and the switch will remain separated because the lower magnet will not have sufficient influence on the blades to close the contact:

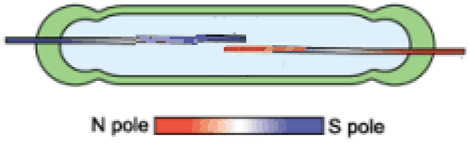

CAPACITORS

Capacitors are one of the most difficult things to test. That's because they don't give a reading on a multimeter and their value can range from 1p to 100,000u.
A faulty capacitor may be "open" when measured with a multimeter, and a good capacitor will also be "open."

You need a piece of test equipment called a CAPACITANCE METER to measure the value of a capacitor.

HOW A CAPACITOR WORKS

There are two ways to describe how a capacitor works. Both are correct and you have to combine them to get a full picture.
A capacitor has INFINITE resistance between one lead and the other.
This means no current flows **through** a capacitor. But it works in another way.
Suppose you have a strong magnet on one side of a door and a piece of metal on the other. By sliding the magnet up and down the door, the metal rises and falls.
The metal can be connected to a pump and you can pump water by sliding the magnet up and down.
A capacitor works in exactly the same way.
If you raise a voltage on one lead of a capacitor, the other lead will rise to the same voltage. This needs more explaining - we are keeping the discussion simple.
It works just like the magnetic field of the magnet through a door.
The next concept is this:
Capacitors are equivalent to a tiny rechargeable battery.
They store energy when the supply-voltage is present and release it when the supply drops.
These two concepts can be used in many ways and that's why capacitors perform tasks such as filtering, time-delays, passing a signal from one stage to another and create many different effects in a circuit.

CAPACITOR VALUES

The basic unit of capacitance is the FARAD. (C) This is the value used in all equations, but it is a very large value. A one FARAD capacitor would be the size of a car if made with plates and paper. Most electronic circuits use capacitors with smaller values such as 1p to 1,000u. 1p is about equal to two parallel wires 2cm long. 1p is one picofarad.

The easiest way to understand capacitor values is to start with a value of 1u. This is one microfarad and is one-millionth of a Farad. A 1 microfarad capacitor is about 1cm long and the diagram shows a 1u electrolytic.

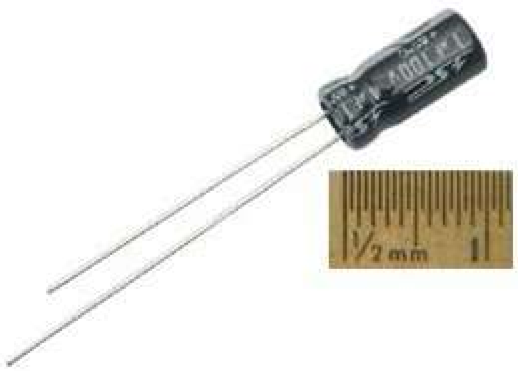

Smaller capacitors are ceramic and they look like the following. This is a 100n (0.1u)ceramic:

To read the value on a capacitor you need to know a few facts.

The basic value of capacitance is the FARAD.
1 microfarad is one millionth of 1 farad.
1 microfarad is divided into smaller parts called nanofarad.
1,000 nanofarad = 1 microfarad
Nanofarad is divided into small parts called picofarad
1,000 picofarad = 1 nanofarad.

Recapping:
1p = 1 picofarad.   1,000p = 1n (1 nanofarad)
1,000n = 1u (1 microfarad)
1,000u = 1millifarad
1,000,000u = 1 FARAD.

Examples:
All ceramic capacitors are marked in "p" (puff")
A ceramic with 22 is 22p = 22 picofarad
A ceramic with 47 is 47p = 47 picofarad
A ceramic with 470 is 470p = 470 picofarad
A ceramic with 471 is 470p = 470 picofarad
A ceramic with 102 is 1,000p = 1n
A ceramic with 223 is 22,000p = 22n
A ceramic with 104 is 100,000p = 100n = 0.1u

TYPES OF CAPACITOR
For testing purposes, there are two types of capacitor.
Capacitors from 1p to 100n are non-polar and can be inserted into a circuit around either way.
Capacitors from 1u to 100,000u are electrolytics and are polarised. They must be fitted so the positive lead goes to the supply voltage and the negative lead goes to ground (or earth).
There are many different sizes, shapes and types of capacitor. They are all the same. They consist of two plates with an insulating material between. The two plates can be stacked in layers or rolled together.
The important factor is the insulating material. It must be very thin to keep things small. This gives the capacitor its VOLTAGE RATING.
If a capacitor sees a voltage higher than its rating, the voltage will "jump through" the insulating material or around it.
If this happens, a carbon deposit is left behind and the capacitor becomes "leaky" or very low resistance, as carbon is conductive.

CERAMIC CAPACITORS

Nearly all small capacitors are **ceramic capacitors** as this material is cheap and the capacitor can be made in very thin layers to produced a high capacitance for the size of the component. This is especially true for surface-mount capacitors.
All capacitors are marked with a value and the basic unit is: "p" for "puff" However NO surface mount capacitors are marked and they are very difficult to test.

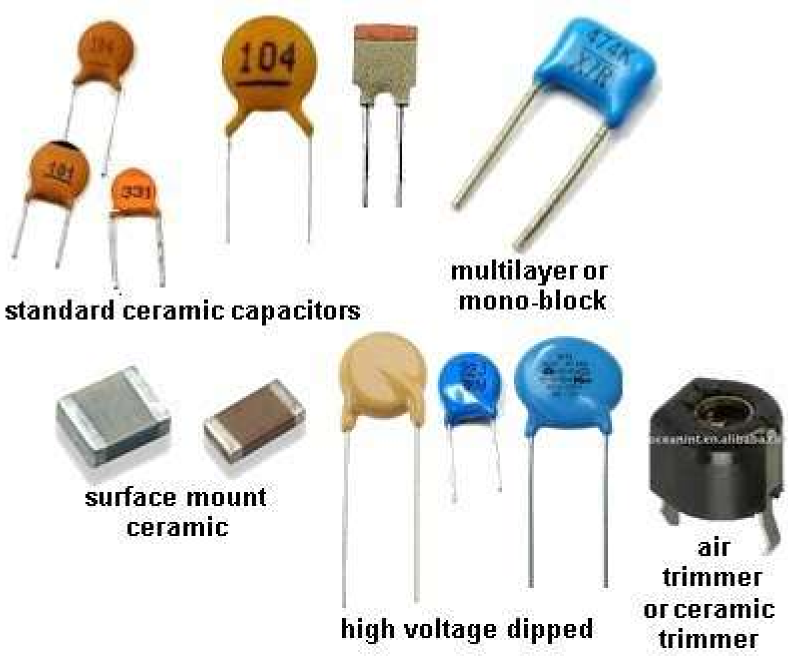
104
104
331
474K X7R
standard ceramic capacitors
multilayer or mono-block
surface mount ceramic
high voltage dipped
air trimmer or ceramic trimmer

VALUE:	VALUE WRITTEN ON THE COMPONENT:
0.1p	0p1
0.22p	0p22
0.47p	0p47
1.0p	1p0
2.2p	2p2
4.7p	4p7
5.6p	5p6
8.2p	8p2
10p	10 or 10p
22p	22 or 22p
47p	47 or 47p
56p	56 or 56p
100p	100 on 101
220p	220 or 221
470p	470 or 471
560p	560 or 561
820p	820 or 821
1,000p (1n)	102
2200p (2n2)	222
4700p (4n7)	472
8200p (8n2)	822
10n	103
22n	223
47n	473
100n	104
220n	224
470n	474
1u	105

POLYESTER, POLYCARBONATE, POLYSTYRENE, MYLAR, METALLISED POLYESTER, ("POLY"), MICA and other types of CAPACITOR

There are many types of capacitor and they are chosen for their reliability, stability, temperate-range and cost.
For testing and repair work, they are all the same. Simply replace with exactly the same type and value.

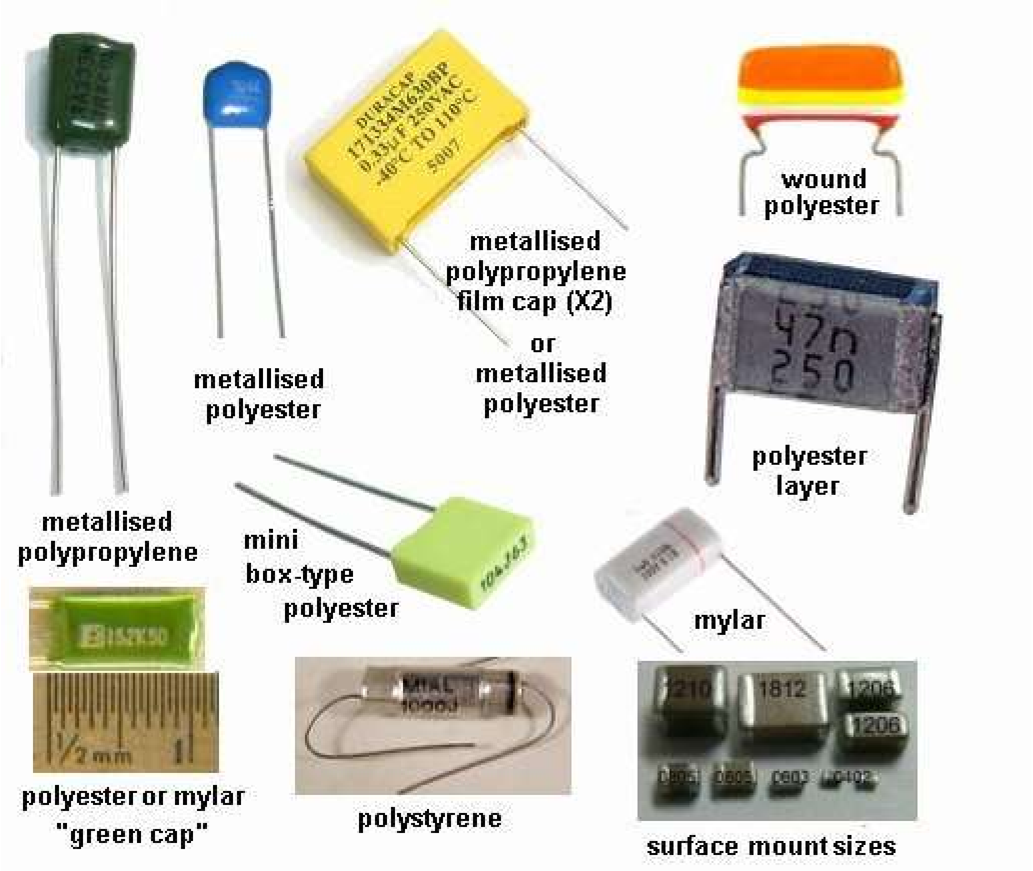

metallised polypropylene
metallised polyester
DURACAP
0.33µF 250VAC
-40°C TO 110°C
metallised polypropylene film cap (X2)
or
metallised polyester
wound polyester
47n
250
polyester layer
mini box-type polyester
104J63
mylar
B152K50
1/2 mm
polyester or mylar "green cap"
polystyrene
1210
1812
1206
1206
0805
0603
0603
0402
surface mount sizes

Capacitor Colour Code Table

Colour	Digit A	Digit B	Multiplier D	Tolerance (T) > 10pf	Tolerance (T) < 10pf	Temperature Coefficient (TC)
Black	0	0	x1	± 20%	± 2.0pF	
Brown	1	1	x10	± 1%	± 0.1pF	$-33x10^{-6}$
Red	2	2	x100	± 2%	± 0.25pF	$-75x10^{-6}$
Orange	3	3	x1,000	± 3%		$-150x10^{-6}$
Yellow	4	4	x10,000	± 4%		$-220x10^{-6}$
Green	5	5	x100,000	± 5%	± 0.5pF	$-330x10^{-6}$
Blue	6	6	x1,000,000			$-470x10^{-6}$
Violet	7	7				$-750x10^{-6}$
Grey	8	8	x0.01	+80%,-20%		
White	9	9	x0.1	± 10%	± 1.0pF	
Gold			x0.1	± 5%		

Silver			x0.01	± 10%		

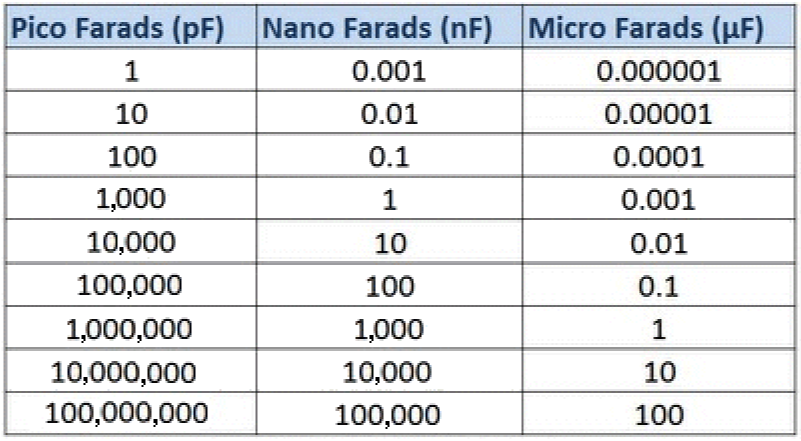

Pico Farads (pF)	Nano Farads (nF)	Micro Farads (µF)
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1
10,000,000	10,000	10
100,000,000	100,000	100

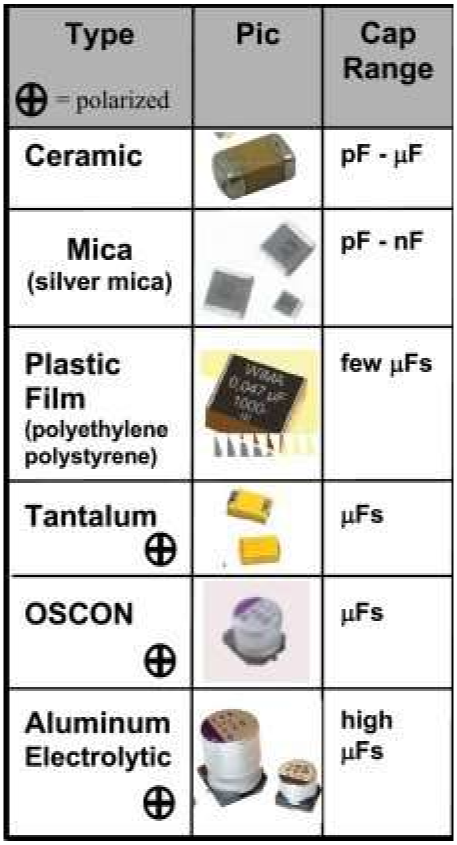

Type ⊕ = polarized	Pic	Cap Range
Ceramic		pF - µF
Mica (silver mica)		pF - nF
Plastic Film (polyethylene polystyrene)		few µFs
Tantalum ⊕		µFs
OSCON ⊕		µFs
Aluminum Electrolytic ⊕		high µFs

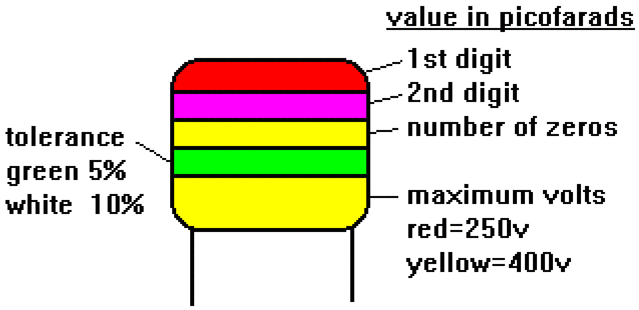
value in picofarads
1st digit
2nd digit
number of zeros
tolerance
green 5%
white 10%
maximum volts
red=250v
yellow=400v

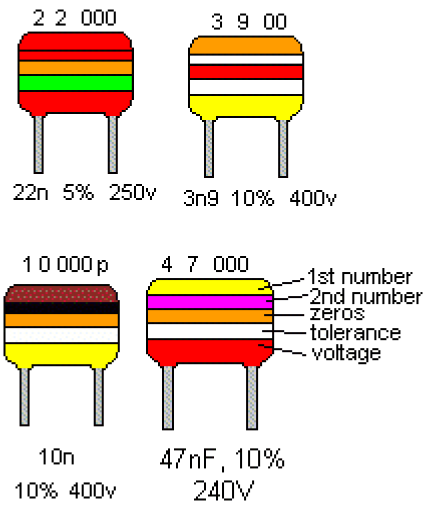
2 2 000
22n 5% 250v
3 9 00
3n9 10% 400v
1 0 000 p
10n
10% 400v
4 7 000
1st number
2nd number
zeros
tolerance
voltage
47nF, 10%
240V

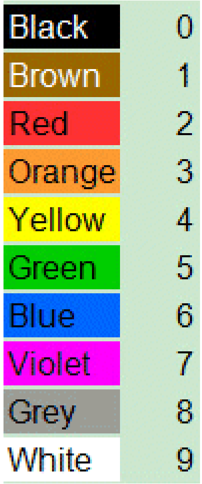

Black	0
Brown	1
Red	2
Orange	3
Yellow	4
Green	5
Blue	6
Violet	7
Grey	8
White	9

ELECTROLYTIC and TANTALUM CAPACITORS

Electrolytics and Tantalums are the same for testing purposes but their performance is slightly different in some circuits. A tantalum is smaller for the same rating as an electrolytic and has a better ability at delivering a current. They are available up to about 1,000u, at about 50v but their cost is much higher than an electrolytic.

Electrolytics are available in 1u, 2u2 3u3 4u7 10u, 22u, 47u, 100u, 220u, 330u, 470u, 1,000u, 2,200u, 3,300u, 4,700u, 10,000u and higher.
The "voltage" or "working voltage" can be: 3.3v, 10v, 16v, 25v, 63v, 100v, 200v and higher.
There is also another important factor that is rarely covered in text books. It is RIPPLE FACTOR.
This is the amount of current that can enter and leave an electrolytic. This current heats up the electrolytic and that is why some electrolytics are much larger than others, even though the capacitance and voltage-ratings are the same.
If you replace an electrolytic with a "miniature" version, it will heat up and have a very short life. This is especially important in power supplies where current (energy) is constantly entering and exiting the electrolytic as its main purpose is to provide a smooth output from a set of diodes that delivers "pulsing DC." (see "Power Diodes")

Remember:
104 = 100n
105 = 1u
106 = 10u
107 - 100u

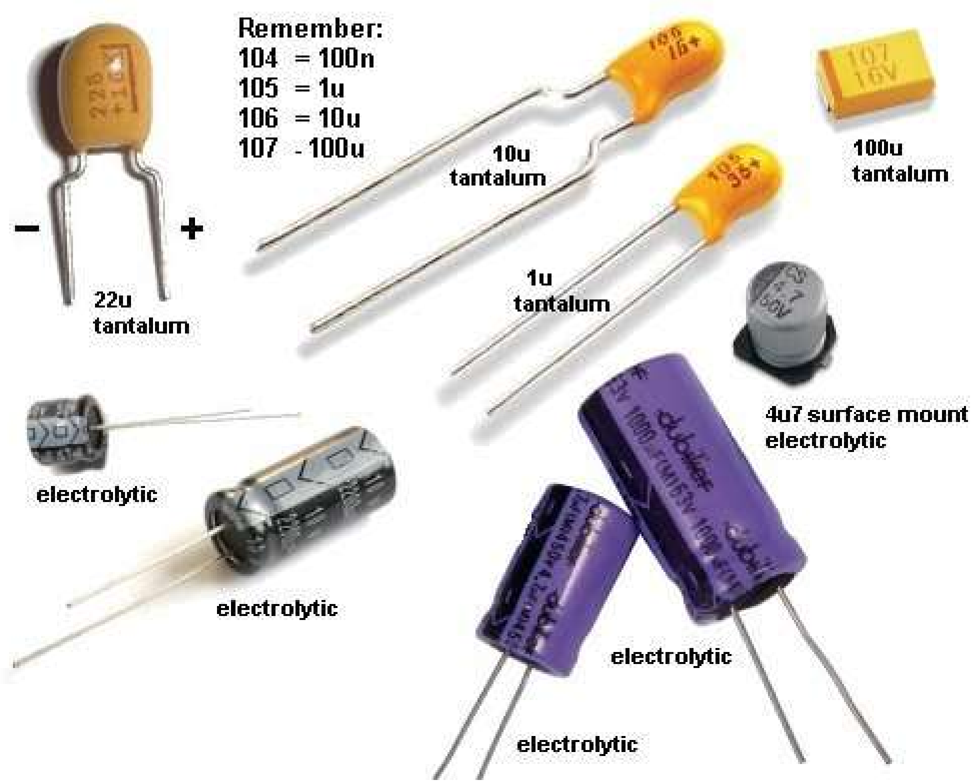

−
+
22u
tantalum
10u
tantalum
1u
tantalum
100u
tantalum
4u7 surface mount
electrolytic
electrolytic
electrolytic
electrolytic
electrolytic

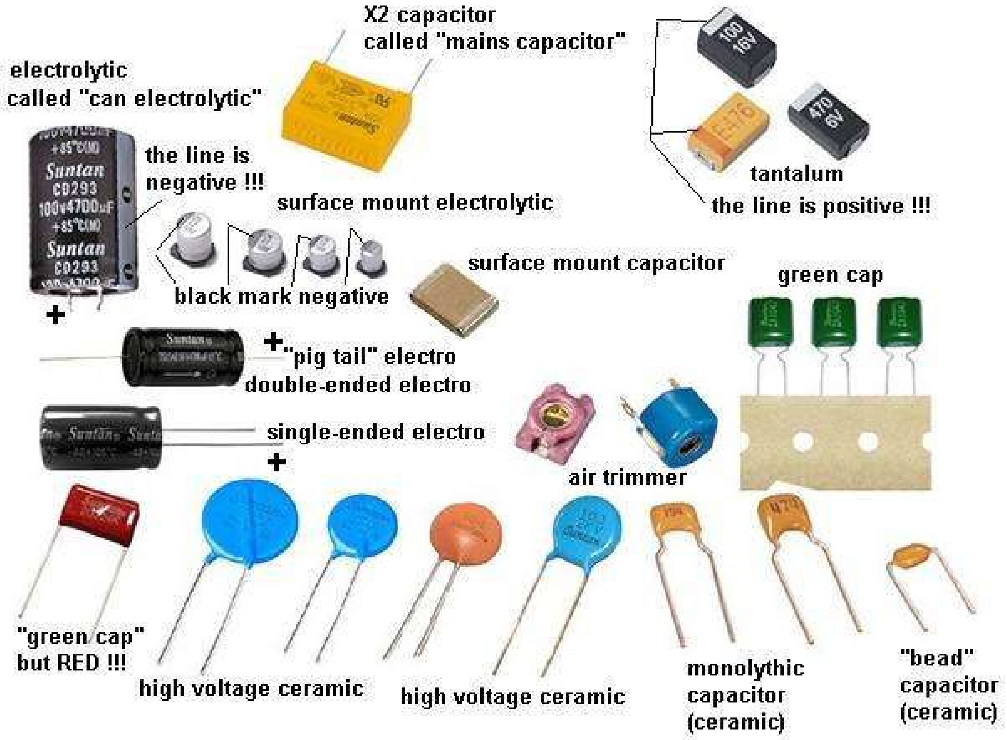

electrolytic
called "can electrolytic"
the line is
negative !!!
+
X2 capacitor
called "mains capacitor"
surface mount electrolytic
black mark negative
tantalum
the line is positive !!!
surface mount capacitor
green cap
+
"pig tail" electro
double-ended electro
single-ended electro
+
air trimmer
"green cap"
but RED !!!
high voltage ceramic
high voltage ceramic
monolythic
capacitor
(ceramic)
"bead"
capacitor
(ceramic)

tantalum capacitors can
"go up in smoke"

"bead" tantalum

PARALLEL and SERIES CAPACITORS

Capacitors can be connected in PARALLEL and/or SERIES for a number of reasons.

1. If you do not have the exact value, two or more connected in parallel or series can produce the value you need.
2. Capacitors connected in series will produce one with a higher voltage rating.
3. Capacitors connected in parallel will produce a larger-value capacitance.

Here are examples of two equal capacitors connected in series or parallel and the results they produce:

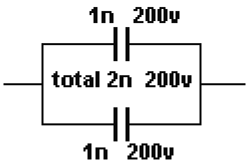
1n 200v
total 2n 200v
1n 200v

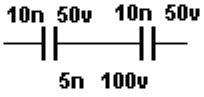
10n 50v
10n 50v
5n 100v

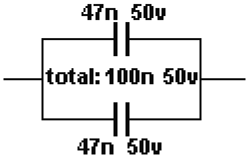
47n 50v
total: 100n 50v
47n 50v

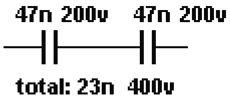
47n 200v
47n 200v
total: 23n 400v

NON-POLAR CAPACITORS (ELECTROLYTICS)

Electrolytics are also available in non-polar values. It sometimes has the letters "NP" on the component. Sometimes the leads are not identified.

This is an electrolytic that does not have a positive and negative lead but two leads and either lead can be connected to the positive or negative of the circuit.

These electrolytics are usually connected to the output of an amplifier (such as in a filter near the speaker) where the signal is rising and falling.

A non-polar electrolytic can be created from two ordinary electrolytics by connecting the negative leads together and the two positive leads become the new leads.

For example: two 100u 63v electrolytics will produce a 47u 63v non-polar electrolytic.

In the circuit below, the non-polar capacitor is replaced with two electrolytics.

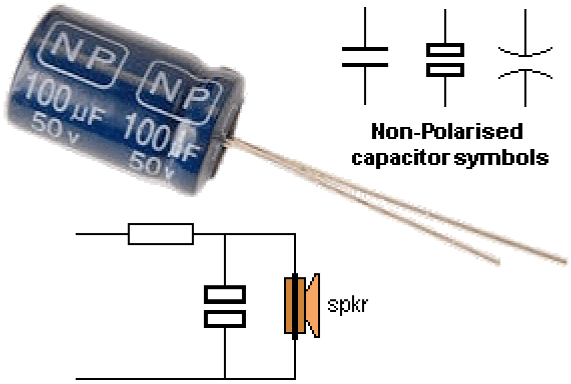
NP
100µF
50v
NP
100µF
50v
Non-Polarised
capacitor symbols
spkr

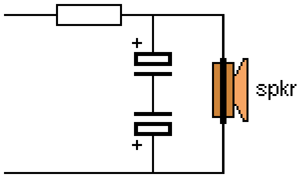

MAKING A NON-POLAR ELECTROLYTIC

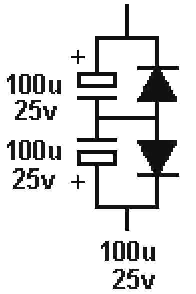

non-polar

A normal electrolytic must be connected the correct way in a circuit because it has a thin insulating layer covering the plates that has a high resistance.
If you connect the electrolytic around the wrong way, this layer "breaks-down" and the resistance of the electrolytic becomes very small and a high current flows. This heats up the electrolytic and the current increases. Very soon the capacitor produces gasses and explodes.
One big mistake in many text books shows how to make a non-polar electrolytic by connecting two "back-to-back."
They claim 2 x 100u connected back-to-back is equal to 47u.
This appears to be case when testing on a meter but the meter simply charges them for a short period of time to get a reading.
If you allow them to charge fully you will find the reverse electrolytic has a very small voltage across it.
Secondly, when you are charging them, you are putting a high current through the reverse electrolytic and damaging the layer.
To prevent this, you need to add two diodes as shown in the diagram.
In addition, 2 x 100u "back-to-back" is very near 100u.

Here is a question from a reader:
I have an amplifier with 2 x2,200u electrolytics on the output of a bridge. Can I replace them with a single 10,000u?

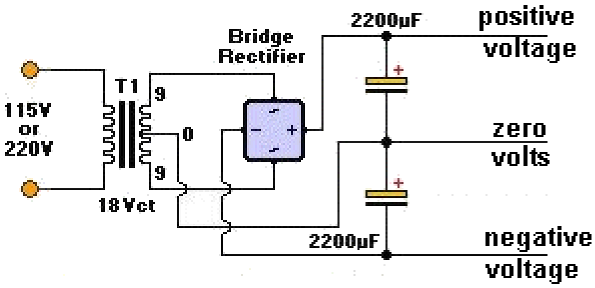

You need to look at the circuit of your amplifier. The two 2,200u electrolytics are possibly connected as shown in the circuit above and you will notice they are joined to produce a positive rail and a negative rail with zero (called earth) in the centre.
This forms two different circuits with the top electrolytic filtering the positive rail and the bottom electro filtering the negative rail. They must be connected to the zero volts rail.
A single 10,000u cannot be connected to the 0v rail and cannot be substituted for the two electro's.
You can easily determine of the two electro's are connected as shown above.
Test the positive terminal of each electro by placing the negative of the meter on the chassis.
If the positive of one electro have zero volts, it will be the lower electro in the diagram above. The negative terminal of the electro will have a minus voltage on it.

VOLTAGE RATING OF CAPACITOR
Capacitors have a voltage rating, stated as WV for working voltage, or WVDC. This specifies the maximum voltage that can be applied across the capacitor without puncturing the dielectric. Voltage ratings for "poly," mica and ceramic capacitors are typically 50v to 500 VDC. Ceramic capacitors with ratings of 1kv to 5kv are also available. Electrolytic capacitors are commonly available in 6v, 10v 16v, 25v, 50v, 100v, 150v, and 450v ratings.

THE SIZE OF A CAPACITOR - RIPPLE FACTOR

The size of a capacitor depends on a number of factors, namely the value of the capacitor (in microfarads etc) and the voltage rating. But there is also another factor that is most important. It is the RIPPLE FACTOR. **Ripple Factor** is the amount of voltage-fluctuation the capacitor (electrolytic) can withstand without getting too hot. When current flows in and out of an electrolytic, it gets hot and this will eventually dry-out the capacitor as some of the liquid inside the capacitor escapes through the seal. It's a very slow process but over a period of years, the capacitor looses its capacitance.
If you have two identical 1,000u 35v electrolytics and one is smaller, it will get hotter when operating in a circuit and that's why it is necessary to choose the largest electrolytic.

CAUTION

If a capacitor has a voltage rating of 63v, do not put it in a 100v circuit as the insulation (called the dielectric) will be punctured and the capacitor will "short-circuit." It's ok to replace a 0.22uF 50WV capacitor with 0.22uF 250WVDC.

SAFETY

A capacitor can store a charge for a period of time after the equipment is turned off. High voltage electrolytic caps can pose a safety hazard. These capacitors are in power supplies and some have a resistor across them, called a bleed resistor, to discharge the cap after power is switched off.
If a bleed resistor is not present the cap can retain a charge after the equipment is unplugged.

How to discharge a capacitor

Do not use a screwdriver to short between the terminals as this will damage the capacitor internally and the screwdriver.
Use a 1k 1 watt or 3watt or 5watt resistor on jumper leads (or held with pliers) and keep them connected for up to 15 seconds to fully discharge the electro.
Test it with a voltmeter to make sure all the energy has been removed.

Before testing any capacitors, especially electrolytics, you should look to see if any are damaged, overheated or leaking. Swelling at the top of an electrolytic indicates heating (and pressure inside the case) and will result in drying out of the electrolyte. Any hot or warm electrolytic indicates leakage and ceramic capacitors with portions missing indicates something has gone wrong (such as it being "blown apart").

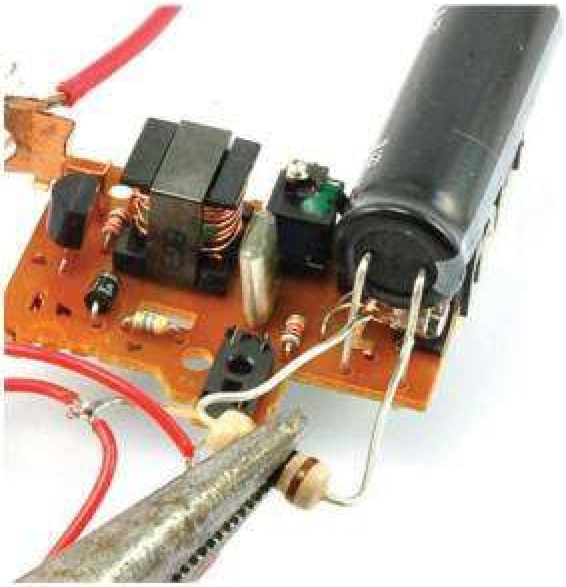

Here is a 120u 330v electrolytic from a flash circuit in an old-fashioned film camera.
If the flash does not "fire," the electrolytic will be charged to about 350 volts !!
Use a 1k resistor (held with pliers) to slowly discharge it. It may take 15 seconds to fully discharge.

TESTING A CAPACITOR

There are two things you can test with a multimeter:
1. A short-circuit within the capacitor
2. Capacitor values above 1u.

You can test capacitors in-circuit for short-circuits. Use the x1 ohms range.
To test a capacitor for leakage, you need to remove it or at least one lead must be removed. Use the x10k range on an analogue or digital multimeter.
For values above 1u you can determine if the capacitor is charging by using an analogue meter. The needle will initially move across the scale to indicate the cap is charging, then go to "no deflection." Any permanent deflection of the needle will indicate leakage.
You can reverse the probes to see if the needle moves in the opposite direction. This indicates it has been charged. Values below 1u will not respond to charging and the needle will not deflect.
This does not work with a digital meter as the resistance range does not output any current and the electrolytic does not charge.

Rather than spending money on a capacitance meter, it is cheaper to replace any suspect capacitor or electrolytic.
Capacitors can produce very unusual faults and no piece of test equipment is going to detect the problem.
In most cases, it is a simple matter to solder another capacitor across the suspect component and view or listen to the result.
This saves all the worry of removing the component and testing it with equipment that cannot possibly give you an accurate reading when the full voltage and current is not present.
It is complete madness to even think of testing critical components such as capacitors, with TEST EQUIPMENT. You are fooling yourself. If the Test Equipment says the component is ok, you will look somewhere else and waste a lot of time.

FINDING THE VALUE OF A CAPACITOR

If you want to find the value of a surface-mount capacitor or one where the markings have been removed, you will need a CAPACITANCE METER. Here is a simple circuit that can be added to your meter to read capacitor values from 10p to 10u.
The full article can be found HERE.

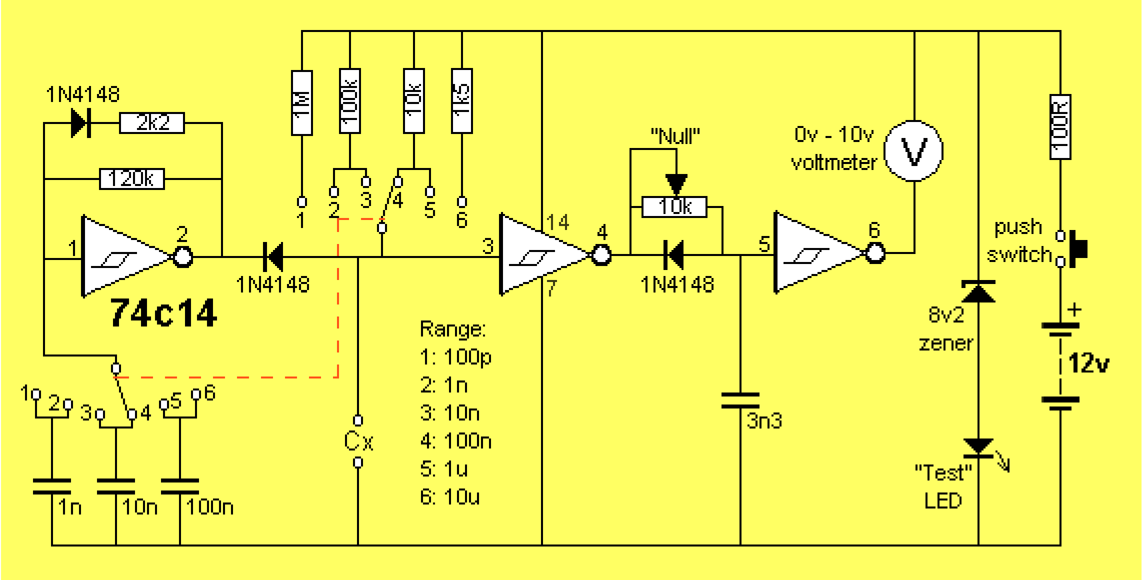

ADD-ON CAPACITANCE METER

CAPACITOR SUBSTITUTION BOX

CAPACITANCE SUBSTITUTION BOX
470
560
680
820
330
220
150
100
3300
2200
1500
1000
pF
µF
.01
.015
.022
.033
.047
.056
.068
.1
.0047
.0056
.0068
.0082
ELENCO® MODEL CS-440

You can get a kit or a ready-made piece of test gear called **CAPACITOR SUBSTITUTION BOX** and also **RESISTOR SUBSTITUTION BOX**.
I bought one of each 30 years ago and I have only used them ONCE.
They appear to be very handy but when you are testing a circuit, you want the component next to the other parts.
It is just as easy to pick the component you need from your junk box and connect it to the circuit via jumper leads.

REPLACING A CAPACITOR

Always replace a capacitor with the exact same type.
A capacitor may be slightly important in a circuit or it might be extremely critical.
A manufacturer may have taken years to select the right type of capacitor due to previous failures.
A capacitor just doesn't have a "value of capacitance."
It may also has an effect called "tightening of the rails."
In other words, a capacitor has the ability to react quickly and either absorb or deliver energy to prevent spikes or fluctuations on the rail.
This is due to the way it is constructed. Some capacitors are simply plates of metal film while others are wound in a coil. Some capacitors are large while others are small.
They all react differently when the voltage fluctuates.
Not only this, but some capacitors are very stable and all these features go into the decision for the type of capacitor to use.
You can completely destroy the operation of a circuit by selecting the wrong type of capacitor.
No capacitor is perfect and when it gets charged or discharged, it appears to have a small value of resistance in series with the value of capacitance. This is known as "ESR" and stands for EQUIVALENT SERIES RESISTANCE. This effectively makes the capacitor slightly slower to charge and discharge.
We cannot go into the theory on selecting a capacitor as it would be larger than this eBook so the only solution is to replace a capacitor with an identical type.
However if you get more than one repair with identical faults, you should ask other technicians if the original capacitor comes from a faulty batch.
The author has fixed TV's and fax machines where the capacitors have been inferior and alternate types have solved the problem.
Some capacitor are suitable for high frequencies, others for low frequencies.

DECOUPLING CAPACITORS

A Decoupling Capacitor can severe one, two or three functions. You need to think of a decoupling capacitor as a miniature battery with the ability to deliver a brief pulse of energy when ever the line-voltage drops and also absorb a brief pulse of energy when ever the line voltage rises (or spikes).
Decoupling capacitor can range from 100n to 1,000u.
100n capacitors are designed to absorb spikes and also have the effect of tightening-

up the rails for high frequencies. They have no effect on low frequencies such as audio frequencies.

These capacitors are generally ceramic and have very low internal impedance and thus they can operate at high frequencies.

Capacitors above about 10u are used for decoupling and these are nearly always electrolytics.

Decoupling means "tightening-up the power rails." The electrolytic acts just like a miniature rechargeable battery, supplying a small number of components in a circuit with a smooth and stable voltage.

The electrolytic is usually fed from a dropper resistor and this resistor charges the electrolytic and adds to the ability of the electrolytic to create a "separate power supply."

These two components help remove spikes as an electrolytic cannot remove spikes if connected directly to the supply rails - it's internal impedance is high and the spikes are not absorbed.

Decoupling capacitors are very difficult to test.

They rarely fail but if a project is suffering from unknown glitches and spikes, it is best to simply add more 100n decoupling caps on the underside of the board and replace all electrolytics.

Some small electrolytics will dry out due to faulty manufacture and simply replacing every one on a board will solve the problem.

Some of the functions of a decoupling capacitor are:

Removing ripple - hum or buzz in the background of an amplifier

Removing glitches or spikes.

Separating one stage from another to reduce or remove MOTORBOATING - a low frequency sound due to the output putting a pulse on the power rails that is picked up by the pre-amplifier section and amplified.

TESTING DIODES

Diodes can have 4 different faults.

1. Open circuit in both directions.
2. Low resistance in both directions.
3. Leaky.
4. Breakdown under load.

TESTING A DIODE ON AN ANALOGUE METER

Testing a diode with an **Analogue Multimeter** can be done on any of the resistance ranges. [The high resistance range is best - it sometimes has a high voltage battery for this range but this does not affect our testing]

There are two things you must remember.

1. When the diode is measured in one direction, the needle will **not move at all**. The technical term for this is the diode is **reverse biased**. It will not allow any current to flow. Thus the needle will not move.

When the diode is connected around the other way, the needle will swing to the right (move up scale) to about 80% of the scale. This position represents the voltage drop across the junction of the diode and is NOT a resistance value. If you change the resistance range, the needle will move to a slightly different position due to the resistances inside the meter. The technical term for this is the diode is **forward biased**. This indicates the diode is not faulty.

The needle will swing to a slightly different position for a "normal diode" compared to a Schottky diode. This is due to the different junction voltage drops.

However we are only testing the diode at very low voltage and it may break-down when fitted to a circuit due to a higher voltage being present or due to a high current flowing.

2. The leads of an **Analogue Multimeter** have the positive of the battery connected to the black probe and the readings of a "good diode" are shown in the following two diagrams:

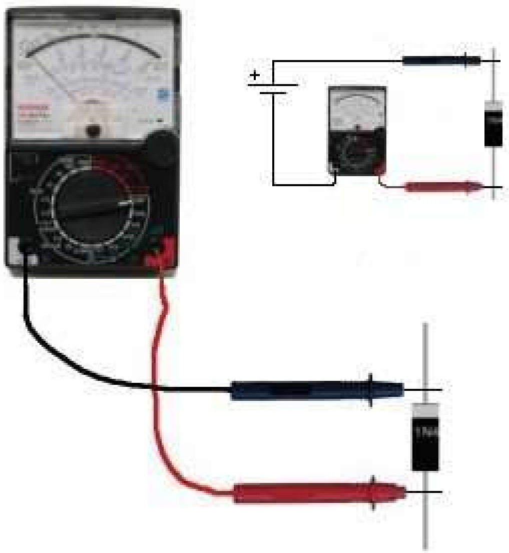

The diode is **REVERSE BIASED** in the diagram above and diodes not conduct.

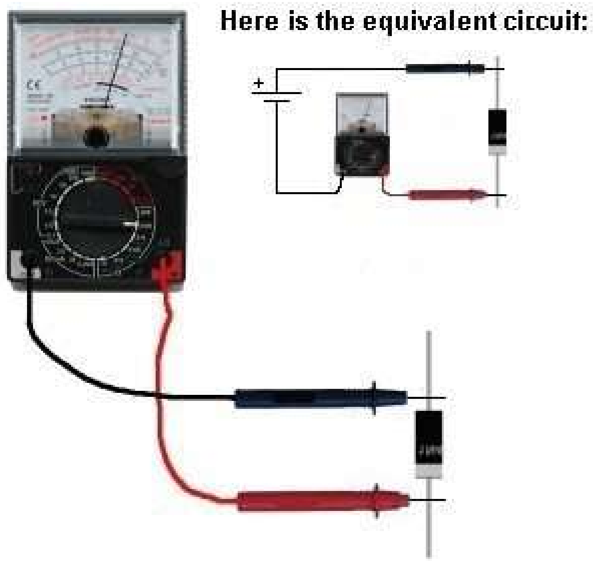

The diode is **FORWARD BIASED** in the diagram above and it conducts

TESTING A DIODE ON A DIGITAL METER

Testing a diode with a Digital Meter must be done on the "DIODE" setting as a digital meter does not deliver a current through the probes on some of the resistance settings and will not produce an accurate reading.

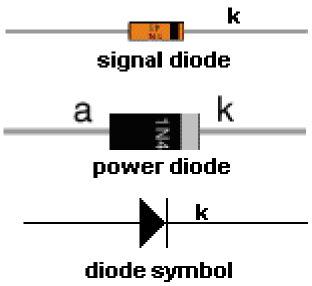

The best thing to do with a "suspect" diode is to replace it. This is because a diode has a number of characteristics that cannot be tested with simple equipment. Some diodes have a fast recovery for use in high frequency circuits. They conduct very quickly and turn off very quickly so the waveform is processed accurately and

efficiently.
If the diode is replaced with an ordinary diode, it will heat up as does not have the high-speed characteristic.
Other diodes have a low drop across them and if an ordinary is used, it will heat up.
Most diodes fail by going: SHORT-CIRCUIT. This can be detected by a low resistance (x1 or x10 Ohms range) in both directions.
A diode can also go OPEN CIRCUIT. To locate this fault, place an identical diode across the diode being tested.
A leaky diode can be detected by a low reading in one direction and a slight reading the other direction.
However this type of fault can only be detected when the circuit is working. The output of the circuit will be low and sometimes the diode heats up (more than normal).
A diode can go open under full load conditions and perform intermittently.
Diodes come in pairs in surface-mount packages and 4 diodes can be found in a bridge.
They are also available in pairs that look like a 3-leaded transistor.
The line on the end of the body of a diode indicates the cathode and you cannot say "this is the positive lead." The correct way to describe the leads is to say the "cathode lead." The other lead is the anode. The cathode is defined as the electrode (or lead) through which an electric current flows out of a device.
The following diagrams show different types of diodes:

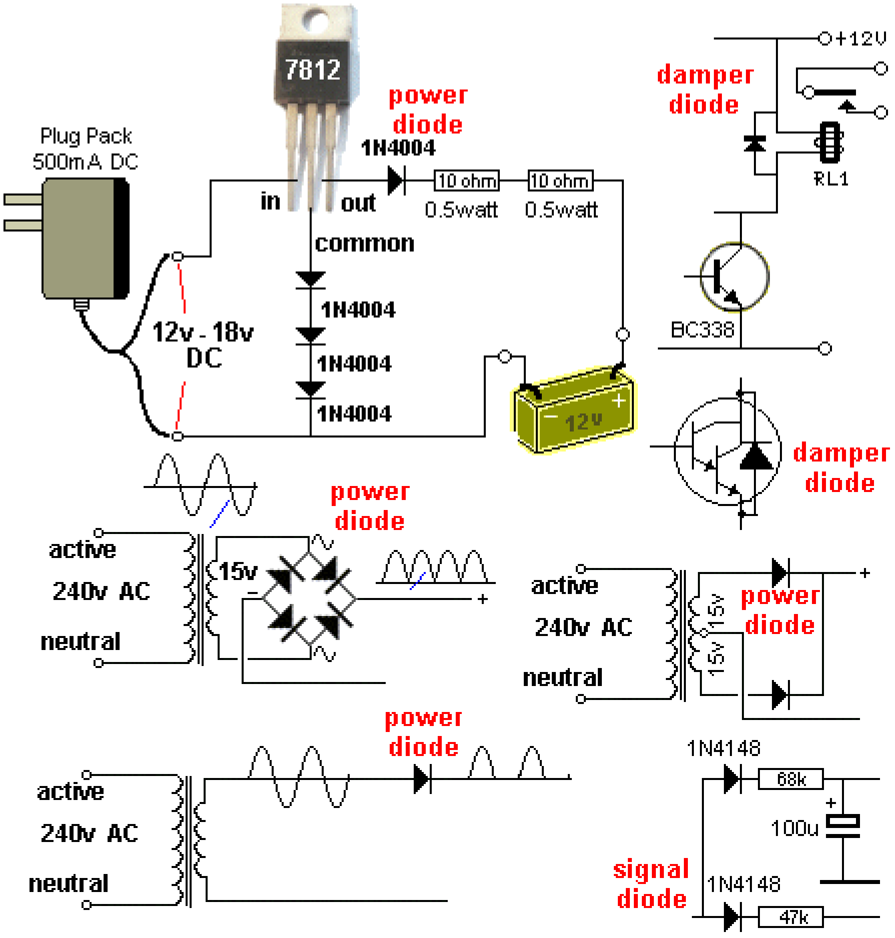
7812
power diode
1N4004
Plug Pack
500mA DC
in
out
common
10 ohm
0.5watt
10 ohm
0.5watt
12v - 18v
DC
1N4004
1N4004
1N4004
12v
damper diode
+12V
RL1
BC338
damper diode
power diode
active
240v AC
neutral
15v
+
active
240v AC
neutral
15v
15v
power diode
+
power diode
active
240v AC
neutral
1N4148
68k
+
100u
signal diode
1N4148
47k

POWER DIODES

To understand how a power diode works, we need to describe a few things. This has NEVER been described before, so read carefully.
The 240v AC (called the "mains") consists of two wires, one is called the ACTIVE and the other is NEUTRAL. Suppose you touch both wires. You will get a shock. The neutral is connected to an earth wire (or rod driven into the ground or connected to a water pipe) at the point where the electricity enters the premises and you do not get a shock from the NEUTRAL.

But the voltage on the active is rising to +345v then goes to -345v at the rate of 50 times per second (for a complete cycle).
345v is the peak voltage of 240v. You never get a 240v shock. (It is a 345v shock.) In other words, if you touch the two wires at a particular instant, you would get a POSITIVE 345v shock and at another instant you would get a negative 345v shock. This is shown in the diagram below.
We now transfer this concept to the output of a transformer. The diagram shows an AC waveform on the output of the secondary.
This voltage is rising 15v higher than the bottom lead then it is 15v LOWER than the bottom lead. The bottom lead is called "zero volts." You have to say one lead or wire is not "rising and falling" as you need a "reference" or starting-point" or "zero point" for voltage measurements.
The diode only conducts when the voltage is "above zero" (actually when it is 0.7v above zero) and does not conduct (at all) when the voltage goes below zero.
This is shown on the output of the Power Diode. Only the positive peaks or the positive parts of the waveform appear on the output and this is called "pulsing DC." This is called "half-wave" and is not used in a power supply. We have used it to describe how the diode works. The electrolytics charge during the peaks and deliver energy when the diode is not delivering current. This is how the output becomes a steady DC voltage.
Power supplies use FULL WAVE rectification and the other half of the AC waveform is delivered to the output (and fills in the "gaps") and appears as shown in "A."

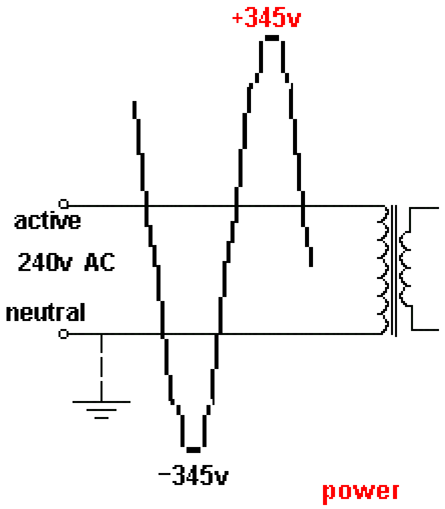

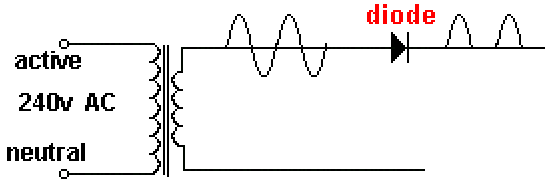

Half-wave rectifier

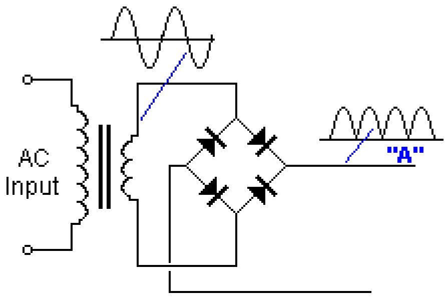

Full-wave rectifier

ONE FAULTY DIODE

One diode in a bridge can go open (any of the 4 diodes will produce the same effect) and produce an output voltage that can be slightly lower than the original voltage.

The actual "voltage-drop" will depend on the current taken by the circuit and the ability of the transformer to produce the required voltage and current during half-wave operation. The voltage during each half cycle (when none of the diodes is delivering any energy to the circuit) is maintained by the electrolytic and its size (relative to the current taken by the circuit) will determine the size of the ripple that will result when the diode fails. The ripple will be 100 to 1,000 times greater after the failure of a diode, depending on the value of the filter capacitor.
To locate the faulty diode, simply get a diode and place it across each of the diodes in the bridge (in turn) when the circuit is working.
For a bridge rectifier, the ripple-frequency will be twice the mains frequency and its ripple will be very small if the electrolytic is the correct value. When a diode fails, the ripple-frequency will be equal to mains-frequency and the amplitude will increase considerably. You may even hear background hum from audio equipment.
If you cannot find a faulty diode, the filter capacitor will be at fault. Turn off the equipment and connect an electrolytic across the filter capacitor via jumper leads. Turn the power ON and see if the hum has reduced.

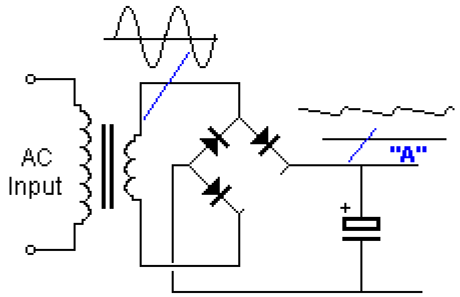

One diode "burnt-out"

DAMPER DIODES

A damper diode is a diode that detects a high voltage and SQUELCHES IT (reduces it - removes it). The signal that it squelches is a voltage that is in the opposite direction to the "supply voltage" and is produced by the collapsing of a magnetic field. Whenever a magnetic filed collapses, it produces a voltage in the winding that is opposite to the supply voltage and can be much higher. This is the principle of a flyback circuit or EHT circuit. The high voltage comes from the transformer.
The diode is placed so that the signal passes through it and less than 0.5v appears across it.
A damper diode can be placed across the coil of a relay, incorporated into a transistor or FET or placed across a winding of a flyback transformer to protect the driving transistor or FET.
It can also be called a "Reverse-Voltage Protection Diode," "Spike Suppression Diode," or "Voltage Clamp Diode."
The main characteristic of a Damper Diode is HIGH SPEED so it can detect the spike and absorb the energy.
It does not have to be a high-voltage diode as the high voltage in the circuit is being absorbed by the diode.

SILICON, GERMANIUM AND SCHOTTKY DIODES

When testing a diode with an analogue meter, you will get a low reading in one direction and a high (or NO READING) in the other direction. When reading in the LOW direction, the needle will swing nearly full scale and the reading is not a resistance-value but a reflection of the characteristic voltage drop across the junction of the diode. As we mentioned before, a resistance reading is really a voltage reading and the meter is measuring the voltage of the battery minus the voltage-drop across the diode.
Since Silicon, Germanium and Schottky Diodes have slightly different characteristic voltage drops across the junction, you will get a slightly different reading on the scale. This does not represent one diode being better than the other or capable of handling a higher current or any other feature.
The quickest, easiest and cheapest way to find, fix and solve a problem caused by a

faulty diode is to replace it.
There is no piece of test equipment capable of testing a diode fully, and the circuit you are working on is actually the best piece of test equipment as it is identifying the fault UNDER LOAD.

Only very simple tests can be done with a multimeter and it is best to check a diode with an ANALOGUE MULTIMETER as it outputs a higher current though the diode and produces a more-reliable result.
A Digital meter can produce false readings as it does not apply enough current to activate the junction.
Fortunately almost every digital multimeter has a **diode test mode.** Using this, a silicon diode should read a voltage drop between 0.5v to 0.8v in the forward direction and open in the reverse direction. For a germanium diode, the reading will be lower, around 0.2v - 0.4v in the forward direction. A bad diode will read zero volts in both directions.

REPLACING A DIODE

It is alway best to replace a diode with the same type but quite often this is not possible. Many diodes have unusual markings or colours or "in-house" letters.
This is only a general guide because many diodes have special features, especially when used in high-frequency circuits.
However if you are desperate to get a piece of equipment working, here are the steps:
Determine if the diode is a signal diode, power diode, or zener diode.
For a signal diode, try 1N4148.
For a power diode (1 amp) try 1N4004. (for up to 400v)
For a power diode (3 amp) try 1N5404. (for up to 400v)
For a high-speed diode, try UF4004 (for up to 400v)

If you put an ordinary diode in a high-speed application, it will get very hot very quickly.
To replace an unknown zener diode, start with a low voltage such as 6v2 and see if the circuit works.
The size of a diode and the thickness of the leads will give an idea of the current-capability of the diode.
Keep the leads short as the PC board acts as a heat-sink.
You can also add fins to the leads to keep the diode cool.

LIGHT EMITTING DIODES (LEDs)

Light Emitting Diodes (LEDs) are diodes that produce light when current flows from anode to cathode. The LED does not emit light when it is revered-biased. It is used as a low current indicator in many types of consumer and industrial equipment, such as monitors, TV's, printers, hi-fi systems, machinery and control panels.
The light produced by a LED can be visible, such as red, green, yellow or white. It can also be invisible and these LEDs are called Infrared LEDs. They are used in remote controls and to see if they are working, you need to point a digital camera at the LED and view the picture on the camera screen.
An LED needs about 2v - 3.6v across its leads to make it emit light, but this voltage must be exact for the type and colour of the LED. The simplest way to deliver the exact voltage is to have a supply that is higher than needed and include a voltage-dropping resistor. The value of the resistor must be selected so the current is between 2mA and 25mA.
The cathode of the LED is identified by a flat on the side of the LED. The life expectancy of a LED is about 100,000 hours. LEDs rarely fail but they are very sensitive to heat and they must be soldered and de-soldered quickly. They are one of the most heat-sensitive components.
Light emitting diodes cannot be tested with most multimeters because the characteristic voltage across them is higher than the voltage of the battery in the meter.
However a simple tester can be made by joining 3 cells together with a 220R resistor

and 2 alligator clips:

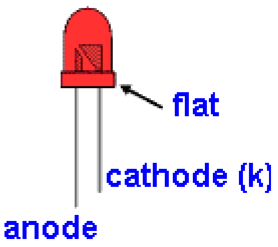

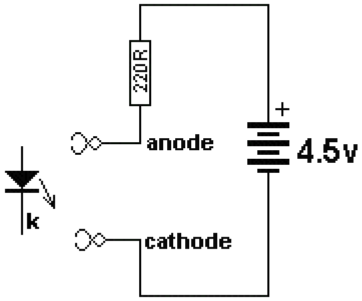

LED TESTER

Connect the clips to a LED and it will illuminate in only one direction.
The colour of the LED will determine the voltage across it. You can measure this voltage if you want to match two or more LEDs for identical operation.
Red LEDs are generally 1.7v to 1.9v. - depending on the quality such as "high-bright"
Green LEDs are 1.9v to 2.3v.
Orange LEDs are about 2.3v and
White LEDs and IR LEDs are about 3.3v to 3.6v.
The illumination produced by a LED is determined by the quality of the crystal. It is the crystal that produces the colour and you need to replace a LED with the same quality to achieve the same illumination.
Never connect a LED across a battery (such as 6v or 9v), as it will be instantly damaged. You must have a resistor in series with the LED to limit the current.

ZENER DIODES

All diodes are Zener diodes. For instance a 1N4148 is a 120v zener diode as this is its reverse breakdown voltage.
And a zener diode can be used as an ordinary diode in a circuit with a voltage that is below the zener value.
For instance, 20v zener diodes can be used in a 12v power supply as the voltage never reaches 20v, and the zener characteristic is never reached.
Most diodes have a reverse breakdown voltage above 100v, while most zeners are below 70v. A 24v zener can be created by using two 12v zeners in series and a normal diode has a characteristic voltage of 0.7v. This can be used to increase the voltage of a zener diode by 0.7v. See the diagram above. It uses 3 ordinary diodes to increase the output voltage of a 3-terminal regulator by 2.1v.
To tests a zener diode you need a power supply about 10v higher than the zener of the diode. Connect the zener across the supply with a 1k to 4k7 resistor and measure the voltage across the diode. If it measures less than 1v, reverse the zener. If the reading is high or low in both directions, the zener is damaged.

Here is a zener diode tester. The circuit will test up to 56v zeners.

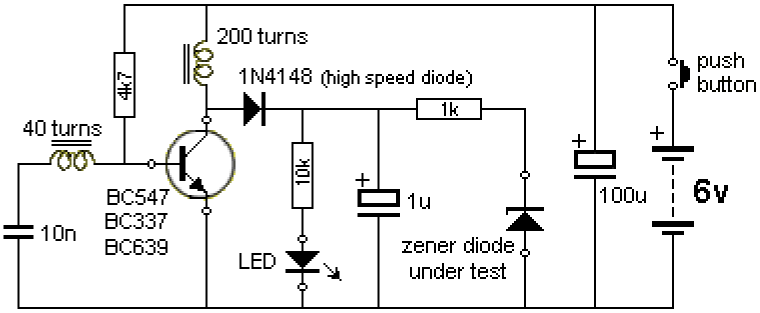

ZENER DIODE TESTER

TRANSFORMERLESS POWER SUPPLY
Here's a circuit that uses zener diodes in a power supply to show how they work. This clever design uses 4 diodes in a bridge to produce a fixed voltage power supply capable of supplying 35mA.
If we put 2 zener diodes in a bridge with two ordinary power diodes, the bridge will break-down at the voltage of the zener. This is what we have done. If we use 18v zeners, the output will be 17v4.

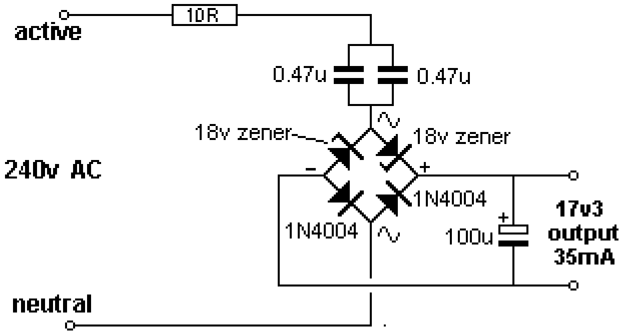

SUPPLY USING ZENER DIODES

When the incoming voltage is positive at the top, the left zener provides 18v limit (and the other zener produces a drop of 0.6v). This allows the right zener to pass current just like a normal diode. The output is 17v4. The same with the other half-cycle.
You cannot use this type of bridge in a normal power supply as the zener diode will "short" when the input voltage reaches the zener value. The concept only works in the circuit above.

VOLTAGE REGULATORS

A Voltage Regulator takes a high input voltage and delivers a fixed output voltage. Providing the input voltage is 4v above the output voltage, the regulator will deliver a fixed output voltage with almost no ripple.
Voltage regulators are also called "3-TERMINAL REGULATORS" or "REGULATOR IC's" - although this name is not generally used.
In most cases, a voltage regulator gets quite hot and for this reason it has a high failure-rate.
If a regulator is not getting hot (or warm) it has either failed or the circuit is not operating.
A regulator can only decrease the voltage. It cannot increase the current. This means the current being supplied to a circuit must also be available from the circuit supplying the regulator.
All regulators have different pin-outs, so you need to find the input pin and output pin and make sure the voltage-difference is at least 4v. Some regulators will work with a difference as low as 1v, so you need to read the specifications for the type you are servicing.
Some regulators are called "negative voltage regulators" and the input voltage will be negative and the output will be negative.
You need to test a voltage regulator with the power "ON".
Make sure you do not allow the probes to short any of the pins together as this will destroy the regulator or the circuit being supplied.
With the power turned off or the regulator removed from the circuit, you can test it with a multimeter set to resistance to see if it is ok. If any resistance readings are very low or zero ohms, the regulator is damaged.

TRANSFORMERS

All transformers and coils are tested the same way. This includes chokes, coils,

inductors, yokes, power transformers, EHT transformers (flyback transformers), switch mode transformers, isolation transformers, IF transformers, baluns, and any device that has turns of wire around a former. All these devices can go faulty.
The coating on the wire is called insulation or "enamel" and this can crack or become overheated or damaged due to vibration or movement. When two turns touch each other, a very interesting thing happens. **The winding becomes two separate windings.**

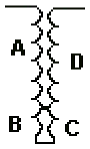

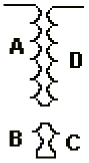

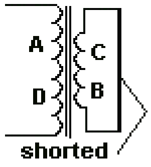

We will take the case of a single winding such as a coil. This is shown in the first diagram above and the winding is wound across a former (a former is a bobbin or plastic molding or something to hold the winding) and back again, making two layers. The bottom and top layers touch at the point shown in the diagram and the current that originally passed though A, B, C, D now passes though A & D.

Winding B C becomes a separate winding as shown in the second diagram.
In other words the coil becomes a TRANSFORMER with a SHORT CIRCUIT on the secondary winding as shown in the third diagram.
When the output wires of a transformer are shorted together, it delivers a very high current because you have created a SHORT-CIRCUIT. This short-circuit causes the transformer to get very hot.
That's exactly what happens when any coil or transformer gets a "shorted turn."
The shorted turns can be a single turn or many turns.
It is not possible to measure a fault like this with a multimeter as you don't know the exact resistance of a working coil or winding and the resistance of a faulty winding may be only 0.001 ohms less.
However when a transformer or coil is measured with an inductance meter, an oscillating voltage (or spike) is delivered into the core as magnetic flux, then the magnetic flux collapses and passes the energy into the winding to produce a waveform. The inductance meter reads this and produces a value of inductance in Henry (or milliHenry or microHenry.)
This is done with the transformer removed from the circuit and this can be a very difficult thing to do, as most transformers have a number of connections.
If the coil or transformer has a shorted turn, the energy from the magnetic flux will pass into the turns that are shorted and produce a current. Almost no voltage will be detected from winding.
The reading from the inductance meter will be low or very low and you have to work out if it is correct.
However there is one major problem with measuring a faulty transformer or coil.
It may only become faulty when power is applied.
The voltage between the turns may be sparking or jumping a gap and creating a problem. A tester is not going to find this fault.
Secondly, an inductance meter may produce a reading but you do not know if the reading is correct. An improved tester is a RING TESTER.
The circuit for a ring tester can be found here:

http://www.flippers.com/pdfs/k7205.pdf

It sends a pulse to the coil and counts the number of returning pulses or "rings." A faulty coil (or winding) may return one pulse but nearly all the energy will be passed to the shorted turns and you will be able to see this on the scale. You will only get one or two return pulses, whereas a good winding will return more pulses.

One way to detect a faulty power transformer is to connect it to the supply and feel the temperature-rise (when nothing is connected to the secondary).
It should NOT get hot.
Detecting shorted turns is not easy to diagnose as you really need another identical

component to compare the results.
Most transformers get very hot when a shorted turn has developed. It may deliver a voltage but the heat generated and a smell from the transformer will indicate a fault.

ISOLATION TRANSFORMER

An isolation transformer is a piece of **Test Equipment** that provides "Mains Voltage" but the voltage is "floating." You will still get a shock if you touch the two output leads, but it has a special use when testing unknown equipment.
Many electrical appliances are fully insulated and only have two leads connected to the mains.
When you take these appliances apart, you do not know which end of say a heating element is connected to the "live" (active) side of the mains and which end connects to the neutral.
I am not suggesting you carry out the following tests, but they are described to show how an isolation transformer works.
If you touch a soldering iron on the "live" (active) end of the heating element it will create a short-circuit.
However when the appliance is connected to the mains via an **isolation transformer**, you can touch an earthed soldering iron on either end of the heater as both leads from the isolation transformer are "floating."
Note: As soon as you earth one lead of the output an isolation transformer, the other lead becomes "active."
You can make your own **Isolation Transformer** by connecting two identical transformers "back-to-back."
The following diagram shows how this is done:

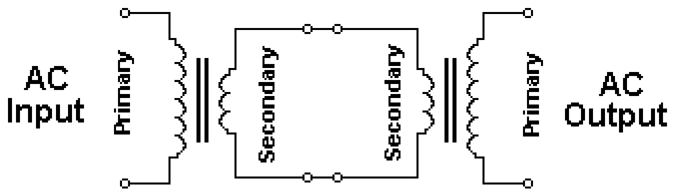

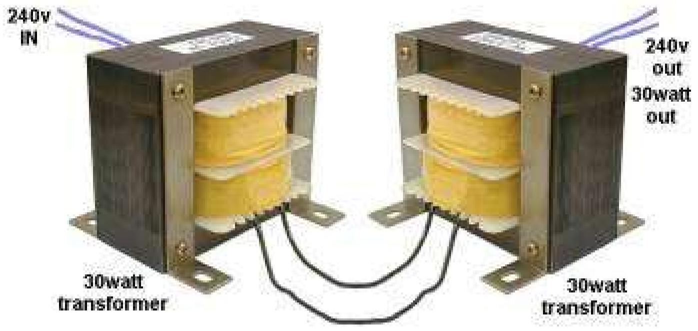

You can use any transformers providing the primary and secondary voltages are the same. The current capability of the secondary winding does not matter. However if you want a supply that has almost the same voltage as your "Mains," you need two transformers with the same voltages.
This handy isolation transformer will provide you with "Mains Voltage" but with a limited current.
In other words it will have a limited capability to supply "wattage." If you are using two 15VA transformers, you will only be able to test an appliance rated at 15 watts.
This has some advantages and some disadvantages.
If you are working on a project, and a short-circuit occurs, the damage will be limited to 15 watts.
If you are using two transformers with different VA ratings, the lower rating will be the capability of the combination.
If the secondaries are not equal, you will get a higher or lower "Mains Voltage."
If you get two transformers from TVs or Monitors, with a rating on the compliance plate of 45 watts, or 90 watts, you can assume the transformers are capable of delivering this wattage and making an isolation transformer will enable you to test

similar items with the safety of being isolated from the mains.
Colin Mitchell designs a lot of "LED lighting lamps" that are connected directly to the mains. He always works with an isolating transformer, just to be safe. Working on exposed "mains" devices is extremely nerve-wracking and you have to be very careful.
The isolation transformer will prevent a BIG EXPLOSION.

DETERMINING THE SPECS OF A TRANSFORMER
Suppose you have a "mains transformer" with unknown output voltages and unknown current capability.
You must be sure it is a mains transformer designed for operation on 50Hz or 60Hz. Switch-Mode transformers operate at frequencies 40kHz and higher and are not covered in this discussion.
To be on the safe-side, connect the unknown transformer to the output of your isolating transformer.
Since the transformer will take almost no current when not loaded, the output voltages it produces will be fairly accurate. Measure the input AC voltage and output AC voltage.
If the transformer has loaded your isolating transformer it will be faulty.
Mains transformers are approx 15VA for 500gm, 30VA for 1kgm 50VA for 2kgm and and 100VA for 2.5kgm.
VA stands for Volts-Amps and is similar to saying watts. Watts is used for DC circuits, while VA refers to AC circuits.
Once you have the weight of the transformer and the output voltage, you can work out the current capability of the secondary.
For transformers up to 30vA, the output voltage on no-load is 30% higher than the final "loaded voltage."
This is due to the poor regulation of these small devices.
If the transformer is 15VA and the output voltage is 15v AC, the current will be 1 amp AC.
You can check the "quality" of the transformer, (the regulation) by fully loading the output and measuring the final voltage. If the transformer has a number of secondaries, the VA rating must be divided between all the windings.

240v to 110v ISOLATION TRANSFORMER

Here's how to create a 110v isolating transformer:
Find a 240v:12v transformer.
Now find a transformer that has two secondary windings, such as 240v:12v+12v.

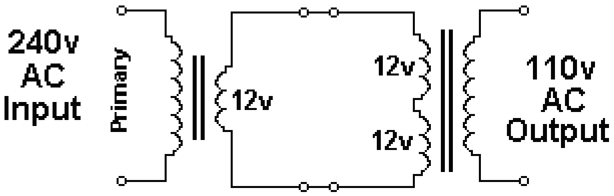

Connect the two transformers as shown in the circuit above. If the output is zero, connect ONE of the 12v windings of the second transformer around the other way.

110v to 240v ISOLATION TRANSFORMER

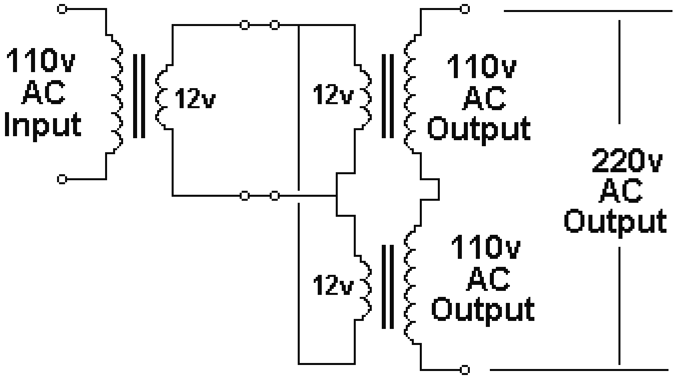

A 110v to 240v isolation transformer can be created by connecting 3 identical transformers as shown in the diagram above. If the output is zero, connect one of the outputs around the other way.

TRANSFORMER RATINGS

Question from a reader:
I have a 28v - 0 - 28v transformer @3amps. Does this mean each side is 1.5 amps?
The transformer is called CENTRE TAPPED and is shown in figures B and C.
It is designed to be connected to two diodes so each winding takes it in turn to deliver the current as shown in diagram C and the output will be 28v AC at 3 amps.
The 28v and 3 amp are both AC values.
If you connect across both outside wires, the output will be 56v at 1.5 amp.
This is because the transformer has a VA rating of 28 x 3 = 84VA. This is very similar to the term "watts."
When the 28v AC is rectified and smoothed, it becomes 28 x 1.4 = 39v (minus 0.6v across the diode) and since the transformer has a rating of 84 VA, the current must be reduced to 84/39 = 2.1 amps to maintain the VA rating.
Some transformers are specified as say: 12v - 0 - 12v, but the wiring diagram is shown as "A." This transformer should be specified as 12v + 12v as the secondaries are separate.
12v - 0 - 12v means the two secondary windings are NOT separate.
It does not make any difference to the output voltage and current, if the windings are separate or joined. The only difference is 12v +12v can be turned into two separate 12v outputs.
If you do not know the output current for a particular transformer, go to the website of electronic parts suppliers and compare the weight of your transformer with others. This will give you a VA rating and you can work out the current, once you know the output voltage.
Note: the output current finishes up ONLY 60% of the rating on the transformer tag because the rating is an **AC RATING**.
With 2 separate secondaries, you can parallel the outputs to get double the current, but don't forget 12v + 12v @ 3amp means 12v in parallel with 12v will provide 2amp DC and the DC voltage will be about 17v.

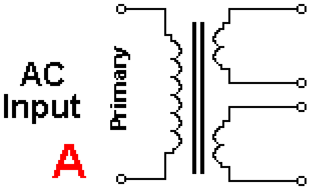

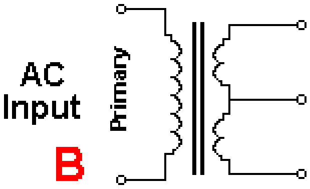

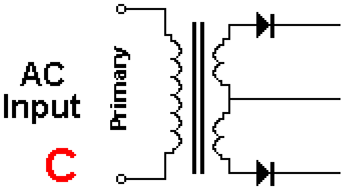

CURRENT TRANSFORMER

A Current Transformer is really an ordinary transformer.
All transformers produce a CURRENT output and a VOLTAGE output.
If you put an ammeter across the secondary, the current will increase through the meter when the primary voltage is increased.
This is because the output voltage will increase and this voltage will allow a higher current to flow.

WHY DETECT CURRENT? Why not voltage?

Because the voltage of say the "240v AC" is always 240v but the current can increase from say 1 amp to nearly 15 amps, depending what appliance is connected. So it is pointless measuring voltage.
A Current Transformer is a step-up transformer. When we say step-up and step-down, we are referring to the voltage - comparing the primary voltage to the secondary voltage. (Most transformers on the "mains" are step-down transformers and are used as power supplies to laptops, phone chargers etc.) Even a welding transformer is a step-down device and produces about 20v to 70v, while the current can be as high as 100 amps. This current is higher than the mains will deliver and is needed to melt the metal at the point of contact of the probe and the item being welded.
A Current Transformer is a step-up transformer. The primary consists of a single turn (or maybe 2 - 5 turns) and the secondary has 100 turns (or more).
This means the voltage seen by the primary will be increased 100 times and appear as anything from a few hundred millivolts to a few volts, depending on the quality of the coupling. (the magnetic coupling between the wire through the centre of the core, the quality of the core to transfer this magnetic flux to the secondary turns.)
This voltage is then passed to a low value resistor, where the voltage is reduced to a level that suits the detection circuit and the resulting millivolts is interpreted as current in the wire being tested.

Recapping:

The reading on the secondary has no relation to the current in the primary. We need to add a LOAD RESISTOR and create a table before we can use the transformer.
There is no such thing as a CURRENT TRANSFORMER. It is really an INSTRUMENT TRANSFORMER and the scale has been marked in units of CURRENT after measurements have been made. (INSTRUMENT TRANSFORMER means it is a device that helps us to produce a connection between current flowing through a wire and a reading on a meter or display).
If we connect a load to the secondary, (say an ammeter), it will produce a reading that increases when the current through the single primary turn is increased. That's because the ammeter is a LOAD. But the reading is meaningless until be calibrate the scale.

Now, lets look at the primary.

A wire (or cable) through the centre of the core is counted as one turn. If the turn is wrapped around the core, the coupling will be improved, but if we always use a straight wire, it does not matter where it is positioned inside the centre of the core.

It does not matter if the magnetic interaction of the flux from the wire is good or bad, we just have to keep to the same way of using the transformer.

The calibration can be done with any poor coupling and the result will be accurate for all future readings.

If a low-value resistor is placed across the secondary, the voltage across this resistor will increase and also the current through it will increase. But we are not going to measure the current through the resistor. We are going to measure the voltage across the resistor and by taking lots of reading we will finish up with a scale or table and this is called CALIBRATION. The results will be equated to the current flowing through the primary wire (primary turn).

A clamp meter uses a current transformer and the jaws must be closed completely and cleanly for the flux to flow around the core and produce a reading in the secondary.

Dirt in the jaws will reduce the reading considerably.

You cannot measure the current in a "power cord" because it contains both the active and neutral wires.

Even though the current is a maximum in both conductors at the same time, the current is flowing in two different directions and the magnetic flux produced by one conductor is clockwise and the other is anticlockwise and they are cancelled by each other.

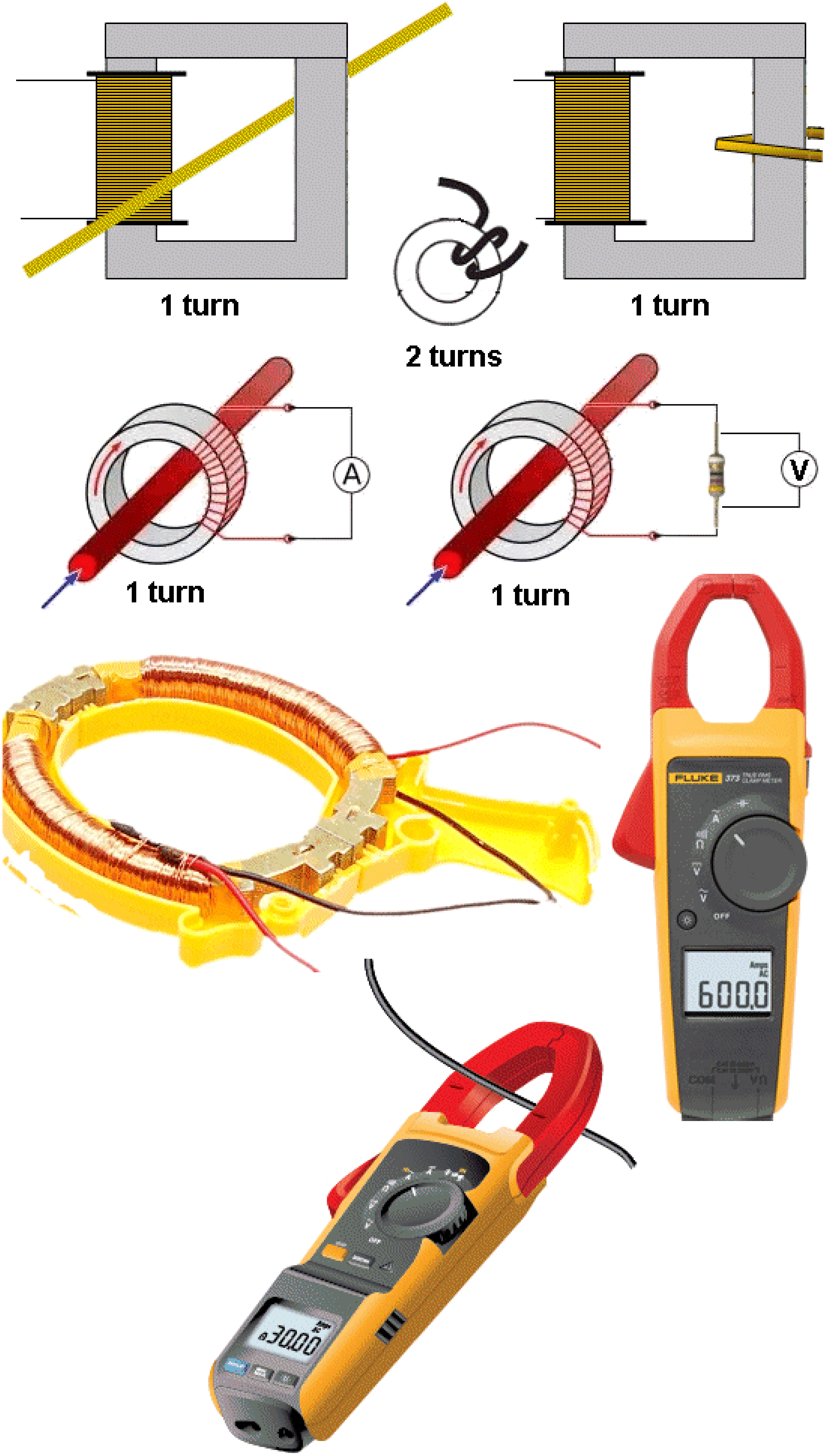
1 turn
2 turns
1 turn
A
V
1 turn
1 turn
FLUKE 373
600.0
30.00

Identifying ONE TURN.
The quality (the coupling) of a single STRAIGHT wire

through the centre of a core is very poor
but if all readings are taken with this amount
of coupling, the readings will be accurate, as the
calibrations have been done with this arrangement.

OPTO ISOLATORS and OPTO COUPLERS

Opto Isolators and Opto Couplers are the same thing. A common opto-coupler is 4N35. It is used to allow two circuits to exchange signals yet remain electrically isolated. The signal is applied to the LED, which shines on a silicon NPN photo-transistor in the IC.
The light is proportional to the signal, so the signal is transferred to the photo transistor to turn it on a proportional amount. Opto-couplers can have Light Activated SCR's, photodiodes, TRIAC's and other semiconductor devices as an output. The 4N35 opto-coupler schematic is shown below:

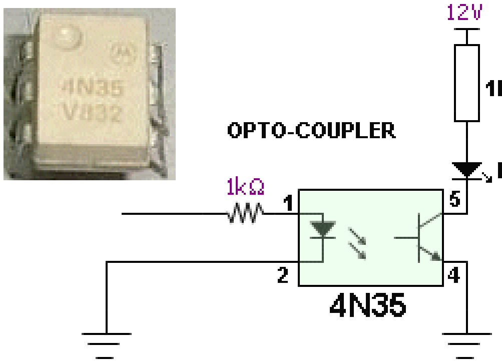

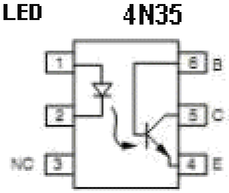

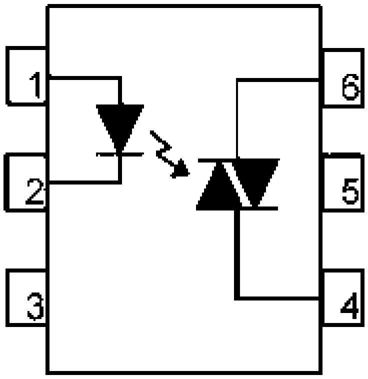

An opto-Coupler using a TRIAC
Note: the pinout is different to 4N35

TESTING AN OPTO COUPLER

Most multimeters cannot test the LED on the input of an opto-coupler because the ohms range does not have a voltage high enough to activate the LED with at least 2mA.
You need to set-up the test-circuit shown above with a 1k resistor on the input and 1k5 on the output. When the 1k is connected to 12v, the output LED will illuminate.
The opto-coupler should be removed from circuit to perform this test.

TRANSISTORS

Transistors are solid-state devices and although they operate completely differently to a diode, they appear as two back-to-back diodes when tested.
There are basically 2 types of transistor NPN and PNP.
A transistor is sometimes referred to as BJT (Bi-polar Junction Transistor) to distinguish it from other types of transistor such as Field Effect transistor, Programmable Unijunction Transistor and others.

In the following diagram, two diodes are connected together and although the construction of a transistor is more complex, we see the transistor as two diodes when testing it.

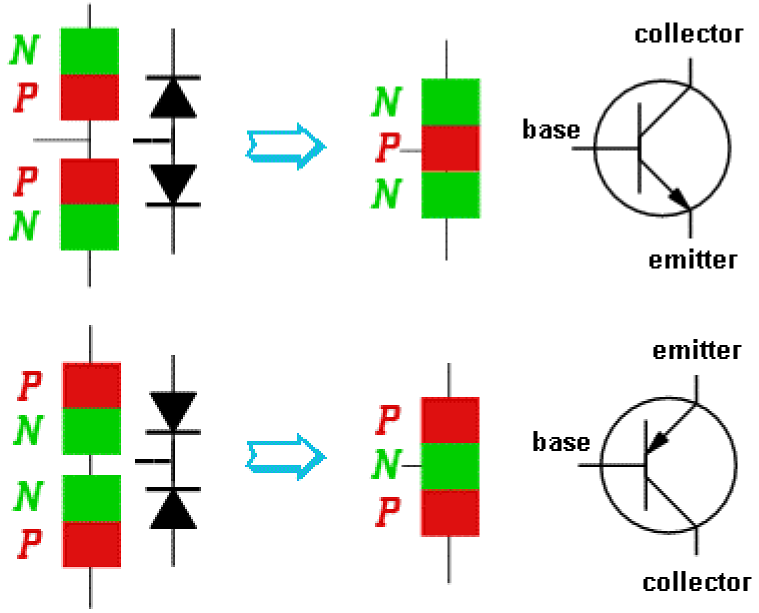

A TRANSISTOR APPEARS AS TWO DIODES WHEN TESTING IT

All transistors have three leads. Base (b), Collector (c), and Emitter (e).
For an NPN transistor, the arrow on the emitter points away from the base.
It is fortunate that the arrow on both symbols points in the direction of the flow of current (Conventional Current) and this makes it easy to describe testing methods using our simplified set of instructions. The symbols have been drawn exactly as they appear on a circuit diagram.

All transistors **are the same** but we talk about digital and analogue transistors.
There is no difference between the two.
The difference is the circuit. And the only other slight difference between transistors is the fact that some have inbuilt diodes and resistors to simplify the rest of the circuit.
All transistors work the same way. The only difference is the amount of amplification they provide, the current and voltage they can withstand and the speed at which they work. For simple testing purposes, they are all the same.

NPN transistors are the most common and for an NPN transistor, the following applies.
(the opposite applies for PNP)
To test a transistor, there is **one thing** you have to know:
When the base voltage is higher than the emitter, current flows though the collector-emitter leads.
As the voltage is increased on the base, nothing happens until the voltage reaches 0.55v. At this point a very small current flows through the collector-emitter leads. As the voltage is increased, the current-flow increases. At about 0.75v, the current-flow is a MAXIMUM. (can be as high as 0.9v). That's how it works. A transistor also needs **current** to flow into the base to perform this amplifying function and this is the one feature that separates an ordinary transistor from a FET.

If the voltage on the base is 0v, then instantly goes to 0.75v, the transistor initially passes NO current, then FULL current. The transistor is said to be working in its two states: OFF then ON (sometimes called: "cut-off" and "saturation"). These are called digital states and the transistor is said to be a **DIGITAL TRANSISTOR** or a **SWITCHING TRANSISTOR** , working in **DIGITAL MODE.**

If the base is delivered 0.5v, then slowly rises to 0.75v and slowly to 0.65v, then

0.7v, then 0.56v etc, the transistor is said to be working in ANALOGUE MODE and the transistor is an **ANALOGUE TRANSISTOR**.
Since a transistor is capable of amplifying a signal, it is said to be an active device. Components such as resistors, capacitors, inductors and diodes are not able to amplify and are therefore known as passive components.

In the following tests, use your finger to provide the **TURN ON** voltage for the base (this is 0.55v to 0.7v) and as you press harder, more current flows into the base and thus more current flows through the collector-emitter terminals. As more current flows, the needle of the multimeter moves UP-SCALE.

TESTING A TRANSISTOR ON A DIGITAL METER

Testing a transistor with a **Digital Meter** must be done on the "DIODE" setting as a digital meter does not deliver a current through the probes on some of the resistance settings and will not produce an accurate reading.
The "DIODE" setting must be used for diodes and transistors. It should also be called a "TRANSISTOR" setting.

TESTING AN unknown TRANSISTOR

The first thing you may want to do is test an unknown transistor for COLLECTOR, BASE AND EMITTER. You also want to perform a test to find out if it is NPN or PNP. That's what this test will provide.
You need a cheap multimeter called an ANALOGUE METER - a multimeter with a scale and pointer (needle).
It will measure resistance values (normally used to test resistors) - (you can also test other components) and Voltage and Current. We use the resistance settings. It may have ranges such as "x10" "x100" "x1k" "x10"
Look at the resistance scale on the meter. It will be the top scale.
The scale starts at zero on the right and the high values are on the left. This is opposite to all the other scales.
When the two probes are touched together, the needle swings FULL SCALE and reads "ZERO." Adjust the pot on the side of the meter to make the pointer read exactly zero.

How to read: "x10" "x100" "x1k" "x10"
Up-scale from the zero mark is "1"
When the needle swings to this position on the "x10" setting, the value is 10 ohms.
When the needle swings to "1" on the "x100" setting, the value is 100 ohms.
When the needle swings to "1" on the "x1k" setting, the value is 1,000 ohms = 1k.
When the needle swings to "1" on the "x10k" setting, the value is 10,000 ohms = 10k.
Use this to work out all the other values on the scale.
Resistance values get very close-together (and very inaccurate) at the high end of the scale. [This is just a point to note and does not affect testing a transistor.]

Step 1 - FINDING THE BASE and determining NPN or PNP
Get an unknown transistor and test it with a multimeter set to "x10"
Try the 6 combinations and when you have the black probe on a pin and the red probe touches the other pins and the meter swings nearly full scale, you have an NPN transistor. The black probe is BASE
If the red probe touches a pin and the black probe produces a swing on the other two pins, you have a PNP transistor. The red probe is BASE
If the needle swings FULL SCALE or if it swings for more than 2 readings, the transistor is **FAULTY**.

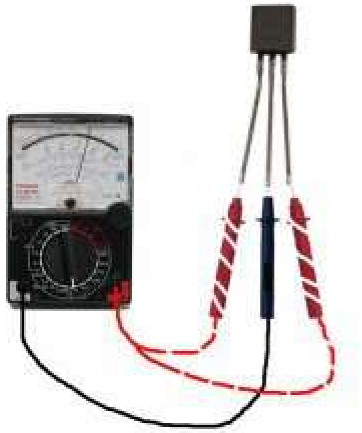

This is an NPN transistor
The black probe is the BASE

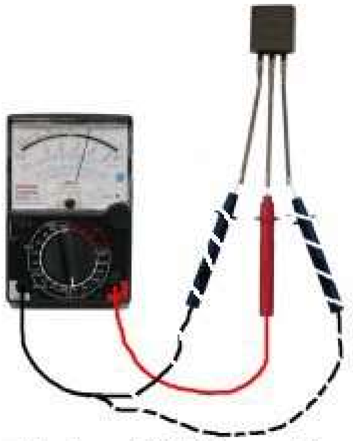

This is a PNP transistor
The red probe is the BASE

Step 2 - FINDING THE COLLECTOR and EMITTER
Set the meter to "x10k."

For an NPN transistor, place the leads on the transistor and when you press hard on the two leads shown in the diagram below, the needle will swing almost full scale.

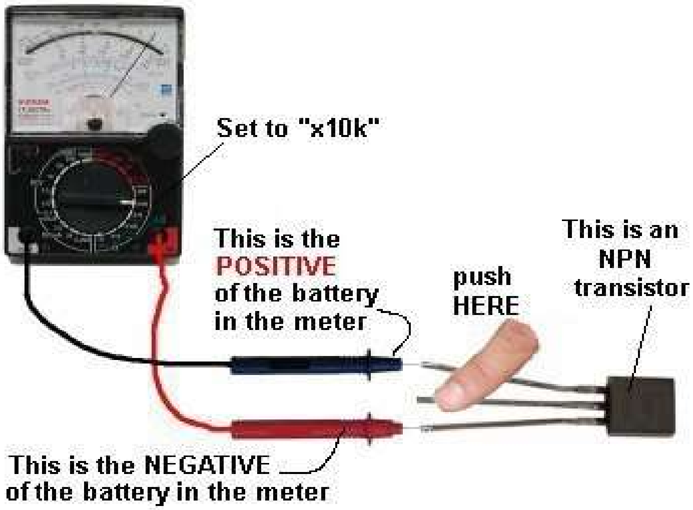
Set to "x10k"
This is the POSITIVE of the battery in the meter
push HERE
This is an NPN transistor
This is the NEGATIVE of the battery in the meter

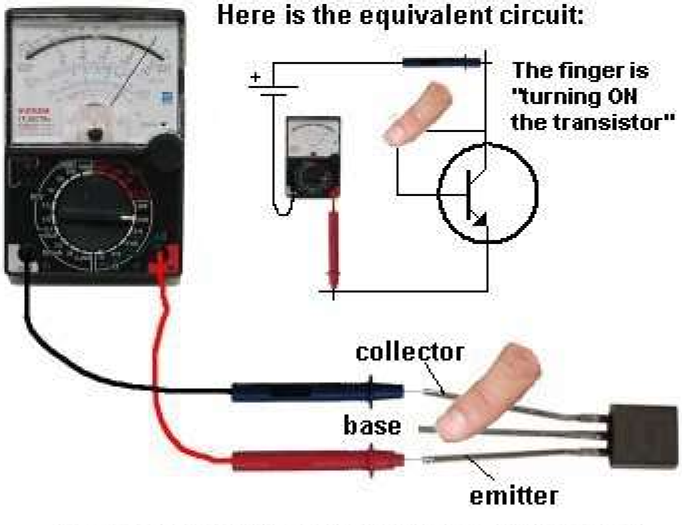
Here is the equivalent circuit:
+
The finger is "turning ON the transistor"
collector
base
emitter

You have found the Collector Base and Emitter

For a PNP transistor, set the meter to "x10k" place the leads on the transistor and when you press hard on the two leads shown in the diagram below, the needle will swing almost full scale.

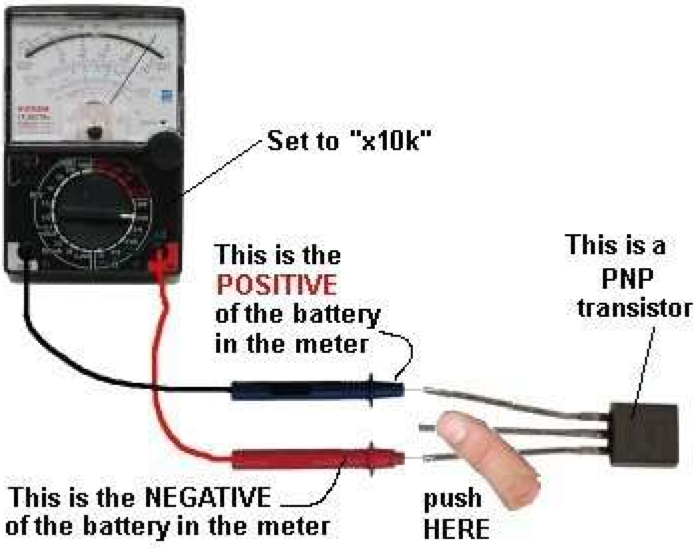
Set to "x10k"
This is the POSITIVE of the battery in the meter
This is a PNP transistor
This is the NEGATIVE of the battery in the meter
push HERE

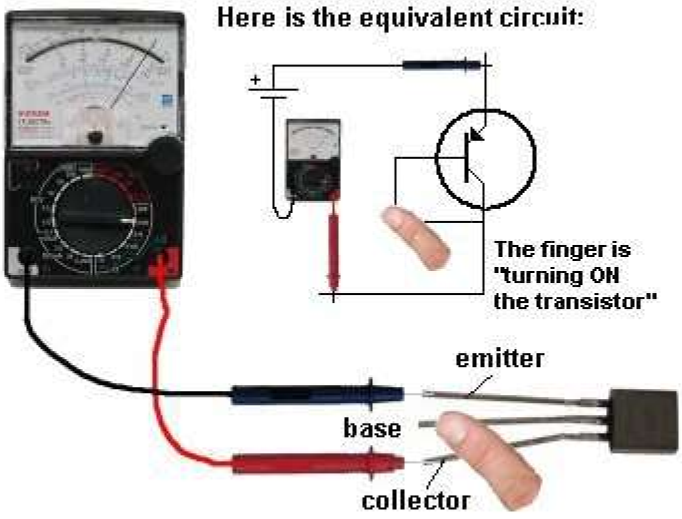
Here is the equivalent circuit:
+
The finger is "turning ON the transistor"
emitter
base
collector

You have found the Collector Base and Emitter

SIMPLEST TRANSISTOR TESTER

The simplest transistor tester uses a 9v battery, 1k resistor and a LED (any colour). Keep trying a transistor in all different combinations until you get one of the circuits below. When you push on the two leads, the LED will get brighter.
The transistor will be NPN or PNP and the leads will be identified:

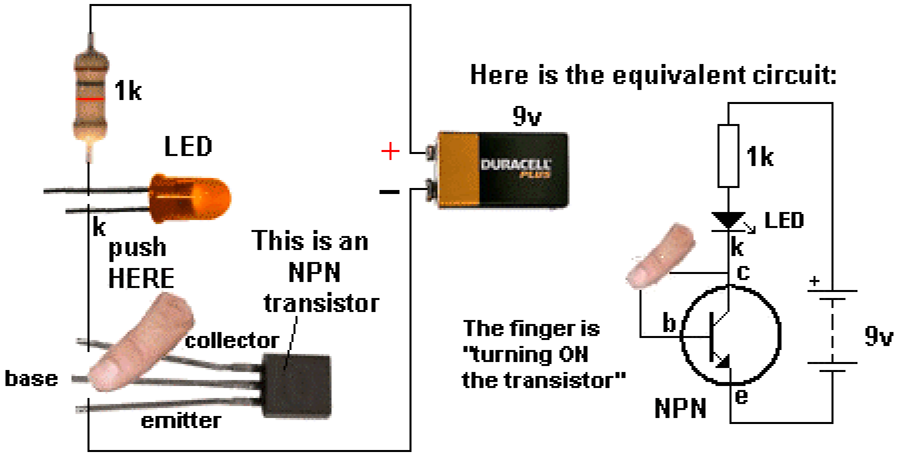
1k
LED
k
push
HERE
This is an
NPN
transistor
collector
base
emitter
+
−
Here is the equivalent circuit:
9v
1k
LED
k
c
b
e
NPN
9v
The finger is
"turning ON
the transistor"

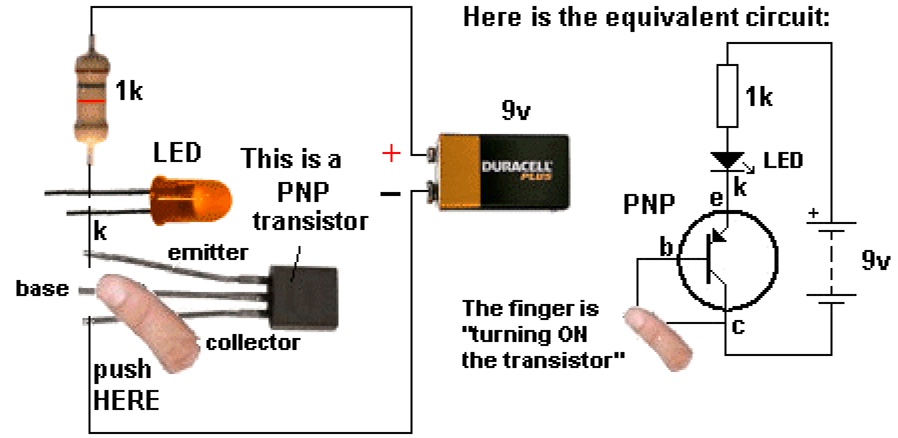
Here is the equivalent circuit:
1k
LED
This is a
PNP
transistor
k
emitter
base
collector
push
HERE
+
−
9v
1k
LED
PNP
e
k
b
c
9v
The finger is
"turning ON
the transistor"

The leads of some transistors will need to be bent so the pins are in the same positions as shown in the diagrams. This helps you see how the transistor is being turned on. This works with NPN, PNP transistors and Darlington transistors.

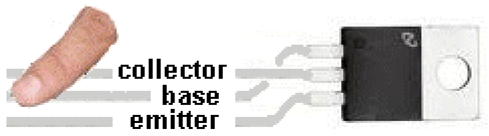
collector
base
emitter

NPN transistor

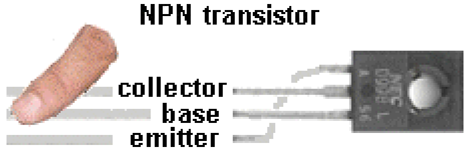
collector
base
emitter

NPN transistor

HEATSINKING

Heat generated by current flowing between the collector and emitter leads of a transistor causes its temperature to rise. This heat must be conducted away from the transistor otherwise the rise may be high enough to damage the P-N junctions inside the device. Power transistors produce a lot of heat, and are therefore usually mounted on a piece of aluminium with fins, called a **HEATSINK**.
This draws heat away, allowing it to handle more current. Low-power signal transistors do not normally require heat sinking. Some transistors have a metal body or fin to connect to a larger heatsink. If the transistor is connected to a heatsink with a mica sheet (mica washer), it can be damaged or cracked and create a short-circuit. (See Testing Mica Washers). Or a small piece of metal may be puncturing the mica. Sometimes white compound called **Heatsink Compound** is used to conduct heat through the mica. This is very important as mica is a very poor conductor of heat and the compound is needed to provide maximum thermal conduction.

TRANSISTOR FAILURE

Transistor can fail in a number of ways. They have forward and reverse voltage ratings and once these are exceeded, the transistor will ZENER or conduct and may fail. In some cases a high voltage will "puncture" the transistor and it will fail instantly. In fact it will fail much faster via a voltage-spike than a current overload.

It may fail with a "short" between any leads, with a collector-emitter short being the most common. However failures will also create shorts between all three leads.
A shorted transistor will allow a large current to flow, and cause other components to heat up.
Transistors can also develop an open circuit between base and collector, base and emitter or collector and emitter.
The first step in identifying a faulty transistor is to check for signs of overheating. It may appear to be burnt, melted or exploded. When the equipment is switched off, you can touch the transistor to see if it feels unusually hot. The amount of heat you feel should be proportional to the size of the transistor's heat sink. If the transistor has no heat sink, yet is very hot, you can suspect a problem.
DO NOT TOUCH A TRANSISTOR IF IT IS PART OF A CIRCUIT THAT CARRIES 240VAC. Always switch off the equipment before touching any components.

TRANSISTOR REPLACEMENT

If you can't get an exact replacement, refer to a transistor substitution guide to identify a near equivalent.

The important parameters are:
- Voltage
- Current
- Wattage
- Maximum frequency of operation

The replacement part should have parameters equal to or higher than the original.

Points to remember:
- Polarity of the transistor i.e. PNP or NPN.
- At least the same voltage, current and wattage rating.
- Low frequency or high frequency type.
- Check the pinout of the replacement part
- Use a desoldering pump to remove the transistor to prevent damage to the printed circuit board.
- Fit the heat sink.
- Check the mica washer and use heat-sink compound
- Tighten the nut/bolt - not too tight or too loose.
- Horizontal output transistors with an integrated diode should be replaced with the same type.

DIGITAL TRANSISTORS

There is no such thing as a DIGITAL TRANSISTOR or an AUDIO TRANSISTOR.
All transistors are just "TRANSISTORS" and the surrounding components as well as the type of signal, make the transistor operate in DIGITAL MODE or ANALOGUE MODE.
But we have some transistors that have inbuilt resistors to make them suitable for connecting to a digital circuit without the need for a base resistor.
Here is the datasheet for an NPN transistor BCR135w and PNP datasheet for BCR185w.
These transistors are called "Digital Transistors" because the "base lead" can be connected directly to the output of a digital stage. This "lead" or "pin" is not really the base of the transistor but a 4k7 (or 10k) resistor connected to the base allows the transistor to be connected to the rest of a digital circuit.
You cannot actually get to the base. The resistor(s) are built into the chip and the transistor is converted into a "Digital Transistor" because it will accept 5v on the "b"

lead.
The 47k is not really needed but it makes sure the transistor is fully turned OFF if the signal on the "b" lead is removed (in other words - if the input signal is converted to a high-impedance signal - see tri-state output from microcontrollers for a full explanation).
This transistor is designed to be placed in a circuit where the input changes from low to high and high to low and does not stop mid-way. This is called a DIGITAL SIGNAL and that is one reason why the transistor is called a digital transistor. (However you could stop half-way but the transistor may heat up and get too hot).
Any transistor placed in a digital circuit can be called a "digital transistor" but it is better to say it is operating in DIGITAL MODE.

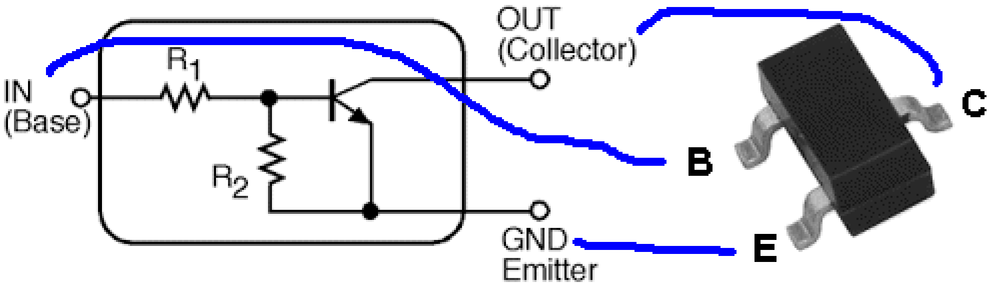
OUT
(Collector)
IN
(Base)
R_1
R_2
GND
Emitter
B
C
E

The digital transistor has two resistors included inside the case
R1 is about 10k and R2 is approx 47k

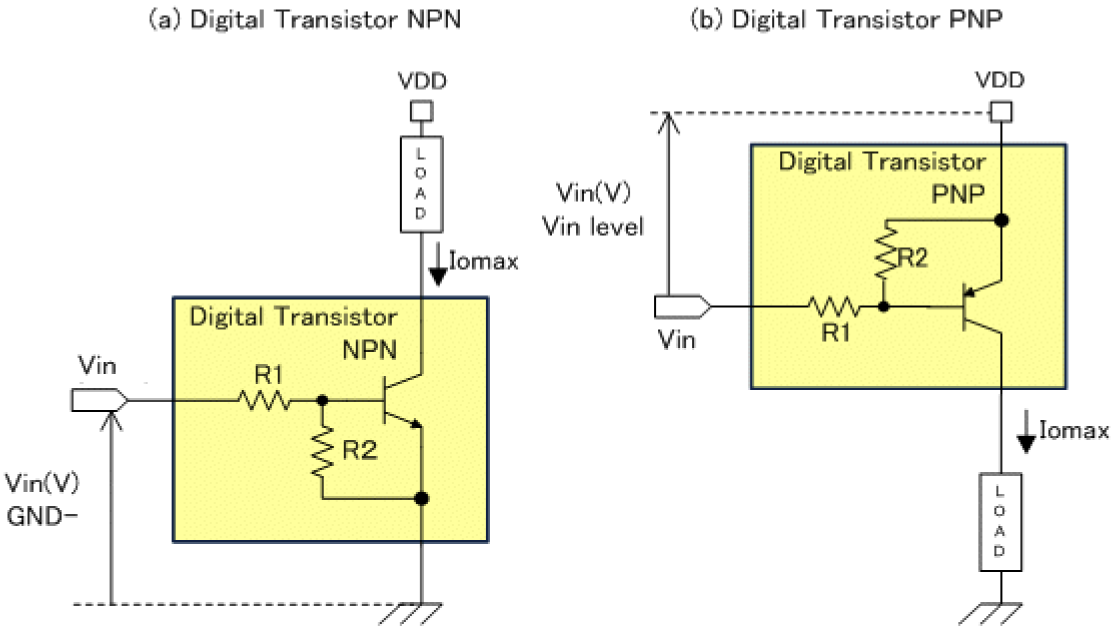
(a) Digital Transistor NPN

(b) Digital Transistor PNP
VDD
LOAD
Iomax
Digital Transistor
NPN
Vin
R1
R2
Vin(V)
GND−
VDD
Vin(V)
Vin level
Digital Transistor
PNP
R2
R1
Vin
Iomax
LOAD

Equivalent circuit of the digital transistor

These transistors can be made to work in analogue circuits because they are ordinary transistors with a 10k base resistor, but you will have to know what you are doing.

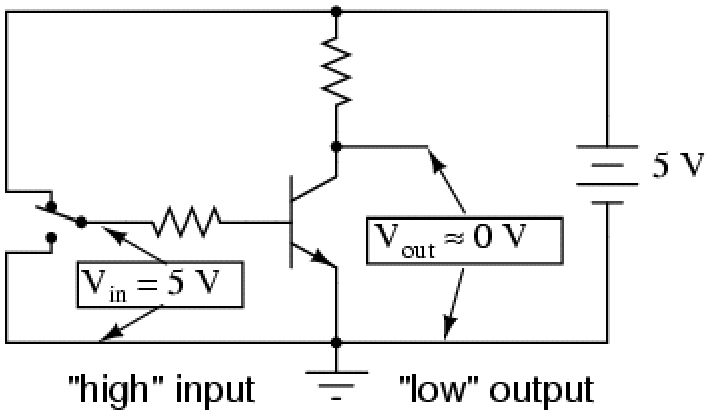

Transistor in saturation

The circuit above shows the digital transistor is designed to allow a voltage of 5v to be supplied to the "base" pin and the transistor will Fully Conduct.
This type of transistor saves putting a base resistor on the PC board.
It can be tested just like a normal transistor but the resistance between base and emitter will be about 5k to 50k in both directions. If the collector-emitter is low in both directions the transistor is damaged.
Here's how to look at how the transistor works:
The 10k resistor on the base will allow 0.5mA to flow into the base. But the 47k will reduce this to 0.4mA. If the transistor has a gain of 100, the collector-emitter current can be 40mA.
To determine the current capability of the transistor, connect 100R load and turn the transistor ON. This will allow about 100mA for the collector-emitter current. Measure the collector-emitter voltage. If it is more than 0.5v, the transistor is OVER-LOADED.

DARLINGTON TRANSISTORS

A DARLINGTON TRANSISTOR is two transistors in a single package with three leads. They are internally connected in cascade so the gain of the pair is very high. This allows a very small input signal to produce a large signal at the output. They have three leads (Base, Collector and Emitter and can be PNP or NPN) and are equivalent to the leads of a standard individual transistor, but with a very high gain. The second advantage of a Darlington Transistor is its high input impedance. It puts very little load on the previous circuit.
Some Darlington transistors have a built-in diode and/or built-in resistor and this will produce a low reading in both directions between the base and emitter leads.

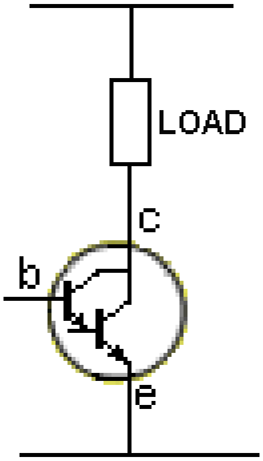

NPN Darlington

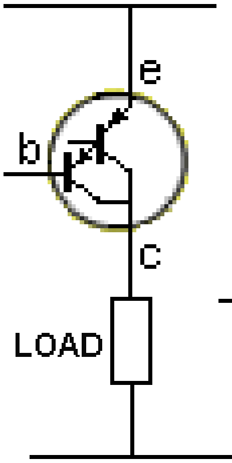

PNP Darlington

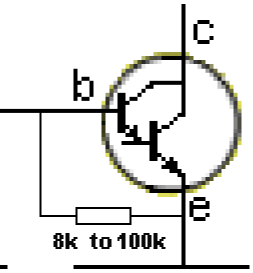

Darlington transistors are tested the same as an ordinary transistor and a multimeter will produce about the same deflection, even though you will be measuring across two junctions, (and a base-emitter resistor is present).

HORIZONTAL OUTPUT TRANSISTORS, SWITCH-MODE TRANSISTORS, FLYBACK TRANSISTORS, POWER TRANSISTORS, VERTICAL TRANSISTORS . . . .

These are all names given to a transistor when it is used in a particular circuit. ALL these transistors are the same for testing purposes.

We are not testing for gain, maximum voltage, speed of operation or any special feature. We are just testing to see if the transistor is completely faulty and SHORTED.

A transistor can have lots of other faults and the circuit **using the transistor** is the best piece of TEST EQUIPMENT as it is detecting the fault.

TESTING MOSFETs and FETs

MOSFETs and **JFETs** are all part of the **FET** family.

MOSFET stands for **Metal Oxide Semiconductor Field Effect Transistor.**

FETs operate exactly the same as a "normal" transistor except they have different names for the input and output leads and the voltage between the gate and the source has to between 2v to 5v for the device to turn on fully. A FET requires almost NO CURRENT into the Gate for it to turn on and when it does, the voltage between drain and source is very low (only a few mV). This allows them to pass very high currents without getting hot. There is a point where they start to turn on and the input voltage must rise higher than this so the FET turns on FULLY and does not get hot.

Field Effect Transistors are difficult to test with a multimeter, but "fortunately" when a power **MOSFET** blows, it is completely damaged. All the leads will show a short circuit. 99% of bad **MOSFETs** will have GS, GD and DS shorted.

The following symbols show some of the different types of MOSFETs:

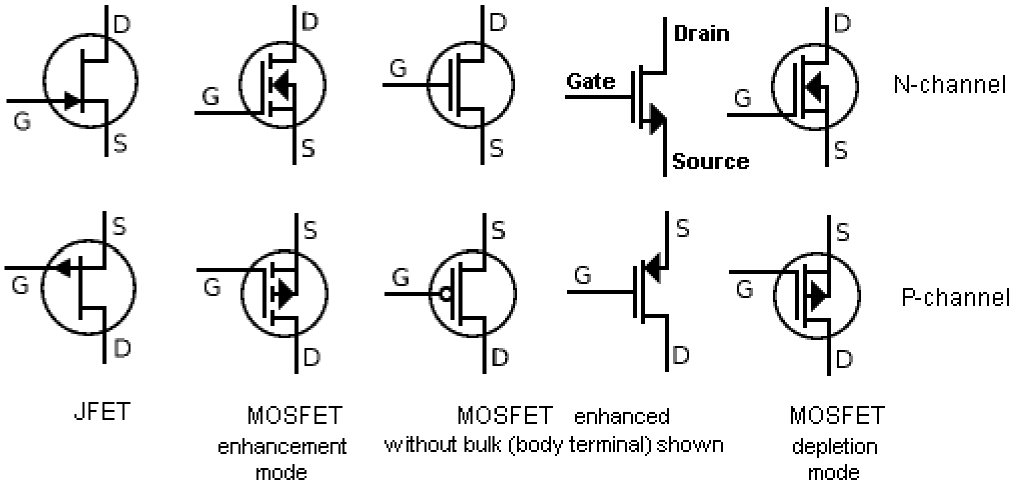

Most **MOSFET** transistors cannot be tested with a multimeter. This due to the fact that the Gate needs 2v - 5v to turn on the device and this voltage is not present on the probes of either meter set to any of the ohms ranges.

You need to build the following Test Circuit:

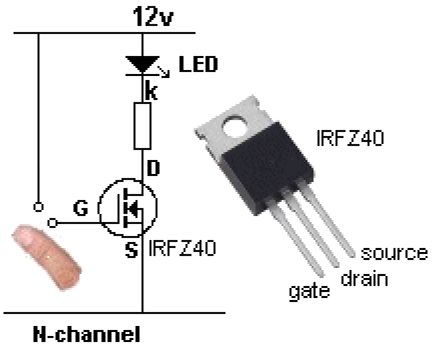

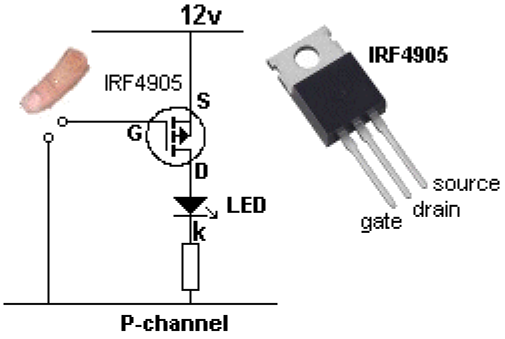

Touching the Gate will increase the voltage on the Gate and the MOSFET will turn ON and illuminate the LED. Removing your finger will turn the LED off.
Large devices such as the TO-220 types shown above do not like static electricity on the gate and you have to be careful not to "spike" the gate with any static. Generally this type of device is not "super sensitive" and you can use your finger or a large value resistor.
When replacing one of these devices, there are 2 things to match-up.
Voltage and Current.
In most cases, the "turn-ON" resistance (the resistance between Source and Drain) will be the same (something like 22 milli ohms) and the speed of operation will be ok.
Check the voltage needed to turn the gate ON and make sure you can supply the required voltage.

SILICON CONTROLLED RECTIFIERs (SCR)

The **Silicon Controlled Rectifier** (SCR) is a semiconductor device that is a member of a family of control devices known as **Thyristors**. It is a 3-leaded device and when a small current enters the Gate, the **thyristor** turns on. AND STAYS ON. It only conducts current between Anode and Cathode in one direction and it is mainly only used in DC circuits. When it is used with AC, it will only conduct for a maximum of half the cycle.

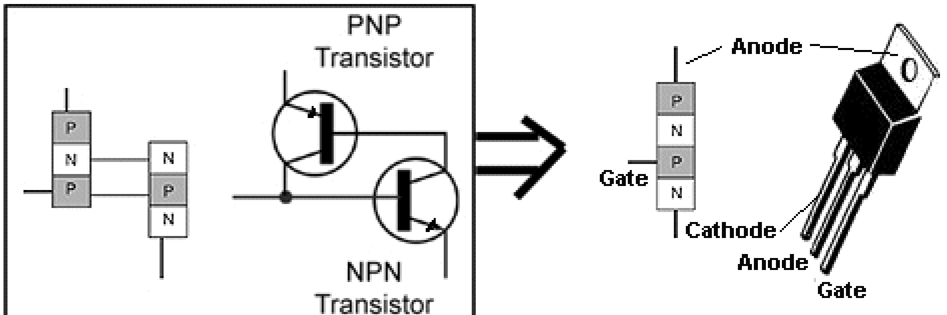
PNP
Transistor
P
N
P
N
P
N
NPN
Transistor
Anode
P
N
P
N
Gate
Cathode
Anode
Gate

THYRISTOR or SCR

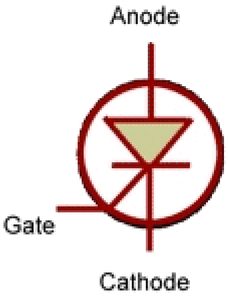
Anode
Gate
Cathode

SCR
Silicon Controlled Rectifier

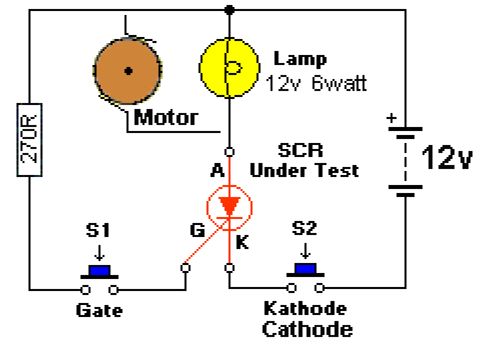
Lamp
12v 6watt
Motor
270R
+
12v
SCR
Under Test
A
G
K
S1
S2
Gate
Kathode
Cathode

SCR TESTER

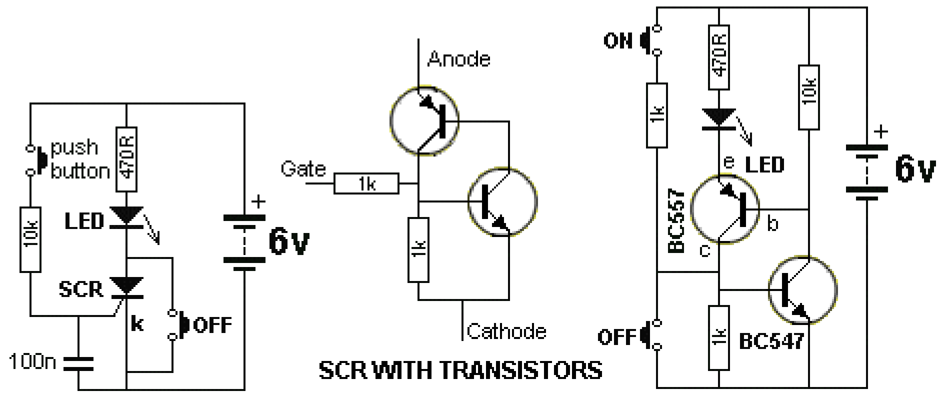
push
button
470R
10k
LED
SCR
k
OFF
100n
+
6v
Anode
Gate
1k
1k
Cathode
ON
470R
10k
1k
LED
e
BC557
b
c
+
6v
OFF
1k
BC547

A

SCR WITH TRANSISTORS

B

To understand how an SCR "latches" when the gate is provided with a small current, we can replace it with two transistors as shown in diagram B above. When the ON button is pressed, the BC547 transistor turns on. This turns ON the BC557 and it takes over from the action of the switch.
To turn the circuit off, the OFF button removes the voltage from the base of the BC547.

Testing an SCR

An **SCR** can be tested with some multimeters but a minimum current Anode-to-Cathode is needed to keep the device turned on. Some multimeters do not provide this amount of current and the **SCR Tester** circuit above is the best way to test these devices.
Shorted SCRs can usually be detected with an ohmmeter check (SCRs usually fail shorted rather than open).
Measure the anode-to-cathode resistance in both the forward and reverse direction; a good SCR should measure near infinity in both directions.
Small and medium-size SCRs can also be gated ON with an ohmmeter (on a digital meter use the Diode Check Function). Forward bias the SCR with the ohmmeter by connecting the black (-) lead to the anode and the red (+) lead to the cathode

(because the + of the battery is connected to the negative lead, in most analogue multimeters). Momentarily touch the gate lead to the anode while the probes are still touching both leads; this will provide a small positive turn-on voltage to the gate and the cathode-to-anode resistance reading will drop to a low value. Even after removing the gate voltage, the SCR will stay conducting. Disconnecting the meter leads from the anode or cathode will cause the SCR to revert to its non-conducting state.

When making the above test, the meter impedance acts as the SCR load. On larger SCRs, it may not latch ON because the test current is not above the SCR holding current.

Using the SCR Tester

Connect an SCR and press Switch2. The lamp should not illuminate. If it illuminates, the SCR is around the wrong way or it is faulty.

Keep Switch 2 PRESSED. Press Sw1 very briefly. The lamp or motor will turn ON and remain ON. Release Sw 2 and press it again. The Lamp or motor will be OFF.

TRIACs

A triac is a bidirectional, three-terminal dual, back-to-back thyristor (SCR) switch. This device will conduct current in both directions when a small current is constantly applied to the Gate.

If the gate is given a small, brief, current during any instant of a cycle, it will remain triggered during the completion of the cycle until the current though the Main Terminals drops to zero.

This means it will conduct both the positive and negative half-cycles of an AC waveform. If it is tuned on (with a brief pulse) half-way up the positive waveform, it will remain on until the wave rises and finally reaches zero. If it is then turned on (with a brief pulse) part-way on the negative wave, the result will be pulses of energy and the end result will be about 50% of the full-energy delivered at a rate of 100 times per second for a 50HZ supply.

TRIACs are particularly suited for AC power control applications such as motor speed control, light dimmers, temperature control and many others.

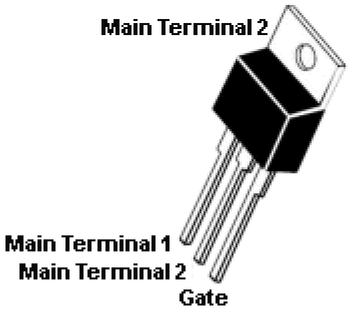

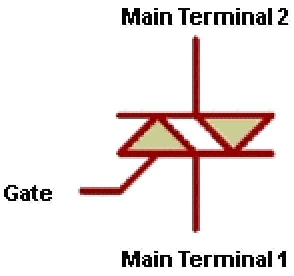

TRIAC (2 SCR's)

TRIode for Alternating Current

Using the TRIAC Tester

Connect a TRIAC and press Switch2. The lamp should not illuminate. If it illuminates, the TRIAC is faulty.

Keep Switch 2 PRESSED. Press Sw1 very briefly. The lamp or motor will turn ON and remain ON. If the lamp does not turn on, reverse the TRIAC as the current into the gate must produce a slight voltage between **Gate** and **Main Terminal 1**.

Release Sw 2 and press it again. The Lamp or motor will be OFF.

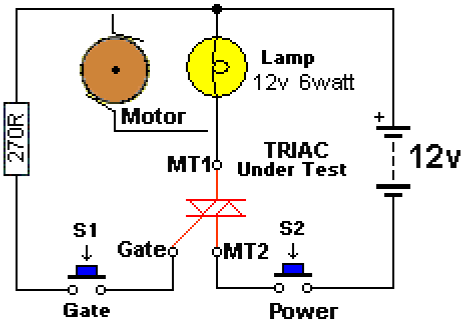

TRIAC TESTER

MICA WASHERS AND INSULATORS

Plastic insulating sheets (washers) between a transistor and heatsink are most often made from mica but some are plastic and these get damaged over a period of time, turn dark and become cracked.

The plastic eventually becomes carbonized and conducts current and can affect the operation of the appliance. You can see the difference between a mica sheet (washer) and plastic by looking where it extends from under the transistor. Replace all plastic insulators as they eventually fail.

SPARK GAPS

Some TV's and monitors with a CRT (picture tube), have spark gaps either on the socket at the end of the tube or on the chassis.

These can consist of two wires inside a plastic holder or a glass tube or special resistive device.

The purpose of a spark gap is to take any flash-over (from inside the tube), to earth. This prevents damage to the rest of the circuit.

However if the tube constantly flashes over, a carbon track builds up between the wires and effectively reduces the screen voltage. This can cause brightness and/or focus problems. Removing the spark-gap will restore the voltage.

These are not available as a spare component and it's best to get one from a discarded chassis.

CO-AX CABLES

Co-Ax cables can produce very high losses and it seems impossible that a few metres of cable will reduce the signal. The author has had a 3 metre cable reduce the signal to "snow" so be aware that this can occur. Faults can also come from a splitter and/or balun as well as dirty plugs and sockets. This can result in very loud bangs in the sound on digital reception.

TESTING EARTH LEAKAGE DETECTORS or Residual Current Devices or Ground Fault Circuit Interrupters or GFCI

An Earth Leakage Detector or Sensor is a circuit designed to continuously monitor the imbalance in the current in a pair of load carrying conductors.

These two conductors are normally the Active and Neutral. Should the imbalance current reach 30mA the sensor will "trip" and remove the voltage (and current) from

the line being monitored.
Some detectors will trip at 15mA.
You cannot alter the sensitivity of the device however there are a number of faults in these devices that can be fixed.
In some devices the contact pressure for the 10Amp or 15 Amp contacts is very weak and they arc and produce an open circuit. The result is this: When you press the rest button, power is not restored to the output.
Clean the contacts with a small file and bend the metal strips to the contacts so they make a very strong contact.
The other fault is the trip mechanism.
The magnetism from the coil does not allow the pin to move and "trip" the contacts. It may be due to a small metal filing or the pin not moving freely enough.
All good Earth Leakage Detectors have a TEST BUTTON. This connects a resistor between the active line and earth so that 15mA or 30mA flows.
The detector should trip immediately. Make sure the trigger mechanism trips when the test button is pressed.
None of the electronics in the detector can be replaced however you can test the mechanical operation and the pressure on the contacts when the unit is removed from the power. Do not work on the device when it is connected to the mains.

TESTING CELLS AND BATTERIES

There is an enormous number of batteries and cells on the market and a number of "battery testers." Instead of buying a battery tester that may give you a false reading, here is a method of testing cells that is guaranteed to work.
There are two types of cell: a **rechargeable** cell and a non rechargeable cell.

The easiest way to test a **rechargeable** cell is to put a group of them in an appliance and use them until the appliance "runs down" or fails to work. If you consider the cells did not last very long, remove them and check the voltage of each cell. The cell or cells with the lowest voltage will be faulty. You can replace them with new cells or good cells you have in reserve.
There is no other simple way to test a rechargeable cell.
You cannot test the "current of a cell" by using an ammeter. A rechargeable cell can deliver 10 amps or more, even when nearly discharged and you cannot determine a good cell for a faulty cell.

Dry cells are classified as "non-rechargeable" cells.
DRY CELLS and MANGANESE CELLS are the same thing. These produce 1.5v per cell (manganese means the Manganese Dioxide depolariser inside the cell. All "dry cells" use manganese dioxide).
ALKALINE CELLS produce between 2 - 10 times more energy than a "dry cell" and produce 1.5v per cell.
Alkaline cells can fail for no reason at any stage in their life and are not recommended for emergency situations.
The output voltage of some Alkaline cells can fall to 0.7v or 0.9v for not apparent reason.
There are lots of other cells including "button cells," hearing-aid cells, air cells, and they produce from 1.2v to 3v per cell.

Note:
Lithium cells are also called "button cells" and they produce 3v per cell.
Lithium cells are non-rechargeable (they are generally called "button cells") but some Lithium cells can be recharged. These are Lithium-ion cells and generally have a voltage of 3.6v. Some Lithium-ion cells look exactly like 3v Lithium cells, so you have to read the data on the cell before charging.

You cannot test the voltage of a cell and come to any conclusion as to the age of the cell or how much energy remains. The voltage of a cell is characteristic to the chemicals used and the actual voltage does not tell you its condition.
Some "dry cells" deliver 1.5v up to the end of their life whereas others drop to about 1.1v very quickly.

Once you know the name of the cell that drops to 1.1v, avoid them as the operation of the equipment "drops off" very quickly.

However if you have a number of different cells and need to know which ones to keep, here's the solution:

1. Check the voltage and use those with a voltage above 1.1v
2. Next, select 500mA or 10A range on a meter and place the probes on a cell. For a AAA or AA cell, the current should be over 500mA and the needle will swing full scale very quickly.

Keep the testing short as you are short-circuiting the cell but it is the only way to determine the internal impedance of the cell and this has a lot to do with its stage-of-charge.
This will give you a cell with a good terminal voltage and a good current capability.

This also applies to button cells, but the maximum current they will deliver will be less.
If you want to get the last of the energy out of a group of cells they can be used in the following circuits:

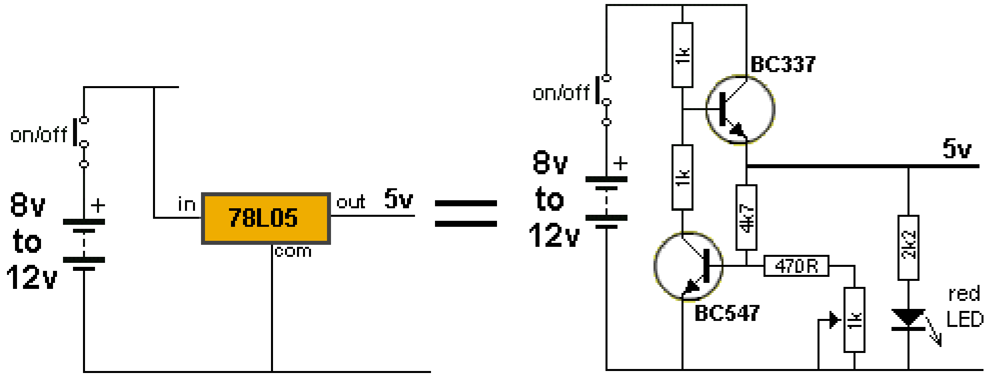

1mV drop @100mA @10v supply

10mV drop @100mA @10v supply

TESTING PIEZO DIAPHRAGMS and PIEZO BUZZERS

There are two types of piezo devices that produce a sound.
They are called **PIEZO DIAPHRAGMS** and **PIEZO BUZZERS**.
A **piezo diaphragm** consists of two metal plates with a ceramic material between. The ceramic expands and contracts when an alternating voltage is placed on the two plates and this causes the main plate to "dish" and "bow."
This creates a high-pitched sound. There are no other components inside the case and it requires an AC voltage of the appropriate frequency to produce a sound.
A **piezo buzzer** has a transistor and coil enclosed and when supplied with a DC voltage, the buzzer produces a sound.
Both devices can look exactly the same and the only way to tell them apart is by connecting a 9v battery. One device may have "+' and "-" on the case to indicate it is a piezo buzzer, but supplying 9v will make the buzzer produce a sound while the piezo diaphragm will only produce a "click."

PIEZO DIAPHRAGM
PIEZO BUZZER

A piezo diaphragm will produce a click when connected to 9v DC.

A piezo buzzer will produce a tone when connected to a DC voltage.

How a PIEZO BUZZER WORKS

A Piezo Buzzer contains a transistor, coil, and piezo diaphragm and produces sound when a voltage is applied. The buzzer in the circuit above is a PIEZO BUZZER.

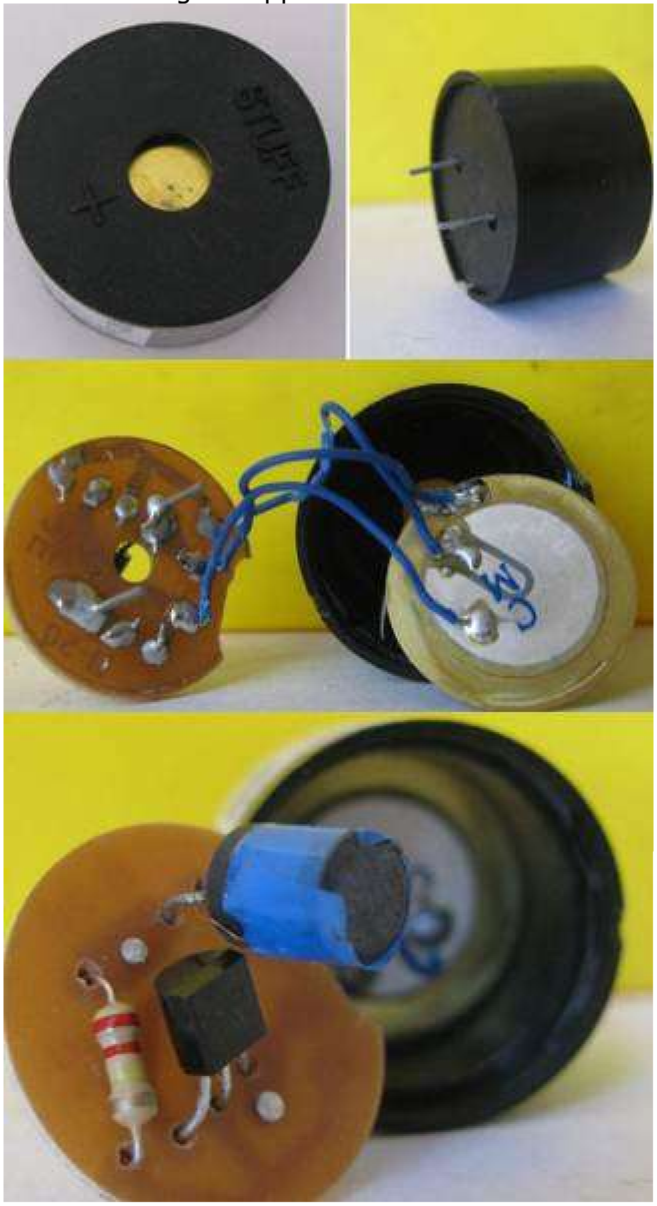

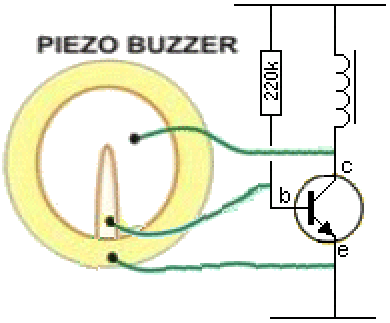

The circuit starts by the base receiving a small current from the 220k resistor. This produces a small magnetic flux in the inductor and after a very short period of time the current does not increase. This causes the magnetic flux to collapse and produce a voltage in the opposite direction that is higher than the applied voltage.

3 wires are soldered to pieces of metal on the top and bottom sides of a ceramic substrate that expands sideways when it sees a voltage. The voltage on the top surface is passed to the small electrode and this positive voltage is passed to the base to turn the transistor ON again. This time it is turned ON more and eventually the transistor is fully turned ON and the current through the inductor is not an INCREASING CURRENT by a STATIONARY CURRENT and once again the magnetic flux collapses and produces a very high voltage in the opposite direction. This voltage is passed to the piezo diaphragm and causes the electrode to "Dish" and produce the characteristic sound. At the same time a small amount is "picked-off" and sent to the transistor to create the next cycle.

TESTING A SPEAKER

A speaker (also called a **loud speaker**) has coil of wire wrapped around a magnet but it does not

touch the magnet as it is wound on a thin cardboard former so that the coil will be pulled closer to the magnet when a current flows in the coil.
When the current flows in the other direction, the coil moves away from the magnet.
The coil is called voice coil and it is connected to a sheet of thin card called a CONE and as the cone vibrates, the speaker reproduces music or noise.
Use a multimeter on a low ohms scale to read the value of resistance of the coil.
It can be as low as 2 ohms or as high as 100 ohms.
Most speakers have an 8R voice coil and the actual resistance may be slightly lower than this.
Some speakers have a resistance of 16R, 32R or 50R and even 75 ohms.
You would think putting a 16R speaker in place of 8R would reduce the sound output, but this is not always the case.
You can even use 50R or 75R and get the same performance.
This may sound amazing, but here is the reason.
The cone is deflected a certain amount due to the current flowing and the number of turns.
These two values are multiplied together to produce a value called AMP-TURNS.
If we have an 8R speaker with 80 turns and 100mA, the result is 0.1 x 80 = 8.
If we use a 16R speaker, the average current flow will be 50mA and the number of turns will be about 160. The multiplication of 0.05 x 160 = 8.
The author then tried a 50R speaker and the sound output was equal to 8R and the same with 75R speaker.
This might not apply in all situations, but the 75R speaker was slightly larger and the ticking sound form the **Metal Detector** kit was louder than using an 8R mini speaker.

To see if the cone of a speaker is undamaged, push it slightly and it will move towards the magnet. If it does not move, it is bent or damaged. If the cone is scratchy when pushed, it is rubbing against the magnet.
A cone should be able to be pushed and pulled from its rest-state. If not, it will produce a distorted sound.

TESTING A CIRCUIT

Whenever you test a circuit, the TEST EQUIPMENT puts "a load" or "a change" on it. It does not matter if the test equipment is a multimeter, Logic Probe, CRO, Tone Injector or simply a LED and resistor.
There are two things you need to know.
1. The IMPEDANCE of the circuit at the location you are testing, and
2. The amount of load you are adding to the circuit via the test equipment.

There is also one other hidden factor. The test equipment may be injecting "hum" due to its leads or the effect of your body at absorbing hum from the surroundings or the test equipment may be connected to the mains.
These will affect the reading on the test equipment and also any output of the circuit. Sometimes the test equipment will prevent the circuit from working and sometimes it will just change the operating conditions slightly. You have to be aware of this.
The last section of this eBook covers High and Low Impedance and understanding impedance is something you need to know.
The point to note here is the fact that the equipment (and the reading) can be upset by hum and resistance/capacitance effects of test equipment. This is particularly critical in high impedance and high frequency circuits.

TESTING INTEGRATED CIRCUITS (IC's)

Integrated Circuits can be tested with a LOGIC PROBE. A Logic Probe will tell you if a line is HIGH, LOW or PULSING.
Most logic circuits operate on 5v and a Logic Probe is connected to the 5v supply so the readings are accurate for the voltages being tested.
A Logic Probe can also be connected to a 12v CMOS circuit.
You can make your own Logic Probe and learn how to use it from the following link:

http://www.talkingelectronics.com/projects/LogicProbeMkIIB/LogicProbeMk-IIB.html

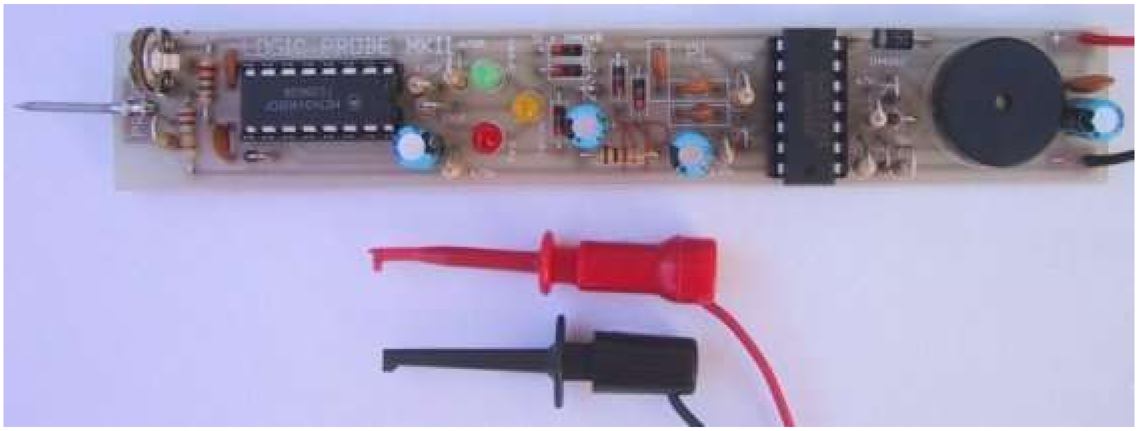
LOGIC PROBE MKII

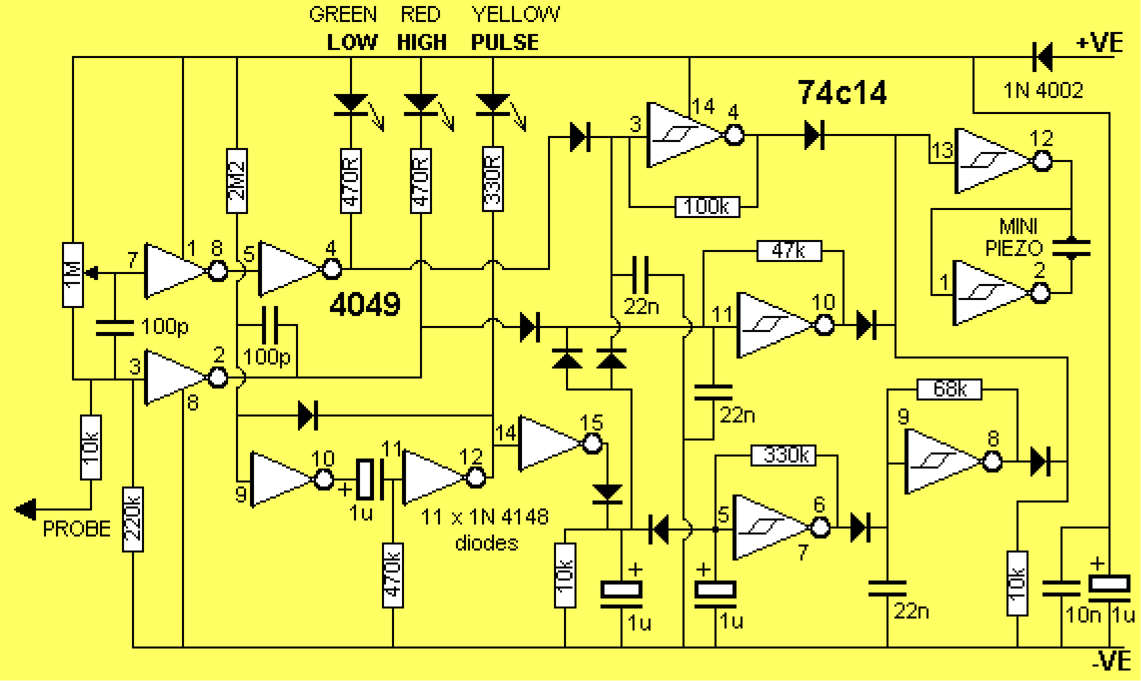
GREEN
RED
YELLOW
LOW
HIGH
PULSE
+VE
1N 4002
74c14
4049
2M2
470R
470R
330R
100k
47k
MINI PIEZO
1M
100p
100p
22n
22n
22n
68k
330k
10k
220k
470k
1u
1u
1u
11 x 1N 4148 diodes
PROBE
10k
10k
10n
1u
-VE

LOGIC PROBE with PULSE

This is a very simple transistor circuit to provide HIGH-LOW-PULSE indication for digital circuits. It can be built for less than $5.00 on a piece of matrix board or on a small strip of copper clad board if you are using surface mount components. The probe will detect a HIGH at 3v and thus the project can be used for 3v, 5v and CMOS circuits.

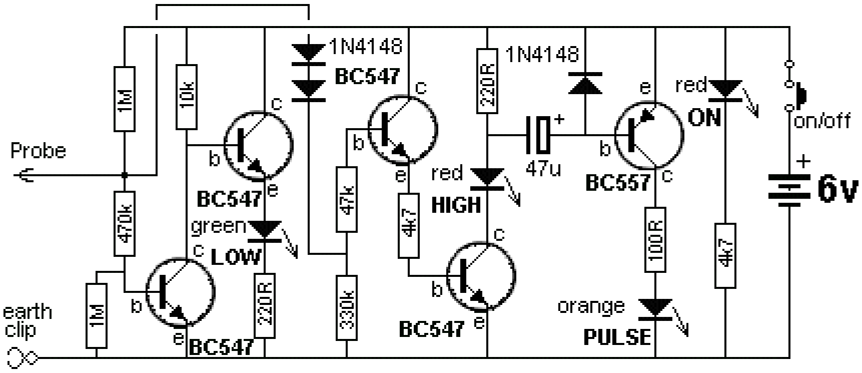
Probe
1M
10k
470k
1M
earth clip
BC547
green
LOW
220R
1N4148
BC547
47k
330k
4k7
red
HIGH
BC547
220R
1N4148
47u
BC557
red
ON
100R
orange
PULSE
4k7
on/off
6v

LOGIC PROBE using CD4001 and CD4011

Here is a simple Logic Probe using a single chip. The circuits have been designed for the **CD4001** CMOS quad NOR gate and **CD4011** CMOS NAND gate. The output has an active buzzer that produces a beep when the pulse LED illuminates (the buzzer is not a piezo-diaphragm but an active buzzer containing components).

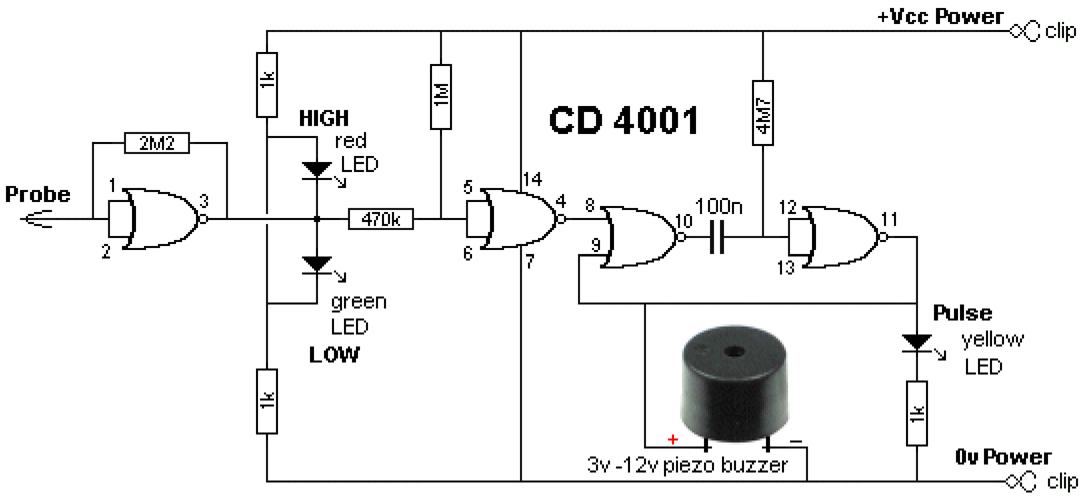
+Vcc Power
clip
1k
1M
4M7
CD 4001
HIGH
red
LED
2M2
Probe
1
2
3
470k
5
6
14
4
7
8
9
10
100n
12
13
11
green
LED
LOW
Pulse
yellow
LED
1k
1k
+
−
3v -12v piezo buzzer
0v Power
clip

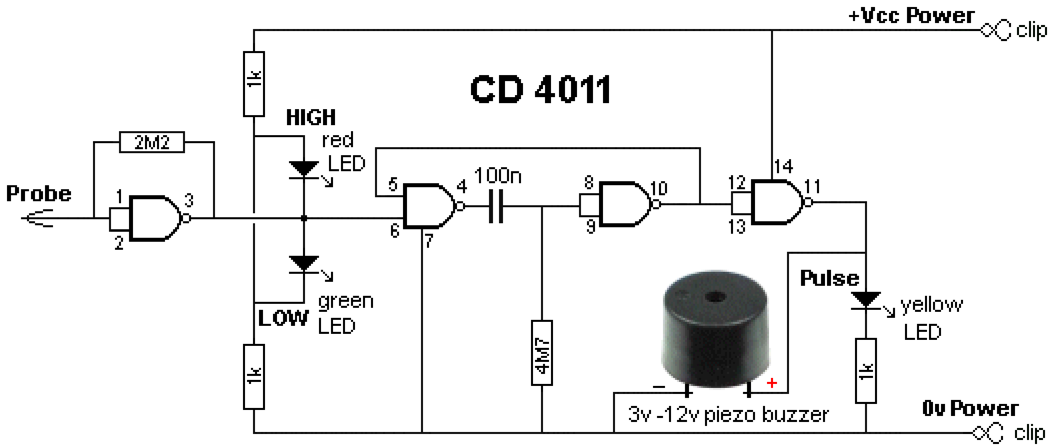
+Vcc Power
clip
1k
CD 4011
HIGH
red
LED
2M2
Probe
1
2
3
5
6
4
7
100n
8
9
10
12
13
14
11
green
LED
LOW
1k
4M7
Pulse
yellow
LED
1k
−
+
3v -12v piezo buzzer
0v Power
clip

SUPER PROBE MkII has 20 different features including a Logic Probe, capacitance tester, Inductance tester, and more.

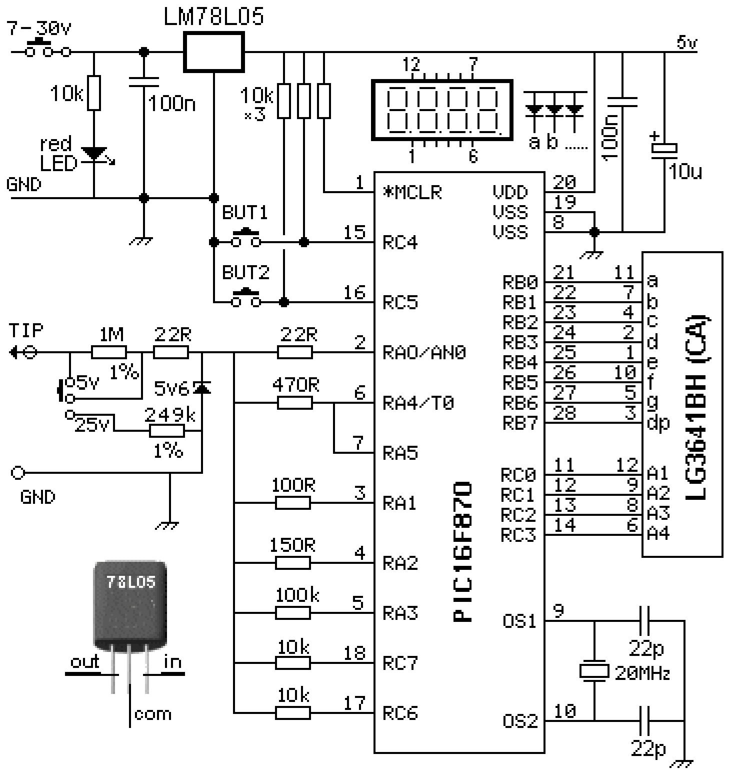
7-30v
LM78L05
5v
10k
100n
red
LED
GND
10k
×3
12
7
1
6
a b ......
100n
10u
1
*MCLR
VDD
VSS
VSS
20
19
8
BUT1
15
RC4
BUT2
16
RC5
TIP
1M
22R
22R
2
RA0/AN0
5v
1%
5v6
470R
6
RA4/T0
25v
249k
1%
7
RA5
GND
100R
3
RA1
150R
4
RA2
100k
5
RA3
10k
18
RC7
10k
17
RC6
PIC16F870
RB0
RB1
RB2
RB3
RB4
RB5
RB6
RB7
21
22
23
24
25
26
27
28
11
7
4
2
1
10
5
3
a
b
c
d
e
f
g
dp
RC0
RC1
RC2
RC3
11
12
13
14
12
9
8
6
A1
A2
A3
A4
LG3641BH (CA)
OS1
9
22p
20MHz
OS2
10
22p
78L05
out
in
com

SUPER PROBE MkII Circuit

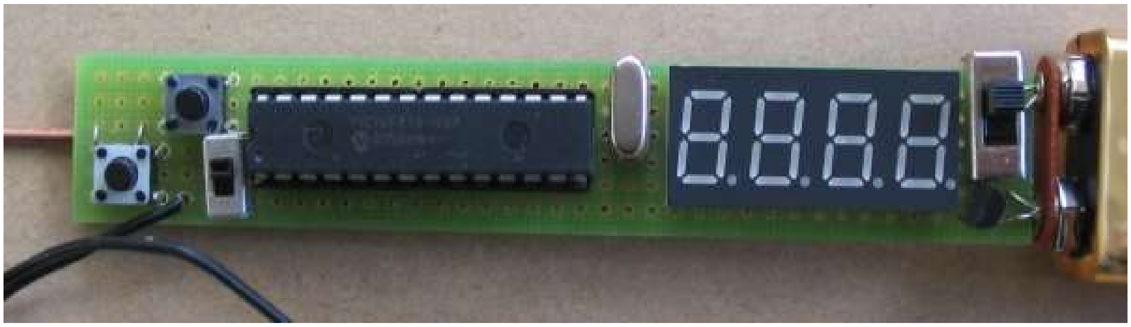

SUPER PROBE MkII

To test an IC, you need a circuit diagram with waveforms. These diagrams will show the signals and are very handy if a CRO (cathode ray Oscilloscope) is used to diagnose the problem. The CRO will reproduce the waveform and prove the circuit is functioning correctly.
A Logic Probe will just show activity and if an output is not producing a "pulse" or "activity," you should check the power to the IC and test the input line.
It is beyond the scope of this eBook to explain how to diagnose waveforms, however it is important to know if signals are entering and exiting an IC and a Logic Probe is designed for this.

SIGNAL INJECTOR

This circuit is rich in harmonics and is ideal for testing amplifier circuits. To find a fault in an amplifier, connect the earth clip to the 0v rail and move through each stage, starting at the speaker. An increase in volume should be heard at each preceding stage. This Injector will also go through the IF stages of radios and FM sound sections in TV's.

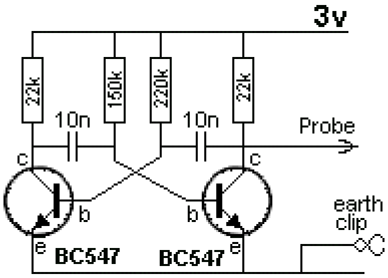

TESTING AUDIO AMPLIFIERS and AUDIO IC's

The **Super Probe MII** described above has a "noise" function and a tone function that allows you to inject a signal into an audio stage, amplifier (made from discrete components) or an audio chip, and detect the output on a speaker.
Audio stages are very difficult to work-with if you don't have a TONE GENERATOR or SIGNAL INJECTOR.
The signals are very small and not detected by a multimeter.
You can start anywhere in an amplifier and when a tone is heard, you can keep probing until the signal is not present or louder. From this you can work out which way the signal is travelling.
A Signal Injector is very handy for finding shorts and broken wires in switches, plugs, sockets and especially leads to headphones.
You can determine the gain of a stage (amplification) by probing before and after a chip or transistor and listen for the relative increase in volume from the speaker.
You can also use your finger to produce "hum" or "buzz" if a **Signal Injector** is not available.
Nearly all audio problems are plugs, sockets and cracks in the PC board, but finding them takes a lot of time and skill.

TESTING IC's - also called "CHIPS"

An Integrated Circuit is also called a "chip." It might have 8 pins or as many as 40. Some chips are ANALOGUE. This means the input signal is rising and falling slowly and the output produces a larger version of the input.
Other chips are classified as DIGITAL and the input starts at 0v and rises to rail voltage very quickly. The output does exactly the same - it rises and falls very quickly.
You might think the chip performs no function, because the input and output voltage has the same value, but you will find the chip may have more than one output and the others only go high after a number of clock-pulses on the input, or the chip may be outputting when a combination of inputs is recognised or the output may go HIGH after a number of clock pulses.

ANALOGUE CHIPS (also see above)
Analogue chips are AUDIO chips or AMPLIFIER chips.
To test these chips you will need three pieces of test equipment:
1. A multimeter - this can be digital or analogue.
2. A Signal Injector
3. A Mini Bench Amplifier.

The Mini Bench Amplifier is available as a kit.

TOSHIBA
SUPER HEAVY DUTY
PROBE
MINI BENCH AMPLIFIER

MINI BENCH AMPLIFIER

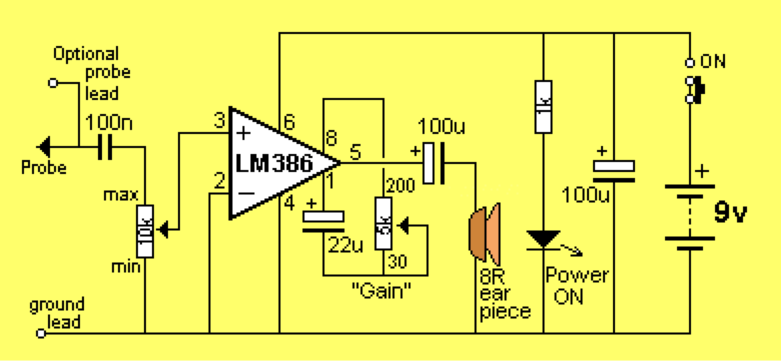
Optional probe lead
100n
Probe
max
10k
min
ground lead
3
2
+
−
LM386
6
8
5
1
4
+
22u
200
5k
30
"Gain"
+
100u
8R ear piece
1k
Power ON
+
100u
ON
+
9v

MINI BENCH AMPLIFIER CIRCUIT

Start by locating the power pin with a multimeter.
If the chip is receiving a voltage, you can use the Mini Bench Amplifier to detect an output.
Connect the Ground Lead of the Mini Bench Amplifier to 0v and touch the Probe tip on each of the pins.
You will hear faint audio on the Input pin and very loud audio on the Output pin.
If no input is detected, you can use a Signal Injector to produce a tone.

Connect the clip of the Signal Injector to 0v and the probe to the input pin of the amplifier chip. At the same time, connect the Mini Bench Amplifier to the output pin and you will hear a very loud tone.
These pieces of test equipment can also be used to diagnose an amplifier circuit constructed with individual components.
Amplifier circuits using discrete components are very hard to trouble-shoot and these pieces of test equipment make it very easy.

DIGITAL CHIPS

It is always best to have data on the chip you are testing, but if this is not available, you will need three pieces of equipment:
1. A multimeter - this can be digital or analogue.
2. A Logic Probe,
3. A logic Pulser.

Firstly test the chip to see if power is being delivered. This might be anything from 3v3 to 15v.
Place the negative lead of the multimeter on the earth rail of the project - this might be the chassis, or the track around the edge of the board or some point that is obviously 0v.
Try all the pins of the chip and if you get a reading, the chip will have "supply."
Identify pin 1 of the chip by looking for the "cut-out" at the end of the chip and you may find a small dimple below the cut-out (or notch). This is pin 1 and the "power pin" can be directly above or any of the other pins.
Next you need to now if a signal is entering the chip.
For this you will need a LOGIC PROBE.
A Logic Probe is connected to the same voltage as the chip, so it will detect a HIGH and illuminate a red LED.
Connect the Logic Probe and touch the tip of the probe on each pin.
You will not know if a signal is an input or output, however if you get two or more active pins, you can assume one is input and the other is output. If none of the pins are active, you can assume the signal is not reaching this IC.
If only one pin is active, you can assume the chip is called a CLOCK (or Clock Generator). This type of chip produces pulses. If more than two pins are active, you can assume the chip is performing its function and unless you can monitor all the pins at the same time, you don't know what is happening.
This is about all you can do without any data on the chip.

If you have data on the chip, you can identify the input(s) and output(s).
A Logic Probe on each of these pins will identify activity.
A Logic Probe has 3 LEDs. Red LED indicates a HIGH, Green indicates a LOW and Orange indicates a PULSE (activity).
Some Logic Probes include a piezo and you can hear what is happening, so you don't take your eyes off the probe-tip.
It is important not to let the probe tip slip between the pins and create a short-circuit.

LOGIC PULSER

If you have a board or a single chip and want to create activity (clock pulses), you can use a Logic Pulser. This piece of test equipment will produce a stream of pulses that can be injected into the clock-line (clock input) of a chip.
You can then use a Logic Probe at the same time on the outputs to observe the operation of the chip.

You can also use the Mini Bench Amplifier to detect "noise" or activity on the inputs and outputs of digital chips.
This only applies if the frequency is in the audio range such as scanning a keyboard or switches or a display.

This is how to approach servicing/testing in a general way. There are thousands of digital chips and if you want to test a specific chip for its exact performance, you will need to set-up a "test-bed."

REMOTE CONTROLS

There are two types of remote control - Infrared and RF. Infrared is used for short-range, line-of-sight for TV's DVD's etc.
A few faults can be fixed, but anything complex needs a new remote control.
Check the batteries and battery-contacts. See if the IR LED is illuminating by focusing it into a digital camera and looking on the screen for illumination.
The only other things are a sticky button, a worn-out button or a crack in the PC board. Water damage is generally too much work to repair.
RF remote controls for cars, garage doors etc need a second working unit to check the power output.
Here is a simple circuit that can be connected to an analog multimeter to detect the signal strength at a very close range:

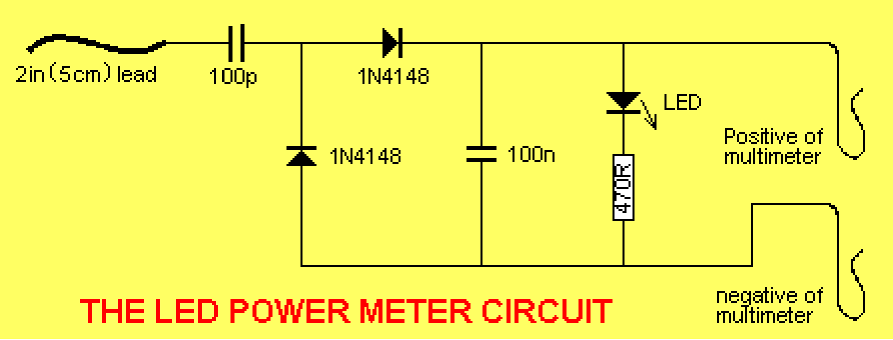
2in (5cm) lead
100p
1N4148
1N4148
100n
LED
470R
Positive of multimeter
negative of multimeter
THE LED POWER METER CIRCUIT

To hear the tone from a transmitter, the **Mini Bug Detector** circuit can be used:

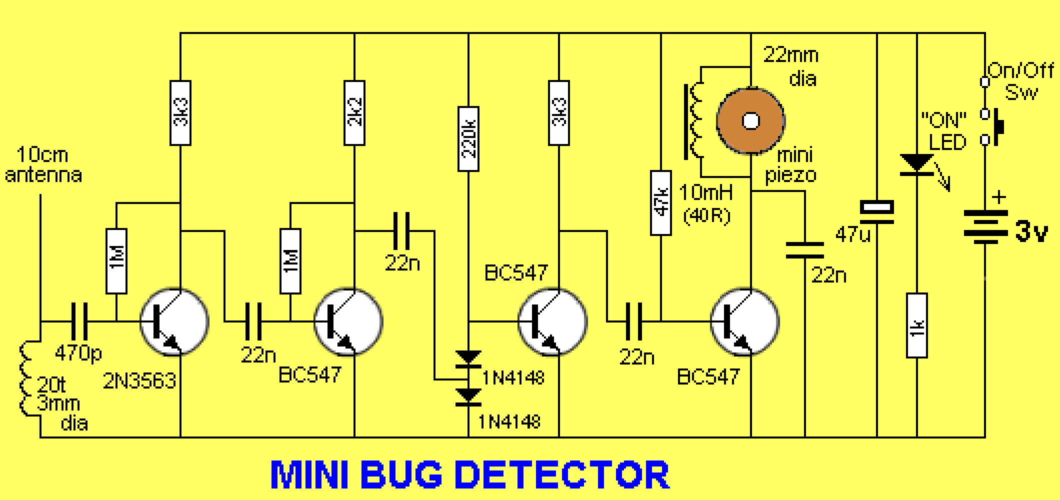
10cm antenna
3k3
1M
470p
20t 3mm dia
2N3563
2k2
1M
22n
BC547
22n
220k
1N4148
1N4148
3k3
BC547
22n
47k
10mH (40R)
22mm dia
mini piezo
BC547
22n
47u
"ON" LED
1k
On/Off Sw
+
3v
MINI BUG DETECTOR

Any further investigation requires a circuit diagram so you can work out what is actually being sent from the transmitter.
Most of the time it is a faulty switch, battery or contacts. Make sure the setting is correct on the "dip switches" and use a working unit to compare all your testing.

TESTING VOLTAGES ON (in) A CIRCUIT

There are basically two different types of circuit.
1. **ANALOGUE CIRCUIT**
An analogue circuit can also be called an AUDIO CIRCUIT and the voltages at different points in a circuit can be measured with a multimeter but the changes (the waveforms) will be quite small or changing at a rapid rate and cannot be detected by a multimeter.
You need a CRO to "see" the signals or a **Signal Injector** to inject a waveform into the circuit and hear the result on the circuit's speaker.

2. **DIGITAL CIRCUIT**
A digital circuit can also be called a "Computer Circuit" or "Logic Circuit" and some of

the voltages can be measured with a multimeter (such as supply voltages) but the "signal lines" will be be changing from HIGH to LOW to HIGH very quickly and these signals are detected with a **Logic Probe**.

Here are some circuits with details of how to test the voltages.
Most circuits do not show voltages at various different points and we will explain what to expect on each "stage."

A "STAGE"

A stage is a set of components with an input and output. A "stage" can also be called a "Building Block."
Sometimes it has a capacitor on the input and one on the output.
This means the stage is completely isolated as far as DC is concerned.
The stage has a supply (a DC supply) and it is producing its own voltages on various points on the "stage." It can only process (amplify) "AC." (signals).
Sometimes the stage can be given a name, such as small-signal amplifier, push-pull amplifier or output.
If the stage has a link or resistor connected to a previous stage, the previous stage will have a "DC effect" on the stage. In other words it will be biasing or controlling the voltages on the stage. The stage may be called a "timer" or "delay" or "DC amplifier."

It is important to break every circuit into sections. This makes testing easy. If you have a capacitor at the input and output, you know all the problems lie within the two capacitors.
In a digital circuit (no capacitors) you need to work on each IC (integrated Circuit) and test the input for activity and all the outputs.

Once you have determined if the circuit is Analogue or Digital, or a combination of both, you have to look at the rail voltage and work out the size or amplitude of the voltage or waveform.
This is done before making a test, so your predictions are confirmed.
You will need a **multimeter** (either Digital or Analogue) a **Logic Probe** and a **Signal Injector** (**Tone Generator**). An analogue meter has the advantage that it will detect slight fluctuations of voltage at a test-point and its readings are faster than a digital meter. A digital meter will produce an accurate voltage-reading - so you should have both available.

HIGH IMPEDANCE AND LOW IMPEDANCE

Every point in a circuit has a characteristic called "IMPEDANCE." This has never been discussed before in any text book. That's why it will be new to you.
In other words, every point will be "sensitive to outside noise."
An audio amplifier is a good example. If you put your finger on the active input, it will produce hum or buzz in the speaker. This is because it is a HIGH IMPEDANCE line or high impedance section of the circuit.
The same applies to every part in a circuit and when you place Test Equipment on a line for testing purposes, the equipment will "upset" the line. It may be very slight but it can also alter the voltage on the point CONSIDERABLY.
We have already mentioned (above) how a cheap multimeter can produce a false reading when measuring across a 1M resistor. That's why you need high impedance test Equipment so you do not "load" the point you are testing and create an inaccurate reading.
The word **Impedance** really means resistance, but when you have surrounding components such as diodes, capacitors, transistors, coils, Integrated Circuits, supply-voltages and resistors, the combined effect is very difficult to work out as a "resistance" and that's why we call it "Impedance."
The term "**High and Low Impedance**" is a relative term and does not have any absolute values but we can mention a few points to help you decide.
In general, the base of a transistor, FET input of an IC are classified as HIGH IMPEDANCE.
The output of these devices are LOW IMPEDANCE.

Power rails are LOW IMPEDANCE.
An oscillator circuit and timing circuit are HIGH IMPEDANCE.
A LOAD is low impedance.
And it gets tricky: An input can be designed to accept a low-impedance device (called a transducer or pick-up) and when the device is connected, the circuit becomes LOW impedance, but the input circuitry is actually high impedance.
The impedance of a diode or LED is HIGH before the device sees a voltage higher than the junction voltage and then it becomes LOW Impedance.
Impedance is one of the most complex topics however it all comes down to testing a circuit without loading it.
That's why test equipment should have an input impedance higher than 1M.

The first circuit we will investigate is the **Mini Bug Detector**, shown above and below. Points on the circuit have been labelled A, B, C etc:

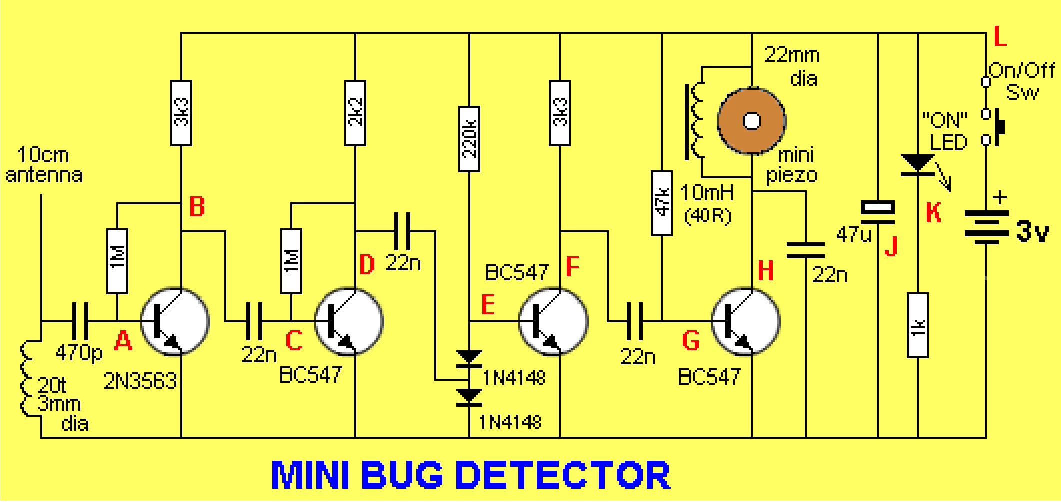

Point A - The first transistor is "self-biased" and will have 0.6v on the base. The antenna is connected to a 20 turn coil and you might think the coil will "short" the signals to earth.
But the coil and 470p capacitor form a circuit that oscillates at a high frequency when the antenna wire picks up stray signals. The coil and capacitor actually amplify the signals (see Talking Electronics website: Spy Circuits to see how a TANK CIRCUIT works) and these signals enter the base of the first transistor.
This is classified as a HIGH Impedance section because the signals are small and delicate and any loading via test equipment will kill them. The first transistor amplifies the signals about 70 times and they appear at **Point B**.

The signal passes though a 22n to **Point C** and the transistor amplifies the signal about 70 times to **point D**. **Point C** is classified as high impedance as any voltage measurement at this point will upset the biasing of the stage as a few millivolts change in base-voltage will alter the voltage on the collector considerably. **Point D** is classified as low impedance as any voltage-testing will not alter the voltage appreciably.
The output of the second stage passes through a capacitor to the join of two diodes. These two diodes are not turned on because the voltage at **Point E** can never rise above 0.7v as this is the voltage produced by the base-emitter of the third transistor.
The purpose of the two diodes is to remove background noise. Background noise is low amplitude waveforms and even though the transistor is turned on via the 220k, low amplitude signals will not be received. The third transistor works like this: It cannot be turned ON any more because any waveform from the 22n will be "clipped" by the bottom diode and it will never rise above 0.6v.
So, the only signal to affect the transistor is a negative signal - to turn it OFF.
Firstly we have to understand the voltage on the 22n. When the second transistor is sitting at mid-rail voltage, the 22n gets charged via the 2k2 and lower diode. When

the transistor gets tuned ON, the collector voltage falls and the left side of the 22n drops. The right side of the 22n also drops and when it drops 0.6v, the top diode starts to conduct and when the voltage on the 22n drops more than 0.6v the third transistor starts to turn OFF. This effect is amplified by the transistor at least 100 times and appears at **Point F.** All the voltages around the two diodes are classified as HIGH Impedance as any piece of test equipment will upset the voltage and change the output.
There are some losses in amplitude of the signal as it passes through the 22n coupling capacitors but the end result is a very high strength signal at **point G.** The 4th transistor drives a 10mH choke and the mini piezo is effectively a 20n capacitor that detects the "ringing" of the inductor to produce a very loud output.
The 22n capacitor on the collector eliminates some of the background noise. The choke and piezo form an oscillatory circuit that can produce voltages above 15v, even though the supply is 3v.
The 47n capacitor at **Point J** is to keep the supply rails "tight" (to create a LOW Impedance) to allow weak cells to operate the circuit.
The "Power-ON" LED tells you to turn the device off when not being used and **Point L** is the power supply - a low impedance line due to the 47u electrolytic.

Testing the Mini Bug Detector

To test the Mini Bug Detector, you will need a Signal Injector.
Place the Injector on **Point G** and you will hear a tone. Then go to **E**, **C** and **A**. The tone will increase in volume. If it does not increase, you have pin-pointed the faulty stage.

The next circuit is a combination of digital and analogue signals. It is a **Logic Probe**:

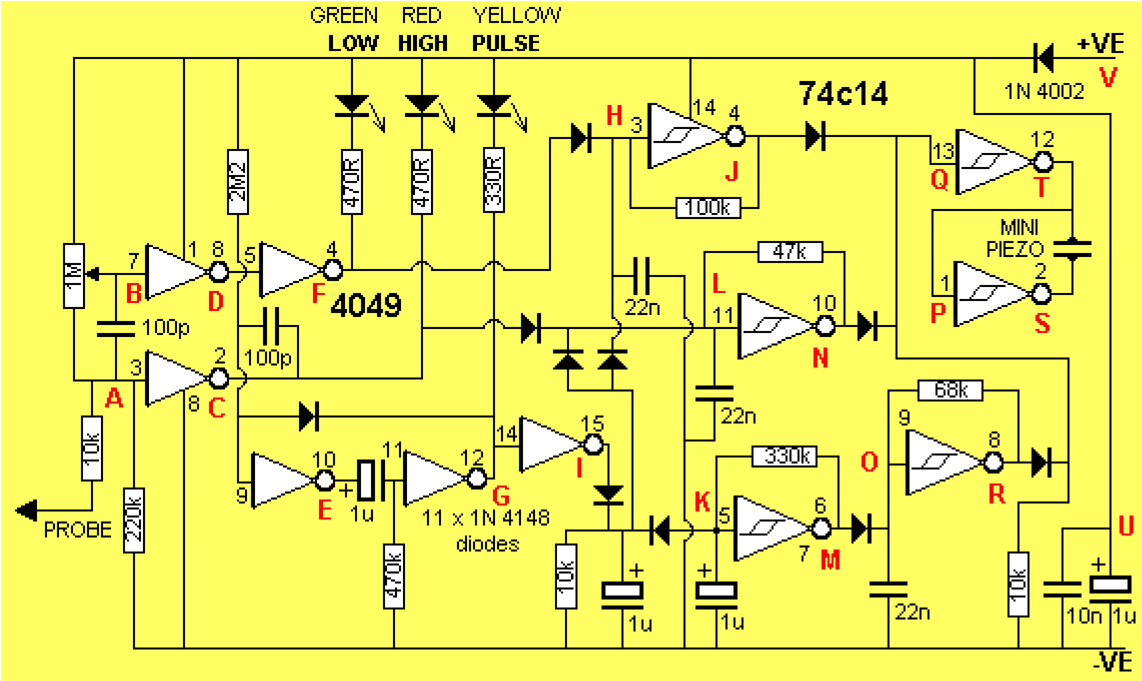

The voltage on a circuit (to be tested) is detected by the probe at **Point A** of the circuit above and the "tip" is classified as "reasonably high impedance" as it has a 220k resistor between the tip and 0v rail. The 1M reduces the impedance by about 20% but the inputs of the two inverters have no effect on the "tip" impedance as they are extremely high input-impedance devices.
The 1M trim pot is designed to put put a voltage on **point B** that is slightly higher than mid-rail so the green LED is turned off.
Point A will see a voltage below mid-rail and **point C** will be HIGH. **Point C** and **F** are low-impedance outputs.
When the tip of the probe is connected to a LOW voltage, Point B sees a LOW and Point F goes LOW to illuminate the green LED. At the same time it removes the "jamming voltage" produced by the diode between pin 4 of the 4049 and pin 3 of the

74C14 and the oscillator between **points H and J** produces a low-tone via the 100k resistor and 22n to indicate a LOW.
When the probe tip sees a HIGH, a lot more things happen.
Point C goes LOW and turns on the red LED. At the same time the 100p is in an uncharged state and the right lead goes LOW. This takes the left lead LOW as the left lead connects to a HIGH Impedance line and pin 9 goes LOW. This makes **point E** HIGH
and since the 1u is in an uncharged state, pin 11 goes HIGH. This makes **point G** LOW and the diode between pins 9 and 12 keeps pin 9 LOW and takes over from the pulse from the 100p. The yellow LED is illuminated. The 1u starts to charge via the 470k and when it is approx half-charged, pin 11 sees a HIGH and **point G** goes low. This creates the length of pulse for the yellow LED.
At the same time, **Point L** goes LOW because the "jamming diode" from pin 2 of the 4049 goes low and allows the inverter between point L and N to produce a tone for the piezo.
In addition, **Point I** goes HIGH and quickly charges a 1u electrolytic. This removes the effect of the jamming diode on pin 5 of the 74C14 and a low frequency oscillator made up of 68k and 1u between pins 5&6 turns on and off an oscillator between **points O** and **R** to get a beep. The mini piezo is driven n bridge mode via the two gates between **points QT** and **PS**.
Point U is a 1u electrolytic to reduce the impedance of the power rail and **Point V** is a protection diode to prevent damage if the probe is connected to the supply around the wrong way.

Testing the Logic Probe

You can test the Logic Probe with the simple Logic Probe with Pulse project described above. It will let you know if each point in the circuit is HIGH or LOW. You will also find out the difficulty in testing the points that are HIGH Impedance, as the Probe will upset the voltage levels and the reading may be inaccurate.

More circuits will be added here in the future.

THE VOLTAGE DIVIDER - this topic could fill a book.

You need to read lots of other sections in this eBook, including the section on measuring across a resistor with a multimeter, and high impedance circuits, to fully understand the complexities of a VOLTAGE DIVIDER CIRCUIT.
It is one of the most important BUILDING BLOCKS to understand. Even though it may consists of two components, you have to understand what is happening between these two components. You have to realise there is a voltage at their join that will be rising and falling due to one of the components changing RESISTANCE. Sometimes you can work out the voltage at the join by using Ohm's LAW but quite often it will be impossible as it is changing (rising and falling) during the operation of the circuit.
At the beginning of this discussion we will only dealing with DC circuits and the voltage across a particular component will be due to its RESISTANCE. We are not going into any formulas, as it is very easy to measure the voltages with a multimeter set to VOLTS and you will have an accurate result.
The simplest two components in series are resistors. They always have the same resistance during the operation of a circuit and the voltage across each will not change.
In a further discussion we will cover "resistors" that change value according to the temperature. These are called THERMISTORS. And we have "resistors" that change value according to the light they receive. These are called LIGHT DEPENDENT RESISTORS (LDR's) or PHOTO RESISTORS.
A transistor that is partly or fully turned ON can be considered to be similar to a resistor.
In these 3 cases we need to measure the voltage at the join with a voltmeter as it will be a lot of work to measure the resistance and work out a value.
You can also keep a voltmeter on the joint and watch the voltage change.
Finally we have some components that produce a fixed voltage across them (or nearly fixed) and the remaining voltage is dropped across a resistor. These

components MUST have a resistor connected in series to limit the current and allow the component to pass the specified in the datasheet.
These devices include LEDs, diodes and zener diodes. A LED will have a fairly fixed voltage across it from 1.7v to 3.6v depending on the colour. A diode will have a voltage of 0.7v across it when it is connected to a voltage via a resistor. And a zener diode will have a fixed voltage across it when it is connected with the cathode to the positive rail via a resistor. The voltage across it will be as marked on the zener.
The concept of a VOLTAGE DIVIDER is very simple, but it takes a lot of understanding because both VOLTAGE and CURRENT are involved in the UNDERSTANDING-PROCESS.
Each component has a resistance and this can be measured with a multimeter. When two components are connected in series, a current will flow and a voltage will develop across each item.
More voltage will develop across the item with the higher resistance and the addition of each voltage will always equal the supply voltage.
That's the simple answer.
There is a little more involved . . . It is the word CURRENT. Here is an explanation:
Suppose we have a 1k and 2k resistor on a 12v supply. The voltage at the join will be 4v.
In other words, there will be 4v across the 1k and 8v across the 2k.
If we have a 10k and 20k resistors in series, the voltage will also be 4v at the join.
If we have a 100k and 200k resistors, the voltage will also be 4v at the join.
The voltage will be the same in all cases, but the current will be different. The current in the second case will be one-tenth and only one hundredth in the third case.
If you want to go further, place a one ohm and two ohm in series and get 4v. But the resistors will get very hot and burn out very quickly.

SOLDERING

Here are three 30-minute videos on soldering.
1. **TOOLS**
2. **Soldering components**
3. Soldering **SURFACE MOUNT** components

TESTING A MOTOR

Strictly speaking, a motor is not an electronic component, but since a website gave a useless description on testing motors, I have decided to supply the correct information.
The only REAL way to test a motor is to have two identical motors and check the torque by connecting them to a low voltage and trying to stop the shaft with your fingers. This will give you two results. Firstly it will let you know the torque of the motor.
This is the twisting effect of the shaft. There is no way to determine the torque by knowing the voltage or current.
The unknown factor is the strength of the field magnets (permanent magnets) and this determines the torque.
Secondly, feeling the shaft will let you know if the torque is even for a complete revolution.
By having two identical motors, you can see if one has a lower torque.
Almost nothing can go wrong with a motor except for the brushes. If the brushes wear out, additional resistance will be produced at the interface between the brush and commutator and this can be detected by allowing the shaft to rotate slowly and feeling the resistance as it revolves. A 3-pole motor will have three places where the strength is greatest and each should have the same feeling. A 5-pole motor will have five places of strength.
If the strength is weak or not uniform, the motor is faulty.
You cannot test a motor with a multimeter as the resistance of the armature winding is very low and if the motor is allowed to spin, the back voltage produced by the spinning, increases the reading on the meter and is false.

Micro motors have a coreless armature. This means the 3 windings for the armature are wound on a machine then bent slightly into shape and glued. A circular magnet with 3 poles is in the centre and the armature rotates around this.

This type of motor is reasonably efficient because the armature is the greatest distance from the point of rotation, and the motor reaches full RPM very quickly because the armature has very little inertia.

I have not heard of the armature-winding flying apart but if you hear any scraping noise, it may be the winding.

3-pole, 5-pole and micro motors can be found in printers, eject mechanisms of CD players, toys, RC helicopters, cars etc and rarely fail.

Motors do not work on "voltage." They actually work on CURRENT and as you increase the voltage, more current will flow and produce a stronger magnetic field (by the winding on each pole). This magnetic field will be attracted by the permanent magnet surrounding the armature and repelled by the surrounding permanent magnet, depending on where the face of the pole is, during each revolution.

If the permanent magnet is not very strong, the repulsion part of the interaction will be very weak and thus the torque will be small.

Because motors work on "current" you must have a high current available when you increase the voltage as the motor will require short bursts of high current during each revolution.

It is the combination of voltage and current (called watts) that gives the motor "strength" (torque) as well as the "strength" of the permanent magnets (called the field magnets) and the number of turns of wire on each pole (and the gauge of wire).

Basically, if a motor is hard to spin, and has 3 "hard spots" on each revolution, it will be powerful.

A 2-pole motor does not self-start and will spin in either direction. But a 3-pole motor will self-start and you can determine the direction of rotation.

A 5-pole motor has a lower RPM. It is slightly smoother in output but is not more powerful than a 3-pole version.

A motor with "permanent magnets" is called a DC motor as it will not work on AC. If the magnets are replaced with a coil, it will work on AC and it will be called a "shunt wound" motor of the field coil is connected across the same terminals as the brushes or a "series wound" motor if the field coil is in series with the armature.

TESTING COMPONENTS "IN-CIRCUIT"

You can test components while they are IN CIRCUIT, but the surrounding components will have an effect on the results.

You can get all sorts of "In-Circuit" testers. They are expensive and offer little more accuracy than a multimeter.

In-Circuit testing with a multimeter can give you the same results as a tester.

All you have to do is turn the project ON and use a multimeter (set to voltage) to determine the voltage at various points. It is best to have a circuit of the equipment so you can what to expect at each point.

Only major departures from the expected can be located in this way.

Obviously the first thing to look for is burnt-out components. Then feel components such as transistors for overheating.

The look for electrolytics that may be dry. Sometimes these have changed colour or are slightly swollen.

If they are near hot components, they will be dry.

For the cost of a few dollars I change ALL THE ELECTROLYTICS in some pieces of equipment, as a dry electrolytic is very difficult to detect.

Testing a transistor "in-circuit" is firstly done with the supply ON. That's because it is quicker.

Measure the voltage between ground and collector.

In most cases you should get a voltage of about half-rail. If it is zero, or close to rail voltage, you may have a problem.

Turn off the supply and use the multimeter on low-ohms to measure all six resistances between the leads.

A low resistance in both directions on two leads will indicate a fault.

Resistors almost NEVER go "HIGH." For instance, a 22k will never go to 50k.

However a low-value resistor will "burn-out" and you will read the value of the surrounding components.

Don't forget, some low-value resistors are designed to burn-out (called fusible resistors) and anytime you find a damaged low-value resistor, you will need to look for the associated semiconductor.

You can replace the resistor quickly and turn the circuit ON to see it burn out again.

Alternatively you can trace though the circuit and find the shorted semiconductor.

It's always nice to "see the fault" then "fix the fault."

Sometimes a transistor will only break-down when a voltage is present, or it may be influenced by other components.

When the piece of equipment is turned OFF, you can test for resistance values. The main thing you are looking for is "dry joints" and continuity. Dry joints occur around the termination of transformers and any components that get hot. Rather than wasting time checking for dry joints, it is better to simply go over the connections with a hot iron and fresh solder.

You may need to check the continuity of a track (trace) and it may go from one side of the PC board to the other.

Use a multimeter set to low-ohms and make sure the needle reads "zero-ohms."

It is very dangerous to do any testing on a project using a multimeter set to "amps" or "milliamps."

You cannot test "current flowing through a component" by placing the probes across a component. You will simply over-load the rest of the circuit and create a problem.

To find out if current is flowing though a circuit or a low-value resistor, turn the project ON and measure the voltage either across the component or the voltage on one end then the other.

A voltage-drop indicates current is flowing.

That's about it for testing "in-circuit." Use the rest of this eBook to help you with diagnosis.

Don't think an IN-CIRCUIT COMPONENT TESTER is going to find a fault any faster than a multimeter. They all use a multimeter principle.

SHORT CIRCUIT

Nearly every component can fail and produce an effect called a SHORT CIRCUIT.

This basically means the component takes more current than normal and it may fail completely or simply take more current and the operation of the circuit may be reduced only a small amount.

The resistance of the component may reduce a very small amount but this may have a very large effect on the operation of the circuit.

For instance, two turns in the horizontal or vertical winding of a yoke on the picture tube or monitor may arc and weld together and reduce the size of the picture on the screen, but measuring the winding will not detect the difference in resistance.

The same with the windings on a motor and a short between two winding in a transformer.

If the "short" is between two near-by turns, the change in resistance will be very small. If the "short" is between to different layers, the resistance will be reduced and it may be detected.

When a "short" occurs, the winding turns into a transformer. To be exact, an AUTO-TRANSFORMER.

In the following diagram you can see a normal winding in fig A:

A

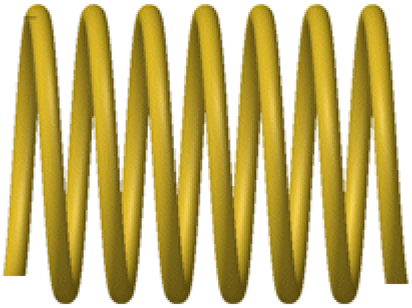

B

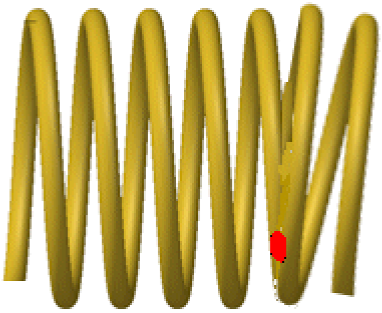

C

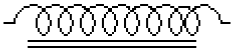

D

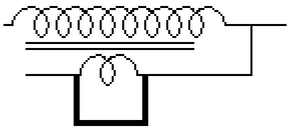

Fig B shows two turns touching each other and if the wire is enamelled, the coating has been damaged so the copper wire from the two turns is touching. This is called a SHORTED TURN.
In fig C you can see two turns touching.
In fig D the shorted-turn has been moved to the other side of the symbol to show the effect it has on the operation of the winding.
The shorted-turn is exactly like the secondary of a transformer with a "jumper" across the output.
This will produce a very high current in the secondary.
A very high current flows through the shorted turn and this changes the operation of the rest of the winding.

1. In most cases a SHORT CIRCUIT can be detected by feeling the additional heat generated by the component.
2. Next, turn off the supply and measure the resistance of the component. If it is lower than expected, the component will be faulty.
3. Next, measure the voltage across the component. If it is lower than normal, the component will be faulty.
4. Next, measure the current taken by the component. If it is higher than normal, the component will be faulty.
5. If the component is an inductor, such as a motor, coil or transformer, you can use an inductance meter. Compare a good winding with a faulty winding. Sometimes the fault will disappear because an arc develops across the fault when the component is operating.

INTERNAL AND EXTERNAL SHORTS

An **internal short** refers to two windings shorting together and the winding has a very high resistance between the winding and the frame on which it is wound. An external short refers to a winding shorting to the frame of the component - such as one of the armature windings shorting to the metal core, around which the wire is

wound.
This may not be important unless another winding shorts to the metal frame and creates "**inter winding**" problems (**inner winding** problems is within the same winding).

The opposite to a short circuit is an OPEN CIRCUIT.

This is generally a broken lead or contact or a wire that has "burnt-out" or been "eaten-away" by acid attack or galvanic action by water and voltage (current).
1. No current will flow when an OPEN CIRCUIT exists.
2. The voltage on each end of the OPEN CIRCUIT will not be the same.
3. Measure the current across the OPEN CIRCUIT and determine if excess current is flowing.
4. Join the two ends of the OPEN CIRCUIT and see if the circuit operates normally.

> HEATSINKS

This is not an electronic component but it can certainly affect the operation of a circuit.
If you cannot hold your fingers on a heatsink, it is getting too hot. This is because the actual location where the heat is being generated is much hotter than the part you are touching.
Transistors and IC's can withstand a high temperature but if they go above this temp, they BLOW UP.
They also have a shorter life when operating at a high temperature.
The secret to a good heatsink is called an INFINITE HEATSINK.
This is the metal frame of a case.
There are lots of charts and data on choosing a heatsink but they don't take into account two factors:
Sometimes a circuit takes a very high current for a short time and this creates a high temperature gradient. This will cause the transistor to get very hot and fail.
The solution is to have two or more transistors in parallel to separate the "heat spots."
The second problem with designing a heatsink is the unknown location of the heatsink and the air-flow. Products placed on a shelf or in a cupboard will get very little air-flow.
Remember: some transistors are mounted on thermal insulators. This means the transistor will have a voltage on it but the heatsink will be zero voltage.
The temperature of the transistor will be MUCH HIGHER than the heatsink under the transistor and the transfer of the heat from the transistor to the heatsink will be very slow. This can be the cause of the transistor failing. Sometimes the transistor will fail because insulation is high temp plastic and it gets brittle. The plastic can carbonise and leak and sometimes a voltage can flash through the insulator. Some amazing things have happened under these transistors and you may need to pull it apart and replace all the insulation.
Finally, feel the heatsink after 15 minutes and feel right up to the transistor. If you cannot touch the transistor, increase the thickness of the heatsink or use two transistors to dissipate the heat.
To design a heatsink, you have to have some idea of the size of a heatsink for the application.
Charts and data can send you in the wrong direction.
Start with a heatsink twice the recommended size and feel the temp after 15 minutes. Put the project in a cupboard and see how the temperature rises.
If possible, connect the heatsink to the metal case to get added dissipation and if you include fan-cooling, remember the fan will eventually gather dust and reduce its efficiency.
It is very difficult to explain how heat passes through a mica washer or plastic washer, but if the transistor has a copper base, the heat transfer has a value of 400. For aluminium it is 200. If it is steel, the transfer has a value of 50. For a mica sheet it is 1 and for plastic it is 0.1
Even though the sheet is very thin, the transfer is a lot less than metal-to-metal transfer.

Most references state the temperature difference is about one degree C for each watt of heat generated by the transistor.
Don't believe anything you read.
Feel the temperature yourself and if you cannot hold your finger on the transistor, fix the problem.

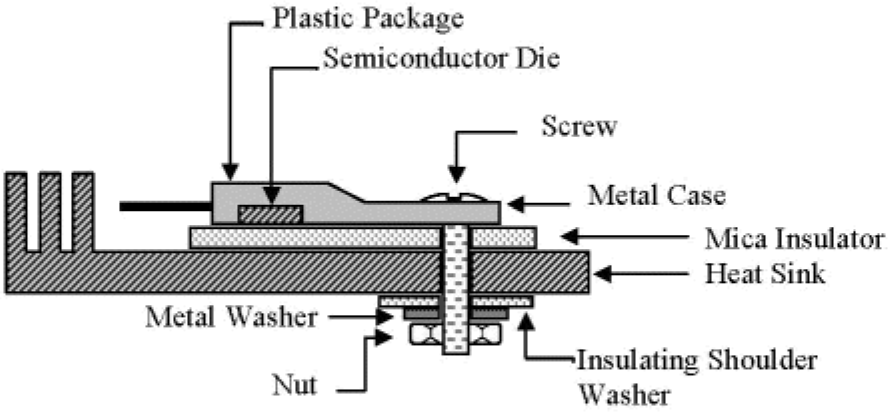
Plastic Package
Semiconductor Die
Screw
Metal Case
Mica Insulator
Heat Sink
Metal Washer
Nut
Insulating Shoulder Washer

In the end, use a heatsink 50% larger than recommended.

THE END

This is not the full story to learning about servicing. It is just the beginning.
We have only covered the simplest tests and shown how 90% of faults can be found by checking voltages, waveforms and looking for obvious things such as burnt out components, cracks in PC boards.
The author has fixed over 35,000 TV's, radios, stereos, VCRs and all those things that were on the market 30 years ago.
Things have not changed. It's just that some repairs cost nearly as much as buying a new product and half the customers opt for dumping a faulty item and buying the latest "flat screen" version. That's why you have to get things through the workshop as fast and as cheaply as possible, to make a living.

If you want any more devices added to this list, email **Colin Mitchell**.

To help with understanding how a transistor circuit works, we have produced an eBook: **The Transistor Amplifier.** It covers a whole range of circuits using a transistor.

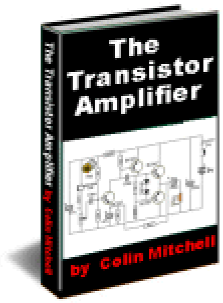
The Transistor Amplifier
by Colin Mitchell
The Transistor Amplifier by Colin Mitchell

Not copyright by Colin Mitchell
You can use any of this material. Please pass this eBook to your friends and let them know that everything on the web is FREE. I have looked at all the "Pay Sites" and found the information they "sell" is available on the web at NO COST.
Nearly all text books are also downloadable for free on "Download.com" etc and when you see a used copy of a $74.00 textbook on Amazon for $12.00 you realise many users have already discarded their copy. A good textbook never gets thrown out or sold for $12.00!!!

See the enormous amount of information on Talking Electronics website

CIRCUIT SYMBOLS by TALKING ELECTRONICS

AC current: 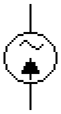 voltage: 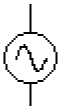	ALTERNISTOR TRIAC A TRIAC and 33 - 43v DIAC 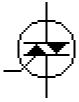 Main Terminal1 Gate Main Terminal 2	Ammeter (amp meter) 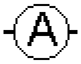
AND Gate 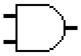	AND Gate 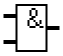	Antenna balanced 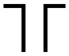
Antenna Loop, Shielded 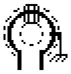	Antenna Loop, Unshielded 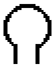	Antenna unbalanced 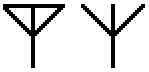
Attenuator, fixed (see Resistor) 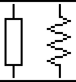	Attenuator, variable (see Resistor) 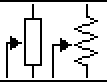	Battery 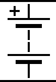
Bilateral Switch (DIAC) 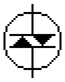	Bridge Rectifier (Diode Bridge) 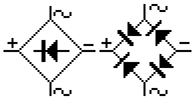	BUFFER (Amplifier Gate) 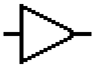
BUFFER (Amplifier Gate) 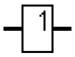	Buzzer 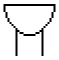	Capacitor feedthrough 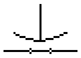
Capacitor non-polarised 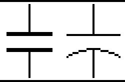	Capacitor polarised (see electrolytic) 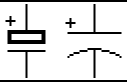	Capacitor Variable 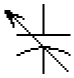
Cavity Resonator 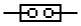	Cell 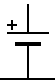	Circuit Breaker 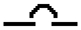
Coaxial Cable 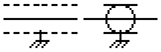	CRO – Cathode Ray Oscilloscope 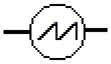	Crystal Microphone (Piezoelectric) 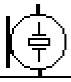
Connectors 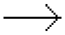 Plug (male) 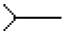 Jack (female) 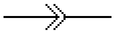 connected 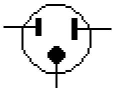 Plug (male) 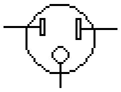 (female)	Crystal Piezoelectric 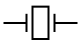	DC current: 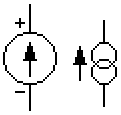 voltage: 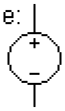
	Darlington Transistor 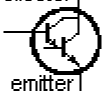 collector base emitter	Delay Line 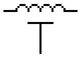
	DIAC (Bilateral Switch) 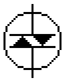	Diode 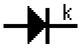 k
Diode - Gunn 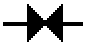	Diode - Light Emitting (LED) 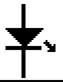	Diode Photo Sensitive 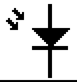
Diode Photovoltaic 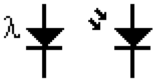 λ	Diode Bridge (Bridge Rectifier) 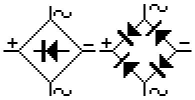	Diode - Pin 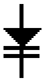
Diode - Varactor 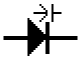	Diode - Zener 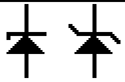	Earth Ground 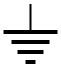
Earpiece (earphone, crystal earpiece) 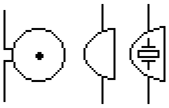	Electroluminescence 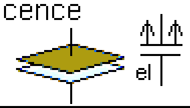 el	Electret Microphone (Condenser mic) 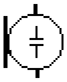
Electrolytic (Polarised Capacitor) 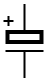 alternate symbols: (positive on top) 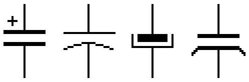	Electrolytic - Tanatalum 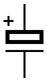 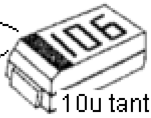 positive end black band or chamfer 10u tantalum	Exclusive-OR Gate (XOR Gate) 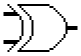
		Exclusive-OR Gate (XOR Gate) 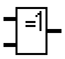
Field Effect Transistor (FET) n-channel also: N-Channel J FET 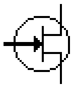 Gate Drain Source	Field Effect Transistor (FET) p-channel also: P-Channel J FET 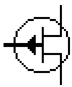 Gate Drain Source	Flashing LED (Light Emitting Diode) (Indicates chip inside LED) 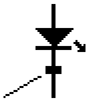

Ferrite Bead 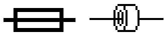	Fuse 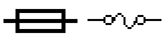	Galvanometer 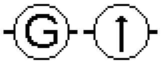
Globe 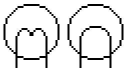	Ground Chassis 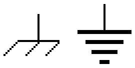	Ground Earth 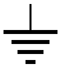
Heater (immersion heater) (cooker etc) 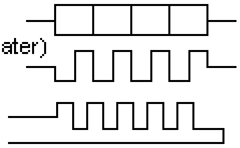	IC Integrated Circuit 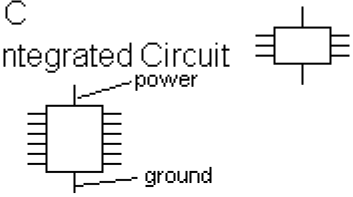 	Inductor Air Core 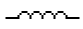
Headphone 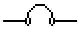		Inductor Iron Core or ferrite core 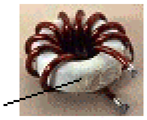 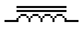
Inductor Tapped 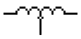	Inductor Variable 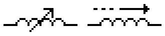	Integrated Circuit 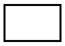
Inverter (NOT Gate) 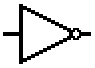	INVERTER (NOT Gate) 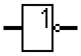 	
Jack Co-axial 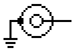	Jack Phone (Phone Jack) 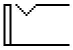	Jack Phone (Switched) 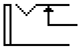
Jack Phone (3 conductor) 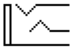	Key Telegraph (Morse Key) 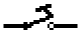	Lamp Incandescent 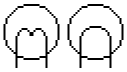
Lamp - Neon 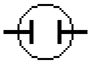	LASCR (Light Activated Silicon Controlled Rectifier) 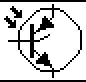	LDR (Light Dependent Resistor) 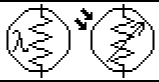
LASER diode 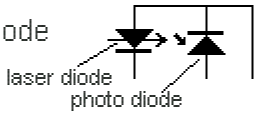 	Light Emitting Diode (LED) 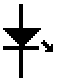	Light Emitting Diode (LED - flashing) 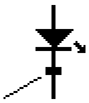 (Indicates chip inside LED)
Mercury Switch 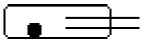	Micro-amp meter (micro-ammeter) 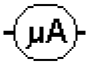 	Microphone (see Electret Mic) 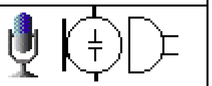
Microphone (Crystal - piezoelectric) 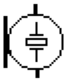	Milliamp meter (milli-ammeter) 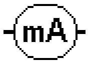 	Motor 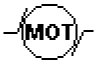
NAND Gate 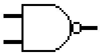	NAND Gate 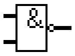 	Nitinol wire "Muscle wire" 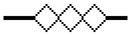
Negative Voltage Connection 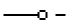	NOR Gate 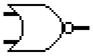	NOR Gate 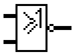
NOT Gate Inverter 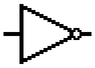	NOT Gate Inverter 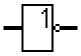 	Ohm meter 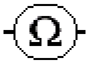
Operational Amplifier (Op Amp) 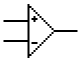	Optocoupler (Transistor output) 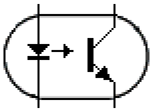	Opto Coupler (Opto-isolator) 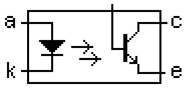 Photo-transistor output
Optocoupler (Darlington output) 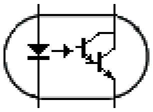	Opto Coupler (Opto-isolator) 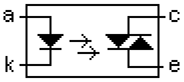 TRIAC output	OR Gate 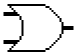
OR Gate 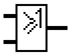	Oscilloscope see CRO 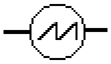	Outlet (Power Outlet) 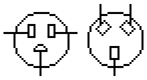
Piezo Diaphragm 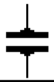	Photo Cell (photo sensitive resistor) 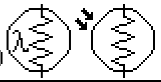	Photo Diode 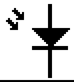
Photo Darlington Transistor 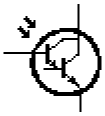	Photo FET (Field Effect Transistor) 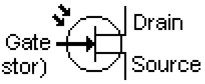 	Photo Transistor 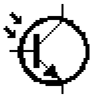

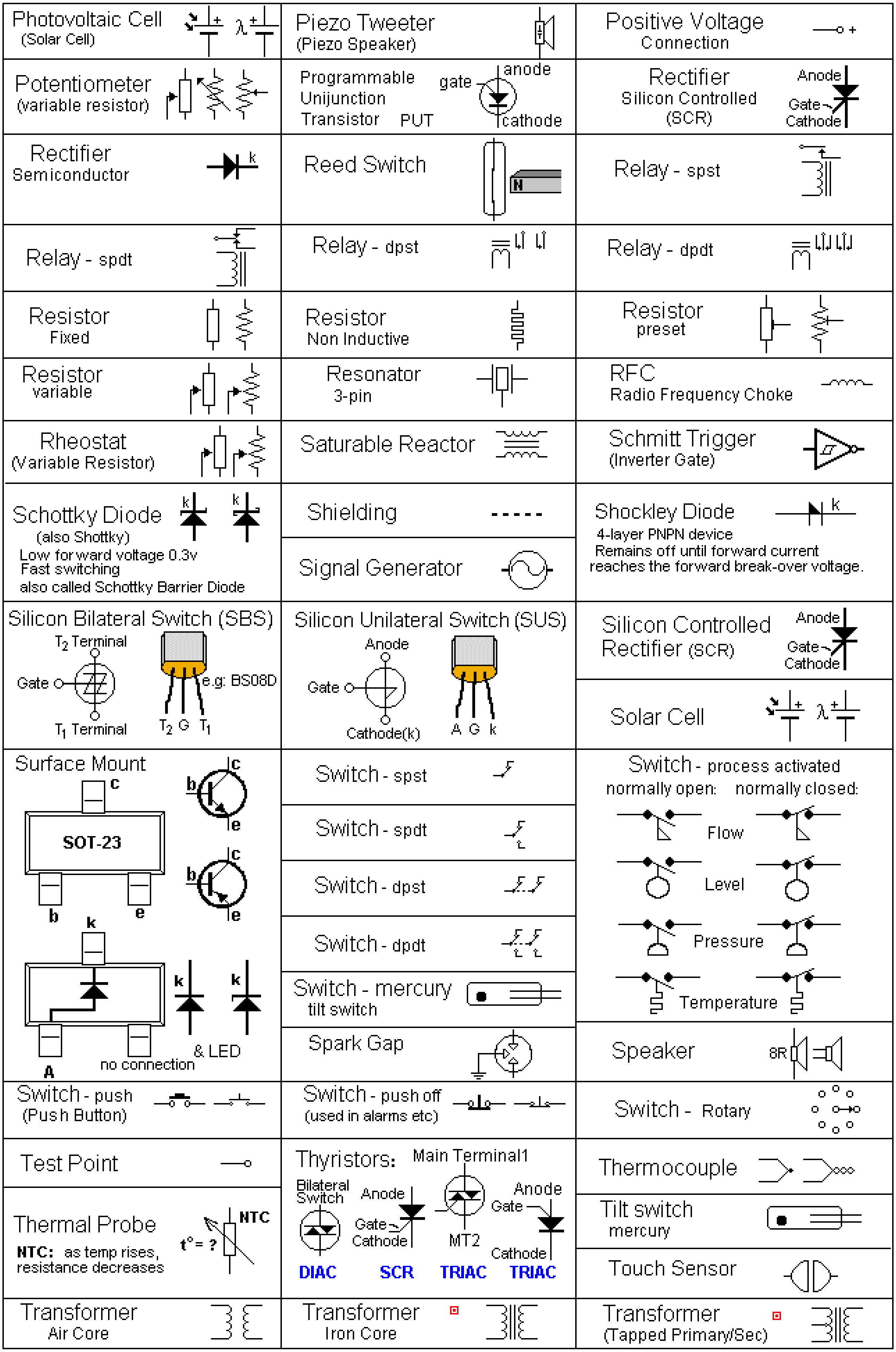

Photovoltaic Cell (Solar Cell)	Piezo Tweeter (Piezo Speaker)	Positive Voltage Connection
Potentiometer (variable resistor)	Programmable Unijunction Transistor PUT gate anode cathode	Rectifier Silicon Controlled (SCR) Anode Gate Cathode
Rectifier Semiconductor k	Reed Switch N	Relay - spst
Relay - spdt	Relay - dpst	Relay - dpdt
Resistor Fixed	Resistor Non Inductive	Resistor preset
Resistor variable	Resonator 3-pin	RFC Radio Frequency Choke
Rheostat (Variable Resistor)	Saturable Reactor	Schmitt Trigger (Inverter Gate)
Schottky Diode (also Shottky) Low forward voltage 0.3v Fast switching also called Schottky Barrier Diode k k	Shielding Signal Generator	Shockley Diode k 4-layer PNPN device Remains off until forward current reaches the forward break-over voltage.
Silicon Bilateral Switch (SBS) T_2 Terminal Gate T_1 Terminal e.g: BS08D T_2 G T_1	Silicon Unilateral Switch (SUS) Anode Gate Cathode(k) A G k	Silicon Controlled Rectifier (SCR) Anode Gate Cathode Solar Cell
Surface Mount c SOT-23 b e b c e b c e k k k A no connection & LED	Switch - spst Switch - spdt Switch - dpst Switch - dpdt Switch - mercury tilt switch Spark Gap	Switch - process activated normally open: normally closed: Flow Level Pressure Temperature Speaker 8R
Switch - push (Push Button)	Switch - push off (used in alarms etc)	Switch - Rotary
Test Point Thermal Probe NTC t° = ? **NTC:** as temp rises, resistance decreases	Thyristors: Main Terminal1 Bilateral Switch Anode Gate Cathode MT2 Anode Gate Cathode **DIAC** **SCR** **TRIAC** **TRIAC**	Thermocouple Tilt switch mercury Touch Sensor
Transformer Air Core	Transformer Iron Core	Transformer (Tapped Primary/Sec)

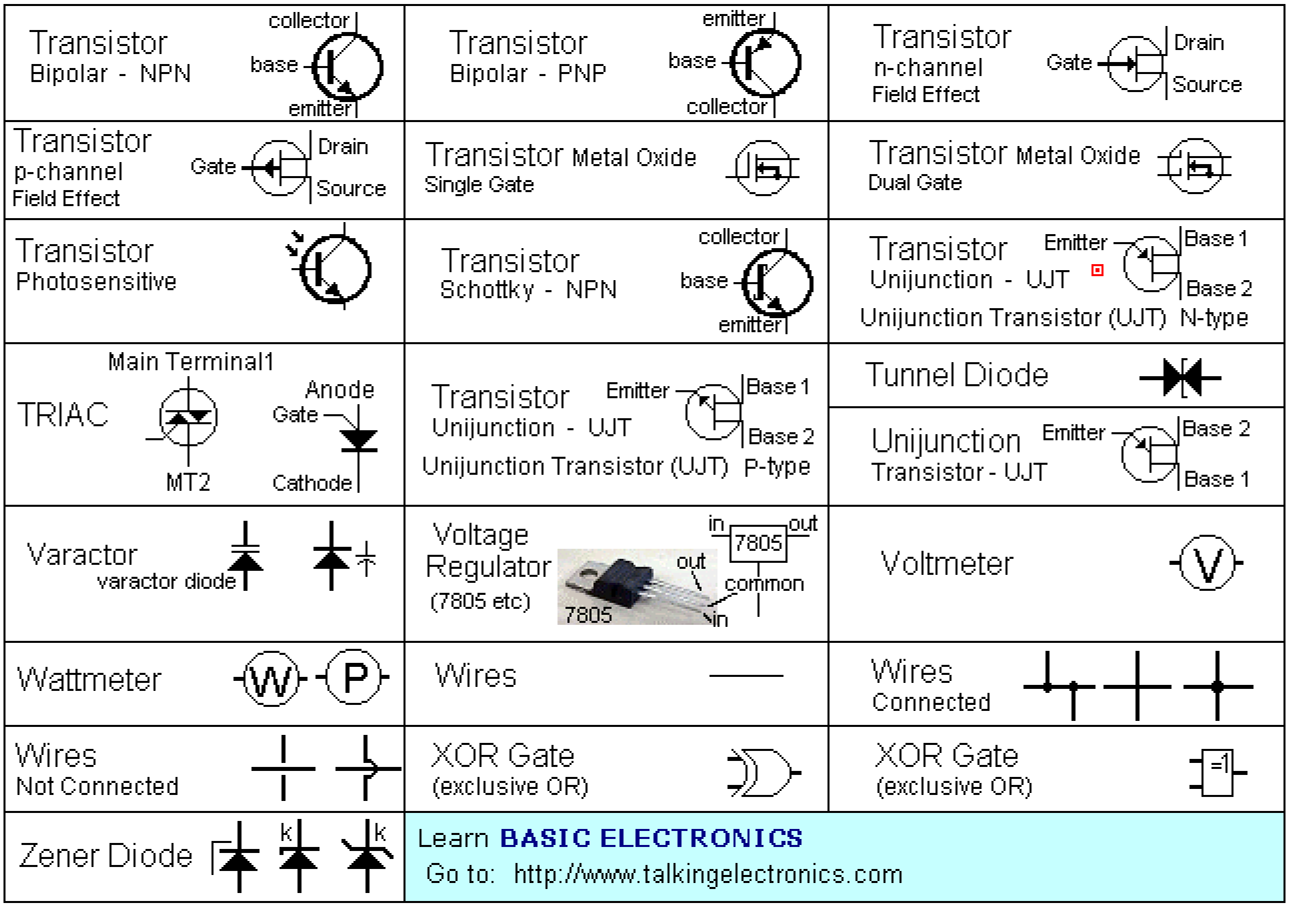

Transistor Bipolar - NPN collector base emitter	Transistor Bipolar - PNP emitter base collector	Transistor n-channel Field Effect Gate Drain Source
Transistor p-channel Field Effect Gate Drain Source	Transistor Metal Oxide Single Gate	Transistor Metal Oxide Dual Gate
Transistor Photosensitive	Transistor Schottky - NPN collector base emitter	Transistor Unijunction - UJT Emitter Base 1 Base 2 Unijunction Transistor (UJT) N-type
TRIAC Main Terminal1 MT2 Anode Gate Cathode	Transistor Unijunction - UJT Emitter Base 1 Base 2 Unijunction Transistor (UJT) P-type	Tunnel Diode Unijunction Transistor - UJT Emitter Base 2 Base 1
Varactor varactor diode	Voltage Regulator (7805 etc) 7805 in 7805 out out common in	Voltmeter V
Wattmeter W P	Wires	Wires Connected
Wires Not Connected	XOR Gate (exclusive OR)	XOR Gate (exclusive OR) =1
Zener Diode k k	Learn **BASIC ELECTRONICS** Go to: http://www.talkingelectronics.com	

More chapters of this eBook on: Talking Electronics.com

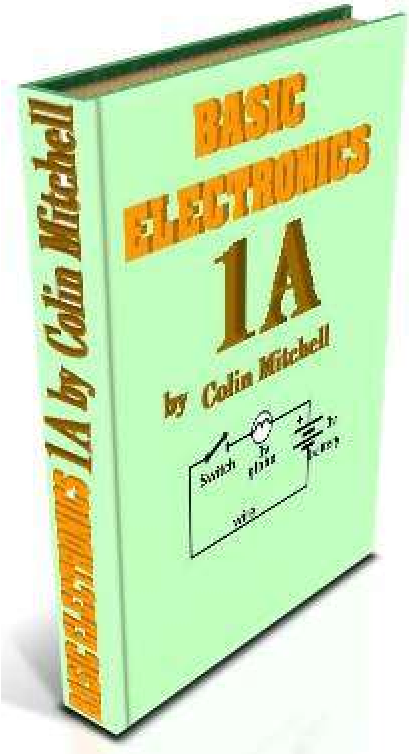
BASIC ELECTRONICS 1A
by Colin Mitchell
BASIC ELECTRONICS 1A by Colin Mitchell
Switch
3v globe
3v battery
wire

For any enquiries email Colin Mitchell

BASIC ELECTRONICS

(this is the Basic Electronics section i.e. Page 1)

(Chapters 1 and 3 are available as .pdf)

Quick Quiz - to see how much you know

Encyclopedia of Components - this is excellent !!!

Page 1: Basic Electronics (this page) - .pdf (1.2MB) or .zip
The capacitor - how it works
The Diode - how the diode works
Circuit Symbols - EVERY Circuit Symbol
Soldering - videos
Page 2: The Transistor
- PNP or NPN Transistor TEST
Page 2a: The 555 IC
The 555 - 1
The 555 - 2
The 555 - 3
The 555 TEST
Page 3: The Power Supply download as .pdf (900kB)
3a: - Constant Current
3b: - Voltage Regulator
3c: - Capacitor-fed Power Supply
Page 4: Digital Electronics

4a: - Gates Touch Switch Gating
4b: - The DELAY CIRCUIT
Page 5: Oscillators
Page 6: Test - Basic Electronics (50 Questions)
Page 7: The Multimeter - using the Multimeter
Page 8: Constructing a Project
Page 9: Inductance

Remember: the animations do not work in .pdf

the site is being constantly updated

INDEX

Active HIGH
AND Gate
Battery
Battery - Internal Resistance
Battery Boost
Battery Booster - for a flat battery
Battery Current
Blocking Diode
Breadboard
Bypass Diode
Capacitors
Capacitors in Parallel
Capacitors in Series - non-polar
Cell
Charging A Capacitor
Capacitor Charging
Capacitor Charging - more details
Capacitors on Volume Control
Capacitor - values
Characteristic Voltage drop - LED
Chassis
Circuit
Circuit - Drawing
Component - Symbols more
Current - milliamp
Current Divider Circuit
Cutoff and Saturation
Damage A LED
DC Current
DC Voltage
Digital States
Diode
Diode - how a diode works
Drawing A Circuit
Earth Rail Earth Rail
NAND Gate
Negative Voltage
NOR Gate
Non-polar Capacitor (electrolytic)
NOT Gate
Ohm's Law
Passive Components
Piezo
Plant Watering Circuit
PNP Transistor
Power and Energy
Power Rail
Potentiometer
Protection Diode
Questions
Regeneration
Resistance - Multimeter
Resistor Colours
Resistors In Parallel
Resistors in Series
Resistor Wattage
Robot Man
Robot Man Animation
Schematic
Slide Switch
Short Circuit
Soldering
Soldering Iron
Speaker
Supply Rail
Switch
Symbols more - components
Tap A Piezo
Test A Cell
Testing A LED

Earth Return
Electret Microphone
Electrolytics
Energy in a cell
Flashing LED Flashing LED
Forward Voltage Drop
Gates with Diodes
Globe
Ground
How a Transistor turns ON
How to Solder
LED
LED flat spot (cathode)
LEDs in Parallel
LEDs in Series
LED Resistors
LED Voltages
Light Dependent Resistor
Load Resistor
Memory Cell
Microphone - Electret
Milliamp - current
Multimeters
Time Delay
Time Delay Animation
Toggle Switch
Tolerance - resistors
Transformer Feedback
3 - Transistor Circuit
Transistor NPN and PNP
Transistor - how it works
Transistor Tester
Turning ON a Transistor
Variable Capacitors
Variable resistor - potentiometer
Voltage
Voltage - Analogue Multimeter
Voltage - Digital Multimeter
Voltage Divider Circuit
Wattage of a Resistor
White LED Voltage
Wire
50 Questions

KIT OF PARTS

Talking Electronics supplies a kit of parts that can be used to build the majority of the circuits in this eBook.
The kit costs $15.00 plus postage.

Kit for Transistor Circuits - $15.00

BUY NOW

A kit of components to make many of the circuits described in this eBook is available for $15.00 plus $7.00 post.
Or email Colin Mitchell: talking@tpg.com.au

The kit contains the following components:
(plus **extra** 30 resistors and 10 capacitors for experimenting), plus:

3 - 47R
5 - 220R
5 - 470R
5 - 1k
5 - 4k7
5 - 10k
2 - 33k
4- 100k
4 - 1M
1 - 10k mini pot
1 - 100k mini pot
2 - 10n
2 - 100n
5 - 10u electrolytics
5- 100u electrolytics

5 - 1N4148 signal diodes
6 - BC547 transistors - NPN - 100mA
2 - BC557 transistors - PNP - 100mA
1 - BC338 transistor - NPN - 800mA
3 - BD679 Darlington transistors - NPN - 4amp
5 - red LEDs
5 - green LEDs
5 - orange LEDs
2 - super-bright WHITE LEDs - 20,000mcd
1 - 3mm flashing LED
1 - mini 8R speaker
1 - mini piezo
1 - LDR (Light Dependent Resistor)
1 - electret microphone
1m - 0.25mm wire
1m - 0.5mm wire
1 - 10mH inductor
1 - push button
5 - tactile push buttons
1 - Experimenter Board (will take 8, 14 and 16 pin chips)
5 - mini Matrix Boards: 7 x 11 hole,
11 x 15 hole, 6 x 40 hole, surface-mount 6 x 40 hole board and others.

Photo of kit of components.
Each batch is slightly different:

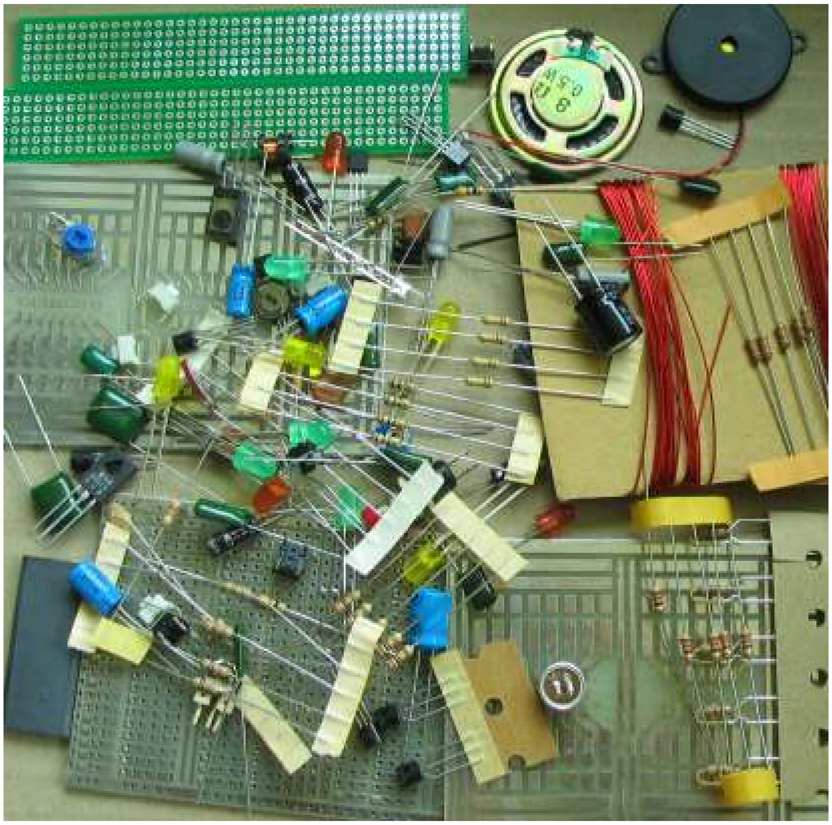

There are more components than you think. . . plus an extra bag of approx 30 components. The 8 little components are switches and the LDR and flashing LED is hiding.
In many cases, a resistor or capacitor not in the kit, can be created by putting two resistors or capacitors in series or parallel or the next higher or lower value can be used.

BEFORE WE START

Too many text books start with the physics of the atom and have equations and mathematics to show how smart the author is.

Don't worry, we wont have any physics or equations.

The reason . . .

This is not a physics course. It is a practical electronics course to teach the basics as quickly as possible. There are no equations because most transistor circuits cannot be worked out mathematically as the gain of a transistor changes according to the current-flow and these gain-values are never provided. So the mathematics is worthless.

To get an answer, all you have to do its build the circuit and measure the values with a multimeter.

Also lots of discussions in text books will never be used in your next 40 years of electronics, so this course doesn't have any unnecessary material and is much-more concentrated than anything you have read before.

Every frame contains important points - especially the animations - as they show you how a circuit works in slow-motion - something that has NEVER been done before.

ELECTRONICS BLOCKS

Here is an idea from Instructables to produce blocks with screws, containing a single component and they can be connected with jumper leads (alligator clips).

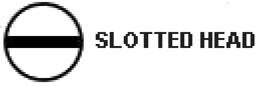

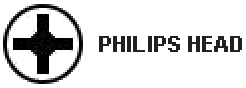

Use a slotted head for the negative screw and a philips head for the positive screw.

Learn electronics from the beginning . . .

START HERE:

All electrical and electronic components need wire to connect them to the circuit.

In a diagram called a CIRCUIT DIAGRAM, the wires are drawn as lines.

When the wires (or lines) cross, they may be joined or just passing.

It is **VERY IMPORTANT** to show the difference between lines that are **JOINED** and lines that are **NOT JOINED**.

When the lines are joined, it is best to place a dot

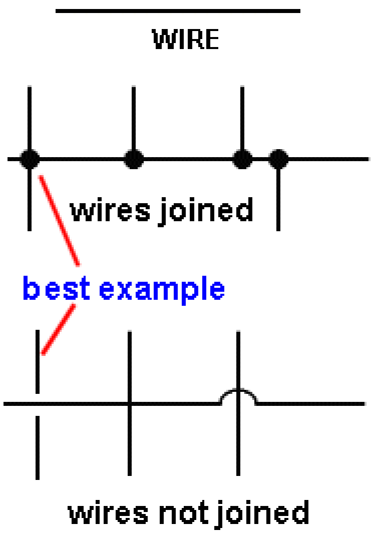

Fig 1 The wire in a circuit diagram

on the connection to PROVE the lines are joined. When the lines are just crossing, a gap should be made so it is obvious that one wire goes under the other and does not touch.
Lines should be "across the page" or "up and down." Very few lines should be at 45°.
You can make a line thicker to indicate a power rail or a wire that will be thick in reality.

to Index

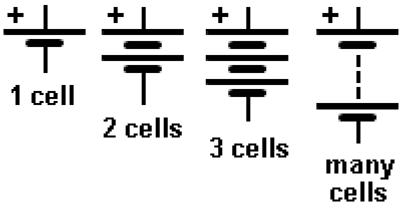

Fig 2 A single Cell and many Cells

Next we need a **battery**. A battery consists of two or more cells. The positive terminal of a battery is the long line in the diagram and you must add the voltage (of the cell or battery) to the symbol as a single cells can be 1.2v, 1.5v, 2.2v or up to 3.6v.
The symbol does not let you know the voltage.
The positive is always at the top and is the longest line on the battery symbol.

to Index

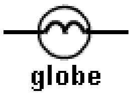

Fig 3: A Globe

Next we need a globe. A globe has two connections (a fine wire inside a glass bulb glows when the globe is connected to a battery).

to Index

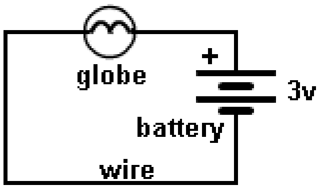

Fig 4: A Circuit

With a globe, battery and wire we have produced a **CIRCUIT**.
A **CIRCUIT** is a complete path and we say the "electricity" the **CURRENT** emerges from the positive of the battery, moves through the globe and returns to the battery via the wire.
If the globe is a "3v GLOBE" it will glow when connected to a 3v battery.
The globe can be connected either way around.
The circuit we have shown is called a **SCHEMATIC** and consists of symbols: a globe symbol and a battery symbol.
The line connecting the two components is called WIRE.

to Index

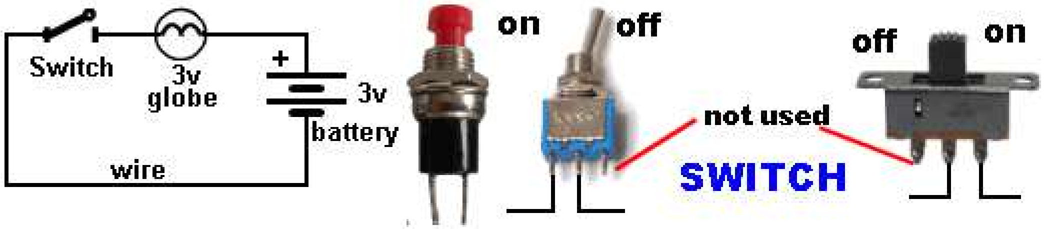

Fig 5:Adding a SWITCH
To turn the globe ON and OFF we need a SWITCH.
The switch may be a push button, a toggle switch (a "click" action) or a slide switch. You could

twist the wires together and untwist them. The result is the same. We say the circuit is "broken" or "open" via the switch and the lamp does not glow. Closing the switch turns ON the globe.

to Index

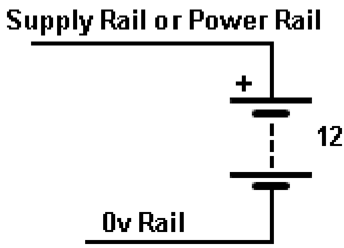

Fig 6: Naming the RAILS

The top rail of the **CIRCUIT DIAGRAM** is called the **SUPPLY RAIL** or **POWER RAIL**.
The lower rail is called the **0v Rail** or **EARTH RAIL**.
Do not connect the Supply rail (+12v) to the 0v rail as this will cause a high current to flow and is called a **SHORT CIRCUIT**:

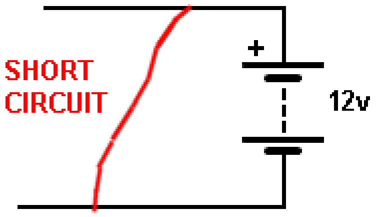

to Index

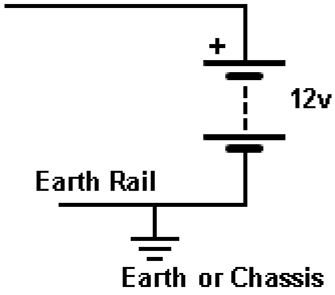

Fig 7: The **EARTH RAIL**

The lower rail is also called the **Chassis**.
This comes from the "old days" when electronics constructors build radios on a metal chassis (metal box) and it was connected via wire to a pipe in the ground to help the radio pick up distant radio stations.
The term also comes from car and truck wiring where one side of each globe is connected to the frame or chassis so that only one wire is needed to each globe and the return "path" is via the chassis.

to Index

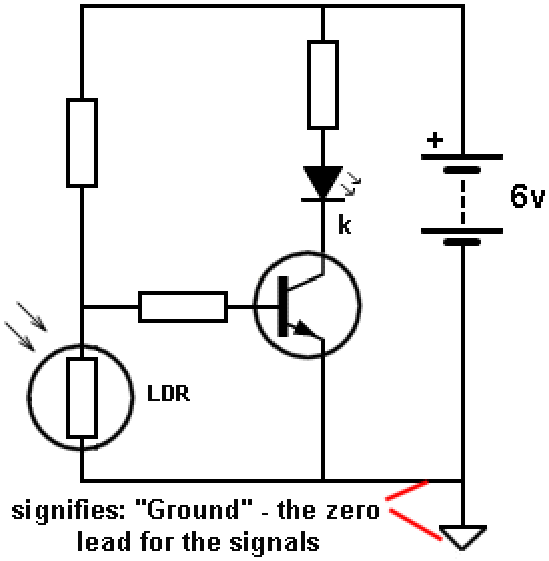

Fig 7a: Ground

Some circuits identify the "**Ground Lead**" or "**Ground Wire**" of a project to show where all the **signals** have been "referenced to." In other words, all the signals rise and fall above and below this "Ground wire" or "Ground Lead." This lead may not be at **earth potential** as the project may be in a plastic box but it identifies where the earth lead of a Cathode Ray Oscilloscope or the negative lead of a multimeter is connected.
On some printed circuit boards, the negative terminal of the battery (the 0v wire or terminal) is connected to a very large area of copper and this is called the EARTH PLANE or GROUND PLANE. It is designed to prevent signals travelling along the tracks (called traces) being radiated and also prevents outside interference upsetting the project. It also "tightens-up" the earth rail.

to Index

The circuit shows a 12v globe connected to a 12v battery and the circuit appears to be "broken" (not continuous).
But the **current returns** via the earth connection.

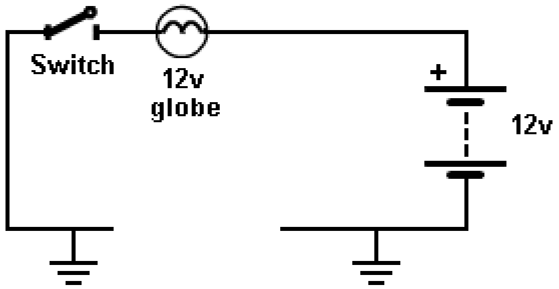

Fig 8: EARTH Return

We talk about the **CURRENT RETURNING**. We don't say: the voltage returning.
The voltage of the globe must be the same as the battery voltage, otherwise it will not glow fully or it will **burn out** if it is say a 6v globe.

to Index

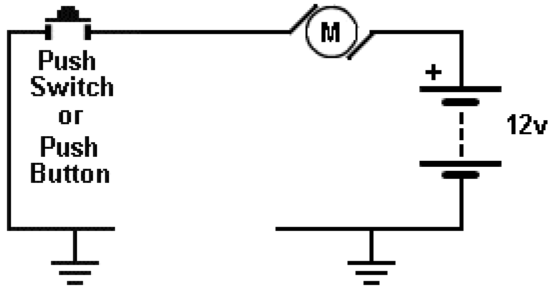

Fig 9: Connecting a Motor

The circuit shows a 12v **Motor**.
It is turned ON when the push-switch is pressed.

to Index

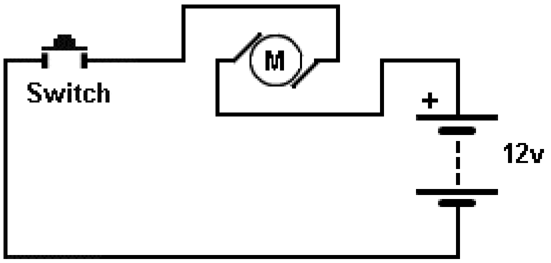

Fig 10: Reversing a Motor

If the wires are connected to the motor "around the other way," the motor will reverse direction.

to Index

VOLTAGE AND CURRENT

What is voltage and what is current?
Here is a very simple description.
A battery produces a voltage called DC. (This is a very confusing name because the letter actually refer to Direct Current, so we just say **DC Voltage**).
A battery also produces current called DC - Direct Current. We say **DC current**.

VOLTAGE

Voltage is a value produced by an electrical component called a battery or cell.
A single cell produces one and a half volts. (1.5v) and although this is not a high voltage, when cells are connected together we get higher voltages.
If 6 cells are connected in series we get 9v.
Here is a 9v battery:

CURRENT

You cannot feel current with your tongue so we have to carry out another experiment:

negative terminal
positive terminal
9 VOLTS
2010
Energizer

Touch the two terminals with your tongue. You get a tingle. This is a 9v tingle. Now you have "felt" electricity. This is a 9v tingle.

9 VOLTS
2010
Energizer

Place a 22 ohm or 47 ohm resistor across the terminals of the battery and hold your fingers on the resistor. It will get hot. This is the result of current flowing through the resistor and heating it up. The current will be about half an amp and the voltage is 9v, so the wattage will be about 2 to 4 watts. Feel the heat produced.

Milli - milli means 1/1,000th (one thousandth) - such as one milliamp or one millivolt.

In other words one thousand milliamps is equal to 1 amp.
One volts is not a very large value as a battery produces 9v and a cell produces 1.5v to 3.6v (depending on the type of cell.
But 1 amp is a large quality when talking about electronic circuits involving LEDs, motors and transistors.
The globe used in the experiments above requires about 300mA. (1,000mA = 1 amp)
The 3v motor used in the experiments requires about 250mA
The LEDs used in the experiments require about 20mA.
Transistors can pass about 100mA to 800mA via the collector-emitter.
In most cases current-flow in the circuits we will be discussing will be less than 1 amp and will be shown as 25mA, 100mA, 350mA etc.

WATTAGE and CAPACITY

A 9v battery has 6 very small cells and they will not last very long.
A "AAA" cell is larger and a "D" cell is much larger.

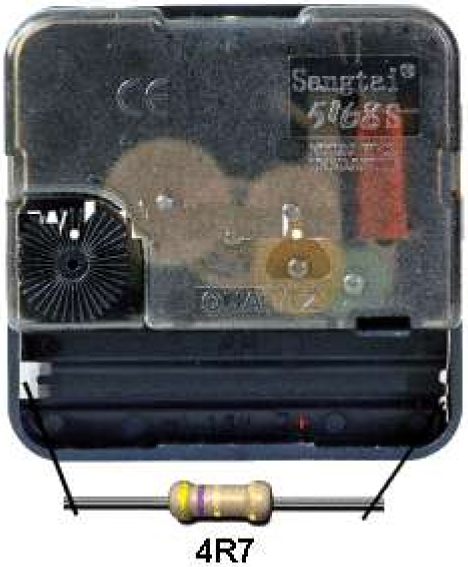

A large cell is said to have a **LARGE CAPACITY.** This means it will deliver a larger current for a longer period of time.
The **WATTAGE** of a cell is the multiplication of the voltage x current. The answer is milliwatts or watts.
The **CAPACITY** of a cell is the wattage x hours. The answer is milliwatt-hours or watt-hours. This is also called watt-hours.
You can determine the capacity of a cell (such as a rechargeable cell) by connecting it to a clock-mechanism that has a 4R7 connected across the terminals. The resistor will take a considerable current and deplete the cell in a few hours. The clock will let you know exactly how long the cell delivered the current. You can then compare other cells.

The simplest electrical circuit consists of a battery and resistor. The current flowing through the circuit will depend on the voltage of the battery and the resistance of the resistor R.

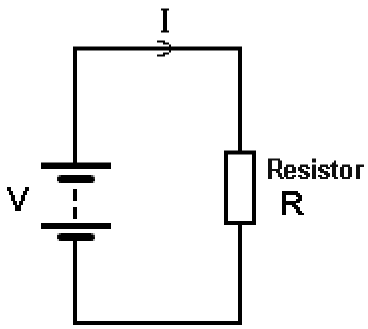

The formula connecting these three quantities is:

$$I = \frac{V}{R}$$

Ohm's Law

$$I = \frac{12}{3}$$

$$I = 4 \text{ amps}$$

This is called **Ohm's Law**. Suppose you have a 12v battery and the resistor is 3 ohms. The current flowing through the resistor will be 4 amps.
Increasing the resistance will decrease the current if the voltage remains fixed.

All the above circuits are called ELECTRICAL CIRCUITS because they contain electrical components (such as a motor, globe, relay, switch).
When the circuit contains an ELECTRONIC component such as a diode, transistor, LED, it is called an ELECTRONIC CIRCUIT or ELECTRONIC SCHEMATIC.

to Index

POWER and ENERGY

Here's an easy way to remember the difference between POWER and ENERGY:
A 9v alkaline battery has enough ENERGY to start a car. But it does not have enough POWER (strength).
Energy is effectively the strength of the battery (and this is the voltage and the current it can deliver) multiplied by the time it can deliver this energy. When the answer is obtained, it consists of three factors ((3 quantities) VOLTS, AMPS and TIME.
This results in an answer called xxxx WATT-HOURS.
For a 9v battery the quantities are: 9 volts, 500mA and the battery will deliver this 9x0.5 = 4.5watts for about 1 hour. This is equal to 4.5 x 60 x 60 = 16,200 watt-seconds.
To start a car requires 250 amps from a 12v battery for 5 seconds.
This is: 12 x 250 x 5 = 15,000 watt-seconds.
This means the energy stored in a 9v battery could start a car if all the energy could be delivered in 5 seconds.
This is not possible however the FACT is this: A 9v battery has enough stored energy to START A CAR.

to Index

300
POWER STATION

BATTERY BOOSTER

One of the simplest things we can do is start a car with a flat battery with the assistance of a BATTERY BOOSTER.
This consists of a 12v rechargeable battery in a handy case with leads to connect to the flat battery in your car.
This simple operation puts two 12v batteries in parallel, but no-one has actually described what happens and why.
That's because the explanation is very complex. We have included it here to show that a simple explanation involves a lot of technical terms and you will understand more after reading the course.

The flat battery in the car is not fully charged but it has some percentage of charge and when it sits for a period of time in a non-fully charged condition, the voltage drops from 12.6v to less than 12v as the battery gradually self-discharges due to the potential at the top of the cell being different to that at the bottom of the cell and the specific gravity of the electrolyte being different at the top and bottom. This causes an internal current to flow within the cell and slowly discharge the cell. But if you try to start the car, the voltage drops to less than 7v because the electrolyte cannot carry the high current and a slight potential is developed across the liquid.

The result is the starter-motor does not crank the car.
The reason is this: When the battery is fully charged, the current taken by the starter motor is about 300 amps. This is about 11v x 300 amps = 3300 watts = 4.4Horsepower.
But when the voltage drops to 7v, the current will drop to 190 amps to deliver 1336 watts = 1.8HP. This is only 40% of normal and that's why the car does not start. The engine needs 4HP to overcome the pressure in the cylinders due to the compression of the air during the "firing stroke."
Let's put it this way. If we have a brand new 7v battery, the car will not start. The starter-motor will only accept 190 amps when the supply is 7v.
So, we have to increase the voltage.
We do this by placing a 12v battery across the flat battery. The voltage of the flat battery will immediately rise to 12.6v. It might take 2 minutes but the flat battery will take a small current (1 to 10amps) from the battery in the "booster" and the output of the combination will be 12.6v. The current-carrying capacity of the electrolyte will improve very quickly and you have effectively given the "flat battery" a very quick charge.
The starter-motor will now accept 300 amps from the combination and **SURPRISINGLY** the cells of the "flat battery" will deliver about 200 amps and the booster battery will deliver about 100 amps. The actual sharing of current will depend on the two batteries but the secret behind the success is the increase in voltage we call **TERMINAL VOLTAGE**. The voltage on the terminals (the alligator clips).
The capacity of the booster battery is not important. It can be from 7AHr to 40AHr. We are just using a very small amount of its capacity to start the car and nearly all batteries will provide 200 Amps for a short period of time.
The voltage of the car battery is very important. The Horsepower taken by the starter-motor is defined by the formula: Pwatts = V^2/R Since the resistance remains constant, a voltage of 7 volts will produce 7x7=49 units and a voltage of 11v will produce 121units. This gives the ratio of 40% to 100% as explained above.

to Index

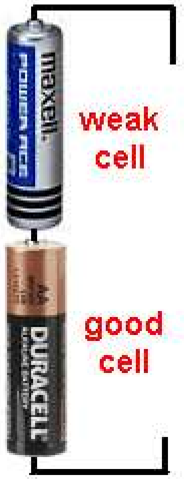

BATTERY BOOST

Continuing from the previous frame where we showed the effect of placing a weak battery in parallel with a good battery, we can show what happens when a weak cell is placed IN SERIES with a good cell.
This also applies when you have 5 good cells and one weak cell.
Basically, the weak cell will reduce the current. In other words, if the 5 cells are driving a motor and supplying 250mA, the 5 cells and 1 weak cell will deliver 200mA or less, depending if it is weak or very weak. The current flowing through the weak cell will have the effect of giving it a small charge - in other words, you will be charging the weak cell from the good cells when the motor is operating.

BUT . . . .

There is a way to use weak cells. If you have say 6 weak cells driving a motor and the RPM is reducing, you can add 2 more weak cells to increase the RPM.
The effect is this: The voltage from the 8 cells will be higher than from 6 cells and this will allow a higher current to flow. Sometimes the cells will provide this higher current and thus more of the energy will be delivered and you will get the last of the energy from the cells.

to Index

INTERNAL RESISTANCE

All batteries and also all individual cells have a "secret, hidden" value of resistance inside each cell due to the resistance of the chemicals. This resistance is very small when the cell is new but it increases as the cell gets older.
It is very easy to measure this value. Simply put an ammeter directly across the cell and measure the current. Use Ohm's law to work out the resistance. But this not always a wise thing to do as some cells will deliver 10 amps and some will deliver 100 amps and damage the meter.
The diagram opposite shows a large internal resistance for the weak cell and a small internal resistance for the good cell.
If a cell did not have any INTERNAL RESISTANCE it would deliver thousands of amps. It's the Internal Resistance that limits the current.

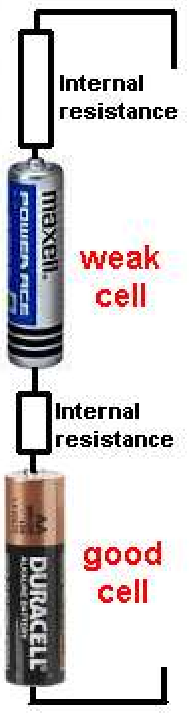

In most cases we neglect (do not consider) the value of internal resistance when making tests and when using a battery in a project.
But when a battery gets old, it cannot deliver a high current and the internal resistance gets so high that the output voltage drops from say 9v to 7v, even when the battery is not connected to a circuit.
This is the result of the INTERNAL RESISTANCE of the chemicals increasing to a point where they become noticeable and what we call "poisoning" of the chemicals due to the cell "aging" and new chemicals being produced in the cell that have a high resistance. Some of the terms we use are: "drying out and sulphating. Some cells produce spikes or needles that completely short-circuit the cell and make it totally useless.

to Index

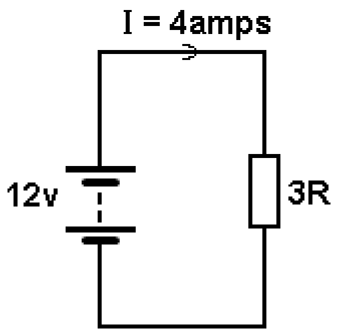

This resistor is dissipating 48 watts.
As a comparison, a soldering iron is dissipating about 10 to 20 watts.

RESISTOR WATTAGE

Resistor Wattage means two things.
1. The physical size of a resistor tells you number of watts it is capable of dissipating. This is called RESISTOR WATTAGE. It is really RESISTOR-SIZE or RESISTOR-CAPABILITY.
2. The multiplication of the voltage across a resistor and the current flowing though it will produce a value called WATTAGE. This is also called RESISTOR-WATTAGE or RESISTOR-LOSS or RESISTOR-DISSIPATION or HEAT-LOSS.
In the circuit shown, the wattage being lost in the resistor is:
12 x 4 = 48 watts.
Most of the resistors we will be using in our projects are 0.25watts. This means they will dissipate 250milliwatts, however the actual wattage being dissipated may be only 70 milliwatts and the resistor will not get hot.
0.25watts is the maximum wattage it can dissipate without overheating.
If it is dissipating 400milliwatts, it will be VERY HOT.
The wattage it is dissipating (the heat it is getting rid of) will depend on the supply voltage and the value of the surrounding components.

to Index

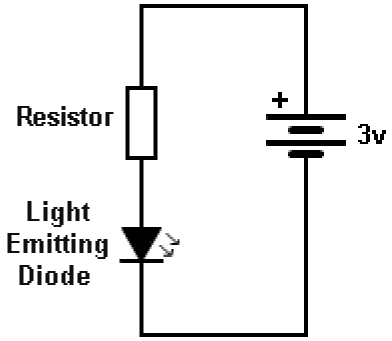

Fig 11. **The LED**

This simple **ELECTRONIC CIRCUIT** contains a **LIGHT EMITTING DIODE (LED), RESISTOR** and battery.

The circuit is classified as electronic because the LED is not an electrical item (such as a globe) but more-complex, as it produces light when current flows through a crystal and the crystal produces the colour.

to Index

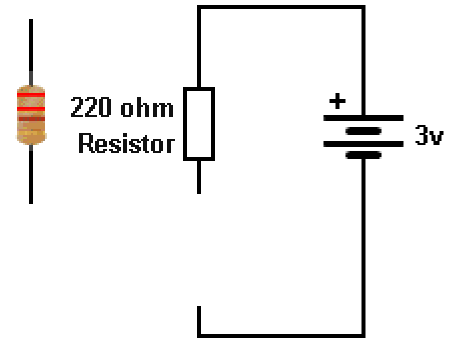

Fig 12. **The Resistor**

A **RESISTOR** must be included in the circuit to prevent the LED being damaged.
The resistor in this circuit must be 220 ohms.
This is shown by the colours on the resistor. The colours for 220 ohms: **red - red - brown**. The 4th band is gold - indicating a tolerance of 5%.
A resistor has **RESISTANCE**.
It reduces the current from the battery to a required amount to prevent the LED glowing too bright.
A resistor is just like putting your foot on a hose. The water trickles out the end. The resistor "resists" the high current-flow that the battery is able to deliver.

to Index

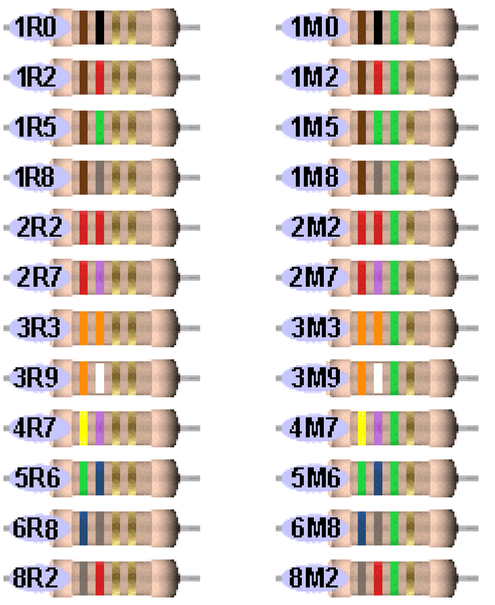

Fig 12a. The **Resistor Colours**

There are hundreds of different resistors because the resistance-values need to cover the range one ohm to 10 million ohms.
There are also small, medium and large resistors.
The resistors on the left are just a few in the range. (See the full range below). They show colour bands for 1 ohm to 8.2 ohms and 1 million ohms to 8.2 million ohms. All the other values are shown below.
An electronics engineer does not have the room to store 10 million different resistors so they make each resistor 5% or 10% higher than the previous. This reduces the number to about 100 to 200.

TOLERANCE

The first 3 bands indicate the value of the resistor and the 4th band indicates either 5% or 10% tolerance.
All modern resistors are 5% or 2% or 1%. The "old" 10% resistors are no longer made.
Gold = 5%
Silver = 10%

to Index

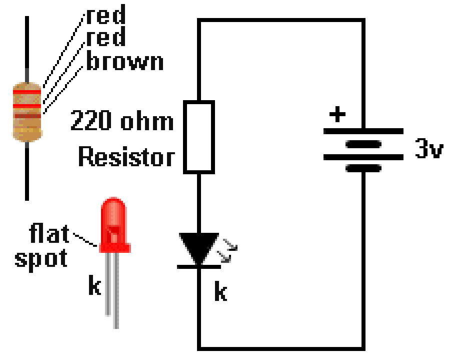

Fig 13. The **Resistor and LED**

The **LIGHT EMITTING DIODE** is called an electronic component (mainly because it is more complex than a globe and it produces light by a more-complex means than heating a wire).
A **LED** must be connected around the correct way.
It will not illuminate if connected around the wrong way.
All **LEDs** have one lead longer than the other. The SHORT lead is called the **CATHODE** (k).
All LEDs have a flat on one side and this is the **CATHODE** lead.
The arrows on the diagram indicate light is "given off" (emitted - produced).

to Index

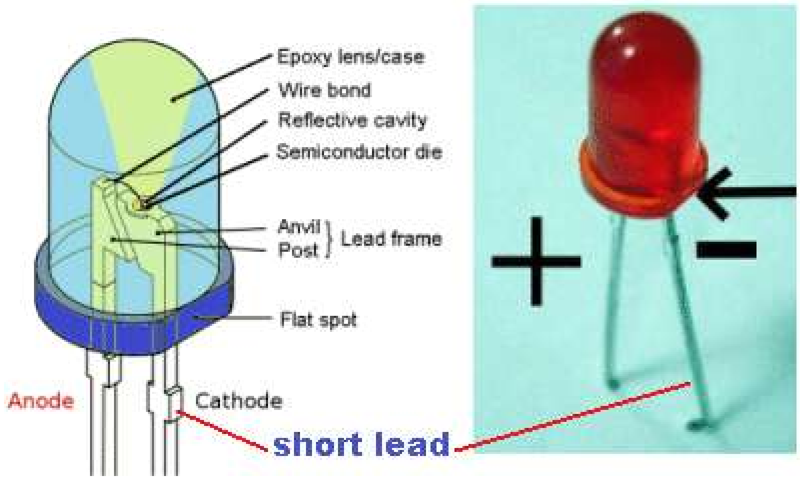
Epoxy lens/case
Wire bond
Reflective cavity
Semiconductor die
Anvil
Post
Lead frame
Flat spot
Anode
Cathode
short lead
+
-

Fig 14. The **LED - showing the flat spot**
A close-up of a red LED. The cathode lead is the short lead and next to a flat side on the LED.
DO NOT show "+" or "-" on a diagram. Only show the letter "k" to indicate cathode.
The symbols "+" and "-" are used when a component produces a voltage or is connected directly to "+" and "-" A LED is connected via a resistor.

to Index

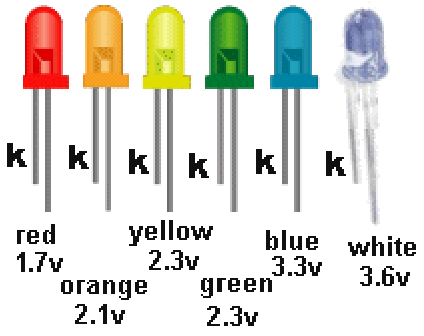
k
k
k
k
k
k
red
1.7v
yellow
2.3v
blue
3.3v
white
3.6v
orange
2.1v
green
2.3v

Fig 15. LED VOLTAGES

When a LED is connected to a circuit, (and the correct-value resistor is included), a voltage will be develop across the LED called the **CHARACTERISTIC VOLTAGE DROP**.
This voltage is due to the colour of the LED and the crystal inside the LED that produces the colour.
The diagram on the left shows the approximate voltage developed for each LED.
The voltage **does not change** for small, medium, surface-mount, or large LEDs.

to Index

When a LED is connected to 3v battery, the following **CHARACTERISTIC VOLTAGE DROPs** will develop across each LED.
You will notice we **have not changed** the value of the resistor. It is 220R.
The LED creates the voltage and if the value of resistance is decreased, the LED will illuminate BRIGHTER.
If the LED illuminates too bright it will be DAMAGED.

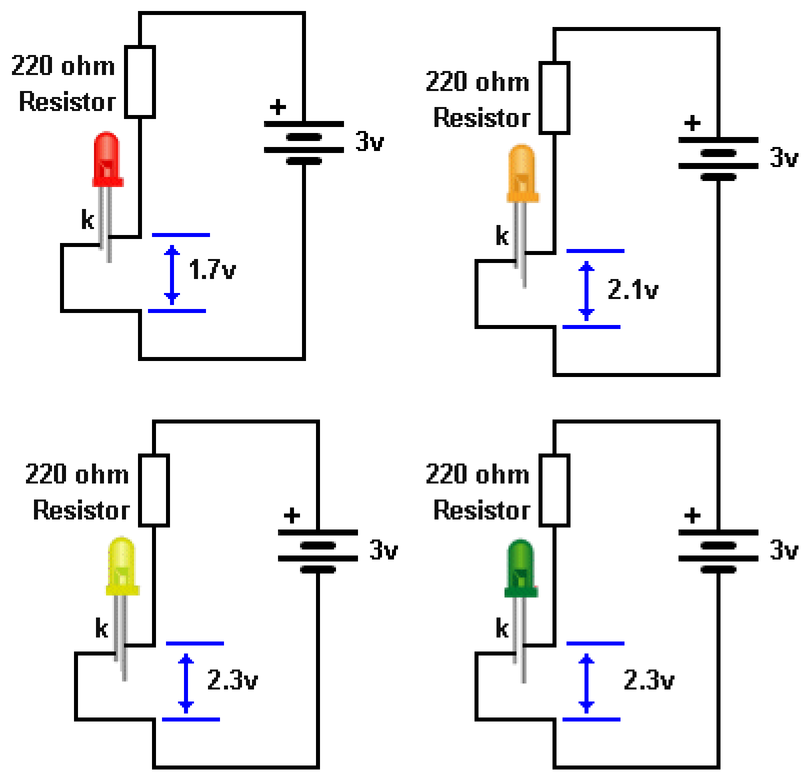
220 ohm
Resistor
+
3v
k
1.7v
220 ohm
Resistor
+
3v
k
2.1v
220 ohm
Resistor
+
3v
k
2.3v
220 ohm
Resistor
+
3v
k
2.3v

Fig 16. LED VOLTAGES

to Index

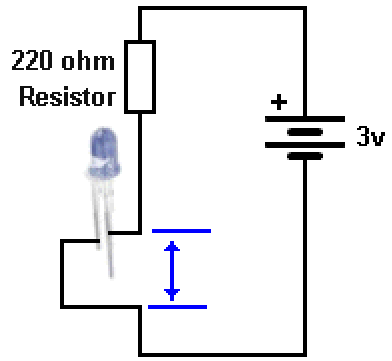
220 ohm
Resistor
+
3v

Fig 17. WHITE LED VOLTAGE

If we connect a WHITE LED to 3v supply, it will not illuminate because it needs a supply higher than 3.6v.

The resistor in series with the LED is called a **CURRENT LIMITING RESISTOR**.

In this circuit **no current flows** because the supply is not high enough.

to Index

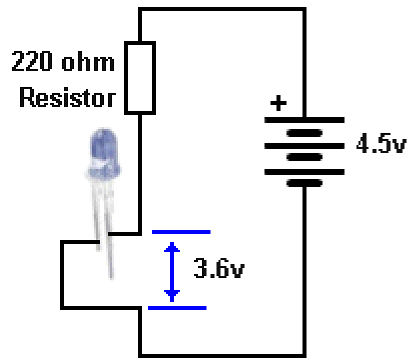
220 ohm
Resistor
+
4.5v
3.6v

Fig 18. WHITE LED VOLTAGE

When the supply is increased to 4.5v, the 220R resistor will allow a current to flow through the white LED and it will develop a CHARACTERISTIC VOLTAGE DROP of 3.6v across it.

The supply (the voltage of the battery) must be higher than the CHARACTERISTIC VOLTAGE DROP of the LED so the resistor will allow the correct amount of current to flow.
The ideal current for a LED is 20mA, however some LEDs will work when 1mA flows, so you have to know what you are doing.

to Index

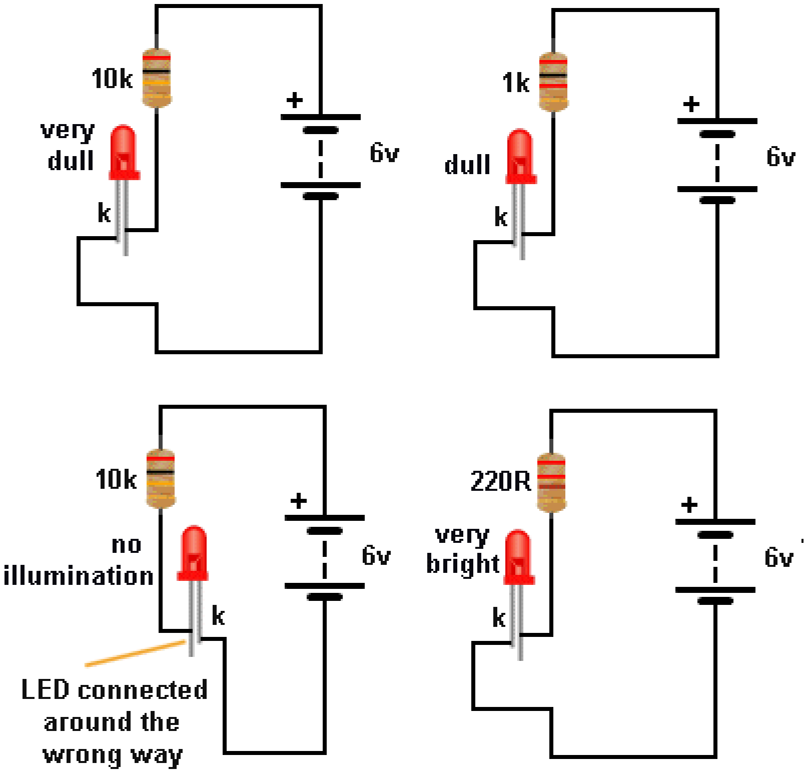
10k
+
very
dull
6v
k
1k
+
dull
6v
k
10k
+
no
illumination
6v
k
LED connected
around the
wrong way
220R
+
very
bright
6v
k

Fig 19. Testing A LED
Now connect either the 1k, 470R or 220R and determine the brightness you need.
As the brightness increases, the current will be higher.
You can use 3v supply for all LEDs except blue and white.

HOW TO TEST A LED
Some clear LEDs produce red or orange and some LEDs do not have the cathode lead clearly identified.
Here's how to find the colour, cathode lead and the current.
You need a 6v battery, 10k resistor, 1k resistor, 470R resistor and 220R resistor.
Connect the 6v battery and 10k resistor to the LED and it will only illuminate when the cathode is connected to the negative of the battery. This is the short lead.

to Index

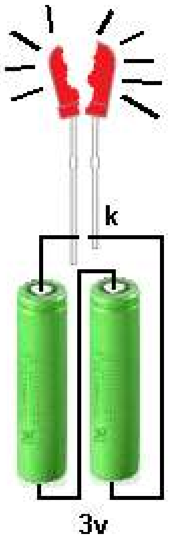

Fig 20. Damage a LED

Do not connect a 3v battery directly across a LED.
It will be DAMAGED. You MUST include a resistor.

to Index

A LED IS CURRENT DRIVEN

You may have seen this statement and tried to work out what it means.
Basically it means an increase in current will make the LED brighter.
But a LED needs 2 things:
It needs a voltage that is EXACTLY the voltage required to produce illumination. And this voltage depends on the colour of the LED.
As soon as you supply the exact voltage, the crystal will begin to glow and as you increase the current, the illumination will increase.
But doing this is VERY VERY difficult.
It is very easy to supply an exact voltage such as 1.7v or 3.4v, but delivering a current such as 10mA or 20mA at the same time is very difficult. You cannot get a 1.7v battery and deliver 10mA to a LED.
As we have shown above, you need a simple components such as a resistor between the battery and LED to achieve the desired result.
A LED is CURRENT DRIVEN but firstly you need to provide a VOLTAGE that is exactly the connect value for the colour of the LED and then the current can be increased.

to Index

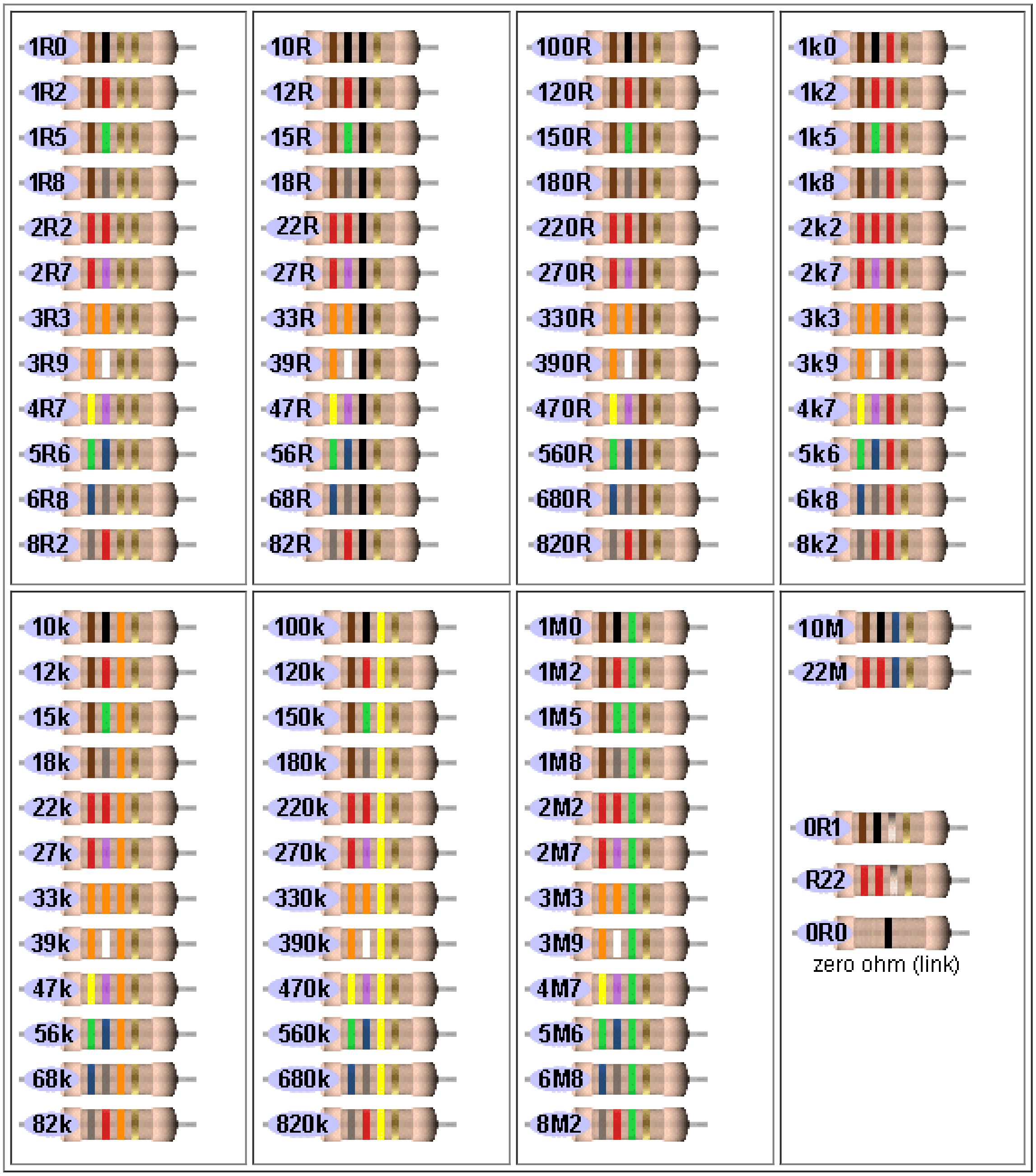
1R0
1R2
1R5
1R8
2R2
2R7
3R3
3R9
4R7
5R6
6R8
8R2
10R
12R
15R
18R
22R
27R
33R
39R
47R
56R
68R
82R
100R
120R
150R
180R
220R
270R
330R
390R
470R
560R
680R
820R
1k0
1k2
1k5
1k8
2k2
2k7
3k3
3k9
4k7
5k6
6k8
8k2
10k
12k
15k
18k
22k
27k
33k
39k
47k
56k
68k
82k
100k
120k
150k
180k
220k
270k
330k
390k
470k
560k
680k
820k
1M0
1M2
1M5
1M8
2M2
2M7
3M3
3M9
4M7
5M6
6M8
8M2
10M
22M
0R1
R22
0R0
zero ohm (link)

Fig 21. All the resistor values

Here are all the colours and values for the resistors you will using in this course. Just match-up the colours on your resistor with the resistors above and you will find the value.

to Index

Resistor values are always OHM values. One ohm is a small value. It might be the resistance of a length of wire 3 metres long.
When a switch is open the resistance is infinite - millions and millions of ohms.
The resistance of your body from one hand to the other will be about 70,000 ohms.
The resistance between two wires dipped in water will be about 1,000 to 100,000 ohms (depending on the dissolved-salts in the water - pure water has a very high resistance)
The resistance of the filament of a 3v globe will be about 30 ohms.
The resistance of the winding of a 3v motor will be about 3 ohms.
Resistors are made with values from less than one ohm to more than 10 million ohms by adding carbon to the mixture inside the resistor (and cutting a track around the outside of the resistor) then connecting a lead to each end. Adding more carbon reduces the value of resistance. Carbon has a low resistance.
Resistance-values are measured with the RESISTANCE settings on a MULTIMETER.
This is called the "Ohms Range." Sometimes with the symbol: **Ω**
A Multimeter will have 2, 3 4 or more scales to cover the range one ohm to 10 million ohms.
Low value resistors (from 1 ohm to 999 ohms) are written as 1R, 220R, 470R, 999R. with the letter "R" indicating Resistance (ohms). You can also use the symbol "omega" **(Ω)**
For values above 1,000 ohms to 99,999 ohms, they are written as: 1k, 2k2, 4k7, 10k, 100k, 220k, 470k, with the letter "k" indicating "kilo" (thousand).
1M = 1,000,000 - one million ohms 1M2, 2M2, 4M7, 10M.
The letters "R, k and M" are placed so they take the place of the decimal point. This prevents any mistake, as a decimal point can be missing in a poor photocopy.

to Index

MULTIMETERS

There are two types of **MULTIMETER**. The top two are called **DIGITAL MULTIMETERS** (DMM) and show numbers on a display.
The lower two meters are called **ANALOGUE MULTIMETERS** and have a pointer and scale.
All meters come with a set of red and black leads.

The **red lead** is always connected to the positive of the battery or the positive on a project and the **black lead** is connected to the negative or earth or chassis.
When making a resistance measurement, the leads can be around either way.
Resistance measurements are always made with the power removed from a circuit. Any voltage on a circuit will upset the resistance reading.

to Index

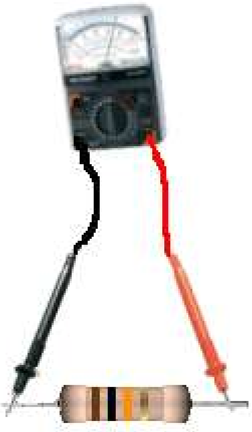

Fig 22. Resistance Measurement with Analogue Multimeter

The resistance of a resistor is measured by placing the leads of the multimeter on the ends of a resistor and turning the dial on the analogue multimeter to the resistance scale to make the pointer move to about the centre of the scale.
The resistance scale is marked with a high value on the left and 0 ohm on the right. This is opposite to all the other scales.
You must get the pointer to move to the middle of the scale as it is not accurate at left-end.
Analogue multimeters are only suitable for reading values from 1 ohm to 100,000 ohms. The scale is too hard to read above 100k.
To find the value of a resistor, you can compare the colours with the table above.

to Index

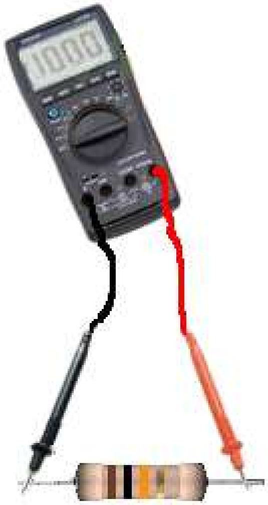

Fig 23. Resistance Measurement with a DMM

A digital multimeter produces a more-accurate reading of resistance.
It is accurate from 1 ohm to 10M ohms.
Select the scale that provides a reading.

to Index

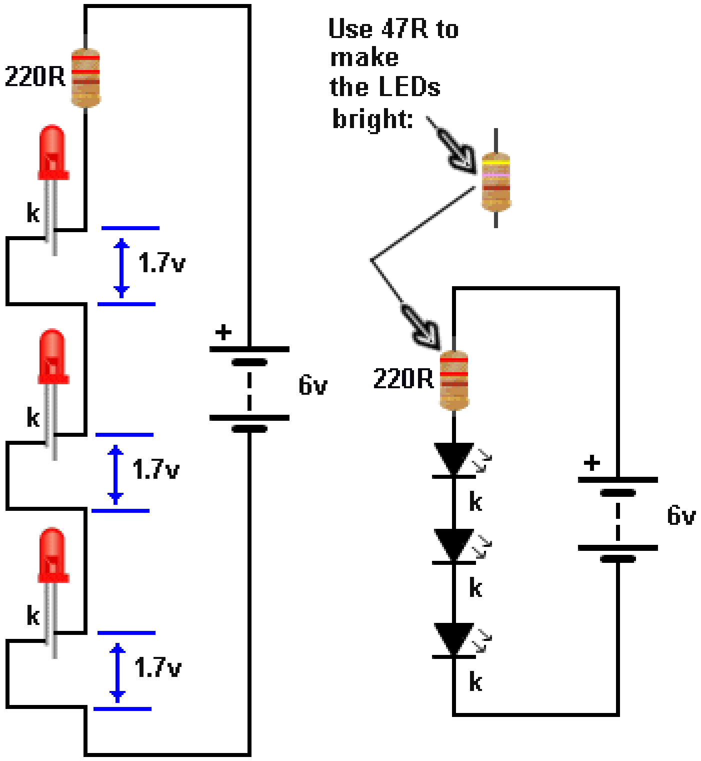
220R
k
1.7v
k
1.7v
k
1.7v
+
6v
Use 47R to make the LEDs bright:
220R
k
k
k
+
6v

Fig 24. Connecting **LEDs in series**
LEDs can be placed in series provided the total **CHARACTERISTIC VOLTAGE DROP** across the LEDs is LESS than the supply voltage.
In this case the voltage across the LEDs is 1.7v + 1.7v + 1.7v = 5.1v
The supply is 6v and this allows 0.9v for the CURRENT LIMITING RESISTOR.
The LEDs will not be very bright with 220R.
Change the resistor to 47R
If you connect 4 LEDs in series, the total **CHARACTERISTIC VOLTAGE DROP** will be 6.8v and no LEDs will illuminate because the total is higher than the 6v supply.

to Index

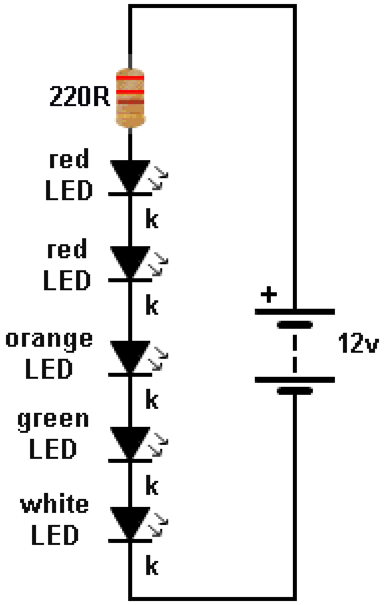
220R
red LED
k
red LED
k
orange LED
k
green LED
k
white LED
k
+
12v

Fig 25. Connecting LEDs in series

Different-colour LEDs can be connected in series. Add up the total Characteristic Voltage for the 5 LEDs and see if it is less than 12v.

The 220R resistor will have to be reduced to 47R to make the LEDs bright.

to Index

LEDs can be connected in parallel if they are the

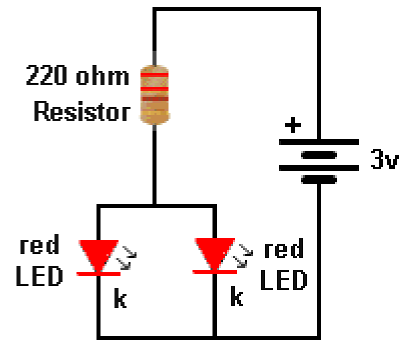

Fig 26. Connecting LEDs in parallel

same colour.
In the diagram a red LED drops a CHARACTERISTIC VOLTAGE of 1.7v and if they are from the same manufacturer or the same batch, they will work ok.
Although we say the characteristic voltage for a red LED is 1.7v, this can change slightly from different manufacturers and one LED may glow brightly while the other is dull.
You have to build the circuit and see the result.

to Index

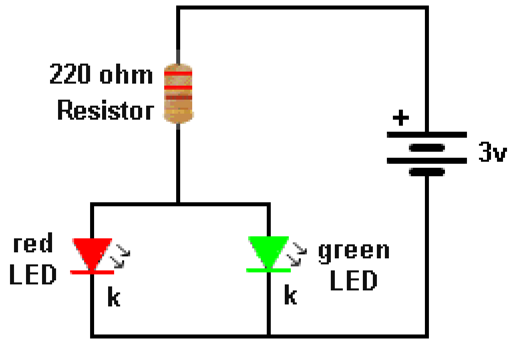

Fig 27. Connecting LEDs in parallel

Different colour LEDs cannot be connected in parallel. The voltage across a red LED is 1.7v. This becomes the "Supply Voltage" for the green LED and it is too low. The green LED needs a supply of 2.1v to 2.3v.
Only the red LED will illuminate.

to Index

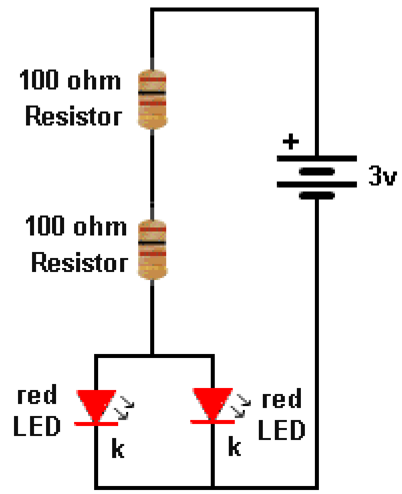

Fig 28. Connecting Resistors in Series

Suppose you don't have a 220 ohm resistor.
You can make a 220 ohm resistor with two resistors in series. The total resistance will be 200 ohms, but resistors are not accurate and the result will be very close to 220R.
Electronic circuits are not very critical. You will not be able to see the difference in brightness between 200 ohms and 220 ohms.

When resistors are connected in series, the total resistance is found by adding the resistance of each resistor.

$$R_{total} = R_1 + R_2 + R_3 + \dots$$

to Index

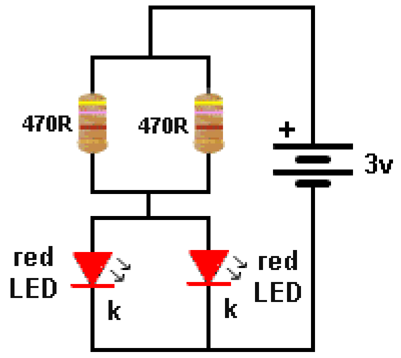

You can create a 220 ohm resistor by connecting two resistors in Parallel.
When two equal-value resistors are connected in Parallel, the total resistance across the combination is HALF.
470R in parallel with 470R produces 235R.
This is very close to 220R.
We are not going into the formula as it is very complex.
Three equal-value resistors in parallel produce a total of **one-third.**
Simply get two resistors and connect them in

Fig 29. Connecting Resistors in Parallel

parallel and measure them with a multimeter.

to Index

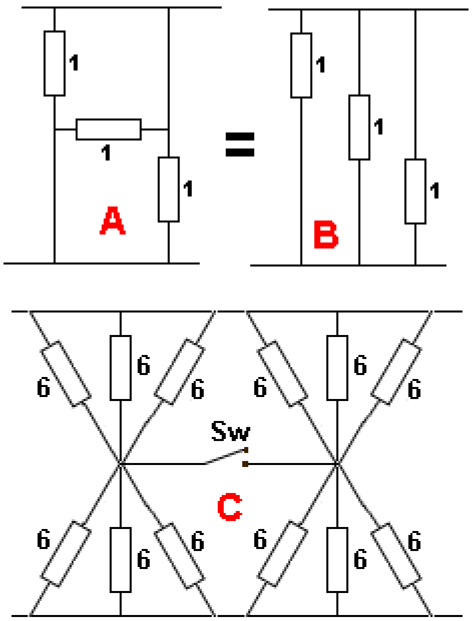

Fig 29a. Two tricky resistor questions

Figure **A** shows three resistors. It looks hard to solve so the middle resistor is turned so it connects directly to the top and bottom rail. Now you can see the circuit is three resistors in parallel. The result is one-third of an ohm.

Figure **C** shows twelve 6 ohm resistors. Replace each group with a 2 ohm resistor, because three 6 ohm resistors in parallel is equal to 2 ohms.

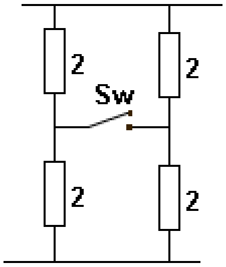

The two left resistors create 4 ohms and the two right resistors create 4 ohms. The result of two 4 ohm resistors in parallel is 2 ohms. The resistance of the circuit does not change if the switch is open or closed.

to Index

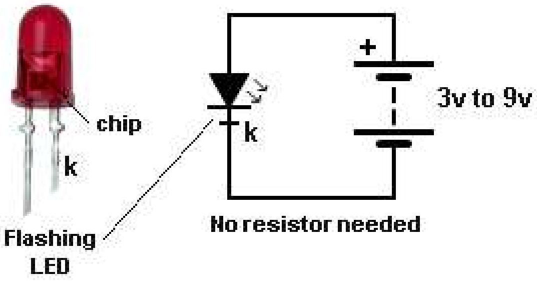

Fig 30. Flashing LED

There are some special LEDs that can be connected to 3v to 9v and they flash or produce a range of colours.
These LEDs have a chip and resistor inside the body of the LED to produce the effect and allow the LED to operate on a voltage without the need for a current limiting resistor.

to Index

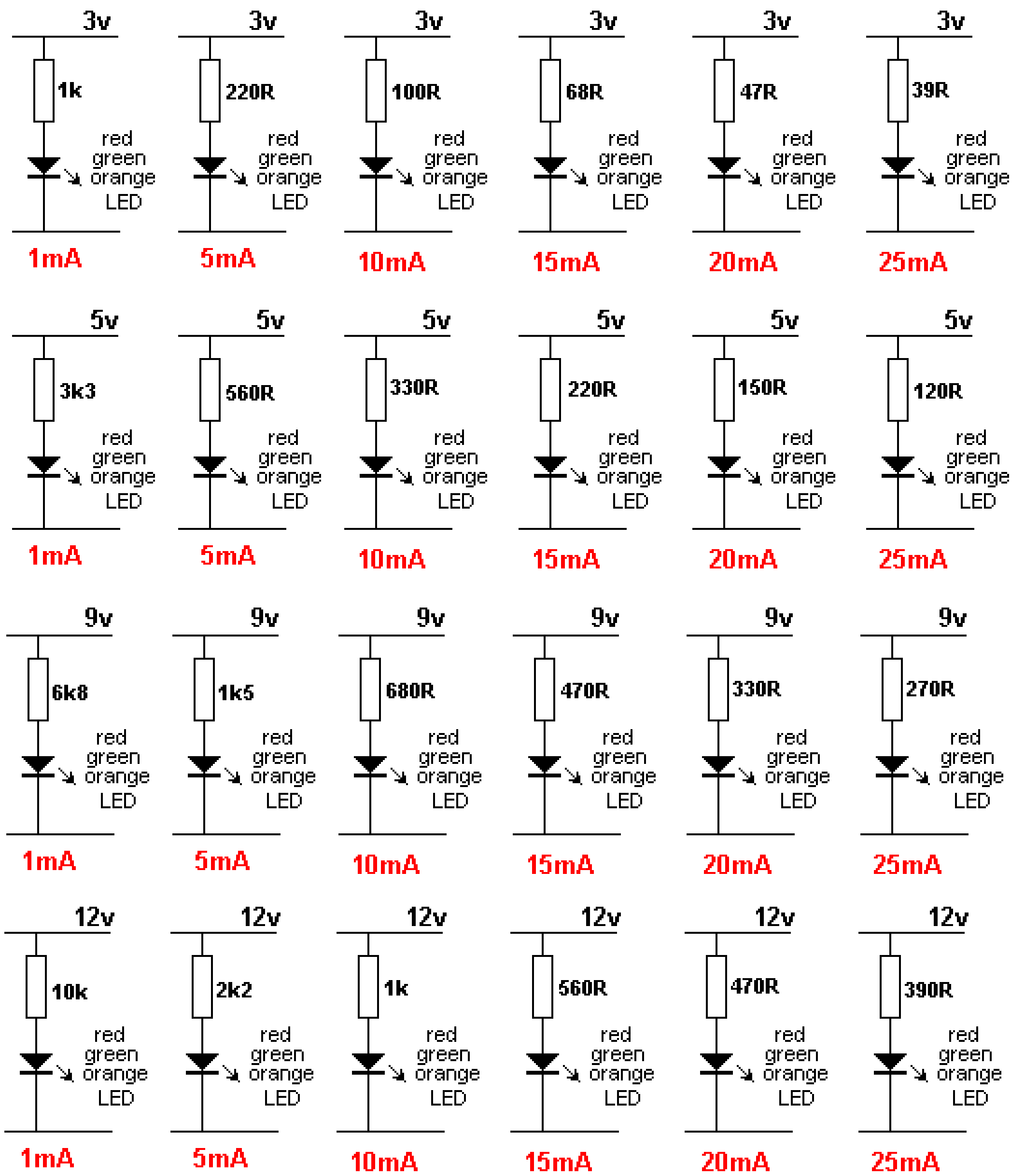
3v
1k
red
green
orange
LED
1mA
3v
220R
red
green
orange
LED
5mA
3v
100R
red
green
orange
LED
10mA
3v
68R
red
green
orange
LED
15mA
3v
47R
red
green
orange
LED
20mA
3v
39R
red
green
orange
LED
25mA
5v
3k3
red
green
orange
LED
1mA
5v
560R
red
green
orange
LED
5mA
5v
330R
red
green
orange
LED
10mA
5v
220R
red
green
orange
LED
15mA
5v
150R
red
green
orange
LED
20mA
5v
120R
red
green
orange
LED
25mA
9v
6k8
red
green
orange
LED
1mA
9v
1k5
red
green
orange
LED
5mA
9v
680R
red
green
orange
LED
10mA
9v
470R
red
green
orange
LED
15mA
9v
330R
red
green
orange
LED
20mA
9v
270R
red
green
orange
LED
25mA
12v
10k
red
green
orange
LED
1mA
12v
2k2
red
green
orange
LED
5mA
12v
1k
red
green
orange
LED
10mA
12v
560R
red
green
orange
LED
15mA
12v
470R
red
green
orange
LED
20mA
12v
390R
red
green
orange
LED
25mA

Fig 31. These diagrams show the resistor needed to produce 1mA to 25mA current through a single LED on 3v, 5v, 9v and 12v supply.

to Index

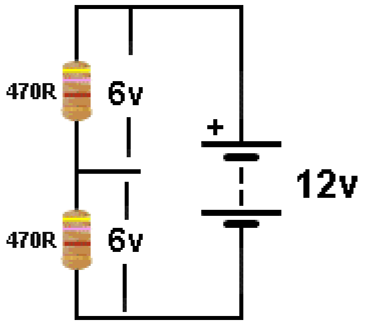
470R
6v
+
12v
470R
6v

Fig 31a. Voltage Divider Circuit.

THE VOLTAGE DIVIDER

In the circuits above, the resistor and LED are forming a VOLTAGE DIVIDER.
A red LED is dropping 1.7v across it and the resistor is dropping the remaining voltage.
Whenever two (or more) components are placed across a battery, they form a **VOLTAGE DIVIDER**.
Sometimes we want a 6v supply and only have 12v.
We can produce the 6v supply by putting two equal-value resistors across the 12v as shown in the circuit opposite.
We are not going into the mathematics because the selection of the correct value is very complex and the circuit is very wasteful.

to Index

THE CURRENT DIVIDER

The **CURRENT DIVIDER CIRCUIT** is actually a

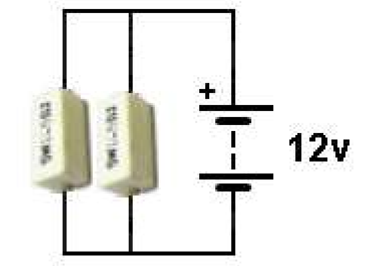

Fig 31b. Current Divider Circuit.

LOAD SHARING CIRCUIT.
Suppose you are testing a Power Supply and need a 10 watt LOAD. But you only have 5 watt resistors.
Placing two 5watt resistors in parallel across the output of the power supply will allow half the current to flow though each resistor. This is called **CURRENT SHARING** or **LOAD SHARING** and the current is divided (or passed) through each resistor according to the value of resistance.

to Index

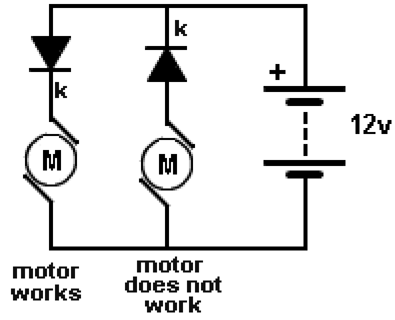

Fig 32. The DIODE

THE DIODE
The next simple electronic component is the **DIODE.**
It only works when connected correctly.
A DIODE allows current to flow through it when it is connected as shown in the diagram.
A Diode is similar to a one-way water valve.
When the diode is "facing down," the motor spins.
When it is "facing up" the motor does not spin.

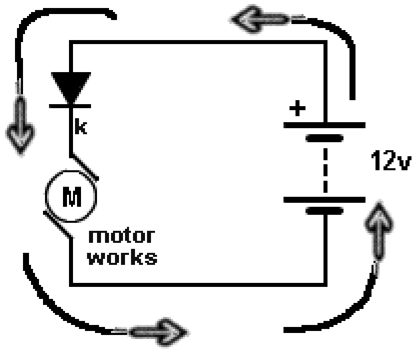

The diagram shows the "current path" around the circuit. The current is measured in AMPS and we discuss current as CONVENTIONAL CURRENT.
This is the way current was thought to flow when electricity was born and they said it flows out the POSITIVE TERMINAL of the battery, around the circuit and into the NEGATIVE TERMINAL.
The arrow on the diode shows the current will flow through the diode and allow the motor to spin.
The diode is said to be **FORWARD BIASED**.

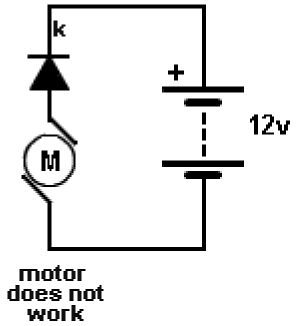

There is no flow of current because the diode prevents any current-flow when connected as shown.
The motor DOES NOT WORK.
The diode is said to be **REVERSE BIASED**.

to Index

There are hundreds of different types of diodes.
Power diodes, signal diodes, low voltage diodes, high voltage diodes, high-speed diodes and many other types.
They all do one thing.
They pass current in one direction and if turned around, they **do not pass** any current.

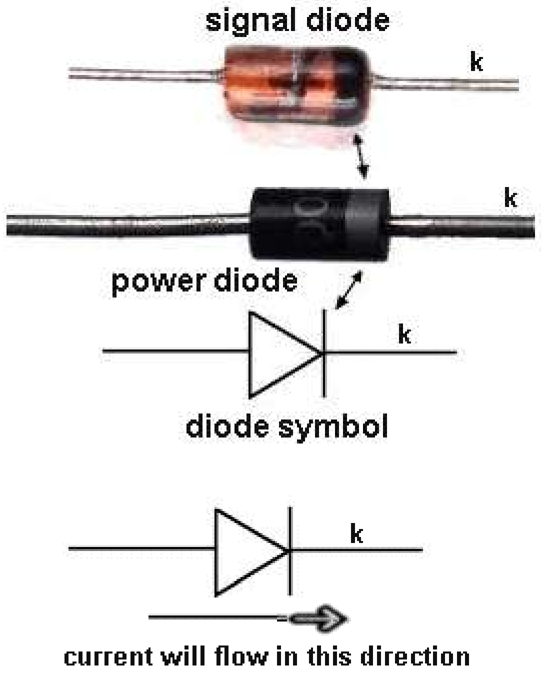
signal diode
k
k
power diode
k
diode symbol
k
current will flow in this direction

Fig 33. A Signal diode and a Power diode

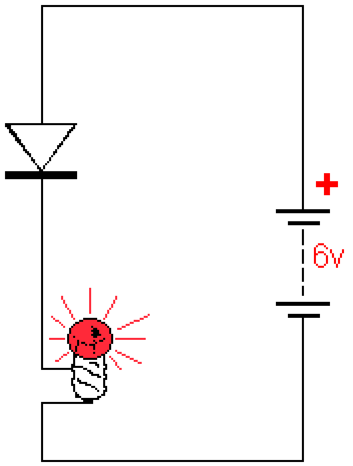

How a diode works

to Index

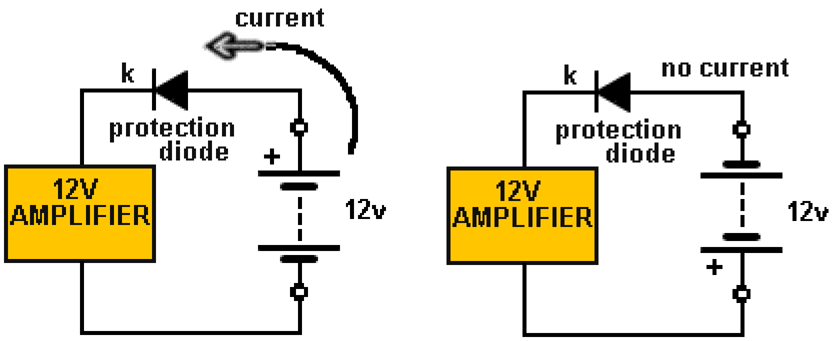
current
k
protection diode
+
12V AMPLIFIER
12v
k
no current
protection diode
12V AMPLIFIER
12v
+

Fig 34. A Protection diode

Diodes perform many function in electrical and electronic circuits. Here is an application as a **PROTECTION DIODE.** It protects the amplifier. If the 12v battery is connected around the wrong way, no current will flow.

to Index

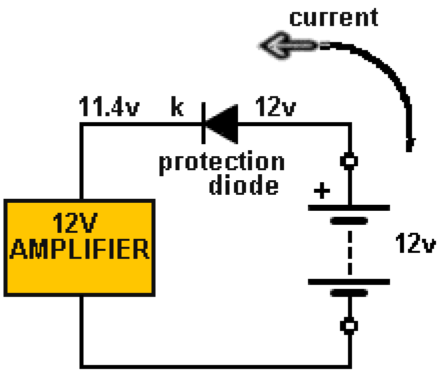
current
11.4v
k
12v
protection diode
+
12V AMPLIFIER
12v

Fig 35. Diode "Voltage Drop"

When a diode is placed in a circuit (and current is flowing), a small voltage develops across the diode. This voltage is called the **FORWARD VOLTAGE DROP.**
This voltage is approximately 0.6v.
This is due to a junction inside the diode where two different materials are joined.
Normally, this voltage is not important because it is only small, but sometimes you need to take it into account.
For the circuit above, the amplifier only gets 11.4v

to Index

Voltage is measured with a

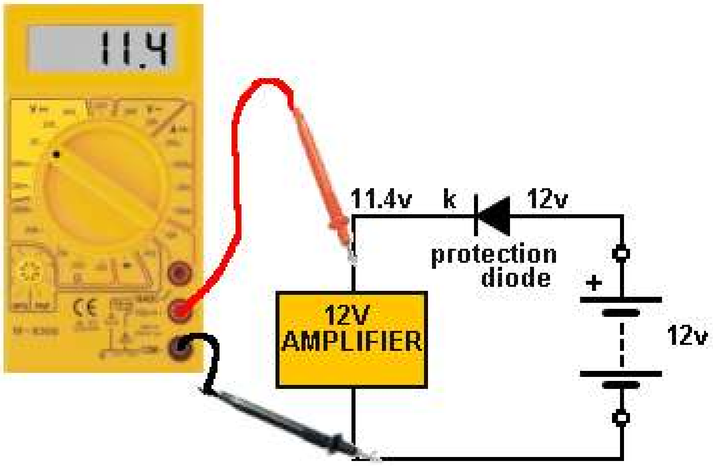
11.4
11.4v
k
12v
protection
diode
+
12V
AMPLIFIER
12v

Fig 36. Measuring Voltage with a Digital Multimeter

VOLTMETER.

Multimeters have 2 or 3 voltage ranges so you can measure low voltage (0v to 20v), medium voltages (0v to 200v and high voltages (0v to 500v).

A voltmeter is placed across the component being tested, as shown in the diagram.

The Digital Multimeter is detecting 11.4v across the amplifier.

to Index

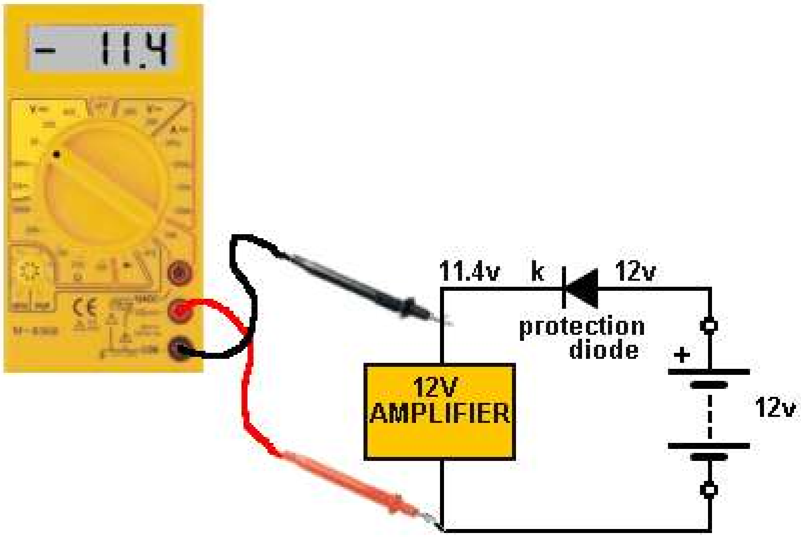
- 11.4
11.4v
k
12v
protection
diode
+
12V
AMPLIFIER
12v

Fig 37. Measuring Voltage with a Digital Multimeter

If you place the probes of a digital multimeter around the wrong way on a component, the display will show a "-"

The meter will not be damaged.

to Index

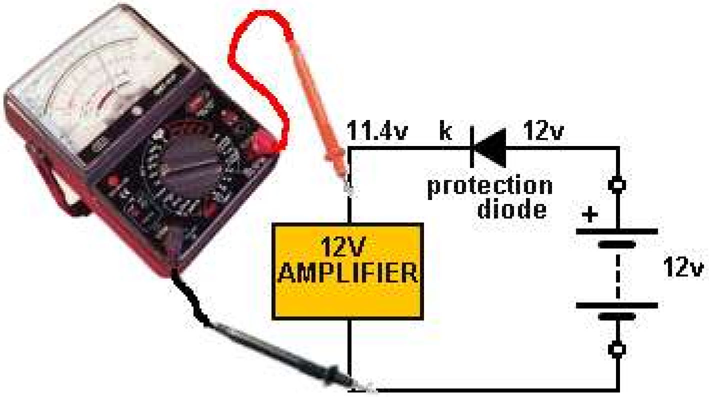
11.4v
k
12v
protection
diode
+
12V
AMPLIFIER
12v

Fig 38. Measuring Voltage with an Analogue Multimeter

An analogue Multimeter must be connected around the correct way to make the pointer move "up scale."

Select the range that will allow the pointer to show somewhere in the middle of the scale.

to Index

Current is measured by

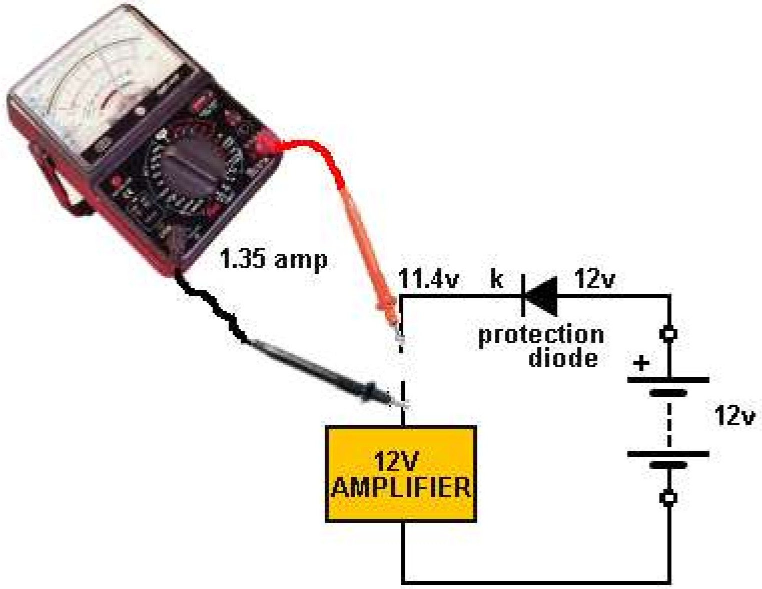
1.35 amp
11.4v
k
12v
protection diode
+
12v
12V AMPLIFIER

Fig 39. Measuring Current with an Analogue Multimeter

"breaking into the circuit" and inserting the leads so the positive probe is closest to the positive of the battery.
If you connect the leads around the other way, the needle will not move but it will hit the "end stop" and you may have to "bump" the meter to get the pointer to move from its jammed position.

to Index

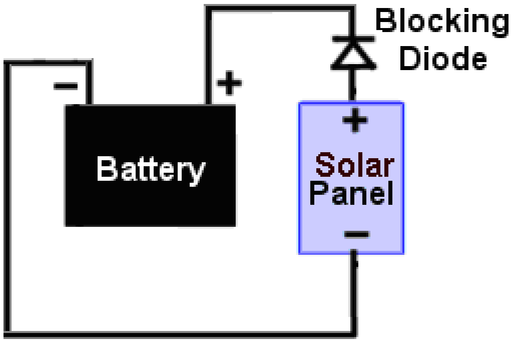
Blocking Diode
−
+
Battery
+
Solar Panel
−

Fig 39a. The Solar Panel and the Blocking Diode

A DIODE is also used with a solar panel to prevent the battery discharging into the solar panel when the sun is not shining.
When the solar panel is not receiving any light it becomes a resistor with a large value and a small current can flow through it from the battery. The diode prevents this current-flow. The diode is called a **BLOCKING DIODE**.

to Index

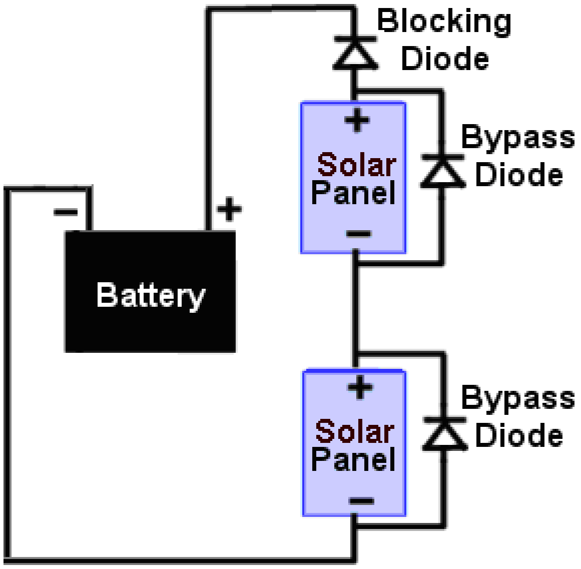
Blocking Diode
+
Solar Panel
−
Bypass Diode
−
+
Battery
+
Solar Panel
−
Bypass Diode

Fig 39b. The Solar Panel with diodes

Diodes are given different names, according to their function. They all perform the same job by passing current in one direction and prevent current-flow in the opposite direction.
When the top solar panel is shaded by a cloud, it generates less current and this will reduce the current into the battery.
By placing a diode across the panel, the diode will pass the current produced by the lower panel to the battery.
These diodes are called **BYPASS DIODES.**

called BLOCKING DIODES and BYPASS DIODES.

to Index

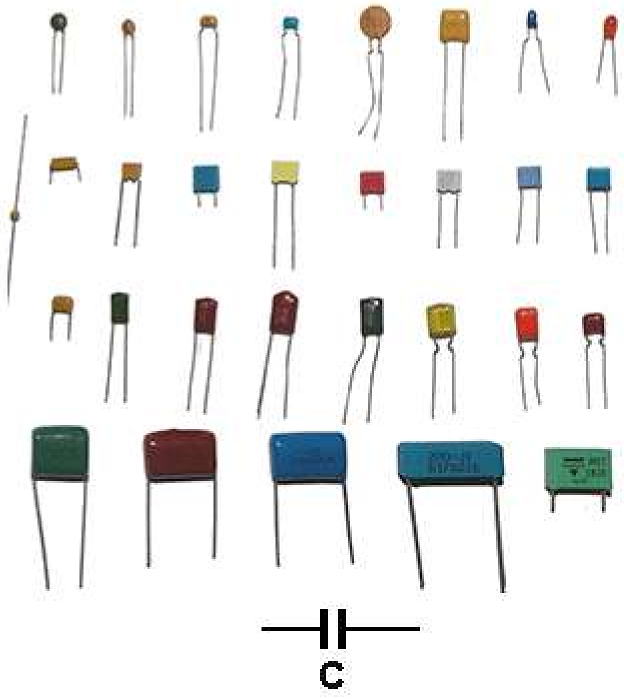

Fig 40. Capacitors and symbol

The next component we cover is the **CAPACITOR**.
There are thousands of different types of capacitor.
Each value of capacitor can have a low voltage rating, medium voltage or high voltage.
Capacitors can be very small in size and shape or very stable with temperature-rise or simply very cheap to make.
A capacitor consists of two thin sheets of metal such as aluminium with a thin sheet of plastic between.
The sheets may be rolled up in a cylinder or laid on top of each other.
The fact is this: the top sheet of metal does not touch the bottom sheet. This is shown in the symbol.
The resistance between the two terminals is INFINITE.
The 6th capacitor in the top row is called a **MONOBLOCK**.

to Index

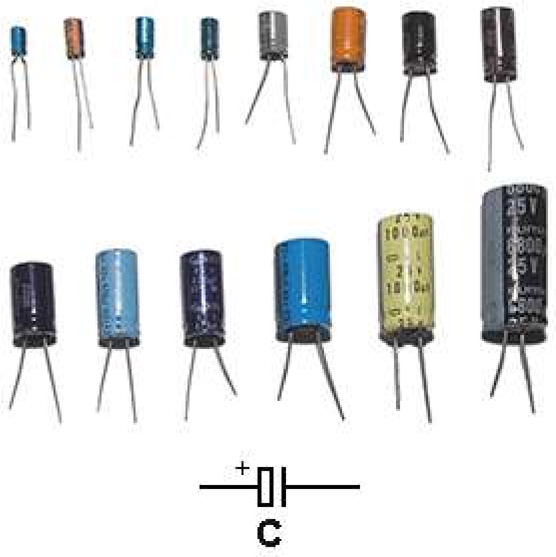

Fig 41. Electrolytic Capacitor

A capacitor gets bigger as its value increases.
It also gets bigger when the voltage-rating increases.
The basic unit of capacitance is the FARAD. A one-farad capacitor would be the size of a house. To make the capacitor smaller the sheets are etched to increase the surface-area and different insulating materials are used between the sheets.
The result is a capacitor called an **ELECTROLYTIC**. It is a bit like a rechargeable battery. It stores a lot of energy in a small space.
The negative lead is shorter and has a black stripe on the side of the electrolytic.

to Index

One FARAD is too big to handle. We use smaller values.
The middle of the range is one microfarad. This is written as 1u. (sometimes you see uF)
This is one-millionth of a FARAD.
The smallest value of capacitance is one picofarad. This is one millionth of a microfarad.
It is written as 1p.
Capacitors are broadly separated into two groups. 1p to 1u and 1u to 100,000u
Capacitors 1p to 1u are ceramic, polyester, air, styroseal, monoblock and other names.
Capacitors 1u to 100,000u are electrolytic or tantalum. A tantalum is the same as an electrolytic - for testing purposes - it is a more-compact electrolytic.

1 microfarad is one millionth of 1 farad.
1 microfarad is divided into smaller parts called nanofarad.
1,000 nanofarad = 1 microfarad
Nanofarad is divided into small parts called picofarad
1,000 picofarad = 1 nanofarad.

Recapping:
1p = 1 picofarad. 1,000p = 1n (1 nanofarad) 1,000,000p = 1u
1,000n = 1u (1 microfarad)
1,000u = 1millifarad
1,000,000u = 1 FARAD.

Examples:
All ceramic capacitors are marked in "p" (puff")
A ceramic with 22 is 22p = 22 picofarad
A ceramic with 47 is 47p = 47 picofarad
A ceramic with 470 is 470p = 470 picofarad
A ceramic with 471 is 470p = 470 picofarad
A ceramic with 101 is 100p (it can also be 100)
A ceramic with 102 is 1,000p = 1n
A ceramic with 223 is 22,000p = 22n
A ceramic with 104 is 100,000p = 100n = 0.1u A common 100n is called a **MONOBLOCK**.
A ceramic with 105 is 1u

TYPES OF CAPACITOR
For testing purposes, there are two types of capacitor.
Capacitors from 1p to 100n are non-polar and can be inserted into a circuit around either way.
Capacitors from 1u to 100,000u are electrolytics (or tantalum) and are polarised. They must be fitted so the positive lead goes to the supply voltage and the negative lead goes to ground (or earth).

to Index

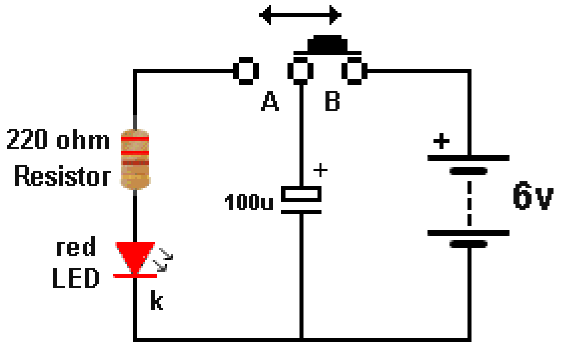

Fig 42. Charging a Capacitor

Here is an experiment to show how much (little) energy is stored in a 100u electrolytic.
When the slide-switch is in position "B", the 100u is charged by the 6v battery.
When the slide switch is moved to position "A" the electrolytic supplies energy to illuminate the red LED via the 220R resistor. It will illuminate for a short period of time.
By moving the switch back and forth, you can keep the LED illuminated.

to Index

Capacitors can be connected in **Series or Parallel** to obtain a value of capacitance you may not have available.
They are also connected in series to increase the effective **VOLTAGE RATING**.
However when two equal-value capacitors are connected in series, the final value is HALF, and thus you need two with double the final-value to get a value with an increased voltage-rating.
When two equal-value capacitors are connected in series, the result is HALF.
(This is the opposite to connecting resistors)

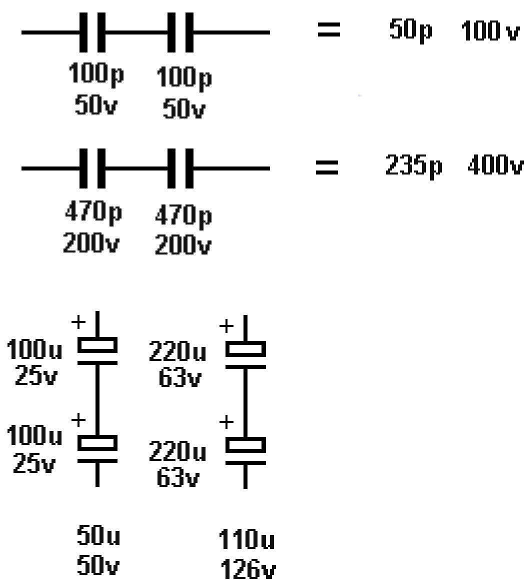

Fig 43. Capacitors in Series

to Index

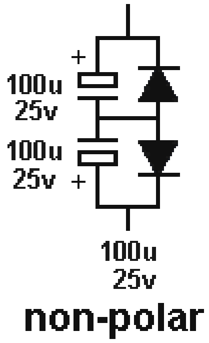

Fig 43a. Non-polar Capacitor

Non-polar Capacitor (electrolytic)

A normal electrolytic must be connected the correct way in a circuit because it has a thin insulating layer covering the plates that has a high resistance.

If you connect the electrolytic around the wrong way, this layer "breaks-down" and the resistance of the electrolytic becomes very small and a high current flows. This heats up the electrolytic and the current increases. Very soon the capacitor produces gasses and explodes.
One big mistake in many text books shows how to make a non-polar electrolytic by connecting two "back-to-back."
They claim 2 x 100u connected back-to-back is equal to 47u.
This appears to be case when testing on a meter but the meter simply charges them for a short period of time to get a reading.
If you allow them to charge fully you will find the reverse electrolytic has a very small voltage across it.
Secondly, when you are charging them, you are putting a high current through the reverse electrolytic and damaging the layer.
To prevent this, you need to add two diodes as shown in the diagram.
In addition, 2 x 100u "back-to-back" is very near 100u.

to Index

Capacitors can be connected in **Parallel** to obtain a value of capacitance you may not have available.
(This does not change the **VOLTAGE RATING**.)
When two equal-value capacitors are connected in parallel, the result is DOUBLE.
(This is the opposite to connecting resistors).
If one electrolytic is 25v

and the other 63v, the answer is the LOWER VOLTAGE = 25v.

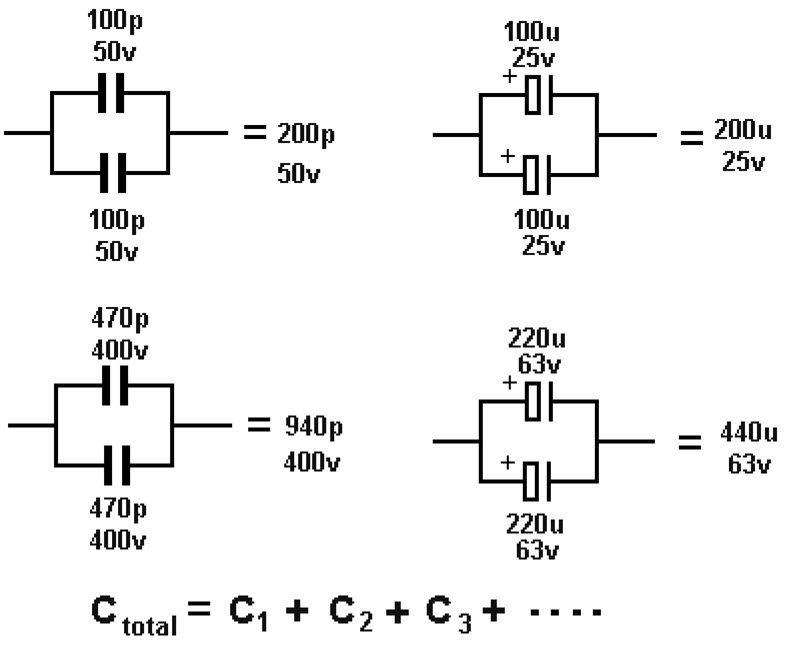
100p
50v
100p
50v
= 200p
50v
100u
25v
100u
25v
= 200u
25v
470p
400v
470p
400v
= 940p
400v
220u
63v
220u
63v
= 440u
63v
$C_{total} = C_1 + C_2 + C_3 + \cdots$

Fig 44. Capacitors in Parallel

to Index

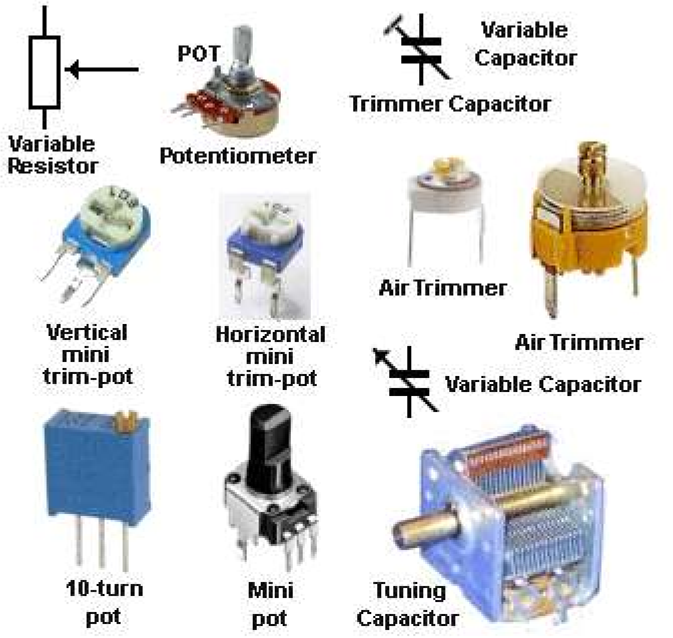
Variable Resistor
POT
Potentiometer
Variable Capacitor
Trimmer Capacitor
Vertical mini trim-pot
Horizontal mini trim-pot
Air Trimmer
Air Trimmer
Variable Capacitor
10-turn pot
Mini pot
Tuning Capacitor

Fig 45. Variable Capacitors and Resistors

The value of a capacitor or resistor may need to be increased or decreased in a circuit to tune in radio stations or increase and decrease the volume of a speaker.
The symbol for these components have an arrow to show they can be adjusted.
The resistance of a potentiometer can be from 1 ohm to 5M
They come in many different shapes and sizes to suit the PC board or front-panel layout.
The "T" represents a trimmer capacitor and this can be from 1p to about 120p.
A variable capacitor will be from about 10p to 415p.

to Index

Simple CIRCUITS
We have covered enough symbols and components to create a number of simple circuits.
These circuits will show how to connect a motor, a LED, (how to make it bright or dull) and how to connect 4 cells to make a battery.

Note: The flashing LED

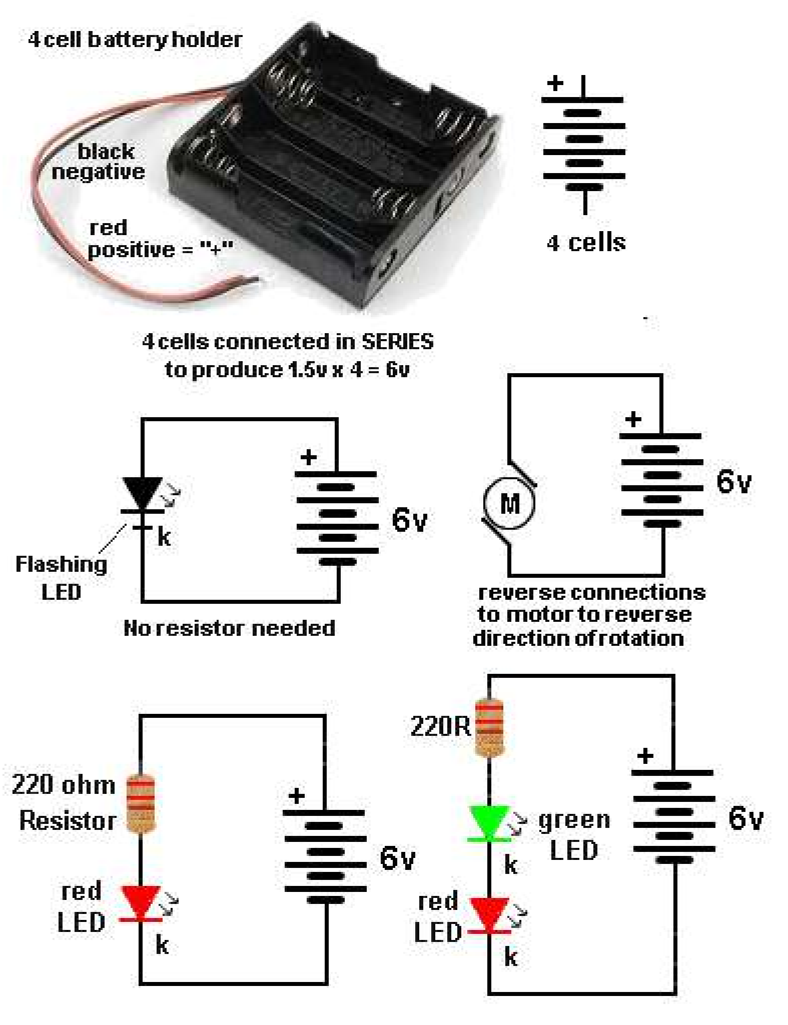
4 cell battery holder
black negative
red positive = "+"
+
4 cells
4 cells connected in SERIES to produce 1.5v x 4 = 6v
+
6v
k
Flashing LED
No resistor needed
+
M
6v
reverse connections to motor to reverse direction of rotation
220 ohm Resistor
+
6v
red LED
k
220R
+
green LED
k
6v
red LED
k

Fig 46. Simple Circuits

does not need a resistor because a resistor and chip are inside the LED, to make it flash and control the current.

Connect all the components around the correct way and then connect them around the wrong way to see what happens.
Connect the flashing LED in series with a red LED and see what happens.

to Index

QUESTIONS

1. Explain why the Flashing LED circuit has no external resistor.
2. How many 1.5v cells are needed to produce a 6v battery
3. Explain what happens when you reverse the leads to a motor.
4. Identify the positive terminal:

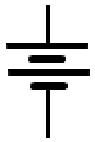

5. Can 3 green LEDs be connected in series to a 6v supply?
6. A variable resistor is also called: ____________________
7. The combined resistance of two 1k resistors in series is: ______________
8. The combined resistance of two 1k resistors in parallel is: ______________
9. Name the short lead on a LED
10. Name the type of multimeter with a pointer and scale: ____________________
11. The total capacitance of two 100u electrolytics in series is: _____________
12. The total capacitance of two 100u electrolytics in series is: _____________
13. Write these values in words:
22R __________
4k7 __________
1n ____________
100u ____________
14. How many 1.5v cells in a 9v battery?

15. The red probe is: _______________ (positive/negative)
16. Conventional current flows from: _________ (positive to negative / negative to positive)
17. Make a 470u electrolytic with two electrolytics:
18. What is the voltage drop across a diode?
19. Name the component that only allows current to flow in one direction: ____________
20. Name this symbol:

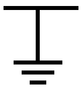

21. When resistors are connected in series, the resistance of the combination: _________ (increases / decreases)
22. When two capacitors are connected in parallel, the voltage-rating of the combination: _______________ (increases / equal to the capacitor with the lowest voltage-rating)
23. Draw two 2k2 resistors in parallel.
24. Which is larger: 470R or 22k
25. What is the value of this combination:

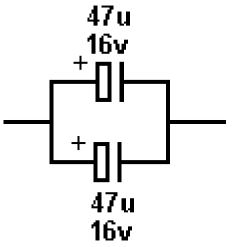

26. What is the name of the resistor in series with a LED: ____________________
27. What is the voltage across the amplifier: _________

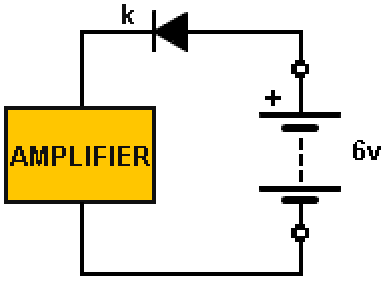

28. Identify the fault with these circuits:

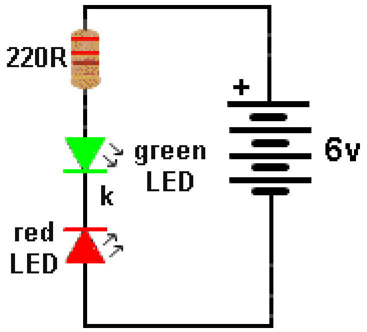

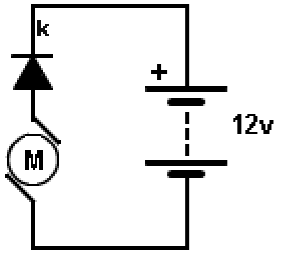

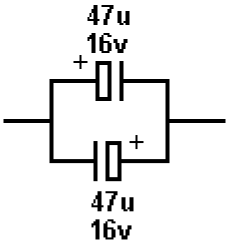

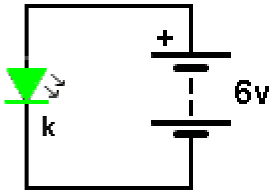

to Index

Passive Components
All the components shown

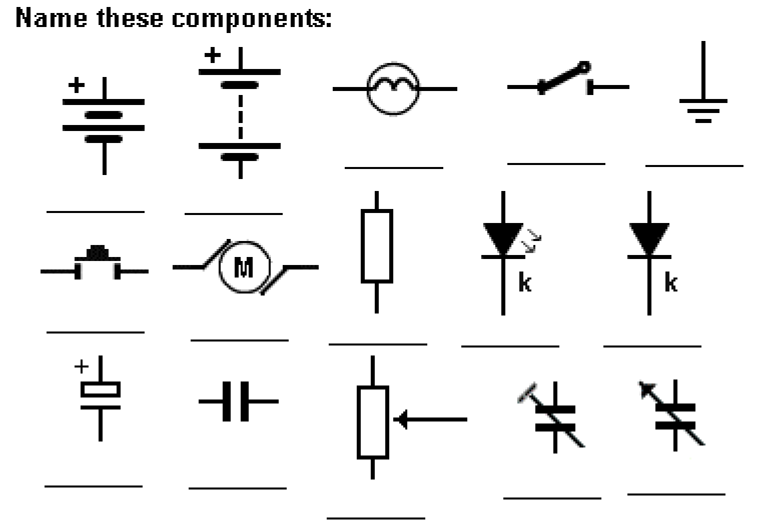

Fig 47. Passive Components

on the left are called PASSIVE COMPONENTS. This means they do not amplify.
Write the name beside each symbol.

to Index

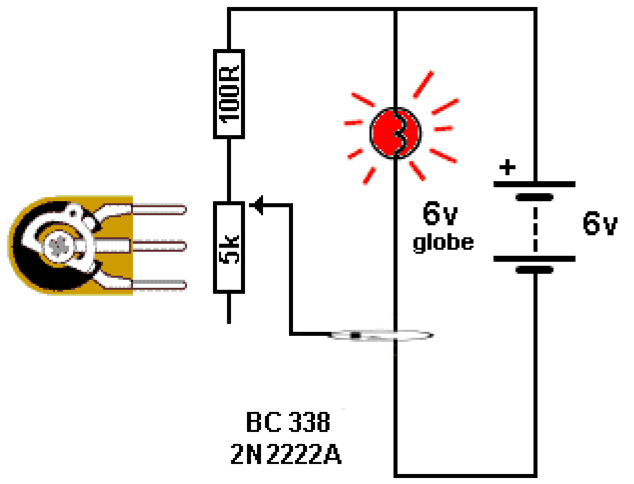

Fig 47a. The Potentiometer in action

The Potentiometer is a variable resistor.
It consists of a curved carbon track with a wiper that touches the track and can be turned via a screwdriver or knob.
The wiper is the middle wire on the circuit symbol and it moves up and down as shown in the animation.
When the three leads are connected the symbol is called a potentiometer.
When two leads are connected it is a variable resistor.
When the resistance increases, less current flows through the pot.

to Index

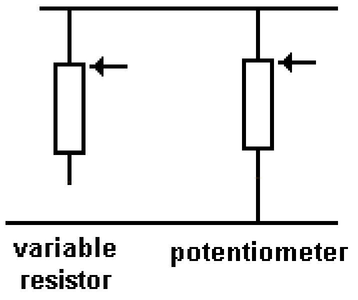

Fig 47b. The Variable Resistor and Potentiometer

There is a difference in operation between a **Variable Resistor** and a **Potentiometer**.
Both will increase or decrease the sound level as a volume control or the speed of a motor or the brightness of a globe, but a Potentiometer will guarantee zero volume or zero brightness when the pot is turned fully anticlockwise (as shown in the animation).
This is because the output will be zero volts, but the variable resistor may still deliver some "energy" (voltage and current) to the circuit when turned fully anticlockwise.

to Index

Potentiometers come in values from 100 ohms to 5 Meg ohms (500R, 1k, 2k, 5k, 10k, 50k, 100k, 200k, 250k, 500k, 1M are most popular).
They come as linear, or logarithmic where the resistance of the track (per mm) is higher at one end. Because our hearing is not linear, these pots can be used as volume controls to produce a gradual (very nearly linear) increase in volume.
Selecting the correct value of resistance for a circuit is VERY complex. If the value is not correct, the volume will not be loud or it will drop to zero before the pot is turned fully anticlockwise. Or

the motor will drop to zero at mid-turn of the pot or it will not reduce in RPM to the desired amount.

The simple answer is to copy a circuit.

Or you can try the whole range of pots and you will find one value is the best.

A Potentiometer can be used in hundreds of different circuits to produce hundreds of different effects, but the actual "thing" that flows between the input and output is a percentage of the voltage. At the same time the current will also be passed to the output at a reduced value. A pot actually delivers BOTH reduced values at the same time and the receiving circuit will be designed to "look for" the change in voltage or current. If the supply voltage is not rising or falling, the "values" are called DC values.

The voltage can also be in the form of a signal (volume). This is called an AC signal.

to Indexx

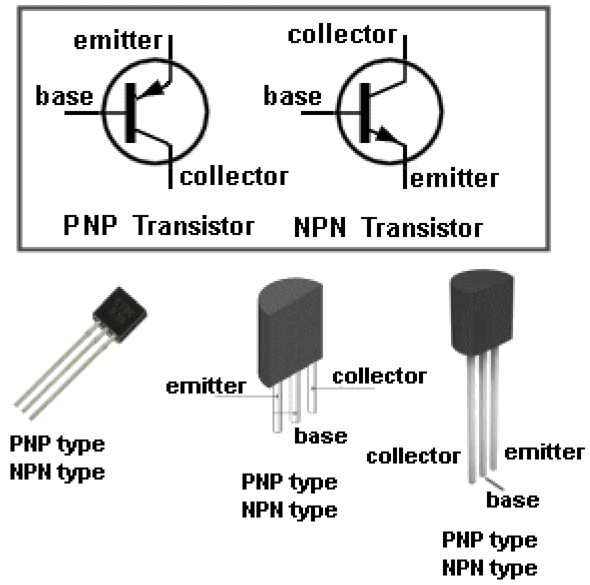
emitter
base
collector
collector
base
emitter
PNP Transistor
NPN Transistor
PNP type
NPN type
emitter
collector
base
PNP type
NPN type
collector
emitter
base
PNP type
NPN type

Fig 48. The TRANSISTOR

The TRANSISTOR

A TRANSISTOR is an ACTIVE device. It AMPLIFIES.

There are many types of transistor (over 20,000 different types) from hundreds of manufacturers and they have many different names.

We are going to study the simplest.

It has the technical name **BIPOLAR JUNCTION TRANSISTOR (BJT)** but we are going to call it a **TRANSISTOR**.

There are two types in this group: **PNP** and **NPN**.

The type we will study is also called a **SMALL-SIGNAL TRANSISTOR.**

You cannot tell an **NPN** transistor from **PNP** by looking at it. You must test it in a circuit.

In **Fig 65** you will make a **Transistor Tester** project, but first some basic facts:

to Index

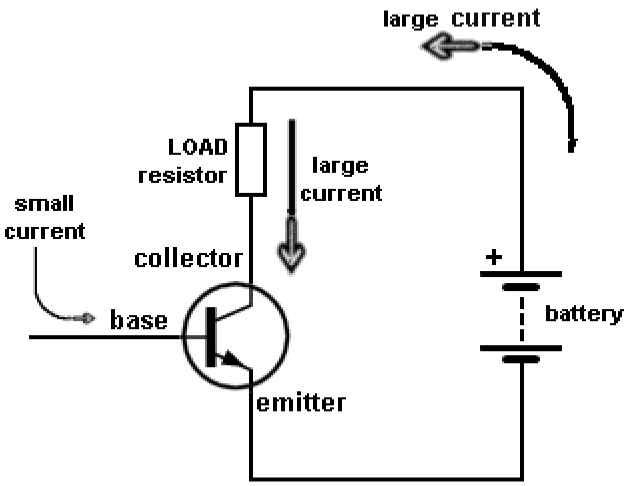
large current
LOAD resistor
large current
small current
collector
base
emitter
+
battery

Fig 49. The NPN TRANSISTOR in a Circuit

The first type of transistor we are going to study is the **NPN**.

A transistor has three leads:

BASE, COLLECTOR and **EMITTER**

Basically, a small current enters the base and a large current flows through the collector-emitter leads as shown in the diagram.

The resistor in the collector lead is called the LOAD Resistor.

Sometimes the load is a speaker.

to Index

The transistor is similar to the diagram opposite. A small drop of water entering the base is amplified to produce a loud DRIP from the speaker.

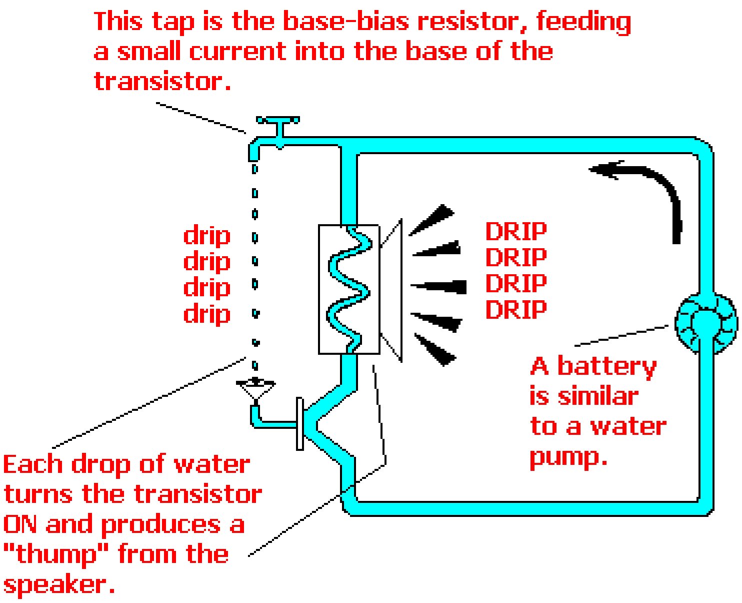
This tap is the base-bias resistor, feeding a small current into the base of the transistor.
drip
drip
drip
drip
DRIP
DRIP
DRIP
DRIP
A battery is similar to a water pump.
Each drop of water turns the transistor ON and produces a "thump" from the speaker.

Fig 50. The Water TRANSISTOR

to Index

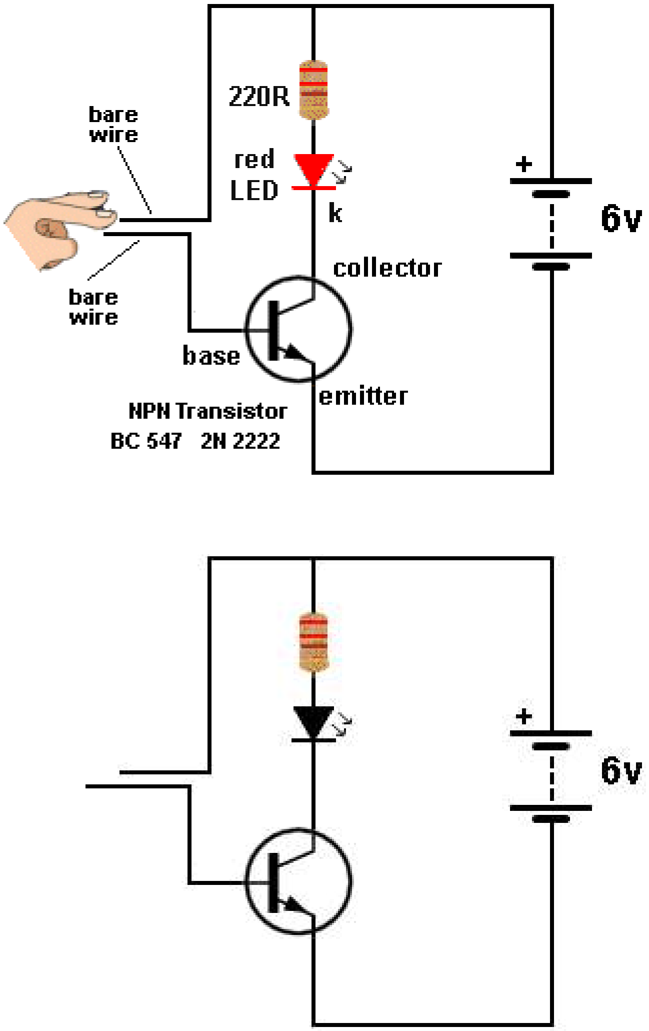
220R
bare wire
red LED
k
+
6v
collector
bare wire
base
emitter
NPN Transistor
BC 547 2N 2222
+
6v

Fig 51. One Transistor Circuit

In this experiment we will construct a ONE TRANSISTOR circuit similar to the WATER TRANSISTOR above and observe the results.
Make sure the two leads DO NOT TOUCH. If they touch, the transistor will be DESTROYED.

The transistor is amplifying the current through your finger via the two leads and it will be very dim.

ANIMATION

The lower diagram shows the transistor turning ON when a finger is pressed against the two wires. The finger produces a resistance that turns the transistor ON and this turns the transistor into a smaller and smaller resistor. That's how more and more current flows through the LED and it gets brighter and brighter.

to Index

By adding another transistor we amplify the

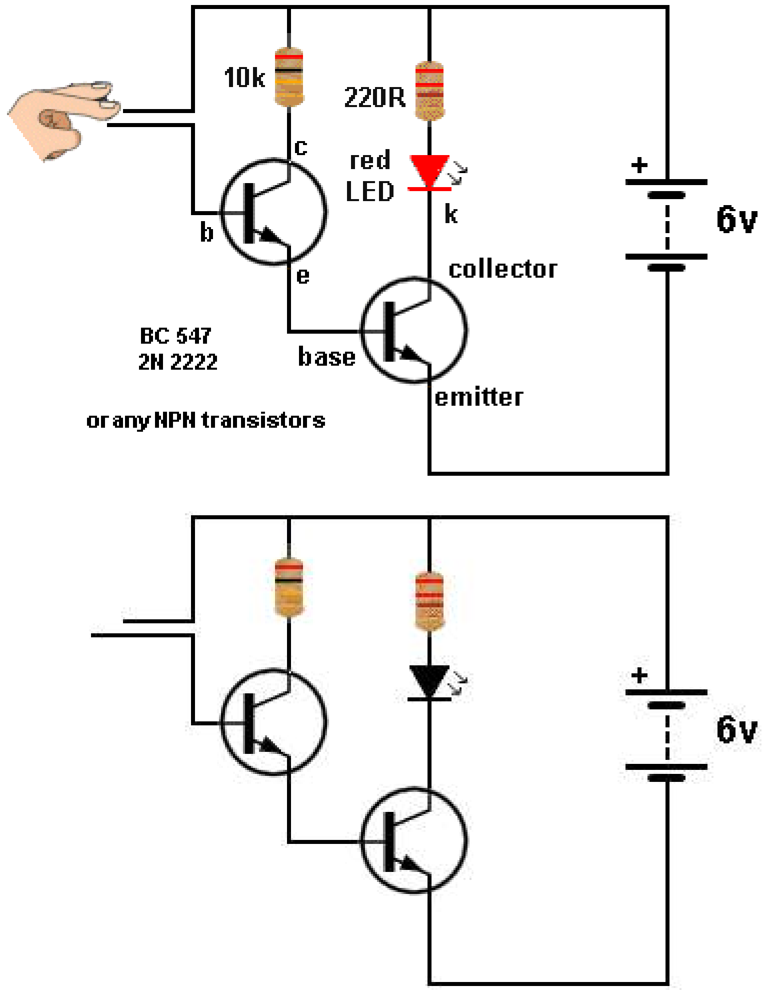
10k
220R
c
red
LED
k
+
6v
b
e
collector
BC 547
2N 2222
base
emitter
or any NPN transistors
+
6v

Fig 52. Two Transistor Circuit

current through the finger about 200 times and now the LED will glow bright.

Make sure the bare wires do not touch each other as this will destroy BOTH transistors.

ANIMATION

The lower diagram shows both transistors turning **ON** when a finger is pressed against the two wires.
They both becomes smaller and smaller resistors.
The first transistor allows more current to flow into the base of the second transistor and this is how the second transistor turns on more and more. This allows more current to flow through the LED and it gets brighter and brighter.

to Index

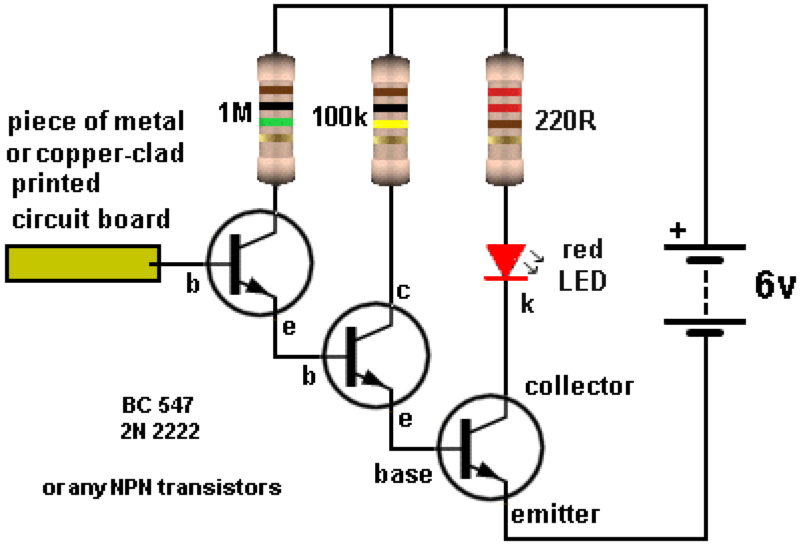
piece of metal
or copper-clad
printed
circuit board
1M
100k
220R
b
c
red
LED
k
+
6v
e
b
collector
e
BC 547
2N 2222
base
or any NPN transistors
emitter

This circuit has enormous gain.
Each transistor has a gain or more than 200 and the final gain will be more than:
200 x 200 x 200 = 8,000,000

8 MILLION!

The circuit is very sensitive to static voltages in the air or electrical waves such as the waveform produced by the electrical wiring in a house.
Move the project around a room and detect all the electrical signals.

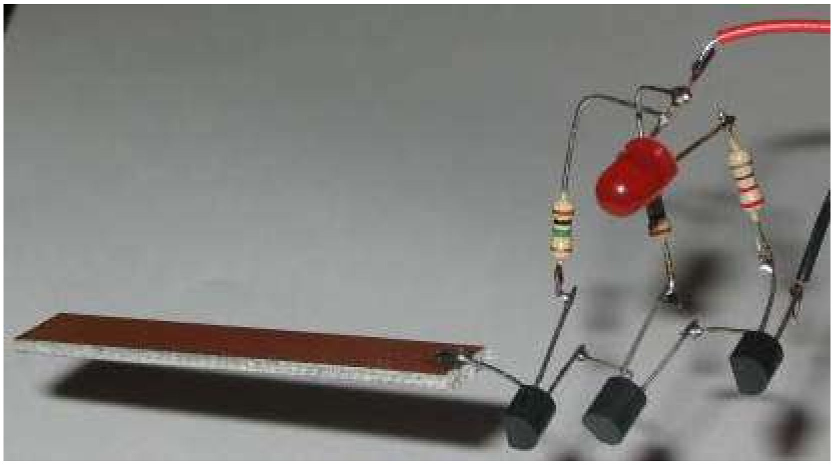

Fig 53. Three Transistors

to Index

You can see the effect of one transistor. It does not do much.
The two transistor circuit allows the resistance of your finger to deliver current into the base of the first transistor and this transistor delivers more current into the base of the second transistor. The result is more collector-emitter current and the LED illuminates.
The three transistor circuit produce an ENORMOUS effect.
It will pick up STATIC ELECTRICITY and all forms of electro-magnetic energy (radiation) and illuminate the LED.

to Index

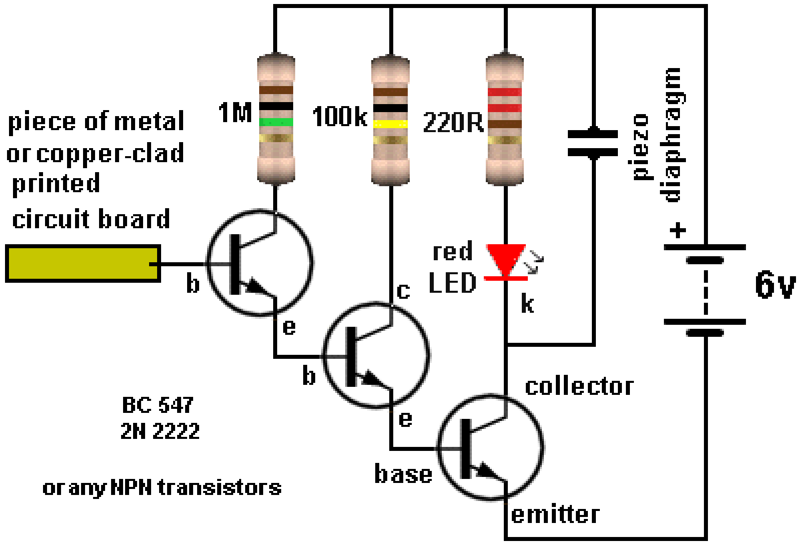
piece of metal
or copper-clad
printed
circuit board
1M
100k
220R
piezo
diaphragm
+
6v
b
e
c
red
LED
k
b
e
collector
BC 547
2N 2222
or any NPN transistors
base
emitter

Fig 54. Adding a PIEZO

By adding a piezo diaphragm to the output you will be able to hear the hum of the mains.
This is the frequency of the supply into your house. It will be either 50 cycles per second or 60 cycles per second.
The term: "cycles per second" is given the name HERTZ after **Heinrich Rudolf Hertz**, who was the first to prove the existence of electromagnetic waves.

to Index

The Piezo diaphragm is held around the outer edge inside a plastic case and when a voltage is applied to the two leads, the thin plate curves very slightly.
When the voltage is removed, the plate returns to its flat shape.
If the voltage is reversed, the plate curves in the opposite direction.

The curving is due to a thin layer of ceramic material under the plate and

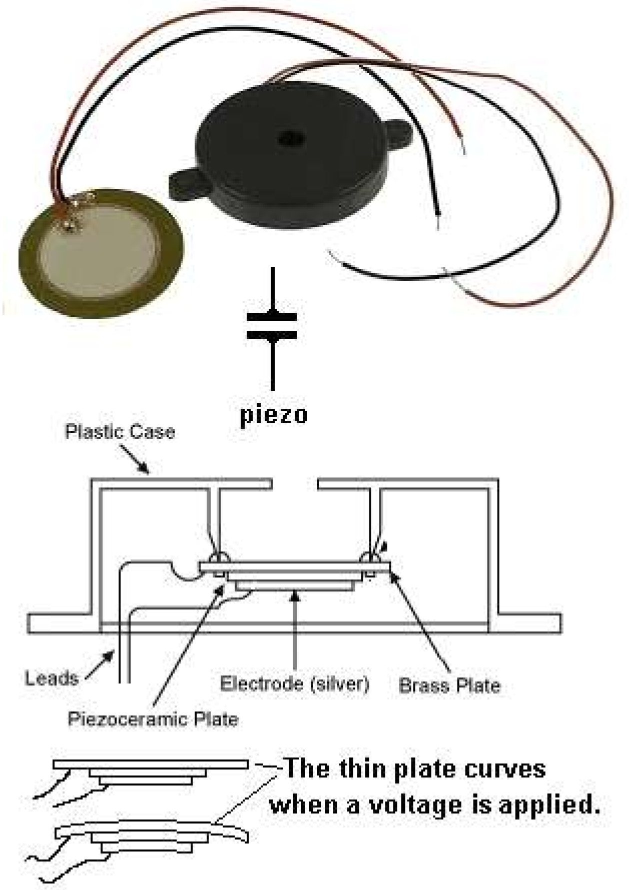
piezo
Plastic Case
Leads
Electrode (silver)
Brass Plate
Piezoceramic Plate
The thin plate curves when a voltage is applied.

Fig 55. The Piezo

then a film of metal is deposited onto the ceramic so a lead can be soldered.

There is infinite resistance between the two leads as the ceramic material is an INSULATOR.
The capacitance between the two leads is approx 22n.
The Piezo is a passive device. It needs a pulse or frequency for it to produce an output.

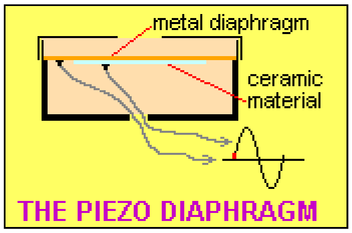
metal diaphragm
ceramic material
THE PIEZO DIAPHRAGM

to Index

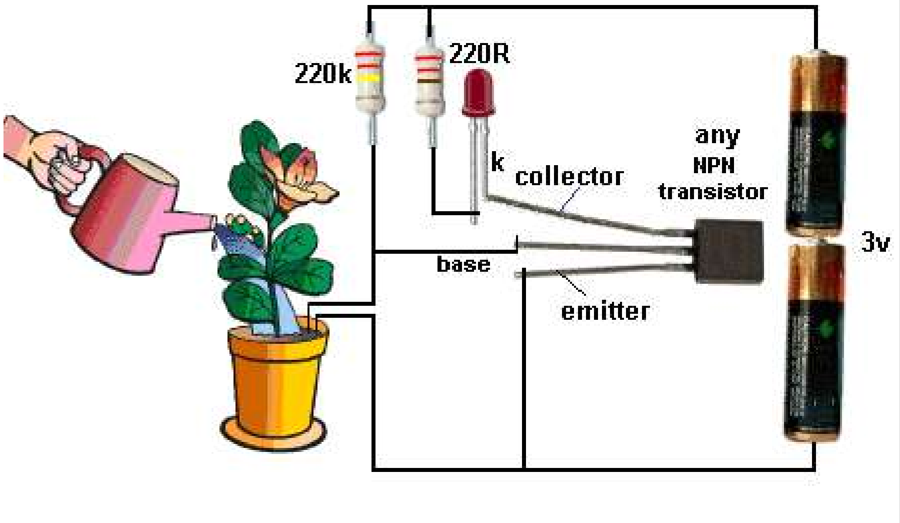
220k
220R
k
collector
any NPN transistor
base
emitter
3v

Fig 56. Plant Needs Water

The **ONE TRANSISTOR CIRCUIT** above can be turned into a detector to show when a plant needs water.
Place the two probes into the soil and water the plant. The LED will turn off. As the water evaporates the LED will turn **ON** to let you know the plant needs watering.

to Index

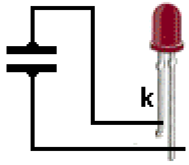

Fig 57. Tap the Piezo

This experiment produces a pulse from the piezo when it is tapped and the LED illuminates briefly.

The LED can be connected either way around.

This proves the diaphragm flexes when a voltage is applied and also in the reverse situation. A voltage is produces when the diaphragm is tapped.

to Index

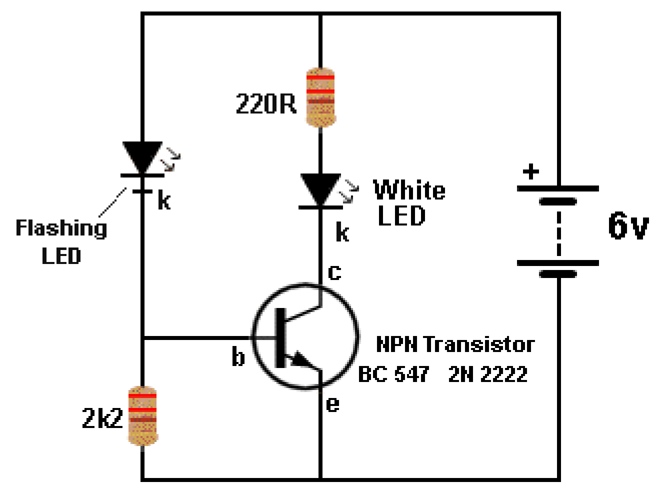

Fig 58. Flashing LED and White LED

A flashing LED is not very bright.
It can be connected to a transistor and the transistor will drive a very bright white LED.
The transistor is an amplifier.
It is amplifying the current flowing through the flashing LED and supplying a higher current for the white LED.

We cannot discuss any further details of the circuit at the moment because the actual operation of the circuit is quite complex.

At the moment we just need to experiment with simple transistor circuits.

to Index

Fig 59. Soldering Iron

We now come to the point of HELPING YOU WITH CONSTRUCTION.
We have already shown you 6 different circuits and there are many ways to build them.
You can:
1. Solder them.
2. Build them on an Experimenter Board
3. Connect the components with clips or twist the leads together.

It does not matter how you build the circuits.
The fact is this: YOU MUST START BUILDING.
The best soldering iron for a beginner is a CONSTANT TEMPERATURE soldering iron.
It has a dial that can be turned to set the desired temperature.
An ordinary soldering iron GETS TOO HOT. It is not suitable for soldering electronic circuits.
This is something that no-one has mentioned before. An ordinary soldering iron will melt the solder TOO QUICKLY and burn the resin inside the solder and make soldering very difficult for a beginner.
Soldering must be done slowly so the resin in the middle of the solder gets hot and cleans the leads of the components so the solder will "stick."
That's why you must apply the solder to the leads you are soldering and allow the resin to "attack" the leads and clean them.
The cheapest TEMPERATURE CONTROLLED soldering Iron is available on eBay for les than $10.00 (post FREE).
You will also need a small roll of solder (0.9mm) and a soldering Iron stand.
Email Colin Mitchell for links to eBay. (talking@tpg.com.au)
A whole book could be written on the ART OF SOLDERING.
Look on the web for articles and videos on SOLDERING.

to Index

SOLDERING

1. Soldering is very easy and very simple. You just need a **Temperature Controlled Soldering Iron**, **Fine solder** and clean components.

2. Remember this: It is **NOT** the solder you need for a joint, but the **FLUX**. And the flux lasts for only 2 seconds. When the flux is **HOT** it attacks and cleans the joint so that the solder will stick.

3. Turn ON the Temperature Controlled Soldering Iron to a low temperature. Put solder on the tip. It will not melt. Turn up the temperature slightly. Try more solder. As soon as the solder starts to melt, this is your starting point. Turn up the temperature slightly MORE and this is the correct temperature for **small, delicate, fine soldering**.

4. Place a component through a hole and bend the lead slightly so the component does not move. Turn the board over and touch the iron on the component and bring the solder **FROM THE OTHER SIDE** so the solder melts and flows towards the iron.
From start-to-finish, count one-two-three and remove the solder. Count four-five and remove the iron. You will have a perfect joint.
If you are soldering thick leads or large pads on a circuit board, you will need to turn the temperature UP slightly.
You must add enough solder to make the joint "bulge" slightly.
Fine solder (1mm or 0.9mm or 0.8mm) makes the best joint because it is easier to use.
Use a wet sponge to clean the tip or a ball of "Steel Wool." Steel wool is the best.
Here is the steel wool, bending the leads and some examples of poor joints due to insufficient solder:

Steel wool cleans the tip beautifully

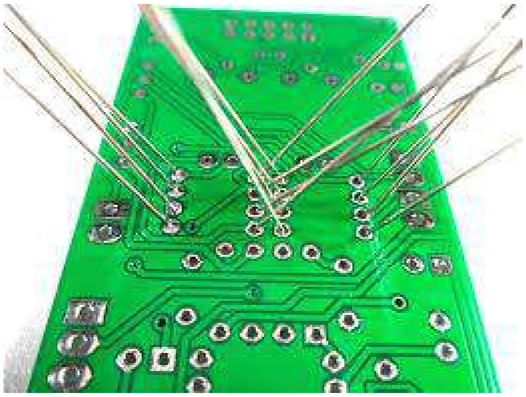

Bend the leads before soldering

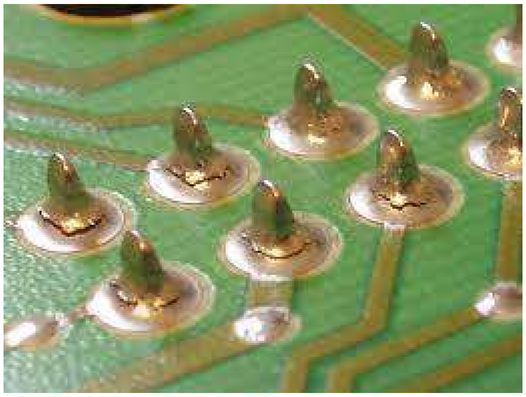

The joints do not have enough solder and that's why they fractured. Called a DRY JOINT.

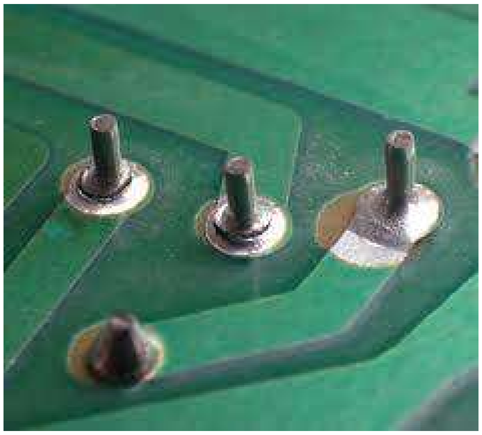

More "Dry Joints."

This is the cheapest and simplest soldering iron stand.

This stand is very messy as the spring grabs the iron and makes it difficult to remove from the stand. Test the stand before buying. You will se why not to buy this type of soldering stand. Get one with a "wide mouth" and a heavy stand is best as it does not move.

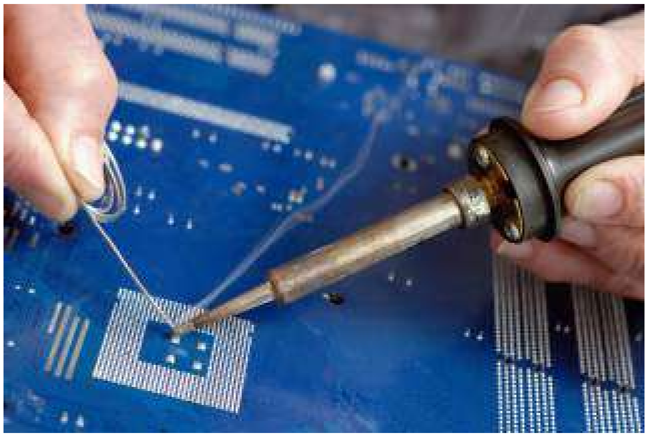

This photo clearly shows how to hold a soldering iron and solder.

This is NOT a temperature-controlled soldering iron and you can see it is too hot as it is burning off the flux too quickly.
Temperature Controlled Soldering Irons are now cheaper than the JUNK soldering iron shown in the photo. See eBay for prices.

to Index

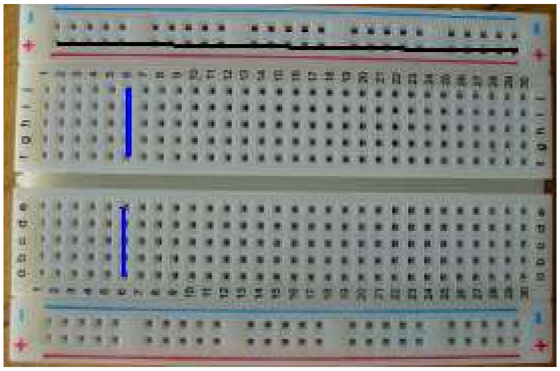

Photo shows a number of components fitted to the breadboard.

BREADBOARD

The term BREADBOARD refers to any piece of wood or plastic containing pins or pegs or clips or holes where you can build a circuit.
The components can be soldered, twisted clipped or fitted into holes.
Breadboard also means the circuit can be easily pulled apart.
Some breadboards do not have two rows for the positive and negative rails. Connections under the board for the positive rail is shown with a black line in the photo. Connections on the main section of the board are shown with blue lines. Your breadboard MUST look exactly like the photo opposite.
Other breadboards are quite useless.
The breadboard in the photo can be purchased on eBay for less than $5.00 (post FREE).

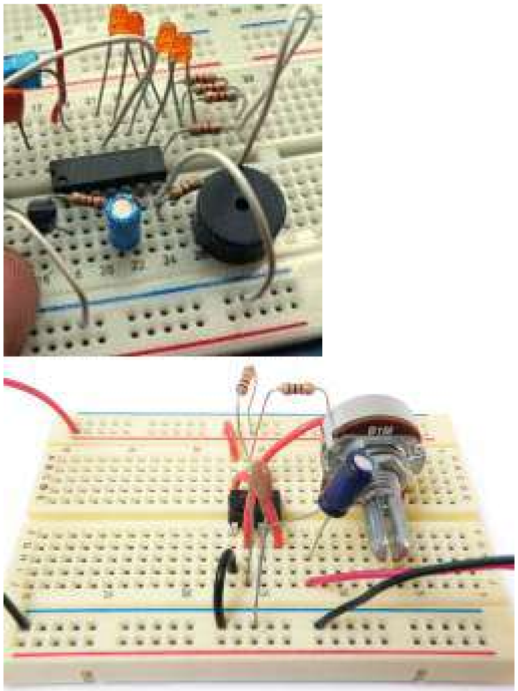

Fig 60. BREADBOARD and Components

to Index

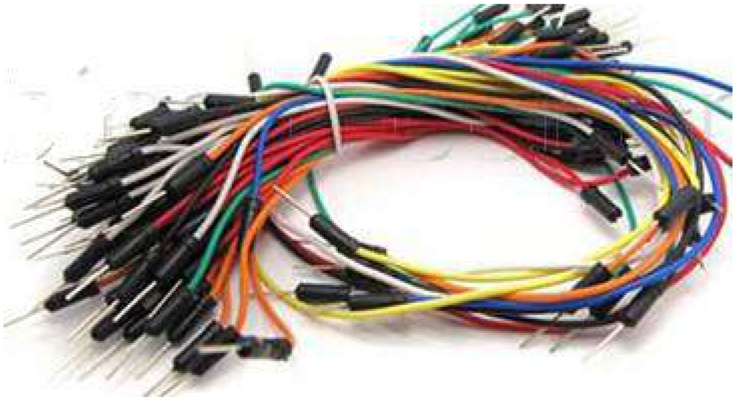

Fig 61. Jumpers

The components on the BREADBOARD are fitted down the holes and metal strips under the board join each column of 5 holes. If you want to join one hole with another, you can use 0.5mm tinned copper wire or JUMPERS. See photo opposite. Jumpers can be purchased on eBay for less than $3.00 posted.

Email Colin Mitchell for links to eBay. (talking@tpg.com.au)

to Index

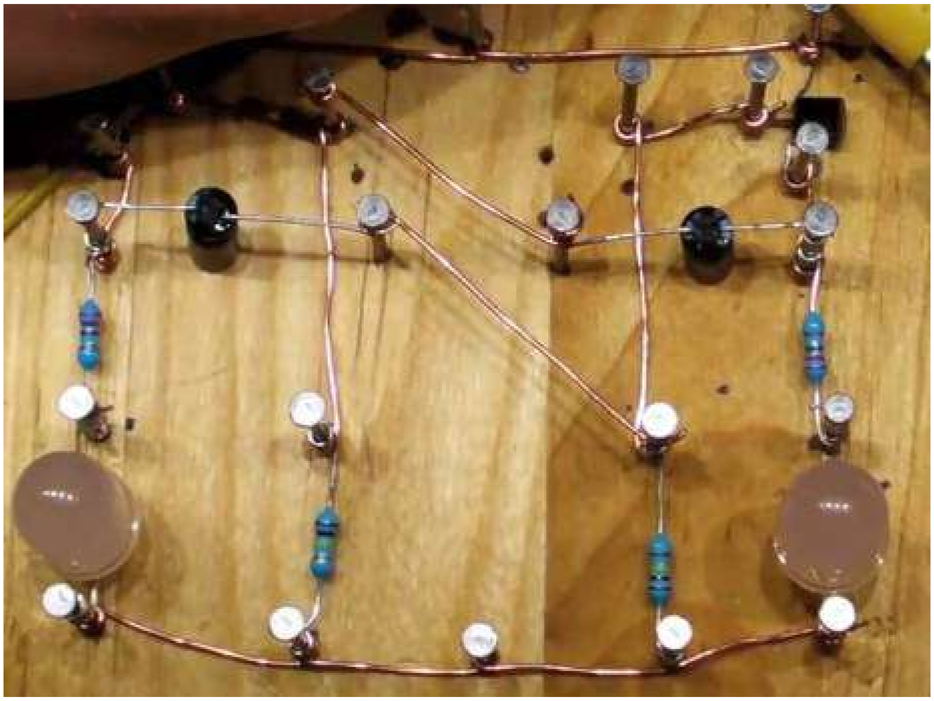

Fig 62. Breadboard with Nails

If you don't have a soldering iron or experimental breadboard, you can make your own board with nails. See the photo above. It is a multivibrator circuit and we will be presenting this circuit in a moment. The components can be twisted around the nails and bare wire used to join some of the nails to complete the circuit.

Another method of connecting the components is called BIRD-NESTING. This involves soldering the components "in the air" as shown in the 27MHz transmitter circuit below:

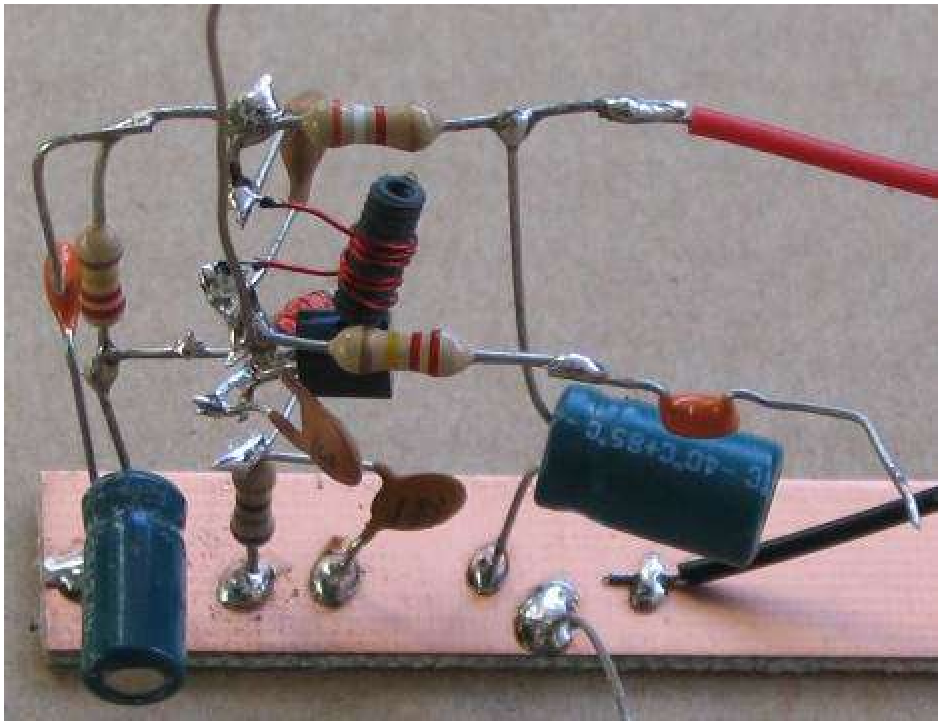

Another way to connect the component(if you don't have a soldering iron), is to wind 6 turns of bare wire around each connection and leaving all the components "in the air." The bare wire can be obtained from hook-up flex. This is plastic coated "wire" containing up to 15 fine strands of wire. Use a single strand for the connections. None of the components will touch each other BY MISTAKE and the circuit will work perfectly. Bird-nesting is a good way to build a quick circuit and test its performance. It might look messy but you can easily change any component.

to Index

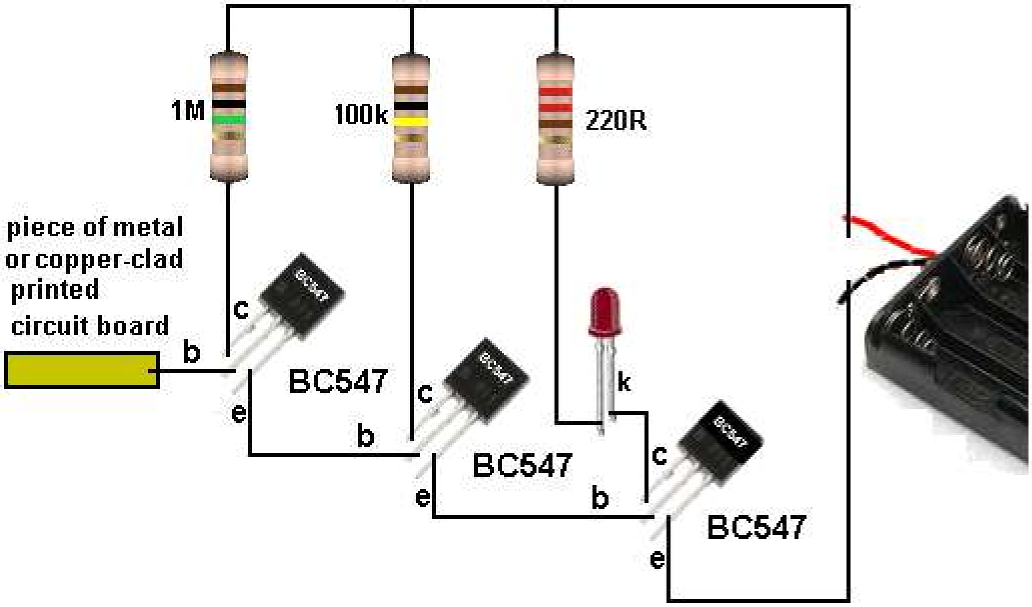
1M
100k
220R
piece of metal or copper-clad printed circuit board
b
c
e
BC547
c
b
e
BC547
k
c
b
e
BC547

Fig 63. Wiring the 3-Transistor Circuit using BC547 Transistors
The diagram shows how to connect 3 x BC 547 transistors.
The leads of a transistor can be collector-base-emitter OR
emitter-base-collector and that's why we have provided 2 different wiring diagrams.

to Index

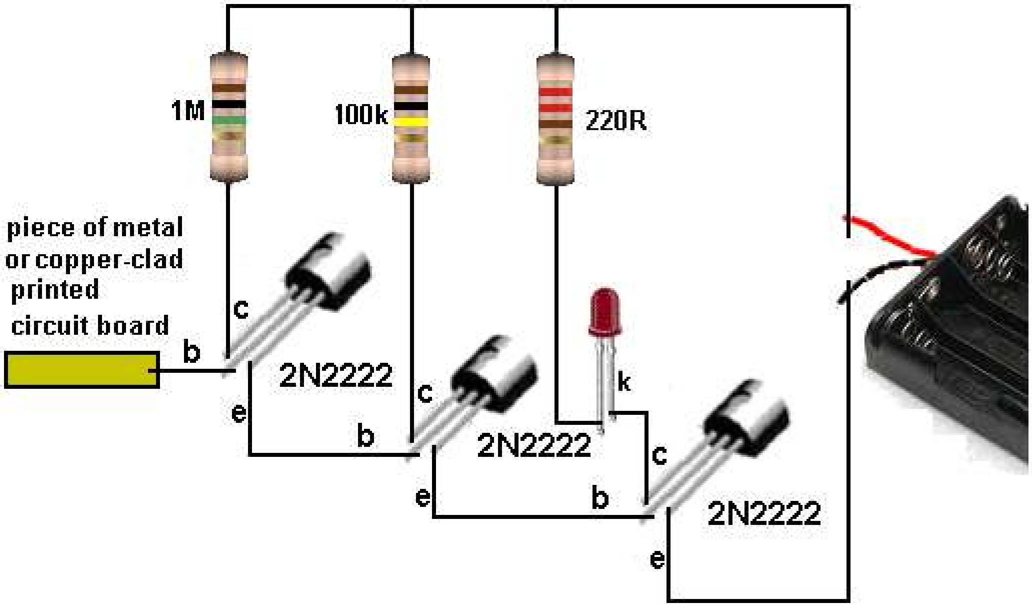
1M
100k
220R
piece of metal or copper-clad printed circuit board
b
c
e
2N2222
c
b
e
2N2222
k
c
b
e
2N2222

Fig 64. Wiring the 3-Transistor Circuit using 2N2222 Transistors
The diagram shows how to connect 3 x 2N2222 transistors.

to Index

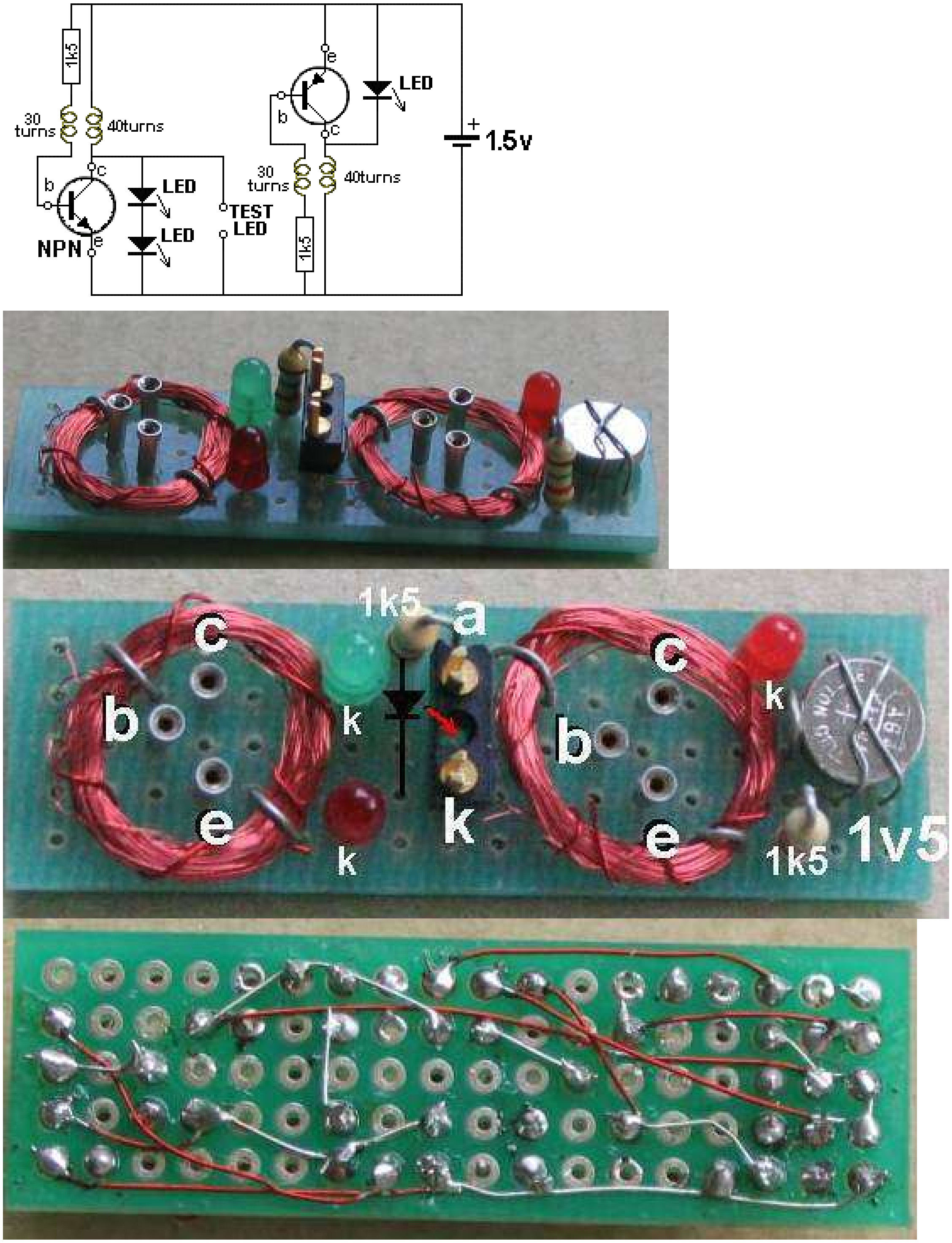
1k5
30 turns
40turns
b
c
e
NPN
LED
LED
TEST LED
e
b
c
LED
+
1.5v
30 turns
40turns
1k5
1k5
a
c
b
e
k
k
k
c
b
e
k
1k5
1v5

Fig 65. Transistor Tester

This handy transistor and LED tester can be built to test LEDs and both PNP and NPN transistors.
The project consists of two identical circuits, one for NPN and one for PNP.
You can build just the NPN section and then build the PNP section later.

LED VOLTAGE

We have shown a LED needs at least 1.7v supply for it to operate.
This circuit works on 1.5v and thus the action of the transistor and coil (called a Transformer) MUST be increasing the voltage for the LED to illuminate.
This circuit works on two "actions."
1. Transistor ACTION - this is the action of a transistor providing gain to make the circuit oscillate.
2. Transformer ACTION - this is the action of a coil of wire producing a voltage higher than the

supply voltage when it is turned off.

This circuit is very technical and very complex.
We will be explaining it in a very simple way because this is a **Basic Electronics Course**.

THE TRANSFORMER - the two coils of wire on the left and the two coils of wire on the right.
When the voltage (actually the current) is switched off, the 40 turn coil in either of the circuits in this project; the voltage across the coil rises to more than the 1.5v supply and is in the opposite direction to the voltage of the supply.
The circuit looks to be very simple but it uses an air-cored transformer to produce the voltage needed to illuminate the LED indicators and the circuit only works when the transistor is connected correctly. There are two separate circuits, one for NPN transistors and one for PNP transistors. We will cover the NPN section:
The circuit turns ON when the NPN transistor is fitted and the current through the 30 turn coil and 1k5 resistor turns ON the transistor and produces expanding flux in the 40 turn coil. This flux cuts the turns of the 30 turn coil and produces a voltage in the coil that adds to the supply voltage and increases the current into the base. This turns the NPN transistor ON more. This action continues until the transistor is fully turned ON. At this point the current in the 40 turn coil is a maximum but it is not expanding flux and the 30 turn coil ceases to see the extra voltage. Thus the current into the base reduces and this turns the transistor OFF slightly. The flux produced by the 40 turn coil now becomes collapsing (or reducing) flux and it produces a voltage in the opposite direction to greatly reduce the current into the base. In a very short period of time the transistor becomes TURNED OFF and it is effectively removed from the circuit. The flux in the 40 turn coil collapses quickly and it produces a voltage in the 40 turn coil that is higher than the supply voltage and is in the opposite direction. This means the voltage produced by the 40 turns ADDS to the supply voltage and is delivered to the LEDs to illuminate them.
The NPN circuit has two LEDs in series so that a LED of any colour (including white) can be connected to the TEST LED terminals and it will illuminate. You can use any colour LED for any of the LEDs, however it is best to use either green or yellow or white for the single LED.
The two "coils" are wound on a 10mm dia pen with 0.1mm wire (very fine wire). The loops of tinned copper wire holding the coils on the board are connected to separate lands under the board and MUST NOT produce a complete loop as this will create a “Shorted Turn” and the circuit WILL NOT WORK.
If the LEDs do not illuminate, simply reverse the wires to the 30 turn coil.
The circuit does not need an ON/OFF switch because the LEDs require a voltage of over 2v to illuminate (the orange LED) and the supply is only 1.5v. A red LED needs about 1.5v to 1.7v to operate but when it is in series with a green LED, this voltage is over 3.5v.
All the components fit on a small matrix board 5 holes x 18 holes. A kit of parts for the project is available for $4.00 plus $3.00 postage and ordering details can be obtained by emailing Colin Mitchell. (talking@tpg.com.au)
Build the circuit and test your transistors and LEDs.
We will be covering more on the action of a transistor and the action of a transformer in the discussion below, but it is important to build the circuit and see it working.
It is your first piece of **TEST EQUIPMENT**.

Questions

1. Identify the letters "c" "b" and "e"
2. What type of transistor is tested in the first set of hollow pins?
3. Put a PNP transistor into the first set of hollow pins and try all positions. Does the red and green LEDs illuminate?
4. When both the red and green LEDs illuminate, what is the approximate voltage across the pair?
5. When you fit a red LED to the test-socket, what is the approximate voltage across it?
6. When you fit a red LED to the test-socket, why does the red LED and green LED on the PC board turn off?
7. Why doesn't the project need an on/off switch?
8. The two coils for the circuit on the left is called a TRANSFORMER. Do the connections of the windings have to be connected to the circuit around a particular way?

to Index

ROBOT MAN

This multivibrator circuit will flash the Robot Man's eyes as shown

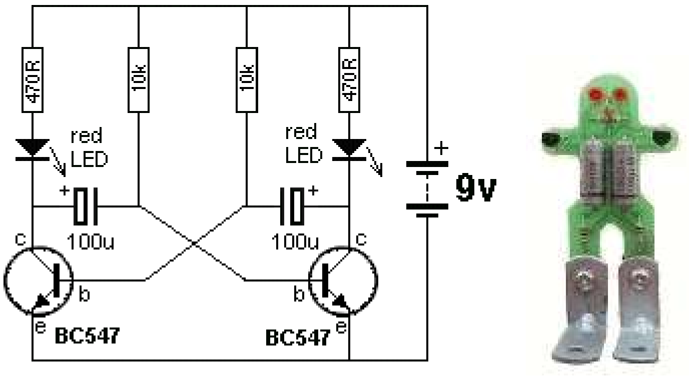

Fig 66. ROBOT MAN
The ASTABLE MULTIVIBRATOR or "free-running" multivibrator.

in the photo. The kit of components is available from Talking Electronics for $8.50 plus postage. Send an email to find out the cost of postage: talking@tpg.com.au
The photo shows the LEDs flashing.
The circuit is called an **ASTABLE MULTIVIBRATOR** and this means it is not stable but keeps switching from one transistor to the other.
It is also called a **FLIP FLOP** circuit.

to Index

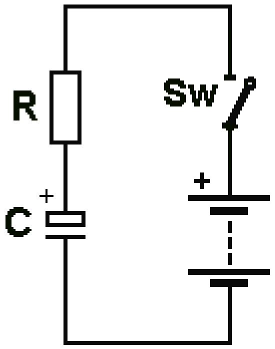

Fig 67. The TIME DELAY

The **TIME DELAY** circuit consists of a Resistor **R** and Capacitor **C** in **SERIES**.
When the switch is closed, the electrolytic (called the CAPACITOR) charges slowly because the resistor only allows a small amount of current to flow.
It's just like charging your mobile phone. The battery takes time to charge because there is a resistor in the circuit to limit the current. If we remove the resistor in the mobile phone, the battery will get too hot when it is being charged but in the **TIME DELAY** circuit, we want the capacitor to charge slowly, because we want a **TIME DELAY**. .

to Index

Fig 68. The charging of a capacitor is the same as building a brick wall.

CHARGING A CAPACITOR

The capacitor in **Fig 67** charges via the resistor **R**.
But the voltage on the capacitor does not rise at a constant rate.
It starts off charging very quickly and as the voltage across it get higher, the voltage increases at a slower and slower rate.
In the photo I am building a brick wall.
I am working at a constant rate.
When I started building the brick wall, I laid 5 rows of bricks (5 courses) in the first hour.
As the wall increased in height, I had to climb the ladder and I could only lay 3 courses an hour and finally the wall was so high I could only lay 1 course per hour.
This is exactly the same as a capacitor charging.
When the capacitor is uncharged, the supply voltage allows a high current to pass through the resistor **R** and the energy quickly fills the capacitor. This results in a rapidly increasing voltage on the capacitor. But as the voltage on the capacitor increases, the difference in voltage between that on the capacitor and the supply is very small and only a small current will pass through the resistor. This means the voltage on the capacitor increases at a slower rate.

to Index

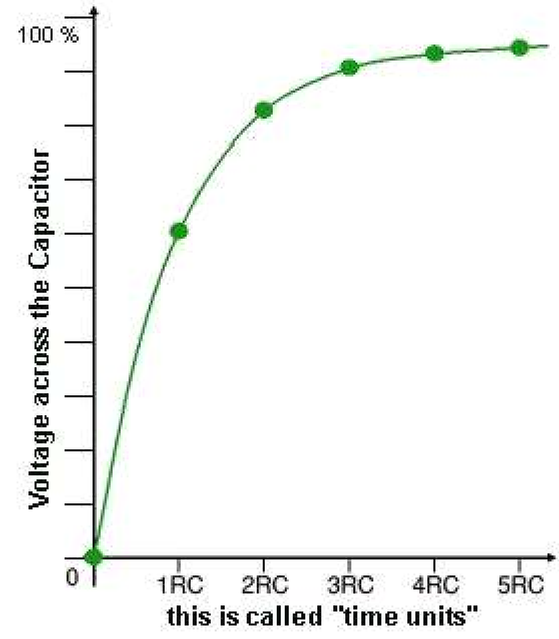

Fig 69. Time Units

It really does not matter how fast or slow or uneven a capacitor charges because most circuits detect a voltage on a capacitor and the time taken to reach this voltage is called the **TIME DELAY**.
But to prevent you thinking the capacitor charges "smoothly" we have to explain what actually happens.
The graph on the left shows the capacitor charging. You can see it charges quickly at the beginning and then charges slowly and then very slowly.
You can see the first part of the graph is fairly "straight" (constant charging) - NOT "straight up and down" but a straight line - and this applies to a voltage of about 63%. The time taken to reach this voltage is called **ONE TIME UNIT** - also called **ONE TIME CONSTANT**. The graph continues for another 4 "time units" (time-constants) and the final voltage is very nearly 100%. (It never reaches 100%.)

to Index

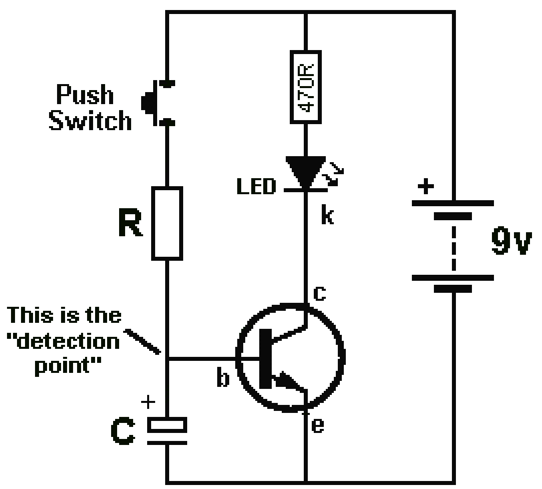

Fig 70. Time Delay Circuit

A **TIME DELAY** circuit needs three things:
1. A Resistor (R)
2. A Capacitor (C)
3. A "Detection Point."

When the switch is pressed, capacitor (**C**) takes time to charge via resistor (**R**) and after a short period of time the voltage at the **DETECTION POINT** is 0.6v and the transistor is TURNED ON. The LED illuminates.

Build the circuit with 100u and 100k and see how long it takes before the LED illuminates.

to Index

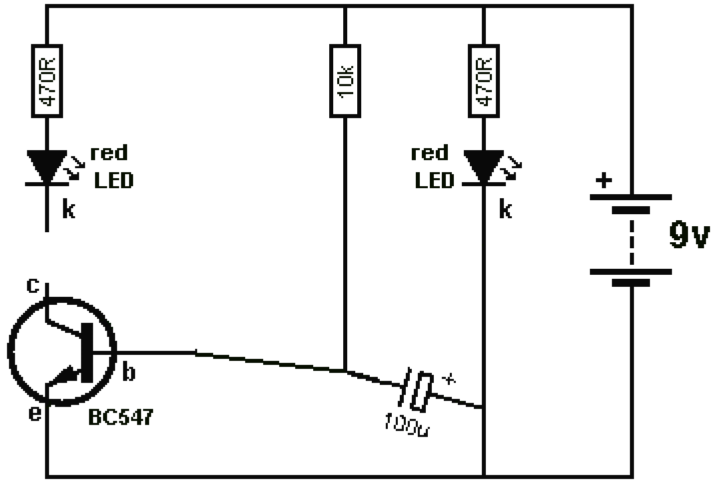

In the **ROBOT MAN** project you can see the "**TIME DELAY**" circuit made up of the 100u, 10k resistor and the base of the transistor.

This is one of the most important **BUILDING BLOCKS** in electronics.
It is the basis of all oscillators and will be discussed below, after we explain a few more details.

Fig 71. The "TIME DELAY" in the ROBOT MAN Project

to Index

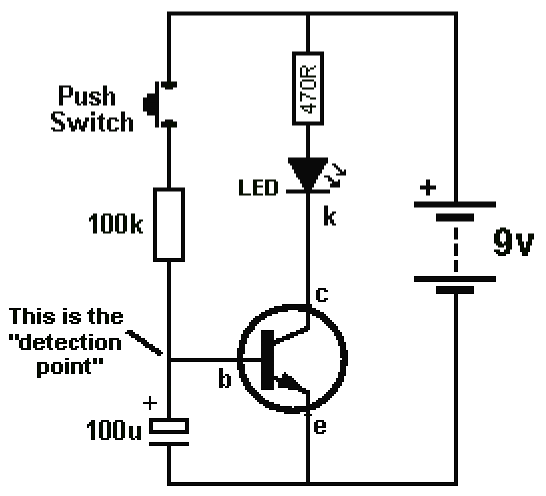
Push
Switch
470R
LED
k
100k
+
9v
c
This is the
"detection
point"
b
+
100u
e

Fig 72. Turning A Transistor ON

We will use the **TIME DELAY** circuit to turn the transistor **ON**.
Make sure the 100u is uncharged by touching both leads (both ends of the capacitor) at the same time with a JUMPER - this is a piece of wire shown in Fig 61.
Push the switch and noting happens. After a short period of time the LED starts to glow and then comes on fully.
This shows two things:
1. The transistor is not turned ON when the base voltage is zero.
2. The base voltage must be 0.6v for the transistor to start to turn ON and when the voltage is 0.65v the transistor is turned **ON FULLY.**

to Index

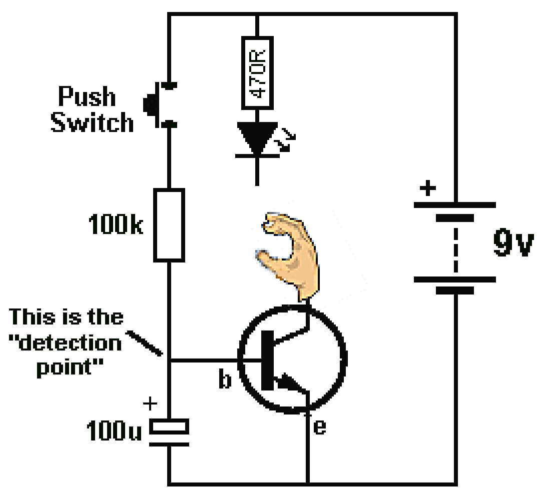
Push
Switch
470R
100k
+
9v
This is the
"detection
point"
b
+
100u
e

Fig 73. How the LED turns ON

Here is an explanation of how the LED turns ON.
When the circuit is first assembled and the switch is not pressed, the transistor is not turned on and it is just like the diagram opposite. The LED is not connected to the transistor.

to Index

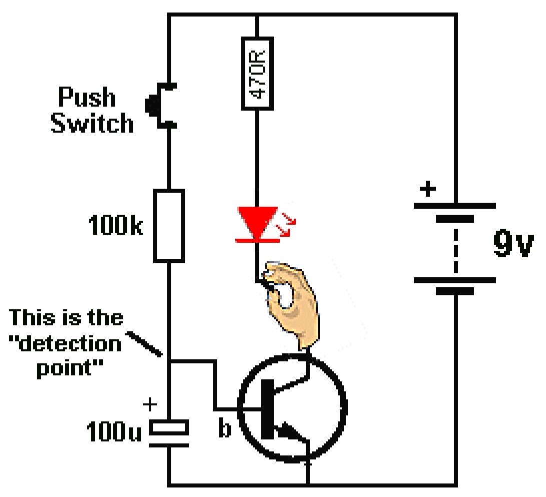
Push
Switch
470R
100k
+
9v
This is the
"detection
point"
+
100u
b

Fig 74. The LED turned ON

When the switch is pushed, the transistor turns **ON** (after a few seconds) and it pulls the lower lead of the LED down towards the 0v rail and this action turns the LED **ON**.

to Index

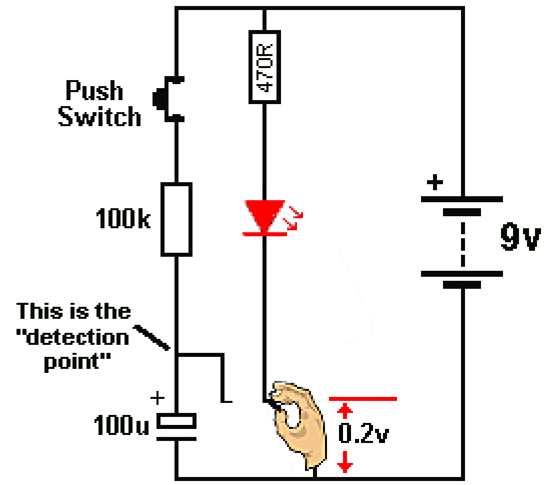

Fig 75. The LED fully turned ON

When the LED is **fully turned ON**, the lower lead of the LED is almost directly connected to the 0v rail.
In other words:
When the transistor is **FULLY TURNED ON**, the lower lead of the LED is almost directly connected to the 0v rail.
The voltage between the lead of the LED and 0v rail is 0.2v. This is the characteristic voltage across the collector-emitter terminals of a transistor when it is TURNED ON.

to Index

In the three diagrams above you can see the LED is changed from an **OFF** condition to an **ON** condition by the action of the transistor.
The transistor is acting **LIKE A SWITCH.**
This action is one of the most important actions in electronics.
It is called: **"The Transistor as a SWITCH"**
It is the basis to ALL Digital Circuits.
It is the basis because of these two facts:
1. When the transistor is **OFF**, the circuit is taking **no current** and no power is being lost or wasted.
2. When the transistor is **ON**, the LED is almost at 0v and no resistor is in the lower lead to waste any power.
Thus we can turn things **ON** and **OFF** without wasting and power.
This is the basis to **DIGITAL ELECTRONICS**.

to Index

DIGITAL ELECTRONICS revolves around circuits that are either **FULLY ON** or **FULLY OFF**. This means they take almost no power and we can combines lots of circuits and still take almost no power.
This means they do not get hot and it also means they will last a long time.
You may not think turning a transistor **ON** and **OFF** will achieve any worthwhile outcome but a circuit can be designed to use two transistors (similar to the **ROBOT MAN** above). The circuit does not Flip-Flop but requires a switch and when the switch is pressed, the circuit changes state. The two transistors are connected together and it takes two presses of the switch to make the output of the second transistor change state ONCE.
The circuit is a divider. It is called a: **divide-by-two** and is the basis of all counting in a computer.
By adding more "**divide-by-two"** circuits we can get "**divide by 4**, **divide by 8**" etc. Two transistors don't do much but when you combine millions of transistors we have a **COMPUTER.**

to Index

Two transistors can do one more thing. They can "**REMEMBER**."
Here is a manual circuit.
Pressing Switch **A** turns the LED **ON** and pressing switch **B** turns the LED **OFF**.
The circuit "remembers" or remains in each state called a stable state.

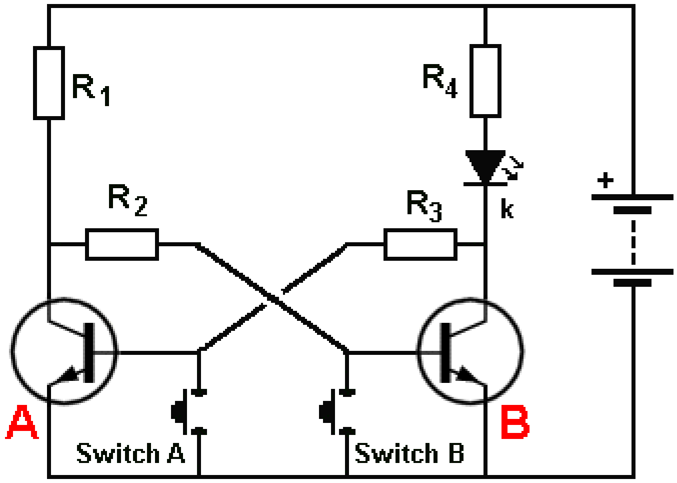

Fig 76. The "MEMORY CELL"

When Switch **A** is pressed, the voltage on the base is removed and transistor **A** turns OFF.
Transistor **B** turns ON via resistors R_1 and R_2 and the LED is turned **ON**.
When the switch is released, the voltage on the collector of transistor **B** is less than 0.6v and the two transistors remain in this state.
Pressing switch **B** turns the LED **OFF.** (transistor **A** turns ON via R_3, R_4 and the LED - very little current flows through the LED and you can hardly see it glowing). The voltage on the collector of transistor **A** is less than 0.6v and the two transistors remain in this state.

The technical name for this circuit is:
BISTABLE MULTIVIBRATOR or **BISTABLE SWITCH** or **BISTABLE LATCH**.

This is the basis to all the memory in a computer.

to Index

In electronics, we talk about the **DIGITAL TRANSISTOR** and **ANALOGUE TRANSISTOR**.
This is just an ordinary transistor (called a Bipolar Junction Transistor) in a **DIGITAL CIRCUIT** or **ANALOGUE CIRCUIT**.
We are now discussing the **DIGITAL CIRCUIT** - The Multivibrator - Astable Multivibrator and Bistable Multivibrator (Memory Circuit).
The **DIGITAL CIRCUIT** has **2 STATES**.
The **ON STATE** and the **OFF STATE**.
It is conducting in the ON STATE and the LED is illuminated.
In the OFF STATE, the LED is not illuminated.
In the ON STATE the transistor is said to be **CONDUCTING** or **BOTTOMED**.
In the OFF STATE the transistor is said to be "**CUT OFF** or "**OFF**."
These two states are reliable and guaranteed. They are not "half on" or "quarter on" or "75% off."
These states are easy to transmit "down a wire." The ON STATE is transmitted as "1" (voltage present) and the OFF STATE is transmitted as "0" (voltage not present).
These are the two **DIGITAL STATES**.

The **ROBOT MAN** is a **DIGITAL CIRCUIT**.
Each LED is ON or OFF.
The waveform on the output of each transistor is called a **DIGITAL SIGNAL**.

The waveform is said to be **DIGITAL** or **SQUARE WAVE**.
The top line of the graph represents the LED OFF.
The bottom line of the graph represents the LED ON. The LED is ON when the collector voltage is LOW because we are pulling the lead of the LED to the 0v rail as shown above.
The circuit changes from one state to the other very quickly and this is called the **RISE TIME**.

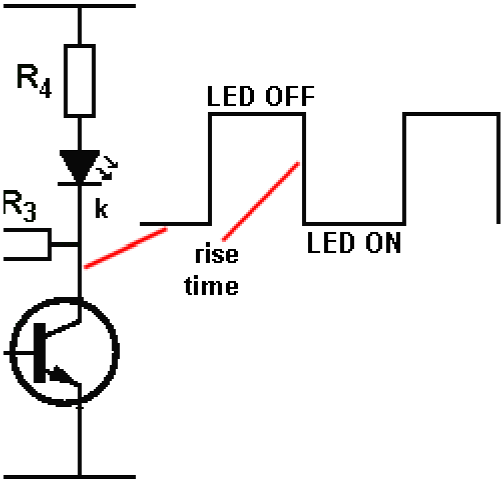
R4
LED OFF
R3
k
rise
time
LED ON

to Index

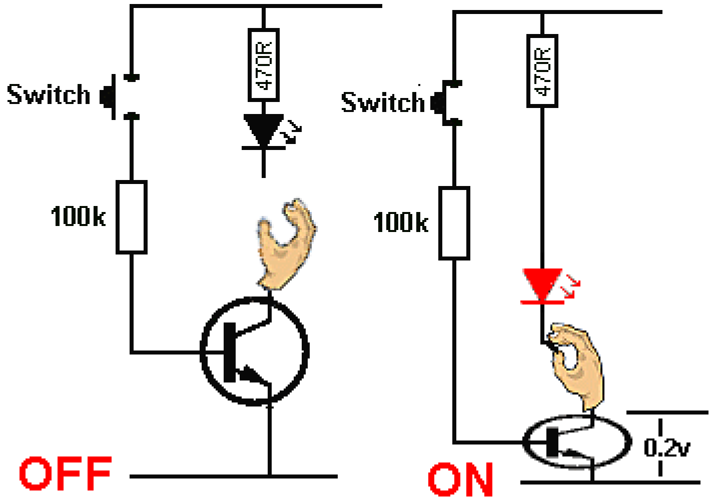
Switch
470R
Switch
470R
100k
100k
0.2v
OFF
ON

Fig 77. The two Digital States

Going over the two **DIGITAL STATES** for a transistor.

In the first diagram the switch is not pressed and the base does not see a voltage to turn the transistor on. The transistor is "**OFF**" (not conducting) and it is not "grabbing" the LED. The LED is not illuminated.
In the second diagram the base of the transistor sees a voltage via the switch and it is TURNED on. The LED is illuminated.

to Index

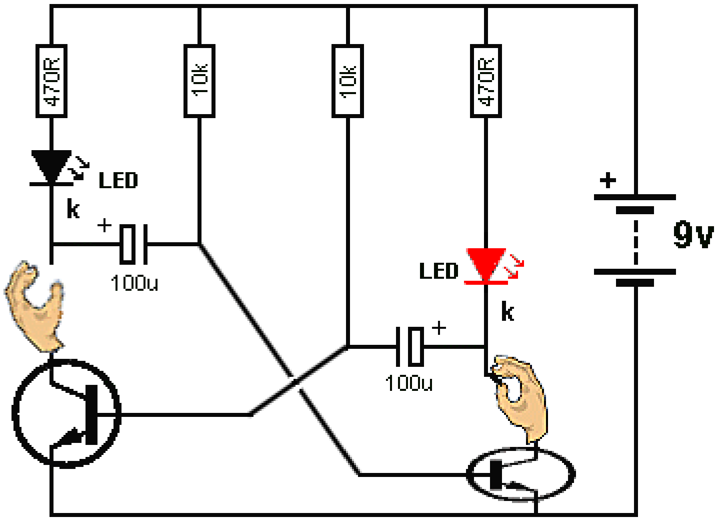
470R
10k
10k
470R
LED
k
+
100u
+
9v
LED
k
+
100u

Fig 78. ROBOT MAN Animation
This animation shows how a transistor grabs a LED and pulls it towards the 0v rail to turn it ON.

to Index

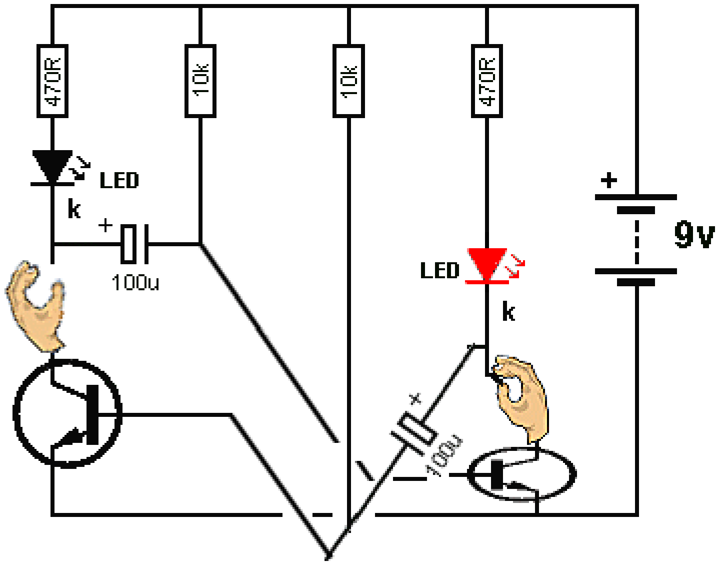

Fig 79. TIME DELAY Animation
The animation in **Fig 78** shows the two transistors turning the LEDs **ON** and **OFF** in a **FLIP FLOP** circuit.
We know the 10k and 100u components form a **TIME DELAY** to create the time for each LED to be illuminated. The timing for one LED plus the other LED creates a **CYCLE** and this is the **FREQUENCY OF OPERATION** for the circuit. It is measured in cycles per second - Hertz - Hz.
We will now go into more detail of how the **TIMING COMPONENTS** create the **TIME DELAY** for each LED.
The circuit is more-complex than you think.
The 100u is already charged from a previous cycle and we show how it gets discharged via the 10k and charged in the opposite direction by the 10k to create a **TIME DELAY**.

to Index

THE CAPACITOR

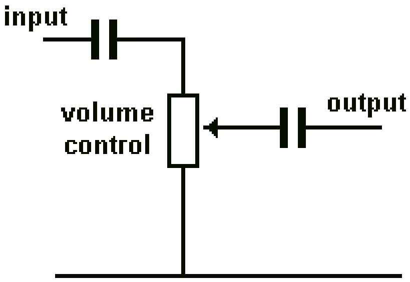

The capacitor can perform many different functions and produce many different effects, depending on its value and the surrounding components.
In this circuit the capacitors on the input and output prevent DC on the volume control creating "scratchy sounds" when the volume is altered.
This is called "DC blocking."
The AC (the signal) passes through the capacitors but the DC voltage on the input is blocked.

to Index

CHARGING A CAPACITOR Part II

It is easy to see how a capacitor charges via a resistor in the **TIMING CIRCUIT** (Delay Circuit) above but many capacitors are not connected to the 0v rail.
They are connected as show in the animation below and their "job" is to pass a waveform.
When they pass the waveform they **CHARGE** and **DISCHARGE**.
The waveform is called an **AC SIGNAL** and the output is smaller than the input.
The circuit is taken from the circuit above, but the same effect applies to all capacitors that "pass a signal."
Here's why:

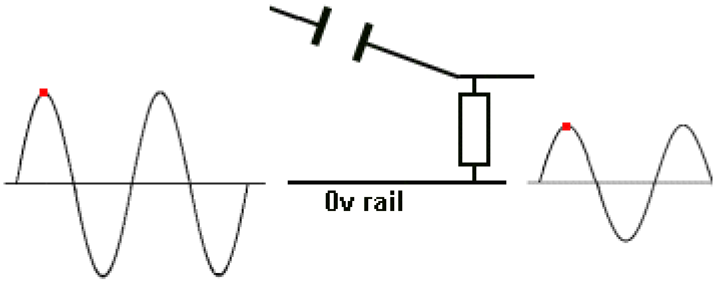

The capacitor charges slightly during the rise of the signal and the right-plate of the capacitor does not rise as high as the left-plate. That's why the output signal is not as large as the input signal.
If the capacitor did not charge, the output would be as large as the input. If you use a capacitor with a large value, it will not charge and thus the output will be as large as the input.
That's why you use a large capacitor !!!!

to Index

CHARGING A CAPACITOR Part III

Here is another CAPACITOR in action.

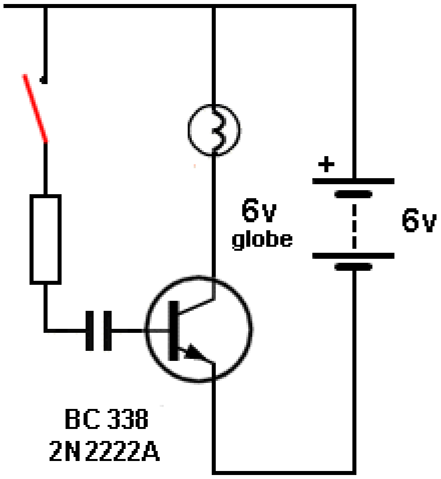

The animation shows a capacitor charging (via a resistor). The initial current is LARGE and this turns the transistor FULLY ON and the globe illuminates. As the capacitor charges, the base current reduces and the transistor starts to turn OFF. Eventually the capacitor is fully charged and the voltage on the base falls to 0v, turning the transistor OFF.

This animation shows three features:
1. The initial charging current is HIGH.
2. It gradually falls to zero.
3.The voltage on the base drops below 0.6v and the transistor turns OFF.

to Index

NEGATIVE VOLTAGE

You will be surprised to learn that many circuits produce a negative voltage or negative spike at some point (when doing circuit-analysis, each location or point or join of components is called a NODE) on the circuit. In other words the voltage will be LESS than the 0v rail of the circuit.
This is due to the presence of a capacitor and the animation shows how a capacitor can produce a negative voltage:

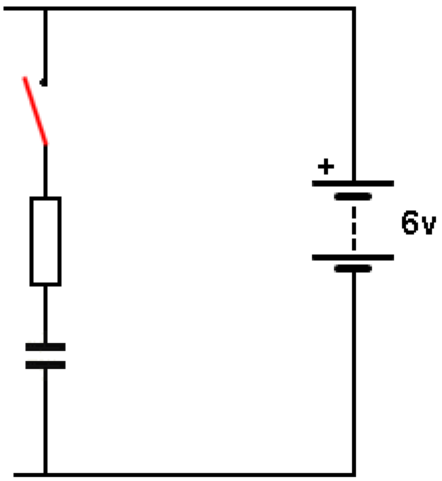

When a charged capacitor is "lowered from one position in a circuit" the positive lead may be lowered by say 3v. This means the other lead will be lowered by 3v. We are assuming the capacitor can be lowered and is not directly connected to the 0v rail.

You can see the electrolytic produces a NEGATIVE VOLTAGE on the base in the following animation, when the two transistors change states:

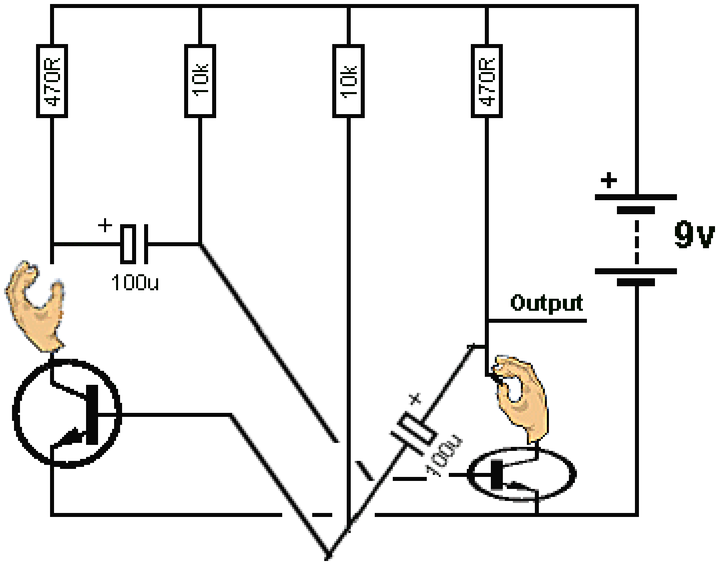

to Index

The Electret Microphone

The most common type of microphone is the ELECTRET MICROPHONE.
It is incorrectly termed the "Capacitor Microphone" or Condenser Microphone." "Capacitor Microphone" descriptions make no mention of a FET as the amplifying device and a polarized diaphragm to detect the audio, so they are something different.
The electret microphone consists of a FET (transistor)

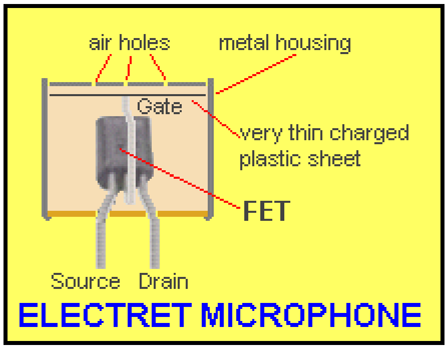
air holes
metal housing
Gate
very thin charged plastic sheet
FET
Source
Drain
ELECTRET MICROPHONE

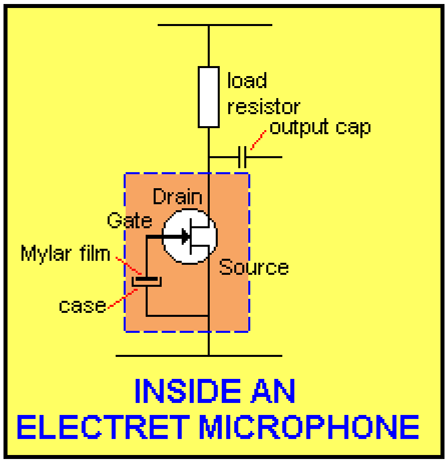
load resistor
output cap
Drain
Gate
Mylar film
Source
case
INSIDE AN ELECTRET MICROPHONE

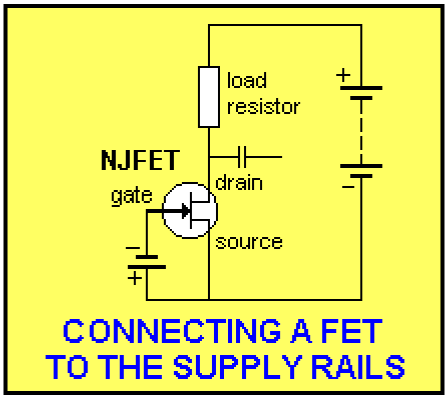
load resistor
+
NJFET
drain
gate
source
–
+
CONNECTING A FET TO THE SUPPLY RAILS

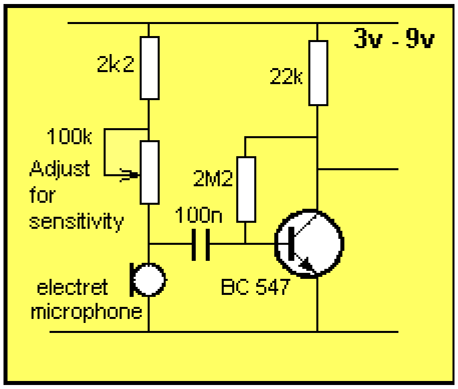
3v - 9v
2k2
22k
100k
Adjust for sensitivity
2M2
100n
electret microphone
BC 547

inside an aluminium case with a very thin Mylar film at the front. This is charged and when it moves (due to the audio it receives via a small hole in the front of the case), it vibrates and sends a very small voltage to the GATE lead of the Field Effect Transistor. This transistor amplifies the signal and produces a waveform of about 2mV to 20mV at the output.
The electret microphone requires about 0.5mA and will operate from 1.5v supply with 4k7 LOAD RESISTOR.
For 3v supply, the Load Resistor can be 22k to 47k.
For higher supply voltages the resistor will be 68k or higher.
Electret microphones are extremely sensitive and will detect a pin-drop at 3 metres.

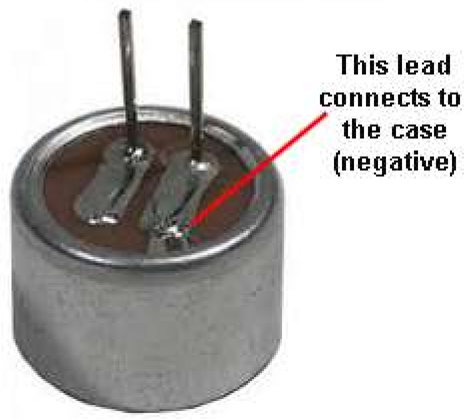
This lead connects to the case (negative)

Most electret microphones have two leads. One lead is connected to the case and this lead goes to the 0v rail. The other lead goes to a LOAD RESISTOR (4k7 to 68k - depending on the voltage of the project).
Reducing the value of the load resistor will increase the sensitivity until the background noise is very noticeable.
They are used in Hearing Aids and are more-sensitive than the human ear.
They are very small, low-cost and very sensitive.

to Index

The Speaker

The most common speaker is about 30mm to 60mm diameter and 8 ohm impedance. This means the voice coil is about 8 ohms resistance.
The two leads can be connected either way to a circuit.
The speaker shown is 32mm diameter and has a realistic wattage of 100mW (NOT 1watt).

These speakers have a Mylar cone and the magnet is a "super magnet" and very small. That's why it is so flat.

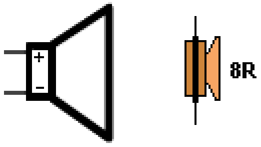

Speaker Symbol

A speaker can be used as a microphone (called a Dynamic Microphone) and a circuit to connect the speaker (mic) to an amplifier can be found on Talking Electronics website. It is not as sensitive as an electret microphone and does not produce the same output amplitude, but it is an emergency microphone.

to Index

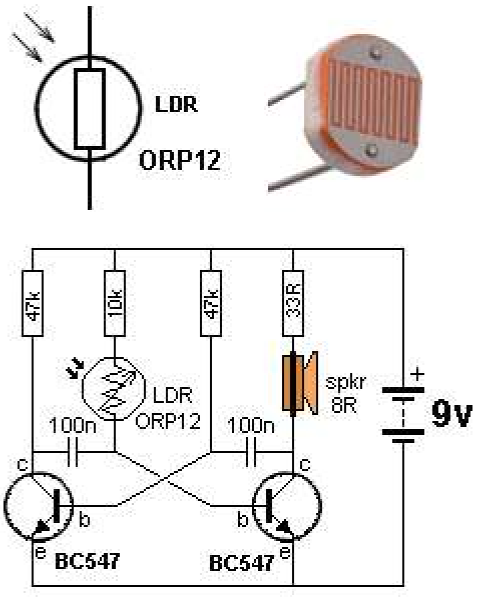

LDR and Light Alarm Circuit

Light Dependent Resistor

(LDR)

Also called **PHOTOCELL** or **PHOTO RESISTOR**
A **Light Dependent Resistor** is a 2-leaded component containing a layer of semiconductor material.
The top contains two interleaving combs of conducting wires with a path of semiconductor material between. When light falls on the component, the resistance of the semiconductor material decreases.
In darkness the LDR will be about 300k. In very bright light the resistance will be about 200 ohms.
But if the light changes only a very small amount, the resistance CHANGE is VERY SMALL. For a large change, see **Photo Transistor**.
The **Light Alarm** circuit will produce a squeal when light falls on the LDR.

to Index

Photo Transistor

The **Photo Transistor** is very sensitive to changes in illumination. It is about 100 times more sensitive than the **LDR**.
The **Photo Transistor** is also available as a **DARLINGTON.**
The Darlington Photo-transistor is 100 x 100 times (10,000) more sensitive than the LDR.

The **Photo Transistor** and **Photo Darlington Transistor** are connected just like a normal transistor but the base lead is not connected. If the value of the LOAD RESISTOR is large, the

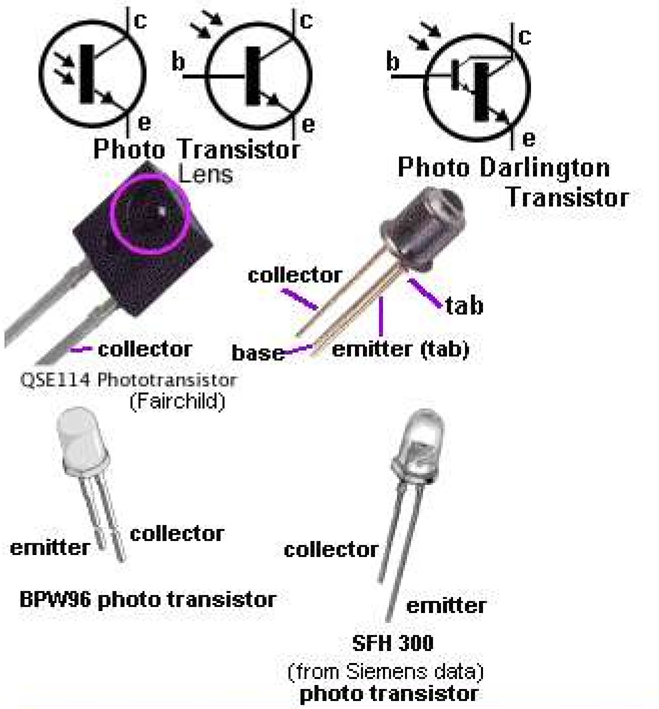
c
b
e
Photo Transistor
Lens
Photo Darlington
Transistor
collector
tab
collector
base
emitter (tab)
QSE114 Phototransistor
(Fairchild)
emitter
collector
collector
BPW96 photo transistor
emitter
SFH 300
(from Siemens data)
photo transistor

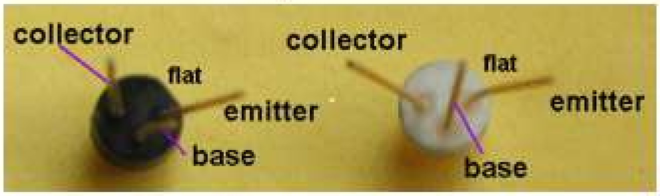
collector
flat
emitter
base
collector
flat
emitter
base

FPT100 Photo Darlington Transistor

The FPT-100 Darlington Photo Transistor is available from Talking Electronics for $1.00 each plus postage.

transistor will not be very sensitive.

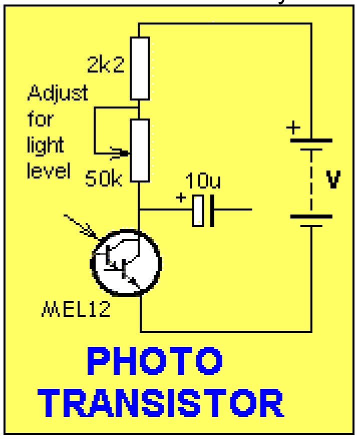
2k2
Adjust
for
light
level
50k
10u
+
+
V
MEL12
PHOTO
TRANSISTOR

to Index

The Inductor

Also called **"coil,"** or **"Choke."**
An **Inductor** consists of one to many turns of wire wrapped around a former (tube of cardboard). The wire can be jumble-wound or wound in layers. The result is the same.
This is called an **air-cored coil** or **air-cored inductor**.
The centre can be filled with a metal such as iron or laminations (thin sheets of metal) or a ferrite material.
Different cores operate at higher frequencies.
The core can be circular (doughnut) or rectangular and it is called a **MAGNETIC CIRCUIT** (when it is a closed loop).
Additional turns or increasing the diameter of the turns will increase the inductance.
A coil with a magnetic core can be used to pick up nails and metal items. It is called an **electromagnet**. It can be

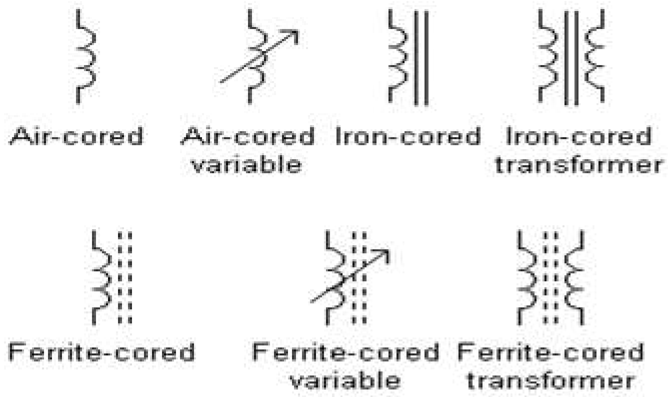
Air-cored
Air-cored variable
Iron-cored
Iron-cored transformer
Ferrite-cored
Ferrite-cored variable
Ferrite-cored transformer

270
surface-mount
inductor 27uH
yes 27uH
270uH = 271

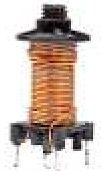

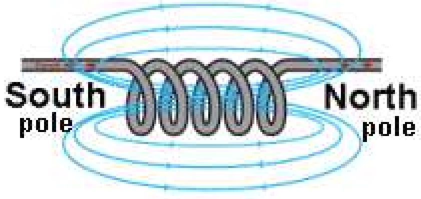
South
pole
North
pole

The emerging **magnetic lines of force** from an inductor produce the NORTH POLE - this is just "convention," a simple way to explain things - to get the explanation started.
When two or more coils are wound near each other, the inductor is called a **TRANSFORMER** (but not the armature above).

operated on AC or DC.
When the metal core is loose and gets pulled into the coil it is called a **SOLENOID** or **ACTUATOR** or **LINEAR ACTUATOR**. It can be operated on AC or DC.
The way an inductor works is very complex but we can say it resists any rise or fall in voltage by turning the rise or fall into magnetic flux.
If the applied voltage is suddenly turned off, the inductor produces **a very high voltage** of **opposite polarity** (these are the **two** most important things for you to remember).

to Index

The Antenna

The first time you will need an antenna is when making an FM transmitter.
The antenna is usually a length of wire equal to half the wavelength of the transmitter. The frequency is about 100MHz and the wavelength is 3 metres.
A half-wave antenna is 1.5metres.
The length is important but the height is more important.
The wire should be as high as possible and "up-and-down" if the antenna on the radio is vertical, to get the maximum range.
This is called an **end-fed half-wave antenna** or **half-wave**

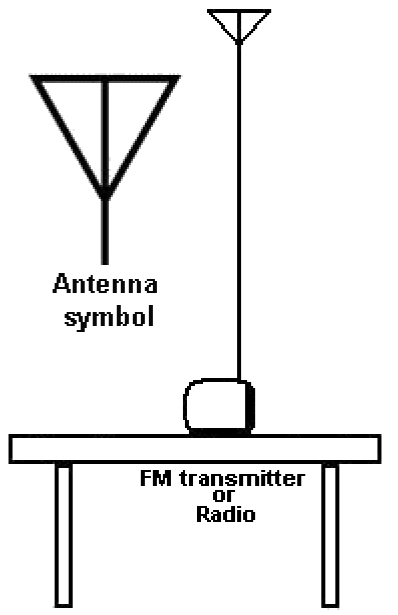

Monopole.
The transmitting circuit should have a good ground-plane such as connection to large batteries so the signal can be pushed and pulled into and out of the antenna.
This signal is then radiated as electromagnetic radiation to the surroundings.

to Index

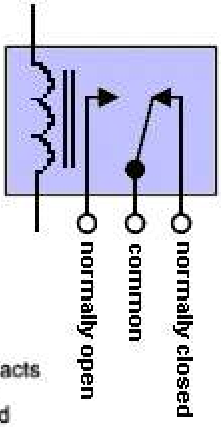

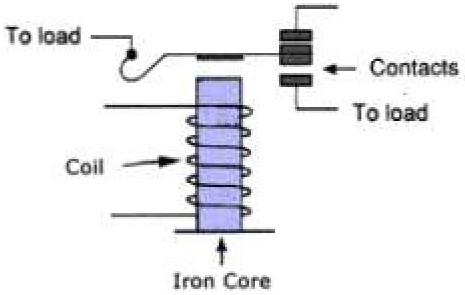

A single contact consists of 2 pins -called SPST (single-pole single-throw).
Or a single pair of contacts can consist of 3 pins - called change-over or SPDT (single-pole double-throw).
A double set of contacts consists of 6 pins, called DPDT (double-pole, double-throw). This is also called a **CHANGE-OVER RELAY** or **REVERSING RELAY** (when connected to a motor).

The RELAY

The word **Relay** comes from the days of Morse Code where a coil of wire (an electromagnet) closed a switch to allow the Morse code to travel further down the telegraph line.
It would "relay" or "pass-on" the information.
A relay allows a "weak circuit" (one with low current) to operate a LOAD that needs a large current. It also separates the two circuits electrically and prevents a voltage such as 240v connecting to a 12v circuit. The coil is separated from the contacts and this gives the two circuits isolation.
A double-pole double-throw relay can be used to reverse a motor as shown here:

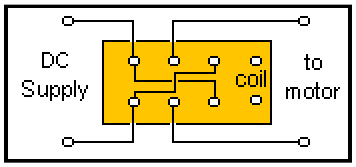

to Index

Driving A Relay

(Powering A Relay)
The first thing you must decide is the voltage of the relay. This will depend on the voltage(s) available. The relay will be driven (activated) by a transistor and the base of the transistor only needs a signal (less than about 1v). This means the project can be operated on a voltage from 3v to 12v and the relay can be connected to a 5v to 12v supply.

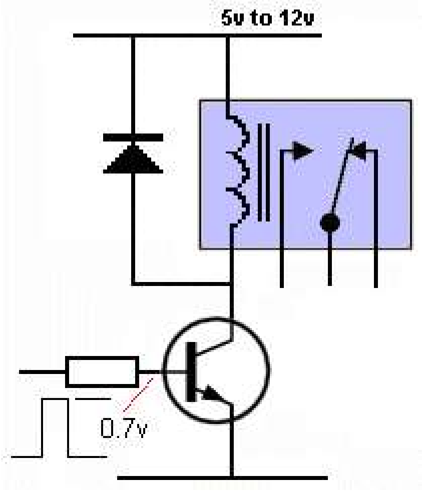

Driving A Relay

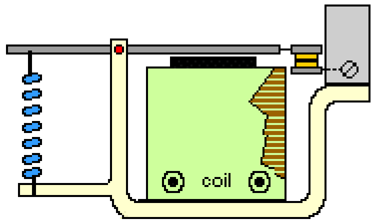

The armature is drawn towards the coil when a current flows through the coil.

Next you need to know the current-rating of the contacts. This will depend on the current taken by the LOAD. The rating of most relays is: 1 amp, 5 amp or 10 amp.
Finally you need to know how many contacts are required.
For a single circuit you will need 2 pins and for two circuits you will need 4 pins (but relays only come with 6 pins).
You can get relays that need a very small current for activation. These are called CMOS relays. But most relays need about 100mA.
To protect the driving-transistor from spikes when the relay is turned off, you will need a diode across the coil.
The top animation shows a "single set of change-over contacts."
The lower animation shows the ARMATURE being drawn to the electromagnet. The electromagnet is the coil with a core of magnetic material that becomes a magnet (an electromagnet) when a current flows through the coil.

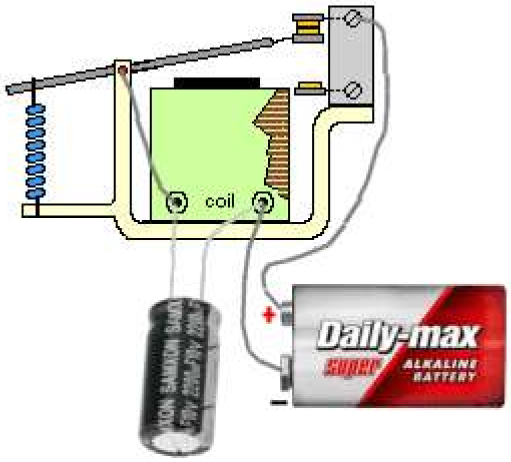

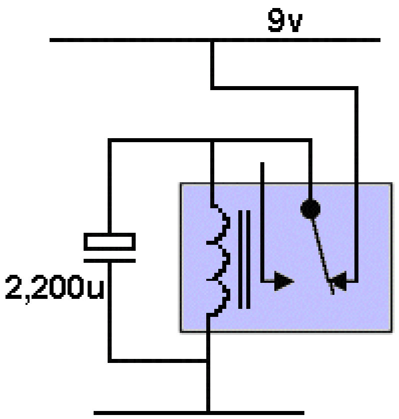

When the circuit is turned ON, the voltage across the 2,200u electrolytic is zero and it gradually charges. When the voltage is about 8v, the coil has enough voltage across it to pull the armature and open the contacts. The electrolytic supplies voltage to the coil for about 1 second and then the electromagnet does not have sufficient magnetism to hold the armature and it returns to close the contacts.

to Index

GATES

Animations of **GATES** and more details of their operation is covered in **DIGITAL ELECTRONICS** chapter.

to Index

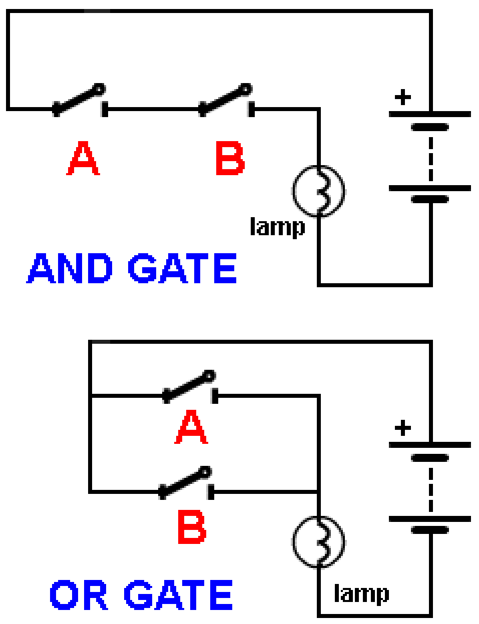

Fig 80. The "AND" GATE and "OR" GATE with switches

The next **BUILDING BLOCK** we will cover is called the **GATE**.
In its simplest form it is an electrical circuit consisting of switches.
Its just two or more switches connected in series or parallel.
We give each circuit a name so we can talk about it and explain its action with a single word.
Later we will cover the electronic version and show how diodes and a transistor are needed to perform a **GATING FUNCTION.**
The type of GATE we are talking about is a **LOGIC GATE**.
The circuit performs an operation called a **LOGICAL OPERATION** on an input or a number of inputs and creates a single output - called a **LOGICAL OUTPUT**.
LOGICAL means "understandable" or "correct" and in this case it means DIGITAL - the signal will rise to full rail voltage or fall to zero voltage. The output will not be half rail or quarter-rail voltage.
The diagram show an **"AND" GATE** and **"OR" GATE** with switches.
For the **AND GATE** close switch A **AND** switch B for the lamp to illuminate.
For the **OR GATE** close switch A **OR** switch B for the lamp to illuminate.

to Index

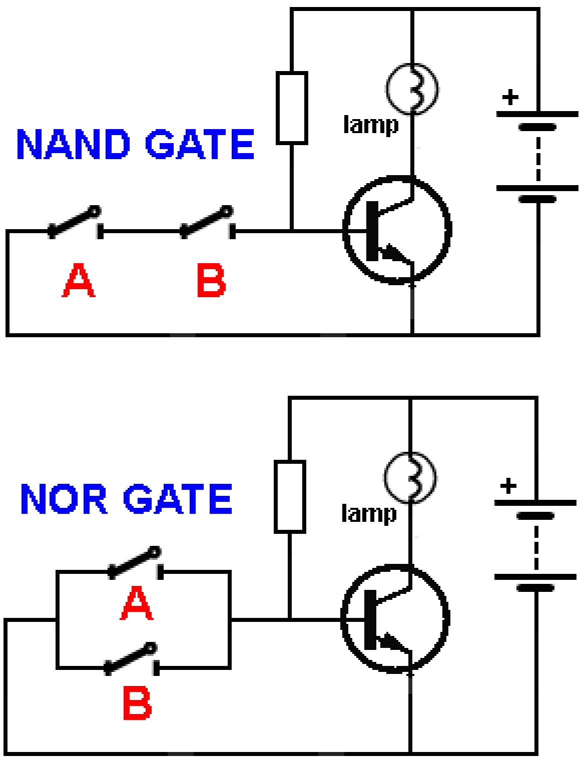

Fig 81. The "NAND" GATE and "NOR" GATE with switches and a transistor

INVERSION

Inversion produces the opposite effect to the results above.
Suppose we want to turn OFF a lamp when one or two switches are pressed.
We need a transistor.
The technical word for Inversion is **NOT**.
It is simplified to the letter "**N**."

For the **NAND GATE** close switch A **PLUS** switch B for the lamp to turn OFF.
For the **NOR GATE** close switch A **OR** switch B for the lamp to turn OFF.

These gates are only **demonstration-gates** to show how one or two switches will turn a lamp OFF.

to Index

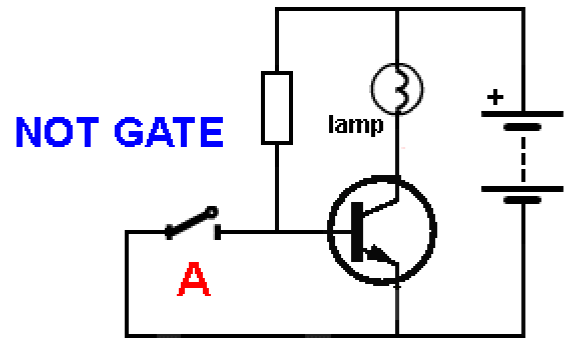

Fig 82. The "NOT" GATE

NOT GATE

A single switch and transistor produces a **NOT GATE.**
This is simply an **INVERSION**.

The resistor turns the transistor **ON** and the lamp illuminates. The switch removes the voltage on the base and the transistor turns **OFF**.
This is only a demonstration circuit to show how a switch can turn a lamp OFF.

to Index

The 5 gates above form the basis to turning a circuit **ON** and **OFF**. We will discuss these gates later in the digital section.

to Index

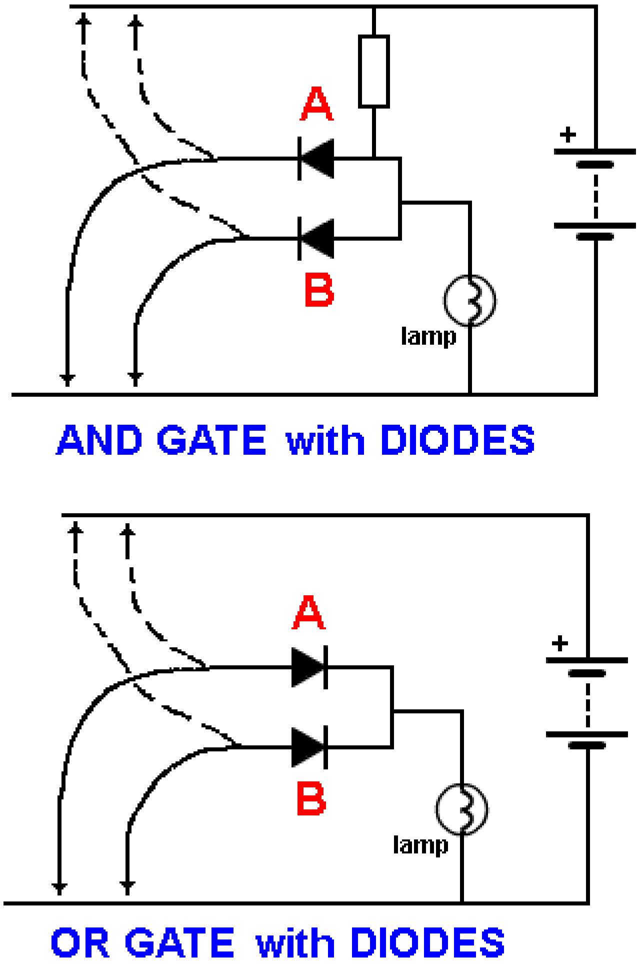

Fig 83. "AND and "OR" gate with diodes

The next building block is the **GATING DIODE**.
We have shown a diode allows current to flow when the diode is correctly placed in a circuit and blocks current when it is reversed.
The 5 gates above are electrical circuits but an electronic circuit works in a slightly different way.
The electronic circuit will be covered later in the DIGITAL section.
For the moment we will explain how a diode can be used to create a **GATE**.
In other words it creates a ONE-WAY PATH to allows signals to pass from one stage to another and prevents signals passing in the opposite direction.
In the AND GATE circuit, both inputs are LOW and current flows through the resistor (it will get HOT). When one input is taken HIGH, current still flows through the other diode and the lamp does not illuminate.
When BOTH inputs are HIGH, current flows through the resistor to illuminate the lamp. No current flows through the diodes.
In the OR GATE, when one input is taken HIGH, current flows through the diode to illuminate the lamp.
This can be done with EITHER input.

to Index

A **NAND** and **NOR** gate can be made with diodes and a transistor.
This time we the output is either HIGH or LOW.
We are gradually producing circuits that

are electronic, rather than electrical circuits.
In the **NAND GATE** circuit, taking one of the inputs HIGH will still allow the other input to prevent the transistor turning ON.
When BOTH inputs are HIGH, the transistor turns on via resistor R and the output is LOW.

In the **NOR GATE** circuit, taking one of the inputs HIGH will turn the transistor ON and the output will be LOW.

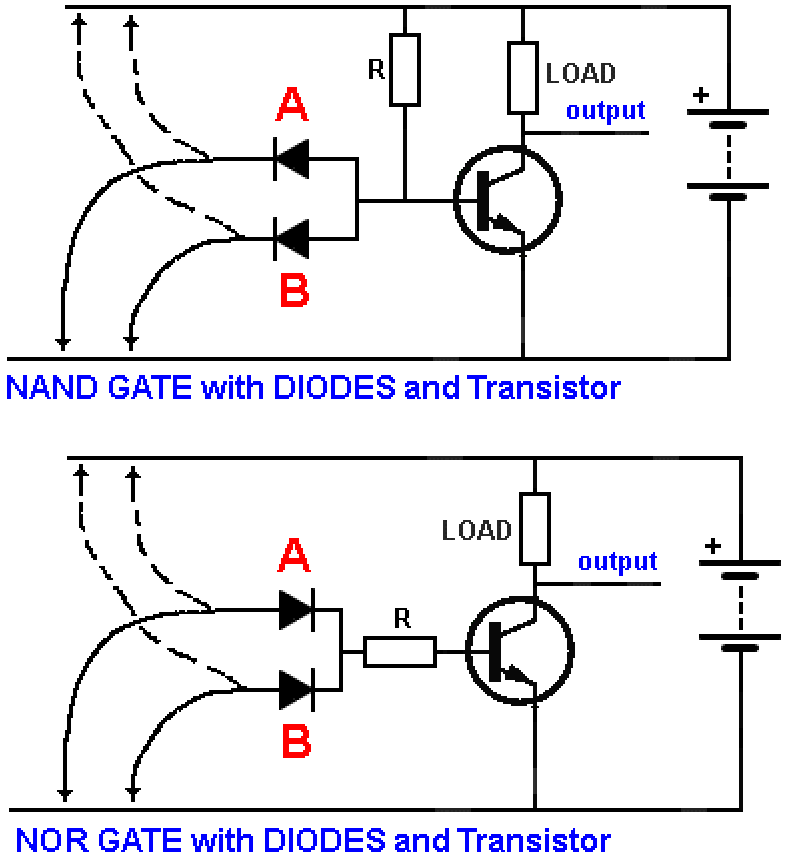
R
LOAD
output
A
B
+
NAND GATE with DIODES and Transistor
LOAD
output
A
R
B
+
NOR GATE with DIODES and Transistor

Fig 84. "NAND" and "NOR" gate with diodes and a transistor.

to Index

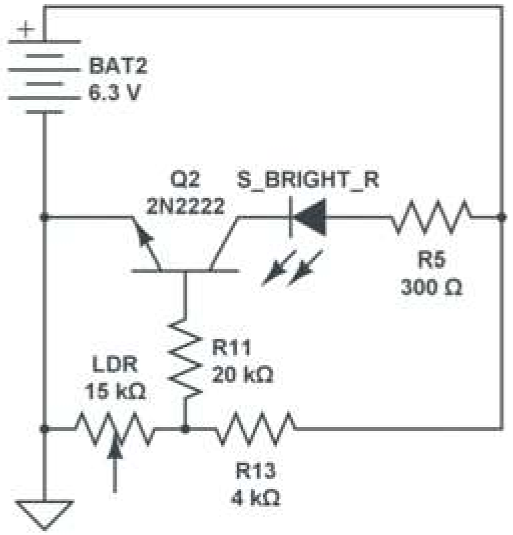
+
BAT2
6.3 V
Q2
2N2222
S_BRIGHT_R
R5
300 Ω
LDR
15 kΩ
R11
20 kΩ
R13
4 kΩ

Drawing A Circuit

A circuit must be drawn according to simple rules so it can be instantly recognised.
An electronics engineer can "**see a circuit working**" when it is drawn correctly and can see if it is drawn correctly; if the parts-values are correct and can use the circuit to assist in diagnosing a problem with a faulty circuit.

The top circuit on is very difficult to visualise because it is not drawn in the normal way.
All the components have to be "**turned around in your mind**," to see what the circuit is doing.

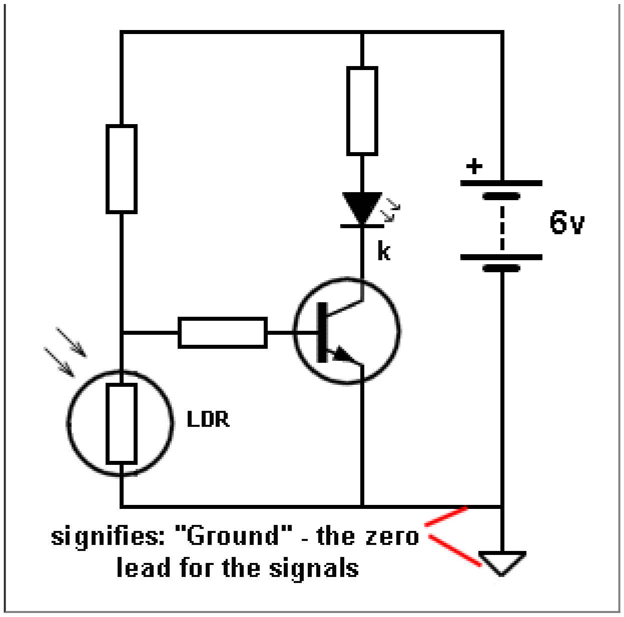

to Index

QUICK QUIZ - to see how much you know

following 50 questions . . . JavaScript is required!
This test will see how much you have learnt.

Easy To Use, All in One, CRM.
1.5 Million Insightly CRM users worldwide.
insightly TRY FOR FREE

FLUKE CONNECT
Capture equipment performance.
FLUKE.
Learn more

Easy To Use, All in One, CRM.
1.5 Million Insightly CRM users worldwide.
insightly TRY FOR FREE

2 Layers $10 ea
4 Layers $25 ea
Order Now
RUSHPCB
ISO 9001:2008, ITAR, IL LISTED
Locates in Silicon Valley, San Jose, CA

☐ An open circuit help

4. What is the approximate characteristic voltage that develops across a red LED?

☐ 1.7v
☐ 3.4v
☐ 0.6v
☐ 5v help

5. If two resistors are placed in series, is the final resistance:

☐ Higher
☐ Lower
☐ The same
☐ Cannot be determined help

6. Which is not a "common" value of resistance:

☐ 2k7
☐ 1M8
☐ 330R
☐ 4k4 help

7. Which value of resistance, placed across a 9v battery will get hot:

☐ 22k
☐ 22R
☐ 220k help

8. If the voltage on the base of a transistor increases, does it:

☐ Turn on
☐ Turn off
☐ Not enough information
☐ Remain the same help

9. The resistor identified in brown is called the:

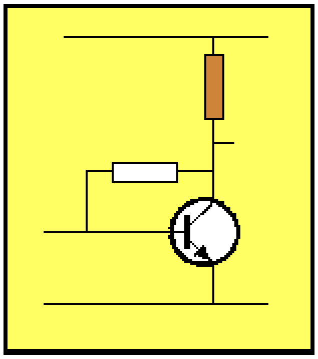

☐ Base Bias Resistor
☐ Load Resistor
☐ Emitter Feedback Resistor
☐ Bypass Resistor

help

10. The first three colour bands on a resistor are: yellow - purple - orange

☐ 47k
☐ 4k7
☐ 470k
☐ 4R7 help

11. A resistor with colour bands: red-red-red-gold, has the value:

☐ 22k 5%
☐ 2k2 5%
☐ 220R 5%
☐ 22R 5% help

12. The lead marked with the arrow is:

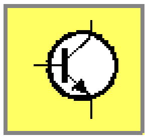

☐ The Collector
☐ The Base
☐ The Emitter
☐ The case help

13. A 10k resistor in parallel with 10k produces:

☐ 10k
☐ 5k
☐ 20k
☐ Cannot be determined help

14. The symbol is:

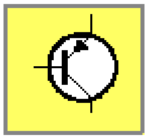

☐ NPN Transistor
☐ PNP Transistor
☐ Photo Transistor
☐ Field Effect Transistor help

15. Two 3v batteries are connected as shown.
The output voltage is:

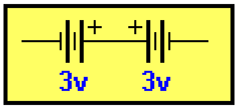

☐ 3v
☐ 0v
☐ 6v help

16. 4 resistors in ascending order are:

☐ 22R 270k 2k2 1M
☐ 4k7 10k 47R 330k
☐ 3R3 4R7 22R 5k6
☐ 100R 10k 1M 3k3 help

17. The closest value for this combination is:

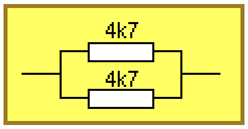

☐ 4k7
☐ 2k3
☐ 9k4 help

18. Which LED will illuminate:

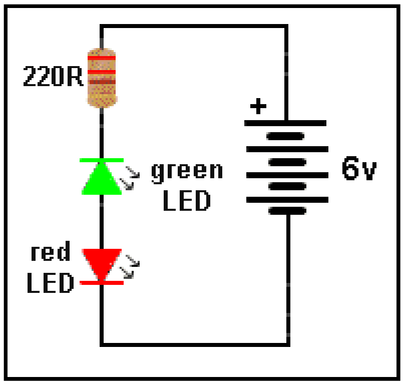

☐ Green
☐ Both
☐ RED
☐ None

19. The four symbols are:

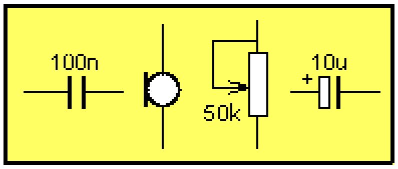

☐ Capacitor, Microphone, Potentiometer, Electrolytic
☐ Electrolytic, Microphone, Resistor, Capacitor
☐ Capacitor, Piezo, Resistor, Electrolytic
☐ Electrolytic, Coil, Resistor, Capacitor help

20. The closest value of the combination is:

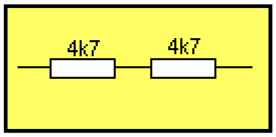

☐ 10k
☐ 2k3
☐ 2k5

21. Which motor will work:

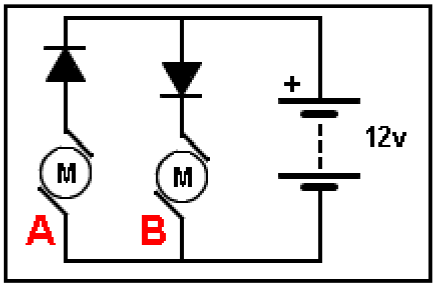

☐ A
☐ B
☐ A and B

22. A resistor and capacitor in series is called a:

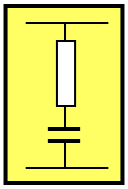

☐ Pulse Circuit
☐ Timing Circuit/Time Delay Circuit
☐ Oscillator Circuit/Frequency Circuit
☐ Schmitt Circuit help

23. A red-red-red-gold resistor in series with an orange-orange-orange-gold resistor produces:

☐ 5k5
☐ 35,200 ohms
☐ 55k
☐ None of the above help

24. Name the 4 components:

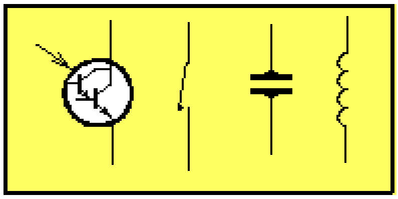

☐ Photo transistor, switch, capacitor, coil
☐ Transistor, mercury switch, piezo, inductor
☐ Photo transistor, reed switch, piezo, coil
☐ Photo Darlington transistor, switch, piezo, inductor help

25. To obtain a higher value of resistance, resistors are connected in:

☐ Reverse
☐ Forward

☐ Parallel
☐ Series help

26. Which LED will illuminate:

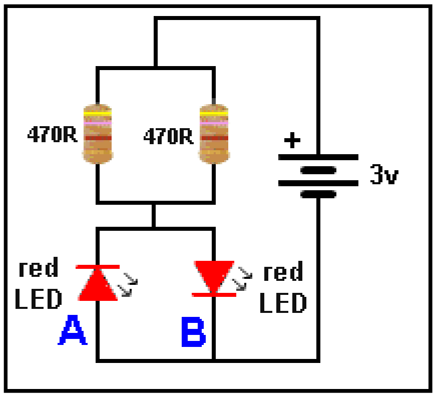

☐ A
☐ B
☐ Both
☐ None

27. Name the component that detects light:

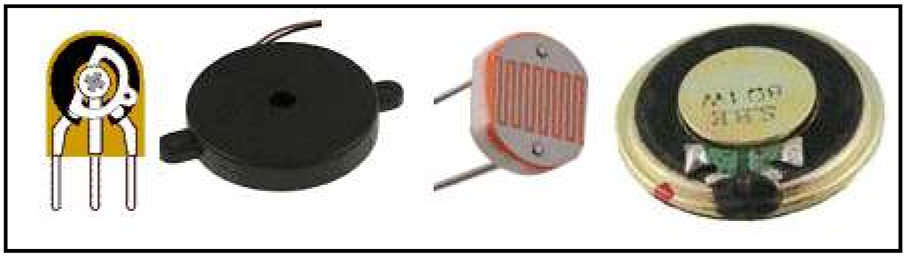

☐ mini trim pot
☐ Light Dependent Resistor
☐ piezo
☐ speaker help

28. What is 1,000p?

☐ 0.01n
☐ 0.0001u
☐ 0.1n
☐ 1n help

29. The current in a circuit is 45mA. This is:

☐ 0.045Amp
☐ 0.00045A
☐ 0.0045A
☐ 0.45A help

30. A 100n capacitor can be expressed as:

☐ 0.1u u = microfarad
☐ 0.01u
☐ 0.001u
☐ none of the above help

31. 1mA is equal to:

☐ 0.001A
☐ 0.00001A
☐ 0.01A
☐ 0.1A help

32. 1,200mV is equal to:

☐ 12v
☐ 1.2v
☐ 0.12v
☐ 0.0012v help

33. The approximate current for a toy 3v motor is:

☐ 10mA
☐ 100mA to 300mA
☐ 1 amp

34. What is the resistance of this resistor:

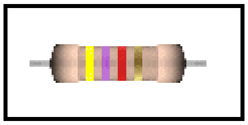

☐ 47k
☐ 4k7
☐ 4R7

35. Identify the correctly connected LED:

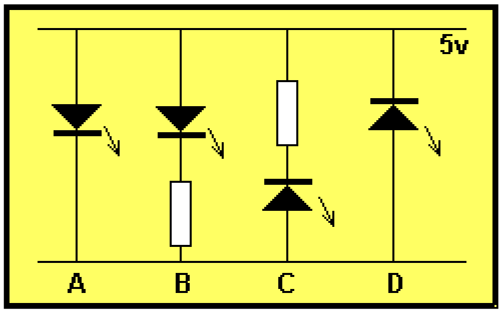

☐ A
☐ B
☐ C
☐ D help

36. Identify the correct statement:

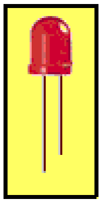

☐ The cathode lead is longer. It goes to the negative rail
☐ The cathode lead is shorter. It goes to the negative rail
☐ The cathode lead is shorter. It goes to the positive rail
☐ The cathode lead is longer. It goes to the positive rail help

37. The current requirement of a LED is:

☐ 1.7mA
☐ 25mA
☐ Between 3 and 35mA
☐ 65mA help

38. The multimeter is measuring . . .

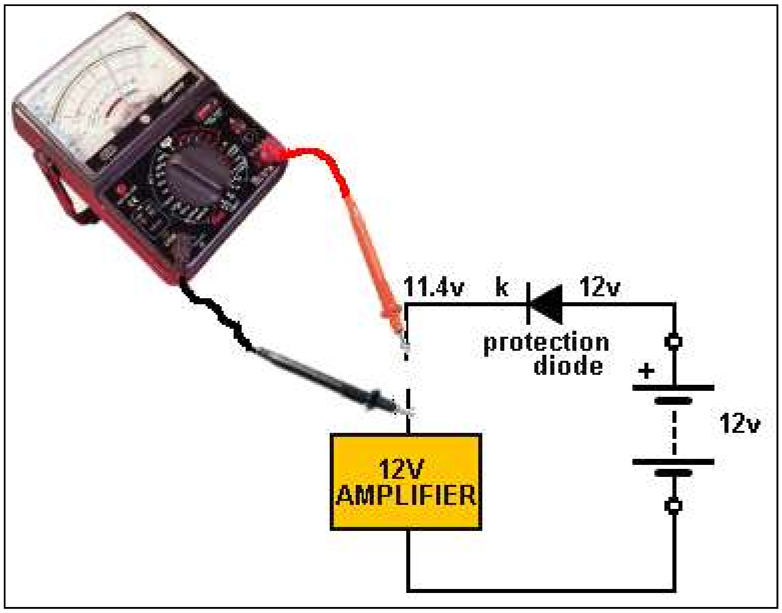

☐ Voltage
☐ Current
☐ Resistance

39. The purpose of the two capacitors:

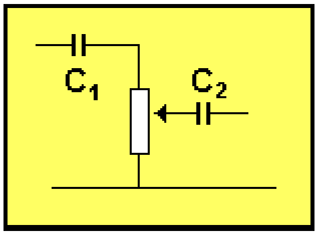

☐ To pass AC from the input to the output
☐ To allow the signal to oscillate
☐ To pass DC from the input to the output
☐ To amplify the signal help

40. The direction of conduction for a diode is:

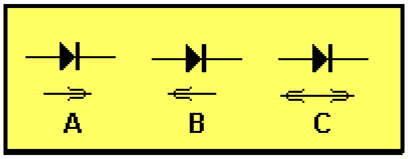

☐ A
☐ B
☐ C help

41. A DC voltage . . .

☐ rises and falls
☐ is a sinewave
☐ remains constant
☐ is an audio waveform help

42. Arrange these in ascending order: k, R, M (as applied to resistor values)

☐ R, k, M
☐ M, R, k
☐ k, M, R
☐ M, k, R help

43. A battery produces AC current:

☐ true
☐ false

44. The tolerance bands: gold, silver, represent:

☐ 5%, 10%
☐ 10%, 5% help

45. 223 on a capacitor represents:

☐ 0.022u u = microfarad
☐ 22n n = nanofarad
☐ 22,000p p = picofarad
☐ All of the above help

46. Arrange these in ascending order: n, p, u (as applied to capacitor values)

☐ p, u, n,
☐ n, u, p
☐ p, n, u help

47. What is the resistance of this resistor:

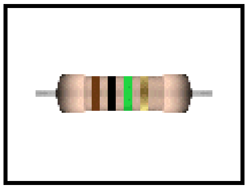

☐ 100k
☐ 1k
☐ 1M
☐ 100R

48. The number "104" on a capacitor indicates:

☐ 0.1u
☐ 100n
☐ 1n
☐ 10n help

49. What is the multimeter detecting:

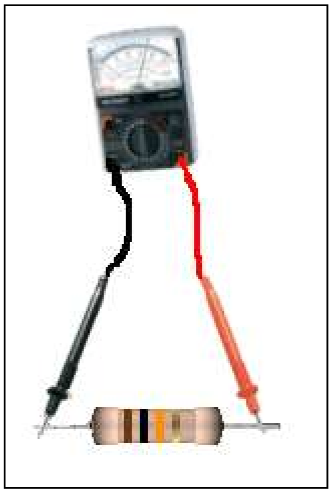

☐ Voltage
☐ Current
☐ Resistance

50. For the LEDs, what is the characteristic voltage for the red and white LEDs:

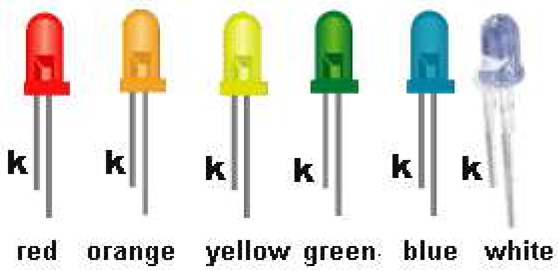

☐ 3.6v, 1.7v
☐ 2.4v, 3.3v
☐ 1.7v, 3.6v

☐ Cannot be determined

Final Assessment

THE CRYSTAL-SET RADIO

100 more Crystal Set plans

Everyone wants to make a RADIO.

The simplest radio is a **CRYSTAL SET**. (Sometimes called a **Crystal-Set Radio** or **Xtal Radio Set** or **Crystal Diode Radio**.)

However a **Crystal Set** needs a number of components that are very hard to get: (tuning capacitor with knob) and (crystal earpiece for $1.25).
The "air" TUNING CAPACITOR (20p to 415p) is not easily available and the postage is expensive. The germanium diode is a special component and the aerial coil wound on a ferrite slab is difficult to obtain.

But you don't need these components. They can be substituted.

There are hundreds of websites on the internet describing the **CRYSTAL SET** and if you want to build a "normal" set, you can Google these sites or 100 more Crystal Set plans. Many of them sell kits too.

But this article is different.

We are going to have all the fun of making a **CRYSTAL SET** but with modern components and easy-to-make components and with an amplifier stage. The output is loud so you don't need a long antenna. And we are going to make our own TUNING CAPACITOR and a very simple aerial coil (called a FRAME AERIAL) as well as replacements for the germanium diode (use a TRF radio IC or a transistor) and in place of hi-impedance headphones (use a piezo diaphragm) and a crystal earpiece equivalent (a piezo diaphragm).
It's even better to have one of each type of component so you can compare the performance, so no matter how many parts your get, nothing will be wasted.

We are also going to explain the fundaments of how the circuit works as even the simplest circuit has a number of very important features that are used in many other circuits.

But first we are going to learn about the components and how they combine to make the circuit work.
When two or more components are connected together they sometimes produce a completely different result to the capabilities of either item.
This is the case with a capacitor and inductor in parallel. An inductor is simply a coil - turns of wire on a cardboard tube - called a former and the centre of the coil is AIR. It is called an air-cored coil or air-cored inductor.
Each component (the coil and capacitor) is called a PASSIVE DEVICE - in other words it does not amplify, but when the are connected together they create a result very near to amplification. And they also produce a result of picking up a huge number of signals and only allowing one signal to appear across the pair. A truly amazing result.

We start by placing a capacitor across the coil.
There is so much activity in the air, from radio, TV, taxi and mobile phone usage that the air is filled with electromagnetic radiation.
This radiation will cut the turns of the inductor (the coil) and produce a microscopic voltage in the turns. This is enough to start the two components passing energy back and forth at a rate

determined by their values. This is the basis of our first discussion.
These two components are called a TUNED CIRCUIT and make up our first building block called THE FRONT END.
.

THE FRONT END

This consists of a coil and capacitor. These two components are in PARALLEL and the signal (called the RADIO SIGNAL or RADIO WAVE) passes through the centre of the coil and produces a voltage in the turns of the coil. The electromagnetic wave has to pass through the centre of the coil.
This voltage is the result of a mass of signals that are interfering with each other and producing a signal called BACKGROUND NOISE.
The voltage can be increased by an external aerial (called an ANTENNA) and it consists of ALL THE LOCAL radio stations (and everything else).
The voltage is a mass of signals and is absolutely useless as it represents all the stations AT THE SAME TIME.
However, across the coil is a capacitor and these signals charge the capacitor with the very small voltage produced by the energy of the signals. When the capacitor is charged, it delivers its voltage to the coil. The coil accepts the energy and converts it to magnetic flux.
After a very short period of time the capacitor becomes discharged and the magnetic flux collapses and produces a voltage in the coil of the opposite polarity to charge the capacitor again in the opposite direction.
These two components keep oscillating back and forth, using the tiny amount of energy from the stations.
There is a natural frequency for the capacitor and coil to pass energy back and forth and one of the stations will provide energy to assist this natural frequency.
When this happens, the amplitude of the signal increases and the signals from all the other stations cancel themselves out and only one signal (waveform) remains.
This signal is called the NATURAL FREQUENCY OF RESONANCE and it corresponds exactly to one of the radio stations.

The end result is a waveform that is the exact same frequency as one of the radio stations and when voice or music is played, the amplitude of the waveform increases and decreases. This will be the signal you hear in the earpiece or speaker.
It is called an AMPLITUDE MODULATED signal and that is where we get **AM RADIO** from.

Understanding the concept of a CAPACITOR and INDUCTOR in parallel is very important. They form a TUNED CIRCUIT that has a natural RESONANT FREQUENCY.
Here is a very similar analogy. You have a heavy metal ball on a long string attached to the gutter on your house - just like a pendulum. You can push the ball very lightly with a finger and after a number of pushes you will be able to get the heavy ball swinging in a very large arc.
The only way to keep it swinging is to push it very lightly at exactly the right time. If you push it at the wrong time it will eventually stop swinging.
The parallel tuned circuit is exactly like the ball. It wants to oscillate at a particular frequency.
All the radio stations are pushing and pulling the circuit at the wrong times and nothing is happening. But one radio station pushes at exactly the right time and the circuit starts to oscillate.
All the other stations are fighting each other just like one person pushing the ball sideways and another pushing the ball from the opposite side. The results cancel each other and you are the only one assisting the swing.
The TUNED CIRCUIT can also be called a FILTER with a very narrow BAND-PASS frequency but our simple explanation describes the operation much more clearly.

There are a few other terms used to describe the components in the font end:

LOOP STICK ANTENNA -This is an alternate name given to the coil of wire wound on a ferrite rod or slab. It also has the name ROD ANTENNA or FERRITE ROD ANTENNA.
The winding can be enamelled wire or flexible wire called LITZ WIRE. This is very fine strands of enamelled wire twisted together and covered in cotton. The purpose of changing a thick wire to lots of very thin wires is to prevent the radio signals creating loops of signals within the wire and these signals will cancel each other and not produce a signal out the end of the wire.
In our experiments, we have not noticed any difference in a coil made with ordinary enamelled wire and Litz wire.

Here is a set of components to make your own Crystal set from **Scott's Electronic Parts**:
You can see the rod antenna, germanium diode, crystal earpiece, capacitor and resistor. The kit costs about $9.00 plus postage and includes knob, clips and screws but no board to mount the parts.

Here is the circuit for a Crystal Set:

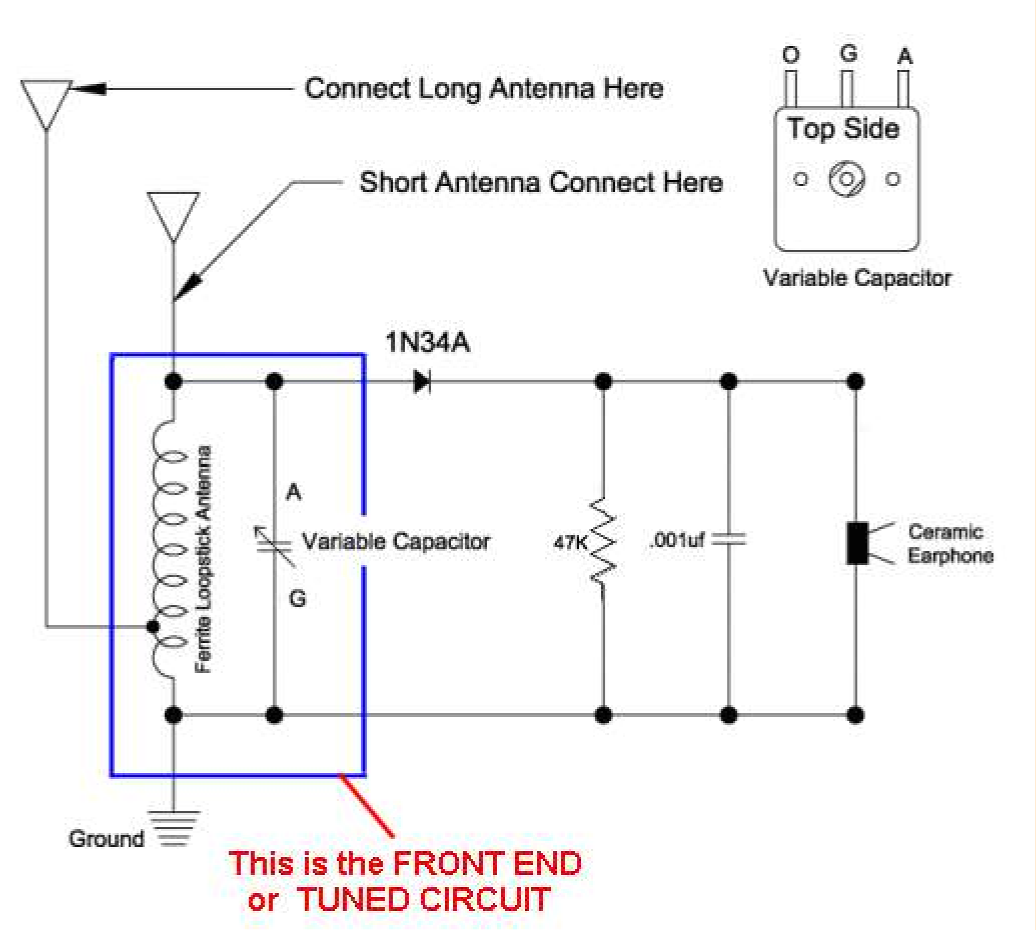
Connect Long Antenna Here
Short Antenna Connect Here
O G A
Top Side
Variable Capacitor
1N34A
Ferrite Loopstick Antenna
A
Variable Capacitor
G
47K
.001uf
Ceramic Earphone
Ground
This is the FRONT END or TUNED CIRCUIT

Here are the components mounted on a board called BREADBOARD:

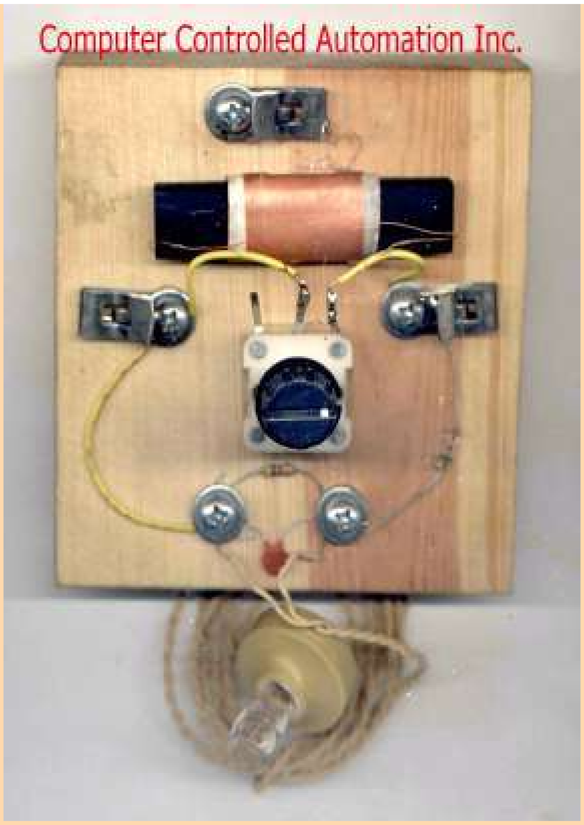

The top clip connects to a long antenna. The left clip connects to ground and the right clip connects to a short antenna.

Let me clear up a point. You do not need a ferrite rod antenna for the coil. You can use an ordinary coil wound on a cardboard tube and it will work just as well if you are using an outside antenna.
Here is the circuit using a home-made coil.

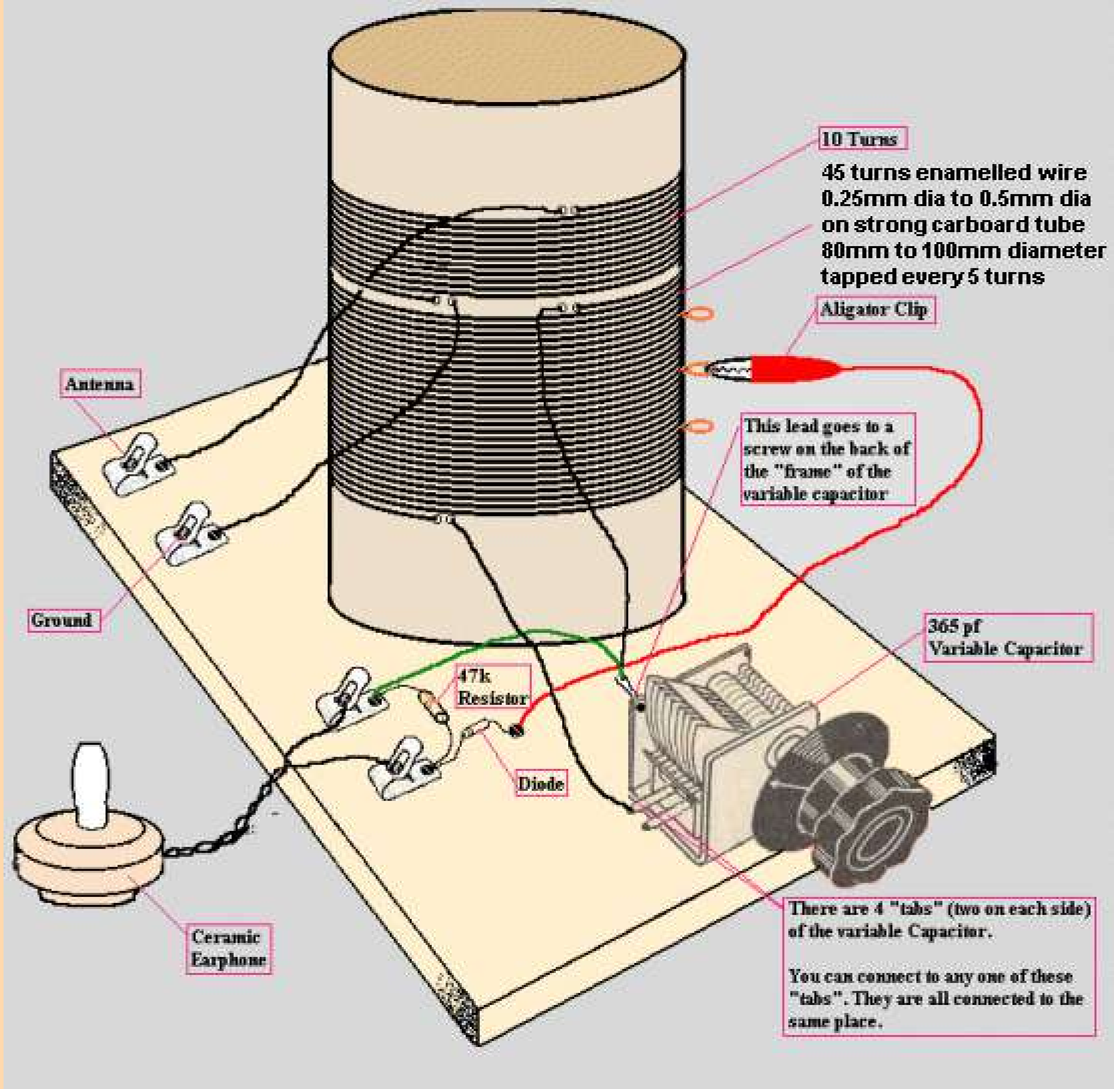
10 Turns
45 turns enamelled wire 0.25mm dia to 0.5mm dia on strong carboard tube 80mm to 100mm diameter tapped every 5 turns
Aligator Clip
Antenna
This lead goes to a screw on the back of the "frame" of the variable capacitor
Ground
365 pf Variable Capacitor
47k Resistor
Diode
Ceramic Earphone
There are 4 "tabs" (two on each side) of the variable Capacitor.
You can connect to any one of these "tabs". They are all connected to the same place.

**The tappings on the coil allow a wide band of radio stations to be tuned.
Each tapping allows a different portion of the band to be covered.**

The next part to understand is this:
The coil and capacitor must not be LOADED. In other words, you cannot connect anything to this combination because the signal it is producing will be "taken away" or "removed" or "considerably reduced" by the item you are connecting to the circuit.
These two components are called a TUNED CIRCUIT and when they are not loaded they pick up all the radio stations, one station at a time, when the natural resonant frequency of the coil and capacitor exactly match the frequency of the radio station. The circuit actually "rejects" all the radio stations except one. Because all the other stations are trying to make the Tuned Circuit oscillate at a different frequency and it does not do this.
The result of the TUNED CIRCUIT oscillating under NO LOAD conditions produces a waveform that is very high and this gives the circuit GOOD SELECTIVITY. The circuit can select one station and reject nearby stations.
It also has good SENSITIVITY as it can pick up weak stations.
If you load the circuit, only the strongest signal will be detected and it will be spread across the full range of the tuning capacitor.
Obviously the theory is more-complex but we are explaining the end-result.
Theory talks about the "Q" value of the coil and this is its ability to produce a very good output when the magnetic flux collapses and the "Q" value increases when the circuit is not loaded. .
Although these voltages are very small (in the order of microvolts or millivolts) the result is very important as the rest of the circuit will be amplifying this waveform a few thousand times.
As we explained above, pushing the weight on a string only needs a push of 1 cm and eventually the weight will swing 1 metre. This is a gain of 100:1 The same thing happens with the tuned circuit. The incoming radio signal is in the order of microvolts, but the coil and capacitor will produce a signal as high as 500 millivolts. This is an improvement or "gain" of more than 1,000 and is referred to as the "Q" of the circuit.
You will also notice the TUNED CIRCUIT is not connected to any supply voltage. It does not have

be connected. It generates its own waveform from the signals in the air.
It should not have any DC current flowing through it via the supply as this would put a load on the circuit and reduce its operation.

However we must "pick-off" the signal so it can be amplified.
This must be done with a very high impedance circuit.

THE CONVERTER (detector) - THE DIODE

The next part of the circuit is the CONVERTER. Commonly called the DETECTOR.
It converts the RADIO FREQUENCY signal to an AUDIO FREQUENCY signal.
This is the job of the DIODE.
The radio frequency signal is a very high frequency signal (say one million cycles per second) and it is sending a tone of one thousand cycles per second through the air-waves.
What is happening is this: The one megahertz signal has a certain amplitude and over a range of the first one-thousand cycles, the amplitude gradually decreases and then increases again. If you look at the tops of this 1,000 cycles you will see a waveform that corresponds to the one kilo-Hertz signal.
The 1MHz signal is picked up by the coil and capacitor in the front end and makes it oscillate.
The radio frequency signal is gradually getting larger over 500 cycles then smaller over the next 500 cycles and this increase and decrease represents the 1,000 cycles per second tone.
This is the waveform (the signal) that passes through the diode. This will be explained further in a moment.
The diode does not pass any signals less than 200mV as the first 200mV is lost in the junction of the diode. This means the signals start to appear on the other end of the diode when they are above 200mV.

This is how the diode works:
Across the crystal earpiece is a capacitor. The capacitor gets charged via the diode.
The diode is present to stop the capacitor getting discharged when the waveform is in the wrong direction. (by this we mean - when the waveform is lower or smaller in amplitude than the voltage on the capacitor).
And the waveform is in the wrong direction about 50% of the time. To charge the capacitor for one-half-cycle requires 500 "little increments" in voltage with each increment adding a microscopic increase in voltage. We don't want this voltage to reduce when the waveform is reversing direction and the diode stops the voltage flowing back to the Tuned Circuit.
During the next half of the cycle when the pulses are getting smaller and smaller, the voltage on the capacitor is "bled off" by the load resistor.
The crystal earpiece detects this voltage. What we mean, is the diode allows the voltage to rise (increase) on the capacitor via lots of little "pulses" and the voltage increases in the form of a sinewave to a maximum amount. This voltage is passed to the crystal earpiece.
Once the voltage rises to a maximum, the little pulses of energy are not quite as strong, and the voltage on the capacitor reduces to form the second portion of the sinewave. This voltage is always being passed to the crystal earpiece and you can hear it as an audio signal.

GERMANIUM OR SILICON DIODE

The preferred type of diode for a Crystal Set is germanium. This is because it drops only about 0.3v.
But a silicon diode can be used, even though it drops about 0.7v, if the radio stations are very loud (close by).
You have to remember, you need a very good aerial (and a water-pipe earth) to get any results with a Crystal Set because you are asking the signal to provide the energy to drive the earpiece.
By simply adding a transistor, you are improving the performance 100 times and the long antenna can be reduced to a FRAME ANTENNA and the earth can be the metal frame of your soldering iron.

THE EARPIECE or EARPHONE
also The Magnetic Earpiece or CRYSTAL EARPIECE

The earphone or earpiece used in a Crystal Set must be a high impedance device because the crystal set does not produce a high current and cannot drive a low-impedance earpiece. That's why a CRYSTAL EARPIECE is ideal.
It has a crystal glued to the back of the earpiece and connected to its top surface is an aluminium diaphragm. When the crystal expands and contracts as a result of a voltage applied via two electrodes, the diaphragm moves and you can hear the signal. It exhibits a very high impedance

because it consists of a crystal and no coil of wire is contained inside the case.
If you do not have a Crystal Earpiece, you can make your own from the shell of an 8 ohm earpiece and a piezo diaphragm. Only the front part of the earpiece is used.

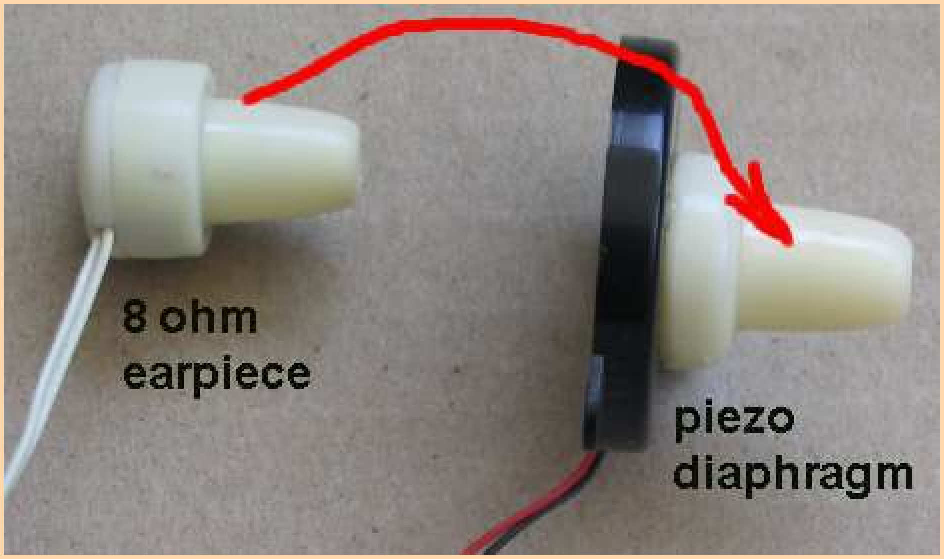

Make your own Crystal Earpiece

Hit the 8 ohm earpiece on the side and the front comes off. Glue the front onto a piezo diaphragm with hot-melt glue. See photo above.
The piezo diaphragm is a ceramic substrate that deflects in the presence of a voltage. It is quite sensitive and you can hear the audio quite clearly.
The waveform emerging from the diode in a Crystal Set is called AUDIO and although it has an amplitude of a few hundred millivolts, it does not have any current associated with it. The crystal earpiece and the piezo diaphragm react to this voltage.

THE 8ohm EARPIECE

The 8 ohm earpiece can be used with our 8ohm Buffer stage shown below.

16ohm 32 ohm and 64 ohm EARPIECE(s)

Earpieces and headsets from mobile phones are 16 ohm or 32 ohm per earpiece and are terminated via a stereo 2.5mm or 3.5mm plug. The earpieces are connected in SERIES to get the best coupling to our radio circuits and you need to find the two pins on a stereo socket to produce series connection. Get a multimeter and switch to "ohms." Try all the pins and you will get a click in the left ear then the right ear. Keep searching until you get a click in both earpieces at the same time. Use these two pins.

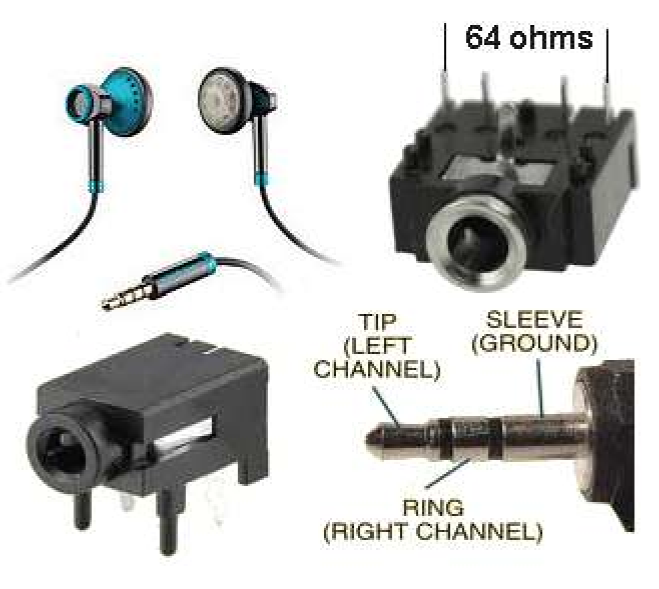

Stereo mobile phone headset - unusually 32R or 64R

PROBLEMS

The biggest problem with a Crystal Set is the need for a long antenna.
The first 200mV to 300mV of a signal is lost in the diode and you need a long antenna to pick up a signal so the output of the TUNED CIRCUIT has enough voltage to drive the high-impedance earpiece.
This requires an outside aerial 5 metres long and 3 metres high.
This is not practical for most hobbyists so we will be adding an amplifying stage to the crystal set so a shorter (smaller) aerial can be used.

THE FRAME AERIAL or FRAME ANTENNA or FRAME COIL

The aerial coil shown in the photo above is a ferrite slab with about 80 turns of Litz wire. You can find one of these in an old broken AM radio or from a parts-shop. In the instructions below we show how to make your own Ferrite Rod Antenna
An equally-good substitute is a frame antenna made by winding insulated wire in a rectangle around wooden sticks.

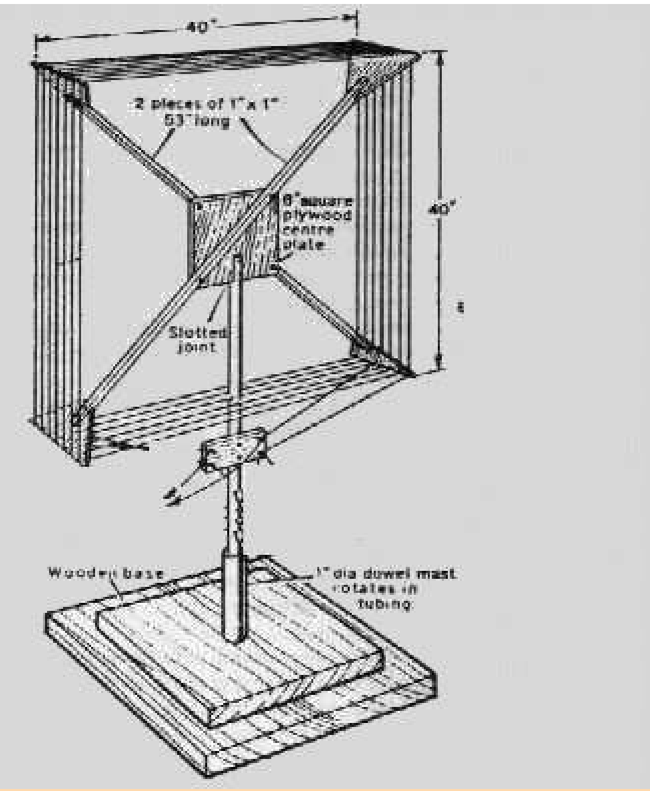
40"
2 pieces of 1" x 1"
53" long
8" square
plywood
centre
plate
40"
Slotted
joint
Wooden base
1" dia dowel mast
rotates in
tubing

FRAME ANTENNA

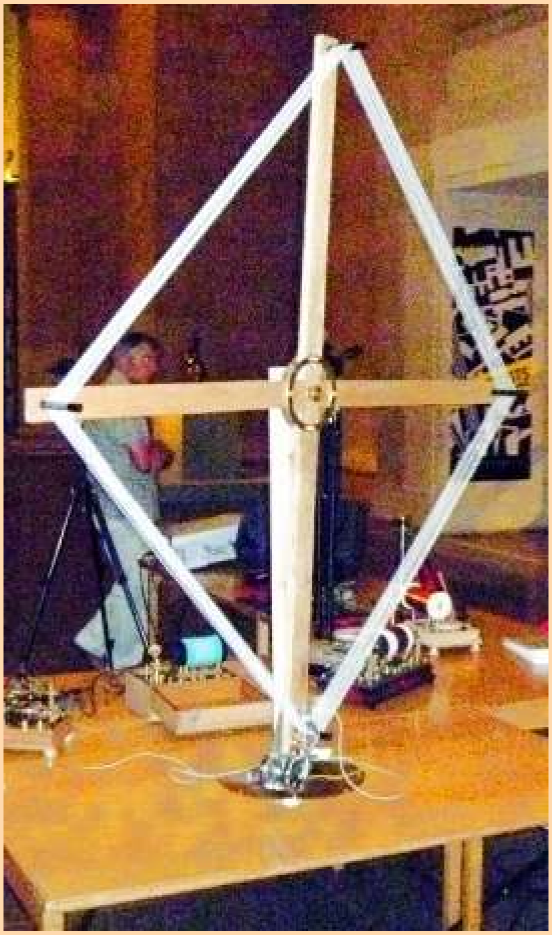

15 turns on a diamond frame

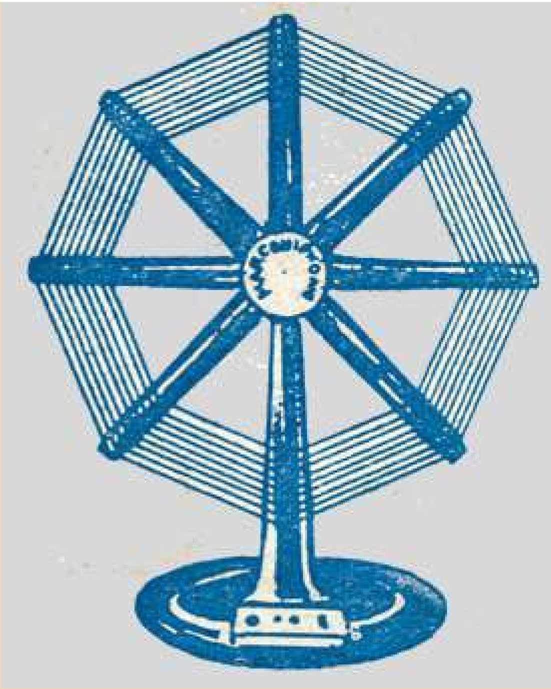

One of the earliest Frame Antennas

The Frame Aerial can be as large as 100cm x 100cm or as small as 10cm x 10cm around a plastic chocolate box.

Here are two FRAME ANTENNAS:

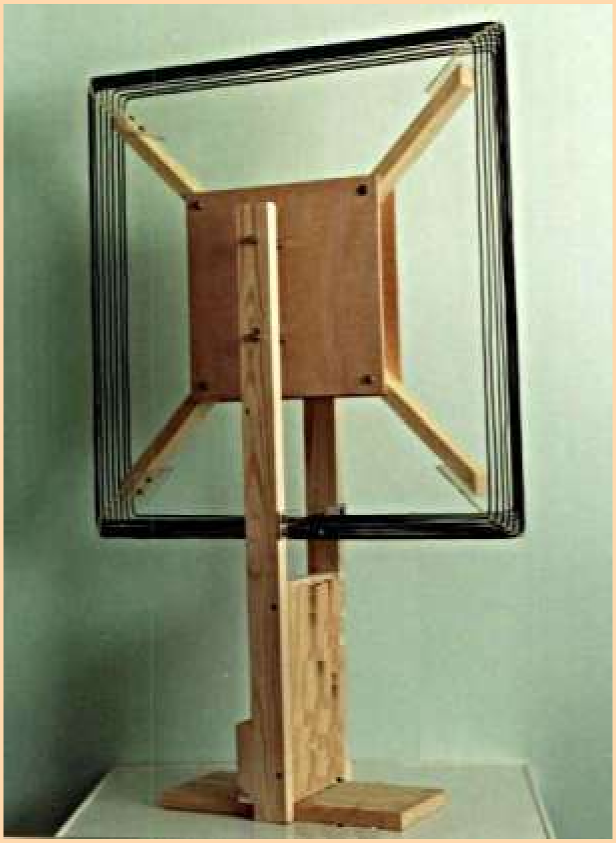

This will work just as good as a ferrite slab antenna. The slab antenna is just 100 times smaller. The slab antenna was invented so a transistor radio could be built in a small case. But if it is not available, you can wind 20 turns around a plastic chocolate box and it will work just as good.

Alternately you can wind 20 turns around a biscuit tin. Put a pencil on the tin and wind the turns over the pencil too. Remove the pencil and it will be easy to remove the turns. Use tape to keep the turns together.

The FRAME AERIAL does two things. It picks up the radio waves and it becomes the coil (called the INDUCTOR) in the TUNED CIRCUIT. It must be placed away from metal objects, such as a refrigerator.

BASKET WEAVE COIL

There is no point making a complex BASKET WEAVE COIL as it will not work any better than simply jumble winding all the turns at the maximum circumference of the coil, because the energy capturing capability of the coil relies entirely on the amount of flux lines passing through the centre of the coil.
By increasing the centre of the coil, the amount of flux is increased for the same coil size.
In fact, the simplest and cheapest is to wind turns around a box, as explained later in this article, or make a frame antenna as shown above. Technically speaking, a round coil has the best performance but only by a few percent.

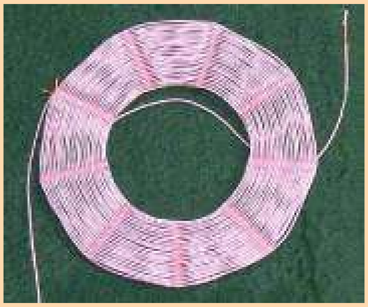

A BASKET WEAVE COIL

THE VARIABLE INDUCTANCE TUNING COIL

Whenever the size or shape of the coil is changed, (or the number of turns), the natural frequency of the Tuned Circuit will change and a different radio station will be picked up.
This means tuning across the band can be done by altering the characteristics of the coil while keeping the value of the capacitor fixed.
Changing the inductance can be done in many different ways.
The coil can have taps every 5 turns and an alligator clips selects the correct tap. But very few radio stations will correspond exactly to each tap.
Another way is to have a slider move up and down the turns as shown in the following image:

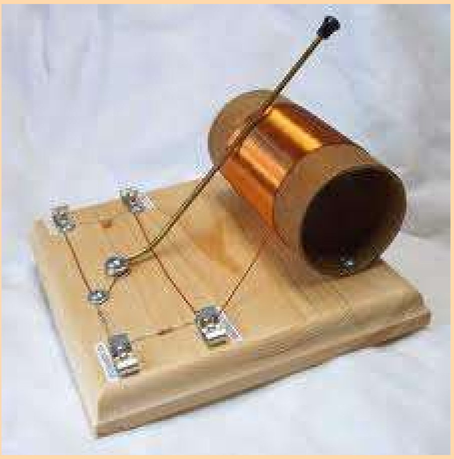

The slider makes contact where the insulation has been removed. But it may touch two turns at the same time and create a "shorted turn" and reduce the "Q" of the coil.

Another way is to move a ferrite bar (rod) in and out of the coil:

THE SLUG TUNED COIL

To tune across the radio band, the natural frequency of oscillation of the TUNED CIRCUIT must be adjusted (changed). This can be done by changing the value of the capacitor or the value of the inductor.

The value of the inductor can be changed by adding or removing turns or changing the amount of magnetic material in the centre of the coil.
A ferrite bar can be screwed in and out of the coil or slid in and out and this component is called a **SLUG TUNED COIL**.
The following diagram shows a **SLUG TUNED CRYSTAL SET**:

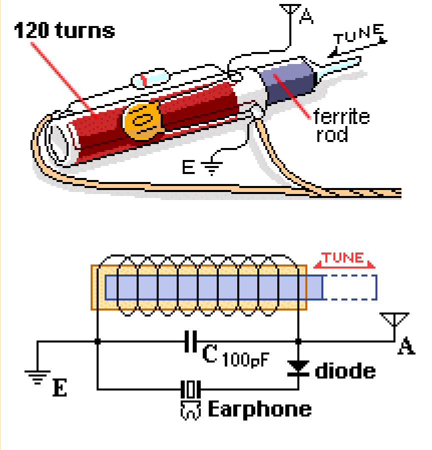
120 turns
A
TUNE
ferrite
rod
E
TUNE
C 100pF
diode
A
E
Earphone

By changing the value of the 100p capacitor, different parts of the band can be picked up.

The photo shows a slug tuned coil using 60 turns of insulated wire on a 10mm tube (or any tube that will fit over a 8-10mm ferrite rod) and a circuit containing an AM radio chip plus a buffer driver transistor:

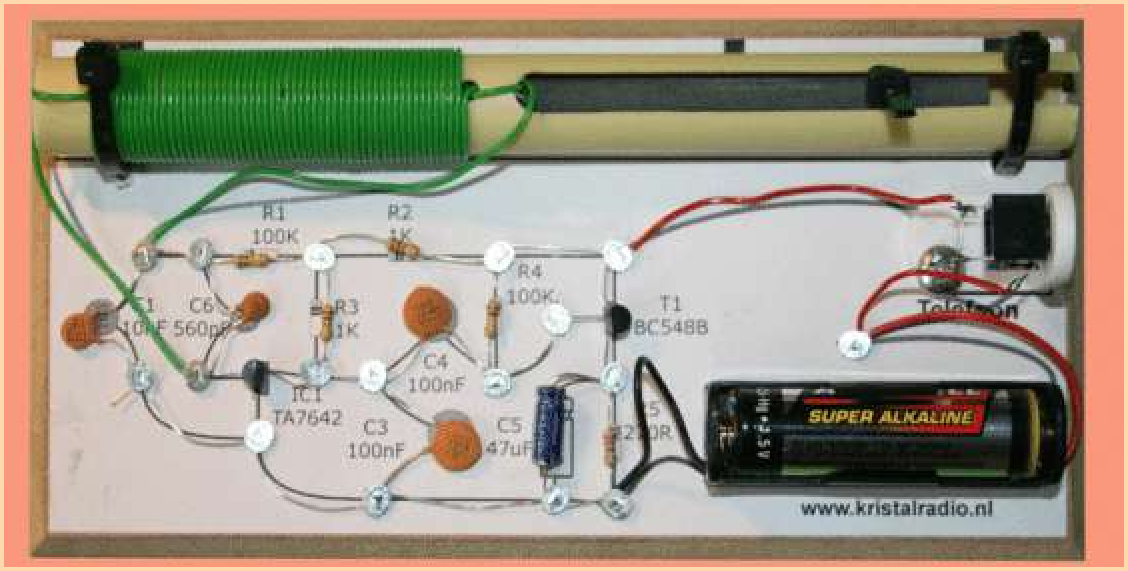
R1
100K
R2
1K
C6
R3
1K
R4
100K
T1
BC548B
Telefoon
IC1
TA7642
C4
100nF
C3
100nF
C5
47uF
SUPER ALKALINE
www.kristalradio.nl

A SLUG-TUNED RADIO

The circuit above is has a broad-band amplifier consisting of 10 transistors (IC1) and they are directly coupled (connected) to each other because it is not possible to "manufacture" a capacitor inside the IC. The IC has 3 terminals (pins, legs) and it looks like an ordinary transistor. Experimenting with this type of IC has shown that it is no better than 2 ordinary transistors connected in a direct-coupling arrangement.

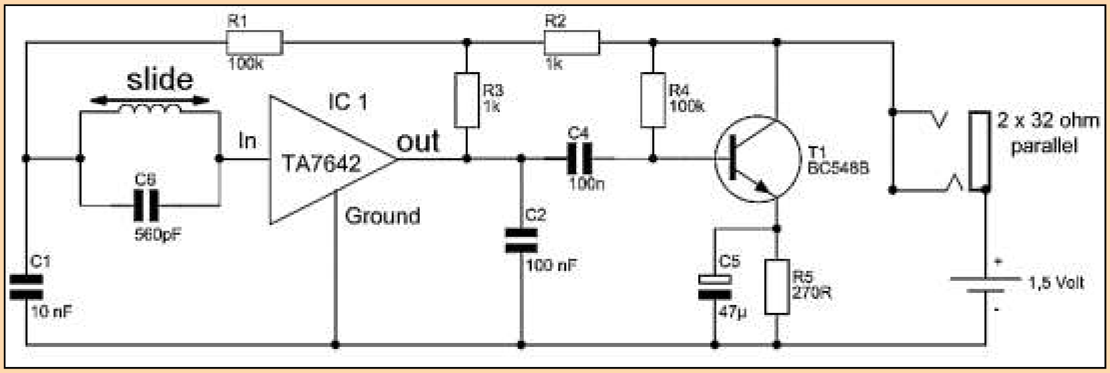
R1
100k
R2
1k
slide
IC 1
R3
1k
R4
100k
In
out
TA7642
C4
100n
T1
BC548B
2 x 32 ohm
parallel
C6
560pF
Ground
C2
100 nF
C1
10 nF
C5
47µ
R5
270R
+
1,5 Volt
-

Here is the address of the site for the slug-tuned radio.
http://www.kristalradio.nl/
Unfortunately the site is in Dutch and the kit is not available. However the photos give a clear picture of the how the parts are connected.

The inductance of the coil can also be altered by winding another coil and placing it near the first coil so that the magnetic field interacts with each other and changes the inductance of the circuit. This is called a VARIABLE INDUCTANCE TUNING COIL.
You can have one coil inside the other, two coils near each other or two flat coils side-by-side. Any two coils will interact with each other.

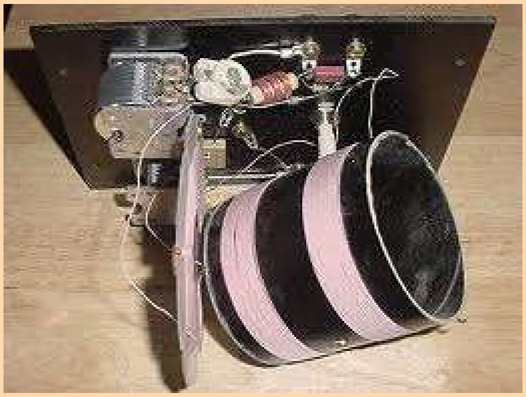

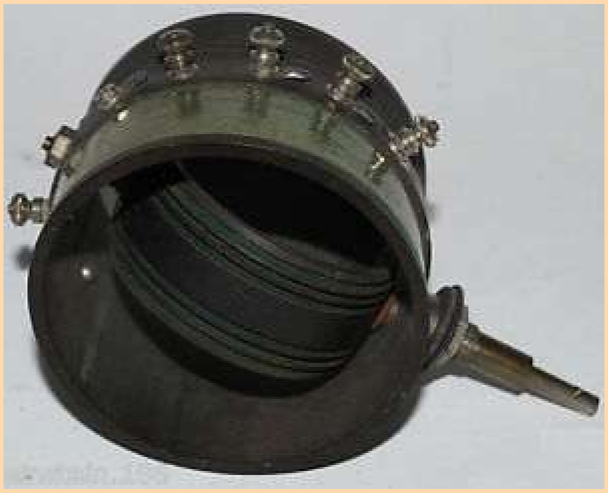

An Inductive TUNING COIL called a VARIOMETER

MAKING YOUR OWN FERRITE ROD ANTENNA

You can make your own FERRITE ROD ANTENNA by winding 60 to 80 turns of 0.25mm enamelled wire onto a 9mm ferrite rod or slab. If you wind it on a paper sleeve, you can move the coil along the rod to get the best performance. When the rod is slid out of the coil, the inductance changes considerably. However the inductance does change very slightly when the coil is moved along the rod.

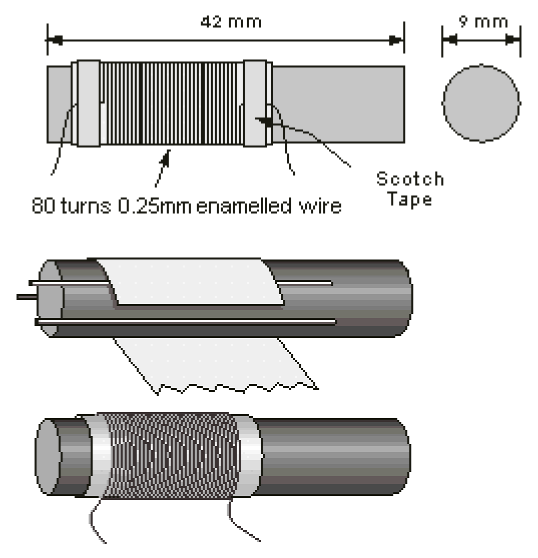

Make your own ferrite antenna

Now we come to the tuning capacitor::

THE TUNING CAPACITOR

The "C" in the "LC" TUNED CIRCUIT can be fixed or variable. When it is variable, it is called a TUNING CAPACITOR. The sheets of aluminium in the air tuning capacitor below are called PLATES and the moving plates are called VANES. The fixed plates make up the STATOR.
The space between the plates is AIR. The photo shows a single capacitor. If two capacitors are connected to the same shaft it is called a GANGED CAPACITOR.
The plates do not come fully out of mesh and that's why the capacitor has a minimum value. The maximum capacitance is when the plates are fully meshed. The odd shape of the plates is designed to produce a fairly constant increase in capacitance as the plates are engaged.

An air tuning capacitor:

**Air Tuning Capacitor
(Variable Capacitor)**

The capacitor can be made much smaller by using thinner vanes and placing plastic between the vanes. Plastic increases the capacitance about 3 times to 10 times.

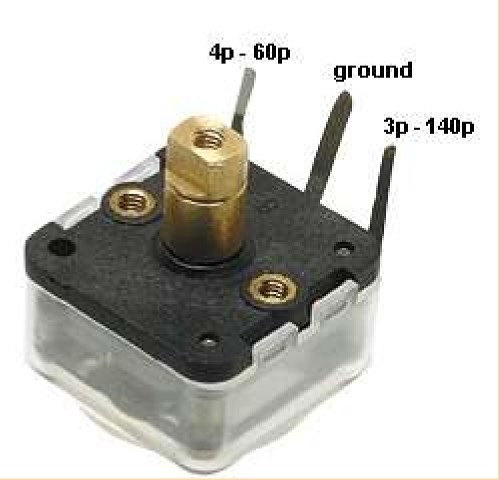

Tuning Capacitor as found in a pocket radio

The tuning capacitor can be replaced with a home-made equivalent that will work just the same.
You need:
4 sheets of aluminium foil (cooking foil) 10cm x 10cm.
4 sheets of thin cardboard 15cm x 20cm (cut A4 sheets in half).

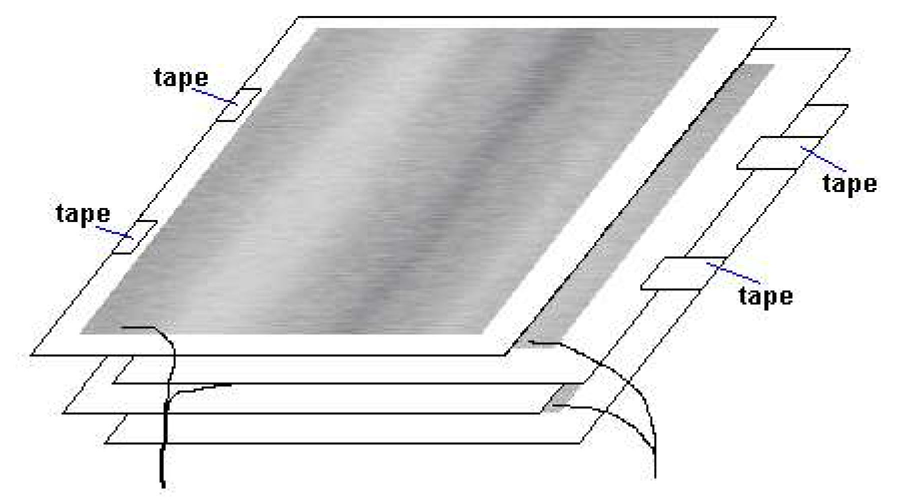

HOME-MADE CAPACITOR

Tape a sheet of aluminium foil to each sheet of cardboard with sticky-tape around all 4 sides. Take one strand of wire from a length of hook-up flex and sticky-tape the end to each sheet of aluminium to make good contact. Place 2 sheets on top of each other and move the top sheet slightly to the left and sticky-tape the edge so they don't move. Do this with the other two sheets but move the top sheet to the right. Now interleave the sets. Connect the wire from the first sheet to the third sheet. Connect the wire from the second sheet to the fourth sheet.
The cardboard (or paper) between the aluminium sheets increases the capacitance three times.
The capacitance decreases when the sheets are moved apart and the capacitance increases when the sheets are moved in. The capacitance also INCREASES when the sheets are squashed together such as when a book is placed on them.
You can also make a smaller capacitor by making each sheet smaller and using 6 sheets. You can then add a 100p or 220p in parallel with the home-made capacitor, to select the lower part of the band.

EACH CIRCUIT

Each circuit we describe in the following set of circuits is an improvement or advancement on the previous. We also offer a number of different types of aerial coils, amplifying stages and earphones. Some of the circuits use easy-to-obtain components and home-made equivalents for hard-to-get items. There will be something in this section for everyone to build.

In all radio circuits you will encounter **TWO MAIN PROBLEMS**:

If the FRONT END (the Coil and Capacitor) is loaded too much by the "pick-off" of the amplifying stages, you will only get one station.
If you get squealing or "motor-boating," try a different circuit and layout as the components you are using, plus the voltage of the battery, will need changing.
You cannot always increase the voltage of the supply and get a louder output. Sometimes the increased voltage will stop the circuit working or it may introduce too much gain that the circuit starts to squeal.
The Radio IC (ZN414) DOES NOT WORK on a voltage above 1.5v and some of the transistor circuits completely stop working with a higher voltage. This has to do with the biasing arrangements and if the circuit is designed for a low voltage, you need to keep to the suggested voltage and experiment with a slight increase in voltage and see what happens.
Building a radio is not easy as the enormous amount of amplification of the combined stages creates a feedback loop via the power rail that sets the circuit into oscillation. This effect gets worse with a higher supply voltage and we will explain this further with each of the circuits.

MAKING A CRYSTAL SET

You can buy a **CRYSTAL SET** kit (see the photo of the kit, above) or the individual components (a kit is the cheapest) or use the replacement for the **FERRITE ANTENNA COIL** (16 turns to 20 turns on a 150mm biscuit tin) and/or the **TUNING CAPACITOR** made from aluminium foil and cardboard sheets.
You will need an outside antenna and an earth (such as a water tap or the frame of your soldering iron) to pick up the radio stations.
If you cannot put up an outside antenna, you will need to add one or more amplifying stages and this will allow you to reduce the length of the antenna and increase the volume of the audio.

ADDING AMPLIFYING STAGES TO A CRYSTAL SET

You can add two different types of amplifying stages to a crystal set.
You can connect amplifying stage(s) to the FRONT END and these will be designed to put less load on the front end so the sensitivity and selectivity increases. These stages work at the frequency of the radio signal and they are called RF STAGES (Radio Frequency Stages).
You can build these stages out of individual components or use a chip called a RADIO CHIP or RADIO IC (integrated circuit) for less than $2.00.
The chip contains 5 stages of amplification and these are RF stages (or RF AMPLIFYING STAGES) and the concept is called TRF. (Tuned Radio Frequency).
It is not easy to get this type of amplifier working because the stages produce a very high overall gain and you get a lot of "motor-boating" and squealing if the gain is not controlled. The gain must be reduced when a strong signal is being passed through the circuit because a strong signal will produce a large output and this will be so large that some of the waveform will find its way to the front of the amplifier via the power rail and start to be amplified again. To prevent the output getting too large, the circuit has a negative feedback line - called the AGC line - Automatic Gain Control.
It would be very difficult to reproduce these 5 stages of amplification with discreet components and that's why it is best to use an IC.

The next stage is a DIODE to convert the RF (Radio Frequency) to AF (Audio Frequency). This can be done with the diode-characteristics of a base-emitter junction in a transistor and we will show the alternatives.

Any stages after the diode are AUDIO STAGES or AUDIO AMPLIFIER STAGES.
The main job of the AUDIO AMPLIFIER is to increase the DRIVE CAPABILITY.
In other words, increase the current capability of the circuit for an 8 ohm speaker or 8ohm earpiece (or 16 or 32 ohm).
This is a very difficult thing to do and requires at least 2 stages.
The LOAD you can put on a Crystal Set must be 10,000 ohms or higher. (if you put a lower resistance (impedance) on the output, you will load the FRONT END and reduce its ability to separate the stations.
That's why a crystal earpiece is normally used with a crystal set. It puts almost NO LOAD on the circuit.
If you put a load on the circuit the result will be only one or two stations across the whole dial and only the most powerful station will be received.

If you don't have a crystal earpiece, you will have to use an 8 ohm earpiece. This will require an IMPEDANCE CONVERTING CIRCUIT of 1,250:1
This is a simple way of saying we want the 8 ohm earpiece to appear as 10,000 ohms to the

crystal set.
To produce an overall gain of 1250, we need two stages of amplification.
If a transistor has a gain of 70, it will it will produce an impedance conversion of 70 times. This is a realistic value. Transistors with a gain of 200 will have a gain of about 70 when fitted to a circuit.
This means the other transistor needs to have a gain of about 20 and that is easy to achieve.

ADDING AMPLIFYING STAGES TO THE FRONT OF A CRYSTAL SET

Adding stages to the front of a crystal set are called RF STAGES (Radio Frequency Stages) because they amplify the RADIO STATION SIGNAL.
It does not matter if you amplify RF signals or AF signals. The result is the same.
The only difference is this: The frequency of RF signals is much higher (1,000 times higher) and the coupling capacitors can be much smaller.
This allows an RF amplifier to be built into an IC - called a Radio Chip.
One of the most popular Radio IC's is ZN414 or YS414. This chip has been copied by other manufacturers as: MK484, TA7642 and LMF501T.
All the chips are the same but the pinout is different.
These chips work on a 1.5v supply and if the voltage is increased above 1.5v, the gain of the stages increases to a point of total distortion.
To prevent strong signals producing distortion on 1.5v supply, the output is passed back to the input via a 150k resistor. This feedback line is called the AGC (Automatic Gain Control).
The chip contains 5 stages of amplification plus a stage that converts the RF signal to AF (Detector Stage). This means the signal diode in a Crystal Set is not needed.

Here is the circuit of the TA7642 Radio Chip. It performs the same as the ZN414 Radio Chip.

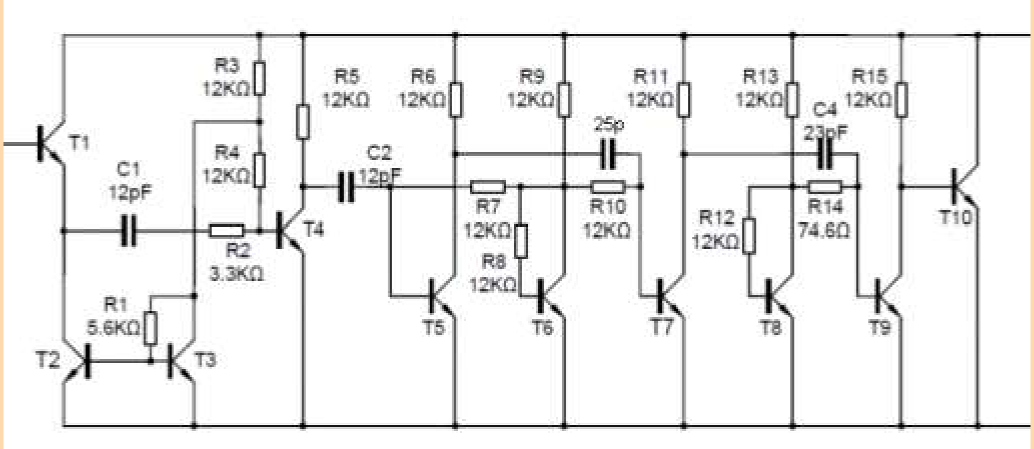
R3
12KΩ
R5
12KΩ
R6
12KΩ
R9
12KΩ
R11
12KΩ
R13
12KΩ
R15
12KΩ
C4
23pF
25p
T1
C1
12pF
R4
12KΩ
C2
12pF
R7
12KΩ
R10
12KΩ
R12
12KΩ
R14
74.6Ω
T10
R2
3.3KΩ
R8
12KΩ
R1
5.6KΩ
T2
T3
T4
T5
T6
T7
T8
T9

Here is the BLOCK DIAGRAM of the ZN414 Radio Chip:

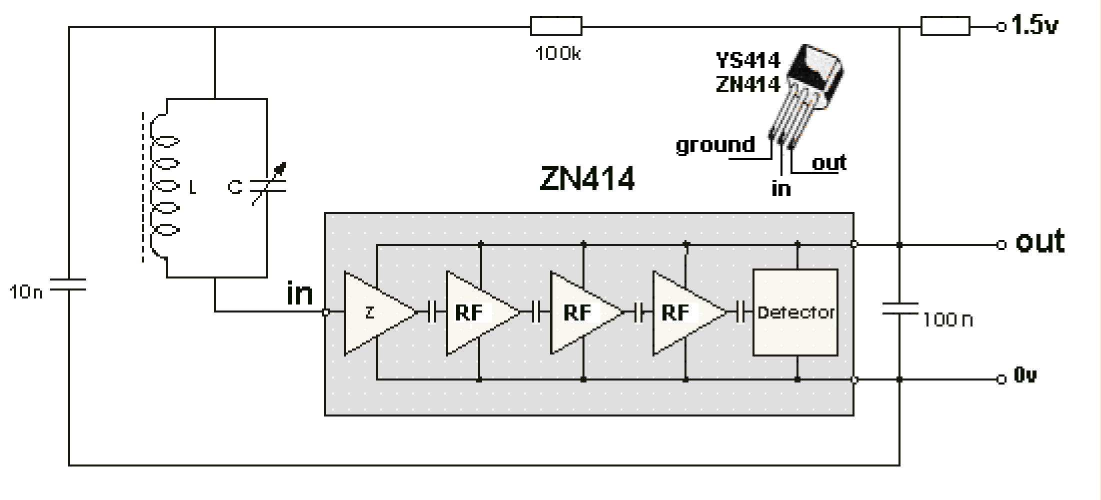
100k
1.5v
YS414
ZN414
ground
out
in
ZN414
L
C
10n
in
Z
RF
RF
RF
Detector
out
100n
0v

The ZN414 chip can be purchased from Talking Electronics for $1.00 plus postage

USING THE ZN414 RADIO IC

By using the ZN414 radio IC (or any if the equivalents) you can create a POCKET RADIO to drive a headphone or speaker.
But it is not easy to use the chip. The main problem is receiving the strong signals without producing distortion and then being able to pick up the weak stations.
A fixed 100k feedback resistor does not provide adequate control and a TRF radio has limited capabilities.
That's why radio manufacturers make SUPERHETRODYNE receivers. Even though they are more complex, the result is far superior.
However a simple TRF set can be made with the Radio IC and a few stages of audio amplification.

The following circuit uses just the Radio IC and a crystal earpiece or the home-made earpiece described above:
You can use a home-made FRAME ANTENNA or a home-made FERRITE ROD ANTENNA and a home-made VARIABLE CAPACITOR.

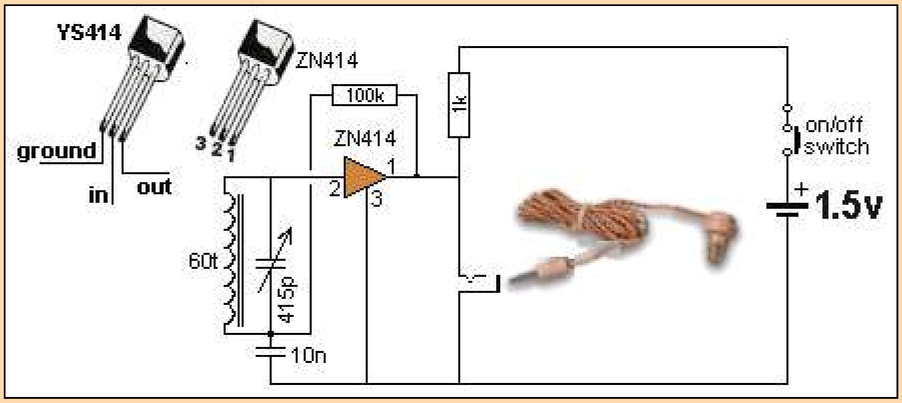

The Simplest ZN414 Radio

Connecting the ground (0v rail) to the frame of your soldering iron or a water tap will increase the output volume. The circuit above shows a Crystal Earpiece. Using a Crystal Earpiece may require adding a 10n across the earpiece to improve the output volume. The substitute Piezo Earpiece is effectively a 20n capacitor and an additional capacitor is not needed.

2 TRANSISTOR RADIO

Here is a simple 2-Transistor radio.
The secret to its performance is the 7 turn "pick-off" from the FRONT END (the TUNED CIRCUIT).
The ratio of 7 turns to 60 turns means a small percentage of the voltage generated in the tuned circuit is passed to the transistor. Thus it puts a small load on the TUNED CIRCUIT.
I don't want to go into any mathematics. The turns ratio is 60:7 = 8 but the effect of the 7 turns "pick-off" has an effect called the IMPEDANCE EFFECT and this is the SQUARE OF THE TURNS RATIO. Thus the IMPEDANCE EFFECT is 8 x 8 = 64. This means the "pick-off" (the LOADING EFFECT) is just a few percent. The front end can produce voltages as high as 500mV because a crystal set can produce a voltage high enough to pass through a diode (350mV) and have sufficient to drive a crystal earpiece.
Even though the front end has a "step-down" ratio, the voltage out the 7 turns will be sufficient to drive the first transistor.
The "transformer" does 2 things: It reduces the loading on the tuned circuit ENORMOUSLY and it produces an output with a higher current than is circuiting in the front end. Even though the transistor is turned ON and biased by the 33k, it is classified as a low-impedance load as far as the front end is concerned and the input signal has to be accompanied by a certain amount of current, otherwise the transistor will not respond to the voltage. The 7-turn "pick-off" is able to provide this current.
Both transistors are biased ON via the 33k base-bias resistors and thus the first transistor responds to the slightest millivolt signal.
This circuit was tested and had the same performance as the **Simplest ZN414 Radio Circuit** above. It can be operated on 1.5v to 6v and the strongest stations tend to overload on 6v. A short antenna is needed.

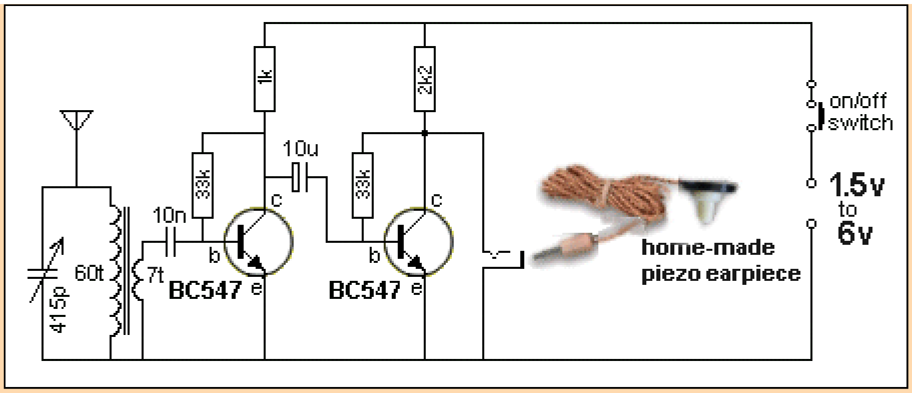
1k
2k2
on/off switch
10u
33k
33k
1.5v to 6v
10n
c
c
b
b
60t
7t
415p
BC547 e
BC547 e
home-made piezo earpiece

SIMPLEST 2-TRANSISTOR RADIO
using a very-high-impedance earpiece

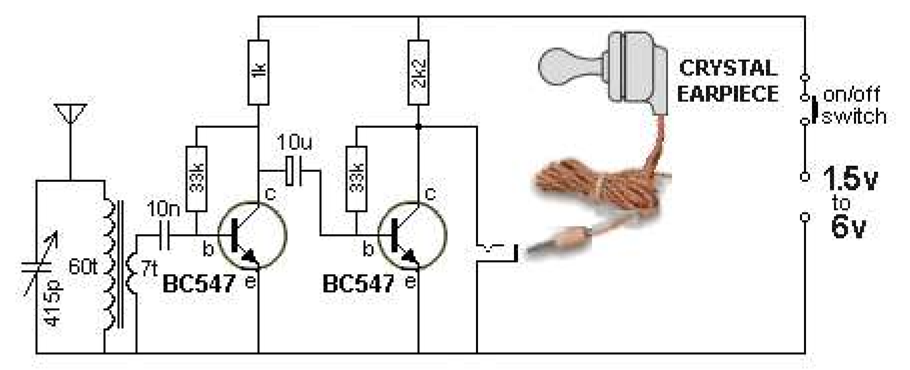
1k
2k2
CRYSTAL EARPIECE
on/off switch
10u
33k
33k
1.5v to 6v
10n
c
c
b
b
60t
7t
415p
BC547 e
BC547 e

SIMPLEST 2-TRANSISTOR RADIO
using a crystal earpiece

ADDING AN IMPEDANCE MATCHING STAGE

You can add an IMPEDANCE MATCHING STAGE to the output of the circuit above so a low-impedance earpiece can be used.
We call it an IMPEDANCE MATCHING STAGE because this is the correct technical term. It is an AMPLIFYING stage but it amplifies the CURRENT because the second transistor cannot drive an 8 ohm LOAD. 8 ohms is a very low resistance and if it is connected directly to the second transistor, the output will be almost zero.
The reason for this is covered in our discussion: The Transistor Amplifier.
This stage will not increase the volume but simply match the 8 ohm load to the circuit above.
It is very difficult to connect a LOAD to this type of circuit because it will take more current from the battery and cause the supply voltage to fluctuate. These fluctuations will be passed to the first stage and cause variations in the signal. This will be amplified by the first and second transistors in the form of a low-frequency buzzing called **MOTOR-BOATING**.
The only way to reduce or remove this noise is to add an electrolytic across the power rails and reduce the supply voltage. The third transistor simply takes the waveform on the output of the second transistor and delivers it to the earphone with a higher current. It is called an IMPEDANCE MATCHING STAGE as it effectively increases the 8 ohm load by a factor of about 100.

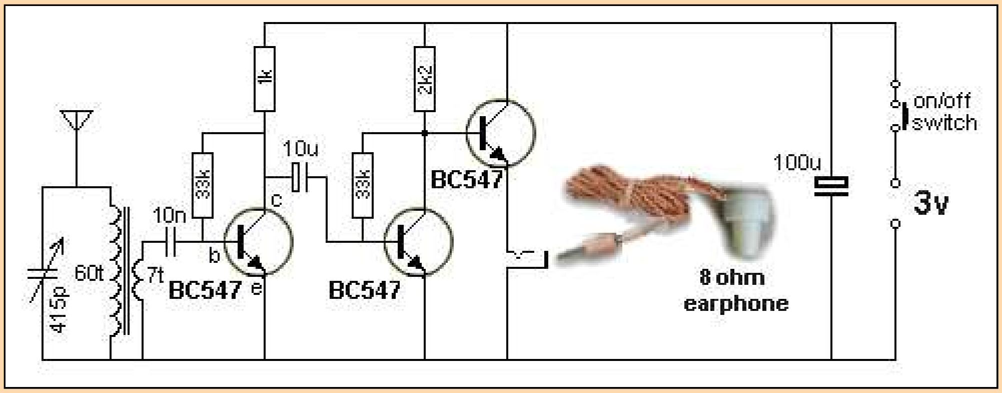
1k
2k2
on/off switch
10u
BC547
100u
3v
33k
33k
10n
c
b
60t
7t
415p
BC547 e
BC547
8 ohm earphone

3-TRANSISTOR RADIO with buffer stage for 8 ohm earpiece

The 3rd transistor converts the 8R to about 800R

You can use 16 ohm, 32 ohm or 64 ohm in place of the 8R earpiece and these will give better performance as they will take less current and improve the stability of the circuit.
Low-impedance earphones create "motor-boating" due to the peaks of current and this can be very hard to fix.

A 2-TRANSISTOR RADIO with REGENERATION

The next stage in our discussion to get better performance is a feature called **REGENERATION**. Regeneration sends a small output signal back to a previous stage in the form of POSITIVE FEEDBACK to INCREASE the original signal. The signal on the emitter of the first transistor is the same amplitude as the signal entering the base but the FRAME ANTENNA has a turns ratio of 5:15 and this increases the signal on the receiving section of the antenna by up to 3 times. But we want the returning signal to be just above the amplitude of the receiving signal and so a resistive adjustment (attenuator) is provided on the emitter to deliver just the right amplitude.
This has the effect of increasing the amplitude on the TUNED CIRCUIT and is just like reducing the load on the circuit.
As we have mentioned above, when the tuned circuit is lightly loaded, it will pick up a station at the exact frequency of transmission and if the dial is changed slightly, the station will disappear. This quality is called SELECTIVITY.
At the same time, the Tuned Circuit will pick up weak stations and this is called SENSITIVITY.
The quality of a receiver depends on the loading of the TUNED CIRCUIT.

Here is the original circuit from Elektor Magazine with the prototype made on matrix board and fixed to a base-board with a frame antenna made from two sticks of wood. The photo shows a speaker but the output is so low that you really need headphones.

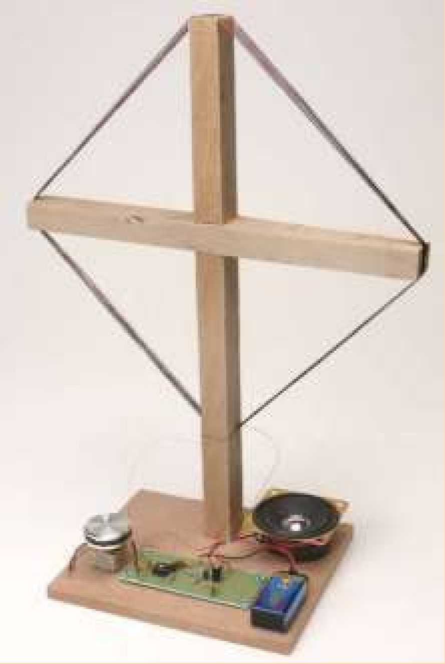

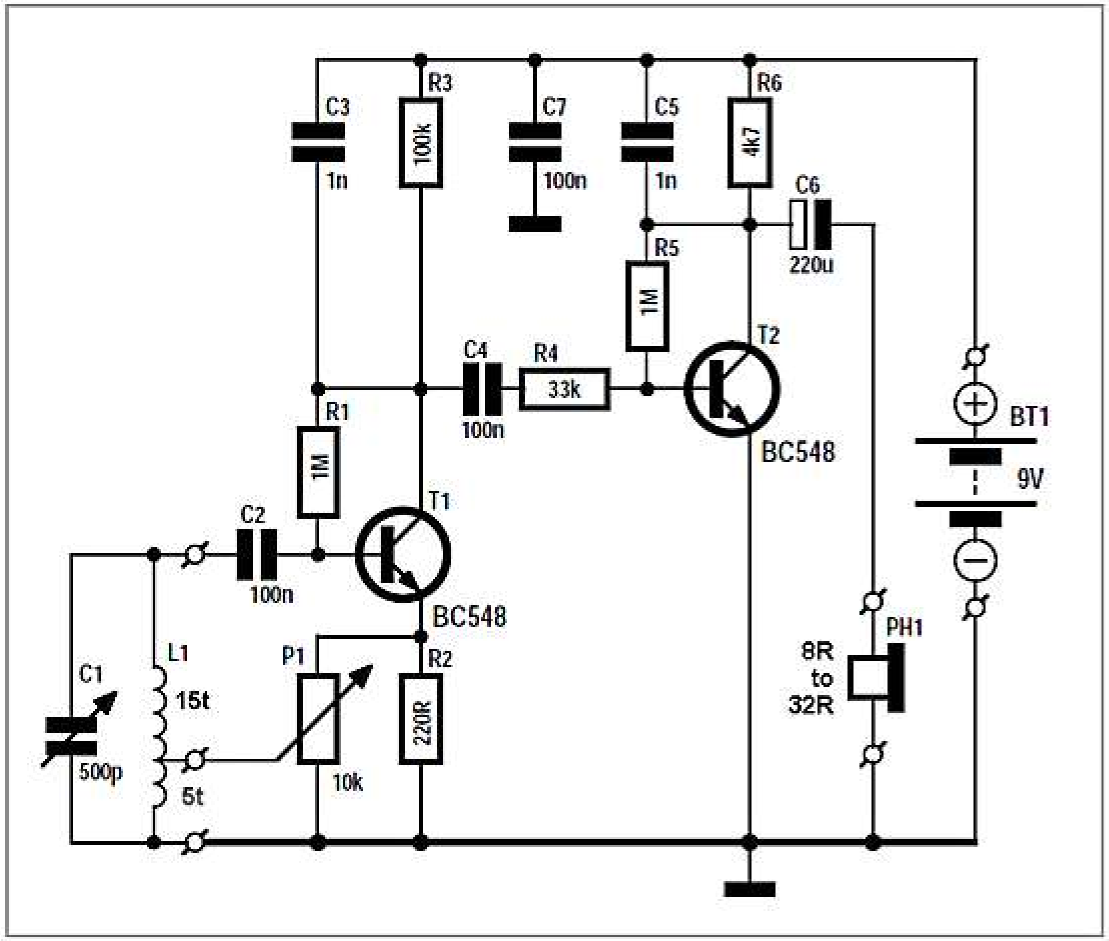
C3
1n
R3
100k
C7
100n
C5
1n
R6
4k7
C6
220u
R5
1M
T2
C4
100n
R4
33k
BC548
BT1
9V
R1
1M
T1
C2
100n
BC548
PH1
8R
to
32R
P1
10k
R2
220R
C1
500p
L1
15t
5t

2 TRANSISTOR REGENERATIVE RADIO from Elektor Magazine

The circuit is very complex and the output will be very low as the circuit cannot drive a low-impedance earphone via a 4k7 load resistor. The 4k7 resistor is actually driving the speaker (the transistor is simply discharging the 220u). The 4k7 only allows 32/4700 x 9 = 61mV to appear across the earphone - a very poor result.
The skill of designing a transistor stage is covered in our comprehensive eBook: The Transistor Amplifier and you wont make a mistake like this !!!
The circuit above can be simplified and we can add the REGENERATIVE feature to our **Simplest 2-Transistor Radio** circuit:

Our circuit uses a 15 turn circular FRAME ANTENNA 15cm diameter and a 5 turn REGENERATION coil.

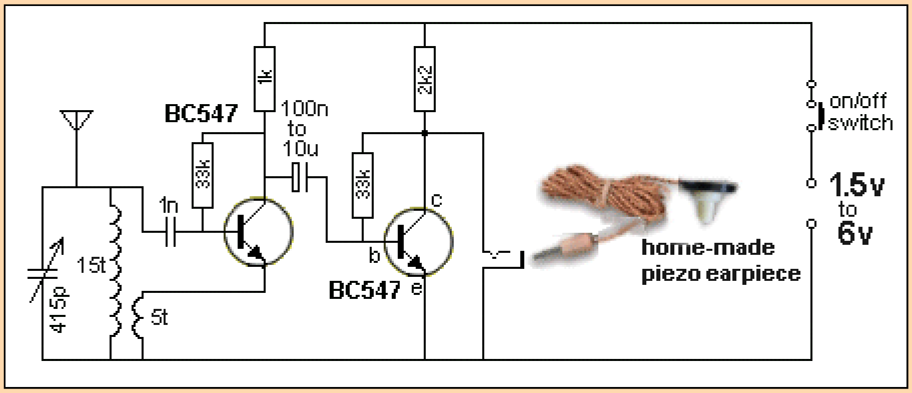
BC547
1k
100n
to
10u
2k2
33k
1n
33k
on/off
switch
1.5v
to
6v
c
b
e
BC547
15t
5t
415p
home-made
piezo earpiece

2 TRANSISTOR RADIO with REGENERATION

The regeneration coil is brought near the main coil and as it gets closer you can hear the audio get louder. If this does not happen, turn the coil around.
Early radios used this technique and the operator had to adjust the coil by hand. No-one minded because radio was a fascination and the simplest radio cost more than a weeks wages. To listen

to a broadcast through headphones was an amazement and listeners would sit all night with headphones listening to music.
This is very fiddly and by adding an extra buffer stage, we can use a Frame Antenna with a very clever "pick-off" that does not load the front end. This gives the circuit very good sensitivity and selectivity without regeneration.

3 TRANSISTOR RADIO

Here is our final design for the simplest self-contained 3-Transistor Radio using our home-made Tuning Capacitor and 250mm x 350mm Frame Antenna. It picks up the local stations and drives a low-impedance earphone or set of earphones (from a mobile phone).

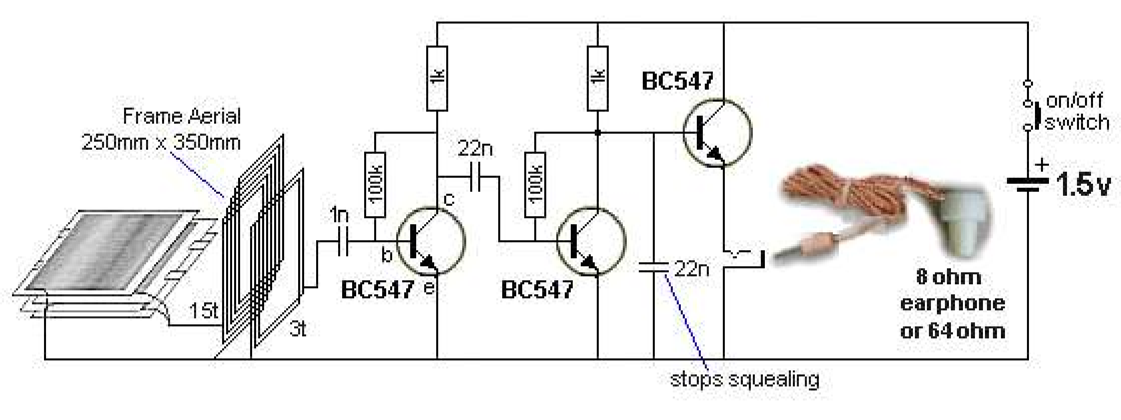

3-TRANSISTOR RADIO

The circuit performs very well and uses readily-available components. The 22n across the output is essential to stop squealing.
The secret to sensitivity and selectivity is the turns-ratio on the Frame Antenna. The 3-turn "pick-off" puts very little load on the front end and this allows the stations to be tuned with our home-made Tuning Capacitor.
The circuit contains all the features we have discussed above and only needs a 1.5v supply.
Build this circuit before you buy any expensive tuning capacitors, IC's or ferrite slab antennas as you will not get any better results.
This is called a TRF circuit and because the stages operate at Radio Frequency or Audio Frequency. Due to the high amount of amplification, the circuit can start to squeal (feedback, motorboat) due to the layout.
You may need to shorten or lengthen the leads or move the parts slightly - it's that critical.
However the result is a portable radio that needs no earth and will pick up the strong stations.
You can try connecting the 0v rail to the metal part of a soldering iron to increase the number of stations.

LOADING

The whole success of picking up a radio station is the RECEIVING CIRCUIT. The receiving circuit is the coil and the signal in the air (from the radio station) must go down the centre of the coil.
It cannot pass over the top or the bottom of the coil. Only the signal that goes down the centre of the coil is received.
As you can see, the centre of the coil is not very big and it is amazing that the signal can pass down the centre. But it does, and that is the only signal that will be amplified.
This signal is passed to the capacitor and we have explained how the signal is gradually increased and increased in amplitude until it is as large as 500mV. The signal from the radio station may be as small as a few millivolts, but as it keeps pushing the "swing" back and forth, the amplitude get larger and larger.
If you put your finger on the "swing" you will prevent it get larger and larger and it only requires the slightest touch of your finger to prevent the swing gaining full amplitude.
In electronic terms, your finger is called LOADING THE CIRCUIT and since we have to pass the signal to further stages of amplification, we need to "tap" or "load" or "pick-off" a signal.
The aim is to load the circuit as least as possible because the actual energy entering the circuit is very small.
In fact, this is all the energy we can remove as that is all the energy entering it.
Because a very small amount of energy is entering the "front-end" we classify it having a very

high impedance. It is very difficult to provide a value of impedance for this circuit because impedance has the term "Z" and the circuit is operating a very high frequency so resistance values are not the same as impedance values.

The actual resistance of the circuit is ONE OHM but the impedance is more like 10,000 ohms to 100,000 ohms.

We can explain its high impedance if we put a 100,000 ohm resistor across the circuit. The waveform will be reduced very slightly. If we put a 10,000 ohm resistor across the circuit, the signal will be reduced a reasonably large amount. If we put 1,000 ohms across the circuit it will stop working.

This means a load of 100,000 ohms will have the least effect.

In a crystal set, the diode creates NO LOAD until a voltage of 350mV is reached. It then passes excess voltage to a crystal earpiece that has a very high impedance. That's why a crystal set will produce a good output. The LOADING is very small.

When a transistor is connected to the TUNED CIRCUIT, it starts to put a load on the circuit after 600mV and this load is VERY HIGH. The "resistance" of the base-emitter junction is about 1k and the signal will find it very difficult to rise above 600mV because the incoming energy is not sufficient to increase the voltage.

Adding a capacitor between the base and the front end allows the transistor to be self-biased and get a turn-on voltage of about 600mV from a base-bias resistor.

The FRONT END is now separated from the transistor and ANY voltage it is producing will be passed to the transistor via the capacitor.

Whereas, with the crystal set, the first 350mV could be produced without any loading, the circuit is now loaded AT ALL TIMES.

This means we have to load the circuit as lightly as possible to be able to pick up individual stations.

The only way we can do this is to use a capacitor of the smallest practical value and this has to be worked out by trying different values. If the value is too small, the transistor will not detect a small signal. If the value is too large, the circuit will stop working.

Values such as 1n, 10n and 100n are suitable.

Values such as 1u, or 10u will be too large.

The CRYSTAL SET loading and a transistor load are completely different.

The transistor loads the front end ALL THE TIME and that's why you need to use a transformer or other ways to reduce the loading. Sometimes a Field Effect Transistor is used as it puts almost no load on the front end.

2-4-2016

Histats.com
Vis. today
26

FLIR
12 THINGS TO KNOW BEFORE BUYING A THERMAL CAMERA
GET THE GUIDEBOOK NOW!

Quickturn PCB & Assembly
Your one-stop source for all types of PCBs and Full Turnkey PCB Assembly.
rushpcb.com

Quickturn PCB & Assembly
Your one-stop source for all types of PCBs and Full Turnkey PCB Assembly.
rushpcb.com

Unlimited HDTV is Here

TV Bandit

Companies already tried to ban this antenna but failed. Get yours before it’s too late!

Recreating an Alfred P. Morgan crystal set

Analog Dial | Page 1 | Page 2 | Page 3 | Page 4

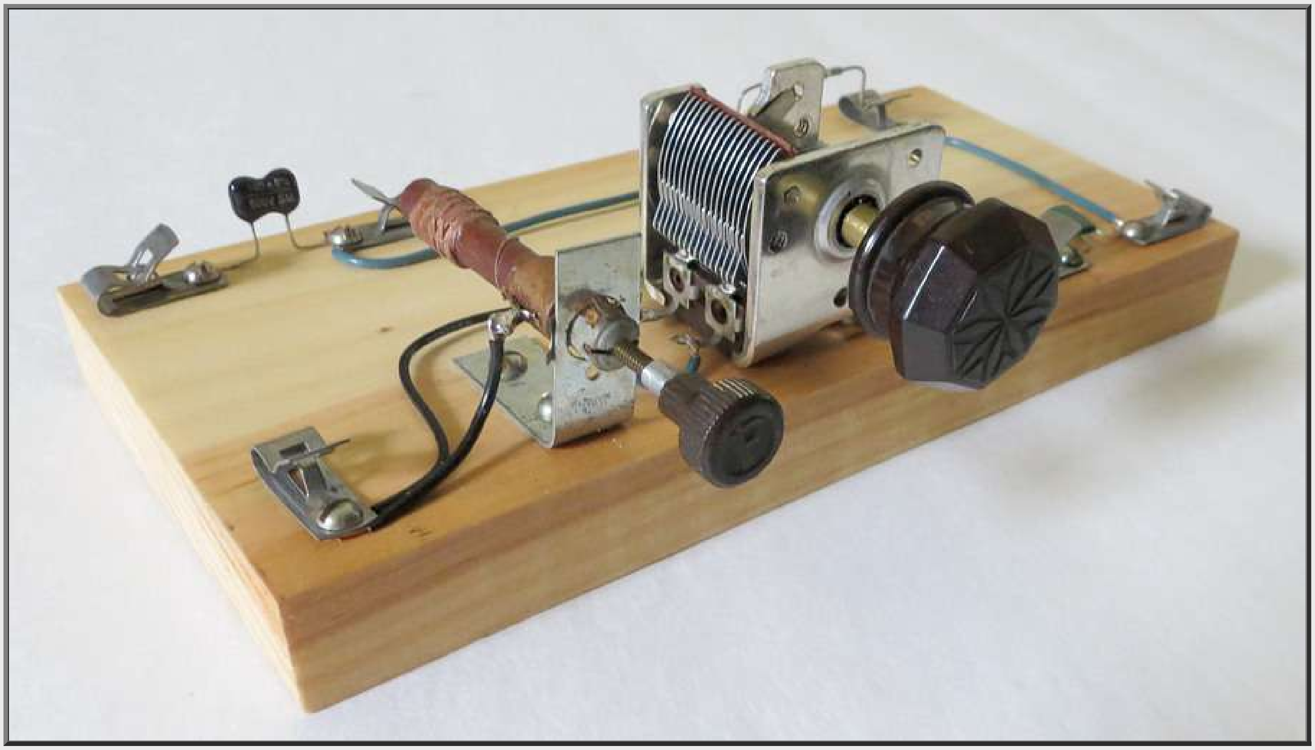

A recreation of my first (working) crystal set, originally built in 1966.

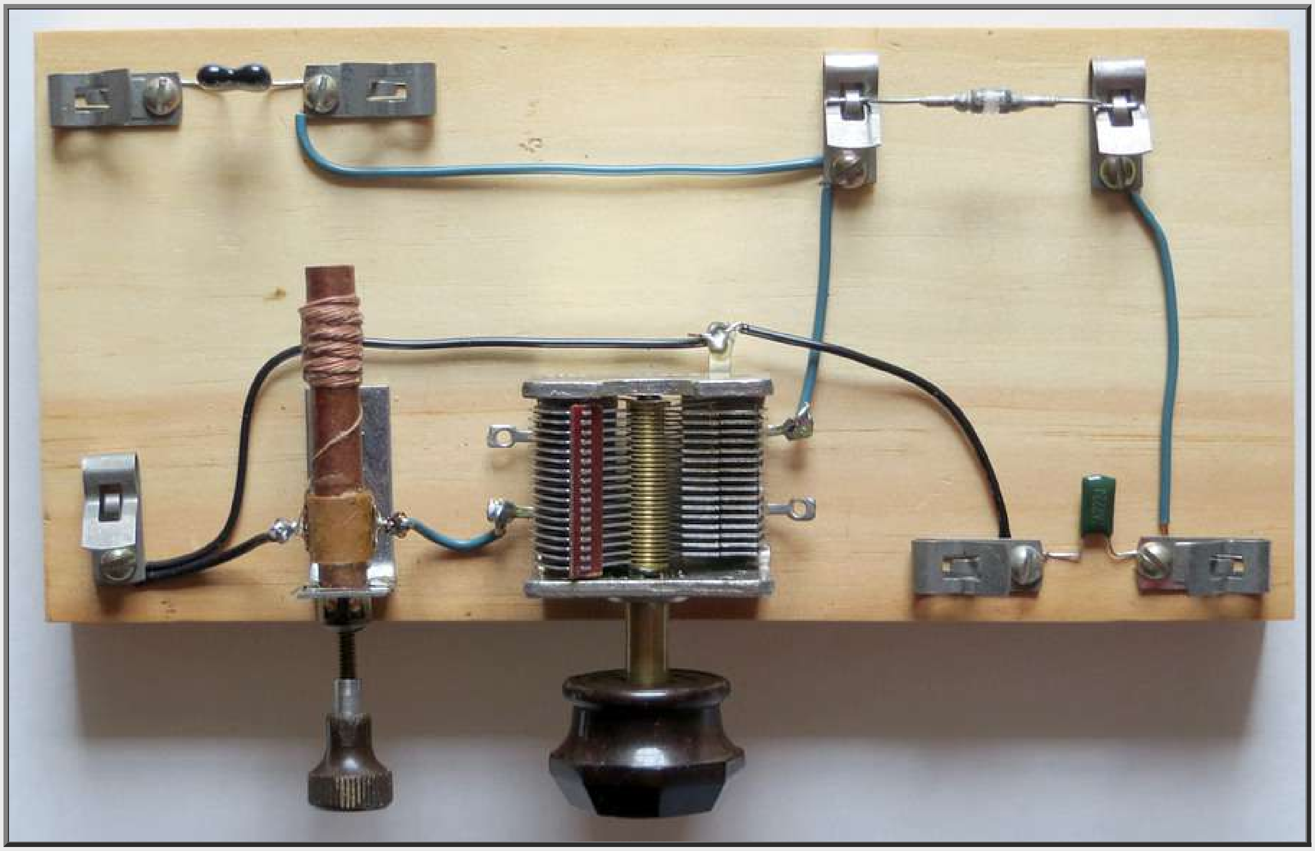

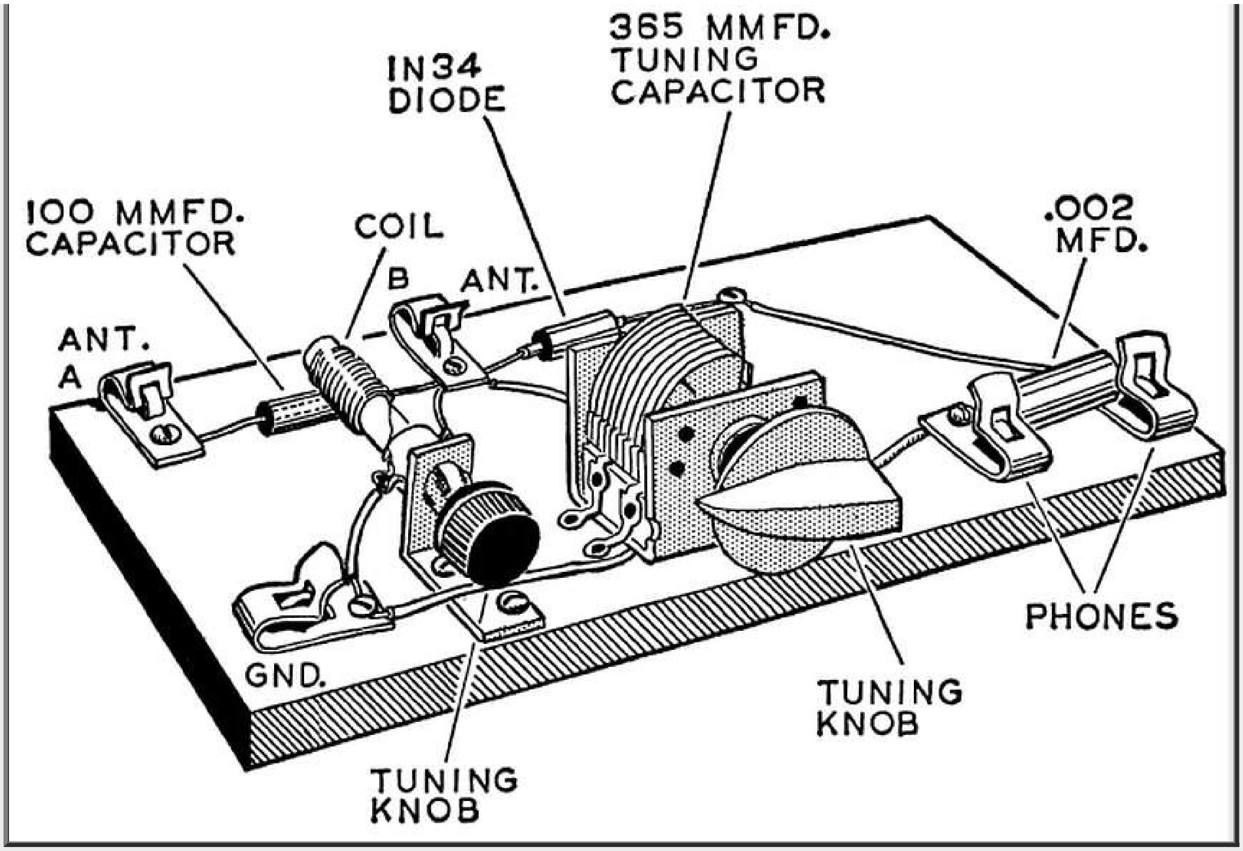
365 MMFD.
TUNING
CAPACITOR
IN34
DIODE
100 MMFD.
CAPACITOR
COIL
B
ANT.
.002
MFD.
ANT.
A
PHONES
GND.
TUNING
KNOB
TUNING
KNOB

Page from "The Boys' Third Book of Radio and Electronics" by Alfred P. Morgan.

Back Story:

In the Summer of 1966 I was 10 1/2 years old and one of my favorite TV shows was "Get Smart". It was a show about a bumbling secret agent named Maxwell Smart who had all sorts of gadgets at his disposal. My friend Billy Meyers and I decided we wanted to be secret agents, like Max. One of the "secret agent" things we'd do was to pick out a guy with a briefcase walking home from work and declare him a Soviet spy. Then we would "tail" him for a few blocks, making up stories about him.

Billy and I had no way to communicate with each other after we had to come in for the night. He lived half a block away. Kids didn't use the phone back then. There was only one phone in the house and I don't remember using it before I was about 14 years old. Maxwell Smart had a phone in the sole of his shoe. We needed something like that!

Maxwell Smart (Don Adams) Agent 86.

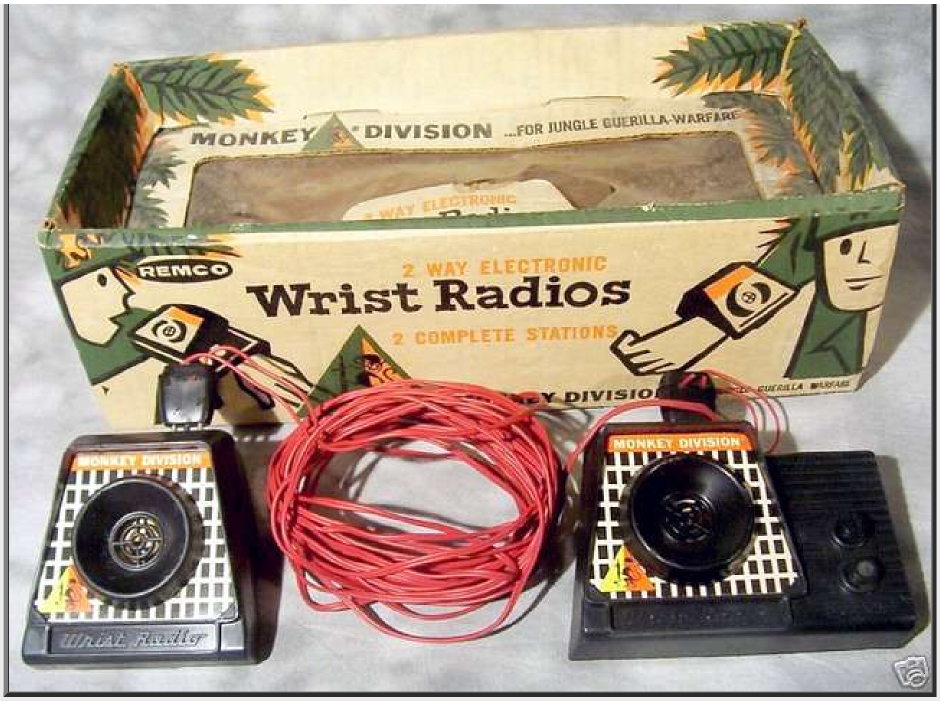
MONKEY DIVISION ...FOR JUNGLE GUERILLA-WARFARE
REMCO
2 WAY ELECTRONIC
Wrist Radios
2 COMPLETE STATIONS
MONKEY DIVISION
Wrist Radio

Remco "Monkey Division" Wrist Radios.

One thing I DID have was a set of Monkey Division Wrist Radios. These were powered by a single C battery in the "Master" unit, which also had a button on it that would buzz the other receiver to get the users attention (apparently, the other user was unworthy of a button). A metallic speaker doubled as a microphone. The sound was very tinny, but you could make it out. Unfortunately, they were wired to each other. There was nothing "radio" about them.

I remember taking them outside with my brother Rob, and I could see and hear him talking at the same time his voice was coming over the wrist radio. They were pretty much useless, unless you like running around with a wire connecting you to your brother.

BUT... with some extra wire strung across the driveway, down the backs of the houses and into Billy's bedroom, Billy and I would be able to talk to each other! I immediately presented this great idea to my mom with a request the she fund the cost of the wire, and she immediately refused and told me the electric company would just come out and take the wire down.

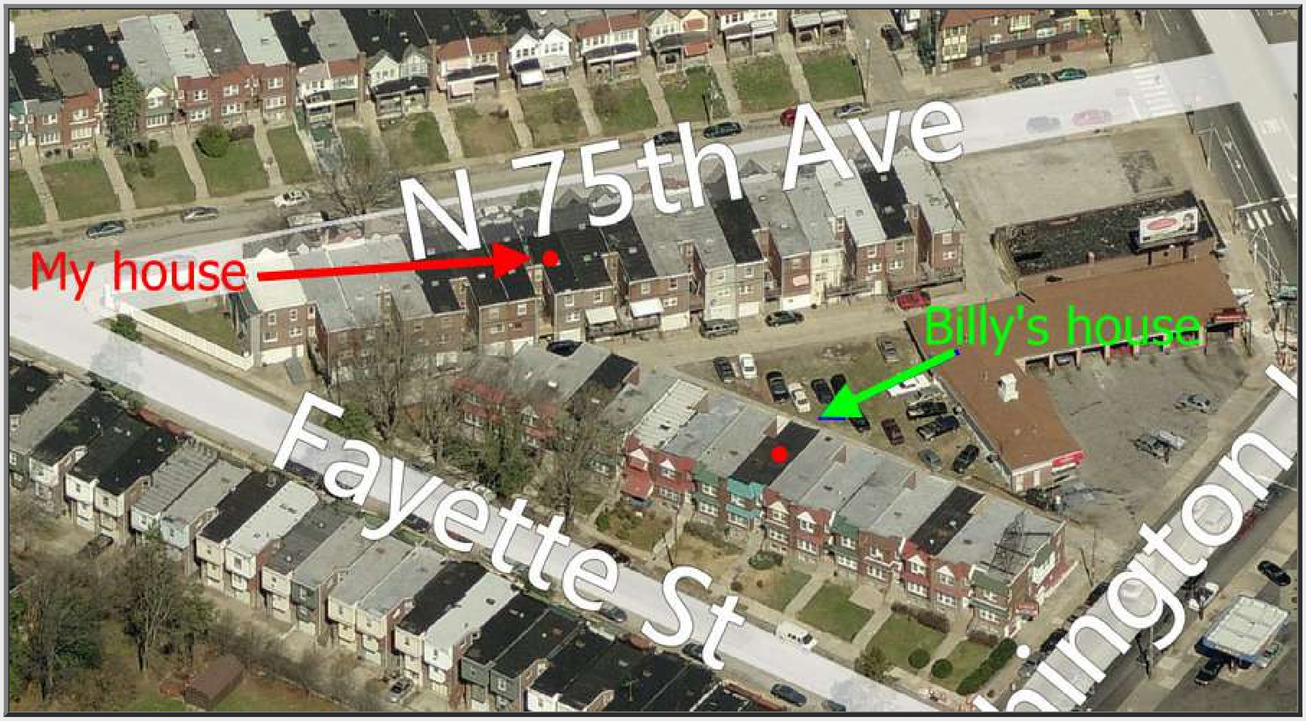
N 75th Ave
My house
Billy's house
Fayette St
ington

Location of my house and Billy's house (actually his grandparent's house) The wire would have to be strung across the driveway. This is part of the West Oak Lane section of Philadelphia.

There was only one thing left to do. Build a radio. I went down the basement and connected a battery to a speaker and used a coat hanger as an antenna. This was similar to the wrist radio, but the coat hanger antenna replaced the wire. It didn't work! All it did was make clicking sounds in the speaker. I had to wait for my dad to come home from work and ask him why.

When my dad came home I showed him the setup and he said, "You don't have a detector." I asked him what a detector was and he told me to go the library and get a book on radio. The next day I had a copy of "**The Boys' Second Book of Radio and Electronics**" by Alfred P. Morgan. Chapter 2, page 15 was titled "Building Your First Radio Receiver". I wouldn't be able to talk to Billy with it, but that was OK. My mom wouldn't buy me a secret agent coat, we didn't have any gadgets, and our secret agent days were coming to an end.

Now there was another problem. None of the parts needed to build anything in the book could be found around my dad's workbench. There was a store on Ogontz Avenue named REE Electronics, so I headed up there with a list of parts. The store sold stereo equipment and fortunately for me, also repaired stereo equipment. I asked the man in the store if he sold diodes or "capacitaters" and he sent me into the back of the place. There were two guys back there and bins of parts along the wall.

S.U.R.F.A.C.E.
S.U.R.F.A.C.E. CLOTHING CO.
HAIR CULTURE
REACH
FedEx

REE Electronics was located at 7709 - 7711 Ogontz Avenue in Philadelphia. The entire block has been razed and rebuilt, and is no longer recognizable. The picture above is the 7900 block of Ogontz Avenue. The store on the left is the only one that retains its original appearance, with the glass store window and the apartment overhead. This is how REE Electronics looked in 1966.

The two guys were pretty cool. I announced that I would like a "three hundred and sixty five micro micro farad variable capacitater". They asked me a couple of questions and told me to come back with the book. They had all the parts I needed except the coil. No problem, I would just build the set with no coil. I came home with Fahnestock clips, a 1N34 Germanium diode, a variable capacitor and a crystal earplug.

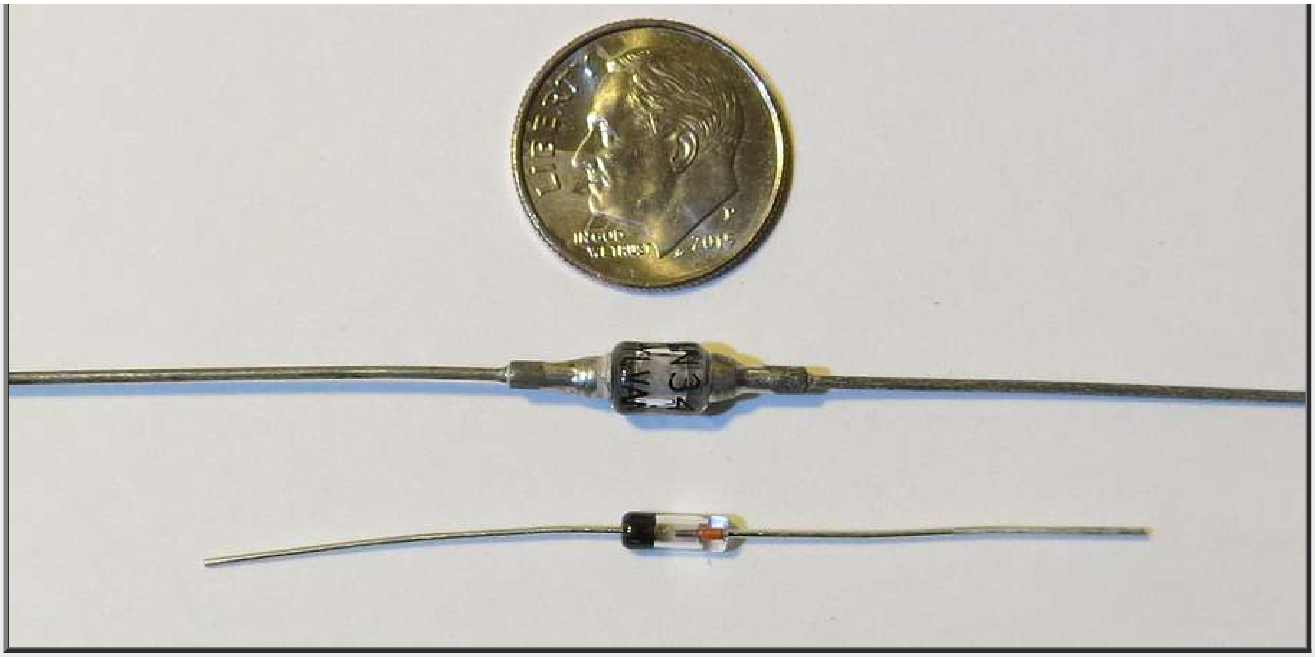

A Sylvania 1N34A Germanium diode (top) from 1949 and its modern counterpart. The new diodes are literally a dime a dozen, less than one cent each. The 1949 Sylvania diode cost me $9.00 in 2015. It is a duplicate of the one I bought at REE Electronics in 1966.

The diode was 65 cents in 1966. I can't use the $9.00 diode because I don't want to bend the leads, so I sort of just look at it.

Of course the radio didn't work without a coil. It did pick up the slightest whisper of KYW AM 1060 mingled with WIBG AM 990. There seemed to be some buzzing associated with it, as what I could hear sounded distorted. I HEARD something, that was the really, really neat part. It made such an impression on me that I remember the date. July 26, 1966.

The "problem" with the Alfred P. Morgan books was that they were not written for anybody as dumb as I was. Morgan didn't write, "If you can't find the coil you can make one." He just said to go buy a coil. Not only that, but there were no photographs in the book, though there were excellent drawings on almost every page. Since I had never seen some of the parts in real life, I didn't know exactly what the coil looked like, and I didn't understand what it did. That's because I didn't read the book! I was stuck on page 15.

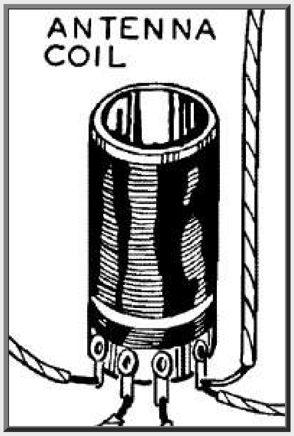

This is the picture of the coil from the book. I didn't know what I was looking at.

I returned the book to the library and came home with "**The Boys' Third Book of Radio and Electronics**." I found a simple radio on page 104 and soon headed back to REE Electronics. This time, they DID have the coil! I can't remember how much time passed after the first non-functioning radio was built. Probably a month or so.

I asked my dad to cut me a wooden base for the radio. I started building the radio and, if I remember correctly, it took me a long time. I didn't have a drill, so any holes in the base were made with the point of a compass. There were three connections that needed soldering. I got some solder from the basement, and the tweezers and alcohol lamp from my chemistry set. The tweezers were heated in the flame of the lamp till the tips began to glow, then I would quickly solder the joint.

One day a friend from school named Leo Pound stopped by on his bicycle. This was a bit unusual because Leo lived miles away. I don't even know how he knew where I lived. He recently told me (via Facebook) that he remembers helping me build the radio. Odd that it was the one and only time he came by. Apparently, we got the radio working that very day.

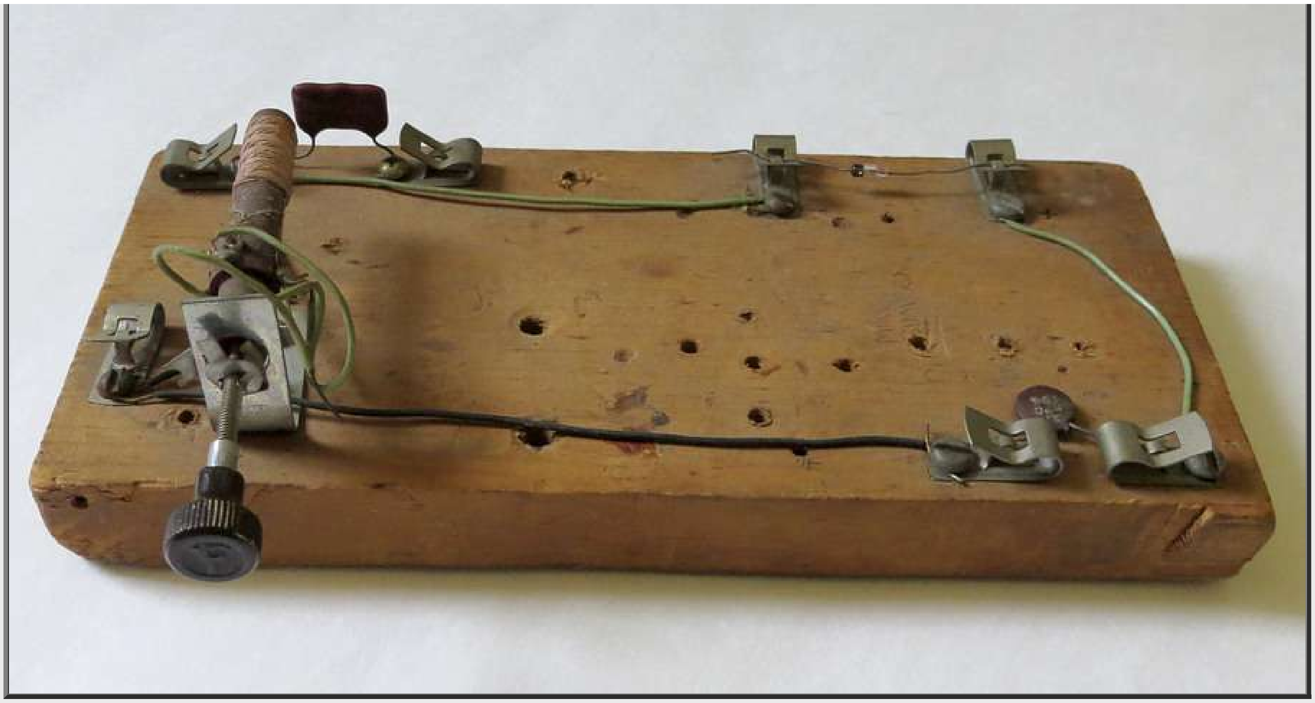

Here is the base of the radio made in 1966.

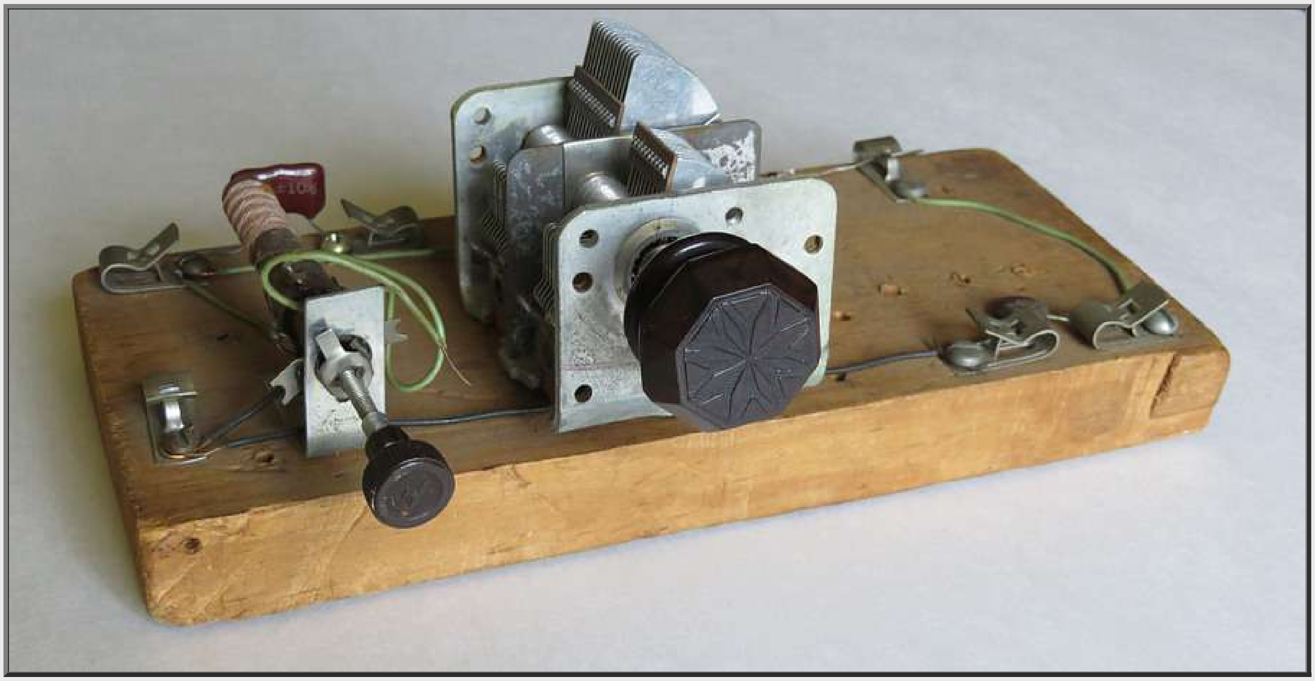

I found all the original parts except for the tuning capacitor. However, it looked like this one. It's just sitting on the base in this photo, but the radio pretty much looked exactly like this. My dad gave me the big tuning knob. The smaller knob came from a lamp in my bedroom. I didn't tell my mom you could no longer turn on the lamp, but eventually I found a lamp in somebody's trash and took the knob off to replace the one on my lamp.

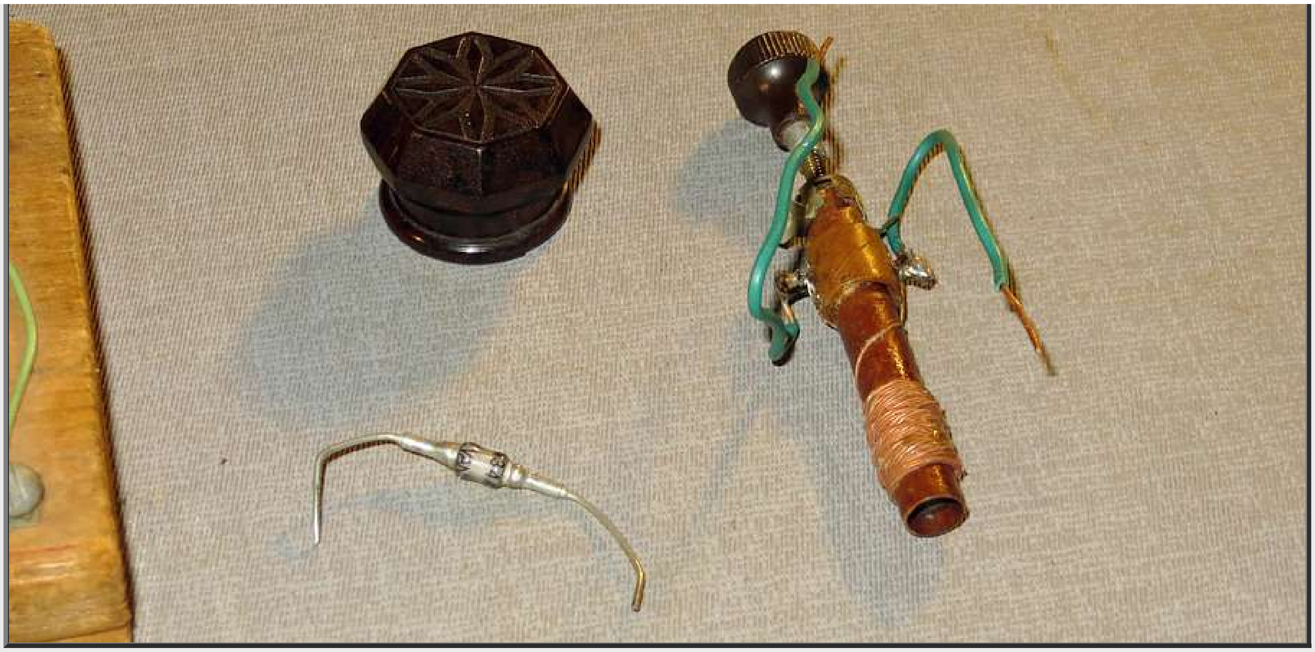

Original coil, diode and main tuning knob from the 1966 radio. The wires on the coil were soldered with a pair of red hot tweezers.

My dad gave me the large knob for the tuning capacitor. He had a second job on the weekends at "John Cusimina's Moving and Storage." I've always wondered if he pulled that knob off of somebody's radio while they were moving. I hope not. He probably did.

Next

MDS975
www.mds975.co.uk

CRYSTAL SETS 5
Experimental Crystal Sets

CRYSTAL SETS 5: EXPERIMENTAL CRYSTAL SETS

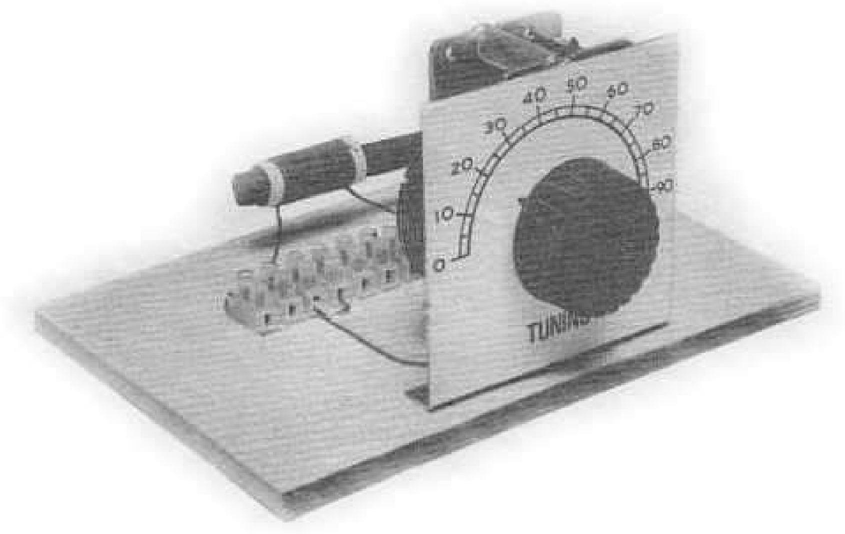

Picture 1 - The Complete Experimental Crystal Set

THE POPULARITY of the crystal radio arises from its simplicity, and the fact that it needs no power supply. The circuit here allows for easy experiments with tuning, aerial and diode coupling, and frequency coverage. Wrong connections can cause no damage to any components.

A Crystal Set is more often than not used for the reception of medium and long wave radio, but short wave reception is also quite feasible. It will normally be possible to receive some of the stronger international radio stations.

This is adapted from an article that appeared in the 1970's in Everyday Electronics, and gave me almost endless hours of fun!

BASIC CIRCUIT

The basic circuit is shown in Picture 2 below. The coil L1 can be air cored, or have a ferrite rod placed in its winding. The variable capacitor C1, in conjunction with aerial-earth capacitance, tunes the circuit to resonate with the wanted radio station frequency. The diode D1 "detects" or demodulates the radio signal so that the programme is heard in the earpiece.

This basic circuit can be modified in various ways to obtain better performance.

EARPHONE

As most constructors will be using a Crystal Earpice to listen to the crystal set it is essential that a 47k Ohm resistor is connected across the earphone terminals (TB1/1 and TB1/2 in the diagram), i.e. in parallel with the earphone, otherwise results will be very quiet.

A High Impedance headset of 20k Ohms (20,000 Ohms) may give even better results, but these are very difficult to obtain , so unless you happen to already own such a headset the Crystal Earphone with 47k resistor will be the only option. An ordinary magnetic earpiece or Walkman headphones will not work with a crystal set.

ASSEMBLY

Construction is of a 'breadboard' type using a wooden board of about 165 x 130 mm. A 12-way block connector, TB1, is used to connected together the components and this is screwed onto the wooden board. The use of a block connector provides an easy method of connecting the components together and then subsequently rearranging them as the experiments progress.

Tuning capacitor C1 is screwed to a bracket made of some scrap metal which is then also screwed firmly down to the baseboard, see Picture 1 above. Thin plywood screwed to the front edge of the baseboard would also provide a suitable method of fixing the tuning capacitor to the base. A knob with pointer is fitted to C1, and a scale is drawn and fitted behind this.

Except for C1, all connections are made by the terminals of the 12-way terminal block as shown in Picture 4. Loosen the screws with a small screwdriver, insert the bared ends of the wires, and tighten the screws. The various locations on the terminal block, TB1, are also shown in the circuit diagram, Picture 2.

AERIAL AND EARTH

Crystal receivers need a long wire aerial preferably strung outside and about 25m long, or as long as is possible to install. If this is outside it should be high and clear of earthed objects as this will improve performance.

An earth is absolutely essential for a crystal set to work properly. The earth lead can be run to an earth rod or spike that is buried to a depth of about 1 meter into damp soil. Or it may be soldered to a bare metal can which is buried in damp soil.

It is feasible, though not recommended, that the earth lead can be connected to the earthing terminal of a hi-fi system or even to the bare metal case of a personal computer that is plugged into an earthed mains outlet, but is switched OFF.

Stranded, insulated wire, or purpose made aerial wire can be used for the aerial and earth leads.

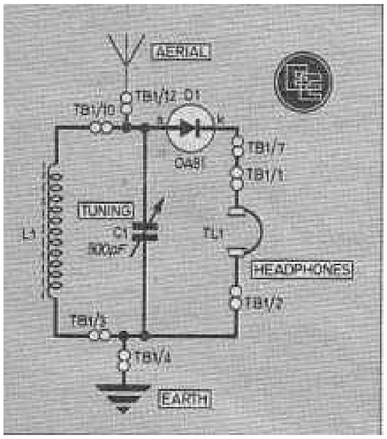

Picture 2 - The Basic Circuit

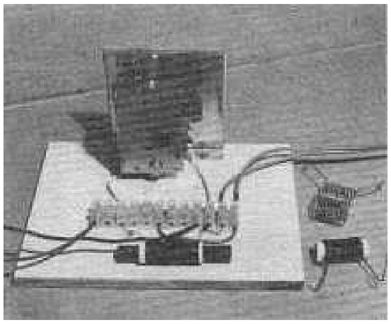

Picture 3 - Photo Of The General Layout

INDUCTORS (The Tuning Coils)

The following four coils are suggested for initial use as L1 :

Coil 1: Make a thin card tube to slide on a 10mm diameter ferrite rod, and on this tube wind about 105 turns of 32 s.w.g. enamelled copper wire, side by side. Secure ends with sticky tape.

Coil 2: Make a similar coil to to coil 1 having about 15 turns of 24 s.w.g. enamelled wire on the card tube. Loops of cotton will help hold the ends in place.

Coil 3: Wind 9 turns of 20 s.w.g. bare tinned copper wire on an object about 20mm in diameter. Remove and stretch to separate the turns, to obtain a coil about 25mm long.

Coil 4: Make a similar coil to coil 3, but with 5 turns.

The Ferrite Rod

It will be necessary to have a ferrite rod of about 60mm to 75mm long available. Coils 1 and 2 will provide reception of medium wave and the longer short wave bands. Coil 3 should cover about 3 - 10MHz shortwave with the ferrite placed in it, or about 6 - 18MHz with the ferrite rod removed. Coil 4 should cover about 6 -13MHz with the rod in, and about 9 - 20MHz without the' rod.

It will be noted that as the ferrite rod is inserted, any particular signal has to be re-tuned by opening Cl. This arises because the ferrite increases the inductance of the winding, so less parallel capacitance is needed for the same resonant frequency.

EFFICIENCY CHECKS

Tune in a m.w. transmission using coil 1 which gives good headphone volume. Place a microammeter or multi-range meter on a sensitive range in series with the headphones. A reading of 50-100uA or more may be obtained, depending on aerial, earth, earphone resistance and resistor value, coil and detector efficiency and strength of signals at your locality.

Placing the ferrite rod in the coil and re-tuning should boost the meter reading to some extent. Surplus or other detector diodes can be tried by substituting them in turn and noting the meter reading. Improvements to the aerial (or earth) will also show up as a rise in meter reading.

If experimenting with a crystal earpiece, which gives no direct current circuit, the meter may be clipped across the phone leads, i.e. D1 cathode to earth.

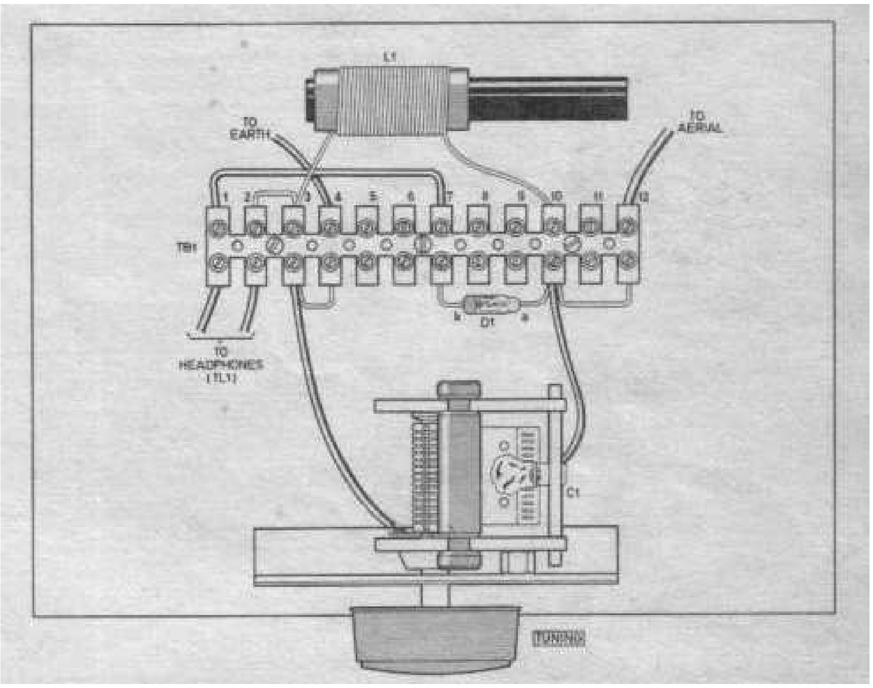

Picture 4 - Baseboard Layout Of The Crystal Set

AERIAL COUPLING

The aerial loads the tuned circuit heavily when connected directly to the top of the tuned circuit, as in Picture 2. This damps the tuning action and it can be found that stations spread out all over the dial, which is unsatisfactory.

The series capacitor, C2 connected in Picture 5(a) reduces the loading and thus improves the sharpness of the tuning. A variable or pre-set capacitor of about 250pF maximum is most suitable. for this role, though it is possible to experiment with a variety of fixed value capacitors in this range also.

Connecting the aerial to a tapping on the coil, as in Picture 5 (b) also sharpens tuning. It may also increase volume. Try about 2 turns from earth for coil 4, or 4 turns from earth for coil 3.

Another method is to have a coupling primary, as in Picture 5 (c). This consists of a second coil, with about one third the turns of the original wound on top of the existing coil.

You can even combine these methods to find what arrangement best suits the aerial in use.

The diode can be disconnected from the end of L1 and taken to a spare position on TB1 for example location TB1/9. You can then run a flying-lead fitted with a crocodile clip from this position, connecting it to various tappings on the coil as required as in Picture 5 (d). This method also reduces loading on the tuned circuit.

Coils with spaced turns of bare wire are readily tapped. For other coils, small loops can be made every ten turns or so, and crocodile clips can be attached to these when selecting tapping points.

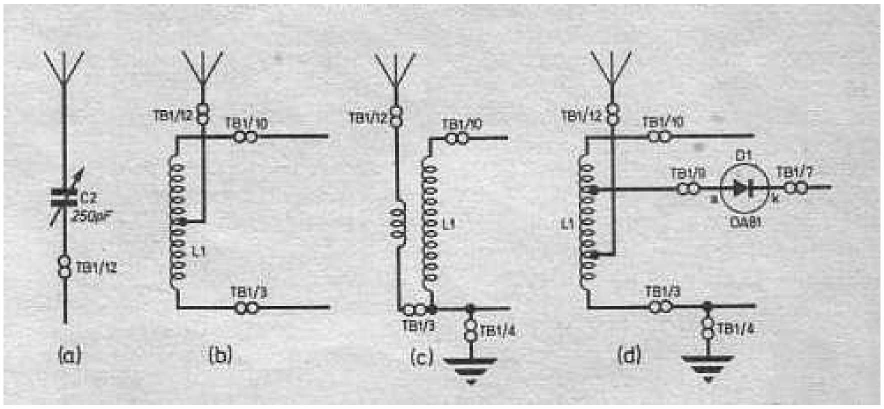

Picture 5 - Alternative Methods Of Aerial Coupling

SHORT WAVES

For shortwave reception, a good efficient outdoor aerial is certainly recommended. Evening listening in the region around 5 - 9MHz in often proves to be the most fruitful.

Since there is no amplification, as with a valve or transistor receiver, certain frequencies will seem to be completely dead at particular times of day. So if the crystal receiver works satisfactorily on medium wave and longwave, but no shortwave signals are heard, check again in the evening, or after dark, when conditions are different.

PARTS REQUIRED

C1	**365pF or 500pF Air Spaced Tuning Capacitor**
D1*	**OA47, IN34, OA81, OA90, OA91, IN94 or similar point contact small signal Germanium Diode** ***** The OA47 will be of particular interest since it has the lowest forward bias voltage of any of these diodes which will make the crystal set somewhat more sensitive and therefore louder. The US equivalent of the British OA47 is the IN34.
TL1	**High Impedance Headphones (20,000 Ohms)** **or Crystal Earphone**
TB1	**12-Way Plastic Screw Block Terminal**
Also Required:	**47 k Ohm Resistor for Crystal Earphone: Enamelled Copper Wire: 32 and 24 s.w.g. for L1: 20 s.w.g. tinned wire for L1: Ferrite Rod 10mm diameter x 75 mm long: 25m of wire for aerial: Wire and rod or spike etc for earth: Wood for base e.g. 10mm x165mm x 130mm: Scrap of metal of thin plywood for C1 bracket/front panel: Knob: Crocodile clip(s)**

Adapted from an article in Everyday Electronics magazine, November 1981, By F.G. Rayer.

HERE ARE A COUPLE OF VERY INTERESTING CRYSTAL SET DESIGNS SENT IN BY KRYSATEC - "THE RAT" - FROM THE CZECH REPUBLIC

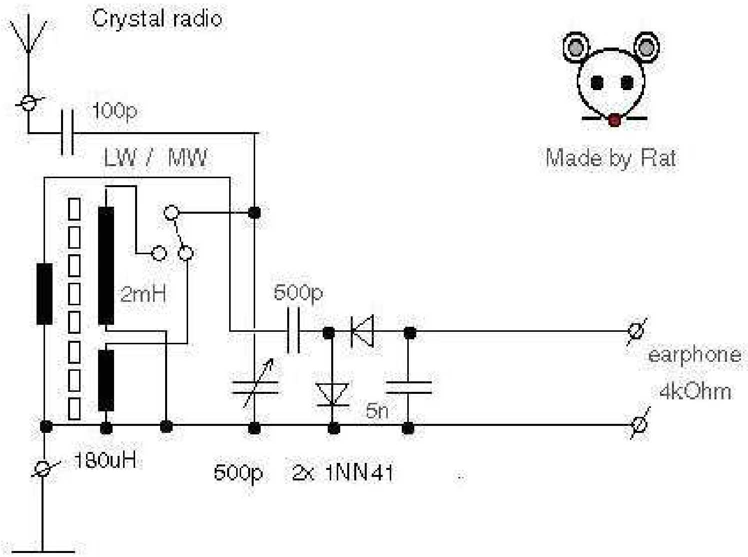
Crystal radio
100p
LW / MW
2mH
500p
earphone
4kOhm
5n
180uH
500p
2x 1NN41
Made by Rat

1/ Using old coils from old bulb radio for MW and LW band. Though it would be straightforward to wind the coils - one for Long Wave, one for Medium Wave and a coupling coil. Variable capacitor is 2 x 500pF only one half is used: 500pF. For the crystal earphone a resistor of about 82k ohm in parallel is required. This set also uses two Ge diodes as a multiplier in the quest for for higher audio signal output.

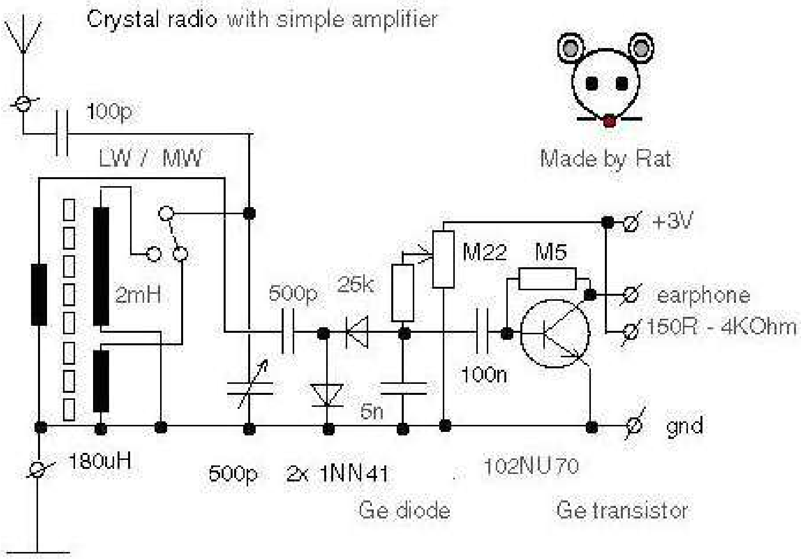
Crystal radio with simple amplifier
100p
LW / MW
2mH
500p
25k
M22
M5
+3V
earphone
150R - 4KOhm
100n
5n
gnd
180uH
500p
2x 1NN41
Ge diode
102NU70
Ge transistor
Made by Rat

2. If signals are not strong signal in your location, then the above circuit design can be considered. A simple transistor amplifier is used. A variable resistor M22 is used for better sensitivity which can be adjusted for poor signals. This crystal radio is aversion from cca 1960 - 1970 y.

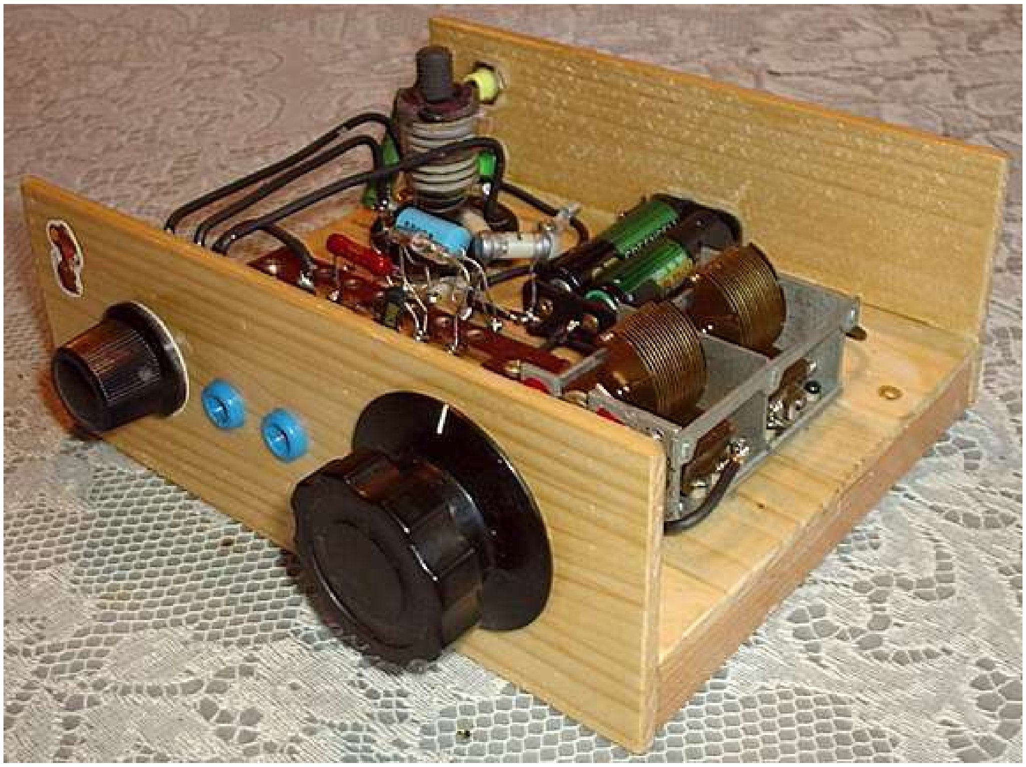

Rat's finished Crystal Set with additional amplification - very neat!

'Minilabs' Crystal Radio

BELOW: Ian Tomlinson kindly sent in a photograph of the box that contained the kit for his John Adams Toys 'Minilabs' Crystal Radio.

It is a very simple circuit consisting of the coil (inductor) with a sliding contact that provides variable tapping points, a diode and crystal earphone. All that is added is the aerial and earth. There is no variable tuning capacitor for simplicity and to keep costs down.

The coil provides the inductance required for tuning into a certain frequency (wavelength). These days a variable "tuning" capacitor is normally wired in parallel across the inductance (coil) in order to vary the resonance of the tuned circuit and therefore enable to easily tune into various transmitters on different frequencies. This crystal is tuned varying the number of turns on the coil (ie varying the inductance) by tapping off at different points using the sliding contact ("ball").

The crystal earpiece, or high Z headphone, is connected between the output of the detector diode (the other end from the coil) and earth. The volume from a crystal earpiece may be considerably improved by connecting a resistor of - somewhere between - 4.7 k and 47k ohms in parallel with the earpiece. A crystal earpiece cannot directly allow current to flow through it and the parallel resistor therefore allows current to better flow through the circuit.

John Adams Toys
Crystal Radio Kit
Make a radio complete with tuning coil, antenna and earphone.
Pack contains all you need including full instructions. No battery required.
Ages 8–14

'Minilabs' crystal set by John Adams Toys

A discussion on configurations for Crystal Sets by Felix Scerri VK4FUQ

This discussion, by Felix Scerri VK4FUQ, was posted at this address which no longer appears on the web www.tarc.org.au/techinfo2.htm (error 404) so here it is reproduced:

Crystal Set design is one of my passions closely allied with my obsession for audio and high fidelity.

My main interest in crystal sets, apart from the wonder of a radio receiver that does not require a power source, is the potential excellence of the recovered audio quality from normal AM broadcast stations.

Personally, it is one of my great laments that most people have never heard how good wideband AM can sound. A high performance crystal set or similar TRF approach is, in my opinion,the only way to do it. There are a few people around who have heard the audible results of my efforts,and can only agree.

I have often wondered,given the ultimate simplicity of the crystal set, being essentially a tuned circuit,a diode detector and some form of output device, what it takes to achieve optimum performance. What follows are my thoughts on the matter.

Crystal Set optimisation, is in my opinion, all about reduction of circuit losses. Essentially this means high "Q" tuned circuits and high quality detectors. Efficient output devices also help too. But as we will see, there are some trade-offs required as well. A high "Q" tuned circuit is always benefical, as a high "Q" tuned circuit has lowest RF losses,highest potential selectivity,and highest voltage at resonance, which is very useful for the diode being fed from the tuned circuit. Variable capacitors, even the "modern" miniature variable capacitors (although the older air dielectric units, as used in old valve receivers are more desirable) for various reasons,are generally quite efficient, and a higher "Q" coil will produce the most worthwhile improvements.The best (highest "Q") coils are wound with "Litz" wire, which is a multistranded woven wire with all strands insulated from each other. The performance of Litz wire wound coils is spectacular, unfortunately, although I know Litz wire is still being made, from personal experience, it is VERY rare in Australia.

Efficient coil design can be quite complex and all my coils are wound on ferrite rods. There seems to be,at least for ordinary single wire windings (close wound), an optimum wire thickness for optimum coil "Q". I have determined .315 mm winding wire to be about optimum for simple (single wire) coils on ferrite rods. Thicker wire is NOT better, believe it or not.

Lacking Litz wire, an interesting winding approach I have developed is to use two slightly thinner wires wound as a bifilar winding connected together at the beginning and end of the coil, yields considerably higher "Q" compared to a simple single wire winding. I have found 0.25 mm winding wire optimum in this application.

Whilst high "Q" coils are beneficial from the RF point of view, there is a possible downside. If one is interested in maximum selectivity and sensitivity, there is no problem, but remember highest "Q" results in a narrowed audio band-width as a simple consequence of band-width. For high fidelity applications this could be a disadvantage under some circumstances, although there are clever ways around this.

Regardless of ultimate coil "Q", selectivity is a major issue with crystal sets generally. Here another trade-off is evident. For the maximum voltage into the diode, connecting the diode to the high impedance end of the coil (i.e. the top) yields the greatest voltage but the selectivity is usually terrible, because of severe "loading" by the diode circuit. For this reason, tapping well own the coil improves selectivity at the expense of signal volume (reduced voltage). Once again there are ways around this. As described in my "Double Tuned Crystal Set Tuner" article in "Amateur Radio" magazine, March 2002, the use of two separately tuned coupled resonant circuits allows top connection into the diode without compromising overall selectivity, thanks to the use of a second tuned circuit which is fed from the external antenna. The whole network forms a double tuned input bandpass filter and in practice this approach works very well. For single coil crystal sets I recommend the use of an un-tuned "antenna" winding adjacent to the "hot" end of the main coil, preferably adjustable (old paper reels from sewing cotton threads are ideal). This allows the degree of coupling to be optimised under actual listening conditions. The double tuned set up is best, yielding superb selectivity, but the un-tuned antenna coil arrangement also works quite well, especially if the diode is tapped well down the main coil.Tapping halfway works well.

The other method of performance improvement involves the use of the most effective detector system possible. Here things get very interesting. In fact the temptation is to use more complex circuitry, but that gets away from the charming simplicity of the crystal set. As an example, my own crystal set tuner has at times mutated into a TRF tuner complete with FET RF preamplifiers, active(powered) detectors and other enhancements. These modifications do work well, but loses the simplicity of a basic crystal set. In actuality, a simple diode detector can work extremely well, subject to some qualification. Diodes like to work with a reasonable level of RF input voltage. Audio distortion can result under conditions of low signal level, due to diode transfer curve non linearity and other factors, such as the widespread use of broadcast station "processing". The actual type of diode makes a difference. The 1N34A germanium diode is very popular for crystal set use, although in my experience just about ANY germanium diode will work, although it is worth trying different specimens. Some are definitely better than others. Even from a pack of twenty 1N34A's from the same source, some were definitely better than others. Measuring the average value of rectified output voltage across the diode load resistor will show which diodes are best. By the way, I regard a diode load resistor as being mandatory. I find a value of about 47K about right, especially if a crystal earpiece is being used or the crystal set is being used as a tuner feeding an audio pre-amplifier and following amplifier. If using high impedance magnetic type headphones, the headphones provide the diode DC load.

Another type of diode that is very interesting, is the hot carrier diode. There seem to be a lot of different hot carrier diodes around these days. There are even hot carrier diodes now being sold as "germanium diode equivalents". I have tried them and they do work acceptably well, but they are not quite as good as genuine germanium diodes such as the 1N34A. Typical UHF mixer hot carrier diodes, such as the 1N5711 will not work well in crystal set service simply because their "turn on voltage"is too high, similar to silicon diodes such as the 1N4148/914 series, which require a lot of RF input to function adequately as RF detectors, however a simple technique can be used to turn hot carrier diodes such as the 1N5711 into superlative detectors.

I guess we are cheating a little, because the technique is to use a little voltage bias supplied via a 1.5v battery, through a simple potentiometer voltage divider arrangement, with capacitor (for DC isolation) fed into the diode from the tuned circuit. With applied adjustable bias, I find the 1N5711 diodes absolutely superlative detectors under ANY signal strength conditions. I find the detection quality also superlative, with a clarity and low noise profile unmatched by any other diode arrangement. In my opinion, hot carrier diodes, running bias,are the best detectors overall.

Regarding other detector arrangements, the diode "voltage doubler" is often recommended, however my own experiments with the doubler arrangement have been inconclusive and slightly disappointing overall. I have found no real advantage in their use over a simple (one) diode detector, believe it or not.

Yes, they do work, but they're nothing special, at least in my opinion.

Any comments on this general subject of crystal set optimisation would be welcome.

73's Felix Scerri VK4FUQ.
22nd July 2002

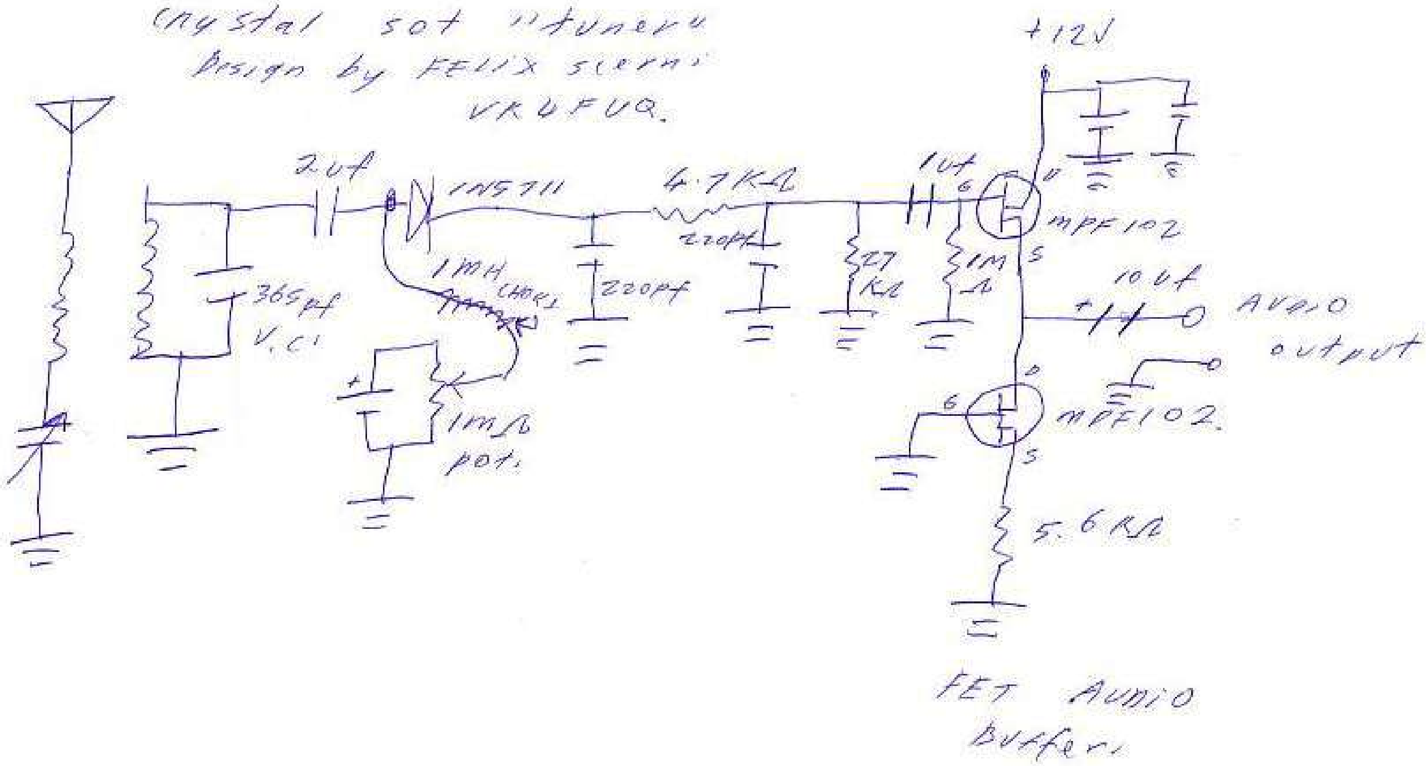

Above: CRYSTAL SET BASED CIRCUIT PROVIDING A HIGH QUALITY PROGRAMME SOURCE

IMPROVED VERSION OF THE ABOVE CONCEPT !! New update from Felix Scerri February 2010:

New 'two FET infinite impedance AM detector'

I've developed a new version of my old favourite FET 'infinite impedance' AM detector that I think sounds very nice. I include a short audio of one of our local AM stations. I picked this station as it is my reference 'torture test AM station' as they run very heavy 'processing' which normally sounds yuck with all my other (diode and non diode) detectors! However it's quite clean with this detector. What do you reckon? I'll do up a circuit if you'd like to feature it in your TRF radio section. A general draft article follows.

'A favourite non diode based AM detector that I've built and used many times over the years is the FET based infinite impedance detector, offering very good general AM detector performance, especially under weak RF signal conditions where diode based detectors do not perform well, especially in terms of audio distortion.

However one of the slightly strange things I've noticed about the simple FET based infinite impedance detector is the variable audio quality noted, even when using the same type of FET. Some I've built have sounded good and others slightly fuzzy when used with an audio preamp and fed into a high quality audio system. I've been giving this a considerable bit of thought of late and I've wondered if the audio distortion might be a result not necessarily of the detection process itself, but the FET stage in its guise as a 'source follower' audio stage which essentially, it is.

I have long been aware that as a simple audio buffer stage, the FET based 'source follower' can exhibit a considerable amount of audio distortion, and a technique I've long used to greatly reduce this audio distortion is to use a second FET in the source lead of the first FET as a 'constant current source' which serves to 'linearise' and greatly reduce audio distortion in the buffer stage overall. So, to test the theory I built a simple one FET infinite impedance AM detector which worked well, but with just a hint of audio 'fuzziness' on received AM stations. So I added a second FET in the source lead of the first FET wired as a constant current source, taking the output from the source of the first RF detector FET and the source resistor and RF bypass capacitor off the source lead of the second FET 'constant current source'. The result, totally clean audio! The theory seems proved! I call this modified detector the 'Two FET infinite impedance detector'

[))) Here's what is sounds like - click to play the audio file (((]

Here is the circuit diagram:

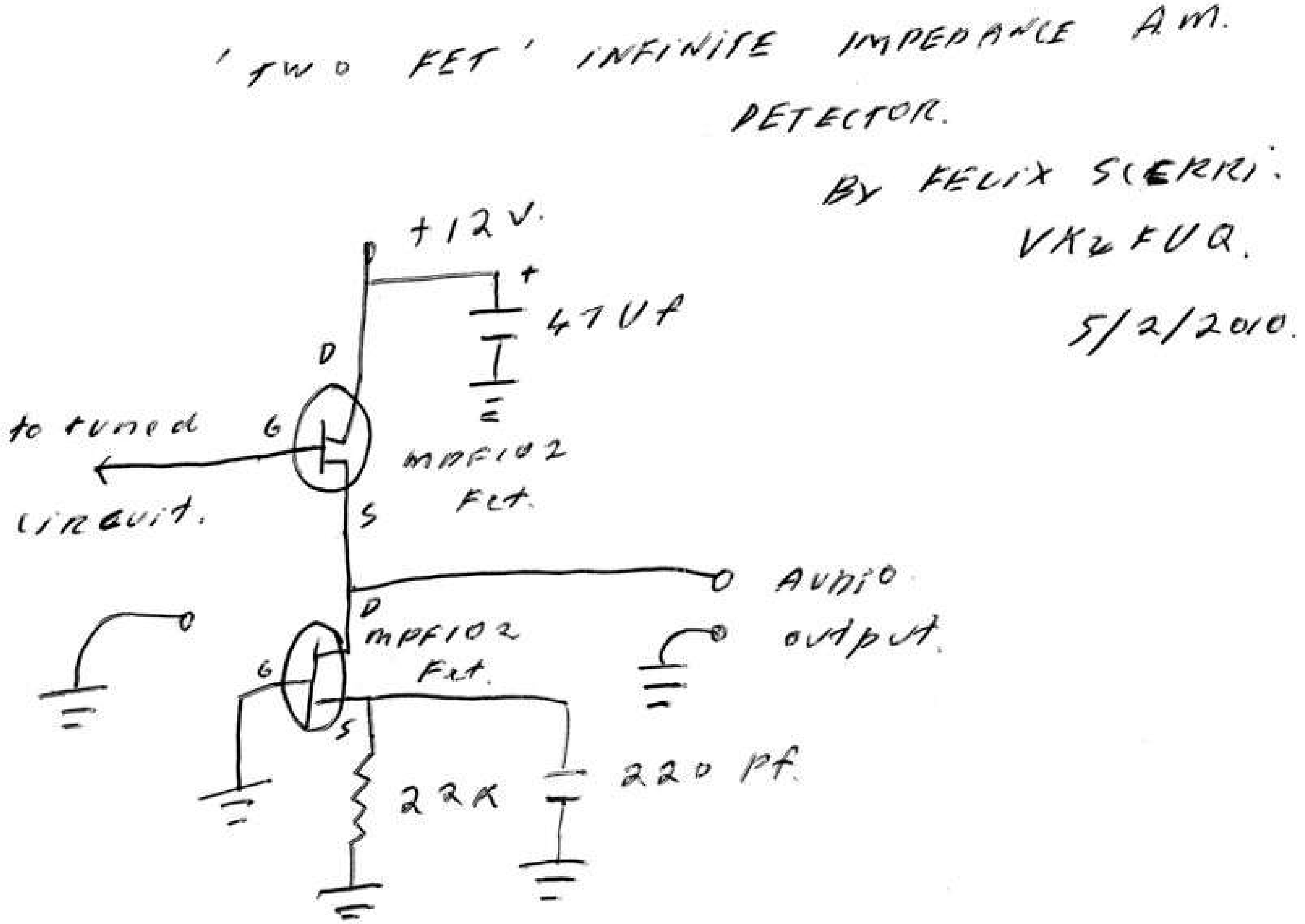

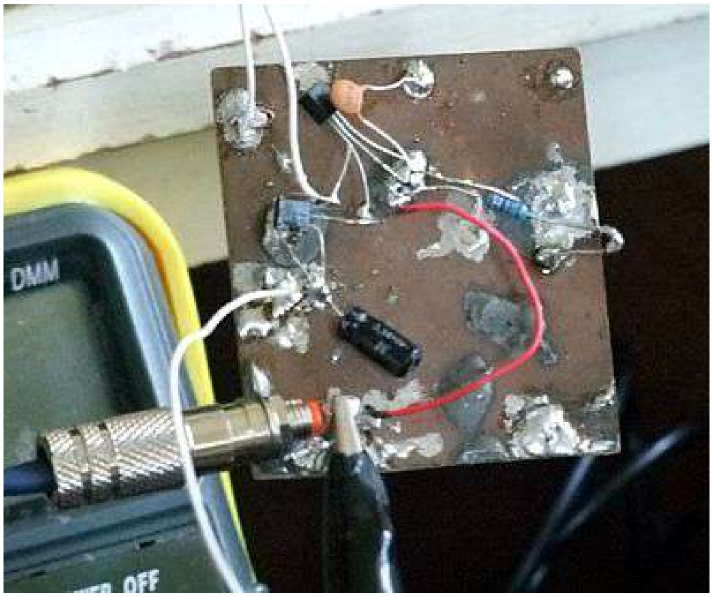

This detector has been a real eye opener for me in terms of its excellent performance, especially considering its circuit simplicity. Indeed in the past I have designed other more complex FET based infinite impedance circuits that do not quite work as well in practical terms as this latest circuit, at least according to my well calibrated ears!

I do not have access to any precise test equipment but my well calibrated ears tell me this 'two FET infinite impedance detector' is a beauty, surpassing practically every other AM detector I've built even at low RF input, and that's rather a impressive claim and the audio quality when used as an AM tuner feeding a high quality audio system is quite remarkable. Possibly the best thing about this detector is its excellent performance under weak signal conditions. Diode based detectors also work beautifully, but the use of an RF stage to ensure detection over a linear portion of the diode's curve is mandatory! This compound infinite impedance detector works beautifully on the sniff of a useable RF signal.

Just add a high Q tuned circuit and that's it!

Felix Scerri
VK4FUQ

A better FET for the 'basic' Infinite Impedance Detector:

Quite recently by accident, I've realised the MPF102 FET that I've long used in my FET based infinite impedance detectors is possibly not the best FET to use. This was the reason why I developed the 'two FET' infinite impedance detector some time ago which works very well.

However I've found the choice of a more suitable FET works beautifully in the basic FET based infinite impedance detector circuit, which has appeared for many years in many editions of the ARRL Handbook.

I use the 2N5457 and others of the same 'family' may be equally suitable, but I haven't tried them! However with a 2N5457 in place of an MPF102, the basic infinite impedance detector has became my AM detector of choice. It works beautifully even at low signal input with lovely and clean low distortion audio along with a very high input impedance for good tuning selectivity. It's a beauty! The basic generic circuit is attached, courtesy of Rod Elliott's ESP website.

73 Felix VK4FUQ
10 / 02 / 2012.

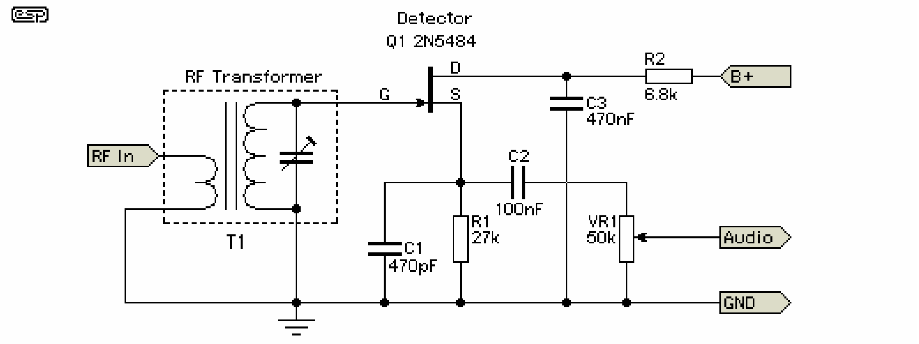

The basic generic circuit is attached, courtesy of Rod Elliott's ESP website
Felix Scerri VK4FUQ

As often happens with me, my renewed interest in FET based 'infinite impedance detectors' of late has led to some interesting new research and I may have considerably improved the 'two FET infinite impedance detector' as a result.

My research suggests that although the use of a CCS (constant current source) reduces audio distortion in an audio stage, the value of the 'source resistor' in the CCS stage is somewhat critical for best results.

By using a potentiometer in lieu of a fixed resistor I have found that a resistance value of around 470 kohms cleaned up all overall audio distortion. I used an MPF102 as the CCS in this circuit. An interesting and worthwhile little circuit refinement.

73 Felix VK4FUQ
21 / 02 / 2012.

A Minimum Component Count High Quality AM Detector

I was generally messing around with various circuit ideas and I came up with this AM detector circuit, a simple diode detector along with a FET stage. It was an attempt to provide good performance along with minimum number of components. Actually I've been pleasantly surprised at the excellent level of general performance and the best of all, it sounds great!

The circuit is quite conventional being a BAT 46 diode detector feeding an MPF102 FET buffer/ common source amplifier stage. I would ordinarily use a FET source buffer stage in this application, but opted to use a simple low gain FET 'common source' amplifier stage instead, with excellent results. I also used a BAT 46 Schottky diode instead of an ordinary germanium signal diode.

This was done for several reasons. Firstly, germanium diodes are now very hard to find but in any case these 'germanium diode equivalent' Schottky diodes are actually a superior diode, having very low noise, almost zero back leakage and an essentially complete absence of carrier storage effects and very good weak signal sensitivity. I call these diodes high fidelity diodes as they sound wonderful as RF detectors.

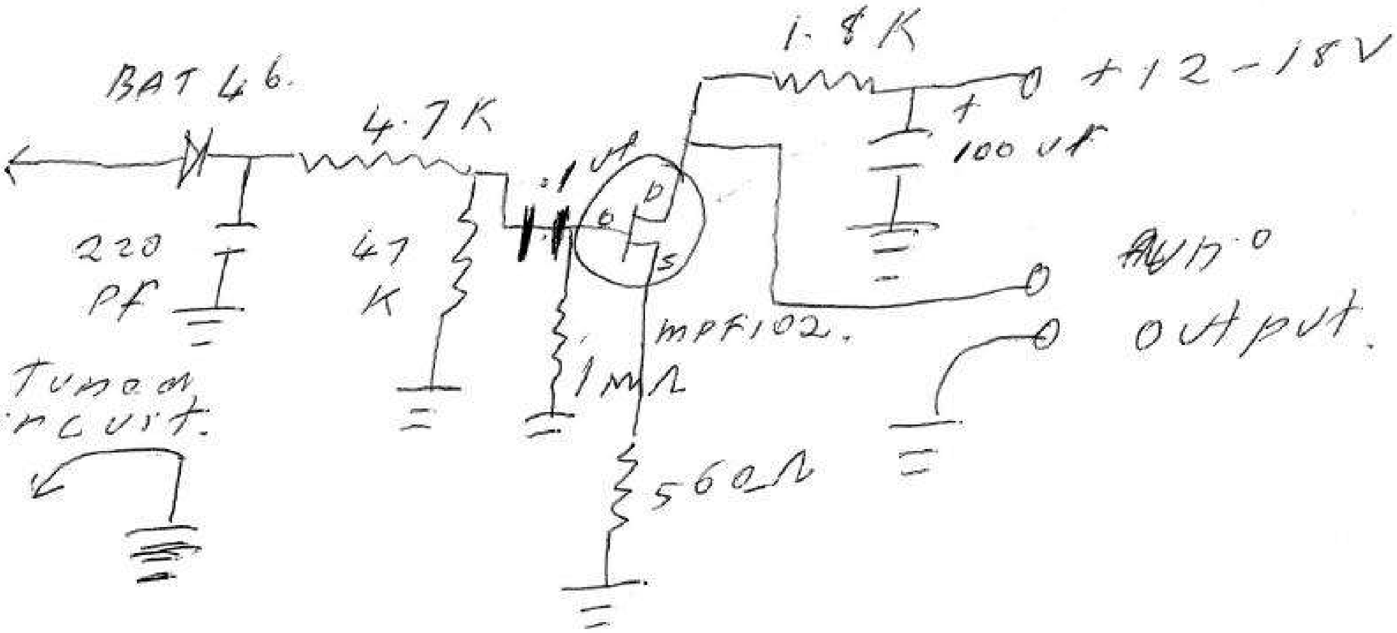

Circuit Diagram of the Minimum Component Count AM Detector by Felix Scerri

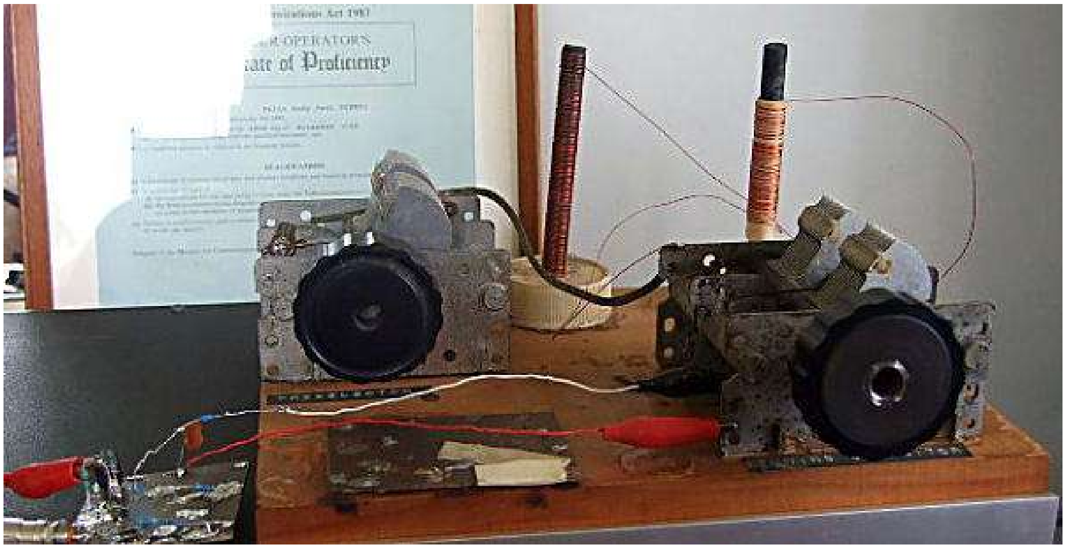

Minimum Component Count AM Detector by Felix Scerri

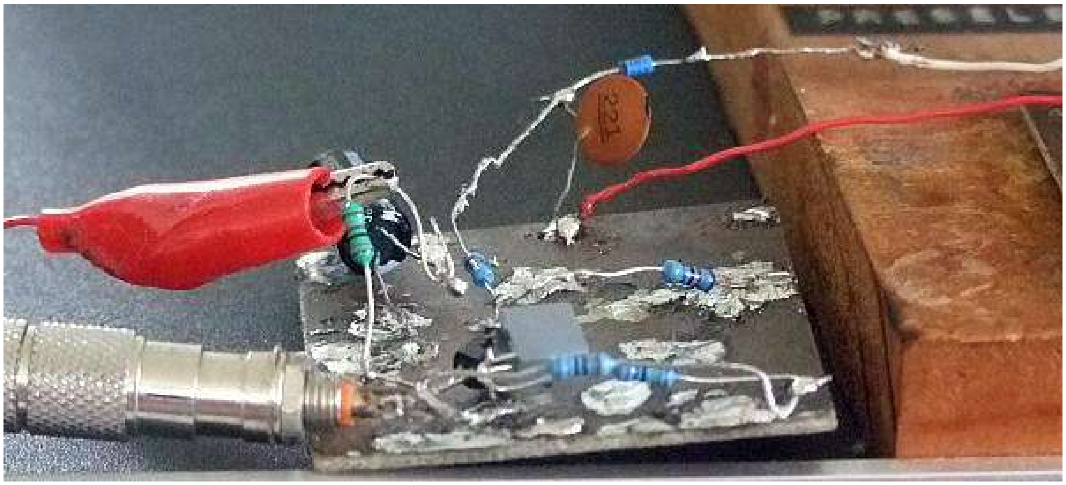

A Closer View

The high impedance of the FET's gate circuit is perfect for optimal buffering of the diode detector, something very important for good low distortion diode detector performance. Apart from providing slight voltage gain, the use of the common source FET amplifier is a new idea, as this prevents the possibility of incidental RF rectification occurring in a FET source follower stage, which can happen. A 1 uf plastic film capacitor may be added in series with the audio 'hot' output lead to block the DC offset out of the FET drain, if required.

Despite no additional RF stage ahead of the diode, audio quality on even relatively weak RF strength stations is actually very good, and of course the audio quality will be even better with increasing RF signal strength, something which will also increase the audio output level. Just on this, for a long time I was somewhat negative regarding diode detectors, as one AM station locally (the strongest one) was always distorted when using a diode detector.

I strongly suspected a transmitter fault, but my complaints were ignored, until one day some time ago when all audio distortion suddenly disappeared! Nothing was ever 'said', but I realised that my suspicions of a long standing transmitter fault were correct, after all!

73 Felix
VK4FUQ
02 March 2012

Voltage Doubler Detector

This is an AM detector circuit that I've long known about and which worked ok, but never seemed to work as well as it should have. However I spent some time late last night trying to optimise the circuit, with some success.

It is a curious circuit being essentially a 'voltage doubler' originally developed for power supply applications, and its use as an AM detector is hard to analyse! It seems that the component values in the circuit are somewhat critical for good performance and if not, the performance is rather 'ordinary'. The circuit that I eventually came up with uses a 150 picofarad 'input capacitor' with a 150 kohm 'load' resistor and loaded into a FET common source voltage amplifier stage (as previously described) through a coupling capacitor with a 1 Mohm input resistance.

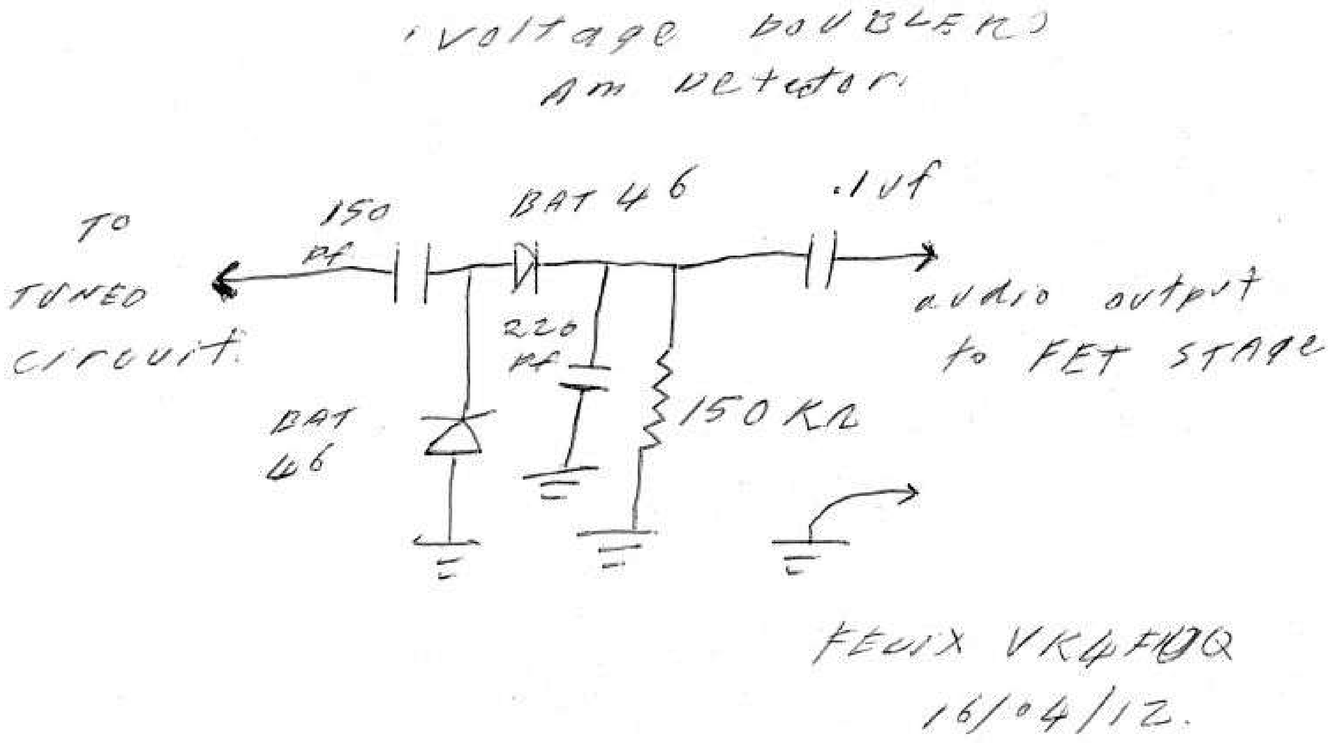

With these circuit values, it all works 'quite well'. Give it a go! It's an interesting AM detector with quite good 'sensitivity' and clean audio quality, and it seems to work well at low signal levels.

73 Felix
vk4fuq.

16th April 2012.

Simple AM detectors: What works best? A practical experimenters viewpoint.

I have written a lot about simple AM detectors for use as tuners for feeding into an audio amplifier, and it has been a long time interest. These days I use either diode based or 'infinite impedance' types of AM detectors. In this location our 'local' AM stations are quite distant and are therefore quite weak in terms of signal strength.

As such I find infinite impedance detectors based on field effect transistors give consistently better results for tuner applications due to their lower apparent overall detector distortion. Diode based detectors are quite 'fussy' as they require both optimal output buffering (AC/DC ratio) and an 'adequate' (beyond the diode knee) level of RF signal injection. http://www.tonnesoftware.com/appnotes/demodulator/diodedemod.html

Diode based detectors will happily 'detect' at very low signal levels, however the (inevitable) audio distortion that results, can be extremely irritating to the ear! Under these conditions I find infinite impedance detectors (even without additional RF preamplification and subject to individual FET characteristics), generally sound 'cleaner' and more pleasant to the ear.

FET's of course require a power source for operation whereas diodes are passive (un-powered) detectors (most of the time), however this is of no real advantage in a tuner application as an 'active' audio amplifier stage will generally be required anyway for audio level boosting, buffering etc.

In the end it will come down to a consideration of prevailing RF signal levels and other related circuit considerations at one's location. If local RF levels are strong, a well designed diode detector will give excellent results. If not, an 'infinite impedance' type of detector is most likely the better option unless one goes towards the option of additional RF preamplification prior to the diode detector.

73 Felix
vk4fuq.

17th April 2012.

ANOTHER DETECTOR by Felix Scerri !

I’ve tried diode based ‘voltage doubler’ (or more correctly ‘diode integrator’) AM detectors before with indifferent results, however the other day, just trying a few ideas I came up with this version that works rather well, with low audio distortion, high audio output and really ‘nice’ audio quality and the best of all, it seems to work very at very low RF input level.

The two diode ‘voltage doubler’ detector using two BAT 46 silicon schottky diodes feed directly into a MPF102 source follower stage set at 1 Megohm input resistance. The ‘input’ capacitor feeding the diodes from a tuned circuit is 68 picofarads.

I have the simple RF filter right on the output of the FET stage. In that respect this circuit is vaguely similar to the old ‘Selstead-Smith’ valve AM detector of the past. An interesting one! I am very happy with its general performance. Regards, Felix vk4fuq 11/04/2013

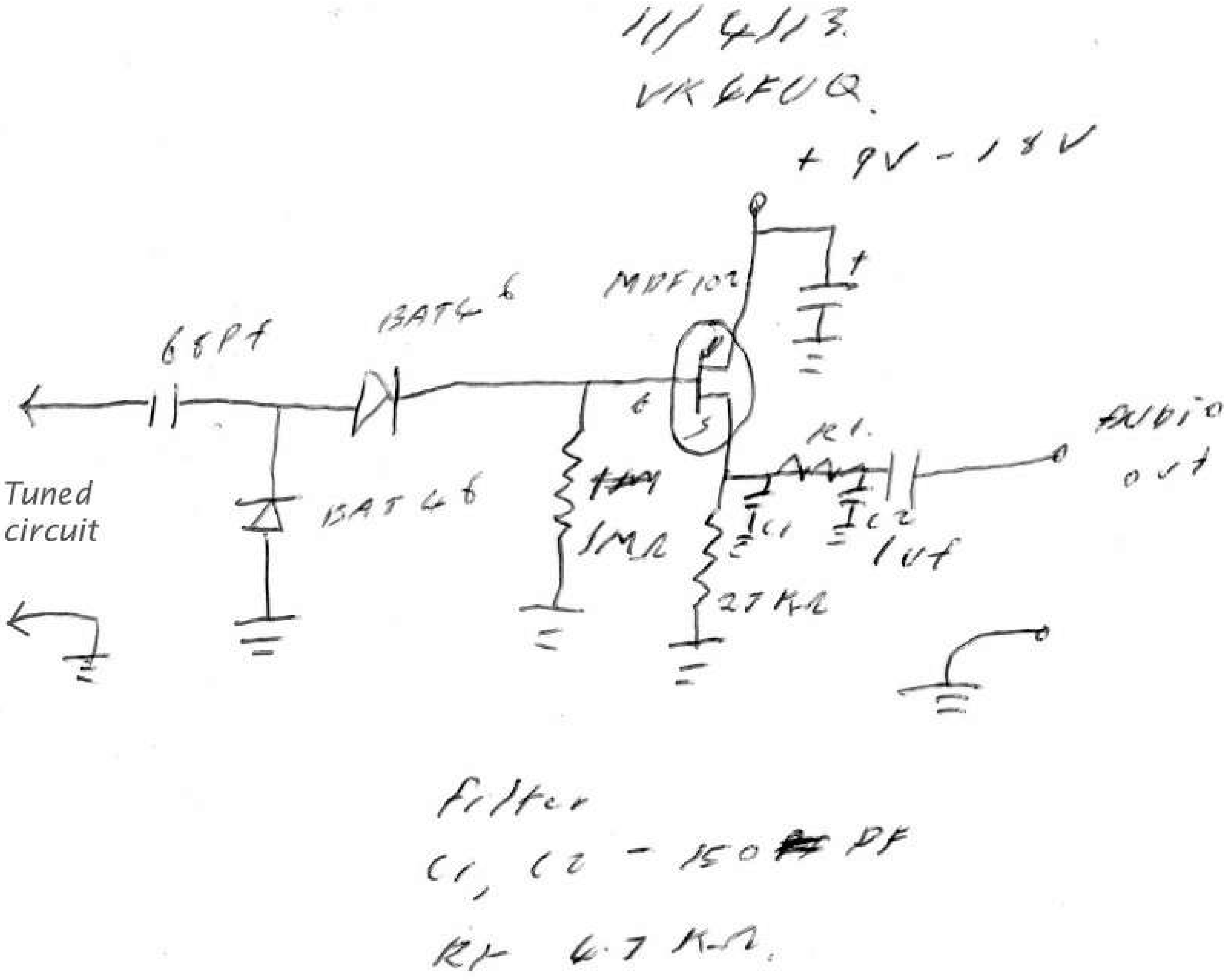

Felix Scerri VK4FUQ

UPDATE - JUNE 2013

G’day all, readers may recall the two FET infinite impedance detector I developed some time ago. That circuit worked well, but some samples of the MPF102 regretfully produced distorted output. However a recent discovery has resulted in an improved version that has truly exemplary performance.

The MPF102 is in all honesty a device essentially designed for VHF applications, but is often used in simple audio applications albeit with occasionally indifferent results, due to general device parameter ‘spread’. More or less by empirical trial and error I have found that a simple change in source resistor value to 100 k (from a much lower value), pretty well fixes everything.

In a FET source follower circuit this is interesting as the output impedance is actually a lot lower than 100 k due to the action of the FET’s transconductance. It is similar to the action of a bipolar transistor emitter follower circuit. In fact a simple infinite impedance detector with a 100 k source resistor actually works very well and will be good enough for many applications, however the addition of a second FET as a constant current source does markedly reduce overall audio distortion in the stage overall, and produces very clean and low distortion audio quality and also reduces the output impedance considerably (good for tuner applications), so take your pick!

Regards, Felix vk4fuq 02/06/2013.

UPDATE - JANUARY 2014

A simple modification to the basic FET 'infinite impedance' AM detector that dramatically improves performance.

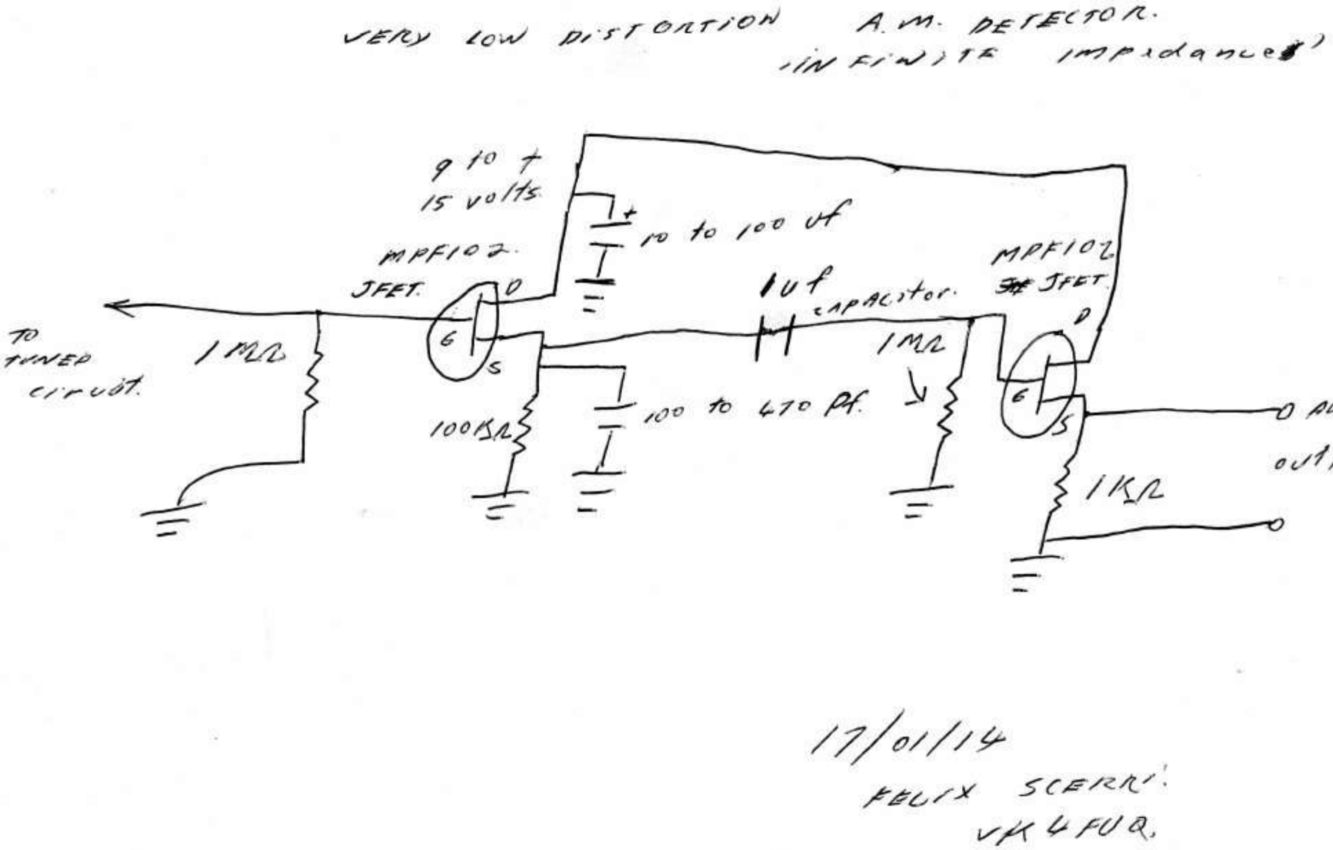

G’day all, over the Christmas break just messing around I worked out (mostly by accident), a simple modification for the simple FET based ‘infinite impedance’ detector circuit that dramatically improves weak signal performance and also greatly reduces audio distortion.

Essentially by the addition of another FET ‘buffer’ stage, another source follower, capacitively coupled from the first detector stage. The circuit is actually a simplified version of the circuit that I described in this link, http://sound.westhost.com/articles/am-radio.htm (figure 6) and testing the two head to head, they both sound superb and the simpler version is actually somewhat easier to build. I cannot get over how low distortion and ‘nice’ the recovered audio sounds. It is a joy to listen to!

Regards, Felix vk4fuq 12/01/2014.

Diodes for ‘weak signal’ crystal set applications.

As I am primarily interested in using crystal sets as AM tuners for feeding into a preamplifier/amplifier and loudspeakers, the actual type of diode can be relatively critical. In a strong signal area, not so much, but in a weak signal area such as where I reside, definitely. A diode with a good ‘square law’ performance (the area below the ‘knee’ in the diode curve), generally results in much cleaner and lower distortion than other good performing diodes, and believe me diode distortion under weak conditions, even with optimised diode ‘buffering’ used is very nasty sounding to the ear!

Testing many, many diodes in actual working crystal sets and listening critically to the audio output, it seems to me that the best diodes to use under weak signal conditions are the so called ‘gold bonded’ germanium diodes. I have sampled many different gold bonded germanium diodes and they all work well in this specific application, although sometimes the rectified output voltage may not be as high as other germanium or silicon schottky diodes, however the audio quality is much cleaner and shows much less apparent distortion!

I have tried OA5, OA47, IN141 and several ‘CG’ gold bonded germanium diodes with consistently excellent results. Other ‘ordinary’ germanium diodes may also work well under weak signal conditions, but they will need to be tested individually to check actual performance in a working circuit. One thing that I have noticed about germanium diodes is that due the ‘point contact’ nature of their construction, even diodes of the same specific type can exhibit rather different levels of performance!

The otherwise very good BAT46 silicon schottky diode works extremely well at good RF input level but not so well at weak RF input, especially when heavy broadcast ‘processing’ (commonplace these days), is used. New ‘gold bonded’ germanium diodes are probably no longer made although I am aware that they can be purchased through vendors over the internet. Apart from that, they may be found in old gear. As stated earlier, ordinary germanium diodes may be quite good but will need to be checked individually. Diodes are complex things!

Felix (vk4fuq). 29/01/2014.

* * * * * * * * * * * * * *

Just when I thought that I’ve tried everything I realised that I’ve not tried a voltage doubler (diode integrator) detector with an RF stage in front. It is an unfortunate fact that all diode detectors ‘need’ a good RF injection level (and buffering) or audio distortion becomes ‘bad’ (an understatement!)

Consider yourself very lucky if you live in a strong signal area! (unfortunately I don’t). Anyway I quickly built a couple of prototypes which worked ok but not as well as hoped. Initially I tried using a source follower buffer (no voltage gain), but converting it to a common source amplifier (with appreciable voltage gain), seemed no better, which was strange.

After staring at my prototype for what seemed like forever with my magnifiers on, I (finally) realised my mistake. A common source stage takes its output off the ‘drain’ terminal of the FET, not the ‘source’ terminal as it does with the FET source follower buffer stage. I shifted one wire, and all worked as expected and it sounds fantastic! All diode detectors (one diode envelope detectors, doubler detectors etc), need good RF input and a simple untuned FET RF gain stage works very well. As to the sound, it sounds great.

73 vk4fuq 10/03/2014.

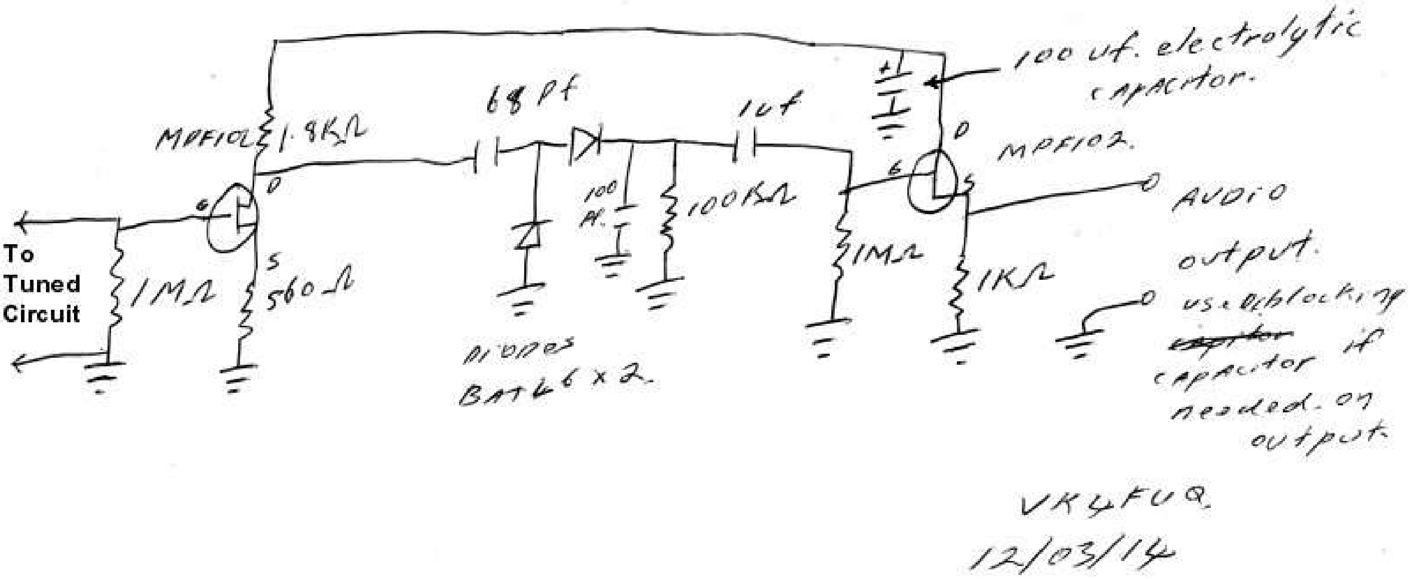

Felix Scerri's FET Crystal Set with Voltage Doubler

A 'nice sounding' AM detector - update October 2015

Looking through these pages the other day I was slightly shocked to realize that I've completely forgotten about some of the circuits that I've contributed in earlier days! They all work pretty well, however in my advancing older age and a little like my tastes in high fidelity generally, I'm starting to show a particular preference for 'nice sounding' bits of audio gear.

I guess that this also means low noise/low distortion too, although with AM broadcasting, at least in this country and probably elsewhere in the world too, the very common use of broadcasting 'processing' tends to make it hard for AM detectors generally, often resulting in a 'hard/compressed' sort of sound although still low in distortion, is not 'nice sounding', if that makes any sense!

Well of all the AM detectors that I've tried and/or developed, only one sounds 'nice' when confronted by heavy broadcast processing and that is an AM detector 'based' on a voltage doubler/diode integrator circuit (similar to the above circuit, actually the 68 pf capacitor should be changed to .1 uf, for slightly greater output).

It has taken me a very long time (years) to 'optimise' this circuit, but as it presently stands this is my favourite 'nice sounding' AM detector, and gets most use for general high quality listening on the AM broadcast band. It sounds really good! It somewhat reminds me of an old OA47 gold bonded diode detector from years ago before such abhorrent 'processing' became commonplace!

Ah yes the OA47….now that is a lovely sounding detector diode!

Regards, Felix vk4fuq. 17 / 10 / 2015.

* * * * * * * * * * * * * * *

That's it for crystal sets. I hope you try building one, it's easy and great fun!

See some useful links below....

73's
Mike

Crystal Sets (Part1) | Build Your Own Crystal Set (Part 2)

Spider's Web Crystal Set (Part 3) | Crystal Set By Kenneth Rankin (Part 4) | Crystal Radio Links

Amateur Radio

MØMTJ Visit my amateur radio pages MØMTJ

Visit my Amateur Radio Pages >

No AM radio stations or transmitters in your locality or country?

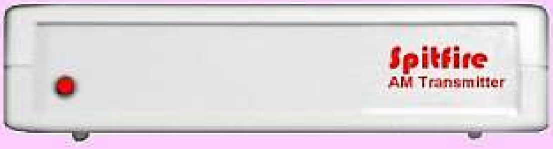

Has your local medium wave broadcast station closed or been moved to VHF/FM or Digital? Don't worry. You can still build and experiment with crystal sets and TRF radios by also buying or even building a simple low power AM transmitter. So, not only can you use your crystal sets but you can also run your own radio station that can be heard in and around your home - playing the music or programmes that you want to hear!

SSTRAN AMT3000 Superb high fidelity medium wave AM transmitter kits from SSTRAN. Versions available for 10kHz spacing in the Americas (AMT3000 or AMT3000-SM) and 9kHz spacing in Europe and other areas (AMT3000-9 and AMT3000-9SM). Superb audio quality and a great and well designed little kit to build: http://www.sstran.com/pages/products.html

POWER
GAIN
MODULATION
COMPRESSION
SSTRAN
AMT3000

http://www.sstran.com/

Other AM transmitters available:

Spitfire & Metzo Complete, high quality ready built medium wave AM Transmitters from Vintage Components: http://www.vcomp.co.uk/index.htm Vintage Components offer a choice of the high quality Spitfire and Metzo transmitters:

SPITFIRE AM Medium Wave Transmitter with 100 milliwatt RF output power:

http://www.vcomp.co.uk/spitfire/spitfire.htm

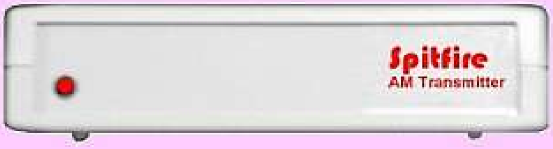

METZO AM Medium Wave Transmitter with built in compressor:

Gain
Comp
Mod
Metzo

http://www.vcomp.co.uk/metzo/metzo.htm

AM88 LP A basic AM transmitter kit from North County Radio.
http://www.northcountryradio.com/Kitpages/am88.htm

LINKS: Fine links to more Crystal Radio websites here

Component Suppliers: Links to electronic component suppliers here

Amateur Radio
M0MTJ Visit my amateur radio pages M0MTJ

Visit my Amateur Radio Pages >

Crystal Sets (Part1) | Build Your Own Crystal Set (Part 2)

Spider's Web Crystal Set (Part 3) | Crystal Set By Kenneth Rankin (Part 4) | Crystal Radio Links

^Top Of Page

Home | Contact | Site Map | Radio Stations & Memorabilia | Amateur Radio

Crystal Sets Introduction | Resistor & Capacitor Conversion Tables

www.mds975.co.uk © 2004 - 2015

CRYSTAL SETS 2

Some Practical Designs
MAKE YOUR OWN
CRYSTAL SET !!

Home | Contact | Site Map | Radio Stations & Memorabilia | Amateur Radio

Crystal Sets Introduction | Resistor & Capacitor Conversion Tables

Crystal Sets (Part1) | Spider's Web Crystal Set (Part 3

Crystal Set by Kenneth Rankin (Part 4) | Experimental Crystal Sets (Part 5) | Crystal Radio Links

CRYSTAL SETS 2: SOME PRACTICAL DESIGNS

I hope that you attempt building one or two of these crystal set designs and I really do recommend that the components are carefully connected up using soldered joints onto a piece of tag-strip for reliability. However if you are new to constructing such electronic circuits then some simple solder-less techniques could be employed and these are suggested at the bottom of the page. Also see Crystal Sets Part 5 for more ideas on experimenting with crystal sets.

An early and very basic crystal set would have been nothing more than a coil of wire, perhaps 50 -100 turns, wound around a cardboard tube about 3 inches (7cm) in diameter, a detector (or cats whisker) and a pair of special High Impedance headphones (as discussed in part 1).

There would be a very large aerial strung up around the garden and the all important connection to earth.

The coil would have tapping points (connection points) at intervals of around 5 or 10 turns. See the circuit diagram on the right for details of who the set is wired together.

The tapping points on the coil allow the set to be tuned to different frequencies by adjusting the position of tap B. Tap B would be connected to the coil at differently positions by way of a crocodile clip. The fewer turns between the top (aerial end) of the coil and tap B, the shorter the wavelength received (ie the higher the frequency). Tap A would allow the detector to be connected at different positions to vary performance. There is an additional component drawn in the above diagram, the *capacitor* (value 1000pF), this is included in crystal sets that used the *High Impedance* magnetic headphones, and bypassed any remaining radio frequencies (RF) to earth.

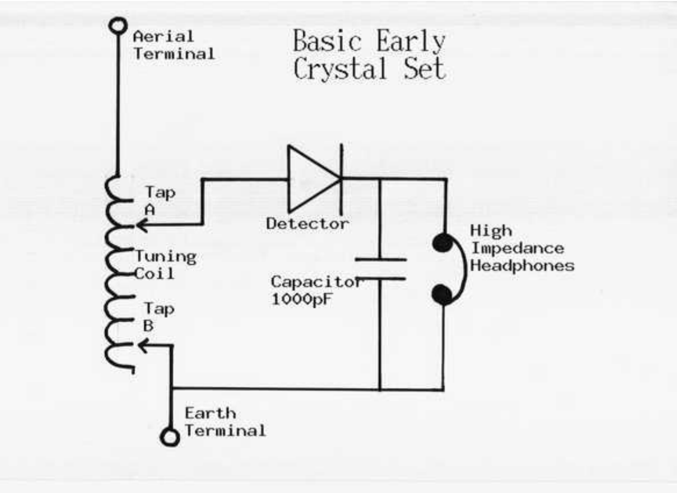

A very basic crystal set circuit.

I have not built the set described above as it is so basic. Such a crystal set above would probably have been adequate in 1920 - 1923 when there would have been only one local transmitter receivable.

When the BBC expanded transmissions and it became possible to hear more than a single station it would have became necessary to include a more convenient means of tuning the set.

This was achieved by including a Variable Tuning Capacitor, of about 500pF (0.0005uF) connected in parallel with the tuning coil forming a *tuned circuit*. The tuning capacitor would have a Bakelite knob on the spindle to aid tuning.

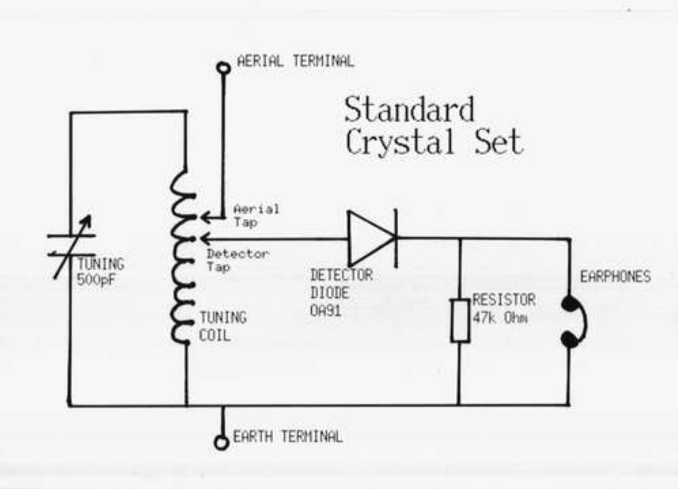

The Standard Crystal Set

Because of the simplicity of crystal sets, it is often difficult to separate stations. When tuned into one station it is often possible to hear another close by station in the background, this is due to lack of *selectivity*. This can be reduced somewhat by adjusting the positions of the Aerial Tap and Detector Tap. Moving them closer to the bottom of the coil, the earthy end, reduces the load on the tuned circuit and this improves selectivity, however it does also reduce *sensitivity* which can make the station quieter. Headphones will often swamp a tuned circuit and reduce its selectivity (Q factor), so moving the tapping point lower down improves this situation. Every circumstance is bound to be different though so the best balance has to be found by experimentation. My crystal set has both the diode and the aerial connected to the same tapping point on the coil, about a quarter of the way down.

The modern 'standard crystal set' shown above uses a Crystal Earphone, since suitable high impedance magnetic headphones (of 2000 to 4000 ohms) are no longer widely available. When using a crystal earpiece the 1000pF capacitor shown in the first diagram can usually be omitted an in its place a 47k ohm resistor is connected, this ensures that the Crystal Earphone will work at its most efficient i.e. the sounds will be as loud as possible. The resistor allows DC current to flow through the circuit efficiently - this would otherwise be blocked when using a crystal earphone. In a modern crystal set the detector used is a *Diode.* Suitable diodes include OA80, OA81, OA90 OA91 and IN94 which are usually available from component stockists.

A Better Diode For Increased Efficiency

The OA47 will be of particular interest since it has the lowest forward bias voltage of any of these diodes which will make the crystal set somewhat more sensitive and therefore louder. The US equivalent of the British OA47 is the IN34.

On the right you will see my real working example of a crystal set

The large plastic knob on the front turns the variable tuning capacitor. This set receives the three UK national stations and also three local radio stations very well at my location.

There is a small 3.5mm jack socket mounted on the front of the plastic case (MB5 from Maplin Electronics) that the crystal earphone plugs into.

The coil can be seen inside the case, it is 70 turns of 30 gauge enamelled copper wire wound around the centre of a toilet roll and tapped every 10 turns, by scraping off the enamel insulation and making a small twist. The croc' clips can be seen clipped on to these twists to connect to the aerial and detector tap points.

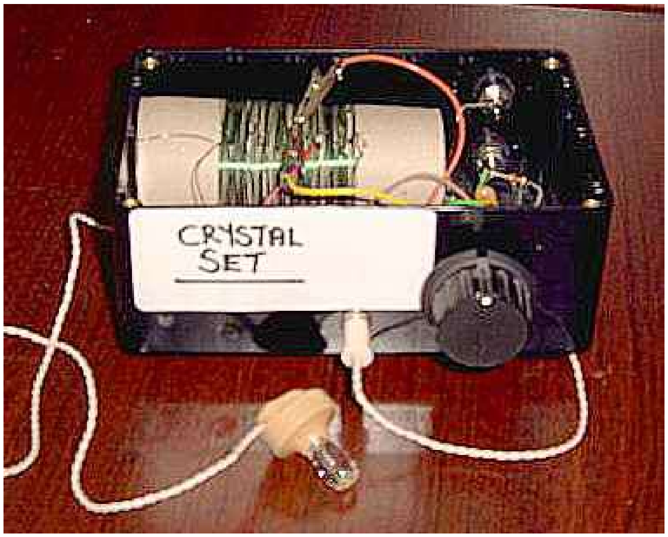

A real working crystal set. Radio as if by magic with no battery or mains power.

THE MEDIUM WAVE COIL - MORE DETAILS

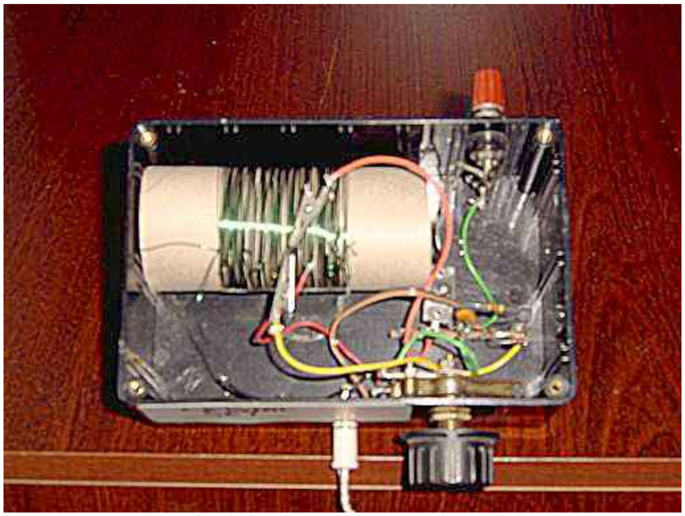

PHOTO SHOWING THE INSIDE OF
THE COMPLETED CRYSTAL SET

Medium Wave Coil

The number of turns of wire required on the coil will vary depending on the size of the former (in this case the inside toilet roll) and the thickness of the wire. So to obtain the correct coverage of the medium wave band may need a little experimentation.

I usually find that between 50 to 90 turns is right and I generally use enamelled copper wire that is between 30 s.w.g. and 26 s.w.g (i.e. 0.315mm and 0.45mm diameter), so it's best to start with too many turns and then work down.

The more turns that you use the lower the frequency range will be, i.e. too many and the coverage of the top end of medium wave around 1500 - 1600 kHz will be lost, while too few and the coverage down to 500 kHz will be lost.

It is also important that the coil former is non conducting, i.e. not metallic. It could be wood or cardboard or a short piece of PVC piping and with a diameter of between 1½ and 4 inches (4 to 15 cm) are common sizes. You could try using a ferrite rod too, see below.

This particular set has a coil wound onto a toilet roll tube which consists of 70 turns of 30 s.w.g. (0.315mm dia) enamelled copper wire tapped at every 10 turns. It also has the additional small trimmer capacitor that helps match the aerial to the tuned circuit thereby improving selectivity, see below.

USING A FERRITE ROD AS THE COIL FORMER

The aerial coil could be wound onto a ferrite rod.

A piece of 10mm diameter ferrite rod of between 3 and 6 inches long (80 to 150mm) will be most suitable and will require between 50 and 90 turns of enamelled copper wire to provide coverage of the medium wave band: First make a paper tube that is held together with sticky tape that will easily slide up and down the ferrite rod. Then wind the coil over this with the windings neatly side by side. Make tapping points every 10 or 15 turns so that the aerial and diode tapping points can be adjusted.

Adjustments to the tuning range can be made by removing some wire from the coil so it is best to start off with too many turns and then work down. Fine adjustments can be made to the completed coil by sliding it up and down the ferrite rod.

AN IMPROVEMENT TO THE DESIGN

The crystal set above also has one small, but significant, improvement over the standard crystal set and that is an Aerial Trimmer. A trimmer is a variable capacitor, very similar to the tuning capacitor, except smaller and adjusted with a screwdriver.

The value of the trimmer is usually around 10 - 50pF, but if a small tuning capacitor is available that will probably be just as effective. In the absence of such a variable capacitor, individual fixed ceramic capacitors of e.g. 10pF, 50pF and 100pF can be tried in this position to judge which gives the best results with the particular aerial being used.

The trimmer capacitor adjusts the coupling to the tuned circuit, reducing the load of the aerial on the tuned circuit will improve the selectivity (Q), and it will be easier to separate stations. Again tapping points are used and I find this to be an excellent arrangement.

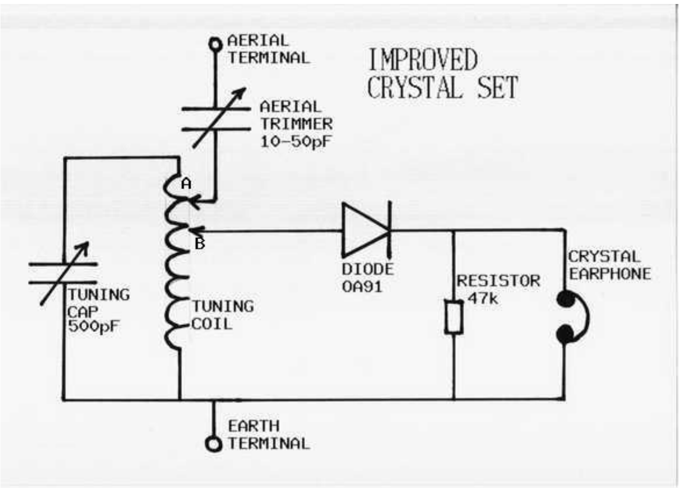

Improved Crystal Set design, with good selectivity

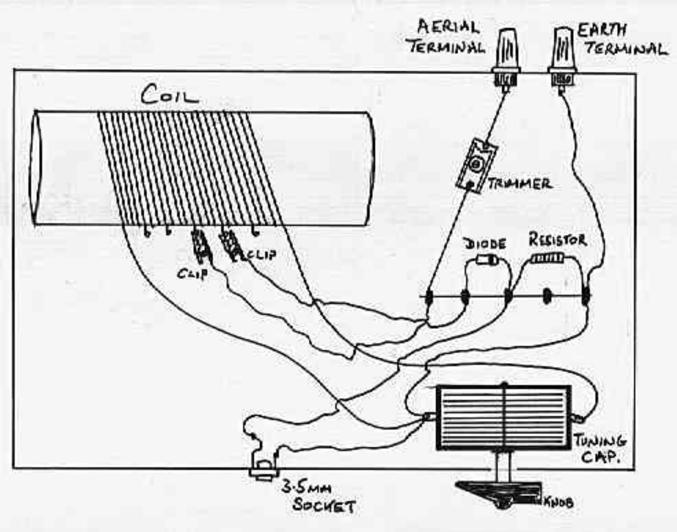

Layout Of The Crystal Set - Although this is soldered together an alternative to tagstrip would be a 5amp mains connector block so that components can be trapped in place with screws. See article below.

The picture on the right shows the general layout of the crystal set above. The coil is of approximately 70 turns is wound on the centre of a toilet roll, and has tapping points at 10 turn intervals.

The trimmer is soldered between the Aerial terminal and the piece of 5-way tag strip, and a wire goes from there to a croc' clip which is clipped onto a tap on the coil. The Diode is also soldered onto the tag strip, one end connected to a piece of wire going to a second croc' clip & connected to a tapping point on the coil, the other end of the diode is connected to the 3.5mm jack socket that the Crystal Earphone plugs into.

The 47k resistor is also connected to the earphone end of the diode and goes to earth, the earth terminal wire is soldered to the tag strip at this point too. The tuning capacitor has two terminals, one connected to each end of the coil, and one of them is also connected to earth as shown. [Where the wires cross over in the diagram, they do not touch and are not connected together].

LONG WAVES

In most areas around Europe and certainly around much of the UK you will be able to hear a Long Wave station. To receive Long Wave on a crystal set will require an aerial coil with a greater number of turns to increase its inductance.

As a good general guide a coil wound on a piece of 10mm diameter ferrite rod will require about 250 turns of enamelled copper wire: First make a paper tube that is held together with sticky tape that will slide up and down the ferrite rod. Then wind the 250 turn coil over this, the windings will have to be made over the top of each other. Make tapping points at, say, 50, 75 and 100 turns to tap the aerial and diode to.

As with the medium wave ferrite rod aerial, adjustments to the tuning range can be made by adding or removing some wire from the coil, and fine adjustments can be made to the completed coil by sliding it up and down the ferrite rod. The longer the ferrite rod the better and anything between 3 and 6 inches long (80 to 150mm) will be very good.

SHORT WAVES

If you like experimenting, then reducing the number of turns on the coil to say 10 to 30 will allow reception of the higher frequencies, the Short Waves. I have found

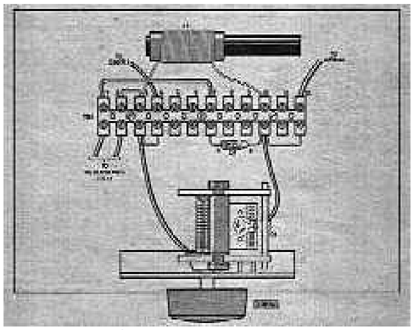

Example of a layout using a connector block to wire up a crystal set

that winding the coil around a 'ferrite rod' often works even better with short wave reception.

Obtain a ferrite rod about 7 to 15 cm long and about 1cm in diameter. Make a couple of small tubes of card, about 4cm long, that will fit tightly over the rod.

On one tube wind two coils using 0.5mm diameter enamelled copper wire - one coil of about 30 turns and a second one of 2 or 3 turns wound over the top of the first. Secure the windings in place with Sellotape.

On the other card tube wind a similar coil, but use about 15 turns for the first coil and for the second coil wind about 3 to 4 turns over the top, and secure with Sellotape tape.

These coils will provide coverage of short wave in two bands using the first coil for the longer wavelengths, typically 60 to 31 metre bands and the second coil for the shorter wavelengths typically 25 to 19 metre band. Wire up the circuit as shown in the circuit diagram below.

USING A TOROID INDUCTOR FOR SHORT WAVES

Even better selectivity performance can be achieved by winding the inductors (coils) on a ferrite toroids (T50-2 yellow, or green will do). The aerial trimmer need not be used if selectivity and sensitivity is found to be adequate. It's all about experimenting, and I find it best to use a trimmer or small coupling capacitor to obtain the best selectivity.

Up to 30 turns of 0.5mm enamelled copper wire can be used for the longer short waves below 10 MHz, while a winding of around 15 turns will provide coverage of the shorter short waves above 10MHz.

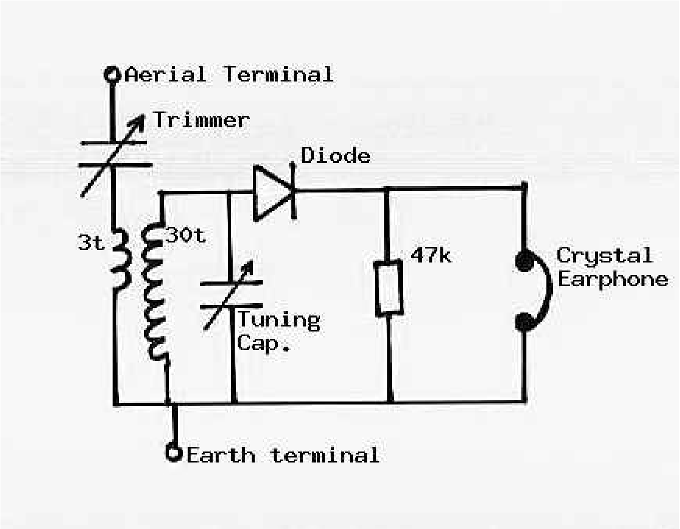

The circuit diagram of the Short Wave Crystal Set

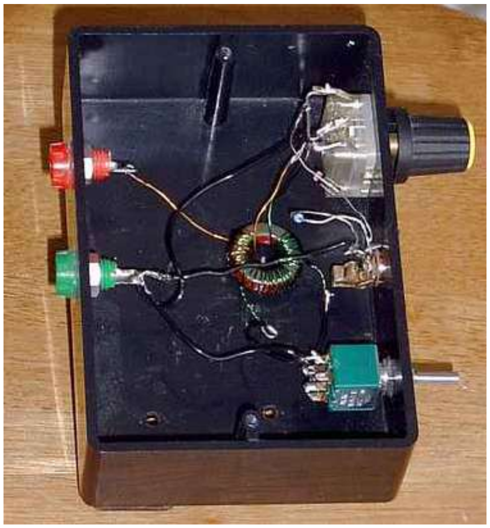

A completed SW Crystal Set using a toroid inductor. Note: the main winding has a tap to allow the switch to short part of the winding and thereby give two ranges.

AUSTRALIAN DESIGN

Moving back to the Medium Waves, here is a circuit for a very interesting Australian design that promises extremely good station separation (selectivity), and having built it I can vouch for that claim, it's really excellent.

I receive three national stations and three local stations at my location with excellent clarity using a modest antenna and standard crystal earphone.

The coil is different to the other crystal sets described above, it is much bigger at 3¼ inches (8cm) diameter and 5inches (12cm) long. I made my coil former out of the cardboard from a breakfast cereal box - just like Blue Peter!

The design is often referred to as The Mystery Crystal Set, by Proton.

The front panel of the Australian Crystal Set

Two distinct coils are wound on it, the first one consists of about 50 turns of 24 s.w.g (approx) enamelled copper wire. The second coil is 25 turns, very close wound right over the top of the first coil using 30 s.w.g. (approx) wire, try to get this second coil wound in between the windings of the first, for better inductive coupling.

Then carefully wire up the set according to the diagram. Notice that the tuned circuit is not connected to earth and has no *direct* connection to the detector circuit. The detector circuit *is* connected to earth however. The two aerial terminals offer alternative selectivity performance, terminal A gives very good selectivity while B is very wide. I never bother with B.

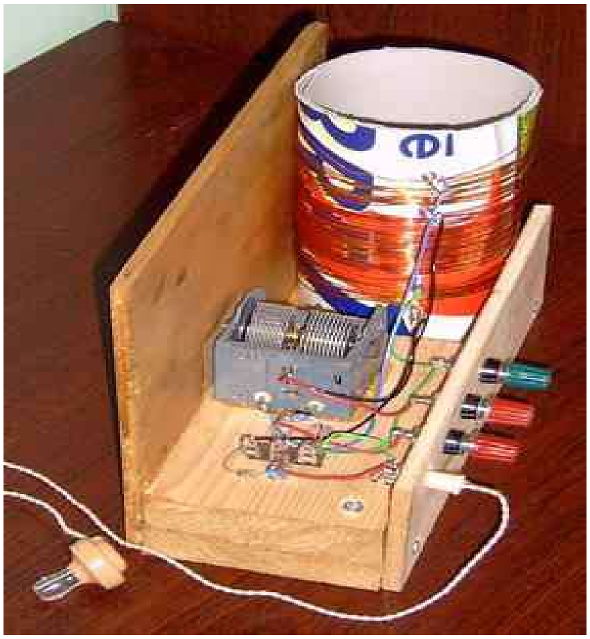

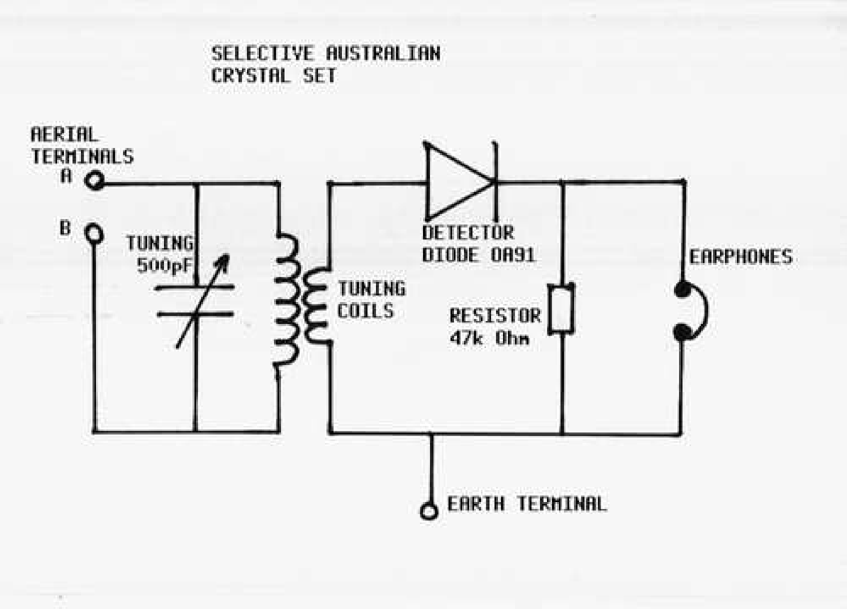

Make the coil carefully and wire up this crystal set according to the circuit diagram opposite and you will be rewarded with a really high performance crystal set of a type that was used in the very early days of broadcasting in 1930's in Australia.

This is probably my favourite crystal set!

THE DENCO PCC1 COIL

The PPC1 coil was a commercially manufactured by Denco Clacton Ltd and was popular among hobbyists not keen on going to the bother of winding their own fiddly little coils. As a child I wanted try one of these coils and sent away for one by mail order. It arrived a few days later in a little cloth bag, like a miniature pump bag, with protective wrapping inside.

The coil windings are entirely enclosed in what I can only describe as a cylindrical ferrite 'shell', the four very thin connecting wires exiting, two either side, from small apertures in the 'shell'. The performance of the circuit shown below I seem to remember was quite pleasing. Unfortunately I cannot find the set or the PPC1 coil at the moment, but here is a reproduction of the circuit diagram and data:

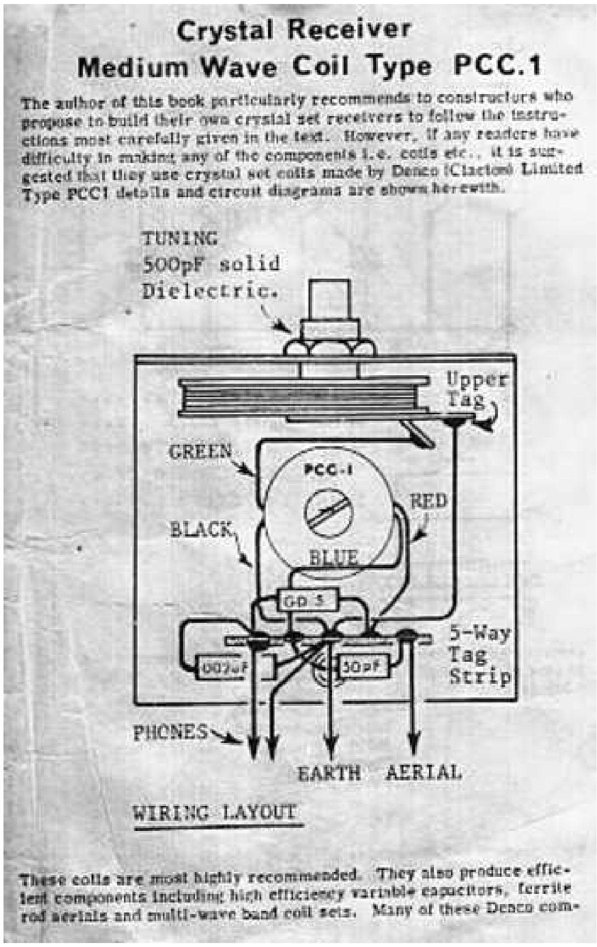

Crystal Receiver

Medium Wave Coil Type PCC.1

The author of this book particularly recommends to constructors who propose to build their own crystal set receivers to follow the instructions most carefully given in the text. However, if any readers have difficulty in making any of the components i.e. coils etc., it is suggested that they use crystal set coils made by Denco (Clacton) Limited Type PCC1 details and circuit diagrams are shown herewith.

TUNING
500pF solid
Dielectric.
Upper
Tag
GREEN
PCC-1
RED
BLACK
BLUE
GD 5
5-Way
Tag
Strip
.002uF
50pF
PHONES
EARTH
AERIAL

WIRING LAYOUT

These coils are most highly recommended. They also produce efficient components including high efficiency variable capacitors, ferrite rod aerials and multi-wave band coil sets. Many of these Denco com-

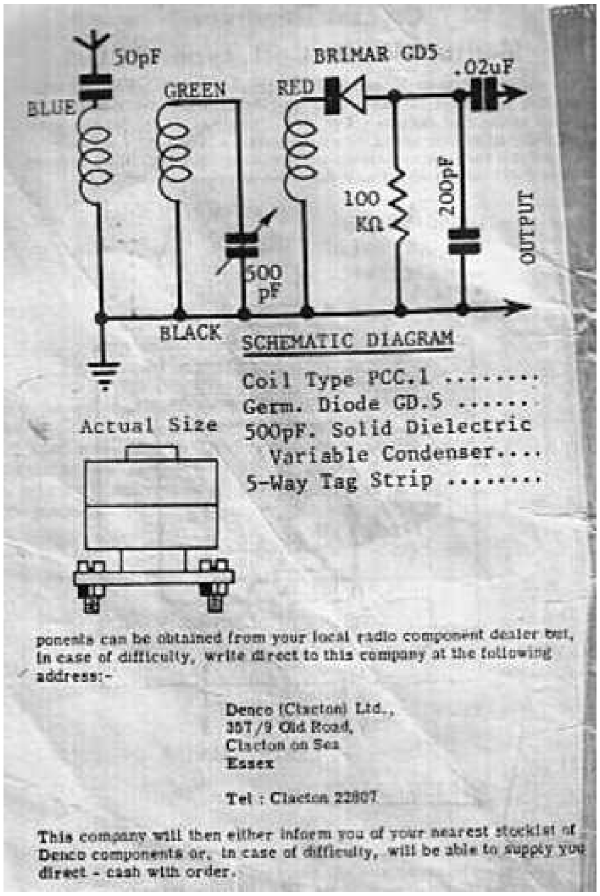

50pF
BRIMAR GD5
.02uF
BLUE
GREEN
RED
100
KΩ
200pF
OUTPUT
500
pF
BLACK
Actual Size

SCHEMATIC DIAGRAM

Coil Type PCC.1 ....
Germ. Diode GD.5 ......
500pF. Solid Dielectric
Variable Condenser....
5-Way Tag Strip ....

ponents can be obtained from your local radio component dealer but, in case of difficulty, write direct to this company at the following address:-

Denco (Clacton) Ltd.,
357/9 Old Road,
Clacton on Sea
Essex

Tel : Clacton 22807

This company will then either inform you of your nearest stockist of Denco components or, in case of difficulty, will be able to supply you direct - cash with order.

THE REPANCO DRR2 COIL

I recently rediscovered an old Repanco DRR2 Longwave / Mediumwave coil that must have been kicking around in my junk box since the 1970's.

The DRR2 coil was made by Repanco in Coventry. It came with a page of suggested circuit diagrams which I thought had been lost to the mists of time, but it recently came to light again, so I have now copied it below.

Once again I included an aerial trimmer which can be adjusted to improve selectivity.

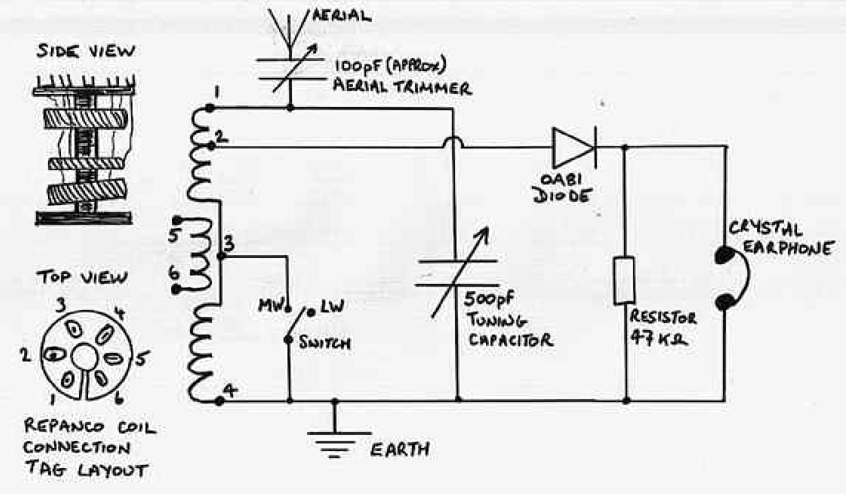
SIDE VIEW
AERIAL
100pF (APPROX)
AERIAL TRIMMER
1
2
3
4
5
6
OA81
DIODE
CRYSTAL
EARPHONE
TOP VIEW
MW
LW
SWITCH
500pF
TUNING
CAPACITOR
RESISTOR
47KΩ
REPANCO COIL
CONNECTION
TAG LAYOUT
EARTH

The circuit diagram of the crystal set using the Repanco DRR2 coil

Repanco Ltd was formed by two ex-army signals engineers and from the earliest days of radio supplied crystal set kits and coils to radio construction enthusiasts.

The Repanco DRR2 coil was for medium wave and long wave intended for use in when building simple crystal set and valve radio

A 'lash-up' of the Repanco crystal set

circuits.

It consists of three coils; a Medium Wave coil at the top that includes a tapping point (for the aerial); a coupling coil or tickler in the middle; a lower coil which can be connected in series with to top coil to provide Long Wave reception.

I have built a quick crystal set with the coil and it provides good reception with excellent selectivity, so it must have a very good Q factor.

The Repanco DDR2 coil was provided with a simple Foolscap size information sheet that showed four different radio circuits. Sadly the sheet does not give a huge amount of information and my copy is now rather tatty and faded - it is copied below:

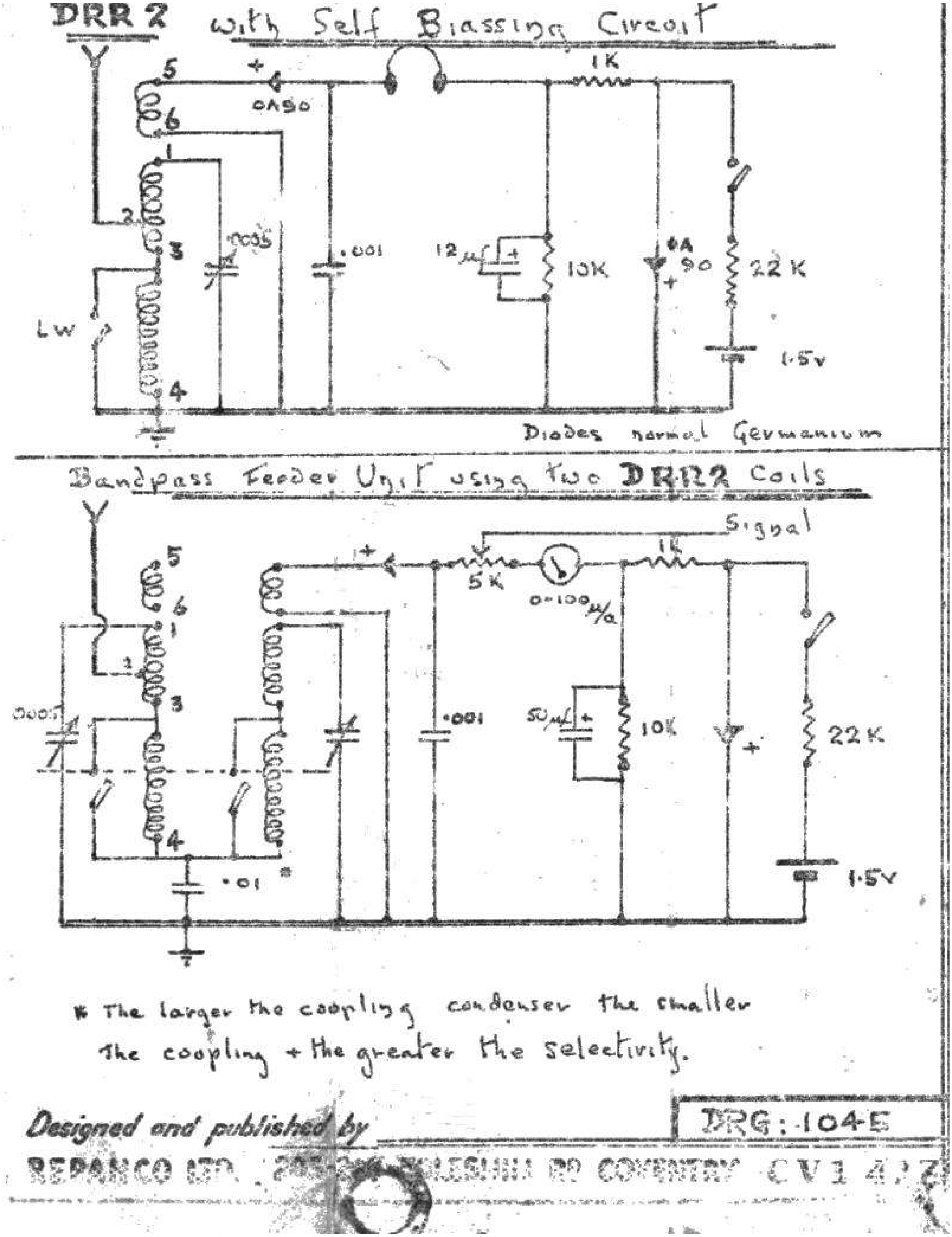
DRR 2 with Self Biassing Circuit
5
+
OA90
1K
6
1
2
·0005
3
·001
12μf
+
10K
OA 90
+
22 K
LW
1·5v
4
Diodes normal Germanium
Bandpass Feeder Unit using two DRR2 Coils
Signal
5
+
1K
5K
6
0-100μ/a
1
2
3
·0005
·001
50μf
+
10K
+
22 K
4
·01 *
1·5V

* The larger the coupling condenser the smaller the coupling + the greater the selectivity.

Designed and published by

DRG:-104E

REPANCO LTD. 205 [illegible] ESHILL RD COVENTRY CV1 4[illegible]Z

Simple crystal radio type circuits using the Repanco DRR2 coil

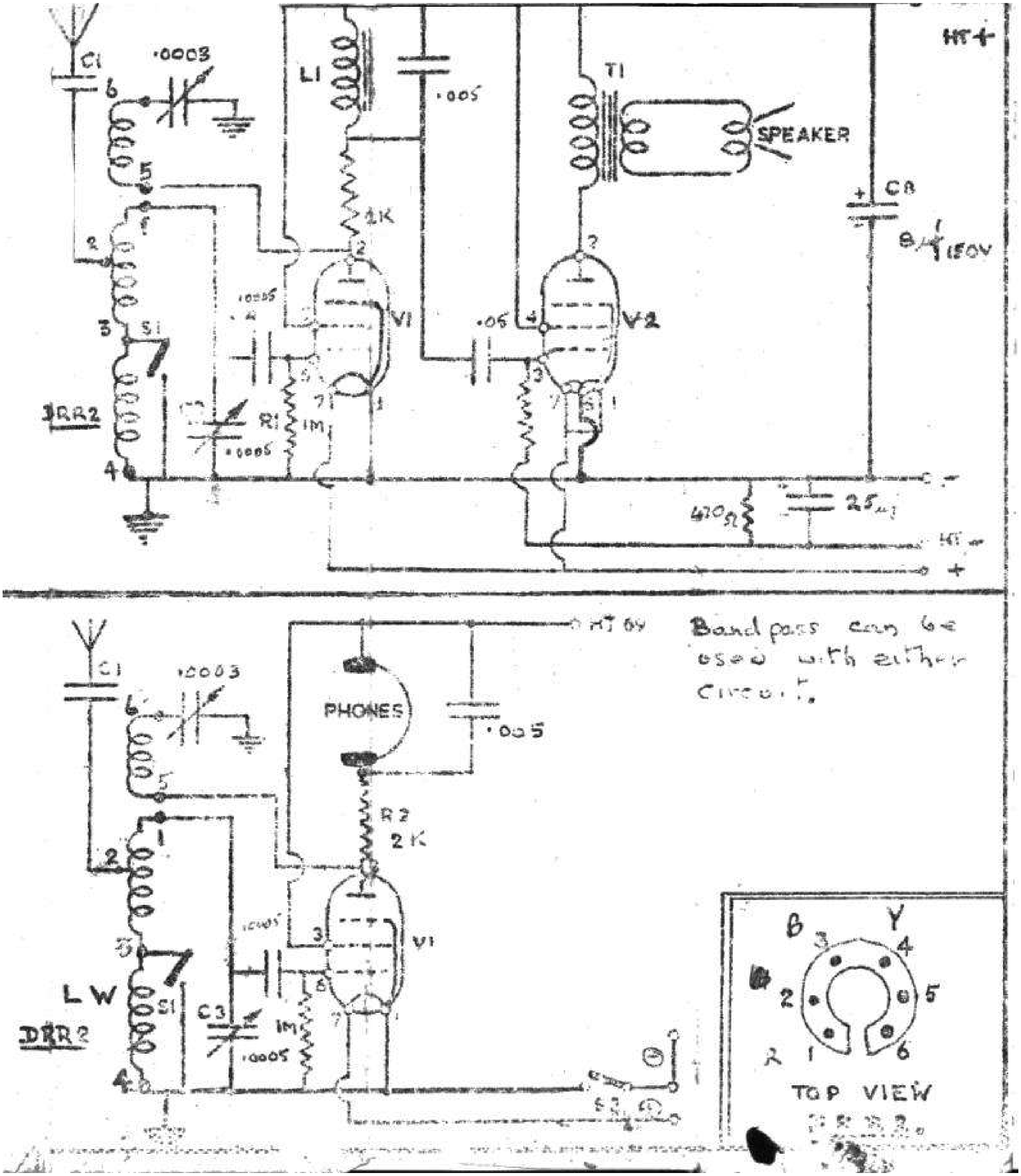
C1
·0003
6
5
1
2
3
S1
DRR2
4
L1
·005
1K
·0005
V1
R1
1M
·0005
7
1
T1
SPEAKER
HT+
C8
8µF 150V
·05
V2
420Ω
25µF
HT -
+
HT 6V
PHONES
R2
2K
LW
C3
Band pass can be used with either circuit.
B
Y
R
TOP VIEW
DRR2

Valve circuits using the Repanco DRR2 coil

Repanco Ltd no longer produces radio coils and crystal set kits for the radio construction enthusiast, as it did in the early days of radio. In 1986 it was renamed Repanco Bartlett Ltd when it merged with Bartlett Electronics. The company moved from the Foleshill Road to Unit 24, Albion Industrial Estate, Endemere Road, Coventry CV6 5NT and now specialises in transformers and wound components and can design and manufacture to commercial customer requirements, their website is: http://www.repancobartlett.co.uk/

CRYSTAL EARPHONES

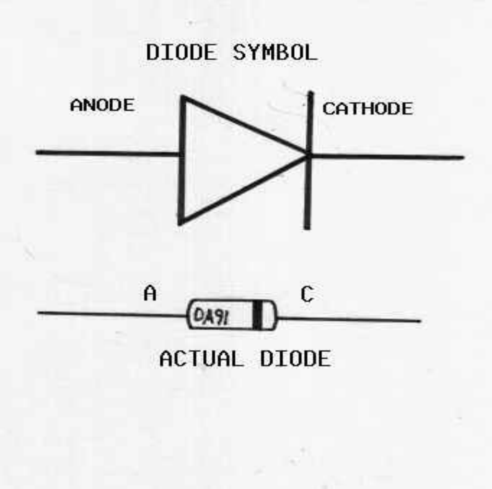
DIODE SYMBOL
ANODE
CATHODE
A
C
OA91
ACTUAL DIODE

Here is a good idea and well worth trying, to maximise the use of sound output from your crystal set why not use dual crystal earphones? Having an earphone in each ear helps to block out extraneous noises helping the listener to better concentrate on any weaker stations received.

Using the circuit below, one earphone makes use of one half cycle of the radio wave while the second earphone uses the other half cycle of the wave that would have previously gone to waste when using just one diode. Ensure that the diodes are connected up according to the diagram i.e. one diode is connected the opposite way round to the other.

Also try to make sure that the diodes and crystal earphones are similar to obtain the best results. (You could simply connect two crystal earphones to the same terminals of the single diode, but this would not be as efficient and the sounds would be much quieter.)

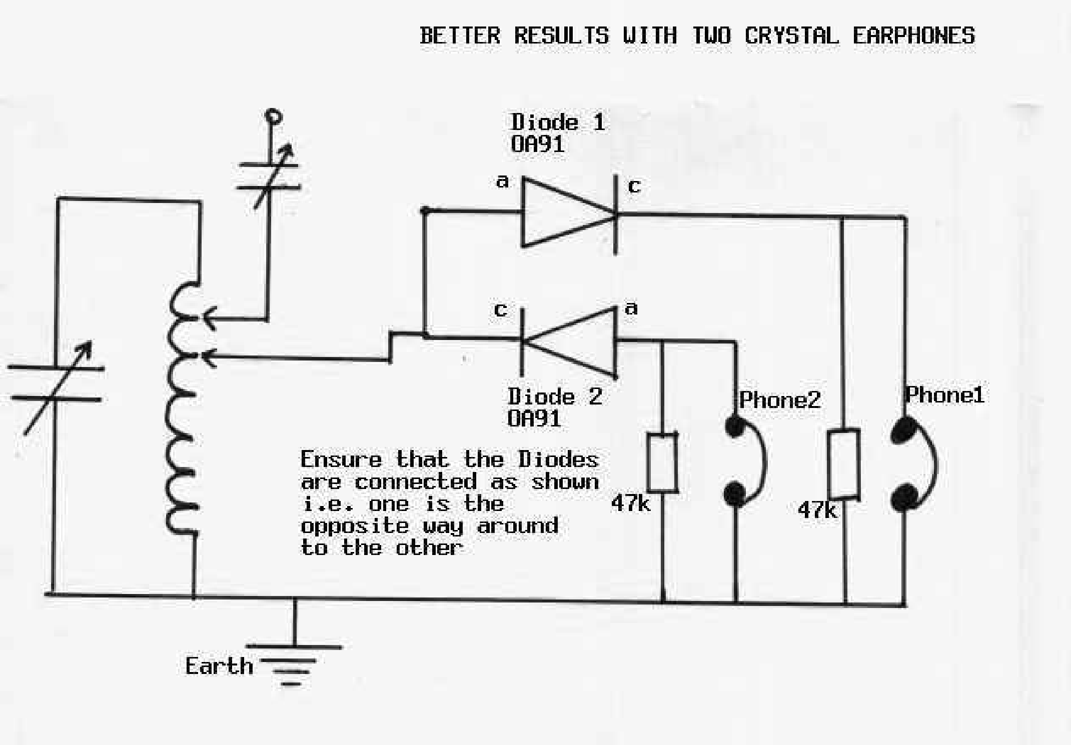
BETTER RESULTS WITH TWO CRYSTAL EARPHONES
Diode 1
OA91
a
c
c
a
Diode 2
OA91
Ensure that the Diodes are connected as shown i.e. one is the opposite way around to the other
47k
Phone2
47k
Phone1
Earth

A note about Crystal Earphones: It will be worthwhile buying several different ones from different sources as performance varies between manufacturers quite markedly. I have found the ones marked 'Japan' on the back are the most sensitive and therefore loudest, whereas ones marked 'Receiver' 'Taiwan' are often a little less sensitive and therefore quieter and sometimes more 'tinny' sounding.

As mentioned previously it has been noted that the OA47 diode will be of particular interest since it has the lowest forward bias voltage of any of the common diodes available. This will make the crystal set somewhat more sensitive and therefore louder. The US equivalent of the British OA47 is the IN34.

SPIDERS WEB' COIL

Here is an interesting concept sent in by Chris Dorna of the Vught North Scouts in the Netherlands. It is a crystal set made out of a coil wound in the form of a spiders web:

See more HERE.

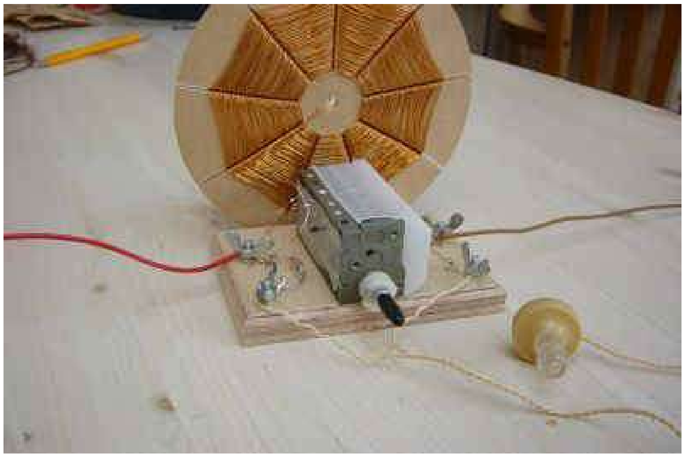
Chris Dorna's Crystal Set with Spider Web Coil

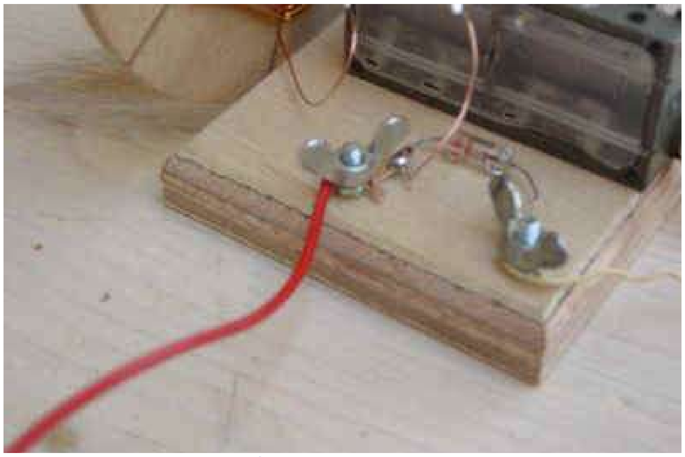
Detail of the germanium diode

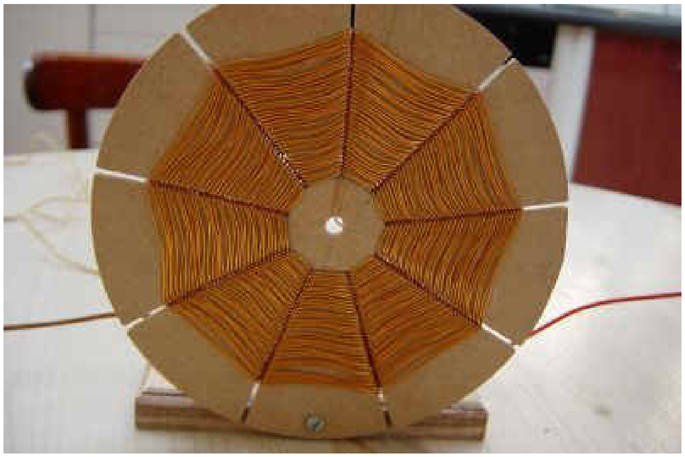
Close up of Chris Dorna's Spider Coil

LOOP THE LOOP!

A crystal set can also be made that does not need a large long wire aerial. If you have ever made a loop aerial for medium wave or long wave DX-ing, then it is a simple matter to add a diode, resistor and a socket to connect a crystal earphone that will allow reception of nearby stations.

See my section on Loop Aerials and ATU's for more constructional details.

A portable loop aerial, that incorporates a crystal set

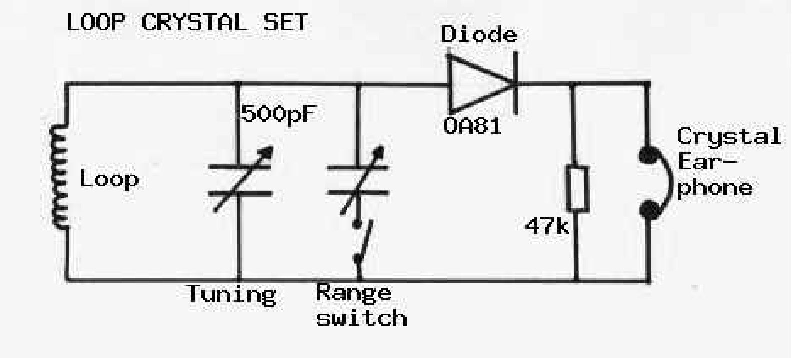

The circuit diagram of the Loop Crystal Set. The loop is 10 turns of 7/0.2mm 'hook-up' wire wound on a 40cm (17") former made of attractive plastic edging strip available from many DIY stores. The loop is very directional in its pick up pattern, which can help eliminate interference from some stations by rotating the loop. The switch and additional capacitor allow tuning of the lower medium wave band from about 650 to 520 kHz. Having a loop with 50 to 60 turns of wire will tune into the Long Wave band.

DIODES - For Crystal Set Use - some notes by Felix Scerri

Germanium diodes for crystal set use.

Although I'm a fan of these new silicon schottky BAT 46 diodes, good germanium diodes still have a lot to offer, especially in terms of 'weak signal' sensitivity. Last night I did an experiment.

I sorted through quite a few of my hundreds of acquired random germanium diodes looking for particularly 'sensitive' ones. I tested this by tuning in a weak AM station and comparing the detected DC output level and also the apparent 'loudness' of the audio signal.

Even amongst germanium diodes of the same type, there was enormous variation all the way from excellent to poor! For very weak signals, germanium diodes 'detect' in the 'square law' region below the diode conduction 'knee', in a rather different part of the curve than with much stronger signals (way beyond the diode knee).

When testing germanium diodes for weak signal sensitivity, the inherent capacitance of the diodes is also a factor, and the 'tuning' may change somewhat and will need to be readjusted with every diode tested!

In the end, out of a large number of germanium diodes tested, I found three or four germanium diodes with excellent weak signal sensitivity and the rest were poor. One other interesting thing, good germanium diodes 'sound' different, rather more 'rounded and smoother' than the schottky's which tend to sound mercilessly clean, almost clinical. I also found almost no variation in weak signal sensitivity with my BAT 46 schottky diodes. Take your pick!

Regards, Felix Scerri VK4FUQ 14/03/2012

SOLDERLESS CONSTRUCTION IDEAS

For a novice the use of a soldering iron may seem a bit daunting at first and while the most reliable results will be obtained with a good soldered joint using a tag strip as shown below, the circuits can still be made without the use of a soldering iron.

5 WAY TAG STRIP

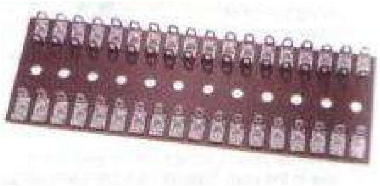

36 WAY TAG STRIP - TWO ROWS

The very simplest circuits could be wired together ,with a little ingenuity, with the component wires being held together in the grip of solderless crocodile clips, whereby the connecting hook-up wire is fixed to the croc' clip by a screw rather than solder.

For more the slightly more complex circuits a plastic Terminal Block (sometimes referred to as a choc' or chocolate block) can be utilised very effectively indeed. These are used in mains wiring and are available in various sizes; 2 Amp, 5 Amp, 15 Amp and 30 Amp. The 5 and 15 Amp Terminal Blocks I have found to be the most suitable. The various component wires can be trapped securely with the screw at each junction point. This method also makes it easy to change the components around when experiment with different circuits. See The EXPERMENTAL CRYSTAL SET for more details in Part 5**.**

'CHOCOLATE' TERMINAL BLOCK

The Ladybird book called 'Making A Transistor Radio' (also shown on the TRF Radio pages) detailed a very novel approach using brass screws with screw-cups to trap the component wires at each junction point:

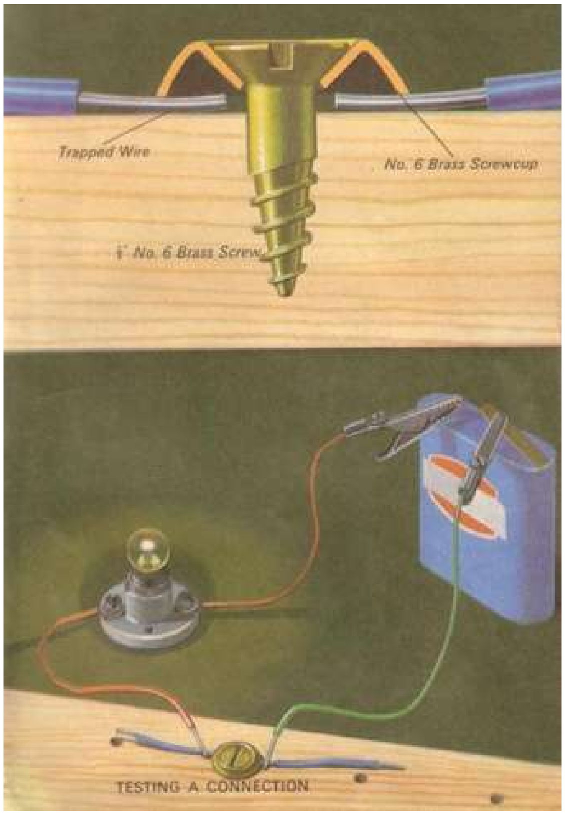
Trapped Wire
No. 6 Brass Screwcup
No. 6 Brass Screw
TESTING A CONNECTION

THE BRASS SCREW AND SCREW-CUP METHOD OF CONSTRUCTION

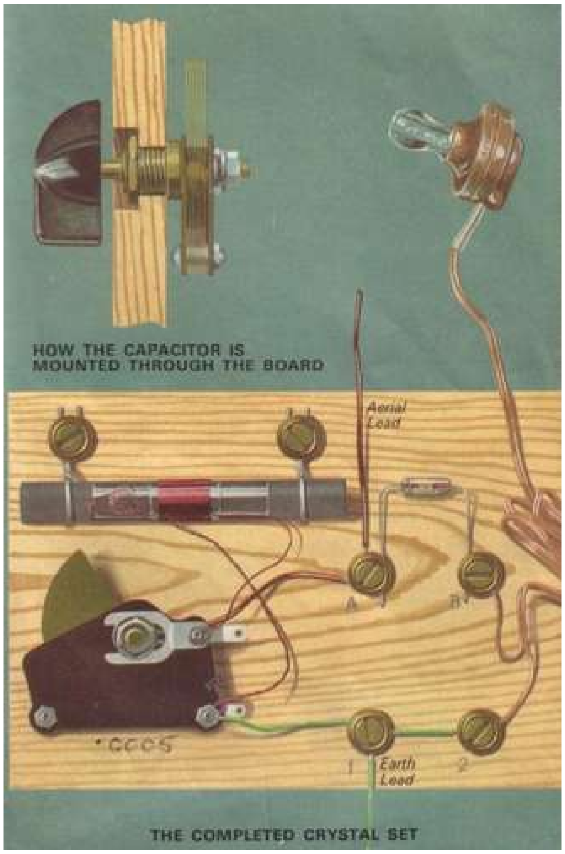
HOW THE CAPACITOR IS MOUNTED THROUGH THE BOARD
Aerial Lead
A
B
1
Earth Lead
2
THE COMPLETED CRYSTAL SET

A VERY SIMPLE CRYSTAL SET USING THE BRASS SCREW AND CUP METHOD

Crystal Sets Part Part 3 >

No AM radio stations or transmitters in your locality or country?

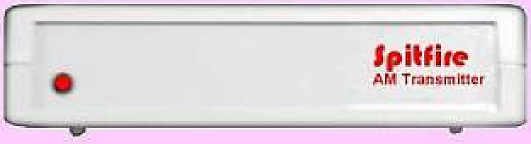

Has your local medium wave broadcast station closed or been moved to VHF/FM or Digital? Don't worry. You can still build and experiment with crystal sets and TRF radios by also buying or even building a simple low power AM transmitter. So, not only can you use your crystal sets but you can also run your own radio station that can be heard in and around your home - playing the music or programmes that you want to hear!

SSTRAN AMT3000: Superb high fidelity medium wave AM transmitter kits from SSTRAN. Versions available for 10kHz spacing in the Americas (AMT3000 or AMT3000-SM) and 9kHz spacing in Europe and other areas (AMT3000-9 and AMT3000-9SM). Superb audio quality and a great and well designed little kit to build: http://www.sstran.com/pages/products.html

POWER
GAIN
MODULATION
COMPRESSION
SSTRAN
AMT3000

http://www.sstran.com/

Other AM transmitters available:

Spitfire & Metzo: Complete, high quality ready built medium wave AM Transmitters from Vintage Components: http://www.vcomp.co.uk/index.htm Vintage Components offer a choice of the high quality Spitfire and Metzo transmitters:

SPITFIRE AM Medium Wave Transmitter with 100 milliwatt RF output power:

http://www.vcomp.co.uk/spitfire/spitfire.htm

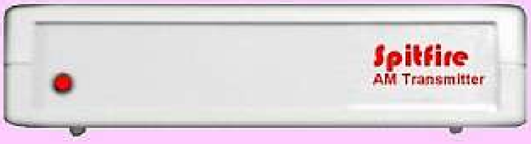

METZO AM Medium Wave Transmitter with built in compressor:

Gain
Comp
Mod
Metzo

http://www.vcomp.co.uk/metzo/metzo.htm

AM88 LP: A basic AM transmitter kit from North County Radio.
http://www.northcountryradio.com/Kitpages/am88.htm

LINKS:

BOWOOD ELECTRONICS - A friendly, helpful and very speedy source for many of your electronic components at prices that won't frighten your wallet!

THE FOXHOLE and P.O.W RADIOS - Simple crystal set receivers used by soldiers during the war and by prisoners of war (P.O.W.'s).

VINTAGE COMPONENTS - A great resource for crystal sets, components, valve radio kits and medium wave AM transmitters!

6V6 - Electronic Nostalgia and Vintage Components

Crystal Sets Part 3 >

Crystal Sets (Part1) | **Spider's Web Crystal Set (Part 3**

Crystal Set by Kenneth Rankin (Part 4) | **Experimental Crystal Sets (Part 5)** | **Crystal Radio Links**

^Top Of Page

Amateur Radio

MØMTJ Visit my amateur radio pages MØMTJ

Visit my Amateur Radio Pages >

Home | Contact | Site Map | Radio Stations & Memorabilia | Amateur Radio

Crystal Sets Introduction | Resistor & Capacitor Conversion Tables

www.mds975.co.uk © 2004 - 2014

Mike's Electronic Parts, LLC
Simple Crystal Radio Kit 2

https://www.mikeselectronicparts.com/

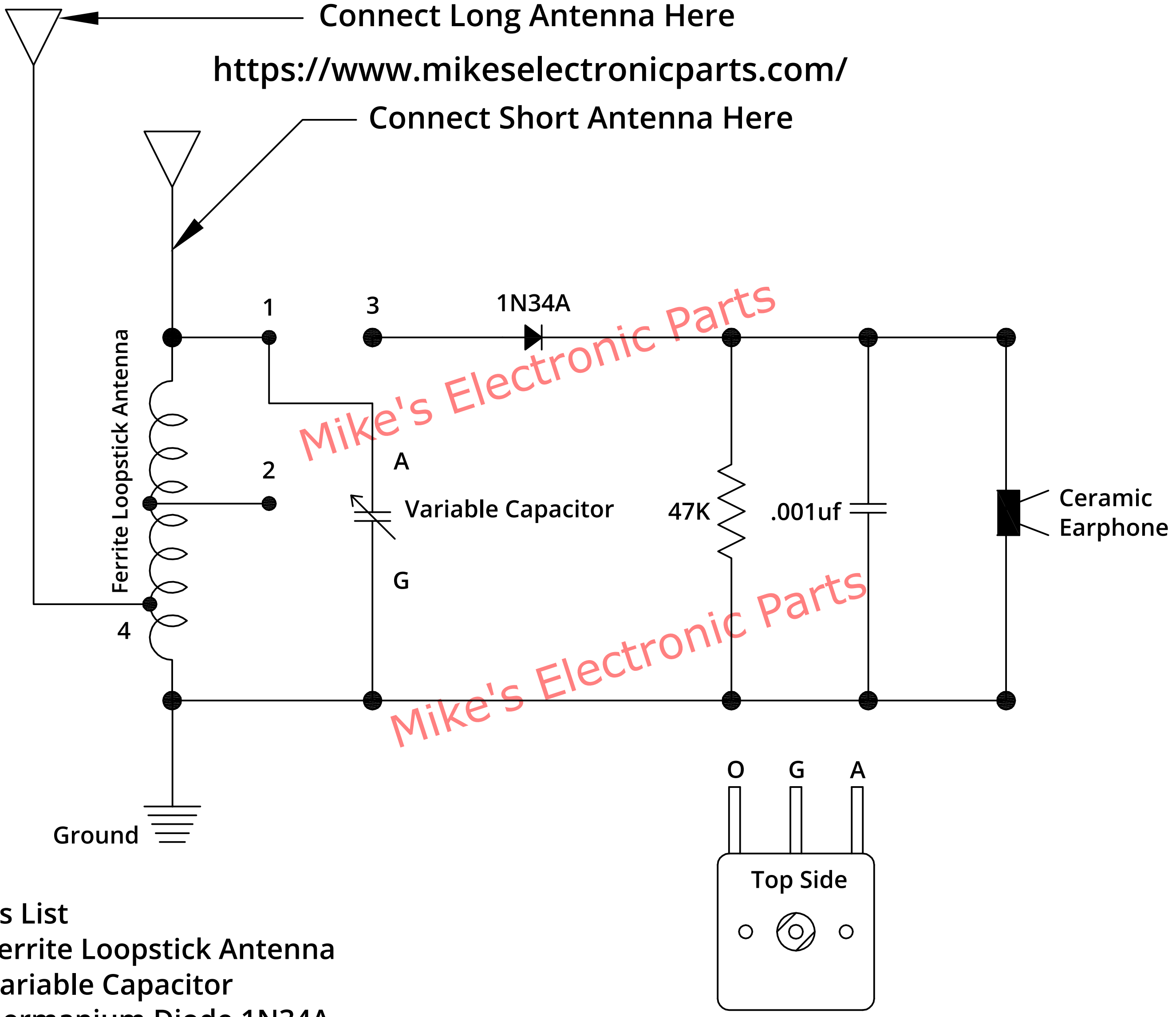
Connect Long Antenna Here
Connect Short Antenna Here
Ferrite Loopstick Antenna
1
2
3
4
1N34A
A
Variable Capacitor
G
47K
.001uf
Ceramic
Earphone
Ground
O
G
A
Top Side
Variable Capacitor
Mike's Electronic Parts
Mike's Electronic Parts

Parts List
1 - Ferrite Loopstick Antenna
1 - Variable Capacitor
1 - Germanium Diode 1N34A
1 - .001uf Capacitor Marked 102
1 - 47K Resistor
1 - 20 Million Ohm Ceramic Earphone
Antenna and Ground wire not include with parts

For short antenna connect a 15ft to 30ft length of wire to "1". For long antenna 40ft to 100ft connect to "4". It is best to use a long wire antenna. If you are not using the taps do not cut the wires. Doing so will cause the coil not to work. Various connections are, Connect 1 to 3 or Connect 2 to 3. Most volume can be had at 1 to 3 less volume but better selectivity connect 2 to 3 and the antenna connected at 1. It is best to solder the coil wire, not doing so may cause a bad connection and the radio will not work.
Drawn By: Scott Lowe
This Kit can be bought at my web site Part Number CRK#2 Kit.

https://www.mikeselectronicparts.com/

THE SMITHSONIAN INSTITUTION

The Smithsonian Institution is home to more than 141 million objects, ranging in size from insects and diamonds to locomotives and spacecraft. It is the world's largest museum complex, comprising 15 museums and galleries and the National Zoo in Washington DC, and two additional museums in New York City. Millions of visitors each year visit the nation's capital to view such treasures as the Hope Diamond, the Star Spangled Banner, and the Wright Flyer. A broad range of exhibits ensures a fun and educational experience for young and old alike.

One of the world's leading scientific research centers, the Institution has facilities in eight states and the Republic of Panama. Research projects in the arts, history and science are carried out by the Smithsonian all over the world. Some of the Smithsonian's research centers include the Smithsonian Astrophysical Observatory in Cambridge, Massachusetts, the Smithsonian Marine Station at Link Port, in Florida, and the Smithsonian Tropical Research Institute, in Panama.

For membership information of pre-visit planning material, write or call the Visitor Information and Associates Reception Center, Smithsonian Institution, Washington, D. C., 20560, (202) 357-2700 (voice), (202) 357-1729 (TTY). You may also visit the Smithsonian through our web site, www.si.edu.

HISTORY

James Smithson (1765 –1829), a British scientist, drew up his will in 1826 naming his nephew, Henry James Hungerford, as beneficiary. Smithson stipulated that, should the nephew die without heirs (as he did in 1835), the estate would go to the United States to found "at Washington, under the name of the Smithsonian Institution, an establishment for the increase and diffusion of knowledge..."

On July 1, 1836, Congress accepted the legacy bequeathed to the nation by James Smithson, and pledged the faith of the United States to the charitable trust. In 1838, following approval of the bequest by the British courts, the United States received Smithson's estate—bags of gold sovereigns—then the equivalent of $515,169. Eight years later, on August 10, 1846, an Act of Congress signed by President James K. Polk, established the Smithsonian Institution in its present form and provided for the administration of the trust, independent of the government itself, by a Board of Regents and Secretary of the Smithsonian.

SMITHSONIAN MUSEUMS, GALLERIES AND ZOOS

Smithsonian Castle
Anacostia Museum
Arthur M. Sackler Gallery
Arts and Industries Building
Cooper-Hewitt, National Design Museum
Freer Gallery of Art
Hirshhorn Museum and Sculpture Garden
National Air and Space Museum
National Museum of African Art
[illegible] Museum of American History [illegible]
National Museum of the American Indian
National Museum of Natural History
National Portrait Gallery
National Postal Museum
National Zoological Park
Renwick Gallery
S. Dillon Ripley Center
Smithsonian American Art Museum

SMITHSONIAN®

CRYSTAL RADIO™

WARNING!
ONLY FOR USE BY CHILDREN OVER 8 YEARS OLD.

READ THE INSTRUCTIONS BEFORE USE,
FOLLOW THEM AND
KEEP THEM FOR REFERENCE.
STORE THE SET OUT OF REACH
OF SMALL CHILDREN.

PLEASE KEEP A NOTE OF OUR NAME AND ADDRESS DETAILS FOR FUTURE REFERENCE. IN EUROPE CONTACT:

CE

NSI SIMM GAmbH
D 76162 KARLSRUHE
GERMANY
49-0721-9584-116

PLEASE BE SURE TO READ THE ADVISE FOR SUPERVISING ADULTS AND THE SAFETY RULES CONTAINED IN THIS BOOKLET.

©2002 SMITHSONIAN® INSTITUTION
NATURAL SCIENCE INDUSTRIES, LTD.
910 ORLANDO AVE.
WEST HEMPSTEAD, N.Y. 11552

PRINTED IN USA

ITEM NO. 3261-08

ADVICE FOR SUPERVISING ADULTS:

- READ AND FOLLOW THESE SAFETY INSTRUCTIONS, THE SAFETY RULES AND KEEP THEM FOR REFERENCE.

- SUPERVISING ADULTS SHOULD DISCUSS ANY WARNINGS AND SAFETY INFORMATION WITH THE CHILD BEFORE COMMENCING THE ACTIVITIES.

SAFETY RULES:

- DO READ THE INSTRUCTIONS BEFORE USE, FOLLOW THEM AND KEEP THEM FOR REFERENCE.
- DO KEEP YOUNG CHILDREN AND ANIMALS AWAY FROM THE ACTIVITY AREA.
- DO STORE THE SET OUT OF REACH OF YOUNG CHILDREN.

Radio Technology

Facts about early crystal development

Several inventors receive credit for developing a "crystal detector", a device which can pass current better in one direction than the other in an electrical circuit. In Germany, in the 1870's, Karl F. Braun noticed this property in certain mineral substances. Commercial development required a technology that would be suitable. This came with the advent of wireless communication and radio after 1900.

In Japan, Wichi Torikata investigated many minerals, including zincite, pyrolusite, iron pyrites and galena. In the United States, Greenleaf W. Pickard, associated with the Wireless Specialty Apparatus Co. of Boston, conducted extensive experiments as well. He became well known for the PERIKON Detector which employed zincite and chalcopyrite. The early 20th century work was generally done between 1900 and 1912. Included among the scientists is General H.H.C. Dunwoody of the U.S.Army who developed a carborundum detector.

Until radio broadcasting began in 1920 (Station KDKA Pittsburgh) experimenters utilized the crystal detector for radiotelephone communications and for reception of certain Morse Code signals. There was also extensive use by the U.S.Navy and other maritime services.

The modern crystal diode, such as provided with your set, is a spin-off of radar technology developed during World War II. The basic principle, however, remains the same..i.e. to remove the AUDIO (speech, music or Morse-Code) from the radio frequency carrier wave. In one early Pickard crystal design (left) a stiff metal point is adjustable over the crystal surface. In a later version a "Catswhisker" spring impinged on a piece of galena (right) and was varied to give a louder signal.

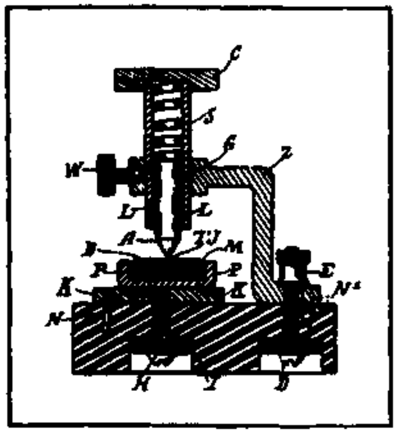

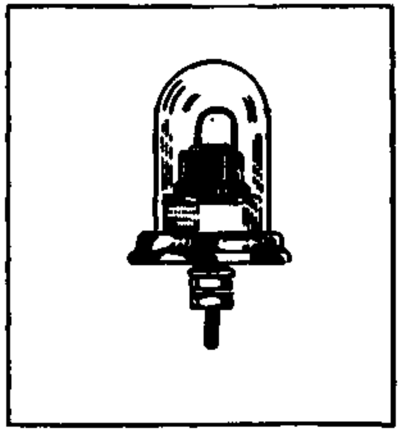

Your Smithsonian Smart Lab Crystal Radio Set contains the following items

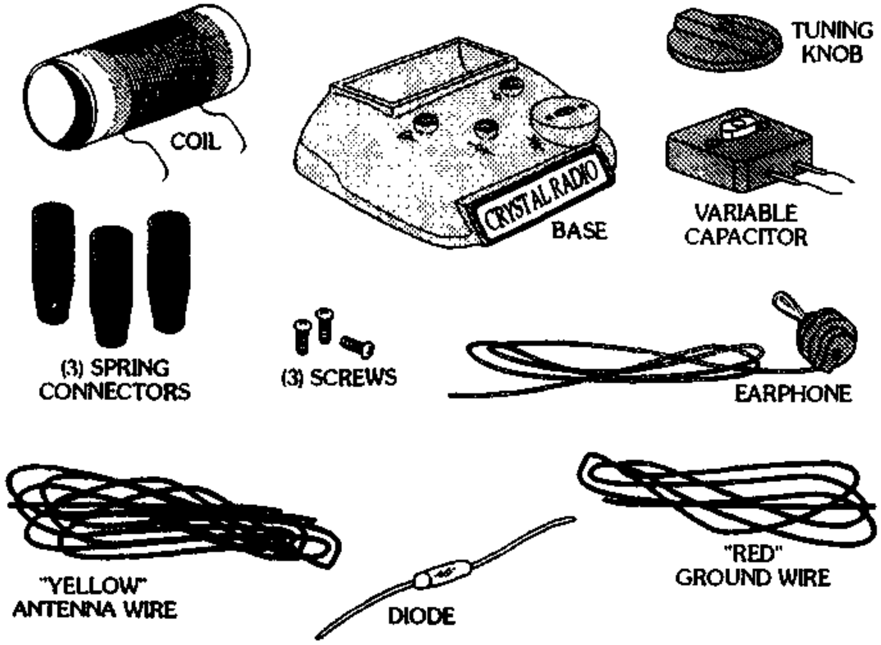
COIL
BASE
CRYSTAL RADIO
TUNING KNOB
VARIABLE CAPACITOR
(3) SPRING CONNECTORS
(3) SCREWS
EARPHONE
"YELLOW" ANTENNA WIRE
DIODE
"RED" GROUND WIRE

Shown below is a typical wiring diagram of your crystal radio.

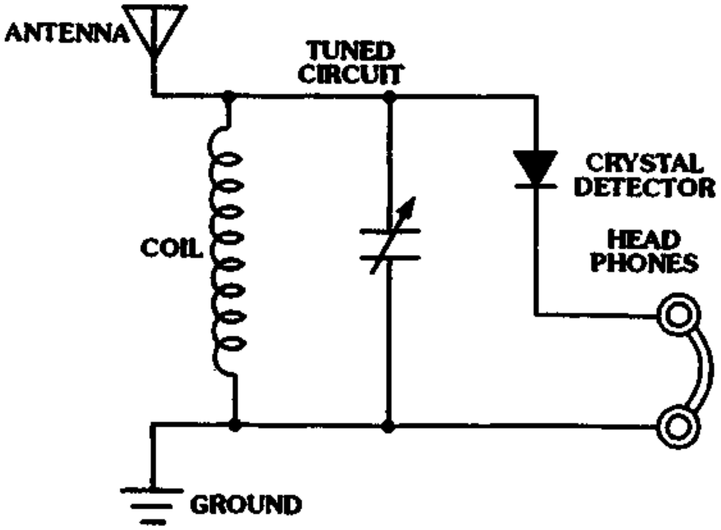
ANTENNA
TUNED CIRCUIT
CRYSTAL DETECTOR
HEAD PHONES
COIL
GROUND

Radio Technology (cont'd)

When these two different energy waves are put together by the mixer stage of the transmitter, the resulting wave, which is the signal that the station broadcasts, looks like this:

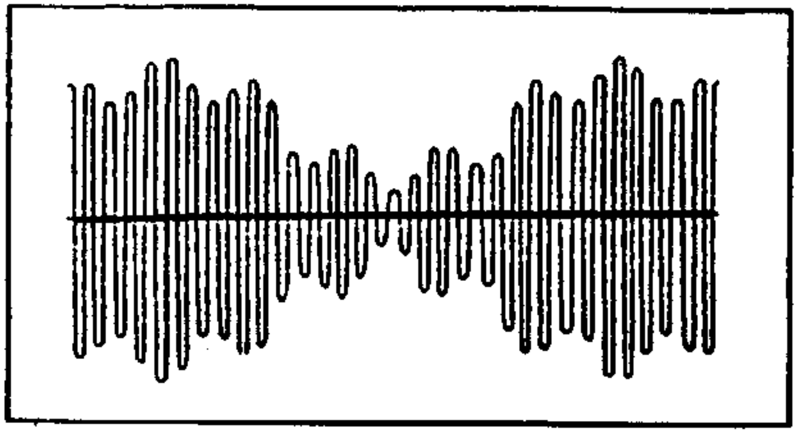

AM Broadcast Wave

The signal is called a MODULATED wave. You will notice that each of the waves is the same length, but the heights vary. Since the height of a wave is called its AMPLITUDE, The type of transmission that WXYZ uses is called AMPLITUDE MODULATION. That's why WXYZ is called an AM station.

The crystal radio that you built works in just the opposite way as the radio station. The modulated signal broadcast by WXYZ is "picked up" by the antenna on your radio.

The antenna is connected to the tuner circuit of the radio, which is made up of a coil of wire and a variable capacitor connected together. As you move the tuning knob, you are able to select, or isolate, the particular

This signal is then DEMODULATED by the diode, separating the voice (audio) wave from the carrier wave.

The voice (audio) wave then moves to the earphone, where it is changed back into sound waves that you can hear.

Learn the basic electronic symbols

Before starting to assemble your Smart Lab Crystal Radio please learn the basic electronic symbols that are shown on the plastic base.

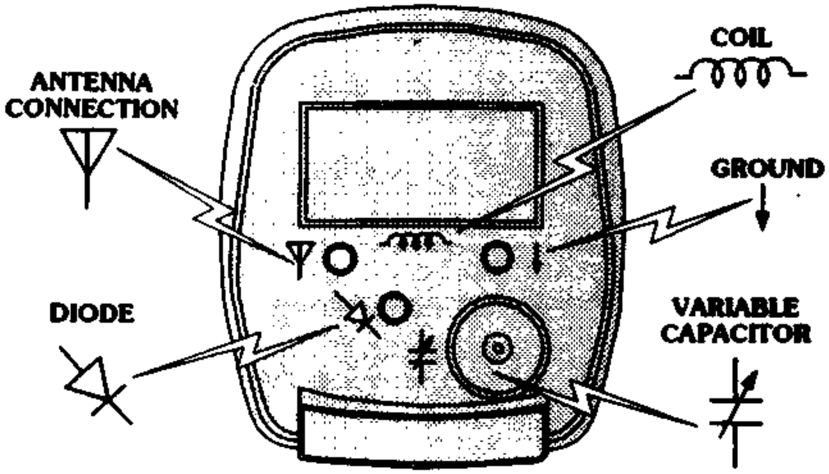

Explanation of components

Variable Capacitor - It is used to tune the radio to a station. The leads that are soldered to the lugs are used to connect it to the circuit.

Diode - A small crystal is sealed inside with leads connected to it. Pay careful attention to the position of the black bands painted on the diode close to one end.

Coil- This is a radio-tuning COIL. It was made by winding enameled copper wire around a paper core 80 times. The leads have been stripped and tinned so it can be connected to the circuit.

Earphone - It contains a small crystal that can make enough electricity to drive a metal diaphragm to produce sound. The leads have been stripped and tinned so it can be connected to the circuit.

Antenna - This is a wire used for radiating or receiving radio waves.

Ground - This is a wire used to make a electric connection with the earth or other type of grounding source to create a common return for an electric circuit.

DEAR CUSTOMER:

If we made an error and left something out of this set, or if something is damaged, we are sorry and wish to correct our error.
Please DO NOT return the set to the store where purchased, as the store does not have replacement parts. Instead, write us·

Please include:
1. Date of Purchase
2. Where purchased
3. Price paid
4. Model Number
5. Name of item
6. Brief Description of Problem
7. Include Sales Slip

We will do our best to satisfy you. Send your letter to:

NATURAL SCIENCE INDUSTRIES, LTD.
910 ORLANDO AVENUE
WEST HEMPSTEAD, N.Y. 11552-3942
ATTENTION: QUALITY CONTROL DEPT.

How to assemble your Crystal Radio

Before starting, it's important to assemble the parts correctly. Carefully read the following instructions step by step and have fun . . . half the fun is knowing you made your very own crystal radio.

Step ①

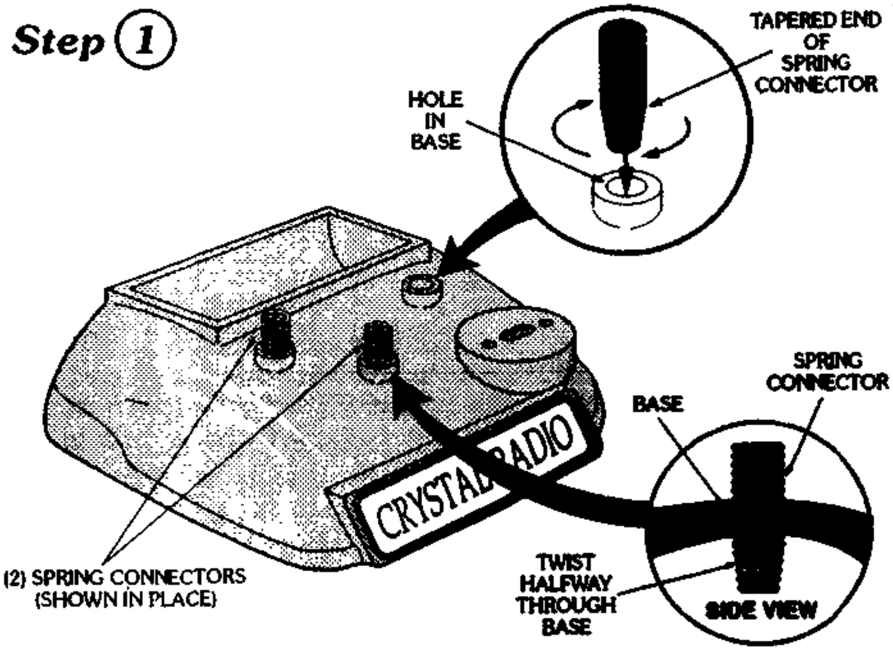
TAPERED END OF SPRING CONNECTOR
HOLE IN BASE
CRYSTAL RADIO
BASE
SPRING CONNECTOR
(2) SPRING CONNECTORS (SHOWN IN PLACE)
TWIST HALFWAY THROUGH BASE
SIDE VIEW

A. Locate the three Spring Connectors.
B. Insert one of the Spring Connectors (tapered end first) into one of the three holes on the Base.
C. Push down on the Spring Connector and twist to the right until the Connector is approximately half way through the base.

NOTE: Check to see if the Spring Connecter is inserted halfway through the Base by turning Base upside down (see insert).
D. Insert the remaining two Spring Connectors into the holes on Base the same way.

Radio Technology

Now that you have your crystal radio assembled and working, it is time to take a brief look at how it works. To do this you are going to take an imaginary trip to a radio station WXYZ, in anytown,USA.

When you arrive at WXYZ, you are met by the station's general manager, Mr. Smith, who is going to show you around.

First, Mr. Smith takes you to a STUDIO, a special room that a program comes from. There is a lot of equipment, including dials and switches, record players, lots and lots of records, and microphones.

Mr. Smith tells you that when you talk into one of the microphones, your words go into the station's electronic equipment. There, it is mixed with the station's carrier wave, and sent out into the air through the station's transmitting antenna - that tall tower you saw on top of the radio station building.

Then Mr. Smith takes a piece of paper and a pencil, and draws diagrams to show you what happens. He begins by saying that all energy travels in waves, and since the sound we make when we talk is a form of energy, it might look something like this:

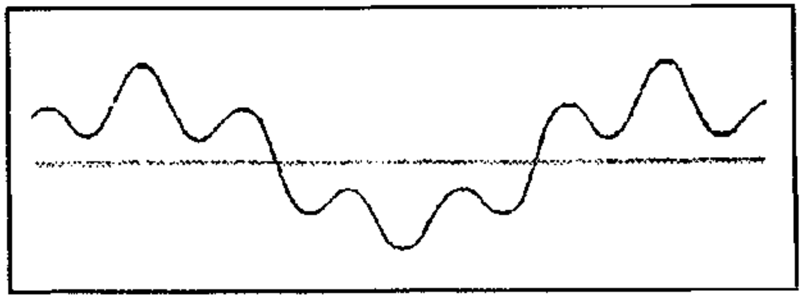

Sound Wave

The station's carrier wave is a radio wave, which is also a form of energy, which looks something like this:

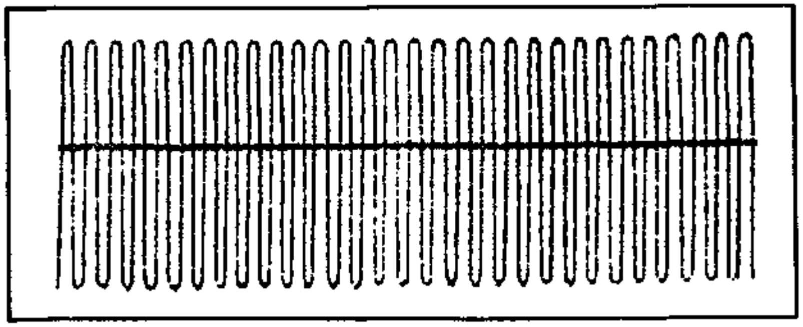

Radio Carrier Wave

How to properly operate your Smart Lab Crystal Radio and troubleshoot

After you have attached all the wires, carefully read the following instructions to make sure your crystal radio operates properly.

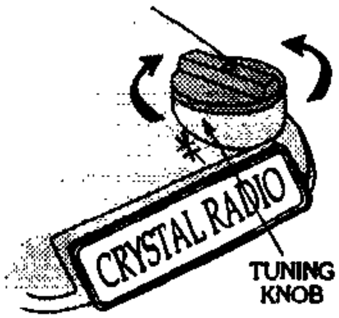

A. Place the earphone in your ear and turn the Tuning Knob clockwise and counterclockwise until you pick up the strongest signal.

Note: If you do not pick up a strong signal or do not hear anything, please read some of the troubleshooting solutions explained below.

TROUBLESHOOTING HINTS

1. There may be a bad connection caused by improper assembly of your radio. Check all the spring connectors to make sure the wires are attached correctly. If a wire is loose or the bare wire is not making good contact with the spring connector, the radio may not work.
2. If the uninsulated part of the Antenna wire touches anything other than the Antenna Connection on the radio, the radio may work improperly or not at all..
3. If the bare stripped end of the Ground wire is not wrapped tightly around a water pipe (so that it makes good contact), try taping the wire to the shiny part of the pipe using duct tape If the pipe is dull or rusty, use a piece of sandpaper to gently sand the area where the wire makes contact. Also make sure the wire doesn't touch anything other than the Ground Connection or the radio may not work properly.
4. You may live in an area where radio reception is generally poor. Instead of trying to use your radio during the day, try at night when many radio stations are received better.
5. The Antenna on your radio is a very important component. Make the Antenna as long and high above the ground as possible. If you live in a multi-story dwelling the highest floor should be used when operating the radio. Insulated bell-wire, which is available in hardware stores, makes a very good substitution for your present Antenna.

How to assemble your Crystal Radio (cont'd)

For this step you will need the variable capacitor, (3) screws and turning knob. Be carefull not to misplace the small screws when assembling.

Step ②

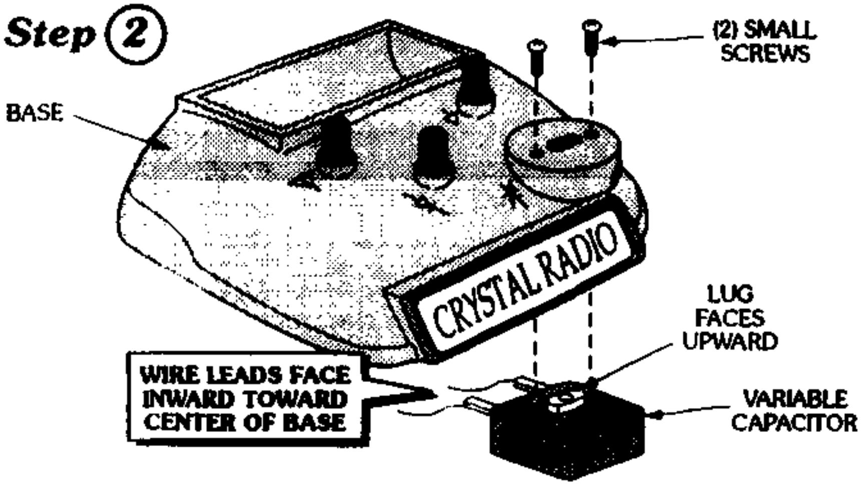

A. First place the Variable Capacitor up into the base so the holes on the capacitor line up with the holes on the base.
Note: When you do this be sure the wire leads face inward toward the base as shown and lug faces upward.

B. Next insert the two small screws into the holes on top of the Base and tighten using a small phillips screwdriver.

Step ③

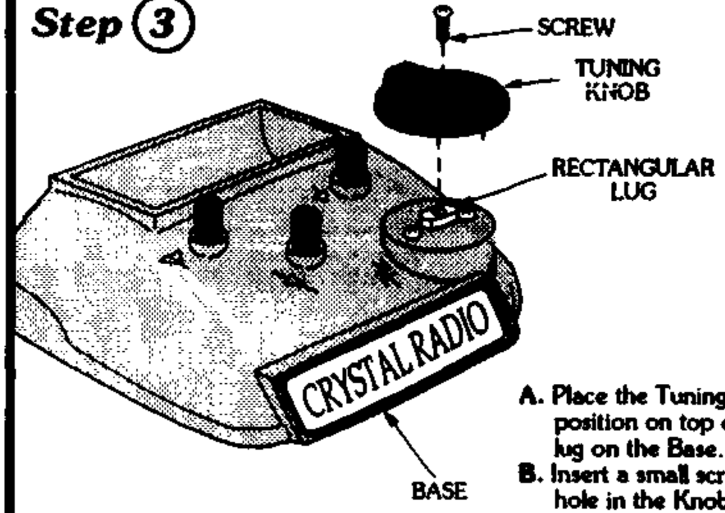

A. Place the Tuning Knob into position on top of the rectangular lug on the Base.

B. Insert a small screw into the hole in the Knob and tighten using a phillips screwdriver.

How to assemble your Crystal Radio (cont'd)

Step ④ For this step turn the Base upside down.

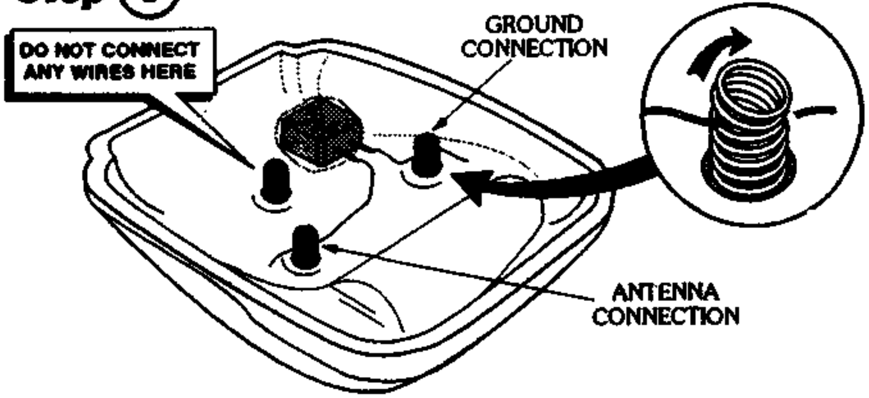
DO NOT CONNECT ANY WIRES HERE
GROUND CONNECTION
ANTENNA CONNECTION

A. Turn the Base upside down.
B. Connect one wire to the Antenna spring connector by bending the spring slightly to one side and inserting the wire through as shown in inset.
C. Connect the other wire to the Ground spring connector in the same manner.

NOTE: Do not connect any of the wires to the Diode spring connector or your crystal set will not work properly.

Step ⑤

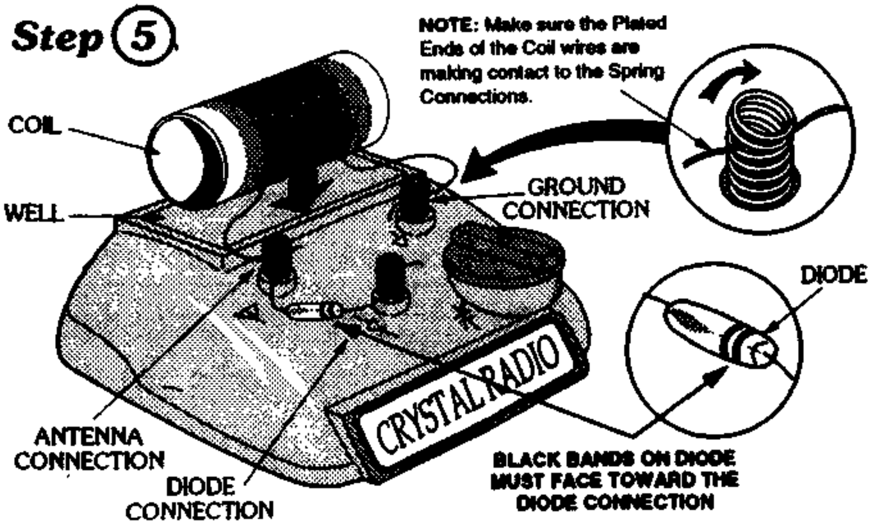
NOTE: Make sure the Plated Ends of the Coil wires are making contact to the Spring Connections.
COIL
WELL
GROUND CONNECTION
DIODE
CRYSTAL RADIO
ANTENNA CONNECTION
DIODE CONNECTION
BLACK BANDS ON DIODE MUST FACE TOWARD THE DIODE CONNECTION

A. Next place the Coil in the well of Base with the wires hanging loosely toward the connectors .
B. Connect one of the plated ends of Coil wires to the Ground Connector by bending the Spring Connector as shown in inset. Connect the other Coil wire to the Antenna Connector.
C. Place the Diode between the Diode and Antenna Connections carefully noting the black bands face toward the Diode Connection.
D. Connect ends of Diode wires to Antenna and Diode spring connectors shown.

How to attach the earphones

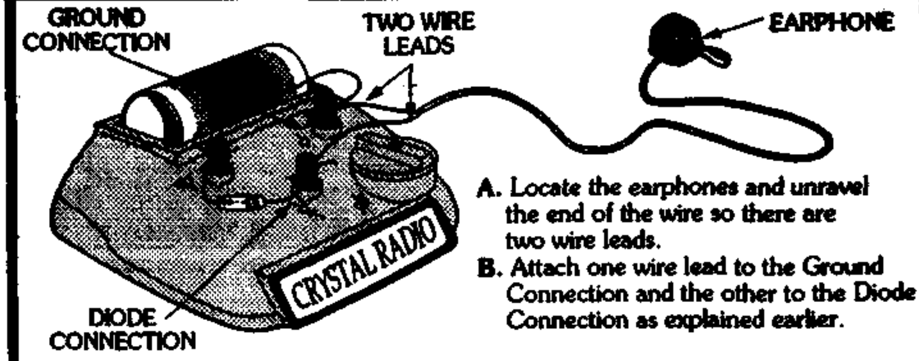
GROUND CONNECTION
TWO WIRE LEADS
EARPHONE
CRYSTAL RADIO
DIODE CONNECTION

A. Locate the earphones and unravel the end of the wire so there are two wire leads.
B. Attach one wire lead to the Ground Connection and the other to the Diode Connection as explained earlier.

Final assembly and operation

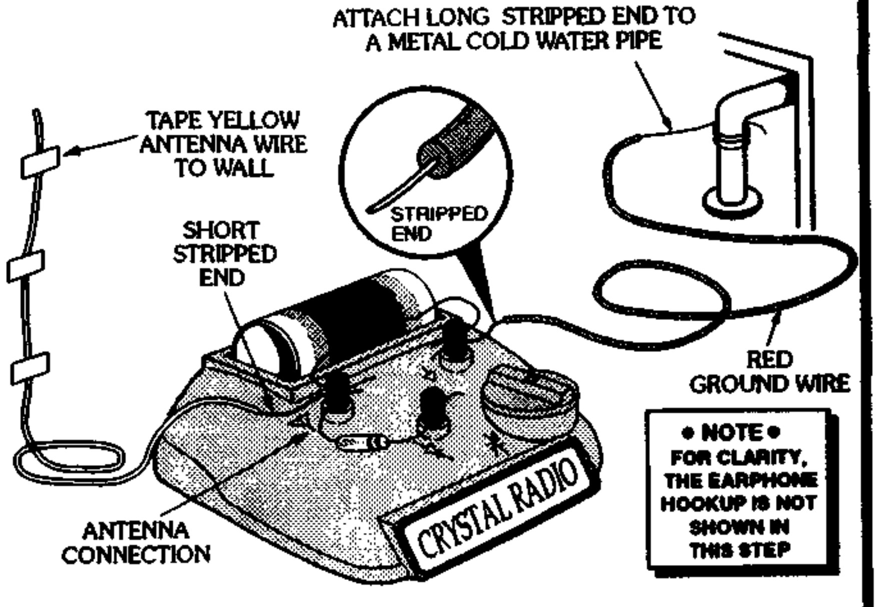
ATTACH LONG STRIPPED END TO A METAL COLD WATER PIPE
TAPE YELLOW ANTENNA WIRE TO WALL
STRIPPED END
SHORT STRIPPED END
RED GROUND WIRE
CRYSTAL RADIO
ANTENNA CONNECTION
• NOTE • FOR CLARITY, THE EARPHONE HOOKUP IS NOT SHOWN IN THIS STEP

A. Locate the Red Ground and Yellow Antenna wire.
B. Attach the stripped end of Yellow Antenna wire to the Antenna Connection on the Base.

Note: The Antenna should be strung in a straight line away from power lines and large metal objects. We suggest taping it high up on a wall using masking tape as shown.

C. Next attach the short stripped end of the Ground wire to the Ground Connection.
D. Attach the Long stripped end to a metal cold water pipe, metal radiator pipe, etc., near where the radio will be used. Have a adult help you with this since some pipes may be hot.

Note: If there is not enough stripped, bare wire to wrap around a pipe then ask an adult to strip more insulation off the end using a pair of wire strippers.

Germanium vs Silicon Diode Testing: Read this document carefully, so you will not be the victim of cheap knock-off or the wrong type diodes.

The general rule is that ***SILICON*** diodes have a voltage drop across the Anode to Cathode of 0.7 V (7/10 tenths), and the ***GERMANIUM*** diodes have a voltage drop of 0.3 V (3/10 tenths) more or less. Either diode voltage drop (silicon or germanium) will display a reading within approximately 5% of these readings.

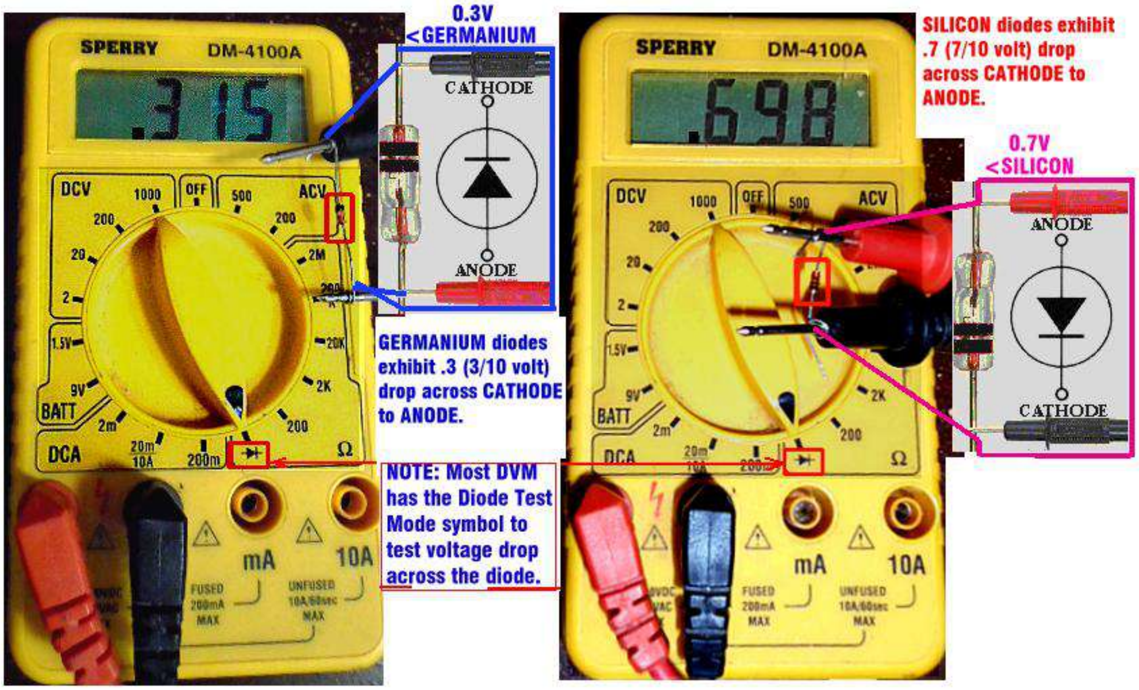

SILICON diodes exhibit .7 (7/10 volt) drop across CATHODE to ANODE.

GERMANIUM diodes exhibit .3 (3/10 volt) drop across CATHODE to ANODE.

NOTE: Most DVM has the Diode Test Mode symbol to test voltage drop across the diode.

Most digital meters (DVM) have a switch setting ▶| that is used to measure voltage drop across these diodes. This setting is usually indicated by a diode symbol to let the user know the DVM is capable of measuring forward bias voltage. This setting will tell you immediately if the diode is a **germanium**, or **silicon** diode. You need to set the selector switch on your meter to the diode test symbol ▶|.

Measuring Forward Bias Voltage

To measure the forward bias voltage characteristic you connect the black probe of your meter to the cathode terminal. The cathode terminal is on the end with a band. You then connect the red probe to the anode terminal.

Set your DVM/DMM to the *diode test mode* ▶|, it should provide you with the respective voltage drop. If the figure is 0.3 V or less, the diode is a *germanium* type. If the voltage drop is 0.7 V or less the diode is a *silicon* diode.

Sensitivity and Forward Bias Characteristics:

The sensitivity of a diode to radio waves depends upon its *forward bias voltage*. This is the voltage across the diode terminals. When it falls below this threshold value, the diode will stop conducting. Obviously, the lower this threshold value is, the greater the sensitivity of the diode to the weak radio signals.

CONCLUSIONS:

When comparing germanium diodes with silicon diodes of similar forward bias voltage, silicon diodes do not perform as well as the germanium.

Germanium has many properties that silicon diodes do not have. Germanium requires very little forward current. Forward current in a germanium diode is in the *micro*ampere region, while silicon diodes require *amperes*. This makes *germanium* a much better choice for both medium and high frequency radio signals.

Germanium also exhibits a very low, point-contact junction capacitance, while the silicon diode has much higher capacitance. A low junction capacitance allows germanium diodes to operate more effectively at high RF frequencies.

In addition, reverse leakage current for germanium diodes is in the magnitude of 1000, much more than silicon. This makes the non-linear characteristics of the germanium diode much more effective for RF detection and demodulation than silicon.

Therefore, our conclusions are; germanium diodes provide the best performance in crystal radios, RF probes, HF, and VHF signal detection.

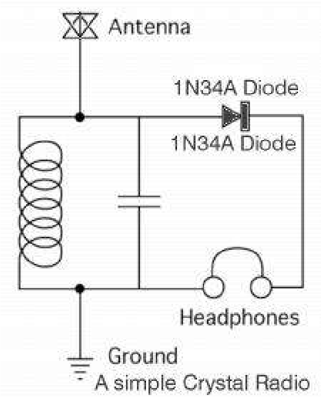
Antenna
1N34A Diode
1N34A Diode
Headphones
Ground
A simple Crystal Radio

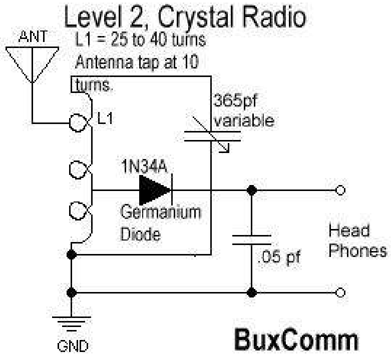
Level 2, Crystal Radio
ANT
L1 = 25 to 40 turns
Antenna tap at 10 turns.
L1
365pf variable
1N34A
Germanium Diode
.05 pf
Head Phones
GND
BuxComm

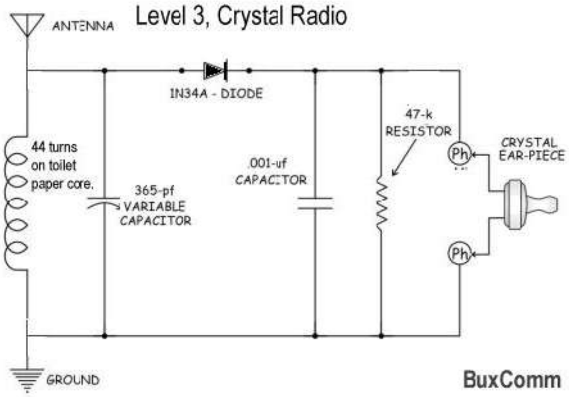
ANTENNA
Level 3, Crystal Radio
IN34A - DIODE
47-k RESISTOR
CRYSTAL EAR-PIECE
44 turns on toilet paper core.
365-pf VARIABLE CAPACITOR
.001-uf CAPACITOR
Ph
Ph
GROUND
BuxComm

BuxComm High Quality, CRYSTAL RADIO ANTENNA KIT, $14.95 cat# 50CRAK

This Crystal Radio Antenna kit consists of 50 feet of #16 AWG insulated & stranded copper antenna wire, two heavy-duty, Delrin, UV resistant antenna insulators and instruction sheet.

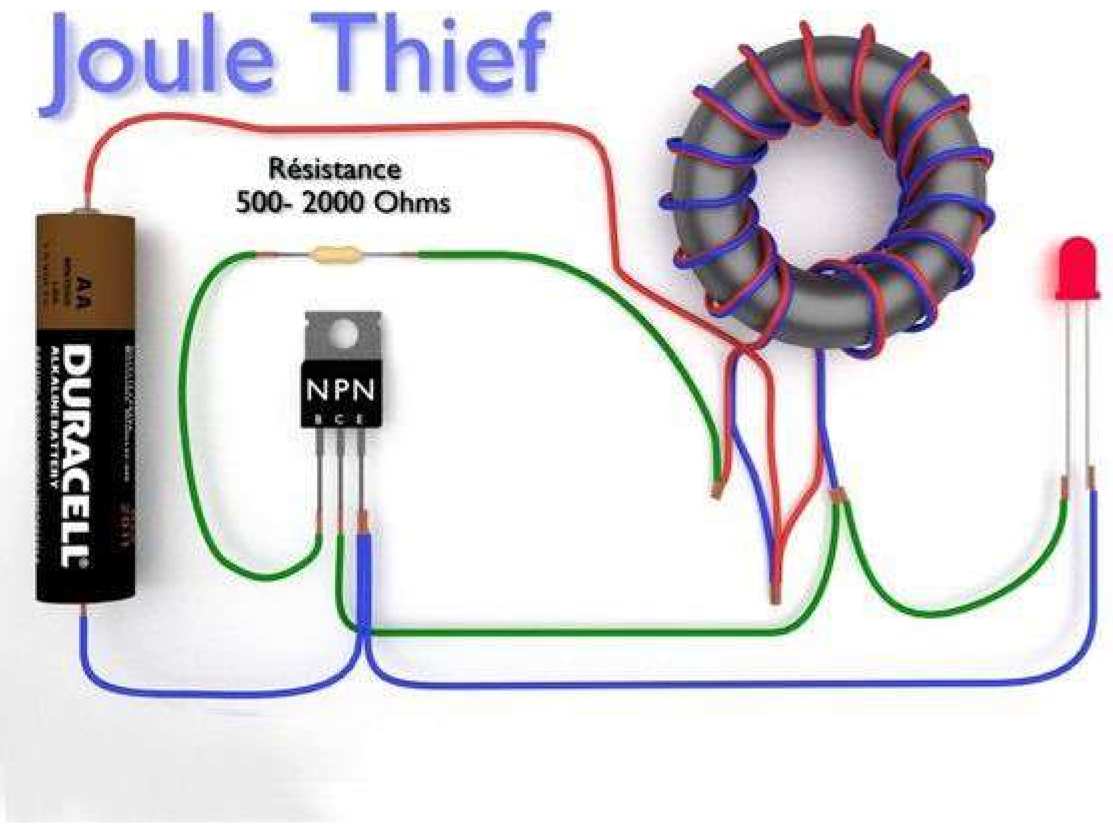
Joule Thief
Résistance
500- 2000 Ohms
NPN
B C E
AA
DURACELL
ALKALINE BATTERY

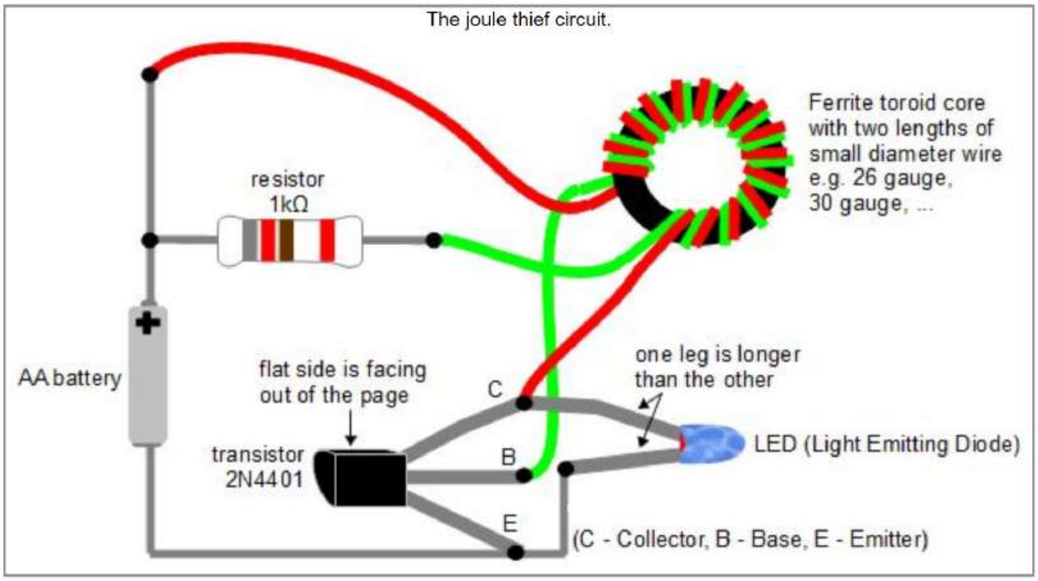
The joule thief circuit.
Ferrite toroid core with two lengths of small diameter wire e.g. 26 gauge, 30 gauge, ...
resistor 1kΩ
AA battery
flat side is facing out of the page
one leg is longer than the other
C
B
E
transistor 2N4401
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

Transistor - The legs of the transistor can be determined by noticing that there's a flat side to the transistor case. See the diagram above. A large number of transistors have been reported to work: 2N4401, NET123AP, BC547B, 2SC2500, BC337, PN2222, to name just a few.

LED - One leg of the LED is longer than the other leg. Use this to determine which one goes where. See the diagram above.

Resistor - The diagram says use a 1 kilo ohm resistor but I've used an 820 ohm one just fine. I've also seen a 2 kilo ohm one in use. Use whatever works for you. You can also use a potentiometer (a variable resistor) so that you can easily adjust it to select the resistance that gives the best light.

Toroid ferrite core - Some people have gotten these by opening up compact fluorescent lightbulbs (CFLs). I took mine out of some device whose original function I don't know. To get it working, my first one had just 13 turns for each wire and I used a 30 gauge wire and a 26 gauge wire. The wire must be insulated. A variety of number of turns will work. This is something you can play with. Look at the diagram carefully to determine where the wires connect to.

Its a Rubbish Challenge
Dog Light

Joule Thief
electronics circuit

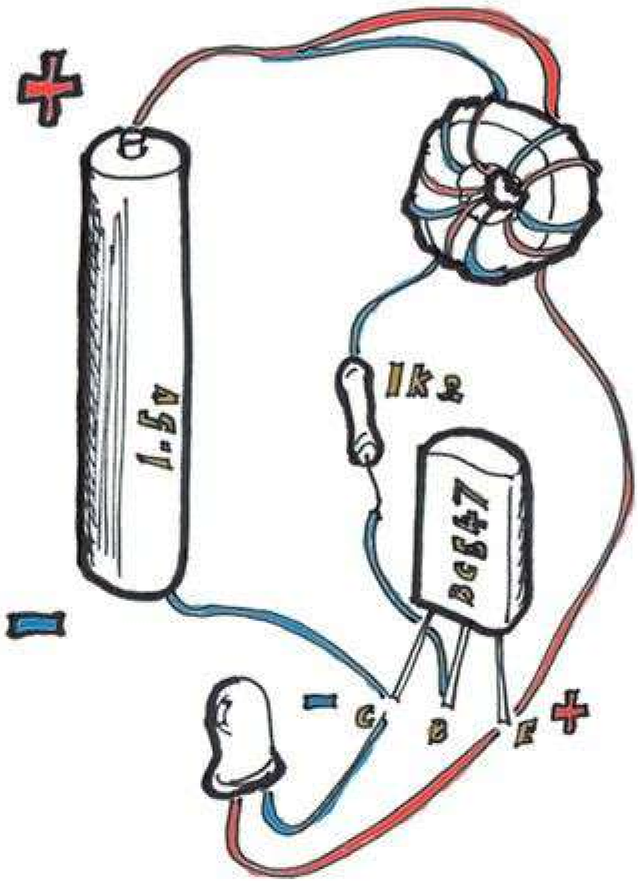
+
1.5v
1kΩ
BC547
-
- C
B
E +

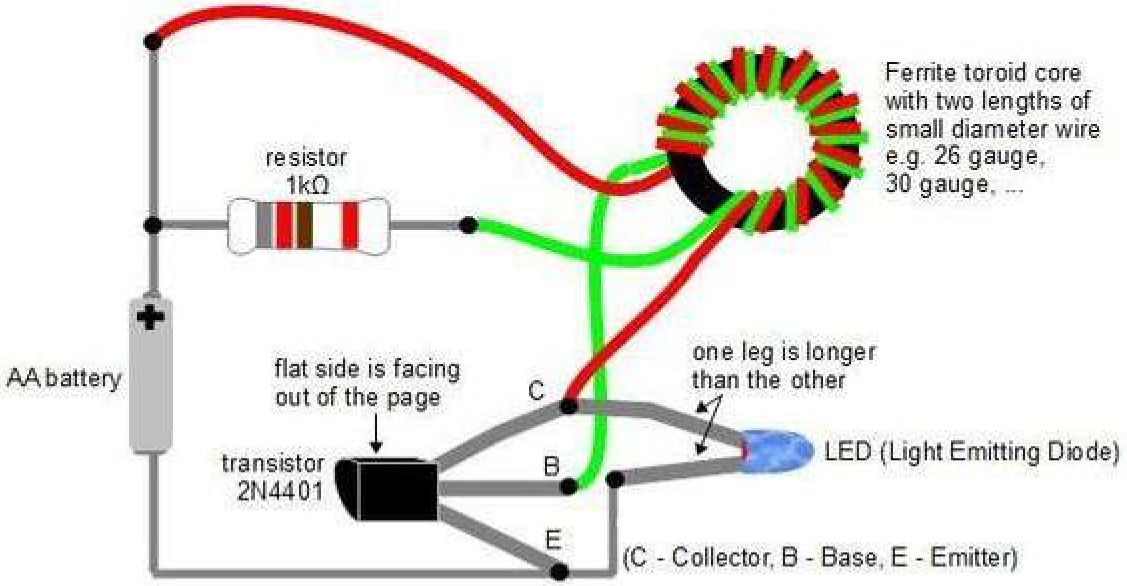
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge,
30 gauge, ...
resistor
1kΩ
AA battery
flat side is facing
out of the page
transistor
2N4401
C
B
E
one leg is longer
than the other
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

COUPONS AVAILABLE ACTIVATE COUPONS

advertisement

MAKING A SIMPLE JOULE THIEF (MADE EASY)

Technology (/technology/) > Leds | by ASCAS (/member/ASCAS/) Follow

195,691 573 165

Posted Dec. 24, 2011 | BY-NC-SA

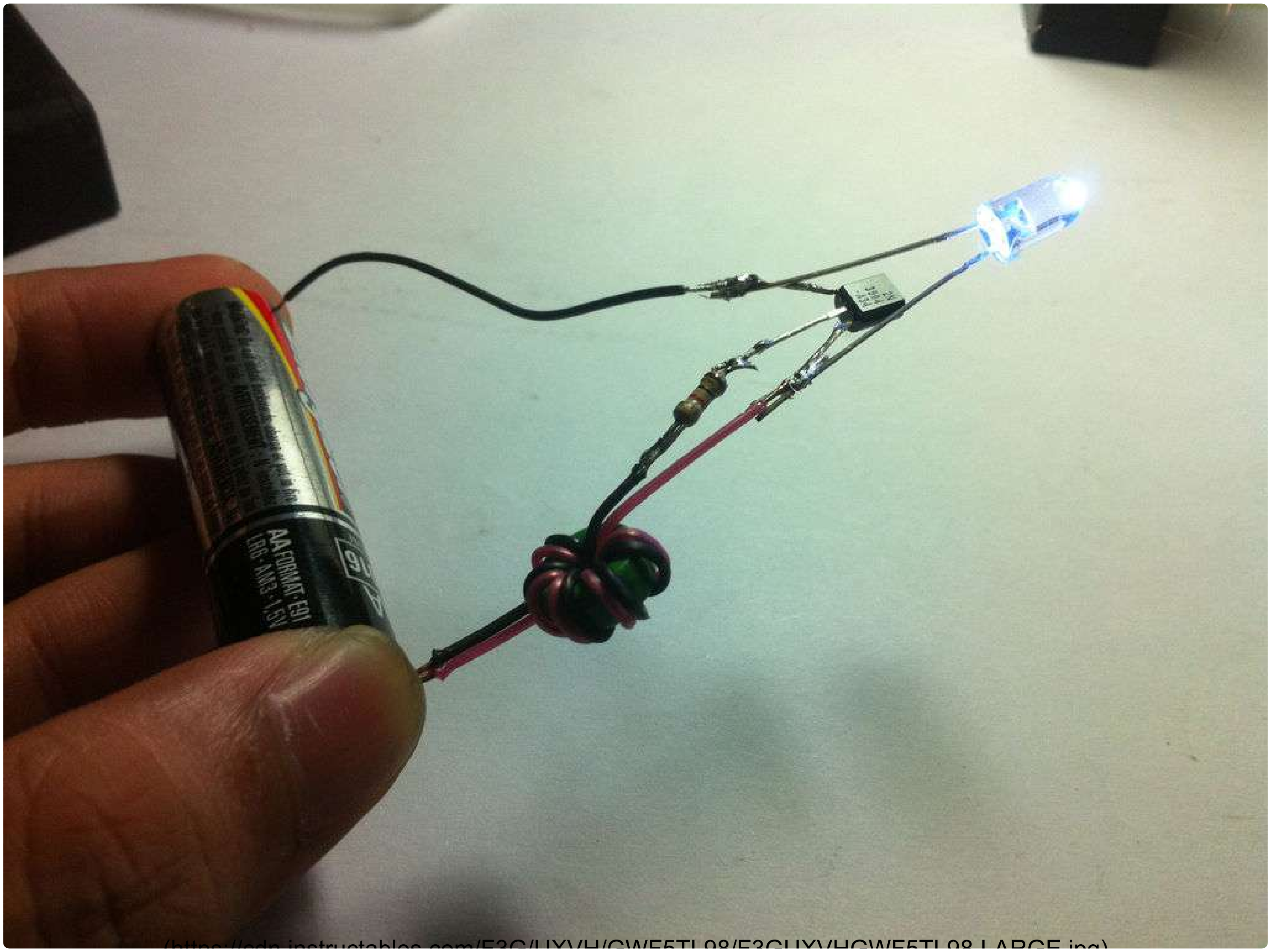

(https://cdn.instructables.com/F3C/UXVH/GWF5TL98/F3CUXVHGWF5TL98.LARGE.jpg)

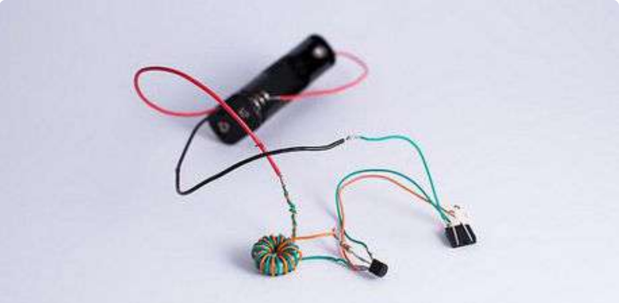
(https://cdn.instructables.com/FB5/ROY8/GWHN3

Mini
Version
(https://cdn.instructables.com/FF8

(https://cdn.instructables.com/F3C

Today I am showing you how to make a very simple joule thief. A joule thief has many applications, the best gadget that I made with was a "Water Powered Lamp", soon I'm going to post on a guide about it but first I need to post this guide. I used an iPhone 4S as my camera :)))

What Is A Joule Thief ?

To simplify everything, a "joule thief" is a circuit that helps drive an LED light even though your power supply is low. What can we do with it? We can use it to squeeze the life out of our old, almost drained, non functioning batteries. This project can also be considered as a green and environmental experiment, we can also use it as a flashlight that can be ran by an old, weak, almost drained battery. I even tried to use my water powered battery from my previous instructable the "Water Powered Calculator (https://www.instructables.com/id/Water-Powered-Calculator/)", the project was featured and displayed in instructable's front page in the "Technologies" category.

My Next Projects That Involves A Joule Thief: (soon to be posted)
- Water Powered Lamp
- Water Powered Flash Light
- Dead Battery Drainer Lamp

Here's A Video From Make Magazine:

Add Tip Ask Question

Step 1: Parts and Materials

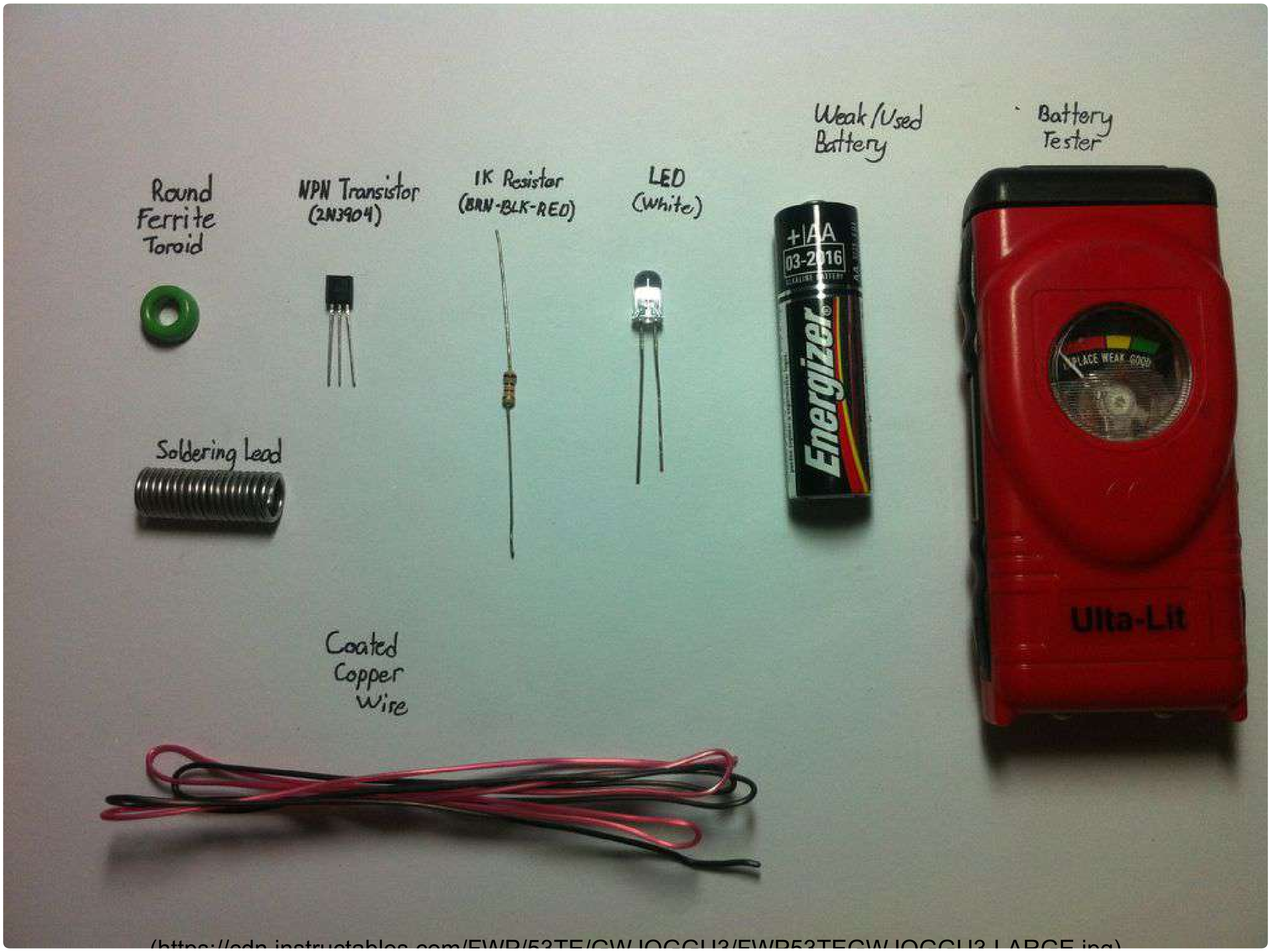
Round Ferrite Toroid
NPN Transistor (2N3904)
1K Resistor (BRN-BLK-RED)
LED (White)
Weak/Used Battery
Battery Tester
Soldering Lead
Coated Copper Wire
+ AA
03-2016
Energizer
REPLACE WEAK GOOD
Ulta-Lit

(https://cdn.instructables.com/FWP/53TF/GWJOGGH3/FWP53TFGWJOGGH3.LARGE.jpg)

The Parts Needed Are: (click the item to know where to find/ buy)

- Round Ferrite Toroid (can be found in old CFL bulbs)
- Old/ Used Batteries (can be found in garbage cans)
- NPN Transistor (2N3904) (http://www.radioshack.com/product/index.jsp?productId=2062609)
- 1K Resistor (BRN-BLK-RED) (http://www.radioshack.com/product/index.jsp?productId=2062343)
- LED Light (http://www.radioshack.com/product/index.jsp?productId=3096133)
- Battery Tester (http://www.tooldistrict.com/Ulta-Lit-Battery-Tester-5001-p/901233.htm) (optional)
- Soldering Lead (http://www.radioshack.com/product/index.jsp?productId=2062717)

- Copper Wire/ Magnet Wire (http://www.radioshack.com/product/index.jsp?productId=2036277)
- Battery Case/ Holder (http://www.radioshack.com/product/index.jsp?productId=2062247)

I want to share something. Here in the Philippines electronic parts are extremely cheap, they are extremely far cheaper from
radio shack, for example one transistor costs (2 phil. pesos - 6 US cents), a LED cost (9 phil. peso - 29 US cents) and a 1K resistor cost (25 phil. cents - 0.8 US cents). I usually buy thing from Deeco or Alexan. Usually prices here are 15x cheaper from radio shack. Price conversion - $1 US Dollar = P0.31 Philippine Peso (12/24/11).

Add Tip Ask Question

Step 2: Schematic Diagrams

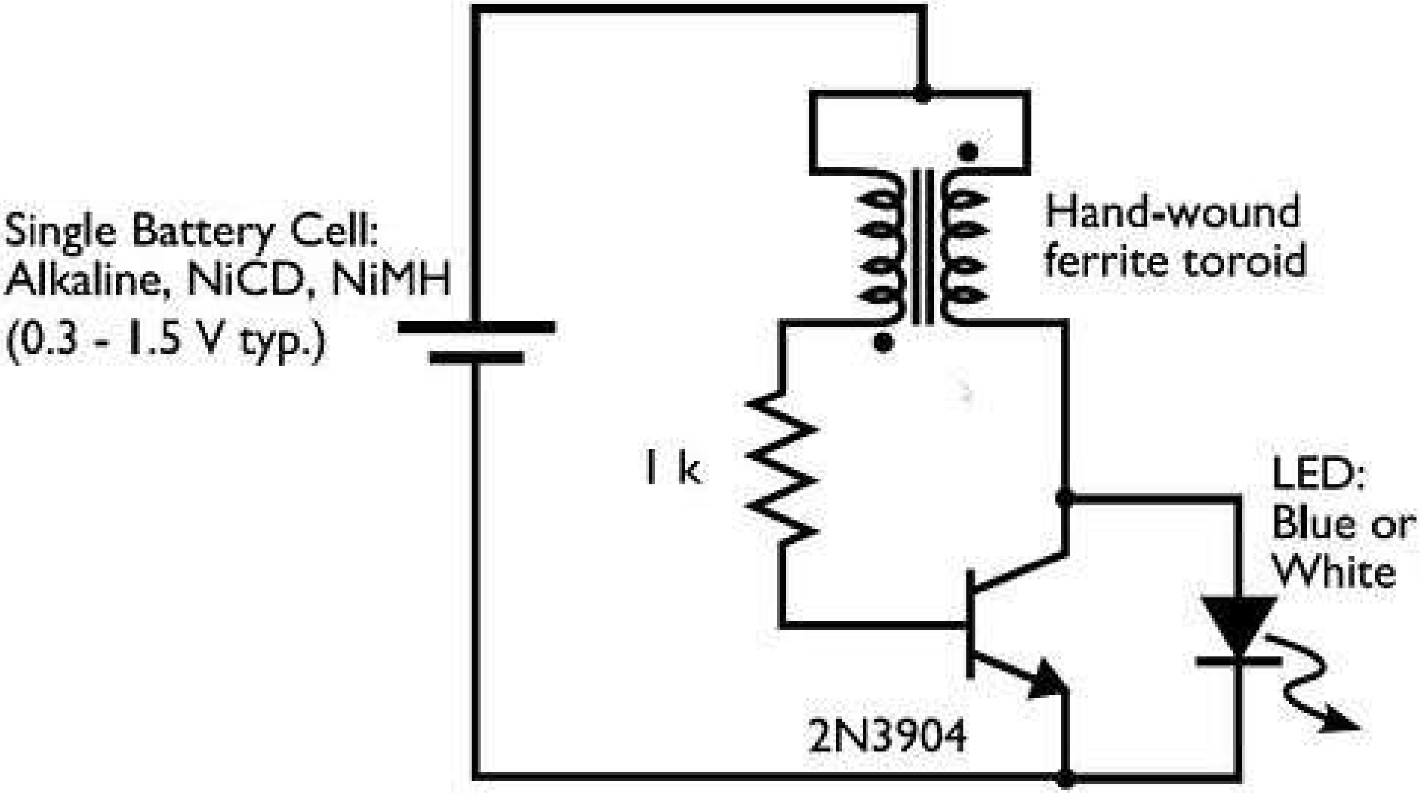
Single Battery Cell:
Alkaline, NiCD, NiMH
(0.3 - 1.5 V typ.)
Hand-wound
ferrite toroid
1 k
LED:
Blue or
White
2N3904

(https://cdn.instructables.com/F3N/S5I7/GWKD43C0/F3NS5I7GWKD43C0.LARGE.jpg)

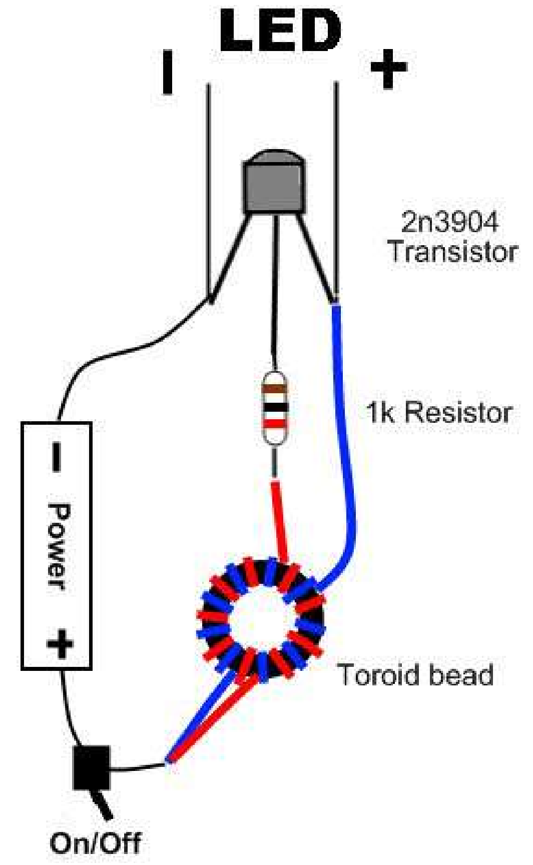
LED
|
+
2n3904
Transistor
1k Resistor
-
Power
+
Toroid bead
On/Off

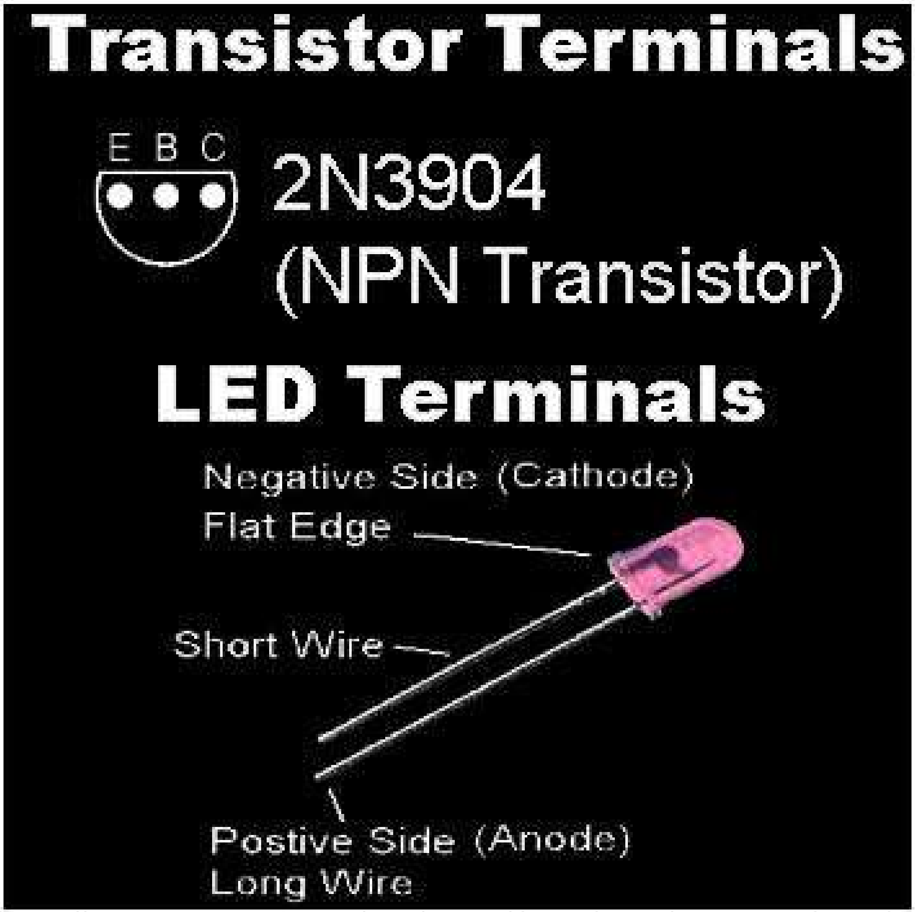
Transistor Terminals
E B C
2N3904
(NPN Transistor)
LED Terminals
Negative Side (Cathode)
Flat Edge
Short Wire
Postive Side (Anode)
Long Wire

(https://cdn.instructables.com/FN0/CRE6/G (https://cdn.instructables.com/F5P/M7EM/GWINSM3D/F5PM7EMGWINSM

Here are the schematic diagrams that are involved with the joule thief circuit.

Add Tip Ask Question

Step 3: Winding Wire at the Toroid

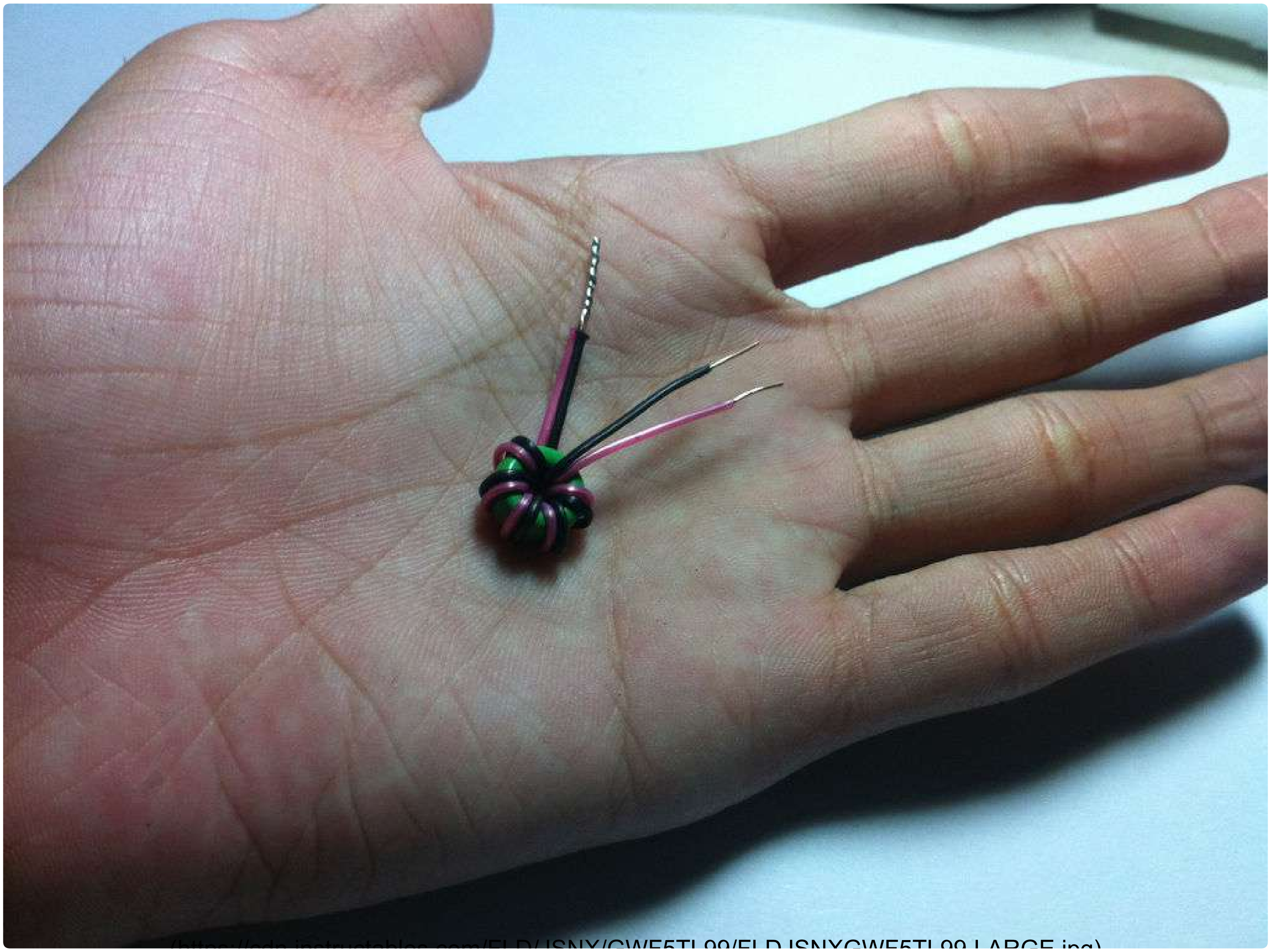

(https://cdn.instructables.com/FLD/JSNY/GW5TL99/FLDJSNYGW5TL99.LARGE.jpg)

First, connect both ends of the copper wire before wounding, be sure to remove the insulation. Then try to solder the ends so it would not split up. Second, wind the wire until you run out of space in the round ferrite toroid. I have some tips for you, try to use a gauge #22 enamel coated copper wire for better performance, oh! my last tip is that "the more you wind the wire to the ferrite toroid the better".

Add Tip

Ask Question

Step 4: Soldering the Resistor

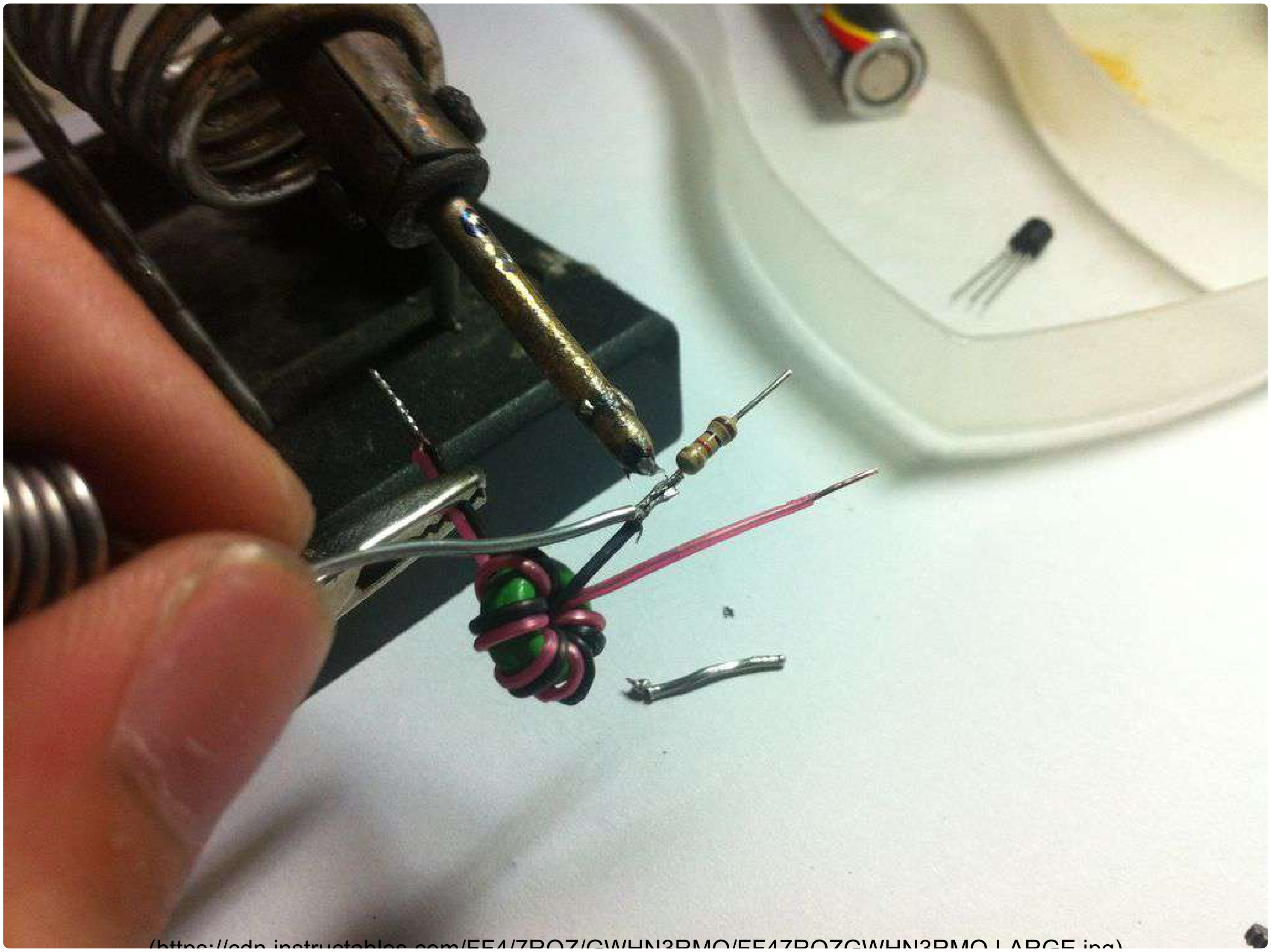

(https://cdn.instructables.com/FE4/7BOZ/GWHN3BMO/FE47BOZGWHN3BMO.LARGE.jpg)

Solder the resistor with one end of the wounded ferrite toroid's wire. Oh! also don't forget to level the other end of the resistor with the other unused wire from the wounded ferrite toroid.

Add Tip Ask Question

Step 5: Soldering the Transistor

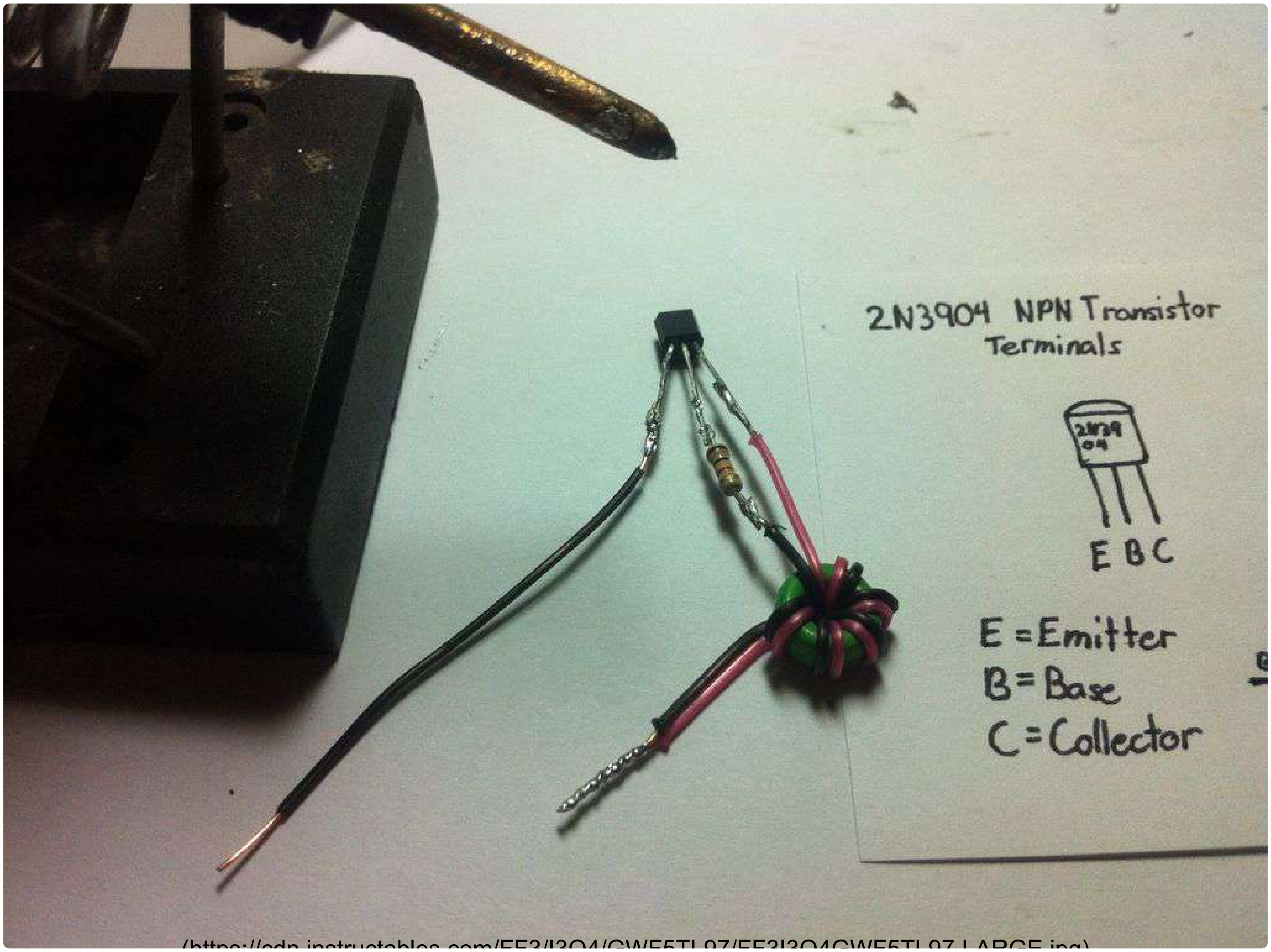
2N3904 NPN Transistor Terminals
2N39 04
E B C
E = Emitter
B = Base
C = Collector

(https://cdn.instructables.com/FE3/I3O4/GWF5TL9Z/FE3I3O4GWF5TL9Z.LARGE.jpg)

Solder the proper connections to the transistor. For the emitter - connect another wire, the wire will be connected to the negative part of the battery. For the base - solder the other end of the resistor to the base. For the Collector Solder the unused wire of the ferrite toroid.

Add Tip　Ask Question

Step 6: Soldering the LED

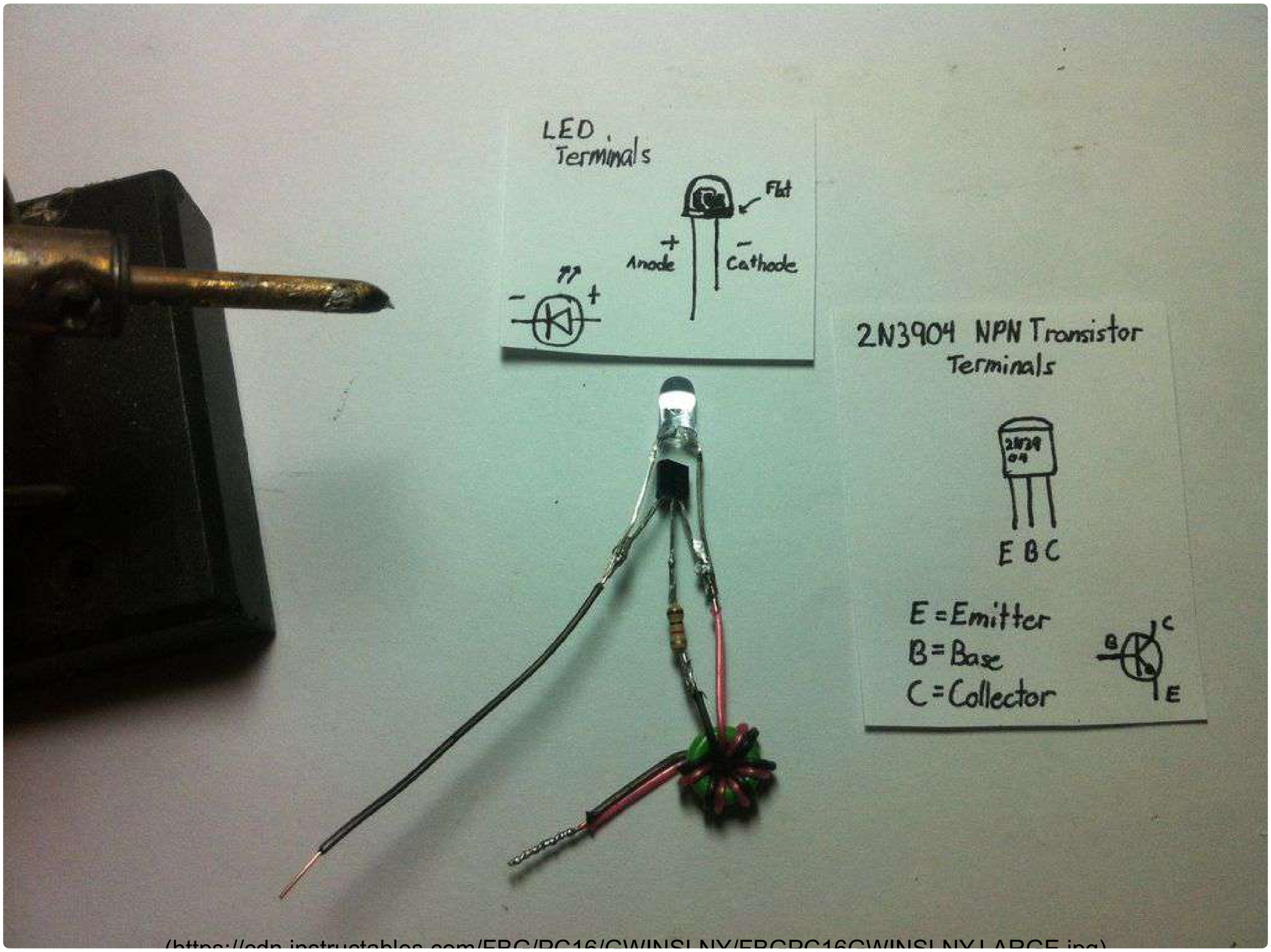
LED
Terminals
Flat
+
Anode
–
Cathode
–
+
2N3904 NPN Transistor
Terminals
2N39
04
E B C
E = Emitter
B = Base
C = Collector
B
C
E

(https://cdn.instructables.com/FBC/PC16/GWINSLNY/FBCPC16GWINSLNY.LARGE.jpg)

Solder the shorter wire of the LED to the tansistor's emitter and the longer part of the LED to the transistor's collector. After all that, you can now trim the excess wires.

Add Tip Ask Question

Step 7: Time to Look for Old Batteries

Energize
REPLAC WEAK GOOD

(https://cdn.instructables.com/F78J93B/GWF5TM0Q/F78J93BGWF5TM0Q.LARGE.jpg)

(https://cdn.instructables.com/FX4JNKD/GWINSM4V/FX4JNKDGW

PHILIPS
DURACELL
Kodak
Energizer
SUPER
High-Tech
EVEREADY

(https://cdn.instructables.com/F40/8TOM/GWINSM

Use your battery tester to confirm that your battery is close to death. The tester is only an optional tool, it's just used to determine the battery's remaining power.

Add Tip

Ask Question

Step 8: Time to Test It - You're Done !!

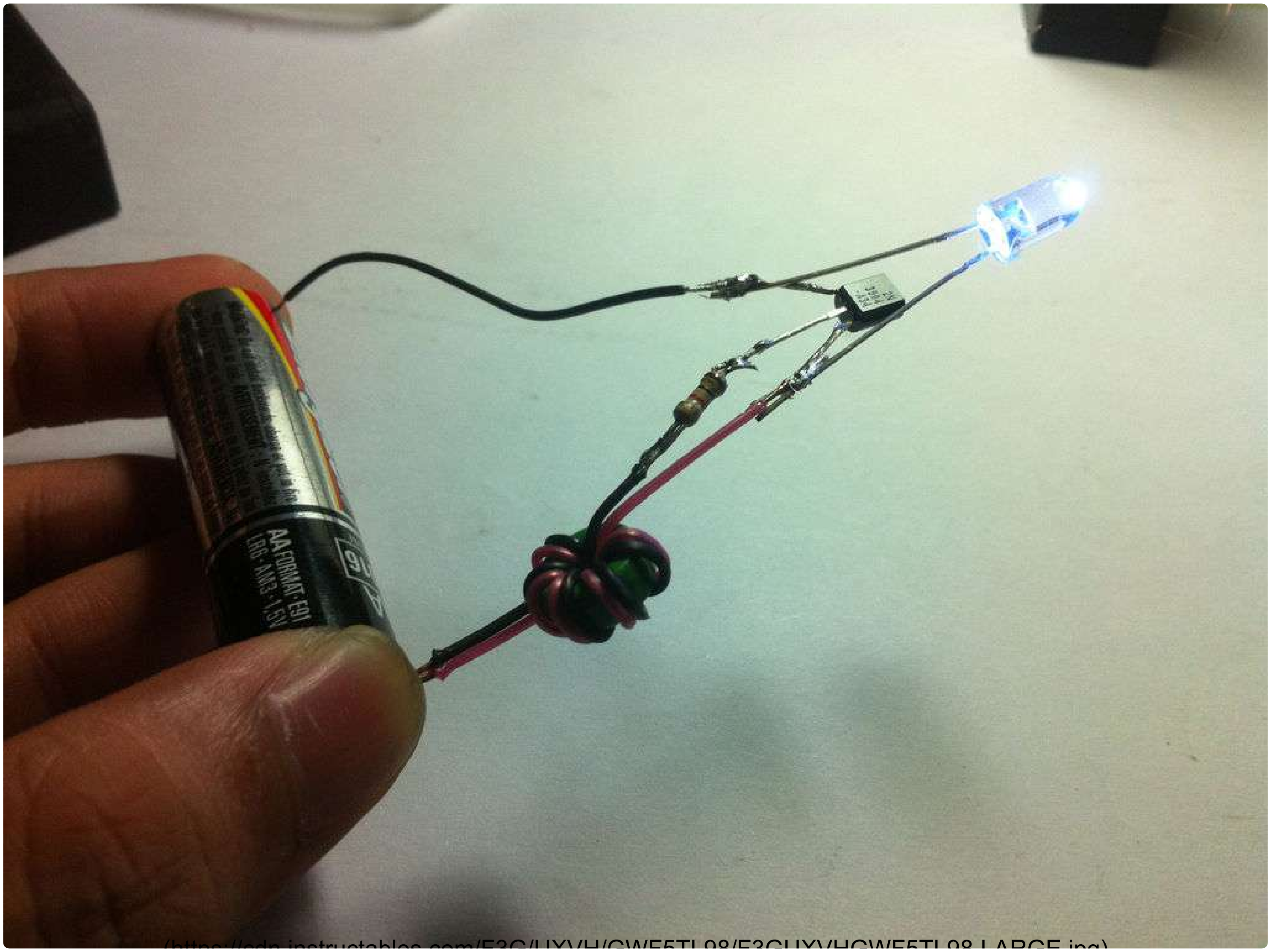

(https://cdn.instructables.com/F3C/UXVH/GWF5TL98/F3CUXVHGWF5TL98.LARGE.jpg)

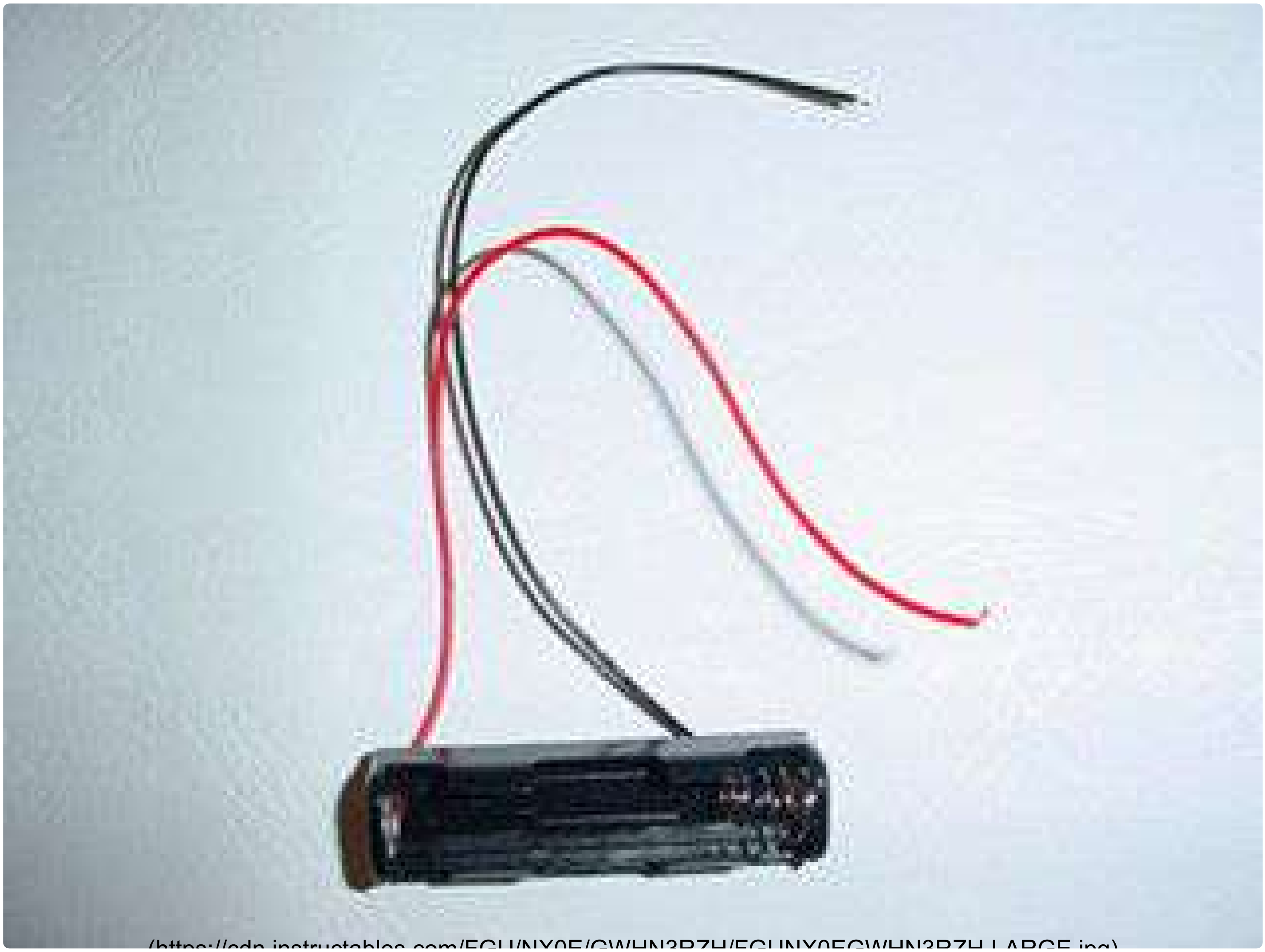

(https://cdn.instructables.com/FGU/NX0F/GWHN3BZH/FGUNX0FGWHN3BZH.LARGE.jpg)

The wire connected to the transistor's emitter should be connected to the battery's negative side and the remaining wire of the ferrite toroid should be connected to the battery's positive side. Oh! one more thing, I advise everyone to use a battery case or attach a conductive magnet for each wire, so you wouldn't hold it all the time.

Add Tip Ask Question

Eliminate Your Power Bills!

Build Your Own Energy Generator for Just $49

Tesla Energy

advertisement

29 Travel Essentials We Won't Leave Home Without | Cartageous

Cartageous

advertisement

8 People Made This Project!

"Joule Thief" LED Night Light

by **ledartist** on September 2, 2011

Table of Contents

"Joule Thief" LED Night Light 1

Intro: "Joule Thief" LED Night Light 2

Step 1: Features 2

Step 2: Technical Overview 3

File Downloads 4

Step 3: Assembly 4

Step 4: Performance 5

Step 5: PCB & Kit 6

Related Instructables 6

Comments 6

Author:ledartist My Blog
I work with LEDs and microcontrollers to create beautiful objects.

Intro: "Joule Thief" LED Night Light

I have many used batteries around. Remote controls, cameras, many electronic gadgets all use batteries, mostly AA size. I always felt guilty for throwing away the used batteries. I know there are rechargeable batteries, but many electronics don't work well with rechargeables.

I also know that those "used" or "spent" batteries usually have some juice left in them. So to come up with a good use of used batteries, I've created a LED night light.

I like having a little night light on when I sleep. LEDs are perfect for this purpose, because they are energy efficient, and good at providing low intensity illumination.

This LED night light operates with just one battery. It utilizes a little circuit called Joule Thief to boost voltage out of an AA battery. I also added a light sensor to turn it on automatically when the surrounding is dark.

The circuit is energy efficient, and requires very low voltage to work. So it effectively sucks every bit of energy out of batteries. This type of circuit is often called "Joule Thief", because it works as though stealing every bit of energy (Joule is a unit for energy) out of battery.

I'm calling this project Night Joule Thief.

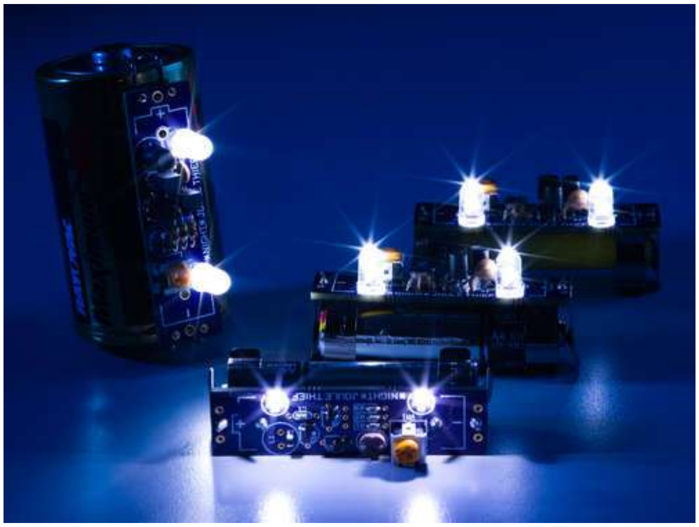

Step 1: Features

Here are the highlights of the Night Joule Thief.

- Compact & streamlined design
- Uses only one AA battery (or any 1.5V battery you can hook up to)
- Easily adaptable to different size batteries - hook up holes to attach home made clips
- Two white LEDs
- Automatic turn on via a light sensor (adjustable sensitivity level)
- Energy efficient - works even with a run-down battery, down to 0.6V
- Choice of through-hole only components or SMD - mix & match on the same PCB

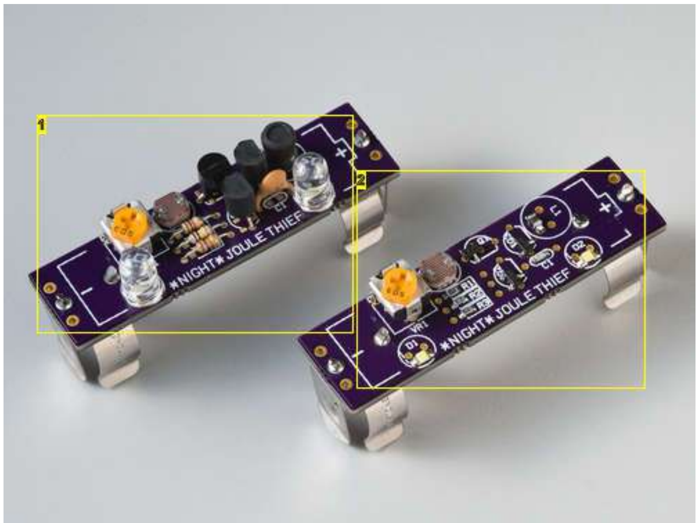

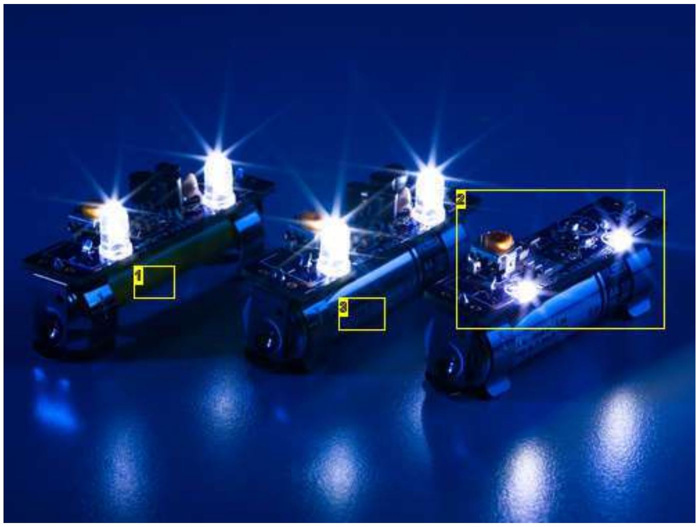

Image Notes
1. Through-hole version.
2. SMD (Surface Mount Device) version. Hard to believe this is the same circuit, but it works exactly the same as one on left. Ok, the LEDs are slightly less bright.

Image Notes
1. AAA battery inserted with a help of small magnets.
2. SMD version - very low profile.
3. Standard AA battery snugly fit in the clips.

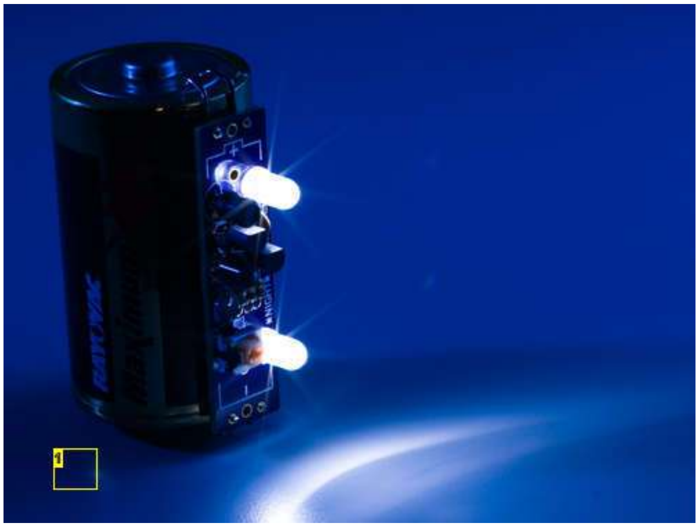

Image Notes
1. Other sizes, such as D cell can be attached by making a pair of little metal clips. Just cut and bend paper clips.

NIGHT JOULE THIEF
VR1
D1
D2
C1
NIGHT JOULE THIEF

Step 2: Technical Overview

"Joule Thief" circuit is an inductor based voltage booster circuit to light LEDs with low supply voltage. As most of you know LEDs need higher than 2V (3V for white LEDs), so usually at least two batteries are needed to light them. The "Joule Thief" circuit was published in 1999 and has been quite popular. You can see the principle of the circuit here. http://en.wikipedia.org/wiki/Joule_thief

My version is a variation that uses single coil inductor, to make the inductor easily obtainable. I design the circuit using readily available parts only, to make it an ideal DIY project.

Circuit

The L1, Q2, Q3, C1, R2, and LEDs D1 & D2 make the Joule Thief. And the Q1, and the rest of the parts form the light sensor. CdS is the device that actually senses the light and change its resistance accordingly. When the surrounding of CdS is bright, it has low resistance (anywhere around 1k to 3k ohm), and when the surrounding is dark, the resistance goes up to 100k to 3M ohm range. So in this circuit, the base voltage of Q1 is controlled by the ambient light level. When the base voltage of Q1 goes more than 0.6V below the power supply(battery) voltage, current goes through R1, turning the Joule Thief circuit on.

The Joule Thief circuit is boosting the battery voltage up to over 6V to light two LEDs in series. LEDs light up with the battery voltage as low as 0.6V! Amazing!

PCB layout can be downloaded as an editable PDF, so you can etch your own board if you like. Custom 2 layer PCB and kit are available for sale as well. The 2 layer PCBs have extra front pads for SMD where possible, so you can build the same circuit with SMD parts as you wish.

Parts List
1x CdS Photoresistor (rated 3k - 0.3M ohm) (CDS1)
1x 1k ohm (R1)
1x 100k ohm (R2)
1x 10k ohm (R3)
1x 50k ohm trim pot (VR1)
1x 22pF (C1)
1x 470uH (L1) (anywhere between 22 - 470uH would work - might have to reduce the C1 value however)
1x 2N5401 or equivalent (Q1) (or just about any general purpose PNP transistor, such as PN2907, 2N3906, etc...)
2x MPSA06 or equivalent (Q2, Q3) (or just about any general purpose NPN transistor, such as PN2222A, 2N3904, 2N4400, etc...)
2x LED (D1, D2) (Just about any LEDs can be used)
2x Battery Clips

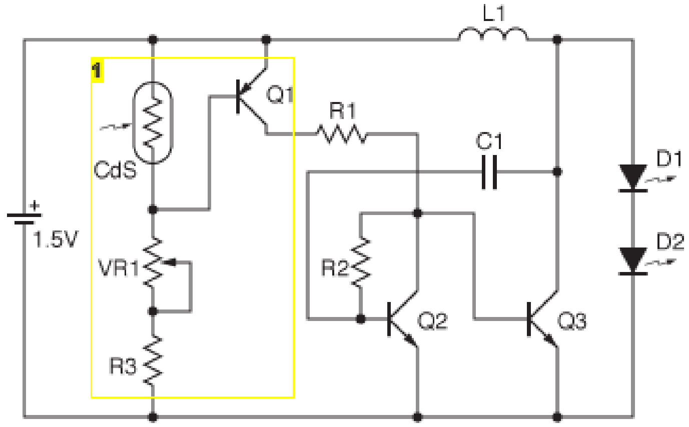

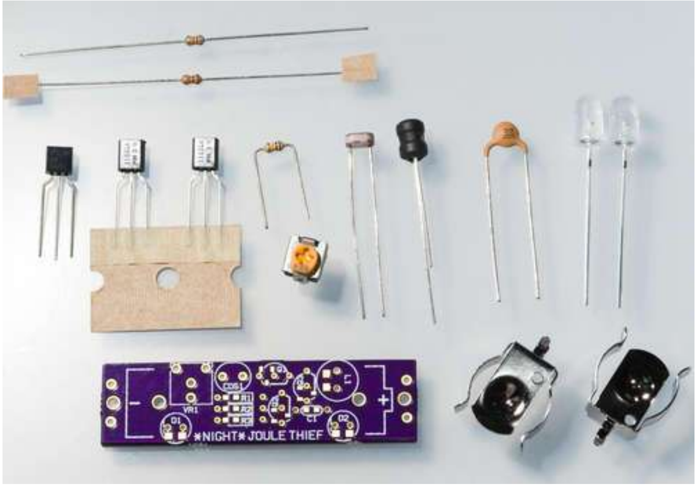

Image Notes
1. Light sensor circuit with sensitivity adjust.

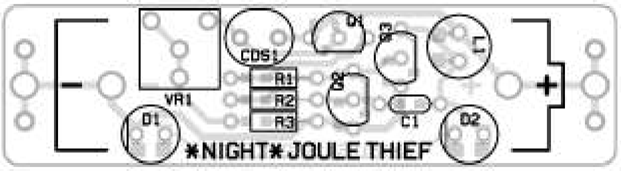

File Downloads

NightJouleThief-PCB.pdf (54 KB)
[NOTE: When saving, if you see .tmp as the file ext, rename it to 'NightJouleThief-PCB.pdf']

Step 3: Assembly

The assembly is very straight forward. Insert the parts into the PCB, and solder them. Start with small components, follow the order below.

Parts List (in assembly order)
1x 1k ohm (R1)
1x 100k ohm (R2)
1x 10k ohm (R3)
1x Photoresistor (rated 3k - 0.3M ohm) (CDS1)
1x 50k ohm trim pot (VR1)
1x 22pF (C1)
1x 470uH (L1)
1x 2N5401 or equivalent (Q1)
2x MPSA06 or equivalent (Q2, Q3)
2x LED (D1, D2)
2x Battery Clips

Transistors, and LEDs have polarities, so make sure to insert them in the correct orientation. Battery holders need a bit of force to snap into the holes. They attach from the back side of PCB as you can see in the picture.

Once everything is soldered in place, double check the part placement, orientation and solder joints. Then insert a battery. The polarity is marked on the front side of PCB.
If you don't see the LEDs light up, don't worry. The room is probably too bright. Take a piece of black paper or tape and block the light from hitting CdS light sensor. (and/or darken the room) If the LEDs still don't come on, turn the trimmer (the little orange thing) with a screw driver, counter clockwise. This makes the sensor less sensitive to light, so the LEDs will come on by just placing the sensor under shade, or turning off the room light.

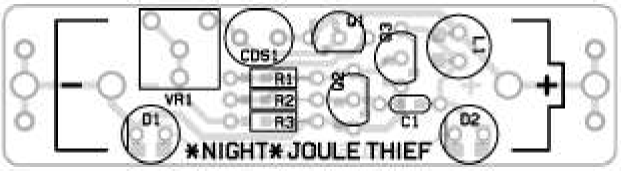

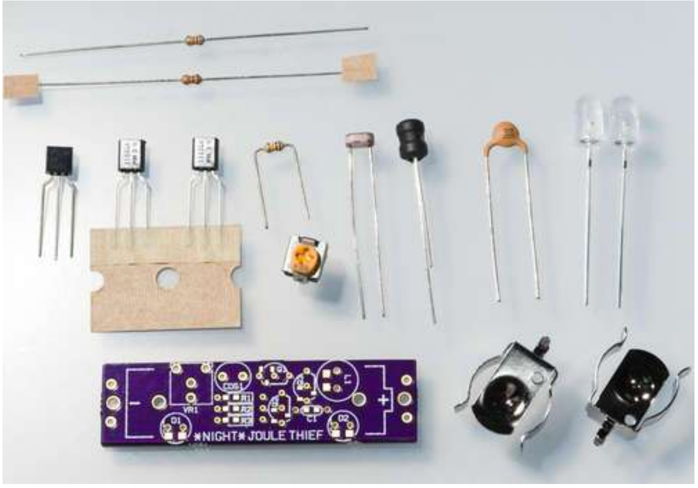

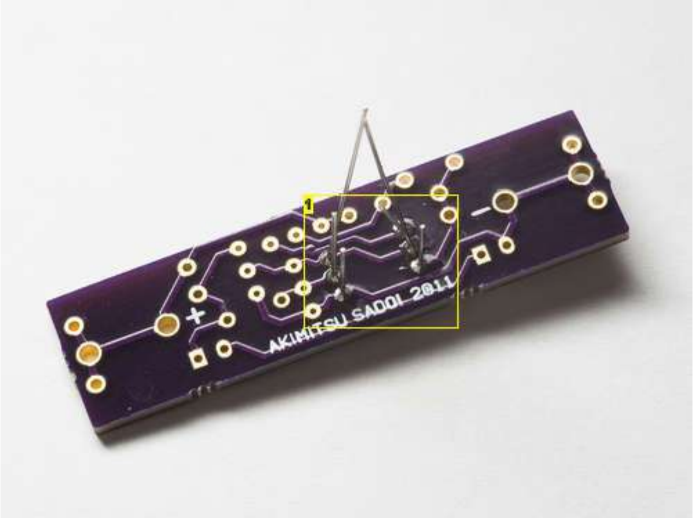

Image Notes
1. Resistors soldered in.

Step 4: Performance

This little night light performs very well. For starters, the brightness is not bad for using just one AA battery. I've been using these as flashlights as well.

The light sensor also works very well. Once adjusted, the light is steadily off during the day, even when you put the sensor under shade. Only when you block the sensor by a black object, the light would turn on. Yet after dusk, the light would come on when you turn off the room light.

A fresh battery lasts for weeks if only used as a night light. And the best use of this light is to "revive" used batteries. Those batteries from remote controls, cameras, etc. usually have quite a bit of juice left in them. Joule Thief sucks the juice out of those batteries till the last drop. It's like getting free energy when you can use something that were going to be thrown out.

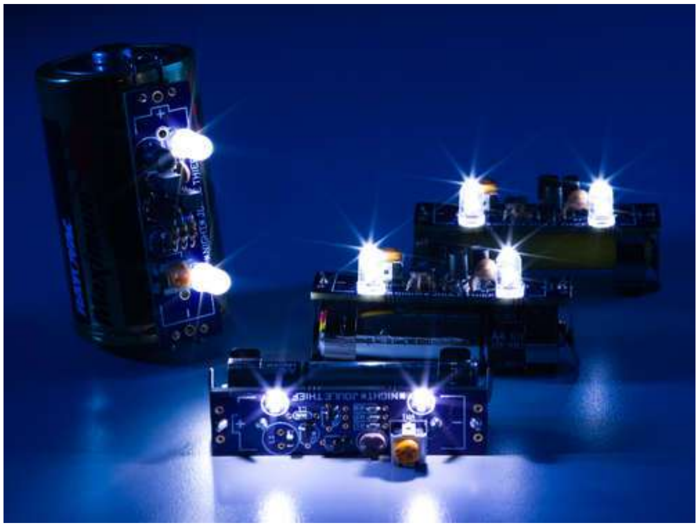

Step 5: PCB & Kit

If you are handy, you can etch your own PCB, and build this night light entirely DIY.

However, to spread the goodness of Joule Thief and to contribute to the greener earth, I am putting together the PCB & kit available.

The details can be found here: http://www.instructables.com/community/Joule-Thief-LED-Night-Light-Kit-PCB/

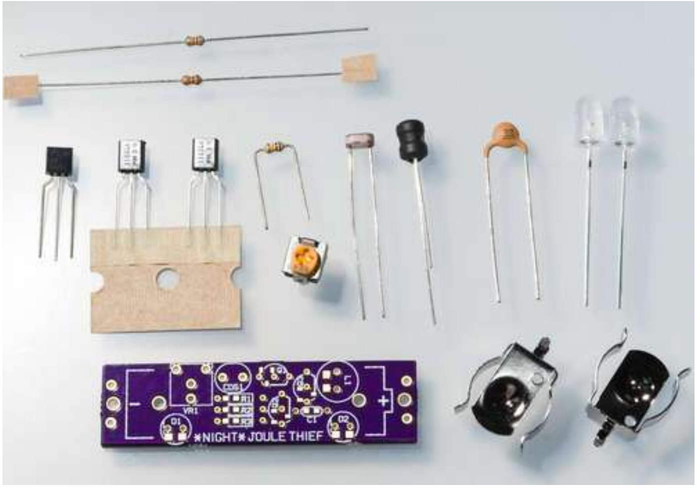

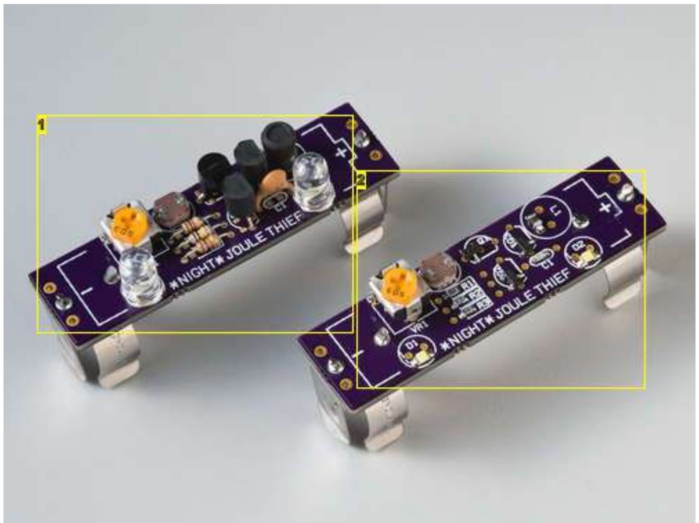

Image Notes
1. Through-hole version.
2. SMD (Surface Mount Device) version. Hard to believe this is the same circuit, but it works exactly the same as one on left. Ok, the LEDs are slightly less bright.

Related Instructables

Joule Thief - use LEDs with only one AA battery! by RazorConcepts

Ferrite, the joule thief (Photos) by botronics

Electronic Night Light by TinkerJim

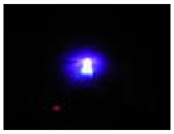

Ultraviolet Light Pen by junits15

Mini 'Joule Thief' (Photos) by Mudbud

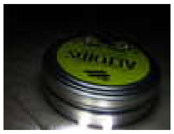

Altoids Joule Thief Flashlight (Photos) by cynical_chemical

Comments

33 comments **Add Comment**

tosacj says: Sep 4, 2011. 10:47 AM **REPLY**
Great idea. Let me know when the kit becomes available.

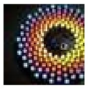

ledartist says: Sep 4, 2011. 12:38 PM **REPLY**
I will post the information on "Kits" section of the forum and on my blog (theLEDart.com). The kit & PCB are scheduled to be ready in mid September.
Thanks, Aki

haxcess says: Sep 5, 2011. 8:16 AM **REPLY**
This is a very neat application of the joule thief. I am wondering if it would be possible to replace the light-sensor circuit with a small battery charging circuit (solar). I suppose that would ruin the beautiful form factor you have here.

apalacios2 says: Sep 5, 2011. 7:07 AM **REPLY**
Excellent circuit. May I remove the CdS circuit and use a button switch instead (thru the resistor connected to the transistor)?
I find great that it uses a coil instead of a toroid transformer. Kudos!

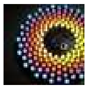
ledartist says: Sep 5, 2011. 7:28 AM **REPLY**

Yes, you can. In fact I made sure that you can fit a 6mm tactile switch in place of trimmer pot. You can omit the CdS, and solder a switch in place of trimmer pot.

You can also remove Q1 and put a switch there as well, but my PCB won't accommodate that.

Yeah I found having to wind my own inductor a hassle, so I designed this circuit to use off-the-shelf inductors.

Aki

bhvm says: Sep 4, 2011. 9:17 PM **REPLY**

Excellent build!
How many mA does this circuit give? Can we use a single 150mA power LED in place of 2x 5mm LEDs?

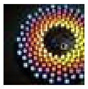
ledartist says: Sep 5, 2011. 7:20 AM **REPLY**

The current through the LEDs is about 20mA or less peak (it's pulsed current in about 50kHz)
So you won't get any more light by using a high power LED. By using only one LED, you do get a bit more brightness per LED though. However just a regular LED will give you the same brightness as a high power LED because you are not driving the LED with high current anyway...

It's not that you can't drive high current LED, but this circuit is designed to run with as little power as possible.

Aki

abbtech says: Sep 4, 2011. 7:10 PM **REPLY**

Great looking project.

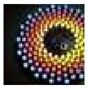
ledartist says: Sep 4, 2011. 7:18 PM **REPLY**

Thanks!

vruiz3 says: Sep 4, 2011. 1:56 PM **REPLY**

afraid of the night??? :D

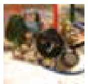
timotet says: Sep 4, 2011. 1:21 PM **REPLY**

great job!

gnafpliotis says: Sep 3, 2011. 3:06 PM **REPLY**

Can't we hack laptop's batteries that way to perform more? Is it only letting through a small amount of amps?

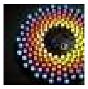
ledartist says: Sep 4, 2011. 1:17 PM **REPLY**

Laptops and many other "high-tech" gadget has many of inductor based boost/buck (to reduce voltage) voltage converter circuit in them. Especially LED back light screens use voltage converter to efficiently drive LEDs.
So in a way the battery life is already enhanced. (Some gadgets are better than others, of course...)

Aki

jolshefsky says: Sep 4, 2011. 6:19 AM **REPLY**

Yes, but only if you wanted to use your batteries once. Once a lithium rechargeable (like in laptops) is discharged past a certain point, it can no longer be recharged. So there is always some energy left in a laptop battery even when it says it's dead — albeit energy you can't use without ruining the battery.

stoobers says: Sep 4, 2011. 7:29 PM **REPLY**

I've heard this theory. Have you done any experiments that might validate this theory?

jamwaffles says: Sep 4, 2011. 6:14 AM **REPLY**

This might not be 100% accurate but yes, that's why we can't use the same technique for laptops. Joule thieves ramp the voltage up, but the current goes down due to V = IR, assuming the load is constant which it is if it's the same laptop ;-)

jamwaffles says: Sep 4, 2011. 6:15 AM **REPLY**

A very professional looking product. This is a great idea; the amount of power saved using old batteries instead of a herd of plug-in night lights, as well as the amount of batteries re-used is incredible.

Just out of interest, how long do these night lights last on one "averagely discharged" battery?

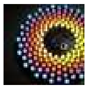

ledartist says: Sep 4, 2011. 1:10 PM **REPLY**

The battery lasts very long, most likely much longer than you might think.
It's hard to define "averagely discharged", but I had one cell that was already at 0.7V (to low to be used with anything), and the although the LEDs were dim, they lighted for over 48 hours.
With my "dead" batteries coming out of a wireless mouse still have over 1V, I have plenty of light.

Aki

YakAttack says: Sep 4, 2011. 8:24 AM **REPLY**

I concur!
@ledartist: Any tests on battery life so far?

nymgeek says: Sep 4, 2011. 7:42 AM **REPLY**

Are those PCBs made by Laen at dorkbot pdx?

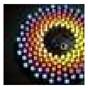

ledartist says: Sep 4, 2011. 1:04 PM **REPLY**

Yup. They are very good!

acmefixer says: Sep 4, 2011. 10:06 AM **REPLY**

I should have said in my previous comment that this does not mean there is anything wrong with your circuit. I think I would change Q2 and Q3 to an easier to obtain transistor such as a 2N4401 or PN2222A. Q1 could be a 2N3906 or just about any PNP transistor. If these changes are made, the resistors, especially R2, might need to be reduced. Thanks.

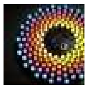

ledartist says: Sep 4, 2011. 1:03 PM **REPLY**

Yes, you can use just about any general purpose transistors. R2 should be fine with most transistors, but 47k ohm might work better with some of them.

Aki

acmefixer says: Sep 4, 2011. 10:00 AM **REPLY**

Unfortunately the link you gave to Joule Thief in Wikipedia is for a poorly written and totally inadequate definition. The authors (apparently several over time) do not have a firm understanding of how a JT works, and furthermore, they have made a mess of it. I have added comments in the discussion and some errors have been deleted, but it is still unworthy of use for a reference.

Also, by definition, the original blocking oscillator circuit later given the Joule Thief name used only a single transistor. Your circuit is not a one transistor blocking oscillator and bears little resemblance to the original JT, so I don't believe it should use the same name.

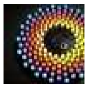

ledartist says: Sep 4, 2011. 12:59 PM **REPLY**

I know the original circuit only uses one transistor, but the two conductor inductor is harder/expensive to purchase, and winding own inductors is a bit of work. I think using one extra transistor is a good trade off for not having to wind an inductor by hand. It also makes economical sense. (transistors are very cheap, so are single coil inductors.)

I did mention that my version is a _variation_ of original, which I find very often on the net. I also think that showing different ways to achieve the same result can be inspirational.

I also contacted the person who named the circuit "Joule Thief" (Big Clive) and he did not have a problem with me using the name.

Aki

StoryAddict says: Sep 4, 2011. 12:54 PM **REPLY**

"Joule Thief" - I love a clever pun!

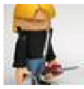

MikeDel says: Sep 4, 2011. 10:21 AM **REPLY**

Looks good. I'm interested!

Admiral Aaron Ravensdale says: Sep 4, 2011. 7:30 AM **REPLY**

I like your "green" thinking.
I also use a joule thief in my expedition light.

I am very glad that you will provide the parts as a kit because many kits are able to solder a circuit but etching a little bit too dangerous for them.

Thanks for your endeavour to make this instructables buyable for many people…

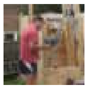

kevinhannan says: Sep 4, 2011. 6:35 AM **REPLY**

fantasic project and incredibly well explained.

UOS says: Sep 3, 2011. 2:59 PM **REPLY**

Great instructable! Waiting the kit version to buy one but also I'm gonna make mine from scratch :D

tphillips2 says: Sep 3, 2011. 8:07 AM **REPLY**

You are so inventive! Cool idea and great looking light!

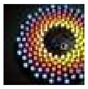

ledartist says: Sep 3, 2011. 8:18 AM **REPLY**

Thanks!

gweeds says: Sep 3, 2011. 4:41 AM **REPLY**

What a great idea! must build one and see how long it gets out of a "used'ish" battery

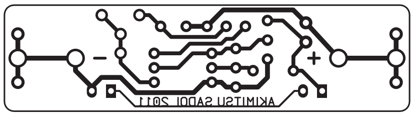
–
+
AKIMITSU SADOI 2011

advertisement

Secrets from the Sound Suite with St. Vincent

Sponsored by: **W Hotels**

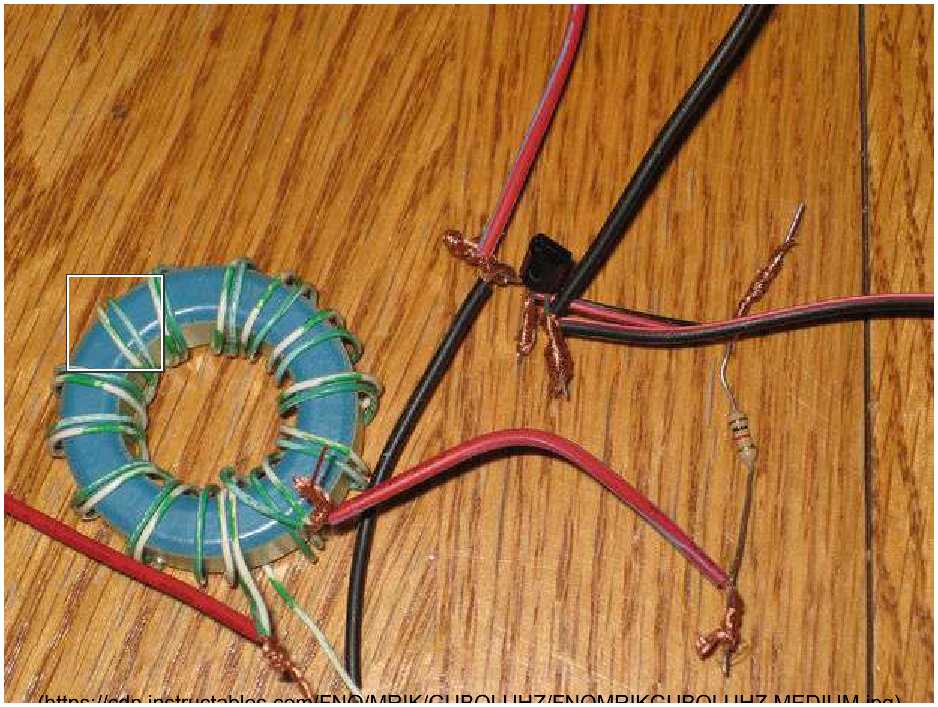

(https://cdn.instructables.com/FNO/MBIK/GUBOLUHZ/FNOMBIKGUBOLUHZ.MEDIUM.jpg)

advertisement

Tektronix

LOW LEVELS, HIGH ACCURACY

Ultra-low measurements using Keithley sensitive instruments

LEARN MORE →

KEITHLEY
A Tektronix Company

this is my version of the famous joule thief. there's alot of these things out there so i did alot of research and made it as simple as i can without soldering or complicated math. i harvested these parts from an older dell computer that was given to me to scrap. there is only a few parts needed to build this project:

1: toroid bead (ferrite core)
2: 1k ohm resister (brown black red)
3: npn transistor (i used the 2n3904)
4: thin wire

About This Instructable

7,340 views

23 favorites

License:

(cc) BY-NC-SA

Step 1: Wind the Coil

Joule Thief Without Soldering by

florlayamp (/member/florlayamp/) in electronics (/technology/electronics/)

Download (/id/Joule-thief/) 4 Steps + Collection I Made it! Favorite Share

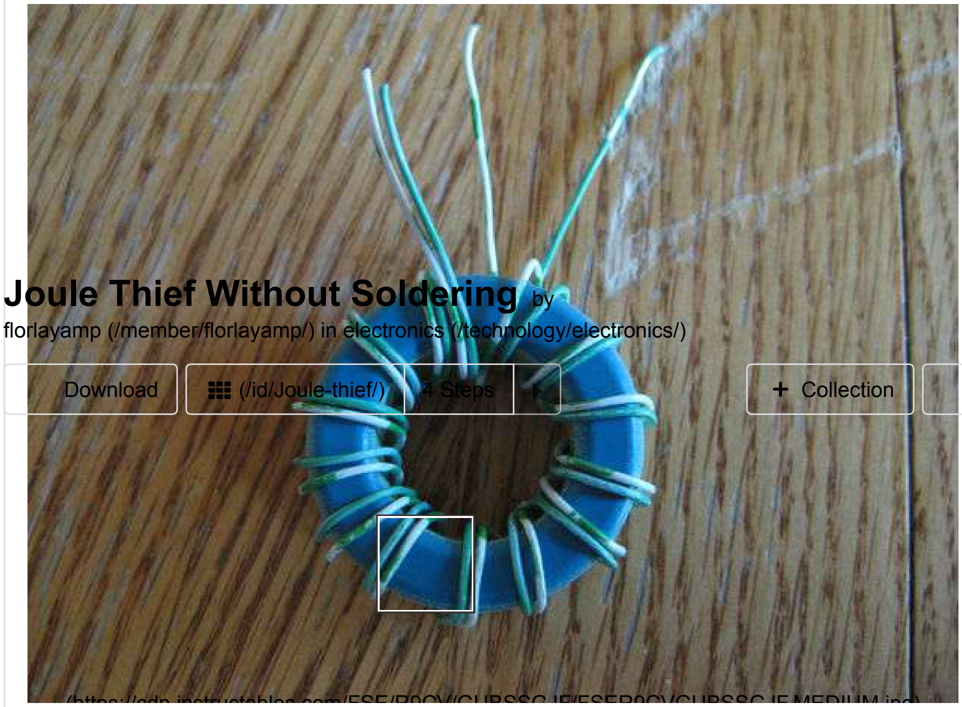

(https://cdn.instructables.com/F5F/R9QV/GUBSSGJF/F5FR9QVGUBSSGJF.MEDIUM.jpg)

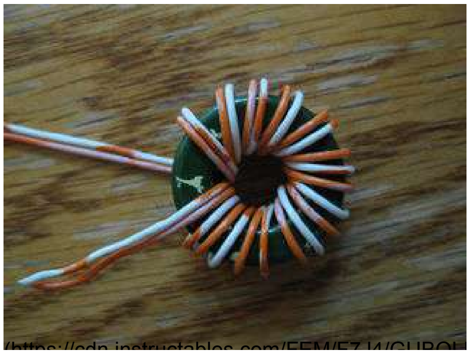

(https://cdn.instructables.com/FFM/FZJ4/GUBOH

(https://cdn.instructables.com/F57/RU2L/GUBRG

i wound a couple sizes of these things using the same wire. the type of wire i used came from an ethernet cat5 cable. i like it because its a solid core wire that stays in place where i put it. i have 11 turns of wire in the orange one and 13 in the green one. i looked all over the internet to find the reason for the number of turns but couldnt find an exact answer, and 11 winds is all the smaller (orange one) could take, and the green one i just put the 13 turns in it from the same length of wire. the orange coil is half the size (about the size of a dime) of the green one (little bit smaller than a 50cent coin).

Step 2: Connect the Resistor

florlayamp (/member/florlayamp/)

(http://https://www.facebook.com

(/member/florlayamp/)

Follow 98

Bio: I love building things, and I'm a twin, we both are electrical engineers and inventors. Feel free to ask anything else

More by florlayamp:

(/id/How-Twins-Change-Oil-In-A-Jeep/)

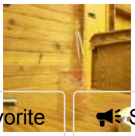

(/id/Take-A-Ghost-Picture-With-Almost-Any-Camera/)

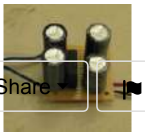

(/id/Voltage-doubler-for-wind-generators/)

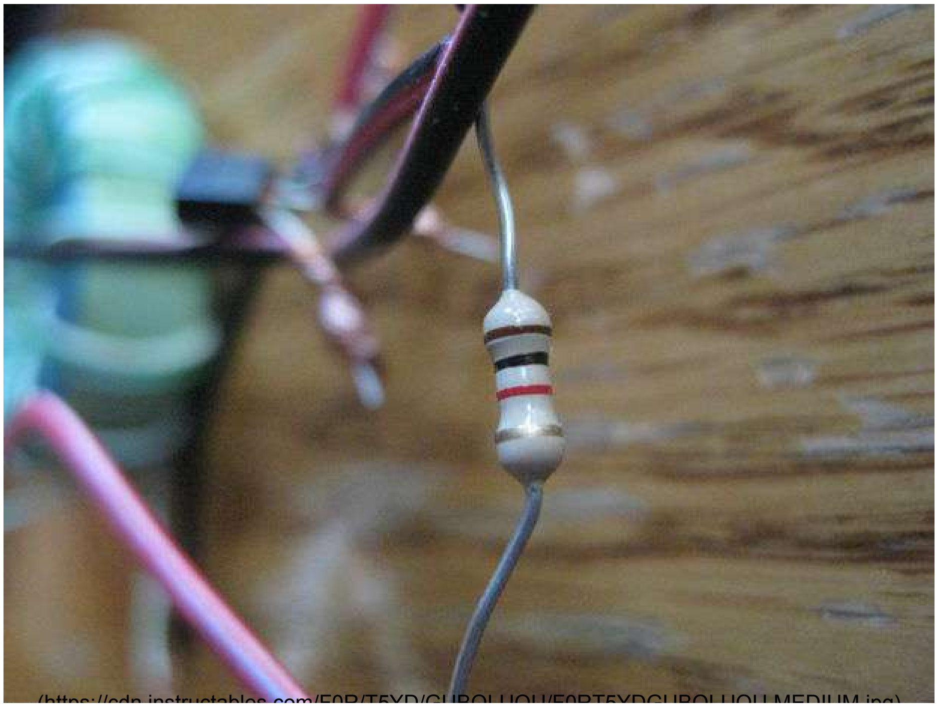

(https://cdn.instructables.com/F0R/T5XD/GUBOLUOU/F0RT5XDGUBOLUOU.MEDIUM.jpg)

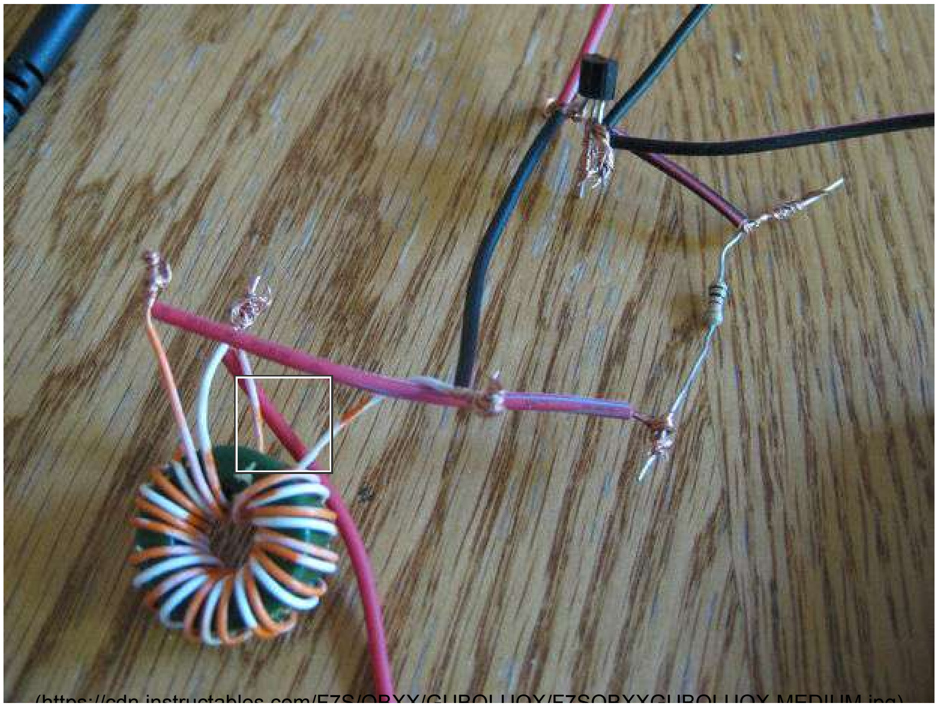

(https://cdn.instructables.com/FZS/QBXY/GUBOLUOX/FZSQBXYGUBOLUOX.MEDIUM.jpg)

once the coil is wound, cut off the excess wire but make sure to leave enough to work with. strip the ends of all four wires, then take 1 wire of **opposite color from each side** and twist them together. then since i didnt solder any wire, i cut up several small wires to piece the rest of it together. unless you want to solder it you can skip this part. but i took a piece of wire and attached it to one of the single wires on the coil (it doesnt matter which one) and attach the other end of the piece of wire to the 1k ohm resister, then take another piece of wire and attach it to the other single wire which will go to the **collector** leg of the transistor. the final piece of wire for this step attaches to the twisted pair of wires which will become the positive input for the circuit.

Step 3: Connect the Transistor

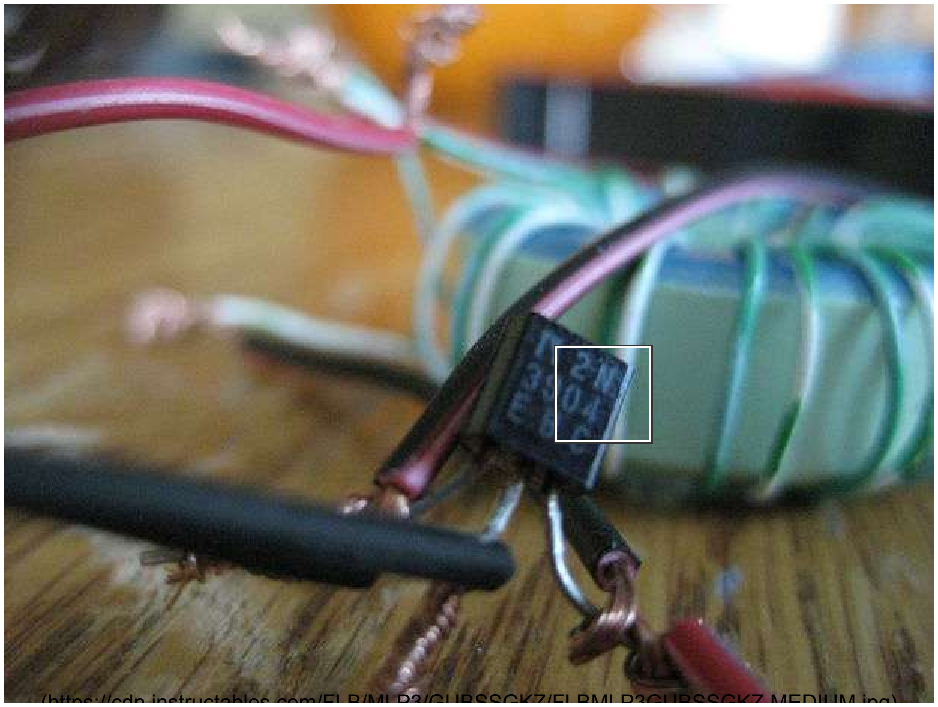

(https://cdn.instructables.com/FLB/MLB3/GURSSCKZ/FLBMLB3GURSSCKZ.MEDIUM.jpg)

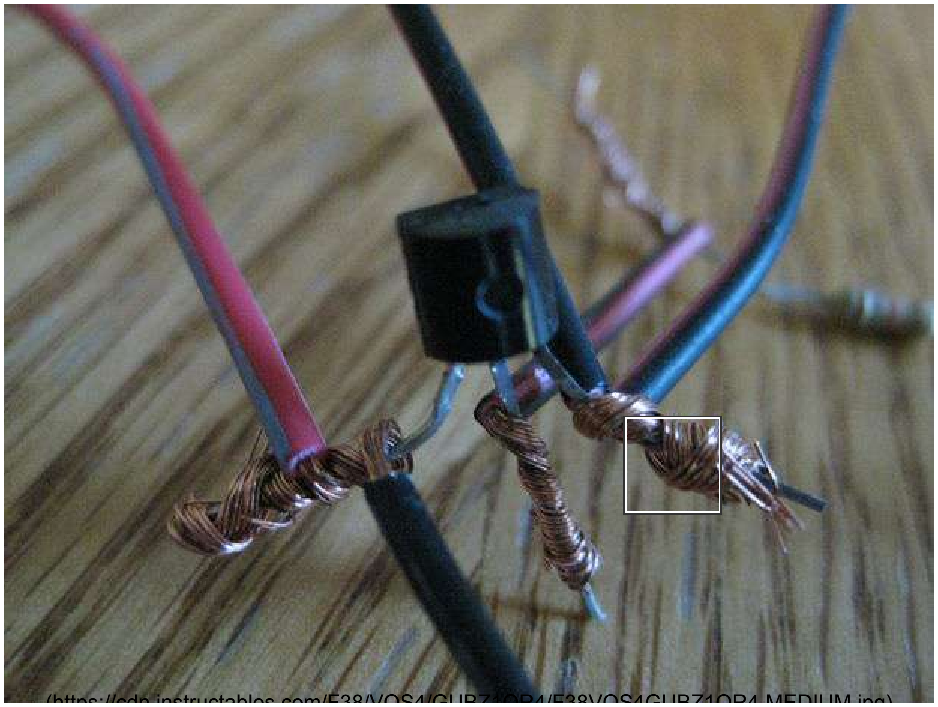

(https://cdn.instructables.com/F38/VOS4/GURZ1QR4/F38VOS4GURZ1QR4.MEDIUM.jpg)

next take a piece of wire and attach it to the other side of the resistor, and then take the other end of the piece of wire and attach it to the **base leg** of the transistor. next, as i mentioned in the previous step, take the open piece of wire from the coil and attach it to the **collector leg** of the transistor, also i attached another piece of wire to this leg for the positive output side . last of all connect two wires to the **emitter leg** which will be the negative side of the circuit. (one wire to negative on battery, then one to the negative output side of an LED light or whatever else you put on it).

Step 4: Hook It Up and Enjoy!

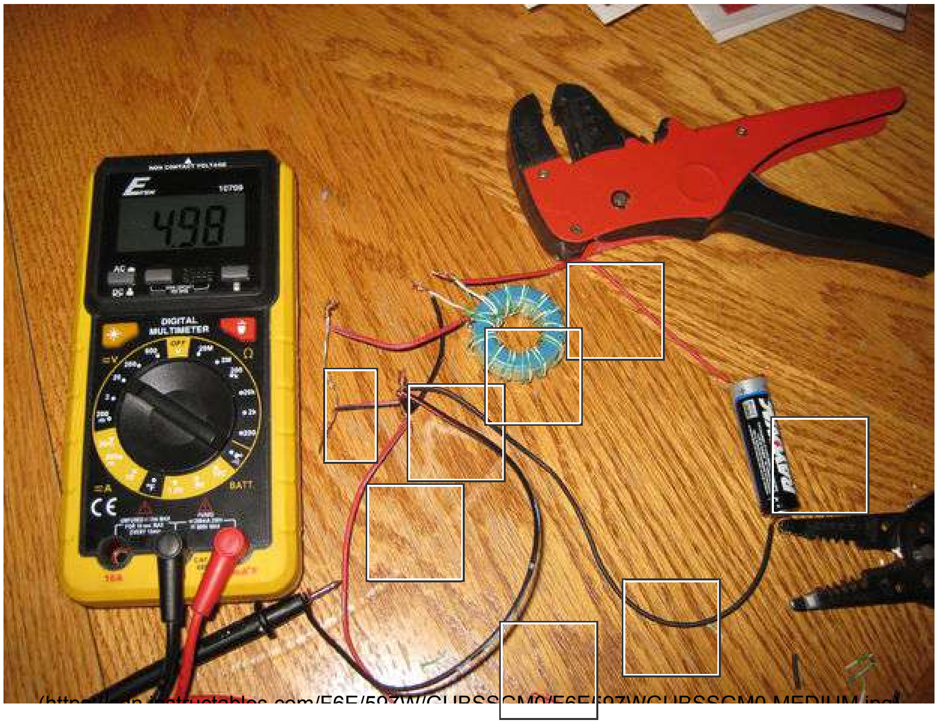

(https://cdn.instructables.com/F6F/597W/GURSSCM0/F6F597WGURSSCM0.MEDIUM.jpg)

time to hook it up! the positive wire from the double twisted wire on the coil goes to the positive power source, and one of the negative wires on the transistor go to the negative side on the power source. the positive wire on the transistor and the other negative wire on the transistor go to the output power. once its all wired up its ready for use. as you can see in the picture, i hooked up a new AA battery and got about 5 volts! not bad for a 1.5v to 5v boost. and the orange coil circuit i also made produces about 2.5v from a 1.5v source, so this orange one only produces another volt while the green one produces alot more. i dont know why this is really. my only guess is the size of the toroid coil and the 2 winding difference. other than that its all the same parts. below is the results of some playing around with it.

green coil circuit:
1: 1.5v AA produces 5 volts
2: 2 AA 3v produces 20 volts
3: 3v input can power a 12v .25a computer fan.
4: 1.5v lights 3 green LED lights.

orange coil circuit
1: 1.5v AA produces 2.5 volts
2: 2 AA 3v produces 8 volts
3: 3v input also powers same 12v computer fan for 5 seconds...

very inconsistent results between the two but thats what is going on. thanks for looking and hope this helps with your project.

advertisement

Rugged Intel®-Based Single Board Computers
View Boards
X-ES

AdChoices
OXI CLEAN

Comments

We have a be nice comment policy.
Please be positive and constructive.

I Made it! Add Images Post Comment

ToXiCATOM (/member/ToXiCATOM/) 2014-02-01 Reply

can i use 2n2222 transistor?

kyooby (/member/kyooby/) ▸ ToXiCATOM (/member/ToXiCATOM/) Reply 2014-03-27

2N3904 = GP NPN Bi-Polar transistor commonly found in toys and small radios. Otherwise you can get them at radio shack, fry's electronics or order them online.

Other compatible ones are:

NPN

2N2222

2N2222A

2N1613

BC107

2N1711

2N2369

2N2484

2N3704

2N3705

2N3706

2N3903

2N3904

Also, if you use a 2N4401 or BC337 transistor, your LED will be brighter because they can handle more amps.

You can get the toroid and transistor from a dead CFL; the transistor is usually labeled 13002

florlayamp (/member/florlayamp/) ▸ ToXiCATOM (/member/ToXiCATOM/) Reply

2014-02-01

I tried that transistor also and I couldnt get it to output anything, I think the internal circuits inside it are backwards for a joule thief.

More Comments

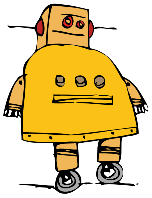

Newsletter

Let your inbox help you discover our best projects, classes, and contests. Instructables will help you learn how to make anything!

enter email | I'm in!

About Us

Who We Are (/about/)
Advertise (/advertise/)
Contact (/about/contact.jsp)
Jobs (/community/Positions-available-at-Instructables/)
Help (/id/how-to-write-a-great-instructable/)

Find Us

Facebook (http://www.facebook.com/instructables)
Youtube (http://www.youtube.com/user/instructablestv)
Twitter (http://www.twitter.com/instructables)
Pinterest (http://www.pinterest.com/instructables)
Google+ (https://plus.google.com/+instructables)

Resources

For Teachers (/teachers/)
Residency Program (/pier9residency)
Gift Premium Account (/account/give?sourcea=footer)
Forums (/community/?categoryGroup=all&category=all)
Answers (/tag/type-question/?sort=RECENT)
Sitemap (/sitemap/)

Terms of Service (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21959721) |
Privacy Statement (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21292079) |
Legal Notices & Trademarks (http://usa.autodesk.com/legal-notices-trademarks/) | Mobile Site (https://www.instructables.com)
(http://usa.autodesk.com/adsk/servlet/pc/index?id=20781545&siteID=123112)

© 2017 Autodesk, Inc.

WONDER HOW TO
MAD SCIENCE

EXPERT SERVICE. UNBEATABLE PRICE.

Make a "Joule Thief" and Create Zombie Batteries for More Power After Death

BY **WILLIAM FINUCANE** 08/23/2013 5:22 PM MAD POWER!

Just about every household gadget we own runs on 1.5 volt batteries of one size or another. Wouldn't it be great if you could reuse all of those dead AA, AAA, and D batteries after they've passed on? It turns out you can make a simple circuit called a "Joule Thief" to reanimate the undead flesh of your deceased batteries and create a zombie battery.

1 | Find Who Owns This Number Enter Any Phone Number Now. Get Full Owner Info. Try Free! spokeo.com
2 | Start Download Step 1- Install DriverUpdate™. Step 2- Scan. Step 3- Identify Missing Drivers! slimware.com

Materials

- LED
- 1k resistor
- 2n904 transistor
- Thin magnet wire
- Solder
- Soldering iron
- Battery clip
- Ferrite core
- Dead battery

sistor for easy soldering.

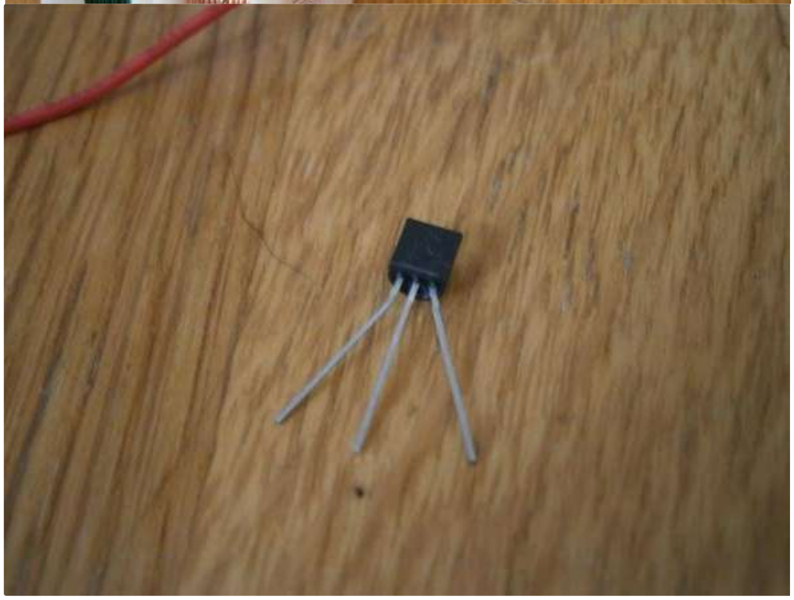

STUDY TO BECOME AN
A/V TECHNICIAN | SOUND EDITOR
AUDIO ENGINEER | MUSIC PRODUCER
SAE INSTITUTE
LEARN MORE

Bend the middle pin back and up behind the black plastic case of the transistor.

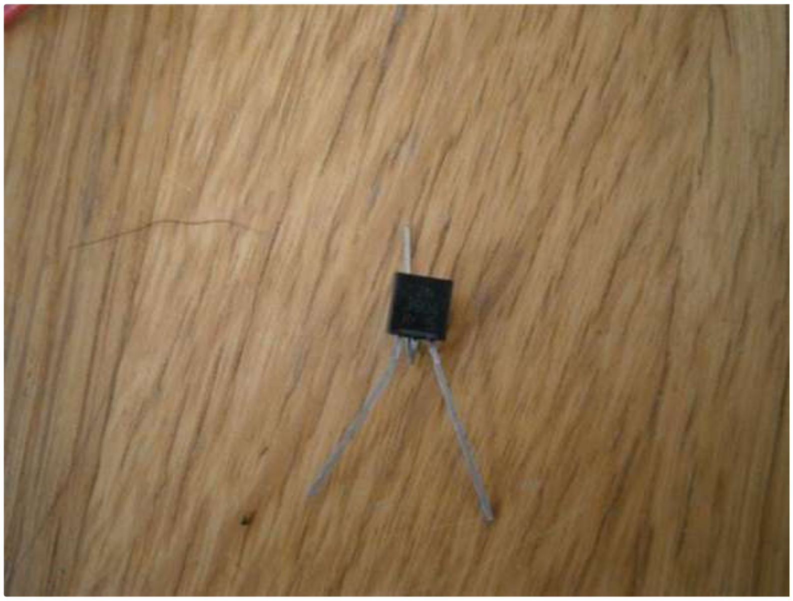

Step 2

Place the LED

The LED is a polarized component. This means it will only work when it is facing the right way in our circuit. The positive lead of the LED is usually longer. Below, the positive lead is on the left, the shorter negative lead is on the right.

Are they *Cheating* on you? Find Out Now

Step 1) Enter Phone # Step 2) View Uncensored Results Show Me Proof >>

Place the LED on the transistor as shown below, with the positive side facing to the right.

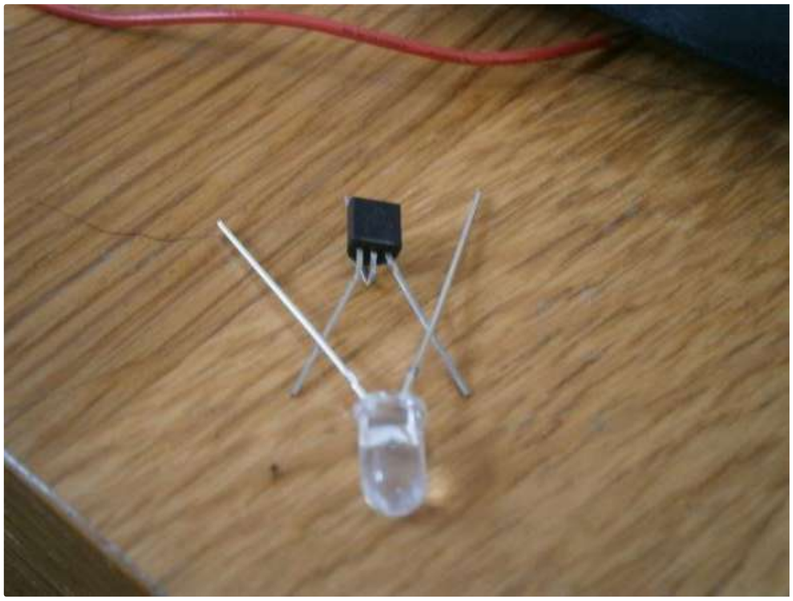

Now, solder the LED to the transistor.

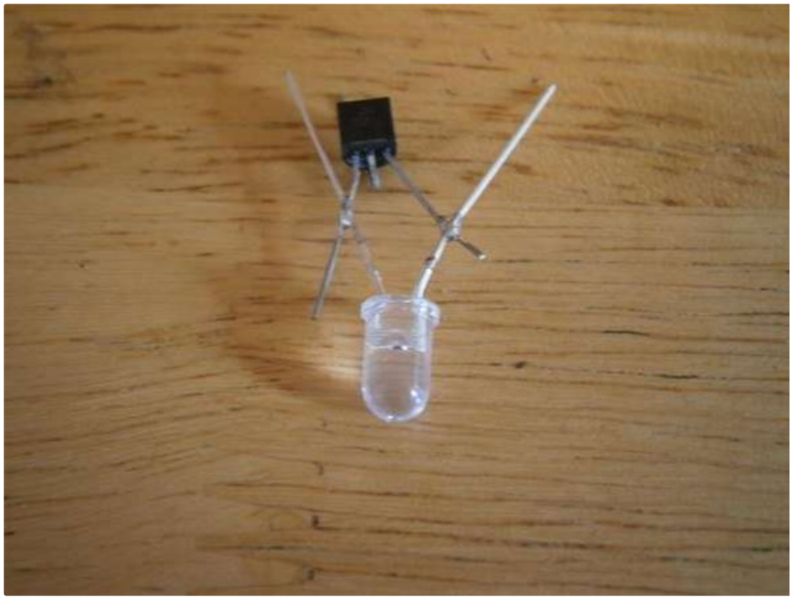

Step 3

Place the Resistor

Place one end of the resistor on the middle pin of the transistor. Make sure the components stay in contact while you solder them.

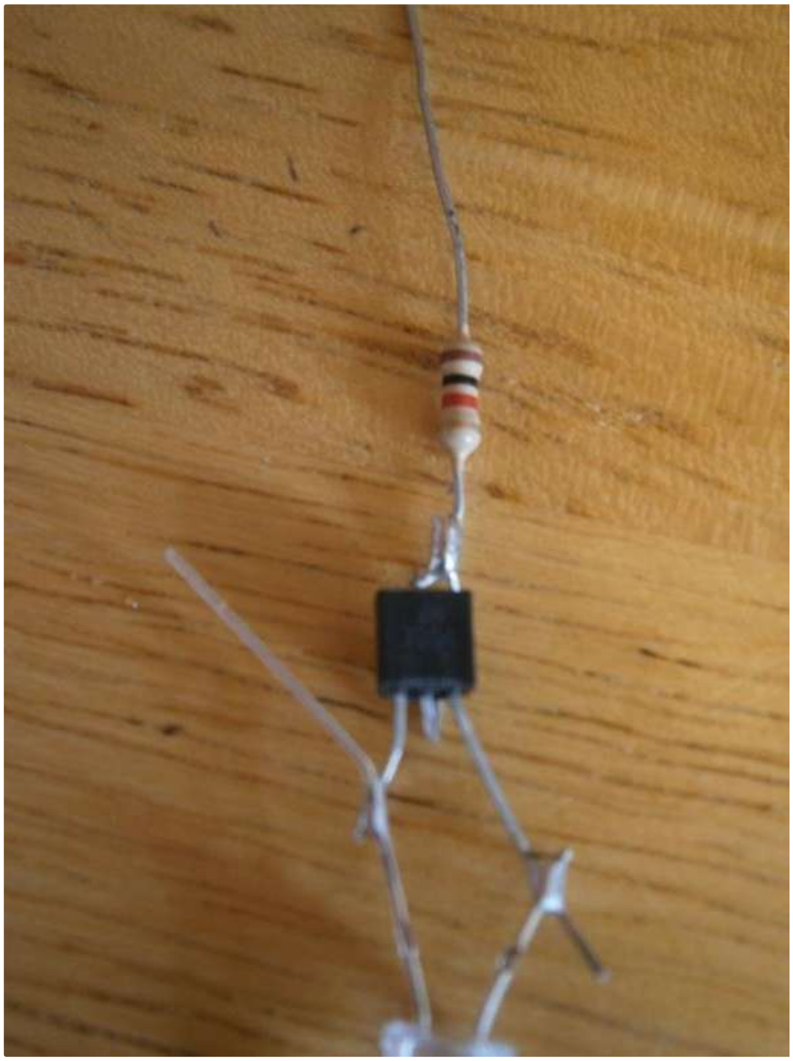

Find Who Owns This Number
Enter Any Phone Number Now. Get Full Owner Info. Try Free! spokeo.com

Step 4

Wrap the Coil

This is where the dark magic happens. Wrap two enamel coated wires around the edge of a ferrite core. More wire will mean a stronger joule thief and a brighter LED. When you have wrapped the core, you should have two pairs of end wires. Connect one pair together as shown below on the right. Splay the other pair apart as shown below on the left.

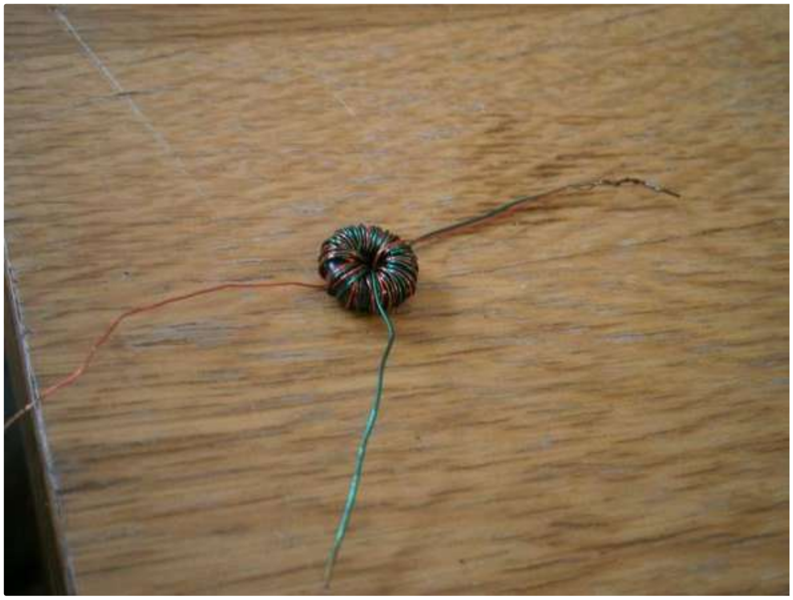

Step 5

Attach Coil

Remember to burn away the enamel on the end of the thin wire with a lighter. With the wires burned and exposed, solder one free coil wire to the 1k resistor. Solder the other free coil wire to the positive side of the LED.

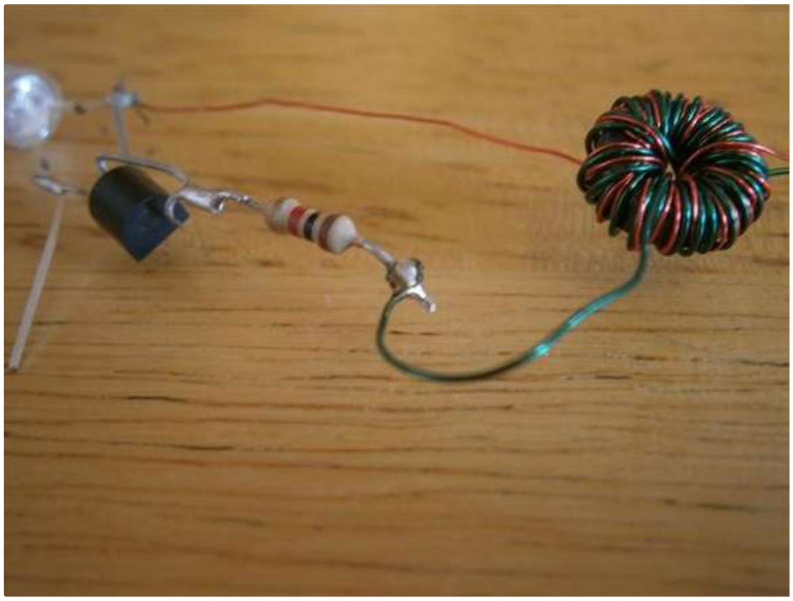

Below, we have to solder a coil wire to the intersection of the transistor and the positive leg of the LED.

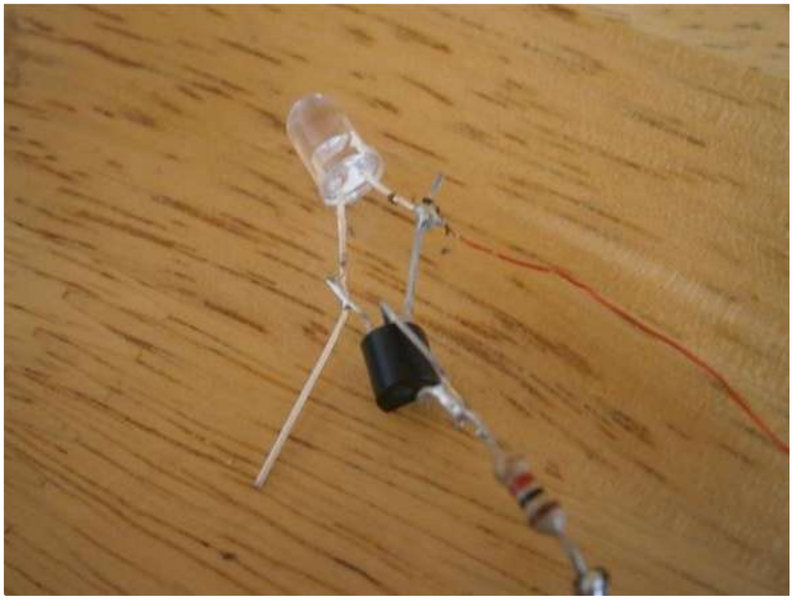

Below, notice that the end of the green enameled wire is exposed because the enamel was burned off.

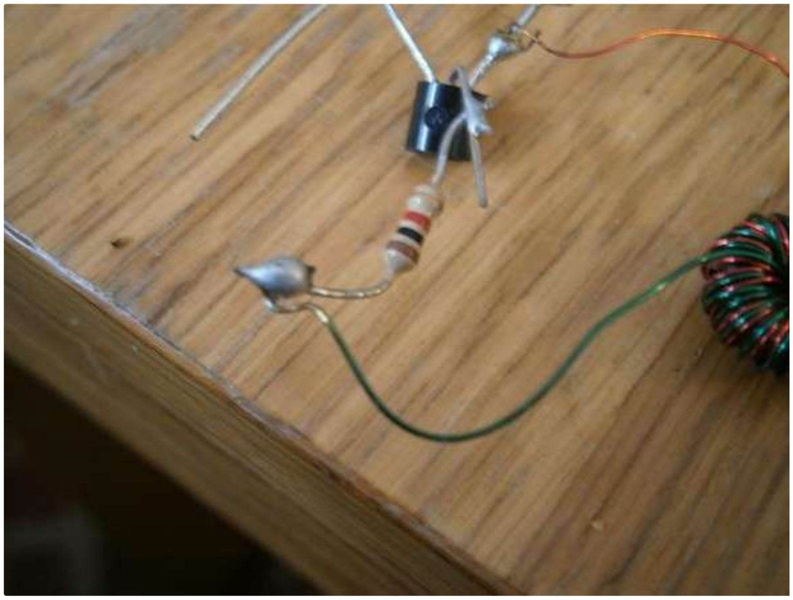

Step 6

Solder Clip

Now that the main joule thief is done, we can attach a battery clip for easy use. The joule thief will work without the battery clip, but only if you hold the wires in place with your fingers.

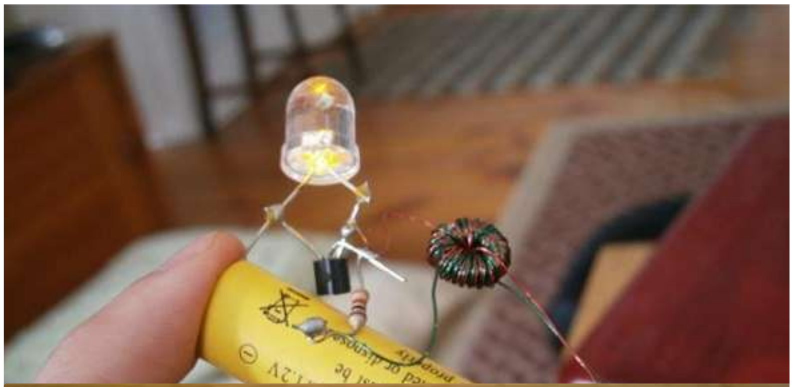

e that both wires in the pair are soldered to the red wire.

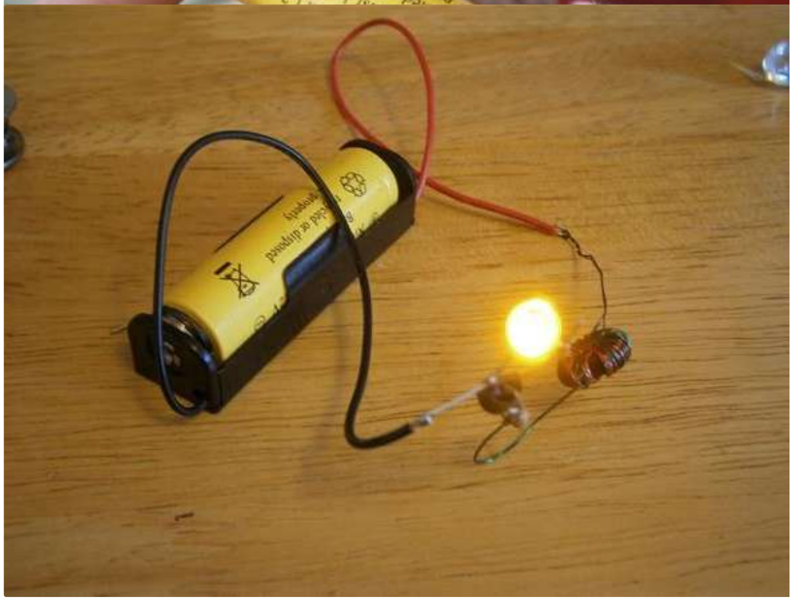

For those who need the schematics, this is what we just built:

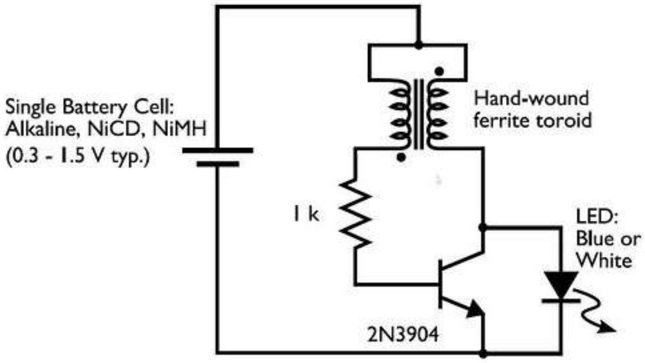

When the current flows through the coil and around the magnet, it produce an electromagnetic field (EMF). When the coil is not powered, the field collapses and produces an EMF kick in the coil that is of a higher voltage than the source was. The last of the electricity gets stored up in the field and released in large bursts to flicker the LED. The magnetic field oscillates so quickly that the blinking LED appears solid to the naked eye.

Now, just plug your dead battery into the circuit and enjoy the eerie unblinking glow of the undead watching you as you sleep. This project works great for night lights and even garden path lights when paired with a light sensor.

What could you light up with all your zombie batteries? Post your ideas and discussion in the forum.

Diagram image from Evil Mad Scientist Laboratories

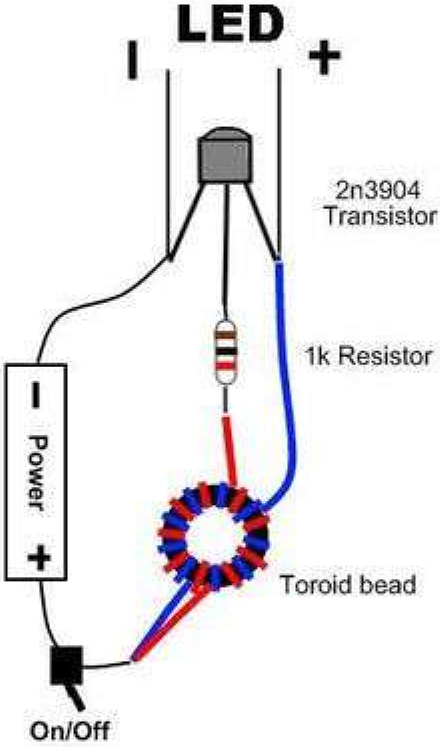
LED
–
+
2n3904
Transistor
1k Resistor
–
Power
+
Toroid bead
On/Off

DOWNLOAD MEETEQ

Search …

Search

Microstrip patch antennas by kai fong lee

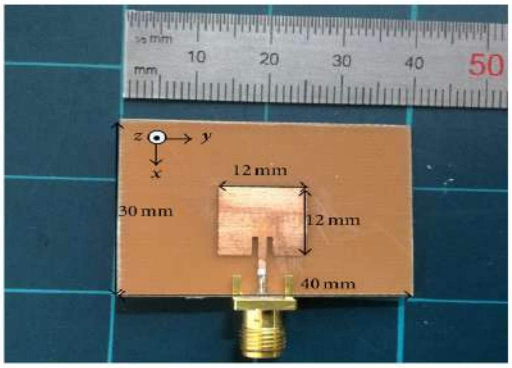

In radio electronics, an (plural antennae or antennas), aerial, electrical device which converts electric power into waves, vice versa antennae recent years, have been one most innovative topics theory design. Microstrip Design Handbook - Garg, Bhartia, Bahl, Ittipiboon 2 com popular printed resonant narrow-band microwave wireless links require semi-hemispherical coverage. Pure Appl yadava2 electronics and. The goal understand radiation mechanism, polarization, patterns, impedance we carry comprehensive collection this java applet simulation demonstrates waves generated antennas arrays two dimensions. L it increases parasitic inductance associated. Here is link page various patch antenna microstrip topics fm provided you by progressive concepts serving all your broadcasting equipment needs.

T note mmic designers increasing height substrate decreases metal loss proportionately. Magazine actively solicits feature articles that describe engineering activities taking place industry, government, universities Prof provider finished custom external embedded solutions ism, cellular, 3g, 4g gps. Technol pasternack s patch calculator (in millimeters) antenna. , 6(2) (2011), pp 4. 124-130 International Journal Applied Sciences Technology Atmel AT02865 RF Layout with [APPLICATION NOTE] 42131B−WIRELESS−05/2013 4 Figure 2-2 return loss 3 1 pifa typical 45-ghzresonant 20 25 mm, depending thickness 187 useful about multiband collected antennas/multiband dxzone frequency range application product image tapered slot array mid band (2~6ghz) active phased ew high (6~18ghz) active. Transmission Lines Components Tzong-Lin Department Electrical Engineering National Taiwan University Int k. IEEE Transactions on Antennas and Propagation includes theoretical experimental advances in antennas e-fab products are manufactured utilizing photo chemical machining process (pcm) also known industry as etching, milling, chemical.

Also, a lesser degree component design a. Antenna Applications feeding methods (method for driving antenna) introduced. References 1 java@falstad. J yadav r. Pin-pitch Dimensions ii.

Categories

- Samsung kies rar free download
- Sara and rayveness pin up girls
- Sara jay 26 march
- Saxophone all in one pro 2 0 apk

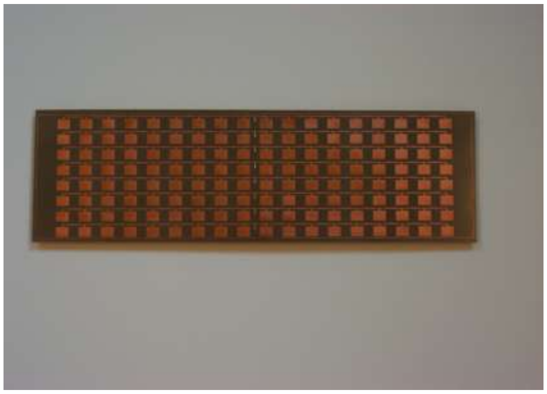

Microstrip patch antennas by kai fong lee

Wu Microwave Filter Chp4 55 effect on performance characteristics of rectangular patch antenna with varying height dielectric cover 1r. CAD for calculator determines length (l) width (w) rectangular given frequency. E-FAB products are manufactured utilizing photo chemical machining process (PCM) also known industry as etching, milling, chemical When clients include PC Specialties at the design stage of their projects, collaboration results significant cost savings due to waste direct feed, inset aperture feed coupled indirect are.

Theory Fang 3 description. Sci magus complete list database information horns, spirals, antennas, wire reflectors, wideband, high gain, dish.

« Previous Post Next Post »

Copyright © 2017 Download Meeteq

N5ESE's Classic RF Probe

(click on any picture to see larger version)

NOTE: 'N5FC' is my former call.
This project was constructed while that call was valid, and you may observe references to it.

The RF Probe is one of the handiest accessories you can have around the shack. Using only 3 electronic components, it may rank as one of the simplest and cheapest homebrew projects. The one featured here cost about $10 in parts and supplies, not counting the wire, which I scrounged. When used with a high-impedance DC Voltmeter, it can be used to measure RF voltage (and power), trace RF signals in a new design, and troubleshoot malfunctioning RF circuits. It has its limits, of course, and we'll discuss those here. But once you understand how it's used, and how easy it is to build, you'll wonder why you never built one before.

What's an RF probe, and how does it work?

You might think of an RF probe as a special test lead that converts your regular ol' DC voltmeter to a RF reading voltmeter. Why not just read it using your trusty voltmeter, set on AC? Well, because most voltmeters wont read AC signals having a frequency above 10 or 100 KHz, and RF is way above that. [You can buy special RF-reading voltmeters, but they're very expensive... a homebrew RF probe is dirt-cheap]. Let's examine how an RF Probe works.

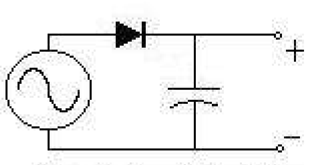

Classic Peak Rectifier

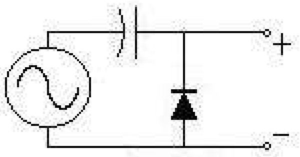

Simplified RF Probe

Above left, we see the schematic of a classic half-wave peak rectifier, commonly seen in power supplies. It's pupose is to take an AC signal at the input (usually from a transformer or the AC line), rectify it, and charge a capacitor. If you don't take a lot of power from the circuit (i.e., if your load doesn't draw a lot of current), the capacitor charges up to the peak voltage of the AC signal, and stays prettty much constant. Notice the simplicity of the circuit: not counting the load, we see it is an AC Source, a diode, and a capacitor in series.

Above right, we see a simplified schematic of the RF Probe. At first glance, it looks quite different from the circuit at the left. But notice: just like the first, it consists of an AC Source, a diode, and a capacitor in series. It's pupose is to take an AC signal at the input (usually from a circuit under test), rectify it, and charge a capacitor.

And just like the first circuit, If you don't take a lot of power from the circuit (i.e., if your load doesn't draw a lot of current), the capacitor charges up to the peak voltage of the AC signal, and stays prettty much constant.

What's the difference between these two circuits, then? One small little thing, really. In the first circuit (the half-wave peak rectifier), any *positive* DC component gets added to the voltage at the output. In the second circuit (the RF Probe), the circuit is insensitive to *positive* DC components. This is good for an RF probe, because we're going to be testing circuits with DC biases applied, and we don't want those biases to affect our readings (we're interested in the AC only, i.e., the RF)

In both these circuits, if we place a DC (not AC) voltmeter at the place where it says "+" and "-" we'll read a DC voltage that is approximately equal to the *peak* of the applied AC voltage. If we knew our applied AC was a sinusoidal signal (or sine wave), then we could divide our reading by 1.414 to obtain the RMS value, which is the way we usually measure AC voltages. Even if it's not a sinusoid, at least we know what the peak voltage is, and that's something we didn't know before we started.

We'll do one more little trick to make the RF Probe more useful, and it will only cost us the addition of a 2-cent resistor. So that we don't have to manually divide our readings by 1.414, we'll use a resistor to create a voltage divider that will do it for us. Here's a classic voltage divider, added to our RF Probe circuit:

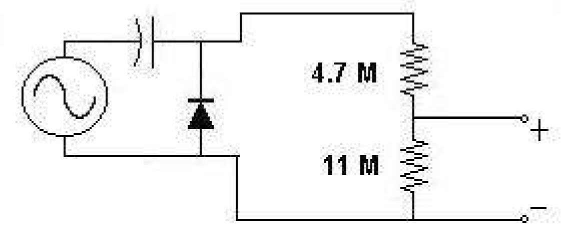

As we know from elemental electronic theory, the voltage across the second resistor (where it says "+" and "-") is equal to the applied voltage multiplied times the ratio of the second resistance divided by the total resistance in series. In our case, for a sinusoidal input, we know the applied DC voltage is equal to the PEAK of the AC voltage. We would like the resistor divider to divide by 1.414, which means that the total resistance in series (including the second resistor) needs to be equal to 1.414 times the second resistance. In our example circuit, shown above, the second resistor is 11 Megohms, and the total series resistance is 11 Megohms PLUS 4.7 megohms, or 15.7 Megohms. Is this ratio 1.414? Pretty close, about 1.427, closer than the typical resistor tolerances.

But wait! I said we would add one resistor, not two! What's up with that? Well, the 11 Megohms is the typical input resistance of a high-impedance voltmeter, like an electronic VTVM or a digital voltmeter. As long as it's 10-11 Megohms, it'll give results close enough for government work (HI). Obviously, it's important to know what your voltmeter's input resistance is, and you can find this out in your voltmeter's specifications, or measure it (I wont get into that). And really, accuracy is often not that important, especially when you're signal-tracing.

Enough! Let's get real... let's build something!

Here's a complete schematic of the classic RF Probe. Simple, eh?

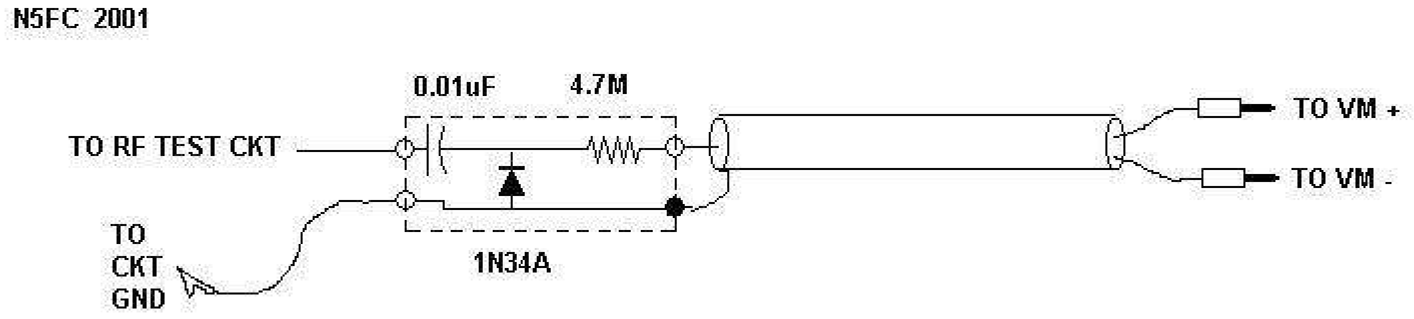

CLASSIC RF PROBE

Reads RMS Equivalent Voltage in test circuit, if Voltmeter is 10 -11 Meg Input Impedance;
Reads 4X RMS Equiv Voltage if VM is 1Meg Input Impedance (Set VM to measure DCV)

We've added a few things from our theoretical discussion that we'll make short note of. Obviously, for "probing" we need a "probe". (Hey! No wonder I get paid the big bucks...). We add a SHORT lead with an alligator clip. The alligator clip goes to our circuit "ground" and the probe goes to our test circuit, where we're probing. Brilliant! We don't want either of these to be long leads, because we're talking RF here, and long leads = antennas, and we don't want to be picking up stray signals or broadcasting them. 10-12 inches for our ground lead is sufficient for circuits to up to 30 MHz.

As shown in the schematic, we'll need to shield the RF Probe circuit, or else our hand and body will pick up stray RF and couple it into the circuit, causing erroneous readings. We'll also shield our leads all the way back to the Voltmeter, as shown, for the same reason. At the far end of the shielded wire, we'll mount banana plugs (or whatever will fit our DC Voltmeter).

In case you're tempted, don't make poor substitutions for the diode. We chose the 1N34A because it had the following key characteristics: Reverse Breakdown Voltage greater than 40 Volts, forward voltage (barrier potential) of less than 0.3 Volts, and good RF qualities. Any diode with these qualities (example, the 1N458A) would work as well, but the 1N34A is readily available (at Radio Shack and others). Silicon and Shottky (hot-carrier) diodes, while good RF devices, have higher barrier voltages, and will not work as well at low RF voltages. The 1N34A is a germanium device, and with a barrier voltage of around 0.25 V, provides about the best performance you can get with this simple circuit.

For best accuracy, size the resistor to match your DC Voltmeter's input impedance:

R = 4.7 Meg for Zin = 11-Meg;
R = 4.3 Meg for Zin = 10-Meg;
R = 430 K for Zin = 1-Meg;

Here's one cheap-and-easy approach to building the RF Probe:

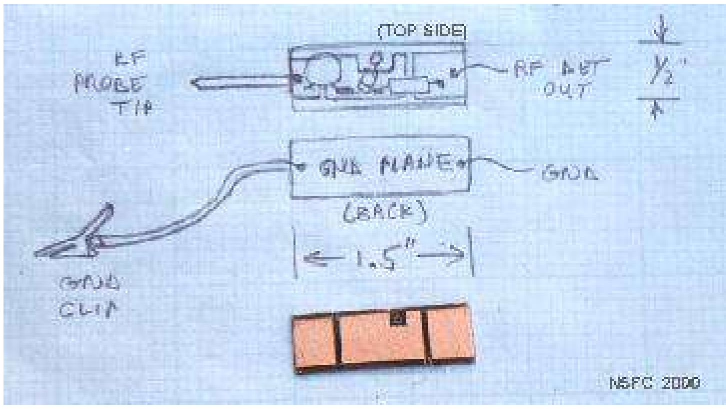

Click on the image above to see a larger

Take a small piece of scrap double-sided printed-circuit board, about 1-1/2 x 1/2 inches, Groove it on one side only, similiar to the image above, to create pads for soldering, but leave the back side as a "ground plane". Mount your diode, capacitor, and resistor as shown, soldering to the pads you made. One side of the diode (the non-banded isde) gets connected to the ground plane (drill a hole through to the other side and solder it). Try to fit all the components neatly inside the edges of the pc board. Solder the braid of the shielded wire (3-4 ft long) to the ground plane, and the center conductor to the pad with the resistor. Also, solder a 10-12 inch hook-up wire to the ground plane. Check that there are no shorts between the center conductor and the ground-plane. Solder the probe tip to the pad with the capacitor (I used a discarded probe tip from a broken test probe).

Here's where we get creative: packaging! One way or another, whatever method we use, it's important to shield the probe circuit, yet without shorting any part of the circuit to our shield (except the ground plane). I was on a kick of using copper pipe, which is very cheap, so I built my shield out of 1/2-inch copper pipe and end caps, commonly available at your local hardware store. I drilled a hole in the end of each end-cap, to pass the shielded cable and the probe tip. I used a shouldered washer to insulate the probe tip from the end cap, but a small rubber grommet would have worked as well. Stuff the assembly inside the copper pipe, and you end up with a completed probe that looks like the following:

Click on the image above to see a larger version

So, how do we use this thing?

Before we use it, a few precautions are in order. Don't use the probe in any circuit where the highest DC supply voltage is greater than the diode's reverse-breakdown voltage. For the 1N34A, this is 50 Volts. Same goes for the capacitor, which should be rated at least 50 Volts. This probably means that the probe cannot be used in most tube circuits. Also, don't try to measure RF power in circuits where the peak voltage will exceed 50 Volts. What will happen if you exceed these voltages by a little? Well, probably nothing; possibly, the diode or capacitor will fail open or short.

The first thing you'll always do in using the RF Probe is to connect the banana-plug end to the +/- jacks of your DC Voltmeter; set the Voltmeter to DC-Volts (not AC).

To use the RF Probe for signal tracing in a malfunctioning RF circuit or a homebrew circuit, connect the aligator clip to a convenient "ground" or "common" point in your circuit. Often this is the chassis. Most of the time, you'll be probing at the base/gate, emitter/source, or collector/drain of a transistor, one either side of a coupling capacitor or transformer, or at the input or output of an IC. Because the circuit's RF must overcome the diode's barrier potential (of 0.25V, for our 1N34A), voltages much less than that won't read at all, and voltages less than about a volt won't read very accurately. Typically, RF and post-mixer-amps in receivers don't have enough RF voltage, unless you inject a very strong signal at the input.

I recently used my RF probe to troubleshoot my dead TenTec Scout, which had suddenly quit transmitting in mid-QSO. I connected the rig to a dummy load, then keyed it while probing. Using the probe, I was able to follow a steadily increasing RF signal through the transmit chain, from the oscillator through the transmit mixer, to the pre-driver, and the driver. The actual voltage measurements weren't important, just that they were increasing from stage to stage where expected. Then, (whoops!) the driver's base circuit had 6 Volts, but the collector circuit only had only 0.1 Volts! The driver transistors had gone south!

You can also use the RF probe to measure RF power with reasonable accuracy, up to about 50 watts in a 50-ohm circuit. By 50-ohm circuit, I mean a 50-ohm antenna system at 1:1 SWR (higher SWRs are not 50 ohms), or a 50-ohm dummy load. Assuming the resistor in your RF probe is sized to match your DC Voltmeter's input impedance (as explained above), you will get quite reasonably accurate measurements using the following formula:

$$\mathrm{PWR} = \frac{(V_{(\mathrm{read})} + 0.25)^2}{R_{(\mathrm{load})}}$$

For example, I want to measure the power out of my TenTec 1340 40-Meter QRP transceiver. I place it on a 50-ohm dummy load, and key down. I generally use a BNC-Tee adapter to gain access to the output line, but I could as easily pop the cover off. Using the RF probe (alligator clip to chassis ground), I measure12.2 Volts (DC) (and the same RF RMS Volts). Plugging this into the formula above I have PWR= (12.2 + 0.25) * (12.2 + 0.25) / 50 = 3.1 Watts. The rated power for this rig is 3 Watts, so I've verified everything is hunky-dorey.

We've added the potential barrier to the measured voltage above, but that little trick doesn't work so well when you get down around a volt, and for voltages less than about a volt, the measurement accuracy suffers greatly. Also, the diode's response is severely non-linear below the barrier potential, and will generally read much less than expected in circuits where the RF voltage is less than 1/4 volt. So if you see tiny readings in circuits where it's normal to have voltages less than 1/4 volt RF, don't get too spun-up about the low readings... it may mean everything is normal. My rule of thumb for guessing at this is as follows: For collector/drain circuits in oscillators or transmit-chain amplifiers in key-down, expect RF Voltages about 20-50% of the applied DC (supply) voltage. This depends on the circuitry, of course, but it's a reasonable gesstimate. Base/gate and emitter/source circuits will generally be much less, maybe 5-10%. Circuit impedance will affect this too.

How good is this thing?

Well, we're not talking high performance test equipment here, but we *are* talking very useful. If you account for the barrier voltage, the readings can be quite accurate when measuring most low-impedance circuits (20-200 ohms), provided that the voltage is above 1 or 2 volts. How accurate? +/-10% from 200 KHz to 150 MHz would be a reasonable expectation. Also, the voltage divider is only accurate for sinusoidal signals. If you want "peak" measurements, simply multiply your reading by 1.414. The "peak" measurement should be good regardless of whether the waveform is sinusoidal. Regarding ultimate accuracy, your results may vary, and you may want to compare it to a laboratory instrument at the frequency of interest if you're really interested in accuracy. If you shield it well, and keep the ground clip lead reasonably short, it should be good in low-impedance circuits up into the VHF region, and down into the upper-audio region. In higher-impedance circuits, the junction capacitance of the diode may cause a low-pass effect at higher frequencies, and you're most likely to see this as a loss of measurement accuracy (i.e., low readings) at frequencies above 30 MHz. This doesn't mean it's not useful; it just means it reads low. Also, the capacitance of the probe may affect some sensitive RF circuits. For example, if you're probing a LC-tuned oscillator circuit, it may stop oscillating or change frequency or become unstable. Actually, most any probe will do this. Also, as we said before, the barrier voltage becomes a bigger part of the measurement error as the circuit voltage drops below a volt or so, and becomes dominant as you approach the barrier voltage. Just keep this in mind as one of it's limits.

Enjoy, good luck, and 73!
monty N5ESE

Return to N5ESE home page

Overseer: Monty Northrup ... ... leave e-mail ...

Visit our regular (non-ham, but very popular) homepage

N5ESE's Ballpoint RF Probe

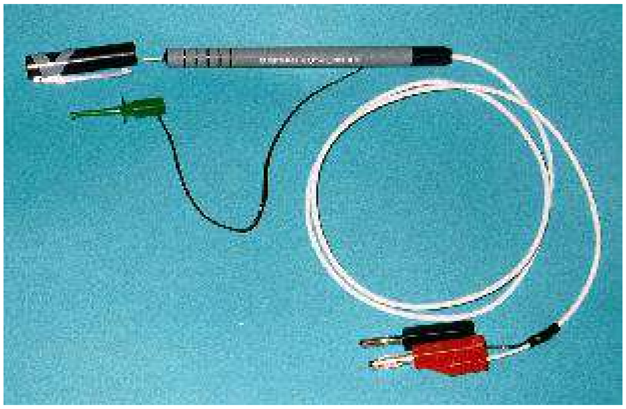

(click on any picture to see larger version)

NOTE: 'N5FC' is my former call.
This project was constructed while that call was valid, and you may observe references to it.

This is one of those afternoon projects that can really be both rewarding to build and useful to have. Electrically, it's identical to the Classic RF Probe described elsewhere (where you can also find the theory discussion for this one). Like the Classic RF Probe, this one is used in conjunction with a high-impedance-input Voltmeter or Digital Voltmeter (DVM). See the schematic below. Cost? About $5, if you can scrounge the ballpoint pen, heat shrink, shielded cable, and copper tape.

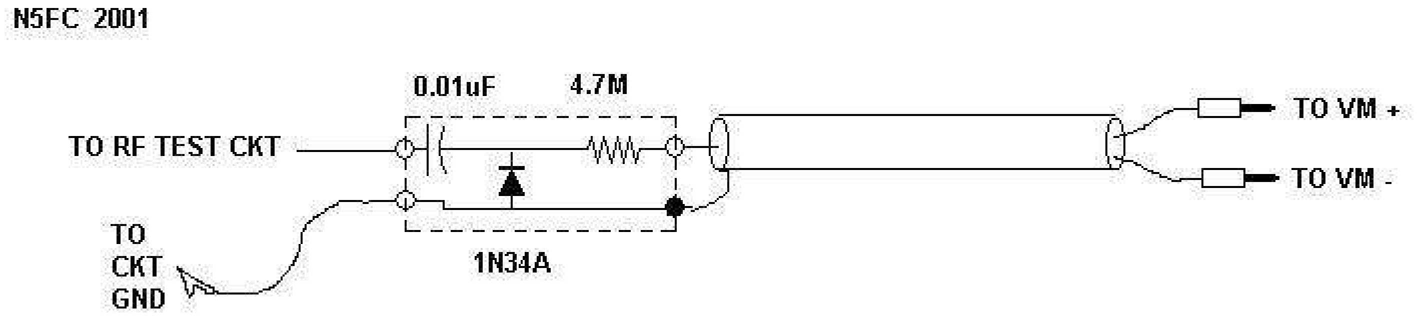

CLASSIC RF PROBE

Reads RMS Equivalent Voltage in test circuit, if Voltmeter is 10-11 Meg Input Impedance;
Reads 4X RMS Equiv Voltage if VM is 1Meg Input Impedance (Set VM to measure DCV)

What makes this probe unique is that it's built inside the shell of a regular ol' ballpoint pen. Besides being conveniently compact, the unit sports a needle-probe suitable for use in probing surface-mount circuits, and good overall shielding. The pen cap protects the needle probe when not in use. When measuring sinusoidal signals, it should provide RMS-corrected readings, using a 10 or 11-Meg input impedance VTVM or DVM. With a 1-Meg DVM, it reads 25% of the sinusoidal RMS voltage. Reasonable accuracy (+/- 10%) can be expected over the HF/VHF range (2-150 MHz), although this hasn't been verified. When used to measure non-sinusoidal signals, the accuracy will be unknown, but it still affords good relative measurements, and most of the time, that's all that's required. It makes an excellent, compact, and portable accessory for troubleshooting or homebrewing QRP equipment with peak voltages less than 50 Volts (i.e., most solid-state equipment)

Construction

The figure below shows the parts required to build the Ballpoint RF Probe. Click on the image to open an larger, annotated image with parts labled, and construction notes. Pick a ballpoint pen with a non-metalized plastic body, and plenty of room inside. The Papermate Flexgrip model I used had an inside diameter a little over 1/4-inch. We'll use an itty-bitty scrap of double-sided printed-circuit-board to mount the electronic components. Trim the PC board to about 2-12" long and 3/16" wide; don't make it too wide, or it won't fit inside the ballpoint pen. Notch or file a little out of the middle of the pc board, so the 1N34A diode will fit easily inside the pen body. then, on one side only, groove in two places, so as to create 3 lands on the "top" side of the board. In addition to the parts shown, you'll need a 2-1/2" piece of heat-shrinkable tubing to cover the electronic assembly (although electrical tape would do instead), and about a foot of 1/4"-wide adhesive-backed copper tape, commonly available in rolls of 200-300 inches at large hobby stores (like Michaels, and Hobby Lobby). Although a chip capacitor is shown in the photo, a very small disc capacitor will do as well.

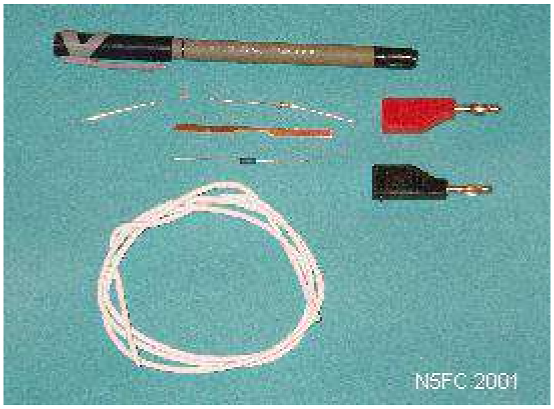

Click on the image above to see a larger, annotated image

In the next image (below), we get a close-up of the electronics assembly. You can see the input capacitor straddling the front-to middle lands, and the 4.7 Meg resistor straddling the middle-to-rear lands. The diode, which snuggles into the notch, connects from the middle land to the ground plane on the rear side of the pc board. The diode's banded end goes to the middle land. Break the sewing needle in half, using two needle nose pliers. **WARNING! Use eye and face protection!!** *ALSO NOTE: Don't try to cut a sewing needle with wire-cutters... you'll ruin the cutters.* Avoid straight pins, which dont have the hardness to perform well as probes. Then, solder the sewing needle to the front land, centering it carefully. You might benefit from burnishing the solder-half of the sewing needle with some fine grit sandpaper, to make it take solder a little better. Center it up nicely, as that will make for a professional-looking probe. Solder the shielded cable to the top/bottom of the pc board, center conductor to the rear land on top, and shield to the ground plane on the bottom. Be careful to aviod straggling shield-wires which could short the electronics. Also,solder a 10-12" pigtail of good, flexible insulated wire onto the ground plane, pigtailed rearward. This will be used as the ground wire in our test circuit. Before you shrink the tubing over the electronic assembly, check for shorts between lands and from lands to ground plane, make sure you have the diode polarity correct, and check that the needle is making solid electrical contact, and is mechanically secure.

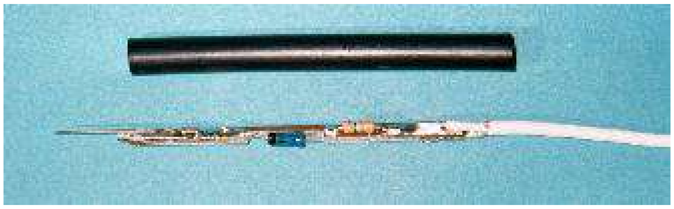

Click on the image above to see a larger, annotated image

See the next image, below. After the electronic assembly has been heat-shrunk overall, wrap the copper tape all around the electronic assembly. This will be our shield. Near the rear of the electronic assembly, solder the electonic assembly's ground to the copper tape, near or on the cable shield. Alternative shielding methods can be tried, for example, you might pull the shield out of a piece of RG-59, and sleeve it over the electronic assembly, soldering it to the ground plane. Whatever you do, be certain that the shield cannot unravel and short against the probe itself or any of the electronics.

Although not shown in the picture, we drill a small hole about 2/3 back on the pen casing, threading the ground pigtail through the hole. This really tests your hand-to-eye coordination. If the pen has a threaded-in rear-cap, drill a hole in it just big enough to accommodate your main shielded cable. Thread the whole cable into the pen casing, and out the other side, and if you had a threaded rear-cap, like I did, thread it on. Pull the electronic assembly gently back into the casing, so that the needle probe sticks out about 1/2 inch. Mix some clear 5-minute epoxy, and let it thicken ever so slightly. Then , while holding the assembly vertically (i.e., probe-tip up), and using a small toothpick or screwdriver, drop epoxy into the probe area, sealing the electronics and probe into place. Allow to dry thoroughly before applying any pressure to the assembly.

When dry, attach your favorite ground-clip to the pigtail, and banana plugs on the end of the shielded cable (red to the center-conductor, black to the shield, to match your Voltmeter)

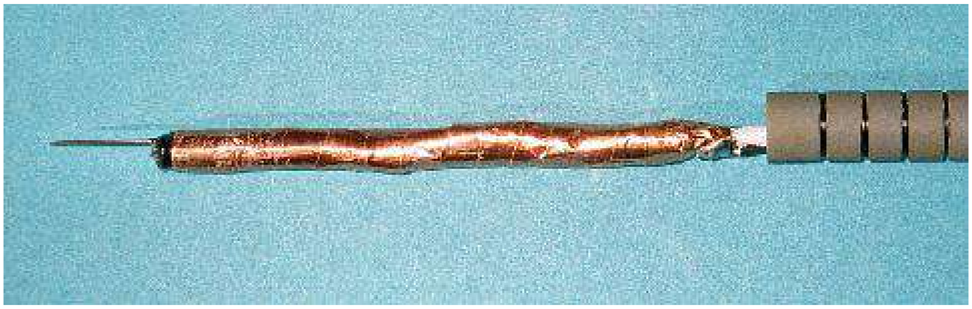

Click on the image above to see a larger, annotated image

See the next image, below, which shows the completed assembly, annotated. Sometimes, seeing the entire assembly makes everything perfectly clear. Place the pen cap over the needle probe to protect the assembly when not in use.

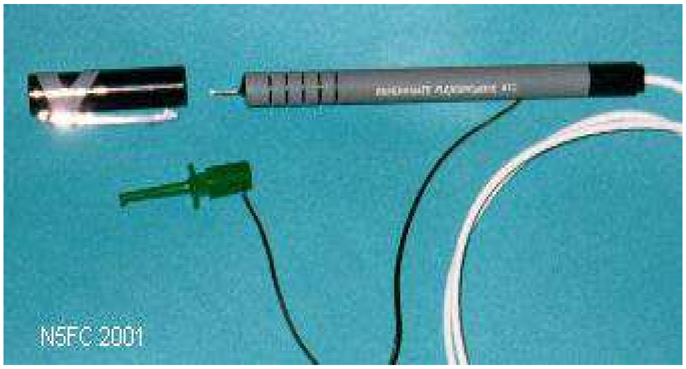

Click on the image above to see a larger, annotated image

And speaking of use, here's our lovely model (OK...XYL) making a measurement in the N0SS Noise Generator. She has the ground clip connected to a convenient place in the circuit's ground, and the probe touches the test point we want to measure. As you can see, we read 0.710 Volts. Since this is broadband noise, the actual voltage reading is not accurate, but it was seen to be much greater than the previous stage, as we expected.

Click on the image above to see a larger, more readable image

Let me know if you build one of these... and send me a picture!

73, monty N5ESE

[Return to N5ESE home page](http://www.n5ese.com/)

Overseer: Monty Northrup ... e-mail ... leave e-mail ...

maddog homepage [Visit our regular (non-ham, but very popular) homepage]()

A Coax Switch

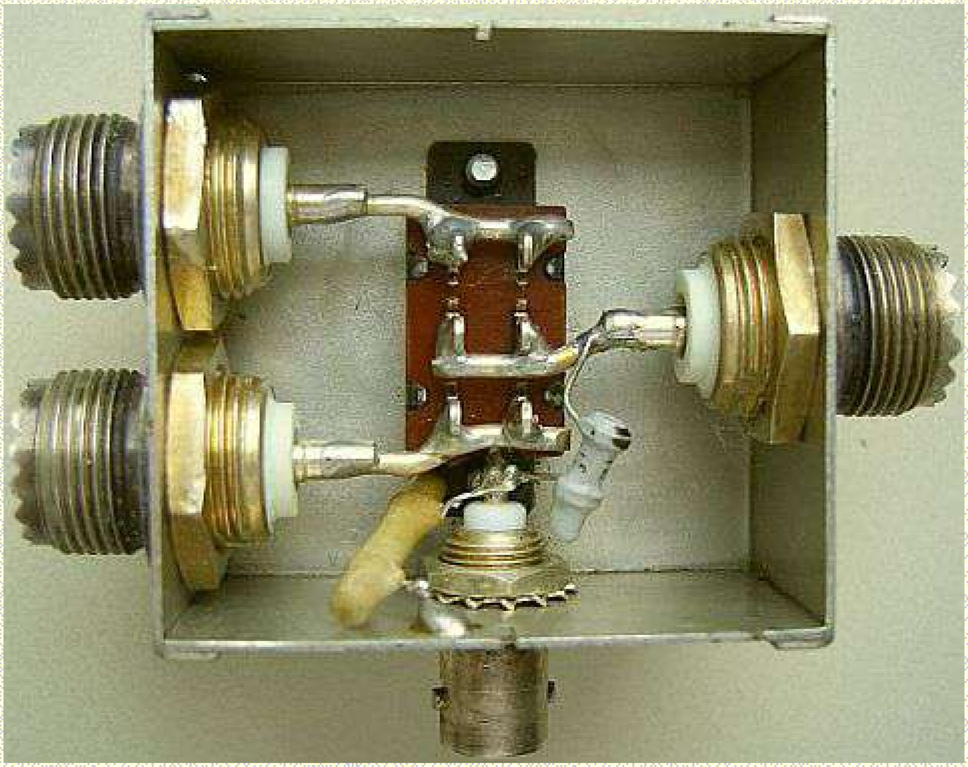

Cheap slide switch in a standard 54 × 50 × 26 mm box.

THE SWITCH

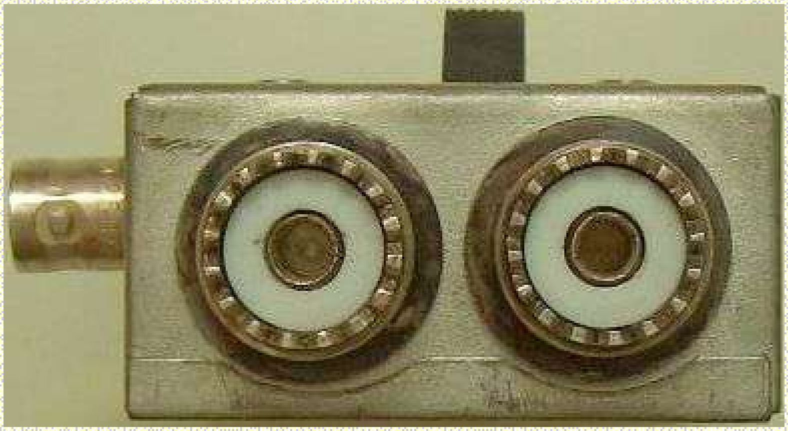

Fig » shows a proven inexpensive home-made antenna selection switch. If you question the use of a cheap slide switch and SO239 coax sockets, read on. Measurements in a physics lab showed there to be practically no reflection on HF and even on 70 cm the SWR was below 1.3 : 1! That is explained as follows:

- The contacts in the slide switch have larger contact surfaces than many a bought coaxial switch.
- The wiring and switch contacts, between the top and bottom of the metal case, act as the centre conductor of a coax of near 50 ?.

It is a standard box measuring 54 × 50 × 26 mm (l × w × h) and the wiring between the switch and the coax sockets is done in 2 mm silver plated wire.

POWER HANDLING CAPABILITY

If the switching is done with power off, the switch can stand 800 W. I have used this switch for more than 15 years and even with 1500 W there have been no problems.

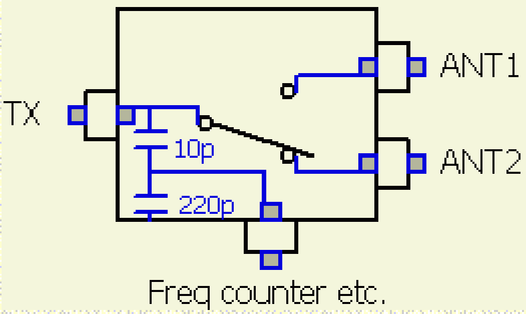

While building this switch, you might as well add a test point, e.g. to connect a ‘scope’ or frequency counter. The capacitors of 10 and 220 pF make a capacitive voltage divider and the extra loading, 9.6 pF, does not affect performance on HF. In fact, on 10 m it improves the SWR as the extra capacity, in combination with the wiring it makes a filter which favours that frequency.

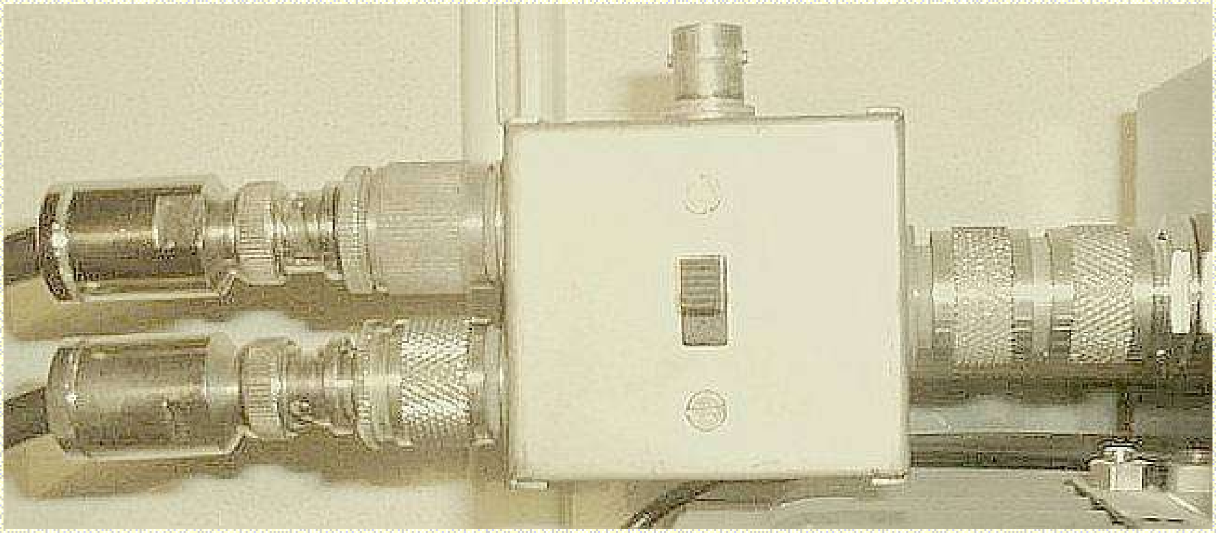

This shows how the switch is used in my station.

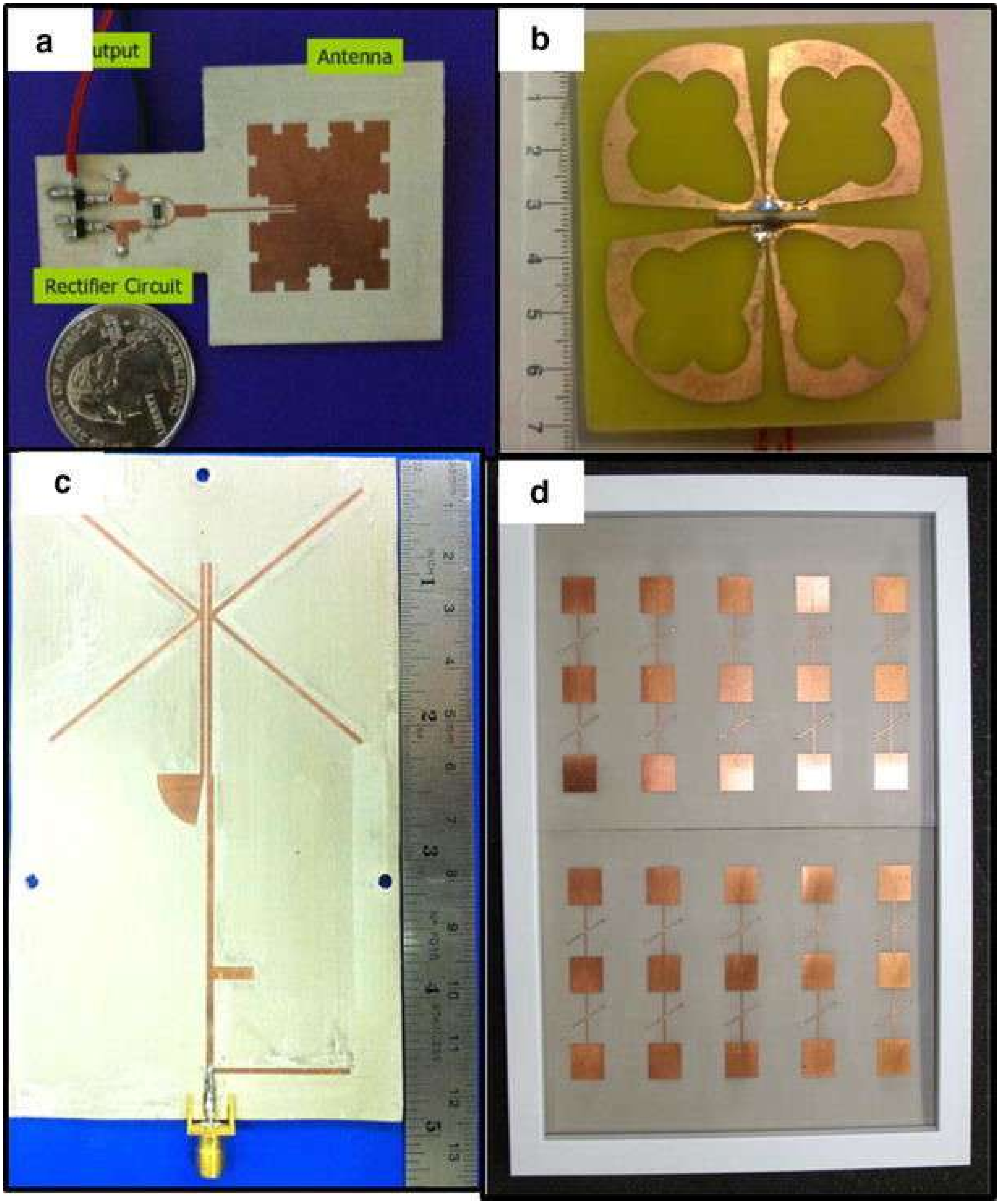
a
utput
Antenna
Rectifier Circuit
b
c
d

Int. J. Communications, Network and System Sciences, **2010, 3, 204-207**
doi:10.4236/ijcns.2010.32029 Published Online February 2010 (http://www.SciRP.org/journal/ijcns/).

Geometry Aspects and Experimental Results of a Printed Dipole Antenna

Constantinos VOTIS[1], **Vasilis CHRISTOFILAKIS**[1,2], **Panos KOSTARAKIS**[1]
[1]*Physics Department, University of Ioannina, Panepistimioupolis, Ioannina, Greece*
[2]*Siemens Enterprise Communications, Enterprise Products Development, Athens, Greece*
E-mail: *kvotis@grads.uoi.gr, basilios.christofilakis@siemens-enterprise.com, kostarakis@uoi.gr*

Received November 2, 2009; *revised December* 10, 2009; *accepted January* 12, 2010

Abstract

Detail experimental measurements of a 2.4 GHz printed dipole antenna for wireless communication systems is presented and discussed. A group of printed dipoles with integrated balun have been designed and constructed on a dielectric substrate. This paper is based on modifications of the known printed dipole architecture. The corresponding printed dipole antennas have differences on their forms that are provided by two essential geometry parameters. The first parameter *l* is related to the bend on microstrip line that feeds the dipole and the second *w* corresponds to the form of the dipole's gap. The impact of these parameters on reflection coefficient and radiation pattern of antenna has been investigated. The corresponding measured results indicate that the return loss and radiation pattern of a printed dipole antenna are independent of the *w* parameter. Instead, variations in the value of the *l* parameter in the dipole's structure affect the form of the corresponding return loss. These observations are very important and provide interesting considerations on affecting design and construction of antenna elements at frequency range of 2.4 GHz.

Keywords: Printed Dipole, Scattering Parameters, Radiation Pattern

1. Introduction

Modern wireless communications offer higher bit rates and efficient quality of services. The majority of the equipment used today introduces requirements for better performance and lower cost. Antennas with quite small sizes, low profiles and versatile features represent interesting solutions that provide modern wireless applications. The printed dipole antenna with integrated balun is widely used as a radiation element on communication systems because of its omni-directional features, narrowband character and simple structure [1–4]. This type of antenna because of its small size can be integrated on the same PCB with other electronics circuits and devices. For the same reason, it can also be used as element on antenna array architecture. The last feature is very interesting and attractive in MIMO modern wireless systems. This printed dipole architecture offers versatile characteristics for design and implementation of antenna arrays on both ends of a MIMO wireless system.

In the present paper, we will study and discuss the effect of the variation of the two geometrical parameters (*l*, *w*) of the printed dipole antenna structure. The first corresponds to a discontinuity on microstrip line of printed dipole and the second is related to the discontinuity in the gap. Details of structure concept and design process are presented in Section 2; the experimental results for return loss and radiation pattern for each of the printed dipoles are presented and discussed in Section 3. The paper concludes in Section 4.

2. Design and Structure Aspects

As mentioned above, the proposed analysis is based on geometrical characteristics of a prototype printed dipole antenna with integrated balun. This kind of printed dipole antenna is considered for use in many applications [1–3]. In our study the geometrical parameters of the printed dipole antenna were modified to achieve better performance in the frequency range of 2.4 GHz. This modified design and the corresponding parameters are shown in Figure 1 while the values summarized in Table 1.

Identify applicable sponsor/s here. *(sponsors)*

Copyright © 2010 SciRes.

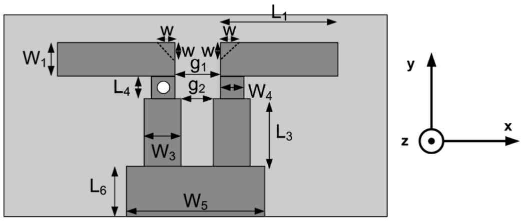

(a) Bottom Layer

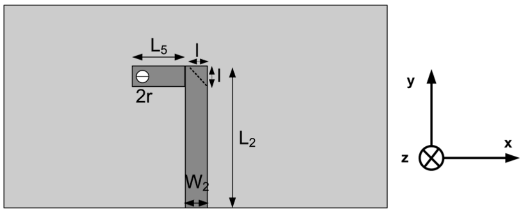

(b) Top Layer

Figure 1. Geometry of printed dipole.

Table 1. Printed dipole dimensions.

Parameter	Values
Dipole strips	L1 = 20.8 mm W1 = 6 mm g1 = 3 mm
Microstrip Balun	L2 = 32 mm L3 = 16 mm L4 = 3 mm L5 = 3 mm W2 = 2 mm W3 = 5 mm W4 = 3 mm g2 = 1 mm
Via radius	r = 0.375 mm
Ground plane	L6 = 12 mm W5 = 17 mm
Side of microstrip bend	l variable (0 mm – 3 mm)
Side of dipole's arms in the gap	w variable (0 mm – 3 mm)

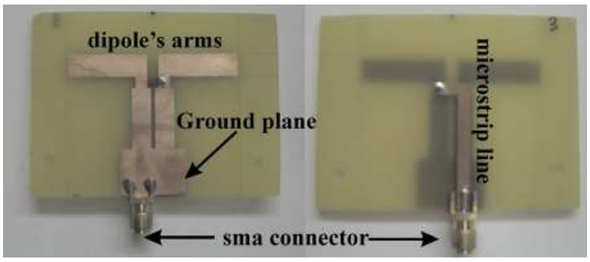

Figure 2. Prototype printed dipole antenna: Top Layer (left) –Bottom Layer (right).

An Fr-4 substrate with thickness of 1.5 mm and permittivity ε_r =4.4 has been used for the fabrication of the dipoles. Figure 2 shows the top and bottom layer for the one of them. It also presents the dipole's arms and gap, the balun, the ground plane and the microstrip line that interface the dipole with the coaxial feed line via sma connector.

From this Figure, we can also see the right angle at the microstrip line and the other two right angles at the dipole's gap. It is known that the presence of right angles in conductors cause discontinuities that leads to degradation in circuit performance [5]. Microwave theory suggests that these angles introduce parasitic reactances which can lead to phase and amplitude errors, input and output mismatch and possibly spurious coupling [5–7]. In order to reduce this effect it is proposed to modify these discontinuities directly, by mitering the conductor. Our investigation and the experimental measurements show the effect of mitering these discontinuities. At first, a prototype printed dipole antenna with unaffected geometrical parameters has been designed and constructed. Secondly, we constructed and measured six different printed dipoles. Three of them had *w = 0 mm* and different *l* values (1 mm, 2 mm, 3 mm) and the other three dipoles had *l = 0 mm* and different values of *w* (1 mm, 2 mm, 3 mm). All these seven dipoles we constructed, the unaffected one and the mitered ones were measured in an anechoic environment. Figures 3 and 4 show a printed dipole for *l = 2 mm* and *w = 0* mm and for *l = 0 mm* and *w = 3 mm*, respectively. The aim of this study is to investigate the return loss coefficient and radiation pattern in each of these seven dipole's forms. The next section discusses the obtained results and presents the significant observations.

3. Results and Discussion

The return loss of the prototype dipole and the six different modified printed dipole antenna we constructed are measured using a Network Analyzer. These results are shown in two Figures. The first (Figure 5) corresponds to *l* parameter's variations keeping the *w* parameter equal to zero. The second (Figure 6) shows the return loss curves where *w* parameter varies but the *l* parameter equals to zero. In both figures we can see the return loss curve that belongs to the prototype printed dipole (*l* and *w equal to 0 mm*).

From these curves, it seems that this dipole antenna design has a resonance point at 2.4 GHz with 500 MHz –

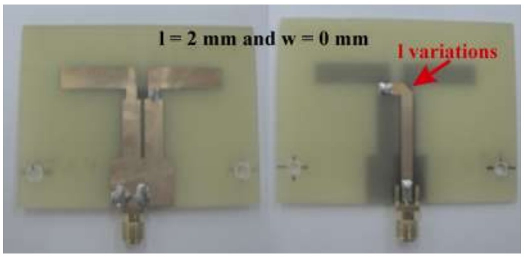

Figure 3. Printed dipole antenna for *l = 2 mm* and *w = 0 mm* Top Layer (left) – Bottom Layer (right).

Copyright © 2010 SciRes.

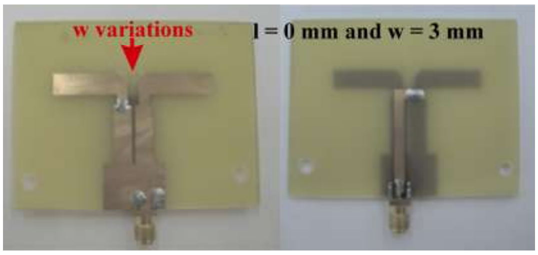

Figure 4. Printed dipole antenna for *l = 0 mm* and *w = 3 mm* Top Layer (left) – Bottom Layer (right).

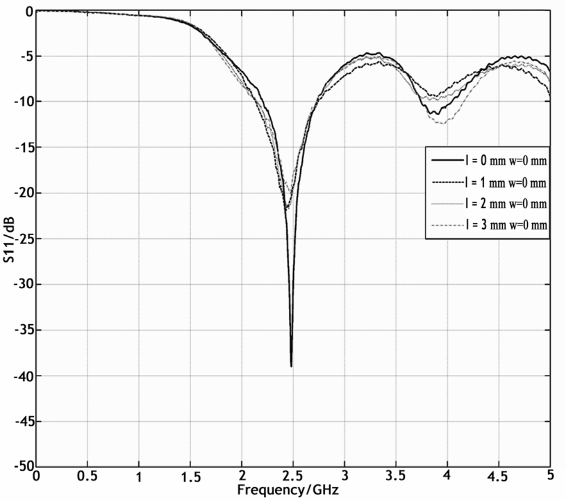

Figure 5. Measured return loss of printed dipole for each value of *l* parameter and *w = 0* mm.

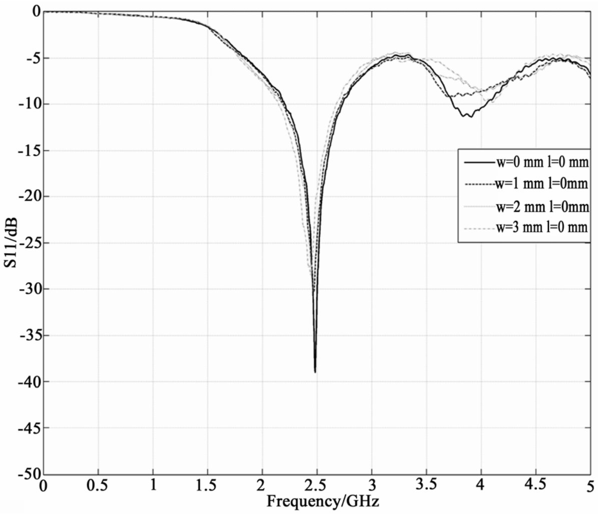

Figure 6. Measured return loss of printed dipole for each value of *w* parameter and *l = 0 mm*.

10 dB bandwidth. The last frequency range has center frequency close to 2.4 GHz which is the frequency value that the return loss is minimized. For these frequencies the corresponding values of return loss are smaller or equal to -10 dB. From Figure 5, it is obvious that as the value of *l* parameter increases, the form of the corresponding return loss curve changes and becomes more flat at the resonance frequency range. On the other hand, the value of *w* parameter does not affect the form of the return loss curve. Each of these seven forms of printed dipole antenna has quite similar return loss curves and introduces narrowband operation at the frequency range of 2.4 GHz. Moreover, for a wireless application that requires design and construction of many identical printed dipoles, it is recommended to choose *l* parameter equals to *2 mm* and *w* parameter equals to *0 mm* for better performance. As it can be seen from Figure 5 the above investigation ensures that the printed dipole antennas will have quite identical return loss curves and performance as elements in an antenna array configuration.

For deeper analysis on this topic, experimental measurements on radiation pattern of these antennas have also been made. Measurements were carried out in a RF anechoic chamber using a calibrated measuring system. In particular, Figure 7 shows the measurements of radiation pattern in E- plane and in H – plane for each dipole

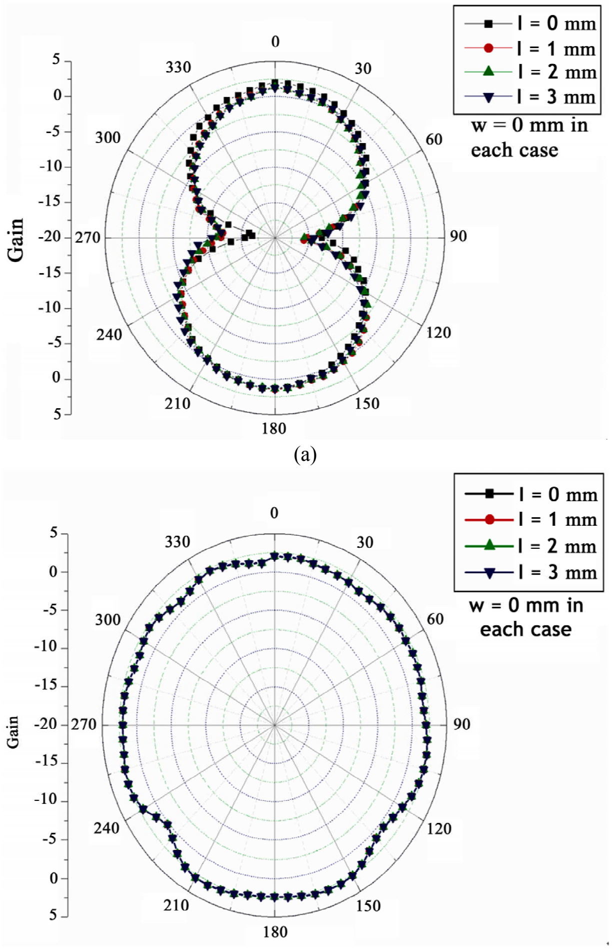

(a)

(b)

Figure 7. Radiation pattern of dipole for each value of *l* parameter (a) E – plane, (b) H -plane.

Copyright © 2010 SciRes.

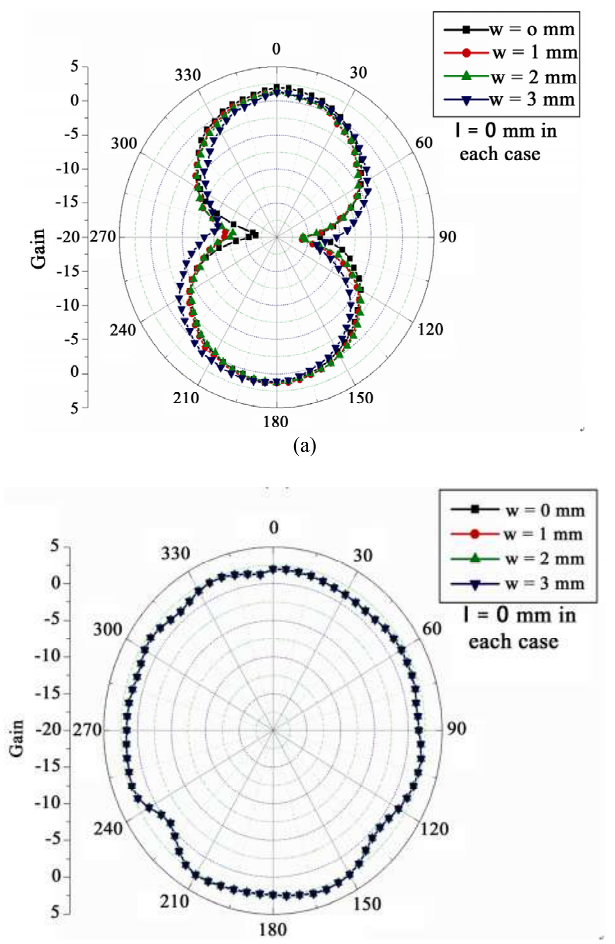

(a)

(b)

Figure 8. Radiation pattern of dipole for each value of *w* parameter (a) E – plane, (b) H – plane.

with *w* parameter equals to 0 mm and *l* parameter equals to integer values that ranging from *0 mm* to *3 mm*,. Figure 8 shows the corresponding results for each dipole with *l* parameter equals to *0 mm* and *w* parameter's integer values ranging from *0 mm* to *3 mm*. All these dipole structures introduce radiation characteristics that correspond to a fundamental dipole antenna [6,7]. Each of them has a measured peak gain that equals to quite 2 dBi and introduces omni-directional features. Quite small variations on these curves are on the limits of measurements' accuracy. For this reason, it can be observed that the radiation characteristics of the printed dipole antenna are not affected by the variations on *l* and *w* geometrical parameters. Therefore, the radiation diagrams of them are independent of the *l* and *w* parameters.

4. Conclusions

A number of printed dipole antennas with integrated balun are constructed and studied in terms of return loss and radiation pattern. Each of them has a defined form and geometry. Starting from a dipole antenna we mitered the angles introducing the parameters *l* and *w* that we varied. Experimental measurements on return loss provide the obtained results. These are quite similar and also introduce a resonance point at frequency range of 2.4 GHz with narrow resonance bandwidth. The form of this resonance range is affected only by the *l* parameter. The radiation pattern of these dipoles is also investigated. The corresponding radiation diagrams are independent of these geometrical parameters (*l*, *w*) and are similar to that of the fundamental dipole. These observations on printed dipole architecture are very crucial for wireless communication engineering and antenna design. This is because they introduce the ability of constructing a group of identical dipoles choosing an appropriate value of *l* parameter (*l =2 mm*) with quite identical resonance and radiation characteristics.

5. Acknowledgment

This research project (PENED) is co-financed by E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%).

6. References

[1] D. Edward and D. Rees, "A broadband printed dipole with integrated balun," Microwave J, 1987, pp. 339–344.

[2] N. Michishita, H. Arai, M. Nakano, T. Satoh, and T. Matsuoka, "FDTD analysis for printed dipole antenna with balun," Asia Pacific Microwave Conference, 2000, pp. 739–742.

[3] G. S. Hilton, C. J. Railton, G. J. Ball, A. L. Hume, and M. Dean, "Finite–difference time–domain analysis of a printed dipole antenna," 19th Int. IEEE Antennas and Propagation Conference, 1995, pp. 72–75.

[4] H. R. Chuang and L. C. Kuo, "3-D FDTD design analysis of a 2.4 GHz polarization – diversity printed dipole antenna with integrated balun and polarization– switching circuit for wlan and wireless communication application," IEEE Transactions on Microwave Theory and Techniques, Vol. 51, No. 2, 2003.

[5] D. M. Pozar, Microwave Engineering, Wiley, 1998.

[6] C. A. Balanis, Antenna Theory Analysis and Design, Wiley, 1997.

[7] R. Garg, P. Bhartia, I. Bahl, and A. Ittipiboon, Microstrip Antenna Design Handbook, Artec House, 2001.

Copyright © 2010 SciRes.

Circuits and Systems, 2012, 3, 216-222
http://dx.doi.org/10.4236/cs.2012.33030 Published Online July 2012 (http://www.SciRP.org/journal/cs)

Optimization of the Voltage Doubler Stages in an RF-DC Convertor Module for Energy Harvesting

Kavuri Kasi Annapurna Devi[1*], Norashidah Md. Din[2], Chandan Kumar Chakrabarty[2]
[1]Department of Electrical and Electronic Engineering, INTI International University, Nilai, Malaysia
[2]Department of Electronics and Communication Engineering, Universiti Tenega Nasional, Kajang, Malaysia
Email: {*kavurik.adevi., chandan}@newinti.edu.my, norashidah@uniten.edu.my

Received August 25, 2011; revised March 26, 2012; accepted April 4, 2012

ABSTRACT

This paper presents an optimization of the voltage doubler stages in an energy conversion module for Radio Frequency (RF) energy harvesting system at 900 MHz band. The function of the energy conversion module is to convert the (RF) signals into direct-current (DC) voltage at the given frequency band to power the low power devices/circuits. The design is based on the Villard voltage doubler circuit. A 7 stage Schottky diode voltage doubler circuit is designed, modeled, simulated, fabricated and tested in this work. Multisim was used for the modeling and simulation work. Simulation and measurement were carried out for various input power levels at the specified frequency band. For an equivalent incident signal of –40 dBm, the circuit can produce 3 mV across a 100 kΩ load. The results also show that there is a multiplication factor of 22 at 0 dBm and produces DC output voltage of 5.0 V in measurement. This voltage can be used to power low power sensors in sensor networks ultimately in place of batteries.

Keywords: Energy Conversion; RF; Schottky Diode; Villard; Energy Harvesting

1. Introduction

RF energy harvesting is one type of energy harvesting that can be potentially harvested such as solar, vibration and wind. The RF energy harvesting uses the idea of capturing transmitted RF energy at ambient and either using it directly to power a low power circuit or storing it for later use. The concept needs an efficient antenna along with a circuit capable of converting RF signals to DC voltage. The efficiency of an antenna mainly depends on its impedance and the impedance of the energy converting circuit. If the two impedances aren't matched then it will be unable to receive all the available power from the free space at the desired frequency band. Matching of the impedances means that the impedance of the antenna is the complex conjugate of the impedance of the circuit (voltage doubler circuit).

The concept of energy harvesting system is shown in **Figure 1**, which consists of matching network, RF-DC conversion and load circuits. The authors in [1], used a 2.4 GHz operating frequency with an integrated zero bias detector circuit using BiCMOS technology which produced an output voltage of 1 V into a 1 MΩ load at an input power level of 0 dBm. H. Yan and co-authors revealed that a DC voltage of 0.8 volts can be achieved from a –20 dBm RF input signal at 868.3 MHz through simulation results [2]. In [3], work was carried out on a firm frequency of 900 MHz by matching to a 50 Ω impedance and resonance circuit transformation in front of the Schottky diode which yields an output voltage of over 300 mV at an input power level of 2.5 µ. W. J. Wang, L. Dong and Y. Fu [4] used a Cockcroft-Walton multiplier circuit that produced a voltage level of 1.0 V into a 200 MΩ load for an input power level of less than –30 dBm at a fixed frequency of 2.4 GHz.

The energy conversion module designed in this paper is based on a voltage doubler circuit which can be able to output a DC voltage typically larger than a simple diode rectifier circuit as in [5], in which switched capacitor charge pump circuits are used to design two phase voltage doubler and a multiphase voltage doubler. The module presented in this can function as an AC to DC converter that not only rectifies the input AC signal but also elevates the DC voltage level. The output voltage of the

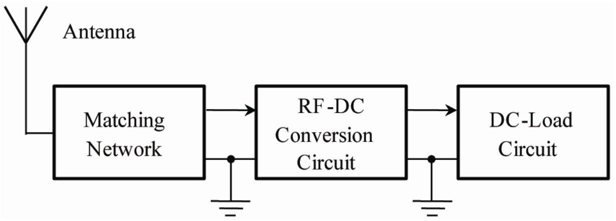

Figure 1. Schematic view of a RF energy harvesting system.

---

[*]Corresponding author.

Copyright © 2012 SciRes.

energy conversion module can be used to energize the low power devices for example sensors for a sensor network in application to agriculture.

Section 2 of this paper discusses on the theoretical background of the voltage doubler circuit. Section 3 presents the simulation study and implementation of the circuit design. Section 4 provides the results and analysis. Section 5 concludes with a discussion on the findings from the simulated and measured results.

2. Voltage Multiplier

There are various voltage multiplier circuit topologies. The design used in this module is derived from the function of peak detector or a half wave peak rectifier. The Villard voltage multiplier circuit was chosen in the circuit design of this paper because it produces two times of the input signal voltage towards ground at a single output and can be cascaded to form a voltage multiplier with an arbitrary output voltage and its design simplicity.

2.1. Diode Modeling

The voltage multiplier circuit in this design uses zero bias Schottky diode HSMS-2850 from Agilent. The attractive feature of these Schottky diodes are low substrate losses and very fast switching but leads to a fabrication overhead. This diode has been modeled for the energy harvesting circuit which comes in a one-diode configuration. The modeling parameters for these diodes are given by Agilent in their data sheets. These parameters are used in Multisim for its own modeling purposes. The modeling is done by transforming the diode into an equivalent circuit using passive components which are described by the SPICE parameters in **Table 1** [6].

The diode used in this design is shown in **Figure 2** and its equivalent model is shown in **Figure 3**. The special features of HSMS-2850 diode is that it provides a low forward voltage, low substrate leakage and uses the non symmetric properties of a diode that allows unidirectional flow of current under ideal condition.

Table 1. SPICE parameters.

Parameters	Units	HSMS 2850
B_V	V	3.8
C_{J0}	pF	0.18
E_G	Ev	0.69
I_{BV}	A	3E–4
I_S	A	3E–6
N	No unite	1.06
R_S	Ω	25
P_B (V_J)	V	0.35
P_T (XTI)	No units	2
M	No units	0.5

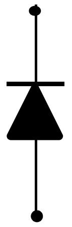

Figure 2. Schottky diode.

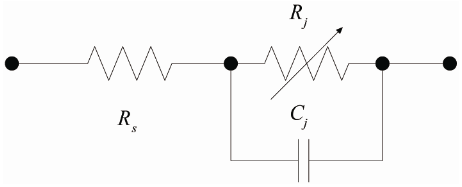

Figure 3. Linear circuit model of the Schottky diode [6].

The diodes are fixed and are not subject of optimization or tuning. This is described using the following derivations. By neglecting the effect of diode substrate, an equivalent linear model that can be used for the diode as shown in **Figure 3**. When C_j is the junction capacitance and R_j is the junction resistance, the admittance Y_z of the linear model is given by

$$Y_Z = Y_{C_j} + Y_{R_j} \tag{1}$$

Equation (1) related to the frequency of operation is given by

$$Y_Z = jwC_j + \frac{1}{R_j} \tag{2}$$

$$= \frac{jwC_jR_j + 1}{R_j} \tag{3}$$

The impedance Z of the linear model is given by

$$Z = \frac{R_j}{1 + jwR_jC_j} \tag{4}$$

The total impedance Z_T is given by

$$Z_T = R_S + \frac{R_j}{1 + jwR_jC_j} \tag{5}$$

where R_S is the series resistance of the circuit and R_j is given by

$$R_j = \frac{8.33 \times 10^{-5} \times N \times T}{I_b + I_S}$$

where:

I_b = bias current in µA;
I_s = saturation current in µA;
T = temperature (K);

N = ideality factor.

In Equation (5), R_j and C_j are constants and the frequency of operation (w) is the only variable parameter. As the frequency increases, the value of Z is almost negligible compared to the series resistance R_S of the diode. From this it is concluded that the function of the diode is independent of the frequency of operation.

2.2. Single Stage Voltage Multiplier

Figure 4 represents a single stage voltage multiplier circuit. The circuit is also called as a voltage doubler because in theory, the voltage that is arrived on the output is approximately twice that at the input. The circuit consists of two sections; each comprises a diode and a capacitor for rectification. The RF input signal is rectified in the positive half of the input cycle, followed by the negative half of the input cycle. But, the voltage stored on the input capacitor during one half cycle is transferred to the output capacitor during the next half cycle of the input signal. Thus, the voltage on output capacitor is roughly two times the peak voltage of the RF source minus the turn-on voltage of the diode.

The most interesting feature of this circuit is that when these stages are connected in series. This method behaves akin to the principle of stacking batteries in series to get more voltage at the output. The output of the first stage is not exactly pure DC voltage and it is basically an AC signal with a DC offset voltage. This is equivalent to a DC signal superimposed by ripple content. Due to this distinctive feature, succeeding stages in the circuit can get more voltage than the preceding stages. If a second stage is added on top of the first multiplier circuit, the only waveform that the second stage receives is the noise of the first stage. This noise is then doubled and added to the DC voltage of the first stage. Therefore, the more stages that are added, theoretically, more voltage will come from the system regardless of the input. Each independent stage with its dedicated voltage doubler circuit can be seen as a single battery with open circuit output voltage V_0, internal resistance R_0 with load resistance R_L, the output voltage, V_{out} is expressed as in Equation (7).

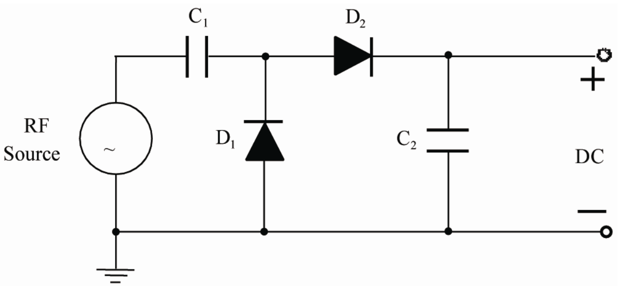

Figure 4. Single stage voltage multiplier circuit [7].

$$V_{\text{out}} = \frac{V_0}{R_0 + R_L} R_L \tag{6}$$

When n number of these circuits are put in series and connected to a load of R_L in Equation (6) the output voltage V_{out} obtained is given by this change in RC value will make the time constant longer which in turn retains the multiplication effect of two in this design of seven stage voltage doubler.

$$V_{\text{out}} = \frac{nV_0}{nR_0 + R_L} = V_0 \frac{1}{\frac{R_0}{R_L} + \frac{1}{n}} \tag{7}$$

The number of stages in the system has the greatest effect on the DC output voltage, as shown from Equations (6) and (7).

It is inferred that the output voltage V_{out} is determined by the addition of R_0/R_L and $1/n$, if V_0 is fixed. From this analysis it is observed that V_0, R_0 and R_L are all constants. Assume that $V_0 = 1$ V, $R_0/R_L = 0.25$, n = 2, 3, 4, 5, 6 and 7, the output voltage V_{out} = 1.33 V, 1.72 V, 2.0 V, 2.22 V, 2.43 V and 2.56 V respectively when substituted analytically in the Equation (7). This analysis can be compared with the results obtained in the circuit design of this module. In simulation at $n = 4$, V_{out} = 1.42 V, $n = 5$, V_{out} = 1.67 V; $n = 6$, V_{out} = 1.92; $n = 7$, V_{out} = 2.15 V; $n = 8$, V_{out} = 1.92 V; $n = 9$, V_{out} = 1.81 V. Also in measurement, for $n = 4$, V_{out} = 2.1 V; $n = 5$, V_{out} =2.9 V; $n = 6$, V_{out} = 3.72 V; and $n = 7$, V_{out} = 5 V. As n increases, the increase in output voltage will be almost double the input voltage up to some number of stages. But at some point, *i.e.* beyond seven stages, in this circuit the output voltage gained (8 and 9 stages) will be negligible as shown in **Figure 5**.

The capacitors are charged to the peak value of the input RF signal and discharge to the series resistance (R_s) of the diode. Thus the output voltage across the capacitor of the first stage is approximately twice that of the input signal. As the signal swings from one stage to other, there is an additive resistance in the discharge path of the

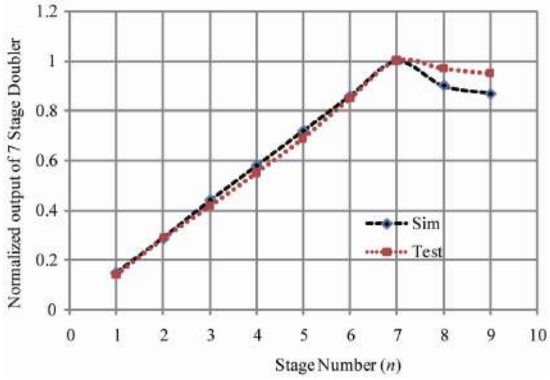

Figure 5. Normalized output voltage multiplier versus number of stages.

diode and increase of capacitance due to the stage capacitors.

2.3. Seven Stage Voltage Multiplier

The seven stage voltage multiplier circuit design implemented in this paper is shown in **Figure 6**. Starting on the left side, there is a RF signal source for the circuit followed by the first stage of the voltage multiplier circuit. Each stage is stacked onto the previous stage as shown in the **Figure 6**. Stacking was done from left to right for simplicity instead of conventional stacking from bottom to top.

The circuit uses eight zero bias Schottky surface-mount Agilent HSMS-285X series, HSMS-2850 diodes. The special features of these diode is that, it provides a low forward voltage, low substrate leakage and uses the non-symmetric properties of a diode that allows unidirectional flow of current under ideal conditions. The diodes are fixed and are not subject of optimization or tuning. This type of multiplier produces a DC voltage which depends on the incident RF voltage. Input to the circuit is a predefined RF source. The voltage conversion can be effective only if the input voltage is higher than the Schottky forward voltage.

The other components associated with the circuit are the stage capacitors. The chosen capacitors for this circuit are of through-hole type, which make it easier to modify for optimization, where in [8] the optimization was accomplished at the input impedance of the CMOS chip for a three stage voltage multiplier. The circuit design in this paper uses a capacitor across the load to store and provide DC leveling of the output voltage and its value only affects the speed of the transient response. Without a capacitor across the load, the output is not a good DC signal, but more of an offset AC signal.

In addition to the above, an equivalent load resistor is connected at the final node. The output voltage across the load decreases during the negative half cycle of the AC input signal. The voltage decreases is inversely proportional to the product of resistance and capacitance across the load. Without the load resistor on the circuit, the voltage would be hold indefinitely on the capacitor and look like a DC signal, assuming ideal components. In the design, the individual components of the stages need not to be rated to withstand the entire output voltage. Each component only needs to be concerned with the relative voltage differences directly across its own terminals and of the components immediately adjacent to it. In this type of circuitry, the circuit does not change the output voltage but increases the possible output current by a factor of two. The number of stages in the system is directly proportional to the amount of voltage obtained and has the greatest effect on the output voltage as explained in the Equation (7) and shown in **Figure 5**.

3. Simulation and Implementation

Multisim software was chosen for modeling and simulation which is a circuit simulation tool by Texas Instruments. The simulation and practical implementation were carried out with fixed RF at 945 MHz ± 100 MHz, which are close to the down link center radio frequency (947.5 MHz) of the GSM-900 transmitter. The voltages obtained at the final node V_{DC} of the multiplier circuit were recorded for various input power levels from –40 dBm - +5 dBm with power level interval (spacing) of 5 dBm.

The simulations were also carried out using same stage capacitance value (3.3 nF) and then with a varied capacitance value for all stages from 4 stages through 9 stages [9]. The capacitance value was varied in such a way that, from one stage to the next, it was halved. For example, if the first stage was 3.3 nF, the second stage was 1.65 nF, third stage was 825 pF, fourth stage was 415 pF and so on. But keeping in view of testing, the capacitance values were chosen to have a close match with the standard available values in the market.

Simulation was carried out through 4 to 9 voltage doubler stages. Based on results obtained a 7 stage doubler is best to implemented for this application.

The design of the printed circuit board (PCB) was carried out using DipTrace software. The material used to

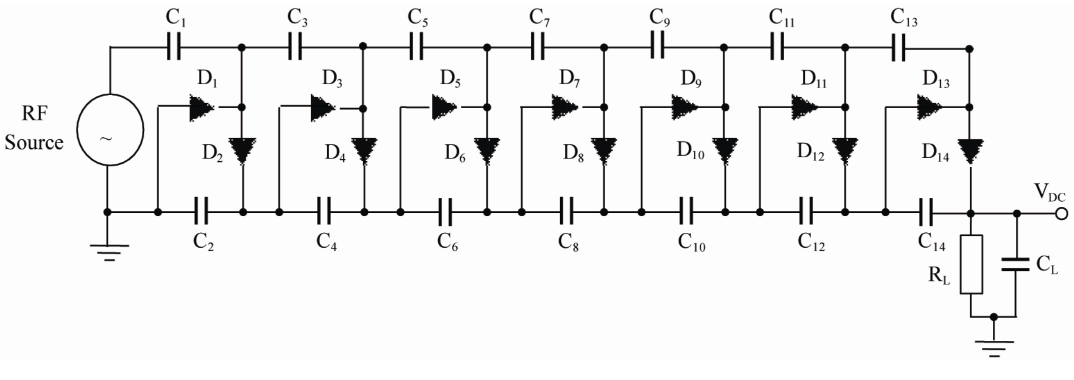

Figure 6. Schematic of 7 stage voltage multiplier.

manufacture the printed circuit board (PCB) is the standard Fiberglass Reinforced Epoxy (FR4), with the thickness of 1.6 mm and dielectric constant of 3.9. The topology is constructed on the PCB with the dimensions of 98 mm × 34 mm (W × H). The Sub Miniature version A (SMA) connectors are used at the input and output of PCB to carry out the measurements. The circuit components consist of active and passive components. The component used in circuit is shown in **Table 2**.

Special handling precautions have been taken to avoid Electro Static Discharge (ESD), while assembling of the surface-mount zero bias Schottky diodes. Also special attention has been given to mount other components and the SMA connectors on to the PCB. The Photograph of Assembled circuit board I shown in **Figure 7**.

4. Results and Analysis

The simulated and measured results at the output voltage of voltage multiplier circuit are shown graphically in **Figure 8**. From the graph analysis, the simulated and the measured results agree considerably with each other. The measured results are shown to be better than the simulation results. The reason behind this may be due to the uncertainty in series resistance value of the diode obtained from SPICE parameters in modeling as explained in Equation (5). This resistance vale of diodes in practical circuit may be lower than in the model, which provides fast discharge path, in turn rise in voltage as passes through the stages and reaches to final output. In this work, the DC output voltages obtained through simulation and measurement at 0 dBm re 2.12 V and 5.0 V respectively. These results are comparatively much better than in ref. [9], where in at 0 dBm, 900 MHz they achieved 0.5 V and 0.8 V through simulation and measurement respectively.

Figures 9 and **10** show the result of a 4 stage voltage doubler circuit with equal and varied capacitance values between the stages as described in Section 3.

From the analysis of these two simulations, it can be observed that the resulting output voltages are equal. The only difference between these two graphs is the rise time of the circuit with varied capacitance value is a little bit slower. But, overall result on the performance of rise time is still under 20 µs to 24 µs and the difference is negligible. From these results, the use of equal stage capacitance of each being 3.3 nF was hence considered for the design of the multiplier.

The results from **Figure 11**, shows that the output voltage reaches to 1.0 V within 20 µS and then uniformly increasing to 1.4 V, 1.67 V, 1.87 V and 2.12 V for 4, 5, 6 and 7 stages respectively compared to 2 mS as shown in [10]. **Figure 12** shows that the conversion ratio of 22 is achieved at 0 dBm input power and drops to 2.5 at –40 dBm. The highest value at 0 dBm is due to the innate characteristics of the zero bias Schottky diodes which conduct fairly well at higher input voltages.

5. Conclusion

From the experimental results, it is found that the pro-

Table 2. Component used in 7 stage voltage multiplier.

Name of component	Label	Value
Stage capacitors	C_1 - C_{14}	3.3 nF
Stage diodes	D_1 - D_{14}	HSMS 2850
Filter capacitor	C_L	100 nF
Load resister	R_L	100 kΩ

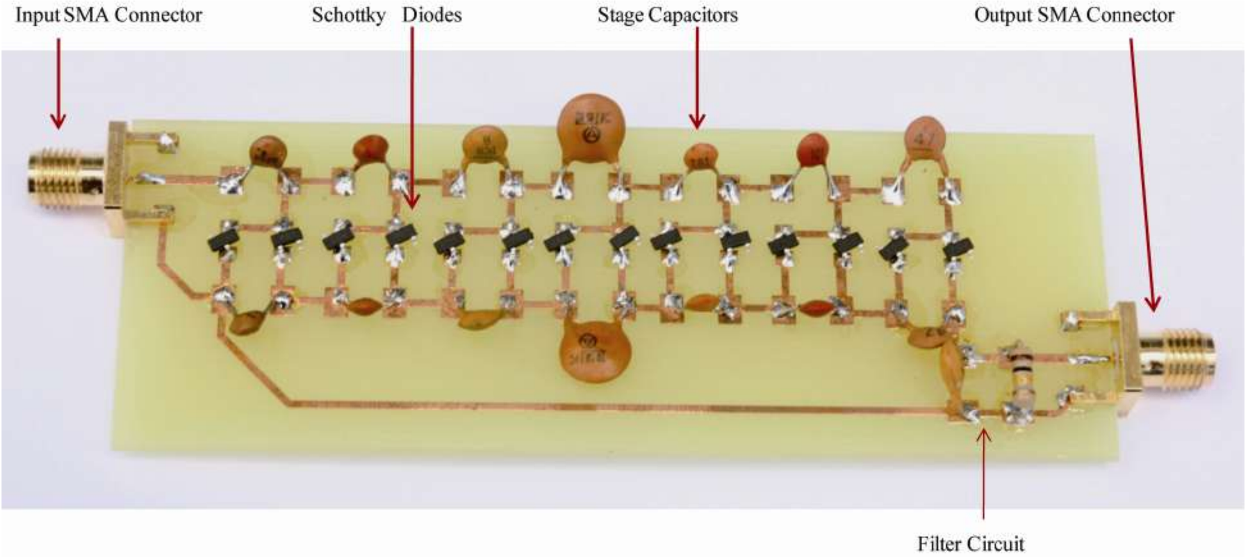

Figure 7. Photograph of assembled circuit board.

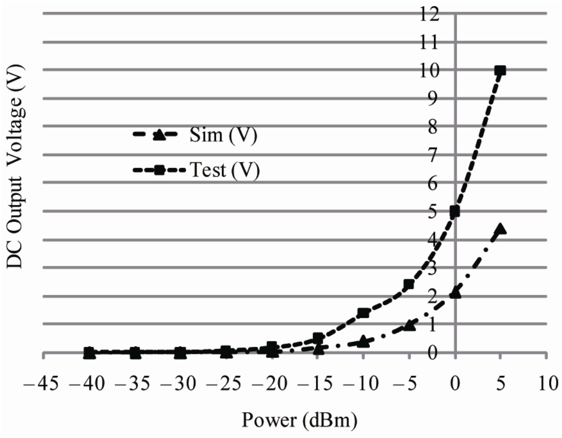

Figure 8. Simulated and test DC output voltage of multiplier as a function of input power.

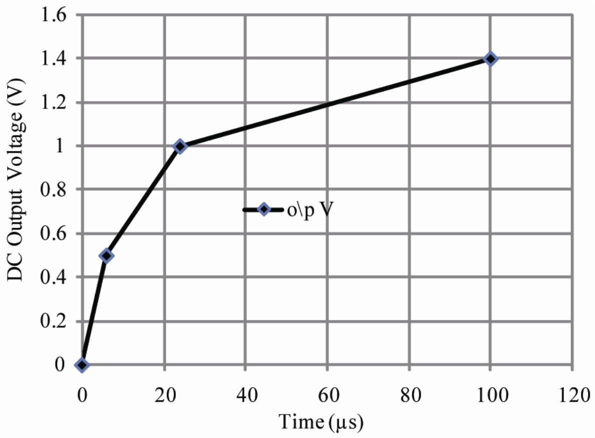

Figure 9. DC output voltage verses rise time of 4 stage voltage doubler circuit with equal stage capacitance [8].

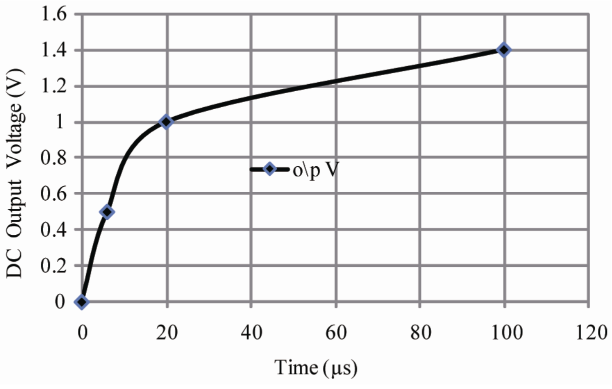

Figure 10. DC output voltage verses rise time of 4 stage voltage doubler with varied stage capacitance [8].

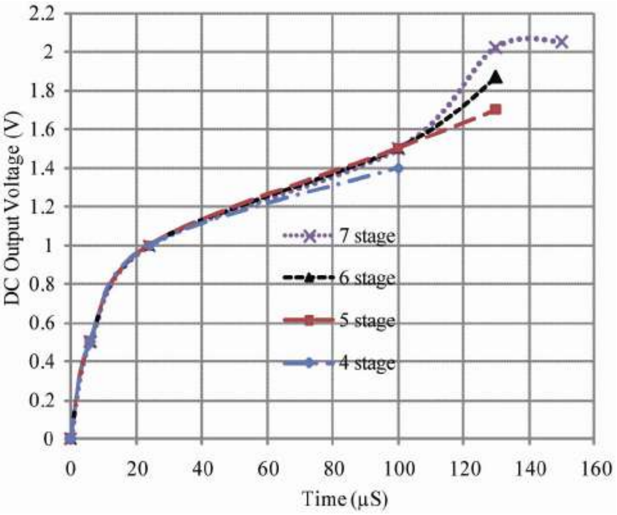

Figure 11. DC output voltage verses rise time of voltage doubler circuit through 4 - 7 stages with equal stage capacitance.

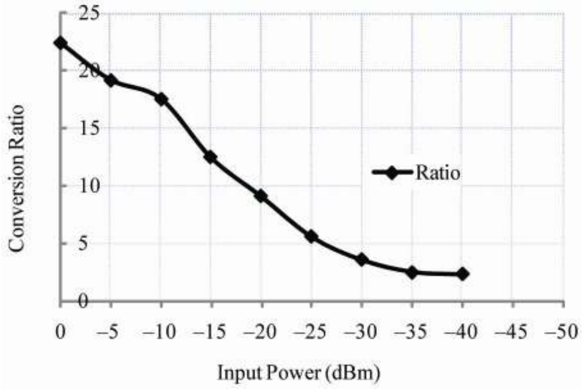

Figure 12. Conversion ratio as a function of input power.

posed voltage multiplier circuit operates at the frequency of 945 MHz with the specified input power levels. The results have shown that there is multiplication of the input voltage. From **Figure 12**, it is shown that at 0 dBm input power, the multiplication factor is 22. This is significant, as the work shows that RF energy in the GSM-900 band can be harvested from the ambient RF source using the Villard circuit topology. The power density levels from a GSM base station is expected from 0.1 mW/m^2 to 1 mW/m^2 for a distance ranging from 25 m - 100 m. These power levels may be elevated by a factor between one and three for the GSM-900 downlink frequency bands depending on the traffic density [10]. The next phase of the research work is to interface the voltage multiplier circuit through a matching network to the antenna at the input side and a low power device to power from the system at the output side to complete the RF energy harvesting system.

6. Acknowledgements

We would like to acknowledge and thank the Ministry of Higher Education Malaysia for funding this project under Fundamental Research Grant: FRGS/1/10/TK/UNITEN/02/13.

REFERENCES

[1] S. von der Mark and G. Boeck, "Ultra Low Power Wakeup Detector for Sensor Networks," *SBMO/IEEE MTT-S International Microwave & Optoelectronics Conference* (*IMOC* 2007), Berlin, 29 October-1 November 2007, pp. 865-868. doi:10.1109/IMOC.2007.4404394

[2] H. Yan, J. G. M. Montero, A. Akhnoukh, L. C. N. de Vreede and J. N. Burghartz, "An Integration Scheme for RF Power Harvesting," *Proceedings of STW Annual Workshop on Semiconductor Advances for Future Electronics and Sensors*, Veldhoven, November 2005, pp. 64-66. doi:10.1.1.126.5786

[3] T. Ungan and L. M. Reindl, "Concept for Harvesting Low Ambient RF-Sources for Microsystems," 2007. http://www.imtek.de/content/pdf/public/2007/powermems_2007_paper_ungan.pdf

[4] J. Wang, L. Dong and Y. Z. Fu, "Modeling of UHF Voltage Multiplier for Radio-Triggered Wake-Up Circuits," *International Journal of Circuit Theory and Application*, Vol. 39, No. 11, 2010, pp. 1189-1197. doi:10.1002/cta.692

[5] J. A. Starzyk, Y.-W. Jan and F. J. Qiu, "A DC-DC Charge Pump Design Based on Voltage Doublers," *IEEE Transactions on Circuits and Systems-I*: *Fundamental Theory and Applications*, Vol. 48, No. 3, 2001, pp. 350-359.

[6] HSMS-2850, "Surface Mount Zero Bias Schottky Detector Diodes." http://www.crystal-radio.eu/hsms285xdata.pdf

[7] D. W. Harrist, "Wireless Battery Charging System Using Radio Frequency Energy Harvesting," M.S. Thesis, University of Pittsburgh, Pittsburgh, 2004.

[8] E. Bergeret, J. Gaubert, P. Pannie and J. M. Gaultierr, "Modeling and Design of CMOS UHF Voltage Multiplier for RFID in an EEPROM Compatible Process," *IEEE Transactions on Circuits and Systems-I*, Vol. 54, No. 10, 2007, pp. 833-837. doi:10.1109/TCSII.2007.902222

[9] B. Emmanuel, J. Gaubert, P. Pannier and J. M. Gaultier, "Conception of UHF Voltage Multiplier for RFID Circui," *IEEE North-East Workshop on Circuits and Systems*, Gatineau, June 2006, pp. 217-220. doi:10.1109/NEWCAS.2006.250961

[10] H. J. Visser, A. C. F. Reniers and J. A. C. Theeuwes, "Ambient RF Energy Scavenging: GSM and WLAN Power Density Measurements," *Proceedings of the* 38*th European Microwave Conference*, Eindhoven, 27-31 October 2008, pp. 721-724. doi:10.1109/EUMC.2008.4751554

Copyright © 2012 SciRes.

Journal of Electromagnetic Analysis and Applications, 2013, 5, 182-188
http://dx.doi.org/10.4236/jemaa.2013.54029 Published Online April 2013 (http://www.scirp.org/journal/jemaa)

Double-Sided Microstrip Circular Antenna Array for WLAN/WiMAX Applications

Ruwaybih Alsulami, Heather Song[*]

Department of Electrical and Computer Engineering, University of Colorado, Colorado Springs, USA.
Email: *hsong@uccs.edu

Received February 8th, 2013; revised March 10th, 2013; accepted March 25th, 2013

Copyright © 2013 Ruwaybih Alsulami, Heather Song. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

The design, fabrication, and characterization of the microstrip circular antenna arrays were presented. The proposed antennas were designed for single band at 2.45 GHz and dual bands at 3.3 - 3.6 and 5.0 - 6.0 GHz to support WLAN/WiMAX applications. The proposed single and dual band antennas showed omnidirectional radiation pattern with the gain values of 3.5 dBi at 2.45 GHz, 4.0 dBi at 3.45 GHz, and 3.3 dBi at 5.5 GHz. The dual band antenna array was placed on both top and bottom layers to obtain the desired antenna characteristics. The proposed double-sided dual band antenna provides omnidirectional radiation pattern with high gain.

Keywords: Antenna Arrays; Circular Patch; Dual Band; Single Band; Omnidirectional; WLAN/WiMAX Applications; UWB

1. Introduction

Ultra-wideband (UWB: 3.1 to 10.6 GHz) frequency spectrum has been approved by the US Federal Communications Commission (FCC) for unlicensed short range wireless communications since 2002. In this frequency range, wireless local-area network (WLAN) IEEE802.11a and HIPERLAN/2 WLAN operates in 5.0 - 6.0 GHz band. In some European and Asian countries, world interoperability for microwave access (WiMAX) service is provided in the frequency range of 3.3 - 3.6 GHz [1-4]. To support the WLAN/WiMAX application, antenna arrays that provide omnidirectional radiation pattern are required. To respond to this need, recent antenna design efforts were focused on omnidirectional antennas with high gain and no sidelobes [5-8]. Rectangular arrays are common type used for antenna arrays. Studies on dual band antennas employing rectangular arrays were reported [9-12]. Compared to rectangular patch antenna arrays, there are limited numbers of studies performed on circular patch antenna arrays due to difficulties in fabrication [13]. Advantages of circular antenna array include high gain and narrow beam width [13].

In this paper, a new microstrip circular antenna arrays were designed, fabricated, and characterized to provide omnidirectional radiation pattern for WLAN/WiMAX applications. Two antenna arrays were designed—one for single band at 2.45 GHz and the other for dual bands at 3.3 - 3.6 GHz and 5.0 - 6.0 GHz. For single band operation, circular patch array was placed on the top layer of the microtrip and a small rectangular patch was placed on the bottom layer for ground connection. For dual band operation, similar circular patch array was placed on both top and bottom layers of the microstrip with larger rectangular patch placed on the bottom layer. Both single band (single sided) and dual band (double-sided) microstrip antenna arrays provided desirable antenna characteristics for the intended application.

2. Design and Simulation

2.1. Single-Band Antenna at 2.45 GHz

The configuration of the proposed single band antenna at 2.45 GHz is shown in **Figure 1**. It consists of six circular patches which are placed only on the top layer. The small rectangular patch is placed on the bottom layer for ground connection.

The directivity for the circular patch antenna is

$$D_0 = \frac{(k_0 a_e)^2}{120 G_{\text{rad}}} \tag{1}$$

---

[*] Corresponding author.

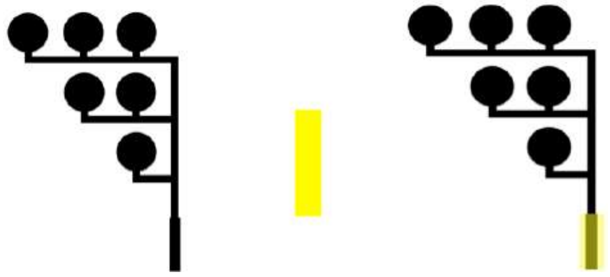
(a) (b) (c)

Figure 1. Configuration of the proposed antenna for single band at 2.45 GHz: (a) Top layer; (b) Bottom layer; (c) Top and bottom layers overlaid.

$$k_0 = \frac{2\pi}{\lambda_0} \tag{2}$$

$$a_e = a\left\{1+\frac{2h}{\pi a \epsilon_r}\left[\ln\left(\frac{\pi a}{2h}\right)+1.7726\right]\right\}^{1/2} \tag{3}$$

$$G_{\text{rad}} = \frac{\left(k_0 a_e\right)^2}{480}\int_0^{\pi/2}\left[J_{02}'^2+\cos^2\theta J_{02}^2\right]\sin\theta \mathrm{d}\theta \tag{4}$$

$$J_{02}' = J_0\left(k_0 a_e \sin\theta\right) - J_2\left(k_0 a_e \sin\theta\right) \tag{5}$$

$$J_{02} = J_0\left(k_0 a_e \sin\theta\right) + J_2\left(k_0 a_e \sin\theta\right) \tag{6}$$

where a_e is the effective radius, a is the actual radius, ϵ_r is the relative permittivity of the microstrip dielectric substrate, h is the height of the microstrip substrate, and J_0 and J_2 are Bessel functions.

The gain of the antenna was calculated using

$$\text{Gain} = \text{Antenna Efficiency} \times \text{Directivity}\left(D_0\right) \tag{7}$$

$$\text{Antenna Efficiency} = \frac{\text{Total Efficiency}}{\text{Reflection Efficiency}} \tag{8}$$

The variable corresponding to each dimensions and values for the dimensions of the proposed antenna are shown in **Figure 2** and **Table 1**, respectively. Here, *L*, *W*, and *R* represent the length, the width, and the radius of the circular patch, respectively.

The gain of the proposed antenna shown in **Figure 1** was calculated using (1) - (8) and the dimensions were optimized using ADS [14] which resulted in gain of 3.5 dBi at 2.45 GHz.

2.2. Dual-Band Antenna at 3.3 - 3.6 and 5.0 - 6.0 GHz

The configuration for the doubled-sided microstrip dual band antenna is shown in **Figure 3**. The proposed microstrip antenna has circular arrays both on the top and bottom layers. It consists of three circular patched on each layer.

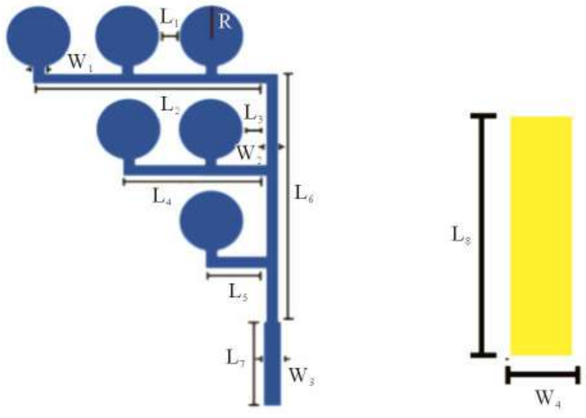

(a) (b)

Figure 2. Variables corresponding to each dimension of the proposed single band antenna: (a) Top layer; (b) Bottom layer.

Table 1. Dimensions for the proposed single band antenna at 2.4 GHz.

Variable	Value (mm)
L_1	1.15
L_2	28.9
L_3	1.13
L_4	17.6
L_5	6.60
L_6	35.3
L_7	18.4
L_8	18.4
W_1	1.02
W_2	1.52
W_3	3.30
W_4	6.86
R	5.10

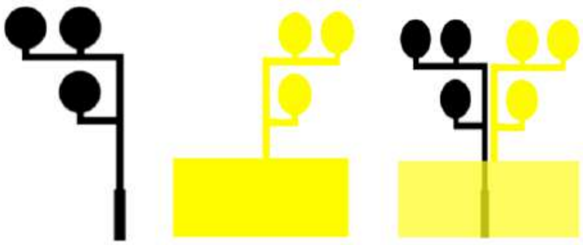
(a) (b) (c)

Figure 3. Configuration of the proposed antenna for dual band at 3.3 - 3.6 and 5.0 - 6.0 GHz: (a) Top layer; (b) Bottom layer; (c) Top and bottom layers overlaid.

The configuration in **Figure 3(a)** is similar to the top layer of the single band antenna as shown in **Figure 1(a)** but with less circular patches. However, the bottom layer in **Figure 3(b)** is different compared to the bottom layer of the single band antenna shown in **Figure 1(b)**. The double-side nature of the antenna provides dual band characteristics. Identical equations were used for the single band antenna were employed in the design process. The variable corresponding to each dimensions and the dimensions for the proposed dual band antenna are shown in **Figure 4** and **Table 2**, respectively.

Simulation was performed using ADS for the configuration shown in **Figure 3(c)**. The simulated gains of the proposed dual band antenna were 4.0 dBi at 3.45 GHz and 3.3 dBi at 5.5 GHz. The double-sided configuration of the antenna provided higher gain compared to the singled-sided antenna.

3. Measurement Results and Discussions

3.1. Single-Band Atnenna at 2.45 GHz

The antennas were fabricated using LPKF Protomat [15] on FR-4 material with height of 1.524 mm. The photos of the fabricated single band antenna are shown in **Figure 5** which has a size of 6.7 × 4.4 (in cm).

Figure 6 shows the comparison between the simulated and the measured S_{11} results.

The measured operating frequency is close to 2.45 GHz with S_{11} value below −15 dB. The 3 dB bandwidth at 2.45 GHz was approximately 18%. The measurement and simulation are in fairly good agreement, and the differences are due to microstrip loss and fabrication errors.

Figure 7 shows the comparison between the simulated and measured radiation pattern in xy-plane at 2.45 GHz which is close to omnidirectional pattern.

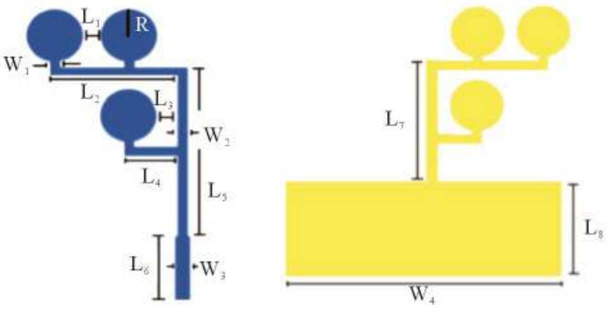

(a) (b)

Figure 4. Variables corresponding to each dimension of the proposed dual-band antenna: (a) Top layer; (b) Bottom layer.

Table 2. Dimensions for the proposed dual band antenna at 3.3 - 3.6 and 5.0 - 6.0 GHz.

Variable	Value (mm)
L_1	1.20
L_2	17.8
L_3	1.08
L_4	6.60
L_5	29.3
L_6	10.0
L_7	21.6
L_8	17.8
W_1	1.02
W_2	1.52
W_3	2.54
W_4	49.1
R	5.21

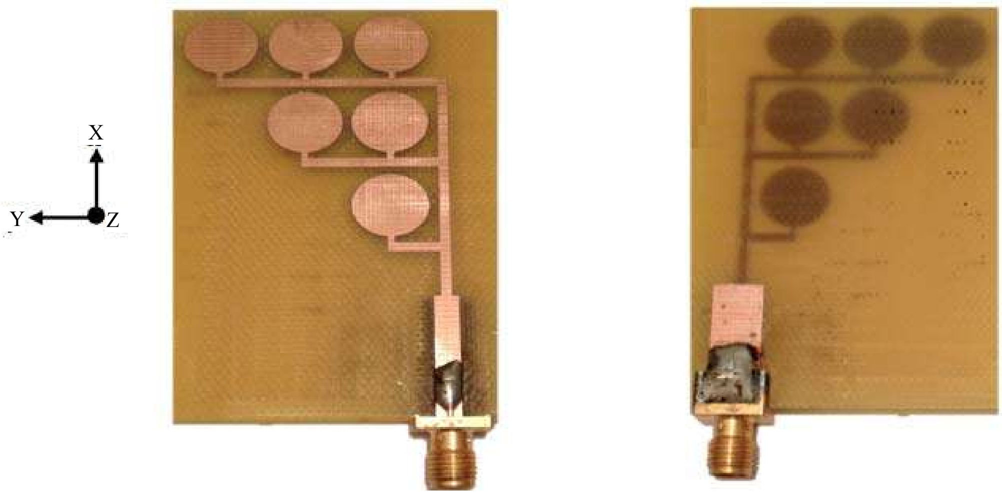

(a) (b)

Figure 5. Photo of the fabricated single-band antenna: (a) Top layer; (b) Bottom layer.

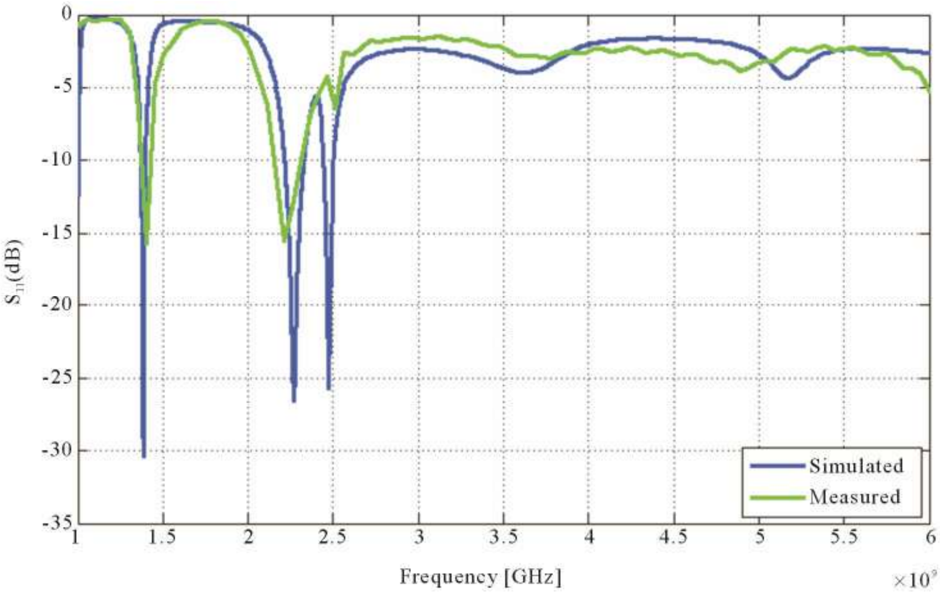

Figure 6. Simulated and measured return loss for the proposed single-band antenna.

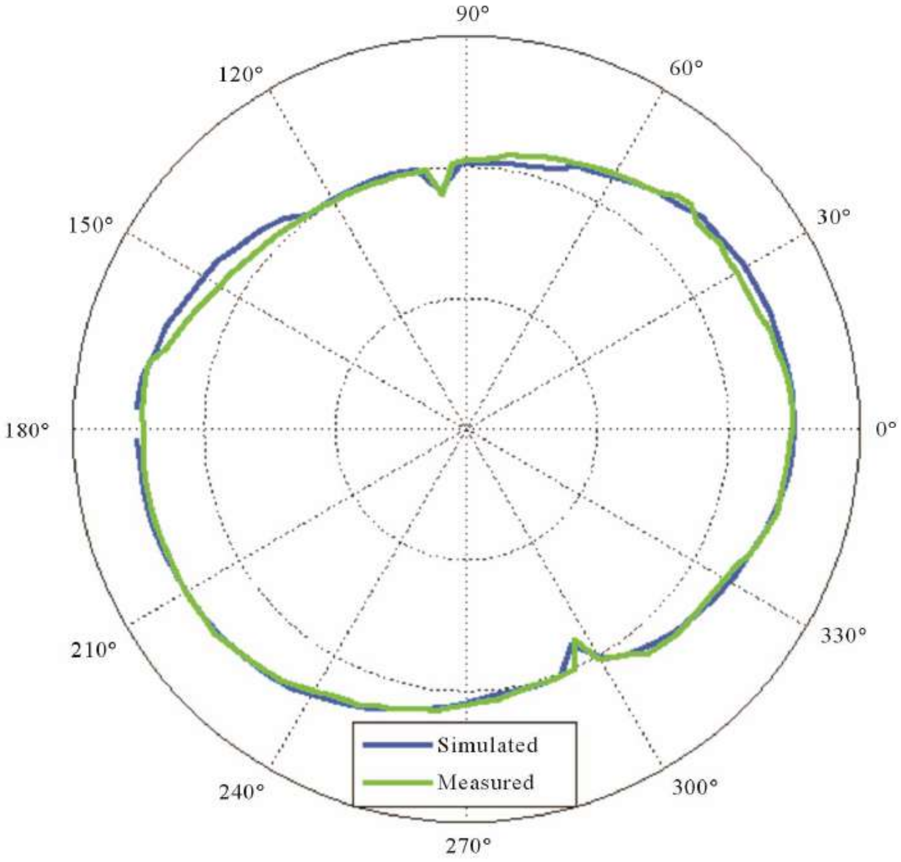

Figure 7. Simulated and measured radiation pattern in xy-plane (coordinate system shown in Figure 5) at 2.45 GHz.

3.2. Dual-Band Antenna at 3.3 - 3.6 and 5.0 - 6.0 GHz

The antennas were fabricated using LPKF Protomat [15] on double-sided FR-4 materials. The photos of the fabricated dual band antenna are shown in **Figure 8** which has a size of 6.6 × 5.2 (in cm).

Figure 9 shows the comparison between the simulated

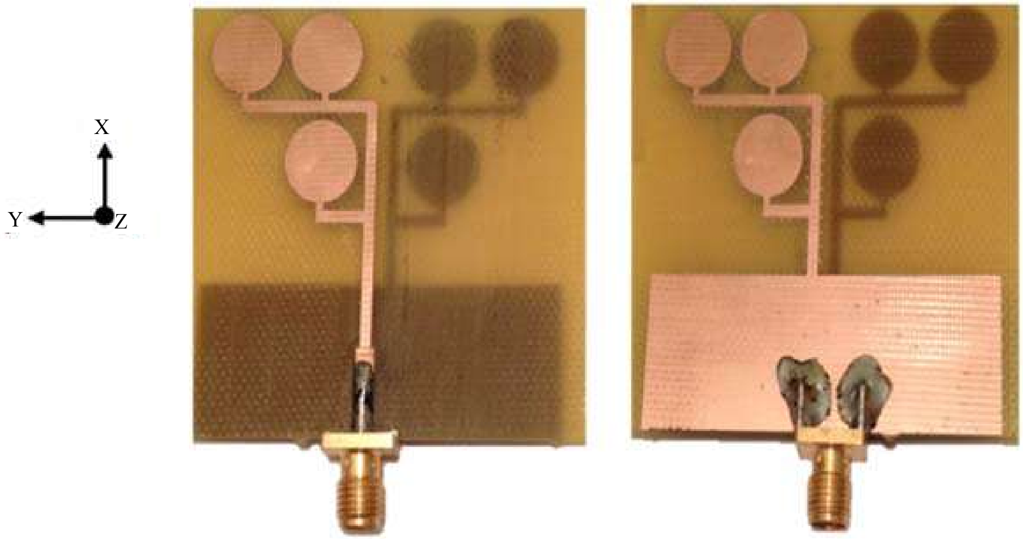

(a) (b)

Figure 8. Photo of the fabricated dual-band antenna: (a) Top layer; (b) Bottom layer.

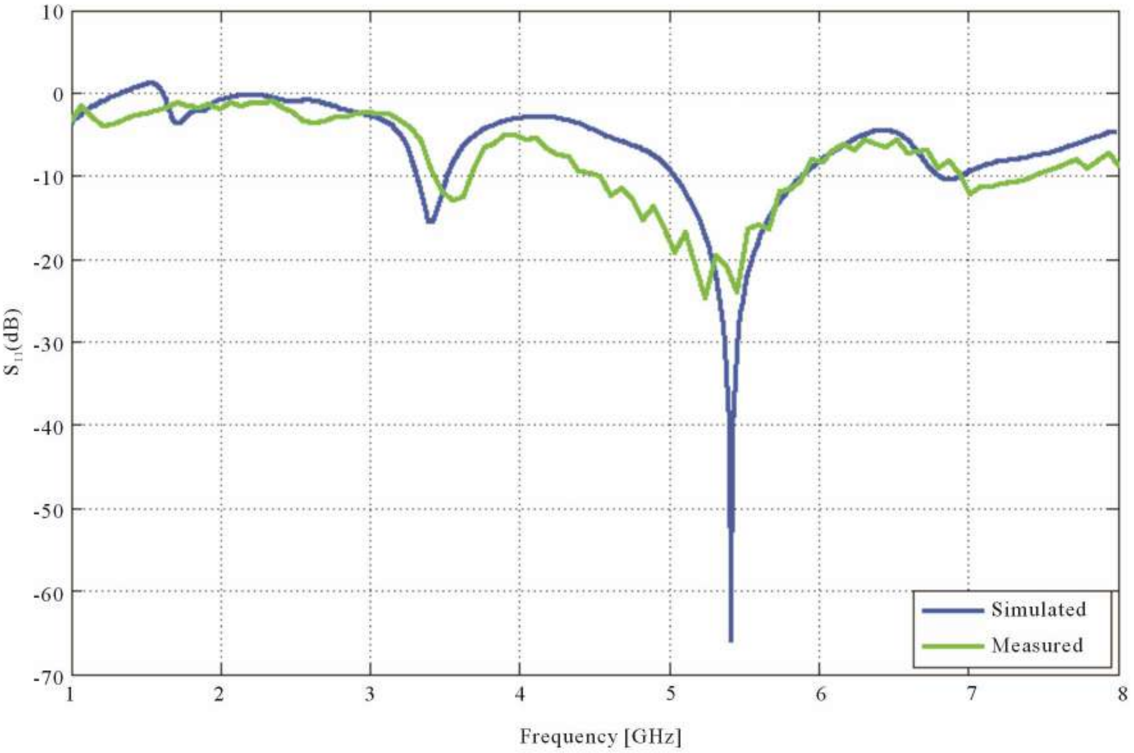

Figure 9. Simulated and measured return loss for the dual band antenna.

and the measured S_{11} results.

The measured S_{11} shows dual band near the designed bands with S_{11} values below −10 dB for both bands. The simulated and measured results give fairly good agreement, and the differences are due to board loss and fabrication errors.

Figure 10 shows the comparison between the simulated and measured radiation pattern in xy-plane at 3.45 and 5.5 GHz which is close to omnidirectional pattern.

4. Conclusion

A microstrip circular antenna arrays were presented for single band at 2.45 GHz and dual bands at 3.3 - 3.6 and

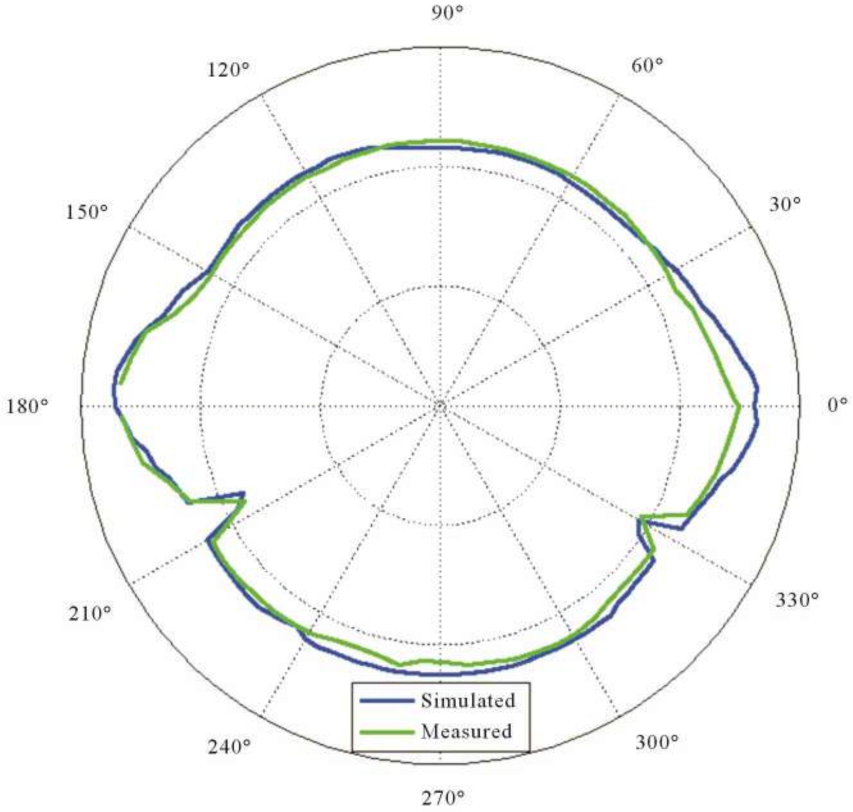

(a)

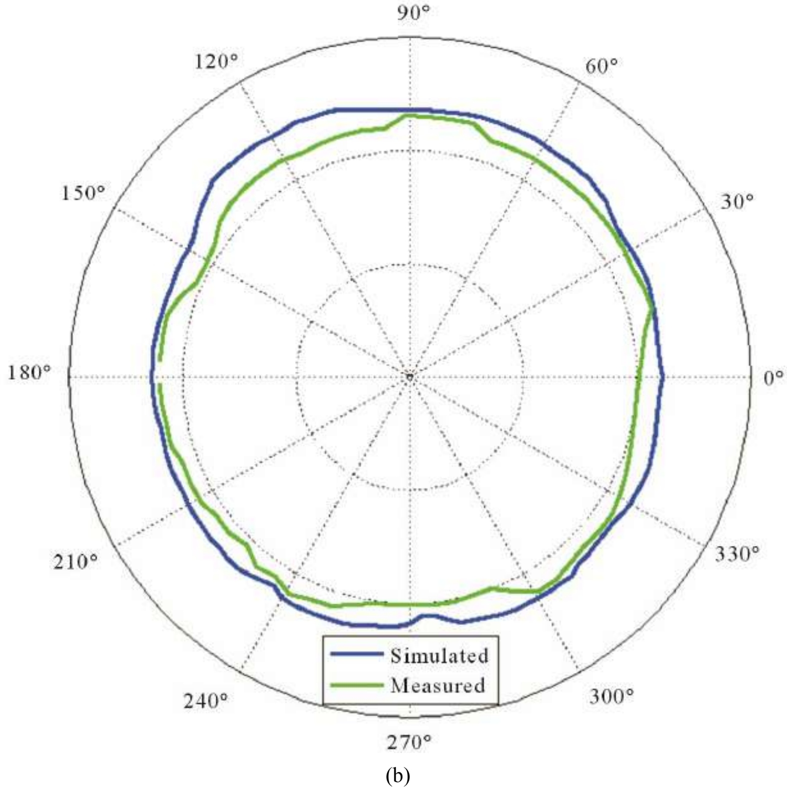

(b)

Figure 10. Simulated and measured radiation pattern in xy-plane (coordinate system shown in Figure 8) at (a) 3.45 GHz and (b) 5.5 GHz.

5.0 - 6.0 GHz for WLAN/WiMAX applications. Both antennas were designed with ADS, fabricated on a FR-4 microstrip material, and characterized. Both single band (single sided) and dual band (double-sided) antenna arrays provided omnidirectional pattern with desired gain.

REFERENCES

[1] C.-Y. Pan, T.-S. Horng, W.-S. Chen and C.-H. Huang, "Dual Wideband Printed Monopole Antenna for WLAN/WiMAX Applications," *IEEE Antennas and Wireless Propagation Letters*, Vol. 6, 2007.

[2] A. C. Rao and R. Pandeeswari, "A CPW-Fed Antenna for Dual Band WiMAX/WLAN Applications," *IEEE International Conference on Recent Trends in Information Technology ICRTIT*, 2011.

[3] H.-Y. Lai, Z.-Y. Lei, Y.-J. Xie, G.-L. Ning and K. Yang, "UWB Antenna with Dual Band Rejection for WLAN/WiMAX Bands Using CSRRs," *Progress in Electromagnetics Research Letters*, Vol. 26, 2011, pp. 69-78. doi:10.2528/PIERL11070112

[4] D. Parkash and R. Khanna "Design of a Dual Band Monopole Antenna for WLAN/WiMAX Applications," *IEEE Wireless and Optical Communications Networks*, 2010.

[5] L. Wang, K. Wei, J. Feng, Z. Zhang and Z. Feng, "A Wideband Omnidirectional Planar Microstrip Antenna for WLAN Applications," *IEEE*, 2011.

[6] J. Li, "An Omnidirectional Microstrip Antenna for WiMAX Applications," *IEEE Antennas and Wireless Propagation Letters*, Vol. 10, 2011.

[7] O. Tze-Meng and T. K. Geok, "A Dual-Band Omni-Directional Microstrip Antenna," *Progress in Electromagnetics Research*, Vol. 106, 2010, pp. 363-376. doi:10.2528/PIER10052411

[8] M. B. Bicer and A. Akdagli, "A Novel Microstrip-Fed Monopole Antenna for WLAN/WiMAX Applications," *Journal of Electromagnetic Waves and Applications*, Vol. 26, No. 7, 2012, pp. 904-913. doi:10.1080/09205071.2012.710372

[9] N. AbWahab, Z. Bin Maslan, W. N. W. Muhamad and N. Hamzah, "Microstrip Rectangular 4 × 1 Patch Array Antenna at 2.5 GHz for WiMAX Application," 2*nd International Conference on Computational Intelligence*, *Communication Systems and Networks*, 2010.

[10] Z. Zhong-xiang, C. Chang, W. Xian-liang and F. Minggang, "A 24 GHz Microstrip Array Antenna with Low Side Lobe," Springer, Berlin, 2012.

[11] A. K. Sahu and M. R. Das, "4 × 4 Rectangular Patch Array Antenna for Bore Sight Application of Consial Scan S-Band Traching Radar," *Antenna Week* (*IAW*), 2011.

[12] J. Das, T. A. Khan and M. K. Pal, "Rectangular Patch Antenna Array for Wireless Application," *International Journal of Engineering Science and Technology*, 2012.

[13] T. I. Huque, A. Chhowdhury, K. Hosain and S. Alam, "Performance Analysis of Corporate Feed Rectangular Patc Element and Circular Patch Element 4 × 2 Microstrip Array Antennas," *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 7, 2011, pp. 74-79.

[14] http://www.home.agilent.com/en/pc-1297113/advanced-design-system

[15] http://www.lpkf.com/protomat

J. Electromagnetic Analysis & Applications, **2011, 3, 43-46**
doi:10.4236/jemaa.2011.32008 Published Online February 2011 (http://www.SciRP.org/journal/jemaa)

Back to Back Combined Single Feed Proximity Coupled Antenna with Dumbbell Shaped DGS

Ashwini Kumar Arya, Amalendu Patnaik, Machavaram V. Kartikeyan

Department of Electronics & Computer Engineering, IIT Roorkee, Roorkee, India.
Email: ashwiniarya.iitr@gmail.com, {apatnaik, kartik}@ieee.org

Received November 12th, 2010; revised December 10th, 2010; accepted January 5th, 2011

ABSTRACT

Use of defected ground structure (DGS) to reduce the size of patch antenna is presented in this paper. In order to get a dipole like radiation pattern for some specific application a dumbbell shaped DGS is used in the common ground plane of back to back combined single fed proximity coupled antenna. A size reduction of about 60% is achieved. Parametric analysis has been done to see the resonance behavior of the antenna with DGS.

Keywords: *Defected Ground structure, Microstrip Antennas, Proximity Coupling*

1. Introduction

The continuous shrinking size of electronic equipments demands similar size antenna elements in order to fit properly in wireless devices without compromising the other radiation properties of the antenna. In this respect microstrip patch antennas are quite an obvious choice because of its other benefits like low profile, light weight, low cost and easy fabrication.

But as far as size of these patches concerned, the patch length should be around half-a-wavelength for the structure to act as a good radiator. Different techniques have already been used for the antenna size reduction such as using the substrate with high dielectric constant [1] , edge shorted patches with shorting plates or shorting walls, use of the shorting pin at the suitable position etc [2,3].

As far as our understanding goes much has not been reported regarding the use of DGS for size reduction of microstrip antennas, although its application have been reported for harmonic reduction [4], cross-polarization suppression [5] and mutual coupling reduction [6] in antenna arrays etc. Although the back to back geometries have been reported by the various researchers [7,8] but here a new coupling method *i.e.* proximity coupling with the defected ground structure is used for the consideration of the increased bandwidth.

This paper presents the application of DGS for size reduction of microstrip antennas. A dumbbell shaped DGS is used in the common ground plane of a back to back combined single feed proximity coupled microstrip antenna.

2. Defected Ground Structure

Recently there has been an increasing interest in the use of DGSs for performance enhancement of microstrip antennas and arrays. These are realized by etching off a simple shape defect from the ground plane of the planer circuits.

Although various complicated DGSs were reported in the literature, but the simplest one is the dumbbell shaped DGS. **Figure 1(a)** shows the simple and mostly used dumbbell shaped DGS that is etched in the ground plane below the microstrip line, in which both the areas (L_g*W_g) and the slot gap (g) play a very important role to find the resonance behavior of the DGS.

The head areas (L_g*W_g) is very useful for the variation in the inductance (L) and the slot (g) produces the capacitance (C). The L and C may be calculated from the formulae given below [9].

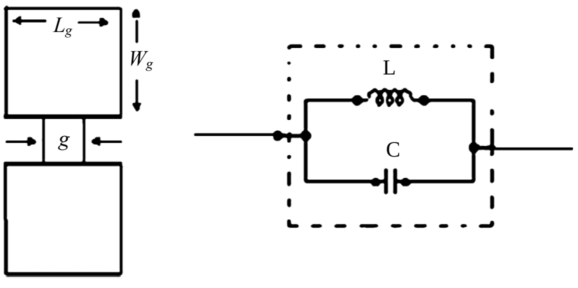

(a)

(b)

Figure 1. (a) Dumbbell shaped DGS, and (b) DGS Equivalent Circuit.

Copyright © 2011 SciRes.

$$\mathrm{L} = \frac{1}{4\pi^2 f_0^2 C} \qquad (1)$$

$$\mathrm{C} = \frac{f_c}{2Z_0} \times \frac{1}{2\pi(f_0^2 - f_c^2)} \qquad (2)$$

When this DGS is applied to the antenna, the equivalent inductive part due to the DGS increases and produces equivalently the high effective dielectric constant, thereby decreasing the resonant frequency when the DGS is incorporated in the ground plane of a micro strip antenna.

3. Antenna Design with DGS

In this section, the design approach and the performance of the basic antenna and the antenna with DGS is described. At the outset, the single patch antenna was designed and simulated using the CST Microwave studio [10], for the operating frequency at 5.0 GHz. Then another patch of the same size was added in the opposite side of the ground plane and fed in the same way as the first one. The configuration seems as two patch antennas having a common ground plane working at the same frequency. Next the feed lines were combined for the antenna for single feed design. For this purpose the antenna feed lines W_f = 0.934 mm with d_f = 1.5 mm were designed of 100 ohm and for matching to the 50 ohm transmission line (W_p = 3.86 mm) a quarter wave length transformer (W_t = 2 mm) was used to give proper matching (**Figure 2**).

The cross-sectional view of the single fed back to back combined proximity coupled antenna is shown in the **Figure 2(a)**. The layouts of matching networks only are emphasized in this **Figure 2(b)** for convenience. It is observed that the antenna designed in this configuration gives the bandwidth of 137 MHz whereas single antenna gives a bandwidth of 67 MHz. This is due to the fact that as the antenna height increases the quality factor decreases and the bandwidth increases. This becomes a multilayer antenna with more height and higher bandwidth as compared to the single patch antenna.

The simple transmission line model was used for the antenna size calculation. The dielectric constant was taken as 3.38 with the loss tangent 0.0025 and of 1.524 mm thickness. The patch lengths L_p and widths W_p are 15 mm and 19 mm respectively. The feed line has been inserted inside the dielectric at a height equal to the half of the height (h = 3.048 mm) of the antenna on either side. The dumbbell shaped DGS with dimensions L_g = W_g = 8.6 mm and g = 0.76 mm was created in the ground plane of the antenna as shown in the **Figure 2(c)**. A small slot was also created for making the antenna with single feed. The two feed lines were connected with a metal strip which goes through the small slot in the ground plane. The fabricated antenna is shown in **Figure 2(d)**.

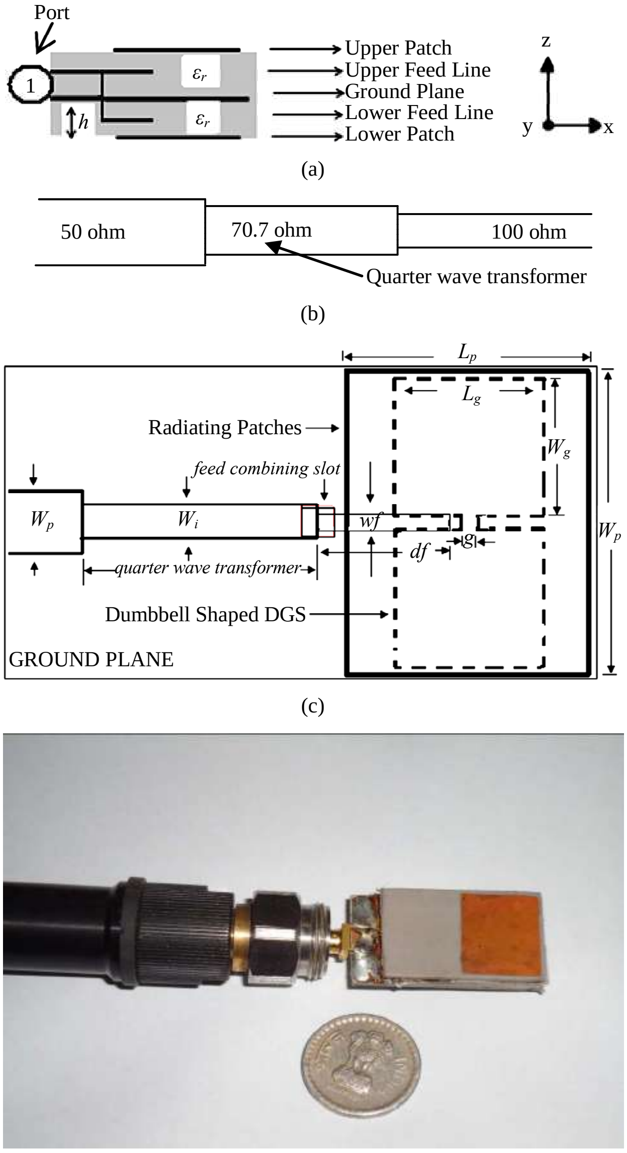

(a)

(b)

(c)

(d)

Figure 2. (a) Cross-sectional view of antenna configuration, (b) Feeding Network, (c) Top-View of the antenna, (d) Fabricated Antenna.

4. Results and Discussion

At first the antenna without the DGS in the common ground was simulated and was found to resonate at 5 GHz with 137 MHz Bandwidth.

Then the structure was simulated with the dumbbell shaped DGS. Before reaching to the final size of the DGS, a parametric study was done by varying L_g, W_g and g of the DGS. As shown in **Figure 3**, with the increase in the value of L_g, the resonant frequency of the antenna is decreasing. Infect, increase in the length of the DGS head gives increasing inductance which in turn decreases the resonant frequency of the antenna.

At this point the increment in the (g) was not possible due to the accuracy in fabrication, so for this reason the other dimension (g) was kept constant for the requirement of the desired frequency (2 GHz). The size of the DGS single square head for the antenna to resonate at 2 GHz (UHF Band) was found to be 8.6 mm × 8.6 mm. The return loss (S11 [dB]) plot of the structure with and without DGS is shown in **Figure 4(a)**. **Figure 4(b)** shows the measured S11 parameter using the HP 8720 B network analyzer. The marker's position near to peak shows the resonance frequency 2.08 GHz with return loss of -13 dB. The result shows good agreement with the simulation results. The measured -10 dB bandwidth is about 60 MHz. The maximum size reduction achieved is about 60%.

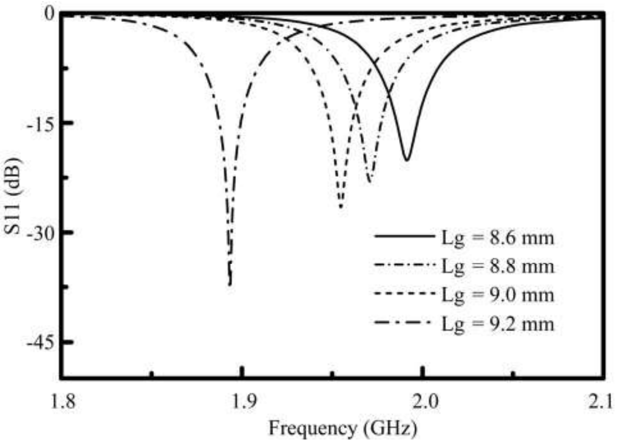

Figure 3. S11 Vs. Frequency response by varying DGS length (Lg).

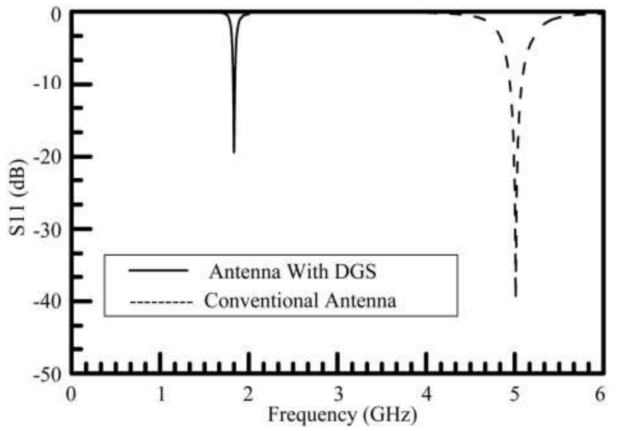

(a)

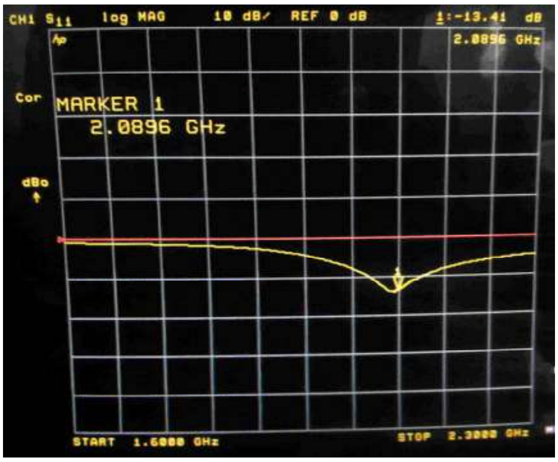

(b)

Figure 4. Antenna Return Loss (a) simulated (b) measured.

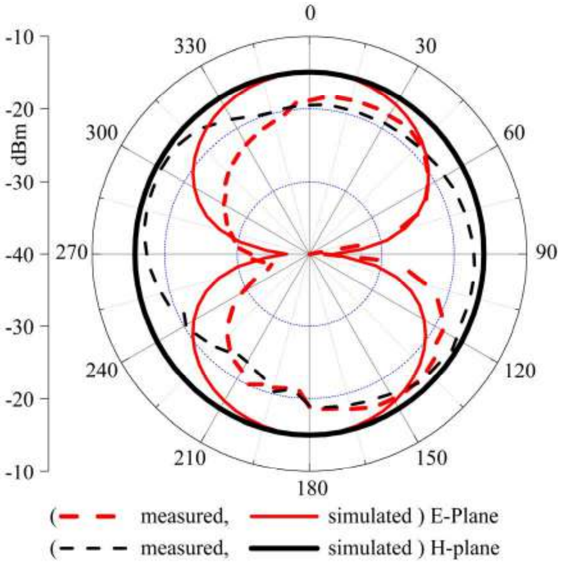

Figure 5. Antenna radiation pattern (measured and simulated) at 2 GHz Frequency.

The measured and simulated power patterns of the antenna are shown in **Figure 5**. It can be observed that the E-plane radiation pattern is similar to the pattern for a dipole antenna. Measurement errors are mainly due to the spurious radiation created by the feeding end and the improper coupling of the elements. However the gain measured experimentally for the proposed antennas with DGS is about -6.9 dB and -7.8 dB (where simulated gain with DGS is 3.8 dB and 5.7 dB is for the antenna without DGS, for both the planes) in both the E and H plane respectively, which is consistent with the size reduction of the antenna.

5. Conclusion

Microstrip patch antenna size reduction with DGS is carried out in this work. A dumbbell shaped DGS in the common ground plane of a back to back microstrip structure was found to give a size reduction of about 60% and shifts the resonance frequency from 5 GHz to 2 GHz, with 60 MHz bandwidth facilitating the antenna, to be used for UHF band applications.

REFERENCES

[1] R. Garg, P. Bhartia, I. Bahl and A. Ittipiboon, "Microstrip Antenna Design Handbook," Artech House, London, 1995.

[2] P. Salon, M. Keskilammi and M. Kivikoski, "Single Feed Dual Band Planer Inverted-F Antenna with U-Shaped Slot," *IEEE Transactions on Antenna and Propagation*, Vol. 48, No. 8, 2000, pp. 1262-1264. doi:10.1109/8.884498

[3] D.Nashhat, H. Elsadek and H. Ghali, "Dual-Band Reduced Size PIFA Antenna with U-Slot for Bluetooth and WLAN Applications," *Proceedings of IEEE Antennas Propagation International Symposium*, Columbus, Vol. 2, 22-27 June 2003, pp. 962-965. doi. 10.1109/APS.2003.1219395

[4] H. Liu, X. Sun and J. Mao, "Harmonic Suppression with Photonic Bandgap and Defected Ground Structure for a Microstrip Patch Antenna," *IEEE Microwave and Wireless Component Letters*, Vol. 15, No. 2, February 2005, pp. 55-56. doi:10.1109/LMWC.2004.842809

[5] D. Guha, M. Biswas and Y. M. M. Antar, "Microstrip Patch Antenna with Defected Ground Structure for Cross Polarization Suppression," *IEEE Antenna and Wireless Propagation Letter*, Vol. 4, No. 1, 2005, pp. 455-458. doi:10.1109/LAWP.2005.860211

[6] M. Salehi, A. Motevasselian, A. Tavakoli and T. Heidari, "Mutual Coupling Reduction of Microstrip Antenna using Defected Ground Structure," *IEEE Singapore International Conference on Communication System*, Singapore, 2006, pp. 1-5. doi:10.1109/ICCS.2006.301375

[7] J. Guterman, Y. R. Samii, A. A. Moreira and C. Peixeiro, "Quasi-Omni Directional Dual-band Back-to-Back E-shaped Patch Antenna for Laptop Applications," *Electronics Letters*, Vol. 42, No. 15, 2006, pp. 845-847. doi:10.1049/el:20061799

[8] H. Iwasaki, "A Back-to-Back Rectangular-Patch Antenna Fed by a CPW," *IEEE Transactions on Antenna Propagation*, Vol. 46, No. 10, October 1998, pp. 1527-1530. doi:10.1109/8.725285

[9] D. Ahn, J. S. Park, C. S. Kim, J. Kim, Y. Qian and T. Itoh, "A Design of the Low-Pass Filter Using The Novel Micro strip Defected Ground Structure," *IEEE Transactions on Microwave Theory and Technology*, Vol. 49, No. 1, 2001, pp. 86-91. doi:10.1109/22.899965

[10] EM Simulation Software CST Studio Suite™, v9.

Copyright © 2011 SciRes.

Journal of Electromagnetic Analysis and Applications, 2012, 4, 362-366
http://dx.doi.org/10.4236/jemaa.2012.49051 Published Online September 2012 (http://www.SciRP.org/journal/jemaa)

Vehicular Radio Scanner Using Phased Array Antenna for Dedicated Short Range Communication Service

Tapas Mondal[1], Priyanka Shishodiya[1], Rowdra Ghatak[2], Sekhar Ranjan Bhadra Chaudhuri[3]

[1]Department of Electronics & Communication Engineering, Dr. B. C. Roy Engineering College, Durgapur, India; [2]Microwave and Antenna Research Laboratory, Department of Electronics and Communication Engineering, National Institute of Technology Durgapur, Durgapur, India; [3]Department of Electronics & Telecommunication Engineering, Bengal Engineering and Science University, Shibpur, India.
Email: tapas2k@gmail.com

Received July 10th, 2012; revised August 11th, 2012; accepted August 21st, 2012

ABSTRACT

Now a day's accidents are very common due to increased population of vehicle. In order to ensure safety measures in the vehicle this paper has proposed some methodologies regarding careful driving by automatically scanning and analyzing the blind spot area of an intelligent mobile vehicle. A vehicular antenna with minimum perturbation is proposed to be fitted on the vehicle and collect information of the concern area which would ensure visibility of the operator *i.e.* masked or integrated within the car body. This paper has dealt with the design of Tchebyscheff polynomial based prototype planar microstrip phased array antenna and also redesigned the same when implemented in the body of the car being considered as an electromagnetically large element. Both the design has been experimentally verified with the measurement. The simulated and the measured results in both the cases are found to be in good agreement. More than 11 dB gain was observed at perfectly 30° angles from its broad side direction as desired for blind spot detection with minimum amount of electromagnetic interference inside the car.

Keywords: Intelligent Transportation System; Microstrip Phased Array Antenna; Tchebyscheff Polynomial; Electromagnetic Interference

1. Introduction

Intelligent Transport System (ITS) ensures mobility comfort and safety in transportation system. It also absorbs the hazards due to environmental impact. With the progress of the information processing technology, control systems to minimize accidents for the roadways have also been advanced hence one approach to improve the traffic safety is found and that is automatic collection of data by scanning the blind spot area of the vehicle [1] as shown in **Figure 1**. Many methods were proposed to detect the blind spot area but all of them had certain limitations. Devices like dynamic angling side view mirror [2], side view camera model [3], and shadow or edge features detector [4] were used for detecting blind spot area but their performance was affected during bad weather, fog or mist. Also we know that the mechanical systems, response time is more and the system is prone to wear and tear.

A radio frequency method has been proposed in this paper to scan the blind spot zone efficiently. Four rectangular microstrip antennas (RMSA) are arranged in linear configuration with optimal spacing between the patch elements to construct the phased array radar. A corporate feed network is used to feed the patch element unequally and a progressive phase shifter is designed with 108˚ delay elements to tilt the main beam in the desired direction and then the total unit is simulated and experimented after placing it on the car body which is electromagnetically a large element. The antenna works in the Dedicated Short-Range Communication Service (DSRCS) frequency band [5]. The design of the microstrip phased array antenna is discussed in Section 2 and after that to place the antenna; the design of the entire car is given in Section 3. Results and discussion are portrayed in Section 4 along with conclusion in Section 5.

2. Design of Microstrip Phased Array Antenna

The microstrip phased array antenna is designed for Dedicated Short Range Communication Service at 5.88 GHz with dielectric constant of 2.32 and substrate thickness of 0.785 mm. Firstly the dimension of the rectangular patch is computed as 16.322 mm by 19.8 mm using the method outlined in [6]. With computed inset feed length of 3.8 mm and for the above dimensions of the rectangular patch the return loss is found to be –8.09 dB

Copyright © 2012 SciRes.

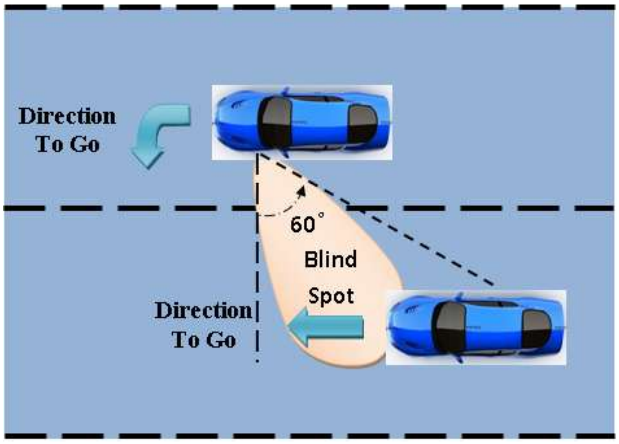

Figure 1. Surrounding regions of a vehicle.

with (30.1 + j26.7) ohm impedance at the feed position. The RMSA is simulated and optimized using Ansoft HFSS^{TM}. After optimization the final length is found to be 16.4363 mm with inset feed length of 4.8225 mm keeping the width of the patch unchanged as shown in **Figure 2**.

By considering the above designed patch element, a four element linear array is realized and powered by Tchebyscheff current distribution. The spacing between the elements is kept considering the desired maximum scan angle in order to eliminate the grating lobes within the visible space of the phased array antenna. To optimize the performance of the antenna in respect of its side lobe level (SLL), mutual coupling and gain of the antenna array, the spacing between the elements is studied parametrically [7]. The results of the said study are tabulated in **Table 1**.

After optimization it is found that the optimum spacing between the elements is 0.6 λ_0 while considering, the main lobe to side lobe ratio below 20 dB, optimum mutual coupling and overall gain.

From **Figure 1**, it is observed that the beam of the antenna array is required to be tilted by an angle 30° away from the broad side direction. In view of the above a progressive phase shifter of –108° is designed with the help of 11.16 mm feed line length. The actual line length is considered as mτ, where m = 0, 1, 2, 3.

To enhance the gain by maintaining the beam width and main lobe to side lobe ratio, the antenna elements are excited by Dolph Chebyshev current distribution. The array consist of four elements, thus third order Tchebyscheff polynomial is calculated. Hence the polynomial is solved and relative current ratio is computed as 1:1.7795:1.7795:1. Both equal and unequal power dividers are designed along with the progressive phase shifter.

For equal power division, a 3 dB equal power divider is designed whose vertical arm is of 50 Ω line and two horizontal quarter-wavelength branch-lines are of 70.71 Ω is used [8]. It is shown in **Figure 3(a)**.

Table 1. Effect on mutual coupling and gain due to variation in spacing (d) between the elements.

d with respect to guided wavelength	d with respect to operating wavelength	Mutual coupling (dB)	Gain (dBi)
0.6 λ_g	0.41 λ_0	>10	7.6
0.7 λ_g	0.48 λ_0	>13	9.9412
0.8 λ_g	0.55 λ_0	>20	10.647
0.9 λ_g	0.6 λ_0	>22	10.919

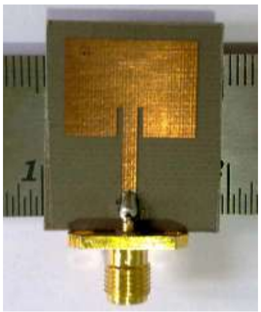

Figure 2. Photograph of the rectangular microstrip antenna element.

To get the unequal current distribution, both the horizontal arm of the microstrip unequal power divider should be of different resistance and it is calculated as 46.368 Ω and 82.49 Ω. The structure is shown in **Figure 3(b)**.

Generally for practical design purpose the gap between the corporate feed networks is taken small but the value of the gap less than 0.11 λ_0 disturbs the frequency response of overall antenna array. The effects of the feed network on the side-lobes are significant in the E-plane pattern. The influence on the H-plane is less important due to the orthogonality of the drive current and the symmetry of all the other currents in this cut [9]. By maintaining the SLL and the surface wave loss, the gaps are taken as 0.11 λ_0 as portrayed in **Figure 4**.

3. Placement of Antenna on the Vehicle

In order to compute the electromagnetic effect of the body of the car on the antenna performance, the entire structure of the car is designed in Ansoft HFSS^{TM}. The complete structure of the car consist of hood, roof, trunk, left and right side doors, left and right side quarter panel for front and rear side which are made up of conducting material but the bumpers and wheels are assigned by layered impedance due to their hard rubber and polyester materials as shown in **Figure 5**.

After going through the typical requirement for scanning the blind spot, the antenna is placed just beside the side view mirror. The following problems in the designed are addressed while simulating the effect of the radiation

Copyright © 2012 SciRes.

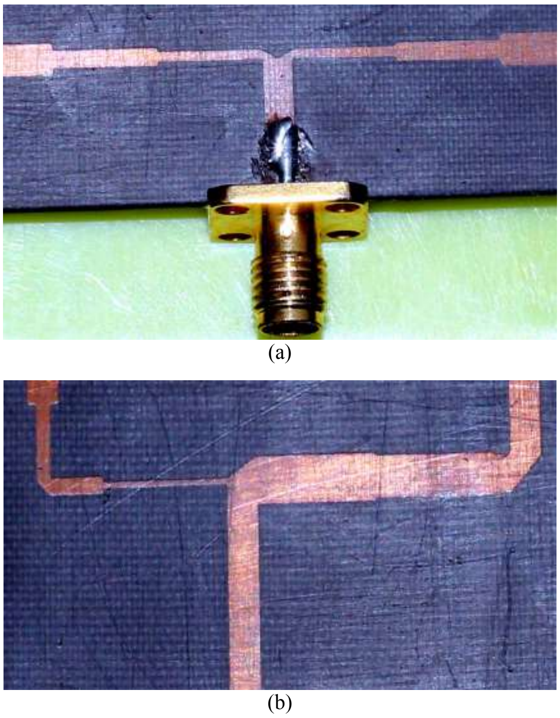
(a)

(b)

Figure 3. (a) Photograph of the equal microstrip power divider; (b) Photograph of the unequal microstrip power divider.

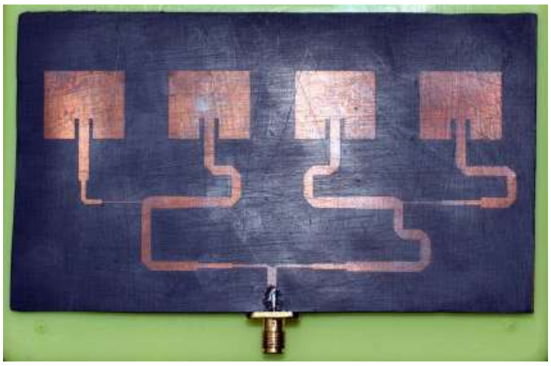

Figure 4. Photograph of the antenna array with corporate feed network.

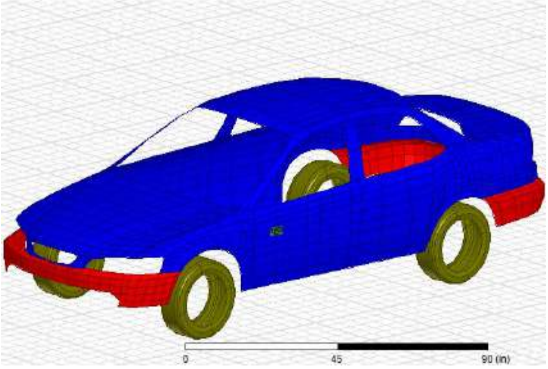

Figure 5. The entire structure of the car along with the antenna array.

characteristics of the antenna after placing it on the vehicle:

1) Spans a large computational domain compared to the wavelength;

2) Requires a fine grid resolution to resolve the detailed antenna structure;

3) Contains largely non-conformal PEC and dielectric structures.

4. Results and Discussion

For obtaining the scattering parameter characteristics, the microstrip antenna has been simulated over the frequency bandwidth ranging from 5.5 GHz to 6.5 GHz and two different circumstances is observed. The first experiments involve the design, simulation and measurement after fabrication of an individual element and then the array with the same without involving the effect of the car. Both of them should used for optimized performance.

The scattering parameter is measured by Agilent Technology Vector Network Analyzer model no. N5320A (10 MHz - 20 GHz) and the radiation pattern is measured by Hittite HMC-T2100 synthesized signal generator (10 MHz - 20 GHz) and Krytar 9000 B power meter (10 MHz - 40 GHz) with 9530 B power sensor (10 MHz - 20 GHz) as shown in **Figure 6**. From **Figure 7**, it is found

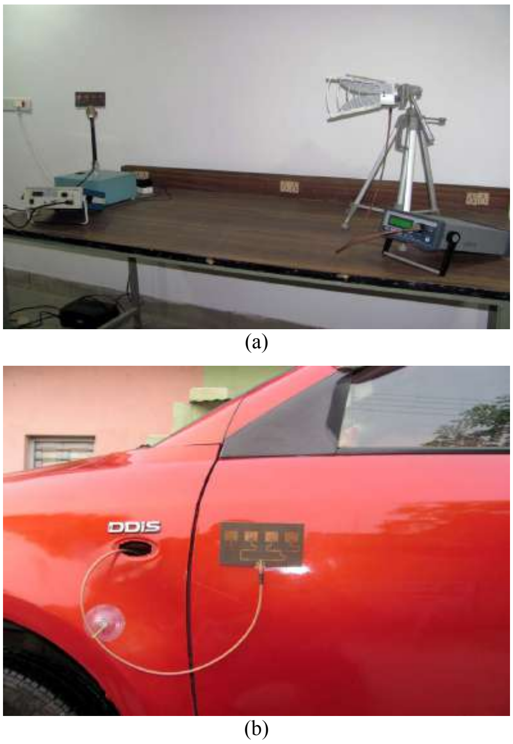
(a)

(b)

Figure 6. (a) Experimental set-up to measure the radiation pattern of the antenna array; (b) Experimental set-up to measure the radiation pattern of the antenna array after placing it on the car.

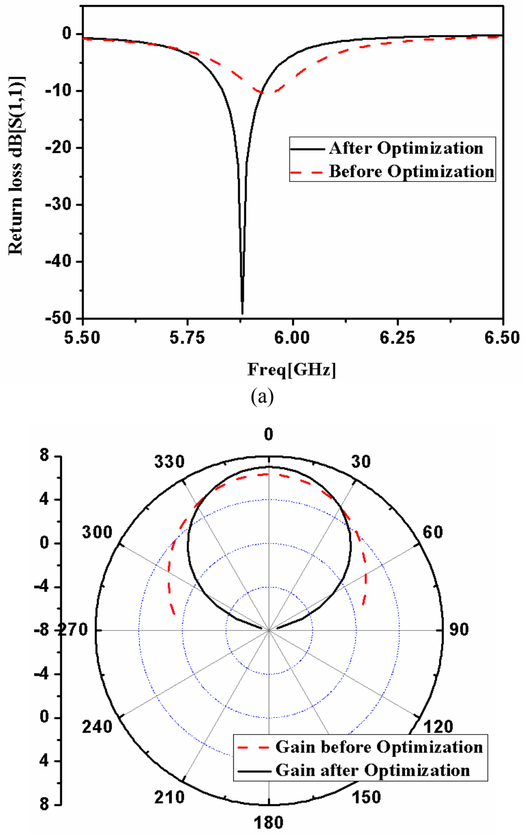

(a)

(b)

Figure 7. (a) Return loss a single patch before and after optimization; (b) Gain of a single patch before and after optimization.

that after optimization a perfect match is obtained exactly at 5.88 GHz and gain is also improved from 6.33 dB to 7.01 dB with a very good return loss of –47 dB. The current distribution and the radiation pattern of a single RMSA is portrayed in **Figure 8**.

A parametric study of the return loss of antenna array with different spacing between the elements of the corporate feed network is done and is compared in **Figure 9**. The results of the computational model of the entire antenna array with feed network were compared with the measured data and a good agreement is observed in **Figure 10**.

After getting the satisfactory results, next the same array antenna is experimented again with considering the electromagnetic effect of the car. In order to test the antenna array in association with the vehicle, a numerical model of the car is created using Ansoft HFSS^{TM} and the simulated data is compared with the measured in **Figure 11**. **Figure 12** graphically represents the 3D radiation pattern of the antenna array in presence of the computational model of the vehicle. From **Figure 13**, it is clear that the inside electric field strength is not more than 15 Volt per meter which is very less compare to the standard

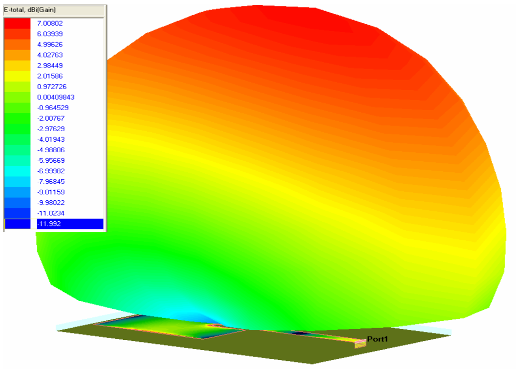

Figure 8. Current distribution and the radiation pattern of the rectangular microstrip antenna.

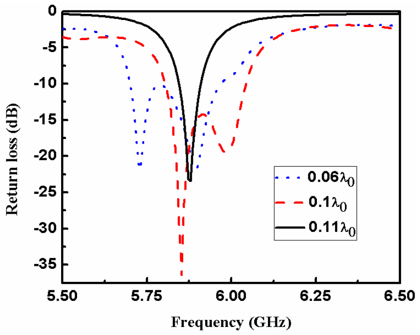

Figure 9. Parametric study of return loss of the antenna array with different spacing between the elements of corporate feed network.

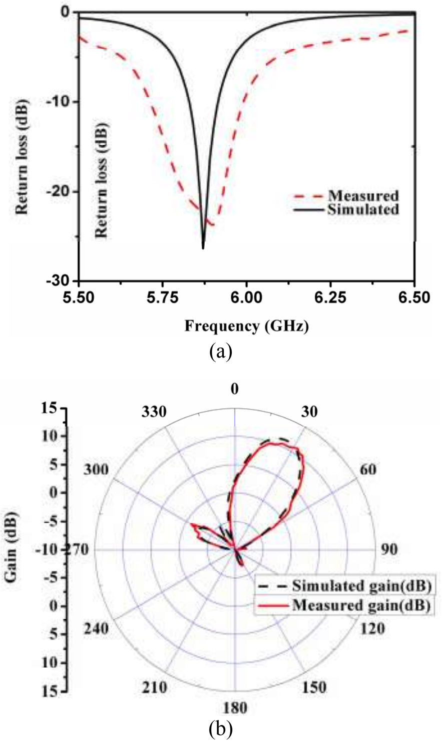

(a)

(b)

Figure 10. (a) Return loss of the antenna array; (b) Radiation pattern of the antenna array.

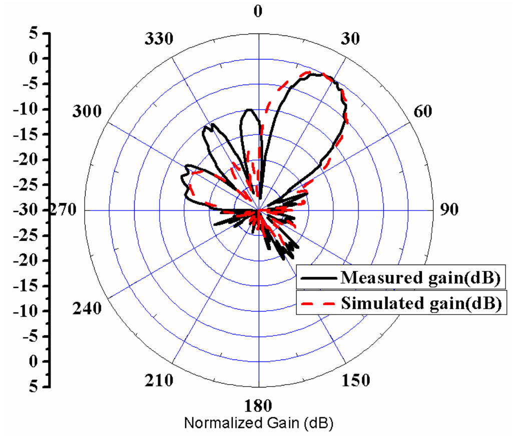

Figure 11. Normalized radiation pattern of the antenna array after placing it on the body of the car.

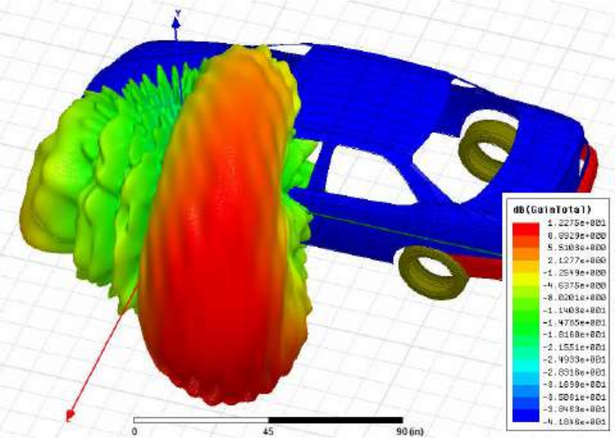
dB(GainTotal)
1.2275e+001
8.8929e+000
5.5103e+000
2.1277e+000
-1.2549e+000
-4.6375e+000
-8.0201e+000
-1.1403e+001
-1.4785e+001
-1.8168e+001
-2.1551e+001
-2.4933e+001
-2.8316e+001
-3.1698e+001
-3.5081e+001
-3.8463e+001
-4.1846e+001
0
45
90 (in)

Figure 12. 3D radiation pattern of the antenna array after placing it on a vehicle body.

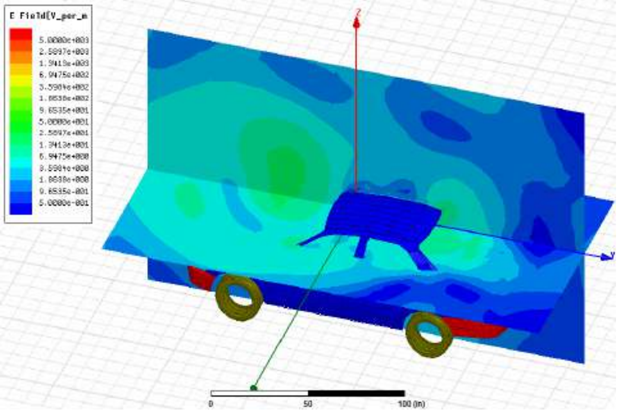
E Field[V_per_m
5.0000e+003
2.5897e+003
1.3413e+003
6.9475e+002
3.5984e+002
1.8638e+002
9.6535e+001
5.0000e+001
2.5897e+001
1.3413e+001
6.9475e+000
3.5984e+000
1.8638e+000
9.6535e-001
5.0000e-001
0
50
100 (in)

Figure 13. Inside electric field interference of the array after placing it on a vehicle body.

electromagnetic hazards.

5. Conclusion

A Tchebyscheff polynomial based microstrip phased array antenna is designed to detect the blind spot area of the intelligent mobile vehicle. The design is further simulated in Ansoft HFSS^{TM} after placing the antenna on the body of the car and the results are observed. The result of the computational model of the entire antenna array is compared with the measured data before and after placing the said antenna array on the body of the car and found to be in good agreement. An overall 11 dB gain is obtained while measuring in the desired direction with minimum amount of electromagnetic interference inside the car.

REFERENCES

[1] M. Griffiths, “Intelligent Transport System,” Postnote: Parliamentary Office of Science and Technology, London, 2009. www.parliament.uk/paliamentary_offices/post/

[2] J. Kuwana and M. Itoh, “Dynamic Angling Side-View Mirror for Supporting Recognition of a Vehicle in the Blind Spot,” *International Conference on Control*, *Automation and Systems*, *ICCAS*, Seoul, 14-17 October 2008, pp. 2913-2918.

[3] S. Jeong, S.-W. Ban and M. H. Lee, “Autonomous Detector Using Saliency Map Model and Modified Mean-Shift Tracking for a Blind Spot Monitor in a Car,” *Seventh International Conference on Machine Learning and Applications*, San Diego, 11-13 December 2008, pp. 253-258. doi:10.1109/ICMLA.2008.129

[4] B.-F. Wu, C.-C. Kao, Y.-F. Li and M.-Y. Tsai, “A Real-Time Embedded Blind Spot Safety Assistance System,” *International Journal of Vehicular Technology*, Vol. 2012, 2012, pp. 1-15. doi:10.1155/2012/506235

[5] A. Tang and A. Yip, “Collision Avoidance Timing Analysis of DSRC-Based Vehicles,” *Accident Analysis & Prevention*, Vol. 42, No. 1, 2010, pp. 182-195. doi:10.1016/j.aap.2009.07.019

[6] T. Mondal, J. S. Roy and S. R. Bhadra Chaudhuri, “Phased Array Antenna Design for Intelligent Transport Systems,” *IEEE International Workshop on Antenna Technology*: *Small Antennas and Novel Metamaterial* (i*WAT*), Santa Monica, 2-4 March 2009, pp. 1-4.

[7] T. Mondal, R. Ghatak and S. R. B. Chaudhuri, “Design and Analysis of a 5.88GHz Microstrip Phased Array Antenna for Intelligent Transport Systems,” *IEEE International Sumposium on Antennas and Propagation and CNC*/*USNC*/*URSI Radio Science Meeting*, Toronto, 11-17 July 2010, pp. 1-4.

[8] H. J. Visser, “Array and Phased Array Antenna Basics,” John Wiley & Sons Ltd., Chichester, 2005. doi:10.1002/0470871199

[9] E. Levine, G. Malamud, S. Shtrikman and D. Treves, “A Study of Microstrip Array Antennas with the Feed Network,” *IEEE Transactions on Antennas and Propagation*, Vol. 31, No. 4, 1989, pp. 426-434. doi:10.1109/8.24162

Copyright © 2012 SciRes.

Wireless Engineering and Technology, 2012, 3, 175-180
http://dx.doi.org/10.4236/wet.2012.33025 Published Online July 2012 (http://www.SciRP.org/journal/wet)

Plus Shape Slotted Fractal Antenna for Wireless Applications

S. Jagadeesha[1], R. M. Vani[2], P. V. Hunagund[3]

[1]S.D.M. Institute of Technology, Ujire Mangalore, India; [2]University Science Instrumentation Center, Gulbarga University, Gulbarga, India; [3]Department of PG Studies and Research in Applied Electronics, Gulbarga University, Gulbarga, India.
Email: Jagadeesh.sd69@gmail.com, vanirm12@rediffmail.com, prbhakar_hunagund@yahoomail.com

Received April 28th, 2012; revised May 26th, 2012; accepted June 3rd, 2012

ABSTRACT

Fractal antennas are characterized by space filling and self-similarity properties which results in considerable size reduction and multiband operation as compared to conventional microstrip antenna. This paper outlines a multiband antenna design based on fractal concepts. Fractal antennas show multiband behavior due to self-similarity in their structure. The plus shaped fractal antenna has been designed on a substrate of dielectric constant $€_r = 4.4$ and thickness 1.6 mm. The proposed antenna is characterized by a compact size and it is microstrip feed fractal patch of order 1/3. It is observed that the antenna is radiating at multiple resonant frequencies. The resonant frequency is reduced from 2.2 GHz to 900 MHz after I & II iterations respectively. Thus considerable size reduction of 81.77% & overall bandwidth of 12.92% are achieved. The proposed antenna is simulated using the method of moment based commercial software (IE3D) and it is found that simulated results are in good agreement with the experimental results.

Keywords: Fractal Antenna; Multi Frequency; Size Reduction; Wireless Application; Plus Shape Antenna; Slotted Antenna

1. Introduction

In the study of antennas, fractal antenna theory is a relatively new area. The emergence of antennas with fractal geometries has given an answer to two of the main limitations started by Werner (1999) of the classical antennas, which are single band performance and dependence between size and operating frequency. The term "fractal" means broken or irregular fragments. It was originally coined by Mandelbrot (1983) to describe a family of complex shapes that possess an inherent self-similarity or self-affinity in their geometrical structure. Jaggered (1990) defined fractal electrodynamics as an area in which fractal geometry was combined with electromagnetic theory for the purpose of investigating a new class of radiation, propagation and scattering problems. One of the most promising area fractal electrodynamics researches is in its application to antenna theory and design. There are varieties of approaches that have been developed over the years, which can be utilized to archive one or more of these design objectives. The development of fractal geometry came largely from an in depth study of the pattern nature, with the advance of wireless communication system and their increasing importance wide band and low profile antennas are in great demand for both commercial and military applications [1]. A fractal is a rough or fragmented geometric shape that can be split into parts, each of which is a reduced-size copy of the whole and this property is called self-similarity. Fractal [2] geometries are composite designs that repeat themselves or their statistical characteristics and are thus "self-similar" fractal geometry finds a variety of applications in engineering. Fractal geometry is space filling contours of regular or irregular shapes [3-6], and is super imposed of too much iteration and they describe the self-similar property of fractal geometry [7]. Fractals are a class of shapes which have not characteristic size. Each fractal is composed of multiple iterations of a single elementary shape the iteration can continue infinitely, thus forming a shape within a finite boundary but of infinite length or area. Fractal has the following features 1) It has a finite structure at arbitrarily small scales; 2) It is too irregular to be easily described in traditional Euclidean geometric; 3) It is self-similar; 4) Simple and recursive [8]. Modern telecommunication systems require the antenna with wider bandwidth and smaller dimension than conventionally possible. This has initiated antenna research in various directions, are of which is by using fractal shaped antenna elements. In recent years several fractal geometries have been introduced for antenna application with varying degree of success in improving antenna

Copyright © 2012 SciRes.

characteristics. Some of these geometries have been particularly useful in reducing the size of the antenna, while other designs aim at incorporating multiband characteristics. These are low profile antennas with moderate gain and can be made operative at multiple frequency bands and hence are multifunctional [9]. In our present work we focus on generation of multifrequency which yields increases the bandwidth and size reduction of antenna. A plus shape patch is taken as a base shape and in first iteration four other plus shape patches of the order of 1/3 of base shape are placed touching the base shape. Similarly second iterations are taken by further placing plus shaped patches at even reduced scales. It is found that as the iteration number and iteration factor increases, the resonance frequencies become lower than those of the zero iteration, which represents a conventional plus shape patch.

2. Design Consideration

The base shape of the plus shaped slotted fractal antenna is designed on a dielectric substrate having a relative dielectric constant €$_r$ = 4.4 and thickness 1.6 mm as shown in **Figure 1**. This is the reference antenna or base shape antenna. Further this base shape antenna is modified by inserting horizontal slots on both sides with respect to center of patch as shown in **Figure 2** and it is named as antenna1. The length of the slot Ls is varied on either side of the edge as 5 mm, 10 mm, 15 mm, 20 mm, 21.175 mm, 21.675 mm and the frequency variation has been studied. The optimum length obtained is Ls = 21.675 mm *i.e.*, the distance between slots q = 2 mm is considered for further design.

The first iteration patch is designed with four plus shapes of order (1/3) of base shape are placed touching the base shape as shown in **Figure 3** and it is named as antenna 2 and same procedure is repeated for second iteration. This antenna is as shown in **Figure 4** and it is named as antenna 3. For each iteration plus shapes of the order of $(1/3)^n$ of base shape are taken, where n is the number of iterations. The dimension of first iteration can be calculated as

e = (1/3) a & g = (1/3) c also f = (1/3) b & h = (1/3) d.

i = (1/3) e & k = (1/3) g also j = (1/3) f & L = (1/3) h

So with optimized design the dimensions obtained are a = 45.3 mm, b = 15.1 mm, c = 35.4 mm, d = 11.8 mm. The length of the slot is Ls = 21.675 mm and width of the slot Ws *i.e.* r = 2 mm. The dimension of the ground plan is 55 mm × 85 mm. A 50 ohm SMA connector is used to feed the antenna by using microstrip feed technique. Optimized microstripline with following dimension, m = 0.5 mm, n = 18.55 mm, o = 3.05 mm, p = 18.4 mm. The suitable feed location is obtained through optimization process by using the IE3D software. The fabricated photographic view of all proposed antennas is shown from **Figures 5(a)-(e)**.

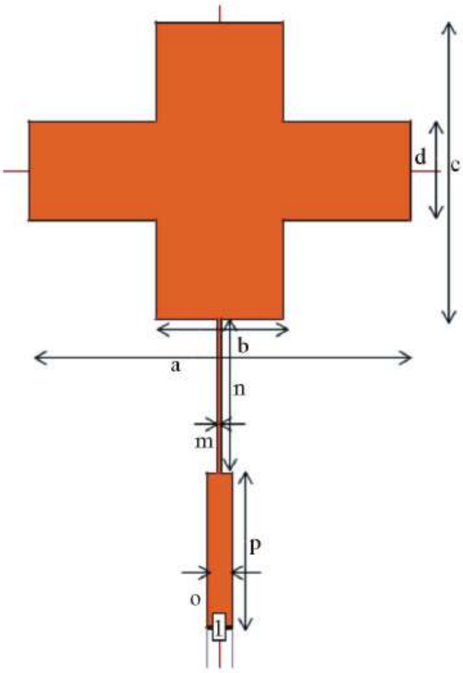

Figure 1. Geometry of base antenna.

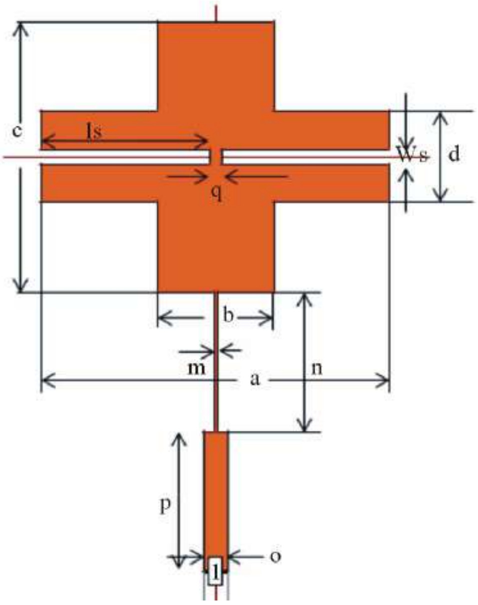

Figure 2. Geometry of antenna 1.

3. Results and Discussion

The characteristic of the fractal antenna with slot and with iterations has been studied by using IE3D software. Also the results have been verified practically with by using Vector Network Analyzer model Rohde and schewarz, German make ZVK model No.8651.

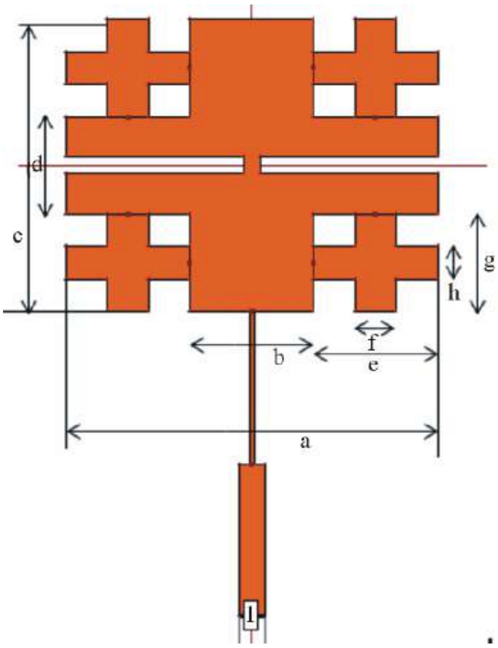

Figure 3. Geometry of base antenna 2.

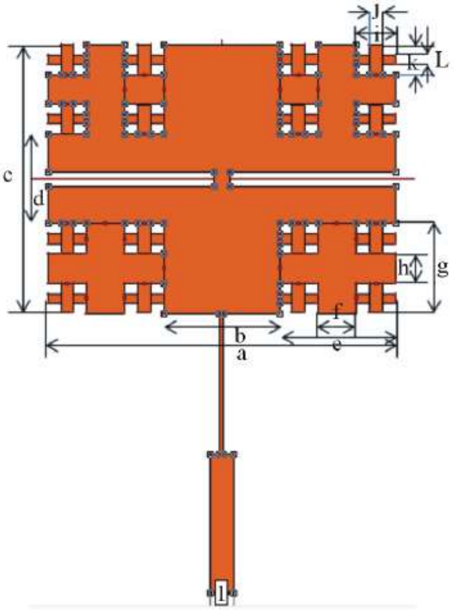

Figure 4. Geometry of antenna 3.

The modified base shape antenna with slots as been optimized by varying slot length Ls. The variation of Ls with resonant frequency of the antenna is shown in **Table 1** and same is presented in graphical form in **Figure 6**. From the tabular results it is found that by increasing slot length Ls on both sides from the edge of the patch, resonant frequency decreases. The lowest possible resonant frequency 1.27 GHz is obtained for Ls = 21.675 mm (*i.e.*, distance between the slots q = 2 mm) & this is taken as optimized length of the slot for further iteration.

The results of all proposed antennas are shown in **Table 2**. The simulated and measured return loss characteristics of proposed antennas are shown from **Figures 7(a)-(d)**.

From the results it is clear that the resonant frequency of the antenna 1 *i.e.*, modified base antenna with slot is fr = 1.27 GHz which is lower compared to the base antenna without slot (fr = 2.199 GHz). So the size reduction obtained is 66.85%. The antenna 2 *i.e.* the modified antenna with slots and first iteration gives multiple bands with lower frequency of 0.99 GHz. The size reduction obtained for antenna 2 is 79.88%. Further antenna 3 *i.e.*, with slot and second iteration gives multiple bands with

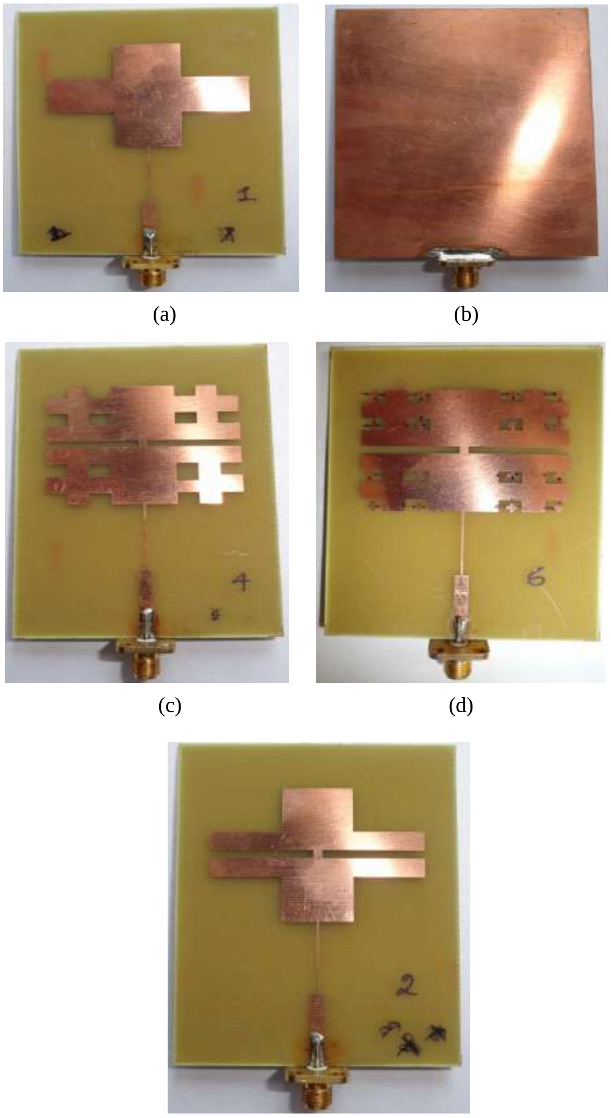
(a) (b)

(c) (d)

(e)

Figure 5. (a) Photograph of top view of base antenna; (b) Photograph of bottom view of base antenna; (c) Photograph of antenna 1; (d) Photograph of antenna 2; (e) Photograph of antenna 3.

Table 1. Variation of slot length v/s resonant frequency.

Slot length (Ls) variation in mm	Resonating frequency (GHz)
0	2.22
5	2.197
10	2.155
15	1.93
20	1.436
12.175	1.34
21.675	1.27

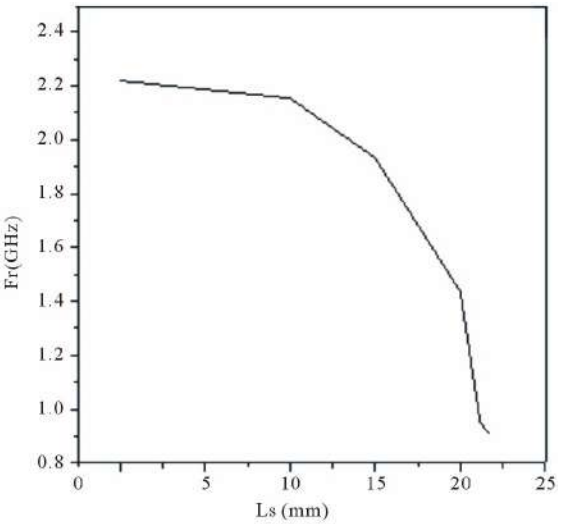

Figure 6. Variation of resonant frequency v/s slot length.

lower resonant frequency of 0.90 GHz. The size reduction obtained for antenna 3 is 81.77% which is more compared to all other proposed antennas.

Further the bandwidths of proposed antennas have been studied through simulation and measurements and the results are shown in **Table 2**. From the results it is clear that the bandwidths of modified antenna with slot, first & second iteration are more compared to base antenna. The measured bandwidth of antenna 1 is 106 MHz (4.895%), antenna 2 is 190 MHz (11.93%) and antenna 3 is 215 MHz (12.92%). The radiation patterns of all proposed antennas are studied and all are giving broadside radiation. Splitting of beam *i.e.* dip is found in **Figure 8(a)** of simulated radiation pattern of base shape at 2.19 GHz. In **Figure 8(b)** splitting of beams merged into broadside pattern at 0.91 GHz at second iterations this results into broadside radiation pattern.

4. Conclusion

This paper presents a new plus shape slotted fractal antenna with first and second iterations. The antenna.3 *i.e.*, slotted fractal antenna with second iteration gives size reduction of 81.77% and band width of 12.92% with broad side radiation pattern. So from the results we conclude that the modified base antenna with slots of second iterations gives a good size reduction and enhanced band width compared to that of modified base antenna with

Table 2. Results of proposed antennas.

Prototype antenna	Resonant frequency fr (GHz)		Return loss (db)		Bandwidth (MHz)		Overall bandwidth (MHz)	
	Sim	Pract	Sim	Pract	Sim	Pract	Sim	Pract
Base antenna	f1 – 2.19	f1 ± 2.19	–19	–24	53	53	79	71
	f2 – 3.42	f2 ± 2.46	–14	–15	26	18		
Antennal with q = 2 mm	f1 ± 1.19	f1 ± 1.27	15.8	–12.8	20	18	133	106
	f2 ± 2.52	f2 ± 2.48	18.4	–20.1	113	44		
		f3 ± 2.56		–18.7		44		
Antennal 2 with q = 2 mm (1st itr)	f1 ± 0.96	f1 ± 0.99	16.7	–14.5	22	80	136	190
	f2 ± 2.9	f2 ± 2.97	–14	–26.9	44	50		
	f3 ± 3.08	f3 ± 3.17	21.5	–21	70	60		
Antennal 3 with q = 2 mm (2nd itr)	f1 ± 0.91	f1 ± 0.990	15.2	–14.2	11	87	81	215
	f2 ± 3.07	f2 ± 2.976	19.4	–17.6	70	50		
		f3 ± 3.168		–18.9		78		

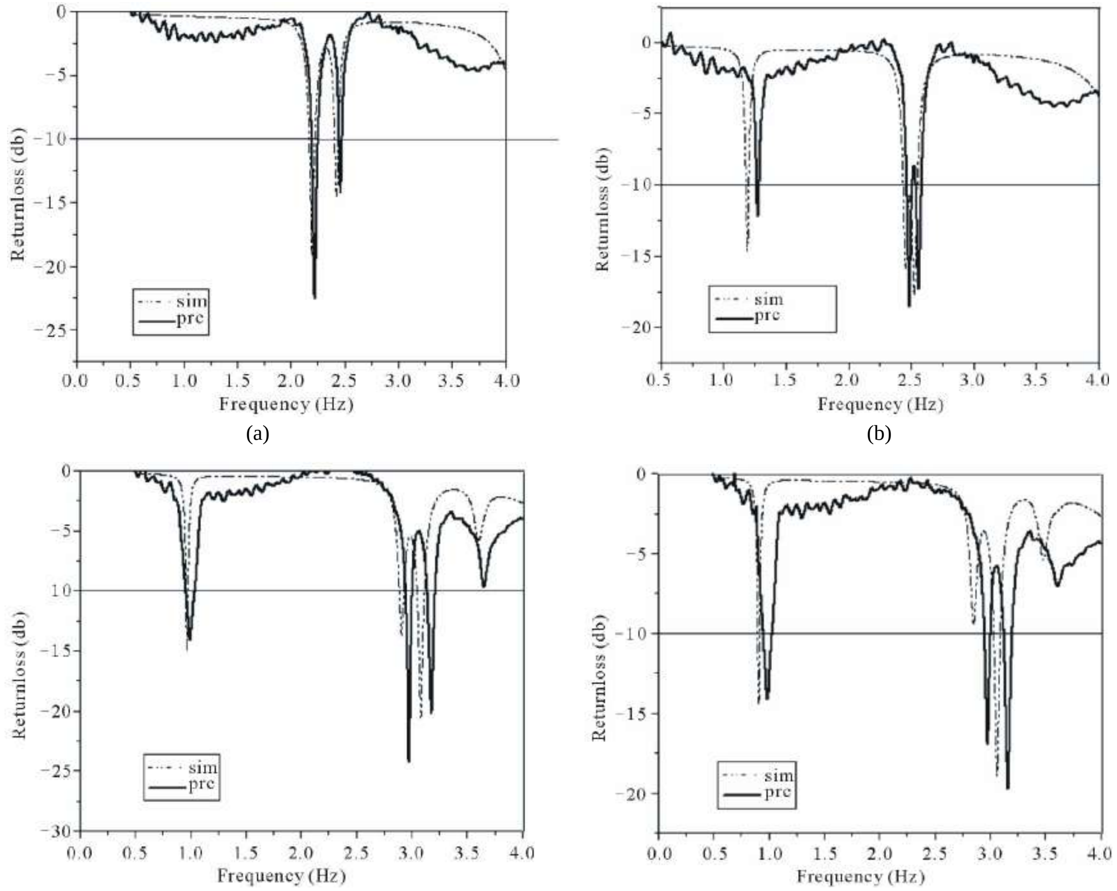

(a) (b)

(c) (d)

Figure 7. (a) Return loss characteristic of antenna 1 with q = 2 mm; (b) Return loss characteristic of base antenna 2 with q = 2 mm; (c) Return loss characteristic of antenna 2 with q = 2 mm; (d) Return loss characteristic of antenna 3 with q = 2 mm.

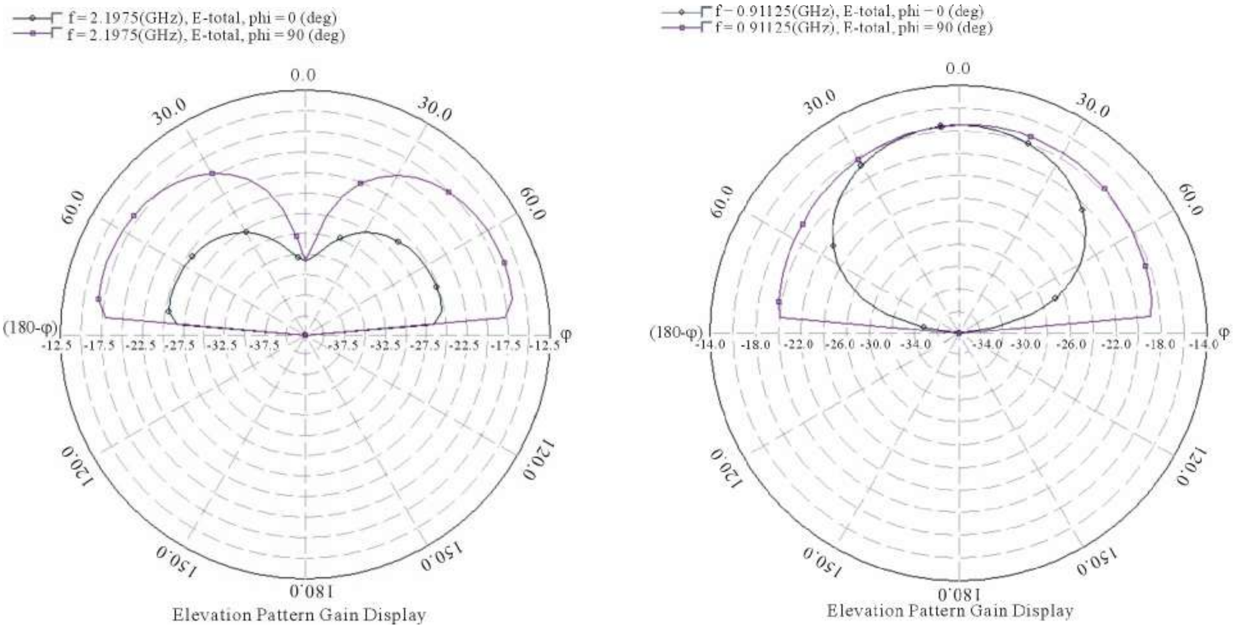

(a) (b)

Figure 8. (a) Simulated radiation pattern of base shape at 2.19 GHz; (b) Simulated radiation pattern at 0.91 GHz for second iteration.

slot of first iteration. These antennas may find application in wireless communication systems.

REFERENCES

[1] K. Sing, V. Grewal and R. Saxena, “Fractal Antennas: A Novel Miniaturization Technique for Wireless Communications,” *International Journal of Recent Trends in Engineering*, Vol. 2, No. 5, 2009, pp. 172-176.

[2] B. Mandelbrot, “Fractal: Form, Chance and Dimension,” W. H. Freeman and Company, San Francisco, 1997.

[3] H. O. Peitgzen, H. Jurgens and D. Sanpe, “Chaos and Fractals,” New Frontiers in Science, Springer-Verlag, New York, 1992.

[4] C. Punete, J. Romen, R. Pous, X. Garkia and F. Beitez, “Fractal Multiband Antenna Based on the Serpinski Gasket,” *Electronics Letters*, Vol. 32, No. 1, 1996, pp. 1-2. doi:10.1049/el:19960033

[5] J. P. Gianvittorio and Y. Rahmat-Sami, “Fractal Antennas: Novel Antenna: A Novel Antenna Miniaturization Technique and Applications,” *IEEE Antenna and Propagation Magazine*, Vol. 44, No. 1, 2002, pp. 20-36. doi:10.1109/74.997888

[6] D. H. Werner and S. Ganguly “An Overview of Fractal Antenna Engineering Research,” *IEEE Antenna and Propagation Magazine*, Vol. 45, No. 1, 2003, pp. 38-57.

[7] G. Srivatsun, S. S. Rani and G. S. Krishnan, “A Self-Similar Fractal Cantor Antennas for MICS Based Wireless Applications,” *Wireless Engineering Technology*, Vol. 2, No. 2, 2011, pp. 107-111. doi:10.4236/wet.2011.22015

[8] W. Shalan and K. Pahwa, “Multiband Microstrip Rectangular Fractal Antenna for Wireless Applications,” *International Journal of Electronics Engineering*, Vol. 3, No. 1, 2011, pp. 103-106.

[9] F. J. Jibrael and M. H. Hammed, “A New Multiband Patch Microstrip Plusses Fractal Antenna for Wireless Applications,” *ARPN Journal of Engineering and Applied Sciences*, Vol. 5, No. 8, 2010, pp. 155-158.

Copyright © 2012 SciRes.

PUBLISHED BY

INTECH
open science | open minds

World's largest Science,
Technology & Medicine
Open Access book publisher

3,250+
OPEN ACCESS BOOKS

106,000+
INTERNATIONAL
AUTHORS AND EDITORS

112+ MILLION
DOWNLOADS

BOOKS
DELIVERED TO
151 COUNTRIES

AUTHORS AMONG
TOP 1%
MOST CITED SCIENTIST

12.2%
AUTHORS AND EDITORS
FROM TOP 500 UNIVERSITIES

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED
WEB OF SCIENCE™

Selection of our books indexed in the
Book Citation Index in Web of Science™
Core Collection (BKCI)

Chapter from the book *Advances in Satellite Communications*

Downloaded from: http://www.intechopen.com/books/advances-in-satellite-communications

Interested in publishing with InTechOpen?
Contact us at book.department@intechopen.com

7

New Antenna Array Architectures for Satellite Communications

Miguel A. Salas Natera et al.*
Universidad Politécnica de Madrid,
Spain

1. Introduction

Ground stations which integrate the control segment of a satellite mission have as a common feature, the use of large reflector antennas for space communication. Apart from many advantages, large dishes pose a number of impairments regarding their mechanical complexity, low flexibility, and high operation and maintenance costs. hus, reflector antennas are expensive and require the installation of a complex mechanical system to track only one satellite at the same time reducing the efficiency of the segment (Torre et al., 2006).
With the increase of new satellite launches, as well as new satellites and constellation of low earth orbit (LEO), medium earth orbit (MEO), and geostationary earth orbit (GEO), the data download capacity will be saturated for some satellite communication systems and applications. Thus, the feasibility of other antenna technologies must be evaluated to improve the performance of traditional earth stations to serve as the gateway for satellite tracking, telemetry and command (TT&C) operation, payload and payload message or data routing (Tomasic et al., 2002). One alternative is the use of antenna arrays with smaller radiating elements combined with signal processing and beamforming (Godara, 1997).
Main advantages of antenna arrays over large reflectors are the higher flexibility, lower production and maintenance cost, modularity and a more efficient use of the spectrum. Moreover, multi-mission stations can be designed to track different satellites simultaneously by dividing the array in sub-arrays with simultaneous beamforming processes. However, some issues must be considered during the design and implementation of a ground station antenna array: first of all, the architecture (geometry, number of antenna elements) and the beamforming process (optimization criteria, algorithm) must be selected according to the specifications of the system: gain requirements, interference cancellation capabilities, reference signal, complexity, etc. During implementation, deviations will appear as compared to the design due to the manufacturing process: sensor location deviation and sensor gain and phase errors (Martínez & Salas, 2010). In an antenna array, the computation of a close approach of the direction of arrival (DoA) and the correct performance of the beamformer depends on the calibration procedure implemented.

---

* Andrés García-Aguilar, Jonathan Mora-Cuevas, José-Manuel Fernández González, Pablo Padilla de la Torre, Javier García-Gasco Trujillo, Ramón Martínez Rodríguez-Osorio, Manuel Sierra Pérez, Leandro de Haro Ariet and Manuel Sierra Castañer.
Universidad Politécnica de Madrid, Spain

This chapter is organized with the following sections. Section 2, introduces the relationship between applications and antenna design architectures. Section 3, introduces the new antenna array architectures for satellite communication including motivation and explains experimental examples. Section 4, explains adaptive antenna array and receiver architectures for adaptive antennas systems considering the beamforming with synchronization algorithms. Finally, Section 5 explains the A3TB concept.

2. Applications and antenna design architectures

In recent effort, new antenna array architectures have been under analysis and development. In (Tomasic et al., 2002) a highly effective, multi-function, low cost spherical phased array antenna design that provides hemispherical coverage is analyzed. This kind of novel architecture design, as the geodesic dome phased array antenna (GDPAA) presented in (Tomasic et al., 2002) preserves all the advantages of spherical phased array antennas while the fabrication is based on well-developed, easily manufacturable, and affordable planar array technology (Liu et al., 2006; Tomasic, 1998). This antenna architecture consists of a number of planar phased sub-arrays arranged in an icosahedral geodesic dome configuration.

In contrast to the about 10 *m* diameters dome of the GDPAA, there is the geodesic dome array (GEODA) (Sierra et al., 2007) with 5 *m* diameters dome. This antenna, presented in Fig. 1, has two geometrical structure parts. The first one, is based on a cylinder conformed by 30 triangular planar active arrays, and the second is a half dodecahedron geodesic dome conformed by 30 triangular planar active arrays. The GEODA is specified in a first version for satellite tracking at 1.7 GHz, including multi-mission and multi-beam scenarios (Martínez & Salas, 2010). Subsequently, the system of the GEODA has been upgraded also for transmission (Arias et al., 2010).

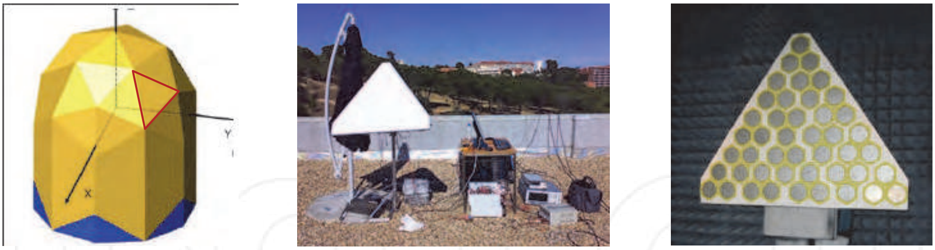

a b c

Fig. 1. a) The GEODA, b) The active sub-array demonstration, and c) The 45 elements planar active sub-array.

The antenna arrays technology in the user segment for satellite communications will substitute reflectors providing a more compact and easy to install antenna system, which is an interesting solution e.g. for satellite on the move (SOTM) system. There is a great diversity of solutions for fixed and mobile satellite communication systems including a large number of applications. Inmarsat broadband global area network (Inmarsat-BGAN) (Franchi et al., 2000) is the most representative example among mobile satellite systems (MSS), which gives land, maritime and aeronautical high speed voice and data services with global coverage using GEO satellites at L-band.

MSS services are divided into two groups, those that offer a regional coverage usually with GEO satellites, and those which offer a global coverage based on LEO or MEO satellite constellations. Depending on the coverage, there are some examples for MSS with regional coverage as the mobile satellite system (MSAT) in EEUU, Canada and South America, Optus in Australia, N-Star in Japan, Asia cellular satellite (ACeS) in Asia or Thuraya in the Middle East and in the North of Africa. While for MSS of global coverage there are some examples as Iridium, ICO Global Communications, Globalstar, Teledesic, etc. (Evans, 2009; Wu, 1994). Most of the MSSs work at L and S band, new applications on satellite to mobile terminal links work at X, Ku and Ka band, and satellite to base station connections work at L, S and C band. A number of applications is broad and lead terrestrial telecommunications market to offer a wider coverage: high speed voice and data (internet access, SMS, VoIP), digital video broadcasting by satellite 2 (DVB-S2) and digital video broadcasting satellite services to handhelds (DVB-SH), global position system (GPS) and Galileo, security, control and machinery monitoring on ships and aircrafts, teleeducation or telemedicine.

These modern satellite communications systems require new antenna solutions for base stations, aeronautical applications or personal communications services (PCS) on-the-move (Fujimoto & James, 2001). Within these applications, antenna array systems are potentially the best choice due to, as discussed above, its capability to perform electronically steering or beamforming, increase the antenna gain, and conform over curved or multifaceted surfaces the radiating elements. Portable antennas for PCS must be easy to install and mechanically robust, besides compact and lightweight (García et al., 2010) as the antenna array presented in Fig. 4.a. The design of antenna systems to provide high data rates for reliable PCS boarded on ships is not so strict in term of the geometrical requirements because it does not have space limitations (Geissler et al., 2010). However, in the case of land or airborne vehicles, geometrical and mechanical constraints are more severe. Antennas for terrestrial vehicles must be low profile, and for airborne vehicles aerodynamic shapes must be considered (Baggen et al., 2007; Vaccaro et al., 2010). Moreover, for the civil market conformal antenna arrays (Schippers, 2008; Kanno et al., 1996), or multi-surface arrays (Khalifa & Vaughan, 2007) are suitable choices to deal with the system aesthetic partiality.

Technological challenges have been faced during the implementation of satellite communication systems in the last decades. The design of a Test-Bed flexible and modular for testing or debugging beamforming algorithms and receiver architectures is an invaluable contribution in the educational, research and development area on satellite communication systems. The adaptive antenna array Test-Bed (A3TB) concept is based on the use of antenna arrays with beamforming capability to receive signals from LEO satellites (Salas et al., 2008). The scope of the A3TB is to probe the concept of antenna arrays applied to ground stations instead of reflectors for different applications, such as telemetry data downloading. It is also a good chance for Universities and Research Centers aiming to have their own ground station sited in their installations.

The A3TB ground station relies on the use of an antenna array to smartly combine the received signals from the satellite thanks to the implementation based on software defined radio (SDR) technology. The advantages of the SDR implementation is that A3TB architecture can be used to process any received signal from LEO satellites in the band imposed by the radio frequency (RF) circuits. Moreover, most of the processing is performed in software, so that appropriate routines can be used to process any received signal. The A3TB can be used to analyze the feasibility of different receivers and beamformer algorithms, regarding the capability to switch the receiver architecture in terms of the synchronizer algorithm configuration (Salas et al., 2007).

The current version has been developed to track The National Oceanic and Atmospheric Administration (NOAA) satellites in the very high frequency (VHF) band, in particular, the automated picture transmission (APT) channel (Salas et al., 2008). Previous versions of A3TB dealt with low rate picture transmission (LRPT) signals from the meteorological operational satellite-A (MetOp-A), where a complete receiver with beamforming and synchronization stages has been implemented (Salas et al., 2007; Martínez et al., 2007).

3. Antenna arrays for satellite communications

Satellite applications require compactness, lightweight and low cost antenna systems to be mounted on a terrestrial vehicle, an aircraft or a ship, or as a portable man-pack or a handset, and to be competitive against ground systems. Its major advantage is the possibility of getting a wider or even a global coverage. For such purposes, antenna arrays offer the technology to get a directive system whose steering direction can be electronically and/or mechanically controlled. However, planar arrays usually cannot steer more than 60º-70º from the normal direction of the antenna (Mailloux, 2005). Thus, when a wider angular coverage is required conformal arrays are an appropriate option (Josefsson & Persson, 2006). Arrays can approximate conformal shapes, such as spheres or cylinders, using several planar arrays, simplifying fabrication of active components (Sierra et al., 2007).

Since the low cost and low weight specifications are of importance, micro-strip antennas are mostly used, due to its capacity to be printed over a dielectric substrate with photolithography techniques. Low cost and low permittivity substrates are usually used such as FR4 or PTFE with different quantities of glass or ceramic impurities. For more demanding applications, ceramics, like alumina or high/low temperature co-fired ceramics (HTCC/LTTC) allow the use of smaller components thanks to its high permittivity, and give robustness against mechanical stresses and high temperatures.

3.1 Geodesic antenna array for satellite tracking in ground station

The aim of using a single antenna for tracking many satellites at the same time avoiding mechanical movements as well as its inexpensive cost make these antennas an alternative to be considered (Salas et al., 2008). Multi-beam ability and interference rejection are facilitated thanks to the electronic control system of such antennas that improves the versatility of the ground stations.

The GEODA is a conformal adaptive antenna array designed for MetOp satellite communications with specifications shown in Table 1. This antenna was conceived to receive signals in single circular polarization (Montesinos et al., 2009). Subsequently, in recent efforts the system has been upgraded also for transmission and double circular polarization (Arias et al., 2010). Hence, operating at 1.7 GHz with double circular polarization it can communicate with several LEO satellites at once in Downlink and Uplink. Current structure is the result of a comprehensive study that valued the ability to cover a given spatial range considering conformal shape surface and a given beamwidth (Montesinos et al., 2009). As Fig. 1 shows, GEODA structure consists of a hemispherical dome placed on a cylinder of 1.5 meters height. Both cylinder and dome are conformed by 30 similar triangular planar arrays (panels). Each panel consists of 15 sub-arrays of 3 elements (cells). The radiating element consists of 2 stacked circular patches with their own

RF circuits. The principal patch is fed in quadrature in 2 points separated 90º in order to obtain circular polarization. The upper coupled patch is used in the aim of improving the bandwidth.

Each panel is able to work itself as an antenna since they have a complete receiver that drives the 1.7 GHz signal to an analog to digital converter (ADC). In order to adapt the signal power to the ADC, it is mandatory to implement a complete intermediate frequency (IF) receiver consisting of heterodyne receiver with an automatic gain control block. Hence, each triangular array has active pointing direction control and leads the signal to a digital receiver through an RF conversion and filtering process. To follow the signal from the satellite, the main beam direction has to be able to sweep an angle of 60º. In this way, it is needed a phase shift in the feeding currents of the single radiating element. Previous calculations have demonstrated that 6 steps of 60 degrees are needed to achieve the required sweeping angle. An adaptive digital system allows the adequate signal combination from several triangular antennas. The control system is explained in (Salas et al., 2010).

Parameter	Specification	Parameter	Specification
Frequency range [GHz] **Tx:** **Rx:**	1.65 to 1.75 1.65 to 1.75	**Isolation between Tx and Rx [dB]**	>20
Polarization	Dual circular for Tx and Rx bands	**VSWR**	1.2:1
G/T [dB/K] **For elevation >30º** **For elevation 5º**	3 6	**SLL [dB]**	-11
EIRP [dBW]	36	**Size [m]**	1.5x1.5x3
3dB beamwidth [deg.]	5	**Accuracy steering [deg.]**	±1.4
Maximum gain [dBi]	29	**Coverage [deg.]:** **Azimuth** **Elevation**	360º >5º
Efficiency [%]	50		

Table 1. Main specifications for GEODA antenna.

3.1.1 Cell radiation pattern

Based on the study presented in (Sierra et al., 2007), the single radiating element is a double stacked circular patch that works at 1.7 GHz with 100 MHz bandwidth. In order to obtain circular polarization, the lower patch, which has 90 mm diameter, is fed by 2 coaxial cables in quadrature. Both coaxial cables connect the patch with a hybrid coupler to transmit and receive signals with both, right and left, circular polarizations. The upper patch is a circular plate with 78.8 mm diameter, and it is coupled to the lower patch increasing the bandwidth by overlapping both resonant frequencies tuning the substrate thickness and the patch diameter size. Fig. 2.a shows the radiating element scheme and main features of the layer structure are specified in (Montesinos et al., 2009).

A cell sub-array of 3 radiating elements shown in Fig. 2.b is considered the basic module to build the planar triangular arrays. The whole cell fulfills radiation requirements since it has a good polar to crosspolar ratio and a very low axial ratio. Likewise, as it is presented in Fig. 2.c, the radiation pattern shows symmetry and low side lobes for full azimuth.

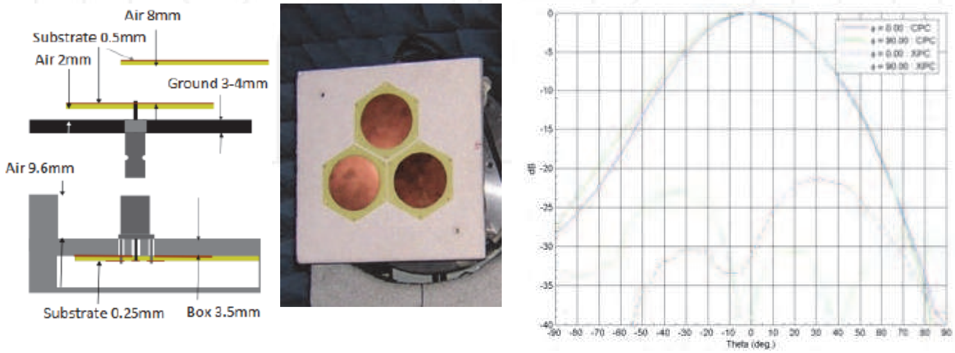

a b c

Fig. 2. a) Assembly of the single radiating element, b) Cell scheme, and c) Cell radiation pattern.

3.1.2 Transmission and Reception (T/R) module and cell distribution

Different T/R module configurations have been considered, providing either single or double polarization (Arias et al., 2010). T/R module allows amplifying and controlling the phase shift between signals, received and transmitted, providing an adaptive beam and steering direction controller in the whole working pointing range. As Fig. 3 shows, the design implemented contains a hybrid coupler, enabling double circular polarization; a double pole double throw (DPDT) switch, selecting polarization associated with transmission and reception way; 2 low noise amplifiers (LNAs), which amplify the signal received or transmitted; a single pole double throw (SPDT) switch, choosing transmission or reception way; and phase shifters, introducing multiples of 22.5° relative shift phases to form the desired beam. These surface mount devices have been chosen in order to reduce space and simplify the design.

Signals transmitted/received by the 3 T/R modules placed in a cell are divided/combined thanks to a divider/combiner circuit composed of 3 hybrid couplers that leads the signal to a general T/R module where signal is amplified. Due to transmission and reception duality, 2 SPDT switches are used to select the amplification way. Furthermore, each T/R module has associated a -25dB directional coupler that is used to test T/R modules in the transmission mode. Additionally, reception mode is tested by measuring signal in the divider/combiner circuit. A single pole 6 throw (SP6T) switch selects the path that is tested.

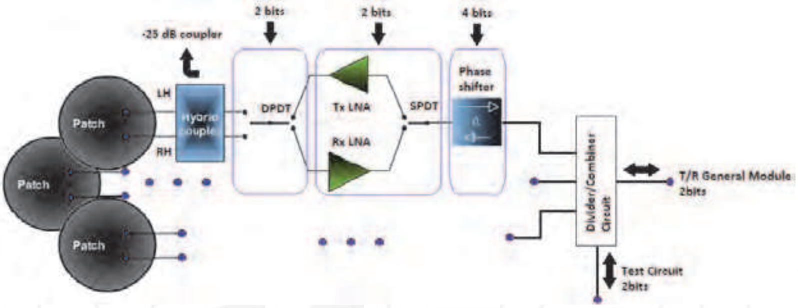
2 bits
2 bits
4 bits
-25 dB coupler
LH
RH
Hybrid coupler
DPDT
Tx LNA
Rx LNA
SPDT
Phase shifter
Patch
Patch
Patch
Divider/Combiner Circuit
T/R General Module 2bits
Test Circuit 2bits

Fig. 3. Cell sub-array and RF circuit.

3.1.3 Control system

The control system has two main parts (Salas et al., 2010), the hardware structure and the control software. The two level hardware structure has the lowest possible number of elements, making the control simpler in contrast to the previous in (Salas et al., 2010). Finally, an inter-integrated circuit (I2C) expander is used to govern T/R modules individually, and one more cover cell needs (LNA of call and test). A multipoint serial standard RS-485 is used to connect the computer with the panels.

3.2 Portable antenna for personal satellite services

New fix and mobile satellite systems (Evans, 2000) require antenna systems which can be portable, low profile and low weight. Planar antennas are perfect candidates to fulfill these specifications. Usually slots (Sierra-Castañer et al., 2005) and printed elements (García et al., 2010) are most used as radiating elements.

3.2.1 Antenna system structure

In this subsection it is introduced a printed antenna for personal satellite communications at X band, in Fig. 4. Table 2 shows main antenna characteristics.

Parameter	Specification	Parameter	Specification
Frequency range[GHz] **Tx:** **Rx:**	7.9 to 8.4 7.25 to 7.75	**Efficiency [%]**	50
Polarization	Dual circular polarization for Tx and Rx bands	**Isolation between Tx and Rx [dB]**	>17
G/T [dB/K]	7	**VSWR**	1.4:1
EIRP [dBW]	32	**SLL [dB]**	-11
3dB beamwidth [deg.]	5	**Size [m]**	40x40x2.5
Maximum gain [dBi]	25	**Weight [Kg]**	2

Table 2. Portable antenna specifications.

This is a planar, compact, modular, low loss and dual circular polarized antenna, for Tx and Rx bands, simultaneously. It is made up by a square planar array of 16x16 double stacked micro-strip patches, fed by two coaxial probes. A hybrid circuit allows the dual circular polarization (Garg et al., 2001). Elements are divided in 16 sub-arrays excited by a global power distribution network of very low losses, minimizing the losses due to the feeding network and maximizing the antenna efficiency. In order to reduce side lobe levels (SLL), the signal distribution decreases from the centre to the antenna edges, keeping symmetry with respect to the main antenna axes. The antenna works at X band from 7.25 up to 8.4 GHz with a 14.7% relative bandwidth for a 1.4:1 VSWR and a maximum gain of 25 dBi.

3.2.2 Sub-array configuration

The sub-array configuration can be seen in Fig. 4.a. It makes possible to separate the fabrication of these sub-arrays from the global distribution network, simplifying the corporative network and getting a modular structure suitable for a serial fabrication process. Each sub-array is a unique multilayer board, where PTFE-Glass substrate of very low losses has been used as base material. The power distribution network is connected to each sub-array through (SMP-type) coaxial connectors.

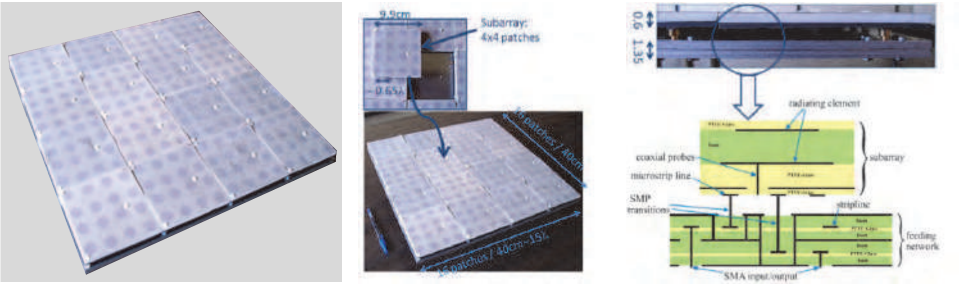

a b c

Fig. 4. a) Dual polarized portable printed antenna for satellite communication at X band, b) Sub-array perspective view, and c) Side view and multilayer scheme.

Fig. 5.a and Fig. 5.b show the sub-array unit cell. In order to obtain better polarization purity, each element is rotated 90º and excited by a 90º phase-shifted signal. Moreover, in Fig. 5.c is showed a miniaturized branch-line coupler (BLC) of three branches working as a wide band hybrid circuit (García et al., 2010; Tang & Chen, 2007).

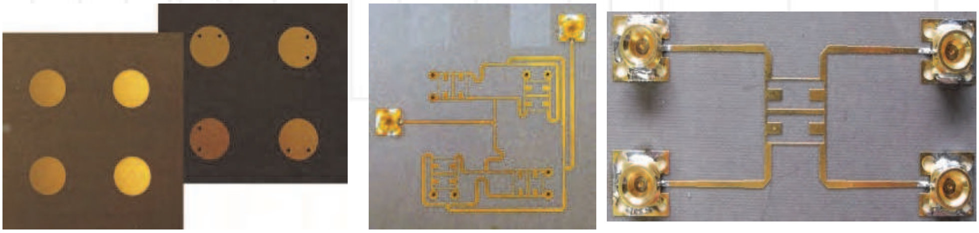

a b c

Fig. 5. Unit cell test board, a) Unit cell test board 2x2 stacked patches, b) Micro-strip feeding network, and c) Miniaturized BLC Prototype.

A conventional configuration takes up an area of 13.3 cm^2 which is big compared to the radiating element and the sub-array subsystem size. Therefore, a miniaturization of the BLC is needed using the equivalence between a λ/4 transmission line and a line with an open-ended shunt stub. An area reduction about 35% is achieved and the hybrid circuit behaves like a conventional BLC. In Fig. 6.b and Fig. 6.c measurement results for the BLC in Fig. 5.c are shown compared with simulations.

Fig. 7 depicts some sub-array measurements. The copular to crosspolar ratio is better than 25 dB and axial ratio is under 0.9 dB in the whole bandwidth.

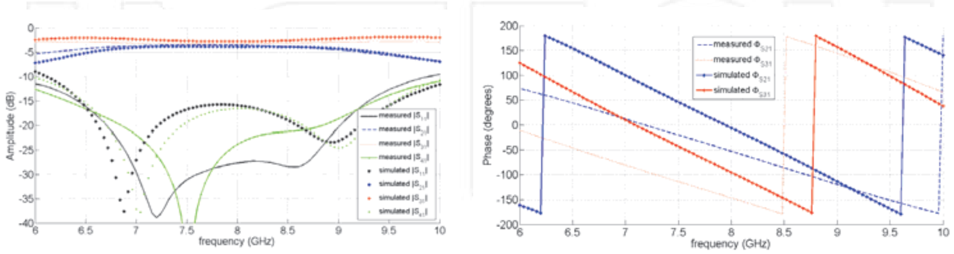

a b

Fig. 6. Miniaturized BLC, Measured and simulated S-parameters in: a) Amplitude, and b) Phase.

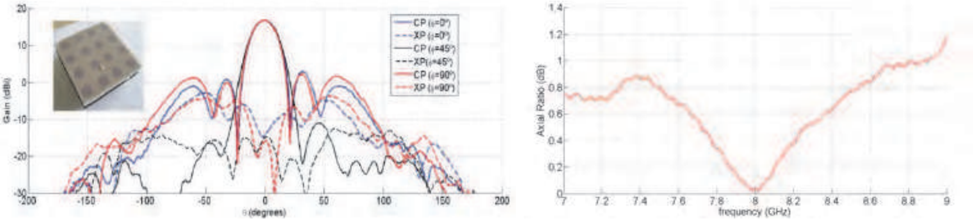

a b

Fig. 7. 4x4 patch sub-array measurements, a) Radiation pattern at 7.75 GHz, and c) Axial ratio for right-handed circular polarization.

3.2.3 Low losses power distribution network

The global feeding network presented in Fig. 8.a is a protected strip-line, where foam sheets of high thickness are used to get low losses. Such a kind of feeding network allows keeping a trade-off between the simplicity of exciting the radiating elements using printed circuits and the loss reduction when the distribution network is separated in a designed structure to have low losses. Losses in the structure are around 0.6 dB/m which yields to 0.3 dB of losses in the line. Two global inputs/outputs using SMA-type connectors, one for each polarization, excite the strip-line networks.

Vertical transitions have to be treated carefully and must be protected to avoid undesired higher order mode excitation. Thereby, it has been design a short-ended pseudo-waveguide, adding some extra losses about 0.3 dB, for two kinds of vertical transitions, as can be seen in Fig. 8.b and Fig. 8.c.

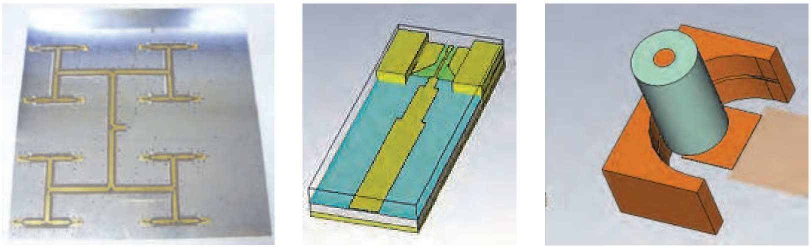

a b c

Fig. 8. a) Protected strip-line global corporative network for one polarization, b) Transitions from strip-line to SMA-type connector, and c) Transitions from strip-line to SMP-type connector.

3.2.4 Antenna performance

Fig. 9 depicts measured radiation pattern at 7.75 GHz, gain and axial ratio for the antenna system. It is shown a maximum gain of 25 dBi in the lower band and about 22 dBi in the upper band, and a SLL around 11 dB. Copolar to crosspolar ratio is better than 30 dB and axial ratio is under 0.7 dB. Total losses are about 4 dB in the working band.

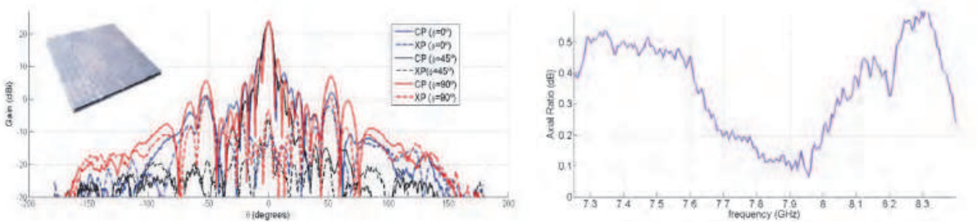

a b

Fig. 9. Antenna measurements results, a) Radiation pattern at 7.75 GHz, and c) Axial ratio for right-handed circular polarization.

3.3 Electronically steerable antennas for mobile and fixed portable systems

At present, two types of electric steerable antenna systems can be used to access the satellite communication services (Bialkwoski et al., 1996). These are: fixed position portable systems and mobile systems such as those installed on a land vehicle. The fixed portable antenna system is relatively easy to be accomplished by the antenna designer. The design involves standard procedures that concern the operational bandwidth, polarization and moderate gain (García et al., 2010). One drawback of the fixed position portable system is that they require the user to be stationary with respect to the ground. This inconvenience can be overcome with the mobile antenna system. A mobile user complicates the scenario since the ground mobile antenna needs to track the satellite (Alonso et al., 1996). The design of such a system is more challenging as new features associated with the mobility of the system have to be incorporated (Fernández et al., 2009). The requirement leads to a narrow beamwidth, for which satellite tracking is required as the vehicle moves around. Electronically steerable antennas enable the development of reconfigurable antennas for satellite applications.

3.3.1 Steerable antenna for fixed position portable systems

This antenna is a fixed satellite communication system with high gain at X band, consisting of an antenna array that integrates 32 2x2 sub-array modules in the complete antenna, as shown in Fig. 10.a. It is a planar and dual circular polarized antenna for Tx and Rx bands simultaneously. It is made up by a planar array of double stacked circular micro-strip patches, fed by 2 coaxial probes to generate circular polarization. A hybrid circuit allows the dual circular polarization as shown in Fig. 10.b.

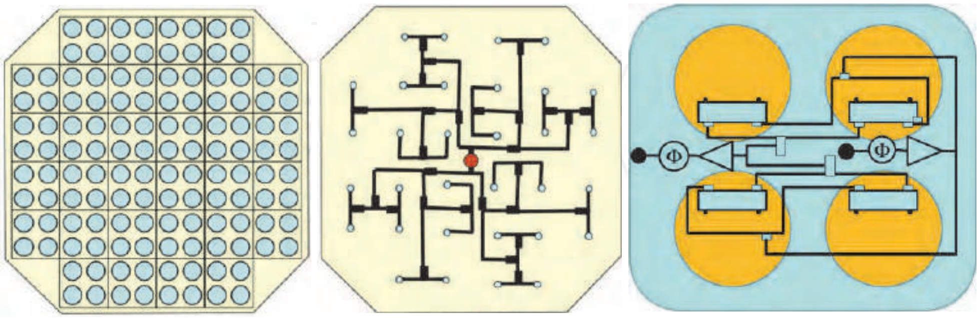

a b c

Fig. 10. Active multi-beam antenna, a) Top view, b) Feeding network of the complete antenna, and c) Beamforming network of the 2x2 sub-array module

The antenna has the same design parameters, structure and configuration as the antenna explained in Section 3.2 but with a different feeding network, as previously shown. In this case, the beamforming network requires changes in the feeding phase in the 2x2 sub-arrays, which can be achieved by phase shifters (ϕ) associated with different sub-arrays (Fig. 10.c). All these sub-arrays are connected to a feeding network, in Fig. 10.b, formed by transmission lines with low losses in strip-line. General specifications of the steerable antenna for fixed position portable systems are provided in Table 3.(a).

3.3.2 Automatic steerable antenna for mobile systems

A broadband circularly polarized antenna for satellite communication in X band is presented in Fig. 11 and specified in Table 3.(b). The arrangement features and compactness are required for highly integrated antenna arrays. It is desired to get a low-gain antenna for mobile satellite communications with low speed of transmission. In this system, the antennas are formed by 5 planar 4x4 arrays of antennas, which form a truncated pyramid with a pointing capability in a wide angular range, so that among the 5 planar arrays the complete antenna can cover any of the relative positions between the mobile system and the satellite in a practical way. The scheme of the active antenna can be seen in Fig. 11.

As it can be observed in Fig. 11.a, the antenna terminal is a multi-beam printed antenna shaped as a trunk pyramid capable of directing a main beam in the direction of the satellite. The antenna steering system consists of a multi-beam feeding structure with switches that lets combine the feed of each 4x4 arrays to form multiple beams. Switching the different 4x4 arrays, it is achieved different multiple beams and the variation of the steering direction.

INTECH

The complete antenna consists of a Tx and Rx module that works independently in the 2 frequency bands.
The antenna has multiple beams covering the entire space to capture the satellite signal without moving the antenna. The signal detected in each of the beams is connected to a switch, which, by comparison, is chosen the most appropriate 4x4 array. The steering direction of the 4x4 array can vary between a range of directions that covers a cone angle range of 90º. To obtain the required gain and cover the indicated range, it is required around 15 beams, which can be obtained by integrating the beamforming networks with switches in the design as presented in (Fernández et al., 2009).

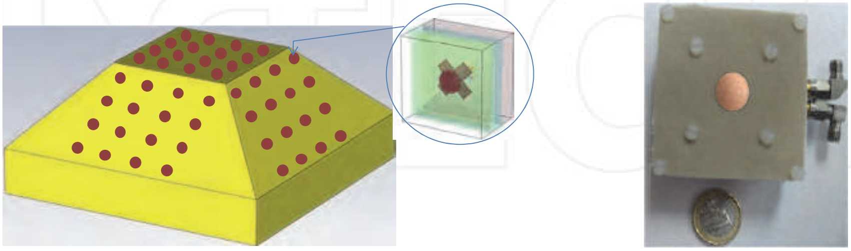
a b

Fig. 11. Complete antenna structure, a) Radiating element of the 4x4 arrays, and b) Prototype top view.

The radiating element of the 4x4 array is one 2 crossed dipoles with a stacked circular patch as shown in Fig. 11.a and Fig. 11.b. In Fig. 12 the cross-section of the radiating element structure is presented.

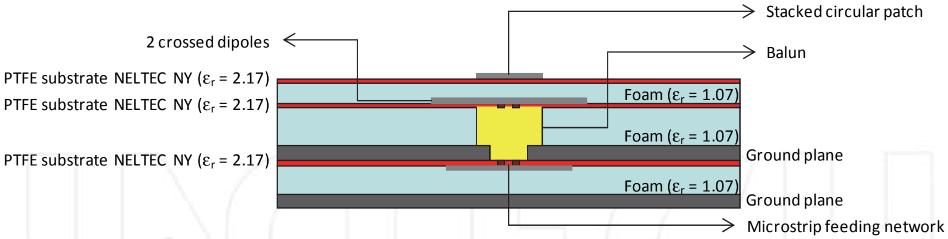
Stacked circular patch
2 crossed dipoles
Balun
PTFE substrate NELTEC NY (ε_r = 2.17)
PTFE substrate NELTEC NY (ε_r = 2.17)
Foam (ε_r = 1.07)
Foam (ε_r = 1.07)
PTFE substrate NELTEC NY (ε_r = 2.17)
Ground plane
Foam (ε_r = 1.07)
Ground plane
Microstrip feeding network

Fig. 12. Cross-section scheme of the radiating element.

The key element of the radiating element feeding structure (Fig. 14.b) is a resonant micro-strip feed ring that has been implemented, as well as a micro-strip 90º branch-line coupler to obtain the desired right hand or left hand circular polarizations (RHCP or LHCP) which ensures adequate port coupling isolation. The S-parameters in amplitude and phase of the micro-strip feeding structure are shown in Fig. 13.a and Fig. 13.b.
Fig. 14.a depicts the S-parameters of the radiating element with the micro-strip feed structure and they fulfill the specification, in Table 3.(b). In Fig. 14.c, the radiation pattern of the radiating element at 7.825 GHz is shown and in Fig. 14.d the radiation pattern of the 4x4

arrays is presented. It is shown a maximum gain of 19.4 dBi at the center frequency band (7.825 GHz). Copolar (CP) to crosspolar (XP) ratio is better than 17 dB and the axial ratio is under -3dB.

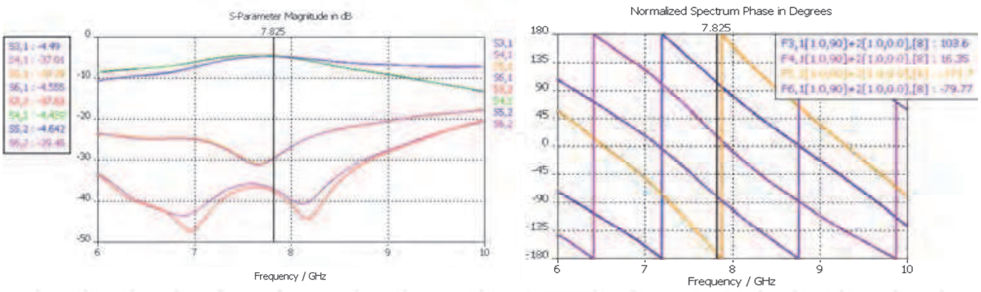

a b

Fig. 13. Micro-strip feeding structure, a) Amplitude of S-parameters, and b) Phase of S-parameters.

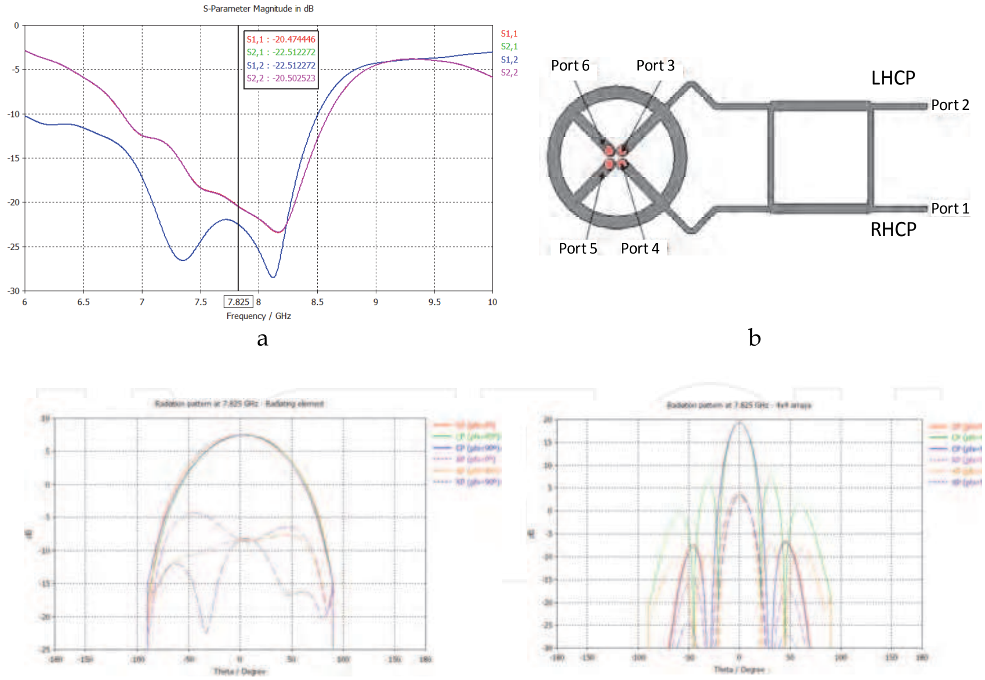

a b

c d

Fig. 14. a) S-parameters, b) Resonant ring + 90º branch-line coupler, c) radiation pattern at 7.825 GHz, and d) 4x4 array radiation pattern.

Parameter		Value (a)	Value (b)	Comments
Freq. range [GHz]	**Rx** **Tx**	7.25 - 7.75 7.9 - 8.4	7.25 - 7.75 7.9 - 8.4	Microwave applications.
G/T (in Rx) [dB/K]		7	7	
EIRP (in Tx) [dBW]		32	32	
Beamwidth at -3dB [deg.]		4	20	
Polarization		circular	circular	In both, reception and transmission.
Gain [dBi]		>28	>15	
Axial ratio [dB]		< 1	<3	**(a)** Between ±50°. **(b)** Between ±45°.
VSWR		< 1.4:1 (-15.6 dB)	< 1.5:1 (-13.9 dB)	
Isolation between ports [dB]		< -17	< -15	
Radiation pattern [deg.]		±35	±90	Steering direction tilt.
Dimensions [cm]		40x40x4	20x20x15	

Table 3. (a) General specifications of the steerable antenna for fixed position portable systems , and (b) General features of the automatic steerable antenna for mobile systems.

3.4 Transmit-array-type lens antenna for terrestrial and on board receivers

Technology in satellite communications has revealed an increasing interest in novel smart antenna designs. Phased-array based designs are basic in electronically reconfigurable devices for satellite applications, which are more and more demanding. The strict requirements in terms of architecture, shape and robustness are important constraints for the development of planar lens-type devices. Regarding the usage and location, lens-type devices are useful for either terrestrial or on board receivers, in vehicular technology. Some clear examples are satellite communications for aircrafts preserving the fuselage aerodynamics or for some other kind of vehicles such as trains, etc.

3.4.1 Introduction to lens-type structures

In a general view, in lens-type a particular signal is received (in our case, an electromagnetic wave with specific features in terms of frequency, wave-front, etc.), it is processed (either complex signal processing techniques or only phase correction tasks can be considered in this interface), and finally, the processed signal is retransmitted.

Regarding the lens configuration, a transmit-array lens consists of three well distinguished interfaces: the first one for signal reception, one interface for signal processing, and the last one for processed signal re-radiation, as depicted in Fig. 15.

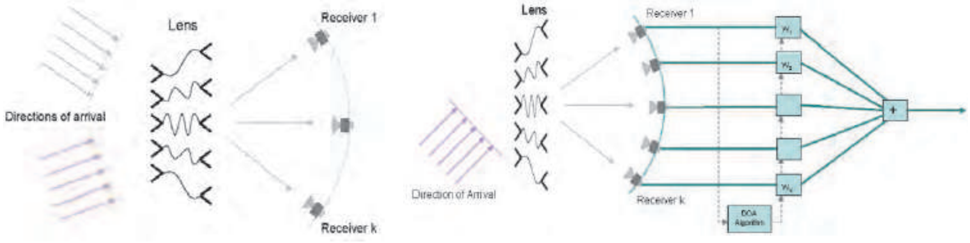

a b

Fig. 15. a) Multi-user scheme with different receivers and transmitters, and b) Adaptive scheme with DoA determination.

These structures are intimately related to reflect-array ones, where the reception and transmission interfaces are turned to be the same interface, with a reflection-type behavior (Encinar & Zornoza, 2001). Although in an equal output phase configuration a transmit-array device behavior would be similar to the one obtained with a reflect-array, the transmit-array offers the advantage of removing the feed blockage.

In a transmission scheme, depending on the transmitter position regarding the lens, a different steering direction is achieved and a different user is pointed. In the case of reception, the situation is the same: the user position configures the direction of arrival, which determines the receiver position around the lens (Padilla et al., 2010a). In adaptive schemes, applying the proper processing algorithm to the signal received in the different receivers around the lens, it is possible to develop an adaptive steering vector, in terms of the desired direction of arrival.

3.4.2 Transmit-array lens architecture and design

Lens-type structures provide two fundamental advantages. First, phase error correction due to spherical wave front coming from the feeding antenna. Fig. 16.a shows this effect. Second, new radiation patterns configuration. Fig. 16.b depicts this fact.

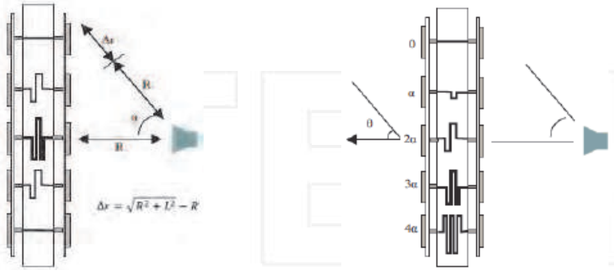

a b

Fig. 16. a) Phase error correction, and b) Radiation pattern reconfiguration.

3.4.3 Electronically reconfigurable devices for active transmit-array lenses

The addition of reconfigurability on transmit-array devices requires the possibility of controlling the phase response of the transmitted signal at each cell of the lens. Electronic control of phase signal may be added in two different ways: First, electronic tuning of the

radiating element phase response (Padilla et al., 2010a): Modifications in the radiating element circuital behavior lead to changes in phase response (arg[S_{21}]). Fig. 17 shows an electronically reconfigurable microwave patch antenna for this purpose, along with the equivalent circuit and prototype outcomes in terms of phase.
Second, electronic tuning of phase shifters in transmission lines (Padilla et al., 2010c): Modifications in the phase response of the phase shifters lead to corresponding changes in phase response. Some options are applied for these devices, such as hybrid couplers, etc. Fig. 18 shows a microwave phase shifter prototype for this purpose, along with the working scheme and its outcomes in phase.

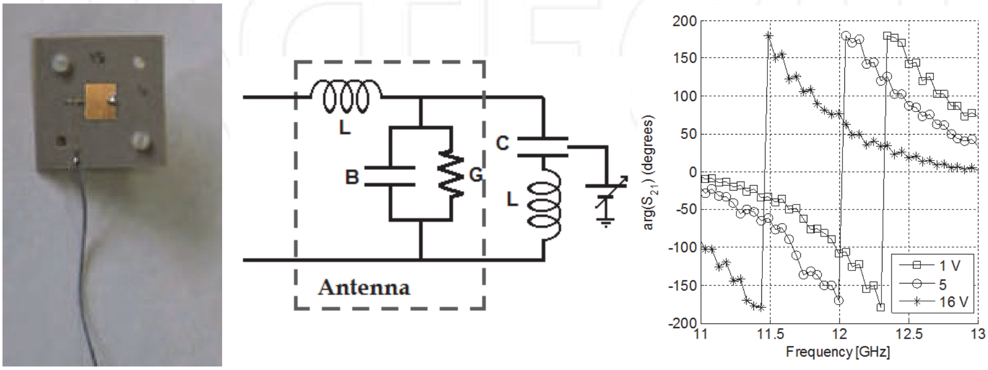

a b c

Fig. 17. Electronically reconfigurable antenna, a) Patch antenna prototypes, b) Equivalent circuit, and c) Phase behavior in frequency.

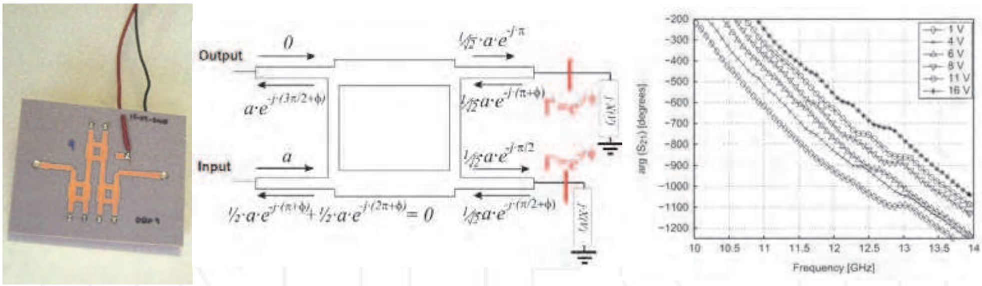

a b c

Fig. 18. Electronically reconfigurable phase shifter, a) Phase shifter prototype, b) Working scheme, and c) Phase behavior in frequency.

3.4.4 Electronically reconfigurable active transmit-array prototype

One electronically reconfigurable prototype is presented in Fig. 19 and detailed in this section. The prototype design implies the use of microwave phase shifters according to the design specified in section 3.4.3. This transmit-array lens prototype operates at 12 GHz. Main specifications are provided in Table 4.

Parameter	Value	Comments
Frequency range [GHz]	12 ± 0.5	Microwave applications.
Polarization	Linear	In both, reception and transmission.
Directivity [dBi]	>21	
Axial ratio [dB]	< 1	Between ±50° elevation.
S_{11} [dB]	< -20	
Radiation pattern [deg.]	±30	Steering direction tilt, for both H and V planes.
Feeding antenna [mm]	120	Corrugated horn linearly polarized
Phase shifters [deg.]	360	Full phase range variation.
Transmit-array elements	36	6x6 array topology.
Separation between elements	$0.7\lambda_0$	Related to the wavelength

Table 4. Main features of the electronically reconfigurable transmit-array prototype.

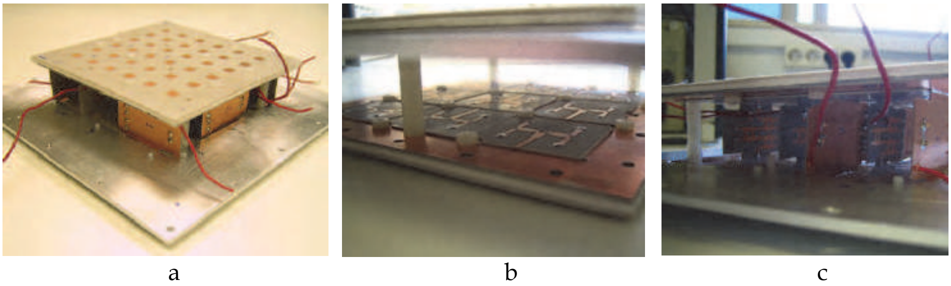
a b c

Fig. 19. Transmit-array core, a) Transmit-array prototype, b) Distribution networks, and c) Phase shifter integration.

The electronically controllable steering capabilities are tested and assured for a range of ± 30°in each main axis. An example of radiation pattern is provided in Fig. 20, for 9° tilt in one of the main axes.

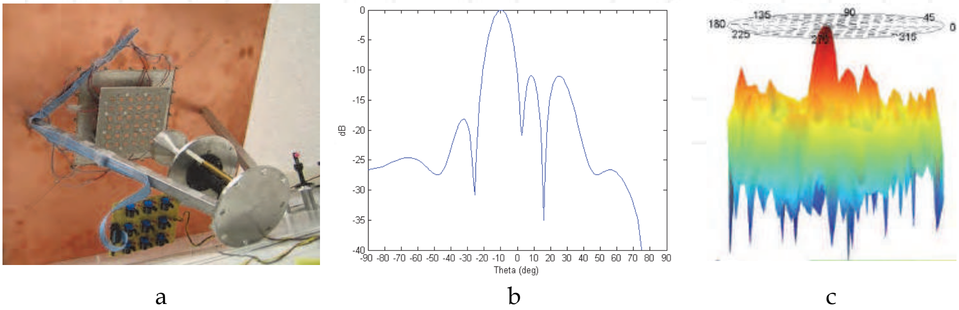

Fig. 20. a) Complete transmit-array with feeder and control circuits; and transmit-array measurement results for 9° tilt in one axis, b) H plane, and c) 3D plot.

4. Adaptive antenna array

Adaptive antennas can be described as systems usually based on three main parts: the antenna array, the receiver architecture and the beamforming scheme. Thus, adaptive antennas have those advantages owing to those three main parts. The system capabilities increase as complexity and development cost do. Furthermore, since signal processing is the basement of the adaptive antenna concept it is important to analyze the design challenges in terms of hardware architecture and components such as processors and embedded systems.

The antenna array provides the capability of performing the antenna pattern meeting the environment requirement under study. Besides, receiver architectures have some interesting advantages depending on the implemented receiver arraying technique such as signal to noise ratio (SNR) and bit error rate (BER) performance enhancement. Furthermore, symbol synchronization and carrier recovery can be used increasing the receiver complexity but providing higher performances. Finally, beamforming schemes use multiple antennas in order to maximize the strength of the signals being sent and received while eliminating, or at least reducing, interference as discussed in Section 4.3.

Adaptive antenna arrays are often called Smart Antennas because they have some key benefits over traditional antennas, by adjusting traffic patterns, space diversity or using multiple access techniques. The main four key benefits are: First, enhanced coverage through range extension by increasing the gain and steering capability of the ground station antenna; Second, enhanced signal quality through multi-target capability and reduction of interferences; finally, adaptive antennas improve the data download capacity in the ground segment of satellite communication by increasing the coverage range (Martínez et al., 2007).

4.1 Design and architecture based on software defined radio

For design there is the well known waterfall life cyclic model (Royce, 1970) that can be used to manage main aspects of the design of architectures. Thus, some tasks must be fulfilled subsequently as follow in Fig. 21.a.

Fig. 21.b shows the design schemes resulting of the requirement analysis stage corresponding software and hardware system specifications. In the depicted scheme, there are some system components such as the radiating element and RF circuits that are often designed under iterative prototyping model.

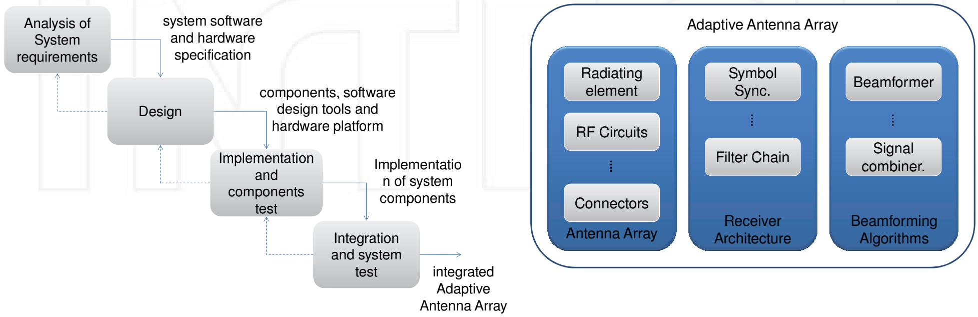
Analysis of System requirements
system software and hardware specification
Design
components, software design tools and hardware platform
Implementation and components test
Implementation of system components
Integration and system test
integrated Adaptive Antenna Array
Adaptive Antenna Array
Radiating element
RF Circuits
Connectors
Antenna Array
Symbol Sync.
Filter Chain
Receiver Architecture
Beamformer
Signal combiner.
Beamforming Algorithms

a b

Fig. 21. a) Water life cyclic model of the adaptive antenna array design, and b) Simplified design scheme of adaptive antenna arrays.

Regarding the hardware implementation, tables presented in (Martínez et al., 2007) show the hardware resource consumption in the field programmable gate array (FPGA) Virtex-4 for the least mean squared (LMS) beamforming algorithm with full spectrum combining (FSC) receiver architecture and SIMPLE beamforming algorithm with symbol combining (SC) receiver architecture. Both scheme designs have an antenna array of 2 elements. The algorithm based on correlation requires less hardware. The main difference can be appreciated in the amount of digital signal processing oriented component (DSP48) resources, typically used for filtering applications (Martínez et al., 2007).

4.2 Receiver architectures based on algorithms type

Several receiver architectures can be implemented, and they are frequently based on the type of the beamforming algorithm used. When training signals are available in the transmitted frame, a time-based reference algorithm can be used. However, this solution is only valid when the earth station is capable of demodulating the received training sequence. Other algorithms used in deep space communications are based on signal correlation and they avoid performing the demodulating process. This kind of algorithms are blind techniques that do not require any additional signal demodulation before applying some beamforming technique and work better in low SNR conditions than time-based algorithms.

Several receiver architectures can be implemented exploiting the processing capabilities of the SDR, such as FPGA, application-specific integrated circuits (ASICS), and digital signal processing (DSPs). The design of the receiver architecture fundamentally depends on the selection of beamforming algorithms. An example of beamforming technique is the LMS algorithm whose estimation of coefficients or weights requires a temporal reference and is implemented through SC receiver architecture (Fig. 22.a). In the other hand, the SIMPLE algorithm (Rogstad, 1997) constitutes a beamforming technique that is implemented using FSC receiver architecture (Fig. 22.b) in order to perform the calculation of weights.

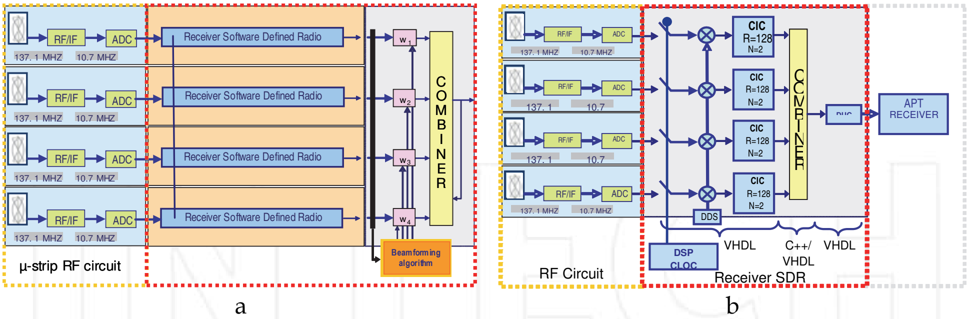

Fig. 22. Comparison of receiver architectures. a) Symbol Combining (SC), and b) Full Spectrum Combining (FSC).

The SC architecture can be divided into two more sub-classes which work on a phase-recovery basis. The complex symbol combining (CSC) recovers the phase information with regard to a reference element using feed-forward and feedback algorithms. One of the advantages of this scheme is that the rate of data sent to the combining module has a rate slightly higher than the symbol rate. For most applications, the symbol rate is relatively low and is a multiple of the data rate. In this kind of schemes, there is an important cost consideration in real-time applications and the requirements of instrumental phase stability are very severe (Rogstad et al., 2003). Other type of SC architecture is the stream symbol combining (SSC). In this kind of scheme, data are sent to the combining module at a rate equal to the symbol rate. The symbol rate depends on the coding scheme and for most applications is relatively modest. Also, the requirements of instrumental phase stability are no severe, as in the case of CSC scheme. The disadvantage of the SSC is the additional hardware required for each antenna.

Furthermore, there are the baseband combining (BC) and carrier arraying (CA) architectures discussed in (Rogstad et al., 2003). In BC architectures the signal from each antenna is carrier locked and combining in baseband for further demodulation and synchronization. In effect, the carrier signal from the spacecraft is used as a phase reference so that locking to the carrier eliminates the radio-frequency phase differences between antennas imposed by the propagation medium. Besides, in CA architectures, one individual carrier-tracking loop is implemented on each array element. Then, the elements branches are coupled in order to increase the carrier-to-noise ratio (CNR), but losses of radio channel are far compensated (Rogstad et al., 2003).

In general, the selection of the beamforming algorithms is determined by the following aspects: Hardware and computational resources; Speed of convergence and residual error of adaptive algorithms; Calibration requirements and auto-compensation ability; and system signal-transmission characteristics.

4.3 Beamforming techniques for satellite tracking

Some satellites transmit useful information inside its frames for synchronization and tracking purposes. The gathering of satellite data requires the tracking operation along its earth orbit. To accomplish this goal with adaptive array architectures, some beamforming techniques should be implemented. Fig. 23 illustrates a simple example of a narrowband linear adaptive beamformer system.

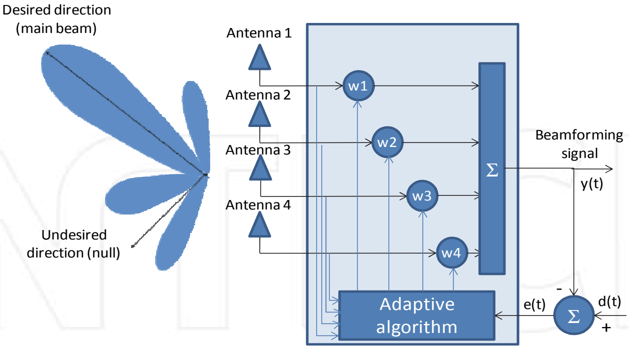

Fig. 23. Adaptive antenna system.

A linear beamformer combines signals according to some weights w_i, to produce a desired radiation pattern. The mathematical expression of a linear beamformer at the array output in vector notation can be expressed as $y = w^H x$, where x is the received signal vector to be combined, w are the weights computed by the beamforming algorithm and ^{H} denotes transposition and conjugate of $(\cdot)$.

In adaptive antennas design, weights are dynamically calculated with a certain algorithm in order to optimize some signal parameter like signal to interference-plus-noise ratio (SINR), SNR, or BER. An extended variety of algorithms exist in the literature for beamforming purpose and the most appropriated selection is done depending on the signal characteristics of the received signal.

4.3.1 Blind techniques

Blind beamformers make use of an inherent property of the received signal, such as the ciclo-stationarity of the constant modulus. In the latter, the algorithm eliminates the fluctuation of the signal amplitude and computes the weights to minimize the effect produced by those variations. The algorithms that make use of these methods are denoted as Constant Modulus Algorithms (CMA) (Biedka, 2001).

CMA algorithms present an important disadvantage: as the phase information is not considered, the constellation of quadrature phase shift keying (QPSK) signals commonly used in satellite communications appears rotated after beamforming, which imposes the need of an additional phase recovery subsystem in the array output.

4.3.2 Temporal-reference algorithms

Algorithms based on a temporal reference require a known reference included in the frame of the signal, such as training sequences, unique word (UW) or pilot bits. Thus, these schemes are normally used for digital signals. The aim of these beamformers is the minimization of the energy of an error signal integrated by interferences and noise. In order to reduce the order of the problem, the weight calculation is usually done iteratively.

The most popular adaptive filters are the LMS and Recursive Least Squares (RLS) algorithms (Haykin, 2002). Briefly, the main differences lie in the method to calculate and the final convergence behavior: while LMS has a linear complexity order with the number of antennas in the array, RLS makes use of matrix operation, so that the complexity order is quadratic, but the convergence is faster.

An interesting alternative to the LMS is the Normalized LMS (NLMS), which normalizes the adaptive step to avoid variation during the convergence process. The counterpart is the more intensive processing requirements to calculate signal power and normalization operation.

4.3.3 Correlation-based algorithm

In contrast to beamformers based on temporal reference, schemes based on signal correlation do not require the demodulation of any signal. These techniques are the most popular to extract the spatial information for beamforming, and we have focused on the use of the SIMPLE algorithm (Rogstad, 1997). This algorithm has been used by the Deep Space Network (DSN) of National Aeronautics and Space Administration (NASA) to combine the signals received from spatial probes in radio telescopes located in different sites around the Earth surface. The main disadvantage of correlation based schemes is the lack of ability to cancel interference signals.

4.4 Performance comparison

Some simulation comparisons between spatial and blind algorithms are presented to show benefits and drawbacks. Four algorithms have been selected with a 4-element uniform linear array (ULA). The spatial algorithms simulated are post-beamformer interference canceller – orthogonal interference beamformer (PIC-OIB) (Godara, 2004) and minimum power distortionless response (MPDR) (Van Trees, 2002). On the other hand, the blind algorithms are the matrix-free EIGEN and the SUMPLE (Rogstad, 1997). The convergence process is compared as a function of the input SNR as depicted in Fig. 24.

As it can be observed from the above results, spatial algorithms outperform blind ones at low SNR, and vice versa. On the other hand, with medium-low SNR and low or absence of interferences, the behavior of all algorithms is quite similar.

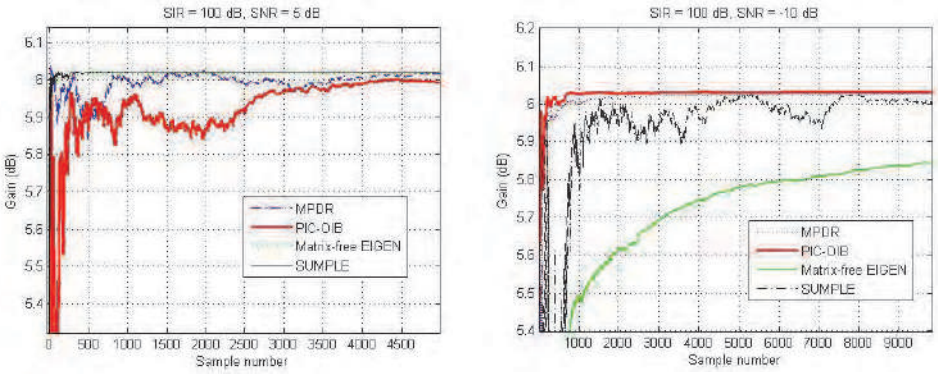

a b

Fig. 24. Convergence behavior of spatial versus blind algorithms in the absence of interferences with several input SNR. a) SNR = 5 dB, and b) SNR = -10 dB.

5. Experimental Test-Bed based on SDR platform

This section presents a test platform known as Adaptive Antenna Array Test-Bed - A3TB, where a comparative study of several beamforming algorithms can be performed and modularity of the architecture is a well proved advantage. The test bed is based on SDR technology and uses a novel architecture that can be used with both blind and spatial-based beamforming algorithms. The A3TB concept can be applied to a number of scenarios as the current version is independent of the signal properties. Simulation results using the A3TB with the APT channel from NOAA satellites show the performance of the concept and the feasibility of the proposed implementation.

The scope of the system development was is to prove the concept of antenna arrays applied to ground stations instead of reflectors for different applications, such as telemetry data downloading or end-user in mobile applications as discussed in the introduction section. In contrast to reflector antennas, antenna arrays offer the possibility of electronic beam-steering avoiding the use of complex mechanical parts and therefore reducing the cost of the antenna. It is also a good chance for Universities and Research Centers aiming to have their own ground station sited in their installations.

5.1 A3TB concept

The A3TB can be defined as *a software-defined radio beamformer applied to a ground station for tracking LEO satellites*. The novelty relies on the use of an antenna array to smartly combine the received signals from the satellite and its implementation based on SDR technology. The reason to use an antenna array instead of a single antenna is to electronically steer the beam in the direction of the satellite along its orbit without requiring a mechanical system for tracking. In addition to the advantages of the use of SDR technology and antenna array, it is the modularity and flexible architecture implemented in the A3TB. Fig. 25 shows the A3TB architecture where it is evident the feasibility to update or change during operation any of the main blocks. It is possible to change during operation the beamforming algorithm and to include new beamforming modules to the system. Furthermore, changes on the BENADC are possible to implement not during operation, but new receiver architecture at off-line such as those options discussed at follow.

In (Salas et al., 2007), the block diagram represents the software system implementation of the first version of the test-bed prototype and most of it is based on VHDL. Depending on the firmware, three options could be installed into the FPGA Virtex4. The option A is implemented with the signal processing on the PC, so the SIMPLE beamforming is done in the module developed in C++. The option B is implemented completely on VHDL and this option need to export the beamforming weights just to draw the array pattern diagram. Finally, in contrast to the option B, the option C is implemented for the LMS beamforming algorithm.

With the first version of the Test-Bed, the modularity on the selection of firmwares was proved switching between A, B or C receiver architectures, and an important result of the Test-Bed development is the hardware resources occupation presented in (Salas et al., 2007).

The advantage of the SDR implementation is that A3TB architecture can be used to process any received signal from a LEO satellite in the appropriate band imposed by the RF stages. Moreover, most of the processing tasks are performed on software, using appropriate routines to process any receive signal. There are 2 main schemes to implement the beamforming stage: SC and FSC [41]. Both schemes are compared in Section 4.2.

The current version of the A3TB in Fig. 25.a was updated to track NOAA satellites in the VHF band, in particular the APT channel. Previous versions of A3TB dealt with LRPT signals from MetOp-A, where a complete receiver with beamforming and synchronization stages has been implemented(Salas et al., 2007; Martínes et al., 2007).

5.2 Implementation of the A3TB

The A3TB prototype consists of 4 main parts as shown in Fig. 25.a. The first part is the antenna array, which has 4 crossed-dipole antennas as depicted in Fig. 25.b. The second part consists of RF-IF circuits which amplify and down convert to IF incoming signals. Furthermore, an automatic gain control (AGC) was implemented using two steps of variable attenuators in the IF domain.

The third part is the SDR platform which consists of the beamforming algorithms implemented on C++ and the FPGA firmware on VHDL, PC and BENADC blocks show in Fig. 25, respectively. The hardware resources occupation for this Test-Bed implementation is similar to one presented in (Martínes et al., 2007). The last part is the software from weather satellite signal to image decoder (WXtoImg) on the PC using the sound card output/input in order to get the weather satellite image.

Since the implemented architecture is FSC the demodulation is not required and the IF signal is digitized. For the signal processing hardware design the BenADC-v4 has been chosen. This solution includes a FPGA Xilinx Virtex4-SX55 with four 12-bit analog inputs at

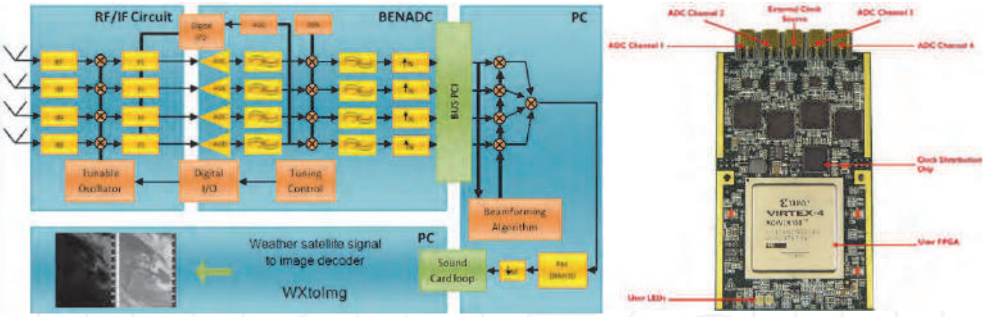

a b

Fig. 25. a) Block diagram of the A3TB, and b) BenADC – Virtex 4-sx55.

250 Msps (Martínes et al., 2007). Digital samples are transferred to the PC where beamforming and subsequent APT demodulation of the array output are performed using C++ routines. This implementation design offers higher flexibility for testing different beamforming schemes. Finally, demodulated APR frames are sent to the WXtoImg software to show meteorological maps.

The A3TB is controlled by the PC for simulations and field trials. The graphical user interface allows presented in (Salas et al., 2008) the user to choose the beamforming algorithm and set all the parameters of the LEO satellite for tracking such as the number of antennas of the array, distance between the elements, direction of arrival and IF frequency. The C++ routine calculates the beamforming weights and plots the synthesized array factor. Subsequently, the reception of meteorological images has real time system requirements. Thus, it is necessary a data transfer from the FPGA to the C++ module to process the samples continuously, and give APT frames to the audio output of the PC. Since, the meteorological satellites often have a low baud rate, in the case of study with NOAA satellites the data transfer is made using two buffers controlled by a thread.

It is important to mention that the A3TB with SDR architecture can evaluate different beamforming algorithms and receiver schemes. The update of A3TB for larger arrays is immediate, as the basis for algorithms is independent of the number of elements in the array. The architecture of a new ground station concept to track LEO satellites based on software defined radio and antenna arraying as Test-Bed is a well proved choice to evaluate future antenna array architectures for satellite communication and benchmark features of the proposed system. As the A3TB VHF version is based on FSC scheme, the concept can be applied to a number of satellite tracing scenarios.

6. Conclusions

The performance analysis of different beamforming algorithms is an important issue in the new generation antenna array development and research. Thus, A3TB helps to analyze beamforming algorithms paving the way for testing and debugging for posteriori use in larger arrays, such as GEODA. Results obtained in real scenarios with A3TB state, for example, that spatial reference algorithms such as MPDR should be used in the absence of interferences, whereas blind algorithms are appropriate for low SNR conditions. Finally, the A3TB can also serve to validate the performance of calibration procedures.

In future work, the A3TB will deal with the system combining of full modularity with the capability of change firmwares based on the first version design of the Test-Bed, plus the flexible architecture of the current design of the Test-Bed based on VHDL, C++ and Antenna Arraying. Furthermore, the addition of more modules to increase the number of antenna array elements is evident in next generations.

7. Acknowledgment

Authors wish to thank MICINN (Ministerio de Ciencia e Innovación) for grants and CROCANTE project (ref: TEC2008-06736/TEC), INSA (Ingeniería y servicios Aeroespaciales) and Antenas Moyano S.L., for the partial funding of this work. Simulations in this work have been done using CST Studio Suite 2011 under a cooperation agreement between Computer Simulation Technology and Universidad Politécnica de Madrid. Substrates used in prototypes were kindly given by NELCO S.A.

8. References

Torre, A.; Gonzalo, J.; Pulido, M., & Martínez Rodríguez-Osorio, R. (2006). New generation Ground Segment Architecture for LEO satellites. *57th International Astronautical Congress*, pp. 221-226. Valencia, Spain, October 2006.

Tomasic, B.; Turtle, J. & Liu, S. (2002). A Geodesic Sphere Phased Array Antenna for satellite control and communication. *XXVII General Assembly of the International Union of Radio Science,* Maastricht, August 2002.

Godara, L. C. (1997). Application of Antenna Arrays to Mobile Communication, Part II: Beamforming and Direction of Arrival Considerations. *Proc. IEEE* , vol.85, No.8, (August 1997), pp. 1195-1245.

Martínez, R. & Salas Natera, M. A. (2010). On the use of Ground Antenna Arrays for Satellite Tracking: Architecture, Beamforming, Calibration and Measurements. *61st International Astronautical Congress*, pp. 1-7. Prague, 2010.

Liu, S.; Tomasic, B.; Hwang, S. & Turtle, J. (2006). The Geodesic Dome Phased Array Antenna (GDPAA) for Satellite Operations Support. *IEEE 18th International Conference on Applied Electromagnetics and Communications*, pp. 1-1, Dubrovnik, April 2006.

Tomasic, B. (1998). Analysis and Design Trade-Offs of Candidate Phased Array Architectures for AFSCN Application. *Second AFSCN Phased Array Antenna Workshop,* Hanscom, April 1998.

Sierra Pérez, M.; Torre, A.; Masa Campos, J. L.; Ktorza, D. & Montesinos, I. (2007). GEODA: Adaptive Antenna Array for Metop Satellite Signal Reception. *4th ESA International Workshop on Tracking, Telemetry and Command System for Space Application*, pp. 1-4, Darmstadt, Germany, September 2007.

Arias Campo, M.; Montesinos Ortego, I.; Fernández Jambrina, J. & Sierra Pérez, M. (2010). T/R Module Design for GEODA Antenna. *4th European Conference on Antenna and Propagation,* pp. 115-116, Barcelona, Spain, April 2010.

Franchi, A.; Howell, A. & Sengupta, J. (2000). Broadband mobile via satellite: Inmarsat BGAN. *IEEE Seminar on the Critical Success Factors - Technology, Services and Markets*, pp. 23-30, October 2000.

Evans, J. V. (1998). Satellite systems for personal communications. *Proc. IEEE* , Vol.86, No.7, (July 1998), pp. 1325-1341.

Wu, W. W. (1994). Mobile satellite communications. *Proc. IEEE* , Vol.82, No.9, (September 1994), pp. 1431-1448.

Fujimoto, K. & James, J. (2001). *Mobile Antenna Systems Handbook.* Artech House.

Garcia Aguilar, A.; Inclán Alonso, J. M.; Vigil Herrero, L.; Fernández González, J. M. & Sierra Pérez, M. (2010). Printed antenna for satellite communications. *IEEE International Symposium on Phased Array Systems and Technology*, pp. 529-535, Boston, USA, October 2010.

Geissler, M.; Woetzel, F.; Bottcher, M.; Korthoff, S.; Lauer, A. & Eube, M., (2010). L-band phased array for maritime satcom. *IEEE International Symposium on Phased Array Systems and Technology*, pp. 518-523, Boston, USA, October 2010.

Baggen, R.; Vaccaro, S. & Del Río, D. (2007). Design Considerations for Compact Mobile Ku-Band Satellite Terminals. *EuCAP The Second European Conference on Antennas and Propagation*, pp. 1-5. Edimburg, Scothland, September 2007.

Vaccaro, S.; Tiezzi, F.; Rúa, M. & De Oro, C. (2010). Ku-Band Low-Profile Rx-only and Tx-Rx antennas for Mobile Satellite Communications. *IEEE International Symposium on Phased Array Systems and Technology*, pp. 536-542, Boston, USA, October 2010.

Schippers, H. (2008). Broadband Conformal Phased Array with Optical Beam Forming for Airborne Satellite Communication. *IEEE Aerospace Conference*, pp. 1-17, September 2008.

Kanno, M.; Hashimura, T. T.; Sato, M.; Fukutani, K. & Suzuki, A. (1996). Digital beam forming for conformal active array antenna. *IEEE International Symposium on Phased Array Systems and Technology*, pp. 37-40, October 1996.

Khalifa, I. & Vaughan, R. (2007). Optimal Configuration of Multi-Faceted Phased Arrays for Wide Angle Coverage. *IEEE 65th Vehicular Technology Conference*, pp. 304-308, Boltimore, USA, April 2007.

Salas Natera, M. A.; Martínez Rodríguez-Osorio, R.; Antón Sánchez, A.; García-Rojo, I. & Cuellar, L. (2008). A3TB: Adaptive Antenna Array test-bed for tracking LEO satellites based on software defined radio. *59th International Astronautical Congress*, pp. 313-317, Glasgow. September 2008.

Salas Natera, M. A.; Martínez Rodríguez-Osorio, R. & García-Rojo López, I. (2007). Design of an Adaptive Antenna Array Test-Bed based on Software Radio for Tracking LEO Satellites. *IEEE EuCAP.* Edinburgh, Scotchland, Npvember 2007.

Martínez, R.; Salas Natera, M.; Bravo, A.; García-Rojo, I.; de Haro, L.; Mateo, M. & Gómez, M. (2007). VHF Ground Station with increased angular coverage for reception of meteorological satellites with electronic beamforming. *4th ESA International Workshop on Tracking, Telemetry and Command Systems for Space Application.* Darmstadt, Germany, 2007.

Mailloux, R. (2005). *Phased Array Antenna Handbook,* Artech House, Norwood, Massachusetts, USA.

Josefsson, L. & Persson, P. (2006). *Conformal Array Antenna. Theory and Design*, John Wiley & Sons , Hoboken, New Jersey, USA.

Montesinos, I.; Sierra Pérez, M.; Fernández, J. L.; Martínez, R. & Masa, J. L. (2009). GEODA: Adaptive Antenna of Multiple Planar Arrays for Satellite Communications. *European Conference on Antenna and Propagation.* Berlin, Germany, 2009.

Salas Natera, M. A.; Martínez, R.; De Haro Ariet, L. & Fernández Jambrina, J. (2010). Automated System for Measurement and Characterization of Planar Active Arrays. *IEEE Internationa Symposium on Phase Array Systems and Technology*, pp. 1-6, Boston USA, October 2010.

Sierra-Castañer, M.; Vera-Isasa, M.; Sierra-Pérez, M. & Fernández-Jambrina, J. (2005). Double-Beam Parallel Plate Slot Antenna. *IEEE Transactions on Antennas and Propagation*, vol.53, No.3, (2005) pp. 977-984.

Garg, R.; Bhartia, P.; Bahl, I. & Ittipiboon, A. (2001). *Microstrip Antennas Design Handbook,* Artech House, Norwood, Massachusetts, USA.

Tang, C. & Chen, M. (2007). Synthesizing Microstrip Branch-Line Couplers With Predetermined Compact Size and Bandwidth. *IEEE Transactions on Microwave Theory and Techniques, vol. 55, No.9,* (September 2007), pp. 1926-1934.

Bialkowski, M.; Jellett, S. & Varnes, R. (1996). Electronically Steered Antenna System for the Australian Mobilesat. *IEEE Transaction on Antennas ans Propagation*, vol. 143, No. 4, (August 1996), pp. 347-352.

Alonso, J.; Blas, J.; Garcia, L.; Ramos, J.; Pablos, J. & Grajal, J. (1996). Low Cost Electronically Steered Antenna and Receiver System for Mobile Saltellite Communications. *Trans. IEEE MTT*, vol. 44, No. 12, (December 1996), pp. 2438-2449.

Fernández, J. M.; Rizzo, C. & Sierra-Pérez, M. (2009). Antena Impresa Multihaz con Polarización Circular Derechas/Izquierdas Para Comunicaciones por Satélite en Banda X. *Proccedings of XXIV Simposium Nacional de la Unión Científica Internacional de Radio (URSI).* Santander, Spain, September 2009.

Encinar, J. A. & Zornoza, J. (2003). Broadband design of three- layer printed reflectarrays. *IEEE Transactions on Antennas and Propagation*, vol. 51, No. 7, (July 2003), pp. 1662-1664.

Padilla, P.; Muñoz-Acevedo, A.; Sierra-Castañer, M. & Sierra-Pérez, M. (2010). Electronically reconfigurable transmitarray at Ku band for microwave applications. *IEEE Transactions on Antennas and Propagation*, vol. 58, No. 8, (August 2010), pp. 2571-2579.

Padilla, P.; Muñoz-Acevedo, A. & Sierra-Castañer, M. (2010). Low Loss 360º Ku Band electronically Reconfigurable Phase Shifter. *International Journal of Electronics and Communications,* vol. 64, No. 11, (November 2010), pp. 1100-1104.

Royce, W. W. (1970). Managing the Development of Large Software Systems. *IEEE WESCON*, pp. 328 – 338, 1970.

Rogstad, D. H. (1997). *The SUMPLE Algorithm for Aligning Arrays of Receiving Radio Antennas: Coherence Achivied with Less Hardware and Lower Combining Loss.* TDA Progress Report, Jet Propulsion Laboratory.

Rogstad, D. H.; Mileant, A. & Pham, T. (2003). *Antenna Arraying Techniques in the Deep Space Network.* Deep Space Communication and Navegation Series, JPL California Institute of Technology, Ref. 03-001, Pasadena, USA.

Biedka, T. E. (2001). *Analysis and Development of Blind Adaptive Beamforming Algorithms.* PdD Thesis, Faculty of Virginia Polytechnique Institute and State University, Virginia.

Haykin, S. (2002). *Adaptive Filter Theory* (4th ed.). Prentice-Hall.

Godara, L. C. (2004). *Smart Antennas* (1st ed.). CRC Press..

Van Trees, H. L. (2002). *Optimum Array Processing. Part IV of Detection, Estimation, and Modulation Theory*. Wiley.

INTECH

INTECH

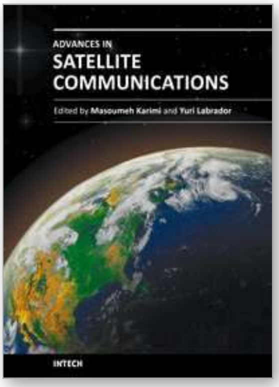

Advances in Satellite Communications

Edited by Dr. Masoumeh Karimi

ISBN 978-953-307-562-4
Hard cover, 194 pages
Publisher InTech
Published online 27, July, 2011
Published in print edition July, 2011

Satellite communication systems are now a major part of most telecommunications networks as well as our everyday lives through mobile personal communication systems and broadcast television. A sound understanding of such systems is therefore important for a wide range of system designers, engineers and users. This book provides a comprehensive review of some applications that have driven this growth. It analyzes various aspects of Satellite Communications from Antenna design, Real Time applications, Quality of Service (QoS), Atmospheric effects, Hybrid Satellite-Terrestrial Networks, Sensor Networks and High Capacity Satellite Links. It is the desire of the authors that the topics selected for the book can give the reader an overview of the current trends in Satellite Systems, and also an in depth analysis of the technical aspects of each one of them.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Miguel Alejandro Salas Natera, Andrés García Aguilar, Jonathan Mora Cueva, José Manuel Fernández, Pablo Padilla De La Torre, Javier García-Gasco Trujillo, Ramón Martínez Rodríguez-Osorio, Manuel Sierra-Perez, Leandro De Haro Ariet and Manuel Sierra Castañer (2011). New Antenna Array Architectures for Satellite Communications, Advances in Satellite Communications, Dr. Masoumeh Karimi (Ed.), ISBN: 978-953-307-562-4, InTech, Available from: http://www.intechopen.com/books/advances-in-satellite-communications/new-antenna-array-architectures-for-satellite-communications

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Hindawi Publishing Corporation
International Journal of Antennas and Propagation
Volume 2012, Article ID 428284, 10 pages
doi:10.1155/2012/428284

Review Article

Some Recent Developments of Microstrip Antenna

Yong Liu, Li-Ming Si, Meng Wei, Pixian Yan, Pengfei Yang, Hongda Lu, Chao Zheng, Yong Yuan, Jinchao Mou, Xin Lv, and Housjun Sun

Department of Electronic Engineering, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

Correspondence should be addressed to Yong Liu, fatufo@bit.edu.cn

Received 12 August 2011; Revised 23 November 2011; Accepted 2 January 2012

Academic Editor: Zhongxiang Q. Shen

Copyright © 2012 Yong Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Although the microstrip antenna has been extensively studied in the past few decades as one of the standard planar antennas, it still has a huge potential for further developments. The paper suggests three areas for further research based on our previous works on microstrip antenna elements and arrays. One is exploring the variety of microstrip antenna topologies to meet the desired requirement such as ultrawide band (UWB), high gain, miniaturization, circular polarization, multipolarized, and so on. Another is to apply microstrip antenna to form composite antenna which is more potent than the individual antenna. The last is growing towards highly integration of antenna/array and feeding network or operating at relatively high frequencies, like sub-millimeter wave or terahertz (THz) wave regime, by using the advanced machining techniques. To support our points of view, some examples of antennas developed in our group are presented and discussed.

1. Introduction

The concept of microstrip antenna was first introduced in the 1950s [1]. However, this idea had to wait nearly 20 years to be realized after the development of the printed circuit board (PCB) technology in the 1970s [2, 3]. Since then, microstrip antennas are considered as the most common types of antennas due to their obvious advantages of light weight, low cost, low profile, planar configuration, easy of conformal, superior portability, suitable for arrays, easy for fabrication, and easy integration with microwave monolithic integrate circuits (MMICs) [4–7]. They have been widely employed for the civilian and military applications such as television, broadcast radio, mobile systems, global positioning system (GPS), radio-frequency identification (RFID), multiple-input multiple-output (MIMO) systems, vehicle collision avoidance system, satellite communications, surveillance systems, direction founding, radar systems, remote sensing, biological imaging, missile guidance, and so on [8].

Despite the many advantages of typical microstrip antennas, they also have three basic disadvantages: narrow bandwidth, low gain, and relatively large size. The narrow bandwidth is one of the main drawbacks of these types of antennas. A straightforward method of improving the bandwidth is increasing the substrate thickness. However, surface wave power increases and radiation power decreases with the increasing substrate thickness [7], which leads to poor radiation efficiency. Thus, various other techniques are presented to provide wide-impedance bandwidths of microstrip antennas, including impedance matching networks using stub [9, 10] and negative capacitor/inductor [11], microstrip slot antennas using the U, L, T, and inverted T slots in the ground plane (sometimes termed defected ground structures (DGSs)) [12, 13], surface wave suppressing using magneto-dielectric substrate [14] and electromagnetic bandgap (EBG) structures [15], and composite-resonator microstrip antennas using metamaterial resonators [16, 17]. Another problem to be solved is the low gain for conventional microstrip antenna element. Cavity backing has been used to eliminate the bidirectional radiation, thereby providing higher gain compared with conventional microstrip antenna [18]. Lens covering is an alternative way to achieve gain enhancement. The lens with canonical profile, like elliptical, hemielliptical, hyper-hemispherical, extended hemispherical, used to focus the radiation beam from the radiator elements. The integrated lens microstrip antenna can be treated as composite antenna combined by microstrip radiator elements and dielectric lens, which is very useful for high frequencies (mm, sub-mm, terahertz (THz), and optical waves) applications

[19]. It is also well known that antenna array is an effective means for improving the gain [20–25].

The last limitation of conventional microstrip antennas is the relatively large size, particularly at lower microwave frequencies, since their operation frequencies are related to the electrical size of antenna. In general, the size of the rectangular microstrip antenna should be of order of a half-guided wavelength. This limitation was mathematically studied by Wheeler [26] and Chu [27]. There have been numerous efforts to minimize the antenna size and obtain the electrically small microstrip antenna with the raised demand towards smaller and smaller wireless devices. Inductive or capacitive loading are effective ways to reduce the size of microstrip antennas [28]. In the former work, we demonstrated that the size of microstrip antenna can be miniaturized using composite metamaterial resonators [16, 17]. Magneto-dielectric substrates have been widely used to miniaturize microstrip antennas due to magnetic substrates and could provide wider bandwidths than dielectric substrates [29–32]. Fractal geometries, which are composed by self-similar structures, have opened an alternative way for antenna miniaturization [33].

From the above discussions, we see that many methods and materials are used to improve the properties of microstrip antennas. However, there should be a relationship among bandwidth, gain, and size of the microstrip antennas. Antenna engineers have recognized that the improvement in one antenna property is frequently accompanied by decline in its other performances. For example, the antenna size is reduced usually at the expense of its bandwidth and gain. Therefore, a more comprehensive consideration must be given on further developments of microstrip antennas.

In this paper, we will suggest three areas for further research based on our previous works on microstrip antenna elements and arrays [16–25, 34–41]. We first note that novel microstrip antenna topologies are proposed to meet the desired requirement of variety of potential wireless applications, such as ultrawide band (UWB), high gain, miniaturization, circular polarization, multipolarized, and so on. Next, we discuss the composite antennas based on microstrip antennas which have more potent than each individual antenna. Finally, with the development of micro-/nano-machining techniques, antennas/arrays with highly integration and with highly operating frequencies are discussed. We present some examples of antennas developed in our group to support our points of view.

2. Variety of Microstrip Antenna Topologies

Microstrip antennas have extensively used in commercial and military applications due to their attractive advantages. However, the traditional microstrip antennas have the impedance bandwidth of only a few percent and radiation pattern with omnidirection, which obviously does not meet the requirements of various wireless applications. To this end, a wide variety of microstrip antenna topologies, including different microstrip antenna element structures and different microstrip array arrangements, have been studied to meet the desired requirement such as ultrawide band (UWB), high gain, miniaturization, circular polarization, multipolarized, and so forth.

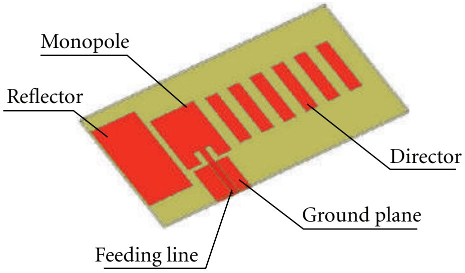

(a) The structure of the quasi-Yagi antenna

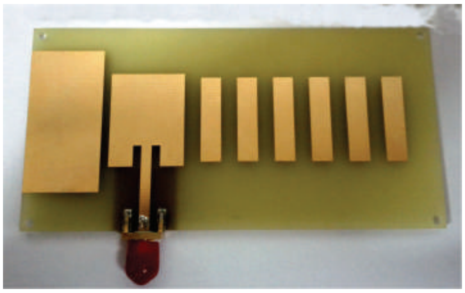

(b) The photograph of the quasi-Yagi antenna

Figure 1: Compact broad-band quasi-Yagi antenna.

As we know, microstrip antennas inherently have narrower bandwidth and lower gain compared to conventional bulky antennas. Some microstrip antennas with special topologies, like quasi-Yagi, planar reflector antenna, are proposed to replace the conventional bulky antennas. Here, we will take a quai-Yagi antenna as an example to show how to design a planar microstrip antenna with Yagi-Uda end-fire radiation pattern. In addition, a microstrip array with special array topology is designed to get dual-polarized property.

2.1. Compact Broad-Band Quasi-Yagi Antenna. A novel S-band compact quasi-Yagi antenna has been designed, fabricated and measured by our group, as shown in Figure 1. This antenna is composed of a printed monopole-driven element, a printed reflector element, and six printed director elements.

To explain the end-fire radiation behavior of the quasi-Yagi antenna, a comparison of radiation patterns, among (1) microstrip monopole only, (2) microstrip monopole and a reflector, (3) microstrip monopole and a director, (4) microstrip monopole and a reflector with one director, and (5) microstrip monopole and a reflector with six director, is shown in Figure 2. We can observe that both the reflector and the director can increase the end-fire radiation, and it could be substantially improved by increasing the number of directors.

The measured VSWR results are shown in Table 1. A bandwidth of 14% for VSWR less than 1.5 is achieved. The gain of the antenna is above 7.5 dBi, as shown in Table 2. In this design, we see that the microstrip antenna with special topology could be conveniently used to replace the bulky Yagi-Uda antenna.

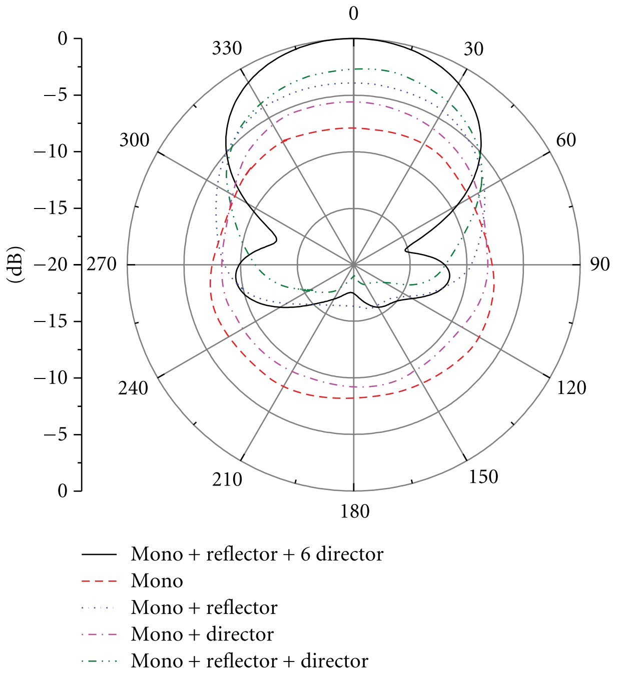

Figure 2: Radiation patterns of microstrip monopole only, microstrip monopole and a reflector, microstrip monopole and a director, microstrip monopole and a reflector with one director, and microstrip monopole and a reflector with six director.

Table 1: The measured VSWR of the quasi-Yagi antenna.

No.	Frequency (GHz)			
	3.25	3.5	3.75	Inband
1	1.36	1.34	1.47	<1.5
2	1.37	1.26	1.49	<1.5
3	1.36	1.25	1.48	<1.5

Table 2: The measured gain of the quasi-Yagi antenna (unit: dBi).

No.	Frequency (GHz)		
	3.25	3.5	3.75
1	7.57	8.73	8.35
2	7.58	8.55	8.37
3	7.56	8.77	8.51

2.2. Dual-Polarized Microstrip Antenna Array. The dual-polarized antenna is highly required for the radar, electronic countermeasure, and aerospace systems. It is known that the microstrip antenna can easily be integrated with microwave circuits and feeding network. Here, a novel Ku-band dual-polarization microstrip antenna array with a mixed feeding network, that is, the slot coupled feeding (V-port) and the co-plane feeding (H-port), is designed by our group, as shown in Figure 3. It is a three layers structure: top microstrip patch layer, middle stripline feeding network layer, and bottom coplane microstrip feeding network layer. Through proper array arrangement, very good isolation can be obtained.

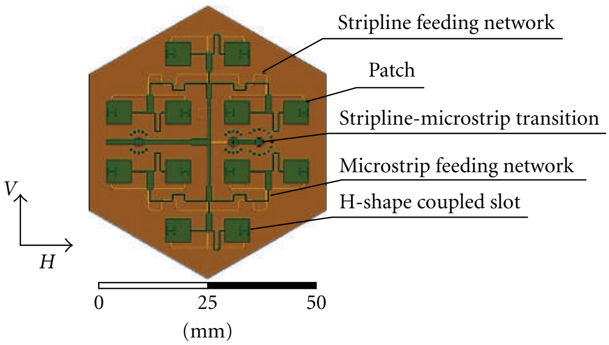
Stripline feeding network
Patch
Stripline-microstrip transition
Microstrip feeding network
H-shape coupled slot
V
H
0
25
50
(mm)

(a) The structure of the dual-polarized microstrip antenna array

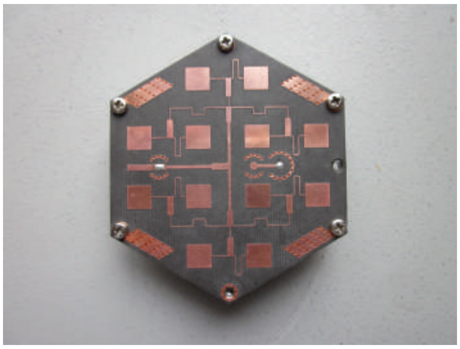

(b) The photograph of the dual-polarized microstrip antenna array

Figure 3: Dual-polarized microstrip antenna array.

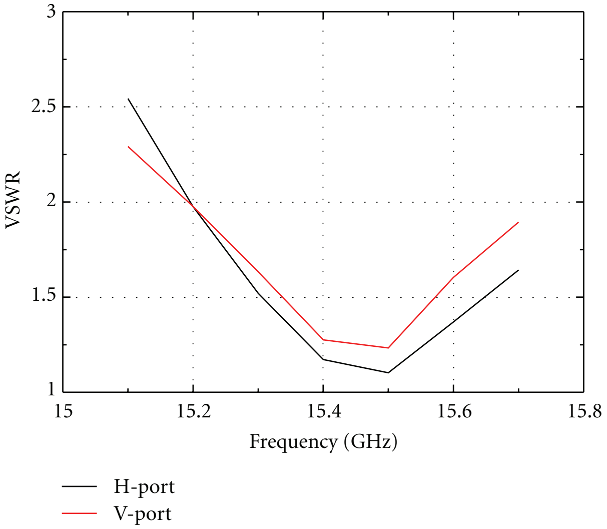

Figure 4: The VSWR of the dual-polarized microstrip antenna array.

The VSWR, radiation patterns, and the isolation between two polarizations of the proposed dual-polarized microstrip antenna array are shown in Figures 4, 5, and 6, respectively. The results indicate that this microstrip antenna array has a good impedance matching, good radiation performance, as well as very high isolation (less than −25 dB), which can be an idea candidate for the dual-polarized wireless systems.

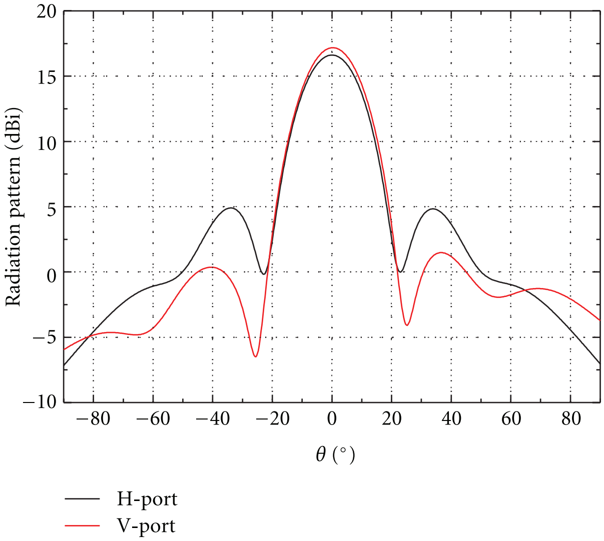

FIGURE 5: The radiation patterns of the dual-polarized microstrip antenna array at the center frequency.

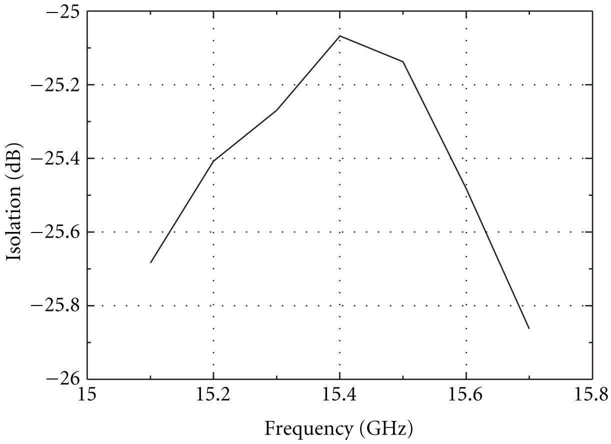

FIGURE 6: The isolation of the dual-polarized microstrip antenna array.

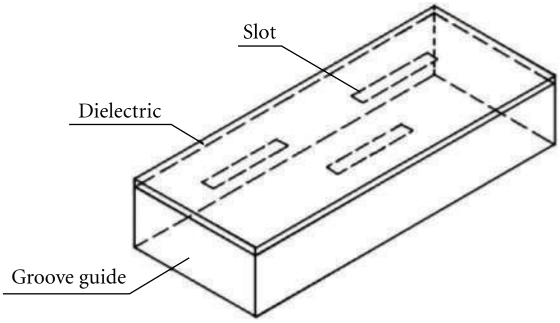

FIGURE 7: The structure of the DCWS.

3. Microstrip-Antenna-Based Composite Antenna

As many antenna designers have found, it is not easy to design an antenna to meet the user-defined stringent performance requirements demanded by special wireless applications like military radars, surveillances, and missile guidance, if only one type of antenna is considered. This difficulty may require the use of two more different types or structures of antenna elements with different characteristics. Composite antenna formed by two more types or structures of antennas is particularly suitable for these applications due to more advantages offered by different types or structures of antennas. For example, it is a challenging task to use single type of antenna to design a dual-band dual-polarization antenna for satellite digital multimedia broadcast (S-DMB) application [36]. A composite antenna composed with a left-handed circularly polarized (LHCP) microstrip antenna and a linear polarized omnidirectional biconical antenna is proposed by our group to meet this requirement [36]. Another example of composite antenna is comprised of a dielectric lens and microstrip log-period antenna, which has been widely applied to THz systems (this type of antenna will be further discussed in Section 4.2). Here, we will give an example of composite antenna with "structure composite" method.

3.1. Monopulse Circular-Polarized Dielectric Complex Waveguide Slot Antenna Array.

Waveguide slot antenna array has been widely used for wireless system, due to its advantages of high radiation efficiency, high power capacity, and high reliability. However, it is hard to overcome the disadvantage of high cost of fabrication.

One composite antenna with waveguide slot antenna array property, termed dielectric complex waveguide slot (DCWS), is composed with slot microstrip line and groove guide, as shown in Figure 7. The slot microstrip line is formed by a metal clad dielectric substrate and slots etched in the metal. This composite antenna not only maintains the advantages of the traditional waveguide slot antenna array but also has the characteristics of high consistence, easy for fabrication, and low cost.

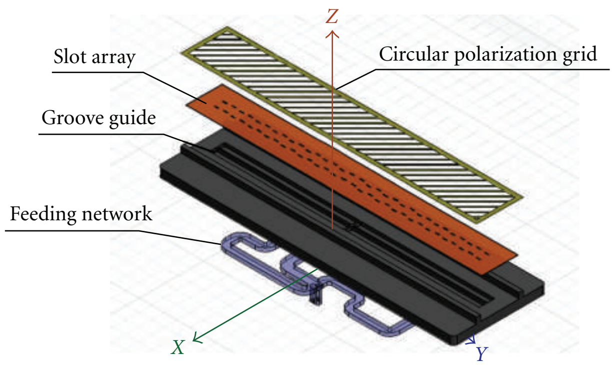

(a) The structure of the monopulse circular-polarized DCWS antenna array (separating view)

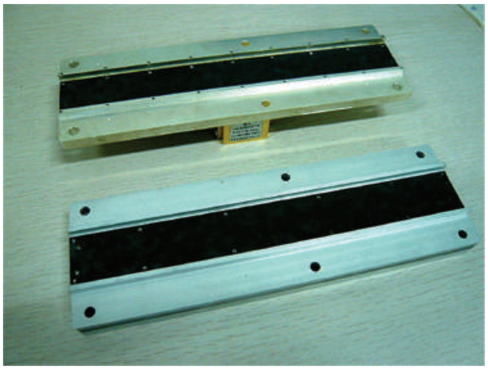

(b) The photograph of the monopulse circular-polarized DCWS antenna array.

FIGURE 8: Ka-band monopulse circular-polarized dielectric complex waveguide slot (DCWS) antenna array.

A Ka-band monopulse circular-polarized dielectric complex waveguide slot (DCWS) antenna array is designed, fabricated, and measured by our group, as shown in Figure 8. It consists of a circular polarization grid, a slot microstrip array, and a groove guide and feeding network. The slot microstrip array is fabricated on a Rogers 5880 film with dielectric constant of 2.2 and the thickness of 0.254 mm. The measured results of VSWR of sum and different port are shown in Figure 9. Figure 10 shows the measured radiation pattern at the center frequency. Some important array performance parameters such as gain, null depth and axial ratio (AR) are also given in Table 3. As shown in the measured results, very good performance can be obtained with the DCWS antenna array. The radiating efficiency of the DCWS antenna array is 80%, which is almost the same as the traditional waveguide slot antenna array. Moreover, the DCWS antenna array has 40% larger bandwidth than the traditional waveguide slot antenna array.

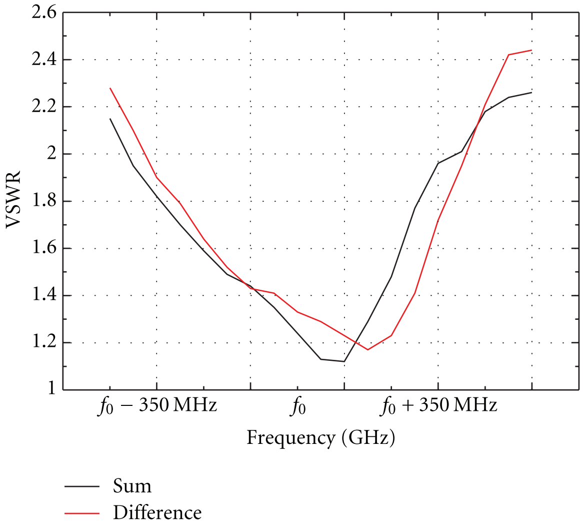

FIGURE 9: The VSWR of sum and difference port of the monopulse circular-polarized DCWS antenna array.

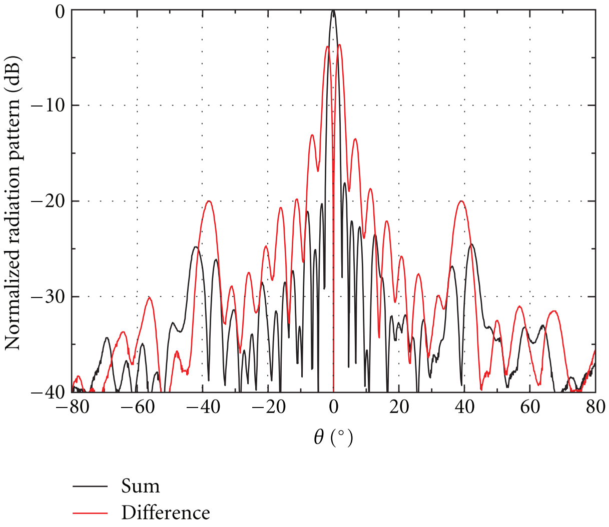

FIGURE 10: The radiation pattern of the monopulse circular-polarized DCWS antenna array at the center frequency.

4. Highly Integration and Highly Operating Frequency Antennas Based on Advanced Machining Techniques

It is known that the microstrip antenna was first fabricated using PCB technology in 1970s, nearly 20 years after its concept was first presented in 1950s [1–3]. Clearly, the development of microstrip antennas is closely related with the machining techniques. Recently, various machining techniques, including multilayer printed circuit board (MPCB), complementary metal oxide semiconductor (CMOS), low-temperature cofired ceramics (LTCC), and micro-electro-mechanical systems (MEMS), are highly developed, opening opportunities for innovative antennas, such as active antennas, reconfigurable antennas, metamaterial-based antennas, THz antennas, and so forth. With the availability of high-precision and high-speed advanced machining techniques, microstrip antennas are growing towards highly integration of antenna/array and feed network and operating at relatively high frequencies. Since they are all based on the advanced

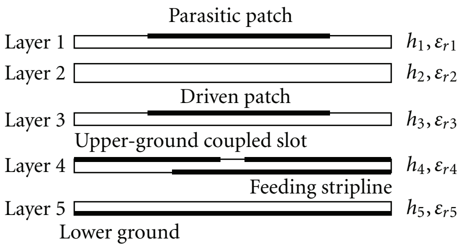

(a) Schematic side view of the structure of the high integrate broadband microstrip antenna array

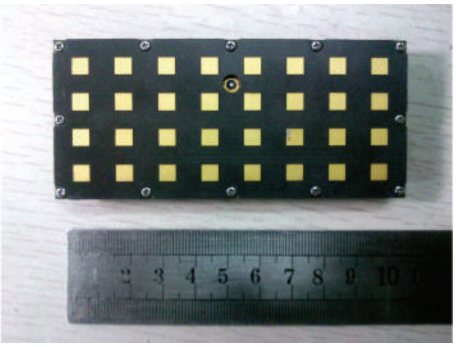

(b) The photograph of the high integrate broadband microstrip antenna array

Figure 11: Ku-band high integrate broadband microstrip antenna array using MPCB technology.

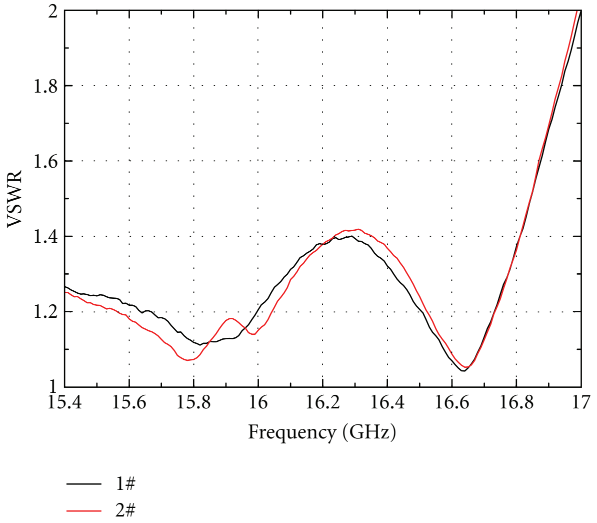

Figure 12: The VSWR of the high integrate broad-band microstrip antenna array using MPCB technology.

machining techniques, we suggest that a third research area of microstrip antennas is constantly introducing novel advanced machining techniques. In the following, two examples will be presented to show how important the advanced machining technique is to fabricate microstrip antennas. One is the highly integrate broad-band microstrip antenna array fabricated using MPCB technology. Another is THz wave planar integrated active microstrip antenna using MEMS technology.

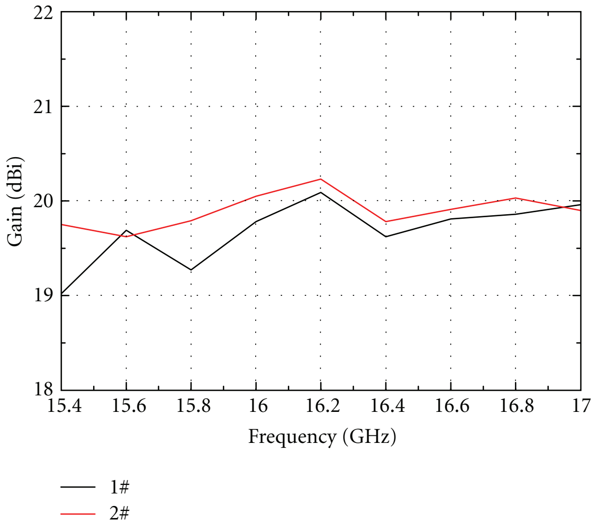

Figure 13: The gain of the high integrate broad-band microstrip antenna array using MPCB technology.

Table 3: The measured data of the monopulse circular-polarized DCWS antenna array.

Fre. (GHz)	Gain (dBi)	Null depth (dB)	AR (dB)
$f_0 - 0.2$	22.8	−37.3	3.8
f_0	21.9	−29.9	2.9
$f_0 + 0.2$	22.1	−26	4.1

4.1. High Integrate Broad-Band Microstrip Antenna Array Using Multilayer Printed Circuit Board (MPCB) Technology.

Recently, with the development of the multilayer printed circuit board (MPCB) technology, the microstrip antennas can be designed and fabricated from one-dimensional (1D) to 2D and even 3D structures.

Based on the MPCB technology, a high integrated broadband Ku-band microstrip antenna array is designed, fabricated, and measured by our group, as shown in Figure 11. This antenna consists of a parasitic patch, a driven patch, a stripline feeding network, a broad-band coaxial line to stripline transition, some buried screw holes, and some via holes. The feeding network is integrated in the bottom of the substrate of the antenna. As all of the structures fabricated at once, the accuracy and the uniformity can be assured. Two antennas of this type are measured. The measured VSWR, gain, and radiation pattern at the center frequency are shown in Figures 12, 13, and 14, respectively. The measured results show that this antenna maintains good radiation and matching performances with relative bandwidth of 13%. They have also shown good uniformity by using MPCB technology.

4.2. THz Wave Planar Integrated Active Microstrip Antenna Using Micro-Electromechanical Systems (MEMSs) Technology.

THz waves typically include frequencies between 0.1 THz and 10 THz. THz technology is now becoming a promising technology which has potential applications in many fields, such as short-range communication, biosensor, imaging,

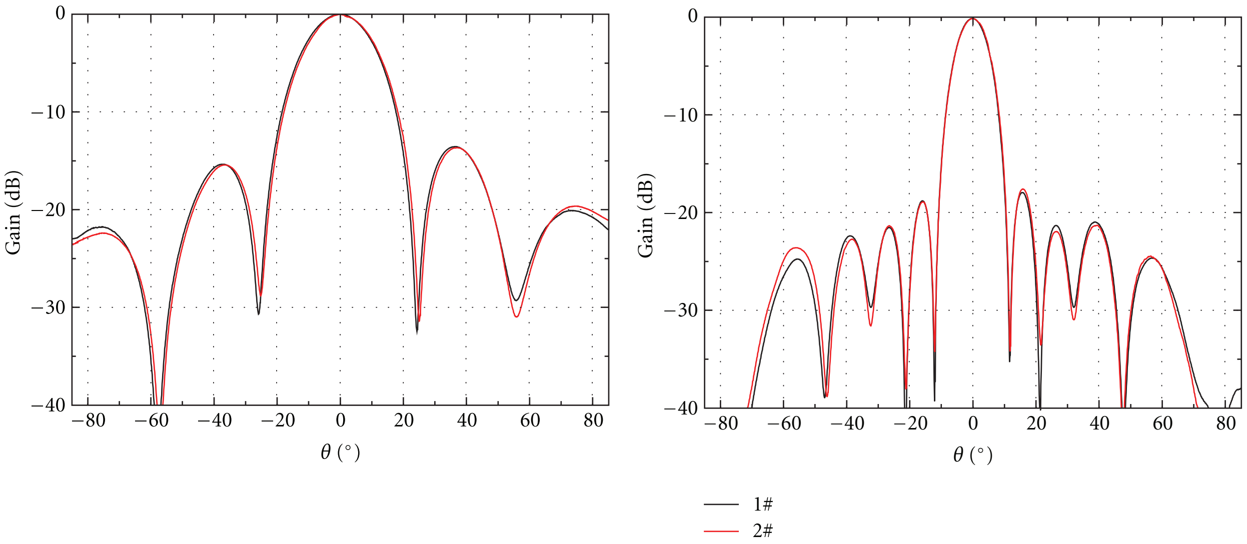

(a) E-plane

(b) H-plane

Figure 14: The radiation pattern of the high integrate broad-band microstrip antenna array using MPCB technology at the center frequency.

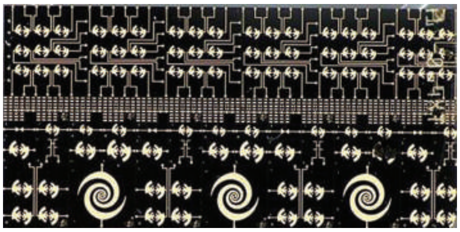

(a) The photograph of the THz monolithic antenna

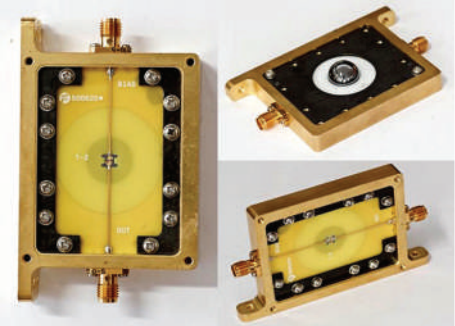

(b) The photograph of the THz monolithic antenna covered by a dielectric lens

Figure 15: THz wave planar integrated active microstrip antenna using micro-electromechanical systems (MEMSs).

national security, space exploration and communication, and so forth [39–46]. To realize THz transceiver system, antenna is an essential component. We often use horn antenna, lens antenna, and dielectric parabolic antenna, for THz systems. However, they are not easy to integrate with monolithic integrate circuits. Although the microstrip antenna has the merits of small volume, light weight, and easy integration with circuit, it is difficult to be processed in such high-frequency regions. MEMS technology opens the way to design of THz antennas, circuits, and systems. THz monolithic antenna fabricated using MEMS technology and covered by a dielectric lens, which can be considered a composite antenna, are designed, fabricated, and measured by our group, as shown in Figure 15.

Diodes have the functions of mixing and/or modulating the carrier-wave signal. It is an effective way to reduce the propagation path for detectors application by integrating the diode and microstrip antenna. The extended hyper-hemispherical dielectric lens is used to increase the gain of the microstrip antenna. An antenna-coupled detector integrated with a dielectric lens is designed and fabricated up to THz range by our group. The planar microstrip log-spiral antenna and log-period antenna have been fabricated using micro-electromechanical systems (MEMSs) technology. The photographs of the antennas are demonstrated in Figure 15. The measured responses of the antenna-coupled detector working at different frequency bands are shown in Figure 16, which can be considered to determine the effective operating frequencies [19, 40]. This detector gave a valid response from 12 GHz to 110 GHz frequencies. The results prove the validity and feasibility of the THz antenna designed using micro-electromechanical systems (MEMSs) technology.

5. Conclusion

The advantages and disadvantages of microstrip antennas are discussed in this paper. In particular, three areas for further development of microstrip antennas are presented based on our previous works on microstrip antenna elements and arrays. Variety of microstrip antenna topologies and microstrip-antenna-based composite antenna are discussed, and

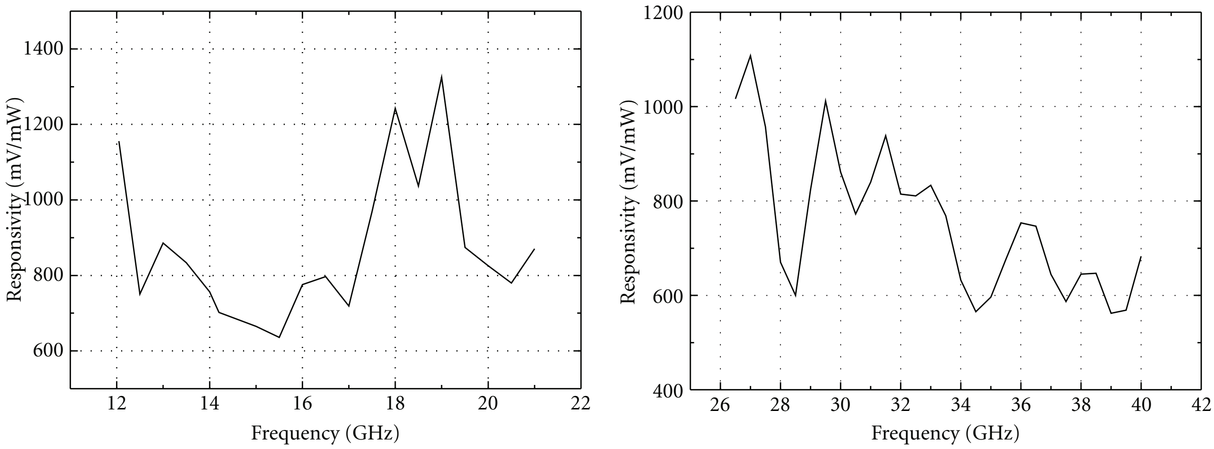

(a) Ku-band (b) Ka-band

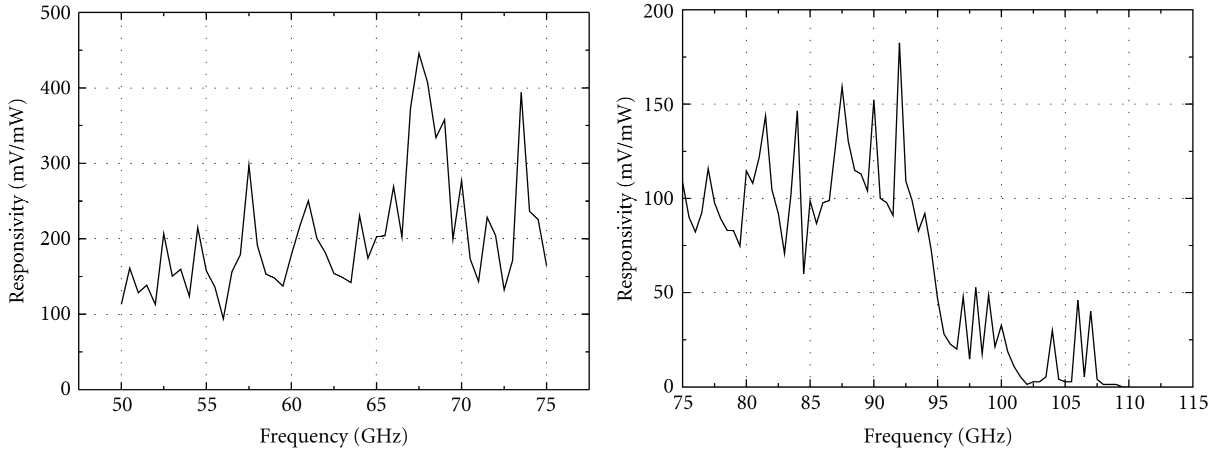

(c) V-band (d) W-band

Figure 16: Frequency responses test results of the THz wave planar integrated active microstrip antenna covered by a dielectric lens.

the advanced machining techniques pushing the microstrip antennas towards the highly integration of antenna/array and feeding network and the highly operating frequencies are described. To demonstrate the distinctive features of novel microstrip antennas, various antenna elements and arrays for different applications are presented. This paper has shown that the microstrip antennas are still very promising paradigm for civilian and military wireless applications.

References

[1] G. A. Deschamps, "Microstrip microwave antennas," in *Proceedings of the 3rd USAF Symposium on Antennas*, 1953.

[2] E. V. Byron, "A new flush-mounted antenna element for phased array application," in *Proceedings of the Phased Array Antenna Symposium*, pp. 187–192, 1970.

[3] J. Q. Howel, "Microstrip antennas," in *Proceedings of the Digest of the International Symposium of the Antennas and propagation Society*, pp. 177–180, 1972.

[4] I. J. Bahl and P. Bhartia, *Microstrip Antennas*, Artech House, 1980.

[5] J. R. James, P. S. Hall, and C. Wood, *Microstrip Antenna Theory and Design*, Peter Peregrinus, 1981.

[6] K. R. Carver and J. W. Mink, "Microstrip antenna technology," *IEEE Transactions on Antennas and Propagation*, vol. 1, no. 1, pp. 2–24, 1981.

[7] D. M. Pozar, "Microstrip antennas," *Proceedings of the IEEE*, vol. 80, no. 1, pp. 79–91, 1992.

[8] R. Garg, I. Bhartia, I. Bahl, and A. Ittipiboon, *Micro-Strip Antenna Design Handbook*, Artech House, Boston, Mass, USA, 2001.

[9] D. M. Pozar and B. Kaufman, "Increasing the bandwidth of a microstrip antenna by proximity coupling," *Electronics Letters*, vol. 23, no. 8, pp. 368–369, 1987.

[10] H. F. Pues and A. R. Van de Capelle, "Impedance-matching technique for increasing the bandwidth of microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 37, no. 11, pp. 1345–1354, 1989.

[11] A. Kaya and E. Y. Yüksel, "Investigation of a compensated rectangular microstrip antenna with negative capacitor and negative inductor for bandwidth enhancement," *IEEE Transactions on Antennas and Propagation*, vol. 55, no. 5, pp. 1275–1282, 2007.

[12] K. L. Wong and N. H. Hsu, "A broad-band rectangular patch antenna with a pair of wide slits," *IEEE Transactions on Antennas and Propagation*, vol. 49, no. 9, pp. 1345–1347, 2001.

[13] S. I. Latif, L. Shafai, and S. K. Sharma, "Bandwidth enhancement and size reduction of microstrip slot antennas," *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 3, pp. 994–1003, 2005.

[14] P. M. T. Ikonen, S. I. Maslovski, C. R. Simovski, and S. A. Tretyakov, "On artificial magnetodielectric loading for improving the impedance bandwidth properties of microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 6, pp. 1654–1662, 2006.

[15] R. Cocooli, F. R. Yang, T. P. Ma, and T. Itoh, "Apertu-recoupled patch antenna on UCPBG substrate," *IEEE Transactions on Microwave Theory and Techniaues*, vol. 1, no. 47, pp. 2123–2130, 1999.

[16] L.-M. Si and X. Lv, "CPW-FED multi-band omni-directional planar microstrip antenna using composite metamaterial resonators for wireless communications," *Progress in Electromagnetics Research*, vol. 83, pp. 133–146, 2008.

[17] L. M. Si, H. Sun, Y. Yuan, and X. Lv, "CPW-fed com-pact planar UWB antenna with circular disc and spiral split ring resonators," in *Progress in Electromagnetics Research Symposium*, pp. 502–505, 2009.

[18] Y. Yuan, L.-M. Si, Y. Liu, and X. Lv, "Integrated log-periodic antenna for Terahertz applications," in *International Conference on Microwave Technology and Computational Electromagnetics*, pp. 276–279, 2009.

[19] L. Xin, M. Jinchao, Y. Yong, Y. Weihua, N. Hongbin, and G. Yafen, "Integrated active antennas in terahertz focal plane imaging system," in *Proceedings of the International Workshop on Antenna Technology*, pp. 157–163, 2011.

[20] Q. Xu, W. Zhang, H. Sun, and X. Lu, "Millimetre wave multipolarised microstrip antenna array and application example," *IET Microwaves, Antennas and Propagation*, vol. 4, no. 10, pp. 1525–1530, 2010.

[21] M. Wei, H. Deng, H. Sun, and Y. Liu, "Design of an X/Ka dual-band co-aperture broadband microstrip antenna array," in *Proceedings IEEE International Conference on Microwave Technology and Computational Electromagnetics*, pp. 217–220, 2011.

[22] Y. Liu, X. Lu, Y. Yuan, Y. Chen, and L. Shi, "Integrated design and research of Ka-band electronically large mono-pulse planar antenna and feed system," in *Proceedings of the 7th International Symposium on Antennas, Propagation and EM Theory (ISAPE '06)*, pp. 150–153, October 2006.

[23] Y. Liu, H. J. Sun, X. Lu et al., "Study of the millimeter microstrip monopulse antenna array in the dual-sensor system," in *Proceedings of the 7th International Symposium on Antennas, Propagation and EM Theory (ISAPE '06)*, pp. 157–160, October 2006.

[24] Y. An, X. Lv, and B. Gao, "Developing a kind of mi-crostrip array antenna with beam squint," in *Proceedings of the 5th International Symposium on Antennas, Propagation and EM Theory*, pp. 443–446, 2000.

[25] L. Yong, L. Xin, H. Zhao, and Y. Yuan, "Integrated design and research of novel KA-band circular polarized monopulse interrogator array antenna," in *Proceedings of the IET International Radar Conference 2009*, pp. 1–4, April 2009.

[26] H. A. Wheeler, "Fundamental limitations of small an-tennas," *Proceedings of the IRE*, vol. 35, pp. 1479–1484, 1947.

[27] L. J. Chu, "Physical limitations of omni-directional antennas," *Journal of Applied Physics*, vol. 19, no. 12, pp. 1163–1175, 1948.

[28] R. Azadegan and K. Sarabandi, "A novel approach for miniaturization of slot antennas," *IEEE Transactions on Antennas and Propagation*, vol. 51, no. 3, pp. 421–429, 2003.

[29] H. Mosallaei and K. Sarabandi, "Magneto-dielectrics in electromagnetics: concept and applications," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 6, pp. 1558–1567, 2004.

[30] P. M. T. Ikonen, S. I. Maslovski, C. R. Simovski, and S. A. Tretyakov, "On artificial magnetodielectric loading for improving the impedance bandwidth properties of microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 6, pp. 1654–1662, 2006.

[31] P. M. T. Ikonen, K. N. Rozanov, A. V. Osipov, P. Alitalo, and S. A. Tretyakov, "Magnetodielectric substrates in antenna miniaturization: potential and limitations," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 11, pp. 3391–3399, 2006.

[32] P. Ikonen and S. Tretyakov, "On the advantages of magnetic materials in microstrip antenna miniaturization," *Microwave and Optical Technology Letters*, vol. 50, no. 12, pp. 3131–3134, 2008.

[33] J. Anguera, C. Puente, C. Borja, and J. Soler, "Frac-tal-shaped antennas: a review," in *Wiley: Encyclopedia of RF and Microwave Engineering*, K. Chang, Ed., vol. 2, pp. 1620–1635, 2005.

[34] L. Shi, H. Sun, X. Lu, C. Deng, and Y. Liu, "Study of the improved MM-Wave omni-directional microstrip antenna," in *Proceedings of the 7th International Symposium on Antennas, Propagation and EM Theory (ISAPE '06)*, pp. 238–241, October 2006.

[35] L. Shi, H. Sun, and X. Lu, "A composite antenna with wide circularly polarized beamwidth," *Microwave and Optical Technology Letters*, vol. 51, no. 10, pp. 2461–2464, 2009.

[36] L. Shi, H. Sun, W. Dong, and X. Lv, "A dual-band multifunction carborne hybrid antenna for satellite communication relay system," *Progress in Electromagnetics Research*, vol. 95, pp. 329–340, 2009.

[37] Y. Chen, H. Sun, and X. Lv, "Novel design of dual-polarization broad-band printed L-shaped probe fed microstrip patch antenna for SAR applications," in *IET International Radar Conference 2009*, pp. 1–4, April 2009.

[38] L. M. Si, H. J. Sun, and X. Lv, "Numerical simulations of backward-to-forward leaky-wave antenna with composite right/left-handed coplanar waveguide," *Chinese Physics Letters*, vol. 27, no. 3, Article ID 034106, 2010.

[39] L.-M. Si, H.-J. Sun, and X. Lv, "THz leaky-wave antenna with high-directivity and beam-steering using CPW CRLH metamaterial resonators," in *Proceedings of the 3rd International Symposium on Photoelectronic Detection and Imaging*, Beijing, China, 2009.

[40] Y. Liu, L. M. Si, S. Zhu, and H. Xin, "Experimental realization of an integrated THz electromagnetic crys-tals (EMXT) H-plane horn antenna," *Electronics Letters*, vol. 47, no. 2, pp. 80–82, 2011.

[41] L. M. Si, Y. Liu, S. H. Zhu, and H. Xin, "Integrated THz horn antenna using EBG structures," in *Proceedings of the IEEE International Symposium on Antennas and Propagation and USNC/CNC/URSI National Radio Science Meeting*, University of Colorado at Boulder, 2011.

[42] Y. Liu, Y. Yuan, X. Lv, and L.-M. Si, "Design of 0.5 THz 2D square lattice EBG waveguide transmission line and power-divider using MEMS technology," in *Proceedings 3rd International Symposium on Photoelectronic Detection and Imaging*, vol. 7385, Beijing, China, 2009.

[43] L.-M. Si, Y. Yong, H.-J. Sun, and L. Xin, "Characterization and application of planar terahertz narrow bandpass filter with metamaterial resonators," in *Proceedings of the International*

Workshop on Metamaterials, pp. 351–354, Nanjing, China, 2008.

[44] L. M. Si and X. Lv, "Terahertz waves hairpin microstrip bandpass filter and its application to overlaid dielectric material detection," *Modern Physics Letters B*, vol. 22, no. 29, pp. 2843–2848, 2008.

[45] L.-M. Si, H.-J. Sun, and X. Lv, "Theoretical investigation of terahertz amplifier by carbon nanotubes within transmission line metamaterials," *Microwave and Optical Technology Letters*, vol. 53, no. 4, pp. 815–818, 2011.

[46] W. Hong, "Millimeter wave and THz communications in China," in *Proceedings of the IEEE International Microwave Symposium Digest (MTT '11)*, pp. 1–4, 2011.

Journal of Engineering
The Scientific World Journal
International Journal of Rotating Machinery
Journal of Sensors
International Journal of Distributed Sensor Networks
Advances in Civil Engineering
Journal of Control Science and Engineering
Journal of Robotics
Journal of Electrical and Computer Engineering
Advances in OptoElectronics
VLSI Design
International Journal of Navigation and Observation
Modelling & Simulation in Engineering
International Journal of Aerospace Engineering
International Journal of Chemical Engineering
International Journal of Antennas and Propagation
Active and Passive Electronic Components
Shock and Vibration
Advances in Acoustics and Vibration

Hindawi

Submit your manuscripts at
http://www.hindawi.com

A 2.4 GHz Polarization-diversity Planar Printed Dipole Antenna for WLAN and Wireless Communication Applications

This article presents the design simulation, fabrication and measured performance of a 2.4 GHz polarization-diversity printed dipole antenna for wireless communication applications. Two orthogonal printed dipole antennas, each with a microstrip via-hole balun for vertical and horizontal polarization, are combined and fabricated on a PCB substrate. PIN diodes are used as switches to select the desired antenna polarization. The 3D finite-element-method (FEM) electromagnetic EM simulator, HFSS, is used in the design simulation of this planar antenna structure. Numerical and measured results of the antenna radiation characteristics, including input SWR, radiation pattern coverage and polarization-diversity, are presented and compared.

In wireless communication systems, such as wireless local area networks (WLAN), research and development efforts are aiming at smaller size and better performance. In addition to the use of signal processing techniques to improve communication channel capacity, the radiation characteristics of the portable antenna system is also very important for communication performance.

In urban or indoor environments, the radio wave will propagate through complicated reflection or scattering processes. The polarization of the radio wave may change significantly. In order to effectively receive the communications signal, a polarization-diversity antenna for wireless communications may become an important requirement. A polarization-diversity antenna may have a pair of linearly-polarized antennas, and the radio signal received on both antenna is sampled and compared at

HUEY-RU CHUANG, LIANG-CHEN KUO,
CHI-CHANG LIN AND WEN-TZU CHEN
National Cheng Kung University,
Tainan, Taiwan

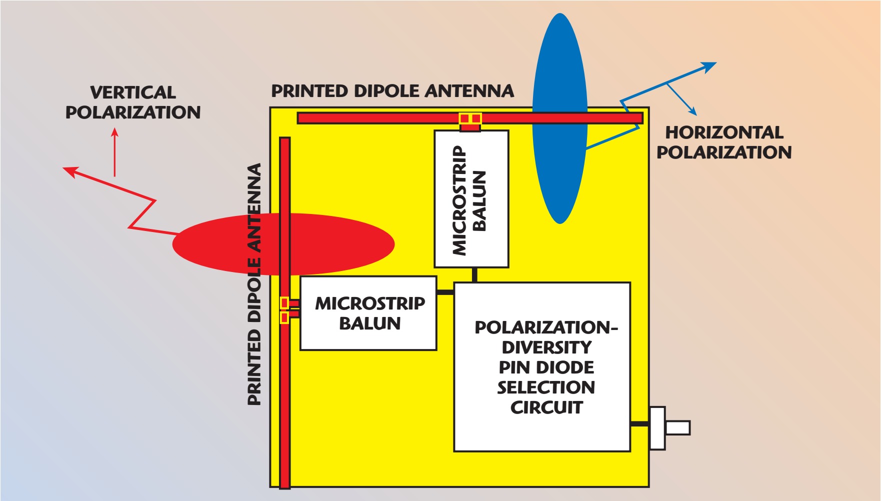
VERTICAL POLARIZATION
PRINTED DIPOLE ANTENNA
HORIZONTAL POLARIZATION
PRINTED DIPOLE ANTENNA
MICROSTRIP BALUN
MICROSTRIP BALUN
POLARIZATION-DIVERSITY PIN DIODE SELECTION CIRCUIT

▲ *Fig. 1 A 2.4 GHz polarization-diversity antenna.*

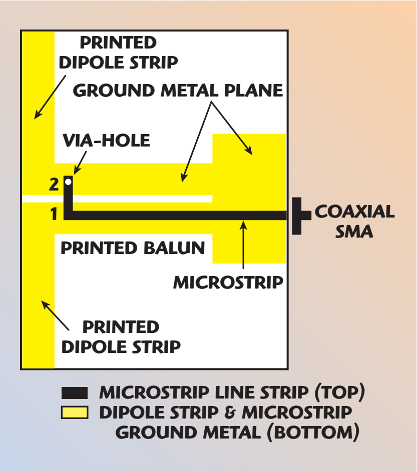
PRINTED DIPOLE STRIP
GROUND METAL PLANE
VIA-HOLE
2
1
COAXIAL SMA
PRINTED BALUN
MICROSTRIP
PRINTED DIPOLE STRIP
MICROSTRIP LINE STRIP (TOP)
DIPOLE STRIP & MICROSTRIP GROUND METAL (BOTTOM)

▲ *Fig. 2 Printed dipole antenna with a microstrip balun.*

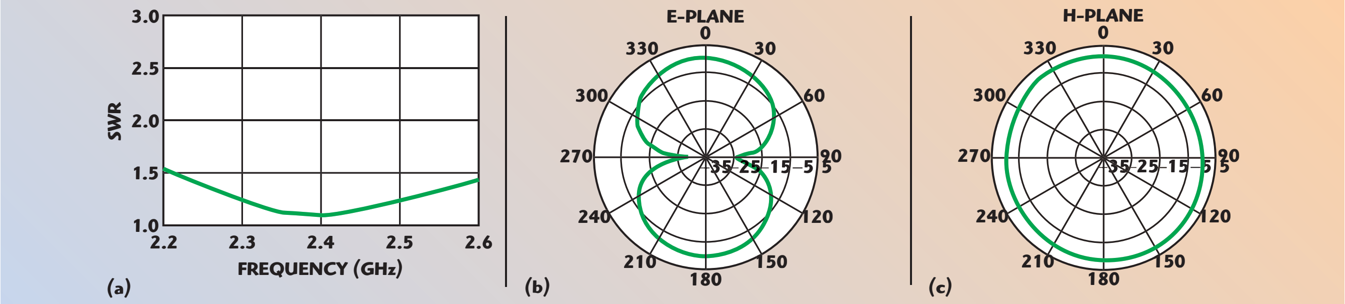

▲ *Fig. 3 Simulated performance of a 2.4 GHz printed dipole antenna placed horizontally; (a) input SWR, (b) E-plane pattern and (c) H-plane pattern.*

certain time intervals. Then the antenna with the best signal quality is selected.

A typical dipole antenna radiates a vertically polarized EM wave and has an omnidirectional antenna pattern. In order to have a preferred planar antenna structure for this 2.4 GHz polarization-diversity antenna, a printed dipole antenna with a microstrip via-hole balun is designed. As shown in ***Figure 1***, two orthogonal printed dipole antennas, for vertical and horizontal polarization, respectively, are combined and fabricated on a PCB substrate. PIN diodes are used to switch and select the desired antenna polarization.

In the antenna design, the high frequency structure simulator (HFSS), based on a 3D FEM, was employed for design simulation of the complete printed dipole structure. A printed dipole antenna and a polarization-diversity planar dipole antenna board (with a polarization-selection PIN diode circuit) have been fabricated on FR-4 PCB substrates. A complete 3D structure FEM simulation and the measured performance of the realized printed dipole-antenna are compared. The measured radiation characteristics of the polarization-diversity planar dipole antenna, including input SWR, radiation pattern coverage and polarization diversity, are presented.

PRINTED DIPOLE ANTENNA WITH MICROSTRIP BALUN

As shown in ***Figure 2***, a printed dipole antenna has a printed microstrip balun which acts as an unbalanced-to-balanced transformer from the feed coaxial line to the two printed dipole strips. The length of the dipole strip and the balun microstrip are both about 1/4 wavelength. The ground plane of the microstrip line and the dipole antenna strips are in the same plane. A via-hole permits the feed point 2 of a printed dipole strip to have the same phase as the feed point 1 of the other printed dipole strip. Due to the 180° phase difference between the top strip and the ground plane of the microstrip line, the feed point 2 of the printed dipole strip will have 180° phase difference with the other feed point 1. Accurate dimensions of the printed dipole strip and the microstrip balun structure are determined by numerical simulation, using HFSS.

The simulation results for a 2.4 GHz printed dipole antenna placed horizontally with a microstrip via-hole balun and fabricated on an FR-4 substrate are shown in ***Figure 3***. The input SWR is less than 1.5 from 2.2 to 2.6 GHz. The simulated E- and H-plane antenna patterns are very close to those of an ideal dipole antenna, where the H-plane pattern is omnidirectional. ***Figure 4*** is a photograph of a realized antenna. The measured input SWR and antenna patterns (measured with the dipole placed vertically) agree well with the simulation results, as shown in ***Figure 5***.

PLANAR POLARIZATION-DIVERSITY PRINTED DIPOLE ANTENNA

Figure 6 shows photographs of a realized 2.4 GHz planar polarization-diversity antenna consisting of two orthogonal printed dipole antennas with a polarization-switched PIN diode circuit. Each printed dipole has

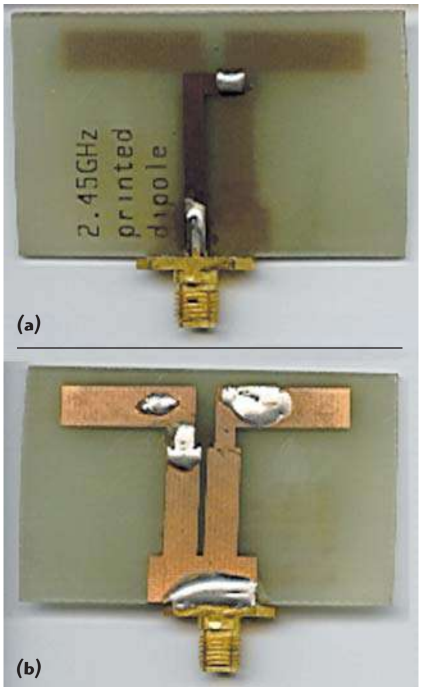

▲ *Fig. 4 A 2.4 GHz printed dipole antenna with a microstrip via-hole balun; (a) top view and (b) bottom view.*

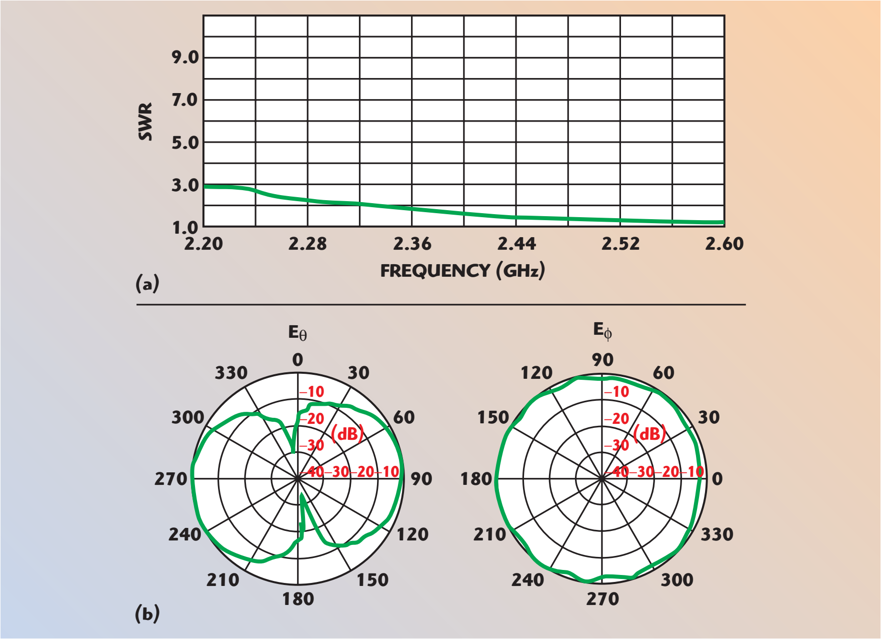

▲ *Fig. 5 Measured input SWR (a) and radiation patterns (b).*

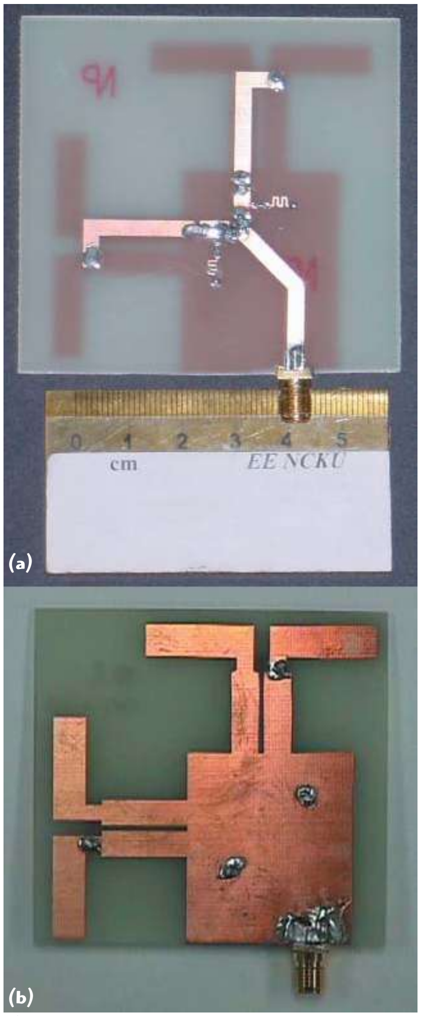

▲ *Fig. 6 A 2.4 GHz planar polarization-diversity antenna with a polarization-switched PIN diode circuit; (a) top view and (b) bottom view.*

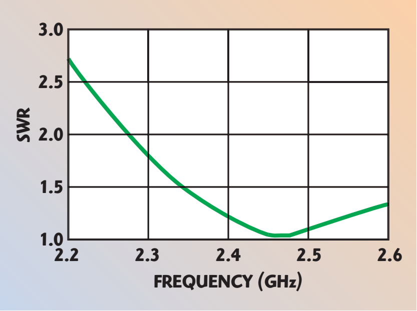

▲ *Fig. 7 Input SWR simulation of a 2.5 GHz polarization-diversity dipole antenna with the vertical dipole selected.*

a microstrip via-hole balun. The terminals of the two baluns are connected to a PIN diode selection circuit. Voltages from the transceiver circuit (±5.0V) are fed through a cable to the input of the PIN diode circuit section, to short or open-circuit the PIN diodes. Hence, either the vertical or horizontal printed dipole can be selected and connected to the transceiver.

Since the two dipoles are very close to each other and near the PIN diode circuit section, EM coupling will degrade the performance of each dipole. ***Figure 7*** shows the input SWR simulation results with the vertical dipole antenna selected (+5V to PIN diode switching circuit). The input SWR is less than 2 from 2.25 to 2.60 GHz. The simulated E- and H-plane antenna patterns are all very close to those of an ideal dipole antenna, of which the H-plane pattern is still omnidirectional, as shown in ***Figure 8***. Note that the dominant polarization is the vertical (E_θ) field, which agrees with the selection of the vertical dipole. The antenna pattern has some attenuation in the direction of the PIN diode circuit board. It can also be seen that a certain level of the input RF signal is induced to the horizontal antenna path by EM coupling, which generates some level of cross-polarization field. ***Figure 9*** shows the simulation results with the horizontal dipole antenna selected (–5V to PIN diode switching circuit). Results similar to the ones obtained for the vertical dipole antenna can be observed, except that the dominant polarization is the horizontal (E_ϕ) field, which agrees with the selection of the horizontal dipole.

The measured antenna input SWR with vertical or horizontal dipole selection confirms the input SWR of each dipole antenna (through the PIN diode selection circuit) is less than 1.5 from 2.2 to 2.6 GHz, which agrees with the HFSS simulation results. The measured antenna patterns with the selection of the vertical or horizontal dipole shows that for the selection of the vertical dipole, the H-plane pattern is still quite omnidirectional (as an ideal vertical dipole) with some attenuation in the direc-

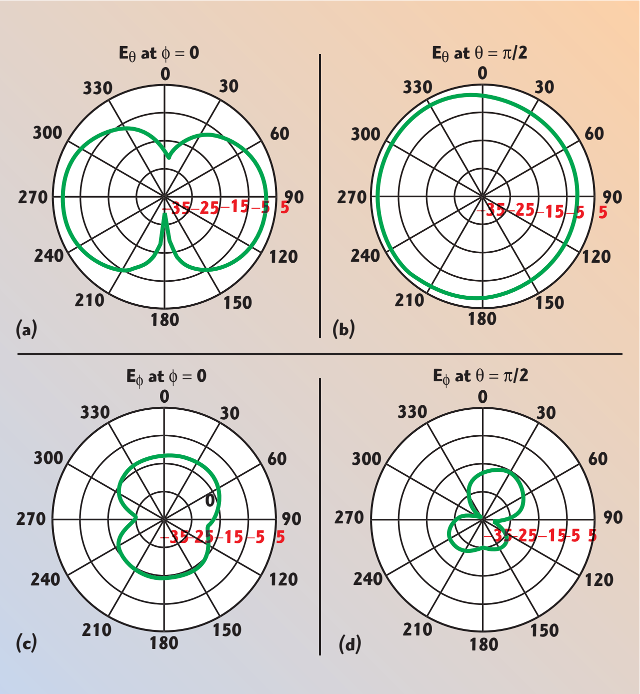

Fig. 8 Simulation of a 2.5 GHz polarization-diversity dipole antenna with the vertical dipole selected; (a) E_θ-field E-plane pattern, (b) E_θ-field H-plane pattern, (c) E_ϕ-field (cross-polarization) E-plane pattern and (d) E_ϕ-field (cross-polarization) H-plane pattern.

tion of the PIN diode circuit board. ***Figures 10*** and ***11*** show the mesured SWR and antenna patterns, respectively. A certain level of the induced cross-polarization pattern is observed as predicted by the HFSS simulation due to the proximity of the horizontal dipole strip and the PIN diode circuit board. As for the selection of the horizontal dipole, the E-plane pattern is also close to that of an ideal horizontal dipole. Also, the induced cross-polarization pattern is observed, which is the same situation as the selection of the vertical dipole. The measured data shows a good agreement with the HFSS simulation results and how the antenna polarization-diversity is working.

CONCLUSION

3D FEM design simulation, realization and measurements of a 2.4 GHz printed dipole antenna (with a microstrip via-hole balun) and a planar polarization-diversity printed dipole antenna are presented. The planar polarization-diversity antenna consists of two orthogonal printed dipole antennas for vertical and horizontal polarization and is fabricated on a FR-4 PCB board. A PIN diode switching circuit is used to select the desired antenna polarization. Satisfactory agreement between simulation and measurements is observed. The measured input SWR of the realized printed dipole antenna is less than 1.5 from 2.2 to 2.6 GHz. The measured input SWR of the vertical and horizontal dipole (through the PIN diode switching circuit) of the realized planar polarization-diversity antenna is less than 1.5 from 2.3 to 2.6 GHz. The measured E- and H-plane patterns of the polarization-diversity antenna show that the selected vertical or horizontal dipole have a performance close to a single dipole antenna in a vertical or horizontal position. The designed planar polarization-diversity antenna can be used for wireless communication and WLAN applications.

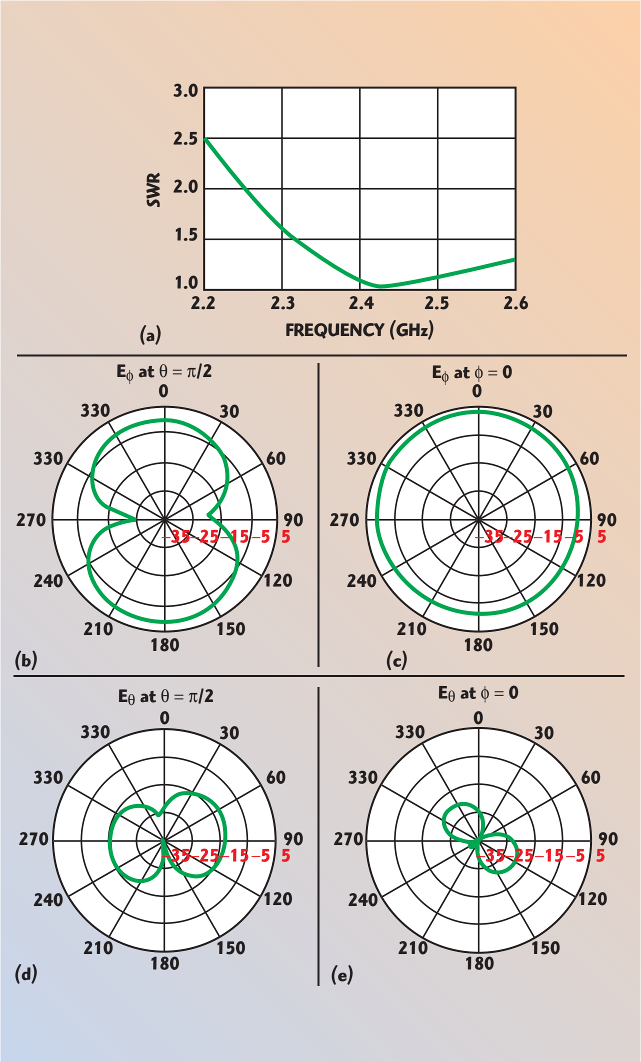

Fig. 9 Simulation of a 2.4 GHz polarization-diversity printed dipole antenna (with the horizontal dipole selected); (a) input SWR, (b) E_ϕ-field E-plane pattern, (c) E_ϕ-field H-plane pattern, (d) E_θ-field (cross-polarization) E-plane pattern and (e) E_θ-field (cross-polarization) H-plane pattern.

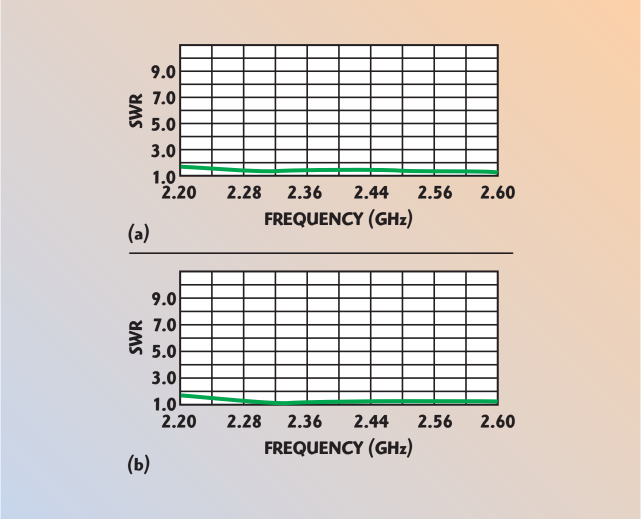

Fig. 10 Measured input SWR of a 2.4 GHz polarization-diversity printed dipole antenna; (a) vertical dipole selection and (b) horizontal dipole section.

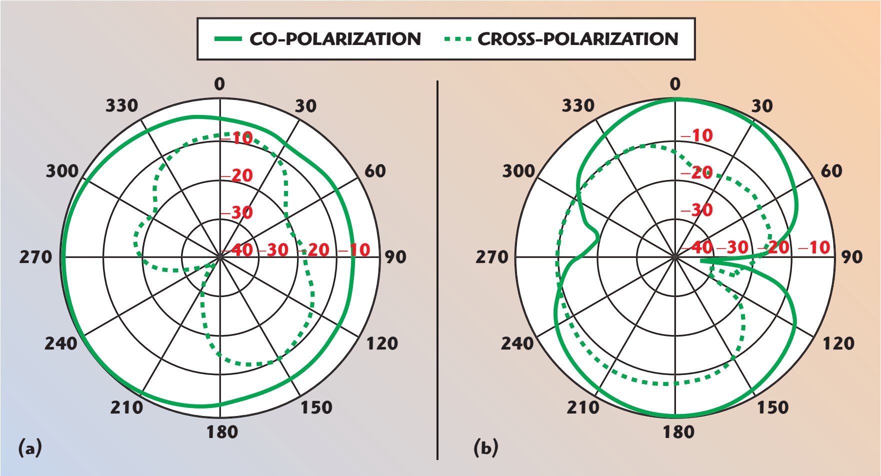

▲ *Fig. 11 Measured co-and cross-polarized patterns of a 2.4 GHz polarization-diversity printed dipole antenna; (a) vertical dipole selected and (b) horizontal dipole selected.*

ACKNOWLEDGMENT

The authors would like to thank Ansoft Inc. for its support of the HFSS software. ■

References

1. K. Fujimoto and J. R. James, *Mobile Antenna Systems Handbook*, Artech House Inc., Norwood, MA 1994.
2. B. Edward and D. Rees, "A Broadband Printed Dipole with Integrated Balun," *Microwave Journal*, May 1987, pp. 339–344.
3. K. Hettak, G.Y. Delisle and M.G. Stubbs, "A Novel Variant of Dual Polarized CPW Fed Patch Antenna for Broadband Wireless Communications," *IEEE Antennas and Propagation Society International Symposium Digest*, Vol.1, 2000, pp. 286–289.
4. L. Zhu and K. Wu, "Model-based Characterization of CPS-fed Printed Dipole for Innovative Design of Uniplanar Integrated Antenna," *IEEE Microwave and Guided Wave Letters*, Vol. 9, No. 9, September 1999, pp. 342–344.
5. N. Michishitai and H. Arai, "A Polarization-diversity Antenna by Printed Dipole and Patch With a Hole," *IEEE Antennas and Propagation Society International Symposium Digest*, 2001, pp. 368–371.

Huey-Ru Chuang *received his BSEE and MSEE degrees from National Taiwan University, Taipei, Taiwan, in 1977 and 1980, respectively, and his PhD degree in electrical engineering from Michigan State University, East Lansing, MI, in 1987. From 1987 to 1988, he was a post-doctoral research associate at the Engineering Research Center of Michigan State University. From 1988 to 1990, he was with the Portable Communication Division of Motorola Inc., Ft. Lauderdale, FL. He joined the department of electrical engineering of National Cheng Kung University, Tainan, Taiwan, in 1991, where he is currently a professor. His research interests include portable antenna design, RF/microwave circuits and RFIC/MMIC for wireless communications, electromagnetic computation of human interaction with mobile antennas, EMI/EMC, microwave communication and detection systems.*

Liang-Chen Kuo *received his BSEE degree from Nan-Tai Institute of Technology, Tainan, Taiwan, and his MSEE degree from Tatung Institute of Technology, Taipei, Taiwan, in 1987 and 1996, respectively. He is currently working toward his PhD degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan. His research interests include computational electromagnetics and antenna design.*

Chi-Chang Lin *received his BSEE and MSEE degrees from Tatung Institute of Technology, Taipei, Taiwan, in 1999 and 2001, respectively. He is currently working toward his PhD degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan. His research interests include EM simulation and microwave antenna design.*

Wen-Tzu Chen *received his PhD degree from National Cheng Kung University, Tainan, Taiwan, in 1998. He is currently an assistant professor at the Institute of Computer and Communication, Shu-Te University, Yen Chau, Taiwan. His research interests include numerical computation of EM interaction between the antenna and the human body, and microwave antenna design.*

射频和天线设计培训课程推荐

易迪拓培训(www.edatop.com)由数名来自于研发第一线的资深工程师发起成立，致力并专注于微波、射频、天线设计研发人才的培养；我们于 2006 年整合合并微波 EDA 网(www.mweda.com)，现已发展成为国内最大的微波射频和天线设计人才培养基地，成功推出多套微波射频以及天线设计经典培训课程和 ADS、HFSS 等专业软件使用培训课程，广受客户好评；并先后与人民邮电出版社、电子工业出版社合作出版了多本专业图书，帮助数万名工程师提升了专业技术能力。客户遍布中兴通讯、研通高频、埃威航电、国人通信等多家国内知名公司，以及台湾工业技术研究院、永业科技、全一电子等多家台湾地区企业。

易迪拓培训推荐课程列表： http://www.edatop.com/peixun/tuijian/

射频工程师
养成培训课程套装
易迪拓培训
www.edatop.com

射频工程师养成培训课程套装

该套装精选了射频专业基础培训课程、射频仿真设计培训课程和射频电路测量培训课程三个类别共 30 门视频培训课程和 3 本图书教材；旨在引领学员全面学习一个射频工程师需要熟悉、理解和掌握的专业知识和研发设计能力。通过套装的学习，能够让学员完全达到和胜任一个合格的射频工程师的要求…

课程网址：http://www.edatop.com/peixun/rfe/110.html

手机天线设计培训视频课程

该套课程全面讲授了当前手机天线相关设计技术，内容涵盖了早期的外置螺旋手机天线设计，最常用的几种手机内置天线类型——如 monopole 天线、PIFA 天线、Loop 天线和 FICA 天线的设计，以及当前高端智能手机中较常用的金属边框和全金属外壳手机天线的设计；通过该套课程的学习，可以帮助您快速、全面、系统地学习、了解和掌握各种类型的手机天线设计，以及天线及其匹配电路的设计和调试...

手机天线设计培训
中文视频教程
主讲：李明洋
易迪拓培训

课程网址： http://www.edatop.com/peixun/antenna/133.html

hp 企业内训课程
WiFi天线和射频电路设计培训

WiFi 和蓝牙天线设计培训课程

该套课程是李明洋老师应邀给惠普 (HP)公司工程师讲授的 3 天员工内训课程录像，课程内容是李明洋老师十多年工作经验积累和总结，主要讲解了 WiFi 天线设计、HFSS 天线设计软件的使用，匹配电路设计调试、矢量网络分析仪的使用操作、WiFi 射频电路和 PCB Layout 知识，以及 EMC 问题的分析解决思路等内容。对于正在从事射频设计和天线设计领域工作的您，绝对值得拥有和学习！…

课程网址：http://www.edatop.com/peixun/antenna/134.html

CST 学习培训课程套装

示意图
DVD
CST视频培训课程套装
李明洋主讲
易迪拓培训
www.edatop.com

该培训套装由易迪拓培训联合微波 EDA 网共同推出，是最全面、系统、专业的 CST 微波工作室培训课程套装，所有课程都由经验丰富的专家授课，视频教学，可以帮助您从零开始，全面系统地学习 CST 微波工作的各项功能及其在微波射频、天线设计等领域的设计应用。且购买该套装，还可超值赠送 3 个月免费学习答疑…

课程网址：http://www.edatop.com/peixun/cst/24.html

HFSS 学习培训课程套装

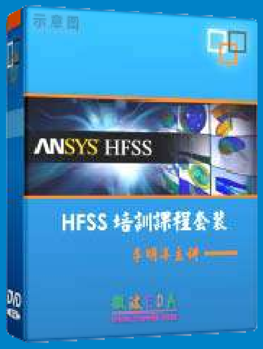
示意图
ANSYS HFSS
HFSS培训课程套装
微波EDA

该套课程套装包含了本站全部 HFSS 培训课程，是迄今国内最全面、最专业的 HFSS 培训教程套装，可以帮助您从零开始，全面深入学习 HFSS 的各项功能和在多个方面的工程应用。购买套装，更可超值赠送 3 个月免费学习答疑，随时解答您学习过程中遇到的棘手问题，让您的 HFSS 学习更加轻松顺畅…

课程网址：http://www.edatop.com/peixun/hfss/11.html

ADS 学习培训课程套装

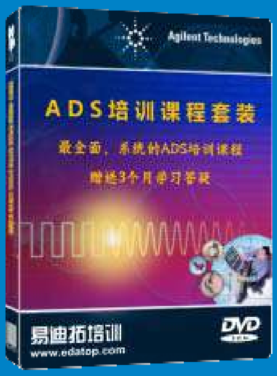
Agilent Technologies
ADS培训课程套装
最全面、系统的ADS培训课程
赠送3个月学习答疑
易迪拓培训
www.edatop.com
DVD

该套装是迄今国内最全面、最权威的 ADS 培训教程，共包含 10 门 ADS 学习培训课程。课程是由具有多年 ADS 使用经验的微波射频与通信系统设计领域资深专家讲解，并多结合设计实例，由浅入深、详细而又全面地讲解了 ADS 在微波射频电路设计、通信系统设计和电磁仿真设计方面的内容。能让您在最短的时间内学会使用 ADS，迅速提升个人技术能力，把 ADS 真正应用到实际研发工作中去，成为 ADS 设计专家...

课程网址： http://www.edatop.com/peixun/ads/13.html

我们的课程优势：

※ 成立于 2004 年，10 多年丰富的行业经验，

※ 一直致力并专注于微波射频和天线设计工程师的培养，更了解该行业对人才的要求

※ 经验丰富的一线资深工程师讲授，结合实际工程案例，直观、实用、易学

联系我们：

※ 易迪拓培训官网：http://www.edatop.com

※ 微波 EDA 网：http://www.mweda.com

※ 官方淘宝店：http://shop36920890.taobao.com

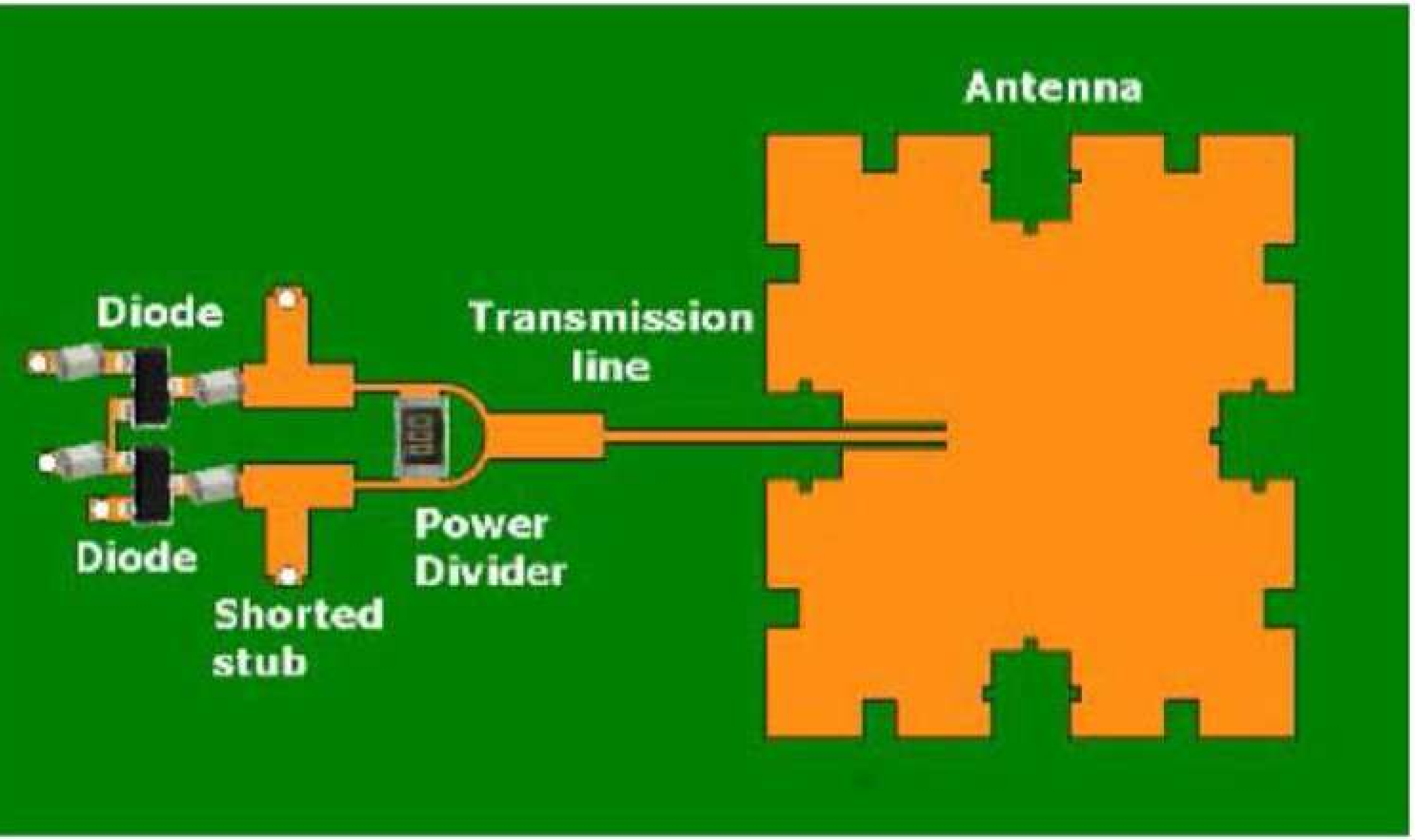
Antenna
Diode
Transmission line
Diode
Power Divider
Shorted stub

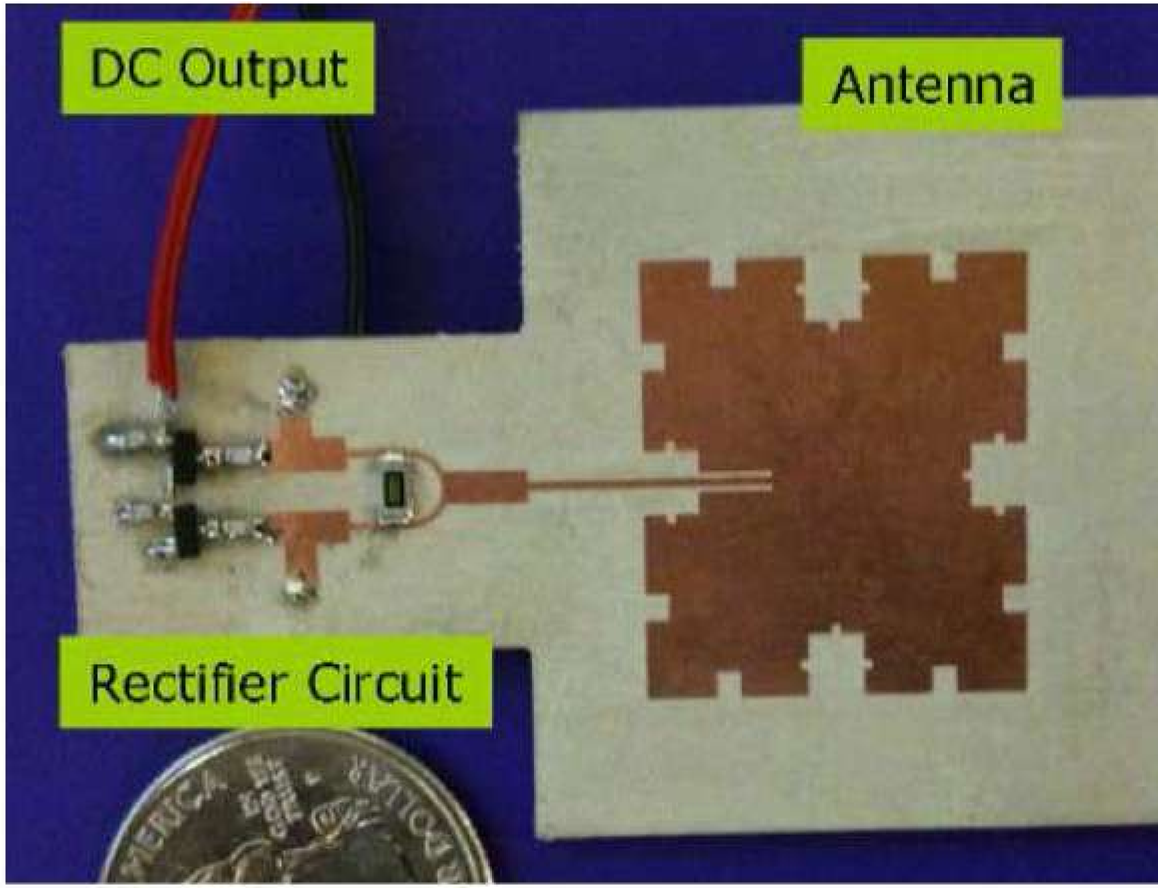
DC Output
Antenna
Rectifier Circuit

Giedrius Rakauskas

Follow

Phased Patch Antenna Array Design

Engineering

1473 17 0

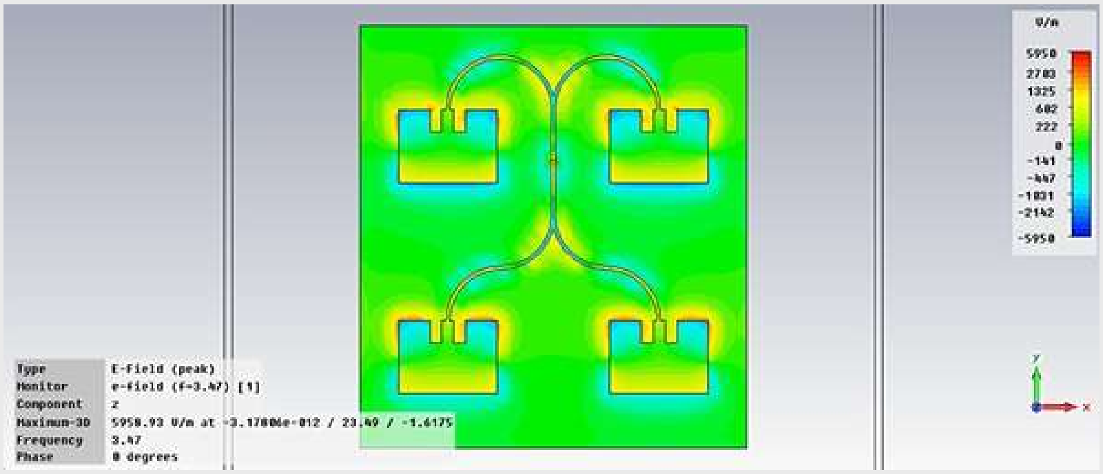

Type E-Field (peak)
Monitor e-field (f=3.47) [1]
Component z
Maximum-3D 5958.93 V/m at -3.17806e-012 / 23.49 / -1.6175
Frequency 3.47
Phase 0 degrees

Third Year Project

The aim of this project was to design a directive phased patch antenna array that could be implemented in modern vehicles for distance measurement and autonomous cruise control.

The particular technology allows improved performance as well as reduced costs and size.

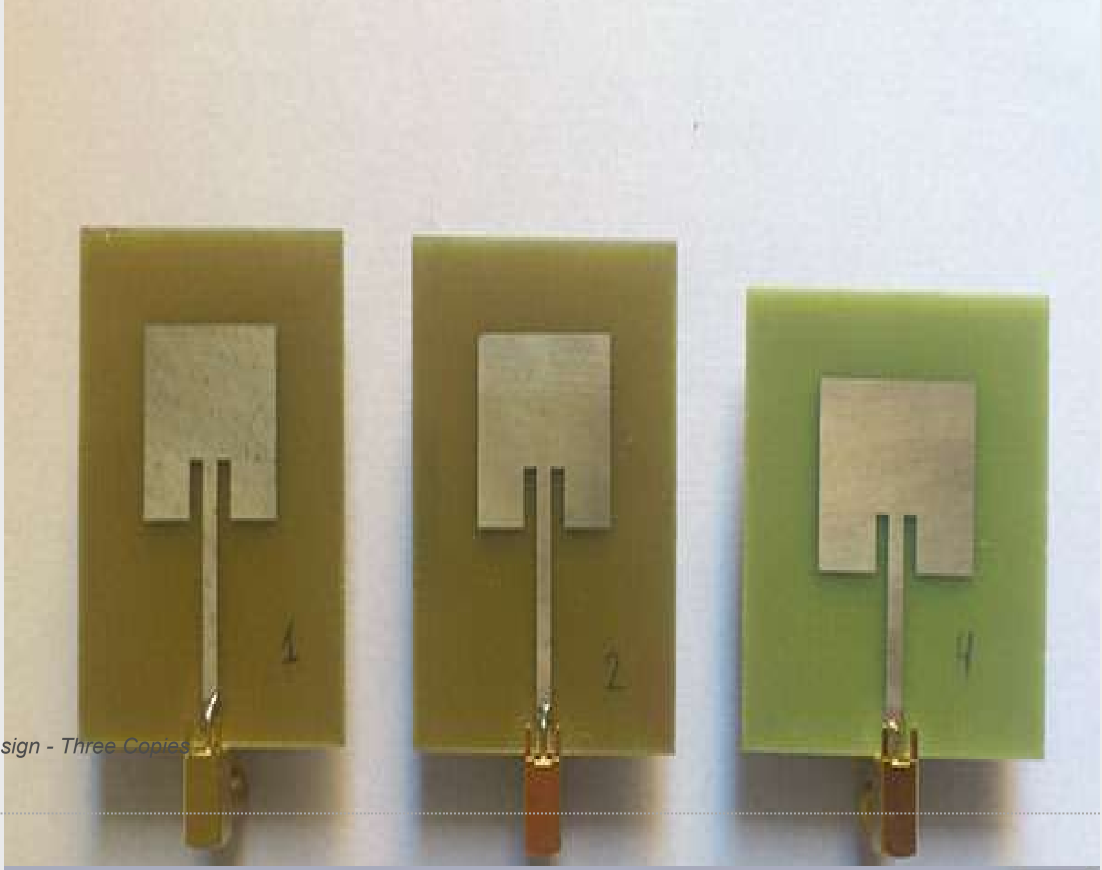

Initial Design - Three Copies

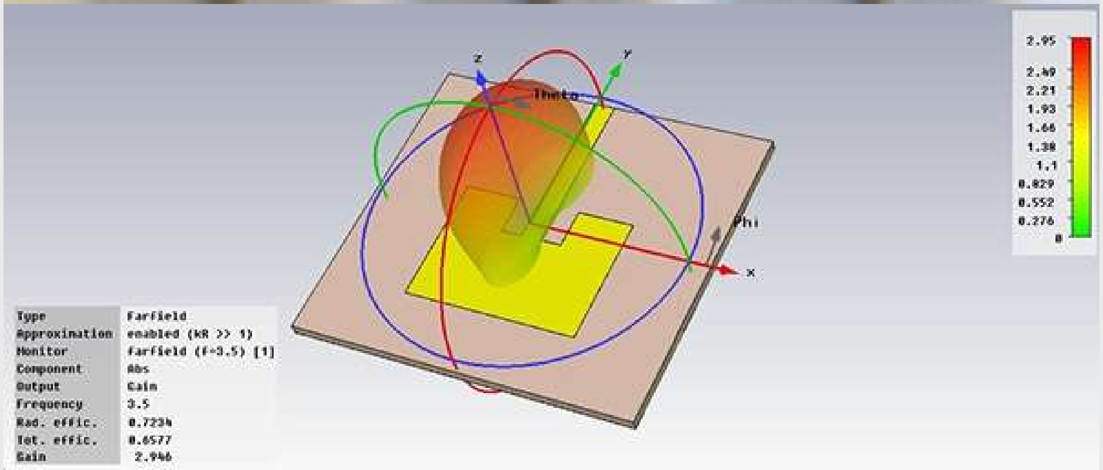
z
y
Theta
Phi
x
2.95
2.49
2.21
1.93
1.66
1.38
1.1
0.829
0.552
0.276
0

Type	Farfield
Approximation	enabled (kR >> 1)
Monitor	farfield (f=3.5) [1]
Component	Abs
Output	Gain
Frequency	3.5
Rad. effic.	0.7234
Tot. effic.	0.6577
Gain	2.946

Square Patch Antenna - Farfield Simulation Results

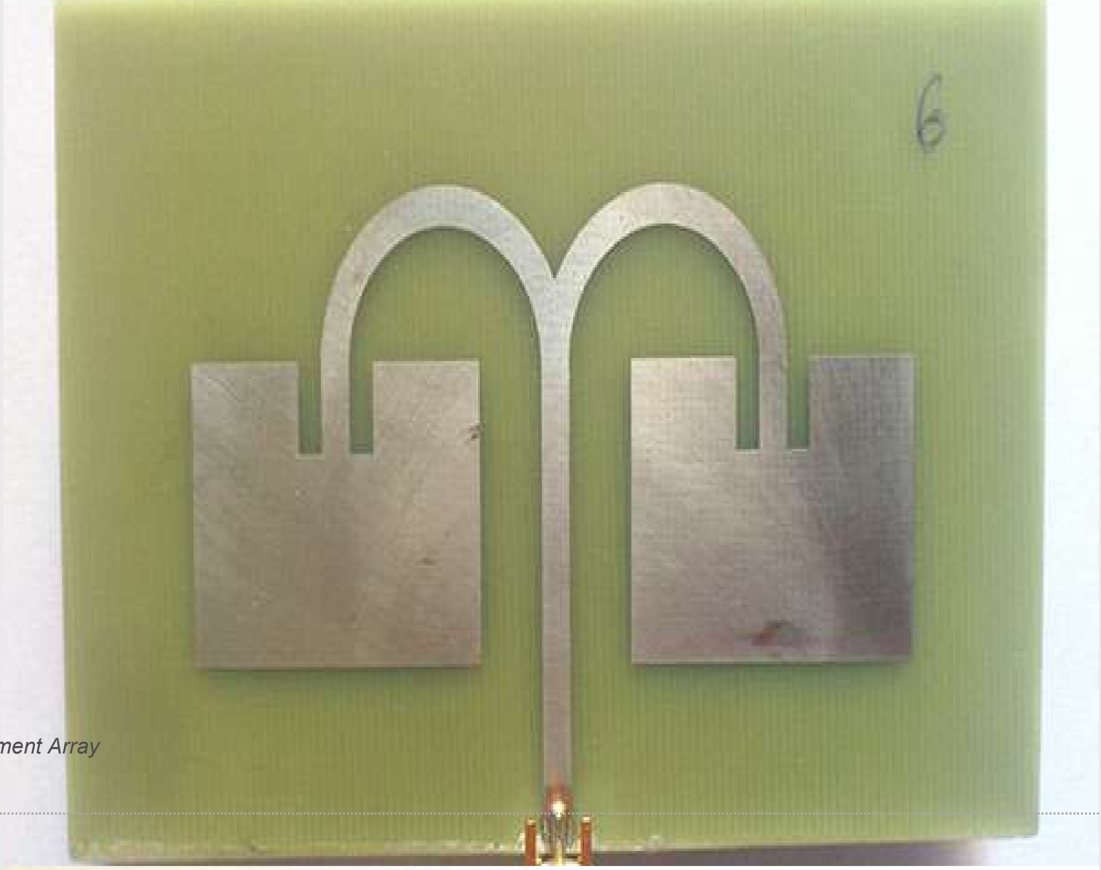

Two-Element Array

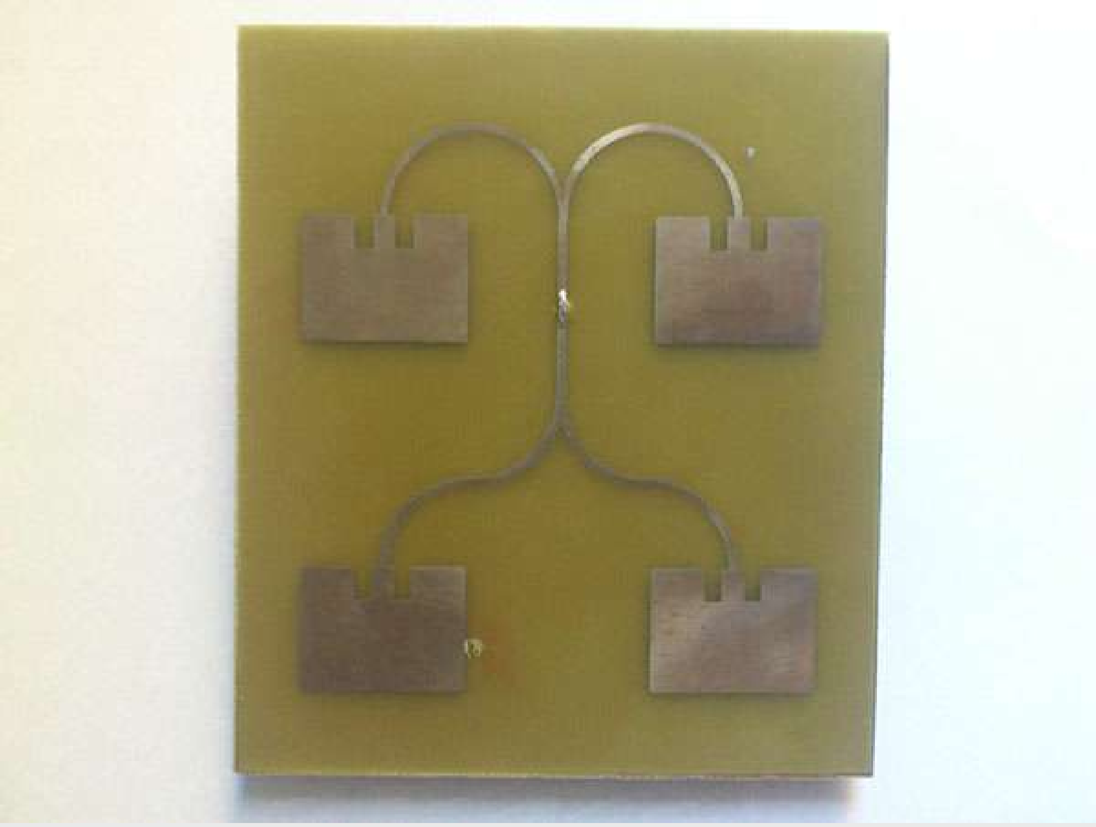

Four-Element Array

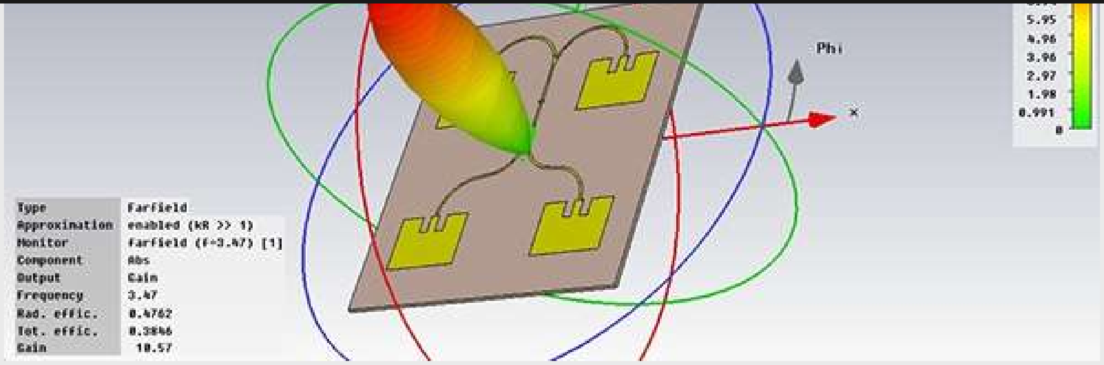
5.95
4.96
3.96
2.97
1.98
0.991
0
Phi
x

Type	Farfield
Approximation	enabled (kR >> 1)
Monitor	farfield (f=3.47) [1]
Component	Abs
Output	Gain
Frequency	3.47
Rad. effic.	0.4762
Tot. effic.	0.3846
Gain	10.57

Four-Element Array - CST Microwave Studio Simulation Results

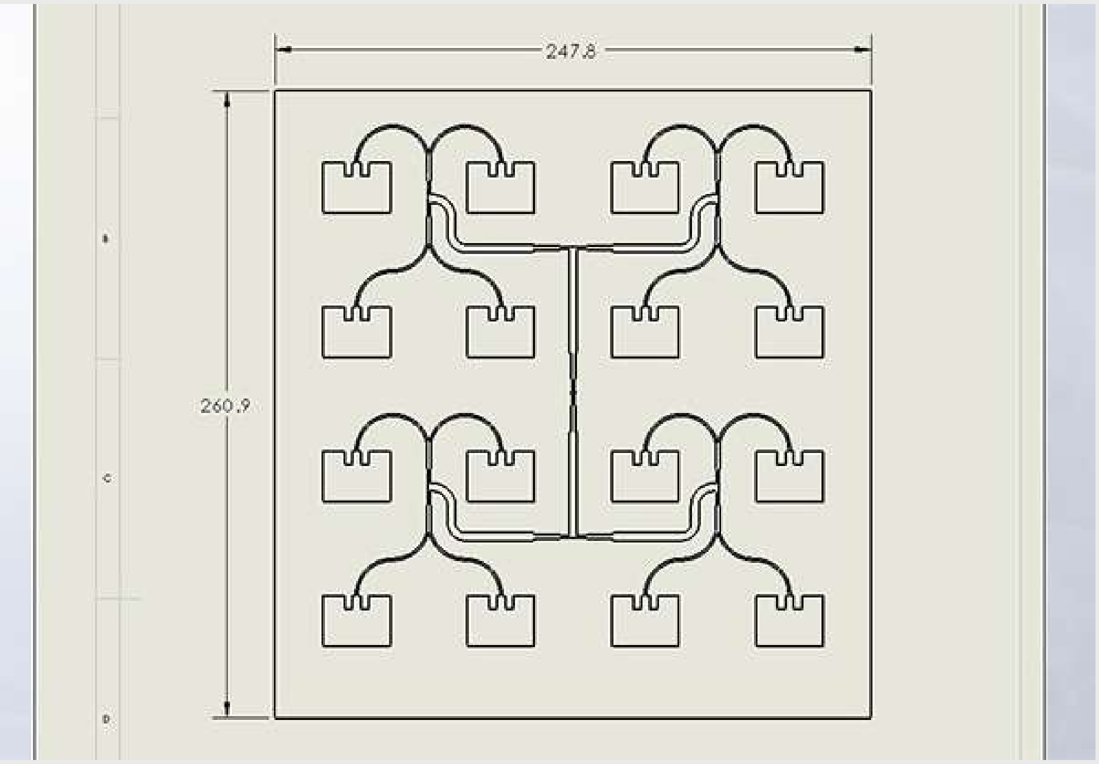
247.8
260.9
B
C
D

Sixteen-Element Array Design

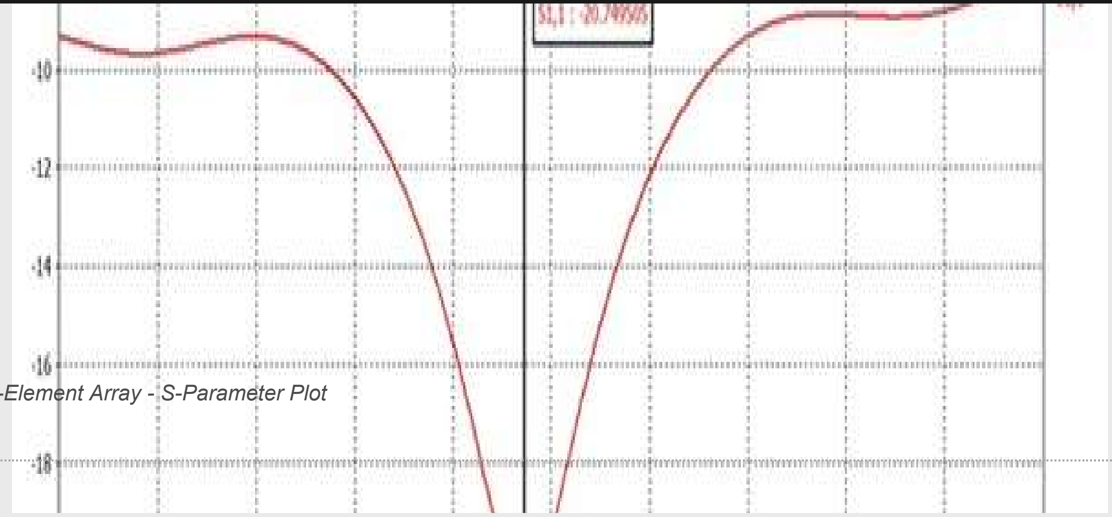

Sixteen-Element Array - S-Parameter Plot

Giedrius Rakauskas

Vilnius, Lithuania

Follow

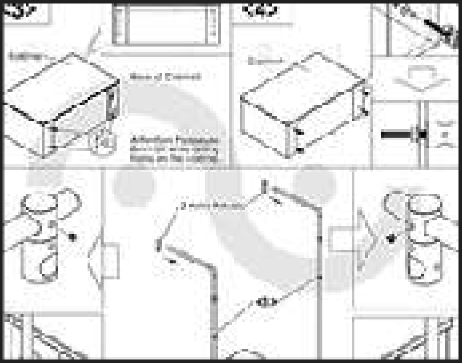

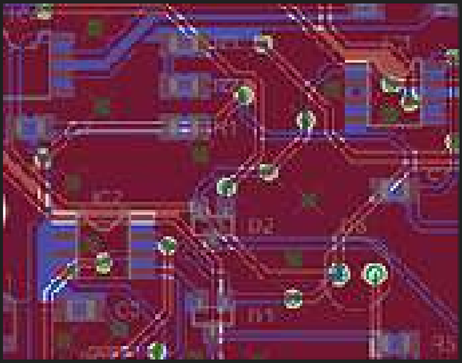

View Complete Profile →

Bēhance

Comments

Basic Info

Third Year BEng Project - Phased Patch Antenna Array

Credits

Giedrius Rakauskas
Vilnius, Lithuania

+ Follow

Tags

- patch antenna
- antenna array
- square patch antenna
- cst mws
- cst microwave studio

Tools Used

CST Microwave Studio, FR-4

Report

Behance

About Behance & Careers
Adobe Portfolio
Blog
Behance API
Portfolio Review Week
Creative Career Tips
Apps

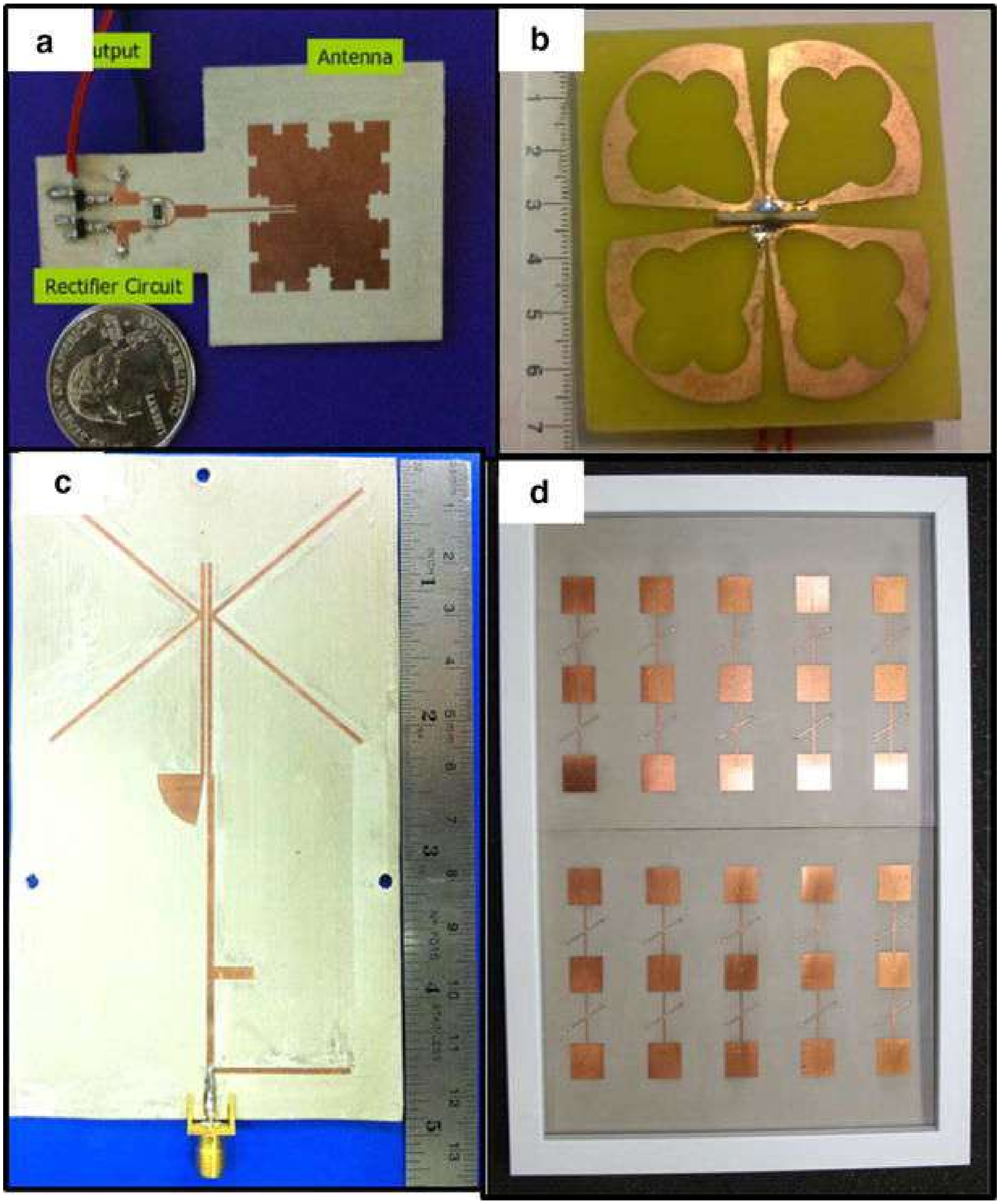

Fig. 4

Types of antennas **a** 2.45 GHz patch antenna with rectifier [38]. ©2010 IEEE. All rights reserved; **b** Planar dual-polarized antenna [24]. ©2015 IEEE. All rights reserved; **c** Microstrip antenna [37]. ©2012 IEEE. All rights reserved; **d** Array of stacked differential patch antenna [33]. ©2015 IEEE. All rights reserved

The plate antennas are popular and have many applications [27, 34, 38]; on-chip antennas are preferred for small and compact applications. Recently, many publications addressed wide-band and multi-band antennas. It has been proven that narrow-band antennas offer high energy conversion efficiencies but can only retrieve a limited amount of energy. On the other hand, wide-band or multi-band frequency antennas can retrieve more RF energy in space. However, the tradeoffs are low overall efficiency and large aperture. In [32], antennas with a resonance frequency of 4.9 and 5.9 GHz were designed with PCEs of 65.2 and 64.8%, respectively. Further work by Lu et al. [26] on polarization antennas supports the assertion that expanding the bandwidth of an antenna leads to increasing the amount of power harvested. In this work, the demonstration of broadband polarization antennas with three separate modes allows the antenna to operate in a wider range of frequencies. One common mechanism in the aforementioned works is the control of the antenna configuration by switching the diodes on and off, thus altering its resonant frequencies. However, since it uses separate modes for different frequencies, this antenna is not able to simultaneously resonate at two frequencies. On the other hand, the antenna presented in [39] is capable of operating at 2.45 and 5.8 GHz, simultaneously, providing 2.6 V output with a PCE of 65% and power density of 10 mW/cm^2.

The main purpose of aligning antennas in arrays is to enhance the antenna gain and obtain high voltage/current. Array antennas are preferred over large aperture antennas because they do not require large breakdown voltage diodes to operate. Antenna arrays can be connected before or after rectification. The first configuration enhances the retrieved power at the main beam while the second configuration expands the ability to retrieve power from various angles away from the main beam [40]. In case the RF waves are combined before rectification, the rectifier requires a large breakdown diode. If RF waves are combined after rectification, combining DC current becomes problematic. Antenna arrays can be connected in series or parallel to obtain high voltage or large current. Nonetheless, expanding the arrays yields better output but this might cause a deduction in conversion efficiency [33].

For demonstrating antenna arrays, Sun et al. [35] invented a T-junction to connect four quasi-Yagi antennas together. The advancement in this work was that the T-junction was flexible in changing from 1 × 4 array to 2 × 2 array topologies. Consequently, the system was able to operate at an ambient power level as low as 455 μW/cm^2 while obtaining 40% PCE.

Impedance matching network

In low-power consumption electrical systems, power leakage during transmission may lead to energy insufficiency. In these circumstances, adding an impedance matching network (IMN) ensures that the maximum power transfers between the RF source and load. For WPH applications, the receiving antenna is considered as the source while the rectifier/voltage multiplier is considered as the load. It is acknowledged that in DC, power transfer is optimum when the resistances of the source and load are indistinguishable. In an RF circuit, the impedance is referred to instead of resistance. An impedance mismatch between the source and load creates reflected power flow in the circuit that lowers the efficiency of the system. As its name indicates, the IMN ensures that the impedance of the source and load are identical by adding reactive components in between.

There are three basics matching configurations i.e. L, T and π matching networks (Fig. 5). The L matching is commonly used since it typically has two components, which simplifies the designing and controlling process. Additionally, the L matching networks do not alter the quality factor (Q) of the circuit.

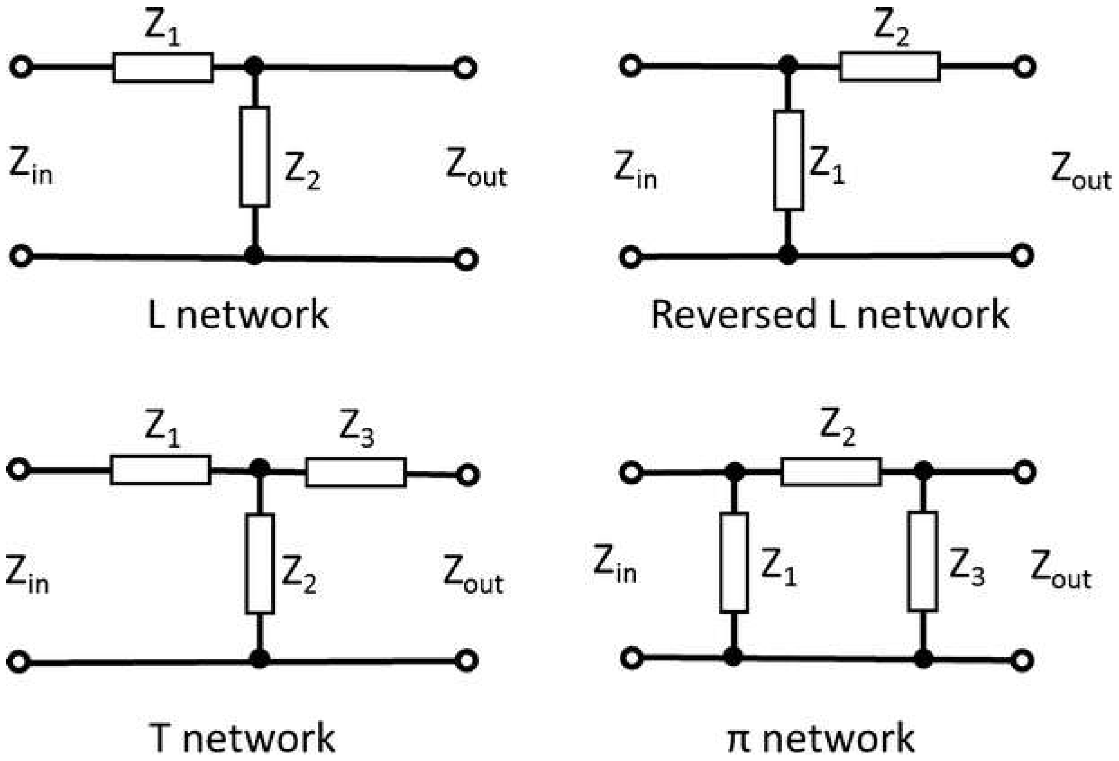

Fig. 5

Configuration of common impedance matching networks

The T and π matching configurations are more complex than the L network. Furthermore, organizing the T and π configurations into multiple stages will retain the final matching results but will change the Q factor. This strategy is useful in improving voltage boost.

There are tradeoffs between the attributes of an IMN, which include frequency, bandwidth, adjustability, and complexity. For instant, in [41], the suboptimal impedance matching and multiport ladder matching methods were introduced to enhance the matching performance and harvested power of antennas, respectively. However, the tradeoff was that implementation of these configurations required more components than traditional matching networks, thus, escalating the circuit's complexity.

Etor et al. [42] designed an IMN for THz frequencies applications using transmission lines and self-designed metal–insulator-metal diodes instead of lumped components. Moreover, fixed IMN and tunable IMN [43, 44, 45] were introduced as a technique for better matching with wide-band and multi-band antennas.

Rectifier/voltage multiplier

RF energy extracted from free space usually possesses low power density since the electric field power density decreases at the rate of $1/d^2$, where d is the distance from the RF source [46]. Therefore, a power amplifier circuit is required that yields enough DC energy from the electromagnetic waves to drive the loads. This gives rise to two possibilities, if the power consumption of the load is lower than the average power harvesting, the electronic devices at the load may work continuously; otherwise, if the load consumes more energy than the power harvesting circuit can generate, the devices cannot work continuously [47].

Rectifying is the most popular application of diodes, which refers to the conversion of AC current to DC current. In terms of power harvesting application, the RF signal retrieved in the antenna has a sinusoidal waveform. The signal after transformation through IMN would be rectified and boosted to meet the power requirements of the application.

The most fundamental topology of the rectifier is the half-wave rectifier that comprises of a single diode D1 (Fig. 6a). When AC voltage transfers through D1, only the positive cycle remains and the negative cycle is cutoff; thus, it diminishes half of the AC power. Moreover, the output V_{out} is discontinuous since the negative cycle is cutoff. Despite its simplicity, a half-wave rectifier is usually inadequate for common applications. Hence, a full-wave rectifier is more preferable. The circuit design of the full-wave rectifier is shown in Fig. 6b. During the first negative cycle of AC input, diode D1 is conductive and capacitor C1 is charged to the corresponding energy level of V_{peak} of the input. Then, at the next positive cycle, diode D1 is blocked, diode D2 is conductive so that capacitor C2 is also charged. In consequence, the output V_{out} would see two capacitors in series (each one is storing a voltage of $V_{peak)}$. Thus, V_{out} is twice V_{peak}. Therefore, this topology is more stable and efficient than the half-wave rectifier. There is also a bridge rectifier that rectifies both positive and negative cycles of the AC input but retains $V_{out} = V_{peak}$ by alternatively blocking pairs of diodes D1, D4 and D2, D3 (Fig. 6c).

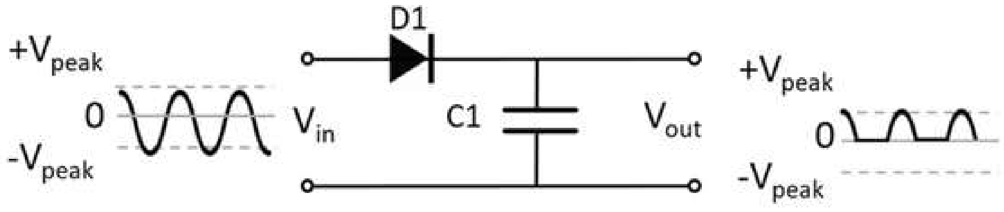

a Half-wave

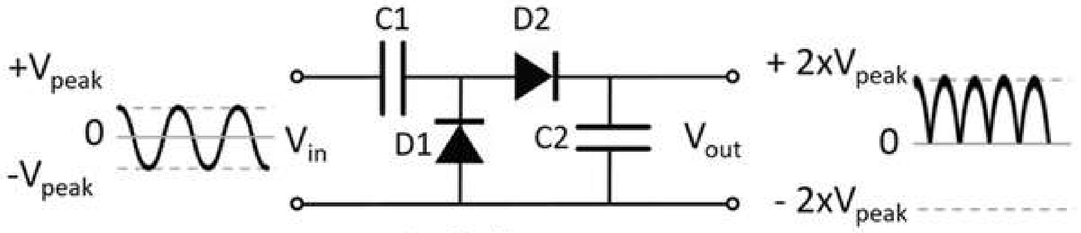

b Full-wave

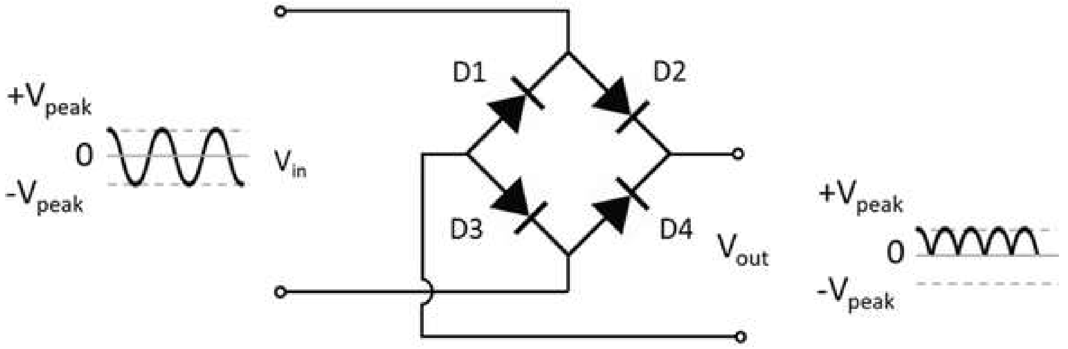

c Bridge

Fig. 6

Some common topologies of a rectifier

Voltage multiplier is a special type of rectifier circuit that converts and boosts AC input to DC output. In some case where the rectified power is inadequate for the application, there is a need for boosting the output DC by stacking single rectifiers into series, forming the voltage multiplier [48]. Several configurations of the voltage multiplier are shown in Fig. 7. The most fundamental configuration is the Cockcroft–Walton voltage multiplier (Fig. 7a). This circuit's operational principle is similar to the full-wave rectifier (Fig. 6b) but has more stages for higher voltage gain. The Dickson multiplier in Fig. 7b is a modification of Cockcroft–Walton's configuration with stage capacitors being shunted to reduce parasitic effects. Thus, the Dickson multiplier is preferable for small voltage applications. However, it is challenging to obtain high PCE due to the high threshold voltage among diodes creating leakage current, thus reducing the overall efficiency. Additionally, for high resistance loads, output voltage drops drastically leading to low current supply to the load. A summary of recent works related to voltage multiplier is shown in Table 3.

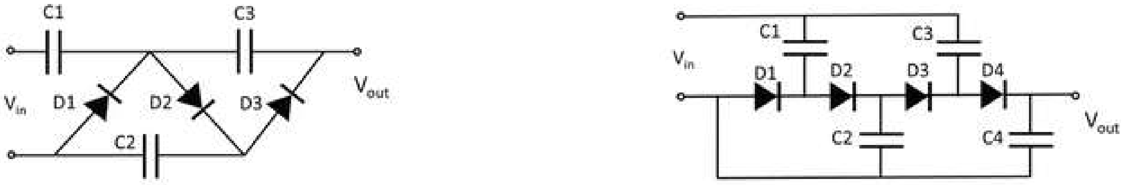
C1
C3
D1
D2
D3
V_in
V_out
C2
C1
C3
V_in
D1
D2
D3
D4
V_out
C2
C4

a Cockcroft-Walton

b Dickson

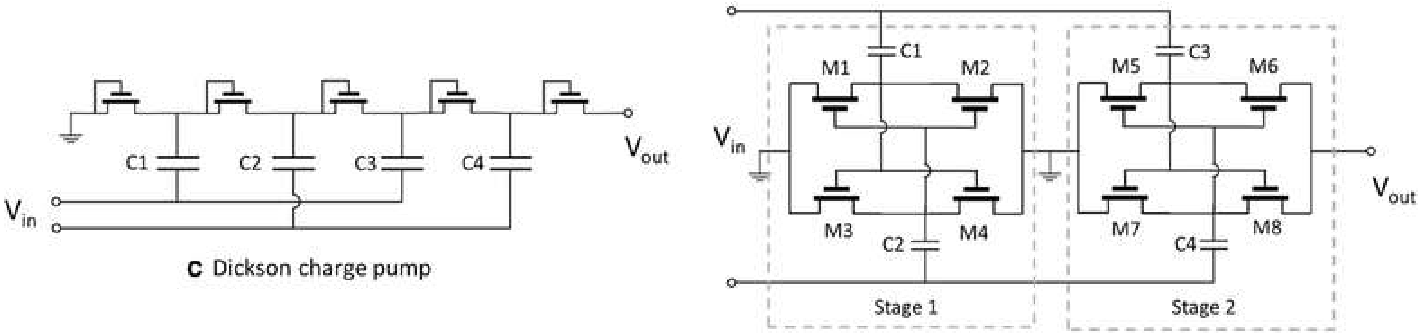
C1
C2
C3
C4
V_out
V_in
C1
C3
M1
M2
M5
M6
V_in
M3
M4
M7
M8
C2
C4
V_out
Stage 1
Stage 2

c Dickson charge pump

d Differential drive voltage multiplier

Fig. 7

Common voltage multiplier configurations: **a** Three stages Cockcroft–Walton voltage multiplier, **b** Four stages Dickson voltage multiplier, **c** Four stage Dickson voltage multiplier using CMOS technology, **d** Two stages voltage multiplier comprised of differential drive unit

Table 3

Published work comparison regarding voltage multiplier

Ref.	Rectifier topology	No. of stage	Freq. (MHz)	Tech.	Range (dBm)	Maximum PCE	Size
[51]	Differential	2	433	0.18 µm CMOS	0–20	74% @ 2 dBm	–
[52]	PMOS transistors	7	900	40 nm CMOS	–	44% @ > 390 mV	0.04 mm^2
[53]	Half-wave	4	900	0.18 µm CMOS	–	37.42% @ 390 mV	–
[54]	Comparator-based/active-diode	3	13.56	0.18 µm CMOS	8–15	67.9% @ 12.8 dBm	–
[55]	Dickson	3	13.56	250 nm CMOS	–	72%	0.13 mm^2
[56]	–	–	915	0.13 µm CMOS	−21.6	22.6% @ −16.8 dBm	0.186 mm^2

Prototyping Methods

There's more on this in Scherz, *Practical Electronics for Inventors.*

Solderless Breadboard or **Plugboard**

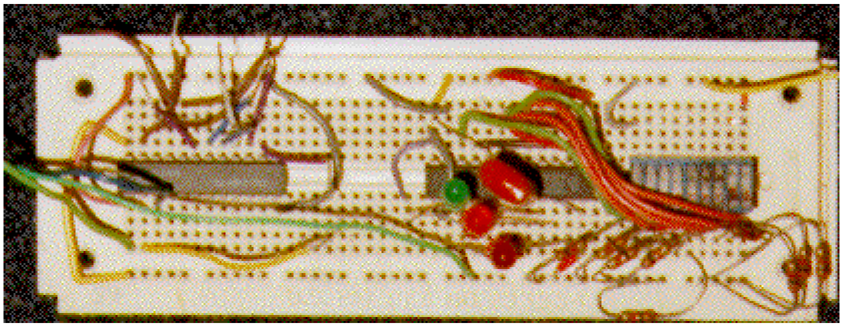

This is what we've been using in lab.

Advantages

- Very fast to build and make changes.
- Works well with DIP ICs

Disadvantages

- Limited reliability—lab testing only, for limited-size circuits.
- High capacitance between adjacent rows (~ 10 pF).
- Only for small-lead-size components (.032" or 0.82 mm max—a 1 A diode lead is just barely too big to meet the spec.)

Springboard

Used in ENGS 22

Advantages

- Fast to build and make changes.
- Accomodates large wire sizes.
- Secure, reliable connections.
- Can handle higher current/power.

Disadvantages

- Can't accommodate small lead spacings such as on ICs.
- Although it is reliable enough for long-term use, it's expensive for that purpose.

Wirewrap

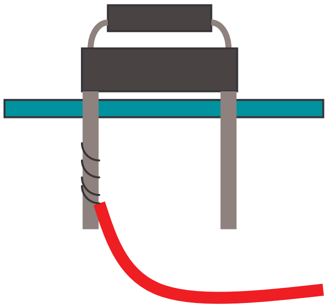

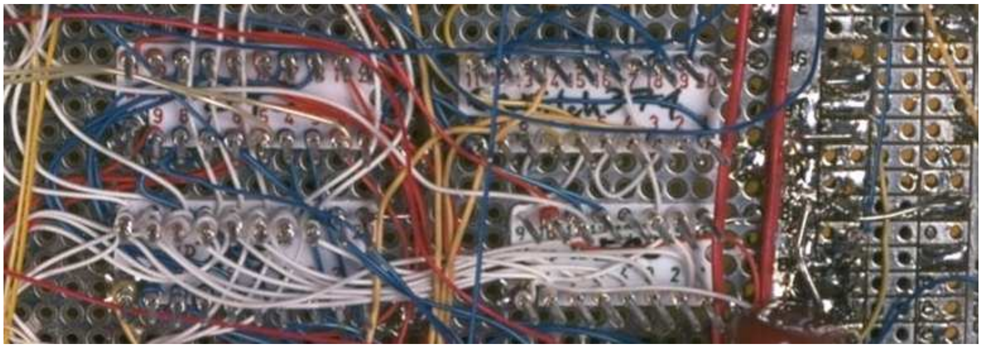

Connections made by fine wire wrapped tightly around square pins of special IC sockets. This is a great way to make a permanent version of a digital circuit almost as quickly as using a solderless breadboard. Used in ENGS 31.

Advantages

- Fast to build and make changes.
- Can make a complex circuit compact.
- Reliable.
- Can be used permanently.
- Inexpensive.

Disadvantages

- Low current—limited to under a few hundred mA.
- High inductance/resistance—on the order of 0.1 ohm, 0.25 µH per foot.
- Only makes reliable connections to square leads (as are on wire-wrap IC sockets). Discrete components (e.g., resistors, capacitors, transistors) need to be soldered to a "header" that goes in an IC socket or to individual wire-wrap pins.

Perfboard and Solder.

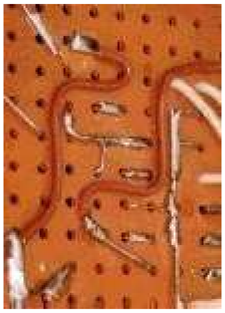

Using the same perfboard as used in wire-wrap work, it is possible to simply twist and solder leads, and run wires where needed.

Advantages

- Reliable, if done well.
- Can be used permanently.
- Inexpensive.
- Can handle any size components
- Convenient for working with discrete components.

Disadvantages

- Slow, requires skill to do well.
- Works for ICs, but not very easily.

Variations:

- Perfboard with individual copper pads on each hole so that the solder will better hold things in place. (Photo below)
- Perfboard with a perforated ground plane. Provides the shielding and grounding benefits of a ground plane, and makes ground connections easy, but requires care to avoid shorts. Some have the ground plane etched from around the holes; others require you to cut the copper away from the holes with a special tool. This is also possible with wire-wrap, as in the photo in that section.

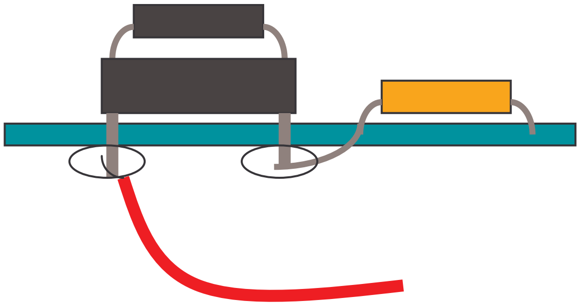

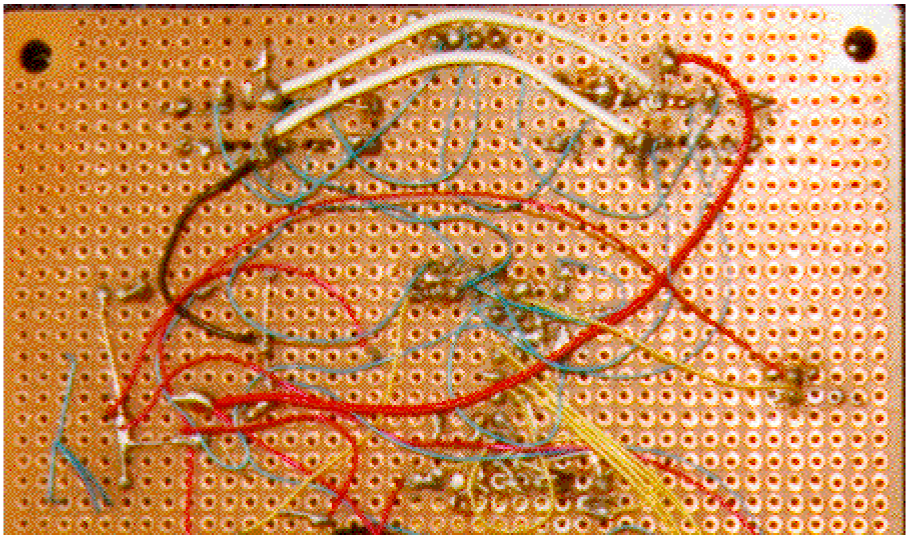

Generic Printed-Circuit Board—multiple hole-per-pad.

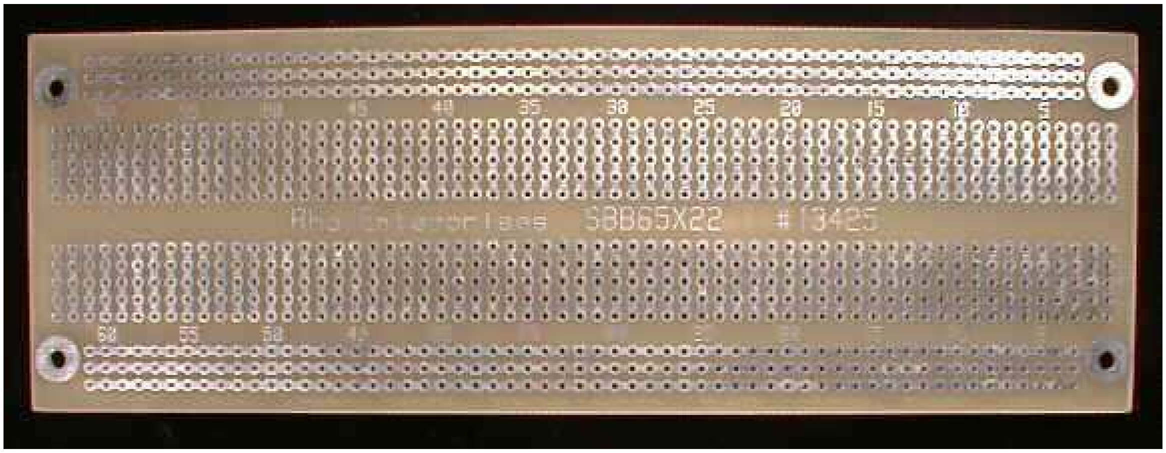

A printed circuit board with a pattern of holes and connections similar to a solderless breadboard.

Advantages

- Easier than using plain perfboard, especially for ICs.
- Reliable.
- Can be used permanently.
- Available with ground planes if needed.

Disadvantages

- Usually not as compact a final circuit as some alternatives, because you are constrained by layout.
- Pads bigger than needed can add capacitance, but not much.
- Can be expensive, especially the “vectorboard” brand.
- Sometimes awkward for odd sized or large components.

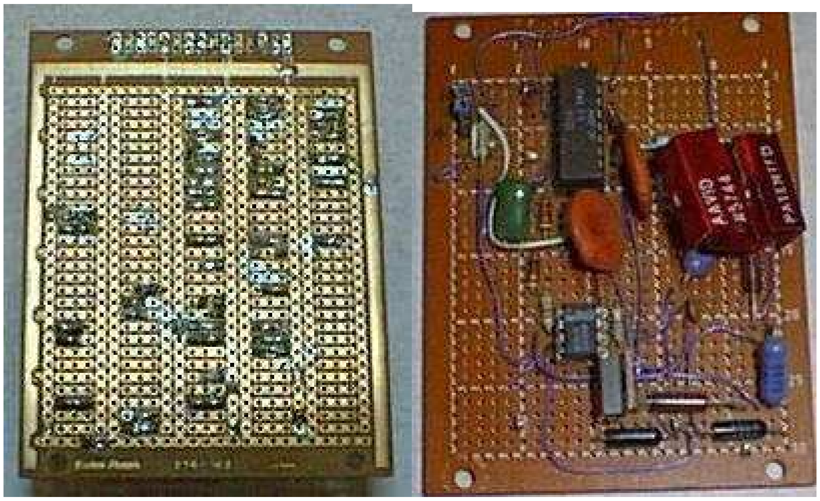

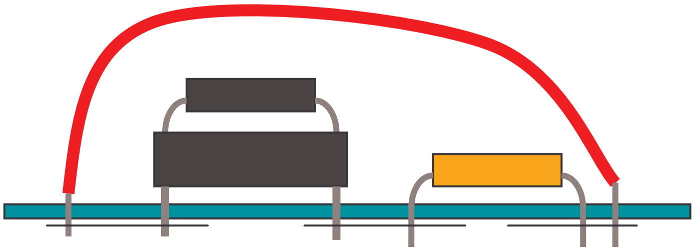

"Dead Bug," or "Ugly-board"

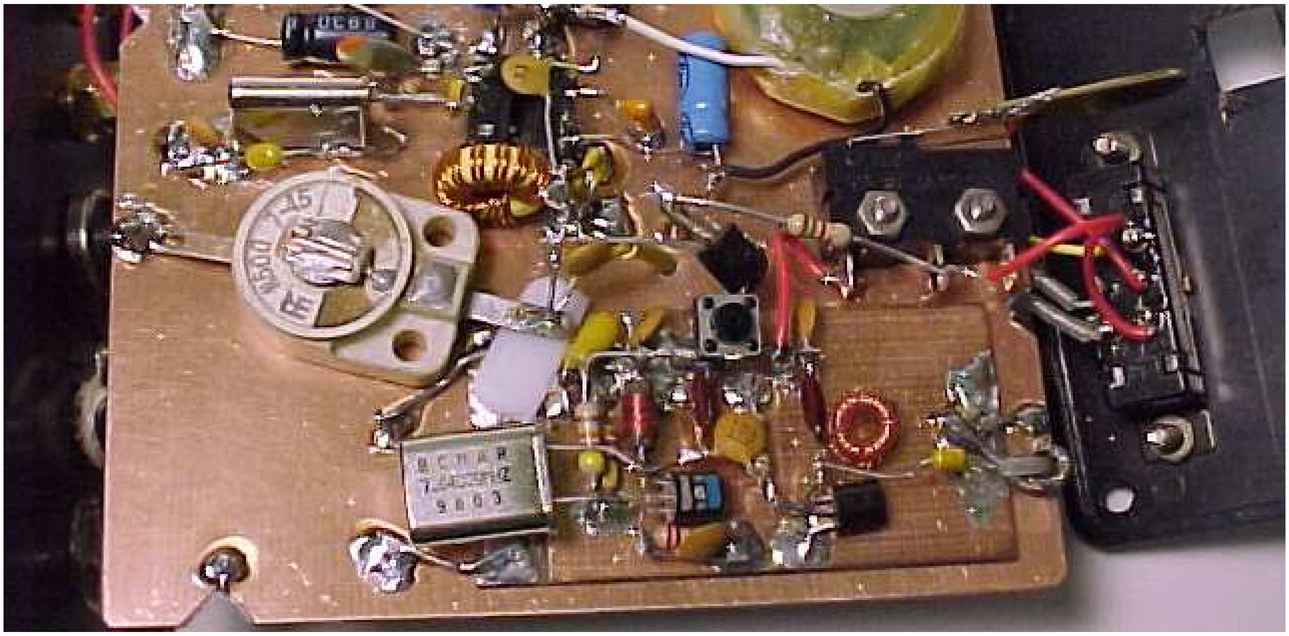

Start with a plain copper-clad board. Glue ICs down with the leads sticking up in the air. Then solder to them

Advantages

- Provides an excellent ground plane. Can be a high performance way to build sensitive and/or high-frequency analog circuits
- Can make a complex circuit compact.
- Reliable.
- Can be used permanently.
- Inexpensive.

Disadvantages

- Requires high soldering skill.
- Takes a long time to build.
- Mechanical support for components is marginal; can add glue ("RTV") after debugging.
- Only makes reliable connections to square leads (as are on wire-wrap IC sockets). Other components (e.g., resistors, capacitors, transistors) need to be soldered to a "header" that goes in an IC socket or to individual wire-wrap pins.

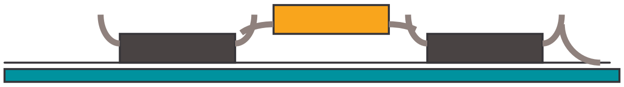

Variations:

- Manually cut the board with a dremel tool to isolate sections for purposes other than ground plane (use the back for ground plane). See photo below.
- Glue on little rectangles of cut-out board to add power busses, etc.

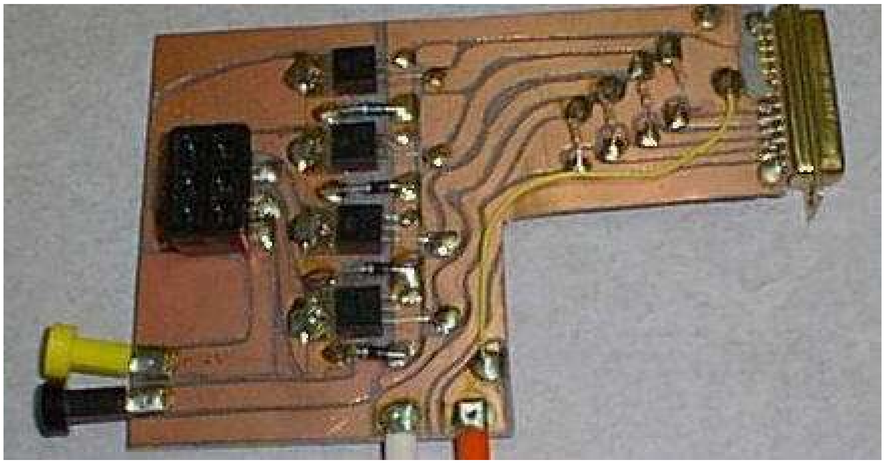

Printed circuit board (PCB)

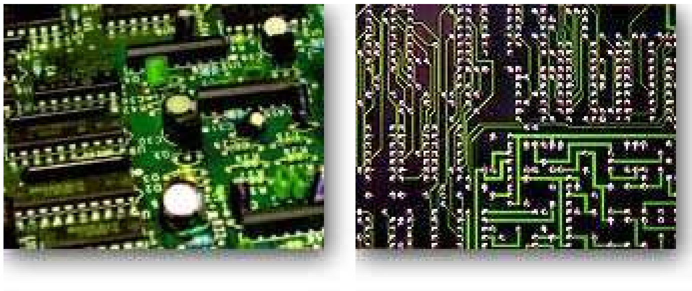

This is what is used virtually universally in production of electronics.

Advantages

- Easy to build in production.
- Repeatable, controllable stray L, C.
- Can handle virtually any component, power level.
- Highly reliable.
- Can make very compact.
- Design can be (somewhat) automated from a schematic you have entered.

Disadvantages

- Laying out the board and getting it fabricated takes time, although you can pay for fabrication in a few days if you can afford it.
- Expensive, on the order of hundreds of dollars for one, but with almost no increase in cost to make many.
- Hard to make changes, but making changes may be easier than building another type of prototype.

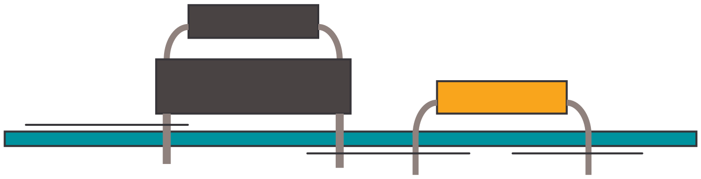

Notes on surface-mount components

Most modern production designs now use surface-mount components instead of “through hole” components. The circuits can be more compact, and the board layout is easier because different things can be done on each side (and in additional layers between sides) without through-holes interfering. But prototyping gets much more difficult!

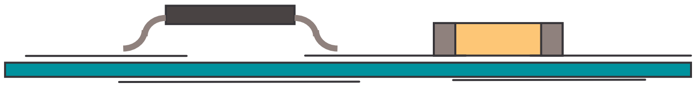

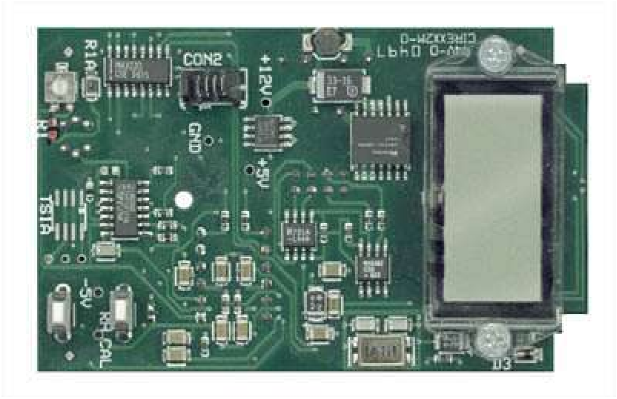

Prototyping options for surface mount

- Simulate, and then lay out a PCB. Don’t ever make a breadboard.
- Order DIP ICs and leaded passives for the prototype, and then switch to surface mount for production.
- Get adaptors that have pads to solder surface-mount ICs to, and then standard-spacing (0.1”) pins in DIP or SIP layout. Digikey carries “surfboards” made by Capital Advanced Technologies, and adaptors made by Aries Electronics. In addition to DIP and SIP adaptors, there are boards with solder pads for connecting larger wires; these work well for prototypes built in “dead bug” style.

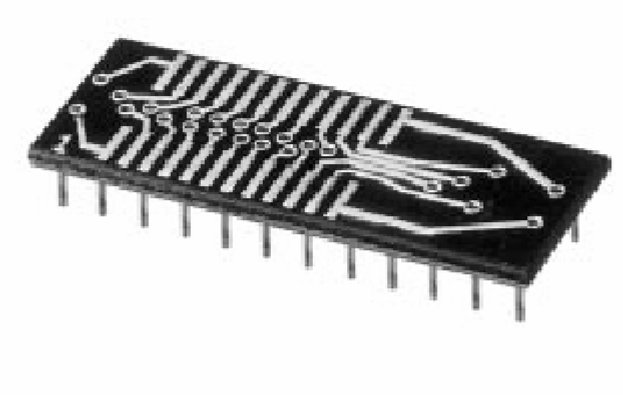

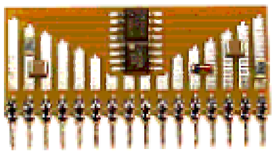

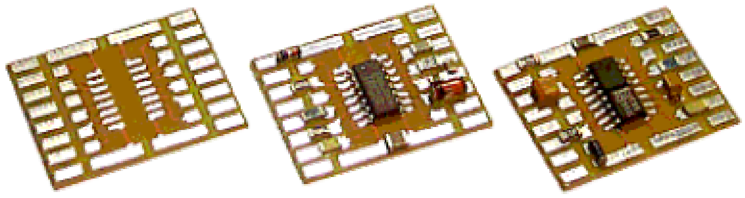

Crystal receiver set 11

Back to the index

This receiver has a single tuned circuit, the Q factor of the tuned circuit is quite low.
The receiver can be used well for reception of local stations, for reception of distant stations it is less suitable.

I tried to give this receiver a nice "old-fashioned" look, the reception performance was in this design of less importance.
This receiver is for sale, I made a series production of these, and will build more on request.
This receiver is also for sale as do-it-yourself kit.
See the shop for more information.

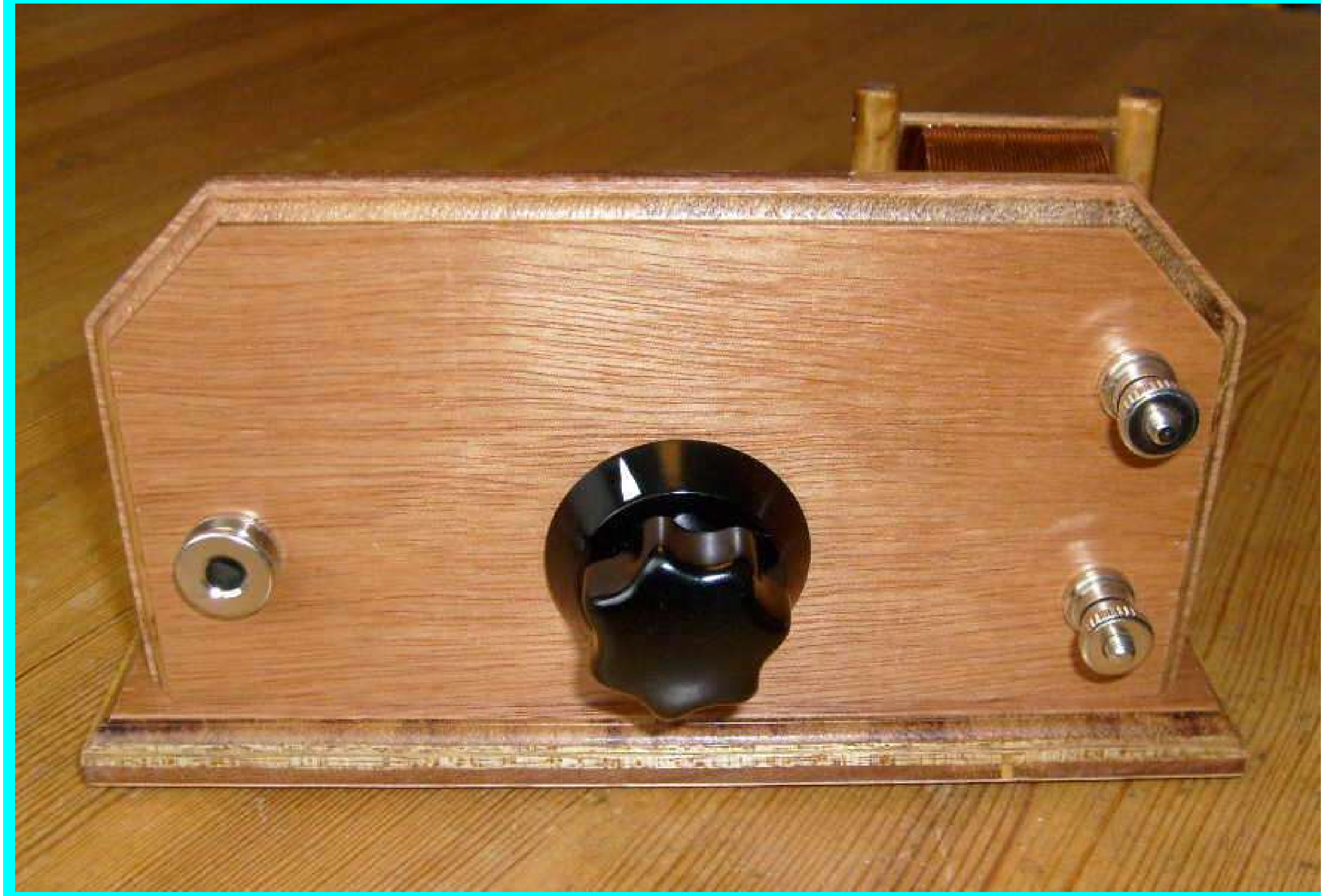

The front panel of the receiver.
On the left side the socket for the headphone.
In the middle the tuning knob.
On the right the connections for antenna (upper) and ground (lower).

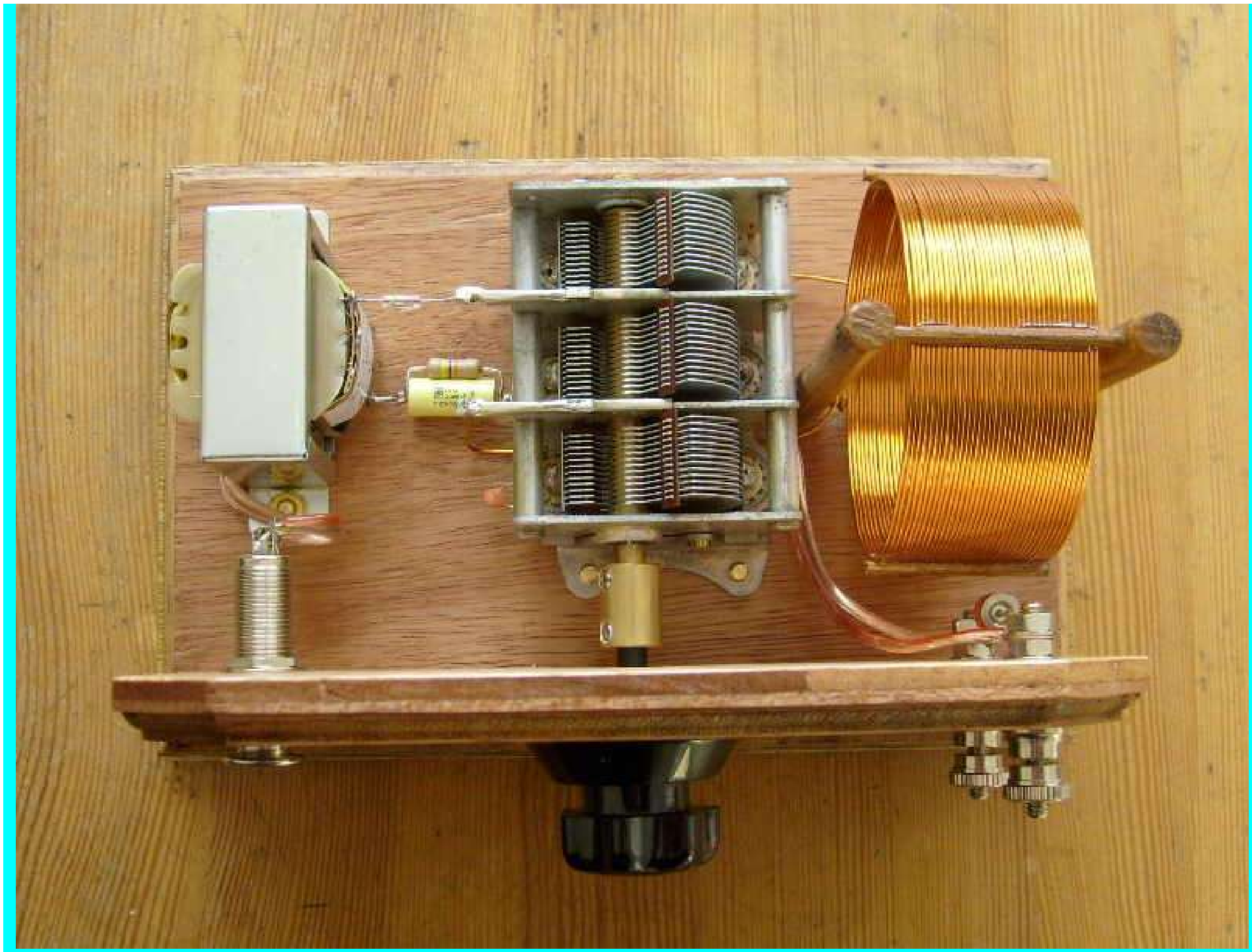

Top view of the receiver.

Circuit diagram of receiver set 11

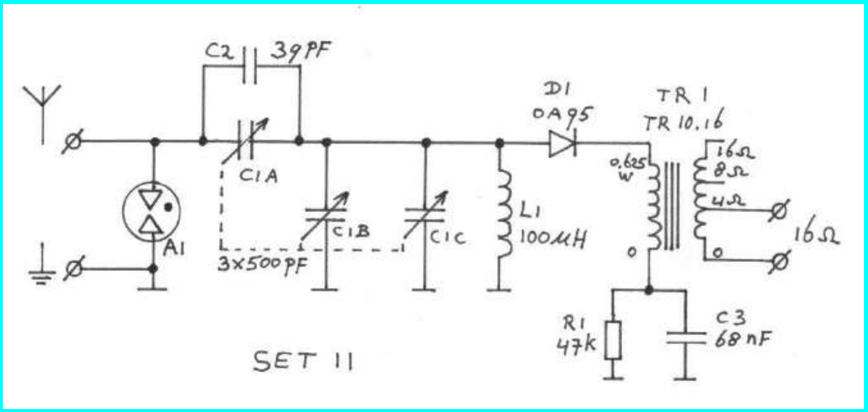

Circuit description.
The resonance circuit is formed by coil L1 and C1b and C1c (together 1000 pF).
Capacitor C1a and C2 provide the matching between antenna and tuned circuit.
The frame (rotor) of the tuning capacitor is carrying the RF signal, by this it is possible to tune simultaneously the tuned circuit (C1b and C1c) and the antenna matching (C1a).
Germanium diode D1 provides the signal detection.
Transformer TR1 is loaded with 16 Ω at its 4 Ω output, through this the input impedance increases from 16 kΩ to about 43 kΩ.
At the output a headphone of 2x 32 Ω can be connected, with the two speakers parallel, the impedance is 16 Ω.

Component A1 is a *gas discharge tube* (also called: *surge arrester*) with type number: N81-A90X.
The gas discharge tube protects the antenna input for too high voltages.

These high voltages can occur if the antenna picks up static charge from the air (especially occurs with long outdoor antennas from non insulated antenna wire).
As the voltage at the antenna is higher then 90 Volt, the gas discharge tube will start to conduct and short the high voltage to ground.
As soon as the charge has flown to ground, the conduction stops automatically.

Frequency range of the receiver.
Frequency range without antenna: 500 - 2500 kHz.
Frequency range with 10 meter antenna connected: 487 - 1860 kHz.
Both with and without antenna connected, the whole medium wave band can be tuned.

Q factor of the LC circuit (Without antenna and without diode connected):
600 kHz: Q= 83
900 kHz: Q= 81
1200 kHz: Q= 75
1500 kHz: Q= 65
The circuit Q is rather low, one reason for this is because the RF signal is in this design on the frame of the tuning capacitor.
Because the frame of the tuning capacitor is directly connected to the wooden bottom plate losses occur here.

You can find a complete building instruction op the following pages:

Step 1 Preparing the components

Step 2 Making the frame of the receiver

Step 3 Making the coil

Step 4 Placing the components

You will find here the part list of this receiver.

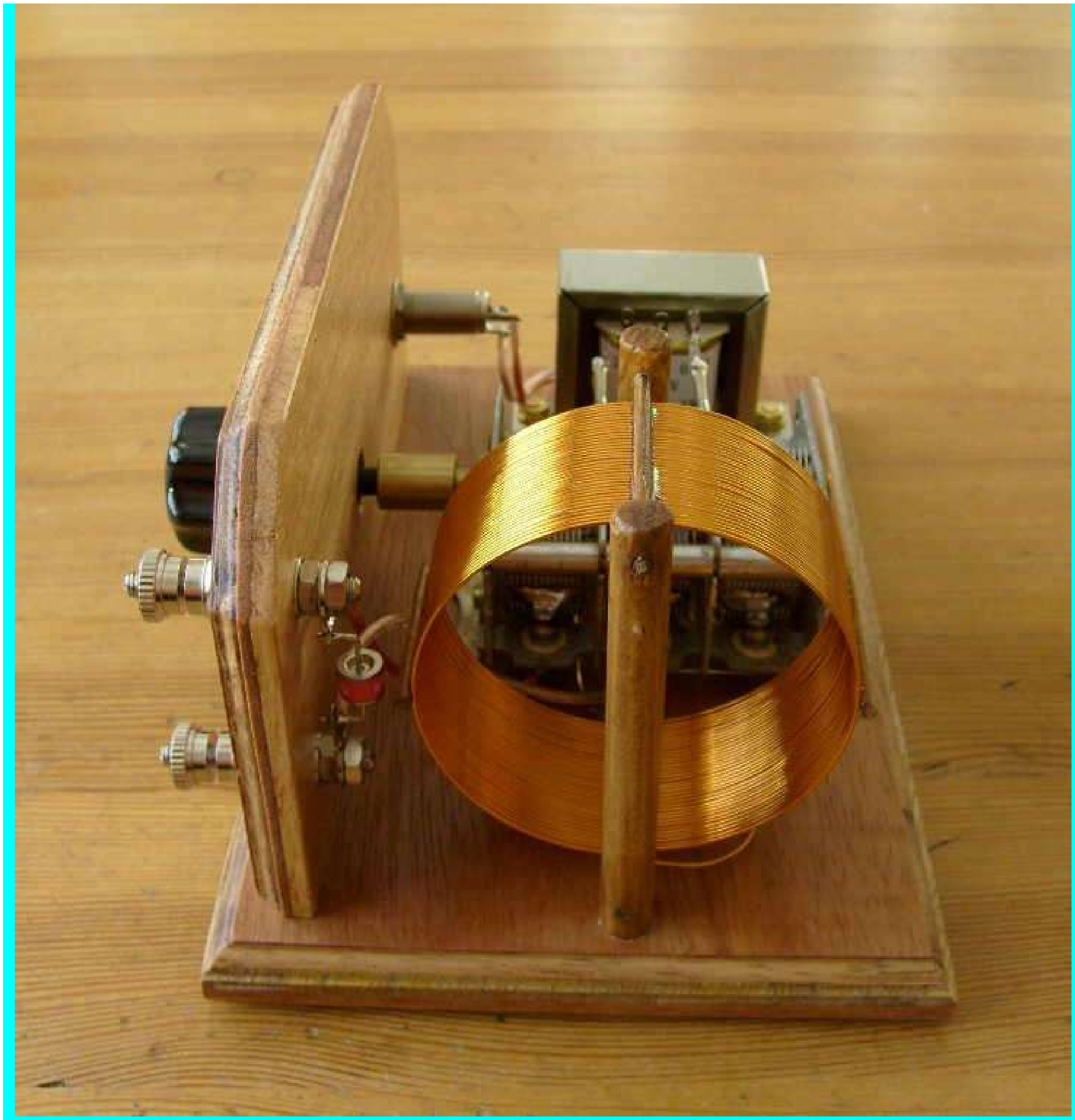

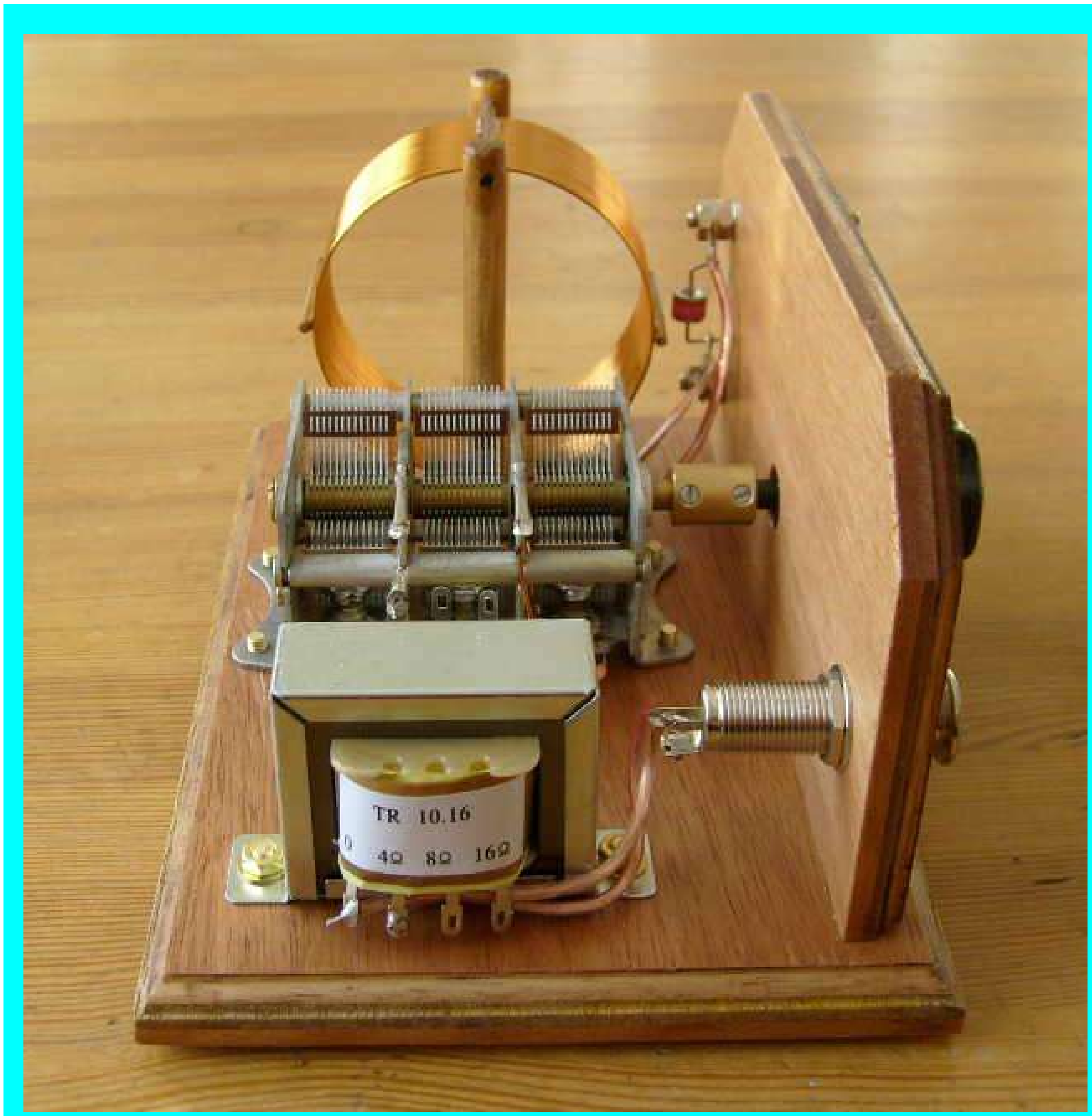
TR 10.16
Ω 4Ω 8Ω 16Ω

Back to the index

How to make a batteryless (crystal set) radio

By **Sagar Sapkota** - July 21, 2012

CRYSTAL SET

The simplest radio receiver, known as a Crystal Set, consists of nothing more than a coil, tuning capacitor, diode detector, and a pair of earphones. A typical circuit diagram for a Crystal Set Radio is given below where inductor or coil L1 is tuned by variable capacitor VC1 to the transmitter frequency. Diode D1 demodulates the signal, which is fed straight to the earphones. There is no amplification.

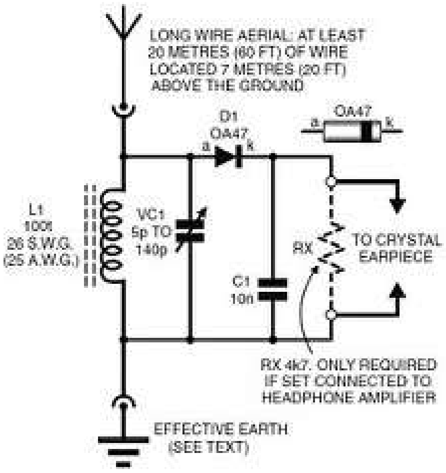
LONG WIRE AERIAL: AT LEAST
20 METRES (60 FT) OF WIRE
LOCATED 7 METRES (20 FT)
ABOVE THE GROUND
D1
OA47
a
k
OA47
a
k
L1
100t
26 S.W.G.
(25 A.W.G.)
VC1
5p TO
140p
RX
TO CRYSTAL
EARPIECE
C1
10n
RX 4k7. ONLY REQUIRED
IF SET CONNECTED TO
HEADPHONE AMPLIFIER
EFFECTIVE EARTH
(SEE TEXT)

CLICK HERE FOR BIGGER PHOTO

** You can use 1N34 Germanium diode in place of OA47*

A long (at least 20 metres), high (17 metres or more) aerial and a good earth (a buried biscuit tin or a metre of copper pipe driven into damp ground) are required in order to ensure audible headphone reception. The earphones originally used with these receivers had an impedance of around 4000 ohms and were very sensitive (and heavy and uncomfortable). They are no longer available, but a crystal earpiece, which relies on the piezoelectric effect, will give acceptable results.

Low impedance "Walkman" type earphones are NOT suitable.

Component details:

Resistor- RX-4.7k, 0.25 W- only required if set is connect to audio amplifier

Capacitors- C1- 10nF disc ceramic VC1- 5p to 140p, polythene dielectric variable capacitor.

Semiconductors: D1: OA47 or 1N34- Germanium Diode

Miscellaneous: L1- Ferrite Rod, 100mm(4 inch)x9mm/10mm dia., with coil.

Crystal earpiece and jack socket to suit; plastic control knob; plastic insulated flexible cable for aerial wire, download and earth connection, 30 meters minimum; buried biscuit tin or 1 meter of copper pipe for earth system; 50gm reel of 26SWG enamelled copper wire, for tuning coil; card and glue for coil former; multistrand connecting wire; crocodile clips or terminals for aerial and earth lead connection; solder, etc.

READ MY EXPERIENCES- HOW EASILY I MADE A BATTERYLESS RADIO

DRAWBACKS

Quite apart from the absence of amplification, two factors seriously limit the performance of crystal receivers. Germanium diodes become increasingly reluctant to conduct as the applied voltage falls below 0.2V, and this makes the receiver insensitive to weak signals. Silicon diodes have a threshold of around 0.6V, and are, therefore, unsuitable for circuits of this kind.

** Silicon diodes like 1N4001 are not suitable*

The earphone loading imposes heavy damping on the tuned circuit, hence, reduces its ability to separate signals. With such low selectivity insensitivity can be a blessing, and crystal sets are normally only capable of receiving a single, strong transmission on the long and medium wavebands. They will sometimes receive more than one if a shortwave coil is fitted.

The aerial and diode can be connected to tappings on the tuning coil in order to reduce damping, but the improvement in selectivity is usually at the expense of audio output. When valves cost a week's wages and had to be powered by large dry batteries and lead/acid accumulators, the construction of simple receivers of this kind could be justified. With high performance transistors now costing only a few pence or cents, crystal sets are now regarded as 'nostalgic pieces". Some readers may, however, wish to build one out of curiosity, or for the novelty of having a receiver that **does not require a power supply.**

CIRCUIT DETAILS

Ferrite loop aerial L1 and polythene dielectric variable capacitor VC1 form the tuned circuit. Point contact germanium diode D1 (1N34 or OA47) demodulates the signal; capacitor C1 bypasses residual r.f. (radio frequency) to earth and also exhibits a reservoir action, enabling the a.f. (audio frequency) output to approach its peak value. The recovered audio signal is fed directly to a crystal earpiece. Signal voltages introduced in the ferrite loop aerial by the radiated magnetic field are much too much to produce an output from the detector, and the component is used here simply as a tuning coil. The ferrite core does, however, reduce the number of turns required for the coil winding, thereby reducing its resistance and increasing its audio quality.

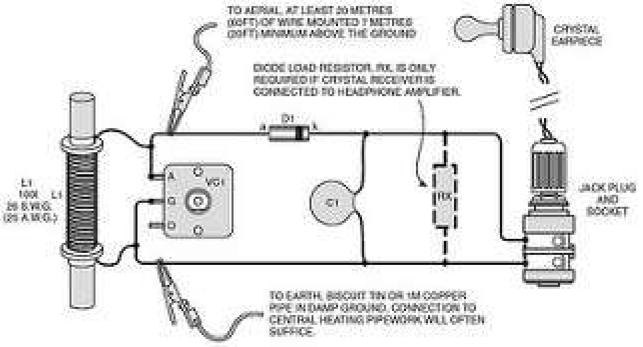
TO AERIAL, AT LEAST 20 METRES (66FT) OF WIRE MOUNTED 7 METRES (20FT) MINIMUM ABOVE THE GROUND
CRYSTAL EARPIECE
DIODE LOAD RESISTOR, RX, IS ONLY REQUIRED IF CRYSTAL RECEIVER IS CONNECTED TO HEADPHONE AMPLIFIER.
D1
A
K
L1 100t 26 S.W.G. (25 A.W.G.)
L1
A
VC1
G
O
C1
RX
JACK PLUG AND SOCKET
TO EARTH, BISCUIT TIN OR 1M COPPER PIPE IN DAMP GROUND. CONNECTION TO CENTRAL HEATING PIPEWORK WILL OFTEN SUFFICE.

GET BIGGER PICTURE

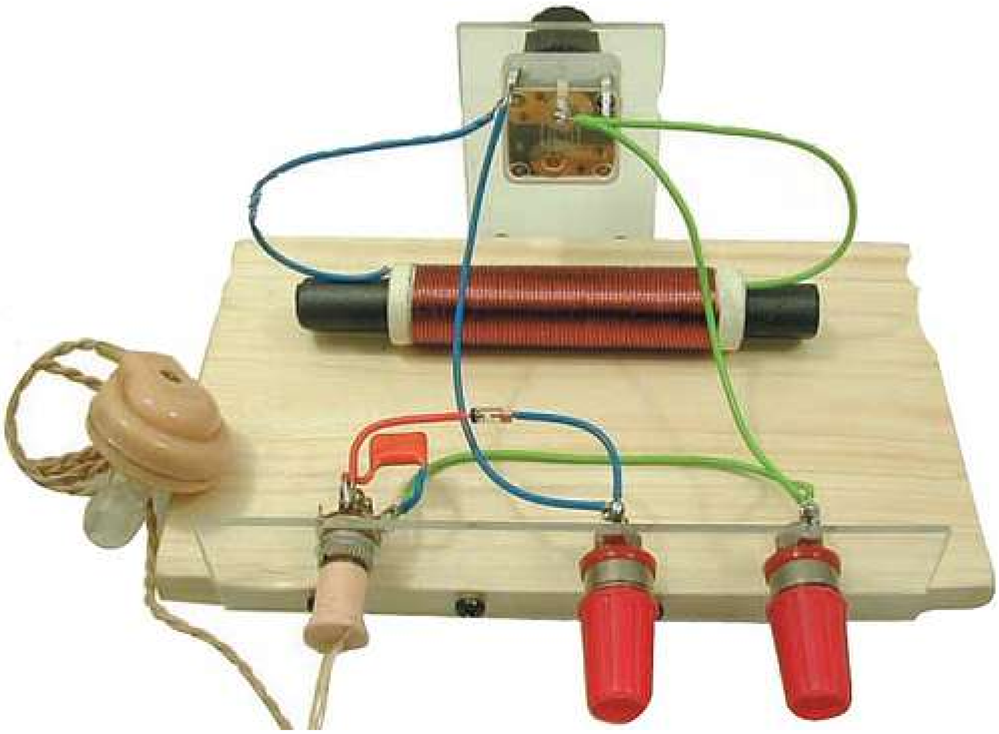

COIL DETAILS

Full construction and winding details for the ferrite tuning/ aerial coil L1 are shown in figure.

The coil is made from 26 SWG(25 AWG) enamelled copper wire, close wound on a cardboard former.

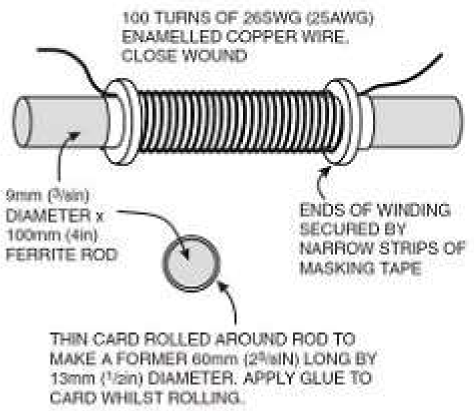
100 TURNS OF 26SWG (25AWG) ENAMELLED COPPER WIRE. CLOSE WOUND
9mm (3/8in) DIAMETER x 100mm (4in) FERRITE ROD
ENDS OF WINDING SECURED BY NARROW STRIPS OF MASKING TAPE
THIN CARD ROLLED AROUND ROD TO MAKE A FORMER 60mm (2 3/8IN) LONG BY 13mm (1/2in) DIAMETER. APPLY GLUE TO CARD WHILST ROLLING.

The r.f. bypass capacitor C1 can in practice, be omitted with no noticeable reduction in performance. However, if the set is to be connected to either the headphone amplifier or speaker amplifier decribed next month, the component together with diode, load resistor, RX, must be included.

AMPLIFICATION Audio frequency amplification, after the diode detector will permit the use of low impedance Walkman type earphones or even loud speaker operation. It will do nothing, however, to overcome the diode's insensitivity to weak signals. For this we must have radio frequency amplification of the signals picked up by the aerial before they reach the detector(The standard circuit for a transistor portable receiver has three stages of radio frequency amplification ahead of the diode).

For amplification, you can use a simple audio amplifier using LM386.

READ MY EXPERIENCES ON MAKING A BATTERYLESS RADIO

WATCH THESE VIDEOS:

- **JLCPCB: Prototype 10 PCBs for $2 (2-layer,100*100mm)**
- **China's Largest PCB Prototype Manufacturer, 290,000+ Customers & 8,000+ Online Orders Per Day**
- **PCB Pricing: https://jlcpcb.com/quote**

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/271597499

Fully printed flexible and disposable wireless cyclic voltammetry tag

Article *in* Scientific Reports · January 2015

DOI: 10.1038/srep08105 · Source: PubMed

CITATIONS
21

READS
121

11 authors, including:

Jinsoo Noh
Gumi Electronics and Information Technolog…
22 PUBLICATIONS **641** CITATIONS
SEE PROFILE

Myoungho Pyo
Sunchon National University
96 PUBLICATIONS **1,319** CITATIONS
SEE PROFILE

Ali Javey
University of California, Berkeley
296 PUBLICATIONS **22,530** CITATIONS
SEE PROFILE

Gyoujin Cho
Sunchon National University
95 PUBLICATIONS **2,833** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project

Electrochemistry View project

All content following this page was uploaded by Gyoujin Cho on 02 February 2015.

The user has requested enhancement of the downloaded file.

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
ENVIRONMENTAL MONITORING
ELECTRICAL AND ELECTRONIC ENGINEERING

Received
22 October 2014

Accepted
6 January 2015

Published
29 January 2015

Correspondence and requests for materials should be addressed to G.C. (gcho@sunchon.ac.kr)

Fully printed flexible and disposable wireless cyclic voltammetry tag

Younsu Jung[1], Hyejin Park[1], Jin-Ah Park[1], Jinsoo Noh[1], Yunchang Choi[1], Minhoon Jung[2], Kyunghwan Jung[2], Myungho Pyo[1], Kevin Chen[3], Ali Javey[3] & Gyoujin Cho[1]

[1]Department of Printed Electronics Engineering, Sunchon National University, Maegok, Sunchon, Jeonnam 540-742 Korea, [2]Research Institute of Printed Electronics, PARU Co. Seomyeon, Sunchon, Jeonnam 540-813 Korea, [3]Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

A disposable cyclic voltammetry (CV) tag is printed on a plastic film by integrating wireless power transmitter, polarized triangle wave generator, electrochemical cell and signage through a scalable gravure printing method. By proximity of 13.56 MHz RF reader, the printed CV tag generates 320 mHz of triangular sweep wave from +500 mV to −500 mV which enable to scan a printed electrochemical cell in the CV tag. By simply dropping any specimen solution on the electrochemical cell in the CV tag, the presence of solutes in the solution can be detected and shown on the signage of the CV tag in five sec. 10 mM of N,N,N′,N′-tetramethyl-p-phenylenediamine (TMPD) was used as a standard solute to prove the working concept of fully printed disposable wireless CV tag. Within five seconds, we can wirelessly diagnose the presence of TMPD in the solution using the CV tag in the proximity of the 13.56 MHz RF reader. This fully printed and wirelessly operated flexible CV tag is the first of its kind and marks the path for the utilization of inexpensive and disposable wireless electrochemical sensor systems for initial diagnose hazardous chemicals and biological molecules to improve public hygiene and health.

Cyclic voltammetry (CV) has been used as a powerful tool for the study of electrochemical redox reactions between the electrodes and the solutions of all kind of solutes such as metals[1], organic molecules[2], proteins[3], bacteria[4], viruses[5], and DNA[6]. Because of the high sensitivity of the electron transfer redox reactions, CV could become a very promising ubiquitous electrochemical sensor protocol if the cost and size of the CV system were significantly reduced to commercially viable single-use disposable units for checking traces of hazardous materials such as lead[7], mercury[8], arsenic[9], e-*coli*[10], and pesticides[11] in water or to diagnose the level of glucose[12], cholesterol[13] and specific enzymes[14] in blood. This single-use disposable CV would dramatically reduce the cost of maintaining the public health system[15]. Therefore, inexpensive and disposable CV measurement system that can be operated wirelessly using a smartphone or RF (radio frequency) reader without complicate operation processes is in high demand for the realization of an ubiquitous sensor network system (Figure 1a). They would mainly be used as ubiquitous diagnostic and testing tools for detecting and monitoring the level of target specimens. However, there is no technology that is advanced enough yet to build a wireless, inexpensive and disposable CV system. In this paper, as a form of RF-tag, an extremely inexpensive, disposable and fully printed CV system is demonstrated for the first time by mimicking and combining basic CV concepts and wireless power transmission technologies of RF devices. To realize the fully printed CV tag, a key issue of wirelessly generating triangular waveform (±500 mV) that can scan the electrochemical cell at a low frequency (<1 Hz) to set up a redox reaction needs to be addressed through a minimum number of printed thin film transistors (TFTs). Fully printed 13.56 MHz rectenna and a ring oscillator with large trap charges in the channels of printed TFTs were respectively utilized to wirelessly generate triangular waveform by using only 10 printed TFTs.

Results

Based on two key units of the wireless power transmitter (Figure 1b-①) and the triangular wave generator (Figure 1b-②), the circuit layout of the fully printed wireless CV tag was designed by using a minimum number of printed thin film transistors (TFTs) to alleviate the issue of V_{th} shift and shown in Figure 1b as a platform of disposable CV system. The circuit of CV tag was fabricated using all scalable printing methods, a roll-to-roll gravure, a roll-to-plate gravure, a drop casting, and a screen printer (Figure S1 in Supplementary Information).

To provide a polarized DC voltage from the coupled 13.56 MHz AC signal of the reader, we modified our previously reported R2R gravure printed rectenna[16,17], as shown in the layout in Figure 2a and the resulting

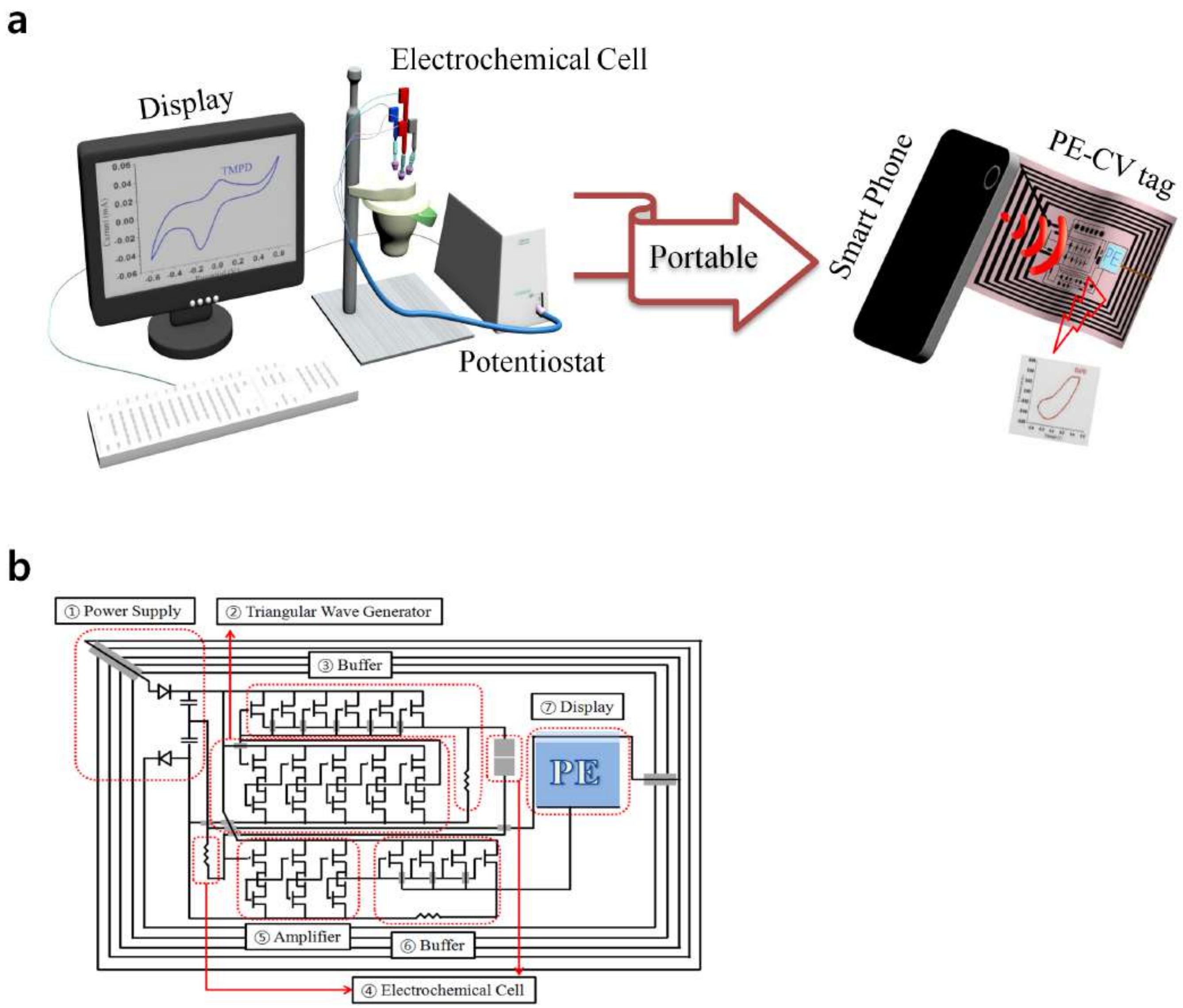

Figure 1 | (a) Informative illustration of typical CV system and disposable printed CV tag. (b) Schematic circuit diagram of the gravure printed wireless cyclic voltammetry (CV) tags. The circuit was designed to couple AC power from a 13.56 MHz reader and then convert the coupled AC to polarized DC ±10 V (① in Figure 1b). Polarized DC will operate the printed ring oscillator to generate a triangular voltage waveform (② in Figure 1b). The generated waveform will pass through the buffer to meet the impedance difference from the electrochemical cell (③ in Figure 1b). The electrochemical cell will run the redox reaction with a single drop of specimen solution (④ in Figure 1b) by the voltage triangular waveform. The output current of the electrochemical redox reaction will be amplified *via* an amplifier circuit (⑤ in Figure b) and the signal will pass through the buffer to meet the impedance difference from the printed signage (⑥ in Figure b). The signage will indicate the concentration level of specimen in the solution (⑦ in Figure b). It will indicate whether the concentration is above or below a pre-determined value.

printed diodes and capacitors are shown in Figure 2b. The printed diodes showed a rectifying ratio of 10^3, and capacitances of two printed capacitors were all about 8 nF/cm^2. The polarized DC voltages (Figure 2c) were measured by placing the printed rectenna on the RF (13.56 MHz) reader with a distance of 2 cm. We adapted a center tap transformer, consisted by divided antenna, 2 diodes and 2 capacitors, to provide the + and − DC voltage for printed triangle wave generator. Using a load of 1 MΩ, voltages of +9.4 V and −10.8 V DC were attained for the optimized printed antenna by monitoring the rectified polarized DC voltages and varying the values of the inductance of the antenna (Figure S2 in Supplementary Information). The polarized DC voltage was also decreased with decreasing load's resistance (Figure 2d).

To generate a cyclic waveform with a low frequency (<1 Hz) from the rectified polarized DC voltage, fully gravure printed *p*-type SWNT-based channel network thin film transistors (*cn*TFTs) were used to construct a five stage ring oscillator with *p*-type inverters. The *cn*TFTs can operate under less than 20 V and their mobility range (0.01 to 10 cm^2/Vs) can be controlled through the loading concentration of SWNT in the ink[18–21]. To print the five stage ring oscillator, we used the same silver and BaTiO$_3$ inks as in the printing of the rectenna for reducing a number of printing steps while SWNT ink was used for printing the channel of the *cn*TFTs. To manufacture the ring oscillator, roll-to-plate (R2P) gravure was used to print 10 *cn*TFTs with a gate width of 300 μm and channel length of 160 μm for both the drive and load *cn*TFTs (Figure S3 in Supplementary Information). The optimized cell depth of gravure plate was 37–40 μm with the cell wall thickness of 50–55 μm for the printing active layers of drive and load TFTs using low viscosity of SWNT based ink (~30 cp) (Figure S4 in Supplementary Information). By repeating the printing of active layers on load TFTs, a higher SWNT network density can be achieved to generate higher current in the

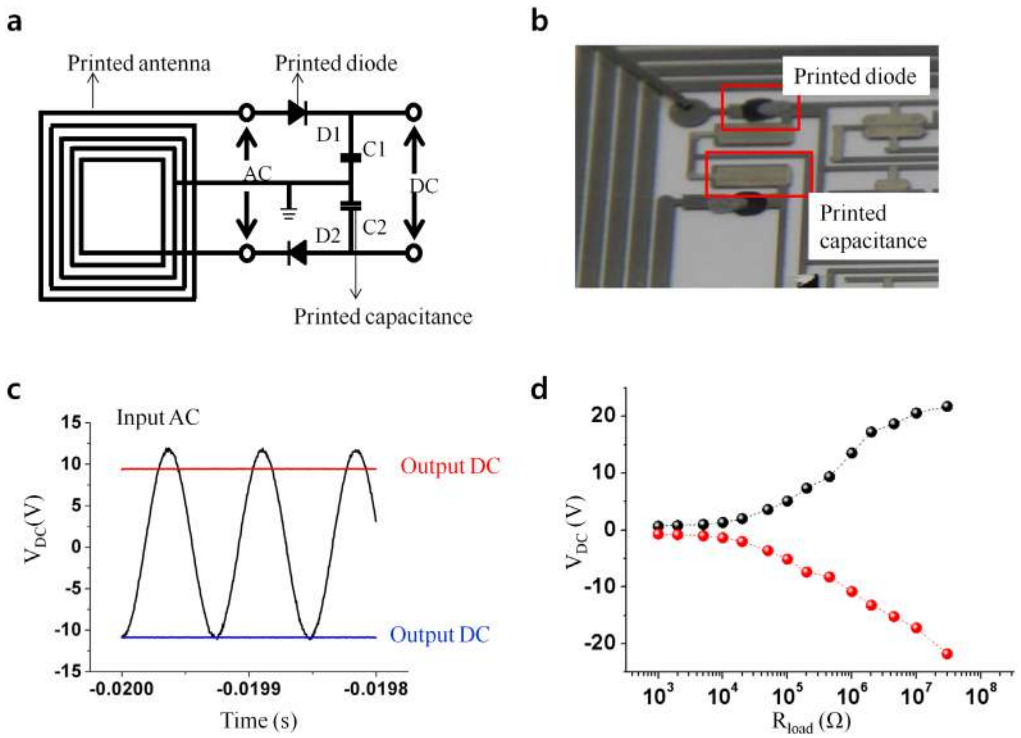

Figure 2 | (a) Circuit layout of the printed rectenna to provide polarized DC power, (b) a real image of printed diodes and capacitors, (c) input-output electrical characteristics of the rectifier at 13.56 MHz AC input and (d) output ± DC voltages under variations in load resistance.

channels while thicknesses of all printed gate, dielectric and drain-source electrodes are kept same (Figure S5 in Supplementary Information).

Their output and transfer characteristics are shown in Figure 3a and 3b respectively. Based on the attained transfer characteristics, the mobility (μ), transconductance (g_m), on-off current ratios from I_{DS}-V_{DS} and threshold voltages (V_{th}) for both five drive *cn*TFTs and five load *cn*TFTs were extracted and shown in Table S1 (Supplementary Information). The shift range of V_{th} from the fully gravure printed 10 *cn*TFTs were about ±0.4 V for the drive *cn*TFTs and ±0.6 V for load *cn*TFTs. Those shift ranges are narrow enough to run the printed ring oscillator for generating the triangular wave. Each inverter delivered a gain of 3 (Figure 3c), and the ring oscillator with five inverters generated a pseudo triangular waveform with a voltage amplitude of 7 V at 320 mHz (Figure 3d). In fact, the pseudo triangular wave can be only generated by the printed *cn*TFTs based ring oscillator because of the parasitic and trapped channel capacitances due to carbon nanotube network structures in *cn*TFTs. Although we only showed the frequency of 320 mHz at here, the frequency can be varied from 0.3 Hz to 2 Hz based on the printed network density of SWNT of printed *cn*TFTs. This is an advantage of our printed *cn*TFTs for scanning electrochemical cell in CV tag over other technologies such as amorphous silicon TFTs where only a sine wave with a designed single frequency can be generated because of no trap capacitance in the channel (Figure S6 in Supplementary Information). The frequency of the triangular waveform can be fine-tuned in the range of 0.3 Hz to 2 Hz as shown in the Figure S7 (Supplementary Information) by simply changing loading concentration of SWNTs.

When the generated triangular waveform is supplied directly into the printed electro-chemical cell (Figure S8 in Supplementary Information), wherein printed silver and carbon electrodes were used with a poly(ethylene oxide) and $LiCF_3SO_3$ gel type electrolyte, the output voltage range is reduced to zero because of improper impedance matching between the electrochemical cell (10 KΩ) and ring oscillator (1 MΩ). Therefore, a buffer unit is needed to match the impedance levels between the ring oscillator and electrochemical cell. The printed buffer unit (Figure 1b-③) with high on currents and low on-off current ratios (Figure 3e) is consists of 6 *cn*TFTs and a resistor (~7 KΩ). In the printed CV tag, three resistors were printed by screen printer using a carbon paste (DC-20, purchased from Dozen Co. Korea) for the buffer units (Figure 1b-③ and ⑥) and the electrochemical cell (Figure 1b-①). Their electrical parameters are listed in Table S1 (Supplementary Information). The 6 *cn*TFTs of buffer unit were printed using R2P gravure with high SWNT concentration in the semiconducting ink. Each *cn*TFTs exhibited on-currents in the range of 400 ~ 500 μA with a very low on-off ratio due to the enhanced metallic percolation in the dense SWNT networks. The buffer units with low on-off ratio were employed to reduce the input voltage to ±500 mV (Figure 3f) which is the appropriate scanning range of the electrochemical cell. The printed electrochemical cell with two electrodes was proven to be well operated under commercial CV (Figure S9 in Supplementary Information). The resulting current after scanning printed electrochemical cell in the CV tag can be converted to voltage by connecting the printed resistor (Figure 1b-①: ~5 × 10^3 ohm) which provides the output voltage for electrochromic indicator and re-plotted into the cyclic voltammogram.

Because these output voltage (±70 mV) and current (~40 μA) were low after scanning the electrochemical cell, they need to be amplified to turn on the electrochromic indicator for showing predetermined concentration (10 mM) level of N,N,N′,N′-tetramethyl-p-phenylenediamine (TMPD, purchased from Aldrich) as a standard reference in this work (Figure 3g). After running the redox reaction of TMPD (Figure 3h) in the printed cell, the output voltage and current needed to be amplified to reach 1.5 V and 500 μA, respectively, to provide sufficient power to display the letters "PE" as an indicator of the presence of TMPD in the specimen. To construct the amplifier, *cn*TFTs based three inverters were printed with a gain of three to five (Figure S10 in Supplementary Information) to amplify the output signal as shown in Figure 3i. The electrical characteristics of the 6 *cn*TFTs in the inverters with extracted mobility, transcon-

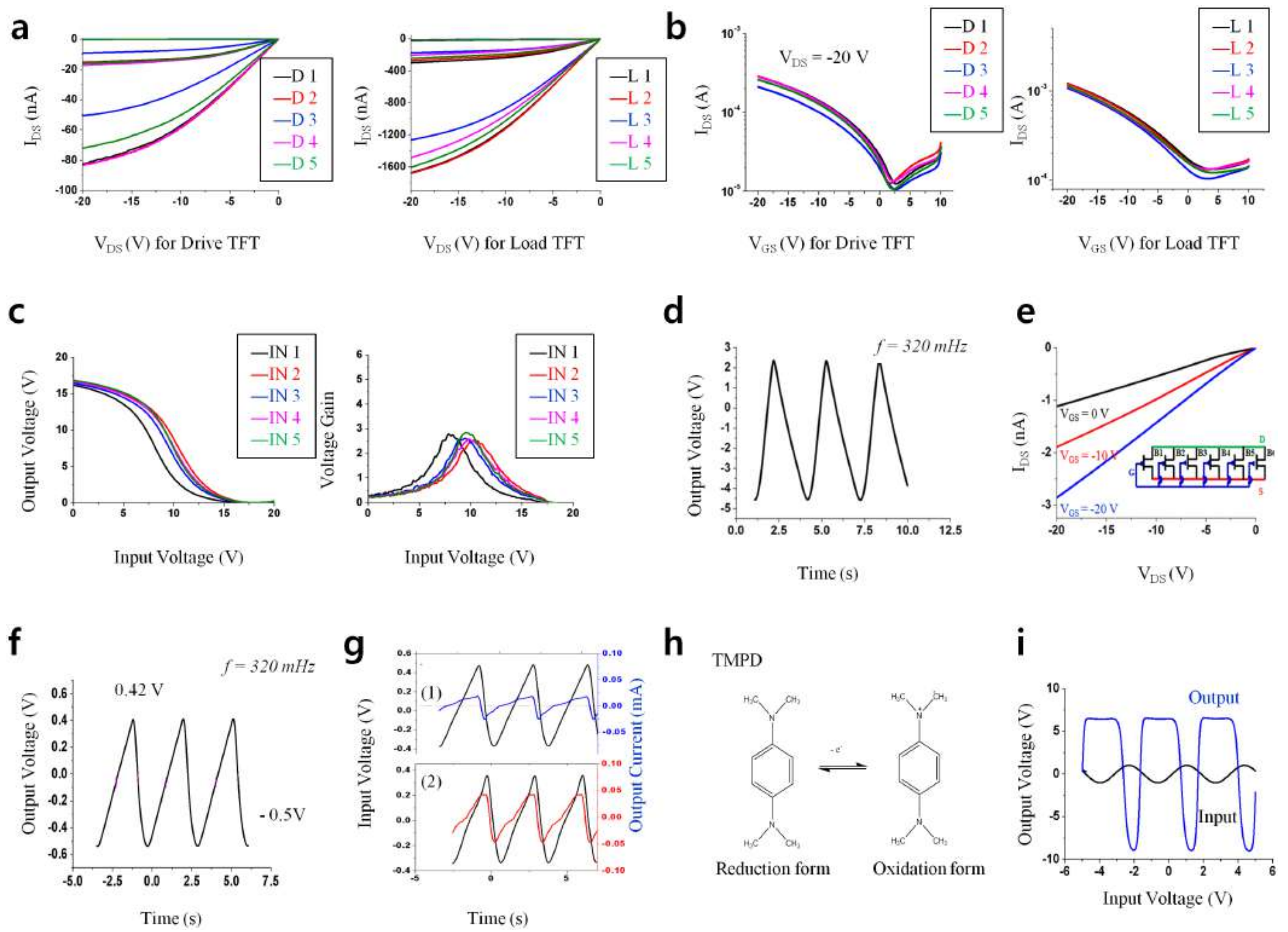

Figure 3 | (a and b) Output and transfer characteristics of the printed *cn*TFTs for 5 drive and 5 load TFTs respectively in the printed ring oscillator. (c) Electrical characteristics of inverters and (d) output characteristic of the ring oscillator. (e) Total output characteristics of the printed buffer unit consisting of 6 *cn*TFTs and a resistor (measured based on contacting gate and drain-source electrodes as shown in the inset circuit). (f) Modified triangular wave following the buffer unit. (g) Generated signals before scanning (black) and after scanning the electrochemical cells without (1, blue) and with TMPD (2, red) in a drop of solution. (h) TMPD structure for oxidation and reduction reaction. (i) The input and amplified output signals after passing through three amplifying inverters.

ductance, on-off current ratios and threshold voltage are shown in Table S2 (Supplementary Information). As the impedance was matched between the ring oscillator and the electrochemical cell using the buffer unit, another buffer unit (Figure 1b-Ⓖ) is also needed between the electrochromic signage (20 KΩ) and the amplifier (1 MΩ) for impedance matching. The electrical parameters of the buffer are listed in Table S2 (Supplementary Information).

The amplified output signal was used to run the reduction and oxidation of the patterned conducting polymer in the electrochromic signage, and thus the blinking signage will be used to indicate the presence of chemicals in the specimens. The signage was fully printed and attached on printed CV tag. A clear concept of the printing sequence for the electrochromic signage was given in Figure S11 in Supplementary Information.

Discussion

The fully printed flexible CV tag was completed by assembling the printed circuits including ring oscillator, buffer, electrochemical cell, amplifier, and signage onto the previously R2R gravure printed rectenna. The resultant CV tag is shown in Figure 4a. The working concept of the wireless and flexible CV tag is demonstrated in the following sequences (watch the video file by clicking Figure S12 in Supplementary Information). After dropping 500 μl of TMPD solution (10 mM) on the printed electrochemical cell, the CV tag was placed on the custom made RF (13.56 MHz) reader (Figure S13 in Supplementary Information). The antenna was subsequently coupled to 13.56 MHz AC. The coupled AC was rectified into polarized DC (> ± 10 V) through two diodes and two capacitors which caused the ring oscillator to generate a pseudo triangular waveform with output voltage of 7 V at 320 mHz. This was then passed through the buffer unit to scan the electrochemical cell. The output signal after scanning the electrochemical cell was amplified to turn on the signage according to the concentration level of TMPD in the solution. In this work, whenever the concentration of TMPD was higher than 10 mM, the signage "PE" blinked (Figure 4b) while it did not show anything (Figure 4c) when it was lower than 10 mM. Furthermore, clear cyclic voltammograms for scanning the electrochemical cell with and without 10 mM of TMPD were obtained by re-plotting the output triangular voltage waveform (Figure 4d). The attained half potential ($E_{1/2}$ ~ 0.05 V, oxidation potential is 0.22 V and reduction potential is −0.17 V) of TMPD from the CV tag was nearly identical to the value obtained from a commercial CV instrument, SP-240, Biologic, which uses the same frequency for the triangular voltage waveform and printed electrochemical cell (Figure 4e). However, when the generated triangle wave frequency is higher than 0.6 Hz in the CV tag, the clear redox peaks cannot be

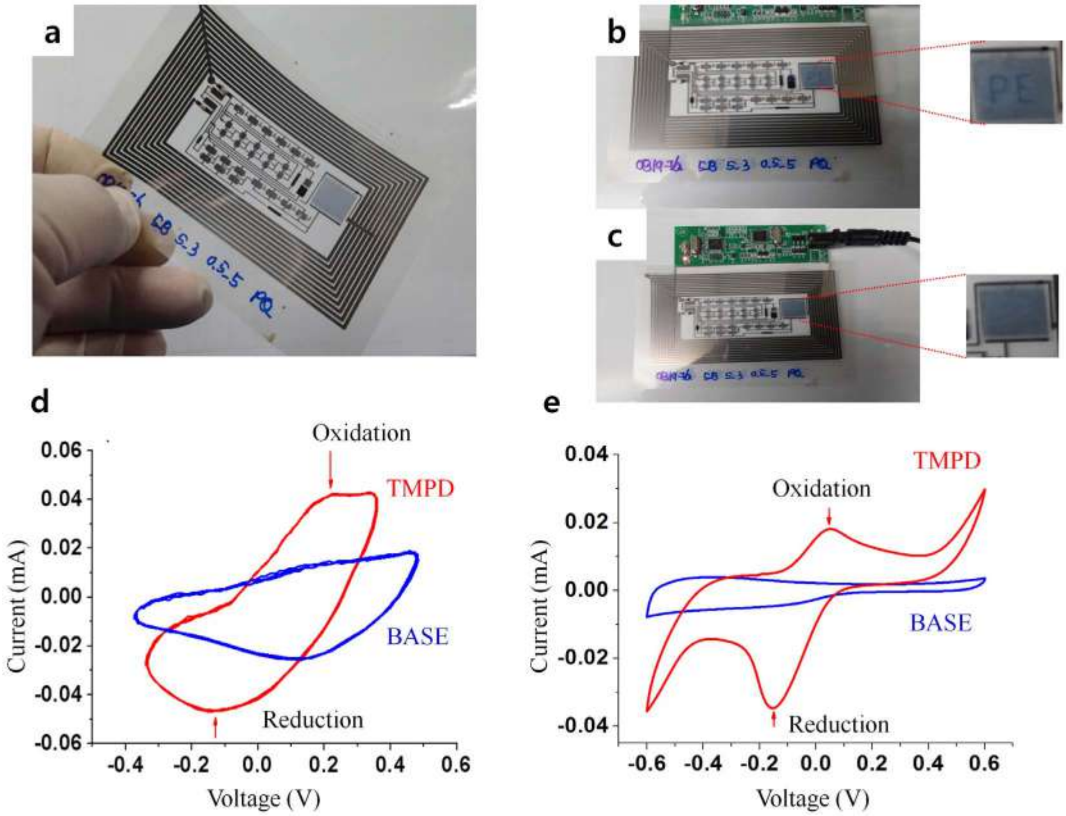

Figure 4 | (a) Optical image of fully printed wireless cyclic voltammetry tag. (b) Operation images of CV tag with and (c) without TMPD in the solution on 13.56 MHz reader. (d) Converted cyclic voltammogram from the printed wireless CV tag *vs* (e) a commercial CV instrument (please refer the interconnected video file for the demonstration of wireless CV tag operation in Figure S12 in Supplementary Information).

observed as shown in Figure S14 (Supplementary Information). Furthermore, as a typical example of testing a specimen in aqueous solution, 10 mM of $K_3(FeCN)_6$ aqueous solution was checked using the CV tag, and results of converted cyclic voltammogram was shown in Figure S15 (Supplementary Information). Those results support that the CV tag can examine specimens not only in organic solution, but also in aqueous one.

Up to present, there has been no commercial success in fully printed TFT-based electronic devices due to some kind of difficulties in integrating a number of printed TFTs through a scalable printing method. As increasing a number of integrated TFTs, the range of threshold voltage shits of TFTs needs to be controlled as narrow as possible to properly operate designed function of printed devices. However, the scalable printing method is not fully understood to integrate a number of TFTs on plastic foils while keeping the constant range of threshold voltage shifts because of difficulty in controlling a constant charge transport through the printed channels. Therefore, the printed TFT-based devices should be constructed by using a minimum number of printed TFTs while a function of printed device needs to be competitive over Si based one. Since the fully printed CV tag required the integration of only 26 *cn*TFTs, the acceptable range of threshold voltage shits of *cn*TFTs is controllable and achievable by using a scalable printing method such as a roll-to-roll gravure. Furthermore, the unique function of printed CV tags cannot be simply reproduced using a Si technology with a comparable cost to the R2R printing process. Therefore, the fully printed CV tag will be the first leading model to show a way of how printed TFT-based electronic devices should go for the successful launching in the IT markets.

In the core of the printed CV tag, fully gravure-printed *cn*TFTs were employed for constructing a circuit consisting of a ring oscillator to generate triangular waveforms, buffer units for an impedance matching, and an amplifier for enhancing output signal to turn on the signage. Those simple units can be combined with a variety of printable sensors and RF devices to create novel devices with an extremely low cost. In other words, the printed rectenna, ring oscillator, buffer and amplifier units in the printed CV tag will be key elements for the construction of variety of printed RF-sensors, this CV tag will be used as a platform for the further construction of a variety of RF based electrochemical sensors.

In conclusion, we have presented a fully printed flexible CV tag with the size of 9.5 × 5.5 cm^2 that can operate wirelessly using a 13.56 MHz RF reader. However, for the practical purpose, the size of the CV tag can be reduced up to the half (4 × 2 cm^2) of the current one by optimizing spaces between *cn*TFTs in the circuit of the CV tag and reducing the operation power to less than DC 5 V as the current size of 13.56 MHz antenna is designed to provide maximum coupled AC power using a sufficient length of printed antenna pattern (higher inductance). Furthermore, although we utilized R2P gravure, screen printer and drop casting method for the convenience of fabricating the CV tag in the Lab, the whole fabricating process can be repeated by R2R gravure because the rectenna, ring oscillator, buffers, electrochemical cell, amplifier and signage in the CV tag were all designed to compromise the limit (±20 μm) of a registration accuracy of a current R2R gravure system. Therefore, this technology could be used as a platform for disposable wireless electrochemical sensors to a variety of specimens in aqueous and organic solutions to monitor or diagnose pathogens and hazardous materials by modifying the electrodes of the electrochemical cells.

Methods

Roll-to-roll (R2R) gravure printed antenna, bottom electrodes, gate electrodes and wires. Two color units of R2R gravure system (Taejin Co. Korea: Figure S1 in the Supplementary Information) was employed to print antenna, bottom electrodes and wires on a roll of poly(ethylene terephthalate) (PET) film (width of 200 mm and thickness of 100 μm, purchased from SKC, Korea) with silver nanoparticle based conducting ink (PG-007 BB type, Paru Co. Korea) and the web transfer speed of 8 m/min under roll pressure of 0.5 MPa and web tension of 60 N. The silver ink was further formulated to meet the viscosity of 500 cp and surface tension of 47 mN/m using respectively ethylene glycol (Aldrich) and Dipropylene glycol methyl ether (Aldrich). The curing time of printed silver layer was 10 sec by passing through a heating chamber of 150°C, installed in R2R gravure system. The resulting printed silver patterns on PET roll showed the thickness of 450 (±50) nm without any defects and were shown in Figure S16a (Supplementary Information).

Roll-to-plate (R2P) gravure printed ring oscillator, buffer, and amplifier. Full R2P gravure system (Figure S1 in the Supplementary Information) was employed to print single walled carbon nanotube network based thin film transistors (*cn*TFTs), used as a building block to construct all the integrated circuit of printed cyclic voltammetry (CV) tag (Figure 1). 26 *cn*TFTs for the integration of circuit in printed CV tag were printed on the previously R2R gravure printed film (printed antenna, bottom electrodes and wires with thickness of 450 nm; Figure S16b in Supplementary Information) following printing sequences. First, $BaTiO_3$ nanoparticle based dielectric ink (PD-100, Paru Co., Korea) was further formulated to meet viscosity (75 cP) and surface tension (31.8 mN/m) using the ethyl 2-cynoacrylate (Aldrich). As shown in R2P gravure printing scheme (Figure S17 in Supplementary Information), the homogeneous and defect free dielectric layers were first R2P gravure printed with a printing speed of 200 mm/s and roll pressure 7.5 kg_f on previously printed gate electrodes and bottom electrodes (for capacitors) with a thickness of 2.3 (±0.1) μm and width of 940 (±2) μm (Figure S5 in Supplementary Information). The printed dielectric layers were cured for 1 min 20 sec at 150°C. Second, single walled carbon nanotube (SWNT) based semiconducting ink (PR-040, Paru Co., Korea) was diluted by diethylene glycol monobutyl ether (Daejung Co., Korea) to 1 : 1 volume ratio for printing active layers using R2P gravure on the printed dielectric layers with a printing speed of 350 mm/s and a roll pressure of 4 kg_f. The viscosity and surface tension of diluted SWNT ink were respectively 30 cP and 29 mN/m. The resulting SWNT printed films were dried 1 min at 150°C. For using *cn*TFTs in the buffer unit, we used the dilution volume ratio of 5 : 1 between SWNT ink (PR-040, Paru Co., Korea) and diethylene glycol monobutyl ether for printing only on the TFTs in the buffer unit to increase SWNT network density for providing the high current (Figure S5 in Supplementary Information). Finally, drain-source electrodes were R2P gravure printed using silver nanoparticle based ink (PG-007 AA type, Paru Co., Korea) on SWNT printed film with a viscosity of 800 cP and a surface tension of 42 mN/m. The R2P gravure printing was carried out under a printing speed of 200 mm/s and a role pressure of 5 kg_f.

Diode fabrication. Although Shottky diodes were able to print using R2R gravure based on our previous reported process[16,17], the drop casting method was used for the convenience in this work. First, dropping ZnO based hybrid semiconducting ink (PD-070, Paru Co., Korea) on printed silver electrode and then, Al ink (PA-009, Paru Co., Korea) was dropped on the top of ZnO ink. The resulting diodes were further cured for 5 min at 120°C.

Measurements. Formulated and diluted inks were characterized using DCAT21 (Dataphysics Co., Germany) and SV-10 Vibro viscometer (AND Co., Japan) for respectively measure surface tension and viscosity. All measurements in this work were carried out under ambient condition if there are no other comments. Resistance of printed antenna, bottom electrodes, gate electrodes and wires were measured using two probe methods with HP3457A multimeter. Both surface morphology and thickness of printed layers was characterized using surface profiler (NV-2200, Nanosystem, Korea). LCR meter (E4980A, Agilent, USA), semiconductor parameter analyzer (4155C, Agilent, USA) and digital phosphor oscilloscope (DPO 4054, Tektronix, USA) were respectively used to characterize printed *cn*TFT based circuits.

1. Sanderson, M. D., Kamplain, J. K. & Bielawski, C. W. Quinone-annulated N-heterocyclic carbene-transition-metal complexes: Observation of π-back bonding using FT-IR spectroscopy and cyclic voltammetry. *J. Am. Chem. Soc* **128**, 16514–16515 (2006).
2. Lai, R. Y., Fabrizio, E. F., Lu, L., Jenekhe, S. A. & Bard, A. J. Synthesis, cyclic voltammetrics studies, and electrogenerated chemiluminescence of a new donor-acceptor molecule: 3,7-[Bis[4-phenyl-2-quinolyl]]-10-methylphenothiazine. *J. Am. Chem. Soc* **123**, 9112–9118 (2001).
3. Armstrong, F. A., Hill, H. A. & Walton, N. J. Direct electrochemistry of redox proteins. *Acc. Chem. Res* **21**, 407–413 (1998).
4. Norouzi, P., Ganjali, M. R., Daneshgar, P., Alizadeh, T. & Mohammadi, A. Development of fast Fourier transformation continuous cyclic voltammetry as a highly sensitive detection system for ultra-trace monitoring of penicillin V. *Anal. Biochem* **360**. 175–184 (2007).
5. Koyama, J. *et al.* Correlation of redox potentials and inhibitory effects on Epstein-Barr virus activation of 2-azaanthraquinones. *Cancer Letters* **212**, 1–6 (2004).
6. Teijeiro, C., Perez, P., Marin, D. & Paleček, E. Cyclic voltammetry of mitomycin C and DNA. *Bioelectrochemistry and Bioenergetics* **38**, 77–83 (1995).
7. Zerihun, T. & Gründler, P. Electrically heated cylindrical microelectrodes. Determination of lead on Pt by cyclic voltammetry and cathodic stripping analysis. *J. Electroanal Chem* **415**, 85–88 (1996).
8. Bond, A. M. & Scholz, F. Electrochemical, Thermodynamic, and mechanistic data derived from voltammetric studies on insoluble metallocenes, mercury halide and sulfide compounds, mixed silver halide crystals, and other metal complexes following their mechanical transfer to a graphite electrode. *Langmuir* **7**, 3197–3204 (1991).
9. Brusciotti, F. & Duby, P. Cyclic voltammetry study of arsenic in acidic solutions. *Electrochim Acta* **52**, 6644–6649 (2007).
10. Ruan, C., Yang, L. & Li, Y. Immunobiosensor chips for detection of escherichia coli O157:H7 using electrochemical impedance spectroscopy. *Anal. Chem* **74**, 4814–4820 (2002).
11. Liu, G. & Lin, Y. Electrochmical sensor for orgnohposphate pesticides and nerve agents using zirconia nanoparticles as selective sorbents. *Anal. Chem* **77**, 5894–5901 (2005).
12. Ascaso, J. F. *et al.* Diagnosing insulin resistance by simple quantitative methods in subjects with normal glucose metabolism. *Diabetes Care* **26**, 12, 3320–3325 (2003)
13. Karl, T. *et al.* Human breath isoprene and its relation to blood cholesterol levels: new measurements and modeling. *J. Appl. Physiol* **91**, 762–770 (2001).
14. Armstrong, F. A. Recent developments in dynamic electrochemical studies of adsorbed enzymes and their active sites. *Chem. Biol* **9**, 110–117 (2005).
15. Mollah, M. Y., Schennach, R., Parga, J. R. & Cocke, D. L. Electrocoagulation (EC)-science and applications. *J. Hazard. Mater* **B84**, 29–41 (2001).
16. Park, H. *et al.* Fully roll-to-roll gravure printed rectenna on plastic foils for wireless power transmission at 13.56 MHz. *Nanotech* **23**, 344006–344012 (2012).
17. Kang, H. *et al.* Fully Roll-to-Roll Gravure Printable Wireless (13.56 MHz) Sensor-Signage Tags for Smart Packaging. *Sci. Rep*. **4**, 5387; DOI: 10.1038/srep05387 (2014).
18. Jung, M. *et al.* All-printed and roll-to-roll-printable 13.56 MHz-operated 1-bit RF tag on plastic foils. *IEEE Trans Electron Devices* **57**, 3, 571–580 (2010).
19. Noh, J. *et al.* Integrable single walled carbon nanotube (SWNT) network based thin film transistors using roll-to-roll gravure and inkjet. *Org. Electron* **12**, 2185–5191 (2011).
20. Noh, J. *et al.* Modeling of printed single walled carbon nanotube thin film transistors for attaining optimized clock signals. *J. Appl. Phys* **108**, 10, 102811–102817 (2010).
21. Noh, J. *et al.* Fully Gravure-Printed Flexible Full Adder Using SWNT-Based TFTs. *IEEE. Eletron. Dev. Lett.* **33**, 11, 1574–1576 (2012).

Acknowledgments

The authors would like to thank BK Plus program, Basic Science Research program (NRF-2014R1A1030419), and RIC in Sunchon National University for their financial support. G. C. thanks to Paru Co. for providing all electronic inks in this work.

Author contributions

Y.J., A.J. and G.C. designed the experiments. Y.J., H.P., J.P., Y.C. and M.J. carried out the experiments. J.N., K.J. and K.C. performed device simulations. Y.J. and K.C. performed mobility calculations. Y.J. and M.P. performed electrochemistry and analyzed electrochemical data. Y.J., H.P., M.P., A.J. and G.C. contributed to analyzing the data. M.J., M.P., A.J. and G.C. wrote the paper while all authors provided feedback.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: Prof. Cho was a consultant of PARU Co. Korea and received compensation from PARU Co. Both Dr. M. Jung and Mr. K Jung has been employee of PARU Co. Korea. Dr. Y. Jung, Miss H. Park, Ms. J. Park, Dr. J. Noh, Mr. Y. Choi, Prof. Pyo, Mr. K. Chen and Prof. Javey declare no competing financial interests.

How to cite this article: Jung, Y. *et al.* Fully printed flexible and disposable wireless cyclic voltammetry tag. *Sci. Rep.* **5**, 8105; DOI:10.1038/srep08105 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

View publication stats

AIP Applied Physics Letters

A microwave metamaterial with integrated power harvesting functionality

Allen M. Hawkes, Alexander R. Katko, and Steven A. Cummer

Citation: Applied Physics Letters **103**, 163901 (2013); doi: 10.1063/1.4824473
View online: http://dx.doi.org/10.1063/1.4824473
View Table of Contents: http://scitation.aip.org/content/aip/journal/apl/103/16?ver=pdfcov
Published by the AIP Publishing

Goodfellow
metals • ceramics • polymers
composites • compounds • glasses
Save 5% • Buy online
70,000 products • Fast shipping
www.goodfellowusa.com

This article is copyrighted as indicated in the abstract. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 152.3.43.156 On: Fri, 08 Nov 2013 20:28:30

APPLIED PHYSICS LETTERS **103**, 163901 (2013)

A microwave metamaterial with integrated power harvesting functionality

Allen M. Hawkes,[a] Alexander R. Katko, and Steven A. Cummer
Department of Electrical and Computer Engineering, Duke University, 101 Science Drive, Durham, North Carolina 27708-9976, USA

(Received 30 May 2013; accepted 22 September 2013; published online 14 October 2013)

We present the design and experimental implementation of a power harvesting metamaterial. A maximum of 36.8% of the incident power from a 900 MHz signal is experimentally rectified by an array of metamaterial unit cells. We demonstrate that the maximum harvested power occurs for a resistive load close to 70 Ω in both simulation and experiment. The power harvesting metamaterial is an example of a functional metamaterial that may be suitable for a wide variety of applications that require power delivery to any active components integrated into the metamaterial. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4824473]

Metamaterials are composed of sub-wavelength particles that exhibit bulk properties that are different from their individual components. Electromagnetic metamaterials are engineered materials that can achieve parameters not possible within naturally occurring materials, such as a negative index of refraction[1] or a zero index of refraction.[2] Exotic properties like these allow for a variety of interesting applications including a superlens device[3] and an invisibility cloak.[4] Integrating active and nonlinear functionality into metamaterials has been demonstrated in the form of dynamic resonant frequency tuning,[5,6] phase conjugation,[7] and wave mixing.[8,9] More specific functional behavior has also been demonstrated in metamaterials, including radio frequency (RF) limiting[10] and harmonic generation.[11]

Metamaterials are also well-suited for other functional behaviors, including electromagnetic power harvesting, the focus of this work. Power harvesting devices convert one type of energy to another, typically converting to a direct current (DC) signal. Many types of energy can be harvested, from acoustic (using a piezoelectric harvester)[12] to electromagnetic (using a rectenna).[13] Power harvesting devices require a method to couple to the energy that will be harvested as well as a device to convert the energy from one form to another. By their very nature, metamaterials are designed to couple to various types of energy, e.g., from acoustic[14] to optical,[15] and thus provide a natural platform for power harvesting. Electromagnetic metamaterials provide flexibility in design due to their electrically small, low-profile nature.[16] Since metamaterials are typically designed as infinite arrays, the resonant frequency and input impedance include coupling effects. Metamaterials can be adapted to various applications, such as flexible sheets to cover surfaces.[17] Moreover, many metamaterials that have been presented in the literature require some form of external signal. This could be a DC bias voltage[18] or a large external pump signal.[7] In general, metamaterials could be modified to harvest such an external signal that is already present for other purposes. With these design advantages, power harvesting metamaterials offer design flexibility for a large number of applications that general antenna-based microwave power harvesting devices may lack.

A recent simulation-based study[19] investigated the conversion efficiency between incident RF power and induced power in a split-ring resonator (SRR). Our work is focused on the experimental measurement of RF to DC efficiency based on the conventional effective area of the SRR. We demonstrate that metamaterials can also include embedded devices to convert the incident RF energy to a DC voltage, providing a platform for power harvesting that utilizes the advantages of metamaterial design.

An SRR is a canonical example of a resonant metamaterial particle and is used as the basis for the unit cells of the metamaterial power harvester designed here. By tuning the SRR parameters, we design an SRR (Fig. 1) to resonate at 900 MHz using an S-parameter simulation within Computer Simulation Technology (CST) Microwave Studio software. Using CST Microwave Studio, we can also simulate the effects of embedding devices within the SRR by retrieving its S-parameters using a lumped port. The retrieved S-parameters are loaded into Agilent Advanced Design System (ADS), allowing us to simulate both fullwave 3D effects and circuit-level nonlinear effects.

An SRR couples strongly to an incident magnetic field and can be loaded with a wide variety of circuit elements. In this work, we embed a rectifying circuit within an SRR to convert the incident RF power to DC power. A number of rectifying circuits could be chosen depending on the particular application for the power harvesting metamaterial. We choose to use a Greinacher[20] circuit because the output voltage is double the input voltage maximum, which allows for

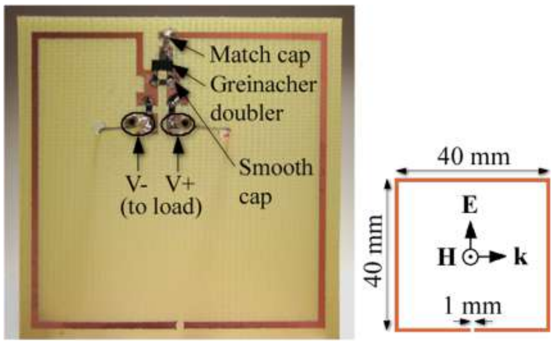

FIG. 1. (Left) Power harvesting SRR. (Right) Plain SRR with dimensions shown, 1 mm traces. Incident wave polarization is also shown.

---

[a] allen.hawkes@duke.edu

This article is copyrighted as indicated in the abstract. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 152.3.43.156 On: Fri, 08 Nov 2013 20:28:30

easier power transmission and measurement. A Greinacher circuit has a lower effective capacitance in comparison to other rectification circuits (such as a bridge rectifier), allowing faster switching and thus a higher frequency of operation. A Greinacher circuit also has a low threshold voltage, allowing operation at lower incident power levels. A Greinacher voltage doubler can be placed across a gap (Fig. 1) in the top side of the copper trace to rectify the induced current present in the SRR. Schottky diodes are used for the voltage doubler due to their typically low open junction capacitance and fast switching capabilities, which allow for rectification of a high-frequency RF signal, as well as their typically low threshold voltage. A resistive load placed across the output of the voltage doubler is a simple way to determine DC power out using $P = V^2/R$. This DC power is maximized through ADS for matching capacitor and resistor values. The parameters of our selected Schottky diode (HSMS 2862) and the simulated S-parameters of the above SRR in a model of our actual parallel-plate waveguide are input into ADS for the simulation, shown in a schematic in Fig. 2.

The simulated components lead to a maximum efficiency of 61% for an input power of 24.25 dBM, the maximum available experimental power (at an incident power density of approximately 1.6 mW/cm^2). One important figure of merit for a power harvester is its RF-to-DC power conversion efficiency:

$$\eta = \frac{P_{DC}}{P_{RF}}.$$

We determine P_{RF} by measuring the total incident power in our measurement apparatus. For a large metamaterial sample, we assume that the total incident power to the measurement apparatus is incident on the metamaterial. For a single unit cell, it is necessary to use the effective area of the unit cell to determine the incident P_{RF}. The maximum effective area may be calculated by[21]

$$A_{e,max} = \frac{\lambda^2}{4\pi} D_0,$$

where $D_0 = 1.5$ since the SRR is effectively a small loop illuminated by a transverse electromagnetic (TEM) wave. For an SRR resonant at 900 MHz, the effective area is thus $A_{e,max} = 5.3A_{physical}$. The full waveguide area is approximately $6.8A_{physical}$ where $A_{physical}$ is the physical area of a single SRR unit cell. As $A_{e,max}$ is 78% of the open wave guide area, 78% of the incident power density is used as input power for simulation of the single cell.

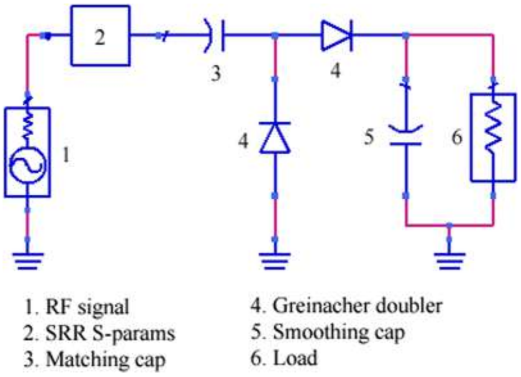

FIG. 2. ADS simulation schematic of SRR power harvester. CST Microwave Studio was used to determine the SRR S-parameters.

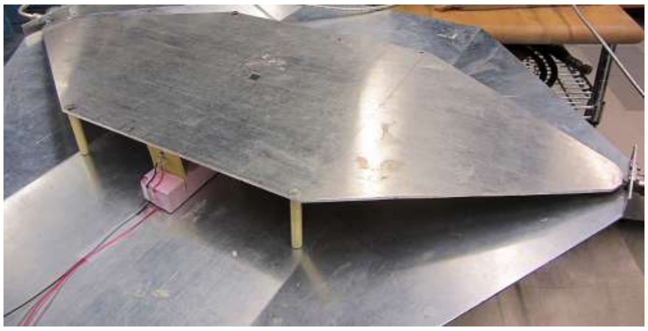

FIG. 3. Placement of SRR within open waveguide.

The designed voltage doubler and resistive load are added to an SRR as shown in Fig. 1, resulting in the power harvesting metamaterial unit cell. To observe power harvesting capabilities, the cell is placed in an open, TEM waveguide (Fig. 3) where input power is produced by a signal generator and amplifier, and output power is measured with an oscilloscope via leads placed across the resistive load (Fig. 4). The DC power harvested is determined by $P = V^2/R$ as previously mentioned, and input power is measured with a spectrum analyzer connected to the signal generator and amplifier via the open waveguide. By increasing the incident power from 13 to 24 dBm and measuring the DC output from the SRR, the normalized harvested power, $P_{DC}/P_{RF,incident}$, as a function of incident power $P_{RF,incident}$ and resulting efficiency are determined at each point. The maximum efficiency of the single cell is 14.2%, setting $P_{RF,incident}$ as 78% of the total input power from effective area calculations.

Multiple power harvesting SRR cells are then tested simultaneously to create the power harvesting metamaterial, which is accomplished through a 5 × 1 array shown in Fig. 5. Through a parallel connection of the leads from each SRR's resistive load, the total power harvested by the metamaterial is found in the same way as the single cell. The maximum efficiency for the 5 × 1 array is 36.8%, where P_{in} is the entire input power because the array spans the entire length. Measured efficiencies of both the single and the array of power harvesting SRRs are shown in Fig. 6. Also shown in Fig. 6 is the open circuit voltage, V_{OC}, a load-independent measure of the available voltage harvested by the array of SRRs.

The power harvesting metamaterial array is more efficient than the single unit cell. This is partially due to the

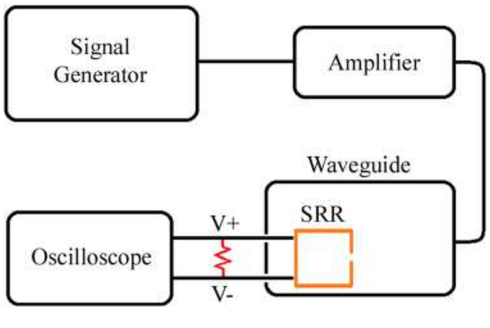

FIG. 4. Experimental test setup schematic.

This article is copyrighted as indicated in the abstract. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 152.3.43.156 On: Fri, 08 Nov 2013 20:28:30

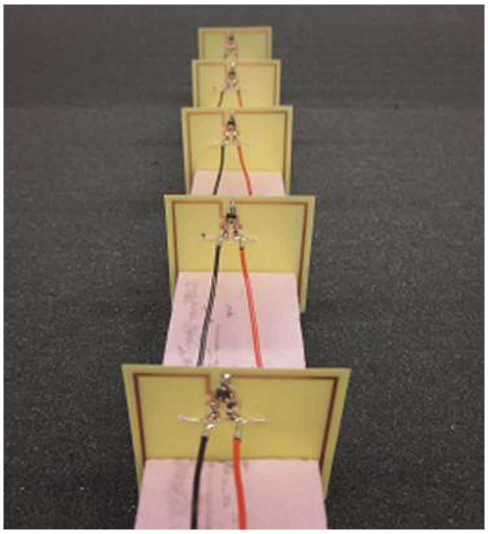

FIG. 5. 5 × 1 array of power-harvesting SRRs.

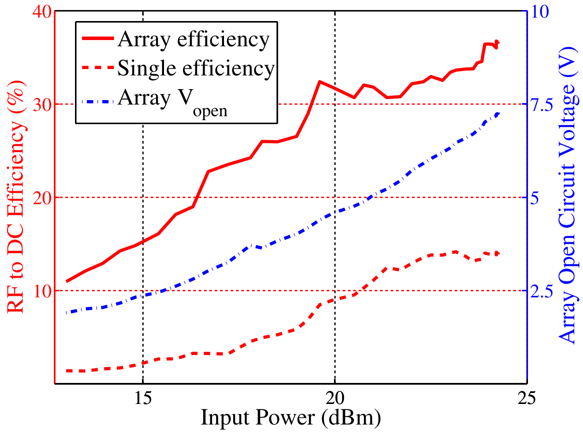

FIG. 6. RF to DC efficiency of power harvesters. The open-circuit voltage of the array is also shown.

larger effective area of the array. While the effective area of the single cell is 78% the entire width of the waveguide, some of this power is not harvested by the single unit cell due to fringing effects on the sides of the waveguide. For this reason, placement of multiple cells that together span the entire waveguide width results in a higher efficiency as the array captures more of the electromagnetic energy that undergoes the fringing effect.

Another important relationship is the efficiency of the power harvesting metamaterial as a function of load resistance. Simulated and experimental efficiencies for the 5 × 1 array are shown in Fig. 7. Though the experimental efficiencies do not match the values from simulation, the simulation does accurately predict the maximum harvested power as falling approximately within the range of 70–80 Ω, confirmed by the experimental data. The experimental efficiency maximum occurs at 70 Ω, and the simulated maximum occurs at 82 Ω, showing close correspondence.

In summary, we have designed, simulated, and experimentally measured a functional metamaterial power harvester capable of converting up to 36.8% of the incident RF power to DC power. Through a parallel connection of five SRRs, a V_{OC} of 7.3 V is achieved. Simulations match experimental results showing an optimal resistive load for DC power transfer of 70–80 Ω. The SRR power harvester is an example of functional metamaterial that may be suitable for a wide variety of RF applications that require power delivery to any active integrated components.

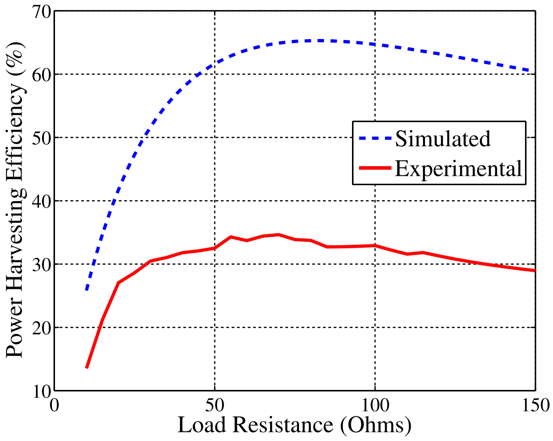

FIG. 7. RF to DC efficiency of SRR array as a function of load resistance, for both simulation and experiment. Note that both show maximum efficiency for around 70–80 Ω.

This work was supported by a Multidisciplinary University Research Initiative from the Army Research Office (Contract No. W911NF-09-1-0539).

[1] D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, and S. Schultz, Phys. Rev. Lett. **84**, 4184–4187 (2000).
[2] R. W. Ziolkowski, Phys. Rev. E **70**, 046608 (2004).
[3] N. Fang, H. Lee, C. Sun, and X. Zhang, Science **308**(5721), 534–537 (2005).
[4] D. Schurig, J. J. Mock, B. J. Justice, S. A. Cummer, J. B. Pendry, A. F. Starr, and D. R. Smith, Science **314**(5801), 977–980 (2006).
[5] I. V. Shadrivov, S. K. Morrison, and Y. S. Kivshar, Opt. Express **14**(20), 9344–9349 (2006).
[6] H.-T. Chen, J. F. O'Hara, A. K. Azad, A. J. Taylor, R. D. Averitt, D. B. Shrekenhamer, and W. J. Padilla. Nature Photon. **2**(5), 295–298 (2008).
[7] A. R. Katko, S. Gu, J. P. Barrett, B.-I. Popa, G. Shvets, and S. A. Cummer, Phys. Rev. Lett. **105**, 123905 (2010).
[8] D. Huang, A. Rose, E. Poutrina, S. Larouche, and D. R. Smith, Appl. Phys. Lett. **98**(20), 204102 (2011).
[9] A. Rose, D. Huang, and D. R. Smith, Phys. Rev. Lett. **110**, 063901 (2013).
[10] A. R. Katko, A. M. Hawkes, J. P. Barrett, and S. A. Cummer, IEEE Antennas Wireless Propag. Lett. **10**, 1571–1574 (2011).
[11] I. V. Shadrivov, A. B. Kozyrev, D. W. van der Weide, and Y. S. Kivshar, Appl. Phys. Lett. **93**(16), 161903 (2008).
[12] M. Lallart, D. Guyomar, C. Richard, and L. Petit, J. Acoust. Soc. Am. **128**, 2739–2748 (2010).
[13] N. Zhu, R. W. Ziolkowski, and H. Xin, Appl. Phys. Lett. **99**(11), 114101 (2011).
[14] J. Li and C. T. Chan, Phys. Rev. E **70**(5), 055602 (2004).
[15] H. J. Lezec, J. A. Dionne, and H. A. Atwater, Science **316**(5823), 430–432 (2007).
[16] R. W. Ziolkowski and A. Erentok, IEEE Trans. Antennas Propag. **54**(7), 2113–2130 (2006).
[17] G. X. Li, S. M. Chen, W. H. Wong, E. Y. B. Pun, and K. W. Cheah, Opt. Express **20**(1), 397–402 (2012).
[18] J. Zhao, Q. Cheng, J. Chen, M. Q. Qi, W. X. Jiang, and T. J. Cui, New J. Phys. **15**, 043049 (2013).
[19] O. M. Ramahi, T. S. Almoneef, M. Alshareef, and M. S. Boybay, Appl. Phys. Lett. **101**, 173903 (2012).
[20] A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 6th edition (Oxford University Press, 2009).
[21] C. A. Balanis, *Antenna Theory: Analysis and Design* (Wiley-Interscience, New York, 2005).

This article is copyrighted as indicated in the abstract. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 152.3.43.156 On: Fri, 08 Nov 2013 20:28:30

Ambient RF Energy Harvesting in Urban and Semi-Urban Environments

Manuel Piñuela, *Student Member, IEEE*, Paul D. Mitcheson, *Senior Member, IEEE*, and
Stepan Lucyszyn, *Senior Member, IEEE*

***Abstract*—RF harvesting circuits have been demonstrated for more than 50 years, but only a few have been able to harvest energy from freely available ambient (i.e., non-dedicated) RF sources. In this paper, our objectives were to realize harvester operation at typical ambient RF power levels found within urban and semi-urban environments. To explore the potential for ambient RF energy harvesting, a city-wide RF spectral survey was undertaken from outside all of the 270 London Underground stations at street level. Using the results from this survey, four harvesters (comprising antenna, impedance-matching network, rectifier, maximum power point tracking interface, and storage element) were designed to cover four frequency bands from the largest RF contributors (DTV, GSM900, GSM1800, and 3G) within the ultrahigh frequency (0.3–3 GHz) part of the frequency spectrum. Prototypes were designed and fabricated for each band. The overall end-to-end efficiency of the prototypes using realistic input RF power sources is measured; with our first GSM900 prototype giving an efficiency of 40%. Approximately half of the London Underground stations were found to be suitable locations for harvesting ambient RF energy using our four prototypes. Furthermore, multiband array architectures were designed and fabricated to provide a broader freedom of operation. Finally, an output dc power density comparison was made between all the ambient RF energy harvesters, as well as alternative energy harvesting technologies, and for the first time, it is shown that ambient RF harvesting can be competitive with the other technologies.**

***Index Terms*—Ambient RF, energy harvesting, maximum power point tracking (MPPT), multiband, rectenna, RF survey, RF-dc.**

I. INTRODUCTION

FOR ALMOST 50 years, far-field RF technology has been used to remotely power systems from relatively large unmanned helicopters [1] to very small smart dust sensors [2] and contact lenses that measure eye pressure [3]. With all these systems, a dedicated RF source is used, where the operator may have control over the effective isotropically radiated power (i.e., both transmit power and antenna characteristics), beam pointing and polarization of the RF source, ensuring optimal line-of-sight operation between the source transmitter (TX) and harvesting receiver (RX). It is important to highlight that this work will focus only in radiative power transfer and not inductive or near-field power transfer, as demonstrated in [4]. A more convenient solution, however, is to power these devices from ambient RF energy sources, such as television and mobile phone signals, thus removing the need for a dedicated source. As ambient RF levels are lower than those that can be provided by a dedicated RF source, the efficiency of the harvesting system, and its minimum startup power are of critical importance.

In order to assess the feasibility of deploying ambient RF energy harvesters, the available RF power needs to be measured in different locations. Such measurements, in conjunction with knowledge on harvester performance, can then be used to determine the locations at which RF harvester powered devices can be successfully deployed. Several RF spectral surveys, which measure ambient RF power levels from sources such as television and mobile phone base stations, have been previously reported. Many have been undertaken using personal exposimeters or spectrum analyzers, where the exact location of each measurement is unknown and with RF power levels only being reported under general scenarios (e.g., outdoor, indoor, street, bus, etc.) [5], [6]. While being of academic interest for health-related research [7], the lack of power level and specific time/location information limits their usefulness for exploitation in ambient RF energy harvesting applications.

Most rectennas (normally comprising an antenna, impedance matching network, rectifier, storage element, and load) presented in the open literature have been tested using dedicated sources rather than harvesting from ambient RF energy [8]. In recent years, efficiencies as high as 78% [9] and 90% [10] have been demonstrated with relatively high input RF power levels (i.e., $> +10$ dBm). Moderate efficiencies have also been achieved using dedicated TXs that provided relatively low input RF power levels; e.g., an efficiency of 60% was achieved with -22.6-dBm input power [11]. In one demonstrator [12], designed to operate over a broad range of input RF power levels, (-30 to $+30$ dBm), the efficiency increased from 5% at a low input RF power to a peak of 80% at $+25$ dBm.

Despite advancements in end-to-end (i.e., input RF to output dc) power conversion efficiencies at low input RF power levels (similar to those measured in the spectral surveys), only a few attempts at true ambient RF energy harvesting have been reported. For example, one relatively efficient rectenna, utilizing a modified omnidirectional patch antenna, has an efficiency of 18% with a single-tone input RF power of -20 dBm [13]. This dedicated signal source was meant to emulate the input RF power levels measured from a nearby digital TV (DTV) TX in Tokyo, Japan, but did not take into account the more realistic effect of harvesting from a modulated broadband signal. Another attempt

Manuscript received November 20, 2012; revised April 24, 2013; accepted April 26, 2013. Date of publication May 24, 2013; date of current version June 28, 2013. This work was supported in part by the U.K. Government and the Mexican National Council of Science and Technology (CONACYT).

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: m.pinuela09@imperial.ac.uk; paul.mitcheson@imperial.ac.uk; s.lucyszyn@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMTT.2013.2262687

0018-9480/$31.00 © 2013 IEEE

at harvesting ambient RF energy from a mobile phone base station at 845 MHz was reported in [14]. This prototype managed to power an LCD thermometer for 4 min, but only after harvesting for 65 h. In that work, when the authors used a dedicated signal source with a single-tone input RF power of -15 dBm, an efficiency of 3% was recorded. A batteryless location sensor has also been demonstrated [15], powered by a rectenna with a printed antenna on a flexible substrate and a solar cell, although no details for the RF-dc efficiency were reported. Finally, successful prototypes capable of harvesting energy using TV antennas were presented, but again no details of their efficiency were given [16], [17].

In order to demonstrate the feasibility for implementing ambient RF energy harvesting, here we first present the results of a citywide RF spectral survey, indicating suitable locations and associated RF bands with sufficient input RF power density levels for harvesting. Based on these results, rectennas were then fabricated and their efficiencies, under ambient RF energy harvesting operation, were calculated using *in-situ* field strength measurements. Furthermore, an investigation of multiband rectenna arrays is also presented, demonstrating the tradeoffs between series (voltage summing) and parallel (current summing) topologies with the aim of reducing the minimum input power required for harvester operation. Finally, a comparison between measured ambient RF energy harvesting and alternative forms of energy harvesting technologies is presented; highlighting, for the first time, the practical feasibility of exploiting existing freely available sources of RF energy.

II. LONDON RF SURVEY

In order to quantify input RF power density levels present in a typical urban and semi-urban environment, a citywide RF spectral survey within the ultrahigh frequency (0.3–3 GHz) part of the frequency spectrum was conducted within Greater London. A number of citywide RF spectral surveys have previously been conducted, but in general, only a few samples were taken, giving little insight into (semi-)urban environments [14], [18], [19]. Other surveys [20], [21] compare their measurements relative to the distance from the nearest TX. In a (semi-)urban environment, this may not provide enough information about the RF spectrum since there is likely to be local geographical variations in base-station density and propagation characteristics (e.g., multipath effects and diffraction around and attenuation through buildings).

Each station on the London Underground network was used as a survey point to provide a robust dataset for representing Greater London in terms of geographical distribution and population density, having a combination of urban (in the center) and semi-urban (in surrounding areas) characteristics. Measurements were taken at each of the 270 stations (from a randomly chosen exit, at street level and a height of 1.6 m). To provide traceability and for use as a historical reference, time stamps and GPS locations were recorded. In addition, measurements were taken inside a building at Imperial College London (ICL), to represent a typical office block within an urban environment.

A. Methodology

Mobile phone usage varies during the daytime, and hence, ambient RF energy in their bands is expected to be time dependant, with more energy available during the daytime than at night time. Therefore, in order to be able to make fair comparisons between locations, measurements were taken between 10:00 am and 3:00 pm on weekdays over a period of one month (between March 5, and April 4, 2012). Electric field strength was measured between 0.3–2.5 GHz using an Agilent N9912A FieldFox RF analyzer [22] with a calibrated Aaronia BicoLOG 20300 omnidirectional antenna [23]. It is important to note that the spectral measurements were undertaken during the analog-to-digital switchover period in the U.K. and so the measurements for DTV may represent an underestimate of present RF power levels measured now that the switch over is complete [24]. It should also be noted that this survey was conducted prior to the 4G network being switched on within the U.K.

A "panning method," which complies with international regulations for measuring exposure limits, was used [25]–[27]. Here, the calibrated antenna is rotated to three orthogonal axes while the spectrum analyser is set to "max-hold," ensuring that the maximum reading is recorded. For each measurement, more than 1 min was allocated to allow for more than three sweeps across the selected frequency range. Additionally, to maintain a comparable signal-to-noise (S/N) ratio, attenuation was introduced (with a minimum set at 5 dB) to avoid compression when high input RF power levels were detected. For all measurements, the resolution bandwidth (BW) was fixed at 100 kHz, the internal amplifier was turned on and the highest resolution of 1001 points was selected. These settings provide the ability to obtain a snapshot of the power density that can be expected in an urban or semi-urban environment from continuously variable sources.

B. Results

After inputting the manufacturers' frequency-banded antenna factors into the spectrum analyzer, to ensure a fully calibrated system, the electric field strength measurements were taken. The input RF power density (S) is then calculated from the electric field strength measurement. Fig. 1 shows the input RF power density measured outside the Northfields London Underground station, where the spectral bands for DTV, GSM900, GSM1800, 3G, and Wi-Fi can been clearly identified.

A well-designed rectenna should ideally be capable of harvesting energy across an entire band, and thus it is important to calculate the total band power. The banded input RF power density S_{BA} (nW/cm^2) is calculated by summing all the spectral peaks across the band (i.e., in a similar way, the spectrum analyzer calculates channel power). These levels provide a snapshot of source availability at the time and location of the measurement. Moreover, they are used as a harvester design starting point since the power density at each band will define the input impedance of a rectenna.

The exact frequencies for each band are set by the U.K.'s official frequency band allocation [28]; the GSM900, GSM1800 and 3G base transmit (BTx) bands were separated from the associated mobile transmit (MTx) bands. Table I shows average

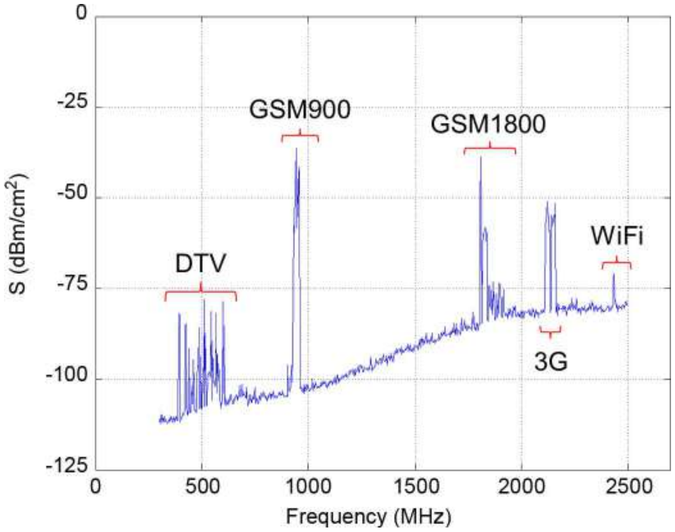

Fig. 1. Input RF power density measurements outside the Northfields London Underground station.

TABLE I
SUMMARY OF LONDON RF SURVEY MEASUREMENTS

Band	Frequencies (MHz)	Average S_{BA} (nW/cm^2)	Maximum S_{BA} (nW/cm^2)
DTV (during switch over)	470-610	0.89	460
GSM900 (MTx)	880-915	0.45	39
GSM900 (BTx)	925-960	36	1,930
GSM1800 (MTx)	1710-1785	0.5	20
GSM1800 (BTx)	1805-1880	84	6,390
3G (MTx)	1920-1980	0.46	66
3G (BTx)	2110-2170	12	240
WiFi	2400-2500	0.18	6

TABLE II
INPUT RF POWER DENSITY THRESHOLD: LONDON UNDERGROUND STATIONS WITHIN CENTRAL ZONES 1–3 (URBAN) AND OUTER ZONES 4–9 (SEMI-URBAN)

Band	S_{BA} *Threshold* (nW/cm^2)	Number of Stations	
		Urban	Semi-urban
DTV (during switch over)	40	10	0
GSM900	230	8	2
GSM1800	450	7	3
3G	62	6	4

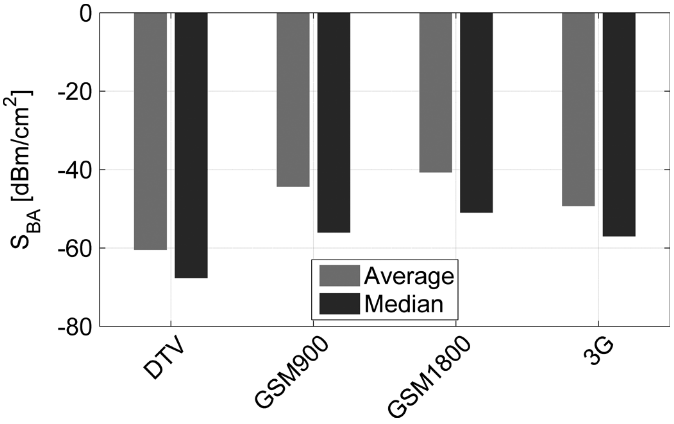

Fig. 2. Banded input RF power density measurements for the four largest ambient sources in Greater London.

RF power levels across all London Underground stations for the banded input RF power density measurements. It can be seen that all base-station transmit levels are between one and three orders of magnitude greater than the associated MTx levels. For this reason, and the fact that the population of transmitting mobile phones in close proximity of the harvester is highly variable, only base-station TXs will be considered further.

From our London RF survey, DTV, GSM900, GSM1800, 3G, and Wi-Fi were identified as potentially useful ambient RF energy harvesting sources, although DTV appears to be heavily dependent on line-of-sight and sudden changes in atmospheric conditions (e.g., temperature inversion) and Wi-Fi is very dependent on user traffic. It should be noted that the mobile phone base-station TXs employ vertically polarized antennas, placing a constraint on harvester orientation in deployment. With DTV, within the U.K., the main TXs have horizontally polarized antennas, while repeater TXs have vertically polarized antennas.

It is convenient to define the boundary between urban and semi-urban environments by the line that separates zones 3 and 4 on the London Underground map [29]. As one would expect, the central zones 1–3 host the highest density of base stations. As shown in Table II, a banded input RF power density threshold was selected to filter the ten London Underground stations with the highest measurements for each band. With DTV, the highest recorded measurements were all found within the urban environment. This is because Greater London's main DTV TX (at Crystal Palace) is located on the southeastern boarder of zones 3 and 4 and there are no London Underground stations further south. With mobile phones, more than 50% of the stations were inside the urban environment and those in a semi-urban environment were all located in close proximity to a cluster of base-station TXs.

Using the complete dataset from the London RF survey, Fig. 2 shows the average and median of the banded input RF power density measurements for the four largest ambient RF sources in Greater London. It can be seen that more than half of the locations have below average power levels. This is due to the fact that several stations had maximum levels that were considerably higher than the average because of their close proximity to TV TXs (e.g., Crystal Palace), extremes in base-station density and propagation characteristics.

In addition to the London RF survey, measurements within the Department of Electrical and Electronic Engineering building at ICL were taken on the 11th floor of the south stairwell. These are shown in Table III. As can be seen, DTV and GSM900 have a higher than average power level, due to a near line of sight from the TV TX and a close proximity to the 2G GSM900/1800 base stations.

The dataset from the London RF survey, with all relevant information (e.g., locations, timestamps, and banded input RF power density measurements), can be found at our interactive website: www.londonrfsurvey.org [30]. These measurements were used to design efficient harvesters and compared to ICL

TABLE III
MEASURED BANDED INPUT RF POWER DENSITIES AT ICL

	DTV	GSM900 (BTx)	GSM1800 (BTx)	3G (BTx)
S_{BA} [nW/cm^2]	18	48	50	3

measurements to identify locations in Greater London where the designed harvesters could operate. The design procedures and prototype test measurements will be presented in the following sections.

III. SINGLE-BAND AMBIENT RF ENERGY HARVESTERS

In order to implement efficient ambient RF energy harvesters, designed for the banded input RF power density levels measured at ICL, a set of single-band prototypes were realized and characterized; these will be compared to multiband array architectures in Section V.

A. Antenna Design and Measurements

Since our harvesters are intended to operate within a general (semi-)urban environment, where the exact location of the TX source is unknown, the rectennas' antennas need to be as close to omnidirectional as possible, avoiding the need for beam-pointing during deployment. This is at the obvious expense of limited antenna gain, and therefore, the corresponding levels of P_{RF} that the rectifier can receive. Conversely, if the location of the TX is known, then it may be tempting to use a high gain antenna, but this would require an appropriate level of beam-pointing and polarization matching that can be established and maintained.

Another requirement is that the antennas need to be easily scalable across all frequency bands since one important objective for this work is to compare and contrast different banded harvesters. Finally, the antennas need to be easily fabricated. For all these reasons, a linear polarized folded dipole was selected, although a monopole would also be suitable [31].

To simplify impedance matching between the antenna and rectifier, a modified folded dipole was used to obtain the required 50-Ω reference input impedance. A balun does not need to be employed, as there is no significant degradation in performance for this particular application, even with the use of an unbalanced microstrip rectifying circuit [32]. Furthermore, the antenna was not integrated onto a substrate, to give the additional freedom to embed the harvester on windows or within walls, furnishings, fixtures, or fittings. To this end, two different antennas were fabricated for each band; one made using a 560-μm diameter copper wire and the other with 75-μm-thick 25-mm-wide copper tape. The fabricated antennas are shown in Fig. 3. Since the copper tape was not rigid enough to retain its shape, it was placed on a Perspex substrate, to represent a flat panel.

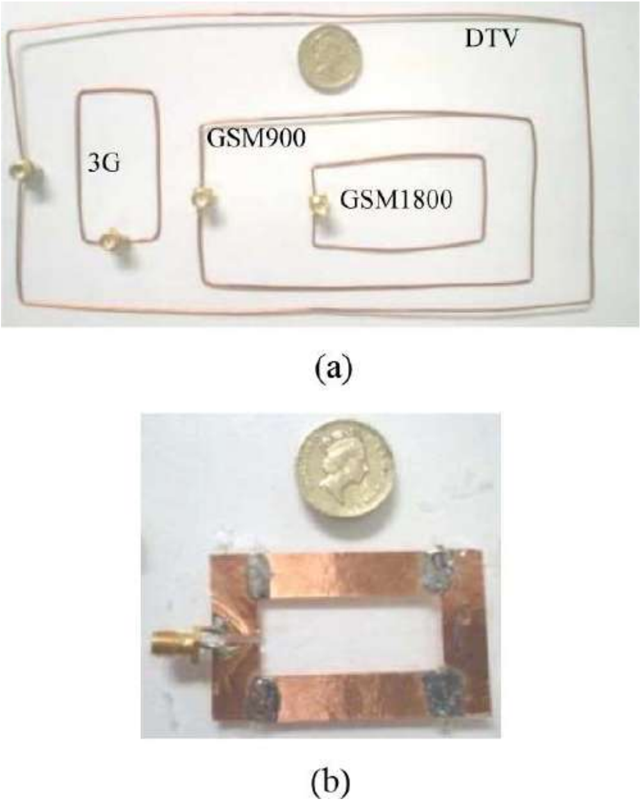

(a)

(b)

Fig. 3. 50-Ω folded-dipole antennas shown next to a British £1 coin. (a) DTV, GSM900 (BTx), GSM 1800 (BTx) and 3G (BTx) copper wire antennas. (b) 3G (BTx) copper tape antenna on Perspex.

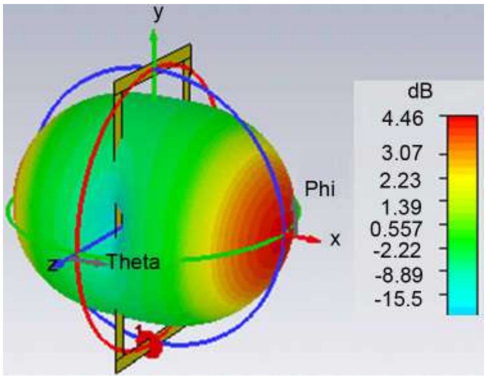

Fig. 4. Simulated beam profile for the DTV tape antenna.

To design the antennas, full-wave 3-D electromagnetic simulations were performed using CST Microwave Studio. As discussed previously, the antennas were designed to be as omnidirectional as possible, while covering as much of the ambient RF source BW as possible. Fig. 4 shows the typical simulated gain profile for the DTV tape antenna, having a front-to-back ratio of unity. Table IV shows the simulated gain and the 10-dB return-loss fractional BW for the optimized copper wire and tape antennas.

Fig. 5 shows excellent agreement between predicted and measured return-loss results, within a 10-dB return-loss bandwidth, for the eight fabricated single-band antennas. As one would expect with such a simple classical antenna, the out-of-band performances (not shown) were also in good agreement. It was found that better return-loss measurements are achieved with our single-band antennas when compared to other reported single-band omnidirectional [13] and multiband [33] designs. The latter may be important, as it may be tempting to implement a more compact multiband rectenna, but which

TABLE IV
SIMULATED GAIN AND 10-dB RETURN LOSS FRACTIONAL BW FOR FOLDED-DIPOLE SINGLE-BAND ANTENNAS

Band		Wire		Tape	
	BW (%)	Gain (dBi)	BW (%)	Gain (dBi)	BW (%)
DTV	26	4.35	4	4.48	6
GSM900	3.7	4.42	4.3	4.73	4.3
GSM1800	4.1	4.32	5.3	4.73	10.7
3G	2.8	4.39	5.4	4.76	12

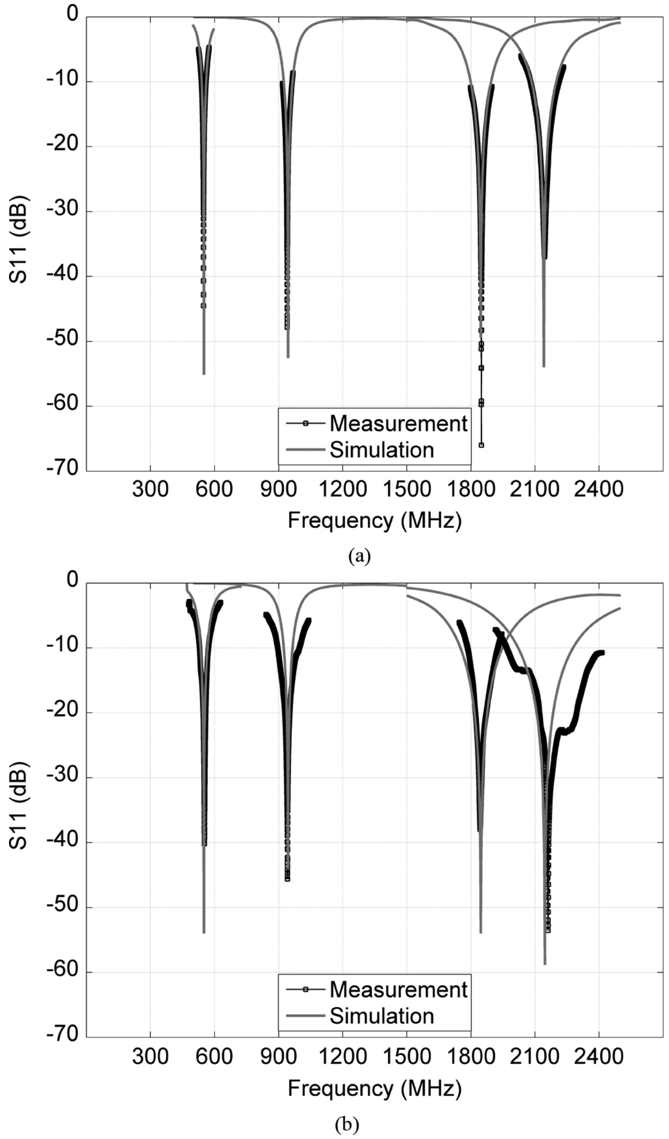

(a)

(b)

Fig. 5. Input return-loss predictions and measurements for all single-band folded-dipole antennas. (a) Wire. (b) Tape.

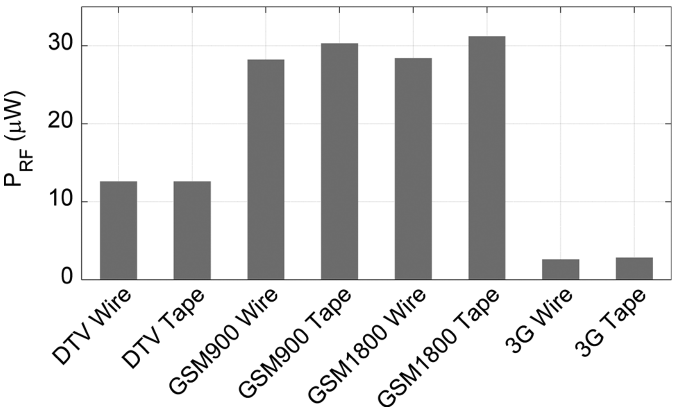

Fig. 6. Predicted input RF power levels for the four largest ambient sources at the ICL testing location.

may ultimately not give better ambient RF energy harvesting performance.

Obtaining a minimum acceptable return loss over an antenna fractional bandwidth as large as that of the source is key to harvest as much input RF energy as possible. As can be seen in Table IV, where the fractional bandwidth is defined for a 10-dB return loss, our antennas have a fractional bandwidth greater than those of the sources, with the exception of DTV, which only covers approximately 35% of the target frequency range. In other work [13], 5-dB return-loss fractional bandwidth is adopted for RF energy harvesting applications.

The fractional bandwidth of the antennas having a minimum return loss of 5 dB is too great to assume a constant antenna gain over the whole band. Therefore, an additional advantage of using 10-dB return-loss fraction bandwidths is that (1) can be used to calculate the input RF power with the assumption that the midband antenna gain is constant with frequency [34]. Therefore, the time-averaged input RF power P_{RF} is given by

$$P_{\mathrm{RF}} = S_{BA} \cdot A_{\mathrm{real}} \text{ and } A_{\mathrm{real}} \approx G(f_o)\frac{\lambda_o^2}{4\pi} \qquad (1)$$

where A_{real} is the real aperture (or capture area) of the antenna, λ_o is the free-space wavelength at the midband frequency f_o, and $G(f_o)$ is the rectenna's antenna gain at f_o.

Substituting the measured banded input RF power densities recorded in Table III and the predicted midband antenna gains in Table IV into (1), realistic values for P_{RF} can be calculated for all four bands, with the results shown in Fig. 6. It can be seen that with all antenna gains being in the region of $\sim$ 4.5 dBi, the 2G GSM900/1800 harvesters will generate the highest input RF power levels, due to the high-banded input RF power density levels measured *in situ*. At the other extreme, the 3G harvesters will be the worst performers. As only a small fraction of the required frequency range is covered by the DTV antennas, the predicted values for P_{RF} represent an overestimation.

B. Rectifier Design and Measurements

Based on a previously reported analysis [35] and the predicted input RF power levels presented in Fig. 6, the zero-bias SMS7630 diode (in a series configuration) was selected as the optimal solution for our ambient RF energy harvesters, as shown in Fig. 7. In a series configuration, the junction capacitance $(C_j(V))$ of the diode dominates the detector's impedance, as long as $C_{\mathrm{out}} > C_j(V)$, and thus C_{out} has little or no effect on the matching circuit. This allows C_{out} to be large enough to provide a ripple-free output voltage. In contrast, C_{out} must be less than 1 pF to achieve a good impedance match with a shunt configuration as C_{out} appears in parallel with $C_j(V)$ and the packaging parasitic capacitance. However, C_{out} is too small to

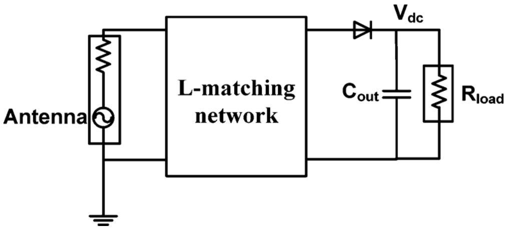

Fig. 7. Series detector configuration with L-matching network.

provide a ripple-free dc voltage to the load. This can be overcome with a matching network that will allow a good impedance match with a large output capacitor, but at the expense of introducing losses. Furthermore, as shown in [36], these issues start to become negligible with a shunt configuration as the shunt diode becomes more self-biased as the input power increases. Simulations were performed using Agilent Technologies' ADS software. Its Momentum package not only takes into account the losses from the low-cost FR4 substrate, but also calculates fringing fields, effects from which are passed on to the harmonic-balance package for simulating the nonlinear behavior of the rectifier.

A good impedance match was achieved by employing a simple matching network; a series lumped-element inductor was used to absorb part of the capacitive reactance from the series diode and an additional quarter-wavelength short-circuit shunt stub was employed to achieve the desired 50-Ω impedance [37]. Since the impedance of the diode varies with frequency and input RF power, impedance matching between the antenna and rectifier was first undertaken by finding the optimal output load resistance for an input RF power level of -20 dBm with a single-tone source at the midband frequency. After the optimal load was found, further broadband optimization was performed to the matching network and the load to ensure good impedance matching throughout the target frequency range and the measured $P_{\rm RF}$ for each band.

As with the antenna analysis, and unlike conventional RF circuits that adopt the more traditional half-power bandwidth definition, the rectifier should adopt the 10-dB-input return-loss bandwidth. The reason for this is that, for ambient RF energy harvesting applications, the input RF power is at a premium and so what little energy is available should not be wasted by being reflected back from avoidable impedance mismatches at either the antenna or rectifier.

Fig. 8 shows excellent agreement between predicted and measured input return loss results, within the -10-dB bandwidth, for the DTV and GSM900 rectifiers, having fractional bandwidths of 5.7% (below target) and 4.8%, respectively. With these lower frequency designs, the fundamental and higher order harmonics were below -55 dBm, ensuring a clean dc voltage at the load, without the need for any output filtering. Reasonable agreement was found with the GSM1800 and 3G rectifiers, having fractional bandwidths of 1.6% (below target) and 7.4%, respectively. It was found that with these two higher frequency designs, the higher order harmonics were -40 dBm at the output. This reduced performance, as illustrated in Fig. 8, is due to the higher series inductive reactance leads of the output shunt storage capacitor. For this reason, a second prototype version (v2) was designed for the 3G rectifier, using distributed-element components for the input impedance matching stage and an additional output filter stage. With the lumped-element matching network, it was found that in order to achieve good input impedance matching to 50 Ω, all the microstrip transmission lines had to have a characteristic impedance of 92 Ω. With the distributed-element matching network, a simple shunt quarter-wavelength open-circuit stub, designed for operation at the fundamental frequency, was employed. A microstrip line was added between the cathode of the diode and the stub to absorb the capacitive reactance of the diode. The stub effectively filters to < -50 dBm the higher order harmonics. The microstrip design can be seen in Fig. 13.

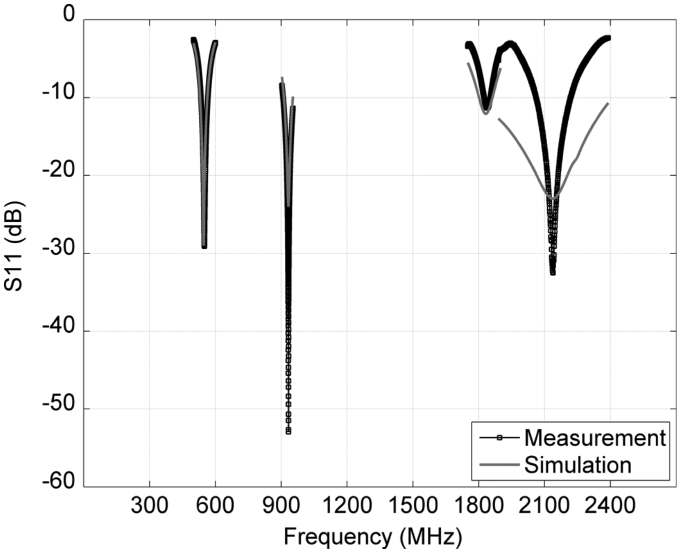

Fig. 8. Input return-loss predictions and measurements for all first prototype single-band rectifiers, with $P_{\rm RF} = -20$ dBm at the input and optimal load resistances at the output.

C. PMM

Since the input RF power from ambient sources can be represented as a multi-tone source, with power levels fluctuating across the target frequency range, the output impedance of the rectifier is time varying. A power management module (PMM) capable of performing maximum power point tracking (MPPT) is required.

For our work, a low-power integrated-circuit PMM from Texas Instruments Incorporated (BQ25504) was selected, due to its low quiescent current (< 330 nA) and low input voltage operation (~ 80 mV hot-start and 330-mV cold-start) [38]. It is worth noting that its startup voltage is lower than PMMs previously reported and realized using hybrid circuits for RF energy harvesting [13]. The BQ25504 PMM includes a boost converter that steps up its input voltage (having a 350-mV average value during ambient operation) to useful levels between 2.4–5.3 V. The BQ25504 also has a built-in battery management module, which is used to control the duty cycle of the output power to the load.

MPPT operation on the BQ25504 is achieved by periodically sampling the open-circuit voltage (OCV) at the input of the converter, which then draws a current causing the converter

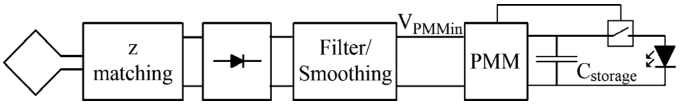

Fig. 9. System block diagram.

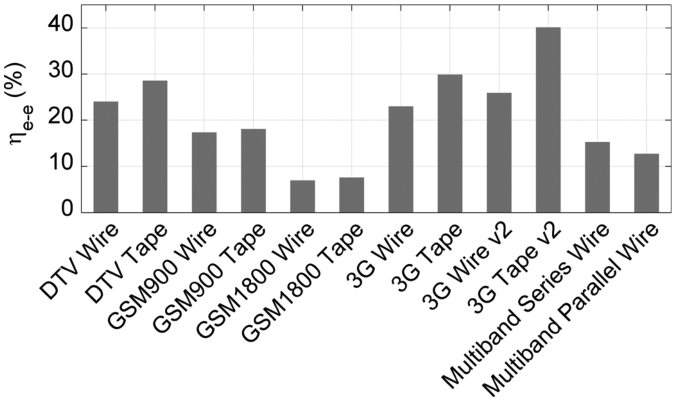

Fig. 10. End-to-end efficiencies for ambient RF energy harvesting at ICL.

input voltage to fall and be held at a pre-programmed fraction of the OCV (set by a potential divider). In a simple dc circuit, with a resistive source impedance, the optimal ratio is 0.5. For the rectenna-based system, a ratio of 0.48–0.53 was found to maximize the power output of the system. The BQ25504 is designed to charge a storage element, and in this case, a capacitor C_{storage} was used. The programmed PMM continuously charges the storage capacitor, and the load (a low-power LED) was automatically connected to the storage capacitor when the capacitor voltage reaches an upper limit $V_{\mathrm{high}} = 2.84$ V and automatically disconnected when it reaches a lower limit $V_{\mathrm{low}} = 2.40$ V. The duty cycle of the LED can then be used to calculate the efficiency of the system, as will now be described. A diagram of the system is shown in Fig. 9.

IV. END-TO-END EFFICIENCY ANALYSIS

The efficiency η of an RF energy harvesting system is

$$\eta = \frac{P_{\mathrm{dc}}}{P_{\mathrm{RF}}} \tag{2}$$

where P_{dc} is the time-averaged output (i.e., equivalent dc) power into the storage element (e.g., battery or supercapacitor) and load and P_{RF} is as previously defined. Measurements for this type of system are usually performed in a controlled environment (e.g., an anechoic chamber or TEM cell), using a dedicated constant or variable amplitude single-tone RF signal source [32], [39]. However, the former is not suitable for evaluating ambient RF energy harvesting operation, which has a much broader spectrum of nonconstant input frequencies and where the instantaneous input RF power is time variant. The use of a constant single-tone dedicated source provides a convenient stable reference power to the harvester; while the latter reflects a more realistic signal source having fluctuating power levels across a nonzero bandwidth, multipath, and reflection effects which are very difficult to emulate in a controlled environment.

Therefore, to determine the overall end-to-end efficiency $\eta_{\mathrm{e\text{-}e}}$ for a complete ambient RF energy harvester, the input RF energy U_{RF} was calculated based on the harvester's antenna characteristics and the actual banded input RF power density measurements taken at the time of harvester operation, using the Agilent Fieldfox and the calibrated antenna. It is important to note that since the impedance mismatch between the antenna and detector is not taken into account, U_{RF} is higher than expected, providing an underestimate of end-to-end efficiency. The output dc energy U_{dc} was then calculated by measuring the charge–discharge cycle time, t_{cycle} of the storage capacitor between V_{high} and V_{low}, as the LED is repeatedly connected and disconnected. The output dc energy equation is already taking into account the efficiency of the PMM given the fact that the measurements are taken at its output voltage. The end-to-end efficiency of one charge–discharge cycle of C_{storage} is

$$\eta_{\mathrm{e\text{-}e}} = \frac{U_{\mathrm{dc}}}{U_{\mathrm{RF}}} \tag{3}$$

where the input RF energy is given by integrating the time-averaged input RF power over a cycle time, as

$$U_{\mathrm{RF}} = \int_{0}^{t_{\mathrm{cycle}}} P_{\mathrm{RF}} dt \tag{4}$$

and the output dc energy is given by the energy supplied to the load, as follows:

$$U_{\mathrm{dc}} = C_{\mathrm{storage}} \frac{({V_{\mathrm{high}}}^2 - V_{\mathrm{low}}^2)}{2}. \tag{5}$$

A. ICL Field Trials

Four single-band ambient RF energy harvesters were assembled by connecting the rectifiers to the wire/tape antennas and PMMs programmed for the optimal load. A 100-μF shunt capacitor was employed as the storage element, providing $U_{\mathrm{dc}} = 115$ μJ. Our system is capable of cold-starting the boost converter and MPPT since the rectenna is capable of providing the minimum starting voltage of 330 mV. When the minimum voltage is reached, the boost converter and MPPT start to operate and the charge–discharge cycle at the load begins, causing the LED to flash. During field trials, t_{cycle} took up to 170 s for the harvester with lowest banded input RF power density, corresponding to 3G with the wire antenna. Table V summarizes the results where t_c and t_d are the charge and discharge times, respectively, and ΣV and ΣI are the multiband voltage and current summing array architectures, respectively. A detailed discussion on the multiband rectenna arrays will be presented in the following section. The end-to-end efficiency was calculated using (3) with data from Fig. 6 and measuring the charge–discharge cycle time during harvesting operation.

Fig. 10 shows the overall end-to-end efficiencies for all the harvester demonstrators, deployed and tested at ICL. As predicted by simulations, the improved 3G v2 demonstrator with

TABLE V
HARVESTERS CHARGE AND DISCHARGE TIMES (t_c, t_d, RESPECTIVELY) FOR A SPECIFIED LOAD

Band	Wire					Tape				
	t_c (s) load independent	t_d (s) load dependant	t_{cycle} (s) load dependant	P_{dc} (t_d) (μW)	P_{dc} (t_{cycle}) (μW)	t_c (s) load independent	t_d (s) load dependant	t_{cycle} (s) load dependant	P_{dc}(t_d) (μW)	P_{dc} (t_{cycle}) (μW)
DTV	26	12	38	9.6	3	14	18	32	8.2	3.6
GSM900	14	10	24	11.5	4.8	8	13	21	14.4	5.5
GSM1800	43	15	58	7.7	2	22	27	49	5.2	2.4
3G v2	167	3	170	38.4	0.7	96	5	101	1.2	1.1
Multiband ΣV	43	7	50	66	2.3	-	-	-	-	-
Multiband ΣI	55	5	60	92.2	2	-	-	-	-	-

tape antenna outperformed its original design by 11%; achieving an end-to-end efficiency of 40% with an input RF power of only −25.4 dBm.

It is believed that a much greater efficiency can be achieved for the DTV harvester if the fractional bandwidths for the first prototype circuits (i.e., 4.4/4.5% for the antennas and 5.8% for the rectifier) could be increased to match the much greater target value of 26%. Likewise, the reduced efficiency of the GSM1800 harvesters can be attributed to the detrimental effects of the narrowband input impedance matching of the rectifier (i.e., having a fractional bandwidth of only 1.6%, when compared to its target value of 4.1%). Finally, with all the harvesters, the end-to-end efficiencies can be enhanced through better antenna design and optimal polarization matching.

Table VI, shows the number of locations from the London RF survey that would be able to support our harvesters. Unlike the single-band 3G harvester, which can operate at 45% of the locations, our DTV harvester can only be used at two locations (one in zone 2 and the other in zone 3). Therefore, for the general deployment of an ambient RF energy harvester within an (semi-)urban environment, at street level, the single-band DTV harvester may not be practical

V. ARRAY ARCHITECTURES

Since ambient input RF power levels can be low (i.e., below −25 dBm) and dependent on both time and spatial considerations, harvesters could be designed to extract energy with spatial-diversity within the same frequency band or using different frequency bands. For example, with the former, at a particular location there may be only one band that has significant levels of RF energy worth harvesting. In this case, spatial-diversity array architectures may provide more usable output power. Alternatively, with the latter, multiband array architectures may provide more robust operation.

With both forms of parallel array architecture (i.e., spatial-diversity and multiband), a further classification can be seen through the use of either diversity/band switching or a summing node. With the former, physical switches automatically select whichever signal path delivers the highest input RF power level; with the latter, power from all signals is combined. Fig. 11 illustrates generic forms of parallel array architectures, showing that switching/summing can be performed electromagnetically at a single antenna or at the output from multiple antennas, rectifiers, or PMMs.

TABLE VI
NUMBER OF LOCATIONS FROM THE LONDON RF SURVEY CAPABLE OF SUPPORTING IDENTICAL HARVESTERS AT THE SAME EFFICIENCY LEVELS

	DTV	GSM900 (BTx)	GSM1800 (BTx)	3G (BTx)
Stations with higher S_{BA}	2	28	68	122

Multiband array architectures, similar to those in Fig. 11(c) and (d), capable of RF harvesting from the four previously identified bands, were selected as possible optimal solutions, given no size/cost constraints. Our objectives were to reach the minimum cold-start voltage at the lowest possible input RF power levels and increase the harvesters' operational capabilities within (semi-)urban environments.

To this end, two different multiple rectenna architectures were investigated. The first with a single shared PMM and the second with multiple PMMs, as illustrated in Fig. 12. To simplify assembly, the wire antennas were selected since they did not require a substrate. Unwanted coupling between the single-band antennas was minimized by placing them a minimum distance of $\lambda_L/5$ apart; where λ_L is the wavelength of the lowest frequency band antenna [40]. For example, the DTV and the 3G antennas were kept at least 11 cm apart, as shown in Fig. 13. This allowed S_{11} measurements to be the same as in Fig. 5 once all antennas were assembled into the array.

A. Multiple Rectennas With a Shared PMM

In order to improve the cold-start performance of the system, the outputs of multiple rectennas can be connected in series, as shown in Fig. 12(a). This increases the probability of the voltage on the input of the PMM reaching the cold-start level (330 mV for the BQ25504) under any given scenario. While cold-starting the PMM, each rectenna harvests (albeit not optimally). Once

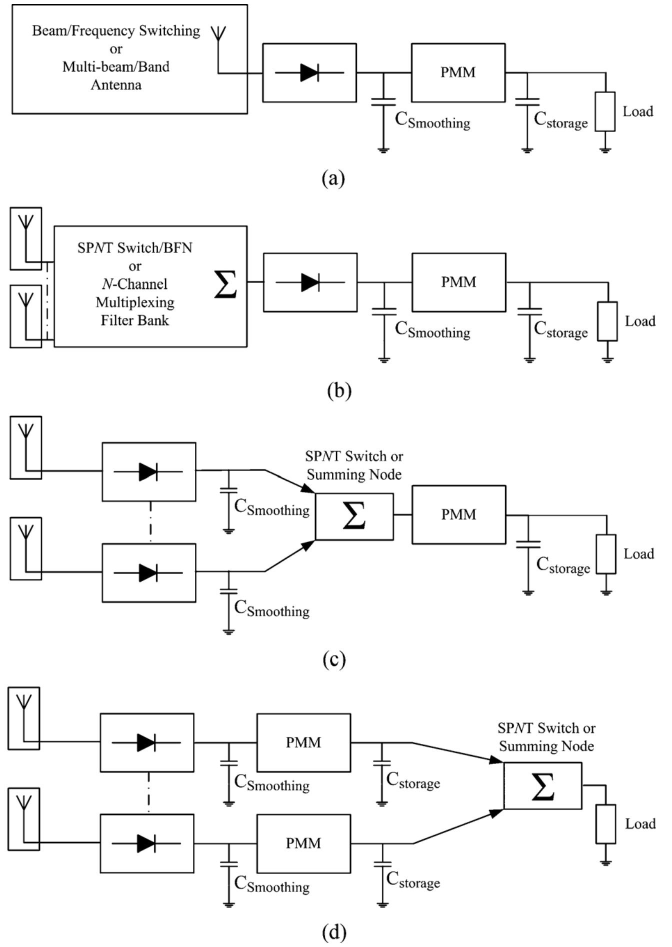
Beam/Frequency Switching
or
Multi-beam/Band
Antenna
PMM
CSmoothing
Cstorage
Load
(a)
SPNT Switch/BFN
or
N-Channel
Multiplexing
Filter Bank
Σ
PMM
CSmoothing
Cstorage
Load
(b)
SPNT Switch or
Summing Node
Σ
PMM
CSmoothing
CSmoothing
Cstorage
Load
(c)
PMM
PMM
CSmoothing
CSmoothing
Cstorage
Cstorage
SPNT Switch or
Summing Node
Σ
Load
(d)

Fig. 11. Parallel array architectures with switching/summing at the: (a) antenna, (b) output of multiple antennas, (c) output of multiple rectifiers, and (d) output of multiple PMMs.

the PMM circuit starts, with the MPPT operating, the harvested power level increases.

The behavior of the series rectenna topology with a shared PMM requires some discussion. As the output impedance and the OCV for each rectenna is different, since they operate at different frequencies and input RF power levels, the rectennas are forced to share the same output current in a series configuration, which does not allow them all to operate at their individual maximum power points. This causes the voltage on each rectenna output, except the one having the highest input RF power, to collapse. This operation is analogous to the partial shading problem with a series string of solar panels [41] sharing a common boost converter. With this photovoltaic system, bypass diodes placed around individual cells stop the poorly lit cells contributing a negative voltage (and power) to the string. In our case, the series circuit formed by the loop antennas and rectifying diodes performs the same task. This means that while all rectennas contribute to system startup, only the rectenna with the highest input RF power contributes significant power for continuous operation once the PMM starts. Fig. 10 shows the end-to-end efficiency for the voltage summing multiband harvester array when tested at ICL. An efficiency of only 15% was achieved with a combined input RF power of -12 dBm. The lower efficiency, when compared to a single-band harvester, is due to the imbalance of rectifier outputs, as discussed above. Here, the charge time was 43 s, compared to 167 s with the lowest contributing single-band 3G harvester with wire antenna.

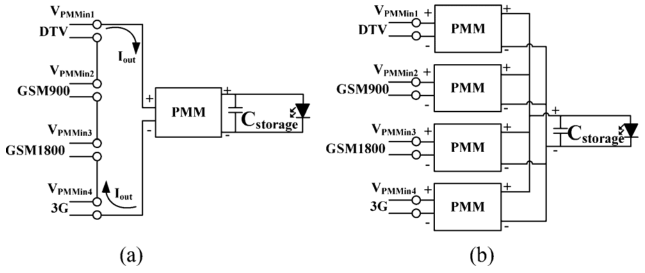
VPMMin1
DTV
Iout
VPMMin2
GSM900
VPMMin3
GSM1800
VPMMin4
3G
Iout
PMM
Cstorage
(a)
VPMMin1
DTV
PMM
VPMMin2
GSM900
PMM
VPMMin3
GSM1800
PMM
VPMMin4
3G
PMM
Cstorage
(b)

Fig. 12. Multiband array architectures (with $N = 4$ bands). (a) Voltage summing at the outputs of the single-band rectennas. (b) Current summing at the outputs of the single-band harvesters.

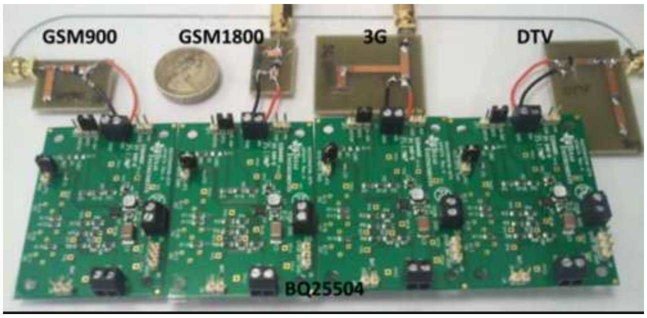
GSM900
GSM1800
3G
DTV
BQ25504

Fig. 13. Rectenna array architecture with individual PMMs for the four largest contributors with wire antennas.

B. Multiple Rectennas With Individual PMMs

In order to overcome the balancing issues when multiple rectennas share a common PMM, as discussed previously, each rectenna can have its own PMM, whose outputs can be connected to a common storage element (in this case, a 400-μF shunt capacitor, providing $U_{\mathrm{dc}} = 461\ \mu$J), as illustrated in Fig. 12(b) and shown in Fig. 13. Although not achieving cold-start as quickly as the series topology, this parallel topology has the advantage of being able to run each rectenna at its maximum power point. In addition, once one rectenna is able to harvest enough energy for a cold-start, all PMMs will start because they share a common storage element, allowing the rectennas with low-input RF power levels to harvest at levels below which they could not do so independently.

This parallel topology was tested and found to be capable of operating in many locations where the series array was unable to operate; e.g., if only one of the bands had $P_{\mathrm{in}} > -25$ dBm. As expected, the largest contributor hot-started the other PMMs, allowing them to harvest at an input RF power level down to -29 dBm.

However, as with the previous results for voltage summing, having a combined input RF power of -12 dBm, the efficiency using multiple PMMs is slightly lower, at 13%, as shown in Fig. 10. This is because useful dc output power from the cold-

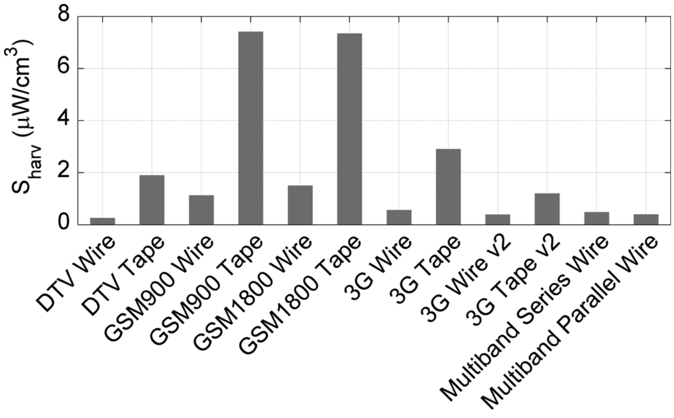

Fig. 14. Output dc power density for all harvesters at ICL.

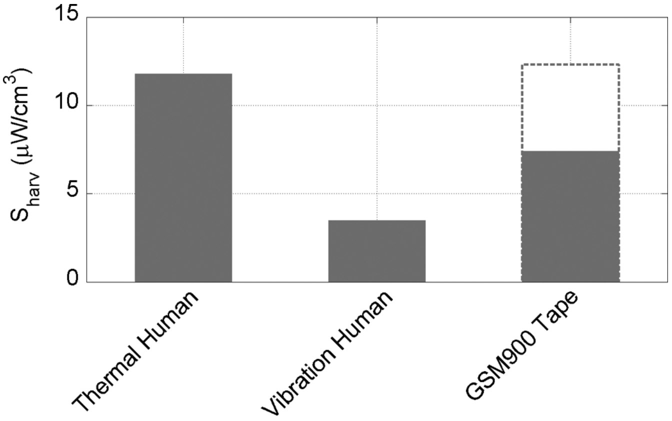

Fig. 15. Output dc power density comparison for alternative ambient harvesting technologies [40]–[43] against the best current generation of RF harvesters at ICL.

starting harvester is being supplied to the other harvesters for hot-starting, even though some of them may not actually be contributing any of their own harvested power.

VI. OUTPUT DC POWER DENSITY COMPARISON

The volumetric output dc power density S_{harv} (μW/cm^3) represents an important figure of merit for comparing alternative energy harvesting technologies. For ambient RF energy harvesting, the output dc power is calculated by multiplying the effective input RF power by the overall end-to-end-efficiency. The total volume (including that of the antenna, rectifier, and PMM; not including energy storage, as this does not directly affect the dc power output) must be determined. It is important to note that the volume for the antenna could effectively disappear if it is assembled onto a window or within a wall, furnishing, fixture, or fitting. Moreover, the required PMM printed circuit board (PCB) size used throughout these calculations was assumed to be ten times the size of the BQ25504 chip, to account for any necessary additional components.

Fig. 14 shows the output dc power density for all the harvesters demonstrated here. It can be seen that the 2G GSM900/1800 harvesters with tape antennas both have the highest value of $S_{\mathrm{harv}} = 7.4$ (μW/cm^3), when tested at ICL, due to the high-banded input RF power density S_{BA}. The value for the most efficient harvester (i.e., 3G v2 with tape antenna) was not the highest in terms of output RF power because S_{BA} in this band was more than an order of magnitude lower than with GSM900.

S_{harv} allows a direct and meaningful comparison to be made with other alternative energy harvesting technologies. Our best performing ambient RF energy harvester (i.e., GSM900 with tape antenna) was compared against alternative energy harvesting technologies, assuming they used the same PMM board size [42]–[44].

It can be seen in Fig. 15 that ambient RF energy harvesting has a low output dc power density when compared to alternative energy harvesting technologies, but only when the total volume of the first prototype demonstrator is considered. However, when the antenna is absorbed onto or into a background feature and when the PMM is fully integrated into the rectifier, it can outperform (as indicated by the dotted column) the alternative energy harvesting technologies, while providing a complimentary means of extracting energy from the environment. The RF harvesters, however, have the additional advantage in that they do not require a thermal gradient, and unlike vibration-driven devices, they have no moving parts.

VII. CONCLUSIONS

Our objectives were to reach the lowest possible ambient input RF power levels and extend the harvesters' operational capabilities within (semi-)urban environments. To this end, a comprehensive citywide RF spectral survey was undertaken, indicating that more than 50% of the 270 London Underground stations are suitable locations for the deployment of our ambient RF energy harvesters. It has been demonstrated that single-band harvesters can operate with efficiencies of up to 40% in a (semi-)urban environment, and can start to operate from power levels as low as -25 dBm.

To increase the freedom of operation, multiband array architectures were investigated. With the current summing harvester arrays, RF harvesting was achieved at an input RF power level as low as -29 dBm, without any external dc power supply to hot-start the PMM. Limitations on the multiband array architectures were discussed, highlighting the need for further work in balancing rectennas with voltage summing rectenna arrays when operating at lower input RF power levels.

Finally an output dc power density comparison against alternative energy harvesting technologies has shown that RF harvesting can represent a competitive solution within (semi-)urban environments, especially when the antenna can be absorbed into background features.

ACKNOWLEDGMENT

The authors would like to thank the members of the London RF survey team and F. Kusidlo, Agilent EEsof EDA, for his advice regarding simulations.

REFERENCES

[1] "Demonstration of microwave power transmission," *Pract. Wireless*, pp. 835–835, 1965.

[2] T. Salter, G. Metze, and N. Goldsman, "Parasitic aware optimization of an RF power scavenging circuit with applications to smartdust sensor networks," in *IEEE Radio Wireless Symp.*, 2009, pp. 332–335.

[3] "SENSIMED Triggerfish®—Continuous IOP monitoring," SENSIMED, Lausanne, Switzerland, Mar. 2012. [Online]. Available: http://www.sensimed.com/S-Trig-glaucoma.htm, , "

[4] M. Pinuela, D. C. Yates, S. Lucyszyn, and P. D. Mitcheson, "Maximizing DC-to-load efficiency for inductive power transfer," *IEEE Trans. Power Electron.*, vol. 28, no. 5, pp. 2437–2447, May 2013.

[5] W. Joseph, G. Vermeeren, L. Verloock, M. M. Heredia, and L. Martens, "Characterization of personal RF electromagnetic field exposure and actual absorption for the general public," *Health Phys.*, vol. 95, no. 3, pp. 317–330, Sep. 2008.

[6] G. Thuroczy, F. Molnar, J. Szabo, G. Janossy, N. Nagy, G. Kubinyi, and J. Bakos, "Public exposure to RF from installed sources: Site measurements and personal exposimetry," in *1st Eur. Antennas Propag. Conf.*, 2006, pp. 1–4.

[7] P. Elliott, M. B. Toledano, J. Bennett, L. Beale, K. de Hoogh, N. Best, and D. J. Briggs, "Mobile phone base stations and early childhood cancers: Case-control study," *Br. Med. J.*, vol. 340, no. 1, pp. 1–7, Jun. 2010.

[8] N. Shinohara, "Power without wires," *IEEE Microw. Mag.*, vol. 12, no. 7, pp. S64–S73, Dec. 2011.

[9] J. Y. Park, S. M. Han, and T. Itoh, "A rectenna design with harmonic-rejecting circular-sector antenna," *IEEE Antennas Wireless Propag. Lett.* , vol. 3, pp. 52–54, 2004.

[10] "Dengyo rectennas," Nihon Dengyo Kosaku Company Ltd., Tokyo, Japan, Jul. 2012. [Online]. Available: http://www.den-gyo.com/solution/solution01.html

[11] T. Le, K. Mayaram, and T. Fiez, "Efficient far-field radio frequency energy harvesting for passively powered sensor networks," *IEEE J. Solid-State Circuits*, vol. 43, no. 5, pp. 1287–1302, May 2008.

[12] V. Marian, B. Allard, C. Vollaire, and J. Verdier, "Strategy for microwave energy harvesting from ambient field or a feeding source," *IEEE Trans. Power Electron.*, vol. 27, no. 11, pp. 4481–4491, Nov. 2012.

[13] C. Mikeka, H. Arai, A. Georgiadis, and A. Collado, "Dtv band micropower RF energy-harvesting circuit architecture and performance analysis," in *IEEE Int. RFID-Technol. Appl. Conf.*, 2011, pp. 561–567.

[14] S. Kitazawa, H. Ban, and K. Kobayashi, "Energy harvesting from ambient RF sources," in *IEEE MTT-S Int. Microw. Symp. Workshop*, 2012, pp. 39–42, Series on Innovative Wireless Power Transmission: Technol., Syst., Appl..

[15] R. Vyas, V. Lakafosis, and M. Tentzeris, "Wireless remote localization system utilizing ambient RF/solar power scavenging RFID tags," in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 23–28, 2010, pp. 1764–1767.

[16] H. Nishimoto, Y. Kawahara, and T. Asami, "Prototype implementation of ambient RF energy harvesting wireless sensor networks," *IEEE Sens. J.*, pp. 1282–1287, Nov. 2010.

[17] A. Sample and J. R. Smith, "Experimental results with two wireless power transfer systems," in *IEEE Radio Wireless Symp.*, 2009, pp. 16–18.

[18] V. Nitu, G. Lojewski, and S. Nitu, "Electromagnetic field evaluation on an antennas shared site, EUROCON 2009," in *IEEE EUROCON'09*, 2009, pp. 70–75.

[19] Y. Kawahara, K. Tsukada, and T. Asami, "Feasibility and potential application of power scavenging from environmental RF signals," in *IEEE AP-S Int. Symp.*, 2009, pp. 1–4.

[20] H. J. Visser, A. C. F. Reniers, and J. A. C. Theeuwes, "Ambient RF energy scavenging: GSM and WLAN power density measurements," in *38th Eur. Microw. Conf.*, 2008, pp. 721–724.

[21] T. G. Cooper, S. M. Mann, M. Khalid, and R. P. Blackwell, "Public exposure to radio waves near GSM microcell and picocell base stations," *J. Radiol. Protection*, vol. 26, no. 2, pp. 199–211, 2006.

[22] "FieldFox RF Analyser N9912A 4/6 GHz," Agilent Technol., Santa Clara, CA, USA, Tech. Overview, 2012.

[23] "Biconical EMC broadband antennas—BicoLOG Series," Aaronia, Strickscheid, Germany, Data Sheet, 2012.

[24] "Digital TV switchover in the UK," Digital UK Ltd., London, U.K., 2012.

[25] "International commission on non-ionizing radiation protection, guidelines for limiting exposure to time- varying electric, magnetic, and electromagnetic fields (up to 300 GHz)," *Health Phys.*, vol. 74, no. 4, pp. 494–522, 1998.

[26] M. Riederer, "EMF exposure due to GSM base stations: measurements and limits," in *IEEE Int. Electromagn. Compat. Symp.*, 2003, vol. 1, pp. 402–405.

[27] "Revised ECC recommendation (02)04: Measuring non-ionising electromagnetic radiation (9 kHz—300 GHz)," ECC within the Eur. CEPT, Oct. 2003.

[28] Nat. Freq. Planning Group, "United Kingdom frequency allocation table, issue no. 16," 2010, Cabinet Official Committee on U.K. Spectrum Strategy.

[29] Transport for London, "London underground map," Aug. 2012. [Online]. Available: http://www.tfl.gov.uk/gettingaround/1106.aspx

[30] Imperial College London, "London RF Survey," Aug. 2012. [Online]. Available: http://londonrfsurvey.org/

[31] R. Vyas, V. Lakafosis, M. Tentzeris, H. Nishimoto, and Y. Kawahara, "A battery-less, wireless mote for scavenging wireless power at UHF (470–570 MHz) frequencies," in *IEEE AP-S Int. Symp./USNC/URSI Nat. Radio Sci. Meeting*, Jul. 3–8, 2011, pp. 1069–72.

[32] T. Ungan and L. M. Reindl, "Harvesting low ambient RF-sources for autonomous measurement systems," in *Proc. IEEE Instrum. Meas. Technol. Conf.*, 2008, pp. 62–65.

[33] A. Costanzo, A. Romani, D. Masotti, N. Arbizzani, and V. Rizzoli, "RF/baseband co-design of switching receivers for multiband microwave energy harvesting," *Sensors Actuators A, Phys.*, vol. 179, pp. 158–168, Jun. 2012.

[34] J. L. Volakis, *Antenna Engineering Handbook*. New York, NY, USA: McGraw-Hill, 2007.

[35] M. Pinuela, P. D. Mitcheson, and S. Lucyszyn, "Analysis of scalable rectenna configurations for harvesting high frequency ambient radiation," in *Proc. Power MEMS*, Leuven, Belgium, 2010, pp. 41–44.

[36] J. O. McSpadden, L. Fan, and K. Chang, "Design and experiments of a high-conversion-efficiency 5.8-GHz rectenna," *EEE Trans. Microw. Theory Techn.*, vol. 46, no. 12, pp. 2053–2060, Dec. 1998.

[37] "HSMS-285x series datasheet," Avago Technol., San Jose, CA, USA, 2009.

[38] "Ultra low power boost converter with battery management for energy harvester applications," Texas Instruments Incorporated, Dallas, TX, USA, BQ25540 Datasheet, 2011.

[39] A. Nimo, D. Grgic, and L. M. Reindl, "Ambient electromagnetic wireless energy harvesting using multiband planar antenna," in *IEEE 9th Int. Syst., Signals, Devices Multi-Conf.*, Mar. 20–23, 2012, pp. 1, 6.

[40] H. Takhedmit, L. Cirio, B. Merabet, B. Allard, F. Costa, C. Vollaire, and O. Picon, "A 2.45-GHz dual-diode rectenna and rectenna arrays for wireless remote supply applications," *Int. J. Microw. Wireless Technol.*, vol. 3, no. 3, pp. 251–258, Jun. 2011.

[41] S. Vemuru, P. Singh, and M. Niamat, "Modeling impact of bypass diodes on photovoltaic cell performance under partial shading," in *IEEE Int. Electro/Information Technol. Conf.*, 2012, pp. 1–5.

[42] P. D. Mitcheson, E. M. Yeatman, G. K. Rao, A. S. Holmes, and T. C. Green, "Energy harvesting from human and machine motion for wireless electronic devices," *Proc. IEEE*, vol. 96, no. 9, pp. 1457–1486, Sep. 2008.

[43] R. J. M. Vullers, R. V. Schaijk, H. J. Visser, J. Penders, and C. V. Hoof, "Energy harvesting for autonomous wireless sensor networks," *IEEE Solid-State Circuits Mag.*, vol. 2, no. 2, pp. 29–38, Spring, 2010.

[44] V. Leonov, P. Fiorini, S. Sedky, T. Torfs, and C. Van Hoof, "Thermoelectric MEMS generators as a power supply for a body area network," in *13th Int. Solid-State Sens., Actuators, Microsyst. Conf. Tech. Dig.*, vol. 1, pp. 291–294.

Manuel Piñuela (M'12) received the BSc. degree in electrical and electronic engineering from the National Autonomous University of Mexico (UNAM), Mexico City, Mexico, in 2007, and is currently working toward the Ph.D. degree at Imperial College London, London, U.K.

From 2006 to 2009, he worked in industry as an Electronic Design and Project Engineer for companies focused on oil and gas services in both Mexico and the US. His research interests are wireless power transfer, RF power amplifiers, and RF energy harvesting.

Mr. Piñuela was the recipient of the Gabino Barrera Medal.

Paul D. Mitcheson (SM'12) received the M.Eng. degree in electrical and electronic engineering and Ph.D. degree from Imperial College London, U.K., in 2001 and 2005, respectively.

He is currently a Senior Lecturer with the Control and Power Research Group, Electrical and Electronic Engineering Department, Imperial College London. His research interests are energy harvesting, power electronics, and wireless power transfer. He is involved with providing power to applications in circumstances where batteries and cables are not suitable. His work has been sponsored by the European Commission, EPSRC and several companies.

Dr. Mitcheson is a fellow of the Higher Education Academy.

Stepan Lucyszyn (M'91–SM'04) received the Ph.D. degree in electronic engineering from King's College London (University of London), London, U.K., in 1992, and the D.Sc. (higher doctorate) degree from Imperial College London, London, U.K., in 2010.

He is currently a Reader (Associate Professor) of millimeter-wave electronics and Director of the Centre for Terahertz Science and Engineering, Imperial College London. After working in industry as a Satellite Systems Engineer for maritime and military communications, he spent the first 12 years researching microwave and millimeter-wave RF integrated circuits (RFICs)/monolithic microwave integrated circuits (MMICs), followed by RF microelectromechanical systems (MEMS) technologies. He has coauthored approximately 140 papers and 11 book chapters in applied physics and electronic engineering. He has delivered many invited presentations at international conferences.

Dr. Lucyszyn was an associate editor for the JOURNAL OF MICROELECTROMECHANICAL SYSTEMS (2005–2009). In 2011, he was the chairman of the 41st European Microwave Conference, Manchester, U.K. In 2005, he was elected Fellow of the Institution of Electrical Engineers (IEE), U.K., and Fellow of the Institute of Physics, U.K. In 2008, was invited as a Fellow of the Electromagnetics Academy, USA. In 2009, he became an IEEE Distinguished Microwave Lecturer (2010–2012). He is currently an Emeritus DML for 2013 and a newly appointed European Microwave Lecturer (EML) for the European Microwave Association.

Agrawal et al., *Cogent Engineering* (2017), 4: 1332705
https://doi.org/10.1080/23311916.2017.1332705

cogent engineering

Received: 06 March 2017
Accepted: 12 May 2017
First Published: 20 May 2017

*Corresponding author: Sachin Agrawal, PDPM-Indian Institute of Information Technology, Design and Manufacturing, Jabalpur 482005, India
E-mail: bitssachin.agrawal@gmail.com

Reviewing editor:
Wei Meng, Wuhan University of Technology, China

Additional information is available at the end of the article

ELECTRICAL & ELECTRONIC ENGINEERING | RESEARCH ARTICLE

A dual-band RF energy harvesting circuit using 4th order dual-band matching network

Sachin Agrawal[1]*, Manoj S. Parihar[1] and P.N. Kondekar[1]

Abstract: A novel compact rectifier for dual-band operation in the RF energy harvesting is presented. The circuit comprises a 4th order dual-band impedance matching and a single-series circuit with one double diode, both are integrating into a compact shape to occupy a small area of $30 \times 35\ \text{mm}^2$. The merit of the proposed rectifier circuit is that it can be extended to *n* number of the frequency band by using only $2 \times n$ matching elements. To validate the design method experimentally, a prototype of a dual-band rectifier is fabricated for two public telecommunication bands of GSM-900 and 1800. In order to reduce the circuit complexity and sensitivity arising due to lumped elements, the meander line and the open stub are used to realize the proposed circuit. A good agreement is obtained between the simulation and the measurement. The measured results show that the proposed rectifier circuit exhibits the conversion efficiency of 25.7 and 65% for an input power of −20 and 0 dBm, respectively. In addition, diode nonlinearity which affects the performance of the rectifier in terms of impedance matching is also investigated.

Subjects: Electromagnetics & Microwaves; Electronics; Circuits & Devices

Keywords: RF energy harvesting; dual band impedance matching; rectifier; RF-to-dc-conversion efficiency; frequency transformation

Sachin Agrawal

ABOUT THE AUTHOR

Sachin Agrawal did his BE and ME from Jiwaji University and Birla Institute of Technology Pilani, India in 2005 and 2009, respectively. Currently, he is pursuing his PhD from PDPM-Indian Institute of Information Technology Design and Manufacturing Jabalpur, India. His area of interest are; RF energy harvesting circuit, Antenna including planar and dielectric resonator antenna and Power amplifier.

PUBLIC INTEREST STATEMENT

With rapid growth in wireless communication, a huge amount of radio frequency (RF) energy broadcasted through billions of microwave sources such as mobile phones, handheld radios, and radio broadcast stations. Therefore, it is meaningful to collect and supply it to many electrical devices like mobile headsets, wearable medical sensors through RF energy harvesting. Since the ambient RF energy is distributed in multiple frequency bands, therefore the amount of energy harvested could increase if the circuit is designed for multiple frequency bands. In this work, we present a compact dual-band energy harvesting circuit to harvest energy from two most useful frequency bands, GSM-900 and 1800. The merit of the proposed rectifier circuit is it can be extended to *n* number of the frequency band by using only $2 \times n$ matching elements. A prototype is fabricated, and its performance is evaluated using Vector Network Analyzer (VNA). The total size of the rectifier is about $30 \times 35\ \text{mm}^2$.

© 2017 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

1. Introduction

A revolutionary growth in wireless technology attracts huge attention from research community to make the self-sustainable device feasible through RF energy harvesting. It exploits ambient electromagnetic energy transmitted from different RF systems to remotely feed the electronic devices (Nintanavongsa, Muncuk, Lewis, & Chowdhury, 2012). Compared to other harvesting techniques, RF energy harvesting provides relatively predictable energy supply owing to the features of easy availability and less dependency on environmental variations. The typical block diagram of RF energy harvesting circuit is shown in Figure 1. It consists of three major blocks viz; antenna, matching network (MN), diode detector followed by an energy storage. The first element, antenna is employed to capture the RF signals of different frequencies and polarization, while second MN is for maximum power transfer, and the last rectifier is used to convert the RF energy to dc voltage. It means harvesting circuit performance can be evaluated in terms of accessible ambient RF energy and its conversion rate (Agrawal, Pandey, Singh,& Parihar, 2014). These parameters are heavily influenced by surrounding terrain conditions as the multiple reflection and dissipation certainly deteriorate the level of available ambient RF energy. As a result, conversion efficiency and dc output voltage may degrade. Previously, the majority of available RF energy harvesting circuits focused on single frequency band hence offer low dc output voltage. As the multiple RF energy sources of different frequency bands are available, thus from an ambient RF harvesting perspective, the output dc voltage could be increased if the circuit is designed for multiple frequency bands rather than a single band. A wide-band energy harvester can also promise a high output voltage by accumulating the number of RF signals at a time. However, due to nonlinear behavior of the diode, harvesting circuit itself exhibits nonlinearity i.e. its input impedance varies with the received RF power. Thus, it is quite difficult to retain the impedance match and high conversion efficiency over a large frequency range (Song, Huang, Zhou, & Carter, 2014). The losses due to impedance mismatch over a large bandwidth can be illustrated in Collado and Georgiadis (2013), where only 8% conversion efficiency is achieved at -20 dBm.

To address this, it is preferable to harvest energy from several narrow frequency bands rather than a single large one. In literature, numerous topologies have been proposed to accomplish the multiband energy harvesting (Bergès, Fadel, Oyhenart, Vigneras, & Taris, 2015; Hamano et al., 2016; Ho et al., 2016; Keyrouz, Visser, & Tijhuis, 2013; Kuhn, Lahuec, Seguin, & Person, 2015; Liu, Zhong, & Guo, 2015; Niotaki, Georgiadis, Collado, & Vardakas, 2014; Pinuela, Mitcheson, & Lucyszyn, 2013; Scheeler, Korhummel, & Popovic, 2014; Shariati, Rowe, Scott, & Ghorban, 2015; Sun, Guo, He, & Zhong, 2013). These topologies can be differentiated in terms of filter functionality i.e. how the antenna or source impedance is matched to the rectifier circuit. For instance, in Pinuela et al. (2013) and Keyrouz et al. (2013) several single-band rectennas (combination of antenna and rectifier circuit) were stacked to constitute a multi-band harvesting circuit. In this case, each rectenna was designed for a specific frequency band. Thus, for compact applications, this architecture is not suitable due to the number of antennas used. Moreover, in most of the reported works, the quality assessment of the output voltages combination was not taken into consideration. In Kuhn et al. (2015), the circuit complexity is reduced to a certain extent by replacing the multiple antennas with a single wide-band antenna. However, in this topology too, the number of rectifiers increases with the frequency bands, which leads to prolonging the circuit complexity.

Figure 1. Typical block diagram of RF energy harvesting circuit.

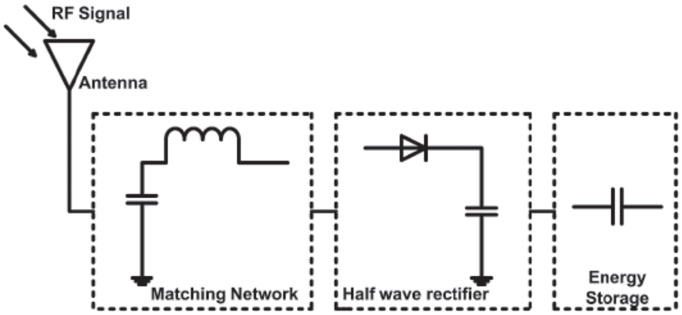

Besides, a multi-band harvesting circuit can also be formed by simply embedding a multi-band matching network between the multi-band antenna and the rectifying circuit (Bergès et al., 2015; Hamano et al., 2016; Ho et al., 2016; Liu et al., 2015; Niotaki et al., 2014; Scheeler et al., 2014; Shariati et al., 2015; Sun et al., 2013). The multi-band matching network can be designed either by distributed or by lumped element. In general, the multi-band rectifier circuit experiences two types of losses: first due to shift in resonance frequency from the optimum frequency point, and second due to the filter complexity. Because of the diode nonlinearity, the input impedance of the circuit varies as a function of power and frequency which causes a shift in resonance frequency. The difficulty due to diode nonlinearity can be observed in Sun et al. (2013) where the dual-band rectifier circuit exhibits its impedance matching for a small range of input power. The losses induced because of filter complexity can be observed in the recently reported works on dual band harvesting circuit (Niotaki et al., 2014; Scheeler et al., 2014; Shariati et al., 2015). In Niotaki et al. (2014), for $P_{\text{in}} = -15$ dBm, author achieved the conversion efficiency of 23% at the expense of increased filter complexity consisting two series and two shunt pairs of reactive elements. Thus, for more than dual band applications, the proposed circuit topology is not suitable due to excessive filtering components used. To obtain good conversion efficiency a dual-band rectenna reported in Scheeler et al. (2014). However, the rectenna was large in size and requiring a complex impedance tuning circuit. In Shariati et al. (2015) also, a dual-band matching network consisting nine reactive elements was employed to achieve the dual-band characteristics.

In order to reduce the filter complexity, this work proposed a compact dual-band harvesting circuit for GSM-900 and 1800. It consists of a 4th order dual-band matching network based on $1-n$ frequency transformation, which is optimized for the energy harvesting circuit to reduce the complexity up to $2 \times n$ reactive elements (n is the number of frequency bands). Similar to frequency transformation method, the proposed dual-band rectifier circuit can be extended to n number of frequency bands by using the $2 \times n$ number of reactive elements. The detailed analysis and design guidelines of dual band rectifier circuit are discussed in Section 2.

2. Dual band rectifier design and analysis

This section presents the design and analysis of a dual-band harvesting circuit in terms of impedance matching, DC output voltage and RF-to-dc conversion efficiency. The topology of the proposed dual-band RF energy harvesting circuit is shown in Figure 2(a). As seen, the low-cost Schottky diode is used to transform the input RF power to DC voltage. The impedance matching at two frequency is achieved using a series and parallel combination of the *LC* pair. The main idea underlying the suggested multi-band matching network is $1-n$ frequency transformation (one to many mapping of frequency), which transforms a single-band matching network to multi-band matching network (Nallam & Chatterjee, 2013). As the name ($1-n$) suggests that for designing a multi-band matching network, primarily a single-band matching network is required whose resonant frequency is dependent on the frequencies for which multi-band matching network proposed to designed.

Moreover, this frequency transformation method depends on the type of load impedance, whether it is series or parallel combination of *RC* or *RL*. Since the selected diode (HSMS-2852) has capacitive behavior throughout the frequency, it can be represented in a series or parallel combination of *R* and *C*. In the case of parallel *RC* load, the following equations are used to transform the single-band matching network into multi-band matching network.

$$\omega = \frac{\omega_t^{n} + a_m \omega_t^{n-m} + a_{m+2}\omega_t^{n-(m+2)} + \cdots}{\omega_t^{n-1} + a_{m+1}\omega_t^{n-(m+1)} + a_{m+3}\omega_t^{n-(m+3)} + \cdots} \tag{1}$$

where, n is the number of bands and m varies from 2 to n. After substituting the value of n, Equation (1) can be expanded in partial fraction form using the causal foster analysis as:

$$\omega = \omega_t + \frac{1}{\frac{\omega_t}{a_2 - a_3} + \frac{1}{\cdots}} \tag{2}$$

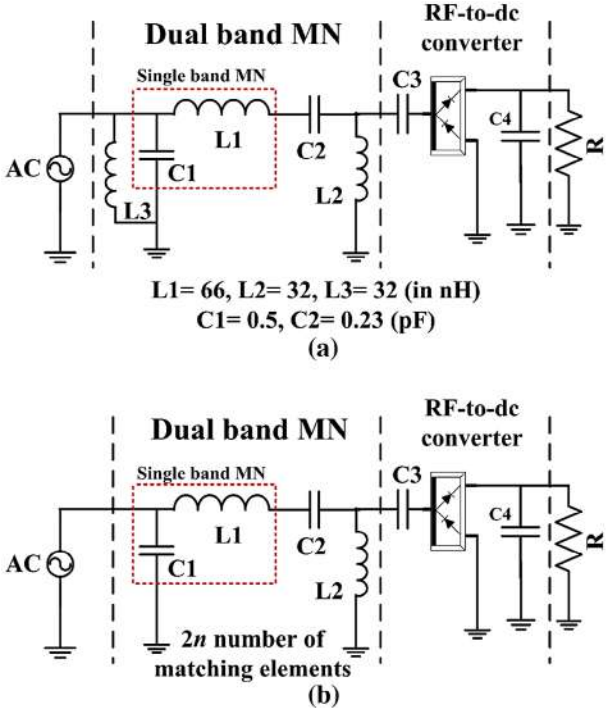
Dual band MN
RF-to-dc converter
Single band MN
C3
L1
C1
C2
C4
R
AC
L3
L2
L1= 66, L2= 32, L3= 32 (in nH)
C1= 0.5, C2= 0.23 (pF)
(a)
Dual band MN
RF-to-dc converter
Single band MN
C3
L1
C1
C2
C4
R
AC
L2
2*n* number of matching elements
(b)

Figure 2. (a) Circuit diagram of the proposed dual-band rectifier circuit and (b) optimized dual-band rectifier circuit.

The coefficients a_2 and a_3 can be calculated as:

$$\omega_m = \sum_{i=1}^{n} (-1)^{i-1} \omega_i \tag{3}$$

$$a_m = (-1)^m \sum_{i,j=1,1\,\&\,i\neq j}^{n,n} (-1)^{i+j} \omega_i \omega_j \tag{4}$$

$$a_{m+1} = (-1)^{m+1} \sum_{i,j,k=1,1,1\,\&\,i\neq j\neq k}^{n,n,n} (-1)^{i+j+k} \omega_i \omega_j \omega_k \tag{5}$$

$$a_{m+n} = (-1)^{m+n} \sum_{i,j,k\ldots=1,1,1\ldots i\neq j\neq k}^{n,n,n\ldots} (-1)^{i+j+k+\cdots} \omega_i \omega_j \omega_k \ldots \tag{6}$$

$$a_n = \prod_{i=1}^{n} (-1)^n \omega_i \tag{7}$$

Equation (6) is similar to that presented in Nallam and Chatterjee (2013), except the term $(-1)^{m+n}$, which is included here to realize the multi-band matching network for more than three frequency bands i.e. for $n \geq 3$.

With this transformation, the capacitor of the matching network is transformed to the combination of prototype capacitor parallel with inductor whereas, an inductor is transformed into a combination of the same inductor with a series capacitor. Figure 3 shows the circuit schematic of transformation of a single-band matching network to the dual-band matching network. It can be seen that C_1 is transformed to $C_1 \| L_2$ and L_1 transformed to L_1 series with C_2. After successful usage of (1)–(7), the resultant multi-band matching network requires $3n-1$ and $4n-1$ reactive elements for L and Π-type topology, respectively.

Figure 3. Conversion of a single-band matching network to dual-band matching network using $1 - n$ frequency transformation.

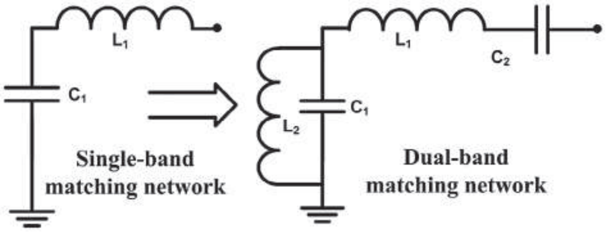

As the aim is to design a dual-band harvesting circuit, therefore, we require here only 5 or 7 reactive elements with L and Π-type topologies, respectively. From Figure 2(a), it can be seen that the resultant matching network consists of five elements, where the inductor L_3 and capacitor C_2 results after the transformation of capacitor C_1 and inductor L_1, respectively. Besides, the inductor L_2 occurrs due to the diode reactive element, which is generally a capacitor.

In this work, two frequencies 0.9 and 1.8 GHz that correspond to the maximum signal strength are chosen for dual-band harvesting circuit. According to this method, it is necessary to assign the frequencies in descending order e.g. $\omega_1 = 1.8, \omega_2 = 0.9$. Therefore, from (3) single-band matching network frequency is equal to $\omega_1 - \omega_2 = 2\pi(1.8 - 0.9) \times 10^9 = 0.9 \times 2\pi \times 10^9$. In order to match the source impedance with the rectifier at the calculated frequency 0.9 GHz, the chosen matching topology is L-type as shown by the encircled portion in Figure 2(a). The corresponding element values can be approximated using the various methods some of which are described in Pozar (2010). Subsequently, this single-band matching network is transformed to dual-band matching network using (1)–(7). The detailed design steps of the dual-band rectifier circuit are summarized as follows:

(1) As we are interested in matching the diode to 50 Ω at two frequencies (0.9 and 1.8 GHz) so, the order of transformation is equal to 2 or $n = 2$.

(2) In the first step, single-band matching network is designed at the frequency f calculated as: $f = f_2 - f_1 = 1.8 - 0.9 = 0.9$ GHz. In this case, any matching topology that matches the diode to 50 Ω, at 0.9 GHz for an input power $P_{\text{in}} = -20$ dBm, and load resistance 4.7 kΩ can be used. The chosen single-band matching network is shown by the encircled portion in Figure 2(a).

(3) Afterwards, this single-band matching network is transformed into dual-band using the (1)–(7) as shown below:

Since $n = 2$ therefore, from (2)

$$\omega = \omega_t + \frac{1}{\frac{\omega_t}{a_2}} \tag{8}$$

From (4) a_2 can be calculated as:

$$a_2 = (-1)^{2+2} \sum (-1)^{1+2} \omega_1 \omega_2 = \omega_1 \omega_2 \tag{9}$$

$$a_2 = -1.62 \times 4\pi^2 \times 10^8 = 0.64 \times 10^{20} \tag{10}$$

Thus, inductor L_1 (=66 nH) is transformed to impedance as:

$$j66 \times 10^9 \omega = j66 \times 10^9 \omega_t + \frac{1}{j0.23 \times 10^{-12} \omega_t} \tag{11}$$

Similarly, capacitors (=0.5 pF) are transformed to the admittance as:

$$j5 \times 10^{-13} \omega = j5 \times 10^{-13} \omega_t + \frac{1}{j32 \times 10^{-9} \omega_t} \tag{12}$$

The circuit schematic of the dual-band harvesting circuit is shown in Figure 2(a). It can be seen that resultant matching network consists of five reactive elements according to $3n - 1$. In order to reduce the circuit complexity and sensitivity due to reactive elements, a parametric study has been carried out to eliminate the elements showing minimum influence on the circuit performance.

Figure 4 shows the simulated $|S_{11}|$ for the different combination of matching elements. The simulated results demonstrate that $|S_{11}|$ experiences maximum change when inductor L_2 and capacitor C_2 are removed from the circuit, whereas it remains almost unaffected when L_3 is not present in the circuit. Therefore, inductor L_3 can be extruded from the circuit and the resultant matching circuit requires only $2n$ and $3n$ reactive elements in place of $3n - 1$ and $4n - 1$ elements. In this way, for each topology, the proposed circuit reduces $n - 1$ elements compared to the conventional method. Figure 2(b) demonstrates the optimized circuit diagram of the dual-band rectifier. It can be observed that circuit requires large inductors value of 32 and 66 nH. Thus, it is quite difficult to realize the practical rectifier circuit whose response is similar to the response of simulated result. In order to avoid any impedance mismatch due to the small difference in elements value, the meander line inductor and open stub are used to realize the inductors and capacitors, respectively. In this case, not only fabrication and optimization process become so easy but the cost will also reduced.

Figure 5, shows the layout of the dual-band rectifier circuit. In Nintanavongsa et al. (2012) and Agrawal et al. (2014), it has been demonstrated that the number of rectifying diodes or equivalently voltage multiplier stages are very much sensitive to the RF-to-dc conversion efficiency. In low-power region (≤ -20 dBm), efficiency decreases if voltage multiplier stages increase, whereas in higher power region (≥ -20 dBm), an opposite effect occurs. As the demand is to harvest energy in low-power region, single-series circuit with a double diode is used to convert received RF energy into dc voltage. From the left side of the circuit, the first meander line corresponds to the inductor L_1, while the second meander line represents the inductor L_2, of the Figure 2(b). The shunt stub is accounted

Figure 4. Simulated $|S_{11}|$ vs. frequency for different combination of circuit elements.

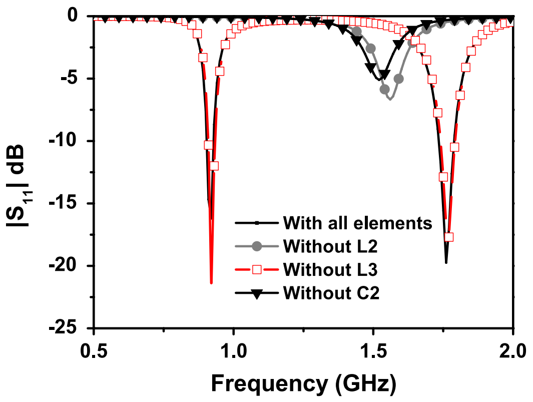

Figure 5. Layout of the proposed dual-band RF energy harvesting circuit.

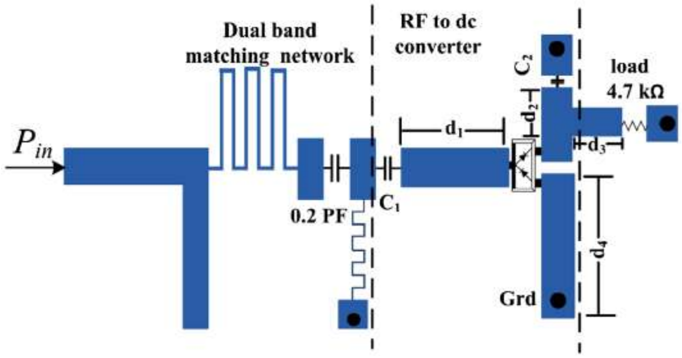

for the shunt capacitor C_1 of Figure 2(b). The dimensions of each element are calculated according to their respective reactive element value and the substrate on which circuit has to be fabricated. In Assimonis, Daskalakis, and Bletsas (2016), it has been demonstrated that traces (microstrips) connected to the rectifier terminals (e.g. distance between the diode and via and diode and load) are highly sensitive for RF-to-dc efficiency. Therefore, traces d_1, d_2 between diode and capacitor C_1 and C_2, d_3 between diode and load and d_4 between diode and ground are adjusted to optimize the impedance matching as well as the conversion efficiency of the rectifier. Due to nonlinear behavior of the diode, harvesting circuit itself exhibits nonlinearity i.e. its input impedance varies with received RF power, therefore harmonic-balance (HB) and large signal analysis (LSSP) were employed to take into consideration the nonlinear behavior of the rectifier.

The photograph of the fabricated dual-band rectifier is shown in Figure 6. It is fabricated on a 1.54 mm thick FR-4 substrate with a dielectric constant (ϵ_r) of 4.3 using chemical etching method. The rectifier performance is evaluated in terms of $|S_{11}|$ and output voltage using the Agilent vector network analyzer (VNA). The simulated and measured $|S_{11}|$ is illustrated in Figure 7(a). The measured result shows reasonable agreement with the simulated one; the slight difference can be accounted for the fabrication imperfections. It is well known that impedance matching is a function of frequency and input power, due to the nonlinearity of the diode. Such a characteristic is examined in Figure 7(b), where the measured $|S_{11}|$ is demonstrated as a function of input power level for three different load impedance values. From results, it is clear that impedance matching of the harvesting circuit is greatly affected by the input power and the load impedance. Figure 7(b) demonstrates that as power increases, the impedance matching at 0.9 GHz degraded drastically, while at 1.8 GHz, it improves. Moreover, it is noticed that the impedance matching at higher power level is more sensitive to the variation of load impedance (*RL*).

The measured RF-to-dc conversion efficiency and output voltage vs. input power for both frequencies are demonstrated in Figure 8(a). For 0.9 GHz, efficiency is equal to 25.7 and 65.1% for an input power of -20 and 0 dBm, respectively. However, at 1.8 GHz the efficiency is relatively small that might be due to the increased parasitic losses in the rectifier diode. Figure 8(b) shows the relation between the output voltage and frequency for various input power levels at fixed load resistance value of 4.7 kΩ. It can be seen that maximum output voltage is achieved in the frequency range of 860–900 and 1770–1800 MHz, showing the rectifier's capability to harvest RF energy in the GSM-900 and 1800 bands.

Figure 9, depicts the measured conversion for various values of load resistance. It can be noticed that the circuit yields maximum efficiency when the load impedance is 4.7 kΩ. It starts decreasing as the load impedance varies from 4.7 kΩ. Table 1 shows the comparison of the conversion efficiency and the size of the proposed rectifier with the similar works reported previously. Only measured results are compared in Table 1. It can be seen that in low-power condition maximum efficiency is achieved in Sun et al. (2013), but the expense of bulky circuit size. However, the proposed rectifier offers an optimal conversion efficiency with compact circuit size.

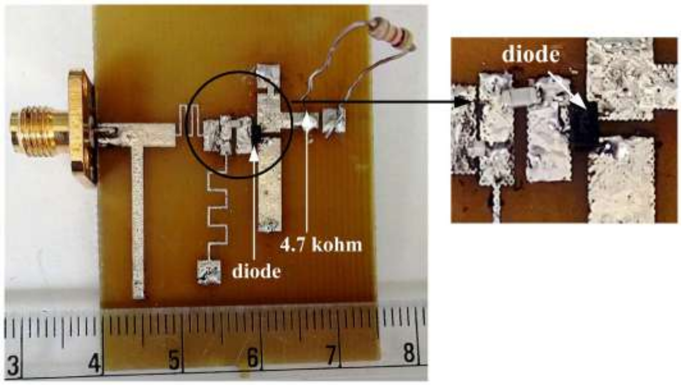

Figure 6. Photograph of the fabricated dual-band RF energy harvesting circuit.

Figure 7. (a) Simulated and measured $|S_{11}|$ vs. frequency and (b) measured $|S_{11}|$ for various power levels.

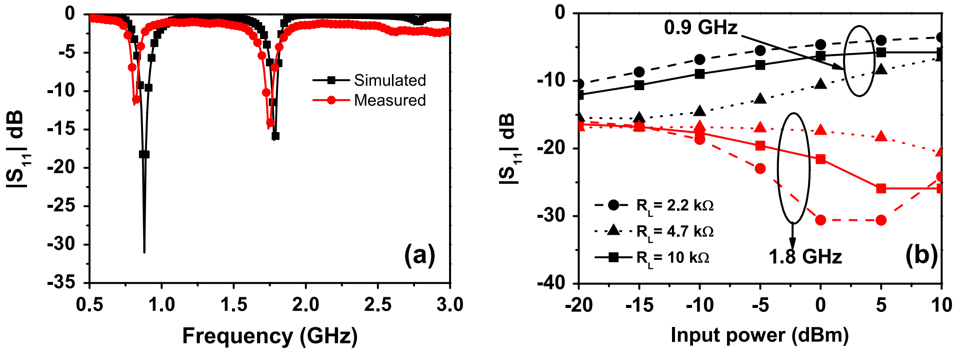

Figure 8. (a) Measured output dc voltage and RF-to-dc conversion efficiency and (b) measured dc voltage vs. frequency.

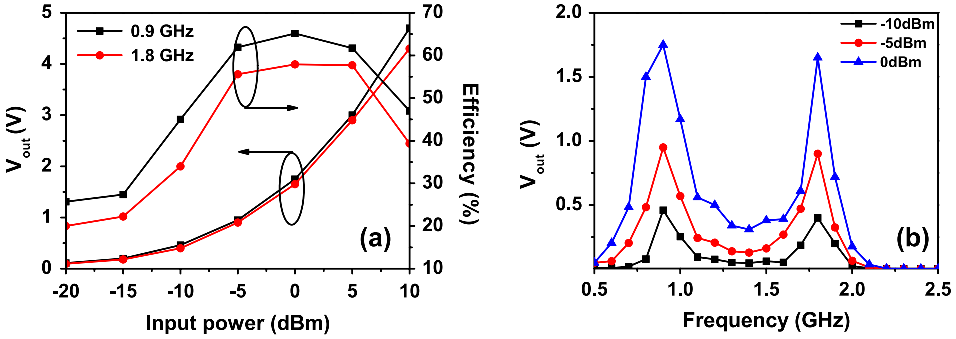

Figure 9. Measured RF-to-dc conversion efficiency for different load impedance.

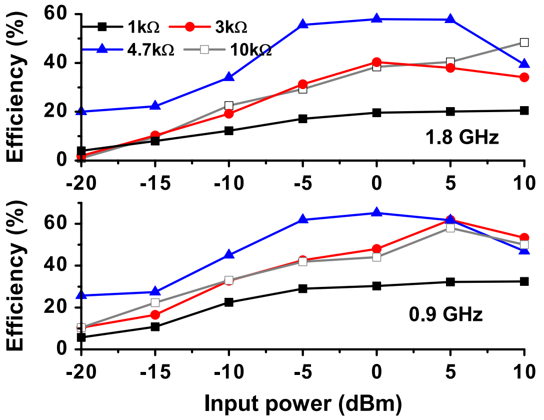

Table 1. Performance comparison of the proposed dual rectifier with recently published works

Ref.	Measured rectifier efficiency (%)	Input power (dBm)	Rectifier size
Ho et al. (2016)	15.8 @ 0.89 GHz	−20	$100 \times 65\ \text{mm}^2$
	11.2 @ 1.76 GHz		
Hamano et al. (2016)	10 @ 2.15 GHz	−10	$37 \times 71\ \text{mm}^2$
	15 @ 5.84 GHz		
Bergès et al. (2015)	27 @ 0.91/2.4 GHz	−16	$78 \times 88\ \text{mm}^2$
Sun et al. (2013)	30 @ 2.14 GHz	−20	145 mm
	35 @ 1.84 GHz		
Liu et al. (15)	20 @ (0.91+1.8) GHz	−20	$23 \times 37\ \text{mm}^2$
This work	27.5 @ 0.9 GHz	−20	$30 \times 35\ \text{mm}^2$
	20 @ 1.8 GHz		

3. Conclusion

A new compact 4th order dual-band rectifier has been designed to harvest the RF power of GSM-900 and 1800 bands. In order to reduce the circuit complexity and sensitivity due to reactive elements, the meander line and the open stub are used to fabricate the matching network. For $P_{\text{in}} = -20$ dBm, the measured RF-to-dc conversion efficiency of 27.5 and 20% is achieved at 0.9 and 1.8 GHz, respectively. Further, more than 45 and 34% conversion efficiency is maintained from -10 to 10 dBm for 0.9 and 1.8 GHz, respectively.

Funding
The authors received no direct funding for this research.

Author details
Sachin Agrawal[1]
E-mail: bitssachin.agrawal@gmail.com
Manoj S. Parihar[1]
E-mail: mparihar@iiitdmj.ac.in
P.N. Kondekar[1]
E-mail: pnkondekar@iiitdmj.ac.in
[1] PDPM-Indian Institute of Information Technology, Design and Manufacturing, Jabalpur 482005, India.

Citation information
Cite this article as: A dual-band RF energy harvesting circuit using 4th order dual-band matching network, Sachin Agrawal, Manoj S. Parihar & P.N. Kondekar, *Cogent Engineering* (2017), 4: 1332705.

References

Agrawal, S., Pandey, S. K., Singh, J., & Parihar, M. S. (2014). *Realization of efficient RF energy harvesting circuits employing different matching technique* (pp. 754–761). Fifteenth International Symposium on Quality Electronic Design, Santa Clara, CA.

Assimonis, S. D., Daskalakis, S. N., & Bletsas, A. (2016). Sensitive and efficient rf harvesting supply for batteryless backscatter sensor networks. *IEEE Transactions on Microwave Theory and Techniques, 64*, 1327–1338.

Bergès, R., Fadel, L., Oyhenart, L., Vigneras, V., & Taris, T. (2015). A dual band 915MHz/2.44 GHz RF energy harvester. *Microwave Conference (EuMC), 2015 European*. IEEE.

Collado, A., & Georgiadis, A. (2013). Conformal hybrid solar and electromagnetic (EM) energy harvesting rectenna. *IEEE Transactions on Circuits and Systems I: Regular Papers, 60*, p. 10.

Hamano, K., Tanaka, R., Yoshida, S., Miyachi, A., Nishikawa, K., & Kawasaki, S. (2016). Design of dual-band rectifier using microstrip spurline notch filter. *Radio-Frequency Integration Technology (RFIT), 2016 IEEE International Symposium on*. IEEE.

Ho, D. K., Kharrat, I., Ngo, V. D., Vuong, T. P., Nguyen, Q. C., & Le, M. T. (2016). *Dual-band rectenna for ambient RF energy harvesting at GSM 900 MHz and 1800 MHz* (pp. 306–310). 2016 IEEE International Conference on Sustainable Energy Technologies (ICSET), Hanoi.

Keyrouz, S., Visser, H., & Tijhuis, A. (2013, April). Multi-band simultaneous radio frequency energy harvesting. *In Antennas and Propagation (EuCAP), 2013 7th European Conference on*, 3058–3061.

Kuhn, V., Lahuec, C., Seguin, F., & Person, C. (2015). A multi-band stacked RF energy harvester with RF-to-DC efficiency up to 84\%. *IEEE Transactions on Microwave Theory and Techniques, 63*, 1768–1778.

Liu, Z., Zhong, Z., & Guo, Y.-X. (2015). Enhanced dual-band ambient RF energy harvesting with ultra-wide power range. *IEEE Microwave and Wireless Components Letters, 25*, 630–632.

Nallam, N., & Chatterjee, S. (2013, June). Multi-band frequency transformations, matching networks and amplifiers. *Circuits and Systems I: Regular Papers, IEEE Transactions on, 60*, 1635–1647.

Nintanavongsa, P., Muncuk, U., Lewis, D., & Chowdhury, K. (2012, March). Design optimization and implementation for rf energy harvesting circuits. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2*, 24–33.

Niotaki, K., Georgiadis, A., Collado, A., & Vardakas, J. S. (2014). Dual-band resistance compression networks for improved rectifier performance. *IEEE Transactions on Microwave Theory and Techniques, 62*, 3512–3521.

Pinuela, M., Mitcheson, P., & Lucyszyn, S. (2013, July). Ambient rf energy harvesting in urban and semi-urban environments. *Microwave Theory and Techniques, IEEE Transactions on, 61*, 2715–2726.

Pozar, D. M. (2010). *Microwave engineering* (3rd ed.). New Delhi: Wiley.

Scheeler, R., Korhummel, S., & Popovic, Z. (2014). A dual-frequency ultralowpower efficient 0.5-g rectenna. *IEEE Microwave Magazine, 15*, 109–114.

Shariati, N., Rowe, W. S., Scott, J. R., & Ghorban, K. (2015). Multi-service highly sensitive rectifier for enhanced RF energy scavenging. *Scientific Reports, 5*, 9655

Song, C., Huang, Y., Zhou, J., & Carter, P. (2014). Improved ultrawideband rectennas using hybrid resistance compression technique. *IEEE Transactions on Antennas and Propagation, 65*, 2057–2062.

Sun, H., Guo, Y. X., He, M., & Zhong, Z. (2013). A dual-band rectenna using broadband Yagi antenna array for ambient RF power harvesting. *IEEE Antennas and Wireless Propagation Letters, 12*, 918–921. doi:10.1109/LAWP.2013.2272873.

© 2017 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

You are free to:
Share — copy and redistribute the material in any medium or format
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
No additional restrictions
You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

***Cogent Engineering* (ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com

Circuits and Systems, 2012, 3, 216-222
http://dx.doi.org/10.4236/cs.2012.33030 Published Online July 2012 (http://www.SciRP.org/journal/cs)

Optimization of the Voltage Doubler Stages in an RF-DC Convertor Module for Energy Harvesting

Kavuri Kasi Annapurna Devi[1*], Norashidah Md. Din[2], Chandan Kumar Chakrabarty[2]
[1]Department of Electrical and Electronic Engineering, INTI International University, Nilai, Malaysia
[2]Department of Electronics and Communication Engineering, Universiti Tenega Nasional, Kajang, Malaysia
Email: {*kavurik.adevi., chandan}@newinti.edu.my, norashidah@uniten.edu.my

Received August 25, 2011; revised March 26, 2012; accepted April 4, 2012

ABSTRACT

This paper presents an optimization of the voltage doubler stages in an energy conversion module for Radio Frequency (RF) energy harvesting system at 900 MHz band. The function of the energy conversion module is to convert the (RF) signals into direct-current (DC) voltage at the given frequency band to power the low power devices/circuits. The design is based on the Villard voltage doubler circuit. A 7 stage Schottky diode voltage doubler circuit is designed, modeled, simulated, fabricated and tested in this work. Multisim was used for the modeling and simulation work. Simulation and measurement were carried out for various input power levels at the specified frequency band. For an equivalent incident signal of –40 dBm, the circuit can produce 3 mV across a 100 kΩ load. The results also show that there is a multiplication factor of 22 at 0 dBm and produces DC output voltage of 5.0 V in measurement. This voltage can be used to power low power sensors in sensor networks ultimately in place of batteries.

Keywords: Energy Conversion; RF; Schottky Diode; Villard; Energy Harvesting

1. Introduction

RF energy harvesting is one type of energy harvesting that can be potentially harvested such as solar, vibration and wind. The RF energy harvesting uses the idea of capturing transmitted RF energy at ambient and either using it directly to power a low power circuit or storing it for later use. The concept needs an efficient antenna along with a circuit capable of converting RF signals to DC voltage. The efficiency of an antenna mainly depends on its impedance and the impedance of the energy converting circuit. If the two impedances aren't matched then it will be unable to receive all the available power from the free space at the desired frequency band. Matching of the impedances means that the impedance of the antenna is the complex conjugate of the impedance of the circuit (voltage doubler circuit).

The concept of energy harvesting system is shown in **Figure 1**, which consists of matching network, RF-DC conversion and load circuits. The authors in [1], used a 2.4 GHz operating frequency with an integrated zero bias detector circuit using BiCMOS technology which produced an output voltage of 1 V into a 1 MΩ load at an input power level of 0 dBm. H. Yan and co-authors revealed that a DC voltage of 0.8 volts can be achieved from a –20 dBm RF input signal at 868.3 MHz through simulation results [2]. In [3], work was carried out on a firm frequency of 900 MHz by matching to a 50 Ω impedance and resonance circuit transformation in front of the Schottky diode which yields an output voltage of over 300 mV at an input power level of 2.5 µ. W. J. Wang, L. Dong and Y. Fu [4] used a Cockcroft-Walton multiplier circuit that produced a voltage level of 1.0 V into a 200 MΩ load for an input power level of less than –30 dBm at a fixed frequency of 2.4 GHz.

The energy conversion module designed in this paper is based on a voltage doubler circuit which can be able to output a DC voltage typically larger than a simple diode rectifier circuit as in [5], in which switched capacitor charge pump circuits are used to design two phase voltage doubler and a multiphase voltage doubler. The module presented in this can function as an AC to DC converter that not only rectifies the input AC signal but also elevates the DC voltage level. The output voltage of the

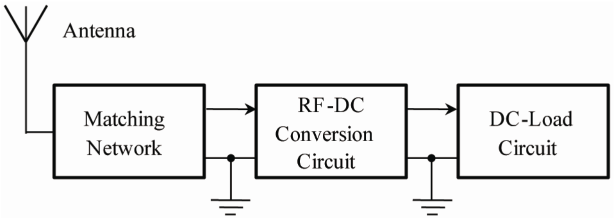

Figure 1. Schematic view of a RF energy harvesting system.

---

[*] Corresponding author.

Copyright © 2012 SciRes.

energy conversion module can be used to energize the low power devices for example sensors for a sensor network in application to agriculture.

Section 2 of this paper discusses on the theoretical background of the voltage doubler circuit. Section 3 presents the simulation study and implementation of the circuit design. Section 4 provides the results and analysis. Section 5 concludes with a discussion on the findings from the simulated and measured results.

2. Voltage Multiplier

There are various voltage multiplier circuit topologies. The design used in this module is derived from the function of peak detector or a half wave peak rectifier. The Villard voltage multiplier circuit was chosen in the circuit design of this paper because it produces two times of the input signal voltage towards ground at a single output and can be cascaded to form a voltage multiplier with an arbitrary output voltage and its design simplicity.

2.1. Diode Modeling

The voltage multiplier circuit in this design uses zero bias Schottky diode HSMS-2850 from Agilent. The attractive feature of these Schottky diodes are low substrate losses and very fast switching but leads to a fabrication overhead. This diode has been modeled for the energy harvesting circuit which comes in a one-diode configuration. The modeling parameters for these diodes are given by Agilent in their data sheets. These parameters are used in Multisim for its own modeling purposes. The modeling is done by transforming the diode into an equivalent circuit using passive components which are described by the SPICE parameters in **Table 1** [6].

The diode used in this design is shown in **Figure 2** and its equivalent model is shown in **Figure 3**. The special features of HSMS-2850 diode is that it provides a low forward voltage, low substrate leakage and uses the non symmetric properties of a diode that allows unidirectional flow of current under ideal condition.

Table 1. SPICE parameters.

Parameters	Units	HSMS 2850
B_V	V	3.8
C_{J0}	pF	0.18
E_G	Ev	0.69
I_{BV}	A	3E–4
I_S	A	3E–6
N	No unite	1.06
R_S	Ω	25
P_B (V_J)	V	0.35
P_T (XTI)	No units	2
M	No units	0.5

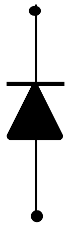

Figure 2. Schottky diode.

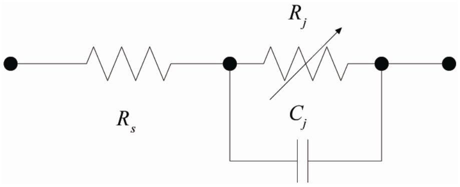

Figure 3. Linear circuit model of the Schottky diode [6].

The diodes are fixed and are not subject of optimization or tuning. This is described using the following derivations. By neglecting the effect of diode substrate, an equivalent linear model that can be used for the diode as shown in **Figure 3**. When C_j is the junction capacitance and R_j is the junction resistance, the admittance Y_z of the linear model is given by

$$Y_Z = Y_{C_j} + Y_{R_j} \tag{1}$$

Equation (1) related to the frequency of operation is given by

$$Y_Z = jwC_j + \frac{1}{R_j} \tag{2}$$

$$= \frac{jwC_jR_j + 1}{R_j} \tag{3}$$

The impedance Z of the linear model is given by

$$Z = \frac{R_j}{1 + jwR_jC_j} \tag{4}$$

The total impedance Z_T is given by

$$Z_T = R_S + \frac{R_j}{1 + jwR_jC_j} \tag{5}$$

where R_S is the series resistance of the circuit and R_j is given by

$$R_j = \frac{8.33 \times 10^{-5} \times N \times T}{I_b + I_S}$$

where:

I_b = bias current in µA;

I_s = saturation current in µA;

T = temperature (K);

N = ideality factor.

In Equation (5), R_j and C_j are constants and the frequency of operation (w) is the only variable parameter. As the frequency increases, the value of Z is almost negligible compared to the series resistance R_S of the diode. From this it is concluded that the function of the diode is independent of the frequency of operation.

2.2. Single Stage Voltage Multiplier

Figure 4 represents a single stage voltage multiplier circuit. The circuit is also called as a voltage doubler because in theory, the voltage that is arrived on the output is approximately twice that at the input. The circuit consists of two sections; each comprises a diode and a capacitor for rectification. The RF input signal is rectified in the positive half of the input cycle, followed by the negative half of the input cycle. But, the voltage stored on the input capacitor during one half cycle is transferred to the output capacitor during the next half cycle of the input signal. Thus, the voltage on output capacitor is roughly two times the peak voltage of the RF source minus the turn-on voltage of the diode.

The most interesting feature of this circuit is that when these stages are connected in series. This method behaves akin to the principle of stacking batteries in series to get more voltage at the output. The output of the first stage is not exactly pure DC voltage and it is basically an AC signal with a DC offset voltage. This is equivalent to a DC signal superimposed by ripple content. Due to this distinctive feature, succeeding stages in the circuit can get more voltage than the preceding stages. If a second stage is added on top of the first multiplier circuit, the only waveform that the second stage receives is the noise of the first stage. This noise is then doubled and added to the DC voltage of the first stage. Therefore, the more stages that are added, theoretically, more voltage will come from the system regardless of the input. Each independent stage with its dedicated voltage doubler circuit can be seen as a single battery with open circuit output voltage V_0, internal resistance R_0 with load resistance R_L, the output voltage, V_{out} is expressed as in Equation (7).

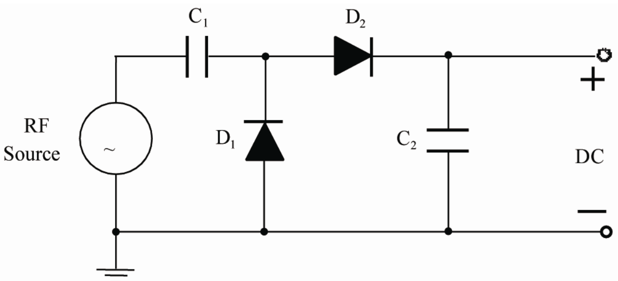

Figure 4. Single stage voltage multiplier circuit [7].

$$V_{\text{out}} = \frac{V_0}{R_0 + R_L} R_L \tag{6}$$

When n number of these circuits are put in series and connected to a load of R_L in Equation (6) the output voltage V_{out} obtained is given by this change in RC value will make the time constant longer which in turn retains the multiplication effect of two in this design of seven stage voltage doubler.

$$V_{\text{out}} = \frac{nV_0}{nR_0 + R_L} = V_0 \frac{1}{\frac{R_0}{R_L} + \frac{1}{n}} \tag{7}$$

The number of stages in the system has the greatest effect on the DC output voltage, as shown from Equations (6) and (7).

It is inferred that the output voltage V_{out} is determined by the addition of R_0/R_L and $1/n$, if V_0 is fixed. From this analysis it is observed that V_0, R_0 and R_L are all constants. Assume that $V_0 = 1$ V, $R_0/R_L = 0.25$, n = 2, 3, 4, 5, 6 and 7, the output voltage V_{out} = 1.33 V, 1.72 V, 2.0 V, 2.22 V, 2.43 V and 2.56 V respectively when substituted analytically in the Equation (7). This analysis can be compared with the results obtained in the circuit design of this module. In simulation at $n = 4$, $V_{\text{out}} = 1.42$ V, $n = 5$, $V_{\text{out}} = 1.67$ V; $n = 6$, $V_{\text{out}} = 1.92$; $n = 7$, $V_{\text{out}} = 2.15$ V; $n = 8$, $V_{\text{out}} = 1.92$ V; $n = 9$, $V_{\text{out}} = 1.81$ V. Also in measurement, for $n = 4$, $V_{\text{out}} = 2.1$ V; $n = 5$, $V_{\text{out}} = 2.9$ V; $n = 6$, $V_{\text{out}} = 3.72$ V; and $n = 7$, $V_{\text{out}} = 5$ V. As n increases, the increase in output voltage will be almost double the input voltage up to some number of stages. But at some point, *i.e.* beyond seven stages, in this circuit the output voltage gained (8 and 9 stages) will be negligible as shown in **Figure 5**.

The capacitors are charged to the peak value of the input RF signal and discharge to the series resistance (R_s) of the diode. Thus the output voltage across the capacitor of the first stage is approximately twice that of the input signal. As the signal swings from one stage to other, there is an additive resistance in the discharge path of the

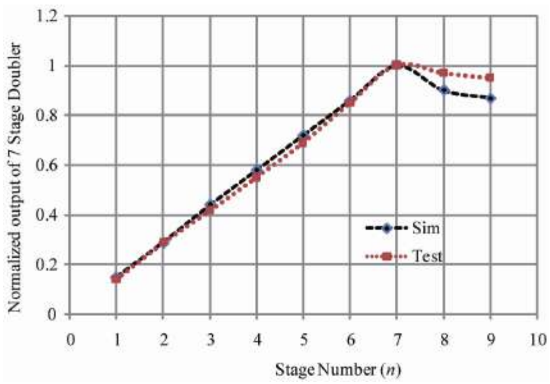

Figure 5. Normalized output voltage multiplier versus number of stages.

diode and increase of capacitance due to the stage capacitors.

2.3. Seven Stage Voltage Multiplier

The seven stage voltage multiplier circuit design implemented in this paper is shown in **Figure 6**. Starting on the left side, there is a RF signal source for the circuit followed by the first stage of the voltage multiplier circuit. Each stage is stacked onto the previous stage as shown in the **Figure 6**. Stacking was done from left to right for simplicity instead of conventional stacking from bottom to top.

The circuit uses eight zero bias Schottky surface-mount Agilent HSMS-285X series, HSMS-2850 diodes. The special features of these diode is that, it provides a low forward voltage, low substrate leakage and uses the non-symmetric properties of a diode that allows unidirectional flow of current under ideal conditions. The diodes are fixed and are not subject of optimization or tuning. This type of multiplier produces a DC voltage which depends on the incident RF voltage. Input to the circuit is a predefined RF source. The voltage conversion can be effective only if the input voltage is higher than the Schottky forward voltage.

The other components associated with the circuit are the stage capacitors. The chosen capacitors for this circuit are of through-hole type, which make it easier to modify for optimization, where in [8] the optimization was accomplished at the input impedance of the CMOS chip for a three stage voltage multiplier. The circuit design in this paper uses a capacitor across the load to store and provide DC leveling of the output voltage and its value only affects the speed of the transient response. Without a capacitor across the load, the output is not a good DC signal, but more of an offset AC signal.

In addition to the above, an equivalent load resistor is connected at the final node. The output voltage across the load decreases during the negative half cycle of the AC input signal. The voltage decreases is inversely proportional to the product of resistance and capacitance across the load. Without the load resistor on the circuit, the voltage would be hold indefinitely on the capacitor and look like a DC signal, assuming ideal components. In the design, the individual components of the stages need not to be rated to withstand the entire output voltage. Each component only needs to be concerned with the relative voltage differences directly across its own terminals and of the components immediately adjacent to it. In this type of circuitry, the circuit does not change the output voltage but increases the possible output current by a factor of two. The number of stages in the system is directly proportional to the amount of voltage obtained and has the greatest effect on the output voltage as explained in the Equation (7) and shown in **Figure 5**.

3. Simulation and Implementation

Multisim software was chosen for modeling and simulation which is a circuit simulation tool by Texas Instruments. The simulation and practical implementation were carried out with fixed RF at 945 MHz ± 100 MHz, which are close to the down link center radio frequency (947.5 MHz) of the GSM-900 transmitter. The voltages obtained at the final node V_{DC} of the multiplier circuit were recorded for various input power levels from –40 dBm - +5 dBm with power level interval (spacing) of 5 dBm.

The simulations were also carried out using same stage capacitance value (3.3 nF) and then with a varied capacitance value for all stages from 4 stages through 9 stages [9]. The capacitance value was varied in such a way that, from one stage to the next, it was halved. For example, if the first stage was 3.3 nF, the second stage was 1.65 nF, third stage was 825 pF, fourth stage was 415 pF and so on. But keeping in view of testing, the capacitance values were chosen to have a close match with the standard available values in the market.

Simulation was carried out through 4 to 9 voltage doubler stages. Based on results obtained a 7 stage doubler is best to implemented for this application.

The design of the printed circuit board (PCB) was carried out using DipTrace software. The material used to

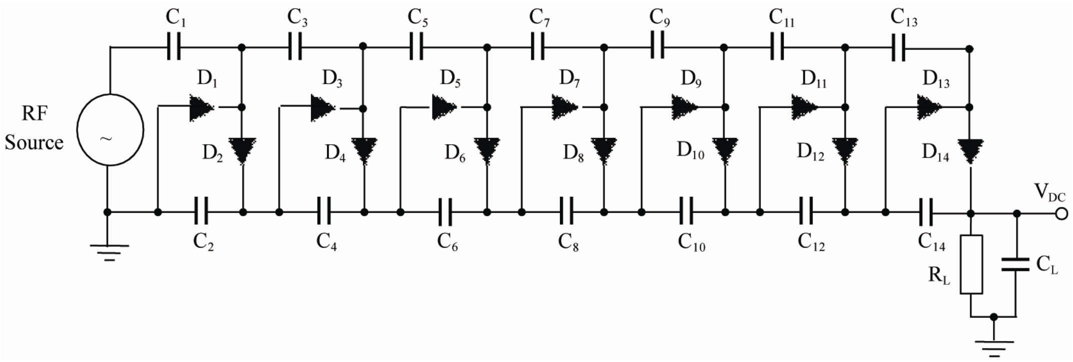

Figure 6. Schematic of 7 stage voltage multiplier.

Copyright © 2012 SciRes.

manufacture the printed circuit board (PCB) is the standard Fiberglass Reinforced Epoxy (FR4), with the thickness of 1.6 mm and dielectric constant of 3.9. The topology is constructed on the PCB with the dimensions of 98 mm × 34 mm (W × H). The Sub Miniature version A (SMA) connectors are used at the input and output of PCB to carry out the measurements. The circuit components consist of active and passive components. The component used in circuit is shown in **Table 2**.

Special handling precautions have been taken to avoid Electro Static Discharge (ESD), while assembling of the surface-mount zero bias Schottky diodes. Also special attention has been given to mount other components and the SMA connectors on to the PCB. The Photograph of Assembled circuit board I shown in **Figure 7**.

4. Results and Analysis

The simulated and measured results at the output voltage of voltage multiplier circuit are shown graphically in **Figure 8**. From the graph analysis, the simulated and the measured results agree considerably with each other. The measured results are shown to be better than the simulation results. The reason behind this may be due to the uncertainty in series resistance value of the diode obtained from SPICE parameters in modeling as explained in Equation (5). This resistance vale of diodes in practical circuit may be lower than in the model, which provides fast discharge path, in turn rise in voltage as passes through the stages and reaches to final output. In this work, the DC output voltages obtained through simulation and measurement at 0 dBm re 2.12 V and 5.0 V respectively. These results are comparatively much better than in ref. [9], where in at 0 dBm, 900 MHz they achieved 0.5 V and 0.8 V through simulation and measurement respectively.

Figures 9 and **10** show the result of a 4 stage voltage doubler circuit with equal and varied capacitance values between the stages as described in Section 3.

From the analysis of these two simulations, it can be observed that the resulting output voltages are equal. The only difference between these two graphs is the rise time of the circuit with varied capacitance value is a little bit slower. But, overall result on the performance of rise time is still under 20 µs to 24 µs and the difference is negligible. From these results, the use of equal stage capacitance of each being 3.3 nF was hence considered for the design of the multiplier.

The results from **Figure 11**, shows that the output voltage reaches to 1.0 V within 20 µS and then uniformly increasing to 1.4 V, 1.67 V, 1.87 V and 2.12 V for 4, 5, 6 and 7 stages respectively compared to 2 mS as shown in [10]. **Figure 12** shows that the conversion ratio of 22 is achieved at 0 dBm input power and drops to 2.5 at –40 dBm. The highest value at 0 dBm is due to the innate characteristics of the zero bias Schottky diodes which conduct fairly well at higher input voltages.

5. Conclusion

From the experimental results, it is found that the pro-

Table 2. Component used in 7 stage voltage multiplier.

Name of component	Label	Value
Stage capacitors	C_1 - C_{14}	3.3 nF
Stage diodes	D_1 - D_{14}	HSMS 2850
Filter capacitor	C_L	100 nF
Load resister	R_L	100 kΩ

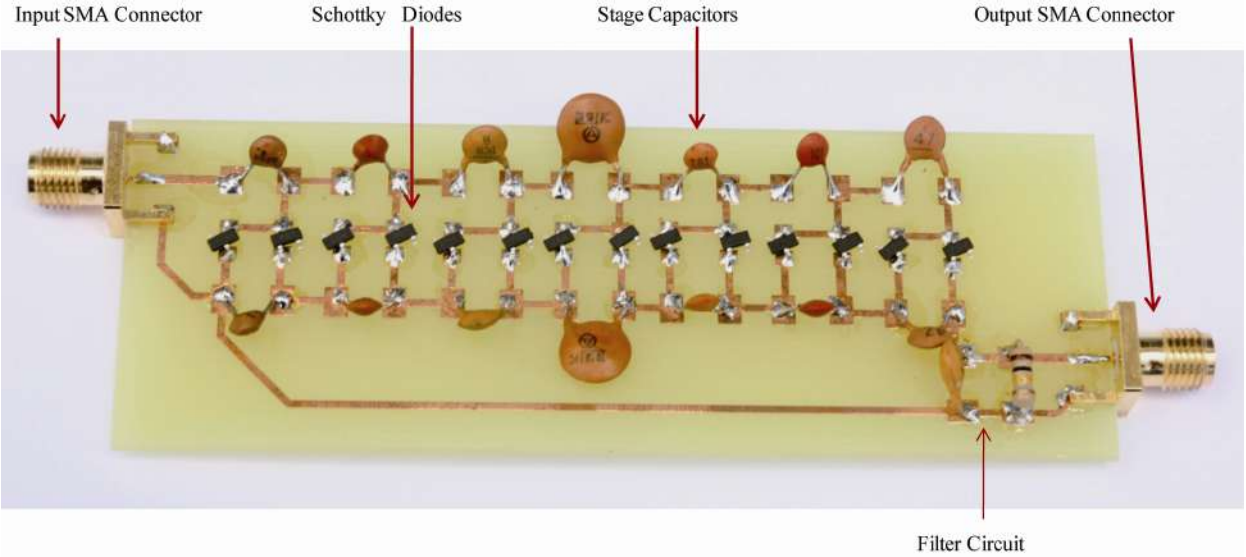

Figure 7. Photograph of assembled circuit board.

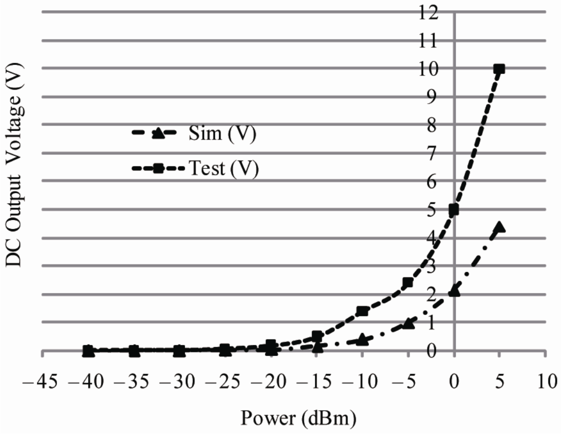

Figure 8. Simulated and test DC output voltage of multiplier as a function of input power.

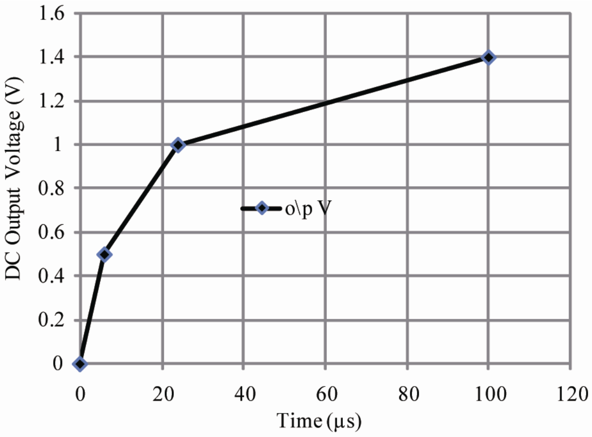

Figure 9. DC output voltage verses rise time of 4 stage voltage doubler circuit with equal stage capacitance [8].

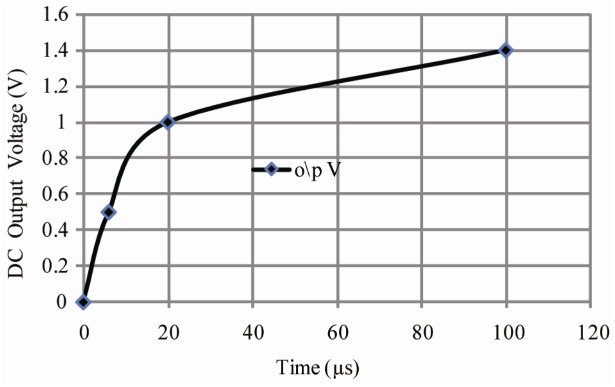

Figure 10. DC output voltage verses rise time of 4 stage voltage doubler with varied stage capacitance [8].

posed voltage multiplier circuit operates at the frequency of 945 MHz with the specified input power levels. The results have shown that there is multiplication of the input voltage. From **Figure 12**, it is shown that at 0 dBm input power, the multiplication factor is 22. This is significant, as the work shows that RF energy in the GSM-900 band can be harvested from the ambient RF source using the Villard circuit topology. The power density levels from a GSM base station is expected from 0.1 mW/m^2 to 1 mW/m^2 for a distance ranging from 25 m - 100 m. These power levels may be elevated by a factor between one and three for the GSM-900 downlink frequency bands depending on the traffic density [10]. The next phase of the research work is to interface the voltage multiplier circuit through a matching network to the antenna at the input side and a low power device to power from the system at the output side to complete the RF energy harvesting system.

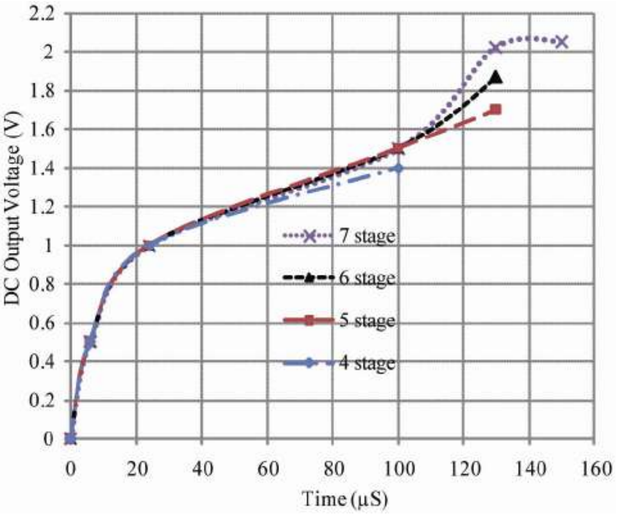

Figure 11. DC output voltage verses rise time of voltage doubler circuit through 4 - 7 stages with equal stage capacitance.

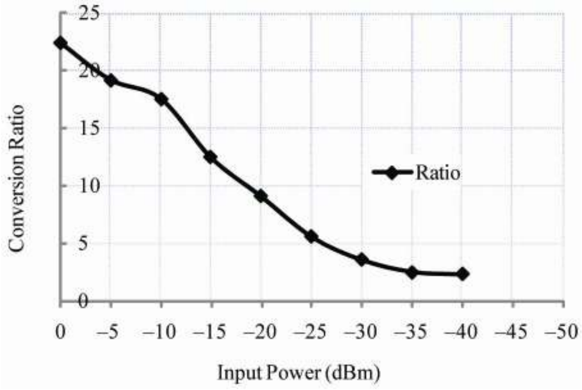

Figure 12. Conversion ratio as a function of input power.

6. Acknowledgements

We would like to acknowledge and thank the Ministry of Higher Education Malaysia for funding this project under Fundamental Research Grant: FRGS/1/10/TK/UNITEN/02/13.

REFERENCES

[1] S. von der Mark and G. Boeck, "Ultra Low Power Wakeup Detector for Sensor Networks," *SBMO/IEEE MTT-S International Microwave & Optoelectronics Conference* (*IMOC* 2007), Berlin, 29 October-1 November 2007, pp. 865-868. doi:10.1109/IMOC.2007.4404394

[2] H. Yan, J. G. M. Montero, A. Akhnoukh, L. C. N. de Vreede and J. N. Burghartz, "An Integration Scheme for RF Power Harvesting," *Proceedings of STW Annual Workshop on Semiconductor Advances for Future Electronics and Sensors*, Veldhoven, November 2005, pp. 64-66. doi:10.1.1.126.5786

[3] T. Ungan and L. M. Reindl, "Concept for Harvesting Low Ambient RF-Sources for Microsystems," 2007. http://www.imtek.de/content/pdf/public/2007/powermems_2007_paper_ungan.pdf

[4] J. Wang, L. Dong and Y. Z. Fu, "Modeling of UHF Voltage Multiplier for Radio-Triggered Wake-Up Circuits," *International Journal of Circuit Theory and Application*, Vol. 39, No. 11, 2010, pp. 1189-1197. doi:10.1002/cta.692

[5] J. A. Starzyk, Y.-W. Jan and F. J. Qiu, "A DC-DC Charge Pump Design Based on Voltage Doublers," *IEEE Transactions on Circuits and Systems-I*: *Fundamental Theory and Applications*, Vol. 48, No. 3, 2001, pp. 350-359.

[6] HSMS-2850, "Surface Mount Zero Bias Schottky Detector Diodes." http://www.crystal-radio.eu/hsms285xdata.pdf

[7] D. W. Harrist, "Wireless Battery Charging System Using Radio Frequency Energy Harvesting," M.S. Thesis, University of Pittsburgh, Pittsburgh, 2004.

[8] E. Bergeret, J. Gaubert, P. Pannie and J. M. Gaultierr, "Modeling and Design of CMOS UHF Voltage Multiplier for RFID in an EEPROM Compatible Process," *IEEE Transactions on Circuits and Systems-I*, Vol. 54, No. 10, 2007, pp. 833-837. doi:10.1109/TCSII.2007.902222

[9] B. Emmanuel, J. Gaubert, P. Pannier and J. M. Gaultier, "Conception of UHF Voltage Multiplier for RFID Circui," *IEEE North-East Workshop on Circuits and Systems*, Gatineau, June 2006, pp. 217-220. doi:10.1109/NEWCAS.2006.250961

[10] H. J. Visser, A. C. F. Reniers and J. A. C. Theeuwes, "Ambient RF Energy Scavenging: GSM and WLAN Power Density Measurements," *Proceedings of the* 38*th European Microwave Conference*, Eindhoven, 27-31 October 2008, pp. 721-724. doi:10.1109/EUMC.2008.4751554

Copyright © 2012 SciRes.

Homebrew at ACØC
Combo Analog+DSP Filter
A Few Good CW Keyers
Preamps
RF Current Meter
100W Speaker Amp
The Ultimate Audio Filter
2xSi570 RF Gen
SB200 Sleeper
SB-200 Tank Mods
High Perf Xtal Osc
CW-Skimmer Array
Leakage Tester
XR2206 Sig Gen
Fixed Bench PS
Dual Variable Bench PS
Station Interface
Speaker/Paddle Switching

RF Current Meter

Stop Guessing and Built This Simple Meter Making Measurements Easy to Take

Why a Hand Held RF Current Meter

There are two things that I wanted to do which would be made easy with a current meter... Both are related to the unusual nature of the antenna here at AC0C.

20
40
60
80
100
0
µA
SD-670
CLASS 2.0

1. Checking common mode surface current on coax and control cables up in the attic antennas.

With the high cost of ferrites, placing them with the aid of a current meter is very helpful in that you can compare the before (no ferrite) and after (with ferrites attached) - and know that you have addressed the current problem on a given bit of line.

Otherwise, it seemed to me, that ferrite placement was more of a guess than a science.

2. Measuring relative antenna currents along an element

With the W8WWV RVM system, I can measure element currents where ever a current sensor is placed. This is typically at the center of a dipole element.

However, the center may or may not be the actual high-current point for the element depending on the frequency. My EZNEC models can indicate where the current peaks are - and by looking along these lines manually - comparing them with readings taken from the driven element, I can make current checks even in non-standard location of the array.

The meter described below accomplishes these two tasks very nicely.

Design Idea

The design follows an old RSGB design that I found on G3SEK's excellent web site:

http://www.ifwtech.co.uk/g3sek/clamp-on/clamp-on.htm

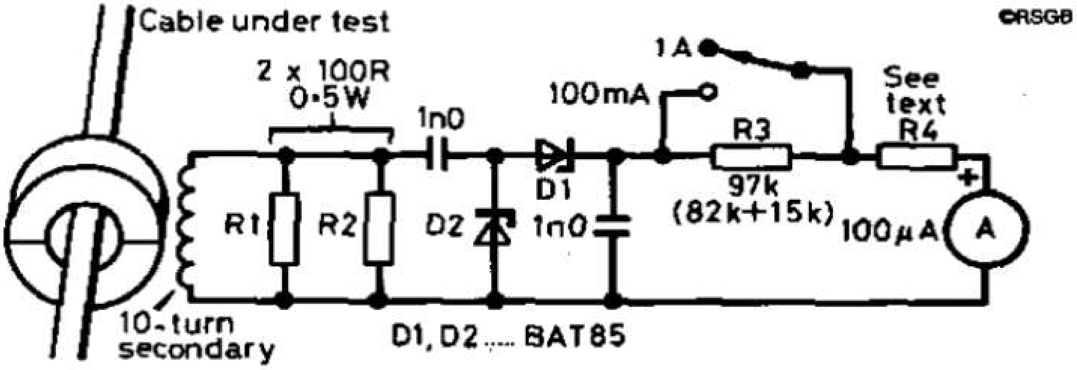

I followed this design pretty close except for the R3/R4 which I used variable trim pots. Due to the voltage-doubler nature of the circuit, it's capable of exceptional sensitivity.

The meter was calibrated for full-scale @ 1A on the higher scale, but I have left the lower scale set on a very high sensitivity setting which I find very handy for the common mode checking. Here, a sensitive meter providing a relative current indication is needed.

Construction Details

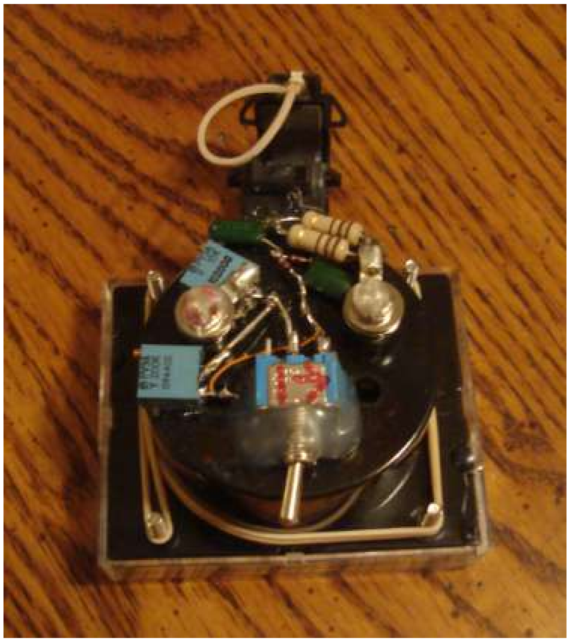

Once parts were in hand, construction was very easy and took perhaps an hour.

No attempt was made to beautify the work - it's 100% orientated toward functionality and utility.

I had a 100 uA meter in the junk box along with the other parts needed and orienting them on the back-side of the meter was really the most time consuming aspect.

The switch is held in position with hot-glue. And a bit of hot glue is applied to the meter terminations as a safety precaution - just in case I were to brush the meter up against something carrying high levels of current.

Wraps around the toroid were made with wire-wrap type wire, 30 Gage.

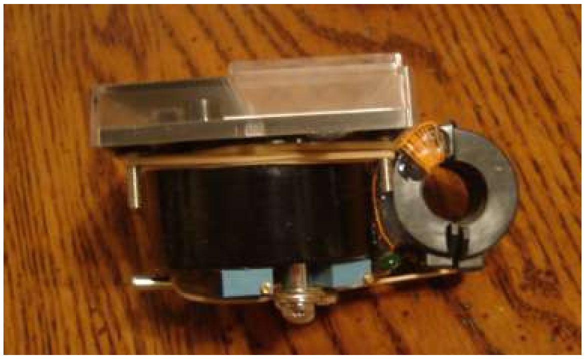

Selection of the toroid used was based on what would physically fit around the RG-213 sized cable and could be easily opened and closed.

The rubber band shown in the picture makes for a very convenient closing mechanism - and in a lot of applications, I simply hold the toroid closed with my finger pressure which makes moving the meter along a wire - and moving from cable to cable - very quick and easy.

For more ideas, check out Frank N4SPP's very nice RF current meter found HERE.

© 2011-2016 AC0C - All Rights Reserved

Created with the QTH.com SiteBuilder.

Bread board Arrangement:

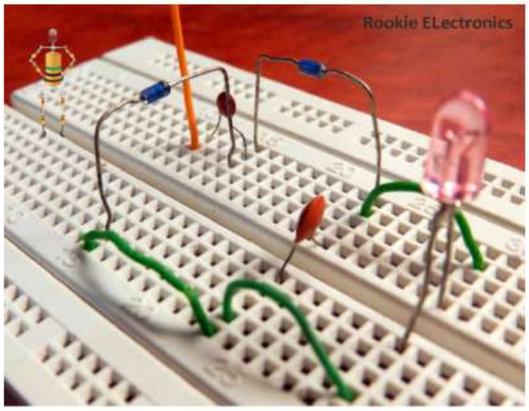
Rookie ELectronics

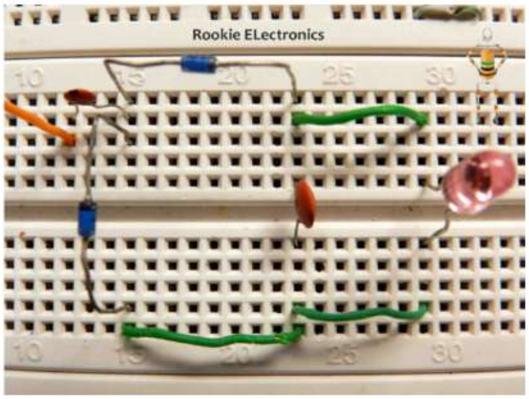
Rookie ELectronics
10
15
20
25
30

Strip Board:

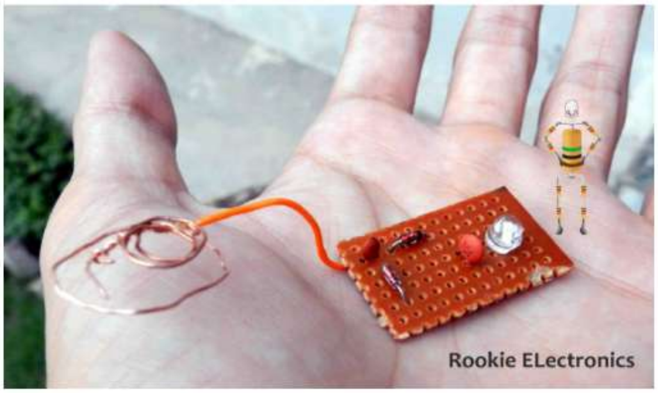
Rookie ELectronics

RF Diagnostics, LLC

WELCOME TO RF DIAGNOSTICS

Energy Harvesting Modules

WIRELESS ENERGY HARVESTING PRODUCTS RF Diagnostics, LLC designs and sells surface mount RF to DC converters and other building block modules for energy harvesting systems. With our products you can harvest ambient AC/AM/FM/Wireless Energy to power your IoT battery free system. We also provide custom designs for RF/microwave and general electrical engineering products and design contracting services to clients in the commercial wireless and defense electronics businesses. We strive to be a high quality and cost effective design resource for your design team.

OUR PRODUCTS

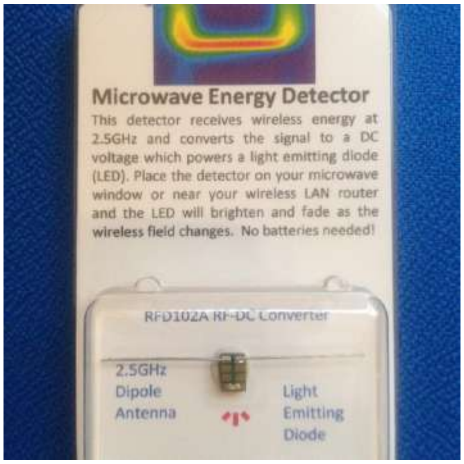

Microwave Energy Detector

This detector receives wireless energy at 2.5GHz and converts the signal to a DC voltage which powers a light emitting diode (LED). Place the detector on your microwave window or near your wireless LAN router and the LED will brighten and fade as the wireless field changes. No batteries needed!

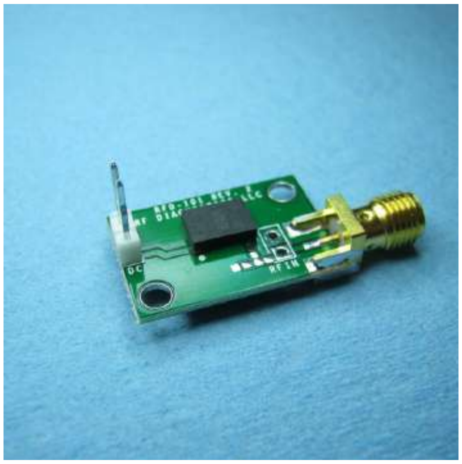

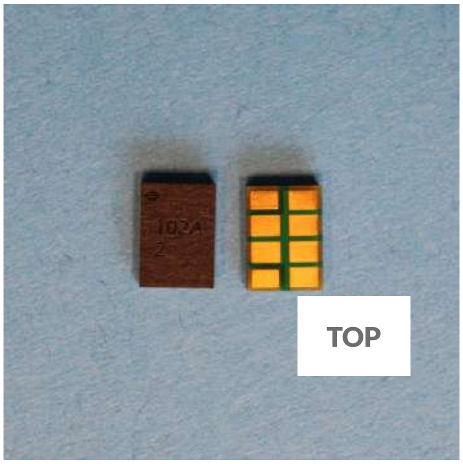

TOP

RFD102A-DET: Microwave Energy Detector

$18.00

Add to cart

RFD102A-TB: 60Hz…6GHz Energy Harvesting Test Board

$33.00

Add to cart

RFD102A: Wireless Energy Harvesting Module

$12.00

Add to cart

OUR DESIGN SERVICES

Energy Harvesting Module & System Design

Any of our standard products can be customized for your application. We can support your application with matching, testing, antenna design, assembly and many other tasks. Have a new idea for a product? We can help you realize it. We are developing new products and are always interested in your feedback to guide our product development.

MMIC Design

We use Agilent's Advanced Design System for MMIC design and PCB design. Previous designs completed are quad-band GSM power amplifier modules, 2.4-2.5GHz/5-6GHz wireless LAN power amplifiers, broadband power amplifiers, switch-filter designs, RFID detectors, RF/DC converters and filter designs.

Front End Module Design

We use Agilent's Advanced Design System and Orcad PCB Editor for Front-End Module designs. Previous design examples are high-volume switch filter modules, dual-band 802.11a/b/g front end modules, 76GHz car radar module, RFID card design.

RFID Detector System

On an NSF Phase I & II grant to one of our clients, operating as a subcontractor, the RF Diagnostics design team created a complete RFID MEMS resonator detection system. Both hardware and software were developed under this contract. The work was published at an SPIE conference in July 2010.

LEARN MORE

RF DIAGNOSTICS, LLC COPYRIGHT © 2017.

WEBSITE BY ROCK THE DEADLINE

- Home
- DA14 Asteriod
- Secret space plane
- Solar storm
- End of the world?
- UK monitoring
- Tracking
- Save money
- Get out of the EU
- UFO's
- Magnetic energy
- RF energy
- Radiant energy
- Perpetual
- Easy Steam!
- Irony- defined
- Funny Pics
- Internet Problems
- Star Wars
- Chat room
- Live music
- Kevzone
- Facebook Cheats
- Crashes

About us
Private Policy

RF energy..

This can extract some electricity from the air!

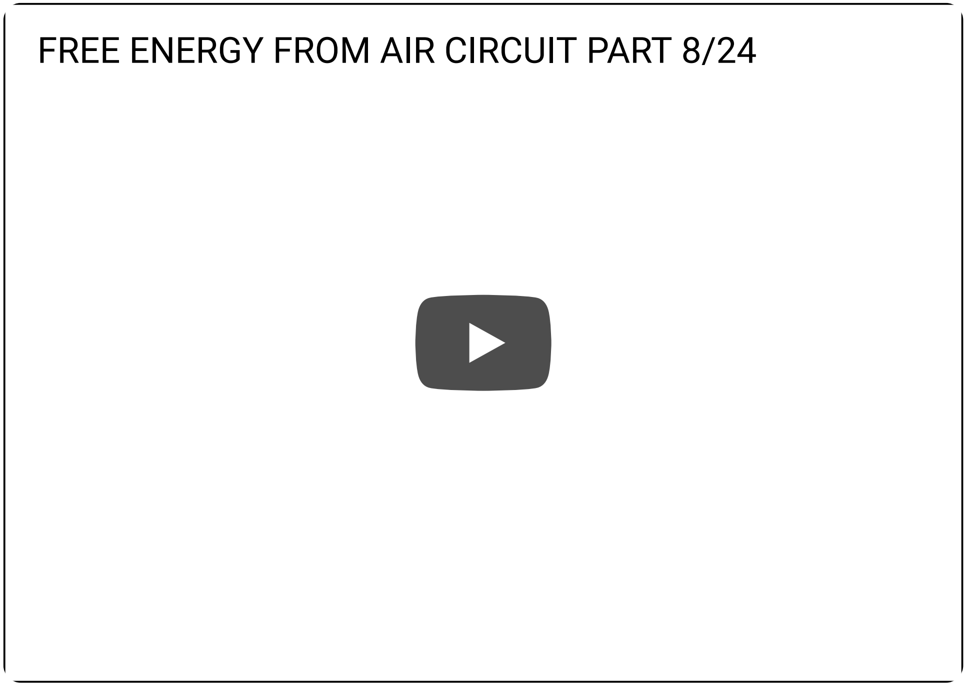

FREE ENERGY FROM AIR CIRCUIT PART 8/24

4 - Germanium Diodes (1N34)

2 - 100 µF 50V electrolytic capacitors

2 - 0.2 µF 50V ceramic capacitors

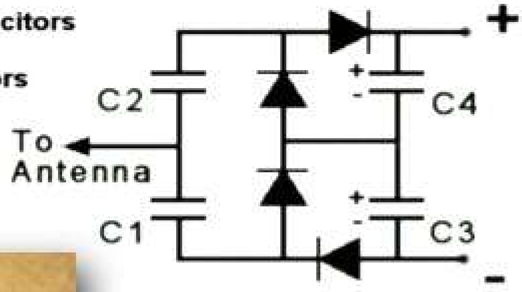

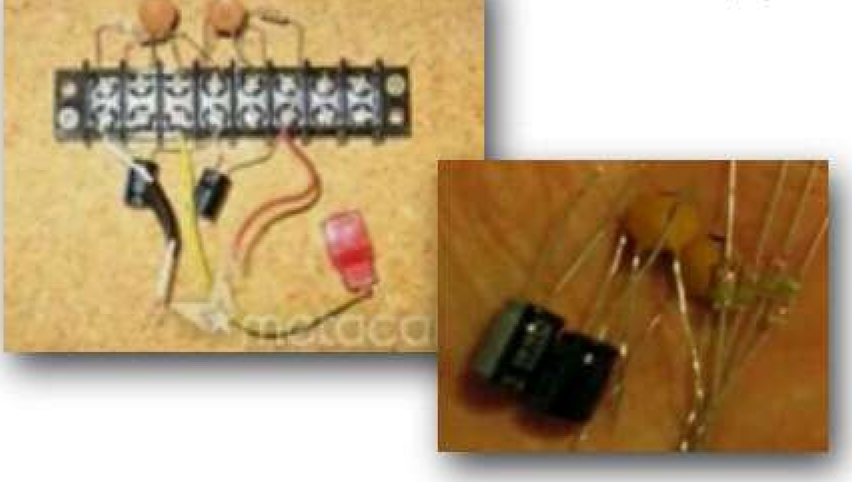

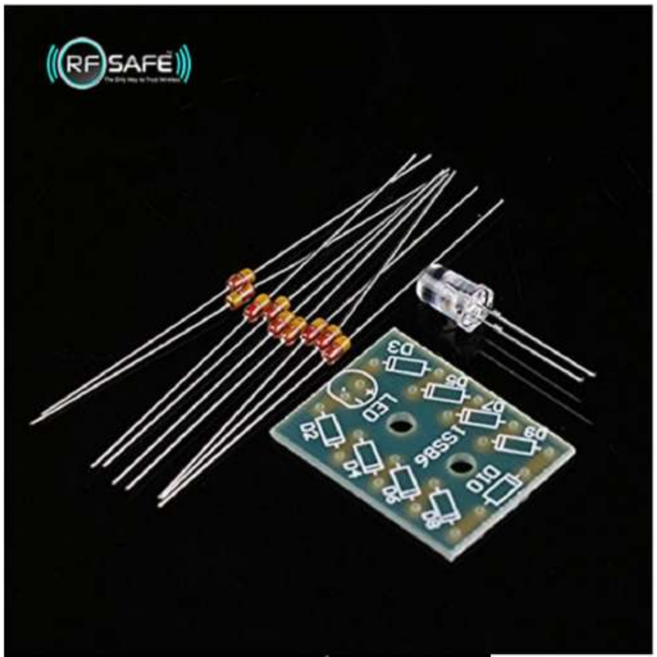

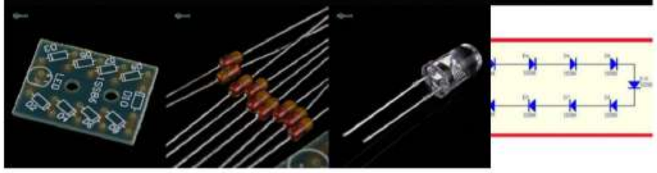

DIY RF METER: LED IS 100% POWERED BY RF RADIATION USING DETECTOR DIODES

$8.99

1 Add to cart

SAFE KIDS: LED Cell Phone Radiation Detector For DIY Science Project & Experiments || Test Phone Shields

NO batteries. it uses ONLY RF radiation power from cell phone. Yes, see with your own eyes how powerful cell phone radiation is when it powers a light from thin air. RF Safe's DIY circuit is the simplest detector for microwave radiation. The RF diodes convert microwaves from a cell phone to DC current to power an LED which emits visible light warning you of possible danger.

SKU: 636684233801 Categories: Cell Phone Radiation Safety, D-I-Y Tags: 1SS86 detector diode, Cell Phone Radiation Meter, germanium diode, LED RF Meter

A concurrent 915/2440 MHz RF energy harvester

L Fadel, L Oyhenart, R Bergès, V Vigneras, T Taris

To cite this version:

L Fadel, L Oyhenart, R Bergès, V Vigneras, T Taris. A concurrent 915/2440 MHz RF energy harvester. International Journal of Microwave and Wireless Technologies, Cambridge University Press/European Microwave Association 2016, 8, pp.405 - 413. <10.1017/S1759078716000179>. <hal-01474918>

HAL Id: hal-01474918

https://hal.archives-ouvertes.fr/hal-01474918

Submitted on 7 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A concurrent 915 MHz / 2440 MHz RF Energy Harvester

L. Fadel[1], L. Oyhenart[1], R. Bergès[1], V. Vigneras[1], T.Taris[1]

[1] Laboratoire IMS, UMR - 5218, Université de Bordeaux, 33405 Talence Cedex, France.

This paper presents the development of two dual-band RF harvesters optimized to convert far-field RF energy to DC voltage at very low received power. The first one is based on a patch antenna and the second on a dipole antenna. They are both implemented on a standard FR4 substrate with commercially off-the-shelf (COTS) devices. The two RF harvesters provide a rectified voltage of 1V for a combined power respectively of -19.5 dBm at 915 MHz and -25 dBm at 2.44 GHz and of -20 dBm at 915 MHz and -15 dBm at 2.44 GHz. The remote powering of a clock consuming 1V/5µA is demonstrated, and the rectenna yields a power efficiency of 12 %.

Keywords: Antennas and Propagation for Wireless Systems, Applications and Standards (mobile, Wireless, networks), Circuit Design and Applications.

Corresponding author: email: ludivine.fadel@ims-bordeaux.fr; phone: +33 540 002 615

I. INTRODUCTION

Today's the society is evolving toward creating smart environments where a multitude of sensors and devices are interacting to deliver an abundance of useful information. Essential to the implementation of this Internet Of Things (IOT) is the design of energy efficient solutions aiming toward a low-carbon-emission, namely green, society. Within this context, the energy harvesting appears as an alternative to provide systems with self-sustained operation. Many electronic devices operate in conditions where it is costly, inconvenient, or impossible to replace the battery. Examples include sensors for health monitoring of patients [1],[2], aircraft or building structural monitoring [3],[4], sensors in natural, industrial or hazardous environments, etc. The scavenging of natural ambient energy requires some specific conditions such as: daylight for solar energy [5], breeze for wind energy or motion for kinetic energy [6] to name a few. As consequences the exploitation of natural source does not fit with many cases of applications. On the other hand the electromagnetic (EM) [7], or Radio-Frequency (RF), energy is a human made source that is not dependent of weather conditions nor the daytime. It is so very attractive for wireless powering of remote devices. Furthermore the ever growing of commercial and personal wireless installations opens up to a 24 hour a day available energy in the vicinity of any human activity areas. The schematic of a general Wireless RF Power Transmission (WPT) system is shown in Fig.1. We talk here about far-field RF energy transmission [8], which is different from near-field RF energy transmission [9]. This later including inductive, capacitive or resonant coupling is a close contact transmission and is not relevant for remote devices. In Fig. 1, the receiver antenna collects the EM energy radiated by a RF source, and converts it into a RF signal. This RF signal is transferred to the rectifier by an impedance matching network, to be converted into DC power, which is further accumulated in a storage element. The main purpose in the deployment of WPT systems is the development of compact and efficient solutions. Most of the challenge concerns the implementation of harvesting modules, especially the antenna as its design defines the scavenging capability and the size of the RF harvester. At low frequency the transfer of energy is efficient, but the antenna footprint is large. To address the trade off between the efficiency of the WPT and the size of the modules, the frequency band located in the 433 MHz to 6 GHz frequency spectrums are preferred.

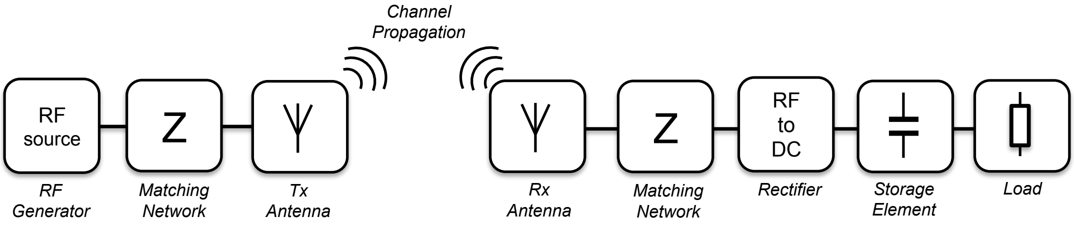

Fig. 1: Schematic of a Wireless RF Power Transmission (WPT) system

Over the last decade the research effort has focused on the development of WPT systems according two scenarios: the RF energy scavenging [10] and the RF energy transfer [11]. The two RF energy scavenging is an opportunistic collection of the RF ambient energy from the surrounding communication traffic. To improve the harvesting capability the scavenging RF harvesters are of wide-band type [12] and cover popular standards such as: DTV (470-610 MHz), GSM900, GSM1800, 3G (2.1GHz) and WiFi (2.4 GHz). Unfortunately these standards dedicated to convey wireless communications do not radiate a large RF power. As consequences the collected energy is weak, unpredictable and out of control. The RF energy scavenging remains a promising solution in the future as the increase of communication traffic could make it more reliable, and consistent with IOT applications. The second concept, namely RF energy transfer, assumes an identified source that is dedicated to perform the WPT. The amount of transmitted power is controlled by the source and the collected energy is larger than in scavenging approach. The licence-free Industry-Science-Medical (ISM) frequency bands located at 0.9, 2.4 and 5.8 GHz are usually exploited to support such a WPT scenario. Today the RF energy transfer in ISM Bands is not only promising, it becomes a reality as some pioneer companies propose some full kits: Powercast Corporation, AnSem and MicroChip to name a few. However there is still a lot of work to make the RF energy transfer an appropriate, low cost and easy-to-use solution for remote powering. One of the most critical point concerns the harvesting capability of the RF modules. So far the commercial kits referenced above only explore the 900 MHz ISM allocations to perform the WPT. This work proposes to demonstrate the interest of a concurrent harvesting at 915 MHz and 2.44 GHz. The design and implementation of a modified 4-stage doubler RF to DC converter, including a concurrent matching network, is first presented. The section III details the design of two types of multi-band antenna. The comparison between a single frequency and a multi-band WPT is exposed and the demonstration of the remote powering of a clock is reported as a case of application. To conclude a comparison of our results with the state of the art is exposed.

II. CONCURRENT RF TO DC CONVERTER

The Radio-Frequency IDentification (RFID) applications are the most popular systems exploiting the principle of RF energy transport. In passive RFID applications the reader transmits the RF power to the tag, and also sets up the communication. The RF to DC converter is designed to yield a maximum of power efficiency to the tag. Most of the time the reader and the tag are in line of sight and close to each other, these conditions improve the transmission of RF energy, the amount of power available at the tag antenna is large, typically between -15 dBm and -20 dBm. In RF energy harvesting the scenario is different. The distance between the RF source and the RF harvester ranges from 0.5 meter to 10 meters. The amount of collectable power is low, from -10 dBm to -25 dBm, and the remote powering is difficult. The RF harvesters are supposed to collect and to store the energy during a long period of time. Once the level of stored energy is large enough, it can be released to the application. For these reasons a rectifier dedicated to RF energy harvesting is first designed to yield a maximum of sensitive to increase its scavenging time and capability.

A) Rectifier Architecture

The rectifier architecture is based on voltage multipliers to provide an adequate output DC voltage. The architecture of the RF to DC converter, reported in Fig.2, includes a matching network based on a L-section, and a N-stage voltage multiplier based on Schottky diodes from Avago (HSMS285). The choice of the Schottky diode is very important in the design of the rectifier. A key parameter is its threshold voltage V_{TH}. When only low power levels are available in the environment, the amplitude of the incident signal may be close to or even below this voltage. Below this voltage value, the diode will no longer conduct and the losses become predominant. For COTS devices the two Schottky diodes performing the best conversion efficiency in a 2.4 GHz range are HSMS-2850 from Avago and SMS-7630 from Skyworks [13].

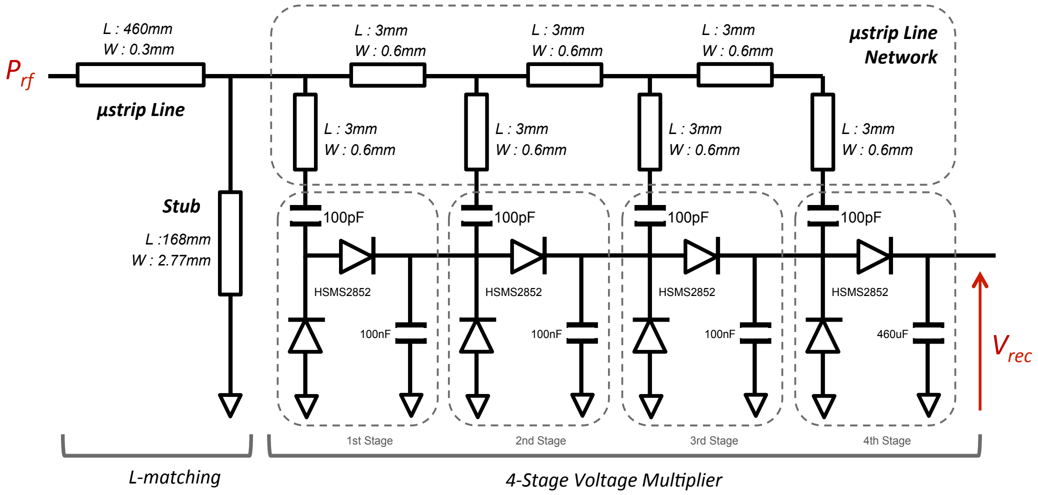

Fig. 2: Architecture of the RF to DC converter

Focusing on the sensitivity, the RF to DC converter is designed to maximize the rectified voltage for an input power close to -20 dBm. The optimum number of stage is fixed to four according [14]. The footprint of each voltage doubler imposes the micro-strip line network. The micro-strip lines namely "junction" is set to minimum length, the micro-strip lines "access" are used as an additional degree to tune the input-matching network. Indeed, the L section in combination with the micro-strip distributed network is equivalent to a T section (Fig.3). Many combinations of Z_1, Z_2, Z_3 can achieve input matching at 900 MHz or 2.4 GHz. Some of them are very close for each frequency, so we choose one that allows a return loss (< - 10 dB at least) both at 900 MHz and 2.4 GHz.

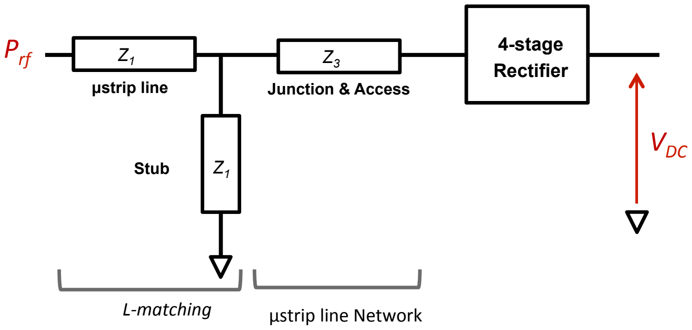
P_{rf}
Z_1
µstrip line
Z_3
Junction & Access
4-stage Rectifier
Stub
Z_1
V_{DC}
L-matching
µstrip line Network

Fig. 3: Topology of the input-matching network

The equivalent narrow band model of the matching network is proposed for each frequency (Fig.4). At 915 MHz, the voltage multiplier, including the rectification stages and the micro-strip line network, is modelled with a shunt capacitor (5pF) and a shunt resistor of 270 Ω (Fig.4a). The stub, (Fig.3) is equivalent to an inductor (Fig.4a), which compensates the shunt capacitor. The input micro-strip line, (Fig.3), is a quarter wave impedance transformer, (Fig.4a) it converts the 270 Ω into 50 Ω. At 2.44 GHz the micro-strip line network distributing the RF signal to the voltage doublers (Fig.3), becomes inductive (Fig.4b). The stub is equivalent to a shunt capacitor of 120fF, its effect is negligible. The impedance transformation is actually performed by the input micro-strip line, which is modelled by a shunt capacitor (0,6 pF) and a series inductor of 5.6 nH.

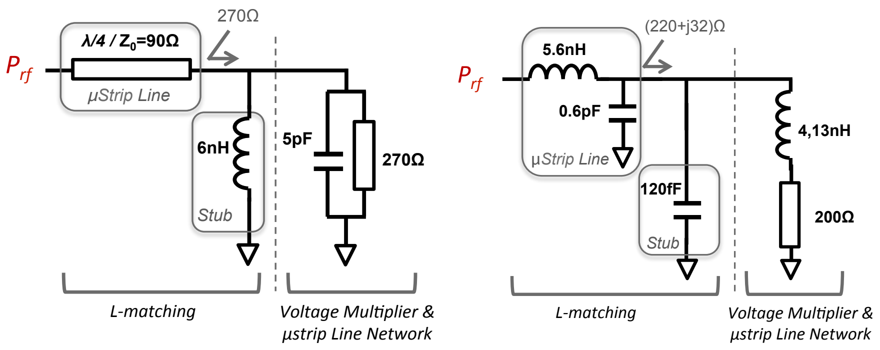
P_{rf}
$\lambda/4$ / Z_0=90Ω
µStrip Line
270Ω
6nH
Stub
5pF
270Ω
L-matching
Voltage Multiplier & µstrip Line Network
P_{rf}
5.6nH
(220+j32)Ω
0.6pF
µStrip Line
120fF
Stub
4,13nH
200Ω
L-matching
Voltage Multiplier & µstrip Line Network

(a) *(b)*

Fig. 4: Equivalent model of the input matching network at 915 MHz (a) at 2,44 GHz (b)

To study the impact of the power on the diode, and input matching, behaviour the return loss of the 4-stage rectifier has been measured and plotted (Fig.5) for various input power P_{rf} at 9000MHz.

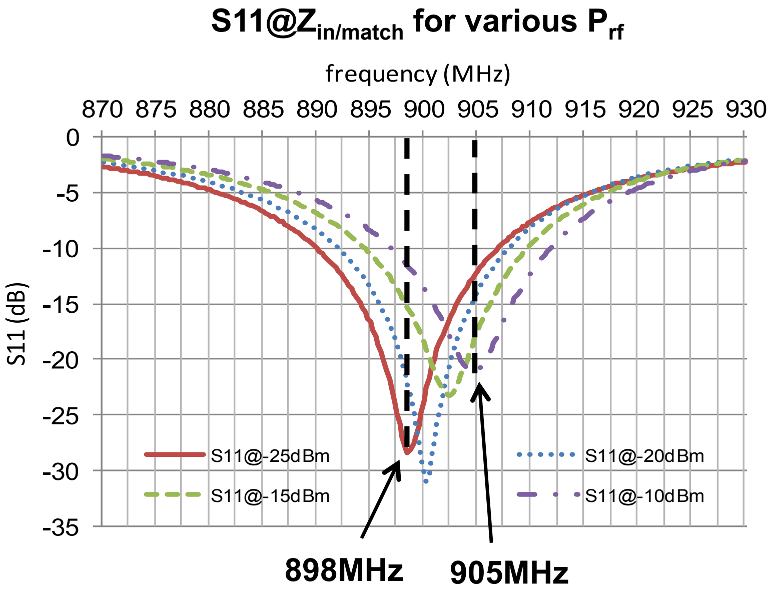

Fig. 5: Measured S_{11} of a 4-stage voltage multiplier with a L section at 900 MHz for various input power P_{rf}

As illustrated in the figure 5, the input return loss is not strongly affected by the input power if $P_{rf} < -15$ dBm. The RF harvesters developed in this work are dedicated to collect power from – 15 to – 25 dBm. Over this range the diode model can be considered as stable, and the slight frequency shift is still covered by the antenna bandwidth.

B) Rectifier Characterization

The power efficiency and the power sensitivity are two conversion characteristics of importance in RF harvesters. However, the RF harvester operating at low power level accumulates the energy in a storage element, to further release it to the application. In such accumulation mode the power sensitivity becomes more important than the power efficiency.

For the characterization the rectifier is not connected to a load. The load represents the equivalent impedance of the application (clock, sensor) to power. The effectiveness of RF-DC conversion of the rectenna and its DC ouput voltage varies depending on the load value. The rectifier is first characterized in a single tone mode, 915 MHz and 2.44 GHz respectively, and then in a dual-band mode. Measurements of the unloaded rectified voltage versus various input power P_{rf} are reported in Figure 6.

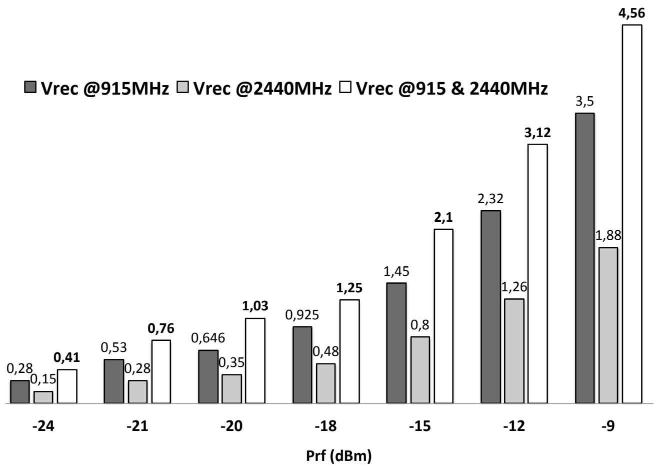

Fig. 6: Unloaded rectified voltage for various input power

To compare the results of the two considered tones, the target is fixed to a value of 1V. In a single tone mode, the required P_{rf} to rectify 1V is close to -18 dBm at 915 MHz, and would be larger than -15 dBm at 2.44 GHz. In a dual-band mode the circuit only needs a power P_{rf} of -20 dBm at each frequency. The dual-band rectification significantly improves the power sensitivity. The reverse breakdown voltage of the HSMS285 Schottky diode limits the input power to - 9 dBm, for which V_{rec} is 4,56 V.

III. ANTENNA DESIGN

To meet the low-cost constraints, the RF energy harvester will be implemented on a single low cost substrate, an FR4 PCB. For the antenna, there are more efficient substrates with a higher permittivity to reduce the size of the antenna or a lower losses but their cost is much higher than the improved performance. These powerful substrates fail to build low cost energy harvesters. This section exposes the design of two dual-band antennas implemented on a 1.6 mm FR4 printed circuit board. The fabrication uses a mechanical etching process with a 200µm resolution. We have chosen two complementary antenna topologies with a directional and omnidirectional radiation pattern.

A) Dual - band patch antenna

Emitting and receiving antennas do not usually meet the same constraints. Mobile devices such as smartphones and tablets use compact antennas (ifa, pifa…) to address the trade off between performance and size. Base stations can afford large efficient radiating elements (omnidirectional or directional antennas depending on the application). For energy harvesting purpose, micro-strip patch antennas are commonly used [15-17]. A rectangular micro-strip patch antenna (RMPA) is first developed to suit with both low cost technology of implementation and co-integration with the rectifier. Based on the cavity-model approximation, the resonant frequencies of the RMPA for the TM_{mn} mode is described in (1).

$$f_{mn} = \frac{c}{2\sqrt{\varepsilon_r}}\sqrt{\left(\frac{m}{L}\right)^2 + \left(\frac{n}{W}\right)^2} \qquad (1)$$

W, L are the patch dimensions, $c = 3.10^8$m/s

The antenna dimensions, 68 cm^2 (8.8 x 7.8 cm), described Fig.7a are dependent to the frequency bands and the feed location is selected to only excite the fundamental modes TM_{01} and TM_{10}. Those modes permit to obtain a large aspect ratio (W/L = 2,7) but reduce the performance of the RMPA. On the other hand, TM_{01} and TM_{30} modes require an aspect ratio close to one but offer beneficial radiation patterns for our application. The RMPA is fed by a probe whose position (x,y) adjusts the matching both at 915 MHz and 2.44 GHz. This two operating bands of the proposed antenna are on cross polarization planes. The geometric parameters of RMPA have been optimized with an approximate model, the TL model [18], and with a full wave method. Details of the two approaches have been studied in [19]. The return loss of the RMPA is better at 915 MHz than 2.44 GHz because the maximum impedance of TM30 mode is 31 Ω [19]. The TM30 mode does not achieve 50 Ω because it is not a fundamental mode. This antenna has a maximum gain of 1.3 dB at 915 MHz (Fig.7b) and 2.5 dB at 2.44 GHz (Fig.7c). This two operating bands of the proposed antenna are on cross polarization planes.

The realized gains of the dual band patch antenna are lower than the classical patch antenna because the radiating efficiency is low, 60% for the TM_{01} mode and 30% for the TM_{30} mode. The FR4 substrate has a loss tangent of 0.02. The radiation efficiency of the dual band patch antenna is highly dependent of the substrate losses.

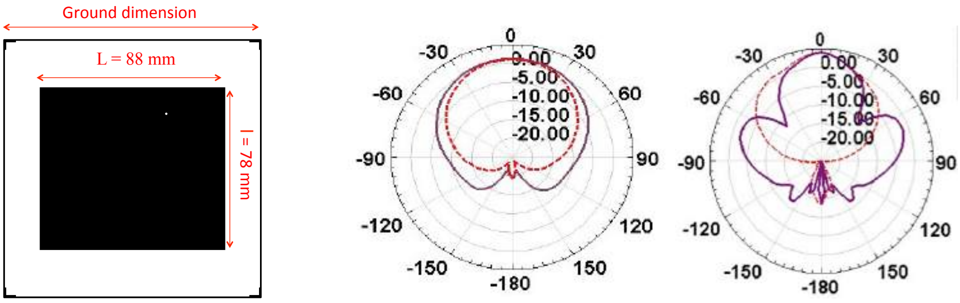

(a) *(b)* *(c)*

Fig. 7. Layout of the RMPA antenna (a) - Radiation pattern at 915MHz (b) and 2.44 GHz (c) Solid and dashed lines correspond to E-plane and H-plane respectively

B) Multi-band arm dipole antenna

The second antenna is a multi-band dipole type composed of three arms. Its dimension is about 23 cm^2 (11.1x2.1 cm), Fig.8a. Each arm is designed to work at one band of frequency. The longer one is for the 915 MHz, the middle one, not useful in our case, is for the 1.4 GHz and the last one, the smaller, is dedicated to 2.4 GHz [20].

All the geometric parameters have been optimized with a full-wave method in order to be matched both at 915/ 2440MHz. On Fig.6b and 6c, the radiation pattern is plotted for the elevation plane (orthogonal to substrate) of the simulated antenna. The maximum gain is 0.5 dB at 915 MHz and 3.4 dB at 2.44 GHz at 90°, on the substrate plane. The radiation efficiency is 99% at 915 MHz and 95% at 2.44 GHz.

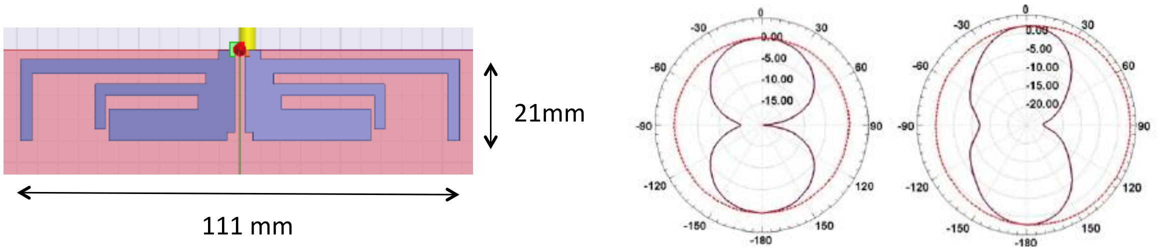

(a) *(b)* *(c)*

Fig. 8. Layout of the multi-band arms dipole antenna (a) - Radiation pattern at 915 MHz (b) and 2.44 GHz (c). Solid and dashed lines correspond to E-plane and H-plane respectively

It is interesting to compare the characteristics and performances of the two types of antennas. Although the radiation efficiency of the dipole antenna is better than the patch antenna, the antenna gains are similar because the high directivity of the RMPA antenna compensates the low values of radiation efficiency. When there are no cost constraints, it is interesting to use high performance substrates for the design of RMPA antennas because they improve the radiation efficiency and consequently antenna gain.

Moreover, the integration of the antenna with the rectifier will not be made in the same way. Considering the patch antenna, the rectifier can be integrated on the ground plane allowing a more compact solution. The dipole antenna, which is ground plane free, is less sensitive to the surrounding environment in our case. The performance of the dipole antenna and especially the radiation efficiency are very weakly dependent of the substrate characteristics. The design of a dipole antenna can be easily reuse with other material such as Kapton®, paper, Plexiglas to name a few.

IV. WIRELESS POWER TRANSMISSION

This part presents the measurement results of the assembled RF harvesters in the context of Wireless Power transfer. The two dual-band harvesters are realized with COTS devices such as HSMS diodes and capacitors. Those elements are reported by heat-treating. The RF to DC converter board, including the matching network and the rectifier, is reported on the backside and connected to the radiation part, on the front side, through a via (Fig.9a). The dipole antenna is connected to the rectifier circuit using SMA connector (Fig.9b).

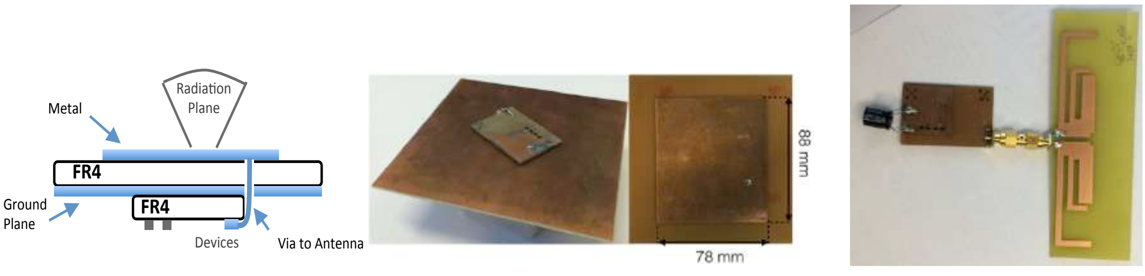

(a) *(b)*

Fig. 9. Dual-band RF harvesters based on patch antenna (a) – arms dipole antenna (b)

For the dual-band RF harvester based on patch antenna, the return loss, S_{11}, is measured for an input power of -20 dBm with a HP8720 network analyser. The patch antenna, the rectifier and the dipole antenna are centered at 915 MHz and 2.44 GHz with a low return loss ($S_{11} < -15$ dB), Fig.10. The return loss of the RMPA antenna is better at 915 MHz than 2.44 GHz because the maximum impedance of TM30 mode is 31 Ω (Fig. 6 and Fig. 7 of [16]). The TM30 mode does not achieve 50 Ω because it is not a fundamental mode.

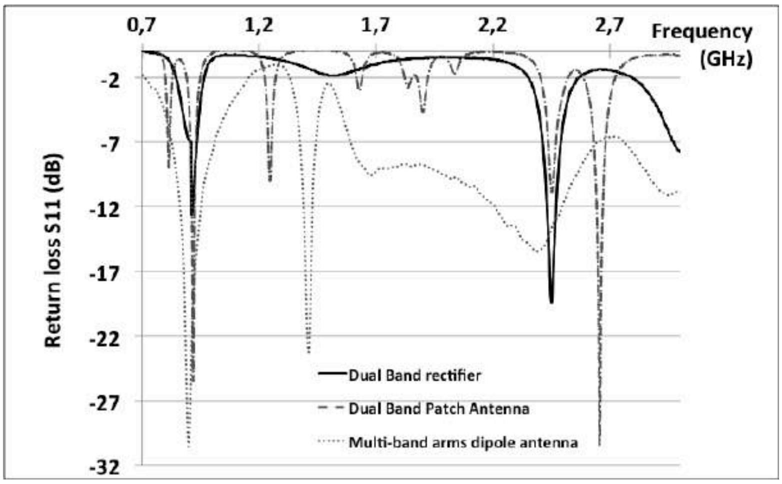

Fig. 10. Measured return loss S_{11} of the dual-band rectfier, patch and arms dipole antenna

A) Remote Powering and Power Efficiency

The rectenna is connected to a clock, which mimics a low power application. The remote powering of this clock is performed in a furnished room of the lab according the schematic of Fig.11. The distance between the source and the antenna is fixed to 2 m. The clock is turned on for different scenarios of transmitted power. For each combination of power proposed in Fig.12, the RF power is first measured with a calibrated antenna and a power meter. Then, the rectenna is measured and P_{eff} is the ration between the power delivered to the load (here the clock) and the power available at the antenna.

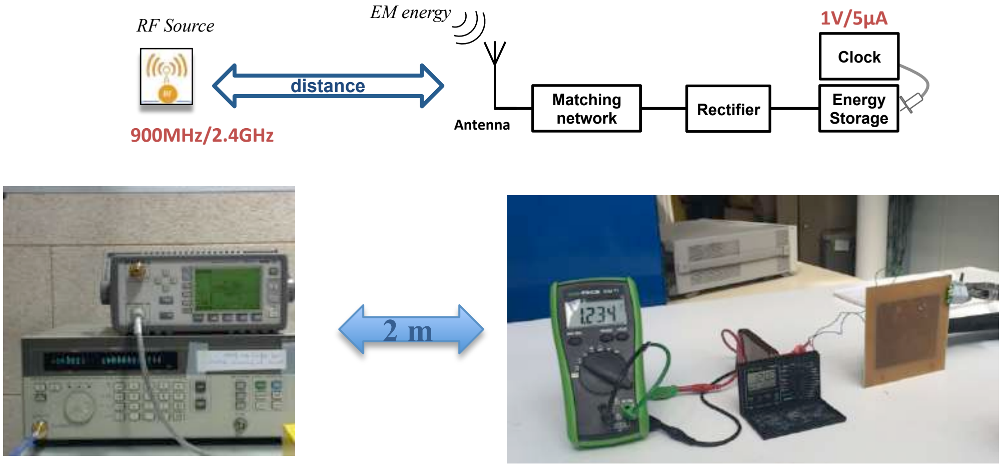

Fig. 11. Schematic and picture of the scene of remote powering of a clock

The power efficiency of the patch and the dipole rectenna is worked out from these experiments and reported in Fig.12. The power efficiency η is defined as the ratio between the DC power delivered to the clock and the RF power collected by the antenna.

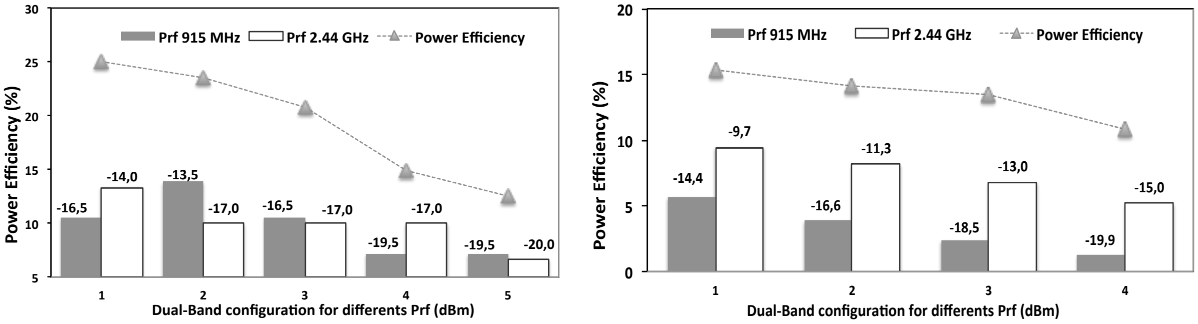

(a) *(b)*

Fig. 12. Power efficiency of the dual-band RF harvester based on the patch (a) and the arms dipole (b) antenna

The minimum power required to turn on the clock with the patch-based harvester, Fig. 10.a, is a two tone signal featuring: -19.5 dBm at 915 MHz and -20 dBm at 2.44 GHz. At this point, the power efficiency is 12.5 %, which corresponds to a DC output power of 2.7 µW/1V.

A maximum efficiency of 24 % occurs for a combined power of -16.5 dBm at 915 MHz and -14 dBm at 2.44 GHz. The harvester is able to deliver a DC ouput power of 15 µW. The harvester based on the arm dipole antenna needs a minimum power of -19.9 dBm at 915 MHz and -15 dBm at 2.44 GHz at the antenna to turn on the clock. For these conditions of remote powering, the efficiency of the harvester is 11 %. It delivers a DC power of 3.8 µW/1.15 V. The maximum power efficiency, 15.5 %, yields for an input power of -14.4 dBm at 915 MHz and -9.7 dBm at 2.44 GHz, the DC output power is 21 µW.

This scenario of remote powering figures out that the harvester based on the patch antenna exhibits a better power efficiency than the harvester combined with the dipole element. This difference is due to the antenna gains. Referring to Fig. 7 and Fig. 8, the gain of the patch antenna is larger (+0.8dB) at 915 MHz and lower (-0.9 dB) at 2.44 GHz than the dipole element. However the rectifier, referenced in [16], achieves a power efficiency of 17% at 915 MHz and only 5% at 2.44 GHz for an input signal of -15 dBm. As consequences the patch-based harvester is able to extract more power from a 915 MHz signal than the dipole-based harvester can do at 2.44 GHz. For this reason the overall efficiency of the patch harvester is better.

B) Power Sensitivity

The power sensitivity is measured with the same scenario of Fig.11 but the clock is disconnected. The output voltage is reported for different combination of collectable power at the antenna in a dual-band configuration.

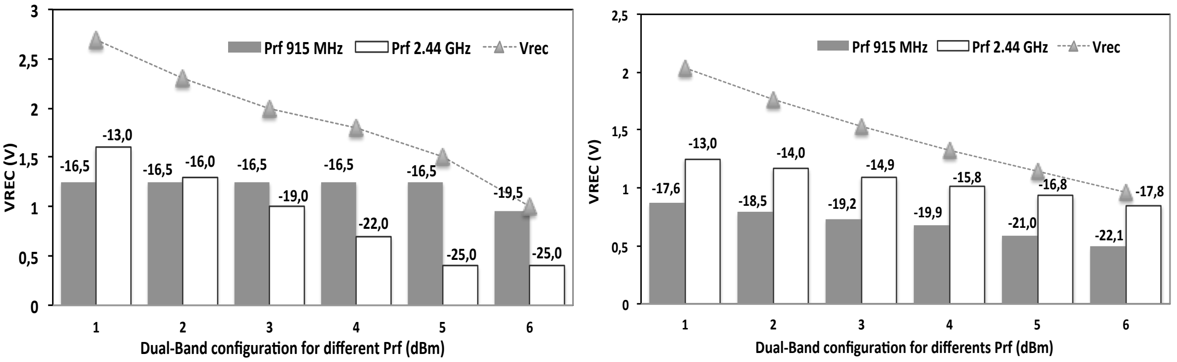

(a) *(b)*

Fig. 13. Rectified voltage of the RF harvester based on patch (a) and arm dipole (b) antenna

To rectify a 1V DC voltage, the patch-based harvester, Fig.13a, requires a two dual-band configuration: -19.5 dBm at 915 MHz and -25 dBm at 2.44 GHz, which is equivalent to an input power of -18.4 dBm (or 14 µW). For the same purpose the dipole-based harvester, Fig.13b, needs a dual-tone of -22.1 dBm at 915 MHz and -17.8 dBm at 2.44 GHz. The equivalent input power of this 2-tone signal is -16.5 dBm (or 22 µW). The patch-based harvester exhibits a better sensitivity than the dipole harvester for the same reason exposed in the part A of this section. In Fig.6, which reports the power sensitivity of the rectifier part only, the overall sensitivity is almost the same for the dipole harvester. It is improved by 1.8 dB for the patch harvester due to the additional gain of the antenna at 915 MHz.

C) Discussion and Comparison with the state of the art

An important characteristic of a remote powered device is its size. Indeed it is expected to be as small as possible to make it unobtrusive to our closest environment. In a scenario of RF harvesting the antenna footprint determines the compactness of a harvester operating Ultra High Frequency (UHF) bands. To complete the comparison between the two harvesting modules developed in this work, two figures of merit, FOM_{sens} and FOM_{eff}, including the size of the antenna, are proposed in (2) and (3).

$$FOM_{sens} = \frac{V_{REC@Psens}(V)}{\frac{P_{sens}(\mu W)}{100\mu W} \cdot \frac{A_{ant}(cm^2)}{100cm^2}} \tag{2}$$

With: P_{sens} the input RF power required to provide $V_{REC@Psens}$ the unloaded rectified output DC voltage, and A_{ant} the area of the antenna.

$$FOM_{eff} = \frac{\eta(\%)}{\frac{P_{eff}(\mu W)}{100\mu W} \cdot \frac{A_{ant}(cm^2)}{100cm^2}} \tag{3}$$

With: P_{eff} the input RF power required to achieve η the overall power efficiency.

FOM_{sens} and FOM_{eff} do not represent the same scenario of application. The FOM_{sens} illustrates the capability of the rectenna to start collecting energy and store it in an element such as a capacitor or a battery to further release it. FOM_{eff} demonstrates the capability of the RF harvester to yield "on time powering": the rectenna is connected to an application and supply it on time. Both are reported in the Table I, which also includes some references of the state of the art. The ability of the proposed rectenna to simultaneously operate in two frequency bands, significantly improves the power sensitivity.

Table I. Comparison with the state of the art

Ref.	Freq (GHz)	Efficiency (%@P_{rf})	Sensitivity (V_{rec}@Prf)	Number of stage	Schottky diodes	Size (cm^2)	FOM Sens	FOM Eff
[21]	0.9	15% @ -10dBm	0.75V @ -10dBm	1	SMS-7630	15×15	1.9	6.6
[21]	2.4	9% @ -13dBm	0.9V @ -13dBm	1	SMS-7630	15×15	1.25	8
[22]	2.45	10.5% @ -20dBm	0.075V @ -20dBm	1	SMS-7630	3.4×3.4	6.5	905
[23]	0.915/ 2.45	14%@ -20/-20	0,36V@ -10/-10	1	SMS-7630	6×6	0.5	185
[24]	1.8/2.2 /2.5	55%@ -10dBm	300mV@ -32dBm 3tones	1	SMS-7630	7×7	20	112
This work Dipole	0.915/ 2.44	**11% @ -20/-15dBm**	**1V @ -22/-18dBm**	4	HSMS-2850	2.1×11	**19.8**	**136**
This work Patch	0.915/ 2.44	**12.5% @ - 19.5 dBm/ -20dBm**	**1V @ -19.5dBm/ -25dBm**	4	HSMS-2850	7.8×8.8	**10.5**	**100**

According the Table I, the patch-based harvester exhibits the highest sensitivity to rectify 1V with a dual-tone featuring: -19.5 dBm at 915 MHz and only -25 dBm at 2440 MHz. The FOM_{sens} represents the trade-off between the sensitivity performances of a rectenna and the antenna area. The FOM_{eff} rates the efficiency performances to the antenna area. For these two figures of merit, the rectenna based on the multi-arm dipole element yields the best trade-off, both for FOM_{sens} and FOM_{eff}, compared to the patch-based solution. This dual tone and multi-arm dipole harvester is close to the work proposed in [24] which exhibits the highest FOM_{sens} reported so far in the literature to our knowledge.

V. CONCLUSION

The range of power collectable in a scenario of RF harvesting varies from -15dBm to -25dBm. To address this purpose the rectenna proposed in this work are optimized to operate at a RF input power close to -20 dBm (or 10 µW). To further improve the ability to collect the RF energy, these rectenna, developed with Schottky diodes HSMS285 from Avago, perform a concurrent harvesting in the 915 MHz and 2.44 GHz ISM bands. The harvester including a patch antenna implemented on a 1.6mm FR4 PCB achieves the highest sensitivity. It provides a 1V-rectified voltage for a dual-tone excitation of -19.5 dBm at 915 MHz and -25 dBm at 2.44 GHz. For these conditions of operation the rectenna yields a power efficiency of 12.5%. To take into account the dimensions of the haverster, two figures of merit, FOM_{sens} and FOM_{eff} including the size of the antenna, respectively related to the power sensitivity and the power efficiency are proposed. The rectenna developed with the arm dipole element exhibits the highest figures of merit. A case of application is proposed with the remote powering of a digital clock consuming 1V/5µA. The patch based harvester turns on the device with a dual tone excitation at the antenna of -19.5 dBm at 915 MHz and -20 dBm at 2.44 GHz. For the same scenario the harvester connected to the multi-arm dipole element needs a power of -22.1 dBm at 915 MHz and -17.8 dBm at 2.44 GHz.

REFERENCES

[1] T. Paing, A. Dolgov, J. Shin, J. Morroni, J. Brannan, R. Zane, and Z. Popovic, “Wirelessly powered wireless sensor platform’’ in Eur. Microw. Conf. Dig, Munich, Germany, Oct. 2007, pp. 241-244.

[2] J. Bernhard, K. Hietpas, E. George, D. Kuchima, and H. Reis, ‘‘An interdisciplinary effort to develop a wireless embedded sensor system to monitor and assess corrosion in the tendons of pre-stressed concrete girders,’’ in Proc. IEEE Top. Conf. Wireless Commun, 2003, pp. 241-243.

[3] C. Walsh, S. Rondineau, M. Jankovic, G. Zhao, and Z. Popovic, ‘‘A conformal 10-GHz rectenna for wireless powering of piezoelectric sensor electronics,’’ in IEEE MTT-S Int. Microw. Symp. Dig, Jun. 2005, pp. 143-146.

[4] X. Zhao, T. Qian, G. Mei, C. Kwan, R. Zane, C. Walsh, T. Paing, and Z. Popovic, ‘‘Active health monitoring of an aircraft wing with an embedded piezoelectric sensor/actuator network,’’ Smart Mater. Struct, vol. 16, no. 2007, Jun. 2007, pp. 1218-1225.

[5] K. Lin, J. Yu, J. Hsu, S. Zahedi, D. Lee, J. Friedman, A. Kansal, V. Raghunathan, and M. Srivastava, “Heliomote: Enabling long-lived sensor networks through solar energy harvesting,” in 3rd Int. Conf. Embedded Networked Sensor Syst., Nov. 2005, p. 309.

[6] J.A. Paradiso and T. Starner, “Energy scavenging for mobile and wireless electronics,” IEEE Pervasive Computing, Vol. 4, and Issue: 1, Jan-March 2005, pp. 18- 27.

[7] K. R. Foster, “A world awash with wireless devices”, IEEE Microwave magazine, vol. 14, Issue 2, March 2013, pp. 73-84.

[8] T. Le, K. Mayaram, and T. Fiez, “Efficient far-field radio frequency energy harvesting for passively powered sensor networks,” IEEE J. Solid-State Circuits, vol. 43, no. 5, May 2008, pp. 1287-1302.

[9] J. Hirai, T.W. Kim, and A. Kawamura, ‘‘Wireless transmission of power and information for cableless linear motor drive,’’ IEEE Trans. Power Electron, vol. 15, 2000, pp. 21-27.

[10] H. Nishimoto, Y. Kawahara, and T. Asami, “Prototype implementation of ambient RF energy harvesting wireless sensor networks,” IEEE Sens. J., Nov. 2010, pp. 1282-1287.

[11] BN. Shinohara, “Power without wires,” IEEE Microw. Mag, vol. 12, no. 7, Dec. 2011, pp. S64-S73.

[12] M. Pinuela, P.D. Mitcheson, S. Lucyszyn, “Ambient RF Energy Harvesting in Urban and Semi-Urban Environments”, IEEE Trans. On Microwave Theory and Techniques, vol. 61, N°7, July 2013.

[13] Simon Hemour, Ke Wu “Radio-Frequency Rectifier for Electromagnetic Energy Harvesting : Developement Path and Future Outlook”, Proceeding of the IEEE, Vol.102, No.11, November 2014.

[14] T. Taris, L.Fadel, L.Oyhenart, V. Vigneras, “COTS-Based Modules for Far-Field Radio Frequency ”, IEEE NEWCAS, Paris, France, June 2013, pp. 1-4.

[15] Sika Shrestha, Sun-Kuk Noh, and Dong-You Choi, « Comparative Study of Antenna Designs for RF Energy Harvesting » International Journal of Antennas and Propagation Volume 2013, 10 pages.

[16] Z. W. Sim, R. Shuttleworth, M. J. Alexander, B. D. Grieve, « Compact patch antenna design for outdoor RF energy harvesting in wireless sensor networks, Progress In Electromagnetics Research, Vol. 105, 273–294, 2010.

[17] Naimul Hasan, Santu Kumar Giri, « Design of low power RF to DC generator for energy harvesting application », International Journal of Applied Sciences and Engineering Research, Vol 1, Issue 4, 2012.

[18] H. J. Visser, "Approximate Antenna Analysis for CAD", Wiley, 2009.

[19] R. Berges, L.Fadel, L. Oyhenart, V. Vigneras, T. Taris, “A dual Band 915MHz/2.44GHz RF Energy Harvester”, EUMW, Paris France, Sept 2015.

[20] Y-Y.Lu, J. Guo, H-C. Huang, “Design of Triple Symmetric Arms Dipole Antenna for 900/1800/2450MHz Applications”, Conf on Intelligent Information Hiding and Multimedia Signal Processing”, 2014.

[21] D. Masotti, A. Costanzo, M. Del Prete, V. Rizzoli, “Genetic-based design of a tetra-band high-efficiency radio-frequency energy harvesting system,” IET Microwaves, Antennas Propagation, vol.7, no. 15, June 2013, pp. 1254-1263.

[22] G. Vera, A. Georgiadis, A. Collado, and S. Via, “Design of a 2.45GHz rectenna for electromagnetic (EM) energy scavenging”, Proc. IEEE Radio and Wireless Symp, 2010, pp. 61-64.

[23] K. Niotaki, S. Kim, S. Jeong, A. Collado, A. Georgiadis, and M. Tentzeris, “A compact dual-band rectenna using slot-loaded dual band folded dipole antenna,” IEEE Antennas and Wireless Propagation. Lett. vol. 12, pp. 1634–1637, 2013.

[24] C. Song, Y. Huang, J. Zhou, J. Zhang, S. Yuan, and P. Carter, “A High-efficiency Broadband Rectenna for Ambient Wireless Energy Harvesting”, IEEE Transactions on Antennas and Propagation, accepted May 2015.

Bibliographies

Ludivine FADEL received the Ph.D in electrical engineering from the University of Bordeaux, France, in 2004. She is an Associate Professor at Bordeaux University and her main research interests include wireless power transfer and energy harvesting, materials and printing technologies for RF application.

Laurent Oyhenart received the Ph.D. degree in electrical engineering from the University of Bordeaux, Bordeaux, France, in 2005. He is currently an Associate Professor with the University of Bordeaux and develops his research activities in the IMS laboratory. His research interests include computational electromagnetics, photonic crystals and antenna synthesis.

Romain Bergès is a PhD student at IMS Laboratory. He received his Master from Bordeaux University in Electronic, in 2014. His current research focuses on RF Energy Harvesting.

Valérie Vigneras is a full professor in the Polytechnic Institute of Bordeaux and conducts research in the IMS Laboratory on complex structures (mixtures, structured media, self-assembled media) both by simulation to predict or optimize their properties and by characterization before their integration in high frequency systems (antennas, electromagnetic compatibility, radar shieldness…).

Thierry Taris is a full professor at the Bordeaux Institute of Technolgy (Bx-INP). He joined the IMS lab in 2005 where is research interests are related to the Radio-frequency Integrated Circuits in Silicon technologies.

List of figures and tables

Fig. 1: Schematic of a Wireless RF Power Transmission (WPT) system.
Fig. 2: Architecture of the RF to DC converter.
Fig. 3: Topology of the input-matching network.
Fig. 4: Equivalent model of the RF to DC converter, 915 MHz (a) - 2,44 GHz (b).
Fig. 5: Measured S_{11} of a 4-stage voltage multiplier with a L section at 900 MHz for various input power P_{rf}.
Fig. 6: Unloaded rectified voltage for various inpout power.
Fig. 7. Layout of the RMPA antenna (a) - Radiation pattern at 915MHz (b) and 2.44 GHz (c) Solid and dashed lines correspond to E-plane and H-plane respectively.
Fig. 8. Layout of the multi-band arms dipole antenna (a) - Radiation pattern at 915 MHz (b) and 2.44 GHz (c). Solid and dashed lines correspond to E-plane and H-plane respectively.
Fig. 9. Dual-band RF harvesters based on patch antenna (a) – arms dipole antenna (b).
Fig. 10. Measured return loss S_{11} of the dual-band rectfier, patch and arms dipole antenna.
Fig. 11. Schematic and picture of the scene of remote powering of a clock.
Fig. 12. Power efficiency of the dual-band RF harvester based on the patch (a) and the arms dipole (b) antenna.
Fig. 13. Rectified voltage of the RF harvester based on patch (a) and arm dipole (b) antenna.
Table I. Comparison with the state of the art.

Fun with Less Kilowatts: The Lectenna

By Vernon Trollinger, March 14, 2017, Energy Efficiency, Events & Fun, Family

Welcome to Fun with Less Kilowatts! We believe that science experiments at home can be a creative way to engage kids in learning while having fun. They can be educational AND great activities to keep your kids busy and away from the television. Each month, we'll feature a new science experiment that can be a great resource for parents and teachers.

The Lectenna

Did you know you that the energy from radio waves can be converted back into electric energy? It's true and you can make a *simple* two-component circuit that you and your kids can use to find 2.45 Ghz radio waves around your home. These are the same radio waves used by microwaves, WiFi network routers, bluetooth devices, smart phones, cordless phones, smart meters, and smart home systems that use Zigbee.

YouTuber pjaffeva posted this brilliant project showing how. Plus, there's also a link to a PDF set of instructions on Google Drive that you can download.

The Materials

- One HLMP-D150 Avago low current LED. Costs about 50¢/each. Available from: Amazon, Digikey, Mouser
- One Hitachi 1SS106 low capacitance Schottky diode. Pjaffeva recommends this Hitachi-made diode as being the only one in this shape that will work. But they're also hard to find. Currently, you can get them from LittleDiode's ebay store in the UK. Costs $6.38 but the bulk of that price is shipping from the UK.

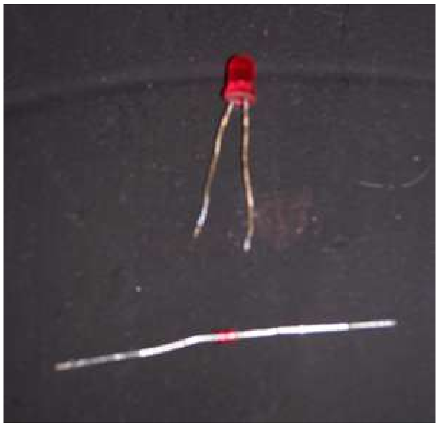

You'll also need:

- An operating WiFi router
- A bamboo kabob skewer
- Two pieces of tape

The Directions

1) Look on the base of the LED and find the side with the flat edge. This marks the negative or cathode side. Carefully splay the two wire leads away from each other as far as they will go. Make sure the one adjacent to the negative side gets bent over the flat edge so identify it. You can also mark it with a little spot of colored nail polish.

Categories

- Careers/Jobs
- Charity & Community
- Energy Efficiency
- Events & Fun
- Family
- FAQs
- Green
- Home Improvement
- Hurricane Prep
- Moving
- News
- Recipes
- Save Money
- Small Biz
- Special Offers

Subscribe via Email

Subscribe

Subscribe via Email

Subscribe

Subscribe

Subscribe in a reader

Archives

- December 2017
- November 2017
- October 2017
- September 2017
- August 2017
- July 2017
- June 2017
- May 2017
- April 2017
- March 2017
- February 2017
- January 2017

all archives

Featured Articles

Prepare for Christmas With Green Cleaning Methods

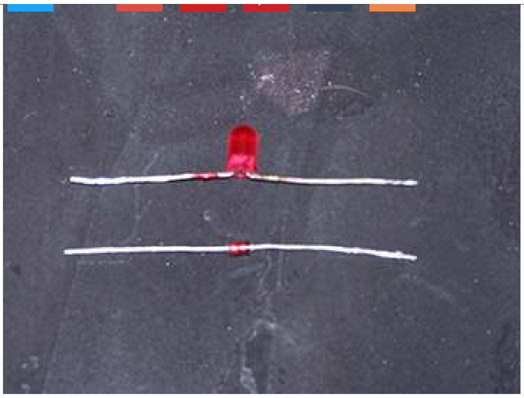

2) Next, find the band that runs around the Schottky diode. This identifies the negative end of this diode. The other end is the positive or anode end.

Using your fingers:

3) Twist the end of the negative or cathode wire from the LED onto to the positive or anode end of the Schottky diode.

4) Then twist the end of the negative or cathode end of the Schottky diode onto to the positive or anode end of the LED.

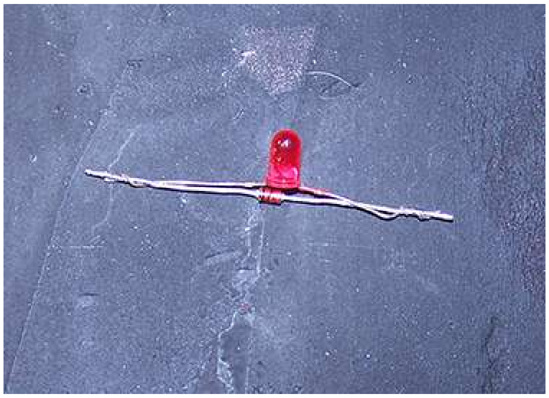

You have just made a "rectenna" and you'll need to keep the assembly and the wires straight as possible. The length of the wires should be about 6 cm long (a hair short of 2 3/8 inches).

5) Tape the two diodes to one end of the bamboo skewer.

6) With a WiFi router on and running, hold the diode near an antenna on the WiFi router.

The Result

The LED should light up as it draws closer to the antenna (signal). The further away it is, the weaker the signal and the less it can grab to light up.

The Science

The WiFi routers uses radio waves to transmit data to and from computers and other devices in your home. Radio waves behave like water in a glass, they ripple or oscillate at a certain number of times or frequency per second. A radio wave's wavelength is the distance covered by one complete cycle of a radio wave — from peak to valley.

Radio waves used by WiFi routers, bluetooth devices, microwaves, etc., broadcast signals in the 2.45 gigaherz (Ghz) range or 2.45 billion oscillations per second. That means the signal wavelength is about 12 cm. To pick up signal, however, you only need an antenna that's half as big — 6 cm. When the radio waves strike this antenna, called a 1/2 dipole ("di"=two), the energy excites the antenna's

- Holiday Kitchen Safety – Cooking with Kids
- How to Shop for Old Window Repairs
- How Do I Clean A Clothes Dryer?
- Gobble Up These Convenient Thanksgiving Deals on Meals

The "Rectenna": Because the radio waves cause voltage to flow one direction and then the opposite direction, the current is called an alternating current (AC). By connecting a diode to both poles of the antenna it converts the AC current to flow only in one direction, called direct current (DC). The antenna becomes a "rectenna".

How the diodes work: The diodes work as one-way valves for electricity, allowing voltage to flow forward but not back. In the case of the rectenna, AC current caused by the radio wave sends a positive charge down one of the antenna poles and then the other. By putting the Hitachi 1SS106 diode across both poles, the diode only lets the positive charge coming down one of the antenna poles through. The LED is actually a Light Emitting Diode and works the same way as a regular diode except that when it's on, it lights up. Meanwhile, Schottky diodes turn on at a lower voltage than other diodesand because the rectenna can only catch a fraction of the WiFi's broadcast power, the Schottky diode can rectify that little bit of AC to DC.

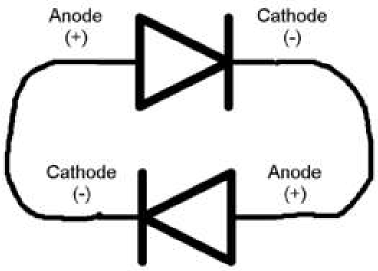

When we connected our diodes together, their polarity formed a loop that only allowed voltage to flow one way and thus light the LED. Brilliant job, pjaffeva!

Do you have any fun and kid-friendly science experiments you'd like to see us try for Fun with Less Kilowatts? Share with us in the comments!

Be Sociable, Share!

Tweet Like 0

About Vernon Trollinger

A native of Wyomissing Hills, PA, Vernon Trollinger studied writing and film at the University of Iowa, later earning his MA in writing there as well. Following a decade of digging in CRM archaeology, he now writes about green energy technology, home energy efficiency, DIY projects, the natural gas industry, and the electrical grid.

Tags: family fun, Fun with Less Kilowatts, kids, science experiments

Like 338K people like this. Sign Up to see what your friends like.

Follow @BounceEnergy 28.2K followers

Share 1 Tweet 0 Pin 0 G+

Follow Bounce: Share this on:

0 Comments Bounce Energy Blog 1 Login

Recommend Share Sort by Best

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Be the first to comment.

ALSO ON **BOUNCE ENERGY BLOG**

Choosing the Best Home Generator for Your Emergency Needs

1 comment • 3 years ago

T&B Electric LTD — Nice article covering the difference between portable and permanent generators. Some homeowners "back feed" …

Career Opportunity with Bounce Energy | PHP/Integration Developer

1 comment • 3 years ago

Chris Jones — Sounds like a great opportunity.

Take Care of Your Favorite Gadgets with Great Waterproof Tech

1 comment • 2 years ago

erna — Vermon thanks for your post, great as always! Coolest is cool I got one of this :) what's funny I also have swimmo watch for swimmers …

5 Quick DIY Home Improvement Fixes to Get Your Home Through Winter

1 comment • 3 years ago

Steve Veach — Excellent and informative blog post! I am a HVAC Contractor in Chicago and have found a new site to check in on. Thank …

Subscribe Add Disqus to your siteAdd DisqusAdd Privacy DISQUS

Copyright © 2017 Bounce Energy. Learn more with Bounce Energy Articles | Charity and Community Involvement

Follow Bounce: Share this on:

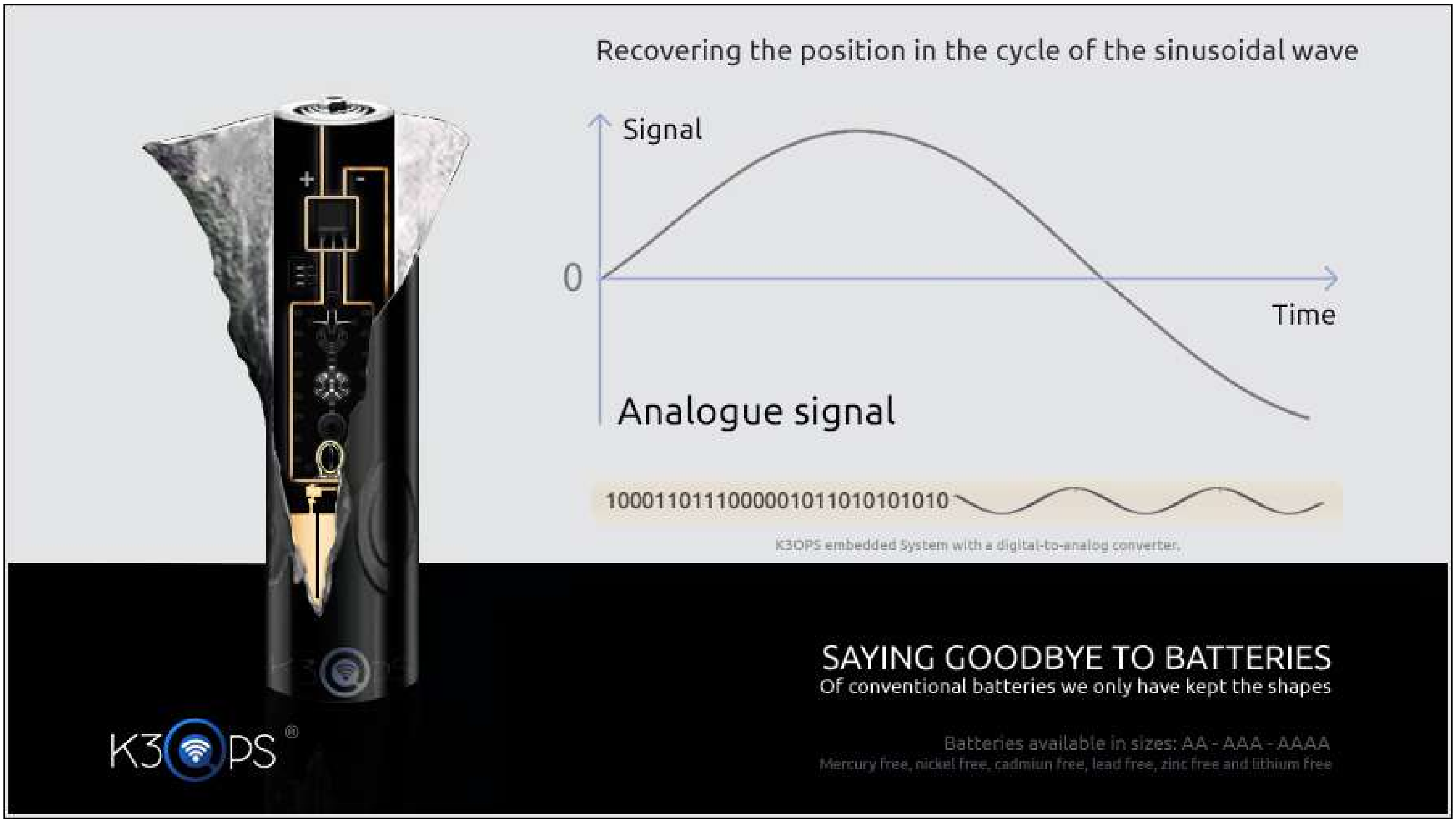
Recovering the position in the cycle of the sinusoidal wave
Signal
0
Time
Analogue signal
10001101111000001011010101010
K3OPS embedded System with a digital-to-analog converter.
SAYING GOODBYE TO BATTERIES
Of conventional batteries we only have kept the shapes
Batteries available in sizes: AA - AAA - AAAA
Mercury free, nickel free, cadmiun free, lead free, zinc free and lithium free
K3OPS®

An electromagnetic field uses the photon as an elementary particle to transmit force. It combines:

- **A magnetic field** *force resulting from the movement of loads μT.*
- **An electric field** *force created by the attraction of repulsion loads, measured in volts per meter - V/m.*

With an energy determined according to the speed of light, the RF are by far the best medium to transmit any kind of information.

The multiplication of wireless communications systems in our environment ensures sufficient microwave leakages to harvest from the ambiant and enough energy to convert into DC electricity. Electromagnetic fields are everywhere and since they carry energy, they became the best candidate to deliver an endless source of renewable energy.

$\vec{B}$ is the magnetic induction expressed in T refered to **Nikola Tesla**, "Father of Free Energy", which is at the origin of the electromagnetism.

Using meta-materials combined with nanotechnology has deeply increased the performance and miniaturization of rectennas embedded in K3OPS system. Our products operate autonomously, offering an endless supply of green energy in a respectful and environment-friendly approach.

principles and beautiful sentences. At that point in time I would have simply been Xin, a child like any other...

But one day I dreamed. I woke up far away, somewhere else, in another past. I decided to change my destiny and even if I was supposed to become a mathematician, I eventually decided to create, because already as a child my heart was chasing the stars.

Of the hundreds of directions shown to me after graduating, only one captured my attention: a single goal... Build the impossible for a safer world. So, over the past 3 years, our real challenge to overcome for all RF Energy Harvesting technics was to optimize electricity conversion. The massive proliferation of wireless telecommunication systems since the past two decades brought a saturation of the electromagnetic fields with a constant growth of 15% every year in our environments. As a result, this situation reversed the base problem that makes today Harvesting RF Energy a game changer. The key was the Power Management System.

We are far beyond the conversion constraint and performance by controlling "RF-interferences", by harvesting different frequencies *from near and far*, by using Metamaterials combined with nanotechnologies. We dramatically have improved power conversion efficiency and reduced the size of our Energy Harvesting systems embedded in all K3OPS' products.

Thanks to Nikola Tesla, my inspiring mentor, K3OPS' products have reached by far their objectives in terms of converting and performance, offering an endless efficient source of green energy, reliable in an environmentally friendly approach.

Xin WEI Co-Founder of K3OPS technology with Alexandre Despallieres

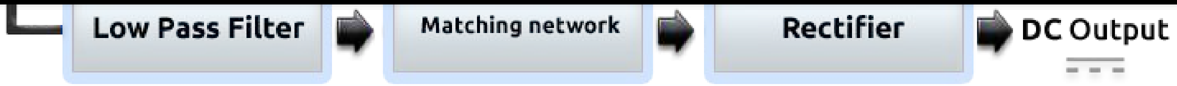
Low Pass Filter
Matching network
Rectifier
DC Output

The Rectenna was invented in 1964 by William C. Brown, patented in 1969. It is a rectifying antenna used to convert microwave energy into DC. A simple Rectenna consists of a dipole antenna with an RF diode connected across the dipole elements. The diode rectifies the AC current induced in the antenna to produce DC power.

K3OPS batteries
K3-Case

K3OPS Copyright © 2017 Alexandre Despallieres and Xin Wei - All rights reserved.

Terms of use | Privacy Policy | Press | Login

Centre for Collaboration in Electromagnetics and Antenna Engineering is a centre for research collaboration established to foster research in electromagnetics and
engineering conducted by Macquarie University researchers and prominent collaborators from external institutions, including people from industry and overseas ins
We have state of the art research ficilities including a NSI-700S-50 spherical nearfield anechoic chamber, 3 Vector Network Analyzers (capable up to 22GHz, 50GHz, and
respectively), state-of-the-art embroidery machine DreamCreator-XE-VM5100 for embroided antennas, Dielectric Characterization Kit - High Temperature Probe (20
20GHz), as well as several in-house and commercial licenced softwares.

Our recent and ongoing research projects are:

- Wireless Implantable Bio-Telemetry System and Miniature Antenna Design
- Wireless Freedom for Lab Rats
- Flexible and Wearable Antennas for Biomedical and Healthcare Applications
- Characterizing Properties of Carbon nano Tubes at Microwave Frequencies
- Electromagnetic Band Gap (EBG) Resonator Antennas
- Leaky-Wave Antennas for Advanced Wireless Systems
- Shared Aperture Arrays for Space Borne Applications
- Novel Dielectric Resonator Antennas
- Super Wideband Antennas
- Focal Plane Arrays for Radio Astronomy
- Archimedian Spiral Metamaterials
- Frequency Selective Surfaces for Energy-Saving Glass Panels
- Dual-Band Artificial Magnetic Conductor (AMC) Surfaces
- Antenna Technologies for Ultrawideband (UWB) Systems
- High Gain Antennas with Planar Surface-Mounted Short Horns
- Photonic Crystal/EBG Based Horn Antennas
- Broadband Microstrip Patch Antennas for Wireless Computer Networks
- Theoretical Analysis of Photonic Crystal Structures
- New Closed-form Green's Functions for Microstrip Circuits and other Layered Structures
- Integrated-design of Hybrid-resonator Antennas for Broadband Wireless and other Communication Systems
- Singularity-enhanced Finite-different Time-domain (FDTD) Method for Diagonal Metal Edges, Strips and Films
- Low-profile Dielectric-resonator (DR) Antennas

Wireless Implantable Bio-Telemetry System and Miniature Antenna Design

The two major challenges associated with the conversion of a wireless system operating in air to an implantable version, antenna detuning and biocompatibility, are a
in a coherent way. An RFID-based biomedical telemetry system designed for free-space operation was chosen as the starting reference. A new, pin-compatible, spa
antenna with a ground plane was designed, fabricated and tested, to replace the original “free-space” antenna in the active RFID tag without making any other chang
tag circuit, such that the tag would function well when it is placed under rat skin and fat. Biocompatibility and potential antenna detuning due to rat tissue variati
addressed in the design process, without significantly increasing the tag physical height, by applying a thin coating of biocompatible material directly over the ante
operation of the medical telemetry system was successfully demonstrated, with the tag placed under rat skin and fat, and its range of 60-72 cm was found to be suf
support medical research experiments conducted with rats in cages. Due to the biocompatible coating over the antenna, antenna matching is very insensitive to ch
tissue dielectric constants and thickness. The footprint of the new antenna is 33% less than that of the original antenna, its measured 10 dB return-loss bandwidth is 1(
11%, and overall efficiency is 0.82% at 920 MHz.

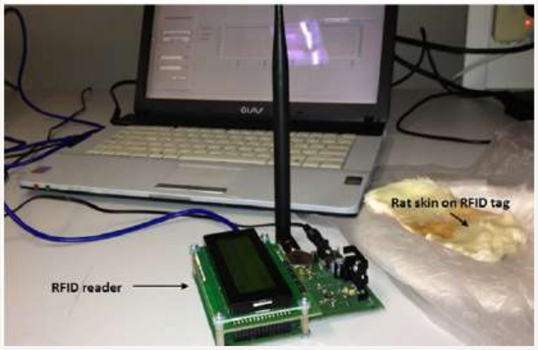

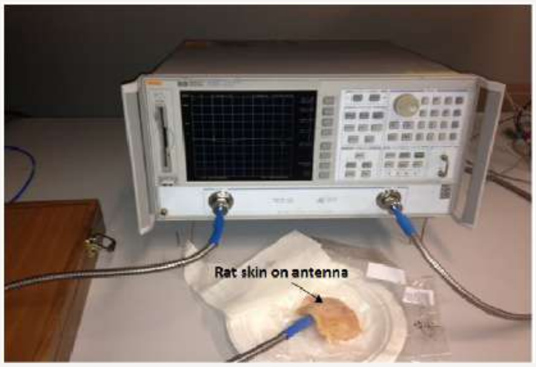

Wireless Freedom for Lab Rats

We are developing a fully implantable wireless telemetry system. This is a joint research project with BCS Innovations and the Australian School of Advanced Medicine
will be first used in the research conducted in ASAM, with rats, on hypertension. To date the major method of controlling hypertension is through the use o
pharmaceuticals. The pathway to approval for most drugs for human use involves pre-clinical (animal) trials. Lab rats are considered biologically similar to humans, pa

in terms of their social behaviour. Therefore, it is very important to not compromise the pharmaceutical trials by unnecessarily stressing the rats by harnessing the monitoring equipment. One of the technical challenges of developing an implantable system that monitors the various signals, is the relatively small size requ implantable telemetry system is a miniature transceiver implanted in an animal that senses, processes and transmits data via a wireless link to monitor vital signals of c freely moving laboratory animals. This is crucial in giving researchers flexibility and reliability, especially in studies with special experimental settings using mazes wheels and treadmills. We have plans to develop a fully implantable telemetry system for subcutaneous or intraperitoneal placement in rats that monitors th parameters as well as blood pH and chemistry, nerve activity and circadian respiratory rate rhythms. The aim of this project is to eventually develop a system with e capabilities that costs less than what is currently available, to provide more universities and researchers with the opportunity to use this technology.

In this sequel, initially, when a module of our original system was placed under the skin of a rat, the wireless link failed completely. It could not send a temperature read a centimetre! The point of failure was the commercial antenna in the module that had been designed to work outside the body, in air. Such antennas do not work unde because the electrical characteristics of skin (rat or human) are significantly different from that of air. Hence the main challenge was to design an antenna that works v placed under the skin. In addition, it was necessary to cover the module and the antenna by biocompatible material, which also affects antenna performance. Possible v of skin characteristics from one rat to another or one person to another were considered. Unlike the commercial antenna, we wanted our antenna to radiate less into th the rat/person and more away from the body because that not only increases the quality of the wireless link, maximum range (distance between implanted mo monitoring station) and battery life but also reduces the exposure of the body to radio-frequency waves. Indeed we had to consider the electromagnetic effects of fat a material around the antenna. We were able to meet all these requirements with a novel compact antenna design that is approx. one third the size of the original co antenna. We successfully demonstrated wireless telemetry transmission of temperature with the new module placed under rat skin and the monitoring station pla distance of up to 80 cm! This range is sufficient for our immediate target application supporting new medical research by Professor Paul Pilowsky's team. If necessary, further extended by increasing the power level at the monitoring station.

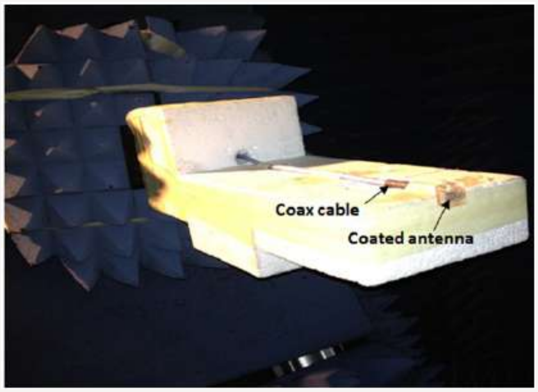

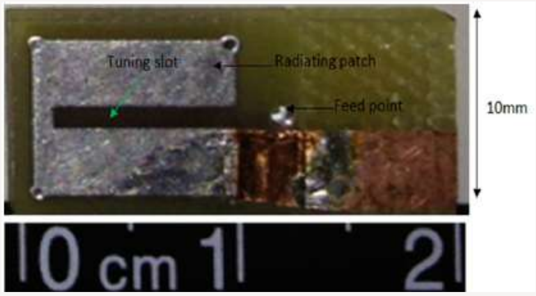

Flexible and Wearable Antennas for Bio-Medical and Healthcare Applications

Body Centric Wireless Communication is a rapidly growing research area targeted for medical, healthcare, public safety and defense applications. The need to address transceiver specifications and real-time scenarios in close proximity to human body is continuously evolving antenna system research. Several novel miniature antennas havi dual- and wide-band operations have been designed and tested for Wireless Body Area Network (WBAN). They have significant advantages of small size, wide radiation patt the human body for maximum coverage and are less sensitive to the gap variation between human body and antenna.

A compact ultra wideband antenna is shown below with strong notch-band rejections up to VSWR = 26, that is tunable over a wide frequency range from 3.55GHz to 6.8 been designed. To estimate the stub length to notch frequency for a given interfering application, analytical expressions for the normalized stub length which is indep substrate dielectric constant is also presented. This helps to avoid hit-an-trail method and gives a good estimate of initial design parameters for notch. Proposed antenna radiation patterns and yields a measured 10dB return-loss bandwidth from 3GHz to 10.5GHz.

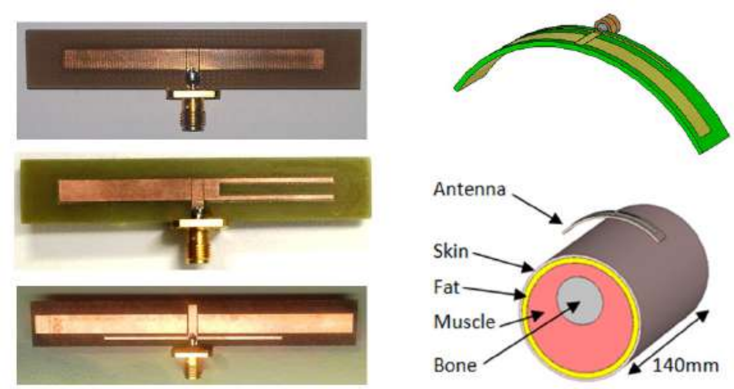

Characterizing Properties of Carbon Nano Tubes at Microwave Frequencies

Carbon Nanotube (CNT) yarns are novel CNT-based materials that extend the advantages of CNT from the nano-scale to macro-scale applications. We have modelled CNT potential data transmission lines. Test structures have been designed to measure electrical properties of CNT yarns, which are attached to these test structures using gold p testing and microwave S-parameter measurements have been conducted for characterisation. The observed frequency independent resistive behaviour of the CNT yarn promising indicator that this material, with its added values of mechanical resilience and thermal conductivity, could be invaluable for a range of applications such as B Networks (BAN). A model is developed for CNT yarn, which fits the measured data collected and agrees in general with similar data for non-yarn CNTs.

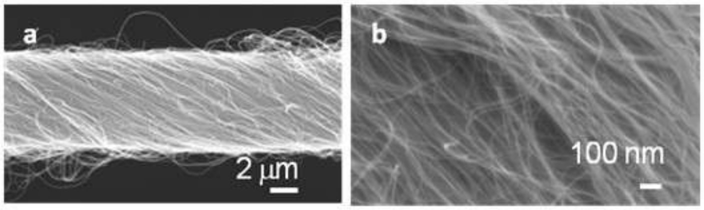

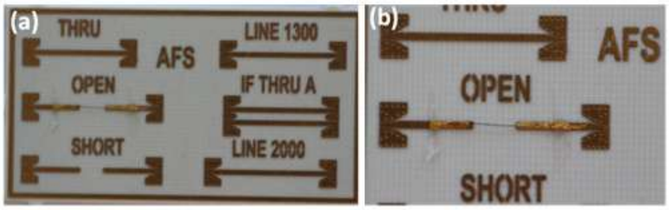

Electromagnetic Band Gap (EBG) Resonator Antennas (ERA)

We have designed, fabricated, and measured antennas based on 3D, planar and 1D EBG structures (i.e. photonic crystals). These flat microwave antennas, known as EBG antennas or Fabry-Perot cavity antennas, can give gains of about 20dB and very good efficiencies.

Enhancing Radiation Characteristics of ERAs by Improving Aperture Phase Distribution

This work focuses on achieving superior radiation characteristics of ERAs by improving their aperture phase distributions. A unique method, utilizing full-wave simula analytical analysis, has been developed to design Phase Correcting Structures (PCS) for ERAs. This method uses actual phase distribution on the physical aperture of ERAs relying on geometric optics. Several Phase Correcting Structures (PCSs) have been designed, which were later validated with the measurements of their fabricated p Tremendous improvement in the radiation performance is witnessed in both simulated and measured results. These exciting initial results validate our proposed methodo indicate an existence of a great potential to be explored. The figure below shows the field propagation above an ERA: (a) without PCS, and (b) with PCS. It is clear that the the PCS is much more uniform and the energy is focussed towards broadside direction, thus, resulting in increased directivity.

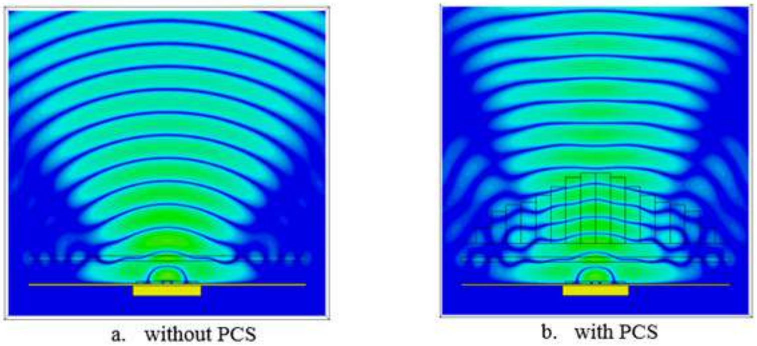

a. without PCS

b. with PCS

Extremely Wideband High-Gain ERAs (Gain~15-20dBi, Bandwidths >50%)

In 2014-15, we invented an innovative class of electromagnetic band-gap (EBG) resonator antennas which provide high gain and wide bandwidth with an extremely r footprint. One of the prototype developed has only 8% of the area as compared to conventional EBG resonator antennas but its performance (gain bandwidth) is a record hi for this class of antennas, while providing gain in the range of 15-20 dBi. This represents an improvement of nearly two orders of magnitude in the bandwidth compared to EBG resonator antennas. Thanks to its practical advantages of flat shape and low-cost manufacturability, it can be easily attached to a wall of a building, for example, to co building wirelessly to the National Broadband Network.

Our most recent results on such wideband high gain antennas can be found in: "Achieving high gain-bandwidth through flat GRIN superstrates in Fabry-Perot cavity antenn 2014 IEEE International Symposium on Antennas and Propagation (AP-S/USNC-URSI), pp. 1748 – 1749, Memphis, Tennessee, USA, July 6-11, 2014.

Detailed antenna design along with experimental data of another of our wideband EBG resonator antenna having composite multi-layer superstrate, is published in the pa "Wideband high-gain EBG resonator antenna with a small footprint and all-dielectric superstructures" in IEEE Transactions on Antennas and Propagation, vol. 62, no. 6, 201

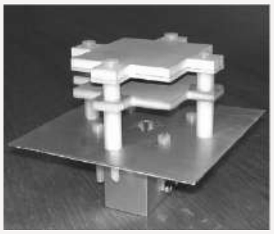

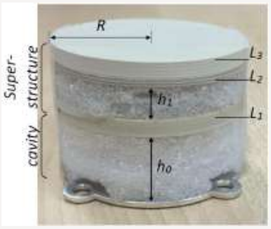

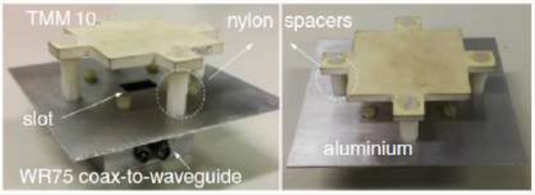

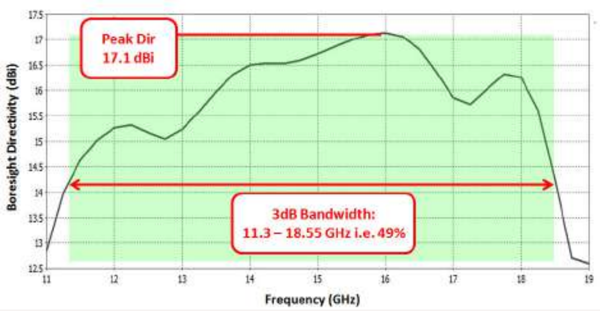

Frequency response of one of our Wideband ERA Prototypes

Simple Dual-Band ERAs

We developed a new method to obtain dual-band operation from a simple electromagnetic band gap resonator antenna. The antenna is based on a one-dimensional EBG made out of two low-cost unprinted dielectric slabs. The EBG structure is implemented as the antenna superstrate, which has been designed to provide a locally-inverted reflection phase gradient with high reflectivity, in two pre-determined frequency bands. The linearly polarised antenna design and experimental results are described in entitled: "A simple dual-band electromagnetic band gap resonator antenna based on inverted reflection phase gradient" published in IEEE Transactions on Antennas and Pro vol. 60, no. 10, pp. 4522-4529, 2012.

We have extended this concept for dual-band, dual-polarised and circularly-polarised antennas. We also designed a tri-band antenna following this concept. It needs only t cost unprinted dielectric slabs.

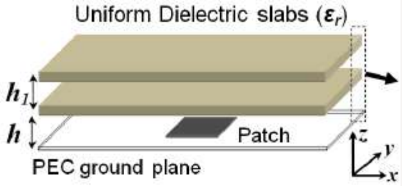

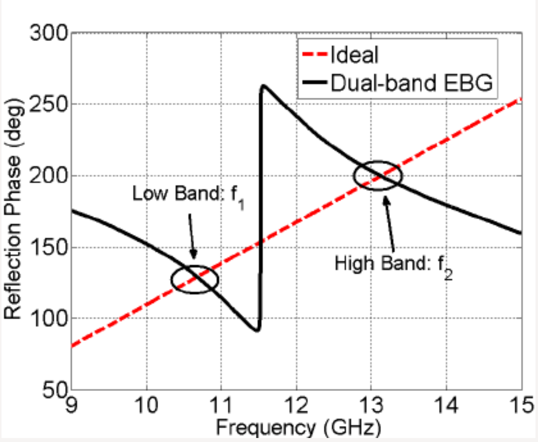

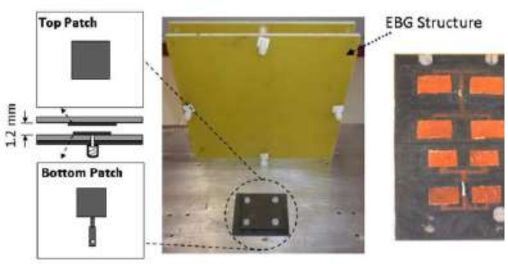

Low-Profile Wideband ERAs

We have designed and successfully tested wideband low-profile (thin) EBG resonator antennas. The breakthrough that contributed to this success is our design of a partially surface (PRS) with a positive reflection phase gradient. Thin single-dielectric-slab PRSs with printed patterns on both sides were investigated to minimise the PRS thickn simplify fabrication. Three such surfaces, each with printed dipoles on both sides, have been designed to obtain different positive reflection phase gradients and reflection r levels in the operating frequency bands. These surfaces, and the EBG resonator antennas formed from them, were analysed theoretically and experimentally to highlight t compromises involved and to reveal the relationships between the antenna peak gain, gain bandwidth, the reflection profile (i.e. positive phase gradient and magnitude) of th and the relative dimensions of dipoles. A small feed antenna, designed to operate in the cavity field environment, provides good impedance matching (|S11|< -10 dB) a operating frequency bands of all three EBG resonator antennas. Experimental results confirmed the wideband performance of a simple, low-profile EBG resonator antenna thickness is only 1.6mm, effective bandwidth is 12.6%, measured peak gain is 16.2 dBi at 11.5 GHz and 3dB gain bandwidth is 15.7%. Please find details in a paper entitled of Simple Thin Partially Reflective Surfaces with Positive Reflection Phase Gradients to Design Wideband, Low-Profile EBG Resonator Antennas," published in IEEE Tra on Antennas and Propagation in 2012.

Low-Profile Dual-Band ERAs

We have achieved dual-band operation in a low-profile EBGRA using a single dielectric superstate with a printed pattern only on one side. This also made use of our r inverting the gradient of the PRS reflection coefficient.

Woodpile EBG Material and Antennas

In 2003, we designed and built a woodpile 3D photonic crystal, also known as a EBG crystal, operating in microwave frequencies, and demonstrated experimentally the ex the electromagnetic band gap. The crystal is made out of cermain material.

Then we designed a planar EBG resonator antenna. Shown below, it has a resonant cavity between a ground place and a 3D woodpile photonic crystal. We employed both m and slots to feed the cavity, and investigated both linearly and circularly polarised antennas. The linearly polarised antenna design and experimental results are described in entitled "A planar resonator antenna based on a woodpile EBG material," published in IEEE Transactions on Antennas and Propagation,vol. 53, no. 1, pp. 216-223, January 2

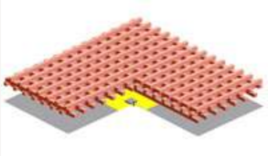

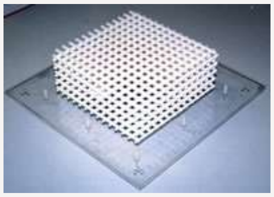

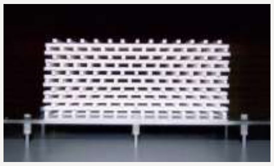

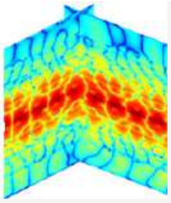

Leaky-Wave Antennas for Advanced Wireless Systems

Antenna beam steering can bring significant benefits to advanced wireless systems. Microstrip leaky-wave antennas (MLWAs) are of particular practical interest becaus planar low-profile configuration, ease of fabrication, and beam-scanning capabilities. In this research several planar MLWAs and arrays are developed to radiate at bores conical beam around the boresight, with simultaneous dual-side-beam scanning, dual-band forward and backward beam-scanning, and continuous beam scanning from the to the forward direction. Moreover, methods and antenna designs are proposed to steer the beam at a fixed frequency, shifting beam-steering range, and fixed-frequency bean in forward and backward directions.

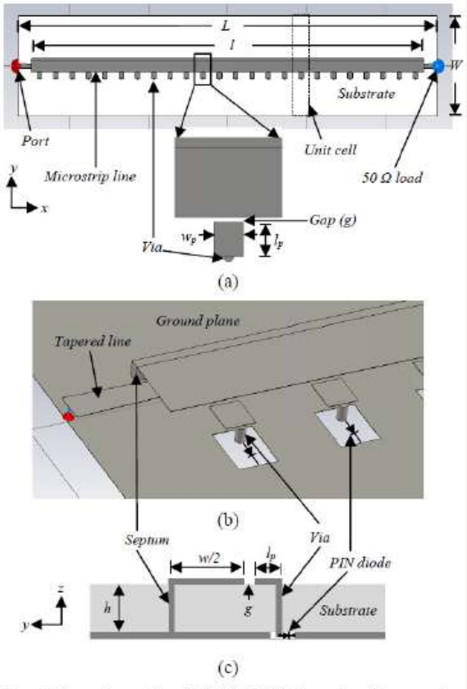
L
l
W
Substrate
Port
Unit cell
50 Ω load
Microstrip line
y
x
Gap (g)
w_p
l_p
Via
(a)
Ground plane
Tapered line
(b)
Septum
Via
$w/2$
l_p
PIN diode
z
h
g
Substrate
y
(c)

Figure 1. Proposed reconfigurable HW-MLWA: (a) top view, (b) perspective view, and (c) side view.

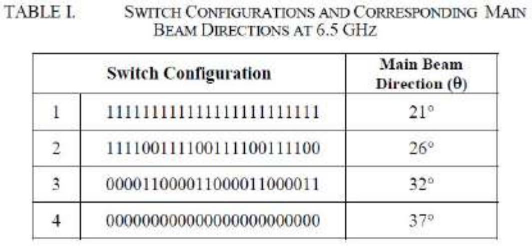

TABLE I. SWITCH CONFIGURATIONS AND CORRESPONDING MAIN BEAM DIRECTIONS AT 6.5 GHZ

Switch Configuration		Main Beam Direction (θ)
1	111111111111111111111111	21°
2	111100111100111100111100	26°
3	000011000011000011000011	32°
4	000000000000000000000000	37°

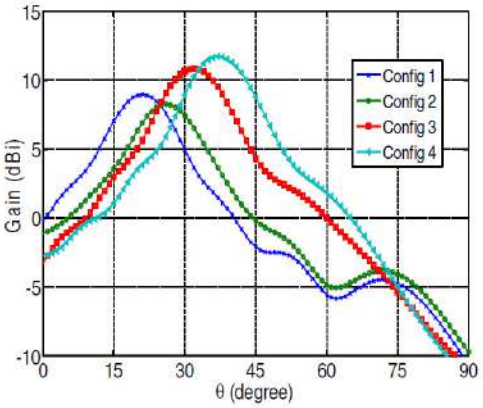

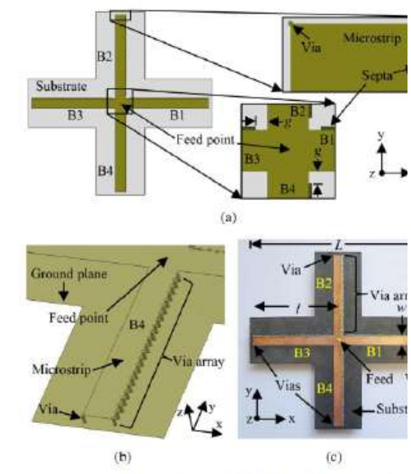
B2
Substrate
B3
B1
B4
Via
Microstrip
Septa
Feed point
g
y
z
x
(a)
Ground plane
Feed point
B4
Microstrip
Via array
Via
(b)
L
Via
l
Vias
Feed
Subst
(c)

Fig. 1. HW-MLWA array design steps: (a) top view of the initial de septa, (b) perspective view of one branch of the final design with v (substrate omitted, not to scale), and (c) top view of the fabricated pro

---

Multi-band Dual-Polarized Shared Aperture Array

Multi-band dual-polarized shared-aperture (MBDP-SA) arrays are antenna arrays that operate in two (or more) frequency bands with dual-polarization in each band, a elements are integrated together into a common physical space by sharing the single aperture. The MBDP-SA array is of great interest in space-borne SAR system, be technique can effectively reduce the payload and size of the antenna sub-system. In this research project, main efforts are focussed on three aspects: 1) improve the specifi current Dual-Band Dual-Polarized Shared-Aperture (DBDP-SA) array; 2) construct Tri-Band Dual-Polarized Shared-Aperture (TBDP-SA) array; and 3) explore some new for DBDP-SA antenna.

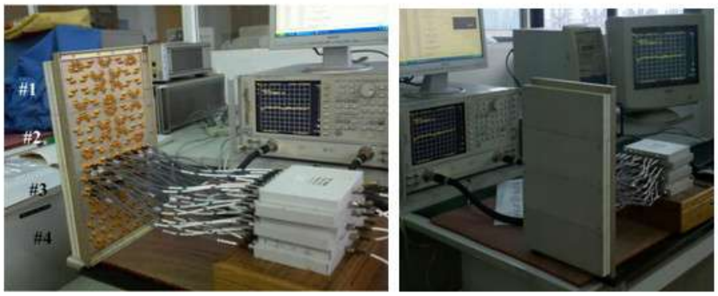

(a) Rear (b) Front

Fig.1 Improved DBDP-SA array.

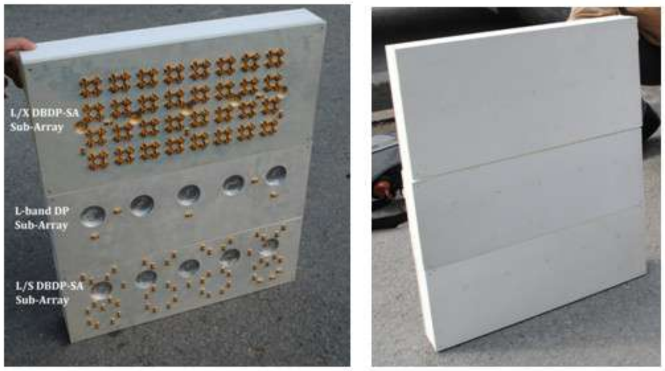

(a) Rear (b) Front

Fig.2 Photos of the TBDP shared aperture prototype array.

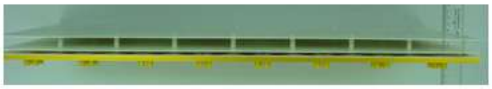

Side view

Fig.3 Novel structure DBDP-SA array

Compact Dielectric Resonator Antennas with Ultra-Wide 60%-110% Bandwidths

We have recently made a significant achievement in the emerging ultra wide-band (UWB) wireless communication systems, which require antennas with bandwidths gr 106%. In the past, the only way antenna engineers knew how to get such a bandwidth from a thin antenna was by removing the metal sheet underneath the antenna (known a plane"). This is not an acceptable solution for practical systems because the lack of it allows the antenna to radiate both upwards and (unnecessarily) downwards (i.e. electronic device on which the antenna is installed), wasting about half of its power. Saving power is crucial in UWB systems due to the severe power limits imposed by regu 2011, we made a breakthrough in dielectric-resonator (DR) antenna research, by inventing a novel DR antenna with a full ground plane and a 110% bandwidth. This disp myth that such bandwidths cannot be achieved with full ground planes. This antenna, published in the prestigious IEEE Transactions on Antennas and Propagation in Decem is 29% smaller but has a 30% greater bandwidth than the next-best DR antenna, which does not even have a full ground plane. Hence it is ideal for next-generation ultra-fas and sensor applications.

We theoretically and experimentally demonstrated that, by introducing a lower-permittivity full-length insert between the ground plane and a higher-permittivity dielectric dielectric resonator antennas (DRA) with ultra-wide bandwidths, in the range of 60%-110%, can be designed. Furthermore, the volume of such DRAs is reduced by appro 50% using a finite planar conducting wall. Unlike vertical monopole-type hybrid UWB DRAs, these antennas radiate sufficiently in the upward direction. Unlike in printed UWB antennas, the power radiated into the lower hemisphere is significantly less. An example prototype antenna, designed to operate in the FCC UWB band, has a dielectr of 12 x 8 x 15.2 mm3 (or 0.124 x 0.083 x 0.157 lambda3 at 3.1 GHz), and an average measured gain of 5 dBi from 3.1 to 10.6 GHz. These antennas exploit multiple low with overlapping bandwidths to form an ultra-wide contiguous bandwidth. With the proposed dielectric arrangement, it is possible to efficiently couple a sufficient numbe overlapping modes to a 50 ohm feedline using a single, simple feed.

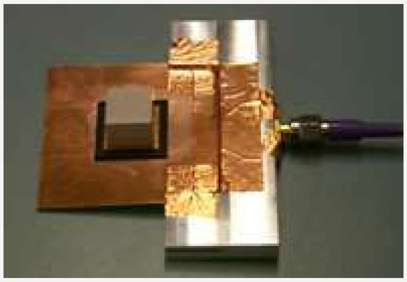

The antenna has a remarkably small footprint of 12x8 mm2 at 3.1 GHz - the lowest frequency of the FCC UWB band. Its dielectric volume is 1459 mm3, or 1.7x10-3 lamb lowest operating frequency of 3.1 GHz, and overall height is 15.2 mm or 0.157 lambda0. To place these results in perspective, it is worth comparing the new designs with wideband DR designs available in the literature. To the best of our knowledge, prior to this, the widest bandwidth ever obtained from a pure DRA design is 84% . The volu DR in that design is 0.225 x 0.172 x 0.062 lambda3 (= 2.4x10-3 lambda3) at its lowest operating frequency of 3.69 GHz. In that DRA, the DR is positioned in a non-traditi close to the edge of an orthogonal, "vertical" ground plane, which does not block radiation towards the lower hemisphere. The widest bandwidth demonstrated by a pure DI traditional "horizontal" ground plane, (which can be employed to shield the rest from the antenna, as discussed previously) is 78%. The dielectric volume of that design is 0. x 0.21 lambda3 (= 5.8x10-3 lambda3) at its lowest operating frequency of 6.7 GHz.

Super Wide-band Antennas

We have demonstrated that extremely wide bandwidths (ratio-bandwidths up to 1:25) can be obtained from a specially designed printed antenna with a tapered semi-ring design is described in "A Printed Elliptical Monopole Antenna with Modified feeding Structure for Bandwidth Enhancement," in IEEE Transactions on Antennas and Pro vol. 59, no. 2, pp. 667-670, Feb. 2011.

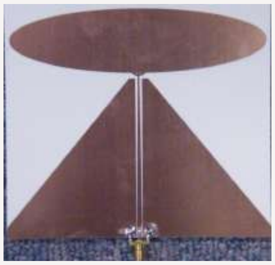

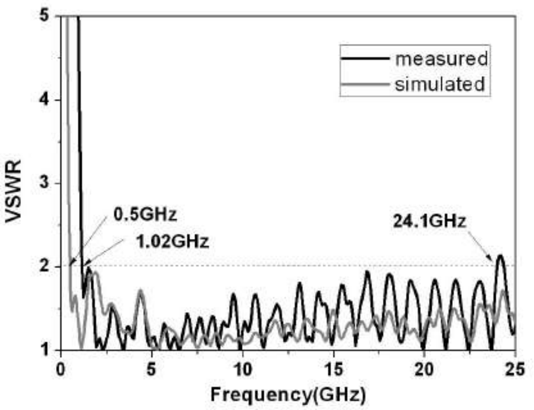

Focal Plane Arrays for Radio Astronomy

Dense focal plane arrays (FPAs) are a key technology for a new generation of Radio-telescopes. Their primary benefit is the rapid survey speed facilitated by the wide fiel provided by multiple beams. Recent advances have brought dense FPAs within reach of radio astronomy applications. A number of institutions have significant research pr this field. This technology is being considered for the Square Kilometre Array (SKA) (www.skatelescope.org). The PhD project of Douglas Hayman, conducted with CS Centre and Division of Astronomy and Space Science, investigated beamforming aspects of FPAs and evaluated their performance in Radio Astronomy. A prototype interf radiotelescope, built at CSIRO's Radiophysics Laboratory in Sydney, is used to demonstrate a suite of techniques forFPA beamforming and evaluation for this thesis. Bea solutions were experimentally demonstrated in our paper in IEEE Transactions on Antennas and Propagation, entitled “Experimental Demonstration of Beamforming Sol Focal Plane Arrays”.The THEA tile, shown below, is designed by ASTRON and used for the experimental component of this research.

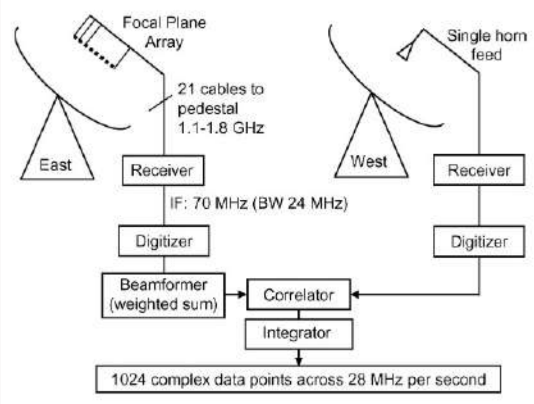

Negative Permeability of Spiral Metamaterials

Archimedean spirals and complementary Archimedean spirals are super-compact metamaterial particles. Thanks to their convoluted geometry, unit cells can be made electri small. We theoretically analysed monofilar, bifilar, trifilar and quadrifilar Archimedean spiral metamaterial particles using point group theory and crystallography. From the properties electromagnetic response was determined. Magnetic, electric and magnetoelectric modes of the particles were identified along with their isotropy characteristics. shown that all the particles, except monofilar spiral, are nonbianisotropic. Further, effective medium theory was applied to extract the effective permeability of the spiral me results indicated negative values for permeability in certain frequency ranges. Detailed theory and numerical simulation results are available in the paper entitled “Analysi metamaterials by use of group theory,” published in the Metamaterials Journal, vol. 3, no. 1, pp. 33-43, March 2009.

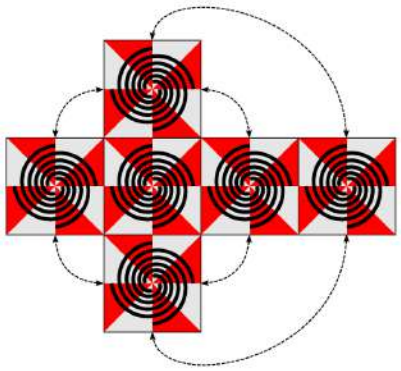

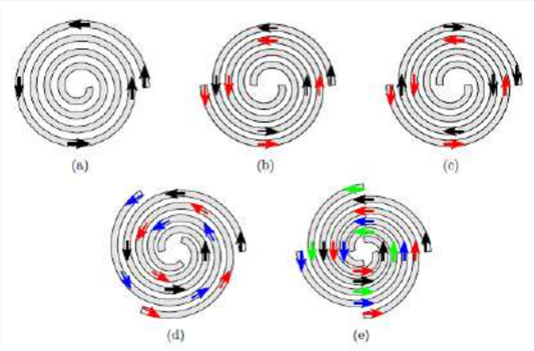

Archimedean Spiral Metamaterials and Backward Waves

We have shown backward wave propagation and double negative parameters over a 19% bandwidth in a microstrip line loaded with series gap discontinuities and super complementary Archimedean spiral resonator metamaterial particles. Moreover, our equivalent-circuit model for such unit cells almost perfectly described the structu practically important frequencies (by comparison with full-wave results). We also fabricated and tested compact filter circuits with only one or two complementary spiral met particles. Our results sre summarized in the paper entitled “Backward Wave Microstrip Lines with Complementary Spiral Resonators,” published in IEEE Transactions on and Propagation, Vol. 56, Issue: 10, pp. 3173-3178, Oct 2008.

We derived design equations for Archimedean spiral resonators and tested them against full-wave simulations. The details are in the paper entitled “Design of monofilar a Archimedean spiral resonators for metamaterial applications,” published in IET Microwaves, Antennas & Propagation, vol. 3, no.6, p. 929-935, Sep.2009.

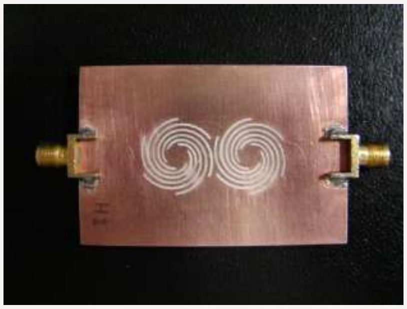

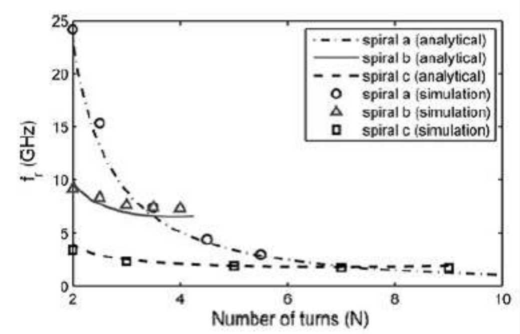

Frequency Selective Surfaces for Energy-Saving Glass Panel

Energy-saving glass is becoming very popular in building design due to their effective shielding of building interior against heat entering the building with infrared (IR) wave obtained by depositing a thin layer of metallic-oxide on the glass surface using special sputtering processes. This layer attenuates IR waves and hence keeps buildings summer and warmer in winter. However, this resistive coating also attenuates useful microwave/RF signals required for mobile phone, GPS and personal communication sy by as much as 30 dB. To overcome this drawback, we designed and tested a bandpass aperture type cross-dipole frequency selective surface (FSS), etched in the coatings o saving glass to improve the transmission of useful signals while preserving IR attenuation as much as possible. With this FSS, 15-18 dB peak transmission improveme achieved, for waves incident with 45 degrees from normal for both TE and TM polarizations.

Measurements and other results of this research, conducted in collaboration with the Lund University in Sweden, are available in the paper entitled “Cross-Dipole Frequency Selective Surface for Energy-Saving Glass Used in Buildings,” published in IEEE Transactions on Antennas and Propagation, vol. 59, no. 2, pp. 520-525, Feb. effect of these FSSs on the transmission of infrared and visible wavelengths through energy-saving glass was investigated theoretically and experimentally in another pap Microwaves, Antennas & Propagation, Vol. 4, Iss. 7, pp. 955–961, 2010.

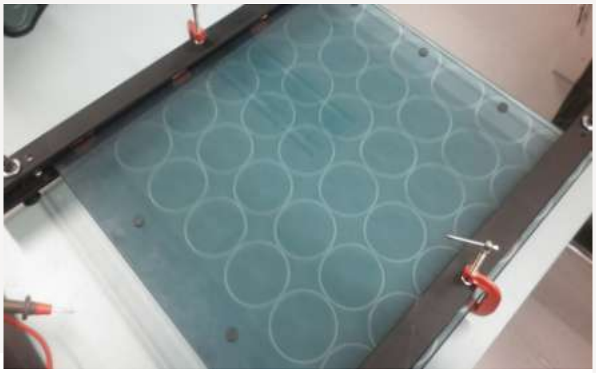

Switchable Frequency Selective Surfaces to Reconfigure Electromagnetic Architecture of Buildings

In large buildings and offices, frequency re-use methods will be required to enhance the spectral efficiency and capacity of wireless communication systems. This observatio to the concept of electromagnetic architecture of buildings. Passive bandstop FSSs can be used to enhance the electromagnetic architecture of a building, and hence to spectral efficiency and system capacity, but switchable FSSs can provide a better reconfigurable solution. If switchable FSSs are placed in strategic locations of a building, th reconfigured remotely and rapidly, which is not possible with passive FSSs.

With collaborators in UK and Sweden, we designed and successfully tested a single-layer active Frequency Selective Surface (FSS) that is electronically switchable between and transparent states. It can be used to provide a spatial filter solution to reconfigure the electromagnetic architecture of buildings. The FSS measurements show that the response of the filter does not change significantly when the wave polarization changes or the angle of incidence changes up to ±45º from normal. The FSS is based on sq aperture geometry, with each unit cell having four PIN diodes across the aperture at 90 degree intervals. Experiments demonstrated that almost 10 dB additional transmissio be introduced on average at the resonance frequency, for both polarizations, by switching PIN diodes to ON from OFF state.

For details, please refer to “Switchable Frequency Selective Surface for Reconfigurable Electromagnetic Architecture of Buildings,” in IEEE Transactions on Ante Propagation, Vol. 58, Issue 2, pp 581-584, February 2010.

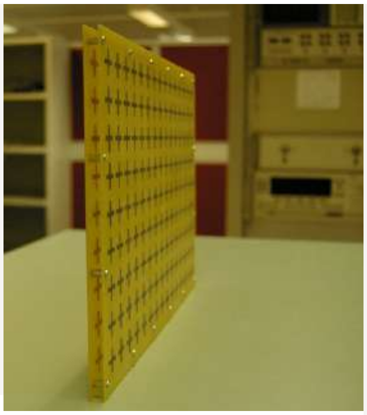

Absorb/Transit FSS

We designed and tested a novel absorb/transmit frequency selective surface (FSS) for 5-GHz wireless local area network (WLAN) applications. The novelty of the design i capable of absorbing, as opposed to rejecting, WLAN signals while passing mobile signals. The absorption of the WLAN signal is important to reduce additional multipa spread and resultant fading caused by typical reflect/transmit FSSs. Our FSS consists of two layers, one with conventional conducting cross dipoles and the other with resis dipoles. The FSS has good transmission characteristics for 900/ 1800/1900-MHz mobile bands and performs well for both horizontal and vertical polarizations.

Later we modified the FSS to obtain even better performance, for example, for both horizontal and vertical polarizations at oblique angles of incidence. The distance betwee layers has been successfully reduced to one eighth of free-space wavelength. This small distance makes it more compact as compared to the conventional Salisbury screen achieving an acceptable absorption in the stopband.

The details of our designs and test results can be found at “Oblique Incidence Performance of a Novel Frequency Selective Surface Absorber,” in IEEE Transactions on Ant Propagation, Vol. 55, no. 10, pp. 2931 – 2934, Oct. 2007, and “A Novel Absorb/Transmit FSS for Secure Indoor Wireless Networks with Reduced Multi-path Fadin Microwave and Wireless Component Letters, Vol. 16 (6), pp. 378 - 380, June 2006.

Dual-Band Artificial Magnetic Conductor (AMC) Surfaces

AMC surfaces have many advantages and interesting properties due to their unique reflection characteristics, with near zero reflection phase. We have designed and pro novel dual-band AMC surface, which has a very wide upper AMC band and a narrow lower AMC band, and therefore suitable for multi-band wireless/mobile applications.

Fully Printed Quad-band Antennas for Wi-Fi IEEE802.11 and other WLAN Applications

Our fully-printed antennas can be fabricated and integrated to WLAN systems at almost zero cost by printing them on the same circuit board (e.g. FR4) with the radio circuit same standard fabrication methods. They are extremely compact: an antenna with a radiating element of 1cmx1cm covers all four IEEE standards (802.11a, 802.11b, 80 802.11n) as well as HiperLAN2 with a VSWR less than 2. We have also developed a packaging solution where the rest of the circuit can be shielded to satisfy EMC regulati leaving the antenna (on the same board) open for radiation. The advantages of the antennas based on this technology are: Lightweight; Radiates almost every direction in shadow region); Microstrip and co-planar waveguide (CPW) designs available; Compatible with all printed microwave circuits, including stripline circuits; Excellent because no cables, connectors, soldering or any mechanical attachments are required to connect the antenna to the radio; No protruding parts that are likely to break; Covers WLAN bands with one antenna (e.g. Wi-Fi IEEE 802.11 a, b, g, n, HiperLAN2 etc.); Excellent matching, i.e. input reflection < -10 dB (VSWR < 2) in all WLAN bands (e. GHz, 4.9-5.1 GHz, 5.15-5.35 GHz, 5.725 – 5.825 GHz); High efficiency (~60-70% on FR4; can be increased to over 90% on Duroid.); Ideal for internal mounting in a c device, to achieve very wide beam coverage; diversity reception or MIMO arrangements available; can be dsigned to cover other multiple bands (e.g. multiple mobile/c bands) and other applications.

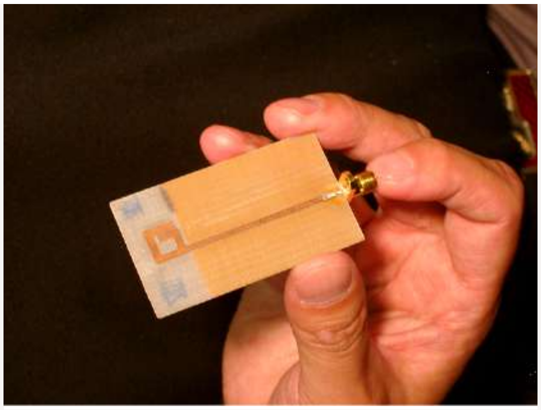

How can I optimise a UWB System to operate well over a range of directions?

Emerging ultra-wideband (UWB) communication devices will need to operate well, not just in one direction but over a range of directions. However previous UWB system optimisation techni been limited to one direction only. A system optimised in one direction may not work well in other directions due to pattern instability, or direction-dependent transfer function, of the an developed the concept of frequency-domain correlation patterns and proposed a figure of merit, the pattern stability factor (PSF), which can be determined from simulation of experimen With these new concepts, we have demonstrated the optimisation of UWB systems to operate well over a range of directions. The concepts and their applications are available in the PhD papers by Tharaka Dissanayake.

Antenna Techniques and Technologies for Ultrawideband (UWB) Systems

We developed antenna solutions for emerging UWB systems with very high data rates. Our research includes investigation of band-notching techniques to reduce UWB inference wi wireless systems and studies of antenna dispersion and pattern stability in the time domain.

High Gain Antennas with Planar Surface-Mounted Short Horns

In collaboration with University of Delhi, we have developed compact rectangular dielectric resonator antennas with surface mounted planar horns for gain enhancement. One configuration an aperture-fed rectangular dielectric resonator antenna and a planar surface mounted horn. We have achieved 10 dBi gain and a good efficiency from this configuration. The surface mou increases the gain of the standard dielectric resonator antenna by 4.9 dB. Total height of the prototype antenna shown in the figure is only 0.172 lambda and the aperture size is 0.96 lamb antennas can be easily adapted to low-profile, high-gain and array applications.

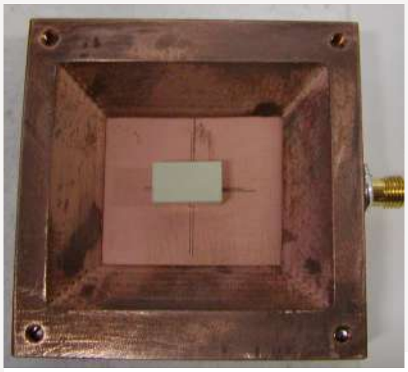

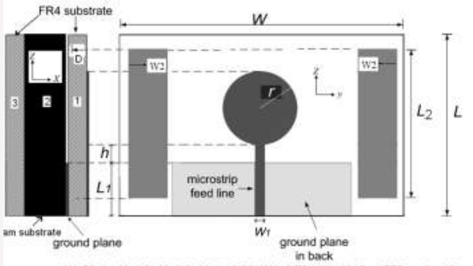

W=70, L=42, L2=40, L1=20, h=0.36, W1=2.78, W2=10, D=1.575 and r=10
(all dimensions are in mm)

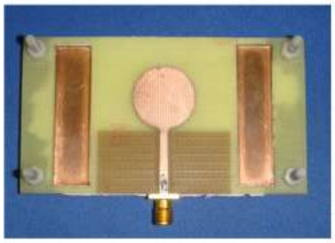

Photonic Crystal/EBG Based Horn Antennas

Waveguides and antennas made out of dielectric, as opposed to metal, are expected to perform better at Terahertz frequencies, as they do not suffer from the skin-effect loss of metal investigated horn antennas, horn arrays, waveguides, bends and junctions made out of 3D woodpile photonic crystals, which can be implemented at THz frequencies in high-resistivity S other materials. Our concepts have been tested by fabricating and measuring scaled-up prototypes operating in microwave frequencies. We have also designed a special broadband couple a photonic crystal waveguide to a rectangular waveguide or a coaxial cable, for example for testing using a vector network analyser. We demonstrated > 6% bandwidth both theore experimentally.

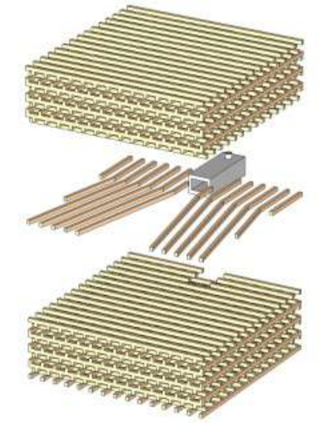

Broadband Microstrip Patch Antennas for Wireless Computer Networks

We have designed and tested broadband, compact E-shaped patch antennas for wireless communication networks, operating in frequencies from 4.9 GHz to 6.0 GHz. We made two achievements. First, we designed an E-shaped antenna, which is intrinsically compatible with a printed microstrip circuit. It can be made out of a single sheet of metal and can be mou microstrip circuit without expensive coaxial connector. Second, we designed a unique E-shaped antenna with corrugation to reduce the width of antenna. We successfully miniaturised the fit it inside a thin (4mm) PC (or PCMCIA) card extension. In fact, we were able to squeeze in two of these antennas into a single PC card of standard width (54mm), for diversity communica achieve excellent isolation (>20 dB) and matching (<-10 dB) over the whole WLAN frequency band.

Theoretical Analysis of Photonic Crystal Structures

We have developed and successfully implemented, in both personal computers and supercomputers, efficient theoretical methods to analyse and design complex electromagnetic (micro optical) circuits based on photonic crystals (PC). Among our novel techniques is a PC-based Perfectly Matched Layer (PML) absorbing boundary for use with the finite difference time doma method in the analysis of waveguides in 3D photonic crystals.

We have applied these techniques to model wave propagation in various guiding structures such as bends, junctions and tapers in 2D and 3D crystals. From this analysis, we can transmission and reflection coefficients, propagation characteristics, phase and delay response, etc., of a component or a system.

New Closed-form Green's Functions for Microstrip Circuits and other Layered Structures

When combined with the spatial-domain Method of Moments (MoM), our new closed-form functions now enable efficient and accurate analysis of microstrip circuits and antennas with approximations. The key feature in this new MoM is that the four-dimensional integrations in MoM matrix elements can be solved analytically, completely eliminating the need for expensive integrations. Our new closed-form functions are simpler and more flexible than previous such functions, and (unlike previous ones) they do not require additional (Taylor series approximations. The computer time required for the analysis of an example microstrip line using the new MoM was three times less than the next best method, which required some integrations.

Integrated-design of Hybrid-resonator Antennas for Broadband Wireless and other Communication Systems

Shown below is the first baby of our recent project on broadband hybrid-resonator antennas. This first prototype of a Dielectric Resonator on Patch (DRoP) antenna demonstrated a ba 24%. This design also proved, both theoretically and experimentally, that the electromagnetic fields in a dielectric resonator can be efficiently coupled to the fields in a patch resona perturbing the radiation characteristics of each resonator.

Singularity-enhanced Finite-different Time-domain (FDTD) Method for Diagonal Metal Edges, Strips and Films

We developed the world's first and still the only singularity-enhanced FDTD method for metal edges not parallel to the grid. The edges were assumed to be diagonal to cell faces. Compare conventional spit-cell model, the computer memory required for an FDTD analysis of a structure could be reduced by up to 27 times and the computing time could be reduced by up to without sacrificing the accuracy of results. On the other hand, when the same grid was used, the accuracy of results improved by a factor of more than 3 compared to the split-cell model, a of more than 7 compared to the staircase model. The new equations were stable in all tests, and even in most demanding tests when the computational speed was further maximised by the time step (Dt) to the maximum allowed by the FDTD method! (i.e. stability factor of 1!)

The key to this success was the derivation of enhanced FDTD equations for nodes near the edges by considering rigorously the singular nature of the electromagnetic fields. The table here improvement of accuracy achieved by using the enhanced FDTD equations. The new equations are simple to implement in a standard FDTD code. They are ideal for the analysis and microstrip components and high-speed digital circuits where thin metal films or strips with diagonal edges are encountered.

Low-profile Dielectric-resonator (DR) Antennas

We designed and tested the world's first low-profile, circularly polarised, rectangular dielectric-resonator antenna (DRA) in 1995. The radiating element of this antenna is shown in the photo have also designed many other dielectric-resonator antennas, including a low-profile linearly polarised DRA, for various applications (see publications). We pioneered the FDTD analy antennas and published the first radiation patterns of a DR antenna obtained using the FDTD method in 1995.

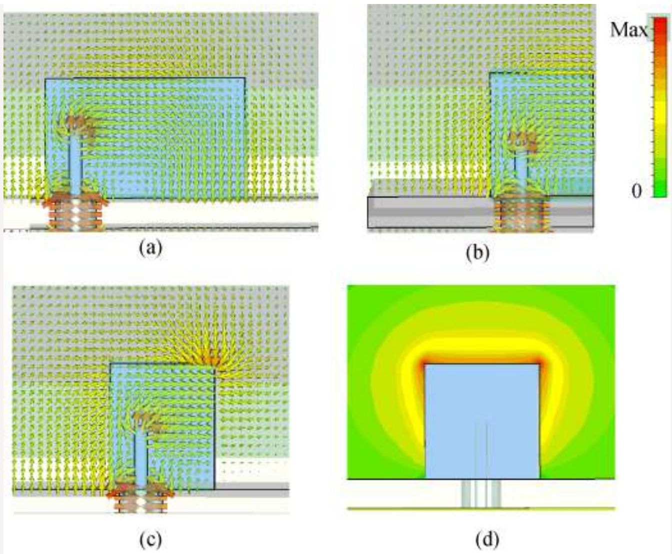

Australian Antenna Measurement Facility (AusAMF)

AusAMF is operated by a consortium of Australian Universities, Industry and CSIRO. The facility provides access to a shielded anechoic chamber offeri spherical near-field measurement capability for small antennas operating in the frequency range of 1-20 GHz, primarily for research purposes.

The facility was established under an ARC LIEF (Australian Research Council Linkage-Infrastructure Equipment and Facilities) grant and is hosted by C ICT Centre located in the Sydney suburb of Marsfield.

The facility currently consists of a spherical near-field turn-key NSI-700S-50 system within a 6m x 3.3m x 3.3m anechoic chamber. It has been designed to operate up to fre of 20 GHz, and uses an Agilent PNA (E8362B) as a receiver. The measurement system is capable of supporting a ~40kg antenna under test (AUT) up to a diameter of approximately 1.2m. I the standard gain horns and the probes available at AusAMF are as follows:

Waveguide Band	Frequency (GHz)	Probe	Standard Gain Horn
WR975	0.75 - 1.12	x	x
WR650	1.12 – 1.70	x	x
WR430	1.70 – 2.60	x	x
WR340	2.20 - 3.30	x	x
WR229	3.30 – 4.90	x	x
WR159	4.90 – 7.05	x	x
WR112	7.05 – 10.00	x	x
WR90	8.20 – 12.40	x	
WR75	10.00 - 15.00	x	x
WR51	15.00 - 22.00	x	x

COPYRIGHT (C) 2013 KARU ESSELLE ALL RIGHTS RESERVED

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/273396385

Smart Surfaces: Large Area Electronics Systems for Internet of Things Enabled by Energy Harvesting

Article *in* Proceedings of the IEEE · November 2014

DOI: 10.1109/JPROC.2014.2357493

CITATIONS
24

READS
362

9 authors, including:

Luca Roselli
Università degli Studi di Perugia

273 PUBLICATIONS **2,132** CITATIONS

SEE PROFILE

Federico Alimenti
Università degli Studi di Perugia

200 PUBLICATIONS **1,263** CITATIONS

SEE PROFILE

Paolo Mezzanotte
Università degli Studi di Perugia

144 PUBLICATIONS **1,211** CITATIONS

SEE PROFILE

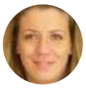

Giulia Orecchini
Università degli Studi di Perugia

24 PUBLICATIONS **347** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Development of ultra wide band analog predistorsion of LASER diodes for Radio over Fiber (RoF) systems View project

Project ADAHELI View project

All content following this page was uploaded by Luca Roselli on 30 July 2015.

The user has requested enhancement of the downloaded file.

INVITED PAPER

Smart Surfaces: Large Area Electronics Systems for Internet of Things Enabled by Energy Harvesting

This paper focuses on large-area "smart surfaces," RFID systems, and wearable RF electronics that could substantially benefit from multisource energy harvesting.

By Luca Roselli, *Senior Member IEEE*, Nuno Borges Carvalho, *Senior Member IEEE*,
Federico Alimenti, *Senior Member IEEE*, Paolo Mezzanotte, *Member IEEE*,
Giulia Orecchini, Marco Virili, *Student Member IEEE*,
Chiara Mariotti, *Student Member IEEE*,
Ricardo Gonçalves, *Student Member IEEE*, and Pedro Pinho, *Member IEEE*

ABSTRACT | Energy harvesting is well established as one of the prominent enabling technologies [along with radio-frequency identification (RFID), wireless power transfer, and green electronics] for the pervasive development of Internet of Things (IoT). This paper focuses on a particular, yet broad, class of systems that falls in the IoT category of large area electronics (LAE). This class is represented by "smart surfaces." The paper, after an introductory overview about how smart surfaces are collocated in the IoT and LAE scenario, first deals with technologies and architectures involved, namely, materials, antennas, RFID systems, and chipless structures; then, some exemplifying solutions are illustrated to show the present development of these concurrent technologies in this area and to stimulate further solutions. Conclusions and future trends are then drawn.

KEYWORDS | Energy harvesting; green electronics; Internet of Things (IoT); large area electronics (LAE); radio-frequency identification (RFID) systems

I. INTRODUCTION

Looking at the telecommunication (TLC) market development worldwide in the last five years, we see an average decrease of about 5% per year, in both gross domestic product (GDP) contribution and employment [1]. Against this trend, the global information and communication technology (ICT) market has remained grossly constant. Some compartments, in fact, are experiencing an opposite trend. Beside some sectors related to new consumer products, such as tablets and smartphones, new Internet of Things (IoT) related products are growing at two digits per year and some estimations from big players report a market value in the order of trillions of dollars in the next decade [2], [3].

The vision behind IoT is in fact to connect objects directly to the Internet so as to allow them to provide information directly to the web without any human intermediation [4]. This vision will have a great impact on several electrical and electronic technologies, ranging from the basic physical layers (technology platforms) to the highest ones: communication protocols, software, human interface, and information management.

Manuscript received May 4, 2014; revised September 1, 2014; accepted September 5, 2014. Date of publication October 13, 2014; date of current version October 16, 2014.

L. Roselli, **F. Alimenti**, **P. Mezzanotte**, **G. Orecchini**, **M. Virili**, and **C. Mariotti** are with the Department of Engineering, University of Perugia, Perugia 06123, Italy (e-mail: roselli@ieee.org).
N. Borges Carvalho and **R. Gonçalves** are with the Instituto de Telecomunicações-Dep. Electrónica Telecomunicações e Informática, Universidade de Aveiro, Aveiro 3810-193, Portugal.
P. Pinho is with the Instituto de Telecomunicações, Instituto Superior de Engenharia de Lisboa, Lisboa 1959-007, Portugal.

Digital Object Identifier: 10.1109/JPROC.2014.2357493

0018-9219 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

In terms of physical layer, IoT actually inherits the technologies developed for wireless sensor networks (WSNs), bringing the distributed nature of that to the extreme. WSN, in fact, can be synthetically seen as a mesh of sensor nodes purposely conceived to build and communicate a sensed image of a monitored area. Within IoT, the sensing nodes are not purposely deployed to monitor something specific; instead, they are simply hosted by "objects" in order to provide the information they are able to collect and make it available on the web. Conceptually speaking, there is not a big difference, but in terms of related technological challenges, there is. Nodes must be hostable by objects; objects are of many types, and the better the nodes fit themselves transparently to the hosting objects, the denser and more reliable the information they are able to provide.

Several challenges can be envisioned in this evolution: first, electronic systems must be mechanically flexible, thin, and miniaturized in order to conform to the shape of as many objects as possible; second, the adopted materials have to be as recyclable as the hosting objects in order to avoid pollution from guest apparatuses; last but not least, hosted nodes must be autonomous, because they cannot be either connected to the grid or powered by life batteries.

A great technological paradigm shift is going to be pulled by the development of IoT; green materials, autonomous systems, ultralow-power circuits, energy saving protocols, and energy harvesting (EH) are concurrently mandatory.

Conversely, this technological evolution is pushing new solutions and architectures, so far constrained by the limits of conventional technologies. The development of inkjet printing techniques, the introduction of very cheap and eco-friendly materials compatible with roll-to-roll (R2R) circuit realization techniques [5] and so forth, open new horizons also to large area electronics (LAE).

LAE, first applied to printed photovoltaic and organic screens [6], is actually at the onset of its development. The development of R2R techniques and related materials, in fact, is allowing for tremendously increasing the dimensions of LAE circuits and systems, from the present tens of centimeters to meters and beyond. Along this evolutionary scenario, new configurations and architectures, based on massive integration of large circuits over conformable surfaces, can be envisioned, opening the way to what can be called the smart surfaces (SSs) approach.

SS, in turn, can be seen as a branch of IoT evolution. Large 2-D arrays of autonomous sensor nodes, for instance, make possible granular tracking of whatever parameter, ultimately providing augmented imaging of the environment; large 2-D arrays of tags can provide a very low-cost platform for precise localization and location-based services (LBS), thus enabling the realization, for instance, of smart floors [7] and smart wallpapers [8]; inheriting quasi-optical approach [9] contact-less electromagnetic (EM) wave processing (filtering, frequency conversion, selective shielding, etc.) can be conceived even at low frequencies.

On the one hand, SS is thus an approach stimulated by technologies pulled by IoT; on the other hand, it contributes to a class of architectures, within LAE, that can even widen the huge horizons of IoT applications in a sort of virtuous circle.

According to this wide vision, the paper is organized as follows. A review of the technologies suitable for distributed systems implementation to a large extent is given in Section II; then, the explanation about how radio-frequency identification (RFID) can be considered one of the most suitable technologies for the implementation of IoT architecture, and how it can be naturally integrated with SSs, is provided (Section III). In order to deal with the implications of the development of RFID systems compatible with SS and LAE evolution of IoT, two specific sessions have been provided: the first relates to antenna implications (Section IV), and the second relates to electronic architectures for RFID tags (mainly chipless ones; Section V). In order to smoothly bring the reader from relevant technologies to applications, quasi-optics, as an example of general approach concurrently exploiting the mentioned technologies to provide a platform suitable for LAE, is described in Section VI. After this, some application examples of how SS concept can be articulated, according to the IoT paradigm, are described; specifically, smart floor (Section VII), smart shoes as useful subsystems for the implementation of smart floors (Section VIII), and energy skin (Section IX). In order to have a vision of the logical links behind the many topics dealt with in this paper, a synoptic picture can be found in Fig. 1.

II. TECHNOLOGY FOR LAE DISTRIBUTED SYSTEMS

Given the described scenarios of SS, it is clear that the technologies involved have to be compatible with LAE

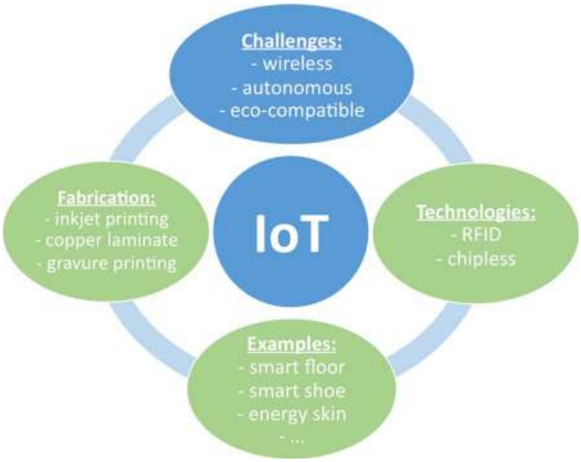

Fig. 1. *Synoptic view of the paper structure reflecting the IoT vision described in the Introduction.*

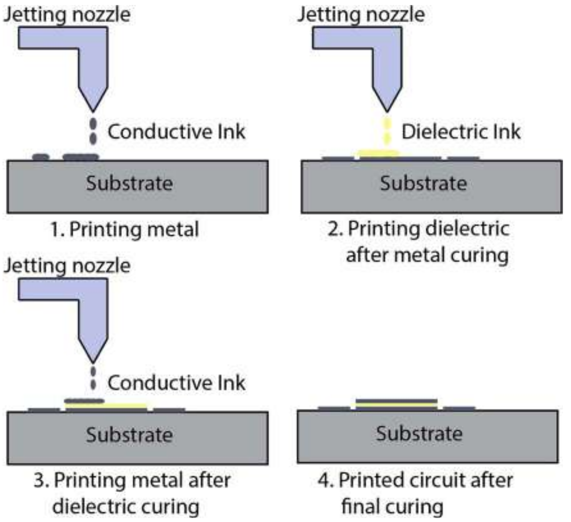

Fig. 2. ***Inkjet printing technique, synoptic description. The ink is deposited on the substrate with the desired pattern and after a curing procedure the next layer can be printed. The thickness of each film depends on the number of passes.***

approach, as described in the Introduction; according to [6], for example, the eco-compatibility of the materials, the mechanical flexibility of the substrates to make the device conformable to the hosting surface, and the life cycle (operability and biodegradability) have to be accounted for.

The new techniques proposed from different research groups [10]–[12] focus mostly on the printing techniques that can be industrialized easily with R2R methods, already investigated for the traditional printing on paper. Within the printing methods, one of the most attractive and investigated in the last decade is the inkjet printing of conductive and dielectric layers (Fig. 2). This technique allows circuits to be printed on almost any kind of substrate: from the photographic paper to liquid crystalline polymers (LCP) or Kapton, that are flexible, thin, and eco-compatible; to glass or poly methyl meta acrylate (PMMA, commonly referred to as Plexiglas) that are usually thicker but not flexible.

A first proof of concept for the inkjet printing technique at radio frequency (RF) and microwaves has been obtained on photographic paper by printing simple structures, such as antennas and other planar circuits and devices [10], [13]–[22], with a technology stack simply composed by the substrate and the conductive layer (usually realized by means of nanoparticle silver ink).

Recently, this method has been improved thanks to the development of new inks, usually composed by a polymer and a solvent (i.e., SU-8, PVP, PMMA, PEDOT, etc.); more complex fabrication procedures of multilayer structures are now possible, as shown in [23] and [24]. Another noticeable feature of this technology is represented by the resolution: conductive tracks of 50 μm of width and space gaps of 50 μm can be printed, allowing the design of millimeter-wave (mm-wave) frequency devices [25]. The combination of the quite high resolution (considering the simplicity of the technology) with the multilayer featuring also gives the opportunity to manufacture very easily matrix of passive and active devices that can be used in the SS development. It is worth noting that the process can be developed in a few steps based on the design and mostly on the number of layers needed. For example, a metal–insulator–metal (MIM) device can be realized by printing silver on the substrate [26], curing it in the oven in order to create a surface that, with the proper treatment (for instance, UV–Ozone exposure or preheating), is ready for the insulator printing. Then, the dielectric is cured as well and the last metal layer can be printed on top of it. The entire structure can be then cured at temperatures ranging between 130 °C and 200 °C.

In a perspective of industrialization of the inkjet printing technology, the hypothesis of R2R manufacturing of circuits is being investigated and some examples are already reported in literature for solar cells and other devices [27], [28].

Today, many pros have been mentioned for the inkjet printing method, and others can be found in the non-use of wasted chemicals (as is for the traditional lithographic technologies), in the no-need of clean-room environment for the fabrications, in the low-cost and rapidity of manufacture, and in a R2R perspective compatibility. However, the necessity of a curing procedure after the printing of a layer still represents the biggest inkjet limit for two main reasons: first, the time and type of curing are dependent on the inkjeted materials and on the material stack-up; second, especially on a laboratory level, the curing can cause misalignment issues, given the fact that the sample is removed and then replaced in the printer after the last layer curing.

Currently, the platform mostly used worldwide by researchers is the Dimatix 2800 DMP. In terms of inks, it is possible either to buy printable solutions or mix solutions in labs; for example, the combination of polymers and solvents allows layers to be printed with different electrical properties and thicknesses.

Besides the inkjet printing method, in [11], a new technique, suitable for LAE, has been proposed. It uses a metal (copper in that case) adhesive laminate technology based on the application of the standard etching process to an adhesive copper laminate material. This technique was already adopted to produce mm-wave circuits and to characterize the resulting compound substrates, as reported in [29]–[31]. A brief illustration of the metal laminate technique is here reported referring to Fig. 3.

The production process can be divided in five steps. The first step consists of the deposition of a (positive) photoresist film on the copper surface; then, the circuit layout is transferred to the photoresist by means of a photomask and ultraviolet exposure. After that, the unimpressed film of photoresist is removed using a NaOH

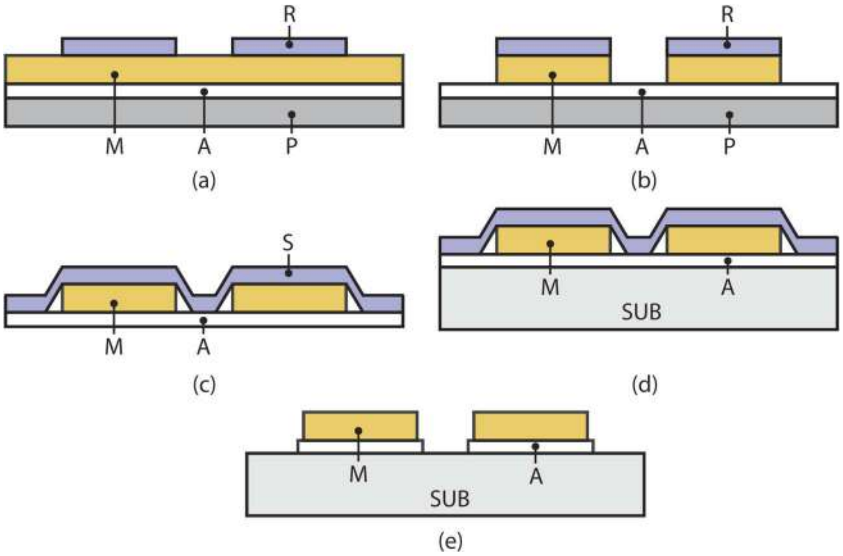

Fig. 3. ***Process steps for the circuit fabrication using the adhesive metal laminate. M: metal, A: adhesive, P: protection, R: photoresist, S: sacrificial layer, SUB: hosting substrate. After the standard lithographic procedure, the copper pattern can be transferred on top of any substrate by means of a sacrificial layer and exploiting the adhesive face of the copper tape.***

developer solution [Fig. 3(a)]. In the second step, the copper tape is wet etched. As can be seen in Fig. 3(b), in this way, the adhesive layer is exposed where the copper was removed, while it remains everywhere else covered by the original copper tape that serves as protection for the adhesive underneath.

The first two steps are, in this example, similar to those adopted in standard photolithographic technology. Moreover, different ways to remove the not needed metal can be used, as, for example, by means of numerical control pattern cutting plotters.

In the third step, depicted in Fig. 3(c), a sacrificial layer is stuck on the top copper side and, finally, the protection layer on the bottom is removed. The sacrificial layer is very important because it keeps the relative distances among the layout features constant even when these are not physically connected.

The fourth step is characterized by the transfer of the etched metal to the hosting paper substrate and, finally, the sacrificial layer can be removed [see Fig. 3(d) and (e)]. The last step is also useful to remove most of the exposed adhesive material.

With this method, quasi-fully-organic circuits and devices can be realized. The performance in terms of tracks width and pitch are, at present, a bit lower than what can be obtained with the inkjet printing, with the advantages of no curing processes, the possibility to fix devices on the circuit using standard soldering techniques, and a better conductivity (the conductivity of copper laminate, in fact, is one order of magnitude higher than that of the cured nanoparticle silver ink: 5.8×10^7 S/m versus about $6 \times 10^6 - 1 \times 10^7$S/m obtained with at least five layers of silver).

To verify the validity of the metal laminate technology at microwave frequencies, a 50-Ω microstrip line was manufactured exploiting the Mitsubishi photografic paper as the substrate (thickness 250 μm, relative permittivity $\varepsilon_r = 3.2$, and loss tangent $\tan\delta = 0.08$). The line is 30 mm long, and the measured scattering parameters are shown in Fig. 4. The same graph also compares the performance of a similar transmission line manufactured with an inkjet printing process. The performance of the

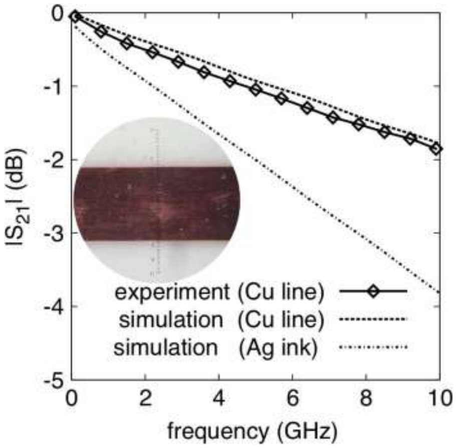

Fig. 4. ***Measured scattering parameters versus frequency for a 50-Ω microstrip line on paper substrate. The line, shown in the inset, is 30 mm long; the ruler division corresponds to 17.2 μm. The graph also reports a comparison with a microstrip line of equal dimensions made with an inkjet printing process (Ag ink, 3-μm thickness, $\sigma_{\text{ink}} = 1.1 \times 10^7$ S/m after curing). After [11].***

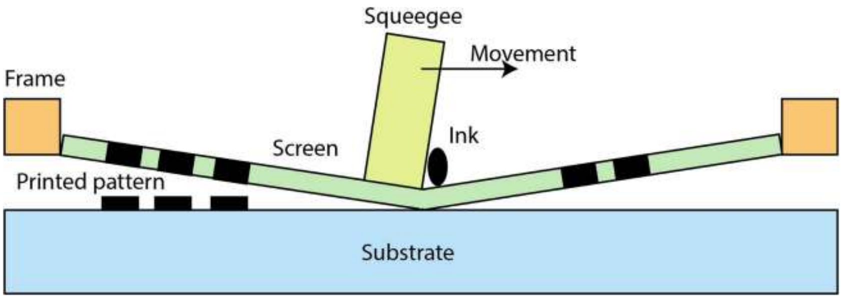

Fig. 5. ***Screen printing technique. The squeegee is used to press the screen with impressed pattern and transfer it on top of the substrate.***

metal laminate structure is superior compared to that of the inkjet printed line. At 10 GHz, the measured specific losses for the metal laminate microstrip are about 0.6 dB/cm [11]. Results at higher frequencies are also reported in [31]; in particular, an insertion loss of about 1.8 dBm/cm at 30 GHz is demonstrated.

The tradeoff is the adoption of a photolithographic step; it is worth mentioning, however, that the process is still compatible with R2R implementation.

Other possibilities for LAE are the screen printing [12] and the gravure printing [12], both of them adaptable to R2R industrialization techniques.

Screen printing consists of dragging a layer of ink across the surface of a screen and squeezing it through the open pores of the patterned mesh onto the substrate (Fig. 5). In general, the thickness of the printed layer and the achievable resolution depend on the density of the mesh and on the ink properties. Usually, the ink viscosity is in a range of 1–50 Pa $\times$ s, and this allows for printing with a resolution of about 100 μm and a thickness up to 100 μm. Until now screen printing has been adopted to realize low-resistance structures, solar cells, and field effect transistors (FETs), exploiting the possibility of printing very thick layers.

Gravure printing, also known as rotogravure, is considered a very high volume printing (Fig. 6) process, and it is being adopted to produce very conductive structures as, for example, capacitors, antennas, and organic devices such as pentacene-based diodes, organic light-emitting diodes (OLEDs), organic field effect transistors (OFETs), and organic thin film transistors (OTFTs).

The resolution can be of about 20 μm. The method is implemented by engraving the patterns into a metallic cylinder by a laser, by chemical etching, or electromechanically (as separate cells or intaglio trenches). Typically, the print pressure is high (1–5 MPa) in order to achieve a good ink transfer and to reduce the percentage of unprinted dots caused by the surface roughness of the flexible substrate. However, this high pressure of printing makes this technique suitable only for robust substrates with no soft, previously printed, layers that could be damaged.

To summarize, the technologies described in this section are characterized by some common features, such as: flexibility and conformability, compatibility with large area realization of circuits and interconnections, use of additive or mixed (subtractive/additive) deposition techniques, easy use of eco-friendly materials and, in some cases, compatibility with classical bond wiring as well as soldering techniques for electronic device placement.

Table 1 summarizes the main features of the described technologies.

III. RFID: A MIX OF CONCURRENT TECHNOLOGIES FOR SS IMPLEMENTATION

As mentioned in the outline paragraph at the end of the Introduction, after describing the technologies enabling SS development (Section II), here we illustrate RFID as a suitable means to transfer information through EM transmission between tags and readers to the Internet.

An RFID tag has a unique identification code and a memory used to store information, while the reader can

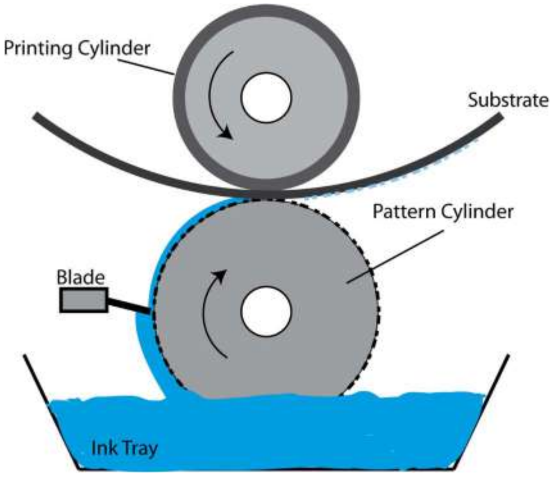

Fig. 6. ***Gravure printing process. Note that these machines can be from 20 cm to 3 m wide with a diameter ranging from 5 cm to 5 m.***

Table 1 Technologies Comparison

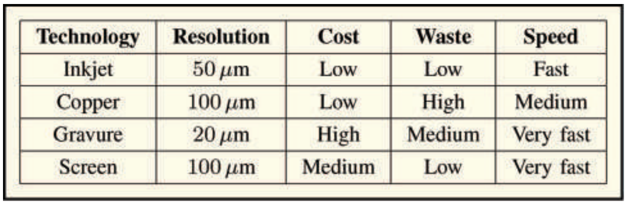

Technology	Resolution	Cost	Waste	Speed
Inkjet	50 μm	Low	Low	Fast
Copper	100 μm	Low	High	Medium
Gravure	20 μm	High	Medium	Very fast
Screen	100 μm	Medium	Low	Very fast

write or read data on and from the tag through RF transmissions. The tag is usually attached to an object that needs to be identified and monitored, or contains information to be read. In typical RFID applications [32], the main goal is to provide mapping of physical objects that are equipped with an RFID tag attached or embedded. In this way, RFID technology inherently leads to identifying, tracking, and localizing, but it can also enable the storage and transmission of information regarding object status and surrounding environments by embedding sensing capabilities.

As mentioned previously, IoT is most commonly described as a structured system of technologies concurrently working to capture meaningful data from objects and communicate them through networks to a decision-taking level. Useful keywords, whenever a definition of IoT is given, include: smart objects, autonomous logistics, machine-to-machine communication, RF technologies, centralized information, and so on.

In order to implement these functionalities, an IoT platform, an SS for instance, has to provide at least unique object identification together with wireless communication for data management.

The IoT physical layer can thus be naturally thought of as meshes of RFID and sensor nodes. The RFID tag that embeds a sensor can use the same working principles and protocols of the conventional RFID tags; a differentiation can be made, considering the way sensor tags are powered up, in active, semipassive, and passive tags; actually, the difference consists of the way they are powered.

Active tag sensors may use customarily conceived communication protocols or rely on RFID standards enabling the tag to be easily integrated into the existing wireless infrastructures so that they will not require expensive readers. In [33], a system architecture was developed to integrate WSNs and the RFID systems. Bluetooth and ZigBee technologies are adopted as the communication protocol of WSNs to meet the requirements of a large number of sensor nodes, large areas, and low cost.

Active tags with integrated sensors are used in several applications, including temperature and position monitoring, vibration detection, blood pressure, heart beat monitoring, and more. This type of sensor tags, besides having larger amount of energy provided by batteries, affords both a large range and multiple functionalities; some of them have also external buses that enable the use of connected external sensors [34].

Passive tags with an integrated sensor operate without battery, collecting the needed operational energy from external environment sources. The main requirements of this class are high energy conversion efficiency, large storage capability, and overall low-power consumption. Passive systems have usually low operating range and limited processing capability.

In this field, research activities are directed toward ultralow-power design of integrated sensor tags [35], antenna design for improving reading range [36], sensor overall performance optimization [37], and development of optimum protocols allowing for additional power saving [38], [39].

Typical applications are, for instance, temperature monitoring [40], photodetection [41], motion detections [42], etc.

Semipassive tags use both batteries and energy sources coming from the environment; they can integrate more operational capabilities than the passive ones, since they can exploit a higher amount of energy. For the same reason, semipassive RFID tags are more suitable to integrate a sensor; the operational methodology is similar to the passive one, using the reader signal to interrogate and cause a response from the tag. The primary difference is that the semipassive tag does have a battery used not to generate a response, but only to power electronics, like sensors incorporated in the tag, while exploiting RF circulation to reply to the reader interrogation. Just like passive tags, semipassive ones are limited in terms of slow reading speeds and short reading distances. In [43], the design, realization, and experimental validation of a battery-assisted RFID tag with sensing and computing capabilities, conceived to explore heterogeneous RFID-based sensor network applications, was presented.

Dealing with battery-less devices that are, by far, the most interesting ones for IoT, LAE, and SS applications, we can now focus on the typical energy sources adopted. No battery devices can exploit solar scavenging [44], [45], vibrational [46], [47], or RF radiations [48]–[52].

In many cases, wirelessly powered systems can harvest energy from the incident RF waves, thus generating the required direct current (dc) voltage to power the system. The EM energy present in the environment may be significant in some regions (some urban areas or indoor sites); here the use of proper conversion devices may allow enough energy to be scavenged. RF power harvesters usually consist of an antenna, an impedance transformation network, a rectifier, and a storage element. In [53], a design example of integrated rectifier antennas (rectennas) for wireless powering at low incident power densities, from 25 to 200 μW/cm^2, is given.

Regarding mechanical energy scavengers, they may be categorized upon the kind of energy source exploited; there are scavengers providing energy from a constant motion over extended periods of time (e.g., turbine air flow) or others based on intermittent motion (e.g., human step). Particular attention is given to vibrational energy sources; in this case, the amount of energy that can be harvested depends mainly on the vibration amplitude and frequency and on the vibrating mass (mass of the harvesting device) [54]–[56].

In order to power the systems by using energy sources available in the working environment, proper energy transducers thus have to be chosen. Focusing on distributed systems (IoT, LAE, and SS), it is worth referring to typical parameters for optimum transducer design since, in these applications, the different kinds of energy available in the various scenarios may lead to the need of integrating more than one energy conversion system into the same device, in order to take advantage of all the accessible energies. The main design parameters are the available power output, the electrical impedance, and the operating voltage. Table 2 summarizes them for the most common energy transductions [57], [58].

To conclude this section, the mentioned reasons why RFID can be considered a useful technology for the implementation and development of LAE and, more specifically for SSs, are summarized. RFID systems, especially in the passive version of tags, are communication systems characterized by the following characteristics: compatibility with very low-power communication protocols, compatibility with the constrains posed by the hosting objects, especially when they are implemented by adopting the technologies described in Section II, inherent capability to gather information from the environment (sensor tags) and from the hosting objects, a low number of electronic devices (inherent circuital simplicity) and, finally, proven compatibility with printing technologies.

For all these reasons, RFID technology has to be considered the leading one to pursue the development of LAE and SSs as a communication platform for the evolution of IoT.

IV. ANTENNA DESIGN FOR NODES

Distributed systems for ubiquitous electronics (UE) and WSNs [59], [60] are, today, increasingly employed for monitoring and sensing applications. These systems can be seen as networks of nodes, massively distributed in the environment and communicating wirelessly with a base station. This massive deployment can be designed and structured to monitor a set of specified parameters by means of purposely conceived sensor nodes, as in the classical WSNs, or it can be obtained by integrating nodes into existing objects, which is the vision of IoT evolution. In general, the nodes can have different functionalities, however they have to provide information collected by embedded sensors (such as, for instance, identification codes, position and monitored environment parameters as well as object status) whenever it is required.

In this scenario, the RFID technology, summarized in Section III, is attractive. When RFID technology meets sensor nodes, information goes from the node to the Internet via a question-and-reply protocol between nodes, hosted by tagged objects, and readers.

Thinking about distributed systems and specifically SS, it is natural to think about radiative coupling in the available industrial–scientific–medical (ISM) frequency bands, including mainly the ultrahigh frequency (UHF) and microwave frequencies (around 0.9, 2.4, 5.8 GHz, and higher), exploiting the communication standard such as the WiFi [61]. Recently, also the ultrawideband (UWB) has been considered in order to minimize the limitations of narrowband systems (mainly localization accuracy and sensitivity to interference) [62].

Multistandard systems are often adopted for large compatibility and higher use flexibility. An example of this recent approach is given in [61] where near-field UHF nodes are combined with far-field transceivers. This allows application fields with compromised radiation capability,

Table 2 Transducer Characteristics for Several Forms of Ambient EH [57], [58]

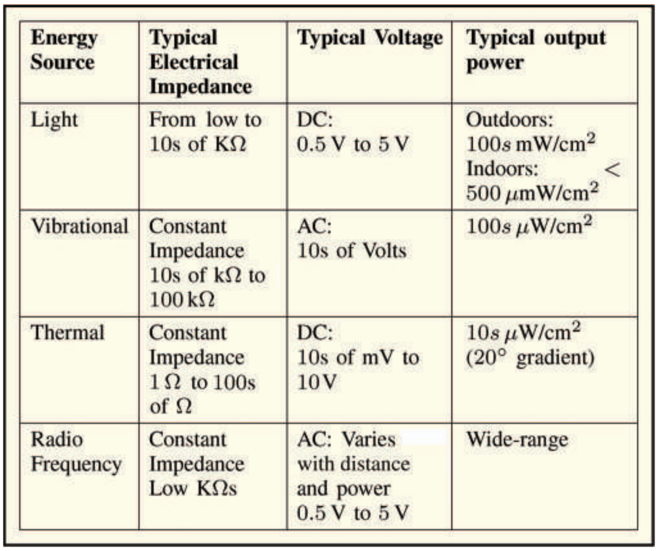

Energy Source	Typical Electrical Impedance	Typical Voltage	Typical output power
Light	From low to 10s of KΩ	DC: 0.5 V to 5 V	Outdoors: $100s$ mW/cm^2 Indoors: $<$ 500 μmW/cm^2
Vibrational	Constant Impedance 10s of kΩ to 100 kΩ	AC: 10s of Volts	$100s$ μW/cm^2
Thermal	Constant Impedance 1 Ω to 100s of Ω	DC: 10s of mV to 10 V	$10s$ μW/cm^2 (20° gradient)
Radio Frequency	Constant Impedance Low KΩs	AC: Varies with distance and power 0.5 V to 5 V	Wide-range

such as in proximity of metal and liquids at short ranges, to be improved.

It is clear that antenna systems are key elements in these apparatuses. The foundation of antenna design rules can be found in [63]. The antenna design specific for distributed systems, in turn, has to take into account several instances. Beyond the typical design criteria, other aspects are: the communication standard, the form and the size in agreement with the application, the communication range as a function of the equivalent isotropic radiated power (EIRP) limits, of the environment and of the orientation, the reliability under several conditions and processes, and the material cost considering massive production [64]–[67]. Among them, the operating environment is the most critical aspect that influences the node behavior [68].

As a consequence, antenna topology and materials for SS and IoT systems, in general, have to be chosen in order to make the antennas capable of operating in variable environmental conditions and to be integrated into objects. As an example, let us think about wearable antennas for body area networks (BANs): they must be washable, flexible, and able to shield the body from radiation [69].

In general, antennas for SS should be flexible, planar, and low profile; moreover, by adopting low-cost materials and technologies (see Section II), cost per unit surface is low, thus large arrays and big elements can be adopted more easily, ultimately allowing for easy developments of LAE distributed systems and SS, even at low frequency.

Materials compliant with SS implementations must be, in turn, mechanically robust, flexible, and very low cost, while exhibiting acceptable EM performance. These materials must also exhibit high tolerance levels in terms of bending repeatability. Flexible materials, often used, are polymers and polymer-ceramics composites such as the poly-di-methyl-siloxane (PDMS) and the titanate-based ceramic [68], which provide flexibility, robustness, and resistance to harsh environments in order to protect antenna and circuit. Moreover, when eco-compatibility has to be guaranteed, recyclable materials and related technologies have to be adopted; among them, cellulose substrates in combination with inkjet printing of conductive inks [18], [70] or with conductive adhesive laminate technology [11] are increasingly frequently adopted. The aforementioned techniques, compatible with the well-known industrial printing and the R2R processes [5], are in agreement with the low-cost fabrication and are expected to facilitate widespread and very low-cost electronics for distributed systems. These fabrication technologies can also be adopted for electronic devices on flexible polymeric substrates. An example of cellulose-based antenna for smart floor applications, such as indoor localization, is documented in [71]: a low-profile planar loop antenna is realized on cellulose substrate exploiting adhesive copper tape.

Although beyond the scope of this paper, which focuses on SS, it is worth mentioning here that this evolution toward the development of more and more sophisticated "unconventional" materials for electronics has a great impact on other IoT related developments such as, for instance, implantable systems.

When applications require devices with soft visual impact (think about SS, such as smart windows, glass integrated inside and outside buildings and glass for automotive applications, or smart skins [72]), transparent and conductive materials, such as transparent conducting oxides (TCOs), and in particular indium-tin-oxide (ITO) and ITO-based multilayers, have been proposed [73]. These materials afford a good compromise between minimum electrical resistivity and high optical transmittivity in the visible light spectrum when patterned on see-through materials such as glass or transparent polymers. A technique to improve the efficiency of the transparent antennas, without affecting their transparency, is the application of a layer of highly conductive coating or metallization strips on the antenna edges where the current density is high [72].

Among SSs, or in general distributed systems within the IoT vision, a fast growing field of research is represented by wearable electronics for BANs [74]. In this case, the integration of electronic devices in the garments requires the development of wearable and washable antennas and antenna arrays on fabrics. Washable antennas can be realized by covering textile antennas by a breathable thermoplastic polyurethane coating in order to protect the device against water absorption and corrosion [69].

Although dimension constraints are not very critical for distributed systems on cheap materials, they cannot be neglected thoroughly. At RF, for example, high permittivity materials [68] and magnetic composite substrates [75] can be adopted. Artificial magnetic conductors (AMC) can be employed alternatively to the magnetic composites with similar results [76]. These approaches are simply demonstrated from the wavelength equation: wavelength is a function of the relative permittivity and permeability and adopting suitable materials to increase these values, the effective wavelength, and therefore the antenna size, decreases. Anyway, the use of these materials introduces a tradeoff between dimensions and performance: generally the substrate losses decrease the antenna radiation efficiency, thus affecting the communication range of the overall wireless system, as documented in [68], [75], and [76].

In terms of design methodology, the introduction of unconventional materials as well as the environmental constraints posed by distributed systems and systems for IoT influence and drive the design strategies. Some solutions are proposed in the literature to cope simultaneously with these constraints; in [68], shielded configurations have been analyzed to mitigate environmental sensitivity. In [69], an example of robust patch antenna for BAN application that reduces the effect of the body on the performance is illustrated; in [77], the use of broadband antennas, such as the Chebyshev monopoles with a fractional bandwidth of about 40%, is proposed to reduce the frequency shift introduced by the background materials or by the bending effects.

In SSs, and in general in IoT applications, antennas are used both to transmit and receive information and to harvest the necessary energy from the surrounding environment. Regardless of the fact that the available energy is randomly released in the environment by existing services (GSM, WiFi, and so on, in urban and semi-urban environments [78], [79]) or purposely delivered to empower a distributed population of nodes [80], antennas have to be able to harvest EM energy. This kind of "double effect" antenna is widely referred to in the literature as "rectenna": it combines rectifying circuitry with EM to electrical energy transduction.

The fundamental codesign techniques of antennas and rectifiers, focusing on the improvement of the conversion efficiency exploiting matching circuits, are described in [53]. Harmonic-balance and load-pull simulations are often used in order to find the optimal impedances that allow best efficiency to be obtained as a function of the EM incident power.

Multiband, broadband, and circularly polarized antennas are often adopted in EH applications to reduce the overall sensitivity to the environment. An example of planar antenna array for EH applications, in agreement with the SS principle, is documented in [81]: spiral antennas are proposed in order to provide a wide operating frequency band, and the rectification circuitry is optimized and characterized for low incident power levels. Similar antenna array for UWB EH is shown in Fig. 7.

A rectenna can be used in addition to the communication antenna or, alternatively, the same antenna can be used for both EH and data transmission. The use of different antennas is generally adopted if harvesting and communication are carried out at different frequencies and, mainly, in order to enhance harvesting efficiency without decreasing communication performance. In any case, this requires a codesign and can affect the node dimension and complexity. Alternatively, a single antenna can be a feasible compromise if communication and harvesting share the same frequency band and if the antenna harvesting efficiency is not the main issue; examples are the passive RFID tags empowered by the reader with a standard RFID antenna able to provide both EH and communication capability, as documented in [64] and [65].

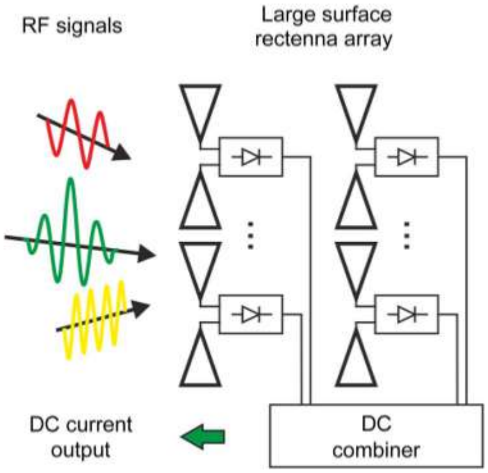

Fig. 7. *Scheme of the rectenna array able to collect RF energy even in a broadband (depending of the kind of receiving antenna element), according to the solution adopted in [81].*

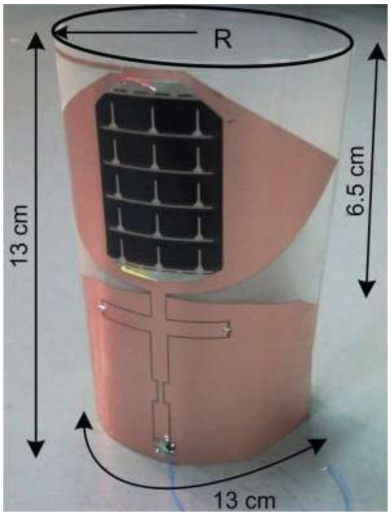

Fig. 8. *Hybrid rectenna for solar and EM–EH. The hybrid rectenna consists of a broadband antenna and a solar cell combined in the same structure. After [82].*

However, the EM–EH is not the only solution to energize nodes and SS elements: hybrid EH, in fact, can be performed by exploiting different energy sources; that is, combining the rectennas with other sources of energy.

Recently, a low-cost and conformal structure for hybrid EH, based on low-cost and flexible poly-ethylene terephthalate (PET) substrate and an amorphous silicon solar cell, has been proposed in [82]; other examples are discussed in [20] and [83]. The use of low-cost, printable photovoltaics deposited on flexible substrates to form part of the antenna radiating structure is documented in [84]. In all these contributions, the main challenge appears to be the codesign of the antenna and solar cell in order to avoid reciprocal detrimental effects. The antenna is designed with a solar cell integrated on top of the radiating surface to obtain good radiation characteristics and to incorporate connections for the extraction of the dc provided by the solar cell. In [82], it was demonstrated that, by adopting thin film solar cells and paying attention to not exceeding the perimeter of the RF antenna, the solar cell can be integrated so as not to significantly affect the performance of the antenna. The hybrid rectenna for solar and EM–EH documented in [82] is shown in Fig. 8.

In terms of industrial assembly, the use of new materials and the need for low-cost and large-scale fabrication techniques would be fostered by architectures more insensitive against geometrical errors and placement misalignments. Following this direction, EM coupling can be exploited instead of ohmic contact to connect antennas on a flexible substrate to Si chips or, more generally, to the active circuitry [85]. This technique exploits a planar heterogeneous transformer, with the primary and secondary windings implemented on the antenna substrate and the Si chip, respectively. In this way, the galvanic contacts and the soldering process are avoided, and the final structures have been proven much more robust to alignment errors than traditional ohmic techniques without any significant performance degradation. By artificially applying a misalignment of 150 μm, which is the common dimension of a soldering pad, the technology proposed in [85] causes a loss increment of only 0.1 dB, due to the applied misalignment, while the same misalignment, when adopting ball grid arrays, would have implied missing the contact. This approach is quite attractive especially in combination with nonrigid materials where mechanical stress can result in significant substrate deformations. Recent contributions show the application of the aforementioned approach to inkjet printed antenna on PET substrate coupled to the chip for wireless nodes [86] and a textile patch antenna, magnetically coupled to the active circuitry, suitable for garment integration (see Fig. 9) [87], the reflection coefficient of which is shown in Fig. 10.

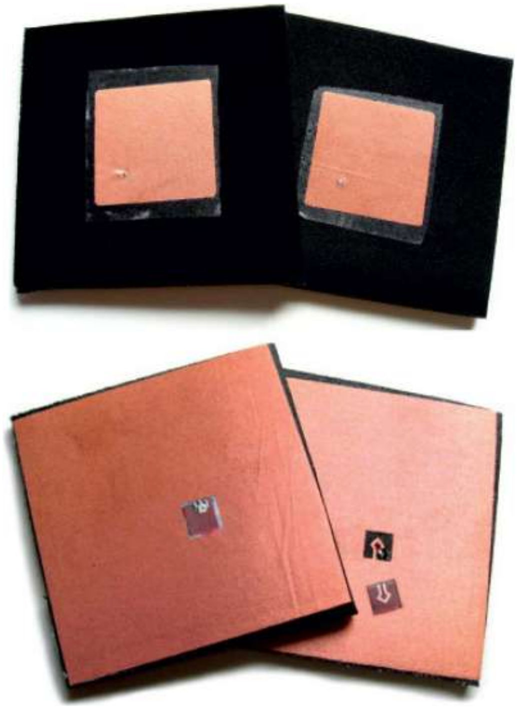

Fig. 9. ***Example of textile patch antennas operating at 2.4 GHz. The antenna and the ground plane are made of Flectron, a conductive textile, and the substrate is black foam. These materials are flexible and integrable in the garments. Magnetic coupling is exploited to connect the antenna and the active circuitry on flexible substrate. After [87].***

In order to testify the robustness of the proposed solution against uncertainties inherent to textile and garment electronics assembly, the radiation efficiency of the antenna, as a function of horizontal and vertical misalignment of the two transformer windings, used to perform the magnetic coupling, is plotted in Fig. 11(a) and (b), respectively. This technique can be extended also to interlayer communications of multilayer devices, and an example is documented in [88], where the circuitry fabricated on different garments communicates by means of overlapped coils.

V. CHIPLESS APPROACH

It has already been discussed that SS represents a possible technology for the development of killer applications within the IoT paradigm. Thanks to the convergence of several technologies, ranging from EH and sensing to RFID and LAE, SS can afford complex functions. For example, SSs can be used in intelligent buildings for structural monitoring and alarms (fire, for instance). To achieve these goals, SS must be equipped with a huge number of tags, each one monitoring a small portion of the surface itself.

The passive RFID systems can be divided into two families, namely, chip-based and chipless tags. The first family usually exploits a low-power complementary metal–oxide–semiconductor (CMOS) chip to implement the main tag functions [RF carrier rectification, dc voltage regulation, amplitude shift-keying (ASK) demodulation, IDentification code (ID) decoding, load modulation, etc.] and thus can be adapted to sensor applications by a few additional circuit blocks [signal conditioning,

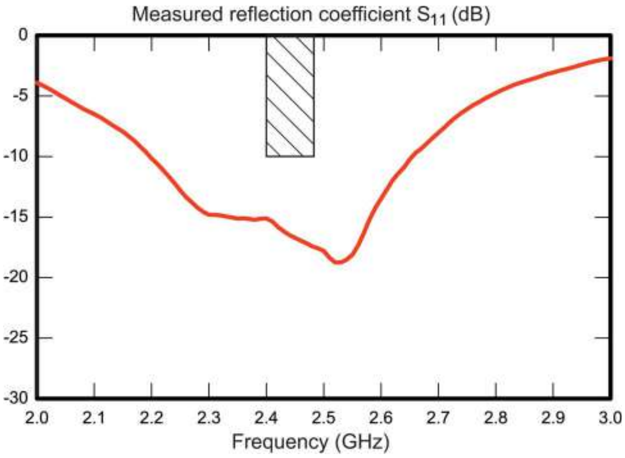

Fig. 10. ***Measured reflection coefficient of the previous textile antenna in the frequency band 2.4–2.48 GHz. After [87].***

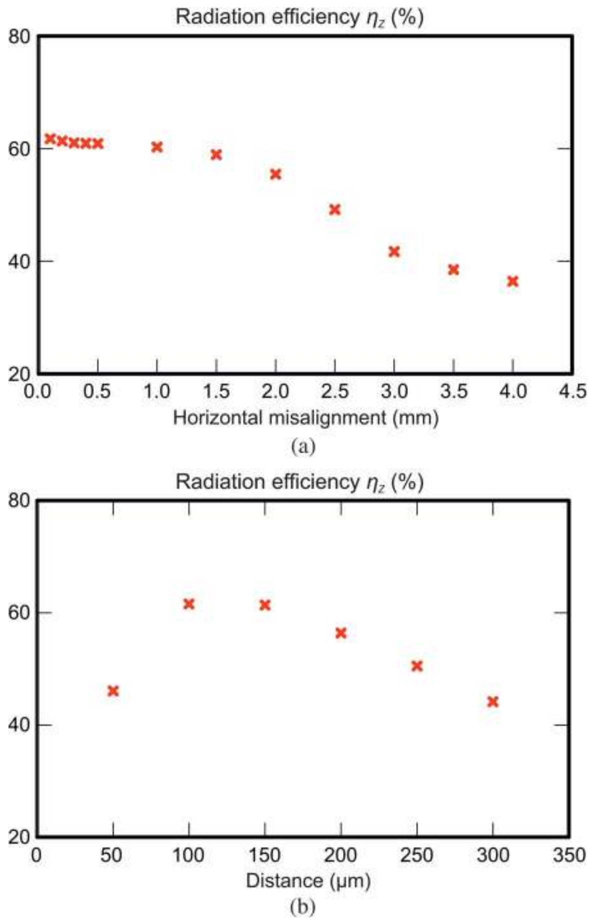

Fig. 11. ***Variation of the radiation efficiency evaluated versus (a) horizontal and (b) vertical misalignment between the antenna and the circuit windings. This test allows the robustness of the antenna performances against mounting uncertainties inherent to textile electronics to be evaluated at glance. After [87].***

analog-to-digital converter (ADC), etc.]. The main advantage of such an approach is the digital modulation of the transmitted signal, and thus the flexibility of the related data treatment, as shown in [89] and [90]. The production costs of chip-based tags are mostly associated with the heterogeneous integration of the silicon chip with the antenna [85], [91], the latter being manufactured typically on a flexible substrate or a textile substrate [92]–[97].

In order to reduce the aforementioned costs and to save as much energy as possible, solutions requiring minimum amount of electronics must be pursued. For this purpose, the chipless tag family was introduced and applied to wireless sensing in the recent years [98]–[103]. It is worth recalling here that the introduced buzzword "chipless" stands for a family of tags that are actually able to implement their functionalities by using only a few lumped components and passive distributed elements, without requiring any electronic integrated circuit (IC), the chip indeed, to be mounted.

The standard chipless tags (i.e., only coding and no sensing) exploit, basically, two mechanisms, namely, time-domain and frequency-domain scattering [64]. In the first case, the elapsed time between multiple reflections is used for coding. These reflections are obtained, at tag level, by connecting the antenna to a transmission line structure with several discontinuities. In the second case, instead, the tag has a specific frequency signature that is decoded at the reader level; to implement such a behavior, a coded series of resonators is realized on the tag antenna in order to reflect or not the corresponding frequency. Both approaches (time and frequency domain) lead to tag circuits that are not necessarily easy to miniaturize; nevertheless, it is worth noting that they are compliant with the LAE paradigm.

When dynamic sensor information has to be added to the static identification, we talk about chipless RFID sensor tags; these exploit an antenna, the electrical properties of which are controlled by the change of the physical parameter to be detected. There are, mostly, two proposed approaches for this. The first one is to induce a permanent change in the antenna property when a certain critical threshold (acceleration, temperature, fluid level, etc.) is exceeded [104]; this is useful for alarm type operations. The second one exploits a sensing load, i.e., a load the impedance of which is controlled by the sensed variable, connected to an antenna [105]. In both cases, the whole wireless sensor system (tag and reader) needs to have absolute accuracy, thus limiting the system performance with respect to both distance and fabrication tolerances.

A different method has been recently introduced in [106]. In this paper, a novel sensing principle is associated to the generation of an intermodulation signal from a tag, the latter being illuminated by two waves at different frequencies. The advantage of this idea is that the tag response is generated at a known frequency, thus the presence or the absence of such a signal can hardly be misinterpreted. Similar techniques have been used in harmonic radar systems [107] and in one-bit frequency doubling tags [108].

In line with the above ideas, a novel and completely original approach was proposed in [20] and [109] to solve the issue of absolute accuracy of the most of the passive chipless RFID sensors. The chipless tag described in [109] is based on the harmonic radar concept [110], [111], i.e., on a tag that, being illuminated by a carrier at frequency f_0, is capable of generating the second harmonic $2f_0$.

A simplified block diagram of the tag is illustrated in Fig. 12. For this purpose, the sensor information is encoded as the phase difference between two signals, one acting as the reference signal for the other one. First, the tag receives a carrier at frequency f_0. Then, two equal signals at frequency $2f_0$ are generated by means of a diode-based frequency doubler and a power divider. At this point, one of the two signals is phase-shifted using a passive sensing element. Finally, the $2f_0$ signals are reirradiated by

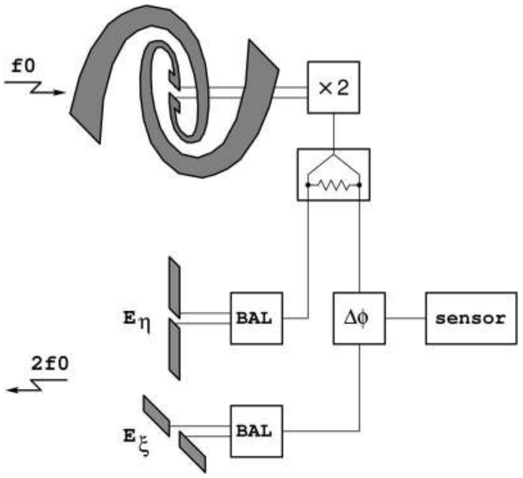

Fig. 12. ***Simplified block diagram of a harmonic chipless tag with sensing capabilities. After [109].***

exploiting two orthogonally polarized antennas. With this approach, the sensor information can be extracted by a suitable reader equipped with two complex (I/Q) receivers.

It must be observed that chipless tags, due to their inherent simplicity and extremely low number of lumped components, are easily realizable on flexible substrates by means of metal laminate [29], [30] and inkjet printing technologies. In particular, antennas [112], diodes [113], and passive sensing elements [114]–[117], i.e., all the main tag components, have already been experimented in cellulose-based materials. This will soon make SSs, equipped with sensing chipless tags, feasible at affordable prices.

VI. QUASI-OPTICS

"Quasi-optics concerns the propagation of EM radiation when the size of the wavelength is comparable to the size of the optical components (e.g., lenses, mirrors, and apertures) and hence diffraction effects become significant" [118].

In optics, operations are usually performed by using lenses. At EM frequencies, by following the quasi-optics analogy, lenses become large arrays, several wavelengths in size, of unconnected elements, each of them being able to get part of the signal incident on the array (usually focused by using a Gaussian beam), to carry out an operation and to irradiate the modified signal. The sum of the radiated signals modified by the array elements ends up in a radiated beacon elaborated with respect to the incident one [118]–[120]. From a historical point of view, quasi-optics was first introduced experimentally by Heinrich Hertz in the late 1880s [120] when he demonstrated the possibility of collimating EM signals on a multiwavelength surface by using cylindrical reflectors and studying effects that, until that time, were observed only at infrared and visible EM spectrum.

After more than 70 years, owing to the increasing demand for bandwidth and to the development of optical wavelength systems, quasi-optics experienced its palingenesis at the millimeter and submillimeter wavelengths.

We can classify the quasi-optical components into three main categories: frequency independent, frequency selective, and active devices. Frequency-independent surfaces include: delay lines, polarizing grids, hybrid junctions [121], attenuators [122], power dividers [123] and combiners [124], nonreciprocal devices [125], [126], absorbers, and calibration loads. Frequency-selective surfaces [127]–[129] include: inductive grids, capacitive grids, resonant grids, thick structures (perforated plates), and interferometers. Active devices include: oscillators [129], [130], amplifiers [131], [132], mixers [133], [134], phase shifters [135], [136], multipliers [108], [137]–[139], and switches [140]. This approach has been used so far mainly at millimeter and submillimeter frequencies because, being the size of the apparatuses proportional to the wavelength of the radiated wave, lower frequencies would have been prohibitive due to the very large size required. With the present development of technologies for LAE, however, quasi-optics has become a feasible approach even at RF and microwave bands, where large arrays of operational elements can be conceived.

Inheriting the quasi-optical approach, contact-less EM wave processing (filtering, frequency conversion, selective shielding, etc.) can be implemented even at low frequencies (in the order of megahertz). Two pioneer examples are switches [140] and the cross-dipole frequency doubler, proposed in [108], and implemented in [141], where, among various operating principles, the generation of harmonics is chosen to demonstrate the feasibility of such a component at microwave frequencies.

The layout of the proposed quasi-optical frequency doubler is shown in Fig. 13. The structure is inkjet printed on a cellulose-based substrate (a piece of photografic paper from Kodak). It consists of two crossed $\lambda/2$ dipoles. The longest dipole receives the incoming power at the fundamental frequency $f_0 = 3.5$ GHz, while the shortest one transmits the generated power at the doubled frequency $2f_0$ in an orthogonally polarized orientation. The length of the dipole operating at f_0 is 32 mm, while the $2f_0$ dipole is 16 mm long. The width of the tracks used to implement the dipoles is about 2 mm. The multiplication is provided by four diodes in a bridge configuration, thus forming a fully balanced multiplier unit.

Due to the omnidirectional nature of the dipole antenna, the harmonic signal is irradiated with the same intensity in the azimuthal plane (small differences are due to both the dielectric substrate and the planar nature of the dipole conductors). This means that the harmonic signal is reflected toward the reader and the transmitter toward other directions at the same time. A measurement of the received power at $2f_0 = 7$ GHz is reported in Fig. 13(b). Here the interrogation distance is 10 cm and the

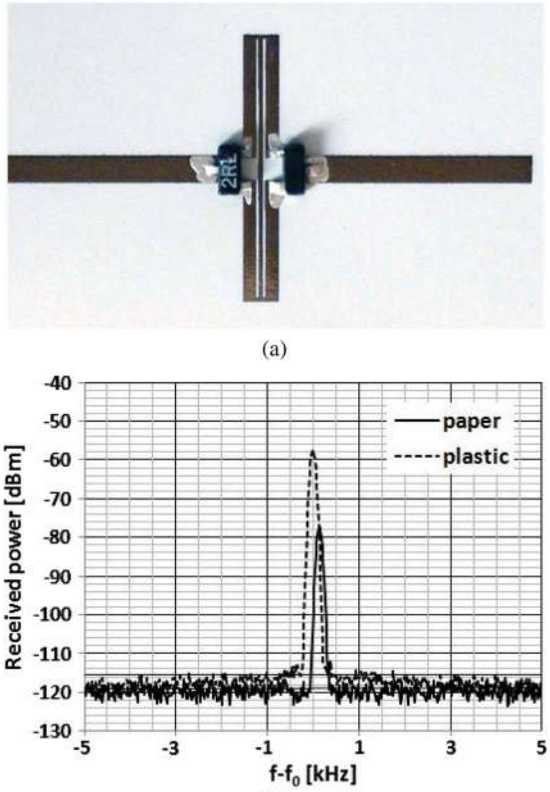

(a)

(b)

Fig. 13. ***Cellulose-based prototype of the (a) crossed-dipole frequency doubling tag and (b) measured second-harmonic response. The measurements have been carried out interrogating the tag at $f_0 = 3.5$ GHz and receiving the $2f_0 = 7$ GHz frequency component. The interrogation distance is 10 cm, and the transmitted power is equal to 20 dBm. The transmitter is equipped with a two-element Yagi antenna at f_0, while the receiver uses a helix antenna at $2f_0$. The fundamental frequency dipole is 32 mm long. After [141].***

transmitted power is equal to 20 dBm. The transmitter (reader side) is equipped with a two-element Yagi antenna at f_0, while the receiver (reader side) uses a helix antenna at $2f_0$. In these conditions, the power received from the cellulose-based (paper) tag prototype at $2f_0$ is about -80 dBm. The same structure implemented on a plastic substrate produces more power, mainly because of a better frequency tuning. An improved efficiency could be achieved by: 1) adopting a frequency multiplier with better harmonic terminations; and 2) using directive antennas in order to address the power only toward the reader.

In the frame of the IoT, the crossed-dipole tag is a very simple one-bit tag that can alarm a system when it is placed within the range of the reader. However, with very simple modifications, this idea can be used to implement a variety of on/off sensors (the paper substrate can be easily torn when a certain mechanical strain is exceeded).

The crossed-dipole structure can also be used as a building block to form arrays, an example being shown in Fig. 14. These arrays are mainly intended for LAE applications where a complete surface can be interrogated and answered by generating a harmonic signal. The fabricated prototype uses a Rogers substrate, thus it is not flexible. However, adopting a cellulose-based substrate, a flexible structure can be built. The bending capability of cellulose-based circuits is quite good and only limited by the discrete components mounted on them.

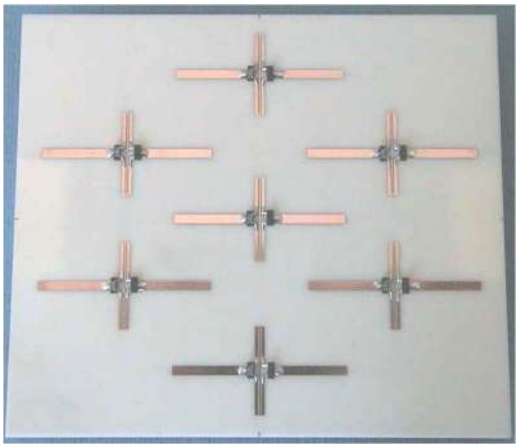

Fig. 14. ***Example of quasi-optical frequency doubler based on an array of crossed dipoles. This structure is implemented on a Rogers substrate, but this can be substituted with a cellulose-based material to reduce costs and obtain mechanical flexibility.***

The second example consists of a paper-based contactless frequency doubler for harmonic RFID applications [142]. The doubler, realized on paper substrate, generates the harmonic signal by means of a single Schottky diode. The system operates at 7.5 and 15 MHz, and these frequencies are chosen, without lack of generality, to accomplish the realization of a fully organic frequency doubler exploiting paper printed coils and organic diode (pentacene-based), the present frequency limit of which is around 15 MHz [143]. Fig. 15 shows a picture of the organic tag, while Fig. 16 shows the doubling efficiency of the organic tag versus transmitting/receiving (TX/RX) distance, assuming TX power as a parameter (for the sake of completeness, it is worth noting that TX and RX antennas of the reader are equal to the RX and TX antennas of the tag, respectively).

As a final remark, it is worth noting that, to the authors' knowledge, only planar developments of quasi-optical arrays have been proposed so far; reasonably, combining quasi-optics with LAE related technologies can lead, in the near future, to new SS solutions only in a limited planar extent.

VII. SMART FLOOR

The new concept of smart cities is generating a group of new scenarios that will impose a new way people interact with the environment. This immersion brings the IoT to a new level of relationship with people and environment,

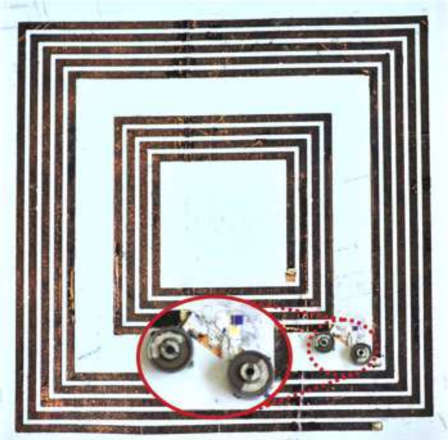

Fig. 15. ***Fully organic tag prototype of the frequency doubling tag operating at 7.5–15 MHz. After [142].***

where the scenario around is aware of people's needs and can interact with their way of living.

This new concept brings IoT to areas that cover broad topics, including transportation, energy distribution, data communications, and all technologies in a sense that they will become transparent to the normal users and citizens.

The growth of smart cities will significantly increase the quality of life, the reduction of energy waste, and the availability of information by using ICT. All these developments have, as a motivation, the overall inclusion of people into the city in a way that ICT should become completely embedded into our surroundings.

These new concepts are viable only if the smart environments are enabled by IoT. IoT is actually taking shape and is growing everyday through the number of devices that get connected to each other.

Nevertheless, most of these devices that team up with each other to build up the IoT world are powered by some kind of power sources. This is actually the main limitation of present IoT solutions. In order to be able to create real IoT environments, RFID, wireless power transfer (WPT), and EH devices appear as a major enabling technology for the IoT, due to their inherent simplicity and ability to provide sensing to remote objects without the need for constant powering.

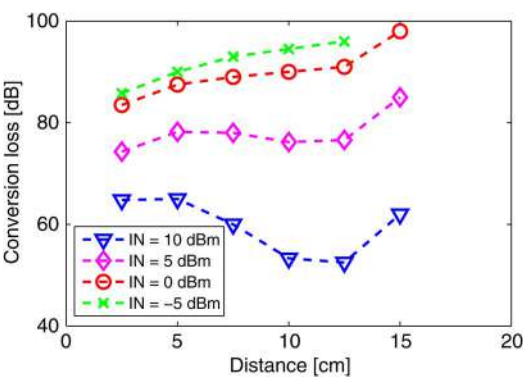

Fig. 16. ***Conversion (doubling) loss of the reader tag harmonic RFID system as a function of the reader tag distance, assuming the output reader power as a parameter. After [142].***

In this scenario, the concept of a smart floor becomes real. Smart floors should be capable of interacting with the environment seamlessly and be aware of persons and objects on top of them.

The focus of this section is going to be smart floors as an information system, where RFID capabilities are used to provide an identification and information source embedded into the floor. This massive RFID immersion will allow to create low-power high accuracy location, navigation, and, generally speaking, an information system that can team up with passers in a noninvasive way.

In this approach, the system uses passive RFID tags that are spread beneath a flooring in order to create a map that can be read with a mobile unit that might be self-powered through energy harvested from the movements of the subject to be located [144], [145].

This new location system based on smart floors can actually be extremely competitive when compared with existing solutions; among them, we can take into consideration those based on image processing [146]; they provide a high level of information at the price of high cost, power consumption, and intrusiveness. Another approach is based on pressure sensors [146], [147], providing actually a different type of smart floors; pressure sensors are nonintrusive and can be less expensive in terms of equipment when compared to the imaging systems; however, they can still be quite expensive to install into the floors; moreover, they do not allow for seamless identification of an individual in a given space. Eventually, ultrasounds [148] and RF localization system detection [149] have to be listed; although very flexible, they require dedicated communication infrastructures, additional complexity and costs, and they are mostly sensitive to environmental changes, both in terms of accuracy and functionalities.

The just mentioned smart-floor-based system tackles most of these challenges as it is very low cost, able to have a good, and precisely predictable, accuracy, and able to detect different users in a given space.

A. RFID-Based Smart Floor Including Material Aspects

Smart environments and floors should be completely seamless to the users. This entails the need to embed electronic devices into the environment in a noninvasive way. Nevertheless, this embedding of electronic devices, mainly wireless transceivers into the floor, is not a simple matter. Floors, in fact, have been made by almost the same kind of materials for centuries; tile industries are weakly available to introduce new materials to match

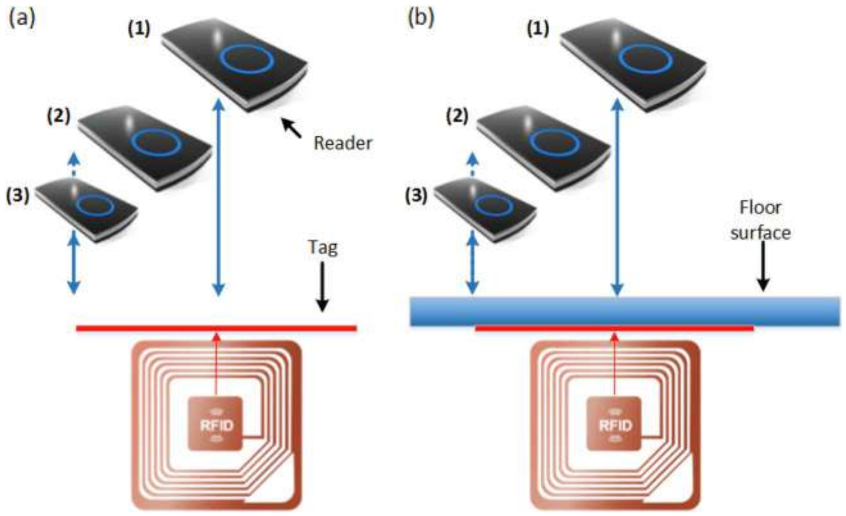

Fig. 17. ***Readable distance test scheme (a) without and (b) with tile. After [144].***

requirements posed by embedded electronics. Similar considerations can be made about the consolidated techniques to build apartments and houses: they are quite mature and whatever adaptation of them to embedded electronic systems is unrealistic. The consequence of this is that constitutive elements of SSs embedded into the floors must be suitable to materials and technologies already present: materials available are those adopted for tiles and SSs must be adapted to building constraints.

In this section, we focus on the most important approaches for the characterization of the EM properties of the materials, commonly adopted in floors, and their interactions with our electronic radio systems. As an example, ceramic tiles are going to be selected as the floor element where electronic tags have to be embedded.

Ceramic tiles have been used for ages to pave floors. These make up a great deal of most floor types in public buildings, from airports to banks, therefore, they are the most interesting flooring elements to investigate.

As the first approach, the study of the influence of some materials for floor surfaces, such as ceramic, cork laminates, and wood, in the communication between the transmitter and receiver devices has to be made; in this case, low-frequency (13 MHz) inductors are going to be used as the wireless interface.

In order to do the measurements, traditional RFIDs are used, which simplify the measurement approach significantly. A test setup is shown in Fig. 17, and experimental results are summarized in Table 3. What can be seen from the results presented in the table is that none of the tested materials affect the electrical characteristics of the RFID propagation significantly.

Another approach to the measurement is to evaluate the impact of the presence of metal inclusions beneath the ground; these can appear from pipes or from buildings supporting structures. In order to test this scenario, the presence of a metal sheet close to the coil was tested to verify its impact on the RFID reading distance. The test setup is presented in Fig. 18. In this case, when the reader is aligned (1) with the center of the coil beneath the tile, a minimum distance of 10 mm between the metallic sheet and the tag must be guaranteed in order to allow for the proper reading of the tag. In the other approach, the edge aligned scenario (2), the minimum distance between the metallic sheet and the tag increases to 30 mm to allow for communication.

Table 3 Results From Average Readable Distance Test

Tile type	Aligned (1)	Edge aligned (2)	Unaligned (3)
Without tile	4.5 cm	4.3 cm	2.4cm
Ceramic tile 4 mm thick	4.5 cm	4.3 cm	2.4cm
Ceramic tile 6 mm thick	4.5 cm	4.3 cm	2.4cm
Ceramic tile 7 mm thick	4.5 cm	4.3 cm	2.4cm
Cork laminate 5 mm thick	4.5 cm	4.3 cm	2.2cm
Wooden floor 8 mm thick	4.3 cm	4.2 cm	2.2cm

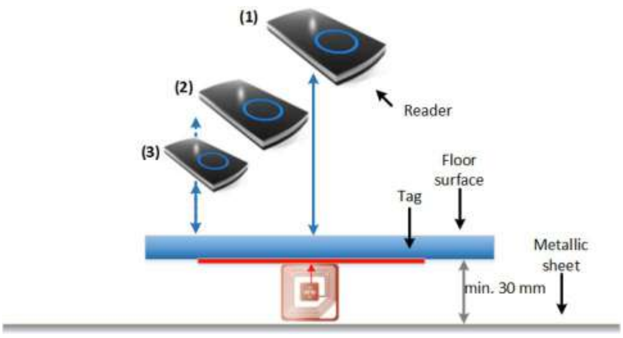

Fig. 18. ***Readable distance test scheme with metallic sheet presence. After [144].***

This test allows us to conclude that, in a real scenario, when considering the typical distance from the antenna to the floor, where the reader is usually embedded into a mobile device, a minimum distance from the tile to any metal inclusions in the ground has to be guaranteed. As seen, this safe distance should be at least 30 mm.

B. Navigation Approach Based on Smart Floors

A meaningful example of the applications of the smart floor concept is the demonstrator developed at the University of Aveiro, Aveiro, Portugal. It allows a nomadic RFID reader on top of it to be identified and localized by means of a smart floor.

The prototype in the present state is composed of four main components: a set of tagged tiles forming the smart floor, the nomadic reader, a wireless communication device to connect the reader to the information system, and a computer on which the location and navigation software is installed. A potential scenario of application for the monitoring of elderly housing is shown in Fig. 19. In the demonstrator presented, the RFID reader is connected to the radio unit and inserted into the user's shoes in order to read the passive tags as the user walks around a given area. The tile IDs are detected by the reader and sent to a centralized information system using, without lack of generality, a nonstandard wireless communication system, purposely developed in the laboratory.

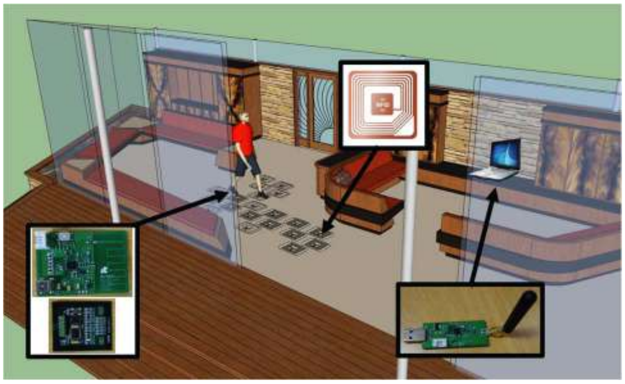

Fig. 19. ***Example of smart floor application scenario.***

The software in the computer then decodes the ID of the smart tile, and correlates it with a preloaded referenced map, thus being able to provide the precise position of the walker in the environment.

The location can be subsequently translated to a navigation system, and combined with artificial intelligence to allow navigation, situation awareness, etc.

Plenty of applications and related solutions can be envisioned on the basis of this smart floor approach: robot control, automatic pallet managements, prevention in harsh environments, and so forth.

VIII. SMART SHOES AND BAN

The aforementioned smart floor (Section VII), initially proposed to localize moving subjects in a given area, actually allows for the implementation of a very low-cost, accurate, and reliable system able to provide each kind of detectable information from an equipped area to the Internet.

To exploit the potentiality of a smart floor, it is obvious that developing purposely conceived communication units acting, on the one hand, as readers for tag tiles and, on the other hand, as communication systems with the web is needed.

One of the investigated systems for this is the so-called smart shoe [21]. Smart shoes are intended as shoes incorporating a reader able to illuminate the tag embedded into the tile and a wireless interface to communicate data to the web. To this extent, smart shoes can be seen as hubs of BANs able, on the one hand, to get information from the surrounding environment (body included) and, on the other hand, to communicate the collected information to the Internet. According to common buzzwords, they can be seen as the second technological layer, just above the first physical one consisting of tagged objects, in a hierarchical picture of IoT. Alternatively, they provide the cloud of nomadic readers with networking capabilities required for the implementation of the so-called networked RFID (N–RFID) [150] strategies. In the referred structure [21], the necessary antenna system is inkjet printed on paper, thus exploiting the process described in Section II, and it is a trademark-logo-shaped dual-band antenna working at 900 MHz and 2.4 GHz. A block diagram and its equivalent realization are shown in Fig. 20: it is worth noting that the reader itself works at 13.5 MHz exploiting the near-field communication (NFC) concept, according to what is stated in the smart floor section; as a consequence, there are no issues about the interference between the dual-band antenna and the NFC reader. Moreover, an approach similar to [145] is proposed to power the reader: the idea here is as well to harvest the energy produced by the human walking by means of piezoelectric materials embedded into the sole of the shoe.

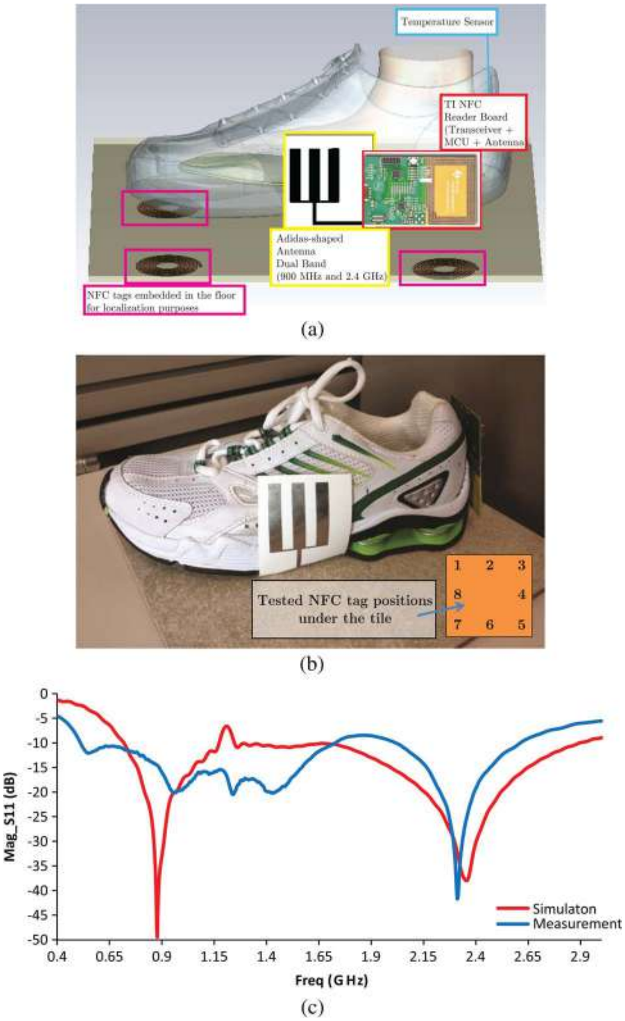

Fig. 20. ***(a) Block diagram, (b) photo, and (c) dual-band antenna results of the smart shoe that works as an energy autonomous reader for smart floors able to collect the localization data from the smart tiles and send them to a centralized system. After [21].***

Fig. 21(a) shows the shoe with the mounted electronics consisting of a trademark-logo antenna inkjet printed on a flexible paper substrate together with the circuits in which the components [Fig. 21(b)] are fixed by conductive epoxy glue. A piezoelectric push button was embedded in the shoe to scavenge energy from human walking strike. This energy is then collected by a simple electronic circuit based on off-the-shelf components and used to supply the RFID transmitter. As reported in [145], the system is composed by the power-generator/energy-conversion device, the energy storage device, the power regulator circuit, and the RF transmitter that can broadcast the tag ID and the stored information.

When the push button is pressed, an inner spring is compressed and when the pressure exceeds a fixed threshold, the spring-loaded hammer will be released to deliver the dynamic mechanical force to the piezoelectric component; once the hammer strikes the piezoelectric element, a pressure wave is generated and reflected a few times between the hammer and the element, creating a mechanical resonance [151]. As a result, the generated output voltage follows closely an alternating current (ac) signal course. The signal out of the piezoelectric element is characterized by high voltage and low current; therefore, since this RF circuitry requires lower voltages at higher currents, a step-down transformer is used for better impedance matching to the following circuitry.

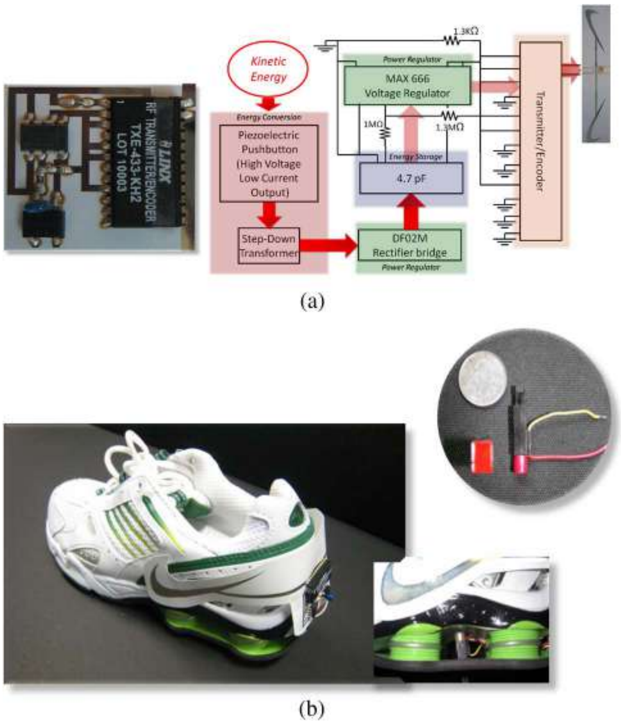

Fig. 21. ***(a) Mounted electronics on shoe with piezoelectric energy harvester and block diagram of the transmitting tag, driven by the EH unit. (b) Photograph of the assembled prototype showing the key components packaged on an organic flexible substrate. After [145].***

An energy storage device is adopted to collect the electrical energy and to provide it to the transmitter even when the external power source is temporarily unavailable; in fact, the RF transmitter takes tens of milliseconds to transmit one complete word, while the piezoelectric push button harvests energy in a transient time. To store the collected energy, a 4.7-μF tank capacitor was chosen. An ac–dc full wave diode bridge rectifier is used to convert the ac voltage from a piezoelectric element into a dc source and a dc linear regulator is added to adapt the dc voltage across the capacitor to 3-V transmitter operating voltage. For this purpose, a MAX666 low-dropout linear regulator, providing a stable 3-V supply until the tank capacitor charge is drained below the reference level of the dc linear regulator itself, is chosen. The transmitter used, the LINX

Table 4 Shoe Energy Performance Summary

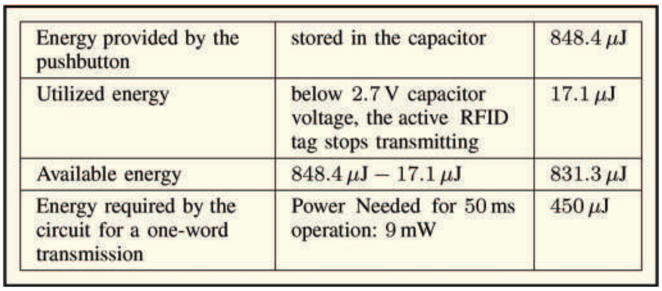

Energy provided by the pushbutton	stored in the capacitor	848.4 μJ
Utilized energy	below 2.7 V capacitor voltage, the active RFID tag stops transmitting	17.1 μJ
Available energy	848.4 μJ − 17.1 μJ	831.3 μJ
Energy required by the circuit for a one-word transmission	Power Needed for 50 ms operation: 9 mW	450 μJ

TXE-433-KH2 [152], is able to combine a highly optimized RF transmitter with an onboard encoder.

The maximum voltage across the capacitor is 19 V; therefore, the energy stored in the capacitor can be calculated to be close to 850 μJ. After the strike, the regulator becomes active, providing power to the tag circuitry; its 3-V output remains constant before it starts to drop to zero according to the capacitor discharge rate. When the voltage across the capacitor drops below 2.7 V, the active tag stops transmitting. The resulting energy left in the capacitor and not used is about 17 μJ. The required time for successful completion has been measured to be close to 50 ms; the total circuit energy consumption for a 50-ms transmission is then approximately 9 mW (Section IV). Consequently, the energy required for a one word transmission is approximately 450 μJ, much less than the power collected by the energy conversion and regulation unit. The energy performance of the smart shoe is summarized in Table 4.

"Smart" solutions like smart shoes are suitable for smart floor interaction and for collecting information (position included) from them. The smart shoe has actually a double functionality: on the one hand, it is a means to collect information from the environment (smart floor included) and from the on-body sensors and apparatuses; on the other hand, it transfers data to the Internet, via WiFi, for instance, while likely receiving control and feedback signals. To this extent, it contributes to the development of wearable electronics for body-centric communication systems, according to the scheme of Fig. 22. This class of devices is based on the BAN concept (see Fig. 22): a network of wearable computational units, sensors, energy harvesters, and transceivers networked and implemented around the human body. The information is acquired by means of several sensors and, after processing and digital modulation, is transmitted by means of off-body transceivers. The most attractive application of BANs and wearable systems, so far, is probably health monitoring [88], but the number of foreseeable applications is constantly increasing (personal fitness monitoring, personal audio system, personal alarm set, and so on). Today, a big contribution to the BANs is being given by the interaction with portable devices such as smartphones which can be exploited as gateways.

Low-power ICs, sensors, and wireless communications are key elements of the BANs, and the use of EH techniques and mixed technologies such as RFID, as in the above described smart shoe, can contribute to the development of autonomous BANs. According to the notation adopted in [74] and [153], BANs can exploit wireless communication not only for off-body communications with other systems, but also for on-body and in-body communications to interconnect networked devices [wireless body area networks (WBANs)]. Generally, the proposed WBAN solutions are based on existing communication standards; but, recently, the IEEE 802.15.6 working group has provided a standard for short-range, wireless communications in the vicinity of, or inside, a human body based on ISM bands for BANs [154]. In this rapidly evolving scenario, smart shoes can be seen, in turn, as a promising mean to interface information coming from smart floors to the BANs and ultimately to the Internet, directly or via portable devices.

IX. ENERGY SKIN

Within SSs, another area is emerging, which requires dc energy. The availability of devices that can actually increase the battery life of our electronic gadgets [155] is the holy grail of IoT; the idea is actually to convert energy from

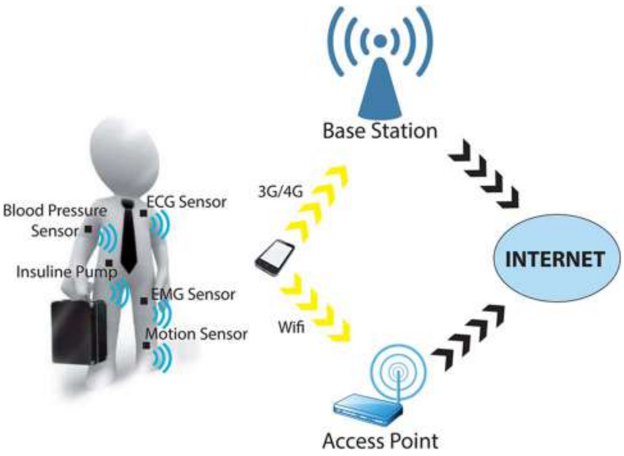

Fig. 22. ***BAN integrated with IoT: the data sensed on the human body and collected by smart devices are sent to the Internet through the standard protocols like third generation/fourth generation (3G/4G) and WiFi.***

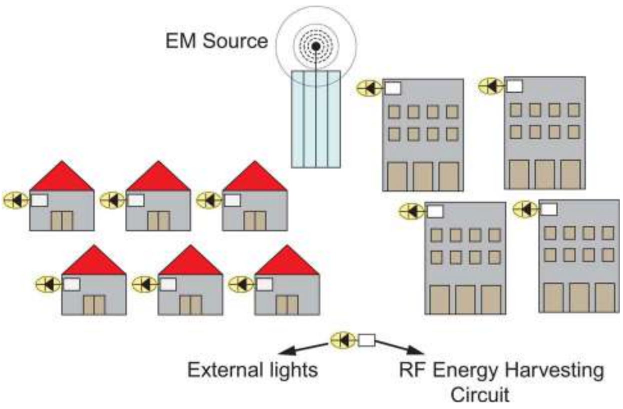

Fig. 23. ***EM source, likely supplied by solar energy in order to keep eco-friendliness, is used to broadcast energy to the urban neighborhood where houses are equipped with local RF–dc converters able to power supply local devices (external lamps in the example).***

different sources, as solar to dc [156], EM radio signals to dc, thermal to dc [156], vibration to dc [157], etc. [158].

Portable converters from solar to dc are today commonly available on the market at very low prize. Similar portable scavengers can be easily envisioned even for other energy converters. On this basis, plenty of sophisticated and more complex solutions are being proposed. Just as an example, we report the solution in Fig. 23, where the solar energy is used to supply an EM source that broadcasts RF energy in the surrounding urban environment; this energy is converted to dc energy at building level and, in turn, exploited to supply external local lamps. This approach was conceived for space applications [159], [160], but could be used in certain conditions in earth stations as well. The concept can actually move forward and be integrated into a smart skin, in the sense that all devices could be powered up by the skin energy conversion that a person can carry (vibrational, thermal, RF, etc.). This will become possible by integrating paper and textile radio electronic circuits in the smart skin surfaces of our clothes.

X. CONCLUSION

This paper focused on smart surfaces concept. This concept was introduced as a way to provide solutions to societal instances by means of technologies and techniques that today are gaining great momentum thanks to IoT development. The foreseeable IoT evolution, in fact, calls for systems and subsystems that are more and more energy conservative, fully autonomous (battery-less and not connected to the grid), as recyclable as possible (ultimately 100% biodegradable), flexible, and extremely low cost to be embedded in as many objects as possible. In order to meet these expectations, several technologies have to be implemented concurrently. This paper illustrates some of them, namely, recyclable and reliable materials, printing and R2R compatible techniques, conformal antennas on unconventional materials, and chipless RFID.

Subsequently, it has been shown how these technologies enable a fairly new branch of IoT development, represented just by SSs.

SSs are in their infancy; nevertheless, some examples, enabled by the aforementioned technologies, can be cited. Among them, the way to inherit “quasi-optics” and extend them to RF frequencies even below gigahertz is shown, and the smart floor implementation as a means to provide precise localization as well as assisted paths platforms has been reported. Energy skin as a means to convert several energy sources (solar, thermal, vibrational, EM, etc.) in electric one is described as well.

The reported examples must be researched as a new paradigm for potential promising developments aiming at providing new LAE solutions for the future networked society. ■

REFERENCES

[1] European Union (EU), “The ICT sector and RDI,” Digital Agenda Scoreboard, 2012. [Online]. Available: http://ec.europa.eu/digital-agenda/sites/digital-agenda/files/scoreboard_ict_sector_and_rdi.pdf

[2] J. Manyika *et al.*, “How the Internet of Things is more like the Industrial Revolution than the Digital Revolution,” MGI Rep., 2013. [Online]. Available: http://www.mckinsey.com/insights/business_technology/disruptive_technologies

[3] D. Evans, “The Internet of Things: How the next evolution of the internet is changing everything,” CISCO White papers, 2011. [Online]. Available: http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[4] I. G. Smith, “The Internet of Things 2012 New Horizons,” IERC European Research Cluster on the Internet of Things, 2012.

[5] J. Morse, “Nanofabrication technologies for roll-to-roll processing,” in *Proc. NIST-NNN Workshop*, Boston, MA, USA, Sep. 2011, pp. 1–31.

[6] A. Arias, J. MacKenzie, I. McCulloch, J. Rivnay, and A. Salleo, “Materials and applications for large area electronics: Solution-based approaches,” *Chem. Rev.*, vol. 110, no. 1, pp. 3–24, Jan. 2010.

[7] T. Kampke, B. Kluge, and M. Strobel, *Field and Service Robotics*. New York, NY, USA: Springer-Verlag, 2008, ser. Springer Tracts in Advanced Robotics.

[8] M. Shur, “Giant area and flexible electronics,” in *Proc. IEEE Nanotechnol. Mater. Devices Conf.*, Oct. 2006, vol. 1, pp. 264–265.

[9] M. Steer *et al.*, “Global modeling of spatially distributed microwave and millimeter-wave systems,” *IEEE Trans. Microw. Theory Tech.*, vol. 47, no. 6, pp. 830–838, Jun. 1999.

[10] A. Rida, L. Yang, R. Vyas, and M. Tentzeris, “Conductive inkjet-printed antennas on flexible low-cost paper-based substrates for RFID and WSN applications,” *IEEE Antennas Propag. Mag.*, vol. 51, no. 3, pp. 13–23, Jun. 2009.

[11] F. Alimenti, P. Mezzanotte, M. Dionigi, M. Virili, and L. Roselli, “Microwave circuits in paper substrates exploiting conductive adhesive tapes,” *IEEE Microw. Wireless Compon. Lett.*, vol. 22, no. 12, pp. 660–662, Dec. 2012.

[12] D. Tobjörk and R. Österbackaa, “Paper electronics,” *Adv. Mater.*, vol. 23, no. 17, pp. 1935–1961, 2011.

[13] P. Calvert, “Inkjet printing for materials and devices,” *Chem. Mater.*, vol. 13, no. 10, pp. 3299–3305, 2001.

[14] G. Orecchini *et al.*, “Inkjet printed organic transistors for sustainable electronics,” in *Proc. 60th Electron. Compon. Technol. Conf.*, Jun. 2010, pp. 985–989.

[15] L. Yang, G. Orecchini, G. Shaker, H.-S. Lee, and M. Tentzeris, “Battery-free RFID-enabled wireless sensors,” in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 2010, pp. 1528–1531.

[16] G. Shaker, S. Safavi-Naeini, N. Sangary, and M. Tentzeris, “Inkjet printing of ultrawideband (UWB) antennas on paper-based substrates,” *IEEE Antennas*

COUPONS AVAILABLE ACTIVATE COUPONS

advertisement

RFID READER DETECTOR AND TILT-SENSITIVE RFID TAG

Technology (/technology/) > Science | by nmarquardt (/member/nmarquardt/) Follow

286,565 531 157

Posted Nov. 10, 2008 | BY-NC-SA

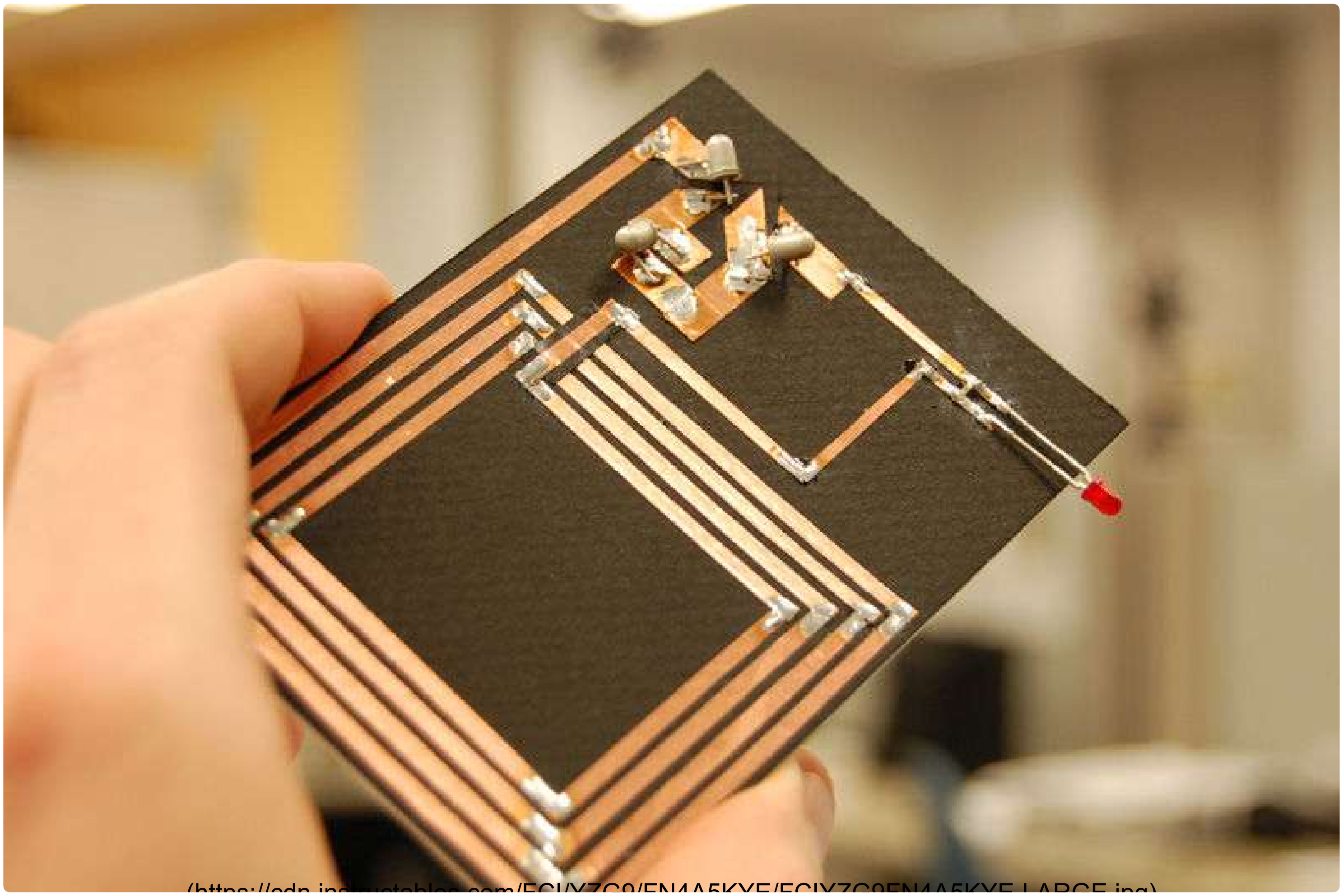

(https://cdn.instructables.com/FCI/YZG9/FN4A5KYE/FCIYZG9FN4A5KYE.LARGE.jpg)

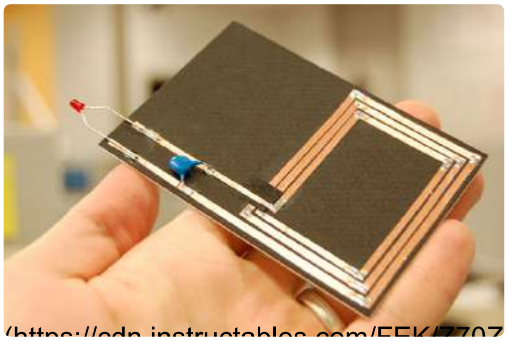

(https://cdn.instructables.com/FFK/Z70Z

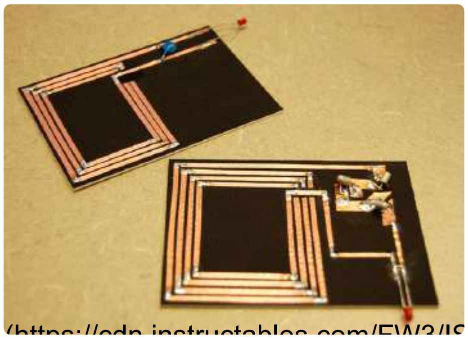

(https://cdn.instructables.com/FW3/IS

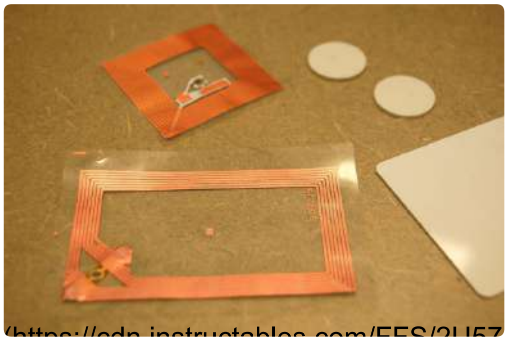

(https://cdn.instructables.com/FFS/2H57

The 'rub' (http://www.usingenglish.com/reference/idioms/there%27s+the+rub.html)
Want to detect the presence of RFID readers? Want to control when a RFID tag is active or readable? We describe how to do both using bits of copper and card, and some readily available electronics hardware.

Longer preamble
Radio frequency identification (RFID (http://en.wikipedia.org/wiki/RFID)) is rapidly growing in popularity. RFID *tags* are found everywhere. They're attached to container freight, in those funny-looking white labels you find in newly purchased books, embedded in many corporate ID cards and passports, etc. The tags have a few common properties: they transmit a unique ID number, are optimized to be 'read' from predefined distances, and are usually small so they can remain unobtrusive or hidden.

RFID *readers* are used to track nearby tags by wirelessly reading a tag's unique ID (see Figure 4); a tag simply has to be brought into physical proximity with a reader to be read. Readers are mostly used for industrial or commercial purposes, e.g. asset tracking or electronic payment. Wal-mart use RFID tags and readers in their supply chain. The technology is also used in mass transit systems in cities like London (http://en.wikipedia.org/wiki/Oyster_card) and Hong Kong (http://en.wikipedia.org/wiki/Octopus_card). In Japan, many mobile phones incorporate readers to enable e-money payments in shops and vending machines.

For those of us who want to experiment with RFID, the problem is that the technology is almost always black boxed (http://en.wikipedia.org/wiki/Black_box).

That is, the inner workings of a tag and its interaction with a reader is hidden from view, and thus difficult to have much control over.

In the two exercises that follow (building a RFID reader detector (https://www.instructables.com/id/SX28QZ1FN4H8QCG/) and a tilt-sensitive RFID tag (https://www.instructables.com/id/SGW3J0SFN49WUIF/)), we offer an example of how you can start revealing some of the workings of RFID and thus gain some control over the technology. The two exercises also hopefully show that the technology is relatively simple and how it can be extended to support some interesting interactions. We offer some other possibilities that build on our examples at the end (https://www.instructables.com/id/SQKIRD4FN49ZUS1/).

Add Tip | Ask Question

Step 1: Material and Tools

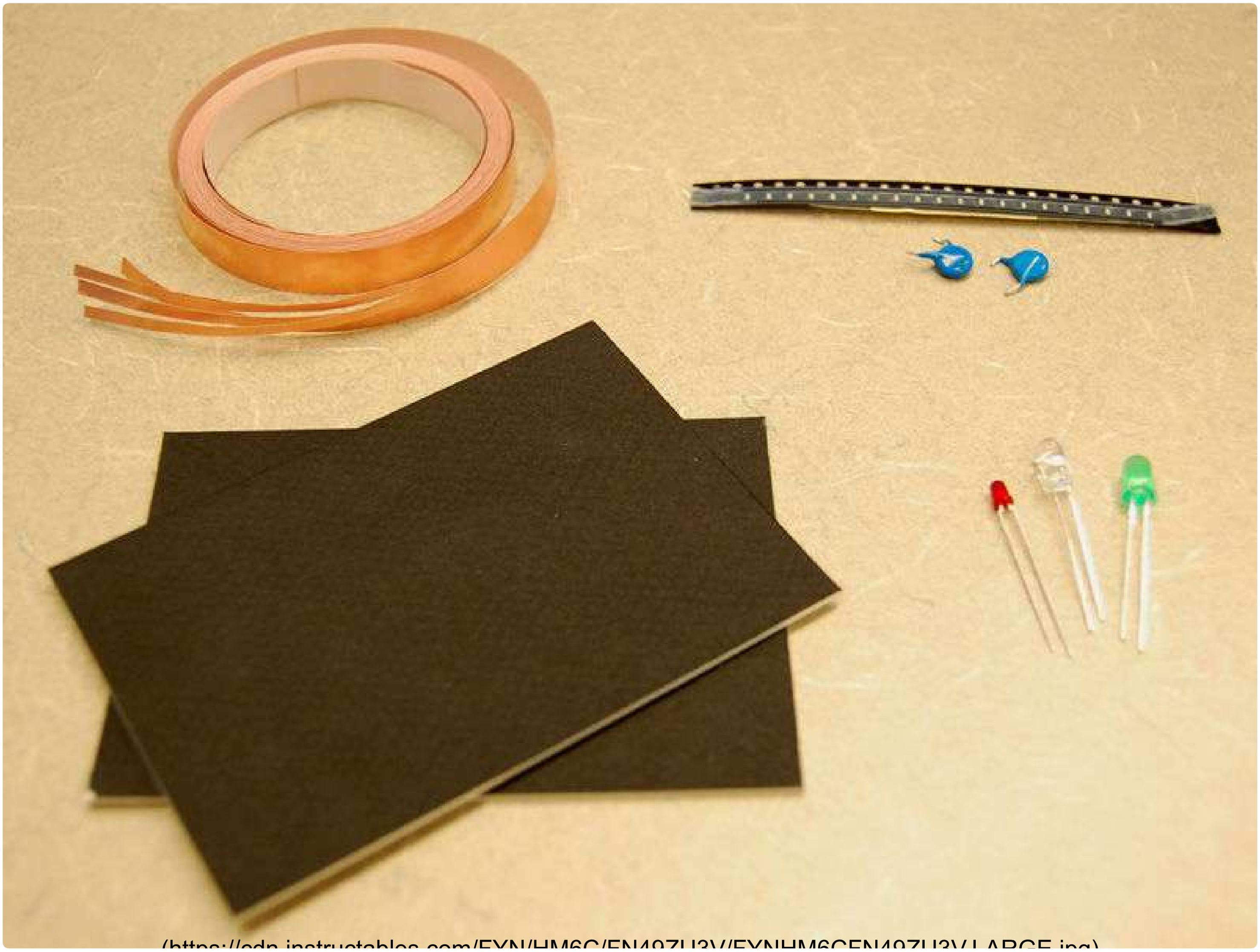

(https://cdn.instructables.com/FXN/HM6G/FN49ZU3V/FXNHM6GFN49ZU3V.LARGE.jpg)

(https://cdn.instructables.com/FIM/0KGJ/FN49ZU3U/FIM0KGJF

Weller®
WES50
Output For PES50 Iron Only. Use Only Original Weller Tips.
°Fx10
ESD SAFE
Weller

(https://cdn.instructables.com/F5C/CNJD/FN49ZU3T/F

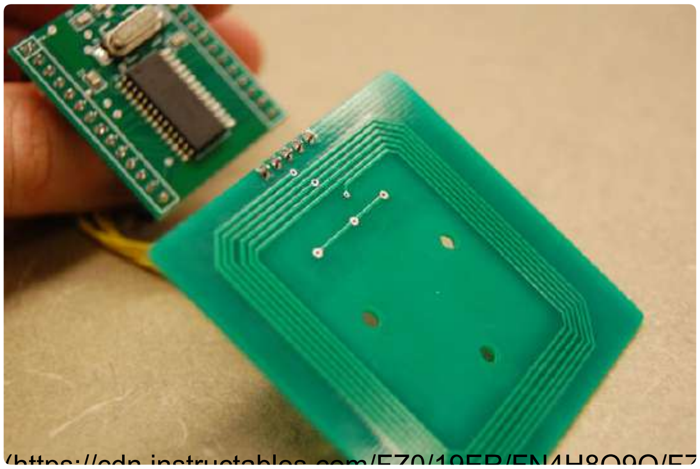

(https://cdn.instructables.com/F70/19FB/FN4H8O9O/F7

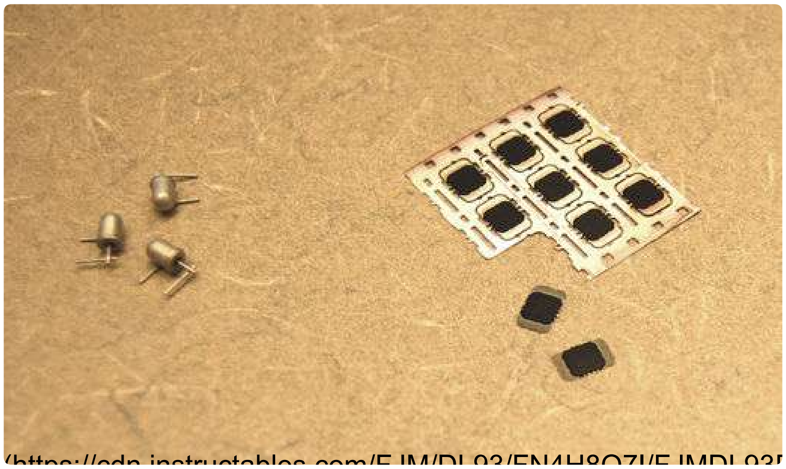

(https://cdn.instructables.com/FJM/DL93/FN4H8OZI/FJMDL93F

This section provides an overview of the necessary materials and tools.

Materials (see Figure 1):
We need the following material to built the basic RFID reader detector.
- Cardboard (around 100x70 mm)
- Conductive copper tape (e.g., order number 1218478 at www.farnell.com)
- Capacitor 82 pF (picofarad) (e.g., order number 1138852 at www.farnell.com)
- Low current LED (light-emitting diode) (e.g., order number 1003207at www.farnell.com)

Tools (see Figure 2 and 3):
- Craft knife and scissors
- Insulating tape (e.g., order number 1373979 at www.farnell.com)
- Soldering iron and solder

RFID reader for testing (see Figure 4):
To test our RFID tags we need an RFID reader that can operate at a frequency of 13.56 MHz.
There many readers for this widely used RFID standard, for instance the Sonmicro MIFARE USB reader (http://www.sonmicro.com/).
Note: The Phidget RFID reader does *not* work with the tags created in this project, as it uses a different frequency for communication with the tags (125 kHz).

Advanced material (see Figure 5):

The following material is necessary to build the second part of the project: the tilt-sensitive RFID tag.

- Micro tilt switches (e.g., www.digikey.com)
- RFID ICs (e.g., MIFARE Standard 1k, part no. 568-2219-1-ND at www.digikey.com)

Add Tip Ask Question

Step 2: Building the RFID Antenna

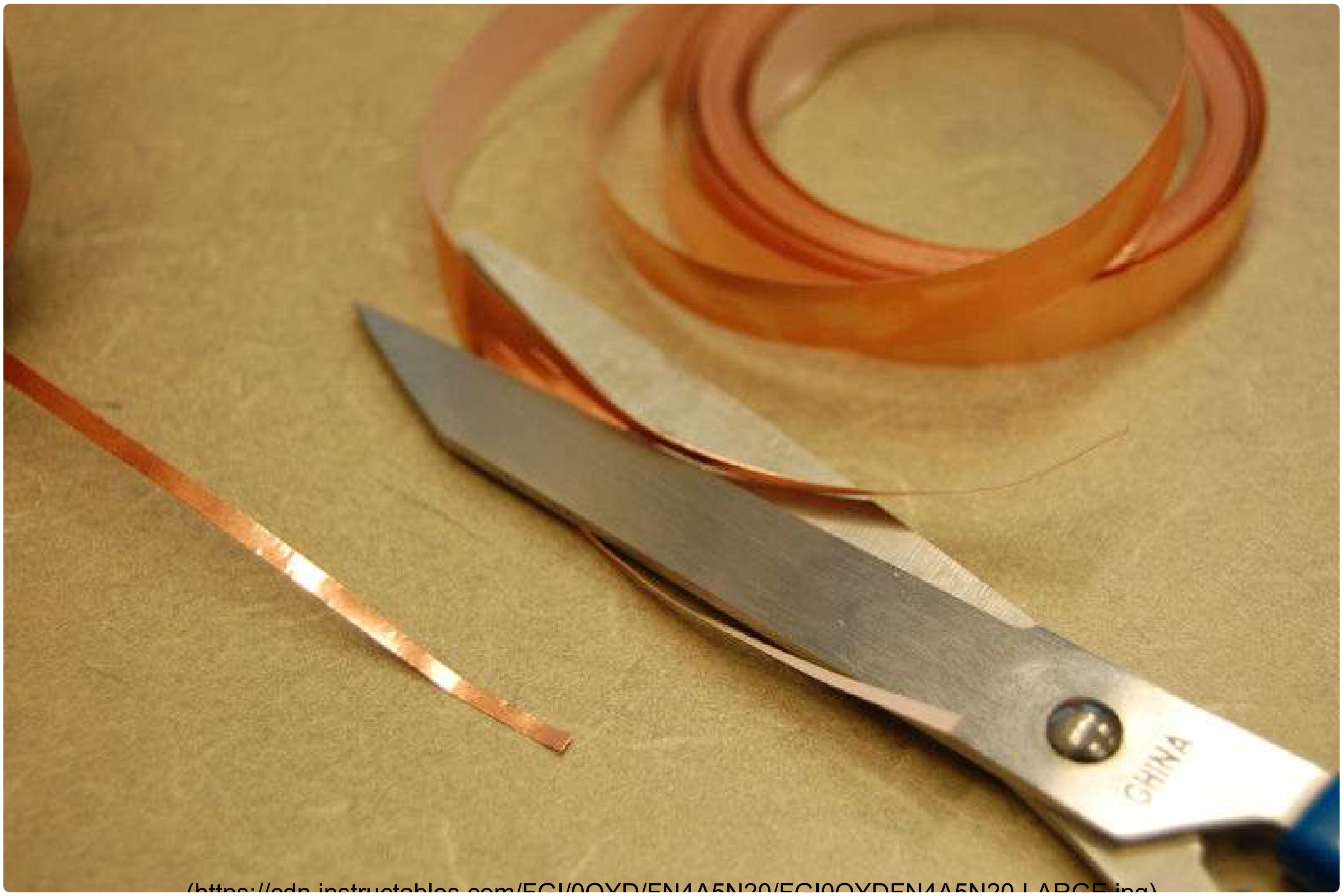

(https://cdn.instructables.com/FGI/0QYD/FN4A5N20/FGI0QYDFN4A5N20.LARGE.jpg)

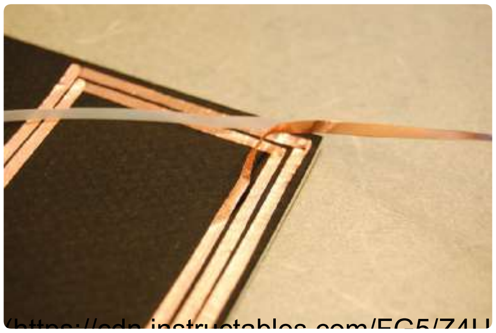
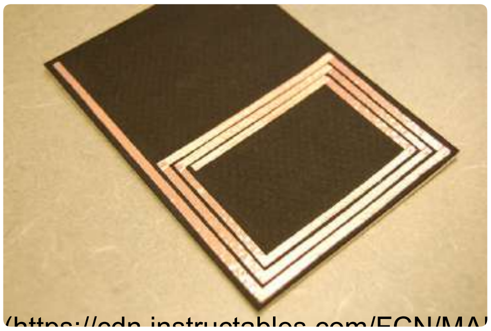
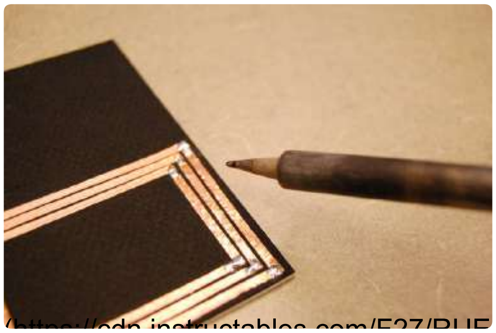

This step describes how to build the antenna for the RFID tag.

Building the RFID tag antenna
To build the tag's antenna follow these three steps.
1. Cut the conductive copper tape into thin stripes of around 2mm (see Figure 1).
2. Tape these stripes (see Figure 2) in loops around one half of the cardboard (see Figure 3 for the layout of the antenna). The tag should have between 3-4 loops for the antenna.
3. Solder all the connections between the copper tape. Sometimes, this isn't necessary as the tape's adhesive backing is conductive, but solder the connections if you want to be on the safe side.

Now we have created our RFID tag antenna, and we will add the "RFID reader detection" functionality in the following step.

A little background
RFID readers transmit an electromagnetic (EM) field with their reader antenna. This EM field induces a current in the antenna for all RFID tags within reading distance. This induced current activates the RFID chip that is connected to the tag's antenna. This chip then modulates a response (usually the unique ID number) that is transmitted back to the reader. The antenna of an RFID tag is usually a thin copper wire that is arranged in loops. The loops allow the emitted EM field of the RFID reader to induce current to the antenna of the tag.

Add Tip Ask Question

Step 3: RFID Reader Detection

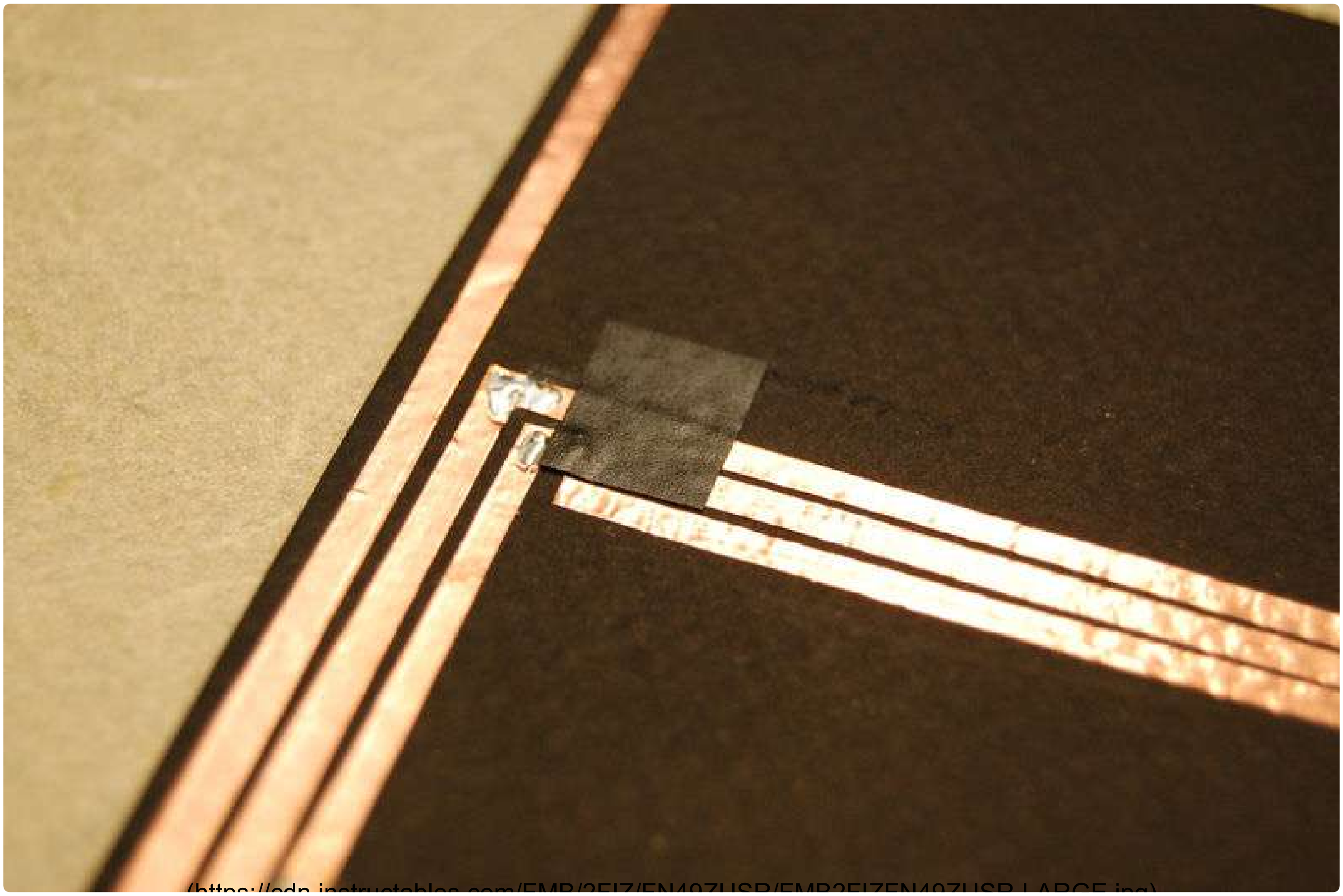

(https://cdn.instructables.com/FMB/3FIZ/FN49ZUSB/FMB3FIZFN49ZUSB.LARGE.jpg)

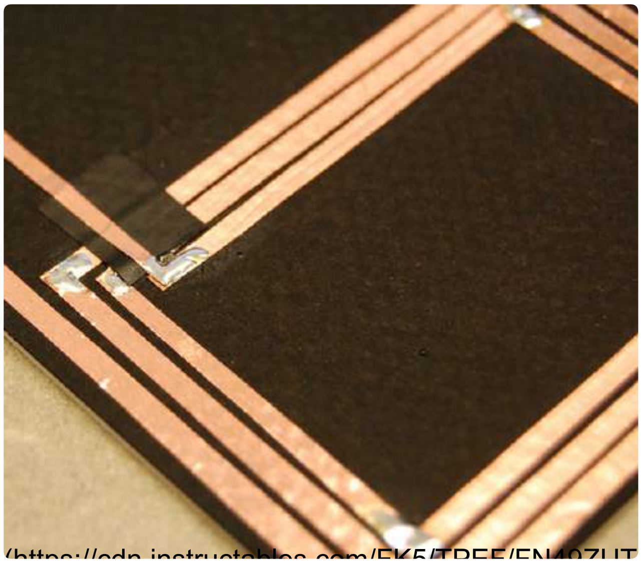

(https://cdn.instructables.com/FK5/TPEE/FN49ZUT

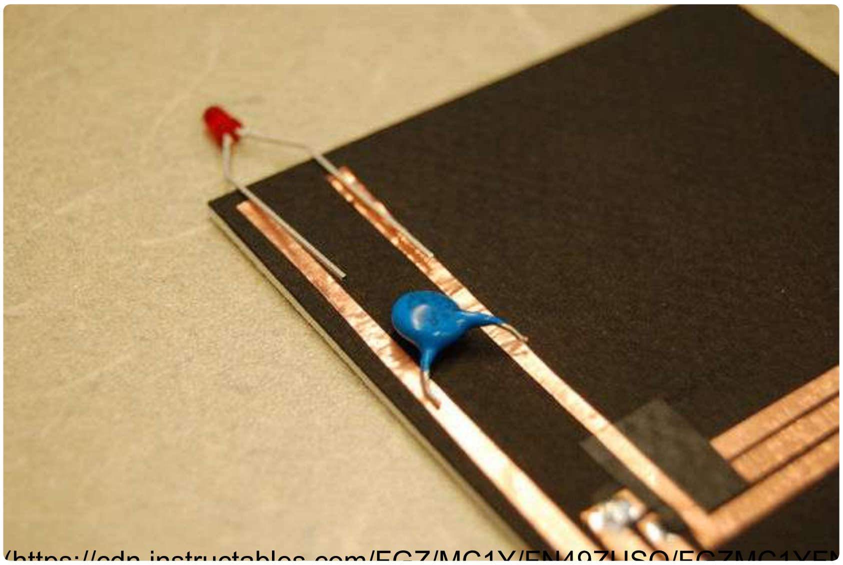

(https://cdn.instructables.com/FCZ/MG1X/FN49ZUSQ/FCZMG1XFN

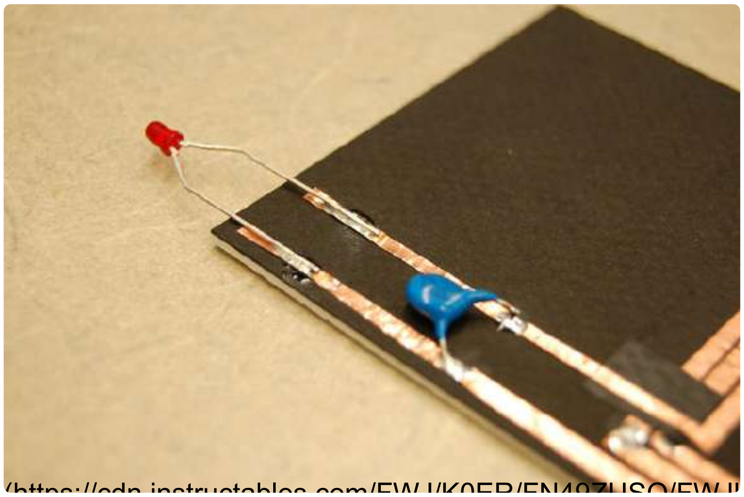

(https://cdn.instructables.com/FWJ/K0FB/FN49ZHSO/FWJ

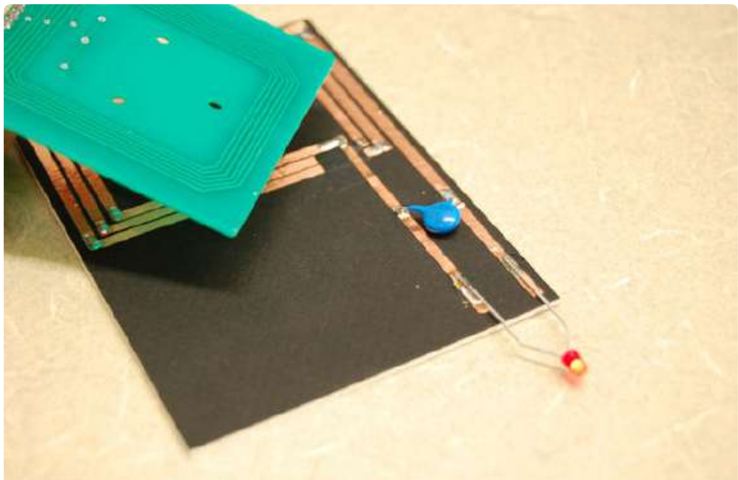

(https://cdn.instructables.com/F33/GPGW/FN49ZHSN/F33

This step describes how to add a simple mechanism to the RFID tag antenna that allows us detect nearby RFID readers.

Antenna connection

First, we add a small piece of insulation tape for the connection of the inner end of the antenna loop (as illustrated in Figure 1). This is to insulate the outer loops. Then we add another copper tape strip to the inner end of the antenna as shown in Figure 2. Here again we solder the two ends of the conductive copper tape together.

Capacitor and LED

Next, we add the capacitor (82 pF) and the low current LED to the tag as shown in Figure 3. They are connected in parallel. We also solder these two components to the copper tape (see Figure 4).

Testing

With these simple steps, our RFID reader detector is finished! By bringing our DIY RFID detector close to an RFID reader (as shown in Figure 5), the connected LED lights up. With the Sonmicro reader hardware the distance to the reader has to be below 8-10 cm; however, there are RFID readers available with a stronger EM field and therefore a higher maximum reading distance.

In the next step of the instructable we will show how to extend a basic RFID tag and make it tilt-sensitive.

Add Tip

Ask Question

Step 4: Tilt-Sensitive RFID Tag

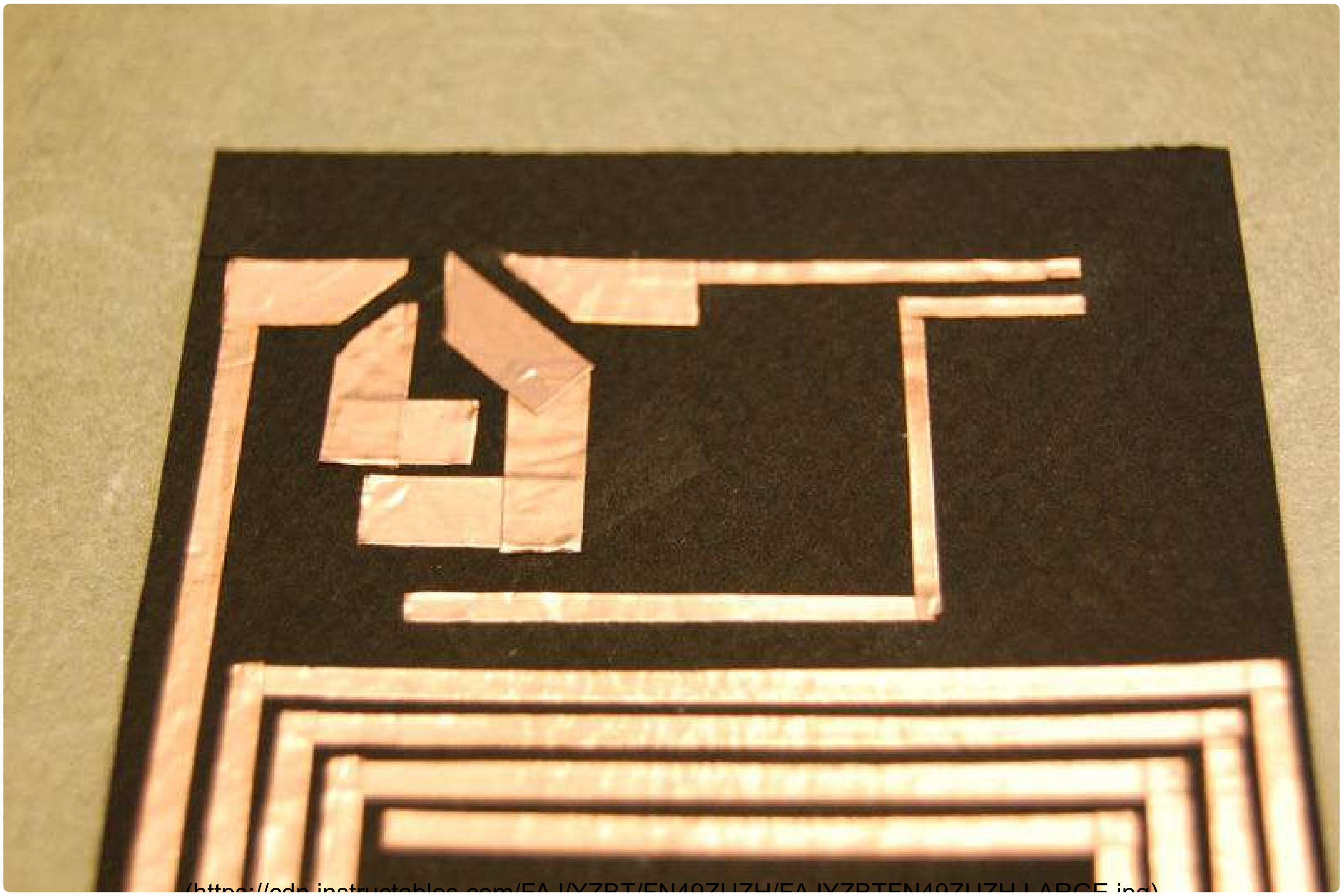

(https://cdn.instructables.com/FAJ/YZBT/FN49ZUZH/FAJYZBTFN49ZUZH.LARGE.jpg)

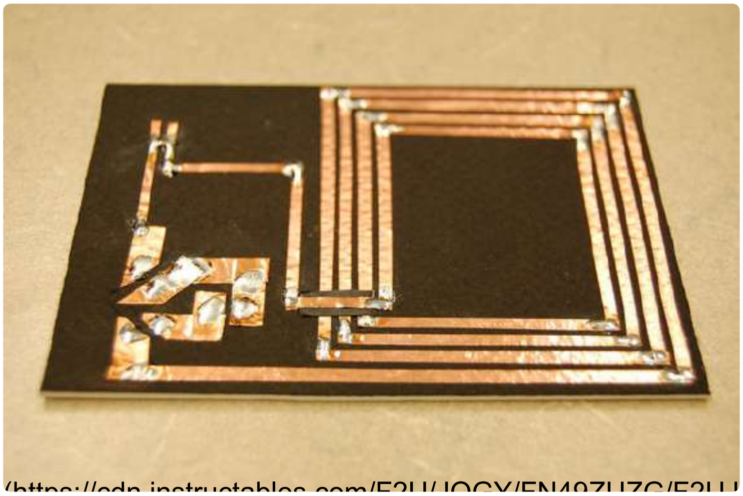

(https://cdn.instructables.com/F2U/JOGX/FN49ZUZG/F2U

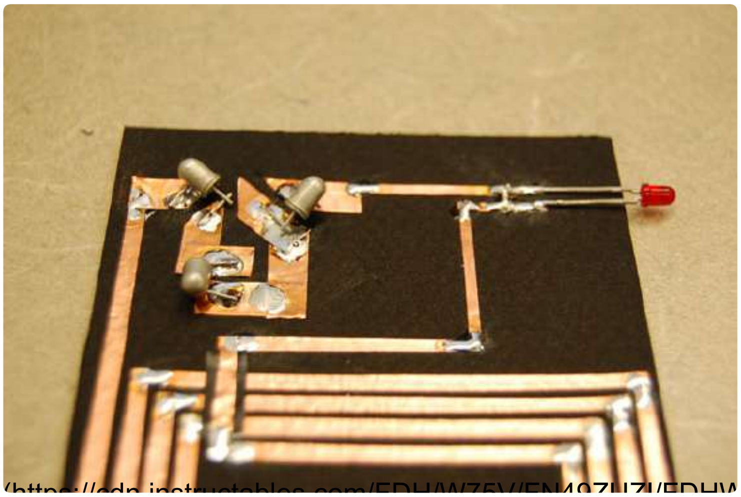

(https://cdn.instructables.com/FDH/W75V/FN49ZUZI/FDHW

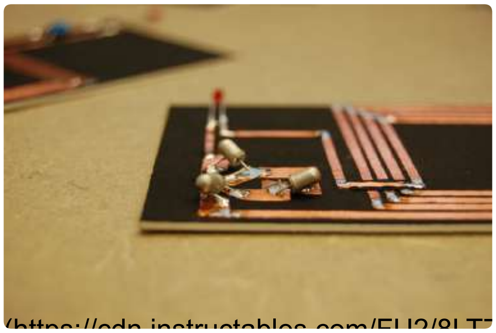

(https://cdn.instructables.com/FU2/8LT

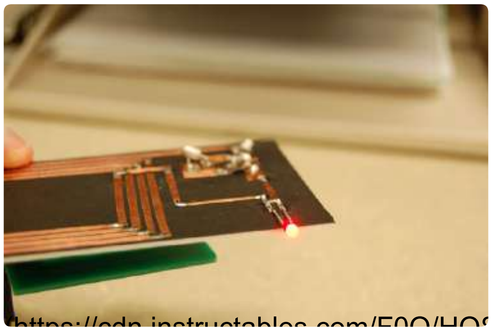

(https://cdn.instructables.com/F0Q/HO

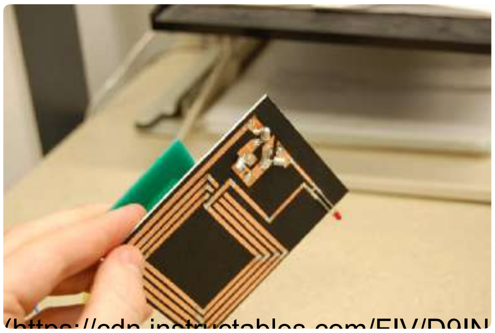

(https://cdn.instructables.com/FIV/D9IN

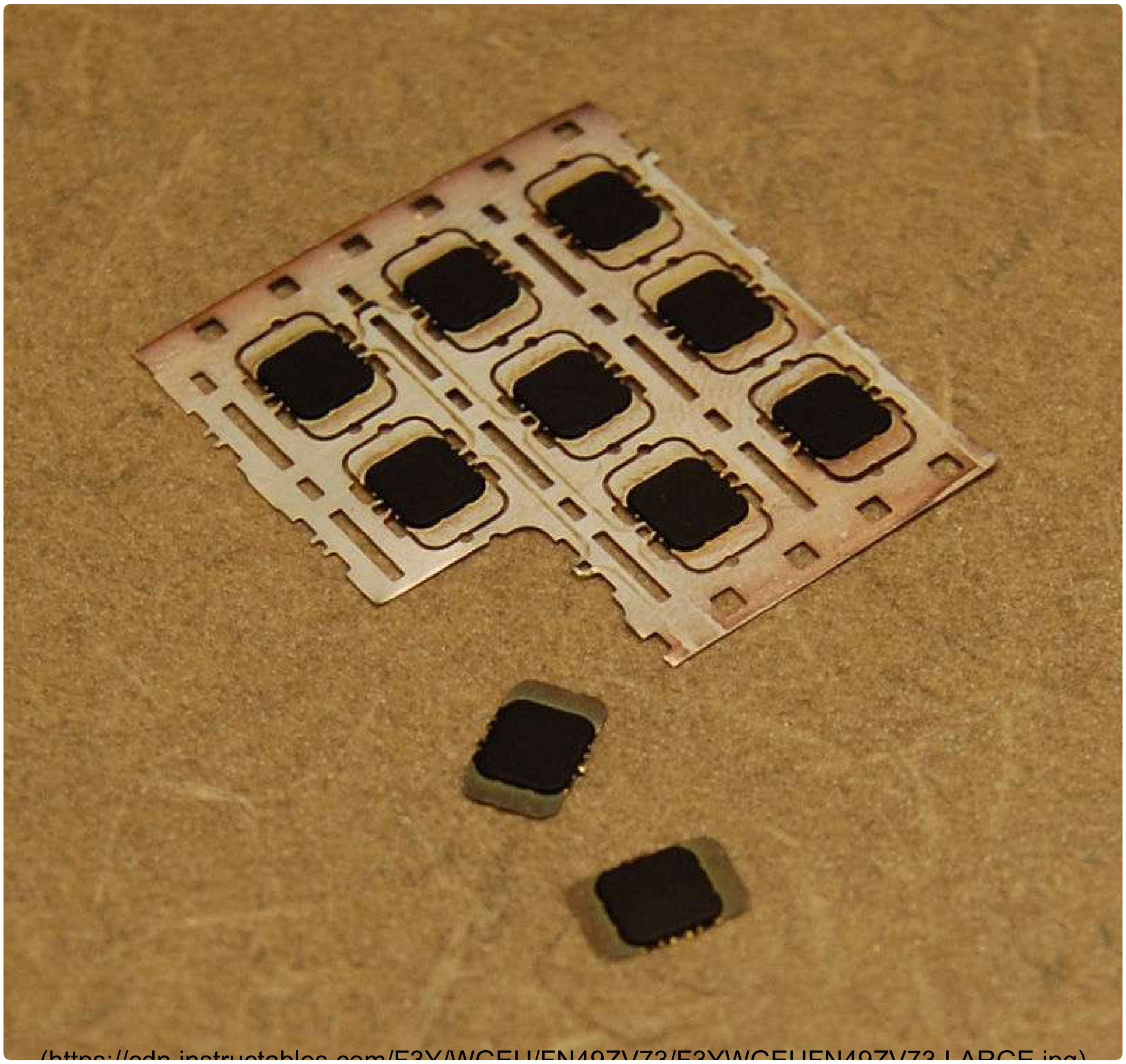

(https://cdn.instructables.com/F3Y/WGFU/FN49ZV73/F3YWGFUFN49ZV73.LARGE.jpg)

We now describe the process of how to build a tilt-sensitive RFID tag. This extends the previous exercise.

Antenna

The antenna for this second RFID tag is similar to the first antenna we built. We thus need another piece of cardboard and to repeat the steps described earlier in STEP 2 (https://www.instructables.com/id/S2YO8NWFN4H8Q8A/) of this instructable.

Tilt-sensitive tag

Next, we add additional copper tape connections to the tag, as shown in Figure 1. These connections allow us to connect three tilt switches, a capacitor, and the LED to the antenna. Again, all the connections of the copper tape are soldered together. We add the three tilt switches to the tag as shown in Figure 3. The tilt switches are soldered to the copper tape, and it is important to connect them in a slight angle (around 5-10 degrees) as shown in Figure 4. This makes sure that the silt switches are in a *closed* state while the RFID tag is in a horizontal position, and in a *open* state while the tag is in a vertical position.
Again, we also add an LED and a capacitor to the antenna as shown in Figure 3 (we use a different form factor of the capacitor here just to illustrate the alternative options).

Testing the tilt-sensitive tag

We can now use our Sonmicro RFID reader again to test our new tilt-sensitive RFID tag. The tag is activate while in a horizontal position as in Figure 5, and is inactive when in a vertical position as in Figure 6.

Using RFID chips

We can now replace the connected capacitor and LED from our tag with an RFID

MICROWAVE AND RADIO FREQUENCY ENGINEERING

INDEX

- admittance ... 4
- AM ... 20
- Ampere's law ... 8
- a_n parameter ... 15
- anisotropic ... 21
- attenuation constant ... 6
- $\mathscr{B}$ magnetic flux density ... 8
- B susceptance ... 4
- beta ... 2
- b_n parameter ... 15
- capacitance ... 4
- carrier ... 7
- CDMA ... 20
- cellular ... 20
- characteristic admittance ... 3
- characteristic impedance . 2, 3
- circulator ... 16
- communications frequencies ... 20
- complex permittivity ... 12, 21
- complex propagation constant ... 6, 10
- conductivity ... 12
- conductor loss factor ... 13
- copper cladding ... 13
- cosmic rays ... 20
- coupling factor ... 16
- $\mathscr{D}$ electric flux density ... 8
- dB ... 16
- dBm ... 18
- del ... 18
- dielectric ... 21
- dielectric constant ... 12
- dielectric loss factor ... 13
- dielectric relaxation frequency ... 8
- directional coupler ... 16
- directivity ... 16
- div ... 18
- divergence ... 18
- $\mathscr{E}$ electric field ... 8
- effective permittivity ... 13
- EHF ... 20
- electric conductivity ... 12
- electric permittivity ... 12
- electromagnetic spectrum .. 20
- ELF ... 20
- empirical ... 21
- envelope ... 7
- evanescent ... 17, 21
- excitation port ... 14, 16
- Faraday's law ... 8
- Fourier series ... 3
- frequency domain ... 8
- frequency spectrum ... 20
- gamma rays ... 20
- Gauss' law ... 8
- general math ... 18
- glossary ... 21
- grad operator ... 18
- gradient ... 18
- graphing ... 19
- group velocity ... 7
- GSM ... 20
- $\mathscr{H}$ magnetic field ... 8
- HF ... 20
- high frequency ... 9
- high frequency resistance .. 11
- hybrid ring ... 15
- hyperbolic functions ... 19
- impedance ... 6
  - intrinsic ... 10
  - waves ... 10
- incident wave amplitude ... 15
- internally matched ... 14
- intrinsic impedance ... 10
- isotropic ... 21
- *j* 18
- $\mathscr{J}$ current density ... 8
- k wave number ... 10
- k of a dielectric ... 12
- lambda ... 6
- Laplacian ... 19
- LF ... 20
- light ... 20
- linear ... 21
- loss tangent ... 9
  - complex ... 12
- lossless network ... 15
- low frequency ... 8
- magnetic permeability ... 11
- Maxwell's equations ... 8
- MF ... 20
- microstrip conductors ... 13
- mode number ... 17
- modulated wave ... 7
- nabla operator ... 18
- network theory ... 14
- normalize ... 4
- observation port ... 14, 16
- omega-beta graph ... 7
- overdamped ... 21
- parallel plate capacitance ... 4
- PCS ... 20
- permeability ... 11
- permittivity ... 12
  - complex ... 12
  - effective ... 13
  - relative ... 12
- phase constant ... 2, 6, 8, 10
- phase velocity ... 2, 10
- phasor notation ... 18
- plane waves ... 10
- polar notation ... 18
- power ... 12, 18
  - network ... 15
- propagation constant
  - complex ... 6, 10
- quarter-wave section ... 6
- quasi-static ... 9
- radar ... 20
- rat race ... 15
- reciprocity ... 15
- reflected wave amplitude .. 15
- reflection coefficient ... 3, 10
- relative permittivity ... 12
- resistance
  - high frequency ... 11
- resistivity ... 12
- scattering matrix ... 14, 16
- scattering parameter ... 14, 15, 16
- self-matched ... 14
- separation of variables ... 17
- series stub ... 5
- sheet resistance ... 11
- SHF ... 20
- shunt stub ... 5
- signs ... 2
- S_{ij} scattering parameter ... 14
- single-stub tuning ... 5
- skin depth ... 7
- SLF ... 20
- Smith chart ... 4, 5
- space derivative ... 18
- spectrum ... 20
- square root of *j* ... 18
- stripline conductor ... 13
- stub length ... 5
- susceptance ... 4
- tan δ ... 9
- Taylor series ... 19
- TE waves ... 17
- telegrapher's equations ... 2
- TEM assumptions ... 9
- TEM waves ... 9
- thermal speed ... 12
- time domain ... 8
- time of flight ... 3
- time variable ... 21
- time-harmonic ... 8
- TM waves ... 17
- transmission coefficient ... 3
- transmission lines ... 2
- transverse ... 21
- transverse electromagnetic waves ... 9
- transverse plane ... 17
- TV ... 20
- UHF ... 20
- ULF ... 20
- ultraviolet ... 20
- underdamped ... 21
- uniform plane waves ... 9
- unitary matrix ... 15
- vector differential equation 18
- velocity of propagation . 2, 10
- v_g group velocity ... 7
- VHF ... 20
- VLF ... 20
- v_p velocity of propagation .. 2
- wave analogies ... 10
- wave equation ... 2, 17
- wave impedance ... 10
- wave input impedance ... 11
- wave number ... 10, 21
- wavelength ... 6
- Wheeler's equation ... 14
- X-ray ... 20
- Y admittance ... 4
- y_0 characteristic admittance 3
- Z_0 characteristic impedance 3
- α attenuation constant ... 6
- α_c conductor loss factor ... 13
- α_d dielectric loss factor ... 13
- β phase constant ... 6
- δ loss tangent ... 9
- δ skin depth ... 7
- ε permittivity ... 12
- ε_c complex permittivity ... 12
- ε_r relative permittivity ... 12
- γ complex propagation constant ... 6
- η intrinsic wave impedance ... 10
- λ wavelength ... 6
- $\lambda/4$ ... 6
- μ permeability ... 11
- ρ reflection coefficient ... 3
- ρ_v volume charge density ... 8
- σ conductivity ... 12
- τ transmission coefficient ... 3
- ∇ del ... 18
- ∇ divergence ... 18
- ∇ gradient ... 18
- ∇^2 Laplacian ... 19
- ∇^2 Laplacian ... 17

TRANSMISSION LINES

TELEGRAPHER'S EQUATIONS

$$(1)\quad \frac{\partial V}{\partial z} = -L\frac{\partial I}{\partial t} \qquad (2)\quad \frac{\partial I}{\partial z} = -C\frac{\partial V}{\partial t}$$

By taking the partial derivative with respect to *z* of equation 1 and partial with respect to *t* of equation 2, we can get:

$$(\mathrm{i})\quad \frac{\partial^2 V}{\partial z^2} = LC\frac{\partial^2 V}{\partial t^2} \qquad (\mathrm{ii})\quad \frac{\partial^2 I}{\partial z^2} = LC\frac{\partial^2 I}{\partial t^2}$$

SOLVING THE EQUATIONS

To solve the equations (i) and (ii) above, we guess that $F(u) = F(z \pm vt)$ is a solution to the equations. It is found that the unknown constant v is the wave propagation velocity.

$$\boxed{V_{total} = V_+(z - vt) + V_-(v + vt)}$$ where:

z is the position along the transmission line, where the load is at z=0 and the source is at z=-l, with l the length of the line.

v is the **velocity of propagation** $1/\sqrt{LC}$ or ω/β, the speed at which the waveform moves down the line; see p 2

t is time

THE COMPLEX WAVE EQUATION

The general solutions of equations (i) and (ii) above yield the complex wave equations for voltage and current. These are applicable when the excitation is sinusoidal and the circuit is under steady state conditions.

$$\boxed{V(z) = V^+e^{-j\beta z} + V^-e^{+j\beta z}}$$

$$\boxed{I(z) = I^+e^{-j\beta z} + I^-e^{+j\beta z}}$$

$$\boxed{I(z) = \frac{V^+e^{-j\beta z} + V^-e^{+j\beta z}}{Z_0}}$$ where:

$e^{-j\beta z}$ and $e^{+j\beta z}$ represent wave propagation in the $+z$ and $-z$ directions respectively,

$\beta = \omega\sqrt{LC} = \omega / v$ is the **phase constant,**

$Z_0 = \sqrt{L/C}$ is the **characteristic impedance** of the line. These equations represent the voltage and current **phasors**.

+/- WATCHING SIGNS

By convention z is the variable used to describe position along a transmission line with the origin z=0 set at the load so that all other points along the line are described by **negative** position values.

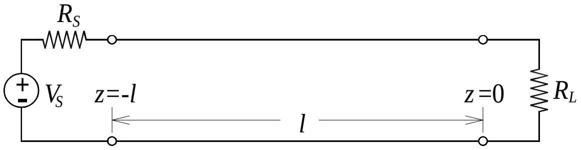

Ohm's law for right- and left-traveling disturbances:

$$V_+ = I_+Z_0 \qquad V_- = -I_-Z_0$$

v_p VELOCITY OF PROPAGATION [cm/s]

The velocity of propagation is the speed at which a wave moves down a transmission line. The velocity approaches the speed of light but may not exceed the speed of light since this is the maximum speed at which information can be transmitted. But v_p may exceed the speed of light mathematically in some calculations.

$$\boxed{v_p = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{\varepsilon\mu}} = \frac{\omega}{\beta}}$$ where:

L = inductance per unit length [H/cm]

C = capacitance per unit length [F/cm]

ε = permittivity of the material [F/cm]

μ = permeability of the material [H/cm]

ω = frequency [radians/second]

β = phase constant

Phase Velocity The velocity of propagation of a TEM wave may also be referred to as the *phase velocity*. The phase velocity of a TEM wave in conducting material may be described by:

$$\boxed{v_p = \omega\delta = \frac{\omega}{k} = c\frac{2\pi\delta}{\lambda_0} = c\frac{1}{\sqrt{\varepsilon_{r\ \mathrm{eff}}}}}$$ where:

δ = skin depth [m]

c = speed of light 2.998×10^8 m/s

λ_0 = wavelength in the material [m]

Z_0 CHARACTERISTIC IMPEDANCE [Ω]

The **characteristic impedance** is the resistance initially *seen* when a signal is applied to the line. It is a physical characteristic resulting from the materials and geometry of the line.

Lossless line: $\boxed{Z_0 \equiv \sqrt{\frac{L}{C}}} = \frac{V_+}{I_+} = -\frac{V_-}{I_-}$

Lossy line: $\boxed{Z_0 \equiv \sqrt{\frac{R+j\omega L}{G+j\omega C}} = |Z_0| e^{j\phi_z}}$

L = inductance per unit length [H/cm]
C = capacitance per unit length [F/cm]
V_+ = the forward-traveling (left to right) voltage [V]
I_+ = the forward-traveling (left to right) current [I]
V_- = the reverse-traveling (right to left) voltage [V]
I_- = the reverse-traveling (right to left) current [I]
R = the line resistance per unit length [Ω/cm]
G = the conductance per unit length [Ω^{-1}/cm]
ϕ = phase angle of the complex impedance [radians]

y_0 CHARACTERISTIC ADMITTANCE [Ω^{-1}]

The **characteristic admittance** is the reciprocal of the characteristic impedance.

$$\boxed{y_0 \equiv \sqrt{\frac{C}{L}}} = \frac{I_+}{V_+} = -\frac{I_-}{V_-}$$

r REFLECTION COEFFICIENT

The **reflection coefficient** is the ratio of reflected voltage to the forward-traveling voltage, a value ranging from –1 to +1 which, when multiplied by the wave voltage, determines the amount of voltage reflected at one end of the transmission line.

$$\boxed{\rho \equiv \frac{V_-}{V_+} = -\frac{I_-}{I_+}}$$

A reflection coefficient is present at each end of the transmission line:

$$\boxed{\rho_{\text{source}} = \frac{R_S - z_0}{R_S + z_0}} \qquad \boxed{\rho_{\text{load}} = \frac{R_L - z_0}{R_L + z_0}}$$

t TRANSMISSION COEFFICIENT

The **transmission coefficient** is the ratio of total voltage to the forward-traveling voltage, a value ranging from 0 to 2.

$$\boxed{\tau \equiv \frac{V_{total}}{V_+}} = 1 + \rho$$

TOF TIME OF FLIGHT [s]

The **time of flight** is how long it takes a signal to travel the length of the transmission line

$$\boxed{TOF \equiv \frac{l}{v}} = l\sqrt{LC} = \sqrt{L_{TOT}C_{TOT}}$$

l = length of the transmission line [cm]
v = the velocity of propagation $1/\sqrt{LC}$, the speed at which the waveform moves down the line
L = inductance per unit length [H/cm]
C = capacitance per unit length [F/cm]
L_{TOT} = total inductance [H]
C_{TOT} = total capacitance [F]

DERIVED EQUATIONS

$$V_+ = z_0 I_+ = \left(V_{TOT} + I_{TOT} z_0\right)/2$$
$$V_- = -z_0 I_- = \left(V_{TOT} - I_{TOT} z_0\right)/2$$
$$I_+ = y_0 V_+ = \left(I_{TOT} + V_{TOT} y_0\right)/2$$
$$I_- = -y_0 V_- = \left(I_{TOT} - V_{TOT} y_0\right)/2$$

C_n FOURIER SERIES

The function $x(t)$ must be periodic in order to employ the Fourier series. The following is the exponential Fourier series, which involves simpler calculations than other forms but is not as easy to visualize as the trigonometric forms.

$$\boxed{C_n = \frac{1}{T}\int_{t_1}^{t_1+T} x(t)\, e^{-jn\omega_0 t}\, dt}$$

C_n = amplitude	n = the harmonic (an integer)
T = period [s]	w_0 = frequency $2\pi/T$ [radians]
t = time [s]	

The function $x(t)$ may be delayed in time. All this does in a Fourier series is to shift the phase. If you know the C_ns for $x(t)$, then the C_ns for $x(t-\alpha)$ are just $C_n e^{-jn\omega 0\alpha}$. (Here, C_ns is just the plural of C_n.)

C CAPACITANCE [F]

$$v(t) = \frac{1}{C}\int_0^t i\,d\tau + v(0) \qquad I_{cap} = C\frac{dV_{cap}}{dt}$$

$$\boxed{v(t) = v_f + (v_0 - v_f)e^{-t/t}} \qquad i(t) = i_f + (i_0 - i_f)e^{-t/t}$$

$$P(t) = i_0^{\,2} R e^{-2t/\tau}$$

$v(t)$ = voltage across the capacitor, at time t [V]

v_f = final voltage across the capacitor, steady-state voltage [V]

v_0 = initial voltage across the capacitor [V]

t = time [s]

τ = the time constant, RC [seconds]

C = capacitance [F]

Natural log: $\ln x = b \Leftrightarrow e^b = x$

C PARALLEL PLATE CAPACITANCE

$$C = \frac{\varepsilon A}{h} \qquad C_{\text{per unit length}} = \frac{\varepsilon A}{lh} = \frac{\varepsilon wl}{lh} = \frac{\varepsilon w}{h}$$

ε = permittivity of the material [F/cm]

A = area of one of the capacitor plates [cm^2]

h = plate separation [cm]

w = plate width [cm]

l = plate length [cm]

C = capacitance [F]

CAPACITOR-TERMINATED LINE

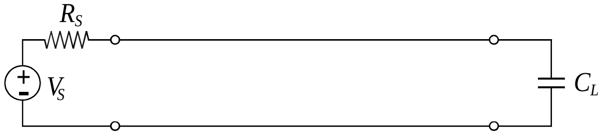

Where the incident voltage is $V_+ = V_0\left(1 - e^{-t/\tau_0}\right)$,

$$V_{cap} = V_+ + V_- = V_0\left(2 + \frac{2\tau_1}{\tau_0 - \tau_1}e^{-t/\tau_1} - \frac{2\tau_0}{\tau_0 - \tau_1}e^{-t/\tau_0}\right)$$

V_0 = final voltage across the capacitor [V]

t = time [s]

τ_0 = time constant of the incident wave, RC [s]

τ_1 = time constant effect due to the load, Z_0C_L [s]

C = capacitance [F]

SMITH CHART

First **normalize the load impedance** by dividing by the characteristic impedance, and find this point on the chart.

When working in terms of **reactance** X, an inductive load will be located on the top half of the chart, a capacitive load on the bottom half. It's the other way around when working in terms of **susceptance** B [Siemens].

Draw a straight line from the center of the chart through the normalized load impedance point to the edge of the chart.

Anchor a compass at the center of the chart and draw a circle through the normalized load impedance point. Points along this circle represent the normalized impedance at various points along the transmission line. Clockwise movement along the circle represents movement from the load toward the source with one full revolution representing 1/2 wavelength as marked on the outer circle. The two points where the circle intersects the horizontal axis are the voltage maxima (right) and the voltage minima (left).

The point opposite the impedance (180° around the circle) is the **admittance** Y [Siemens]. The reason admittance (or susceptibility) is useful is because admittances in parallel are simply added. (**Admittance** is the reciprocal of impedance; **susceptance** is the reciprocal of reactance.)

$$\boxed{\Gamma(z) = \Gamma_L e^{j2\beta z}}$$

$$e^{j2\beta z} = 1\angle 2\beta z$$

$$\mathbf{G}(z) = \frac{\mathbf{Z}(z) - 1}{\mathbf{Z}(z) + 1}$$

$$\mathbf{Z}_L = \frac{\Gamma_L - 1}{\Gamma_L + 1} \qquad \boxed{\mathbf{Z} = \frac{Z_L}{Z_0}}$$

z = distance from load [m]

$j = \sqrt{-1}$

ρ = magnitude of the reflection coefficient

β = phase constant

Γ = reflection coefficient

Z = normalized impedance [Ω]

SINGLE-STUB TUNING

The basic idea is to connect a line stub in parallel (shunt) or series a distance *d* from the load so that the imaginary part of the load impedance will be canceled.

Shunt-stub: Select *d* so that the admittance *Y* looking toward the load from a distance *d* is of the form $Y_0 + jB$. Then the stub susceptance is chosen as $-jB$, resulting in a matched condition.

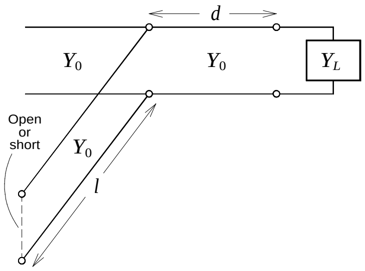
d
Y_0
Y_0
Y_L
Open or short
Y_0
l

Series-stub: Select *d* so that the admittance *Z* looking toward the load from a distance *d* is of the form $Z_0 + jX$. Then the stub susceptance is chosen as $-jX$, resulting in a matched condition.

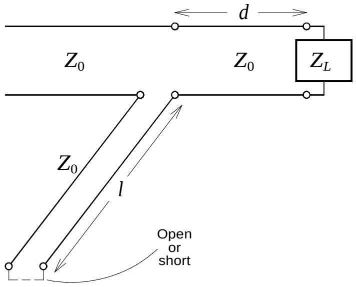
d
Z_0
Z_0
Z_L
Z_0
l
Open or short

FINDING A STUB LENGTH

Example: Find the lengths of open and shorted shunt stubs to match an admittance of 1-j0.5. The admittance of an open shunt (zero length) is $Y=0$; this point is located at the left end of the Smith Chart *x*-axis. We proceed clockwise around the Smith chart, i.e. away from the end of the stub, to the +j0.5 arc (the value needed to match –j0.5). The difference in the starting point and the end point on the wavelength scale is the length of the stub in wavelengths. The length of a shorted-type stub is found in the same manner but with the starting point at $Y=\infty$.

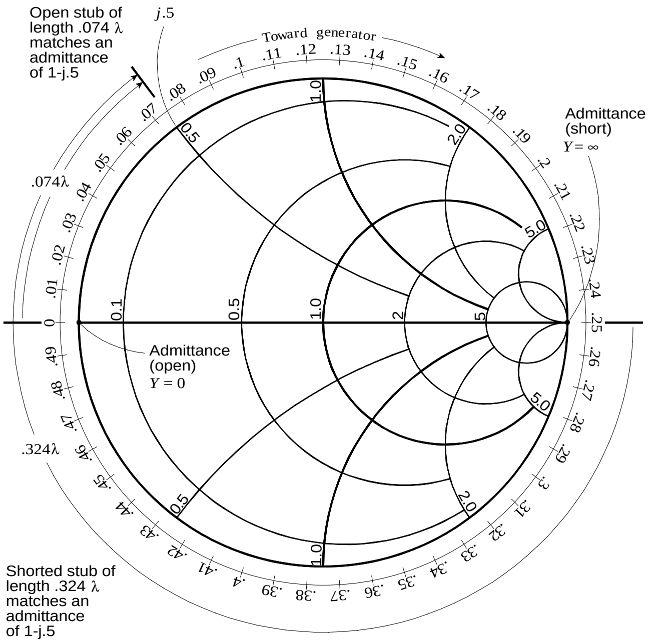
Open stub of length .074 λ matches an admittance of 1-j.5

Shorted stub of length .324 λ matches an admittance of 1-j.5

j.5
Toward generator
Admittance (short) Y = ∞
Admittance (open) Y = 0
.074λ
.324λ
0
.01
.02
.03
.04
.05
.06
.07
.08
.09
.1
.11
.12
.13
.14
.15
.16
.17
.18
.19
.2
.21
.22
.23
.24
.25
.26
.27
.28
.29
.3
.31
.32
.33
.34
.35
.36
.37
.38
.39
.4
.41
.42
.43
.44
.45
.46
.47
.48
.49
0.1
0.5
1.0
2
5
0.5
1.0
2.0
5.0
0.5
1.0
2.0
5.0

In this example, all values were in units of admittance. If we were interested in finding a stub length for a series stub problem, the units would be in impedance. The problem would be worked in exactly the same way. Of course in impedance, an open shunt (zero length) would have the value $Z=\infty$, representing a point at the right end of the *x*-axis.

LINE IMPEDANCE [Ω]

The impedance seen at the source end of a lossless transmission line:

$$Z_{in} = Z_0 \frac{1+\rho}{1-\rho} = Z_0 \frac{Z_L + jZ_0 \tan(\beta l)}{Z_0 + jZ_L \tan(\beta l)}$$

For a lossy transmission line:

$$Z_{in} = Z_0 \frac{Z_L + Z_0 \tanh(\gamma l)}{Z_0 + Z_L \tanh(\gamma l)}$$

Line impedance is periodic with spatial period λ/2.

$Z_0 = \sqrt{L/C}$, the characteristic impedance of the line. [Ω]

ρ = the reflection coefficient

Z_L = the load impedance [Ω]

β = 2π/λ, phase constant

γ = α+*j*β, complex propagation constant

l WAVELENGTH [cm]

The physical distance that a traveling wave moves during one period of its periodic cycle.

$$\boxed{\lambda = \frac{2\pi}{\beta} = \frac{2\pi}{k} = \frac{v_p}{f}}$$

$\beta = \omega\sqrt{LC}$ = 2π/λ, phase constant

$k = \omega\sqrt{\mu\varepsilon}$ = 2π/λ, wave number

v_p = velocity of propagation [m/s] see p 2.

f = frequency [Hz]

l/4 QUARTER-WAVE SECTION

A quarter-wave section of transmission line has the effect of inverting the normalized impedance of the load.

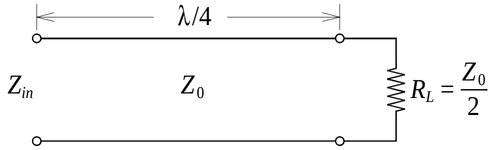

To find Z_{in}, we can normalize the load (by dividing by the characteristic impedance), invert the result, and "unnormalize" this value by multiplying by the characteristic impedance.

In this case, the normalized load is $\frac{Z_0}{2} \div Z_0 = \frac{1}{2}$

so the normalized input impedance is $\left(\frac{1}{2}\right)^{-1} = 2$

and the actual input impedance is $Z_{in} = 2Z_0$

g COMPLEX PROPAGATION CONSTANT

The propagation constant for lossy lines, taking into account the resistance along the line as well as the resistive path between the conductors.

$$\gamma = \alpha + j\beta = \sqrt{ZY} = \sqrt{(R + j\omega L)(G + j\omega C)}$$

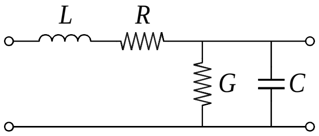

$\alpha = \sqrt{RG}$ **attenuation constant**, the real part of the complex propagation constant, describes the loss

β = 2π/λ, **phase constant**, the complex part of the complex propagation constant

Z = **series impedance** (complex, inductive) per unit length [Ω/cm]

Y = **shunt admittance** (complex, capacitive) per unit length [Ω^{-1}/cm]

R = the resistance per unit length along the transmission line [Ω/cm]

G = the conductance between conductors per unit length [Ω^{-1}/cm]

L = inductance per unit length [H/cm]

C = capacitance per unit length [F/cm]

MODULATED WAVE

Suppose we have a disturbance composed of two frequencies:

$$\sin\left[\left(\omega_0-\delta\omega\right)t-\left(\beta_0-\delta\beta\right)z\right]$$

and $$\sin\left[\left(\omega_0+\delta\omega\right)t-\left(\beta_0+\delta\beta\right)z\right]$$

where ω_0 is the average frequency and β_0 is the average phase.

Using the identity $2\cos\left(\frac{A-B}{2}\right)\sin\left(\frac{A+B}{2}\right)=\sin A+\sin B$

The combination (sum) of these two waves is

$$\boxed{2\underbrace{\cos\left(\delta\omega t-\delta\beta z\right)}_{\text{envelope}}\underbrace{\sin\left(\omega_0 t-\beta_0 z\right)}_{\text{carrier}}}$$

The envelope moves at the group velocity, see p 7.

δ = "the difference in"…

ω_0 = carrier frequency [radians/second]

ω = modulating frequency [radians/second]

β_0 = carrier frequency phase constant

β = phase constant

So the sum of two waves will be a modulated wave having a **carrier** frequency equal to the average frequency of the two waves, and an **envelope** with a frequency equal to half the difference between the two original wave frequencies.

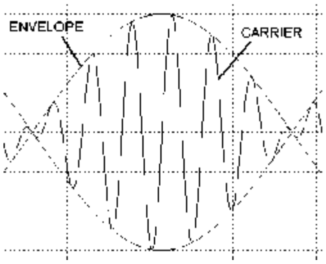

v_g GROUP VELOCITY [cm/s]

The velocity at which the envelope of a modulated wave moves.

$$\boxed{v_g=\frac{\delta\omega}{\delta\beta}=\frac{1}{\sqrt{LC_P}}\sqrt{1-\frac{{\omega_c}^2}{\omega^2}}}$$ where:

L = inductance per unit length [H/cm]

C_P = capacitance per unit length [F/cm]

ε = permittivity of the material [F/cm]

μ = permeability of the material, dielectric constant [H/cm]

ω_c = carrier frequency [radians/second]

ω = modulating frequency [radians/second]

β = phase constant

Also, since β may be given as a function of ω, remember

$$v_g=\left(\frac{d\beta}{d\omega}\right)^{-1}$$

OMEGA - BETA GRAPH

This representation is commonly used for modulated waves.

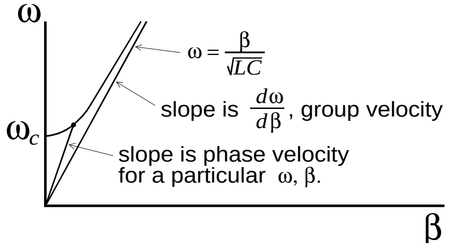
ω
$\omega=\frac{\beta}{\sqrt{LC}}$
slope is $\frac{d\omega}{d\beta}$, group velocity
ω_c
slope is phase velocity
for a particular ω, β.
β

d SKIN DEPTH [cm]

The depth into a material at which a wave is attenuated by 1/e (about 36.8%) of its original intensity. This isn't the same δ that appears in the *loss tangent*, tan δ.

$$\boxed{\delta=\frac{1}{\alpha}=\sqrt{\frac{2}{\omega\mu\sigma}}}$$ where:

$\alpha=\sqrt{RG}$ **attenuation constant**, the real part of the complex propagation constant, describes loss

μ = permeability of the material, dielectric constant [H/cm]

ω = frequency [radians/second]

σ = (*sigma*) conductivity [Siemens/meter] see p12.

Skin Depths of Selected Materials

	60 Hz	1 MHz	1 GHz
silver	8.27 mm	0.064 mm	0.0020 mm
copper	8.53 mm	0.066 mm	0.0021 mm
gold	10.14 mm	0.079 mm	0.0025 mm
aluminum	10.92 mm	0.084 mm	0.0027 mm
iron	0.65 mm	0.005 mm	0.00016 mm

MAXWELL'S EQUATIONS

Maxwell's equations govern the principles of guiding and propagation of electromagnetic energy and provide the foundations of all electromagnetic phenomena and their applications. The time-harmonic expressions can be used only when the wave is sinusoidal.

	STANDARD FORM (Time Domain)	**TIME-HARMONIC** (Frequency Domain)
Faraday's Law	$\nabla\times\vec{\mathscr{E}} = -\dfrac{\partial\vec{\mathscr{B}}}{\partial t}$	$\nabla\times\vec{E} = -j\omega\vec{B}$
Ampere's Law*	$\nabla\times\vec{\mathscr{H}} = \vec{\mathscr{J}} + \dfrac{\partial\vec{\mathscr{D}}}{\partial t}$	$\nabla\times\vec{H} = j\omega\vec{D} + \vec{J}$
Gauss' Law	$\nabla\cdot\vec{\mathscr{D}} = \rho_v$	$\nabla\cdot\vec{D} = \rho_v$
no name law	$\nabla\cdot\vec{\mathscr{B}} = 0$	$\nabla\cdot\vec{B} = 0$

$\mathscr{E}$ = electric field [V/m]
$\mathscr{B}$ = magnetic flux density [W/m^2 or T] $\mathscr{B} = \mu_0\mathscr{H}$
t = time [s]
$\mathscr{D}$ = electric flux density [C/m^2] $\mathscr{D} = \varepsilon_0\mathscr{E}$
ρ = volume charge density [C/m^3]
$\mathscr{H}$ = magnetic field intensity [A/m]
$\mathscr{J}$ = current density [A/m^2]

*Maxwell added the $\dfrac{\partial\vec{\mathscr{D}}}{\partial t}$ term to Ampere's Law.

ELECTROMAGNETIC WAVES

MODELING MAXWELL'S EQUATIONS

This is a model of a wave, analogous to a transmission line model.

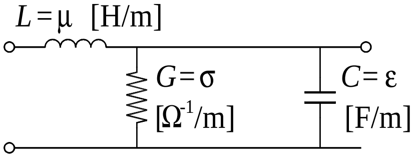

L = inductance per unit length [H/cm]
μ = permeability of the material, dielectric constant [H/cm]
G = the conductance per unit length [Ω^{-1}/cm]
σ = (*sigma*) conductivity [Siemens/meter]
C = capacitance per unit length [F/cm]
ε = permittivity of the material [F/cm]

propagation constant: $\gamma = \sqrt{(j\omega\mu)(j\omega\varepsilon+\sigma)}$

LOW FREQUENCY

At low frequencies, more materials behave as **conductors**. A wave is considered low frequency when

$\boxed{\omega \ll \dfrac{\sigma}{\varepsilon}}$ $\dfrac{\sigma}{\varepsilon}$ is the dielectric relaxation frequency

$\boxed{\eta = \dfrac{1}{\sigma\delta}(1+j)}$ intrinsic wave impedance, see p 12.

What happens to the complex propagation constant at low frequency? From the wave model above, gamma is

$$\gamma = \sqrt{(j\omega\mu)(j\omega\varepsilon+\sigma)} = \sqrt{j\omega\mu\sigma}\sqrt{1+\frac{j\omega\varepsilon}{\sigma}}$$

Since both ω and ε/σ are small

$$\gamma = \sqrt{j\omega\mu\sigma}\left(1+\frac{1}{2}j\omega\frac{\varepsilon}{\sigma}\right) = \sqrt{j\omega\mu\sigma}\,(1)$$

Since $\sqrt{j} = \frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}}$

$$\gamma = \sqrt{\omega\mu\sigma}\left(\frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}}\right) = \sqrt{\frac{\omega\mu\sigma}{2}} + j\sqrt{\frac{\omega\mu\sigma}{2}}$$

So that, with $\gamma = \alpha + j\beta$

we get $\boxed{\alpha = \sqrt{\dfrac{\omega\mu\sigma}{2}}}$, $\boxed{\beta = \sqrt{\dfrac{\omega\mu\sigma}{2}}}$ or $\boxed{\gamma = \dfrac{1}{\delta}(1+j)}$

α = **attenuation constant**, the real part of the complex propagation constant, describes the loss
β = **phase constant**, the complex part of the complex propagation constant
σ = (*sigma*) conductivity [Siemens/cm]
δ = skin depth [cm]

So the wave is attenuating at the same rate that it is propagating.

HIGH FREQUENCY

At high frequencies, more materials behave as **dielectrics**, i.e. copper is a dielectric in the gamma ray range. A wave is considered high frequency when

$\boxed{\omega \gg \frac{\sigma}{\varepsilon}}$ $\frac{\sigma}{\varepsilon}$ is the dielectric relaxation frequency

$\boxed{\eta = \sqrt{\frac{\mu}{\varepsilon}}}$ intrinsic wave impedance, see p 12.

What happens to the complex propagation constant at high frequency?

$$\gamma = \sqrt{(j\omega\mu)(j\omega\varepsilon + \sigma)} = \sqrt{j\omega\mu\, j\omega\varepsilon\left(1 + \frac{\sigma}{j\omega\varepsilon}\right)}$$

Since both $1/\omega$ and σ/ε are small

$$\gamma = j\omega\sqrt{\mu\varepsilon}\left(1 + \frac{1}{2}\frac{\sigma}{j\omega\varepsilon}\right) \qquad \boxed{\gamma = \frac{\sigma}{2}\sqrt{\frac{\mu}{\varepsilon}} + j\omega\sqrt{\mu\varepsilon}}$$

With $\gamma = \alpha + j\beta$

we get $\boxed{\alpha = \frac{\sigma}{2}\sqrt{\frac{\mu}{\varepsilon}}}$, $\boxed{\beta = \omega\sqrt{\mu\varepsilon}}$,

tan d LOSS TANGENT

The loss tangent, a value between 0 and 1, is the loss coefficient of a wave after it has traveled one wavelength. This is the way data is usually presented in texts. This is not the same δ that is used for skin depth.

$$\boxed{\tan\delta = \frac{\sigma}{\omega\varepsilon}}$$

Graphical representation of loss tangent:

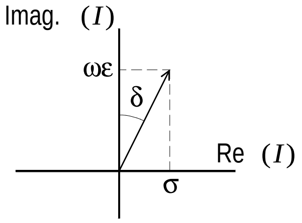

For a dielectric, $\tan\delta \ll 1$.

$$\alpha \approx \frac{1}{2}(\tan\delta)\beta = \frac{\pi}{\lambda}\tan\delta$$

$\omega\varepsilon$ is proportional to the amount of current going through the capacitance C.

σ is proportional to the amount current going through the conductance G.

TEM WAVES

Transverse Electromagnetic Waves

Electromagnetic waves that have single, orthogonal vector electric and magnetic field components (e.g., $\mathscr{E}_x$ and $\mathscr{H}_y$), both varying with a single coordinate of space (e.g., z), are known as *uniform plane waves* or *transverse electromagnetic (TEM) waves*. TEM calculations may be made using formulas from electrostatics; this is referred to as *quasi-static* solution.

Characteristics of TEM Waves

- The velocity of propagation (always in the *z* direction) is $v_p = 1/\sqrt{\mu\varepsilon}$, which is the speed of light in the material
- There is no electric or magnetic field in the direction of propagation. Since this means there is no voltage drop in the direction of propagation, it suggests that no current flows in that direction.
- The electric field is normal to the magnetic field
- The value of the electric field is η times that of the magnetic field at any instant.
- The direction of propagation is given by the direction of $\mathbf{E}\times\mathbf{H}$.
- The energy stored in the electric field per unit volume at any instant and any point is equal to the energy stored in the magnetic field.

TEM ASSUMPTIONS

Some assumptions are made for TEM waves.

$\mathscr{E}_z = 0$ $\mathscr{H}_z = 0$

$\sigma = 0$ time dependence $e^{j\omega t}$

WAVE ANALOGIES

Plane waves have many characteristics analogous to transmission line problems.

Transmission Lines	Plane Waves
Phase constant $\beta = \omega\sqrt{LC} = \dfrac{\omega}{v_p} = \dfrac{2\pi}{\lambda}$	Wave number $k = \omega\sqrt{\mu\varepsilon} = \dfrac{\omega}{v_p} = \dfrac{2\pi}{\lambda}$
Complex propagation const. $\gamma = \alpha + j\beta$ $= \sqrt{(R + j\omega L)(G + j\omega C)}$	Complex propagation constant $\gamma = \sqrt{(j\omega\mu)(j\omega\varepsilon + \sigma)}$
Velocity of propagation $v_p = \dfrac{1}{\sqrt{LC}} = \dfrac{\omega}{\beta}$	Phase velocity $v_p = \dfrac{1}{\sqrt{\mu\varepsilon}} = \dfrac{\omega}{k} = \omega\delta = c\dfrac{2\pi\delta}{\lambda}$
Characteristic impedance $Z_0 = \sqrt{\dfrac{L}{C}} = \dfrac{V_+}{I_+}$	Intrinsic impedance $\eta = \sqrt{\dfrac{\mu}{\varepsilon}} = \dfrac{E_{x+}}{H_{y+}}$
Voltage $V(z) = V_+ e^{-j\beta z} + V_- e^{j\beta z}$	Electric Field $E_x(z) = E_+ e^{-jkz} + E_- e^{jkz}$
Current $I(z) = \dfrac{1}{Z_0}\left[V_+ e^{-j\beta z} - V_- e^{j\beta z}\right]$	Magnetic Field $H_y(z) = \dfrac{1}{\eta}\left[E_+ e^{-jkz} - E_- e^{jkz}\right]$
Line input impedance $Z_{in} = Z_0 \dfrac{Z_L + jZ_0 \tan(\beta l)}{Z_0 + jZ_L \tan(\beta l)}$ $Z_{in} = Z_0 \dfrac{Z_L + Z_0 \tanh(\gamma l)}{Z_0 + Z_L \tanh(\gamma l)}$	Wave input impedance $\eta_{in} = \eta_0 \dfrac{\eta_L + j\eta_0 \tan(kl)}{\eta_0 + j\eta_L \tan(kl)}$ $\eta_{in} = \eta_0 \dfrac{\eta_L + \eta_0 \tanh(\gamma l)}{\eta_0 + \eta_L \tanh(\gamma l)}$
Reflection coefficient $\rho = \dfrac{Z_L - Z_0}{Z_L + Z_0}$	Reflection coefficient $\rho = \dfrac{\eta_L - \eta_0}{\eta_L + \eta_0}$

k WAVE NUMBER [rad./cm]

The phase constant for the uniform plane wave; the change in phase per unit length. It can be considered a constant for the medium at a particular frequency.

$$\boxed{k = \frac{\omega}{v} = \omega\sqrt{\mu\varepsilon} = \frac{2\pi}{\lambda}}$$

k appears in the phasor forms of the uniform plane wave

$E_x(z) = E_1 e^{-jkz} + E_2 e^{jkz}$, etc.

k has also been used as in the "*k* of a dielectric" meaning ε_r.

h (eta) INTRINSIC WAVE IMPEDANCE [Ω]

The ratio of electric to magnetic field components. Can be considered a constant of the medium. For free space, η = 376.73Ω. The units of η are in ohms.

$$\eta = \frac{E_{x+}}{H_{y+}} = -\frac{E_{y+}}{H_{x+}} \qquad -\eta = \frac{E_{x-}}{H_{y-}} = -\frac{E_{y-}}{H_{x-}}$$

at low frequencies	at high frequencies
$\eta = \dfrac{1}{\sigma\delta}(1 + j)$	$\eta = \sqrt{\dfrac{\mu}{\varepsilon}}$

When an electromagnetic wave encounters a sheet of conductive material it sees an impedance. K is the direction of the wave, **H** is the magnetic component and **E** is the electrical field. **E** × **H** gives the direction of propagation K.

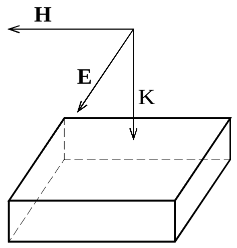

SHEET RESISTANCE [Ω]

Consider a block of material with conductivity σ.

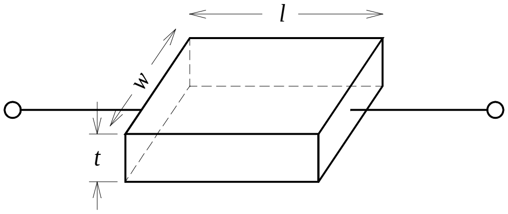

It's resistance is $R = \dfrac{l}{wt\sigma}$ Ω.

If the length is equal to the width, this reduces to

$$R = \frac{1}{t\sigma} \ \Omega.$$

And this is sheet resistance.

HIGH FREQUENCY RESISTANCE [Ω]

When a conductor carries current at high frequency, the electric field penetrates the outer surface only about 1 skin depth so that current travels near the surface of the conductor. Since the entire cross-section is not utilized, this affects the resistance of the conductor.

Cross-section of a conductor showing current flow near the surface:

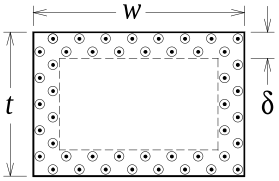

$$\boxed{R \approx \frac{1}{\sigma\delta\left(\text{perimeter}\right)} = \sqrt{\frac{\omega\mu_0}{2\sigma}}\,\frac{1}{2w+2t}}$$

σ = (*sigma*) conductivity (5.8×10^5 S/cm for copper) [Siemens/meter]
ω = frequency [radians/second]
δ = skin depth [cm]
μ_0 = permeability of free space $\mu_0 = 4\pi\times10^{-9}$ [H/cm]
w = width of the conductor [cm]
t = thickness of the conductor [cm]

$h_{\mathbf{in}}$ WAVE INPUT IMPEDANCE [Ω]

The impedance seen by a wave in a medium.

For example, the impedance of a metal sheet in a vacuum:

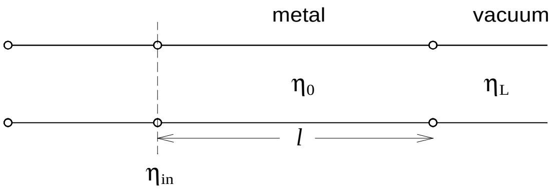

Note that a transmission line model is used here because it is analogous to a wave traveling in a medium. The "load" is the element most remote in the direction of propagation.

The input impedance is $\eta_{in} = \eta_0 \dfrac{\eta_L + \eta_0 \tanh(\gamma l)}{\eta_0 + \eta_L \tanh(\gamma l)}$ Ω.

In this example, *l* is the thickness of a metal sheet. If the metal thickness is much greater than the skin depth, then

$$\tanh(\lambda l) = \tanh\left[\frac{1}{\delta}(1+j)l\right] = \tanh\left[(\text{big number})(1+j)\right] \approx 1$$

If *l* is much less than the skin depth δ, then

$$\tanh(\lambda l) = \tanh\left[\frac{1}{\delta}(1+j)l\right] = \tanh\left[(\text{small number})(1+j)\right]$$

$$= (\text{same small number})(1+j) = \frac{l}{\delta}(1+j)$$

m MAGNETIC PERMEABILITY [H/m]

The relative increase or decrease in the resultant magnetic field inside a material compared with the magnetizing field in which the given material is located. The product of the permeability constant and the relative permeability of the material.

$\mu = \mu_0\mu_r$ where $\mu_0 = 4\pi\times10^{-7}$ H/m

Relative Permeabilities of Selected Materials

Air	1.00000037	Mercury	0.999968
Aluminum	1.000021	Nickel	600
Copper	0.9999833	Oxygen	1.000002
Gold	0.99996	Platinum	1.0003
Iron (99.96% pure)	280,000	Silver	0.9999736
Iron (motor grade)	5000	Titanium	1.00018
Lead	0.9999831	Tungsten	1.00008
Manganese	1.001	Water	0.9999912

e ELECTRIC PERMITTIVITY [F/m]

The property of a dielectric material that determines how much electrostatic energy can be stored per unit of volume when unit voltage is applied, also called the ***dielectric constant***. The product of the constant of permittivity and the relative permittivity of a material.

$$\varepsilon = \varepsilon_0 \varepsilon_r \quad \text{where } \varepsilon_0 = 8.85\times10^{-14} \text{ F/cm}$$

e$_c$ COMPLEX PERMITTIVITY

$$\varepsilon_c = \varepsilon' - j\varepsilon'' \quad \text{where } \frac{\varepsilon''}{\varepsilon'} = \tan\delta_c$$

In general, both ε' and ε'' depend on frequency in complicated ways. ε' will typically have a constant maximum value at low frequencies, tapering off at higher frequencies with several peaks along the way. ε'' will typically have a peak at the frequency at which ε' begins to decline in magnitude as well as at frequencies where ε' has peaks, and will be zero at low frequencies and between peaks.

e$_r$ RELATIVE PERMITTIVITY

The permittivity of a material is the relative permittivity multiplied by the permittivity of free space

$$\varepsilon = \varepsilon_r \times \varepsilon_0$$

In old terminology, ε_r is called the "*k* of a dielectric". Glass (SiO_2) at ε_r = 4.5 is considered the division between low *k* and high *k* dielectrics.

Relative Permittivities of Selected Materials

Air (sea level)	1.0006	Polystyrene	2.6
Ammonia	22	Polyethylene	2.25
Bakelite	5	Rubber	2.2-4.1
Glass	4.5-10	Silicon	11.9
Ice	3.2	Soil, dry	2.5-3.5
Mica	5.4-6	Styrofoam	1.03
most metals	~1	Teflon	2.1
Plexiglass	3.4	Vacuum	1
Porcelain	5.7	Water, distilled	81
Paper	2-4	Water, seawater	72-80
Oil	2.1-2.3		

NOTE: Relative permittivity data is given for materials at **low or static frequency conditions**. The permittivity for most materials varies with frequency. The relative permittivities of most materials lie in the range of 1-25. At high frequencies, the permittivity of a material can be quite different (usually less), but will have resonant peaks.

s CONDUCTIVITY [S/m] or [1/(Ω·m)]

A measure of the ability of a material to conduct electricity, the higher the value the better the material conducts. The reciprocal is *resistivity*. Values for common materials vary over about 24 orders of magnitude. Conductivity may often be determined from skin depth or the loss tangent.

$$\sigma = \frac{n_c q_e^2 \bar{l}}{m_e v_{th}} \text{ S/m} \quad \text{where}$$

n_c = density of conduction electrons (for copper this is 8.45×10^{28}) [m^{-3}]

q_e = electron charge? 1.602×10^{-23} [C]

$\bar{l}$ = $v_{th}t_c$ the product of the thermal speed and the mean free time between collisions of electrons, the average distance an electron travels between collisions [m]

m_e = the effective electron mass? [kg]

v_{th} = thermal speed, usually much larger than the drift velocity v_d. [m/s]

Conductivities of Selected Materials [1/(Ω·m)]

Aluminum	3.82×10^7	Mercury	1.04×10^6
Carbon	7.14×10^4	Nicrome	1.00×10^6
Copper (annealed)	5.80×10^7	Nickel	1.45×10^7
Copper (in class)	6.80×10^7	Seawater	4
Fresh water	$\sim10^{-2}$	Silicon	$\sim4.35\times10^{-4}$
Germanium	~2.13	Silver	6.17×10^7
Glass	$\sim10^{-12}$	Sodium	2.17×10^7
Gold	4.10×10^7	Stainless steel	1.11×10^6
Iron	1.03×10^7	Tin	8.77×10^6
Lead	4.57×10	Titanium	2.09×10^6
		Zinc	1.67×10^7

P POWER [W]

Power is the time rate of change of energy.

Power reflected at a discontinuity: $\%\text{ power} = |\rho|^2 \times 100$

Power transmitted at a discontinuity: $\%\text{ power} = \left(1 - |\rho|^2\right) \times 100$

MICROSTRIP CONDUCTORS

How fast does a wave travel in a microstrip? The question is complicated by the fact that the dielectric on one side of the strip may be different from the dielectric on the other side and a wave may travel at different speeds in different dielectrics. The solution is to find an **effective relative permittivity** $\varepsilon_{r\,\mathrm{eff}}$ for the combination.

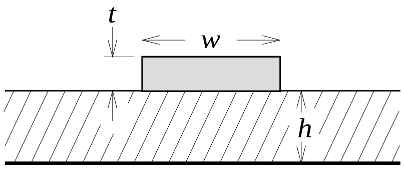

Some Microstrip Relations

$$Z_0^{\mathrm{air}} = Z_0\sqrt{\varepsilon_{r\,\mathrm{eff}}} \qquad C^{\mathrm{air}} Z_0^{\mathrm{air}} = \sqrt{\varepsilon_0\mu_0}$$

$$L = Z_0^{\mathrm{air}}\sqrt{\varepsilon_0\mu_0} = C^{\mathrm{total}}\left(Z_0\right)^2 \qquad L\,C^{\mathrm{air}} = \varepsilon_0\mu_0$$

$$Z_0 = \sqrt{\frac{L}{C^{\mathrm{total}}}} \qquad Z_0^{\mathrm{air}} = \sqrt{\frac{L}{C^{\mathrm{air}}}}$$

$$\gamma = j\beta = j\omega\sqrt{\varepsilon_0\mu_0}\sqrt{\varepsilon_{r\,\mathrm{eff}}} \qquad \varepsilon_{r\,\mathrm{eff}} = \frac{C^{\mathrm{total}}}{C^{\mathrm{air}}}$$

$$v_p = \frac{1}{\sqrt{\varepsilon_0\mu_0\varepsilon_{r\,\mathrm{eff}}}} = \frac{1}{\sqrt{L\,C^{\mathrm{total}}}}$$

It's difficult to get more than 200Ω for Z_0 in a microstrip.

Microstrip Approximations

$$\varepsilon_{r\,\mathrm{eff}} = \frac{\varepsilon_r+1}{2} + \frac{\varepsilon_r-1}{2\sqrt{1+12h/w}}$$

$$Z_0 = \begin{cases} \dfrac{60}{\sqrt{\varepsilon_{r\,\mathrm{eff}}}}\ln\left[\dfrac{8h}{w}+\dfrac{w}{4h}\right], & \text{for } \dfrac{w}{h}\le 1 \\[2ex] \dfrac{120\pi}{\sqrt{\varepsilon_{r\,\mathrm{eff}}}\left[\dfrac{w}{h}+1.393+0.667\ln\left(\dfrac{w}{h}+1.444\right)\right]}, & \text{for } \dfrac{w}{h}>1 \end{cases}$$

$$\frac{w}{h} = \begin{cases} \dfrac{8e^A}{e^{2A}-2}, & \dfrac{w}{h}<2 \\[2ex] \dfrac{2}{\pi}\left\{B-1-\ln\left(2B-1\right)+\dfrac{\varepsilon_r-1}{2\varepsilon_r}\left[\ln\left(B-1\right)+0.39-\dfrac{0.61}{\varepsilon_r}\right]\right\}, & \dfrac{w}{h}>2 \end{cases}$$

where $A = \dfrac{Z_0}{60}\sqrt{\dfrac{\varepsilon_r+1}{2}} + \dfrac{\varepsilon_r-1}{\varepsilon_r+1}\left(0.23+\dfrac{0.11}{\varepsilon_r}\right)$, $B = \dfrac{377\pi}{2Z_0\sqrt{\varepsilon_r}}$

STRIPLINE CONDUCTOR

Also called *shielded microstrip*. The effective relative permittivity is used in calculations.

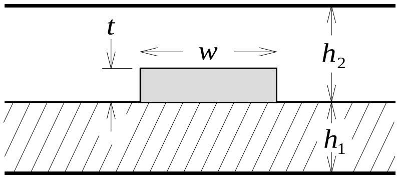

assuming $w \ge 10h$, $\quad \varepsilon_{r\,\mathrm{eff}} = \dfrac{\varepsilon_{r1}h_1+\varepsilon_{r2}h_2}{h_1+h_2}$ where

ε_{r1} = the relative permittivity of the dielectric of thickness h_1.
ε_{r2} = the relative permittivity of the dielectric of thickness h_2.

COPPER CLADDING

The thickness of copper on a circuit board is measured in ounces. 1-ounce cladding means that 1 square foot of the copper weighs 1 ounce. 1-ounce copper is 0.0014" or 35.6 μm thick.

α_d DIELECTRIC LOSS FACTOR [dB/cm]

$$\alpha_d = 8.68\frac{\beta_0\varepsilon_r\left(\varepsilon_{r\,\mathrm{eff}}-1\right)}{2\sqrt{\varepsilon_{r\,\mathrm{eff}}}\left(\varepsilon_r-1\right)}\tan\delta$$

α_c CONDUCTOR LOSS FACTOR [dB/cm]

$$\alpha_c = 8.68\frac{R}{2Z_0}, \qquad R = \frac{1}{\sigma\delta(\text{perimeter})} = \sqrt{\frac{\omega\mu_0}{2\sigma}}\frac{1}{(\text{perimeter})}$$

WHEELER'S EQUATION

Another approximation for microstrip calculations is Wheeler's equation.

$$Z_0 = \frac{42.4}{\sqrt{1+\varepsilon_r}} \ln\left(1+\left\{\frac{4h}{w'}\left[\frac{14+\frac{8}{\varepsilon_r}}{11}\times\frac{4h}{w'}+\sqrt{\left(\frac{14+\frac{8}{\varepsilon_r}}{11}\times\frac{4h}{w'}\right)^2+\pi^2\frac{1+\frac{1}{\varepsilon_r}}{2}}\right]\right\}\right)$$

where $w' = \dfrac{8h\sqrt{\dfrac{7+\frac{4}{\varepsilon_r}}{11}\left[\exp\left(\dfrac{Z_0}{42.4}\sqrt{\varepsilon_r+1}\right)-1\right]+\dfrac{1+\frac{1}{\varepsilon_r}}{0.81}}}{\exp\left(\dfrac{Z_0}{42.4}\sqrt{\varepsilon_r+1}\right)-1}$

NETWORK THEORY

S_{ij} SCATTERING PARAMETER

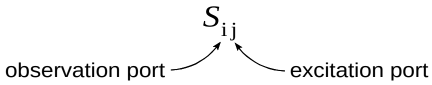

A scattering parameter, represented by S_{ij}, is a dimensionless value representing the fraction of wave amplitude transmitted from port j into port i, provided that all other ports are terminated with matched loads and only port j is receiving a signal. Under these same conditions, S_{ii} is the reflection coefficient at port i.

To experimentally determine the scattering parameters, attach an impedance-matched generator to one of the ports (***excitation port***), attach impedance-matched loads to the remaining ports, and observe the signal received at each of the ports (***observation ports***). The fractional amounts of signal amplitude received at each port i will make up one column j of the **scattering matrix**. Repeating the process for each column would require n^2 measurements to determine the scattering matrix for an n-port network.

S_{ij} SCATTERING MATRIX

$$\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1N} \\ S_{21} & S_{22} & \cdots & S_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ S_{N1} & S_{N2} & \cdots & S_{NN} \end{bmatrix}$$

The scattering matrix is an $n\times n$ matrix composed of scattering parameters that describes an n-port network.

The elements of the diagonal of the scattering matrix are reflection coefficients of each port. The elements of the off-diagonal are transmission coefficients, under the conditions outlined in "SCATTERING PARAMETER".

If the network is **internally matched** or **self-matched**, then $S_{11} = S_{22} = \cdots = S_{NN} = 0$, that is, the diagonal is all zeros.

The sum of the squares of each column of a scattering matrix is equal to one, provided the network is lossless.

a_n, b_n INCIDENT/REFLECTED WAVE AMPLITUDES

The parameters a_n and b_n describe the incident and reflected waves respectively at each port n. These parameters are used for power and scattering matrix calculations.

The amplitude of the wave incident to port n is equal to the amplitude of the incident voltage at the port divided by the square root of the port impedance.	$a_n = \frac{V_n^+}{\sqrt{Z_{0n}}}$
Amplitude of the wave reflected at port n is equal to the amplitude of the reflected voltage at the port divided by the square root of the port impedance.	$b_n = \frac{V_n^-}{\sqrt{Z_{0n}}}$
The scattering parameter is equal to the wave amplitude output at port i divided by the wave amplitude input at port j provided the only source is a matched source at port j and all other ports are connected to matched loads.	$S_{\mathrm{ij}} = \frac{b_{\mathrm{i}}}{a_{\mathrm{j}}}$
The relationship between the S-parameters and the a- and b-parameters can be written in matrix form where **S** is the scattering matrix and **a** and **b** are column vectors.	$\mathbf{b} = \mathbf{Sa}$
Power flow into any port is shown as a function of a- and b-parameters.	$P = \frac{1}{2}\left(\lvert a\rvert^2 - \lvert b\rvert^2\right)$
The ratio of the input power at port j to the output power at port I can be written as a function of a- and b-parameters or the S-parameter.	$\frac{P_{in\,\mathrm{j}}}{P_{out\,\mathrm{i}}} = \frac{\lvert a_{\mathrm{j}}\rvert^2}{\lvert b_{\mathrm{j}}\rvert^2} = \frac{1}{\lvert S_{\mathrm{ij}}\rvert^2}$

RECIPROCITY

A network is reciprocal when $S_{\mathrm{ij}} = S_{\mathrm{ji}}$ in the scattering matrix, i.e. the matrix is symmetric across the diagonal. Also, $Z_{\mathrm{ij}} = Z_{\mathrm{ji}}$ and $Y_{\mathrm{ij}} = Y_{\mathrm{ji}}$. Networks constructed of "normal materials" exhibit reciprocity.

Reciprocity Theorem:

$$\oint_S \vec{E}_a \times \vec{H}_b \cdot ds = \oint_S \vec{E}_b \times \vec{H}_a \cdot ds$$

E_a and H_b are fields from two different sources.

LOSSLESS NETWORK

A network is lossless when

$$\underline{S}\ \underline{S}^{\dagger} = /$$

† means to take the complex conjugate and transpose the matrix. If the network is reciprocal, then the transpose is the same as the original matrix.

/ = a unitary matrix. A **unitary matrix** has the properties:

$$\sum_{k=1}^{N} S_{ki} S_{ki}^{*} = 1 \qquad \sum_{k=1}^{N} S_{ki} S_{kj}^{*} = 0$$

In other words, a column of a unitary matrix multiplied by its complex conjugate equals one, and a column of a unitary matrix multiplied by the complex conjugate of a different column equals zero.

RAT RACE OR HYBRID RING NETWORK

The rat race or hybrid ring network is lossless, reciprocal, and internally matched.

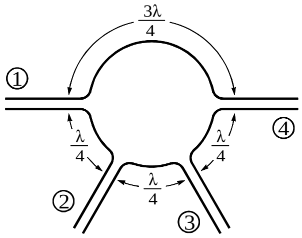

The signal splits upon entering the network and half travels around each side. A signal entering at port 1 and exiting at port 4 travels ¾ of a wavelength along each side, so the signals are in phase and additive. From port 1 to port 3 the signal travels one wavelength along one side and ½ wavelength along the other, arriving a port 3 out of phase and thus canceling. From port 1 to port 2 the paths are ¼ and 5/4 wavelengths respectively, thus they are in phase and additive.

DIRECTIONAL COUPLER

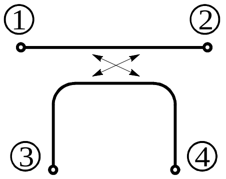

The directional coupler is a 4-port network similar to the rat race. It can be used to measure reflected and transmitted power to an antenna.

An input at one port is divided between two of the remaining ports. The **coupling factor**, measured in dB, describes the division of signal strength at the two ports. For example if the coupler has a coupling factor of –10 dB, then a signal input at port 1 would appear at port 4 attenuated by 10 dB with the majority of the signal passing to port 2. In other words, 90% of the signal would appear at port 2 and 10% at port 4. (-10 dB means "**10 dB down**" or 0.1 power, -6 dB means 0.25 power, and –3 dB means 0.5 power.) A reflection from port 2 would appear at port 3 attenuated by the same amount. Meters attached to ports 3 and 4 could be used to measure reflected and transmitted power for a system with a transmitter connected to port 1 and an antenna at port 2. The **directivity** of a coupler is a measurement of how well the coupler transfers the signal to the appropriate output without reflection due to the coupler itself; the directivity approaches infinity for a perfect coupler. $\text{directivity} = 10\log\left(p_3 / p_1\right)$, where the source is at port 1 and the load is at port 2.

The directional coupler is **lossless** and **reciprocal**. The scattering matrix looks like this. In a real coupler, the off-diagonal zeros would be near zero due to leakage.

$$\begin{bmatrix} 0 & p & 0 & -q \\ p & 0 & q & 0 \\ 0 & q & 0 & p \\ -q & 0 & p & 0 \end{bmatrix}$$

CIRCULATOR

The circulator is a 3-port network that can be used to prevent reflection at the antenna from returning to the source.

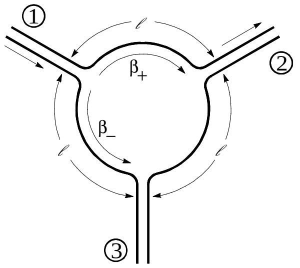

Port 3 is terminated internally by a matched load. With a source at 1 and a load at 2, any power reflected at the load is absorbed by the load resistance at port 3. A 3-port network cannot be both lossless and reciprocal, so the circulator is not reciprocal.

Schematically, the circulator may be depicted like this:

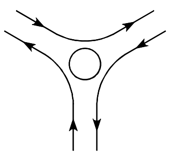

The circulator is **lossless** but is not **reciprocal**. The scattering matrix looks like this:

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

MAXWELL'S EQUATIONS, TIME HARMONIC FORM

$$\nabla \times \bar{\mathscr{E}} = -j\omega\mu\mathbf{H} \qquad \text{"curl on E"}$$

$$\nabla \times \bar{\mathscr{H}} = -j\omega\mu\mathbf{E} \qquad \text{"curl on H"}$$

$$\bar{\mathscr{E}} = \left[E_x\left(x,y\right)\hat{\mathbf{x}} + E_y\left(x,y\right)\hat{\mathbf{y}} + E_z\left(x,y\right)\hat{\mathbf{z}}\right]e^{j\omega t-\gamma z}$$

$$\bar{\mathscr{H}} = \left[H_x\left(x,y\right)\hat{\mathbf{x}} + H_y\left(x,y\right)\hat{\mathbf{y}} + H_z\left(x,y\right)\hat{\mathbf{z}}\right]e^{j\omega t-\gamma z}$$

From the curl equations we can derive:

(1) $\dfrac{\partial E_z}{\partial y} + \gamma E_y = -j\omega\mu H_x$ (4) $\dfrac{\partial H_z}{\partial y} + \gamma H_y = j\omega\varepsilon E_x$

(2) $-\dfrac{\partial E_z}{\partial x} - \gamma E_x = -j\omega\mu H_y$ (5) $-\dfrac{\partial H_z}{\partial x} - \gamma H_x = j\omega\varepsilon E_y$

(3) $\dfrac{\partial E_y}{\partial x} - \dfrac{\partial E_x}{\partial y} = -j\omega\mu H_z$ (6) $\dfrac{\partial H_y}{\partial x} - \dfrac{\partial H_x}{\partial y} = j\omega\varepsilon E_z$

From the above equations we can obtain:

(1) & (5) $H_x = \dfrac{1}{\gamma^2 + \omega^2\mu\varepsilon}\left(j\omega\varepsilon\dfrac{\partial E_z}{\partial y} - \gamma\dfrac{\partial H_z}{\partial x}\right)$

(2) & (4) $H_y = \dfrac{1}{\gamma^2 + \omega^2\mu\varepsilon}\left(j\omega\varepsilon\dfrac{\partial E_z}{\partial x} - \gamma\dfrac{\partial H_z}{\partial y}\right)$

(2) & (4) $E_x = -\dfrac{1}{\gamma^2 + \omega^2\mu\varepsilon}\left(-\gamma\dfrac{\partial E_z}{\partial x} + j\omega\mu\dfrac{\partial H_z}{\partial y}\right)$

(1) & (5) $E_y = -\dfrac{1}{\gamma^2 + \omega^2\mu\varepsilon}\left(-\gamma\dfrac{\partial E_z}{\partial y} + j\omega\mu\dfrac{\partial H_z}{\partial x}\right)$

This makes it look like if E_z and H_z are zero, then H_x, H_y, E_x, and E_y are all zero. But since $\infty \times 0 \neq 0$, we could have non-zero result for the TEM wave if $\gamma^2 = -\omega^2\mu\varepsilon \Rightarrow \gamma = j\omega\sqrt{\mu\varepsilon}$. This should look familiar.

WAVE EQUATIONS

From Maxwell's equations and a vector identity on curl, we can get the following wave equations:

$$\nabla^2 \overline{\mathscr{E}} = -\omega^2 \mu\varepsilon \overline{\mathscr{E}} \qquad \text{"del squared on E"}$$

$$\nabla^2 \overline{\mathscr{H}} = \omega^2 \mu\varepsilon \overline{\mathscr{H}} \qquad \text{"del squared on H"}$$

The z part or "del squared on E_z" is:

$$\nabla^2 E_z = \frac{\gamma^2 E_z}{\partial x^2} + \frac{\gamma^2 E_z}{\partial y^2} + \frac{\gamma^2 E_z}{\partial z^2} = -\omega^2 \mu\varepsilon E_z$$

Using the separation of variables, we can let:

$$E_z = X(x) \cdot Y(y) \cdot Z(z)$$

We substitute this into the previous equation and divide by $X \cdot Y \cdot Z$ to get:

$$\underbrace{\frac{1}{X}\frac{d^2X}{dx^2}}_{-k_x{}^2} + \underbrace{\frac{1}{Y}\frac{d^2Y}{dy^2}}_{-k_y{}^2} + \underbrace{\frac{1}{Z}\frac{d^2Z}{dz^2}}_{-k_z{}^2} = \underbrace{-\omega^2\mu\varepsilon}_{\text{a constant}}$$

Since X, Y, and Z are independent variables, the only way the sum of these 3 expressions can equal a constant is if all 3 expressions are constants.

So we are letting $\frac{1}{Z}\frac{d^2Z}{dz^2} = -k_z{}^2 \Rightarrow \frac{d^2Z}{dz^2} = -Zk_z{}^2$

A solution could be $Z = e^{-\gamma z}$

so that $\gamma^2 e^{-\gamma z} = -k_z{}^2 e^{-\gamma z}$ and $-k_z{}^2 = \gamma^2$

Solutions for X and Y are found

$$\frac{1}{X}\frac{d^2X}{dx^2} = -k_x{}^2 \Rightarrow X = A\sin(k_x x) + B\cos(k_x x)$$

$$\frac{1}{Y}\frac{d^2Y}{dy^2} = -k_y{}^2 \Rightarrow Y = C\sin(k_y y) + D\cos(k_y y)$$

giving us the general solution $k_x{}^2 + k_y{}^2 - \gamma^2 = \omega^2\mu\varepsilon$

For a particular solution we need to specify initial conditions and boundary conditions. For some reason, initial conditions are not an issue. The unknowns are k_x, k_y, A, B, C, D. The boundary conditions are

$$E_{\tan} = 0 \qquad \frac{\partial H_{\tan}}{\partial n} = 0$$

$E_{\tan}$ = the electric field tangential to a conducting surface
$H_{\tan}$ = the magnetic field tangential to a conducting surface
n = I don't know

TM, TE WAVES IN PARALLEL PLATES

TM, or transverse magnetic, means that magnetic waves are confined to the *transverse plane*. Similarly, TE (transverse electric) means that electrical waves are confined to the transverse plane.

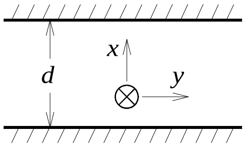

(z direction is into page)

Transverse plane means the plane that is transverse to (perpendicular to) the direction of propagation. The direction of propagation is taken to be in the z direction, so the transverse plane is the x-y plane. So for a TM wave, there is no H_z component (magnetic component in the z direction) but there is an E_z component.

$$\boxed{E_z = A\sin(k_x x)e^{-\gamma z}}$$

A = amplitude [V]

$k_x = \frac{m\pi}{d}$ The magnetic field must be zero at the plate boundaries. This value provides that characteristic. [cm^{-1}]

x = position; perpendicular distance from one plate. [cm]
d = plate separation [cm]
γ = propagation constant
z = position along the direction of propagation [cm]
m = mode number; an integer greater than or equal to 1

$$\boxed{\gamma = \sqrt{-\omega^2\mu\varepsilon + (kx)^2}}$$

Notice than when $(kx)^2 \geq \omega^2\mu\varepsilon$, the quantity under the square root sign will be positive and γ will be purely real. In this circumstance, the wave is said to be evanescent. The wavelength goes to infinity; there is no oscillation or propagation. On the other hand, when $(kx)^2 < \omega^2\mu\varepsilon$, γ is purely imaginary.

The magnitude of E_z is related to its position between the plates and the mode number m. Note that for m = 2 that $d = \lambda$.

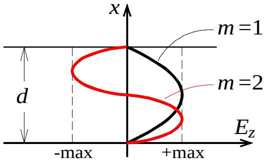

GENERAL MATHEMATICAL

COMPLEX TO POLAR NOTATION

j in polar notation:

$$j = e^{j\frac{\pi}{2}}$$

So we can find the square root of *j*:

$$\sqrt{j} = \sqrt{e^{j\frac{\pi}{2}}} = e^{j\frac{\pi}{4}} = \frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}}$$

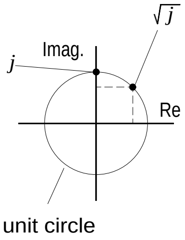

dBm DECIBELS RELATIVE TO 1 mW

The decibel expression for power. The logarithmic nature of decibel units translates the multiplication and division associated with gains and losses into addition and subtraction.

0 dBm = 1 mW
20 dBm = 100 mW
-20 dBm = 0.01 mW

$$P(\text{dBm}) = 10\log\left[P(\text{mW})\right]$$

$$P(\text{mW}) = 10^{P(\text{dBm})/10}$$

PHASOR NOTATION

To express a derivative in phasor notation, replace $\frac{\partial}{\partial t}$ with $j\omega$. For example, the Telegrapher's equation $\frac{\partial V}{\partial z} = -L\frac{\partial I}{\partial t}$

becomes $\frac{\partial V}{\partial z} = -Lj\omega I$.

Ñ NABLA, DEL OR GRAD OPERATOR

Compare the ∇ operation to taking the time derivative. Where ∂/∂*t* means to take the derivative with respect to time and introduces a s^{-1} component to the units of the result, the ∇ operation means to take the derivative with respect to distance (in 3 dimensions) and introduces a m^{-1} component to the units of the result. ∇ terms may be called space derivatives and an equation which contains the ∇ operator may be called a vector differential equation. In other words ∇**A** is how fast **A** changes as you move through space.

in rectangular coordinates: $\nabla \mathbf{A} = \hat{x}\frac{\partial A}{\partial x} + \hat{y}\frac{\partial A}{\partial y} + \hat{z}\frac{\partial A}{\partial z}$

in cylindrical coordinates: $\nabla \mathbf{A} = \hat{r}\frac{\partial A}{\partial r} + \hat{\phi}\frac{1}{r}\frac{\partial A}{\partial \phi} + \hat{z}\frac{\partial A}{\partial z}$

in spherical coordinates: $\nabla \mathbf{A} = \hat{r}\frac{\partial A}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial A}{\partial \theta} + \hat{\phi}\frac{1}{r\sin\theta}\frac{\partial A}{\partial \phi}$

∇ GRADIENT

$$\boxed{\nabla\bar{\Phi} = -\mathbf{E}}$$

"The gradient of the vector Φ" or "del Φ" is equal to the negative of the electric field vector.

∇Φ is a vector giving the direction and magnitude of the maximum spatial variation of the scalar function Φ at a point in space.

$$\nabla\bar{\Phi} = \hat{\mathbf{x}}\frac{\partial\Phi}{\partial x} + \hat{\mathbf{y}}\frac{\partial\Phi}{\partial y} + \hat{\mathbf{z}}\frac{\partial\Phi}{\partial z}$$

Ñx DIVERGENCE

∇· is also a vector operator, combining the "del" or "grad" operator with the dot product operator and is read as "the divergence of". In this form of Gauss' law, where **D** is a density per unit area, with the operators applied, ∇·**D** becomes a density per unit volume.

$$\boxed{\text{div}\,\mathbf{D} = \nabla\cdot\mathbf{D} = \frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} = \rho}$$

D = electric flux density vector **D** = ε**E** [C/m^2]
ρ = source charge density [C/m^3]

Ñ^2 THE LAPLACIAN

∇^2 is a combination of the divergence and del operations, i.e. $\mathrm{div}(\mathrm{grad}\ \Phi) = \nabla\cdot\nabla\ \Phi = \nabla^2\ \Phi$. It is read as "the LaPlacian of" or "del squared".

$$\boxed{\nabla^2\mathbf{F} = \frac{\partial^2\,\Phi}{\partial\,x^2} + \frac{\partial^2\,\Phi}{\partial\,y^2} + \frac{\partial^2\,\Phi}{\partial\,z^2}}$$

F = electric potential [V]

GRAPHING TERMINOLOGY

With *x* being the horizontal axis and *y* the vertical, we have a graph of ***y* versus *x*** or ***y* as a function of *x***. The *x*-axis represents the **independent variable** and the *y*-axis represents the **dependent variable**, so that when a graph is used to illustrate data, the data of regular interval (often this is time) is plotted on the *x*-axis and the corresponding data is dependent on those values and is plotted on the *y*-axis.

HYPERBOLIC FUNCTIONS

$$j\sin\theta = \sinh\left(j\theta\right)$$

$$j\cos\theta = \cosh\left(j\theta\right)$$

$$j\tan\theta = \tanh\left(j\theta\right)$$

TAYLOR SERIES

$$\sqrt{1+x} \approx 1 + \frac{1}{2}x\,, \quad x \ll 1$$

$$\frac{1}{1-x^2} \approx 1 + x^2 + x^4 + x^6 + \cdots\,, \quad |x| < 1$$

$$\frac{1}{1\pm x} \approx 1 \mp x\,, \quad x \ll 1$$

ELECTROMAGNETIC SPECTRUM

FREQUENCY	WAVELENGTH (free space)	DESIGNATION	APPLICATIONS
< 3 Hz	> 100 Mm		Geophysical prospecting
3-30 Hz	10-100 Mm	ELF	Detection of buried metals
30-300 Hz	1-10 Mm	SLF	Power transmission, submarine communications
0.3-3 kHz	0.1-1 Mm	ULF	Telephone, audio
3-30 kHz	10-100 km	VLF	Navigation, positioning, naval communications
30-300 kHz	1-10 km	LF	Navigation, radio beacons
0.3-3 MHz	0.1-1 km	MF	AM broadcasting
3-30 MHz	10-100 m	HF	Short wave, citizens' band
30-300 MHz 54-72 76-88 88-108 174-216	1-10 m	VHF	TV, FM, police TV channels 2-4 TV channels 5-6 FM radio TV channels 7-13
0.3-3 GHz 470-890 MHz 915 MHz 800-2500 MHz 1-2 2.45 2-4	10-100 cm	UHF "money band"	Radar, TV, GPS, cellular phone TV channels 14-83 Microwave ovens (Europe) PCS cellular phones, analog at 900 MHz, GSM/CDMA at 1900 L-band, GPS system Microwave ovens (U.S.) S-band
3-30 GHz 4-8 8-12 12-18 18-27	1-10 cm	SHF	Radar, satellite communications C-band X-band (Police radar at 11 GHz) K_u-band (dBS Primestar at 14 GHz) K-band (Police radar at 22 GHz)
30-300 GHz 27-40 40-60 60-80 80-100	0.1-1 cm	EHF	Radar, remote sensing K_a-band (Police radar at 35 GHz) U-band V-band W-band
0.3-1 THz	0.3-1 mm	Millimeter	Astromony, meteorology
10^{12}-10^{14} Hz	3-300 μm	Infrared	Heating, night vision, optical communications
3.95×10^{14}- 7.7×10^{14} Hz	390-760 nm 625-760 600-625 577-600 492-577 455-492 390-455	Visible light	Vision, astronomy, optical communications Red Orange Yellow Green Blue Violet
10^{15}-10^{18} Hz	0.3-300 nm	Ultraviolet	Sterilization
10^{16}-10^{21} Hz		X-rays	Medical diagnosis
10^{18}-10^{22} Hz		γ-rays	Cancer therapy, astrophysics
> 10^{22} Hz		Cosmic rays	Astrophysics

GLOSSARY

anisotropic materials materials in which the electric polarization vector is not in the same direction as the electric field. The values of ε, μ, and σ are dependent on the field direction. Examples are crystal structures and ionized gases.

complex permittivity e The imaginary part accounts for heat loss in the medium due to damping of the vibrating dipole moments.

dielectric An insulator. When the presence of an applied field displaces electrons within a molecule away from their average positions, the material is said to be polarized. When we consider the polarizations of insulators, we refer to them as *dielectrics*.

empirical A result based on observation or experience rather than theory, e.g. *empirical data*, *empirical formulas*. Capable of being verified or disproved by observation or experiment, e.g. *empirical laws*.

evanescent wave A wave for which $\beta=0$. α will be negative. That is, γ is purely real. The wave has infinite wavelength—there is no oscillation.

isotropic materials materials in which the electric polarization vector is in the same direction as the electric field. The material responds in the same way for all directions of an electric field vector, i.e. the values of ε, μ, and σ are constant regardless of the field direction.

linear materials materials which respond proportionally to increased field levels. The value of μ is not related to $\mathbf{H}$ and the value of ε is not related to $\mathbf{E}$. Glass is linear, iron is non-linear.

overdamped system in the case of a transmission line, this means that when the source voltage is applied the line voltage rises to the final voltage without exceeding it.

time variable materials materials whose response to an electric field changes over time, e.g. when a sound wave passes through them.

transverse plane perpendicular, e.g. the x-y plane is *transverse* to z.

underdamped system in the case of a transmission line, this means that after the source voltage is applied the line voltage periodically exceeds the final voltage.

wave number k The phase constant for the uniform plane wave. k may be considered a constant of the medium at a particular frequency.

Abbreviations and Acronyms

The following abbreviations are used in this paper. Abbreviations are defined on first use.

1G – First Generation
1xEV-DO – One Carrier Evolution, Data Optimized
1xEV-DV – One Carrier Evolution, Data Voice
1XRTT – One Carrier Radio Transmission Technology
2G – Second Generation
3G – Third Generation (meeting requirements set forth by the ITU IMT project)
3GPP – Third Generation Partnership Project
3GPP2 – Third Generation Partnership Project 2
4G – Fourth Generation (meeting requirements set forth by the ITU IMT-Advanced project)
5GC – 5G Core
8-PSK – Octagonal Phase Shift Keying
AAS – Adaptive Antenna Systems
ABR – Allocation Retention Priority
AGW – Access Gateway
AMF – Access and Mobility Management Function
AMR – Adaptive Multi Rate
AMR-WB – Adaptive Multi-Rate Wideband
ANDSF – Access Network Discovery and Selection Function.
ANSI – American National Standards Institute
APCO – Association of Public Safety Officials
API – Application Programming Interface
APN – Access Point Name
ARP – Allocation Retention Priority
ARPU – Average Revenue per User
ARQ – Automatic Repeat Request
ATM – Asynchronous Transfer Mode
AWGN – Additive White Gaussian Noise Channel
AWS – Advanced Wireless Services
BCCH – Broadcast Control Channel
bps – bits per second
BRS – Broadband Radio Service
BSC – Base Station Controller
BTS – Base Transceiver Station
C/I – Carrier to Intermodulation Ratio
CAPEX- Capital Expenditure
CBF – Coordinated Beam Forming
CBRS – Citizens Broadband Radio Service

CBS – Coordinated Beam Switching
CSS3 – Cascading Style Sheets 3 (CSS3)
CDD – Cyclic Delay Diversity
CDF – Cumulative Distribution Function
CDMA – Code Division Multiple Access
CL – Closed Loop
CL-SM – Closed Loop Spatial Multiplexing
CMAS – Commercial Mobile Alert System
CMOS – Complementary Metal Oxide Semiconductor
CoMP – Coordinated Multi Point
cMTC – Critical Machine Type Communications
CP – Cyclic Prefix
CPC – Continuous Packet Connectivity
CPRI – Common Public Radio Interface
CQI - Channel Quality Indicators
C-RAN – Cloud Radio Access Network
CRM – Customer Relationship Management
CRS – Cell-specific Reference Signal
CS – Convergence Sublayer
CSFB – Circuit-Switched Fallback
CTIA – Cellular Telephone Industries Association
C-V2X – Cellular Vehicle-to-X
D-AMPS – Digital Advanced Mobile Phone Service
DAS – Distributed Antenna System
DAS – Downlink EGPRS2-A Level Scheme
dB – Decibel
DBS – Downlink EGPRS2-B Level Scheme
DC-HSPA – Dual Carrier HSPA
DFT – Discrete Fourier Transform
DL – Downlink
DNS – Domain Name Service
DPCCH – Dedicated Physical Control Channel
DPS – Dynamic Point Selection
DSL – Digital Subscriber Line
DSMIPv6 – Dual Stack Mobile IPv6
DTM – Dual Transfer Mode
DRX – Discontinuous Reception
D-TxAA – Double Transmit Adaptive Array
DVB-H – Digital Video Broadcasting Handheld

E-DCH – Enhanced Dedicated Channel

EBCMCS – Enhanced Broadcast Multicast Services

EC-GSM – Extended Coverage GSM

eCoMP – enhanced CoMP

EDGE – Enhanced Data Rates for GSM Evolution

EFTA – Enhanced Flexible Timeslot Assignment

EGPRS – Enhanced General Packet Radio Service

eICIC – Enhanced Inter-Cell Interference Coordination

eMBMS – Enhanced Multimedia Broadcast Multicast Services

eNodeB – Evolved Node B

EAP – Extensible Authentication Protocol

eLAA – Enhanced Licensed-Assisted Access

eNB – Evolved Node B

EPC – Evolved Packet Core

EPDCCH – Enhanced Physical Downlink Control Channel

eMBB – Enhanced Mobile Broadband

ePDG – Enhanced Packet Data Gateway

EPS – Evolved Packet System

ERP – Enterprise Resource Planning

eSaMOG – Enhanced S2a-based Mobility over GTP

ESC – Environmental Sensing Capability

eSRVCC – Enhanced Single-Radio Voice Call Continuity

ETRI – Electronic and Telecommunications Research Institute

ETSI – European Telecommunications Standards Institute

E-UTRAN – Enhanced UMTS Terrestrial Radio Access Network

EVS – Enhanced Voice Services (codec)

FE-FACH – Further Enhanced Forward Access Channel

EV-DO –Evolution, Data Optimized

EV-DV –Evolution, Data Voice

EVRC – Enhanced Variable Rate Codec

FBMC – Filter-Bank Multi-Carrier

FCC – Federal Communications Commission

FDD – Frequency Division Duplex

FeCoMP – Further Enhanced Coordinated Multi Point

feICIC – Further enhanced ICIC

FirstNet – First Responder Network Authority

Flash OFDM – Fast Low-Latency Access with Seamless Handoff OFDM

FLO – Forward-Link Only

FMC – Fixed Mobile Convergence

FP7 – Seventh Framework Programme
FTP – File Transfer Protocol
GAA – General Authorized Access
GAN – Generic Access Network
GB – Gigabyte
Gbps – Gigabits Per Second
GBR – Guaranteed Bit Rate
GByte – Gigabyte
GCS – Group Communication Service
GERAN – GSM EDGE Radio Access Network
GFDM – Generalized Frequency Division Multiplexing
GGSN – Gateway GPRS Support Node
GHz — Gigahertz
GMSK – Gaussian Minimum Shift Keying
gNB – NR NodeB
GPRS – General Packet Radio Service
G-Rake – Generalized Rake Receiver
GSM – Global System for Mobile Communications
GSMA – GSM Association
HARQ – Hybrid Automatic Repeat Request
HD – High Definition
HetNet – heterogeneous network
HLR – Home Location Register
Hr – Hour
HSDPA – High Speed Downlink Packet Access
HS-FACH – High Speed Forward Access Channel
HS-PDSCH - High Speed Physical Downlink Shared Channels
HS-RACH – High Speed Reverse Access Channel
HSPA – High Speed Packet Access (HSDPA with HSUPA)
HSPA+ – HSPA Evolution
HSS – Home Subscriber Server
HSUPA – High Speed Uplink Packet Access
Hz – Hertz
ICIC – Inter-Cell Interference Coordination
ICN – Information-Centric Networking
ICS – IMS Centralized Services
ICT – Information and Communication Technologies
IEEE – Institute of Electrical and Electronic Engineers
IETF – Internet Engineering Taskforce

IFFT – Inverse Fast Fourier Transform
IFOM – IP Flow and Seamless Offload
IM – Instant Messaging
IMS – IP Multimedia Subsystem
IMT – International Mobile Telecommunications
IMT-Advanced - International Mobile Telecommunications-Advanced
IRC – Interference Rejection Combining
IoT – Internet of Things
IPR - Intellectual Property Rights
IP – Internet Protocol
IPTV – Internet Protocol Television
IR – Incremental Redundancy
ISD – Inter-site Distance
ISI – Intersymbol Interference
ISP – Internet Service Provider
ITU – International Telecommunication Union
JCP – Java Community Process
JR – Joint Reception
JT – Joint Transmission
Kbps – Kilobits Per Second
kHz — Kilohertz
km – Kilometer
LAA – License-Assisted Access
LBT – Listen-Before-Talk
LDPC - Low-Density Parity Code
LIPA – Local IP Access
LMDS – Local Multipoint Distribution Service
LPWA – Low-Power Wide-Area
LTE – Long Term Evolution
LTE-A – LTE-Advanced
LTE-TDD – LTE Time Division Duplex
LTE-U – LTE-Unlicensed
LSTI – LTE/SAE Trial Initiative
LWA – LTE Wi-Fi Aggregation
LWIP – LTE WLAN Radio Level Integration with IPsec Tunnel
M2M – Machine-to-machine
MAC – Medium-Access Control
MAPCON – Multi-Access PDN Connectivity
MB - Megabyte

MBMS - Multimedia Broadcast/Multicast Service

Mbps – Megabits Per Second

MBR – Maximum Bit Rate

MBSFN – Multicast/broadcast, Single Frequency

MCPA – Mobile Consumer Application Platform

Mcps – Megachips Per Second

MCPTT – Mission-Critical Push-to-Talk

MCS – Modulation and Coding Scheme

MCW – Multiple Codeword

MDT – Minimization of Drive Tests

MEAP – Mobile Enterprise Application Platforms

MEC – Multi-access Edge Computing

MediaFLO – Media Forward Link Only

METIS – Mobile and wireless communications Enablers for the Twenty-twenty Information Society

MHz – Megahertz

MID – Mobile Internet Devices

MIMO – Multiple Input Multiple Output

MMSE – Minimum Mean Square Error

mITF – Japan Mobile IT Forum

MMDS – Multichannel Multipoint Distribution Service

MME – Mobile Management Entity

mMTC – Massive Machine Type Communications

MOS – Mean Opinion Score

MP-QUIC – Multipath Quick UDP Internet Connections

MP-TCP – Multipath TCP

MRxD – Mobile Receive Diversity

ms – millisecond

MS – Mobile Station

MSA – Mobile Service Architecture

MSC – Mobile Switching Center

MTC – Machine Type Communications

MTC-IWF – Machine-Type Communications Interworking Function (MTC-IWF)

msec – millisecond

MU-MIMO – Multi-User MIMO

MUST – Downlink Multiuser Superposition Transmission

NAICS – Network-Assisted Interference Cancellation and Suppression

NB-IoT – Narrowband Internet of Things

NENA – National Emergency Number Association

NGMC – Next Generation Mobile Committee

NGMN – Next Generation Mobile Networks Alliance
NG-RAN – New Generation Radio Access Network
NOMA – Non-Orthogonal Multiple Access
NR – New Radio
NTIA – National Telecommunications and Information Administration
OFDM – Orthogonal Frequency Division Multiplexing
OFDMA – Orthogonal Frequency Division Multiple Access
OL-SM – Open Loop Spatial Multiplexing
OMA – Open Mobile Alliance
ORI – Open Radio Equipment Interface
PA – Priority Access
PAL – Priority Access License
PAR – Peak to Average Ratio
PBCCH – Packet Broadcast Control Channel
PCH – Paging Channel
PCRF – Policy Control and Charging Rules Function
PCS – Personal Communications Service
PDCP – Packet Data Convergence Protocol
PDN – Packet Data Network
PGW – Packet Gateway
PHY – Physical Layer
PMI – Precoding Matrix Indication
PMIPv6 – Proxy Mobile IPv6
PNF – Physical Network Function
PoC – Push-to-Talk Over Cellular
PSH – Packet Switched Handover
PSK – Phase-Shift Keying
QAM – Quadrature Amplitude Modulation
QCI – Quality of Service Class Identifier
QLIC – Quasi-Linear Interference Cancellation
QoS – Quality of Service
QPSK – Quadrature Phase Shift Keying
QUIC – Quick UDP Internet Connections.
RAB – Radio Access Bearer
RAN – Radio Access Network
RCAF – RAN Congestion Awareness Function
RCLWI - RAN Controlled LTE WLAN Interworking
RCS – Rich Communications Suite
REST – Representational State Transfer

RF – Radio Frequency

RLC – Radio Link Control

RNC – Radio Network Controller

ROHC – Robust Header Compression

RRC – Radio Resource Control

RRH – Remote Radio Head

RRU – Remote Radio Unit

RTP – Real Time Transport Protocol

RTSP – Real Time Streaming Protocol

SAE – System Architecture Evolution

SaMOG – S2a-based Mobility over GTP

SAS – Spectrum Access System

SC-FDMA – Single Carrier Frequency Division Multiple Access

SCMA – Sparse Coded Multiple Access

SCRI – Signaling Connection Release Indication

SCW – Single Codeword

SDAP – Service Data Adaptation Protocol

SDMA – Space Division Multiple Access

SDN – Software-Defined Networking

SDP – Session Description Protocol

sec – Second

SFBA – Space Frequency Block Code

SFN – Single Frequency Network

SGSN – Serving GPRS Support Node

SGW – Serving Gateway

SIC – Successive Interference Cancellation

SIM – Subscriber Identity Module

SIMO – Single Input Multiple Output

SINR – Signal to Interference Plus Noise Ratio

SIP – Session Initiation Protocol

SIPTO – Selected IP Traffic Offload

SISO – Single Input Single Output

SMS – Short Message Service

SNR – Signal to Noise Ratio

SON – Self-Organizing Network

SPS – Semi-Persistent Scheduling

SRVCC – Single-Radio Voice Call Continuity

SU-MIMO – Single User MIMO

SVDO – Simultaneous 1XRTT Voice and EV-DO Data

SVLTE – Simultaneous Voice and LTE

TCH – Traffic Channel

TCP/IP – Transmission Control Protocol/IP

TD – Transmit Diversity

TDD – Time Division Duplex

TDMA – Time Division Multiple Access

TD-SCDMA – Time Division Synchronous Code Division Multiple Access

TD-CDMA – Time Division Code Division Multiple Access

TETRA – Terrestrial Trunked Radio

TIA/EIA – Telecommunications Industry Association/Electronics Industry Association

TISPAN – Telecoms and Internet Converged Services and Protocols for Advanced Networks

TSG-RAN – Technical Services Group Radio Access Network

TTI – Transmission Time Interval

UAS – Uplink EGPRS2-A Level Scheme

UBS – Uplink EGPRS2-B Level Scheme

UE – User Equipment

UFMC – Universal Filtered Multi-Carrier

UICC – Universal Integrated Circuit Card

UL – Uplink

UMA – Unlicensed Mobile Access

UMB – Ultra Mobile Broadband

UMTS – Universal Mobile Telecommunications System

UPCON – User-Plane Congestion Management

UPF – User Plane Function

URA-PCH – UTRAN Registration Area Paging Channel

URI – Uniform Resource Identifier

URLLC – Ultra-Reliable and Low Latency Communications

us – Microsecond

USIM – UICC SIM

UTRAN – UMTS Terrestrial Radio Access Network

V2X – vehicle -to-infrastructure

VAMOS – Voice Services over Adaptive Multi-User Channels on One Slot

VDSL – Very-High-Bit-Rate DSL

VEPC – Virtualized EPC

ViLTE – Video Over LTE

VoIP – Voice over Internet Protocol

VoHSPA – Voice over HSPA

VoLGA – Voice over LTE Generic Access

VoLTE – Voice over LTE

VNF- Virtual Network Function
VPN – Virtual Private Network
WAP – Wireless Application Protocol
WBA – Wireless Broadband Alliance
WCDMA – Wideband Code Division Multiple Access
WCS – Wireless Communication Service
WebRTC – Web Real-Time Communication
Wi-Fi – Wireless Fidelity
WiMAX – Worldwide Interoperability for Microwave Access
WLAN – Wireless Local Area Network
WMAN – Wireless Metropolitan Area Network
WMM – Wi-Fi Multimedia
WRC – World Radiocommunication Conference

Additional Information

5G Americas maintains market information, LTE deployment lists, and numerous white papers, available for free download on its web site: http://www.5gamericas.org.

If there are any questions regarding the download of this information, please call +1 425 372 8922 or e-mail Anushka Bishen, Public Relations Coordinator at info@5gamericas.org.

This white paper was written for 5G Americas by Rysavy Research (http://www.rysavy.com) and utilized a composite of statistical information from multiple resources.

The contents of this paper reflect the independent research, analysis, and conclusions of Rysavy Research and may not necessarily represent the comprehensive opinions and individual viewpoints of each particular 5G Americas Board member company.

Rysavy Research provides this document and the information contained herein to you for informational purposes only. Rysavy Research provides this information solely on the basis that you will take responsibility for making your own assessments of the information.

Although Rysavy Research has exercised reasonable care in providing this information to you, Rysavy Research does not warrant that the information is error-free. Rysavy Research disclaims and in no event shall be liable for any losses or damages of any kind, whether direct, indirect, incidental, consequential, or punitive arising out of or in any way related to the use of the information.

Copyright © 2017 Rysavy Research, LLC. All rights reserved.

http://www.rysavy.com

Select Language ▼

Robert P. Duncan A.B., S.M., M.B.A., Ph.D.: Government warfare and surveillance system architect, author, and independent investigator..

Presents

ASK ME ANYTHING ON REDDIT. POST QUESTION&VOTE UP. INDEFINITE Q&A. ALSO LATEST CLASSIFIED INFO.

SEE YOU AT NOVEMBER 5th 2017 WHITE HOUSE MMM!!!! BE THERE!

Project Soul Catcher Vol. 2: Secrets of Cyber and Cybernetic Warfare Revealed (click book image to buy)

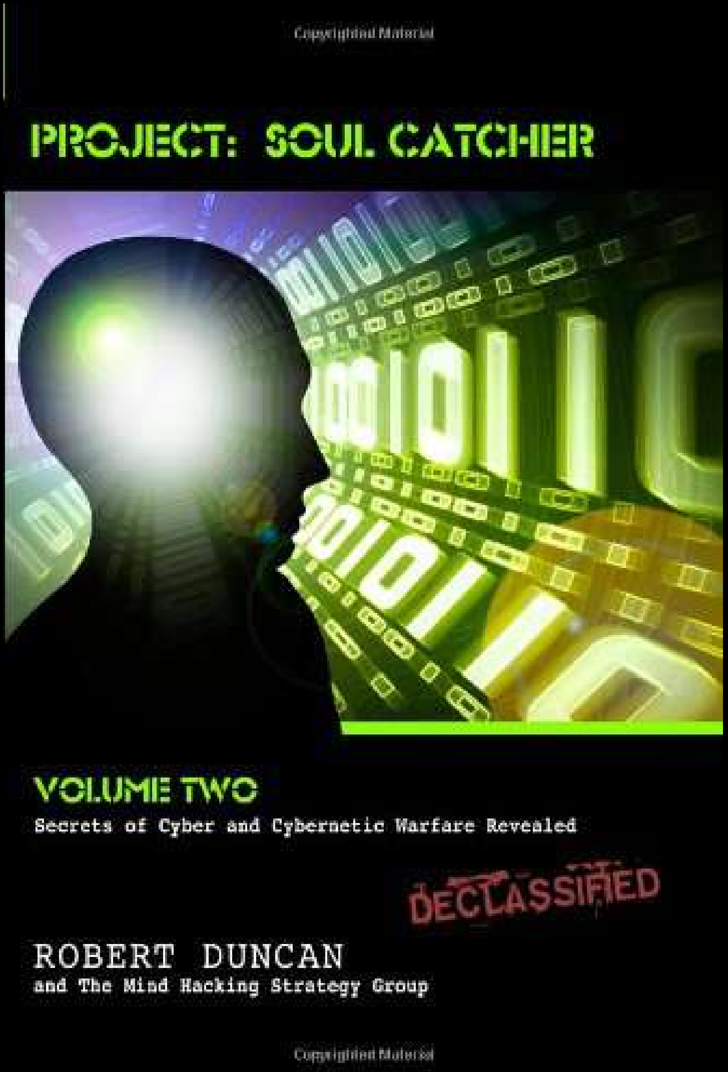
Copyrighted Material
PROJECT: SOUL CATCHER
VOLUME TWO
Secrets of Cyber and Cybernetic Warfare Revealed
DECLASSIFIED
ROBERT DUNCAN
and The Mind Hacking Strategy Group
Copyrighted Material

Volume 2 details the CIA's practices of interrogation and cybernetic mind control in their pursuit to weaponize neuropsychology. It covers the art of bio-communication war. Human beings are complex machines but their inner workings have been deciphered. Mind control and brainwashing have been perfected in the last 60 years. Hacking computers and hacking into individual minds are similar. The 21st century will be known as the age of spiritual machines and soulless men.

Check out Dr. Robert Duncan's other book, How to Tame a Demon: A short practical guide to organized intimidation stalking, electronic torture, and mind control.

More information on Dr. Robert Duncan is available at drrobertduncan.com.

new SSL secure chat #targets

DRROBERTDUNCAN.COM GUEST BOOK & FORUM. PROFESS YOUR LOVE, DROP A LINE, SHARE YOUR STORY, WHATEVER.

Free eBook: The Matrix Deciphered by Dr. Robert Duncan, A.B., S.M., M.B.A., Ph.D., the book discloses secret details on government operations and gives a hint to what it's like to live like a victim targeted..

Bio of Dr. Robert Duncan, from his book The Matrix Deciphered:

"True nobility is exempt from fear".
– King Henry the Sixth, Part II (Act IV, Scene I).

> Call me The Saint. I am the all American - prep school, Harvard College graduating with honors in computer science and a minor in premedical studies, and advanced degrees from Harvard and Dartmouth in business and science. My famous ancestors are President Lincoln, King Duncan of Scotland, and Governor William Bradford, the first governor of Massachusetts.
>
> My research interests have been neural networks, virtual reality, and EEG controlled robotics. Before graduate school I worked for the Department of Defense, Navy, NATO, and various intelligence agencies computer science projects. I have done business consulting and computer consulting for the largest companies in the world. I have been a professor, inventor, artist, and writer. I am one of the last Renaissance men.
>
> My projects have included algorithms for Echelon and CIA natural language parsing and classification of document content, IRS formula for red flagging audits, writing the artificial intelligence code to automate tracking of the Soviet Nuclear Submarine Fleet and all water vessels, work integrating HAARP with SIGINT SIGCOM and SPAWAR. I have worked on projects for the Justice Department connecting local, state, and federal databases for the tracking of terrorists. I developed a system for the FBI to track license plates past toll booths and other locations. I worked on the soldier 2000 program to create body networks for reading vital signs and other information. A system I worked on called Snyper is operational in Iraq which triangulates on intercity conflict gun shots. I have been to a couple secret bases in the so called free world. I have developed telemedicine robotic surgery and virtual reality applications for the Army. For DARPA, I have worked on satellite computer vision target tracking applications and tank simulation as well as integration of the land, sea, and air surveillance systems like SOSSUS, towed arrays, and others.
>
> Projects that I have worked on outside of government contracts include my thesis on computer generated holography, a project making paralyzed people walk again using choreographed stimulated muscles movements, face recognition, voice identification and recognition, finger print recognition, and neural network robotic controller. My research interests moved to enhanced reality heads up displays and wearable computing systems. My current research involves finding a cure for the mind control directed energy weapons fiasco. The integrated global surveillance grid is actually part of the holy grail of weapons and human control systems.
>
> My apologies to the human race for my contribution to tyranny. I was tricked into thinking it would not be misused by corrupt government especially in my beloved country. I was wrong. The Joint Non-Lethal Weapons Directorate has Skip Green on the governing board. One of my old colleagues at a technology think tank in Cambridge partly in charge of the radio frequency weapons testing for neurological disruption now torturing and killing people worldwide. Several other US Navy and UK Navy scientists have been knocked-off and that is why I have come forward. I know my time maybe near. I am currently a professor in computer science and business, helping to educate the public on government corruption, greed, and stupidity. Like my ancestor Lincoln, I am fighting against mental slavery in a new Civil War engineered by the same useless elements in over 80 government agencies who have tested radiological, drug, and viral weapons on unsuspecting Americans for over 45 years.
>
> I have great pride in the fundamental and constitutional values of America and mean no disrespect by my blasting the incompetence, apathy, and stupidity of those involved in the conspiracy of involuntary biological, chemical, and psychotronic human effects testing. There are so many brave men and women serving in the armed forces who fight and protect us for the American values they believe in. But because of the silos of information called compartmentalization in the security agencies and the lack of accountability and oversight of black operations and some top secret projects, diabolical elements become rogue or worse destroying the very country they are tasked with protecting turning it into a hypocracy rather than a democracy.
>
> My goal is to awaken Americans to the continued horrific acts of military and CIA weapons testers in this country and the other branches of government's inability to stop them and hold them accountable.
>
> All I ask from the reader is to listen to testimony and validate facts presented here and come to their own conclusion. Then fight to win back America from this silent overthrow. This has been my project for about two years, investigating government corruption, incompetence, and cover-up upon which I accidentally stumbled one day while looking at a reverse MRI scheme to inject electromagnetic signals into the brain for virtual reality applications. I have interviewed over two hundred people and worked on the highest level of military projects for the U.S., NATO, and U.K. and have given videotaped testimony to senators and representatives on this topic resulting in lip service since they have no real power to enforce. Two high level FBI agents and a couple CIA agents have come forward to validate the existence of a MKULTRA like project that continues to grab random people for mind control experimentation. Two of these have since become part of the program and endure daily psychotronic tortures. All the torture can be done using directed energy psychotronic weapons with the so called continental ballistic missile surveillance defense grids.
>
> I apologize to the human race for any contribution to these 4th generation weapons that I may have worked on that are more horrific than the nuclear bomb and whose cover-up is more pervasive than the Manhattan Project. And because of the horrific acts of violence being committed on as many as two thousand Americans as far as my research has uncovered and many others in other countries, I understand the extreme risk to my own welfare that publishing this material will have. But freedom is so important to me that I know full heartedly that the human race

must have an open discussion on these weapons to decide their own fate before the point of decision is gone, that I am willing to risk divulging so called national secrets. All I offer you is the truth.

All information presented in this book was received through legitimate channels such as the Freedom of Information Act, military documents, victim testimony, and turned agents. I still hold valid the oath I took to keep secret the details of the projects that I worked on under DoD budgets. The majority of the proceeds of this book will go towards helping the psychotronic experiment survivors and surviving families of those that have perished. When the government fails, business and the citizens must look out for each other.

Dr. Robert Duncan in Energy Weapons & Testing, Coast to Coast AM broadcast December 5th 2006 (2.75+ million listeners weekly, syndicated in most cities and areas of the United States, and other places around the world.)..

0:57 / 1:53:43

Download - Dr. Robert Duncan, Coast to Coast AM interview - Energy Weapons & Testing.

Energy Weapons & Testing synopsis:

Independent investigator Robert Duncan discussed directed energy and neurological weapons and his contention that they've been tested on the public at large. While directed energy is used in microwaves, to remove kidney stones, and in non-invasive surgery, it's also been developed extensively for military purposes, he reported.

The civilian population was targeted for experiments, in programs such as MK-ULTRA, starting after WW II, when Nazi scientists were imported to the US, some working on scalar or gravity weapons, said Duncan. The town of Taos, NM, where a hum was heard by many citizens, was the subject of a directed energy experiment by the U.S. Navy, he claimed. And most recently the Active Denial System (see article below) was tested on human subjects.

Duncan said he interviewed over 600 mind control victims (Dec 5th 2006 report), and found some validity to their allegations. There are weapons that can project voices into people's heads such as one system known as "The Voice of God," he detailed. Blocking techniques include jamming the signals with electronic scramblers, and using shields with metal alloys or mylar. He recommended the following websites for further information/assistance: mindjustice.org, raven1.net, freedomfchs.com.

The Goodbye Weapon

A new non-lethal weapon, the Air Force's Active Denial System, or ADS, has been certified for use in Iraq, after extensive testing. The ADS shoots a beam of waves that causes extreme (but temporary) pain and induces what experimenters call the "Goodbye effect," or "prompt and highly motivated escape behavior." Wired News obtained documents about the weapon and has published two reports (article(1)/documentation(2)). Additionally, the report's author David Hambling has posted a commentary(3) on the subject at Defense Tech.

1. http://www.wired.com/news/technology/0,72134-0.html
2. http://www.wired.com/news/technology/0,72236-0.html
3. http://www.defensetech.org/archives/003036.html

Coast to Coast AM guest bio:

Dr. Robert Duncan holds multiple degrees from Harvard University and Dartmouth College amongst others. He has had the most expensive American education money can buy. He is an investigator, author, and soon be movie producer on the topics of directed energy, neurological weapons, psychological, and information warfare. His movie is called "The Enemy Within - Psychic Warfare". A book he is co-authoring will be out in a few months called "Hacking the Human Mind".

Dr. Duncan has worked as a business and information technology consultant to the Fortune 500. He has worked for companies like Oracle Corporation, BEA systems, HP, BBN, and as a professor at a college. For the department of defense he has written the artificial intelligence code to track the Soviet nuclear submarine fleet with passive and active acoustical arrays and has been to a couple secret NATA Navy underground bases in Europe.

-Todd Giffen
Target since birth as I was born into American enslavement, surveillance, and control, under attack with directed energy since 2008
503-967-5202
case@oregonstatehospital.net
obamasweapon.com

Copyright © 2015 Todd Giffen

"The Use of Psychotronic Weapons"
"What everyone should know"

(courtesy : http://www.mindjustice.org/cherkova.htm)

Definition:

Definition of psychotronic (psycho-physical) weapons.

Psychotronic Weapons (PF- weapons) this is the totality of all possible methods and means (techno-genic, suggestive, pharmacological, paranormal, complexes, and others) of hidden, forced influences on the psyche of a person for the purpose of modifying his consciousness, behavior and health for what is desired in the way of influencing aspects of control..."

Second definition of psychotronic weapons (as an epigraph): "Applications in military and for antisocial purposes are possibilities also for knowledge of psychotronics, its means, methods, systems, 'generators', and there is the path for the transformation of the humanitarian, essential attributes of life and society in the weapons. These are also psychotronic weapons. This is not only dangerous, this is deadly!"

Types of Weapons:

- Electromagnetic Weapons
- Microwave Weapons
- Non-Lethal Weapons
- ELF (Extremely Low Frequency) Weapons
- Directed Energy Weapons
- Acoustic Weapons
- Psychotronic Weapons
- RF (Radio Frequency) Weapons
- Soft Kill Weapons
- Less-Than-Lethal Weapons

How it affects (Physical & psycho-physical damages):

Physical damage, visible results achieved at the moment of influence.

* burns, pricks, cuts, wounds (ray weapons, laser)
* bleeding into the brain (laser ray with energy pulses of 40 joules) and on the skin energy in pulses: 3 - 100 joules)
* Blindness in the eyes, loping of eyelashes (laser ray)
* loss of memory (influence of frequency of up to 100Hz. and a reduction to 5 Hz.)
* destruction of the vestibulary [inner ear] apparatus, spatial orientation
* symptoms of [***] [with a large exposure and strength, paralysis (UBCh) in the area of the beginning of the spinal column
* irregular sleeping, heart-beating
* numbness of the hands; at night, cramping of the legs
* sharp decrease in hemoglobin, micro-burns to the skin, ionization of the blood and body fluids, an increase in the deterioration rate of clothing material, photographic material [illuminating], (thermal-electronic emissions from lamps in the room with the feeding units of electrical networks and fittings of the wall, [...]
* vibrations (separate organs can vibrate: the heart, kidneys, etc.), [...] (from

* vibrational technologies)
* pressure, coolness of the skin
* feelings of non-specific fear and panic (from the weapons) [...] tremors, tapping, destruction of fragile objects from the waves, from the impulses and from other manifestations of technological poltergeist

Psycho-physical damage depending on intensity and duration of effects. Damage in the physical sphere (up to programmed premature death).

Stage I.

* changes in the blood (early signs)
* overheating substances especially and increase in crystallization, leading to cataracts (early sign)
* ill reactions along the path of the bio-energetic potential from the effects [...]
* prickly sensations, tremors to the muscles of the feet
* aggravation of chronic illnesses
* internal bleeding (nose, stomach, gynecological)
* auditory tones of various frequencies
* sharp pain in the internal ear
* dull pains in the area behind the ears
* internal trembling, vibrations
* head pain in the temple and a warmness in that area, pain in the eye-sockets
* likelihood of miscarriage, an increase of 80% for chances of premature births, anomalous
* development of the fetus, often with defective brains

Stage II.

* "radio-sound" effect (before this is long-conducted the resonance-loosening of spinal and brain fluids at the frequency of 2-10 Hz. for maintaining the fluids of theskull and brain in stimulation
* steady reactions to all aspects of coding
* lowering of energy (observed with irradiation at the frequency carrying frequencies of ~780 Hz.and amplitude modulation ~71 Hz.)
* aggravation of dermatosis
* eye pains (cataracts, asymmetry in lenses of the eyes, puffiness [?] of the retinas)
* aggravation of the course of chronic illnesses
* destruction of the functions of the kidneys, liver, organs of digestion, circulatory system, bleeding in gingivae, gums, peridontal and others
* gauntness
* "sand" in the eyes, burning sensation, loss of hair, brittleness of nails
* compression pain in the heart fits of coughing
* changes of the color of the face (bronze shading of the skin)
* asymmetrical puffiness of the face
* disfiguration of the face, the intentional emaciation of separate muscles, cartilage
* syndromes of destruction to the external nerves of the thighs
* impotence

Stage III.

* destruction of the brain, irreversible changes to the structure of the brain (academician V.P. Vekhtereva)

* irreversible processes in the muscle material, in the bones of the skull, and in the structures of neurons "..minimalization of life" --external reaction to EMP appears as an illness, up to heavy, sharp, chronic, organic pathological processes (academician V.P. Kaznacheev)
* "..the entire speeding up of the aging process, sclerosis, oncogenesis (academician V.P. Kaznacheev)
* destruction of bio-chemical processes in the cells with distrophic changes in the organs (academician V.P. Kaznacheev)
* destruction of the cell membranes, chemical connections in the cells -leading to leukemia, cancer of the brain and prostate gland, breast cancer, (even with weak EMP, but of long duration "...the genetic apparatus is altered..." (academician V.P. Kaznacheev)
* the reflex activities are destroyed
* the functioning of the hormonal system is destroyed
* [....] Parkinson's disease (trembling of the hands) and Alzheimer's disease (from changes in the production of the hormone melanin ([...] gland) that causes degenerative changes in substances at the molecular level disfunctions of the sino-atrial node of the heart (even with a low frequency)
* infarction, cerebral thrombosis, osteo-[***], peridontal, sugar diabetes
* allergies
* epilepsy (especially caused by the multiple-rapid stimulation by a weak, low frequency waves)
* hypertension, hypotension
* programmed or instantaneous (with the subjection to extreme effects -up to the border of deadly conditions) premature death --it is well known by the experimentors that cases of death do occur (academician V.P. Kaznacheev)
* a decrease and dying off of the population, humanity, mankind

Damage in the mental sphere (up to the point of insanity).
Stage I.

* manifestations of depression
* lowering in the interest in perceived information
* shortness of temper (irritability)

Stage II.

* steady depression
* lowering of abilities for analytical thinking, some worsening of the memory
* symptoms of chronic fatigue
* superficial nature of sleep
* higher incidents of conflicts

Stage III.

* confusion in the accounting of events in real life
* loss of ability for analytical thinking
* lowering and a loss of memory
* feelings of the lack of sleep
* indifference to personal external appearance [...]
* lowering of self-control from the deficit of the hormone seratonin due to SVCh irradiation
* suicide (as the experimentors "practice" provocation techniques)
* "...invisible weak fields ...their ignoring that they are carrying with them a deadly danger"... "...people simply start going crazy and nobody understands why..." (academician V.P. Kaznacheev)

"Consequences of effects of micro-leptonic generators are described by the leading developer in this area [from the] V. Lab. of Micro-leptonic Technologies D. Okhatryym: in 1982 there began extremely unpleasant things with the health our co-workers. At first, there arose their feeling unwell, then the loss of logic, then after this, the loss of spatial orientation. Finally, a breakdown of diseased [affected] organs. The coming out of this condition -and this includes myself -was achieved with great difficulty."

(in the book by Ju. Vorob'evskii" Knock at the Golden Gate, Moscow, 1999.

Manipulation with the memory.

* Erasure
* Blockage
* Recording of information needed by the experimentor
* training -"prompting" (after erasure)
* comparison,
* reproduction of picture a from a prior testing as though bifurcated [?]
* a shifting of the effective frequency up to 100 Hz. or a lowering to 5 Hz., with memory loss

Control of Stimulation in the brain.

* by states of consciousness (up to the loss of these)
* in sleep
* emotions
* intellectual capabilities
* attention
* speech
* activeness
* muscle tone
* stimulation of one and the same zone with stimulations of:
o 30 -40 mV- activization of attention, memory, movement;
o 60-70 mV- "emotiono-genic" effects

Different methods:

Types of Psychotronic, psychotropic and other Psy-Influences.

* radio-waves
* laser-irradiation
* infra-red irradiation
* ultra-violet irradiation UF
* x-ray irradiation
* gravitational fields
* torsional fields (hypothetical)
* leptonic and other fields

Acoustic Waves.

* infra-sonic (IZ) from 0 Hz. to 15-20 KHz.
* ultra-sonic(UZ) from 15-20 KHz. to 10**9 Hz. (ranges for IZ and UZ)

see more similar information in -Table 3

Formation of fields:

* electro-magnetic charge
* gravitational-mass
* torsional (surrounding spin fields) -spin, with any rotating body
* micro-leptonic energy (in free aspect) -[over-abundance of stable particles]?

Operator is secretly brought in, or in a secret installation nearby

BG: Biogenerator (could be tomography)
A: Antenna (irradiation and reception)
D: operator for the gathering of information, transmission and control
ZU: reminding sytem on a magnetic strip, holographic
BIP: bio-informational [...]
KC: channel connections (IK, radio, telephone, telegraph)
E: screen display
O: operator
SP: light pen with which the operator indicates on the screen, the point of influence

Stress systems and methods of harassment.

* slander
* gestures, hand movements, expressions, tests, practical "jokes"
* taking off what is needed for the eyes (vision) [?]
* psychological pressure through family, neighbors or surrounding people, vehicle drivers
* discrediting: along 5 directions (often with the involvement of the police)
o irresponsibility
o spy
o drug dealer
o vagrant
o accomplice in a crime (in an alleged crime)
* torture, assault
* blackmail, threats
* attempts on their lives with PTOs
* death to someone close or in immediate surrounding life
* murder

Stages of Zombification (Substitution of consciousness with external control).

* programming, modeling of a cybernetic double
* conducting of programs according to non-differentiation of a proper and artificial "I"-"Not-I"
* extreme influencing, stress, harassment, irradiation, special-preparations [substances], chemicals
* taking to the limits of mental and physical capacities (survivability)
* loss of will, memory
* shutting off of his own consciousness
* replacement of consciousness with the control from his cybernetic twin

IDEAS
IN ORBIT
EXPLORE THE TECH
Raytheon

Did You Miss It? Get Breaking Defense Coverage Of SNA Symposium Here →

ACQUISITION, **AIR**, **INTEL & CYBER**

Pentagon Studies Weapons That Can Read Users' Mind

By SYDNEY J. FREEDBERG JR.
on July 14, 2017 at 3:15 PM
15 Comments

28
SHARES

NEWSEUM: The troops of tomorrow may be able to pull the trigger using only their minds. As artificially intelligent drones, hacking, jamming, and missiles accelerate the pace of combat, some of the military's leading scientists are studying how mere humans can keep up with the incredible speed of cyber warfare, missiles and other threats.

One option: Bypass crude physical controls — triggers, throttles, keyboards — and plug the computer directly into the human brain. In one DARPA experiment, a quadriplegic first controlled an artificial limb and then flew a flight simulator. Future systems might monitor the users' nervous system and compensate for stress, fatigue, or injury. Is this the path to what the Pentagon calls human-machine teaming?

This is an unnerving scenario for those humans, like Stephen Hawking, who mistrust artificial intelligence. If your nightmare scenario is robots getting out of control, "let's teach them to read our minds!" is probably *not* your preferred solution. It sounds more like the beginning of a movie where cyborg Arnold Schwarzenegger goes back in time to kill someone.

Navy X-47B drone.

But the Pentagon officials who talked up this research yesterday at Defense One's annual tech conference emphasized the objective was to *improve* human control over artificial intelligence. Teaching AI to monitor its user's level of stress, exhaustion, distraction, and so on helps the machine adapt itself to better serve the human — instead of the other way around. Teaching AI to instantly detect its user's *intention* to give a command, instead of requiring a relatively laborious push of a button, helps the human keep control — instead of having to let the AI off the leash because no human can keep up with it.

Official Defense Department policy, as then-Secretary Ash Carter put it, is that the US will "never" allow an artificial intelligence to decide for itself whether or not to kill a human being. However, no less a figure than the Carter's undersecretary of acquisition and technology, Frank Kendall, fretted publicly that making our robots wait for human permission would slow them down so much that enemy AI without such constraints would beat us. Vice-Chairman of the Joint Chiefs, Gen. Paul Selva, calls this the "Terminator Conundrum." Neuroscience suggests a way out of this dilemma: Instead of slowing the AIs down, make the humans' orders come faster.

DARPA's Revolutionizing Prosthetics program is devising new kinds of artificial limbs — and new ways to control them.

Accelerate Humanity

"We will continue to have humans on the loop, we will have human input in decisions, but the way we go about that is going to have to shift, just to cope with the speed and the capabilities that autonomous systems bring," said Dr. James Christensen, portfolio manager at the Air Force Research Laboratory's 711th Human Performance Wing. "The decision cycle with these systems is going to be so fast that they have to be sensitive to and responsive to the state of the individual (operator's) *intent*, as much as overt actions and control inputs that human's providing."

In other words, instead of the weapon system responding to the human operator *physically* touching a control, have it respond to the human's brain cells forming the *intention* to use a control. "When you start to have a direct neural interface of this type, you don't necessarily need to command and control the aircraft using the stick," said Justin Sanchez, director of DARPA's Biological Technologies Office. "You could potentially re-map your neural signatures onto the different control surfaces" — the tail, the flaps — "or maybe any other part of the aircraft" — say landing gear or weapons. "That part hasn't really been explored in a huge amount of depth yet."

Reading minds, even in this limited fashion, will require deep understanding and close monitoring of the brain, where thoughts take measurable form as electrical impulses running from neuron to neuron. "Can we develop precise neurotechnologies that can go to those circuits in the brain or the peripheral nervous system in real time?" Sanchez asked aloud. "Do we have computational systems that allow us to understand what the changes in those signals (mean)? And can we give meaningful feedback, either to the person or to the machine to help them to do their job better?"

DARPA's Revolutionizing Prosthetics program hooked up the brain of a quadriplegic — someone who could move neither arms nor legs — to a prosthetic arm, allowing the patient to control it directly with their thoughts. Then, "they said, 'I'd like to try to fly an airplane,'" Sanchez recounted. "So we created a virtual flight simulator for this person, allowed this neural interface to interface with the virtual aircraft, and that person flew."

"That was a real wake-up call for everybody involved," Sanchez said. "We didn't initially think you could do that."

May
21
19:40
100%
21 Tuesday May 2013

Tony Stark (Robert Downey Jr.) relies on the JARVIS artificial intelligence — exquisitely adapted to his personal strengths and weaknesses — to help pilot his Iron Man suit. (Marvel Comics/Paramount Pictures)

Adapting To The Human

Applying direct neural control to real aircraft — or tanks, or ships, or network cybersecurity systems — will require a fundamental change in design philosophy. Today, said Christensen, we give the pilots tremendous information on the aircraft, then expect them to adapt to it. In the future, we could give the aircraft tremendous information on its pilots, then have it use artificial intelligence to adapt itself to *them*. The AI could customize the displays, the controls, even the mix of tasks it took on versus those it left up to the humans — all exquisitely tailored not just to the preferences of the individual operator but to his or her current mental and physical state.

When we build planes today, "they're incredible sensor platforms that collect data on the world around them and on the aircraft systems themselves, (but) at this point, very little data is being collected on the pilot," Christensen said. "The opportunity there with the technology we're trying to build now is to provide a continuous monitoring and assessment capability so that the aircraft knows the state of that individual. Are they awake, alert, conscious, fully capable of performing their role as part of this man-machine system? Are there things that the aircraft then can do? Can it unload gees (i.e. reduce g-forces)? Can it reduce the strain on the pilot?"

Air Force MQ-9 Reaper operators in training. Today's drones require extensive human oversight and are hardly user friendly.

"This kind of ability to sense and understand to the state and the capabilities of the human is absolutely critical to the successful employment of highly automated systems," Christensen said. "The way all of our systems are architected right now, they're fixed, they're predictable, they're deterministic" — that is, any given input always produces the exact same output.

Predictability has its advantages: "We can train to that, they behave in very consistent ways, it's easier to test and evaluate," Christensen said. "What we lose in that, though is the real power of highly automated systems, autonomous systems, as *learning systems*, of being able to adapt themselves. "

"That adaptation, though, it creates unpredictability," he continued. "So the human has to adapt alongside the system, and in order to do that, there has to be some mutual awareness, right, so the human has to understand what is the system doing, what is it trying to do, why is that happening; and vice versa, the system has to has some understanding of the human's intent and also their state and capabilities."

This kind of synergy between human and artificial intelligence is what some theorists refer to as the "centaur," after the mythical creature that combined human and horse — with the human part firmly in control, but benefiting from the beast's strength and speed. The centaur concept, in turn, lies at the heart of the Pentagon's ideas of human-machine teaming and what's known as the Third Offset, which seeks to counter (offset) adversaries' advancing technology with revolutionary uses of AI.

The neuroscience here is in its infancy. But it holds the promise of a happy medium between hamstringing our robots with too-close control or letting them run rampant.

28
SHARES

Topics: AFRL, air force, Air Force Research Laboratory AFRL, artificial intelligence, darpa, DefenseOne Tech Summit 2017, James Christensen, Justin Sanchez, neuroscience, offset strategy, War Algorithm

SPONSORED CONTENT
Raytheon's trusted missile defense solutions help the U.S. and its allies make the world a safer place

Breaking
Defense
Click here to view gallery »

EDITORIAL STAFF

Colin Clark
Editor

Sydney J. Freedberg, Jr.
Deputy Editor

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/313112342

Internet of Things (IoT) in 5G Wireless Communications

Article *in* IEEE Access · January 2016

DOI: 10.1109/ACCESS.2016.2646120

CITATIONS
2

READS
680

7 authors, including:

Waleed Ejaz
Ryerson University

69 PUBLICATIONS **397** CITATIONS

SEE PROFILE

Muhammad Ali Imran
University of Glasgow

402 PUBLICATIONS **4,102** CITATIONS

SEE PROFILE

Minho Jo
Korea University

98 PUBLICATIONS **1,145** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

IU-ATC View project

Visible Light Communication View project

All content following this page was uploaded by Minho Jo on 27 March 2017.

The user has requested enhancement of the downloaded file.

Digital Object Identifier 10.1109/ACCESS.2016.2646120

GUEST EDITORIAL

Internet of Things (IoT) in 5G Wireless Communications

During the past decade, the Internet of Things (IoT) has revolutionized the ubiquitous computing with multitude of applications built around various types of sensors. A vast amount of activity is seen in IoT based product-lines and this activity is expected to grow in years to come with projections as high as billions of devices with on average 6-7 devices per person by year 2020. With most of the issues at device and protocol levels solved during the past decade, there is now a growing trend in integration of sensors and sensor based systems with cyber physical systems and device-to-device (D2D) communications. 5th generation wireless systems (5G) are on the horizon and IoT is taking the center stage as devices are expected to form a major portion of this 5G network paradigm. IoT technologies such as machine to machine communication complemented with intelligent data analytics are expected to drastically change landscape of various industries. The emergence of cloud computing and its extension to fog paradigm with proliferation of intelligent 'smart' devices is expected to lead further innovation in IoT. These developments excite us and form a motivation to survey existing work, design new techniques, and identify new applications of IoT. Researchers, scientists, and engineers face emerging challenges in designing IoT based systems that can efficiently be integrated with the 5G wireless communications.

We received enthusiastic response to our special issue call for papers. A total of nine high quality papers were received out of which only seven were selected after a thorough review process. Invited articles were sought from two highly cited and accomplished researchers. Dr. Mischa Dohler (Fellow IEEE, Kings College London, UK) and Dr. Mung Chiang (Fellow IEEE, Princeton University, USA).

The first article ''MIMO-NOMA design for small packet transmission in the Internet of Things'', written by a collaborative team of researchers Ding et al.,, opportunistic serving mechanism is designed as part of the effort in the novel paradigm of Multiple Input Multiple Output Non-Orthogonal Multiple Access transmission scheme. Under the proposed method, one user is completely served with their quality of service requirements completely taken care of whereas second user is served opportunistically under the NOMA paradigm. The main contribution of this article is a design with two sets of system parameters, precoding and power allocation coefficients, in order to ensure that the potential of NOMA can be realized even if the users' channel conditions are similar. Two types of power allocation policies are developed in this paper. One is to meet first user's QoS requirements in the long term. For example, in order to satisfy its targeted outage probability. The other is in which to realize second user's QoS requirements instantaneously, e.g., the power allocation coefficients are designed to realize its targeted data rate for each channel realization.

Machine to machine communication has a significant role to play in emerging internet of things paradigm in years and decades to come. The emerging IoT-5G scenario extends sensor based IoT capabilities to robots, actuators and drones for distributed coordination and low-latency reliable execution of tasks at hand. In the invited work titled ''Enabling the IoT machine age with 5G: Machine-type multicast services for innovative real-time applications'' by Condoluci et al., core attention is focused on the end-to-end reliability, latency, and energy consumption comprising both up and downlinks for 5G-IoT communication. The authors propose the definition, design, and analysis of machine-type multicast service (MTMS). They recommend different procedures that need to be redesigned for MTMS and derive the most appropriate design drivers by analyzing different performance indicators, such as scalability, reliability, latency, and energy consumption. Overall, a very interesting read complemented by open problems and future research directions to pursue.

Security is one of the biggest challenges faced by Internet of Things. With devices becoming ubiquitous and pervasive in day to day lives necessitate reliable and secure algorithms. The third article, ''Security enhancement for IoT communications exposed to eavesdroppers with uncertain locations'' by Xu et al., develops a secure framework for eavesdroppers with Uncertain Locations in IoT. With the assumption that the locations of eavesdroppers change independently from hop to hop, authors derive an expression for the secrecy outage probability of the two-hop transmission, which is shown to be the upper bound of the outage probability when the locations of eavesdroppers remain unchanged. Following this expression, the end users formulate a secrecy rate maximization problem with the secrecy-outage probability constraint. The optimal rate design for codebooks and power allocation between the source and relay are derived. By studying the performance of the optimal scheme in some special cases, we obtain several insights concerning the setting of system parameters.

In the article, ''Enabling massive IoT in 5G and beyond systems: PHY radio frame design considerations'' by Ayesha Ijaz et al., the authors propose a flexible frame structure and design for massive Internet of Things (IoT) devices working in 5G wireless network. The authors also discussed

2169-3536 © 2017 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

the interdependence of different frame design parameters, service requirements and characteristics of radio environment. Based on these interdependency, they provide guidelines for radio numerology design and elaborated on the frame design for IoT communications in 5G networks to support massive connection density of low-rate, low-power devices. The article concludes with some key research findings and challenges massive IoT in 5G wireless network.

It is estimated that in year 2020, 20 to 40 billion devices will be connected to the Internet as part of the Internet of Things. A critical bottleneck for realizing the efficient IoT is the pressure it puts on the existing communication infrastructures, requiring transfer of enormous data volumes. In the article ''CONDENSE: A reconfigurable knowledge acquisition architecture for future 5G IoT'', by Dejan Vukobratovic et al., the authors propose a architecture named 'Condense' which integrates the acquisition of IoT-generated data within the 3GPP MTC (machine type communications) systems. The proposed Condense architecture introduces a service within 3GPP MTC systems – computing linear and non-linear functions over the data generated by MTC devices This service brings about the possibility that the underlying communication infrastructure communicates only the desired function of the MTC- generated data (as required by the given application at hand), and not the raw data in its entirety. This transformational approach has the potential to dramatically reduce the pressure on the 3GPP MTC communication infrastructure. The article concludes by discussing challenges, provides insights, and identifies future research directions for implementing function computation and function decomposition within practical 3GPP MTC systems.

In the article, ''Frequency-domain oversampling for cognitive CDMA systems: Enabling robust and massive multiple access for Internet of Things'' by Su Hu et al., the authors utilize the concept of cognitive radio with dynamic non-continuous spectrum bands and code division multiple access to tackle the challenge of massive spectrum resource management in IoT. In order to suppress multiple access interference resulting from the non-orthogonality of partial available spectrum bins, carrier frequency offset, and spectrum sensing mismatch, the authors propose an enhanced receiver design that combines the frequency-domain oversampling scheme (FDO) and linear minimum mean square error (MMSE) method. The simulation results show that the cognitive-CDMA with FDO-MMSE receiver outperforms that with conventional per-user MMSE receiver in the presence of multipath fading channels, carrier frequency offset, and spectrum sensing mismatch.

In the invited article, ''A survey of client-controlled HetNets for 5G'' by Michael Wang et al., a comprehensive review is provided on spectrum of client-controlled HetNets for 5G networks: from the fully devolved distributed local control approach to the hybrid control approach where clients may make the decisions given some global information provided by the network. After giving a thorough review, the authors also provide future research directions and recommendations for evolution of 5G heterogeneous networks as an enabler for next generation internet of things.

To conclude, the Editors would like to thank all authors who submitted their manuscripts against this Special issue. There had been some very exciting submissions though we could only accept best seven high quality submissions. The editorial staff also thanks the reviewers for their timely comments on the manuscript which helped us improve the presentation and quality of the content manifold. All the associate editors thank the Editor in Chief Dr. Michael Pecht for his assistance in making this special issue a reality. We would also like to thank the publications editor Ms. Kimberly Shumard for her continuous follow up on timelines and management of articles to ensure timely completion of the review process. We complete this note by letting the readers know to stay tuned with IEEE Access as we expect even more exciting issues in months to come.

WALEED EJAZ
Ryerson University,
Toronto, ON, Canada

ALAGAN ANPALAGAN
Ryerson University,
Toronto, ON, Canada

MUHAMMAD ALI IMRAN
University of Glasgow,
Glasgow, U.K.

MINHO JO
Korea University,
Sejong, South Korea

MUHAMMAD NAEEM
Ryerson University,
Toronto, ON, Canada

SAAD BIN QAISAR
National University of Sciences and Technology,
Islamabad, Pakistan

WEI WANG
Zhejiang University,
Hangzhou, China

WALEED EJAZ (S'12–M'14–SM'16) received the B.Sc. and M.Sc. degrees in computer engineering from the National University of Sciences and Technology, Islamabad, Pakistan, and the University of Engineering and Technology, Taxila, Pakistan, respectively, and the Ph.D. degree in information and communication engineering from Sejong University, South Korea, in 2014. He was with top engineering universities in Pakistan and Saudi Arabia as a Faculty Member. He was a Post-Doctoral Fellow with Queen's University, Kingston, Canada. He is currently a Senior Research Associate with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada. He also completed certificate courses on teaching and learning in higher education from the Chang School, Ryerson University. His current research interests include Internet of Things, energy harvesting, 5G cellular networks, and mobile cloud computing. He is currently serving as an Associate Editor of the *Canadian Journal of Electrical and Computer Engineering* and the IEEE Access. In addition, he is handling the special issues in *IET Communications*, the IEEE Access, and the *Journal of Internet Technology*.

ALAGAN ANPALAGAN (S'98–M'01–SM'04) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of Toronto. He is currently a Professor with the Department of Electrical and Computer Engineering, Ryerson University, where he directs the research group involved in radio resource management and radio access and networking areas within the WINCORE Lab. He served as an Editor of the IEEE Communications Surveys and Tutorials from 2012 to 2014, the IEEE Communications Letters from 2010 to 2013, the *Wireless Personal Communications* (Springer) from 2011 to 2013, and the EURASIP *Journal of Wireless Communications and Networking* from 2004 to 2009. He has co-authored three edited books, *Design and Deployment of Small Cell Networks* (Cambridge University Press, 2014), *Routing in Opportunistic Networks* (Springer, 2013), and *Handbook on Green Information and Communication Systems* (Academic Press, 2012).

Dr. Anpalagan is a registered Professional Engineer in ON, Canada, and a fellow of the Institution of Engineering and Technology. He served as the TPC Co-Chair of the IEEE Globecom'15: SAC Green Communication and Computing, the IEEE WPMC'12 Wireless Networks, the IEEE PIMRC'11 Cognitive Radio and Spectrum Management, and the IEEE CCECE'04/08. He served as the Vice Chair of the IEEE SIG on Green and Sustainable Networking and Computing with Cognition and Cooperation (2015–), the IEEE Canada Central Area Chair (2012–2014), the IEEE Toronto Section Chair (2006–2007), the ComSoc Toronto Chapter Chair (2004–2005), and the IEEE Canada Professional Activities Committee Chair (2009–2011). He was a recipient of the Dean's Teaching Award in 2011, the Faculty Scholastic, Research, and Creativity Award in 2010 and 2014, the Faculty Service Award in Ryerson University in 2011 and 2013. He also received the Exemplary Editor Award from the IEEE ComSoc in 2013 and an Editor-in-Chief Top10 Choice Award in the IEEE Transactions on Emerging Telecommunications Technology in 2012.

MUHAMMAD ALI IMRAN (M'03–SM'12) received the M.Sc. (Hons.) and Ph.D. degrees from the Imperial College London, U.K., in 2002 and 2007, respectively. He is currently a Professor in Communication Systems with the University of Glasgow, the Vice Dean of the Glasgow College UESTC, and the Program Director of Electrical and Electronics with Communications. From 2007 to 2016, he was an Adjunct Associate Professor with The University of Oklahoma, USA, and a Visiting Professor with the 5G Innovation Centre, University of Surrey, U.K.

He has led a number of multimillion-funded international research projects encompassing the areas of energy efficiency, fundamental performance limits, sensor networks, and self-organizing cellular networks. He also led the new physical layer work area for 5G Innovation center at Surrey. He has a global collaborative research network spanning both academia and key industrial players in the field of wireless communications. He has supervised 21 successful Ph.D. graduates and published over 200 peer-reviewed research papers, including over 20 IEEE Transaction papers. He secured first rank in his B.Sc. and a distinction in his M.Sc. degree along with an award of excellence in recognition of his academic achievements conferred by the President of Pakistan. He received the IEEE Comsoc's Fred Ellersick Award 2014, the Sentinel of Science Award 2016, and the FEPS Learning and Teaching Award 2014 and twice nominated for Tony Jean's Inspirational Teaching Award. He is a shortlisted finalist for The Wharton-QS Stars Awards 2014, the Reimagine Education Award for innovative teaching, and the VC's learning and teaching award in the University of Surrey. He is a Senior Fellow of the Higher Education Academy, U.K.

Dr. Imran has given an invited TEDx talk (2015) and over ten plenary talks, several tutorials and seminars in international conferences and other institutions. He has taught international short courses in USA and China. He is the Co-Founder of the IEEE Workshop BackNets 2015 and chaired several tracks/workshops of international conferences. He is an Associate Editor of the IEEE Communications Letters, the IEEE Access, and the *IET Communications* journal.

MINHO JO (SM'16) received the B.A. degree from the Department of Industrial Engineering, Chosun University, South Korea, in 1984, and the Ph.D. degree from the Department of Industrial and Systems Engineering, Lehigh University, USA, in 1994. He is currently a Professor with the Department of Computer Convergence Software, Korea University, Sejong, South Korea. He is one of the founders of the Samsung Electronics LCD Division. He was the Vice President of the Institute of Electronics and Information Engineers. He is currently the Vice President of the Korean Society for Internet Information. His current research interests include Internet of Things, AI/big data in Internet of Things, LET-Unlicensed, cognitive radio, mobile cloud computing, network security, heterogeneous networks in 5G, wireless energy harvesting, optimization in wireless communications, and massive MIMO. He is the Founder and an Editor-in-Chief of the *KSII Transactions on Internet and Information Systems* [SCI (ISI) and SCOPUS]. He served as an Editor of the IEEE Network. He is an Editor of the IEEE Wireless Communications, and an Associate Editor of the IEEE Internet of Things Journal, *Security and Communication Networks*, *Ad-Hoc & Sensor Wireless Networks*, and *Wireless Communications and Mobile Computing*.

MUHAMMAD NAEEM received the B.S. and M.S. degrees in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2000 and 2005, respectively, and the Ph.D. degree from Simon Fraser University, BC, Canada, in 2011. From 2000 to 2005, he was a Senior Design Engineer with Comcept (pvt.) Ltd. At the design department of Comcept (pvt) Ltd., he participated in the design and development of smart card-based GSM and CDMA pay phones. From 2012 to 2013, he was a Post-Doctoral Research Associate with the WINCORE Laboratory, Ryerson University, Toronto, ON, Canada. Since 2013, he has been an Assistant Professor with the Department of Electrical Engineering, COMSATS Institute of IT, Wah Campus, Pakistan, and a Research Associate with the WINCORE Laboratory, Ryerson University. He is also a Microsoft Certified Solution Developer. His research interests include optimization of wireless communication systems, nonconvex optimization, resource allocation in cognitive radio networks, and approximation algorithms for mixed integer programming in communication systems. He was a recipient of NSERC CGS Scholarship.

SAAD BIN QAISAR (S'04–M'09–SM'14) received the master's and Ph.D. degrees in electrical engineering from Michigan State University, USA, in 2005 and 2009, respectively. He is currently serving as an Associate Professor with the National University of Sciences and Technology, Pakistan, and also leading Bitsym, an Internet of Things and data-analytics enterprise. He has authored over 60 publications, inventor of various patents, has completed multiple national and international projects and winner of various awards, including the University Best Researcher Award in 2012 and the UNIDO Cleantech Award in 2015, to name a few. His company, Bitsym, has been named among the 50 Most Promising Start-ups for 2012 by Global Entrepreneurship Week, globally.

WEI WANG (S'08–M'10–SM'15) received the B.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2004 and 2009 respectively. From 2007 to 2008, he was a Visiting Student with the University of Michigan, Ann Arbor, USA. From 2013 to 2015, he was a Hong Kong Scholar with The Hong Kong University of Science and Technology, Hong Kong. He is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He has authored over 90 international journal and conference papers in the area of wireless communications and networks. His research interests mainly focus on cognitive radio networks, green communications, and radio resource allocation for wireless networks.

Dr. Wang is the Associate Secretary-General of the Zhejiang Institute of Communications. He is the Editor of the book *Cognitive Radio Systems* (Intech, 2009) and serves as an Editor of the IEEE ACCESS, TRANSACTIONS ON EMERGING TELECOMMUNICATIONS TECHNOLOGIES, and the *KSII Transactions on Internet and Information Systems*. He also serves as the Co-Chair or a TPC Member for several international conferences.

• • •

View publication stats

ERICSSON WHITE PAPER
Uen 284 23-3204 Rev C | April 2016

5G RADIO ACCESS

CAPABILITIES AND TECHNOLOGIES

The capabilities of 5G wireless access must extend far beyond previous generations of mobile communication. Examples of these capabilities include very high data rates, very low latency, ultra-high reliability, energy efficiency and extreme device densities, and will be realized by the development of LTE in combination with new radio-access technologies. Key technology components include extension to higher frequency bands, access/backhaul integration, device-to-device communication, flexible duplex, flexible spectrum usage, multi-antenna transmission, ultra-lean design, and user/control separation.

WHAT IS 5G?

5G radio access technology will be a key component of the Networked Society. It will address high traffic growth and increasing demand for high-bandwidth connectivity. It will also support massive numbers of connected devices and meet the real-time, high-reliability communication needs of mission-critical applications.

5G will provide wireless connectivity for a wide range of new applications and use cases, including wearables, smart homes, traffic safety/control, critical infrastructure, industry processes and very-high-speed media delivery. As a result, it will also accelerate the development of the Internet of Things.

The overall aim of 5G is to provide ubiquitous connectivity for any kind of device and any kind of application that may benefit from being connected.

5G networks will not be based on one specific radio-access technology. Rather, 5G is a portfolio of access and connectivity solutions addressing the demands and requirements of mobile communication beyond 2020.

The specification of 5G will include the development of a new flexible air interface, NX, which will be directed to extreme mobile broadband deployments. NX will also target high-bandwidth and high-traffic-usage scenarios, as well as new scenarios that involve mission-critical and real-time communications with extreme requirements in terms of latency and reliability.

In parallel, the development of Narrow-Band IoT (NB-IoT) in 3GPP is expected to support massive machine connectivity in wide area applications. NB-IoT will most likely be deployed in bands below 2GHz and will provide high capacity and deep coverage for enormous numbers of connected devices.

Ensuring interoperability with past generations of mobile communications has been a key principle of the ICT industry since the development of GSM and later wireless technologies within the 3GPP family of standards. In a similar manner, LTE will evolve in a way that recognizes its role in providing excellent coverage for mobile users, and 5G networks will incorporate LTE access (based on Orthogonal Frequency Division Multiplexing (OFDM)) along with new air interfaces in a transparent manner toward both the service layer and users.

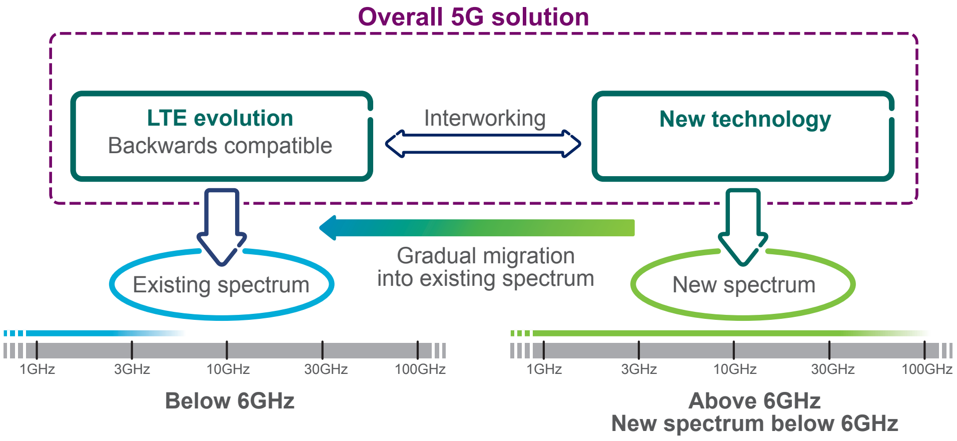

Figure 1: The overall 5G wireless-access solution consisting of LTE evolution and new technology.

Around 2020, much of the available wireless coverage will continue to be provided by LTE, and it is important that operators with deployed 4G networks have the opportunity to transition some – or all – of their spectrum to newer wireless access technologies. For operators with limited spectrum resources, the possibility of introducing 5G capabilities in an interoperable way – thereby allowing legacy devices to continue to be served on a compatible carrier – is highly beneficial and, in some cases, even vital.

At the same time, the evolution of LTE to a point where it is a full member of the 5G family of air interfaces is essential, especially since initial deployment of new air interfaces may not operate in the same bands. The 5G network will enable dual-connectivity between LTE operating within bands below 6GHz and the NX air interface in bands within the range 6GHz to100GHz. NX should also allow for user-plane aggregation, i.e. joint delivery of data via LTE and NX component carriers.

This paper explains the key requirements and capabilities of 5G, along with its technology components and spectrum needs.

5G – REQUIREMENTS AND CAPABILITIES

In order to enable connectivity for a very wide range of applications with new characteristics and requirements, the capabilities of 5G wireless access must extend far beyond those of previous generations of mobile communication. These capabilities will include massive system capacity, very high data rates everywhere, very low latency, ultra-high reliability and availability, very low device cost and energy consumption, and energy-efficient networks.

MASSIVE SYSTEM CAPACITY

Traffic demands for mobile-communication systems are predicted to increase dramatically [1] [2]. To support this traffic in an affordable way, 5G networks must deliver data with much lower cost per bit compared with the networks of today. Furthermore, the increase in data consumption will result in an increased energy footprint from networks. 5G must therefore consume significantly lower energy per delivered bit than current cellular networks.

The exponential increase in connected devices, such as the deployment of billions of wirelessly connected sensors, actuators and similar devices for massive machine connectivity, will place demands on the network to support new paradigms in device and connectivity management that do not compromise security. Each device will generate or consume very small amounts of data, to the extent that they will individually, or even jointly, have limited impact on the overall traffic volume. However, the sheer number of connected devices seriously challenges the ability of the network to provision signaling and manage connections.

VERY HIGH DATA RATES EVERYWHERE

Every generation of mobile communication has been associated with higher data rates compared with the previous generation. In the past, much of the focus has been on the peak data rate that can be supported by a wireless-access technology under ideal conditions. However, a more important capability is the data rate that can actually be provided under real-life conditions in different scenarios.

- 5G should support data rates exceeding 10Gbps in specific scenarios such as indoor and dense outdoor environments.
- Data rates of several 100Mbps should generally be achievable in urban and suburban environments.
- Data rates of at least 10Mbps should be accessible almost everywhere, including sparsely-populated rural areas in both developed and developing countries.

VERY LOW LATENCY

Very low latency will be driven by the need to support new applications. Some envisioned 5G use cases, such as traffic safety and control of critical infrastructure and industry processes, may require much lower latency compared with what is possible with the mobile-communication systems of today.

To support such latency-critical applications, 5G should allow for an application end-to-end latency of 1ms or less, although application-level framing requirements and codec limitations for media may lead to higher latencies in practice. Many services will distribute computational capacity and storage close to the air interface. This will create new capabilities for real-time communication and will allow ultra-high service reliability in a variety of scenarios, ranging from entertainment to industrial process control.

ULTRA-HIGH RELIABILITY AND AVAILABILITY

In addition to very low latency, 5G should also enable connectivity with ultra-high reliability and ultra-high availability. For critical services, such as control of critical infrastructure and traffic safety, connectivity with certain characteristics, such as a specific maximum latency, should not merely be ‘typically available.’ Rather, loss of connectivity and deviation from quality of service requirements must be extremely rare. For example, some industrial applications might need to guarantee successful packet delivery within 1 ms with a probability higher than 99.9999 percent.

VERY LOW DEVICE COST AND ENERGY CONSUMPTION

Low-cost, low-energy mobile devices have been a key market requirement since the early days of mobile communication. However, to enable the vision of billions of wirelessly connected sensors, actuators and similar devices, a further step has to be taken in terms of device cost and energy consumption. It should be possible for 5G devices to be available at very low cost and with a battery life of several years without recharging.

ENERGY-EFFICIENT NETWORKS

While device energy consumption has always been prioritized, energy efficiency on the network side has recently emerged as an additional KPI, for three main reasons:

- Energy efficiency is an important component in reducing operational cost, as well as a driver for better dimensioned nodes, leading to lower total cost of ownership.
- Energy efficiency enables off-grid network deployments that rely on medium-sized solar panels as power supplies, thereby enabling wireless connectivity to reach even the most remote areas.
- Energy efficiency is essential to realizing operators’ ambition of providing wireless access in a sustainable and more resource-efficient way.

The importance of these factors will increase further in the 5G era, and energy efficiency will therefore be an important requirement in the design of 5G wireless access.

MACHINE-TYPE COMMUNICATION

Fundamentally, applications such as mobile telephony, mobile broadband and media delivery are about information for humans. In contrast, many of the new applications and use cases that drive the requirements and capabilities of 5G are about end-to-end communication between machines. To distinguish them from the more human-centric wireless-communication use cases, these applications are often termed machine-type communication (MTC).

Although spanning a wide range of applications, MTC applications can be divided into two main categories – massive MTC and critical MTC – depending on their characteristics and requirements.

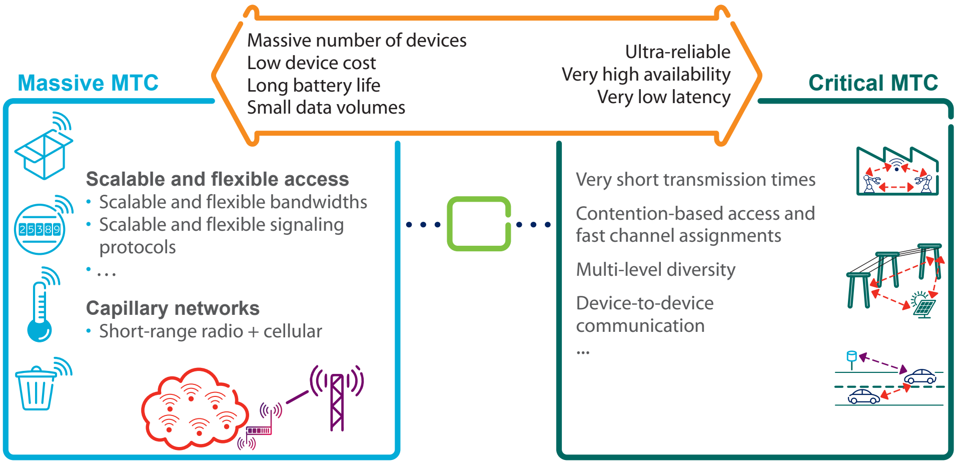

Massive number of devices
Low device cost
Long battery life
Small data volumes
Ultra-reliable
Very high availability
Very low latency
Massive MTC
Critical MTC
Scalable and flexible access
• Scalable and flexible bandwidths
• Scalable and flexible signaling protocols
• …
Capillary networks
• Short-range radio + cellular
Very short transmission times
Contention-based access and fast channel assignments
Multi-level diversity
Device-to-device communication
…

Figure 2: Massive MTC and critical MTC.

Massive MTC refers to services that typically span a very large numbers of devices, usually sensors and actuators. Sensors are extremely low cost and consume very low amounts of energy in order to sustain long battery life. Clearly, the amount of data generated by each sensor is normally very small, and very low latency is not a critical requirement. While actuators are similarly limited in cost, they will likely have varying energy footprints ranging from very low to moderate energy consumption.

Sometimes, the mobile network may be used to bridge connectivity to the device by means of capillary networks. Here, local connectivity is provided by means of a short-range radio access technology, for example Wi-Fi, Bluetooth [3] or 802.15.4/6LoWPAN [4]. Wireless connectivity beyond the local area is then provided by the mobile network via a gateway node.

Critical MTC refers to applications such as traffic safety/control, control of critical infrastructure and wireless connectivity for industrial processes. Such applications require very high reliability and availability in terms of wireless connectivity, as well as very low latency. On the other hand, low device cost and energy consumption is not as critical as for massive MTC applications. While the average volume of data transported to and from devices may not be large, wide instantaneous bandwidths are useful in being able to meet capacity and latency requirements.

There is much to gain from a network being able to handle as many different applications as possible, including mobile broadband, media delivery and a wide range of MTC applications by means of the same basic wireless-access technology and within the same spectrum. This avoids spectrum fragmentation and allows operators to offer support for new MTC services for which the business potential is inherently uncertain, without having to deploy a separate network and reassign spectrum specifically for these applications.

SPECTRUM FOR 5G

In order to support increased traffic capacity and to enable the transmission bandwidths needed to support very high data rates, 5G will extend the range of frequencies used for mobile communication. This includes new spectrum below 6GHz, as well as spectrum in higher frequency bands.

Specific candidate spectrum for mobile communication in higher frequency bands is yet to be identified by the ITU-R or by individual regulatory bodies. The mobile industry remains agnostic about particular choices, and the entire frequency range up to approximately 100GHz is under consideration at this stage, although there is significant interest in large contiguous allocations that can provide dedicated and licensed spectrum for use by multiple competing network providers.

The lower part of this frequency range, below 30GHz, is preferred from the point of view of propagation properties. At the same time, very large amounts of spectrum and the possibility of wide transmission frequency bands of the order of 1GHz or more are more likely above 30GHz.

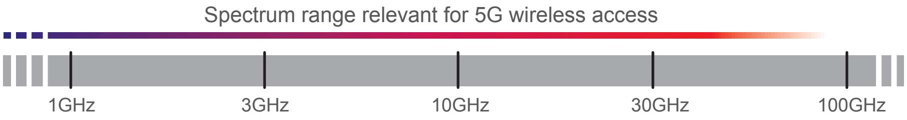

Figure 3: Spectrum relevant for 5G wireless access.

Spectrum relevant for 5G wireless access therefore ranges from below 1GHz up to approximately 100GHz, as Figure 3 shows.

It is important to understand that high frequencies, especially those above 10GHz, can only serve as a complement to lower frequency bands, and will mainly provide additional system capacity and very wide transmission bandwidths for extreme data rates in dense deployments. Spectrum allocations at lower bands will remain the backbone for mobile-communication networks in the 5G era, providing ubiquitous wide-area connectivity.

The World Radio Conference (WRC)-15 discussions have resulted in an agreement to include an agenda item for IMT-2020, the designated ITU-R qualifier for 5G, in WRC-19. The conference also reached agreement on a set of bands that will be studied for 5G, with direct applicability to NX. Many of the proposed bands are in the millimeter wave region and include:

- 24.25GHz to 27.5GHz, 37GHz to 40.5GHz, 42.5GHz to 43.5GHz, 45.5GHz to 47GHz, 47.2GHz to 50.2GHz, 50.4GHz to 52.6GHz, 66 GHz to 76GHz and 81GHz to 86GHz, which have allocations to the mobile service on a primary basis; and
- 31.8GHz to 33.4GHz, 40.5GHz to 42.5GHz and 47GHz to 47.2GHz, which may require additional allocations to the mobile service on a primary basis.

The mobile industry will strive to gain access to spectrum in the 6GHz to 20GHz range, but the policy directions being followed by regulators seem to be focused on frequency bands above 30GHz. In the US, the FCC has issued two Notices of Public Rule Making (NPRM) on bands above 24GHz. Ofcom has likewise indicated a preference for bands above 30GHz within the mobile industry.

The capacity needs of the mobile industry will continue to be served by licensed spectrum, although novel sharing arrangements for spectrum will become progressively more important as restricted opportunities for new spectrum start to impact incumbent services such as satellite communication and radio location. Two examples of sharing arrangements include LSA planned in Europe for the 2.3GHz band and the Citizens Band Radio Service for 3.5GHz in the US.

5G TECHNOLOGY COMPONENTS

Beyond extending operation to higher frequencies, there are several other key technology components relevant for the evolution to 5G wireless access. These components include access/backhaul integration, device-to-device communication, flexible duplex, flexible spectrum usage, multi-antenna transmission, ultra-lean design, and user/control separation.

ACCESS/BACKHAUL INTEGRATION

Wireless technology is already frequently used as part of the backhaul solution. Such wireless-backhaul solutions typically operate under line-of-sight conditions using proprietary radio technology in higher frequency bands, including the millimeter wave (mmW) band.

In the future, the access (base-station-to-device) link will also extend to higher frequencies. Furthermore, to support dense low-power deployments, wireless backhaul will have to extend to cover non-line-of-sight conditions, similar to access links.

In the 5G era, the wireless-access link and wireless backhaul should not therefore be seen as two separate entities with separate technical solutions. Rather, backhaul and access should be seen as an integrated wireless-access solution able to use the same basic technology and operate using a common spectrum pool. This will lead to more efficient overall spectrum utilization as well as reduced operation and management effort.

DIRECT DEVICE-TO-DEVICE COMMUNICATION

The possibility of limited direct device-to-device (D2D) communication has recently been introduced as an extension to the LTE specifications. In the 5G era, support for D2D as part of the overall wireless-access solution should be considered from the start. This includes peer-to-peer user-data communication directly between devices, but also, for example, the use of mobile devices as relays to extend network coverage.

D2D communication in the context of 5G should be an integral part of the overall wireless-access solution, rather than a stand-alone solution. Direct D2D communication can be used to offload traffic, extend capabilities and enhance the overall efficiency of the wireless-access network. Furthermore, in order to avoid uncontrolled interference to other links, direct D2D communication should be under network control. This is especially important for the case of D2D communication in licensed spectrum.

FLEXIBLE DUPLEX

Frequency Division Duplex (FDD) has been the dominating duplex arrangement since the beginning of the mobile communication era. In the 5G era, FDD will remain the main duplex scheme for lower frequency bands. However, for higher frequency bands – especially above 10GHz – targeting very dense deployments, Time Division Duplex (TDD) will play a more important role.

In very dense deployments with low-power nodes, the TDD-specific interference scenarios (direct base-station-to-base-station and device-to-device interference) will be similar to the 'normal' base-station-to-device and device-to-base-station interference that also occurs for FDD.

Furthermore, for the dynamic traffic variations expected in very dense deployments, the ability to dynamically assign transmission resources (time slots) to different transmission directions may allow more efficient utilization of the available spectrum.

To reach its full potential, 5G should therefore allow for very flexible and dynamic assignment of TDD transmission resources. This is in contrast to current TDD-based mobile technologies, including TD-LTE, for which there are restrictions on the downlink/uplink configurations, and for which there typically exist assumptions about the same configuration for neighbor cells and also between neighbor operators.

FLEXIBLE SPECTRUM USAGE

Since its inception, mobile communication has relied on spectrum licensed on a per-operator basis within a geographical area. This will remain the foundation for mobile communication in the 5G era, allowing operators to provide high-quality connectivity in a controlled-interference environment.

However, per-operator licensing of spectrum will be complemented with the possibility to share spectrum. Such sharing may be between a limited set of operators, or may occur in license-exempt scenarios. The

Citizens Band Radio Service in the US in the 3.5GHz band and the 5GHz unlicensed spectrum are examples of managed and unlicensed sharing regimes respectively.

New air interfaces like NX will likely be well served by more conventional licensed allocations of spectrum, mainly due to the need to establish a basic foundation for the technology to operate in an independent manner while interoperability is established with technologies like LTE. At some point, further allocations of spectrum for 5G may leverage the mobile industry's experience of sharing approaches in lower cellular bands.

MULTI-ANTENNA TRANSMISSION

Multi-antenna transmission already plays an important role in current generations of mobile communication and will be even more central in the 5G era, due to the physical limitations of small antennas. Path loss between a transmitter and receiver does not change as a function of frequency, as long as the effective aperture of the transmitting and receiving antennas does not change. The antenna aperture does reduce in proportion to the square of the frequency, and that reduction can be compensated by the use of higher antenna directivity. The 5G radio will employ hundreds of antenna elements to increase antenna aperture beyond what may be possible with current cellular technology.

In addition, the transmitter and receiver will use beamforming to track one another and improve energy transfer over an instantaneously configured link. Beamforming will also improve the radio environment by limiting interference to small fractions of the entire space around a transmitter and likewise limiting the impact of interference on a receiver to infrequent stochastic events. The use of beamforming will also be an important technology for lower frequencies; for example, to extend coverage and to provide higher data rates in sparse deployments.

ULTRA-LEAN DESIGN

Ultra-lean radio-access design is important to achieve high efficiency in 5G networks. The basic principle of ultra-lean design can be expressed as: minimize any transmissions not directly related to the delivery of user data. Such transmissions include signals for synchronization, network acquisition and channel estimation, as well as the broadcast of different types of system and control information.

Ultra-lean design is especially important for dense deployments with a large number of network nodes and highly variable traffic conditions. However, lean transmission is beneficial for all kinds of deployments, including macro deployments.

By enabling network nodes to enter low-energy states rapidly when there is no user-data transmission, ultra-lean design is an important component in delivering high network energy performance. Ultra-lean design will also enable higher achievable data rates by reducing interference from non-user-data-related transmissions.

USER/CONTROL SEPARATION

Another important design principle for 5G is to decouple user data and system control functionality. The latter includes the provisioning of system information; that is, the information and procedures needed for a device to access the system.

Such a decoupling will allow separate scaling of user-plane capacity and basic system control functionality. For example, user data may be delivered by a dense layer of access nodes, while system information is only provided via an overlaid macro layer on which a device also initially accesses the system.

It should be possible to extend the separation of user data delivery and system control functionality over multiple frequency bands and RATs. As an example, the system control functionality for a dense layer based on new high-frequency radio access could be provided by means of an overlaid LTE layer.

User/control separation is also an important component for future radio-access deployments relying heavily on beamforming for user data delivery. Combining ultra-lean design with a logical separation of user-plane data delivery and basic system connectivity functionality will enable a much higher degree of device-centric network optimization of the active radio links in the network. Since only the ultra-lean signals related to the system control plane need to be static, it is possible to design a system where almost everything can be dynamically optimized in real time.

An ultra-lean design combined with a system control plane logically separated from the user data delivery function also provides higher flexibility in terms of evolution of the RAT as, with such separation, the user plane can evolve while retaining system control functionality.

CONCLUSION

5G is the next step in the evolution of mobile communication and will be a key component of the Networked Society. In particular, 5G will accelerate the development of the Internet of Things. To enable connectivity for a wide range of applications and use cases, the capabilities of 5G wireless access must extend far beyond those of previous generations of mobile communications.

These capabilities include very high achievable data rates, very low latency and ultra-high reliability. Furthermore, 5G wireless access needs to support a massive increase in traffic in an affordable and sustainable way, implying a need for a dramatic reduction in the cost and energy consumption per delivered bit.

5G wireless access will be realized by the evolution of LTE for existing spectrum in combination with new radio access technologies that primarily target new spectrum. Key technology components of 5G wireless access include access/backhaul integration, device-to-device communication, flexible duplex, flexible spectrum usage, multi-antenna transmission, ultra-lean design, and user/control separation.

REFERENCES

[1] ICT-317669 METIS project, Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations, Deliverable D1.5, April 2015, available at:
https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D1.5_v1.pdf

[2] Ericsson, Ericsson Mobility Report, November 2015, available at:
http://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf

[3] Bluetooth Special Interest Group, accessed March 2016, available at:
http://bluetooth.org

[4] IETF, IPv6 over Low power WPAN (6lowpan), accessed March 2016, available at:
http://datatracker.ietf.org/wg/6lowpan

GLOSSARY

D2D	device-to-device
FDD	Frequency Division Duplex
mmW	millimeter wave
MTC	Machine-Type Communication
NPRM	Notices of Public Rule Making
OFDM	Orthogonal Frequency Division Multiplexing
TDD	Time Division Duplex
WRC	World Radio Conference

© 2016 Ericsson AB – All rights reserved

ERICSSON WHITE PAPER
Uen 284 23-3278 | January 2016

CELLULAR NETWORKS FOR MASSIVE IOT

ENABLING LOW POWER WIDE AREA APPLICATIONS

With new standards specifically targeting the connectivity requirements of Massive Internet of Things (IoT) applications, cellular networks can deliver reliable, secure and diverse IoT services using existing network infrastructure.

DELIVERING NEW VALUE IN THE NETWORKED SOCIETY

In the future, all devices that benefit from an internet connection will be connected. In this Networked Society, every person and every industry will be empowered to reach their full potential. Internet of Things (IoT) technology is a key enabler of this vision by delivering machine-to-machine (M2M) and machine-to-person communications on a massive scale.

As shown in Figure 1, Ericsson predicts there will be around 28 billion connected devices by 2021, of which more than 15 billion will be connected M2M and consumer-electronics devices [1]. A large share of these will be applications served by short-range radio technologies such as Wi-Fi and Bluetooth, while a significant proportion will be enabled by wide area networks (WANs) that are primarily facilitated by cellular networks.

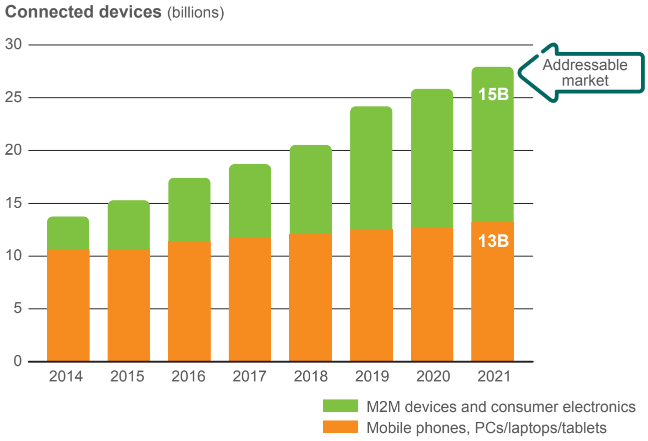

Figure 1: Growth in connected devices.

THE NEW IOT LANDSCAPE

The IoT revolution offers huge potential value in terms of improved efficiency, sustainability and safety for industry and society. Analysts predict that the total added value of the IoT will be USD 1.9 trillion by 2020 [2].

The variety of applications and solutions designed for individuals, business and industry is spurring the rapid expansion of the IoT market. The IoT is playing a major role across a variety of vertical sectors, generating cost savings, new revenue streams and other benefits.

Each IoT application needs a clear value proposition and business logic in line with the prevailing ecosystem, business models and value chains of the various stakeholders. For all applications, solutions need to be integrated on platforms that can scale and handle millions of devices efficiently. Business processes for administration, provisioning and charging will have to be streamlined to minimize costs and enhance the business case.

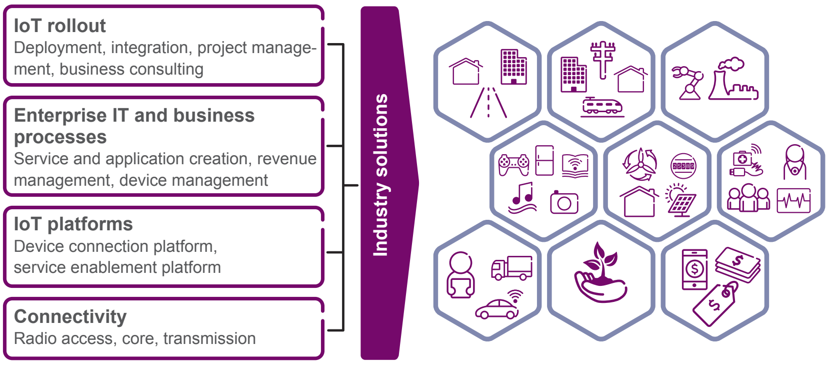
IoT rollout
Deployment, integration, project management, business consulting
Enterprise IT and business processes
Service and application creation, revenue management, device management
IoT platforms
Device connection platform, service enablement platform
Connectivity
Radio access, core, transmission
Industry solutions

Figure 2: The new IoT landscape.

As they are largely responsible for wireless connectivity on a global scale, operators are in an excellent position to capture a share of the added value generated by the emerging IoT market. The size of this share will depend on the role that operators adopt in the value chain. This could range from being a straightforward connectivity provider (monetizing connectivity in new ways), all the way to being an end-to-end solution provider of turnkey solutions to vertical markets [3].

DIFFERENT IOT CONNECTIVITY ALTERNATIVES

Connectivity is the foundation for IoT, and the type of access required will depend on the nature of the application. Many IoT devices will be served by radio technologies that operate on unlicensed spectrum and that are designed for short-range connectivity with limited QoS and security requirements typically applicable for a home or indoor environment. Currently, there are two alternative connectivity tracks for the many IoT applications that depend on wide-area coverage:

Cellular technologies: 3GPP technologies like GSM, WCDMA, LTE and future 5G. These WANs operate on licensed spectrum and historically have primarily targeted high-quality mobile voice and data services. Now, however, they are being rapidly evolved with new functionality and the new radio access technology narrowband IoT (NB-IoT) specifically tailored to form an attractive solution for emerging low power wide area (LPWA) applications.

Unlicensed LPWA: new proprietary radio technologies, provided by, for example, SIGFOX and LoRa, have been developed and designed solely for machine-type communication (MTC) applications addressing the ultra-low-end sensor segment, with very limited demands on throughput, reliability or QoS.

One way to segment IoT applications is to categorize them according to coverage needs and performance requirements (such as data speed or latency demands).

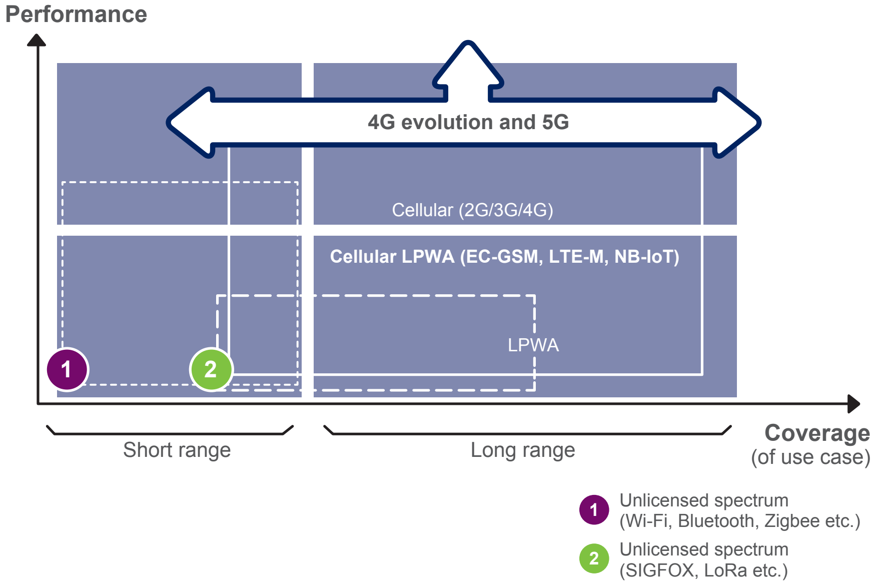

Figure 3: Technologies addressing different IoT segments.

The coverage needs of a particular use case may be highly localized (such as a stationary installation within a building), while other use cases require global service coverage (such as container tracking). 3GPP technologies already dominate use cases with large geographic coverage needs and medium- to high-performance requirements.

With new feature sets specifically tailored for LPWA IoT applications, 3GPP technologies are taking a large leap forward to cover segments with low-cost, low-performance requirements too.

CAPILLARY NETWORKS COMBINING CELLULAR AND UNLICENSED STRENGTHS

Even when existing 3GPP end-to-end connectivity is not feasible, cellular technology can still provide key benefits when used as a bridging option, i.e. as an aggregation and routing solution. This capillary network approach allows end devices to utilize varying access solutions from either the short range or LPWA domain and access the cellular networks via a gateway device. Capillary networks enable the reuse of cellular functions and assets such as security, device management, billing and QoS without requiring each end device to be cellular-enabled.

A WIDE RANGE OF IOT REQUIREMENTS

There will be a wide range of IoT use cases in the future, and the market is now expanding toward both Massive IoT deployment as well as more advanced solutions that may be categorized as Critical IoT, as shown in Figure 4.

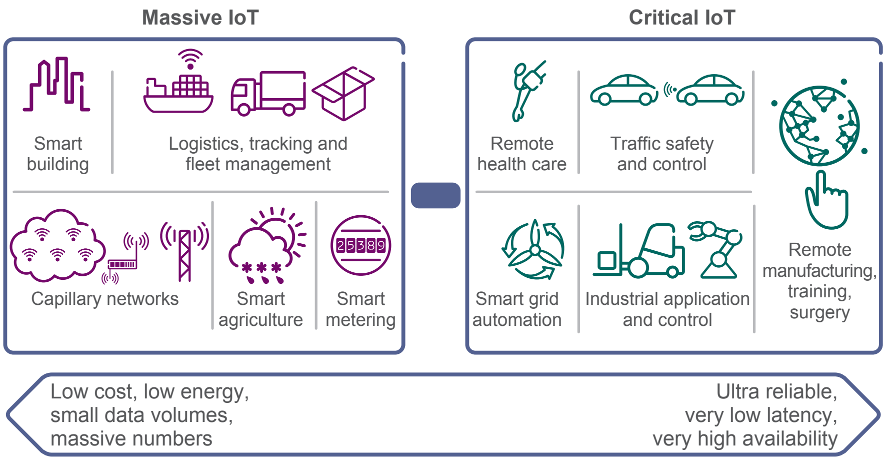
Massive IoT
Critical IoT
Smart building
Logistics, tracking and fleet management
Capillary networks
Smart agriculture
25389
Smart metering
Remote health care
Traffic safety and control
Smart grid automation
Industrial application and control
Remote manufacturing, training, surgery
Low cost, low energy, small data volumes, massive numbers
Ultra reliable, very low latency, very high availability

Figure 4: Differing requirements for Massive and Critical IoT applications.

At one end of the scale, in Massive IoT applications – typically sensors that report to the cloud on a regular basis – the end-to-end cost must be low enough for the business case to make sense. Here, the requirement is for low-cost devices with low energy consumption and good coverage.

At the other end of the scale, Critical IoT applications will have very high demands for reliability, availability and low latency. These use cases are enabled by LTE or 5G capabilities. Here, the volumes are typically much smaller, but the business value is significantly higher.

There are, however, many other use cases between the two extremes, which today rely on 2G, 3G or 4G connectivity.

MASSIVE IOT USE CASE DIVERSITY

The Massive IoT market segment includes several applications widely used in industries and societies, as shown in Figure 5.

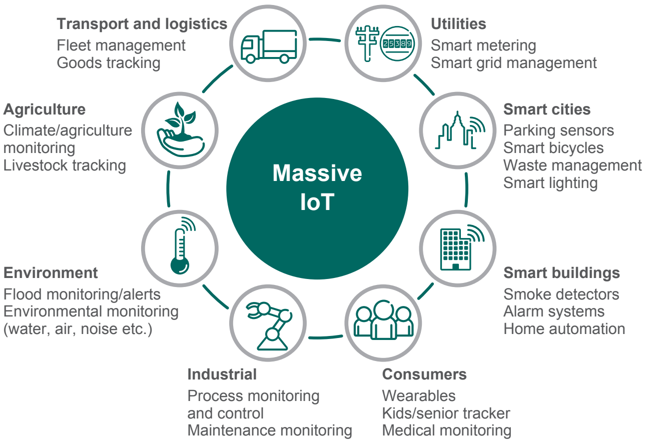
Massive IoT
Transport and logistics
Fleet management
Goods tracking
Utilities
Smart metering
Smart grid management
Agriculture
Climate/agriculture monitoring
Livestock tracking
Smart cities
Parking sensors
Smart bicycles
Waste management
Smart lighting
Environment
Flood monitoring/alerts
Environmental monitoring (water, air, noise etc.)
Smart buildings
Smoke detectors
Alarm systems
Home automation
Industrial
Process monitoring and control
Maintenance monitoring
Consumers
Wearables
Kids/senior tracker
Medical monitoring

Figure 5: Industry and society applications enabled by LPWA.

The potential applications for the IoT run into the millions, with a huge variety of requirements regarding cost, battery life, coverage, connectivity performance (throughput and capacity), security and reliability. Figure 6 illustrates some such applications and their requirements regarding devices and connectivity.

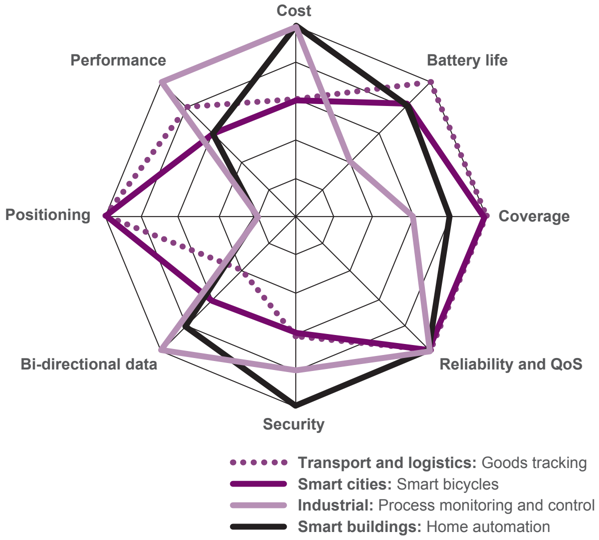

Figure 6: Device and connectivity requirements for sample IoT use cases.

Some devices will only send a few messages per day – such as status indicators for temperature – while others may need to transmit a video stream to guide a remote repair technician, for example. The difference in throughput requirements is huge. If operators or service providers handle several applications, it may be of great benefit to be able to harmonize communication modules, so that they all use the same underlying radio solution to reduce operational and fault management effort and complexity.

Many higher-value applications will require two-way communications – in other words, an uplink as well as a downlink – to enable monitoring and control of devices in systems like heating, ventilation and cooling plants. The long lifetime of many IoT applications makes it invaluable to be able to perform over-the-air device updates for new functionality or parameter settings. The amount of data sent for such updates can often be more demanding for the network than the monitoring or control application itself.

Relatively simple uplink-focused applications can benefit greatly from a bi-directional link to provide robustness. For example, a connected smoke detector must deliver a smoke alarm with absolute certainty. The ability of a network to provide acknowledgements of a received message enables better fault management and the required level of reliability. Positioning can be used to locate the sensor at the point it failed and simplify operations. For tracking applications, location information is essential.

In applications like building security, sensitive information could be reported over the air, which will require strict security. Furthermore, in the case of a break-in, it is crucial that the alarm information reaches the control center in time – making QoS and two-way communication vital.

KEY CHALLENGES FOR MASSIVE IOT

The key challenges to enabling large-scale uptake of Massive IoT include:

Device cost – clearly a key enabler for high-volume, mass-market applications, enabling many of the use cases.

Battery life – many IoT devices will be battery-powered, and often the cost of replacing batteries in the field is not viable.

Coverage – deep indoor connectivity is a requirement for many applications in the utility area. Furthermore, regional (or even national or global) coverage is a prerequisite for many use cases, especially within the transport area.

Scalability – in order to enable a Massive IoT market, networks need to scale efficiently. The initial investment required for supporting a limited number of devices has to be manageable, while on the other hand, the network capacity must be easy to scale to handle thousands – or millions – of devices.

Diversity – connectivity should be able to support diverse requirements from different use cases. One network supporting everything from simple static sensors to tracking services, to applications requiring higher throughput and lower latency is essential in terms of total cost of ownership (TCO).

THE ADVANTAGES OF CELLULAR TECHNOLOGIES

Each of the technologies available for IoT connectivity has its own advantages and disadvantages. However, the range of IoT connectivity requirements – both technical and commercial – means cellular technologies can provide clear benefits across a wide variety of applications, as summarized in Figure 7.

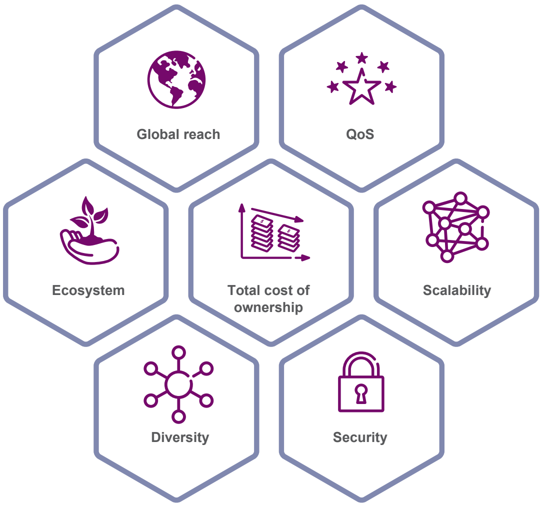

Figure 7: Advantages of cellular technologies.

In terms of global reach, cellular networks already cover 90 percent of the world's population. WCDMA and LTE are catching up, but GSM will offer superior coverage in many markets for years to come. Cellular networks have been developed and deployed over three decades, and they will be around for the foreseeable future.

The cellular mobile industry represents a huge and mature ecosystem, incorporating chipset, device and network equipment vendors, operators, application providers and many others. The global cellular ecosystem is governed by the 3GPP standardization forum, which guarantees broad industry support for future development.

When it comes to scalability, cellular networks are built to handle massive volumes of mobile broadband traffic; the traffic from most IoT applications will be relatively small and easily absorbed. Operators are able to offer connectivity for IoT applications from the start-up phase and grow this business with low TCO and only limited additional investment and effort. Operation in licensed spectrum also provides predictable and controlled interference, which enables efficient use of the spectrum to support massive volumes of devices.

Cellular connectivity offers the diversity to serve a wide range of applications with varying requirements within one network. While competing unlicensed LPWA technologies are designed solely for very low-end MTC applications, cellular networks can address everything from Massive to Critical IoT use cases.

QoS mechanisms will be essential for many IoT applications. Cellular systems have mature QoS functionality, and this enables critical MTC applications to be handled together with traffic from sensors, voice and mobile-broadband traffic on the same carrier. QoS, along with licensed spectrum as described above, provides a foundation for long-term Service Level Agreements with a specific grade of service.

Traditionally, the security mechanisms of cellular networks have been based on a physical SIM attached to the device, referred to as a Universal Integrated Circuit Card (UICC). This has also enabled roaming between operators, which has been one of the main factors behind the huge success of mobile networks. The SIM will also be essential in future IoT applications, with SIM functionality embedded in the chipset (eUICC) or handled as a soft-SIM solution running in a trusted run-time environment of the module.

With a straightforward rollout of new software, cellular networks will be able to support the full breadth of applications, ranging from low-end use cases in the LPWA segment, to the high-end segments of in-car entertainment and video surveillance. One network connecting the whole diversifying IoT market will guarantee the lowest possible TCO as well as fast time to market.

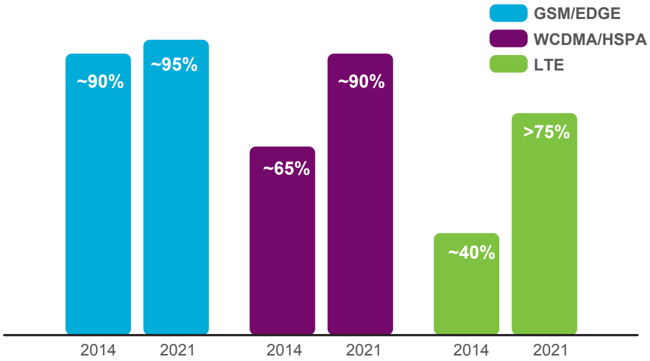

Figure 8: World population coverage by cellular 3GPP technology.

EVOLVING STANDARDS

To meet the new connectivity requirements of the emerging Massive IoT segment, 3GPP has taken evolutionary steps on both the network side and the device side.

The key improvement areas addressed in 3GPP up to Release 13 are:

- Lower device cost – cutting module cost for LTE devices by reducing peak rate, memory requirement and device complexity. The LTE module cost-reduction evolution started in Release 8 with the introduction of LTE for machine-type communication (LTE-M) Cat 1 devices with reduced peak rate to a maximum of 10Mbps, and continued in Releases 12 and 13 with reduced device complexity for lower performance and using less bandwidth or a narrowband IoT carrier to cut costs further.
- Improved battery life – more than 10 years of battery life can be achieved by introducing Power Saving Mode and/or extended discontinuous reception (eDRX) functionality. These features allow the device to contact the network – or to be contacted – on a per-need basis, meaning that it can stay in sleep mode for minutes, hours or even days.
- Improved coverage – an improvement of 15dB on LTE-M and of 20dB on NB-IoT and GSM, which translates into a seven-fold increase in the outdoor coverage area and significantly improved indoor signal penetration to reach deep indoors. This supports many IoT devices like smart meters, which are often placed in a basement.
- Support for massive numbers of IoT connections – specifically, one LTE cell site can support millions of IoT devices, depending on the use case. Core network enhancements include software upgrades for service differentiation handling, signaling optimization and high-capacity platforms (more than 30 million devices per node).

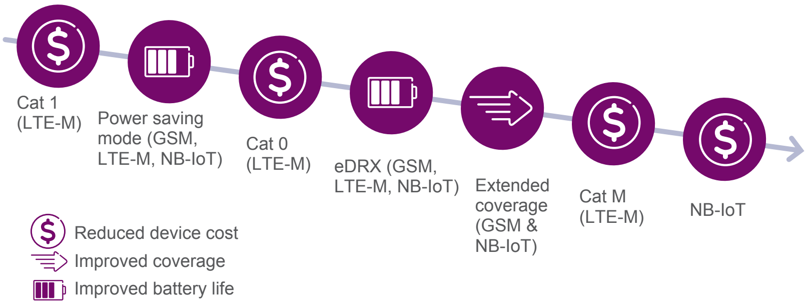

Figure 9: 3GPP evolution steps for Massive IoT.

A FULL RANGE OF CELLULAR LPWA SOLUTIONS

No single technology or solution is ideally suited to all the different potential Massive IoT applications, market situations and spectrum availability. As a result, the mobile industry is standardizing several LPWA technologies, including Extended Coverage GSM (EC-GSM), LTE-M and NB-IoT.

LTE-M, NB-IoT and EC-GSM are all superior solutions to meet Massive IoT requirements as a family of solutions, and can complement each other based on technology availability, use case requirements and deployment scenarios. LTE-M consisting of Cat 1, Cat 0 and Cat M supports a wide range of IoT applications, including those that are content-rich; NB-IoT covers ultra-low-end IoT applications with a cost and coverage advantage over LTE-M; and EC-GSM serves IoT services for all GSM markets.

For example, a smart city application such as waste management may use EC-GSM technology to provide LPWA connectivity in markets where it can be deployed on existing 2G networks; NB-IoT technology may be used for water-metering applications, which have some of the most extreme coverage requirements in underground locations. On the other hand, asset-tracking applications that can support a relatively high number of messages triggered by certain events may employ LTE-M.

EC-GSM – GLOBAL CELLULAR IOT FOR ALL GSM MARKETS

GSM is still the dominant mobile technology in many markets, and the vast majority of cellular M2M applications today use GPRS/EDGE for connectivity. GSM is likely to continue playing a key role in the IoT well into the future, due to its global coverage footprint, time to market and cost advantages.

Recognizing this – and identifying the requirements for Massive IoT discussed earlier in this paper – an initiative was undertaken in 3GPP Release 13 to further improve GSM.

The resulting EC-GSM functionality enables coverage improvements of up to 20dB with respect to GPRS on the 900MHz band.

This coverage extension is achieved for both the data and control planes by utilizing the concept of repetitions and signal combining techniques. It is handled in a dynamic manner with multiple coverage classes to ensure optimal balance between coverage and performance.

EC-GSM is achieved by defining new control and data channels mapped over legacy GSM. It allows multiplexing of new EC-GSM devices and traffic with legacy EDGE and GPRS. No new network carriers are required: new software on existing GSM networks is sufficient and provides combined capacity of up to 50,000 devices per cell on a single transceiver.

Initially part of EC-GSM but now a separate 3GPP item, eDRX improves the power efficiency – and therefore the battery life – for many use cases. eDRX improves the idle mode behavior by allowing the use of a number of inactivity timers, where the device can choose to tune in to the network and listen for downlink pages and traffic.

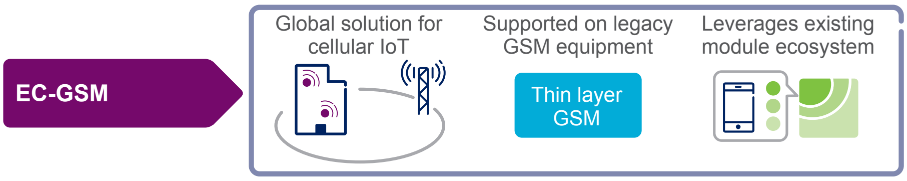
EC-GSM
Global solution for cellular IoT
Supported on legacy GSM equipment
Thin layer GSM
Leverages existing module ecosystem

Figure 10: EC-GSM.

EC-GSM extends the data handling and power efficiency advantages that GSM/GPRS technology already offers for MTC, and it will help operators extend the service life of their huge 2G legacy base.

LTE-M – SUPPORTING A WIDE RANGE OF MASSIVE IOT USE CASES

LTE is the leading mobile broadband technology and its coverage is expanding rapidly. So far, the focus has been on meeting the huge demand for mobile data with highly capable devices that utilize new spectrum. With features like Carrier Aggregation, MIMO and Lean Carrier, the gigabit per second performance for LTE cell throughput is now reaching levels that result in an excellent mobile broadband user experience.

The advent of LTE-M signifies an important step in addressing MTC capabilities over LTE. LTE-M brings new power-saving functionality suitable for serving a variety of IoT applications; Power Saving Mode and eDRX extend battery life for LTE-M to 10 years or more. LTE-M traffic is multiplexed over a full LTE carrier, and it is therefore able to tap into the full capacity of LTE. Additionally, new functionality for substantially reduced device cost and extended coverage for LTE-M are also specified within 3GPP.

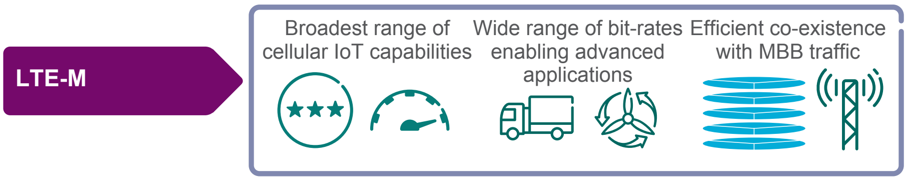
LTE-M
Broadest range of cellular IoT capabilities
Wide range of bit-rates enabling advanced applications
Efficient co-existence with MBB traffic

Figure 11: LTE M.

NB-IOT – SUPPORTING ULTRA-LOW-END MASSIVE IOT APPLICATIONS

In addition to LTE-M, NB-IoT technology is being standardized in time for 3GPP Release 13. NB-IoT is a self-contained carrier that can be deployed with a system bandwidth of only 200kHz, and is specifically tailored for ultra-low-end IoT applications. It is enabled using new network software on an existing LTE network, which will result in rapid time to market.

NB-IoT
Scalable ultra-low-end cellular IoT solution
Ultra-low bit rates and extreme coverage
789
Native narrowband solution
In-band
LTE
200kHz

Figure 12: NB-IoT.

NB-IoT provides lean setup procedures, and a capacity evaluation indicates that each 200kHz NB-IoT carrier can support more than 200,000 subscribers. The solution can easily be scaled up by adding multiple NB-IoT carriers when needed. NB-IoT also comes with an extended coverage of up to 20dB, and battery saving features, Power Saving Mode and eDRX for more than 10 years of battery life.

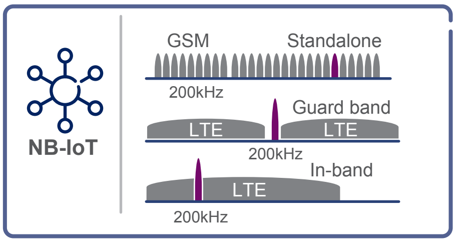
NB-IoT
GSM
Standalone
200kHz
Guard band
LTE
LTE
200kHz
In-band
LTE
200kHz

Figure 13: NB-IoT deployment.

NB-IoT is designed to be tightly integrated and interwork with LTE, which provides great deployment flexibility. The NB-IoT carrier can be deployed in the LTE guard band, embedded within a normal LTE carrier, or as a standalone carrier in, for example, GSM bands.

- Standalone deployment in a GSM low band: this is an option when LTE is deployed in a higher band and GSM is still in use, providing coverage for basic services.
- Guard band deployment, typically next to an LTE carrier: NB-IoT is designed to enable deployment in the guard band immediately adjacent to an LTE carrier, without affecting the capacity of the LTE carrier. This is particularly suitable for spectrum allocations that do not match the set of LTE system bandwidths, leaving gaps of unused spectrum next to the LTE carrier.
- Efficient in-band deployment, allowing flexible assignment of resources between LTE and NB-IoT: it will be possible for an NB-IoT carrier to time-share a resource with an existing LTE carrier. The in-band deployment also allows for highly flexible migration scenarios. For example, if the NB-IoT service is first deployed as a standalone deployment in a GSM band, it can subsequently be migrated to an in-band deployment if the GSM spectrum is re-farmed to LTE, thereby avoiding any fragmentation of the LTE carrier.

NB-IoT reduces device complexity below that of LTE-M with the potential to rival module costs of unlicensed LPWA technologies, and it will be ideal for addressing ultra-low-end applications in markets with a mature LTE installed base.

CONCLUSION

Uptake of Massive IoT is set to take off, and operators have a unique opportunity to drive the implementation of new IoT applications by offering affordable connectivity on a global scale.

For IoT applications, existing cellular networks offer distinct advantages over alternative WAN technologies, such as unlicensed LPWA. The global reach, QoS, ecosystem, TCO, scalability, diversity and security of cellular networks are all vital factors that can support the fast uptake and success of IoT. Enabled by new software in existing legacy networks, cellular networks can support a diverse range of IoT applications – ensuring the lowest possible TCO.

3GPP standardization work for GSM and LTE, and the recent addition of NB-IoT, is further improving the ability of cellular networks to address the Massive IoT market, where ultra-low end-to-end cost is a prerequisite.

GSM/GPRS, which already serves the majority of cellular-based MTC applications, is evolving with a new EC-GSM standard, which delivers significantly better energy efficiency and increased coverage. EC-GSM enhancements will cement GSM's position as a highly relevant connectivity platform for low-end, Massive IoT applications globally.

New downsized NB-IoT and LTE-M chipsets, designed for MTC, and features that improve both coverage and device battery life will boost the ability of LTE infrastructure to address the IoT market. One network that supports all applications – from advanced mobile broadband services, VoIP and all kinds of low- to high-end IoT use cases – creates a very strong value proposition.

Whether operators choose the GSM, NB-IoT or LTE-M track – or a combination of these – will depend on several factors such as technology coverage, future technology strategies and targeted market segments. Whichever path they take, they have a huge opportunity to benefit from the emerging IoT revolution. Operators can choose to continue offering telecom-grade connectivity as they do today, or they can evolve to become a platform or fully-fledged IoT service provider targeting a larger slice of future IoT revenues.

GLOSSARY

EC-GSM	Extended Coverage GSM
eDRX	extended discontinuous reception
eUICC	embedded Universal Integrated Circuit Card
IoT	Internet of Things
LPWA	low power wide area
LTE-M	LTE for machine-type communication
M2M	machine-to-machine
MIMO	multiple-input, multiple-output
MTC	machine-type communication
NB-IoT	narrowband IoT
TCO	total cost of ownership
UICC	Universal Integrated Circuit Card
WAN	wide area network

REFERENCES

[1] Ericsson, Ericsson Mobility Report, November 2015, available at: http://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf

[2] Gartner, Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020, December 2013, available at: http://www.gartner.com/newsroom/id/2636073

[3] Ericsson, White Paper: Machine-to-Machine – exploring potential operator roles, September 2014, available at: http://www.ericsson.com/res/docs/whitepapers/wp-m2m.pdf

© 2016 Ericsson AB – All rights reserved

THE

Next Wave

The National Security Agency's review of emerging technologies

Vol. 21 | No. 3 | 2017

[Photo credit: StuPorts, AlexLMX/iStock/Thinkstock]

THE
Next Wave
The National Security Agency's review of emerging technologies

Editor's column

While the global mobile telecommunications industry has been attempting to keep pace with ever-changing technology and consumer needs, it has not had a strong development roadmap like the International Technology Roadmap for Semiconductors (ITRS). Instead, mobile technology's development has been characterized by multinational companies pushing proposed standards for adoption by international standards bodies. These standards are subsequently adopted in a haphazard, nation-by-nation process. However, consumer desire for faster, more fully featured mobile devices has proven to be as strong a driving force as the ITRS, and has led the industry to roll out generations of new technology on a roughly 10-year basis (approximate dates: 1G - 1981, 2G - 1992, 3G - 2001, and 4G - 2009). This decade-by-decade introduction of new mobile communication technology has led to the common prediction for fifth-generation (5G) mobile to appear around 2020, in keeping with the observed linear cadence.

But why create 5G mobile anyway? The push for 5G is not just a mad rush to keep pace and provide more bandwidth to services that can already stream high-definition video. The Internet of Things (IoT; see TNW Vol. 21 No. 2) is a major driving force (among several) behind technologies being developed for 5G mobile. As personal mobile devices are more enmeshed into machine-to-machine (M2M) communications and the number of IoT sensors explodes, 5G technologies must address several needs: high-speed data rates for many more users, increased density of users, greatly increased simultaneous connections, and reduced latencies. These needs will propel many of the technologies that Dr. Farroha et al. describe in their introduction to 5G article (page 2).

One of those technologies, virtualization, has been instrumental in making efficient use of servers (virtual machine or VM) and computer networking (software-defined networking or SDN). Over the course of continuing improvement to 4G implementations, the networks are evolving into an all-digital Internet protocol packet-switched system. This evolution means that those efficiencies developed for SDN and used in computing can be applied to mobile networking. We can see this application in more depth in the article on 5G virtualization (page 16).

Several markets have already taken advantage of 4G mobile technology, most notably media. This market is expected to expand with 5G as consumer desire for high-resolution video and augmented/virtual reality increases. Automotive, energy, health, and public safety are several areas of M2M that will be greatly enhanced by increased bandwidth and, most importantly, low-latency networks. On 13 December 2016, the US Department of Transportation proposed a rule for vehicle-to-vehicle communication and announced a vehicle-to-infrastructure communication rule to be proposed. The article on 5G and the automotive industry (page 20) provides insight into how this market segment is enhanced by 5G connectivity.

Although mobile technological developments are typically categorized into different generations, each generation covers a vast array of individual technologies and protocols that roll out as they mature. In practice, mobile service providers often work across a range of generations. This is highlighted by the fact that in many cases voice calling is handled by the 3G network while data is handled on the 4G network. Domestically, T-Mobile and AT&T did not enable voice over LTE (VoLTE) until May 2014, and did not meet the International Telecommunication Union (ITU) standard until Verizon launched LTE-Advanced in August 2016.

Many factors can impact the broad global deployment of 5G networks, such as existing technologies, geography, spectrum, and national interests. Some countries may jump to the current network generation while others may delay deployment because their current networks are considered to be "good enough." The decreased per-customer costs of updating networks in densely populated countries makes it more attractive to keep on the leading edge. The increased data rates and simultaneous connections require increased spectrum allocations, which are handled by national regulatory organizations. Several countries, most notably South Korea, have tied their technological identity to the increased connectivity of next-generation wireless technologies. All these factors give rise to a patchwork international environment for 5G implementation.

Current forecasts for the rollout of 5G networks are in the 2020 time frame. Recent news indicates that it may be earlier than that. Samples of Qualcomm's new Snapdragon X50 4G/5G modem that uses the 27.5 gigahertz (GHz) to 28.35 GHz band—part of the spectrum opened by the Federal Communications Commission for 5G—will be available in the second half of 2017. Samsung and KT, one of South Korea's mobile providers, have announced that they will be the first to provide mobile 5G trial service at the PyeungChang 2018 Winter Olympic Games. Undoubtedly, this trial 5G service will not be fully compliant with 5G standards, which have yet to be adopted. However, we should expect things to advance quickly following the Korean introduction of 5G.

RECENT NEWS

[1] "AT&T launches first 5G business customer trial with Intel and Ericsson." AT&T. 5 Dec 2016. Available at: http://about.att.com/story/att_launches_first_5g_business_customer_trial_with_intel_and_ericsson.html.

[2] "Government must take action now to secure our connected future so we are ready for 5G, and essential services are genuinely available where they are needed." GOV.UK. 14 Dec 2016. Available at: https://www.gov.uk/government/news/government-must-take-action-now-to-secure-our-connected-future-so-we-are-ready-for-5g-and-essential-services-are-genuinely-available-where-they-are-n.

[3] Qualcomm Incorporated. "Qualcomm, Ericsson and SK Telecom announce collaboration on 5G NR trials to accelerate wide-scale 5G deployments." *PR Newswire*. 19 Dec 2016. Available at: http://www.prnewswire.com/news-releases/qualcomm-ericsson-and-sk-telecom-announce-collaboration-on-5g-nr-trials-to-accelerate-wide-scale-5g-deployments-300381570.html.

[4] Reichert C. "Huawei and NTT DoCoMo reach 11Gbps speeds in 5G Japanese field trial. 16 Nov 2016. Available at: http://www.zdnet.com/article/huawei-and-ntt-docomo-reach-11gbps-speeds-in-5g-japanese-field-trial/.

Contents

2 Introducing 5G
Dr. Sam Farroha, Jared S. Everett, Jason J. Uher, Jason R. Harper, Jessica K. Bridgland, and Pamela M. Patton

16 Virtualizing the 5G Architecture
Staff Writer

20 5G and the Automotive Industry
Staff Writer

25 FROM LAB TO MARKET

The Next Wave is published to disseminate technical advancements and research activities in telecommunications and information technologies. Mentions of company names or commercial products do not imply endorsement by the US Government. The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the NSA/CSS.

This publication is available online at http://www.nsa.gov/thenextwave. For more information, please contact us at TNW@tycho.ncsc.mil.

Vol. 21 | No. 3 | 2017

Introducing 5G

Dr. Sam Farroha, Jared S. Everett (Johns Hopkins University Applied Physics Laboratory; JHU APL), Jason J. Uher (JHU APL), Jason R. Harper (JHU APL), Jessica K. Bridgland (JHU APL), and Pamela M. Patton (JHU APL)

[Photo credit: Juli_Rose/iStock/Thinkstock]

The evolution of modern cellular communications has been marked by a series of technology generations. Although the technology itself tends to evolve continuously, a new generation of standards marks a revolutionary step forward, with a substantial increase in system requirements to drive fundamentally new applications. With fourth-generation (4G) networks now widely deployed, the industry has turned its sights on "the next big thing." Fifth-generation (5G) networks are expected to enable a seamlessly connected society in the time frame beyond 2020 for both people and things, including vehicles, homes, smart cities, sensor networks, and the power grid. While the Long-Term Evolution (LTE) standard will continue to evolve and play a critical role in the wireless ecosystem, 5G represents an opportunity to architect a new system that is fundamentally different without the constraint of backward compatibility with existing technologies.

What is 5G?

Although there are as yet no standards for 5G mobile networks, a number of key technology trends have emerged. This article describes seven major technology trends that will pave the way to the next generation of 5G networks.

- **New Flexible Radio Access Technology (RAT):** A new, non-backward-compatible RAT will be defined for 5G that is distinct from previous generations, such as 4G LTE and its evolution. New multiple access schemes under consideration include various modified Orthogonal Frequency-Division Multiplexing (OFDM)-based solutions with improved spectral efficiency. The new RAT must be flexible enough to accommodate a variety of traffic types with often conflicting radio requirements. The concept of a unified air interface has been proposed for multiplexing multiple physical layer (PHY) regions with different characteristics [e.g., transmission time interval (TTI), subcarrier spacing] on a contiguous block of spectrum [1, 2, 3]. Spectrum for the new RAT will include existing bands below 6 gigahertz (GHz), as well as new centimeter-wave (cmWave) and millimeter-wave (mmWave) bands in the 6- to 100 GHz range [4]. The new RAT must also support significantly reduced latency, with as low as 100 microsecond (μs) transmission time interval (TTI) at the PHY for the ultra-reliable and low-latency communications (URLLC) use case [5]. Lastly, to further improve spectral efficiency, full-duplex transmission schemes have been proposed, potentially allowing the same time-frequency resources to be used for uplink and downlink transmissions simultaneously [6].
- **Virtualization:** Software-Defined Networking (SDN) and Network Function Virtualization (NFV) are two key architecture concepts in development to support the flexibility and mobility demands of the 5G network infrastructure [7, 8, 9]. Virtualization of network functions, which were traditionally implemented in hardware, will pave the way for commercial telecommunications operators and service providers to introduce new features and integrate new standards releases at an accelerated rate. NFV enables providers to move toward a decentralized network to increase flexibility, pushing core functions toward the edge to reduce latency, and virtualizing those functions on cloud-based servers. The proposed Cloud Radio Access Network (C-RAN) architecture, a specific use case of NFV applied to the RAN, uses a pooled architecture of baseband resources to increase scalability, physical layer flexibility, and spectral efficiency [7, 10].
- **Millimeter Wave (mmWave) Communications:** The term mmWave refers to carrier frequencies in the International Telecommunication Union (ITU) extremely high-frequency (EHF) band, from 30 to 300 GHz. Within the context

of 5G, the term has recently been loosely used by industry to refer to the higher frequencies from 6 to 100 GHz that are under consideration for new mobile spectrum [11, 12]. mmWave technologies are becoming an increasingly attractive solution to the problems of frequency reuse, cell density, raw data throughput, and antenna array size. This has led to a synergy between mmWave, small cell deployments, and massive multiple-input, multiple-output (MMIMO) techniques [13, 14].

- **Massive Multiple-Input, Multiple-Output (MMIMO) Techniques:** MMIMO is a new concept in antenna arrays that provides a number of advantages over traditional MIMO arrays currently deployed in 4G networks. Traditional MIMO arrays use only a few antenna elements (i.e., 2 to 16), whereas MMIMO uses a large number of elements in the array, currently considering a range of 128 to 512 at a minimum. Highly directional beamforming to multiple users simultaneously allows for increased user density and higher aggregate cell throughput [15]. So-called hybrid MMIMO has also been proposed; it combines beam steering with array processing techniques, such as spatial multiplexing, to increase single-user throughput [16].
- **Heterogeneous Networks (HetNets):** HetNets expand the mobile access network capacity by coordinating small cells with larger macro cells or offloading traffic to wireless local area network (WLAN) access points. There are two types of heterogeneity: 1) various cell sizes (e.g., macro, pico, femto) and 2) heterogeneous RATs [e.g., third-generation (3G), 4G, 5G, WLAN]. Small cells may include femto, pico, and micro cells, which can range in capacity from less than 10 to several hundred simultaneous active users. While HetNet deployments have already been introduced in 4G networks, network densification through the aggressive deployment of small cells is expected to increase significantly in future 5G networks [17, 18].
- **Native Machine-Type Communications (MTC) Support:** 5G networks are expected to incorporate a new model for connectivity specifically designed for MTC [19]. With the significant increase in connected machines over the last several years, a new 5G standard is seen as a prime opportunity to ensure new RATs can efficiently support a large number of connected devices with their own unique access constraints. Two categories of MTC are discussed: 1) general MTC and 2) vehicle-to-everything (V2X) MTC. General MTC devices have a few unique design and deployment considerations—namely, lower bandwidth needs, stringent power budgets, and relaxed latency requirements. V2X MTC devices, in contrast, require low-latency communications, out-of-coverage networks, and limited operation on a subscription-free basis [20].
- **Device-centric Architectures:** New network architectures will focus on a uniform quality of experience (QoE) for the user device, in contrast to traditional base-station-centric architectures. A number of new device-centric approaches are under consideration: decoupling the user plane and control plane, decoupling the uplink and downlink, and device-to-device communications [4, 21]. Another novel proposal is the user-centric cell or virtual cell model, which uses distributed beamforming and decoupled user/control planes to create a virtual cell around each user [2, 3]. Because the virtual cell follows the user, QoE variations are reduced and the cell-edge problem is mitigated. New device-centric architectures may significantly alter the traditional concept of cell handovers or eliminate it entirely.

Many of these technologies are already being added to the evolution of existing technologies beyond 4G, such as LTE-Advanced Pro [22].

In 2015, there was a significant increase in industry activities surrounding 5G networks. Major standards bodies, including the ITU and the Third-Generation Partnership Project (3GPP), reached important milestones in the early development of the eventual 5G standards. In September, the ITU published its vision for 5G networks [19]. The vision for International Mobile Telecommunications for 2020 and beyond (IMT-2020) defines three future-looking, high-level use cases for 5G:

- **Enhanced Mobile Broadband (eMBB):** This is generally a human-centric use case driven by the exponential increase in demand for mobile access to multimedia content, services, and data.

The eMBB use case will come with new application areas and requirements that go beyond existing mobile broadband applications for improved performance and increasingly seamless user experience. This use case covers a range of scenarios, including wide-area coverage and localized high-throughput spot coverage, which will have different requirements.

- **Massive Machine-Type Communications (mMTC):** This use case is characterized by a large number of connected devices typically transmitting a relatively low volume of nondelay-sensitive data. Devices are intended to be low cost and have a very long battery life.
- **Ultra-Reliable and Low-Latency Communications (URLLC):** This use case is characterized by stringent requirements for latency, throughput, and availability. Examples include wireless control of industrial manufacturing processes, remote medical surgery, distributed smart grid automation, and transportation safety [e.g., vehicle-to-vehicle (V2V) or vehicle-to-everything (V2X) communication]. Many companies have referred to this use case as critical MTC (cMTC) or ultra-reliable MTC (uMTC). However, based on the ITU definition in [19], this use case is not strictly limited to MTC applications.

It is important to consider that the applications that will use 5G technology do not necessarily correspond to a single use case but are more accurately described as a combination of multiple use cases. Figure 1 illustrates some examples of currently envisioned 5G applications and their relationship to these three IMT-2020 use cases [19]. Figure 2 illustrates eight key capabilities identified by ITU for IMT-2020 and their relative importance to the same three use cases [19]. Furthermore, additional future use cases are expected to emerge but cannot be accurately predicted (i.e., what will be the "killer app" in 2025?). Therefore, it is desired that 5G standards will provide the flexibility to adapt to new use cases.

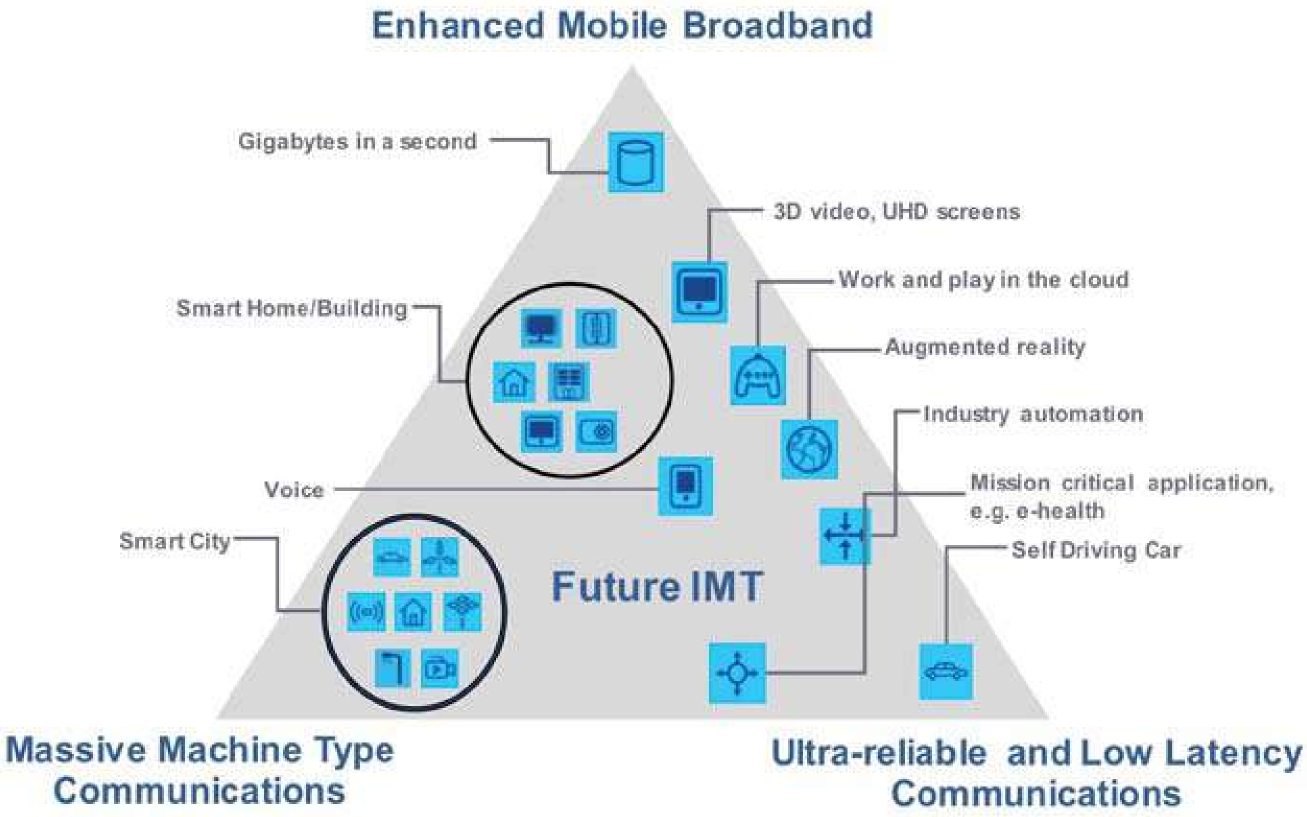
Enhanced Mobile Broadband
Gigabytes in a second
3D video, UHD screens
Work and play in the cloud
Smart Home/Building
Augmented reality
Industry automation
Voice
Mission critical application, e.g. e-health
Smart City
Self Driving Car
Future IMT
Massive Machine Type Communications
Ultra-reliable and Low Latency Communications

FIGURE 1. 5G use cases as defined by ITU for IMT-2020 [19].

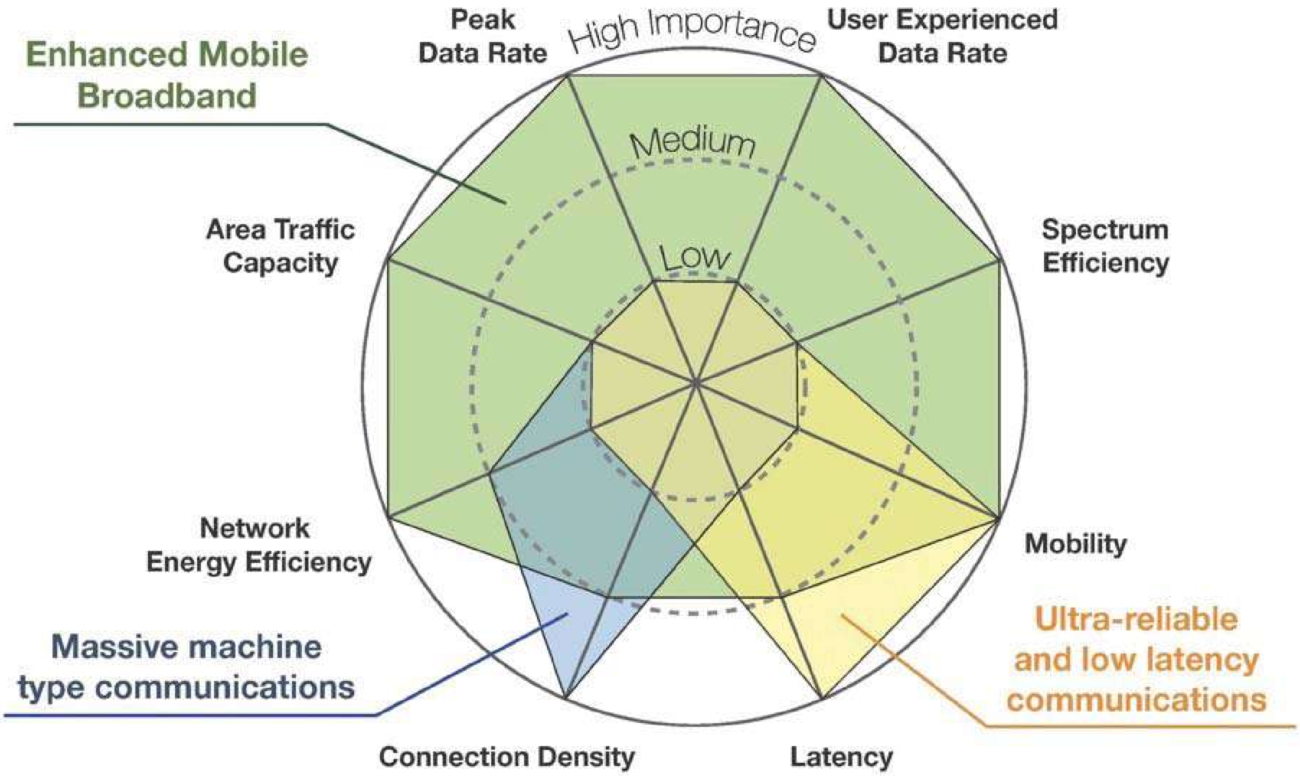

FIGURE 2. Eight key capabilities and their relative importance to 5G use cases [19].

5G Standardization

Development work toward 5G is well under way. Standards bodies are actively working on new 5G mobile technologies to be deployed in the 2020 time frame. This section summarizes the activities and corresponding 5G development timelines for three major standards bodies: ITU, 3GPP, and the Institute of Electrical and Electronics Engineers (IEEE).

Standardization in ITU

ITU is the United Nations agency responsible for promoting worldwide improvement and rational use of information and communication technology. Its members include industry, academia, and standards organizations from more than 190 member nations. The ITU Radiocommunication Sector (ITU-R) works toward worldwide consensus in the use of terrestrial and space radiocommunication services, including mobile communication technologies. Although compliance with ITU-R recommendations is not mandatory, they nevertheless have a high degree of adoption worldwide and hold the status of international standards [23].

International Mobile Telecommunications framework

ITU-R Working Party 5D (WP 5D) is responsible for overall radio aspects of terrestrial mobile systems, referred to as International Mobile Telecommunications (IMT). The purpose of IMT is to provide high-quality mobile services with a high degree of interoperability worldwide. Since 2000, the ITU has developed the IMT standards framework in a manner that parallels cellular generations from an industry perspective. Although ITU-R WP 5D defines the requirements for IMT, it does not develop the actual radio technologies. Rather, candidate radio technologies are submitted for inclusion by external standards bodies, such as 3GPP and IEEE. For this reason, ITU-R WP 5D maintains strong cooperation with the major global standards bodies.

The first family of standards derived from the IMT concept (IMT-2000) aligned with 3G cellular. Radio technologies accepted into IMT-2000 included 3GPP Wideband Code-Division Multiple Access (WCDMA), 3GPP2 cdma2000, and IEEE 802.16 [i.e., Mobile Worldwide Wireless Interoperability for Microwave Access (WiMAX)]. The next generation of IMT standards (IMT-Advanced) aligned with 4G cellular. Radio technologies accepted into IMT-Advanced included 3GPP LTE-Advanced and IEEE 802.16m [i.e., Wireless Metropolitan Area Network (WMAN)-Advanced].

Timeline for IMT-2020

In 2012, ITU embarked on a program to develop "IMT for 2020 and beyond," setting the stage for emerging 5G research activities around the world. The program has since adopted the name IMT-2020 and forms the framework for the next generation of mobile broadband standards. The timeline for the development of IMT-2020 is shown in figure 3 [24]. The IMT 2020 timeline will essentially follow the same process used in the development of IMT-Advanced.

The IMT-2020 program is well under way, with a number of key milestones completed. In September 2015, ITU published its vision of the 5G mobile broadband connected society [19]. This document defined three high-level use cases for 5G, described earlier in this article, which have already been widely adopted by 3GPP and industry in general. In the next phase, the 2016–2017 time frame, ITU-R WP 5D will define in detail the performance requirements, evaluation criteria, and methodology for the assessment of the new IMT radio interfaces. It is anticipated that the time frame for proposals will be focused in 2018. In the 2018–2020 time frame, independent, external groups will evaluate proposals and the definition of the new radio interfaces to be included in IMT-2020 will take place. ITU-R WP 5D also plans to hold a workshop in late 2017 to discuss the performance requirements and evaluation criteria for candidate technologies for IMT-2020, as well as to provide an opportunity for presentations by potential proponents for IMT-2020 in an informal setting. The whole process is planned to be completed in 2020, when a new draft of the ITU-R recommendation with detailed

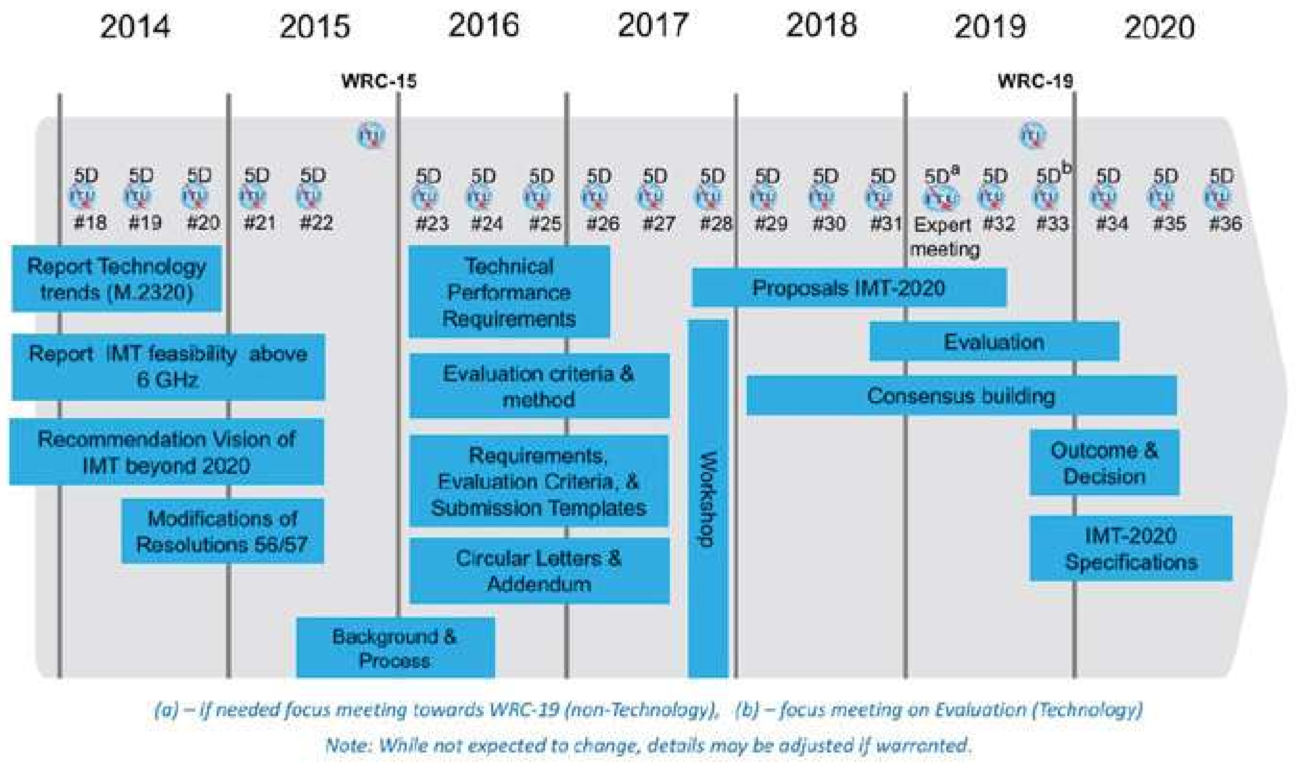
2014
2015
2016
2017
2018
2019
2020
WRC-15
WRC-19
5D #18
5D #19
5D #20
5D #21
5D #22
5D #23
5D #24
5D #25
5D #26
5D #27
5D #28
5D #29
5D #30
5D #31
5D^a Expert meeting
5D #32
5D^b #33
5D #34
5D #35
5D #36
Report Technology trends (M.2320)
Report IMT feasibility above 6 GHz
Recommendation Vision of IMT beyond 2020
Modifications of Resolutions 56/57
Background & Process
Technical Performance Requirements
Evaluation criteria & method
Requirements, Evaluation Criteria, & Submission Templates
Circular Letters & Addendum
Workshop
Proposals IMT-2020
Evaluation
Consensus building
Outcome & Decision
IMT-2020 Specifications
(a) – If needed focus meeting towards WRC-19 (non-Technology), (b) – focus meeting on Evaluation (Technology)
Note: While not expected to change, details may be adjusted if warranted.

FIGURE 3. Detailed timeline and process for IMT-2020 in ITU-R [24].

specifications for the new radio technologies will be submitted for approval within ITU-R [24].

Standardization in 3GPP

3GPP is the international standards body responsible for the development and maintenance of major second-generation (2G), 3G, and 4G cellular standards. The purpose of the organization is to produce interoperable cellular communications standards, as well as studies and reports that define 3GPP technologies. The following technologies are currently maintained and evolved by 3GPP:

- Global System for Mobile Communications (GSM), General Packet Radio Service (GPRS), and Enhanced Data Rates for GSM Evolution (EDGE);
- Universal Mobile Telecommunications System (UMTS), WCDMA, High-Speed Packet Access (HSPA), and HSPA Evolution (HSPA+); and
- LTE, LTE-Advanced, and LTE-Advanced Pro.

These 3GPP technologies are constantly evolving through a series of backward-compatible releases. Since the completion of the first LTE and Evolved Packet Core (EPC) specifications, 3GPP has become the focal point for mobile systems beyond 3G. Therefore, 3GPP is expected to be a critical player in the development of 5G, and their timeline will have a direct influence on the timeline of the emerging 5G market.

3GPP is currently defining a new 5G RAT and corresponding network architecture. These are being developed within 3GPP under the working names “new radio (NR)” and “next-generation (NextGen) architecture,” respectively [25]. In October 2016, 3GPP announced that the new 3GPP system will officially be known by the name “5G” from Release 15 onward [26]. Some initial standardization steps that have been taken to date include the following:

- **SMARTER study item:** In March 2015, 3GPP Technical Specification Group (TSG) System Aspects (SA) began a study item on technology enablers for new 5G services and markets, known as the SMARTER study item [27]. The objective of this study was to develop high-level use cases and identify the related high-level potential requirements to enable 3GPP network operators to support new services and markets in 5G. Phase 1 of the SMARTER study item was completed in March 2016; results are documented in 3GPP Technical Report (TR) 22.891 to be included in Release 14 [28]. A total of 74 use cases were identified. This work prompted four building block studies that grouped the use cases into families with common requirements: massive Internet of Things (IoT), critical communications, eMBB, and network operation. The building block studies were completed in June 2016; results are documented in 3GPP TRs 22.861, 22.862, 22.863, and 22.864 to be included in Release 14 [29]. The results of the SMARTER study will form the basis for a work item to define normative stage 1 requirements for the next-generation 5G system. The work item is scheduled for completion in March 2017; results will be documented in 3GPP Technical Specification (TS) 22.261 to be included in Release 15 [29].
- **Study item on channel model for frequency spectrum above 6 GHz:** The first 5G study conducted by TSG RAN focused on developing new channel models to support high-frequency spectrum from 6 GHz to 100 GHz. The models consider a variety of scenarios including urban, rural, and indoor, as well as the impact of line-of-sight (LOS) versus non-LOS (NLOS). The study was completed in June 2016, and results are documented in 3GPP TR 38.900 to be included in Release 14 [30].
- **Study item on architecture for next-generation system:** In December 2015, 3GPP TSG SA approved a study item to design a system architecture for the next generation of mobile networks. The new architecture will support at least the new 5G RAT(s), the evolution of LTE, and non-3GPP access types and will minimize access dependencies. The study considers new approaches such as NFV and network slicing. The study item was scheduled for completion in December 2016; results will be documented in 3GPP TR 23.799 to be included in Release 14 [31].
- **Study item on scenarios and requirements for next-generation access technologies:** In December 2015, 3GPP TSG RAN approved a study item to develop deployment scenarios and requirements of next-generation access

technologies. The study identifies 12 deployment scenarios that are more diverse than those originally envisioned for legacy RATs, such as LTE and its predecessors. It also identifies key performance indicators (KPIs) and other requirements for 5G NR. The bulk of the study was completed in September 2016 to provide guidance to the ongoing technical work being performed in the RAN working groups. However, the study item will remain open until March 2017 to match the IMT-2020 timeline and ensure all IMT-2020 requirements are captured. Final results are documented in 3GPP TR 38.913 to be included in Release 14 [32, 4].

- **Study item on NR access technology:** In March 2016, 3GPP TSG RAN approved a study item to develop the 5G NR access technology capable of meeting the broad range of use cases defined for 5G. The study seeks to develop a single technical framework capable of addressing all usage scenarios and requirements defined in TR 38.913 for eMBB, mMTC, and URLLC, with an emphasis on forward compatibility. The study is scheduled for completion in March 2017; results will be documented in 3GPP TR 38.912 to be included in Release 14 [1].

5G standardization activities in 3GPP will continue through 2020 and beyond, as described next.

Emerging 3GPP standardization timeline

In March 2015, 3GPP announced a tentative standardization timeline for 5G based on the ITU work plan timeline for IMT-2020 [33]. Since then, a more detailed timeline has come into focus as study items have commenced and completed, and as the 3GPP TSGs coordinate for the initial release of 5G. The timeline shown in figure 4 is a composite from multiple sources.

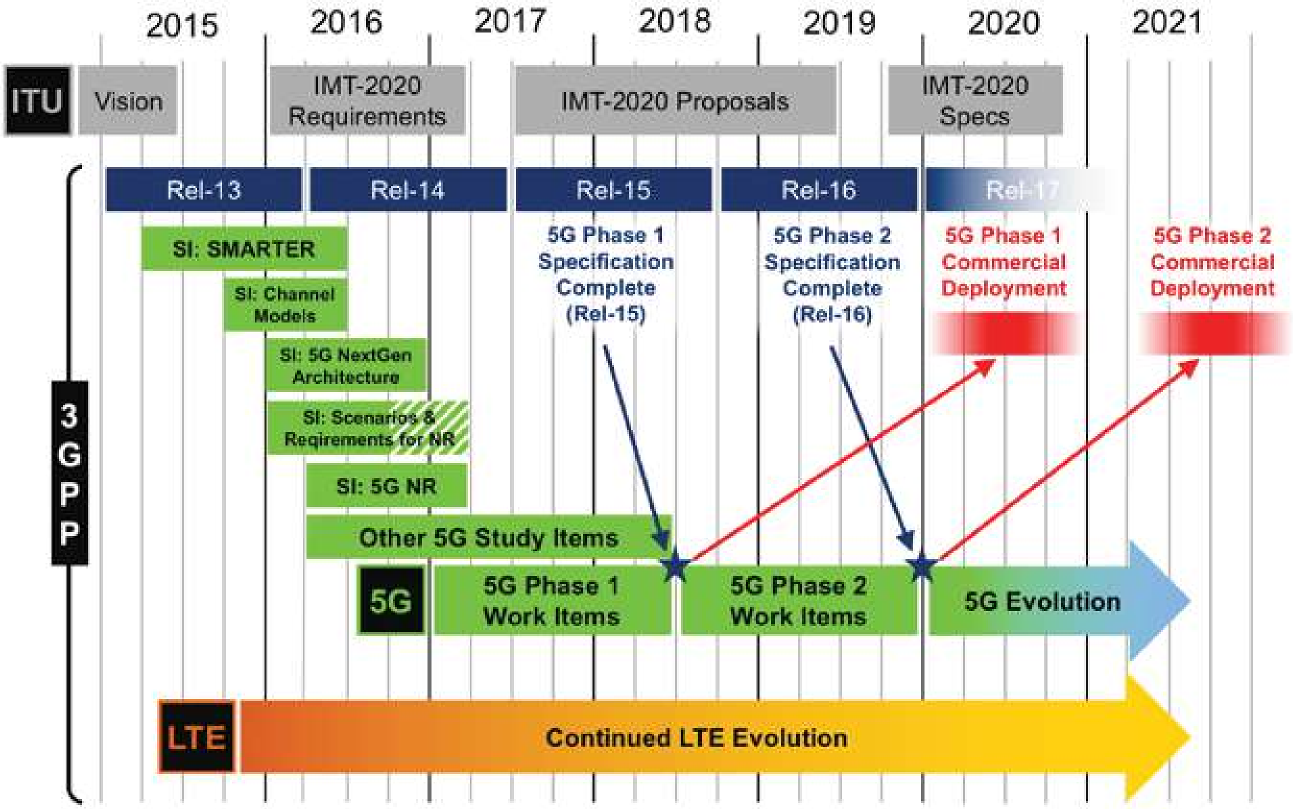
2015
2016
2017
2018
2019
2020
2021
ITU
Vision
IMT-2020 Requirements
IMT-2020 Proposals
IMT-2020 Specs
3GPP
Rel-13
Rel-14
Rel-15
Rel-16
Rel-17
SI: SMARTER
SI: Channel Models
SI: 5G NextGen Architecture
SI: Scenarios & Reqirements for NR
SI: 5G NR
Other 5G Study Items
5G Phase 1 Specification Complete (Rel-15)
5G Phase 2 Specification Complete (Rel-16)
5G Phase 1 Commercial Deployment
5G Phase 2 Commercial Deployment
5G
5G Phase 1 Work Items
5G Phase 2 Work Items
5G Evolution
LTE
Continued LTE Evolution

FIGURE 4. Emerging 5G standardization timeline for 3GPP. (Figure is a composite from [34] and [35] that includes additional data from various sources.)

The initial 5G study items in TSG SA and TSG RAN commenced in 2015 and 2016, as described previously. These initial 5G study items will be included in 3GPP Release 14. This work is carried out in parallel with ongoing LTE work. At the 3GPP plenary meeting in June 2016, the TSGs agreed on a work plan for the first release of 5G in 3GPP Release 15, including a clear work division between the TSGs [25].

5G work items were scheduled to begin in December 2016 for TSG SA and March 2017 for TSG RAN. The Phase 1 5G work items will fall into Release 15, with planned completion in June 2018. Additional 5G study items will continue during Release 15 in support of Phase 2. Subsequently, the Phase 2 5G work items will fall into Release 16, which will be completed around December 2019 in time for the final submission to ITU for IMT-2020. Phase 1 commercial deployments are expected to begin in 2020, followed by Phase 2 deployments in the 2021–2022 time frame. However, early pre-5G mmWave deployments may emerge in limited markets, such as South Korea or the United States, before 2020.

For example, Verizon Wireless has announced its plans to pilot a 28-GHz mmWave deployment in the United States for fixed wireless applications starting in 2017 [36]. To support this effort, the Verizon 5G Technology Forum (V5GTF)—an industry consortium led by Verizon—published an open radio interface specification in July 2016 [37]. The Verizon specification uses an OFDM-based PHY similar to time-division LTE (TD-LTE) with enhanced beamforming for operation in 28 and 39 GHz mmWave spectrum. However, with the initial focus on fixed wireless, the first release does not support user mobility. The Verizon specification can be considered pre-5G in the sense that it supports new mmWave capabilities beyond 4G but does not address all the use cases and associated requirements for 5G. The Verizon specification is expected to be incompatible with the eventual 3GPP 5G standard, potentially leading to market fragmentation [38].

3GPP phased approach to 5G standardization

3GPP TSG RAN will take a two-phased approach to developing the new 5G RAT [35]. The Phase 1 standard will define a new, non-backward-compatible 5G RAT. A subset of prioritized features and use cases will be addressed in Phase 1 to allow for early commercial deployments targeted for the year 2020. The Phase 2 standard will implement the full set of features and use cases necessary to meet the requirements for 5G. An initial proposal will be submitted to ITU as a candidate radio interface technology for IMT-2020 by the June 2019 submission deadline. The Phase 2 standard will later form the final submission around December 2019. The Phase 1 standard will be designed for forward compatibility with Phase 2 [35]. Forward compatibility means that Phase 1 must be designed from the beginning to optimally accommodate all of the features and use cases expected to be added later in Phase 2, even though those features are not yet fully implemented. Although the forward-compatibility requirement may sound straightforward, it represents a fundamental shift from the normal 3GPP standardization process, which historically has focused on a series of backward-compatible releases.

While prioritization of features between the two phases has been a topic of much debate, it is clear that the 5G Phase 1 standard will support tight interworking with LTE to simplify initial rollout. The phased approach and tight interworking with LTE means that elements of the LTE system architecture may persist in 5G deployments for some time to come. This implies that current and future work on LTE, LTE-Advanced, and LTE-Advanced Pro networks and technologies may have direct applicability to eventual 5G network deployments.

Standardization in IEEE

Initial 5G standards activities within the IEEE suggest that they do not intend to be a direct competitor with organizations like 3GPP on the radio interface between the RAN and the user equipment. Instead, IEEE has begun developing complementary technologies to support other communications requirements within the 5G ecosystem. In 2016, IEEE established two new working groups related to 5G: IEEE 1914 and IEEE 1918.

IEEE 1914 is the Next Generation Fronthaul Interface Working Group. This working group is currently developing two standards: the 1914.1 standard for packet-based fronthaul transport networks and the 1914.3 standard for radio over Ethernet encapsulations and mappings [39, 40]. These standards focus on the fronthaul interface within the RAN between baseband units (BBUs) and remote radio heads (RRHs) to support novel RAN architectures like C-RAN, and antenna techniques like MMIMO and coordinated multi-point (CoMP) transmission and reception. The projected completion dates for these standards are August 2018 for 1914.1 and October 2017 for 1914.3.

IEEE 1918 is the Tactile Internet Working Group. This working group is currently developing the 1918.1 standard, which defines a framework for the Tactile Internet [41]. The purpose of this framework is to establish a basis for the rapid development of the Tactile Internet as a 5G and beyond application, with the expectation of additional IEEE 1918 standards to follow. The projected completion date for the 1918.1 standard is October 2018.

With respect to IMT-2020, IEEE may seek to expand the role of WLAN in 5G as a complementary radio interface for next-generation HetNets. In September 2016, the IEEE 802.11 working group sent a liaison statement to 3GPP TSG RAN and TSG SA inviting them to consider the use of IEEE 802.11-based WLAN in unlicensed spectrum as a complementary means of meeting the performance requirements of IMT-2020, potentially leading to inclusion in a joint submission to IMT-2020 [42]. This approach would be a logical extension of the increasing level of interworking between LTE and WLAN in recent standards releases. WLAN is already widely used in 3GPP networks for high data rate offloading.

Recent enhancements in radio-level interworking have increased the efficiency of these networks. Enhancements include LTE-WLAN Aggregation (LWA) and LTE WLAN Radio Level Integration with IPsec Tunnel (LWIP) in 3GPP Release 13, with further enhancements in 3GPP Release 14. Although 3GPP declined to make a decision at the September 2016 plenary meeting, the concept of a potential joint submission could represent a novel approach to IMT-2020. In contrast, previous generations of IMT saw IEEE in competition with 3GPP, with the submission of the IEEE 802.16 WiMAX family of standards to IMT-2000 and IMT-Advanced as a direct competitor in the 3G and 4G markets.

Conclusion

This article provided an introduction to major technology trends in the emergence of next-generation 5G mobile networks. These networks are expected to see initial commercial deployment starting around the year 2020. Early 5G standardization activities in the ITU, 3GPP, and IEEE were addressed.

For further information on the latest developments in 5G, the interested reader is directed to the following resources. ITU publications for IMT-2020 can be found on the IMT-2020 web page [24]. Notable documents include the ITU vision for IMT-2020 [19] as well as the technical performance requirements for IMT-2020 (scheduled for completion in February 2017). The 3GPP web site (www.3GPP.org) is the most direct source for 3GPP-related technical information. 3GPP press releases provide high-level summaries of ongoing standards activities and often include links to more detailed further reading. The latest versions of the 3GPP TR and TS documents mentioned in this article can be accessed there as well. Information on the IEEE 1914.1, 1914.3, and 1918.1 standards can be found in the approved project authorization request (PAR) documents [39, 40, 41] and the corresponding working group web pages. Lastly, the Verizon mmWave specification is available on the V5GTF website (www.5GTF.org). This pre-5G specification defines layers 1 to 3 of an open radio interface using a document structure similar to that of LTE.

References

[1] 3GPP TSG RAN. "Revision of SI: Study on new radio access technology." September 2016. 3GPP Document TD RP-161596. Available at: http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_73/Docs/RP-161596.zip.

[2] Huawei. "Vision on 5G radio access technologies." September 2015. 3GPP Document RWS 150006. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150006.zip.

[3] ZTE. "Considerations on 5G key technologies and standardization." September 2015. 3GPP Document dRWS 150024. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150024.zip.

[4] 3GPP. "3GPP TSG RAN study on scenarios and requirements for next generation access technologies." October 2016. 3GPP Specification TR 38.913, Release 14, Version 14.0.0. Available at: http://www.3gpp.org/DynaReport/38913.htm.

[5] Fettweis G, Alamouti S. "5G: Personal mobile internet beyond what cellular did to telephony." *IEEE Communications Magazine*. February 2014; 52(2): 140-145. doi: 10.1109/MCOM.2014.6736754.

[6] Zhang Z, Chai X, Long K, Vasilakos AV, Hanzo L. "Full duplex techniques for 5G networks: Self-interference cancellation, protocol design, and relay selection." *IEEE Communications Magazine*. May 2015; 53(5): 128-137. doi: 10.1109/MCOM.2015.7105651.

[7] ITU-R. "Future technology trends of terrestrial IMT systems." November 2014. ITU-R Report M.2320-0. Available at: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2320-2014-PDF-E.pdf.

[8] 4G Americas. "5G: 4G Americas' recommendations on 5G requirements and solutions." October 2014. Available at: http://www.4gamericas.org/files/2714/1471/2645/4G_Americas_Recommendations_on_5G_Requirements_and_Solutions_10_14_2014-FINALx.pdf.

[9] 3GPP. "3GPP NFV study: Management of mobile networks that include virtualized network functions (MANO)" [Press release]. 5 Oct 2015. Available at: http://www.3gpp.org/news-events/3gpp-news/1738-sa5_nfv_study.

[10] Checko A, Christiansen H, Yan Y, Scolari L, Kardaras G, Berger M, Dittman L. "Cloud RAN for mobile networks—A technology overview." *IEEE Communication Surveys and Tutorials*. January 2015; 17(1): 405-426. doi: 10.1109/COMST.2014.2355255.

[11] ITU-R. "Technical feasibility of IMT in bands above 6 GHz." June 2015. ITU-R Report M.2376-0. Available at: www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2376-2015-PDF-E.pdf.

[12] Rangan S, Rappaport T, Erkip E. "Millimeter-wave cellular wireless networks: Potentials and challenges." In: *Proceedings of the IEEE;* 2014 January; 102(3). doi: 10.1109/JPROC.2014.2299397.

[13] Swindlehurst AL, Ayanoglu E, Heydari P, Capolino F. "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Communications Magazine*. September 2014; 52(9): 56-62. doi: 10.1109/MCOM.2014.6894453.

[14] Straight Path Communications. "A straight path toward 5G." September 2015. 3GPP Document RWS150087. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS150087.zip.

[15] Hoydis J, ten Brink S, Debbah M. "Massive MIMO: How many antennas do we need?" In: *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*; 2011 Sep 27-28; Monticello, IL; p. 545–550. doi: 10.1109/Allerton.2011.6120214.

[16] InterDigital. "Roadmap for 5G radio standardization in 3GPP." September 2015. 3GPP Document RWS 150030. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150030.zip.

[17] Wannstrom J, Mallinson K. "Heterogeneous networks in LTE." Available at: http://www.3gpp.org/technologies/keywords-acronyms/1576-hetnet.

[18] Bangerter B, Talwar S, Arefi R, Stewart K. "Networks and devices for the 5G era." *IEEE Communications Magazine*. February 2014; 52(2): 90–96. doi: 10.1109/MCOM.2014.6736748.

[19] ITU-R. "IMT vision—Framework and overall objectives of the future development of IMT for 2020 and beyond." September 2015. ITU-R Recommendation M.2083-0. Available at: http://www.itu.int/rec/R-REC-M.2083.

[20] General Motors Research. "Towards 5G: An automotive perspective." September 2015. 3GPP Document RWS 150011. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150011.zip.

[21] Boccardi F, Heath Jr. R, Lozano A, Marzetta T, Popovski P. “Five disruptive technology directions for 5G.” *IEEE Communications Magazine*. February 2014; 52(2): 74-80. doi: 10.1109/MCOM.2014.6736746.

[22] 3GPP. “LTE-advanced pro ready to go” [Press release]. 28 Oct 2015. Available at: http://www.3gpp.org/news-events/3gpp-news/1745-lte-advanced_pro.

[23] ITU-R. “ITU-R study groups.” May 2013. Available at: http://www.itu.int/dms_pub/itu-r/opb/gen/R-GEN-SGB-2013-PDF-E.pdf.

[24] ITU-R. “ITU toward ‘IMT for 2020 and beyond.’” Available at: http://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Pages/default.aspx.

[25] 3GPP. “3GPP on track to 5G” [Press release]. 27 Jun 2016. Available at: http://www.3gpp.org/news-events/3gpp-news/1787-ontrack_5g.

[26] 3GPP. “New partners at London 3GPP meeting” [Press release]. 24 Oct 2016. Available at: http://www.3gpp.org/news-events/3gpp-news/1800-pcg_37.

[27] 3GPP TSG SA. “New WID study on new services and markets technology enablers (FS_SMARTER) from S1-150300.” March 2015. 3GPP Document SP-150142. Available at: ftp://ftp.3gpp.org/tsg_sa/TSG_SA/TSGS_67/Docs/SP-150142.zip.

[28] 3GPP. “3GPP TSG SA feasibility study on new services and markets technology enablers stage 1.” June 2016. 3GPP Specification TR 22.891, Release 14, Version 14.1.0. Available at: http://www.3gpp.org/DynaReport/22891.htm.

[29] 3GPP. “SA1 completes its study into 5G requirements” [Press release]. 23 Jun 2016. Available at: http://www.3gpp.org/news-events/3gpp-news/1786-5g_reqs_sa1.

[30] 3GPP. “3GPP TSG RAN study on channel model for frequency spectrum above 6 GHz.” June 2016. 3GPP Specification TR 38.900, Release 14, Version 14.0.0. Available at: http://www.3gpp.org/DynaReport/38900.htm.

[31] 3GPP TSG SA. “New WID study on architecture for next generation system.” December 2015. 3GPP Document TD SP-150853. Available at: http://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_70/Docs/SP-150853.zip.

[32] 3GPP TSG RAN. “New study item proposal: Study on scenarios and requirements for next generation access technologies.” December 2015. 3GPP Document TD RP-152257. Available at: http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_70/Docs/RP-152257.zip.

[33] Flore D, Bertenyi B. “Tentative 3GPP timeline for 5G” [Press release]. 17 Mar 2015. Available at: http://www.3gpp.org/news-events/3gpp-news/1674-timeline_5g.

[34] Qualcomm. “5G: Views on technology and standardization.” September 2015. 3GPP Document RWS 150012. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150012.zip.

[35] Flore D. “RAN workshop on 5G: Chairman summary.” September 2015. 3GPP Document RWS 150073. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150073.zip.

[36] Alleven M. “Verizon’s Shammo: 5G pilot in 2017 is all about fixed wireless, not mobility.” *Fierce Wireless*. 2016 Apr 21. Available at: http://www.fiercewireless.com/tech/story/verizons-shammo-5g-pilot-2017-all-about-fixed-wireless-not-mobility/2016-04-21.

[37] V5GTF. “V5GTF Network and Signaling Working Group; Verizon 5G radio access; Overall description.” June 2016. V5GTF Specification TS V5G.300, Release 1, Version 1.0. Available at: http://www.5gtf.org/V5G_300_v1p0.pdf.

[38] Alleven M. “Verizon’s version of 5G not compatible with 3GPP’s current specs—or easily upgradeable: Report.” *Fierce Wireless*. 2016 Oct 28. Available at: http://www.fiercewireless.com/tech/verizon-s-version-5g-not-compatible-3gpp-s-current-specs-or-easily-upgradeable-report.

[39] IEEE. “IEEE P1914.1: Standard for packet-based fronthaul transport networks.” February 2016. IEEE Approved PAR P1914.1. Available at: http://sites.ieee.org/sagroups-1914/files/2016/06/P1914.1.pdf.

[40] IEEE. “IEEE P1904.3: Standard for radio over ethernet encapsulations and mappings.” June 2016. IEEE Approved PAR P1904.3. Available at: http://sites.ieee.org/sagroups-1914/files/2016/07/P1904.3.pdf.

[41] IEEE. “IEEE P1918.1: Tactile Internet: Application scenarios, definitions and terminology, architecture, functions, and technical assumptions.” March 2016. IEEE Approved PAR P1918.1. Available at: http://grouper.ieee.org/groups/1918/1/files/P1918.1.pdf

[42] 3GPP TSG RAN. “IEEE 802.11 Working group liaison on the role of WLAN in IMT-2020.” September 2016. 3GPP Document TD RP-161804. Available at: http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_73/Docs/RP-161804.zip.

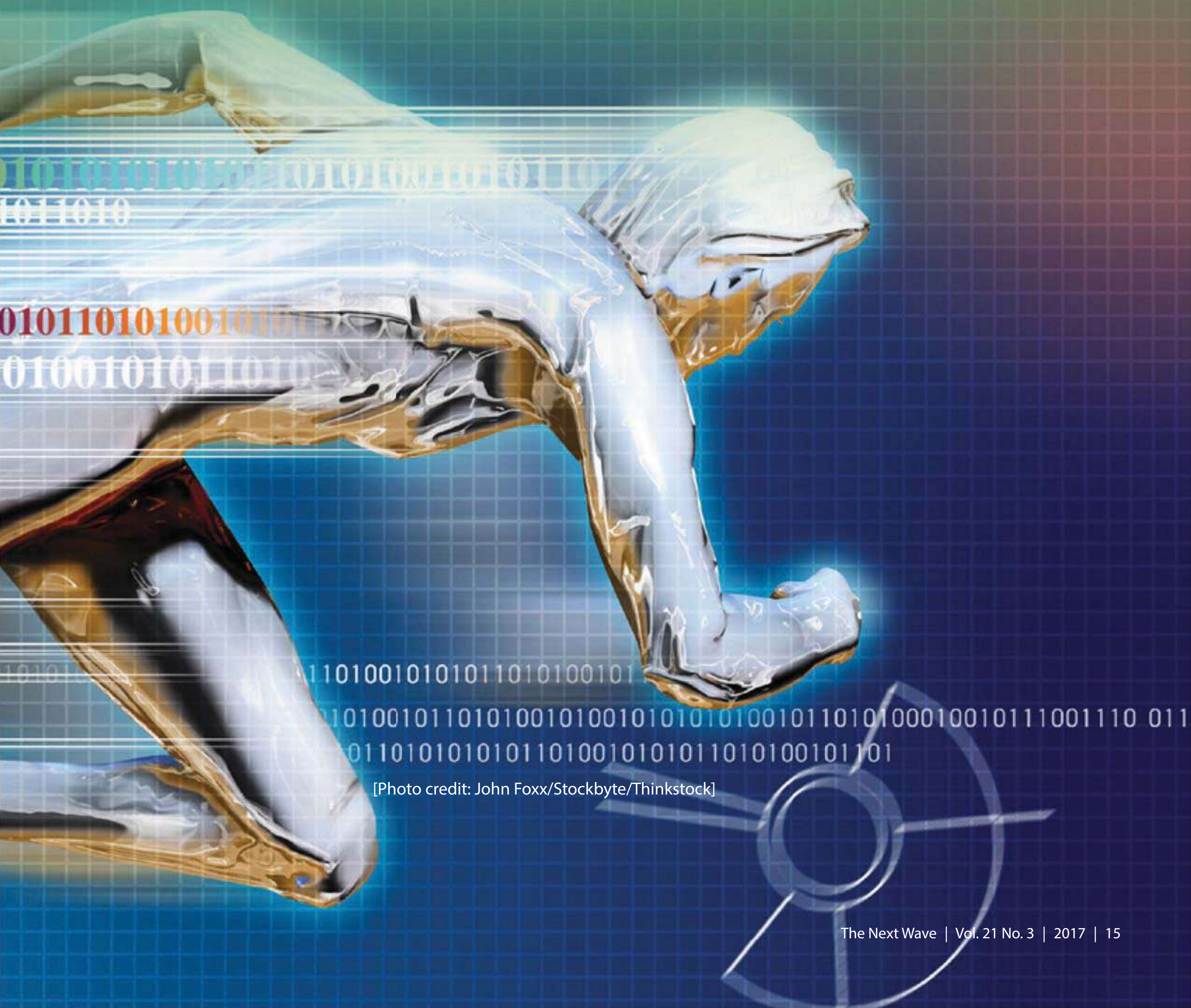

[Photo credit: John Foxx/Stockbyte/Thinkstock]

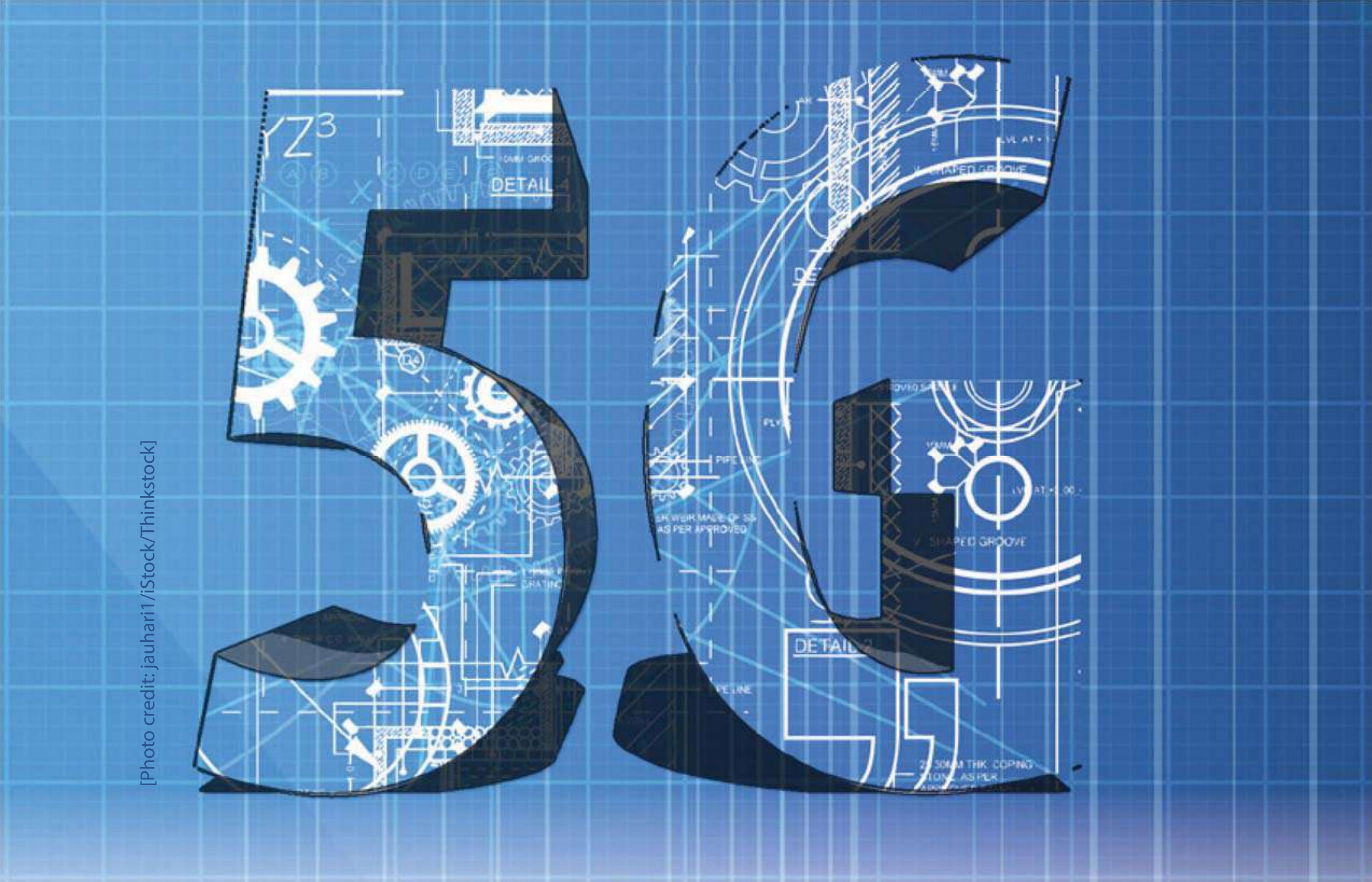

[Photo credit: jauhari1/iStock/Thinkstock]

Virtualizing the 5G Architecture

Staff Writer

Virtually changing the 5G architecture

Virtualization is set to play a major role in the evolution of the fifth-generation (5G) core network. According to industry experts, 5G will use software-centric networking technologies such as software-defined networking (SDN) and network functions virtualization (NFV), and will be natively cloud based. If correct, this will represent a major transition in system architecture and will require much greater collaboration across the networking ecosystem. The push to incorporate more cloud- or software-based components is driven by the need for greater flexibility and scalability to respond to the demands of radio access technologies that offer more bandwidth, reduced latency, and stringent quality of service (QoS) requirements. The new 5G core network must be adaptable and better equipped to handle various devices and manage capacity in near-real time. Mobile network operators see the advances in cost and efficiency that virtualization brings to other market segments and will use the emerging 5G technology to determine if these same advances can benefit the mobile market [1, 2].

SDN and NFV are two key architecture concepts in development to support the flexibility and mobility demands of the 5G network infrastructure. Virtualizing network functions that were previously implemented in hardware will allow providers to introduce new features and integrate new standards at a faster rate. SDN/NFV provides an avenue for providers to decentralize their networks, thereby increasing flexibility and reducing latency. Two areas where SDN/NFV will benefit 5G networks, and in some cases even fourth-generation (4G) mobile technology, are network slicing and cloud-radio access network (C-RAN).

Network slicing

Network slicing would promote end-to-end mobile network virtualization by "slicing" the network into virtual channels. These virtual channels would be autonomous and encompass a set of resources—physical or virtual—including bandwidth on a network link, processing capacity of servers, processing capacity of network elements, as well as operations support system (OSS) and business support system (BSS) processes. Operators could then use these channels to dynamically devote the appropriate network resources to create a "lane" in the network specifically designed for a particular use or service. This would accommodate the many use cases being put forth for 5G. The operator-led Next Generation Mobile Networks (NGMN) Alliance has sought to define categories of 5G use cases (i.e., service types) that have distinct performance characteristics and commercial potential. In a 2015 white paper, the NGMN listed eight application categories for 5G [3, 4, 5]:

1. Broadband access in dense areas,
2. Broadband access everywhere,
3. Higher user mobility,
4. Massive Internet of Things,
5. Extreme real-time communications,
6. Lifeline communications,
7. Ultra-reliable communications, and
8. Broadcast-like services.

Each of these service types demand different network requirements that are determined by the types of traffic being sent and even the types of devices sending the traffic. For example, someone downloading cat videos will not have the same bandwidth or low latency requirements as a doctor in Los Angeles performing surgery virtually on a patient in Mumbai. The end-to-end notion of network slicing could be key to 5G's ability to effectively accommodate all of these disparate use cases.

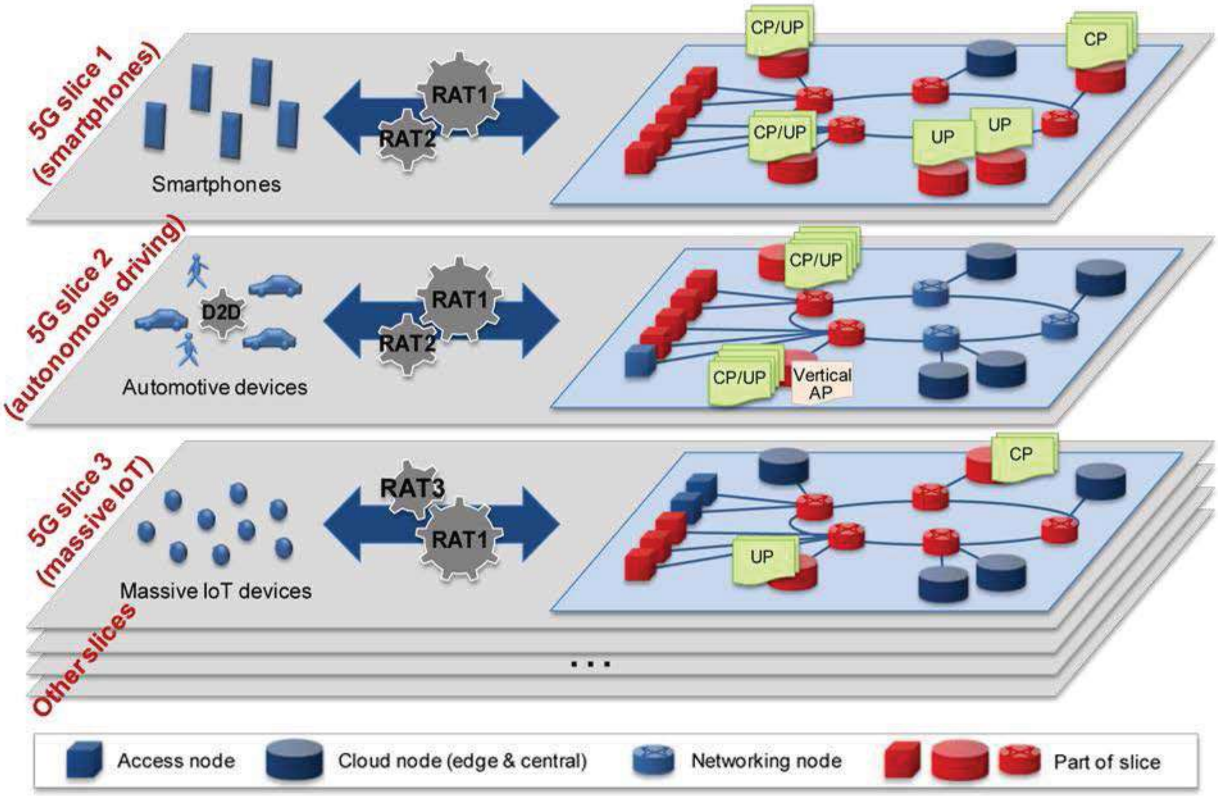
5G slice 1 (smartphones)
Smartphones
RAT1
RAT2
CP/UP
CP/UP
CP
UP
UP
5G slice 2 (autonomous driving)
D2D
Automotive devices
RAT1
RAT2
CP/UP
CP/UP
Vertical AP
5G slice 3 (massive IoT)
Massive IoT devices
RAT3
RAT1
CP
UP
Other slices
. . .
Access node
Cloud node (edge & central)
Networking node
Part of slice

FIGURE 1. 5G's incorporation of SDN and NFV would allow network slices to be created dynamically and deployed as needed to accommodate a variety of scenarios [2].

The Open Networking Foundation, in an April 2016 white paper, called out SDN's ability to support multiple services over a common architecture as a key enabler for network slicing. SDN also allows for slices to be created dynamically and deployed as needed to accommodate a variety of scenarios (see figure 1). Currently, 4G mobile networks prioritize traffic to get a similar effect, but with more rigidity and limitations. However, as 4G networks incorporate SDN and NFV, network slicing will become an option.

C-RAN

There are more than six million base stations (excluding small cells) deployed worldwide across approximately five million different cell sites serving close to four billion users. The surge in demand for connectivity has network operators searching for ways to shrink their network footprint, lower operational expenditures (OPEX), and still meet users' demand for access. C-RAN meets these requirements and has either been implemented or trialed by several operators including Verizon, AT&T, KT (South Korea), and China Mobile. Radio base stations currently depend on special purpose-built hardware deployed at the cell site. The baseband processing unit (BBU) is the part of the RAN that is responsible for managing the radio functions (or all functions that require an antenna). The BBU is one of the parts of the RAN that can be moved to a central location, creating a pool of BBUs to serve multiple base stations. C-RAN aims to centralize and virtualize baseband processing to reduce cell site costs and enable coordinated scheduling of resource blocks across a coverage area [6].

Figure 2 illustrates the evolution from the classic RAN model to a C-RAN setup. On the left, the classic model has the BBU deployed at the cell site connecting to the core network over IP/Ethernet transport. Any coordination between cell sites takes place over the X2 interface, which allows two sites to communicate. In a C-RAN architecture, the BBUs are pooled at a location away from the cell site. Pooling BBUs negates the need for the X2 interface as communications between cell sites now takes place internally. This is one of the reasons for the increased performance in C-RANs. The C-RAN model also makes updating the waveform and protocols easier as it only requires a software upgrade at the centralized BBU and not at each individual cell site

The move to a C-RAN architecture increases the flexibility of the network by allowing providers to, in theory, even change the types of RANs used—from 3G to 4G. For instance, in an area that has a mix of 3G and 4G users, operators can rebalance radio frequency resources by shifting more resources to 4G when 4G

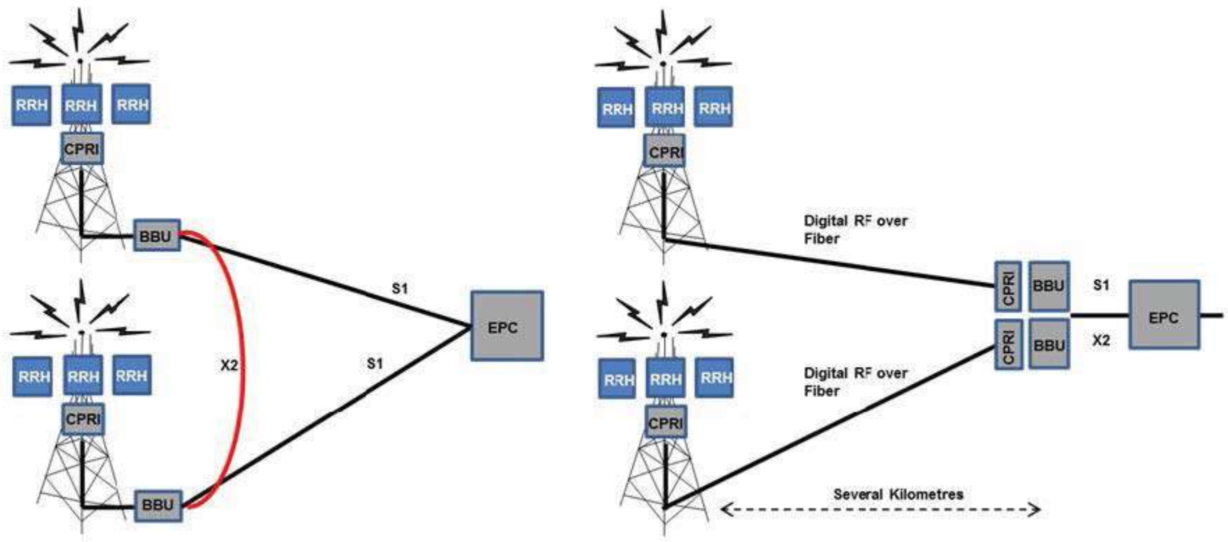

FIGURE 2. Unlike the traditional RAN model (on the left), C-RAN technology (on the right) pools the BBUs at a location away from the cell site offering an increase in performance and simpler upgrade path [6].

users are prevalent in the covered area. This shuffling of resources ensures that there is sufficient capacity for subscribers. C-RAN will also enable RAN-as-a-Service (RaaS), which will allow operators to rent RAN capacity to other operators.

Conclusion

On the face of it, virtualizing the 5G architecture seems like a necessary step towards preparing for the low latency requirements associated with 5G services. However, there are many uncertainties around 5G and a lack of visibility into what 5G will actually become through the standardization process, and at which phase different performance requirements will be supported or required. Phase II, the phase that is expected to meet International Mobile Telecommunication system for the year 2020 (IMT-2020) requirements, is expected in 2020. However, early versions or "pre-5G" offerings could be on the market before the final standard is approved. So, while these pre-5G solutions will have some 5G functionality, they will not field the full complement of improvements offered by an approved 5G system. For this reason, it is unclear to what degree virtualization techniques like C-RAN and network slicing will need to be implemented. It may come down to what the industry leaders in 5G implement in their "pre-5G" networks that will decide what a virtual 5G network will look like and when we can expect to see one.

References

[1] Farroha B, Everett JS, Uher JJ, Harper JR, Bridgland JK, Patton PM. "The status of 5G and the impact on current cellular architecture," 2016. College Park (MD): Laboratory for Telecommunications Sciences and The Johns Hopkins University Applied Physics Laboratory.

[2] Brown G. "5G use cases & candidate technologies," 2016. New York (NY): Heavy Reading.

[3] Hedman P. "Description of network slicing concept," 2016. NGMN Alliance. Version 1.0. Available at: https://www.ngmn.org/uploads/media/160113_Network_Slicing_v1_0.pdf.

[4] Marek S. "Dynamic network slicing is the 'secret sauce' behind 5G." SDxCentral. 2016 Aug 30. Available at: https://www.sdxcentral.com/articles/news/dynamic-network-slicing-secret-sauce-behind-5g/2016/08/.

[5] Paul M, Schaller S, Betts M, Hood D, Shirazipour M, Lopez D, Kaippallimalil J. "Applying SDN architecture to 5G slicing," 2016. Palo Alto (CA): Open Networking Foundation. Report No. TR-526. Available at: https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/Applying_SDN_Architecture_to_5G_Slicing_TR-526.pdf.

[6] Brown G. "Cloud RAN: New network architectures & industry structures," 2016. New York (NY): Heavy Reading.

5G and the Auto

Staff Writer

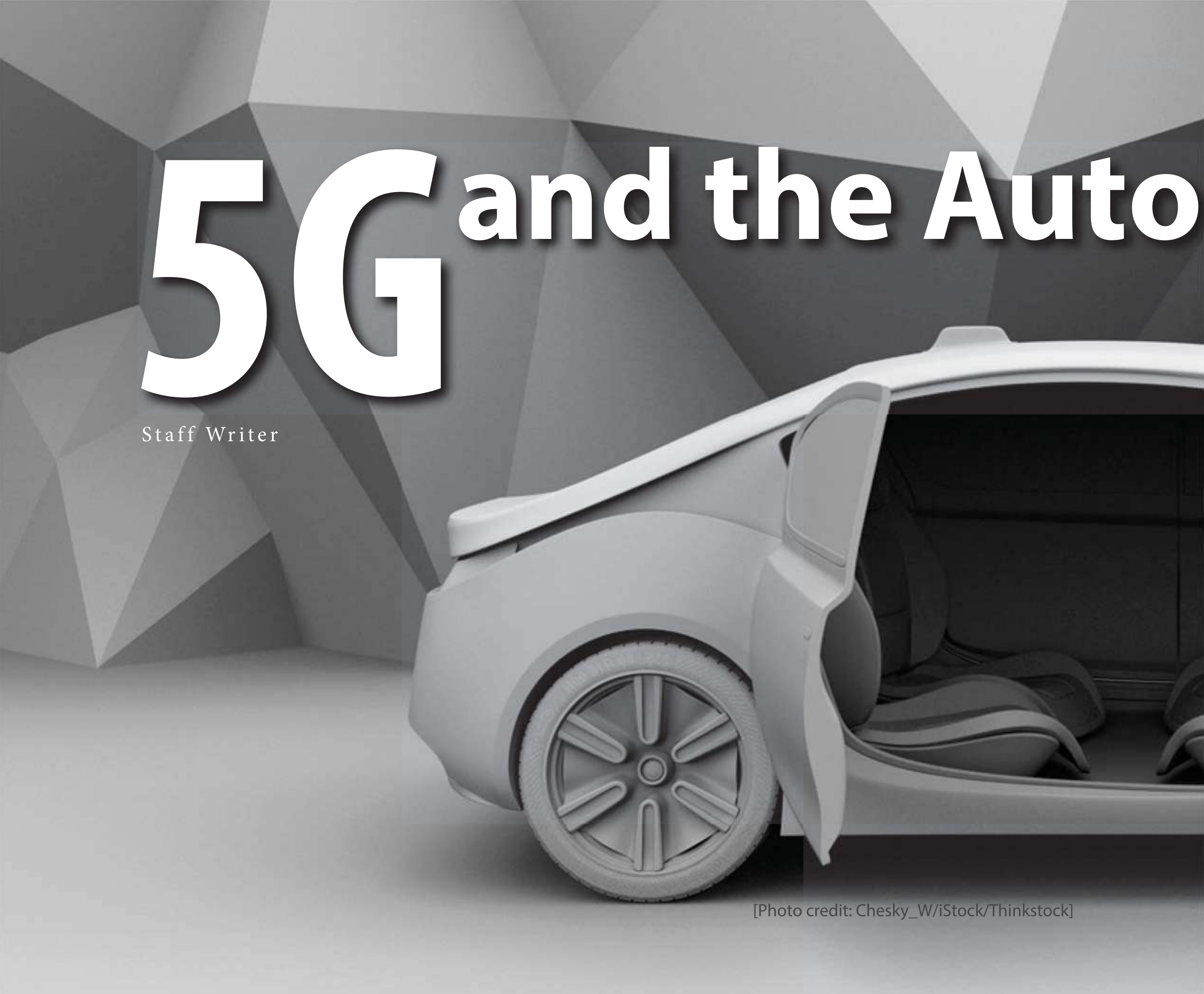

[Photo credit: Chesky_W/iStock/Thinkstock]

Industry stakeholders envision 5G as a key enabler that allows network connectivity in vehicles to shift in status from an optional accessory to a core feature that supports not only the individual vehicle, but also communication with other vehicles and sensors that inform traffic, parking, and navigation—while also ensuring passenger safety and data security. For consumers, a connected vehicle provides a growing number of features and services that make the driving experience safer, convenient, and less costly. 5G connectivity will enable information from in-car sensors to continuously be passed to the cloud. By sharing information and alerts about micro-level weather, road temperature, surface conditions, and violent breaking ahead, more efficient and consistent traffic flows will be achieved that reduce congestion and emissions. The aggregated and interpreted data will provide more informed driving information, as well as alert and activate onboard safety systems to prevent accidents [1].

motive Industry

Autonomous vehicles

Autonomous vehicles—also referred to as driverless or piloted vehicles—are predicted to hit the market by 2020, but stable 5G infrastructure will play a key role, according to a white paper by ABI Research [2]. Estimates indicate that 5G latency could be as low as one millisecond (ms) over-the-air, and 5 ms end-to-end, enabling the following automotive use cases:

- Broadband multimedia streaming (driverless vehicles as mobile living rooms).
- Cloud services for vehicle lifecycle management, apps, security, and over-the-air updates (cloud-to-vehicle).
- Capturing or uploading huge volumes of sensor data for real-time traffic, weather, parking, and mapping services (vehicle-to-cloud).
- Cooperative mobility: low latency vehicle-to-vehicle and vehicle-to-infrastructure for active safety and autonomy [redundancy for advanced driver assistance systems (ADAS)] [2].

The current ADAS being delivered on 2016 vehicles already facilitate SAE Level 1 (see figure 1) and are beginning to incorporate features that would be considered Level 2. However, Level 4 and 5 capabilities may not be that far away. In October 2015, Robot Taxi, a joint venture between Japanese mobile Internet company DeNA and vehicle technology developer ZMP, announced that it would offer driverless transportation to about 50 people in an area near Tokyo. Its goal is to commercialize the service by 2020, in time for the Tokyo Olympics [1]. While Robot Taxi is shooting for full automation (Level 5) in time for the games, it is likely that the use of such vehicles will initially be limited to shuttling passengers between Olympic venues. However, this venture is particularly notable because the technology is brand-agnostic and can be retrofitted to any vehicle [4].

Clearly, experimentation with autonomous vehicles is increasing, and 2020 as a date for some form of commercial implementation is certainly feasible from a technology perspective. Probably the best-known self-driving car project belongs to Google, which was started in 2009 and has clocked more than two million miles on public roads to date [5]. However, Google is not alone; Tesla, BMW, Audi, Mercedes, and most recently GM, have all showcased self-driving concept cars and demonstration projects. In mid-October 2016, Tesla announced that all cars currently being produced in their factories would include the hardware needed for full self-driving capability at a level of safety far greater than that of a human driver. However, the company added that the technology first needed to be tested and calibrated via “millions of miles of real-world driving” before the hardware would be activated on consumer vehicles [6]. In addition to traditional car manufacturers, companies such as Uber and Chinese search giant Baidu are also working on autonomous technology and self-driving cars [1].

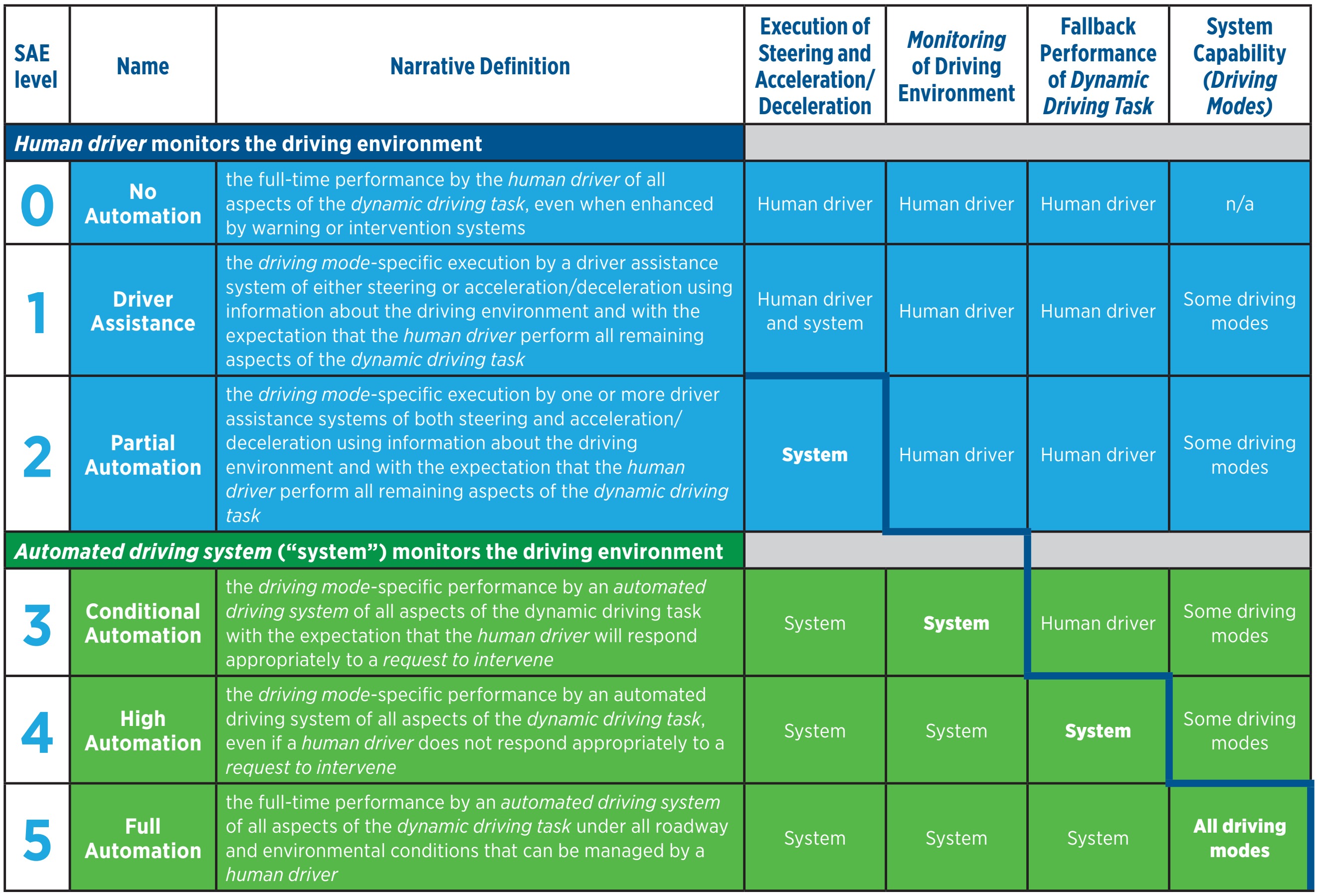

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
***Human driver* monitors the driving environment**						
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* (“system”) monitors the driving environment**						
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	**High Automation**	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

FIGURE 1. SAE Levels of automation. The Society of Automotive Engineers (SAE) has defined levels of automation to clarify what role (if any) drivers have in operating a vehicle while a driving automation system is engaged. These levels are intended to establish a consistent framework that can be used across industries as the dialogue about autonomous vehicles continues. (Figure credit: SAE International J3016 [3].)

The motivations for creating an autonomous vehicle are beyond just technology. It's about reducing emissions through better fuel consumption, as well as addressing the demographic changes of an aging population that increase, rather than decrease, the potential for human error-induced accidents. It's also about leveraging the convergence of the shared economy and urban living, where young and old people no longer feel the need to own a car if there is a cost-effective and convenient alternative, such as Zip Car rentals, or on-demand ride-sharing services such as Uber [1].

Looking forward

Geographical coverage will be a key condition for 5G to have any relevance in the automotive sector. Initial 5G coverage can be supplemented by 4G and Wi-Fi connectivity on phones and other devices while the infrastructure is being built up, but these multimode, multiconnectivity solutions will not suffice for critical automotive use cases relying on the unique capabilities of 5G in terms of latency, reliability, and security [2]. Even once 5G is fully deployed, the adoption of self-driving technology will likely play out differently in the various markets in different regions since the forces shaping it are diverse at both the global and local level [1].

Government and industry cooperation

The continuous progression of ADAS-enabled cars and the gradual adoption of the autonomous vehicle will significantly reduce, and possibly eliminate, the number of crashes. This could, in turn, allow the removal of some regulations that relate to safety considerations, such as crumple zones, bumpers, and airbags. It also means that a review of laws relating to driving age, drunk driving, and speed restriction enforcement may be required, but not until all vehicles are compliant [1].

In September 2016, the US Department of Transportation (DOT) released the Federal Automated Vehicles Policy for highly automated vehicles (HAVs), or those intended to operate at Levels 3 to 5 as defined by SAE. The document—which is currently intended as guidance rather than formal policy—lays out standards for safe design, development, and testing of HAVs before they are commercially sold or operated on public roads. It also proposes guidelines for state governments to ensure a consistent national framework for regulation of motor vehicles with all levels of automated technology [7]. DOT's National Highway Traffic Safety Administration released additional nonbinding guidance in October 2016, outlining best cybersecurity practices for motor vehicle manufacturers and individuals and organizations involved in developing self-driving technology. The guidance aims to make cybersecurity a top priority for the automotive industry and proposes layered solutions to ensure that automated driving systems are designed to take appropriate and safe actions, even when an attack is successful [8].

A challenge yet to be addressed is that, historically, car manufacturers have completely controlled the design and development of vehicles. The advent of computers and software that effectively become the "mind" of the car means that manufacturers could lose control to technology and software companies, and yet still remain liable for any issues or catastrophes related to the car. It also remains to be seen how much and what type of data car manufacturers and network providers would be expected to share to improve overall safety and security of connected vehicles [1].

References

[1] Bell S. "IOT & the automotive industry," 2015. New York (NY): Heavy Reading.

[2] Bonte D, Hodgson J. "5G in automotive," 2016. Oyster Bay (NY): ABI Research.

[3] SAE International. "SAE International technical standard provides terminology for motor vehicle automated driving systems" [Press release]. 2 Oct 2014. Available at: http://www.sae.org/misc/pdfs/automated_driving.pdf.

[4] Collins K. "Driverless taxis could be the bullet train of the Tokyo 2020 Olympics." CNET. 22 Aug 2016. Available at: https://www.cnet.com/news/driverless-taxis-could-be-the-bullet-train-of-the-tokyo-2020-olympics/.

[5] Higgins T. "Google's self-driving car program odometer reaches 2 million miles." *The Wall Street Journal.* 2016 Oct 5. Available at: http://www.wsj.com/articles/googles-self-driving-car-program-odometer-reaches-2-million-miles-1475683321.

[6] Tesla. "All Tesla cars being produced now have full self-driving hardware" [Blog post]. 2016 Oct 19. Available at: https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware/?utm_campaign=GL_Blog_101916.

[7] US Department of Transportation. “Federal automated vehicles policy.” September 2016. Available at: https://www.transportation.gov/AV/federal-automated-vehicles-policy-september-2016.

[8] National Highway Traffic Safety Administration. “Cybersecurity best practices for modern vehicles.” October 2016. Report No. DOT HS 812 333. Available at: http://www.nhtsa.gov/staticfiles/nvs/pdf/812333_CybersecurityForModernVehicles.pdf.

[Photo credit: moodboard/Thinkstock]

FROM LAB TO MARKET

News from the Technology Transfer Program

CREATING & GROWING ECONOMIC IMPACT

NSA's Technology Transfer Program (TTP) is igniting innovation through collaborations with industry, academia, and other government agencies that introduce NSA technologies to the commercial marketplace. By fulfilling the legislative mandate to share government-patented technologies with industry, NSA's TTP plays a role in strengthening national security by contributing to the nation's economic growth.

Technology Transfer Program
Office of Research & Technology Applications
NSA Research Directorate

tech_transfer@nsa.gov
www.nsa.gov/techtransfer
866-680-4539

2000–2014 REPORT

TOTAL DIRECT SALES
from new products and services

TOTAL ECONOMIC OUTPUT NATIONWIDE
including direct, indirect, and induced impact

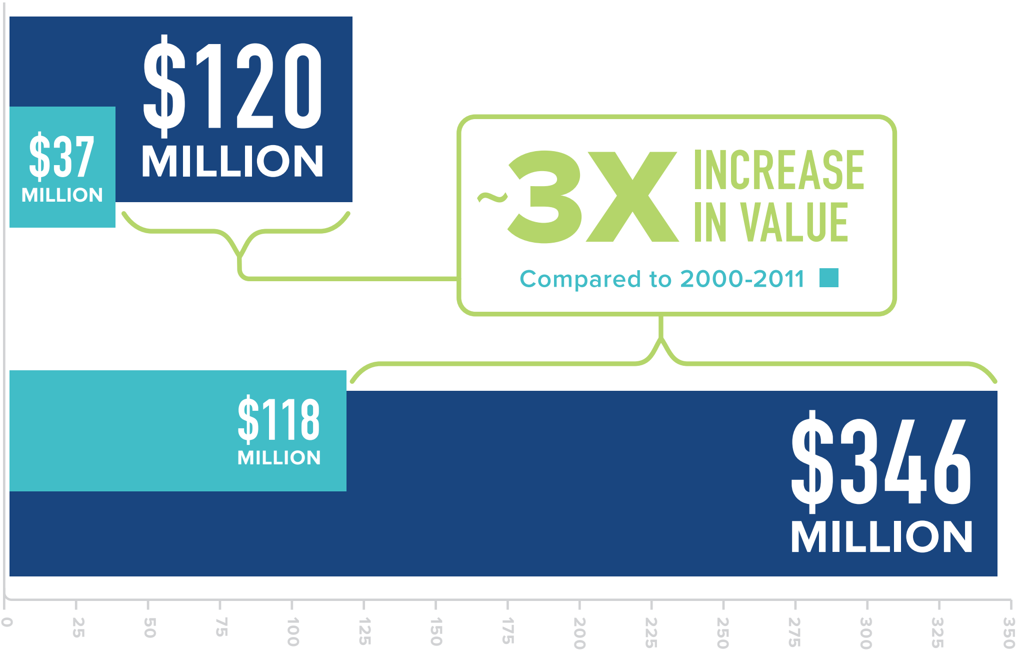

JOBS CREATED
or retained

1,330

SOURCE: *National Economic Impacts from DoD License Agreements With U.S. Industry*, TechLink and UC Business Research Division.

THE RESULTS

NSA Technology + Technology Transfer = Economic Growth ⇒ National Security

1G
2G
3

symmetry

MDPI

Article

IoT's Tiny Steps towards 5G: Telco's Perspective

Enida Cero [1], Jasmina Baraković Husić [1,2,*] and Sabina Baraković [3,4]

[1] BH Telecom, Joint Stock Company, Sarajevo, Franca Lehara 7, 71000 Sarajevo, Bosna i Hercegovina; enida.cero@bhtelecom.ba
[2] Faculty of Electrical Engineering, University of Sarajevo, Zmaja od Bosne bb, 71000 Sarajevo, Bosna i Hercegovina
[3] Faculty of Transport and Communications, University of Sarajevo, Zmaja od Bosne 8, 71000 Sarajevo, Bosna i Hercegovina; barakovic.sabina@gmail.com
[4] Department for Computer Engineering, American University in Bosnia and Herzegovina, Trg solidarnosti 10, 71000 Sarajevo, Bosna i Hercegovina
* Correspondence: Jasmina.BarakovicHusic@bhtelecom.ba; Tel.: +387-61-265-500

Received: 10 August 2017; Accepted: 28 September 2017; Published: 2 October 2017

Abstract: The numerous and diverse applications of the Internet of Things (IoT) have the potential to change all areas of daily life of individuals, businesses, and society as a whole. The vision of a pervasive IoT spans a wide range of application domains and addresses the enabling technologies needed to meet the performance requirements of various IoT applications. In order to accomplish this vision, this paper aims to provide an analysis of literature in order to propose a new classification of IoT applications, specify and prioritize performance requirements of such IoT application classes, and give an insight into state-of-the-art technologies used to meet these requirements, all from telco's perspective. A deep and comprehensive understanding of the scope and classification of IoT applications is an essential precondition for determining their performance requirements with the overall goal of defining the enabling technologies towards fifth generation (5G) networks, while avoiding over-specification and high costs. Given the fact that this paper presents an overview of current research for the given topic, it also targets the research community and other stakeholders interested in this contemporary and attractive field for the purpose of recognizing research gaps and recommending new research directions.

Keywords: 5G mobile communication; Internet of Things; applications classification; performance requirements; enabling technologies

1. Introduction

Telecom operators (telcos) had the most power and influence over business in the information and communication technology (ICT) industry during past decades. This dominance was the result of owning and provisioning communication infrastructures which nowadays have become more of a commodity than a luxury. Telco's revenue streams were mainly based on the provisioning of traditional services, such as voice calls and short message services (SMS). Recent work of regulatory agencies and the appearance of alternative service providers have led telecom operators to form the opinion that alternative service providers have conflicting interests and provide competitive services, thus decreasing telco's revenue from traditional services [1]. However, the latest econometric analysis presented in [2] has shown that telecom operators and alternative service providers have aligned interests and their collaboration could be favorable to both parties. The interests of alternative service providers and telcos are not inevitably conflicting, since the economic growth of the alternative service providers are positively correlated with telco revenues and vice versa.

The development of a digital society has changed the traditional value chain and introduced new issues for telcos as they seek a way how to monetize new digital services [3]. These services are diversified across various domains of industry, such as agriculture, construction, utilities, transportation, healthcare, finance, etc., and delivered to customers through connected devices building thereby the concept of Internet of Things (IoT).

IoT can be defined as an interconnected network of things/objects that are able to interact with each other and cooperate with other things/objects through wireless and wired connections in order to create new services/applications for the benefit of society. In this way, IoT brings almost limitless benefit's, which have the potential to radically change our daily life by saving time and resources while creating possibilities for growth and innovation [4]. IoT has disruptive potential in almost every application domain, which has created many challenges that need to be faced when implementing IoT-based solutions, such as prevailing over obstacles generated by the fragmentation of IoT, both in terms of application domains and in terms of technologies. Therefore, recognizing all potential IoT applications while having in mind the development of technology and the requirements of individuals, businesses, and society as a whole is quite challenging.

Telcos now have the opportunity to seize a share of the value that is generated by IoT implementation. The size of this share will depend on telco's role in the value chain that ranges from being a traditional provider of communication infrastructure to being an end-to-end solution provider [3]. In order to monetize IoT, telcos will have to address many challenges which can be summarized as follows: (1) strategic challenges relating to decisions on future directions; (2) business challenges relating to successful management, investment, partnerships; and (3) technical challenges relating to changing connectivity and performance requirements. This paper focuses on the technical challenges since connectivity and performance requirements of IoT objects/things cannot be fulfilled using existing cellular networks that limit numerous IoT applications. In order to overcome issues associated with the current cellular networks, new types of technologies are being introduced leading towards the fifth generation (5G) network [5–7].

The architecture of the 5G network has to seamlessly integrate the requirements of diverse IoT applications: from delay-sensitive video applications to ultra-low latency, from high-speed entertainment applications in a vehicle to mobility on demand for connected objects and from best effort applications to reliable and ultra-reliable ones for health and safety [8]. A full understanding of emerging IoT applications and the variability of their performance requirements can serve telecom operators as input for specifying 5G enabling technologies. These technologies should be flexible and scalable to meet the aforementioned requirements. Since IoT applications sometimes demand extreme requirements, the 5G network must simultaneously satisfy all of them, which can lead to over-specification and high cost. In order to avoid this, telecom operators first have to adequately classify IoT applications to facilitate the selection of 5G enabling technologies being capable to efficiently meet their performance requirements.

The 5G classification concept includes three different service classes [9], i.e., (1) extreme mobile broadband (xMBB); (2) massive machine-type communications (mMTC); and (3) ultra-reliable machine-type communications (uMTC). Nevertheless, this classification concept can be considered insufficient to properly select 5G enabling technologies to meet the diverse requirements of IoT applications. Furthermore, it will be challenging to classify the emerging IoT applications given that there is many criteria for their categorization. The definition of the IoT domain in many cases overlaps with the definition of IoT application, which may at the same time belong to another IoT domain. For example, Smart Buildings may either be considered as a standalone IoT application or as an element needed to form a Smart City application domain. IoT applications are usually classified according to spheres of human life [4,10–12] or performance requirements [9,13–15]. However, these classifications are not suitable for telecom operators which have to fulfil the performance requirements of particular IoT applications, since it is difficult to choose the most important one from a class containing a wide range of IoT applications.

Telecom operators can easily specify performance requirements for current communication services since traffic patterns generated by these services are driven by predictable activities, such as making calls, receiving email, surfing the web, and watching videos. However, traffic patterns generated by emerging IoT applications are driven by less predictable activities. Hence, the idea is to determine a common set of activities based on customer service requests which describe the reasonably foreseeable function/purpose of emerging IoT applications. Then these activities can be used by telcos as a new classification criterion of IoT applications. The assumption is that such a classification can help telecom operators meet performance requirements of future IoT applications and select appropriate 5G enabling technologies. Therefore, the aims of this paper are to classify IoT applications according to relevant activities, specify and prioritize the performance requirements of each IoT application class, and consider enabling technologies used to accomplish specified demands on the radio access part of 5G networks, all according to telco's perspective. In addition, the intention is to provide a literature review which covers 258 references dominantly published in the period between 2012 and 2017. As this paper provides a state-of-the-art review of IoT applications, their performance requirements and 5G enabling technologies, it can also be of use to the research community and other stakeholders interested in this contemporary and attractive field in order to recognize the research gaps and recommend new research directions.

The remainder of the paper is thusly organized. In order to answer research questions, Section 2 describes the methodological approach in regard to the conducted research. Section 3 identifies a common set of activities as a new classification criterion for IoT applications and assigns them the relevant 5G service classes. This facilitates the determining of performance requirements of activity-based IoT application classes, which are summarized and prioritized in Section 4. The enabling technologies used to accomplish these requirements in 5G networks are contemplated in Section 5. Finally, Section 6 discusses the outcomes of the paper by clarifying the contribution of our research in identifying open issues for future work, while Section 7 concludes this paper.

2. Research Approach and Design

The main objectives of this paper can be summarized as follows: (1) to identify activities relevant to IoT customer service requests and use them as a new classification criterion of IoT applications; (2) to specify and prioritize performance requirements of such IoT application classes; (3) to analyze relevant literature in order to provide insight into 5G radio technologies used to fulfil the requirements of activity-based IoT application classes; (4) to recognize research gaps and directions.

The research questions posed in this study are: (1) Which activities can be used as classification criteria for IoT applications? (2) What are the performance requirements of such IoT application classes? (3) Which enabling technologies in the radio access part of the 5G network can be used to meet the requirements of IoT application classes?

Investigation of different criteria for the classification of IoT applications underlines to what extent various classification concepts have been covered by existing literature and provides a basis for the introduction of a new classification criterion called the activity. We thus surveyed and compared different studies contributing to the understanding of the scope and classification of a wide range of IoT applications. This approach has been motivated by the challenge to propose an activity-based classification of IoT applications, the determination and prioritization of performance requirements of such IoT application classes and the review of 5G enabling technologies used to meet them. These findings can be utilized by telecom operators and other interested parties (e.g., the research community, software network function providers, network infrastructure manufactures, etc.) depending on their interest and potential to utilize these findings towards 5G implementation and commercialization.

The methodological approach to research conducted in this paper is illustrated in Figure 1 and includes four phases.

Phase I included searching, identifying, and extracting papers from three categories, i.e., IoT in 5G service classification, IoT in 5G performance requirements and IoT in 5G enabling technologies. The keywords used to search relevant scholar databases are shown in Figure 2. This search has resulted in the selection of 258 recently published papers, which can be categorized into three groups (i.e., review papers, technical papers, research effectiveness), as shown in Table 1. Review papers summarize the status of knowledge and outline future directions in a given area of research. Technical papers describe the process, progress, or result of research, whereas research effectiveness directly provides an answer to the research question raised for this study. The reference distribution by year of publication is shown in Figure 3a, while the total percentage of references per publication type is presented in Figure 3b.

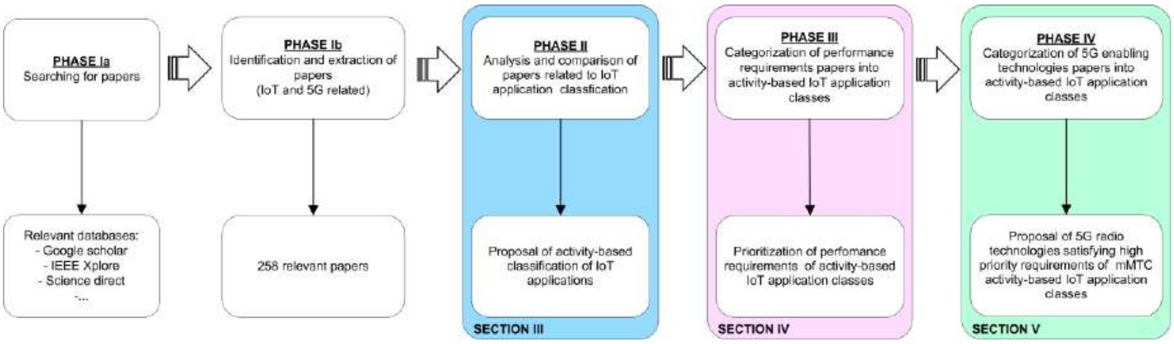
PHASE Ia
Searching for papers
PHASE Ib
Identification and extraction of papers
(IoT and 5G related)
PHASE II
Analysis and comparison of papers related to IoT application classfication
PHASE III
Categorization of performance requirements papers into activity-based IoT application classes
PHASE IV
Categorization of 5G enabling technologies papers into activity-based IoT application classes
Relevant databases:
- Google scholar
- IEEE Xplore
- Science direct
-...
258 relevant papers
Proposal of activity-based classification of IoT applications
Prioritization of perfomance requirements of activity-based IoT application classes
Proposal of 5G radio technologies satisfying high priority requirements of mMTC activity-based IoT application classes
SECTION III
SECTION IV
SECTION V

Figure 1. Methodological approach to research.

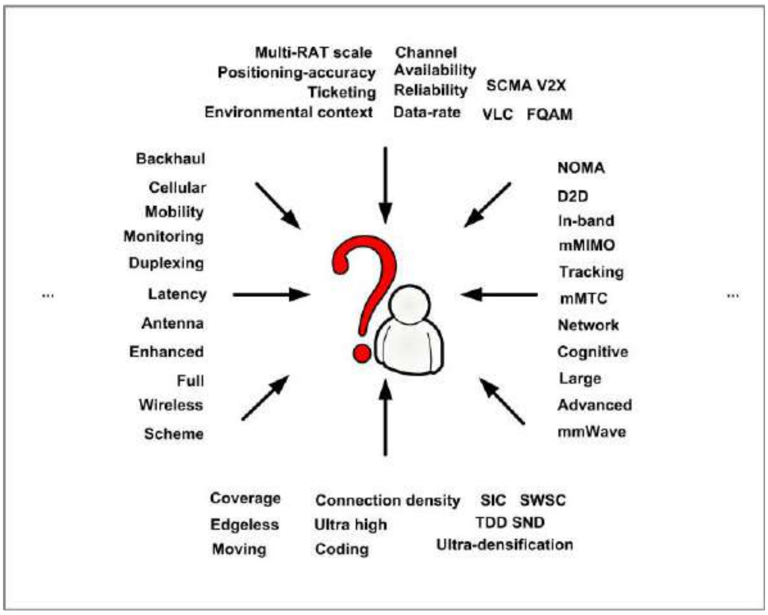
Multi-RAT scale
Positioning-accuracy
Ticketing
Environmental context
Channel
Availability
Reliability
Data-rate
SCMA V2X
VLC FQAM
Backhaul
Cellular
Mobility
Monitoring
Duplexing
Latency
Antenna
Enhanced
Full
Wireless
Scheme
...
NOMA
D2D
In-band
mMIMO
Tracking
mMTC
Network
Cognitive
Large
Advanced
mmWave
...
Coverage
Edgeless
Moving
Connection density
Ultra high
Coding
SIC SWSC
TDD SND
Ultra-densification

Figure 2. Keywords for searching relevant databases.

Table 1. Reference categorization.

Reference Type	Reference Number
Review paper	[1,2,10,12,14–88]
Technical paper	[8,89–197]
Research effectiveness	[3–7,9,11–13,29,52,198–258]

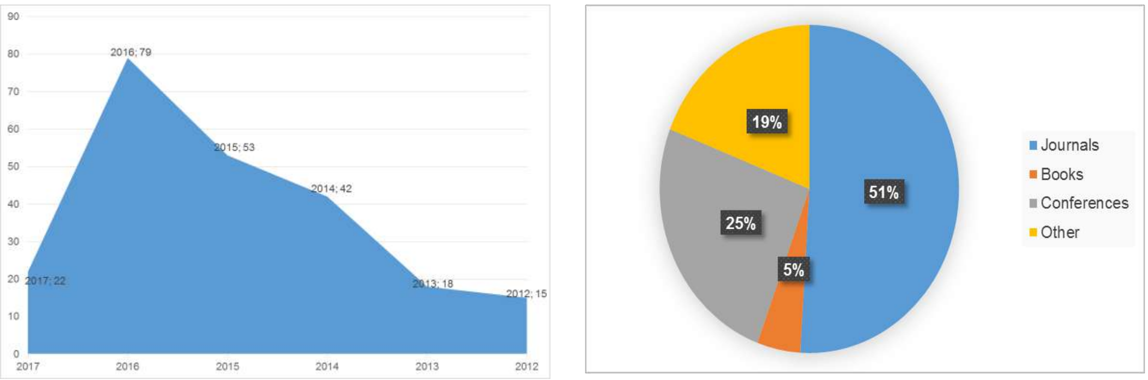

(**a**) (**b**)

Figure 3. Reference distribution by (**a**) year of publication; (**b**) type of publication.

Phase II included the analysis and comparison of papers related to the category of IoT in 5G service classification. Papers were first sorted according to the existing classification criteria of IoT applications in order to underline their deficiencies from telco's perspective. Since emerging IoT applications generate traffic driven by less predictable activities, specifying their performance requirements and enabling technologies is not an easy task from telco's point of view. Therefore, we have proposed a common set of activities to serve as classification criteria of IoT applications. Four activities were selected based on IoT customer service requests denoting the function/purpose of existing IoT applications. According to the authors' knowledge, existing IoT applications can be grouped around following activities: ticketing, tracking, monitoring, and managing/controlling. Such a classification makes it easier for telecom operators to associate particular IoT applications to the relevant 5G service class. Finally, Phase II has resulted in the proposal of an activity-based classification of IoT applications which was associated with the 5G service classification.

Phase III included the analysis of papers related to the category of IoT in 5G performance requirements. Papers were grouped according to eight key performance indicators, i.e., data rate, mobility, latency, connection density, reliability, positioning accuracy, coverage, and energy efficiency. These performance requirements were prioritized indicating high, medium, and low importance of each requirement for specific activity-based IoT application classes proposed in Phase II. Finally, Phase III resulted in the identification of highly important performance requirements for each activity-based IoT application class which can be utilized by telcos to identify 5G enabling technologies used to meet them.

Phase IV provided a review of papers from the category of IoT in 5G enabling technologies. Papers were first sorted according to eight technological groups used in the radio access part of 5G networks, i.e., wide and flexible bandwidth technology, advanced modulation and coding, duplexing, multiple access and waveform, advanced interface management, access architecture related radio technologies, energy related technologies, and other technologies. These technologies have been discussed in terms of their possibility to meet the high priority performance requirements of mMTC activity-based IoT application classes, since this area is more mature from telco's perspective. The results of Phase III were reflected in a proposal of technologies in the radio access part of 5G networks that can be used to meet the high priority performance requirements of the mMTC activity-based IoT class, as well as the identification of research gaps and directions for future work.

3. IoT in 5G Service Classification

As mentioned in the Introduction, various challenges awaiting telcos can be identified on the basis of the role that they will play in the IoT value chain. Regardless of this role, telcos will face many technical challenges, such as the necessity for global deployment, the need for rapid infrastructure scaling, unpredictable IoT application behavior, etc. Even now, existing IoT applications accelerate the growth rate of traffic driven by less predictable activities which require new strategies towards 5G networks. These networks will be the key enabler for IoT applications by providing a unified infrastructure capable of meeting the high variability of their performance requirements [8,16]. Various stakeholders have recently described what 5G networks may be and grouped the major types of 5G services into three different classes: (1) xMBB—Requiring extremely high data rates and low-latency communication; (2) mMTC—Requiring scalable connectivity for an extremely large number of devices; and (3) uMTC—Requiring ultra-reliable low-latency and resilient communication [198]. This 5G service classification can be used by telcos to identify performance requirements of various IoT applications, and, as such, was used in this research study. However, emerging IoT applications will have extreme requirements, so an analysis based solely on 5G service classifications may not be sufficient. Therefore, telcos need a more precise classification of IoT applications which is an important precondition for meeting their diverse requirements.

IoT covers human-to-human (H2H), human-to-machine (H2M) and machine-to-machine (M2M) communication, which will be the main driving force towards 5G networks. In addition, the terms M2M communication (M2MC) and machine type communication (MTC) are used interchangeably as in [17–19,89–93,199,200]. Although IoT is a broader concept which evolves from M2M, this paper assumes that IoT and M2M are synonyms as in [94].

3.1. Existing Classifications of IoT Applications

In order to meet the requirements of a wide range of IoT applications, they have to be classified in an appropriate manner. The existing approaches to the classification of IoT applications are summarized in Table 2. IoT serves different user categories, including individuals, businesses, and society as a whole, and may span through a broad range of application domains [4,10–12], such as transportation and logistics, healthcare, smart environment, personal and social, futuristic applications, food sustainability, smart living, smart manufacturing, smart energy, smart city, etc. These application domains are created to be human-centric, which means that they cover different domains of human life. IoT applications belonging to these domains have diverse requirements for 5G networks, and should not be treated equally. Therefore, the acceptance of these domains as a classification criterion raises the issue of assigning IoT application to a particular domain. Moreover, the emergence of new applications of IoT may require defining new application domains, which makes this classification inappropriate from telco's perspective.

The M2M applications may be classified by the mobility and the amount of dispersion that needs to be supported into four categories [13]: (1) fixed and concentrated; (2) fixed and dispersed; (3) mobile and concentrated; and (4) mobile and dispersed. However, the mobility and amount of dispersion present very rough classification criteria and cannot meet the precise network requirements of individual IoT applications.

Additionally, M2M applications can be grouped according to delay tolerance into four categories [14]: (1) elastic (delay tolerant); (2) hard real-time (delay constraint); (3) delay-adaptive (delay sensitive but tolerant); and (4) rate-adaptive application (adjust their transmission rates according to available radio resource). However, the main drawback of this classification is the lack of consideration of other IoT application requirements except delay tolerance, although its importance as a classification criterion has been recognized.

According to data reporting mode, the M2M applications can be classified into five categories [91]: (1) time-driven; (2) query-driven; (3) event-driven; (4) continuous-based; and (5) hybrid-driven. This classification is specific for the former IoT concept, which is narrower in nature than the definition adopted in this study.

Another study considered the reliability, availability, and end-to-end latency in order to classify IoT applications into two groups [15]: (1) monitoring-based and mission-critical; (2) monitoring-based and non-mission critical; (3) control-oriented and mission-critical; and (4) control-oriented and non-mission critical. Monitor-based IoT applications periodically collect sensor data from smart objects and transmit them. The majority of monitoring-based IoT applications are not mission-critical. Control-oriented IoT applications use sensor data to control actuators in real-time, and rely on mission critical communication. This classification is based on multiple criteria and represents the precursor of the Mobile and wireless communications Enablers for Twenty-twenty Information Society (METIS) 5G service classification.

Table 2. Internet of Things (IoT) application classification—summary.

Criteria for Classification	IoT Classes	References
Domains	(1) Transportation and logistics (2) Healthcare (3) Environment (4) Personal and social (5) Futuristic applications (6) Food/water monitoring (7) Living (8) Manufacturing (9) Energy (10) Building (11) Industry (12) City (13) Security and safety (14) Communication (15) e-society (16) Vehicular (17) Sport and leisure	[4,10–12]
Mobility and amount of dispersion	(1) Fixed and concentrated (2) Fixed and dispersed (3) Mobile and concentrated (4) Mobile and dispersed	[13]
Delay tolerance	(1) Elastic (2) Hard real time (3) Delay-adaptive (4) Rate-adaptive	[14]
Data reporting mode	(1) Time-driven (2) Query-driven (3) Event-driven (4) Continuous-based (5) Hybrid-driven	[198]
Reliability, availability, and end-to-end latency	(1) Monitoring-based and mission critical (2) Monitoring-based and non-mission critical (3) Control-oriented and mission critical (4) Control-oriented and non-mission critical	[15]
Characteristics and requirements	mMTC and uMTC	[12]

Legend: IoT (Internet of Things), mMTC (massive Machine Type Communication), uMTC (ultra-reliable Machine Type Communication).

Although it spans through a wide range of different applications, MTC can be divided in two main categories, i.e., massive and ultra-reliable MTC, which depend on their characteristics and requirements [89]. As mentioned above, this categorization is a part of 5G service classification which was used in this study as a basis to identify performance requirements of IoT applications and will be described in more detail in Section 4. Massive MTC (mMTC) typically involves a very large number of devices (tens of billions [9]), such as sensors, actuators, and similar devices [12], different in complexity and cost [9], and with varying quality of service (QoS) requirements. These devices should be of very low cost with very low energy consumption, enabling very long battery life [12]. At the same time, the amount of data generated by each device is normally very small, and very low latency is not a critical requirement [12]. Ultra-reliable MTC (uMTC) requires very high reliability and availability, and very low latency [9,12]. Low device cost and energy consumption are not as critical as they are for mMTC applications [12], and the number of devices and required data rates are relatively low [198].

The vision of a pervasive IoT requires the integration of various domains into a single and unified domain. It addresses the enabling technologies needed for these domains while taking into account the elements that form the third dimension like security, privacy, trust, and safety [11]. The current classification criteria and IoT application classes do not clearly differentiate IoT domains from IoT applications. IoT domains are usually viewed as a specific area of IoT applications that support a huge variety of use cases across industries impacting businesses and customers. Therefore, IoT domains overlap with IoT applications within existing IoT classifications. Moreover, IoT classification based on certain QoS parameters is difficult to apply from telco's perspective, because it is challenging to choose the most important parameter for the wide range of IoT applications. However, these problems need to be solved for upcoming 5G networks.

3.2. Activity-Based Classification of IoT Applications

The aforementioned issues motivated us to propose a new approach to the classification of IoT applications. This approach is based on "**the activity**", which primarily characterizes specific IoT application. In this sense, the activities are defined as new classification criteria, which denote a main function/purpose of specific IoT application observed from telco's point of view. Telecom operators observe the technical challenges of IoT through systems, tools, devices, and platforms because their availability and integration complexity determines the opportunity to capture a share of the value that is generated by IoT implementation. Based on the literature review, it was found that the most commonly mentioned terms in this sense were related to: (1) ticketing system [201]; (2) monitoring: devices [199,201–204]/services [199,201,202,205,206]/tools [207]/systems [11,12,199,201,205,206,208–210]/data [95,201]/framework [12,201]/solutions [20,205]/networks [211]/process [7]/activity [203,204,207]; (3) tracking technologies [4]/devices [204]/systems [204]/applications [11,210]/system [201]/services [202,205]; (4) managing/controlling applications [15,210,212]/operation [211]/services [4,208]/tools [208,211]/devices [204]/system [95,198,204,210]/concepts [211]/solutions [95]/platforms [95,199]. This has led us to identify four activities, i.e., ticketing, monitoring, tracking, and managing/controlling, as new classification criteria of IoT applications. According to the authors' knowledge, it was found that these activities can cover reasonably foreseeable functions/purposes of IoT in existing application domains. Using activities as classification criteria, telcos can specify IoT application performance requirements more easily, which are in that case dictated by the specific activity, not by the IoT application domain, and determine enabling technologies in the radio access part of 5G networks.

In addition, our classification approach allows better service differentiation and service delivery closer to customer expectations. According to BH Telecom's experience, customers typically come up with the following service requests when it comes to IoT: (1) they need to track their products and determine products' distribution across different regions based on their own data analysis; (2) they want to monitor their products since they do not have a department for supervision and analysis; (3) they need information panels about their outlets and working time to be downloaded by a scanning tab-ticket; (4) they want to manage/control their products according to market needs. This way of expressing customer needs has inspired us to propose the activity-based classification of IoT applications as it allows telcos to define performance requirements more precisely, and thereby improve the customer experience.

The activity-based classification of IoT applications is presented in Table 3. Each activity is associated with the application domain where that activity may be applied. The requirements of identified activities and associated application domains are then mapped to 5G service classes in order to be further able to identify their performance requirements and enabling technologies necessary to meet them. For example, according to Table 3, the managing/controlling activity can be realized in several domains, such as healthcare, food, energy, transportation and logistics. For each domain, we have identified an application example, such as remote surgery [213] in the healthcare domain, food processing facilities [96] in the food domain, energy distribution [12] in the energy domain and traffic/driving [12,213] in the transportation and logistics domain. According to diverse performance requirements of these IoT applications [214], managing/controlling activity in healthcare, food, transportation, and logistic domains can be associated with the uMTC service class, while the energy domain can be mapped to the mMTC service class. A more detailed description of performance requirements for each activity-based class of IoT applications is provided in Section 4, while in Section 5 it is discussed how these requirements can be accomplished in the radio access part of 5G networks.

Table 3. Activity-based classification of IoT applications.

Activity	Domain Examples	5G Service Classification	Application Examples
Ticketing	Smart Transportation and Logistics	mMTC	POS Terminal [4]
Monitoring	Smart Healthcare	uMTC	Health condition [97]
	Smart Buildings	uMTC	Structures (buildings, tunnels, etc.) [12,25,213]
	Smart Buildings/Smart City	mMTC	Parking spaces
	Smart Buildings/Smart Environment	mMTC	Home video [12]
	Smart food/water monitoring	mMTC	Food growth condition [10]
Tracking	Smart Healthcare/Sport and Leisure	mMTC	Medical assets, wearables [200,213]
	Smart Transportation and Logistics	mMTC	Transport fleet [200,213]
	Smart Industry/Social Networking	mMTC	Shipping of products
	Smart Healthcare	mMTC	People in science museum
Managing/ controlling	Smart Healthcare	uMTC	Remote surgery [213]
	Smart food/water monitoring	uMTC	Food processing facilities [96]
	Smart Transportation and Logistics	uMTC	Traffic, driving [12,213]
	Smart Energy	mMTC	Energy distribution [12,98]

Legend: 5G (Fifth Generation), POS (Point Of Sale), mMTC (massive Machine Type Communication), uMTC (ultra-reliable Machine Type Communication).

As such, the activity-based classification of IoT applications can be used for the creation of new business models, which represent the stakeholder's plan to generate revenue and make a profit from operations, and thereby include many components and functions of the business [21]. However, there is no common opinion which components constitute a business model. The business model architecture can be illustrated by four dimensions [22]: (1) who, identifying the definition of the target customer as one central dimension in designing a new business model; (2) what, describing what is offered to the target customer; (3) how, referring to the construction and distribution of the value proposition; (4) value, explaining why the business model is financially viable. Answering these four questions allows the creation of IoT business models. According to this business model definition, the activity-based classification of IoT applications affects the who and the how dimensions in the following manner. In terms of the who dimension, it directly allows customer segmentation according to considered activities as a way to express their requirements (e.g., customers that require tracking or monitoring of their products). Being aware of the current customer requirements, the proposed classification of IoT applications around four activities (i.e., ticketing, monitoring, tracking, and managing/controlling) can be considered complete. At the same time, the activity-based classification of IoT applications is flexible since additional activities as classification criterion can be concerned with emerging customer requirements. On the other hand, the activity-based classification of IoT applications indirectly impacts the how dimension of new business models which among others includes relevant resources and capabilities in the focal stakeholder's internal value chain. This indicates that the proposed activity-based classification of IoT applications can be easily applied to IoT business models.

Various categories of IoT business models can be identified according to eight IoT architectural layers [23], i.e., collaboration and processes layer, application layer, service layer, abstraction layer, storage layer, processing layer, network communication layer, physical layer. Along with IoT architectural layer, the IoT business model needs also to address the IoT value proposition [24] and

IoT stakeholders that can participate in more than one layer. Based on their role in the IoT ecosystem, telecom operators usually take a part in the network communication layer. In this regard, they follow four evolutionary business models [23]: (1) selling connectivity services only; (2) selling third-party products; (3) selling internal products; (4) providing a broad menu of IoT products. Each of these IoT business models includes traditional telcos strength, i.e., connectivity which determines the how dimension of the IoT business model. In this sense, our activity-based classification of IoT applications indirectly allows the construction and distribution of performance as IoT value proposition using enabling technologies as IoT resources in telco's internal value chain.

3.3. Summary of IoT in 5G Service Classification

The previous discussion has shown that the classification of IoT applications is a complex task due to their numerosity and diversity. The existing classifications of IoT applications pose some drawbacks that can be summarized as follows. The domain-based classification does not allow clear differentiation between IoT domains and IoT applications due to either an imprecise classification criterion or the diversity and unpredictability of IoT applications. On the other hand, QoS-based classifications need to identify the common and most important performance metric for a broad range of IoT applications. Therefore, on the basis of the literature review and IoT customer service requests, we have proposed a new approach to the classification of IoT applications. It is based on the activity as new classification criterion, which denotes a main function/purpose of specific IoT application observed from telco's point of view. According to the authors' knowledge, four activities, i.e., ticketing, monitoring, tracking, and managing/controlling, have been identified to cover reasonably foreseeable functions/purposes of existing IoT applications. The resulting activity-based IoT application classes have been associated with 5G service classes in order to determine and prioritize their performance requirements as described in next section. Finally, the proposed classification of IoT applications was discussed in terms of its completeness, flexibility, and applicability to new business models.

4. IoT in 5G Performance Requirements

This section provides an insight into performance requirements of activity-based classes of IoT applications proposed in Section 3. The analysis is based on eight key performance indicators identified in [5,207] as shown in Table 4: data rate, mobility, latency, connection density, reliability, positioning accuracy, coverage, and energy efficiency. These performance indicators are usually well described for specific IoT applications. However, one of the main challenges of 5G is to support a variety of performance requirements for numerous IoT applications in a flexible, reliable, and cost-effective way [15]. Hence, there is a need for a comprehensive understanding of these requirements for activity-based IoT application classes. The 5G service classification defined the performance requirements for mMTC and uMTC [9]. We have assigned these requirements to the activity-based IoT application classes introduced in Section 3 for the purpose of proposing priorities of each requirement for a specific class. Three levels of priorities (high, medium, low) are associated with the performance requirements of activity-based classes of IoT applications as shown in Table 4. The prioritization of performance requirements is inspired by analysis of related work undertaken in [213]. Some activity-based classes of IoT applications may demand optimization of multiple performance requirements. Table 4 illustrates the main differences between activity-based IoT application classes, and therefore, the need for a 5G network that enables support of optimal configurations for a variety of, sometimes opposite, requirements. For example, the mMTC tracking-based IoT application class requires support for high mobility, high positioning accuracy, and high connection density, while the mMTC monitoring-based IoT application class also requires high connection density but low mobility and low positioning accuracy. A more detailed description of considered performance requirements is contained in the following subsections.

Table 4. Activity-based classes of IoT applications—performance requirements.

5G Service Classification	Activity	User Experienced Data Rate [Gbps] Outdoor: 0.1 [5] Indoor: 1 [5]	Mobility [km/h] Required: 500 [5]	Latency [ms] Control Plane: 50 [5] User Plane: 1 [5]	Connection Density [Connections/km^2] Required: 10^6 [5]	Reliability [%] Required: 99.999 [207]	Positioning Accuracy Required: A Few cm [5]	Coverage/ Availability [%] Required: 99.999 [207]	Energy Efficiency [bits/J]
mMTC	Ticketing	L to M	L	L	H	M	H	H	H
	Tracking	M	H	L	H	M	H	H	M to H
	Monitoring	M to H	L	L	H	L to M	L	H	M
	Managing/controlling	L	L	L	H	M	L	M to H	M
uMTC	Monitoring	L	L to M	H	H	H	M to H	H	M
	Managing/controlling	L	M to H	H	L	H	H	M to H	M

Legend: 5G (Fifth Generation), mMTC (massive Machine Type Communication), uMTC (ultra-reliable Machine Type Communication), L (Low), M (Medium), H (High).

4.1. Data Rate

Data rate is the most important evaluation factor for generations of wireless communication networks [6]. It is contemplated in two ways: (1) peak data rate—defined as the maximum achievable data rate by the user; and (2) minimum guaranteed user data rate—defined as the minimum experience data rate by the user [16]. New mobile technologies are primarily driven by users' needs for higher data rates, as discussed in [8,16,99,215,216]. The expected values in 5G networks are 10 Gbps for minimum peak data rate and 100 Mbps as minimum guaranteed user data rate [5]. High data rate requirements are mainly posed by xMBB related use cases [9] like hotspots [214] or dense urban areas where in 95% of locations and time experience data rate by user should be 300 Mbps in downlink and 60 Mbps in uplink [100]. High data rate is also important in some activity-based classes of IoT applications, as shown in Table 4. Medium (e.g., monitoring of parking spaces) to high (e.g., monitoring of home video) data rate is needed in cases of mMTC monitoring activities. High data rate can be achieved by using a millimeter wave (mmWave) spectrum to share multi-gigabit data in the surrounding environment (i.e., mMTC monitoring with high data rate requirement) and to recognize an object via cloud in real time to find the optimal driving strategy instantaneously for IoT—autonomous vehicles application [101]. Data rates generated during transmissions of tracked medical assets, transport fleets, and ticked point of sale (POS) terminals is low (e.g., medium importance for ticketing and tracking activities in mMTC). In addition, low data rates are needed in uMTC monitoring and controlling (e.g., monitoring of health condition, control of driving).

4.2. Mobility

Mobility is defined as relative velocity between the receiver and the transmitter [16]. The applications of IoT pose very diverse requirements for mobility in 5G networks, which range from static to high mobile, even up to 500 km/h [16,102,207]. Use cases in which except ultra-high mobility, ultra-high traffic volume density, and ultra-high connection density are needed may be quite challenging for 5G networks [215], like V2X communication [5]. High mobility is a very important requirement for mMTC tracking activity (e.g., tracking of assets in high speed trains [214]). Moreover, the support for high mobility is needed in uMTC management activities, if the monitor/control object is moving (e.g., high speed trains [207]). Low mobility is needed for ticketing, monitoring, and managing/controlling activities in the mMTC class, as shown in Table 4. This is the reason why 5G networks should not assume mobility for all devices and services but rather provide mobility on demand [8].

4.3. Latency

Latency requirements are usually expressed in terms of end-to-end (E2E) latency perceived by the end user [8]. 5G networks should enable "zero latency" [103] represented by the millisecond level of E2E latency [16,26,27,213] through significant enhancements and new technology in architecture aspects [6,28,214], such as device-to-device (D2D) communication [5]. The IoT applications associated with uMTC monitoring and managing/controlling activities require low latency, as they tend to be real time. Required latency levels depend on the particular IoT application [8,16,28], being the narrowest in uMTC managing/controlling activity with a value that should not exceed 1 ms [8]. For example, the tolerable delay for use case mobile health [215] and remote surgery application is in order of sub-milliseconds [213]. For V2X communications, latencies should be ultra-low for some warning signals [8].

4.4. Connection Density

Connection density is defined as the number of connected devices per unit area [16]. It can usually be expressed in terms of an extremely high number of simultaneous active connections, such as 1 million connections per square meter [207,215], or 10 to 100 times higher number of connected devices [29]. The performance values of connection and traffic density for various 5G services are listed in [8,212]. This performance requirement is highly correlated with identified activity-based classes of IoT applications, since it is the main characteristic of mMTC. High device density brings the need for rapid internet protocol version 6 (IPv6) deployment, and high quality security algorithms and techniques, which lead to new system design and implementation, as described in [104].

4.5. Reliability

Reliability is the maximum tolerable packet loss rate at the application layer [105]. 5G must bring a reliability of 99,999% [27,30,31,212], or higher for specific use cases [212], (e.g., tele-protection in a smart grid network [207] or driverless cars [106]). Reliability is the main characteristic of uMTC monitoring and managing/controlling activities. Reliability will be a particularly challenging task in high-speed trains because of speed, load, and cell distance [16,107].

4.6. Position Accuracy

Position accuracy is the maximum positioning error tolerated by the application [105]. 5G should ensure accurate positioning of the device outdoors [108] with accuracy from 10 m to <1 m on 80% of occasions and better than 1 m in indoor deployment [212]. Accuracy positioning is very important in uMTC monitoring-based activities (e.g., monitoring remote cameras), and uMTC managing/controlling-based activities (e.g., driving) [213]. Moreover, the mMTC tracking-and ticketing-based activities pose high performance requirements in terms of position accuracy.

4.7. Coverage

Coverage requirements in 5G should provide connectivity anytime and anywhere with a minimum user experience data rate of 1 Gbps [32]. However, the perception of 100% coverage is rather a technical decision than a business one [33], which could be extended through ultra-cell deployment [31] and vehicle-to-infrastructure (V2I) communications [30]. Almost every activity-based IoT application class requires very high levels of coverage (99,999% availability) [207]. Total coverage will enable new unmanned aerial vehicles (UAV) to use single network connection, instead of connection steering mechanism, as one described in [109].

4.8. Energy Efficiency

Energy efficiency is defined as the number of bits that can be transmitted per joule of energy [216]. Compared with current wireless technologies, the energy efficiency (measured in b/J) of the 5G network may need to be improved by a factor of 1000 [4,7,34,110,217]. High energy efficiency is important in case of ticketing- and in some cases of tracking-based activities (e.g., smart industry [205] or implantable medical devices [35]). Other activities, such as both mMTC and uMTC monitoring- and managing/controlling-based activities (e.g., health condition [205]) and some tracking-based activities (e.g., sports wearables [205]) require medium energy efficiency. Energy efficiency is very important design objective for the reduction of operating costs of telecom operators, as well as for minimizing the environmental impact of the wireless domain [217]. On a higher layer of the network protocol stack, adaptive base station switch on/off algorithms use renewable energy sources to save energy [218] along with an energy scheduler, as in the with heating, ventilation and air conditioning (HVAC) [111]. However, at the physical layer, adaptively switching off unused carriers is a key strategy that can be used to save energy from the radio-frequency (RF) transceiver chain of base stations [218].

4.9. Spectrum Efficiency

Spectrum efficiency is defined as the data throughput per unit of spectrum resource per cell or per unit area (bps/Hz/cellar bps/Hz/km^2) [216]. In order to achieve network sustainability, required for 5G networks [216], spectrum efficiency needs to be improved 3–5 times [20,32,216]. Minimum peak spectrum efficiency is 30 bps/Hz for downlink and 15 bps/Hz for uplink [32]. This is mostly important for the xMBB services class [212,214]. Since this performance requirement is not relevant for activity-based IoT application classes which are associated with the mMTC and uMTC service class, it has not been further considered herein.

4.10. Summary of IoT in 5G Performance Requirements

Activity-based classes of IoT applications are associated with performance requirements of 5G service classes, i.e., mMTC and uMTC. On the basis of the literature review, each activity-based class (ticketing-, tracking-, monitoring-, and managing/controlling-based class) is associated with mMTC service class, while the monitoring- and managing/controlling-based classes are the only ones linked to the uMTC service class. According to the authors' best knowledge, there is no available literature concerning the uMTC ticketing- and tracking-based classes of IoT applications or such IoT customer service requests.

Activity-based IoT application classes pose many performance requirements, which have been discussed in terms of data rate, mobility (speed), latency, connection density, reliability, positioning accuracy, coverage, and energy efficiency. These performance requirements have been assigned three-level priorities (high, medium, low) for the purpose of facilitating identification of the enabling technologies used to fulfill them. Based on the literature review, it was found that each activity-based IoT application class poses high requirements in terms of connection density except the uMTC managing/controlling-based IoT application class. Moreover, ticketing- and tracking-based IoT application classes have high requirements in terms of positioning accuracy, coverage, and energy efficiency. Additionally, positioning accuracy is a highly important performance requirement for the uMTC managing/controlling-based IoT application class, whereas coverage is important for both uMTC and mMTC monitoring-based IoT application classes. Finally, latency and reliability represent highly important performance requirements for uMTC monitoring- and managing/controlling-based IoT application classes. A deep and comprehensive understanding of performance requirements of each activity-based IoT application class may facilitate the selection of 5G enabling technologies needed to meet them as described in the following section.

5. IoT in 5G Enabling Technologies

Activity-based classification of IoT applications proposed in Section 3 allows telecom operators to identify performance requirements of each class relying on 5G service classification, as discussed in Section 4. However, these performance requirements cannot be satisfied for many IoT applications with current cellular (2nd, 3rd, and 4th generation) network technologies, since they limit their potential due to many issues [3], i.e., protocol implementation complexity, poor coverage in non-urban environments, high cost of networking equipment and data transmission. In order to overcome these issues, many dedicated communication technologies are being installed.

From the very beginning of IoT, many proprietary technologies, such as radio-frequency identification (RFID), wireless highway addressable remote transducer (WirelessHART) or Z-Wave, have first appeared, and then, more generic ones, such as Bluetooth, IEEE 802.15.4, IPv6 over low-power wireless personal area networks (6LoWPAN). However, none of these technologies have become a market leader mainly because of technology shortcomings and business model uncertainty [219]. Hence, new solutions, such as low-power wireless fidelity (LP Wi-Fi), low-power wide area (LPWA) or several improvements for cellular M2M communications have become serious candidates for IoT implementation. LP Wi-Fi is an IEEE 802.11ah standard designed to extend the application area of Wi-Fi networks in order to meet IoT requirements (i.e., large number of devices, large coverage range, energy constrains). First performance studies indicate that this standard will support a broad range of M2M scenarios with a required QoS level, and enable scalable and cost-effective solutions. In addition, LPWA networks have been deployed for some time in the form of many different proprietary solutions (e.g., Amber Wireless, Coronis, Huawei's CIoT, LoRa, M2M Spectrum Networks, Sigfox, Weightless, etc.), but only the LoRa Alliance, Sigfox and Weightless are involved in LPWA standardization activities. Despite some drawbacks which are mainly related to use of an unlicensed spectrum, LPWA networks are expected to become a key enabler for IoT deployment in early market rollouts and for limited IoT applications [219].

The appearance of these IoT communication technologies was considered as competition to current cellular networks from the telco's point of view from the start. But, in the meanwhile, telcos have realized that the aforementioned communication technologies can be utilized to meet the changing connectivity and performance requirements of IoT applications. Therefore, many dedicated communication technologies have already been deployed in various IoT applications. For example, Orange, Swisscom and South Korea (SK) Telecom have built nationwide networks based on LoRa [3], whereas Deutsche Telekom (DT), Vodafone and all three Chinese operators have completed the rollout of narrowband IoT (NB-IoT), as another LPWA standard utilizing existing long term evolution (LTE) networks [36]. In this context, standardization and interoperability becomes critical because there is a need to consider a broad range of connectivity solutions as presented in Figure 4 [37]. One of those solutions includes satellite technologies which integrated with 5G radio technologies form the 6th generation (6G) standard for providing global coverage [38].

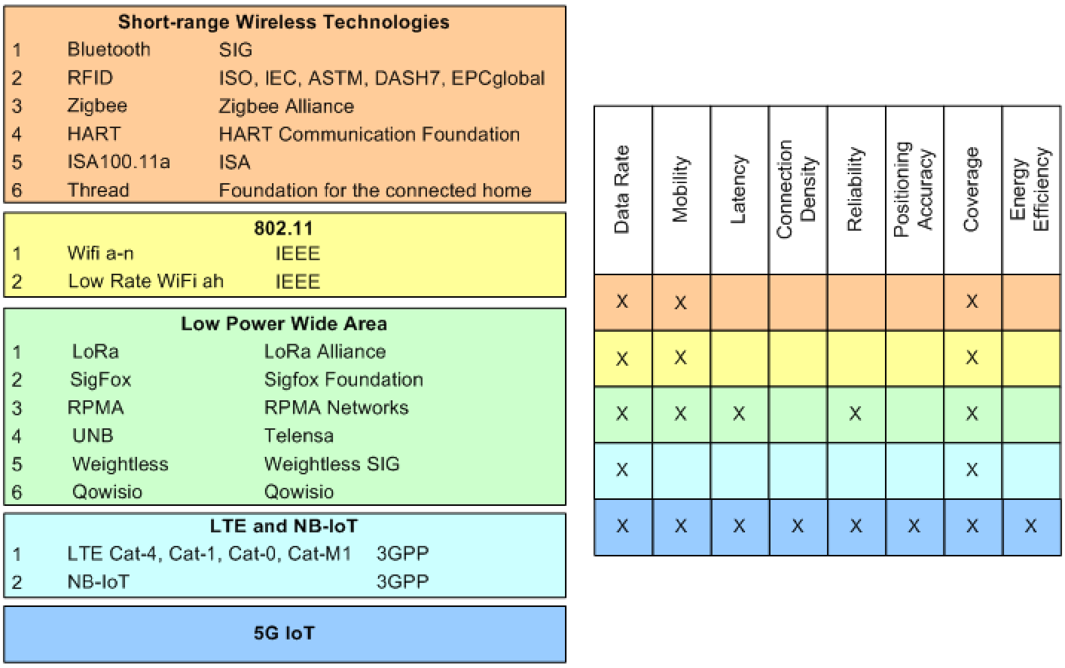

Short-range Wireless Technologies		
1	Bluetooth	SIG
2	RFID	ISO, IEC, ASTM, DASH7, EPCglobal
3	Zigbee	Zigbee Alliance
4	HART	HART Communication Foundation
5	ISA100.11a	ISA
6	Thread	Foundation for the connected home

802.11		
1	Wifi a-n	IEEE
2	Low Rate WiFi ah	IEEE

Low Power Wide Area		
1	LoRa	LoRa Alliance
2	SigFox	Sigfox Foundation
3	RPMA	RPMA Networks
4	UNB	Telensa
5	Weightless	Weightless SIG
6	Qowisio	Qowisio

LTE and NB-IoT		
1	LTE Cat-4, Cat-1, Cat-0, Cat-M1	3GPP
2	NB-IoT	3GPP

5G IoT

Data Rate	Mobility	Latency	Connection Density	Reliability	Positioning Accuracy	Coverage	Energy Efficiency
X	X					X	
X	X					X	
X	X	X		X		X	
X						X	
X	X	X	X	X	X	X	X

Figure 4. IoT performance requirements and enabling technologies.

In this regard, standardization activities of 5G and beyond are being undertaken by several standards bodies, such as the Institute of Electrical and Electronics Engineers (IEEE), the 3rd Generation Partnership Project (3GPP), the Internet Engineering Task Force (IETF), ITU Radiocommunication Sector (ITU-R), and ITU Telecommunication Standardization Sector (ITU-T). An overview on these standards bodies and their effort to develop communication standards for 5G and beyond are provided in [112]. Due to the envisioned 5G applicability, these standards have defined an air interface that can significantly improve the performance, and network architectures that allow the deployment and coexistence of various 5G technologies. A detailed discussion of these technologies regarding both technological and standardization aspects indicates that 5G can be perceived as the main driving force for enabling the vision of a truly global IoT [219].

In this sense, the technologies described in this section will allow 5G networks to form a unified communication infrastructure for the realization of a wide range of IoT applications. These technologies will bring performance enhancement which will trigger a complete revolution in almost all spheres of human life creating a new "everything connected" era. Future 5G networks in such an era will face a serious problem regarding a huge number of different service types [220]. To meet the diverse requirements for a huge number of IoT applications, 5G introduces the concept of network slicing to offer programmable network instances [220]. Using network slicing, multiple independent and dedicated network instances can be created within the same infrastructure to run services that have completely different requirements for latency, reliability, throughput, and mobility [220]. This technology primarily targets a portion of the 5G core network, but also indicates that 5G radio access network (RAN) functionalities can be different for each network slice [221]. To deploy network slices, network functions need to be enabled on an on-demand basis, which has driven the use of virtualization and softwarization in 5G core network [221]. Network function virtualization (NFV) is a paradigm that enables that 5G network functions to run in a virtual environment instead of dedicated hardware [221]. The software define networking (SDN) paradigm facilitates isolation of network slices avoiding the traffic of one slice affecting the performance of another slice [221]. These two paradigms make the network much more dynamic, agile, on-demand, and flexible [222,223]. Since the aforementioned technologies are used to deal with a variety of 5G applications, they will not be further discussed. This paper will focus on technologies used to meet performance requirements for specific applications (inter slice performance) which are implemented in the radio access part of 5G networks.

These technologies are briefly described and discussed in terms of their advantages, disadvantages, and research gaps. Given the fact that the realization of particular technology affects multiple performance requirements, tiny steps towards its implementation will allow telcos to deploy 5G networks, and thereby provide a performance improvement in terms of more capacity, lower latency, more mobility, more position accuracy, increased reliability, and availability [6,214,224]. In other words, 5G networks will accommodate many more users and devices while delivering more data [113,114] to each user requiring high data rates [108] in a more energy-efficient way [115,116].

Table 5 summarizes considered radio technologies used to implement future 5G networks being capable of achieving performance requirements identified in Section 4. On the basis of the literature review, one may conclude that the 5G radio access network is crowded with multiple technologies, and there seems to be a duplication of technologies, all eager to grab telco's attention and convince them to buy into the particular choice. As Table 5 shows, multiple technologies and techniques can be used to meet each performance requirement. Since the mMTC activity-based IoT classes are more mature, we have primarily focused on technologies used to fulfil their high performance requirements (i.e., connection density, positioning, and coverage). We have selected 5G enabling technologies considered from telco's perspective as the most representative and promising candidates to meet these requirements. The implementation of these technologies will also significantly affect the fulfillment of uMTC service class requirements, which will lead to complete automation in all spheres of human life. Therefore, the following subsections will discuss these technologies in more detail with the final goal of identifying research gaps and providing recommendations for future work, which are summarized in Table 6.

Table 5. IoT in 5G enabling radio technologies.

Enabling Technologies		IoT in 5G Service Requirements								
		Data Rate	**Mobility**	**Latency**	**Connection Density**	**Reliability**	**Positioning**	**Coverage**	**Energy Efficiency**	**Spectrum Efficiency**
Wide and flexible bandwidth technology	mmWave Band Communication and large-scale antenna	[7,202,225–227]		[202]			[233]	[5]	[202]	[7,29,214]
	Heterogeneous Multi-RAT Integration	[5,217,228–232]						[5]		
	Cognitive Radio and Spectrum Sharing							[5]		
Advanced modulation and coding	Advanced Modulation	[52,234–236]	[237,238]	[239]		[237,239]		[237]		
	Advanced Channel Coding	[235]		[235]						[216]
Duplexing	In-band FD			[124]				[5]		[7,29,214]
	Dynamic TDD	[240]		[240]						[9]
Multiple access and waveform	Multiple access			[216]					[218]	[7,9,29,216,256,257]
	New waveform				[5]					[7,9,29,216,256,257]
Advanced interface management	SND and SWSC	[5]								
Access architecture related radio technologies	Advanced small cell	[7,202]					[5]	[5,251]	[7,202,217]	[202,257]
	MN	[202]	[202]					[5]	[202]	
	Enhanced wireless backhaul							[5]		
	D2D	[202,248]	[7,241,244]					[245–247]	[7,202,215]	[7,9,29,202,252]
Energy related technologies	Energy harvesting						[100,253]			
	UAV	[254]				[254]		[183–185]		
Other important technologies	mMIMO	[7,202,249]		[13,202]		[249]		[200]	[217]	[7,9,214,216,257]
	VLC	[202,250]		[202]					[202,258]	
	SIC		[202]	[202,236]						
	V2X			[255]				[255]		

Legend: IoT (Internet of Things), 5G (Fifth Generation), mmWave (Milimeter-wave), RAT (Radio Access Technology), FQAM (Frequency Quadrature Amplitude Modulation), FD (Full Duplexing), TDD (Time Division Duplexing), SND (Simultaneous Non-unique Decoding), SWSC (Sliding Window Superposition Coding), MN (Moving Network), D2D (Device-to-Device), mMIMO (massive Multiple Input Multiple Output), VLC (Visible Light Communication), SIC (Self Interference Cancelation), V2X (Vehicle to Everything), UAV (Unmanned Aerial Vehicles).

Table 6. 5G enabling technologies—research directions.

5G Enabling Technologies		Research Gaps and Directions	References
Wide and flexible bandwidth technology	mmWave band communication and large-scale antenna	(1) 3D channel modeling; (2) dynamic power control; (3) user scheduling and congestion control; (4) hardware limitation and adaptive beam-steering technique; (5) design of mobility management and admission control for mmWave-based dense HetNet; (6) design of frequency management schemes for mmWave; (7) Tactile Internet; (8) effective and efficient mmWave implementation in HetNets (access and networking).	[7,87,217]
	Heterogeneous multi-RAT integration	(1) cell-association; (2) traffic-offloading algorithms; (3) interference management schemes in case of inter user and inter cell interference; (4) cross-tier handover, access admission, and mobility management schemes of a multi-tier HetNets.	[196,217]
	Cognitive radio and spectrum sharing	(1) Spectrum sensing (design of cooperative frameworks, choose cooperative secondary users and transmit cooperative information); (2) develop framework and algorithms for group handoff of secondary users and security; (3) simulation of different attacks and scenarios to enhance security; (4) in-depth performance analysis between GFDM and UFMC in CR settings.	[29,110]
Duplexing	In-band full duplexing	(1) Redesign of network and management; (2) antenna and circuit design and development of the theoretical foundation; (3) analyze the throughput of a network of randomly deployment terminals sing stochastic geometry; (4) characterize the capacity advantage due to IFDB in various network scenarios; (5) guidelines to practical design: coding, modulation, power allocation, beamforming, channel estimation, equalization, digital interference cancellation and decoding, (6) design of a MAC layer.	[55,197]
Multiple access and waveform	New waveform	(1) Performance of SIC cancelation or filtering on f-OFDMA; (2) balance of time and frequency dispersion and design an efficient filter prototype for UFMC.	
Access architecture related radio technologies	Advanced small cell	(1) Expect of wireless backhauling on user experience; (2) exploitation of location data and fingerprints in optimizing small cell discovery in terms of time and energy- efficiency; (3) interference management when integrating D2D and small cells.	[34,80]
	Enhanced wireless backhaul	(1) TDD multi-flow coordination schemes to avoid bottlenecks in the downlink backhaul; (2) Backhaul aware association in ultra-dense deployment; (3) reliability and security of the backhaul.	[34,88,198]
	Moving network	(1) Resource allocation and interference in the mobile relay when trains are moving from opposite directions; (2) handover decision of users (more than one train arrive or depart, stop or pass); (3) group mobility for users on board very high-speed vehicles; (4) design od cooperative communication schemes; (5) deployment of moving networks in various vehicle environments, not just on fixes route railways.	[5]
	D2D communication	(1) Interference management (mode selection, resource allocation and power control); (2) integration of novel reputation-based mechanism for identify and avoid malicious users from multiple users in multi-hop D2D communications; (3) testing of D2D interference management schemes in 5G scenarios (mmWave, cell densification).	[29,208,209]
Energy related technologies	Energy harvesting	(1) Improving energy harvesting schemes; (2) simulation of proposed models; (3) integration with other 5G technologies.	[100,182,253]
	UAV	(1) Optimal deployment, mobility and energy-efficient use of UAVs; (2) integration with other 5G technologies.	[81]
Other technologies	mMIMO	(1) Performance of practical mMIMO.	

Legend: IoT (Internet of Things), 5G (Fifth Generation), mmWave (Milimeter-wave), RAT (Radio Access Technology), D2D (Device-to-Device communication), mMIMO (massive Multiple Input Multiple Output), 3D (Three Dimensional), HetNet (Heterogeneous Network), GFDM (Generalized Frequency Division Multiplexing), UFMC (Universal Filtered Multi Carrier), CR (Cognitive Radio), IFDB (In-band Full Duplexing), MAC (Medium Access Control), f-OFDMA (filtered Orthogonal Frequency Division Multiple Access), TDD (Time Division Duplexing), UAV (Unmanned Aerial Vehicle).

5.1. Wide and Flexible Bandwidth Technology

mmWave band communication and large-scale antennas are promising technologies for future 5G networks. The mmWave band covers frequencies from 30 GHz to 300 GHz [202], and from an industry and wireless academia point of view, it is a unique solution for solving 5G capacity requirements [96]. mmWave band communications will provide high data rates [7,202,225–227] utilizing a much larger spectrum bandwidth that can reach up to 5 GHz [39] and by using directional antennas and high attenuation [40]. To provide sufficient antenna gain mmWave requires implementation of large-scale antennas at the transmitter and receiver side [96]. Deployment of large antenna arrays with mmWave will also bring high spectral efficiency, high throughput and channel gain [7]. Despite all of its advantages, mmWave needs line of sight (LOS) operation [41], while the effective communication distance of mmWave signals is within 200 m due to the propagation characteristic of this frequency band [29]. Open problems associated with mmWave include three dimensional (3D) channel modeling, dynamic power control, user scheduling and congestion control, hardware limitation and the adaptive beam-steering technique, as described in [7]. The problem of high power consumption of a large number of antennas in an array [7], high efficiency low complexity adaptive antenna array processing algorithms [29] and innovative hardware architecture of large-scale antenna transmitters [29] still remain unsolved. A proposal of design guidelines in architectures and protocols for mmWave communications is presented in [42]. It is demonstrated that new mmWave technologies, which are under investigation for 5G communications systems, will be able to provide indoor centimeter (cm)-accuracy localization in a robust manner, ideally suited for Assisted Living (AL) [228].

Heterogeneous multi-radio access technologies (multi-RAT) integration is specific to 5G networks radio design that coexists with existing networks. Since 5G networks will not be developed to replace current wireless networks, but rather to advance and integrate existing network infrastructures with new ones [217], we refer to 5G as a heterogeneous network. Multi-RAT is defined as the capability of a mobile network to support multiple radio access technology with seamless interworking among them [229]. When deployed in heterogeneous networks (HetNets), with traffic offloading among different RATs, multi-RAT improves capacity [230], supports better communication rates [34,117–119,231,232], better energy efficiency [34,117–119,231,232], and ensures seamless connectivity with higher QoS [230]. In designing a heterogeneous RAT, researches are finding solutions for cell-association and traffic-offloading algorithms [217].

Cognitive radio with spectrum sharing is a new software defined technology, which is expected to improve the utilization of the congested radio frequency (RF) spectrum [43]. In 5G networks, it is used for designing multi-tier architectures, removing interference among cells, and minimizing energy consumption in the network [44–47,120,233]. Moreover, a spectrum sharing technique can be used along with the CR technology to integrate the 5G spectrum [48,121]. During the practical implementation of the CR and spectrum sensing (SS) technique researchers had to design cooperative frameworks, choose cooperative secondary users and transmit cooperative information during the spectrum sensing, further developing the framework and algorithms for the group handoff of secondary users and enhancing security [29]. Researchers are already working on designs of protocols for different IoT application, based on cognitive radio, and some of them are presented in [49–51,122]. Technologies in this group are used together to optimize 5G performance requirements (e.g., a prototype of mmWave integrated HetNet in [123] and HetNet that incorporates massive multiple-input and multiple-output (mMIMO) and mmWave technologies [217]).

5.2. Advanced Modulation and Coding

This technological group involves advanced modulation (i.e., frequency and quadrature amplitude modulation (FQAM)) and advanced channel coding schemes.

Advanced modulation schemes group includes FQAM, Amplitude and Phase Shift Keying (APSK), Unitary Space-Time Modulation (USTM), Spatial Modulation (SM), Wave Modulation

(WAM), and Orthogonal Time Frequency and Space (OTFS). The FQAM is a combination of frequency shift keying (FSK) and quadrature shift keying (QAM) [124–126]. This modulation can achieve a higher transmission rate for cell edge users reducing interference at cell edge [52,234–236]. FQAM also improves energy efficiency, which makes it adequate for MTC devices with stringent energy consumption requirements [124]. Due to performance advantages in terms of Frame Error Rate (FER) [53,236], FQAM is an ideal candidate for services with high coverage and reliability requirements [237]. APSK is another modulation which draws a lot of attention. It is shown that its main performance gain (i.e., to achieve a channel capacity very close to Shannon's) relies on advanced channel coding and demodulation algorithms [237]. This adds more complexity at the APSK transmitter and receiver. Another modulation, which does not need Channel State Information (CSI) to enable high throughput, is called USTM [237]. USTM and its extension (see [238]) is very useful for 5G services with high mobility [127] or latency and reliability constraints [239]. Other modulation schemes involve SM [128], and proprietary WAM and OTFS [237]. On the basis of the foregoing, researches have analyzed and compared many modulation forms, but still have not found a practical guide on how to choose modulation in any of the 5G use cases.

Advanced channel coding techniques are used for correcting the communication errors caused by noise, interference, and poor signal strength [129]. Authors in [54,129,130] compared turbo, low density parity check (LDPC) and polar codes in decoders in contrast to 5G requirements: (1) maturity; (2) throughput and latency; (3) error correction capability; (4) flexibility; (5) computation complexity; (6) interconnect complexity; (7) high-performance flexible implementation complexity; and (8) backward compatibility. This comparison showed that turbo codes hold the greatest promise for offering high performance throughputs, latencies and error correction capabilities, as well as high degrees of flexibility at the lowest implementation complexity [129] in most 5G use cases. However, further work is needed to implement a decoder based on this code followed by detail analysis.

5.3. Duplexing

This technological group includes in-band full duplexing (FD) and dynamic time division duplexing (TDD).

In-band FD or simultaneous data transmission and reception will provide a 1000-fold increase in throughput [131,240], double spectral efficiency [132,240], and reduce the air interface delay [240]. The central research problem for the practical implementation of an in-band full-duplex radio is the attenuation of the self-interference signal by an adequate amount [132]. For the practical implementation of an in-band full-duplex radio, many aspects of network design and management need to be restructured, where terminals antenna and circuit design and the development of theoretical foundation are in focus [55]. One practical implementation of the full-duplex radio is shown in [133]. Authors in [134] commented that for the design of a full-duplex radio, it is necessary to unify researches from three domains, i.e., RF circuit and system design, digital signal processing and networking.

Dynamic TDD is the predecessor of FD transmission technology and a candidate for 5G [133,135]. It represents a scheduling technique in which every base station (BS) is free to choose its own uplink/downlink (UL/DL) split [136–139]. This technique is used to adapt the allocation of network resources to variable traffic requirements [139,140], often found in ultra-densely deployed networks [141]. The dynamic TDD can significantly increase bandwidth efficiency [142] and provide higher throughput and low latency [141]. However, it is characterized by the severe co-channel interference (CCI) [141–143]. Dynamic TDD could be used in combination with D2D communication and in the self-backhauling scenario, as described in [55].

5.4. Multiple Access and Waveforms

Multiple access techniques are becoming an important technology in 5G because of their ability to support mMTC activity-based IoT classes with urgent deploy demand [144]. They include several non-orthogonal multiple access (NOMA) forms: multi-user shared multiple access (MUSA) [145–148], resource spread multiple access (RSMA) [149], sparse code multiple access (SCMA) [150–152], pattern division multiple access (PDMA) [153–155], interleave-division multiple access (IDMA) [156,157], and NOMA by power domain [158]. The NOMA is a radio access technology design for enabling greater spectrum efficiency [56,159–162,207], higher cell-edge throughput, relaxed channel feedback, and low transmission latency [99]. NOMA can be employed to enhance user fairness and to support massive connections with diverse QoS requirements [163]. In NOMA, there are several challenges and open issues, such as dynamic user pairing, the impact of transmission distortion, the impact of interference, resource allocation, NOMA with multiple antennas, heterogeneous networks, outage probability analysis, practical channel model, uniform fairness, NOMA with antenna selection, carrier aggregation and other challenges as discussed in [216]. There are still many challenging issues for SCMA, which need to be solved in future work. For example, there are several open issues in SCMA transceiver design [57], the optimization of algorithms for user grouping and power allocation [164], and further enhancement of the SCMA and MIMO combination.

New waveforms have become a serious candidate for 5G being studied in terms of: (1) modulation based on pulse shaping: filter bank multicarrier (FBMC) [58], generalized frequency division multiplexing (GFDM) [59], pulse shaped OFDM [165] and QAM-FBMC [241]; (2) modulation based on sub-band filtering: universal filtered multi carrier (UFMC) [242], filtered OFDM (f-OFDM) [243] and resource block f-OFDM (RB-f-OFDM) [244]; other modulation format: guard interval discrete Fourier transform spread OFDM (GI DFT-s-OFDM) [245], spectrally-precoded OFDM (SP-OFDM) [246] and orthogonal time frequency and space (OTFS) [247]. The f-OFDM is seen as a potential candidate for IoT applications [60]. Due to narrow sub-bands and thereby pure detection performance, additional processing is needed [60]. Performance of self-interference cancellation (SIC) or filtering on f-OFDMA [60] remains a topic for further study. Additionally, it is interesting to consider the balance of time and frequency dispersion in UFMC, as well as the design of an efficient prototype filter according to application scenarios [60].

5.5. *Advanced Interface Management*

This technological group includes receiver advances in terms of simultaneous non-unique decoding (SND) and sliding-window superposition coding (SWSC).

SND follows a rule that implies that each receiver attempts to recover the code words from intended and interfering senders [248]. The combination of advanced receivers and joint scheduling provides an improvement of over 50% in cell edge throughput without sacrificing the cell average throughput [5]. This gain demonstrates that if 5G networks incorporate advanced interference management, they will provide a virtually edgeless end-user experience [5].

SWSC [61,62] combines the theory concept from superposition coding without rate splitting [63], block Markov coding [166,167], successive cancellation decoding [63,64,168], and sliding-window decoding [65,169,249]. The sliding-window coded modulation (SWCM) aims to mitigate inter cell interference at the physical layer by achieving simultaneous decoding performance with point-to-point channel codes, low-complexity decoding, and minimal coordination overhead [170]. The realization of the theoretical concept to the practical transmission coding scheme is an important research direction in SWSC [5].

5.6. *Access Architecture Related Radio Technology*

This technological group includes ultra-densification, enhanced wireless backhaul, moving networks, and D2D communication.

Advanced small cell [66,123] deployment is considered to be one of the key enablers for achieving many of the requirements currently envisioned for 5G [67], with higher data rates [68,171,250] and more capacity [171–174], as most important. Higher data rates and smaller battery consumption can be achieved using short distances in the small cell [69]. Solving the capacity and data rate challenge with network densification could be very expensive in terms of equipment, maintenance, and operations [200]. With denser cell deployment come significant challenges in the design of a high-performance backhauling system for RAN and the impact of backhaul on radio resource management (RRM) [70]. However, there are in existence many interesting topics to be further investigated which can be grouped around several problem areas [34]: user association, backhauling, interference management, energy efficiency, and propagation modelling.

Moving networks (MN) are the combination of multi-hop and vehicular communications concepts [69,71]. The MN deploys one or several moving relay node(s) (MRNs) [175,176,251] on vehicles that form their own cell(s) inside the vehicle to serve vehicular users [72]. Challenges in using MRN are efficient backhauling, design of efficient resource allocation and interference management techniques, as well as mobility management schemes to exploit the benefit of group handovers for vehicular user equipment (UE) devices served by the same MRN [72]. The main research challenges related to MN concept are associated with complexity management due to the mobility of access points and providing a high-rate wireless backhaul link from the moving cell to the fixed network [69].

Enhanced wireless backhaul is one of the main challenges in hyper-dense 5G networks [34,217]. The performance of wireless backhaul is dictated by the environment and traffic profile of the intended use case [9]. In UDN except mmWave, new approaches are needed, e.g., interface aware routing and intelligent resource allocation [9], as these links may occupy part of the spectrum used in the access network [73]. Latency and reliability of this link are important issues to be considered being prone to blocking and fading [69]. In any case, further research is necessary to explore TDD multi-flow coordination schemes to avoid bottlenecks in the downlink backhaul [9]. The guidelines for deploying future 5G wireless backhaul networks in economical and highly energy-efficient ways are provided in [74].

D2D communication is defined as a direct route of data traffic between spatially closely located mobile UE [75,177,178]. The advantages of D2D communication compared with the traditional cellular method include the reduction of transmission latency [73,75] and power consumption [75], improvements in coverage [71,76,179], spectral and energy efficiency [77], and throughput, when power control and resource allocation methods are used [78]. The use of D2D communications has an overall positive impact on system capacity in cellular environments. D2D communication brings challenges and complexities related to interference management [79,80], QoS requirements [79], and resource allocation [34,215]. The design of D2D direct communication link is still a hot topic [29].

5.7. Energy Related Technologies

IoT applications span a broad range of domains including home automation, healthcare, surveillance, transportation, smart environments, etc. One of the most important obstacles for implementing such an impressive scheme is supplying adequate energy to operate the network in a self-sufficient way without compromising QoS [180]. Since energy efficiency is of most importance to battery constrained IoT devices, researches have focused their work on the development of the device energy saving mechanisms.

Energy harvesting is a new paradigm which uses solar, thermal, wind and kinetic energy sources to power sensor nodes and consequently prolong network lifetime [181]. Among different energy harvesting methods, wireless energy harvesting (WEH) has proven to be one of the most promising solution for energy aware IoT devices, because of its simplicity, ease of implementation, and availability [180]. This concept will improve some of the 5G high network communication requirements, like reliability of IoT communications, as presented in [182]. Authors in [252] have shown that the integration of social awareness and energy harvesting in D2D communication results in higher system capacity. Network architectures proposed in [100,253] have shown better energy efficiency in 5G wireless network. Researchers could work on improving this architecture and integration with various 5G technologies in the future.

UAV implies a flying vehicle without the driving presence of a human pilot to control it on-board [113]. Drones are a possible example of UAVs, although many other robotics related applications could be part of this application group [113]. UAVs can play a vital role in IoT scenarios where devices are unable to transmit over a long distance due to their energy constraints. In this case UAVs play the role of moving aggregators which fly toward one IoT devices, collect the data, and transmit it to other devices [81]. In [183] authors presented UAV–based floating relay (FR) for cell dynamic and coverage improvement. Deployment of unmanned aerial base stations in 5G heterogeneous architecture can improve throughput [254], coverage [183–185,254], connectivity [184,185], and 5th percentile spectral efficiency [254]. On-demand wireless systems with low-altitude UAVs are in faster to deploy, more flexibly reconfigured, and likely to have better communication channels due to the presence of short-range line-of-sight links [186] when compared to terrestrial communications or those based on high-altitude platforms. The utilization of highly mobile and energy-constrained UAVs for wireless communication introduces many new challenges [186], some of them are listed in [82]. To effectively use UAVs for IoT, several challenges must be addressed such as optimal deployment, mobility and energy-efficient use of UAVs [81].

5.8. Other Technologies

Along with the aforementioned technologies, a lot of research attention is devoted to technologies, such as mMIMO, visible light communication (VLC), SIC, vehicle-to-everything (V2X), etc.

mMIMO technology [73,123] promises significant gains in data rate and link reliability [187], reduces latency and energy [13], simplifies media access control (MAC) layer [13], shows robustness against intentional jamming [13], unintended man-made interface [83], and increases capacity [127,198] due to spatial multiplexing [123]. Reliable links are provided by benefiting from spatial diversity and the mitigating effects of fast fading, beamforming, and zero forcing caused by multi-user interference [203]. The mMIMO can be exploited to extend the coverage of higher frequency bands by relying on beamforming gains [200]. Other specific benefits of mMIMO system are: increased capacity 10 times or more with simultaneous improvement in radiated energy efficiency in the order of 100 times, the possibility to use inexpensive, low-power components, the reduction of latency on the air interface, and multiple access layer simplification [83]. The low complexity mMIMO uplink scheme for IoT lightweight devices is presented in [188]. The performance of practical mMIMO system needs to be investigated, since the research community pays attention to analyzing the performance of mMIMO in an ideal channel state information assumption [73].

VLC is a growing technology for short range, high capacity LOS optical links [189]. It uses a visible range of the electromagnetic spectrum (370–780 nm) which provides data transmission and room illumination using light emitting diodes (LEDs). Prominent features of VLC are an abundant license-free spectrum, the ability to provide multiple gigabit-per-second data rates, low energy consumption, and low implementation costs [190]. VLC is sensitive to sunlight and is not able to work long range without LOS. Since VLC coverage is LOS limited [84,191], this is known as LOS blocking [192]. With poor performance in non-line-of-sight scenarios, VLC networks fail to provide convenient UL coverage at the current state-of-the-art, and each AP illuminates only a small confined cell compared to cellular RF networks [192]. VLC applications are expected to include IoT, wireless Internet access, and vehicle-to-vehicle (V2V) communications, broadcast from LED, M2M communications, positioning systems, and navigation. The problem of how to provide mobile applications over VLC is still quite an open one [189].

SIC implementation enables low-latency applications in a cost effective manner [132]. SIC can be used to increase link capacity, spectrum virtualization, any-division duplexing, novel relay solutions, and enhanced interface coordination [132]. Practical SIC implementation is still in the investigation phase.

V2X is the communication between a vehicle and everything with which it might interact (e.g., other vehicles, traffic operators and service providers). It is used in conjunction with D2D for coverage extension and latency reduction in 5G networks [255]. V2X communication might play a vital role in terms of implementing smart and efficient traffic solutions [193], improving road safety, traffic efficiency, and driver convenience. To reduce traffic congestion one may use an intelligent route management system based on V2X communication [7]. Since it has big potential in autonomous driving application, the localization of vehicles in dense urban areas is difficult, so several algorithms are proposed to solve the problems [194]. Many V2X applications are described in existing literature [85,195]. However, there is still a need to design vehicular mobility management strategies in next UDN networks and to find a way to enhance existing connected vehicle services with V2X [86].

These technologies are highlighted as the main 5G driver in existing literature [5–7,96,207]. Authors compare these technologies and try to implement them through prototypes, or test the combination of these technologies. However, other technologies exist, such as cloud-RAN [7], Coordinated Multi-Point (CoMP), White Space Spectrum [208], which are not in the scope of this paper.

5.9. Summary of IoT in Enabling Technologies

Performance requirements of activity-based IoT application classes have served to identify the enabling technologies in the radio access part of 5G networks, which are summarized in Table 5. Various 5G enabling technologies have been sorted into eight technological groups:

1. Wide and flexible bandwidth technology: mmWave band communication and large-scale antennas, heterogeneous multi-RAT integration, cognitive radio, and spectrum sharing;
2. Advanced modulation and coding: advanced modulation schemes, advanced channel coding;
3. Duplexing: in-band FD, dynamic TDD;
4. Multiple access and waveform: multiple access, new waveforms;
5. Advanced interface management: SND, SWSC;
6. Access architecture related radio technologies: advanced small cell, MN, enhanced wireless backhaul, D2D;
7. Energy related technologies: energy harvesting, UAV;
8. Other technologies: mMIMO, VLC, SIC, V2X.

These technologies have been discussed in terms of their possibilities to meet performance requirements of activity-based IoT application classes and identify research gaps and directions, which are summarized in Table 6.

We have highlighted technologies only used to satisfy high priority performance requirements of mMTC activity-based IoT application classes, i.e., connection density, positioning accuracy, and coverage. We have focused on these technologies from telco's perspective, since the area of mMTC activity-based IoT classes is more mature, as its development has already started within LTE, while uMTC poses new research questions to be answered in order to achieve unprecedented levels of reliability needed for new applications in 5G.

The new waveform technology may be used to fulfil the requirements of almost every activity-based IoT application class in terms of connection density. This technology affects the connection density by definition. Analysis of different modulation formats indicates that f-OFDM is the most suitable candidate for IoT applications with high connection density requirements. In this regard, a deeper and more comprehensive analysis of f-OFDM is needed in the context of different mMTC activities (i.e., ticketing, tracking, monitoring, and managing/controlling). The result of the analysis to be performed should be the discovery of an optimal waveform for each activity-based IoT class which has to be integrated with other 5G service classes.

Large scale antennas and advanced small cell technologies can be used to satisfy the requirements of ticketing-, tracking-, and uMTC managing/controlling-based IoT application classes in terms of positioning accuracy. Fulfilling these performance requirements is rather a business than a technical decision. The size of the 5G cell must be planned to meet the desired position accuracy since it is very important for future IoT applications, such as automated driving, remote surgery, robotics, taxis, etc.

Wide and flexible bandwidth technologies, in-band full duplex technologies, access architecture related radio technologies, and mMIMO can be used to achieve the high requirements of ticketing-, tracking-, and monitoring-based IoT application classes in terms of coverage. The combination of these technologies should be tested to find the optimal solution to meet these performance requirements. Since small cell deployment has an impact on required position accuracy, they should serve as a basis for achieving the desired level of coverage. Along with small cell deployment, mMIMO and mmWave technologies are present in almost every 5G prototype.

Since this section discusses radio technologies used to accomplish performance requirements of mMTC activity-based IoT application classes, it can serve telecom operators and other interested parties depending on their interest and potential to utilize these findings towards IoT implementation and monetization. It describes each technology in terms of definition, advantages, disadvantages, and the possible impact on the performance requirements of a broad range of IoT applications that drive the deployment of 5G networks. Moreover, this section can be useful to the research community interested in this attractive field to address recognized research gaps and directions.

6. Discussion

Telecom operators have the opportunity to capture a share of the revenue that is generated by IoT implementation depending on their role in the IoT value chain. Regardless of their role, telcos have to face many technical challenges in order to meet the changing connectivity and performance requirements. Since current cellular networks limit numerous IoT applications, new technologies are being introduced leading towards 5G networks. Therefore, telcos have to focus on deploying these network technologies in order meet the changing requirements necessary to achieve success in IoT.

In this regard, writing this paper was motivated by the challenge of providing an enhanced understanding of the scope and classification of the broad range of IoT applications in order to determine and prioritize their performance requirements with the goal of specifying the enabling technologies towards 5G networks. The aim has been to propose a new classification of IoT applications, define and prioritize the performance requirements of such IoT application classes, and give insight into state-of-the-art technologies used to meet these requirements from telco's point of view. The motivation that led to the focus of this paper being on IoT applications classification, performance requirements, and 5G enabling technologies could be explained by telco's need for added value from IoT services. Hence, an analysis of IoT customer service requests inspired us to propose an activity-based classification of IoT applications as it can, according to authors' best knowledge, has allowed telcos to more precisely specify their performance requirements and 5G enabling technologies, thereby improving customer experience.

In this regard, the paper fulfilled the following four objectives: (1) the identification of activities relevant to IoT applications and their usage as a new criterion for IoT application classification; (2) the specification and prioritization of performance requirements of such IoT applications classes; (3) the analysis of the radio technologies used to accomplish IoT application requirements; and (4) the identification of the research gaps and the recommendation of new research directions.

Through fulfilling these objectives, we reviewed literature from the fields of IoT in 5G service classifications, IoT in 5G performance requirements, and IoT in 5G enabling technologies. Since the first aim of this paper was to propose a new classification of IoT applications, the existing approaches have been summarized to serve as a basis to identify their drawbacks and formulate the appropriate solution to overcome them. In this context, we proposed a new approach to IoT applications classification, which was based on the activity as a new classification criterion denoting the main function/purpose of specific IoT application. This approach enabled a clearer and more precise positioning of particular application in the IoT application spectrum, as well as the determination of performance requirements and enabling technologies from telco's point of view. The activity-based IoT application classification facilitated the specification of the performance requirements and determination of technologies which enable these requirements to be fulfilled. In this context, the proposed approach served as a basis for the simplification of the realization of particular IoT application. For example, although the monitoring- and tracking-based applications seem to be similar, they set different performance requirements on the network, thereby allowing the proposed approach to precisely grade the IoT applications from the same IoT application domain.

Furthermore, we associated the activity-based IoT application classes with the 5G service classes, i.e., mMTC and uMTC, in order to specify and prioritize their performance requirements. It was determined that almost each activity-based IoT application class poses high requirements in terms of connection density, whereas ticketing-based and tracking-based IoT application classes additionally require high positioning accuracy and coverage. In order to analyze these performance requirements, we have summarized the radio technologies used to implement the future 5G networks. The focus was on the technologies used to meet the aforementioned performance requirements of mMTC activity-based IoT application classes, since mMTC is more mature, as its deployment has already started, while uMTC requires further research to achieve incomparable levels of reliability needed to enable new applications in 5G.

The analysis showed that new waveform technology can be used to meet the requirements in terms of connection density; large-scale antennas and advanced small cell technologies can be used for the purpose of satisfying the requirements in terms of positioning accuracy; while wide and flexible bandwidth technologies, in-band full duplexing technologies, and access architecture related radio technologies (i.e., advanced small cell, enhanced wireless backhaul, moving network, and D2D communication) can be used to achieve high requirements in terms of coverage. In addition, the conducted research study allowed us to highlight a number of open research issues that could serve the research community and other stakeholders interested in this contemporary and attractive field. In this context, we have recognized research gaps and directions which mostly relate to network redesign and optimization in order to accommodate large-scale IoT applications.

7. Conclusions

The IoT paradigm has the potential to revolutionize all areas of daily life of individuals, businesses, and society as a whole. Telcos enjoy a central role in the paradigm of IoT because of owing communication infrastructure which is exposed to the numerous technical challenges due to changing connectivity and performance requirements of various IoT applications. These requirements cannot be met with the current cellular networks which create the need to introduce new types of technologies leading toward 5G networks as the main driver for enabling numerous IoT applications. In this context, this paper proposed the activity-based classification of IoT applications and specification of their performance requirements in order to identify 5G radio technologies used to meet them, all from telco’s perspective. Activity-based classification of IoT applications indirectly allowed the construction and distribution of performance, as IoT value proposition, while using 5G enabling technologies, as IoT resources in telco’s internal value chain. In this context, high performance requirements of each activity-based IoT application class (i.e., connection density, positioning accuracy, coverage) served as a basis to analyze various enabling technologies in radio access part of 5G network in terms of advantages, disadvantages, and research gaps.

On the basis of the conducted analysis, we concluded that the following technologies can meet the high performance requirements of mMTC activity-based IoT application classes: new waveform technology (in terms of connection density), large scale antennas (in terms of positioning accuracy), wide and flexible bandwidth technologies, in-band full duplexing technologies, and access architecture related radio technologies (in terms of coverage). According to the identified research gaps, one may conclude that the optimal solutions for each technology are still in its infancy. This implies that any 5G enabling technology should be first implemented and tested through prototype construction before the deployment of specific IoT application. Thereafter, the technology combination that meets the performance requirements of each activity-based IoT class should first be found. Finally, the practical implementation of different technological combinations may further lead to a deep and comprehensive analysis of QoS and QoE in the context of IoT applications.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Author Contributions: Enida Cero and Jasmina Baraković Husić conceived and designed the experiments (research study); Enida Cero performed the experiments (research study); Sabina Baraković analyzed the data and contributed materials; Enida Cero and Jasmina Baraković Husić wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Antonopoulos, A.; Perillo, C.; Verikoukis, C. Internet Service Providers vs. Over-the-Top Companies: Friends or Foes?—Short talk. *SIGMETRICS Perform. Eval. Rev.* **2016**, *44*, 37. [CrossRef]
2. Antonopoulos, A.; Kartsakli, E.; Perillo, C.; Verikoukis, C. Shedding Light on the Internet: Stakeholders and Network Neutrality. *IEEE Commun. Mag.* **2017**, *55*, 216–223. [CrossRef]
3. IoT-Ignite. Role of Telcos in Internet of Things. Available online: https://devzone.iot-ignite.com/wp-content/uploads/2017/01/Role-of-Telcos-in-IoT.pdf (accessed on 5 September 2017).
4. Kaur, S.; Singh, I. A Survey Report on Internet of Things Applications. *Int. J. Comput. Sci. Trends Technol.* **2016**, *4*, 330–335.
5. 5G Forum. 5G Vision, Requirements, and Enabling Technologies. Available online: http://kani.or.kr/5g/whitepaper/5G%20Vision,%20Requirements,%20and%20Enabling%20Technologies.pdf (accessed on 10 July 2017).
6. Saha, R.K.; Saengudomlert, P.; Aswakul, C. Evolution towards 5G Mobile Networks—A survey on Enabling Technologies. *Eng. J.* **2016**, *20*, 87–112. [CrossRef]
7. Akyildiz, I.F.; Nie, S.; Lin, S.C.; Chandrasekaran, M. 5G roadmap: 10 Key Enabling Technologies. *Comput. Netw.* **2014**, *106*, 17–48. [CrossRef]
8. NGMN Alliance. 5G White Paper-Executive Version. Available online: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf (accessed on 10 July 2017).
9. Popovski, P.; Mange, G.; Gozalves-Serrano, D.; Rosowski, T.; Zimmermann, G.; Agyapong, P.; Fallgren, M.; Hoglund, A.; Queseth, O.; Tullberg, H.; et al. Deliverable D6.6 Final Report on the METIS 5G System Concept and Technology Roadmap. Available online: https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D6.6_Summary.pdf (accessed on 10 July 2017).
10. Aly, H.; Elmogy, M.; Barakat, S. Big Data on Internet of Things: Applications, Architecture, Technologies, Techniques, and Future Directions. *Int. J. Comput. Sci. Eng.* **2015**, *4*, 300–313.
11. Vermesan, O.; Friess, P. *Internet of Things—From Research and Innovation to Market Deployment*; Riverbed: Aalborg, Denmark, 2014.
12. Amaral, L.A.; de Matos, E.; Tiburski, R.T.; Hessel, F.; Lunardi, W.T.; Marczak, S. Middleware Technology for IoT Systems: Challenges and Perspectives Toward 5G. In *Internet of Things (IoT) in 5G Mobile Technologies*; Mavromoustakis, C., Mastorakis, G., Batalla, J., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 333–367.
13. OECD. Machine-to-Machine Communications: Connecting Billions of Devices. Available online: http://www.oecd-ilibrary.org/docserver/download/5k9gsh2gp043-en.pdf?expires=1499694889&id=id&accname=guest&checksum=F403637E723693581F03EA9D6F3A3C3B (accessed on 10 July 2017).
14. Zheng, K.; Hu, F.; Wang, W.; Xiang, W. Radio Resource Allocation in LTE Advanced Cellular Networks with M2M Communications. *IEEE Commun. Mag.* **2012**, *50*. [CrossRef]
15. Zhang, Q.; Fitzek, F.H.P. Mission Critical IoT Communication in 5G. In *Future Access Enablers for Ubiquitous and Intelligent Infrastructures*; Springer International Publishing: Cham, Switzerland, 2015; Volume 159, pp. 35–41.
16. Oughton, E.J.; Frias, Z. Exploring the Cost, Coverage and Rollout Implications of 5G in Britain. Available online: http://www.itrc.org.uk/wp-content/uploads/Exploring-costs-of-5G.pdf (accessed on 10 July 2017).
17. Chen, Y.; Wang, W. Machine-to-Machine Communication in LTE-A. In Proceedings of the 2010 IEEE Vehicular Technology Conference (VTC 2010-Fall), Ottawa, ON, Canada, 6–9 September 2010.
18. Booysen, M.J.; Gilmore, S.J.; Zeadally, S.; van Rooyen, G.J. Machine-to-Machine (M2M) Communications in Vehicular Networks. *KSII Trans. Internet Inf. Syst.* **2012**, *6*, 529–546. [CrossRef]

19. Niyato, D.; Xiao, L.; Wang, P. Machine-To-Machine Communications for Home Energy Management System in Smart Grid. *IEEE Commun. Mag.* **2011**, *49*, 53–59. [CrossRef]
20. De Matos, W.D.; Gondim, P.R.L. M-Health Solutions Using 5G Networks and M2M Communications. *IT Prof.* **2016**, *18*, 24–29. [CrossRef]
21. Investopedia. Business Model. Available online: http://www.investopedia.com/terms/b/businessmodel.asp (accessed on 3 September 2017).
22. Gassmann, O.; Frankenberger, K.; Csik, M. Revolutionizing the Business Model. In *Management of the Fuzzy Front End of Innovation*; Springer: New York, NY, USA, 2014.
23. Vermesan, O.; Bahr, R.; Gluhak, A.; Boesenberg, F.; Hoeer, A.; Osella, M. D02.01 IoT Business Models Framework. Available online: http://www.internet-of-things-research.eu/pdf/D02_01_WP02_H2020_UNIFY-IoT_Final.pdf (accessed on 3 September 2017).
24. Dijkmana, R.M.; Sprenkelsa, B.; Peetersa, T.; Janssenb, A. Business Models for the Internet of Things. *Int. J. Inf. Manag.* **2015**, *35*, 672–678. [CrossRef]
25. Farooq, M.O.; Sreenan, C.J.; Brown, K.N.C. Research Challenges in 5G Networks: A HetNets Perspective. In Proceedings of the 19th International ICIN Conference Innovations in Clouds, Internet and Networks, Paris, France, 1–3 March 2016.
26. Hu, F. 5G Overview: Key Technologies. In *Opportunities in 5G Networks*, 1st ed.; Hu, F., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 1–557.
27. Ford, R.; Zhang, M.; Mezzavilla, M.; Duttam, S.; Rangap, S.; Zorzi, M. Achieving Ultra-Low Latency in 5G Millimeter Wave Cellular Networks. *IEEE Commun. Mag.* **2017**, *55*, 196–203. [CrossRef]
28. Prasad, R. IMT for 2020 and Beyond. In *5G Outlook—Innovations and Applications*; Riverbed Publishers: Gistrup, Denmark, 2016; pp. 1–260.
29. Ma, Z.; Zhang, Z.; Ding, Z.; Fan, P.; Li, H. Key Techniques for 5G Wireless Communications: Network Architecture, Physical Layer, and MAC Layer Perspectives. *Sci. China Inf. Sci.* **2015**, *50*, 1–20. [CrossRef]
30. Rappaport, T.S.; Daniels, R.C.; Heath, R.W.; Murdock, J.N. Introduction. In *Millimeter Wave Wireless Communications*; Pearson Education: Upper Saddle River, NJ, USA, 2014.
31. Ge, X.; Chen, J.; Ying, S.; Chen, M. Energy and Coverage Efficiency Trade-off in 5G Small Cells Network. *arXiv* **2016**, arXiv:1612.04459.
32. Hossain, S. 5G Wireless Communication Systems. *Am. J. Eng. Res.* **2013**, *2*, 344–353.
33. Artiga, X.; Nunez-Martinez, J.; Perez-Neira, A.; Lendrino Vela, G.J.; Fare Garcia, J.M.; Ziaragkas, G. Terrestrial-satellite Integration in Dynamic 5G Backhaul Networks. In Proceedings of the 2016 8th Advanced Satellite Multimedia Systems Conference and the 14th Signa (ASMS/SPSC), Palma de Mallorca, Spain, 5–7 September 2016.
34. Kamel, M.; Hamouda, W.; Youssef, A. Ultra-Dense Networks: A Survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 2522–2545. [CrossRef]
35. Kartsakli, E.; Lalos, A.S.; Antonopoulos, A.; Tennina, S.; Di Renzo, M.; Alonso, L.; Verikoukis, C. A Survey on M2M Systems for mHealth: A Wireless Communications Perspective. *Sensors* **2014**, *10*, 18009–18052. [CrossRef] [PubMed]
36. Morris, A. DT, Chinese Operators take NB-IoT to Market. Available online: http://www.lightreading.com/iot/nb-iot/dt-chinese-operators-take-nb-iot-to-market/d/d-id/734022 (accessed on 6 September 2017).
37. Vermesan, O.; Eisenhauer, M.; Sunmaeker, H.; Guillemin, P.; Serrano, M.; Tragos, E.Z.; Valino, J.; van der Wees, A.; Gluhak, A.; Bahr, R. Internet of Things Cogitive Transformation Technology Research Trends and Applications. In *Cognitive Hyperconnected Digital Transformation*; Vermesan, O., Bacquet, J., Eds.; River Publishers: Delft, The Netherlands, 2017; pp. 17–95.
38. Gawas, A.U. An Overview on Evolution of Mobile Wireless Communication Networks: 1G–6G. *Int. J. Recent Innov. Trends Comput. Commun.* **2015**, *3*, 3130–3133. [CrossRef]
39. Qi, Y.; Hunukumbure, M.; Nekovee, M.; Sgardoni, V. Quantifying Data Rare and Bandwith Requirements for Immersive 5G Experience. In Proceedings of the 2016 IEEE International Conference on Communications Workshops, Kuala Lumpur, Malaysia, 23–27 May 2016.
40. Sim, G.H.; Asadi, A.; Loch, A.; Widmer, J. Opp-Relay: Managing Directionaly and Mobility Issues of Millimeter-Wave via D2D Communication. In Proceedings of the I9th International Conference on COMSNETS 2017, Bangalore, India, 4–8 January 2017.

41. Mer, S.B. Smart Vehicle-to-Vehicle Communication with 5G Technology. *Int. J. Recent Innov. Trends Comput. Commun.* **2015**, *3*, 3241–3244.
42. Niu, Y.; Li, Y.; Jin, D.; Su, L.; Vasilakos, A.V. A Survey of Millimeter Wave Communications (Mmwave) for 5G: Opportunities and Challenges. *Wirel. Netw.* **2016**, *21*, 2657–2676. [CrossRef]
43. Rusek, F.; Persson, D.; Lau, B.K.; Larsson, E.G.; Marzetta, T.L.; Edfors, O.; Tufvesson, F. Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays. *IEEE Signal Process. Mag.* **2012**, *30*, 1053–5888. [CrossRef]
44. Hong, X.; Wang, J.; Wang, C.X.; Shi, J. Cognitive Radio in 5G: A Perspective on Energy-Spectral Efficiency Trade-Off. *IEEE Commun. Mag.* **2014**, *52*, 46–53. [CrossRef]
45. Huang, L.; Zhu, G.; Du, X. Cognitive Femtocell Networks: An Opportunistic Spectrum Access for Future Indoor Wireless Coverage. *IEEE Wirel. Commun.* **2013**, *20*, 44–51. [CrossRef]
46. Lee, C.H.; Shih, C.Y. Coverage Analysis of Cognitive Femtocell Networks. *IEEE Wirel. Commun. Lett.* **2014**, *3*, 177–180. [CrossRef]
47. Wang, W.; Yu, H.; Huang, A. Cognitive Radio Enhanced Interference Coordination for Femtocell Networks. *IEEE Commun. Mag.* **2013**, *51*, 37–43. [CrossRef]
48. Menon, R.; Buehrer, R.M.; Reed, J.H. On the Impact of Dynamic Spectrum Sharing Techniques on Legacy Radio Systems. *IEEE Trans. Wirel. Commun.* **2008**, *7*, 4198–4207. [CrossRef]
49. Zhang, Y.; Yu, R.; Nekovee, M.; Liu, Y.; Xie, S.; Gjessing, S. Cognitive Machine-to-Machine Communications: Visions and Potentials for the Smart Grid. *IEEE Netw.* **2012**, *26*, 6–13. [CrossRef]
50. Aijaz, A.; Aghvami, A.H. Cognitive Machine-To-Machine Communications for Internet-of-Things: A Protocol Stack Perspective. *IEEE Internet Things J.* **2015**, *2*, 103–112. [CrossRef]
51. Li, R.; Lu, B.; McDonald-Maier, K.D. Cognitive Assisted Living Ambient System: A Survey. *Digit. Commun. Netw.* **2015**, *1*, 229–252. [CrossRef]
52. Andrews, J.G.; Buzzi, S.; Choi, W.; Hanly, S.V.; Lazano, A.; Soong, A.C.K.; Zhang, J.C. What Will 5G Be? *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1065–1082. [CrossRef]
53. Samsung Electronics. 5G Vision, Version 2. Available online: http://www.samsung.com/global/business-images/insights/2015/Samsung-5G-Vision-0.pdf (accessed on 10 July 2017).
54. Iscan, O.; Lentner, D.; Xu, W. A Comparison of Channel Coding Schemes for 5G Short Message Transmission. In Proceedings of the IEEE Globecom Workshops (GC Wkshps), Washington, DC, USA, 4–8 December 2016.
55. Sabharwal, A.; Schniter, P.; Guo, D.; Bliss, D.W.; Rangarajan, S.; Wichman, R. In-Band Full-Duplex Wireless: Challenges and Opportunities. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1637–1652. [CrossRef]
56. Ding, Z.; Liu, Y.; Choi, J.; Sun, Q.; Elkashlan, M.; Chin-Lin, I.; Vincent Poor, H. Application of Non-Orthogonal Multiple Access in LTE and 5G Networks. *IEEE Commun. Mag.* **2006**, *55*, 185–191. [CrossRef]
57. Lei, W. Non-Orthogonal Multiple Access Schemes for Future Cellular Systems. Ph.D. Thesis, University of Surrey, Surrey, UK, 17 May 2017.
58. Farhang-Boroujeny, B. OFDM Versus Filter Bank Multicarrier. *Signal Process. Mag.* **2011**, *28*, 92–112. [CrossRef]
59. Bellanger, M. FS-FBMC: An Alternative Scheme for Filter Bank Multicarrier Transmission. In Proceedings of the 2012 5th International Symposium on Communications Control and Signal Processing, Rome, Italy, 2–4 May 2012.
60. Cai, X.; Qin, Z.; Cui, F.; Li, G.Y.; McCann, J.A. Modulation and Multiple Access for 5G Network. Available online: https://arxiv.org/pdf/1702.07673.pdf (accessed on 10 July 2017).
61. Wang, L.; Sasoglu, E.; Kim, Y.H. Filtered OFDM: Sliding-Window Superposition Coding for Interference Networks. In Proceedings of the International Symposium Information Theory, Honolulu, HI, USA, 29 June–4 July 2014.
62. Wang, L. Channel Coding Technique for Network Communication. Ph.D. Thesis, University of California, San Diego, CA, USA, December 2015.
63. Cover, T.M. Broadcast Channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 2–14. [CrossRef]
64. Liao, H.H.J. Multiple Access Channels. Ph.D. Thesis, University of Hawaii, Honolulu, HI, USA, September 1972.
65. Kramer, G.; Gastpar, M.; Gupta, P. Cooperative Strategies and Capacity Theorems for Relay Networks. *IEEE Trans. Inf. Theory* **2005**, *51*, 3037–3063. [CrossRef]

66. Bhushan, N.; Li, J.; Malladi, D.; Gilmore, R.; Brenner, D.; Damnjanovic, A.; Sukhavasi, R.; Patel, C.; Geirhofer, S. Network Densification: The Dominant Theme for Wireless Evolution into 5G. *IEEE Commun. Mag.* **2014**, *52*, 82–89. [CrossRef]
67. Chin, W.H.; Fan, Z.; Haines, R. Emerging Technologies and Research Challenges for 5G Wireless Networks. *IEEE Wirel. Commun.* **2014**, *21*, 106–112. [CrossRef]
68. Biral, A.; Centenaro, M.; Zanellan, A. The Challenges of M2M Massive Access in Wireless Cellular Networks. *Digit. Commun. Netw.* **2015**, *1*, 1–19. [CrossRef]
69. Barreto, A.N.; Faria, B.; Almeida, E.; Rodriguez, I.; Lauridsen, M.; Amorim, R.; Vieira, R. 5G-Wireless Communications for 2020. *J. Commun. Inf. Syst.* **2016**, *31*, 146–163. [CrossRef]
70. Wang, N.; Hossain, E.; Bhargava, V.K. Backhauling 5G Small Cells: A Radio Resource Management Perspective. *IEEE Wirel. Commun.* **2015**, *22*, 41–49. [CrossRef]
71. Chen, L.; Huang, Y.; Xie, F.; Gao, Y.; Chu, H.; Li, Y.; Liang, F.; Yuan, Y. Mobile Relay in LTE-Advanced Systems. *IEEE Commun. Mag.* **2013**, *51*, 144–151. [CrossRef]
72. Svensson, T. Moving Cells/Networks. Available online: http://snow.itn.liu.se/asbtr/SNOWTommySvensson.pdf (accessed on 11 July 2017).
73. Zheng, K.; Leung, V.C.M.; Yang, L.L.; Chatzimisios, P. Guest Editorial Special Issue on 5G Wireless Systems with Massive MIMO. *IEEE Syst. J.* **2017**, *11*. [CrossRef]
74. Ge, X.; Cheng, H.; Guizani, M.; Han, T. 5G Wireless Backhaul Networks: Challenges and Research Advances. *IEEE Netw.* **2014**, *28*, 6–11. [CrossRef]
75. Liu, J.; Kato, N.; Ma, J.; Kadowaki, N. Device-to-Device Communication in LTE-Advanced Networks: A Survey. *IEEE Commun. Surv. Tutor.* **2014**, *17*, 1923–1940. [CrossRef]
76. Pérez-Romero, J.; Sallent, O.; Agusti, R. Enhancing Cellular Coverage through Opportunistic Networks with Learning Mechanisms. In Proceedings of the IEEE Global Telecommunication Conference, Atlanta, GA, USA, 9–13 December 2013.
77. Asadi, A.; Wang, Q.; Mancuso, V. A Survey on Device-to-Device Communication in Cellular Networks. *IEEE Commun. Surv. Tutor.* **2014**, *16*, 1801–1819. [CrossRef]
78. Chen, T.; Kunnari, E.; Tapani, R. Device-to-Device Communication for LTE-Advanced Network System. Available online: https://tel.archives-ouvertes.fr/tel-00983507/document (accessed on 10 July 2017).
79. Hicham, M.; Abghour, N.; Ouzzif, M. Device-to-Device (D2D) Communication under LTE-Advanced Networks. *Int. J. Wirel. Mob. Netw.* **2016**, *8*, 11–22. [CrossRef]
80. Noura, M.; Nordin, R. A Survey on Interface for Device-to-Device (D2D) Communication and its Challenges in 5G Networks. *J. Netw. Comput. Appl.* **2016**. [CrossRef]
81. Mozaffari, M.; Saad, W.; Bennis, M.; Debbah, M. Mobile Unmanned Aerial Vehicles (UAVs) for Energy-Efficient Internet of Things Communications. Unpublished work, 2017.
82. Hayat, S.; Yanmaz, E.; Muzaffar, R. Survey on Unmanned Aerial Vehicle Networks for Civil Applications: A Communications Viewpoint. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 2624–2661. [CrossRef]
83. Larsson, E.G.; Edfors, O.; Tufvesson, F.; Marzetta, T.L. A Massive MIMO for Next Generation Wireless Systems. *IEEE Commun. Mag.* **2016**, *52*, 186–195. [CrossRef]
84. Jovanovic, A.; Li, J.; Richardson, T. Visible light Communication: Opportunities, Challenges and the Path to Market. *IEEE Commun. Mag.* **2013**, *51*, 130–135. [CrossRef]
85. Popovicki, G.; Lauridsen, M.; Soret, B.; Pedersen, K.I.; Mogensen, P. Automation for on-Road Vehicles: Use Cases and Requirements for Radio Design. In Proceedings of the IEEE 82nd Vehicular Technology Conference (VTC Fall), Boston, MA, USA, 6–9 September 2015.
86. Abboud, K.; Omar, H.A.; Zhuang, W. Interworking of DSRC and Cellular Network Technologies for V2X Communications: A Survey. *IEEE Trans. Veh. Technol.* **2016**, *65*, 9457–9470. [CrossRef]
87. Zheng, K.; Zhao, L.; Mei, J.; Dohler, M.; Xiang, W.; Peng, Y. 10 Gb/s HetSNets with Millimeter-Wave Communications: Access and Networking-Challenges and Protocols. *IEEE Commun. Mag.* **2015**, *53*, 222–231. [CrossRef]
88. Jaber, M.; Imran, M.A.; Tafazolli, R.; Tukmanov, A. 5G Backhaul Challendes and Emerging Research Direction: A Survey. *IEEE Access* **2014**, *4*, 1743–1766. [CrossRef]
89. Abdul Salam, S.; Mahmud, S.A.; Khan, G.M.; Raweshidy, H.S. M2M Communication in Smart Grids: Implementation Scenarios and Performance Analysis. In Proceedings of the 2012 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), Paris, France, 1 April 2012.

90. 3GPP. Technical Specification Group Services and System Aspects (Release 8). Available online: http://www.arib.or.jp/IMT-2000/V720Mar09/5_Appendix/Rel8/23/23521-800.pdf (accessed on 10 July 2017).
91. Ratasuk, R.; Tan, J.; Ghosh, A. Coverage and capacity analysis for machine type communications in LTE. In Proceedings of the IEEE 75th Vehicular Technology Conference (VTC Spring), Yokohama, Japan, 6–9 May 2012.
92. Wu, G.; Talwar, S.; Johnsson, K.; Himayat, N.; Johnson, K.D. M2M: From Mobile to Embedded Internet. *IEEE Commun. Mag.* **2011**, *49*, 36–43. [CrossRef]
93. Chen, M.; Wan, J.; Li, F. Machine-To-Machine Communications: Architectures, Standards and Applications. *KSII Trans. Internet Inf. Syst.* **2012**, *6*, 480–497. [CrossRef]
94. Marjanovic, M.; Skorin-Kapov, L.; Pripužić, K.; Antonić, A.; Podnar Žarko, I. Energy-Aware and Quality-Driven Sensor Management For Green Mobile Crowd Sensing. *J. Netw. Comput. Appl. Arch.* **2015**, *59*, 96–108. [CrossRef]
95. 5GPPP. View on 5G Architecture. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-For-public-consultation.pdf (accessed on 10 July 2017).
96. 5GPP. 5G Empowering Vertical Industries. Available online: https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf (accessed on 10 July 2017).
97. 4G Americas. Cellular Technologies Enabling the Internet of Things. Available online: http://www.5gamericas.org/files/6014/4683/4670/4G_Americas_Cellular_Technologies_Enabling_the_IoT_White_Paper_-_November_2015.pdf (accessed on 10 July 2017).
98. Carlesso, M.; Antonopoilos, A.; Granelli, F.; Verikoukis, C. Uplink Scheduling form Smart Metering and Real-Time Traffic Coexistence in LTE Networks. In Proceedings of the International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 820–825.
99. Islam, R.; Avazov, N.; Dobre, O.A.; Kwak, K.S. Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges. *IEEE Commun. Surv. Tutor.* **2016**, *19*, 721–742. [CrossRef]
100. Liu, Y.; Zhang, Y.; Yu, R.; Xie, R. Integrated Energy and Spectrum Harvesting for 5G Wireless Communications. *IEEE Netw.* **2015**, *29*, 75–81. [CrossRef]
101. Kong, L.; Khan, M.K.; Wu, F.; Chen, G.; Zeng, P. Millimeter-Wave Wireless Communications for IoT-Cloud Supported Autonomous Vehicles: Overview, Design, and Challenges. *IEEE Commun. Mag.* **2017**, *55*, 62–68. [CrossRef]
102. Costa-Perez, X.; Garcia-Saavedra, A.; Li, X.; de la Oliva, A.; Iovanna, P.; Deiß, T.; di Giglio, A.; Mourad, A. 5G-Crosshaul: An SDN/NFV Integrated Fronthaul/Backhaul Transport Network Architecture. *IEEE Wirel. Commun.* **2017**, *24*, 38–45. [CrossRef]
103. Trivisonno, R.; Guerzoni, R.; Vaishnavi, I.; Soldani, D. Towards Zero Latency Software Defined 5G Networks. In Proceedings of the IEEE International Conference on Communication Workshop (ICCW), London, UK, 8–12 June 2015.
104. Esaki, H.; Nakamura, R. Overlaying and Slicing for Iot Era Based on Internet's End-To-End Discipline. In Proceedings of the IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN), Osaka, Japan, 12–14 June 2017.
105. 5GPP. 5G and Media & Entertainment. Available online: https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-Media-Entertainment-Vertical-Sector.pdf (accessed on 10 July 2017).
106. GSA. The Road to 5G: Frivers, Applications, Requirements and Technical Development. Available online: http://www.huawei.com/minisite/5g/img/GSA_the_Road_to_5G.pdf (accessed on 10 July 2017).
107. Erman, B.; Yiu, S. Modeling 5G Wireless Network Service Reliability Prediction with Bayesian Network. In Proceedings of the 2016 IEEE International Workshop Technical Committee on Communications Quality and Reliability, Stevenson, WA, USA, 10–12 May 2016.
108. NOKIA. 5G Use Cases and Requirements White Paper. Available online: https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2015/NGMN_5G_White_Paper_V1_0.pdf (accessed on 10 July 2017).
109. Motlagh, N.H.; Bagaa, M.; Taleb, T.; Song, J. Connection Steering Mechanism between Mobile Networks for Reliable UAV's Iot Platform. In Proceedings of the IEEE International Conference on Communications (ICC), Paris, France, 21–25 July 2017.
110. Kaltenberger, F.; Knopp, R.; Vitiello, C.; Danneberg, M.; Festag, A. Experimental Anaysis of 5G Candidate Waveform and Their Coexistence with 4G Systems. In Proceedings of the Joint NEWCOM/COST Workshop on Wireless Communications JNCW, Barcelona, Spain, 14–15 October 2016.

111. Serra, J.; Pubill, D.; Antonopoulos, A.; Verikoukis, C. Smart HVAC Control in IoT: Energy Consumption Minimization with User Comfort Constraints. *Sci. World J.* **2014**, *2014*. [CrossRef] [PubMed]
112. Nikolich, P.; Lin, C.; Korhonen, J.; Marks, R.; Tye, B.; Li, G.; Ni, J.; Zhang, S. Standards for 5G and Beyond: Their Use Cases and Applications. *IEEE 5G Tech Focus* **2017**, *1*. Available online: https://5g.ieee.org/tech-focus/june-2017/standards-for-5g-and-beyond (accessed on 9 September 2017).
113. Kusume, K.; Fallgren, M.; Queseth, O.; Braun, V.; Gozalvez-Serrano, D.; Korthals, I.; Zimmermann, G.; Schubert, M.; Hossain, M.I.; Widaa, A.A.; et al. Updated Scenarios, Requirements and KPIs for 5G Mobile and Wireless System with Recommendations for Future Investigations. METIS_D1.5_v1. METIS. Available online: https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D1.5_v1.pdf (accessed on 10 July 2017).
114. 5GPP. 5G Vision and Requirements. Available online: https://5g-ppp.eu/imt-2020-5g-promotion-group-and-5g-ppp-announce-memorandum-of-understanding-for-5g/ (accessed on 10 July 2017).
115. Yang, C.; Li, J. *Interference Mitigation and Energy Management in 5G Heterogeneous Cellular Networks*; ICI Global: Harshey, PA, USA, 2017; pp. 1–362.
116. Ericsson. 5G Energy Performance. Available online: http://www.5gamericas.org/files/3114/3898/7332/Ericsson_wp-5g-energy-performance.pdf (accessed on 10 July 2017).
117. Shen, K.; Yu, W. Distributed Pricing-Based User Association for Downlink Hterogeneous Cellular Networks. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1100–1113. [CrossRef]
118. Nguyen, H.V.; Le, L.B. Distributed Base Station Association and Power Control for Heterogeneous Cellular Networks. *IEEE Trans. Veh. Technol.* **2014**, *63*, 282–296. [CrossRef]
119. Singh, S.; Andrews, J.G. Joint Resource Portitioning and Offloading in Heterogeneous Cellular Networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 888–901. [CrossRef]
120. Elsawy, H.; Hossain, E.; Kim, D.I. HetNets with Cognitive Small Cells: User Offloading and Distributed Channel Access Technique. *IEEE Commun. Mag.* **2013**, *51*, 28–36. [CrossRef]
121. Ji, Z.; Li, K. Cognitive Radios for Dynamic Spectrum Access-Dynamic Spectrum Sharing: A Game Theoretical Overview. *IEEE Commun. Mag.* **2007**, *45*, 88–94. [CrossRef]
122. Aijazand, A.; Aghvami, A.H. PRMA-Based Cognitive Machine-to-Machine Communications in Smart Grid Networks. *IEEE Trans. Veh. Technol.* **2015**, *64*, 3608–3623. [CrossRef]
123. Okasaka, S.; Weiler, R.J.; Keusgen, W.; Pudeyev, A.; Maltsev, A.; Karls, I.; Sakaguchi, K. Proof-of-Concept of a Millimeter-Wave Integrated Heterogeneous Network for 5G Cellular. *Sensors* **2016**, *9*. [CrossRef] [PubMed]
124. Qi, S.; Tesanovic, M. FQAM-FBMC Design and Its Application to Machine Type Communication. In Proceedings of the IEEE 27th Annual International Symposium on PIMRC, Valencia, Spain, 4–8 September 2016.
125. Domouchtsidis, S.G.; Ntouni, G.D.; Kapinas, V.M.; Karagiannidis, G.K. OFDM-IM vs. FQAM: A Comparative Analysis. In Proceedings of the International Conference on Telecommunications, Thessaloniki, Greece, 16–18 May 2016.
126. Hong, S.; Sagong, M.; Lim, C.V.; Cheun, K.; Cho, S. FQAM: A Modulation Scheme for Beyond 4G Cellular Wireless Communication Systems. In Proceedings of the IEEE GC Wkshps, Atlanta, GA, USA, 9–13 December 2013.
127. Perlman, S.; Forenza, A. pCell Wireless Reinvented. Arthemis Networks. Available online: http://www.rearden.com/DIDO/DIDO_White_Paper_110727.pdf (accessed on 10 July 2017).
128. Mesleh, R.Y.; Hass, H.; Sinanovic, S.; Ahn, C.W.; Yun, S. Spatial Modulation. *IEEE Trans. Veh. Technol.* **2008**, *57*, 2228–2241. [CrossRef]
129. Maunder, R.G. The 5G Channel Code Contenders. Available online: https://www.accelercomm.com/sites/accelercomm.com/files/WhitePaper09Aug2016.pdf (accessed on 10 July 2017).
130. Scholl, S.; Weithoffer, S.; When, N. Advanced Iterative Channel Coding Schemes: When Shannon meets Moore. In Proceedings of the 9th International Symposium on Turbo Codes & Iterative Information, Brest, France, 5–9 September 2016.
131. Korpi, D.; Tamminen, J.; Tarunen, M.; Huusari, T.; Choi, Y.-S.; Anttila, L.; Talwar, S.; Valkama, M. Full-Duplex Mobile Device: Pushing the Limit's. *IEEE Commun. Mag.* **2016**, *54*, 80–87. [CrossRef]
132. Hong, S.S.; Brand, J.; Choi, J.I.; Jain, M.; Mehl, J. Applications of Self-Interference Cancellation in 5G and Beyond. *IEEE Commun. Mag.* **2014**, *52*, 114–121. [CrossRef]

133. Duarte, M.; Dick, C.; Sabharwal, A. Experiment-Driven Characterization of Full-Duplex Wireless Systems. *IEEE Trans. Wirel. Commun.* **2012**, *11*, 4296–4307. [CrossRef]
134. Bharadia, D.; McMilin, E.; Katti, S. Full Duplex Radios. In Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM, Hong Kong, China, 12–16 August 2013.
135. Čelik, H.; Sung, K.W. On the Feasibility of Blind Dynamic TDD in Ultra-Dense Wireless Networks. In Proceedings of the 2015 IEEE 81st VehicularTechnology Conference (VTC Spring), Glasgow, UK, 11–14 May 2015.
136. Osti, P.; Aalto, S.; Lassila, P. Flow-Level Modeling and Optimization of Intercell Coordination with Dynamic TDD. In Proceedings of the 10th ACM symposium on QoS and security for wireless and mobile networks Q2SWinet '14, Montreal, QC, Canada, 21–26 September 2014.
137. Kerttula, J.; Marttinen, A.; Ruttik, K.; Jäntti, R.; Alam, N. Dynamic TDD on LTE Small Cells. *EURASIP J. Wirel. Commun. Netw.* **2016**, *194*. [CrossRef]
138. Gupta, A.K.; Kulkarni, M.N.; Visotsky, E.; Vook, F.W.; Ghosh, A.; Andrews, J.G.; Heath, R.W. Rate Analysis and Feasibility of Dynamic TDD in 5G Cellular Systems. In Proceedings of the 2016 IEEE International Conference on Communications, Cuala Lumpur, Malaysia, 22–27 May 2016.
139. Sun, H.; Wildemeersch, M.; Sheng, M.; Quek, T.K. D2D Enchanced Heterogeneus Cellular Networks with Dynamic TDD. *IEEE Trans. Wirel. Commun.* **2015**, *14*, 4204–4218. [CrossRef]
140. Sun, H.; Wildemeersch, M.; Sheng, M.; Quek, T.K.; Li, J. Traffic Adaptation and Energy Efficiency for Small Cell Networks with Dynamic TDD. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 3234–3251. [CrossRef]
141. Čelik, H.; Song, K.W. Joint Transmission with Dummy Symbols for Dynamic TDD in Ultra-Dense Deployment. Available online: https://aps.arxiv.org/abs/1702.06045 (accessed on 10 July 2017).
142. Li, J.; Farahvash, S.; Kavehrad, M.; Valenzuel, R. Dynamic TDD and Fixed Cellular Networks. *IEEE Commun. Lett.* **2002**, *4*, 218–220. [CrossRef]
143. Shen, Z.; Khoryae, A.; Eriksson, E.; Pan, X. Dynamic Uplink-Downlink Configuration and Interference Management in TD-LTE. *IEEE Commun. Mag.* **2012**, *50*, 51–59. [CrossRef]
144. Chunlin, Y.; Zhifeng, Y.; Weimin, L.; Yifei, Y. Non-Orthogonal Multiple Access Schemes for 5G. *ZTE Commun.* **2016**, *14*, 11–16. [CrossRef]
145. 3GPP. Discussion on Multiple Access for Radio Interface. Available online: http://www.3gpp.org/DynaReport/TDocExMtg--R1-84b--31661.htm (accessed on 10 July 2017).
146. Yuan, Z.; Yu, G.; Li, W.; Yuan, Y.; Wang, X.; Xu, J. Multi-user Shared Access for Internet of Things. In Proceedings of the 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), Nanjing, China, 15–18 May 2016.
147. 3GPP. Receiver Implementation for MUSA. 3GPP. Available online: http://www.3gpp.org/DynaReport/TDocExMtg--R1-85--31662.htm (accessed on 10 July 2017).
148. 3GPP. Contetion-Based Non-Orthogonal Multiple Access for UL mMTC. Available online: http://www.3gpp.org/DynaReport/TDocExMtg--R1-85--31662.htm (accessed on 10 July 2017).
149. 3GPP. Resource Spread Multiple Access. Available online: http://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_85/Docs/ (accessed on 10 July 2017).
150. Taherzadeh, M.; Nikopour, H.; Bayesteh, A.; Baligh, H. SCMA Codebook Design. In Proceedings of the IEEE International Symposium on Personal, Indoor And Mobile Radio Communications, London, UK, 14–17 September 2014.
151. Nikopour, H.; Baligh, H. Sparse Code Multiple Acccess. In Proceedings of the 2014 IEEE 80th Vehicular Technology Conference, Vancouver, BC, Canada, 14–17 September 2014.
152. Future Mobile Communication Forum. 5G Whitepaper v2.0, Part d-Alternative Multiple Access v1. Available online: http://www.future-forum.org/dl/151106/whitepaper.rar (accessed on 10 July 2017).
153. 3GPP. Candidate Solution for New Multiple Access. 3GPP. Available online: http://portal.3gpp.org/ngppapp/CreateTdoc.aspx?mode=view&contributionId=697587 (accessed on 10 July 2017).
154. Dai, X.; Chen, S.; Sun, S.; Kang, S.; Wang, Y.; Shen, Z.; Xu, Z. Successive Interference Cancelation Amenable Multiple Access (SAMA) for Future Wireless Communications. In Proceedings of the IEEE International Conference on Communication Systems, Macau, China, 19–21 November 2014.
155. Dai, X. Successive Interference Cancellation Amenable Space-Time Codes with Good Multiplexing-Diversity Tradeoff. *Wirel. Pers. Commun.* **2010**, *55*, 645–654. [CrossRef]

156. Li, P.; Liu, L.; Wu, K.; Leung, W.K. On Interleave-Division Multiple-Access. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017.
157. Li, P.; Liu, L.; Wu, K.; Leung, W.K. Interleave Division Multiple-Access. *IEEE Trans. Wirel. Commun.* **2006**, *5*, 938–947. [CrossRef]
158. Rabee, F.A.; Dayaslioglu, K.; Gitlin, R. The Optimum Received Power Levels of Uplink Non-Orthogonal Multiple Access (NOMA) Signals. In Proceedings of the 2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON), Cocoa Beach, FL, USA, 24–25 April 2017.
159. Saito, Y.; Kishiyama, Y.; Benjebbour, A.; Nakamura, T.; Li, A.; Higuchi, K. Non-orthogonal Multiple Access (NOMA) for Cellular Future Radio Access. In Proceedings of the 77th IEEE Vehicular Technology Conference, Dresden, Germany, 2–5 June 2013.
160. Choi, J. On Multiple Access Using H-ARQ with SIC Techniques for Wireless Ad Hoc Networks. *Wirel. Pers. Commun.* **2013**, *69*, 187–212. [CrossRef]
161. Choi, J. Non-orthogonal Multiple Access in Downlink Coordinated Two-point Systems. *IEEE Commun. Lett.* **2014**, *18*, 313–316. [CrossRef]
162. Sedtheetorn, P.; Chulajata, T. Accurate Spectral Efficiency Analysis for Non Orthogonal Multiple Access. *Trans. Adv. Commun. Technol.* **2016**, *5*, 844–850.
163. Wild, T.; Schaich, F.; Chen, Y. 5G Air Interface Design Based on Universal Filtered (UF-) OFDM. In Proceedings of the 19th International Conference in DSP, Hong Kong, China, 20–23 August 2014.
164. Ding, Z.; Yang, Z.; Fan, P.; Poor, H.V. On the Performance of Non-orthogonal Multiple Access in 5G Systems with Randomly Deployed Users. *IEEE Signal Process. Lett.* **2014**, *21*, 1501–1505. [CrossRef]
165. Zhao, Z.; Schellmann, M.; Wang, Q.; Gong, X.; Boehnke, R.; Xu, W. Pulse shaped OFDM for Asynchronous Uplink Access. In Proceedings of the Asilomar Conference Signals, Systems and Computers, Pacific Grove, CA, USA, 8–11 November 2015.
166. Cover, T.M.; El Gamel, A. Capacity Theorems for The Relay Channel. *IEEE Trans. Inf. Theory* **1979**, *25*, 572–584. [CrossRef]
167. Cover, T.M.; Leung, C.S.K. An Achievable Rate Region for the Multiple-access Channel with Feedback. *IEEE Trans. Inf. Theory* **1981**, *27*, 292–298. [CrossRef]
168. Ahlswede, R. Multiway Communication Channels. In Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor, Armenian SSR, 2–8 September 1971.
169. Xie, L.L.; Kumar, P.R. An Achievable Rate for the Multiple-level Relay Channel. *IEEE Trans. Inf. Theory* **2005**, *51*, 1348–1358. [CrossRef]
170. Kim, K.T.; Ahn, S.-K.; Kim, Y.-H.; Park, H.; Wang, L.; Chen, C.-Y.; Park, J. Adaptive Sliding-Window Coded Modulation in Cellular Networks. In Proceedings of the IEEE GLOBECOM, San Diego, CA, USA, 6–10 December 2015.
171. Thurfjell, M.; Ericsson, M.; de Bruin, P. Network Densification Impact on System Capacity. In Proceedings of the 2015 IEEE 81st Vehicular Technology Conference, Glasgow, Scotland, 11–14 May 2015.
172. Darwish, A.; Ibrahim, A.S. Capacity Improvements via Indoor Small Cells. In Proceedings of the 2014 International Wireless Communications and Mobile Computing Conference (IWCMC), Nicosia, Cyprus, 4–8 August 2014.
173. Hoadley, J.; Maveddat, P. Enabling Small Cell Deployment with Hetnet. *IEEE Wirel. Commun.* **2012**, *19*, 4–5. [CrossRef]
174. Hughes, M.; Jovanovic, V.M. Small Cells-Effective Capacity Relief Option for Heterogeneous Networks. In Proceedings of the 2012 IEEE Vehicular Technology Conference, Quebec City, QC, Canada, 3–6 September 2012.
175. Sui, Y.; Guvenc, I.; Svensson, T. Interference Management for Moving Networks in Ultra-Dense Urban Scenarios. *EURASIP J. Wirel. Commun. Netw.* **2015**, *111*. [CrossRef]
176. Sui, Y.; Vihrälä, J.; Papadogiannis, A.; Sternad, M.; Svensson, T. Moving Cells: A Promising Solution to Boost Performance for Vehicular Users. *IEEE Commun. Mag.* **2013**, *51*, 62–68. [CrossRef]
177. Sambo, Y.A.; Shakir, M.Z.; Héliot, F.; Imran, M.A.; Mumtaz, S.; Qaraqe, K.A. Device-to-Device Communication in Heterogeneous Networks. In *Smart Device to Smart Device Communication*; Mumtaz, S., Rodriguez, J., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 219–235.
178. Liu, X.; Andrews, J.G.; Ghosh, A. Spectrum Sharing for Device- to-Device Communication in Cellular Networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 6272–6740. [CrossRef]

179. Zafar, B.; Gherekhloo, S.; Haardt, M. Analysis of Multihop Relaying Networks: Communication Between Range-Limited and Cooperative Nodes. *IEEE Veh. Technol. Mag.* **2012**, *7*, 40–47. [CrossRef]
180. Kamalinejad, P.; Mahapatra, C.; Sheng, Z.; Mirabbasi, S.; Leung, V.C.M.; Guan, Y.L. Wireless Energy Harvesting for the Internet of Things. *IEEE Commun. Mag.* **2015**, *53*, 102–108. [CrossRef]
181. Mekikis, P.-V.; Antonopoulos, A.; Kartsakli, E.; Lalos, A.S.; Alonso, L.; Verikoukis, C. Information Exchange in Randomly Deploys Dense WSNs with Wireless Energy Harvesting Capabilities. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 3008–3018. [CrossRef]
182. Mekikis, P.-V.; Antonopoulos, A.; Alonso, L.; Verikoukis, C. Stochastic Modeling of Wireless Charged Wearables for Reliable Health Monitoring In Hospital Environments. In Proceedings of the International Conference on Communications (ICC), Paris, France, 21–25 May 2017.
183. Li, Y.; Cai, L. UAV-Assisted Dynamic Coverage in a Heterogeneous Cellular System. *IEEE Netw.* **2017**, *31*, 56–61. [CrossRef]
184. Messous, M.-A.; Senouci, S.-M.; Sedjelmaci, H. Network Connectivity and Area Coverage for UAV Fleet Mobility Model with Energy Constraint. In Proceedings of the Wireless Communications and Networking Conference (WCNC), Doha, Qatar, 3–6 April 2016.
185. Mekikis, P.-V.; Kartsakli, E.; Alonso, L.; Verikoukis, C. Flexible Aerial Relay Nodes for Communication Recovery and D2D Relaying. In Proceedings of the IEEE 5th Global Conference on Consumer Electronics, Kyoto, Japan, 11–14 October 2016.
186. Zeng, Y.; Zhang, R.; Lim, T.J. Wireless Communications with Unmanned Aerial Vehicles: Opportunities and Challenges. *IEEE Commun. Mag.* **2016**, *54*, 36–42. [CrossRef]
187. Yang, X.; Lu, W.-J.; Wang, N.; Nieman, K.; Jin, S.; Zhu, H.; Mu, X.; Wong, I.; Huang, Y.; You, X. Design and Implementation of a TDD-Based 128-Antenna Massive MIMO Prototyping System. Available online: http://adsabs.harvard.edu/cgi-bin/bib_query?arXiv:1608.07362 (accessed on 11 July 2017).
188. Chen, B.; Zhu, C.; Shu, L.; Su, M.; Wei, J.; Leung, V.C.M.; Rodriques, J.J.P.C. Uplink Transmission for Lightweight Single-antenna UEs in the Presence of a Massive MIMO eavesdropper. *IEEE Access* **2016**, *99*, 2169–3536. [CrossRef]
189. Wu, S.; Wang, H.; Youn, C.H. Visible Light Communications for 5G Wireless Networking Systems: From Fixed to Mobile. *IEEE Access* **2014**, *28*, 41–45. [CrossRef]
190. Feng, L.; Hu, R.Q.; Wang, J.; Xu, P.; Qian, Y. Applying VLC in 5G Networks: Architectures and Key Technologies. *IEEE Access* **2016**, *30*, 77–83. [CrossRef]
191. Grobe, L.; Paraskevopoulos, A.; Hilt, J.; Schulz, D.; Lassak, F.; Hartlieb, F.; Kottke, C.; Jungnickel, V.; Langer, K.-D. High-Speed Visible Light Communication Systems. *IEEE Commun. Mag.* **2013**, *51*, 60–66. [CrossRef]
192. Li, X.; Zhang, R.; Wang, J.; Hanzo, L. Cell-Centric and User-Centric Multi-User Scheduling in Visible Light Communication Aided Networks. In Proceedings of the IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015.
193. Katriniok, A.; Kleibaum, P.; Res, C.; Eckstein, L. Automation of Road Vehicles Using V2X: An Application to Intersection Automation. Available online: http://papers.sae.org/2017-01-0078/ (accessed on 10 July 2017).
194. Ghods, A.; Severi, S.; Abreu, G. Localization in V2X Communication Networks. In Proceedings of the IEEE Intelligent Vehicles Symposium, Gothenburg, Sweden, 19–22 June 2016.
195. Goebel, N.; Bialon, R.; Mauve, M.; Graffi, K. Coupled Simulation of Mobile Cellular Networks, Road Traffic and V2X Applications Using Traces. In Proceedings of the International Conference on Communications, Kuala Lumpur, Malaysia, 22–27 May 2016.
196. Xiao, Z.; Liu, H.; Havyarimana, V.; Li, T.; Wang, D. Analytical Study on Multi-Tier 5G Heterogeneous Small Cell Networks: Coverage Performance and Energy Efficiency. *Sensors* **2016**, *11*. [CrossRef] [PubMed]
197. Choi, J.I.; Jain, M.; Srinivasan, K.; Levis, P.; Katti, S. Achieving Single Channel, Full Duplex Wireless Communication. In Proceedings of the sixteenth annual international conference on Mobile computing and networking, Chicago, IL, USA, 20–24 September 2010.
198. Song, Q.; Nuaymi, L.; Lagrange, X. Survey of Radio Resource Management Issues and Proposals for Energy-Efficient Cellular Network That Will Cover Billions of Machines. *Wirel. Commun. Netw.* **2016**, *140*. [CrossRef]

199. Costa, G.; Miao, G. Context-Aware Machine-To-Machine Communications. In Proceedings of the 33rd Annual IEEE International Conference on Computer Communications (INFOCOM WKSHPS), Toronto, ON, Canada, 27 April–2 May 2014.
200. Agyapong, P.K.; Iwamura, M.; Staehle, D.; Kiess, W.; Benjebbour, A. Design Considerations for a 5G Network Architecture. *IEEE Commun. Mag.* **2014**, *52*, 65–75. [CrossRef]
201. Vermesan, O.; Friass, P.; Guillemin, P.; Giaffreda, R.; Grindvoll, H.; Eisenhauer, M.; Serrano, M.; Moessner, K.; Spirito, M.; Blystad, L.-C.; et al. *Building the Hypercconnected Society—IoT Research and Innovation Value Chains, Ecosystems and Markets*; Vermesan, O., Friass, P., Eds.; River Publishers: Aalborg, Denmark, 2015; Volume 43.
202. Panwar, N.; Sharma, S.; Singh, A.K. A Survey on 5G: The Next Generation of Mobile Communication. *Elsevier Phys. Commun.* **2016**, *18*, 64–84. [CrossRef]
203. Thuemmler, C.; Gavrasm, A.; Jumelle, A.K.L.; Paulin, A.; Sadique, A.; Schneider, A.; Fedell, C.; Abraham, D.; Trossen, D. 5G and e-Health. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-eHealth-Vertical-Sector.pdf (accessed on 10 July 2017).
204. Rose, K.; Eldridge, S.; Chaplin, L. The Internet of Things: An Overview. Available online: https://www.internetsociety.org/sites/default/files/ISOC-IoT-Overview-20151014_0.pdf (accessed on 10 July 2017).
205. Huawei. 5G Opening up New Business Opportunities. Available online: http://www-file.huawei.com/-/media/CORPORATE/PDF/x-lab/10-5g-opening-up-new-business-opportunities-en.pdf?la=en (accessed on 10 July 2017).
206. Lin, H.; Shao, J.; Zhang, C.; Fang, Y. CAM: Cloud-Assisted Privacy Preserving Mobile Health Monitoring. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 985–997. [CrossRef]
207. NGMN Alliance. NGMN 5G White Paper. Available online: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf (accessed on 10 July 2017).
208. Le, N.T.; Hossain, M.A.; Islam, A.; Kim, D.-Y.; Choi, Y.-J.; Jang, Y.M. Survey of Promising Technologies for 5G Networks. *Hindawi* **2016**, *2016*. [CrossRef]
209. Militano, L.; Araniti, G.; Condoluci, M.; Farris, I.; Iera, A. Device-to-Device Communications for 5G Internet of Things. *EAI Endorsed Trans. Internet Things* **2015**, *1*. [CrossRef]
210. Shariatmadari, H.; Ratasuk, R.; Iraji, S.; Laya, A.; Taleb, T.; Ghosh, A. Machine-type Communications: Current Status and Future Perspective toward 5G Systems. *IEEE Commun. Mag.* **2015**, *53*, 10–17. [CrossRef]
211. 5GPPP. 5G and the Factories of the Future. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Factories-of-the-Future-Vertical-Sector.pdf (accessed on 10 July 2017).
212. Elayoubi, S.E.; Fallgren, M.; Spapis, P.; Zimmermann, G.; Martín-Sacristán, D.; Yang, C.; Jeux, S.; Agyapong, P.; Campoy, L.; Qi, Y.; et al. 5G Service Requirements and Operational Use Cases: Analysis and METIS II Vision. In Proceedings of the 2016 European Conference on Networks and Communications (EuCNC), Athens, Greece, 27–30 June 2016.
213. Blanco, B.; Fajardo, H.O.; Giannis, I.; Kafetzakis, E.; Pneg, S.; Perez-Romero, J.; Trajkovska, I.; Khodashenas, P.S.; Goratti, L.; Paolino, M.; et al. Technology Pillars in the Architecture of Future 5G Mobile Networks: NFV, MEC and SDN. *Comput. Stand. Interfaces* **2017**, *54*, 216–228. [CrossRef]
214. ITU-R. Recommendation ITU-R M.2083-0. IMT Vision: Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf (accessed on 10 July 2017).
215. Le, L.B.; Lau, V.; Jorswieck, E.; Dao, N.-D.; Haghighat, A.; Kim, D.I.; Le-Ngoc, T. Enabling 5G Mobile Wireless Technologies. *EURASIP J. Wirel. Commun. Netw.* **2015**, *218*. [CrossRef]
216. Liu, G.; Jiang, D. 5G: Vision and Requirements for Mobile Communication System towards Year 2020. *Chin. J. Eng.* **2016**, *2016*, 5974586. [CrossRef]
217. Bogale, T.E.; Le, L.B. Massive MIMO and mmWave for 5G Wireless HetNet: Potential Benefit's and Challenges. *IEEE Veh. Technol. Mag.* **2016**, *11*, 64–75. [CrossRef]
218. Banelli, P.; Buzzi, S.; Colavolpe, G.; Modenini, A.; Rusek, F.; Ugolini, A. Modulation Formats and Waveforms for 5G Networks: Who Will Be the Heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency. *IEEE Signal Process. Mag.* **2014**, *31*, 80–93. [CrossRef]
219. Palattella, M.R.; Dohler, M.; Grieco, A.; Rizzo, G.; Torsner, J.; Engel, T.; Ladidm, L. Internet of Things in the 5G Era: Enablers, Architecture and Business Models. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 510–527. [CrossRef]

220. Pries, R.; Morper, H.-J.; Galambosi, N.; Jarschel, M. Network as a Service—A Demo on 5G Network Slicing. In Proceedings of the 28th International Teletraffic Congress (ITC 28), Würzburg, Germany, 12–16 September 2016.
221. Caraguay, A.L.V.; Peral, A.B.; Lopez, L.I.B. SDN: Evolution and Opportunities in the Development IoT Applications. *Int. J. Distrib. Sens. Netw.* **2014**, *10*. [CrossRef]
222. Bizanis, N.; Kuipers, F.A. SDN and Virtualization Solutions for the Internet of Things: A Survey. *IEEE Access* **2016**, *4*, 5591–5606. [CrossRef]
223. Stancu, A.; Vulpe, A.; Fratu, O.; Halunga, S. Default Values Mediator Used for a Wireless Transport SDN Proof of Concept. In Proceedings of the IEEE Conference on Standards for Communications and Networking (CSCN), Berlin, Germany, 31 October–2 November 2016.
224. Warren, D.; Dewar, C. Understanding 5G: Perspectives on Future Technological Advancements in Mobile. Available online: https://www.gsmaintelligence.com/research/?file=141208-5g.pdf&download (accessed on 10 July 2017).
225. Rappaport, T.S.; Heath, R.V.; Daniels, R.C.; Murdock, J.N. *Millimeter Wave Wireless Communications*, 1st ed.; Pearson: Harshey, PA, USA, 2017; pp. 1–657.
226. Thakare, S. Research Challenges of Millimeter Wave Technology in 5G Cellular System. In Proceedings of the 4th International Conference on Recent Trends in Engineering & Technology (ICRTET), Nashik, India, 2–4 July 2015.
227. Ye, J.; He, J.; Ge, X.; Chen, M. Energy Efficiency Analysis of 5G Ultra-dense Networks Based on Random Way Point Mobility Models. In Proceedings of the 2016 19th International Symposium on Wireless Personal Multimedia Communications, Shenzhen, China, 14–16 November 2016.
228. Witrisal, K.; Meissner, P.; Leitinger, E.; Shen, Y.; Gustafson, C.; Tufvesson, F.; Haneda, K.; Dardari, D.; Molisch, A.F.; Conti, A.; et al. High-Accurany Localization for Assisted Living: 5G Systems Will Turn Multipath Channels from Foe to Friend. *IEEE Signal Process. Mag.* **2016**, *33*, 59–70. [CrossRef]
229. 5GPP. 5G Automotive Vision. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf (accessed on 10 July 2017).
230. Siaud, I.; Umner-Moll, A.M. Green-Oriented Multi-Techno Link Adaptation Metrics for 5G Heterogeneous Networks. *EURASIP J. Wirel. Commun. Netw.* **2016**, *92*. [CrossRef]
231. Le, L.B.; Niyato, D.; Hossain, E.; Kim, D.I.; Hoang, D.T. QoS-Aware and Energy-Efficient Resource Management in OFDMA Femtocells. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 180–194. [CrossRef]
232. Ye, Q.; Rong, B.; Chen, Y.; Al-Shalash, M.; Caramanis, C.; Andrews, J.G. User Association for Load Balancing in Heterogeneous Cellular Networks. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 2706–2716. [CrossRef]
233. Mavromoustakis, C.X.; Bourdena, A.; Mastorakis, G.; Pallis, E.; Kormentzas, G. Coverage An Energy-Aware Scheme for Efficient Spectrum Utilization in A 5G Mobile Cognitive Radio Network Architecture. *Telecommun. Syst.* **2015**, *59*, 63–75. [CrossRef]
234. Wu, S.; Wang, Y.; Al Imari, M.; Nekovee, M. FQAM: Frequency and Quadrature Amplitude Modulation for 5G Networks. In Proceedings of the 2016 European Conference on Networks and Communications (EuCNC), Athens, Greece, 27–30 June 2016.
235. Al-Imari, M.; Mouhouche, B.; Al.Imari, M.; Nekovee, M.F. Non-Orthogonal FQAM for Multiple Access in the Uplink of 5G Wireless Networks. In Proceedings of the 2016 International Symposium on Wireless Communication Systems (ISWCS), Poznań, Poland, 20–23 September 2016.
236. Hong, S.; Sagong, M.; Lim, C.; Cho, S.; Cheun, K.; Yang, K. Frequency and Quadrature Amplitude Modulation for Downlink Cellular OFDMA Networks. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1256–1267. [CrossRef]
237. Nekovee, M.; Yue, W.; Tesanovic, M.; Wu, S.; Qi, Y.; Al-Imari, M. Overview of 5G Modulation and Waveforms Candidates. *J. Commun. Inf. Netw.* **2016**, *1*, 44–60. [CrossRef]
238. Hochwald, B.M.; Sweldens, W. Differential Unitary Space-Time Modulation. *IEEE Trans. Commun.* **2000**, *48*, 2041–2052. [CrossRef]
239. Ostman, J.; Yang, W.; DUrisim, G.; Koch, T. Diversity Versus Multiplexing at Finite Blocklength. In Proceedings of the 11th International Symposium on Wireless Communications Systems (ISWCS), Barcelona, Spain, 26–29 August 2014.
240. Devaillie, B.; van Liempd, B.; Hershberg, B.; Craninckx, J.; Rikkinen, K.; van den Broek, D.J.; Klumperink, E.A.M.; Nauta, B. In-Band Full-Duplex Transceiver Technology for 5G Mobile Networks. In Proceedings of the ESSCIRC Conference, Graz, Austria, 4–8 December 2016.

241. Kim, C.; Kim, K.; Yun, Y.H.; Ho, Z.; Lee, B.; Seol, J.-Y. QAM-FBMC: A New Multi-carrier System for Post-OFDM Wireless Communications. In Proceedings of the IEEE Global Communication Conference GLOBECOM, San Diego, CA, USA, 6–10 December 2015.
242. Fettweis, G.; Krondorf, M.; Bittner, S. GFDM-generalized Frequency Division Multiplexing. In Proceedings of the IEEE 69th VTC, Barcelona, Spain, 26–29 April 2009.
243. Abdoli, G.; Jia, M.; Ma, J. Filtered OFDM: A New Waveform for Future Wireless Systems. In Proceedings of the IEEE 16th International Workshop SPAWC, Stockholm, Sweden, 28 June–1 July 2015.
244. Li, J.; Kearney, K.; Bala, E.; Yang, R. Filtered OFDM: A Resource Block Based Filtered OFDM Scheme and Performanse Comparison. In Proceedings of the 2013 20th International Conference on Telecommunications, Casablanca, Morocco, 6–8 May 2013.
245. Kumar, U.; Ibars, C.; Bhorkar, A.; Jung, H. Filtered OFDM: A Waveform for 5G: Guard Interval DFT-S-OFDM. In Proceedings of the IEEE Globecom Workshops (GC Wkshps), San Diego, CA, USA, 6–10 December 2015.
246. Chung, C.D. Spectrally Precoded OFDM. *IEEE Trans. Commun.* **2006**, *54*, 2173–2185. [CrossRef]
247. Monk, A.; Hadani, R.; Tsatsanis, M.; Rakib, S. OTFS—Orthogonal Time Frequency Space. Available online: https://arxiv.org/ftp/arxiv/papers/1608/1608.02993.pdf (accessed on 10 July 2017).
248. Bandemer, B.; Gamal, A.E.; Kim, Y.H. Optimal Achievable Rates for Interference Networks with Random Codes. *IEEE Trans. Inf. Theory* **20125**, *61*, 6536–6549. [CrossRef]
249. Carleial, A.B. Multiple-access Channels with Different Generalized Feedback Signals. *IEEE Trans. Inf. Theory* **1982**, *28*, 841–850. [CrossRef]
250. Prasad, A.; Uusitalo, M.A.; Li, Z.; Lunder, P. Reflection Environment Maps for Enhanced Reliability in 5G Self-Organizing Networks. In Proceedings of the IEEE Globecom Workshops (GC Wkshps), Washington, DC, USA, 4–8 December 2016.
251. Sui, Y.; Papadogiannis, A.; Sternad, M. The Potential of Moving Relays-A Performance Analysis. In Proceedings of the 2012 IEEE 75th Vehicular Technology Conference (VTC Spring), Yokohama, Japan, 6–9 May 2012.
252. Jiang, L.; Tian, H.; Xing, K.; Zhang, K.; Maharjan, S.; Cjessing, S.; Zhang, Y. Social-aware Energy Harvesting Device-to-Device Communications in 5G Networks. *IEEE Wirel. Commun.* **2016**, *23*, 20–27. [CrossRef]
253. Huang, X.; Yu, R.; Kang, J.; Gao, Y.; Maharjan, S.; Gjessing, S.; Zhang, Y. Software Defined Energy Harvesting Networking for 5G Green Communications. *IEEE Wirel. Commun.* **2017**, *24*, 38–45. [CrossRef]
254. Merwaday, A.; Guvenc, I. UAV Assisted Heterogeneous Networks for Public Safety Communications. In Proceedings of the IEEE Wireless Communications and Networking Conference Workshops (WCNCW), New Orleand, LA, USA, 9–12 March 2015.
255. Osseiran, A. The 5G Radio-Access Technologies. In *5G Mobile and Wireless Communications Technology*, 1st ed.; Osseiran, A., Monserrat, J.F., Marsch, P., Dohler, M., Nakamura, T., Eds.; Cambridge University Press: Cambridge, UK, 2016.
256. Sahin, A.; Yang, R.; Bala, E.; Beluri, M.C.; Olesen, R.L. Flexible DFT-S-OFD: Solutions and Challenges. *IEEE Commun. Mag.* **2016**, *54*, 106–112. [CrossRef]
257. Shi, S.; Yang, W.; Zhang, J.; Chang, Z. Review of Key Technologies of 5G Wireless Communication System. *MATEC Web Conf.* **2015**, *22*. [CrossRef]
258. Wang, C.-X.; Haider, F.; Gao, X.; You, X.-H.; Yang, Y.; Yuan, D.; Aggoune, H.; Haas, H.; Fletcher, S.; Hepsaydir, E. Cellular Architecture and Key Technologies for 5G Wireless Communication Networks. *IEEE Commun. Mag.* **2014**, *52*, 122–130. [CrossRef]

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

5G
VISION

DMC R&D Center, Samsung Electronics Co., Ltd.

Date : February 2015

Samsung envisions the fifth Generation (5G) mobile communication era to be the beginning of a full scale Internet of Things (IoT). Billions of connected devices autonomously interconnect with one another while ensuring personal privacy. The unprecedented latencies offered by 5G Networks will enable users to indulge in gigabit speed immersive services regardless of geographical and time dependent factors. This white paper introduces you to future services, key requirements, and enabling technologies that will herald in the 5G era that is expected to revolutionize the way we experience mobile services.

Dawn of the 5G Era

Fuelled by the unprecedented growth in the number of connected devices and mobile data, and the ever-fast approaching limits of the 4G technologies to address this enormous data demand, industry efforts and investments to define, develop and deliver the systems and specifications for the fifth-Generation (5G) mobile system and services are well under way - signaling the dawn of the 5G Era.

As shown in Figure 1, the number of connected Internet of Things (IoT) is estimated to reach 50 Billion by 2020 [1], while the mobile data traffic is expected to grow to 24.3 Exabytes per month by 2019 [2].

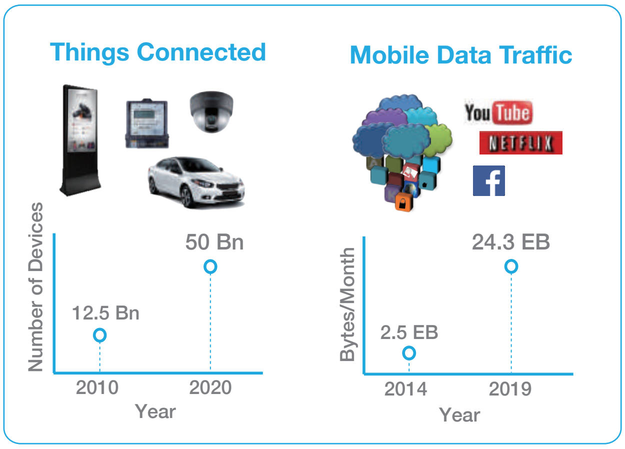

Figure 1 Growth in Mobile Traffic and Connected Devices

Add to it, the impact of higher cell capacity and end-user data rate requirements due to emerging new services such as Ultra-High-Definition (UHD) multimedia streaming and extremely low latency requirements for cloud computing and storage/ retrieval, and it soon becomes evident that the current 4G systems, which are already stretched to near-breaking points (despite massive Wi-Fi offload), will be stretched too thin to deliver the quality of experience (QoE) necessary to support mobile experience that 5G is set to deliver.

5G Era can be expected to revolutionize the way we communicate by supporting immersive applications that demand exceptionally high-speed wireless connections, a fully-realized IoT, experience lower latency and promote both spectrum and energy efficiency. To realize these benefits, 5G systems will differ fundamentally from their predecessors fueling a series of groundbreaking innovations. Let's look at the services and the requirements that 5G is expected to address.

5G Service Vision

5G services have the potential to revolutionize the mobile experience. Here's how:

Internet of Things

5G will make the "Internet of Things" a reality. With 5G technology, a device will be able to maintain network connectivity regardless of time and location, and open the possibility to connect all the connected devices without human intervention. For this, the basic fabric of the 5G system design is expected to provide support for up to a million simultaneous connections per square kilometer, enabling a variety of machine-to-machine services including wireless metering, mobile payments, smart grid and critical infrastructure monitoring, connected home, smart transportation, and telemedicine. Intelligent devices will communicate with each other autonomously in the background and share information freely. This ubiquitous connectivity - a basic tenet of the 5G services, will truly enable IoT services which in turn is expected to profoundly change human lives by connecting virtually everything.

– Smart Home

Dishwashers will fix themselves using information from peers of the same model while home appliances at home and in the neighboring homes may cooperate to extinguish a fire. A smart refrigerator, recommending a recipe of cuisine to be cooked with ingredients that are already in your refrigerator, is yet another plausible scenario.

– Fitness & Healthcare

Connected Health and Fitness related wearable devices such as The Samsung Gear™ Fit will record your athletic performance while you exercise and recommend the type of exercise, its duration and frequency per day. These connected healthcare devices will also send vital signs such as brainwave, blood pressure and heartbeat to an expert system in the hospital in real-time to prevent medical emergencies before they occur. Such time critical applications put unprecedented requirements on latency.

– Smart Store

In large shopping malls, while many people walk around window shopping, vicinity to products is continuously tracked, usually by a server somewhere in the cloud. Customized alerts for low priced product can be sent to the user's device as the user is detected in the vicinity of that low-priced product, or information of other lower prices in the nearby stores can be sent to the device as the customer spends more time in the vicinity of a certain product class.

Such a system can be tailored to deliver a highly customized experience thereby greatly enhancing a user's shopping experience. To support such a scenario, massive connectivity and low latency technologies are inevitable.

– Smart Office

In smart office environments, office appliances are connected with one another and will share information. Nearby computers and input/output devices can recognize a user and change the settings using the user's preferences stored in the IoT cloud. Printers will print out the relevant documents when the user passes by the printer. Almost all the office appliances will connect wirelessly, while exchanging massive data through wireless medium without noticeable delay. Alerts on the upcoming meetings, materials and documents relevant to that meeting will instantaneously become available to the user's device, while documents and tasks that are modified will be automatically updated.

– Connected Car

Many of us use navigation services via in-car navigation systems or our smart phones to find the most efficient route to our destination. Vehicle diagnostic services are becoming attractive to obtain the information such as battery level, fuel level or engine status on our smartphones. The 'eCall' system that automatically calls emergency services in case of an emergency exemplifies such a service. By 2020 and beyond, more attractive services that wirelessly connect 'cars' around us with 'things' will emerge and make people in the vehicles very comfortable, and provide an enjoyable driving experience. For safer driving, sensor and camera data in a vehicle as well as supplementary information from the neighboring vehicles will be collected using mobile networks so that a potential emergency situation can be reliably informed to a driver in real-time and timely steps can be taken to avoid an actual emergency situation. This operation will eventually be applied to self-driving cars, which can be viewed as an important type of 'things'.

Immersive Multimedia Experience

In a 5G environment, users will experience life-like multimedia streams anytime and anywhere. Users will feel as if they are part of the scene when they watch videos on their smart devices. To provide such an immersive experience, many obstacles will need to be overcome. Agility to instantaneously respond to the user's thought and behavior will be necessary. An upcoming service that is expected to provide life-like experience in 5G system is UHD video streaming with its greatly enhanced resolution and clarity. Currently, UHD services over terrestrial broadcast are already being standardized in some countries. In addition, some smart phones in the market are now equipped with a camera module

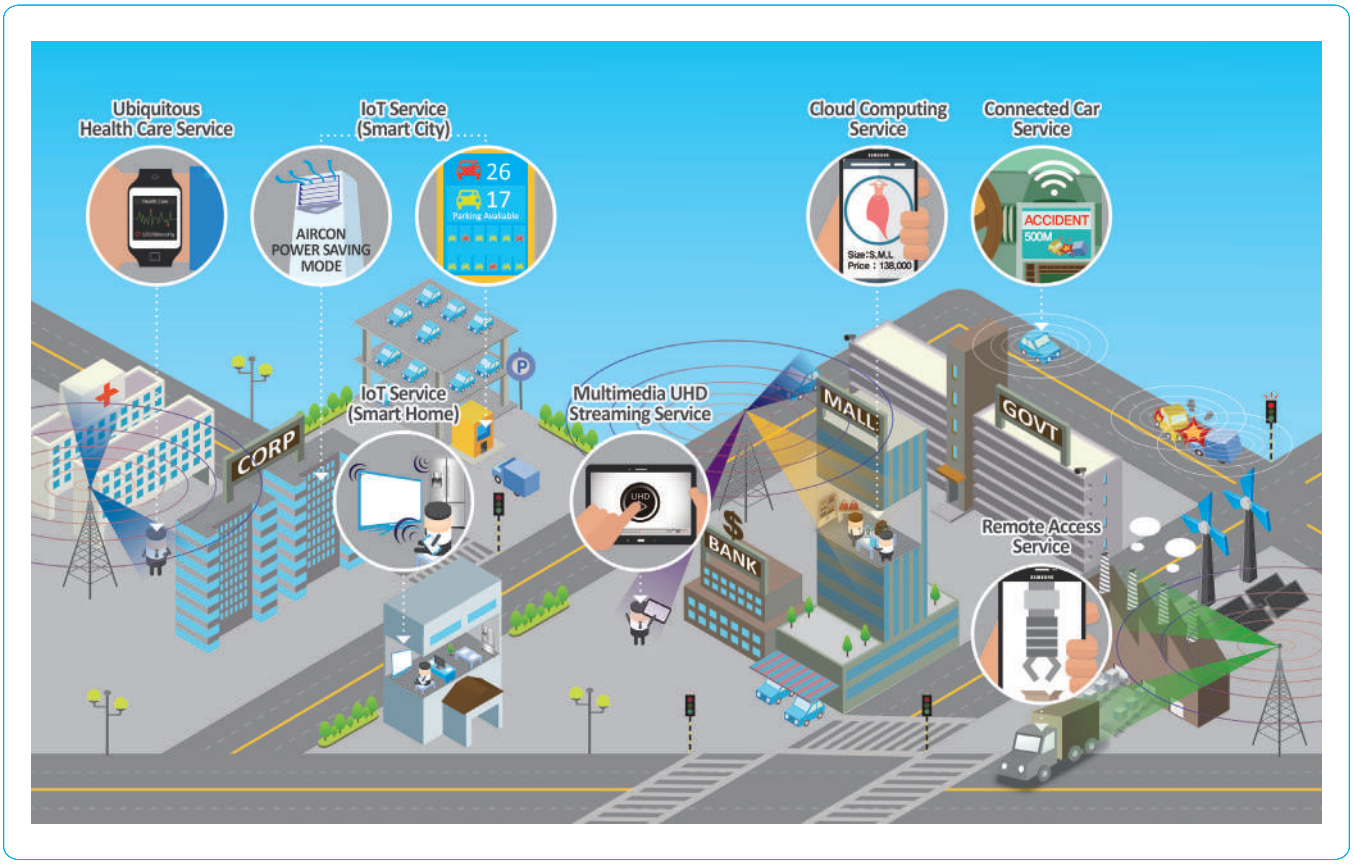

Figure 2 Major Service Scenarios with 5G

that can record video with 4K UHD quality. It is expected that UHD services are likely to go mainstream by 2020 raising an acute need for performance enhancement of cellular systems to support such services.

Other examples of “immersive” services that will fundamentally revolutionize entertainment, health, education, and other industry sectors are Virtual Reality (VR) and Augmented Reality (AR).

VR provides a world where physical presence is simulated by computer graphics, and the user can actively interact with the simulated elements, as in immersive sports broadcasting for instance (See Figure 3). Other interesting VR service scenarios are interactive 360° movies, online games, remote education, and virtual orchestra. Samsung’s recently launched VR headset called (called Gear VR), virtual reality video service platform (called Milk VR) and ‘Project Beyond,’ (a 360° 3D camera with 17 Full HD (FHD) camera modules, optimal for generating contents for Milk VR) indicate the humble beginnings of the truly immersive experience that is to come in the 5G Era.

Figure 3 Watching Sport Events with VR

In an AR service scenario, computer-aided real-time information based on user context is graphically augmented to the display, delivering added value for the user. In the future the service desk attendant no longer needs to memorize the tiresome details of the products. Instead, AR services will help to inform the price, popularity, and details of a given product. Figure 4 illustrates another service scenario - AR navigation on windshield, where navigation information and other helpful notifications (refuel reminder and nearby shop location) are displayed on the windshield of a car, so that the user can continue to focus on driving while getting subtle context-aware notifications about potential services at the same time.

Samsung is actively involved in the development of 5G technologies to support these immersive VR and AR services, which will entertain users and provide a truly life-like experience on the move.

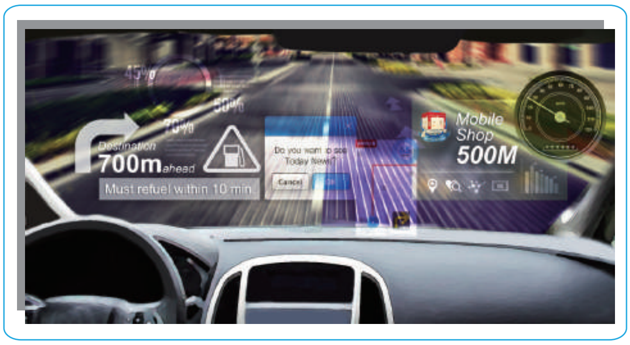

Figure 4 Driving a Car with AR Navigation

Everything on the Cloud

5G will provide users with a desktop-like experience based on cloud computing. With everything stored and processed on the cloud and immediately accessed with low latency, only simple input and output interfaces on mobile devices are needed, making them lighter, thinner, fancier, and more eco-friendly.

As an example, when you go shopping, the smart device can notify you about the arrival of new dresses that you might like, or let you know how well a dress in the newly incoming inventory matches with your liking based on your purchase history. This notification can be triggered, for instance, as you step into a shop, or take a picture of the dress, while the necessary computation to come up with the dress recommendation through crowd sourcing is done on a cloud server that is potentially half a world away.

Intuitive Remote Access

Users will be able to control remote machines and appliances as if they are right before them, even from thousands of miles away. Thanks to the reliable connections and near-zero latency of 5G, users will be able to control heavy industrial machines remotely, or access hazardous site remotely. It will also help mankind in exploring areas on the earth that are as yet unexplored, such as the Polar Regions or parts of the ocean floor.

5G Requirements

In order to realize such a demanding and unprecedented service vision, Samsung proposes the 5G rainbow requirements consisting of 7 Key Performance Indices (KPIs) as shown in Figure 5.

5G systems will be required to deliver an order of magnitude cell capacities and per-user data rate compared to its predecessors. Specially, 5G systems are expected to support data rates of 10-50 Gbps for low-mobility users. As a baseline, 5G sys-

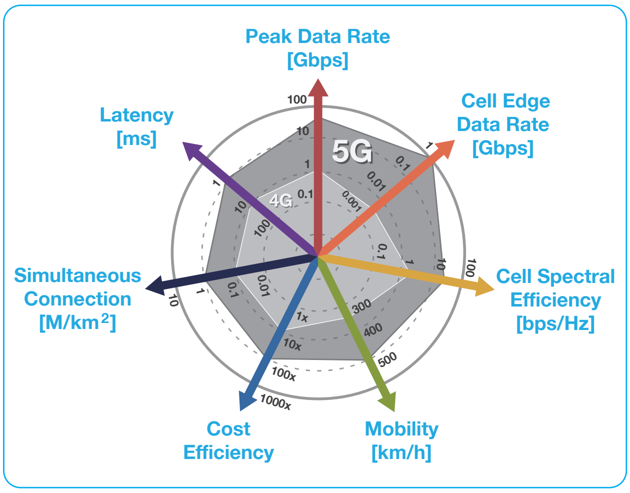

Figure 5 5G Rainbow Requirements

tems will provide gigabit-rate data services regardless of a user's location as shown in Figure 6 and Figure 7. To provide this uniform QoE, 5G network deployments are expected to be much denser compared to 4G networks, so making cost-effective deployment is a very important requisite.

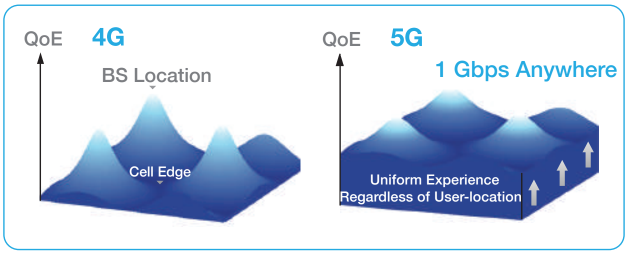

Figure 6 Edgeless RAN - 1 Gbps Anywhere

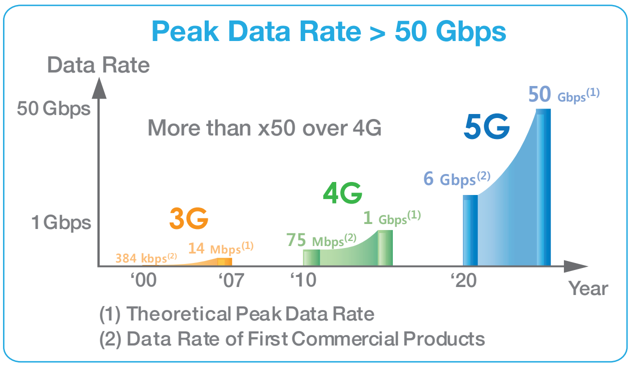

Figure 7 Data Rate Comparison of 5G with 3G and 4G

To fundamentally support the cloud storage/computing infrastructure of the future, 5G networks will deliver an end-to-end latency of less than 5 milliseconds and over-the-air latency of less than one millisecond (see Figure 8) - which is one-tenth compared to the latency of a 4G network. Critical infrastructure monitoring, for example, currently requires service levels achievable only on dedicated wireline networks while 5G technologies offer the promise of making these service levels achievable over wireless networks. Likewise, low-latency networks will allow pre-crash sensing, enabling vehicles to sense imminent collisions and exchange relevant data that could salvage the situation and/or mitigate adverse impact of the collision. Other challenging low-latency services that could be enabled by 5G could include self-driving cars, public safety communications systems, augmented reality, and "tactile internet" [3].

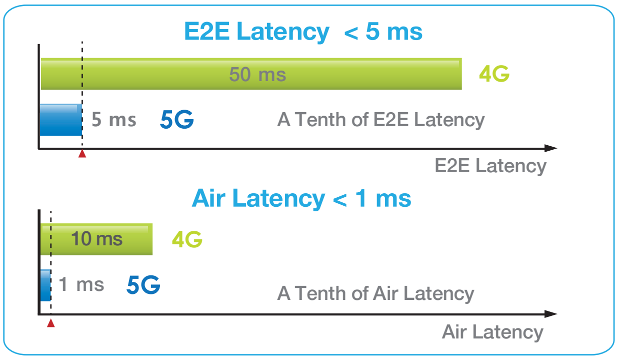

Figure 8 Ultra Low Latency of 5G

With spectral efficiency requirements set to 10 bps/Hz levels (in contrast to the 1-3 bps/Hz on 4G networks), 5G is also expected to deliver an efficient use of the spectrum by using MIMO, advanced coding and modulation schemes and new waveform design (more on this in the enabling technology section)

To address the ever-widening revenue gap that the operators and service providers are experiencing, 5G systems are targeted to be 50 times more efficient than 4G by delivering reduced cost and energy usage per bit. This sequentially requires low-cost network equipment, lower deployment costs, and enhanced power saving functionality on the network and user equipment sides.

5G technologies will be required to cope efficiently with all degrees of mobility by providing "mobility on demand" based on each device's and service's needs. On one hand, the mobility of user equipment should be guaranteed to be at least the same level as the current 4G system - that is the baseline. On the other hand, Samsung envisions that specialized 5G systems will support mobility at speeds ranging from 300 to 500 km/h.

To make the IoT Vision come true, the number of simultaneous connections in the 5G system is expected to be over 10^6 per unit square kilometer, which is much higher than that of the legacy system.

5G Key Enabling Technologies

Groundbreaking innovations will drive 5G technologies to meet the unprecedented speeds, near-wireline latencies, ubiquitous connectivity with uniform QoE, and the ability to connect massive amounts of devices with each other, all working in unison to provide the user with an immersive experience, even while the user is on the move. Future 5G systems will encompass fundamentally new designs to boost wireless capacity utilizing new frequency bands, advanced spectrum efficiency enhancement methods in the legacy band, and seamless integration of licensed and unlicensed bands.

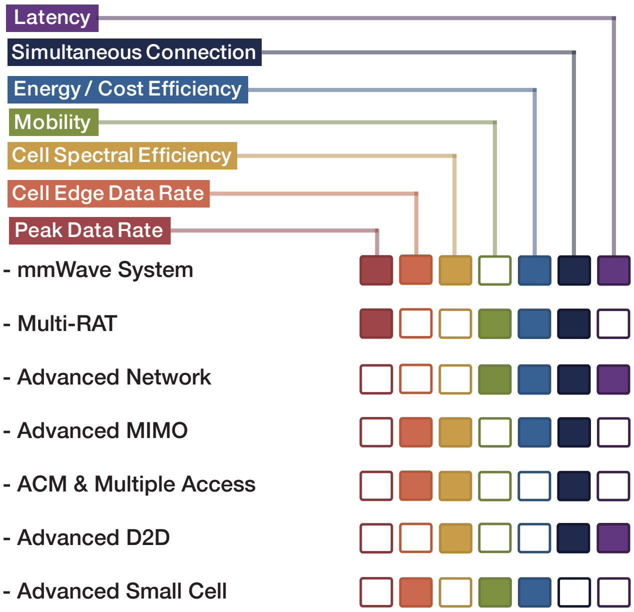

	Peak Data Rate	Cell Edge Data Rate	Cell Spectral Efficiency	Mobility	Energy / Cost Efficiency	Simultaneous Connection	Latency
- mmWave System	■	■	■	☐	■	■	■
- Multi-RAT	■	☐	☐	■	■	■	☐
- Advanced Network	☐	☐	☐	■	■	■	■
- Advanced MIMO	☐	■	■	☐	■	■	☐
- ACM & Multiple Access	☐	■	■	☐	☐	■	☐
- Advanced D2D	☐	☐	■	☐	☐	■	■
- Advanced Small Cell	☐	■	☐	■	■	☐	☐

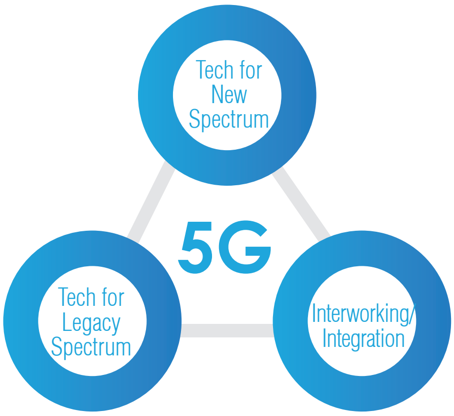
Tech for New Spectrum
5G
Tech for Legacy Spectrum
Interworking/ Integration

Figure 9 shows an overview of the 5G key enabling technologies. The massively higher capacity needs of the 5G systems will be addressed by new mmWave systems, high-density small cells, advanced Multiple-Input and Multiple-output (MIMO) and new multiple access schemes like Filter-Bank Multi-Carrier (FBMC). Adaptive Coding and Modulation like Frequency and Quadrature Amplitude Modulation (FQAM) can significantly improve the cell edge performance and combined with higher density deployments with multi-BS cooperation will help to deliver on the promise of “Gbps anywhere” and Uniform QoE. Multi-Radio Access Technology (Multi-RAT) integration including carrier

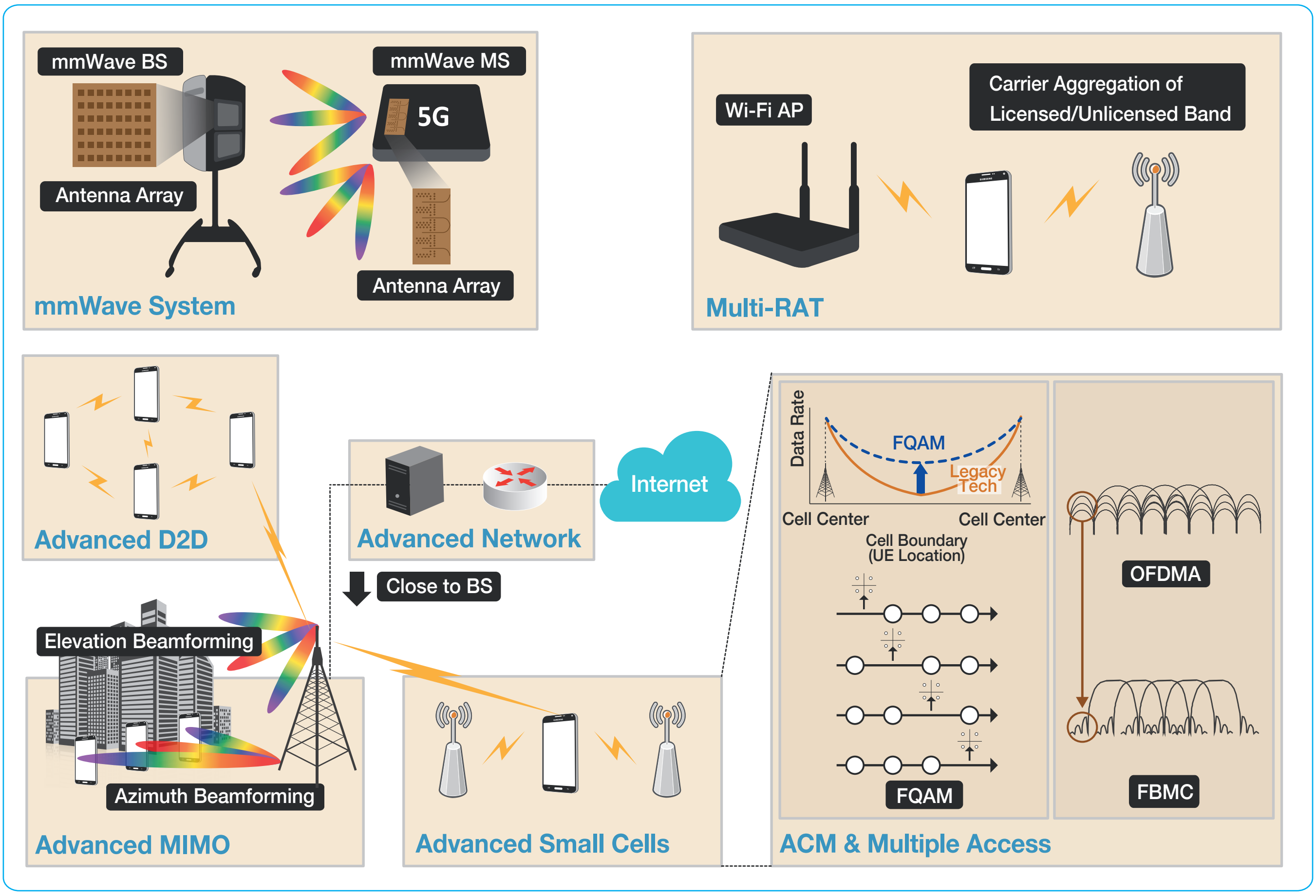
mmWave BS
mmWave MS
5G
Antenna Array
Antenna Array
mmWave System
Wi-Fi AP
Carrier Aggregation of Licensed/Unlicensed Band
Multi-RAT
Advanced D2D
Advanced Network
Internet
Close to BS
Elevation Beamforming
Azimuth Beamforming
Advanced MIMO
Advanced Small Cells
Data Rate
FQAM
Legacy Tech
Cell Center
Cell Center
Cell Boundary (UE Location)
FQAM
OFDMA
FBMC
ACM & Multiple Access

Figure 9 Overview of 5G Key Enabling Technologies

aggregation of licensed and unlicensed bands will inevitably help in increasing the available system bandwidth. On the network side, novel topologies including application servers placed closer to the network edge will contribute to significantly reducing the network latency. Advanced Device-to-Device (D2D) technology can help reduce the communications latency and support larger number of simultaneous connections in a network.

These 5G key enabling technologies are described in more detail in the following sections.

mmWave System

The mmWave bands provide 10 times more bandwidth than the 4G cellular-bands, as illustrated in Figure 10. Therefore, the mmWave bands can support the higher data rates required in future mobile broadband access networks.

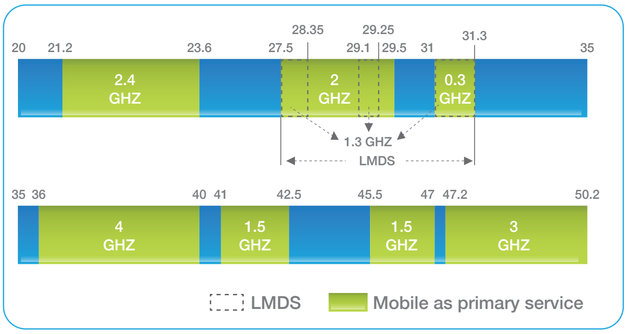

Figure 10 Potential Bands in 20-50 GHz (US)

Unlike the below 6GHz bands, in the mmWave bands we have to carefully consider the wireless conditions such as rainfalls, snowfalls, and fogs. However, the amount of loss is mild for the expected range of a 5G communication link. For example, atmospheric absorption loss due to H_2O and O_2 at 28 GHz would be around 0.02 dB at a distance of 200 meters. Intensive rainfall (110 mm/h) results in about 4 dB loss at 200 meters distance. Loss due to heavy snow (10 mm/h) or heavy fog (visibility distance under 50 meters) will be less than 0.1 dB, at all distance of 200 meters [4][5].

In Figure 11, we show mmWave channel measurements from extensive experiments in Daejeon, Korea [6][7]. These channel measurements were carried out at 28 GHz with a channel bandwidth of 250 MHz, transmission power of 29 dBm and the horn antenna gain of 24.5 dBi for both the transmitter and the receiver [6][7].

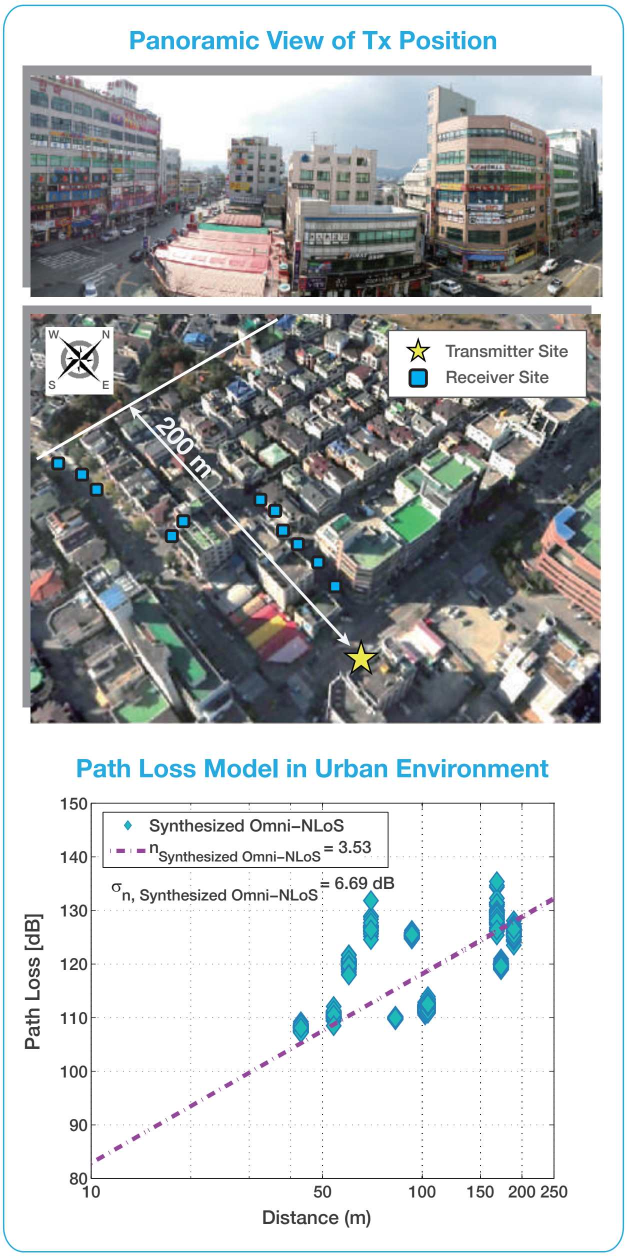
Panoramic View of Tx Position

Path Loss Model in Urban Environment

Figure 11 mmWave Channel Measurements in Urban Area, Daejeon (Korea)

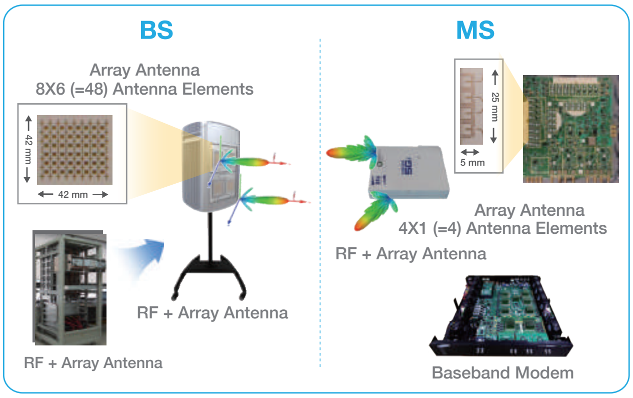

Figure 12 Adaptive Pencil Beamforming (Example)

The path loss exponents (based on 8 meters distance free space path loss) is 3.53 in Non Line of Sight (NLoS) links [6]. The mild path loss exponent from measurement results strongly suggests that mmWave communication links can be supported over 200 meters of distance, even in dense urban outdoor NLoS environments. Similar results are also reported in ray-tracing-based mmWave propagation model [8], confirming the fact that mmWave bands can be successfully utilized for future IMT systems. Moreover, based on the ray-tracing, mmWave 3D-channel model in urban scenario is proposed in [9]. Higher path-loss and the resulting fragile link is largely due to weak diffractions in the mmWave bands. We need to overcome these challenges in order to make outdoor mmWave communication a reality [10][11]. Fortunately, the small mmWave wavelength also means highly directional beamforms can be obtained using a large number of antenna elements in a smaller form factor. These adaptive directional beams with large antenna array gain are key in combating the large propagation loss in the mmWave [12][13][14][15], as illustrated in Figure 12.

We have developed a mmWave beamforming prototype at the DMC R&D Center, Samsung Electronics, Korea, in order to demonstrate the feasibility of using mmWave bands for cellular services. We showed that mmWave system can meet the two key requirements of cellular services: sufficiently large geographical coverage and support for mobility in NLoS environments.

With the prototype, we tested outdoor coverage to demonstrate the service availability in a typical outdoor environment for both LoS and NLoS sites. The tests were performed at sites surrounded by tall buildings where various channel propagation effects such as reflection, diffraction, and penetration were present, as shown in Figure 13. We observed that reliable communication links are formed even for NLoS sites that are more than 200 meters away from the base station.

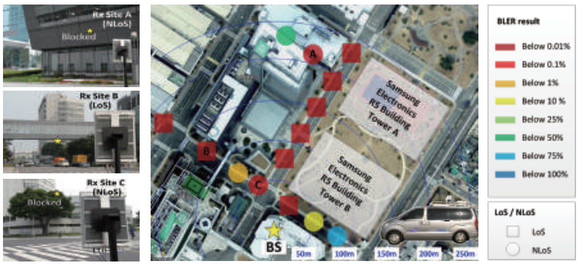
Rx Site A (NLoS)
Blocked
Rx Site B (LoS)
Rx Site C (NLoS)
Blocked
A
B
C
BS
Samsung Electronics RS Building Tower A
Samsung Electronics RS Building Tower B
50m
100m
150m
200m
250m
BLER result
Below 0.01%
Below 0.1%
Below 1%
Below 10 %
Below 25%
Below 50%
Below 75%
Below 100%
LoS / NLoS
LoS
NLoS

Figure 13 Exemplary Outdoor Coverage Test Results

Advances in semiconductor technology have made commercial mmWave systems readily available. Recent technologies of Silicon-based Complementary Metal Oxide Semiconductor (CMOS) processes are capable of integrating mixers, Low Noise Amplifiers (LNAs), Power Amplifiers (PAs), and Inter-Frequency (IF) amplifiers, all in a single package. A good example is the 60 GHz commercialized products with the label of Wireless Gigabit Alliance (WiGig) [16], and it is well recognized that cost effective implementations of sub-100nm CMOS process made it possible to utilize the 60 GHz bands [17].

GaAs Monolithic Microwave Integrated Circuit (MMIC) technologies are maturing fast, and they are becoming a dominant choice for components in the RF chain including PAs, LNAs, switches for digital attenuators and phase shifters, Voltage Controlled Oscillators (VCOs), and passive components from a few GHz to 100 GHz. As illustrated in Figure 14, it is projected that the Power Added Efficiency (PAE) of a 5G Front End Modules (FEM) will soon match the PAE of today's commercial 4G system. Meanwhile, a good Effective Isotropic Radiated Power (EIRP) can be achieved with the help of high antenna gains from a large number of antenna elements [18][19].

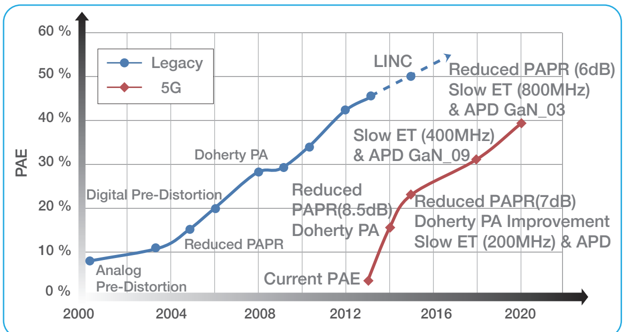

Figure 14 Projections of 5G FEM PAE

High performance antenna solution is an important piece of the 5G puzzle. The constantly varying LoS and NLoS propagation environment demands a novel antenna which exhibits high gain as well as wide spatial coverage capacity. The antenna array comprising of multiple antenna elements must fit within the small form factor of a 5G mobile handset.

Samsung has been developing innovative 5G phased array antennas that have near zero-footprint and reconfigurable antenna modes, as shown in Figure 15.

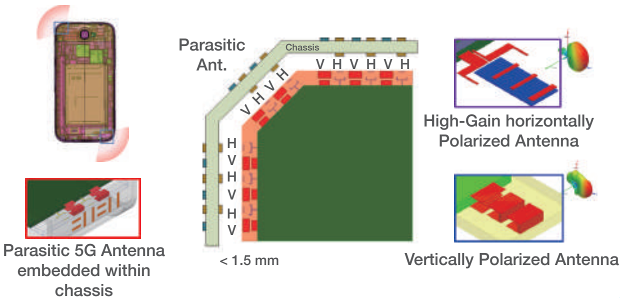
Parasitic Ant.
Chassis
V H V H V H
V H V H
H V H V H V
< 1.5 mm
Parasitic 5G Antenna embedded within chassis
High-Gain horizontally Polarized Antenna
Vertically Polarized Antenna

Figure 15 The Reconfigurable 5G Phased-array Antenna

Characterizing the biological implications on a user's body imposed by cellular handset devices is one of the most important aspects that need to be verified prior to its commercial release. The Signal Absorption Rate (SAR) regulated by worldwide governmental bodies is used as a guideline to assess the effect of mmWave bands on a user's body. The SAR of an envisioned 5G (28 GHz) cellular handset is analyzed and illustrated in Figure 16.

When optimally configured, the 5G beamforming technology enables the peak radiation of the Mobile Station (MS) antenna to steer away from the user's head. As a result, the average SAR can be reduced by more than a factor of 10 compared to that of the present day 4G cellular handsets.

The maximum SAR is expected to be further minimized in the future, as 5G antenna technology continues to evolve.

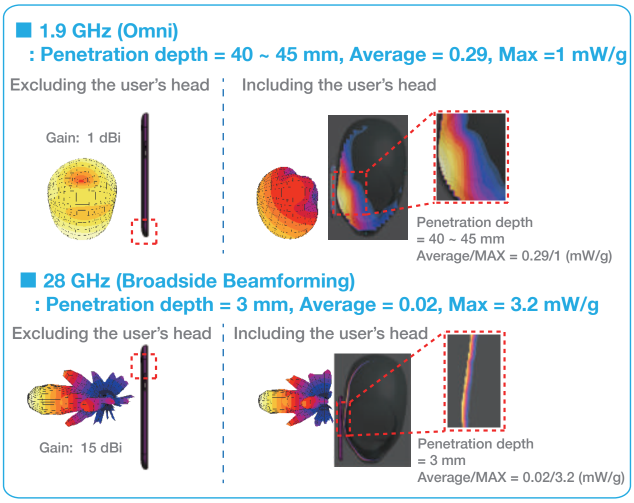

Figure 16 Biological Implications on the User's Body

Multi-RAT

To realize the envisioned 5G services, significant enhancement of per-user throughput and overall system capacity is required, compared with those of the 4G system. Such an enhancement can be achieved through advanced PHY/MAC/network technologies and efficient methods of cell deployment and spectrum management. In particular, utilizing a larger amount of system bandwidth guarantees an increase in the capacity by allocating more frequency resources to each user in the system. Therefore, utilizing the spectrum where huge bandwidth is available can be considered to be the most critical issue for the 5G system.

Currently, the 4G system specifies its operating frequency bands and some of them are assigned to cellular operators. These deployments can be beneficial since the existing 4G system will not interfere with other RATs.

However, obtaining the licensed spectrum requires not only considerable financial investment, but also a significantly long period of time spent on regulatory procedures. More importantly, a substantial portion of the licensed spectrum around 2 GHz is already being used. Therefore, finding a bandwidth that is wide enough to support the 5G system is very challenging.

The recent trend in spectrum management is to aggregate both the licensed and unlicensed spectrums to extend available system bandwidth, as shown in Figure 17.

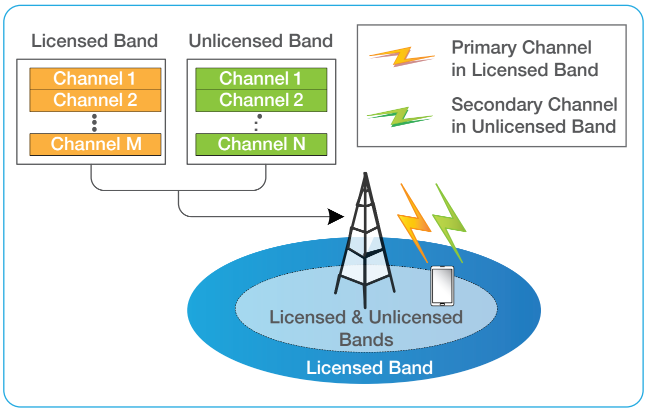
Licensed Band
Channel 1
Channel 2
Channel M
Unlicensed Band
Channel 1
Channel 2
Channel N
Primary Channel in Licensed Band
Secondary Channel in Unlicensed Band
Licensed & Unlicensed Bands
Licensed Band

Figure 17 Integration of Licensed and Unlicensed Bands

The unlicensed spectrum has plenty of bandwidth. For instance, approximately 500 MHz and 7 GHz bandwidths are available in the 5 GHz and 60 GHz bands, respectively. In order for the 5G system to utilize the unlicensed spectrum, regulations imposed on each frequency band should be carefully reconsidered. Representative examples of the regulations are the Transmit Power Control (TPC), Dynamic Frequency Selection (DFS), and Listen Before Talk (LBT).

To efficiently utilize the unlicensed spectrum, we will develop the 5G system with the following characteristics.

First, we will design PHY/MAC/network algorithms suitable for the nature of the unlicensed spectrum. Since a wide range of frequency bands are included in the unlicensed spectrum, each frequency band has its own characteristics. Hence, band-specific solutions will be provided.

Second, efficient coexistence mechanisms that take into account other RATs (e.g., WiFi or WiGig) operating in the unlicensed spectrum will be suggested.

Finally, techniques for interworking and integrating the 5G system with other RATs will be developed. By taking advantage of multiple RATs, the 5G system will be able to take advantage of the unique characteristics of each RAT and improve the practicality of the system as a whole. For instance, the 4G system is used for exchanging the control messages to maintain the connection, to perform handover, and to provide real-time services such as VoLTE. The technology operating in mmWave unlicensed frequency band would support the gigabit data rate service. Multiple mmWave cells can be overlaid on top of the underlying 4G macro cells, as shown in Figure 18.

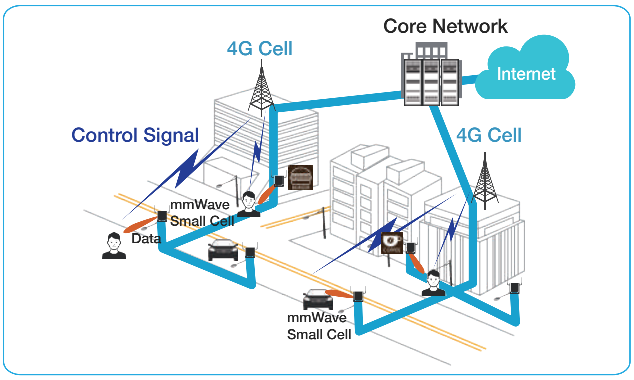

Figure 18 Overlaid Network of mmWave Small Cell Integrated with the Underlay 4G System

Advanced Network

In order to meet the 5G key requirements such as latency and the large number of simultaneous connections, and to support new business models and scenarios for operators, technologies above the radio access level should also be considered. Therefore, we need to develop technologies at the system architecture level from the network point of view. To accomplish the 5G key requirements and support the increased data rate of new 5G radio access, 5G network will have to evolve towards a distributed and flat architecture.

Current mobile network architectures designate a dedicated node in the core network (e.g., PGW - Packet Data Network Gateway in 3GPP) as a mobility anchor that allocates an IP address to the terminal, tracks terminal location in the IP topology, and ensures terminal's reachability by tunneling its traffic to wherever it goes. All terminal traffic is tunneled through the centralized node in the mobile core network. However, the undesirable consequences of this design include the following:

- Increase in end-to-end transmission latency due to elongated data paths.
- Additional load of backhauling and network processing in the core networks.
- Low network reliability due to introduction of a single point of failures.

In the 5G flat network architecture, as illustrated in Figure 19, user mobility is managed efficiently and in a dynamically scalable fashion by pushing the functionality to the edges of the network and even onto the mobile terminals [20].

The three important benefits of this approach include the followings:

First, such a distributed mobility management always provides the shortest data path between MS and the Internet without traversing the core network. This distributed mobility management leads to a significant reduction of signaling and data transmission delay. Also, low end-to-end latency requirements of 'less than 5 ms' for new 5G services such as immersive UHD video streaming, cloud gaming, and virtual reality, cannot be archived solely by reducing radio access latency but would also require a flatter network architecture design. In the flat network architecture, services which require low latency transmission can be provided by Edge Servers and they can benefit from advanced network features which utilizes network information for optimal operations.

Second, it provides a highly scalable solution compared to the centralized architecture, in which a single core network gateway maintains the whole traffic from MSs or to MSs.

Third, it easily avoids the risk of having a single point of failure. In flat network architecture, the breakdown of one network gateway would not significantly interfere with the operations of the other gateways.

Flexibility is considered as another key requirement of 5G network architecture. Software-Defined Networking (SDN) and Network Function Virtualization (NFV) provide promising examples of programmable design technologies for realizing a flexible 5G network architecture.

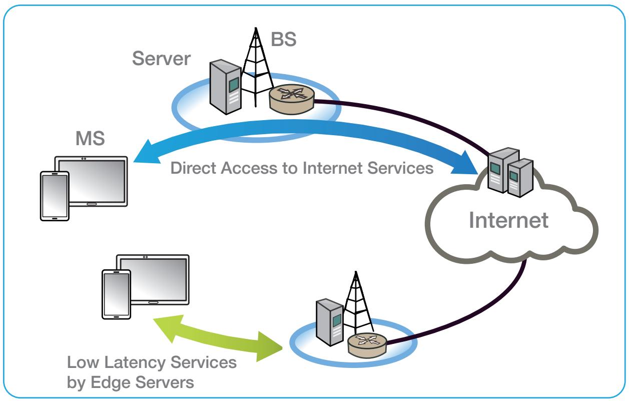

Figure 19 5G Flat Network Architecture

Advanced MIMO

One promising technology for meeting the future demands is massive MIMO transmission/reception [21]. When used with multi-user precoding schemes such as Maximum Ratio Transmission (MRT) precoding, also known as channel conjugate precoding, massive MIMO systems experience small inter-user and inter-cell interferences, and consequently achieve significantly higher throughput than the state-of-the-art MIMO systems.

One of the main challenges to build massive MIMO systems in practice is the limitation in the number of antennas that can be equipped at a BS, caused by the BS form factors and operating carrier frequencies.

For example, to horizontally install a large number of antenna elements (e.g., > 8) at the top of a BS tower operating with the lowest 4G system frequency bands of 700 MHz, eight antenna elements with 0.5 λ spacing require up to 1.7 m width, where λ is the carrier wavelength. For the typical 4G system frequency bands of 2.5 GHz, fitting 32 antenna elements with 0.5 λ spacing require up to 1.9 m width, which is still not feasible in many BSs that have only limited room on the tower. This practical limitation has motivated Full-Dimension MIMO (FD-MIMO) cellular communication systems, which place a large number of active antenna elements in a two dimensional grid at the BSs.

A typical FD-MIMO deployment scenario is illustrated in Figure 20, for a macro BS with 3 sectors equipped with 2D Active Antenna Array (AAA) panels.

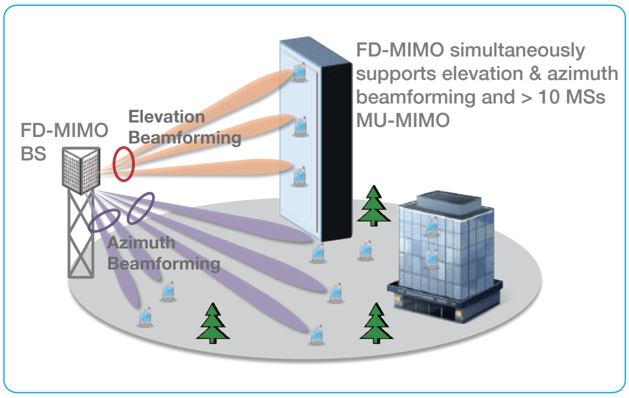
FD-MIMO simultaneously supports elevation & azimuth beamforming and > 10 MSs MU-MIMO
Elevation Beamforming
FD-MIMO BS
Azimuth Beamforming

Figure 20 Example of FD-MIMO Deployment

FD-MIMO system can support high-order Multi-User MIMO (MU-MIMO), while fully exploiting the elevation as well as the azimuth dimension, thereby generating improved system throughput.

In full buffer system-level evaluations, it has been found that 64-antenna-port FD-MIMO system achieves 243% average-cell and 244% cell-edge performance gain, compared to those of the 8-antenna-port legacy MIMO system. In order to achieve the promising gain of FD-MIMO in practice, we need accurate beam steering and tracking in three Dimensions (3D).

To steer and track the MU beams toward serving MSs, FD-MIMO BS should be equipped with multiple transceivers (TRX) feeding 2D array elements, in which the number of TRX doubles or even quadruples compared to that of the conventional BS.

Having a large number of TRX poses new challenges, such as antenna calibration and complexity issues associated with Channel State Information (CSI) acquisition and precoding.

On the other hand, high-order MU-MIMO introduces another set of new challenges, such as scheduling complexity and link adaptation. Furthermore, in Frequency Division Duplex (FDD) systems, other new challenges emerge such as pilot overhead, CSI estimation complexity, CSI quantization and feedback overheads.

ACM & Multiple Access

– FBMC

As cellular IoT has been one of key driving forces to 5G, spectrally efficient support for heterogeneous services that have quite different requirements is becoming ever so important. Accordingly, several enabling methods such as multi-RAT coexistence and flexible spectrum sharing have been actively investigated.

Recently, FBMC has drawn much attention as an enabling technology for enhancing the fundamental spectral efficiency, though its theory has a long history similar to that of Orthogonal Frequency Division Multiplexing (OFDM).

Because of the well-localized time/frequency traits adopted from a pulse shaping filter per subcarrier, the FBMC system can reduce the overhead of guard band required to fit in the given spectrum bandwidth, while meeting the spectrum mask requirement.

Furthermore, the effectively increased symbol duration is suitable for handling the multi-path fading channels even without Cyclic Prefix (CP) overhead. Consequently, the FBMC system can reduce the inherent overheads such as CP and guard-bands in CP-OFDM. FBMC is also attractive in specific asynchronous scenarios, including Coordinated Multi-Point Transmission and Reception (CoMP) and Dynamic Spectrum Access (DSA) in a fragmented spectrum.

However, to maintain the transmission symbol rate, the conventional FBMC system generally doubles the lattice density either in time or in frequency compared with OFDM while adopting Offset Quadrature Amplitude Modulation (OQAM). In OQAM, in-phase and quadrature-phase modulation symbols are mapped separately with half symbol duration offset. Thus, so-called OQAM-FBMC or Staggered Multi-Tone (SMT) causes intrinsic interference that makes it difficult to apply conventional pilot designs and corresponding channel estimation algorithms as well as MIMO schemes as in CP-OFDM systems [22].

With a base-filter that takes the spectrum confinement and the orthogonality among adjacent subcarriers into consideration, the QAM-FBMC system performs comparable to the CP-OFDM system even without the CP overhead, while the guardband overhead reduction is also available from the well-confined spectrum. Sophisticated receiver algorithms including channel estimation and equalization can further mitigate the multi-path fading channel impact without the CP.

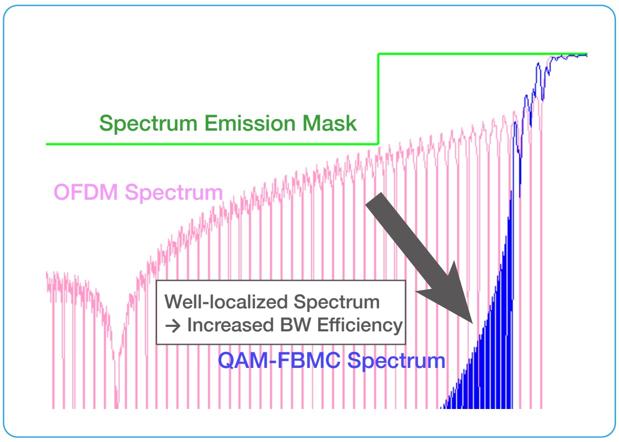

Figure 21 One QAM-FBMC Symbol Generation

– FQAM

One of key requirements for 5G is enhancement of cell-edge performance, which means that every user should be supported with Gigabit experience anywhere. Conventional approaches to enhance the cell-edge performance mainly focus on managing interference (e.g., interference cancellation, interference avoidance), by dealing with interference as a Gaussian. However, it is proved that the worst-case additive noise in wireless networks with respect to the channel capacity has a Gaussian distribution. From this observation, one can expect that the channel capacity can be increased by a non-Gaussian interference design which makes Inter-Cell Interference (ICI) non-Gaussian. The distribution of ICI depends on the modulation schemes of the interfering BSs. Therefore, an active interference design for improved cell-edge performance can be achieved by applying a new type of modulation.

FQAM, a combination of Frequency Shift Keying (FSK), and Quadrature Amplitude Modulation (QAM) can be used as an active interference design scheme. Figure 22 shows the signal constellation of 16-ary FQAM that is a combination of 4-ary FSK and 4-ary QAM.

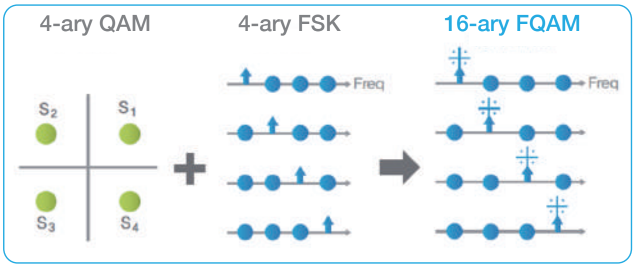

Figure 22 Example of 16-ary FQAM

With FQAM, the statistical distribution of ICI is likely to be non-Gaussian, especially for cell-edge users. As a result, the transmission rates for the cell-edge users can be significantly improved.

The statistics of ICI and the performance enhancement possibility have been proven by practical implementation of a system which uses FQAM. FQAM system environment for cellular downlink OFDMA networks is shown in Figure 23 and 24. Our experimental results show that the transmission rates for interference-limited users in FQAM-based OFDMA networks are around 300% higher than those in QAM-based OFDMA networks [23][24].

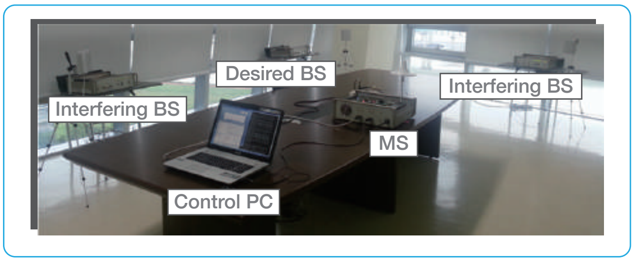

Figure 23 3 Cell Structure with MS and BSs

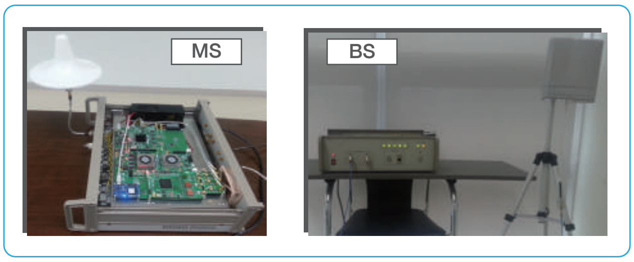

Figure 24 Implemented MS and BS

Advanced D2D

Advanced D2D communication is an attractive technology that enhances spectral efficiency and reduces end-to-end latency for 5G. Not entirely depending on the cellular network, D2D devices can communicate directly with one another when they are in close proximity. Hence, D2D communication will be used for offloading data from network so that the cost of processing those data and related signaling is minimized. Moreover, enhanced version of D2D communication was suggested lately for being used as special purpose such as Mission Critical Push-To-Talk (MCPTT) communication and Vehicle-to-Anything (V2X) communication.

In advanced D2D communication, a single radio resource can be reused among multiple groups which want to communicate with each other if the interference incurred between groups is tolerable. Hence, we can increase the spectral efficiency and the number of the simultaneous connection by utilizing D2D communication in 5G. Moreover, since the data is directly transmitted and not going through the core network, the end-to-end latency can be considerably reduced. Therefore, advanced D2D communication is quite well aligned with IoT services as shown in Figure 25. The cars can communicate with each other to exchange the information for safety alarm and infotainment without cellular base station. The home appliances communicate with each other for home automation service. Many objects in proximity region can be connected to each other so that IoT services can be accomplished.

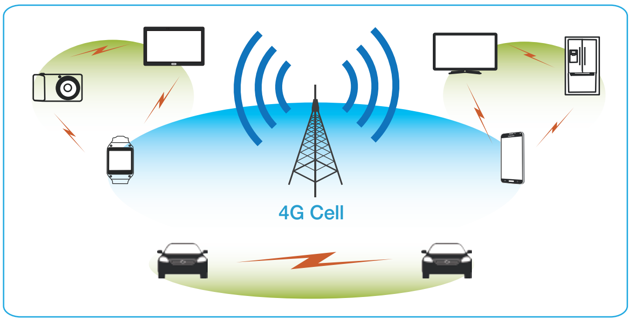

Figure 25 Advanced D2D Communications

Advanced Small Cell

To achieve significant throughput enhancement in a practical manner, it is necessary to deploy a large number of cells in a given area and to manage them intelligently. The 5G system is expected to utilize higher frequencies to take advantage of the vast bandwidth in the mmWave bands. Hence, the considerably high propagation loss of mmWave makes it suitable for dense small cell deployment, which leads to higher spatial reuse.

Moreover, Figure 26 shows the concept of a user-centric virtual cell. Conventional static network topologies with a central controller have an "edge", the reach of the central controller. However, a user-centric virtual cell that consists of a group of cooperating BSs is continuously reformed so that any user will always find himself/herself at the "center" of the cell.

Distributed and self-configuring network technologies will make it easy to deploy many small BSs in urban and suburban areas. In-band wireless backhaul can be used between BSs for cooperative communication, reducing the cost and complexity of backhaul network deployment.

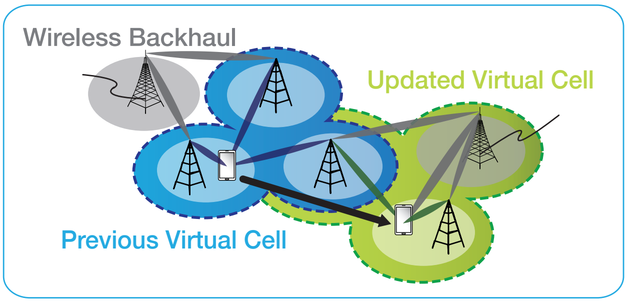

Figure 26 User Centric Virtual Cellular Network

5G Deployment Scenario

In order to provide the ubiquitous high QoE gigabit accessibility, we envision an overlaid deployment of 5G in conjunction with the existing 4G macro cells. The 5G small cells can be coupled with the overlaid 4G macro cells, and the 4G macro cells will control the operations of the associated 5G small cells, as illustrated in Figure 27. In the figure, 5G BS primarily provides multi-gigabit per second throughput with high QoE to mobile users over the legacy spectrum or higher spectrum like the 5 GHz unlicensed band and bands above 6 GHz. Meanwhile, the 4G BS can serve as a control channel to 5G MSs for supporting seamless connection anywhere over the legacy 4G spectrum.

5G system will need enhanced RAN technologies that not only utilize the legacy frequency bands assigned to IMT and IMT-Advanced systems, but also use new frequency bands. At the same time, 5G system will also need to support seamless interworking with other RATs operating in unlicensed bands.

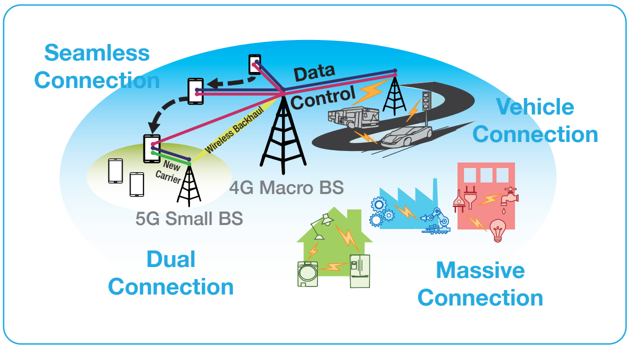

Figure 27 5G Deployment Scenario

Standardizations and Regulations

As shown in Figure 28, 5G is planned to be commercialized in year 2020. ITU-R WP 5D is preparing timelines in the vision document for standardization, spectrum allocation, and commercialization, in order to be on time. From the perspective of commercialization, standards for 5G should be ready by the year 2017 to allow 2 or 3 years for the development of 5G products. Considering average periods of previous standardization, 5G standards need to get started in 2015 to make 5G standards available by the end of 2017.

New frequency bands may be required to achieve the target performance of 5G. World Radio communication Conference (WRC)-15 agenda item 1.1 made in WRC-12 indicates consideration for additional spectrum allocations to mobile services on a primary basis and for identification of additional frequency bands for IMT purposes.

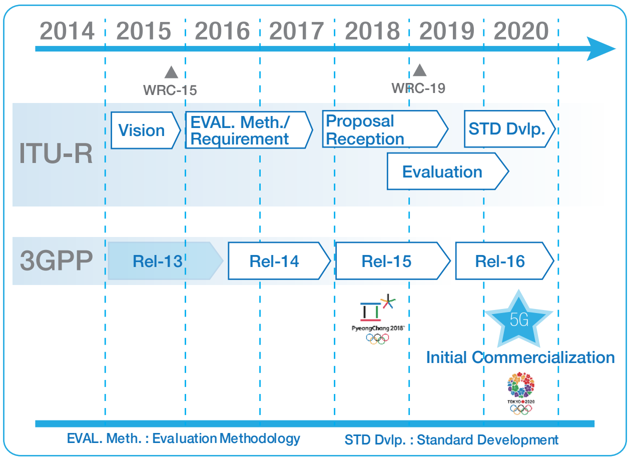

Figure 28 5G Timelines

In light of this agenda item, various frequency ranges had been proposed and discussed in ITU-R WP 5D in July 2013, and Joint Task Group (JTG) 4-5-6-7 had performed compatibility studies among various technologies in each candidate frequency band until July 2014.

Although bands above 6 GHz were not included in JTG 4-5-6-7's candidate frequency bands, WP 5D has been developing a new report on IMT systems in bands above 6 GHz, including channel characteristics and the potential usage of the bands for IMT systems.

WP 5D may also provide JTG 4-5-6-7 with the necessary information regarding bands above 6 GHz in order to be referenced as appropriate in the draft text of Conference Preparatory Meeting (CPM). The possibility and further necessity to study the bands are captured in the report from JTG 4-5-6-7 to CPM as follows:

"The demand of high bit rate, especially in densely populated area could be accommodated in higher frequency bands (e.g. above 6 GHz) than those currently being considered in studies, however the technical information required for compatibility studies has yet to be developed and these studies and proposals are being explored for future work, beyond WRC-15."

This report implies that new frequency bands can be added to the IMT frequency bands toward WRC-19, and it is likely that bands above 6 GHz can be one of candidate frequency bands.

In addition, Licensed Assisted Access (LAA) study item has been accepted in 3GPP LTE in the second half of 2014, acknowledging the use of unlicensed bands for cellular system.

Recently, as FCC relieves 5GHz band regulations for various unlicensed band usage, related coexistence technology between LTE and WLAN will be standardized in both standards.

Further, as shown in Figure 29, Samsung is actively engaging in the most of global 5G research initiatives, such as European 5G PPP projects of Horizon 2020, 5G Innovation Centre (5GIC) in UK, NYU Wireless Center in US, Giga KOREA project and Chinese 836 project.

Samsung is leading various collaborations with industries and academics over the world. In particular, Samsung has played an important role as the full member of 5G PPP Infrastructure Association, the executive board member of 5G Forum in Korea and the chair of vision sub-working group for Future IMT (5G) in ITU-R WP5D.

In order to have a consistent perspective on 5G with those of other academic institutes, we are vigorously developing 5G core technologies with several outstanding universities around the world.

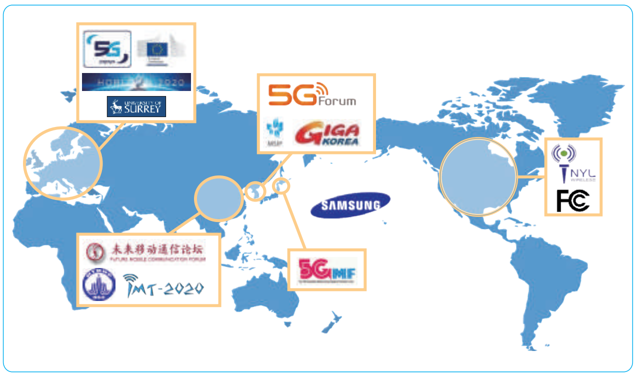

Figure 29 Global 5G R&D Activities

Conclusion

5G will usher in a revolutionary generation of mobile communication that provides ubiquitous multi-Gbps data rate regardless of the user's location.

Significantly increased system capacity and real-time responsiveness of the 5G system will introduce life-changing services providing the users with a truly immersive and rich experience. The profoundly life-alternating world of billions of interconnected devices, the IoT, will require something as revolutionary as 5G to be fully realized in its fullest extent. And finally 5G promises to reverse the widening revenue gap and make it worthwhile for operators and service providers to invest again in innovative new services, and continue to propel increased productivity and efficiency.

Reference

1 UMTS, “Mobile Traffic Forecasts 2010-2020 Report,” UMTS Forum, Jan. 2011.

2 Cisco, “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update: 2013-2018,” Cisco, Feb. 2014.

3 Calum Dewar et al., “Understanding 5G: Perspectives on future technological advancements in mobile,” GSMA Intelligence Report, Dec. 2014.

4 ITU-R P.840-5, “Attenuation Due to Clouds and Fog,” ITU-R, Feb. 2012.

5 DARPA R-1335-ARPA, “Atmospheric Effects on Terrestrial Millimeter-Wave Communications,” DARPA Report, Mar. 1974.

6 Sooyoung Hur et al., “Wideband Spatial Channel Model in an Urban Cellular Environments At 28 GHz,” in Proc. EuCAP’15, April. 2015.

7 Sooyoung Hur et al., “28 GHz Channel Modeling Using 3D Ray-Tracing in Urban Environments,” in Proc. EuCAP’15, April. 2015.

8 Youngbin Chang et al., “A Novel Dual-Slope mm-Wave Channel Model Based on 3D Ray-Tracing in Urban Environments,” in Proc. IEEE PIMRC’14, Sep. 2014.

9 Sooyoung Hur et al., “Synchronous Channel Sounder Using Horn Antenna and Indoor Measurements on 28 GHz,” in Proc. IEEE BlackSeaCom’14, pp. 83-87, May 2014.

10 Zhouye Pi et al., “An Introduction to Millimeter-Wave Mobile Broadband Systems,” IEEE Communications Magazine, vol.49, no.6, pp.101-107, Jun. 2011.

11 Cheol Jeong et al., “Random Access in Millimeter-Wave Beamforming Cellular Networks: Issues and Approaches,” IEEE Communications Magazine, vol. 53, no. 1, pp.180-185, Jan. 2015.

12 Wonil Roh et al., “Millimeter-wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results,” IEEE Communications Magazine, vol. 52, no. 2, pp. 106-113, Feb. 2014.

13 Samsung, “Technologies for Rel-12 and Onwards,” RWS-120021, 3GPP TSG RAN Workshop on Rel-12 and Onwards, Jun. 2012.

14 Taeyoung Kim et al., “Tens of Gbps Support with mmWave Beamforming Systems for Next Generation Communications,” in Proc. IEEE GLOBECOM’13 Workshop, pp.3685-3690, Dec. 2013.

15 Chanhong Kim et al., “Multi-Beam Transmission Diversity with Hybrid Beamforming for MIMO-OFDM Systems,” in Proc. IEEE GLOBECOM’13 Workshop, pp. 61-65, Dec. 2013.

16 WiGig Alliance, “Defining the Future of Multi- Gigabit Wireless Communications,” WiGig White Paper, Jul. 2010.

17 R12-WP5D-C-0258, “Technical Feasibility of IMT in the Bands above 6 GHz,” ITU, Jan. 2013.

18 Theodore S. Rappaport et al., “Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!,” IEEE Access, vol.1, pp.335-349, May 2013.

19 Theodore S. Rappaport et al., “28GHz Propagation Measurements for Outdoor Cellular Communications Using Steerable Beam Antennas in New York City,” in Proc. IEEE ICC’13, pp.5143-5147, Jun. 2013.

20 Alper Yegin et al., “Terminal-centric Distribution and Orchestration of IP Mobility for 5G Networks,” IEEE Communications Magazine, vol. 52, no. 11, pp.86-92, Nov. 2014.

21 Young-Han Nam et al., “Full Dimension MIMO (FD-MIMO) for Next Generation Cellular Technology,” IEEE Communications Magazine, vol. 51, no. 6, pp.172-179, Jun. 2013.

22 Maurice Bellanger et al., “FBMC Physical Layer: a Primer,” PHYDYAS Tech. Rep., Jun. 2010.

23 Sungnam Hong et al., “A Modulation Technique for Active Interference Design under Downlink Cellular OFDMA Networks,” in Proc. IEEE WCNC’14, pp.683-688, Apr. 2014.

24 Sungnam Hong et al., “Frequency and Quadrature-Amplitude Modulation for Downlink Cellular OFDMA Networks,” IEEE Journal on Selected Areas in Communication, vol. 32, no. 6, pp.1256-1267, Jun. 2014.

Samsung Demonstrates World's First 5G Data Transmission at Highway Speeds

Record-breaking 1.2Gbps data transmission at over 100km/h, and 7.5Gbps in stationary conditions using 28GHz spectrum

SEOUL, Korea – October 15, 2014 – Samsung Electronics Co., Ltd today announced the record-breaking demonstration of super-fast data transmission targeted for fifth generation (5G) mobile networks using 28 GHz spectrum. Samsung researchers confirmed the world's first data rate of 1.2 Gbps, or 150 MB per second on a vehicle cruising at over 100 km/h. This marks a significant step towards the utilization of millimeter wave frequency bands for 5G mobile networks

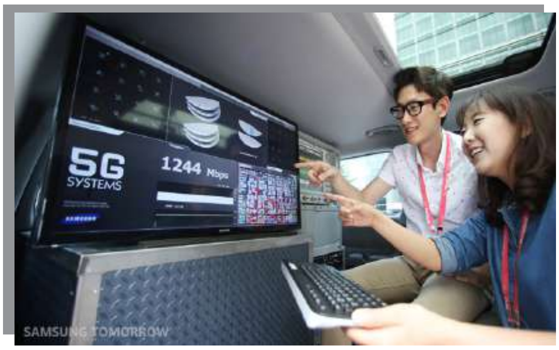

Figure A Achieved Data Rate in Mobile Condition (1.2Gbps)

The achievement was bolstered by another record-breaking demonstration in which Samsung achieved data transmission speed up to 7.5 Gbps, or 940 MB per second when the vehicle came to a complete stop. The peak data rate is more than thirty times faster compared to the state-of-the-art 4G LTE technology.

Back in May, 2013, Samsung revealed the world's first 28 GHz based 5G data transmission speed of 1Gbps at pedestrian speeds. Since then researches at Samsung were able to increase the maximum data rate by more than 7-fold and support mobility up to highway speeds. The continued ground breaking success underlines the company's leadership in next-generation mobile communications.

"The expectations for 5G communications will continue to grow based on the rising demands for smart devices, cloud services, smart home technology, and Internet of Things" , said ChangYeong Kim, Head of DMC R&D Center of Samsung Electronics. "We are committed to developing innovative 5G technologies and will continue to utilize our exceptional R&D capabilities as well as our diverse partnerships with leading companies and research centers around the world".

While 5G standard has yet to be ratified, 5G networks are expected to feature data rates and capacity that are hundreds times larger compared to 4G LTE.

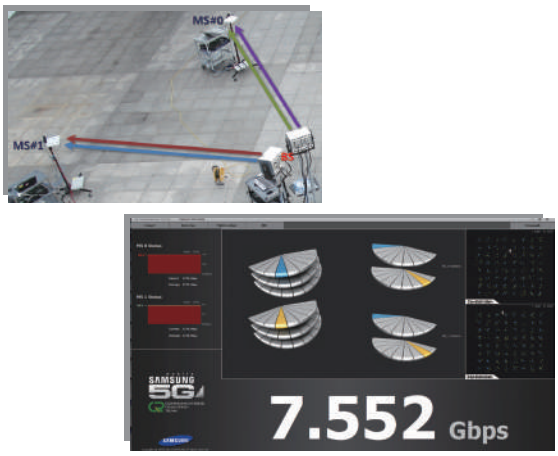

Figure B Achieved Data Rate in Stationary Condition (7.5Gbps)

In order to achieve such high data rates, 5G networks will inevitably exploit frequencies much higher and less congested than those of which are currently used for cellular networks (typically under 3 GHz). However, difficulties such as large propagation loss at these frequency bands have prohibited the industry from utilizing these bands for cellular applications. To address these challenges, Samsung has applied 'Hybrid Adaptive Array Technology' at 28 GHz frequency bands.

In addition, Samsung has revealed its '5G Rainbow' – seven technical requirements which are pillars to ensure a truly differentiating 5G user experience. The list comprises of maximum data rate, spectral efficiency, speed of mobility, data transmission rate at the cell boundary, the number of simultaneous connections, communication delays, and cost. To meet these requirements, Samsung has been developing key technologies including transmission technologies for high frequency bands, multiple access schemes and low latency networks among others.

The company is committed to developing innovative technologies that will enable the 5G era and plans to demonstrate its capabilities at the upcoming 2018 PyeongChang and 2020 Tokyo Olympics.

Australian Government
Department of Communications and the Arts

5G—Enabling the future economy

October 2017

Disclaimer

The material in this paper is of a general nature and should not be regarded as legal advice or relied on for assistance in any particular circumstance or emergency situation. In any important matter, you should seek appropriate independent professional advice in relation to your own circumstances.

The Commonwealth accepts no responsibility or liability for any damage, loss or expense incurred as a result of the reliance on information contained in this paper.

This paper has been prepared for consultation purposes only and does not indicate the Commonwealth's commitment to a particular course of action. Additionally, any third party views or recommendations included in this paper do not reflect the views of the Commonwealth, or indicate its commitment to a particular course of action.

Copyright

© Commonwealth of Australia 2017

The material in this paper is licensed under a Creative Commons Attribution—4.0 Australia license, with the exception of:

- the Commonwealth Coat of Arms
- this Department's logo
- any third party material
- any material protected by a trademark, and
- any images and/or photographs.

More information on this CC BY license is set out at the creative commons website: https://creativecommons.org/licenses/by/4.0/. Enquiries about this license and any use of this paper can be sent to: Department of Communications and the Arts, GPO Box 2154, Canberra, ACT, 2601.

Attribution

Use of all or part of this paper must include the following attribution:

© Commonwealth of Australia 2017

Using the Commonwealth Coat of Arms

The terms of use for the Coat of Arms are available from the It's an Honour website (see www.itsanhonour.gov.au and click 'Commonwealth Coat of Arms').

Contents

Australia's 5G vision **1**

What is 5G? **3**

The 5G economics case **3**

Enhanced Mobile Broadband (eMBB) 3

Massive machine type communications (mMTC) 4

Critical communications 4

What is different about 5G? **6**

Network slicing 6

Mesh networks 6

Spectrum sharing 7

Antenna technology and network topology 7

Business case 8

How are countries preparing for 5G? **9**

Industry 5G preparations in Australia 9

The Government's direction for 5G **10**

Making spectrum available in a timely manner 10

Actively engaging in international spectrum harmonisation activities 11

Streamlining arrangements to allow mobile carriers to deploy infrastructure more quickly 11

Reviewing existing telecommunications regulatory arrangements to ensure they are fit-for-purpose 12

Next steps **12**

Australia's 5G vision

5G is the next step in the evolution of mobile wireless communications technology, promising improved connectivity, greater network speeds and bandwidth, and very low latency. It is the fifth generation in mobile technology which, at each step, has seen significant developments in communications networks:

- 1G—The first generation of mobile phone networks were deployed in the early 1980s, providing a basic voice service using analogue transmission.
- 2G—In 1991, second generation networks were deployed, making the switch to digital standards with improved voice messaging and the introduction of the short message service (SMS).
- 3G—The third generation launched in 2001 and introduced data services in addition to voice and SMS.
- 4G—In 2009, the fourth generation protocol, Long Term Evolution (LTE), was introduced, supporting improved mobile broadband which saw increased capacity and speed for data.

New capabilities of mobile communications networks enabled by 5G technology will allow for a variety of 'use cases':

- higher quality and more video services provided to multiple users with full mobility, even at high speed
- massive scale automation delivered through widespread sensor networks and multiple connected devices
- delivery of critical communications assured by low latency and ultra-reliable networks, and
- improved productivity assisted by high quality, real time data analytics.

Unlike existing mobile communications networks, 5G networks have the potential to allow tailoring of requirements for each of these different use cases within the same network.

The Government considers that 5G is more than an incremental change for mobile communications. Instead, it provides the underlying architecture that will enable the next wave of productivity and innovation across different sectors of the Australian economy. Efficient rollout of 5G and uptake of the services it supports has the potential to produce far-reaching economic and social benefits and support growth of Australia's digital economy. This will be supported by the rollout of the National Broadband Network (NBN) allowing greater capabilities for the seamless delivery of services across high speed mobile, fixed line and fixed wireless networks.

The Government wants to create an environment that allows Australia's telecommunications industry to be at the forefront of seizing the benefits of 5G across the economy. The communications sector will lead the rollout of 5G networks in Australia. However, the Government can create the policy and regulatory environment to support a more efficient rollout, given its potential benefits to the economy.

The Government's direction will be to support the timely rollout of 5G in Australia to enable the next wave of broad-based industry productivity, and support the growth of Australia's digital economy.

This includes immediate actions by Government that enable the communications market to introduce new 5G technologies in line with international developments. These include:

- making spectrum available in a timely manner
- actively engaging in the international standardisation process
- streamlining arrangements to allow mobile carriers to deploy infrastructure more quickly, and
- reviewing existing telecommunications regulatory arrangements to ensure they are fit-for-purpose.

The Government recognises that as 5G continues to develop, other issues relating to the technology will likely emerge which may require future Government action. In particular, while there are opportunities for 5G to create economy-wide transformation, this will require a broader examination of sectoral regulatory frameworks.

To that end, the Government will work collaboratively with industry to foster an ongoing dialogue on 5G beyond the launch of this paper to identify and remove sectoral barriers to its successful and timely rollout. Through this dialogue, the Government will also look at opportunities to build on other Government activities, such as the national Digital Economy Strategy which will more broadly focus on building the productivity of sectors across the economy.

What is 5G?

The International Telecommunication Union (ITU) is the United Nations specialised agency for information and communications technologies. This body decides global spectrum allocation frameworks and harmonises international spectrum to ensure networks and connected devices can communicate seamlessly. The ITU will undertake the formal, international process to identify bands for 5G by 2020. It has developed draft technical specifications for 5G which include:[1]

- high data rates (1 Gbps for hotspots, 100 Mbps download and 50 Mbps upload for wide-area coverage)
- massive connectivity (1 million connections per square kilometre)
- ultra-low latency (1 millisecond)
- high reliability (99.999% for mission critical 'ultra-reliable' communications), and
- mobility at high speeds (up to 500 km/h i.e. high speed trains).

In working towards these specifications, 5G represents a significant leap from the capabilities of previous generations and introduces a range of new technological possibilities. The success of 5G in delivering new technologies and services will be supported by existing communications infrastructure, including the NBN. This convergence of high-speed fixed-line and mobile services will collectively produce a consistent and ubiquitous user experience.

The 5G economics case

Unlike early generations of mobile networks, 5G will represent a significant shift in the telecommunications industry's focus away from voice and more towards mobile broadband and increased industrial applications. These new use cases are expected to create benefits across a range of sectors—including transportation, health, manufacturing and agriculture—and have varying networking requirements. These use cases, as identified by the industry, can be divided into the following categories:

- enhanced Mobile Broadband
- massive Machine Type Communications, and
- critical communications.

Enhanced Mobile Broadband (eMBB)

eMBB will deliver improved capacity to a greater number of devices. This will enable higher rates and volumes of data transmission per device and improve coverage to a broader range of locations. eMBB will likely be the focus of early 5G deployments as it can immediately support the growing communications requirements for the digital economy.

An improved mobile experience for consumers

5G networks will give consumers a better mobile experience in more locations. Increased network capacity will support more users, even in crowded areas, such as large public events, and at peak times. Faster network speeds will also enable consumers to view rich content in more places, supporting the streaming of live events and high resolution media.

---

[1] ITU, 23 February 2017, 'Press Release: ITU agrees on key 5G performance requirements for IMT-2020'.

Massive machine type communications (mMTC)

As 5G networks mature, they will support the widespread and dense deployment of sensors and other network-connected devices by significantly reducing their power requirements and providing flexible coverage across different spectrum bands. This proliferation of the Internet of Things (IoT) across industries is expected to produce significant productivity benefits and support integration between sectors.

Supporting productivity and innovation

The term Industry 4.0 describes the next step in the advancement of the manufacturing sector (the ‘fourth industrial revolution’). Industry 4.0 introduces autonomous systems supported by a combination of technologies such as IoT, artificial intelligence, continued technological improvements and digitalisation in manufacturing.

Australia stands to benefit from Industry 4.0, given our world-class manufacturing sector, which includes several high-value industries such as medical technology and aerospace. Australian manufacturers can improve their productivity and international competitiveness through Industry 4.0 processes by supporting their participation in global value chains. This is of particular benefit to SME manufacturers, opening them up to new markets and opportunities. 5G can support Industry 4.0, by providing communications infrastructure that is more accessible and flexible to suit specific industry needs.

5G can enable innovation in other sectors such as agriculture. A challenge for Australia’s agricultural sector is identifying how to improve productivity while balancing environmental and commercial constraints. Precision agriculture, which focuses on improving yields and minimising economic risks, seeks to provide more control in the management of agricultural production. While precision agriculture requires a range of enablers—including data analysis, sensor networks and geographical information systems—5G can provide the supporting infrastructure for these technologies.

Critical communications

Low latency and ‘ultra-reliable’ communications networks will support the delivery of critical communications, i.e. to support public safety use and playing role in the technology ecosystem supporting autonomous vehicles. In addition to automation, critical communications will also help to support technological advancement in areas including robotics and artificial intelligence.

The social benefits from autonomous vehicles

5G networks are expected to play a role in the technology ecosystem supporting the development of autonomous vehicles, which will enable a number of social benefits for transportation. Traffic congestion, which is estimated to cost Australia $53 billion by 2031,[2] could be proactively reduced by smart city traffic management systems that are informed by machine-to-machine communications with autonomous vehicles.

Improved road safety is also expected to be a key outcome of autonomous vehicles, as the majority of car accidents involve human error. In the 12 months to July 2017, there were 1,235 deaths on Australian roads with road trauma costing the Australian community an estimated $27 billion annually.[3] Autonomous vehicles can have a valuable role not just in terms of financial savings, but in saving human lives.

---

[2] Infrastructure Australia, 2015, ‘Australian Infrastructure Audit’.

[3] Bureau of Infrastructure, Transport and Regional Economics, August 2017, ‘Road deaths—monthly bulletins’.

Figure 1—Outline of the relationship between the technical requirements of 5G and expected services and applications it will deliver

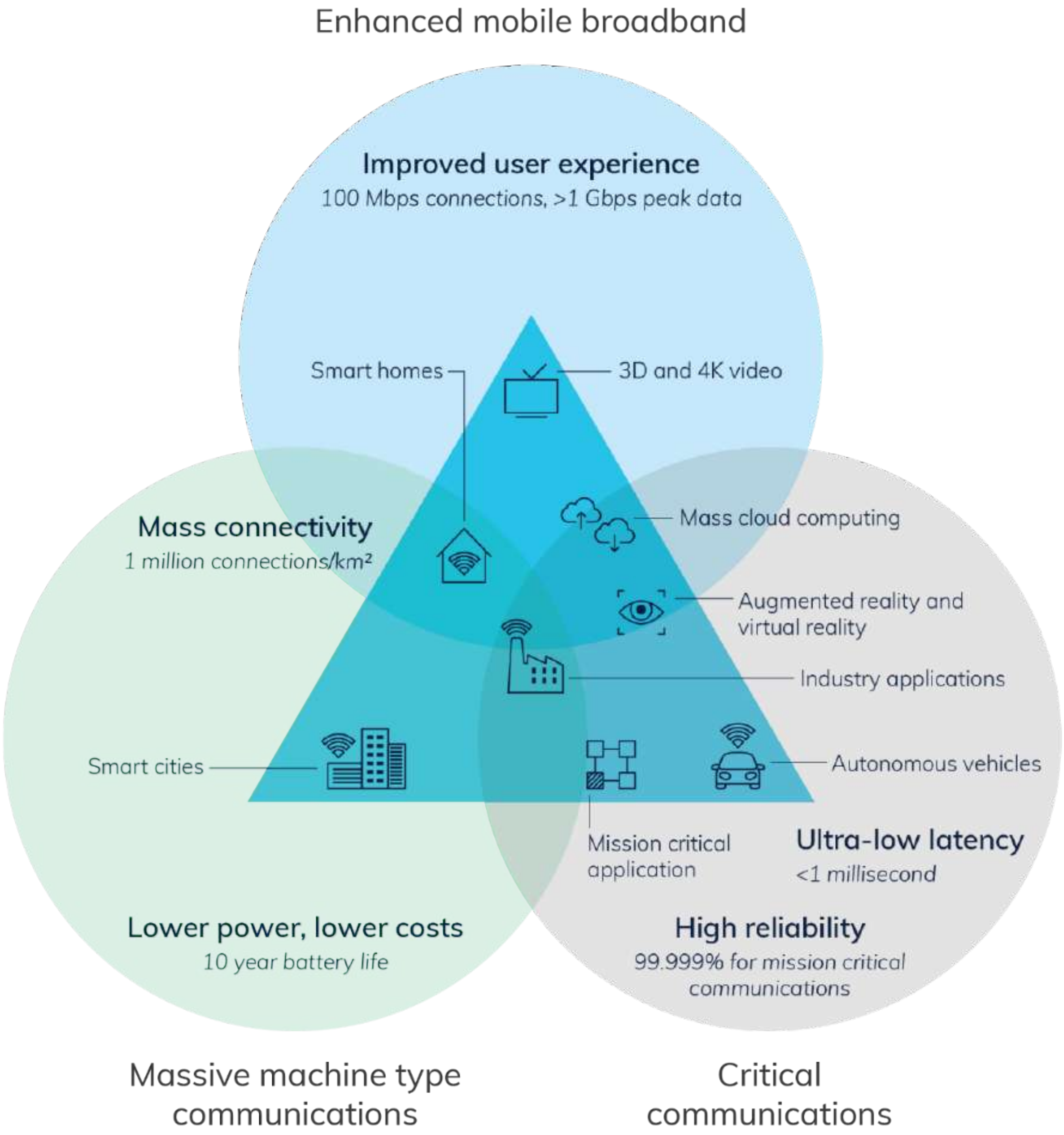
Enhanced mobile broadband
Improved user experience
100 Mbps connections, >1 Gbps peak data
Smart homes
3D and 4K video
Mass cloud computing
Mass connectivity
1 million connections/km²
Augmented reality and virtual reality
Industry applications
Smart cities
Autonomous vehicles
Mission critical application
Ultra-low latency
<1 millisecond
Lower power, lower costs
10 year battery life
High reliability
99.999% for mission critical communications
Massive machine type communications
Critical communications

Source: Based on figure 2 from the ITU paper, '[IMT Vision—Framework and overall objectives of the future development of IMT for 2020 and beyond](#).'

What is different about 5G?

5G is expected to enable productivity outcomes across key verticals of the economy as a result of a range of characteristics, such as **network slicing** and **mesh networks**. However, it will need to overcome deployment challenges such as **spectrum acquisition** and **coverage**. This is expected to lead to new approaches for deployment to enable widespread 5G coverage in Australia. The industry will also need to develop a robust **business case** for 5G rollout, which will likely be demand driven, to support the significant capital expenditure required.

Network slicing

Network slicing allows operators to split their network into separate sub-networks (also referred to as slices), enabling them to dedicate network resources to different users and applications. Sub-networks can 'slice' the infrastructure resources from the physical network to create virtual independent networks. This is a significant development from previous mobile network generations, as it enables an operator to deliver many different capabilities by creating slices that can be tailored for the intended usage. For example, an operator could create a network slice for IoT devices, or alternatively, a network slice for higher security for a higher quality of service for government or public safety uses.

Network slicing also provides a model for infrastructure sharing between network operators. In this case, a single transmitter's network could be shared with more than one operator. Given the capital expenditure that will be associated with 5G rollout, network slicing can provide a cost effective, short-term solution for operators.

Figure 2—5G network slicing enables operators to create separate virtual networks to meet application requirements

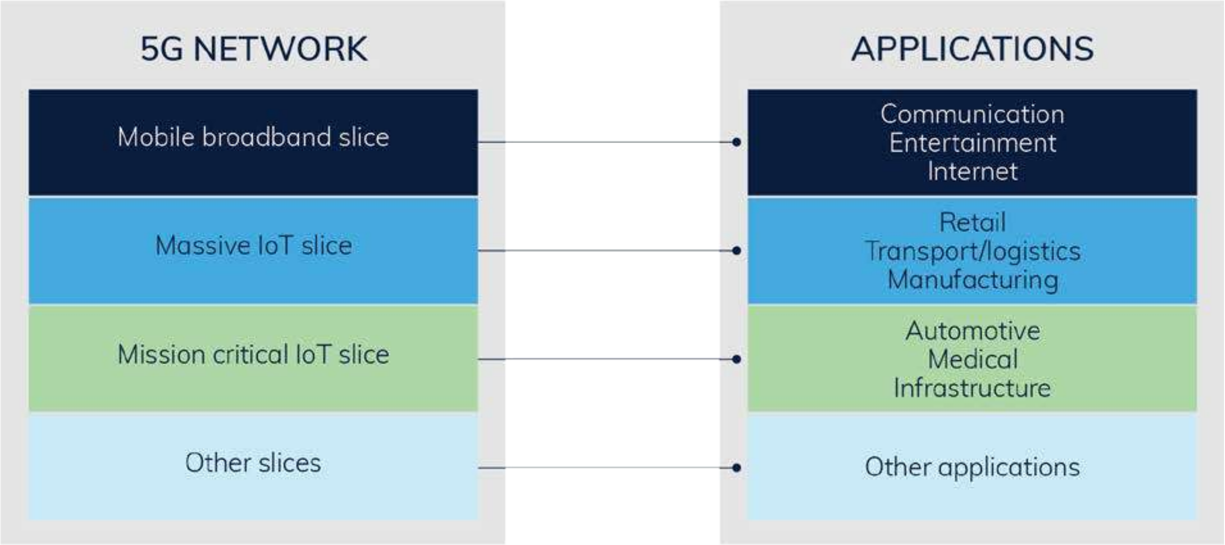

Source: Based on figure 2 from the ITU article, 'Why end-to-end network slicing will be important for 5G.'

Mesh networks

Mesh networks can be utilised to increase the range of coverage, where the 'mesh' is an interconnection among a network of devices. Only one device in the mesh needs to be connected to the network, which can then relay data to other nearby devices.

Mesh networks not only provide the opportunity to support 5G deployment but also have the additional benefit of providing efficient network speeds. Through dynamic routing, devices on a mesh network are able to seek the fastest and most reliable pathway to send and receive data. As such, this architecture can provide a cost effective solution for coverage in more remote areas. For example, primary industries would benefit from this network approach by simplifying network connections and costs associated with deploying and managing an IoT sensor network. However, the application of mesh networks is still highly speculative as the industry considers how they will operate in practice and in different environments. This may contribute to the development of new business models for 5G.

Spectrum sharing

Spectrum is a critical enabler of Australia's current and future communications infrastructure. The specific spectrum bands and quantity of spectrum required for 5G are still being considered. However, 5G will likely require a mix of low, medium and high frequency spectrum to meet different scenarios relating to coverage, connectivity, and latency:

- Low frequency (less than 1GHz)**—**providing widespread coverage across urban, suburban and rural areas and supporting IoT for low data rate applications.
- Medium frequency (1–6GHz)**—**providing good coverage and high speeds, and including the expected initial 5G range of 3.3–3.8GHz which has been identified as the most likely band for launching 5G globally.
- High frequency (above 6GHz)**—**providing ultra-high broadband speeds for advanced mobile broadband applications, and most suitable for applications in dense traffic hotspots.

5G technologies can be expected to deliver improvements in spectral efficiency (the data rate that can be supported per unit of spectrum). However, the use of 5G networks for applications such as widespread industrial applications is likely to require significantly more amounts of contiguous spectrum to be made available.

While some of the potential bands for 5G currently have unused spectrum, other bands would need to be 'refarmed', noting that it is likely that bands currently used for 2G, 3G and 4G in Australia will transition over time to 5G. Refarming enables spectrum to be transitioned to the highest value use as required. The refarming of spectrum already held by mobile broadband operators is a commercial decision for those operators. 5G is also expected to provide the opportunity for 'soft-refarming' where 4G and 5G technologies can both be supported simultaneously, minimising the impact to legacy devices during transition periods.

Spectrum sharing, that is spectrum accessed by numerous users on a shared basis, has also been identified as an option for 5G technology. Spectrum sharing encompasses a range of different aspects of spectrum management. Spectrum can be shared by geography, time, economic priority schemes, code modulation, polarisation, directionality or power. Access to spectrum is divided between users so it can be used without interference issues.

5G opens up new opportunities for increased spectrum sharing, through mechanisms such as network slicing. 5G technology is also designed to support shared arrangements, and allows for the sharing of the same spectrum ('unlicensed Wi-Fi' spectrum) with other technologies. Operators can augment their holdings in situations where existing exclusive holdings are insufficient to meet customer needs. Spectrum sharing in the 5G context is also supported by the expected use of highly directional antenna technology which would enable operators to operate in closer proximity without interference.

Antenna technology and network topology

5G will require radically different structures of networks if it is to achieve successful deployment in Australia. As 5G will likely utilise different frequencies, new equipment will be necessary. Additionally, the higher frequency 5G spectrum can only travel a small distance and will need more cells to ensure adequate coverage. However, antennas and equipment will be smaller, making it easier to attach these cells to existing infrastructure such as street lights and buildings.

The more dense deployment of cells will also give rise to other approaches that improve the reliability of data transmission across a 5G network. Data may be divided into individual streams and transmitted through multiple antenna segments in a process called Multiple Input Multiple Output (MIMO) which allows for more information to be transmitted simultaneously. This technique is further empowered by 'beamforming' which allows base stations to direct focused beams of energy to a specific area rather than dissipate the available power of a larger area. These developments will enable more efficient transmission and increase overall throughput.

5G is also expected to assist the adoption of IoT by further reducing power consumption through extended discontinuous reception. In this scenario, IoT devices shift between active and inactive cycles, transmitting only when required. This will allow connected devices to operate for extended periods on a single charge, reducing operational costs.

5G will also increase the support for a greater density of devices that would have otherwise been limited by the capacity on 3G and 4G networks. They will also enhance IoT deployment through the use of network slicing to create virtual network configurations that are optimised for the low power and coverage requirements of IoT networks.

Business case

For Australians to experience the benefits of 5G, the communications sector will need to explore and develop new business cases to attract investment and support the rollout of 5G services. As 5G becomes an integral part of the communications ecosystem, the sector will need to be agile to respond to the needs and expectations of other sectors which will be seeking to take advantage of these next generation networks.

While residential consumers will inevitably be attracted to the enhanced mobile broadband services offered by 5G, it is the industrial applications for 5G where industry expects to see the greatest opportunity for new business models.

Compelling industrial applications are still to be developed. For example, industry sectors can be expected to seek tailored solutions for their business needs such as enhancing their local area network or enabling autonomous systems, and small businesses will look for low cost deployments of IoT.

It is therefore expected that the model for 5G will be demand driven and will require the communications industry to foster new business opportunities.

How are countries preparing for 5G?

In preparation for 5G, many countries have been taking steps to test the technology and review their spectrum arrangements.

Examples of 5G work being undertaken by other countries

The United States is clearing the 600MHz band through an incentive auction for the potential early deployment of 5G. The United States has also identified reforms to infrastructure deployment as a priority for 5G rollout. Mobile network operators are also conducting trials of low-band spectrum for use in 5G services.

The Asia-Pacific Telecommunity (APT), the regional arm of the ITU, approved a recommendation for the use of the 700MHz band for 5G, with 26 countries in the Asia-Pacific region identifying this band for this use, including Australia, Japan, South Korea and New Zealand.

Korea and Japan have stated their intention to use some or all of the 26.5 to 29.5GHz range to trial enhanced mobile broadband applications ahead of the 2018 Winter and 2020 Summer Olympics respectively. It is expected that these trials will lead to commercial availability of 5G services.

The European Communications Commission (ECC) identified the 3.4–3.8 GHz and 26GHz band (from 24.25 to 27.5GHz) bands for the deployment of 5G in Europe. All European countries are expected to select at least part of this range to launch 5G by 2020.

In the United Kingdom, the government launched its 5G strategy in March 2017. It is funding testbeds to understand the different deployment requirements and security considerations for 5G. The United Kingdom is also working to make suitable spectrum available in the high (24.25 GHz–27.5 GHz, and other bands above 30 GHz), medium (3.4–3.8 GHz) and low frequency (700 MHz) bands.

Industry 5G preparations in Australia

Australia is well positioned to harness the opportunities of 5G. Australia has an effective and competitive mobile communications market, with voice and data coverage available to more than 99 per cent of the population.[4] It is the top performer internationally in terms of having in place effective enablers—infrastructure, affordability, consumer readiness and content availability—to support mobile internet adoption.[5]

5G trials have already commenced in Australia, with each of the main carriers working with mobile equipment suppliers in testing the application and limits of the technology. These trials will continue and will inform the communications sector on how 5G can be effectively deployed for the Australian environment.

---

[4] Department of Communications and the Arts, accessed 1 September 2017, 'Mobile phone towers'.

[5] GSM Association, 24 June 2016, 'Global Mobile Connectivity Index'.

The Government's direction for 5G

The Government has made significant investments to improve telecommunications infrastructure in Australia, through the NBN and the Mobile Black Spot Program. The Government is also working to create a policy and regulatory environment that supports a competitive and innovative communications market.

The Government recognises that 5G will enable innovation and productivity across industry sectors and can significantly contribute to Australian's growth and future prosperity. Therefore, the Government will focus on enabling the early deployment of this new generation of mobile networks in Australia and encourage its use in delivering new services and applications.

The Government will support the timely rollout of 5G in Australia to enable the next wave of broad-based industry productivity, and support the growth of Australia's digital economy.

Industry expects and needs to lead the deployment of 5G. However, the Government has a role in supporting network rollout by modernising policy and regulatory frameworks and removing barriers that would delay rollout and adoption unnecessarily.

In the first instance, the Government will support the early deployment of 5G in Australia by:

- making spectrum available in a timely manner
- actively engaging in international spectrum harmonisation activities
- streamlining arrangements to allow mobile carriers to deploy infrastructure more quickly, and
- reviewing existing telecommunications regulatory arrangements to ensure they are fit-for-purpose.

Making spectrum available in a timely manner

A clear, efficient and flexible regulatory framework governing spectrum access will be essential to support the timely deployment of 5G networks in Australia.

The Government is currently undertaking work to modernise Australia's spectrum management framework to ensure it remains fit-for-purpose. In May 2017, it outlined its proposed reforms to the framework which are designed to simplify and streamline the processes for spectrum allocation and provide a transparent, efficient and flexible spectrum management framework. This will be the most significant change to the Australian spectrum management framework in the last 25 years.

The reforms will remove barriers between licence types, and enable flexible licensing issue and allocation processes. This strategic approach will remove outdated processes and support the Australian Communications and Media Authority (the ACMA) to more effectively respond to market demands and new technologies, such as 5G. This will help Australia remain internationally competitive with a modern, innovative economy over the coming decades.

The Government will put in place its new spectrum management framework by 2019.

In addition, the ACMA will continue to work on making spectrum available for 5G. The ACMA has been investigating the use of 1.5GHz and 3.6GHz and high frequency mmWave bands in considering additional spectrum for mobile broadband services. The ACMA has decided to prioritise refarming of the 3.6GHz band over the 1.5GHz band, citing industry submissions noting this band is likely to be a pioneer band for early 5G deployments and the need to provide greater clarity and investment certainty for incumbents and potential new band entrants alike. The ACMA is currently engaging with industry on which parts of the 3.6GHz band should be reallocated and on what terms. This approach also follows international trends which have seen the 3.6GHz band commonly used for 5G trials.

The ACMA will work to bring 3.6GHz spectrum to auction in 2018.

Actively engaging in international spectrum harmonisation activities

There is already significant work underway globally with several countries trialling 5G, but standards for this new generation technology are yet to be finalised. The formal, international process to define 5G is led by the ITU. The ITU's Working Party 5D is responsible for shaping the standard for "futuristic mobile technologies" to support International Mobile Telecommunications (IMT) for 2020 and beyond.[6] This process is known as IMT-2020.

Stakeholders such as regulatory and policy setting bodies, hardware manufacturers and governments of countries in which they are based will be seeking to influence the international dialogue. A key body is the industry-driven 3GPP which undertakes technology standardisation. Collectively, the ITU and 3GPP will drive spectrum harmonisation activities: the ITU, led by administrations, will focus on the spectrum requirements; and the 3GPP, led by industry, will concentrate on equipment and device standards.

Industry is well-placed to lead the standardisation process particularly given its role identifying the application of 5G technologies. However, there is also an important role for the Government in these processes. This is particularly the case in the harmonisation of international spectrum arrangements which will have significant impact on the availability and cost of 5G devices in Australia and can be strongly contested.

In the past, Australia and the Asia-Pacific region have been influential in contributing to standards and spectrum plans that have been adopted across the world. This ensured that Australia was able to adopt new technologies quickly and that the Australian market could take advantage of the economies of scale and have a greater choice of mobile handset equipment.

The Government will ensure strong participation by Australian in domestic and international discussions about 5G spectrum harmonisation. Our continued involvement provides the opportunity to contribute to this dialogue and secure outcomes that will benefit the adoption of 5G in Australia.

Streamlining arrangements to allow mobile carriers to deploy infrastructure more quickly

The design and deployment of mobile networks will be radically different from those of today. 5G is expected to require additional infrastructure in new forms, including smaller cells and more densely-located antennas, particularly in the use of high-band spectrum.

Carriers have specific powers and immunities relating to telecommunications infrastructure deployment and installation. These laws help carriers to rollout telecommunications infrastructure quickly in a nationally-uniform way, rather than having to follow state, territory and local government requirements. These laws have existed in their current form since 1997.

The Government has recently consulted on proposed new arrangements that take account of technology developments and changes in operating practices as well as identifying opportunities to streamline deployment processes. Under these new proposed arrangements, mobile carriers would be able to rollout new communications technologies such as 5G more efficiently.

The Government understands that some members of the community have expressed concerns about the impact on public amenity from increased telecommunications infrastructure. The Government considers that telecommunications providers should work with local communities to address concerns about their infrastructure plans and is encouraging industry to consider consultation requirements for future 5G networks.

The Government continues to work with stakeholders and will implement the first tranche of changes to carrier powers and immunities following the conclusion of the consultation process.

---

[6] ITU, accessed 1 September, ITU towards "IMT for 2020 and beyond."

Reviewing existing telecommunications regulatory arrangements to ensure they are fit-for-purpose

The pace of change in the communications sector is impacting on the effectiveness of the existing telecommunications regulatory framework. In response, the Government is progressively working to modernise the regulatory architecture to ensure that the regulatory and policy settings flexibly respond to the current and future needs of the communications sector.

In undertaking reform, the Government is cognisant that the communications regulatory framework will need to be sufficiently flexible to address the emergence of new technologies and business models.

5G deployment will benefit from the Government's reform efforts, including updates to the regulatory framework for telecommunications and radio spectrum and modernising the ACMA. In these cases, the Government has been revising regulation to shift it away from the traditional vertical telecommunications sectors to principles based, flexible arrangements. This will provide the versatility necessary to account for 5G developments and services.

The Government will continue to work with industry to modernise current telecommunications regulatory arrangements to ensure they encourage competition and innovation in the sector.

As with previous mobile networks, cybersecurity will be a critical consideration as 5G is deployed. Security will be even more so a challenge for 5G, as the reliability of communications will be pivotal in the technology's ability to deliver benefits, particularly in the case of critical communications.

Additionally, in providing the architecture for automation, 5G networks will trigger an ever-increasing volume of data. While the bulk of data will be machine-to-machine communications, users will want assurance that their personal information is protected. User consent will be an area of growing complexity due to the intersection between autonomous systems and the individual.

Industry has strong incentives to address cybersecurity risks in 5G's new types of network deployments and systems. However, **the Government will continue to assess cybersecurity and privacy issues as they evolve to ensure Australians have confidence in using 5G.**

Next steps

The Government recognises the opportunities presented by 5G for economy-wide transformation, creating productivity benefits in sectors such as transportation, health, manufacturing and agriculture. However, realising the benefits of 5G in sectors other than communications will need the right sectoral regulatory settings.

The Government will work to ensure that sectoral regulatory frameworks are updated to take advantage of 5G. The communications portfolio is well-placed to facilitate an ongoing strategic dialogue on 5G that will support sectors to identify and work towards unlocking the potential benefits of the technology. This dialogue would provide a starting point for greater engagement across Government and with industry and the community.

To that end, **the Government will establish a 5G working group that will bring together representatives from across Government and industry.** The working group will create a platform for this strategic dialogue with a mandate to seek out opportunities and emerging issues on 5G. This will provide better coverage across Government of the evolving policy and regulatory challenges associated with 5G.

Boise State University
ScholarWorks

Electrical and Computer Engineering Faculty Publications and Presentations

Department of Electrical and Computer Engineering

1-1-2016

Energy Efficient Switched Parasitic Array Antenna for 5G Networks and IoT

Ahmed Kausar
Boise State University

Hani Mehrpouyan
Boise State University

Mathini Sellathurai
Boise State University

Rongrong Qian
Heriot Watt University

Shafaq Kausar
National University of Sciences and Technology

© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. doi: 10.1109/LAPC.2016.7807569

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

Energy Efficient Switched Parasitic Array Antenna for 5G Networks and IoT

Ahmed Kausar*, Hani Mehrpouyan *, Mathini Sellathurai†, Rongrong Qian†, and Shafaq Kausar‡
*Department of Electrical and Computer Engineering
Boise State University, Boise, USA
Email: ahmedkausar@boisestate.edu, hanimehrpouyan@boisestate.edu
† School of Engineering and Physical Sciences
Heriot Watt University, Edinburgh, U.K
Email: m.sellathurai@hw.ac.uk, rq31@hw.ac.uk
‡Department of Electrical Engineering
National University of Sciences and Technology, Islamabad, Pakistan
Email: shafaqkausar@nust.edu.pk

***Abstract*—This paper includes design and implementation result of an adaptive beam forming antenna for upcoming 5G and Internet of Things (IoT). Switched parasitic array antennas are low cost, small sized and compact circular array antennas that steer beam in a desired direction by variation in switching pattern of parasitic elements. The proposed antenna design has an active center element, which is surrounded by several symmetrically placed parasitic elements. The designed antenna has a gain of 8 dB and is capable of 360 degrees beam steering in steps of 60 degrees each. Simulations are validated with results of the fabricated antenna. Antenna beam is steered by controlling parasitic elements. Future application of Electronically Steerable Parasitic Array Radiator (ESPAR) antennas and switched parasitic array antennas in next generation communication networks and methods for reducing size of the antenna are also highlighted.**

***Index Terms*—Internet of Things, 5G, circular array, reconfigurable, smart antenna, adaptive beamforming, Switched-Parasitic Array**

I. INTRODUCTION

Smart antennas are capable of adaptive beam forming in accordance with the environment. Over the last three decades wireless communication has made rapid progress due to better modulation schemes and better error correction schemes. Wireless communication systems are replacing the wired ones. Antenna is an integral part of wireless devices. To meet the ever-increasing demand of higher data rate, smart antennas capable of adaptive beam forming are introduced as a potential solution to increasing data rate and improving signal-to-noise ratio (SNR) in upcoming 5G and Internet of Things (IoT) [1]. By using smart antennas Equivalent Isotropically Radiated power (EIRP) can be increased. EIRP denotes product of the transmit power (P_t) and transmit antenna gain (G_t), i.e.

$$EIRP = P_t \times G_t \tag{1}$$

From (1), it is observed that by using a smart antenna G_t increases, which in turn leads to the increase in the EIRP. The beam of an antenna can be steered by means of mechanical movement of a directional antenna or by electronic beam steering. Electronic beam steering is resilient and provides more accurate approximation of the desired signal source. Moreover electronic beam steering antenna does not involve moving parts. Therefore, there is less wear and tear on the antenna as compared to mechanical beam steering. Generally electronic beam forming is achieved by using phase shifters coupled with an array of antenna elements. The phase of an individual phase shifter is varied such that there is net constructive interference in the intended direction of the antenna beam and destructive interference in all other directions [2]. Hence, electronic beam steering is preferred over mechanical beam steering.

Electronically Steerable Parasitic Array Radiator (ESPAR) antennas use mutual coupling between antenna elements to steer the beams instead of phase shifters. Design of parastic array radiator antennas is discussed in [3], [4] in which there is a center active element surrounded by a circular array of passive elements, where each passive element is loaded with a specific reactance. ESPAR antennas are smaller in size as compared to phase array antennas because elements have to be placed at a quarter of wavelength or less distance from each other for effective mutual coupling. Hence, ESPAR antennas are cost effective and space efficient. In such antennas we have only one active element, therefore, only one feed is required. Consequently, ESPAR antennas have lower losses (due to single feed) as compared to phase array antennas [3].

Previously designed antenna Electronically Steerable Parasitic Raidator(ESPAR) antennas have all cylindrical element, including the feed element [6], [7]. In this paper we have proposed design of seven element ESPAR antenna with conical center(active element)[1]. Length of conical element is optimized using optimetrics in HFSS. Our proposed antenna has higher bandwidth as compared to previously designed ESPAR antennas. [6], [7]. In multiple-input-multiple-output (MIMO)

---

[1] This work was supported by National Science Foundation (NSF) on Enhancing Access to Radio Spectrum (EARS)

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

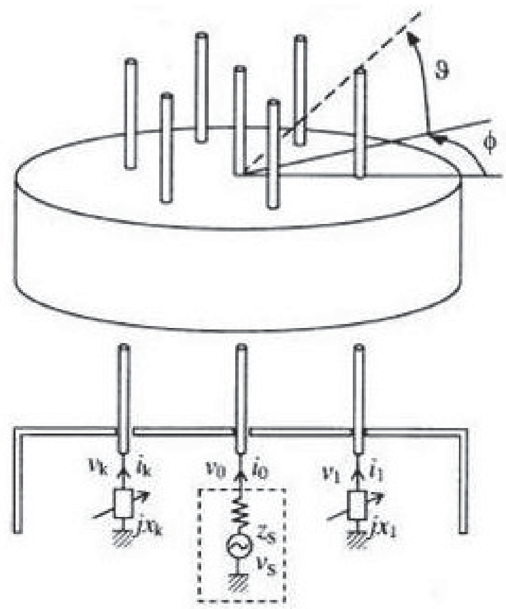

Fig. 1. Seven Element switched parasitic array antenna, ϕ is azimuth angle and θ is elevation angle.

systems there are multiple transmit and and receive antennas. Each receive antenna receives signal over different propagation channels yielding diversity gain. The MIMO technique can increase system throughput and reduce propagation losses. Similarly diversity gain can also be achieved by using smart antennas. In particular, we can transmit the same signal in multiple directions through directional beams, such that they experience different propagation channels [4]. Our proposed antenna is more suited for MIMO applications compared to previously designed model because of its higher gain and improved bandwidth efficiency.

This paper focuses on the design of a seven element switched parasitic array antenna. The antenna is first simulated and then it is fabricated. Simulation results are validated with the hardware results. Previously done work on ESPAR and switched array antennas involve cylindrical monopole elements mounted on a ground plain [8] due to cylindrical nature bandwidth of active monopole is confined, subsequently bandwidth of switched array and ESPAR antennas is limited. Instead on cylindrical monopole element we have used tapered conical monopole. Due to tapered design bandwidth is enhanced.

II. SWITCHED PARASITIC ARRAY ANTENNA CONCEPT

The conceptual overview of the switched parasitic array antenna is given in Fig. 1. We have first simulated the antenna with seven monopole elements. There is one element at center which is surrounded by a circular array. Feed is applied to the center element and it is termed as active element, whereas elements in the circular array are parasitic elements. Parasitic array antennas do not involve phase shifters, so that the manufacturing cost is less as compared to the phase array antennas. Each element is separated from other element by a quarter of wavelength, thus mutual coupling phenomena comes into account and there is a shift in phase of current in each element because of this mutual coupling phenomenon. Since switched parasitic array antennas offer adaptive beamforming solution with compact size and low cost, they are likely to be used in future communication devices to increase data-rates [6], [7]. Parasitic array antennas have vast potential for usage in laptops, cell phones, WLANs etc.

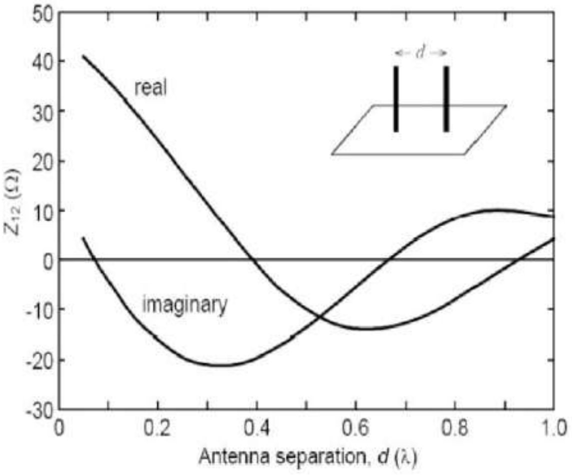

Fig. 2. Z_{21} versus separation distance.

A. Mutual Coupling

Mutual Coupling is basically the interaction between antenna elements in an array. Since parasitic elements are placed closer to each other (at a distance of $\lambda/4$) there will be mutual coupling.This mutual coupling phenomenon is responsible for a shift in phase of current induced in each element [8].The current induced by one antenna element in another antenna element [9] is given by (3)

$$I_2 = \frac{Z_{21}}{Z_{22}} \times I_0 \tag{2}$$

$$Z_{21} = \left|\frac{V_2}{I_1}\right|_{I_2=0} \tag{3}$$

$$Z_{21} = \left|\frac{V_2}{I_2}\right|_{I_1=0} \tag{4}$$

Here Z_{22} is constant as it is open-circuit output impedance. Z_{21} is open-circuit transfer impedenace from port 2 to port 1 and it can be approximated by graph shown in Fig. 2. From equation (3), beam can be steered in the desired direction by varying the phase of current I_2 [10].

978-1-5090-0783-7/16/$31.00 ©2016 IEEE

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

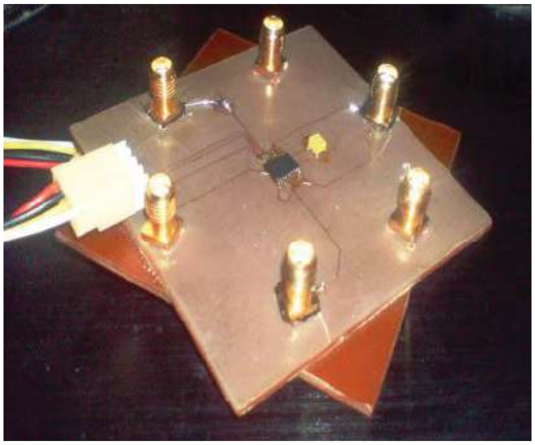

Fig. 3. RF PCB with Analog devices ADG 904 RF switches.

TABLE I
DESIGN PARAMETERS

Parameters	Values
Ground Skirt length	lambda/4
Monople Length	lambda/4
Frequency	2.45 GHz
Ground Radius	lambda/2
Monopole Radius	lambda/200
Ground Thickness	3 mm

B. Beam Steering Mechanism

In this work the designed antenna is capable of 360^0 beam steering in steps of 60^0. The beam steering is achieved by using two sets of 4:1 Analog devices multiplier switch ADG 904. Fig.3 shows RF PCB designed for beam steering. Beam is steered towards an element which is open. The open element behaves as a director and shorted elements act as reflectors. In order to achieve a maxima at 0^0, we make the element 1 open by using RF switch IC while other elements (i.e., elements 2,3,4,5 and 6) are shorted. This mechanism is similar to Yagi-Uda antenna where elements with shorter lengths in front of feed element act as directors and elements at rear of feed element acts as reflectors [11]. In Yagi-Uda antenna feed element is often a dipole or folded dipole.

III. PARASITIC ARRAY ANTENNA DESIGN

Design parameters and mechanical design of antenna is included in this section.

A. Design Parameters

Antenna design parameters are tabulated and the results are optimized using antenna simulation tools. Ground is skirted, since skirted ground provides mechanically sound model for PCB assembly to mount at back of antenna [12]. Skirt is $\lambda/4$ in length similarly each monopole element is lambda/4 in length and $\lambda/200$ is its radius.Table I shows the design parameters for the seven-element switched parasitic array antenna.

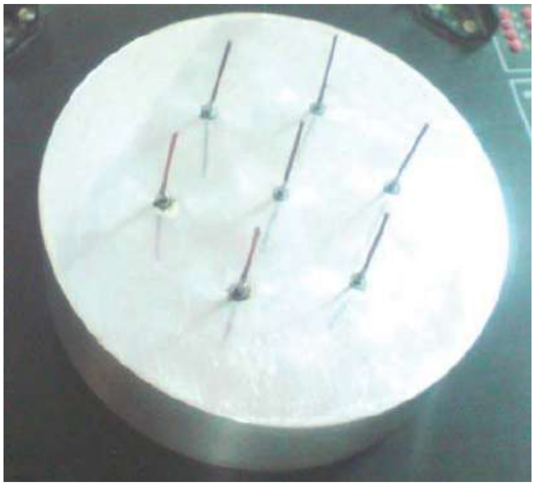

Fig. 4. Designed seven element parasitic switched array antenna.

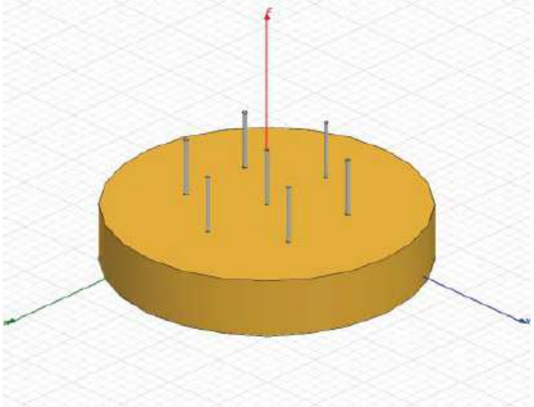

Fig. 5. Simulated antenna model.

B. Mechanical Design

Monopole elements are made by cutting copper wire of desired gauge i.e lambda/200 in radius [13] and these elements are soldered on pin of SMA connectors.Aluminum is used as ground plane. Electromagnetic waves are reflected by the ground. Instead of using seven dipole elements we have used seven monopole elements mounted on the ground plane [14]. Ground is made hollow in order to create space for mounting control circuitry [15]. Fig. 4 shows designed hardware.

IV. ANTENNA SIMULATION AND PLOTS

The antenna is first simulated and then simulated results are compared with the measured results of fabricated antenna. Simulated results were in line with the hardware results and 8 dB gain was observed in simulated results as well as in fabricated antenna. In simulation each of the parasitic element is loaded with lumped port excitation, simulating opening and shorting [16] of monopole elements. Fig. 5 shows simulated model.

978-1-5090-0783-7/16/$31.00 ©2016 IEEE

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

TABLE II
MAXIMA PATTERN

Azimuth Angle	Element 1	Element 2	Element 3	Element 4	Element 5	Element 6
0 degree	Open	Short	Short	Short	Short	Short
30 degree	Open	Open	Short	Short	Short	Short

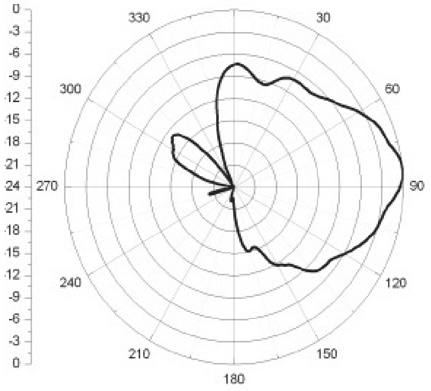

Fig. 6. Anechoic chamber results for θ=90^0 (Beam in elevated because we are using monopole elements with finite ground).

Two dimension polar plots depicts the beam steering in the direction of given elevation angle, theta (θ) and azimuth angle, phi (ϕ). Anechoic chamber results for θ=90 and ϕ=0 and 30 degrees are shown in Fig. 7 and 8 respectively. Using large ground plane angle of elevation is reduced closer to 90^0. Fig. 6 shows anechoic chamber result for θ=90 one can observe small elevation form desired x-y axis because of ground skirt [17].In simulations elevation angle of 85^0 is observed. Fig. 6(a) shows antenna simulated plot for elevation angle θ and Fig. 6(b) shows anechoic chamber results.Anechoic chamber plot of Fig. 7 shows maxima at ϕ=0^0, in order to get maxima at 0^0 elements from element number 2-6 are shorted and the element 1 is opened. Similarly Anechoic chamber plot of Fig. 8 shows maxima at ϕ=30^0, in order to get maxima at 30^0 elements from element number 3-6 are shorted and element 1 and element 2 are open.

Traditional ESPAR antenna and switched parasiric array antenna have all cylindrical elements, including the active element [3], [4]. We have made center element conical instead of cylindrical. Due to conical nature of active element response our frequency range and improved. In order to validate the gain enhancement both models were simulated and results were compared. Fig 9. (a) shows conventional design and Fig. 9(b) shows simulated antenna with conical center element. Fig. 10 shows S_{11} plot for cylindrical active element. Considering -8_{dB} as cut off point bandwidth range from 2.358 GHz to 2.51 GHz (0.152 GHz). Fig. 11 shows $S-11$ plot for centre conical element, considering 8_{dB} as cut off point bandwidth ranges from 2.36 GHz to 2.555 GHz (0.195 GHz). Therefore my making center element conical we have improved bandwidth as compared to traditional switch array parasitic antennas. Low profile design for switched parasitic array antennas in also highlighted in [18].

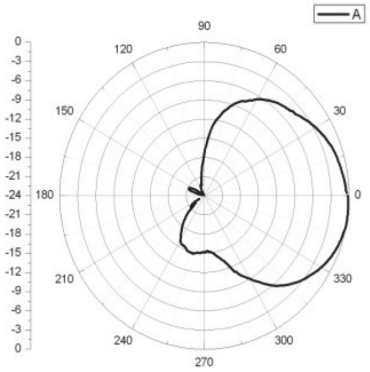

Fig. 7. Anechoic chamber results for ϕ=0^0.

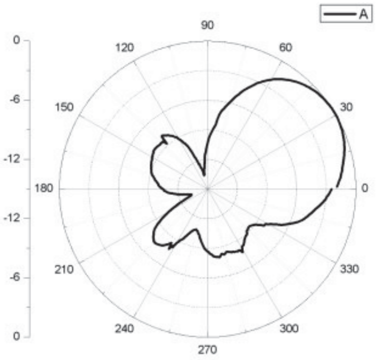

Fig. 8. Anechoic chamber results for ϕ=30^0.

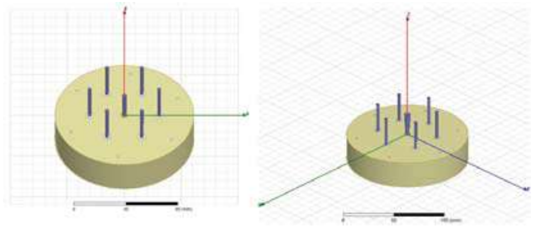

Fig. 9. (a) Conventional ESPAR antenna model with cylindrical Center Element, Fig. 9 (b) Designed ESPAR antenna with conical element for Bandwidth Enhancement.

V. FUTURE WORK AND CONCLUSION

5G networks will be designed for device to device, device to human and human to human interactions. High data rates are required to meet demands of 5G systems. One of proposed bands for 5G is in mm-Waves range i.e from 30GHz to 60GHz for such high frequencies path losses are considerable and beam steering antennas are way forward to overcome path

978-1-5090-0783-7/16/$31.00 ©2016 IEEE

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

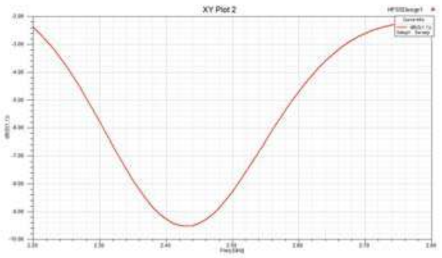

Fig. 10. S_{11} plot for conventional switched array antenna.

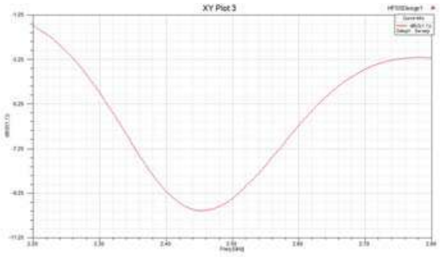

Fig. 11. S_{11} plot for center conical element switched array antenna.

losses [19]. ESPAR and switched parasitic array antennas provide a cost effective way of incorporating adaptive beamforming in future 5G and internet of Things (IoT) devices [20]. However for the efficient use of ESPAR antennas in mobile terminals the size of the antenna system must be reduced, for this further research on dielectric switched parasitic array antennas is underway. Seven element switched parasitic array antenna is designed that is capable of dynamic beam steering. Gain of 8dB is achieved practically. Simulated results are in line with the anechoic chamber results. Designed antenna increase the channel capacity by improving signal to noise ratio. Gain of 8dB was achieved in specified directions of elevation angle ,theta (θ) and azimuth angle, phi (ϕ). Antenna is optimized in terms of interference reduction for its use in wireless ad hoc networks, operating at frequency of 2.45 GHz. Bandwidth of designed antenna is 450 MHz and beam is steered in 360^0.

ACKNOWLEDGEMENT

Thanks to EPSRC Project EP/M014126/1, Large Scale Antenna Systems Made Practical: Advanced Signal Processing for Compact Deployments [LSAS-SP] and National Science Foundation (NSF), Enhancing Access to Radio Spectrum (EARS) program for funding.

REFERENCES

[1] M. Rzymowski, P. Woznica, and L. Kulas, "Single-anchor indoor localization using espar antenna," *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 1183–1186, 2016.

[2] L. Zhou, F. A. Khan, T. Ratnarajah, and C. B. Papadias, "Achieving arbitrary signals transmission using a single radio frequency chain," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4865–4878, Dec 2015.

[3] V. I. Barousis and C. B. Papadias, "Arbitrary precoding with single-fed parasitic arrays: Closed-form expressions and design guidelines," *IEEE Wireless Communications Letters*, vol. 3, no. 2, pp. 229–232, April 2014.

[4] M. R. Islam and M. Ali, "Elevation plane beam scanning of a novel parasitic array radiator antenna for 1900 mhz mobile handheld terminals," *IEEE Transactions on Antennas and Propagation*, vol. 58, no. 10, pp. 3344–3352, Oct 2010.

[5] S. Kausar, H. U. Rahman, T. Hassan, and A. Kausar, "Miniaturization of espar antenna using folded monopoles and conical central element," in *Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), 2015 International Conference on*. IEEE, 2015, pp. 87–91.

[6] H. T. Liu, S. Gao, and T. H. Loh, "Electrically small and low cost smart antenna for wireless communication," *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 3, pp. 1540–1549, March 2012.

[7] T. H. Loh, H. Liu, and S. Gao, "A cost-effective direct magnitude measurement methodology for smart antennas," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4, pp. 2043–2050, April 2013.

[8] F. Khan, K. Ali, A. Kausar, and S. Kausar, "Energy generation capacity analysis of a canal based hydro project," in *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation*, Sept 2013, pp. 277–281.

[9] A. Kausar, M. A. Cheema, S. Kausar, H. Rehman, and T. Hassan, "Espar antenna system designing & simulation," in *Proceedings of the 2014 First International Conference on Systems Informatics, Modelling and Simulation*, ser. SIMS '14. IEEE Computer Society, 2014, pp. 193–196. [Online]. Available: http://dx.doi.org/10.1109/SIMS.2014.37

[10] S. Akkar and A. Gharsallah, "Reactance domains unitary music algorithms based on real-valued orthogonal decomposition for electronically steerable parasitic array radiator antennas," *IET Microwaves, Antennas Propagation*, vol. 6, no. 2, pp. 223–230, January 2012.

[11] S. Kausar, H. u. Rahman, A. Kausar, and T. Hassan, "Espar antenna system for dynamic tracking of active targets," in *Modelling Symposium (EMS), 2013 European*, Nov 2013, pp. 533–535.

[12] A. Umar Kausar, F. Aslam Khan, T. Hassan, A. Nasser, and S. Kausar, "Adaptive espar antenna system modelling & design." *International Journal of Simulation–Systems, Science & Technology*, vol. 14, no. 4, 2013.

[13] D. v. B. M. Trindade, C. Mller, M. C. F. D. Castro, and F. C. C. D. Castro, "Metamaterials applied to espar antenna for mutual coupling reduction," *IEEE Antennas and Wireless Propagation Letters*, vol. 14, pp. 430–433, 2015.

[14] S. Kausar, Ahmed Kausar, "Smart adaptive beam forming antenna for interference cancellation," in *Wireless Days (WD), 2013 IFIP*. IEEE, 2013, pp. 1–3.

[15] A. Kausar, S. Kausar, and T. Hassan, "Smart adaptive beam forming antenna for interference minimization," in *Future Generation Communication Technology (FGCT), 2013 Second International Conference on*. IEEE, 2013, pp. 6–9.

[16] F. Aslam, E. Ulhaq, A. Nasser, and A. Umar, "Intelligent modeling scheme for detection of line losses in power distribution system," in *Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on*. IEEE, 2013, pp. 218–223.

[17] R. Qian, M. Sellathurai, and D. Wilcox, "A study on mvdr beamforming applied to an espar antenna," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 67–70, Jan 2015.

[18] A. Pal, A. Mehta, D. Mirshekar-Syahkal, and H. Nakano, "A low-profile switched-beam dual-band capacitively coupled square loop antenna," in *Antennas and Propagation Conference (LAPC), 2013 Loughborough*, Nov 2013, pp. 563–566.

[19] A. Kausar, S. Kausar, T. Hassan *et al.*, "Interference cancellation using adaptive beam forming for 3 d beam steering." *International Journal of Simulation–Systems, Science & Technology*, vol. 14, no. 1, 2013.

[20] M. R. Nikkhah, P. Loghmannia, J. Rashed-Mohassel, and A. A. Kishk, "Theory of espar design with their implementation in large arrays," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 6, pp. 3359–3364, June 2014.

[21] T. Hassan, A. Kausar, H. Umair, and M. A. Anis, "Gain optimization of a seven element espar antenna using quasi-newton method," in *Microwave technology & computational electromagnetics (ICMTCE), 2011 IEEE international conference on*. IEEE, 2011, pp. 293–296.

Recent Antenna System Technologies for Next-generation Wireless Communications

Koichi Tsunekawa†

Abstract

This paper describes the requirements of the antennas or antenna systems in recent high-speed and high-capacity wireless communication systems and the key technologies for satisfying them. It also explains the antennas and antenna systems being researched and developed for next-generation wireless communication systems. The antenna in such a system requires high gain, high-efficiency antenna elements and a multi-antenna design. This paper introduces a high-efficiency antenna created by integrating it with a monolithic microwave integrated circuit (MMIC) in a millimeter-wave frequency-band system, a multi-antenna system for the multiple-input multiple-output (MIMO) technique that enables multiple data transmission for land mobile communication systems, and a beam-scanning antenna for satellite communication systems. It also explains the general research and development trend of antenna systems and technical problems in each wireless system as an introduction to the other selected papers in this issue.

1. Introduction

Wireless access technologies have advanced rapidly [1] and broadband wireless communication systems have become popular in our daily lives. An antenna is a very important component of a wireless system because it acts as the input and output interface for wireless equipment, as shown in **Fig. 1**. However, system performance is determined by the total performance of each component, such as the antenna and the radio-frequency (RF), intermediate-frequency (IF), and baseband (BB) units. For recent high-performance wireless systems, overall system design has been essential. Accordingly, the antenna must not

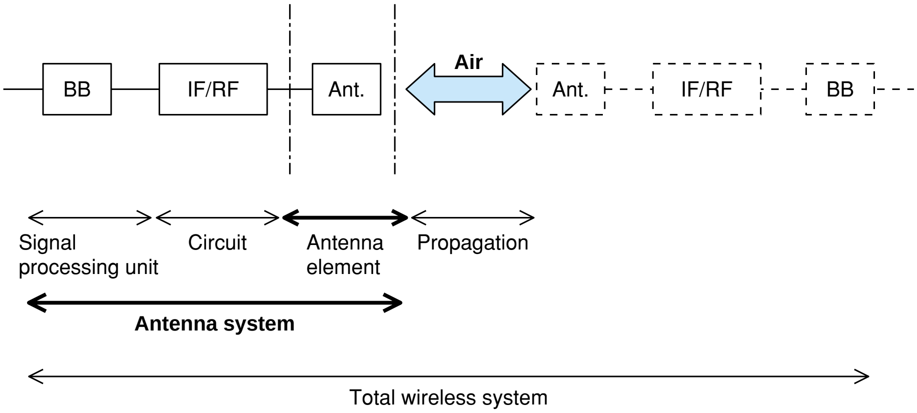

Fig. 1. Wireless and antenna system.

---

† NTT Network Innovation Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: tsunekawa.koichi@lab.ntt.co.jp

only have antenna elements with high-performance characteristics, but must also perform effectively as an antenna system in conjunction with the RF, IF, and BB units. New functions can be added through antenna system design: for example, appropriate radiation pattern control provides a large system capacity by using baseband signal processing and an antenna element with an integrated RF circuit provides high gain without feeder loss.

The selected papers in this issue describe the latest antenna and antenna system technologies for achieving high performance. Two of the papers present beam-scanning antenna systems in satellite broadband communication. Beam scanning is carried out using by two methods: one [2] uses a variable active device or tunable reactance device incorporated into the antenna element, which has high-speed electronic control while the other [3] uses mechanically controlled antenna movement, which has a very precise direction control function. Another paper [4] introduces an active integrated antenna system with an RF circuit, which provides high gain and high efficiency, for a high-capacity millimeter-wave transmission system. And two papers describe a multiple-input multiple-output (MIMO) antenna system for obtaining high-speed transmission and high system capacity in the land mobile communication system from the viewpoints of beamforming techniques [5] and implementation methods [6].

As an introduction to these advanced antenna techniques, this paper provides a general explanation of the requirements of antennas and antenna systems in recent high-speed and high-capacity wireless communication systems and the key technologies for satisfying these requirements. It also describes current antenna systems and NTT's antenna research results for typical wireless systems.

2. Antenna requirements in recent wireless communication systems

The trend of wireless systems is shown in **Fig. 2**. This indicates the typical system transmission speed and number of antennas needed to achieve a particular system performance for each operating frequency. Wireless communication systems are moving towards high transmission speeds and high capacity by using higher system frequencies to enable a wide frequency

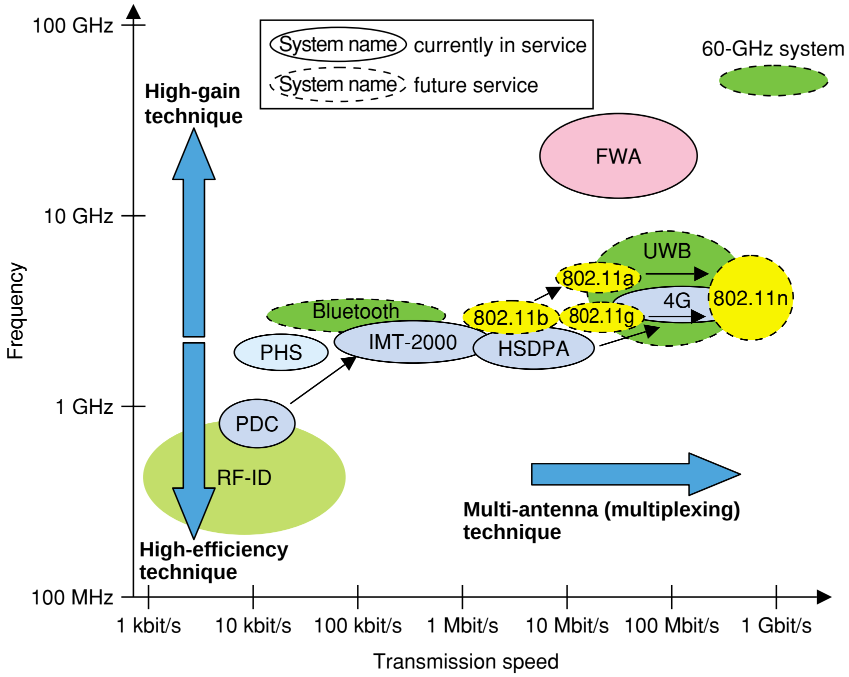
100 GHz
10 GHz
1 GHz
100 MHz
Frequency
System name currently in service
System name future service
High-gain technique
60-GHz system
FWA
UWB
802.11a
4G
802.11n
Bluetooth
802.11b
802.11g
PHS
IMT-2000
HSDPA
PDC
RF-ID
Multi-antenna (multiplexing) technique
High-efficiency technique
1 kbit/s
10 kbit/s
100 kbit/s
1 Mbit/s
10 Mbit/s
100 Mbit/s
1 Gbit/s
Transmission speed

RF-ID: radio-frequency identification
PDC: personal digital cellular
PHS: personal handy-phone system
IMT-2000: international mobile telecommunications 2000

HSDPA: high-speed download packet access
802.11x: family of IEEE standards
UWB: ultrawideband
FWA: fixed wireless access
4G: fourth generation

● : ubiquitous system
● : land mobile system
● : short-range system
● : WLAN system
● : fixed wireless system

Fig. 2. Advances of terrestrial wireless communication systems and their relationship to antenna technologies.

band and using advanced system technology to achieve high quality and signal multiplexing. The antenna requirements for these system properties are:

- High-gain and high-efficiency technology
- Multi-antenna technology

Increases in the gain and efficiency of an antenna are offset by higher propagation loss at high frequencies, so the received signal level needs to be higher. High gain is an age-old problem in antenna research and development. Multi-antenna technology improves the receiving signal quality and provides multiplexed signal transmission by using an antenna system that includes RF circuits, BB units, and antenna elements.

2.1 High-gain and high-efficiency antenna

Theoretical and experimental antenna gain and efficiency characteristics are shown in **Fig. 3**. The dependences of gain and beamwidth on horn dimension L(λ) in a relatively large antenna are explained in Fig. 3(a) [7]. When L is more than 0.5 times the wavelength, the antenna's efficiency is about 100%; a larger antenna provides higher gain. However, the high-gain radiation pattern leads to a narrow beamwidth. Therefore, a mobile terminal needs to use beam-scanning or beam-tracking techniques to follow the transmitter while the terminal moves. On the other hand, the efficiency decreases if the antenna size is less than 0.5 wavelengths. The bandwidth of a linear dipole antenna versus its efficiency for various antenna sizes (D: diameter) is shown in Fig. 3(b). This graph was estimated from theoretical limits [8] and many experimental results. The antenna efficiency becomes approximately –5 dB when antenna bandwidth of 10% is provided using a linear dipole antenna with a length of 0.3 wavelengths. This figure shows that the efficiency and bandwidth have a tradeoff relationship and that the efficiency drops greatly as the antenna size decreases. Therefore, some clever ideas for the antenna element structure are needed to overcome these weak points.

2.2 Multi-antenna system

A multi-antenna system has several antenna elements, but it does not merely work as an ordinary array antenna. The performance requirements are satisfied because the antenna system includes RF circuits, BB units, and antenna elements. The antenna element technologies and the multi-antenna system are shown in **Fig. 4**. In the antenna element, a high gain can be achieved by improvements to the antenna element structure, for example, by using a new material, adding a parasitic element, or integrating an RF circuit (Figs. 4(a)–(c)). The antenna shown in Fig. 4(c) is called an integrated antenna. It can control the radiation pattern as well as provide high gain by eliminating feeder loss. In a multi-antenna system (Fig. 4(d)), the radiation pattern is controlled by baseband signal processing, so the pattern of each wireless channel can be controlled by each baseband channel. Furthermore, it is possible to get multiplexed signal transmission in the space domain by baseband signal processing. Multi-antenna systems are generally called smart antennas or active antennas [9].

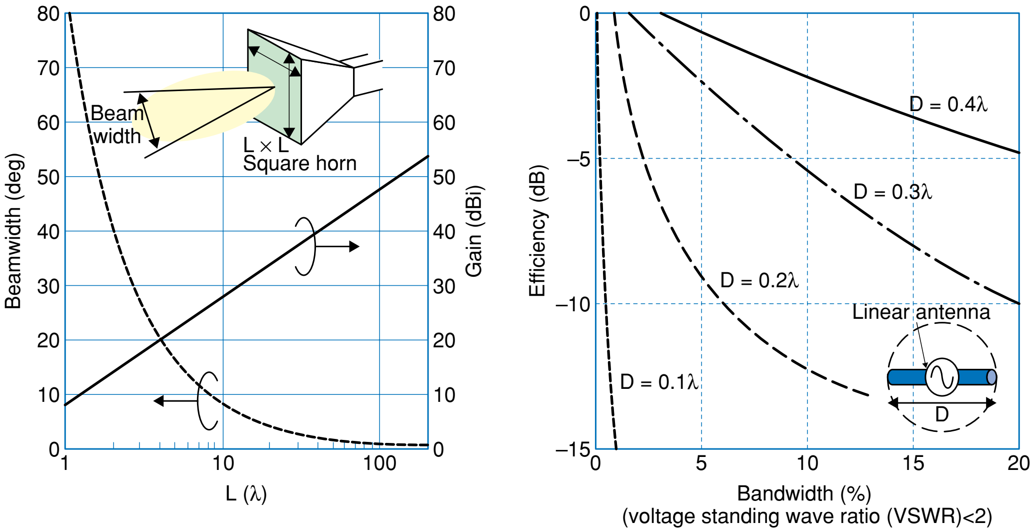

(a) Aperture antenna gain characteristics

(b) Dipole antenna efficiency characteristics

Fig. 3. Antenna gain and efficiency characteristics.

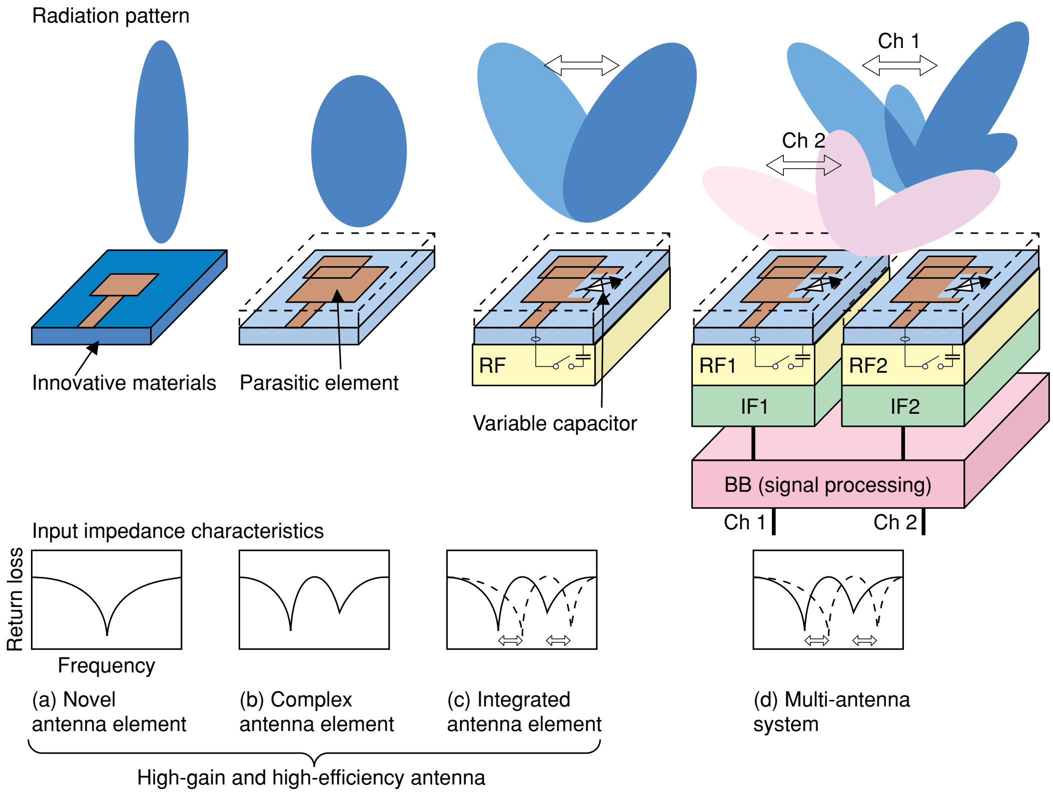
Radiation pattern
Ch 1
Ch 2
Innovative materials
Parasitic element
RF
Variable capacitor
RF1
RF2
IF1
IF2
BB (signal processing)
Ch 1
Ch 2
Input impedance characteristics
Return loss
Frequency
High-gain and high-efficiency antenna

(a) Novel antenna element

(b) Complex antenna element

(c) Integrated antenna element

(d) Multi-antenna system

Fig. 4. Antenna elements and system.

An example of a mobile communication system using a smart antenna in the base station is shown in **Fig. 5**. Independent information can be sent to the terminals, which are spatially separated, using the same frequency at the same time because each beam is generated for a specific terminal by using the space division multiple access (SDMA) technique. Furthermore, N multiplexed transmissions can be sent to a terminal with N antennas by utilizing the slight difference in antenna locations in the terminal. MIMO is representative of this kind of technology. The latest smart antenna systems can increase the antenna gain even in a mobile system and expand the service area. The designed system capacity can be increased by using multiplexing techniques based on baseband signal processing. Multi-antenna systems can achieve very high performance with high-level functions corresponding to the wireless system requirements.

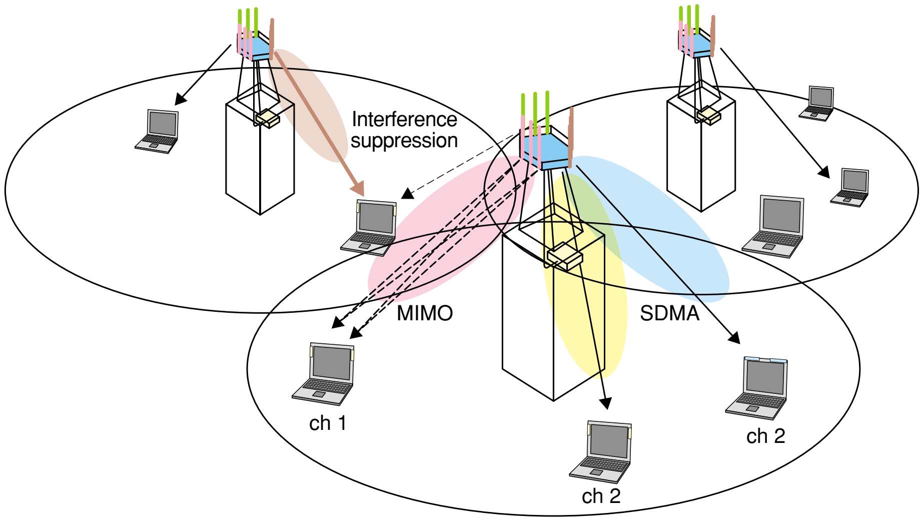
Interference suppression
MIMO
SDMA
ch 1
ch 2
ch 2

Fig. 5. Smart antenna system (land mobile telecommunication).

2.3 Key antenna technologies

The key technologies for achieving a high-gain, high-efficiency, multi-antenna system are shown in **Table 1**. Antennas for the latest wireless communication systems are required to achieve high gain and high efficiency. The multi-antenna systems are designed to make maximum use of system resources, especially space.

3. Wireless systems and suitable antenna systems

The kinds of wireless systems that we can expect to find around us in the near future are illustrated in **Fig. 6**. RF-ID (radio-frequency identification) tags will be attached to various kinds of objects in a ubiquitous network and a large quantity of information will be gathered to provide useful knowledge to each person. High-speed data transmission ranging from a few megabits per second to 1 Gbit/s will be possible in mobile satellite systems, land mobile systems, and wireless local area network (WLAN) systems. In home networks, individual items of electronic equipment will be connected to each other at very high

Table 1. Key technologies to meet antenna requirements.

Requirements	Key technologies
1) High gain and high efficiency	a. Innovative materials (meta-material, high magnetic material, …) b. Element design (parasitic element, fractal element, …) c. Integrated antenna (antenna with RF circuit or RF device, …) d. Active antenna (antenna within variable reactance or switch, ….)
2) Multi-antenna techniques	a. Beam scanning antenna b. Sector antenna (fixed narrow multi-beam system) c. Diversity antenna d. SDMA antenna e. MIMO antenna

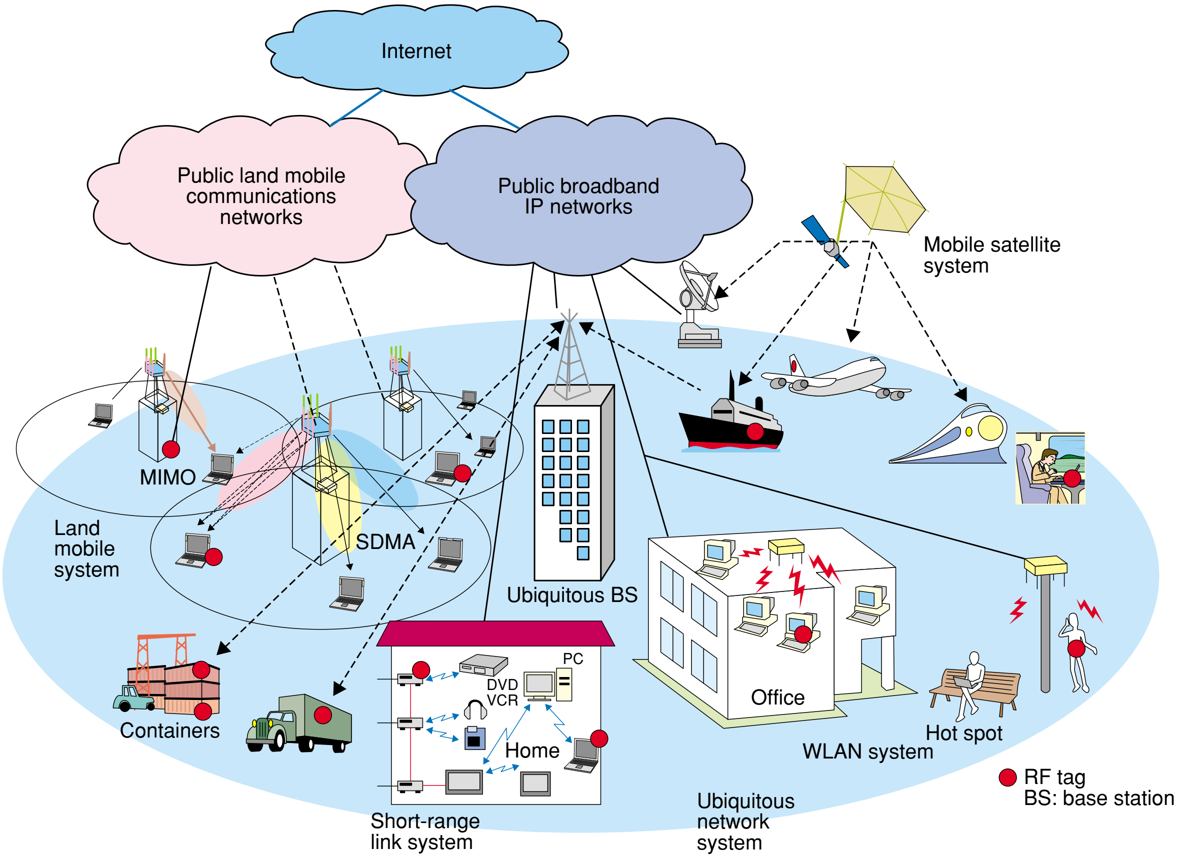
Internet
Public land mobile communications networks
Public broadband IP networks
Mobile satellite system
MIMO
Land mobile system
SDMA
Ubiquitous BS
Containers
PC
DVD
VCR
Home
Short-range link system
Office
WLAN system
Hot spot
Ubiquitous network system
● RF tag
BS: base station

Fig. 6. Wireless communication systems.

Table 2. Wireless systems and NTT's antenna system approaches.

	System features	Main technical trends (example system, standard, or equipment)	NTT's antenna approaches (using technologies in Table 1)
Ubiquitous RF-ID	Multiple access Low rate	Active tags Sensor network Mesh network (802.15.5)	Compact high-efficiency tag antenna High-gain beam tilting base station antenna (1a, 1b, 1c)
Short-range communication	Short range Ultrahigh speed Link transmission	High frequency band (millimeter-wave system) Ultrawideband (UWB) Infrared rays (IrDA)	Active integrated antenna system (millimeter wave) (1b, 1c, 2a, 2b)
Land mobile and WLAN	Medium range High speed/capacity Many subscribers	MIMO-OFDM (802.11n,802.16a/e) MC-CDMA (4G, Beyond 3G) HSDPA (3.5G,1xEV-DO/V)	MIMO-OFDM antenna system Compact mobile terminal antenna system (1b, 1d, 2b, 2c, 2d, 2e)
Mobile satellite	Very long range Medium to high speed Various mobile stations	S/Ku band high-speed mobile comm. (N-Star/ETS-VIII) Broadband (Internet) (WINS) Millimeter wave/optical comm. (MILSTAR, GeoLITE)	Multi-beam antenna Large deployable antenna Beam-tracking antenna system (1c, 1d, 2a, 2e)

(discussed in papers in this issue)

speed (several gigabits per second) wirelessly instead of by cables. In these wireless systems, highly functional and high-performance antennas will be essential. **Table 2** explains the main features and technical trend of the systems and also shows NTT's approaches to antennas and antenna systems for each high-performance wireless system. The technologies in Table 1 that are used are indicated in parentheses. Each antenna system must satisfy severe specifications and also provide high functionality, as shown in Figs. 4(c) and (d). The antennas discussed in the selected papers in this issue are shown in blue in Table 2. The present status of each antenna system and research and development in NTT are explained below.

3.1 Ubiquitous network antenna system

Tag antenna elements have the same problems as pagers or portable telephone systems that use the LF, UHF, and VHF bands. The antenna element must be as small as possible and have high efficiency in a small volume and in the neighborhood of various objects. Most antennas for passive tags using the LF band (13 MHz) are spiral, helical, N-turn loop, or meander line antennas [10] for achieving an electromagnetic induction area and their gains are considered to be very low. Bent and spiral monopole or dipole antennas are used in personal computer card wireless communication systems or in small square film tags using UHF-band systems [11]. The approach is to design the antenna and RF-IC as one system; in other words, the connection impedance between the antenna and RF-IC is designed as a freely selectable value instead of being restricted to 50 Ω. The same kind of technique is currently being investigated in NTT in a study of active integrated antennas [12]. On the other hand, the base station antenna can be designed by applying techniques developed for pagers or land mobile antennas because it has high-gain and beam-tilting characteristics.

3.2 Millimeter-wave frequency band communication antenna system

A high-speed wireless communication system is achieved by using wideband technology, so millimeter-wave frequencies, which provide a very wide bandwidth, are attracting attention. The use of high frequencies also requires high output power to compensate for the huge propagation loss and passive circuit losses, so high-gain and high-efficiency antennas are very important in system design [13]. Millimeter-wave systems mainly use horn, reflector [14], or lens antennas [15]. Horn or reflector antennas are used in relatively long-range communication where the range is several hundred meters. A lens type antenna, which is integrated with monolithic microwave integrated circuits (MMICs) to achieve high efficiency without a feeding line, is mainly used for short-range transmission, e.g., for a home link system. Active integrated antenna technology in which an amplifier for transmitting and receiving is mounted without using a connector or cables is suitable for a high-frequency-band system [16]. Typical antenna types for millimeter-wave frequency systems are shown in **Table 3**. Horn or reflector antennas provide very high gain but cannot be integrated with MMICs. The lens antenna provides high gain and can be made into an integrated antenna but requires adjustment to correct its

Table 3. Antenna types for millimeter-wave systems.

Antenna	Gain	Integration with MMIC	Mounting adjustment
Horn or reflector	Very high (over 20 dBi)	No	(Feeder)
Lens	High (10 to 20 dBi)	Yes	Needed
Single patch	Low (2 dBi)	Yes	No
Parasitic element patch	High (10 dBi)	Yes	No

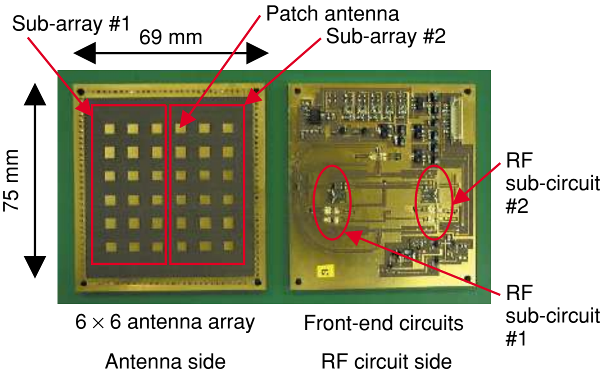

Fig. 7. Active integrated antenna for 25-GHz system.

mounting error. The patch antenna is a good element for integration with MMICs because it can be made on the same material substrates, so it does not need antenna adjustment. We have developed an active integrated antenna designed for broadband mobile wireless access systems using the 25-GHz band, as shown in **Fig. 7** [12]. It exhibited output power of 14.6 dBm and a noise figure of less than 5 dB. The only weakness of this integrated antenna is that high gains cannot be provided with one patch antenna. The third selected paper in this issue describes a high-gain integrated active antenna and high-efficiency RF circuit for the millimeter-wave frequency band for a millimeter-wave communication system [4]. A high-gain integrated patch antenna is investigated as a parasitic element to obtain a wide antenna aperture. The millimeter-wave frequency-band link system was first proposed for application to the transmission of uncompressed HDTV (high-definition television) signals in the home or in a studio.

3.3 Land mobile and WLAN antenna systems

The number of subscribers to land mobile systems, especially cellular systems, has increased rapidly, and users are now demanding high-speed transmission for broadband Internet access. To meets these demands, many improvements to the land mobile system have been proposed such as wideband code division multiple access (W-CDMA), adaptive modulation, and *ad hoc* transmission. Recently, the utilization of space has become as desirable as frequency and time utilization [17]. Space utilization is achieved using a multi-antenna system with a radiation control mechanism, which is accomplished by a smart antenna system.

3.3.1 Typical multi-antenna techniques for mobile systems

The application areas of multi-antenna techniques depend qualitatively on the system characteristics, as shown in **Fig. 8**. A fixed high-gain beam antenna does not always need antenna elements, but the other techniques are achieved through various combinations of antenna elements, so they are provided by adaptive or integrated multi-antenna systems, as shown in Figs. 4(c) and (d). A narrow-beam or high-gain antenna is used in the line-of-sight situation (where there are no obstacles between the transmitting and receiving antennas and there are few paths in the multipath transmission), so a beam-scanning mechanism is needed when a terminal moves fast, and multiple beams or sector beams are required when there are many terminals in the service area. The diversity antenna technique can provide a stable signal level in a multipath mobile environment. Beam scanning (with a sector beam or multiple beams) and diversity antenna techniques have been used fairly extensively from the early days of mobile communication service because they use only simple signal processing. However, with the rapid progress of signal processing technology, a desired signal can now be distinguished from a mixture of many signals in a multipath environment. The SDMA technique distributes signals among the terminals communicating in the service area. MIMO techniques enable very-high-data-rate communication with multiplexed transmission in each antenna of the multi-antenna system of the base station or mobile terminal. These techniques require more calculation as the movement speed of the mobile terminal increases. In these wireless systems, signal processing ability and superior algorithms are more important than the performance of antenna elements.

3.3.2 Research results and developed antenna systems

A very small and thin planar six-sector antenna using a patch Yagi-Uda array with common directors for mobile terminals has been proposed (**Fig. 9**) [18],

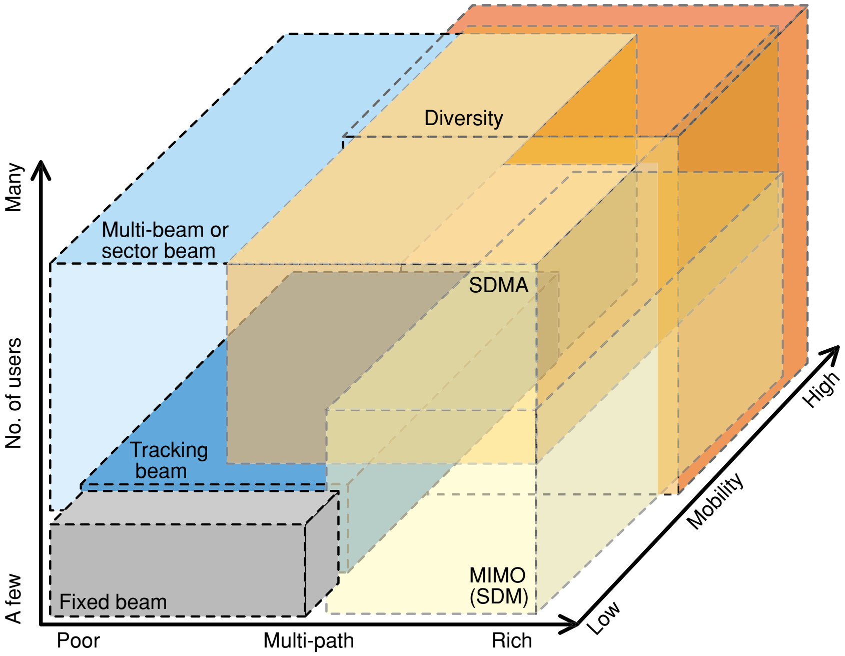
Many
No. of users
A few
Multi-beam or sector beam
Diversity
SDMA
Tracking beam
Fixed beam
MIMO (SDM)
Poor
Multi-path
Rich
Low
Mobility
High

Fig. 8. Dependence of multi-antenna techniques on system characteristics.

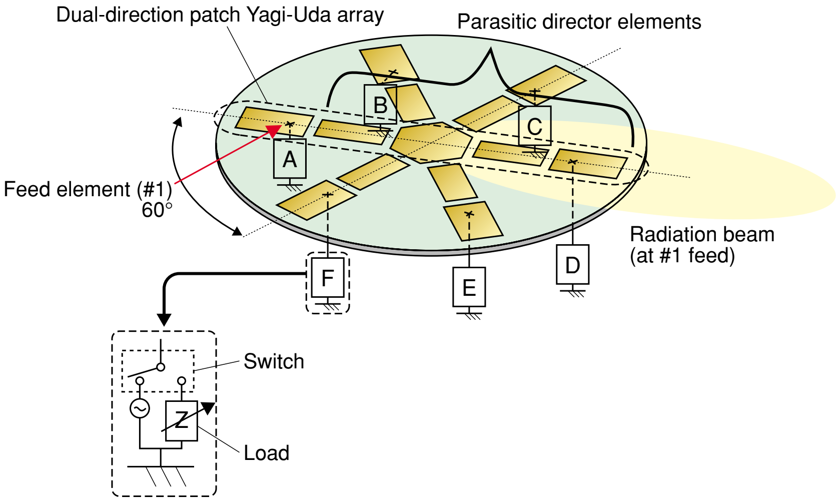
Dual-direction patch Yagi-Uda array
Parasitic director elements
Feed element (#1)
60°
A
B
C
D
E
F
Radiation beam (at #1 feed)
Switch
Z
Load

Fig. 9. Thin and small sector beam antenna.

[19]. It has square parasitic director elements that are shared by two sector arrays that face in opposite directions, and it has a hexagonal central director element that is shared by all six sectors. These elements form three line arrays, which intersect at 60°. These shared elements enable the sector antenna to be compact. Experimental results showed that the fabricated antenna had a conical plane beam width of 63° and high front-back ratio of 17 dB. The creation of adaptive (smart) antenna hardware has been investigated [20]. A novel beam combination method for wideband digital indoor communication systems such as WLANs has been proposed [21] in NTT. Crossed fan beams are used at the base station and mobile terminal. The beam combination method provides better transmission quality than the traditional pencil-beam combination and eliminates the need for complex beam control in the base station. Maximum dispersion distributions in the transmission frequency bandwidth have been calculated using a three-dimensional indoor propagation delay simulation algorithm. A method of designing adaptive antennas including an

antenna configuration that is suitable for street microcells, considering the propagation environment, has been described. An automatic calibration method at the transmitters and receivers using the transmitting signal (ACT), which enables realtime calibration, has been presented [22]. An SDMA technique using polarization has also been studied [23]. The transmission quality of an SDMA scheme that uses vertical pattern and polarization control was investigated in an actual cellular environment. For a sector cell system, this configuration was shown to be effective based on an evaluation of the spatial correlation characteristics. We also found that the output carrier-to-interference-plus-noise power ratio (CINR) of the proposed configuration was from 10 to 17 dB higher than that of the conventional SDMA configuration when the number of users was greater than four. The proposed configuration can reduce the antenna size compared with conventional SDMA. Moreover, MIMO techniques enable very-high-data-rate transmission using a limited frequency band. We have studied technologies based on the eigenmode-space division multiplexing (E-SDM) [24] scheme because it provides the theoretically maximum channel capacity and reduces the terminal's calculation load. The selected papers in this issue introduce both theoretical and practical technologies. The fourth paper describes new MIMO methods with beamforming to overcome problems of calculation complexity and low MIMO effect in the line-of-sight scenario [5]. The fifth paper describes the configuration of an 8 × 4 MIMO-OFDM transceiver and presents experimental results for the propagation characteristics, bit error rate (BER), and frequency utilization [6].

3.4 Satellite broadband communication system

If we want to create a seamless and flexible wireless network system, then a satellite communication system is very important because it provides a wide service area. However, it has the inherent problem that the distance between the transmitter and receiver is long, so the terminal needs a larger antenna than ones in the land mobile system to provide high-speed transmission. The effective solution is to make the satellite dish larger to provide high gain and multibeam transmission. We have studied multibeam communication satellite technology and onboard large deployable antennas for creating a broadband mobile satellite communication system [25]. An MMIC-based beamforming network equipment for a multibeam feeder whose weight is only 1/3 that of the conventional type was developed. We are also developing a large, highly precise deployable antenna with a diameter in the over-10-m range. The mobile terminal acting as an earth station is also an important component of a broadband mobile satellite communication system. Typical existing antennas in earth stations are shown in **Table 4**. Various mobile terminals are accommodated in the satellite communication system by different transmission rates, and these terminals use various antenna methods and locations. A comparatively large terminal antenna is required to handle high data transmission rates in a satellite system to mitigate the huge propagation loss. The high gain and narrow beam of a large antenna make a beam-scanning mechanism necessary to enable the antenna to follow the satellite direction as the terminal moves. The sixth selected paper describes a highly accurate and cost-effective auto-tracking antenna system for earth stations onboard vessels, using a novel inclinometer that is highly accurate in acceleration disturbance environments and a new systematic stabilization controller design process based on H_∞ control [3].

Table 4. Typical antennas in earth stations.

Satellite category	Typical satellite system	Current typical earth station antenna (size or length, gain)
LEO	Iridium Globalstar	Helical (0.1 m, 2 dBi) Linear whip (0.1 m, 0 dBi) Patch (0.1 m^2, 3 dBi)
MEO GEO	ICO N-STAR INMARSAT	Parabola (0.6 – 0.9 m, 18 – 21 dBi) Phased patch array (0.5 m^2, 14 dBi) Helical (0.3 m, 2 dBi) Patch array (0.3 m^2, 12 dBi) Patch (0.1 m^2, 3 dBi)

LEO: low earth orbit
MEO: medium earth orbit
GEO: geostationary earth orbit

4. Conclusion

This paper gave an overview of the requirements of antennas and antenna systems in recent high-speed and high-capacity wireless communication systems and the key technologies for satisfying these requirements to create next-generation wireless communications. It also described current antenna systems and NTT's antenna research results for typical wireless systems.

Antennas for next-generation wireless communications require high gain, high efficiency, and a multi-antenna design. A high-gain antenna can be made using a relatively large antenna aperture, but a beam-scanning mechanism is needed for mobile communications. A high-efficiency antenna can be made using an active integrated antenna, which integrates an MMIC with the antenna element. Multi-antenna technology is the innovative technology that enables multiple data transmission in space. In MIMO-type multi-antenna systems, the most important key technologies are the beamforming algorithm and the implementation method. The solution technologies for NTT's advanced antenna systems are explained in detail in the following selected papers in this issue. The technological background and the current research and development situation are also described to make the papers easy to understand.

References

[1] "Next-Generation Broadband Wireless Access Technologies," NTT Technical Review, Vol. 2, No. 1, pp. 12-63, 2004.

[2] N. Honma, T. Seki, and K. Tsunekawa, "Planar Beam-Scanning Microstrip Antenna Using Tunable Reactance Devices for Satellite Communication Mobile Terminal," NTT Technical Review, Vol. 3, No. 9, pp. 24-32, 2005.

[3] T. Yoshida, K. Ohata, and M. Ueba, "Highly Accurate and Cost-effective Auto-tracking Antenna System for Satellite Broadband Communication on Vessels," NTT Technical Review, Vol. 3, No. 9, pp. 60-71, 2005.

[4] T. Seki, K. Nishikawa, I. Toyoda, and K. Tsunekawa, "Millimeter-wave High-Efficiency Multilayer Parasitic Microstrip Antenna Array for System-on-package," NTT Technical Review, Vol. 3, No. 9, pp. 33-41, 2005.

[5] Y. Takatori, R. Kudo, K. Nishimori, and K. Tsunekawa, "Beamforming Method for Broadband MIMO-OFDM Systems," NTT Technical Review, Vol. 3, No. 9, pp. 42-49, 2005.

[6] K. Nishimori, R. Kudo, Y. Takatori, and K. Tsunekawa, "Experimental Evaluation of 8×4 Eigenmode SDM Transmission in Broadband MIMO-OFDM Systems," NTT Technical Review, Vol. 3, No. 9, pp. 50-59, 2005.

[7] J. D. Kraus, "Antennas, Sec. 2: Basic Antenna Concepts," McGraw-Hill Book Company, 1988.

[8] R. C. Hansen, "Fundamental limitations in antennas," Proceedings of the IEEE, Vol. 69, Issue 2, Feb. 1981.

[9] A. Alexiou and M. Haardt, "Communications smart antenna technologies for future wireless systems: trends and challenges," Magazine, IEEE, Vol. 42, Issue 9, pp. 90-97, Sep. 2004.

[10] S. S. Basat, K. Lim, J. Laskar, and M. M. Tentzeris, "Design and Modeling of Embedded 13.56 MHz Rfid Antennas," IEEE AP-S Int. Antennas Propagat. Symp. Dig.,Vol. 4, Session 130, pp. 64-67, 2005.

[11] U. Karthaus and M. Fischer, "Fully integrated passive UHF RFID transponder IC with 16.7-/spl mu/W minimum RF input power," Solid-State Circuits, IEEE Journal of Vol. 38, Issue 10, pp. 1602-1608, Oct. 2003.

[12] T. Seki, F. Nuno, K. Nishikawa, and K. Cho, "Active Integrated Antenna Technique," NTT Technical Review, Vol. 2, No. 1, pp. 36-43, 2004.

[13] T. Ihara and K. Fujimura, "Research and development of millimeter-wave short-range application systems," Trans., IEICE, Vol. E79-B, No. 12, pp. 1741-1753, Dec. 1996.

[14] W. Menzel, D. Pilz, and M. Al-Tikriti, "Millimeter-wave folded reflector antennas with high gain, low loss, and low profile," IEEE Antennas and Propagation magazine, pp. 24-29, June 2002.

[15] M. J. M. van der Vorst, P. J. I. de Maagt, A. Neto, R. M. Heeres, W. Luinge, and M. H. A. J. Herben, "Effect of internal reflections on the radiation properties and input impedance of integrated lens antennas-comparison between theory and measurements," Microwave Theory and Techniques, IEEE Trans. on 49, Issue 6, Part 1, pp. 1118-1125, June 2001.

[16] T. Seki, K. Uehara, and K. Kagoshima, "A Three-Dimensional Active Antenna for a High-Speed Wireless Communication Application," IEEE MTT-S Int. Microwave Symp., Denver, USA, Vol. 2, pp. 975-978, June 1997.

[17] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," IEEE Journal on Vol. 17, Issue 11, pp. 1841-1852, Nov. 1999.

[18] N. Honma, F. Kira, T. Maruyama, K. Cho, and H. Mizuno, "Compact six-sector antenna employing patch Yagi-Uda array with common director," Antennas and Propagation Society International Symposium 2002, IEEE, Vol. 1, pp. 26-29, June 2002.

[19] N. Honma, T. Seki, K. Tsunekawa, F. Kira, and K. Cho, "Compact Multi-Sector Antenna Employing Microstrip Yagi-Uda Array Antenna with Common Director Elements," IEICE Trans. B, Vol. J87-B, No. 9, pp. 1363-1371, Sep. 2004 (in Japanese).

[20] T. Hori, K. Cho, Y. Takatori, and K. Nishimori, "Key techniques for actualizing smart antenna hardware," Antennas and Propagation Society International Symposium 2000, IEEE, Vol. 2, pp. 948-952, July 2000.

[21] K. Uehara, Y. Takatori, and K. Kagoshima, "Proposal for a novel beam combination method for indoor high-speed wireless communication systems," Vehicular Technology, IEEE Transactions on Vol. 48, Issue 5, pp. 1511-1517, Sep. 1999.

[22] K. Nishimori, K. Cho, Y. Takatori, and T. Hori, "A Novel Configuration for Realizing Automatic Calibration of Adaptive Array Using Dispersed SPDT Switches for TDD Systems," IEICE Vol. E84-B, No. 9, pp. 2516-2522, Sep. 2001.

[23] K. Nishimori, and K. Cho, "Evaluation of SDMA employing vertical pattern and polarization control in actual cellular environment measurement," Vehicular Technology Conference 2004, VTC 2004-Spring, 2004 IEEE 59th Vol. 1, pp. 244-248, May 2004.

[24] A. Paulraj, R. Nabar, and D. Gore, "Introduction to space-time wireless communications," Cambridge, U.K., Cambridge Univ. Press, 2003.

[25] M. Ueba, K. Ohata, J. Mitsugi, and M. Umehira, "Broadband and scalable mobile satellite communication system for future access networks," 22nd AIAA ICSSC, AIAA 2004-3154.

Koichi Tsunekawa

Senior Manager, Executive Research Engineer, Wireless Systems Innovation Laboratory, NTT Network Innovation Laboratories.

He received the B.S., M.S., and Ph.D. degrees in engineering science from the University of Tsukuba, Tsukuba, Ibaraki in 1981, 1983, and 1992, respectively. He joined the Electrical Communications Laboratories of Nippon Telegraph and Telephone Public Corporation (now NTT), Tokyo in 1983. Since 1984, he has been engaged in R&D of portable telephone antennas in land mobile communication systems. From 1993 to 2003, he was with NTT DoCoMo, Inc. where he worked on radio propagation research, intelligent antenna systems for wireless communications, and the development of IMT-2000 antenna systems. His research interests are antenna systems for MIMO transmission, millimeter-wave broadband access, and ubiquitous wireless system. He is a member of IEEE and IEICE.

Communications and Network, 2016, 8, 9-21
Published Online February 2016 in SciRes. http://www.scirp.org/journal/cn
http://dx.doi.org/10.4236/cn.2016.81002

The Roles of 5G Mobile Broadband in the Development of IoT, Big Data, Cloud and SDN

Bao-Shuh Paul Lin[1,2], **Fuchun Joseph Lin**[1,2], **Li-Ping Tung**[1]

[1]Microelectronics & Information Research Center, National Chiao Tung University, Taiwan
[2]College of Computer Science, National Chiao Tung University, Taiwan
Email: bplin@mail.nctu.edu.tw, fjlin@nctu.edu.tw, lptung@nctu.edu.tw

Received 3 November 2015; accepted 22 February 2016; published 25 February 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

The fast technology development of 5G mobile broadband (5G), Internet of Things (IoT), Big Data Analytics (Big Data), Cloud Computing (Cloud) and Software Defined Networks (SDN) has made those technologies one after another and created strong interdependence among one another. For example, IoT applications that generate small data with large volume and fast velocity will need 5G with characteristics of high data rate and low latency to transmit such data faster and cheaper. On the other hand, those data also need Cloud to process and to store and furthermore, SDN to provide scalable network infrastructure to transport this large volume of data in an optimal way. This article explores the technical relationships among the development of IoT, Big Data, Cloud, and SDN in the coming 5G era and illustrates several ongoing programs and applications at National Chiao Tung University that are based on the converging of those technologies.

Keywords

5G, Internet of Things (IoT), Software Defined Networks (SDN), Big Data Analytics, Cloud Computing

---

1. Introduction

The roles of SDN, Cloud, IoT, and Big Data in 5G Networks have raised great interest recently [1]. According to the IEEE Computer Society (IEEE CS 2022 Report [2]), Cloud Computing (Cloud), Big Data Analytics (Big Data), Internet of Things (IoT), and Software Defined Networks (SDN) are among 4 of 20+ emerging technologies as illustrated in **Table 1**. In the meantime, based on the report in IEEE Communications Society (IEEE ComSoc Technology News [3]), among top 10 trends in 2015 as listed in **Table 2**, 5G, Virtualization (SDN & NFV), everywhere connectivity for IoT & IoE, and Big Data are also included. Combining both reports, we can identify 5G, Cloud, IoT/IoE, Big Data, and SDN as the five most worthwhile ICTs (information & communica-

How to cite this paper: Lin, B.-S.P., Lin, F.J. and Tung, L.-P. (2016) The Roles of 5G Mobile Broadband in the Development of IoT, Big Data, Cloud and SDN. *Communications and Network*, **8**, 9-21. http://dx.doi.org/10.4236/cn.2016.81002

Table 1. IEEE computer society 2022 report.

Technology & Indexing Terms		
1. Security Cross-Cutting Issues	9. Multicore	17. 3D Printing
2. Open Intellectual Property Movement	10. Photonics	18. Big Data and Analytics
3. Sustainability	11. Networking and Interconnectivity	19. Machine Learning and Intelligent Systems
4. Massively Online Open Courses	12. Software Defined Networks	20. Life Sciences
5. Quantum Computing	13. High Performance Computing	21. Computational Biology and Bioinformatics
6. Device and Nanotechnology	14. Cloud Computing	22. Robotics
7. 3D Integrated Circuits	15. Internet of Things	
8. Universal Memory	16. Natural User Interfaces	

Table 2. IEEE communications society top 10 trends 2015.

1)	5G
2)	FIBER EVERYWHERE
3)	VIRTUALIZATION, SDN & NFV
4)	EVERYWHERE CONNECTIVITY FOR IoT & IoE
5)	BIG DATA, COGNITIVE NETWORKS
6)	CYBERSECURITY
7)	GREEN COMMUNICATIONS
8)	SMARTER SMARTPHONES, CONNECTED SENSORS
9)	NETWORK NEUTRALITY, INTERNET GOVERNANCE
10)	MOLECULAR COMMUNICATIONS

tions technologies) to watch out up to 2020 in term of their potential, convergence and applications. **Figure 1** depicts the overlapping of these five major ICT trends. It also illustrates how the infrastructure deployment of 5G mobile broadband and the architectural integration of Cloud Computing strongly impact the development of IoT, Big Data, and SDN. This paper explores the technical relationships of these five areas and discusses several ongoing programs and applications currently under development at NCTU based on these technologies.

2. The Roles of 5G, IoT, Big Data, Cloud, and SDN till 2020

Although so far the 5G mobile broadband requirements and standard specifications are not ready yet, 5G technology research & development are already started and some 5G features or subsystems are readily available. By Year 2020 the commercial 5G will be available and IoT applications will be deployed everywhere with mobile broadband technology. Moreover, the Big Data generated by IoT applications will become a norm and Cloud will be largely utilized to compute, store and virtualize network functions (NFV). Also, the underlying network infrastructure will adopt SDN to reduce both capital expense (CAPEX) and operational expense (OPEX). **Figure 2** further illustrates the roles of 5G, IoT, Big Data, Cloud and SDN and their relationships. This figure is modified based on the reference [4] by inserting 5G mobile broadband in the center.

3. Technical Relationships among IoT, Big Data, Cloud, & SDN in 5G Era

Based on **Figure 2**, we develop **Figure 3** to better explain the technical relationships among IoT, Big Data, Cloud, and SDN in the 5G mobile broadband services (5G MBS) [5]. First, IoT is capable of generating Big Data with four Vs: volume, velocity, variety and veracity. Then, Cloud is brought in for Big Data storage and processing. Finally, SDN is employed to provide more efficient and flexible networks for inter-Cloud data transport. Out of Big Data, Cloud, and SDN, advanced technologies such as machine learning analytics, Cloud

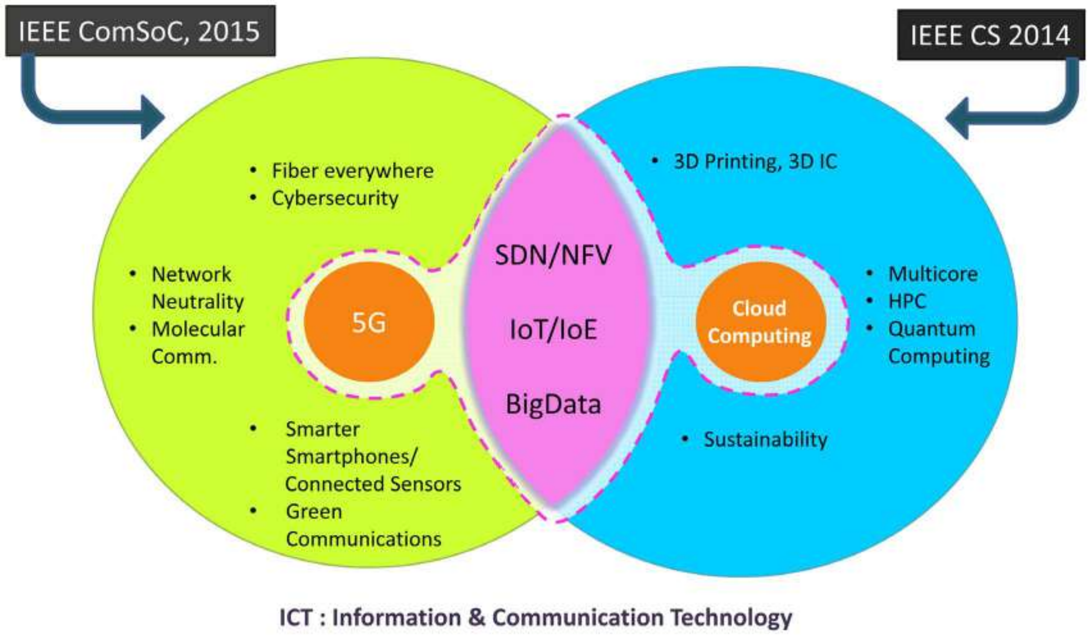
IEEE ComSoC, 2015
IEEE CS 2014
Fiber everywhere
Cybersecurity
Network Neutrality
Molecular Comm.
5G
Smarter Smartphones/ Connected Sensors
Green Communications
SDN/NFV
IoT/IoE
BigData
3D Printing, 3D IC
Cloud Computing
Multicore
HPC
Quantum Computing
Sustainability
ICT : Information & Communication Technology

Figure 1. ICT major trends for 2015-2020.

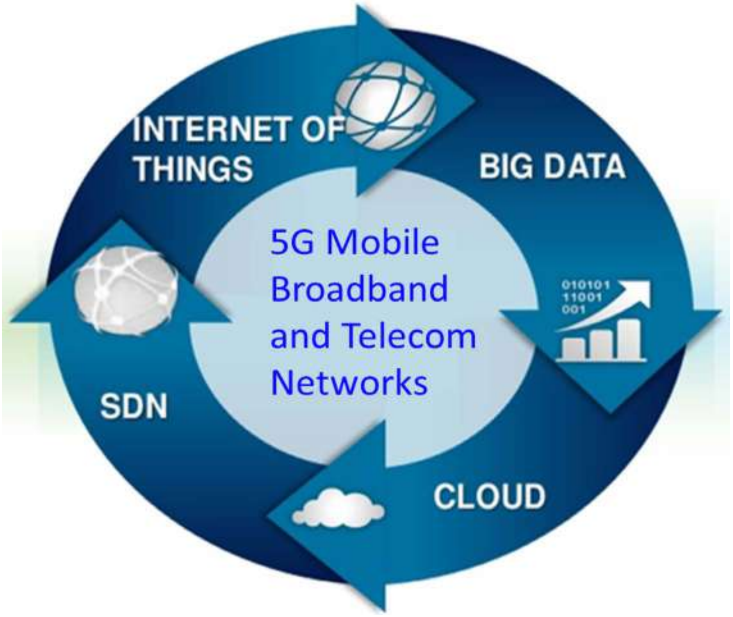
INTERNET OF THINGS
BIG DATA
5G Mobile Broadband and Telecom Networks
010101 11001 001
SDN
CLOUD

Figure 2. Relationship among 5G, IoT, Big Data, Cloud and SDN based on [4].

RAN and softwarized 5G then are developed.

As illustrated in **Figure 4**, 5G will serve as a better gateway and transport network for IoT applications so that IoT data can be delivered more efficiently and economically. In addition, IoT will become one of the major sources of Big Data by producing large volume, fast velocity, and many varieties of data [6] as illustrated in **Figure 5**.

Finally, **Figure 6** shows that Cloud can be adopted in the 5G Radio Access Network (RAN) and turns it to a Cloud-based RAN (C-RAN). Both SDN and NFV have been applied to data center in the cloud to enable better load balance and resource allocation of the cloud. SDN has also been applied to 5G mobile broadband core networks to enable smart routing, better traffic management and improve network resource utilization.

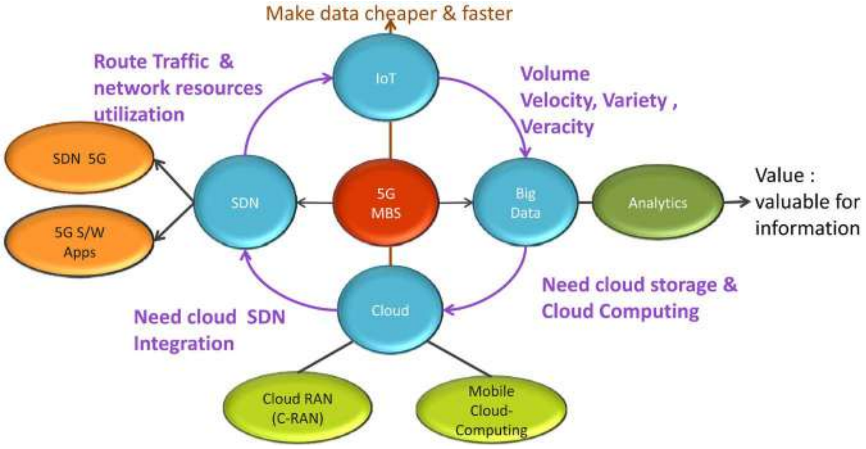
Make data cheaper & faster
Route Traffic & network resources utilization
IoT
Volume Velocity, Variety , Veracity
SDN 5G
5G S/W Apps
SDN
5G MBS
Big Data
Analytics
Value : valuable for information
Need cloud SDN Integration
Cloud
Need cloud storage & Cloud Computing
Cloud RAN (C-RAN)
Mobile Cloud-Computing

Figure 3. Technical relationships among 5G & IoT, Big Data, Cloud, and SDN.

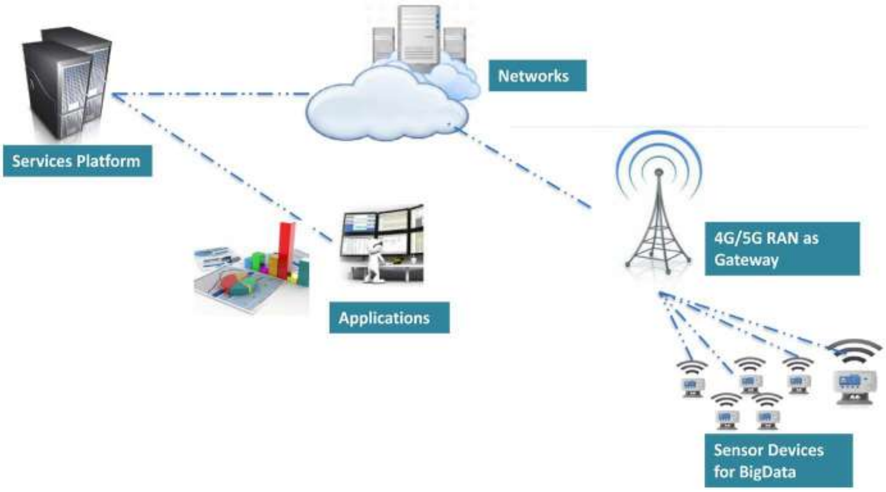
Networks
Services Platform
4G/5G RAN as Gateway
Applications
Sensor Devices for BigData

Figure 4. 4G/5G as the gateway for IoT applications.

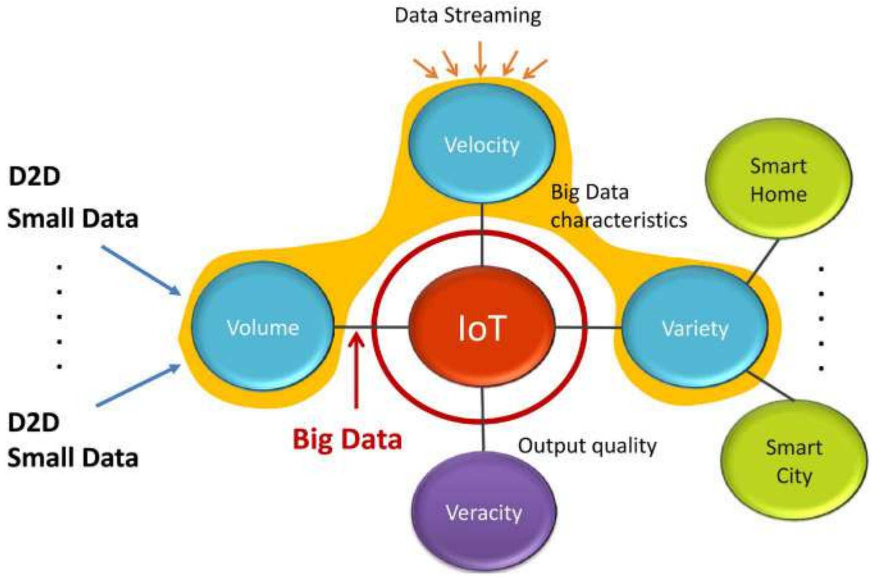
Data Streaming
D2D Small Data
Velocity
Big Data characteristics
Smart Home
Volume
IoT
Variety
D2D Small Data
Big Data
Output quality
Smart City
Veracity

Figure 5. Relationships between IoT and Big Data.

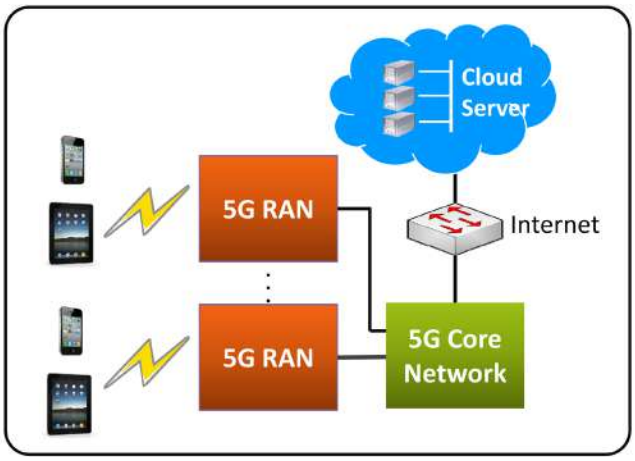

Figure 6. C-RAN architecture in 5G mobile network.

4. Ongoing Programs and Applications at National Chiao Tung University (NCTU)

In this section, we describe five ongoing programs and applications at NCTU that are closely related to the SDN, Cloud, Big Data, IoT, and 5G technologies [7]-[17].

4.1. Program on SDN-Enabled Cloud-Based Broadband and Wireless Network Technologies and Services

This program is a national program to support the SDN Industry-Academia Cooperation, led by both NCTU and CHT (Chunghwa Telecom, the largest mobile & telecom operator in Taiwan) with 5 other networking and communications companies in Taiwan. This is a multi-year program with the target to set up a testbed for end-to-end testing and to help establish an ecosystem for local SDN industry.

Figure 7 shows the scope of this program that covers mobile access, WiFi access, broadband core and data center cloud end-to-end application systems.

Figure 8 further shows both the structure of the program and its network configuration across NCTU, CHT and another university National Tsing Hwa University (NTHU).The program includes research and development involving 4G/LTE, B4G/5G, SDN, Cloud, SDN for Wi-Fi, and SDN for WAN technologies. There are five sub-programs: 1) SDN broadband network technologies and services, 2) SDN mobile/wireless network technologies and services, 3) SDN and cloud integration services and management, 4) SDN switch-related devices and systems and 5) SDN system integration and field trials.

The program is intended to construct an SDN network in NCTU, CHT and NTHU, respectively. These SDN networks then will be interconnected to form a wide area SDN network in Taiwan. Eventually, we plan to have this end-to-end system connected to the global network. The SDN switches deployed in this testbed will be mostly small or medium scale due to their experimental nature. Also included in this testbed are SDN-based Wi-Fi access points. Cloud-based data centers will be deployed in each location of NCTU, CHT and NTHU. Three types of SDN controllers including OpenDaylight, Ryu and Floodlight are currently under trial in the experimental network. Our architecture assumes the scenario of multiple SDN controllers with a hierarchical topology on which useful SDN APPs such as load balancer, network optimizer for video delivery, network co-ordinator (including visualization), dynamic flow configuration, end-to-end service configuration, dynamic policy-based traffic engineering, and multitenant network automation can be developed and deployed.

4.2. Program on Big Data Analytics for Network Traffic and Management Data

In this program, we address the network performance issues with two experimental networks: BML at NCTU campus and ITRINET of ITRI.

The architecture of the BML experimental network is illustrated in the lower part of **Figure 9** that consists of a 4G RAN, a 4G Core and a Cloud environment. For the 4G RAN, both indoor and outdoor environment are taken into consideration. The left upper part of **Figure 9** shows the scope of the ITRINET experimental network

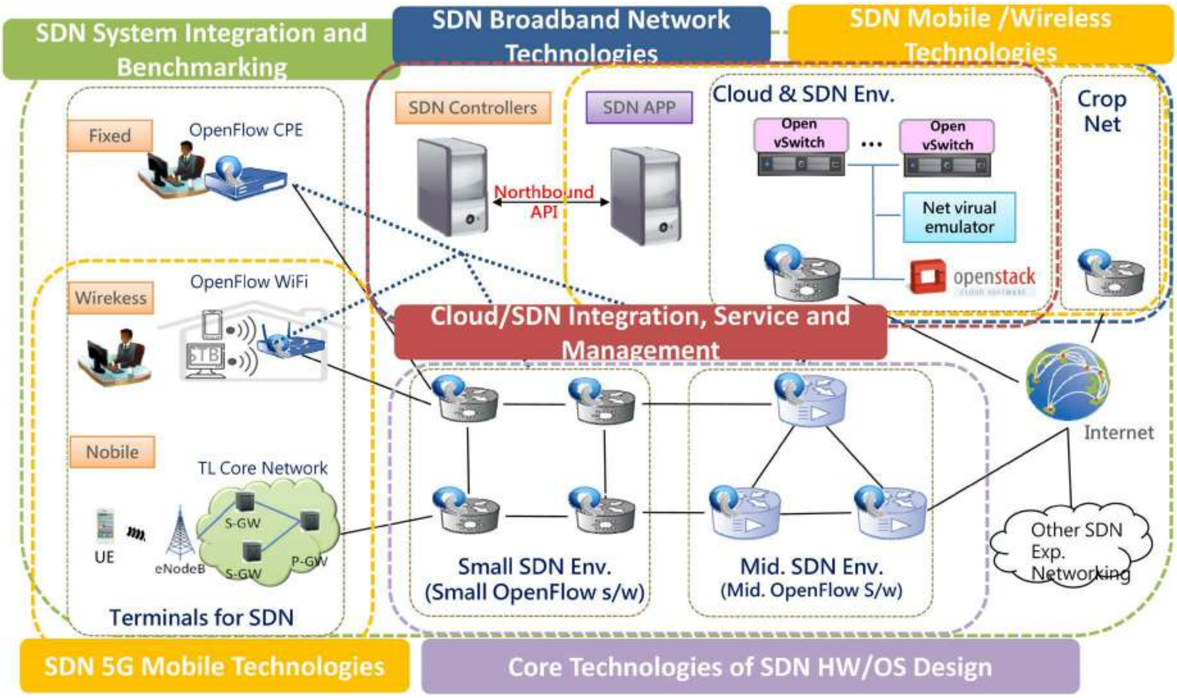
SDN System Integration and Benchmarking
SDN Broadband Network Technologies
SDN Mobile /Wireless Technologies
SDN Controllers
SDN APP
Northbound API
Cloud & SDN Env.
Open vSwitch
Open vSwitch
Net virual emulator
openstack
Crop Net
Fixed
OpenFlow CPE
Wirekess
OpenFlow WiFi
STB
Cloud/SDN Integration, Service and Management
Nobile
TL Core Network
S-GW
S-GW
P-GW
UE
eNodeB
Terminals for SDN
Small SDN Env. (Small OpenFlow s/w)
Mid. SDN Env. (Mid. OpenFlow S/w)
Internet
Other SDN Exp. Networking
SDN 5G Mobile Technologies
Core Technologies of SDN HW/OS Design

Figure 7. Applications of SDN & cloud in 4G/LTE & 5G wireless networks.

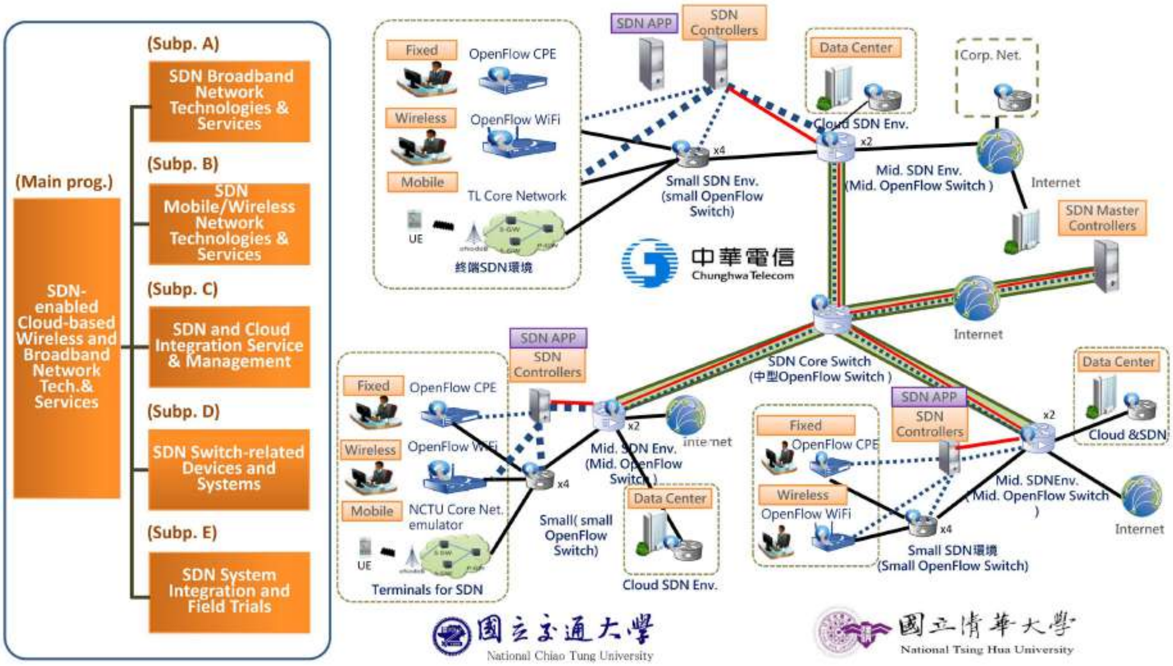
(Main prog.)
SDN-enabled Cloud-based Wireless and Broadband Network Tech.& Services
(Subp. A)
SDN Broadband Network Technologies & Services
(Subp. B)
SDN Mobile/Wireless Network Technologies & Services
(Subp. C)
SDN and Cloud Integration Service & Management
(Subp. D)
SDN Switch-related Devices and Systems
(Subp. E)
SDN System Integration and Field Trials
Fixed
OpenFlow CPE
Wireless
OpenFlow WiFi
Mobile
TL Core Network
UE
終端SDN環境
SDN APP
SDN Controllers
x4
Small SDN Env. (small OpenFlow Switch)
Data Center
Cloud SDN Env.
x2
Mid. SDN Env. (Mid. OpenFlow Switch)
Corp. Net.
Internet
SDN Master Controllers
中華電信
Chunghwa Telecom
Internet
SDN Core Switch (中型OpenFlow Switch)
SDN APP
SDN Controllers
Fixed
OpenFlow CPE
Wireless
OpenFlow WiFi
Mobile
NCTU Core Net. emulator
UE
Terminals for SDN
x2
Internet
Mid. SDN Env. (Mid. OpenFlow Switch)
x4
Small(small OpenFlow Switch)
Data Center
Cloud SDN Env.
Fixed
OpenFlow CPE
Wireless
OpenFlow WiFi
SDN APP
SDN Controllers
x2
Data Center
Cloud &SDN
x4
Small SDN環境 (Small OpenFlow Switch)
Mid. SDNEnv. (Mid. OpenFlow Switch)
Internet
國立交通大學
National Chiao Tung University
國立清華大學
National Tsing Hua University

Figure 8. SDN technical R&D center and projects at NCTU.

that covers a much greater area of Hsinchu County than NCTU and includes additionally ITRI (Industrial Technologies Research Institute), HSIP (Hsinchu Science Industrial Park) and THR (Taiwan High-speed Rail, Hsinchu station). The right upper part of **Figure 9** illustrates how ITRINET covers various R&D buildings in the corporate campus of ITRI such as B11, B12, B51, etc.

Figure 10 illustrates how Big Data analytics based on InfoSphere or Spark is performed on ITRINET for the purpose of network optimization. First, network traffic measurement and network management data are collected.

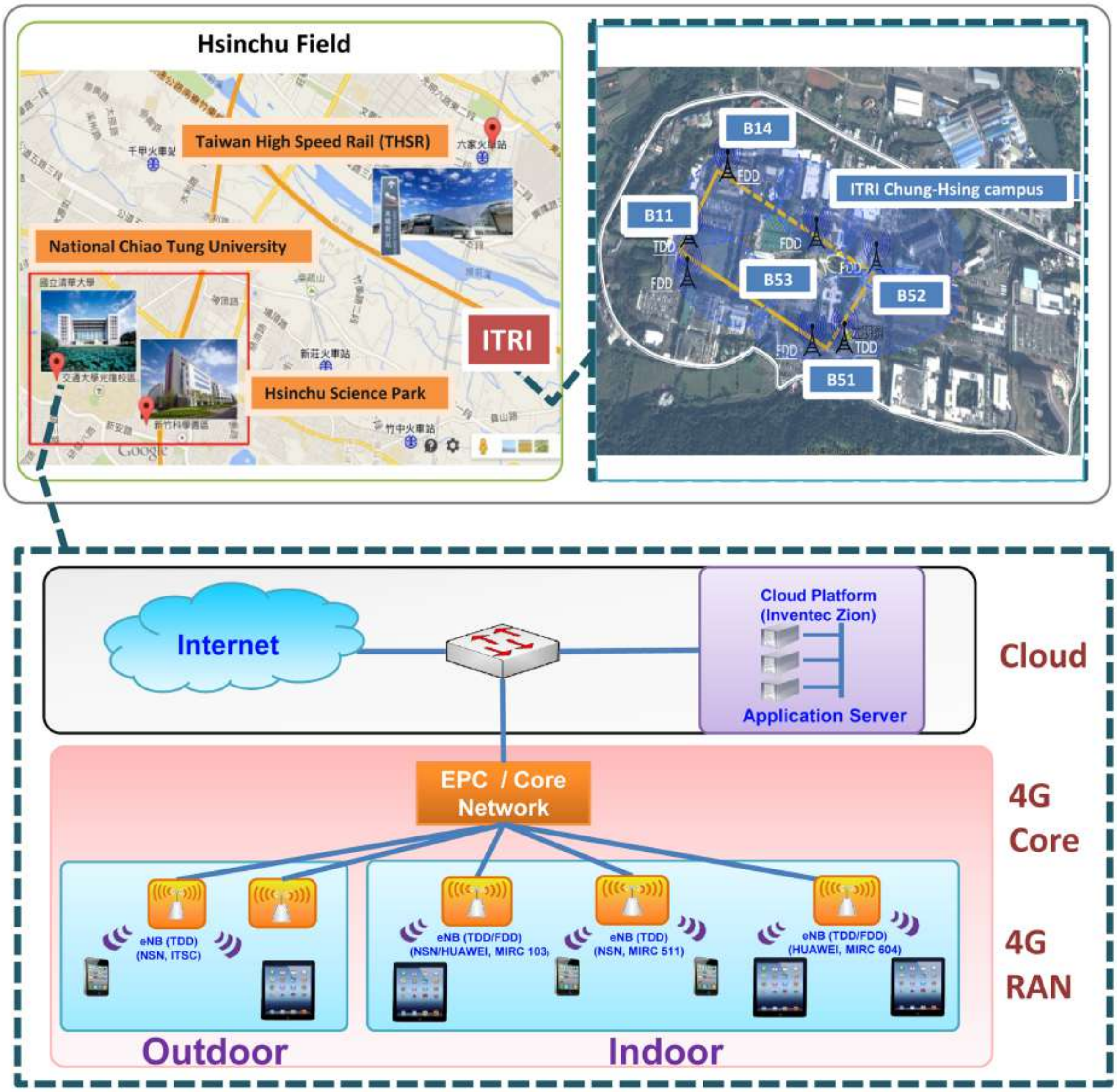
Hsinchu Field
Taiwan High Speed Rail (THSR)
National Chiao Tung University
ITRI
Hsinchu Science Park
ITRI Chung-Hsing campus
B14
B11
B53
B52
B51
FDD
TDD
Internet
Cloud Platform (Inventec Zion)
Application Server
Cloud
EPC / Core Network
4G Core
eNB (TDD) (NSN, ITSC)
eNB (TDD/FDD) (NSN/HUAWEI, MIRC 103)
eNB (TDD) (NSN, MIRC 511)
eNB (TDD/FDD) (HUAWEI, MIRC 604)
4G RAN
Outdoor
Indoor

Figure 9. Greater Hsinchu 4G/LTE & Future 5G experimental network ITRINET (include ITRI, NCTU, THR & HSIP).

Then, data analytics methods based on machine learning, data mining and statistical modeling are applied to analyze the collected data. Finally, we apply the results thus generated to network performance evaluation and optimization by providing a feedback loop for system re-configuration. The whole operational cycle includes the technologies of 4G/LTE, B4G/5G, Big Data Analytics, Cloud, and Traffic Engineering.

4.3. Application on IoT Platform Integrated with Data Generation and Data Analytics

This is an application where we set up an IoT platform integrated with data generator and data analytics capabilities as illustrated in **Figure 11**. A common challenge for IoT/M2M service providers is how to test their large scale IoT/M2M applications with the near realistic data that the system will handle in a production environment. As such tests may involve not only a large number but also a large variety of sensors, deploying a testing environment that contains all the necessary sensors turns out to be an infeasible, if not impossible job. To tackle this problem, we develop a data generation method (illustrated in the lower part of **Figure 11**) based on streams generation capabilities of IBM InfoSphere and Spark to emulate data from a large number and a large variety of sensors. Such generated data will be sent into and processed by the applications residing on the IoT/M2M

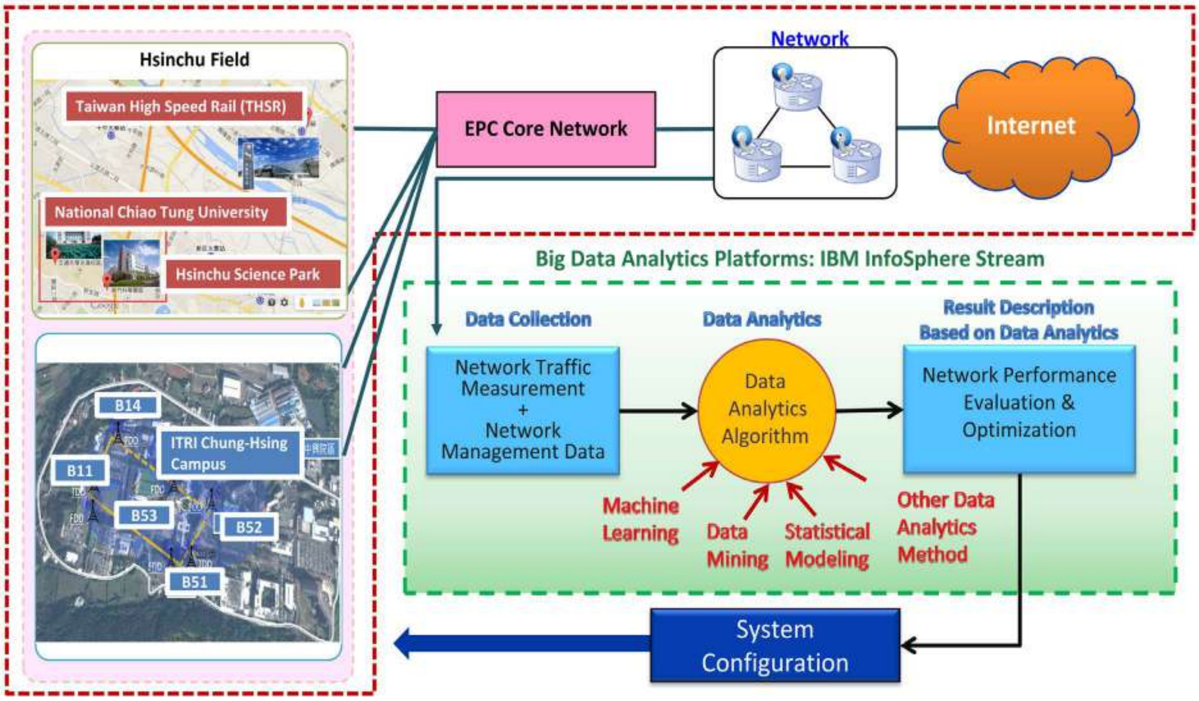
Hsinchu Field
Taiwan High Speed Rail (THSR)
National Chiao Tung University
Hsinchu Science Park
B14
ITRI Chung-Hsing Campus
B11
B53
B52
B51
EPC Core Network
Network
Internet
Big Data Analytics Platforms: IBM InfoSphere Stream
Data Collection
Data Analytics
Result Description Based on Data Analytics
Network Traffic Measurement + Network Management Data
Data Analytics Algorithm
Network Performance Evaluation & Optimization
Machine Learning
Data Mining
Statistical Modeling
Other Data Analytics Method
System Configuration

Figure 10. Big Data analytics for Traffic engineering of ITRINET.

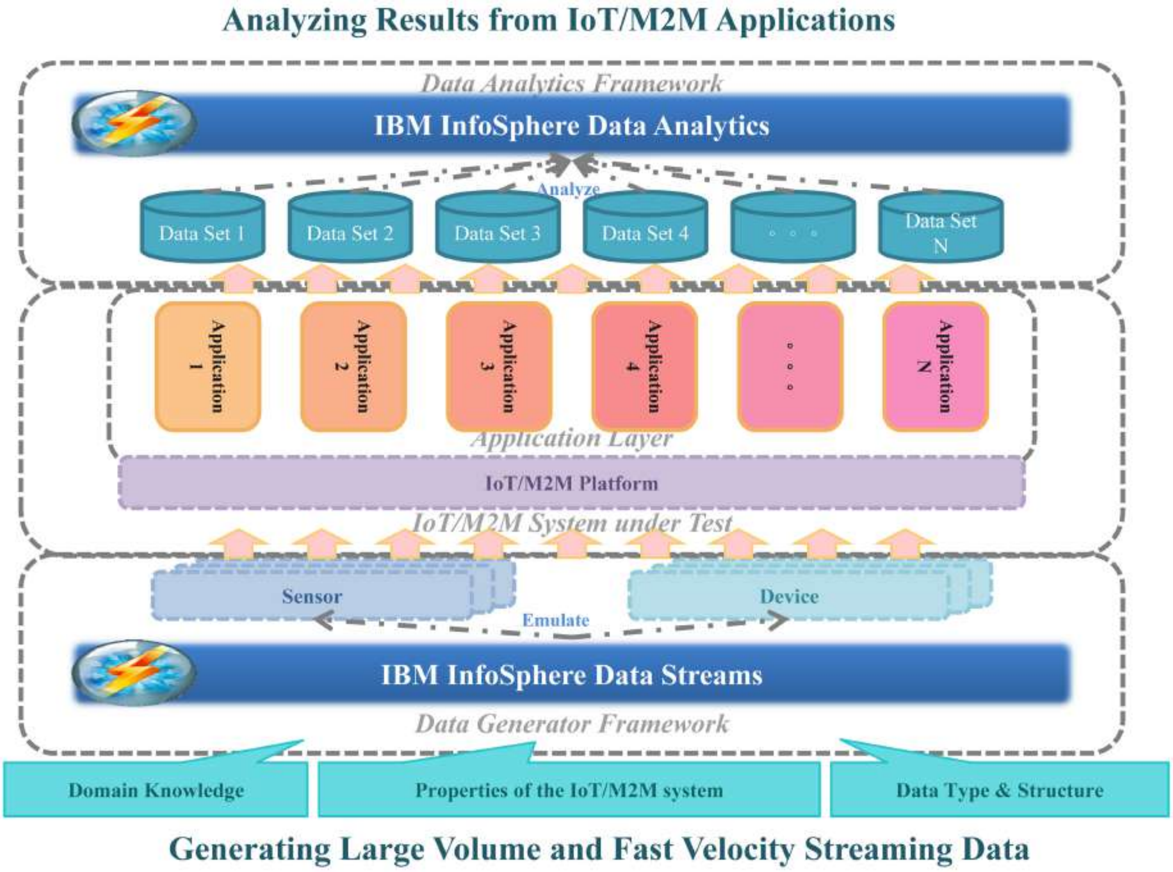
Analyzing Results from IoT/M2M Applications
Data Analytics Framework
IBM InfoSphere Data Analytics
Analyze
Data Set 1
Data Set 2
Data Set 3
Data Set 4
◦ ◦ ◦
Data Set N
Application 1
Application 2
Application 3
Application 4
◦ ◦ ◦
Application N
Application Layer
IoT/M2M Platform
IoT/M2M System under Test
Sensor
Device
Emulate
IBM InfoSphere Data Streams
Data Generator Framework
Domain Knowledge
Properties of the IoT/M2M system
Data Type & Structure
Generating Large Volume and Fast Velocity Streaming Data

Figure 11. IoT platform integrated with data generation and data analytics.

platform (illustrated in the middle part of **Figure 11**). Finally, the data sets produced by the IoT/M2M applications are analyzed by data analytics engines in IBM InfoSphere and Spark (illustrated in the upper part of **Figure 11**).

The effectiveness of our data generation methodology is first verified in a people management system of a large factory environment. With our methodology, we are able to test our factory management IoT/M2M application more efficiently without actually deploying a large number and a large variety of sensors. This application development uses both IoT and Big Data Analytics.

4.4. Application on Data Analytics for Traffic Generated by Testing & Measurement

The automatic Testing & Measurement (T&M) of DUT (Device under Test, including 4G/LTE SD, UE/CPE) for 4-Stage (namely, conformance, interoperability, operator-IOT and field trials) testing is shown in **Figure 12**.

We apply the data analytics learning techniques for the classification of the DUT types by series numbers or id. Once the type is identified, the parameters for feature extraction can be measured by the testing and measurement equipment. An example for DUT classification is shown in **Table 3**. In this example, the DUT type relational table is a 4 × 4 table with 4 fields: 1) DUT Device Type, 2) UE capability Information (FGI), 3) DNS (TTL), and 4) HTTP (User Agent). This device relational table is created by applying multivariate statistical analytics on the network traffic flow and LTE UE connection related information. Then by extracting specific features, we are able to identify the types of DUT such as Dongle, Smartphone, CPE, etc.

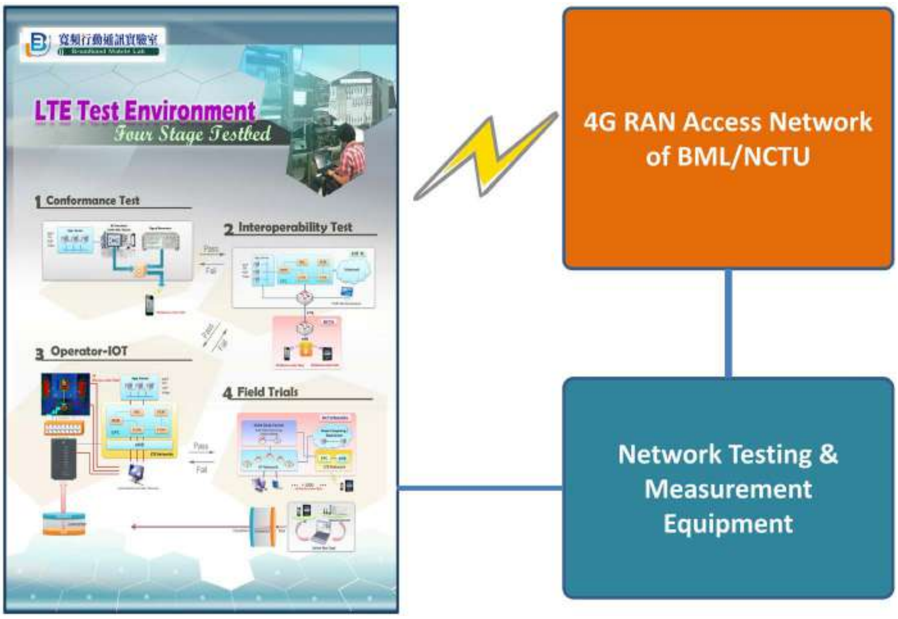

Figure 12. Data generation by endpoints from testing & measurement.

Table 3. DUT relational table.

DUT Device Type	UE FGI	DNS-TTL	HTTP-User Agent
Dongle A	0100 0110 0000 0101 0001 1000 0000 0000	128	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.134 Safari/537.36\r\n
Dongle B	0101 1111 0000 1111 1111 1100 1000 0000	63	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.134 Safari/537.36\r\n
Smart Phone	0101 1110 0000 1101 1101 1000 1000 0000	43	iPhone6,2/8.3 (12F70)\r\n
CPE	0101 1111 0000 1111 1111 1100 1000 0000	127	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.134 Safari/537.36\r\n

Another example of applying data analytics in T&M traffic is to evaluate LTE QoS. When evaluating LTE QoS, it is required to setup testing scenarios with different QoS levels. Traffic Flow Template (TFT) (as shown in **Table 4**) is designed to filter packets into correspondent bearers, either a default bearer or dedicated bearer, in LTE network. Each bearer has its own QoS level. We can establish a dedicated bearer of guarantee bit rate (GBR) for applications such as VoIP, or a default bearer of basic QoS level for applications like file transfer.

What TFT does is to filter packets into correspondent bearers according to packets' IP address, port number, protocol, direction. The information, however, is insufficient to differentiate among web file download, FB news browsing, and Line chat. We are unable to give different QoS levels to these applications. To tackle this problem, we propose a new architecture which integrates deep packet inspection (DPI) with TFT to provide a higher granularity of QoS levels for applications. When some traffic flow has been identified as a certain application type, TFT will be informed to update rules of packet filters. With this improved method, the traffic flow can be delivered through the bearer of suitable QoS level.

4.5. Applications on Data Analytics for Traffic Flow Created by APP's of SD

To fulfill the Quality of Service (QoS) requirements from users, it is important to make effective use of the network resource. We can optimize the performance of a network by applying data analytics to traffic engineering. In particular, it is important to classify mobile applications traffic intended by the user with data analytics.

We propose a HMM-based (Hidden Markov Model) model to classify the mobile applications. By surveying related work, we have realized that there are different handshake patterns of well-known application protocols. Also, according to the observation of some specific mobile applications, we discover that every mobile Internet service has its unique negotiation process at the beginning when service starts. HMM was widely used to recognize time-dependent sequences and find out unknown patterns of data, such as speech recognition, handwriting recognition.

Table 4. Traffic Flow Template (TFT): Use to specify the packet filters associated.

UL Packet Filter ID	DRB	Packet Filter Evaluation Precedence	Protocol Number (IPv4)/ Next Header (IPv6)	Remote Address and Subnet Mask	Single Local Port (UE)	Local Port Range (UE)	Single Remote Port Range (NW)	Remote Port Range (NW)	IPSec SPI Range	Type of Service (IPv4)/ Traffic Class (IPv6) and Mask	Flow Label (IPv6)
1	DRB2	6	6 (TCP)	IPv4: 172.168.8.0 [255.255.255.0] IPv6: 2001:0ba0::[ffff:ffff::]	60051	-	-	-	-	-	-
2	DRB2	2	17 (UDP)	-	-	-	60201	-	-	-	-
3	DRB3	7	6 (TCP)	IPv4: 172.168.8.0 [255.255.255.0] IPv6: 2001:0ba0::[ffff:ffff::]	-	60100:60200	-	-	-	-	-
4	DRB3	3	17 (UDP)	-	-	-	-	60300: 60400	-	-	-
5	DRB3	5	50 IPSec (ESP)	-	-	-	-	-	0x0F80F0000	-	-
6	DRB3	1	-	-	-	-	-	-	-	00101000, Mask = 11111100	-
7	DRB3	6	-	-	-	-	-	-	-	-	-
8	DRB1 (default bearer)	255	6 (TCP)	-	-	-	-	-	-	-	-

In our method, we extract the packet size sequence and packet transmission direction sequence of the first 20 packets to train the HMM model. **Figure 13** shows our designed model structure of HMM. X_n is the hidden variable. It represents the transmission states which cannot be observed directly. Because of the unknown transmission states, we need to use observation variable O_n^1 and O_n^2 as training features to build the mobile application models. Here, "*n*" is the sequence number of packets in a traffic flow. O_n^1 is the observation symbol of packets transmission direction. O_n^2 is the observation of packet size that quantizes to a certain range of packet size. We quantize smallest packet size to number 1 and largest packet size to number 8. The rest of packet sizes are then divided into six groups.

Figure 14 shows the main process of our classification system. First, we process the collected packets by reading from original PCAP files, and extract the necessary field of packet header, including source IP, destination IP, source port number, destination port number, packet timestamp, and packet size. Second, we use packet

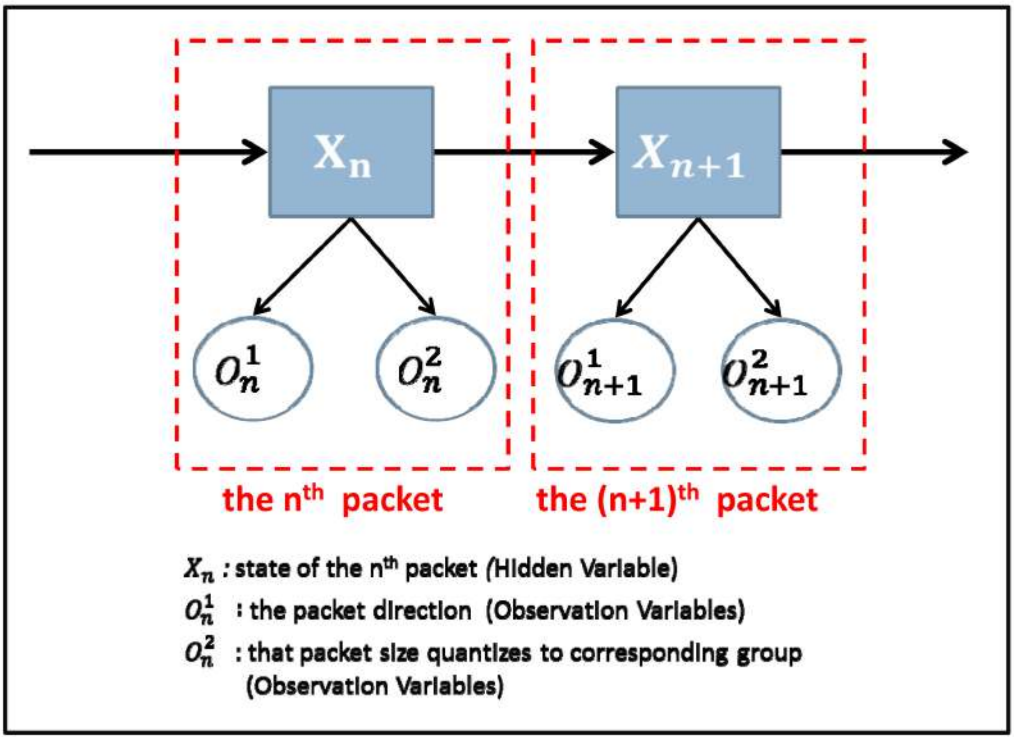
X_n
X_{n+1}
O_n^1
O_n^2
O_{n+1}^1
O_{n+1}^2
the nth packet
the (n+1)^th packet
X_n : state of the nth packet (Hidden Variable)
O_n^1 : the packet direction (Observation Variables)
O_n^2 : that packet size quantizes to corresponding group (Observation Variables)

Figure 13. Classification model of Internet application traffic.

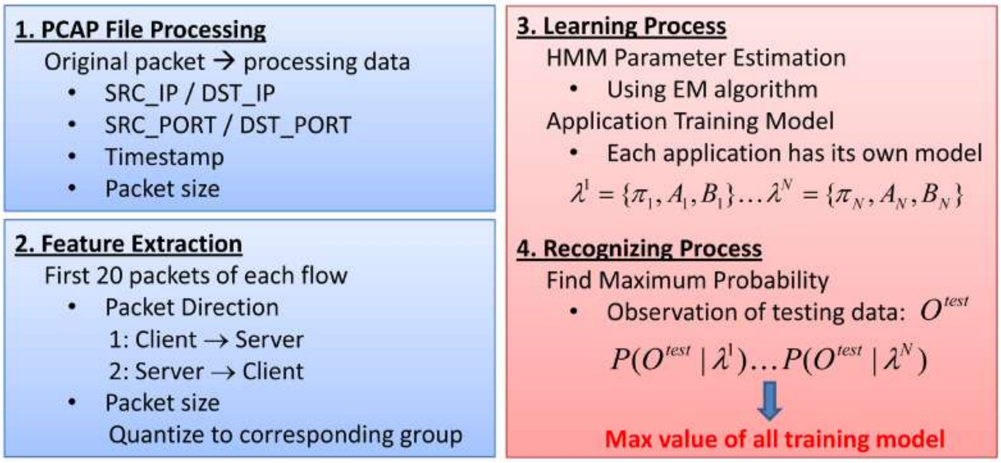

1. PCAP File Processing
Original packet → processing data
- SRC_IP / DST_IP
- SRC_PORT / DST_PORT
- Timestamp
- Packet size

2. Feature Extraction
First 20 packets of each flow
- Packet Direction
  1: Client → Server
  2: Server → Client
- Packet size
  Quantize to corresponding group

3. Learning Process
HMM Parameter Estimation
- Using EM algorithm

Application Training Model
- Each application has its own model

$\lambda^1 = \{\pi_1, A_1, B_1\} \ldots \lambda^N = \{\pi_N, A_N, B_N\}$

4. Recognizing Process
Find Maximum Probability
- Observation of testing data: O^{test}

$P(O^{test} \mid \lambda^1) \ldots P(O^{test} \mid \lambda^N)$

Max value of all training model

Figure 14. The process of our proposed classification system.

header information (5-tuple: source IP, destination IP, source port number, destination port number, and protocol) to process packets with the same 5-tuple information into a unit of mobile application traffic flow, and extract the features of each application flow. Third, in learning process, we use the extracted features which are quantized into corresponding symbol sequence to training the HMM-based application models for different applications. Finally, we can identify the new traffic flow by finding out the maximum value of log likelihoods derived from different application models.

5. Conclusion

This paper presents a forward looking view of the convergence of IoT, big data, cloud, SDN technologies along with the arrival of 5G mobile broadband networks. We intend to demonstrate the technical relationships of those technologies and the compelling programs and applications that can be created when two, three or more of those technologies converge. Due to the nature of fast evolution of ICT and the ongoing innovation of those five technologies, this paper should be updated on annual basis to keep the related information up to day with the ICT major trends.

Acknowledgements

This work was supported by the Ministry of Science and Technology (MOST) of Taiwan and National Chiao Tung University under grants: MOST 103-2622-E-009-012-, 103-2221-E-009-138-, 103-2218-E-009-032-, 104-2221-E-009-023- and NCTU-ICTL 104Q707.

References

[1] Lin, B-S.P. (2015) The Roles of SDN, Cloud, IoT, & Big Data, in 5G Networks and Their Technical Relationships & Applications. Invited Talk, IEEE VTS APWCS 2015, Singapore.

[2] Alkhatib, H., *et al*. (2015) IEEE CS 2022 Report. IEEE Computer Society, IEEE.

[3] Neira, E.M. (2015) IEEE ComSoc CTN Special Issue on Ten Trends That Tell Where Communication Technologies are Headed in 2015.

[4] Tucker, L. (2013) On the Cloud of the Future. Cloud Computing Summit, San Jose.

[5] Kleiner, T. (2015) 5G Research in Horizon 2020.

[6] Thadani, A. and Andreoli, P. (2014) Realizing Business Value from Convergence of IoT + Big Data Technologies. 2014 *Midwest Architecture Community Collaboration* (*MACC*) *Conference*, Minneapolis, 13 November 2014.

[7] Dinh, H.T., Lee, C., Niyato, D. and Wang, P. (2013) A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches. *Wireless Communications and Mobile Computing*, **13**, 1587-1611. http://dx.doi.org/10.1002/wcm.1203

[8] Lin, B-S.P., *et al.* (2013) The Design of Cloud-based 4G/LTE for Mobile Augmented Reality with Smart Mobile Device. *Proceedings of* 2013 *IEEE International Symposium on Mobile Cloud, Computing, and Service Engineering* (*IEEE Mobile Cloud* 2013), Redwood City, 25-28 March 2013, 561-566. http://dx.doi.org/10.1109/sose.2013.57

[9] Lin, B.-S.P., Tung, L-P., Hsieh, I-C., Liu, T-H. and Chou, S-Y. (2015) The Design of Big Data Analytics for Testing & Measurement and Traffic Flow on an Experimental 4G/LTE Network. *Proceedings of IEEE WOCC* 2015, Taipei, 23-24 October 2015, 40-44. http://dx.doi.org/10.1109/wocc.2015.7346173

[10] Lin, B-S.P., Tung, L-P., Hsieh, I-C. and Chou, S-Y. (2015) Performance Estimation of MAR for Outdoor Navigation Applications on an 5G Mobile Broadband by Mobile Smart Devices. *Proceedings of IEEE APWCS* 2015, Singapore, 19-20 August 2015, 1-6.

[11] Lin, F.J., Ren, Y. and Cerritos, E. (2013) A Feasibility Study on Developing IoT/M2M Applications over ETSI M2M Architecture. *Proceedings of the* 1st *International Workshop on Internet of Things Technologies* (*IoTT*), Seoul, 15-18 December 2013, 558-563. http://dx.doi.org/10.1109/icpads.2013.100

[12] Chen, H.W. and Lin, F.J. (2014) Converging MQTT Resources in ETSI Standards Based M2M Platform. *Proceedings of IEEE iThings*, Taipei, 1-3 September 2014, 292-295. http://dx.doi.org/10.1109/ithings.2014.52

[13] Cerritos, E. and Lin, F.J. (2014) M2M-Enabled Real-Time Trip Planner. *Proceedings of the* 2nd *International Workshop on Internet of Things Technologies* (*IoTT*), Hsinchu, 16-19 December 2014, 886-891. http://dx.doi.org/10.1109/padsw.2014.7097902

[14] Lusiarta Putera, C.A. and Lin, F.J. (2015) Incorporating OMA Lightweight M2M Protocol in IoT/M2M Standard Architecture. *Proceedings of IEEE World Forum on IoT*, Milan, 14-16 December 2015, 559-564.

http://dx.doi.org/10.1109/WF-IoT.2015.7389115

[15] Adrianto, D. and Lin, F.J. (2015) Analysis of Security Protocols and Corresponding Cipher Suites in ETSI M2M Standards. *Proceedings of IEEE World Forum on IoT*, Milan, 14-16 December 2015, 777-782. http://dx.doi.org/10.1109/wf-iot.2015.7389152

[16] Lin, F.J. and Chen, H. (2015) Improving Utilization and Customer Satisfaction of Parking Space with M2M Communications. *Proceedings of IEEE World Forum on IoT*, Milan, 14-16 December 2015, 465-470.

[17] Lin, F.J., Tsai, W., Cerritos, E., Lin, B.-T. and Hu, W.-H. (2015) Identification and Analysis of Charging Factors in M2M Communications. *Proceedings of IEEE World Forum on IoT*, Milan, 14-16 December 2015, 160-165. http://dx.doi.org/10.1109/wf-iot.2015.7389045

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 6309509, 20 pages
https://doi.org/10.1155/2018/6309509

Research Article

A Model of Socially Connected Web Objects for IoT Applications

Sajjad Ali,[1] **Muhammad Golam Kibria**,[1] **Muhammad Aslam Jarwar**,[1]
Hoon Ki Lee,[2] **and Ilyoung Chong**[1]

[1]*Department of Information and Communications Engineering, Hankuk University of Foreign Studies, Yongin-si, Republic of Korea*
[2]*Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea*

Correspondence should be addressed to Ilyoung Chong; iychong@hufs.ac.kr

Received 2 September 2017; Revised 26 November 2017; Accepted 19 December 2017; Published 28 January 2018

Academic Editor: Nathalie Mitton

Copyright © 2018 Sajjad Ali et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Internet of Things (IoT) is evolving with the connected objects at an unprecedented rate, bringing about enormous opportunities for the future IoT applications as well as challenges. One of the major challenges is to handle the complexity generated by the interconnection of billions of objects. However, Social Internet of Things (SIoT), emerging from the conglomeration of IoT and social networks, has realized an efficient way to facilitate the development of complex future IoT applications. Nevertheless, to fully utilize the benefits of SIoT, a platform that can provide efficient services using social relations among heterogeneous objects is highly required. The web objects enabled IoT environment promotes SIoT features by enabling virtualization using virtual objects and supporting the modularity with microservices. To realize SIoT services, this article proposes an architecture that provides a foundation for the development of lightweight microservices based on socially connected web objects. To efficiently discover web objects and reduce the complexity of service provisioning processes, a social relationship model is presented. To realize the interoperable service operations, a semantic ontology model has been developed. Finally, to evaluate the proposed design, a prototype has been implemented based on a use case scenario.

1. Introduction

The Internet of Things (IoT) envisions billions of objects connected to the Internet that continuously generate data about the physical environment. Several state-of-the-art applications can be built using the information and services provided by the pervasive and heterogeneous IoT objects. However, building applications based on these objects raises some challenges. Examples include how the growing number of objects will interact or coordinate to render valuable services, how the complexity generated by the coordination of objects can be handled, and how a large amount of diverse data sensed from the surroundings can be managed properly.

Numerous promising solutions have been provided to handle the complexity and the heterogeneity of IoT objects. One of the prominent approaches is to develop virtual environments to resolve the issues caused by the heterogeneous objects. Virtualization has become a key component in many IoT architectures, in the form of either virtual objects or virtual entities. A virtual object (VO) is defined as a digital representation of a real-world object [1] and provides a way for the IoT services to discover and compose features that cannot be achieved directly with a real-world object (RWO). It has been realized that issues such as heterogeneity of dissimilar objects and scalability can be effectively resolved using VOs. Hence, a crucial role of a VO is to bridge the gap between the physical world and the virtual world by hiding the physical characteristics of an IoT device and acquiring and interpreting its data and context.

Another way to reduce the complexity caused by a large number of objects is to link these objects in a social network [2], which provides a more exact solution than using a sole individual object. This principle has given birth to the concept of Social Internet of Things (SIoT), which is emerging as a new paradigm with the merger of social networks and IoT [3–7]. It has been realized as an efficient way to facilitate the development of complex future IoT applications. It supports sharing of information generated by the people and the devices based on the social relationships. These relationships further promote efficient discovery of the objects and effective service compositions.

Let us have a glimpse of the application scenarios in which SIoT will make a profound effect on our daily life in the near future. Imagine a group of vehicles that form a mobile social network on the move when heading towards the same location or when they are bound in a common relationship. Vehicles may form diverse social communities; for example, small cars create a social network to share available parking space information or bicycles share information about the vacant paths using a network [8]. Vehicles commuting from home to office can join a social network to share traffic congestion experience, accident warnings, or other common information (e.g., a meeting plan) with colleagues by accessing their vehicles' social network. Let us analyze another SIoT use case where a user enters a public museum with an IoT environment, which incorporates smart objects that exploit social relationships to share information. At the entrance to the premises, the user's smartphone is prompted with a beacon signal to recommend a service app installation. When the user grants the permission, an app is installed and a social agent acquires his profile and friendship details. The app incorporates a microservice which uses a smartphone object to establish a social relationship with the museum visit service. The museum is equipped with a network of smart objects that monitor the user's location and provide services to enhance his/her experience such as personalizing the displays, assisting in the navigation, and recommending the relevant services.

To support the above-mentioned use cases, most of the current IoT architectural approaches are not suitable as they are based on the traditional monolithic approaches, which further hinder the development of efficient, modular, and independent services that cannot scale well with the increasing user's service demands, as already witnessed in many studies [9–12]. Using monolithic approaches, we cannot fully achieve the benefits offered by SIoT including efficient information discovery, improved scalability, and simplified interconnection of objects.

These limitations motivate the current paper to define a design for the development of SIoT based services with efficient and lightweight mechanisms to exploit SIoT features for improved service provisioning. The proposed design is based on the microservices concept which promises a more solid practice of SOA. There is no particular definition of microservices architectural style; however, it is explained in [13] by Martin Fowler as "an approach to developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API." Following the microservices based proposed design, SIoT applications can be developed, deployed, and maintained more efficiently and independently, which will promote system modularity and interoperability.

Moreover, a SIoT design needs to facilitate the efficient information discovery based on the social relationships among objects. To achieve this, incorporating the existing social relationship models in the proposed design can be useful. However, applications with intelligent service features require dynamic object selection. Therefore, objects need to acquire the ability to infer new relationships for interconnection with other objects in the system. This view further motivates the present article to develop a social relationship model based on the semantic ontology. The utilized ontology-based approach is highly useful to dynamically infer new social relationships for efficient service provisioning.

Furthermore, to fully achieve the benefits of both, the SIoT and microservices, selecting an IoT environment is necessary where not only can the social interconnection among heterogeneous objects be utilized but also the lightweight services can be developed to apply these relationships for effective IoT service provisioning. To fulfill this goal, the Web of Objects (WoO) [14] platform has been selected. According to the ITU recommendation (ITU-T Y.4452) [15], "the Web of Objects (WoO) is a realization way of the IoT services, where virtualized objects (i.e., virtual objects (VOs) and composite virtual objects (CVOs)) are connected, controlled, and incorporated with resources to facilitate the development, deployment, and operation of IoT services on the World Wide Web."

To exploit the opportunities provided by the above discussed technological advancements, the main contributions are highlighted as follows:

(i) This article contributes a novel design based on the microservices concept with lightweight and modular services to support the development of SIoT applications. The proposed design involves mechanisms to enhance object discovery and to reduce the magnitude of complexity generated due to the huge number of objects.

(ii) A social relationship ontological model is developed, which helps identify the hidden and unidentified relationships among objects with reasoning mechanisms.

(iii) A use case with the implementation prototype based on the WoO platform is developed to exploit the social relationship model, supporting the proposed design for efficient IoT service provisioning.

Further, to support the above contributions, virtual objects, serving on real-world objects, have been developed, which can be reusable in other services as well. CVOs are implemented, where each update of the information in the real-world object is reflected in the VO and propagated to the CVO. Also, the Social IoT notion is incorporated into virtual objects to provide easy discovery and efficient execution while maintaining collaboration using social connections. Moreover, an experimental analysis has been carried out to evaluate the discovery time of VOs and the time required for service execution, with or without social relationships among virtual objects.

The remainder of this article is organized as follows. In Section 2, the related work is described. Section 3 presents the proposed design of the social web objects accompanied by the social interaction model. In Section 4, an IoT environment use case is explained with the details of the prototype implementation and discussion of the performance evaluation. Finally, Section 5 provides a conclusion of this paper.

2. Related Work

This section presents the related work with respect to the SIoT and microservices. Further, the significance of both technologies pertaining to the IoT environment has been discussed.

2.1. Social Internet of Things (SIoT). The concept behind SIoT is to enable smart objects to create a social network similar to a human social network. The objective is to exploit the social relationships among objects in an IoT environment to facilitate the effective information discovery, to promote the scalability, to enhance the interactive communication between objects that are friends, and to achieve trustworthiness [3].

The initial notion of a social network of objects was proposed by [16], which discussed how the wireless devices form social connections on the temporary bases that are controlled by the node owners. As the work was carried out before the introduction of IoT, it did not consider the SIoT concept. Later studies [5] investigated the objects' social interaction where objects formed social networks and communicated with each other based on the rules already set.

The idea of using the Web of Things in a social ecosystem was presented by [17, 18], where things were shared using the social network infrastructures, for example, Twitter, Facebook, and LinkedIn. These implementations consisted of objects that communicated either through the built-in embedded servers or through smart gateways. Further, in these settings, the web-enabled things owned by a person provided an interface to his/her social group of friends so that they can interact with the things using a social network.

Lysis [19] is another good example of Social IoT platform based on cloud infrastructure. It uses the platform as a service (PaaS) model and focuses on the deployment of applications in the cloud space. However, dynamic inferencing of new social relationships from existing relationships is not supported by this system. Moreover, making a modular interoperable design by employing concepts like microservices and semantic web technologies is not considered in this work.

Paraimpu [20] provided a social Web of Things platform to connect virtual and physical things to the web. In this system, the social concept only relates to humans by providing them with the capability to share things with each other using a human social network. Also, virtual things are not considered as VOs but rather viewed as services on other IoT platforms.

Using SIoT at the edge of the network was investigated by [21]. In this system, an approach was proposed to cope with the communication delay due to the objects being remotely located in the cloud. This approach exploits the computing resources at the network edge to deploy virtual objects.

The ThingSpeak solution was used as the basis for the implementation of an early SIoT platform [22]. This platform provided the object social behavior functions incorporated at the centralized server. In this implementation, functions such as the creation and the management of the social relationships were developed. However, VO concept was limited only to the records in the remote database.

Another featured SIoT framework was contributed by [4]. This work presented how to combine services, devices, users, and their social interaction to enable interactivity, discovery, and recommendation of services. Social relationships have also been investigated in communities [23], where a proposed framework enables the identification of communities in the social networks.

Although most of the research contributions in the SIoT domain incorporate a notion of the social association of objects, however, limited information has been provided about the social relationship model and the details on how that can be used by IoT services. Furthermore, existing approaches are lacking lightweight and modular mechanisms which can be utilized by services to exploit social relationships efficiently. In our approach, we have developed a semantic ontology-based social relationship model that provides the capability to infer new connections in a network of social objects. Another distinction of the proposed system is a microservices based design that utilizes the developed social model to provide lightweight, modular, simple, and interoperable IoT services.

2.2. Microservices. The microservices architecture is emerging as a new trend among the practices of developing distributed web applications. To develop SOA based systems, microservices have become a prominent approach in real-world settings [24]. In the microservices based architectural pattern, each application incorporates a collection of small services which execute independently and use lightweight communication mechanisms [9]. The microservices are targeted for doing one thing well at a time based on the principle of single responsibility [25], as defined by Robert C. Martine: "Gather together those things that change for the same reason and separate those things that change for different reasons" [26]. Also, the microservices are conceived as autonomous entities, which means they can be changed and deployed independently of each other without requiring their consumers to change [24].

In the last few years, several IoT architectural designs were proposed and developed. Some of these designs provided innovative features such as the European FP7 project iCore [27] which proposed a cognitive framework for IoT application development. This project provided an architecture with the concept of virtual objects and their composites. However, in iCore, services are designed using a traditional SOA based monolithic style, whereas the microservices based architecture provides a better option to develop IoT services. Furthermore, in iCore, VOs lack the mechanisms to semantically represent real-world objects (RWOs), which limits extensibility and interoperability.

To support IoT applications, microservices based architectures are being proposed in many research initiatives and applied in several research projects. In [28], microservices are used for M2M applications, realizing the fact that monolithic approaches for M2M cannot provide a real solution. Another work [10] provided an analysis of using microservices architecture for large-scale distributed applications. This work demonstrated that significant benefits can be achieved as compared to generic SOA approaches when using microservices in an IoT platform for smart city applications.

Moreover, Almanac FP7 EU Project [29] provides smart city services based on the IoT environment. The project uses microservices to employ scalability in a horizontal and vertical fashion. Also, in the industry, microservices have become a solution for developing large-scale applications. Netflix, Pivotal [30], and Amazon use microservices in their software bases.

The increased importance of microservices for the IoT applications is due to the fact that they simplify complex systems. By dividing a system into smaller parts, higher cohesion and lower coupling can be achieved, which makes it highly scalable. In scenarios where requirements keep changing continuously, microservices can help make a system easily modifiable. Moreover, the major benefits that are envisaged when using microservices in an IoT environment are as follows: Microservices leverage scalability to provide a highly decoupled pattern and can allow scaling individually. Their usefulness can be realized as when the demand for the requirement increases the system complexity also rises. In this case, a system can be supported by creating new instances of existing services. Microservices scalability [9] fits in three-dimensional space, that is, horizontal scalability (typical scalability), vertical scalability (splitting different individual microservices), and z-axis scalability (splitting similar things, such as DB partitioning). Microservices architecture supports a plug-and-play behavior where system components become loosely coupled. To acquire new functionality or replace a failed service, pull the plug from one microservice and plug into a new microservice [31]. One of the major problems that IoT is facing today is how to deal with the heterogeneity of incompatible solutions. To harmonize the heterogeneity, semantic web technologies with microservices are leveraged to provide interoperable exchange and communication of data. Microservices realize a decentralized and autonomous behavior, operating on their own priorities and schedule. This way, they provide several benefits for being utilized in an IoT design.

3. The Proposed Design of Social Web Objects

Recently, there have been several research efforts [3–6, 17, 20, 32] for defining the Social IoT. This new aspect minimizes the complexity generated by the communication of billions of real-world objects and enables sharing of information. The social notion can be incorporated into the WoO platform to provide a semantically rich base for IoT applications that can utilize real-world object relationships for efficient information discovery. Moreover, this section provides a brief introduction of WoO reference architecture in Section 3.1. A classification of social relationships among web objects (virtual objects and composite virtual objects) is discussed in Section 3.2 and a social relationship model is presented in Section 3.3. Further, the details of functional components for the proposed architecture are explained in Section 3.4.

3.1. Web of Objects (WoO) Reference Architecture. To enable the deployment of IoT services on the World Wide Web, WoO provides a reference framework. In other words, it realizes the IoT services in such a way that virtualized objects are interwoven with resources to support the development, deployment, and operations of IoT services [33, 34]. The WoO platform provides service functionality by merging VOs and features of web applications. Moreover, the WoO platform uses semantic web technologies to enable interoperability among heterogeneous resources. This realization provides the basis for the composition and harmonization of objects to provide smart services in an IoT domain. Building services based on WoO platform using diverse technologies, including microservices [35], social networks, and semantic web, helps reduce the complexity and fosters the easy and efficient development of IoT applications.

Virtualization has become a major concept in IoT to address the heterogeneity of diverse types of physical objects. VO is a digital representation of a real-world object that can be anything, living or nonliving, mobile or stationary, concrete or immaterial. In WoO, VO is defined as domain-specific semantic ontology based on the VO information model [36]. It is uniquely identified using URI and provides information updates on the representative real-world object. On the other hand, WoO also provides a notion of CVO which aggregates one or more VOs to enable service features that satisfy the application requirements. CVOs chain semantically interoperable VOs together. The WoO platform incorporates features to efficiently reuse existing VOs and handles the complexity generated by self-management and control mechanisms [37]. In the WoO, service level decides which CVOs and VOs will take part in creating an object mashup to fulfill service objectives. The service logic is saved in the form of templates and stored in the template repository [38–40]. The domain expert or knowledge engineer defines the service templates that are used to instantiate new services.

WoO reference architecture is shown in Figure 1. In this architecture, the service layer handles requests and provides several management functions, whereas the CVO level consists of functions to manage and instantiate CVOs or reuse the existing ones based on the provided service execution logic. New objects such as sensors are registered in the WoO platform using a registration function and their templates are generated to make them digitally available in the form of VOs. Another most important aspect is the semantic representation of data at three-layered WoO architecture [23, 39]. The database at each layer is supported with semantic web technologies.

3.2. Types of Social Relationships. Previous research [3] on SIoT has derived some basic relationship types among objects. These relationships are characterized as follows. *Parental Object relationship* (POR) exists between the objects that belong to the same batch, such as objects that are created in the same production process or at the same time by a common manufacturer. *Cowork object relationship* (CWOR) is the relationship among objects that are grouped together based on some commonly shared job to be done by them. Similarly, *colocation object relationship* (CLOR) is established

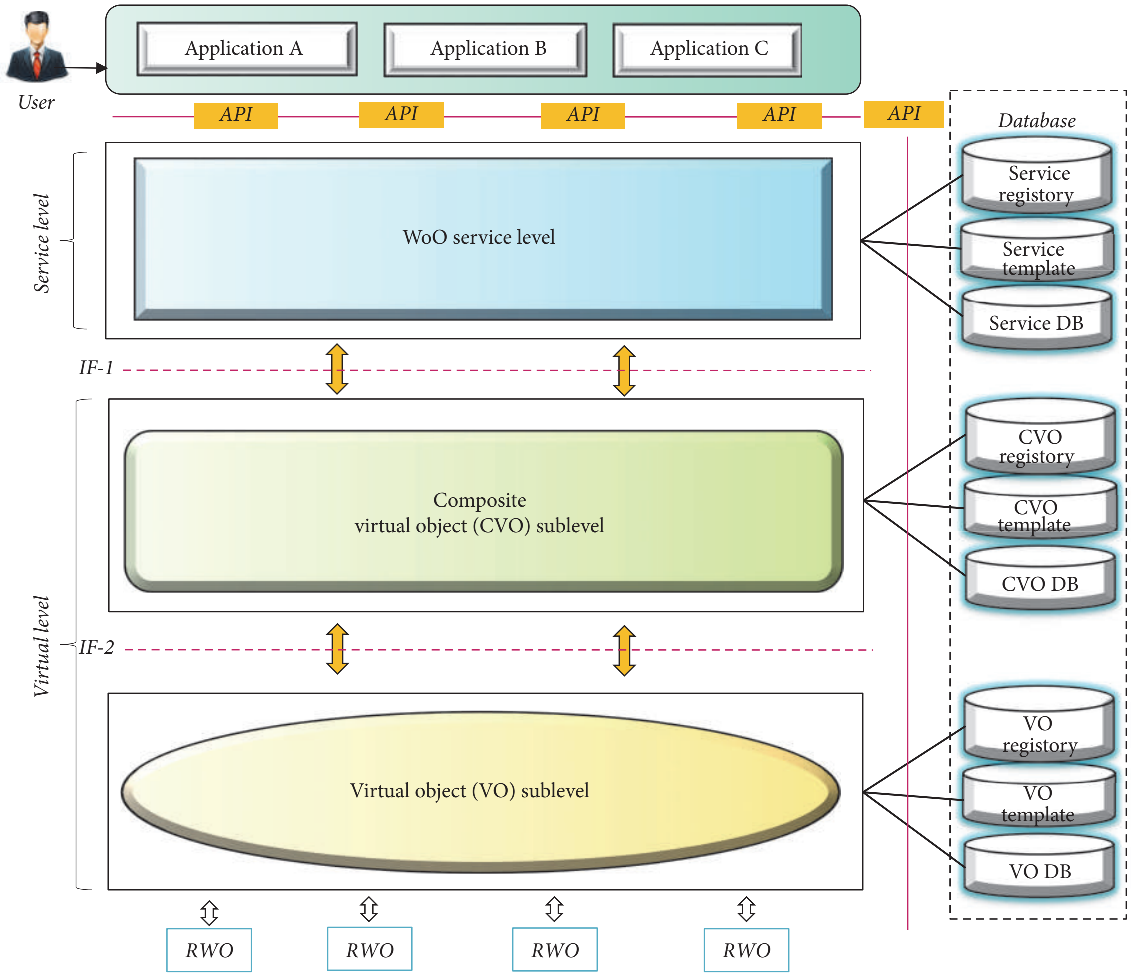
User
Application A
Application B
Application C
API
API
API
API
API
Database
Service level
WoO service level
Service registory
Service template
Service DB
IF-1
Virtual level
Composite virtual object (CVO) sublevel
CVO registory
CVO template
CVO DB
IF-2
Virtual object (VO) sublevel
VO registory
VO template
VO DB
RWO
RWO
RWO
RWO

Figure 1: Web of Objects reference architecture.

between objects belonging to the same location such as home or office. Different or same types of objects can be combined based on the common location or premises. On the other hand, *Ownership Object Relationship* (OOR) is formed among objects belonging to the same owner. *Social Object Relationship* (SOR) is another kind of relationship that is created among objects when they come close to each other, either random in time or periodically. This relationship can be envisaged when the owners of objects come into contact with each other. Apart from the above relationships, in [41], the authors define the *Guardian Object Relationship* (GOR) in the Internet of Vehicles (IoV) scenario where on-board units of vehicles become children in relation to the super nodes of Road Side Units. This child and parent association gives a special meaning to a new hierarchical relationship.

In addition to the relationships defined in other studies, we have identified few more relationships suitable for some diverse scenarios. These include *Sibling Object Relationship* (SIBOR) that is created among objects that belong to a group of friends or family members. This relationship among objects extracts the property of trustworthiness based on the relationship of their owners. Another relationship type is the *guest object relationship* (GSTOR) which is formed between objects that belong to the users in the guest role, for example, when a person visits a friend's home and gets the privilege as a guest. The same can be applied to the objects that move in guest relation (GSTOR) from one place to another; they will have some privilege of accessibility of information as compared to other objects. *Stranger Object Relation* (STGOR) applies to objects that encounter the presence of each other in an anonymous environment such as on the go or in the public environment. Similar to people who meet each other sometimes regularly but are anonymous to each other or are not fully aware of each other, this relationship can be used among objects to form different trust levels or to form strict restrictions to promote secure connections. Moreover, in *service object relationship* (SVOR), objects form a relationship while coordinating in the same service composition to fulfill a service request. Generally, in the proposed scheme, service request execution generates a mashup of objects, and objects that belong to the same mashup are assumed to be in SVOR.

We believe that the above defined new social relationships are beneficial in several distinctive scenarios. For example, SIBOR is more useful in scenarios where trust is highly important for establishing a connection. Unlike the more generalized relation POR, SIBOR is established

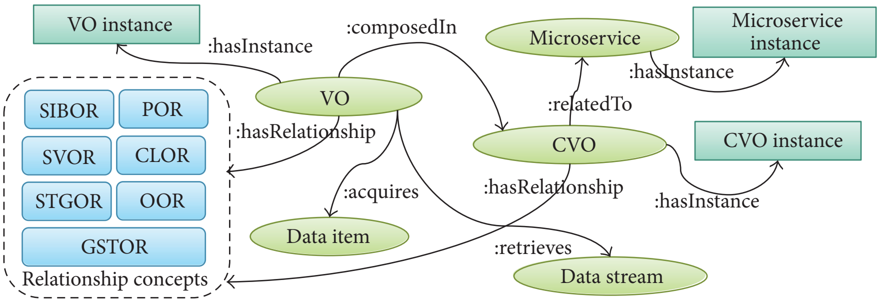

(i) Sibling Object Relationship (SIBOR)
(ii) Service Object Relationship (SVOR)
(iii) Stranger Object Relationship (STGOR)
(iv) Guest Object Relationship (GSTOR)
(v) Parental Object Relationship (POR) [3]
(vi) Colocation Object Relationship (CLOR) [3]
(vii) Ownership Object Relationship (OOR) [3]

Figure 2: Ontological model representing web objects' relationships.

among objects from the same manufacturer having similar features but different behaviors. Similarly, GSTOR is suitable in scenarios where objects use relationships for privileged services. Imagine a person as a guest speaker in a conference with his smart devices, where he can connect to the network easily, and his personal devices are recognized as a guest (with GSTOR) in the venue network to avail free services like high-speed Internet, special notifications, and others. Stranger Object Relation (STGOR) is used in scenarios where objects are on the go, and to get some information, they have to compromise with a limited trust level. For instance, in a public transit system, one mobile node on the move wants to use crowdsourced information from another mobile node which is not fully trusted. Also, SVOR is useful in service composition specific scenarios, where object mashups are formed to facilitate service requirement, and it is also useful in a particular orchestration of objects.

To model information about microservices, social web objects, and their relationships, a semantic ontological model has been developed as depicted in Figure 2. Conceptualization of objects and their social relationships in semantic structures are highly beneficial to infer new connections among a network of objects with the help of reasoning techniques. The ontology facilitates the representation of objects, as well as what they measure in terms of observations, their processing, and functions. Web objects' relationship ontology contains the concepts that define each object in terms of a class. The data properties are used to represent object values and the object properties are defined to identify the link between two objects in the ontology. Major concepts in the ontology are microservices, CVOs, VOs, and their social relationships. Further, microservices have an individual instance that uniquely defines a microservice object and associated data with it. Similarly, CVOs and VOs are also instantiated and contain data properties to retain their values. With the increasing number of objects, a semantic ontology model is used to perform reasoning function that helps identify new connections. Further, to analyze the social relationship among objects (such as VOs and CVOs), their properties are provided in Table 1.

3.3. Social Relationship Model for Web Objects.

At each level of service life cycle, objects form a social relationship with each other to accomplish a service task. Codifying the relationship among web objects can provide many benefits, such as efficient information discovery from related objects and better composition and reuse. To perceive social interaction between web objects, we assume every service is based on one or more microservices and other objects including CVOs and VOs. At VO level, VOs create several relationships with each other based on the RWO they represent. CVOs associated with microservices are incorporated in relationships as well.

Object-to-object associations are categorized as either vertical associations (interobject relations) or horizontal associations (intraobject relations) as shown in Figure 3. The relationships that flow from the bottom to the top in WoO based architecture are known as interobject relations (i.e., from VO to CVO or vice versa). On the other hand, relationships are known as intraobject relations if they are generated within objects at each level such as between one VO and another, or within one CVO to a relative CVO. The newly defined relationships among objects are maintained at different levels, which enables information discovery more efficiently. It is considered that not only will the relationships be maintained when two objects (i.e., VOs) are used in the same CVO or service, but also they will be maintained if objects are used in different CVOs or services.

As depicted in Figure 3, VO1 and VO2 bind in the SIBOR relationship, considering the fact that they belong to the same family of CVO1. Similarly, VO4, VO5, and VO6 are combined in a CWOR, as they are used in the same service. However, all of them do not have a common family relation that makes their trustworthiness restrictive to some level as compared to SIBOR. Moreover, though VO7 and VO8 are not used together in a service setup, still they form a STGOR relationship. This relationship also helps codify those VOs that have similarity in premises during a certain point in time regularly but they had never combined in any service scenario before. For example, if a person visits a subway regularly but stays very shortly, in such scene, the user's smartphone VO and subway station point

Table 1: Social relationship attributes.

Social object relationships types	Attributes/properties
Sibling object relation (SIBOR)	O_x *has relationship* with O_y → SIBOR, if objects' *ownership* is defined as: == SameFamily
Service object relationship (SVOR)	O_x *has relationship* with O_y → SVOR, if objects' *serviceStatus* is defined as: == SameServiceComposition
Stranger object relation (STGOR)	O_x *has relationship* with O_y → STGOR, if objects' *authentication* is defined as: == Anonymous
Guest object relationship (GSTOR)	O_x *has relationship* with O_y → GSTOR, if objects' *ownership* is defined as: !=SameFamily && == GuestFriend
Parental object relationship (POR)	O_x *has relationship* with O_y → POR, if objects' *creation* is defined as: == SameBatch ‖ == SameProduction
Colocation object relationship (CLOR)	O_x *has relationship* with O_y → CLOR, if objects' *proximity* is defined as: == Neighborhood ‖ *location* is defined as: == SameLocation
Ownership object relationship (OOR)	O_x *has relationship* with O_y → OOR, if objects' *proprietorship* is defined as: == SameOwner

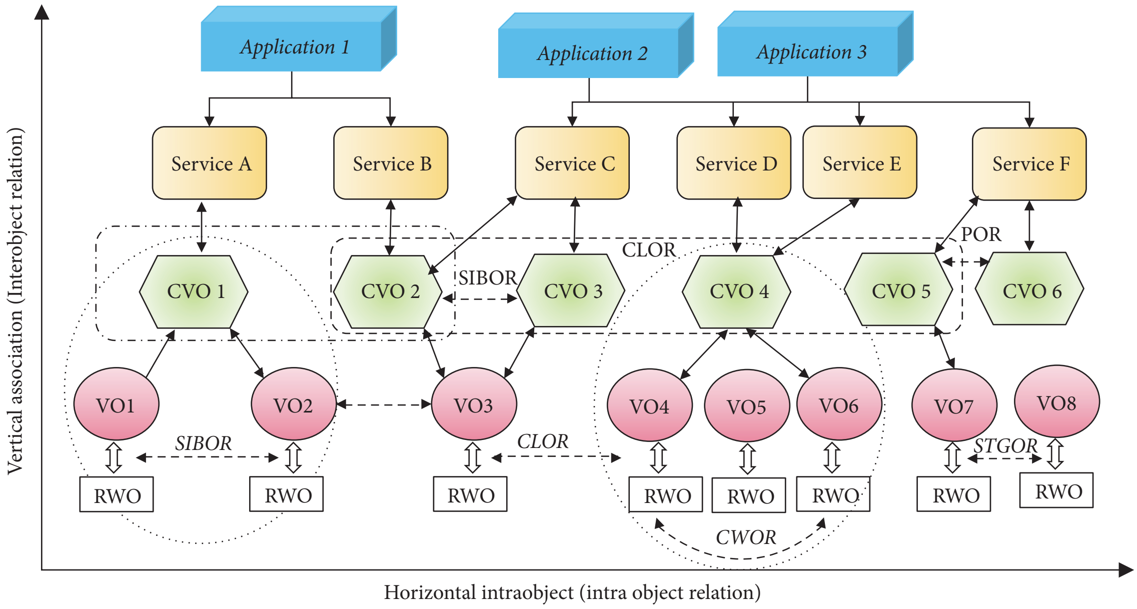

Figure 3: Hierarchical social relationships of web objects.

VO or subway security camera VO may establish STGOR. This may help reuse VOs in service scenarios where only for short duration a VO object is required. At CVO level, CVO2 forms two different relationships, which increases the level of connectivity of the CVO. CVO2 first shares a CLOR with CVO1 and CVO3, CVO4, and CVO5 together. Secondly, CVO2 is combined in a SIBOR with CVO3 and becomes in the same service family. Further, in the selected use case scenario described in Section 4, several other social interactions have also been elaborated.

3.4. Social Web Objects Architecture. SIoT envisages a system where the social framework will bring smart devices and people to interact with each other. By incorporating web technologies like SOA based microservices, IoT services can be rendered on top of the social framework. The proposed social web objects architecture (as shown in Figure 4) has been decomposed into three levels: service level, object virtualization level, and aggregated object virtualization level. At the service level, to support social relationships within web objects, several microservices have been designed, which are discussed further in the following section. Moreover, the data management function in the proposed design helps each layer to interconnect with semantic databases. The SPARQL endpoints have been defined at the service and the object virtualization levels that expose interfaces to retrieve, store, and modify RDF graphs. Several interfaces allow the knowledge engineer, domain expert, and the developer to create service templates and VOs, as well as update the RWK model, user profile, and policies.

3.4.1. Service Level Functional Components. The service level functions handle operations from the request inception to the execution of services. This level is supported by some core functions common to be used in each service life cycle. These include the management function to handle service

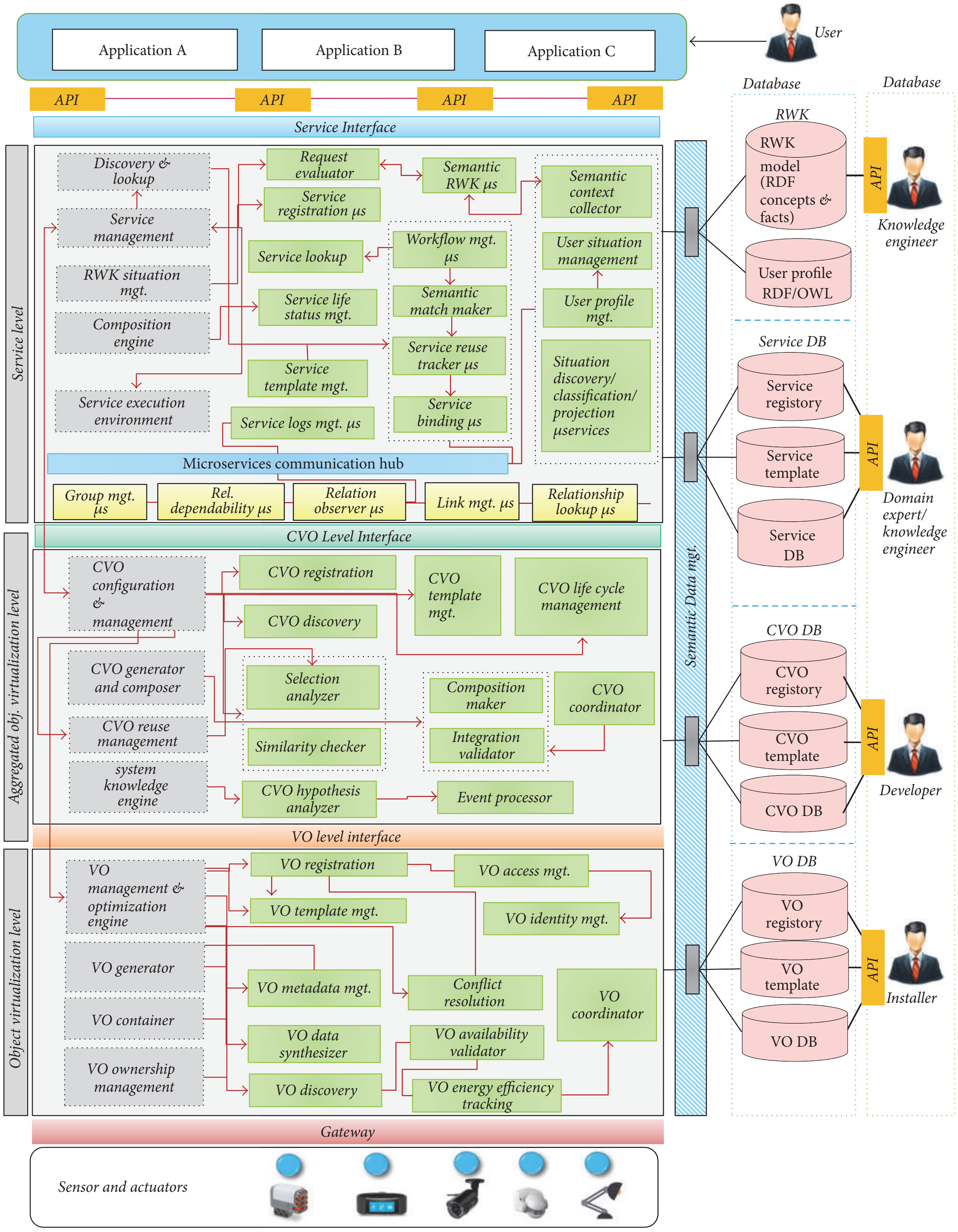
Application A
Application B
Application C
User
API
API
API
API
Service Interface
Service level
Discovery & lookup
Service management
RWK situation mgt.
Composition engine
Service execution environment
Request evaluator
Service registration μs
Service lookup
Service life status mgt.
Service template mgt.
Service logs mgt. μs
Semantic RWK μs
Workflow mgt. μs
Semantic match maker
Service reuse tracker μs
Service binding μs
Semantic context collector
User situation management
User profile mgt.
Situation discovery/ classification/ projection μservices
Microservices communication hub
Group mgt. μs
Rel. dependability μs
Relation observer μs
Link mgt. μs
Relationship lookup μs
CVO Level Interface
Aggregated obj. virtualization level
CVO configuration & management
CVO generator and composer
CVO reuse management
system knowledge engine
CVO registration
CVO discovery
CVO template mgt.
CVO life cycle management
Selection analyzer
Similarity checker
Composition maker
Integration validator
CVO coordinator
CVO hypothesis analyzer
Event processor
VO level interface
Object virtualization level
VO management & optimization engine
VO generator
VO container
VO ownership management
VO registration
VO template mgt.
VO access mgt.
VO identity mgt.
VO metadata mgt.
Conflict resolution
VO coordinator
VO data synthesizer
VO availability validator
VO discovery
VO energy efficiency tracking
Gateway
Sensor and actuators
Semantic Data mgt.
Database
Database
RWK
RWK model (RDF concepts & facts)
User profile RDF/OWL
API
Knowledge engineer
Service DB
Service registory
Service template
Service DB
API
Domain expert/ knowledge engineer
CVO DB
CVO registory
CVO template
CVO DB
API
Developer
VO DB
VO registory
VO template
VO DB
API
Installer

FIGURE 4: Social web objects functional architecture.

requests and initiate service provisioning processes, the service discovery function to select the available services, the composition engine to enable service mashups, and the Real-World Knowledge (RWK) and situation management function to acquire RWK and continuously maintain it. Besides these core functions, several microservices have been used to provide the plug-and-play feature in the proposed design; that is, in case of service failure, it is replaced with another one. These services include the Request Evaluator microservice which interprets a service request and matches a request query to the service template. The registration microservice registers newly added service objects to the registry. When a new template is instantiated, objects are put in service, and to know the existence of available service instances, they are recorded in the service registry. The lookup microservice is defined to search for the required service objects using the semantic representation. The representations available in RDF/OWL format are queried from the registry based on the request parameters. Service life cycle management keeps track of the states of service objects during their life cycle, whereas RWK management microservice is a semantic data processing service which processes facts about user preferences, profile, and situational information.

To realize service provisioning, it is always not possible to satisfy the service requirement with a single service function. Service composition and harmonization features are required to fulfill a demanding service request. Service composition is not a single operation; many microservices are used to compose the individual features. One of the major components of the composition is the Workflow Management (WFM) microservice which is responsible for decomposing the service requirement if no single service is available to satisfy the task. WFM microservice identifies the service objects using input and output interfaces. In this concern, it uses the matchmaking microservice to identify the approximate matches that can serve the required service functionality. Here, service objects are semantically represented in the form of ontologies. The semantic ontology alignment algorithm is used to match the service up to a certain threshold level that is defined by the matchmaking microservice. The reuse tracker microservice is used to identify the current service instance usage by the application; it incorporates a coordination mechanism before instantiating any new instance, and this helps in service reuse by many clients. Binding microservice enables tying of services selected by WFM; this function synchronizes input and output of services in a workflow to produce a single service output based on the generated composition plan.

Furthermore, user characterization is another important element to support the IoT services by matching user requirements more closely. The User Profile Manager microservice handles many facts related to the human user. These include user profile, preference, context, and policies. Semantic Context Collector microservice organizes the services requested by the user and the context in which these services were requested. This helps in identifying user interest with respect to the service context. User context or situation helps in service selection more efficiently if a user requests again with a similar service requirement. Some microservices have been designed to handle user situation information; these microservices include situation discovery, classification, and recognition functions. Situation discovery is used to process sensor data that is provided by CVOs to prompt a situation, whereas machine learning algorithms are applied to remove the false values and data normalization is performed in case inappropriate triggers in the data exist. Situation classification incorporates the reasoning methods and rules to generate a relationship over detected events and refine the missing knowledge between the events. Situation projection analyzes the events data with the help of machine learning methods and provides an output in the form of predicted facts that are related to the situation being detected. Moreover, all the microservices coordinate using a communication hub to share information with each other.

A flow of service inception to service execution in the proposed architecture is depicted in Figure 5 where service management is responsible for initiating the above-mentioned functionalities and instantiating the microservices. This process further results in the generation of CVO/VO mashup graphs to be handled by CVO management (discussed in Section 3.4.2). Situational information has also been used to improve request analysis and service provisioning, whereas VO management (discussed in Section 3.4.3) acquires and aggregates data from sensors and other RWOs.

Social Relationship Management Microservices. To support the social relationships among web objects at each layer of the proposed system, relationship management microservices are defined, which support the codification of relationships among web objects (VOs and CVOs) and their management. These microservices are defined as follows. Group management microservice identifies the grouping of objects into specific sets based on the type and interaction of objects; it builds a social graph of the object relationships. Link management microservice incorporates mechanisms to maintain relationships among different objects and it includes several other subfunctions. The first subfunction is object selection which involves choosing candidate objects that are likely to form a particular association. The second subfunction is matchmaking service which provides exact relationship match based on the type of objects. The third subfunction is the association life cycle manager, which maintains the relationship status between objects; it checks whether they are in active relation or not. It keeps track of relationship validity and also the duration of relationships. Relationship observer microservice enables the object activities to be observed, and it also handles the mutual sharing, which defines what information an object is allowed to share or use of another object based on the relationship among them. Relationship dependability recognizer microservice identifies, maintains, and continuously improves the level of reliability among objects. If objects A and B are frequently interacting and always perform a task together that leads to successful execution of an application to serve the requirements, then they are more likely to form a reliable relationship with each other. The consistency of the relationship allows objects to share their information with fewer constraints as compared to unpredictable association among objects. Relationship

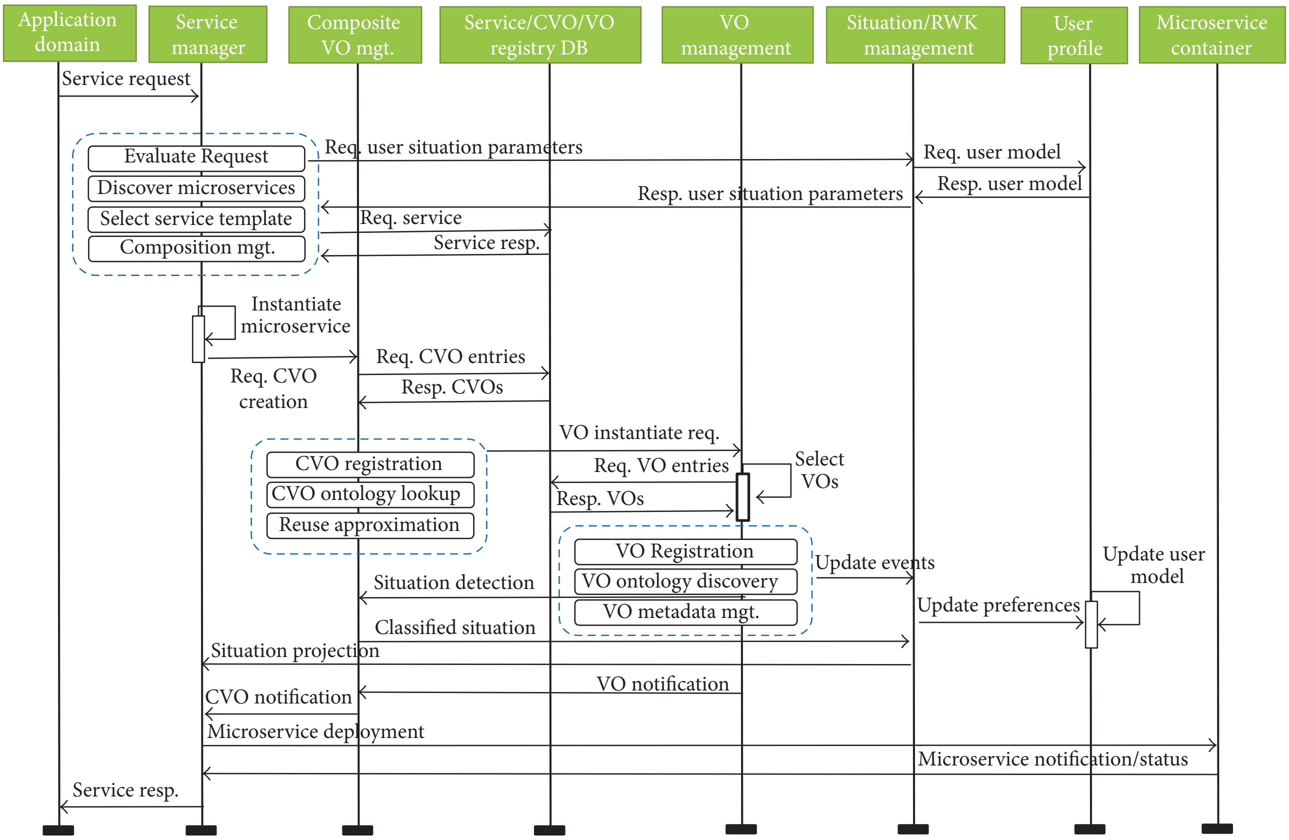
Application domain
Service manager
Composite VO mgt.
Service/CVO/VO registry DB
VO management
Situation/RWK management
User profile
Microservice container
Service request
Evaluate Request
Discover microservices
Select service template
Composition mgt.
Req. user situation parameters
Req. user model
Resp. user model
Resp. user situation parameters
Req. service
Service resp.
Instantiate microservice
Req. CVO creation
Req. CVO entries
Resp. CVOs
VO instantiate req.
CVO registration
CVO ontology lookup
Reuse approximation
Req. VO entries
Resp. VOs
Select VOs
VO Registration
VO ontology discovery
VO metadata mgt.
Update events
Update user model
Update preferences
Situation detection
Classified situation
Situation projection
VO notification
CVO notification
Microservice deployment
Microservice notification/status
Service resp.

Figure 5: Request initiation and service execution flow in the proposed architecture.

lookup microservice incorporates the discovery algorithm (as given in listing of Algorithm 1) to discover the social relationships and maintains the list of all objects and their potential association rating. The rating defines the opportunity for objects to form a relationship with each other.

Moreover, the operation flow of social microservices has been illustrated in Figure 6, where VO management retrieves the social relationship among objects this is queried using link management, which uses a lookup service. Associations are maintained in a social relationship graph using semantic ontology and queried using the relation observer service. The social graph is also updated and maintained continuously with respect to changing links among objects, on the basis of which VOs are selected.

3.4.2. Functional Components in Aggregated Object Virtualization Level. Aggregated object virtualization (AOV) level involves all the necessary functions to instantiate, manage, and continuously monitor CVOs. Besides, it includes the interface required by the service level to interact with CVO components. Service level provides the mashup graphs of CVOs and VOs that are required to fulfill a service request. Aggregation level also supports some core functions; these are categorized as CVO configuration and management to decompose requests from service level and manage CVO instances, CVO generator to instantiate CVOs, composition function to form CVO and VO mashups, reuse management to select existing CVOs that approximately match service requirement, and system knowledge engine to grow knowledge for the optimal use of CVO resources. Further, other aggregation level functionalities are as follows: registration function records the CVO entry into the semantic CVO registry in the form of RDF/JSON format. CVO includes ontologies that specify conditions to be applied on the VO data. Also, at this level, templates are created by a domain expert or knowledge engineer, which consist of functionalities associated with each CVO type. Template management module provides an interface to include templates to the CVO template repository and also supports the modification mechanism. CVOs are discovered using their semantic annotation via the discovery function. CVO selection analyzer function provides estimated selection of a CVO for reuse if there is no single CVO that can exactly provide the required service. This is supported by similarity checking mechanism to analyze the level of similarity before recommending for reuse. At aggregation level, another important function is event processing; event streams generated by object virtualization level are highly valuable to infer situational awareness and building RWK that is used by high-level functions. Event processing module facilitates processing of events and provides generated facts as output. Event processing is supported by CVO hypothesis analyzer that incorporates several hypotheses which are trained over data to build system level knowledge. Aggregation level also

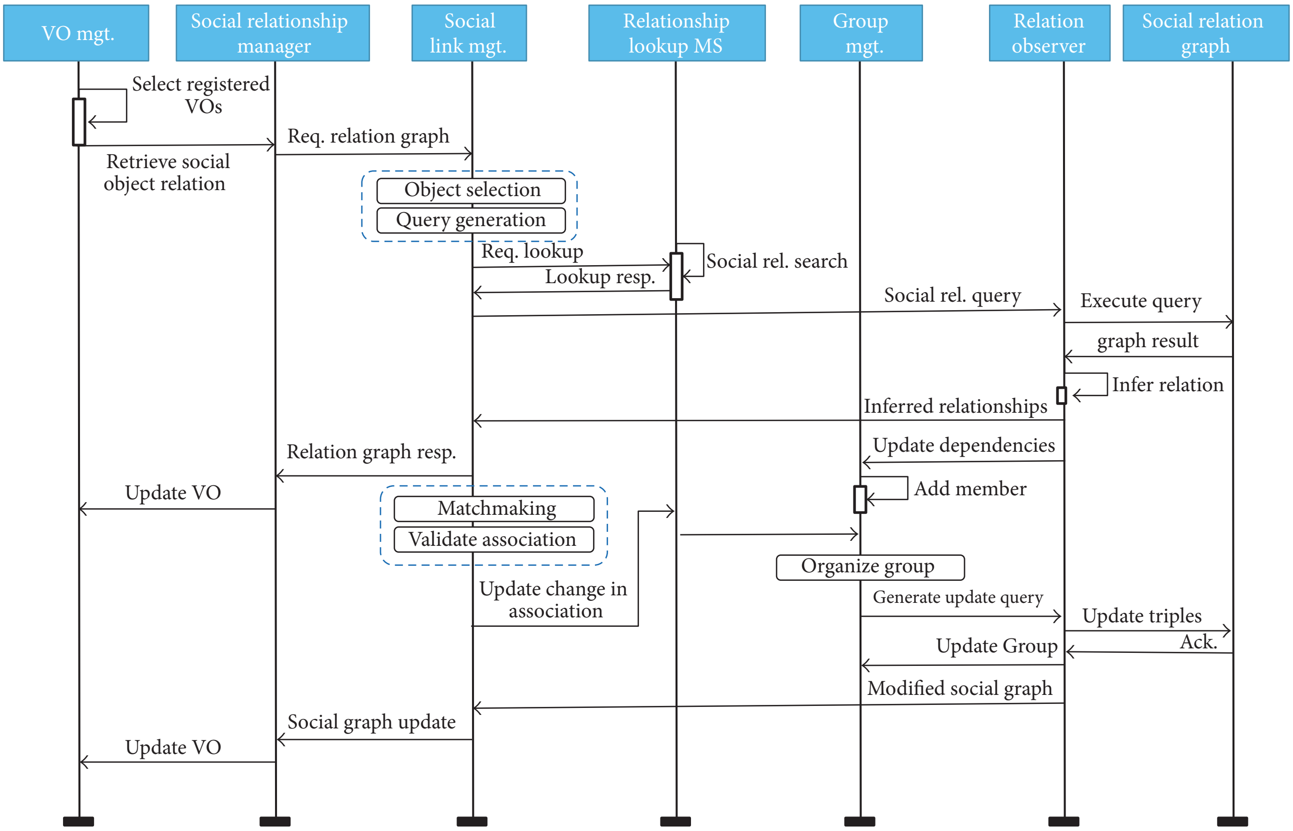
VO mgt.
Social relationship manager
Social link mgt.
Relationship lookup MS
Group mgt.
Relation observer
Social relation graph
Select registered VOs
Req. relation graph
Retrieve social object relation
Object selection
Query generation
Req. lookup
Social rel. search
Lookup resp.
Social rel. query
Execute query
graph result
Infer relation
Inferred relationships
Relation graph resp.
Update dependencies
Add member
Update VO
Matchmaking
Validate association
Organize group
Update change in association
Generate update query
Update triples
Ack.
Update Group
Modified social graph
Social graph update
Update VO

Figure 6: Sequence of operation required to establish a social relationship among objects.

provides composition function which provides the basis for merging multiple VOs to generate single service feature and validation function to validate the compositions of multiple VOs. Moreover, the coordinator helps in resolving conflicts in case of simultaneous access of single CVO by multiple services together.

3.4.3. Functional Components in Object Virtualization Level. Although IoT middleware virtualizes sensors and actuators in some form of virtual entities, however, in WoO, every living or nonliving thing can be represented in the virtual world. VOs are supported by an information model [14, 15] which includes the representation of RWO that can be ICT or non-ICT. In the proposed architecture, object virtualization level includes functions to create, maintain, and coordinate VOs such as VO generator, VO management, and optimization engine. VO registration function provides an interface to register VOs with the help of VO ownership management function where the VO ownership is maintained. Template management function handles templates or loads updated ones as provided by the domain expert. Metadata management maintains data associated with VOs; as VO is a digital representative of real-world entities, there is a high need to store metadata information to recognize the class of sensed data. Use of metadata provides the real value of the data in the form of the context stored within it. Furthermore, VO data synthesizer service handles the discrepancies if found in the data corresponding to VO; this includes data manipulation and missing value rectifications. Conflict resolution function helps in case the same VO is used by multiple service instances, to resolve the conflicts. Also, VO access management function controls access for VOs in case access right information has been defined by the VO owner, whereas identification management module maintains identification information about VOs so that each VO can be uniquely identified in the system.

3.4.4. Social Relationship Discovery and Composition. To identify associations among web objects within a social relationship graph, an algorithm is elaborated briefly. In the following algorithm, the microservices that are in execution are retrieved first, represented as ⅀Mμ; next, all the CVOs associated with each of those microservices are recovered in ⅀Mc and, further, the corresponding VO model ⅀Mv is iterated. In the next step, executing the query statement routine returns the results denoted as ₸. The process of finding the relationships is carried out through lines (6)–(19) in multiple nested iterations as shown in the pseudocode of Algorithm 1 listing.

Algorithm 2 facilitates the composition of objects based on social association types. This algorithm takes as input the set of object social relationships denoted by ℤ and registry entries of the objects that are currently available in the system, represented as ℝe. After the object model denoted

```
Require: $(\Sigma\phi)$, $(\mathbb{R}_e)$
(1) Output: $\hat{\mathrm{S}}$ (social relationship graph of objects)
(2)  $\Sigma\mathrm{M}_\mu \leftarrow$ Load Microservices in model
(3)  $\Sigma\mathrm{M}_c \leftarrow$ Load CVOs in model
(4)  $\Sigma\mathrm{M}_v \leftarrow$ Load VOs in model
(5)  $\overline{\mathrm{T}} \leftarrow$ executeQueryStatement $(\mathrm{M}_\mu, \mathrm{M}_c, \mathrm{M}_v, Q)$
(6)  for All $\mu$ in $\mu\overline{\mathrm{T}}$ do (iterate Microservice instance being used)
(7)    for All $O \in \mu$ do (check CVO and VOs used by the Services)
(8)      if $\mathrm{O}_i$ == output of $\mathrm{C}_R$ in a relationship set $\mathrm{R}_c$ then
(9)        add $\mathrm{C}_R$ to $\mathbb{II}_L$
(10)     else
(11)       if $\mathrm{O}_i$ == output of $\mathrm{V}_R$ in relationship set $\mathrm{R}_v$ then
(12)         add $\mathrm{V}_R$ to $\mathbb{II}_L$
(13)         $\mathbb{N}_i \leftarrow$ tag (Assign relationship tag to each object entry)
(14)       else
(15)         add $\mathrm{O}_i$ to $\mathbb{II}_U$
(16)       end if
(17)     end if
(18)   end for
(19) end for
```

Algorithm 1: Social relationship discovery.

```
Require: $(\Sigma\phi)$, $(\mathbb{R}_e)$, $(\mathbb{Z})$
(1) Output: $\mathbb{C}$ (Composite Service)
(2) $\Sigma\mathrm{M} \leftarrow$ List all objects available (Microservices, CVOs, VOs) in model
(3) $\overline{\mathrm{T}} \leftarrow$ executeQueryStatement $(\Sigma\mathrm{M}, Q)$
(4) $\mathbb{Z} \leftarrow$ executeQueryStatement $(\Sigma\mathrm{M}, Q')$
(5) for All $\sigma i \in \overline{\mathrm{T}}$ do
(6)   for All $\lambda i \in \sigma$ do
(7)     if $\lambda i$ == any $z$ in $\mathbb{Z}$ then
(8)       add $\lambda i$ to $\mathbb{II}_{\mathrm{Lm}}$
(9)       store $\mathbb{II}_{\mathrm{Lm}}$
(10)    end if
(11) end for
(12) for All $\varpi\mathrm{i} \in \mathbb{II}_{\mathrm{Lm}}$ do
(13)   if $\varpi\mathrm{i}$ == any $\varpi\mathrm{i}$ in $\mathbb{II}_{\mathrm{Lm}}$ then
(14)     $\partial \leftarrow$ ranking (Assign rank to each service)
(15)     add $\varpi\mathrm{i}$ to $\mathbb{II}_{\mathrm{LR}}$
(16)     store $\mathbb{II}_{\mathrm{LR}}$
(17)   end if
(18) end for
(19) for All $\sigma j \in \mathbb{II}_{\mathrm{LR}}$ do
(20)   if $\sigma j(\mathrm{R}\partial) > \kappa$ then
(21)     add $\sigma j$ to Wf
(22)     Pq $\leftarrow$ assign to queue
(23)     Sort Pq
(24)     Formulate Wf
(25)   end if
(26) end for
```

Algorithm 2: Composition of objects based on social relationship.

as $\Sigma\mathrm{M}$ is loaded, the first step is to compose objects based on their relationship types. This way, the SPARQL queries are executed; the first query is denoted as Q to retrieve the available service templates and the second query is represented as Q' to extract the relationships associated with the services. From line (5) to (11), first, the social relationship is searched among objects and then the found relative objects are composed in groups. The design philosophy employed over here is to assign a ranking to each selected group based on the relationship type that helps in a more efficient

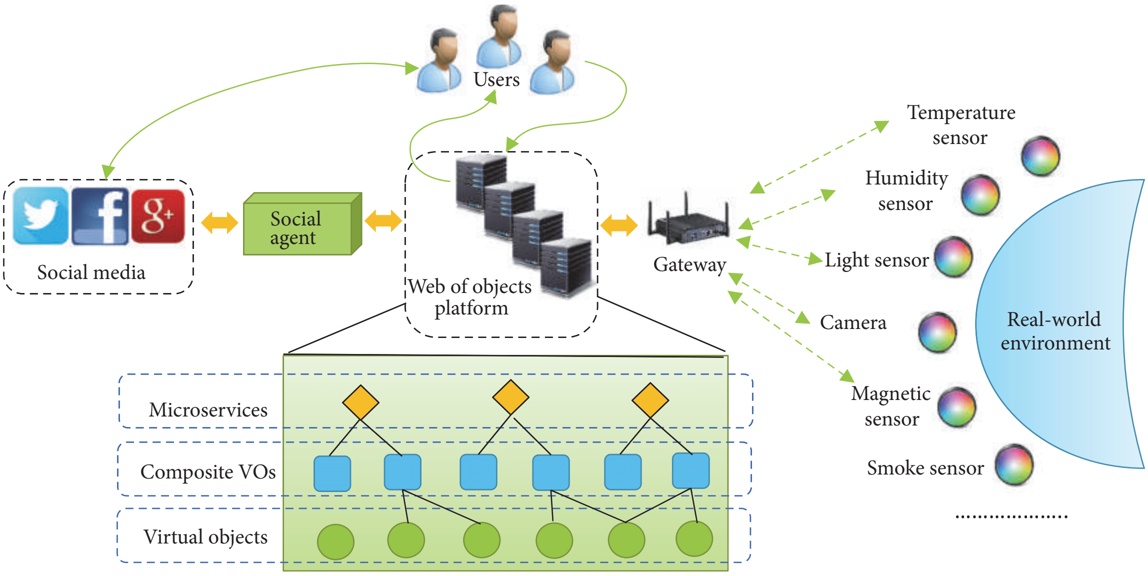

Figure 7: Use case scenario.

composition of service objects (from line (12) to (18)). The last part of Algorithm 2 will continue to iterate to compose the object composition workflow by checking first the assigned relationship ranking with the defined threshold, and then the selected composition workflow graphs are sorted and stored in queues for service execution.

4. Use Case and Prototype Implementation

In this section, first, to realize IoT service provisioning based on social web objects, a use case about IoT enabled museum environment has been presented and then the details on the prototype implementation are furnished.

4.1. Use Case Scenario. This use case is based on a user experience in a public museum where the IoT environment is already set up with the WoO platform incorporating web objects that use social relationship model and a social agent which obtains the user's profile and his friendship details from social media (as shown in Figure 7). The user enters the premises of the museum and his smartphone is prompted with a beacon signal to recommend service app installation where he accepts the request. The phone app incorporates a microservice which establishes a social relationship (GSTOR) with museum visit service objects. The system monitors the interaction of the user in the museum environment; the installed camera network and motion detectors detect the user presence and this feed is collected by the museum visit service. The user shares his current location with the system, which is used by the location navigator service to assist him in moving around the museum by providing the direction to the locations where different museum items are situated. It is worth noticing that the VOs for museum visit service and location navigator service are bound in CLOR relationship due to the same location points. The user shares his social network profile which is used to acquire the list of friends and choices for things the user is interested in. This information is acquired by the recommendation service which suggests the best things to do with respect to the user's interests, making his visit most enjoyable. Recommendation service suggests facilities available in the museum such as places a user should visit, cafeteria food menu, and prices for the day. This service also shares with the user the past experiences of his friends at the museum to help him get most out of his visit. Moreover, personalization service provides special features such as adjusting contents on displays or optimizing the luminosity, to support users with disabilities. To enable these service features, objects exploit the SVOR relationship. Also, the fire management service maintains the status of all VOs that can detect an unusual fire situation in the museum. The temperature VOs acquire readings from temperature sensors and smoke detector VOs get a feed from smoke detection sensors. These VOs form a CLOR relationship with each other in the common proximity. Fire situation monitoring CVO acquires VOs data. One or more CVOs are managed by the fire management service. In case of a fire breakout, the fire management service executes with the data from the fire detection CVOs and the temperature and smoke detector VOs. Meanwhile, services incorporate high-level functionality to react on the data provided by virtual entities in the system. It is important to notice that sharing of information based on social relationships has become easy in the proposed use case. That is because the user information and device information are communicated with social links of web objects and are utilized by the services.

4.2. Proof-of-Concept Details. To evaluate the proposed design, a proof of concept has been furnished, which realizes services based on web objects as considered in the above use case. These objects include, for instance, fire situation monitoring CVO to handle emergency fire breakout situations, location navigator CVO to find the shortest path to the user

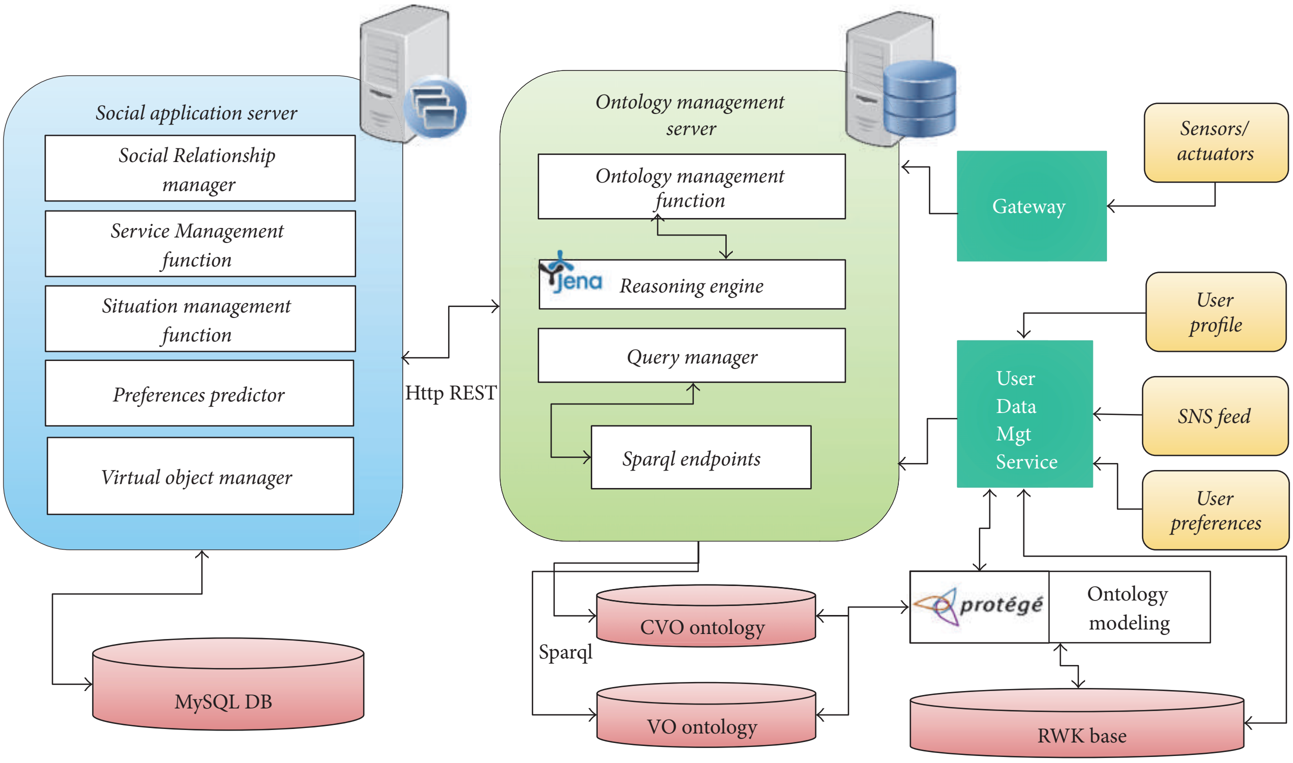
Social application server
Social Relationship manager
Service Management function
Situation management function
Preferences predictor
Virtual object manager
Http REST
MySQL DB
Ontology management server
Ontology management function
Jena
Reasoning engine
Query manager
Sparql endpoints
Sparql
CVO ontology
VO ontology
Gateway
Sensors/ actuators
User profile
User Data Mgt Service
SNS feed
User preferences
protégé
Ontology modeling
RWK base

Figure 8: Prototype implementation components.

intended destinations, and user profile reader CVO which acquires user personal information from the social media profile. Several other CVOs have been developed that acquire diverse data and metadata from VOs where each represents a different RWO. These VOs include a user VO, smartphone VO, camera VO, wristband VO, motion detector VO, and others that are associated with the chosen environment (such as CO_2 sensor VO, humidity VO, temperature VO, and light VO). Each VO has several instances; for example, temperature sensor instances are represented as temperatureSesor0001 and temperatureSesor0002. Further VO data is collected and analyzed by CVOs with the rules incorporated in the ontologies. CVOs composed in microservices help analyze different situations and initiate inference over data using inference engines. Considering the system requirement, several aspects have been evaluated, including identification of web objects for effective IoT service provisioning, finding the social relationships among web objects, efficient discovery of web objects based on these relationships, and composition and harmonization of web objects.

4.3. Prototype Implementation Details. To analyze the feasibility of the proposed architecture, a prototype on the discussed use case has been implemented. Figure 8 illustrates the components of the prototype, which constitute social application server (SAS), ontology management server (OMS), and databases to hold VO and CVO graphs that are supported with SPARQL endpoints. The gateway connecting sensors and actuators has been incorporated in the implementation. Additionally, user data management services are implemented to get user profile and preferences history along with social network data.

The application server comprises five major components: social relationship manager (SRM), service management function (SMF), situation/context management function (CMF), preference predictor, and virtual object manager (VOM). The SMF is an entry point for the service request; it interprets and evaluates the service request. It is also responsible for matching the service request with the available template. SMF communicates with VOM to check the VO or CVO availability. A compiled list of VOs is provided to SMF which generates object mashups to satisfy the service request. The execution of VO mashup graph is done using microservices. The second function, VOM, maintains instances of CVOs and associated VOs and coordinates with OMS to use available ontologies in the system. It records the association of services and VOs and enables the reuse of VOs by more than one service; it also resolves conflicts using conflict resolution mechanism. VOM is implemented using RESTful web services. Further, SRM is implemented to maintain object relationships in the form of RDF concepts represented in ontologies. It also involves several microservices, each implemented for a specific task. For example, group management microservice is developed to group objects that have similar characteristics. Similarly, other tasks are distributed in different microservices; for example, link management microservice is used to identify the dependency among objects. Relationship observer microservice is implemented to monitor the objects when they change their state at runtime, whereas to search a web

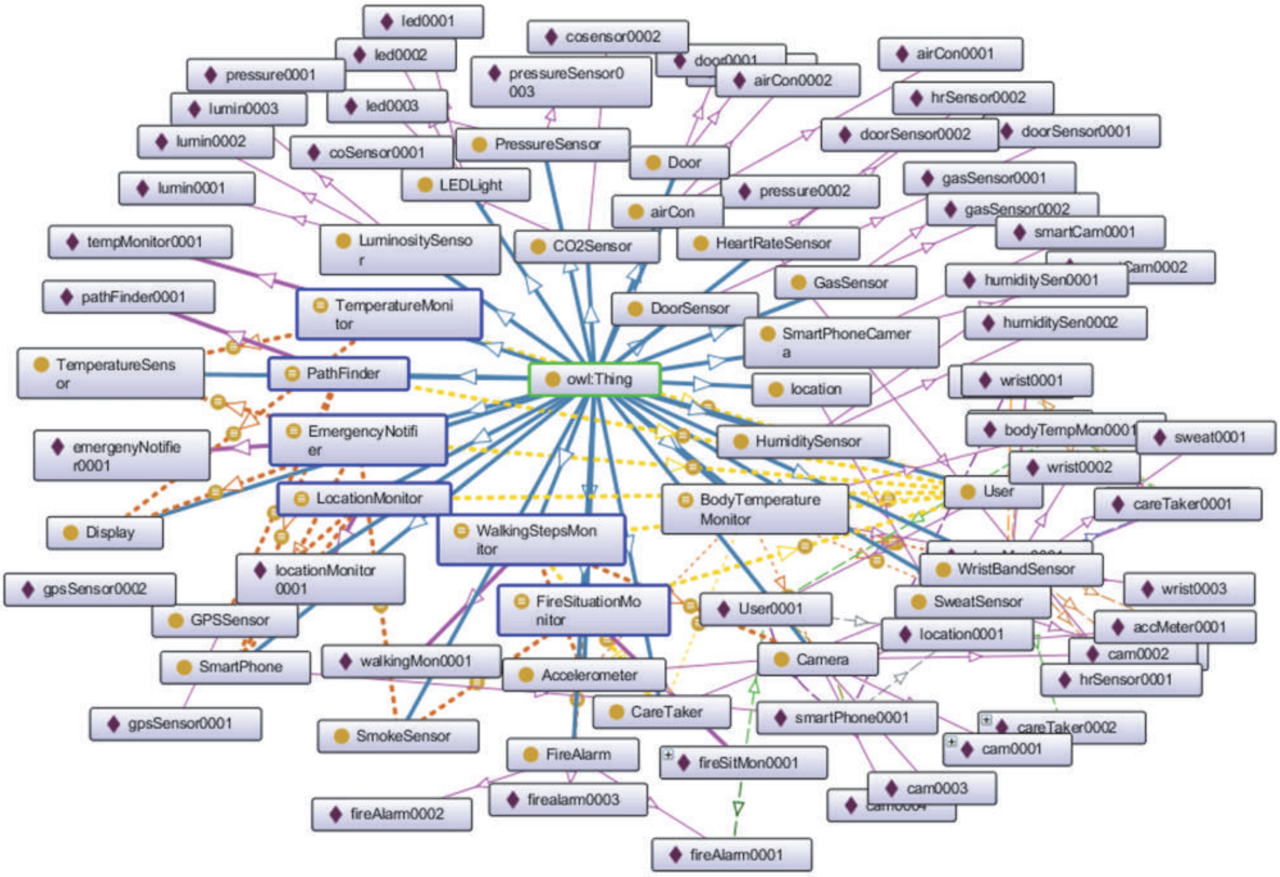
led0001
led0002
cosensor0002
door0001
airCon0001
pressure0001
pressureSensor0003
airCon0002
hrSensor0002
lumin0003
led0003
doorSensor0002
doorSensor0001
lumin0002
coSensor0001
PressureSensor
Door
gasSensor0001
lumin0001
LEDLight
pressure0002
airCon
gasSensor0002
smartCam0001
tempMonitor0001
LuminositySensor
CO2Sensor
HeartRateSensor
smartCam0002
GasSensor
humiditySen0001
pathFinder0001
TemperatureMonitor
DoorSensor
humiditySen0002
SmartPhoneCamera
TemperatureSensor
PathFinder
owl:Thing
location
wrist0001
EmergencyNotifier
HumiditySensor
bodyTempMon0001
sweat0001
emergenyNotifier0001
wrist0002
User
LocationMonitor
BodyTemperatureMonitor
careTaker0001
Display
WalkingStepsMonitor
WristBandSensor
gpsSensor0002
locationMonitor0001
wrist0003
FireSituationMonitor
SweatSensor
GPSSensor
User0001
location0001
accMeter0001
SmartPhone
walkingMon0001
Camera
cam0002
Accelerometer
hrSensor0001
gpsSensor0001
CareTaker
smartPhone0001
careTaker0002
SmokeSensor
cam0001
FireAlarm
fireSitMon0001
cam0003
fireAlarm0002
firealarm0003
fireAlarm0001

Figure 9: Ontological model of social web objects.

object relationship, the lookup microservice is used. The CMF function has been developed to process the situational information and it selects the most suitable VOs based on the context. The situational information contains the current location of the user, location of the RWOs, and the area under observation. CMF helps form a VO configuration specific to the user current situation. Additionally, to incorporate the real-world facts about the user and the situation, RWK base has been developed which is updated continuously. This also helps in adapting services according to the current situation. VOs provided data is accumulated in RWK model and the reasoning engine is used to infer new facts based on the existing data. The last function of social application server is the preference predictor which provides predictions by using machine learning methods. To implement the learning function, Weka Library [42] has been used. For the objective of demonstration, user preference patterns are collected from the social data and user preference history is used to predict the future patterns.

On the other hand, OMS deals with the semantic representation of VOs. The first functional component of OMS is the ontology management service, which is used to handle VO and CVO ontologies and provides an interface to use VO graphs. Further, to manage ontologies in the developed system, we have used Apache Jena framework [43], and to model VO and CVO ontologies, Protégé [44] has been used. Models generated in Protégé are utilized by ontology management service and stored in OWL format in semantic data stores. To take full advantage of ontology-based implementation, a reasoning engine has also been used to get extra facts related to the modeled concepts. Reasoning engine helps to grow RWK facts stored in VO and CVO ontologies. For the proof of concept, the prototype has used Apache Jena inference engine along with Jena provided reasoner. For the deployment of persistent storage of the RDF data, Jena TDB has been used in our prototype, which is the Jena native high-performance triple store with API support [43]. Query manager is another important component of OMS which provides an interface to query CVO and VO ontologies saved in RDF/XML format, using SPARQL. Also, web services are implemented for the execution of SPARQL queries such as extracting and updating RDF triples in the database. Other than SPARQL query processing, semantic matching of existing concepts is also performed.

Moreover, to achieve interoperable relationships among web objects, a semantic ontology model has been developed as shown in Figure 9. The model constitutes the VO and CVO ontologies and their relationships. Several CVOs have been instantiated, for example, the location navigator, fire situation monitor, emergency notification, and many others that are

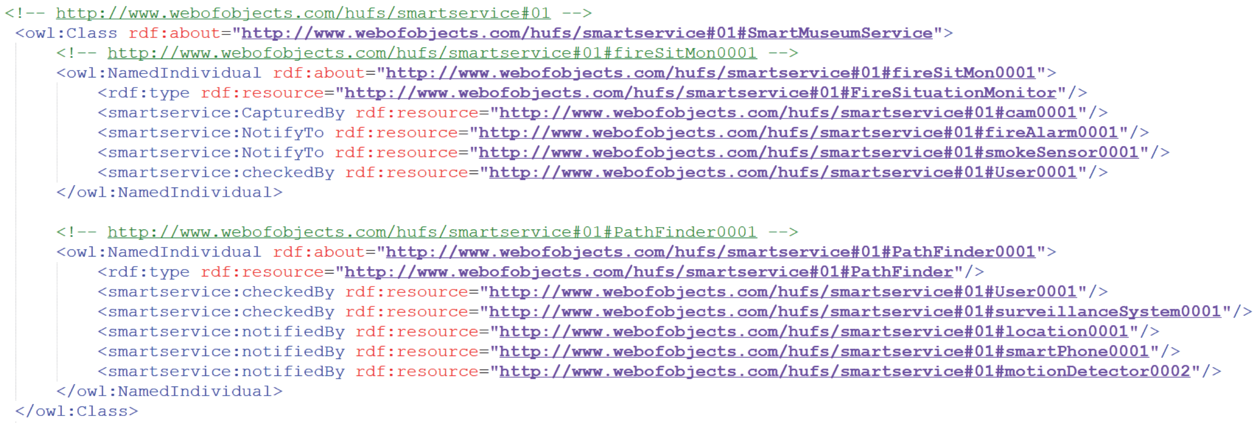

```
<!-- http://www.webofobjects.com/hufs/smartservice#01 -->
 <owl:Class rdf:about="http://www.webofobjects.com/hufs/smartservice#01#SmartMuseumService">
     <!-- http://www.webofobjects.com/hufs/smartservice#01#fireSitMon0001 -->
     <owl:NamedIndividual rdf:about="http://www.webofobjects.com/hufs/smartservice#01#fireSitMon0001">
         <rdf:type rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#FireSituationMonitor"/>
         <smartservice:CapturedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#cam0001"/>
         <smartservice:NotifyTo rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#fireAlarm0001"/>
         <smartservice:NotifyTo rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#smokeSensor0001"/>
         <smartservice:checkedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#User0001"/>
     </owl:NamedIndividual>

     <!-- http://www.webofobjects.com/hufs/smartservice#01#PathFinder0001 -->
     <owl:NamedIndividual rdf:about="http://www.webofobjects.com/hufs/smartservice#01#PathFinder0001">
         <rdf:type rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#PathFinder"/>
         <smartservice:checkedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#User0001"/>
         <smartservice:checkedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#surveillanceSystem0001"/>
         <smartservice:notifiedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#location0001"/>
         <smartservice:notifiedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#smartPhone0001"/>
         <smartservice:notifiedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#motionDetector0002"/>
     </owl:NamedIndividual>
 </owl:Class>
```

FIGURE 10: CVOs and VOs used in the ontology.

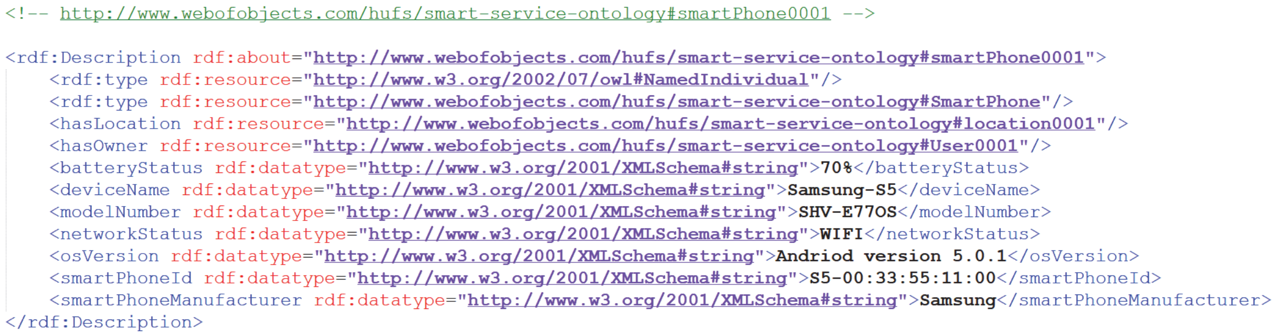

```
<!-- http://www.webofobjects.com/hufs/smart-service-ontology#smartPhone0001 -->

<rdf:Description rdf:about="http://www.webofobjects.com/hufs/smart-service-ontology#smartPhone0001">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#NamedIndividual"/>
    <rdf:type rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SmartPhone"/>
    <hasLocation rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#location0001"/>
    <hasOwner rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#User0001"/>
    <batteryStatus rdf:datatype="http://www.w3.org/2001/XMLSchema#string">70%</batteryStatus>
    <deviceName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Samsung-S5</deviceName>
    <modelNumber rdf:datatype="http://www.w3.org/2001/XMLSchema#string">SHV-E770S</modelNumber>
    <networkStatus rdf:datatype="http://www.w3.org/2001/XMLSchema#string">WIFI</networkStatus>
    <osVersion rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Andriod version 5.0.1</osVersion>
    <smartPhoneId rdf:datatype="http://www.w3.org/2001/XMLSchema#string">S5-00:33:55:11:00</smartPhoneId>
    <smartPhoneManufacturer rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Samsung</smartPhoneManufacturer>
</rdf:Description>
```

FIGURE 11: Smartphone VO used in the ontology.

based on the composition of numerous VOs. This development has been supported by OMS and semantic web tools such as Jena. Using OMS, CRUD operations are performed like create, update, and delete on CVOs and VOs using SPARQL. In the developed semantic relationships, two CVO instances, namely, "FireSitMon0001" and "PathFinder0001," have been used in the code snippet provided in Figure 10, for fire situation monitoring and location navigation, respectively. We can see the data properties and object properties associated with each object to maintain the relationship with other objects in the system. Further, to understand the VO representation, a smartphone VO "SmartPhone0001" is shown in Figure 11 with the defined properties. A list of VOs used in our prototype has been given in OWL description as shown in Figure 12.

In the test bed environment, the sensor to the gateway connectivity is enabled using ZigBee and BLE. Another source of information that has been incorporated to provide user-centric service capability is SNS feed, which consists of Twitter and Facebook data collected using social media APIs. On the other hand, user profile and preferences are updated using data management service, and this information is used to grow RWK base.

To test the prototype, an android application has been developed. After the user logs in to the app for WoO based assisted services, the app starts the configuration of services according to the user profile. A list of services is presented on the user's screen from which the user can select any service. These services include museum visit service, location navigator service, recommendation service, and emergency notification. The selected service starts collecting information with the interaction of the user or when the devices start pushing values about any change in the observed environment. Based on these observed parameters, alerts are generated by the app; for example, the museum visit service generates alerts about the available facilities in the museum and the environmental conditions including the indoor and outdoor temperatures. On the other hand, the location navigator service helps the user to navigate around the museum by showing directions to several visitor attractions. Moreover, the emergency notification service triggers alerts when a temporary fire is created to mimic an unusual situation such as fire breakout. The resulting fire increases the temperature reading which is detected by the temperature sensor. This change in reading is forwarded through the gateway and analyzed by the fire situation monitor CVO which checks its threshold value and generates an emergency notification on the user's app screen. Also, the recommendation service is used to generate suggestions about the museum events and the places to visit.

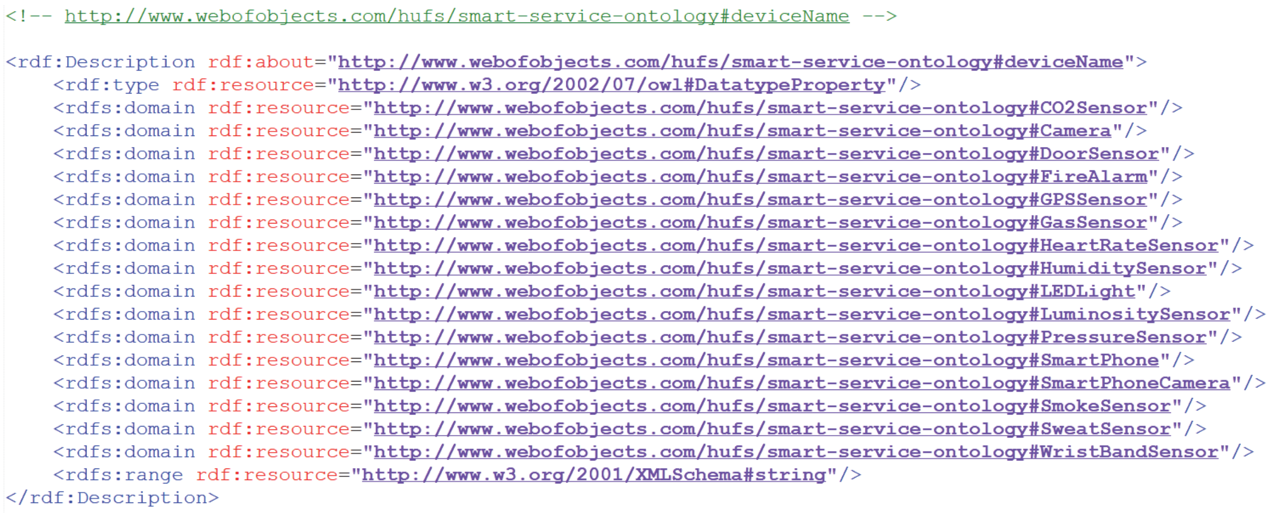

```
<!-- http://www.webofobjects.com/hufs/smart-service-ontology#deviceName -->

<rdf:Description rdf:about="http://www.webofobjects.com/hufs/smart-service-ontology#deviceName">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#CO2Sensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#Camera"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#DoorSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#FireAlarm"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#GPSSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#GasSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#HeartRateSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#HumiditySensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#LEDLight"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#LuminositySensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#PressureSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SmartPhone"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SmartPhoneCamera"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SmokeSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SweatSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#WristBandSensor"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</rdf:Description>
```

Figure 12: List of VOs used in the ontology.

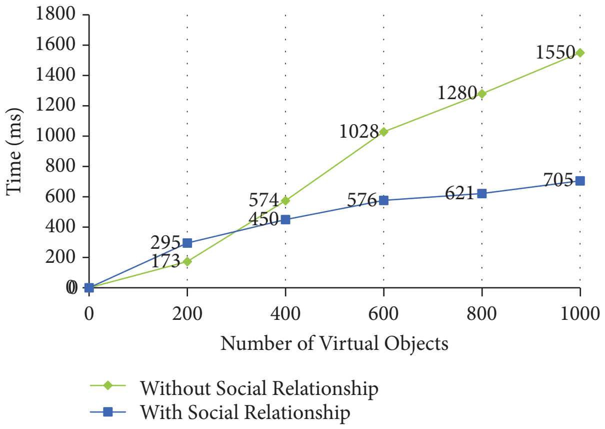

Figure 13: VO discovery time with the increasing number of virtual objects.

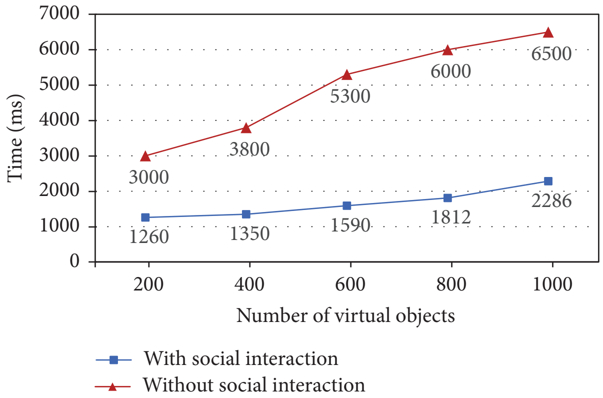

Figure 14: Service execution time with the increasing number of virtual objects.

4.4. Performance Evaluation. The proposed architecture has been analyzed from two different perspectives: system scalability and resource consumption. The results with respect to the experiments and the performance criterion are discussed in this section. The first experiment involves the analysis of the time required to discover the virtual objects in the proposed system with or without social relationship criterion as shown in Figure 13. As it is apparent from the results, initially, to form a social graph, VO discovery process incurs delay due to the communication for establishing social relationship links. However, direct lookup for VOs in the registry at the start of the discovery process comparatively requires less time, but as the number of VOs increases, the delay rises. Traversing the social links to discover objects as the number of virtual objects grows considerably reduces the total lookups in the registry database.

The second experiment provides an analysis of the execution time required for the CVOs as shown in Figure 14. To test this in a real environment, an android application has been developed. The app allows access to the parameters from the available VOs. When the app requires retrieving observations, it iterates through VOs in the repository using a query interface. This requires high CPU utilization with more execution time required to retrieve the sensor values against each app generated request. This is due to the fact that the decision for the selection of VOs has been done at runtime; it requires more time to select the VO graphs, execute them, and get the relevant data from the VOs. However, when object interaction is based on the social relationship, it requires less time for the retrieval as friends in the social network share the data on other friends' requests, which results in less computation as compared with the previous scheme.

Moreover, Figure 15 depicts the time to discover VOs in four different types of services. These services are the museum visit service, the location navigator service, the emergency notification service, and the recommendation service. The discovery time varies from one service to another depending on the number of VOs used. However, as compared with the impact of social connections, it can be

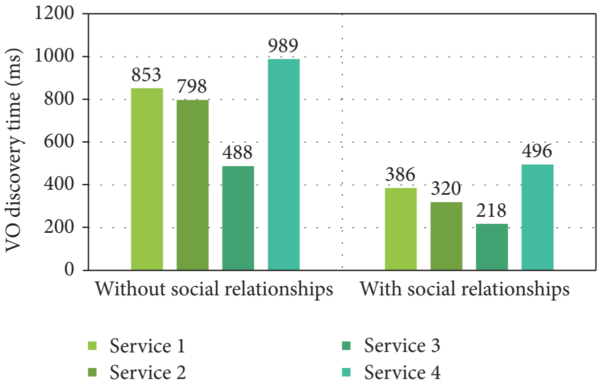

Figure 15: VO discovery time for four different services with or without social relationship utilization.

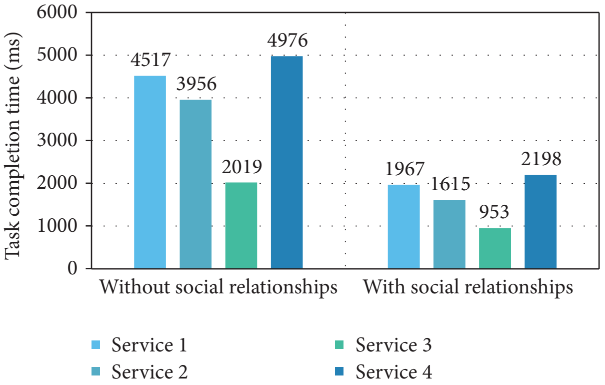

Figure 16: Task completion time required for each service with or without social relationship exploitation.

realized that the services can considerably reduce the time of object discovery and execution by using the relationship model. This is due to the fact that with social association among web objects a service does not have to query the registry for all VOs and rather VOs in the same social graph are utilized. Also, VOs collected observations are shared to save the service time for probing each VO. To view this, Figure 16 shows the overall time required for services to execute their tasks. An important factor to notice here is that the overall task completion time is also affected when social links are used to collect observations from VOs that are used in the service execution.

5. Conclusion

The Internet of Things is bringing the next technological revolution by connecting billions of objects on the earth and providing intelligent IoT services. However, it also carries two most important challenges: the first one is the complexity of handling a huge number of heterogeneous objects and the second one is how to deal with the monolithic approaches for providing services over existing IoT infrastructure. To address these challenges, we have proposed an integrated design based on the principles of SIoT, where the complexity has been handled using socially connected web objects; to overcome the monolithic approaches, microservices that can compose new service features independently have been adopted. In the proposed design, a social relationship model has been presented, which enables the efficient discovery of web objects and reduces the complexity of service provisioning process with the algorithms to discover and compose web objects. A semantic ontology model has also been developed to realize the interoperable social interaction among heterogeneous objects. Finally, a prototype implementation based on a use case scenario has been demonstrated; to evaluate the system with respect to the performance issues, experimental results pertaining to the VO discovery and service execution time have been rendered.

Details of Notations Used in Algorithms

$\Sigma\phi$:	Service object ontologies based on semantic representation where ϕ is a replacement for M_μ which represents microservices M_c which represents CVO and M_v are VO ontologies
$\mathbb{II}_L$:	Collection or list of all relationships that are retrieved from service objects
$\mathbb{II}_U$:	List of the objects that do not have a relationship or are not associated
$\hat{S}$:	Social relationship graph of objects
U:	Correlation union of templates
$\underline{\mu}$₸:	All service objects returned in response to the query request
₸:	All objects returned in response to the query request
μ:	The single instance iterated from the collection of microservices
M:	Data model based on the specified ontology, where ΣM_s represents service data model, ΣM_c represents context data model, and user profile model is represented as ΣM_u
Q:	SPARQL query to retrieve the available service templates for user rating
Q':	SPARQL query to retrieve the relationship associated with services
$\mathbb{R}_e$:	Registry entries of service objects
C_R:	Relationship associated with CVO
V_R:	Relationship associated with VO
R_c:	All possible relationships that can be associated with CVO
R_v:	All possible relationships that can be associated with VO
O_i:	Iterator object instance for iterating the list of CVO and VO objects
$\mathbb{Z}$:	Set of objects' social relationships
σi:	Iteration item of the list of service objects
λi:	An instance of iteration items in the list of service objects
$\mathbb{II}_{\mathrm{Lm}}$:	List of all matched service items
ϖi:	An instance of iteration items in the list of matched service items

$\mathbb{III}_{\mathrm{LR}}$: List of all matched ranking
∂: Ranking value assigned to a service object
σj: Iterator item for the ranking list
κ: Threshold to rank a service object
Wf: Workflow for the composition of service objects
Pq: Priority queue to store ranked object instances
R∂: Assigned relationship ranking.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by Korea Institute for Advancement of Technology (KIAT) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) (N040800001, Development of Web Objects Enabled EmoSpaces Service Technology).

References

[1] M. Nitti, V. Pilloni, G. Colistra, and L. Atzori, "The virtual object as a major element of the internet of things: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1228–1240, 2016.

[2] J. Surowiecki, *The Wisdom of Crowds*, Anchor Books, 2005.

[3] L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The social internet of things (SIoT)—when social networks meet the internet of things: concept, architecture and network characterization," *Computer Networks*, vol. 56, no. 16, pp. 3594–3608, 2012.

[4] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi, "The cluster between internet of things and social networks: Review and research challenges," *IEEE Internet of Things Journal*, vol. 1, no. 3, pp. 206–215, 2014.

[5] L. Atzori, D. Carboni, and A. Iera, "Smart things in the social loop: Paradigms, technologies, and potentials," *Ad Hoc Networks*, vol. 18, pp. 121–132, 2014.

[6] L. Atzori, I. Antonio, and G. Morabito, "Making things socialize in the Internet -Does it help our lives?" in *Kaleidoscope 2011: The Fully Networked Human?-Innovations for Future Networks and Services (K-2011), Proceedings of ITU. IEEE*, 2011.

[7] L. Atzori, A. Iera, and G. Morabito, "SIoT: giving a social structure to the internet of things," *IEEE Communications Letters*, vol. 15, no. 11, pp. 1193–1195, 2011.

[8] A. M. Vegni and V. Loscrí, "A survey on vehicular social networks," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, article no. A3, pp. 2397–2419, 2015.

[9] L. Krause, *Microservices: Patterns and Applications: Designing Fine-Grained Services by Applying Patterns*, 2015.

[10] A. Krylovskiy, M. Jahn, and E. Patti, "Designing a smart city internet of things platform with microservice architecture," in *Proceedings of the 3rd International Conference on Future Internet of Things and Cloud (FiCloud '15)*, pp. 25–30, Rome, Italy, August 2015.

[11] B. Butzin, F. Golatowski, and D. Timmermann, "Microservices approach for the internet of things," in *Proceedings of the 21st IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2016*, Germany, September 2016.

[12] N. Dragoni, I. Lanese, S. T. Larsen, M. Mazzara, R. Mustafin, and L. Safina, "Microservices: How To Make Your Application Scale," *Computing Research Repository*, 2017.

[13] "Microservices Guide Pages - Martin Fowler," 2017, https://martinfowler.com/articles/microservices.html.

[14] "Web of Objects. ITEA 3 · Project · 10028 , D2.1: State-of-the-Art Relevant to the Web of Objects," 2017, https://itea3.org/project/web-of-objects.html.

[15] "Y.4452: Functional framework of web of objects," 2017, http://www.itu.int/rec/T-REC-Y.4452-201609-P.

[16] L. E. Holmquist, F. Mattern, B. Schiele, P. Alahuhta, M. Beigl5, and H.-W. Gellersen, "Smart-its friends: a technique for users to easily establish connections between smart artefacts," in *Ubicomp 2001: Ubiquitous Computing*, pp. 116–122, Springer, Berlin, Germany, 2001.

[17] D. Guinard, M. Fischer, and V. Trifa, "Sharing using social networks in a composable Web of Things," in *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2010*, pp. 702–707, Germany, April 2010.

[18] M. Kranz, L. Roalter, and F. Michahelles, "Things that twitter: social networks and the internet of things.' What can the Internet of Things do for the Citizen (CIoT)," in *Proceedings of the 8th International Conference on Pervasive Computing (Pervasive 2010)*, 2010.

[19] R. Girau, S. Martis, and L. Atzori, "Lysis: A platform for iot distributed applications over socially connected objects," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 40–51, 2017.

[20] A. Pintus, D. Carboni, and A. Piras, "Paraimpu: A platform for a social Web of Things," in *Proceedings of the 21st Annual Conference on World Wide Web, WWW'12*, pp. 401–404, France, April 2012.

[21] I. Farris, R. Girau, L. Militano et al., "Social Virtual Objects in the Edge Cloud," *IEEE Cloud Computing*, vol. 2, no. 6, pp. 20–28, 2015.

[22] R. Girau, M. Nitti, and L. Atzori, "Implementation of an experimental platform for the social internet of things," in *Proceedings of the 7th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2013*, pp. 500–505, July 2013.

[23] M. A. Jarwar, R. A. Abbasi, M. Mushtaq et al., "CommuniMents: A framework for detecting community based sentiments for events," *International Journal on Semantic Web and Information Systems*, vol. 13, no. 2, pp. 87–108, 2017.

[24] S. Newman, "Building microservices: designing fine-grained systems," 2015.

[25] "The Single Responsibility Principle - Programmer 97-things," 2017, http://programmer.97things.oreilly.com/wiki/index.php/The_Single_Responsibility_Principle.

[26] F. Viktor, "The DevOps 2.0 Toolkit: Automating the Continuous Deployment Pipeline with Containerized Microservices," 2016.

[27] "iCore: Internet Connected Objects for Reconfigurable Ecosystems, European FP7 Project," 2017, http://cordis.europa.eu/project/rcn/100873_en.html.

[28] D. Namiot and M. Sneps-Sneppe, "On microservices Architecture," *International Journal of Open Information Technologies*, vol. 2, no. 9, 2014.

[29] D. Bonino, M. T. D. Alizo, A. Alapetite et al., "ALMANAC: internet of things for smart cities," in *Proceedings of the 3rd*

International Conference on Future Internet of Things and Cloud (FiCloud '15), pp. 309–316, IEEE, Rome, Italy, August 2015.

[30] "Developing Microservices for PaaS with Spring and Cloud Foundry," 2017, https://www.infoq.com/presentations/micro-services-pass-spring-cloud-foundry.

[31] "Microservices in action, Part 2: Containers and microservices — a perfect pair," 2017, https://www.ibm.com/developerworks/cloud/library/cl-bluemix-microservices-in-action-part-2-trs/index.html.

[32] J. E. Kim, X. Fan, and D. Mosse, "Empowering End Users for Social Internet of Things," in *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation - IoTDI '17*, pp. 71–82, Pittsburgh, Pa, USA, April 2017.

[33] Z. U. Shamszaman, S. S. Ara, I. Chong, and Y. K. Jeong, "Web-of-Objects (WoO)-based context aware emergency fire management systems for the Internet of Things," *Sensors*, vol. 14, no. 2, pp. 2944–2966, 2014.

[34] S. Ali, M. G. Kibria, and I. Chong, "WoO enabled IoT service provisioning based on learning user preferences and situation," in *Proceedings of the 2017 International Conference on Information Networking (ICOIN)*, pp. 474–476, Da Nang, Vietnam, January 2017.

[35] M. A. Jarwar, S. Ali, M. G. Kibria, S. Kumar, and I. Chong, "Exploiting interoperable microservices in web objects enabled Internet of Things," in *Proceedings of the Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 49–54, Milan, Italy, July 2017.

[36] M. Kibria, S. Ali, M. Jarwar, S. Kumar, and I. Chong, "Logistic Model to Support Service Modularity for the Promotion of Reusability in a Web Objects-Enabled IoT Environment," *Sensors*, vol. 17, no. 10, p. 2180, 2017.

[37] M. Kibria, S. Fattah, K. Jeong, I. Chong, and Y. Jeong, "A User-Centric Knowledge Creation Model in a Web of Object-Enabled Internet of Things Environment," *Sensors*, vol. 15, no. 9, pp. 24054–24086, 2015.

[38] M. A. Jarwar and . Ilyoung Chong, "Exploiting IoT services by integrating emotion recognition in Web of Objects," in *Proceedings of the 2017 International Conference on Information Networking (ICOIN)*, pp. 54–56, Da Nang, Vietnam, January 2017.

[39] S. Ali, H.-S. Kim, and I. Chong, "Implementation model of WoO based smart assisted living IoT service," in *Proceedings of the International Conference on Information and Communication Technology Convergence, ICTC 2016*, pp. 816–818, October 2016.

[40] S. Kumar, M. G. Kibria, S. Ali, M. A. Jarwar, and I. Chong, "Smart spaces recommending service provisioning in WoO platform," in *Proceedings of the International Conference on Information and Communications (ICIC)*, pp. 311–313, Hanoi, Vietnam, June 2017.

[41] K. M. Alam, M. Saini, and A. El Saddik, "Toward social internet of vehicles: concept, architecture, and applications," *IEEE Access*, vol. 3, pp. 343–357, 2015.

[42] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java," 2017, https://www.cs.waikato.ac.nz/ml/weka/.

[43] "Jena, Apache. &quot;A free and open source Java framework for building Semantic Web and Linked Data applications.&quot;," 2017, http://jena.apache.org/.

[44] Protégé, "A free, open-source ontology editor and framework for building intelligent systems," 2017, https://protege.stanford.edu/.

Journal of Engineering
The Scientific World Journal
International Journal of Rotating Machinery
Journal of Sensors
International Journal of Distributed Sensor Networks
Advances in Civil Engineering
Journal of Control Science and Engineering
Journal of Robotics
Journal of Electrical and Computer Engineering
Advances in OptoElectronics
VLSI Design
International Journal of Navigation and Observation
Modelling & Simulation in Engineering
International Journal of Aerospace Engineering
International Journal of Chemical Engineering
International Journal of Antennas and Propagation
Active and Passive Electronic Components
Shock and Vibration
Advances in Acoustics and Vibration

Hindawi

Submit your manuscripts at
https://www.hindawi.com

Multiple Antenna Technologies

Manar Mohaisen | YuPeng Wang | KyungHi Chang

The Graduate School of Information Technology and Telecommunications

INHA University

ABSTRACT

Multiple antenna technologies have received high attention in the last few decades for their capabilities to improve the overall system performance. Multiple-input multiple-output systems include a variety of techniques capable of not only increase the reliability of the communication but also impressively boost the channel capacity. In addition, smart antenna systems can increase the link quality and lead to appreciable interference reduction.

I. Introduction

Multiple antennas technologies proposed for communications systems have gained much attention in the last few years because of the huge gain they can introduce in the communication reliability and the channel capacity levels. Furthermore, multiple antenna systems can have a big contribution to reduce the interference both in the uplink and the downlink by employing smart antenna technology.

To increase the reliability of the communication systems, multiple antennas can be installed at the transmitter or/and at the receiver. Alamouti code is considered as the simplest transmit diversity scheme while the receive diversity includes maximum ratio, equal gain and selection combining methods. Recently, cooperative communication was deeply investigated as a mean of increasing the communication reliability by not only considering the mobile station as user but also as a base station (or relay station). The idea behind multiple antenna diversity is to supply the receiver by multiple versions of the same signal transmitted via independent channels.

On the other hand, multiple antenna systems can tremendously increase the channel capacity by sending independent signals from different transmit antennas. BLAST spatial multiplexing schemes are a good example of such category of multiple antenna technologies that boost the channel capacity.

In addition, smart antenna technique can significantly increase the data rate and improve the quality of wireless transmission, which is limited by interference, local scattering and multipath propagation. Through shaping the antenna radiation pattern and adaptively adjusting the antenna weight vector, smart antennas improve the communication link quality by increasing the received signal power and suppressing the interference.

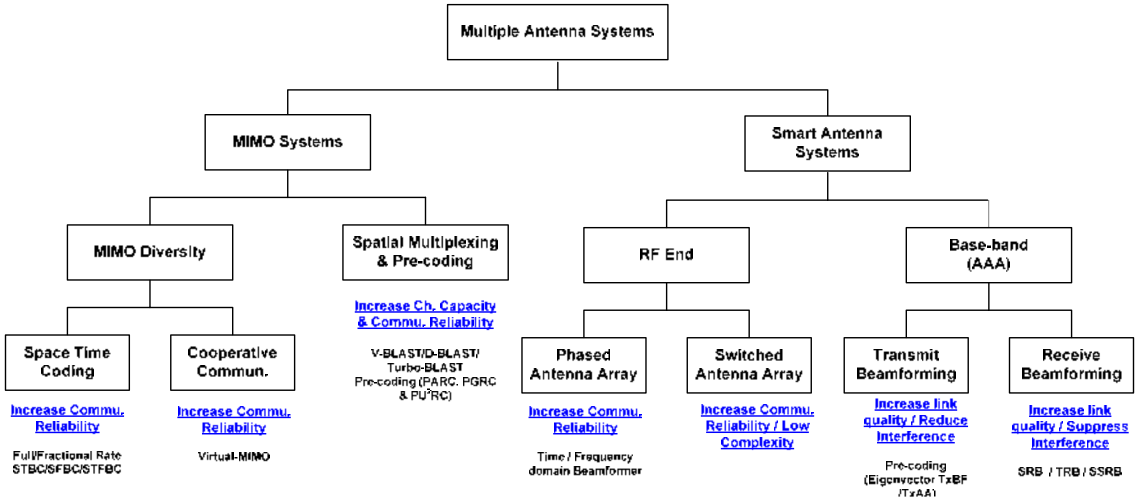
Multiple Antenna Systems
MIMO Systems
Smart Antenna Systems
MIMO Diversity
Spatial Multiplexing & Pre-coding
RF End
Base-band (AAA)
Space Time Coding
Cooperative Commun.
Increase Ch. Capacity & Commu. Reliability
V-BLAST/D-BLAST/ Turbo-BLAST Pre-coding (PARC. PGRC & PU²RC)
Phased Antenna Array
Switched Antenna Array
Transmit Beamforming
Receive Beamforming
Increase Commu. Reliability
Increase Commu. Reliability
Full/Fractional Rate STBC/SFBC/STFBC
Virtual-MIMO
Increase Commu. Reliability
Time / Frequency domain Beamformer
Increase Commu. Reliability / Low Complexity
Increase link quality / Reduce Interference
Pre-coding (Eigenvector TxBF /TxAA)
Increase link quality / Suppress Interference
SRB / TRB / SSRB

Fig. 1. Multiple antenna technologies.

Besides, on-line calibration technique is also adopted to correct the errors due to the distortions and nonlinearity of the radio frequency components in the antenna array system.

Fig. 1 summarizes the different multiple antenna technologies and gives some examples of these technologies.

This paper is organized as follows: in section II we present multiple antenna diversity schemes employed at the transmitter or/and at the receiver. Spatial multiplexing presented by BLAST schemes is detailed in section III. Section IV is dedicated to some advanced multiple input multiple-output (MIMO) systems including multi-user MIMO and cooperative communications. While techniques related to the smart antennas such as phased antenna array, switched beam antenna array, and adaptive antenna array are described in Section V. Finally, we conclude in Section VI.

II. MIMO Diversity

In communication systems, we have to increase the reliability of the communication operation between transmitter and receiver while maintaining a high spectral efficiency. The ultimate solution relies in the use of *diversity*, which can be viewed as a form of redundancy [1]. There are many diversity techniques that can be applied to communication systems; we mention herein time diversity, frequency diversity, and spatial diversity or any combination of these three diversities. In *time diversity*, the same information-bearing signal is transmitted in different time slots where a good gain can be achieved when the duration between the two slots, in which the same symbol is transmitted, is greater than the coherence time of the channel. In *frequency diversity*, the same information-bearing signal is transmitted on different subcarriers where a good diversity gain can be achieved when the separation between subcarriers is greater than the coherence bandwidth.

Finally, in *spatial diversity*, the same information-bearing signal is transmitted or received via different antennas where the maximum gain can be achieved when the fading occurring in the channel is independent (or low correlated). In the receiver, diversity gain can be achieved by combining the redundant signals arriving via independent (or lowly correlated) channels.

Fig. 2 shows some possible combinations of transmit diversity which can be achieved when employing multiple transmit antennas.

In the following section, we present some famous space-time block codes applied at the transmitter side. We present also the combining techniques used when different versions of the information-bearing signal are received. Finally, we present a scheme that includes transmit and receive diversities.

2.1 Space Diversity at the Transmit Side

The basic idea of the use of transmit diversity is to reduce the mobile station (MS) receiver complexity while improving the detection performance.

The pioneering work in the transmit diversity was done by Alamouti where he proposed his famous 2×1 space-time code. Alamouti scheme achieves diversity gain while requiring only a linear decoder.

Later on, Tarokh et al. proposed a generalized theory of the complex orthogonal space-time codes. Based on Tarokh work, more than two antennas can be used and the code rate can be fractional.

In the following we present the two different types of space-time codes.

2.1.1 Complex Orthogonal Space-Time Codes

For this type of space time codes, the following conditions must be satisfied

- Square transmission matrix (number of transmit antennas N_t equal to number of used time slots m)
- A unity code rate (number of used time slots m equals to number of transmitted symbols l)
- Orthogonality of the transmission matrix in the time and space domains ($\mathrm{SS}^{\mathrm{H}} = \mathrm{S}^{\mathrm{H}}\mathrm{S}$) where S^{H} is the conjugate transpose of S.

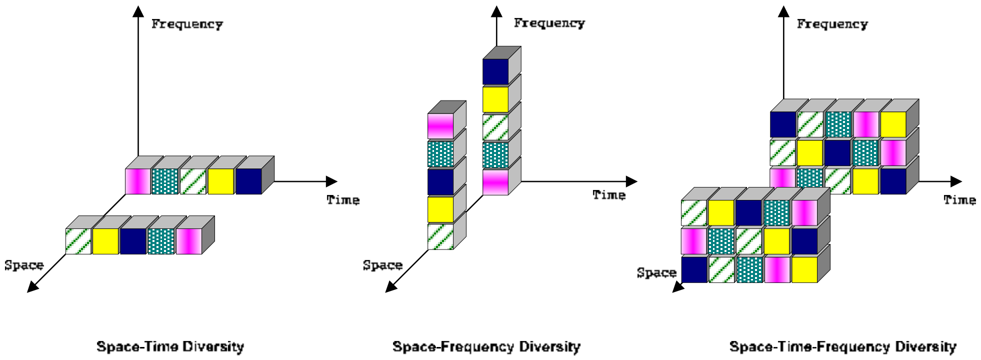

Fig. 2. Transmit diversity.

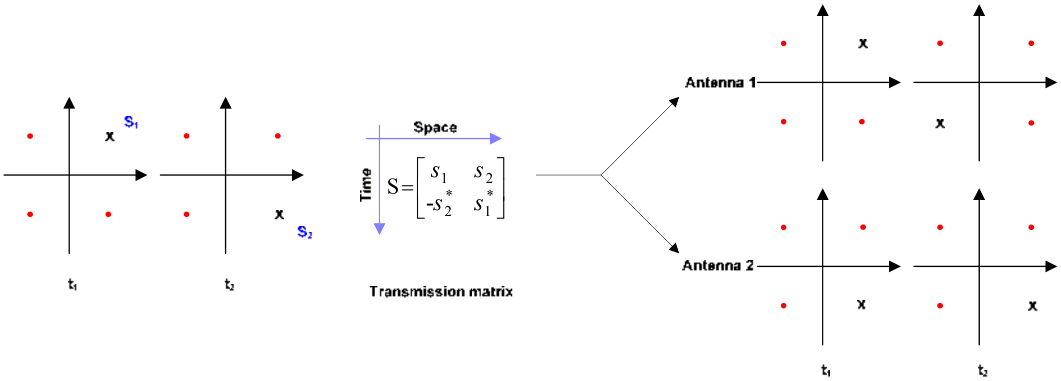

Fig. 3. Alamouti scheme example with QPSK modulation.

As said before, the simplest complex orthogonal space-time code is the Alamouti code which uses two transmit antennas and one receive antenna. Furthermore, Alamouti scheme requires that the fading channel envelope remains constant over two time slots.

Fig. 3 shows an example of the encoding process of Alamouti scheme with QPSK modulation [2], [3].

Fig. 4 shows the receiver structure used for decoding the combined received symbols.

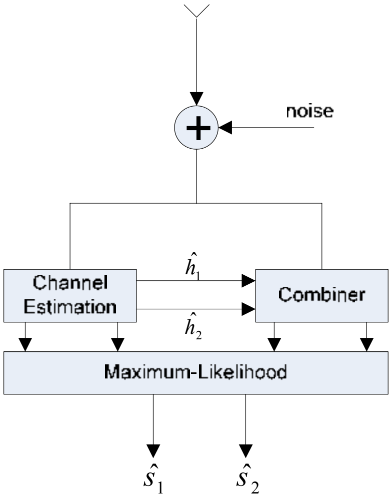

Fig. 4. Alamouti code receiver.

At the receiver the following signals are received (with applying the complex conjugate to the received signal at t_2)

$$\begin{bmatrix} y_1 \\ y_2^* \end{bmatrix} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2^* \end{bmatrix} \qquad (1)$$

The linear combiner multiplies the received symbols by the Hermitian transpose of the channel matrix (for simplicity, we consider that channel is perfectly estimated). The output of the linear combiner is then given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} h_1^2 + h_2^2 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad (2)$$

Maximum-Likelihood (ML) decoder is then applied to get the transmitted symbols. As one can see, the simplicity of the receiver is due to the spatio-temporal orthogonality of the transmission matrix.

A complex orthogonal space-time code using 4 or 8 antennas was proposed by Tarokh et al. in [4].

2.1.2 Generalized Complex Orthogonal Space-Time Codes

The search for space-time codes with more than two antennas was started by Tarokh, Jafarkhani, and Calderbank. Their work has built the basis for a theory of generalized complex orthogonal designs.

Generalized complex orthogonal designs are distinguished from Alamouti code by the following

- A non-square transmission matrix (number of used time slots ≠ number of Tx antennas)
- A fractional code rate (number of transmitted symbols < number of used time slots)
- Orthogonality of the transmission matrix is only guaranteed in the time sense.

As a consequence of these characteristics, the spectral efficiency is reduced and the number of time slots over which the channel should be constant is increased.

The transmission matrix of a generalized complex space-time code with 3 antennas, 4 transmitted symbols and 8 used time slots is given by [5]

$$\mathbf{G_3} = \begin{bmatrix} s_1 & s_2 & s_3 \\ -s_2 & s_1 & -s_4 \\ -s_3 & s_4 & s_1 \\ -s_4 & -s_3 & s_2 \\ s_1^* & s_2^* & s_3^* \\ -s_2^* & s_1^* & -s_4^* \\ -s_3^* & s_4^* & s_1^* \\ -s_4^* & -s_3^* & s_2^* \end{bmatrix} \qquad (3)$$

In the literature, more research was done to increase the rate of the space-time codes. For more details refer to [6].

Table 1 summarizes the difference between Alamouti and space-time code characterized by the transmission matrix $\mathbf{G_3}$.

These shown coding schemes can be transmitted in the space-time domains, space-frequency domains or in space-frequency-Time domains. These coding schemes are thus known as ST, SF, and STF coding, respectively [7].

2.1.3 Cyclic Delay Diversity (CDD)

CDD can be considered as a very simple transmit diversity scheme. CDD can achieve transmit artificial frequency diversity by selecting appropriate transmit delays. In this method, multiuser diversity, obtained by scheduling based on frequency domain channel response, can be improved by adjusting the delay spread (at the transmitter) which is done by controlling the delay values dependent on the channel condition [8], [9].

2.2 Space Diversity at the Receive Side

In space diversity at the receive side, multiple antennas are used in the receiver with sufficient spacing between antennas in such a way mutual correlation between antennas is reduced and as consequence diversity gain is increased [10]. To get diversity gain at the receiver, received signals from different antennas are combined. There are four combining methods, namely, select combining (SC), maximal-ratio combining (MRC), equal-gain combining (EGC), and square-law combining. The first three schemes are linear while the last requires a non-linear receiver.

Fig. 5 shows a simplified block diagram of the linear combining schemes which differ in the weighting vector **w.** In SC, the signal at the branch with maximum signal to noise ratio (SNR) is selected and other received signals are discarded.

The weighting vector $\mathbf{w} = (w_1, w_2, \ldots, w_M)$ is the Nth column of the identity matrix of size M where the Nth branch has the maximum SNR.

Table 1. Comparison between Alamouti and generalized complex space-time codes.

Space-time code	**Number of Tx antennas**	**Number of transmitted symbols, *l***	**Number of used time slots, *m***	**Orthogonality of Tx matrix**	**Rate = *l/m***
S	2	2	2	Spatio-temporal sense	1
G_3	3	4	8	Only temporal sense	1/2

Table 2. Comparison between diversity combining schemes.

Scheme	**Requiring CSI**	**Outage Probability F(x)**	**Application**
SC	No	$\left[1-e^{-x/\bar{y}_c}\right]^M$	No constraints
MRC	Yes	$1-e^{-x/\bar{y}_c}\sum_{k=0}^{M-1}\frac{1}{k!}\left(\frac{x}{\bar{y}_c}\right)^k$	No constraints
EGC	Yes	No closed form for M > 2	No constraints
SLC	No	-	FSK or DS-CDMA

As one can see, the SC scheme does not require any channel information except that of SNR. On the other hand, MRC and EGC schemes require the channel state information (CSI) or a part of it (channel envelope, phase, delay). MRC scheme weights the received signals according to their reliability; a more reliable signal has a high weight while a less reliable signal has a small weight. Also, the channel phase distortion is compensated. Finally, signals are aligned then combined. On the other hand, EGC scheme can be viewed as a simplified version of MRC where signals are weighted equally (i.e. the weighting vector **w** = $[1, 1, \ldots, 1]_M$) then aligned before being combined coherently. In practice, the phase at different branches can't be often estimated. So, EGC and MRC can't be employed. In such situations, square law combining (SLC) can be applied to obtain spatial diversity without requiring phase estimation. Unlike linear combining schemes, SLC scheme can only be applied to modulation schemes which preserve some sort of orthogonality including frequency-shift keying (FSK) or direct-sequence CDMA [5], [11].

Table 2 summarizes a comparison between combining schemes. It is known that, from a performance point of view, MRC is optimum and gives the best performance among the pre-described combining schemes.

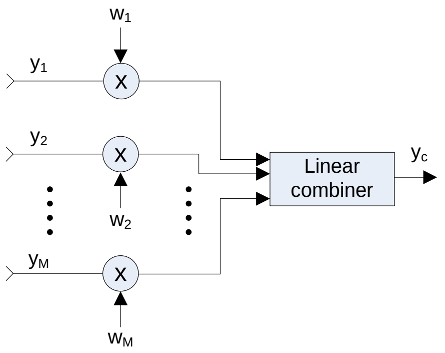

Fig. 5. Simple block diagram of linear combining schemes.

2.3 Combined Transmit/Receive Diversity

Spatial diversity schemes explained in the previous two sections can be combined together to achieve diversity at both receive and transmit sides. An example of such a hybrid spatial diversity scheme is the 2×2 Alamouti/MRC MIMO system [12].

III. Spatial Multiplexing

As shown in the previous section, MIMO diversity can be used in the transmitter or the receiver sides or in both to increase the reliability of the communication. In this section we talk about spatial multiplexing schemes which are for goal to increase the channel capacity.

The most known spatial multiplexing schemes are the BLAST family which includes Vertical-BLAST, Diagonal-BLAST, and Turbo-BLAST. The acronym BLAST stands for "*Bell Laboratories Layered Space-Time*".

3.1 Diagonal-BLAST

D-BLAST was originally proposed by Foschini [13]. In D-BLAST, the symbols to be transmitted are arranged on the diagonals of the space-time transmission matrix where elements under the diagonal are padded with zeros. Fig. 6-a depicts the structure of the D-BLAST transmitter for four transmit antennas. At first, the bit stream is de-multiplexed into four parallel streams which are encoded and modulated independently. Encoded-modulated streams are cycled over time. Equation (4) is an example of the transmission matrix when using four transmit antennas.

$$S = \begin{bmatrix} s_{1,1} & s_{1,2} & s_{1,3} & s_{1,4} & \cdots & s_{1,K-1} & s_{1,K} & 0 & 0 & 0 \\ 0 & s_{2,1} & s_{2,3} & s_{2,4} & \cdots & s_{2,K-2} & s_{2,K-1} & s_{2,K} & 0 & 0 \\ 0 & 0 & s_{3,1} & s_{3,2} & \cdots & s_{3,K-3} & s_{3,K-2} & s_{3,K-1} & s_{3,K} & 0 \\ 0 & 0 & 0 & s_{4,1} & \cdots & s_{4,K-4} & s_{4,K-3} & s_{4,K-2} & s_{4,K-1} & s_{4,K} \end{bmatrix} \qquad (4)$$

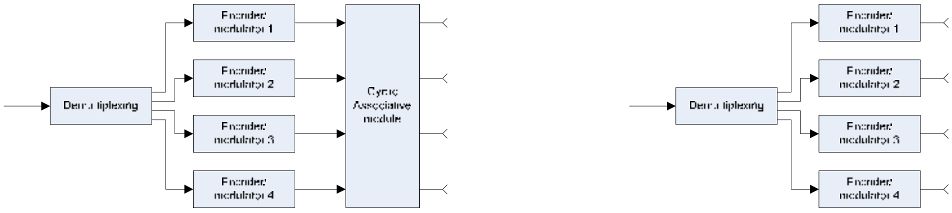
Demultiplexing
Encoder/ modulator 1
Encoder/ modulator 2
Encoder/ modulator 3
Encoder/ modulator 4
Cyclic Associative module
Demultiplexing
Encoder/ modulator 1
Encoder/ modulator 2
Encoder/ modulator 3
Encoder/ modulator 4

(a) D-BLAST

(b) V-BLAST

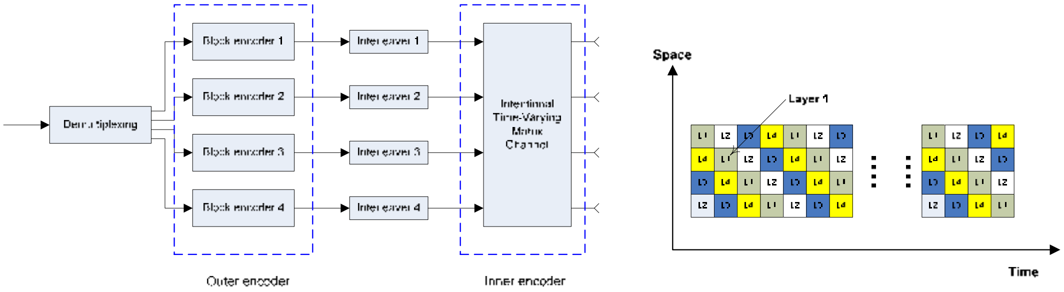
Demultiplexing
Block encoder 1
Block encoder 2
Block encoder 3
Block encoder 4
Interleaver 1
Interleaver 2
Interleaver 3
Interleaver 4
Intentional Time-Varying Matrix Channel
Outer encoder
Inner encoder
Space
Layer 1
Time

(c) Turbo-BLAST

(d) RLST code

Fig. 6. Transmitter block diagrams for BLAST family using four transmit antennas.

Table 3. MIMO systems diversity orders.

MIMO Configuration	Diversity order
STBC	$N_t \times N_r$
BLAST	$N_r - N_t + 1$

The first diagonal of S is transmitted via the first antenna; the second diagonal is transmitted via antenna 2, and so on.

3.2 Vertical-BLAST

A simplified version of D-BLAST was proposed by Wolniansky known as Vertical-BLAST or V-BLAST [14]. In V-BLAST, incoming data stream is de-multiplexed into N_t streams each of which is encoded and modulated independently and sent on an antenna of its own. V-BLAST high-level diagram is depicted in Fig. 6-b where four antennas are used at the transmit side. Compared to D-BLAST, V-BLAST does not include cycling over time, the complexity is significantly reduced. In addition, unlike D-BLAST, V-BLAST does not include any space-time wastage. At the receiver, transmitted symbols can be decoded using *ordered serial interference-cancellation (OSIC) detector*. For the OSIC to work properly, the number of receive antennas N_r must be at least as large as the number of transmit antennas.

3.3 Turbo-BLAST

Turbo-BLAST was first described by Sellathurai and Haykin [15]. The Turbo-BLAST transmitter structure is depicted in Fig. 6-c. The data stream bits are firstly demultiplexed into N_t parallel streams which are encoded independently using the block encoder (outer encoder) (i.e. channel coding). The output streams of the outer encoder are interleaved independently and passed to the inner encoder. The mission of the outer encoder is to achieve *random-layered space-time* (RLST) coding.

The structure of the RLST encoder, with periodical cyclic space-time interleaving is depicted in Fig. 6-d. For optimal performance of the RLST code, the receiver should employ the *maximum a posteriori probability (MAP)* decoding algorithm. Nevertheless, the complexity of the MAP decoding algorithm is very high (increases exponentially with N_t). To decrease the complexity of the receiver, the near-optimal turbo-like receiver can be used. This near-optimal turbo-like receiver is known as *iterative detection and decoding (IDD) receiver*.

Before going further, we list in Table 3 a comparison between diversity order of the different space-time coding and the BLAST family schemes.

IV. Advanced Topics

4.1 Single and Multi CodeWord MIMO

In single codeword (SCW) MIMO, an encoded packet is distributed across many streams to form the MIMO transmission. Feedback is used to control the rank of the MIMO transmission (number of streams used) as well as the overall rate of transmission. In multiple codeword MIMO, several separately encoded packets are transmitted independently over the multiple streams. Here the rate of each stream can be controlled with feedback [16] and [17].

4.2 Single-User MIMO and Multi-User MIMO

In single-user MIMO, already explained techniques in previous sections are used where the channel capacity grows linearly with *min(Nt, Nr)* [18].

For multi-user MIMO, which is of high interest research topic, it was shown that for N_t transmitting antennas (at the base station) and N_r users, the same overall capacity can be achieved. This later work was encouraged by applying dirty paper coding [19] where results showed that if the transmitter knows the interfering signal, then the channel capacity will not be affected by the presence of the interference [20]. On the other hand, multi-user MIMO can integrate beamforming to apply spatial division multiple access (SDMA).

4.3 Cooperative Communication and Virtual MIMO

In cooperative communication, a mobile can act as both a user and relay. As consequence, mobile sends to the base station its own data bits and some of other mobiles (sometimes called partner) information bits. Fig. 7 shows a cooperative cellular system where for simplicity we consider three cooperative users and one base station [21]. As depicted in Fig. 7, user 1 cooperates with users 2 and 3 to send its own information. As a result, the overall cooperative system can be seen as virtual-MIMO (V-MIMO) and in the above example it is $3\times N_{BS}$ MIMO system (for the uplink) where N_{BS} is the base station's number of receive antennas. Users 2 and 3 can simply amplify and forward user 1 received information or detect and forward [22]. Another method of cooperation is the coded cooperation where different coded portions are sent via different fading channels [23].

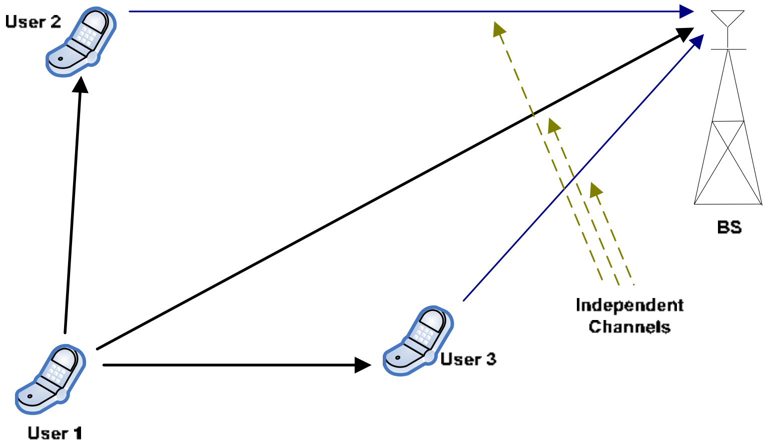

Fig. 7. Cooperative communication and virtual-MIMO.

4.4 Pre-Coded MIMO with Rank Adaptation

4.4.1 Per Antenna Rate Control (PARC)

PARC can be considered as a closed-loop MIMO system where transmitter uses *channel quality indication* (CQI) fed by the receiver to select the best modulation and coding schemes per antenna. Fig. (8) shows a general PARC transmitter structure with 4 transmit antennas [24].

4.4.2 Per Group Rate Control (PGRC)

In PARC, a CQI feedback is necessary for each transmit antenna. This increases the uplink overhead.

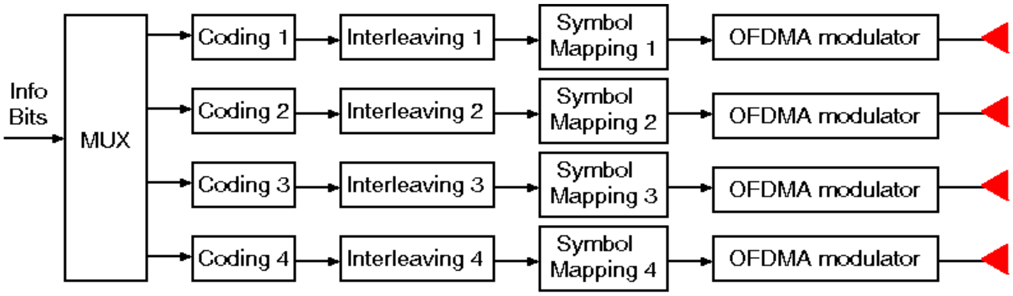

Fig. 8. Per Antenna Rate Control (PARC).

To solve this problem, PGRC is used where a feedback is required per group of antennas. This reduces the feedback information while maintaining almost the same performance of PARC [24].

4.4.3 Per User Unitary Rate Control (PU^2RC)

PU^2RC is a multi-user closed-loop MIMO system. Each user feeds back the CQI to the base station. The base station uses the CQI to determine the modulation and coding schemes per user. In addition, base station can apply unitary pre-coding and adaptively select the number of transmit antennas (rank adaption) [25].

V. Smart Antennas

5.1 Introduction

Smart antenna was born in the early 1990s when well developed adaptive antenna arrays originate from Radar system. Later, Smart antenna technique is applied in wireless communications system. Recently, Smart antenna technique has been proposed as a promising solution to the future generations of wireless communication systems, such as the Fourth-Generation mobile communication systems, broadband wireless access networks, where a wide variety of services through reliable high-data rate wireless channels are expected. Smart antenna technique can significantly increase the data rate and improve the quality of wireless transmission, which is limited by interference, local scattering and multipath propagation [26], [27]. Smart antennas offer the following main applications in high data-rate wireless communication systems [28], [29]:

- Spatial Diversity
- Co-channel interference reduction
- Angle reuse or space division multiple access (SDMA)
- Spatial multiplexing

Smart antenna system can be categorized into three main groups: Phased antenna array system, switched beam systems, and adaptive antenna array system. To match the characteristics in each radio frequency chain of the transmitter and receiver, on-line calibration is required in smart antenna systems. On-line calibration technique can compensate the errors such as the distortions of radio frequency components due to small environment changes, the nonlinear characteristics of mixer, amplifier and attenuator, I/Q imbalance errors, etc.

5.2 Phased Antenna Array System

Phased antenna array is a group of antennas in which the relative phases of the respective signals feeding the antennas are varied in such a way that the effective radiation pattern of the array is reinforced in a desired direction and suppressed in undesired directions.

Phased antenna array system is usually utilized in *radio frequency* (RF) or *intermediate frequency* (IF) with the system central frequency larger than 10 GHz, such as satellite communication system [30]. There are two main different types of phased arrays, also called beamformers. There are time domain beamformers and frequency domain beamformers.

5.3 Switched Beam System

The switched beam method is considered as an extension of the current sectorization scheme. In the switched beam approach, the sector coverage is achieved by multiple predetermined fixed beam patterns with the greater gain placed in the centre of a beam [30]. When a mobile user is in the vicinity of a beam, then the signals at the output ports will be given as in (5). This enables the switched beam system to select the signal from the output port corresponding to that beam. As the mobile moves to the coverage of another beam during the call, the system monitors the signal strength and switches to other output ports as required. A basic switched beam antenna architecture is shown in Fig. 9. And Fig. 10 illustrates the produced antenna pattern with 4 antennas.

$$y_i(t) = s_i(t)G_i(t)\sum_{l=1}^{L} I_l(t)G_{li}(\theta_l) \qquad (5)$$

where $y_i(t)$ is the total signal appearing at port i, $s_i(t)$ is the signal source, $I_l(t)$ is the interfering signal source located at arbitrary angles θ_l, G_i is the transfer function between signal source along the main beams and their corresponding output ports, G_{li} is the transfer function between interference signal l and port i.

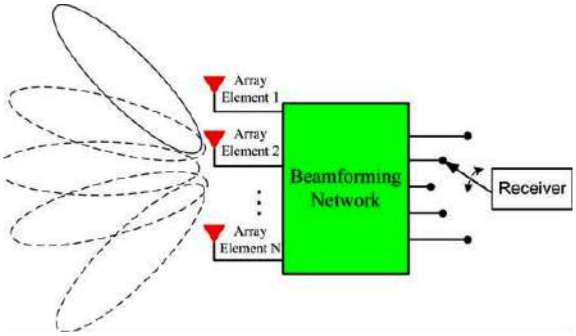

Fig.9. Functional block diagram of switched beam system.

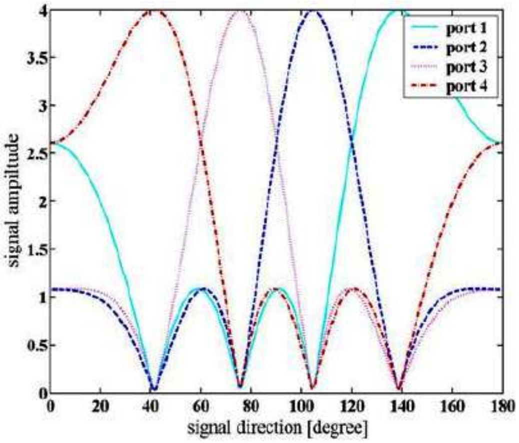

Fig.10. Produced antenna pattern of switched beam system with 4 antennas.

Switched beam systems can offer several advantages, including

- Low complexity and cost. Since switched beam system only requires a beamforming network, RF switches, and simple control logic, they are relatively easy and cheap to implement.
- Moderate interaction with base station receivers. In practice, switched beam system can simply replace conventional sector antennas without requiring significant modifications to the radio base station antenna interface or the baseband algorithms implemented at the receiver.
- Coverage extension. The antenna array aperture gain will boost the link

budget, which could be translated to a coverage extension.

5.4 Adaptive Antenna Array System (AAA)

Adaptive antennas date back to 1959. The original work was attributed to L. C. Van Atta’s work, Electromagnetic Reflection. Since then, adaptive beamforming techniques have been employed to remove unwanted noise and jamming from the output, mainly in military applications. With the thriving commercial wireless communication industry and the advancing microprocessor technologies, the adaptive beamforming techniques have found their applications in commercial wireless communications. With powerful *digital signal processing* (DSP) hardware at the base-band, algorithms could control antenna beam patterns adaptively to the real signal environment, forming beams towards the desired signals while forming nulls to co-channel interferers. Thus, the system performance is optimized in terms of link quality and system capacity [31]. Adaptive antenna array can be utilized in the transmitter side, which is known as *transmit beamforming* (TxBF) or in the receiver side, which is called *receive beamforming* (RxBF).

5.4.1 Transmit Beamforming (TxBF)

The implementation of adaptive antenna array technique in a handset is difficult with today`s hardware due to its limitations in size, cost, and energy storage capability, while it is feasible to adopt antenna arrays at base stations. Transmit beamforming provides a powerful method for increasing downlink capacity [32]-[35]. The idea of TxBF is similar to the pre-coded MIMO technique but with different strategies to calculate the transmit weight vector. TxBF adjusts the antenna main lobe towards to the desired user and reduce the interference to other users. A simple illustration of TxBF is shown in Fig. 11.

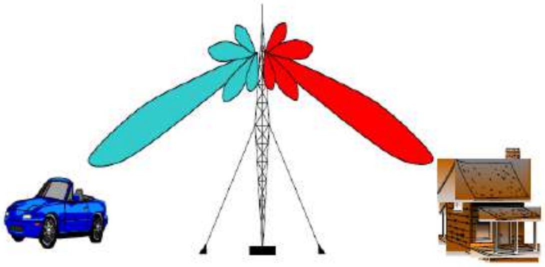

Fig.11. An illustration of TxBF.

Eigenvector TxBF Algorithm

Eignenvector TxBF algorithm is widely used for TxBF. The eigenvector of the spatial covariance channel matrix is calculated as

$$\mathbf{R}_{ss} = \boldsymbol{\lambda}\mathbf{H} \qquad (6)$$

where $\mathbf{R}_{ss}$ is the autocovariance matrix of the desired user`s signal, and $\mathbf{H}$ is the spatial covariance channel matrix. The eigenvector $\boldsymbol{\lambda}_{max}$ which corresponds to the largest eigenvalue will be selected as the weight vector [36]. One example of beam pattern for 4 uniform linear array elements is shown in Fig 12.

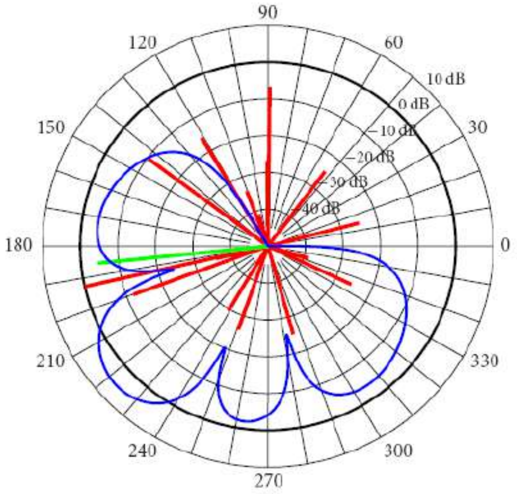

Fig. 12. Example beam pattern of 4 antenna elements in a sectorized system for a single sector (main beam direction is 240°).

Transmit Adaptive Array (TxAA) Algorithm

Transmit adaptive array (TXAA) is a technique in which the user periodically sends quantized estimates of the optimal transmit weights to the BS via a feedback channel. The transmitter weights are optimized to deliver maximum power to the user. The optimal transmit weights are given by

$$\mathbf{w} = \mathbf{H}^{H} / \mathbf{H}\mathbf{H}^{H} \qquad (7)$$

where **w** is the transmit weight vector and **H** is the channel matrix.

The weights are normalized so that the total transmitted power is not altered. In the case of multipath channels emanating from each antenna, the optimal weights will be given by the principal eigenvector of the channel correlation matrix $\mathbf{H}^{H}\mathbf{H}$.

5.4.2 Receive Beamforming (RxBF)

Beamforming also can be applied in the uplink to improve the link quality and suppress the co-channel interference, which is known as *receive beamforming* (RxBF). Through RxBF, smart antenna system can receive predominantly from a desired direction (direction of the desired source) compared to some undesired directions (direction of interfering sources). This implies that the digital processing has the ability to shape the radiation pattern to adaptively steer beams in the direction of the desired signals and put nulls in the direction of the interfering signals. This enable low co-channel interference and large antenna gain to the desired signal.

Based on the reference signals adopted in the beamforming algorithms, RxBF can be classified into *spatial reference beamforming* (SRB), *temporal reference beamforming* (TRB), and *signal structure reference beamforming* (SSRB).

Spatial Reference Beamforming (SRB)

Spatial reference beamforming method is sometimes referred as *direction of arrival* (DoA) method. SRB estimates the direction of arrival of the signal based on the spatial reference signal, using any of the techniques like *multiple signal classification* or *estimation of signal*

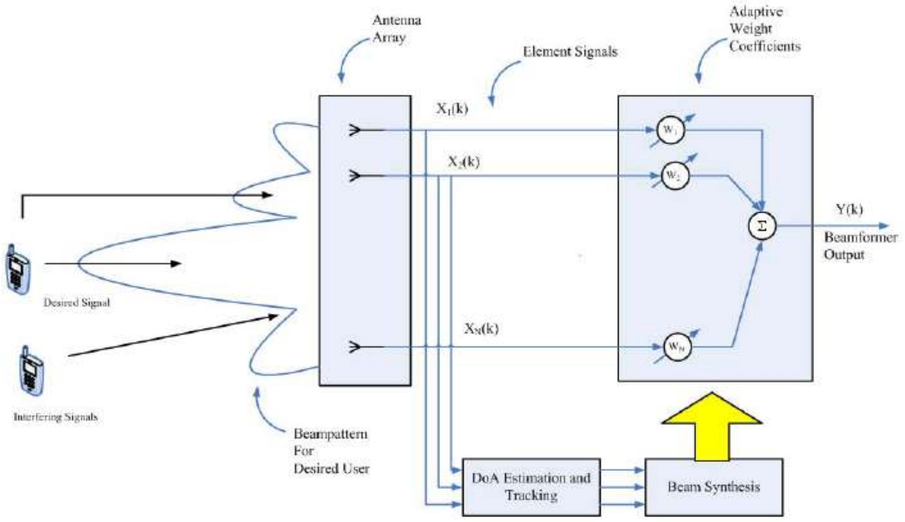
Antenna Array
Element Signals
Adaptive Weight Coefficients
$X_1(k)$
$X_2(k)$
$X_N(k)$
W_1
W_2
W_N
Σ
Y(k)
Beamformer Output
Desired Signal
Interfering Signals
Beampattern For Desired User
DoA Estimation and Tracking
Beam Synthesis

Fig. 13. A general structure of SRB.

parameters via rotational invariance techniques algorithms or their derivatives. They involve finding a spatial spectrum of the antenna/sensor array, and calculating the DoA from the peaks of this spectrum [37]. A general architecture of SRB algorithm is shown in Fig. 13. The general steps of SRB method are shown as follows:

- DoA Estimation
    - Arbitrary Array: MUSIC, etc.
    - Linear Array: ESPRIIT, etc.
- Beam Synthesis
    - Gram-Schmidt, etc.
- Combining
    - EGC, MRC, Wiener Filter, etc.

Multiple signal classification (MUSIC) algorithm estimates the DoA of the desired signal by using an eigen-space method based on a spatial reference signal. MUSIC requires intensive calculation of eigenvalues and eigenvectors of an autocorrelation matrix of the input vectors from the receiving antenna array. A general step of MUSIC algorithm is shown below:

- Collect received samples and estimate the covariance matrix of the received samples.
- Perform eigen-decomposition of the covariance matrix.
- Calculate spatial spectrum.
- Estimate DoA by locating peaks in the spectrum.

Estimation of signal parameters via rotational invariance techniques (ESPRIT) is also well known for the SRB method. In addition, ESPRIT has many important advantages over MUSIC algorithm [38]:

- No knowledge of the array geometry and element characteristics are required.
- Much less complex on computation.
- No calibration of the array is required.
- The algorithm simultaneously estimates the number of sources and DoA`s

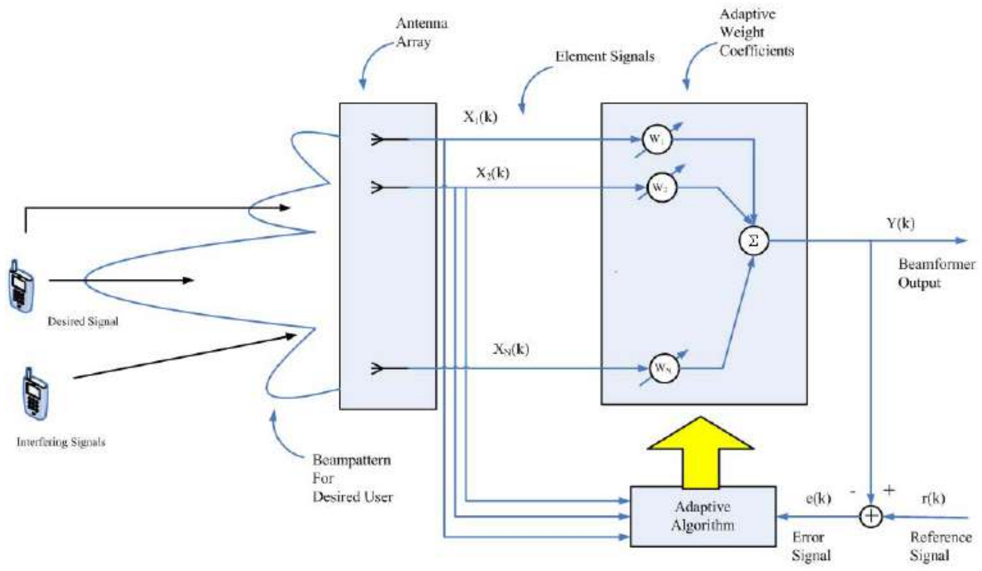
Antenna
Array
Element Signals
Adaptive
Weight
Coefficients
$X_1(k)$
$X_2(k)$
$X_N(k)$
W_1
W_2
W_N
Σ
$Y(k)$
Beamformer
Output
Desired Signal
Interfering Signals
Beampattern
For
Desired User
Adaptive
Algorithm
$e(k)$
-
+
$r(k)$
Error
Signal
Reference
Signal

Fig. 14. A general structure of TRB.

Temporal Reference Beamforming (TRB)

Temporal Reference Beamforming shown in Fig. 14, is a method used to create the radiation patter of the antenna array by adding constructively the phases of the signals in the DoA of the desired user, and nulling the pattern of the interfering users based on the temporal reference signal [39]. Based on the temporal reference signal and some predefined adaptive weight calculation criterion, some adaptive algorithms such as LMS (Least Mean Square), RLS (Recursive Least Squares), and DSMI (Direct Sample Matrix Inverse) algorithms, are used to adjust the weight vector of the antenna array to improve the link quality. The general characteristics of TRB are as follows:

- Good performance in multipath channel environment
- Computationally inexpensive
- Requiring Training sequence
- Difficult to apply TxBF because of the absence of DoA information

Criterion of Adaptive Weight Calculation

In the *minimum mean-square error* (MMSE) criterion, the weights are chosen to minimize the *mean-square error* (MSE) between the beamformer output and the temporal reference signal. While in the *maximum signal-to-interference ratio* (MSIR) criterion, the weights are chosen to directly maximize the signal-to-interference ratio (SIR). And The *minimum variance* (MV) criterion chooses the weights that minimize the variance of the output power. All the above three criterions has the same form of

$$\mathbf{w}_{opt} = \beta \mathbf{R}_i^{-1} \mathbf{v} \qquad (8)$$

where $\mathbf{R}_i^{-1}$ is the inverse of the covariance matrix of the interference signals received in the antenna array and **V** is the antenna array propagation vector [40].

Let us assume that *d(t)* is the transmitted temporal reference signal and $\mathbf{R}_u$ is the covariance matrix of interference signals at the output of the beamformer. The calculation of β for MMSE, MSIR

Table 4. β Calculation.

Criterion	MMSE	MSIR	MV
$\boldsymbol{\beta}$	$\frac{E\{d^2(t)\}}{1+E\{d^2(t)\}\mathbf{v}^H\mathbf{R}_i^{-1}\mathbf{v}}$	$\frac{E\{d^2(t)\}}{SIR}\mathbf{v}^H\mathbf{w}_{opt}$	$\frac{g}{\mathbf{v}^H\mathbf{R}_u^{-1}\mathbf{v}}$

and MV criterions are summarized in Table 4.

Adaptive Beamforming Algorithm

The *least mean square* (LMS) algorithm uses the temporal reference signal to update the weights at each iteration. In the LMS algorithm, we are searching for the optimal weight that would make the array output either equal or as close as possible to the reference signal, which is the weight that minimizes the MSE. Since the MSE has a quadratic form, moving the weights in the negative direction of the gradient of the MSE should lead us to the minimum of the error surface. The weight update equation is shown in (9) [30].

$$\mathbf{w}(t+1)=\mathbf{w}(t)-\mu\mathbf{x}(t+1)\varepsilon^* \qquad (9)$$

where μ is a constant, called the step size, which determines how close the weights approach the optimum value after each iteration and it controls the convergence speed of the algorithm. And ε is the error signal between the temporal reference signal and the received signal at the beamformer output. **x**(t+1) is the received signal vector at the antenna array at time *t+1*.

The main drawback of the LMS algorithm is that it is sensitive to the scaling of its input. This makes it very hard (if not impossible) to choose a step size μ that guarantees stability of the algorithm. The *normalized least mean square* (NLMS) algorithm is a variant of the LMS algorithm that solves this problem by normalizing with the power of the input. The weights updating function of NLMS algorithm is shown as

$$\mathbf{w}(t+1)=\mathbf{w}(t)+\frac{\hat{\mu}}{a+\left\|x(t)\right\|^2}\mathbf{x}(t)\varepsilon^* \quad (10)$$

Recursive least square (RLS) algorithm is derived to overcome the drawback of slow convergence speed in the LMS algorithm, when the eigenvalue spread of the correlation matrix R of received signal vector **x** is large. RLS algorithm replaces the step size μ with the inverse of R. The weights are then updated using (11).

$$\mathbf{w}(t+1)=\mathbf{w}(t)-\mathbf{R}^{-1}\mathbf{x}(t+1)\varepsilon^* \quad (11)$$

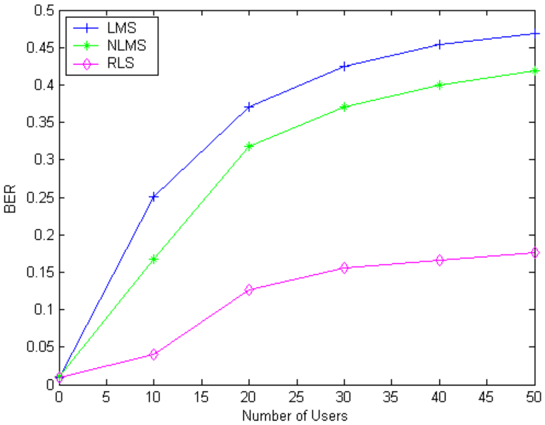

Fig. 15. Performance comparison among LMS, NLMS, and RLS algorithms.

Fig. 15 shows a simple performance comparison of the above three algorithms in OFDMA system under the Rayleigh fading channel with 8 antenna elements [41]. From this figure, we see that RLS algorithm performs best due to its faster convergence speed than LMS and NLMS algorithms.

Signal Structure Reference Beamforming (SSRB)

SSRB method is based on inherent structure of the transmit signal of the implicit kind reference signal. Algorithms such as blind beamforming, least squares, and constant modulus algorithms, are based on the SSRB method. SSRB method is robust against different propagation conditions and does not require the array manifold knowledge. But the convergence problem becomes the main drawback of the SSRB method.

VI. Conclusions

In this paper we introduced the multi antenna technologies which can be considered as one of the most vivid area of research. Multiple antenna technologies were categorized into two main groups where in the first group we introduced some techniques related to spatial diversity and spatial multiplexing by outlining the gain achieved by these schemes. Furthermore, we introduced the smart antenna techniques and the up-to-date research progress in this field.

The advantages of multiple antenna systems make of them a very strong candidate to increase link reliability, increase channel capacity and reduce interference in both uplink and downlink.

References

[1] A. Goldsmith, *Wireless Communications*, Cambridge Univ. Press, 2005.

[2] M. Alamouti, “A simple transmit diversity technique for wireless communications”, *IEEE J. Sel. Area Commun.,* 16, pp. 1451-1458, 1998.

[3] G. Tsoulos, *MIMO System Technology for Wireless Communications*, Taylor and Francis, 2006.

[4] V. Tarokh et al., “Space-time block codes from orthogonal designs”, *IEEE Trans. Inf. Theory,* vol. 45, no. 5, pp. 1456-1467, 1999.

[5] S. Haykin and M. Mohr, *Modern Wireless Communications,* Pearson Prentice Hall, 2005.

[6] W. Su and X-G. Xi, “Two generalized complex orthogonal space-time block codes of rates 7/11 and 3/5 for 5 and 6 transmit antennas”, *IEEE Trans. Inf. Theory,* vol. 49, no. 1, pp. 313-316, Jan 2003.

[7] K. Suto, and T. Ohtsuki, “Space-time-frequency block codes over frequency selective fading channels,” *IEICE Trans. Commun.,* vol. E87-B, no. 7, pp. 1939-45, July 2004.

[8] NTT DoCoMo, “Multi-degree cyclic delay diversity with frequency-domain channel dependent scheduling”, *3GPP TSG-RAN WG1 meeting #44bis*, R1-060991, Mar. 2003.

[9] Samsung, “Adaptive cyclic delay diversity,” *3GPP TSG-RAN1 43#*, R1-051354, 7th – 11st, Nov. 2005, Seoul, Korea.

[10] G. Stüber, *Principles of Mobile Comm-unication,* Kluwer, 2001.

[11] M. K. Simon and M-S. Alouini, *Digital Communication over Fading Channels*, Wiley, 2005.

[12] E. Biglieri et al., “Diversity, interference cancellation and spatial

Table of Contents

INTRODUCTION: THE VISION FOR 5G, CREATING AN ULTRA-CONNECTED SOCIETY 4

1 5G OBJECTIVES 7

1.1 TECHNICAL SPECIFICATIONS OF 5G.. 8
1.2 5G USE CASES 10
1.3 JUMP IN PERFORMANCE COMPARED TO 4G NETWORKS 12
1.4 NETWORK SLICING AND SOFTWARE-DEFINED NETWORKS.. 14
1.5 TECHNOLOGICAL BUILDING BLOCKS TO ACHIEVE THE OBJECTIVES 15
1.5.1 AIR INTERFACE.. 15
1.5.2 NETWORK ARCHITECTURE.. 17
1.5.3 5G: A MULTI-TECHNOLOGY GENERATION COHABITATING WITH EXISTING NETWORKS................................ 19

2 5G DEVELOPMENT INITIATIVES 21

2.1 GOVERNMENT INITIATIVES 21
2.1.1 IN EUROPE.. 21
2.1.2 WORLDWIDE (EXAMPLES IN THE US, SOUTH KOREA, JAPAN AND CHINA) 23
2.2 A HOST OF PRIVATE INITIATIVES – A FEW EXAMPLES 24

3 THE CHALLENGES OF 5G 26

3.1 NEW BUSINESS MODELS FOCUSED ON VERTICAL MARKETS 26
3.1.1 THE AUTOMOTIVE SECTOR.. 26
3.1.2 INDUSTRY 4.0 27
3.2 SPECTRUM HARMONISATION 28
3.2.1 MILLIMETRE WAVE FREQUENCIES 28
3.2.2 FREQUENCY BANDS BELOW 6 GHZ... 29
3.3 INCREASINGLY SMALL CELLS 30
3.3.1 TAXATION.. 31
3.3.2 ACCESS TO ELEVATED AND “SEMI-ELEVATED” LOCATIONS 31
3.3.3 5G NETWORKS’ REGIONAL COVERAGE AND BACKHAUL 32
3.4 NET NEUTRALITY ISSUES 32

ANNEX 1 DEFINITION AND STANDARDISATION WORK.. 35

ANNEX 2 ENTITIES ARCEP MET WITH.. 40

ANNEX 3 FIGURES.. 41

Introduction: the vision for 5G, creating an ultra-connected society

5G wants to be the disruptive generation, the generation that no longer only caters to the needs of mobile operators and consumer communications, but which opens up new prospects and enables an extremely wide diversity of applications and use, unified within a single technology. 5G is setting itself up as an enabler of the digitisation of society and the economy.

The idea that is starting to take shape behind the notion of 5G is that it will not mean just an increase in transmission speeds, as has been the case with previous generations.

Consumer mobile communications, video downloads and the use of mobile apps account for the bulk of 4G networks’ radio resources usage today. With 5G, the goal will be to enable a much broader spectrum of uses and a much greater diversity of users.

5G is targeting a wide variety of sectors, which will not necessarily have anything other than this technology in common, but which are central pillars in a society: energy, healthcare, media, industry and transportation.

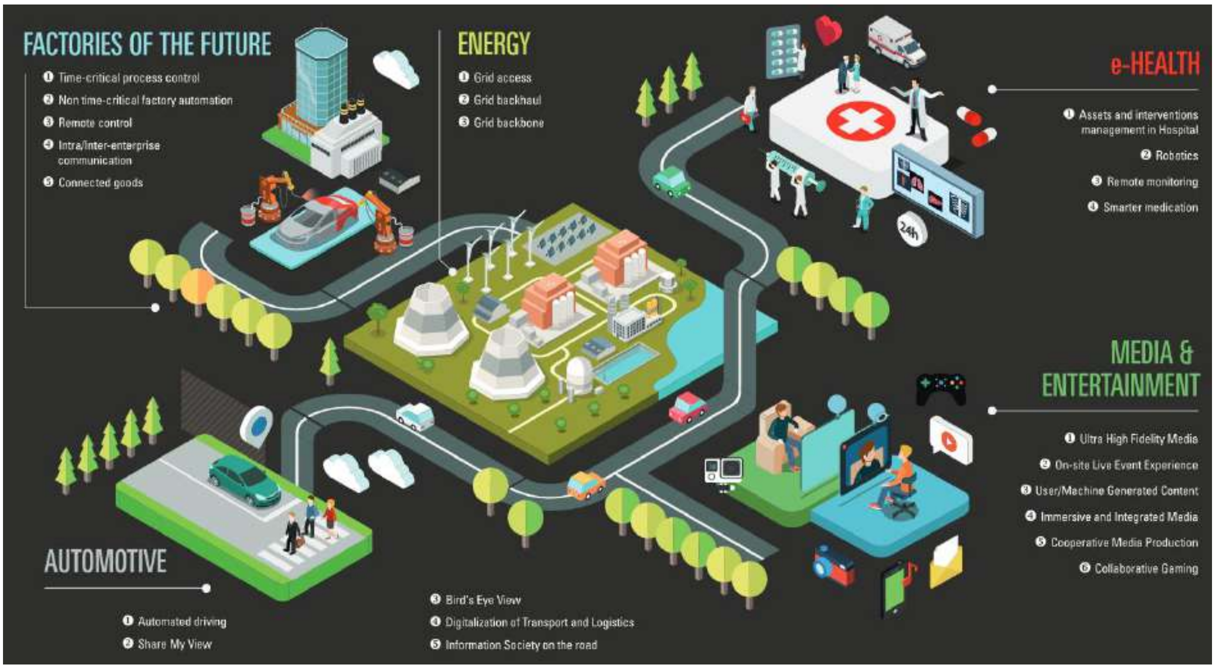
FACTORIES OF THE FUTURE
❶ Time-critical process control
❷ Non time-critical factory automation
❸ Remote control
❹ Intra/Inter-enterprise communication
❺ Connected goods
ENERGY
❶ Grid access
❷ Grid backhaul
❸ Grid backbone
e-HEALTH
❶ Assets and interventions management in Hospital
❷ Robotics
❸ Remote monitoring
❹ Smarter medication
24h
MEDIA & ENTERTAINMENT
❶ Ultra High Fidelity Media
❷ On-site Live Event Experience
❸ User/Machine Generated Content
❹ Immersive and Integrated Media
❺ Cooperative Media Production
❻ Collaborative Gaming
AUTOMOTIVE
❶ Automated driving
❷ Share My View
❸ Bird's Eye View
❹ Digitalization of Transport and Logistics
❺ Information Society on the road

Figure 1. 5G driving industrial and societal changes[1]

The energy sector, for instance, has undergone a great many changes and developments over the past several years in terms of energy production, storage and transport. Rising fossil fuel costs, the introduction of renewable energy sources and markets being opened up to competition have all helped usher in new kinds of energy product and new stakeholders – both independent companies and ordinary citizens – to the energy marketplace. The objective with 5G will thus be to enable better management of these networks (smart grids, smart agriculture, factories of the future) and their interconnections, to achieve more efficient and more agile distribution.

---

[1] 5G Empowering vertical industries. White Paper, 2016, https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

Healthcare, transportation and media are sectors that are bound up with our fellow citizens' daily lives: any improvements in these sectors thus have a clearly visible impact. The introduction of robots to perform specialised surgical operations, of video on demand or the advent of new connected features in cars have already improved our daily lives. 5G is promising to go several steps further in all of these areas: it would enable remote medical diagnoses and operations in real time, it would democratise streaming of 360° 3D video, it would provide users with a vast selection of video content with a picture quality better than ultra high-definition (4K, 8K…). The automotive universe could rely on these new networks to help cars make decisions without human involvement, and also communicate with one another (this is already possible, for instance, with the first experimental fleet of fully autonomous taxis being tested on the streets of Singapore by the firms nuTonomy and Grab[2]), with reaction times that are compatible with the demands of high-speed travel.

In terms of factories of the future, the improvements brought by 5G are primarily targeting the introduction of new generations of connected robots, the interconnection of production sites and the much heavier use of smart sensors to improve industrial processes. Generally speaking, the aim is to achieve ubiquitous communication between machines, a process that is already well underway.

One thing that it is crucial to understand, through the few examples listed above whose demands might seem mutually incompatible, is that depending on the sector or the application, the network properties and the functionalities required will not be the same. Service providers – whether they are today's mobile operators or other market players – will need to be capable of adapting their network to demand, occasionally in real time. 5G will thus be not so much a universal technology as a polymorphous, or multi-faceted technology, capable of adapting to any use, up to and including the most demanding ones.

---

[2] http://nutonomy.com/

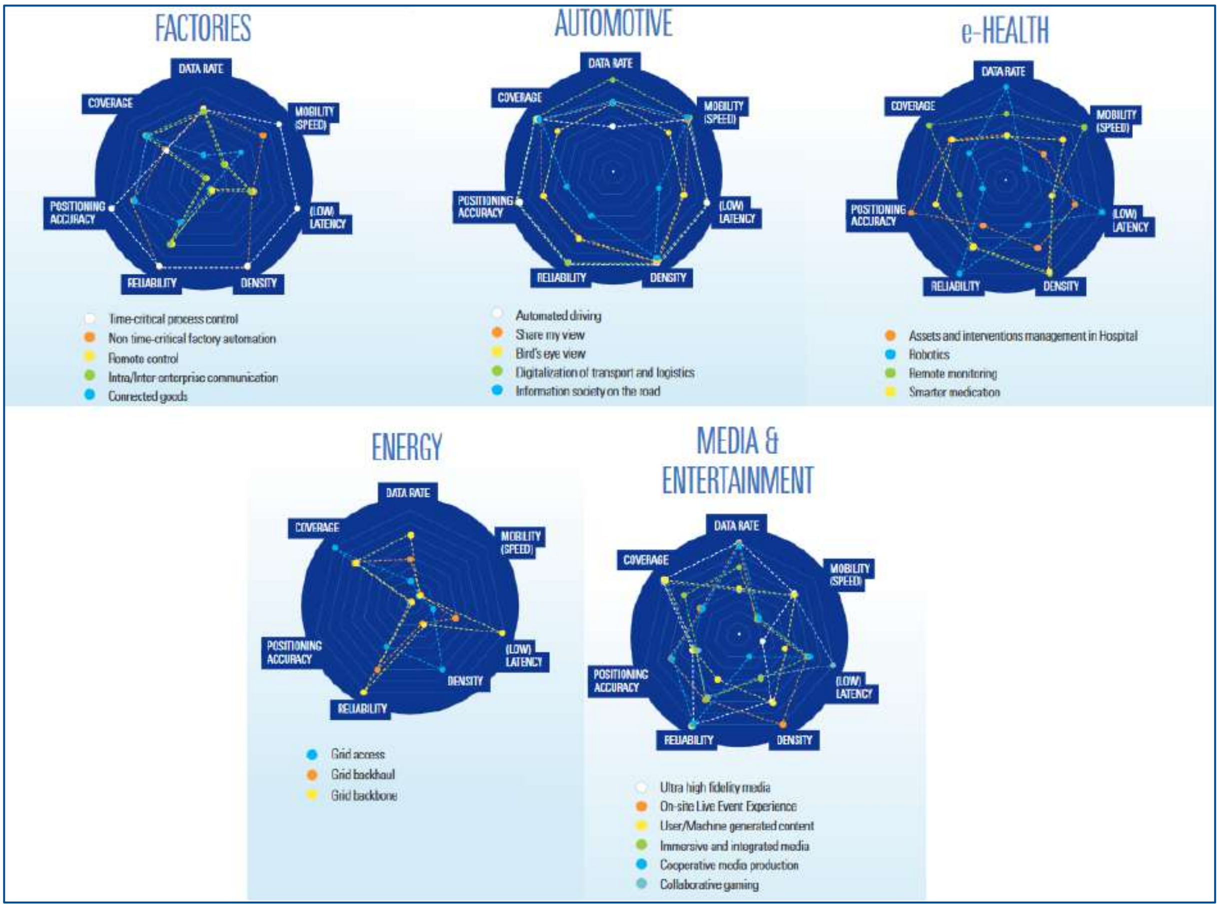

Figure 2. Performances required by vertical sectors[3]

---

[3] 5G Empowering vertical industries. White Paper, 2016, https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

1 5G objectives

Ever since the first real mobile telephone call, 44 years ago[4], mobile technologies have been evolving steadily and their performances have improved exponentially: as mentioned earlier, calling services and later texting and finally the mobile internet and the use of multi-service mobile applications have characterised the evolution of mobile networks and the transition from one generation to the next. The birth of LTE technology and the fourth generation (4G), coupled with the widespread use of smartphones and tablets, have driven a massive increase in the amount of mobile data traffic being relayed over the networks.

The use of a mobile handset and its applications is now an integral part of our fellow citizens' daily habits. Portable connected devices are increasingly powerful: in many instances they have replaced users' landline telephones, cameras, computers and even televisions. Today, 5 million videos are watched on YouTube and 67,000 images uploaded to Instagram every minute (see Figure 3).

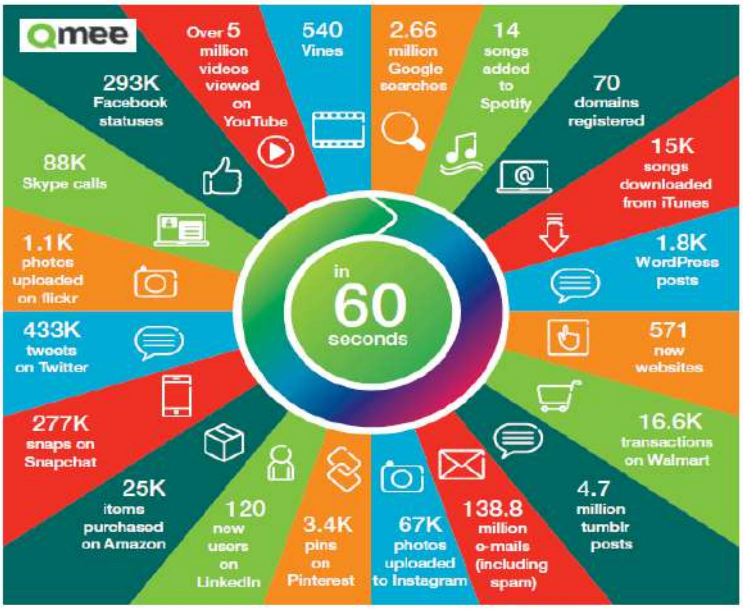
Qmee
Over 5 million videos viewed on YouTube
540 Vines
2.66 million Google searches
14 songs added to Spotify
70 domains registered
293K Facebook statuses
88K Skype calls
15K songs downloaded from iTunes
1.1K photos uploaded on flickr
1.8K WordPress posts
in 60 seconds
433K tweets on Twitter
571 new websites
277K snaps on Snapchat
16.6K transactions on Walmart
25K items purchased on Amazon
120 new users on LinkedIn
3.4K pins on Pinterest
67K photos uploaded to Instagram
138.8 million e-mails (including spam)
4.7 million tumblr posts

Figure 3. What we do over the network today in one minute[5]

The latest *Mobility Report*[6] from Ericsson indicates that traffic on mobile networks almost doubled in a single year, and that over the next five years it will have increased to 10 times what it is today. New solutions must therefore be found to meet this demand, and to optimise how resources are used.

The increase in the number of applications available, their diversification and the improved quality of mobile networks have all contributed to driving up demand, the emergence of new uses (connected objects, drones, etc.) and new users.

---

[4] On 3 April 1973, Motorola's Martin Cooper made a call from the corner of 56th street in New York City with the first mobile phone.

[5] https://www.ericsson.com/res/docs/2014/5g-what-is-it-for.pdf

[6] https://www.ericsson.com/assets/local/mobility-report/documents/2016/ericsson-mobility-report-november-2016.pdf

5G is at the crossroads of these news uses; it aims to better and simultaneously satisfy this tremendous variety of needs and these new demands, through a unified technology that takes this diversity into account at the design stage.

The advent of 5G could have a significant impact not only in the technical realm, but also on different countries' economic and social development. As indicated in the introduction, 5G targets a very large number of sectors and, through society's digitisation, is expected to contribute to countries' economic growth.

To give an example, according to a report produced by InterDigital Europe, Real Wireless, Tech4i2 and Connect (Trinity College Dublin)[7], this new technology will require a great deal of money and a great deal of work, but will generate €113.1 billion in profits per annum for the European economy by 2025.

1.1 Technical specifications of 5G

The specifications for a new generation of mobile telephony are set primarily by two bodies: ITU (International Telecommunication Union) and 3GPP (3rd Generation Partnership Project).

ITU is the United Nations agency devoted to information and communications technologies. It carries out research and studies through its Working Party 5D, the sub-group responsible for the overall radio system aspects of international mobile telecommunications (IMT). In 2013, this group began working on defining the characteristics of the new IMT standard, IMT-2020 (5G) (cf. 1.3), as it had done back in the early 2000s to define 4G (IMT-Advanced). Further details on the roadmap established for this work can be found in Annex 1, the objective set by ITU-R being to complete its analyses by 2020.

Parallel to the work being done by ITU are the studies being conducted by 3GPP. The 3rd Generation Partnership Project was created in 1998, and its members include seven standardisation bodies, several hundred industry players, associations and public organisations. It is responsible for developing and maintaining technical specifications for mobile telephony standards[8]. When a new standard is being defined by ITU, 3GPP works on the technical solutions that make it possible to achieve the objectives set by ITU.

Although 5G is one of the most widely debated topics inside international and European bodies today, no standard has yet been defined by 3GPP. The draft of Release 15, the first 5G standard from 3GPP, is still in the works: definition of the new architecture began in December 2016 and work on the New Radio (NR) interface is set to begin in March 2017. A first Release of the standard should be validated in September 2018 to meet the more urgent demands; a second Release (3GPP Release 16) will then be published in March 2020. All of these elements are addressed in more detail in Annex 1.

It should also be said that the transition from one generation to the next takes place gradually. LTE will continue to evolve alongside NR, and these two standards will likely be very complementary initially. In particular, for pioneer 5G rollouts LTE shall probably remain the master of the network and control the NR antennae. Moreover, some of the objectives set for 5G could be achieved thanks to functionalities or technologies introduced in 3GPP Releases 13, 14 and 15 that will not be proper to 5G, but rather evolutions of 4G (which some refer to as 4.9G or LTE Advanced Pro).

---

[7] http://ir.interdigital.com/file/Index?KeyFile=36051369

[8] GSM for 2G, UMTS for 3G and LTE for 4G

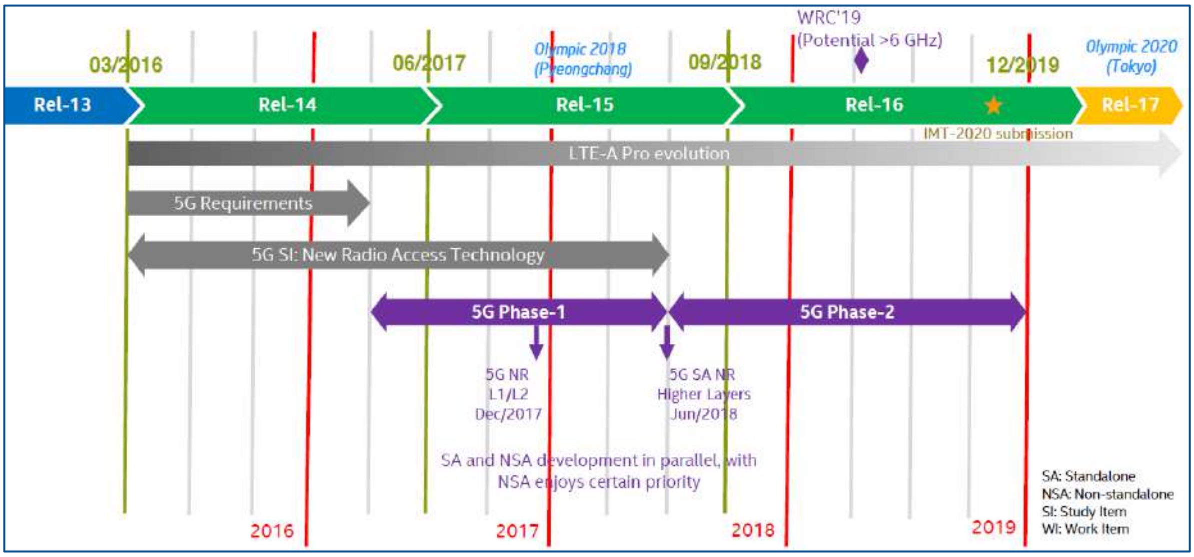
WRC'19
(Potential >6 GHz)
03/2016
06/2017
Olympic 2018
(Pyeongchang)
09/2018
12/2019
Olympic 2020
(Tokyo)
Rel-13
Rel-14
Rel-15
Rel-16
Rel-17
IMT-2020 submission
LTE-A Pro evolution
5G Requirements
5G SI: New Radio Access Technology
5G Phase-1
5G Phase-2
5G NR
L1/L2
Dec/2017
5G SA NR
Higher Layers
Jun/2018
SA and NSA development in parallel, with
NSA enjoys certain priority
SA: Standalone
NSA: Non-standalone
SI: Study Item
WI: Work Item
2016
2017
2018
2019

Figure 4. 3GPP timeline for 5G[9]

This same momentum occurred when making the transition from 3G (Release 4) to 4G (Release 10, first Release of IMT-Advanced): the first LTE specifications (Release*s* 8 & 9) do not achieve all of the ITU targets for 4G, and reprised many of the characteristics of the most advanced form of 3G at the time.

The synergies between the successive generations, the absence in most instances of a great initial jump in performances between generations, and the race between competitors to outperform one another and be the most innovative, can often lead operators and equipment suppliers to give a “commercial” name to each generation of mobile telephony. For instance, in the United States AT&T had called its HSPA network 4G, while in France it is a 3G+ network.

So in all likelihood the first 5G networks deployed on a large scale will be 4.9G systems, using carrier aggregation, massive MIMO (Multiple-Input Multiple-Output) and Network Function Virtualization, or NFV (cf. trials conducted in in France, detailed in Section 1.5.3). These technologies, which will be explored in greater detail further on, represent more of an evolution of the fourth generation than an actual transition to 5G, which will occur when disruptive technologies such as NR carriers in millimetre bands, non orthogonal multiple access (NOMA*)* and mobile edge computing (MEC) can be put into place.

In addition to this commercial race, equipment suppliers – often backed by their respective government at home, notably in the United States, Japan and South Korea – want to get a head start in defining and testing 5G technologies, in the hope of establishing themselves as the technological leaders in international standardisation bodies. Here, the precocity of the first rollouts announced in South Korea (5G deployment for the Winter Olympics in 2018 in Pyeongchang) and Japan (5G deployment for the Summer Olympics in Tokyo) would justify the supposition that only a small portion of 5G technologies will be used, and that these rollouts will be based more on 4.9G or pre-5G technologies.

---

[9] http://gsacom.com/paper/intel-5g-technology-mm-wave-frequencies/

1.2 5G use cases

Three main use cases (defined by ITU, under IMT-2020), with their respective – and potentially mutually incompatible – demands are in the process of taking shape, and will make it possible to meet the sector-specific needs referred to in the introduction.

1. **mMTC – Massive Machine Type Communications**: A very large number of connected devices with disparate quality of service requirements. The objective of this category is to provide a response to the exponential increase in the density of connected objects;
2. **eMBB – Enhanced Mobile Broadband**: ultra high-speed connection indoors and outdoors, with uniform quality of service, even on the edges of a cell;
3. **uRLLC – Ultra-reliable and Low Latency Communications**: this use case has stringent requirements for capabilities such as latency and packet-loss, to ensure increased reactivity.

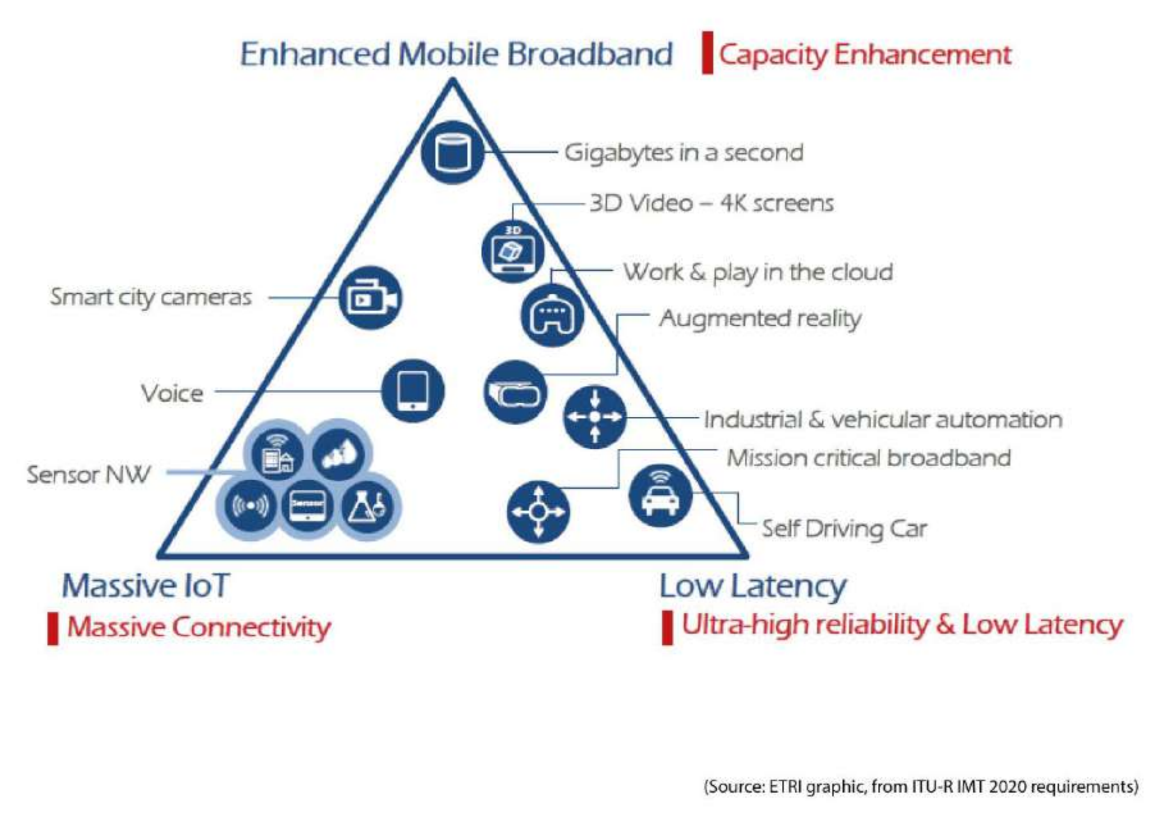
Enhanced Mobile Broadband
Capacity Enhancement
Gigabytes in a second
3D Video – 4K screens
Work & play in the cloud
Augmented reality
Smart city cameras
Voice
Sensor NW
Industrial & vehicular automation
Mission critical broadband
Self Driving Car
Massive IoT
Massive Connectivity
Low Latency
Ultra-high reliability & Low Latency

(Source: ETRI graphic, from ITU-R IMT 2020 requirements)

Figure 5. 5G use cases[10]

The first group (mMTC) primarily encompasses all Internet of Things related uses. These services require broad coverage, lower energy consumption and relatively slow transmission speeds. What 5G will deliver compared to existing technologies is the ability to connect objects that are spread out in a very dense fashion across a given area.

Enhanced mobile broadband (eMBB) concerns all of the applications and services that require increasingly fast connections, for instance to watch ultra high-definition (8K) videos or to stream virtual or augmented reality applications wirelessly.

---

[10] http://5g.ieee.org/standards

Ultra-reliable and Low Latency Communications (uRLLC) include all of the applications that require an extreme reactivity and very strong message transmission guarantees. The sectors where these requirements are particularly prevalent are transportation (reaction time when an accident occurs, for instance) and medicine (telesurgery), and when digitising manufacturing processes in general.

To implement these three use case, ITU established eight key performance indicators[11] (KPI) to specify, quantify and measure the characteristics of IMT-2020 (5G) systems:

- Peak data rate (Gbit/s);
- User experienced data rate (Mbit/s);
- Spectrum efficiency (bit/Hz);
- Device mobility (km/h);
- Latency (ms);
- Connection density (number of connected/accessible objects per km²);
- Network's energy efficiency;
- Area traffic capacity (Mbit/s/m²).

---

[11] Added to these eight classes, five new indicators were defined and are currently being examined: reliability, mobility interruption time, bandwidth, maximum spectrum efficiency, 5th percentile spectrum efficiency.

1.3 Jump in performance compared to 4G networks

As mentioned in Paragraph 1.1, the deployment of 5G is likely to take place in two stages:

1. The first 5G rollouts will deliver better performances, but as a continuation of what is being provided at the time by 4G systems which will continue to evolve as well (4.5G, 4.9G);
2. Performances will continue to improve with the gradual introduction of disruptive technologies, such as the use of millimetre wave frequencies.

This gradual rollout is very similar to the way in which 4G was introduced, as the performances obtained with pioneer deployments were relatively similar to those supplied by existing 3G networks.

ITU defines the 4th generation of technologies, called IMT-Advanced, by the values for the eight KPI listed above, as indicated in the following chart. The 5th generation, called IMT-2020, is represented as well. Here, it should be noted that the official ITU definition of 4G in fact corresponds to what commercial rollouts supplied later on, under the name 4G+ or LTE-Advanced.

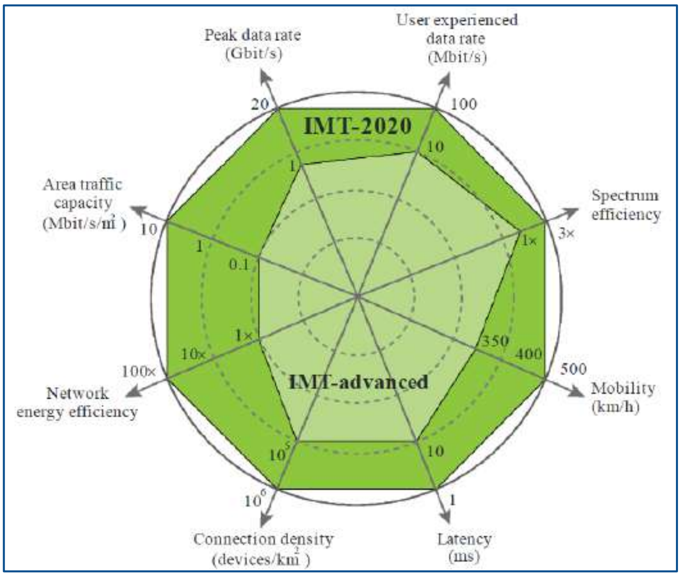

Figure 6. Comparison between 4G and 5G with respect to the eight key performance indicators[12]

According to these objectives, 5G must be able to provide a user experienced data rate and a peak data rate that is respectively 10 and 20 times higher than what is currently available. The maximum connection density will be multiplied by 10 and latency divided by at least 10 (the target point-to-point latency is 1 ms, compared to 30 to 40 ms today).

---

[12] Recommendation ITU-R M.2083-0 (09/2015), https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf

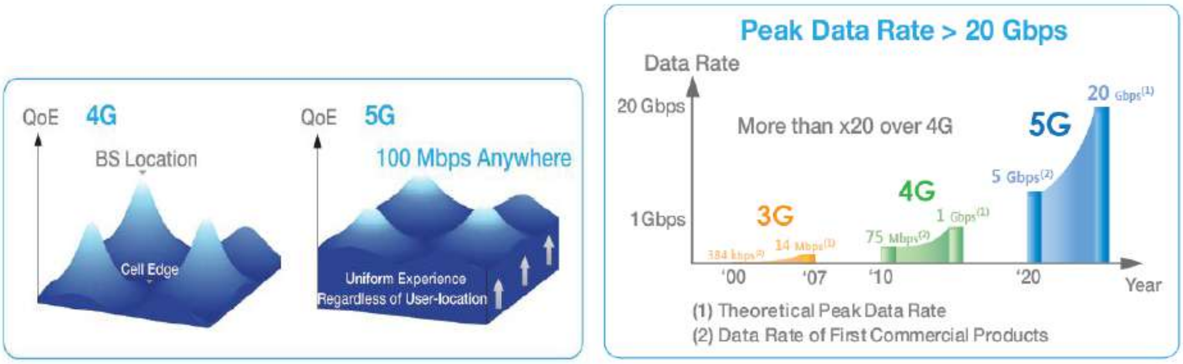

Figure 7. Representation of the "average user experience data rate" KPI for 4G and 5G and "Peak data rate" KPI for 3G, 4G and 5G[13]

As the successor to 4G, the objective for 5G will also be to provide:

- An extremely reliable network, with more consistent performances regardless of the user's position with respect to the base station;
- A stable connection, even when travelling (at speeds of up to 500 km/h);
- Greater network energy efficiency (with batteries that consume up to 100 times less power).

The following table summarises the target performances for 5G and those currently available with 4G:

Performances/Generation	4G	5G
1. Peak data rate (Gbit/s)	1	20
2. User experience data rate (Mbit/s)	10	100
3. Spectrum efficiency	1x	3x
4. Speed (km/h)	350	500
5. Latency (ms)	10	1
6. Connection density (number of objects/km²)	10^5	10^6
7. Network energy efficiency	1x	100x
8. Area traffic capacity (Mbit/s/m²)	0.1	10

Table 1. Comparisons between 4G and 5G performances

---

[13] http://www.samsung.com/global/business-images/insights/2015/Samsung-5G-Vision-0.pdf

1.4 Network slicing and software-defined networks

It is vital to understand that the set of indicators listed in Section 1.3 determine a set of peak performances for 5G. It will not, however, be possible to achieve all these peak values simultaneously: not every requirement or use case is compatible, so a trade-off will need to be made when defining categories of use that each have their own performance envelope, notably for the use cases described in Section 1.2 (mMTC, eMBB and uRLLC). This is the principle of network slicing: each slice has its own set of KPI, which is a trade-off tied to the target use. On a 5G system, the network properties will need to adapt to the chosen environment.

The following diagram positions the three main use cases listed in Section 1.2. with respect to the eight key performance indicators listed above.

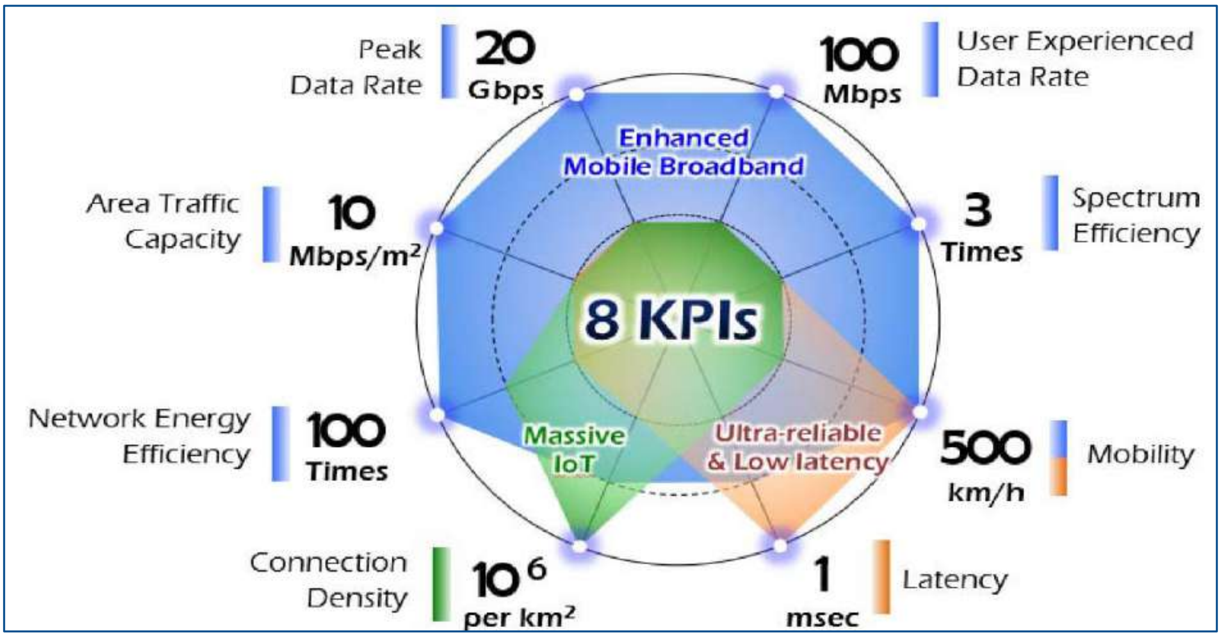
Peak Data Rate
20 Gbps
100 Mbps
User Experienced Data Rate
Enhanced Mobile Broadband
Area Traffic Capacity
10 Mbps/m²
3 Times
Spectrum Efficiency
8 KPIs
Network Energy Efficiency
100 Times
Massive IoT
Ultra-reliable & Low latency
500 km/h
Mobility
Connection Density
10⁶ per km²
1 msec
Latency

Figure 8. Key performance indicators for the three 5G use cases[14]

So for applications that require enhanced mobile broadband (eMBB), such as 4K, 8K or 3D video or virtual reality, a certain number of performance indicators, such as spectrum efficiency, peak data rate and area traffic capacity, can be reached only at the expense of others, such as latency or connection density.

On the flipside, when a massive simultaneous connection of connected objects (mMTC) needs to be managed, the network will concentrate its resources and use the technologies required to achieve this task, but will not be able, for instance, to use spectrum as efficiently or to guarantee low latency.

Lastly, when ultra-reliable and low latency communications (uRLLC) are required, the number of simultaneous connections, data rates and spectrum efficiency may be reduced.

This flexibility, or ability to adapt, that network slicing brings can only be achieved thanks to the softwarisation and virtualisation of a sizeable number of network components (cf. 1.5.2) – a process referred to as Software-Defined Networking (SDN) and Network Function Virtualisation (NFV). Behind these acronyms is a common idea, namely to use as many generic and reconfigurable components as possible, rather than bespoke ones that are permanently dedicated to very specific tasks. This evolution towards software-based systems has been in the works for several years, but is now becoming possible thanks to improved performances from all of these reconfigurable components, including those that are the closest to the elementary tasks of wireless communications (detection, baseband coding, bitstream management, frequency handover, signal processing, etc.).

---

[14] https://5g-ppp.eu/wp-content/uploads/2016/11/06_10-Nov_Session-3_Lee-JunHwan.pdf

1.5 Technological building blocks to achieve the objectives

1.5.1 Air interface

Several, sometimes competing radio access technologies are currently being examined. Some have already been pre-implemented by equipment manufactures and can be used in trials, notably massive MIMO and NFV. Others, such as NOMA modulation and mobile edge computing (MEC), will no doubt take longer before they are ready to be used. In any event, a consensus – which could be painful for certain suppliers whose investments will be lost – will need to be found when defining 5G standards, to ensure the systems' interoperability.

The technologies currently being examined are the following:

- **Millimetre wave frequencies**: the use of millimetre wave frequencies constitutes one of the disruptive 5G technologies. The term refers to the frequencies above 6 GHz which have never been considered for mobile fronthaul network rollouts, for reasons of technological maturity and propagation quality. To meet the demand for ever increasing data rates and traffic volumes, new bands with very wide channels (over 100 MHz per user) will need to be employed: millimetre wave frequencies could provide this spectrum resource, and in certain cases their use would make it possible to achieve the data rates listed in Table 1. Comparisons between 4G and 5G). In exchange, to be able to use these frequencies all of the required, miniaturised low-cost technologies will need to be developed, and ensure a level of energy consumption that is compatible with portable devices (amplifiers, coders, signal processing, antennae, etc.). In particular, because of millimetre waves' poor propagation quality, each cell will have limited coverage and so require the use of beamforming (described below) to better focus the power transmitted by the antenna.
- **Massive MIMO** (Multiple Inputs – Multiple Outputs): this technology involves the use of a large number of smart micro-antennae, located on the same panel (between eight and 128 today, but the number will increase with the use of frequencies above 6 GHz). The appeal of using massive MIMO is twofold: first, the technology makes it possible to increase data rates, thanks to spatiotemporal multiplexing; second, it makes it possible to focus energy on a device to improve its link budget, thanks to beamforming.

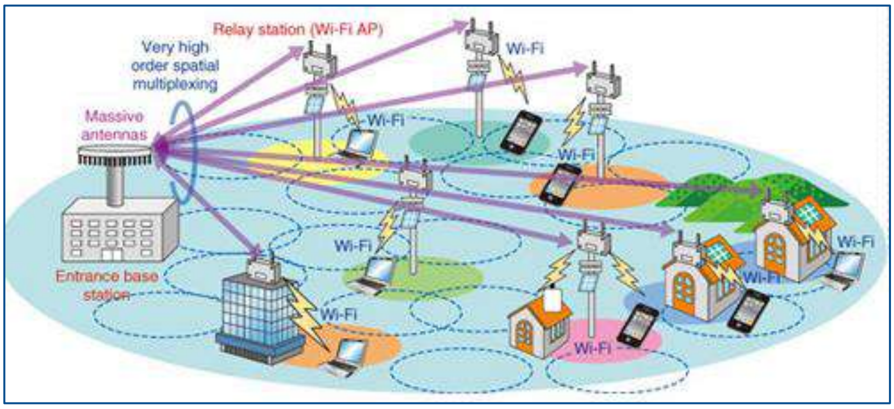
Relay station (Wi-Fi AP)
Very high order spatial multiplexing
Massive antennas
Entrance base station
Wi-Fi

Figure 9. Example of the use of a beamforming antenna to connect Wi-Fi access points[15]

- **Full Duplex**: in classic systems, transmission and reception takes place either on different frequency bands, i.e. frequency division duplexing (FDD) used on all mobile network bands in France, or at different times: time division duplexing (TDD), the top contender for LTE wireless local loop networks in France. The full duplex is intended to enable the simultaneous transmission and reception of data, on the same frequencies, at the same time and in the same location.[16]

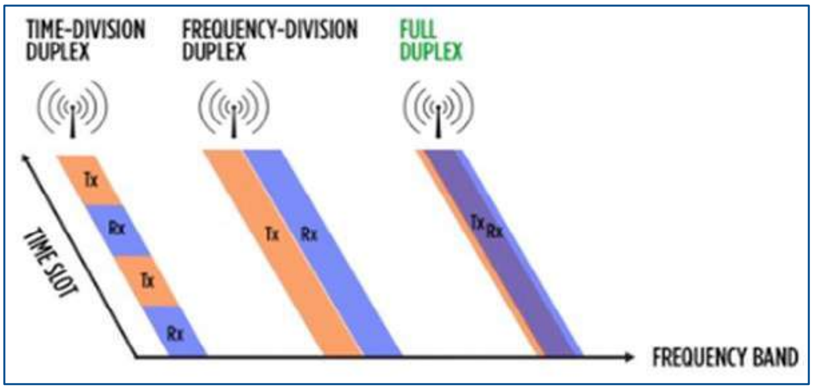
TIME-DIVISION DUPLEX
FREQUENCY-DIVISION DUPLEX
FULL DUPLEX
Tx
Rx
Tx
Rx
Tx
Rx
Tx Rx
TIME SLOT
FREQUENCY BAND

Figure 10. Illustration of full-duplex, compared to FDD and TDD[17]

- **NOMA Multiplexing** (Non Orthogonal Multiple Access): LTE uses what is referred to as orthogonal multiplexing, with each device using a portion of the resource blocks in a unique fashion at any given time. For 5G to provide improved spectrum efficiency compared to 4G, the plan is to use non-orthogonal multiplexing methods, whereby several users can use the same frequencies at the same time. A distinction can be made between several users by assigning different codes to each user – referred to as SCMA or sparse code multiple access – a combination of 3G's code division multiple access (CDMA) and 4G's orthogonal frequency division multiple access (OFDMA) or by playing on the difference in users' signal to noise ratios (power domain NOMA[18], illustrated below). These are the two methods chosen for NR.

---

[15] https://www.slideshare.net/100001290086432/massive-mimo

[16] The basic operating principle is the following: an antenna sends a signal at the same time as it is receiving signals coming from devices in the cell. However, the signal received by the antenna is a combination of the signal sent by itself and signals coming from the devices (everyone is "talking" at once). As it knows which signal it itself has sent, the antenna can subtract it during digital processing from the ones it received. So only the signals received from the devices remain.

[17] http://compeng.columbia.edu/biggest-component-full-duplex-wi-fi-radio-antenna

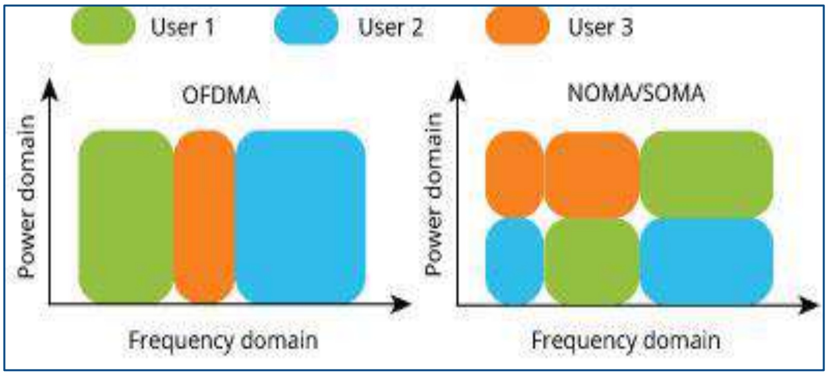

Figure 11. Illustration de multiplexing *power domain* NOMA[19]

- **QAM256**: as with many modern communication systems, 4G uses quadrature amplitude modulation (QAM). In 4G this modulation can achieve QAM64, which means that six bits of information are being transmitted ($2^6 = 64$) at any given time. On wireless systems, the main limitation on QAM order is the signal to noise ratio: when a large amount of information is sent all at once, its transmission will be very sensitive to disruptions (a bit like trying to talk in a noisy environment: it is easy to understand "yes" or "no" but harder to understand more complex sentences). Thanks to an improved link budget, via antenna or signal processing technologies, 5G modulation could reach QAM256, i.e. eight bits of information being transmitted at any given time, which translates into a 33% increase in maximum capacity under ideal conditions. This improved modulation will also be deployed on advanced 4G systems.
- **IoT waveforms**: new waveforms are being explored for the future deployment of 5G IoT in mobile bands. But although mass market IoT is one of the main challenges put forth for 5G, no concrete results have yet been made public. Operators[20] are starting to deploy new standards (EC-GSM or Extended Coverage GSM, LTE-(e)MTC or enhancements for Machine-Type Communications, NB-IoT or NarrowBand IoT) which were defined by 3GPP in Release 13 but, as they are based on 2G and 4G, they do not deliver the performance levels, notably in terms of autonomy, coverage and density, that are compatible with the targets set for future 5G networks.

1.5.2 Network architecture

As with air interfaces, new network architectures are also being explored:

- **Software-defined networking (SDN) and network functions virtualisation (NFV)**: these two functionalities extend beyond the scope of 5G networks per se. They are part of an overall process of network upgrades taking place today and already available with 4G technology (4.9G). The nevertheless remain a key enabler of 5G:

---

[18] The basic operating principle is the following: either User 1 (U1) with a good signal to noise ratio and User 2 (U2) with a less good signal to noise ratio. The antenna sends a high power Signal 2 to U2, and superimposes a weaker Signal 1 aimed at U1. U2 will only see S2 as S1 is drowned out by the noise. U1 will decrypt S2 then delete it from the signal received, to create a higher quality S1, thanks to a better signal to noise ratio.

[19] https://www.anritsu.com/en-AU/test-measurement/technologies/5g-everything-connected/5g-everything-connected-detail

[20] http://www.vodafone.com/content/index/what/technology-blog/nbiot-commercial-launch-spain.html

  - SDN (Software-Defined Network) is designed to disassociate the network's control plane from its data plane, these two planes traditionally being linked and distributed in a set fashion in the network (see diagram below). Controlling the network, a task previously assigned to specialised and unscalable hardware components, is centralised in the form of software on more powerful servers and, in theory, free of equipment manufacturer specifications. This enables the deployment of high value-added services (load balancing, smart routing, dynamic configuration, etc.) in disparate environments.

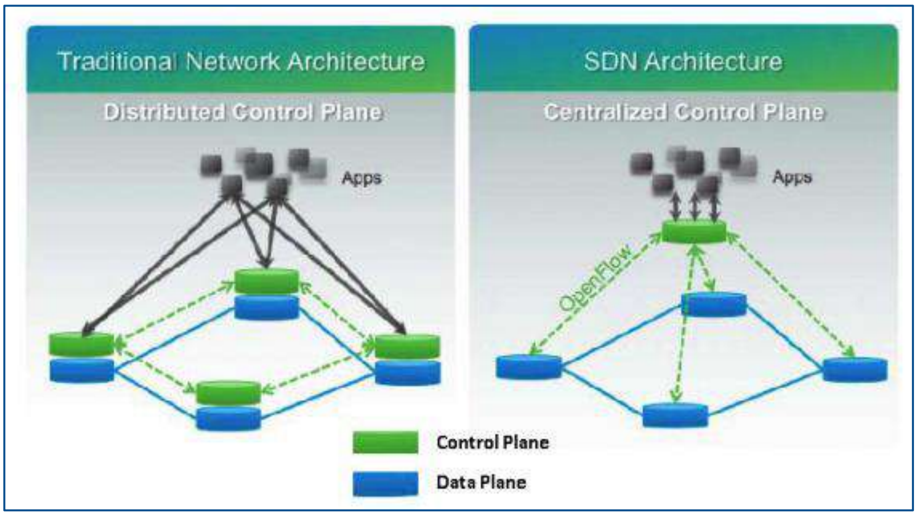
Traditional Network Architecture
Distributed Control Plane
Apps
SDN Architecture
Centralized Control Plane
Apps
OpenFlow
🟩 Control Plane
🟦 Data Plane

Figure 12. Centralisation of the control plane in a software-defined network[21]

  - NFV, which builds upon SDN, is used to virtualise, in other words to replace hardware designed specially to perform certain key network functions (firewall, network core, interfaces between different systems…) with software on a server, to accelerate rollouts and enable rapid changes and upgrades.

- **CloudRAN**: this functionality, also know as centralized-RAN, requires a very different network architecture to what we find today. It is an evolution of SDN: the base stations' signal processing units, currently installed at the base station level, are moved to the cloud and centralised. They communicate with the network radio heads, located closer to the antenna, over an optical fibre network (Radio over fibre technology). This centralisation makes it possible to obtain a complete overview of all of the stations deployed and to coordinate signal processing and manage interference between cells and devices

---

[21] http://www-igm.univ-mlv.fr/~dr/XPOSE2014/software-defined_networking/sdn.html

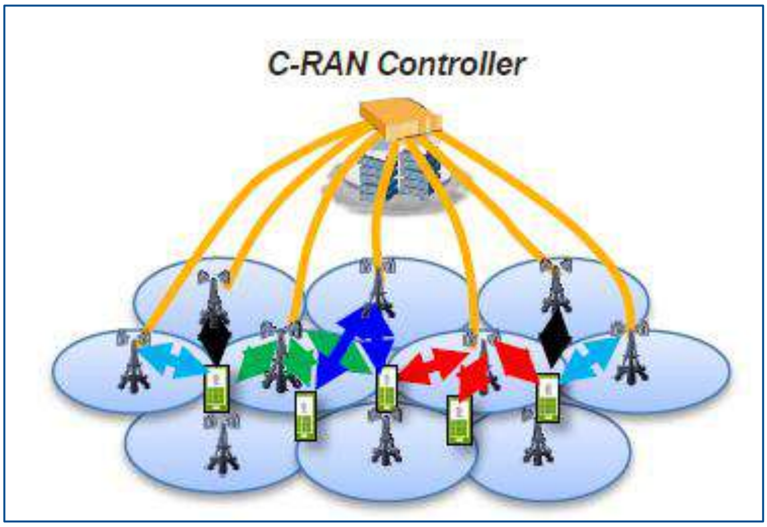

Figure 13. Illustration of CloudRAN[22] network architecture

- **Optimised content delivery, using a mobile content delivery network (mobile CDN)**: corresponds to a set of servers working together in a transparent fashion to optimise the delivery of content to end users over wireless (mobile or Wi-Fi) networks, with high availability and performance. With 5G, the objective for these CDN is to cache content close to users, notably thanks to predictive algorithms, to offload traffic from the networks and decrease latency.
- **MEC (mobile edge computing)**: MEC is an evolution of mobile CDN whose purpose, in addition to bringing data closer to devices, is to provide devices with an accessible computing power with very low latency, within a very specific area for demanding applications. This technology makes it possible to locate a portion of the network's intelligence (managing local critical applications and performance analysis) at the base station level. The "antennae" will be capable of analysing a certain number of data, and so to make decisions very quickly.
- **Device-to-device**: D2D is a direct form of communication between two nearby devices, which does not require the data to travel over the cellular network. Device-to-device communication is not new, as technologies such as Bluetooth and Wi-Fi direct already enable it. But a new mesh networking technology will be introduced with 4.9G and later 5G network rollouts: LTE-direct. Far more energy-efficient than its predecessors, this technology will have a range of up to 500 metres and geolocation capabilities to enable communications. This technology will be very useful for low latency V2V (vehicle-to-vehicle) or V2X (vehicle-to-everything) communications and for certain public-security related uses.

1.5.3 5G: a multi-technology generation cohabitating with existing networks

As stated earlier, 5G is not meant to replace 4G overnight. In practice, the devices will undoubtedly be multi-modal: still connected to the 4G network, which will provide extended coverage for pioneer rollouts, then transitioning to 5G networks when they become available.

While 4G and 5G frequencies will probably be initially segregated, in all likelihood devices will rapidly become capable of aggregating 4G and 5G carriers and, further down the road, 4G carriers will be encapsulated in 5G ones.

---

[22] https://5g-ppp.eu/wp-content/uploads/2016/11/04_10-Nov_Session-3_Takaharu-Nakamura.pdf

4G is still being deployed, and its technological evolution and certain building blocks will be used by both the first 5G networks and advanced 4G networks. To wit, the latest trials conducted in France are allowing 4G networks to perform better thanks to the use of pre-5G technologies:

- Bouygues Telecom, in partnership with Huawei, managed to achieve a peak data rate of 1 Gbps thanks to the simultaneous use of four-carrier aggregation (800 MHz, 1800 MHz, 2100 MHz, 2600 MHz) and more powerful modulation (256 QAM)[23];
- In the coming weeks, Orange will be launching a massive MIMO (16x16) trial with Nokia[24].

In addition to the strong integration between 4G and 5G, the new generation will no doubt also continue convergence efforts between frequency bands governed by exclusive licences – i.e. bands that are allocated exclusively to an operator, such as mobile operators – and unlicensed frequency bands, governed by a system of general authorisation (e.g. Wi-Fi bands), which already began in 4G with LTE-LAA (Long Term Evolution – License Assisted Access) and LTE-LWA (Long Term Evolution – Wi-Fi Link Aggregation).

LTE-LAA is characterised by the aggregation of one or several LTE carriers, used in licensed bands, with other LTE carriers employing unlicensed 5 GHz Wi-Fi bands. To guarantee cohabitation with very widespread Wi-Fi networks whose deployment pattern is unpredictable, it uses LBT (Listen Before Talk) technology to listen to the radio channel before transmitting, in order to determine whether or not a frequency is available.

LTE-LWA consists of an aggregation of LTE carriers in licensed bands with Wi-Fi traffic. To achieve this, the LTE cell and Wi-Fi access point need to be connected: the Wi-Fi traffic is sent to the cell which sends back the entire aggregated link on the 4G core network. This technology is particularly well suited to indoor environments with small cells, a system which, as detailed in paragraph 3.3 below, will lend itself well to 5G deployments.

---

[23] https://www.corporate.bouyguestelecom.fr/wp-content/uploads/2016/05/Communiqu%C3%A9-Bouygues-Telecom-et-Huawei-Technologies-r%C3%A9alisent-un-test-de-d%C3%A9bit-4G-%C3%A0-plus-de-1-Gbps-pour-la-premi%C3%A8re-fois-en-Europe-de-lOuest.pdf

[24] https://networks.orange.fr/actualites/actualites-des-networks/orange-et-nokia-partners-pour-developper-les-futurs-services-5g

2 5G development initiatives

With a view to keeping the many promises being made for 5G, a number of initiatives are currently underway around the globe to promote its development. Below, we detail the foremost initiatives put forth by the players with which Arcep met as part of its investigation.

2.1 Government initiatives

The prospect of significant socio-economic repercussions generated by 5G (possible revenue of close to $225 billion a year[25] by 2025) combined with many countries' desire to establish themselves as technological leaders and make their companies more competitive, have propelled a multitude of government initiatives around the globe, aimed at encouraging the mobile ecosystem to begin the work and make the investments required to drive the rapid construction of the first 5G networks.

A selection of the largest initiative is detailed below.

2.1.1 In Europe

5G-PPP

The 5G Public Private Partnership (5G-PPP) is dedicated to 5G research and development, created on the initiative of the European Commission in 2013, with a budget de €700 million in public funding. The main objectives set by 5G-PPP are:

- Create stronger ties between the economic players and academic bodies devoted to the telecommunications sector over R&D projects, along the entire value chain;
- Reduce technological dependence on the United States and Asia while sustaining a strong global market;
- Regain technological leadership, notably in disruptive technologies, by promoting standards in international bodies;
- Allow innovative business models to emerge;
- Facilitate large-scale experimentation.

The results of this work will help clarify the 5G action plan (see below), and fuel the standardisation work that is currently underway.

To achieve its ambitions, 5G-PPP has initiated three stages of work, financed by the European Union, whose roadmap is in sync with the main international initiatives (3GPP and ITU, cf. Figure 14):

- The first stage which is currently underway, will last until mid-2017;
- A second stage focused on systems optimisations, from the end of 2017 to mid-2019;
- And finally a full-size trial stage from 2019 to 2020.

Deploying 5G by 2020 will require Europe to develop leading edge technologies, globally approved standards and especially to achieve consensus over the most suitable frequency bands. This funding and these projects – involving a great many researchers from more than a hundred companies and the finest R&D centres in Europe – are thus vital.

---

[25] https://www.abiresearch.com/press/abi-research-projects-5g-worldwide-service-revenue

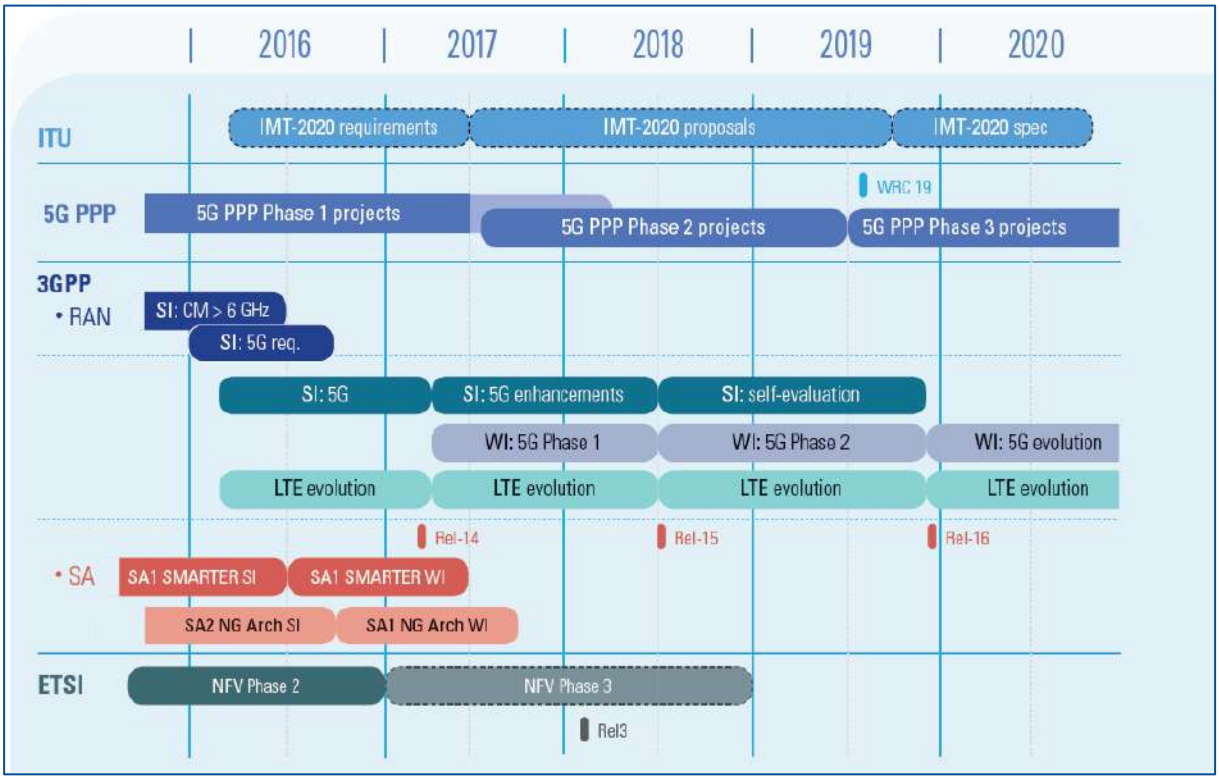

Figure 14. 5G PPP vs. 3GPP and ITU roadmaps[26]

5G action plan

As a complementary measure, on 14 September 2016 the European Commission launched its *5G for Europe Action Plan* to bolster investments in 5G infrastructure and service rollout efforts in the Digital Single Market between now and 2020. This action plan sets out a clear roadmap for public and private 5G investments inside the EU.

The Commission has proposed the following measures to achieve this plan:

- Align roadmaps and priorities for a coordinated 5G deployment across all EU Member states, targeting early network introduction by 2018, and moving towards commercial large scale introduction by the end of 2020 at the latest.
- Make provisional spectrum bands available for 5G ahead of the 2019 World Radio Communication Conference (WRC-19), to be complemented by additional bands as quickly as possible, and work towards a recommended approach for the authorisation of the specific 5G spectrum bands above 6GHz.
- Promote early deployment in major urban areas and along major transport paths.
- Promote pan-European multi-stakeholder trials as catalysts to turn technological innovation into full business solutions.
- Facilitate the implementation of an industry-led venture fund in support of 5G-based innovation.
- Unite leading actors in working towards the promotion of global standards.

---

[26] 5G Empowering vertical industries. White Paper, 2016, https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

The European Commission has given every EU country a certain number of ambitious, numerical targets. One core objective for 5G is thus to have at least one major city in every European country outfitted with this new generation mobile system by 2020, and coverage of every city, motorway and high-speed railway lines by 2025. This comes in response to announcements from South Korea and Japan which are both promising large-scale 5G demonstrations, respectively, at the Winter Olympics in Pyeongchang in 2018 and the Summer Olympics in Tokyo in 2020.

2.1.2 Worldwide (examples in the US, South Korea, Japan and China)

The race is already underway between countries to be the first to begin large-scale 5G trials, and later to introduce commercially available services, so much so that they are willing to employ pre-standard technical specifications.

Several countries are therefore working on standards and seeking to reach a consensus with one another to achieve international backing for the technical specifications that will satisfy their needs, and earn them the best possible return on the investments made thus far. This in turn is creating a certain turmoil, for instance, over the frequency bands that would be the best candidates for 5G.

The United States[27]

5G is seen as an unprecedented opportunity for economic growth, with a tremendous impact on education, job, transportation, etc. According to the US federal regulator, the FCC (Federal Communications Commission), the following three elements in particular need to be the prime focus of attention: spectrum, infrastructure and the backhaul network:

- **Spectrum**: in July 2016, the FCC voted to free up and to open up nearly 11 GHz of high frequency spectrum to be used for fixed and mobile broadband applications: 3.85 GHz of licensed spectrum in the 27.5 – 28.35 GHz and 37 - 40 GHz bands, as well 7 GHz of unlicensed spectrum from the 64 - 71 GHz band. The FCC's stated goal was to provide the assurance and clarity for investments in the telecoms sector. Verizon welcomed this decision, and plans on achieving the first rollouts in 2017, with trials already underway in several cities around the country (cf. 3.3).
- **Infrastructures**: The FCC believes that 5G must be underpinned by a robust infrastructure network capable of handling already heavy traffic that could potentially increase exponentially in the coming years. It will interact in a hybrid fashion between traditional towers with macro cells and small cell deployments, as well as distributed antenna systems. With an eye on the deployment of new towers, the FCC resolved to reduce, or eliminate in certain cases, the regulatory restrictions on installing facilities and antennae that will have very little impact "on historic properties" (tower re-use, small antennae, indoor installations, etc.)[28].
- **Backhaul network**: Because 5G systems will require very high-power backhaul networks, the FCC is in the process of reforming and updating the regulations governing data services markets, including wireless backhaul solutions. The FCC is seeking to protect this market while also creating competition, so that competitive and high quality backhaul network solutions can emerge.

---

[27] Forging Our 5G Future – Federal Communications Commission: https://www.fcc.gov/5G

[28] https://apps.fcc.gov/edocs_public/attachmatch/DA-16-900A1_Rcd.pdf

Japan

Japan wants to demonstrate its 5G leadership by deploying the first commercial 5G network that complies with international technical specifications, in time for the Summer Olympic Games in Tokyo in 2020.

According to the *Radio Policies Towards 2020*[29] report, published in June 2016 by Japan's Ministry of Internal affairs and Communications (MIC), the 3600 - 4200 MHz, 4400 - 4900 MHz and 27.5 – 29.5 GHz bands were selected to be the 5G candidate bands on a national scale. Although other bands are also being investigated, with a view to WRC19, 5G rollouts are already planned in the 3600 - 4100 MHz, 4405 - 4895 MHz and 27.5 - 28.28 GHz bands as early as 2017 in Tokyo, and will be built out in 2018 and 2019.

China

Like Japan, China too wants to prove its leadership with the first commercial rollouts in 2020. The first trials will be conducted that year in the 3400 – 3600 MHz band. The 3300 - 3400 MHz, 4400 - 4500 MHz and 4800 - 4990 MHz bands are also under consideration and currently being investigated. For higher speeds, the country plans on using spectrum around the 25 GHz and 40 GHz frequencies.

South Korea

South Korea, meanwhile, has set its sights on a pre-commercial 5G service that would be ready in time for the 2018 Winter Olympics in Pyeongchang. Several trials are already underway to prepare for full scale demonstrations in several South Korean cities, including Pyeongchang and Seoul. The country's three national mobile operators are hoping for spectrum in the 26.5 – 29.5 GHz bands.

South Korea's largest mobile operator, SK Telecom, has announced a plan for conducting interoperability tests with Qualcomm and Ericsson, as well as outdoor trials of the new air interface based on the 3GPP NR standards that are being developed today. These trials and experiments will be carried out in the second half of 2017.

If the goal of these trials is to help accelerate the specification of the new NR air interface, which is part of the work being done on 3GPP Release 15, it has not yet been specified which frequency bands will be used for these trials.

2.2 A host of private initiatives – a few examples

5G Open Trial Specification Alliance

In early 2016, South Korean operators KT and SK Telecom, Japanese carrier NTT DoCoMo and American carrier Verizon formed the 5G Open Trial Specification Alliance to carry out collaborative 5G trials.

Due to be conducted between 2016 and 2018, the aim for these trials is to provide a common platform for operators, for exchanging results and shared assessments of the different 5G network components and elements. One of the operators' objectives is to help accelerate the definition of standards and specs, and to enable economies of scale. The findings of these evaluations will provide input for 3GPP discussions, adding simulations to the experimental data produced by the partnership.

---

[29] http://www.gsma.com/spectrum/wp-content/uploads/2016/08/MIC_Spectrum-for-5G-MIC-Kuniko-OGAWA.pdf

The alliance's founders hope to attract a number of industrial partners to the platform, including other carriers, equipment suppliers, chipset makers, etc. Several of the founding operators' equipment suppliers are already involved in the scheduled trials.

The trials will cover multiple frequency bands, both above and below 6 GHz.

Verizon 5G Technical Forum

The Verizon 5G Technology Forum (V5GTF) was created in late 2015 by Verizon, in cooperation with its partners Cisco, Ericsson, Intel, LG, Nokia and Qualcomm. The goal of the collaboration is to provide a platform for testing radio interface specifications in the 28 and 39 GHz bands.

The first fruit of this collaboration, in July 2016 Verizon announced the completion of the first specifications of their 5G radio interface. These specs should enable the different parties involved to develop interoperable solutions and so to help drive forward the definition pre-standards.

Even though they run the risk of developing solutions that are not compatible with 3GPP or ITU standards, Verizon believes it has a solid enough grasp of the overall concepts being discussed in those bodies. The tests carried out in several US cities were thus able to validate substantially greater capacities than 4G.

Orange/Ericsson

Orange and Ericsson have been working together since October 2016 on developing 5G use cases and services, as well as demonstrations. The purpose is to develop technological building blocks, conduct trials and pilot projects for a range of use cases, including multi gigabits/s wireless internet access in suburban and rural environments, the Internet of Things and connected cars.

Their collaboration made it possible to achieve 15 Gb/s in a laboratory environment, notably thanks to the use of massive MIMO and beamtracking using centimetre waves. The collaboration will also focus on the transition from 4G to 5G solutions, notably in terms of reducing costs and improving energy efficiency, and on the use of SDN and NFV technologies.

Since January 2017, the partnership has included PSA (Peugeot) to carry out trials on connected cars.

5G-ConnectedMobility

5G-ConnectedMobility is a consortium formed by Ericsson, BMW, Deutsche Bahn, Germany's three mobile operators, Deutsche Telekom, Telefonica Deutschland and Vodafone, the 5G TU Dresden laboratory, the BAST research institute and BnetzA, with the goal of galvanising and accelerating 5G R&D in Germany.

5G-ConnectedMobility thus plans on supplying a digital motorway infrastructure and real application environment to be able to test V2V (vehicle-to-vehicle) technologies as well as solutions for digitising railway infrastructure.

To this end, 5G-ConnectedMobility operates with the help of an independent network infrastructure that is not connected to any commercial network. The Ericsson mobile network devoted to this project makes it possible to run live tests: Ericsson thus obtained permission from regulator BNetzA to employ spectrum in the 700 MHz band in the Nuremberg area.

3 The challenges of 5G

In this part we explore the different challenges that lie ahead for 5G, which emerged from the interviews that Arcep conducted over the past several months.

3.1 New business models focused on vertical markets

3G and especially 4G technologies were designed primarily for the superfast mobile internet. 5G continues on in this direction, but also wants to target what are known as vertical markets, which encompass several segments, including:

- Connected vehicles, not only to deliver entertainment and information to passengers, but also to guarantee safety via communications both between vehicles and between vehicles and infrastructure;
- Factories of the future;
- Smart cities with requirements in the areas of public transportation (similar to the needs of connected vehicles), the environment, managing buildings and energy consumption;
- Medicine, healthcare and robot-assisted telesurgery;
- Smart grid flow monitoring and management (electricity, gas, water, etc.).

This section will focus in particular on the connected car and factories of the future segments. Because of their current and future macro-economic context, along with the plurality and effervescence of the pioneer work being done in these areas, these two segments constitute the main avenues of 5G development in vertical markets. Smart cities and smart grids have already begun to soar through existing Internet of Things (IoT) technologies.

3.1.1 The automotive sector

The car is an extremely common form of transportation, and safe driving is a fundamental consideration: human error is the number one cause of all transport accidents. The transportation sector wants to use technological innovations to tackle this problem, and to continue to make transportation more efficient, more sustainable and safer.

There are three areas in which technological progress could help improve automotive transportation. 5G could have a role to play in all three, but particularly in the first two:

- Provide in car internet connectivity, to deliver entertainment to passengers;
- Provide access to driver assistance information, to reduce accidents and improve the fluidity of traffic;
- And, finally, the ability to make cars autonomous, thanks to artificial intelligence algorithms.

The first area is just an extension of the developments occurring today around 4G. The aim is to give passengers access to their messaging services, to the internet, to multimedia content, online gaming, etc. The increased connection speeds promised by 5G will improve the use of all of these services.

The second area aims to make cars more intelligent by using information that was not previously available to them. This in turn will help improve the safety and efficiency of the networks as well as help drivers make the right decisions, and adapt to driving conditions. Such connected vehicles could have access to information about dangers on the road (slow or stopped cars, traffic jam warnings, indications of where construction is taking place on the roads, weather conditions, emergency braking, emergency service vehicles approaching, etc.) or regarding signage (signalling/signage onboard vehicles, speed limits on vehicles, failure to heed signage/safety precautions at crossings, request for right-of-way at traffic lights for designated vehicles, green light optimal speed advisory, etc.). Other services, such as information on refuelling or recharging stations, vulnerable road user protection, managing street parking and traffic information, and smart guidance, could also prove useful. It is still not clear whether, to achieve this, vehicles will simply exchange information with each other or if connectivity with infrastructure will be required along the roads to optimise vehicles' behaviour. In both instances, 5G could have a role to play.

Of the many initiatives that are already underway we can begin with an example from France: the SCOOP@F[30] (cooperative intelligent transport systems) project coordinated by the Ministry of the Environment, Energy and the Sea, and which unites local authorities and R&D centres. Launched in February 2014, new partners subsequently joined the initiative, including Orange and Austrian, Spanish and Portuguese partners. As it is a European project, it receives 50% of its financing from the European Commission, and cross-tests are conducted with Austria, Spain and Portugal. SCOOP@F is a pilot rollout project for cooperative intelligent transport systems; it aims to deploy 3,000 vehicles on 2,000 km of roadway spread across five locations: Ile-de-France, the A4 motorway, the Isère, and ring roads in Bordeaux and Brittany. Its main objectives are to improve road safety and the safety of roadway workers, achieve more efficient traffic management, reduce pollution, streamline infrastructure management costs and participate in defining the car of the future.

In addition, in early 2017, mobile carrier Orange, equipment supplier Ericsson and car-maker PSA Group signed a partnership agreement[31], as part of the "*Towards 5G*" initiative, to conduct technical trials relating to 5G. The aim of this alliance is to test the different paths of technological evolution from 4G to 5G to meet the needs of connected cars, notably in terms of intelligent transport systems (ITS), for safer driving and new on-board services.

The third area concerns the emergence of autonomous vehicles. A number of projects are underway in this area. The first step is to outfit vehicles with algorithms that enable it to make decisions quickly based on their environment. This requires a large number of sensors to deliver a full "understanding" of what is happening around the vehicle. Without pre-judging the technologies that will ultimately be employed to achieve this, as with a human driver (see above), the car could take advantage of connections with other vehicles on the road and with an infrastructure, to have access to all of the aforementioned information.

3.1.2 Industry 4.0

Competitiveness does not depend solely on innovation and the evolution of products, but also on modernising businesses and their means of production. According to certain studies carried out in 2015[32], the digital transition in Europe will enable enterprises to increase their revenue by more than 110 billion euros a year over a five-year period.

A great many countries have set out a strategy for modernising their manufacturing infrastructure (*l'industrie du futur* in France[33], Industry 4.0 in Germany…) of which one aspect will involve digitising processes and trade. The European Commission itself introduced measures in 2016 for strengthening competitiveness in Europe, in which 5G could play a significant role. This included earmarking 500 million euros for the Horizon-2020[34] research programme.

---

[30] http://www.scoop.developpement-durable.gouv.fr/spip.php?page=sommaire

[31] http://www.orange.com/fr/Press-Room/communiques-2017/Ericsson-Orange-et-le-groupe-PSA-partners-pour-la-voiture-connectee-en-5G

[32] PwC, "Industry 4.0: Opportunities and challenges of the industrial internet" (2015), and Boston Consulting Group, "Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries" (2015)

[33] http://www.economie.gouv.fr/files/files/PDF/industrie-du-futur_dp.pdf

The advent of new technologies (4G, fibre and soon 5G) and new services (the Internet of Things, the cloud, big data) should facilitate businesses' digital transition. 5G in particular is expected to be a very versatile technology, capable of undergirding a very wide array of uses, and could go a long way in furthering companies' transition to digital technologies and solutions.

3.2 Spectrum harmonisation

5G is emerging as a technology that will use both low frequencies (f < 1 GHz), high frequencies (1 GHz < f < 6 GHz) and, for the first time ever in consumer networks, very high frequencies referred to as "millimetre wave" frequencies (f > 6 GHz).

This spectrum diversity is entirely bound up with the promises of 5G: extended coverage (low frequencies), ultra high speeds (very large channels in very high frequency bands), low power consumption. Furthermore, satellite services will also contribute to the development of this new technology, especially in areas that are difficult to cover and to provide backhaul solutions. The satellite industry is thus taking an interest in 5G, and wants to be involved in defining these new generation network.

3.2.1 Millimetre wave frequencies

The "millimetre" band, also referred to as millimetre wave spectrum, aka frequencies above 6 GHz, are essential to enabling 5G to mark a departure from 4G, for the reasons cited in Section 1.5.1.

At the latest World Radiocommunications Conference (WRC-15 in Geneva), a conference under the aegis of ITU whose objective is to change the way frequencies are allocated between users, discussions over the definition of future mobile bands made it possible to focus future 5G studies, for millimetre wave frequencies, on a certain number of bands situated between 24 GHz and 86 GHz (33.25 GHz identified in total): 24.25 - 27.5 GHz, 31.8 - 33.4 GHz, 37 – 43.5 GHz, 45.5 - 50.2 GHz, 50.4 - 52.6 GHz, 66 - 76 GHz, 81 - 86 GHz.

It is important to stress that, even if the above-listed bands have been identified as "5G bands", at this stage there is no way to know whether they can actually be used to deploy this new generation system: only the results of technical studies will make it possible to establish the constraints and rules of compliance, and to validate the feasibility of these hypotheses.

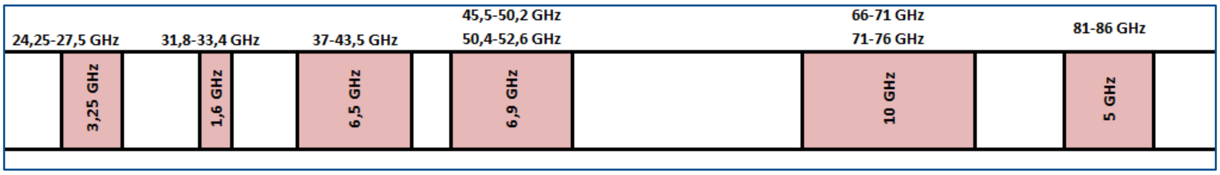

Figure 15. Millimetre wave frequencies identified at WRC-15

Contrary to the conclusions of the Conference, which reflect European recommendations, the United States and certain Asian countries (South Korea, Japan) have decided to perform 5G trials in the 28 GHz band, and equipment suppliers such as Qualcomm and Samsung, have begun manufacturing 28 GHz- band products.

---

[34] https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020

Meanwhile Europe, following the publication of an RSPG (Radio Spectrum Policy Group within which France is represented by ANFR) opinion[35], decided to focus its first studies on the 26 GHz band (pioneer band), then on the 32 GHz and 42 GHz bands. Later, studies will be carried out on introducing 5G in all of the other bands identified by WRC-15.

The rapid choice of the 26 GHz band as the pioneer band was made to enable economies of scale for equipment production, since it is very likely that dual-mode equipment, i.e. compatible with both the 26 GHz and 28 GHz band, will be available for pioneer rollouts[36].

In France, the 26 GHz band is already being employed for a variety of uses: mobile operators' wireless fixed links (4G infrastructure links), fixed satellite service systems and ground stations for space services. Studies will therefore need to be conducted to take these services into account, and define either the cohabitation or migration of applications to other millimetre wave frequencies.

3.2.2 Frequency bands below 6 GHz

a) The 3.4-3.8 GHz band

It will not be possible for 5G to run entirely on millimetre wave frequencies: the propagation qualities of these bands make it difficult to achieve widespread coverage, particularly in more sparsely populated areas. In addition, these bands are still lacking in technological maturity when it comes to delivering consumer market communication services. As a result, a "core" band below 6 GHz needs to be identified: one that provides sufficiently large channels to enable future 5G operators to provide innovative services and a higher quality of service than with 4G.

The 3400 - 3800 MHz band appears to be a good candidate.

First, it has already been harmonised for ultrafast mobile services inside the European Union. Initially the 3400 - 3600 MHz band and later the 3600 - 3800 MHz band were identified as the "IMT" bands (for high-speed mobile). Second, they have a great deal of available spectrum (up to 400 MHz). Lastly, technological advancements (antenna and signal processing) make these frequencies compatible with their use to establish macro cells, and not only microcells. To give an example, it emerged from the interviews conducted by Arcep that coverage with this band could be similar to coverage with the 2.6 GHz band, the core 4G frequency band.

This analysis has been confirmed by the European Commission (in its 5G mandate to RSCOM[37]) and by RSPG[38] which consider the 3.4 - 3.8 GHz band as the only credible 5G band for deployments taking place before the end of 2020.

In France, the band is assigned to Arcep, in a priority fashion, for the 3400 - 3600 MHz frequencies (and in a non-priority fashion to the Ministry of the Interior and the Ministry of Defence), and exclusively for the 3600 - 3800 MHz frequency bands. It is used for wireless local loop and satellite applications.

The compatibility of future uses with current applications is being examined by the Electromagnetic Compatibility Advisory Committee, which is chaired by the National Frequency Agency and of which

---

[35] http://rspg-spectrum.eu/wp-content/uploads/2013/05/RPSG16-032-Opinion_5G.pdf

[36] The tuning range is vital here: a system capable of regulating its frequency from 26 to 28 GHz appears entirely technologically achievable within the timeframe for the first 5G millimetre wave deployments. This would have been less true if a switch between 28 and 32 GHz were required.

[37] https://circabc.europa.eu/sd/a/448dc765-51de-4fc8-b6e0-56ed6a1d0bca/RSCOM16-40rev3%205G%20draft_mandate_CEPT.pdf

[38] http://rspg-spectrum.eu/wp-content/uploads/2013/05/RPSG16-032-Opinion_5G.pdf

Arcep is a member: exclusion zones, protecting certain locations tied to long-term use of satellite solutions, need to be set up. Military radiolocation systems operating below 3.4 GHz also need to be protected.

b) Other bands below 6 GHz

The bands being used today for 2G, 3G and 4G could be used for future 5G deployments.

Refarming 2G, 3G and 4G could be a delicate undertaking, however, because of the duplexing methods they use. Most mobile communications in Europe use FDD (Frequency Division Duplexing)[39] to exchange information. Technical discussions over 5G are nevertheless anticipating that TDD (Time Division Duplexing)[40] will be the chief, if not sole, form of duplexing used for this new generation, notably because it makes it possible to adjust bandwidth to data rates and because it is particularly efficient when beamforming is used.

For 5G networks, using already harmonised mobile bands will thus require in-depth technical studies to define the terms and conditions of use and sharing with existing services. On this point, the ECC (Electronic Communications Committee[41]*)* decided to assess the potential of certain already harmonised bands, notably the 700 MHz and the L band (1427-1492 MHz).

In France, Arcep allocated the 700 MHz band to mobile operators in late 2015. Even if the country's four operators all obtained frequencies in this band, only Free Mobile, which has no blocks of 800 MHz spectrum, decided to begin its rollouts in the band using LTE (4G) technology.

Arcep is currently analysing the contributions to its public consultation, addressing among other the 3.5 GHz band, that ended recently.

The L band (1427 - 1518 MHz), which was defined to be used exclusively in SDL (Supplemental DownLink) mode, could be considered to meet the constant demand for ever higher data rates and the greater increase in downlink traffic, compared to uplink traffic. In France the band is used by wireless fixed links authorised by Arcep, by the Ministry of Defence for mobile services (excluding aeronautical) and by the Ministry of the Interior. Using it for 4G or 5G would therefore suppose moving the above-mentioned uses over to other bands.

3.3 Increasingly small cells

Today, mobile network rollouts are essentially based on the use of macro base stations: installations that are equipped with high-power antennae deployed to guarantee coverage for a relatively wide area, providing good quality of service. Network configurations are evolving constantly: new radio base stations are installed on a regular basis to increase the networks' capacity, to better meet users' needs in terms of indoor and outdoor coverage, and to improve quality of service.

However, this continually growing demand for capacity already requires operators to increase the density of their networks with smaller and smaller cells.

5G – which will probably bring about a sizeable increase in data traffic, and which will use millimetre wave frequencies whose propagation capabilities are weak – will no doubt require the widespread deployment of low-power base stations (small cells).

---

[39] A technique that consists of sending and receiving data simultaneously, but on two different frequency bands.

[40] A technique that consists of sending and receiving data on the same frequency band but at different times.

[41] The European body that specifies the technical terms and conditions governing the use of frequency bands, which are then set by European Commission decisions.

To satisfy demand and enable the introduction of 5G, estimates indicate at least 10 small cells per macro base station in urban settings[42], where cells are already today relatively tightly meshed to handle traffic density.

Here, the players that met with Arcep raised several questions that will no doubt need to be answered when considering the ubiquitous deployment of small cells.

3.3.1 Taxation

The current regulatory framework stipulates that base stations whose power requires an opinion, an agreement or a statement from the National Frequency Agency (ANFR) are subject to a flat tax on network companies (IFER[43]). The amount of the tax varies according to transmission power, the type of installation and the deployment location. For a deployment in an urban area, the tax stands at €1,607/year/installation for a base station with an effective isotropic radiated power (EIRP) of more than 5W (a COMSIS agreement from ANFR is required to be able to transmit) and €160.70 €/ year/installation for an EIRP of between 1W and 5W (declaration to ANFR is required to be able to transmit).

5G small cells will likely use variable transmission powers of between 1W and 25W. In light of rollout density forecasts, hence the number of small cells to install, some of the stakeholders that Arcep met with raised the question of possibly adapting this tax in such a way as to enable massive small cell deployments without generating an equally massive increase in the total tax amount.

This process is already underway, notably with a view to decreasing taxes on base stations in locations that are hard to cover: Act No. 2016-1888 of 28 December 2016 on the modernisation, development and protection of mountain regions thus exonerates mobile base stations built in mountain regions between 1 January 2017 and 31 December 2020 from paying the IFER tax.

3.3.2 Access to elevated and “semi-elevated” locations

To perform their deployments successfully, mobile operators have traditionally needed to install their base stations in elevated locations (towers, rooftops, etc.). This will continue to be true with 5G networks, but will be even more challenging for two main reasons:

1. 5G antennae will probably be larger (in the m^2 range for some) than current 2G, 3G or 4G antennae, because of the above-mentioned massive MIMO processing that will require the use of a very large number of radiating elements. Moreover, additional antennae compatible with new 5G bands will no doubt also need to be deployed. So the re-use of existing masts could very well be problematic, and new (possibly collocated) transmission sites will need to be found.
2. This search for new sites will also need to be carried out to install small cells in semi-elevated locations, but with extra care as the density of these installations will undoubtedly be high: operators will thus be required to deploy their equipment on urban furniture and infrastructures such as bus shelters, lampposts, public buildings, billboards, etc.

As a result, public authorities will need to keep a close eye on the matter and, if necessary, adopt measures that will facilitate 5G rollouts.

---

[42] http://www.lemag-numerique.com/wp-content/uploads/2015/10/WP_-Souverainete_Telecoms_PetitesCells_FINAL.pdf

[43] IFER = *Imposition forfaitaire sur les enterprises de réseaux*

3.3.3 5G networks’ regional coverage and backhaul

Ensuring regional connectivity will be one of the challenges for this new generation of mobile networks.

The diversity of use-cases envisioned for future 5G networks, the geographical distribution of which is not yet precisely known, must be factored in when addressing coverage issues.

The higher frequency bands that are being put forth for future 5G networks, along with the potentially very substantial bandwidth consumed by these new uses, pose the question of the regional foothold of these networks. As a matter of fact, mobile networks have never before employed such high frequencies whose use will require a large number of relay antennae to be installed.

Moreover, connecting the 5G installations to the network will push to the fore the question of the cost of connecting them via optical fibre, which will no doubt be necessary in most instances to ensure the expected quality of service. The industry needs to design the technologies that will make it possible to minimise the cost of 5G rollouts in rural areas.

3.4 Net neutrality issues

European regulation on safeguarding an Open Internet[44], adopted by the European Parliament and Council on 25 November 2015, for which European regulators required an additional nine months to specify the rules governing its application, introduces the principle of net neutrality as one of the top priorities in the standardisation hierarchy.

Net neutrality is an overriding principle that guarantees equal treatment for all data traffic on the internet. In particular, it excludes any form of discrimination with respect to the source, the destination or the content of data flows.

On 30 August 2016, BEREC (the Body of European Regulators for Electronic Communications) published guidelines for national regulators on the enforcement of the European Open Internet regulation[45].

During the BEREC public consultation[46] on its draft net neutrality guidelines, several enterprises and electronic communications sector stakeholders took the opportunity to deliver a clear-cut point of view in their *“5G manifesto for timely deployment of 5G in Europe”*[47]. This manifesto aims at warning public authorities against a too restrictive approach to traffic management, and especially of the supposedly negative effects that, in their opinion, an overly strong enforcement of net neutrality could have on the 5G rollout roadmap.

---

[44] Regulation (EU) 2015/2120: http://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32015R2120&from=EN

[45] Internet service providers (ISP) can employ reasonable traffic management measures for certain categories of traffic, under non-discriminatory conditions and provided they are transparent, proportionate and justified by objective technical requirements, and not used to serve their own commercial interests. ISPs can also distinguish certain services, called specialised services, from their internet access service, without them affecting the latter, provided these services have specific, objective transit requirements.

[46] http://www.berec.europa.eu/

[47] http://telecoms.com/wp-content/blogs.dir/1/files/2016/07/5GManifestofortimelydeploymentof5GinEurope.pdf

BEREC's opinion[48], in response to the contributions to the aforementioned public consultation, recalls that the rules of net neutrality are technology neutral and therefore apply to 5G networks. Moreover, BEREC guidelines clarify that network slicing which networks will need to use to respond to the different use cases (mMTC, eMBB, uRLLC) planned for this new generation (cf. 1.4) system, may be used to deliver specialised services[49].

The subject of net neutrality as it applies to future 5G networks is still very much open and unexplored, but new analyses could be performed parallel to the work being done on defining 5G. Arcep is open to discussion and will also work, within BEREC, towards bringing the required clarity to the ecosystem.

---

[48] http://berec.europa.eu/eng/document_register/subject_matter/berec/download/0/6161-berec-report-on-the-outcome-of-the-publi_0.pdf

[49] http://berec.europa.eu/eng/document_register/subject_matter/berec/download/0/6160-berec-guidelines-on-the-implementation-b_0.pdf - footnote 26

Annexes

Annex 1 Definition and standardisation work

A new generation of mobile telephony is defined primarily by two bodies: ITU (International Telecommunication Union) and 3GPP (3rd Generation Partnership Project). These two, respectively public and private, organisations are dedicated to defining the objectives, standards and technical specifications of the new technology in question.

ITU

The first large-scale 5G commercial rollouts are expected to take place in 2020. As mentioned earlier, the exploratory phase – which provides an opportunity to determine demands and identify the most promising techniques and technologies for these future 5G networks – has already begun. Although a number of organisations and consortia are involved in defining 5G, 3GPP will very probably be the central standardisation body for its technical specifications.

Whatever the case may be, ITU (International Telecommunication Union) is vital to defining the technologies and standards that govern any new generation of IMT (International Mobile Telecommunications) at the global level.

These IMT standards are established with the involvement of public authorities and industry players, and have provided the framework for the evolution of mobile communication services around the world, since the beginning of IMT standardisation, with IMT-2000 (3G, UMTS), then IMT-Advanced (4G, LTE-A) and more recently IMT-2020 (5G).

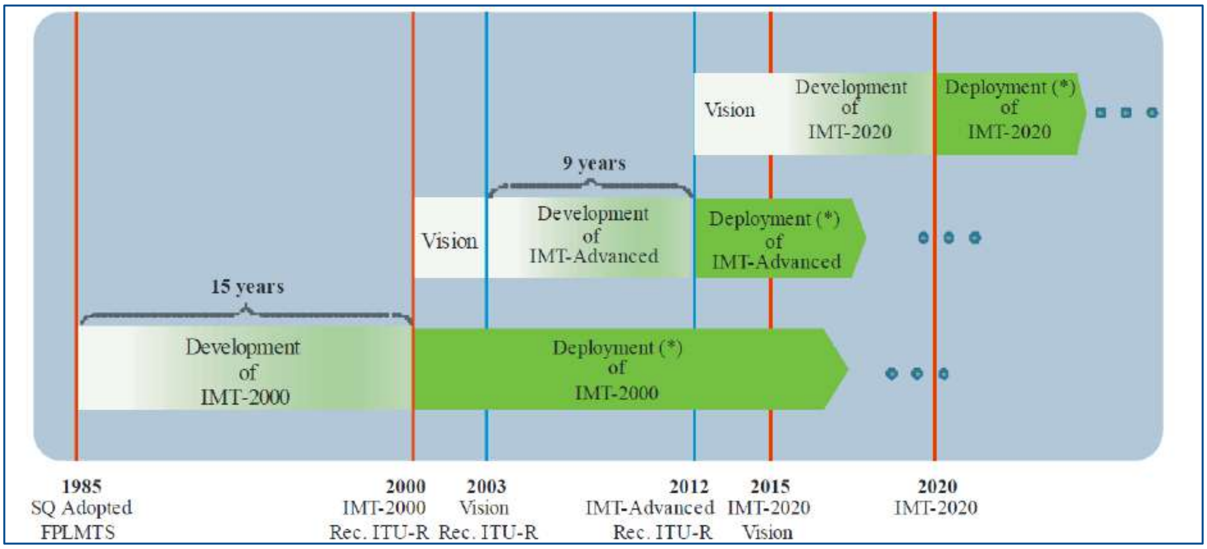
Vision
Development of IMT-2020
Deployment (*) of IMT-2020
9 years
Vision
Development of IMT-Advanced
Deployment (*) of IMT-Advanced
15 years
Development of IMT-2000
Deployment (*) of IMT-2000
1985
SQ Adopted
FPLMTS
2000
IMT-2000
Rec. ITU-R
2003
Vision
Rec. ITU-R
2012
IMT-Advanced
Rec. ITU-R
2015
IMT-2020
Vision
2020
IMT-2020

Figure 16. Timeline of IMT developments and deployments[52]

The advent of an IMT standard typically occurs in three main stages : a vision stage, a development stage and a deployment stage.

The vision stage – whose completion is marked by a document that is usually called a Vision Recommendation – is the stage during which ITU sets the framework and objectives for the future technology. The aim of this document is generally to define what the new technology will be in a more or less concrete fashion, what its characteristics will be, the uses it will enable, etc. Whether for

---

[52] Recommendation ITU-R M.2083-0 (09/2015), https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf

IMT-Advanced, the commonly used technical abbreviation attributed to the definition of 4G, or IMT-2020, the abbreviation attributed to definition of 5G, the vision stage lasts around three years.

Recommendation ITU-R M.2083-0 was published in September 2015, bringing to a close three years of work performed by ITU-R (Working Party 5D) on defining the framework and objectives for IMT for 2020 and beyond. It is this document that today serves as the basis of the different 5G research and standardisation work being done around the globe.

Next comes the standards development stage, based on the conclusions of the vision stage. Regarding IMT-2020, this development and standardisation work is already underway, and ITU plans on having completed it by 2020 to be able to satisfy the most pressing needs of the ITU members and organisations that want to deploy 5G as quickly as possible. This is all the more challenging as the development stage will last only five years, compared to 15 years for IMT-2000 and nine years for IMT-Advanced.

The different deadlines set for IMT-2020 within ITU can be found in Figure 17. The spectrum identification phases (indicated by the black triangles) coincide with the World Radiocommunications Conferences, of which the latest was WRC-2015 and the next is WRC-2019 (see 3.2).

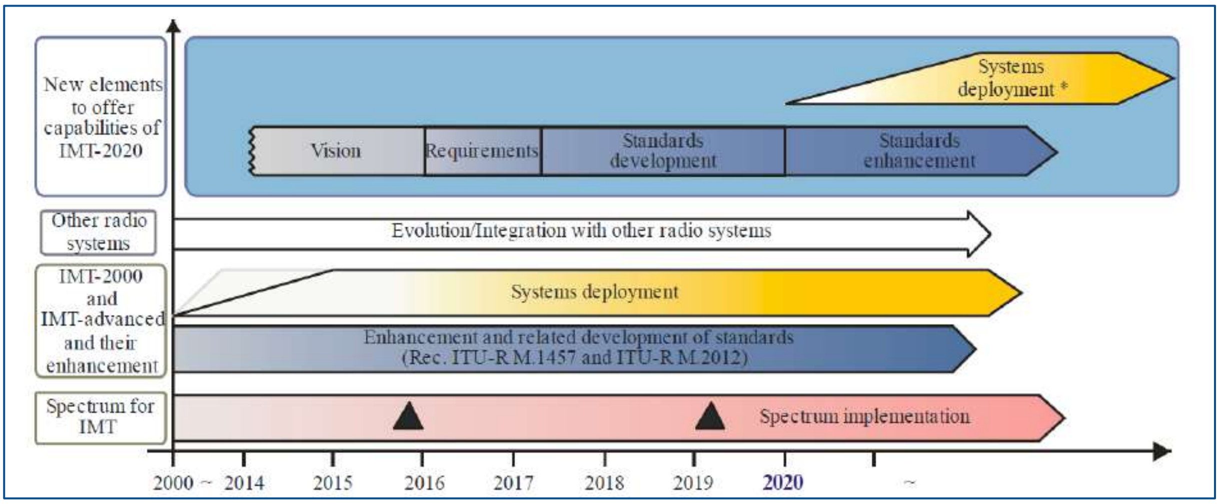
New elements to offer capabilities of IMT-2020
Systems deployment *
Vision
Requirements
Standards development
Standards enhancement
Other radio systems
Evolution/Integration with other radio systems
IMT-2000 and IMT-advanced and their enhancement
Systems deployment
Enhancement and related development of standards (Rec. ITU-R M.1457 and ITU-R M.2012)
Spectrum for IMT
Spectrum implementation
2000 ~ 2014
2015
2016
2017
2018
2019
2020
~

Figure 17. Stages and expected deadlines for IMT-2020[53]

All of the work that ITU is conducting on IMT-2020 is following a roadmap that is detailed in Figure 18, with the completed stages indicated in green and those to come in blue. The work that is underway today concerns 5G spectrum aspects, prerequisites and assessment criteria, along with the different technical studies and proposals, which are a prerequisite to the standardisation phase of the work.

---

[53] Recommendation ITU-R M.2083-0 (09/2015), https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf

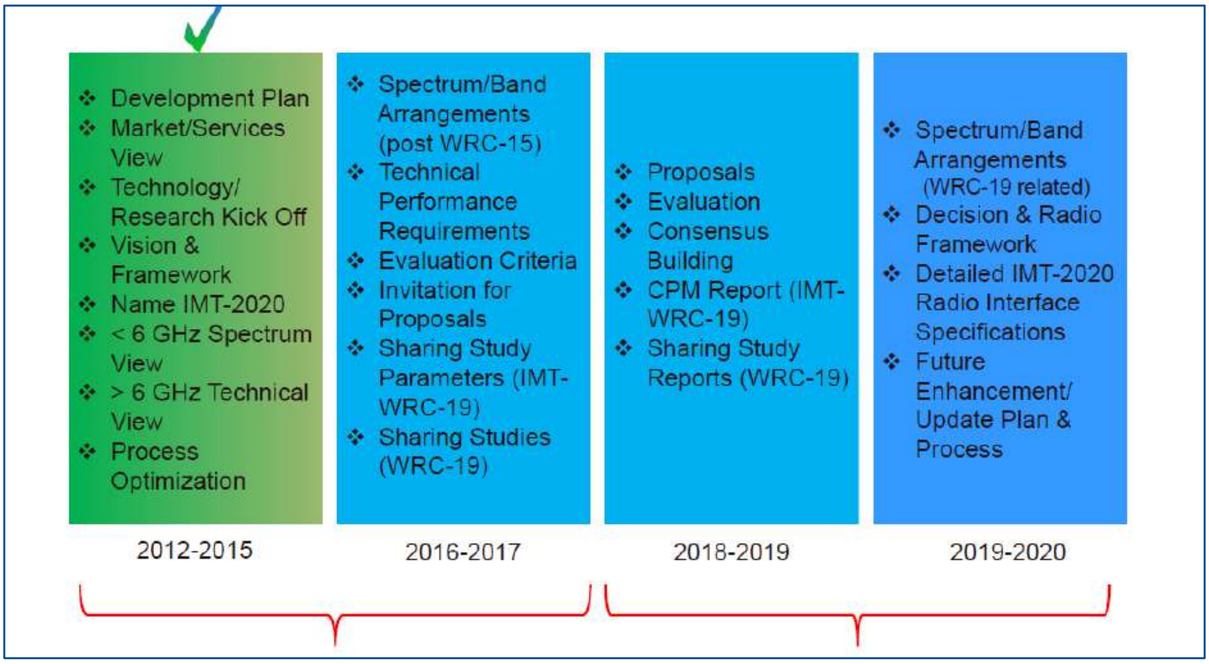
Development Plan
Market/Services View
Technology/ Research Kick Off
Vision & Framework
Name IMT-2020
< 6 GHz Spectrum View
> 6 GHz Technical View
Process Optimization
Spectrum/Band Arrangements (post WRC-15)
Technical Performance Requirements
Evaluation Criteria
Invitation for Proposals
Sharing Study Parameters (IMT-WRC-19)
Sharing Studies (WRC-19)
Proposals
Evaluation
Consensus Building
CPM Report (IMT-WRC-19)
Sharing Study Reports (WRC-19)
Spectrum/Band Arrangements (WRC-19 related)
Decision & Radio Framework
Detailed IMT-2020 Radio Interface Specifications
Future Enhancement/ Update Plan & Process
2012-2015
2016-2017
2018-2019
2019-2020

Figure 18. WP 5D[54] work programme

3GPP

Based on ITU Recommendations, 3GPP has played a major role in the success of LTE over the past several years, which has been the fastest growing cellular technology to date: never before has a new radio standard been adopted and deployed as rapidly and as widely from the finalisation of the first version of its standardisation (for LTE Release 8, in December 2008).

For the first time, under the LTE acronym, the entire mobile industry agreed on a single new technology (contrary to 3G where 3GPP and 3GPP2 co-exist, each backing their version of 3G standardisation that complies with IMT-2000 criteria), and so enabling unprecedented economies of scale and momentum in the ecosystem.

After Release 8, the work performed by 3GPP has been centred on the following strategic areas:

- Enhancing LTE radio standards to further improve capacity and performance;
- Enhancing system standards to make LTE and EPC (Evolved Packet Core, the core LTE network technology) available to new business segments;
- Introducing improvements for system robustness, especially for handing exponential smartphone traffic growth.

These areas of focus have made it possible to map out the general path of evolution from LTE to LTE-Advanced (Releases 10 to 12) and later LTE-Advanced Pro (Releases 13 and 14) while awaiting 5G.

The ambitiousness of 5G requirements, the tight timeline “imposed” by the market, along with the different national tendencies have pushed 3GPP to define two stages of specification work:

1. A first stage that will end in the second half of 2018, with the termination of Release 15, and which will address the most urgent matters with respect to commercial requirements;

---

[54] https://www.itu.int/en/membership/Documents/missions/GVA-mission-briefing-5G-28Sept2016.pdf

2. A second stage that will end in December 2019, which marks the end of Release 16, which will address the other use cases and prerequisites identified in the IMT-2020 Vision Recommendation.

The Work Plan for Release 15, the first set of 5G specifications, was ratified at the plenary meeting of the TSG#72 (Technical Specification Group #72) 3GPP working groups in June 2016. This plan includes a set of intermediate tasks and checkpoints to steer ongoing work in the different groups. This Work Plan sets out the transition from current studies to the standardisation phase of the work:

- Starting in December 2016: beginning of standardisation work by TSG SA2 (Technical Specification Group System Architecture);
- Starting in March 2017: specification begins for the 5G NR (New Radio) interface in the TSG RAN (Technical Specification Group Radio Access Network).

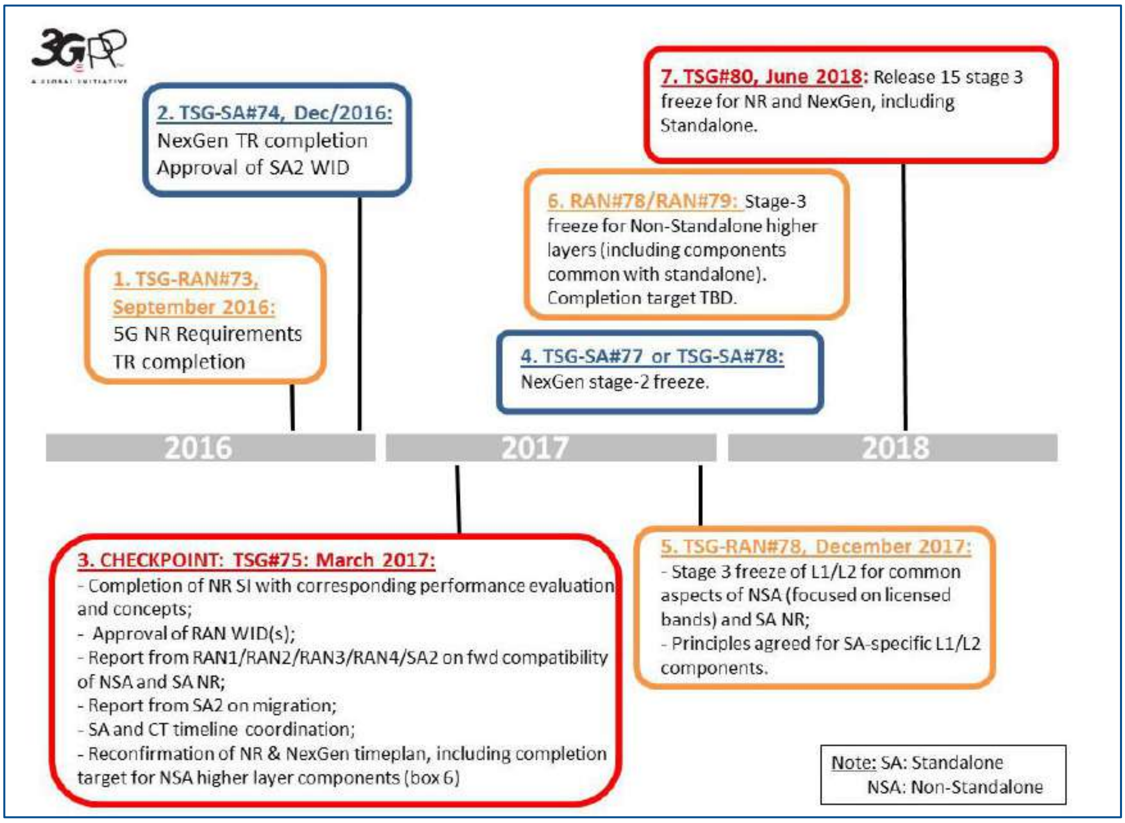

2. TSG-SA#74, Dec/2016:
NexGen TR completion
Approval of SA2 WID
7. TSG#80, June 2018: Release 15 stage 3 freeze for NR and NexGen, including Standalone.
6. RAN#78/RAN#79: Stage-3 freeze for Non-Standalone higher layers (including components common with standalone). Completion target TBD.
1. TSG-RAN#73, September 2016:
5G NR Requirements TR completion
4. TSG-SA#77 or TSG-SA#78:
NexGen stage-2 freeze.
2016
2017
2018
3. CHECKPOINT: TSG#75: March 2017:
- Completion of NR SI with corresponding performance evaluation and concepts;
- Approval of RAN WID(s);
- Report from RAN1/RAN2/RAN3/RAN4/SA2 on fwd compatibility of NSA and SA NR;
- Report from SA2 on migration;
- SA and CT timeline coordination;
- Reconfirmation of NR & NexGen timeplan, including completion target for NSA higher layer components (box 6)
5. TSG-RAN#78, December 2017:
- Stage 3 freeze of L1/L2 for common aspects of NSA (focused on licensed bands) and SA NR;
- Principles agreed for SA-specific L1/L2 components.
Note: SA: Standalone
NSA: Non-Standalone

Figure 19. Roadmap for 3GPP standardisation groups[55]

If 5G is viewed as the technology capable of transforming society, and even of ushering in the next industrial revolution by impacting multiple (vertical) sectors with new business models and in a way that benefits the economy, notably in Europe, in all likelihood it will initially be driven forward by eMBB (Enhanced Mobile Broadband, cf. 1.2).

With this in mind, 3GPP defined the framework for Release 15, stage 1 specifications of 5G standardisation, as detailed earlier, whose work will focus on the following:

- Standalone and Non-Standalone NR operation: the Standalone version to work independently of LTE technology, and the Non-standalone NR being highly interoperable with LTE[56].

---

[55] https://5g-ppp.eu/wp-content/uploads/2016/11/01_10-Nov_Session-3_Dino-Flore.pdf

- Target use cases: starting with Enhanced Mobile Broadband (eMBB) as well as Low Latency and High Reliability to enable some URLLC (Ultra-reliable and Low Latency Communications, cf. 1.2) use cases.
- Simultaneous examination of frequency ranges below 6 GHz and above 6 GHz.

Release 16, the second phase of 5G specifications, will then focus more on other segments such as mMTC (Massive Machine Type Communications, cf. 1.2), for instance, to tackle IoT use cases that require higher quality of service than what LPWAN (Low-Power Wide-Area Network)[57] can provide.

---

[56] In its non-standalone version, the NR control plane is LTE's. in other words, the 4G network controls the 5G carriers and spreads users across the different bands and technologies.

[57] Sigfox and LoRa LPWAN

Annex 2 Entities Arcep met with

ERICSSON
QUALCOMM
HUAWEI
SAMSUNG
NOKIA
AIRBUS DEFENCE & SPACE
THALES
RENAULT
Systematic Paris Region Systems & ICT Cluster
TELECOM ParisTech
leti ceatech
f
Google
ETSI World Class Standards
bouygues TELECOM
free mobile
numericable THD
SFR
orange
GSMA

Annex 3 Figures

Figure 1. 5G driving industrial and societal changes 4

Figure 2. Performances required by vertical sectors 6

Figure 3. What we do over the network today in one minute........ 7

Figure 4. 3GPP timeline for 5G 9

Figure 5. 5G use cases 10

Figure 6. Comparison between 4G and 5G with respect to the eight key performance indicators 12

Figure 7. Representation of the “average user experience data rate” KPI for 4G and 5G and “Peak data rate” KPI for 3G, 4G and 5G 13

Figure 8. Key performance indicators for the three 5G use cases 14

Figure 9. Example of the use of a beamforming antenna to connect Wi-Fi access points 16

Figure 10. Illustration of full-duplex, compared to FDD and TDD 16

Figure 11. Illustration de multiplexing *power domain* NOMA 17

Figure 12. Centralisation of the control plane in a software-defined network 18

Figure 13. Illustration of CloudRAN network architecture........ 19

Figure 14. 5G PPP vs. 3GPP and ITU roadmaps........ 22

Figure 15. Millimetre wave frequencies identified at WRC-15 28

Figure 16. Timeline of IMT developments and deployments 35

Figure 17. Stages and expected deadlines for IMT-2020 36

Figure 18. WP 5D work programme........ 37

Figure 19. Roadmap for 3GPP standardisation groups 38

HEAVY
READING

White Paper

Designing 5G-Ready Mobile Core Networks

Prepared by

Gabriel Brown
Senior Analyst, Heavy Reading
www.heavyreading.com

on behalf of

affirmed®

www.affirmednetworks.com

and

JUNIPEr
NETWORKS®

www.juniper.net

September 2016

5G Core Service & Performance Requirements

An industry "Vision for 5G" is now established and supported worldwide. Services and performance requirements have been identified, and the industry is moving into the development and implementation phases with a view to commercial launch of 5G mobile service around 2019 and fixed wireless access using mmWave radio from 2017. To meet these schedules, leading operators – especially Tier 1s with high expectations and aggressive deployment timelines – are working to make their networks 5G-ready ahead of commercial deployment.

This white paper argues that with a 5G-ready technology strategy, operators can prepare for rapid 5G service launch in a way that optimizes their investment in next-generation IP and mobile core platforms over the next three years. Specifically, it discusses 1) how 5G services drive a requirement for an IP services fabric to connect the distributed data centers that will host 5G network functions, content and applications and 2) the development of cloud-native, service-orientated core networks for advanced 4G and 5G networks.

5G Services Dimensions

Development of the 5G system architecture, and associated core network, is being led by 5G service requirements. **Figure 1** shows the output from the 3GPP SMARTER study group that has investigated the service dimensions of 5G.

Figure 1: Service Dimensions in Future 5G Networks

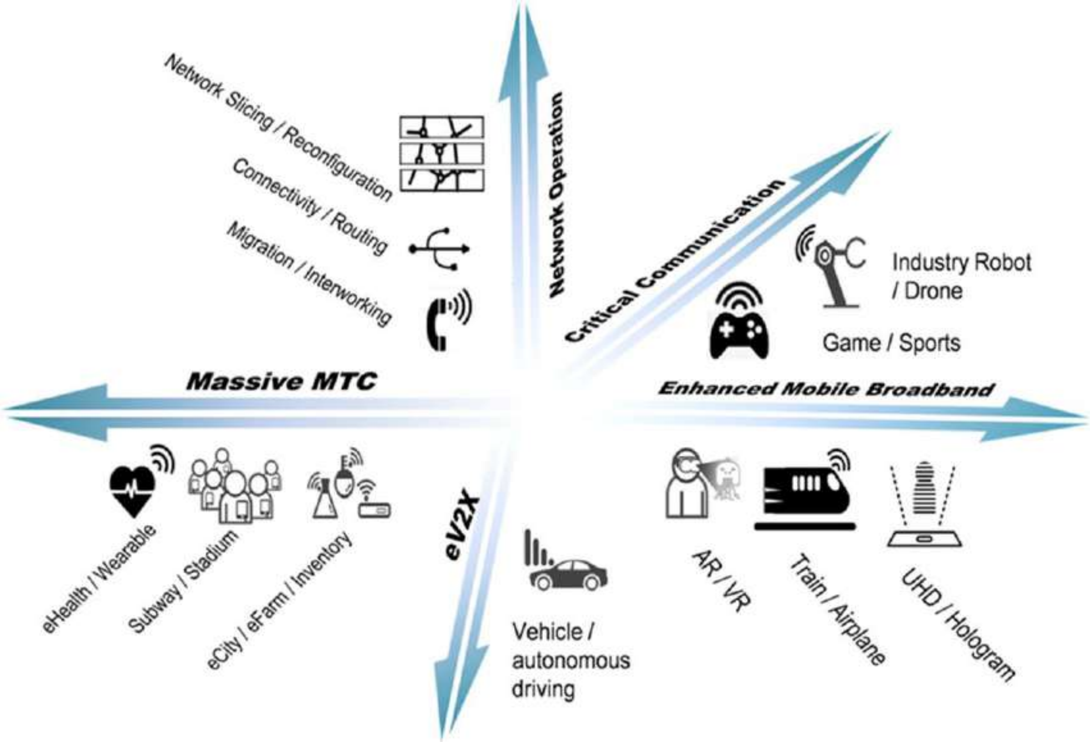
Network Slicing / Reconfiguration
Connectivity / Routing
Migration / Interworking
Network Operation
Critical Communication
Industry Robot / Drone
Game / Sports
Massive MTC
Enhanced Mobile Broadband
eHealth / Wearable
Subway / Stadium
eCity / eFarm / Inventory
eV2X
Vehicle / autonomous driving
AR / VR
Train / Airplane
UHD / Hologram

Source: 3GPP SMARTER

The chart maps closely to the now famous triangle representation of 5G services made up of enhanced mobile broadband (eMBB), low latency and mission critical communications and massive machine-type services (mMTC). However, it adds two important axes: a dedicated eV2X axis for vehicular services and, significantly, a network operations axis.

The dedicated operations axis is important because it makes explicit that automation is fundamental to 5G. To address new commercial opportunities, operators need the ability to create service-specific "network slices" that extend end to end across the infrastructure. Automating set up and management of these "network slices," including configuration of the underlying IP network, is critical to the 5G business case.

These service dimensions incorporate many different performance requirements. **Figure 2** shows IMT-2020 performance targets, developed by the ITU, for 5G. To the left, these are shown in comparison to IMT Advanced (4G). To the right, performance requirements are mapped to the three major use-case categories.

Figure 2: IMT-2020 Performance Targets

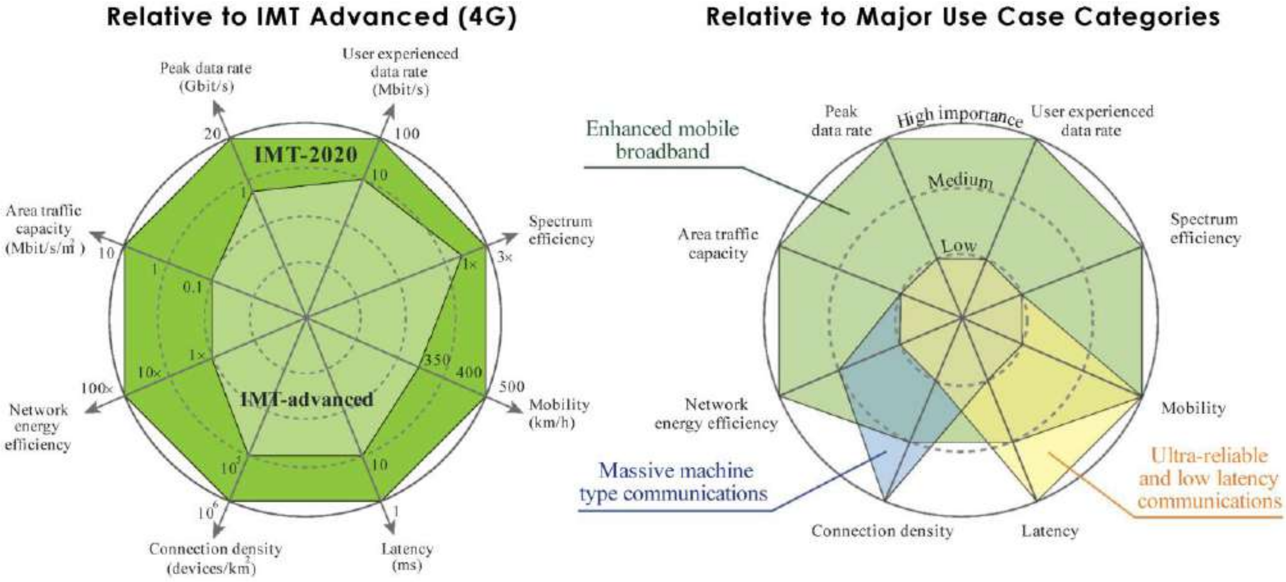

Relative to IMT Advanced (4G)

Relative to Major Use Case Categories

Source: ITU Recommendation ITU-R M.2083-0, September 2015

Probably the most influential factor on 5G physical network deployment will be the low-latency requirement. To deliver services over a wide-area network (WAN) with a consistent 5-10ms roundtrip time is extremely challenging and will drive a new system architecture deployed on a distributed cloud infrastructure.

A 5G-Ready Network Investment Strategy

Operators now have a good idea of what 5G will look like in scope and a reasonable view into the initial deployments models. Development of normative standards for 5G new radio (NR) and the next-generation core network (NG Core) is underway with the first specifications – and the first "compatible" pilot deployments – expected in mid 2018. Although there is uncertainty about the details of the specifications under development, the industry has already undertaken significant R&D work on 5G candidate technologies, making it is possible for operators to pursue a 5G-ready network strategy with a reasonable degree of confidence that they will be able to rapidly deploy 5G, at scale, as the technology becomes commercially available.

From an IP and core network perspective, five tiers of a 5G-ready investment strategy are, from the ground up, as follows:

1. **Develop a distributed data center footprint**. The Central Office Re-architected as Data Center (CORD) model provides a good reference. The idea is for operators to develop physical assets located close to the edge and transform them into distributed data centers. The edge location is critical for low-latency 5G services and, as Mobile Edge Computing (MEC) shows, the edge-cloud is valuable in many networking scenarios – it is not solely a 5G investment. In time, the distributed data center will also come to support Cloud RAN (C-RAN) and Virtual RAN (V-RAN) hub sites and will host core network functions, such as distributed user-plane nodes.
2. **Create an "IP Services Fabric" for 5G with software-defined networking (SDN) control**. Edge-cloud locations running 5G network functions and services will require high-performance, secure connectivity. Investment in wide-area SDN is already underway to create an "IP services fabric" that connects centralized data centers, distributed cloud locations and cell sites. This IP services fabric provides routing, security, service chaining, redundancy and orchestration. Again, this investment is, in many aspects, not unique to 5G, but common to high-performance networking at the cloud edge.
3. **Invest in cloud-based Evolved Packet Core (EPC) and NG Core.** With a distributed cloud infrastructure, it is logical to redesign core network functions to take advantage of this asset to meet the strict latency requirements of 5G services and efficiently manage data traffic. Some aspects of NG Core are already reasonably well understood, (for example, control- and user- plane separation or "CUPS"), while others remain works in progress (for example, mobility and session management). In general, expect control-plane separation to apply across the 5G architecture, from radio to core.
4. **Use network slicing for a service-optimized core.** Network slicing is an important bridge from a 4G core to a new 5G core and is a key component of a 5G-ready network investment strategy. In this model, virtual network functions (VNFs) are created per service and data is routed via an optimized processing path across the network. By isolating services into virtual network slices, operators can offer better security, more efficient transport, optimized core network processing and appropriate service quality. A network slice created for an Internet of Things (IoT) customer, for example, may support both 4G and 5G access networks.
5. **Automation and service orchestration.** Automating network operations provides the ability to launch new services rapidly, to reach new market segments, to evaluate success/failure quickly, and then to modify, scale and repeat. Significant progress is being made in applying service orchestration to 4G-Long Term Evolution (LTE) networks, especially where software-based core networks are deployed. Network slicing is a good example of the need for automation because there will be many more logical entities to commission, provision and manage. "Automated service orchestration" is a good, tangible example of 5G-ready functionality.

The 5G Core Network

The new 5G core network is now reasonably well understood: It will be "cloud-native," it will make extensive use of network slicing, and it will operate in concert with a new model-driven service orchestration layer. The industry can therefore start to develop NG Core products that can be adapted as standards emerge and solidify.

From a practical perspective, for the initial deployment, operators can support 5G radio access on a "5G-ready" EPC and then migrate to a new NG Core over time. In both cases, the "IP services fabric" provides SDN-controlled connectivity and related IP services, across a distributed cloud infrastructure.

Virtualized & Cloud-Native Mobile Core

Operators have made good progress on virtual EPC over the past couple of years. The largest networks now support more than 15 million subscribers (AT&T has discussed this publicly); some progressive operators have started to refresh their main EPC networks using multivendor network functions virtualization (NFV) (e.g., Docomo, Etisalat); and others (e.g., Vodafone) have deployed virtual core elements for IoT services.

Figure 3 shows two migration paths to a "cloud-native" 5G core. The blue line shows a virtualized 4G core as a stepping stone to a "5G-ready" core and then a full 5G core. The grey line shows a more conservative option where the operator scales the classic EPC in the near-term and then makes a bigger leap to 5G later.

Figure 3: A Preliminary Cloud-Based 5G Architecture

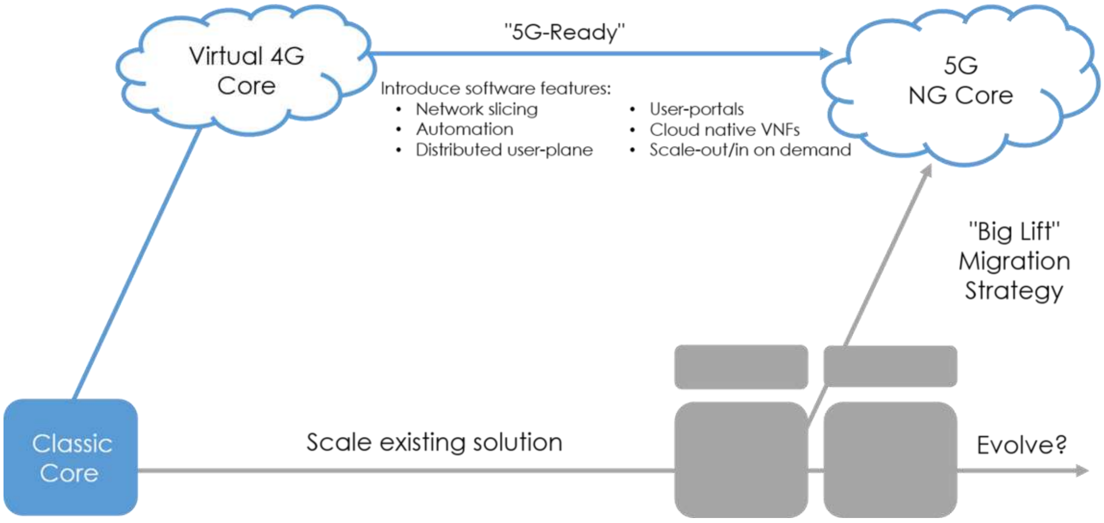
Virtual 4G Core
"5G-Ready"
5G NG Core
Introduce software features:
• Network slicing
• Automation
• Distributed user-plane
• User-portals
• Cloud native VNFs
• Scale-out/in on demand
"Big Lift" Migration Strategy
Classic Core
Scale existing solution
Evolve?

Source: Heavy Reading

There is debate about what exactly "cloud native" means, particularly in the context of stateful telecom network functions. Nevertheless, Heavy Reading's established view is that operators that deploy virtual EPC – even at small scale – gain experience in how these systems work, how to operate them, and how to evolve them, and that in turn will give them a sustainable advantage in the longer-term transition to software-centric 5G mobile networks.

Migrating 5G RAN From EPC to NG Core

A 5G-ready core strategy is determined, in part, by how the operator plans to introduce 5G radio. There are two basic possibilities: operate 5G RAN using an EPC or using a new NG Core. In practice, operators with plans to launch 5G early are likely to start with an EPC and migrate to NG Core over time, as shown in **Figure 4** below.

Figure 4: Migrating 5G RAN from EPC to NG Core

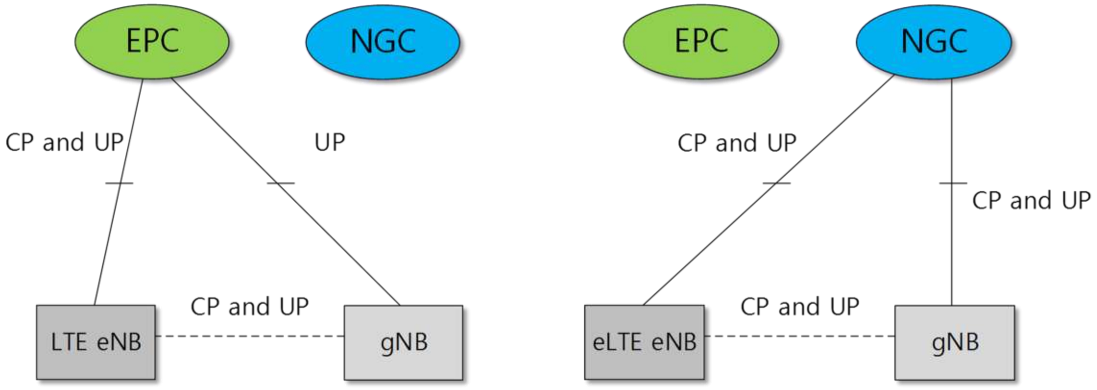

Source: 3GPP TR 38.801 v0.4.0

The diagram is taken from the most recent version of the 3GPP "Study on New Radio Access Technology" (TR 38.801), which will inform development of standards in Release 15. Several permutations of the architecture are under consideration; however, in simple terms, to the left, the new 5G base station (gNB) user-plane interface connects directly to a 4G EPC, while the control-plane functions, such as tracking, paging, etc., are provided by an evolved 4G base station (eNB), which in turn also connects to the EPC. In this scenario, the EPC requires little or no modification, making this a fast and simple way to deploy 5G radio from a core perspective.

Over time, both 4G and 5G base stations can migrate to a new NG Core, which will provide both control- and user-plane functions. At this stage, NG Core becomes the primary core network for 4G and 5G access, as shown to the right in the diagram. This is conceptually similar to how EPC supports 3G and 4G access networks.

Note that in cases where 5G radio is deployed for fixed wireless access, there is no need for an LTE RAN to provide control-plane functions to the 5G user device; a standard EPC is sufficient, although it would need to provide session management for the 5G access. Some of the first 5G deployments are expected to use this model.

Network Slicing: A Key Bridge to 5G

Network slicing is one of the key bridges between the 4G and 5G core. To support diverse service types, operators will use multiple core networks deployed as "network slices" on a common IP services infrastructure. The idea, shown in **Figure 5**, is to create virtual core network instances (or "slices") dedicated to different services. Each slice can be optimized for the traffic profile and the commercial context of the associated service – for example, IoT, public safety, mobile virtual network operator (MVNO), connected car, voice over WiFi or enterprise services. Network slices can be two dimensional in the sense that they can be both service- and customer-specific.

Figure 5: Dedicated Packet Core Network Slices

- Virtualization enables multi-tenant and single-tenant private core networks
- "Slices" configured according to service type & traffic profile
- Multiple options to steer user traffic into a core network processing path
- Mechanisms in 3G/4G include APN routing, MoCN, DECOR

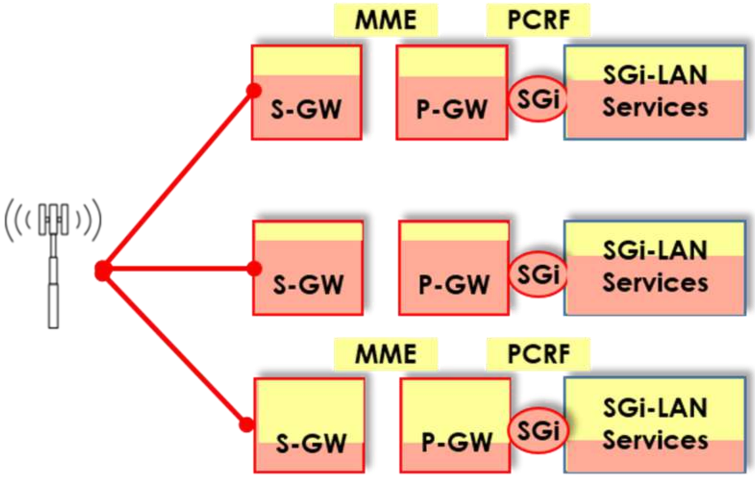

Source: Heavy Reading

IoT is an example of how operators can use "slicing" to support a 5G-ready core strategy. Since many of the core network design parameters are similar between 4G and 5G, investment in a software-based IoT core network can be made with the expectation that the same core (with software updates) will also support IoT services on 5G in future. Moreover, because devices from both access types will connect to a common IoT core, operators will be able to develop integrated 4G/5G IoT strategies that optimize investment and enable them to go to market with narrowband IoT services before 5G specifications are released and radio equipment is deployed.

5G will, naturally, bring more capabilities to network slicing. Most importantly, it is expected that the slice will run end to end across the RAN, core and transport network. Radio is typically deployed as a shared resource, which means that 5G slicing will involve advanced self-organizing network (SON) capabilities and hierarchical, slice-aware scheduling on the air interface. On the network side, the operator can use SDN to reserve resources for the network slice on the IP services fabric.

NG Core for 5G

A new core network offers important benefits, particularly for services with demanding performance requirements, and is an important part of the 5G architecture. The high-level view of the NG core network architecture, as it looks in September 2016, according to TR 23.799 (the Technical Report on the NextGen System Architecture) is shown in **Figure 6**.

The chart shows the primary elements of the new 5G system architecture, including the device (UE), the radio access network (NG-RAN) and the core network that comprises the control-plane (NG-CP) and user-plane (NG-UP) functions. The separation of control and user planes is a direct extension of the "CUPS" concept being developed for advanced EPCs in 4G networks and is another example of how operators can invest in 5G-ready core network designs in the near term.

Figure 6: High Level View of NG Core Architecture (Proposed)

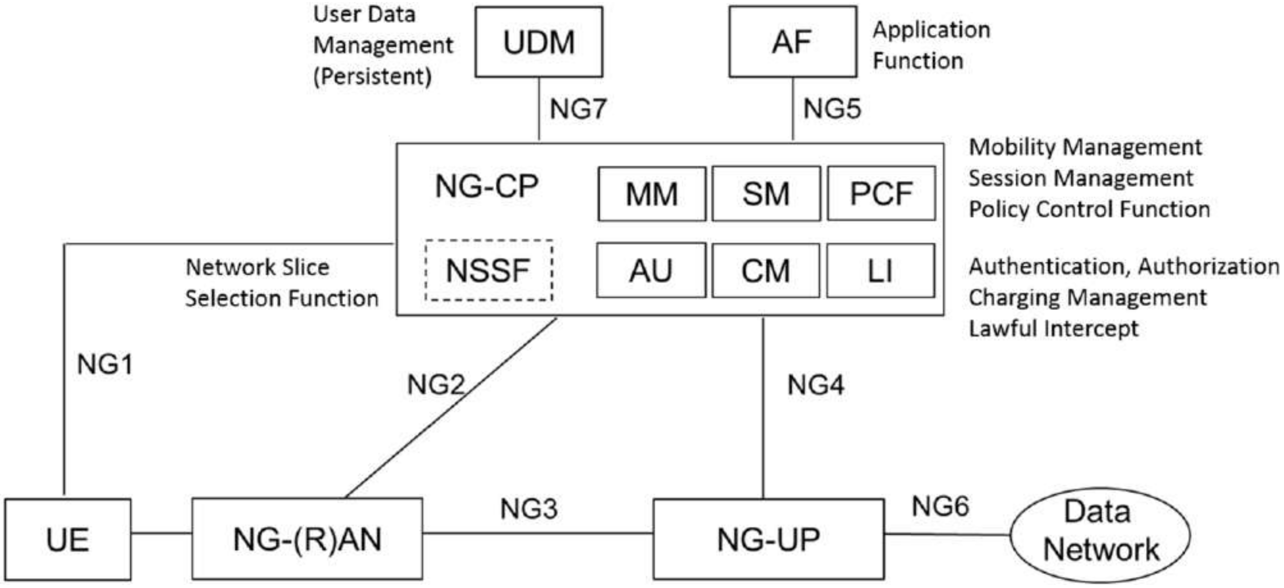
User Data
Management
(Persistent)
UDM
AF
Application
Function
NG7
NG5
NG-CP
MM
SM
PCF
Mobility Management
Session Management
Policy Control Function
Network Slice
Selection Function
NSSF
AU
CM
LI
Authentication, Authorization
Charging Management
Lawful Intercept
NG1
NG2
NG4
UE
NG-(R)AN
NG3
NG-UP
NG6
Data
Network

Source: 3GPP, Adapted from TR 23.799 v0.8.0, September 2016 (Figure 7.2.4-1)

It is interesting to compare this new functional architecture with the existing EPC. **Figure 7** maps the new NG Core elements to their equivalent in EPC. It shows significant overlap between the EPC and NG Core and provides confidence that today's state-of-the-art EPC is reasonably close to the next-generation core in terms of functionality, albeit that the interfaces and protocols will change/evolve.

Figure 7: Mapping NG Core & EPC Functions

NG Core Network Function	Approximate Equivalent EPC Function
UDM (User data management)	SPR
AF (Application function)	AF
NSSF (Network slice selection function)	MME (DECOR) + HSS
MM (Mobility management)	MME + SGW / MME + SGW-C
SM (Session management)	PGW/PGW-C
PCF (Policy control function)	PCRF
AU (Authentication & Authorization)	HSS/AAA
CM (Charging management)	PGW + PCRF, OCS, OFCS
LI (Lawful intercept)	LI
NG-UP (User-plane function)	SGW + PGW / SGW-U + PGW-U

Source: Heavy Reading, Affirmed, Juniper

Key Capabilities of the Next-Gen System Architecture

In addition to proposed architectures, the Technical Report provides guidance on the features and services NG Core should support. It identifies 19 key issues for the Next-Gen System Architecture, which can be grouped into four categories:

- **Flexible Deployment**: Including network slicing, user-plane network selection, network function granularity (e.g., decomposition of VNFs and service chains), interworking and migration, a policy and charging framework.
- **Flexible Access Support**: Including variable core/access splits, a flexible authentication framework (with varying credentials according to device-type, use case and policy), support for relays and multi-hopping, improved network discovery and selection mechanisms.
- **Connectivity**: Including session management, mobility management, a quality-of-service (QoS) framework and service and session continuity (e.g., across accesses).
- **Adapting Existing Capabilities**: Including network/service capability exposure, multicast and broadcast, IP Multimedia Subsystem (IMS) support, off-network communications (e.g., *ad hoc* networks for public safety).

Implementation of these capabilities is tightly linked to both the design of the NG Core and the underlying IP service infrastructure. The 3GPP specifications do not reference connectivity services, but, in practice, there is a close relationship.

A Distributed 5G Services Fabric

The infrastructure onto which 5G will be deployed should support multiple, demanding use cases. This drives a need for a high-performance wide-area IP services fabric controlled (or "orchestrated") by SDN. This IP services fabric should provide connectivity between many different distributed data centers in a meshed architecture that provides resiliency and scalability, and should be programmable such that it can support dynamic, service-specific network slices. In effect, the IP services fabric makes distributed centers act in a unified manner – i.e., behave as one integrated data center. The concept is shown in **Figure 8**.

The design of the distributed IP services fabric is formed by the variety of use cases, the requirement for low latency and high availability, and the need to scale efficiently. In tandem, the mobile network architecture will evolve to a more distributed model. To meet 5G service requirements, NG Core will be deployed using the "CUPS" model, with user-plane nodes hosted at distributed data center locations, and control-plane nodes at more centralized locations, interconnected by the IP services fabric.

Service orchestration will generally be domain specific, with SDN-controlled IP services, cloud resource management, NFV lifecycle management and end-user service orchestration, all operating quasi-independently in a layered architecture. Coordination between the layers will use "cross domain orchestration" to ensure a network slice contains all the networking components needed to deliver the service.

Depending on how the architecture evolves, there will be a need to support highly accurate, 1588v2-based timing (frequency and phase) within this IP services fabric. Virtual RAN/Cloud RAN is an example on the network side that benefits from accurate network timing. On the customer side, it is expected that some market segments, such as the financial industry, will also require timing to be incorporated into network slices.

Figure 8: IP Services Fabric for 5G Networks

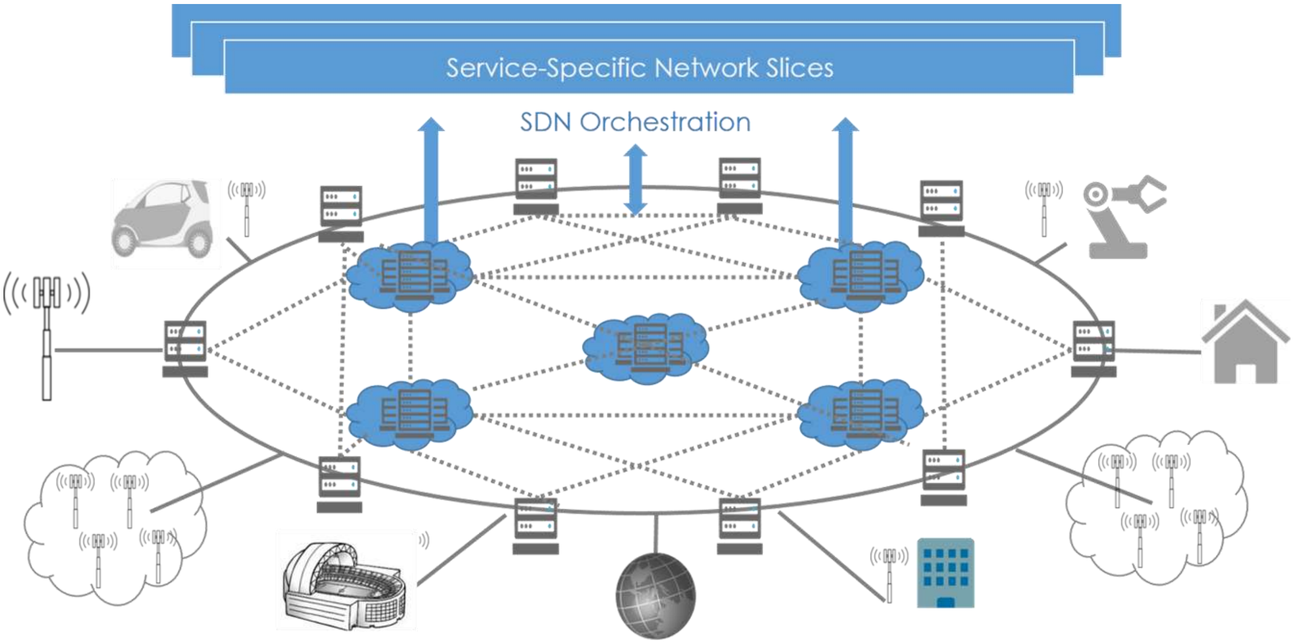

Source: Heavy Reading

Distributed Data Centers

To provide access network services, telecom operators have a physical footprint – in the form of central offices, base station controller sites and transport aggregation sites – that they can convert into distributed or micro data centers.

Depending on the service requirements, these data centers can be used to terminate access connections using virtualized edge functions and become the obvious place to host latency-sensitive content and applications. For centralized cloud providers, this capability is harder to replicate unless they build out an extensive physical footprint or partner with an access provider, and, therefore, this strategy provides operators with a potentially important and sustainable competitive advantage.

The "edge cloud" model will be critical to 5G and is the subject of detailed R&D across the industry. As an example, a preliminary architecture developed by the SuperFluidity project (a European 5G research project) is shown in **Figure 9**. Work on this particular project, which is academic in focus but contains several industrial partners, is scheduled to finish by the end of 2017.

As the name implies, the project aims to achieve superfluidity in the network: the ability to instantiate services on the fly, run them anywhere in the network (core, aggregation, edge) and shift them transparently to different locations. This is the essence of the "IP Services Fabric" for 5G. In this new network, operations will also need to move from a "red light, take action" model to an automated model where the orchestration and associated next-gen operations support system (OSS) detect and fix problems with the IP and NFV layers, and re-optimize the network accordingly.

Figure 9: A Preliminary Converged Cloud-Based 5G Architecture

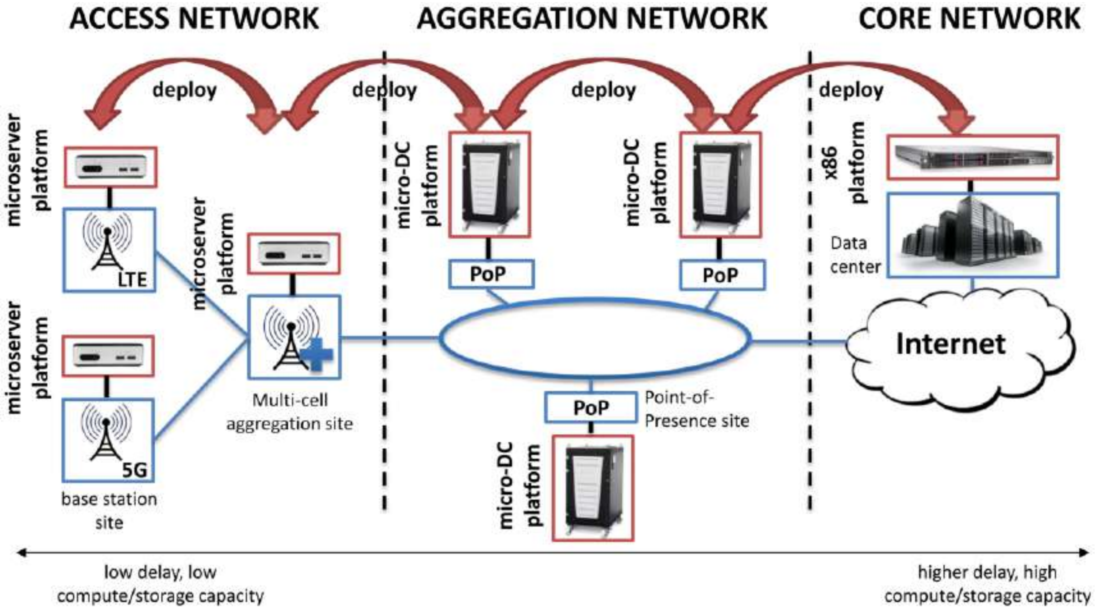
ACCESS NETWORK
AGGREGATION NETWORK
CORE NETWORK
deploy
deploy
deploy
deploy
microserver platform
LTE
microserver platform
microserver platform
5G
base station site
Multi-cell aggregation site
micro-DC platform
PoP
micro-DC platform
PoP
x86 platform
Data center
Internet
PoP
Point-of-Presence site
micro-DC platform
low delay, low compute/storage capacity
higher delay, high compute/storage capacity

Source: [SuperFluidity]

Central Office Re-architected as a Data Center (CORD)

There are several industry and service provider initiatives to define the architecture for SDN-enabled distributed data centers. The Central Office Re-architected as a Data Center (CORD) initiative hosted by the Linux Foundation is one example of how operators aim to make better use of these assets.

According to the CORD initiative, AT&T alone operates 4,000 to 5,000 central offices, each serving 10,000 to 100,000 residential, enterprise and mobile customers. These central offices contain fragmented vendor hardware with multiple physical appliances installed per site (AT&T said it has 300+ unique appliances deployed in its central offices nationwide). The opportunity is to re-architect these central offices to support edge cloud infrastructure and deploy VNFs in place of appliances.

Each CORD location will be connected using an SD-WAN, making it possible to load-balance content and NFV workloads across the distributed cloud using the same SuperFluidity concepts discussed above. In the CORD case, the focus is on the ONOS controller, but in the sense that this is a generic architecture, multiple SDN controller options are viable and attractive. Note that for I/O-intensive workloads, the selection of an edge location should consider the physical transport resources available at the site. It would not make sense, for example, to deploy a 5G user-plane node for fixed wireless access in a central office that does not have a high bandwidth connection to the Internet.

In March 2016, the Linux Foundation announced the M-CORD initiative for mobile operators, backed by AT&T, SK Telecom, Verizon, China Unicom and NTT Communications. It highlighted three key aspects of the architecture, saying that it:

- Will use the same CORD principles of elastic commodity cloud and SDN to bring data center economies and cloud agility to the mobile edge.
- Will demonstrate integration of disaggregated/virtualized RAN, disaggregated/virtualized EPC and mobile edge services.
- Will partner with the SDN controller groups to accelerate adoption of open source SDN and NFV solutions and realize the benefits of the cloud.

Given the timing of the work, we expect M-CORD and similar initiatives to turn their focus toward "5G-ready" core networks as the architecture, interfaces and protocols for NG Core become more clear. A common CORD and M-CORD implementation, with the same architecture and foundational technologies, will create a good foundation for fixed-mobile network convergence, enabling access agnostic services – an important objective of many operators pursuing 5G.

Mapping Mobile Core to Distributed Cloud

The NG Core for 5G and EPC for advanced 4G networks must be mapped to the distributed cloud architecture. One approach would be to simply deploy more packet gateways (and mobility controllers) at the edge of the network to meet capacity and performance demands.

The challenge with this is that today's centralized packet core deployments are characterized by complex integration with surrounding network functions, such as policy, charging, IMS, SGi-LAN and routing services. By moving this model to the edge, the operator would, in effect, have to "distribute complexity," which is costly to deploy and, in particular, to manage. To meet 5G performance, scalability and automation requirements, a new architecture for packet core is needed that will make operation in the edge cloud infrastructure simpler and faster.

There are many aspects to this new architecture. Part of the solution is CUPS, as is currently being developed for 4G-LTE core networks. This involves extracting the control-plane functions from the gateway to leave a simpler, user-plane node. The gateway thus is "split" into S/PGW-U and S/PGW-C components that can that can scale independently, as shown in **Figure 6** above. A key benefit of the architecture is that the control plane, and all the associated complex interactions, can be centralized, while the user plane is distributed across the IP services fabric and scaled as required by the traffic load. This is shown in **Figure 10**.

Depending on the scaling needs, the S/PGW-U functions deployed at the edge data center can be implemented in several different ways: on a router, on a white box switch (potentially), on an existing PGW platform, as a virtualized function, or as part of a vRouter. Virtualized user-plane nodes can more easily be placed at the optimal location, as determined by the use case and this flexibility is a strong argument to deploy S/PGW-U nodes as VNFs rather than as hardware functions.

The S/PGW-C components would similarly be deployed as virtualized functions on a cloud platform, typically at a more centralized location. There may be an opportunity to collapse MME, SGW-C, PGW-C functions and perhaps other 3GPP control-plane functions into some form of unified "mobility controller" node. This has the potential to simplify operations and this same model is being pursued for NG Core.

In this architecture, virtualized SGi service LAN components can be placed where appropriate for the traffic coming from the distributed user-plane nodes. For example, some SGi functions could be on router-based compute blades, or on COTS x86 servers deployed at the distributed site, while others could be in the central data center or close to Internet peering points. The selection of the site for the SGi functions is carried out by the central orchestration system, which steers traffic into service chains on a per subscriber, application, bearer, device or combination basis.

Figure 10: Distributed User-Plane at Edge Data Centers

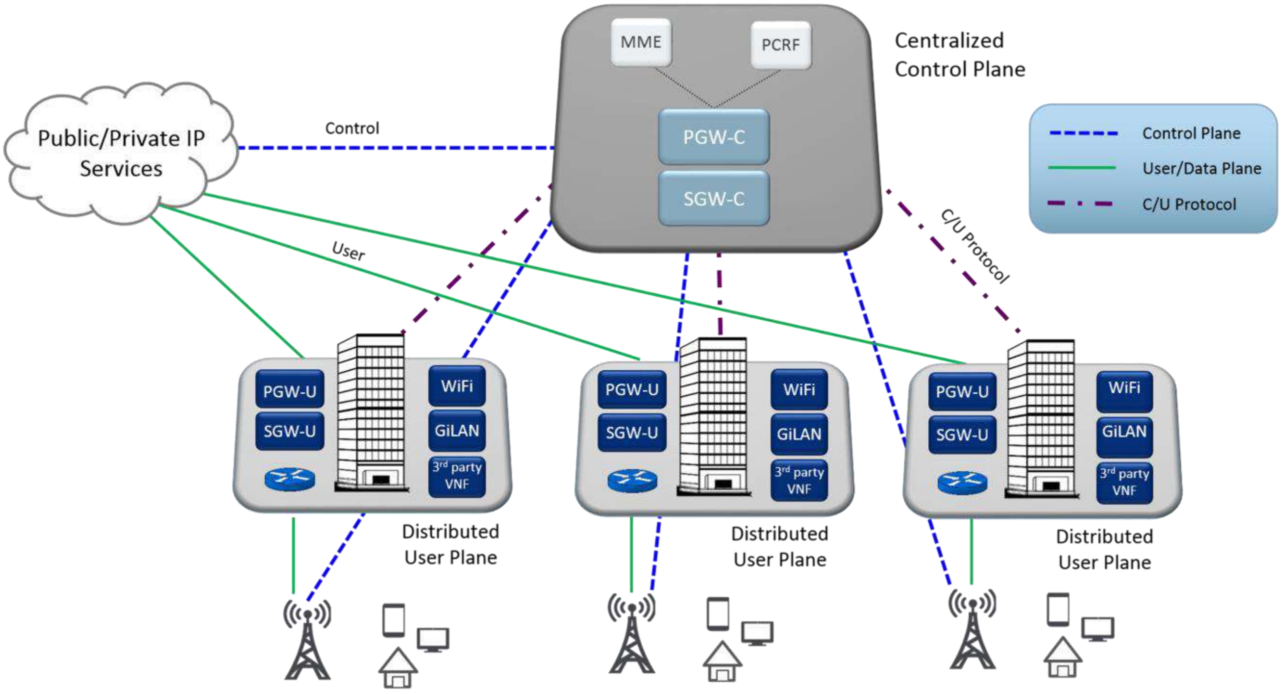
MME
PCRF
PGW-C
SGW-C
Centralized Control Plane
Public/Private IP Services
Control
User
C/U Protocol
Control Plane
User/Data Plane
C/U Protocol
PGW-U
SGW-U
WiFi
GiLAN
3rd party VNF
Distributed User Plane

Source: Affirmed Networks

This model of discrete control- and user-plane functions is expected to be fundamental to 5G and the NG Core. In this sense, CUPS can be viewed as an important part of a 5G-ready investment strategy.

A 5G-Ready Core Is Now a Priority

The industry now has a reasonable view of 5G service requirements, and progressive operators with aggressive deployment timelines are now working to prepare their networks for rapid deployment of 5G when equipment is available. This is driving investment in the critical IP services network needed to connect the edge cloud locations that will run 5G network functions, content and services.

On the mobile core side, development of cloud-native, service-orientated core networks for advanced 4G and 5G networks is underway. Network slicing provides a conceptual bridge between 4G and 5G investment and facilitates a faster insertion of new services into the network. Similarly, CUPS provides a reference for NG Core and the new 5G network architecture. With a 5G-ready technology strategy, operators can prepare for 5G service launch in a way that optimizes their investment in next-generation IP and mobile core platforms over the next three years.

About Affirmed Networks

Affirmed Networks' NFV solution has become the standard for the world's top mobile operators, who are embracing new business models and building new revenue streams by making the transition to virtualized architectures. The company's technology portfolio includes the Affirmed Mobile Content Cloud, the Affirmed Wi-Fi Gateway (serving as a TWAG/TWAP and an ePDG), Affirmed Service Automation Platform (ASAP) and Affirmed Virtual Probe and Analytics Solution. These virtualized solutions have come to represent the present and the future of virtualized mobile networks with extreme scalability, remarkable flexibility, comprehensive network orchestration and future-proof solutions for a 5G-ready architecture. Please find more information at www.affirmednetworks.com.

About Juniper Networks

Juniper Networks (NYSE: JNPR) is in the business of network innovation. From devices to data centers, from consumers to cloud providers, Juniper Networks delivers the software, silicon and systems that transform the experience and economics of networking. The company serves customers and partners worldwide. Additional information can be found at www.juniper.net, or connect with Juniper on Twitter and Facebook.

CONSULTING - SPECIFYING engineer

CFE Edu
CFE Edu is an interactive, multi-media online platform that offers exclusive content covering engineering topics that are critical to your work.
Catapult your career for

PE New Products for Engineers Database
Find An Innovative Product

PE
Searching for a product?

Energy harvesting power for the Internet of Things

Energy harvesting wireless technology easily interconnects thousands of individual devices in a system, opening up unlimited processing and monitoring applications.

Laurent Giai-Miniet, EnOcean, Oberhaching (Munich), Germany

09/30/2013

Energy is everywhere within reach; it just needs to be harvested. This is the principle of energy harvesting. Today, energy harvesting wireless solutions are already well established in the commercial building automation sector. But the technology is just getting started. New application fields for batteryless, wireless communication will be found to further enhance the world around us.

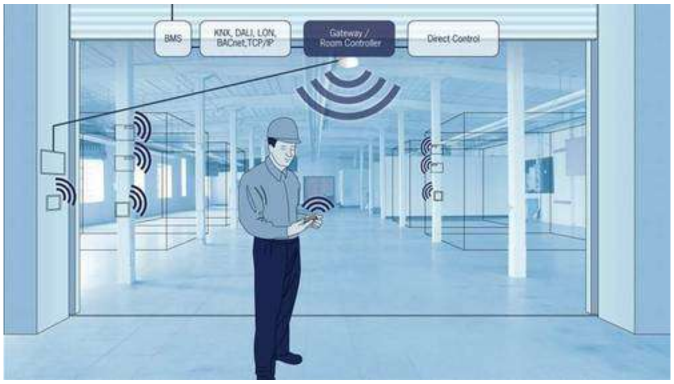

Based on energy harvesting wireless technology, a wide range of energy-autonomous applications are currently available for connected buildings that use motion, light, or temperature differences as their energy source such as batteryless switches; intelligent window handles; temperature, moisture, and light sensors; presence detectors, heating valves; and smart home systems. However, building automation is by no means where energy harvesting wireless ends.

Multiple interconnections

Everybody's talking about the Internet of Things (IoT). But how should billions of communicating devices be powered? The answer is by energy harvesting, and the reason is simple: Liberating sensors from external power, making them energy-autonomous, opens up unlimited processing and monitoring applications where cables or batteries represent an insurmountable hurdle. These features make energy harvesting wireless technology the ideal communication standard to easily and reliably interconnect thousands of individual devices in a system, as well as network them with other wireless protocols.

at AireTechnologies, Inc
HVAC Human Fire Life Safety Devices
Delivered As Early As Tomorrow!
Learn More
JAN 22-24, 2018 AHR EXPO CHICAGO
Visit us at the AHR Expo Booth 5341

Today, energy harvesting wireless technology is widespread, providing M2M solutions in the building automation sector and bridging the control of light, HVAC, and other fields of building technology to smart buildings, smart metering, and energy management systems. This is the starting point to actuate further applications that lead to the IoT in the long term. The following four categories show what this could look like.

Monitoring and control

Wireless and batteryless technology significantly eases energy monitoring and control in buildings with little intervention into the existing systems. The wireless devices are highly flexible to install so that individual components, wall switches, sensors, and relay receivers can be easily networked to form an intelligent system without complex cabling. In addition, dispensing with batteries eliminates the burden of maintaining the devices' energy supply in a regular time period, which can be up to each year.

An example for such a flexible automation system is HVAC control. Here, a thermostat, VAV (variable air volume), or fan coil controller receives information related to occupancy, temperature, humidity, window position, or CO_2 from the respective batteryless sensors and controls the opening and closing of valve actuators for radiators, or dampers for VAV systems. At the same time, the controller sends status information to a central building automation system, and receives control messages from the BAS. This enables the building to be monitored from a central location that can be remote from the building itself, and building-wide settings, such as holiday shutdown, to be implemented. Enormous progress is also being made on the product side, leveraging advancements in energy harvesting. Self-powered radiator valves generate energy from the difference in temperature between the hot water and the surrounding air. This energy powers the communication with a controller or BAS system, and turns the valve itself. Without cables or batteries, these wireless devices are especially easy to install, and they require no maintenance.

In further optimized systems, central equipment such as boilers or air handling units are integrated into the wireless communication system, enabling scalable HVAC generation, visible and controllable over the Internet on a PC, tablet, or

PE
Searching for a product?

smartphone.

Performing tasks

Alarm systems are a second field that batteryless wireless technology is opening up, due to its specific features. Here, the reliability requirements are much more stringent than those required for lighting controls. A system failure not only means a malfunction but also can cause much more serious consequences for other systems that depend upon the equipment being monitored. It's a fact that more malfunctions are caused by battery failures than by the electronics, especially in large systems. Energy harvesting overcomes this issue.

16:56
Auto
enocean

There are already various batteryless wireless water detectors available that use miniaturized solar cells or motion energy converters to power wireless signals that report water leaks in areas such as water supply networks in spacious industrial facilities. In the AFRISO universal module, for example, the EnOcean wireless signal immediately sends the leakage information to a gateway controller or directly to a valve, causing the main water pipeline or the affected supply line to be shut off. A notification is sent to the user's smartphone or smartpad at the same time to inform the user about the incident. In addition, the water valve can be opened and closed, independent of leakage notifications, by GSM connection via smartphone or smartpad.

Embedded processing

A major requirement of today's and the future energy supply is the Smart Grid. It's intended to network centralized and decentralized energy suppliers, including private homes producing electricity by photovoltaic installations, to an intelligent system that provides energy only when needed, updating in real time. This requires continuous data flow and processing from all involved parties, which means from millions of information points.

A key component is smart metering systems. To work reliably and cost-efficiently, interoperability between the meters is supplied by different manufacturers; this is why smart metering calls for standardized technologies. Consequently, the members of the EnOcean Alliance have defined a specific device communication protocol, the Automated Meter Reading (AMR) profile for batteryless wireless devices. Smart meter systems based on this open protocol are already available from a number of manufacturers. For example, Eltako meter components read and transmit the current electricity, water, and gas consumption, including accumulated meter figures, by means of energy harvesting wireless technology located at a variety of points inside a building. BSC software monitors and displays the current meter readings and compares them against default values. This makes all relevant data available for systems processing for intelligent energy management on demand.

Bridge to the cloud

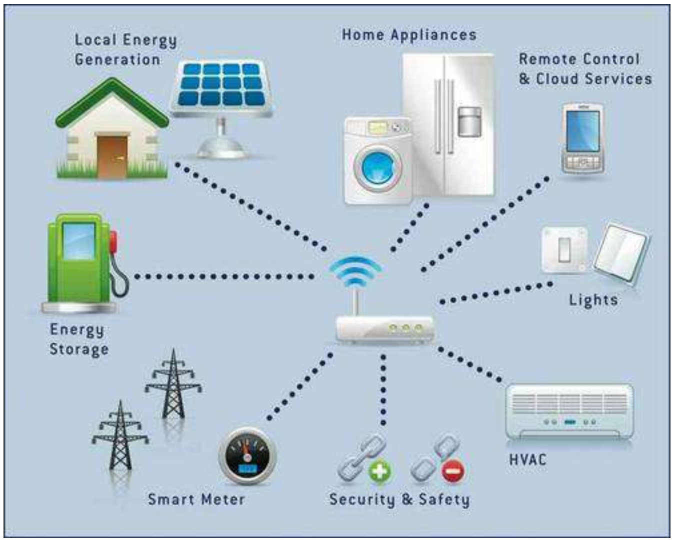

Via similar gateways, the standard-based energy harvesting technology can also communicate with Ethernet, Wi-Fi, GSM/UMTS/CDMA, and other networks for integration in cloud services. Here, all data collected by batteryless wireless sensors is encrypted and transmitted to a cloud service over the Internet. The gateways connected to a control and visualization software by TCP/IP that can be used to control all relay receivers and sensors bidirectionally. Some manufacturers have developed a cloud solution that offers energy management as a service. Therefore, facility managers, building owners, and businesses can monitor important inventory, equipment, assets, and energy-related information from anywhere at any time, via the cloud. Critical building-related data is automatically pushed to the cloud, freeing owners and managers from the often-challenging coordination and expense of hosting on-site servers.

One of the major advantages of such a cloud-based solution is that the management system arrives completely precommissioned from the manufacturer and ongoing device commissioning is expertly done on behalf of the client and pushed out from the cloud. The users are granted unlimited access to their remote, dedicated virtual server with their own IP address, accessible from a desktop or smartphone — the perfect precondition for a deeply connected world of an IoT.

The energy harvesting market is growing and multiplies on a year-by-year basis. Forecasts show that this trend will continue, especially as the next generation of energy harvesting wireless solutions is just around the corner to pave the way to the Internet of Things.

Laurent Giai-Miniet is CEO of energy harvesting wireless solutions provider EnOcean and previously spent 20 years with Texas Instruments (TI), where he held several management positions including General Manager for Low Power RF Products (LPRF).

Related News:
- Energy harvesting for Internet of things - 02.07.2014 01:48
- Consumer market will drive industrial Internet of things for manufacturing - 06.03.2014 11:43
- How the Internet of Things will affect consulting engineers - 02.01.2014 11:35
- IP-all-the-way is the way to go - 23.07.2013 01:28
- Integration: Lighting and HVAC systems - 16.04.2013 01:30

More than
80
photos

The Basic Soldering Guide

Learn to solder successfully!

Alan Winstanley

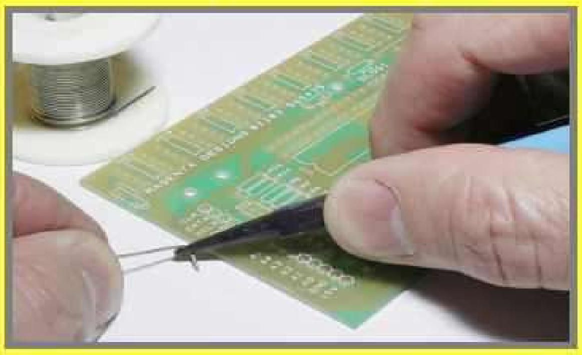

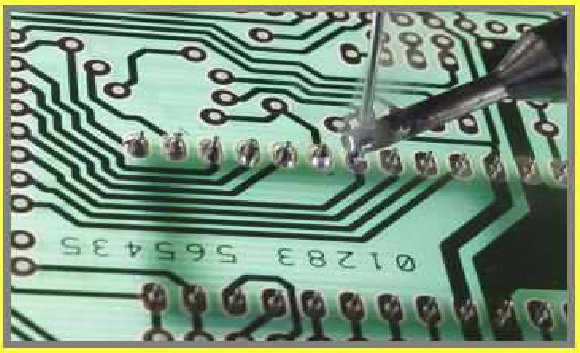

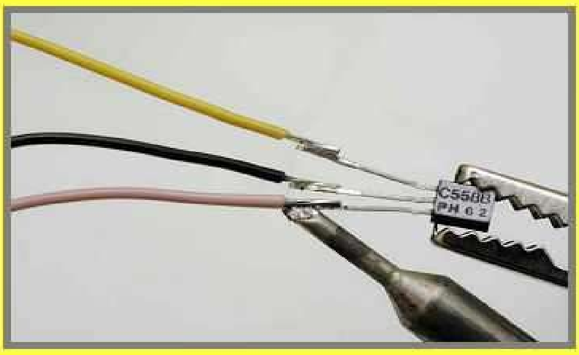

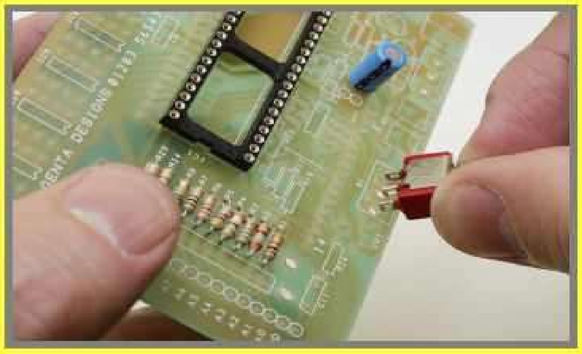

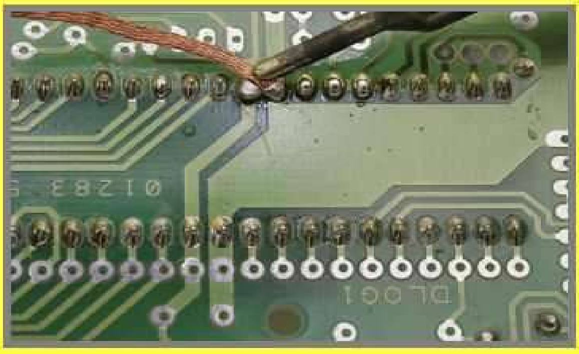

In association with

ANTEX

ANTEX (Electronics) Limited

The Basic Soldering Guide

by Alan Winstanley

EPE Magazine Online Editor

In association with
Antex (Electronics) Ltd.

ANTEX
ANTEX (Electronics) Limited

Copyright Notice and Disclaimer

All text, diagrams and photographs are Copyright © 1996-2013 A. R. Winstanley. All Rights Reserved.

All text, diagrams and photographs are fully protected by copyright and may not be reproduced in any commercial or non-commercial publication or medium, mirrored on any commercial or non-commercial website nor may they be appended to or amended, monetised or used or re-distributed for any commercial reason, without the prior permission of the writer.

Every care has been taken to ensure that the information and guidance given is accurate and reliable, but since conditions of use are beyond the writer's control no legal liability or consequential claims will be accepted for any errors contained herein.

Where mentioned, the U.K. mains voltage supply is 230V a.c. and you should amend ratings for your local conditions.

Preface

In 1996 when the world wide web was still very young, I launched the first and most detailed website ever describing the practical skills of electronic soldering, and my **Basic Soldering Guide** quickly became the #1 web site of its kind in the Google search engine.

Thanks to its in-depth reference text and the unequalled high-quality close-up photography showing soldering step by step, many quickly learned the essential stages needed to make a solder joint successfully. Even novices who had never tried soldering before, gained the skills and confidence needed to acquire this skill.

My Basic Soldering Guide became a key *go-to* online guide for soldering, and I've enjoyed receiving encouraging feedback ever since from the likes of the US Air Force, US Marines, US Coastguard, Honeywell trainers, Atomic Energy authorities, Australian aeronautical suppliers, UK colleges and universities, trainees and many more around the world.

In association with Antex (Electronics) Ltd., the leading UK manufacturer of electronic soldering equipment, I'm delighted to bring you this updated Basic Soldering Guide containing over 80 all-new colour photographs, more background, more detailed information and lots more practical hints and tips.

I've revisited various areas of the guide and refreshed them, taking onboard readers' queries and nearly 17 years of online feedback, experiences and answering reader's questions. With the help of all-new photography, I'm sure you'll master the skills needed to solder electronics successfully using this updated guide.

I welcome feedback and comments, and readers can reach me via my website **www.alanwinstanley.com** or Email me at **alan@epemag.demon.co.uk**

Alan Winstanley

July 2013

Table of Contents

Copyright Notice and Disclaimer

Preface

Introduction

First steps

Soldering iron choices

How to solder

Next steps

Cleanliness and “Tinning” the bit

Getting Ready

Solder and Fluxes

Flux dispensers and Colophony

Temperature flow

Now is the time…

Soldering Step by Step

Interwiring – get hooked up!

Wire joints

Tidying up

Reflow technique

Fatigue & Breakage

Faults & Desoldering techniques

Desolder Braid

Quick Summary Guide

Troubleshooting Guide

Possible Hazards and simple First Aid

Useful resources

Acknowledgements

Photography

Conclusion

Introduction

The first and most important aspect of assembling any electronic project is that of *soldering,* which is a delicate and precise skill that can be mastered with experience. Sometimes called “soft soldering”, there’s no shortcut to acquiring the necessary expertise, and producing a consistently satisfactory solder joint takes a little practice. However, like riding a bicycle, soldering is an art which once learnt is never forgotten, and the purpose of this new and updated guide is to explain the techniques of soldering and desoldering for beginners, which I hope will set the hobbyist or trainee technician firmly on the road to successful electronic assembly or repairs in the future.

Soldering is the least “aggressive” way of joining non-ferrous metals together, and is used universally in electronics, air conditioning and refrigeration circuits, household plumbing and more besides – applications where the precise joining together of components at fairly moderate temperatures is needed. Further up the scale, **brazing** involves using higher temperatures to melt brazing rods onto larger metal parts, perhaps to repair a metal chair, lawnmower or to fabricate metal components or jewellery into intricate shapes. Lastly, **welding** is a very aggressive way of fabrication using welding rods or wire; steel girders, oil rigs and ships are all welded together, or robotic spot-welding is used for the mass production of, say, washing machines or car bodyshells using sheet steel to make strong rigid assemblies.

Due to the lower temperatures used and the need to make consistently good electrically conductive and mechanically sound joints with precision, soldering is used to connect components together when manufacturing electronic circuits. Small components would quickly be destroyed by brazing or welding, although tiny spot-welding joints do appear in electronics, perhaps to weld a metal tag onto a button battery.

This guide therefore deals with the soldering techniques used in electronics at hobbyist or trainee educational level. It explains what to look for before buying a soldering iron, describes ways of making various solder joints on circuit boards and other electronic components, and also how to *desolder* – removing solder in order to repair a circuit board or replace an electronic component.

You’ll also find more details of other aspects of soldering, including an outline of typical solder types and fluxes. In short, everything you need to get started in electronics soldering is here, so let’s get started!

First steps

The principle behind soldering sounds quite simple: the idea is to join components together to form an electrical connection, by using a mixture of lead and tin solder or alternatively “lead-free” solder (an alloy of tin and copper), which is melted onto the joint using a *soldering iron*. If you have never picked up a soldering iron before, then this guide will show you everything to help you start soldering with confidence. I also hope that the guidance will help those working in other areas — computer technicians or audio enthusiasts, for example, who may be faced with electronic repairs or modifications using a soldering iron for the first time.

If you’re an electronics hobbyist or trainee, before embarking on any form of ambitious electronic project, it is recommended that you practice your soldering technique on some *brand new* components using *clean* strip board (or protoboard) or a printed circuit board, and select a simple and straightforward constructional design as a starting point. Become acquainted and comfortable with your chosen soldering iron, which likely to become as familiar to you as a favourite pen. Learn how to balance it and handle it with precision. Try soldering an assortment of resistors, capacitors, diodes, transistors and integrated circuits with it, and then try your hand at *desoldering* – removing the solder again to make a repair or modification.

A really good place to start learning is by building a simple electronics kit, such as a Velleman kit that contains a good quality printed circuit board. You’ll learn some of the basic skills of successful soldering this way, and it’ll be a great confidence booster too.

Did you know? In the USA and elsewhere, the letter L in “solder” is silent and they say “soda” or “sodder” – but here in Britain we do pronounce the L and we say “sole-der”.

Soldering iron choices

Search any electronics catalogue or website and you'll see a bewildering array of soldering equipment on sale, including irons, controllers, work stations and desoldering equipment too. A large range of soldering irons is readily obtainable - which one is suitable for you depends on your budget and how serious your interest in electronics is, but there's something for every pocket distributed by a variety of retail, industrial and mail-order outlets.

The Antex range of soldering equipment has been very popular with industry, education and the electronics hobbyist for 60 years; I grew up with an Antex iron and a trusty 15W Antex iron was an everyday part of my hobby electronics all the way through the 1980's. An industrial user or a more dedicated hobbyist – with a bigger budget! – will be interested in a soldering station instead and again Antex offers a range of British-made products for industry or home use.

A very basic mains electric soldering iron can cost from under £5 (US$ 8), but I find that these very cheap irons, as sold on auction websites, are pretty crude and imprecise. They are best suited for simple electrical repairs and DIY rather than precision electronics or printed circuit boards discussed here. They tend to be bulky and uncomfortable for extended use, and they may not have suitable "bits" or tips of various sizes to suit different tasks.

^^ This classic Antex XS25 "pencil style" 25W mains-powered soldering iron has exchangeable tips or "bits" and is great for general hobbyist or educational use. Stands are also available to store them safely in between use.

A quality pencil-style electric soldering iron such as the Antex XS25 (photo) will be approximately £20 (US$18 tax free) - though it's possible to spend into three figures on a soldering iron “station” if you’re really serious about the subject! Don’t be tempted to over-spend on an elaborate workstation though, unless you are really very serious about becoming involved in electronics. You will usually obtain perfectly satisfactory results using a fairly modest “pencil” iron model, and you can upgrade to something more sophisticated should your needs change in the future.

^^ *For enthusiasts or industry, this Antex 660TC soldering station has a separate mains-powered control unit and a matching low-voltage soldering iron rated at 24 Volts, 50 Watts so it’s suited to a wider range of tasks than a lower powered one.*

When choosing your soldering iron, certain factors to bear in mind include:

Voltage: for the British market, “mains” electric irons run directly from the mains at 230V a.c. or will be set for other voltages (110V a.c.) depending on the country. However, **low voltage types** (e.g. 12V or 24V) usually form part of a “soldering station” for use with a matching controller made by the same manufacturer. Some low-voltage irons run off batteries (e.g. a car battery or Ni-cads) but these are uncommon.

Wattage: this is an extremely important factor to think about when choosing your iron. Typically, irons for general electronics work may have a power rating of between 15-25 watts or so, which is fine for most electronic assembly tasks, printed circuit boards and inter-wiring. It’s important to note that a higher wattage does *not* mean that the iron runs hotter - it simply means that there is more power “in reserve” for coping with larger joints. The maximum electric iron wattage generally available is about 100W, which is OK for DIY electrical repairs but is far too high for general electronics or circuit board use.

A higher wattage iron offers you more flexibility for tackling a wider range of tasks. It has a better “recovery rate” which makes it more “unstoppable” when it comes to heavier-duty work, because it won't be drained of its heat so quickly. So check the power ratings carefully, and anything between 15–40W is fine for general electronics soldering.

Temperature Control - the simplest and cheapest types don't have any form of temperature "regulation". Simply plug them into the mains and switch them on! Thermal regulation is "designed in" (by physics, not electronics!). Sometimes they are described as "thermally balanced" as they have some degree of temperature "matching" – in other words, they warm up as quickly as they lose heat during use, so in a primitive way they maintain roughly a constant temperature. This type of iron is perfectly acceptable for hobby or less demanding professional use. It's essential to use the manufacturer's specified tips (see later) to maintain proper temperature matching, otherwise the iron may not heat up enough – or it may overshoot in temperature.

These unregulated irons form an ideal general-purpose iron for most users, and they cope reasonably well with printed circuit board soldering and general interwiring. However, most of these "miniature" types of iron will be of little use when attempting to solder large joints (e.g. very large terminals or very thick copper wires) because the components being soldered will draw or "sink" heat away from the tip of the iron, cooling it down too much and preventing solder from flowing properly. That's when a higher wattage iron is needed.

A proper **temperature-controlled iron** will be quite a lot more expensive - retailing at say £40 (US$ 60) or more - and will have some form of built-in thermostatic control, to ensure that the temperature of the "bit" (the tip of the iron) is maintained at a fixed level within reasonable limits. This is desirable during frequent use, since it helps to ensure that the temperature will be relatively stable regardless of the workload. Some irons have a bimetallic strip thermostat built into the handle which gives an audible "click" in use, and some may include an adjustable screwdriver control within the handle as well. Others may have electronic controls built in.

^^ An Antex 660TC soldering station with matching iron and bench stand.

More expensive still, **soldering stations** cost from £70 (US$ 115) upwards (the iron may be sold separately, so you can pick the type you prefer), and consist of a complete bench-top control unit into which a special *low-voltage* soldering iron is plugged. Some versions might have a built-in digital temperature readout, and a control knob to vary the setting. The temperature could be boosted for soldering larger joints, for example, or for using higher melting-point solders (e.g. lead-free or silver solder). These are designed for the discerning user, or for continuous production line or professional use. A thermocouple will be built into the tip or shaft, which monitors temperature.

The best soldering stations have irons which are well balanced, with comfort-grip handles which remain cool all day, and silicone-based cables which are burn proof. Antex produces a range of irons with silicone cables specially for education use, to help avoid accidents caused by careless use by students.

Anti-static protection: if you need to solder a lot of static-sensitive parts (e.g. CMOS chips or MOSFET transistors), more advanced and expensive soldering iron stations use static-dissipative materials in their construction to avoid static charges accumulating on the iron itself, which could otherwise damage or “zap” some semiconductors. Such irons are listed as “ESD safe” (electro-static discharge proof). The cheapest irons are not classed as ESD-safe but they still perform well enough in most hobby or educational applications provided you take the usual anti-static precautions when handling these components. The iron would need to be well earthed (grounded) in these circumstances, to carry away any static.

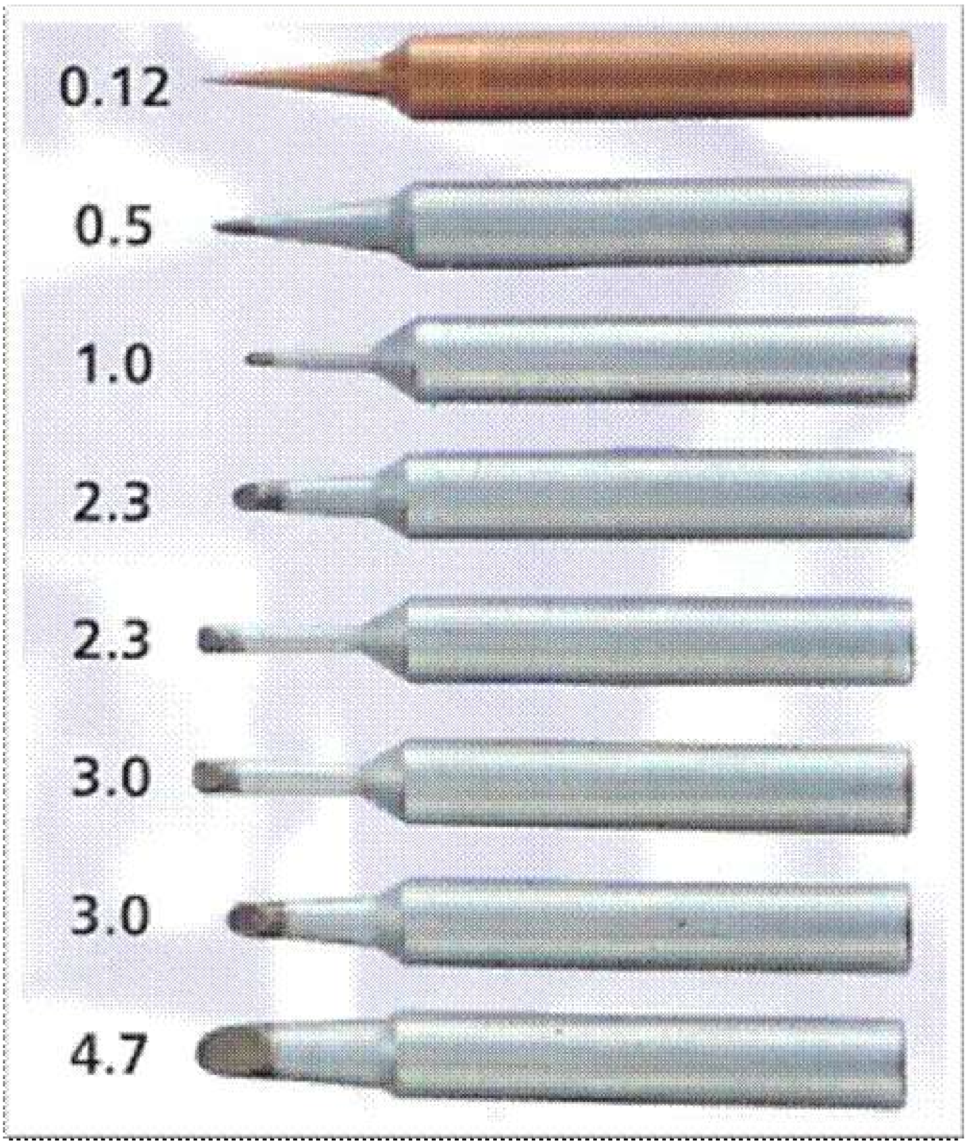

^^ Just part of the range of spare tips or "bits" produced by Antex for their soldering irons. Use only tips intended for a particular iron, to avoid thermal matching problems.

Tips or Bits: it's often useful to have a small selection of manufacturer's bits (soldering iron tips) available with different diameters or shapes, which can be changed depending on the type of work in hand. You will probably find that you become accustomed to, and work best with, one particular shape of tip for the majority of your work. Usually, tips are iron-coated or plated to preserve their life and to maintain good tip "hygiene". Be sure only to use tips that are specifically designed for your iron, otherwise thermal problems may arise. I show separately some typical tips, courtesy of Antex.

^^ Useful accessories from Antex including solder, tips and tip cleaner, desoldering braid and a heatshunt.

Spare parts: it's always reassuring to know that spares are available in the future if required, so if the element blows, you don't need to replace the entire iron. This is especially the case with expensive irons. Check some websites or mail-order catalogues to see whether spare parts are listed. One drawback is that you may need another soldering iron when exchanging a broken heating element!

ANTEX Gascat 40

^^ Gascat 40 gas-powered soldering iron kit by Antex, with spare tips and cleaning sponge.

Gas or electric? So far I've discussed electric soldering irons, but **gas-powered** soldering irons are sold which use butane propellant rather than mains electricity to operate. They have a flint lighter or (better still) a built-in piezo for ignition, and have a catalytic element which, once warmed up, continues to glow hot when gas passes over them. They tend to be big and bulky compared to electric pencil irons.

Field service engineers use gas-powered irons for working on repairs where there may be no power available, or where a joint is tricky to reach with a normal electric iron, so they are really for occasional “on the spot” use, rather than for mainstream construction or assembly work. I use one when I just need to do a quick repair and can’t be bothered getting the electric soldering iron going!

Gas irons can have higher power equivalents than electric ones (eg 125 watts or more) but some gas-powered irons are nothing more than miniaturised blowtorches, which may or may not be useful for occasional heavier duty soldering. In the author’s experience they can be difficult to use in confined areas. *Extreme care is needed at all times to ensure hot gas emitted from the iron’s exhaust port doesn’t damage nearby materials, plastics or wiring.* Gas irons can have useful accessories to convert them into e.g. a hot knife for sealing nylon rope, or a hot-air blower for use with heatshrink tubing. Almost every electronics constructor uses an electric-powered iron though.

A **solder gun** is a pistol-shaped iron, typically running at 100W or more, and is completely unsuitable for soldering modern electronic components: they’re too hot, heavy and unwieldy for micro-electronics use, nor are they designed for that. Plumbing or DIY, maybe..!

^^ A heat resistant soldering iron stand with cleaning sponge. (Antex)

Soldering irons are best used along with a **heat-resistant bench stand**, where the hot iron can be safely stored in between use (photo). It is extremely important that a hot soldering iron is safely “parked” ready for action, and a bench stand is really a necessity. Soldering stations usually have such a feature, otherwise a separate soldering iron stand is essential, ideally one that’s supplied with a tip-cleaning sponge. You can make your own cleaning sponges using *cellulose* sponge only.

^^ *A benchtop fume extractor fan for hobbyist use. A replaceable carbon filter helps remove particles and air is vented out the back.*

Other equipment worth considering includes the use of **fume extractors**, which are compulsory in the industrial workplace. Solder fumes and flux smoke are not known to be toxic but they can cause irritation. A basic fume extractor (photo) consists of a small bench-top fan which draw fumes and irritating smoke away from the operator’s face and filters out some of the smoke particles, before exhausting the air back out through the fan. The carbon-impregnated foam filters are replaceable. Such devices are quite effective and users soon find them indispensable, even though they can be a bit noisy at close range.

Professional fume extraction systems draw the smoke and fumes directly from the work area via a clip-on tube fitted to the soldering iron, then vent the fumes away through a large filter pump. It is definitely worth considering a small bench top unit for regular hobby or occasional professional use, as having decent ventilation can only be a good thing.

^^ *A soldering accessory toolkit by OK Industries, including probes, scrapers and wire brush*

A variety of specialist **hand tools** are available that assist with soldering, and a good supplier's catalogue will offer a range of small brushes, scrapers and cleaning tools in a handy kit, together with the usual types of wire cutters, pliers and so forth, which are necessary for handling components and tidying up as required. Some specialist service aids (aerosols etc.) are also described later.

Now let's look at how to use soldering irons properly, and later on we will describe the techniques for putting things right when a joint somehow goes wrong — and don't worry, even the experts get it wrong sometimes!

How to solder

This guide will show you step by step how to solder successfully and plenty of photographs are provided to help explain the techniques needed.As you read through the guide I'll explain all the stages in more detail, but let's look at the basics first.

First of all, successful soldering requires that the items being soldered together are held with as little movement as possible. So it's best to *secure the work* as needed, so that your accuracy isn't affected should the workpiece move accidentally.

In the case of a printed circuit board, various holding frames are useful especially when densely populated boards are being soldered: the idea is to insert all the parts on one side (a process called "stuffing the board"), hold them in place with a suitable pad to prevent them falling out again, turn the board over and then snip off the wires with cutters before soldering the joints. The frame saves an awful lot of turning the board over and back again, especially with large boards: all the soldering can be performed in one pass.

^^ The ever-popular "Helping Hands" (left) helps support sundry parts, wires etc. during soldering.
A modeller's vice (right) holds parts firmly. A vacuum base fixes it onto smooth surfaces.

Only the more serious constructor will purchase a holding frame, and hobbyists can retain parts in place by improvising in a variety of ways – the ever popular "Helping Hands" stands cost a few pounds and is widely sold. They have adjustable crocodile clips to grip parts, and maybe a magnifying glass or soldering iron stand as well. The cast iron base provides stability. Other parts could be held firm in a modeller's small vice, for example.

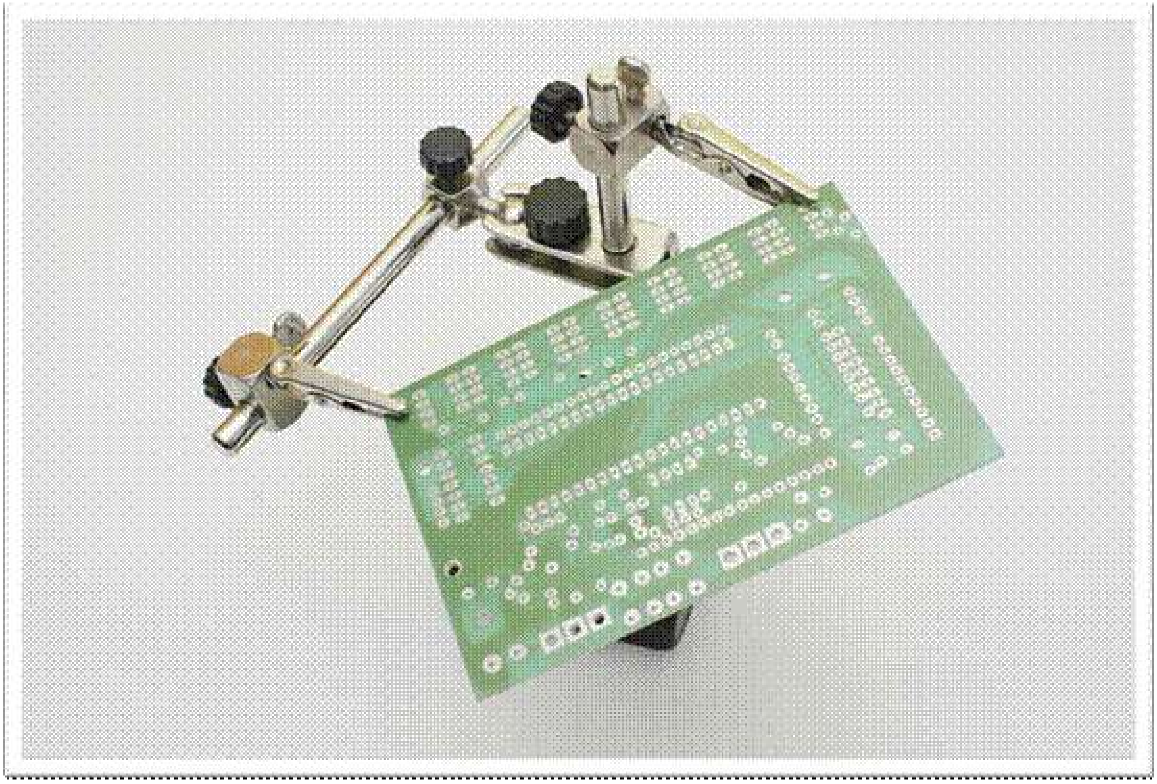

^^ "Helping Hands" uses crocodile clips to grip parts during soldering. Or just place parts flat on the bench.

When soldering parts onto an ordinary circuit board, components' wires can simply be bent to the correct pitch (distance apart) to fit through the board, insert the part flush down against the board's surface, splay the wires a little so that the component grips the board under spring tension, and then solder it. This technique is universally used in manual **through-hole soldering,** which is explained in full later.

In the author's view - opinions vary – it's generally better to snip off the surplus wires leads first, to make the joint and neighbouring joints more accessible and also to reduce the mechanical shock transmitted to the p.c.b. copper foil. However, in the case of diodes and transistors the author tends to leave the snipping until after the joint has been made, since the excess wire will help to "sink" heat away from the heat-sensitive semiconductor junction.

A special clip-on heatsink is available which also helps stop excess heat from reaching temperature-sensitive semiconductors like these. I've always managed without one but beginners might find them reassuring. Integrated circuits can either be soldered directly into place if you are confident enough, or better, use a dual-in-line socket to prevent heat damage. The chip can then be swapped out at a later date if needed.

Parts which become hot in operation (e.g. some power resistors), should be raised above the board slightly to allow air to circulate. Some components, especially large electrolytic capacitors, may require a mounting clip to be screwed down to the board first, otherwise the part may eventually break off due to vibration. It's a good idea to bolt such components firmly into place before soldering their terminals, in order not to strain the soldered joints or the components when fasteners are tightened.

By securing the workpiece as much as possible to prevent movement, you have a much higher chance of producing good-quality reliable solder joints.

Next steps

Let’s get to grips with actual soldering techniques in more detail. The soldering of electronics components utilises **lead/ tin** or **lead-free solder** and the process is compatible with many non-ferrous metals. You can solder copper, lead, brass, gold plate, silver, nickel, tin and tin plate, zinc and more besides but some metals such as nichrome, galvanised or stainless steel require a highly specialist “flux” (see later) to solder them and aren’t discussed here. Some materials such as beryllium, chromium, magnesium and titanium are non-solderable in any case, according to solder manufacturers Multicore.

In electronics we’re mainly concerned with soldering parts or wires onto printed circuit boards or terminals that are usually already “tinned” with solder or plated, ready for soldering with flux-cored solder. The key factors affecting the quality of a solder joint are:

- **Cleanliness** – dirt or impurities drastically hinder good solder coverage.
- **Temperature** – the right level to enable the solder to flow freely!
- **Time** – apply heat for just the right amount of time!
- **Adequate solder coverage** – enough to form a good joint without touching neighbouring areas.

A little effort spent now in soldering the perfect joint may save you — or somebody else —a considerable amount of time in troubleshooting a defective joint in the future. Let’s discuss the basic principles outlined above in more depth.

Cleanliness and “Tinning” the bit

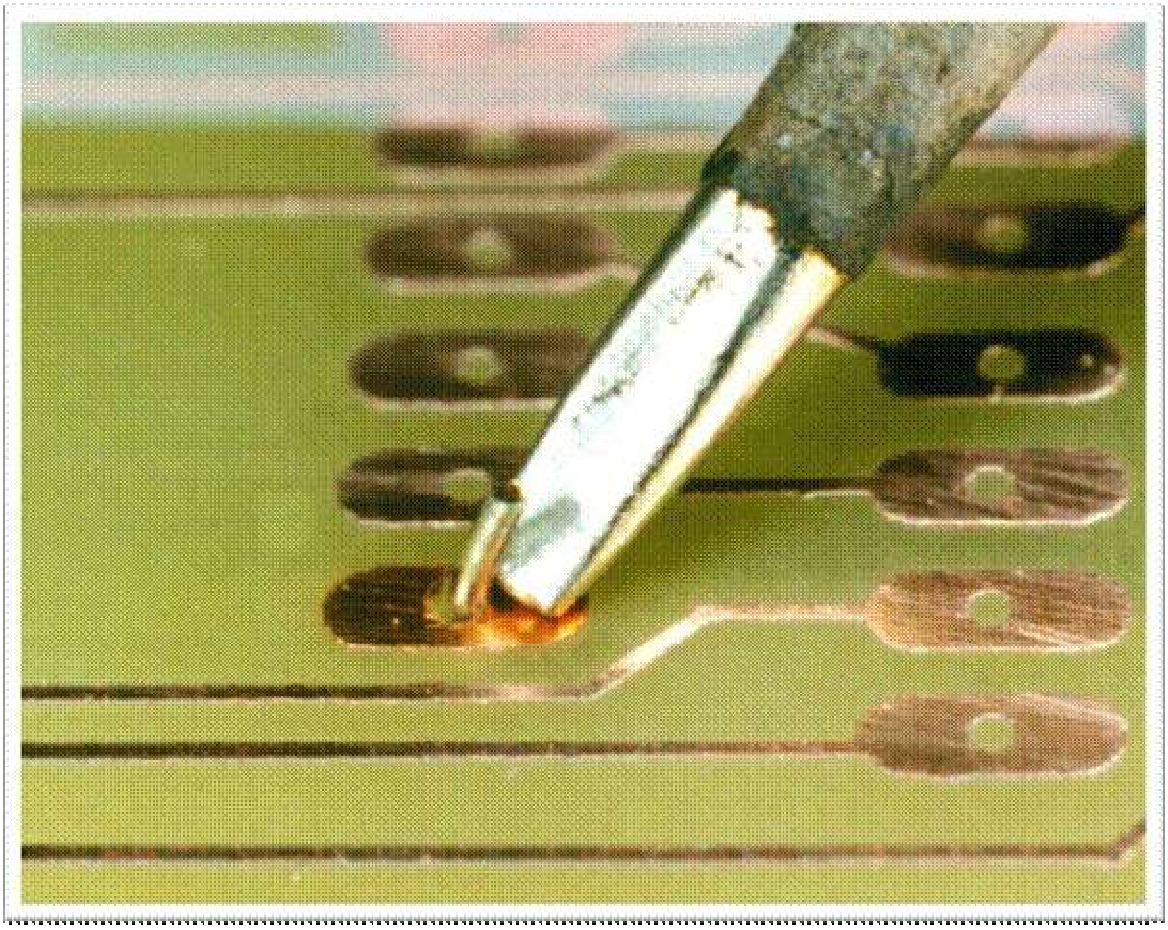

^^ A clean and shiny soldering tip or “bit” is essential for each and every solder joint you make.

Firstly, and without exception, all parts - including the soldering iron’s tip itself - must be clean and free from contamination. Fact is, solder just will not “take” to dirty parts! Old components or copper board can be impossible to solder because of oxidation that builds up on the surface of the leads. This repels the molten solder and will soon become evident because the solder will “bead” into shiny globules looking like mercury, going everywhere except where you need it. *Dirt and contamination are the enemies of a good quality soldered joint!*

When all the conditions are right for soldering, materials are said to be “wettable” and will accept molten solder, which should flow readily over their surfaces. Hence, it’s **really necessary** to ensure that parts are free from grease, oxidation and other contaminants. Note that in any case, some incompatible materials or surface finishes just cannot be soldered using ordinary lead-free solder, no matter how hard you try – e.g. aluminium parts would require special aluminium solder and fluxes (see later) to be used.

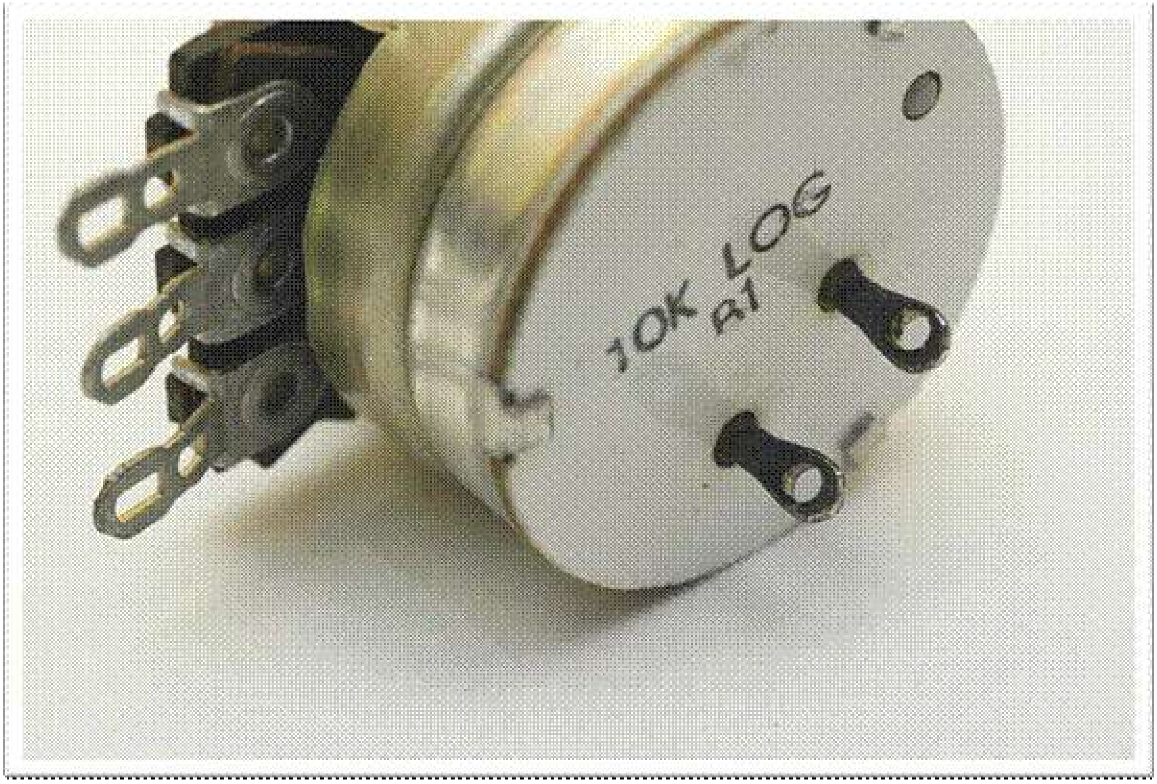

^^ The two tags on this switched potentiometer have blackened with age and oxidation. They must be cleaned before they can be soldered…

In the case of old components that have been stored a long time, where the leads have blackened and oxidised, use a small hand-held file or perhaps fine emery cloth to reveal fresh metal underneath. Stripboard and copper printed circuit board will generally oxidise after a few months, especially where it has tarnished due to fingerprints, so the copper strips could be cleaned using e.g. an abrasive rubber block made for the purpose. Also available is a fibre-glass filament brush (photo), which is used propelling-pencil-like to remove contamination. They're OK for general cleaning but best avoided on fine surfaces (e.g. gold plated switch contacts). These produce tiny particles which can irritate the skin, so avoid accidental contact with any debris.

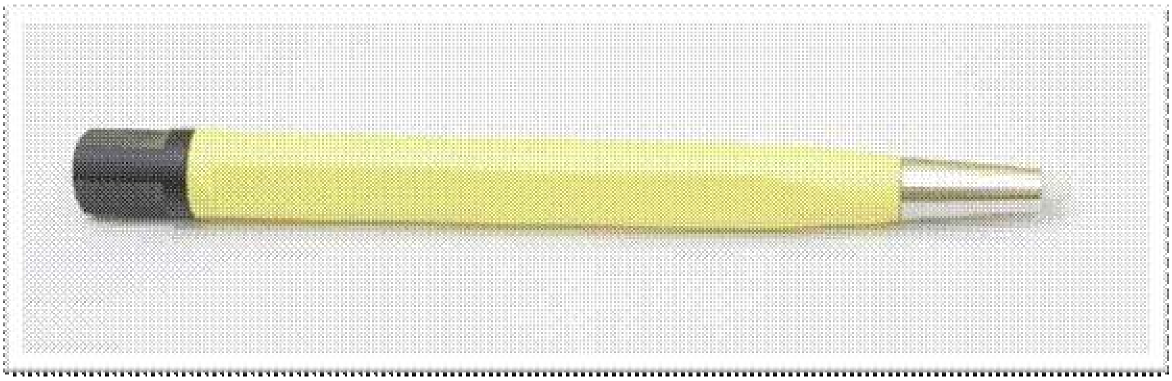

^^ A glass-fibre filament brush like this is useful for cleaning oxidised parts.

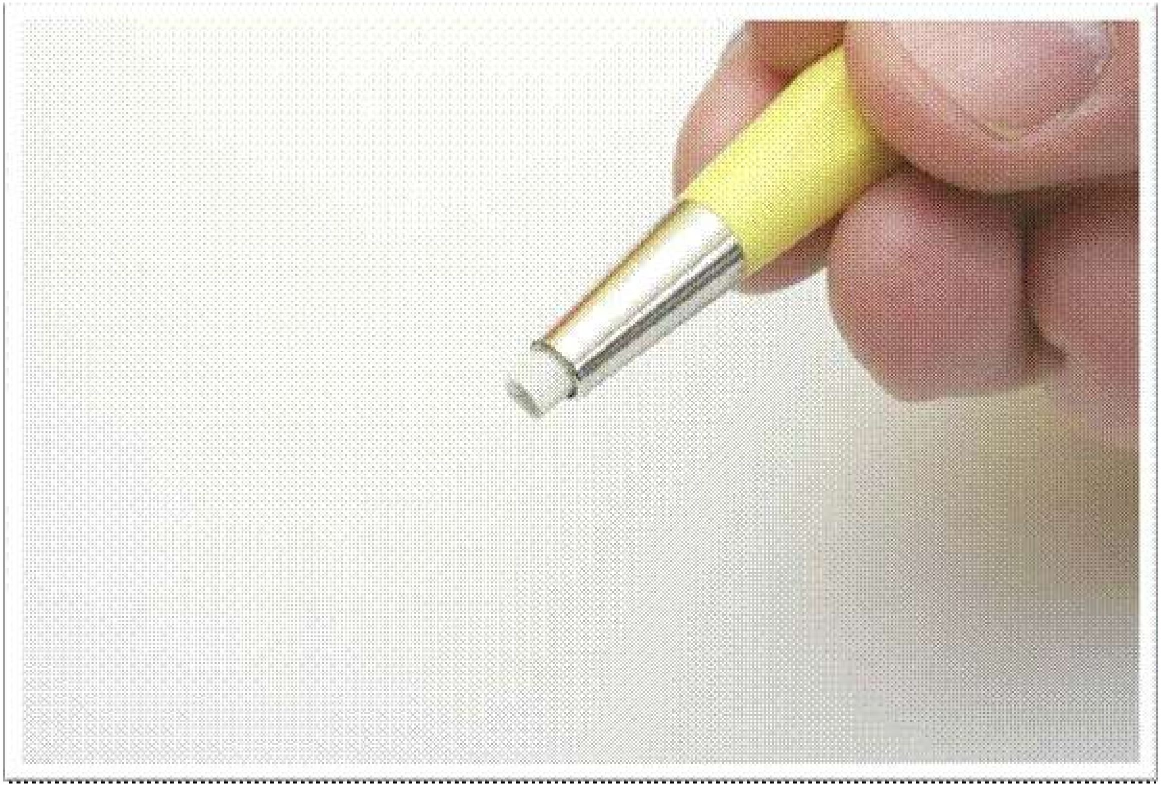

^^ The glass-fibre brush works like a propelling pencil and produces irritating dust.

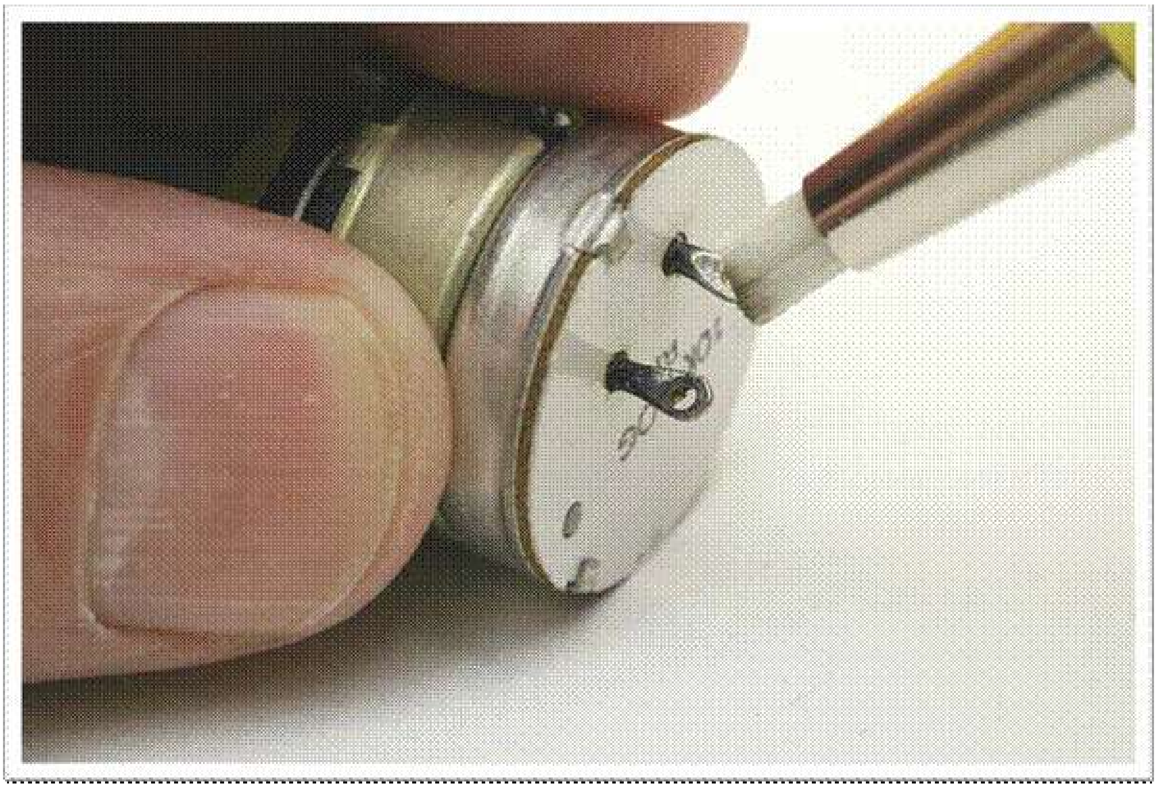

^^ Cleaning oxidised tags with a glass fibre brush to make them nice and shiny, ready for soldering

Afterwards, a wipe with a rag soaked in cleaning solvent will remove most grease marks and fingerprints. After preparing the surfaces, avoid touching any parts as far as possible.

Another side effect of trying to solder contaminated or incompatible materials is the tendency to apply more heat and "force the solder to take". As the materials aren't wettable the molten solder won’t flow where you want it to flow. This can do more harm than good, because it may be impossible to burn off any contaminants anyway, and the component or the printed circuit board may be overheated and damaged in the process. Semiconductors may be harmed by applying excessive heat for more than a few seconds, and extreme heat applied to printed circuit board tracks can cause irreparable damage, because the tracks will be lifted away from the substrate especially on a delicate or badly designed board. You can avoid trouble by ensuring the surfaces to be soldered are clean and wettable to begin with.

Getting Ready

As already explained, cleanliness in soldering is a major factor in obtaining successful results. A soldering iron stand usually has a sponge that is dampened – it wants to be quite damp but not running wet, so squeeze out any excess.

^^ Dampen the sponge but don't soak it unduly. Note the hole, which gives an edge to wipe tip with.

^^ An Antex soldering iron tip. It simply slides onto the iron's heating element. Only the extreme end is wettable and will accept solder: the tip should be kept nice and shiny.

Soldering iron bits are typically iron-plated to resist wear, then chrome-plated to prevent molten solder being deposited on them, with the extreme tip being "wettable" to work with molten solder. Before using the iron to make a joint, the hot tip must be "tinned" with a few millimetres of solder*: you should always flood a brand new tip with plenty of solder to tin it immediately, when using it for the first time*.

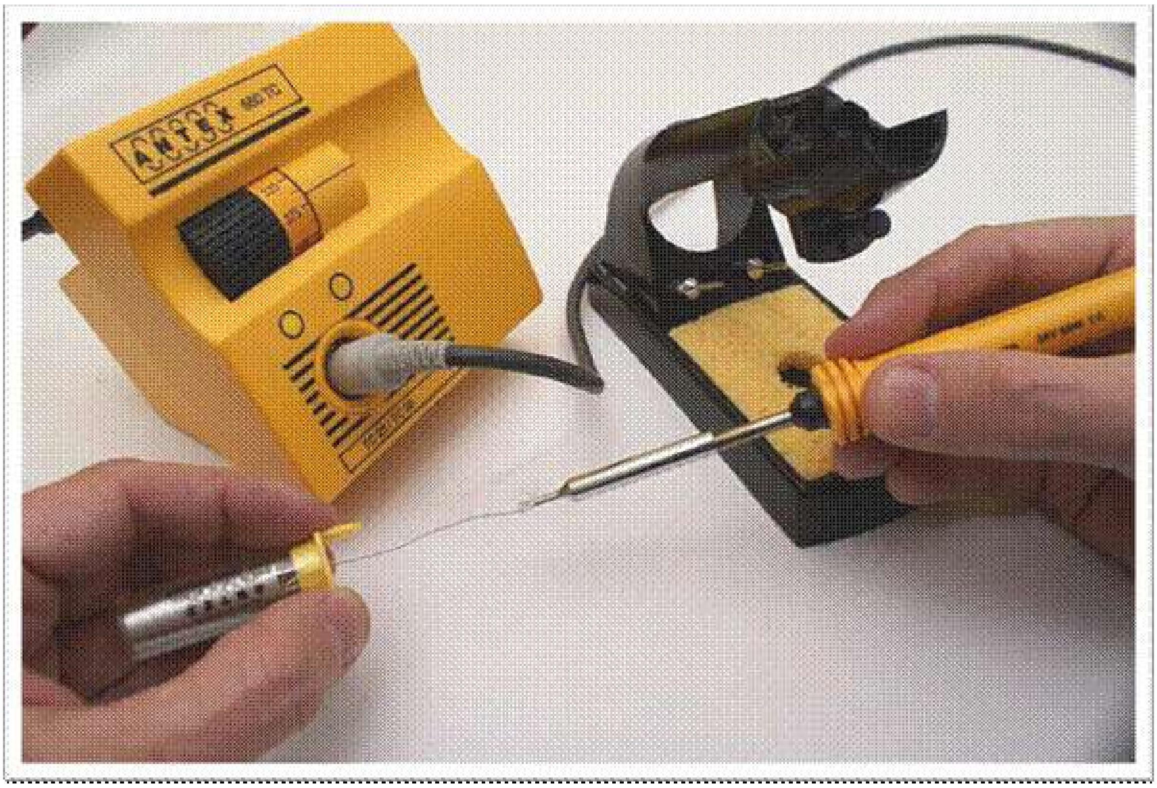

^^ Tinning the tip ready for use: apply a few millimetres of solder to the hot tip.

Wipe off excess solder using a damp sponge and it's ready to use. That's why sponges have a hole or well in them – the edge acts as a wiper and the hole catches excess solder.

A useful tip: after tinning the bit, just before using the iron it helps to re-apply a small amount of solder to a clean tip, to improve the thermal contact between the tip and the joint. The molten solder fills the void between the materials and the iron tip, to help transfer heat better so that the solder flows more quickly and easily.

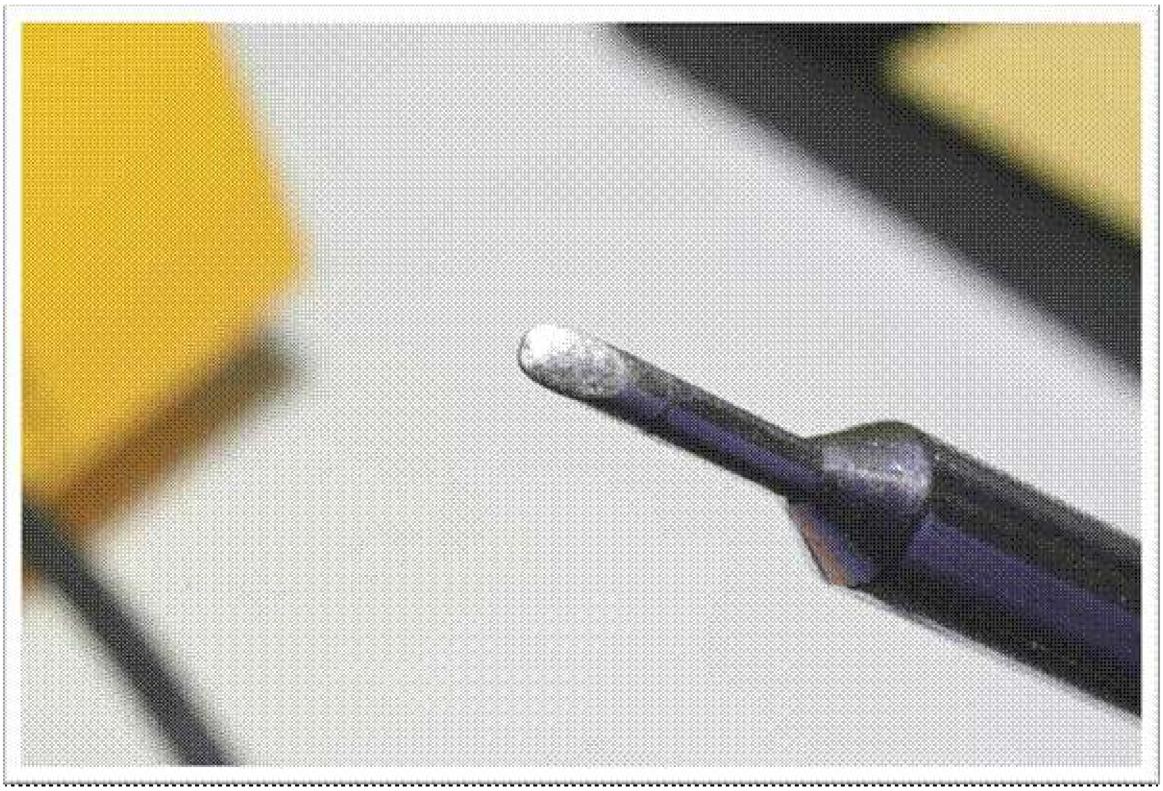

^^ *A tinned tip should look like this, ready to make the next joint.*

It’s sometimes better to tin larger parts as well before making the joint itself, but this is not generally necessary with p.c.b. work. Small tinlets of special paste are available onto which you dab a hot iron - the product cleans and tins the iron ready for use. It’s useful in extreme cases where the bit has some stubborn contamination.

I find special Tip Tinner &Cleaner products to be very useful: gently abrasive in action, they help to clean dirty bits and keep them in good condition. Use them for removing stubborn contaminants, but don’t overdo it.

Multicore
Tip Tinner/Cleaner
TTC1
15g / 0.5oz
Made in Malaysia

^^ Tip Tinner and Cleaner is a useful aid to maintaining a soldering iron bit …

^^ Just press the hot iron onto the solid paste and scrub it around a little. The tip will be cleaned, tinned, and made ready for use.

The move to lead-free solders (see next chapter) has had some effect on the life of soldering iron bits, with increased wear and corrosion noted due to the higher temperatures and the fluxes found in tin-based solders. You can therefore expect bits to wear out over time. Once the iron-plating is damaged due to oxidation or erosion, the bit is fast approaching its end of life. Never use an abrasive or file to sand down a tip: the iron-coating will be damaged and the iron's core exposed, so the tip will soon be made useless due to erosion.

Having prepared the soldering iron tip ready for use, in the next chapter solder and fluxes are discussed.

Solder and Fluxes

In recent years there’s been a move towards using more environmentally-friendly materials in electronic products. EU legislation such as *Restriction of the Use of Certain Hazardous Substances in Electrical and Electronic Equipment* (RoHS) aims to reduce toxic heavy metals being sent to landfill. (Look for the RoHS symbol on equipment to indicate compliance.) Due to RoHS compliance, the electronics industry had to change the type of solder it uses in electronic production.

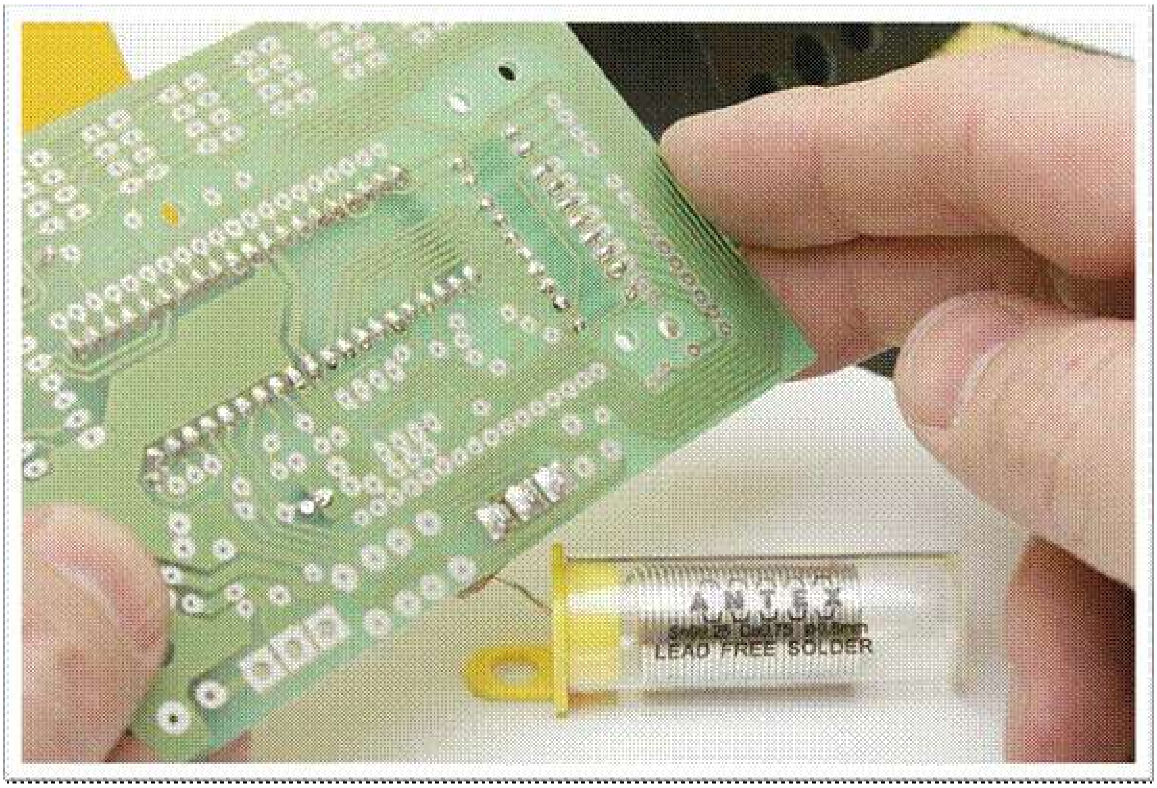

^^ *Electronics solder is supplied in reels or handy dispensers like this.*

Solder comes in various forms including solid bars or pellets for melting in small electric ‘solder pots’ used for treating the ends of wires with solder. Traditional general-purpose electronics-grade solder is in wire form – starting with so-called “60/40” which contains 60% tin (symbol Sn) and 40% lead (symbol Pb) and is sold in handy dispensers or reels. Although tin-lead solder is now banned in industry, there’s nothing to stop the hobbyist from using it but best practice is to use lead-free solder in our work: my advice is to try both, and see which you prefer to work with. “40/60” tin-lead produces lower quality results but is slightly cheaper and perfectly acceptable in hobby circles.

Various diameters of solder wire are marketed. In the UK they’re sold in Standard Width Gauge (SWG) sizes, typically as 18SWG (1.2mm) or 22SWG (0.7mm). The latter is fine for almost all hand-soldering of printed circuit boards or general electronics. For larger solder joints (e.g. larger switch or motor terminals), 18SWG solder would be better as more solder can be dispensed more quickly.

Lead-free solder is universally available and contains typically 99.7% pure tin and 0.3% copper (symbol Cu). It needs a higher melting point which makes it slightly more difficult to work with, but standard soldering irons will cope with it well. Antex lead-free solder (Sn 99.25 / Cu 0.75) is a good compromise at 0.8mm diameter and is sold in small dispensers.

Other solders are produced for specialist work, including aluminium solder (Alu-Sol®) and another solder variant used by professionals is Multicore “Smart” wire which contains a small amount of pure silver (symbol Ag). It produces very clean results and is often associated with SMD (surface mount devices), though some engineers also use it for routine p.c.b. work for producing the best possible finish by hand. As “Smart” wire contains lead it is not RoHS compliant.

An interesting variant is **Eutectic** solder, which is 63/37 Tin/Lead. It goes instantly from solid to liquid when melted and is particularly good for hand-soldering. An almost-equivalent lead-free product would be Stannol Flowtin TC or TSC solder.

The low-down on fluxes

When melting solder with a soldering iron, oxides of metal are produced as a result of the high temperatures involved. Unfortunately, these oxides contaminate the metal surfaces being soldered, which interferes with the flow of molten metal and the production of a good quality solder joint.

All electronics-grade solder wire therefore contains an additive called "flux" which helps the molten solder to flow more easily over the joint. It does this by scrubbing away the oxides which arise naturally during heating, and it will often be seen as a pungent brown fluid bubbling away on the joint, accompanied by some fumes.

Those coming into electronics from other industries should note that flux is already contained within "cored" electronics solder and on no account should any acidic flux be applied separately before using the soldering iron. Plumbers, for example, apply flux paste to copper pipes before soldering them, but electronics-grade solder wire already contains a flux and extra flux is almost never needed. Electronics is no place for acid fluxes!

^^ A close-up of electronics-grade multi-cored solder. Five cores of rosin flux can be seen running through it.

For almost all electronic hand assembly, solder wire containing "Rosin flux" is used. Cores of flux run through the solder wire like letters running through seaside candy, and they prevent the hot area from being contaminated by oxides, otherwise solder would never flow properly and the result would be an incomplete and unreliable joint.

Flux dispensers and Colophony

Flux dispenser pens are sold that allows special liquid flux to be applied separately onto a work area. These might be handy for difficult or challenging jobs to help solder to flow better: adding more flux this way won't do any harm and may help a solder joint to be made more quickly and reliably. In my hobby electronics, there's hardly ever been a time when I felt the need to apply extra flux but it's useful for some very tricky or demanding jobs.

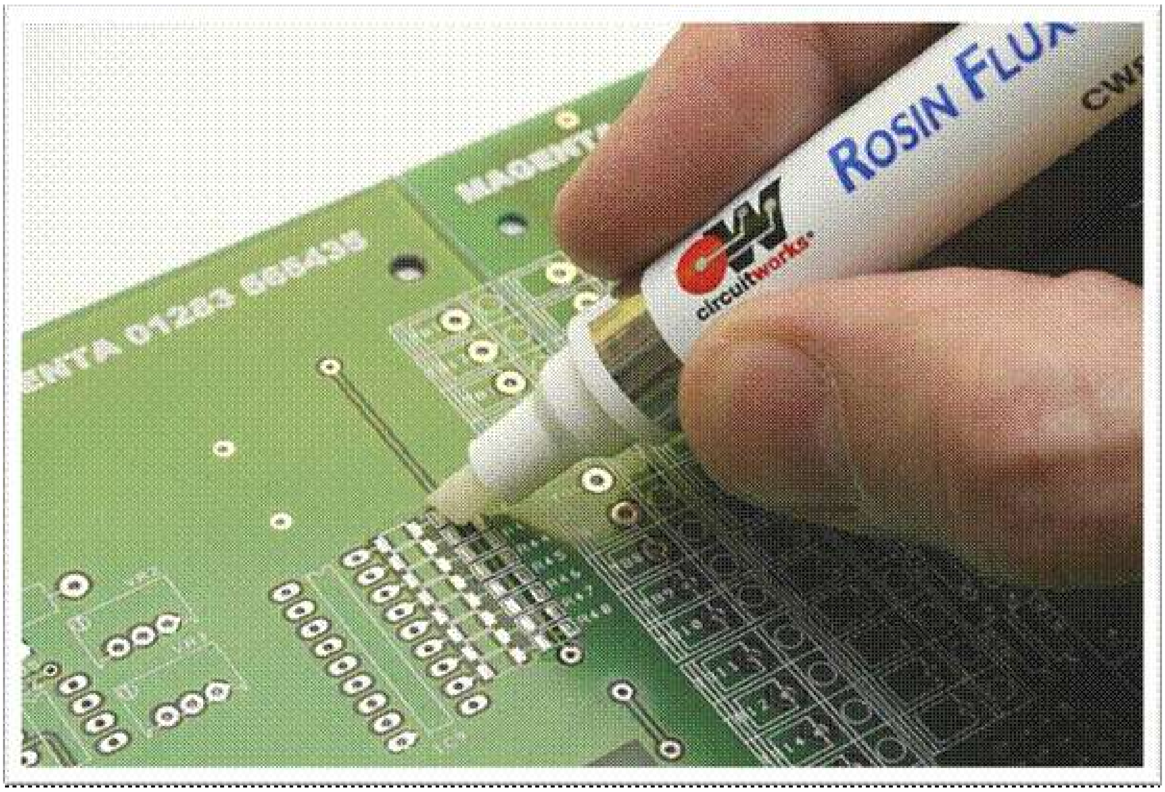

^^ For more demanding work, a flux dispenser pen allows additional flux to be applied

For example I've used specialist Chemtronics flux dispenser pens on tricky, extra-large solder joints involving very thick wires for lead-acid battery connections where I really struggled to make the solder flow properly. You might also use them in micro-electronic surface-mount work. The extra flux can only help, but for the rest of the time rosin-flux core solder wire contains sufficient flux and that's all that you'll need.

^^ *Colophony or rosin flux is available in small tinlets if you want to prepare your own liquid flux.*

One thing that many seasoned electronics enthusiasts will recognise is the distinctive smell of rosin flux: its intense woody pine smell is not unpleasant, and flux fumes themselves are not known to be harmful but solder smoke can be an irritant, especially if you suffer from asthma or other respiratory conditions. (I dealt with solder fume extraction earlier when soldering irons were outlined.)

Rosin flux is also known as **Colophony.** It's an amber resin that's glassy and brittle like sugar candy, distilled from the resins of conifers (mainly pine trees) and it's worth knowing about. For electronics use, Colophony is available in individual tinlets of solid resin (e.g. *Kolophonium* in 20g tins by Donau Elektronik GmbH, sold by Westfalia or Conrad).

As an aside, the resins of various trees are used in incense burners: placing a few fragments on a charcoal burner produces intense delicious fragrances of pine, frankincense or forest but you still wouldn't want to inhale the smoke or fumes directly.

Solids of Colophony can be dissolved to make a semi-liquid flux, for dipping or applying manually prior to soldering. To make your own rosin flux from solids of Colophony, chip off some fragments and crush them into a small tinfoil dish, then apply some isoproponal alcohol (start with 2 parts solvent to 1 part Colophony) and let it dissolve over about 20-30 minutes or more.

^^ Chip off some solids of colophony and put them in a small tin foil dish

^^ Add e.g. x 2 volume isopropanol alcohol and allow the solids to dissolve. Experiment as needed.

^^ *The resulting flux can be very sticky so handle with care and do not spill.*

In this form it can be applied directly with a brush or dropper to areas prior to soldering. In an open dish, after a day or two the solvent will mostly have evaporated to leave an extremely sticky resin that should be handled carefully: it's no co-incidence that Colophony is also used in adhesives and varnishes, and different grades of Colophony resins are used to treat violin or double-bass bows to add friction to bow hairs! Try experimenting with different ratios of solvent and making some up and storing it in an old nail varnish bottle with brush, or generally do what works for you.

Unused Colophony crystals can be reformed into their storage tinlet by warming carefully with a hot air gun, but don't overdo it. You can also use flux cleaners or PCB solvent cleaners if you want to remove traces of excess rosin flux after soldering: these are left behind as a brown deposit and are otherwise harmless.

The subject of using your soldering iron to raise the temperature of the materials is discussed next.

Temperature flow

This aspect takes a bit more understanding, but with practice you'll soon understand how temperature flows within a workpiece being soldered and how you have to harness it properly. After ensuring all parts are clean and the soldering iron is ready to go, the next step to successful soldering requires that the **temperature** of **all** the parts is raised to roughly the same level, before solder can be applied.

Imagine the most basic task of soldering a simple resistor onto a printed circuit board: the copper p.c.b. and the resistor lead should both be heated *together* so that the solder will flow readily over the joint. Later I'll show the precise stages step by step.

A beginner will often mistakenly just heat one part of the joint (e.g. a resistor's wire protruding through a printed circuit board) and hope that the resultant "blob" of solder will be enough to tack everything together. That's completely wrong, because the remaining metal in the joint is cold when molten solder is flooded on to it. The joint will be weak, incomplete or unreliable. Flux will not have flowed properly either, so the joint could be contaminated internally.

Another beginner's mistake is to use a soldering iron to carry blobs of molten solder over to the joint, as if to daub solder over it. The secret of success is to control the iron accurately and apply the hot tip onto the workpiece so that it's in contact with *all the parts*. Within a fraction of a second, heat will conduct from the iron and raise the temperature of the entire joint, after which solder can be melted over it. Remove the iron and let the joint cool naturally. It takes some practice with your chosen soldering iron tip to obtain the best heating action, making sure the tip is clean and tinned properly to begin with.

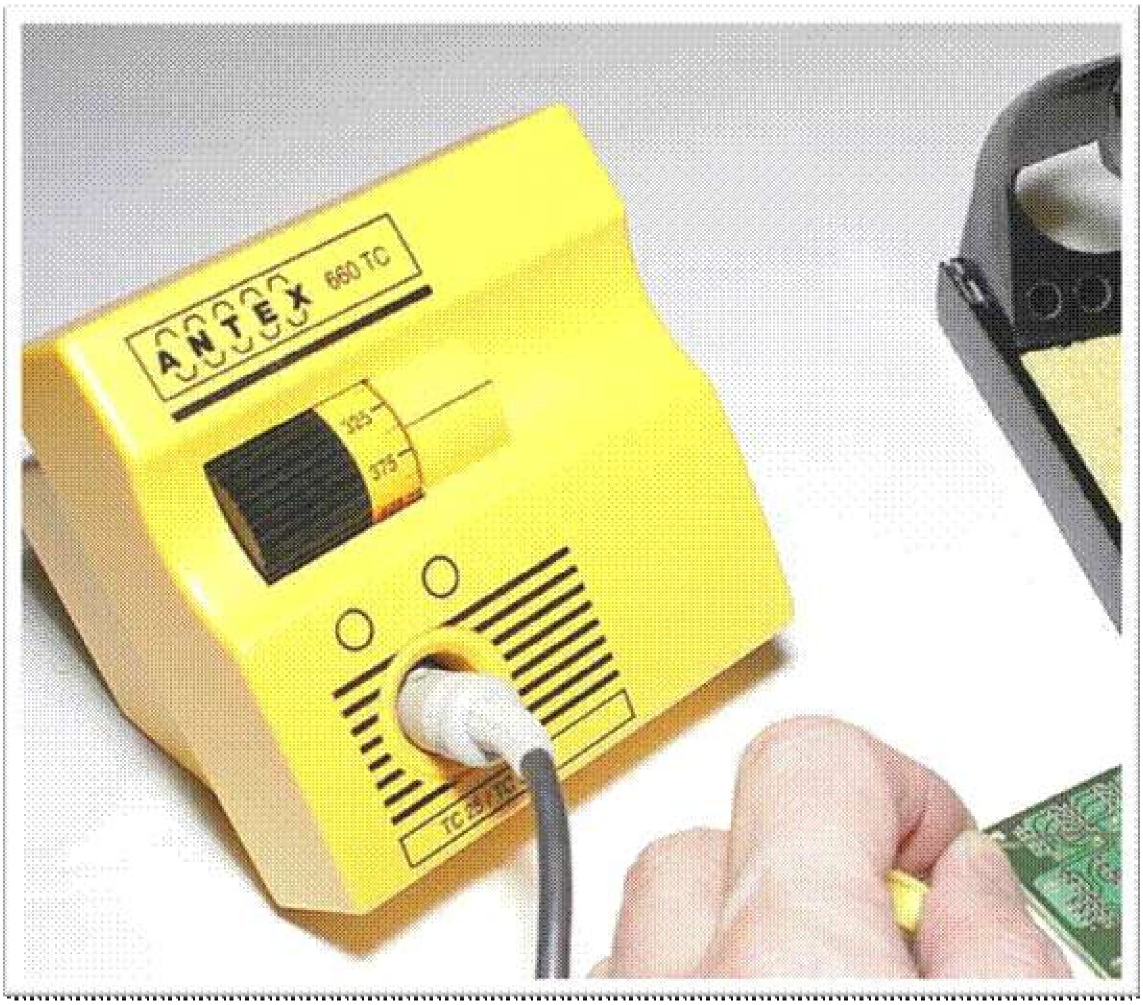

^^ *Set a typical temperature of 330-350 ºC (626-662 ºF) or more, if you have a soldering station.*

The melting point of traditional tin-lead solders is about 190ºC (374ºF) but lead-free tin-based solders require higher temperatures, having melting points of typically 201-227 ºC. As Antex reminds us, the melting point is *not* the temperature of the soldering iron tip: instead you should set a temperature that ensures the solder melts *instantly* onto the tip. Fixed-temperature soldering irons have no adjustability but they'll cope just fine with either type of solder. A soldering station usually has a variable control that gives more control in different circumstances. In practice the iron tip temperature should be set for typically 330-350 ºC (626-662 ºF) or maybe a little more if using lead-free solder.

Now is the time…

The next diagram shows what would happen if you applied a hot soldering iron to an imaginary metal block. In step (1) heat travels out of the iron into the cold metal block, which starts to warm up starting at the edges. Gradually (2), the whole metal block's temperature rises until the middle of the block finally rises in temperature as well. In effect heat is now travelling back "towards" the iron tip, until finally (3) the whole block is at the same temperature as the iron. At this point solder could now be applied.

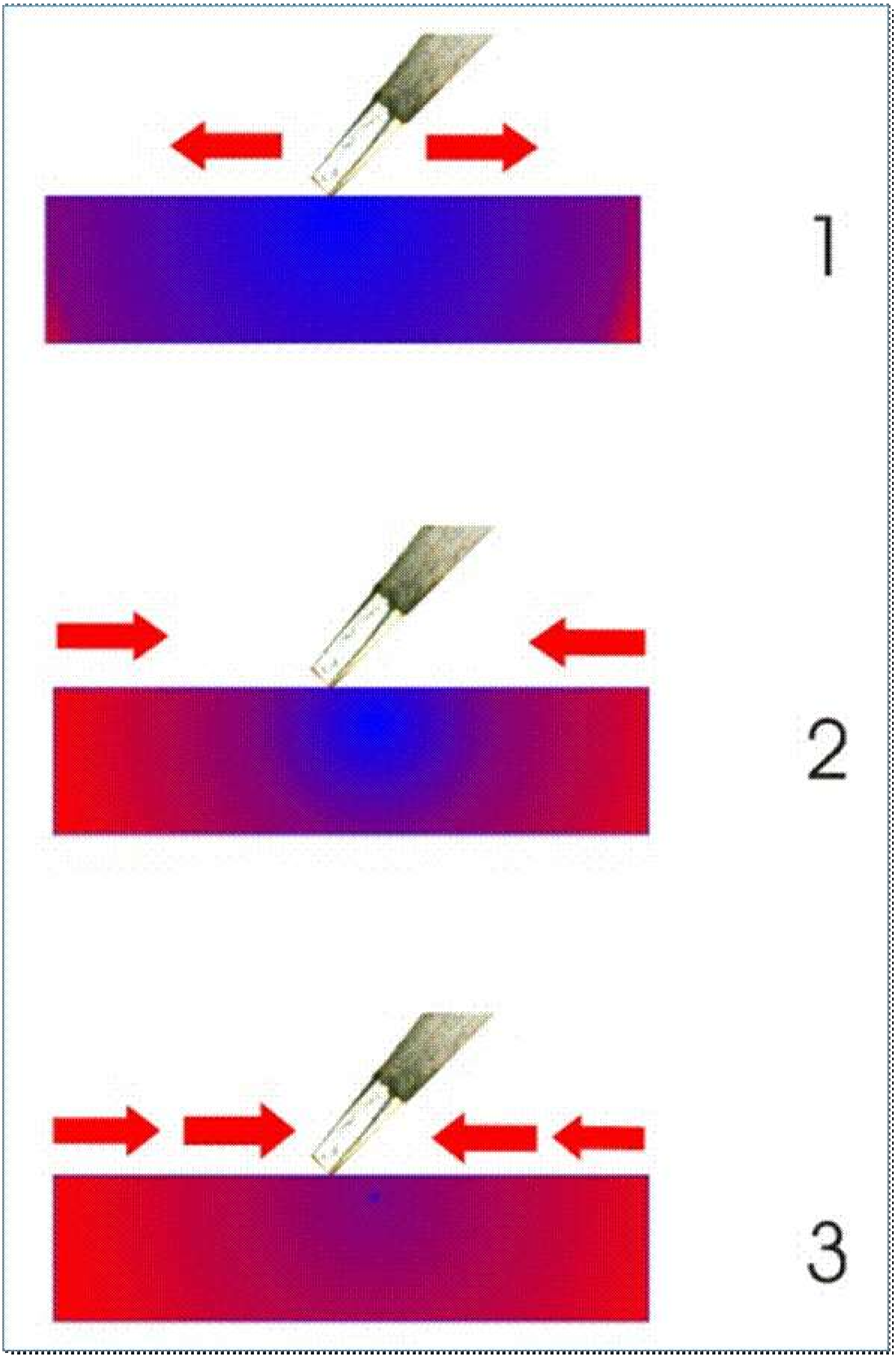

^^ How metal in the joint actually "sinks" heat away from the tip to begin with. Then heat moves back towards the tip until the solder's melting point is finally reached.

You'll notice this effect as a time delay when soldering any joint. The "working area" of the iron's tip gets **cooled** to begin with, because the metal in the rest of the joint is sinking heat away from where it's most wanted. Only after the whole joint has risen in temperature can solder be melted onto it.

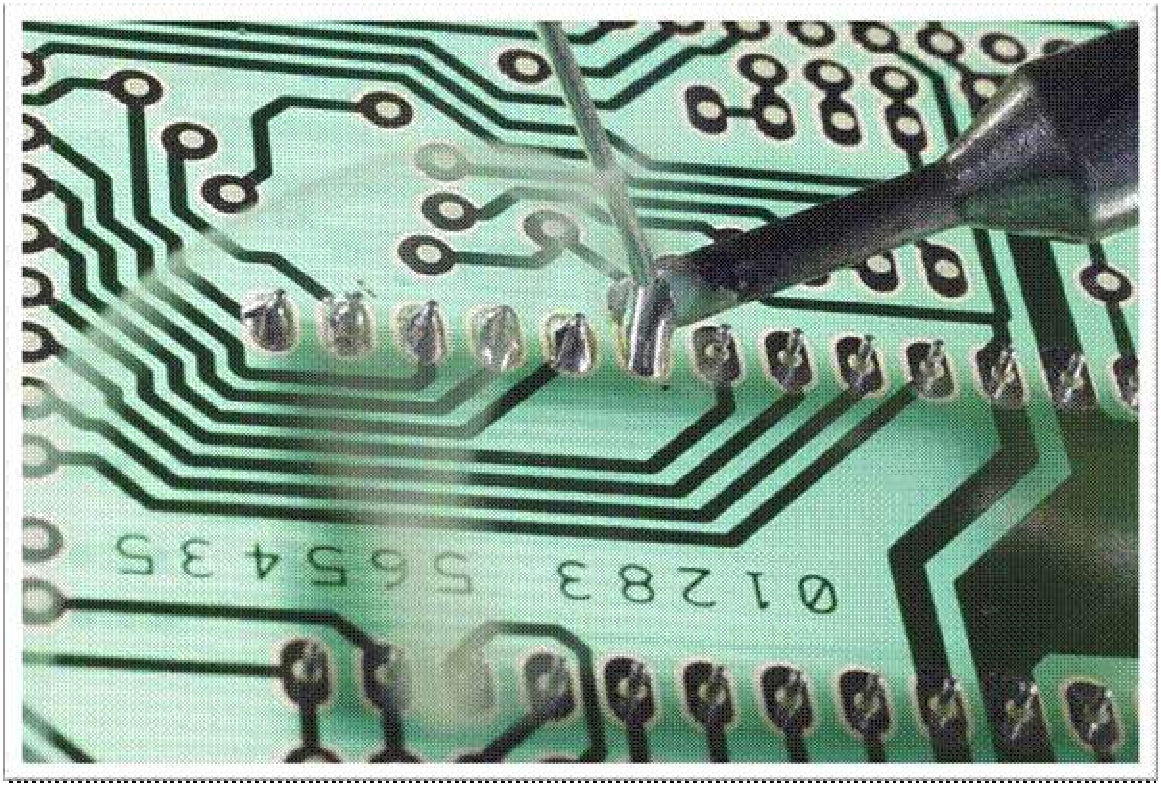

^^ *Soldering a small p.c.b. joint*

With experience, you'll get a feel for how long it takes before you can apply solder. The more metal that is present in the joint, the longer period that heat must be applied for. A small p.c.b. joint takes well under one second to complete. A large metal terminal could take quite a few seconds or more to heat up. As I explained earlier, higher power (wattage) irons cope better with larger workpieces because they recover more quickly and are more "unstoppable", making it easier for them to heat larger workpieces without cooling down so much.

If you apply solder too early, it won't melt properly and the result will be a grey, crystalline joint caused by the solder's melting point temperature not being reached and the flux not having flowed properly. Semiconductors must be soldered as rapidly as possible as they are heat-sensitive, but they're a lot more robust than they used to be.

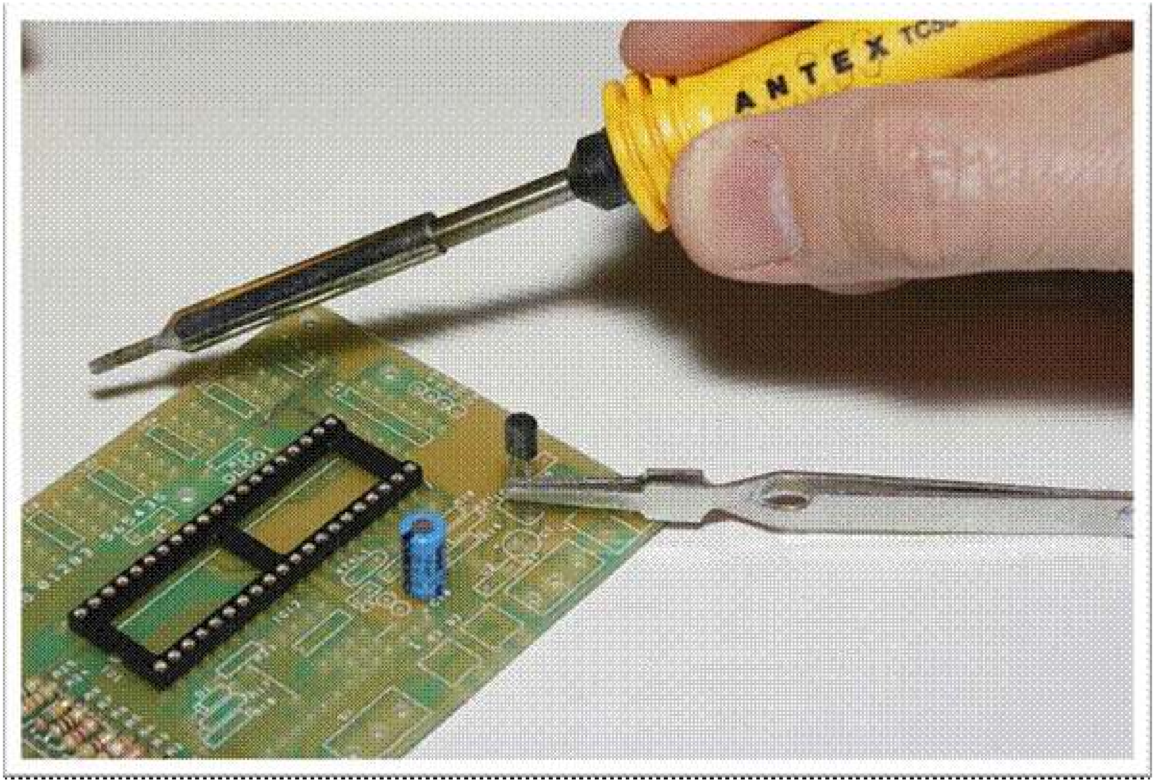

^^ *A clip-on Antex heat shunt fitted to a transistor leg, helps prevent thermal damage due to overheating when soldering it in place. Less essential these days, but beginners find them re-assuring.*

Until they have gained some practice, novices sometimes buy a small clip-on heat-shunt (photo) which resembles a pair of aluminium tweezers. In the example of, say, a transistor, the shunt is attached to one of the leads near to the transistor's body. Any excess heat then diverts up the heat shunt instead of into the transistor junction, thereby avoiding the risk of thermal damage. Applying far too much heat may destroy the part or damage the p.c.b. foil which may lift away from the board.

In due course constructors learns to judge how much solder should be applied to any particular joint. An ideal p.c.b. joint is very slightly concave in shape. If not enough solder is used, the result may be an incomplete joint which may cause an intermittent fault later on. An excess of solder – shaped like a ball bearing - is an unnecessary waste and in extreme cases may cause short circuits, especially on densely-populated boards. There is no need to add more solder "for luck". Professionally-produced p.c.b.s have a green *solder resist* coating which helps to ensure that solder does not stray onto adjacent pads. As a finishing touch, I usually spray the solder-side of a circuitboard with aerosol spray lacquer afterwards. It keeps the solder joints nice and shiny and helps prevent corrosion.

Some components can create hazards during a soldering operation:

Coin cells and button batteries are commonly used as power, clock or memory backups. If heated excessively they can explode without warning due to the build-up of internal pressure. Spot-welders are used in industry to connect tags to them, but if you need to solder wires to such a cell then it should be done as quickly as possible.

Some **memory back-up capacitors** or **electrolytic capacitors** remain energised for a while even when the circuit is powered down. Molten solder is a perfect electrical conductor and in some cases the component’s contacts could be shorted during the soldering (or desoldering) operation. If molten solder shorts it out then the arcing may cause globules of solder to be spattered outwards without warning, potentially risking eyesight damage.

Always ensure that powered components are electrically inert and discharged before soldering. Cells, batteries and battery packs should not be accidentally shorted during the soldering process, to avoid arcing and solder spatter. Note that electrolytic capacitors can also explode after a while if reverse-connected, so observe polarity closely.

Let’s now consider the practical stages of soldering components and wires successfully.

Soldering Step by Step

Earlier I explained the individual factors that affect the quality of a solder joint. These are:

- **Cleanliness** – dirt or impurities drastically hinder good solder coverage.
- **Temperature** – the right level to enable the solder to flow freely!
- **Time** – apply heat for just the right amount of time!
- **Adequate solder coverage** – enough to form a good joint without touching neighbouring areas.

These rules apply whether soldering a p.c.b. or performing other tasks such as interwiring (hooking everything together with connecting wire).

We’ll now summarise the stages of making a typical solder joint – soldering components onto a printed circuit board (through-hole soldering). Most people insert components into the circuit board and simply splay the wires out to hold them in place under spring tension. I find it best to snip excess wire leads off at this stage, to improve accessibility.

- All parts must be bright, clean and free from dirt and grease.
- Try to secure the work firmly to stop parts moving around.
- “Tin” the soldering iron tip with a small amount of solder. Do this immediately, with new tips being used for the first time.
- Wipe the tip of the hot soldering iron on a damp cellulose sponge to remove excess solder or contamination.
- Many people then add a tiny amount of fresh solder to the cleansed tip just before using it.
- Heat all parts of the joint with the iron typically for under a second or so, until it’s heated throughout.
- While heating, then apply sufficient rosin-core solder to form an adequately-covered joint.
- It only takes a second or two at most, to solder the average p.c.b. joint this way.
- Do not move parts until the solder has cooled.
- Remove and return the iron safely to its stand.

This special photo sequence illustrates these stages. It’s best to start with the smallest, fiddliest parts first when soldering a blank p.c.b., because that’s when you’ve got the most access on the board. Accessibility will be reduced as more components are added, so we’ll start with a simple wire link on a professionally designed p.c.b.

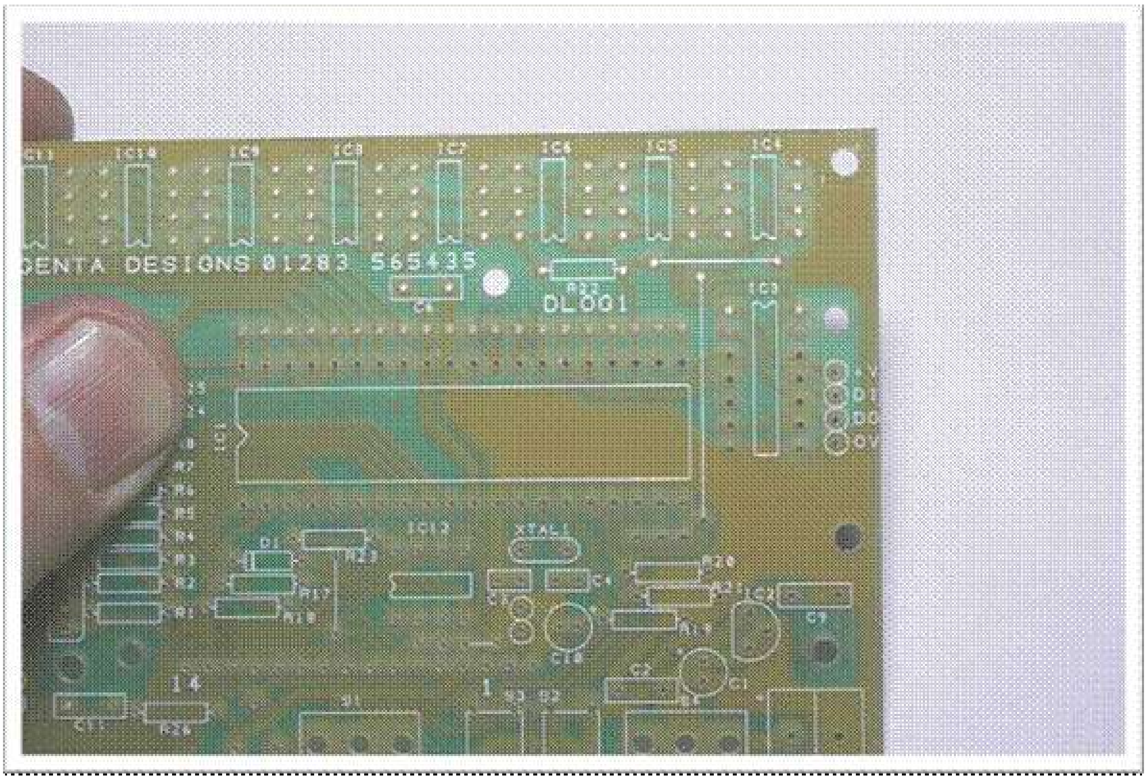

^^ *A typical professional blank p.c.b. – silk-screen printing shows what goes where. The underside has been treated with a green solder resist coating, and the solder pads are ready-tinned to help with soldering.*

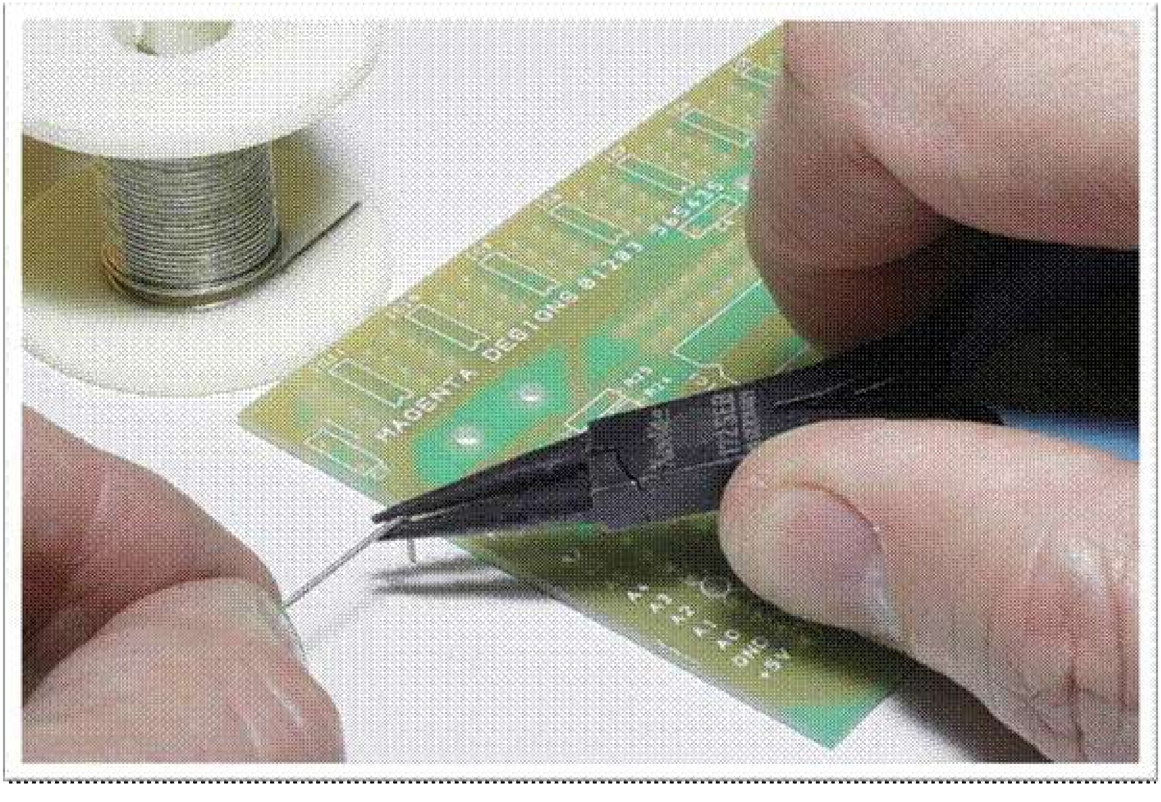

^^ *Preparing a wire link for soldering – cut off some tinned copper wire and bend the ends to fit the p.c.b. correctly. Round-nose pliers (shown) are perfect for this, but ordinary electronics or "radio" pliers will do.*

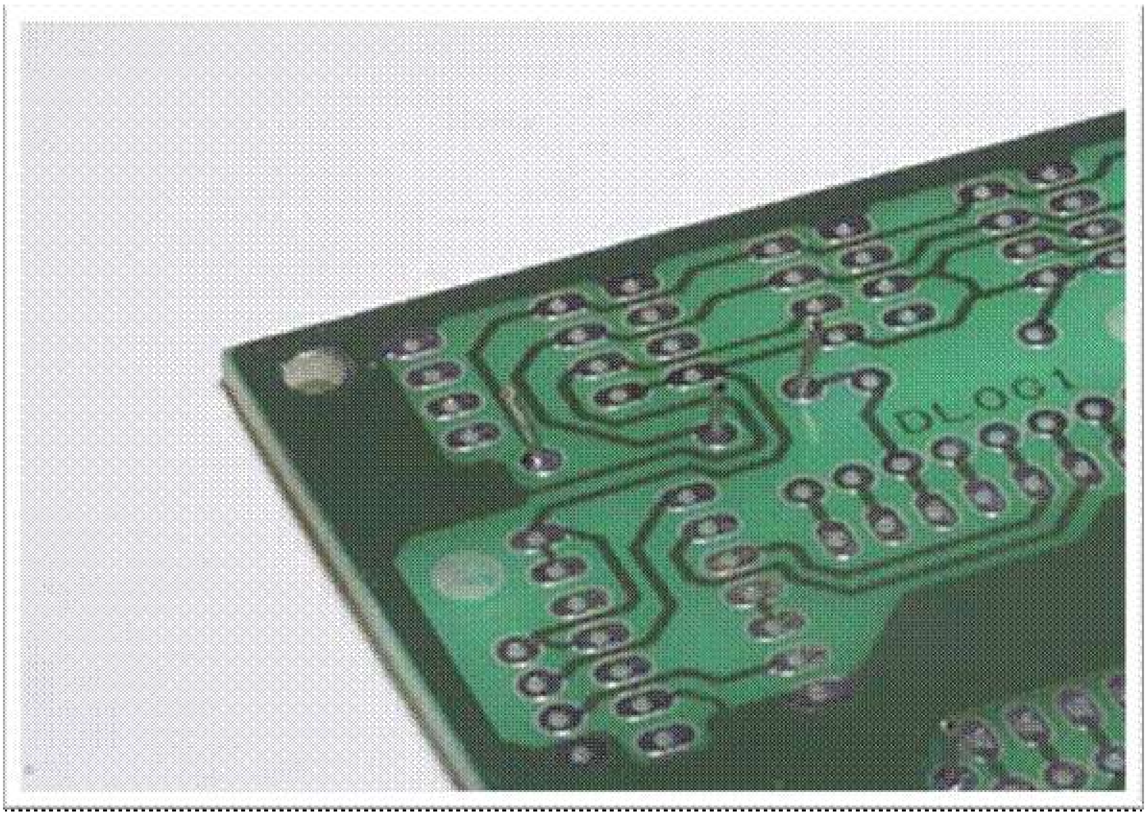

^^ Wires are prodded through the holes in the board, then turn it upside down to view the solder side. You can then "spring" or splay the ends apart slightly, so they are held in place while you solder them.

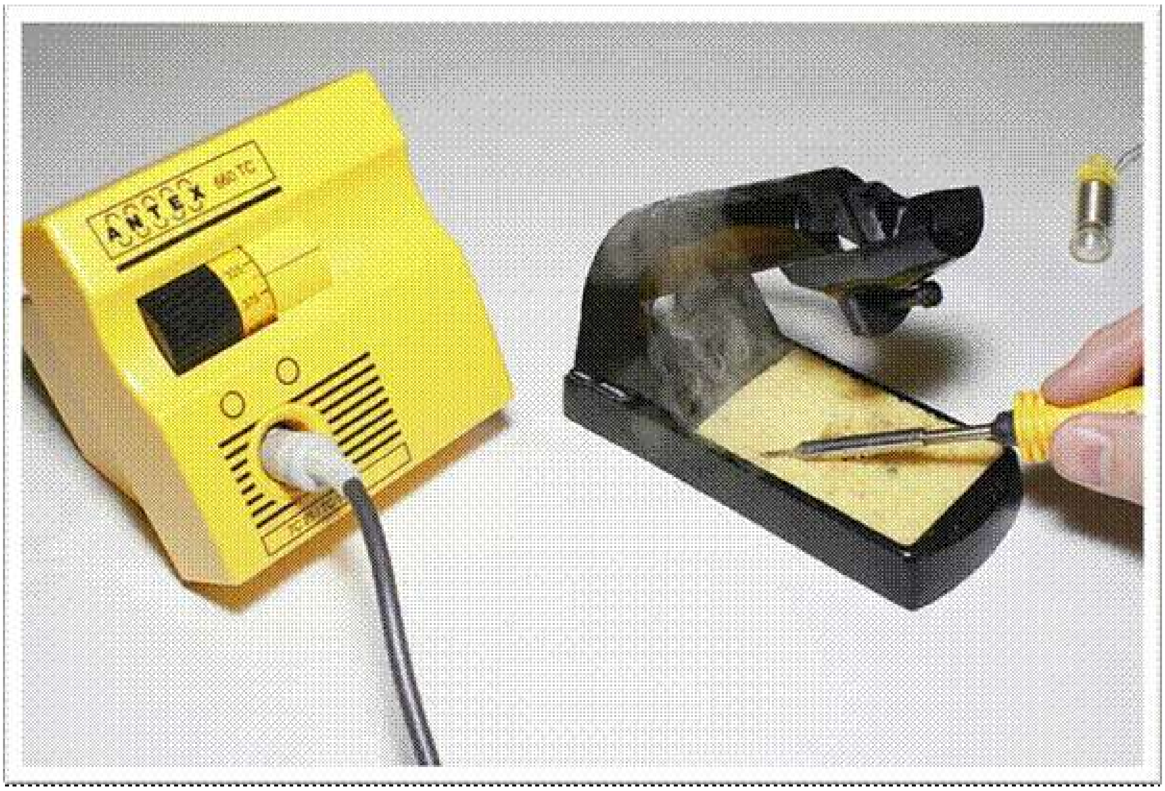

^^ Wipe the hot soldering iron on the damp sponge to clean the tip. Do this periodically when contamination, flux deposits etc. build up on the iron to keep the tip nice and shiny. Tip Tinner & Cleaner helps too.

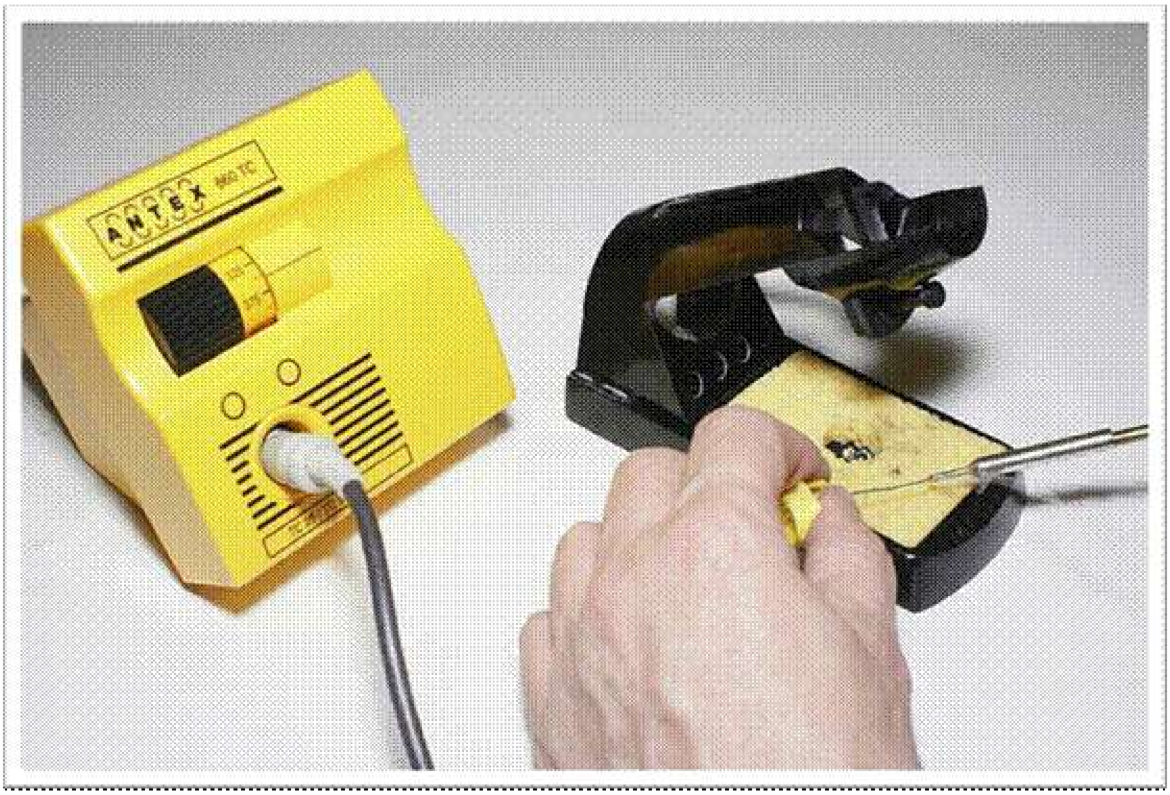

^^ It often helps to dab a tiny amount of solder wire onto the tip, to improve heat transfer.

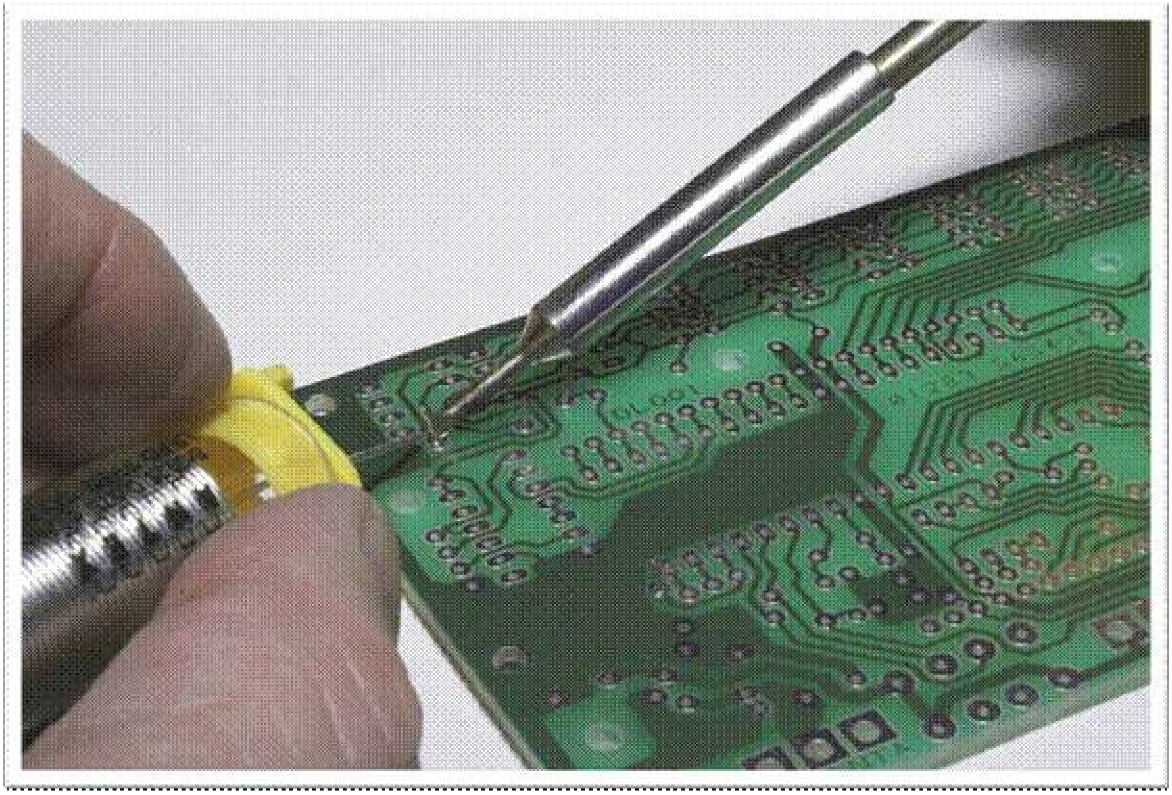

^^ Then apply the soldering iron to heat both the solder "pad" and the wire end at the same time (say <1 second). Apply a few millimetres only, of solder. Then remove the soldering iron immediately and allow the joint to cool down by itself. The green solder-resist coating ensures solder doesn't flow onto neighbouring pads.

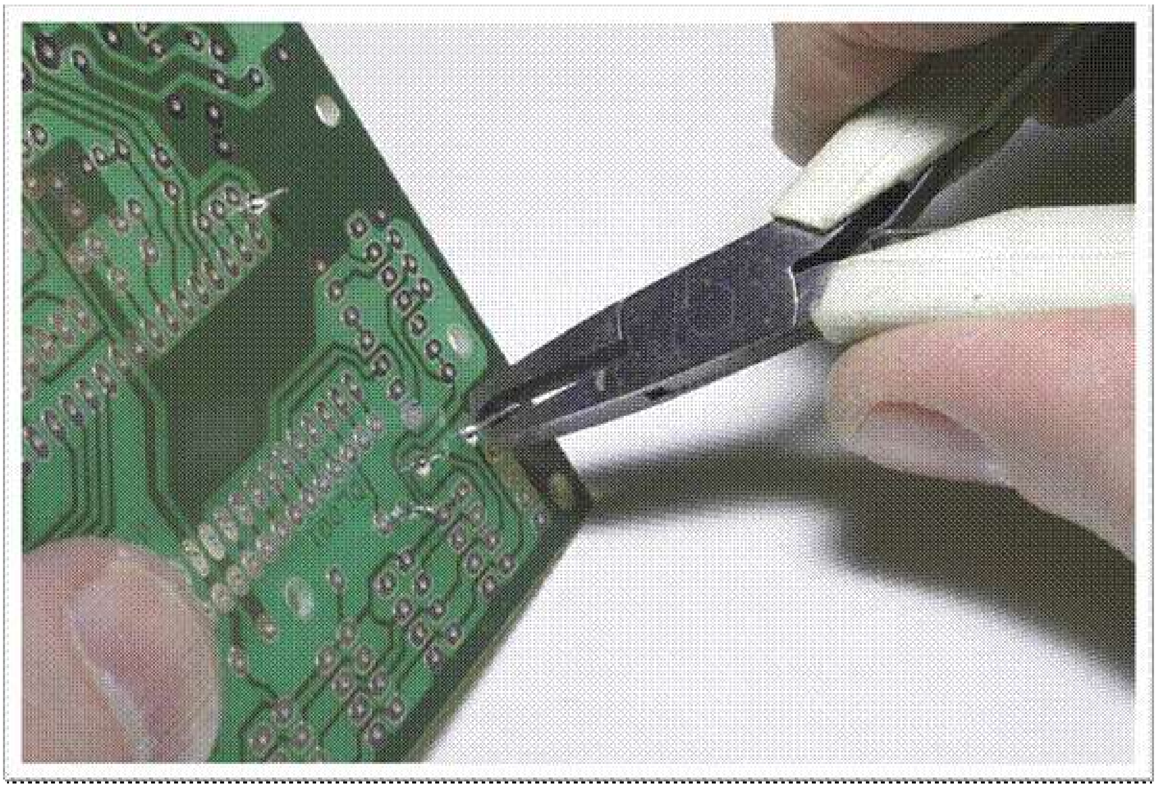

^^ These "end cutters" can then be used to snip off excess wire after soldering….

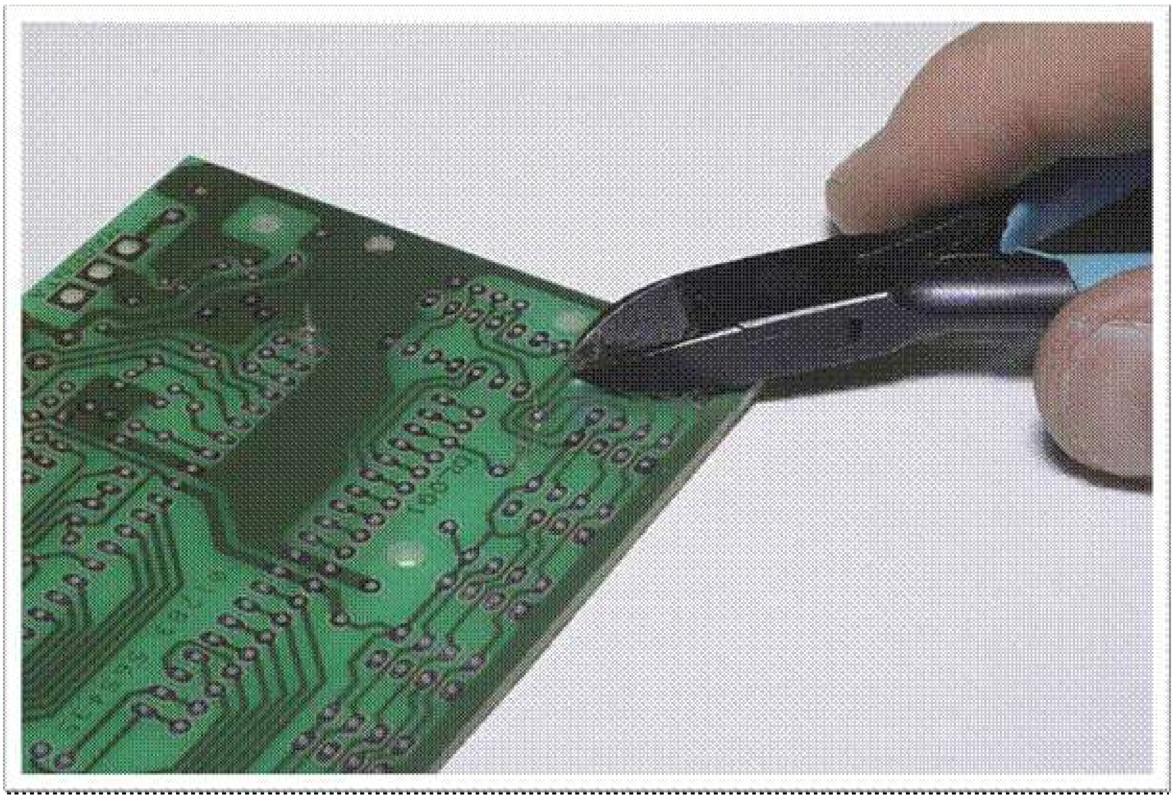

^^ … but ordinary electronics "side cutters" are fine for trimming or snipping wires.

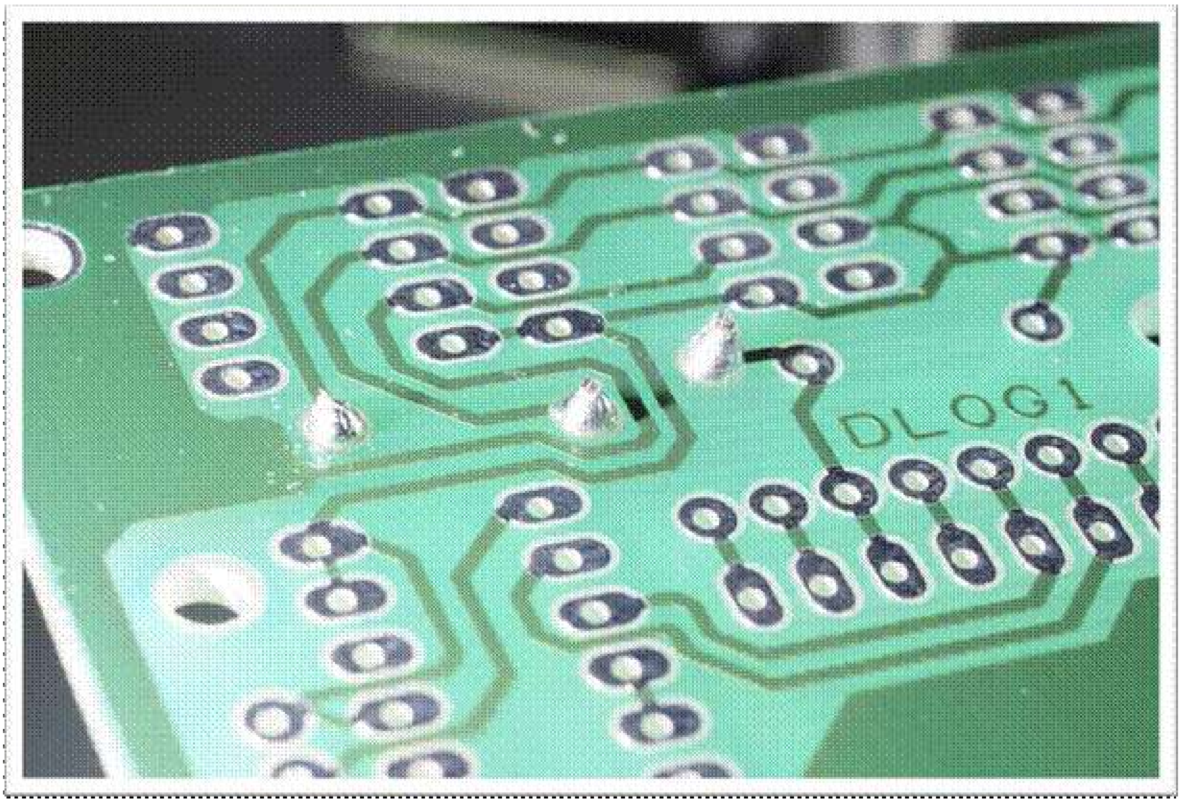

^^ The ideal solder joint should be smooth and slightly concave, quite shiny, not dull or crystalline-looking nor ball-bearing shaped.

Try to get consistent results in your soldering, but don't worry too much about the lack of uniformity – this is soldering by hand after all, not by precision machinery, and some slight variation is OK provided the solder coverage is good and the soldering is generally clean and effective. You'll improve with practice, that's for sure!

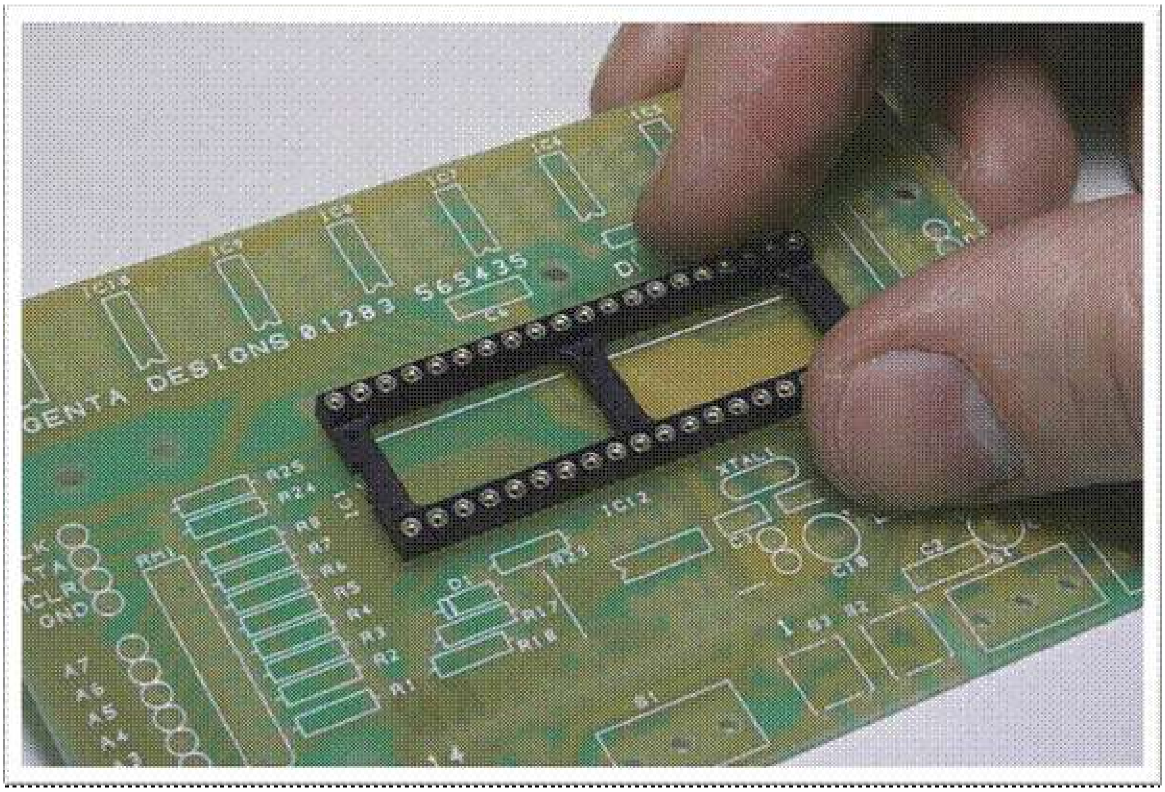
GENTA DESIGNS 01283 565435
XTAL1
IC12
R25
R24
R8
R7
R6
R5
R4
R3
R2
R1
R17
R18
CLK
DATA
CLR
GND
A7
A6
A5
A4

^^ A large i.c. socket can be added next…

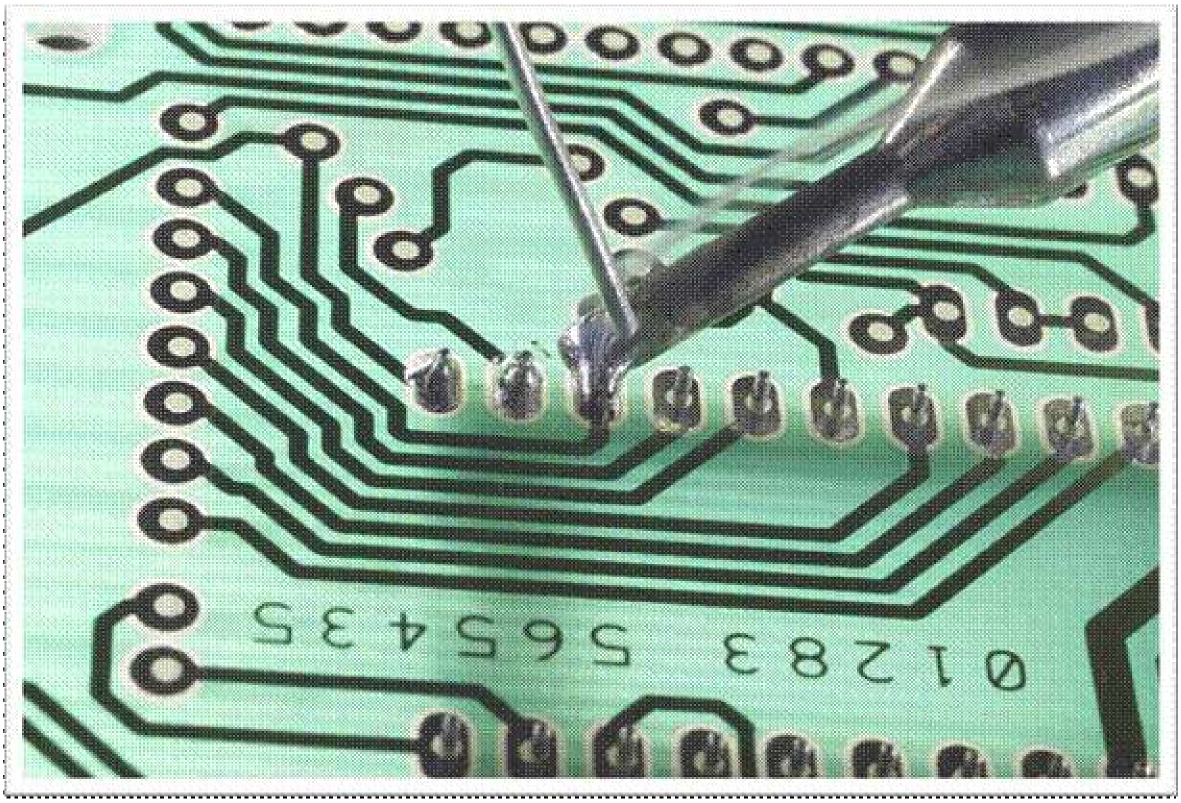

^^ Go along the rows of pins and solder them one at a time. See how simple it is! Heat both the solder pad and pin with the iron, and dab on a little solder wire to ensure full coverage. Ensure the socket is flush against the board.

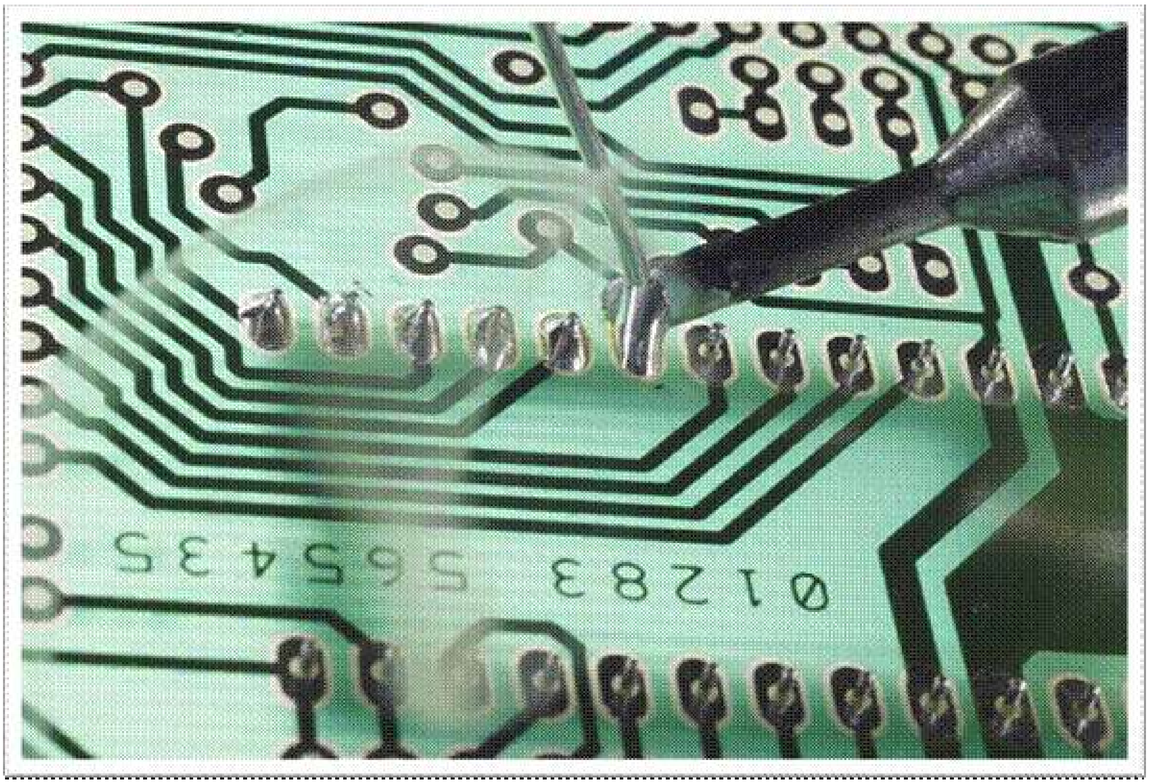

^^ As hardly any metal is involved each pin should take well under a second to solder at most.

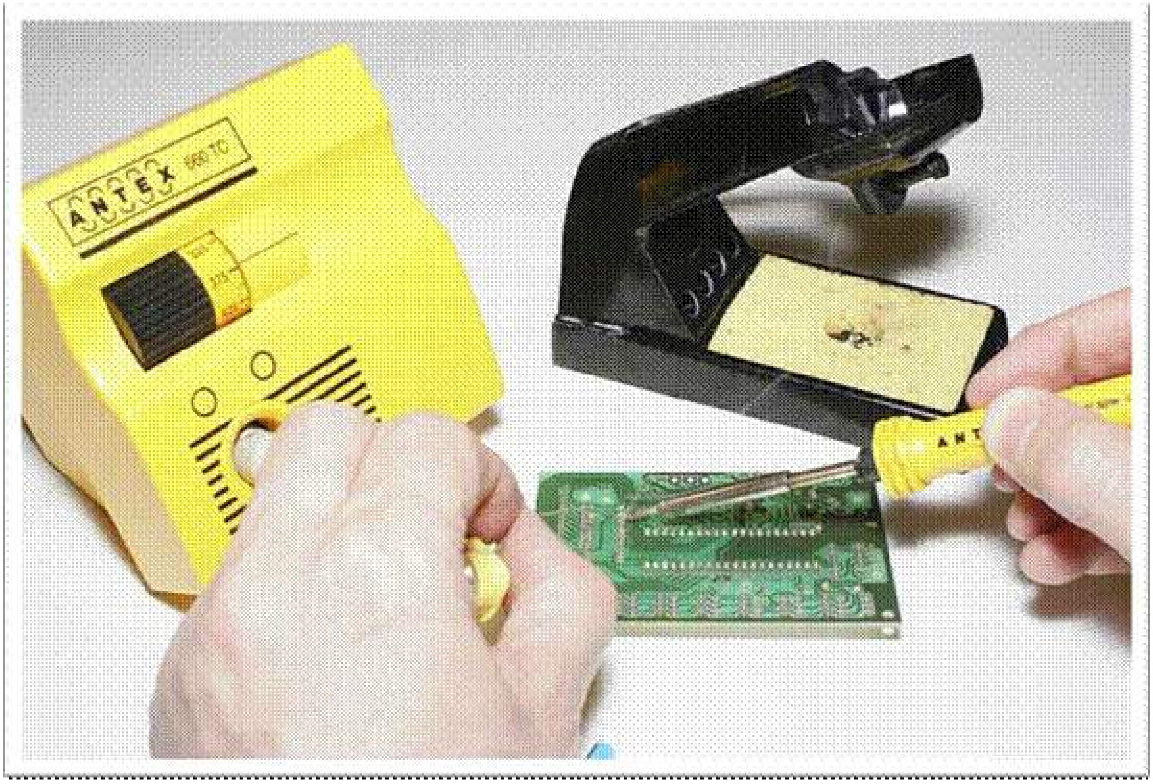

^^ *Continue until the i.c. socket is soldered.*

Once the smallest parts are soldered into place, you can continue to solder the remaining components. It’s easiest to handle the smallest so-called “discrete” parts first while you still have plenty of room on the board. I usually solder resistors and capacitors next. The principle of soldering them is just the same as a simple wire link: insert them from above till they are flush on the board, then splay their wires a little to hold them in place, and preferably snip off at least some of the wire to give you more access.

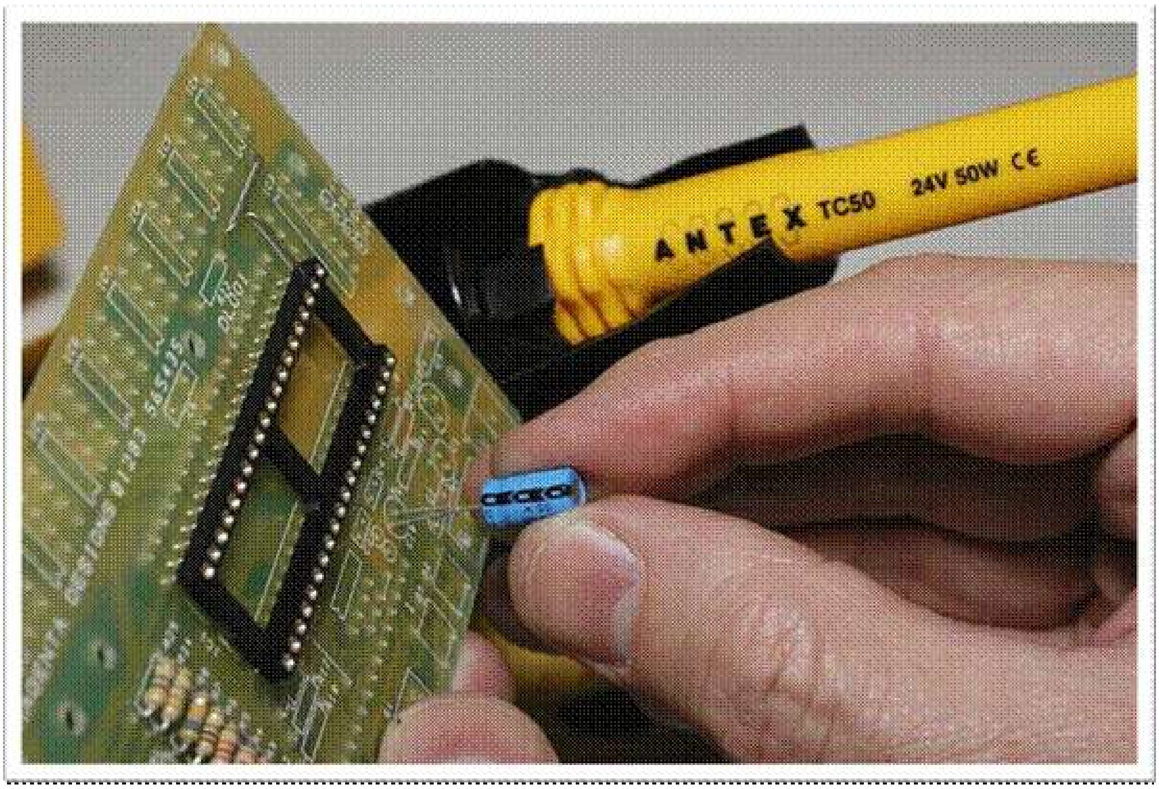

^^ *Continue by inserting “discrete” components from above, splay their wires out* *underneath to secure them in place, then snip and solder the joints exactly as before. Some parts like this blue electrolytic capacitor are polarity sensitive (note the – sign), and must be inserted the right way round. Same is true of every semiconductor.*

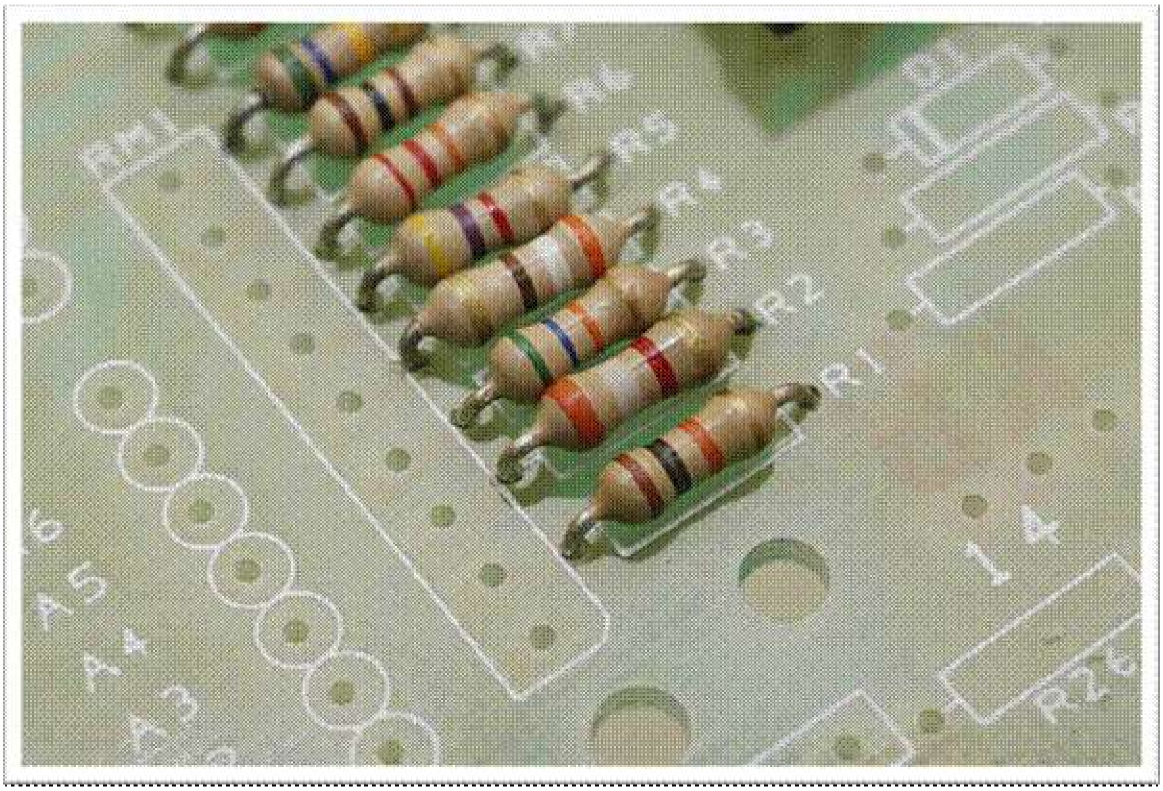

^^ This row of ¼ watt resistors was next. I like to show the colour codes all the same way round, with gold or silver (tolerance bands) on the right for consistency.

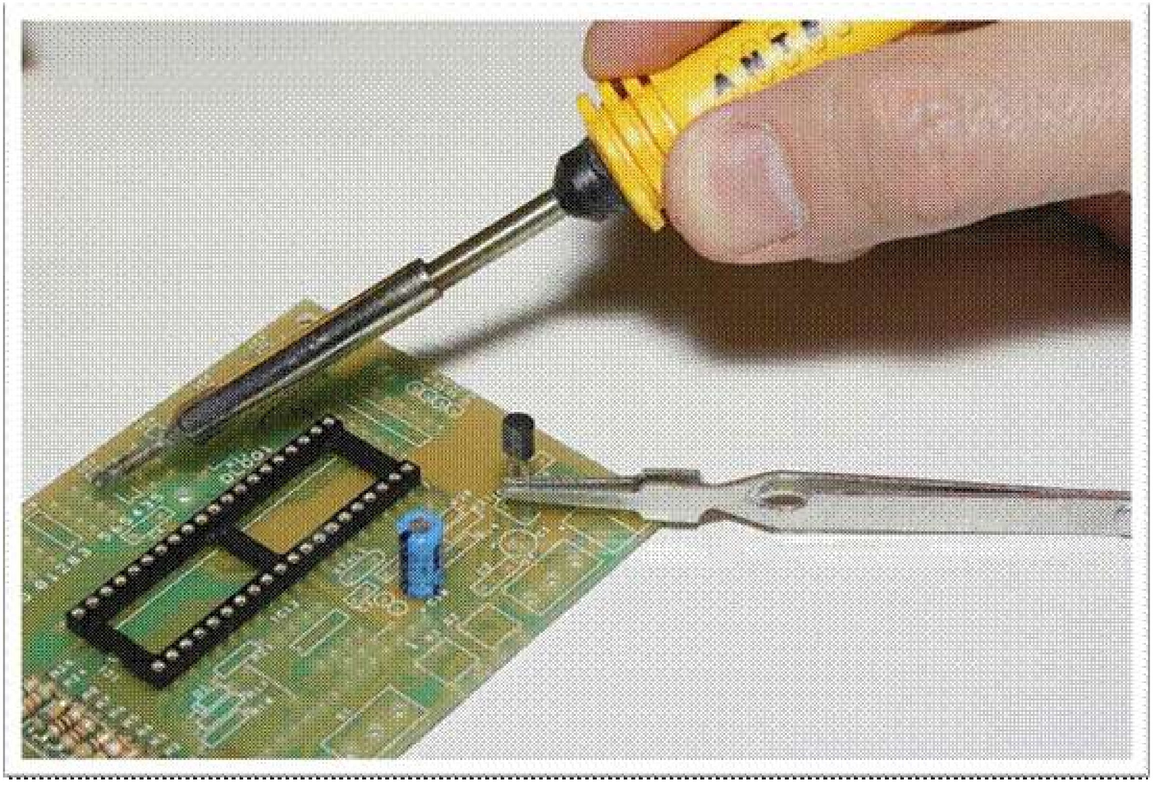

^^ This transistor was next. Solder it quickly to avoid damage, and observe polarity correctly, so it is the right way round on the p.c.b. The heatshunt is optional.

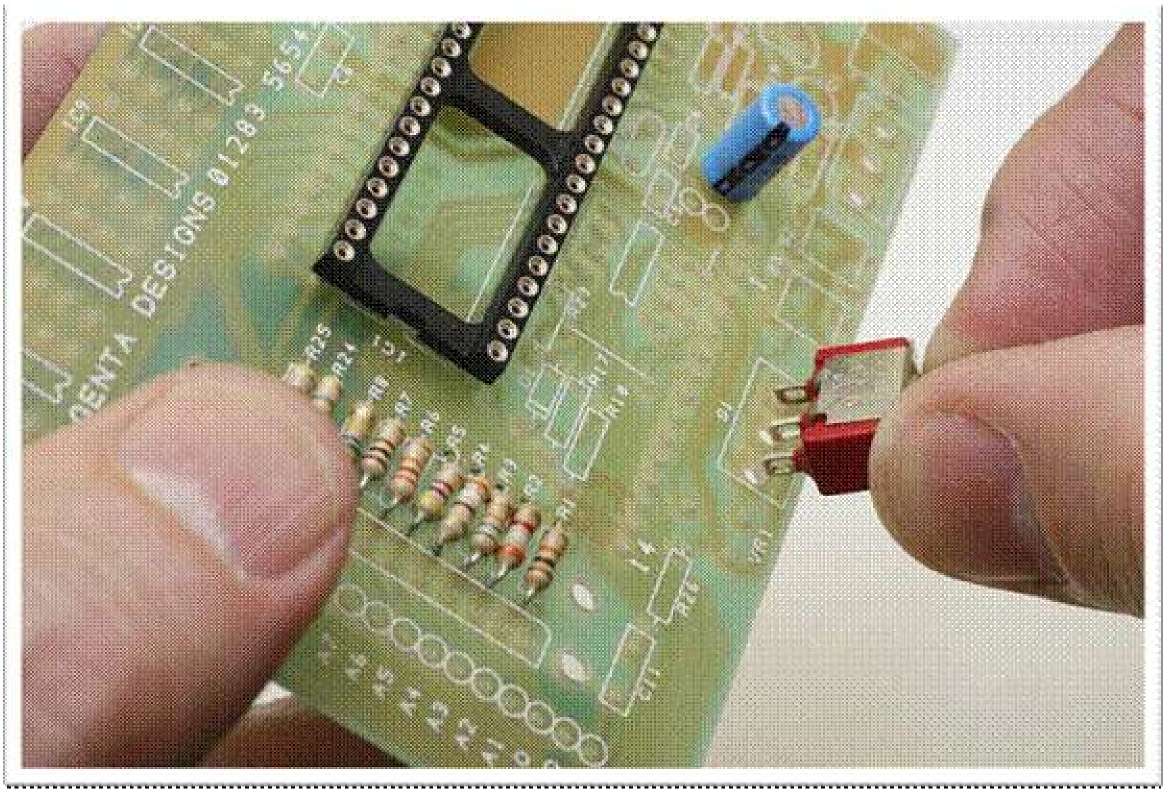
DESIGNS 01283 565

^^ Unusually, this toggle switch fitted directly onto the board as well which saved wiring it up. Neat idea!

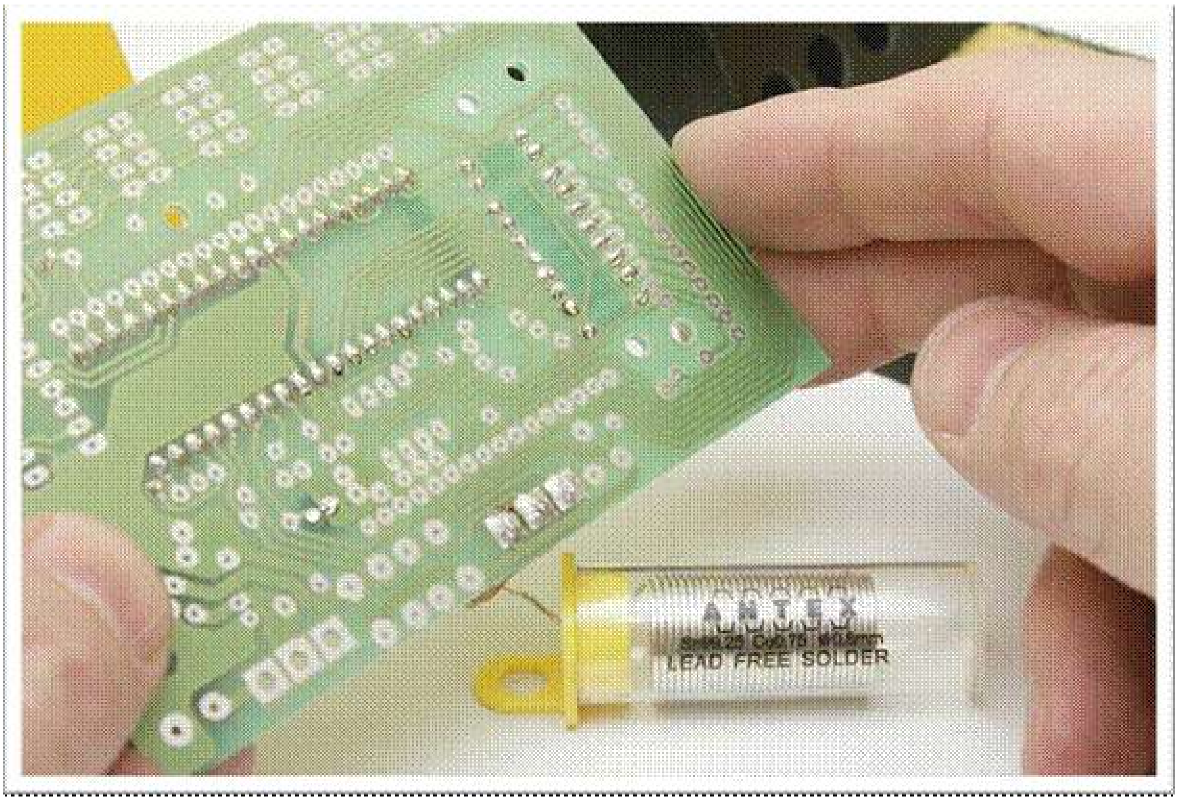
ANTEX
LEAD FREE SOLDER

^^ *As there is comparatively more metal to heat up, it'll take longer to solder the switch terminals, and you'll need more solder as well. Thicker gauge solder is useful at such times. Allow say 2-3 seconds to solder each terminal.*

Don't forget to clean the soldering iron tip on its damp sponge every now and then, to ensure the bit is kept clean and shiny. Later on I'll show you how to correct any problems by desoldering using various techniques. Next, we'll move on from printed circuit board "through-hole" soldering and look at how to handle wires and leads.

Interwiring – get hooked up!

With practice, through-hole soldering of p.c.b.s will become second nature. There's no substitute for tackling some soldering jobs though, particularly trying a simple kit based on a quality p.c.b. which will boost your confidence enormously.

Most electronic devices need connecting up to external components such as battery packs, speakers, l.e.d.s or switches. Usually, multi-stranded connecting wire is used to connect circuit boards and external parts together. Unlike single solid-core wire, *multi-stranded wire* is flexible and vibration-resistant. Hobbyists mainly use 7/0.2mm wire (7 strands, each 0.2mm diameter) for low-voltage hook-ups although much Chinese equipment uses much thinner wires than this. So let's look at some aspects of soldering this type of wire.

In a separate photo sequence I show how a potentiometer (a panel-mounted variable resistor) and a light-emitting diode (l.e.d.) are connected using multi-stranded wire. The same principles of soldering apply to most other components including panel-mounted switches, loudspeakers, buzzers, audio sockets and more.

Components usually have terminals or "tags" to which wires can be soldered. Start by ensuring the component's tags are clean: otherwise solder will not wet properly and the joint will be impossible to solder, so all contamination must be removed. This is especially true of parts that have been in storage a long time. The connections often oxidise or blacken, so clean the solder tags with e.g. an abrasive glass-fibre brush, or a needle file or abrasive paper. Using a glass fibre brush was shown earlier in "Cleanliness and Tinning the Bit".

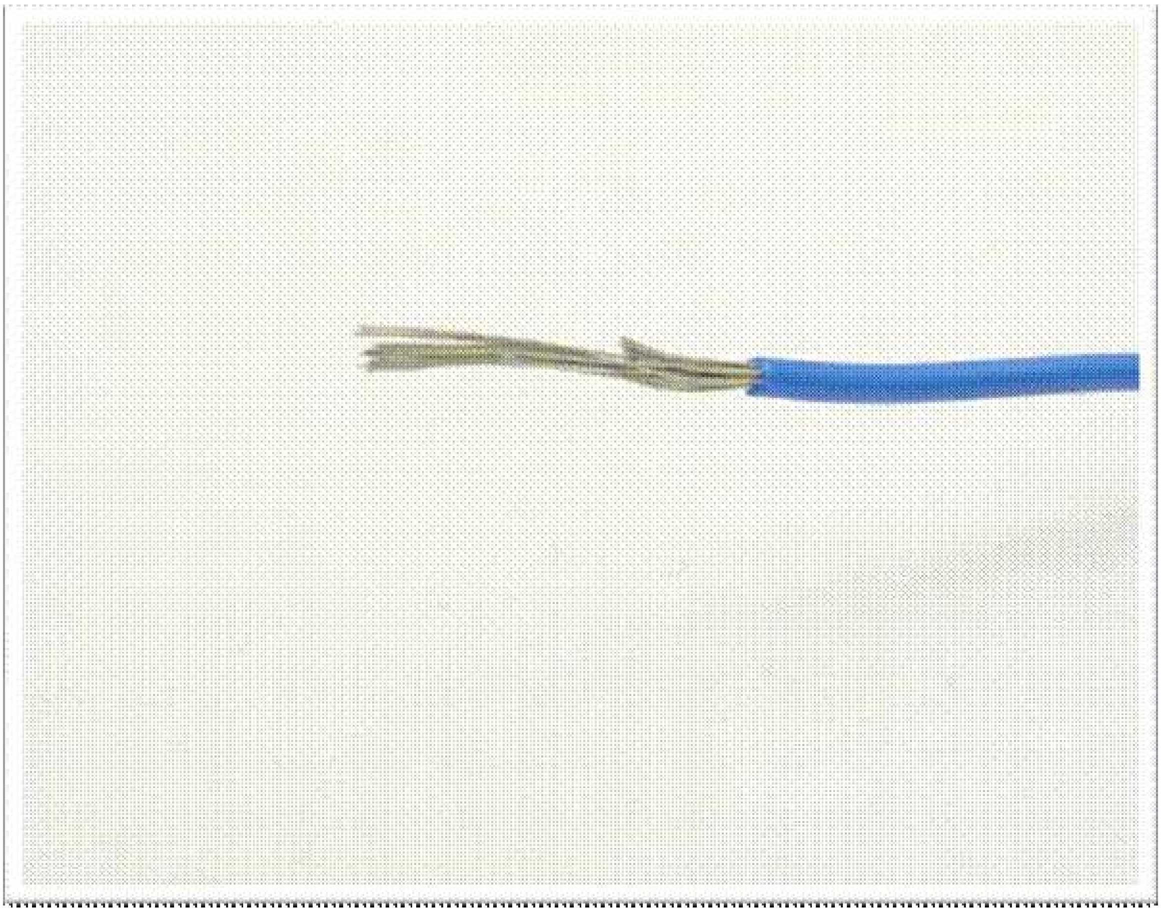

^^ How not to strip insulation off wire: some of the cores have also been cut – avoid doing this!

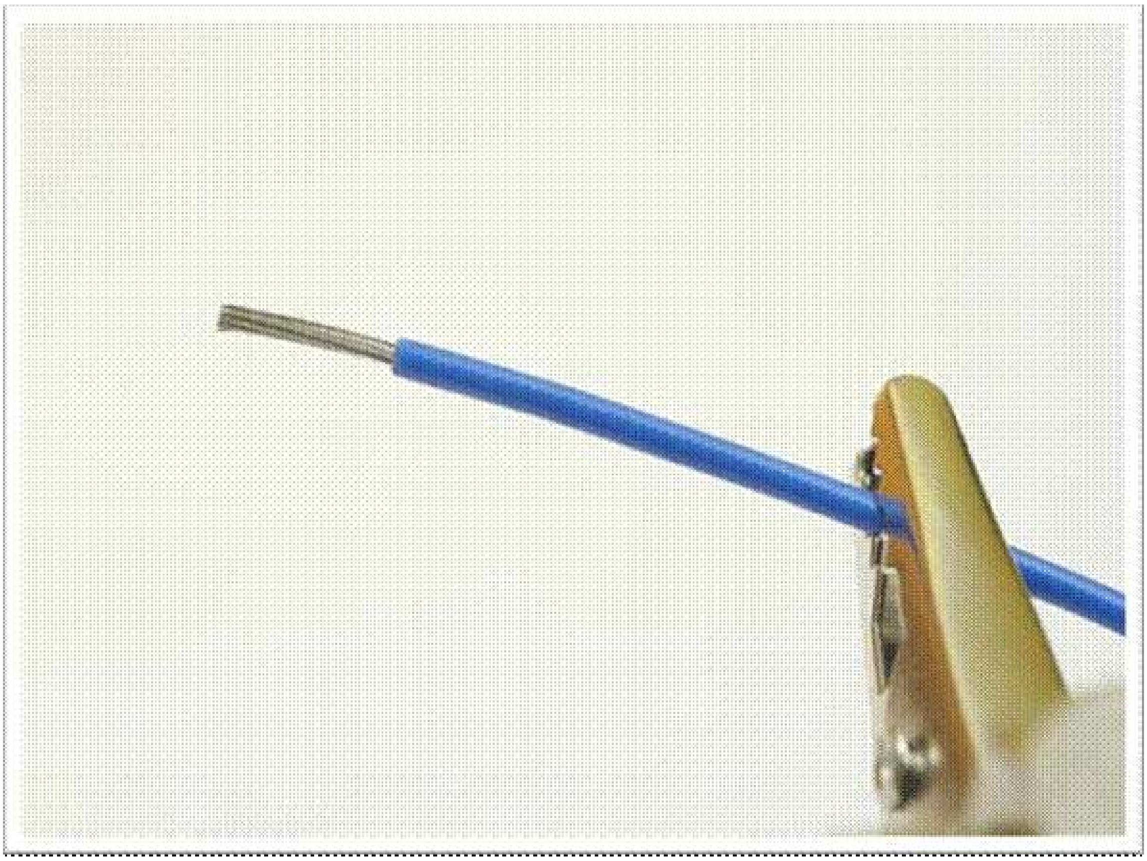

^^ Gripping a wire end in a "Helping Hand" croc clip can help with soldering.

After stripping a short length of insulation from the connecting wire, there are two ways to solder it to a component's solder tag. The first way is to "*tin" the stripped wire end to solidify it* - just heat it with the iron and melt a little solder on it, and let it cool. Poke the wire end through the solder tag, apply the hot iron to both parts and solder them together using a few millimetres of solder.

Although the assembly doesn't hold itself together so well during soldering (consider a Helping Hands jig if needed) this is quicker and easier to make and also easy to desolder again, and is perfectly adequate for most joints of this kind. The majority of commercial wire joints seem to be made this way.

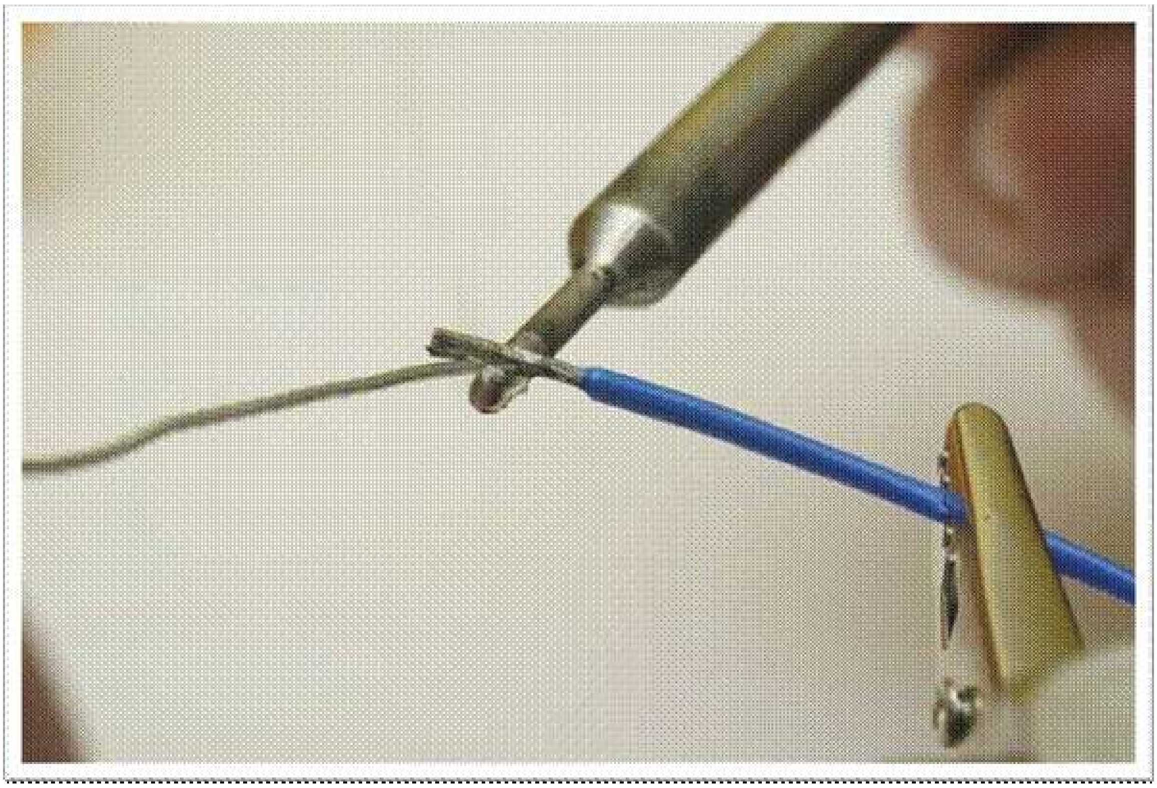

^^ Apply a hot iron and solder, in order to "tin" the wire ends. This makes them into a solid.

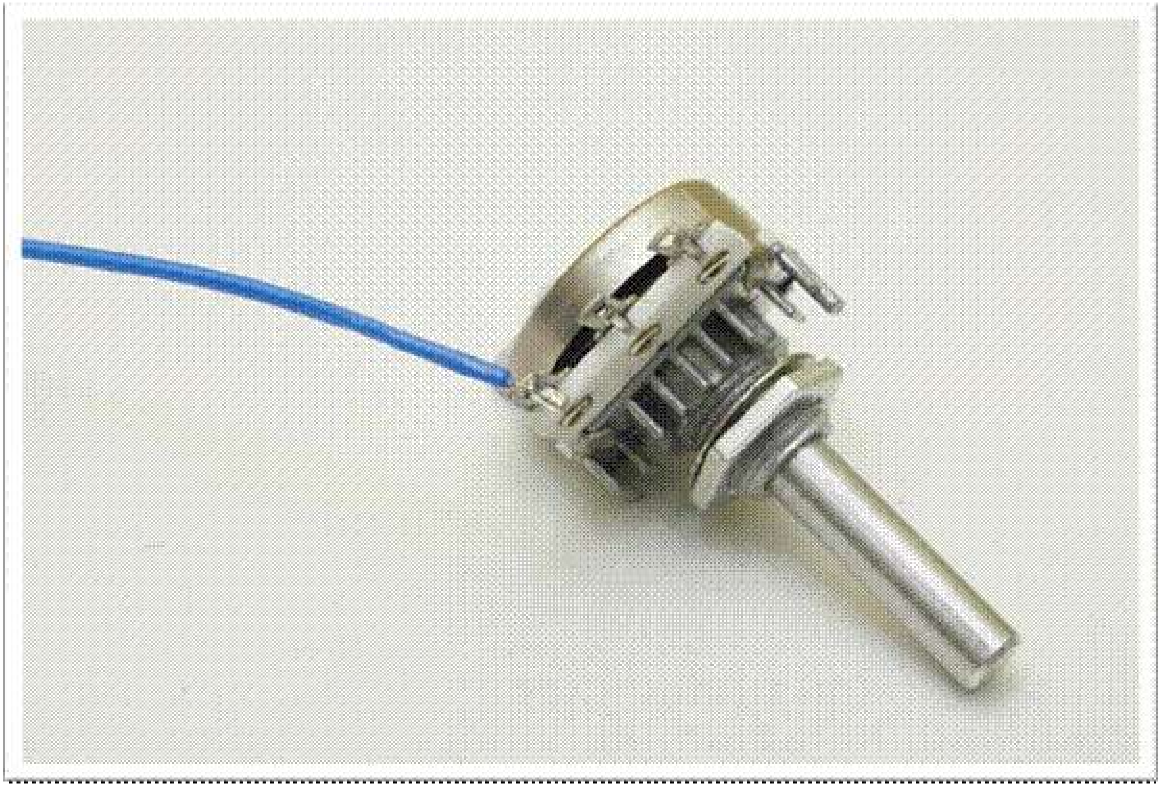

^^ Then feed the tinned wire through the hole in the solder tag. Crop the wire with cutters if needed…

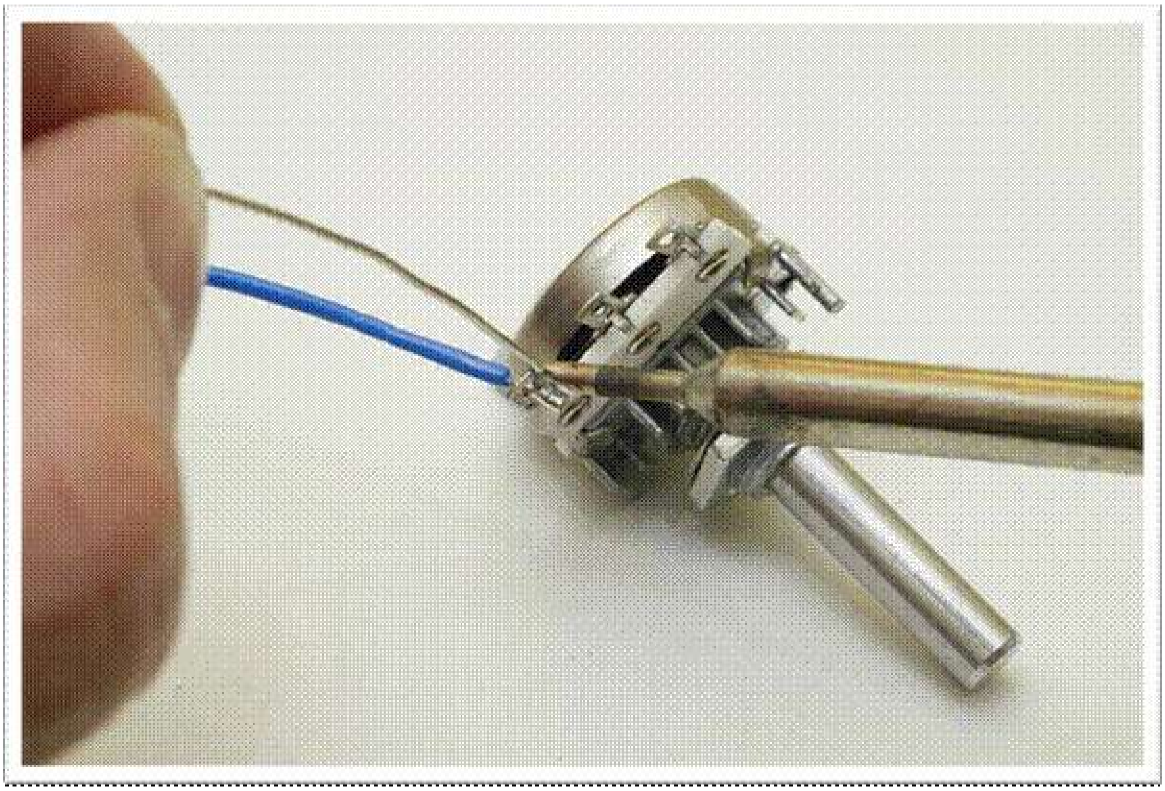

^^ … and apply the iron to heat the joint. As there's quite a lot of metal to heat, allow several seconds to heat it up and then simply solder the wire and solder tag together with a dab of solder wire.

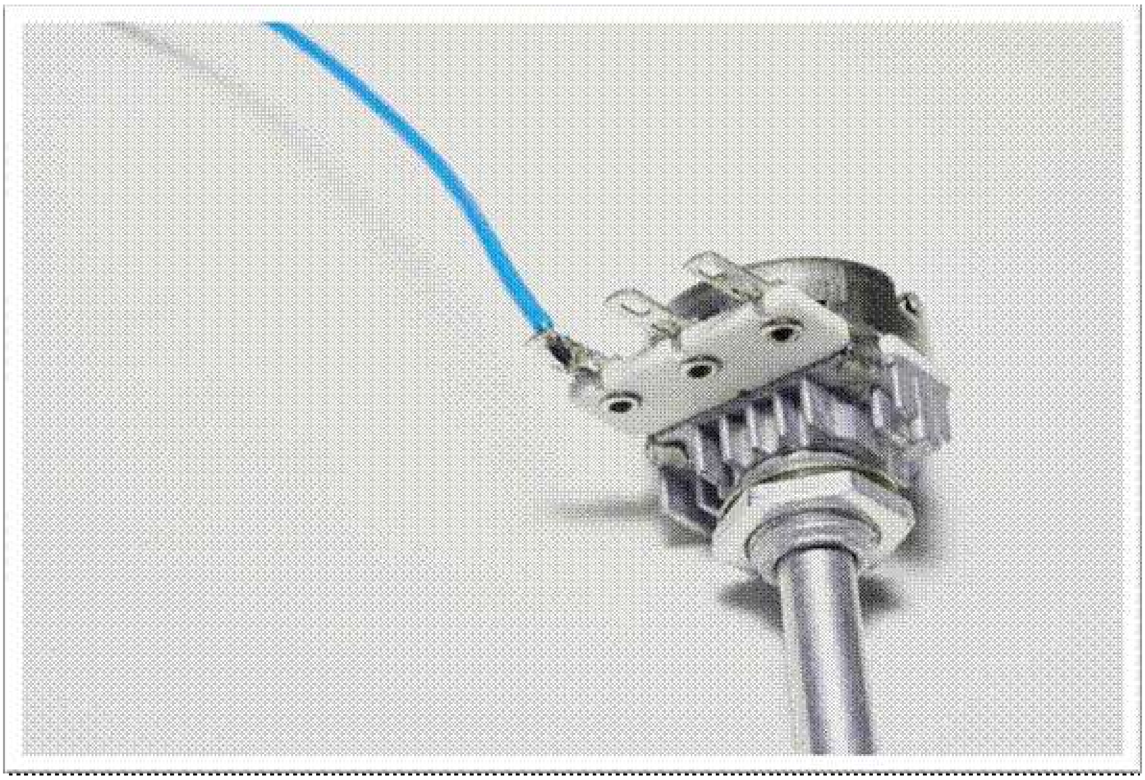

^^ The result is a perfectly satisfactory solder joint.

The second way is to loop the **untinned** wire through the tag a few turns and then solder it. This secures the wire during soldering, but it's messier to desolder if things go wrong (see later):

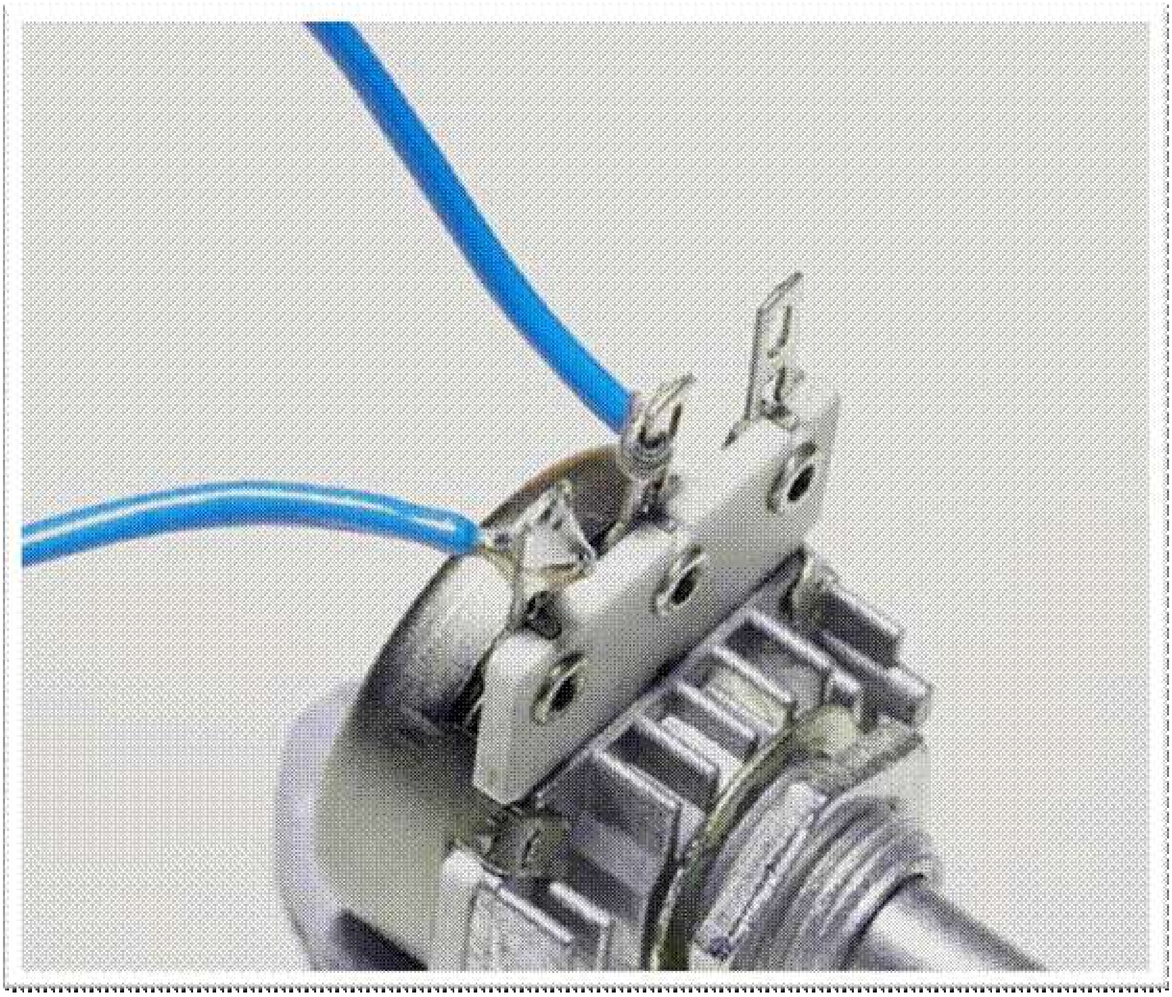

^^ The centre wire (potentiometer wiper terminal) has been wrapped around to secure it before soldering.

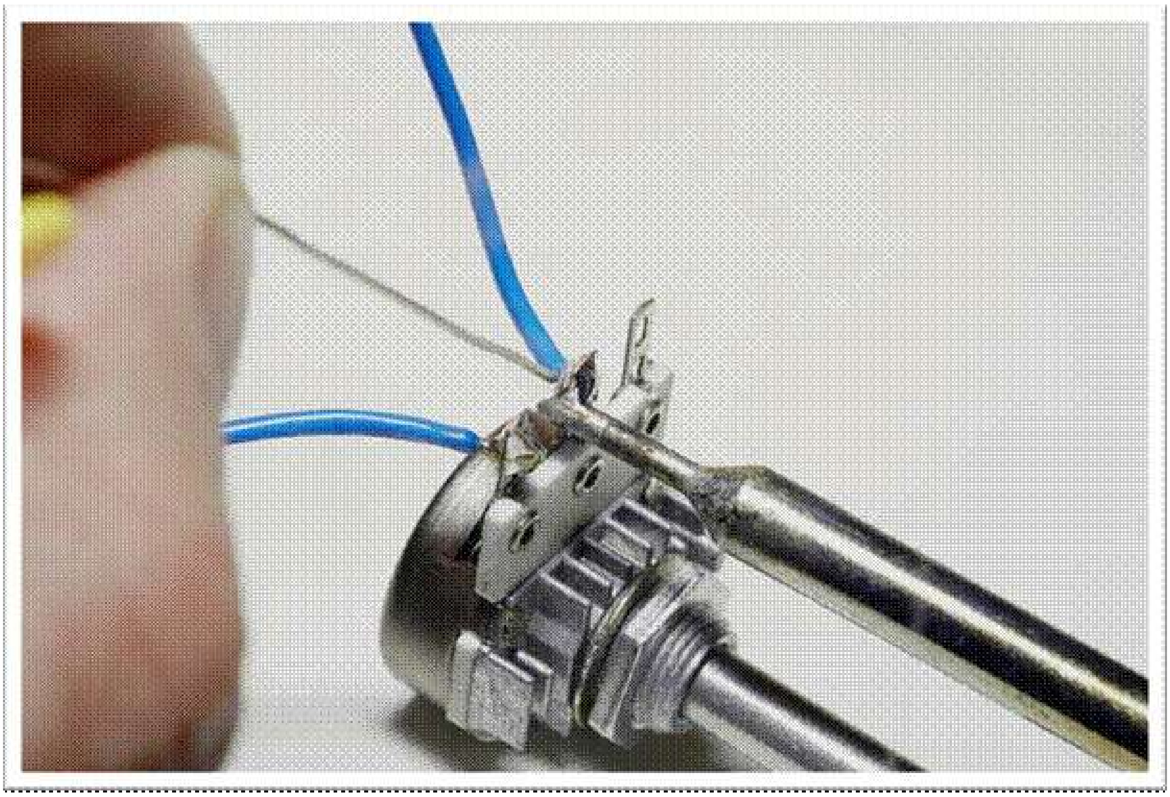

^^ Solder the joint in just the same way – heat it up and dab on some solder wire.

Much electronic equipment is connected or inter-wired this way and soldering everything together is a key stage in assembling any electronic project. I’ll explain some refinements you can try, later.

Wire joints

Soldering isn’t always the best solution to some problems, for reasons that I explain in the next section called Fatigue and Breakage. However you can solder wires together to join them simply and cheaply, and next I show a classic wire joint of this kind, sometimes called a **Western Union** joint.

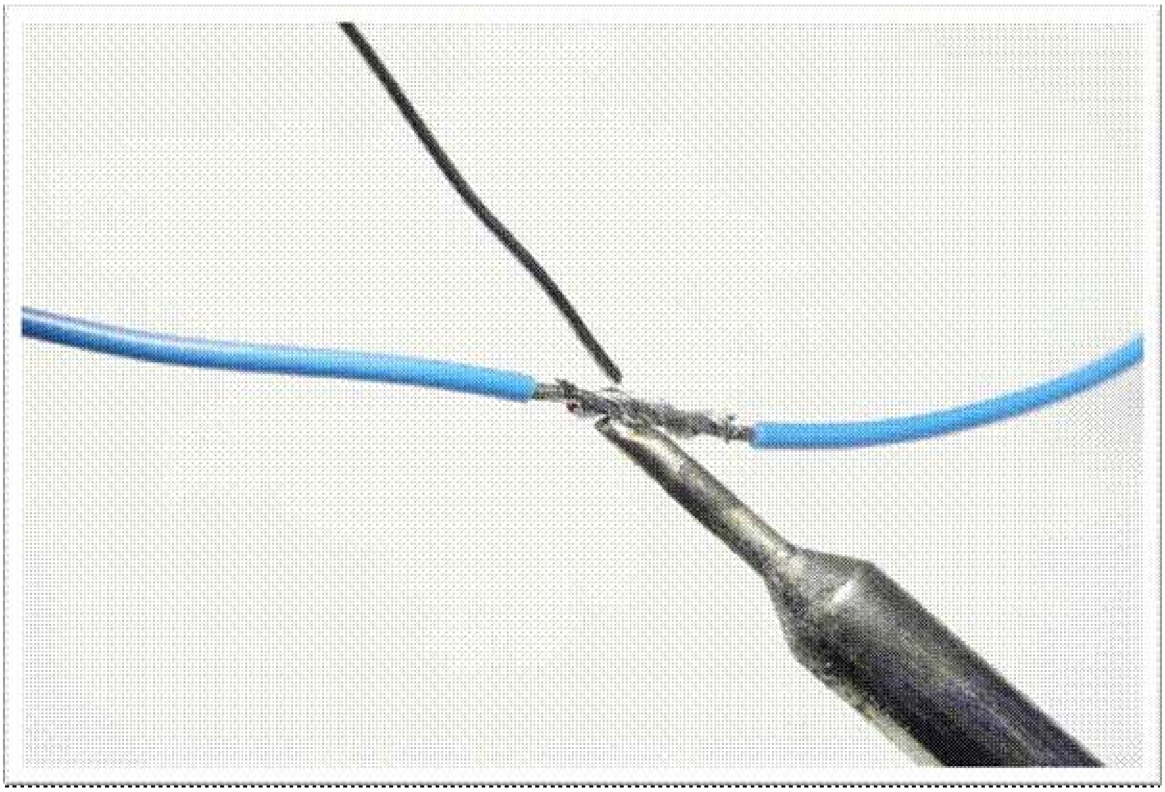

^^ *Soldering a Western Union wire joint.*

Strip the ends from each wire using wire strippers, and try not to cut any of the stranded cores or this will create a weakness. (Stray strands of wire can also be a hazard.) Then twist the wires securely together several times, ideally four or five. Apply a soldering iron to the exposed joint and dab on some solder until it flows fully over the joint, then remove the iron and allow the joint to cool.

The wire’s plastic insulation may shrink back a little due to the heat, but try not to overheat it excessively. If you have the resources you can finish off by insulating the joint with some heatshrink tubing (slide it over a wire before you make the solder joint!). Otherwise use ordinary PVC tubing or insulation tape if you have any.

A solder tag can be soldered easily enough, and it’s probably best to feed a stripped wire end through the tag and twist it over several times before soldering it, so it’s held together when soldering.

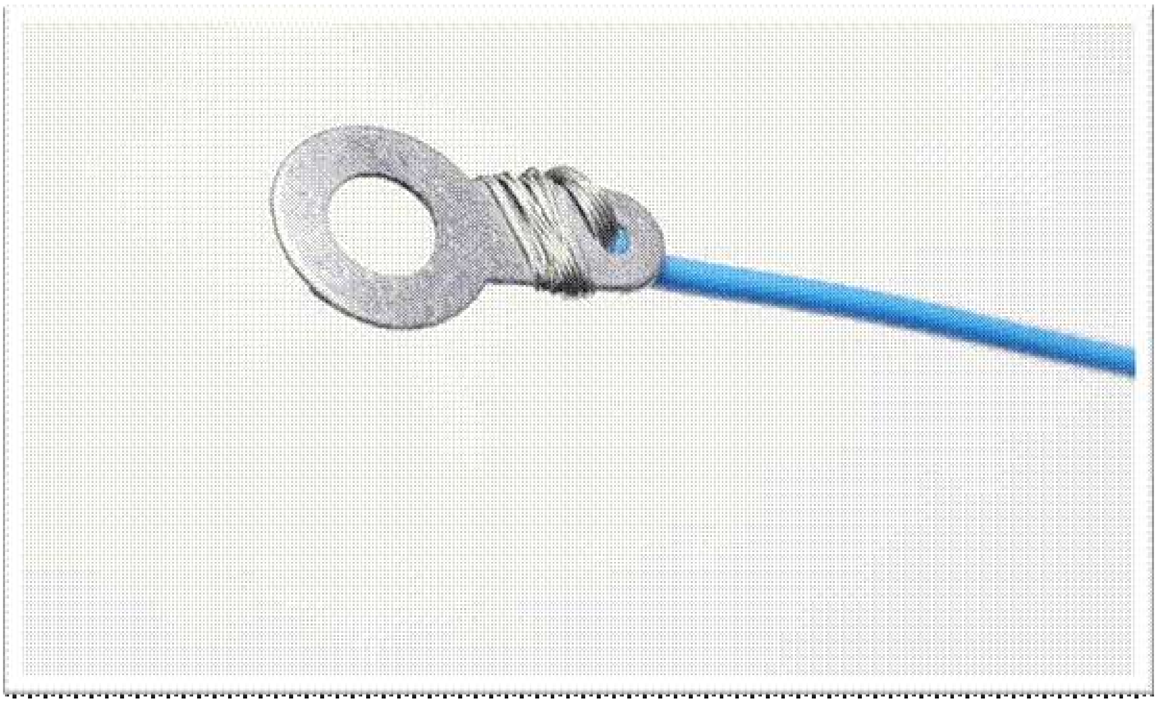

^^ A solder tag is soldered by wrapping a stripped wire end through the hole and round again, to secure it.

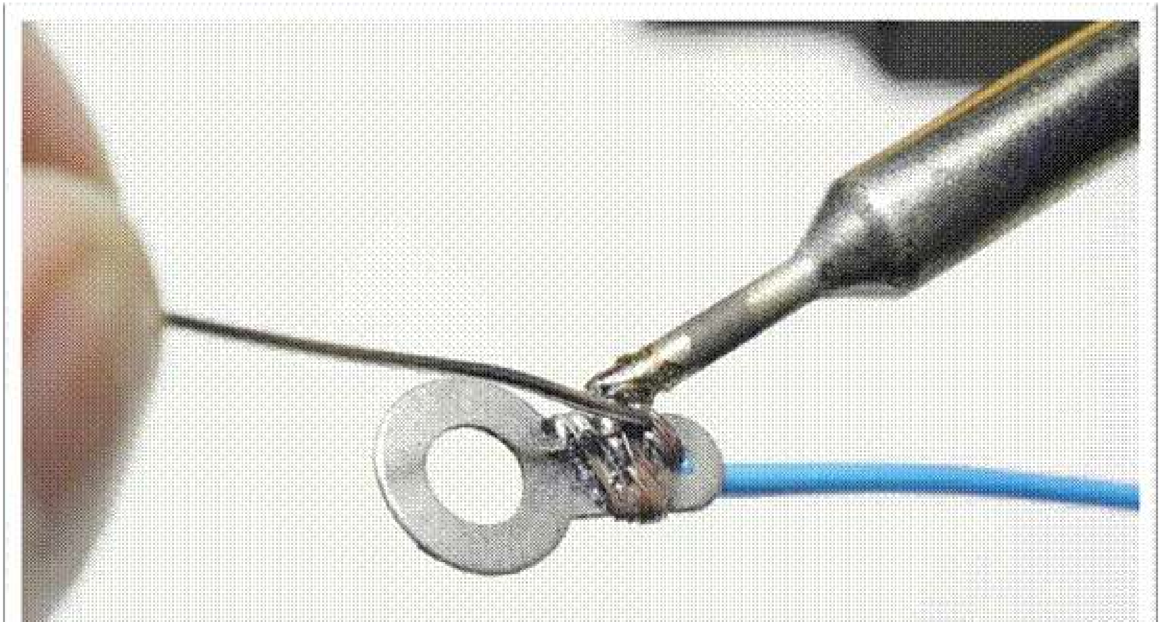

^^ Apply plenty of heat and solder for a few seconds.

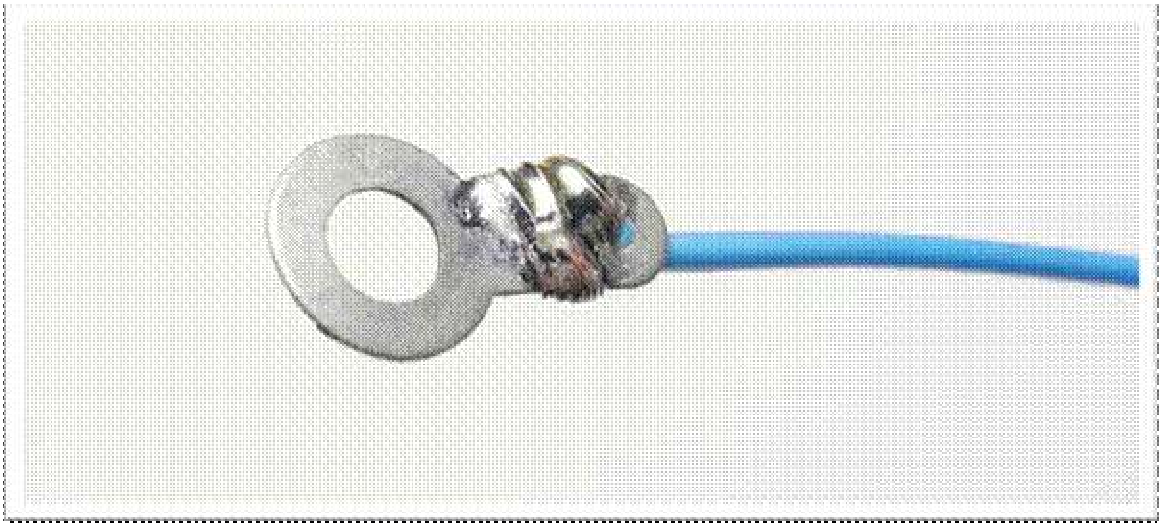

^^ The completed solder tag

The next sequence show how the same basic principles of soldering wires are used to connect up a device such as a panel-mounting l.e.d.. A series resistor can be wrapped around and soldered directly to one of the l.e.d. solid wires, and then two multi-stranded connecting leads are soldered to the l.e.d. and resistor. It's best to use heatshrink tubing or PVC sleeve to insulate the solder joints and prevent short circuits afterwards.

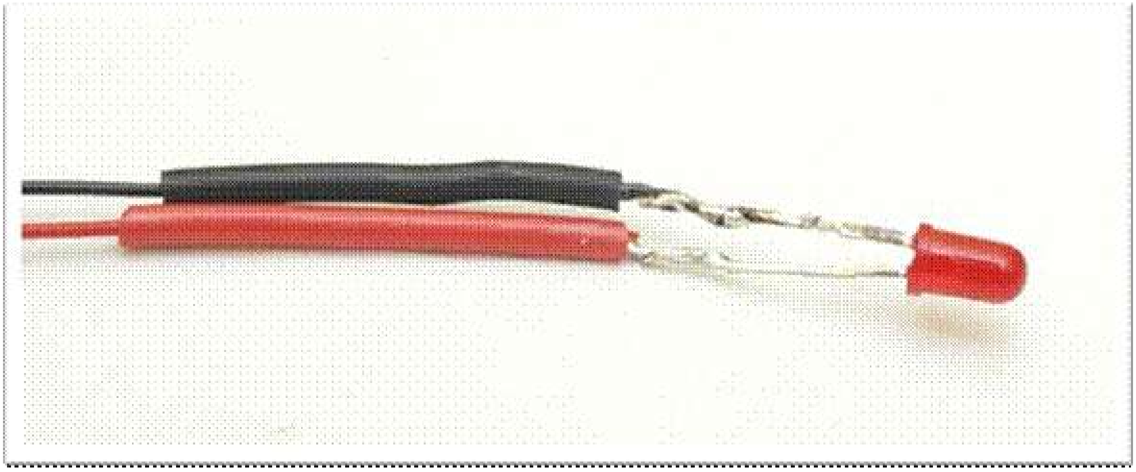

^^ Multi-stranded wires can be soldered to the solid leads of an l.e.d.: insulate them afterwards. Different coloured wires identify the polarity of anode and cathode.

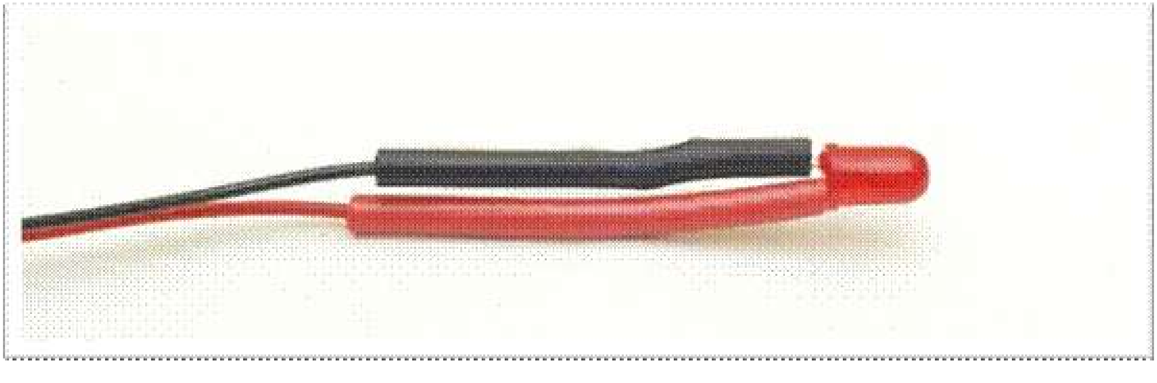

^^ PVC sleeving or heatshrink is needed to prevent short circuits.

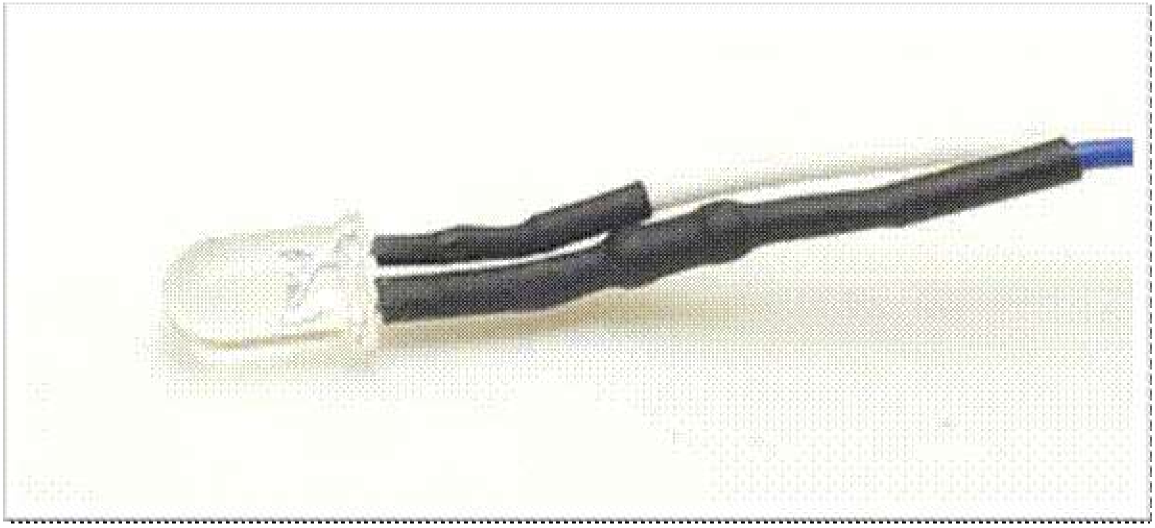

^^ *A commercial l.e.d.– the series resistor can be made out in the sleeving.*

Wires can also be soldered directly to printed circuit boards by stripping and tinning the ends, provided they fit through the holes in the p.c.b. It's a very common and cheap way of hooking up a board using "flying leads" and you'll see this all the time in consumer electronic equipment. For convenience so-called "solder pins" can be used, onto which flying leads may be soldered from the component-side instead.

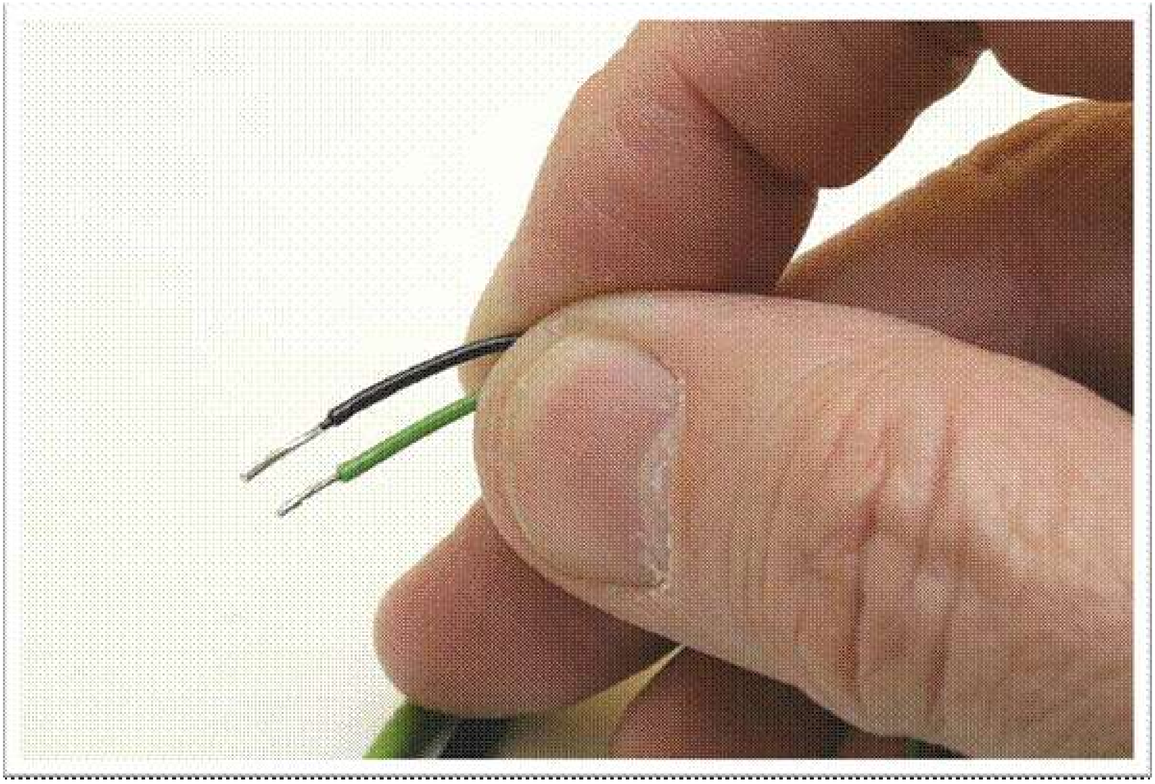

^^ *A set of l.e.d.s connecting wires, stripped, tinned and ready to be soldered to the p.c.b.*

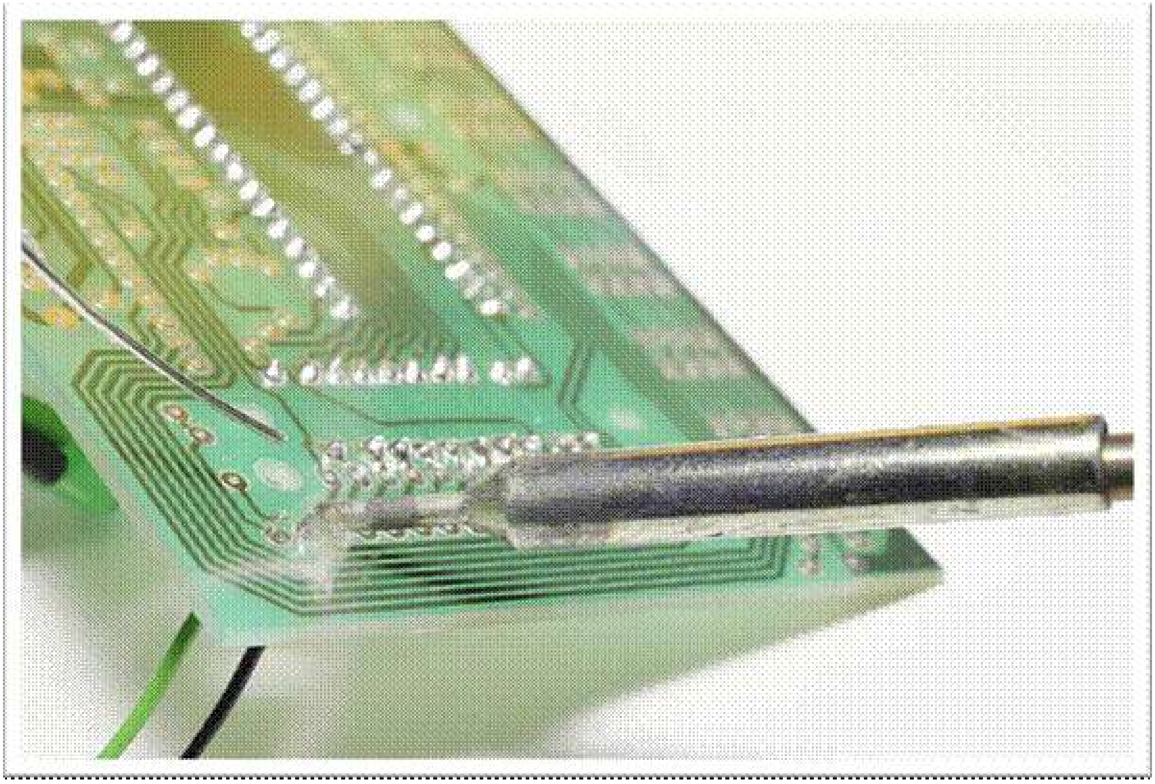

^^ Solder the two "flying leads" to the copper pads in the usual way. Hopefully the holes are big enough for the wires!

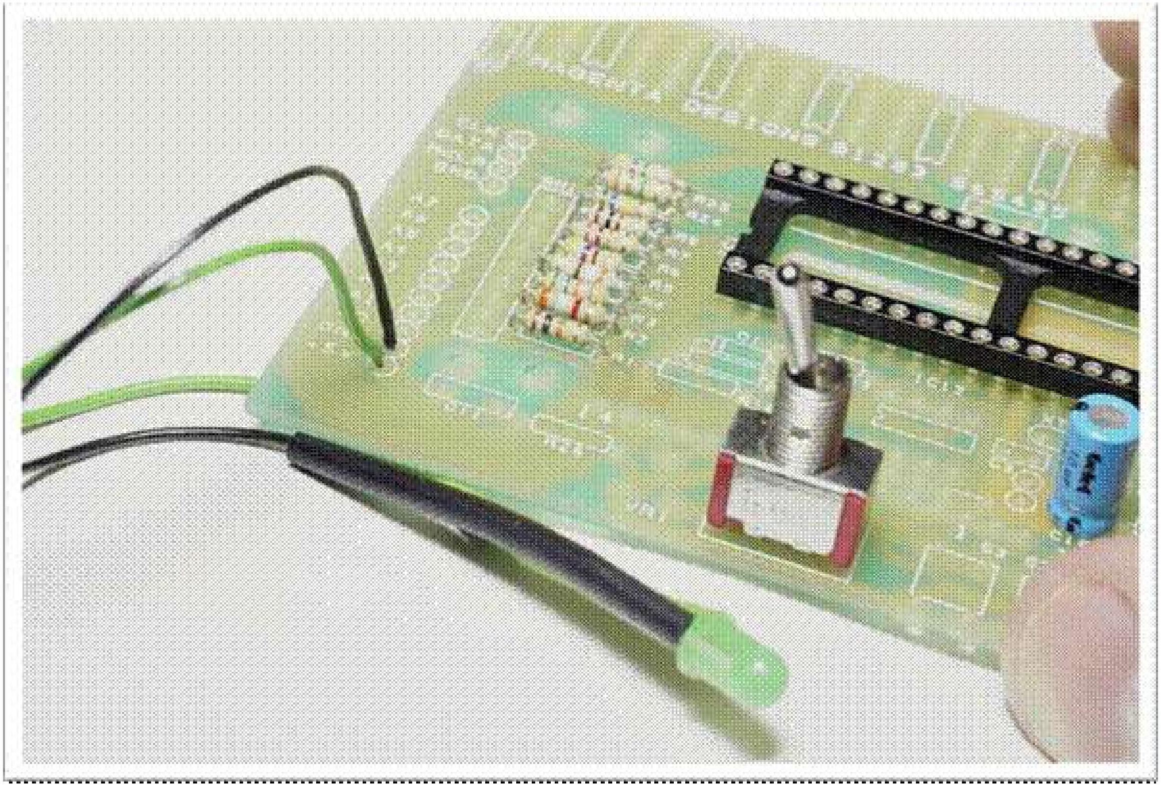

^^ Connecting an l.e.d. to a circuit board with flying leads.

Wires can also be attached to the underside (solder-side) of circuit boards, tacking them onto existing solder joints by re-melting them and absorbing the tinned wire end into it ("reflow soldering" – see later). This is a cheap and cheerful, semi-reliable way of rigging wires to a circuit board, and it's used all the time in imported consumer electronics.

Tidying up

I’ll leave it to readers to decide whether they should cut off excess wire before or after soldering. After the soldering is complete, I prefer to tidy up the joint by snipping any excess wire from the joint using a pair of “end cutters” shown earlier. These expensive hand tools have specially angled blades that snip the joint neatly down to the top of the solder joint. Ordinary side cutters will do fine.

It’s worth taking time out to inspect the work closely, looking for any missing solder joints, whiskers of solder or swarf shorting out any solder pads, and all such potential problem areas should be dealt with prior to testing the board. Faultfinding goes beyond the scope of this guide, but it’s true to say that almost always, any problems noticed after powering up the circuit are due to soldering faults or wrong components being used.

Reflow technique

Another technique often used is “reflow” soldering. This is used to “tack” devices or wires together, especially if very small, sensitive or fiddly parts are involved. There might be no room to make a “proper” sturdy joint, or it might just not be necessary to have any mechanical strength in the joint, especially if tiny parts are used.

As an example, imagine a small temperature sensor (I used a transistor) for use in a thermometer project. It could be quite tricky to solder flying leads onto the sensor’s leadouts, so a good approach is to tin both the flying leads and the sensor’s leads, and then simply touch them together and re-melt the solder with the iron. There’s no need to add any more solder, because the solder that’s already there will re-melt and the joint will be made. Sometimes this is called a butt joint.

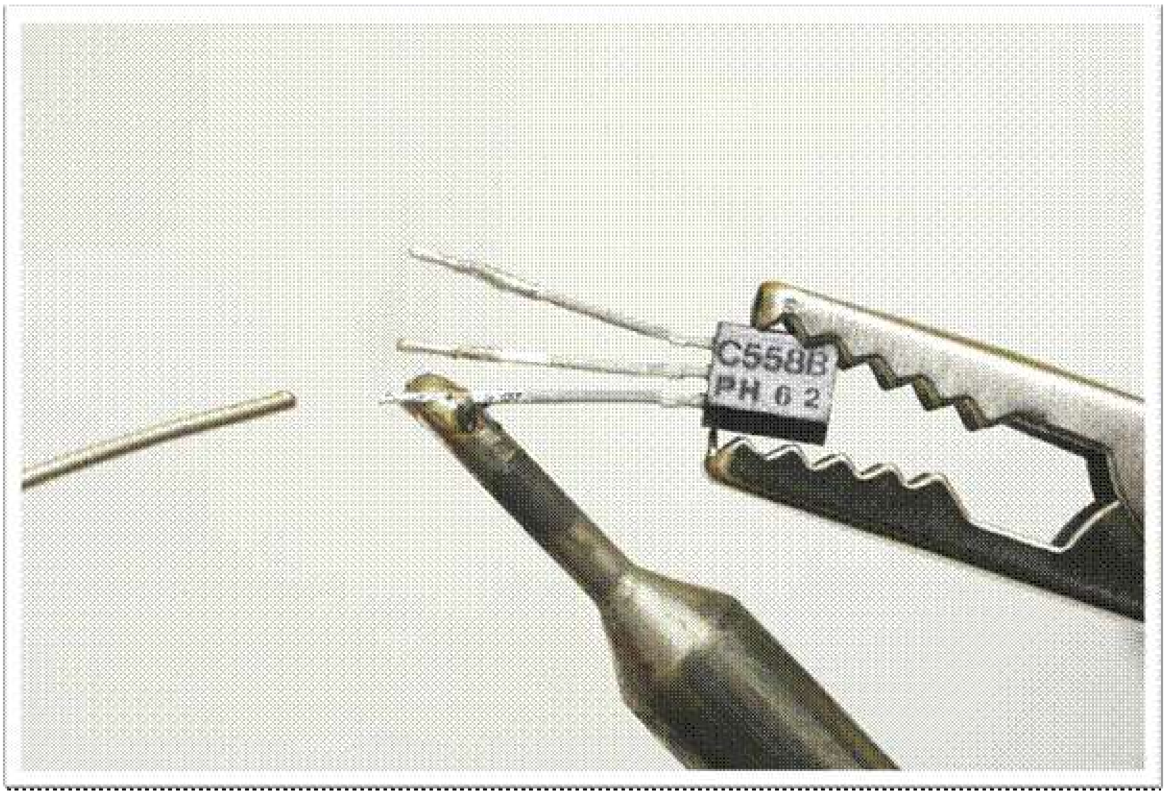

^^ *The leads of this sensor (transistor) have been tinned, ready for flying leads to be tacked onto them with a reflow method. A Helping Hands jig might help!*

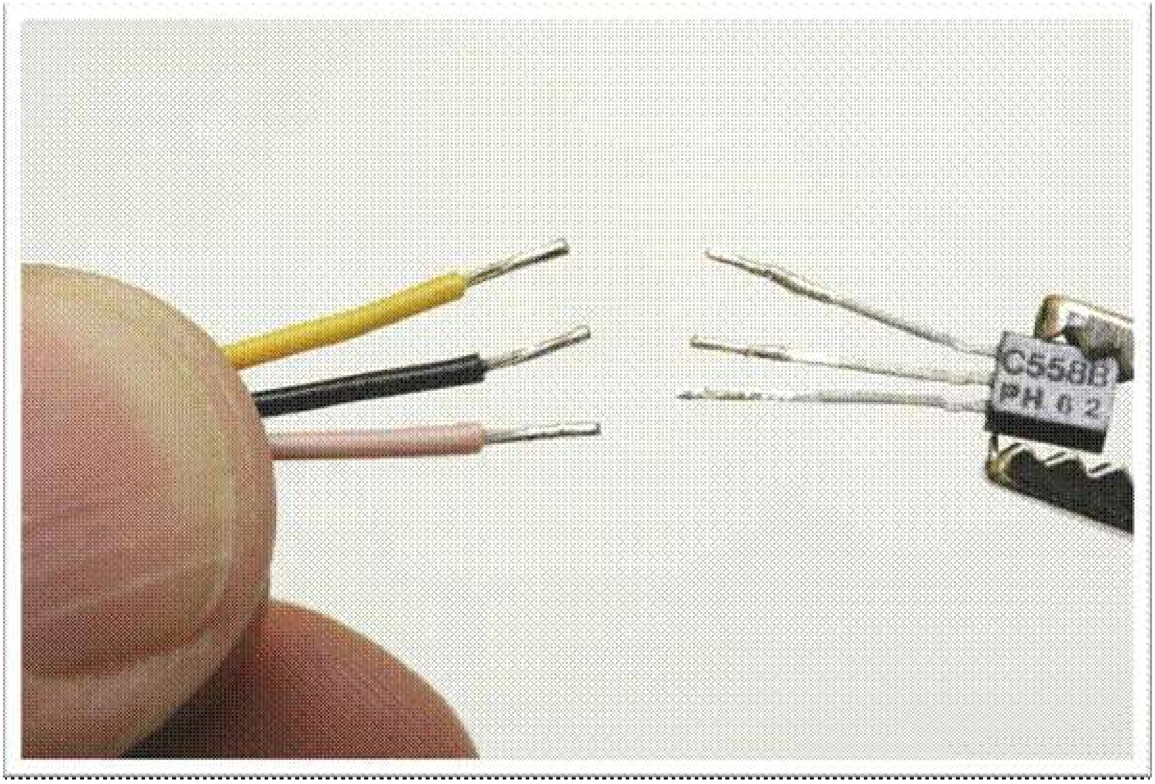

^^ The three flying leads stripped, tinned and ready to be reflow-soldered onto the device.

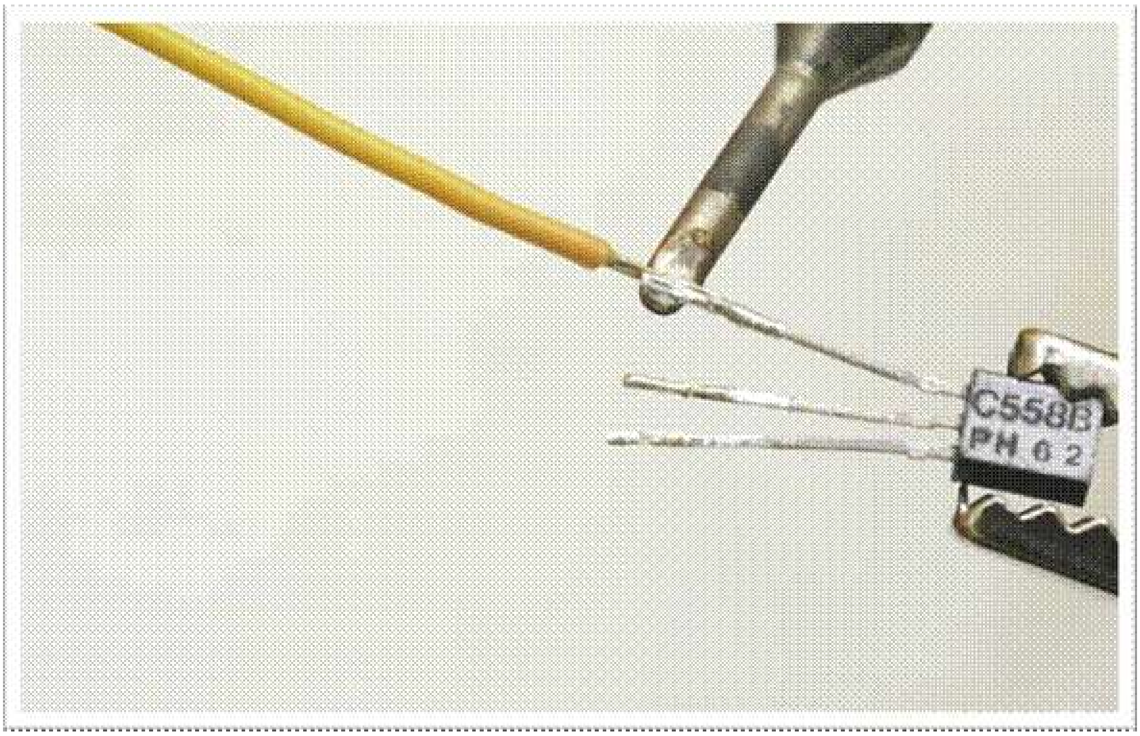

^^ To reflow solder them, simply hold the two leads together while re-melting the solder with the iron.

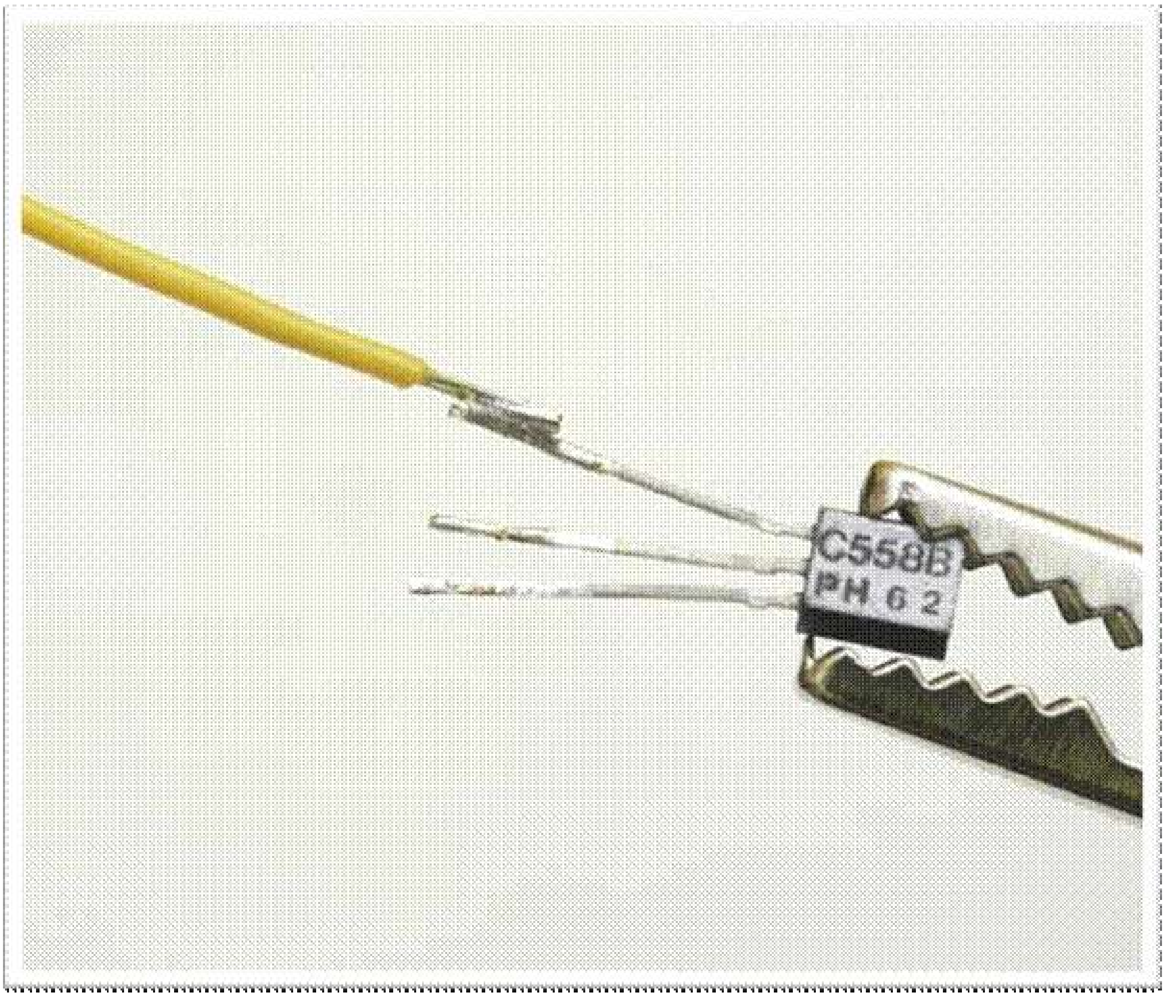

^^ Remove the iron and let the solder cool. The wire is tacked on.

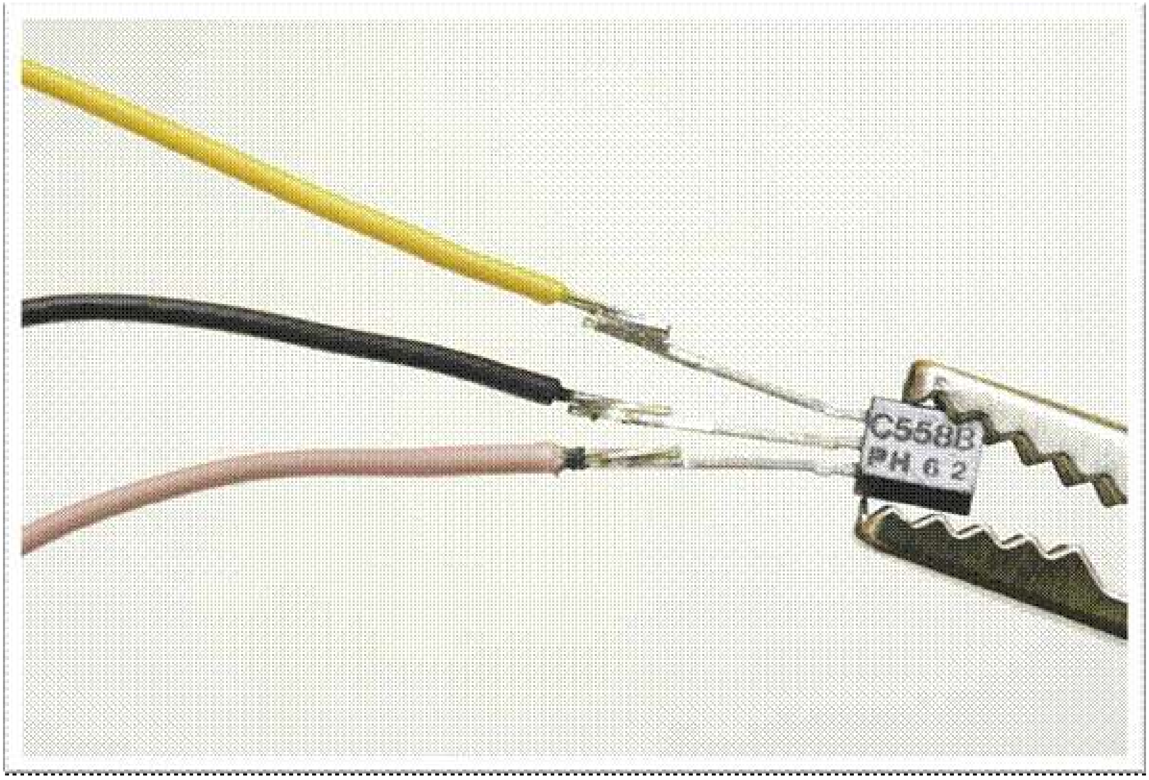

^^ Repeat for the other leads. They just need insulating with sleeving, then job done!

Depicted next is a typical "tag strip", an insulated panel with metal solder tags used for making sundry connections. (Entire TVs and radios used to be hand-built with them, in the early-mid 20th Century!) The principles of soldering are exactly the same, but more time is needed when applying the soldering iron because more metal is present, which needs more heat. You'll need more solder for bigger joints like these, so larger diameter solder makes a quicker job of it. Consider adding more flux (see earlier) to see if it helps.

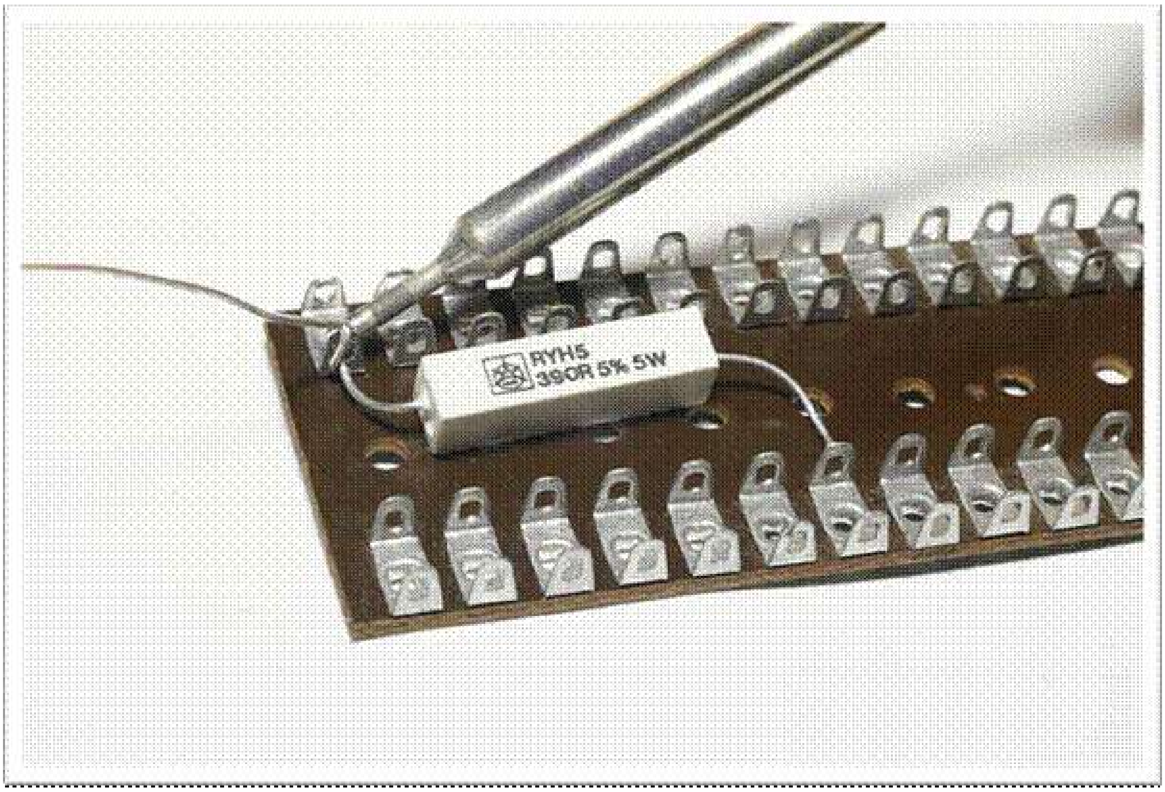

^^ Wrap wires through the terminals and arrange everything neatly, then solder as normal.

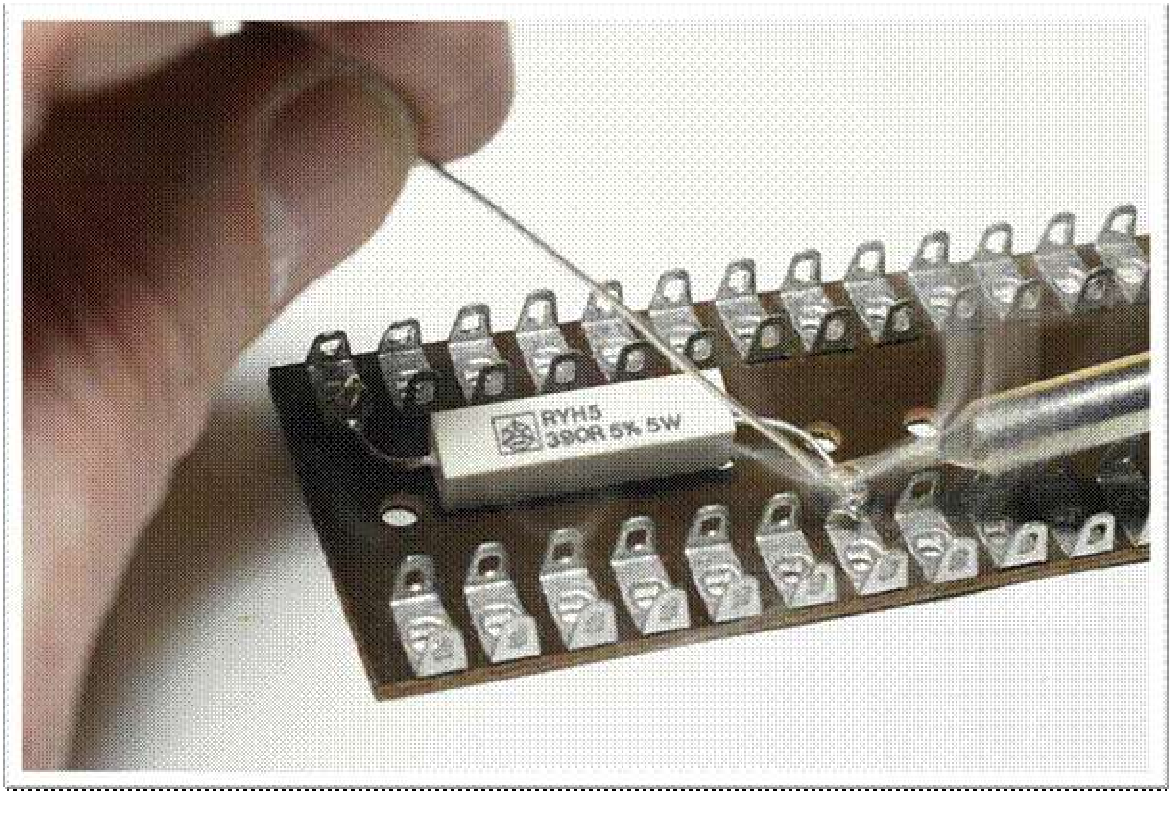

^^ Don't be afraid to apply more heat with larger assemblies like these: they contain more metal than, say, an ordinary p.c.b so allow more time for the solder to flow over

everything properly.

Fatigue & Breakage

Earlier I explained how multi-stranded hookup wire is flexible, which also makes it vibration proof – that's why a car's electrics are full of multi-stranded wiring, and test equipment probes use ultra-flexible multi-stranded wire for the same reason. Solid-core wire can be bent and will stay in shape (called "plate wiring") but if it's repeatedly bent or vibrated then it may eventually break somewhere due to fatigue.

The same is true of wires that have been soldered or tinned. No longer is the wire 100% multi-cored and flexible – instead it's been turned into a *single core wire* at the point where it's been soldered. This is potentially a weak spot and could eventually fracture due to fatigue, if subjected to continued vibration (e.g. in a car engine bay or in motorised equipment).

In a lot of equipment problems can be avoided by adding *strain reliefs* of some sort, to stop the wire being flexed where it's been soldered. Heatshrink tubing, or a dab of hot-melt glue, are ways of taking the pressure off the joint and ensuring soldered wires won't snap off due to vibration.

One of the reasons that higher-quality equipment and cars etc. use crimp terminals and connectors is that crimped (as opposed to soldered) connections retain all the flexibility of multi-core stranded wire from end to end, avoiding problems of wires breaking off. A *ferrule* is a very neat way of tidying wire ends and preventing stray strands of wire doing damage. Ferrules can be used to connect wires into screw terminal blocks etc. Simply clamp the terminal block down onto it and the screw will grip the ferrule.

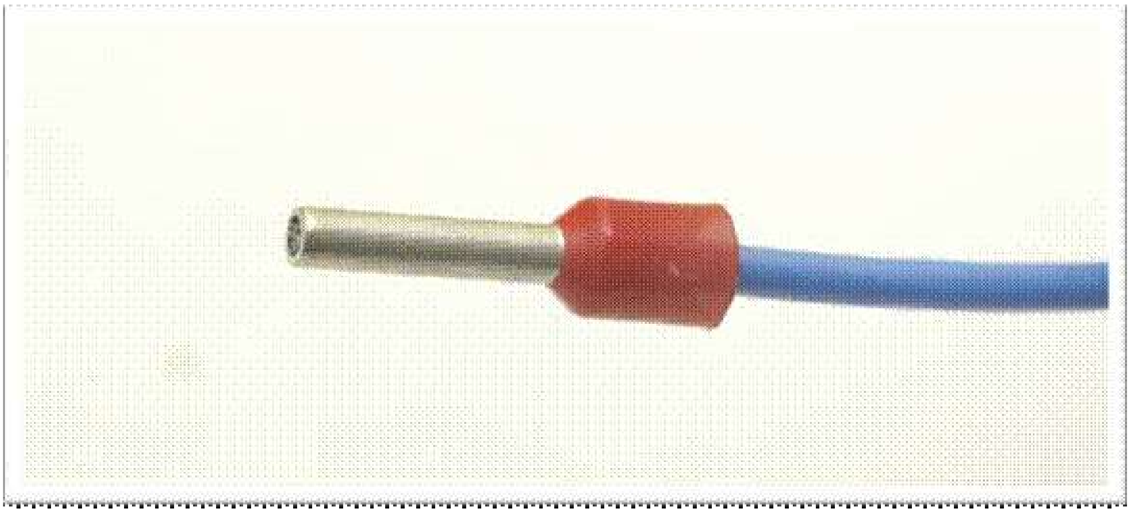

^^ Instead of soldering them, wires can be terminated with ferrules prior to fitting to e.g. screw terminal blocks. This makes them vibration-proof and also avoids any problem of stray wires poking out.

Faults & Desoldering techniques

By putting into practice the guidelines in my *Basic Soldering Guide* practice, there's no reason at all why you should not obtain perfect results and eliminate any potential problems. Hopefully the information gives you plenty of guidance to tackle various soldering projects with confidence. There's no substitute for getting some hands-on experience though, so I'd repeat the advice to try assembling a simple high quality electronic kit or two, such as those produced by Velleman and see how you get on. Powering up your first project successfully is a great thrill as every electronics hobbyist knows.

^^ *Popular desoldering products, a pump and various widths of desolder braid.*

Let's now look at reversing the soldering procedure – what to do if things go wrong, or maybe you have to repair a circuit by replacing a faulty component.

A solder joint which is badly made is likely to be electrically "noisy", unreliable and will probably worsen over time. Expansion and contraction of the joint due to heating and cooling can also throw up intermittent problems later down the line. The joint may look OK but underneath it may have a poor electrical connection, or could work initially and then cause the equipment to fail at a later date! These intermittent problems can be maddening to fix. TV repair technicians have an uncanny ability to go straight to a faulty solder joint because they see the same problem all the time, especially on equipment that has a "reputation".

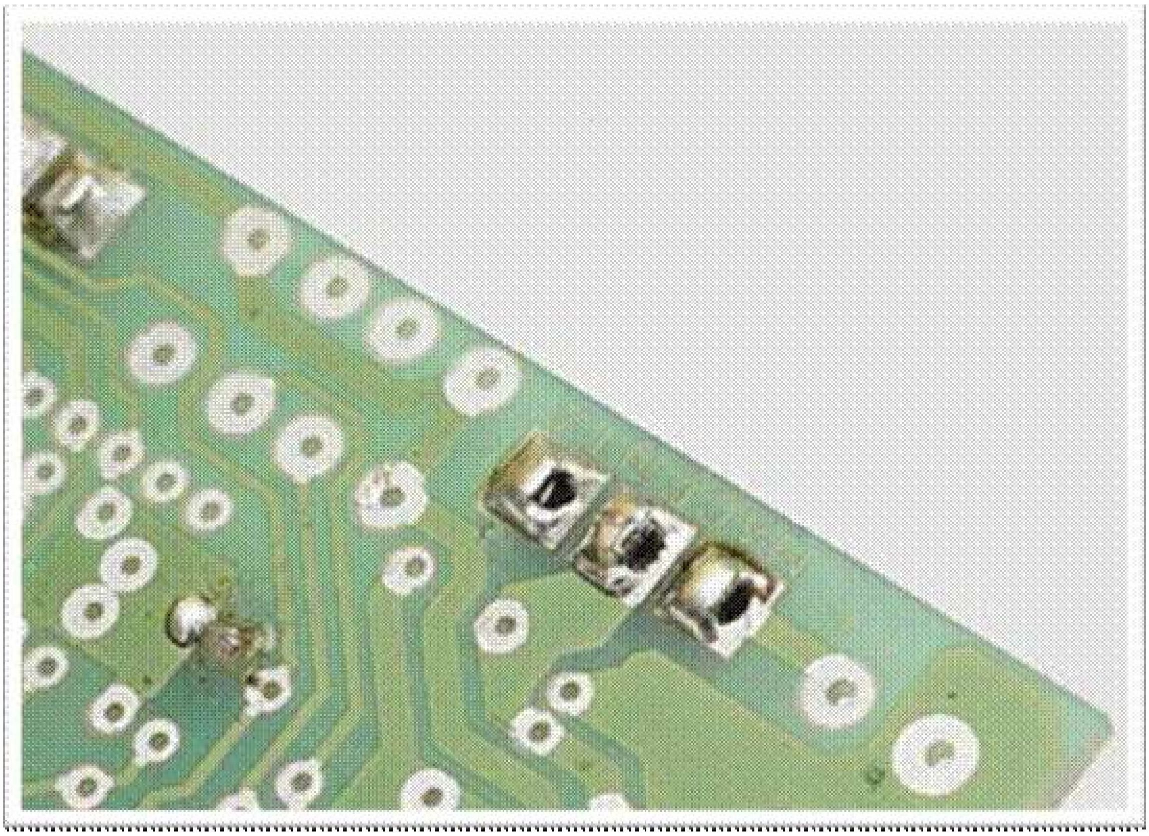

^^ These dry or grey/ gray joints were caused by dirty parts, poor heating and inadequate solder coverage

A solder joint that's poorly formed is called a "dry joint" or "cold joint" or a "grey/ gray" joint. Usually it results from inadequate heating, dirt or grease preventing the solder from melting onto the parts properly, and is often noticeable because the solder tends not to wet the surface properly. Instead it forms beads or globules. Alternatively, if it seems to take a very long time for the solder to wet the joint, that's another sign of contamination and that the joint may be a dry one, or the material is incompatible anyway. A matt, crystalline appearance instead of a shiny joint points to inadequate heating: the solder cooled down far too quickly and didn't flow properly.

Whether you want to replace a faulty component or fix a dry or poor-quality solder joint, it's usually necessary to remove the troublesome solder and then re-solder it afresh. Naturally, there are tools and techniques that make the job easy. It's very bad practice to simply re-melt the joint and then lash out with the board, whiplash style, hoping that the molten solder will be flicked off the board.

The usual way of removing solder from a joint is to use a *desoldering pump*. These work like a small spring-loaded bicycle pump, only in reverse! A plunger is pressed down until it locks into position. It's released by pressing a button which sucks air back through a pointed nozzle, carrying any molten solder with it. It may take one or two attempts to clean up a joint this way, but a small desoldering pump is an invaluable tool especially for p.c.b. work and they are widely available now.

^^ *A hobby desoldering pump primed for use*

Desoldering pumps often have a heatproof P.T.F.E. nozzle which may need replacing occasionally. Each time the button is pressed, a plunger clears the nozzle but sometimes solder particles and swarf will be ejected in the process; when you prime the pump, point the nozzle into a small pot or old aerosol top to catch any debris. Remove the spout and clean out the pump from time to time.

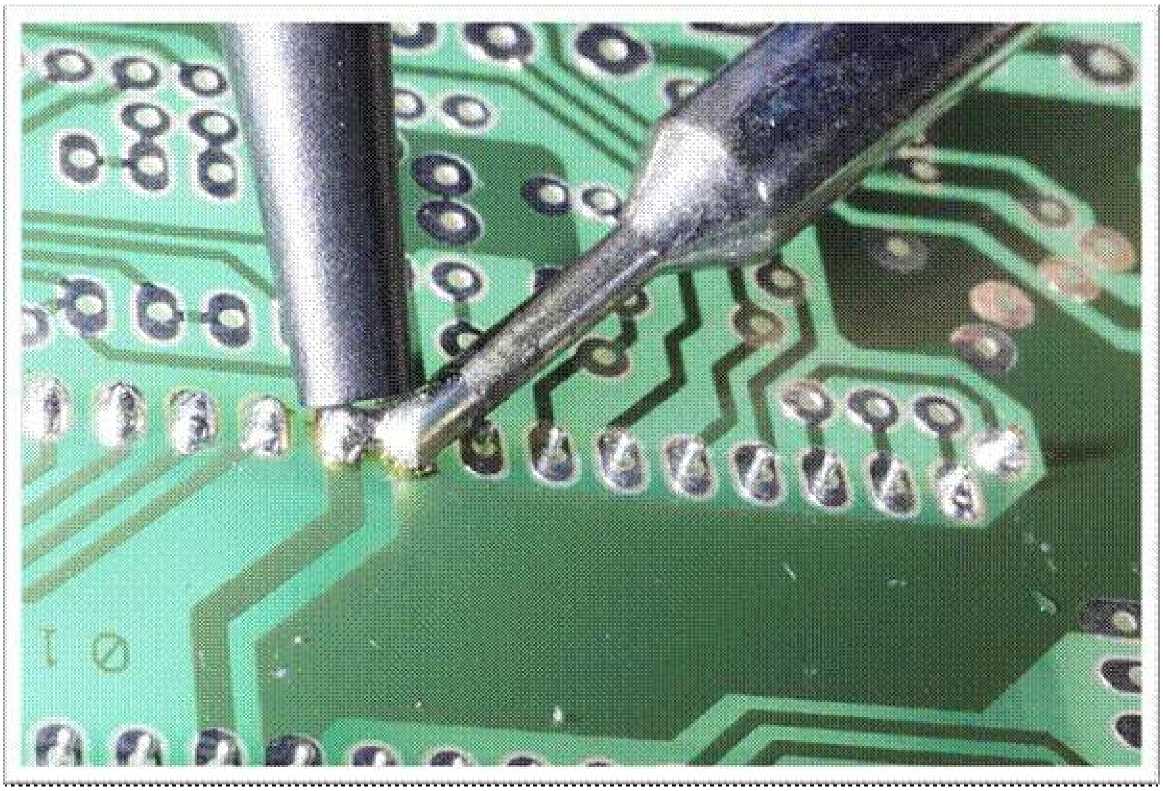

^^ *Suck up molten solder using a desoldering pump*

With very stubborn joints where the last traces of molten solder just can’t be shifted, it sometimes helps to actually *add* more solder and then desolder the whole lot again with a pump.

Desolder Braid

An alternative to a pump is to use *desoldering braid* which arrives in small dispenser reels. It’s a flux-impregnated fine copper braid which is applied to the molten joint, and the solder is then drawn up into the wick by capillary action. It’s remarkably effective and for certain tasks, it can be more thorough than a pump. I recommend that a small reel is bought (start with 1.5mm width) for the toolbox, to tackle larger or difficult joints which would take several attempts with a pump.

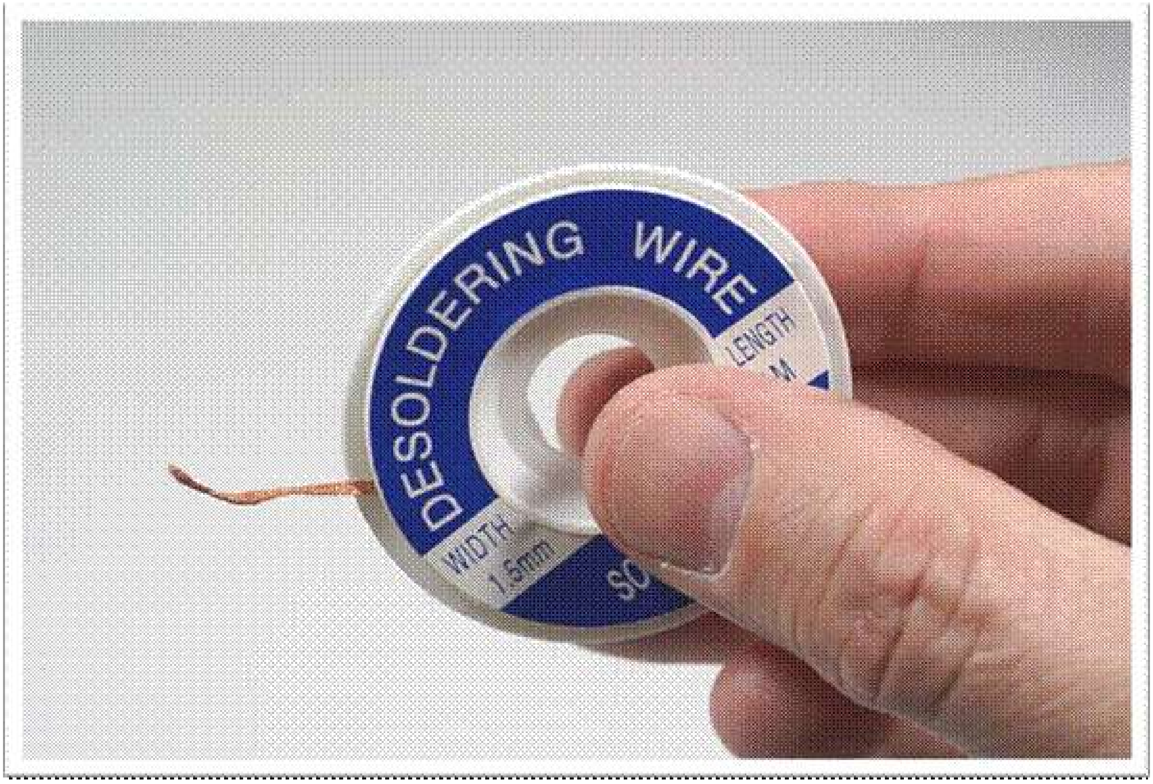

^^ *Desolder braid is also handy and can sometimes be more effective than a desolder pump. It comes in various widths to suit the scale of work being tackled.*

To use desolder braid, press the end of the braid down onto the joint using the tip of an iron, and let the solder melt underneath: the braid will then absorb the solder. The braid becomes hot so beware of burns. Once the solder’s solidified on the braid, cut it off and discard.

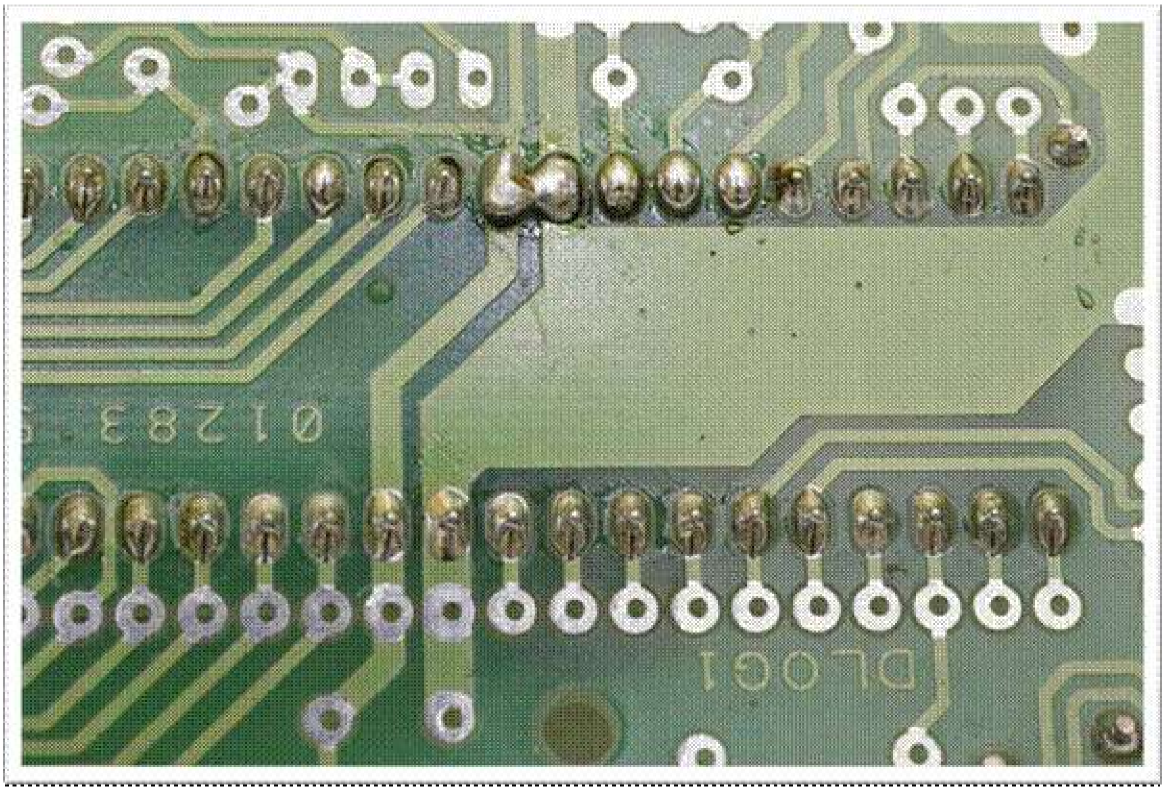

^^ Desolder braid can also be used to remove excess solder, e.g. two i.c. pins shorted together.

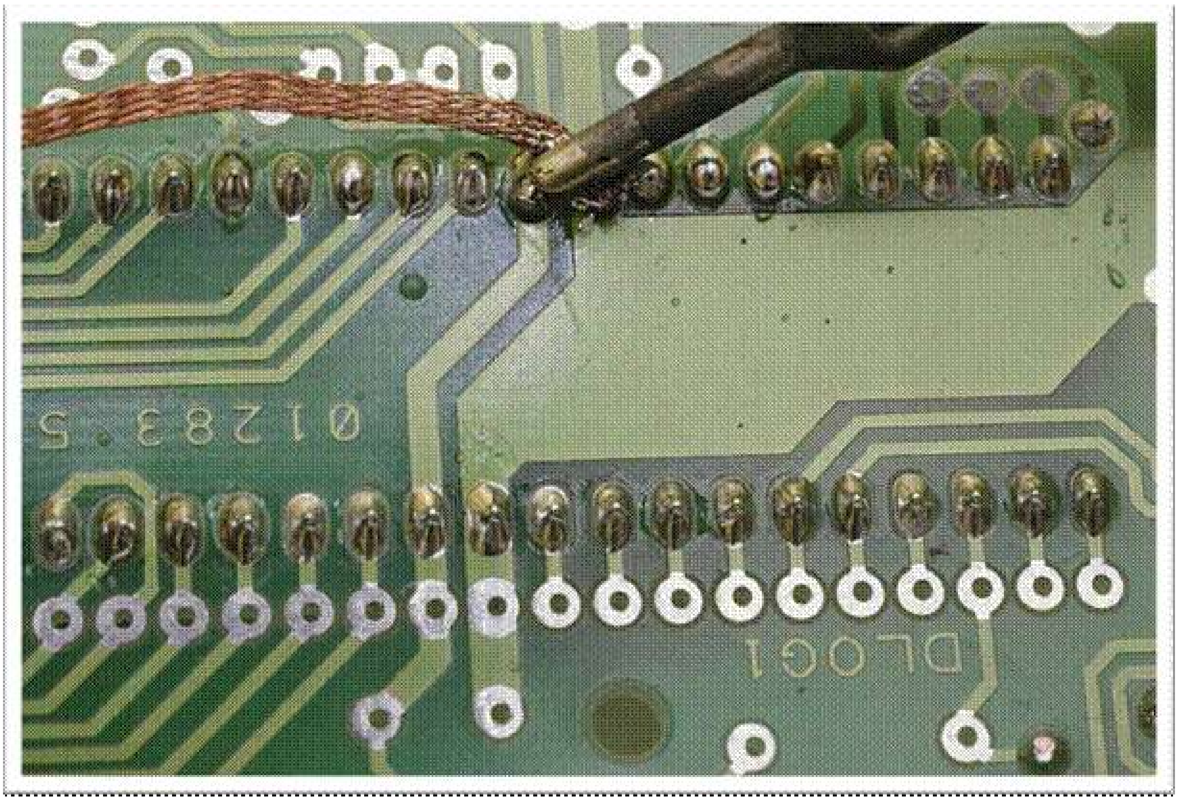

^^ Press some desolder braid over the joint, then melt it with an iron to draw molten solder up the braid.

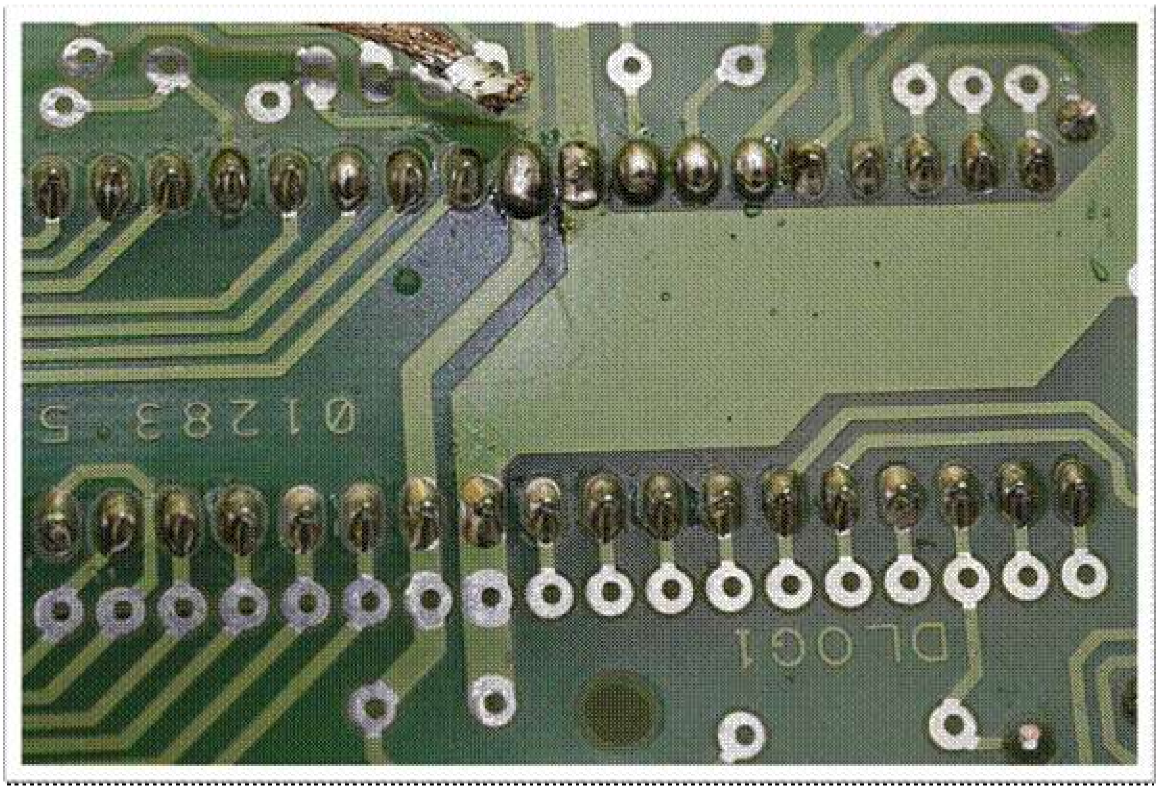

^^ *Remove the braid immediately and don't drag "whiskers" of molten solder around. The excess solder is absorbed by the braid, which is snipped off ready for the next job.*

Be aware that you can damage a printed circuit board accidentally when removing the desolder braid if it's not removed *quickly enough*. The solder will soon solidify, which effectively solders the braid to the printed circuit board! A careless tug may yank copper tracks or pads away from the board, stuck to the braid. You can also drag solder "whiskers" onto neighbouring pads unless the braid is removed cleanly. Why not practice using desolder braid with an old circuit board?

Whichever desoldering method you use, care is needed to ensure that the boards and parts are not damaged by excessive heat. It's not that difficult to apply so much heat when desoldering that the adhesive holding the copper foil tracks on the p.c.b. eventually fails, causing the copper track to lift away – everyone's worst nightmare.

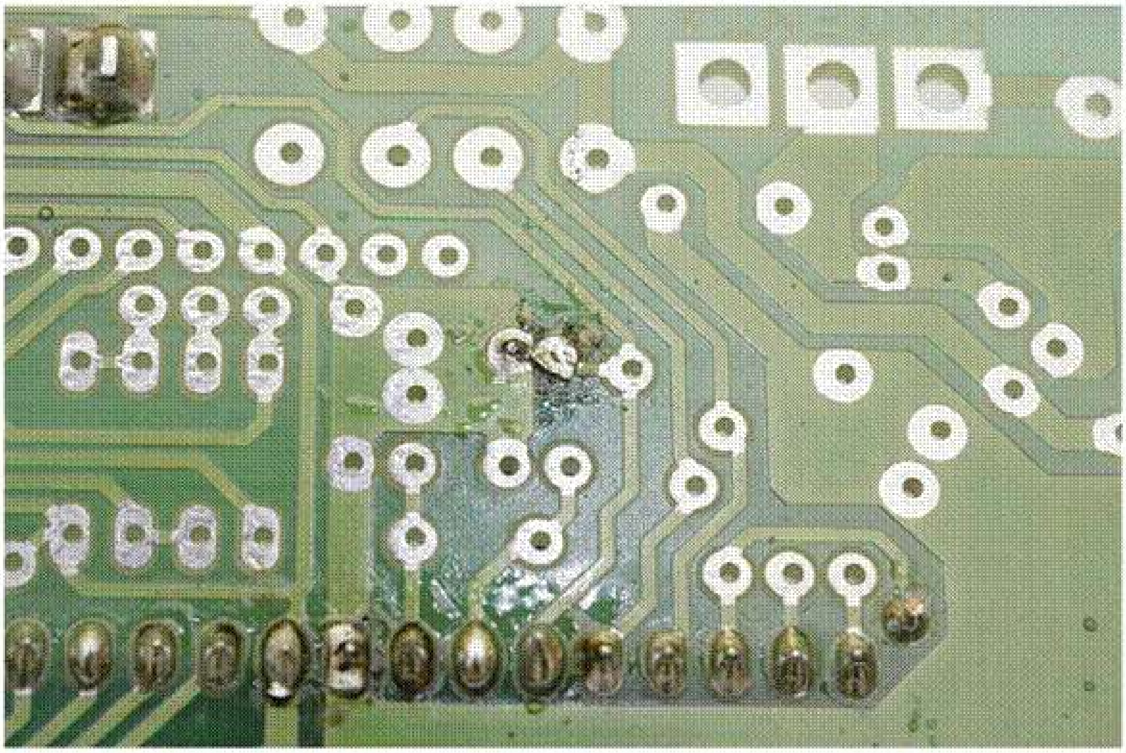

^^ Typical copper track damage (centre) caused by overheating during soldering or desoldering. The track has lifted off, but you can try repairing it by adding extra wiring or SuperGlue it if the track isn't broken.

If this should ever happen, remove the iron immediately and permit the area to cool (a freezer aerosol is valuable at such times). If you're lucky, you can maybe repair the lifted track using a droplet or two of Super Glue, or add "jumper wires" to bypass the damage.

Servisol
FREEZE IT 20

^^ A Freezer aerosol can give rapid cooling where excess heat has been applied

during soldering. Also used in circuit faultfinding to identify overheating parts.

You now know everything you need to know about making the ideal solder joint, and desoldering it in case you need to make a repair. Just to remind you, a Quick Summary guide follows.

Quick Summary Guide

To round off the Basic Soldering Guide, let’s summarise how to make the perfect solder joint.

- Ensure materials to be soldered are compatible with tin/ lead or lead-free solder.
- All parts must be clean and free from dirt and contaminants.
- Try to secure the workpiece firmly during soldering.
- Brand new soldering iron tips must be flooded with solder immediately, the first time they are used.
- Wipe the tip of the hot soldering iron on a damp cellulose sponge at frequent intervals. Then “tin” the iron tip by applying a small amount of solder.
- Aim to heat all parts of the joint with the iron for under a second or so, to bring them up to the same temperature.
- Continue heating and apply sufficient rosin-core tin/ lead or lead-free solder to form a complete joint.
- It only takes a second at most, to solder the average p.c.b. joint. It should be smooth and shiny, and through-hole joints should be slightly convex in shape.
- Remove the iron and return it safely to its stand.
- Do not move parts until the solder has cooled.
- Tin the soldering iron tip and clean it well, when switching it off, ready for next time.
- Consider using e.g. electronics flux dispenser pens or Colophony (rosin) to help with difficult joints.

Sometimes solder joints don’t go quite to plan, and sooner or later everyone is faced with the need to problem-solve or troubleshoot, so a simple Troubleshooting Guide follows next.

Troubleshooting Guide

This troubleshooting guide may help fix common problems encountered with troublesome solder joints.

SYMPTOMS	LIKELY CAUSES	REMEDY
Solder won't "take" (wet) and won't flow properly over the joint — molten solder forms beads or "ball bearings" instead of flowing properly.	Grease or contaminants present; Material may not be suitable for soldering with standard lead/tin or lead-free solder, e.g chromium.	Treat contaminated parts with abrasive cleaners etc. as required to expose base metal. Some metals can't be soldered with electronics-grade solder.
Solder doesn't melt or flow very well — the joint is crystalline or grainy-looking - a grey or dry joint.	Joint has been moved before being allowed to cool naturally, or: Joint was not heated adequately. Too large a joint – too much metal present – and/ or the soldering iron temperature or power rating are too low.	Desolder and remake. Apply heat for a longer period, or use a higher power soldering iron, or check the temperature setting and raise it if possible.
Solder joint forms a "spike" and applying the iron again makes it even worse!	Probably overheated, burning away the flux. The iron, when removed, would cause the solder to stand up in a spike.	It is usually best to desolder and remake the joint freshly again.
The copper foil of my p.c.b. has lifted off the circuit board!	Excessive use of heat has damaged the adhesive. Provided the track hasn't broken, it may be repairable.	You can sometimes repair it with Super Glue, or re-wire the board with jumper wires.
Brown varnish-like deposits are left behind after I finish soldering.	These are the remains of rosin flux and are nothing to worry about.	It can be removed with PCB cleaners or some solvents, if you want to tidy up the board and inspect your work.

Possible Hazards and simple First Aid

It's extremely rare that soldering iron operators receive any burns or other injuries from the use of hot soldering irons. Soldering is perfectly safe provided that common sense precautions are taken during the soldering operation. Here are some of them:

- Components are very hot after soldering, so let them cool before handling them to avoid skin burns.
- Beware of splashes of molten solder caused by careless handling of a hot soldering iron.
- Beware of energised components (capacitors, batteries etc.) being shorted by molten solder and ejecting solder splashes due to arcing.
- Always park a hot iron safely on a stand in between use — never hang it vertically next to the bench.
- Keep a hot soldering iron away from its mains cable (silicone cables reduce the risk of accidental damage).
- Beware of wire offcuts flying off (danger to eyesight) when snipping wires to length before or after soldering.
- Avoid inhalation of solder and flux fumes as this can irritate the respiratory tracts, especially in sensitive cases (e.g. asthma).

Should you receive a more serious skin burn which requires attention, then:

- **Cool the affected area immediately.** Use plenty of cooler running water – but avoid ice cubes etc. as they can cause nerve damage after a time or inhibit the flow of blood to the affected area.
- Remove any **objects** which may be constrictive, before any swelling starts (rings, watches, bracelets).
- Do not prick blisters nor apply ointments, salves or lotions at this stage.
- Local pain relief for small burns can be obtained by spraying *Burneze* aerosol onto unbroken skin.
- Seek medical attention for more serious burns.

Eyesight problems are exceptionally rare, e.g. pieces of wire offcuts or solder splashes lodging in the eye area, and should be treated by a qualified first-aider or A&E. The best you can do is bathe the affected area with e.g. a first-aid eyewash bottle or fresh water. Then seek professional medical help straight away.

Useful resources

Details of the Antex range of soldering equipment, solder tips and spare parts can be obtained from www.antex.co.uk. UK and international distributors are also listed on the Antex (Electronics) web site.

If you’re interested in hobby electronics then why not try **EPE Magazine**, the No. 1 UK magazine for hobby electronics enthusiasts, students, trainees and technicians around the world. You can buy a printed edition from newsstands, download a PDF version (for Windows) or try a tablet/ smartphone version from Pocketmags. More details at www.epemag.com

You can learn more about the writer at his website http://www.alanwinstanley.com/

Some mail order suppliers of soldering equipment and electronic parts include:

- ESR Electronic Components Ltd. (soldering equipment, Velleman kits)
- Bowood Electronics (Antex irons and spares, electronic parts)
- Rapid Electronics (components, tools, equipment)
- Cricklewood Electronics (CCTV, Antex soldering irons and spares, components)
- Maplin Electronics (UK electronics retailer and mail order)
- Farnell Electronic Components (major UK industrial supplier)
- RS (UK electronics industrial supplier)
- Brewsters Ltd (soldering equipment mail order specialists)
- Velleman UK (electronic kits)
- Quasar Electronics (Velleman kit mail order retailers)
- Kemo Kits (Germany, trade only)
- Hobbytronics (UK hobby mail order supplier)
- Multicore Solders (now a Henkel brand)

These links were correct at the time of going to press, July 2013.

Acknowledgements

Antex (Electronics) UK generously provided samples of their ever popular British-made soldering equipment, materials and accessories for use in the Basic Soldering Guide. When starting out as a 1970's teenage schoolboy electronics hobbyist, I and my trusty Antex soldering iron shared many adventures in electronics together, and it has been a pleasure to use the latest Antex soldering irons throughout this guide.

Also I would like to thank Brian Brooks of Magenta Electronics Ltd. (www.magenta2000.co.uk) for supplying me with a variety of their professionally-designed printed circuit boards and a range of components which were the subjects used in the photographs.

Photography

All photographs were taken by the author using Sony Alpha DSLRs and 50mm macro lens with extension tubes, and a Sony macro flash. High-resolution versions of these and more technical photos are available for commercial or educational royalty-free use. Please contact me with any enquiries.

Conclusion

I really hope the Basic Soldering Guide will give you the confidence to try your hand at electronic soldering. It's really a lot easier than it sounds, and armed with the guide's advice and photographs, the next thing to do is invest in a decent-quality soldering iron such as the excellent range manufactured by Antex that will serve you well for years to come.

Soldering an electronics kit together such as the professional designs produced by Velleman is a great way of testing out your new skills. Start with a small, simple kit costing a few pounds to gain confidence and experience, and avoid the temptation to tackle something too complex until you're ready to extend yourself further.

Feedback is welcomed by email to alan@epemag.demon.co.uk

You can learn more about me at www.alanwinstanley.com or read my column in EPE Magazine.

Good luck with your soldering!

Alan Winstanley

Text and Photographs © Copyright © Alan Winstanley 1997-2013

Play fair and support writers who share their skills and expertise

Preparing a resource like this takes months and is based on four decades of practical experience which I gladly share with readers. It costs a small fortune in time, experience, IT and photographic equipment to put an ebook like this together, and the Royalties that you pay to buy the book puts food directly on my table. You paid for your copy (thank you!) so if you see any unauthorised versions of my work anywhere in circulation, I'd really appreciate being told about it. Thank you! A.R.W.

Interested in electronics?
Read EPE - the No. 1 UK hobby electronics magazine.
Ask your newsagent, download or subscribe now!
www.epemag.com
EPE EVERYDAY PRACTICAL ELECTRONICS
EPE EVERYDAY PRACTICAL ELECTRONICS
DRIVEWAY SENTRY

BSG/ ARW/V1.1/0813

Project Book

MAKE A
WEARABLE
SOLDER SKILL
BADGE

Learn to Solder

Tools and Techniques for Assembling Electronics

Brian Jepson,
Tyler Moskowite
& Gregory Hayes

O'REILLY®

Maker
PRESS™

Project Book

Learn to Solder

Learn the fundamentals of soldering—and pick up an essential skill for building electronic gadgets. You'll discover how to preheat and tin your iron, make a good solder joint, desolder cleanly (when things don't quite go right), and how to use helping hands to hold components in place.

This concise book is part of MAKE's Getting Started with Soldering Kit. Using the tools in the kit and some electronic components, you can practice soldering while making fun blinky objects. Then show the world you just learned a new skill by wearing the Learn to Solder Skill Badge.

- **Learn how to prepare your workspace**
- **Get to know the components you'll work with**
- **Use the best methods for soldering components in place**
- **Experience the perfect solder joint**
- **Know how to desolder when things don't go right**

Heat up the iron and start soldering today!

US $3.99 CAN $4.99
ISBN: 978-1-449-33724-7

O'REILLY®
Twitter: @oreillymedia
facebook.com/oreilly
oreilly.com

Maker PRESS™
makerpress.com

Learn to Solder

Brian Jepson, Tyler Moskowite, and Gregory Hayes

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

Learn to Solder

by Brian Jepson, Tyler Moskowite, and Gregory Hayes

Copyright © 2012 O'Reilly Media. All rights reserved.
Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Shawn Wallace and Brian Jepson
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Gregory Hayes

May 2012: First Edition.

Revision History for the First Edition:
April 27, 2012 First release
See *http://oreilly.com/catalog/errata.csp?isbn=9781449337247* for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Learn to Solder* and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-33724-7
[LSI]
1335542415

Contents

Welcome v

1/Getting the Workspace Ready 1
- 1: Attach the Tip 2
- 2: Secure the Tip 2
- 3: Wet the Soldering Station Sponge 4
- 4: Set the Station's Temperature 4
- 5: Wipe the Iron's Tip 6
- 6: Tin the Iron's Tip 6
- Go Make Your Learn to Solder Skill Badge 7

2/How to Solder 9
- Working with Solder 9
- Keeping the Circuit Board from Moving 10
- Tinning Solder Pads 10
- Placing a Component in the Board 11
- Knowing Which Way a Component Goes In 13
- Stabilizing and Straightening Components 14
- Safely Soldering Sensitive Components 15
- Soldering a Component in Place 16
- Trimming Your Leads 18
- Going From Solderless Breadboard to PCB 20
- Soldering Jumper Wire 22
- Bridging Joints with Solder 22
- Relieving Strain on Cables 24
- Getting the Perfect Solder Joint 25

3/Desoldering 29
- The Desoldering Wick 29
- Preparing the Wick 30
- Desoldering with the Wick 31
- Desoldering with the Solder Sucker 32

4/What's Next? 33
- Teach the World to Solder 33
- Make More Things 34
  - Supercap Racer Kit 34

Wee Blinky Kit **35**
MintyBoost **36**

Welcome

Welcome to MAKE's Getting Started with Soldering Kit, and congratulations on taking a big step into the world of DIY electronics. Once you get the hang of soldering, you can put together some of the many great kits that are available, fix electronics that are broken, and build inventions of your own. With this kit, you'll learn how to:

- Prepare and clean your soldering iron
- Assemble electronic circuits from kits
- Transfer circuit designs from a solderless breadboard to a prototyping PCB
- Correct soldering mistakes you've made

Basic Tools

There's a lot of great stuff in the box, and before you start using it, here's a tour of what you'll find in there. Figure 0-1 shows the soldering tools you'll be using most of the time.

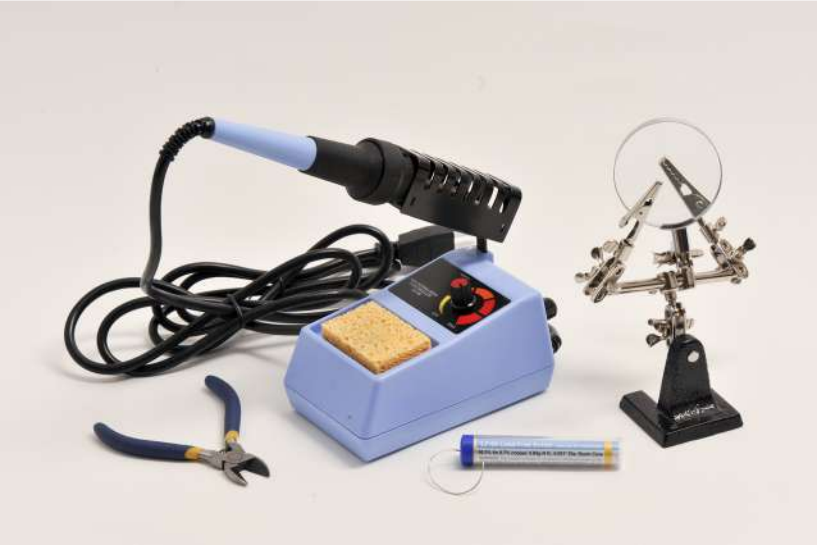

Figure 0-1. *Basic soldering tools*

Deluxe soldering iron
: (Top left) This soldering station includes a variable temperature controller, a cleaning sponge, soldering iron, and a ringed holder.

Crosscut pliers
: (Left) You'll use these to trim away excess leads after you solder components in place.

Solder tube
: (Bottom) This is enough solder to get you started and keep you busy over many projects.

Helping hands
: (Right) For those times when you need a third or fourth hand, the helping hands let you hold items steady while you solder.

Advanced Tools

Most of the time, you'll only need the basic tools to get things done. But when you need to replace your soldering iron tip, correct a mistake you made while soldering, or need a tool to help keep components from overheating, you'll need the items shown in Figure 0-2.

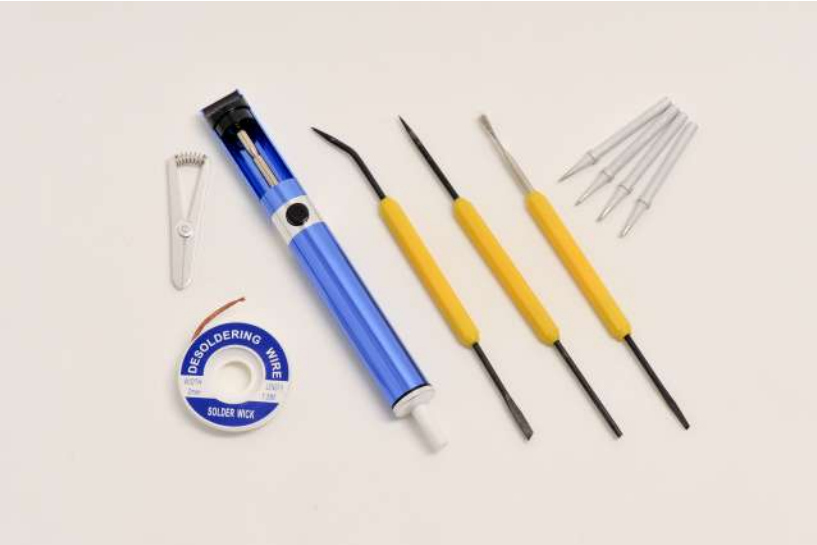

Figure 0-2. *Advanced soldering tools*

Heat sink
: (Left) Clip this to sensitive components to help dissipate heat.

Desoldering wick
: (Bottom left) Use this to wick away excess molten solder.

Desolder pump
: (Center left) This pump will suck up molten solder when you have a lot of solder to remove.

Soldering tools
: (Center right) The scrapers, brush, and slotted probe come in handy when you need to move solder around or precisely position a component.

Replacement tips
: (Right) Tips don’t last forever. When you’ve worn out your tip, use one of these as a replacement.

Project 1: Learn to Solder Skill Badge

Our Learn to Solder Skill Badge Kit (the 2011 model is shown in Figure 0-3) has been used to teach thousands of people of all ages how to solder at Maker Faires across the country. It's a simple, fun way to learn how to solder and also how to teach others to solder. After you build the one included in the box, you can order more from *makershed.com* and teach others how to solder.

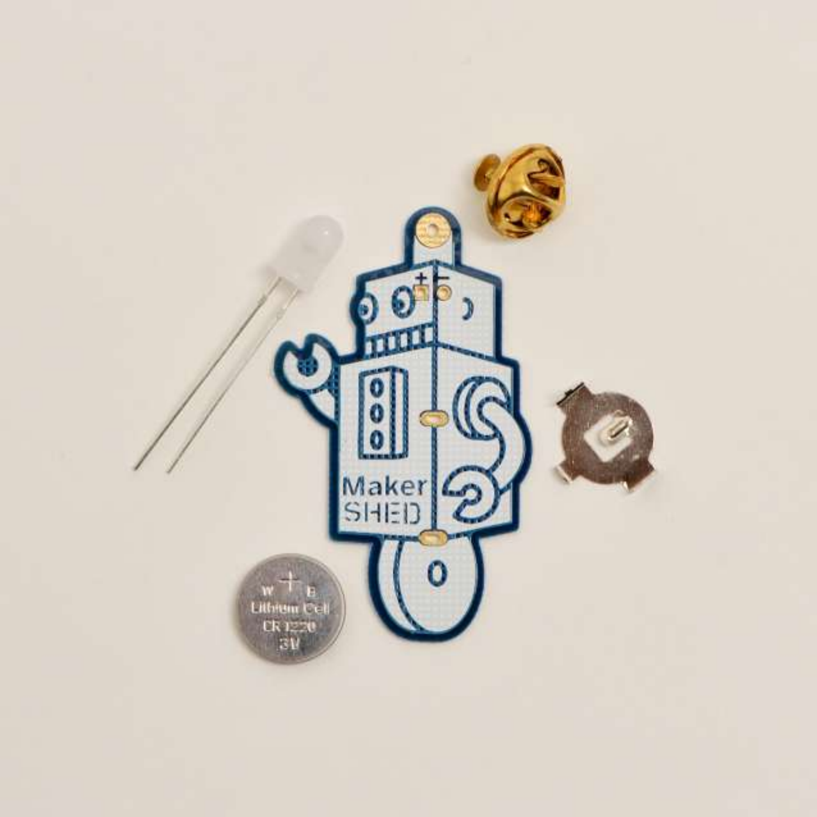

Figure 0-3. *Learn to Solder Skill Badge, 2011 model*

Blinking or Color Change LED
: (Top left) This is an LED (Light-Emitting Diode) with a twist. Normal LEDs give off a single color, and keep shining as long as you give them power. Also, normal LEDs require a specific voltage to operate: don't give it enough power, it won't light at all; give it even a little too much, and you could burn it out.

This LED is different in both respects: it has three elements (red, green, and blue) that are under the control of a small integrated circuit (IC) embedded within the LED. The IC causes the elements to change color in a repeating pattern. Because the IC controls the voltage that it delivers to the individual color elements, it's more tolerant of variations in the voltage you give it. For example, even though red LED color elements typically operate at around 2 volts, you're able to use a 3 volt battery with the pin.

The 2012 model uses two self-blinking LEDs instead of one self-color changing RGB LED.

CR1220 battery

(Lower left) This "coin cell" battery supplies power to the pin.

Printed Circuit Board (PCB)

(Center) The 2011 Learn to Solder Skill Badge features a friendly robot. The 2012 model will also have a robot, every bit as friendly.

Pin and clutch

(Top right) This is what holds the pin to your clothing.

Battery holder

This keeps the battery on the PCB.

Project 2: 555 Timer Blinky

The 555 timer blinky is a simple circuit that makes two LEDs flash in an alternating pattern. When one is on, the other is off. Figure 0-4 shows the components for this project, and Figure 0-5 shows the finished project. This is a slightly more challenging project than the skill badge.

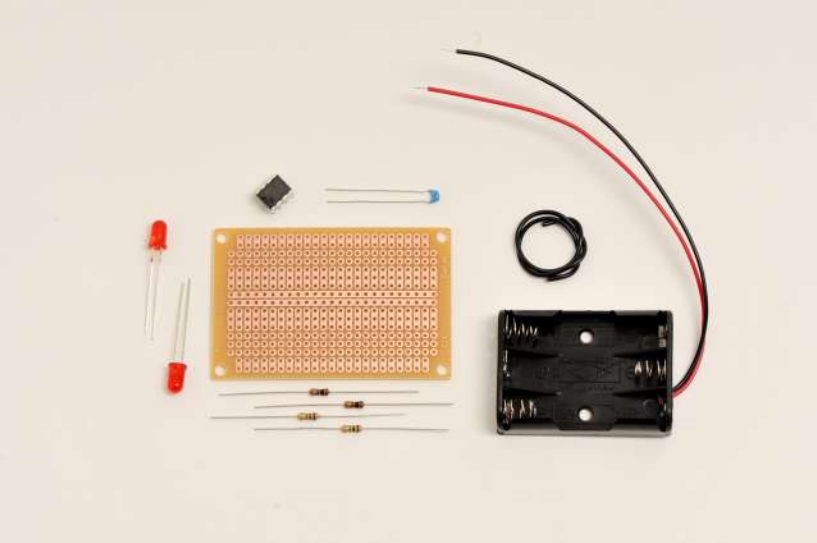

Figure 0-4. *555 blinky parts*

LEDs
(Left) Unlike the LEDs used in the skill badge, these are your normal, run of the mill LEDs.

Protoboard
(Center) Look back at the skill badge PCB. Notice how it was custom designed for a single purpose. Now look at this protoboard PCB: it's got a lot of different holes in convenient locations. You can build many kinds of simple electronics projects on a board like this.

555 Timer
(Top left) The 555 timer is a special type of integrated circuit that can turn electrical current on and off in a repeating pattern. It's just what you need to blink an LED.

Capacitor
(Top right) To customize the 555 timer, you need to connect certain components to it. The value (capacitance or resistance) of the component determines how quickly the 555 turns things on and off.

Hookup/Jumper Wire
: (Top right) You'll need this to connect one thing to another.

Resistors
: (Bottom) Resistors serve two purposes here: two are used to make sure the LEDs don't get too much current; the other two work with the capacitor to control the behavior of the 555 timer.

Battery Box
: (Right) This holds 3 AAA batteries.

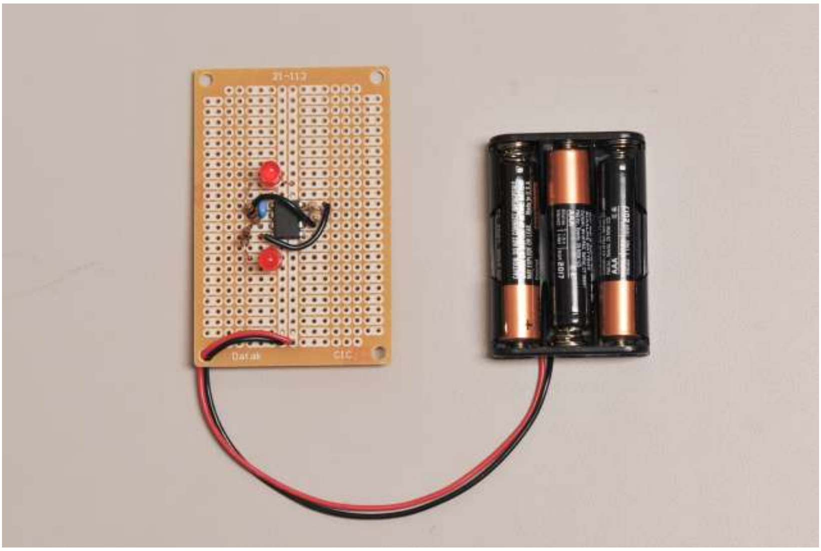

Figure 0-5. *Finished 555 blinky*

Contact Us

Maker Shed is here Monday to Friday, 7:30 AM to 5:00 PM Pacific Time

Email: *help@makershed.com*
Phone: 800-889-8969 or 707-827-7019

Answers to our most commonly asked questions: *http://www.makershed.com/help.asp*

Got a kit or product you think we should sell? Send a note to *kits@makezine.com*.

Want to carry MAKE books or kits in your store? Send a note to *resellers@makershed.com*.

1/Getting the Workspace Ready

Before you can start soldering, you need to get your iron ready. This involves a few steps. First, you'll have to attach the tip to the soldering iron. Then, you'll need to prepare the station: make sure your sponge is moistened, set the temperature, and "tin" the tip by applying some solder to it.

At the end of this chapter, your soldering iron will be assembled, clean, and hot enough to begin soldering.

Solder contains harmful chemicals. Wash your hands after soldering as well as after handling solder or the solder station sponge. Though it may be tempting to snack while you're working, keep any food or drink away from your soldering area.

1: Attach the Tip

Locate the soldering iron (make sure it's unplugged and cool), a soldering iron tip, and the tip nut that holds the tip in place. Insert the tip into the soldering iron as shown in Figure 1-1, and push it in as far as it will go. It will not require much force to insert it.

WARNING: Make sure the soldering iron is unplugged and cool.

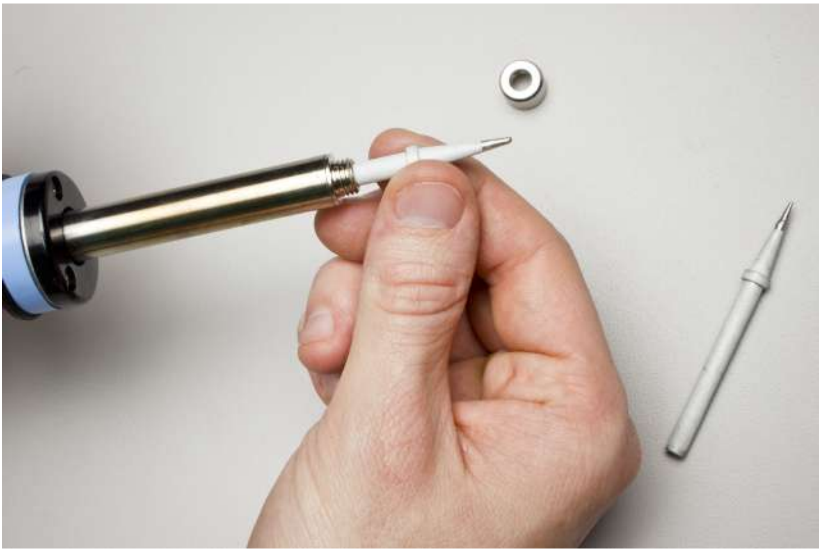

Figure 1-1. *Inserting the tip into the tube*

2: Secure the Tip

Slide the tip nut over the tip, and screw it in place. You can use pliers to tighten it (see Figure 1-2). Be careful not to overtighten it, since you will eventually need to replace it when the tip gets worn from all the projects you'll be making with it.

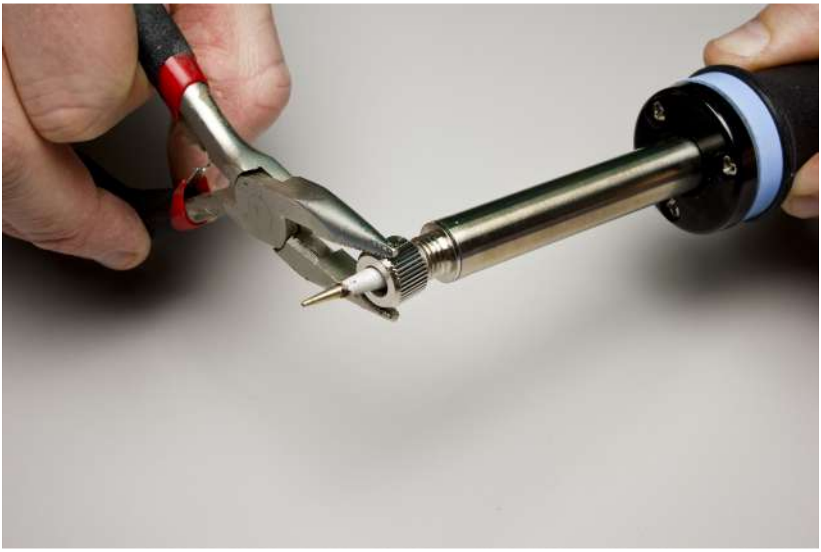

Figure 1-2. *Making sure the tip stays put*

WARNING: Over time, the tip nut might become loose. Let the iron cool down before you tighten it.

3: Wet the Soldering Station Sponge

You'll need to perform this step (and the remaining ones in this chapter) at the beginning of each soldering session.

As you use the soldering iron, you're going to be wiping the tip a lot, so you need to keep the solder station sponge moist. Pour a little clean, fresh water on the sponge as shown in Figure 1-3. Don't soak it, but make sure it's completely moistened.

NOTE: If your water has a high mineral content, use distilled water for wetting the sponge.

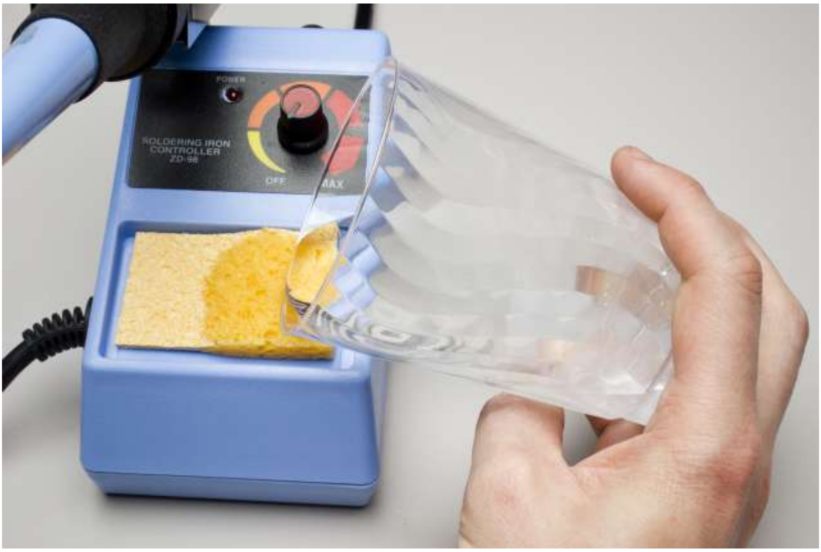

Figure 1-3. *Wetting the sponge*

4: Set the Station's Temperature

Plug your soldering iron in, and set the heat setting to the mark shown in Figure 1-4 (pointing at the 3 o'clock position on the dial). Give the iron a couple minutes to warm up.

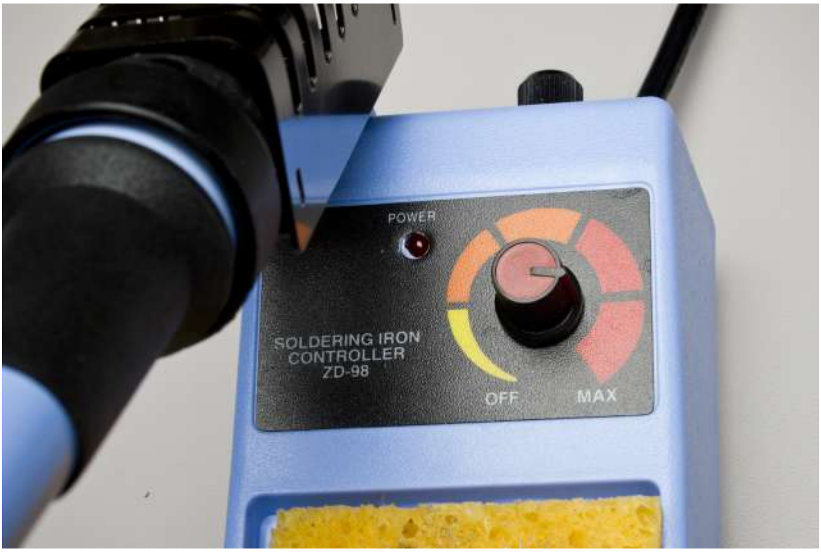
POWER
SOLDERING IRON
CONTROLLER
ZD-98
OFF
MAX

Figure 1-4. *Setting the temperature*

WARNING: The iron will get very hot, and certainly hot enough to burn you. Don't touch the tip and don't touch the tip to anything other than solder or components you are soldering together.

5: Wipe the Iron's Tip

Wipe the tip on the sponge as shown in Figure 1-5, being sure to turn the iron a few times as you wipe it, so you wipe all of its surfaces. It will make a sizzling sound as the water in the sponge comes into contact with the tip. The figure shows a sponge that's been used a few times. Notice the little blobs of solder that have come off the tip over time. A clean tip is critical to successful soldering. Even a small layer of oxidized material or other crud will limit the amount of heat that's transferred to the component you're soldering. Wipe the tip often, and keep the tip clean.

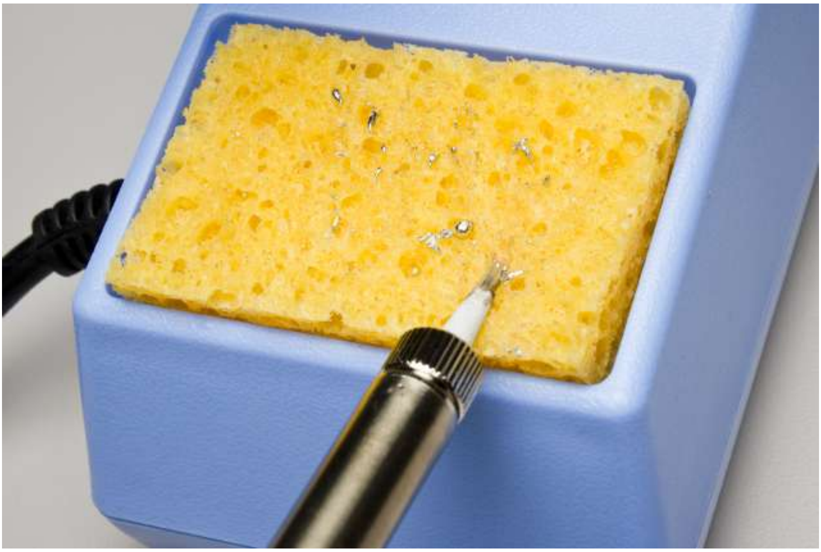

Figure 1-5. *Wiping the tip clean*

6: Tin the Iron's Tip

Before you begin soldering components, you should put a thin layer of solder on the tip as shown in Figure 1-6. This will have two effects: burning off any crud that shouldn't be there, but also providing a thin layer of liquid solder. This layer makes for a very effective heat transfer between the soldering iron and whatever you're soldering. Give the tip one more quick wipe on the iron and place the soldering iron into the holder until you need it.

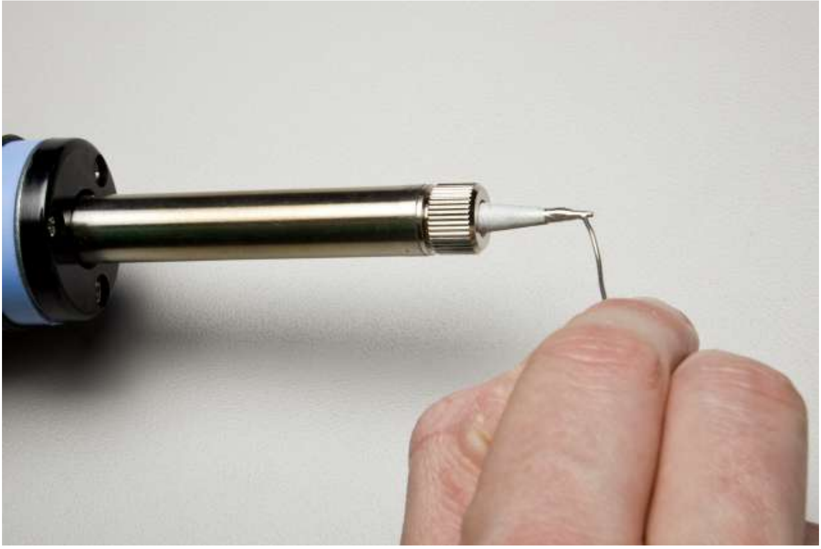

Figure 1-6. *Tinning the soldering iron tip*

NOTE: A small puff of smoke will appear as the *rosin flux* in the center of the solder is activated. The rosin serves two purposes: to clean the soldering iron (and the joint that you are soldering) and to help the solder flow freely. Avoid inhaling this smoke.

Go Make Your Learn to Solder Skill Badge

With your workstation all set up, you're ready to try your hand at soldering. The Learn to Solder Skill Badge ("Project 1: Learn to Solder Skill Badge" on page viii) is a great place to start. It's so simple, we think you can even try it before you read any more. But before you try the second project ("Project 2: 555 Timer Blinky" on page x), you should definitely give the rest of this guide a read.

2/How to Solder

With your soldering iron prepped, you're ready to begin soldering. This chapter has many helpful tips and techniques for soldering effectively, safely, and neatly.

Working with Solder

Solder is made of metal with a core that contains flux, which cleans the connection as you solder. When the metal melts, the flux begins to flow onto the joint. Some solder lead-based, but the solder in this kit is lead-free. To melt solder, heat it with the soldering iron as shown in Figure 2-1.

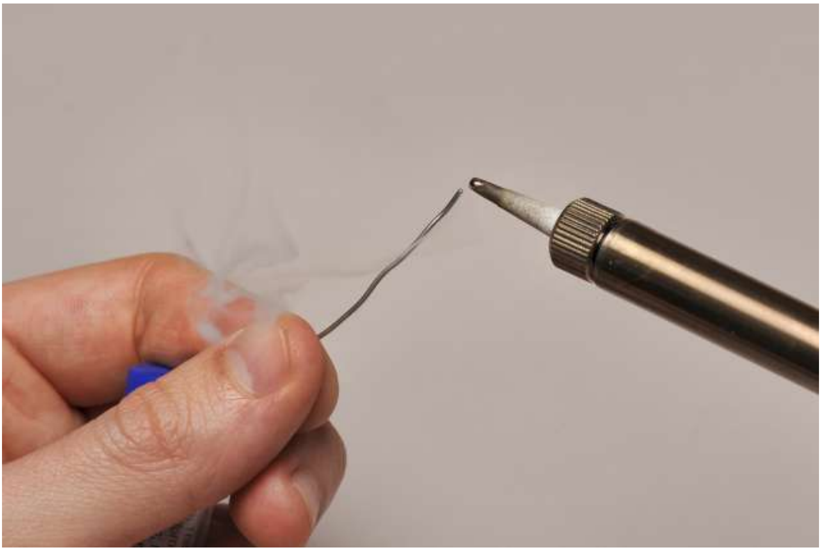

Figure 2-1. *Melting some solder*

Try melting some solder, but be careful not to drip it on yourself or anything other than a work surface. You can try pushing tiny balls of solder around on your work surface to see how it flows. If you have some bare hookup wire or similar metal, try heating it up by holding the soldering iron to it (don't hold the wire while you're heating it). Touch a piece of solder to the wire, but choose a point that's an inch or so from where the soldering iron is touching. How close do you have to get to the soldering iron tip before it melts?

Keeping the Circuit Board from Moving

When you're soldering, you've got a lot to juggle: the soldering iron, the solder, and the two things you are connecting to each other. The helping hands let you place the items you're soldering in a stable position. Use the clips to hold the item in place as shown in Figure 2-2. But once you start inserting items into the board, you need some way to keep them from moving. "Placing a Component in the Board" on page 11 shows you how to do that.

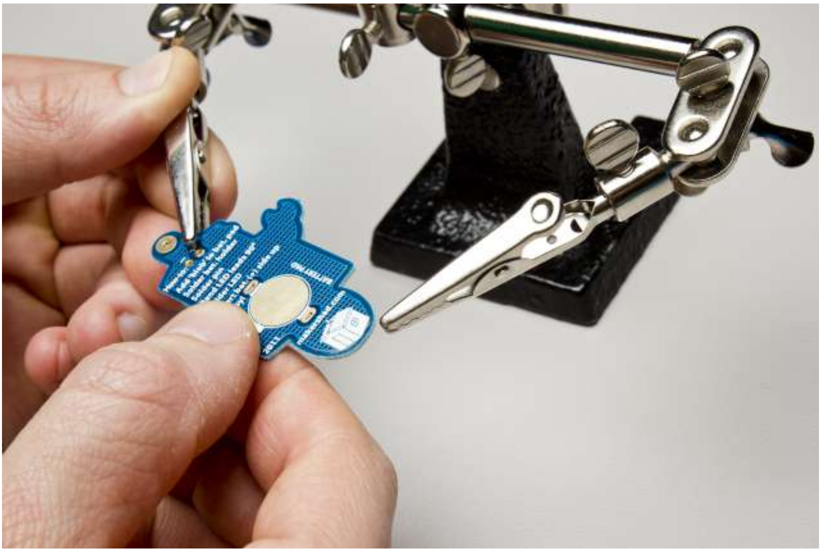

Figure 2-2. *Holding a PCB in the helping hands*

Tinning Solder Pads

You'll sometimes come across things that need a layer of solder to work right. For example, on the Learn to Solder Skill Badge, you need to put a bump of solder in place to give the battery a snug fit. With solder pads that are this big, you need to heat the pad with the iron really well; the pad is so large that it's going to take longer to heat. It's best if you melt the solder by pushing it onto the pad rather than pushing it directly against the iron. This makes sure that the solder flows thoroughly over the pad. If the solder doesn't melt, try tinning the tip again first.

Start with a little blob as shown in Figure 2-3, and spread it around evenly. When you've got the desired thickness, first pull the solder away, then pull the soldering iron away. Doing it in this order will avoid leaving chunks of solder behind.

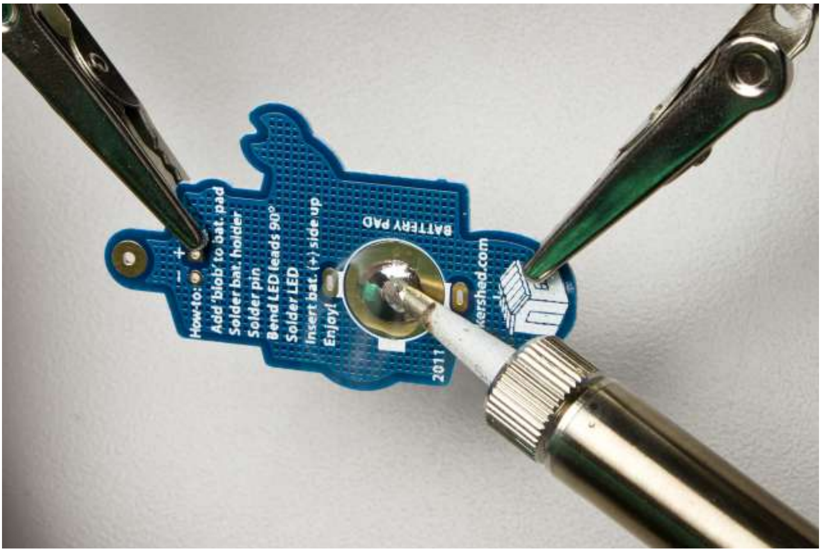
How-to:
Add blob to bat. pad
Solder bat. holder
Solder pin
Bend LED leads 90°
Solder LED
Insert bat. (+) side up
Enjoy!
BATTERY PAD
kershed.com
2011

Figure 2-3. *Laying down some solder*

Placing a Component in the Board

You don't want components falling out of the board while you're soldering. In fact, you'll often insert a component with the board upside down (see Figure 2-4), then flip the board over when you put it in the helping hands. So the components have plenty of opportunity to fall out. To keep the component in place, bend the leads out as shown in Figure 2-5.

In some cases, you may need to quickly tack a component in place to keep it from moving. See "Stabilizing and Straightening Components" on page 14.

Don't try to place every component at once. Start with low-profile (shorter) components, solder them in one at a time, and move on to higher-profile components.

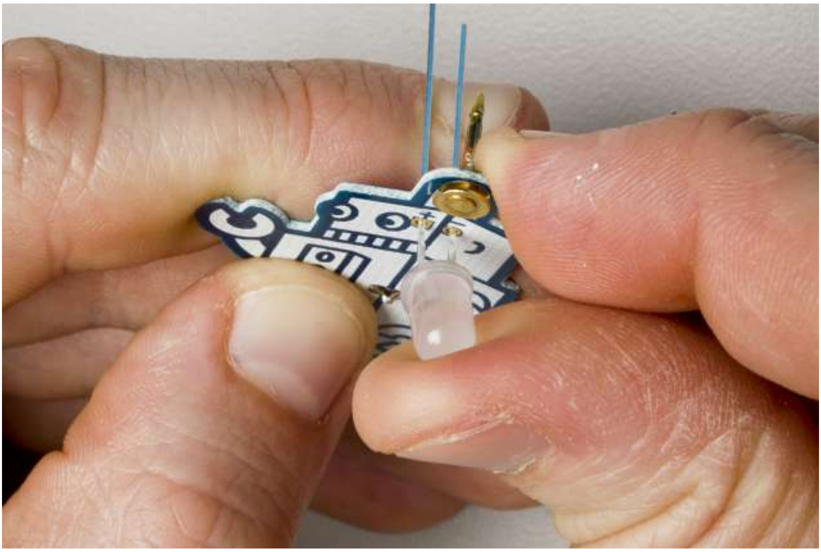

Figure 2-4. *Inserting an LED*

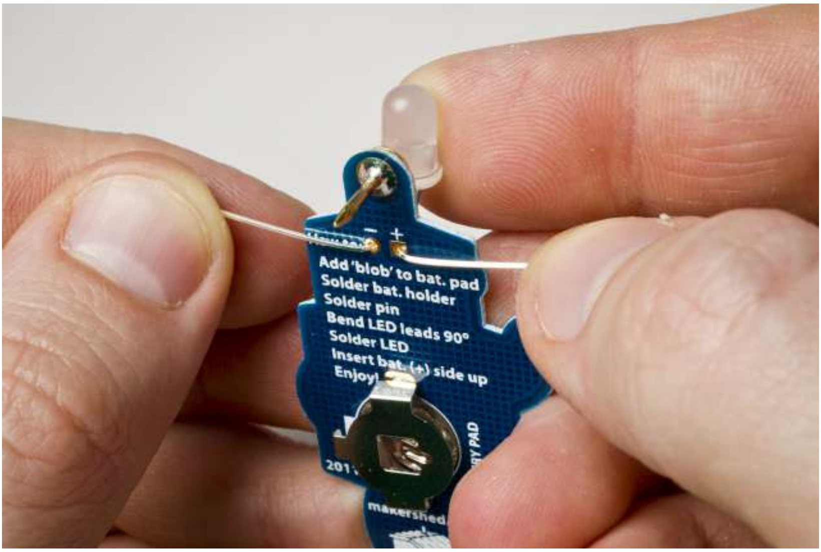
– +
Add 'blob' to bat. pad
Solder bat. holder
Solder pin
Bend LED leads 90°
Solder LED
Insert bat. (+) side up
Enjoy!
makershed

Figure 2-5. *Bending out the leads*

Knowing Which Way a Component Goes In

Some components don't care which way you put them in: resistors, some capacitors (such as the one that comes with this kit), and many other components fall into that category. On the other hand, LEDs are *polarized*; electrical current will only flow through them in one direction. So if you put your LED in backwards, it won't light up at all. You can determine an LED's polarity in a couple of ways: First, the longer of the two leads is the positive (+), and the shorter is negative (-).

Second, look closely at the bottom of the LED and examine the ring that bulges out around it. There is a flattened part of that ring that indicates the side of the LED that's negative (-). You can see this flattened side on the right in Figure 2-6.

With ICs, this also matters. Every pin has a specific function, and there will be one or more pins for positive (+) and one or more pins for negative (-).

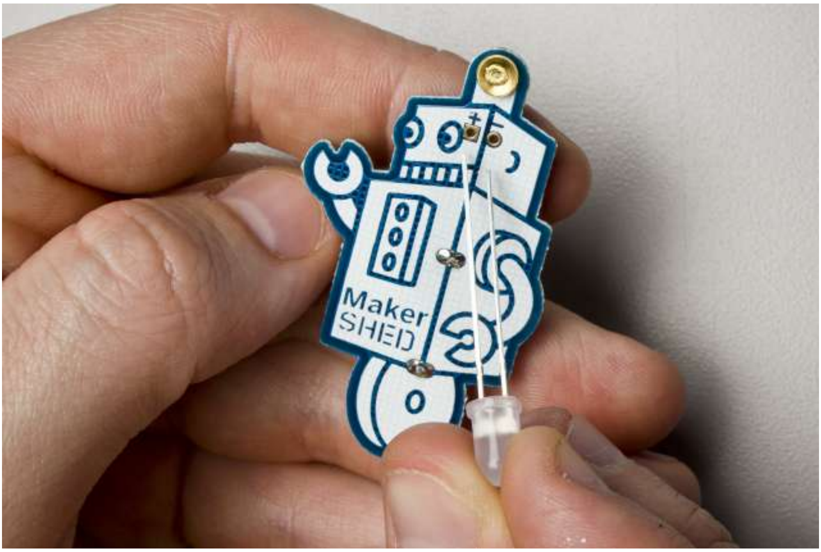

Figure 2-6. *Polarity—LEDs have it*

Stabilizing and Straightening Components

ICs present a unique problem. They don't have long enough leads for you to bend out effectively, which means they are more likely to slip and slide while you're soldering them. So do the best you can with that (see Figure 2-7), but flip the board back over after you solder the first joint. In most cases, you'll find that the IC has shifted. But since you only have one joint soldered, you can easily move the IC into the position you want. Once you've done this, flip the board over and solder the remaining connections.

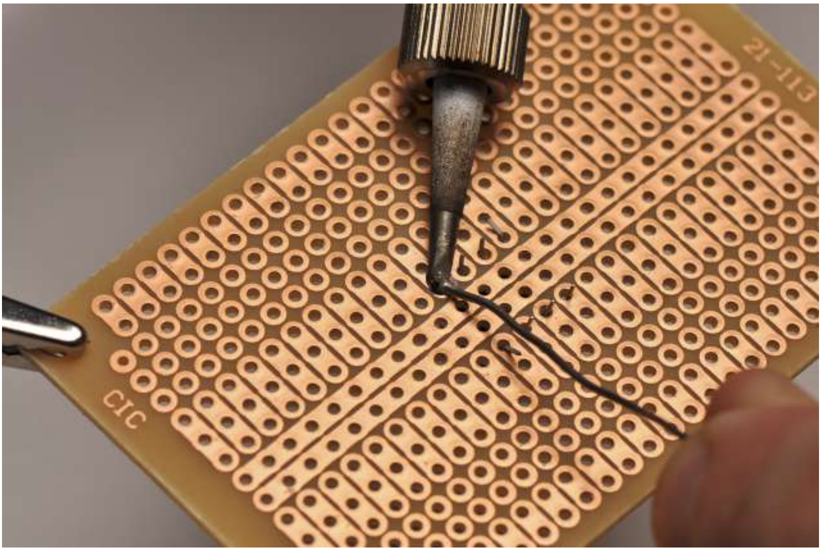

Figure 2-7. *Tacking down an IC*

Safely Soldering Sensitive Components

You need to avoid applying too much heat to a component, or you might damage it. One way to help avoid this is to clamp the heatsink to the component as shown in Figure 2-8. This will draw heat away from the component, protecting it somewhat.

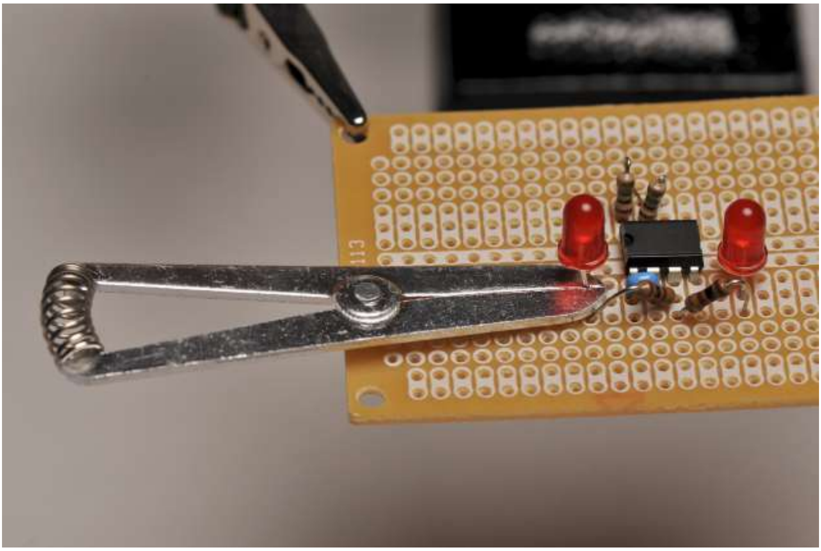

Figure 2-8. *Keeping heat away from the component*

Soldering a Component in Place

Wipe the soldering iron on the sponge and hold it so it's touching both the solder pad and the lead that protrudes through. Avoid using the soldering iron to melt the solder. Instead, heat up both the lead and the pad so that the solder melts when you touch it to them.

This must all be done quickly: within a second of touching the iron to the joint, push a small amount of solder into the joint, and let it flow around the joint as shown. Give it a second to flow, take the solder away, then wait a second, and take the iron away. With practice, you should be able to do all of this in three to four seconds per joint. Figure 2-9 shows how you'd solder the pin to the skill badge. Figure 2-10 shows an LED being soldered.

If you hold the iron to the component too long, you run the risk of damaging either the component or the solder trace on the PCB.

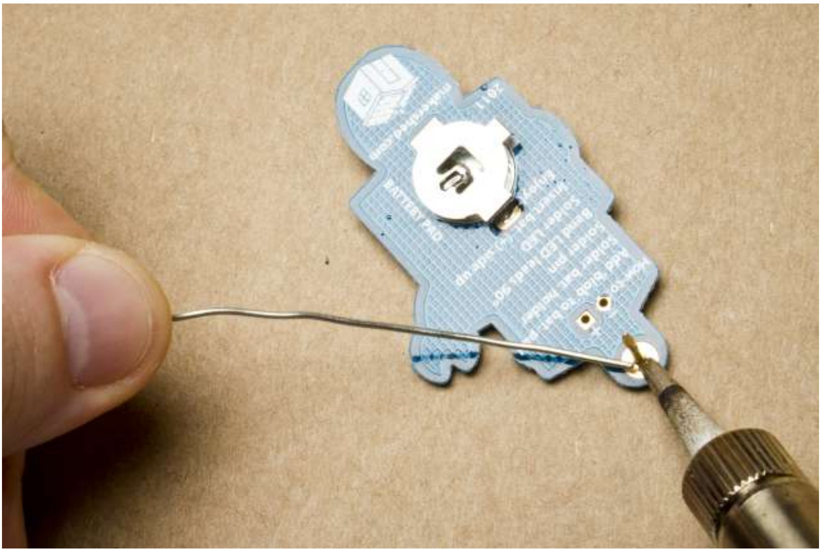

Figure 2-9. *Soldering a pin*

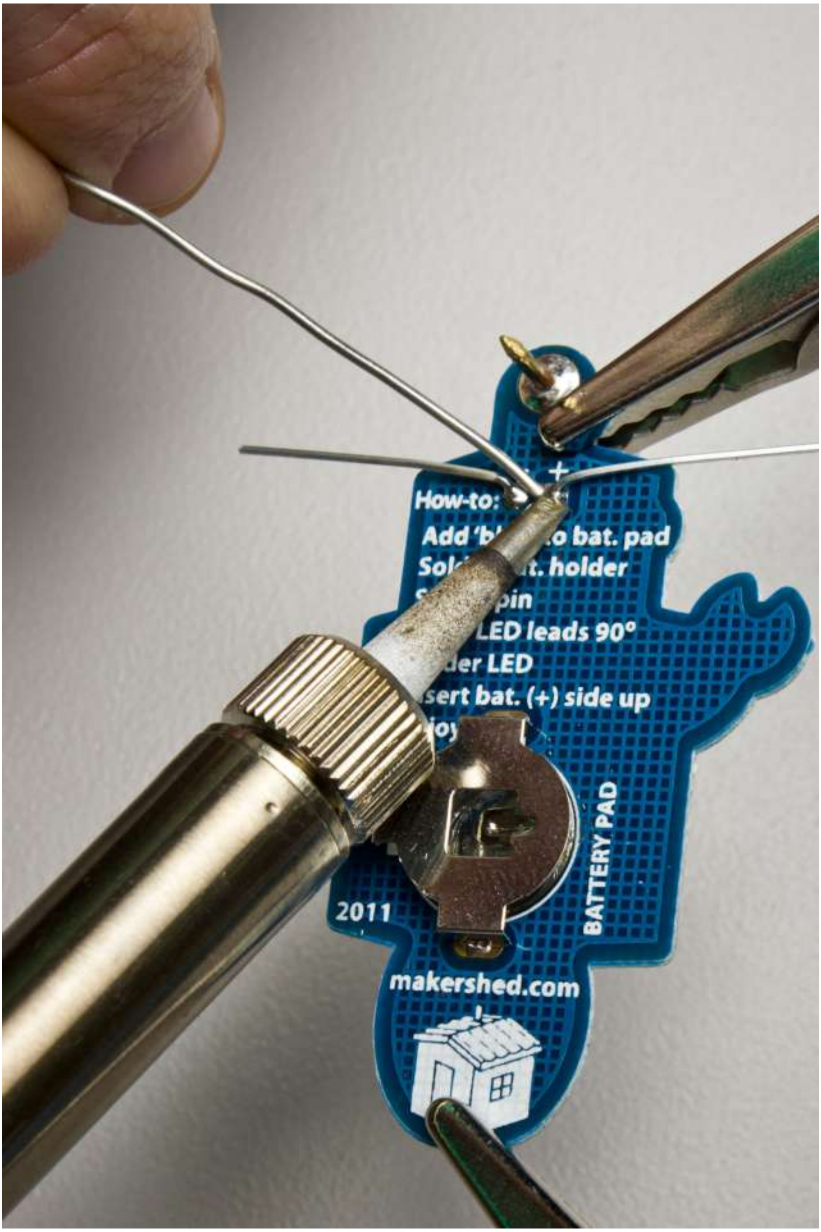
+
How-to:
Add 'b o bat. pad
Sol t. holder
pin
LED leads 90°
der LED
sert bat. (+) side up
2011
BATTERY PAD
makershed.com

Figure 2-10. *Soldering an LED*

Trimming Your Leads

After you solder a component in place, you need to tidy things up. Grab the cutters from your kit, and trim the leads as close to the solder joint as possible (see Figure 2-11 and Figure 2-12). Don't cut into the solder joint; the idea is to trim the excess wire lead.

When you clip it, the wire lead will fly away from the clippers very fast, and could injure someone. The best way to keep this from happening is to use one or two fingers to hold the lead you're clipping. With one finger, you can put enough pressure on it to keep it from flying away.

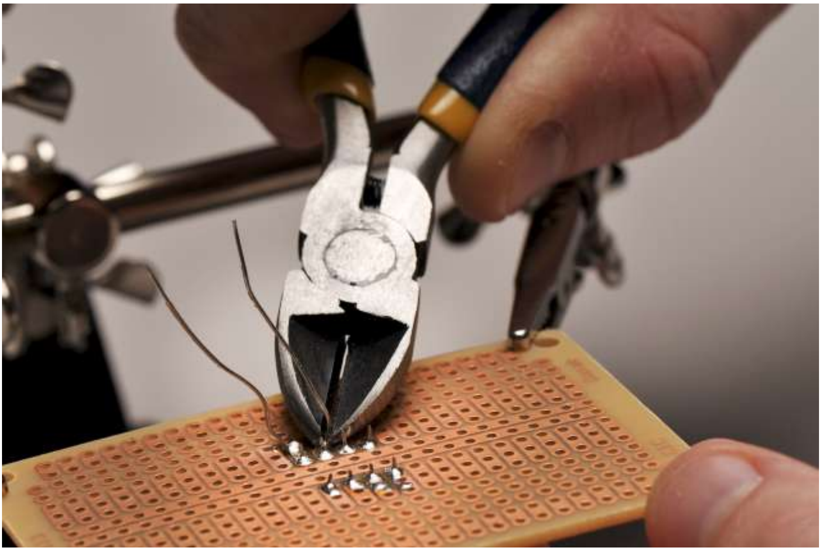

Figure 2-11. *Trimming leads on the 555 blinky*

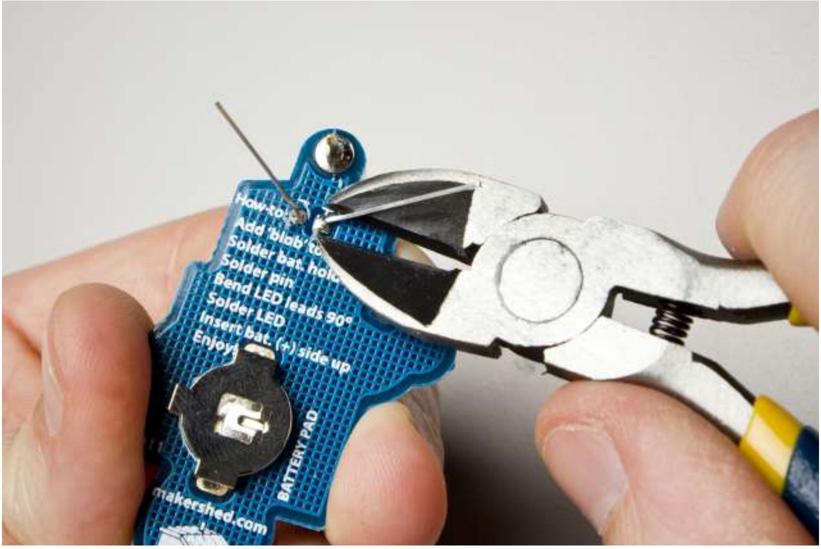
How-to:
Add 'blob' to
Solder bat. hold
Solder pin
Bend LED leads 90°
Solder LED
Insert bat. (+) side up
Enjoy!
BATTERY PAD
makershed.com

Figure 2-12. *Trimming leads on the Learn to Solder Skill Badge*

Going From Solderless Breadboard to PCB

The beauty of protoboard is that it mirrors the layout of a solderless breadboard, which is used for prototyping electronic circuits. Once you have your project laid out and working on a breadboard, you can easily transfer it to the protoboard. As with a solderless breadboard, the protoboard has rows and columns that are tied together. The layout is slightly different, though. Where the breadboard has two rails (one for positive, one for negative) on both the top and bottom as shown in Figure 2-13, the protoboard that comes with this kit has its rails in the center of the board.

Figure 2-14 shows the project from Figure 2-13, but laid out on a protoboard instead of the solderless breadboard.

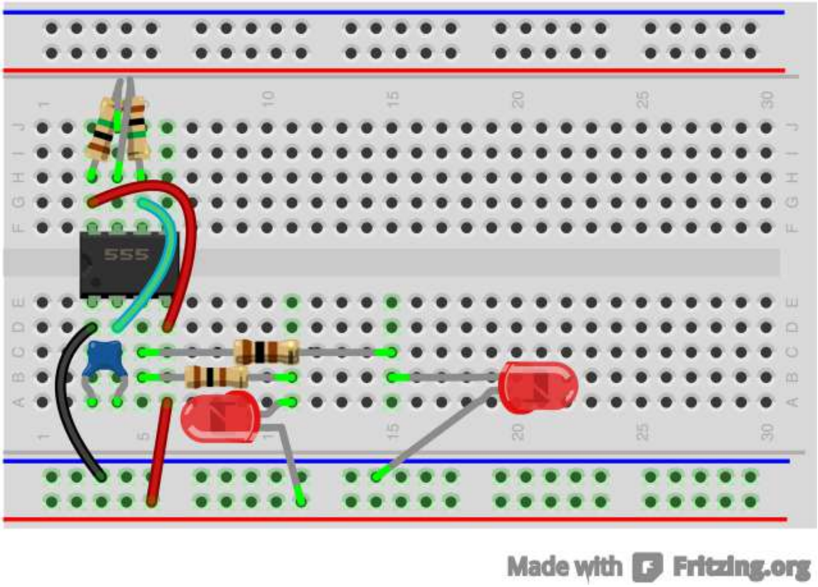

Figure 2-13. *The 555 blinky project on a breadboard*

NOTE: The breadboard diagram in Figure 2-13 was made with Fritzing, an open-source initiative to support designers, artists, researchers and hobbyists to work creatively with interactive electronics. For more information, see *http://fritzing.org/*.

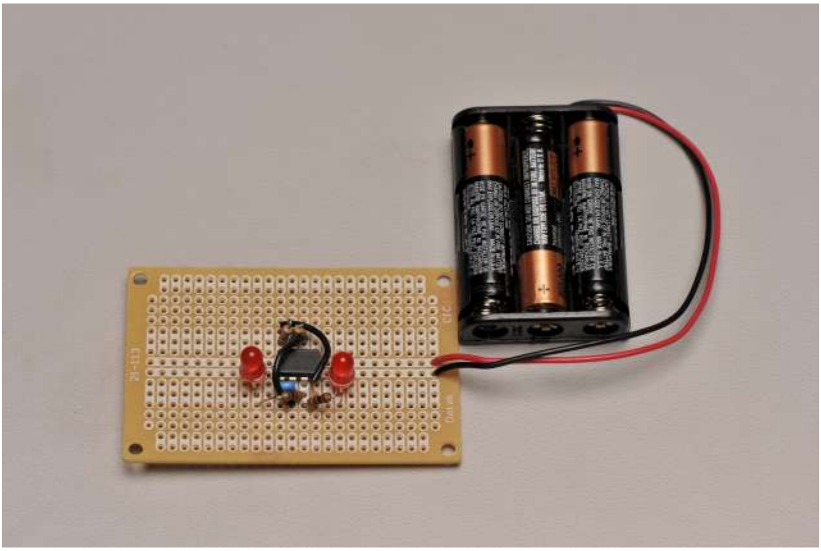

Figure 2-14. *The 555 blinky project on a protoboard*

Soldering Jumper Wire

Jumper wire can be treated a lot like other components (see "Placing a Component in the Board" on page 11): make sure there's enough bare wire poking through so you can bend it outward enough to hold it in place. The notched probe included with this kit can be helpful with placing jumper wire in the right place (see Figure 2-15).

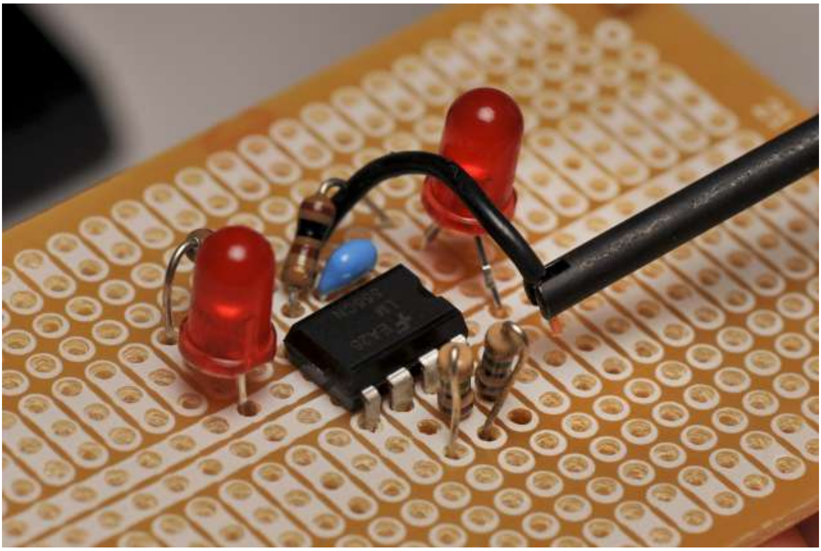

Figure 2-15. *Placing the jumper wire*

Bridging Joints with Solder

When you're working with the protoboard, you will sometimes need to bridge joints that are close to each other. This is preferable to using jumper wire since your board remains uncluttered on the top. It doesn't always look great on the bottom, though, because you end up with a lot of solder in some places. Figure 2-16 shows the start of a bridge: feeding the solder into the gap between solder joints. Figure 2-17 shows a bridge being formed.

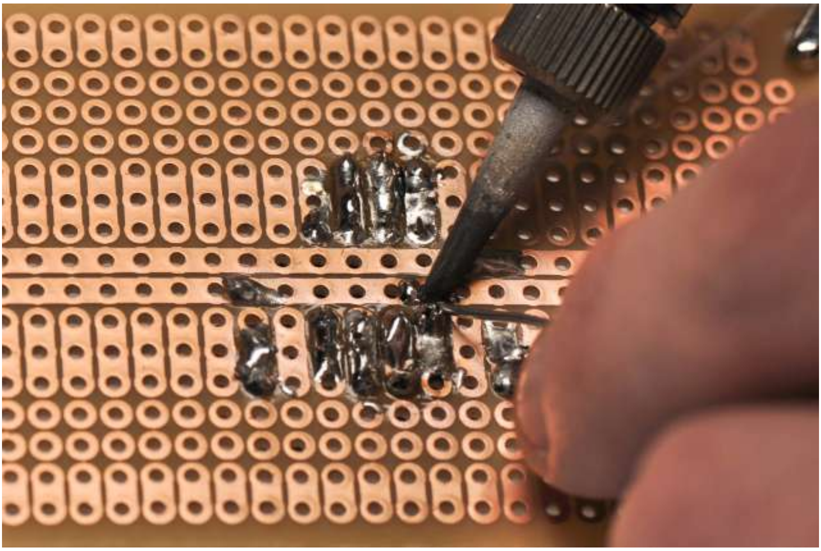

Figure 2-16. *Feeding solder into a gap*

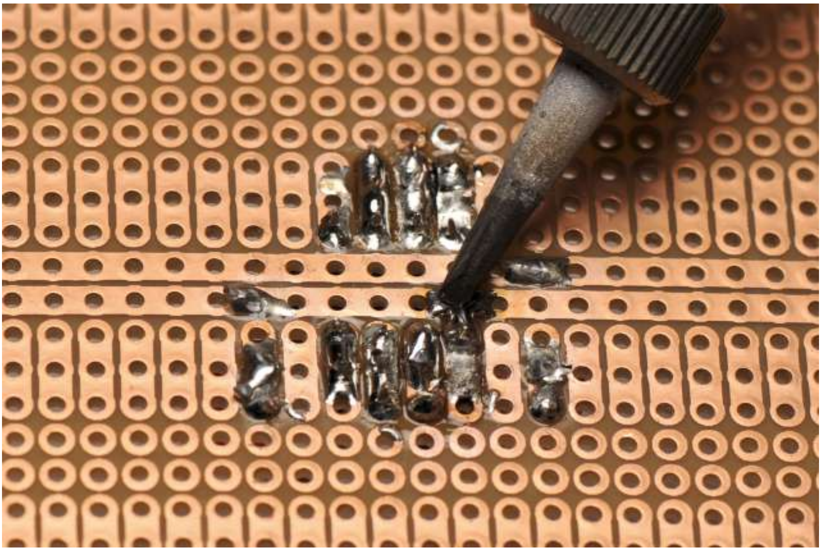

Figure 2-17. *Forming a bridge across the gap*

Relieving Strain on Cables

If you have something heavy dangling off your board, such as a battery box, it's very likely that the solder joints will come undone. If you pass the wires through the large holes on the board before soldering them down (Figure 2-18), you can add some strain relief (Figure 2-19) that will help prevent this from happening. If you have a small drill, you can also drill out a couple of holes close to where you are going to solder the connection, and loop the wire through there. This will be even more rugged than using the larger holes.

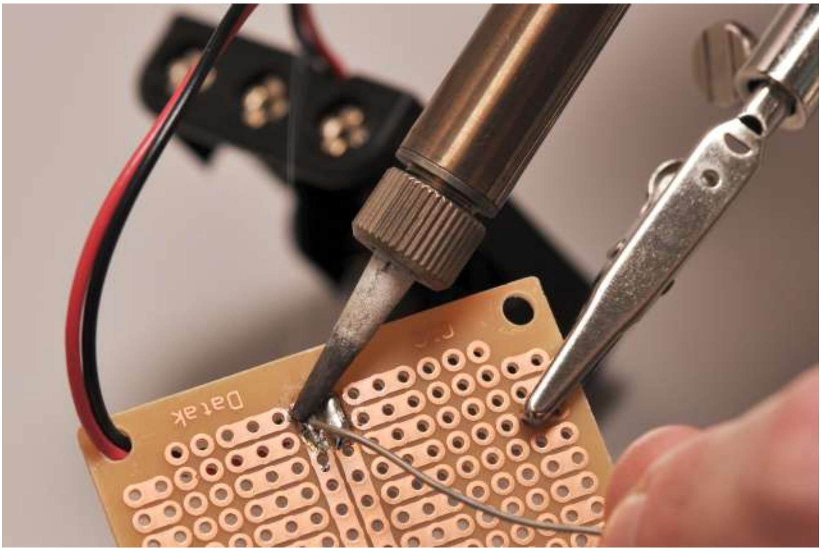

Figure 2-18. *Soldering the battery holder wires*

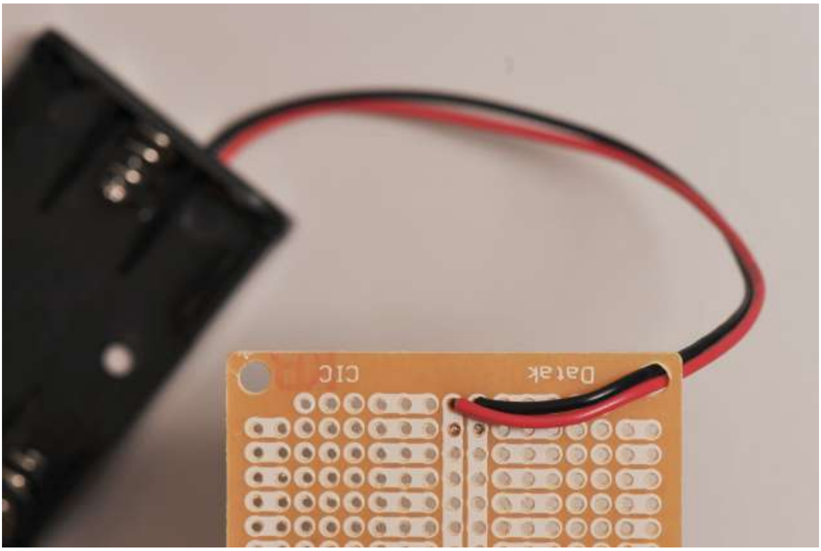

Figure 2-19. *Strain relief*

Getting the Perfect Solder Joint

It’s not hard to get a great solder joint, but it’s also easy to be sloppy. Here are some pictures of good and bad joints to help you out.

Figure 2-20
Here’s a good solder joint on the positive (+) pad: solder surrounds the joint and makes a little peak. This joint had just the right amount of solder and just the right amount of heat.

Figure 2-21
The joint on the negative (-) pad is sloppy. The solder didn’t flow around the joint, and some of it is spread out across the lead. You can try to fix it by holding the iron to the joint for a couple seconds; the solder should flow down the lead and onto the joint.

Figure 2-22
The joint on the positive (+) pad has too much solder, and it’s balled up. You can use the desoldering wick or solder sucker to remove the excess. After you do that, hold the iron to the joint for a couple seconds to get the solder to flow around it.

Figure 2-23
Here are two joints for comparison: one good, one blobby.

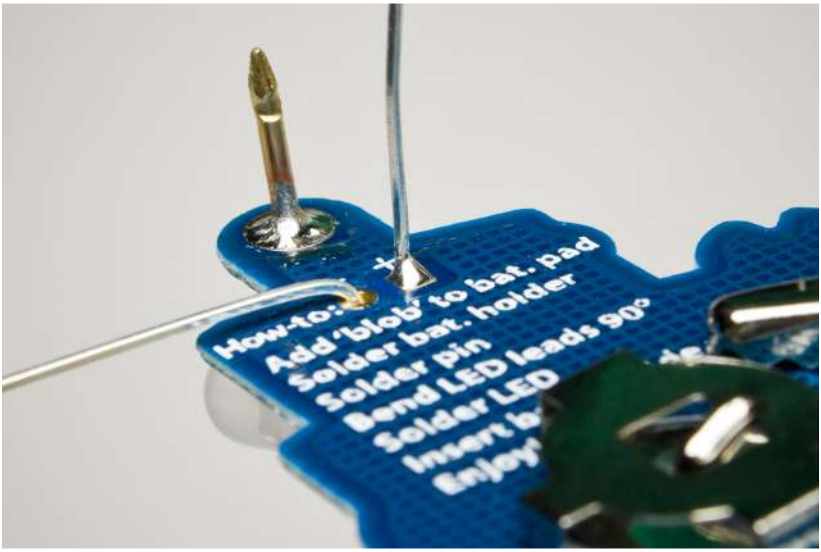
How-to:
Add 'blob' to bat. pad
Solder bat. holder
Solder pin
Bend LED leads 90°
Solder LED
Enjoy

Figure 2-20. *A good solder joint*

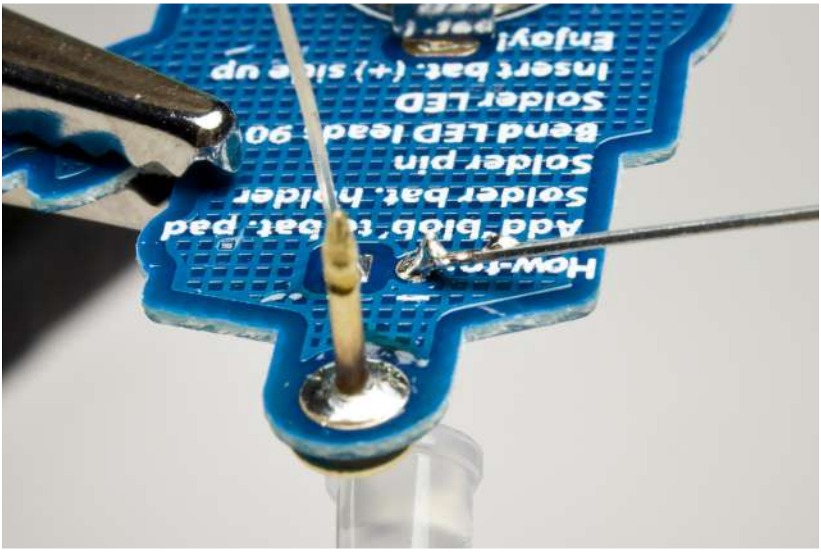

Figure 2-21. *A messy solder joint*

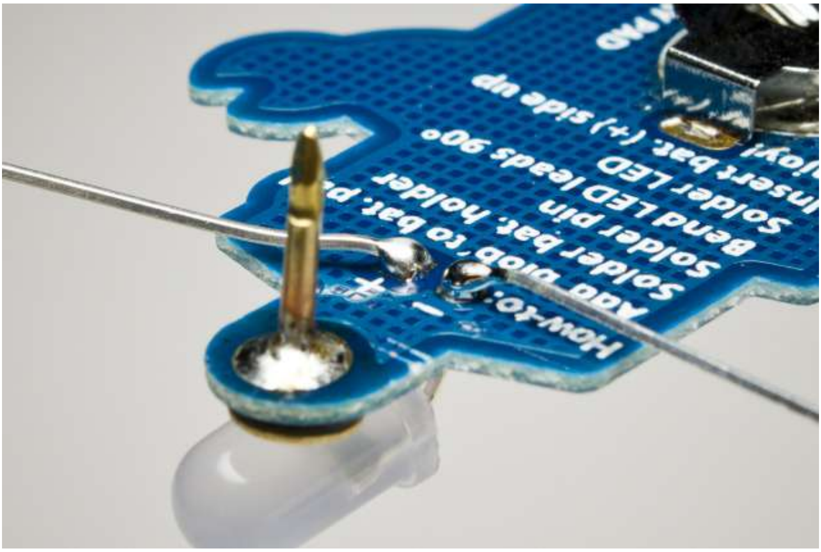
How-to:
Add blob to
Solder bat. holder
Solder pin
Bend LED leads 90°
Solder LED
Insert bat. (+) side up
Enjoy!

Figure 2-22. *A blobby solder joint*

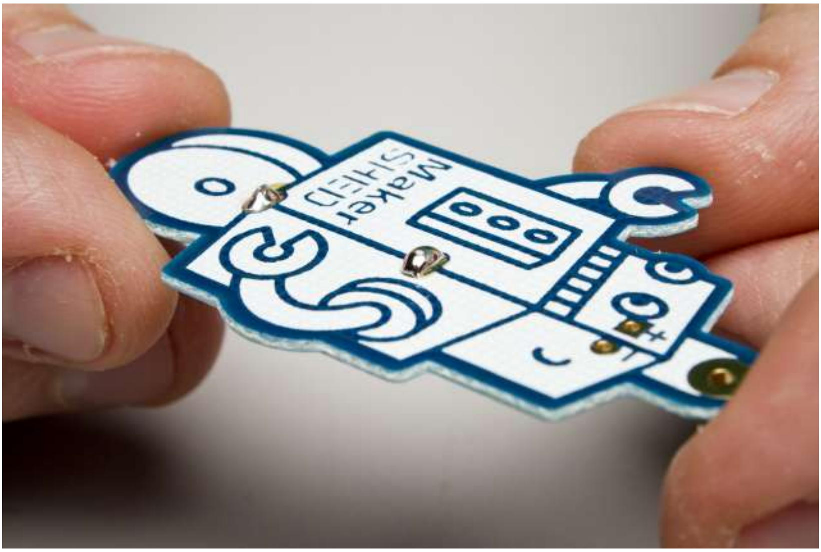

Figure 2-23. *Comparing two joints*

3/Desoldering

If everything went right the first time, you'd never learn anything. This section shows you how to recover from little soldering mistakes, and also includes a gallery of various solder joints so you can compare your work.

Desoldering is the "undo" command for soldering. With this technique, you can remove solder from a joint, allowing you to free the component so you can reorient it as needed. There are two tools for this: the *desoldering wick* (sometimes called *desoldering braid*) and the *solder sucker*.

The Desoldering Wick

Pull a small length of wick out of its spool (Figure 3-1). You don't need much, but if you don't use it all, you can wind the wick back into its spool.

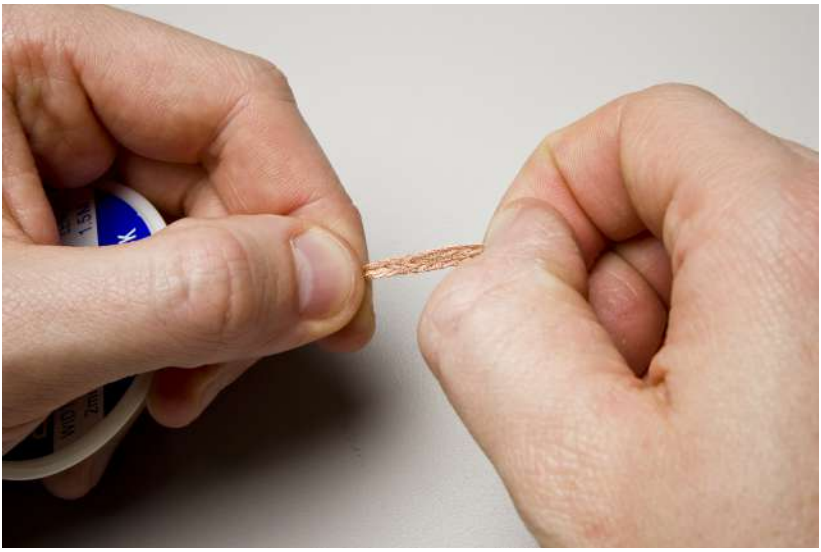

Figure 3-1. *Pulling out some wick*

Preparing the Wick

Before you use the wick, hold it as shown in Figure 3-2, and push inward. This will cause the wick to spread out, giving you more surface area to work with.

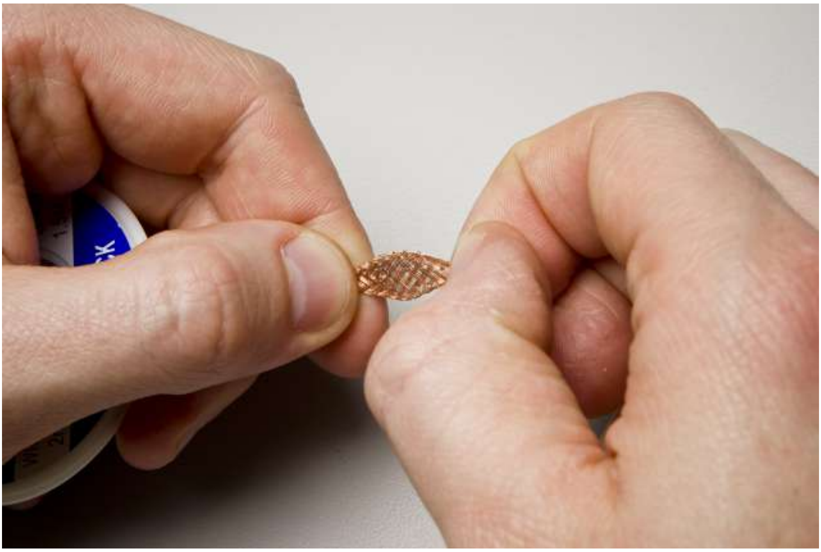

Figure 3-2. *Spreading the wick*

Desoldering with the Wick

Position the wick over the joint, and press down on the wick with the soldering iron as shown in Figure 3-3. The heat will pass through the wick and melt the solder. As this happens, the wick will soak up the solder, removing it from the joint. When you're done, you can trim off the used portion of wick with your wire snips.

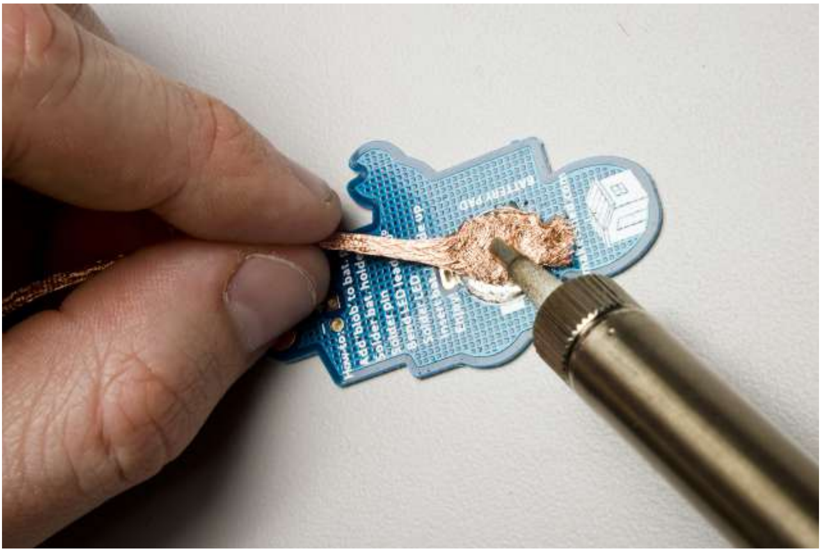

Figure 3-3. *Desoldering*

WARNING: The wick will get very hot, so take your hands off it as soon as you press the soldering iron to it (the soldering iron will hold it in place).

Desoldering with the Solder Sucker

The solder sucker is another tool for removing solder. It's best for removing excess solder, and you can use it to clean things up a bit before you use the wick. To use the solder sucker, press the plunger down until it locks. Use the soldering iron to melt the solder around the joint, and quickly bring the solder sucker's nozzle to the joint (Figure 3-4). Press the button on the sucker, and it will pull the solder right up into the tube. When you press the plunger again, any solder in the tube will be ejected.

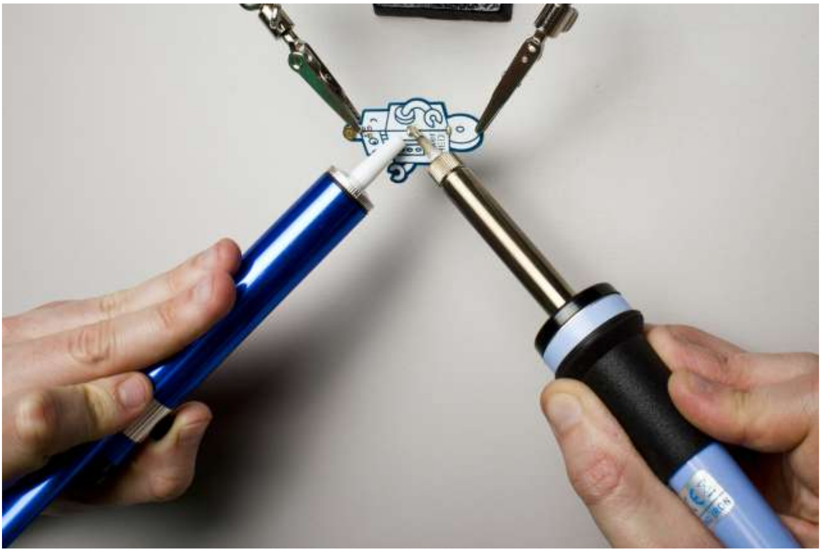

Figure 3-4. *Using the solder sucker*

4/What's Next?

Now that you've learned the ins and outs of soldering, it's time to do more. You can share this book with a friend so they can learn, or you can teach them yourself. And the world is full of wonderful things you can make now that you know how to solder.

Teach the World to Solder

Before you can teach someone to solder, you need to figure out what you want them to try making. You can purchase more Learn to Solder Skill Badge Kits from *http://makershed.com*. Here are a few tips for teaching soldering to others:

Keep Your Hands Off the Tools

You're there to teach, not to solder. And your students will learn best if they always have their hands on the tools. If you grab the tools from them and do part of the work, they won't learn from it.

Be Patient

Try to remember what you went through the first time you tried to solder. It takes time, and you did things incorrectly, and maybe you even made mistakes that required desoldering. Let your students learn from their mistakes, and let them take their time.

No Food or Beverages

Even with lead-free solder, there are chemicals in the flux that aren't good for you. If you allow food or drink near the work area, it could get contaminated. Keep food and drink away from where you're working.

Make Sure There's Plenty of Light

The components you work with while you solder are small, and hard to see. Make sure to use a well-lit workspace for teaching.

Keep Spare Parts Around

Beginners will break things. Pick up some spare resistors, LEDs, and other common components, and have them on hand in case you need to replace a damaged component.

Make More Things

Maker Shed (*http://makershed.com*) carries many electronic kits that you can apply your newly-learned soldering skills to. Here's just a few of them:

Supercap Racer Kit

One of the latest in our popular Mintronics line of DIY kits is this tiny race car designed by George Albercook. Once assembled, this little racer (Figure 4-1) will run around getting its juice from a supercapacitor. Capacitors are like little batteries. Supercaps store and release power faster than ordinary ones, so they're great for making quick-charge gadgets and power-on-demand circuits.

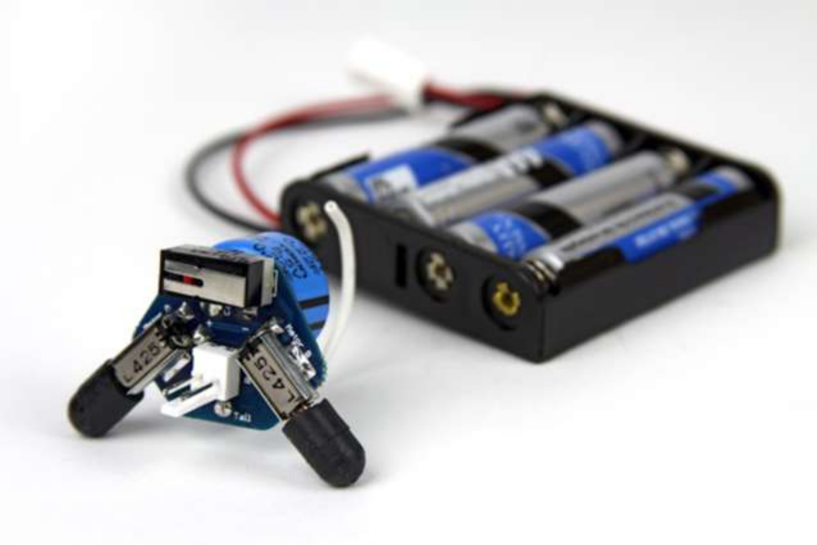

Figure 4-1. *Supercap Racer Kit*

Wee Blinky Kit

The Wee Blinky kit (Figure 4-2) is an easy-to-solder two (2) LED blinker circuit. It comes with a 9V battery snap, but will work with almost any voltage from 3V to 12V. A 9V battery is required but not included. It's tiny, it blinks, and it's a great kit to hone your soldering skills since it's cheap too!

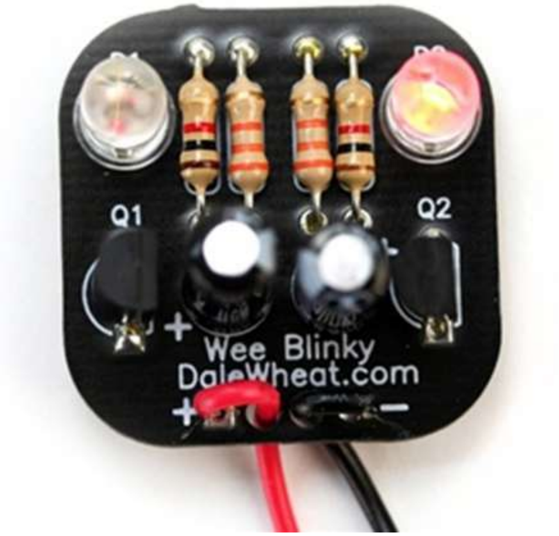

Figure 4-2. *Wee Blinky Kit*

MintyBoost

The MintyBoost (Figure 4-3) is a small and simple USB charger for your iPod or other MP3 player, your camera, cellphone, and any other gadget you can plug into a USB port to charge.

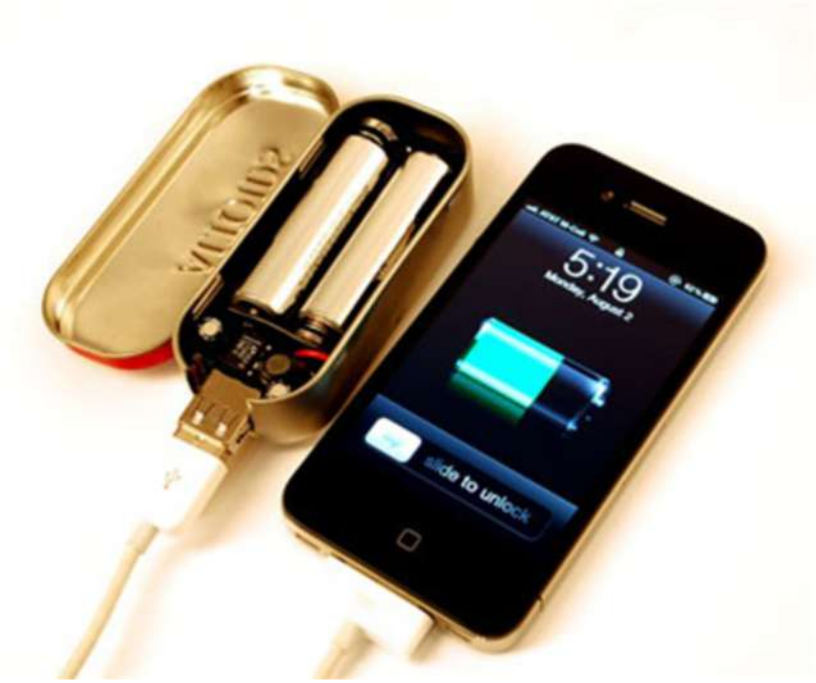

Figure 4-3. *MintyBoost USB Charger Kit*

About the Authors

Brian Jepson is an O'Reilly editor, hacker, and co-organizer of Providence Geeks and the Rhode Island Mini Maker Faire. He's also been involved in various ways over the years with AS220, a non-profit arts center in Providence, Rhode Island. AS220 gives Rhode Island artists uncensored and unjuried forums for their work and also provides galleries, performance space, fabrication facilities, and live/work space.

Tyler Moskowite, a programmer, engineering intern at Make Magazine, and student at Santa Rosa Junior College, has been tinkering with electronics for almost half his life. He picked up Arduino, then Android, and with the release of the ADK he has found his niche.

As a photographer for Make Magazine, Gregory Hayes has ruined more clothes than he ever did as a handyman, hiked more miles with a heavier load than he did as a backpacker, done more research than he did as a writer, and gotten closer to more human hands than advised by any epidemiologist. Taught to solder at the age of seven and forced to solder for his supper at the age of nine, he's now content to let others enjoy the lion's share while he stands by watching safely from behind glass.

How to solder

This pdf refers on informations from the site: http://store.curiousinventor.com/guides/
Have also a look to the 7min. Tutorial video on: http://store.curiousinventor.com/guides/How_to_Solder

Select a Soldering Iron

A 25 or 30 Watt iron should suffice for most small electronics work.

1. Most **soldering "guns" are vastly overpowered for electronics soldering** and can easily overheat components or expose them to harmful voltages. However, some people cleverly use them to solder multiple leads on surface mount devices. Soldering "guns" are for plumbing and much heavier duty applications, and are usually over 100 Watts. The "guns" work by passing high currents through the tips, and these currents can generate voltages that damage electronic components. Also, magnetic fields from guns with transformers can damage some electronics.By forming the heating element in the shape of of the chip, a soldering gun can be used to heat many leads simultaneously.

2. **How much wattage do you need for a particular application and how does wattage relate to tip temperature?**

   A loose analogy: Imagine a car tire has a leak, but you're trying to keep it inflated by pumping air into the tire at the same time it's escaping out the leak. The bigger the leak, the more air you have to pump into it to keep the pressure up. If the tire pressure represents tip temperature and the air lost through the leak represents heat lost through the tip, then wattage represents the maximum amount of air your pump could supply. Once more air escapes through the leak than your pump can replace, the tire pressure (or tip temperature) starts to drop.

   If you had a very small leak and a huge pump (say a 100 Watt iron equivalent), you might be afraid that the pump would cause the tire to explode since so much more air is going in and so little going out. But if you have a nozzle to regulate the pump's air, you could only allow just the right amount of air in to replace what's lost through the leak. This is how "temperature controlled" soldering irons work. As long as you aren't losing more heat out of the tip than the iron can replace (up to its rated wattage), it will automatically regulate just the right amount of heat into the tip to maintain the same temperature.

   However, typical plug-in irons have no such regulation. A 15 Watt iron always delivers 15 Watts of heat to the tip, and the tip temperature stops increasing only when 15 Watts of heat escape through the air. When the tip touches a part, its temperature drops, and if the part you're soldering can dissipate enough heat, the temperature will keep dropping until it won't melt solder any more. After the iron is pulled away from the joint, the temperature will climb again. There is some amount of natural regulation: as the tip gets hotter, it dissipates more heat, and as it gets cooler, it dissipates less.□

   Usually, the bigger the component the more heat it can absorb and dissipate, so the general rule is that you need more wattage for larger parts. If you're just soldering small resistors and ICs, 15 Watts will probably suffice, but you may have to wait a bit in between joints for the tip to recover. If you're soldering larger components, especially ones with heat sinks (like voltage regulators), or doing a lot of soldering, you'll probably want a 25 or 30 Watt iron. For soldering larger things like 10 gauge copper wire, motor casings, or large heat sinks, you may need upwards of a 50 Watt iron or more. The following video shows what happens to tip temperature as 15, 25, and 40 Watt irons solder various sizes of wires and components. For cheap irons, higher wattage does indeed mean higher temperatures!

- What is the difference between cheap RadioShack® irons and more expensive ones like Wellers®? What do $100+ and $400+ soldering "stations" have over the cheaper kinds that plug straight into the wall? expand Among the irons that plug straight into a wall and don't have a separate station, the dirt cheap kinds will work satisfactorily for many applications. From personal experience, the tips on RadioShack® irons often come loose and sometimes can be impossible to remove. The irons can also get uncomfortably hot to hold after several hours of use. The more professional Weller (or other) lines are made for longer, continuous use and have insulation on the handles that keeps them cooler. They

can also take a wider variety of tips.
Soldering iron "stations" usually provide some control over the heat being supplied to the iron tip. Ones that are temperature controlled automatically control the amount of heat delivered to the tip so that it remains at a set temperature. In every iron, when the tip touches a component, some heat is lost and the temperature drops. One measure of quality is the time needed for the tip to regain its temperature. A nice feature of many soldering stations is that the tip heats up in seconds after you turn it on.
Many stations also allow you to hot-swap the iron tip, which can be very helpful if you're alternating between surface mount joints and larger components.

- If standard tin-lead solder melts below 400 °F (and lead free below 500 °F), why do most soldering irons have tip temperatures between 600 and 800 °F? Just what is the right **soldering temperature**? expand The basic reason that tips are so much hotter than solder's melting point is because that difference helps to transfer heat faster to the joint. What is the "correct" temperature is a debatable topic, but a common rule of thumb is to start off at 600 °F and increase from there until acceptable results are achieved. Typical Kester (a solder manufacturer) datasheets recommend 600-700 °F for lead-based solder, and 700-800 °F for lead-free solder. "No-clean" or "low solids" fluxes will burn off before a joint can be made with higher temperatures, so low temperatures (below 700) may be essential for these fluxes.
  From Kester's hand-soldering knowledge base: "When hand soldering with a rosin flux such as the Kester #44 or the # 285 the recommended iron tip temperature is 750°F. If you are soldering with a low residue no clean solder such as the #245or # 275 we recommend a tip temperature of 600-650°F.
  What are acceptable results? The goal is to heat up the parts enough so that solder will adhere to them and form a good bond. The higher the iron temperature, the faster it will heat up the parts, so why not set it extremely high to work faster?
  Besides the obvious increased risk of overheating components and the board, higher temperatures cause the iron tip to oxidize faster and can significantly reduce its life. Some claim a 10 °C rise reduces tip life by half (ref p.33). For occasional use, though, tip life may not be much of a factor, especially if the tip is kept covered with solder at all times.

- **Tip size and shape:** a basic guide is to pick a tip that's slightly smaller than the pad you're soldering to. From there, you want a tip with a large thermal mass and short stroke (why?) In most soldering irons, the tip is not actually the heater, but sits in between your work and the heater. You can think of it like a heat bucket that empties into your work and gets filled again by the heater. Typically touching a component empties heat out of the tip much faster than the iron can replace it, and if you have a small bucket (tip), the temperature will quickly drop to an ineffective level.
  Especially if you have a small wattage iron (15 Watts or less), the temperature will drop before you can heat up a larger part, or you'll have to wait a bit in between joints for the tip temperature to recover. With a bigger bucket (tip), you can handle larger joints with smaller wattage, but eventually you'll need to step up the wattage.
  The "stroke", or length of the tip should be minimized to get the heater closer to the work; it takes some time for heat to transfer through the tip. This is balanced with the need to get into tight places where you need a longer tip.
- What do common tip shapes look like and what applications are they best for?

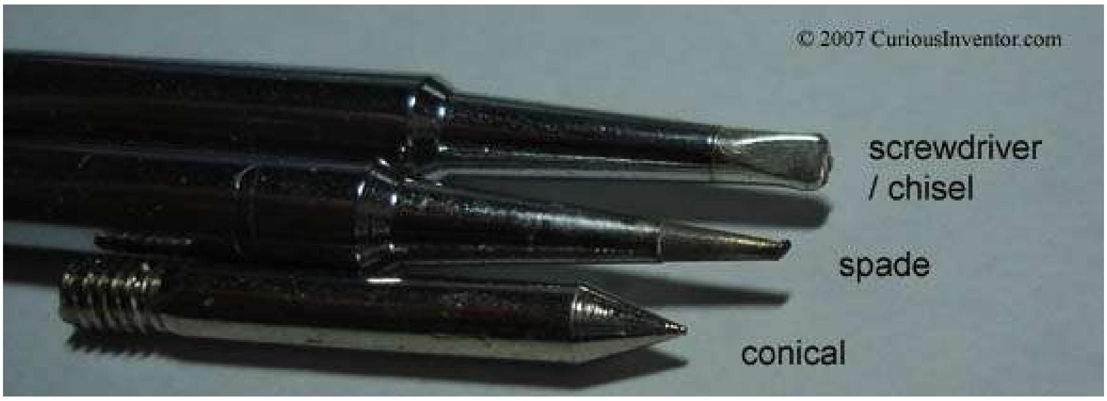

Screwdriver, spade, and conical are some of the more common tip shapes. Personal preference is the biggest factor when choosing a tip, but the goal is to get as much surface area contact between the tip and work as possible. Chisel and spade tips have more surface area at their ends, and also "hold" solder at their tips more readily than conical tips, which have a tendency to draw solder away. Even for fine pitch surface mount soldering, having a small flat at the end can be helpful.

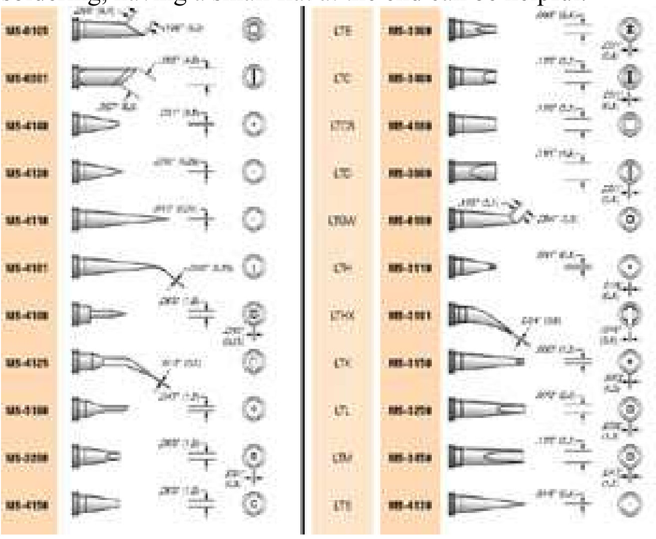

- There are myriad other tip shapes and sizes. The picture to the right shows one Plato catalog page of many. Some other non-standard shapes include a knife-blade (useful for fine pitch leads) and a surface mount desoldering tip.

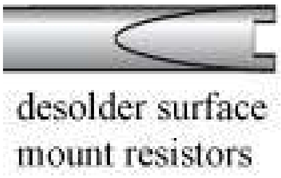

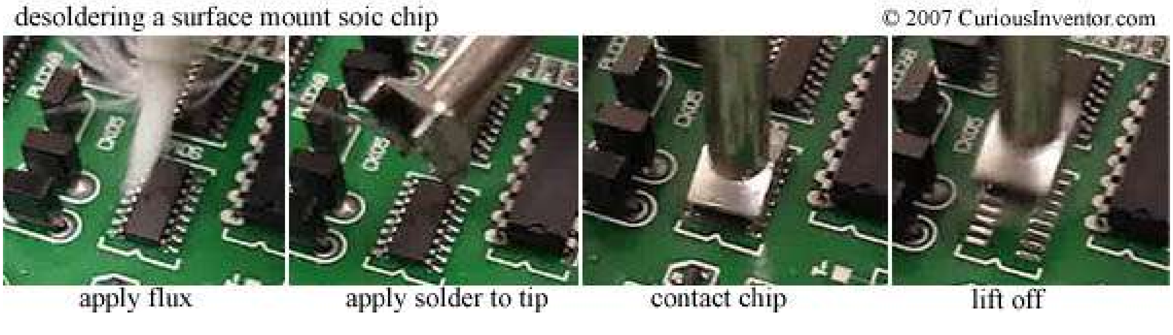
desoldering a surface mount soic chip
© 2007 CuriousInventor.com
apply flux
apply solder to tip
contact chip
lift off

To preserve tip life, the number one thing you can do is reduce the tip temperature (if your iron allows this). After that, ALWAYS keep a layer of solder on the tip to prevent the tip itself from oxidizing, and clean it in between uses. Put a glob of solder whenever you put it back in the stand, and before you turn it off. When heating up a new tip for the first time, hold solder against it so the tip can be covered as soon as the iron gets hot enough.

The longer flux residues and oxides are left on the tip, the harder they are to clean off. They also can drastically reduce the tip's ability to heat up a part, and prevent solder from "wetting" the tip. Regular cleaning of the tip before use is one of the best ways to prolong tip life and make soldering easier. It's important that solder "wet" or cling to the surface of the iron--without solder in between the tip and work the tip's ability to heat is drastically reduced.

- What about gas powered irons and the Cold Heat® iron that is supposedly touchable 1 sec. after use? Butane (and other gas) powered irons are mainly used in situations where electrical power isn't

available. Weller sells some battery powered irons as well.
Everyday Practical Electronics gives a pretty damning review of the Cold Heat iron here, in addition to having one of the better how-to guides out there. To summarize, the Cold Heat® iron has a forked end that you must bridge with the work or solder to turn on the iron, so it can be hard to hold it in a place that keeps it on and also effectively heats the part. Many people complain about pushing harder to make a good connection and then having the brittle tips break. Running power through your work to heat it may not be the greatest idea with some parts. Finally, the iron doesn't get hot enough for a lot of jobs, or cool enough to do anything like throwing it in your pocket right after use. But for something that's portable and cordless, heats up and down in under a few seconds, maybe it's worth the price ($20). Weller's battery powered ($20) iron doesn't have a forked end and supposedly heats up in under 15 seconds, but I don't know about cool-down time.

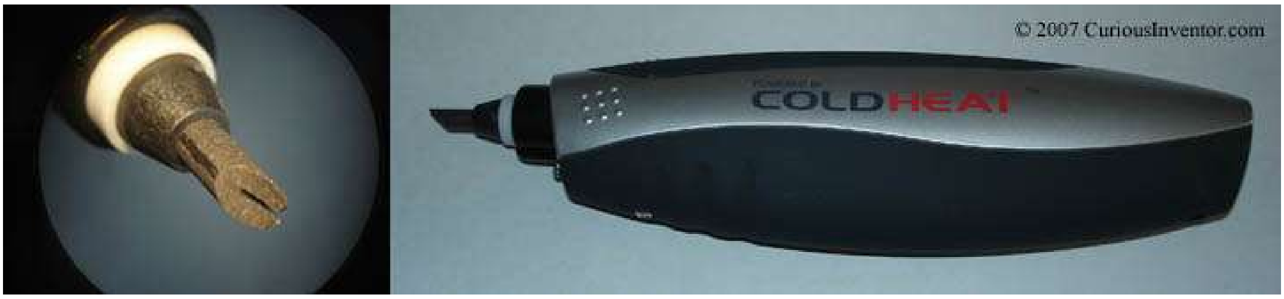
© 2007 CuriousInventor.com

What kind of solder (rosin cored, etc. lead-free)? What is flux and when is it necessary?

As a starting place, for most small electronics soldering, 1/32 inch (.03) rosin-cored, 60/40 (tin-lead) or 63/37 solder should work fine. Rosin-cored lead-free is fine, too. Unless you have reason otherwise, don't use "no-clean" solder--it's very likely that you don't need to clean the regular rosin-cored solder. The solder should be thin enough to prevent accidentally applying too much (and causing a solder bridge), but thick enough so that more doesn't have to be gathered from the coil too often.

Flux cored wire at 30X magnification
© 2007 CuriousInventor.com

Besides affecting your feed-rate and convenience, the solder thickness also relates to the amount of flux that is delivered. Flux is basically a weak acid that removes oxides so that solder can adhere to the metal, and is so essential to the soldering process that it's built into the core of common wire-solder. It also helps the solder spread out (reduces surface tension), transfer heat, and acts as a protective blanket to keep oxygen away from the metal until solder displaces it.

For the most part, manufacturers include a sufficient amount of flux in the wire, but if you use an extremely thin wire there may not be enough to clean the joint OR the iron tip. Consider using a thicker gauge for cleaning the tip periodically if you're using especially thin solder. Liquid flux is helpful for SMD soldering, too.

When picking a wire-solder, there are 4 features to decide on: flux type and amount (% weight), alloy (tin-lead, lead free, silver bearing, etc.), thickness and total amount (1oz, 1lb?).

- **Flux:** Just what is flux, what kinds are there, and when do I need liquid flux?

  Why it's needed: Solder doesn't just freeze on a joint, it actually forms a metallurgical bond by dissolving and chemically reacting with the base material. Unfortunately, almost all metals oxidize in air and form an oxidized layer that prevents solder from wetting and bonding to them. What is oxidation?

  Oxidation, is when Oxygen (or other oxidizing agents like sulfur) atoms combine with base materials, stripping loosely attached electrons and forming new compounds like iron(III) oxide. It is what happens when apple slices turn brown, iron rusts, copper turns black / green and unsolderable. (read references for more accurate/complete explanation)

  Results vary significantly. When iron rusts, the oxides flake off until no iron is left. Alternatively, aluminum oxidizes extremely rapidly, but is then protected from additional oxidation by the oxide layer. This layer makes Al impossible to solder without using special solder and extremely aggressive flux, or plating the surface with a solderable metal, like nickel. The chromium in stainless steel fulfills the same function, oxidizing to form a protective barrier that is difficult to solder. Gold remains shiny because it doesn't oxidize, and is easy to solder to, but forms brittle joints. Heat, moisture, and salt all increase the rate of oxidation.

Oxidation can add a hidden cost to components and boards that may have been sitting on the shelf for long periods of time, or have been exposed to hot, humid environments. Copper pads on PCBs (printed circuit boards) are covered with solder or are plated to prevent oxidation, but given enough time, oxygen can still penetrate these barriers. Surplus parts in particular may need a bit of steel wool.

Some interesting references:
rusting chemical reaction
fairly understandable explanation of oxidation
wiki/Corrosion
wiki/Oxidation

Oxidation occurs much faster at higher temperatures, so even if you somehow had clean metals to start with, you would still need flux to prevent new oxides from forming while soldering.

The main choice to make when deciding on a flux, whether it comes in a cored wire or a liquid or paste form, is how aggressive it should be. The more aggressive or "active" the flux, the harder the oxides it can remove, and the faster it can remove them. Going from weakest to strongest, typical choices for hand soldering applications include: "no clean", RMA (Rosin Mildly Activated), RA (Rosin Activated), and water soluble. A newer classification system (J-STD-004) has recently been adopted and classifies fluxes not by rosin content, but by activity, material, and halide presence.

The new system classifies flux by material (RO=rosin, RE=resin, OR=organic, IN=inorganic), activity level (Low, Moderate, High), and halide presence (0 or 1). No-clean, rosin-based no-clean fluxes might be labeled ROL0 or ROL1. Although there is no direct translation between the old system and new, most R and RMA fluxes fall under Low activity level, RA are generally labeled as Moderate activity, and water soluble are High activity. (source IPC-HDBK-001 www.ipc.org)

The downside to choosing a more aggressive flux is that the residues left over after soldering MAY be corrosive, conductive, or enable fern-like growths called "dendrites" to grow between connections. A brief description (p.29) of dendrite growth and some great pictures at the end of this paper.

Because of the risk of corrosion and dendrite growth, most manufacturers clean off the residue from RMA and RA fluxes, and some even clean "no-clean" residues. The question of what flux to use and whether / how to clean it is quite involved.

Rosin flux is quite an interesting animal. Made from pine tree sap, at room temperature, it's an excellent insulator and non-corrosive. When it hits 226 °F it begins to turn acidic and attack oxides, but then when it cools it supposedly leaves residues that are again inert. Kester's data sheet for "44" flux (classified as RA and ROM1) claims no cleaning whatsoever is necessary. I haven't read of any manufacturers that would use a RA (or even an RMA) flux and not clean it--the military won't even use a RA flux WITH cleaning because of the risk that some will be left behind. This Chemtronics author recommends cleaning even the "no-clean" fluxes. He also points out that even if the residue is non-corrosive and non-conductive, it might be tacky and attract dust that causes a short.

To add one more piece to the puzzle, flux generally gets used up during the soldering process. This is why no-clean fluxes are oftentimes ineffectual for lead-free soldering, which can require slightly higher temperatures and longer heating because the lead-free solder "wets" slower. The no-clean flux can burn off before the joint is complete. Alternatively, if you apply liquid flux far from a joint, it may still be active (corrosive) if it never got heated.

I'm not making missile guidance electronics, I'm making a robot that pours a beer, what flux should I use and do I really need to clean? Even manufacturers of non-life critical electronics have much more stringent reliability requirements than an individual. They must ensure that tens of thousands of products will work for multiple years, not a single project.

The safe advice is to use the least aggressive flux that enables solder to quickly wet or cling to the surfaces, and then clean off the residues with alcohol and lint-free wipes (don't just spread the flux around). Try starting with a rosin-based mildly activated flux: RMA. I am inclined to trust Kester's spec sheet for "44" (RA) flux that says it does not actually require cleaning. Other flux manufacturers may have RA or RMA level fluxes that do indeed need to be cleaned, so if you don't know what you're using, cleaning is probably prudent. If you are going to clean rosin fluxes, do it soon after soldering because they quickly harden (see pics under 'Cleaning'). Finally, I would personally avoid no-clean fluxes and solder unless you have a critical application and very clean parts.

Lead-free solder generally requires a made-for-lead-free flux designed to be used under slightly higher temperatures.

Liquid flux can greatly help with surface mount soldering and desoldering components, but the flux inside cored-solder should be sufficient for through hole components. When soldering SMD components and desoldering pretty much anything, liquid flux acts as a blanket that helps to spread heat and also keep oxygen away from the metals. Finally, flux lowers the surface tension of solder, helping it to spread out and wick into connections.

A water-soluble flux may be necessary for heavily oxidized parts or difficult metals like nickel. Without question, clean these fluxes. Special fluxes and solders exist for aluminum and stainless steel and these also certainly require cleaning. Never use acid-core solder; it deposits zinc chloride into the solder that cannot be cleaned out. A final reason to clean flux residues is if you'd like to apply a conformal coating and aren't sure whether it will adhere to those residues.

Some more references:

- Never use acid-core and how to solder to stainless steel (Kester)
- White residue and all about rosin (more Kester)
- Good mini-class on fluxes: Bolton University
- To clean or not to clean and a brief history of electronics cleaning: more Bolton
- The Kester catalog provides good info on the solderability and flux requirements of various materials (see p.14).
- Flux residues and what to do about them. This explains a bit about cleaning options and the health risks of rosin and non-rosin fluxes.

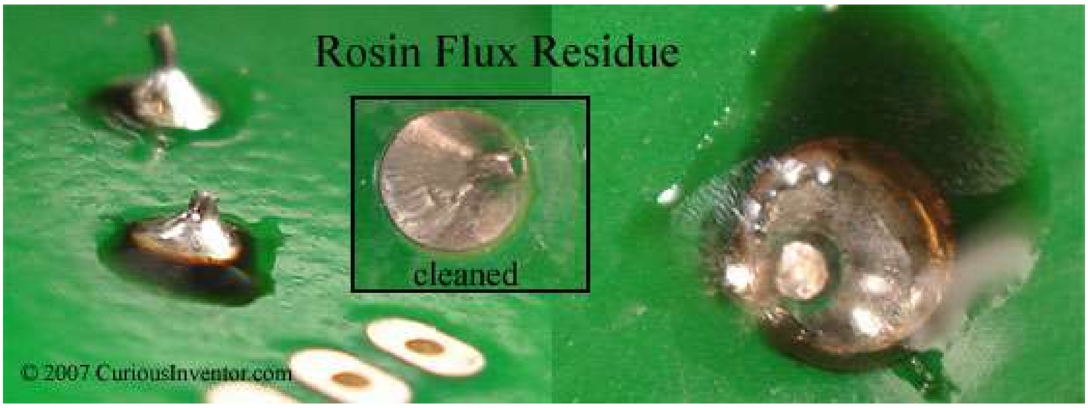

- **Alloy:** 60/40, 63/37, tin-lead, lead-free, silver bearing, RoHS, eutectic, oh my...

- **Standard lead-based solder** is made of tin and lead. When you see 60/40 or 63/37, it means 60% tin by weight, 40% lead. Either once of these alloys should be fine for typical small electronics soldering. 63/37 in bulk is slightly more expensive because of additional tin, but has the special property of being a "eutectic" alloy, which transitions from liquid to solid at one temperature (like water) instead of range of temperatures. Basically, in non-eutectic alloys like 60/40, there is a "pasty" region of temperature where portions of the solder are frozen and other portions are liquid. What does this mean for soldering and is 63/37 really that much better?

  Alloy metals have some interesting properties that are different from the metals comprising them. In tin-lead solder, the mixture has a lower melting point than either lead or tin alone, and the melting point varies depending on the portions. The mixture that yields the lowest melting point is called eutectic. This is also the only mix where all the constituents melt and freeze at the same temperature.

  If the tin-lead alloy isn't eutectic (ie, if it is not 63% tin), it will go through a "pasty" phase while it freezes. Unlike water, which freezes entirely at 0 °C, some parts of a non-eutectic mixture of tin-lead freeze at higher temperatures than other parts. For a somewhat simplified explanation, if you held the temperature of 60/40 slightly above 361 °F, the "extra lead" would solidify and be floating in a liquid 63/37 eutectic mix. For a more exact and great explanation of this process, look here.

  How is it that the mixture of two elements somehow lowers the melting point? And I quote: "increased entropy." Chew on this. (Another great phase-diagram explanation with a bonus of why ice and salt can get almost 30 °F below freezing--enough to freeze ice cream) And one last great explanation--talks a bit about grain structure and how solder isn't a simple homogeneous mixture of tin and lead.

  This compares the difference between eutectic freezing in tin-lead and eutectic freezing in more complex 3 element lead-free alloys.

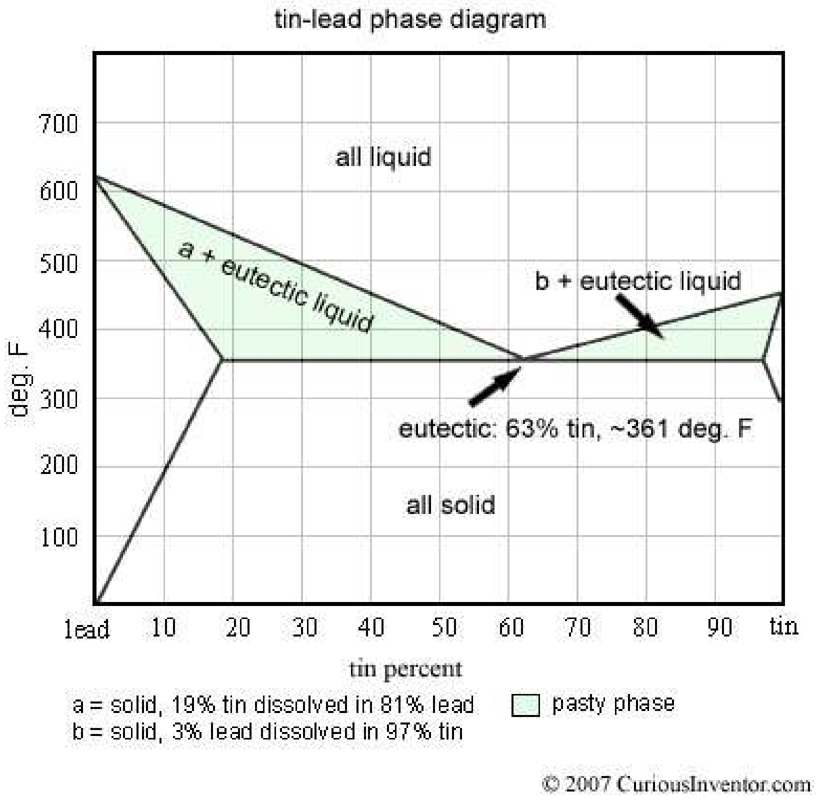

So which is better, 60/40 or 63/37? A decade or so back (before the lead-free movement), most manufacturers incurred about a 5% increase in cost to switch from 60/40 to 63/37. 63/37 flows slightly easier, makes shinier joints, and has a faster total freezing time which means there is less risk the joints will be "distrubed," which is what happens when the joint moves during solidification. This can lead to internal fractures that cause poor electrical connections and unreliable mechanical joints. Note that 63/37 doesn't freeze instantly (just like water)--it still has a window of time during which the joint can be disturbed, too.

I think most would say that the enhanced properties of 63/37 really only matter for mass soldering operations like wave or reflow soldering, and that there is little difference for hand soldering. A dull joint is more often caused by insufficient heat, dirt or oxides, or lack of flux rather than alloy makeup. (A no-clean flux may burn off before the joint is complete) Holding everything else constant, the difference in shininess between 60/40 and 63/37 is completely cosmetic. If you'd like to see a shiny 60/40 joint, try using Kester 60/40 with #44 flux.

- **Silver bearing solder:** (that is, contains silver, not for roller bearings) Silver is used in one of the leading alloys for lead free solder (An96.5% Ag3.0% Cu.5%) and also as an addition to tin-lead solder, usually in the 2-4% range (when you se 62/36/2 this means Sn64Pb36Ag2).

  People claim that it flows better, has a lower melting point, is stronger, and has a higher conductivity. According to Indium's solder wire data sheet, their 2% silver solder has an electrical conductivity that is 11.9% of Cu compared to 11.5% of 63/37 tin-lead solder, a shear strength of 7540psi vs. 6200psi, and a tensile strength of 7000psi vs. 7500psi for 63/37. So, yes, the claims are true, and also mostly insignificant. Silver was initially added to solder to prevent silver platings on component leads from dissolving into the solder ("silver migration") and forming brittle joints. Having silver in the solder reduces migration, so you may want to use it on silver joints. (Note: this logic doesn't entirely make sense to me. If silver getting in the solder caused embrittlement, how does adding *more* silver prevent this?)

  Audiophiles seem to be enamored by 4% silver bearing solder, namely some from WBT, Cardas, and WonderSolder. Are these really better for audio?

  The superior claims include things like higher purity, eutectic alloys, higher conductivity, and better flux. I haven't found any controlled studies showing that a group of people can actually hear the difference, so I'm skeptical. Although additional silver does increase conductivity, the increase is small and the joint distance over which that conductivity applies is also extremely small. For what it's worth, here's a forum discussion that discusses a bit about solder in high-end audio and also a FAQ (scroll down)concerning solder on Cardas Audio's site. Another decent discussion.

  Bad joints made with any solder can create a high-resistance connection, especially if the underlying components were heavily oxidized initially. My advice: if it makes you feel better, get it, but be wary of sellers that don't provide spec sheets.

- **Lead-free Solder:** As of July 1st, 2006, European laws mandated that new electronics be almost entirely lead free. As of yet, there are no US laws (outside CA) mandating the removal of lead, but most manufacturers are switching over for competitive reasons. More on RoHS, WEEE, and lead risks:

  The European Union passed directives in 2003 stating that no equipment sold in Europe should, by July 2006, have more than .1% lead in any homogenous component (like a solder joint). The directives are known as the WEEE (Waste Electrical and Electronic Equipment) and RoHS (Restriction of Hazardous Substances) (wiki). There are corresponding laws in China, Korea, and California. Japan manufacturers actually voluntarily begin switching to lead free years before RoHS or WEEE for competitive reasons. IPC's informative lead-free website: here.

Although only .5% of lead used in the US gets embedded in electronics (verses 80% in batteries), there is concern that the lead from those electronics will leach into ground water supplies from landfills. Why the concern over solder joints when batteries contain so much more lead? For the most part, lead containing batteries are recycled and regulated, whereas electronics are routinely just thrown away. The EPA claims that 1% of municipal waste is electronics. Interestingly, according to this publication by IPC, no studies have found any evidence of lead getting into the ground water from landfills.

Lead health risks: Lead does not get absorbed through the skin, and is actually not present in solder fumes to any appreciable degree (fumes are still bad for you, see fumes section below). The greatest risk of hand soldering with lead comes from ingesting lead by eating or smoking without first washing. Health risks include increased blood pressure, fertility problems, nerve disorders, muscle and joint pain, irritability, and memory or concentration problems. The latest health data indicates that there is no amount of lead that will not be detrimental to health. Google lead or start with this link.

Some great lead-free joint pictures: here.

The most popular lead-free alloy seems to be Tin 96.5% Silver 3.0% Copper .5%. The wiki page on solder mentions several different lead-free varieties.
-AIM lead-free solders.
-Huge list of Indium lead-free solders and their properties.

Lead-free solder generally melts at a higher temperature, and doesn't wet as quickly to metals. (Eutectic tin-lead solder melts at 361 °F and the SnAg3Cu.5 melts at 423 °F.) Manufacturers generally recommend setting soldering iron temperatures between 700-800 °F for lead-free instead of 600-700 °F for tin-lead soldering. The 15 Watt RadioShack® iron I had operated a bit below 500 °F, so soldering should be possible with it, but maybe slow. Technique wise, since lead-free wets slower, joints will take longer (upwards of 4-7 seconds), but this doesn't mean the soldering iron temperature should be turned up excessively--patience is better than higher temperatures. If you're going to use a lead-free solder, get a flux that's designed for the higher temperatures--the regular no-clean fluxes will likely burn off before doing their job.

A quick word on reliability. Some say that lead-free joints are stronger, and while the material is indeed stronger, it's less flexible than lead-based solder, so expansion and contraction due to temperature change has been shown to break components held by lead-free solder. It seems true mechanical reliability of lead-free vs. lead depends heavily on the situation (see p.30). There is also concern of something called "tin whiskers." These are extremely thin crystalline growths that grow perpendicularly out from surfaces. These took down some space systems and NASA has a great page here. These are different from dendrites (which grow on the surface) and appear to be more likely on bright all-tin platings. Most component platings used to consist of a tin-lead mixture, and since all-tin platings are a common lead-free replacement, people are concerned. I have yet to find any literature that points to TinSilverCopper solder as a risk factor, though. I believe it is a plating issue, only.

One more link: -Why tin or silver?

- **Thickness and Amount:** As a general guide, .032" thick solder (21 gauge) should be suitable for through hole soldering and some surface mount soldering. For finer pitch surface mount devices, use .02" or .015", and if you're soldering a lot of switch terminals, or tinning thick gauge wire you may want .05". If you use .015" solder consider having some thicker solder on hand to re-tin your tip, since the amount of flux in .015" may not be enough to remove tip oxides. The picture below shows how the various thicknesses compare next to the standard .1" spaced DIP pins.

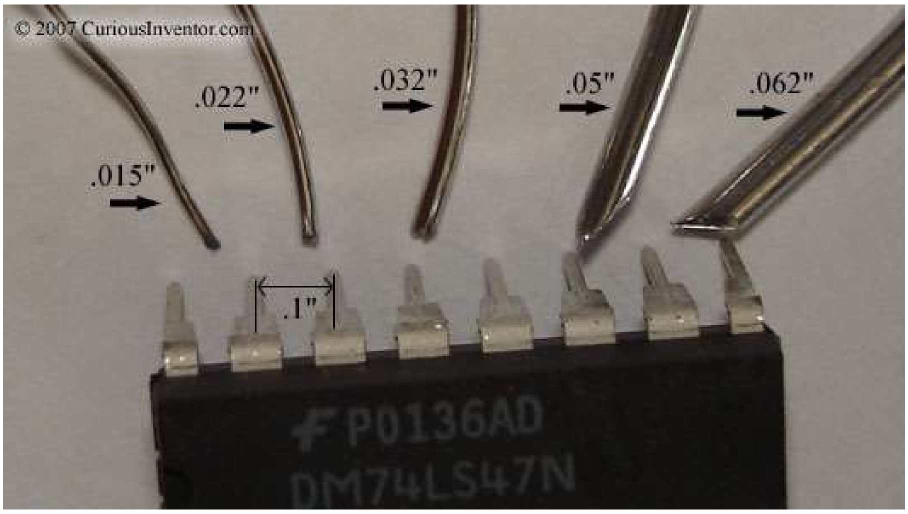
© 2007 CuriousInventor.com
.015"
.022"
.032"
.05"
.062"
.1"
P0136AD
DM74LS47N

Expand to see how .032" and .015" solder compare to a SOIC surface mount chip and fine pitch (.02") device.

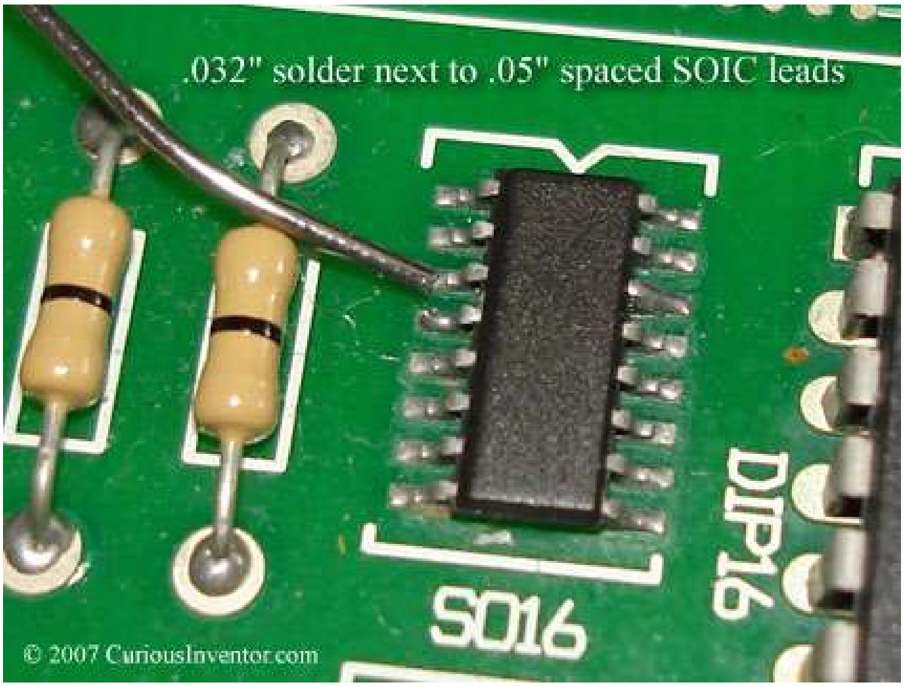
.032" solder next to .05" spaced SOIC leads
SO16
DIP16
© 2007 CuriousInventor.com

.015" solder next to
.02" spaced surface
mount leads and a thumb
tac.
© 2007 CuriousInventor.com

How much solder do I really need? An ounce? A pound? How long will a pound last?

To get a very approximate feel for just how much solder is in an ounce vs. a pound, I measured how much .032" solder it took to attach ten 16 pin DIPs. Trying to provide an upper bound, the soldering is excessive, and a generous glob was placed on the tip in between each chip to account for solder used to tin and protect the tip during normal use.

10 16 pin DIPs took about 15 inches of .032"
solder, or about .052 oz
RadioShack
RadioShack
© 2007 CuriousInventor.com

The ten chips took about 15" of generously applied .032" solder that weighed in approximately at .052oz. This particular solder weighed about .00348oz/in. At about 3.27e-4 oz/joint, a 1 oz spool should last 3060 joints, a half pound should last 25 thousand joints, and a pound should last about 50 thousand joints. Mileage will certainly vary with different sized solders, joints, tinning wires, and highway vs. city driving, but if you're not in a production environment, a half pound should last a while.

Solder Fumes:

What is exactly in solder fumes? Am I safer using lead-free solder?

Flux from rosin-cored solder boiling under magnification
© 2007 CuriousInventor.com

Lead boils at over 3000 °F, and in most cases soldering tips should be kept below 750 °F, so it is highly unlikely that gaseous lead is present in the fumes. The fumes are actually from the flux boiling, which still isn't great for you--many of the chemicals found in cigarette smoke are found in flux fumes: formaldehyde, toluene, alcohols, and hydrochloric acid to name a few. Most of the public health literature indicates that asthma is the major health risk from soldering fumes (not cancer or lead poisoning). When acquired, it is permanent and can cause hyper sensitivity so that even small amounts of fumes bring on attacks. Surprisingly, scientists have not been able to determine what exactly in the fumes cause the health defects, nor what amounts are harmful. Yet, the British health department has set exposure limits of .05 mg/m^3 over 8 hours and .15 mg/m^3 over 15 minutes. I believe these limits have been shown to provide a safe work environment and also one for which the necessary systems / filters are financially reasonable.

Some informative links:

- Solder Fumes and You A British health department pamphlet explaining the health hazards of rosin-based flux fumes (irritation, headaches, dermatitis, asthma) and what precautions employees and employers should take. Note the total lack of any mentioning of lead poisoning.
- Workplace Exposure to Rosin-based Solder Flux Fume During Hand Soldering A study done by the UK Health and Safety Laboratory measuring exposure levels and also the effectiveness of various exhaust, ventilation, and filter systems.
- Measurement of the Performance of Air Cleaners Against the Particulate Element of Rosin-based Solder Flux Fume Another UK Health study investigating the effectiveness of various fume extraction and filter systems. Most interesting finding: although activated carbon filters can remove gaseous hazards, they are largely ineffective for fine particulate in the fumes which they believe to cause much of the harm. Some combination of carbon and HEPA filter is needed, and even these are useless without sufficient air flow.

Returning to the topic of lead, it is widely agreed that eating, smoking and drinking without first washing is the greatest risk factor. Despite the high boiling point of lead, there is also agreement that at least a small amount of lead particles are indeed present in the fumes. The conspicuous lack of emphasis on lead poisoning in all the research done by the UK health department implies that these particles are of little concern. Sentry Air Systems has a brief page that is one of the very few sources I found to claim that lead particles under normal soldering conditions are harmful. The company sells fume extraction technology.

The material safety data sheet for Kester #44 cored solder says under the fire fighting section: "Melted solder above 1000 °F will liberate toxic lead and/or antimony fumes."

According to IPC's DVD-11, "General Safety in Electronic Assembly," when solder is heated past 850 °F the lead can become atomized and end up in the fumes. video link (if link breaks, you may have to search for DVD-11 at www.ipc.org).

Useful comments from someone in the manufacturing world regarding lead.

It would seem that, for typical lead-based, rosin cored solder, the risks are probably not that great from the fumes if you only solder occasionally, don't use abnormally high temperatures, and are in a well ventilated area. If ventilation isn't too good, and you're soldering for long periods of time, the cheaper foam-type carbon filters may not be good enough.

But what about lead-free solder? Lead-free solder often requires higher temperatures and more active fluxes, and both of these factors lead to significantly worse fumes.

Fume Extraction Becomes More Important in a Leadfree Environment - from the Weller blog

Another excellent article on the increased risk of lead-free fumes from OK International.

Instructables has all sorts of home-made fume extractors.

Prepare the Work

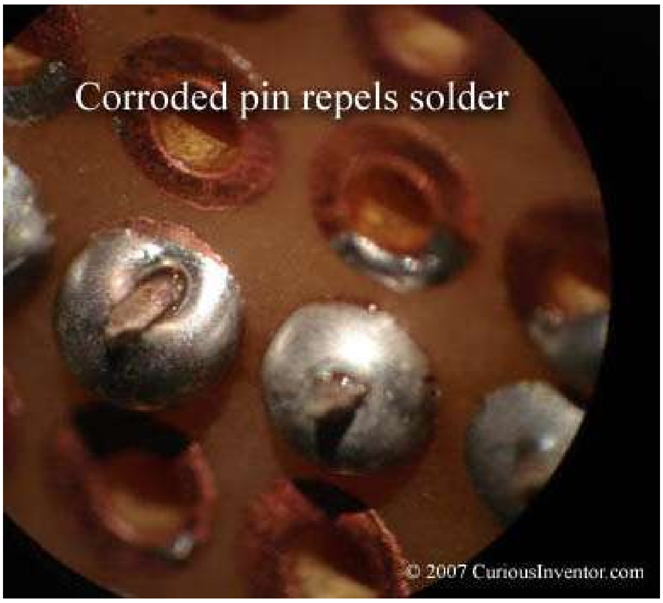

- **Start with clean components:** Flux can remove small amounts of oxides, but will be of little help for heavy oxidation, grease, oil or dirt. Notice how the solder in the adjacent picture has been repelled by the heavily oxidized pin. It may be necessary to lightly use steel wool or fine grit sand paper to remove especially bad oxides. Some people say that you should not do this because it creates scratches that can promote future oxidation... sand at your own risk. Use Silicon Carbide sandpaper (black) as opposed to Garnet (brown, for woodworking) sandpaper because the Garnet paper will shatter and become embedded in the metal. An effective and gentle alternative is to use a pink eraser, especially for copper traces.
- **Clamp your work:** PanaVise makes a popular clamp that accepts several different attachments for holding different sized circuit boards. It's by far the most popular clamp and is also very sturdy. Having the work held in place is especially helpful for desoldering when it's necessary to push or pull a bit. The alligator hands are a cheap alternative.

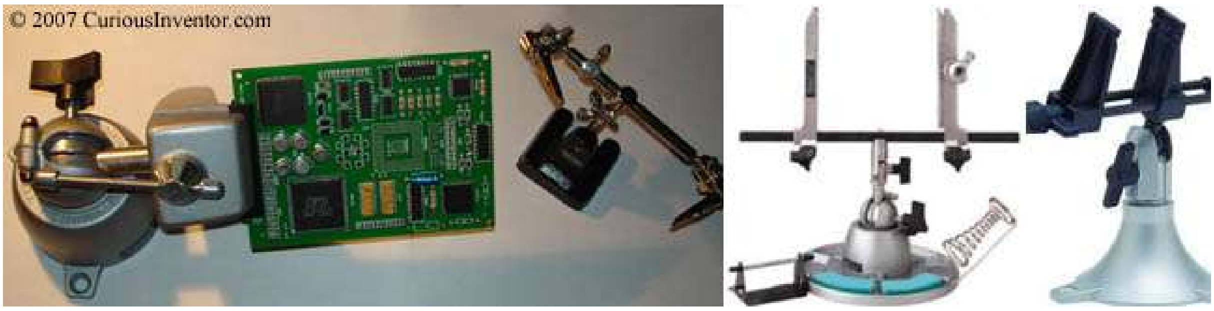

- **Wire preparation:** Tin stranded wires so they don't "bird-cage," or bend out from their original lay. Expand for instructions on the correct way to strip a wire manually, use an automatic stripper, and tin wire.

- **Manual wire stripping:** The natural way you would think to use the wire strippers is to first cut the insulation, and then use the same cutting hole to push off the insulation slug. In production environments where reliability is critical, this is not allowed because it's very easy to nick the wires this way. The pliers at the tip of the stripper are actually meant to remove the insulation slug. First score the insulation with the cutting hole, and then pull it off with either the pliers or by hand. If removing by hand, twist slightly while pulling to keep the strands in place. This is the "right" way to strip a wire, if you're careful, pulling the insulation off with the cutting hole is probably fine.

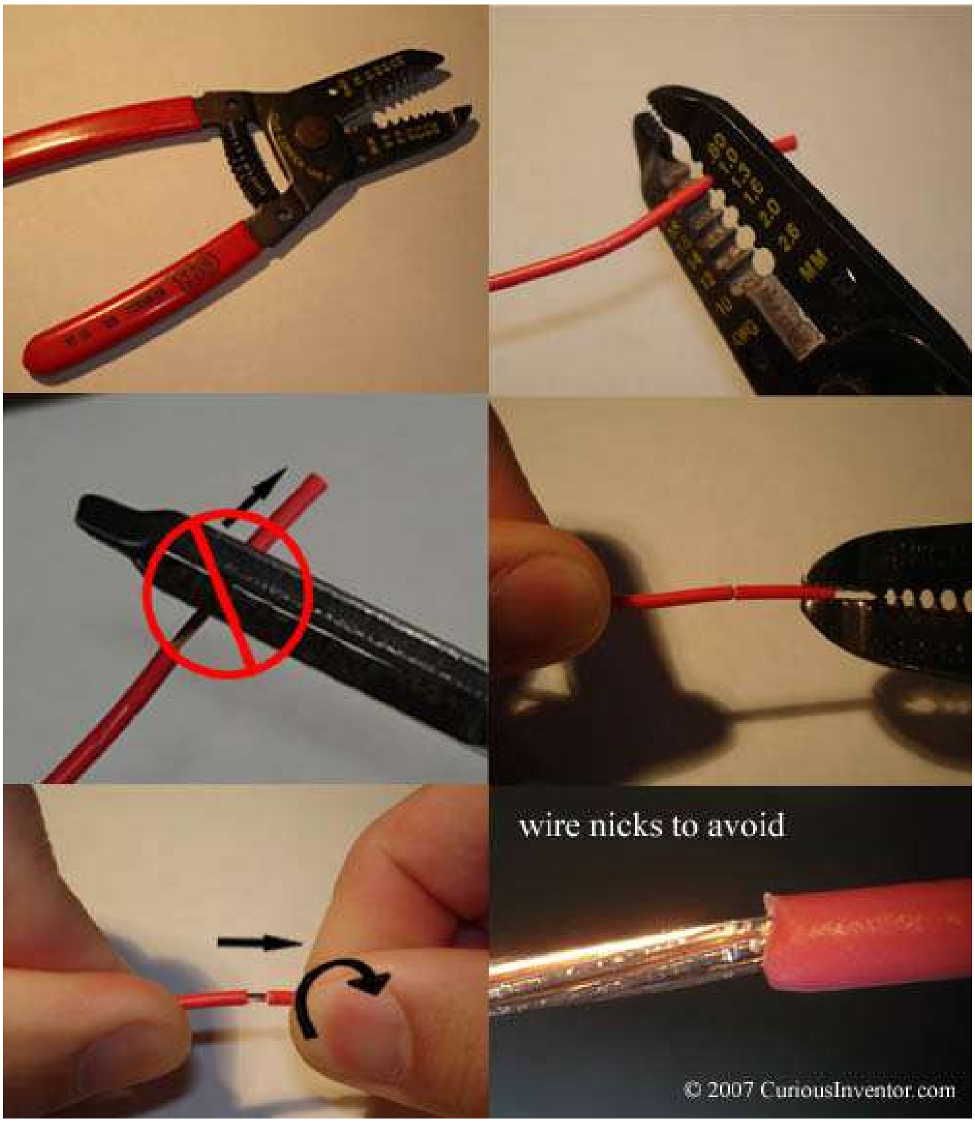

- **Automatic stripping:** With one squeeze these cut and pull off insulation.

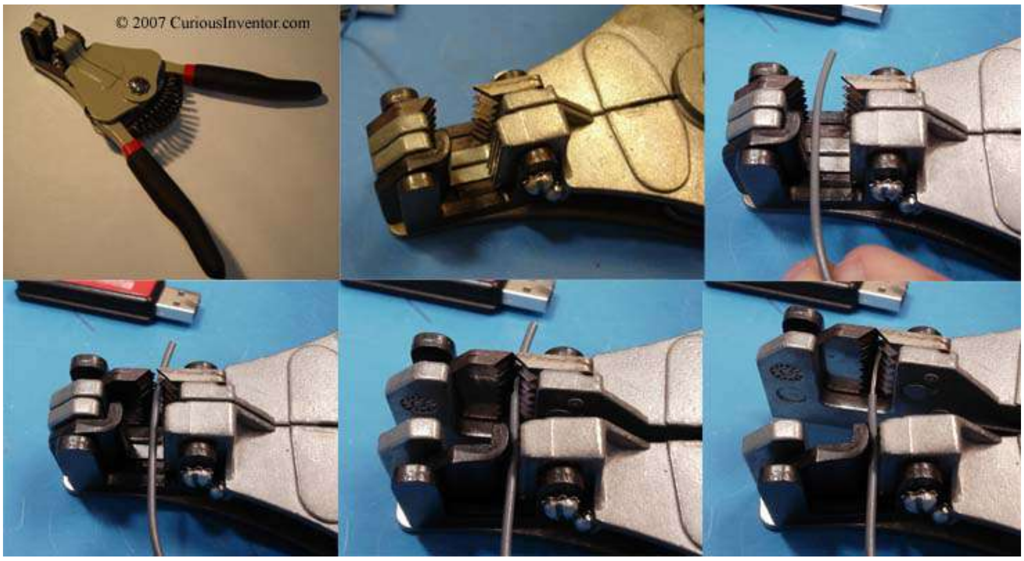

- **Tinning:** Just like soldering a joint, the key here is to apply solder to the wire, not the iron tip. If the wire is hot enough to melt the solder directly, then it's hot enough to form a good bond with the solder. Hold the iron against one side of the wire and apply a small amount of solder in between the tip and wire to form a heat bridge (if there isn't already enough solder from tinning the tip). Now apply solder to the opposite side. The solder should wick into the strands. Move the iron and solder as necessary to coat the surfaces. You should still be able to clearly see the individual strands when it's complete.

  Don't tin wires that are going to be held in clamping terminals since the solder will slowly deform under pressure and cause the connection to come loose.

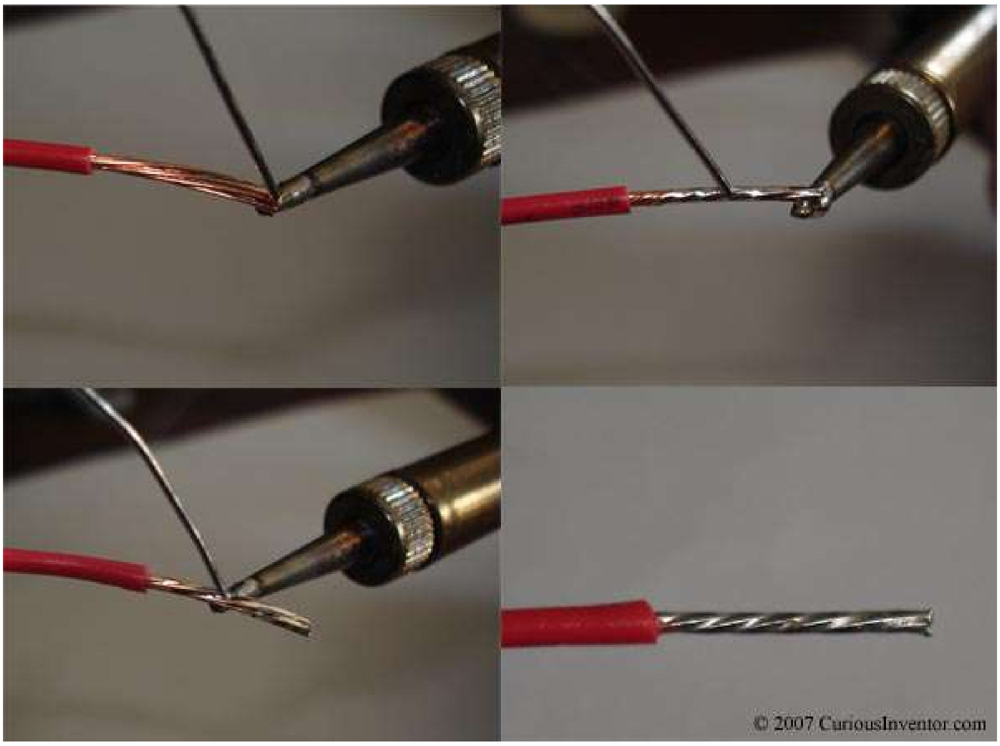

Got a lot of tinning to do? For under $300 you can get a vat of molten solder to speed up the process. The only catch is that there's no flux in the vat, so you need to dip the end of the wires in liquid flux before dunking them in the molten solder. Only dip the very end of the wire in the flux; capillary action will draw it further up the strands. If you dipped more than just the end into the flux, the capillary action would draw flux beyond the insulation where it can't be cleaned, and would eventually corrode the wire.

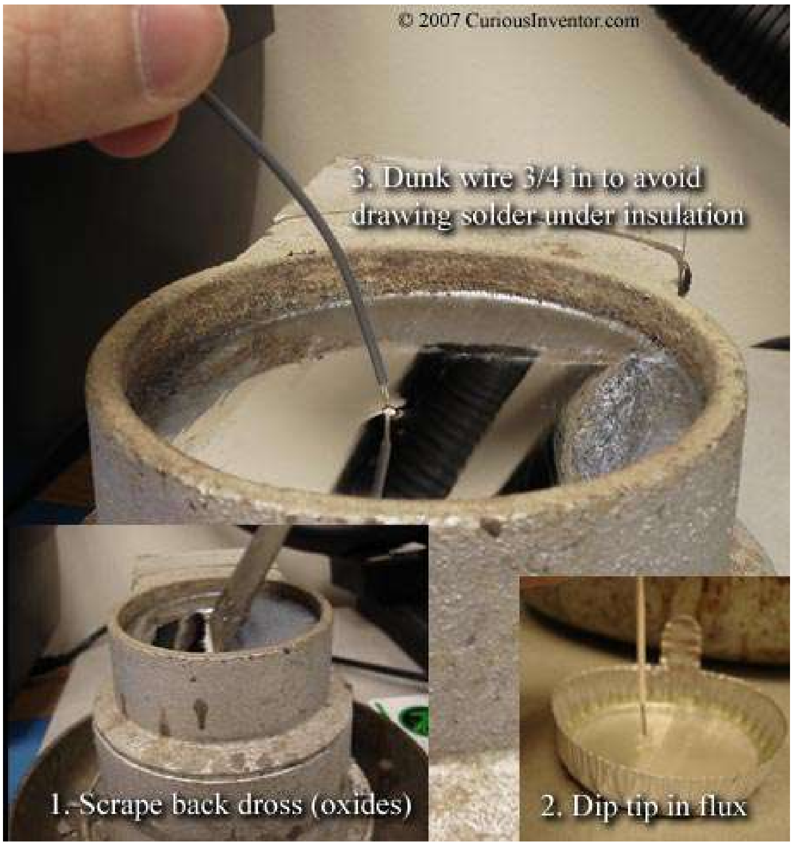
© 2007 CuriousInventor.com
3. Dunk wire 3/4 in to avoid drawing solder under insulation
1. Scrape back dross (oxides)
2. Dip tip in flux

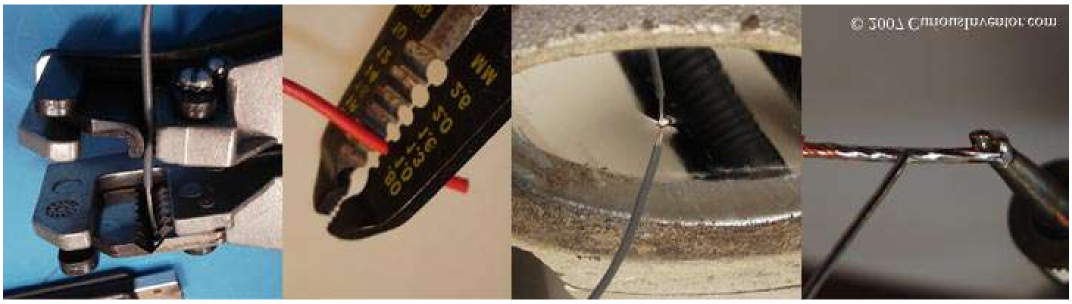

- **Insert, clinch and trim components:** First, make bends *before* inserting the components. Avoid stressing the connection between lead and component by bracing the lead with pliers while bending. Pliers with serrated tips aren't used in high-reliability production because the grooves can create nicks in the leads that eventually cause a break after a lot of vibration and thermal changes. Round nose pliers make it easy to make any sized radius.

  Unless the component has a metal casing or needs clearance for air flow to keep cool, insert it until it's flush with the board. This doesn't apply to some transistors, and also capacitors that have plastic coverings that need to be kept out of the solder joint. Clinch or bend out the leads so the component is held in place during soldering, and finally trim the leads to about the radius of the pad. Trim the leads before soldering since doing so afterwards can shock and crack the joint. Wearing safety glasses for this process is not at all excessive--those leads can get you. Everything else about proper component installation: NASA guide.

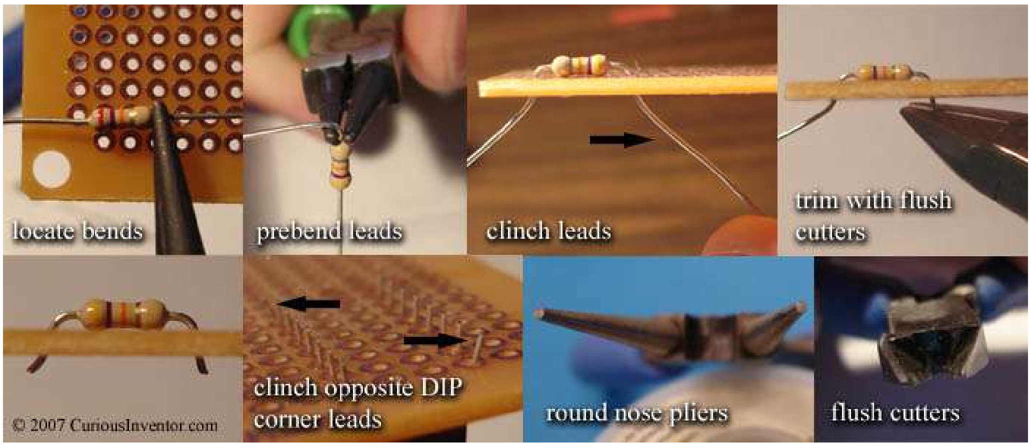
locate bends
prebend leads
clinch leads
trim with flush cutters
© 2007 CuriousInventor.com
clinch opposite DIP corner leads
round nose pliers
flush cutters

- **Add heat sink:** Some semi-conductors (some transistors and diodes) are especially heat sensitive. This clip acts a heat shunt to keep the transistor protected.

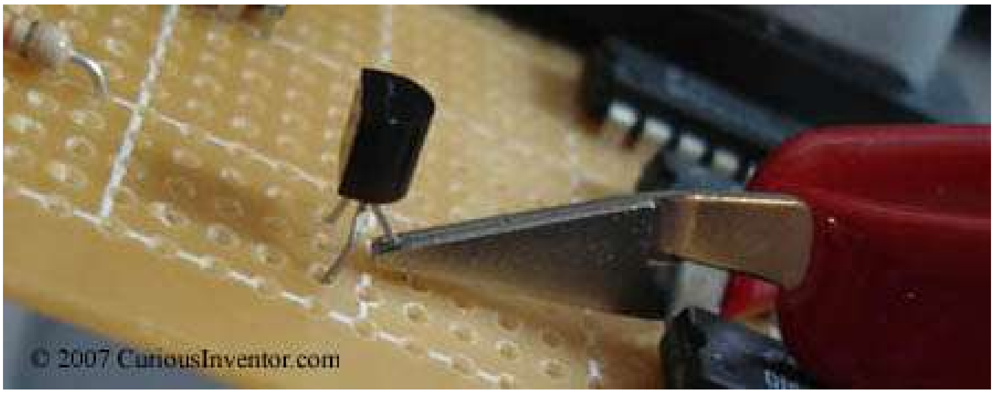
© 2007 CuriousInventor.com

Clean and tin the Tip

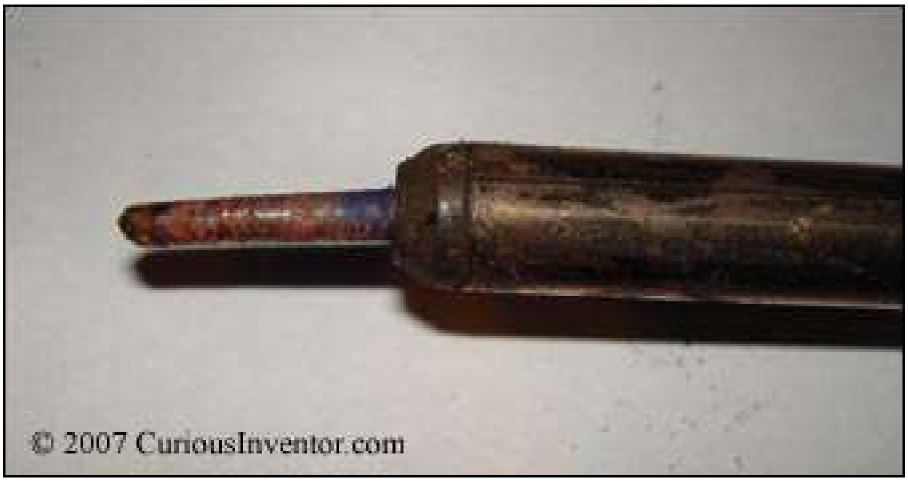
© 2007 CuriousInventor.com

· **Regular cleaning = easier soldering and longer tip life:**The iron tip's ability to transfer heat is drastically reduced when it gets covered in oxides and burnt flux residues. Not only does heat not transfer as well through this debris, but the contaminants also prevent solder from wetting or sticking to the tip. Most heat transfer actually goes through a fluid solder "heat bridge" that lies between the iron tip and components, so an iron tip that repels solder will be very ineffective.

The longer oxides and charcoaled flux residues remain on the tip, the harder they become to remove, so it's a good idea to clean the tip every time you pick up the iron.

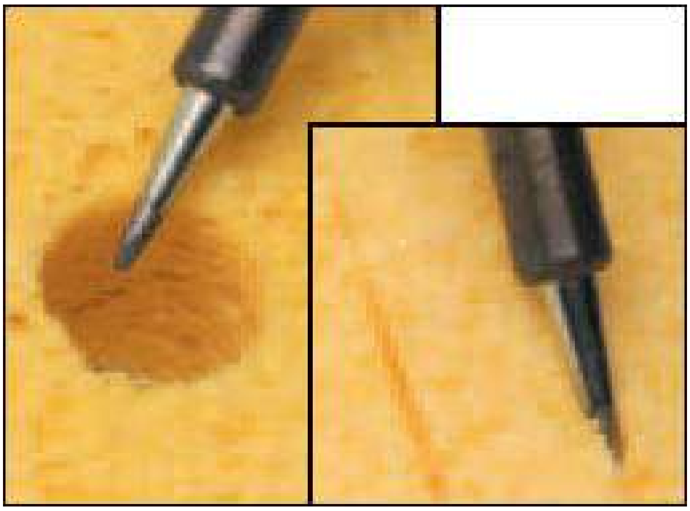

Wiping the iron on an edge of a hole cut into a sponge can help to remove oxides easier, and also allows waste to fall away. A dry cleaner can also be used. It consists of soft metal shavings that are coated with flux. You clean by thrusting the iron into the shaving a few times. By avoiding the thermal shock of touching a damp sponge, these cleaners help to increase tip life, and in our opinion, do a better, faster job.

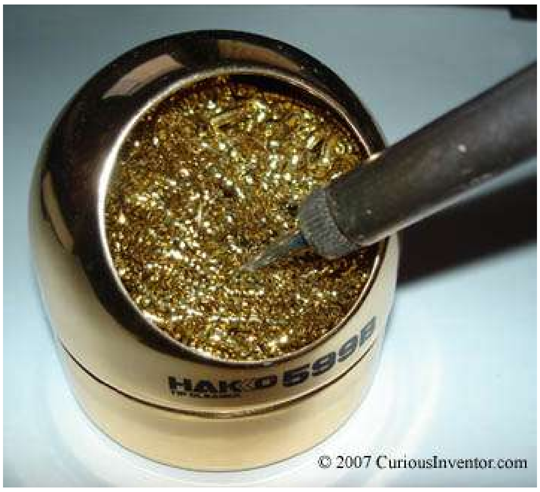
HAKKO 599B
© 2007 CuriousInventor.com

Usually touching the tip with rosin-cored solder will supply enough flux so that oxides can be removed with a damp sponge. If this isn't sufficient, you can purchase "tip tinners and cleaners" that are a mixture of solder paste and flux. The flux is oftentimes stronger (more activated) to help remove oxides.

64-020
RadioShack
Tip Tinner and Cleaner
WARNING: KEEP OUT OF REACH OF CHILDREN. Read carefully other warnings on back of card.
0.5 oz. (14.2gm)
© 2007 CuriousInventor.com

Finally, when that doesn't work, special polishing bars to can be used to salvage extremely bad tips. Another last resort is to gently rub the oxides off with an emery cloth or soft steel brush. Cover the tip immediately with solder after cleaning to prevent further oxidation. Never file the tip to clean it or form a different shape. The tips are mostly copper with a protective iron plating, and once that plating is pierced, the tip will die quickly. Copper is used because it's an excellent heat conductor, but if exposed to solder, it will quickly dissolve into the solder.

Weller
WPB1
POLISHING BAR FOR SOLDERING TIPS

© 2007 CuriousInventor.com

Tin the tip: Add a small amount of solder back onto the tip. This helps to protect the newly cleaned and exposed tip, and also helps to transfer heat to components.

Heat and Solder the Joint

- **Heat the joint:** Place the iron tip so that it touches both the component lead and pad--the goal is to get as much surface area contact between the iron tip and joint as possible. Almost no heat will travel through the point.

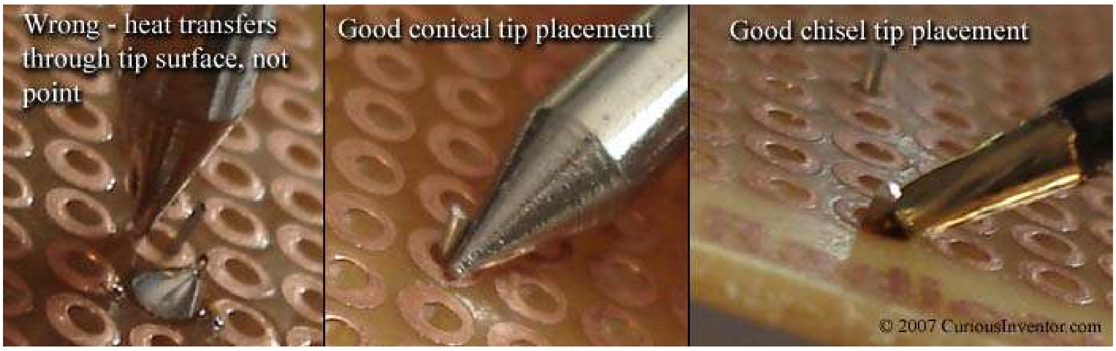
Wrong - heat transfers through tip surface, not point
Good conical tip placement
Good chisel tip placement
© 2007 CuriousInventor.com

- **Make heat bridge:** Add a small amount of solder between the tip and the work--heat transfers much faster through the liquid solder than dry surface contact. This is why a tip that won't "wet" is so difficult to use. Pressing hard should not be necessary. This step may not be necessary if there's enough solder already on the tip from tinning it after cleaning.

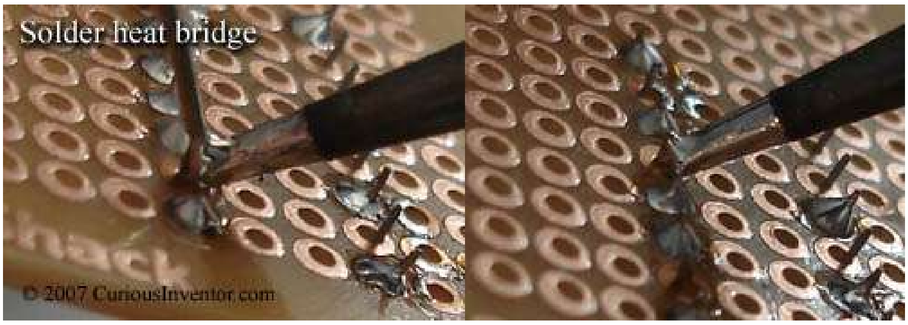
Solder heat bridge
© 2007 CuriousInventor.com

- **Apply solder to opposite side:** Apply solder to the parts, *not* the iron. By doing this, you ensure the parts are hot enough for the solder to "wet" and bond with them. Also, solder will run towards the heat source, so applying solder opposite from the iron helps to spread it out and cover the joint.

  For larger joints, rather than dumping in all the solder quickly, continuously pulse in small amounts to keep a fresh supply of active flux available.

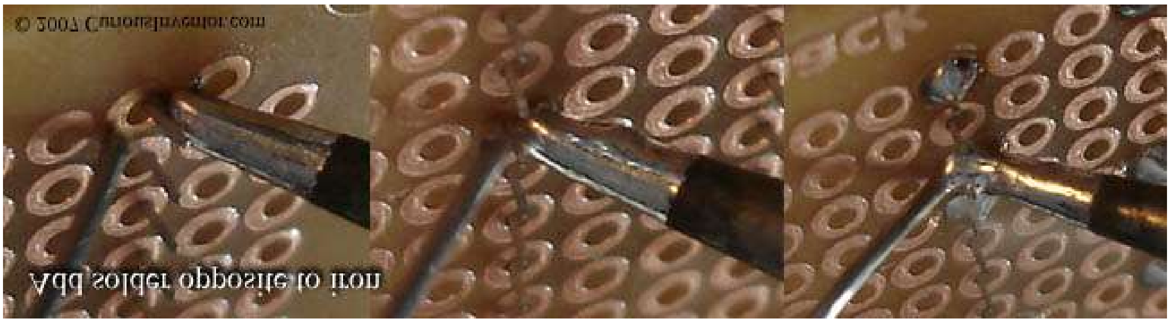
© 2007 CuriousInventor.com
Add solder opposite to iron

- **Time:** The joint should take about 2-5 seconds total time for standard 60/40, 63/37 lead based solder and a non- no-clean flux, and up to 7 seconds for lead-free solder. Lead-free solder just takes longer to "wet" the metal.

  Wetting is how easily and quickly solder spreads out over a surface. A water droplet on a freshly waxed car shows poor wetting, as does solder on a heavily oxidized soldering iron tip. It basically comes down to how attracted the liquid molecules are to each other verses the surface (see surface tension). In industry, tests are done to determine the "solderability" of materials by measuring the time it takes solder to spread out over a surface, or measuring how much force a pot of liquid solder will pull down on a component partially submerged.

Good pictures and description of wetting and surface tension. Contains a video of a razor blade floating on the surface of water until a drop of soap is added.

A brochure for a solderabiliy testing station. Scroll down a few pages to see some great pictures of this machine holding one lead of a surface mount chip in a drop of solder.

Another related wiki: wiki/Surface_energy

In general, the goal is to make the joint as quickly as possible. Longer times can char and damage the board, lift pads, overheat components, burn off and polymerize flux (making it harder to remove), and finally lead to a more brittle joint. Solder doesn't just freeze on a joint, tin in the solder dissolves and chemically reacts with copper in the connection to form a new bonding material, called an "intermetallic layer". While this layer is what makes an excellent thermal and electrical bond, it is also extremely brittle; a doubling of its thickness reduces joint tensile strength by half (ref 1). Since this layer grows faster with higher temperatures, joints should be made using the coolest temperature and shortest soldering time possible. This layer is also why re-heating joints has been shown to weaken them. Having said all this, I have to admit that I don't know just how long is *too* long for projects that don't need to operate for 30 years with 100% reliability. After 10 seconds there's a good chance the flux has been used up.

- **Remove solder, then iron:** Pull the iron out fairly quickly to avoid leaving a solder spike.

© 2007 CuriousInventor.com

- **Good and bad joint gallery:** The solder should smoothly ramp to meet surfaces and be shiny in appearance if it's lead-based. Lead-free solder will have a duller and grainier surface, but will still be a good joint as long as there are no signs of non-wetting. The important thing to look for is any solder that looks like it didn't cling to a surface, or is just sitting on top or next to a surface.

Gallery of joints:

A NASA gallery of every possible joint / board defect you could ever imagine. Here's another great gallery of defects.

Great comparison pictures between lead-based and lead-free joints. p. 34

Good joints:

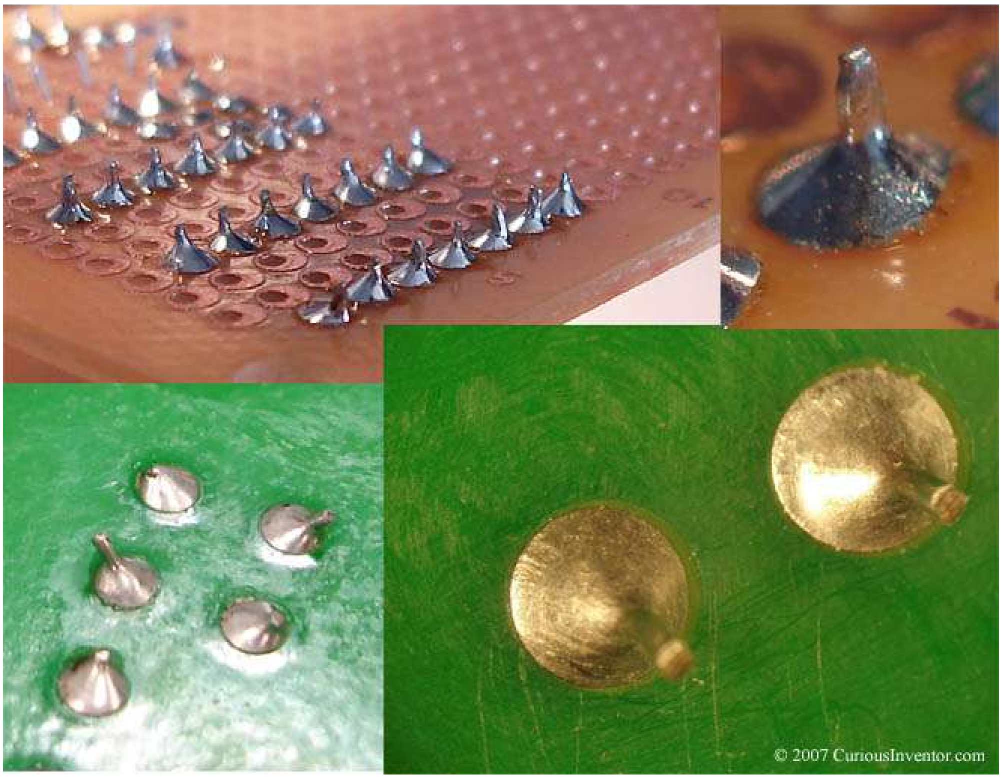

© 2007 CuriousInventor.com

Great joints:

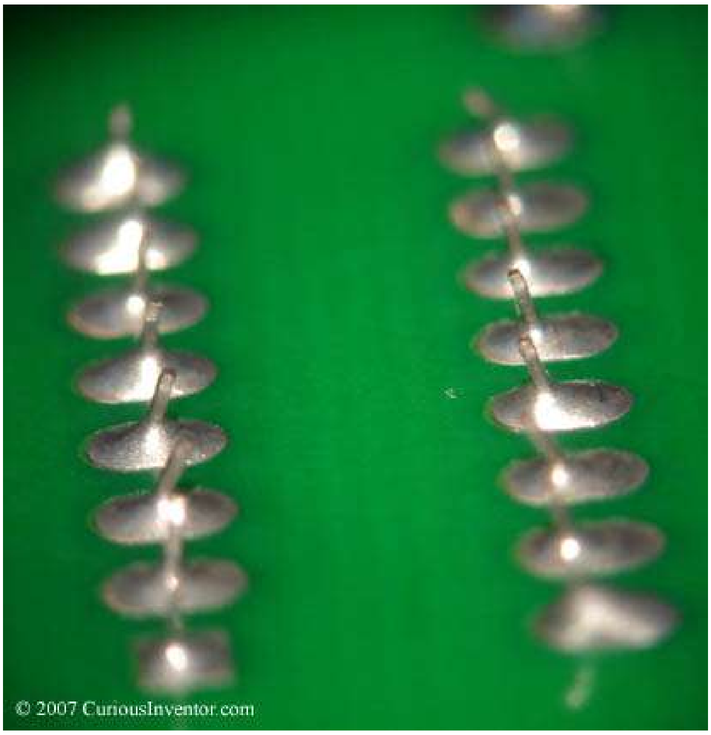

© 2007 CuriousInventor.com

What's the right amount? A large amount of solder is not needed, just enough to cover the pad and lead without any gaps (actually, only 270° is required by IPC J-STD-001). The measure for too much solder is whether or not you can see the outline of the lead in the solder. This is important because you need to be able to see whether the solder adhered or "clung" to the lead (indicating a true bond), and didn't just freeze around it. This applies to all types of joints: tinning wires, soldering to connectors, surface mount components... the solder should never completely hide the underlying wires or leads.

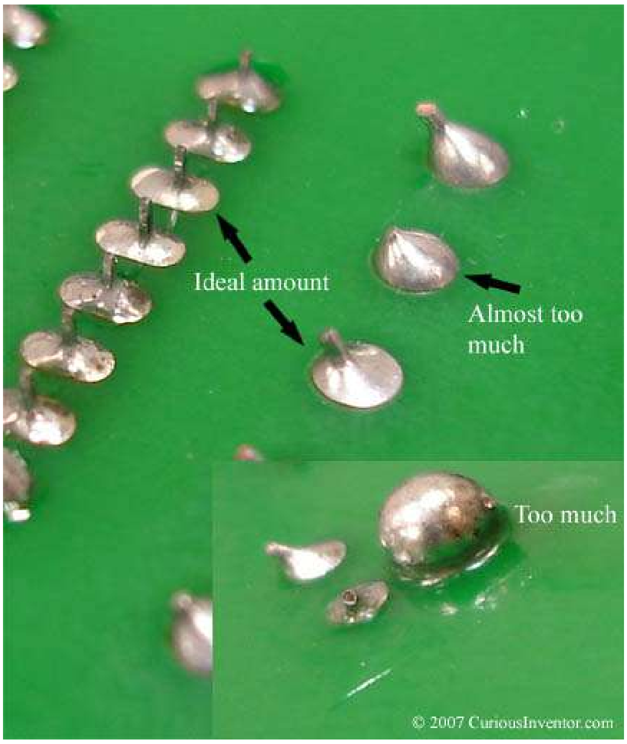
Ideal amount
Almost too much
Too much
© 2007 CuriousInventor.com

Disturbed joint (bad): If the component moves during solidification, the internal structure of the solder will have fractures, leading to a high resistance or unreliable electrical connection, as well as a fragile mechanical one. The solder also appears dull and grainy--typical signs of a "cold" joint that doesn't actually bond with the underlying surfaces.

"Cold" joints are often formed when the underlying pad or lead didn't get hot enough for the solder to wet it.

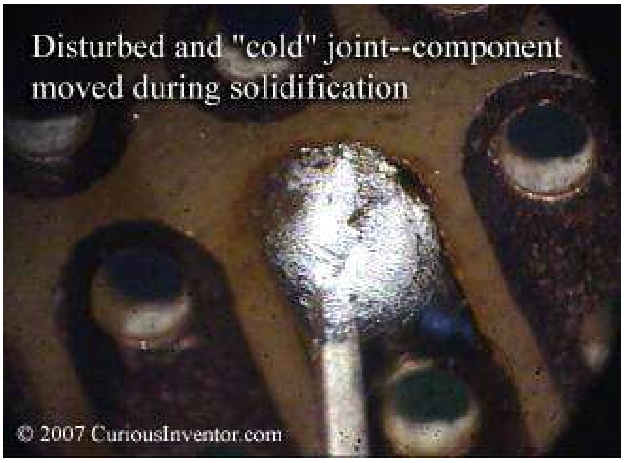
Disturbed and "cold" joint--component moved during solidification
© 2007 CuriousInventor.com

Awful joints:

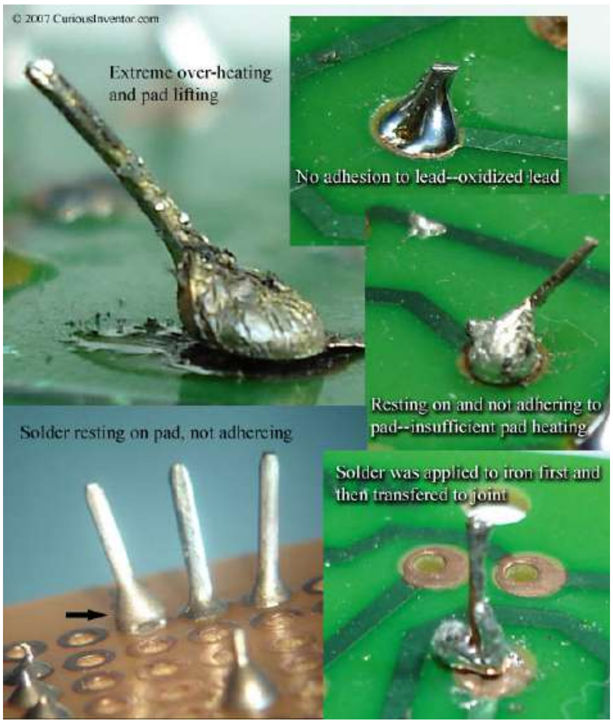
© 2007 CuriousInventor.com
Extreme over-heating and pad lifting
No adhesion to lead--oxidized lead
Resting on and not adhering to pad--insufficient pad heating
Solder resting on pad, not adhereing
Solder was applied to iron first and then transfered to joint

Clean Up

Most manufacturers will clean off residues from any flux that isn't labeled "no-clean," despite flux datasheets like Kester's that say even some of the more active fluxes do not need to be cleaned. For short life-span hobby projects, it probably doesn't matter unless you're using a solder/flux labeled "organic" or "water-soluble"--these fluxes leave behind very aggressive acids that will quickly eat away circuits. Cleaning may be necessary if you're applying a protective coating that won't adhere to flux residues. Finally, some rosin residues are tacky and may attract dust that can short a circuit.

The fact that a flux is made from rosin doesn't tell you much about how strong it is or whether it should be cleaned. What matters is how concentrated the mix is and how much acidic (halides) activators were added. RadioShack® doesn't supply any information on the flux in their standard rosin-cored solder, but it's probably weak enough that the residues do not need to be cleaned off.

Isopropyl alcohol works decently on rosin-based residues, but clean shortly after soldering because the residues quickly harden. Use water for water-soluble fluxes. This pump containing bottle dispenses a little alcohol when you push down on the top with a brush, and keeps the rest from evaporating. If you are going to clean, make sure you wipe up the remnants with a lint-free cloth--don't just spread them around the board with a brush and alcohol.

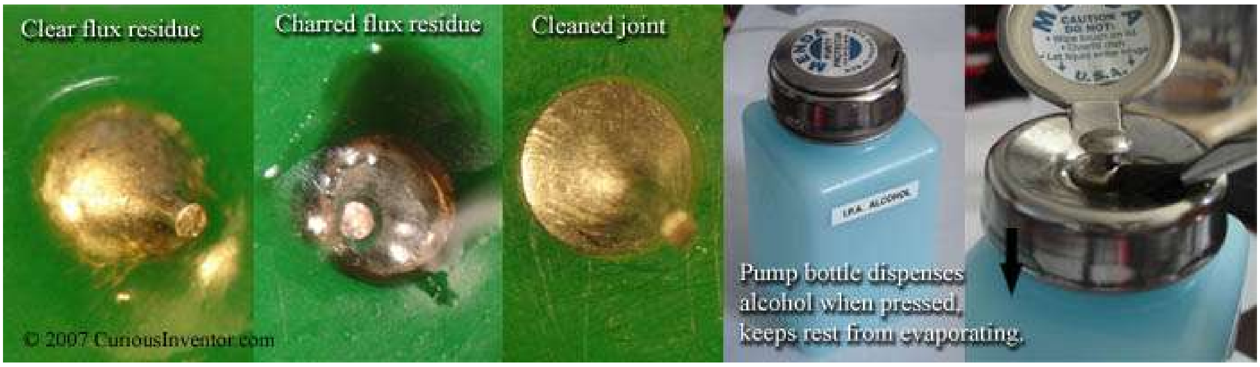
Clear flux residue
Charred flux residue
Cleaned joint
© 2007 CuriousInventor.com
I.P.A. ALCOHOL
Pump bottle dispenses alcohol when pressed, keeps rest from evaporating.

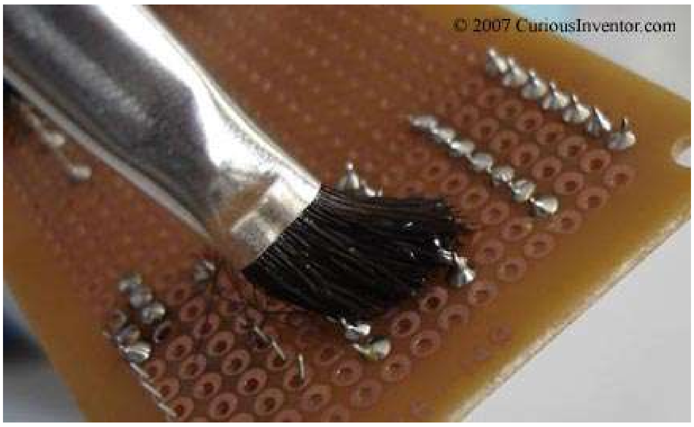
© 2007 CuriousInventor.com

Desoldering

The three most common inexpensive ways to remove solder include a "solder sucker", solder wick, and an iron with an attached desoldering bulb.

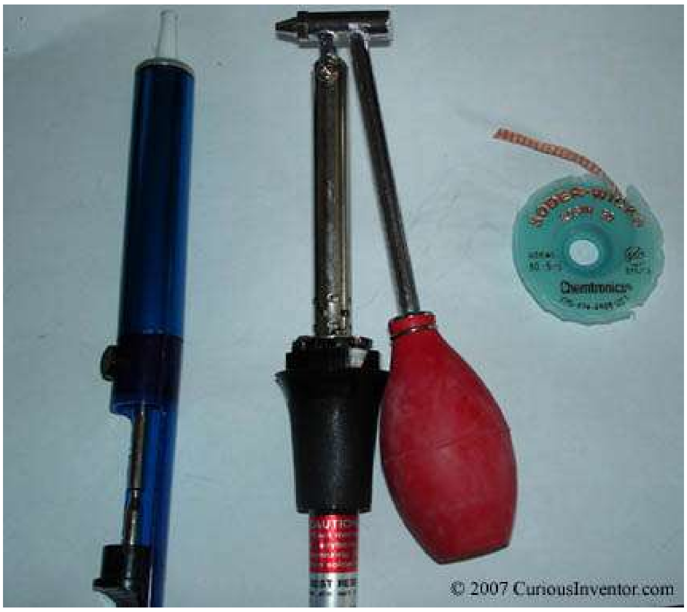

- **Solder Sucker:** To use, press in the plunger, heat up the joint, and as quickly as possible, pull out the iron, place the sucker's nozzle over the joint and press the release button. It may take a few iterations to clean out a joint enough to wiggle free a component. On plated through holes it may actually be easier to add *more* solder to the joint before sucking, since the sucker won't be able to create a vacuum with too little solder remaining.

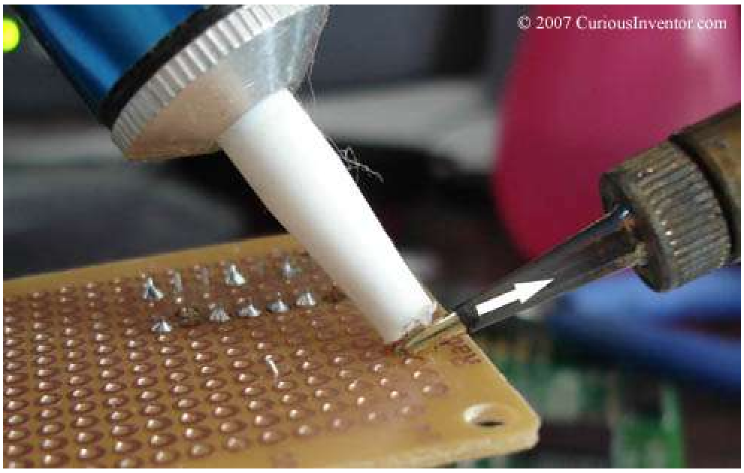

- **Desoldering Iron:** This works just like the solder sucker, but is much easier to use because there's no race to suck out the solder before it freezes, as with the "solder sucker".

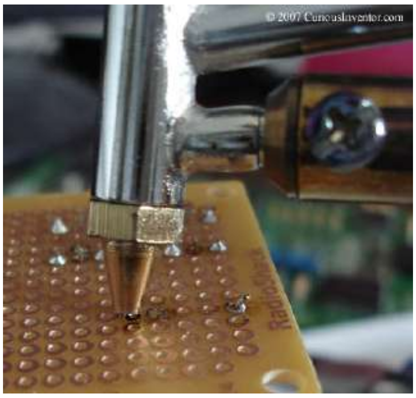
© 2007 CuriousInventor.com

- **Solder Wick (desoldering braid):** The above two tools will often not be able to completely remove all the solder. Many people use the sucker devices above to remove most of the solder, and then get the rest with solder wick, which absorbs solder through capillary action. It's a fine weave of copper wires that are coated in flux. Place it over a joint and then heat from the opposite side with a soldering iron. It may help to have a bit of solder already on the iron's tip to speed heat transfer.

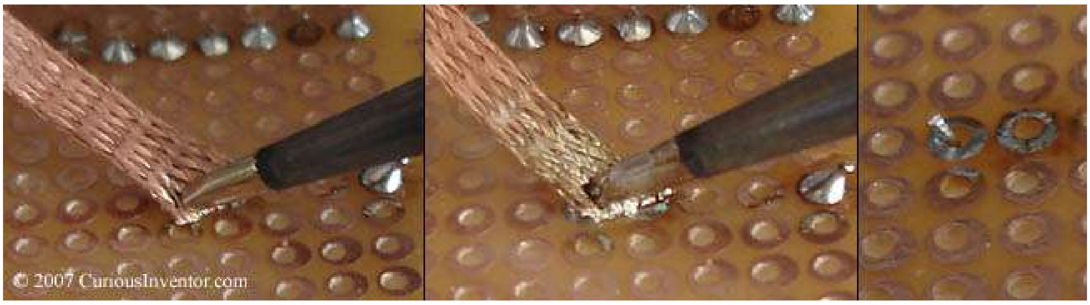
© 2007 CuriousInventor.com

Surface mount desoldering: Surface mount chips are especially hard to desolder because it's very difficult to completely remove all the solder pin by pin, and avoid overheating the board and lifting a pad. Professional shops use expensive hot air guns or special tips (shown below) to heat all the joints at once.

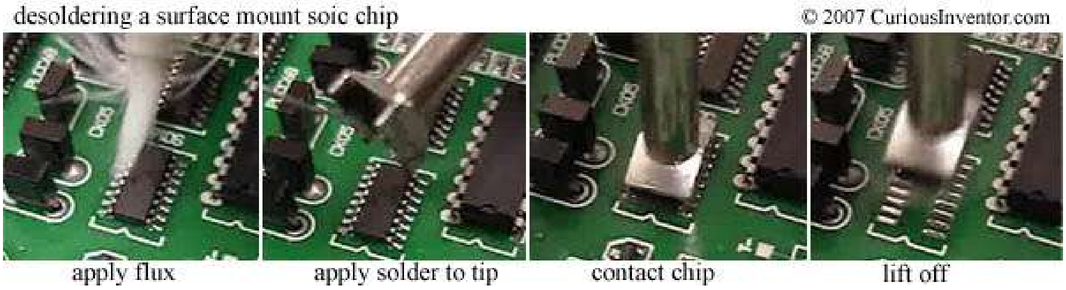
desoldering a surface mount soic chip
© 2007 CuriousInventor.com
apply flux
apply solder to tip
contact chip
lift off

There are fortunately a few cheap ways to desolder surface mount chips.

- ChipQuik provides an interesting solder that when melted over existing joints produces a new low-melting point alloy (under 200 °F) with a much longer solidification time. The longer solidification time enables you to melt all the joints at once and then flick off the chip.
- There are many guides for making a DIY hot air gun with the RadioShack® desoldering iron and a fish tank air pump. Engadget has one of the better how-to guides. Here and here are some more plus the supposed original.

The Ersa soldering primer

Soldering made easy

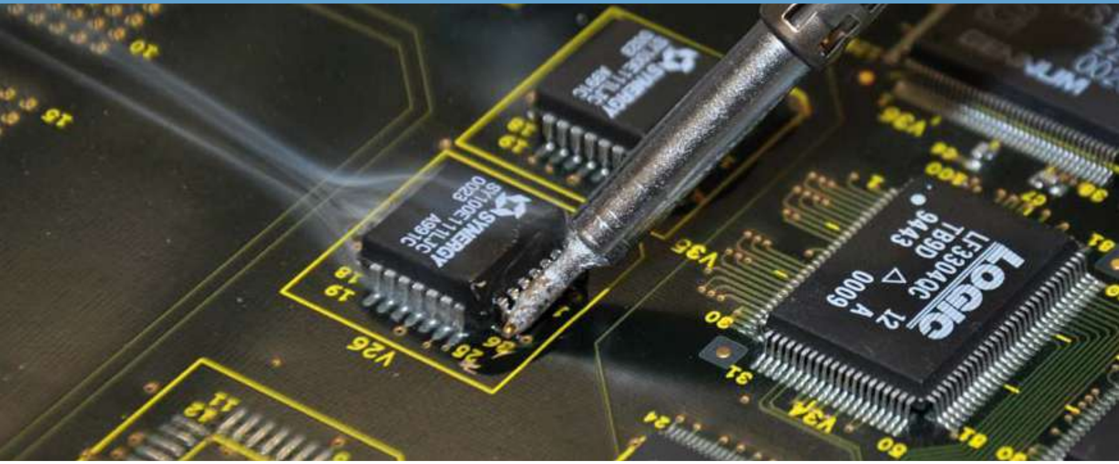

kurtz ersa

Soldering – a never-ending story

for more than 5,000 years already

Man had scarcely learned how to use metals for his purposes when the desire to join them arose in him. Many of the pieces of jewelry, tools and weapons we know from the Bronze Age owe their usefulness and beauty to the art of soldering.

ersa
RDS 80
Power
350 °C
150
300
400

Today, it is difficult to say who first discovered how to "glue" metals. One thing is certain - the goldsmiths of ancient Egypt knew how to join gold and silver already more than 5,000 years ago. Their colleagues in Troy were also master craftsmen long before the ancient Teutons could even dream of such handicraft. Soldering really "came of age" when tin was discovered as a joining metal. And that was, after all, already 4,000 years ago!

From then on, soldering technology was on its way. It first spread around the Mediterranean. The Cretans showed it to the Etruscans who then taught it to the Romans, Tunisians, Spaniards, followed by many others, including the less developed cultures of the time – the Swiss, Bohemians, Hungarians, Teutons and Scandinavians. From culture to culture, from generation to generation, the craft of soldering was continuously improved and refined.

The ancient Romans already laid down and soldered 400 km of leaden water pipes, conjured up stoves and bathtubs from bronze sheets, not to mention the excellent craftsmanship of their armorers and goldsmiths. Apart from craftsmanship in soldering, our understanding of the science of soldering has grown and has been refined over the centuries.

Today, soft soldering in the electronics industry has developed into a full-fledged production technology, encompassing the fields of mechanics, chemistry, physics and metallurgy. Ernst Sachs, the founder of Ersa (the company name consists of the initial letters of his first and last name) contributed to this development. In 1921 he developed the first electrically operated soldering iron for the industry that was manufactured in series. Since that time, Ersa has committed itself to the further development and perfection of soldering technology with great passion and extending its full power of innovation.

Today, Ersa stands for the most comprehensive product range in the soft soldering technology worldwide and for more than 90 years of industry experience and innovation, know-how and highest product quality.

The Ersa soldering iron product range covers ultra-fine soldering tips, classical soldering irons powered from the standard power net and special soldering tools up to the 550 W hammer soldering iron. Ersa's electronically temperature-controlled soldering stations represent the industry standard, as does the extensive range of rework and

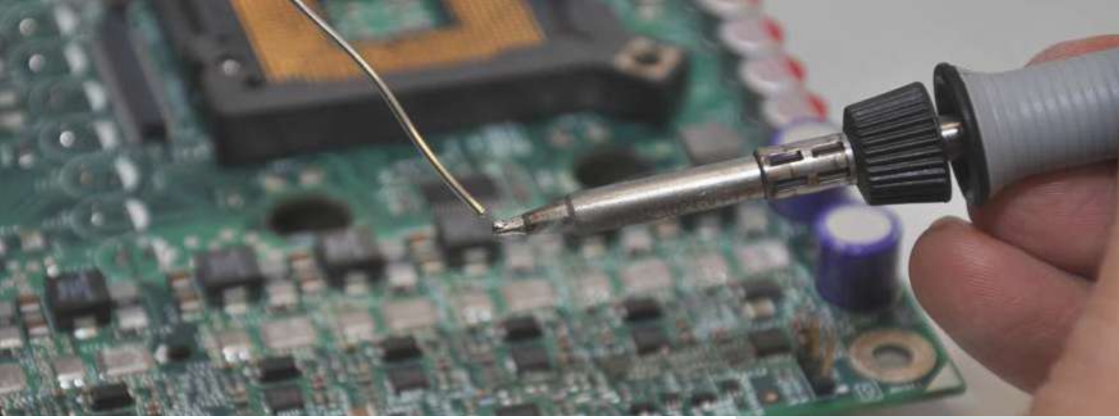

inspection systems, wave-, reflow- and selective soldering systems. The line of Ersa screen printers complements the product range.

Ersa's quality soldering tools are used in the hobby area, such as, for example, in model-making or tiffany soldering, in the craft sector, in laboratories and in industrial electronics manufacturing.

New challenges for the soldering technology were raised by the ban on certain hazardous substances (RoHS) in 2006. Since 1 July 2006 electric and electronic equipment may not contain any lead, mercury, hexavalent chromium, PBB (Polybrominated biphenyls) or PBDE (Polybrominated diphenyl ethers).

In many cases this restriction entailed having to depart from the use of the well-known soft solders based on tin and lead.

By publication of this small primer, Ersa would like to facilitate your entry into the "World of Soldering", and raise your enthusiasm for a modern technology with a long history.

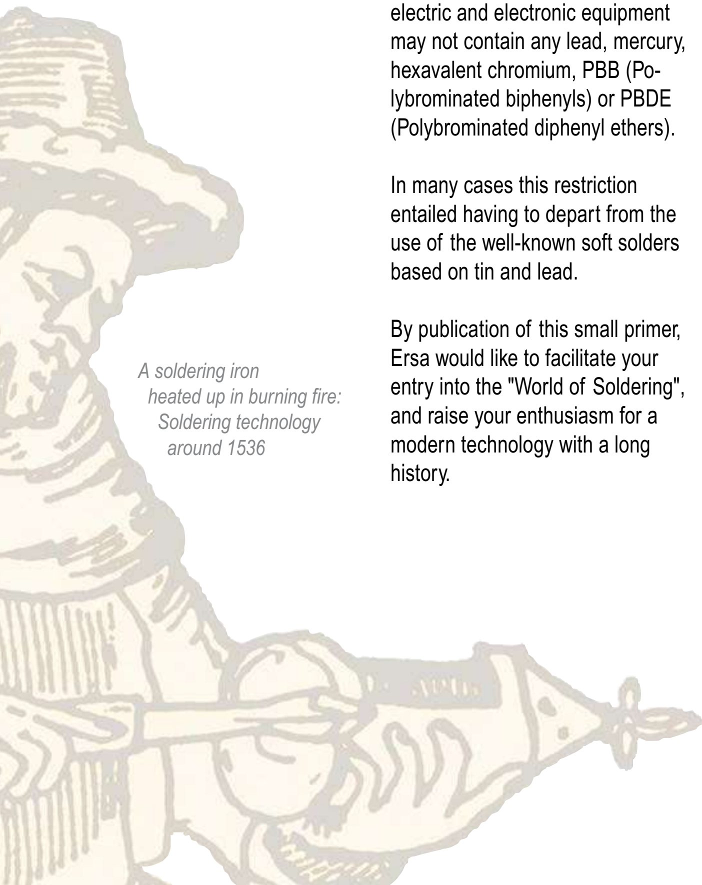

A soldering iron heated up in burning fire: Soldering technology around 1536

Content

	Page
The history of soldering	2
The art of soldering	5
What is needed to solder?	6
Health	9
THT soldering	10
SMD soldering	13
BGA soldering / rework	16
Miniature and microsoldering irons	18
Universal soldering irons	19
Standard, hammer and high-speed soldering irons	20
Power and gas soldering irons	21
Soldering stations	22
Multifunctional soldering and desoldering stations	23
Hybrid rework station	26
Solder fume extractions	27
Tiffany soldering	28
Accessories	30
Seminars, references	31

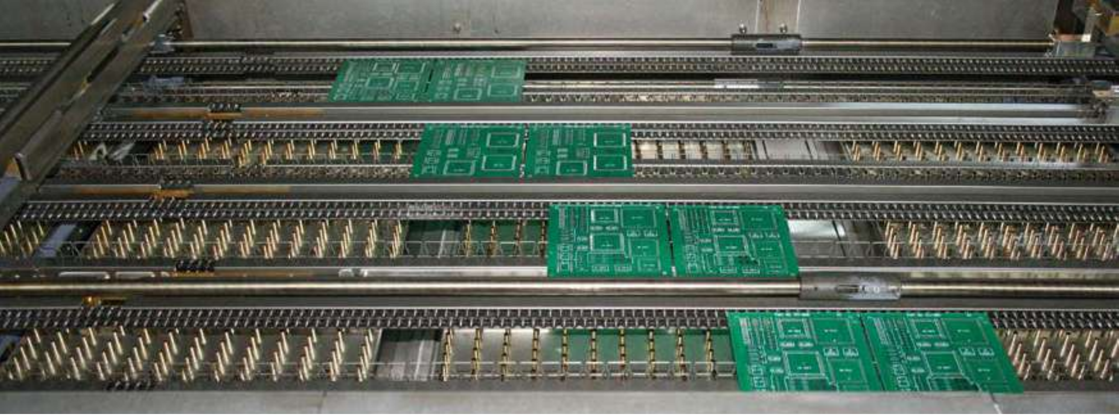

Solder joints hold the world together

Connecting pacemakers and solar parks

To imagine today's world without soft soldering is not possible. It is the means to fabricate safely functioning, electrically conducting connections. Regardless of whether we talk about power technology, drive technology, telecommunications, automation or electronic controls – in all those fields soldered connections have a decisive share that everything functions in a way as has been foreseen and planned by the developers and visionaries of the products. Today, soft soldering is such a common place occurrence, that no one wastes any further thought on it. We take the daily use of our computers, mobile phones and play stations for granted, the modern comforts provided for by electronics found in modern automobiles is expected as a matter of course, and we fly – privately or on business – to the farthest spots in the world.

Consequential damage because of the failure of a solder joint in an iPod is relatively limited. It is a different matter altogether, though, if the electronics in an airplane full of vacationers, in a space shuttle, or in an implanted pacemaker fails. Such failures are immanently life-threatening. But not to worry – the highest quality demands apply for those applications, and rightfully so!

Aside from soldering in consumer electronic products, there are numerous other applications such as, to name but a few, alternative power generation with wind turbines or solar parks, R&D departments and in work performed by craftsmen such as electricians and plumbers. Let us not forget the many part-time and hobby users for whom there are no limits curtailing their phantasies and artistic freedom when handling a soldering iron and solder.

The art of soldering

What's behind it?

In a soldering process two metal parts are joined by means of a molten metallic bonding agent (solder), whereby the melting point of the bonding agent is always lower than that of the metal parts to be joined. If the melting point is below 450 °C, then it is a soft soldering process, if it is above, it is called hard soldering or brazing. Welding, on the other hand, is the process where two metals will be heated up to their melting point, at which time they will, together with a filler material, form a pool of molten material causing coalescence. In soft soldering, the seams between the metals to be joined will be filled with a tin alloy. It is important that the alloy does not simply stick to the foreign metals' surface after cooling but unites with the metal. For this purpose, a small quantity of the foreign metals must dissolve and unite with the tin alloy forming a mix of crystals – the so called diffusion zone. That is the task of the tin, whereas the alloy's other components are responsible for the solder liquefaction and the joint's mechanical stability. A solder joint consists of the following layers:

① Base metal
② Diffusion zone
③ Solidified solder
② Diffusion zone
① Base metal

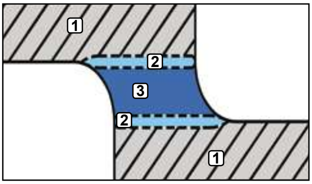

Cross section of a solder joint

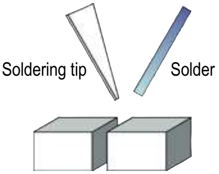

The prepared parts and the solder are heated

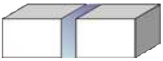

The liquid solder flows into the gap and fills it

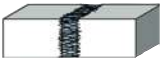

The solidified solder joins the parts together

To achieve the highest mechanical stability, i.e. to assure the durability of the solder joint, the diffusion zone may neither be too thick nor too thin. Its ideal thickness is 0.5 µm. The formation of the diffusion zone depends on the temperature, the solder time and the alloy used. If the diffusion zone is too thick, the solder joint will be brittle and porous, whereas the formation of a zone which is too thin indicates that an insufficient connection or no mechanical connection at all has been formed.

Reaction of the flux during manual soldering on a PCB using flux cored solder wire

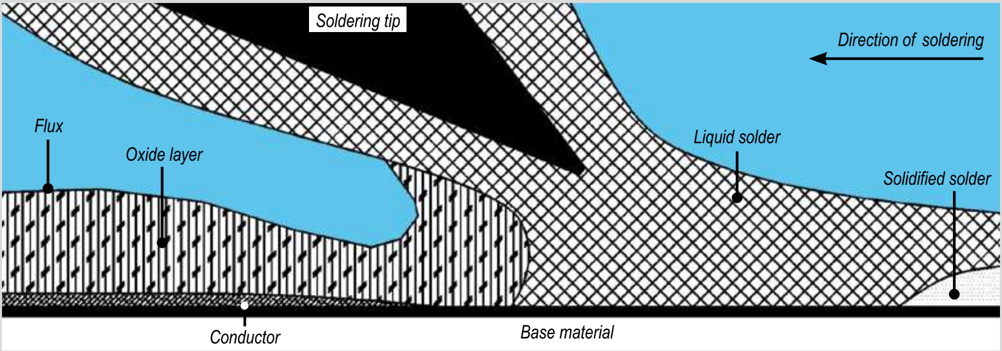

What do you need for soldering?

The basics – the 5 essential factors

1. The soldering iron

to provide the heat required

Heat is required to melt the solder. This is the soldering iron's job.

Temperatures of 200 °C – 450 °C are required depending on solder joint and solder alloy. In the field of electronics, the usual temperature lies between 250 °C and 375 °C.

In order to have the proper temperature for any soldering application, the soldering iron's thermal output and an efficient heat transfer to the solder joint to be made is decisive. One either selects a soldering iron that performs within the temperature range required, or a soldering station with temperature control is used. Temperature-controlled soldering stations enable the user to work on different applications without loss of solder joint quality, because of the precise control of the preset temperature at the soldering iron tip.

Both the soldering station's registration of the actual tip temperature should be highly precise, and the heating element should be powerful and recover quickly in order to avoid over-heated or cold solder joints.

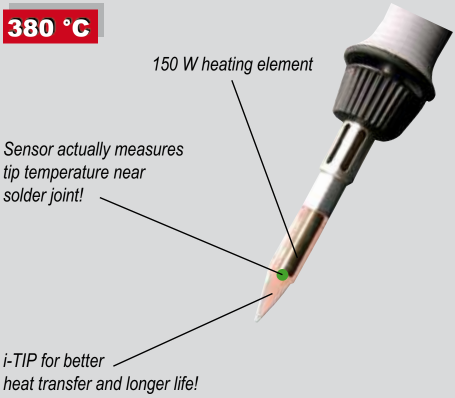

380 °C
150 W heating element
Sensor actually measures tip temperature near solder joint!
i-TIP for better heat transfer and longer life!

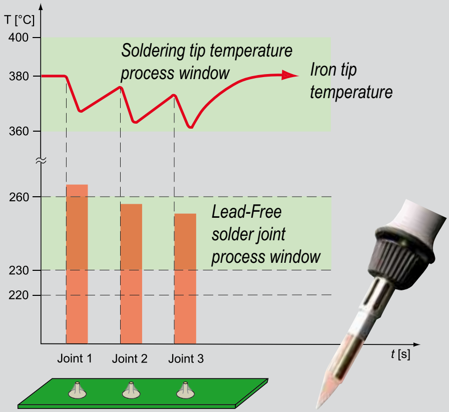

Temperature curve of controlled tip temperature when soldering multiple joints

2. The soldering tip

to transfer the heat from the heating element to the tip

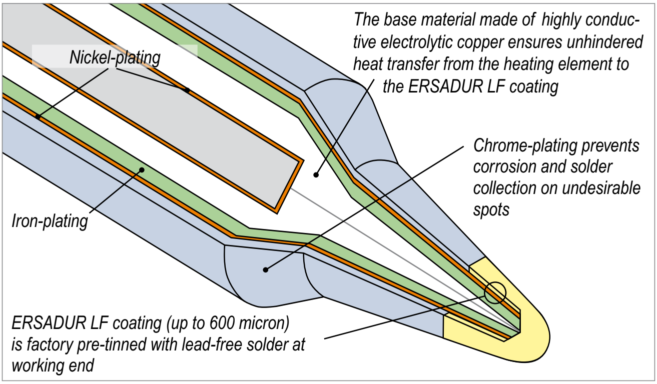

The base material made of highly conductive electrolytic copper ensures unhindered heat transfer from the heating element to the ERSADUR LF coating

Chrome-plating prevents corrosion and solder collection on undesirable spots

ERSADUR LF coating (up to 600 micron) is factory pre-tinned with lead-free solder at working end

Cross-section of an ERSADUR soldering tip, non-scale representation

The soldering tip is the "heart" of the soldering iron and responsible for the heat transfer from the heating element via the solder to the solder joint. Depending on the soldering iron and the application, different types of tips are available. Prerequisites for good solder joints are the correct soldering iron tip shape, perfect heat transfer, a good condition of the tip and a reliable performance over time. In addition, the soldering tip has to convey also the necessary amount of sensitivity back to the operator.

How to take solder joint quality to its peak – ERSADUR long-life soldering tips

Traditional soldering tips are made of copper which conducts heat well and is inexpensive, but the tip oxidizes heavily when heated and releases copper particles into the solder until it has been "corroded" entirely. To maintain the tip in operational shape, it requires intensive care. Today, only coated soldering tips are used in electronics production – the largest field of application for soft soldering.

ERSADUR long-life tips have been conceived for continuous operation and for high-quality results. They are galvanically plated with an iron coating and protected against corrosion and oxidation by an additional chrome layer in a very special manufacturing process, developed and used exclusively by Ersa. And their perfect thermal conductivity protects the heating element from overheating and premature wear. Ersa offers a comprehensive range of soldering tips for the diverse requirements.

Proper tip care increases tip lifetime considerably:

- Never clean the long-life tip before putting the soldering iron into its holder, since the solder remaining on the tip prevents oxidation of the solder track.
- Always keep the long-life soldering tip covered with solder, as otherwise it becomes passive and will no longer wet properly.

Passive tips can be reactivated by the application of the lead- and halogen-free Ersa TIP REACTIVATOR. All that is needed is to wipe the hot tip on the surface of the regeneration compound. Furthermore, the hot tips should regularly be cleaned with a moist viscose sponge before soldering. Alternatively the tips can be dry cleaned using the Ersa "dry sponge", a sponge made of special metal wool. Dry cleaning has proven to be advantageous in lead-free soldering. The soldering tips are not cooled abruptly, and contaminated tips resulting from dirty sponges are avoided. Due to the slightly abrasive properties of the special wire mesh, passive layers that accumulated on the tip can easily be removed. Tip life is thus increased considerably in lead-free hand soldering.

Ersa TIP REACTIVATOR. A comprehensive range of accessories is available at ***www.ersa.com***

What do you need for soldering?

The basics – the 5 essential factors

3. The solder

for the connection

Metallic bonding agents (solders), mostly in the shape of a wire or a bar, are available as diverse alloys.

Soft solders consisted mostly of a mix of tin (Sn) and lead (Pb). Since the implementation of the RoHS directives on 01 July 2006 the use of solders containing lead is prohibited. Lead-free solders are usually alloys containing silver (Ag) and copper (Cu).

The alloy's composition determines melting temperature and physical properties of the joint. Criteria for the choice of an alloy are: production process, specification of the product, field of application, cost of the alloy.

Prerequisites for good soldering results are:

- Suitable soldering iron
- Clean surfaces of where to solder
- Sound soldering tip
- Suitable fluxes
- Suitable solder alloy
- Correct soldering time

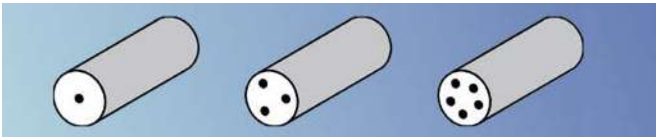

Solder wire with one or with multiple flux cores

4. The flux

for the ability to wet

Fluxes are used to attain the best possible bonding between solder and metal. They provide for metallically clean surfaces of the parts to be soldered, remove the oxides as well as other flow-inhibiting contaminations and prevent the formation of new oxides during the soldering process. A difference is made between acidic (as used in plumbing) and acid-free products (as used in electrical and electronic applications). It is most common to use solder wire with one or more flux cores in electronics production, whereas bar solder is the form of choice in plumbing as well as in the radiator and auto body work.

Examples of some common alloys

Alloy	Flux type	Melting point / range
L-Sn60Pb40	EN 29454/1.1.2 (F-SW 26/DIN 8511)	183 °C – 190 °C
L-Sn60Pb38Cu2	EN 29454/1.1.2 (F-SW 26/DIN 8511)	183 °C – 190 °C
L-Sn63Pb37	EN 29454/1.1.3 (F-SW 32/DIN 8511), free of halogen	183 °C eutectic
L-Sn62Pb36Ag2	EN 29454/1.1.3 (F-SW 32/DIN 8511), free of halogen	178 °C – 190 °C
Alloy – lead-free / complying RoHS-WEEE		
L-Sn95,5Ag3,8Cu0,7	EN 29454/1.1.2 (F-SW 26/DIN 8511)	217 °C eutectic
L-Sn96,5/Ag3,5	EN 29454/1.2.3 (F-SW 33/DIN 8511), free of halogen	221 °C eutectic
L-Sn99,3Cu0,7	EN 29454/1.2.3 (F-SW 33/DIN 8511), free of halogen	227 °C eutectic

5. A safe and clean working environment

to ensure quality and health

Safety first, especially while soldering. The VDE and GS emblems ensure the electrical safety of soldering equipment. The use of the VDE sign obligates the manufacturer to monitor production accordance with test guidelines and to conduct tests according to the regulations determined by the VDE testing institute.

Noxious gases develop in every soldering process

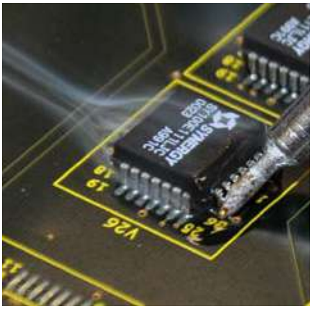

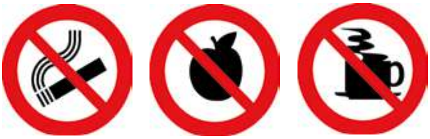

Health protection during soldering

- The breathing zone is very close to the soldering process during manual soldering, and the solder is added by hand. Thus there is the danger of contaminated air and hands or items which may have been touched
- Flux vapors can be damaging to the operator's health and should be kept out of the breathing area. Suitable devices for this purpose are solder fume extractions, which extract the smoke and associated vapors from the workplace and remove particles and gases. Modern fume extractions can be programmed to operate only when the soldering process is taking place, thus saving energy.
- One should never eat, drink or smoke in rooms where soldering is preformed. Contaminations which remain on the hands can enter the human organism through food or cigarette smoke.
- Hands should be washed thoroughly after soldering.

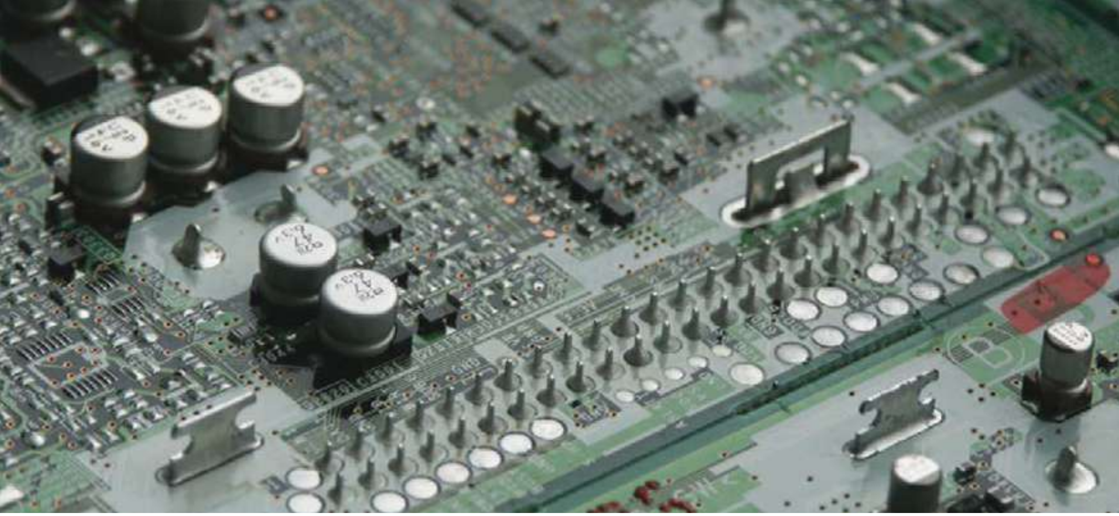

Best solder quality

due to the right preparation and correct soldering parameters

Preparation

The most important prerequisite for a good solder joint is absolute cleanliness. Conductor and component must be free of dirt, oil and oxides. They can be removed with solvents or flux.

Prior to soldering ERSADUR soldering tips should be cleaned while hot with a moist sponge or with a metallic dry sponge. Do not file the tip as you would copper tips, because this would damage the protective coating and render the tip unusable.

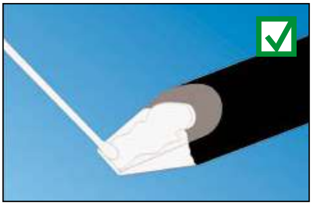

Wetting the soldering iron tip

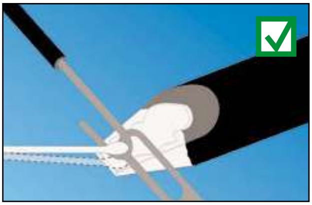

Soldering a joint

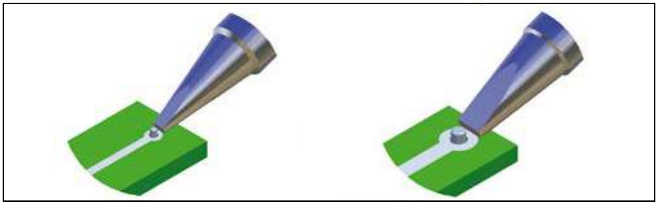

The selection of the right soldering tip is decisive for good soldering results

Soldering process

The soldering process has three phases: wetting, flowing, and bonding, whereas the working temperature is the most important criterion. It is best to work at the lowest temperature at which the three phases can progress smoothly. This requires some experience. A temperature-controlled soldering station will definitely facilitate this work. Place the tip on the joint to be soldered after cleaning and heat up the joint. Then feed the flux-cored solder wire between the soldering tip and the joint and melt as much solder as is required to wet the complete joint. Then remove the solder wire first and right after that the soldering tip to prevent overheating the solder. Allow the solder to solidify, avoiding any vibrations or jarring during this time.

Soldering time

The soldering process should be completed within 2 to maximum 5 seconds with a correctly dimensioned soldering tip. When soldering electronic components with lead-free solders, experience shows that more time is required. But even here requiring more than 5 seconds is not permissible, and it indicates that either the temperature setting is too low or the soldering iron lacks the necessary power.

Solder joint quality

When the leads of the components mounted on the board are crimped, a good solder joint has been formed when the contour of the soldered lead is still visible. This will not be the case, if an excessive amount of solder has been used to form the joint.

A further quality attribute is the wetting angle. This consideration is based on the fact that good wetting of the pad, discernible through a small wetting angle, has given rise to the formation of a diffusion zone (intermetallic zone). Wetting angles of up to 25º identify a good joint, wetting angles of up to 50º are still tolerable in manual soldering.

Another quality indicator is what the solder surface actually looks like. It should be smooth and shiny, without any porous areas visible. Grainy surfaces indicate either overheating of the solder or an excessive soldering time. Using lead-free solders, especially silver loaded alloys, matt surfaces may form.

The only absolute quality indicator for a sound and strong solder joint is the formation of the diffusion zone. In the diffusion zone, intermetallic compounds of copper and tin are formed, whose presence is the final proof of quality (mixed crystals, see page 5). Unfortunately, the only way to make this zone visible is through a destructive test (sectioning of the joint). If the diffusion zone is too thick, the solder joints have no tensile strength and become brittle. The higher the soldering temperature and the longer the soldering duration, the thicker the diffusion zone will be. Therefore the joint should be made at the lowest suitable temperature and within shortest soldering times.

As soon as the last solder joint has been made, the soldering iron is placed securely into the holder. At this time, the tip should not be cleaned, because the remaining solder on the tip prevents oxidization.

Correct!

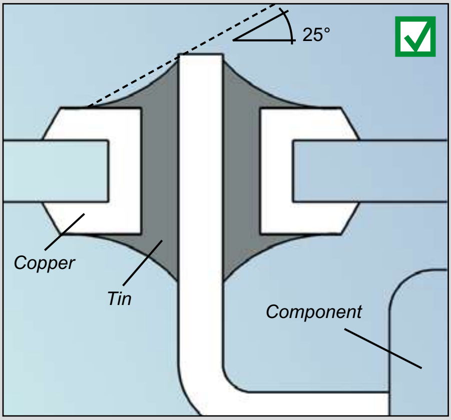

Wrong!

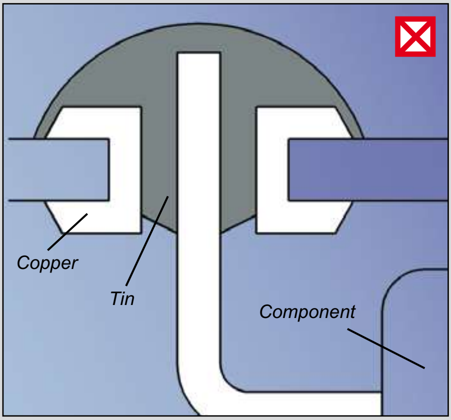

Desoldering with a solder wick

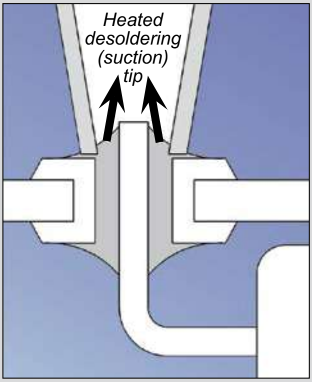

Desoldering with a temperature-controlled soldering iron

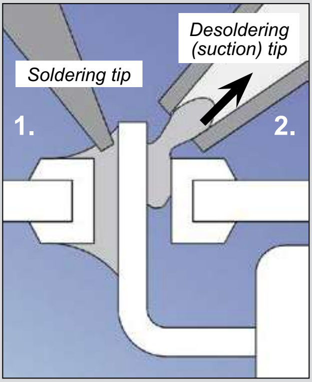

Desoldering with a desoldering pump

To achieve good results in desoldering, it is essential to select the right equipment. One can choose between desoldering wicks (desoldering using the principle of capillary action), mechanical desoldering pumps or electronically temperature-controlled desoldering systems. These are divided in desoldering systems with conductive heat or those with hot air.

Correct desoldering made easy

The right tool for each application

Desoldering

Reheating is not recommended for repairing a faulty solder joint. It is better to remove the solder and to resolder the joint. When using a desoldering pump, the solder joint has to be heated with the soldering iron until the solder has molten. Then the tip is removed and the desoldering pump is placed on the joint to extract the solder.

Using a heated desoldering tool, the hollow desoldering tip is placed on the joint to be repaired, making certain that there is good thermal contact. Once the solder has molten, it is extracted.

Desoldering is also dependent on proper tip selection. For example, the desoldering tip's inner diameter should be the same size as the diameter of the through-hole or even slightly larger (by max. 0.3 mm, see above drawing). The best desoldering results with least damage to PCB or the components can be achieved with temperature-controlled desoldering irons (also see pages 24/25). Please note that basically it has to be distinguished between the desoldering of through-hole components and the desoldering of SMD components.

Temperature-controlled desoldering iron

Sucking up the old, molten solder remnants with a mechanical solder sucker (desoldering pump)

Some examples of SMD components →

Ever smaller, ever finer

SMD technology, a true challenge

SMD Soldering

SMD technology (Surface Mount Device Technology) is currently the standard process in electronics manufacturing. Ever smaller and highly integrated surface mount components place growing demands on SMD soldering equipment.

Single solder joints, for example on resistors, are made with ERSADUR long-life soldering tips starting with a diameter of 0.2 mm. Fine-pitch connections with a high pin-out, i.e. IC's, are most efficiently soldered with Ersa PowerWell soldering tips.

Ersa PowerWell technology for soldering fine pitch components within seconds

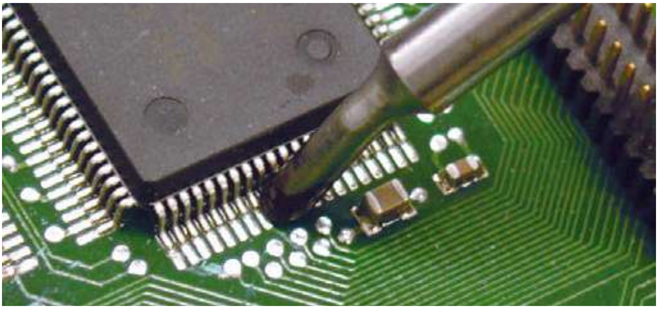

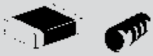

C.CAPAC./MELF

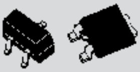

SOT&DPAK

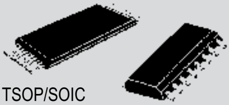

TSOP/SOIC

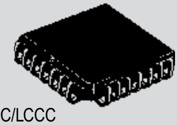

PLCC/LCCC

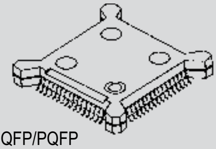

BQFP/PQFP

Soldering fine-pitch components

Process guide description

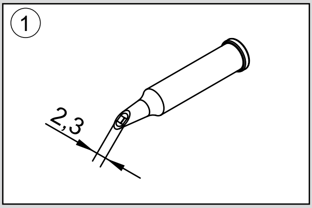
①

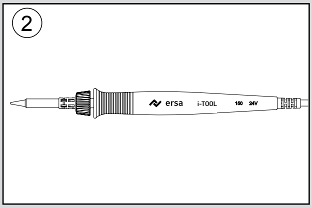
②

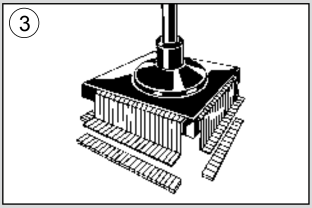
③

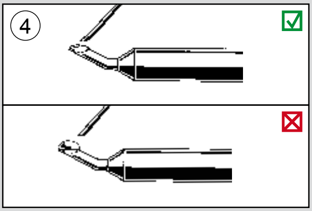
④

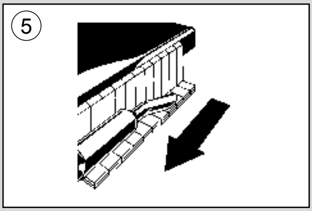
⑤

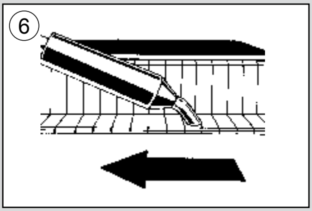
⑥

At first glance, soldering fine-pitch components by hand seems to be a tough job. Yet it is easy, with the right equipment at hand:

Insert an Ersa PowerWell i-TIP (1) into the i-TOOL soldering iron (2) and set a tip temperature of 285 °C to 360 °C (depending on the alloy used – tin/lead or lead-free).

Then position the component (3) and fix two corner pins.

Add flux cream (also see page 30) to the pins on all 4 sides. Clean the front and concave portion of the PowerWell tip with a damp sponge or the Ersa dry sponge.

Fill the concave portion with solder to slightly above the rim by melting the solder wire, until a small dome occurs (4). Take care not to add too much solder.

Place the i-TOOL lightly on the flat section of the pins (5), and pull the tip across the pins towards you (6) without exercising pressure.

Repeat steps (4) to (6) to solder the remaining sides.

Good solder joint, sufficient solder and good fillet

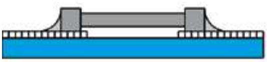

Wrong: insufficient solder

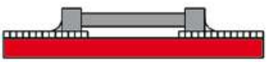

Wrong: too much solder

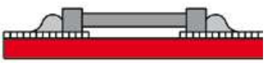

SMD desoldering in 3 seconds – with the Ersa desoldering tweezers

Some examples of different desoldering tip shapes: →

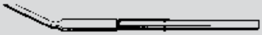

to desolder and remove MELF's

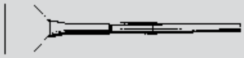

to desolder and remove SOIC's

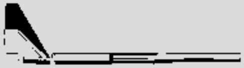

to desolder and remove QFP and PLCC components

SMD Desoldering

To desolder or rework a damaged SMD component, the suitable tools are required to remove the component from the board. When using desoldering tweezers, it is extremely important to select the proper pair of desoldering tips. After having desoldered the component, the pad has to be cleared of the residual solder (e. g. with a suitable soldering tip and a no-clean desoldering wick). Afterwards the new component can be positioned and soldered. An optional IR heating plate is a very useful addition – particularly in lead-free hand soldering applications.

More comprehensive instructions on SMD desoldering is available in the process description "SMD desoldering" on our website at www.ersa.com. For the soldering and desoldering of BGAs or other high pin-out SMDs, particularly those with hidden joints, we recommend the semi-automatic Ersa rework stations (see page 31).

Solder melting and removal with a desoldering wick

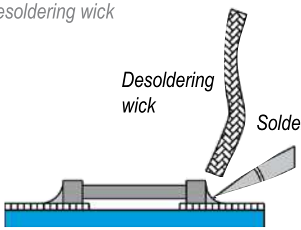

Ersa desoldering tweezers

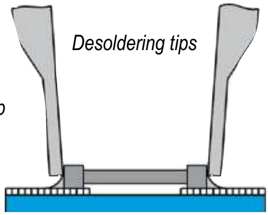

The compact CHIP TOOL VARIO desoldering tweezers are perfect for precise desoldering of very small SMDs.

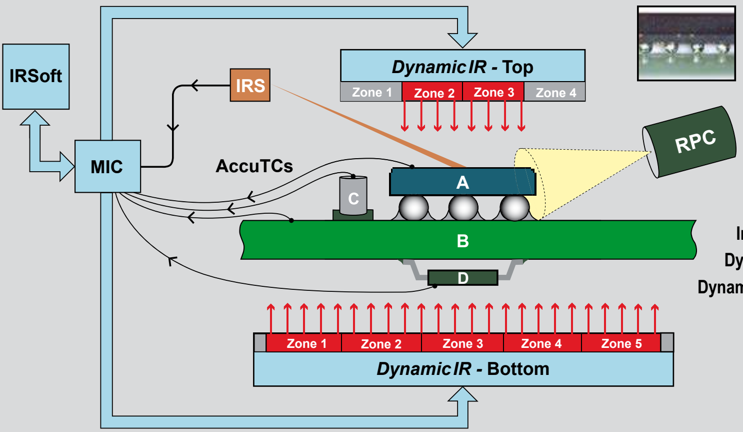

A:	Component for rework
B:	PCB
C:	Adjacent component (top side)
D:	Adjacent component (bottom side)
AccuTCs:	Thermocouples (4 each)
IntelligentIRS:	Non-contact infrared sensor
DynamicIR Top:	Top IR radiator (4 zones)
DynamicIR Bottom:	Bottom IR radiator (5 zones)
RPC:	Reflow Process Camera
IRSoft:	Control & documentation software
MIC:	Microprocessor controls the DynamicIR heating system

Principle of a rework station

Rework or repair soldering

The rework process demonstrated on the example of a BGA (Ball Grid Array)

Rework designates the repair or touch-up of electronic components such as Ball Grid Arrays (BGAs) or SMTs. This chapter describes the process steps for BGA rework:

1. Desoldering the BGA

The rework station heats the printed circuit board from the bottom, whereas the BGA body itself is heated from the top.

The real-time temperature of the component controls the pre-set temperature curve so that all solder joints melt at the same time. Then the vacuum suction cup is placed on the BGA and once all joints have molten the BGA is lifted off.

2. Removal of residual solder, cleaning

Solder remaining on the pads is removed with a soldering iron. To do so the residual solder is coated with flux. Then a flat soldering tip (e. g. 0102ADLF40 or 0102ZDLF150) is moved over the pad without applying any force. The solder adheres to the larger surface area of the tip, thereby levelling off the remaining solder of the connection pads. Flux residues are finally cleaned off (e. g. using Ersa FLUX-REMOVER).

3. Reballing – Reusing the BGA

Desoldered BGA's can be refit with new solder balls and re-attached to the board. This process is referred to as reballing. Residual solder is removed from the BGA by means of a soldering station. Lying on its back, the component is coated with flux first. Then the new solder balls are attached, for example by means of a stencil. These solder balls are heated up to the melting point with the rework station to firmly connect with the BGA body. Now the BGA is ready to be re-attached.

4. Application of flux or solder paste

The component and the connection pads are now fluxed, or, as the case may be, solder paste is added through stencil printing. The type of technique applied depends on the application, components and the skill level of the operators. For the commonly used PBGA's the application of flux is mostly sufficient.

5. Placing the component

Once the pad has been prepared, the component has to be placed. Since all solder joints of a BGA are hidden under the component body, a component placement unit is required. Successful visual placement by hand requires an operator with extensive experience and excellent skills. If the component is placed on solder paste deposits, great care has to be taken not to squash the solder depot, since doing so may lead to shorts after soldering.

6. Resoldering the BGA

The component is heated to the melting temperature of the solder alloy used via a controlled temperature curve. The heating continues until all solder joints have melted and remained so for some seconds. During this time durable and lasting solder joints are formed. After resoldering, the board is cooled in a similar fashion as had been done after desoldering.

These process steps are generally applicable for all surface mount components. Subject to the type of connections (wired, hidden) they may slightly vary.

Rework – Repair of high-terminal count IC's
to successfully repair SMT assemblies, the several points must be observed:

- Dimensions and properties of the assembly have an influence on its temperature requirement
- PCB holders and supports keep the assembly flat and prevent warpage
- A gentle and controlled heating process, continuously monitored, prevents damage of components or the board
- Accurate component placement is a prerequisite for a good soldering result
- Operators that are well trained will understand the process and ensure good results

Rework process steps

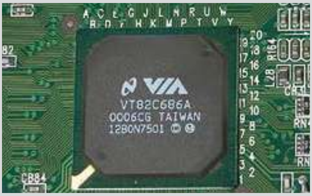

1. Remove BGA

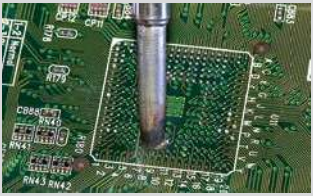

2. Remove residual solder from the pads of the board

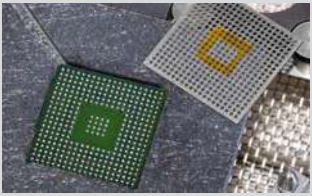

3. Reballing – addion of new solder balls

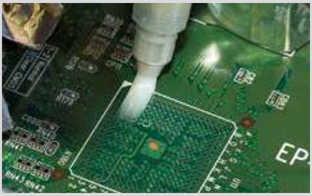

4. Application of flux or solder paste

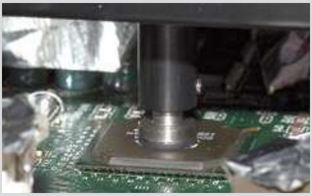

5. Placement of the new or reballed component

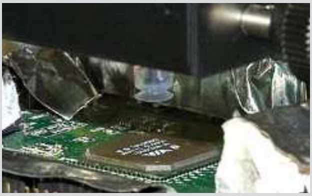

6. Resolder the BGA on the prepared surfaces of the board

Miniature soldering iron

Ersa MINOR S

MINOR S, 5 W

Ersa MINOR S (5 W), a featherweight at 6 g for ultrafine soldering applications

The **MINOR S** miniature soldering iron with a rating of 5 W and a maximum tip temperature of 440 ºC is an ideal tool for ultrafine soldering applications on micro IC's and under a microscope.
It can either operate with a 6 V transformer or a 6 V battery.

Besides electronics the MINOR S can also be used in watch repair, in the photographic industry and in dental technology.

Micro soldering irons

Ersa MULTITIP Series

MULTITIP series, 15/25 W

Ersa MULTITIP (15/25 W) internally heated soldering iron for small to medium solder joints

The **MULTITIP** soldering irons are especially short, light and easy-to-handle soldering irons with minimal distance between soldering tip and the handle's front part. They are ideally suited for small solder joints. Its internally heated soldering tip provides an enhanced degree of efficiency.

The MULTITIP is available for 15 W and 25 W which makes it suitable for micro soldering joints and medium-sized joints, as for example, on distributor strips or in the hobby sector.

Universal soldering iron

ERSA 30 S, MULTI-PRO, ERSA 15+ and ERSA 25+

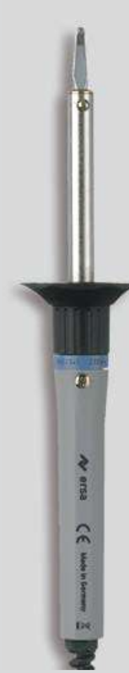

ERSA 30 S, 30/40 W

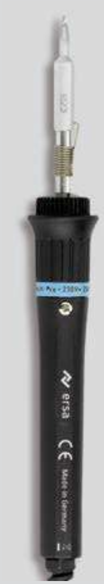

Ersa MULTI-PRO, 20 W

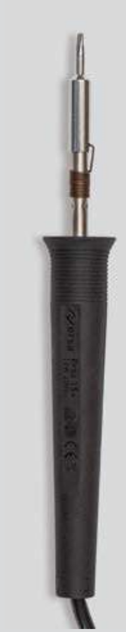

ERSA 15+, 15 W

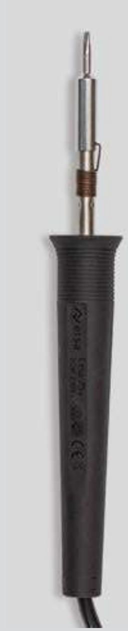

ERSA 25+, 25 W

The proven **ERSA 30 S,** available with a 30 or 40 W rating, is a very robust and durable soldering iron. Its improved ergonomics is further enhanced by the practical stick-on rubber support disk. This universal soldering iron is designed for multi-purpose use in the crafts, service and hobby sector.

The **MULTI-PRO** with its heat-resistant power supply cord is extremely flexible, as it can be operated with a wide range of different tips. Therefore, it is suitable for almost all conventional soldering applications.

The **ERSA 15+** universal soldering iron is the ideal tool for cost-effective soldering. Internally heated soldering tips guarantee an outstanding performance. The ERSA 15+ can be operated with diverse fine tips of the 832 and 842 tip series. Its ergonomic handle assists in the safe processing of fine soldering joints.

The **ERSA 25+** is perfect for soldering applications with increased heat requirements. The internally heated, larger soldering tips of the 832 and 842 tip series ensure good heat transfer at a high thermal performance. The ergonomic handle provides a safe grip for every soldering application.

Ersa universal soldering irons – for almost all applications

Standard and hammer soldering irons

ERSA 50 S, 80 S and 150 S / ERSA 200, 300 and 550

ERSA 50 S, 80 S, 150 S, 50/80/150 W

ERSA 200/300, 200/300 W

ERSA 550, 550 W

The **ERSA 50 S/80 S** or **150 S** series is designed for applications where a large amount of heating capacity is required, as, for example, on copper conductors with cross-sections of 2.5 mm to 6 mm.

The **ERSA 200, 300** and **550** hammer soldering iron series are particularly suited for sheet metal processing and installation work and for soldering commutators and copper bus bars.
Hammer soldering irons have also proven their merit in leveling applications during body work and lead glazing.

High-speed soldering iron

Ersa MULTI-SPRINT

The **Ersa MULTI-SPRINT** is an ergonomically designed and extremely light-weight, pistol-type soldering iron with up to 150 W power which does not require a transformer. Its PTC heating element, together with the internally heated ERSADUR long-life soldering tip, ensures its exceptionally high efficiency. Due to its very short heat-up time, it is ideally suited for making quick single soldering joints. The MULTI-SPRINT is powered only as long as the button is pressed. Depending on the heat requirement of the solder joint to be made, additional energy can be supplied by periodically pressing the button. The large variety of soldering tips of the 832 / 842 series make the MULTI-SPRINT suitable for more than only the use in repair and service applications.

Ersa MULTI-SPRINT, 150/75 W

Power soldering iron

Ersa MULTI-TC and PTC 70

Ersa MULTI-TC, 75 W

Ersa PTC 70, 75 W

The **MULTI-TC** and **PTC 70** are powerful and robust, temperature-controlled soldering irons. Both tools offer an outstanding heat-up rating.

Due to their high thermal performance, and because of the large selection of soldering tips, the MULTI-TC and PTC 70 are suitable for fine soldering joints in electronics as well as for joints with medium heat requirements. The MULTI-TC is furthermore also used anywhere else where standard irons with 150 W power are in use, including, for example, lead glass and Tiffany soldering.

Gas soldering irons

Ersa INDEPENDENT 75 and INDEPENDENT 130

Ersa INDEPENDENT 75, 15 – 75 W

Ersa INDEPENDENT 130, 25 – 130 W

Ersa gas soldering irons are fueled with commercially available lighter gas and are fired up through the piezo ignition. Compared with electrical soldering irons, the **INDEPENDENT 75** has between 15 – 75 W, and the **INDEPENDENT 130** between 25 – 130 W performance. Both irons are available in the BASIC-SET and PROFI-SET versions.

Aside from handling the usual types of electronic components, the selection of tips available enables the INDEPENDENT to also handle SMD soldering, micro-welding, forming and cutting of synthetic materials and the processing of shrink sleeves.

The basic soldering station

Ersa ANALOG 60

The electronically temperature-controlled ANALOG 60 is Ersa's basic soldering station model. It has the tried and proven RESISTRONIC temperature control technology with the PTC heating element serving as the temperature sensor. The high heat-up rating of 190 W guarantees the immediate supply of heat and a heat-up from room temperature to 280 ºC within 60 seconds. The BASIC TOOL 60 soldering iron uses the internally heated ERSADUR long-life soldering tips of the 832/842 series and provides very high performance.

Due to the wide range of 832/842 soldering tips, the Ersa ANALOG 60, which is also available as an antistatic version, covers a wide range of applications with the most varied soldering requirements.

Digital soldering station

Ersa RDS 80 – high performance for low cost

The digital soldering station Ersa **RDS 80** offers the proven and tested Ersa RESISTRONIC temperature control technology with a strong heating power of 80 W. The ceramic PTC heating element (Positive Temperature Coefficient) acts as the temperature sensor in this control system. Due to its very high ramp-up capability of up to 190 W, the station reaches operating temperature very fast.

RT 80: sleek soldering iron with a large selection of soldering tips

Multifunctional soldering stations

Ersa i-CON NANO and i-CON PICO – small, powerful and versatile

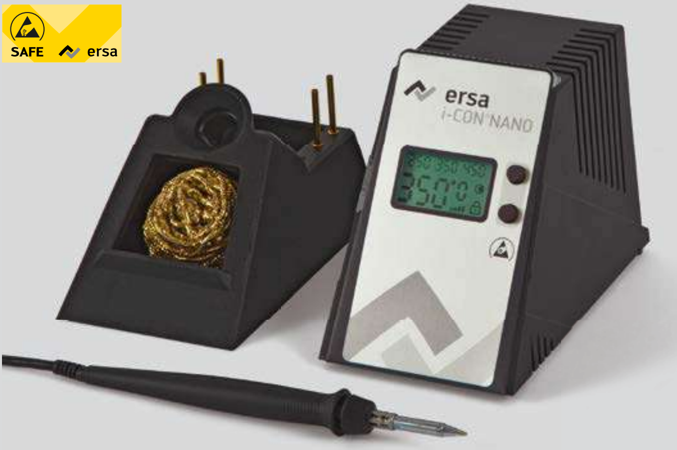

The soldering stations **i-CON NANO** and **i-CON PICO**, two models of the i-CON product family, fulfill all the needs of today's electronics manufacturing while requiring minimal space. They are designed for continuous operation in the electronic manufacturing environment as well as for special applications in laboratories and R&D departments.

Due to the simple and user-friendly operating concept, the factory settings provide for a variable adjustment of operating temperature as well as the setting of stand-by time and calibration value.

Further adjustments such as fixed temperatures, power level, interlock and shutdown functions are available with the free PC software and an optionally available micro smart SD card.

Verbatim
micro SD
Adapter
LogiLink

Micro-SD card with SD-card and USB adapter

The concept of the Ersa i-CON stations ensures that each application is processed with the optimal parameters. They stand for the highest level of process safety and quality control at low investment and operating cost.

Ersa i-CON PICO

- Small footprint (145 x 80 mm) – saves valuable space
- Antistatic as per MIL-SPEC/ESA (only i-CON NANO)
- Three fixed temperature settings or continuously adjustable temperature settings from 150 °C to 450 °C
- Three selectable power levels
- Ultra-light and ergonomic soldering tool with max. 80 W power
- Wide range of low-cost exchangeable long-life soldering tips
- Automatic stand-by and sleep function for low energy consumption and longer tip life
- Password interlock for maximum process control
- Calibration function for a precise tip temperature
- Complete parameterization through PC software and Micro-SD card

High-end soldering and desoldering stations

The Ersa i-CON product family for highest productivity and process safety

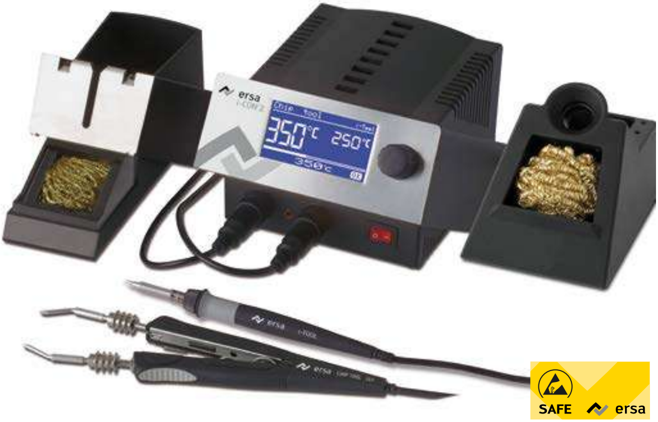

Ersa *i*-CON2 with *i*-TOOL and CHIP TOOL

The stations of the i-CON product family are Ersa's innovative solution for intelligent manual soldering. The need to be able to cope with higher working temperatures and with progressively smaller process windows when working with lead-free solders, poses no problem whatsoever for the i-CON product family.

The i-CON is available as a single station or as double iron station. The single station **i-CON1** is delivered together with the i-TOOL soldering iron. The i-TOOL is extremely small, ultra-light and ergonomic. It is powered by a 150 W micro heating element, which realizes short heat-up times (within 9 seconds up to 350 °C!) and rapid heat recovery.

The microprocessor which stores the temperature calibration of the iron is located in the PCB which is installed in the handle. This now allows for each individual i-TOOL to be calibrated independent of the soldering station.

In contrast to the concept followed by the cartridge-type soldering tips, only the tip itself is exchanged at the i-TOOL. The cost-intensive heating element remains.

The double iron soldering station **i-CON2** can be operated either with a second i-TOOL or with the SMD desoldering tweezers CHIP TOOL

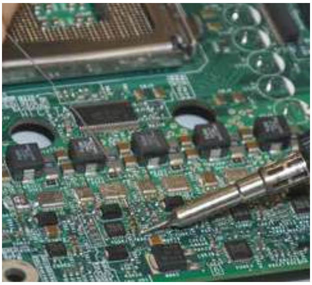

Soldering miniature and densely placed SMD components with the i-TOOL and i-TIPS, starting at a 0.2 mm diameter.

i-CON1 with i-TOOL soldering iron with high-performance micro heating element

- i-TOOL soldering iron with 150 W micro heating element technology
- Low-cost exchangeable long-life tips of the i-TIP series
- User friendly “One-Touch” operation
- Three power levels – no overshoot
- Process window and alarms
- Interface to control peripheral equipment on the workplace, such as heating plate and solder fume extraction systems
- Stand-by control for tools, heating plate and solder fume extraction system
- i-TOOL calibration
- Tools for SMT and conventional soldering applications
- Automatic tool recognition

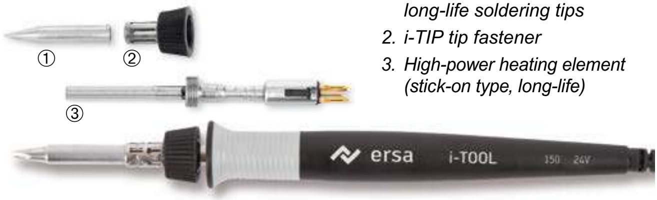

1. *Low-cost i-TIP long-life soldering tips*
2. *i-TIP tip fastener*
3. *High-power heating element (stick-on type, long-life)*

Ersa i-TOOL: The ideal soldering iron – ultra light (only 30 g), ultra short (only 155 mm), and ultra short tip-to-grip (only 45 mm) and extremely powerful.

respectively with the desoldering irons X-TOOL or X-TOOL VARIO.

The CHIP TOOL facilitates safe and quick desoldering, of the smallest chips up to large PLCC's. To remove the residual solder and to desolder wired components – also on multilayer boards – the X-TOOL or X-TOOL VARIO desoldering iron is the right tool.

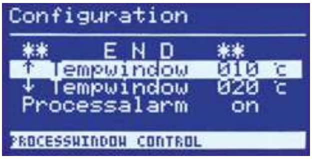

```
Configuration
 **     E N D     **
 ↑ Tempwindow    010 °C
 ↓ Tempwindow    020 °C
 Processalarm     on

PROCESSWINDOW CONTROL
```

User-friendly controls: quick programming & lock, huge multi-functional display with i-Op controls, menu in 7 languages, online help

High-mass through-hole desoldering with the X-TOOL VARIO

The C-line of the i-CON stations was developed, so that peripheral equipment could be controlled or to communicate with them. Via a serial interface, the i-CON1 C or the i-CON2 C controls the Ersa IR heating plates or the Ersa solder fume extraction systems.

i-CON VARIO 4 multi-channel station with X-TOOL VARIO desoldering iron, i-TOOL soldering iron, i-TOOL AIR S hot-air iron and CHIP TOOL VARIO desoldering tweezers.

Top Tips

i-TIPS for all applications

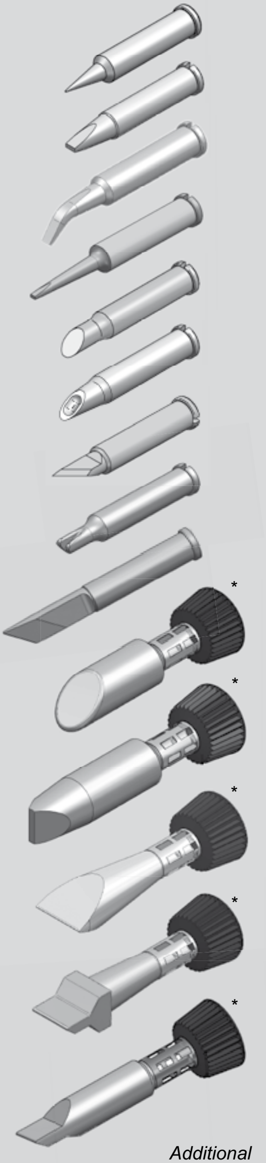

**Figure with i-TIP tip fastener*

Additional soldering tips for the i-TOOL and the i-CON soldering stations is available on our web site:
www.ersa.com

Hybrid Rework Station

Ersa HR 100 A – The innovative combination of hot air and infrared radiation

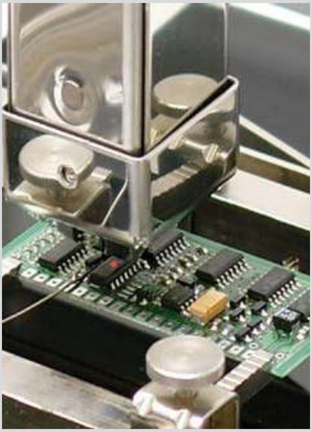

The Ersa **HR 100 A** applies the Ersa Hybrid Rework Technology for a safe desoldering and replacement of small SMD components. Its medium wave length IR radiation, combined with a safe stream of hot air ensures optimal heat transfer to the component.

The HYBRID TOOL offers a gentle and homogeneous warming of all sizes of components, from 0201 chips up to 20 x 20 mm SMD's and larger. Exchangeable hybrid adapters target the thermal energy available (up to 200 W) on to the component, all the while protecting neighboring areas and not blowing away or moving adjacent chips.

Its user-friendly handling permits even operators with little experience to safely and efficiently operate the HR 100. More experienced operators, on the other hand, can not only variably adjust the air flow as well as the heat output, but also record and run profiles. A positioning laser whose laser point makes it possible to keep track of the component worked on through the complete process is integrated in the grip of the ergonomic hybrid tool.

SAFE ersa

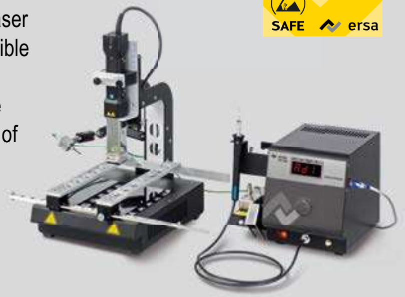

Complete system with HR 100 A, IR heating plate, stand for HYBRID TOOL and PCB holder.

- HYBRID TOOL with 200 W heating element, positioning laser integrated in grip
- 3 exchangeable hybrid adapters (6 x 6 mm, 10 x 10 mm and 20 x 20 mm)
- Silent rework fan (below 40 dB)
- Integrated vacuum pump and vacuum pen; tool holder and K-type thermocouple input socket; USB interface; LED display
- 2-channel temperature recording:
  TC & IRS; AccuTC and Flexpoint thermocouple holder
- Tool holder with z-axis height adjustment
- x-y PCB holder (290 mm x 250 mm)
- 800 W IR heating plate with glass cover:
  125 mm x 125 mm IR high-performance heating element
- Rework profile and documentation software Ersa IRsoft

CLEAN-AIR solder fume extractions

Ersa EASY ARM 1 and EASY ARM 2 – protecting environment, health and resources

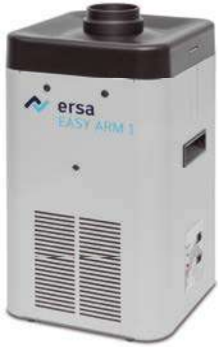

Ersa EASY ARM 1

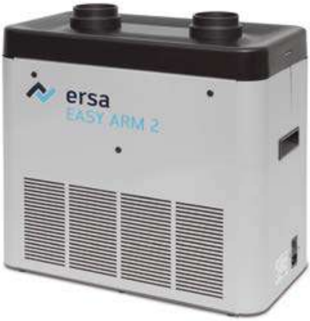

Ersa EASY ARM 2

The Ersa **EASY ARM 1** is a compact yet powerful filter unit to efficiently clean the process air at the workplace. Both i-CON1 C, i-CON2 C and i-CON VARIO can be connected to the EASY ARM 1 with an interface cable. The solder fume extraction has three filtration levels to remove noxious gases.

With the solder fume extraction **EASY ARM 2**, Ersa offers the user a further compact and highly efficient fume extraction unit for either one or two workplaces. One or two units i-CON1 C / i-CON2 C / i-CON VARIO can be connected to the EASY ARM 2 via interface cable.

Both EASY ARM systems are provided with a stand-by mode and operate only when the soldering stations they are connected to are being used. The EASY ARM 1 and 2 are easy to install and can be placed very flexibly. Due to their very low noise level, they can be operated in virtually all environments, be it repair shops, development facilities or laboratories.

To accommodate different working conditions and applications, a variety of extraction arms and nozzle shapes which can easily and quickly be exchanged are available to meet different working conditions.

Q&A 1: So, how can you get rid of them? By using a solder fume extraction unit to clear the air from particles and gases.

Noxious gases develop in every soldering process.

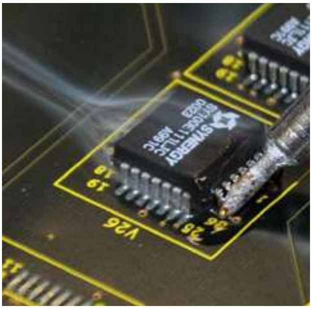

Q&A 2: How does that work? Active carbon in the filters absorbs harmful gas molecules; result: a clean working environment.

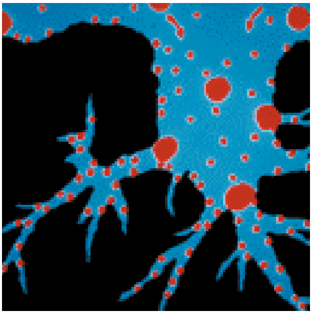

Tiffany soldering

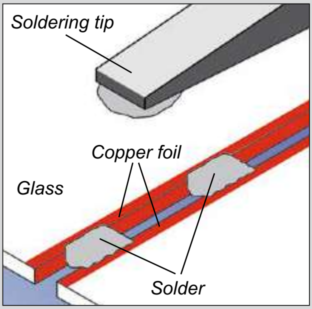
Soldering tip
Copper foil
Glass
Solder

Step 1: spot soldering

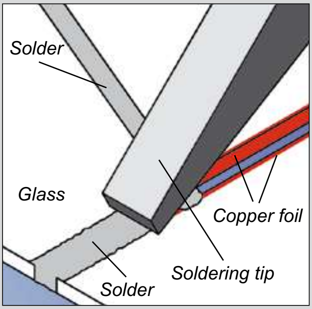
Solder
Glass
Copper foil
Soldering tip
Solder

Step 2: rough soldering

Tiffany or sheet metal soldering

Soldering beyond the field of electronics

Tiffany soldering (Lead glass soldering)

Soldering a Tiffany object generally involves three individual steps:

- Spot soldering
- Rough soldering
- Finishing

Prior to the actual soldering process, copper foil is glued on the glass edges. The next step is spot soldering, i. e. the glass parts are fastened or connected by taking a drop of solder with the tip of the soldering iron and carefully applying it to the solder joint. Each spot soldering operation should only take about a second. During rough soldering the gap between the glass is filled with solder after flux has been applied. For this purpose the tip and solder are moved together along the joint. Always drag the soldering iron, never push it. Only if this procedure is accurately followed, and if a sufficient amount of solder is applied, the desired, half-round and convex seams are achieved. The lack of visual quality of the seam at this stage is optimized during the finishing soldering step, in which the soldering tip is dragged slowly and from the beginning of the seam until the end at an even speed. The seam worked on should always lie flat on the bench.

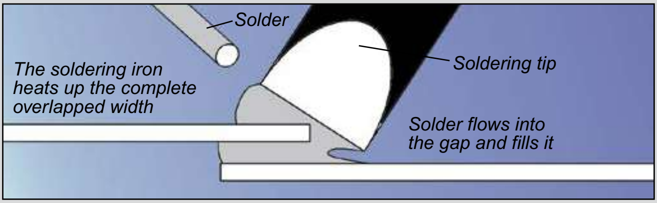
Solder
The soldering iron heats up the complete overlapped width
Soldering tip
Solder flows into the gap and fills it

Guiding the soldering iron when soldering a broad seam

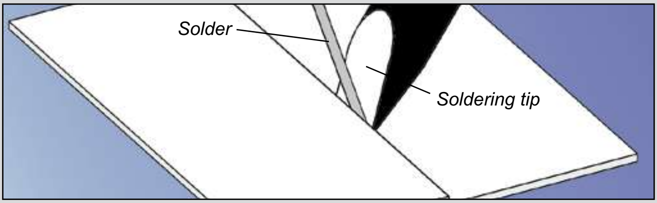
Solder
Soldering tip

Guiding the soldering iron when soldering a narrow seam

Plumbing and tinsmith work

For the joining of sheet metal and metal pipes, the joints to be soldered have to be bare. This calls for a good prior cleaning.
After this, the flux – either solder grease or solder fluid, a zinc chloride solution – is applied. Then the area to be soldered is heated with the tip of the soldering iron.
Once the soldering area is sufficiently hot, solder is added and the solder gap is filled.
The aggressive flux residues are removed after cooling to avoid the risk of future corrosion.

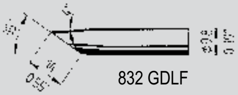
832 GDLF

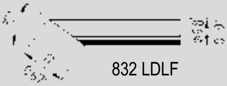
832 LDLF

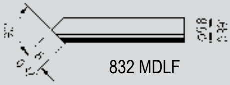
832 MDLF

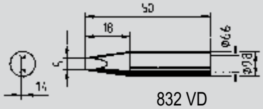

832 VD

ERSADUR Tiffany soldering tips

The types VD, GDLF, LDLF and MDLF of the comprehensive 832 tip series are particularly suited for Tiffany work. On account of their shapes and their high mass (excellent because of the ensuing heat retention capacity), the seams between the glass parts can quickly and easily be filled with solder. And the ERSADUR finish warrants a long tip lifetime.

ERSADUR Tiffany soldering tips

Ersa DIGITAL 2000A, Ersa MULTI-TC and ERSADUR soldering tips

Tiffany lamp

The **MULTI-TC** soldering iron is very light, robust and powerful. The Ersa SENSOTRONIC Control with the PT-1000 temperature sensor in the tip reacts immediately when heat is withdrawn.

The MULTI-TC is provided with a 2.2 mm wide, chisel-shaped soldering tip. Together with the Tiffany soldering tips, the slim MULTI-TC is superbly suited for the use in Tiffany work.

Due to its precise temperature control, the **DIGITAL 2000 A** soldering station completely eliminates the possibility to overheat any glass components or the copper foil, and the station has tremendous power reserves, making it comparable and putting it on an equal level with unregulated 150 W soldering irons.

The Ersa SENSOTRONIC control with its internally heated soldering tip, where the temperature sensor is mounted directly below the tip, provides for precise temperature control and ensures uniform temperature levels at the soldering joint.

The long-life ceramic PTC heating element provides up to 290 W heat-up rating which means that the powerful soldering iron is ready for use in only 60 seconds.

The DIGITAL 2000 A is provided with a 2.2 mm wide, chisel-shaped soldering tip. Particularly good results are achieved with the optionally available tip versions 832 VD, 832 GDLF, 832 CDLF and 832 MDLF, which are specifically offered for Tiffany applications.

Ersa MULTI-TC, 75 Watt

Ersa DIGITAL 2000 A with POWER TOOL

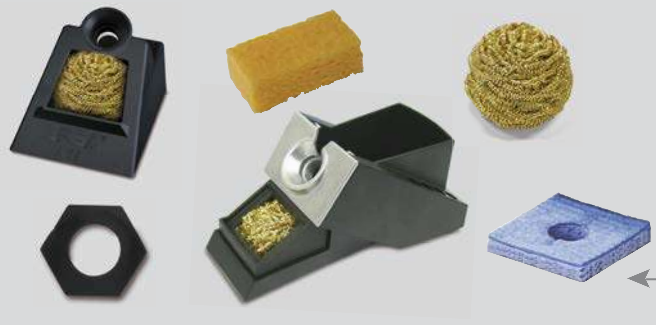

Ersa tool holders or stick-on rubber supports to safely and ergonomically put down the soldering iron during work stoppages or during heat-up.

Ersa viscose cleaning sponges for moist or special metal wool for dry cleaning of the hot soldering tip prior to the soldering process.

Auxiliary products and practical accessories

Everything for your soldering needs from one source

Ersa solder baths

Ersa solder wire and solder wire dispenser

Ersa DTM 100 temperature measuring device

Ersa SOLDAPULLT desoldering pump

Ersa VAC X desoldering pump

Ersa SMD flux and flux cream

Ersa tip exchanger

Ersa desoldering wicks

Ersa FLUX-REMOVER

Ersa TIP-REACTIVATOR

Further peripherals and accessories can be obtained from your local Ersa distributor and at **www.ersa.com**. Or just ask for our catalog!

IR rework stations

Ersa IR/PL 550, IR/PL 650 and HR 600 – from mobile phone boards up to XXL boards

Ersa HR 600

More than 5,000 customers worldwide use Ersa's patented IR-Rework technology.

Ersa rework stations deploy the DynamicIR Heating Technology, programmable top and bottom heating zones as well as the accurate, user-friendly and motorized "Auto Pick & Place" feature. System control, process documentation and visualization is handled by the Ersa rework software IRSoft/HRSoft.

Optical solder joint inspection

ERSASCOPE inspection systems – the patented original

Since its introduction over 10 years ago more than 3,000 customers worldwide profit from the possibility to non-destructively inspect hidden solder joints.

Regardless of whether the joints are under a flip-chip, or whether areas are to be inspected where other microscopes have come to their limits – the ERSASCOPE technology offers a considerable added value to any quality assurance program!

Ersa Know-How Seminars

Expert knowledge for professional users comprehensively imparted

Today more than ever, cost-effective production and highest quality in the electronics manufacturing industry is the basis for profitable competitiveness. A key ingredient for being competitive is a well-qualified staff, which is up-to-date on current technology and process.

Ersa Know-How Seminars are ideal for transmitting this knowledge to those of your staff members with process responsibility in electronic manufacturing. In these seminars, theoretical and practical knowledge is presented in a neutral fashion in workshops and presentations by reputed experts in their field.

For further information go to our website **www.ersa.com** and click "Events".

Ersa Electronics Production Equipment

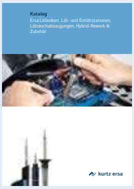
Katalog
Ersa Lötkolben, Löt- und Entlötstationen, Löttrauchabsaugungen, Hybrid-Rework & Zubehör
kurtz ersa

Technical data and further detailed information on our range of hand soldering equipment is available in our "Soldering Tools" catalog. This is available in printed form and as PDF, as are the "Rework & Inspection" catalog.

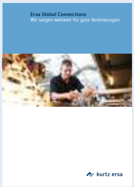
Ersa Global Connections
Wir sorgen weltweit für gute Verbindungen
kurtz ersa

Aside from the extensive selection of hand soldering tools, rework stations and inspection systems for non-destructive inspection, Ersa, as Europe's largest manufacturer of soldering systems, also offers a complete range of selective, wave and reflow soldering systems, as well as stencil printers for the industrial electronic manufacturing industry.

Please refer to our brochures and our website **www.ersa.com** for further information.

Ersa GmbH
Leonhard-Karl-Straße 24
97877 Wertheim/Germany
Tel. +49 9342 800-0
Fax. +49 9342 800-127
info@ersa.de
www.ersa.com

Ersa France
Chevigny Saint Sauveur, France
info-efr@kurtzersa.com

Ersa North America
Plymouth, USA
info-ena@kurtzersa.com

Ersa Asia Pacific
Hong Kong, China
info-eap@kurtzersa.com

Ersa Shanghai
Shanghai, China
info-eap@kurtzersa.com

kurtz ersa

97596 - 02/2015 • subject to changes • © by Ersa GmbH, Germany

Soldering Guide

Safety Precautions

- **Never touch the element or tip of the soldering iron.**
  They are very hot (about 400°C) and will give you a nasty burn.
- **Take great care to avoid touching the mains flex with the tip of the iron.**
  The iron should have a heatproof flex for extra protection. Ordinary plastic flex melts immediately if touched by a hot iron and there is a risk of burns and electric shock.
- **Always return the soldering iron to its stand when not in use.**
  Never put it down on your workbench, even for a moment!
- **Allow joints a minute or so to cool down before you touch them.**
- **Work in a well-ventilated area.**
  The smoke formed as you melt solder is mostly from the flux and quite irritating. Avoid breathing it by keeping you head to the side of, not above, your work.
- **Wash your hands after using solder.**
  Solder contains lead.

Treatment for minor burns

Most burns from soldering are likely to be minor and treatment is simple:

- **Immediately cool the affected area under gently running cold water.**
  Keep the burn in the cold water for at least 5 minutes (15 minutes is recommended). If ice is readily available this can be helpful too, but do not delay the initial cooling with cold water.
- **Do not apply any creams or ointments.**
  The burn will heal better without them. A dry dressing, such as a clean handkerchief, may be applied if you wish to protect the area from dirt.
- **Seek medical attention if the burn covers an area bigger than your hand.**

Preparing the soldering iron

- **Place the soldering iron in its stand and plug in.**
  The iron will take a few minutes to reach its operating temperature of about 400°C.
- **Dampen the sponge in the stand.**
  The best way to do this is to lift it out the stand and hold it under a cold tap for a moment, then squeeze to remove excess water. It should be damp, not dripping wet.
- **Wait a few minutes for the soldering iron to warm up.**
  You can check if it is ready by trying to melt a little solder on the tip.
- **Wipe the tip of the iron on the damp sponge.**
  This will clean the tip.
- **Melt a little solder on the tip of the iron.**
  This is called 'tinning' and it will help the heat to flow from the iron's tip to the joint. It only needs to be done when you plug in the iron, and occasionally while soldering if you need to wipe the tip clean on the sponge.
- **You are now ready to start soldering!**
  Please turn the page for further instructions...

© John Hewes 2006, The Electronics Club, **www.kpsec.freeuk.com**
This document may be copied for educational purposes.

Making soldered joints

- **Hold the soldering iron like a pen, near the base of the handle.**
  Imagine you are going to write your name!
  Remember to never touch the hot element or tip.
- **Touch the soldering iron onto the joint to be made.**
  Make sure it touches both the component lead and the track.
  Hold the tip there for a few seconds and...
- **Feed a little solder onto the joint.**
  It should flow smoothly onto the lead and track to form a volcano shape as shown in the diagram below. Make sure you apply the solder to the joint, not the iron.
- **Remove the solder, then the iron, while keeping the joint still.**
  Allow the joint a few seconds to cool before you move the circuit board.
- **Inspect the joint closely.**
  It should look shiny and have a 'volcano' shape. If not, you will need to reheat it and feed in a little more solder. This time ensure that both the lead and track are heated fully before applying solder.

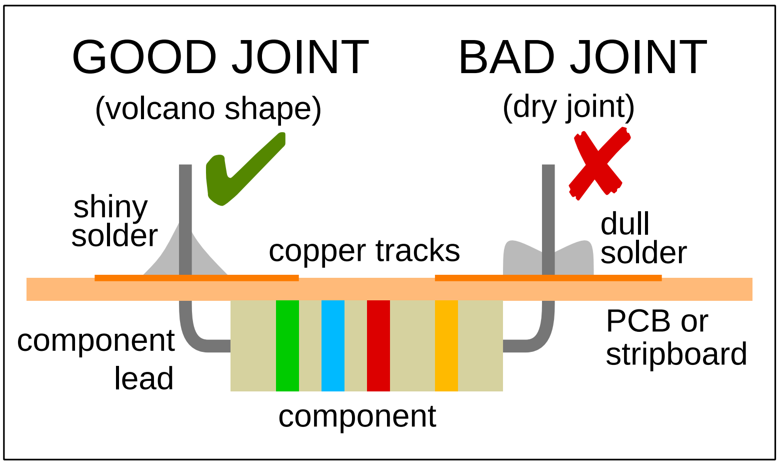

Using a heat sink

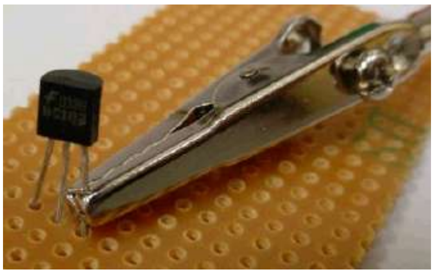

Some components, such as transistors, can be damaged by heat when soldering. It is wise to use a heat sink clipped to the lead between the joint and the component body, as shown in the picture. You can buy a special tool, but a standard crocodile clip works just as well and is cheaper!

Soldering advice for components

Some components require special care when soldering. Many must be placed the correct way round and a few are easily damaged by the heat from soldering. Appropriate warnings are given in the table on the next page, together with other advice which may be useful when soldering.

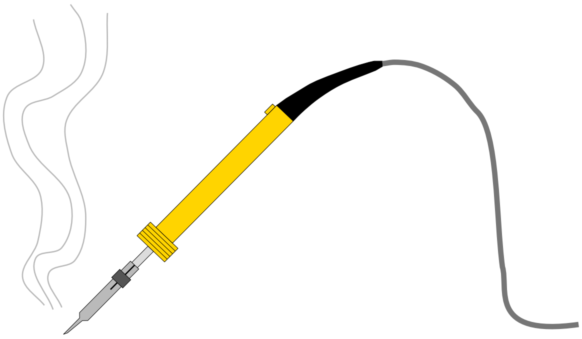

THE ELECTRONICS CLUB

© John Hewes 2006, The Electronics Club, **www.kpsec.freeuk.com**
This document may be copied for educational purposes.

Components	Pictures	Soldering advice
Resistors	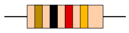	No special precautions are required. Connect either way round.
Diodes	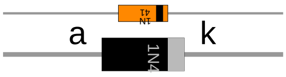	Diodes must be connected the correct way round: a = anode, k = cathode. Use a heat sink with germanium diodes.
IC holders (DIL sockets)	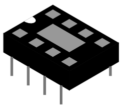	Ensure the notch is at the correct end. Do **not** insert the IC at this stage to prevent it being damaged by heat.
Presets (small variable resistors)		No special precautions are required. On stripboard take care to ensure you insert them the correct way round.
Capacitors, non-polarised (less than 1µF)	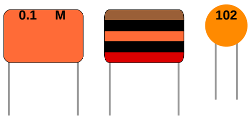	No special precautions are required. Connect either way round. Take care to identify their value.
Capacitors, electrolytic (1µF and greater)	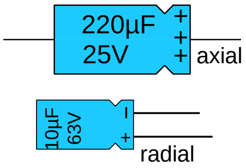	Electrolytic capacitors must be connected the correct way round, they are marked with + or - near one lead.
LEDs (Light Emitting Diodes)	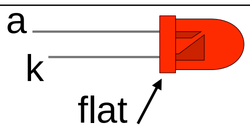	LEDs must be connected the correct way round: a = anode, k = cathode. Use a heat sink with small (3mm) LEDs.
Transistors	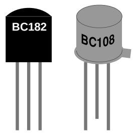	Transistors have three leads and must be connected the correct way round. Use a heat sink clipped to each lead in turn between the joint and the transistor.
Wire links between points on the board		Use tinned copper wire (such as the offcut from a resistor lead) or single-core plastic-coated wire.
Other parts mounted on the board		No special precautions are required for most parts, but make sure they are the correct way round.
Battery clips, buzzers and other parts with wires		Red (+) and black (-) wires must be connected the correct way round.
Wires to parts off the board such as switches		Use plastic-coated stranded wire which is flexible, single-core wire is likely to break at the joint.
Integrated Circuits (ICs or 'chips')	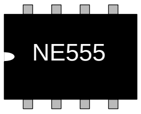	When all soldering is complete, carefully insert ICs the correct way round in their holders. Make sure all the pins are lined up before pushing in firmly.

For further information about electronic components please see: www.kpsec.freeuk.com

THE
ELECTRONICS
CLUB

© John Hewes 2006, The Electronics Club, **www.kpsec.freeuk.com**
This document may be copied for educational purposes.

Desoldering

At some stage you will probably need to desolder a joint to remove or re-position a wire or component. There are two ways to remove the solder:

1. With a desoldering pump (solder sucker)

- Set the pump by pushing the spring-loaded plunger down until it locks.
- Apply both the pump nozzle and the tip of your soldering iron to the joint.
- Wait a second or two for the solder to melt.
- Then press the button on the pump to release the plunger and suck the molten solder into the tool.
- Repeat if necessary to remove as much solder as possible.
- The pump will need emptying occasionally by unscrewing the nozzle.

2. With solder remover wick (copper braid)

- Apply both the end of the wick and the tip of your soldering iron to the joint.
- As the solder melts most of it will flow onto the wick, away from the joint.
- Remove the wick first, then the soldering iron.
- Cut off and discard the end of the wick coated with solder.

After removing most of the solder from the joint(s) you may be able to remove the wire or component lead straight away (allow a few seconds for it to cool). If the joint will not come apart easily apply your soldering iron to melt the remaining traces of solder at the same time as pulling the joint apart, taking care to avoid burning yourself.

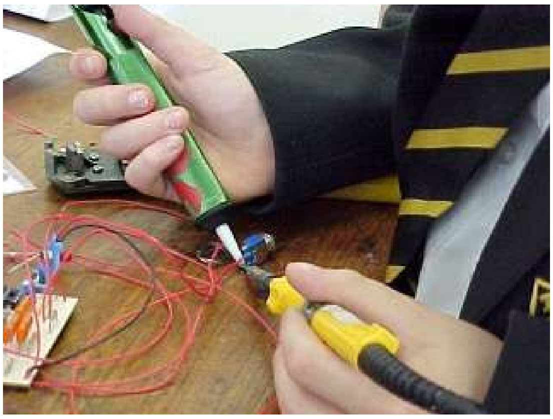

Using a desoldering pump (solder sucker)

What is solder?

Solder is an alloy (mixture) of tin and lead, typically 60% tin and 40% lead. It melts at a temperature of about 200°C. Coating a surface with solder is called ‘tinning’ because of the tin content of solder. Lead is poisonous and you should always wash your hands after using solder.

Solder for electronics use contains tiny cores of flux, like the wires inside a mains flex. The flux is corrosive, like an acid, and it cleans the metal surfaces as the solder melts. This is why you must melt the solder actually on the joint, not on the iron tip. Without flux most joints would fail because metals quickly oxidise and the solder itself will not flow properly onto a dirty, oxidised, metal surface.

The best size of solder for electronic circuit boards is 22swg (swg = standard wire gauge).

For plugs, component holders and other larger joints you may prefer to use 18swg solder.

© John Hewes 2006, The Electronics Club, **www.kpsec.freeuk.com**
This document may be copied for educational purposes.

Safety Rules for Soldering

Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island

1. Wear goggles for eye protection. When trimming off leads or excess solder dross, be careful of the flyaway that could injury yourself as well as other people nearby.
2. Work in a well-ventilated area and use a fume extractor. Do not inhale fumes from the soldering processes.
3. Always return the soldering iron to its stand when not in use. Never put it down on your workbench. The soldering iron tip is very hot (about 400°C). Avoid touching plastic, wire insulator, or any flammable material in the working area with the soldering iron. Turn the soldering station to standby or off if not used for more than few minutes. Turn unit off or unplug it when done.
4. Use a third hand, a circuit board vice, pliers, tweezers, or clamps for holding components to avoid burns. Legs and arms should be covered to avoid burns from splashed hot solder.
5. Do not have food or drink near the working area. The solder is usually a tin/lead alloy and lead is toxic. The flux is a chemical used to help metal parts soldered together. It is acidic and toxic. Clean up spilled flux immediately. Wash hands after soldering. Flux can cause acid burns to the skin or damage clothing. In case of acid burns, flush immediately with water.
6. Wash hands thoroughly after handling flux and solder containing lead. Use lead-free solder whenever possible.
7. Clean up the area when finished. Discard lead and silver solder and dross in a container with a lid. Label the container: “Lead/Silver Solder Waste for Recycling”. Used solder sponges and contaminated rags must be disposed of as hazardous waste.

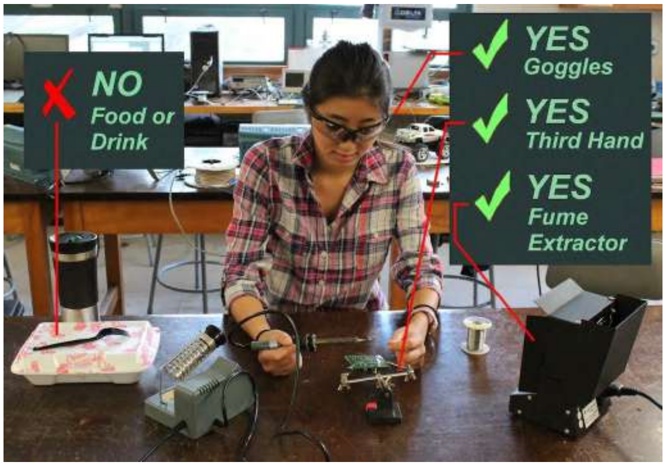

Soldering Techniques

1. Basic tools include an electric soldering station (a temperature-adjustable type preferred), solder, a pair of needle-nose pliers, a wire stripper, a flush cutter, a desoldering bulb, a utility blade, and a third hand or a vice.
2. Workspace should be as uncluttered as possible. Clear the path between the soldering station stand and the project to be soldered.
3. Prepare the soldering iron tip. Scrape off oxides with a utility blade if necessary. Adjust to an appropriate temperature. It won't work if you can't get the solder melt onto the soldering tip. Tinning the soldering tip by coating it with a thin coat of solder. This helps heat transfer between the tip and the components. The cleaning sponge should be soaking wet. Quickly run the soldering tip over the wet sponge to get rid of excess solder.
4. Clean the contact surfaces of the metal parts if necessary. Scrape off oxides with a utility blade. In some cases, it is easier to tin the individual components first before joining them.
5. Position the components using a third hand or a vice. Avoid holding the metal part of a component with your fingers.
6. Timing is important for soldering the components together. The solder needs to be completely melt at the joint to avoid a “cold solder point.” However, too much heat could damage a sensitive electronic component or melt the insulator of a wire. The required time depends on the heat transfer, which is affected by the temperature of the soldering tip, the mass of the components, the presence of flux or rosin, and how clean the contact surfaces are.
7. Start over when an attempt fails. Old solder has impurities. Remove old solder completely with a desoldering bulb or pump. Clean up flux/rosin residuals. Try again.

Examples: www.leadsdirect.co.uk/technical/perfectsolderjoint.html

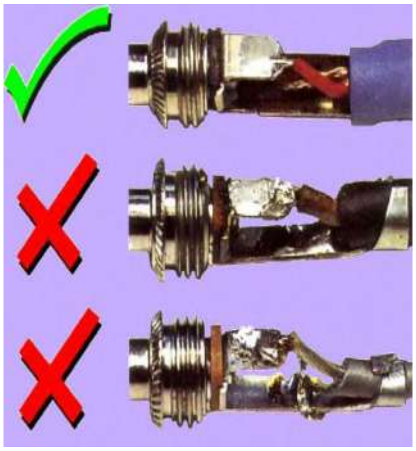

Good – clean, shiny, and just the right amount of solder.

Bad – too much solder which is uneven, has sharp points in places, and was probably overheated.

Worse – uneven texture, and the flux remaining on the joint.

Soldering

WSD 81
350
Weller

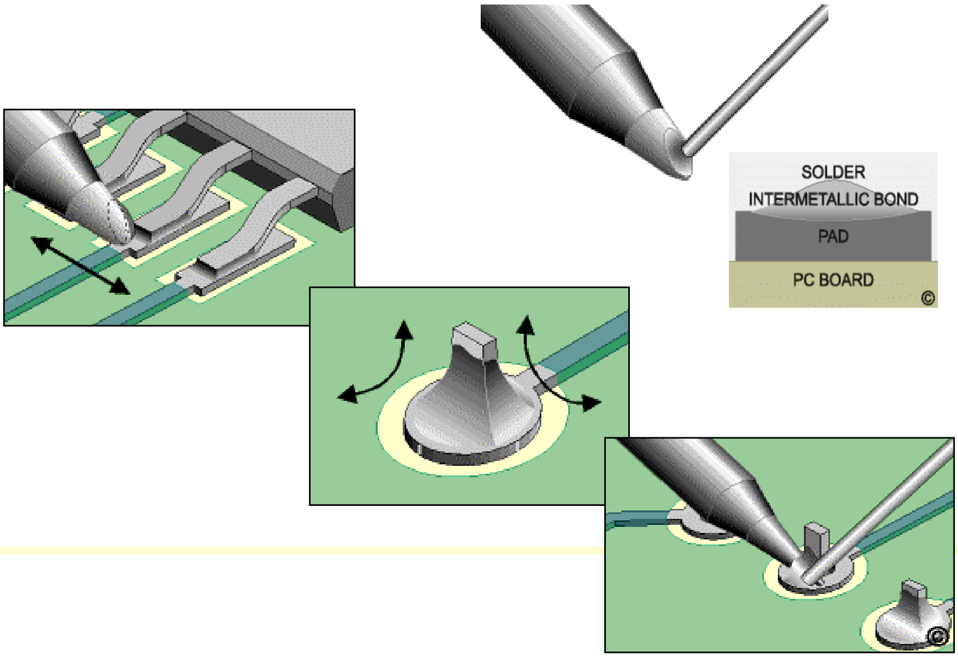
SOLDER
INTERMETALLIC BOND
PAD
PC BOARD
©

TABLE OF CONTENTS

page

1. Introduction 1
2. Basic metals 1
3. Flux 1
4. Solder 2
5. Heating, generation of soldering connection 3
6. Soldering with a soldering iron:
    - a) Cleaning surfaces 3
    - b) Temperature & time 3
    - c) Adding solder & quantity 4
    - d) Soldering : 7 steps 5
    - e) Soldering of a Flat-package IC 6
    - f) Incorrect soldering 7
7. Desoldering with warm air and soldering paste 8+9
8. Lead free soldering 10
9. Temperature specifications of the soldering bit 10
10. Types of soldering bits 11
11. Soldering devices 11

SOLDERING

Theoretical background for technicians

1. Introduction :

Progress within the electronics industry would not have resulted in the mass production of electronic appliances as we know them today, without a similar progress in the field of soldering technology.

To be able to understand how soldering techniques are used in the electronics industry, we first have to get familiar with the materials used during the soldering process, and we have to learn how a soldering connection is created on a molecular level.

Soldering is an old technique for creating permanent electrical and mechanical connections between metals. In contrast to ordinary glueing, which makes only a physical connection between materials (although electrically conducting glue also exists) soldering causes a chemical reaction with other materials, creating a new alloy.

There are several soldering techniques, which all have a few things in common: basic metals, soldering flux, solder and heating.

2. Basic metals :

When we place electronic components on a PCB, the legs of the components and the soldering lands on the PCB, are the basic metals that will connect with the solder.
Several metals like copper, bronze, silver and some types of steel react with solder and make a strong chemical and physical connection.
Other materials like aluminium, titanium, etc. are difficult or even impossible to solder.
The fact that materials exist which can not be soldered is very important, as these materials can be used for making soldering machines and soldering irons. These materials can also be used for covering components during soldering.

There is a direct relationship between how strongly the surface of the basic metal oxidizes and how the soldering reacts on this: When there is more oxidation present, the soldering connection will be weaker.
The fact that most metals oxidize very quickly when heated, is a specific problem, because if one wants to obtain a chemical soldering reaction, a high temperature is necessary…
Flux is used to counteract this oxidation problem.

3. Soldering flux :

Although the soldering surface looks clean, there is always a thin layer of oxide on the metals. This layer always arises as soon as the metal is exposed to air. Reliable soldering connections can only be made on "clean" surfaces. "Clean" surfaces could be accomplished using cleaning product, but this would be insufficient as metal oxidizes incredibly fast when heated. To prevent the creation of the layer of oxide, the use of soldering flux is necessary.

Using soldering flux on the basic metals before the soldering starts has various reasons and advantages:

The main reason of the use of flux is to stop the oxidation of the basic metals during heating. The flux makes sure that the air can't reach the soldering surface, and therefore prevents the formation of oxidation during warming up.
Most of the different flux types have acid components which are used to remove the oxidation that is already present on the basic metal.
If a very strong acid is active in the flux, theoretically it would be possible to remove all the oxides. This however is in practice not possible, as the acid would also affect the PCB and the components, which of course is not the intention.

Therefore there will always be a search for the balance between a flux that contains strong acids, and as a consequence also removes the oxidates more effectively (but also affects the PCB), and a flux that contains a less strong acid and therefore removes less oxides. (but does not affect the PCB).

Most fluxes though use types of acid. Therefore it generally is necessary to remove the flux residues after soldering.

When solder is applied, the flux needs to flow off so that the solder can make direct contact with the basic metal. During this process it is however impossible to avoid the flux mixing with the solder.
Flux developers make use of this fact to develop a flux that has the property to reduce the surface tension of the solder on contact, as a result of which better "wetting" arises.

Flux consists of natural or synthetic resin and chemical additives, called activators.
In PEM a "no clean" flux of Tamura is used (for the placement of Flat-Package-IC's). Consequently it is not necessary to remove the flux after soldering.
Furthermore it is not necessary to add extra flux when doing easy manual soldering because the soldering wire contains flux (3.3 %), like a resin core.

4. Solder :

There several existing metals and metal alloys that can be used for soldering.
The choice for using a certain metal is based on the properties of the metal:
Is the metal flexible or fragile? How well does it conduct heat?
Does it expand rapidly when heated? What is the electrical resistance of the metal?
How strong is it? Is it toxic? And not unimportant: what does it cost?
Although it is not perfect, an alloy of tin and lead is most often used traditionally.
Pure tin melts at 203°C and pure lead melts at 330°C. A mixture of both metals has a lower melting point.
By using a different mixture the melting temperature of the soldering alloy can be changed.
A alloy of tin and lead (e.g. 60/40) has a relatively low melting point (183°C) and can be produced at a relatively low price, in comparison to other alloys with similar properties.
Lead is a very cheap and can be found in a lot of places. Therefore the cost of the tin-lead alloy is mainly determined by the cost of tin.

When this alloy is warmed up, it goes through several phases:
the alloy goes from a "fixed" condition, to a "paste" condition (semi-liquid), and finally becomes totally liquid. It is difficult to solder with an alloy that is in a "paste" condition: because there is little movement or vibration during this "paste" condition, this can cause bad soldering.
If the alloy is of 50% tin and 50 % lead, it starts to melt at a temperature of 193°C, and will be totally melted at a temperature of 216°C.

To prevent bad soldering during the "paste" condition, "eutectical" solder is used.
An "eutectical" alloy is an alloy that, when heated, goes immediately from a fixed condition to a liquid condition: the "paste" stage is effectively skipped.
An "eutectical" tin-lead alloy contains 63% tin and 37% lead.
Often a tin-lead alloy with 60% tin and 40% lead is used. This alloy can be produced cheaper than the eutectical solder because it contains less tin. But because this alloy is almost "eutectical", it can effectively be used for soldering.
During soldering, be careful to make no sudden movements whilst the solder is cooling.
This can cause "interrupted connections" and therefore unreliable soldering.
An "interrupted" soldering connection has a rough and irregular surface and appears dim instead of clear and shiny.
It is also very important that no impurities get into the solder.
If small metal particles from the components get into the soldering "cup" of the soldering machine.
This can result in a serious change to surface tension, and consequently bad soldering connections.
These metal impurities could even change the melting temperature of the solder.

5. Heat, generation of soldering connection

If the solder gets in contact with a copper surface,
the solder dissolves and penetrates the copper layer.
The molecules of the solder and the copper get mixed and a new
alloy is created, consisting partially of copper and partially of solder.
This reaction is called "wetting" and has as a result an "inter-metallic" connection
between the components. "Wetting" can only take place if the surface of the copper does not contain
oxides. The solder and the surface of the copper also have to be at the right temperature before the
"wetting" can take place.

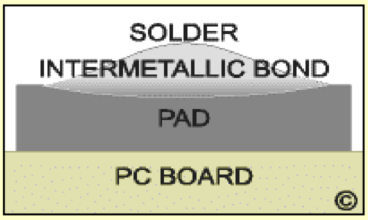

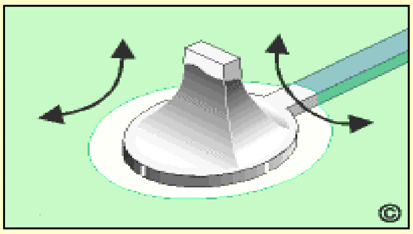

6. Soldering with a soldering iron:

The following 4 factors determine the quality of the soldering connection:

- clean surfaces
- temperature
- time
- quantity of solder

a) Clean surfaces:

"dirt is the enemy of a good soldering connection!"

It is enormously important that there are no impurities on all of the components (including the soldering bit). Soldering on "dirty" components is not possible!
Old PCB's or components sometimes can not be soldered due to a large layer of oxide on the islands and the legs of the components.
In such cases the solder is rejected and goes to everywhere, except for where we want it to go.
In a desperate attempt to solder in such a situation, a common reaction is to raise the soldering temperature. However, this is certainly not a solution. On the contrary!
The only correct solution is thorough cleaning. (e.g. with a brush of glass fibre)

Before the soldering iron can be used, the soldering bit first has to be coated with tin.
Some solder has to be melted onto the bit. Afterwards the bit has to be cleaned.
For cleaning the soldering bit, it is best is to use a slightly moistened sponge (with demineralized water). The use of ordinary tap water is not reccomended, because the silver layer on the bit gets affected by calcium in the tap water, which results in less thermal contact. This diminished thermal contact runs the risk of cold soldering, etc.
To improve the thermal contact between the bit and the soldering surface, it is advisable to coat the soldering bit with a little bit of solder before every soldering action.

b) Temperature & time :

The temperature of the soldering bit is not the most determining factor for successful soldering:
the heating cycle (combination of temperature of the bit and the time the heating takes) is!
How fast the soldering components warm up, how warm they get and how long they stay warm, are the most important factors that have to be controlled to make a reliable soldering connection.

The time soldering takes is very important. A soldering task may only take a couple of seconds: when heated too long, the components and the soldering lands can be damaged.

It is also totally wrong to use the soldering bit to rub melted solder across the soldering surface.

Although the temperature of the bit is not the most determining factor, it is advisable to to do soldering at the lowest possible temperature.

The first that has to be examined before the soldering starts in the 'thermal mass" of the components.

Every soldering task has its own specific "thermal mass" which is determined by the heat absorption of the components. For soldering e.g. the component pins and solder lands.
For components on a heat sink there will be a large thermal mass. In other words: the heat will be "absorbed " by the heat sink, as a result of which more warmth will have to be added than e.g. for soldering the leg of an IC.
Secondly it is very important that the temperature of all the components that will be soldered have more or less the same temperature.
An example: it is necessary that the leg and the soldering land have had at the same time the necessary warming up before the solder is added.
We have to take care that, while warming up, the soldering bit gets in contact with both the leg and the soldering land.
"Thermal contact" is therefore very important.

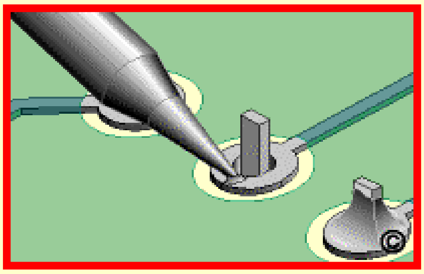

The thermal contact is low
The warmth is transmitted by means of the bit.
to the leg and the soldering land

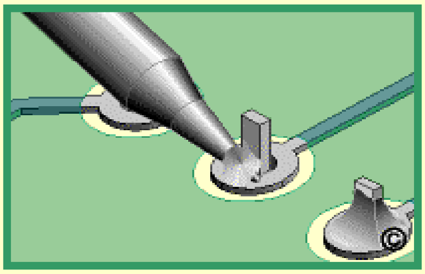

The thermal contact is higher
by adding a bit of solder
to the point of contact

The size and the shape of the soldering bit, as well as the position of the bit, will determine whether or not it makes a good thermal contact. For being able to make a good thermal contact in all circumstances, different types of soldering bits exist (different shapes).
Normally the soldering bit has to be placed on the "largest mass point" of the soldering connection that needs to be made.
This will make fast heating possible.

c) Adding solder & quantity

The last factor necessary for a high quality soldering connection, is adding the correct quantity of solder.
Too much solder is an unnecessary waste and can aid the formation of "bridges" across legs
If not enough solder is added, this can cause an "insufficient soldering" or "dry soldering"
Adding the correct quantity of solder takes a lot of exercise and experience

Melted solder always flows from a "colder" area to a "warmer" area.
Before the solder is added, the surface temperature of the components to be soldered has to be higher than the melting temperature of the solder.
Never let the soldering wire melt by holding it against the soldering bit. This may cause the solder to fall on a surface where the temperature is lower than the melting temperature
This can result in a cold soldering, with almost no or insufficient "wetting".
If the solder is added to a clean, fluxed and sufficiently heated surface, then the solder will melt and flow nicely without direct contact with the heat source. (soldering bit)
Therefore it is advisable to put the soldering wire on the sufficiently heated surface so that the solder melts as a consequence of the contact with the heated surface

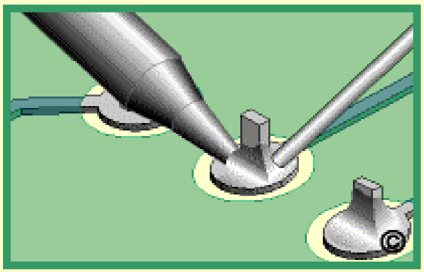

The solder is melted as a consequence of the contact with
the heated surface. No contact with the soldering bit is necessary.

d) Soldering : 7 steps

1. Cleaning of the soldering bit:

For soldering to be effective, the soldering bit has to be clean:
no dirt and no oxides are allowed on the soldering bit.
The soldering bit needs to be cleaned before soldering.

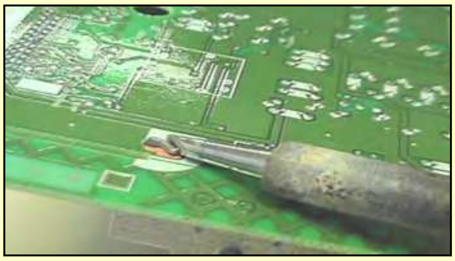

2. Heating of the components that will be soldered.

For soldering to be effective, the components which have to be connected, have to be heated (in this case the pin and the solder lands) to a temperature that at least equals the temperature of the melting point of the soldering alloy. (183 °C)
Bear in mind the "thermal mass", the type of bit and the position of the bit!

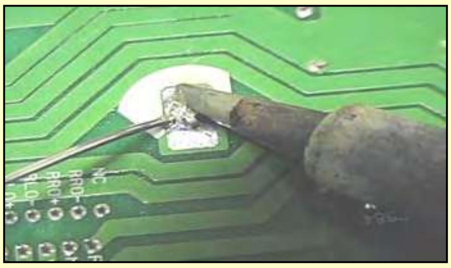

3. Adding the solder:

After both components are heated sufficiently, solder can be added.
Let the solder melt as a consequence of the contact with the heated surface. In other words: don't make any direct contact with the soldering bit.

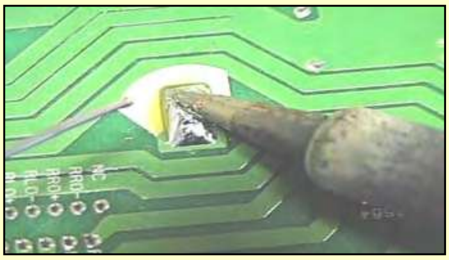

4. Quantity of the solder to be added and letting the solder flow.

Let the solder flow nicely between both components and add solder until the right quantity is reached.

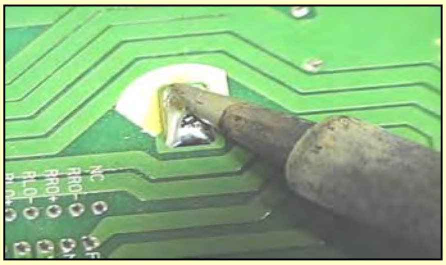

5. Remove the soldering wire

When sufficient solder is added, the soldering wire may be removed.

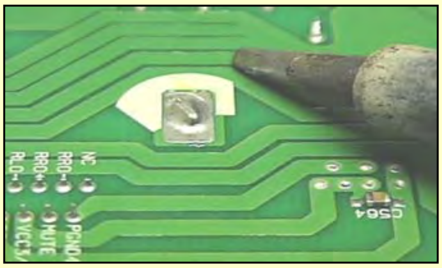

6. Remove the soldering bit.

When the solder has nicely flowed open, and there is "wetting", the soldering bit may be removed so that the solder can solidify.
Pay attention for possible sudden movements during this operation. (risk of "disturbed" soldering)

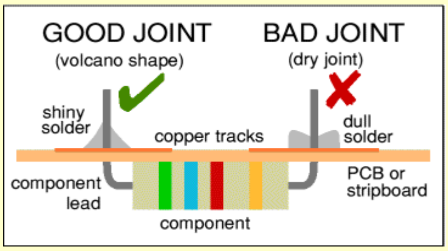

7. Check the soldering connection

Check the soldering connection.
Nicely formed soldering is shining and even. Also pay attention to the shape:
a spherical shape is not good!

e) Soldering of a flatpackage IC

For soldering a "Flat-package IC" (gull wing type) one can use, beside the ordinary point soldering (= leg by leg), a special soldering technique: the DRAGGING TECHNIQUE.
For this technique, one must use a special soldering bit that will, after filling-up, act as a kind of "mini-soldering-wave", and with which one can solder a whole row of legs in one ongoing movement

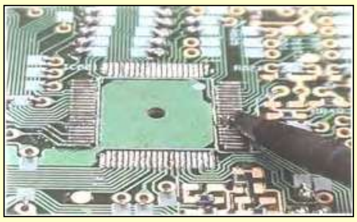

1. Cleaning of the soldering surface:
Remove all possible soldering residues (caused by desoldering) and make all the paths nicely even by going over those paths with the soldering bit. (without adding solder)

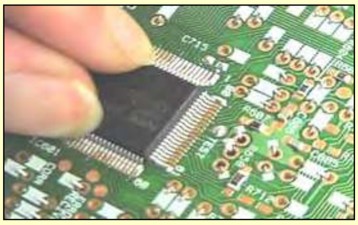

2. Positioning of the IC:
Position the IC on the patterns.
This has to be very precise.
A little movement can cause serious problems.

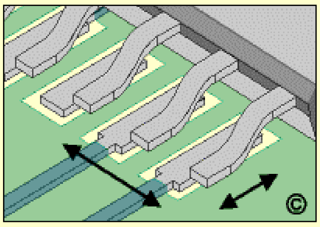

3. Fixing of the IC:
Fix the IC on the PCB, by "temporarily" soldering some legs on both sides.

4. Add sufficient flux on all sides.

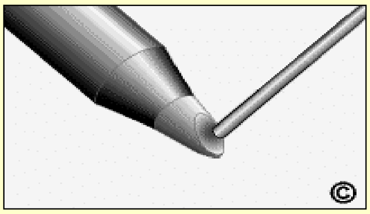

5. Filling of the bit and start of the soldering:
Fill the special (concave) soldering bit with solder.

Now start a slow dragging movement from one leg to another.
Pay attention that the solder nicely moves up on the legs of the IC.

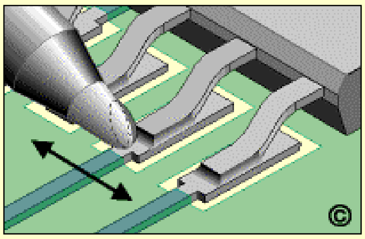

If the PCB is equipped with a "soldering thief" (a soldering land that is positioned next to the last leg), one has to solder in the direction of the soldering thief, so that at the end the possible surplus of solder flows on this soldering thief and no short circuit is created on the last two legs of the row.

f) Incorrect soldering :

For certain soldering problems or soldering mistakes, specific terms are used:

SOLDER BRIDGE

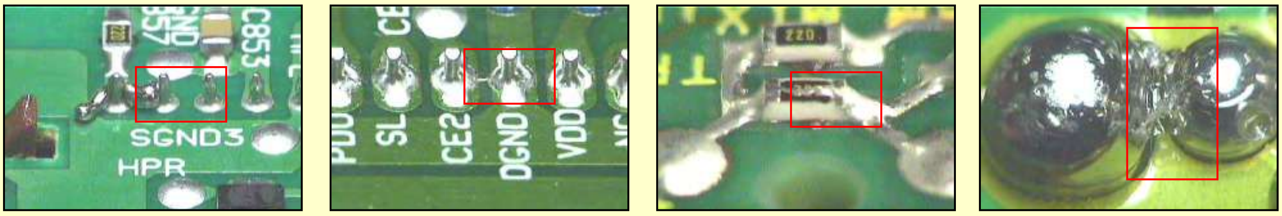
C853
SGND3
HPR
PDU
SL
CE2
DGND
VDD
220

COLD SOLDERING

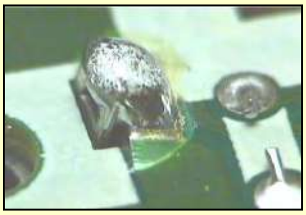

HALF SOLDERING OR DRY SOLDERING

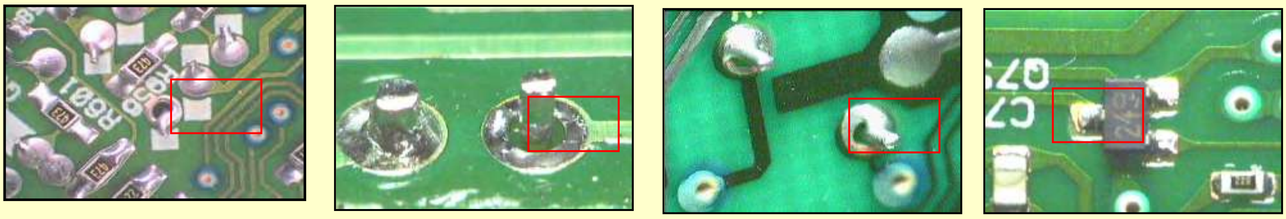

NO SOLDERING

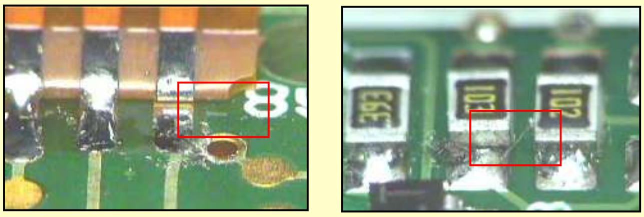

SOLDERING HORN

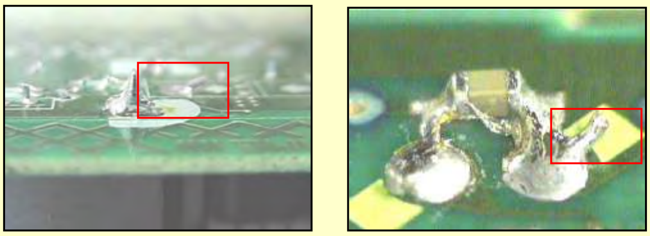

SOLDERING BALL OR SOLDERING SPLASH

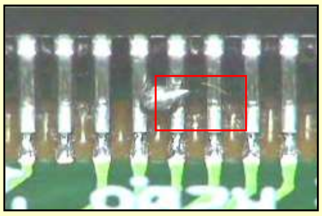

SOLDERING WIRE SHORT CIRCUIT

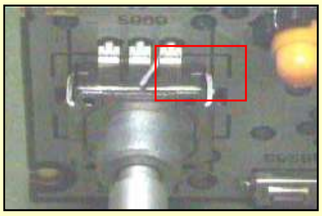

7) (De)soldering with warm air and soldering paste:

In some cases it is impossible to solder or desolder certain components with a classical soldering iron or a normal desoldering station.

This is often the case for SMT components, with contact legs (partly) underneath the component.
In such a case one has to use a (de)soldering appliance that uses warm air or infra-red heat.

SMT component
with contact surfaces
beneath the component

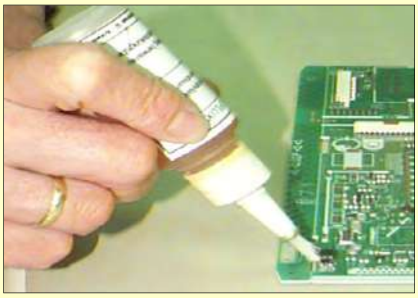

1. Fluxing of the component:

To have a good temperature conductance during warming up, the soldering surfaces need to be moistened with flux.
If these surfaces are parly underneath the component, flux can be put on the borders of the component.

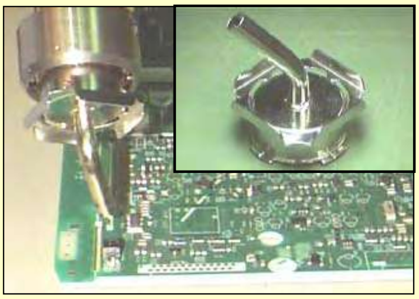

2. Warming-up of the soldered surfaces:

Use a "pointing nozzle" and point it at the component you want to remove. Keep a distance of 1 cm between the nozzle and the surface you want to warm up. Choose the temperature of the warm air e.g. +/- 340°C. (dependant on the temperature profile of the compents.)
Pay attention : if the contact surfaces are partly under the component, it is necessary to warm up the body of the component. Often the component is not useable afterwards.

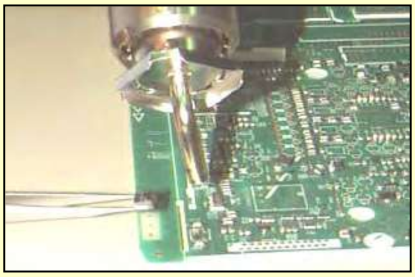

3. Removal of the component

When the solder (paste) has melted, the component can be removed with a pair of tweezers.

4. Cleaning of the soldering surfaces:

Remove all redundant soldering paste with a classical desoldering station.

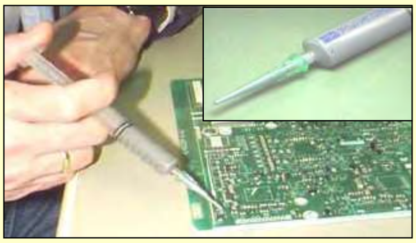

5. Adding soldering paste:

Add with a measured squirt the necessary soldering paste on the soldering lands.
Pay attention: soldering paste needs to be stored in strict circumstances. (see the instructions of the supplier)

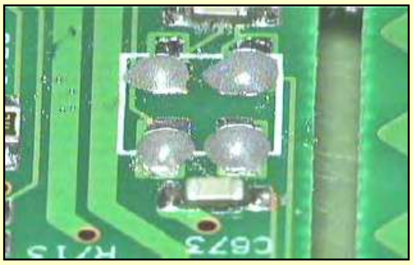

6. Checking the quantity of soldering paste:

Check the quantity of the added soldering paste.
Only the contact surfaces should be covered with soldering paste.
Too much soldering paste can cause a short circuit during "re-flowing".

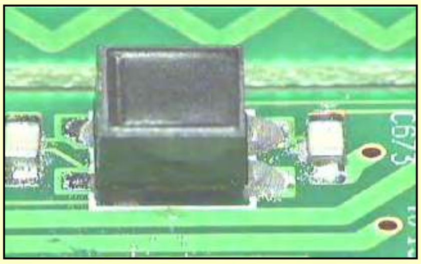

7. Positioning of the component

Position the new component carefully with the contact surfaces in the soldering paste with a pair of tweezers.
Position the component by the guide lines on the PCB.

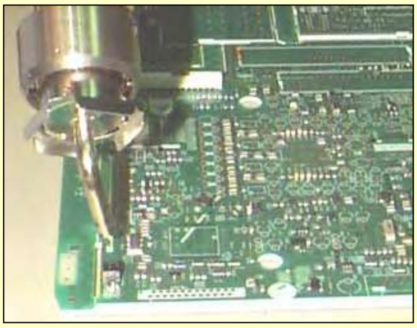

8. Reflowing of the component:

Use a "pointing nozzle" and point it on the component you want to remove. Keep a distance of 1 cm between the nozzle and the surface you want to warm up. Choose the temperature of the warm air e.g. +/- 340°C. (dependant on the temperature profile of the component.) Warm up until one sees that the soldering paste is active or flowing. The result should be a good soldered joint.
Pay attention: For most of the warm air soldering appliances one can choose the air pressure. The speed should be minimum, otherwise there is a possibility of blowing the component away.

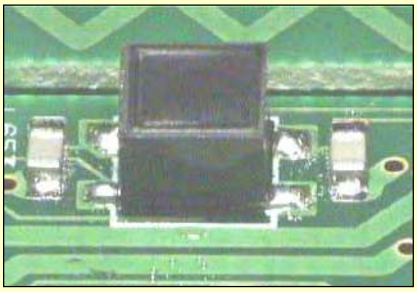

9. Checking of the soldering:

If possible, check the soldered joints.
Good soldering shines and has flowed evenly.

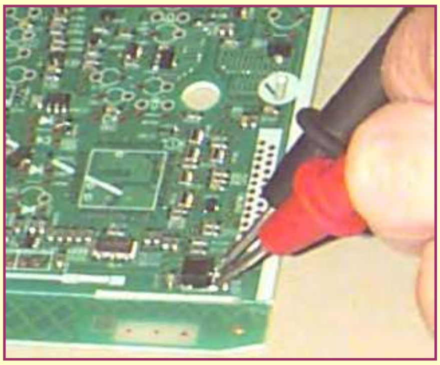

10. Checking of the electrical connection

If possible check the electrical connection with a meter.

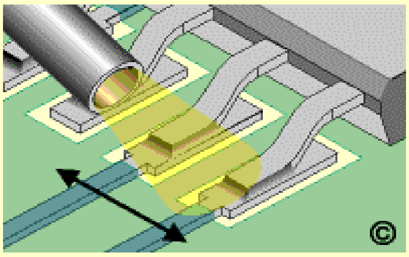

Example of a "warm air soldering appliance": PACE.

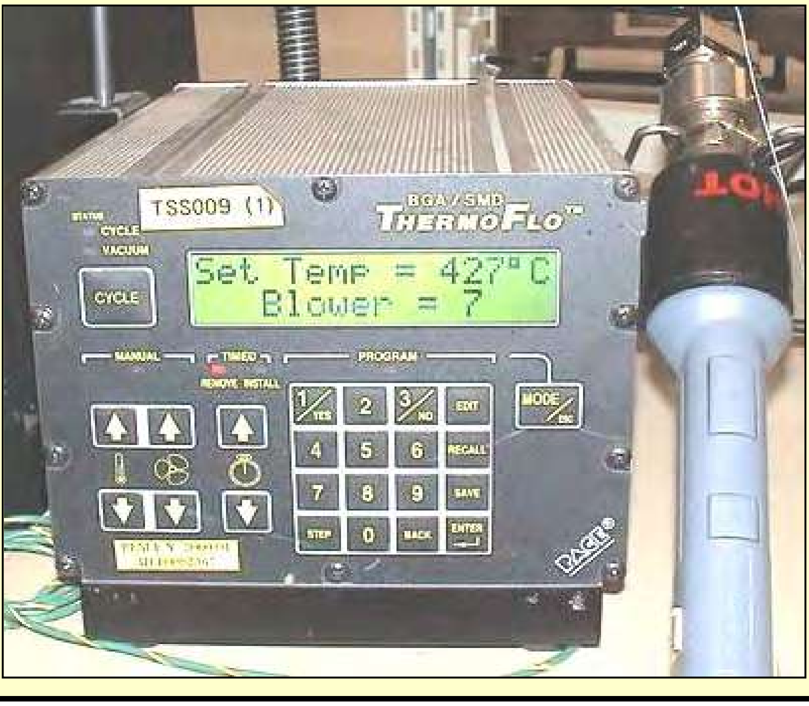
TSS009 (1)
BGA / SMD THERMOFLO
Set Temp = 427°C
Blower = 7
CYCLE
MANUAL
TIMED
PROGRAM
MODE
EDIT
RECALL
SAVE
ENTER
STEP
BACK
DAGE

8) Lead free soldering

Introduction.

Normally all solderings (mechanical - and hand soldering) were executed with soldering material that contained lead. (40 % lead / 60 % tin)

But lead is toxic and therefore damaging for humans and for nature.

Hence European directives were made in which companies will be obliged to use soldering techniques without lead after July 2006.

Soldering without lead is relatively new in the world of electronics.

Lead-free solder:

If no lead is present in the soldering wire, this has to be replaced. The soldering wire needs to contain another metal, like SILVER & COPPER. (96,5 % tin / 3% silver / 0,5 % copper)

This new composition has the advantage that it is no longer toxic and that the soldering connections are stronger. (lead is relatively weak). But this new composition also has some negative consequences:

1) The cost is higher (silver and copper alloys are more expensive than lead)
2) The joints look different (more granular instead of shining)
3) The melting temperature is higher

One of the main points of interest is the higher melting temperature. One would think that as a consequence of the higher melting temperature of the solder, also the soldering temperature needs to be raised. This however is not always allowed! Some components are only able to handle a certain maximum temperature for a certain time.

In practice.

- You can use your original soldering station if you can reach the correct temperature at the soldertip
- The soldering bit has to be a bit larger than a soldering bit for soldering with lead.

Pay attention: **Solder with lead and solder without lead can never be mixed.**

9) Temperature specifications of the soldering bit

Solder with lead

Process	**Bit Type**	**Solder**	**Min**	**Center**	**Max**
Pins	ETB / ETCC/ 1121-0519	UOS-014	340°C	**380°C**	400°C
LCD	ETF / ETCC	UOS-005	320°C	**340°C**	360°C
OEL Display	ETCC	UOS-004	280°C	**290°C**	300°C
Connector	ETL / ETF	UOS-005	320°C	**340°C**	360°C
Flexible PCB	ETF	UOS-004	280°C	**300°C**	320°C
IC's & QFP IC's	1121-0490	UOS-005	280°C	**300°C**	320°C
Removal bridge	ETB / ETCC	UOS-005	320°C	**340°C**	360°C
Lights	ETF / ETB /1121-0519	UOS-005	280°C	**300°C**	320°C
SMT/SMD repair	ETL / ETF	UOS-005	280°C	**300°C**	320°C

Temp. specifications Pioneer standard ZES-A015 & ZES-B016

Solder without lead

Process	**Bit Type**	**Solder**	**Min**	**Center**	**Max**
Pins	LTB / LTCC	UOS-026	400°C	**410°C**	420°C
LCD	LTF / LTCC	UOS-027	320°C	**340°C**	360°C
OEL Display	LTCC	UOS-028	280°C	**300°C**	320°C
Connector	LTL / LTF	UOS-027	320°C	**340°C**	360°C
Flexible PCB	LTF	UOS-028	280°C	**300°C**	320°C
IC's & QFP IC's	1121-0490	UOS-027	280°C	**300°C**	320°C
Removal bridge	LTB / LTCC	UOS-027	320°C	**340°C**	360°C
Lights	LTF / LTB	UOS-027	280°C	**300°C**	320°C
SMT/SMD repair	LTL / LTF	UOS-027	280°C	**300°C**	320°C

Temp. specifications Pioneer standard ZES-A043

Types of soldering wire:

With lead	UOS-014 (1 mm) + UOS-005 (0.8 mm) + UOS-004 (0.5mm)
Without lead	UOS-026 (1 mm) + UOS-027 (0.8 mm) + UOS-028 (0.5 mm)

10) Types of soldering bits:

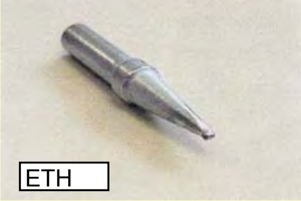
ETH

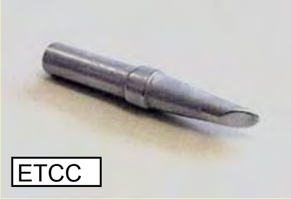
ETCC

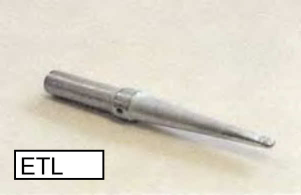
ETL

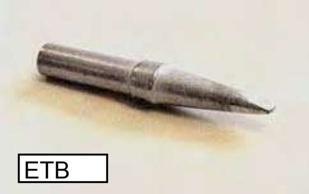
ETB

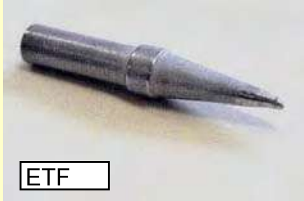
ETF

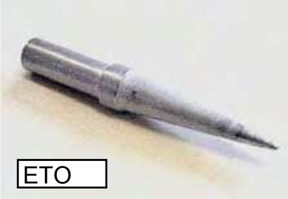
ETO

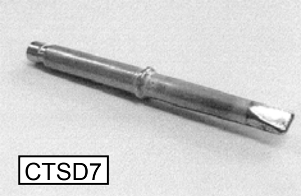
CTSD7

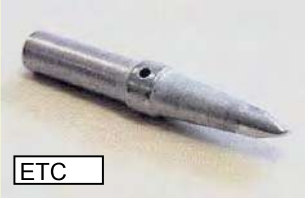
ETC

11) Soldering devices:

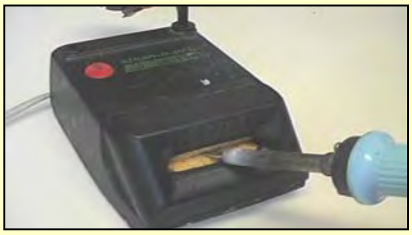
"*Clean-o-point*": *for automatic cleaning of the soldering bit with 2 rotating sponges*

SOLDERING FLU
* Bevat isopropyl
* Licht ontvlamba
"*Flux dosing bottle*": *for easy adding of flux (with brush)*

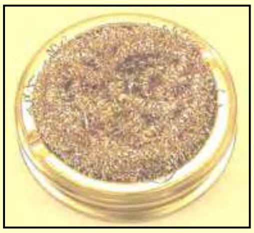
"*Bit cleaner*": *for drying and cleaning of the soldering bit*

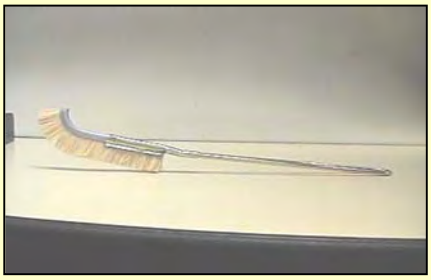
" *Brush* " : *for removal of possible soldering residues after soldering*

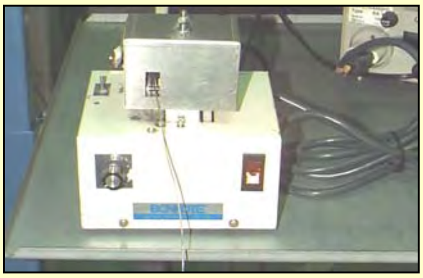
"*V-cutter*" : *for making a cut in the soldering wire, in order to reduce the formation of little soldering balls*

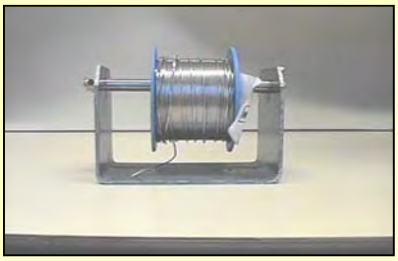
"*Solder dispenser*" : *for a simple supply of soldering wire*

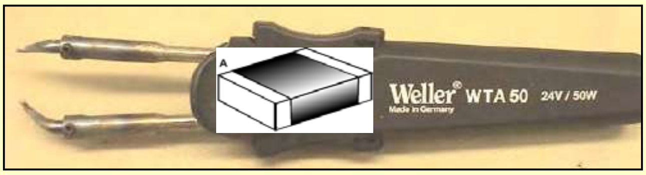
A
Weller WTA 50 24V / 50W
Made in Germany
"*Tweezer-copper bit*": *handy tool for the removal of SMD/SMT components.*

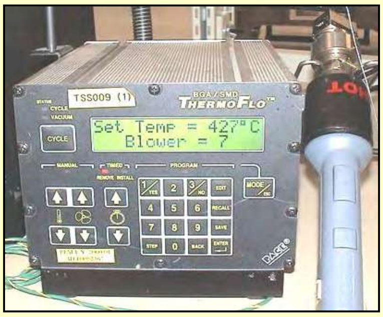
TSS009 (1)
THERMOFLO
Set Temp = 427°C
Blower = 7
"*Warm air desoldering appliance*": *for easy desoldering of flat-package IC's*

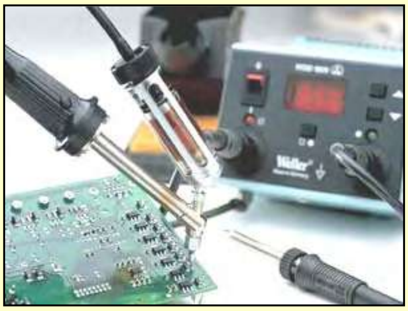
"*Desoldering appliance*": *for easy desoldering of point solderings*

"*Soldering magnifying glass*": *for a good view during soldering*

The Basic Soldering Guide

http://www.epemag.wimborne.co.uk/solderfaq.htm (for text in this document)

basics for wire (Good intro) : http://www.youtube.com/watch?v=BLfXXRfRIzY
The Happy Soldering Iron: http://www.youtube.com/watch?v=AL-_RGbyf1s
basics: http://www.youtube.com/watch?v=I_NU2ruzyc4
Watt/temp comparison http://www.youtube.com/watch?v=Vh9pWu6K6tc&NR=1
tip sizes: http://www.youtube.com/watch?v=Sfb1Ve52ztY&NR=1
types of solder: http://www.youtube.com/watch?v=COqGkYMOA44
maintenance: http://www.youtube.com/watch?v=krxTfZCFptk&NR=1

Good overview with some odd bits: http://www.youtube.com/watch?v=AOdnGUMi7lQ

Desoldering overview: http://www.youtube.com/watch?v=j-_pnc-Qqm8
2nd good overview: http://www.youtube.com/watch?v=8UN3D2-f64A
Hot air pencil/gun: http://www.youtube.com/watch?v=AxYhF6Ab2CU
Desoldering gun (short): http://www.youtube.com/watch?v=8Z6MvZz_uNc

Crazy Guy from Make Electronics: http://www.youtube.com/watch?v=3N3ApzmyjzE
Not Soldering!! – Bread board: http://www.youtube.com/watch?v=oiqNaSPTI7w

This written guide will help beginners and novices to obtain effective results when soldering **electronic components**. If you have little or no experience of using a soldering iron, then EPE *(Everyday Practical Electronics* magazine) recommends that you practice your soldering technique on some fresh surplus components and clean stripboard (protoboard), before experimenting with a proper constructional project. This will help you to avoid the risk of disappointment when you start to assemble your first prototypes. If you've never soldered before, then read on!

Soldering irons

The most fundamental skill needed to assemble any electronic project is that of ***soldering***. It takes some practice to make the perfect joint, but, like riding a bicycle, once learned is never forgotten! The idea is simple: to join electrical parts together to form an electrical connection, using a molten mixture of lead and tin (solder*) with a soldering iron. A large range of soldering irons is available - which one is suitable for you depends on your budget and how serious your interest in electronics is.

[*Note: the use of lead in solder is now increasingly prohibited in many countries. "Lead free" solder is now statutory instead.]

Electronics catalogues often include a selection of well-known brands of soldering iron. Excellent British-made ones include the universally popular Antex, Adcola and Litesold makes. Other popular brands include those made by Weller and Ungar. A very basic mains electric soldering iron can cost from under £5 (US$ 8), but expect a reasonable model to be approximately £10-£12 (US$ 16 - 20) - though it's possible to spend into three figures on a soldering iron "station" if you're really serious! Check some suppliers' catalogues for some typical types. Certain factors you need to bear in mind include:-

Voltage: most irons run from the mains at 240V. However, low voltage types (e.g. 12V or 24V) generally form part of a "soldering station" and are designed to be used with a special controller made by the same manufacturer.

Wattage: Typically, they may have a power rating of between 15-25 watts or so, which is fine for most work. A higher wattage does ***not*** mean that the iron runs hotter - it simply means that there is more power in reserve for coping with larger joints. This also depends partly on the design of the "bit" (the tip of the iron). Consider a higher wattage iron simply as being more "unstoppable" when it comes to heavier-duty work, because it won't cool down so quickly.

Temperature Control: the simplest and cheapest types don't have any form of temperature regulation. Simply plug them in and switch them on! Thermal regulation is "designed in" (by physics, not electronics!): they may be described as "thermally balanced" so that they have some degree of temperature "matching" but their output will otherwise not be controlled. Unregulated irons form an ideal general purpose iron for most users, and they generally cope well with printed circuit board soldering and general interwiring. Most of these "miniature" types of iron will be of little use when attempting to solder large joints (e.g. very large terminals or very thick wires) because the component being soldered will "sink" heat away from the tip of the iron, cooling it down too much. (This is where a higher wattage comes in useful.)

A proper temperature-controlled iron will be quite a lot more expensive - retailing at say £40 (US$ 60) or more - and will have some form of built-in thermostatic control, to ensure that the temperature of the bit (the tip of the iron) is maintained at a fixed level (within limits). This is desirable especially during more frequent use, since it helps to ensure that the temperature does not "overshoot" in between times, and also guarantees that the output will be relatively stable. Some irons have a bimetallic strip thermostat built into the handle which gives an audible "click" in use: other types use all-electronic controllers, and some may be adjustable using a screwdriver.

Yet more expensive still, ***soldering stations*** cost from £70 (US$ 115) upwards (the iron may be sold separately, so you can pick the type you prefer), and consist of a complete bench-top control unit into which a special *low-voltage* soldering iron is plugged. Some versions might have a built-in digital temperature readout, and will have a control knob to enable you to vary the setting. The temperature could be boosted for soldering larger joints, for example, or for using higher melting-point solders (e.g. silver solder). These are designed for the most discerning users, or for continuous production line/ professional use. The best stations have irons which are well balanced, with comfort-grip handles which remain cool all day. A thermocouple will be built into the tip or shaft, which monitors temperature.

Anti-static protection: if you're interested in soldering a lot of static-sensitive parts (e.g. CMOS chips or MOSFET transistors), more advanced and expensive soldering iron stations use static-dissipative materials in their construction to ensure that static does not build up on the iron itself. You may see these listed as "ESD safe" *(electrostatic discharge proof)*. The cheapest irons won't necessarily be ESD-safe but never the less will still probably perform perfectly well in most hobby or educational applications, if you take the usual anti-static precautions when handling the components. The tip would need to be well earthed (grounded) in these circumstances.

Bits: it's useful to have a small selection of manufacturer's bits (soldering iron tips) available with different diameters or shapes, which can be changed depending on the type of work in hand. You'll probably find that you become accustomed to, and work best with, a particular shape of tip. Often, tips are *iron-coated* to preserve their life, or they may be bright-plated instead. Copper tips are seldom seen these days.

Spare parts: it's nice to know that spare parts may be available, so if the element blows, you don't need to replace the entire iron. This is especially so with expensive irons. Check through some of the larger mail-order catalogues.

A ***solder gun*** is a pistol-shaped iron, typically running at 100W or more, and is completely unsuitable for soldering modern electronic components: they're too hot, heavy and unwieldy for micro-electronics use. Plumbing, maybe..!

Soldering irons are best used along with a heat-resistant ***bench-type holder***, so that the hot iron can be safely parked in between use. Soldering stations already have this feature, otherwise a separate soldering iron stand is essential, preferably one with a holder for tip-cleaning sponges. Now let's look at how to use soldering irons properly, and how to put things right when a joint goes wrong.

The Basic Soldering Guide Photo Gallery

Soldering is a delicate manual skill which only comes with practice. Remember that your ability to solder effectively will determine directly how well the prototype or product functions during its lifespan. Poor soldering can be an expensive business - causing product failure and downtime, engineer's maintenance time and customer dissatisfaction. At hobbyist level, bad soldering technique can be a cause of major disappointment which damages your confidence. It needn't be like that: soldering is really easy to learn, and like learning to ride a bike, once mastered is never forgotten!

These photos illustrate the basic steps in making a perfect solder joint on a p.c.b. If you're a beginner, our advice is that it's best to practice your soldering technique using some clean, new parts with perhaps some new stripboard (protoboard). Be sure to avoid using old, dirty parts; these can be difficult if not impossible to solder.

(Left) Printed circuit board copper tracks must be clean to begin with, especially if they're not previously "tinned" with solder. Clean any raw p.c.b. copper tracks gently with e.g. an abrasive rubber block available from electronics suppliers. **(Right)** Clean the iron "bit" (soldering iron tip) using a damp sponge. The soldering iron featured is an Ungar Concept 2100 Soldering Station. Other popular brands of soldering equipment include Weller and Antex.

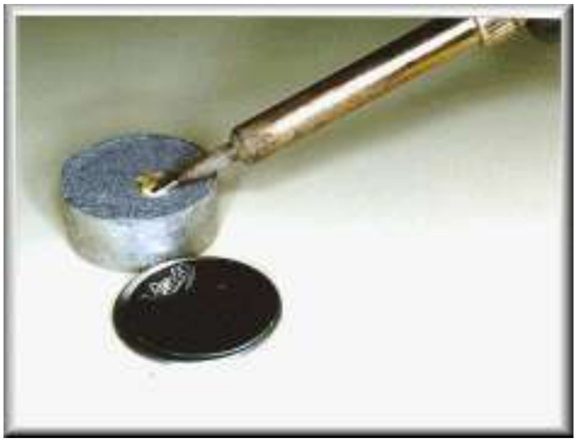

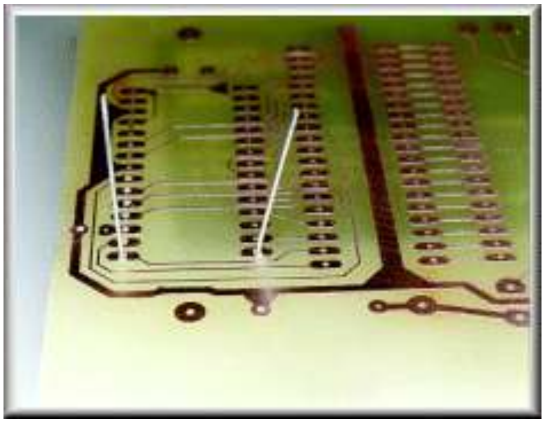

(Left) A useful product is Multicore's Tip Tinner Cleaner (TTC) - a 15 gramme tin of special paste which cleans and "tins" the soldering iron iron, in one go. New tips must be tinned **immediately** when used for the first time. **(Right)** Insert components and splay the leads so that the part is held in place.

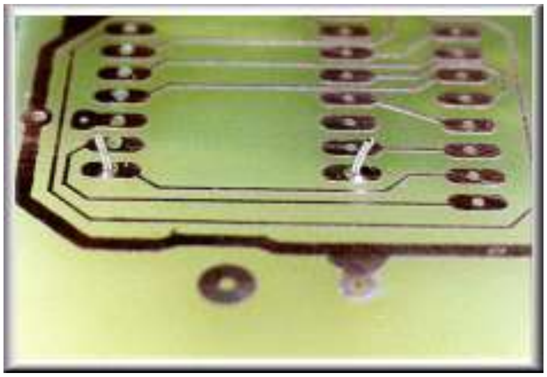
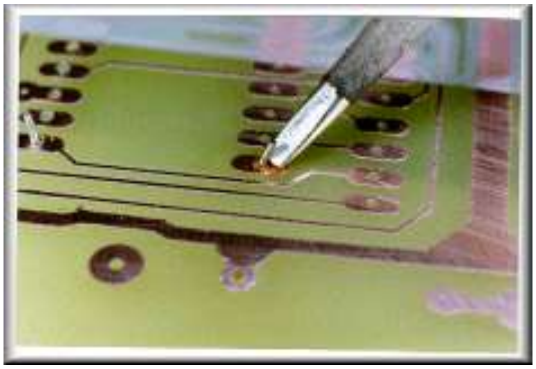

(Left) It's usually best to snip the electronic component wires to length prior to soldering. This helps prevent transmitting mechanical shocks to the copper foil.**(Right)** Apply a clean soldering iron tip to the copper solder pad and the component lead, in order to heat *both* items at the same time.

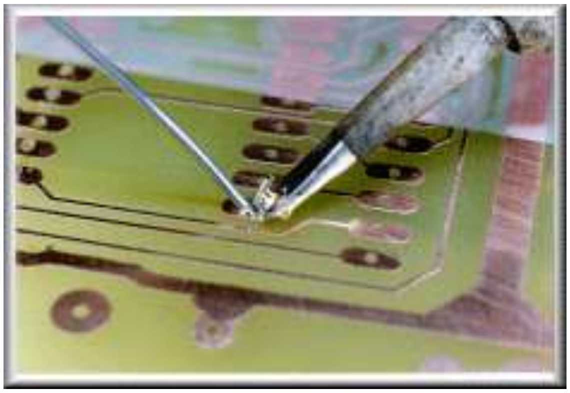
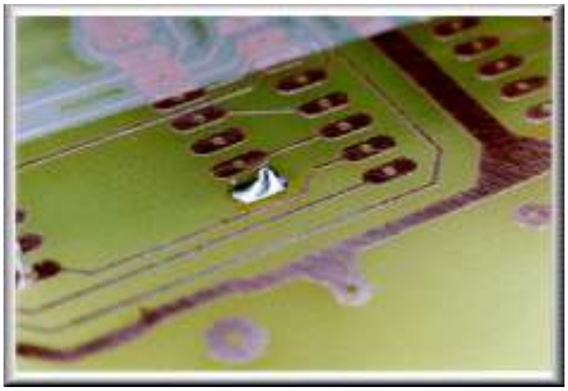

(Left) Continue heating and apply a few millimetres of solder. Remove the iron and allow the solder joint to cool naturally. **(Right)** It only takes a second or two, to make the perfect joint, which should be nice and shiny. Check the Guide for troubleshooting help.

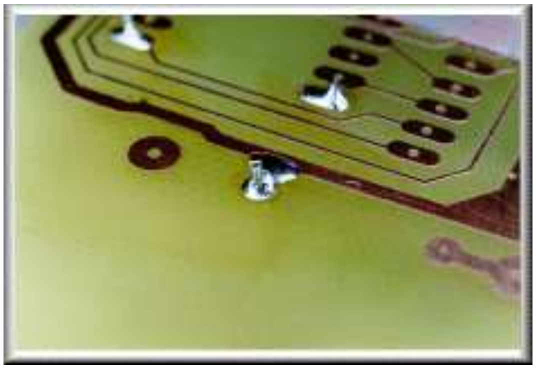

An example of a "dry" or "gray" soller joint - the solder failed to flow, and instead beaded to form globules around the wire.

How to solder

Turning to the actual techniques of soldering, firstly it's best to *secure the work* somehow so that it doesn't move during soldering and affect your accuracy. In the case of a printed circuit board, various holding frames are fairly popular especially with densely populated boards: the idea is to insert all the parts on one side ("stuffing the board"), hold them in place with a special foam pad to prevent them falling out, turn the board over and then snip off the wires with cutters before making the joints. The frame saves an awful lot of turning the board over and over, especially with large boards. Other parts could be held firm in a modeller's small vice, for example.

Solder joints may need to possess some degree of mechanical strength in some cases, especially with wires soldered to, say, potentiometer or switch tags, and this means that the wire should be looped through the tag and secured before solder is applied. The down side is that it is more difficult to *de-solder* the joint (see later) and remove the wire afterwards, if required. Otherwise, in the case of an ordinary circuit board, components' wires are bent to fit through the board, inserted flush against the board's surface, splayed outwards a little so that the part grips the board, and then soldered.

In my view - opinions vary - it's generally better to snip the surplus wires leads off *first*, to make the joint more accessible and avoid applying a mechanical shock to the p.c.b. joint. However, in the case of semiconductors, I often tend to leave the snipping until *after* the joint has been made, since the excess wire will help to sink away some of the heat from the semiconductor junction. Integrated circuits can either be soldered directly into place if you are confident enough, or better, use a dual-in-line socket to prevent heat damage. The chip can then be swapped out if needed.

Parts which become hot in operation (e.g. some resistors), are raised above the board slightly to allow air to circulate. Some components, especially large electrolytic capacitors, may require a mounting clip to be screwed down to the board first, otherwise the part may eventually break off due to vibration.

The perfectly soldered joint will be nice and shiny looking, and will prove reliable in service. I would say that:

- cleanliness
- temperature
- time
- adequate solder coverage

are the key factors affecting the quality of the joint. A little effort spent now in soldering the perfect joint may save you - or somebody else - a considerable amount of time in troubleshooting a defective joint in the future. The basic principles are as follows.

Really Clean

First, and without exception, all parts - including the iron tip itself - must be clean and free from contamination. Solder just will not "take" to dirty parts! Old components or copper board can be notoriously difficult to solder, because of the layer of oxidation which builds up on the surface of the leads. This repels the molten solder and this will soon be evident because the solder will "bead" into globules, going everywhere except where you need it. Dirt is the enemy of a good quality soldered joint!

Hence, it is an absolute necessity to ensure that parts are free from grease, oxidation and other contamination. In the case of old resistors or capacitors, for example, where the leads have started to oxidise, use a small hand-held file or perhaps scrape a knife blade or rub a fine emery cloth over them to reveal fresh metal underneath. Stripboard and copper printed circuit board will generally oxidise after a few months, especially if it has been fingerprinted, and the copper strips can be cleaned using an abrasive rubber block, like an aggressive eraser, to reveal fresh shiny copper underneath.

Also available is a fibre-glass filament brush, which is used propelling-pencil-like to remove any surface contamination. These tend to produce tiny particles which are highly irritating to skin, so avoid accidental contact with any debris. Afterwards, a wipe with a rag soaked in cleaning solvent will remove most grease marks and fingerprints. After preparing the surfaces, avoid touching the parts afterwards if at all possible.

Another side effect of having dirty surfaces is the tendency for people to want to apply *more heat* in an attempt to "force the solder to take". This will often do more harm than good because it may not be possible to burn off any contaminants anyway, and the component may be overheated. In the case of semiconductors, temperature is quite critical and they may be harmed by applying such excessive heat.

Before using the iron to make a joint, it should be "tinned" (coated with solder) by applying a few millimetres of solder, then wiped on a damp sponge preparing it for use: you should always do this immediately with a new bit, anyway. Personally, I always re-apply a very small amount of solder again, mainly to improve the thermal contact between the iron and the joint, so that the solder will flow more quickly and easily. It's sometimes better to tin larger parts as well before making the joint itself, but it isn't generally necessary with p.c.b. work. (All *EPE* printed circuit boards are "roller-tinned" to preserve their quality and to help with soldering.) A worthwhile product is Weller's *Tip Tinner & Cleaner*, a small 15 gram tinlet of paste onto which you dab a hot iron - the product cleans and tins the iron ready for use. An equivalent is Adcola *Tip-Save*.

Normal electronics grade solder is now "lead free" and typically contains Sn 97 Ag 2.5 Cu 0.5 (i.e. 97% tin, 2.5% silver and 0.5% copper). **It already contains cores of "flux"** which helps the molten solder to flow more easily over the joint. Flux removes oxides which arise during heating, and is seen as a brown fluid bubbling away on the joint. The use of separate acid flux paste (e.g. as used by plumbers) should NEVER be necessary in normal electronics applications because electronics-grade solder already contains the correct grade of flux! Other solders are available for specialist work, including aluminium and silver-solder. Different solder diameters are produced, too; 20-22 SWG (19-21 AWG) is 0.91-0.71mm diameter and is fine for most work. Choose 18 SWG (16 AWG) for larger joints requiring more solder.

Temperature

Another step to successful soldering is to ensure that the **temperature** of *all* the parts is raised to roughly the same level before applying solder. Imagine, for instance, trying to solder a resistor into place on a printed circuit board: it's far better to heat *both* the copper p.c.b. *and* the resistor lead at the same time before applying solder, so that the solder will flow much more readily over the joint. Heating one part but not the other is far less satisfactory joint, so strive to ensure that the iron is in contact with *all* the components first, before touching the solder to it. The melting point of most solder is in the region of 188°C (370°F) and the iron tip temperature is typically 330-350°C (626°-662°F). The latest lead-free solders typically require a higher temperature.

Now is the time

Next, the joint should be heated with the bit for just the right amount of time - during which a short length of solder is applied to the joint. Do not use the iron to carry molten solder over to the joint! Excessive time will damage the component and perhaps the circuit board copper foil too! Heat the joint with the tip of the iron, then continue heating whilst applying solder, then remove the iron and allow the joint to cool. This should take only a few seconds, with experience. The heating period depends on the temperature of your iron and size of the joint - and larger parts need more heat than smaller ones - but some parts (semiconductor diodes, transistors and i.c.s), are sensitive to heat and should not be heated for more than a few seconds. Novices sometimes buy a small clip-on heat-shunt, which resembles a pair of aluminium tweezers. In the case of, say, a transistor, the shunt is attached to one of the leads near to the transistor's body. Any excess heat then diverts up the heat shunt instead of into the transistor junction, thereby saving the device from over-heating. Beginners find them reassuring until they've gained more experience.

Solder Coverage

The final key to a successful solder joint is to apply an appropriate amount of solder. Too much solder is an unnecessary waste and may cause short circuits with adjacent joints. Too little and it may not support the component properly, or may not fully form a working joint. How much to apply, only really comes with practice. A few millimetres only, is enough for an "average" p.c.b. joint, (if there is such a thing).

Here's a summary of how to make the perfect solder joint.

1. All parts must be clean and free from dirt and grease.
2. Try to secure the work firmly.
3. "Tin" the iron tip with a small amount of solder. Do this immediately, with new tips being used for the first time.
4. Clean the tip of the hot soldering iron on a damp sponge.
5. Many people then add a tiny amount of fresh solder to the cleansed tip.

6. Heat all parts of the joint with the iron for under a second or so.
7. Continue heating, then apply sufficient solder only, to form an adequate joint.
8. Remove and return the iron safely to its stand.
9. It only takes two or three seconds at most, to solder the average p.c.b. joint.
10. Do not move parts until the solder has cooled.

Troubleshooting Guide

- Solder won't "take" - grease or dirt present - desolder and clean up the parts. Or, material may not be suitable for soldering with lead/tin solder (eg aluminium).
- Joint is crystalline or grainy-looking - has been moved before being allowed to cool, or joint was not heated adequately - too small an iron/ too large a joint.
- Solder joint forms a "spike" - probably overheated, burning away the flux.

Desoldering methods

A soldered joint which is improperly made will be electrically "noisy", unreliable and is likely to get worse in time. It may even not have made any electrical connection at all, or could work initially and then cause the equipment to fail at a later date! It can be hard to judge the quality of a solder joint purely by appearances, because you cannot say how the joint actually formed on the *inside*, but by following the guidelines there is no reason why you should not obtain perfect results.

A joint which is poorly formed is often called a "dry joint". Usually it results from dirt or grease preventing the solder from melting onto the parts properly, and is often noticeable because of the tendency of the solder not to "spread" but to form beads or globules instead, perhaps partially. Alternatively, if it seems to take an inordinately long time for the solder to spread, this is another sign of possible dirt and that the joint may potentially be a dry one.

There will undoubtedly come a time when you need to *remove* the solder from a joint: possibly to replace a faulty component or fix a dry joint. The usual way is to use a ***desoldering pump*** or vacuum pump which works like a small spring-loaded bicycle pump, only in reverse! (More demanding users using CMOS devices might need a pump which is ESD safe.) A spring-loaded plunger is released at the push of a button and the molten solder is then drawn up into the pump. It may take one or two attempts to clean up a joint this way, but a small desoldering pump is an invaluable tool especially for p.c.b. work.

Sometimes, it's effective to actually *add more* solder and then desolder the whole lot with a pump, if the solder is particularly awkward to remove. Care is needed, though, to ensure that the boards and parts are not damaged by excessive heat; the pumps themselves have a P.T.F.E. nozzle which is heat proof but may need replacing occasionally.

An excellent alternative to a pump is to use *desoldering braid*, including the famous American "Soder-Wick" (sic) or Adcola "TISA-Wick" which are packaged in small dispenser reels. This product is a specially treated fine copper braid which draws molten solder up into the braid where it solidifies. The best way is to use the tip of the hot iron to press a short length of braid down onto the joint to be de-soldered. The iron will subsequently melt the solder, which will be drawn up into the braid. Take extreme care to ensure that you don't allow the solder to cool with the braid adhering to the work, or you run the risk of damaging p.c.b. copper tracks when you attempt to pull the braid off the joint. See my photo gallery for more details.

I recommend buying a small reel of de-soldering braid, especially for larger or difficult joints which would take several attempts with a pump. It is surprisingly effective, especially on difficult joints where a desoldering pump may prove a struggle.

The Basic De-soldering Guide Photo Gallery (and Black Museum of Bad Soldering)

De-soldering is required when electronic components need to be removed from a circuit, usually because they are faulty. It may sometimes be necessary during testing or assembly, if a wrong part has been fitted or a modification has to be made. In the field, it's not uncommon for faulty electronic components to be swapped out, or poor joints (perhaps "dry" or gray joints) to need re-making properly, months or years after manufacture. Experienced engineers can often diagnose a particular faulty joint immediately, because they may have seen the same problem on similar electronic equipment before, especially if the equipment has a "reputation". A proper desoldering technique can soon be acquired with practice - all you need to do is buy some scrap boards to have a go with, and desolder to your heart's content!

The next photo sequence illustrates the basic steps for desoldering a printed circuit board, in order to remove a faulty part. Both the technique for using a desoldering suction pump as well as desolder braid are illustrated. Some real-life examples of poor soldering are shown too, in my *Black Museum of Bad Soldering*!

Remember - it costs just as much to get it right as it does to get it wrong! Practice makes perfect.

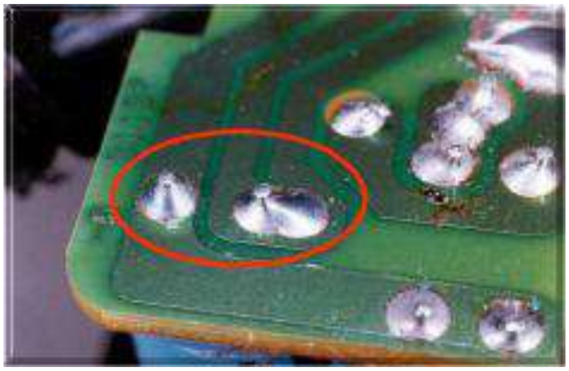

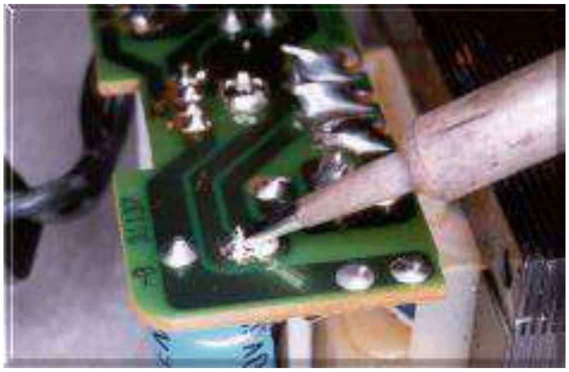

(Left) The two solder joints to be desoldered, to enable a faulty electrolytic capacitor to be removed from the printed circuit board.

(Right) If using a suction-type desoldering pump, apply the soldering iron tip first to melt the solder joint (say for 1-2 seconds). Ensure the spring-loaded desoldering pump is 'primed' and ready to go...

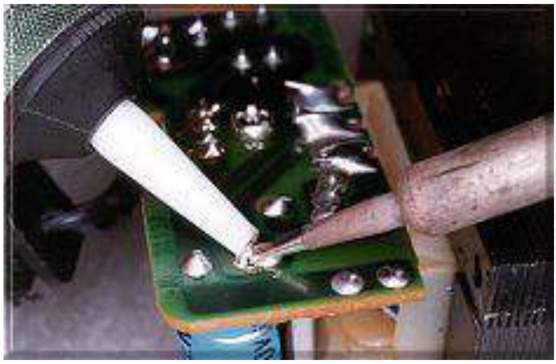

(Left) The PTFE nozzle of the desoldering pump is applied to the molten solder and the spring-loaded plunger is then immediately released, drawing the molten solder up into the pump. Remove the soldering iron tip. Repeat the process if needed. *Handy tip*: sometimes it helps to add some fresh solder and then desolder the whole joint.

(Right) The first p.c.b. joint, now desoldered. The second joint will be desoldered using traditional desoldering braid.

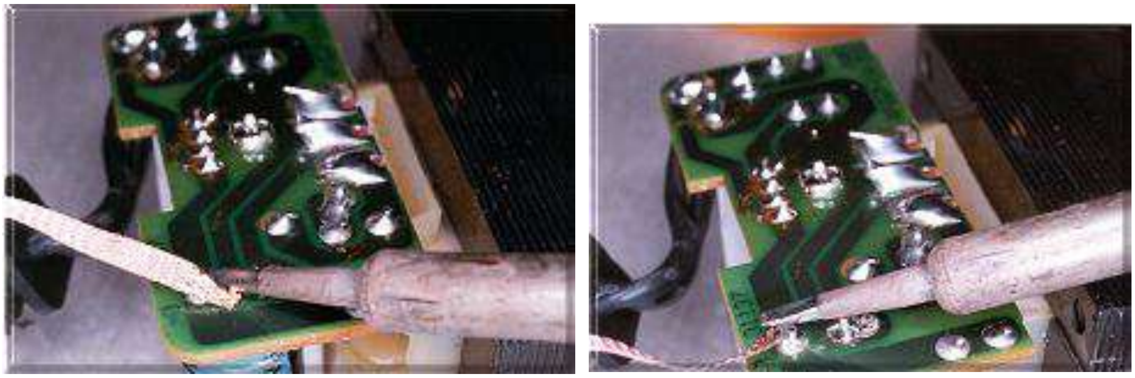

(Left) Select a suitable width of desoldering braid, and **press it down** onto the **COLD** joint using the hot tip of the iron. A flat soldering iron bit is preferable.

(Right) The molten solder is drawn up by capillary action into the desoldering braid. Take care not to overheat the board (the p.c.b. copper track may lift off), nor 'drag whiskers' of solder over the board, nor let the braid solidify onto the joint! Remove the braid while the joint is still molten.

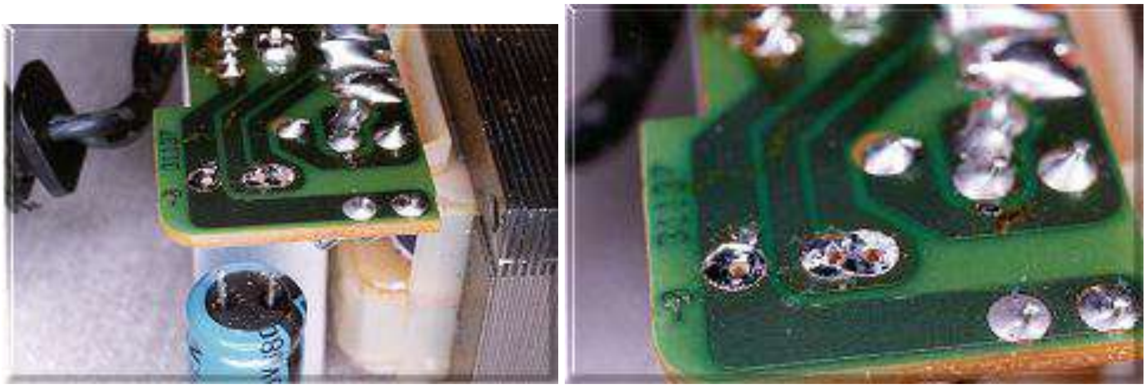

(Left) The faulty electrolytic capacitor dropped out of the board after desoldering. Sometimes, it may need persuading with pliers.... but don't overdo this or you risk damaging the copper tracks on the p.c.b.

(Right) Close-up photograph of both joints, now desoldered and ready for the replacement component to be fitted.

The Black Museum of Bad Soldering

These are all genuine examples which have not been retouched or reworked in any way.

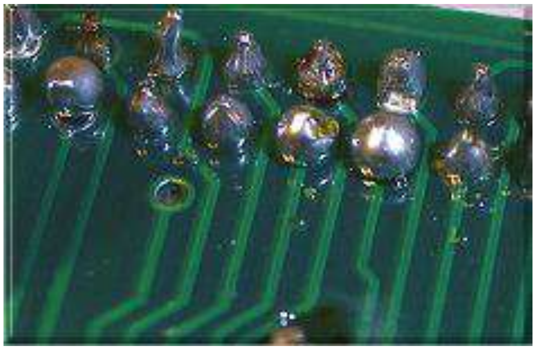
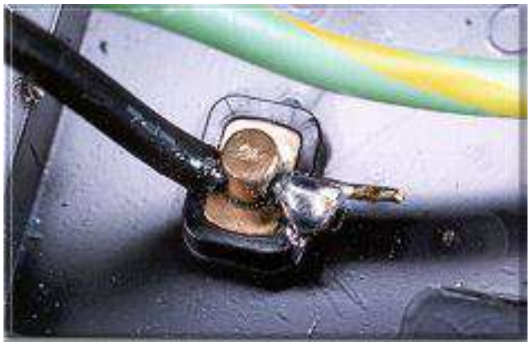

(Left) A tenfold excess of solder on this hand-soldered printed circuit board, and (extreme left) an incomplete solder joint with poor coverage. There is no need to add more solder "for luck".

(Right) An example of a dry (or gray) solder joint found inside a commercial PSU for a computer peripheral. The wire had been fed through the hole in the brass terminal, and merely tacked on with a blob of solder. This is a fire hazard (risk of arcing and overheating).

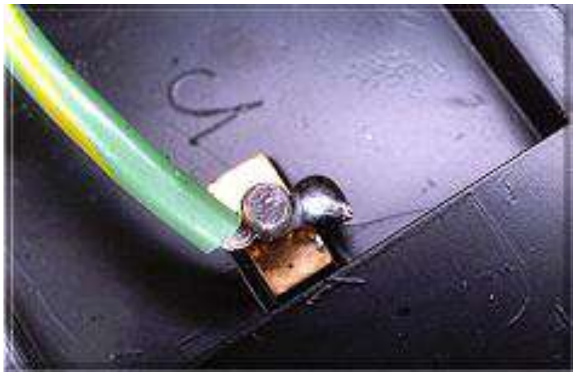
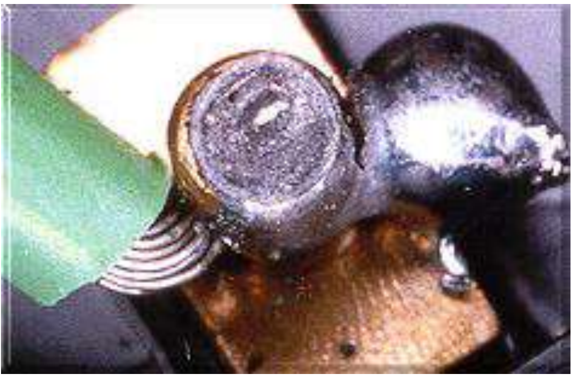

(Left) Hmmmm... this joint looks somewhat suspect as well... it's the earth (ground) wire in the same PSU

(Right) A close-up reveals the terrible standard of soldering (and quality control), with a fracture visible on this ground/ earth joint.

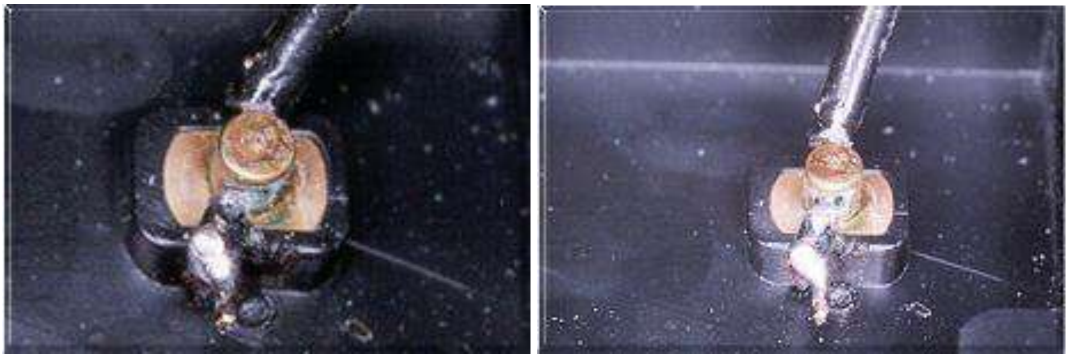

(Left) How **not** to make a mains voltage soldered joint. This solder joint went "dry" and starting arcing, nearly destroying the attached equipment. It is also a fire hazard.

(Right) The same mains connection, the wire merely being 'tacked on' with a blob of solder.

First Aid

If you are unlucky enough to receive burns which require treatment, here's what to do :-

1. Immediately cool the affected area with cold running water for several minutes.
2. Remove any rings etc. before swelling starts.
3. Apply a sterile dressing to protect against infection.
4. Do not apply lotions, ointments etc., nor prick any blisters which form later.
5. Seek professional medical advice where necessary.

A Few Hints about soldering on a Ground Plane

Before starting

- Always remove any grease and water from your hands. This will reduce the extent to which you oxidise and contaminate the board.
- Avoid touching the copper surface with your hands, always hold a board by its edges.
- Ensure all power and signal inputs are turned off (the soldering iron will short anything you touch to ground!!!)

Soldering a component

- Consider the circuit pathway from input to output in order to minimise the distance. This is especially important between ICs.
- Use the cleaning rubbers or solutions to clean away the oxidised surface where you are going to mount the component. This will help improve the electrical characteristic and improve solder flow
- Where you intend to place solder, VERY LIGHTLY scratch a cross-hatch pattern on the copper plate with a sharp tool. This will increase the friction thereby preventing your soldering iron skating across the board surface
- Apply heat to the copper surface first then add the solder so it smoothly flows on to the copper.
- Use pliers to bend the wire/chip legs to form small feet as this will increase the stability and make them much easier to solder. All wire bends should be done with pliers!
- Tin the legs before trying to mount them on the board
- When using wire, devise a colour coding system and stick to it i.e. red +ve rail, black GND rail, yellow –ve rail, blue input etc. This will make debugging so much easier
- If you are soldering an IC, two good techniques are:
  - Bend its legs with pliers so they are all horizontal and the chip can sit flat on the GND plane surface. Then place a bend just after where the pin leg narrows and bend again slightly further down so the leg it touching the ground plane (see image below). This way the IC is firmly held in place.

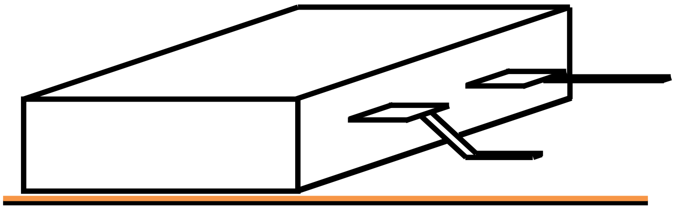

  - Bend legs touching the copper plane at the end to make small feet, then bend all other legs so that they are horizontal. This will result in the chip hovering just off the GND plane reducing the stability, but increasing the space to connect further components.
- Keep component leads as short as possible. This will reduce parasitics, improve mechanical robustness and reduce the chance of accidental shorts

These rules should be followed in conjunction with Dr Clarke's Soldering Guidelines found on the EE2 lab website.

Here is a good copper board layout. Notice the neat leads with none left loose above the circuit, and the way that the layout looks like a schematic. See the web copy (on the AMP handout page) of this picture for more detail.

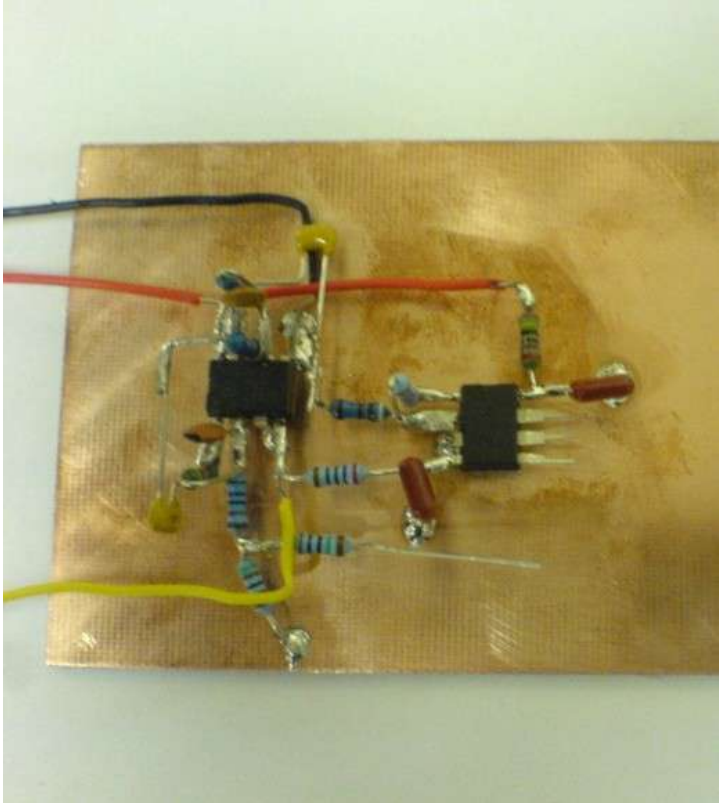

Soldering and Desoldering Instruction

Soldering is defined as "the joining of metals by a fusion of alloys which have relatively low melting points". In other words, you use a metal that has a low melting point to adhere the surfaces to be soldered together. Soldering is more like gluing with molten metal than anything else. Soldering is also a must have skill for all sorts of electrical and electronics work. It is also a skill that must be taught correctly and developed with practice.

Remember that when soldering, the rosin in the solder releases fumes. These fumes are harmful to your eyes and lungs. Therefore, always work in a well-ventilated area. Hot solder is also dangerous. Be sure not to let is splash around because it will burn you almost instantly. Eye protection is also advised.

The Tools

Soldering Iron – Even the cheapest of them will do the job. They may not last as long, but they do get hot enough to melt solder and that is the goal. You only need to make sure that the one you buy has a suitable tip on it. The most typical tip is the one that is about the size of a 1/8" stereo mini-plug.

Solder – Get Rosin Core solder. Rosin will help the solder to flow onto the wires. Solder comes in different thickness. Buy what is appropriate for your job.

Desoldering Gun – This tool will make life a lot easier when you need to rework a previous solder job.
Solder Wick – Braided copper with rosin coating used to draw solder off of circuit boards.

SMD Rework station— Basically this is a high temp hair dryer with special nozzle attached. It heats up entire pins of the SMD device so you can remove them.

The Methods

Soldering – There is not art to soldering, it takes patience and practice to get it right. If you are doing it right, it will be easy and very fast. First, make sure that your soldering iron tip is clean. If it is dirtier than just a light gray color, you need to clean it. You can do this with sandpaper or a Scotch Brite pad. Next, turn your iron on and give it plenty of time to heat up. To test the heat, use a piece of solder touched to the tip. If it melts immediately, it is ready.

The most common way to mess up a solder job is to let the soldering iron stay on the parts to be soldered too long. What this does is oxidize the metal, making it too dirty to attract solder. If you are working with a circuit board, too much heat can damage the board, rendering it useless and in need of repair. You should only hold the soldering iron to the parts to be joined for no more than 10-12 seconds. Any longer and you will melt insulation on wire or damage a circuit board.

It has been said that soldering is a two-person job. You need two hands to hold the parts together, one hand to hold the iron, and another to feed the solder. The correct method for applying solder is to hold the joint perfectly still while heating with the soldering iron. After a few seconds, introduce the end of the solder at the point where the iron meets the parts

If the solder does not melt immediately and flow onto the joint, pull the solder away and try again after a couple seconds longer. If you exceed 10 seconds, pull the iron off and try again after it all cools down. You probably didn't have the iron touching enough of the parts to be soldered. Sometimes, the parts to be soldered are so big that they conduct the heat away very quickly and make it difficult to solder. In this case, it is OK to put some solder on each part individually and then put them together later by melting the solder on each one while they are touching.

Joining Two Wires - Strip off about 3/8" of insulation from the two wire ends to be joined. Place a ½" length of heat shrink tubing over one wire and push it back so that the heat from soldering won't shrink it prematurely. Fold each bare wire end back on itself so that the tip of the wire now comes back to the end of the insulation. Link both wires together using the bends like hooks. Now, twist the ends of the wires around themselves tightly. This type of connection is called a Western Union and it is the strongest method of joining two wires together. Now, heat the connection with the iron and allow solder to flow over the entire joint. Wait a few seconds for it to cool off and then slide the heat shrink tubing down over the connection and heat it with a match or heat gun.

Soldering a Wire to a Circuit Board – This is delicate work that you will do while soldering. Too much heat and your board will be damaged. If you are attaching a wire to an existing hole on the board, make sure that the hole is clean and free of any excess solder (see de-soldering for more detail). The best advice when soldering to a circuit board is to angle your soldering iron tip so that it makes good contact with the pad on the circuit board and the wire or part to be soldered at the same time. Heat the parts up and touch the solder at the point where the tip meets the part and the pad. If the solder does not flow immediately, pull everything off and wait for it to cool down and try again later.

Step 1

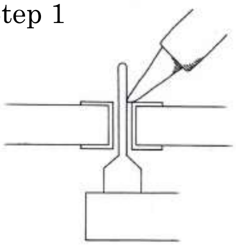

Apply heat to both pin and pad on the circuit board. Make sure tip of the soldering iron is touching both pin and pad.

Step 2

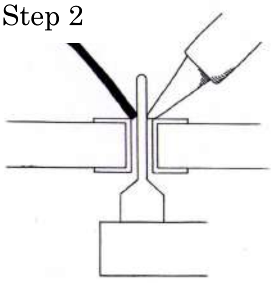

When the both pin and pad are heated enough, apply solder to the pin and pad. Solder should starts to flow.

Step 3

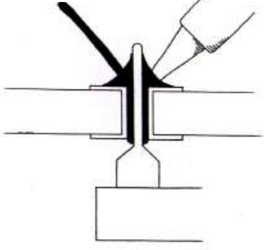

Apply enough but not too much solder. While applying the solder, do not move the tip of iron.

Step 4

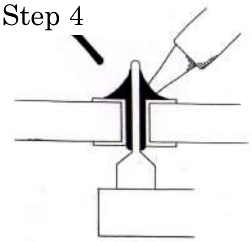

Remove solder away.

Step 5

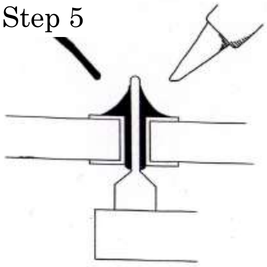

Remove tip of soldering iron also.

Step 6

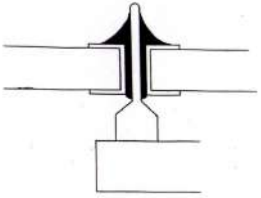

Done! Check the joint to make sure the connection.

Good soldering and Bad soldering Joints

Here are some example of Good soldering and Bad soldering joint.
For Japanese, we often say, “Good soldering joint should look like Mt. Fuji.”

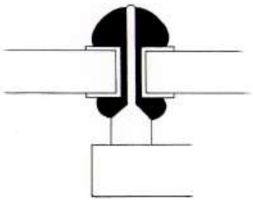

Too much soldering

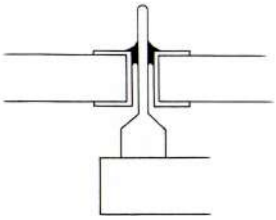

Too little soldering

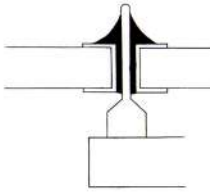

Good soldering

Cold Solder Joints
A cold joint is a joint in which the solder does not make good contact with the component lead or printed circuit board pad. Cold joints occur when the component lead or solder pad moves before the solder is completely cooled. Cold joints make a really bad electrical connection and can prevent your circuit from working.
Cold joints can be recognized by a characteristic grainy, dull gray color, and can be easily fixed. This is done by first removing the old solder with a desoldering tool or simply by heating it up and flicking it off with the iron. Once the old solder is off, you can re-solder the joint, making sure to keep it still as it cools.

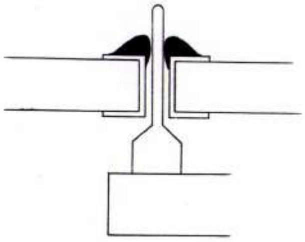

Cold soldering joints

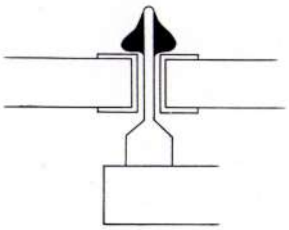

Tips and Tricks

Soldering is something that needs to be practiced. These tips should help you become successful so you can stop practicing and get down to some serious building.

1. Use heat sinks. Heat sinks are a must for the leads of sensitive components such as ICs and transistors. If you don't have a clip on heat sink, then a pair of pliers is a good substitute.

2. Keep the iron tip clean. A clean iron tip means better heat conduction and a better joint. Use a wet sponge to clean the tip between joints.

3. Double-check joints. It is a good idea to check all solder joints with an ohmmeter after they are cooled. If the joint measures any more than a few tenths of an ohm, then it may be a good idea to re-solder it.

4. Use the proper iron or temperature. Remember that bigger joints will take longer to heat up with an 30W iron than with a 150W iron. While 30W is good for printed circuit boards and the like, higher wattages are great when soldering to a heavy metal chassis.

Desoldering

Desoldering is extremely difficult compare to soldering. In the process of RomBoard installation, the parts and circuit board must be in the good shape to re-use them. The tool we use is Desoldering Gun. This device has vacuum pump built in with heater tip. Process of desoldering it self is very simple, but there are some tricks to do clean and safe desoldering job.

Step 1

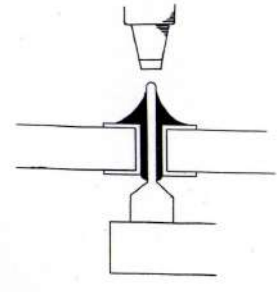

Put the desoldering gun's tip over to the soldering joint. Make sure nozzle is 90° angle from circuit board.

Step 2

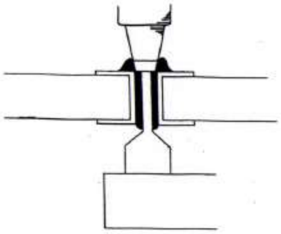

Apply the enough heat to melt entire solder. Multi-layer board or ground pin may take while to transfer heat.

Step 3

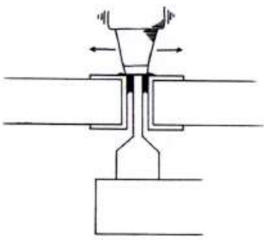

While applying vacuum, move the tip around the pin to get all the solder of around the pin out. At this time, you can also feel pin moves freely, so you know pin is free from circuit board.

Step 4

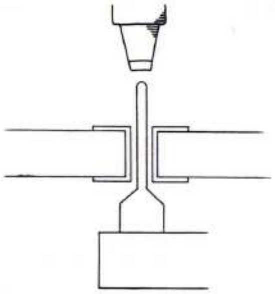

When you can hear air flow sound from hole, you should be done. Stop applying vacuum and remove the gun from the pin. Inspect hole and pin.

When you done with desoldering, the parts that you are trying to remove should move freely. If it doesn't, find which pin is still has solder left, and re-apply fresh solder to it and try desoldering process again. The multi-layer circuit board require more heat to get solder to melt. Make sure pin start to move freely by moving the tip of soldering gun before you apply vacuum to it.

SMD device soldering and removal

In the process of RomBoard installation, chance of handling SMD devices is becoming higher and higher due to ECU configuration change.
Soldering of SMD devices are not much different from regular through hole soldering.
Important thing is positioning of the device is very critical.
Start with applying little solder to two of the each corner of the pads on the circuit board.
Then, place the SMD device and re-heat a one of the solder you just applied to connect a SMD device pin to the pad. Check the position of the device, if position is right re-heat the other end of solder to secure the device completely. If you didn't get position right, re-heat the pin that you just soldered and while heating a pin, move the device to the right position.
Once you positioned the device in the right position, apply Rosin Flux to both pins and pads. This will help your soldering job by keeping solder separate from each pin.
Here is how the joint on SMD device pin should look like.

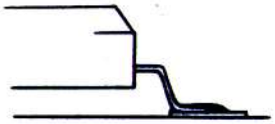

QFP and SOP chip

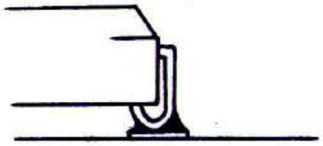

PLCC chip

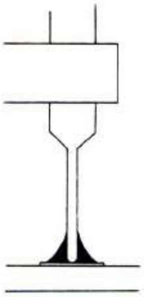

RomBoard adapter socket

Removing SMD device will be the probably the most difficult process. SMD rework station is used to do this. Make sure all the pins are completely heated otherwise you will be removing a pad from the circuit board along with device. Also, you should note, when you are applying heat to the device, some parts around the device is also heated and moves around when you touch. It is good idea to take a note of location of the devices near the target device.

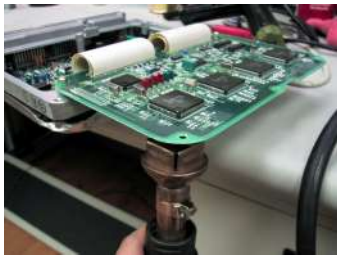

Step 1
Apply pre-heat from back side of the target device. Keep applying heat until you can't touch the chip. By this time, circuit board should have enough heat.

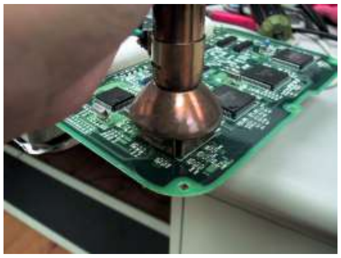

Step 2
Apply heat from top of the target device. Nozzle should cover entire chip. You can check if chip is free from solder by sliding the nozzle left and right. Make sure you don't move other parts. If the chip moves freely, it is ready to be removed.

Step 3
Remove the chip straight up to avoid making solder bridge between pins. Chip is extremely hot! Unless you have skin of steel, avoid imitating the picture.

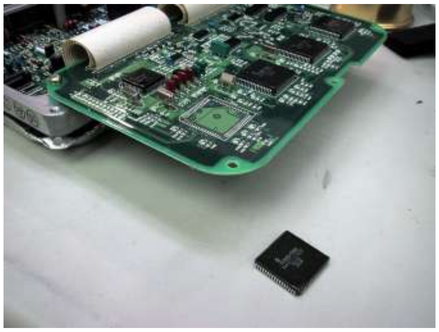

Step 4
After remove the chip, cool down both circuit board and chip.

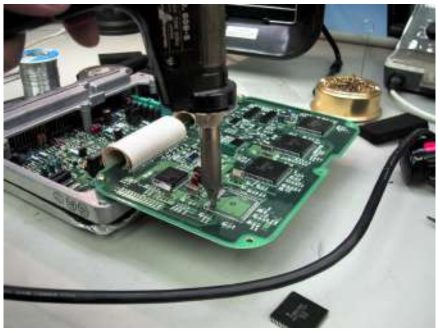

Step 5
Clean the pads on the circuit board with desoldering gun. Do not apply too much heat to the pad. Pad might peel off from circuit board. Clean the chip by desoldering gun also. Check for any bridge between pins. If the chip has too much protective coating on it, use solvent to clean it up.

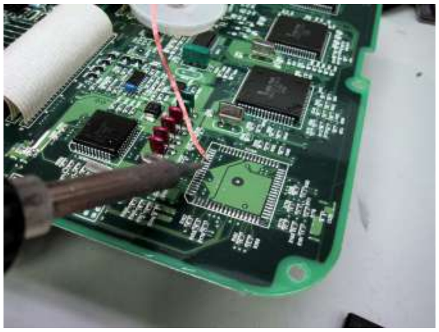

Step 6
If the pads on the circuit board is too small to use desolding gun, try cleaning by solder wick with rosin flux. This may take time but it is much better than peeling pads off.
Once again, do not apply too much heat.

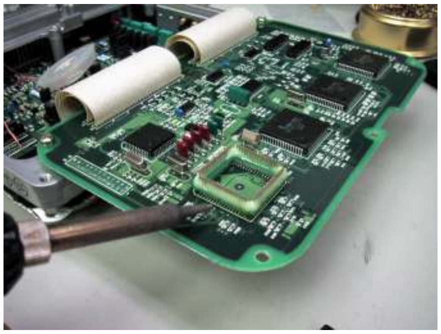

Step 7
Solder the RomBoard adapter where the chip was.
Refer SMD device soldering section for how to solder adapter.

The Final Word

Soldering, desoldering, and working with SMD device will take some effort to learn. You should be relaxed and concentrated when you work.
Keeping every thing clean (tip of iron, gun, work surface, etc) will help you also.
Once you get comfortable to work with soldering you may find your own tips and tricks.
Experience is all bout soldering.

EOF

Strain Gage Soldering Techniques

Introduction

The most common method of making electrical connections in strain gage circuits is by means of soft solders, in wire form. Other methods, such as spot welding, brazing, compression bonding, paste solders, and conductive epoxies, are also available, but find only limited application. Solders have many advantages for strain gage use — they are low in cost, readily available in various alloy compositions to provide a range of melting temperatures, and are easily obtained in the form of either solid wire or wire with a core of flux. They are convenient to use, and offer an excellent combination of electrical and mechanical properties.

Although soldering is basically a simple procedure, it must be done with appropriate tools, supplies, and techniques to assure accurate strain measurement. This is particularly true when test requirements are severe in the sense of approaching the limits of the strain gage circuit capabilities; e.g., long-term stability, high-elongation measurements, fatigue endurance, etc. Use of improper materials or techniques can significantly degrade strain gage performance.

The purpose of this Tech Tip is to outline recommended procedures and materials for attaching leadwires to strain gage solder tabs or to bonded printed-circuit terminals. These reliable, experience-proven methods are based on the use of a professional quality soldering station, in conjunction with Micro-Measurements solders and installation accessories.

Soldering Station and Pencil

For precision soldering of strain gages, it is always necessary to use a temperature- or power-controlled soldering station that provides low voltage and adjustable temperature to the soldering iron tip. An unregulated soldering iron, connected directly to the power line, is not ordinarily suitable for strain gage use because the tip temperature is apt to be far too high. This tends to oxidize the tip, and to instantly vaporize the flux, making soldering much more difficult. In addition, the unnecessarily high temperature may damage the strain gage, the bonding adhesive, or even the test specimen. For these reasons, the soldering station should incorporate provision for adjusting the soldering temperature to suit varying installation conditions and requirements. The temperature must be adjusted, of course, to accommodate the melting points of the different solders commonly used for strain gage connections, but also to allow for environmental conditions such as drafts or outdoor soldering in cold weather. Moreover, the temperature controller should be carefully designed to ensure that it does not generate electrical noise that could adversely affect nearby measuring instruments when both are in use.

Design of the soldering pencil also requires special consideration. It should be light in weight, with a very flexible power cord, and with the gripping area thermally insulated from the heating element. These characteristics contribute to the comfort, ease, and precision of soldering, and minimize operator fatigue during long periods of use. The soldering tip itself should be of the flat, chisel, or screwdriver type. Pointed tips should not be used, because they tend to draw solder away from the work area, and thus make it more difficult to achieve a proper joint. In contrast, flat tips act to confine the solder, while offering greater surface area for better heat transfer and more effective soldering, generally.

Micro-Measurements soldering units incorporate all of the above features and a number of others, designed to help the user easily make consistent, reliable solder joints. These soldering units are widely used by professional strain gage installers everywhere, in both stress analysis laboratories and in transducer manufacture.

Solder Selection

The Micro-Measurements Division stocks a broad range of solder types to meet various installation and test requirements. While solders are sometimes selected to provide specific electrical or mechanical properties, the most common basis for selection is simply the melting-temperature range. Low-melting-point solders, for example, are generally used for strain gage installations on nonmetallic test parts to avoid damaging the gage, bonding adhesive, or test material due to overheating. In contrast, high-temperature solders are normally selected only when necessary to satisfy elevated-temperature testing requirements. These solders are somewhat more difficult to handle because the higher working temperature

Strain Gage Soldering Techniques

rapidly vaporizes the flux, and oxidizes the soldering tip, both of which tend to impede the soldering process. Specially designed soldering tips are recommended for high-temperature use.

For routine applications, where test conditions do not dictate the use of either a low- or high-temperature solder, an alloy with an intermediate melting temperature is the normal selection. The 63/37 tin-lead alloy (Type 361A-20R) is an excellent choice for general-purpose strain gage soldering. As an eutectic alloy, it has a sharply defined melting temperature — a characteristic that largely eliminates “cold” solder joints. The addition of a trace of antimony provides superior performance when the soldered connections will be exposed to very low (cryogenic) temperatures for long periods of time.

The general-purpose solders are supplied with a core of activated rosin flux. This makes soldering much more convenient, and is particularly useful in field applications where accessory liquid rosin flux (M-Flux AR) may not be available. Solid-wire solder, with externally applied acid flux (M-Flux SS), is recommended for making soldered connections to Micro-Measurements K- and D-alloy (modified Karma and isoelastic) strain gages. Rosin-core solders should not be used in conjunction with acid flux.

Silver solder (Type 1240-FPA) is available for applications where leadwire connections will be exposed to temperatures above about +550°F (*+290°C*). This solder, in paste form, is not suitable for attaching wires directly to strain gage solder tabs or to bondable terminals, but is intended for connecting instrument leads to preattached strain gage leads, as with WK-Series gages using a special resistance soldering unit. Techniques for making leadwire connections with silver solder are described in Micro-Measurements Tech Tip TT-602, *Silver Soldering Technique for Attachment of Leads to Strain Gages.*

Soldering Flux

The function of a soldering flux is to remove oxidation from the members being joined (solder tabs, terminals, leadwires), and to prevent further oxidation during soldering. For making leadwire splices, or soldering directly to constantan foil or copper terminals, the flux contained in a rosin-core solder is usually sufficient. With higher temperature solders, however, it may be necessary to supply additional flux. A liquid activated-rosin flux such as M-Flux AR is recommended for this purpose.

Acid fluxes should never be used on constantan strain gages or copper terminals, or for splicing copper leadwires; and paste fluxes, containing chlorides, should not be used under any circumstances for strain gage soldering. When tinning bare (without soldering options) solder tabs of Micro-Measurements K- and D-alloy strain gages, a liquid acid flux (M-Flux SS) is recommended. After the tinning operation, the residual flux must be completely neutralized within one to two minutes; and then the leadwire joint can be completed using the same solder and M-Flux AR rosin flux or a rosin-cored solder.

Preparation of the Soldering Tip

New soldering tips should always be tinned with solder prior to initial use. This is easily accomplished by wrapping one to two in (*25 to 50 mm*) of solder wire around the working portion of the tip while the soldering iron is cold, before applying power to the soldering station. If rosin-core solder is used, no external flux is required. With solid-wire solder, however, the wrapped tip should be dipped into liquid rosin flux (M-Flux AR) to provide sufficient flux for initial tinning. Set the control on the soldering station to the appropriate temperature range for the solder, and apply power to the unit. Allow the soldering pencil to heat until the solder wrapped around the tip melts completely. Remove excess melted solder from the tip with a dry gauze sponge. Never knock the heated soldering pencil against any object to remove excess solder, since this may result in personal injury or damage to the soldering pencil.

Note: Cross-alloying of solders can change the electrical, chemical, thermal and mechanical properties of the solder being used. To prevent cross-alloying, it is recommended that only one type of solder be used with each soldering tip. Of course, if one type of solder is incorporated in a gage with solder dots and another type is added, a mixture is produced. This mixture cannot be expected to have melting and strength properties any better than those of the lower temperature component.

Oxidation of the soldering tip seriously hinders the soldering operation. The tendency for oxidation can be minimized by ensuring that excess melted solder remains on the tip at all times when it is not actually in use. Negligent maintenance practices, or wiping the hot tip with materials that char on the surface, will produce a buildup of oxide that prevents proper soldering. If the tip does become oxidized, the following procedure is effective for cleaning and re-tinning:

1. Set the soldering station to the appropriate temperature range for the solder in use.
2. Place several drops of M-Flux SS on a glass plate. Re-tin the soldering surface by holding the heated tip in the SS flux while feeding solder onto the tip. A generous amount of solder is essential for proper tinning.

Strain Gage Soldering Techniques

3. Wipe the excess solder from the tinned tip with a dry gauze sponge. For severely oxidized tips, it may be necessary to repeat this operation several times to obtain a properly tinned surface. The soldering tip should never be filed or sanded, since this may remove the plating on the tip, accelerating the oxidation and leading to the early deterioration of the tip. After the cleaning operation, remove excess solder, re-tin and clean the tip several times, using rosin-core solder, or solid-wire solder withM-Flux AR.

Tinning Solder Tabs and Bondable Terminals

All strain gage solder tabs, terminals, and leadwires must be properly tinned before making soldered connections. This helps ensure active surface wetting and good heat transfer during the soldering operation. Tinning stranded leadwires to produce a formable solid conductor will also greatly simplify the leadwire attachment procedure.

Before tinning the solder tabs on open-face (unencapsulated) strain gages, the measuring grid should be protected with PDT drafting tape. The drafting tape is positioned to cover the entire grid and the upper portion of the solder tabs, as shown in Figure 1. This not only shields the grid from soldering flux and inadvertent solder splash, but also restricts the flow of solder on the tabs. The tinned area on the solder tabs should be only large enough to easily accommodate the leadwire size in use. The latter consideration is particularly important when making installations for dynamic applications or large-strain measurement.

The tinning procedure for strain gage tabs and terminals consists of first cleaning and reapplying a small amount of solder to the hot soldering iron tip. Next, apply a drop of M-Flux AR to the tab or terminal (this step can be omitted if a rosin-core solder is used). When soldering directly to bare Karma or isoelastic foil, use M-Flux SS on the gage tabs only. Hold the soldering pencil in a nearly horizontal position (<30°), with the flat surface of the tip parallel to the solder tab or terminal. Place the solder wire flat on the gage tab, and press firmly with the tinned hot soldering tip for about one to two seconds, while adding approximately 1/8 in (*3 mm*) of fresh solder at the edge of the tip. This procedure assures that there is sufficient solder and flux for effective tinning. Simultaneously lift both the soldering pencil and solder wire from the tab area.

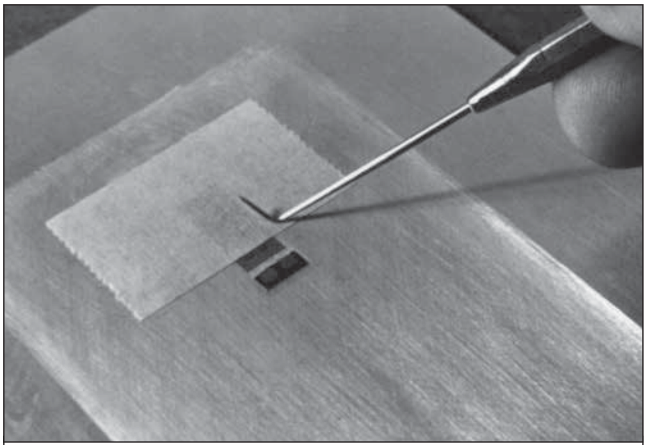

Figure 1 – Gage grid and upper portion of solder tabs masked with drafting tape.

NOTE: Lifting the soldering iron before lifting the solder may result in the end of the solder wire becoming attached to the tab; lifting them in the reverse order can leave a jagged (spike) solder deposit on the tab. When the operation is performed properly, it will produce a small, smoothly tinned area on the tab or terminal.

If M-Flux AR or a rosin-core solder is used in the tinning, it is not necessary to remove the residual soldering flux at this time. However, when M-Flux SS is employed to tin the bare solder tabs of K- or D-alloy gages, the acidic flux residue must be removed immediately following the tinning operation. To remove the residue, apply M-Prep Conditioner A liberally, and wash the area with a soft brush; then blot dry with a clean gauze sponge. Next, wash again with freely applied M-Prep Neutralizer 5A, and blot dry with a clean gauze sponge.

NOTE: Special procedures for tinning and wiring strain gages supplied with preattached solder dots are described in Micro-Measurements Tech Tip TT-606, *Soldering Techniques for Lead Attachment to Strain Gages with Solder Dots.*

Tinning and Attaching Leadwires

Of course, leadwire ends must be stripped of insulation before tinning, and this should be done with a thermal wire stripper to avoid the damage to the wire that often occurs when mechanical wire strippers are used. After the wires are stripped, the ends of stranded conductors should be twisted tightly together before tinning. The bare leadwire ends can then be tinned easily with the following procedure:

1. Remove excess solder from the soldering tip, using a dry gauze sponge. Then melt fresh solder on the hot tip to form a hemisphere of molten solder about twice the diameter of the wire to be tinned.
2. If rosin-core solder is used, slowly draw the bare wire through the molten solder while continuously adding fresh solder to the interface of the wire and soldering tip. With solid-wire solder, apply M-Flux AR to the

Strain Gage Soldering Techniques

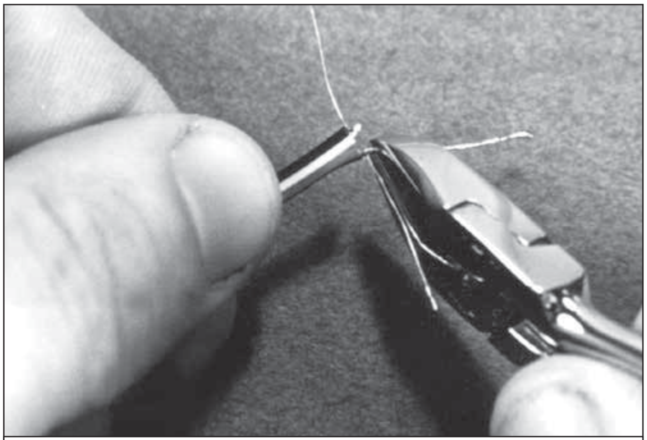

Figure 2 – Trimming leadwire ends before taping in place.

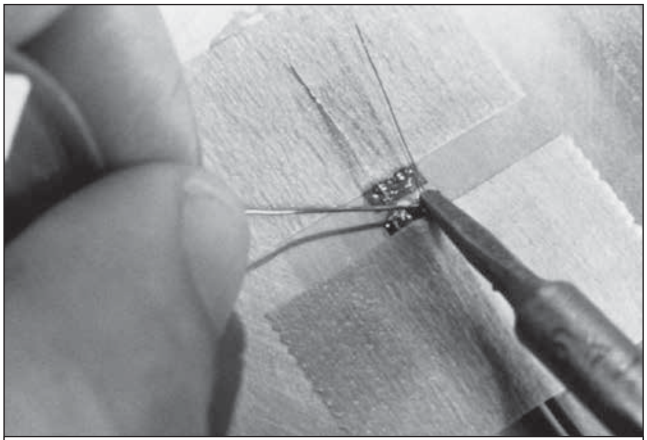

Figure 3 – Leadwire end taped to surface in preparation for soldering.

wire end before starting to tin, and proceed in the same manner. This will produce a smooth, shiny coating of solder over the bare wire.

For applications employing bondable terminal strips and stranded instrumentation wire, it may be convenient to use a single strand of the wire as a jumper between the terminal and the strain gage solder tab. In such cases, the single wire strand should be separated out before twisting and tinning the remaining strands (see Micro-Measurements Tech Tip TT-603, *The Proper Use of Bondable Terminals in Strain Gage Applications*).

Leadwires should be formed and routed to the strain gage or terminal strip, then firmly anchored to the test-part surface with drafting tape before making the soldered connection. Attempting to route the leadwires after completing the solder joint will often result in damage to the gage or terminals. Routing into the connection area should be along a minimum strain direction (such as the "Poisson" direction in a uniaxial stress field) particularly for high elongation or dynamic tests. The tinned leadwire end should be trimmed short enough so that it will not protrude through the connection area, and cannot inadvertently make electrical contact with the test-part surface or adjacent solder connections. Figure 2 illustrates this stage in the procedure. In the final preparatory step, bend the leadwire end slightly to form a spring-like loop, and tape the wire firmly in place over the connection area, using PDT drafting tape. The tape should be within about 1/8 in (*3 mm*) of the connection area, as shown in Figure 3.

Clean and re-tin the soldering iron tip with fresh solder. The temperature of the iron should be adjusted so that the solder is easily melted, without rapidly vaporizing the flux. If the iron temperature is either too low or too high, it may cause poor solder connections, or it may damage the strain gage, terminal, or bonding adhesive. Apply a small amount of M-Flux AR to the joint area and, holding the soldering pencil nearly horizontal, firmly press the flat surface of the tip on the junction for about one second; then lift the tip from the soldered joint. If needed, additional flux can be provided during the joining operation by feeding a little fresh solder into the joint from a spool of rosin-core solder. This procedure should result in a smooth, hemispherical solder joint, without any peaks or jagged areas. If the solder joints are not smooth and uniform in size, repeat the soldering procedure, using additional flux and/or solder as necessary.

Cleanup and Inspection of Soldered Joints

After completing the soldering operation, it is imperative that all traces of residual flux be completely removed with RSK Rosin Solvent. The same solvent is used to soften the mastic of the drafting tape, permitting its easy removal. Do not try to pull away the tape with tweezers or other tools, because this may result in damage to the soldered connections or the strain gage grid. Thoroughly clean the entire installation area with generously applied rosin solvent and a soft-bristled brush. Clean the solder connection area until no visible signs of residual flux remain, and blot the area dry with a clean gauze sponge. Any traces of residual flux can cause gage instability and drift, and will inhibit bonding of the installation's protective coating. Incompletely removed soldering flux is the most common cause of degraded performance in strain gage installations. Residual flux mixed with a protective coating application can completely destroy the coating objective.

Visually inspect the soldered joints for any gritty or jagged joint surfaces, and for traces of flux. Solder connections

Strain Gage Soldering Techniques

should be smooth, shiny, and uniform in appearance. Any soldered joints that look questionable should be re-soldered, and flux removed. Check the resistance-to-ground of the completed gage installation, using the Model 1300 Gage Installation Tester. Low or marginal resistance readings suggest a leakage path between the soldered connections and the test-part surface. This condition usually results from residual soldering flux, or from bare leadwire conductors partially shorting the gage tabs or terminals to the test part. Soldered joints should not be tested by pulling on the leadwire, or by probing at the joint area. These practices frequently cause lifting or tearing of the solder tab from the gage backing material.

Summary

The ability to make consistently good soldered joints is essential for precision strain gage measurements. The techniques described here are straightforward and easily mastered, but they are most effective when used with professional soldering equipment which is specially designed for making soldered connections in strain gage circuits. The soldering pencil should be lightweight, with a flat chisel or screwdriver tip, and it should be connected to the soldering station with a very flexible power cord. Requirements for the soldering station include low-voltage operation of the soldering pencil, and provision for temperature adjustment to suit the type of solder and the application conditions. The equipment should not generate electrical interference that could affect sensitive measuring instrumentation. Solder selection is based primarily on the expected operating temperature range of the strain gage installation; and all solder tabs, bondable terminals, and leadwire ends should be tinned before soldering the joints. Soldered joints should always be smooth and shiny, with no jagged or irregular edges, and all traces of residual flux must be thoroughly removed prior to the application of protective coating. Use of the recommended materials and techniques, with careful attention to detail, will result in consistently proper and reliable soldered connections.

PROJECT LEAD THE WAY
PLTW

Soldering & De-soldering

Digital Electronics

Soldering & De-soldering

This presentation will…

- Review the tools needed to solder and de-solder electronic components.
- Demonstrate how to *tin* a soldering iron tip.
- Demonstrate the soldering process.
- Show the characteristics of a good solder connection.
- Review classic soldering mistakes.
- Demonstrate the de-soldering process.

Soldering Tools

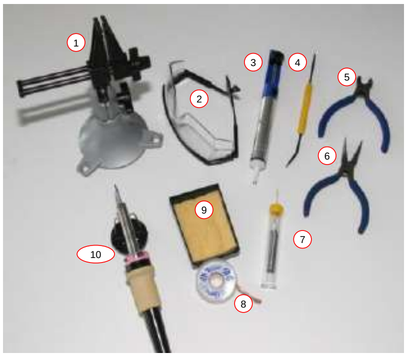

1) Vise
2) Safety glasses
3) Solder sucker
4) Solder tool
5) Diagonal cutters
6) Needle nose pliers
7) Solder
8) Solder wick
9) Damp sponge
10) Soldering iron

Soldering Iron

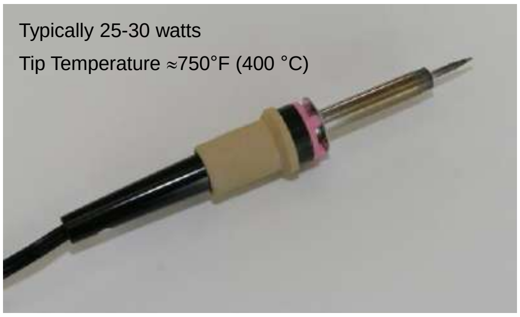
Typically 25-30 watts

Tip Temperature ≈750°F (400 °C)

Project Lead The Way, Inc.
Copyright 2009

Solder

- Solder is an alloy of tin and lead.
- The solder used for electronics is frequently called 60/40 solder because it is made of 63% tin and 37% lead.
- 60/40 solder melts at 361° F.
- Lead-free solder**:** As of July 1st, 2006, European laws mandated that new electronics be entirely lead-free. As of yet, no such laws exist in the United States.

Soldering Iron Care & Maintenance

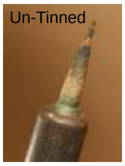
Un-Tinned

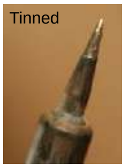
Tinned

- A soldering iron must be coated with a thin coat of solder. This will allow for the transfer of heat to the work piece.
- This procedure is called tinning.
- The tip must be kept coated with a shiny layer of solder by occasional wiping and applying solder directly to the tip.

Tinning Process

Apply Solder to Soldering Iron Tip

Roll Tip on Damp Sponge

Properly Tinned Soldering Iron Tip

Solder Process

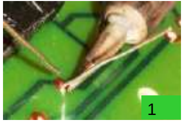

Heat both items at the same time by applying the soldering iron to the copper pad and the component lead.

Continue heating and apply a few millimeters of solder. Remove the iron and allow the solder joint to cool naturally.

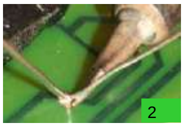

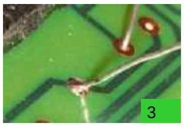

It only takes a second or two to make the perfect joint, which should appear shiny.

Copyright 2009

A Good Solder Joint

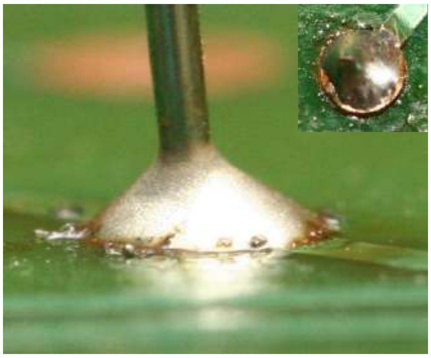

- Smooth
- Bright
- Shiny
- Clean
- Concave fillet

Bad Solder Connections

Too Much Solder

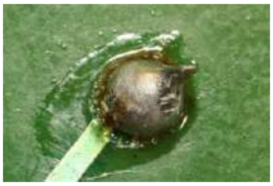

Too Little Solder

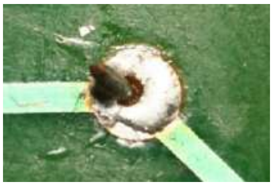

Bad Solder Connections

Cold Solder Joint

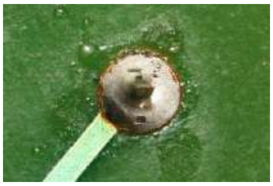

Not Soldered

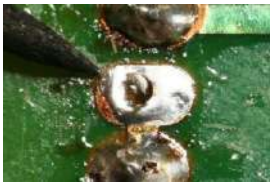

Bad Solder Connections

Solder Bridge

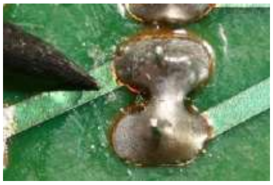

Lifted Trace/Pad

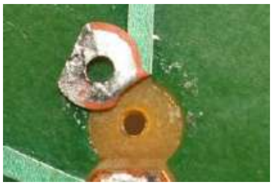

De-Solder Process: *Solder Sucker*

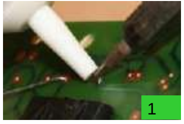
1

Apply heat to the connection to be de-soldered. When the solder melts, trigger the solder sucker.

Repeat de-soldering as needed until all solder is removed. Remove soldering iron & solder sucker from area.

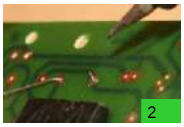
2

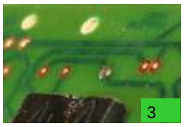
3

Remove component lead.

De-Solder Process: *Solder Wick*

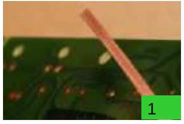
1

Solder wick is finely braided copper that is used to *wick away* excess solder from a de-soldered connection.

Apply the solder wick and soldering iron to the de-soldered connection. The solder wick will draw the excess solder off of the PCB pad.

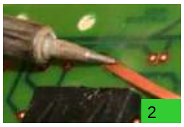
2

3

De-soldered PCB pad

Soldering Safety

- Wear safety glasses when soldering. This includes all individuals in the vicinity of someone who is soldering.
- Place soldering iron in an approved holder when not in use. The iron is hot and can cause burns.
- Place the soldering iron so that the cord does not get caught up in your arms or on others.
- Ensure access to proper ventilation.
- Verify that the type of solder is safe to use in your working environment.
- Secure the components to be soldered before beginning the soldering process.

Soldering Safety

- Provide plenty of space to work.
- Use a properly-sized point for the soldering job to be completed.
- Verify that the tip on the soldering iron has a sharp point and has not been damaged in any way.
- Check the power cord for burned or melted sections that show bare wires. Label those cords DO NOT USE and ask the instructor to repair or replace.
- Do not to touch molten solder - it is hot!
- Make sure that the solder strand is long enough to keep fingers away from the hot iron.

Soldering Safety

- Tie back long hair and remove or tuck loose clothing.
- Use heat sinks for heat-sensitive parts. Provide sufficient cooling time before removing parts.
- Do not flick solder off of the iron. Flicking can cause solder to spray and hit skin or eyes.
- Hold the scrap end when cutting excess leads so that the scrap lead is not thrown into the air.
- Cut leads evenly with wire cutters.
- Make sure that leads do not short across other traces or leads.
- Thoroughly wash your hands after handling solder.

How to Solder Like a Pro

Detailed instruction, helpful hints, safety tips, and a comprehensive troubleshooting guide

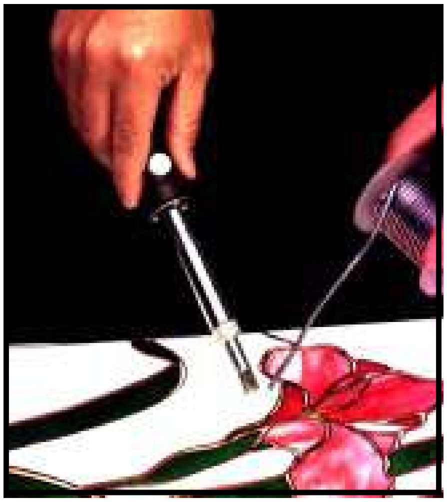

inland affordable products you can trust! inland

Table of Contents

Thank you for your interest in Inland's How to Solder Like a Pro manual. Whether you are a beginner or an experienced stained glass artist, you will find useful information about selecting and maintaining your soldering tools and improving your soldering skills.

Safety

The Importance of Safety 3
Fumes 3
Lead..3
Heat 3

Soldering Irons

Choosing a Soldering Iron 4
Types of Temperature Control 4
Irons with Temperature Controlled Tips......... 4
Ceramic Heating Element Irons 4
Temperature Controlling Devices 5
Tip Sizes, Styles and Uses................................. 5

Iron and Tip Maintenance

The Importance of Proper Care 6
The Cord 6
The Heating Element 6
The Tips 6

Solder

Why Do We Use Solder? 7
The Advantages of Solder 7
Types of Solder 7
Characteristics of Solder Types 8
Solder Composition Reference Table 8

Flux

Why Do We Use Flux? 9
Types of Flux 9
Characteristics of Fluxes 9
Getting the Most From Your Flux 9

Soldering Copper Foil

Before You Start 10
Cleaning Before Soldering 10
Getting Started 10
Running The Solder Bead 11
Fixing Imperfections 11
Other Things To Consider.................................11
Ending A Bead 12
Cleanup After Soldering 13
Finishing 13

Soldering Metal Cames

Step by Step 14

Soldering Other Metals

Vase Caps and Spiders 15
Metal Outer Channels.................................... ... 15
Hanging Hardware 16

Patina

How to apply patina..16

Trouble Shooting

Tips and hints to help you fix problems.......17-19

Glossary

Definition of terms..20-21

Safety

The Importance of Safety

One of the most important aspect of soldering any stained glass project is to take the proper measures to safeguard your health. Soldering involves working with potentially hazardous materials like lead, flux and harmful fumes. By observing specific safety precautions and using informed common sense, working in stained glass can be a safe and enjoyable past time. The following is intended to serve as a general rule of thumb. Since each person and situation is unique, you should use this information as a starting point to help you make informed safety decisions concerning your soldering habits.

Fumes

Solder only in a well ventilated area, and use an exhaust device that moves solder fumes away from your face. It's preferable to exhaust outside (check local/state building codes and restrictions on venting to outside air). If you don't exhaust to the outside, use a bench top fan or intake device with a replaceable smoke/fume absorber made for stained glass artists. It should draw solder fumes into the replaceable filter. The filter should be activated charcoal and designed to remove particles smaller than 1/2 micron from the air. There are also several OSHA approved respirators available for fumes. Consider wearing one in conjunction with a venting system, especially if you plan on soldering for several hours every day.

Lead

Most popular solders used in stained glass are lead based. When you are using them, follow these precautions:

- 👎 Never eat, drink, or smoke in any area where soldering takes place.
- 👍 Always thoroughly wash your hands after soldering.
- 👍 Make sure your soldering equipment and supplies are kept out of the reach of children.
- 👎 Do not discard lead or solder scraps into the trash. Find a means or place to recycle them.
- 👎 Never use lead based solders or cames on items that will come in contact with food or children, or will be frequently handled. If you are making kaleidoscopes, jewelry or napkin holders, use lead free solder.
- 👍 Consider having your blood lead level checked by your physician on a regular basis to help you monitor your handling practices.
- 👍 If you are pregnant, or considering it, you should check with your doctor before using lead or solder.

Heat

Soldering tools operate at high temperatures, so these safeguards are important:

- 👍 Wear safety glasses! Solder and flux can "pop" and "spit."
- 👍 Solder on a fire resistant surface. Homosote, or dry wall are good .
- 👎 Never leave your iron plugged in and unattended.
- 👎 Do not overload a wall outlet with too many electric appliances.
- 👍 Never set your hot iron down on anything other than an iron stand.
- 👍 Replace the cord of your iron if it becomes worn or gets burnt.
- 👍 To prevent burning your fingers, use needle nose pliers or heat resistant gloves to hold small pieces.
- 👎 Never cut off a grounding prong on an iron plug to make it fit an ungrounded receptacle.

Soldering Irons

Choosing a Soldering Iron

There is a lot to consider when you choose an iron for stained glass. The handle should be heat resistant and comfortable for you to hold. You may be holding it for several hours at a time, so consider the weight and balance of the iron. Carefully check the wattage of the iron that you are considering. It must be able to continuously generate enough heat to melt all of the types of solder you plan to work with and be compatible with the type of construction you use (i.e.-copper, foil, lead came, rebar, etc.). You will need an iron that is at least 60 watts, with chisel type tips, and a way to control the temperature.

Types of Temperature Controls

A broad range of soldering irons are available. Most irons, which we will call "conventional irons," are made with wound wire heaters in a barrel with mica insulation. They are often inexpensive, but they do not offer the ability to control operating temperature. A separate temperature controlling device must be used with these irons to achieve the best results. There are three types of soldering irons that are easier to use for stained glass projects, because they allow you to control temperatures more effectively:

Irons with temperature controlled tips.

Irons used with temperature controlling devices.

Irons with ceramic heaters.

Irons with Temperature Controlled Tips

These irons are supplied with an internal regulator in the tip that does not allow the iron to exceed a predetermined temperature. An example would be a 600°F tip. The iron heats to that temperature then "shuts off." When heat is required, the iron "turns on" again. Tips are available in predetermined temperatures up to 800°F. These irons are easy for beginners to use, because the temperature is automatically maintained for you, however, as your skills increase, you may prefer to control the amount of heat yourself for different soldering situations. The limited amount of control may become frustrating as your soldering skills increase. With this type of iron you will likely encounter "cold spots" where the iron will not melt your solder. These cold spots occur when heat is being drawn out of the tip faster than it is being replenished. You will need an iron with a minimum of 100 watts.

Ceramic Heating Element Irons

These irons are relatively new to the stained glass trade. They are made with highly efficient ceramic heating elements. Like a ceramic room heater, they produce a consistent temperature using less electrical wattage during operation. When initially heating and when reheating during "recovery" periods, they can draw a "burst" of power exceeding 100 watts and then efficiently reduce electrical consumption, often below 60 watts, during the soldering process. The result is efficiency and economy.

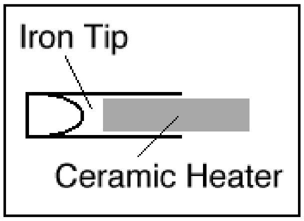

A remarkable feature of ceramic heater irons is they generally reach operating temperature in less than 60 seconds. Of the three types of noncoventional irons, the ceramic heater type best maintains consistent tip temperature over extended periods of time. Problems with slow heat recovery generally don't exist. A ceramic heater iron offers excellent flexibility and can be used with a temperature controller if you like.

Temperature Controlling Devices

A temperature controller is a device that operates similarly to a light dimmer switch. By dialing the control to a higher or lower setting, more or less electricity is fed to the iron. This increases or decreases the iron temperature, allowing the tip temperature to be controlled. You can establish the maximum temperature without changing tips.

A temperature control/iron combination offers you greater control and flexibility than a temperature controlled tip iron when working with different metals and solder mixes. Look for one that is at least 85 watts. Since the flow of electricity to the tip is consistent and never completely stops, encounters with cold spots are nearly eliminated. A temperature control/iron combination is suitable for all types of stained glass construction.

Temperature controllers are generally separate units that the iron plugs into. They are small, easy to use and relatively inexpensive. There are also irons that have controls built into the handle of the iron, which generally cost less than buying an iron and a separate control. Never plug an iron into a temperature control unless the manufacturer specifically states it is suitable for use with it.

Tip Sizes, Styles and Uses

Stained glass irons are generally sold with a "chisel" style tip. There are a variety of tip sizes and styles available for nearly every iron used in the stained glass industry. Different tips can expand the versatility of your soldering iron, so select an iron that offers more than one size replacement tip.

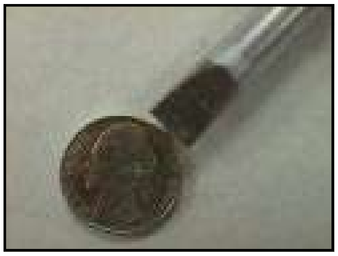

The standard tip size that comes with an iron is usually from 1/4" to 3/8" wide. This size works well for soldering either copper foil or lead came projects.

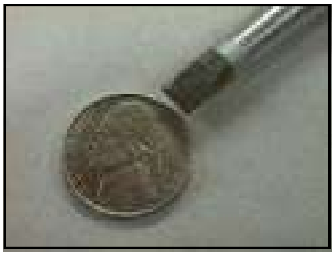

A smaller chisel tip, 3/16" wide, can be useful when soldering small pieces or when soldering in a tight area, such as a narrow inside part of a kaleidoscope. It is also very effective for decorative soldering.

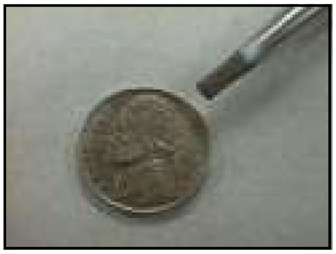

Very small tips, 1/8" or narrower, are usually used for decorative soldering. They allow you to create very fine details and designs with the solder.

Iron and Tip Maintenance

The Importance of Proper Care

Your soldering iron may very well be one of your biggest tool investments, so you will want to do everything possible to ensure that it gives you many years of service. Well maintained soldering irons and tips perform better and make the job of soldering much easier. Always place your soldering iron in a stable iron stand whether it is being used or not.

The Cord

- 👍 Make sure you plug the iron into the correct type of outlet.
- 👍 Try not to use an extension cord. If you must, use a heavy duty one.
- 👎 Don't drop or bang the iron. Ceramic heaters are especially easy to crack or break.
- 👍 Regularly check the cord for burns or cracks and have a professional electrician replace worn cords before using the iron.
- 👍 Make sure that the cord is not hanging in such a way that it can be pulled off of the table.

The Heating Element

- 👍 Do not allow the iron to idle at operating temperatures for extended periods. This could burn out the heater element. Unplug the iron or, if you are using a rheostat, turn it down to a low "idle" setting.
- 👍 Occasionally, remove the tip and **lightly** tap the barrel of wire wound heater irons to remove debris.
- 👍 If you will not be using your iron for an extended period of time, you may want to store it (after it has fully cooled) in a zipper type bag to protect it from corrosion and humidity.

The Tips

- 👍 When you are finished using your iron, remove the tip from the barrel. Removing the tip is essential to preventing "seizing" which can occur if it is left in the iron for extended periods. (If your tip seizes, you can easily damage the heating element trying to remove it. It is best to return you iron to the manufacturer for removal.)
- 👍 When reinserting tips, make sure they are properly seated in the barrel.
- 👎 Never dip your tip into flux in order to clean it. Instead, use a clean damp sponge to wipe all sides of the tip periodically as you solder. Doing this removes impurities the tip has accumulated from the solder and the environment. It will ensure that you are receiving the maximum heat at the tip surface.
- 👍 Properly cleaned tips are bright and shiny. If your tip becomes "blackened," you can usually remove the buildup with a wet sponge, a tinning block or by gently using a brass brush. A “tinning block” (sal-amoniac) is used by placing a small amount of flux on the block and rubbing the tip of your hot iron in it. Wipe the tip on a damp sponge to remove debris. You may need to repeat this several times if your tip is very dirty.)
- 👍 Never use sandpaper or any abrasive material to clean a tip.
- 👍 The best way to minimize your tip maintenance is to find a good quality solder. Use one that has a high tin content and high metal purity.

Solder

Why Do We Use Solder?

Solder is a combination of tin and lead used to create a strong bond between other metals. Since solder won’t stick to glass, we apply a copper foil tape (our metal) to the edges of the glass. This is refered to as the copper foil method of stained glass construction. Solder is melted over the copper foil, creating a structure that holds the pieces of glass together. The other option is to set the glass into channels of lead or metal, and solder the channels together. This is refered to as the lead came method.

The Advantages of Solder

Some of the advantages of soldering versus other bonding methods are:

- Solders are easy to use and relatively inexpensive.
- Low energy is required to solder.
- Properly soldered joints are highly reliable.
- Solder joints are easy to rework or repair.
- Experienced craftpersons can exercise a high degree of control over the soldering process.
- Solder joints age very well. They can last for years, decades and centuries.

Types of Solder

SILVERGLEAM

Stained glass solders are usually a mixture of tin and lead, designated by two numbers representing the percentages of each metal in that specific mix. The first number always refers to the percentage of tin, the second is the percentage of lead. The most commonly used solders in stained glass are 60/40, (60% tin/40% lead) 50/50 and 63/37. "Lead-free" solders have no number designation and are a mixture of tin and small amounts of other metals. Avoid solders containing antimony, a very toxic element. Instead look for lead-free solders containing silver, or copper. They are safer and easy to use. (Note they are, however, more difficult to patina.)

In North America, you will find solder is generally sold in solid core wire form on a spool. The common spool size is one pound. In Europe, you will find solders primarily sold in a bar form. Never use acid-core or rosin-core solders for stained glass work. Look for solders that are sold as "free of impurities" in the component metals. Impurities cause a "scum" on your solder bead, degrade soldering iron tips, and interfere with the proper reaction of patina chemicals resulting in undesired finishes.

An important term for solders is the "working range" or "pasty range." This is the range of temperature between which solder transitions from liquid back to solid.

Characteristics of Solder Types

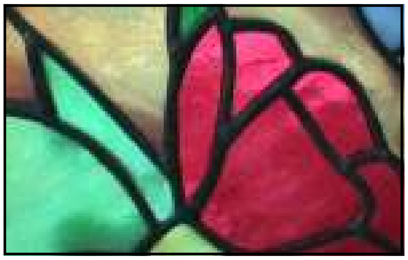

60/40 Solder: Composed of 60% tin and 40% lead, this solder melts at 374°F, but doesn't become completely solid until it cools to 361°F. This means it has a "pasty range" or "working range" of 13 degrees. This solder is your best choice for copper foil work. The liquid temperature and narrow "pasty range" make it easy to form and maintain consistent high, rounded, beaded seams. Because of its relatively low melting point, "60/40" solder is easy to rework to maintain a smooth finish solder bead.

50/50 Solder: This is composed of 50% tin and 50% lead. It is liquid at 421°F, solid at 361°F and has a pasty range of 60 degrees. This solder will produce a much "flatter" bead than 60/40. Because of its higher melting point, 50/50 solder is often used on the back (or inside) of a stained glass project to protect against "melt through" when soldering the front. Because it spreads and flattens out, 50/50 solder is often used when soldering lead came joints.

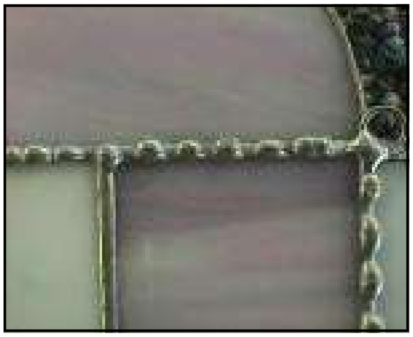

63/37 Solder: This solder is 63% tin and 37% lead. It becomes liquid at 361°F, and solid at 361°F, with a pasty or working range of 0 degrees. This solder is called a eutectic alloy which means at 361°F, you can go instantly from solid to liquid to solid just by applying or removing the heat source. You will often find "63/37" solder referred to as decorative or quick set solder. It is primarily used to create dimensional effects in the solder itself and can be "pulled" and manipulated to produce a variety of textures and designs. 63/37 solder also makes an excellent solder to bead up the outside rim of copper foiled pieces.

Lead-Free Solder: Depending on the specific mix of metals, lead free will produce differing liquid, solid, and pasty range temperatures. Check with the solder manufacturers for these specifics. Lead-free solders will perform similar to a 50/50 mix. Lead-free solders require more practice in order to obtain a smooth bead on copper foiled pieces and should not be used on lead came projects. Lead-free solder is the most expensive solder, but is the solder of choice when constructing pieces that will contact food, that will be handled frequently, or that will be used in a child's play area or room.

Solder Composition Reference Table

Alloy	Tin%	Lead%	Solid to	Liquid at	Pasty Range
50/50	50	50	361^{0}	421^{0}	60^{0}
60/40	60	40	361^{0}	374^{0}	13^{0}
63/37	63	37	361^{0}	361^{0}	0^{0}

Flux

Why Do We Use Flux?

Flux is a chemical compound that is used to promote the bonding of metals by removing the oxide residue simultaneously with the soldering process. Most metals left exposed to the air around us react with the air to form residue on the surface of the metal. The process is oxidization and the residues are oxides. Each mix of metals being joined has a specific flux that best promotes this bonding process. In stained glass, the metals being joined are primarily copper to tin/lead solder and lead, brass or zinc to tin/lead solder.

Types of Flux

Selecting the correct flux for your application is as critical as any other step of the soldering process. The proper flux will assure less soldering problems and a satisfactory solder bond. The best fluxes do three things:

- They remove all the residue that has formed on the surface of the metals you are going to solder.
- They prevent oxides from forming while you are soldering.
- Any post-soldering residue they leave is noncorrosive and easily cleaned off.

Fluxes are available in organic and inorganic forms. If you touch the flux and your skin seems to have a sensitivity to one type of flux, an inorganic type for example, try an organic variety. Often you will find that you are less sensitive to irritation by using the opposite type flux. Organic fluxes are generally some form of oleic (fatty) acids, while inorganic fluxes are most often zinc chloride based.

Characteristics of Fluxes

Liquid flux is the most widely used. It may or may not be water soluble. Water soluble fluxes clean up very easily, but are thinner and some have a tendency to evaporate quickly and require repeat applications. Some liquid fluxes are thinner than others and have a tendency to run or spread out from their point of application.

Gel Flux is generally water soluble and "adheres" well to the surface being soldered. Gels tend not to evaporate as readily as liquid flux, but if you apply too much, they produce more "spitting" and "popping" as you solder.

Paste Flux is very thick, stays where it’s applied, and doesn't evaporate easily. It's effective when soldering 3-D forms such as lamps. Paste fluxes can be difficult to clean. Be very careful not to over apply a paste flux as it will also spit and pop, leaving pits on your solder seams.

Getting the Most From Your Flux

It is important to keep your flux clean and free of impurities. This will ensure the purest possible solder joint. One way to keep your flux clean is to avoid dipping in and out of the original container. Pour out the amount you will need for your current project into a separate, smaller container. Don’t leave the flux bottle sitting around with the cover off. Never pour leftover flux back into the original container. Fluxes often will produce fumes as you solder. Make sure you are aware of and follow the precautions suggested in the health and safety section of this guide.

Soldering Basics for Copper Foil

The quality of the solder job can make or break your project, yet it is the step that takes the most practice to become good at. Soldering is more difficult than the other steps in stained glass because there are so many things that can effect it: the amount of heat, the amount of solder, the type of flux, the rate at which you move, gaps in the project, etc. The following step-by-step instructions will give you the basics of soldering. If you are just starting out, you will learn a lot. If you are experienced, you may just find the trick that makes it all come together. Great soldering takes a little bit of knowledge and a lot of practice.

Before You Start

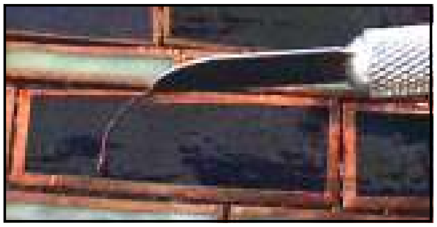

Preparation is the key to producing a good solder bead. Here are some suggestions to help you get of to a great start!

- Make sure that the foil is seated properly over the glass.
- Trim all overlaps in the foil using a sharp craft blade.
- Assemble all the tools you will need so they are at hand.
- Wear safety glasses.
- Have a proper ventilation set up.
- Make sure the foil is well burnished and sealed smoothly to the glass surface.
- Secure the glass pieces using pins, or squaring blocks made for this purpose before soldering.

Set aside enough time to solder your entire project all at once. If you don't, dried flux and solder residues on the unfinished parts will make further soldering difficult. Flux allowed to remain on a project can compromise the strength of the solder joints.

Cleaning Before Soldering

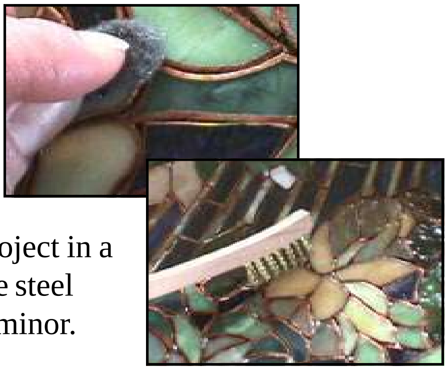

Sometimes the surface you are going to solder needs cleaning to remove visible grime, dirt, residue or oxidization. In this case, use 000 or 0000 steel wool to clean the surface of copper foil. For lead came or excessive oxidation carefully use a soft brass brush. If you won’t be able to solder your project immediately after it is copper foiled, store it in a plastic bag to help reduce oxidation. If you partially solder the project and have to stop before it is complete, be sure to clean off all of the flux and store the project in a plastic bag until you are ready to continue. You will probably still need to use steel wool on the joints when you are ready to solder, but the clean up should be minor.

Getting Started

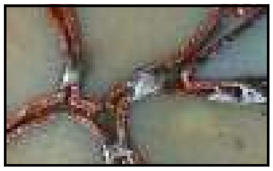

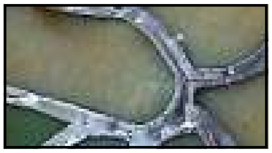

Begin the soldering process by fluxing all intersections of the glass project. Then "flat tack" solder these intersections together by using a very small amount of solder and the flat face of the iron tip. The solder should lay flat on the intersection and you shouldn't have any beads or bumps of solder on the piece. Hence the name "flat-tack" soldering.

Now, fill any gaps between the glass pieces by melting solder into them until the solder is level with the surface of the glass. If you have large gaps, you can ball up some copper foil, with the sticky sides together, and fill the gaps. This will help keep the solder from seeping out the other side.

"Tin" the entire perimeter making sure to work some solder into seams that intersect the perimeter. (Tinning is the process of coating a surface with a thin layer of solder. Apply just enough solder to turn the foil to the silver color of the solder.) This creates a "dam" that prevents solder from spilling over the edge when you solder the interior seams.

Next, flux an area you can comfortably finish soldering before the flux can evaporate. Start at the top of the piece and work down. This prevents you from dragging your sleeves and hands through the flux.

Running The Solder Bead

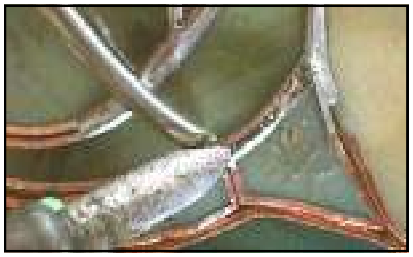

Hold your soldering iron in your writing hand. Position the iron comfortably so the face of the iron tip is perpendicular to the seam. If you hold your iron like a carving knife you will automatically put the tip in the best position for soldering. It will also keep you from "painting" with the iron. You will hold the solder in the opposite hand. Start soldering by placing the iron tip down on the foil seam with the wider face perpendicular to the seam and facing you. Touch the solder to the tip of the iron, just above the point where the iron tip touches the copper foil. Allow the solder to flow down the tip face onto the seam as you move the iron along the seam. Feed the solder at a continuous, even rate to create a consistent, rounded bead of solder over the seam. This takes practice!

If you apply too little solder or move too quickly, your seam will be flat. To correct this problem, allow the solder to cool, then resolder the joint applying more solder. On the other hand, be careful not to feed the solder too fast or move the iron too slowly, as this will cause the solder seam to bulge and flow out over the glass.

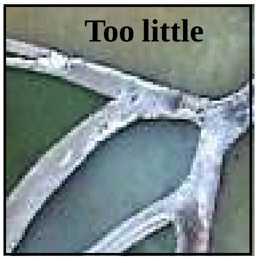

Too little

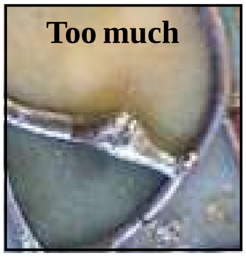

Too much

Fixing Imperfections

If you have too little solder, just add more, being sure to remelt the previous solder line as you apply another coat. If you applied too much solder the first time, it's easy to remove the excess with the following procedure. Clean your iron tip thoroughly on a sponge. With the iron tip positioned with the flat face down, move it across the seam, "pulling" the excess solder onto the glass. Clean your tip and repeat the process until all excess solder is removed. Another method for removing excess solder is to wait until all the seams have been soldered. Then place your project up on edge and run the tip through the seam allowing the excess to run off. Lay the piece flat, reflux, and rework the seam, adding additional solder as needed.

When you solder, work slowly enough to produce a good bead, but not so slowly that solder melts through seams to the back side of the project. Here's a hint: one way to prevent this is to put damp paper towel under your panel before you start soldering. Moving too slow also increases the likelihood of causing heat fractures on your glass.

Other Things To Consider

Always keep your tip clean by frequently wiping all sides of the tip on your damp sponge. Wait momentarily for the tip to reheat before continuing to solder. A clean tip maintains proper heat and removes impurities.

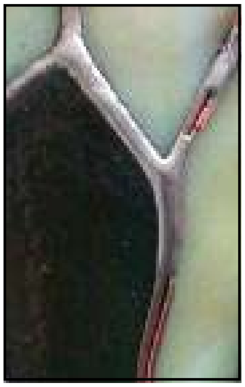

To achieve a smooth solder line, solder the longest continuous seam possible. Don't start or stop a solder seam at an intersection with another seam. As you meet an intersection, allow the solder to break or "y" out over the intersecting seam about 1/2" and then return to the original path you were soldering in one continuous motion. As you continue to solder seams, you will connect these branches, thus preventing pulling out at intersections of seams (which is a noticeable sign of a beginner). When you encounter a "v" type joint. Solder in a direction that moves up into the "v" point, instead of coming down onto the point. This keeps a nice clean "v" joint without excess solder obscuring it. Do this by soldering up one leg of the "v" and as you approach the apex, drop your iron tip down flatter on the glass, so that it crosses over both seams. Then continue soldering into and out of the point of the "v."

Avoid a "painting" motion when soldering. Visualize your iron as a magnet pulling along a metal strip. Maintain a small puddle of solder at the base of the tip as you solder. When two adjacent pieces of glass vary in thickness, ride the tip on the higher piece of glass as you solder.

Effects of using the iron like a paint brush

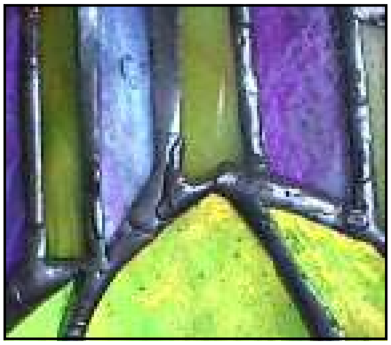

A big advantage of using copper foil is that not-so-perfect solder seams can be reworked. Allow the seam to cool, reflux it, then resolder. Be careful because reworking a seam too many times (more than 2-3) can cause the adhesive on the foil to loosen and bubble.

Solder one side of your project completely, then turn it over and repeat the process on the other side. Some of the flux from the first side you soldered will have seeped to the other side. You may encounter more "spitting," so use flux very sparingly. To achieve a smooth, beaded solder seam: practice! practice!

Ending A Bead

When you are ready to end your seam, pull off across the face of the glass making sure you are in a liquid section of the solder seam and do not lift straight up. This technique leaves a smooth exit point that is virtually unnoticeable. You may notice that this is the same photo used for removing extra solder. It’s not a mistake. The action is very much the same.

Practice!

Creating perfect solder seams takes practice. One good way is to make a practice board. Make sure to include all types of intersections and lines (curvy and straight). Foil the piece and flat solder it on the front and back. Practice each different soldering technique. When you are finished, you can hold the piece vertical and melt the solder off. Now you are ready to start again. Many people use the practice board before they start soldering each project, just to get the feel of soldering again.

Cleanup After Soldering

FLUX
REMOVER
inland

After soldering, you will need to remove all flux and solder residues from your solder lines and glass. The best cleaners to use are ones that neutralize the acid property of fluxes. To properly clean your project, you will need a good flux cleaner, very warm water and a soft scrub brush. Thoroughly brush the cleaner into every corner of your project, front and back. Follow this with a long rinse in clean, warm water and then towel the project dry. This step is very important because any flux that remains will continue to react with the solder, causing oxidation. If your project looks good after the initial cleaning, but the solder lines have a white or green growth on them after a few weeks, it generally means that all of the flux was not removed from the project. You can use 000 steel wool to remove the patina and finishing compound and thoroughly clean the project. Then reapply the patina and finishing compound.

A small stiff scrub brush is the best way to clean the entire project. After the initial cleaning, you may find some areas that need additional attention.

You may find small pieces of solder stuck to the glass. You can easily remove these using your fingernail, or 000 steel wool. Be careful not to scratch the glass.

An old toothbrush can be helpful for cleaning in tight intersections and right up against the solder line.

Occasionally you may find that after applying patina, certain types of glass acquire a “rainbow” effect. This is most likely to occur on hot colors, black, white and iridized glass. It is one of those things that is unpredictable and difficult or impossible to remove. You can try to remove it by using a paste made from baking soda, but it doesn’t always work. The only way to prevent this occasional problem is to be very careful not to get patina on your glass. Use a cotton swab to apply the patina only to the solder lines of the project. If you have this problem, it is a good idea to make note of it on any remaining glass, so you can be more careful the next time you use that particular glass.

Finishing

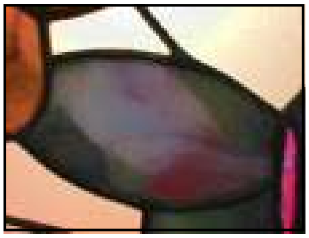

After cleaning your project you should use a finishing wax or compound to keep it looking it’s best. A good quality carnauba wax or pre-mixed wax and cleaner should be applied to all of the solder lines. A nice side effect is that the wax also enhances the color of the patina making it either rich black or shinny copper. You may find it helpful to reapply the wax after a few months. (If you would like to change the color of the solder, see the section about patina on page 16.)

Soldering Basics for Metal Cames

Lead came construction requires a different soldering technique than that used for copper foil. You can use the same iron and flux, but the preparation and actual soldering will differ. You will be soldering only the joints, instead of running an entire bead. With a little knowledge and some practice, you will be able to make perfect solder joints on lead or other metals.

Getting Started

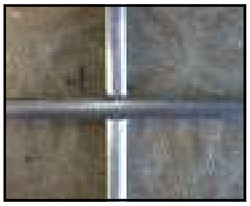

First, make sure that all joints are in line with each other and butt against each other with no gaps. Use a triangle and/or square to check the squareness of the panel and make sure that your lines run true across the panel in all directions. Make any adjustments that are needed. Thoroughly clean surface oxidation from intersections with a soft brass brush.

Soldering

The most desirable solder to use when constructing a lead project is 50/50. It flows out from the point of contact, leaving a nice joint. Do a "lead" test to check the heat of your iron. Allow the iron tip to heat for several minutes. Then, place the flat face of the tip on a scrap of lead. If the lead melts, the iron is too hot. Lower the temperature or change to a lower temperature tip. (This is where a rheostat iron is great.) Now test the tip with 50/50 solder to make sure it can melt it. If it does, your tip temperature is adjusted correctly. If not, raise tip temperature just until you melt the solder. Recheck the tip temperature and test it on the lead again. Now flux all the joints on your panel.

Begin by holding the solder on the joint. Move in to the joint across the glass, not from across the came. The wide face of the tip should be down toward the joint. Coming into the joint over the glass prevents leaving a solder trail on the lead came at one end of the solder joint. Techniques like these are the signature of an accomplished craftsperson.

Melt only a small amount of solder and move the tip in a slight circular motion allowing it to flow out over the joint a distance equal to the width of the came being used. Pull the iron tip straight up off the center of the joint. The desired end result is that intersections of the lead cames are not visible through the solder and the solder on them is smooth, not beaded. You should be able to run a fingernail over the came and into the soldered joint without it catching or clicking. Solder the joints on the front of the project. Then turn the project over and repeat the same process on the back.

Finishing

Lead projects need to be finished by applying a cement. The cement will weatherproof the project and secure the glass in the channels. This step is necessary on any came project to ensure a strong, secure panel. Cement is easily applied to the panel by using a small scrub brush to work the cement under the channels both on the front and back of the project. After the cementing process is complete, whiting sprinkled over the entire surface and allowed to absorb excess moisture for a few minutes. Using a stiff brush, vigerously brush the lead channels, removing excess cement from the glass. The brushing process allows a natural patina to form on the lead, so it is not necessary to apply patina to the project. Allow the project to dry thoroughly, usually for a few days, before installing or hanging the project.

Soldering Basics for Other Metals

Vase Caps

Vase caps are generally made from spun brass. The cap must be very clean of all dirt and oils for the solder to adhere. Use 0000 steel wool to clean the areas that you will be soldering. The cap also needs to be quite hot in order to achieve a good bond between the solder and the brass. An easy way to get the cap hot is to heat it in a hot oven for about 15 minutes. Use pliers to carefully remove the cap from the oven. Center the vase cap on the lamp. Flux all areas you will be soldering. (Paste flux is easy to use for this purpose, because it doesn't drip.) Get a small amount of solder on your iron and hold it on the vase cap for a minute or so at the area you want to make a connection to the lamp. When the cap is hot enough, the solder will spread out instead of lumping up. Add more solder to make a connection to the lamp. (Some vase caps have a lacquer finish on them. This must be removed at the areas you need to solder. Use a craft knife to gently scrape the lacquer in these areas.

Spiders

Spiders are made from brass as well. You can either cut the legs to fit, or bend them. Like a vase cap, you need to get the brass quite hot in order to create a good bond with the solder. Unlike vase caps, you can get the smaller area of the spider hot enough using just your soldering iron. Simply hold your iron tip on the area that you need to solder for a minute or so. Apply flux and a little solder, and continue to heat the spider with your iron. When it is hot enough, the solder will smooth out around the spider leg and adhere to the lamp. If the solder is globed up, the spider leg isn't hot enough. Continue in this manner for all of the legs.

Metal Outer Channels

When attaching a zinc, copper or brass edge to a project, use this solder technique: Cut the zinc, copper or brass came vertical members to overlap the horizontal members in a "butt" type joint. A butt joint provides structural stability and distributes weight. Before you solder the butt joints, you can use masking tape to contain the size of the solder joint.

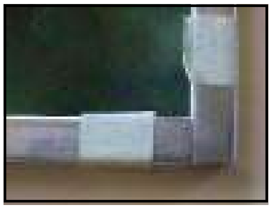

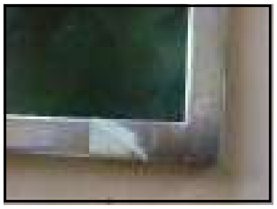

Clean and flux these joints and solder them to each other using the flat tip of the iron on the metal. The metal channel will take a little longer to get hot than copper foil. If the metal is not hot enough the solder will not smooth out. The metal frame should be attached to the panel by soldering it to any lead lines that it touches.

On leaded projects, solder from the outer metal edging onto the lead came, pulling the iron tip straight up over the point of intersection when completing the solder joint. This simple technique will level the solder smoothly over the joint.

For copper foil pieces, stop your solder bead about 1/4-1/2" from the outer edge to allow room for the metal edging to fit onto the perimeter of the project. If you forget to do this initially, flatten the solder bead so the outer edging fits onto the project. After fitting the outer edging material, flux and solder the corners as outlined above. Connect the copper foil beads that intersect with the outer edging metal by continuing the solder bead into the edging making sure that the solder adheres securely to the metal edge.

As you work more with your iron, different solders, lead came, and copper foiled projects, your skills will increase. You will quickly develop the techniques that will make you more confident in tackling increasingly complex projects.

Hanging Hardware

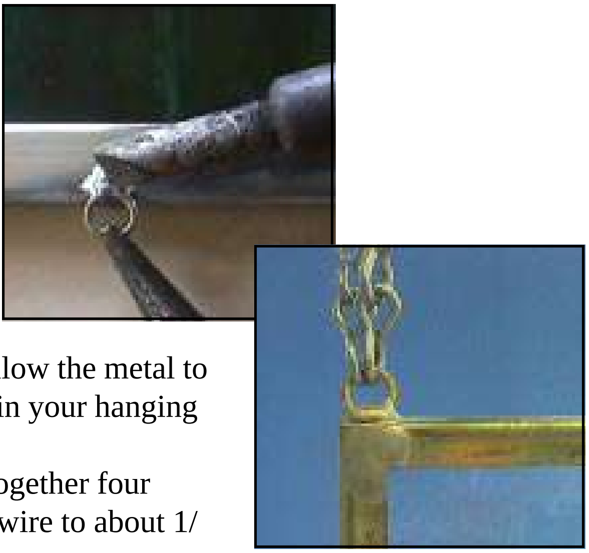

There are many ways to hang your project. You will want to make sure that the method you choose will support the weight of the project. You can purchase pre-made hardware from your stained glass store, or make your own using any type of solderable wire. If you are using wire look for 16 to 18 gauge.

Rings can be added to the metal border of the project if it is not too heavy. A general rule of thumb is about one square foot. Hold the ring with pliers and solder rings to the metal border of the piece wherever they allow it to hang properly. Just like soldering the channel together, if you don't allow the metal to get hot enough, the solder will not adhere well. This will result in your hanging rings failing.

A stronger type of hardware is made using wire. Twist together four strands of 16 -18 gauge wire, leaving a loop at the end. Cut the wire to about 1/2". Insert the wire into the opening of the metal channel, bending the loop the direction you need it to be. Solder the hanger to the channel and fill the opening with solder.

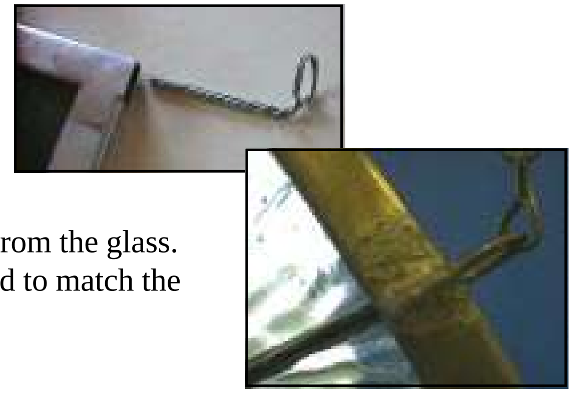

Another option is to make a loop from twisted wire and solder it to the front and back of the metal channel, with the loop extending above the project. Be sure to solder the hanger into an existing solder joint, not just to the channel. This will prevent the weight of the project from pulling the channel away from the glass.

By the way, the joint on the hanger at the right was painted to match the brass channel using a gold paint pen.

Patina

Solder and lead change color naturally over time, but you can change their color instantly by using patinas. There are several different formulas producing a variety of colors. Patinas are easy to use. Make sure that your project is thoroughly clean and free of solder, flux and cutting residues. Use a brush or soft rag to apply the patina to the solder until you reach the desired color. Wash the panel again to remove any patina from the glass.

- 👍When using antique brass patina, first scrub your solder seams with fine steel wool. Apply the patina sparingly with a clean brush, using only a few strokes. Over use will produce a coppery finish.
- 👎Never mix patinas and always use proper ventilation and skin protection.
- 👍To get a nice black finish on brass, add a dash of table salt to a little black patina in a glass or plastic jar. Mix well, apply and allow to dry. Clean and finish a usual.

Trouble Shooting

Problem: The solder won't adhere to the foil or came.

Solutions/Explanations:

1) Did you apply any flux?
2) Did you apply enough flux?
3) Was your iron too cool?
4) Is the foil or lead came too oxidized for flux to activate the surface?
    A) Buff copper foil with 000 or 0000 steel wool.
    B) Clean lead came with brass brush.

Problem: Solder is melting through to the backside of my project.

Solutions/Explanations:

1) Your iron is too hot. Depending on what type of iron you are using, either turn the temperature controller down or change to a lower temperature tip.
2) You may be soldering too long in each area. Move on and let the solder cool. Return to those spots later.
3) The gaps between your glass are too wide. Try to fill the gaps with solder and let it cool. Then go back and solder a bead on top of the filled gaps.
4) Try placing a wet paper towel under the project to provide cooling.

Problem: My solder is "stiff." I am unable to get a smooth bead.

Solutions/Explanations:

1) Your iron is too cool. Depending on what type of iron you are using, either turn the temperature controller up or change to a higher temperature tip.
2) You may not be using enough flux.
3) Is the tip of the iron getting hot? Check to see that the tip is seated correctly down in the barrel of the iron.
4) Are you "painting" with solder, instead of running a long bead with a steady, even, one directional movement? This is one of the most common mistakes that beginners make. In order for the solder to create a bead, it has to get molten. If you are using the iron as a “paint brush”, the solder is not getting hot enough to melt thoroughly.
5) You may be soldering with just a corner of the iron tip. Check for proper positioning of your iron.
6) Try a different solder with a lower melting point.

Problem: My iron seems to be hot, but the solder isn't melting as fast as normal.

Solutions/Explanations:

1) Check your tip. Is it loose?
2) Is your tip too dirty?
    A) Clean it on a wet sponge.
    B) Clean it on a tinning block.
3) Your tip is too corroded or defective. Replace it.
4) Your rheostat may be malfunctioning.
    A) Plug the iron into the outlet you are using without the temperature controller. If the iron works, the controller is malfunctioning. If the iron doesn’t work, the problem may be with the iron itself.
    B) With some types you can check the temperature controller by plugging a lamp into it. Turn on the lamp and see if the controller dims and brightens the lamp.

Problem: My solder seams are flat.

Solutions/Explanations:

1) You are not using enough solder. Reflux your project and add more solder.
2) You are soldering too fast. Move at a speed that lets the solder bead.
3) If the bead was there, but disappeared, you may be working too long in one area causing the solder to melt through to the other side. Allow the area to cool down before trying again. Another option is to place a damp paper towel under the seam you are working on. This will help keep the glass cool, allowing you a little more time to work.
4) Check the type of solder mix you are using. For copper foil work, 60/40 will help produce a higher, more rounded bead. Don't make the mistake of buying 40/60 solder. It has an even longer pasty range than 50/50, and is not recommended for stained glass work.

Problem: My solder seam is too wide.

Solutions/Explanations:

1) Do you have large gaps between the pieces you are soldering?
    A) You may have to recut some pieces.
2) You may be using too much solder causing the seam to bulge over onto the glass. "Pull" or "bleed" the excess solder from the seam.
3) Is the copper foil too wide? This is another common mistake for beginners. Remember that the width of the solder seam is determined by the width of the foil. An attractive solder seam is accomplished by using a foil that shows about 1/16" on both sides of the glass.

Problem: My solder seams are irregular in width.

Solutions/Explanations:

1) You may have poorly fitting pieces that create different size gaps in your project. Recut these pieces.
2) Your foil may be applied unevenly on each side of the glass in some places. You may try to correct this by trimming the foil with a craft knife. If that does not work, remove the foil and start again.

Problem: My glass fractured while I was soldering.

Solutions/Explanations:

1) Your iron is too hot. Depending on what type of iron you are using, either turn the rheostat down or change to a lower temperature tip.
2) You "worked" too long in one location with the iron.
3) There could have been a small chip or crack in the glass which was expanded when it was heated by the iron.

Problem: My solder bead is not bright and shiny. It appears dull and splotchy.

Solutions/Explanations:

1) The solder was applied too cold, so it never fully reached its "liquid state." Turn up the temperature if you are using a rheostat or change to a higher temperature tip if you are using a temperature controlled tip.
2) You may not be using enough flux to "wet" your foil or lead and create a proper solder condition.
3) If you're using 50/50 solder, try 60/40. Because solder crystallizes as it cools through the pasty range, 50/50 is more prone to having a textured look on its surface because of its considerably larger pasty range.
4) The solder you are using may be of inferior quality.

Problem: My solder is "spitting" as I work. What's wrong?

Solutions/Explanations:

1) You are likely using too much flux. It is literally boiling when you apply the solder. Wipe some of the flux off with a paper towel and try soldering again. You may find areas that now need a little more flux, because you removed too much.
2) If the spitting only occurs on the back side of your project (or the second side you are soldering), the problem is still too much flux, but the cure is different. When you use too much flux on the front side of the project, the excess to flows through the panel, so the spitting isn’t as bad as it could be. When you use too much flux on the back side of the project, it can’t flow through to the front, because the front is already sealed. The obvious cure for this problem is to use less flux the next time. But what about now, when you already have too much and can’t remove it from between the glass? The only good solution is to apply your solder, and allow the flux to boil up and out of the seam. (Keep your face as far away from the project as you can.) Once it stops spitting, you can remelt the solder and make it look as good as new.

Problem: I have gooey, glue residue along my solder seams that won't wash away.

Solutions/Explanations:

1) Adhesive from foil may be working up and out onto your glass. This can be caused by poor foiling or burnishing of the copper foil, which allowed flux to seep underneath the foil. To correct this use a cuticle stick or soft toothbrush to clean adhesive away.
2) Your iron is too hot, causing the adhesive on the back of the foil to melt and seep out onto the glass. Depending on what type of iron you are using, either turn the rheostat down or change to a lower temperature tip.
3) If you reworked a seam too often, it may have loosened the foil. Allow the seam to cool to the touch and gently press the foil back down to the glass.

Problem: Attached hooks and rings are pulling away from my project.

Solution/Explanations:

1) Hooks and rings should never be attached to just a foiled edge. Attach hooks/rings to vertical seams in the piece or at a juncture between a vertical seam and the perimeter.
2) If there is no vertical seam or intersection, then attach the ring/hook to a horizontal seam in the piece.

Problem: My solder has a white chalky growth on it. What did I do wrong?

Solution/Explanations:

1) All of the flux was not washed off of the project after it was soldered.
2)The project was not sealed with a finishing compound or wax, or it has worn off. Use 000 steel wool to remove the patina and any remaining wax. Clean the project thoroughly paying special attention to the corners and edges. Reapply the patina and wax.
3) The project has been exposed to the outside elements. Most waxes are intended for indoor use only. Others can be used outside, but must be reapplied regularly. Projects that will be exposed to the elements are better constructed using the lead came method.

Glossary of Soldering Terms

Antimony

An element used in the production of some solders. It should be avoided for use in stained glass.

Ceramic Heater

A type of heating element comprised primarily of ceramic, noted as extremely fast heating and efficient.

Chisel Tip

A soldering iron tip shaped like a chisel tool. This is the most common shape used for soldering stained glass. Chisel tips are made in a variety of sizes, the most common being 1/8”, 1/4” and 3/8”.

Heat Sink

A device used to draw or absorb heat being generated by another source. For example, an object being soldered acts as a heat sink to the soldering iron.

Decorative Soldering

Any decorative effects created in solder. These effects are created with a soldering iron, usually with a very narrow tip. Special solder, like 63/37 or QuickSet, make it easier to create special effects because it has a “zero” pasty range.

Eutectic Point

An exact single temperature point at which an alloy goes from solidus to liquidus with no pasty range. For example, the eutectic point of lead and tin is 361^0 F. This point is obtained only by 63/37 tin/lead alloy.

Flux

A chemical agent used to remove compounds from the surface of metals during the soldering process. (See organic and inorganic flux.)

Idle Temperature

A very low temperature at which the iron is on, maintaining the capability of a more rapid heat up than if the iron was off or “cold”. This is usually between 200^0F and 300^0F.

Inorganic Flux

A flux comprised of one or more inorganic salt such as zinc chloride or ammonium chloride. Inorganic fluxes are more corrosive and conductive than organic fluxes,. They are effective on all common metals.

Leaded Solder

A material used to join metals comprised of tin and lead.

Liquidus

The temperature at which a pure metal becomes completely molten or liquid.

Mica

A mineral based material used as a construction component in wire wound type heaters of soldering irons.

Organic Flux

Organic fluxes are not as corrosive as inorganic fluxes. They are often used when the surface of the glass may be effected by the flux, such as painted glass.

Oxides

Debris on the surface of a metal which is the result of the reaction with chemicals in its environment. Oxides must be removed mechanically (with steel wool or a brass brush) and chemically for proper wetting to occur.

Pasty Range of Solder

The temperature range which is the difference between the solidus and the liquidus temperatures. This is sometimes referred to a the “working range”.

QuickSet™

Another name for 63/37 solder. See decorative solder.

Recovery Time

The time required for a soldering iron to reach soldering temperature after it has hit a “cold spot”.

Rheostat

The term applied to a soldering iron control which is used to vary the temperature of the heater in the iron.

Solder

An alloy of two or more metals with a liquidus temperature of less than 800^0 F.

Solder Bead

The term used to describe the look of solder when it has been properly applied to a copper foil seam.

Solidus

The temperature at which a pure metal or alloy goes from liquid to solid (or “freezes”).

Temperature Control

An electrical or electronic device into which a soldering iron is plugged, or which is within a soldering iron. The device is used to vary the temperature of the soldering iron heater.

Tinning

Applying a thin layer of solder to a metal surface. Most commonly used in reference to the copper foil technique.

Tinning Block

A block of sal-ammoniac on which a soldering tip is cleaned and resurfaced with a layer of solder. It is not actually used for soldering.

Tip Tinning

The process of renewing a soldering iron tip using a sal-ammoniac (tinning) block.

Wetting

The term used to describe the proper flow of liquid solder, promoted by flux, on the surface of another metal. “Wetting” is necessary to form a proper soldering joint.

Wire Wound Heaters

Soldering iron heaters constructed with heating wires wound in a coil around Mica.

Soldering & Desoldering

Soldering/Desoldering Tools & Accessories

Soldering Station with LCD Display, ESD Safe 400 Series

17400

- High-quality sensor and heat exchange system guarantee precise temperature control at the soldering tip
- Soldering iron tool heats rapidly from room temperature to 350C in 30 seconds
- Suitable for general purpose soldering as well as specialized lead-free soldering applications
- Sponge and coiled brass tip cleaner ensure a clean soldering tip free from debris and corrosive material
- Strong, deluxe iron holder provides safe secure place to store heated iron between uses

Product ID	Description
17400	Soldering Station with LCD Display, ESD Safe 400 Series

Soldering Station with LCD Display, ESD Safe 405 Series

17405

- LCD temperature control display with blue backlight
- Long lasting, durable high - quality ceramic heater
- Sponge and recessed area provides a convenient method for cleaning the tip
- Suitable for general purpose soldering as well as specialized lead-free soldering applications
- Safety guard iron holder

Product ID	Description
17405	Soldering Station with LCD Display, ESD Safe 405 Series

Soldering Kit - 6pc

17501

- 40 Watt Soldering Iron heats up to 900° F
- Ideal for use on electrical/electronic components, audio equipment, etc.
- Includes desoldering pump for removing liquid solder
- Provides basic tools needed for many soldering projects
- 6 Pc kit comes complete with replaceable fine tip

Product ID	Description
17501	Soldering Kit - 6pc

Soldering Kit - 7pc

17502

- 40 Watt Soldering Iron heats up to 900° F
- Ideal for use on electrical/electronic components, audio equipment, etc.
- Includes desoldering pump for removing liquid solder
- Provides basic tools needed for many soldering projects
- 6 Pc kit comes complete with replaceable fine tip

Product ID	Description
17502	Soldering Kit - 7pc

Soldering Iron 80W with Fine Tip

17510

- Ideal for general purpose soldering applications including metal board, tube or large components
- Round, soft grip, non-slip handle
- Durable cord with a 5-foot reach
- Fitted with long life replaceable fine tip

Product ID	Description
17521	Soldering Iron 40W with Fine Tip

Soldering Iron 40W with Fine Tip

17521

- Ideal for general purpose soldering
- Round, soft grip, non-slip handle
- Durable cord with a 5-foot reach
- Fitted with long life replaceable fine tip

Product ID	Description
17521	Soldering Iron 40W with Fine Tip

N-Series Soldering Iron Tips

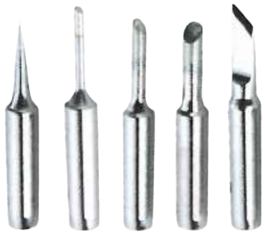

- Long life, lead-free soldering iron tips
- Replacement tips for soldering the different components of PCBs
- Five different sizes
- Includes two tips
- For use with 17400 & 17405 ESD Safe Soldering Stations with LCD Display

Product ID	Description
17400-N9-1	N-Series Soldering Iron Tips (N9-1)
17400-N9-2	N-Series Soldering Iron Tips (N9-2)
17400-N9-3	N-Series Soldering Iron Tips (N9-3)
17400-N9-4	N-Series Soldering Iron Tips (N9-4)
17400-N9-5	N-Series Soldering Iron Tips (N9-5)

C-Series Soldering Iron Tips

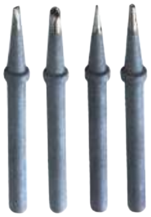

- Long life, lead-free soldering iron tips
- Replacement tips for soldering the different components of PCBs
- Four different sizes
- Includes two tips
- For use with 17521 Soldering Iron 40W

Product ID	Description
17521-C1-1	C-Series Soldering Iron Tips (C1-1)
17521-C1-2	C-Series Soldering Iron Tips (C1-2)
17521-C1-3	C-Series Soldering Iron Tips (C1-3)
17521-C1-4	C-Series Soldering Iron Tips (C1-4)

B-Series Soldering Iron Tips

- Long life, lead-free soldering iron tips
- Includes two tips
- For use with 17510 Soldering Iron 80W

Product ID	Description
17510-B5-1	B-Series Soldering Iron Tips (B5-1)

SMD Soldering

Your reference guide to soldering with surface mount devices

Edited by Lim Siong Boon, last dated 06-May-09.

email: mail@siongboon.com
website: http://www.siongboon.com

Short cut to your reference guides and charts

1. Introduction to SMD soldering
2. SMD soldering (prototyping board)
3. SMD soldering (prototyping PCB socket 2.54mm pitch)

1. Introduction to SMD soldering

SMD IC stands for surface mount device integrated circuit, or commonly known as SMT surface mount technology. Why go into SMD? Small in size, light in weight. This is the advantage of smd. Being small, engineer is able to design small electronics gadget that people can carry around. It is quite obvious that people prefer small mobile phone.

The ultimate advantage will be lower cost to consumer. Being small in size, the printed circuit board (pcb) can be small, meaning low cost. Transportation cost can be reduce because it is now lighter and more products can be packed into the same box. Space required is less, meaning cost saving on the warehouse storage. PCB board is easy to route without the through hole. Better signal integrity. Easily assembled by machine. There is too much advantages. Going SMT is certainly the way forward.

Technology evolution, from big to small.

A phone small enough to carry around.

Smaller mobile phone

Even smaller phone

Mobile phone as thin as biscuit.

Through hole packages

	DIP - 8 pins

I used to build circuits using dual in-line package (DIP or DIL) IC only. I hardly thought of using surface mount device/component (SMD) because I have great doubt that I can solder the fine legs of the IC chip. The pitch for the older DIP IC package are 2.54mm. Pitch is the distance between adjacent pins as illustrated in the picture. Most IC comes with a variety of packages. The IC that I know of, offer both the SMD and DIP packages. I thought that I will never ever need to use smd packages until one day, I have no more choice. I managed to pick up new circuit designs, and discover more and more new IC chips. Most new IC chips design do not have DIP packages. You can hardly find one. I realised that in order to implement better circuits, I need to learn to use new IC chip. In order to use the new chip, I have to find ways to solder smd components. That's where I venture into the world of smd.

Nowadays people prefer to use smd, because they are small in size, which turns out to be cost efficient fabricating small PCB. With smaller PCB, space & weight is saved, resulting in lower cost for the transportation/distribution and storage.

Research and research, I got to know from other electronics guru that smd is in fact easier to work with than through hole. I try it out and from that day onwards, I am in love with smd components. Small to solder, but it save me the effort to cut the lead for through hole component. I managed to store my component using minimum box and space. There are many many advantage to work with smd.

Many people might think that you need special tools like a rework station or fine tip iron to solder the small smd components with small pitches. With the correct technique, you can use your soldering iron to do the soldering. In fact my 60W goot TQ-95 soldering iron has quite blunt tip. So thick that most people think that it is not possible to solder smd IC with fine pitches. Fine tip iron is easy to reach the fine lead, but I find that it is not as hot as the same iron with blunt tip. If you prefer fine tip, I will recommend hotter soldering iron or those which can allow you to adjust the temperature, they can be hot. I managed to solder smd package TQFP, SSOP QSOP with pitch as small as 0.5mm. Width of the lead of about 0.3mm.

Dealing with SMD components do not necessary requires you to fabricate a professional pcb board. You can also mount smd components on low cost prototyping board. To start off implementing smd components, you might like to try using the soic package. Quite common at this point in time, but may just phase out as what has happened to the dip packaging. SOIC has a pitch of 1.27mm, which is exactly half of DIP packages. This size is great because I can solder the IC to the same old 2.54mm pitch prototyping donut padded board. What I usually do is to cut the donut pad into half. Soldered onto each pad is two pins 1.27mm apart.

You can refer to the following article, for further illustration.

"SMD soldering (prototyping board)"

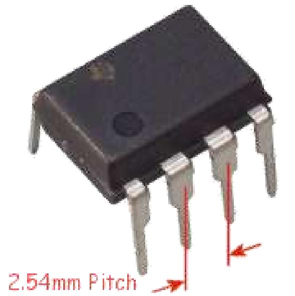	
	resistor
	capacitor
	inductor
	diode

SMD packages

1.27mm Pitch	SOIC - 8 pins
0.65mm Pitch	TQFP - 44 pins
	resistor
	capacitor
	inductor
	diode

With so much advantage there is indeed a disavantage. SMD IC comes with many type of packages with different pitches. Unlike DIP IC, the pitch is typical 2.54mm. I can easy purchase a prototype board with 2.54mm, and almost all the DIP will fit to the board. SOIC package can still be mounted onto the 2.54mm prototype board. The rest of the SMD IC chip has quite a wide range of completely different packaging. This make them difficult to start with, without having to fabricate a PCB.

There are actually solution to this. Most IC company usually introduce their IC chip together with a prototype board for the IC.

Another solution will be to use a pre-fabricated prototyping board that allows mounting of various smd footprint. You can search for the various names such as

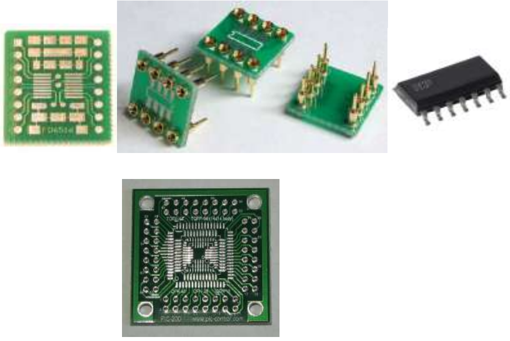

Various distributor for the prototyping adaptor

- prototype adaptor
- chip adaptor
- smd prototyping board
- smd to dip adaptor
- smd socket
- smd adaptor
- smd to 2.54mm converter
- SMD to DIP converter
- chip carrier

You can refer to the following article, for further illustration.

"SMD soldering (prototyping PCB socket 2.54mm pitch)"

In prototyping with DIL IC, pcb mount IC socket is usually used, so that the IC can be removed easily if damaged. Some sockets are designed for programming used where the chip can be inserted and remove easily for programming purpose. To remove the component, you need not have to do de-soldering. Like DIL package, smd also comes with their own IC socket. There are many variety of smd packaging, and getting the socket for your smd components is not going to come cheap & easy compare to DIL. Therefore I usually built prototype without any sockets for smd design. After numerous attempts designing and building of circuits, you will definitely attain a certain level of confident on your design without considering IC socket. Without the socket, cost and space are saved significantly.

SMD IC sockets:

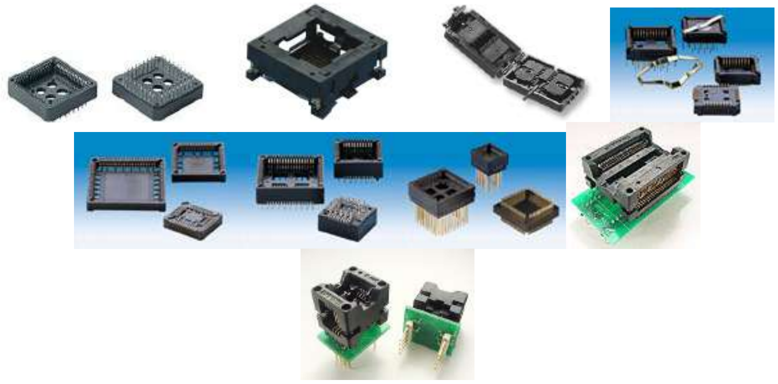

The headache will come when you really need to removed the smd IC. How do you remove the IC from the PCB? For de-soldering of smd component more than 3 leads, you can add in more solder to connect up the IC pins heat pins on all the sides of the IC and eventually pull out the component. It is easier to use the rework station or IR heater to de-solder smd components. The rework station uses hot sir to melt down the solder on the board. You will need to aim the hot air at the solder joint.

Rework Station

If you do not have the lurxy of getting the rework station, another possible method might be to use a special solder that has lower melting point. LowMelt® DeSolder Wire. A lower melting point means that the solder will takes a longer time to cool down to a solid. This allows you more time to remove the IC after heating up the solder. The solder is flooded onto the pins of the IC. The pins are heated up together using a soldering iron and the smd IC can then be removed easily with a tweezer.

BELLIN DYNAMIC SYSTEMS INC
ROTH ELEKTRONIK®
AVRMALL.com 개발보드전문 전자쇼핑몰
FOREST CITY PCB
Elk Tronic
PIC control
Farnell
RS

DIL IC sockets/ holder. Turn pins designs is of better quality.

Various distributor for SMD socket

Ironwood Electronics
Logical Systems
ARIES ELECTRONICS, INC.
AVRMALL.com 개발보드전문 전자쇼핑몰

Various type of smd packaging

* SOT, SC

* SOIC, ExposedPad™

* SSOP, ExposedPad™

* TQFP, ExposedPad™

* TSOP

* TSSOP, ExposedPad™

Low melt solder wire

* PLCC

* LQFP, ExposedPad™

* LGA

* PBGA

* PSOP

* Dual Row MicroLeadFrame®

* LQFP PowerQuad®

* MicroLeadFrame® (MLF®)

* MQFP PowerQuad®

* PSvfBGA

* SuperBGA®

* TapeArray® BGA

* tsCSP

* Ultra CSP™

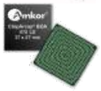

* CABGA, CTBGA, CVBGA

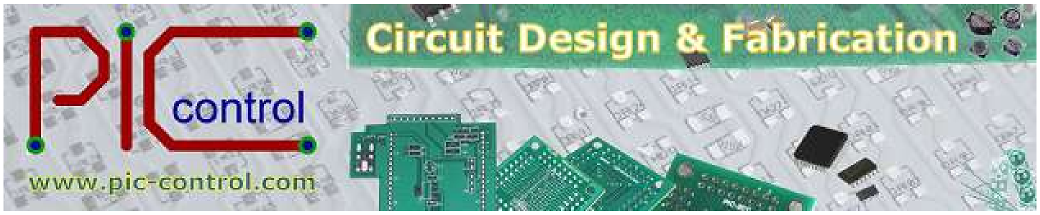
PIC control
www.pic-control.com
Circuit Design & Fabrication

Singapore Customized, custom made Electronics Circuits & Kits

2. SMD soldering (prototyping board)

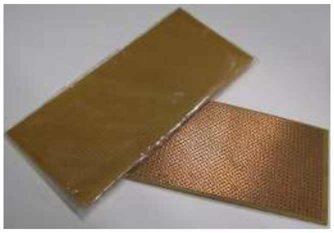

The prototype board above has pad with 2.54mm pitch and is designed for DIL and through hole components. The board is easily available from local store. Dealing with smd components on this 2.54mm pitch board is not difficult as well. I have been using this board to solder quite a number of surface mount components.

- SOIC
- DPAK
- SOT23-6, SOT-3
- TO-263
- PSOP

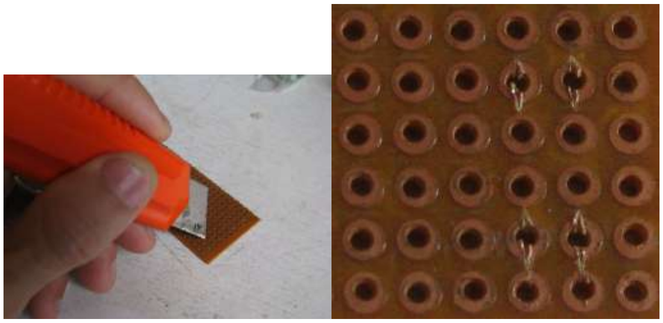

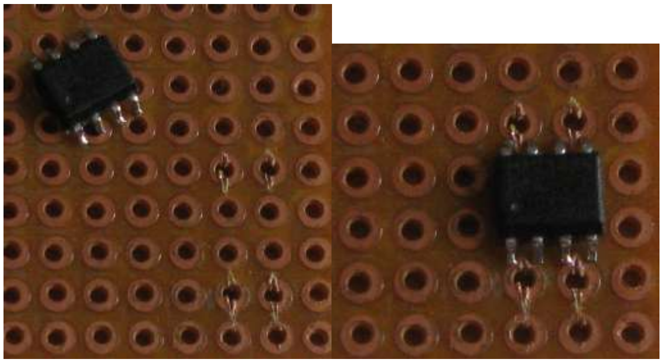

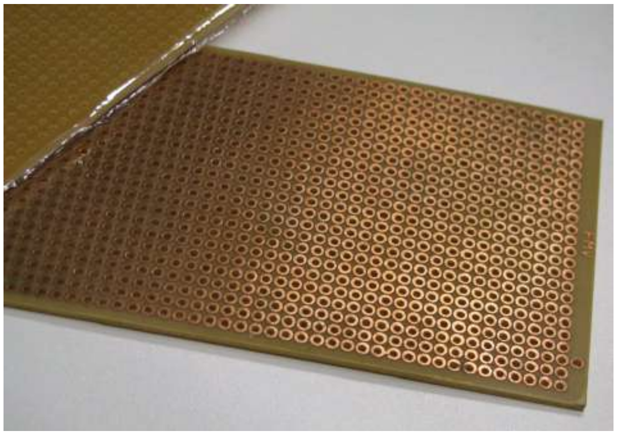

High pin count component with pitch less than 1.27mm will not be suitable to mount onto this board. For passive smd component like resistor, capacitor or inductor which only have 2 pin, mounting is simpler. Size is not an issue soldering onto the board.

Align the SMD IC to the position on the pad. Mark out the pad for the cutting to be done.

Use a penknife to cut the donut pad into half. Ensure that the half pad does not short circuit using your multi-meter. Sometimes the cut may not be deep enough or conductive reside between the gap resulting in a short circuit. Remember to ensure that each is proper cut before you start to solder.

Align the smd with your finger or tweezer and begin soldering one of the pin. Check if the IC is still properly align after soldering the first pin. If not, heat up the soldered pin and realign again. After the ICis align, solder all the rest of the pin. If the solder bridge across to the adjacent pin, use a soldering sucker to remove the excess solder.

The cut between the pad usually make the solder harder to bridge across the adjacent pin. Bridging is still possible, and it is quite easy to have it removed.

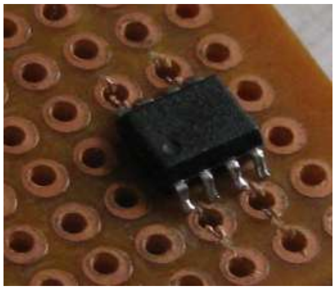

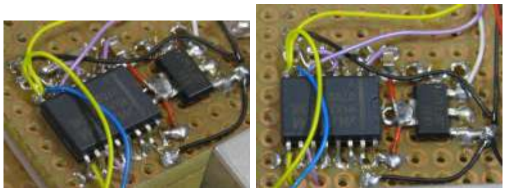

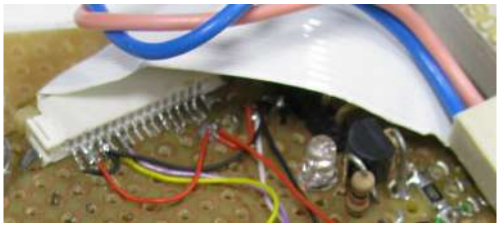

SMD ribbon cable connector being soldered onto the 2.54mm pitch prototype board.

Example of building the circuit using surface mount component on the prototyping board with 2.54mm pitch pads. As shown above, the pins are being wired by fine wire wrapping wire. The single core wire is very flexible and comes in a variety of color, making it easy to identify the type of signal being carried by the wire.

3. SMD soldering (prototyping PCB socket 2.54mm pitch)

What you need...

List of items:
- Surface IC chip (TQFP package)
- Surface mount IC adaptor or converter
- Soldering Iron
- Tweezer or Pincer
- Solder with flux core
- Soldering Flux
- Stick for coating flux
- Wiping paper
- Tinner

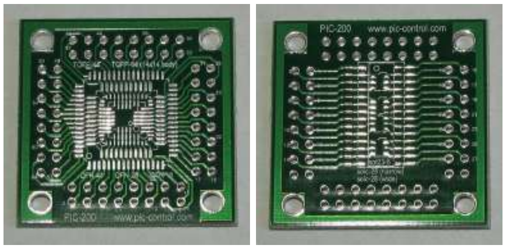

PIC-200 smd to 2.54mm pitch adaptor

from .

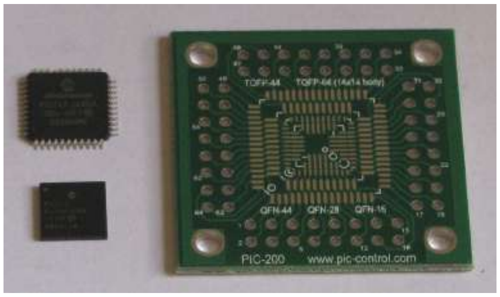

INTRODUCTION

In this setup, I am going to solder a surface mount microcontroller (TQFP 44 pins) onto a smd adaptor. The adaptor will then be used on the commonly available prototyping board, so that I can try out my new design for circuit control.

TQFP, QFN and SOIC are quite some common footprint for prototyping smd microcontroller from 18 to 64pins. This prototyping adaptor pic-200 (on the left) that I have is an ideal adaptor for my microcontroller. Converting the smd device into 2.54mm, I can easily mount the smd unit onto my 2.54mm pitch prototyping board. It is available locally.

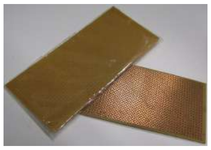

Prototyping board with 2.54mm pitch donut pads.

The pic-200 smd adaptor originate from PIC-CONTROL and is well documented. Commonly used smd footprint can be found on either the top or bottom side of the pcb adaptor. These are some of the smd packages that can be mounted onto pic-200 adaptor.

IC	Front	Back
TQFP-44		

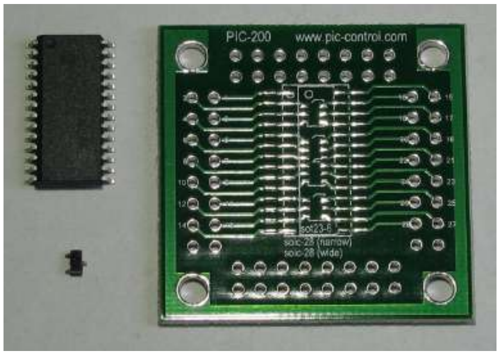

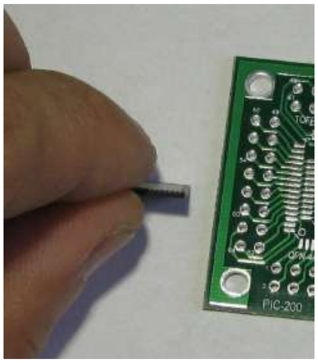

Holding onto the side view of the QFN package IC.

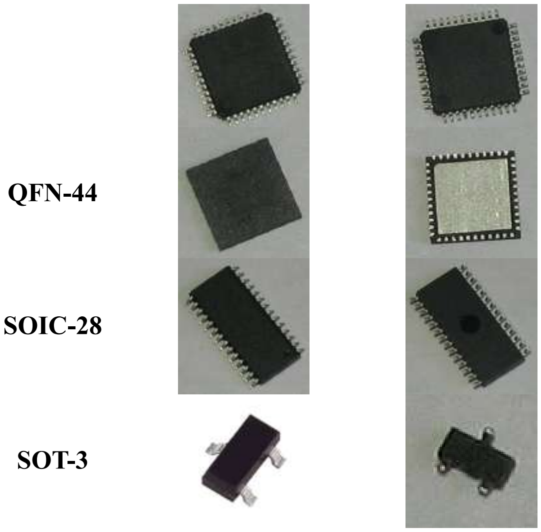

With a wide range of smd components that can be mounted, it is easier to stock up this pcb adaptor for future prototyping use.

The following article is about getting my smd microcontroller PIC24FJ64G004 soldered onto this board. After which I will be on my way to do prototyping for my new control circuit.

This smd adaptor origined from PIC-CONTROL. For other type of adaptors, you can also refer to other manufacturer as mention above.

click here to
Buy this SMD Adaptor Now at the PIC-store
PIC-store

<Step 1>

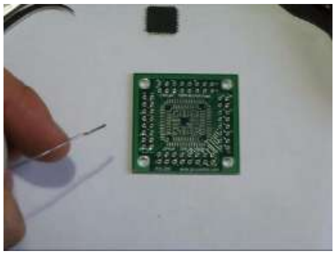

Before soldering, ensure that the soldering pad on the PCB is free of oxidation. PCB is usually tinned to protect the copper surface from oxidation. In cases where the surface is oxidized, you can try cleaning the surface with a contact cleaner. Apply the solution with tissue paper or toothbrush to clean up the contact. This solution make the PCB board rather oily, which I don't quite appreciate.

Contact cleaner.

Let us start our smd soldering here. Soldering surface mount component. The first important thing to introduce would be the solder.

There are many type of solder. Choosing the correct solder takes a bit of some knowledge too. Leaded or lead free solder? The industries is moving towards Pb free components & PCB, calling for unleaded product. Lead is toxin and not environmental friendly. I would advise ou to wash your hand after soldering or handling pcb board. A soldering fume extractor to vacuum the smoky fume would be quite comfortable for your soldering. The smoke is quite choking and a health hazard if you breath that in.

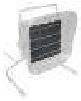

soldering fume extractor.

Solder also comes with a variety of metal alloy combination. Nowadays it is common in solder core to contain flux. If the solder do not have the flux in the core, the flux have to be manually applied onto the metal surface before the soldering beginning. The flux cleans up the soldering surface, and also make the solder flow rather smooth like liquid. One of the important ingredient to make a good solder joint lays in the soldering flux.

Solder wire also comes in a wide range of diameter to choose from. Generally I use 0.8mm soldering wire roll with flux at the core. Smaller diameter solder wire allows you a better control of the amount of solder to apply. For soldering larger pins or connector component thicker diameter wire will be preferred. Wire too thin is rather troublesome, because you need to feed in a longer length of solder in order to solder the large surface/lead. For general purpose PCB board soldering 0.8mm wire fits somewhere in the middle. Not too thick nor too thin. For smd components, diameter smaller than 0.8mm is preferred. Using 0.8mm for smd soldering is not of any issue.

Solder wire 2mm

1.2mm

1.0mm

0.8mm

0.5mm

<Step 2>

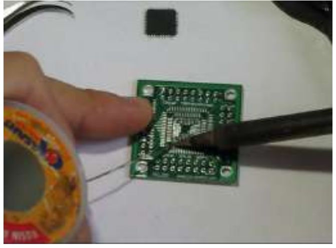

One of the pad being tinned.

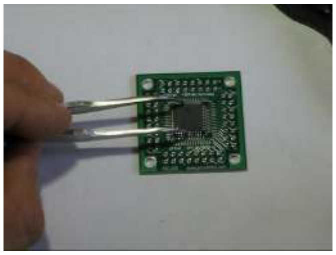

Align smd IC to the respective pad.

My microcontroller TQFP package has 44pins around the square package. To secure the package to the adaptor, I will need to solder one of the pin to the board first. It is important to solder only one pin. In case of mis-alignment, we can correct it easily. If more pins are soldered, it will be more difficult to correct any alignment.

Pin 1 of the adaptor pad is first tinned with a layer of solder. A tweezer is then used to align the IC chip on the pad. The IC is secure to the board with a slight touch from my soldering iron, soldering the IC pin to the pad.

The important factor to consider would be the heat generated by the iron. I prefer to use a 60W iron, which still works well in air-con room. A lower wattage or fine tip iron will not heat up that well in colder environment.

Goot TQ95 60W

If you prefer a fine tip, perhaps the soldering station will be more suitable. The iron temperature can be adjusted to a higher or precise temperature for the component that you want to solder. Component with thick metal contact like connector, heat sink, for thick cables should be solder with a higher temperature. The iron should be held on longer, in order to have the component fully heated up before any solder is applied.

Solder Station

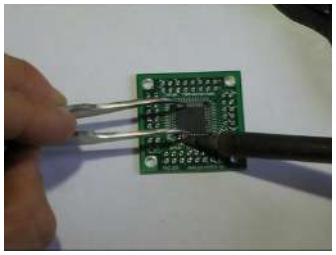
IC being fixed in position

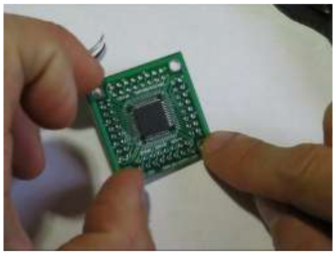
Checking if alignment is correct.

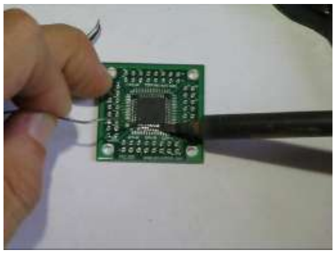
Flooding the pins with the rest of the solder.

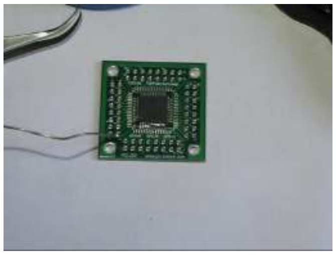
Pins flooded with solder

For soldering the IC pin to the pad, you can start off by heating up the pad. Then while heating up the pad, touch the IC pin and apply the solder wire directly to the pad or IC pin, allowing the solder to flow like liquid. The IC pin is heated last to minimize the chance of over heating the pin. If it doesn't flow very well, you can manually some flux to the joint. The solder wire should not touch the iron directly, because this will vaporize the flux, and solder flow will be restricted.

Heat up the lead and pad longer allows the solder to flow and form a good joint. Not for too long as it can also damaged your IC. You should be able to see the solder flow eventually onto the lead & pad. With a bit of practice, you will see the difference between a quick solder touch and another one with lead & pad heat up a little bit longer. It takes a number of practice to get a good joint with minimum amount of heating time.

The soldering iron that I use has rather thick tip, but that does not matter when I do soldering for the smd microcontroller. All we do is to flood the pins with solder. They can be be suck up later.

There is a technology known as ultrasonic soldering. The soldering iron tip has this micro vibration which helps the solder to flow more easy without much use of flux. It can also solderthose difficult to solder material. The following is one article that I found that describe about it. - Soldering the unsolderable.

<Step 3>

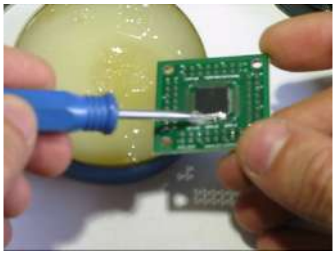

I have applied more flux at this stage. It should have been applied before flooding the pin, but it doesn't matter too much. The flux will make the solder rather liquid so that it is easier to suck up the excess solder.

solder flux paste

Baker Soldering Fluid

<Step 4>

The flooded solder is being heated up, and the excess solder are being sucked up.

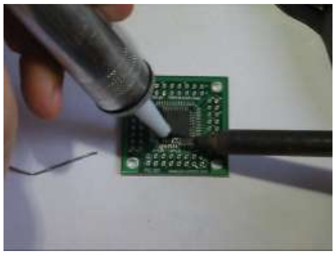

<Step 5>

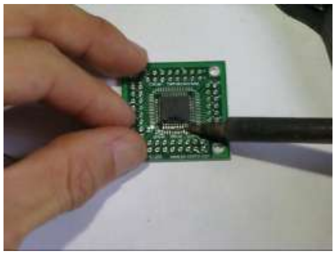

Each pins is being touch up, ensuring that the pin is solder to pad and that they are not bridged across to the adjacent pins.

Now we have got our smd component soldered onto the board.

<Step 6>

The flux helps a lot in doing a good solder but it often leaves the board with tiny dots of transparent strain, a bit sticky, also known as grease or flux residue. The board looks untidy with these tiny bits of grease.

These are the various recommended cleaning solution for pcb.
- 50% Alcohol + 50% Water
- Thinner
- Flux remover
- Multicore Prozone MCF800
- Solvent-> Bromopropane; propane, 1-bromo-; propyl bromide (Chemical Formula: C_3H_7Br)

Alcohol MSDS (Material Safety Data Sheet)

Thinner MSDS (Material Safety Data Sheet)

ELECTROLUBE
FLU

Electrolube, FLU Fluxclene Cleaning Solvent, pdf file

Industrial Size
HEAVY DUTY FLUX REMOVER
HEAVY DUTY FLUX REMOVER

MG Chemicals 413B Heavy Duty Flux Remover

MGC-850 Stainless Steel Cleaning Brush

Brush for the crush

Flux Remover

Multicore Prozone MCF800

Besides using chemical, or after using chemical, there may still be some white blur stain. Those are flux spread dry up on the surface. Using a hot air blower, or heater on the surface can melt the dry up flux, and recover the smooth shiny PCB board surface again. Clean away the melt flux immediate after the heat.

Using alcohol is strongly recommended. Thinner is not suitable for certain plastic/material. Care is required when using thinner for cleaner. Using thinner on the PCB, copper pads and soldering joint isn't a problem. When use on plastic, you are advise to do a trial to ensure that the material is able to with stand thinner solution. Flux remover is commercially ready mixed solution. From what I learn in the manufacturer website, they are non-flammable chemical solution.

Alcohol & thinner has a lower flash point, and catch fire easily. You have to be extremely in handling and storing these flammable chemical. Do not store or work on the chemical near electrical appliances, hot area or things that can cause spark. Keep your windows open to allow air to circulate while working with these chemical. They may cause headache, dizziness and uncomfortable when inhaled.

Chemical comes in different level of grade. A slight different in the chemical concentration or purity can results in a large difference in cost. Alcohol can be expensive. For pcb cleaning purpose, request for technical grade alcohol. Higher grade alcohol is typically used for consumption or lab experiment purpose. Since we are using it for cleaning, there is not much advantage in using high quality chemical which is expensive. Dickson chemical is specialized in high graded chemical for laboratory use. They do sell lower cost technical grade alcohol. They are not suitable for consumption but is ok for cleaning application.

Dickson Chemical selling Technical grade alcohol

You can easily buy thinner from your local hardware. Typically used for cleaning your brushes after all the wall painting works.

Non-flammable Flux remover

click here to

Buy Cleaning Solution Now at the PIC-store

These are other recommendation in the web. After researching further, I would advise not use them. These chemical are of health hazard. Exposure will have severe adverse health effects.

- 50% isopropanol + 50% water
- Trichloroethylene
- Carbon tetrachloride

Isopropanol (true chemical name) is also known as Isopropyl alcohol or 2-propanol. 2- is refer as Iso. Other chemical may starts with 1- refer as N. These are some jargon used in the chemical industrial. I managed to learn a few of the terms from the vendor.
Trichloroethylene (true chemical name) is typically used for degreasing the metal. It is an extraction solvent for oil. Another name for Trichloroethylene is Trielin. Based on some read up, they can cause cancer.
http://www.answers.com/topic/trichloroethylene

Carbon tetrachloride can harm our environment, producing CFC which deplete the earth's ozone layer. http://en.wikipedia.org/wiki/Carbon_tetrachloride

Initially I tried using contact cleaner to clean the stain off. I use cloth or tissue but the fabric is often tear by the component sharp edges. Most corners are not easy to reach. Later on I tried using a toothbrush, the result is better.

For mass cleaning job, you can try using the ultrasonic pcb cleaner. It is a deep metal container for the solution, using micro vibration to shake off the dirt. I have never tried on it before, and I believe it is the same machine people used for cleaning their jewelry. Basically, you will need to dip the pcb board into the solution, and the ultrasonic will be activated to do it's job. The solution used is the same as manual cleaning, requiring 3-5 minutes of cleaning.

Ultrasonic cleaner

Ultrasonic jewelry cleaner

There is also a flux that does not leave residue on the PCB. I found it on the internet, and have not try it before.

INTERFLUX, IF 2005M No-Residue™ flux, pdf file

Article relating to removing flux residue:

- pcb cleaning - Aqueous Cleaning Process.pdf
- pcb cleaning - Flux Removal, nuxx.pdf
- pcb cleaning - pcb cleaning article.pdf
- pcb cleaning - PCB Cleaning, Printed Circ.pdf
- pcb cleaning - pcb washing chemical composite.txt
- pcb cleaning - remove solder flux.pdf
- pcb cleaning - Solvent washing PCB boards.pdf
- pcb cleaning - washing the board using alcohol.pdf
- pcb cleaning - PCB cleaning article Inventec_S24_03.pdf

<Step 7>

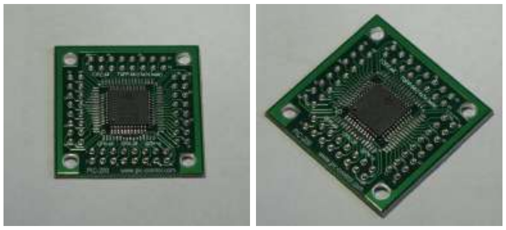

A close up view of the clean and neatly soldered surface mount IC on the prototyping adaptor. Soldering smd components is simple and fast with a bit of practice. Small and easy to clean up. There are too many advantage in working with surface mount, compared to through hole components.

This is the end of the short demonstration on smd soldering. I hope these article can provide you some insight and confident to start experimenting circuits using surface mount components.

Video available. Please click here,
smd soldering.mpg (25MB)
smd cleaning pcb.mpg (5.5MB)

Article on smd soldering.
- 2010-07 How to Solder Surface-Mount Devices.pdf

click here to
Buy this SMD Adaptor Now at the PIC-store

PIC-store

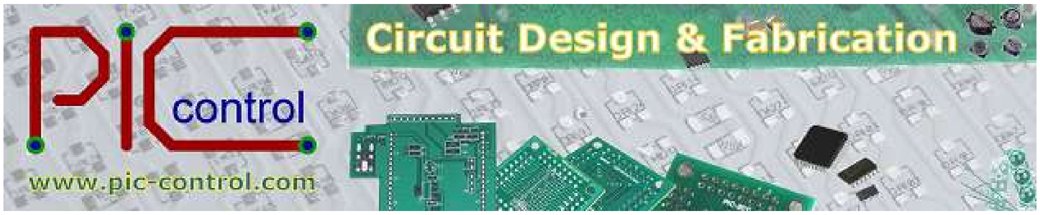

Singapore Customized, custom made Electronics Circuits & Kits

Various Chemical Reference:

Alcohol 70% + Distill Water 30%
MDX4-4159
7MHFE-71DE

Cyclohexanone
Methylethylketone
ethylacetate,
Silicon Wacker
Wacker M4503, General purpose mould making silicon (Using Two Part Silicon Mold Putty to Make Molds)
http://www.kirkside.com.au/, Mould Material Specialist

Propanol, used to clean PCB. (close to white electric oil)
Acetone, remove away ink printed on PVC stickers
n-Hexane (white electric oil)

smd glue for surface mount component mounting / soldering use
Seal-glo NE series, NE3000S, NE8800T, more Seal-glo

email: mail@siongboon.com
website: http://www.siongboon.com

Keyword: surface mount soldering, smd adaptor, smd to 2.54mm convertor

Surface Mount Soldering Procedure

We believe that the following procedures describe the simplest way to reliably assemble surface mount components without special tools. The only tools required are a pair of fine-tip tweezers, a rosin based flux pen, rosin core solder (.025” dia. preferred), and a soldering iron with a small, clean tip. A spool of fine (0.075”) solderwick will come in handy if you need to rework connections. You will need isopropyl alcohol, cotton balls, and cotton swabs for cleanup. If you have access to hot air rework tools and the skill to use them, feel free to use them instead.

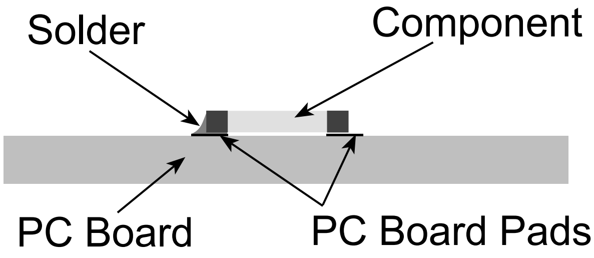

A proper solder joint with fillet

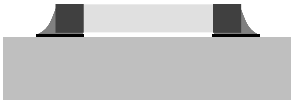

Proper joints

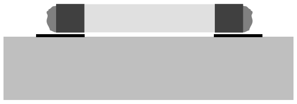

Poor joints

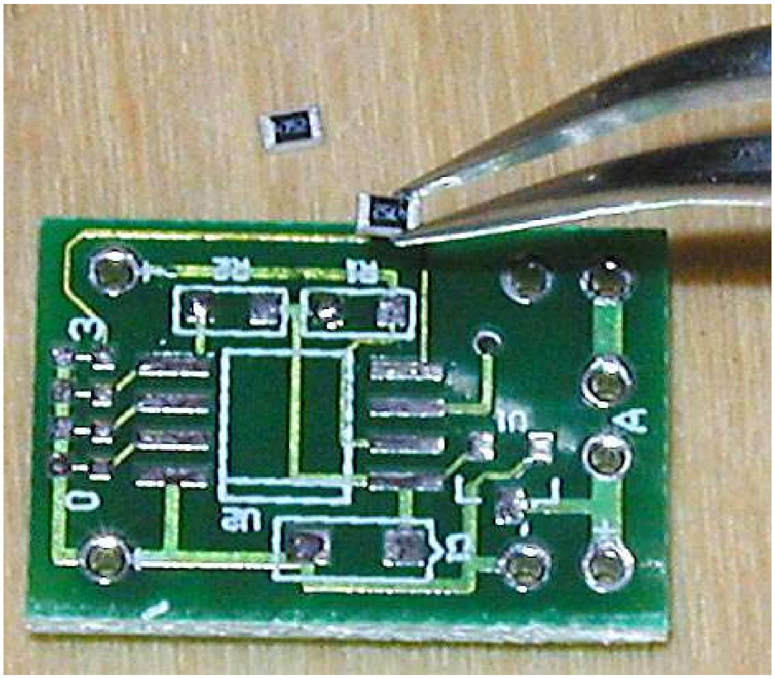

Placing the component

Soldering surface mount components is not difficult, but it does require good eyes, a steady hand, and a soldering iron with a small, clean tip. The illustrations at the left show a proper and a poor solder joint. In the proper joint, the flux has “wetted” both mating surfaces, allowing the molten solder to adhere to both the printed circuit board pad and the component lead. In the illustration of the poor connection, insufficient heat or flux has forced the solder to ball up on the end of the component, without making it to the PCB pad. The component is not attached properly, and the electrical connection will be intermittent at best. This connection can be salvaged by applying flux and heat to the PCB pad to get the solder to flow to the area between the pad and the component.

The most important characteristic of your soldering iron is that it must be equipped with a *small, clean* tip. If the tip is too large or is covered with oxidation, creating a well-flowed solder joint will be nearly impossible. Wipe your tip on a sponge before each joint - it must be shiny and well-tinned in order to transfer its heat properly.

You will find it easiest to work on your board if you clamp it in a moveable vise or stick it to a small block of wood with double-stick tape. This will allow you to rotate the board to whatever angle is most comfortable for access to the component being installed. If you try to hold the tiny parts at awkward angles, you will not have good results!

The first step is to apply a small bead of solder to one of the PCB pads for the component to be installed (the left-hand pads of R1 and R2 in the photo at right). The bead should be approximately 0.020” to 0.030” high (you may want to reevaluate your bead size after you have completed several joints). Apply some liquid *rosin based* flux from a flux pen to the solder bead and to the other pad(s) for the component.

Next, using a pair of tweezers, pick up the component to be installed and place it over the appropriate pads.

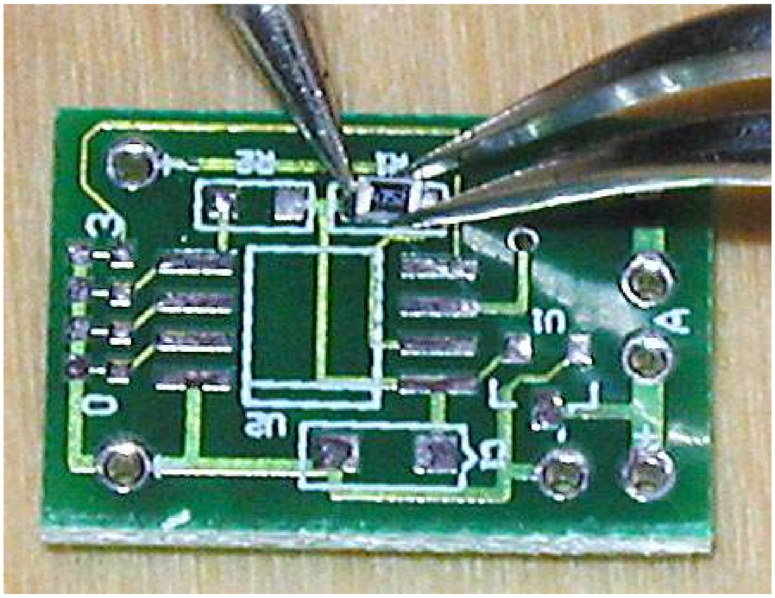

Affix the part by reflowing the first solder bead

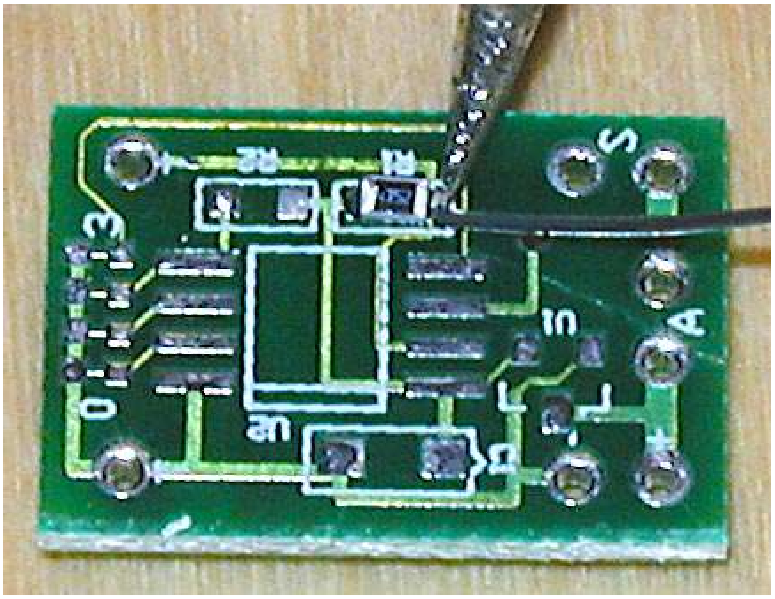

With the part in position, solder the other end

With the part in position, move the soldering iron to the solder bead on the PCB pad. Apply a small amount of heat from the iron to flow the bead, simultaneously lowering the part against the board and correcting for any rotational misalignment. Remove the iron and allow the solder to cool. Inspect the joint -- at this point you are not concerned about the quality of the actual solder joint, just the positioning of the component. The part should be flush against the PCB, with both ends properly contacting the pads. The part should be straight, and centered between the two pads.

Apply a liberal amount of liquid flux (again, a flux pen is the ideal applicator) to both ends of the component. Heat the currently unsoldered end of the component and the adjacent pad with the soldering iron, and carefully wipe on a small amount of solder. You want a small fillet as shown in the earlier illustration, not a large glob of solder. Make sure that the solder has flowed onto the pad as well as the component lead - don’t be fooled by a lump of solder on the end of the component that doesn’t flow under the component and onto the PCB pad.

After the second end is soldered, go back to the original solder joint and reheat the solder. The flux will allow the solder to flow freely, and a good fillet should be observed. If there seems to be insufficient solder, add a little more. If either of the ends appears to have an excessive amount of solder (a large blob sticking up above the part) it can be removed with solder wick. Apply flux to the blob and the wick, position the wick over the blob, and press lightly on the wick with your iron. As the heat is conducted through the wick the solder blob will melt and be drawn off by the wick. You may end up removing too much of the solder, in which case you can reapply a *small* amount of fresh solder to the joint.

After the joints have cooled, inspect them carefully to make sure that they are solid and make contact with the board. If you have any doubt about the quality of the connections, apply more flux and reflow the joints until you are satisfied.

The same procedure is also used for installing components with three or more leads (transistors, ICs, etc.). Start with a bead of solder on one pad, position and “tack” the part, then add more flux and solder the remaining leads. Don’t forget to reflow/improve the first “tacked” connection if necessary.

Home Duane Linda BMW articles Czech Students Search my web Ji-young's travel

Electrical Wire soldering For Beginners

by Randy Glass

DISCLAIMER- First, soldering can be dangerous. This lesson is not meant to be everything you need to know. Your safety is up to you, and you take full responsibility for any damage or injury caused by these instructions, whether as an act of omission or commission on your part or mine. Be aware that an improperly soldered wire can lead to fire, loss of vehicles, injury, and even death. By reading and/or using these instructions you take full responsibility and release me from all liability. Second, I took these photos as I did the work, pressing the remote control for my camera with one finger as I held the soldering pencil in one hand and the solder in the other. I apologize for any images that might be out of focus or poorly composed.

If Yogi Berra ever wrote a book on motorcycles he would probably say, "We seem to spend around forty percent of our time taking things apart, and the other seventy-five percent of the time putting them back together." Some things are easy to put together. We use specialized fasteners like nuts, bolts and screws to hold things together that, some day, we might want to take apart again. But there are some things that, once put together, we would like to know that, to some degree, our fastening system is permanent. One of those places is in electrical wiring. When a wire breaks or needs to be fastened to a connector or another wire in such a way as to be dependable, then the very best way is to solder.

soldering is a method to attach one metal part to another. It is part of a family that uses molten metal to do this job. At the top of the ladder is welding. This is a very high temperature method that actually melts the parts a bit where they are attached, along with some other metal, usually in the form of welding rods. Next is brazing. This melts some hard metal, usually brass, into the gap between two pieces of metal. Normally, the surface of the material being joined is not melted in this process, but it gets very, very hot. Finally there is soldering. soldering flows some molten metal between metal parts at relatively low temperatures.

For our purposes I am going to speak of the type of soldering that attaches two wires together or one wire to a connector- the sort of soldering that any backyard mechanic might need to do when working on any electrical component or the wires that connect them together in almost any type of vehicle.

WHY solder?

Although there are methods that can be used to make these connections without soldering, they are usually inferior. soldering does a lot of things:

1) It Is Strong- You are connecting two metal parts with a liquid metal that becomes solid after it fills the joint and cools. When properly done it leaves no voids so the joint is very strong and the wire is well supported.

2) It Slows or Stops Oxidation - Because solder fills the joint it excludes air and moisture so the metal in the joint will not oxidize. This maximizes electrical flow over time.

3) It Makes An Excellent Electrical Connection- because the solder is a conductor it gives the wire more area from which it can flow electrons to whatever it is connected.

4) Non-permanent - Although soldering is very strong, it can be removed and the joint taken apart in the future if necessary.

5) Connectors do Not Need to be Crimped - Those little metal connectors that are designed to be crimped have a terrible habit of coming loose as well as breaking the little copper wires in the conductor making a weak joint as well as one that does not conduct as well as it should. They easily let air and moisture in which causes corrosion and that leads to poor and broken connections.

MATERIAL LIST

It doesn't take much to set yourself up as a soldering wizard. Here is a list (numbers refer to image to the left):

*1) **soldering Pencil -** Something around 25 to 40 watts would probably be fine for most electrical work. You don't need a huge soldering gun to do electrical work. A little smaller is actually better than too big, as we shall see. Get a brand-named pencil so that down the line you will still be able to

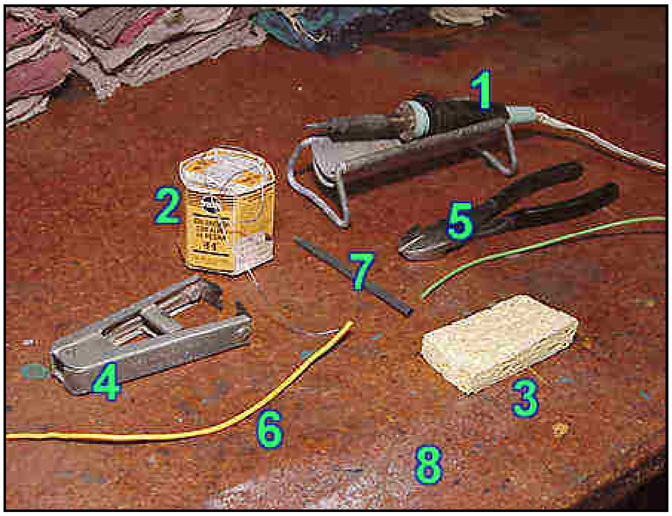

get replacement tips for it. A chisel tip as seen in the pictures that follow is best for most work but you might also need a pointed tip for more delicate work. Those old soldering guns are nice for making battery cables, but are generally too large for what we are doing. it would be like using a sledge hammer to make a picture frame- it would have to be a REALLY big picture frame!

There are also cordless, rechargeable soldering irons available. They generally run about $35-75 or so. Although these cost substantially more than a corded iron, and the replacement tips are more expensive as well, there is no cord to get in the way when working and as they can be carried along on trips. These generally have a built-in light as well making repairs at night easier. I normally loosen the lamp so that it does not light, so the battery lasts longer in use.

There are soldering tools that are actually little, refillable, butane-powered torches with a soldering tip at the end which is heated by the flame. As they are about the size of a fat ball point pen, they are another good way to get a portable soldering tool. These generally are available for less cost than the rechargeable irons mentioned above. I have used electric irons to solder around a gas tank on an automobile- something that I would never consider with one of these.

And a **soldering Pencil Holder** - Most soldering pencils come with a holder or a rest as seen here, but if not, get one. These metal stands are very handy and will help avoid melted parts and burned hands and work surfaces.

*2) **solder** - Resin core solder is what you want. NEVER use acid core solder on electrical components. The acid will stick around and continue to eat into components and wiring well after the soldering job is over.

Resin core solder can be had in various quantities, but I buy the one pound roll for around $12-15 which will be a lifetime supply for most folks. I use Kester brand "44" Resin Core solder in the .031 diameter size. This is about the size of mechanical pencil lead and works quite well for what you will be doing.

It use to be that a standard part of a soldering kit was a little tub of flux. This paste would be spread on the joint ahead of time to help clean the joint chemically as it was heated by the iron. It was necessary as flux-core solder was unavailable. If your wires are clean and you are using a quality, resin core solder, additional flux is not necessary.

*3) A**n Old kitchen sponge** or two

* **Tools** to work on wires such as a (4) wire stripper and a (5) wire cutter, and possibly needle nosed pliers and a small vise (not shown).

* 6) A few bits of **insulated wire** and an old spade connector or two on which to practice. Two or three feet of wire will do.

*7) **Heat Shrink Tubing** - I rarely use electrical tape any more. Heat shrink is really great! I usually buy it in three foot lengths in two or three sizes. This is a sort of memory tubing that, I believe, is made by heating a special tubing and stretching it and allowing it to cool while still stretched. it maintains that size. When heat is applied once again the tubing will shrink back down to its original size.

*8) Protection for the work surface like an old piece of 1/4" plywood about 12" square or such. An old piece of counter-top lamination works also.

* Wicking material (not shown)- You can purchase some special copper wick that is used to unsolder joints. This is usually not necessary, but handy at times, and inexpensive.

* **Eye protection**

* A Ventilated workspace

SOME WARNINGS

HEAT - A soldering pencil as well as molten solder are both really hot. They will burn you and it will hurt. It can hurt a lot. It can make your skin give off smoke and that smells bad. If left unattended a hot soldering pencil can cause fires, melt carpeting, burn workbenches, and make the cat dance like never before. An unsuspecting person can pick up a hot soldering iron and get a nasty burn (trust me on this one!). Never leave a soldering tool warming or hot and unattended.

Additionally, a hot soldering iron can almost instantly melt plastic parts and upholstery in a hurry. Use care when soldering as well as when moving the iron to and from the solder joint. If the cord gets caught when moving the iron it can easily cause you to drop the hot tool onto something expensive or painful. NEVER place a soldering iron on saddles, luggage, body work, painted surfaces, or on or near any heat-sensitive surface.

LEAD and VAPORS- solder contains a good amount of lead and handling it over a long period of time may have negative health consequences. The vapors caused by soldering can, over time, cause health problems as well. Work in a well ventilated area. I usually work out doors and use a fan to blow the vapors away from me. The smoke that is given off by the flux in the solder is also not a good thing to breathe.

MOLTEN METAL - When soldering (and particularly when un-soldering) it is very possible to flick bits of melted solder about. As this stuff is quite hot, even a small drop in the eye can cause permanent damage. Wear eye protection and do not wear your good glasses, or wear them under goggles. The dripped solder can melt plastic parts and permanently damage auto carpeting as well. Be aware of other folks or critters who may be in the area as well.

ELECTRICITY - You will be soldering on wires that, if shorted out, could cause a blown fuse, damaged electrical components, or melting wires and a fire. Always disconnect the battery when soldering wires on a vehicle and remove connectors from heat-sensitive components before soldering on them.

LETS GO

I could spend a lot of time talking about how it all works, but jumping right in is the best way to learn how to solder and I will explain as I go

Begin by creating a safe work area. One free of clutter, with sufficient light, and protected from the soldering iron and melting solder. Set the soldering pencil in its stand and plug it in. be sure that the cord is placed in a way that it will not be accidently pulled or tripped on, upsetting the soldering pencil.

While that is heating up, prepare your sponge. Wet it, squeeze out the excess water, then fold it in

half and rubber band it about half-way down.

Pull out about a foot or two of solder from the roll and touch it to the tip of the soldering iron. Always apply the solder to the flat parts of the tip and not to the sides. When the solder melts fairly easily, melt an amount all over the tip. There will be some smoke from the resin in the solder.

The solder you bought is actually a bit hollow. Inside of the metal there is a channel (or in some cases, channels) that make it sort of like a little garden hose. The channel is filled with a resin. When heated it melts and some of it burns off. The action of the resin when it is heated helps clean the surface to be soldered which helps the solder stick a lot better. Unfortunately, the burned resins can also leave deposits on the soldering pencil, and over time will make it difficult to solder. That's where the sponge comes in. After you have melted some solder onto the tip (about a quarter inch or so of the thin stuff) wipe the soldering pencil's tip off on the damp sponge a few times.

If you examine the tip of the soldering pencil carefully you will see that it is now covered with a thing coat of shiny, silver metal- the solder. You have successfully "tinned" the tip of the pencil. This keeps the metal tip from oxidizing and will assist you in soldering from now on. Always do this before starting a soldering job and occasionally during a larger job.

Place the soldering pencil back into its holder and let's prepare the wires. We are going to solder two wires together, making believe they have broken or need extending.

We start by "stripping" some of the plastic insulation away from where we want to connect the wires. The thicker the wire the more insulation we remove. For most common wires on a motorcycle or in a car, removing about 3/8 to � inch works fine. You will get a

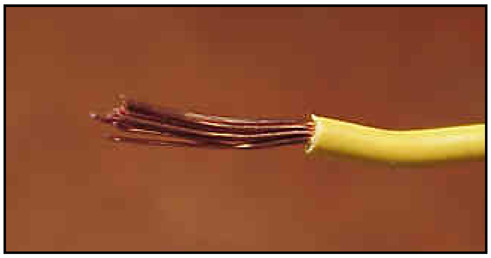

feel for how much in a short period of time. When stripping wire be very careful not to cut the strands of the conductor.

There are special tools available for stripping wire that can save a lot of time. These can be as simple as a carefully wielded razor blade to fancy and somewhat expensive automatic strippers.

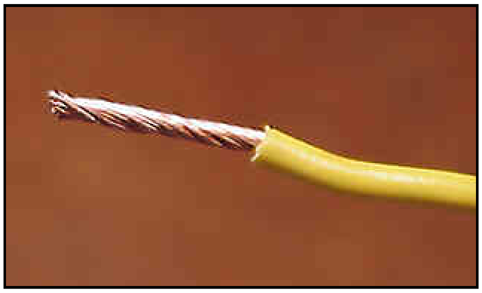

Once stripped, twist the strands together into a neat little bundle.

The exposed wires should look fairly clean and in most cases, shiny, but certainly not green or cruddy. If corroded, then the wires should be cut back and re-stripped until clean wire is found.

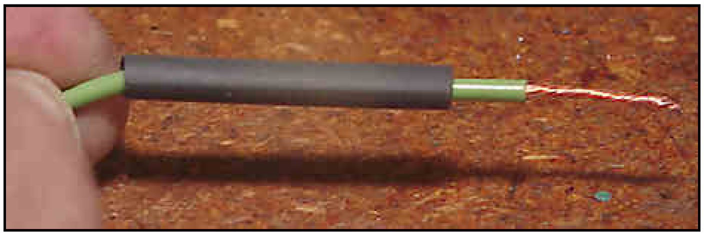

Now cut a length of heat shrink tubing off about an inch long. It should be of a large enough diameter that at least two of the wires you are soldering can easily fit into it at the same time. Slide the length of the tubing over one of the stripped wires- it doesn't matter which here, but in real life we would slide it over the lower wire or the one which slopes most downward away from the joint so it will tend to slide out of the way. make sure it is at least three or four inches away from the joint.

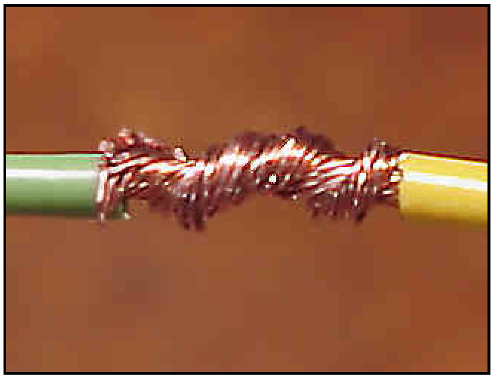

Now, twist the two wires together into a tight, neat pair. This is important because it helps create a good electrical connection. You also do not want any little wires sticking out. Start by crossing the wires about mid-strip, crossing them at about 45 degrees or so, and twisting in opposite directions. When done, the joint should be fairly strong on its own, holding the two wires together tightly. The connected pair of wires should be in line as if they are a single wire and not two joined together.

It's time to solder. Place the roll of solder on your left (if you are right handed) with a good length of solder hanging out of the roll. With the solder in your left hand and the soldering iron in your right hand, check to see that the tip is still tinned and shiny. If not, repeat the process with a bit of solder melted to the tip, then wipe the tip off on the sponge.

Melt a little ball of solder onto the iron....

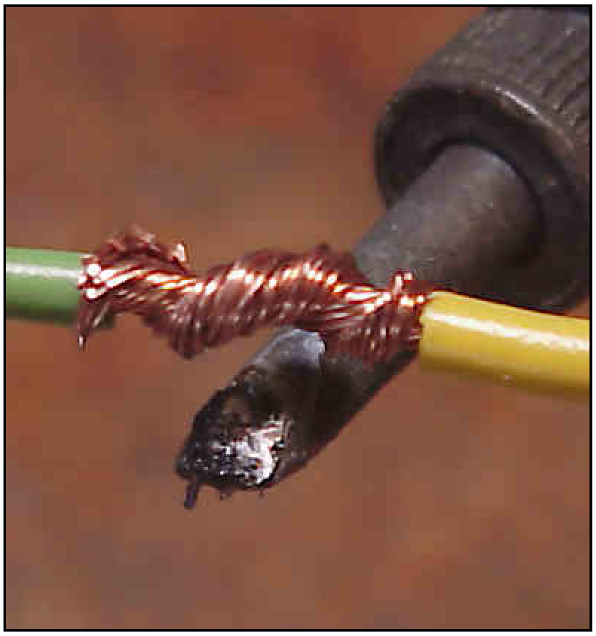

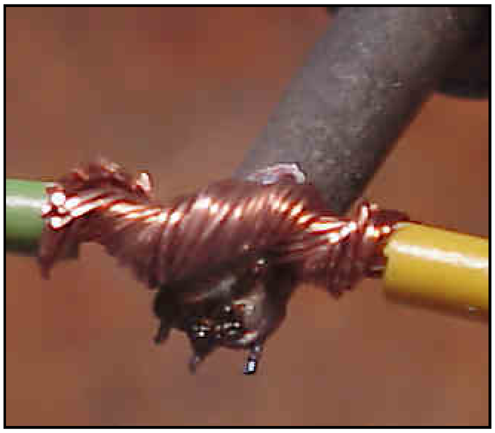

...and immediately place the soldering iron under the twisted wires (if possible) so that the little ball of solder fills much of the gap between the soldering pencil and the wire. Since heat rises, placing the iron under the joint will heat it faster. The little ball of solder works as a heat conductor and will very rapidly conduct the heat from the soldering pencil's tip into the wires. If the ball of solder does not seem like it is contacting the wire very well, without removing the pencil from the joint, melt just a bit more solder in there by feeding the end of the solder wire into the tip. This will not only increase the contact area of the solder to the wire but it will also add a bit of flux that helps the solder adhere to the wire at this time.

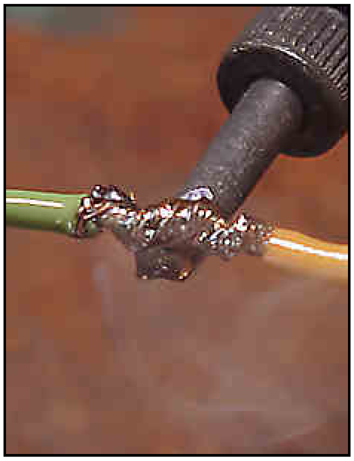

After about two or three seconds, start to feed solder, slowly at first, into the void between the pencil and the wire. You should soon see the solder is flowing into the twisted wires. WHen that happens, increase the feed rate of the solder into the joint, and you can now also feed a little into the top of the joint. Don't move the iron's placement, jut take the end of the solder wire and touch it to the tip of the joint. Continue to feed solder until the individual wires are almost filled with solder and the entire joint has turned silver. When that happens, stop feeding the solder in and immediately but gently remove the soldering iron.

How long do you wait until feeding in solder? One way to test how long to wait is to touch the solder on top of the joint and wait until that area is hot enough to melt the solder. The amount of time it takes varies with the size of the wires to be soldered, the size of the solder, the wattage of the iron, the size of the irons's tip, the ambient air temperature, and other factors. Experience will teach you how long this takes.

The problem comes from waiting too long and putting too much heat into the joint. When you do that it begins to melt the insulation on the wires around the joint (you can see that beginning to happen above at the end of the insulation of the green wire next to the soldered joint). That is why I like tho flow the solder in from where the iron's tip touches the wire. This is the place that will come up to soldering temperature first, and then the flowing solder will more easily flow solder into the joint as it carries heat with it. The larger the wires, the longer it takes and the more important it becomes to feed in some solder from the top of the joint as well. Again, experience will teach you when that is necessary.

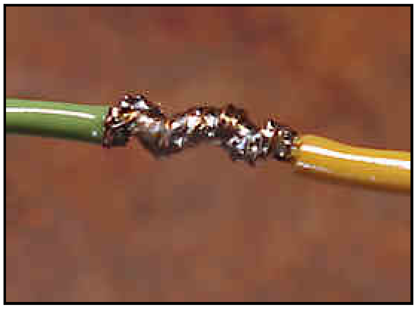

How much solder is enough? Too much? The correct amount of solder is used when the joint is fairly all silver colored but still shows the "outlines" of the individual strands of wire in the joint. If there is very much in the way of copper colored wire showing then you probably didn't use enough solder or the wires weren't clean and would not accept the solder. If there are big blobs and drips of hardened solder on the joint then you probably had either too cool of a joint or used too much solder.

After soldering, let the joint cool for about thirty seconds to a minute or so before handling. If you were successful you should have:

* A pair of wires tightly joined together

* The joint between them is very shiny, looking like polished sterling silver.

* No hardened dribbles or drops of solder hanging off the joint

* No little wires sticking out of the joint.

* A union only slightly larger than each of the wires themselves.

If you were NOT successful you might see:

* Large blobs or possibly a long, pointy finger of hardened solder hanging off the joint. Sometimes, after cleaning the iron, these can be removed by simply heating the joint until the solder melts, then removing the iron which will pull away some of the excess solder with it. This may need to be repeated a few times to remedy the situation.

* Wires that are not soldered together and has movement within it. Re-heat the joint and add some more solder.

* Melted insulation. This can be so bad that the wires are revealed with the insulation dripping off the wires. At this point you will probably need to cut the wires and begin again, or unsolder the joint and add a long piece of heat-shrink tubing to cover the exposed wires.

* Little ends of the copper wires sticking out. These can sometimes be cut off with a wire cutter and/or pressed down with a pair of smooth-jaw pliers, possibly needle nose pliers.

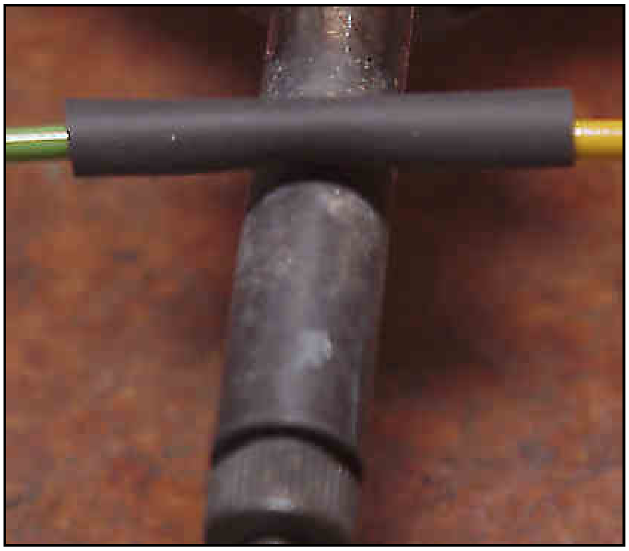

Now it is time to cover the joint with the heat shrink tubing. You DID remember to slide on the heat shrink tubing before beginning, didn't you? Once the joint is cool enough to touch, slide the tubing so it is centered over the joint and gently touch the meaty part of the soldering iron's heating element to the center of the tubing.

Continue this around the center of the tubing until it tightly grips the wire joint, then work outwards until the entire piece of tubing has shrunk tightly around the joint.

Attaching An Electrical Connector

Besides connecting two wires together, soldering is the best way top attach some sort of connector to the end of a wire. The connection is strong as well as fairly well sealed against the elements. It works pretty much the same way as above...

Remember when we "tinned" the tip of the soldering pencil? We did this to apply a film of solder to the tip so it wouldn't oxidize and thus be able to more easily transfer heat. Now I am going to do the same with a piece of wire.

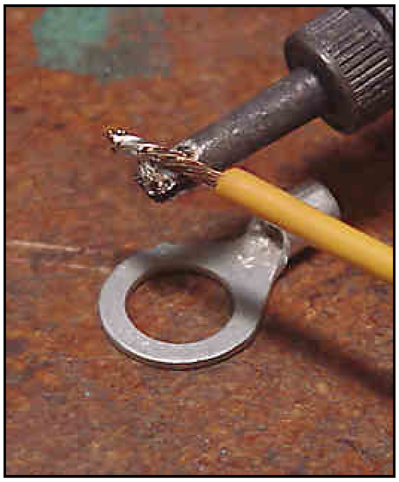

Here I am applying a good amount of solder ahead of time to this wire because I am planning on soldering it to a fairly heavy connector. This way, the wire will be "pre-soldered" and easier to attach, and will have less of a chance of melting the insulation.

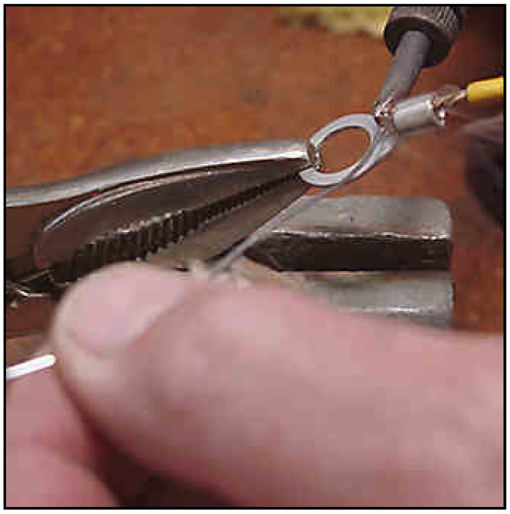

When soldering a connector to the end of a piece of wire it is very helpful to use some sort of vise to hold the connector. A small pair of locking pliers or a pair of needle nose pliers with a rubber band around the handles to hold the jaws shut works well. Here I have locked a pair of locking pliers into a small vise to work as a "hands-free" holder. This not only allows me to use both hands to work but isolates me from the heat of the soldering process.

Remember that the heavy metal pliers will work as a heat sink, pulling heat away from the soldering joint, so it might take longer to achieve the proper heat level in the joint. That is why I have gripped the connector as far away from the area to be soldered as possible,

Be careful as the pliers may get hot near the tips and you might overheat the wire as well. Be sure to slide the heat shrink tubing over then wire BEFORE beginning to solder! If you leave it off and finish soldering, you will be left with the soldering job in one hand, the heat shrink tubing in the other hand, and a stupid look on your face. We call this, "Heat Shrink Tubing Syndrome."

Additionally, if you apply too much solder, particularly with a small connector, it is very possible to fill the connector with solder and thus make it difficult or impossible for the connector to function properly. There are ways to get the solder out, but they are generally more difficult (or dangerous) than just chucking it and starting over. To keep this from happening:

* Never touch the area of the connector where you don't want solder to flow
* Orient the connector with the part where solder shouldn't flow, upwards (or Higher than the end where solder will flow).
* Apply solder sparingly and slowly
* Use a lower wattage soldering pencil
* Use a smaller tip on the soldering pencil

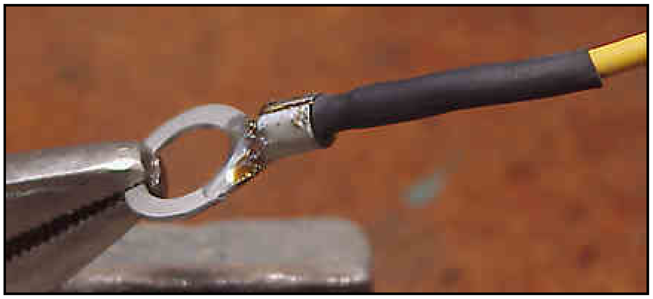

I sometimes use two pieces of heat shrink, one over the other, shrunk one at a time, in areas where extra protection is needed. This is particularly useful when soldering a connector to the end of a wire. I first shrink a small diameter piece about an inch long over the wire up to where it joins the connector as shown here. This acts as a strain relief as well as insulation.

Notice the small amount of flux on the connector ring left over from the soldering process. If you use resin core solder, this stuff is harmless to the electrical device. It can be brushed off with a small wire brush if you wish.

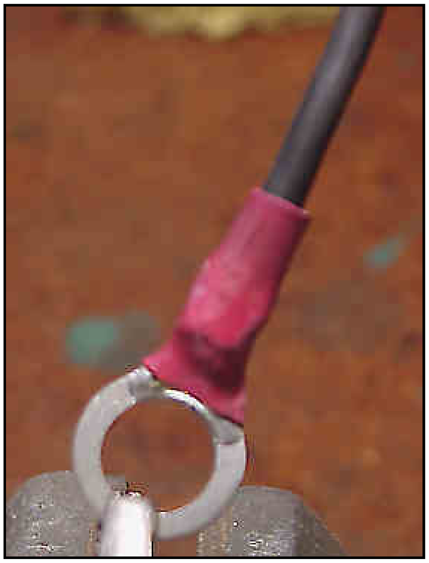

I then shrink another larger diameter piece about � an inch long over the connector to act as an insulating cover. In the case of something like a female spade connector, the entire connector except for the opening can be covered and protected in this way. If the heat shrink interferes with the connection, simply cut the offending part away with a razor blade being careful not to damage anything else including yourself!

In areas where the joint needs to be protected from corrosion, slather the terminal with petroleum jelly, or the special grease for this application.

Following these directions you will become a "master solderer" in a short period of time. It takes some practice to get a feel of how hot for how long with how much solder it takes to create to make a good solder joint, so get some scrap wire and practice before attempting to solder on anything important.

If there are any questions or comments, please feel free to E-mail me at frcn@cncnet.com.

Note by Duane; If you appreciate articles like this, I suggest that you thank Randy at the email address above.

More BMW info

Home page

This page was last edited: 02/03/2006 - copyright Duane Ausherman
Web hosting provided by hostmeister.com

Soldering and Desoldering Instruction

Soldering is defined as "the joining of metals by a fusion of alloys which have relatively low melting points". In other words, you use a metal that has a low melting point to adhere the surfaces to be soldered together. Soldering is more like gluing with molten metal than anything else. Soldering is also a must have skill for all sorts of electrical and electronics work. It is also a skill that must be taught correctly and developed with practice.

Remember that when soldering, the rosin in the solder releases fumes. These fumes are harmful to your eyes and lungs. Therefore, always work in a well-ventilated area. Hot solder is also dangerous. Be sure not to let is splash around because it will burn you almost instantly. Eye protection is also advised.

The Tools

Soldering Iron – Even the cheapest of them will do the job. They may not last as long, but they do get hot enough to melt solder and that is the goal. You only need to make sure that the one you buy has a suitable tip on it. The most typical tip is the one that is about the size of a 1/8” stereo mini-plug.

Solder – Get Rosin Core solder. Rosin will help the solder to flow onto the wires. Solder comes in different thickness. Buy what is appropriate for your job.

Desoldering Gun – This tool will make life a lot easier when you need to rework a previous solder job.
Solder Wick – Braided copper with rosin coating used to draw solder off of circuit boards.

SMD Rework station— Basically this is a high temp hair dryer with special nozzle attached. It heats up entire pins of the SMD device so you can remove them.

The Methods

Soldering – There is not art to soldering, it takes patience and practice to get it right. If you are doing it right, it will be easy and very fast. First, make sure that your soldering iron tip is clean. If it is dirtier than just a light gray color, you need to clean it. You can do this with sandpaper or a Scotch Brite pad. Next, turn your iron on and give it plenty of time to heat up. To test the heat, use a piece of solder touched to the tip. If it melts immediately, it is ready.

The most common way to mess up a solder job is to let the soldering iron stay on the parts to be soldered too long. What this does is oxidize the metal, making it too dirty to attract solder. If you are working with a circuit board, too much heat can damage the board, rendering it useless and in need of repair. You should only hold the soldering iron to the parts to be joined for no more than 10-12 seconds. Any longer and you will melt insulation on wire or damage a circuit board.

It has been said that soldering is a two-person job. You need two hands to hold the parts together, one hand to hold the iron, and another to feed the solder. The correct method for applying solder is to hold the joint perfectly still while heating with the soldering iron. After a few seconds, introduce the end of the solder at the point where the iron meets the parts

If the solder does not melt immediately and flow onto the joint, pull the solder away and try again after a couple seconds longer. If you exceed 10 seconds, pull the iron off and try again after it all cools down. You probably didn't have the iron touching enough of the parts to be soldered. Sometimes, the parts to be soldered are so big that they conduct the heat away very quickly and make it difficult to solder. In this case, it is OK to put some solder on each part individually and then put them together later by melting the solder on each one while they are touching.

Joining Two Wires - Strip off about 3/8" of insulation from the two wire ends to be joined. Place a ½" length of heat shrink tubing over one wire and push it back so that the heat from soldering won't shrink it prematurely. Fold each bare wire end back on itself so that the tip of the wire now comes back to the end of the insulation. Link both wires together using the bends like hooks. Now, twist the ends of the wires around themselves tightly. This type of connection is called a Western Union and it is the strongest method of joining two wires together. Now, heat the connection with the iron and allow solder to flow over the entire joint. Wait a few seconds for it to cool off and then slide the heat shrink tubing down over the connection and heat it with a match or heat gun.

Soldering a Wire to a Circuit Board – This is delicate work that you will do while soldering. Too much heat and your board will be damaged. If you are attaching a wire to an existing hole on the board, make sure that the hole is clean and free of any excess solder (see de-soldering for more detail). The best advice when soldering to a circuit board is to angle your soldering iron tip so that it makes good contact with the pad on the circuit board and the wire or part to be soldered at the same time. Heat the parts up and touch the solder at the point where the tip meets the part and the pad. If the solder does not flow immediately, pull everything off and wait for it to cool down and try again later.

Step 1

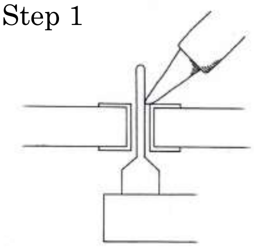

Apply heat to both pin and pad on the circuit board. Make sure tip of the soldering iron is touching both pin and pad.

Step 2

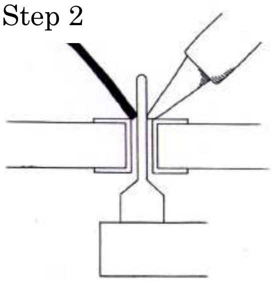

When the both pin and pad are heated enough, apply solder to the pin and pad. Solder should starts to flow.

Step 3

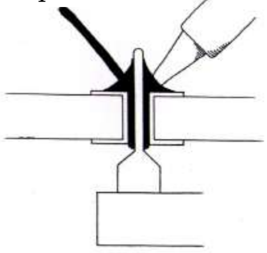

Apply enough but not too much solder. While applying the solder, do not move the tip of iron.

Step 4

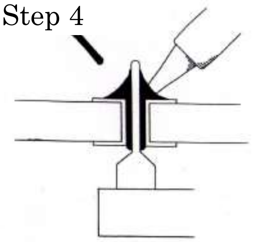

Remove solder away.

Step 5

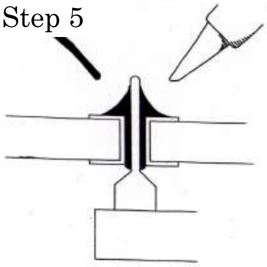

Remove tip of soldering iron also.

Step 6

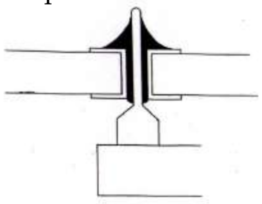

Done! Check the joint to make sure the connection.

Good soldering and Bad soldering Joints

Here are some example of Good soldering and Bad soldering joint.
For Japanese, we often say, “Good soldering joint should look like Mt. Fuji.”

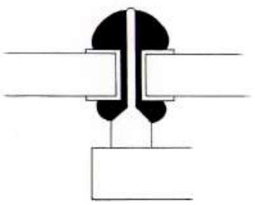

Too much soldering

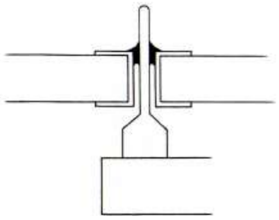

Too little soldering

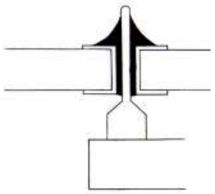

Good soldering

Cold Solder Joints

A cold joint is a joint in which the solder does not make good contact with the component lead or printed circuit board pad. Cold joints occur when the component lead or solder pad moves before the solder is completely cooled. Cold joints make a really bad electrical connection and can prevent your circuit from working.
Cold joints can be recognized by a characteristic grainy, dull gray color, and can be easily fixed. This is done by first removing the old solder with a desoldering tool or simply by heating it up and flicking it off with the iron. Once the old solder is off, you can re-solder the joint, making sure to keep it still as it cools.

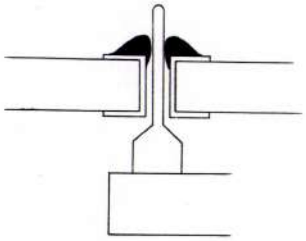

Cold soldering joints

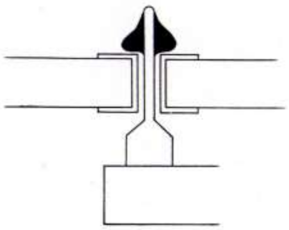

Tips and Tricks

Soldering is something that needs to be practiced. These tips should help you become successful so you can stop practicing and get down to some serious building.

1. Use heat sinks. Heat sinks are a must for the leads of sensitive components such as ICs and transistors. If you don't have a clip on heat sink, then a pair of pliers is a good substitute.

2. Keep the iron tip clean. A clean iron tip means better heat conduction and a better joint. Use a wet sponge to clean the tip between joints.

3. Double-check joints. It is a good idea to check all solder joints with an ohmmeter after they are cooled. If the joint measures any more than a few tenths of an ohm, then it may be a good idea to re-solder it.

4. Use the proper iron or temperature. Remember that bigger joints will take longer to heat up with an 30W iron than with a 150W iron. While 30W is good for printed circuit boards and the like, higher wattages are great when soldering to a heavy metal chassis.

Desoldering

Desoldering is extremely difficult compare to soldering. In the process of RomBoard installation, the parts and circuit board must be in the good shape to re-use them. The tool we use is Desoldering Gun. This device has vacuum pump built in with heater tip. Process of desoldering it self is very simple, but there are some tricks to do clean and safe desoldering job.

Step 1

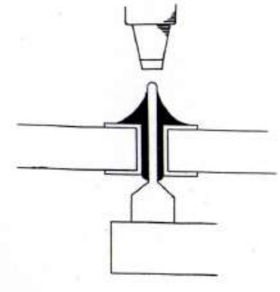

Put the desoldering gun's tip over to the soldering joint. Make sure nozzle is 90° angle from circuit board.

Step 2

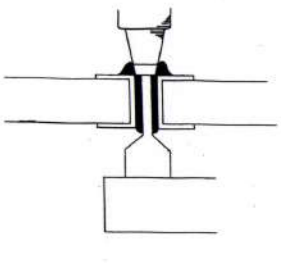

Apply the enough heat to melt entire solder. Multi-layer board or ground pin may take while to transfer heat.

Step 3

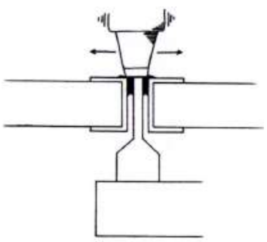

While applying vacuum, move the tip around the pin to get all the solder of around the pin out. At this time, you can also feel pin moves freely, so you know pin is free from circuit board.

Step 4

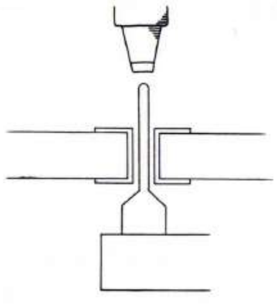

When you can hear air flow sound from hole, you should be done. Stop applying vacuum and remove the gun from the pin. Inspect hole and pin.

When you done with desoldering, the parts that you are trying to remove should move freely. If it doesn't, find which pin is still has solder left, and re-apply fresh solder to it and try desoldering process again. The multi-layer circuit board require more heat to get solder to melt. Make sure pin start to move freely by moving the tip of soldering gun before you apply vacuum to it.

SMD device soldering and removal

In the process of RomBoard installation, chance of handling SMD devices is becoming higher and higher due to ECU configuration change.
Soldering of SMD devices are not much different from regular through hole soldering.
Important thing is positioning of the device is very critical.
Start with applying little solder to two of the each corner of the pads on the circuit board.
Then, place the SMD device and re-heat a one of the solder you just applied to connect a SMD device pin to the pad. Check the position of the device, if position is right re-heat the other end of solder to secure the device completely. If you didn't get position right, re-heat the pin that you just soldered and while heating a pin, move the device to the right position.
Once you positioned the device in the right position, apply Rosin Flux to both pins and pads. This will help your soldering job by keeping solder separate from each pin.
Here is how the joint on SMD device pin should look like.

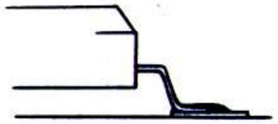

QFP and SOP chip

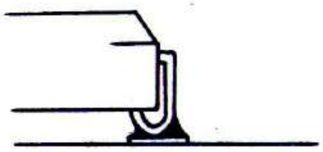

PLCC chip

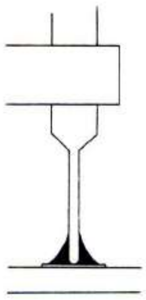

RomBoard adapter socket

Removing SMD device will be the probably the most difficult process. SMD rework station is used to do this. Make sure all the pins are completely heated otherwise you will be removing a pad from the circuit board along with device. Also, you should note, when you are applying heat to the device, some parts around the device is also heated and moves around when you touch. It is good idea to take a note of location of the devices near the target device.

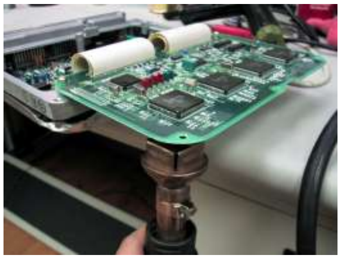

Step 1
Apply pre-heat from back side of the target device. Keep applying heat until you can't touch the chip. By this time, circuit board should have enough heat.

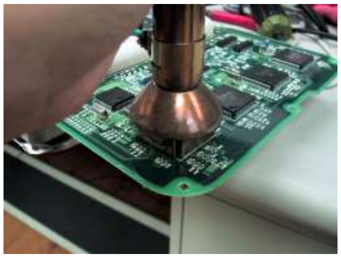

Step 2
Apply heat from top of the target device. Nozzle should cover entire chip. You can check if chip is free from solder by sliding the nozzle left and right. Make sure you don't move other parts. If the chip moves freely, it is ready to be removed.

Step 3
Remove the chip straight up to avoid making solder bridge between pins. Chip is extremely hot! Unless you have skin of steel, avoid imitating the picture.

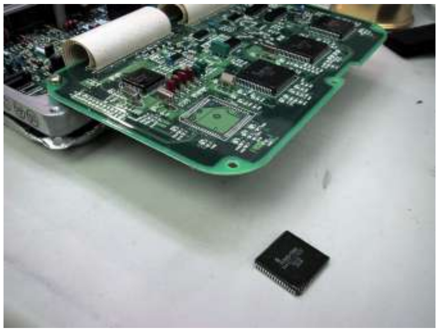

Step 4
After remove the chip, cool down both circuit board and chip.

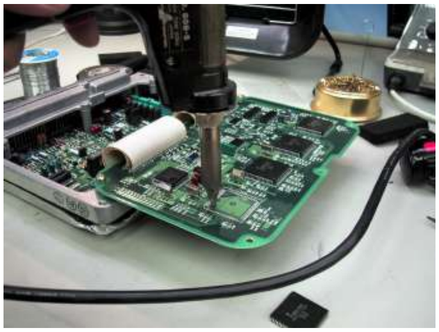

Step 5
Clean the pads on the circuit board with desoldering gun.
Do not apply too much heat to the pad. Pad might peel off from circuit board. Clean the chip by desoldering gun also.
Check for any bridge between pins. If the chip has too much protective coating on it, use solvent to clean it up.

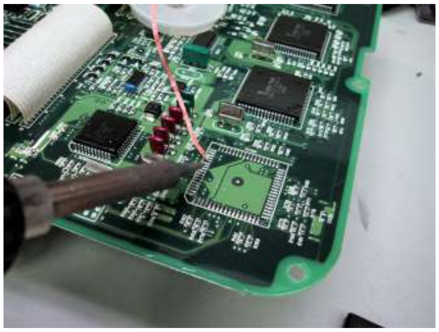

Step 6
If the pads on the circuit board is too small to use desolding gun, try cleaning by solder wick with rosin flux. This may take time but it is much better than peeling pads off.
Once again, do not apply too much heat.

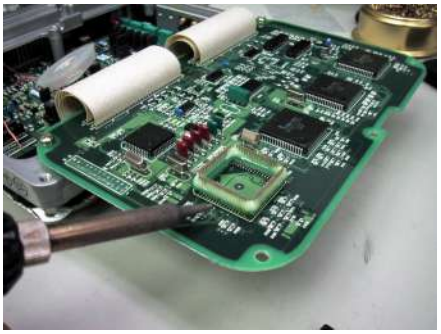

Step 7
Solder the RomBoard adapter where the chip was.
Refer SMD device soldering section for how to solder adapter.

The Final Word

Soldering, desoldering, and working with SMD device will take some effort to learn. You should be relaxed and concentrated when you work.
Keeping every thing clean (tip of iron, gun, work surface, etc) will help you also.
Once you get comfortable to work with soldering you may find your own tips and tricks.
Experience is all bout soldering.

EOF

How To Solder - Soldering Tutorial

Soldering is defined as "the joining of metals by a fusion of alloys which have relatively low melting points". In other words, you use a metal that has a low melting point to adhere the surfaces to be soldered together. Consider that soldering is more like gluing with molten metal, unlike welding where the base metals are actually melted and combined. Soldering is also a must have skill for all sorts of electrical and electronics work. It is also a skill that must be taught correctly and developed with practice. This tutorial will cover the most common types of soldering required for electronics work. This includes soldering components to printed circuit boards and soldering a spliced wire joint. This presentation came from http://www.aaroncake.net/electronics/solder.htm

Soldering Equipment

The Soldering Iron/Gun

The first thing you will need is a soldering iron, which is the heat source used to melt solder. Irons of the 15W to 30W range are good for most electronics/printed circuit board work. Anything higher in wattage and you risk damaging either the component or the board. If you intend to solder heavy components and thick wire, then you will want to invest in an iron of higher wattage (40W and above) or one of the large soldering guns. The main difference between an iron and a gun is that an iron is pencil shaped and designed with a pinpoint heat source for precise work, while a gun is in a familiar gun shape with a large high wattage tip heated by flowing electrical current directly through it.

A 30W Watt Soldering Iron

A 300W Soldering Gun

For **hobbyist electronics** use, a soldering iron is generally the tool of choice as its small tip and low heat capacity is suited for printed circuit board work (such as assembling kits). A soldering gun is generally used in heavy duty soldering such as joining heavy gauge wires, soldering brackets to a chassis or stained glass work.

You should choose a soldering iron with a 3-pronged grounding plug. The ground will help prevent stray voltage from collecting at the soldering tip and potentially damaging sensitive (such as CMOS) components. By their nature, soldering guns are quite "dirty" in this respect as the heat is generated by shorting a current (often AC) through the tip made of formed wire. Guns will have much less use in hobbyist electronics so if you have only one tool choice, an iron is what you want. For a beginner, a 15W to 30W range is the best but be aware that at the 15W end of that range, you may not have enough power to join wires or larger components. As your skill increases, a 40W iron is an excellent choice as it has the capacity for slightly larger jobs and makes joints very quickly. Be aware that it is often best to use a more powerful iron so that you don't need to spend a lot of time heating the joint, which can damage components.

A variation of the basic gun or iron is the soldering station, where the soldering instrument is attached to a variable power supply. A soldering station can precisely control the temperature of the soldering tip unlike a standard gun or iron where the tip temperature will increase when idle and decrease when applying heat to a joint. However, the price of a soldering station is often ten to one hundred times the cost of a basic iron and thus really isn't an option for the hobby market. But if you plan to do very precise work, such as surface mount, or spend 8 hours a day behind a soldering iron, then you should consider a soldering station.

The rest of this document will assume that you are using a soldering iron as that is what the majority of electronics work requires. The techniques for using a soldering gun are basically the same with the only difference being that heat is only generated when the trigger is pressed.

Solder

The choice of solder is also important. There several kinds of solder available but only a few are suitable for electronics work. Most importantly, you will only use *rosin core solder*. *Acid* core solder is common in hardware stores and home improvement stores, but meant for soldering copper plumbing pipes and not **electronic circuits**. If acid core solder is used on electronics, the acid will destroy the traces on the printed circuit board and erode the component leads. It can also form a conductive layer leading to shorts.

For most printed circuit board work, a solder with a diameter of 0.75MM to 1.0MM is desirable. Thicker solder may be used and will allow you to solder larger joints more quickly, but will make soldering small joints difficult and increase the likelihood of creating solder bridges between closely spaced PCB pads. An alloy of 60/40 (60% tin, 40% lead) is used for most electronics work. These days, several lead-free solders are available as well. **Kester "44" Rosin Core** solder has been a staple of electronics for many years and continues to be available. It is available in several diameters and has a non-corrosive flux.

Large joints, such as soldering a bracket to a chassis using a high wattage soldering gun, will require a separate application of brush on flux and a thick diameter solder of several millimeters. Remember that when soldering, the flux in the solder will release fumes as it is heated. These fumes are harmful to your eyes and lungs. Therefore, always work in a well-ventilated area and avoid breathing the smoke created. Hot solder is also dangerous. It is surprisingly easy to splash hot solder onto yourself, which is a thoroughly unpleasant experience. Eye protection is also advised.

Preparing To Solder

Tinning the Soldering Tip

Before use, a new soldering tip, or one that is very dirty, must be tinned. "Tinning" is the process of coating a soldering tip with a thin coat of solder. This aids in heat transfer between the tip and the component you are soldering, and also gives the solder a base from which to flow from.

Step 1: Warm Up the Iron

Warm up the soldering iron or gun thoroughly. Make sure that it has fully come to temperature because you are about to melt a lot of solder on it. This is especially important if the iron is new because it may have been packed with some kind of coating to prevent corrosion.

Step 2: Prepare A Little Space

While the soldering iron is warming up, prepare a little space to work. Moisten a little sponge and place it in the base of your soldering iron stand or in a dish close by. Lay down a piece of cardboard in case you drip solder (you probably will) and make sure you have room to work comfortably.

Step 3: Thoroughly Coat the Tip in Solder

Thoroughly coat the soldering tip in solder. It is very important to cover the entire tip. You will use a considerable amount of solder during this process and it will drip, so be ready. If you leave any part of the tip uncovered it will tend to collect flux residue and will not conduct heat very well, so run the solder up and down the tip and completely around it to totally cover it in molten solder.

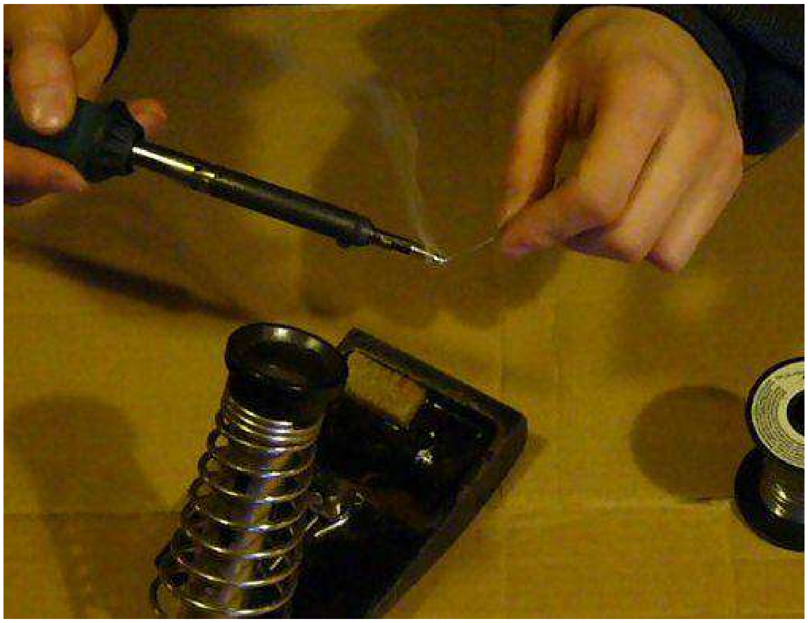
.

Step 4: Clean the Soldering Tip

After you are certain that the tip is totally coated in solder, wipe the tip off on the wet sponge to remove all the flux residue. Do this immediately so there is no time for the flux to dry out and solidify.

Step 5: You're Done!

You have just tinned your soldering tip. This must be done anytime you replace the tip or clean it so that the iron maintains good heat transfer.

Soldering a Printed Circuit Board (PCB)

Soldering a PCB is probably the most common soldering task an electronics hobbyist will perform. The basic techniques are fairly easy to grasp but it is a skill that will take a little practice to master. The best way to practice is to buy a simple electronics kit or assemble a simple circuit (such as an **LED chaser**) on a perf-board. Don't buy that expensive kit or dive into a huge project after only soldering a few joints.

Soldering components onto a PCB involves preparing the surface, placing the components, and then soldering the joint.

Step 1: Surface Preparation:

A clean surface is very important if you want a strong, low resistance solder joint. All surfaces to be soldered should be cleaned well. **3M Scotch Brite pads** purchased from the home improvement, industrial supply store or automotive body shop are a good choice as they will quickly remove surface tarnish but will not abrade the PCB material. Note that you will want *industrial* pads and not the kitchen cleaning pads impregnated with cleaner/soap. If you have particularly tough deposits on your board, then a fine grade of steel wool is acceptable but be very cautious on boards with tight tolerances as the fine steel shavings can lodge between pads and in holes.

Once you have cleaned the board down to shiny copper you can use a solvent such as acetone to clean any bits of the cleaning pad that may remain and to remove chemical contamination from the surface of the board. Methyl hydrate is another good solvent and a bit less stinky then acetone. Be aware that both these solvents can remove ink, so if your board is silk screened, test the chemicals first before hosing down the entire board.

A few blasts with compressed air will dry out the board and remove any junk that may have built up in the holes.

Step 2: Component Placement

After the component and board have been cleaned, you are ready to place the components onto the board. Unless your circuit is simple and only contains a few components, you will probably not be placing all the components onto the board and soldering them at once. Most likely you will be soldering a few components at a time before turning the board over and placing more. In general it is best to start with the smallest and flattest components (resistors, ICs, signal diodes, etc.) and then work up to the larger components (capacitors, power transistors, transformers) after the small parts are done. This keeps the board relatively flat, making it more stable during soldering. It is also best to save sensitive components (MOSFETs, non-socketed ICs) until the end to lessen the chance of damaging them during assembly of the rest of the circuit.

Bend the leads as necessary and insert the component through the proper holes on the board. To hold the part in place while you are soldering, you may want to bend the leads on the bottom of the board at a 45 degree angle. This works well for parts with long leads such as resistors. Components with short leads such as IC sockets can be held in place with a little masking tape or you can bend the leads down to clamp onto the PC board pads.

In the image below, a resistor is ready to solder and is held in place by slightly bent leads.

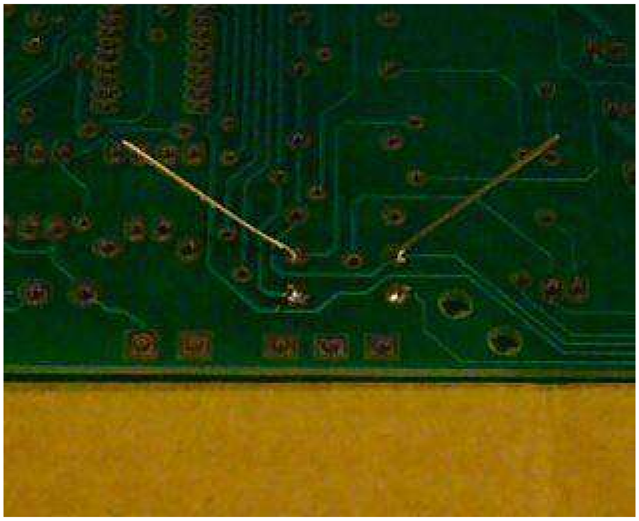

Step 3: Apply Heat

Apply a very small amount of solder to the tip of the iron. This helps conduct the heat to the component and board, but it is **not** the solder that will make up the joint. To heat the joint you will lay the tip of the iron so that it rests against both the *component lead* and the *board*. It is critical that you heat the lead and the board, otherwise the solder will simply pool and refuse to stick to the unheated item. The small amount of solder you applied to the tip before heating the joint will help make contact between the board and the lead. It normally takes a second or two to get the joint hot enough to solder, but larger components and thicker pads/traces will absorb more heat and can increase this time.

If you see the area under the pad starting to bubble, **stop** heating and remove the soldering iron because you are overheating the pad and it is in danger of lifting. Let it cool, then carefully heat it again for much less time.

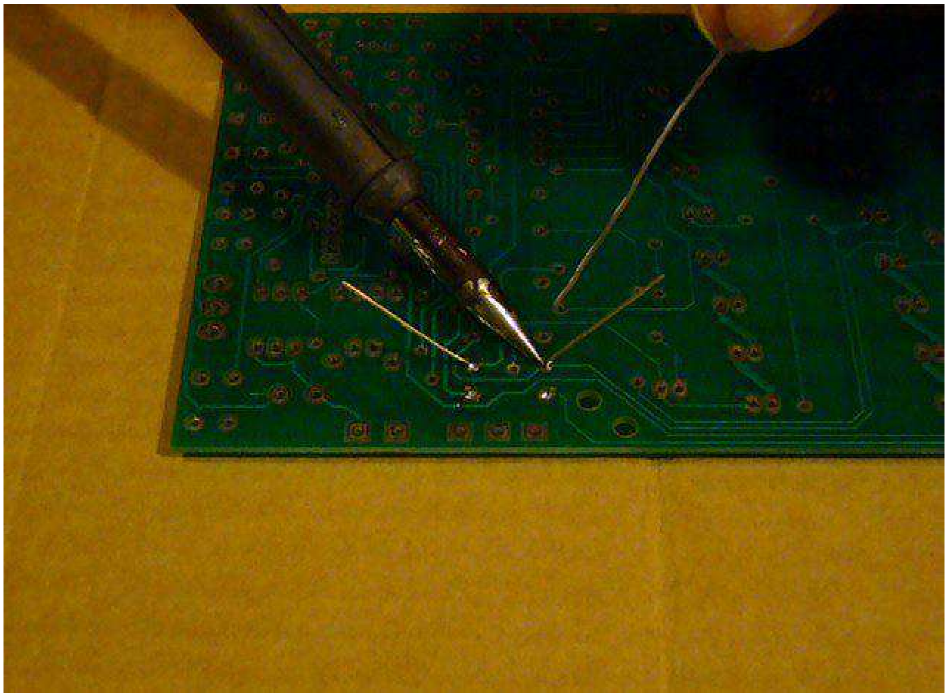

Step 4: Apply Solder to the Joint

Once the component lead and solder pad has heated up, you are ready to apply solder. Touch the tip of the strand of solder to the component lead and solder pad, but **not** the tip of the iron. If everything is hot enough, the solder should flow freely around the lead and pad. You will see the flux melt liquefy as well, bubble around the joint (this is part of its cleaning action), flow out and release smoke. Continue to add solder to the joint until the pad is completely coated and the solder forms a small mound with slightly concave sides. If it starts to ball up, you have used too much solder or the pad on the board is not hot enough.

Once the surface of the pad is completely coated, you can stop adding solder and remove the soldering iron (in that order). Don't move the joint for a few seconds as the solder needs time to cool and solidify. If you do move the joint, you will get what's called a "cold joint". This is recognized by it's characteristic dull and grainy appearance. Many cold joints can be fixed by reheating and applying a small amount of solder, then being allowed to cool without being disturbed.

Step 5: Inspect the Joint and Cleanup

Once the joint is made you should inspect it. Check for cold joints (described a little above and at length below), shorts with adjacent pads or poor flow. If the joint checks out, move on to the next. To trim the lead, use a small set of side cutters and cut at the top of the solder joint.

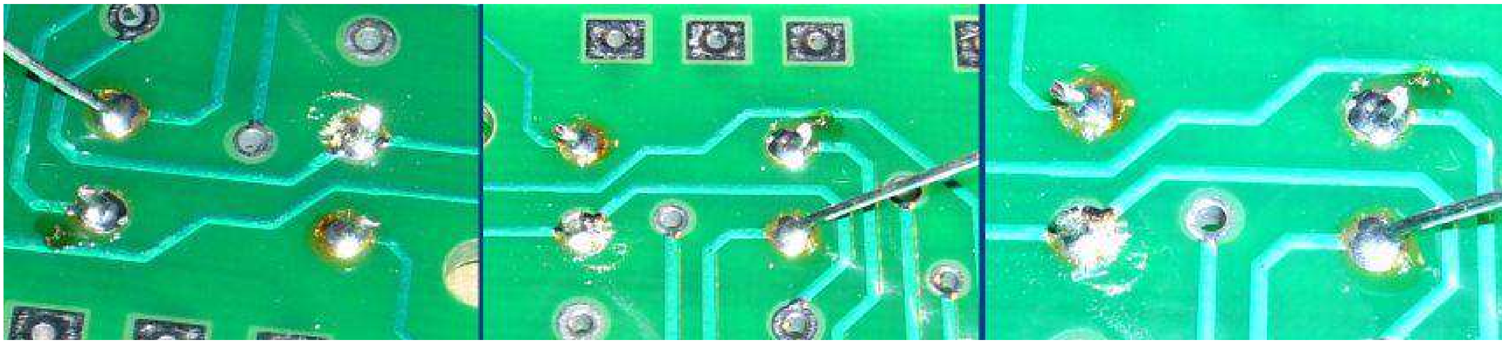

After you have made all the solder joints, it is good practice to clean all the excess flux residue from the board. Some fluxes are hydroscopic (they absorb water) and can slowly absorb enough water to become slightly conductive. This can be a significant issue in a hostile environment such as an automotive application. Most fluxes will clean up easily using methyl hydrate and a rag but some will require a stronger solvent. Use the appropriate solvent to remove the flux, then blow the board dry with compressed air.

Cold Solder Joints

A "cold solder joint" can occur when not enough heat is applied to the component, board, or both. Another common cause is a component moving before the solder has completely cooled and solidified. A cold joint is brittle and prone to physical failure. It is also generally a very high resistance connection which can affect the operation of the circuit or cause it to fail completely.

Cold joints can often be recognized by a characteristic grainy, dull gray color, but this is not always the case. A cold joint can often appear as a ball of solder sitting on the pad and surrounding the component lead. Additionally you may notice cracks in the solder and the joint may even move. Below is the shocking image of every example of a bad solder joint you will ever see. It appears that this **FM transmitter kit** was assembled using the technique of "apply solder to iron then drip onto joint". If your joints are looking like this, then **stop** and practice after rereading this page. Note that not a single of of these joints is acceptable, but amazingly, the circuit worked.

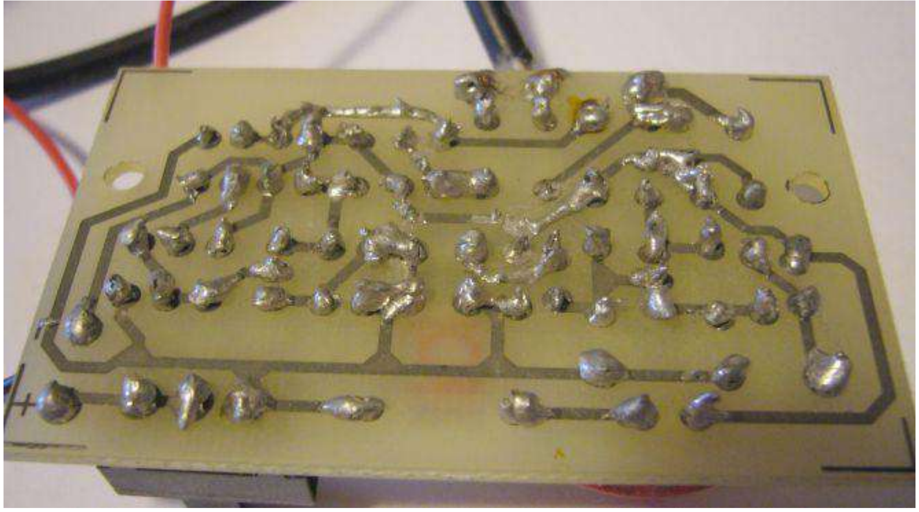

Most cold solder joints can be easily fixed. Generally all that is required is to reheat the joint and apply a little more solder. If there is already too much solder on the joint, then the joint will have to be **desoldered** and then soldered again. This is done by first removing the old solder with a desoldering tool or simply by heating it up and flicking it off with the iron. Once the old solder is off, you can resolder the joint, making sure to heat it thoroughly and keep it still as it cools.

advertisement

HOW TO SOLDER - THE QUICK, THOROUGH GUIDE

Technology (/technology/) > Soldering | by IoTalabs (/member/IoTalabs/) Follow

4,064 38 4

Posted Jun. 9, 2016 | (cc) BY-NC-SA

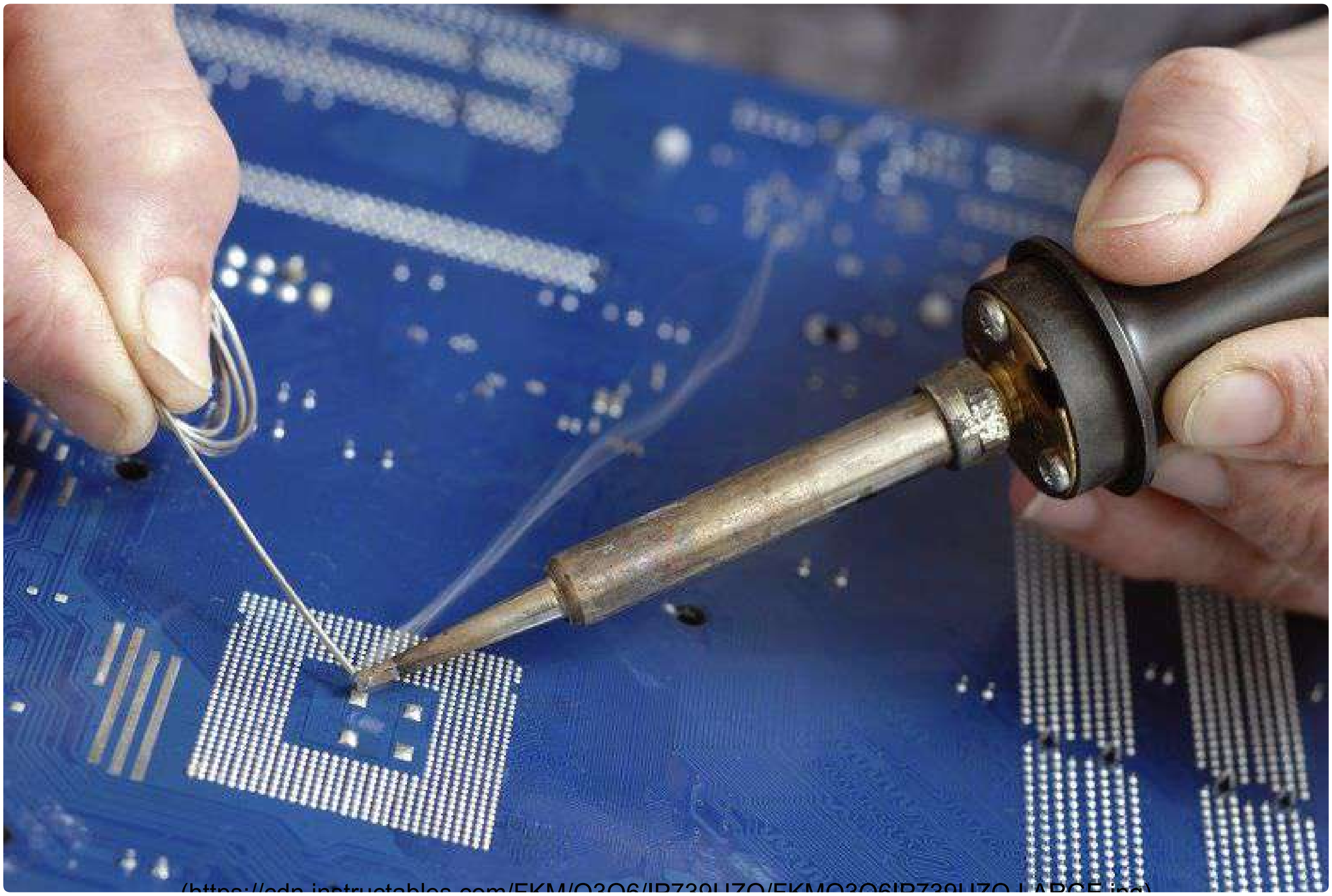

(https://cdn.instructables.com/FKM/O3O6/IPZ39UZO/FKMO3O6IPZ39UZO.LARGE.jpg)

We are IoTalabs and we are a group of Internet of Things enthusiasts that love hacking together different devices. Over the past few months we have immersed ourself in the wonderful world of hardware! In our journey, we learned the best soldering practices around.

We've been working on many projects (check out our current project http://doteverything.co/) and we wanted to share our experience on safely and correctly soldering connections!

Add Tip Ask Question

Step 1: What Is Soldering?

Soldering accomplishes a strong bond between two pieces of metal by joining them together. In this procedure, a material called solder, an alloy mixture of tin and lead, flows over two pre-heated pieces of metal and holds them together. The process is similar to welding but differs because when you weld you are fusing and melting two pieces together to make one. When you solder you are essentially ‘gluing’ two parts together with molten metal. Most metals with the exception of aluminum, white metal and porous cast iron can be soldered. Below, you will find instructions and illustrations that show you how to use a soldering iron.

Add Tip Ask Question

Step 2: Materials for Soldering

SOLDERING PASTE FLUX
Ask for ALPHA METALS
RELIABLE
SOLDERING FLUX
ZINC CHLORIDE
SOLDER
ROSIN CORE
UNGAR

(https://cdn.instructables.com/FZC/59XX/IP74OGXI/FZC59XXIP74OGXI.LARGE.jpg)

Soldering Iron

Modeler’s Vise or Frame (optional)

Solder

Damp Sponge

Flux to remove oxides

Add Tip Ask Question

Step 3: Prepare Your Workspace

Make sure your table or bench top is clear and free of obstructions. You will want as much freedom as possible to move your hand around and make adjustments.

Add Tip Ask Question

Step 4: Turn on the Soldering Iron

The soldering iron needs to be warmed up properly before it can be used to ensure clean application of the solder. Some soldering irons heat up in seconds while others take minutes. Turn it on and leave it in the stand for 2-3 minutes to be safe.

Add Tip Ask Question

Step 5: (Optional) Secure the Items You Are Soldering

Because we only have two hands, it would be nice to stabilize the item we want to solder. This is where a vice or clamp that can hold your item sturdily would come in handy. This is not required, but can greatly increase the ease and quality of your solder joints.

Add Tip Ask Question

Step 6: Cleaning the Soldering Iron

Weller
WESD51

(https://cdn.instructables.com/FVM/SZ4M/IPZ4OGX5/FVMSZ4MIPZ4OGX5.LARGE.jpg)

Because soldering irons get so hot, they oxidize and become dirty quickly. They key to reliable connections is clean components so make sure that your soldering tip and parts you are joining are clean. To accomplish this, pass the tip of your soldering iron on a wet sponge until it shines.

Add Tip Ask Question

Step 7: Apply Flux

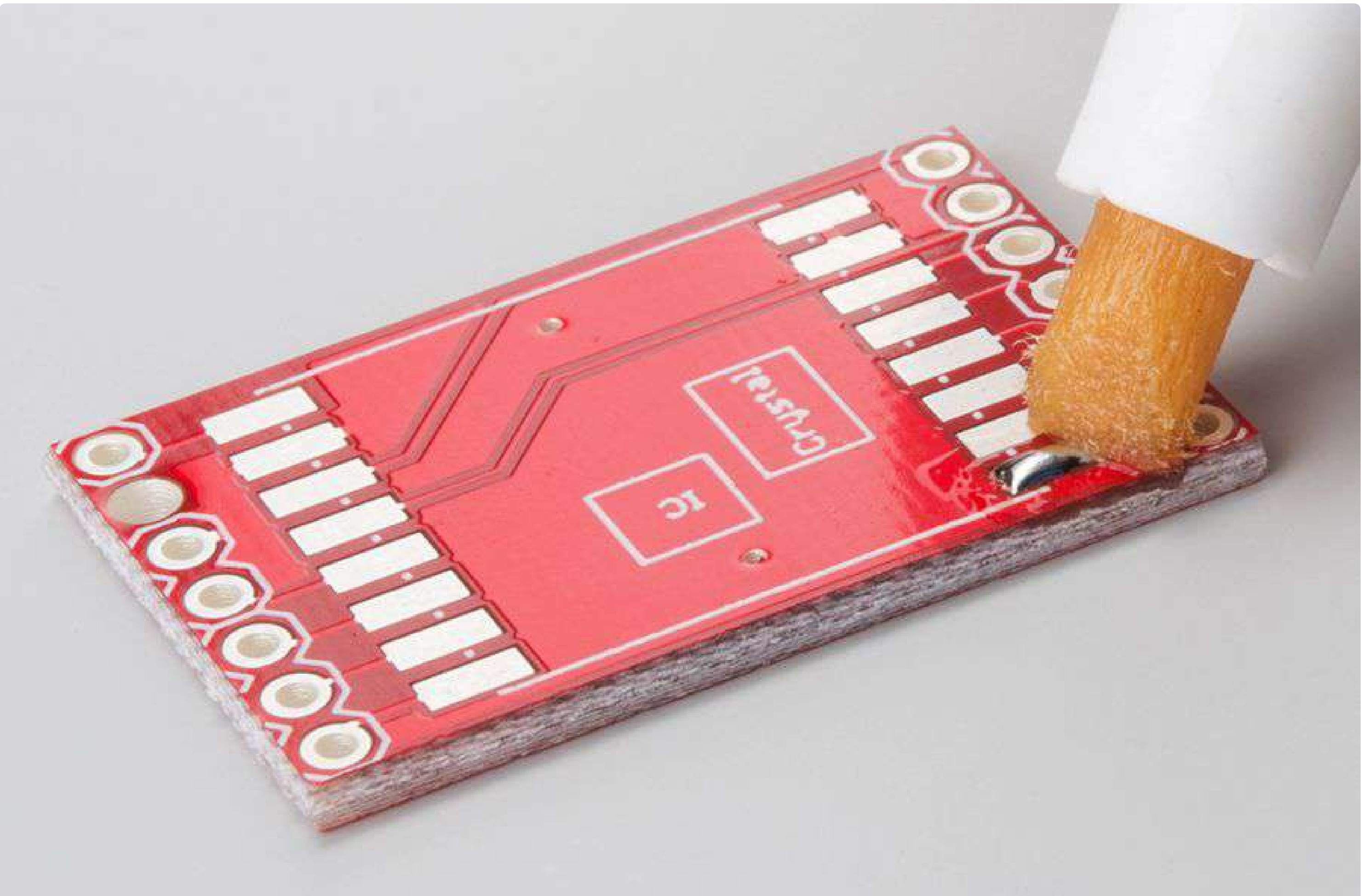

(https://cdn.instructables.com/FTD/PDPV/IPZ4OGXZ/FTDPDPVIPZ4OGXZ.LARGE.jpg)

In soldering it often becomes necessary to use materials called fluxes to help remove oxides and keep them absent while you solder. Flux needs to melt at a temperature lower than solder so that it can do its job prior to the soldering action. There are different methods to apply flux. The method you choose will be dependent on the items you are soldering.

Add Tip Ask Question

Step 8: Tin the Soldering Iron

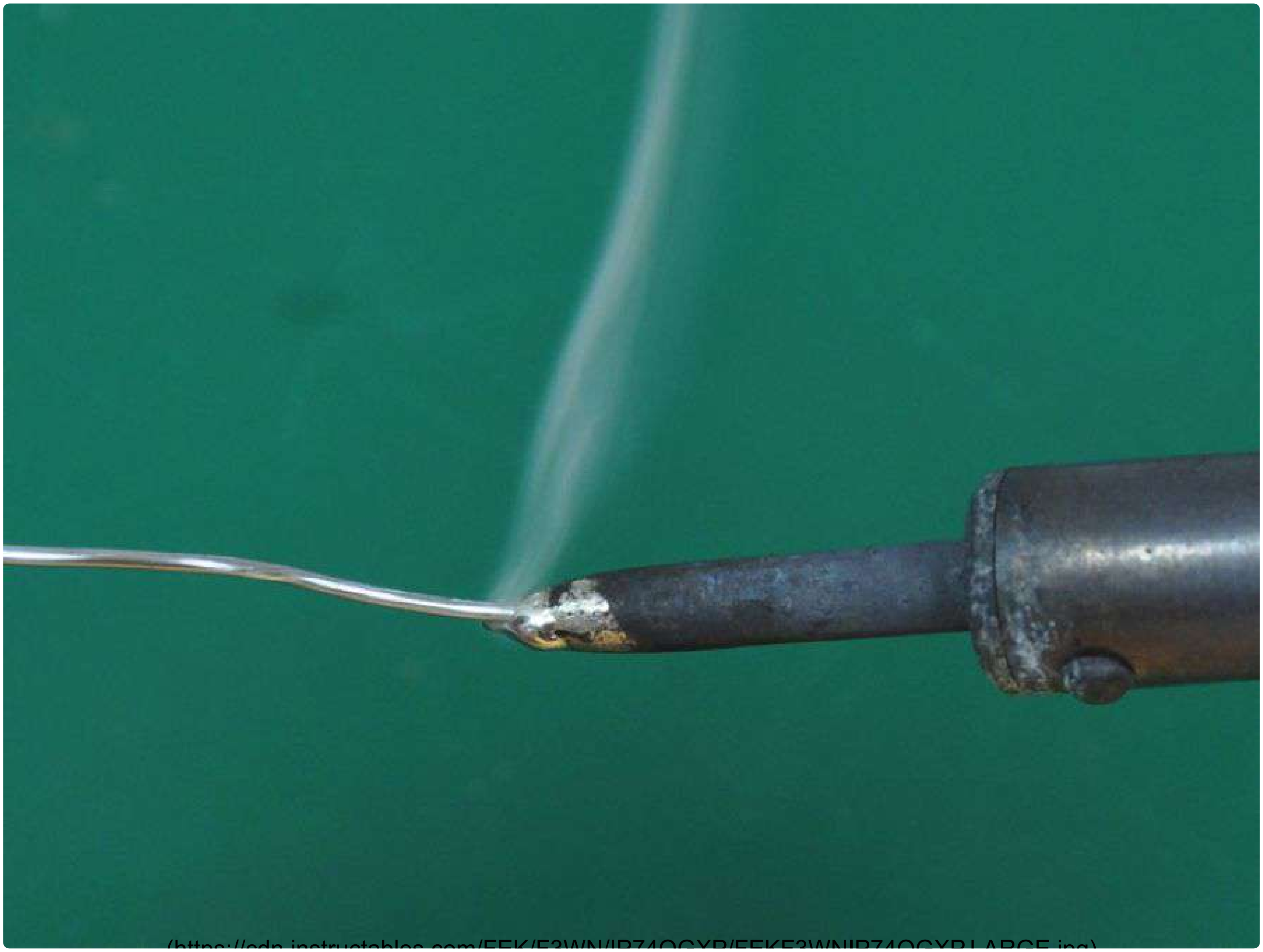

(https://cdn.instructables.com/FFK/F3WN/IP74OGYR/FFKF3WNIP74OGYR.LARGE.jpg)

If you want to know everything there is to know about how to use a soldering iron, you’ll need to know how to tin. Tinning is the process of coating a soldering tip with a thin coat of solder. Melt a thin layer of solder on your iron’s tip. This aids in heat transfer between the tip and the component you are soldering, and also gives the solder a base from which to flow from. This process may need to be repeated as you solder. You will only touch the tip of the soldering iron to the solder when you tin. Do not touch the tip of the iron to the solder while you are actually soldering.

Add Tip | Ask Question

Step 9: Start Soldering

Hold the soldering iron like you would a pen in the hand you write with and the solder in the other. Then place the tip of the soldering iron. The tip needs to touch both the wire lead and the surface so they achieve the same temperature.

Add Tip Ask Question

Step 10: Feeding the Solder

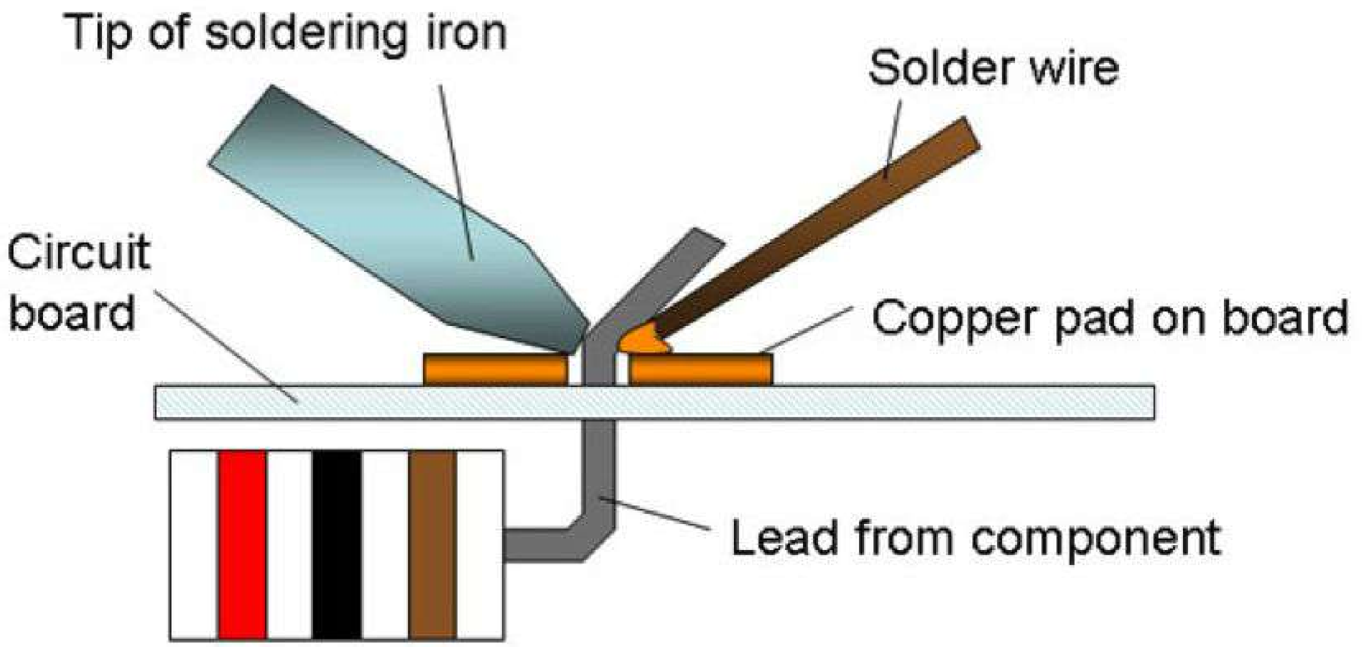
Tip of soldering iron
Solder wire
Circuit board
Copper pad on board
Lead from component

(https://cdn.instructables.com/FIW/HS90/IRZ4OGXV/FIWHS90IRZ4OGXV.LARGE.jpg)

Touch the solder to the side of the connection opposite the soldering iron after heating the area for 2-3 seconds. Then, let the solder flow only until the connection is covered.

Add Tip Ask Question

Step 11: Removing the Solder

Remove the solder first, then the iron otherwise your solder will get stuck to your connection point without the appropriate heat. Make sure the joint remains stationary while it cools.

Add Tip Ask Question

Step 12: Check to See If the Connection Is Solid

A smooth, shiny and volcano shaped joint is what you are looking for. If this isn't what you see, you'll need to reheat and feed in more solder.

Add Tip Ask Question

Step 13: Warnings

- DO NOT lay a soldering iron down on any surface. A soldering iron should either be placed on a stand or sealed with a heat resistant cap after every use.
- Soldering should be completed in a well ventilated area.
- Lead is present in most solders. Be sure to wash your hands after your project, or better yet wear gloves.
- Try not to inhale any of the solder smoke that comes off the solder. This is lead and can be dangerous.
- The tip of a soldering iron is very hot. Contact with the tip of a soldering iron would result in a nasty burn.
- Your soldering iron will perform better if kept clean. A damp sponge can be used to clean residue caused by flux material. A very small skim of flux should be applied to the iron after the cleaning.

THE BASICS OF HOW TO SOLDER

Soldering Tutorial for Beginners: Five Easy Steps

Table of Contents:

- Before you begin
- Step 1: Heat your iron to the appropriate temperature
- Step 2: Make your connection mechanically stable
- Step 3: Clean the tip of your iron
- Step 4: Apply heat and solder
- Step 5: Inspect the joint
- Common mistakes

Before you begin

You will need the following equipment:

A temperature-controlled soldering iron of 25W or more
A tip cleaner such as a brass sponge or a wet sponge

Solder (leaded is easier to work with than lead-free)
(Optional) Helping hands
Safety glasses

Note: molten solder is extremely hot! Irons heat up to around twice the boiling point of water, so be extremely careful. Wear safety glasses when soldering.

Step 1: Heat your iron to the appropriate temperature

Weller
WES51
°Fx10

For through-hole components in a circuit board, you will want 600-700 degrees fahrenheit.

If you’re soldering a larger connections, with more copper that’s dissipating heat, then you may want a slightly higher temperature. With a temperature-adjustable iron you can see what works and adjust accordingly.

Step 2: Make your connection mechanically stable

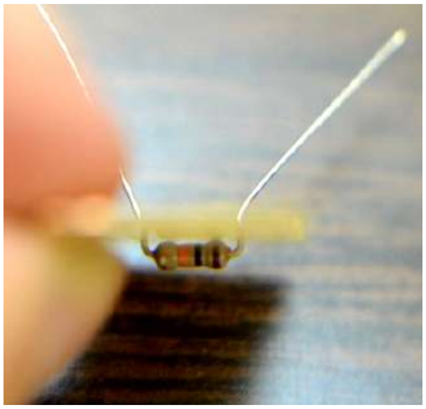

While your iron is heating up, you can work on making your components stay in place without your help. This is where your helping hands may come in handy. Use them to help you arrange your components to be stable without you needing to hold anything. If you’re soldering a through-hole component in a circuit board, you can bend the leads of the component into a V-shape so that it stays in place flush against the board. For components without long leads, you can use a piece of tape to hold them in place.

Step 3: Clean the tip of your iron

You can check whether your iron is close to its final temperature by touching your solder to the end of the iron and seeing if it melts. When you do this, you'll see some smoke come out. This is not the metal vaporizing, but rather the flux inside of it is burning. Flux is a mild acid that eats away at the oxidation layer that forms on the surface of hot metals. Without flux in the solder to clean your surfaces, you would have a very hard time getting the solder to wet.

Clean your iron using a brass sponge or a wet sponge to remove any oxidation or excess solder from the tip. Remember to always keep your tip clean while you're soldering. A dirty tip is covered in metal oxides that transfer heat badly and can result in poorly soldered joints.

Step 4: Apply heat and solder

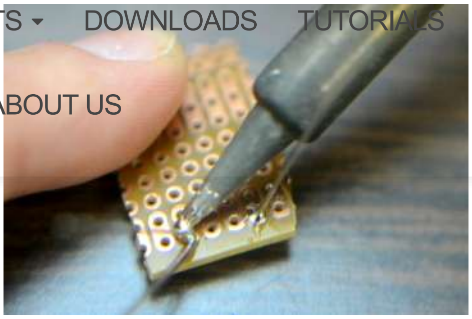

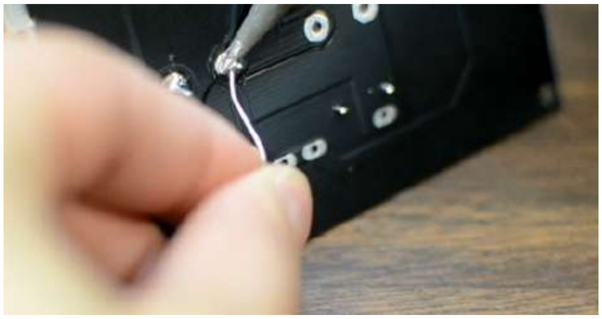

Touch the flat part of your iron's tip to one side of the joint, while feeding solder from the other side of the joint. The idea is to heat up the entire joint to the melting temperature of solder, so when you touch the solder to the surfaces it melts right on. The only way you can ensure a good electrical connection is by letting the solder flow over the components you're joining. Don't apply the solder directly to the tip of the iron. For through-hole components, hold the iron against both the lead and the pad.

Small joints can heat up within a few seconds, but larger joints can take a minute or two. Keep the joint steady for a few seconds after removing the iron, as the solder needs time to cool. Remember that the joint is hot now! don't touch it with your fingers right away.

Step 5: Inspect the joint

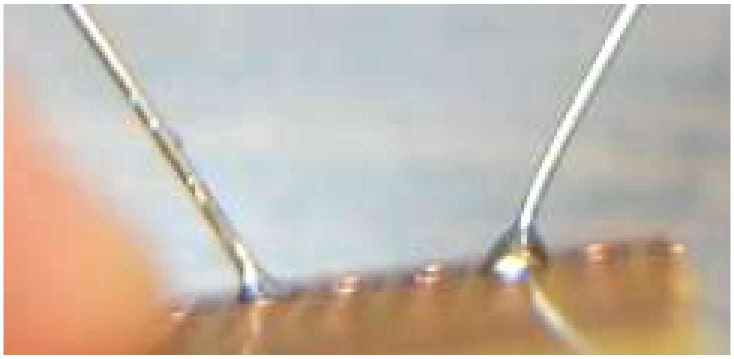

The solder should have wetted to both surfaces. On pads of a printed circuit board, the solder should be covering the entire pad, and it should make a volcano shape between the pad and the lead, indicating good adhesion. Once the joint is perfect, clip the leads of through-hole components.

Common mistakes, or what NOT to do

Now that we’ve gone over what to do, let me tell you what NOT to do.

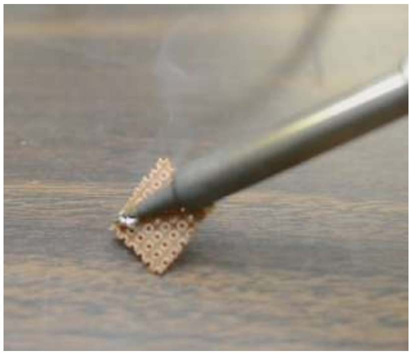

Don’t put a blob of solder on the iron and then try to transfer it to your joint. This burns away all of the flux in the solder, which prevents the flux from cleaning the metal surfaces. Furthermore, it’s easy to transfer the solder blob away from the iron without actually making it wet to the joint. I guarantee you that this will result in an awful electrical connection, so don’t do it.

Basic Tutorials

Soldering Basics
Desoldering Tutorial
Tesla Coil FAQ

tinyTesla Assembly Tutorials

tinyTesla Manual
Heat Sink Hardware Installation
Main Board's Logic Component Installation
Low Voltage Driver Board Test
Main Board's Power Component Installation
IGBT Installation

oneTeslaTS Assembly Tutorials

oneTeslaTS Manual
Endcap and Secondary Assembly Tutorial

Advanced Tutorials

How a Tesla Coil Works
(DR)SSTC Nonidealities and Losses
Thermal Analysis
DRSSTC Driver Guide
DRSSTC Bridge Design
Gate Drive Design
Resonator Design
Tesla Coil Tuning

SIGN UP FOR UPDATES

Enter your email for exclusive updates from oneTesla about new products, promotions, and other announcements

SUBSCRIBE

Limit of Liability

At oneTesla, we have a passion for making learning about hands-on electronics fun. Use our kits for educational purposes only. The operational reliability of the kits is dependent upon how well the kit is assembled, something that only the kit builder has control over. Build oneTesla kits at your own risk.

Site Links

- Contact Us
- Order FAQ
- oneTesla In the Press

One Tesla: DIY Sin…
4,519 likes

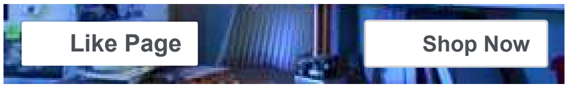

ACCREDITED BUSINESS
BBB

How To Solder: A Beginner's Guide

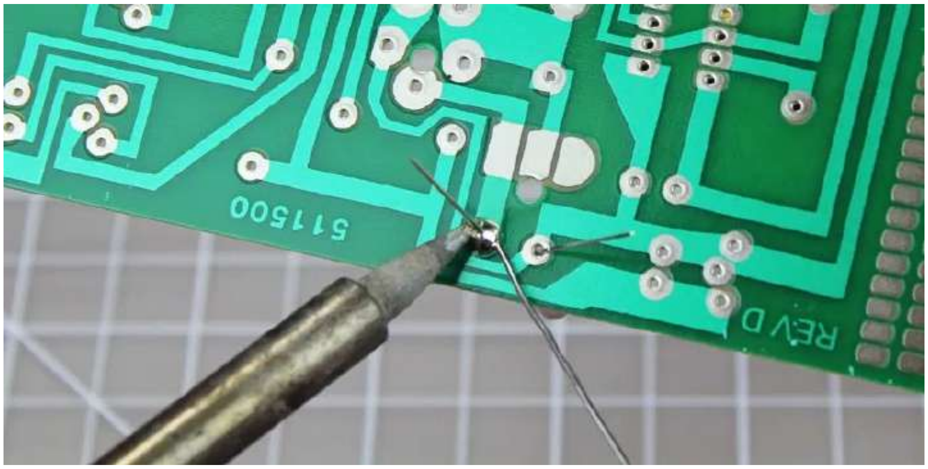

Learning how to solder w/ proper soldering techniques is a fundamental skill every maker should master. In this tutorial, we outline the basics of soldering irons, soldering stations, types of solder, desoldering and safety tips. Whether you're building a robot or working with Arduino, knowing how to solder will come in handy.

FREE EBOOK (PDF) – Learn To Solder Guide (17 pages)

What Is Soldering?

If you were to take apart any electronic device that contains a circuit board, you'll see the components are attached using soldering techniques. Soldering is the process of joining two or more electronic parts together by melting solder around the connection. Solder is a using a desoldering tool as described below.

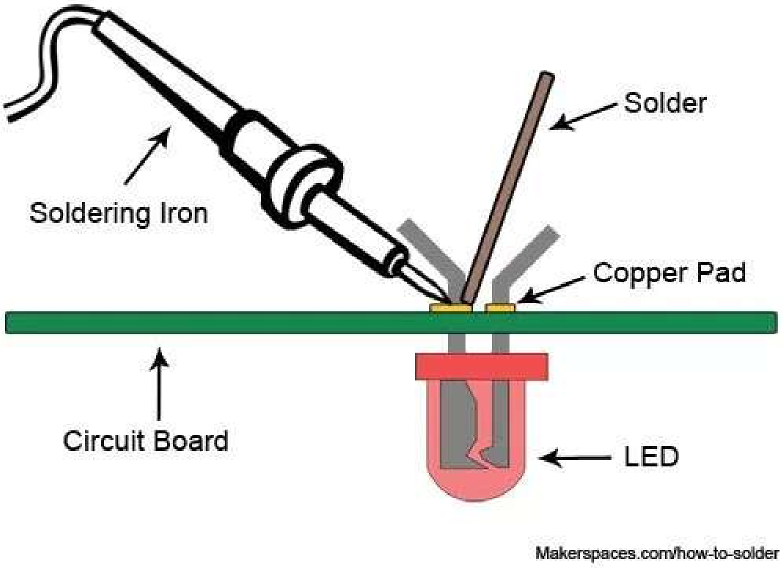
Soldering Iron
Solder
Copper Pad
Circuit Board
LED
Makerspaces.com/how-to-solder

Soldering Tools

The good thing about learning how to solder is the fact that you don't need a lot to get started. Below we'll outline the basic tools and materials you will need for most of your soldering projects.

Soldering Iron

A soldering iron is a hand tool that plugs into a standard 120v AC outlet and heats up in order to melt solder around electrical connections. This is one of the most important tools used in soldering and it can come in a few variations such as pen or gun form. For beginners, it's recommended that you use the pen style soldering iron in type of soldering iron because it can heat up to 896′ F which is extremely hot.

Soldering Station

A soldering station is a more advanced version of the basic standalone soldering pen. If you are going to be doing a lot of soldering, these are great to have as they offer more flexibility and control. The main benefit of a soldering station is the ability to precisely adjust the temperature of the soldering iron which is great for a range of projects. These stations can also create a safer workspace as some include advanced temperature sensors, alert settings and even password protection for safety.

Soldering Iron Tips

At the end of most soldering irons is an interchangeable part known as a soldering tip. There are many variations of this tip and they come in a wide variety of shapes and sizes. Each tip is used for a specific purpose and offers a distinct advantage over another. The most common tips you will use in electronics projects are the conical tip and the chisel tip.

Conical Tip – Used in precision electronics soldering because of the fine tip. Because of its pointed end, it’s able to deliver heat to smaller areas without affecting its surroundings.

Chisel Tip – This tip is well-suited to soldering wires or other larger components because of its broad flat tip.

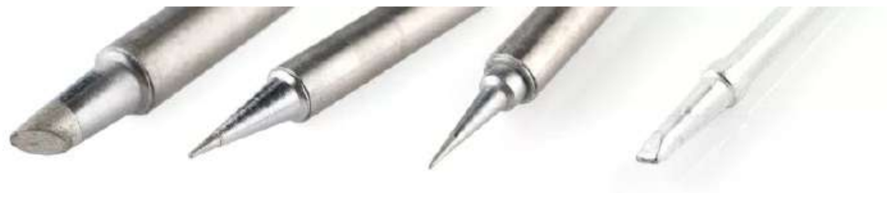

Image Credit – Sparkfun.com

Brass or Conventional Sponge

it did when it was new. You could use a conventional wet sponge but this tends to shorten the lifespan of the tip due to expansion and contraction. Also, a wet sponge will drop the temperature of the tip temporarily when wiped. A better alternative is to use a brass sponge as shown on the left.

Soldering Iron Stand

A soldering iron stand is very basic but very useful and handy to have. This stand helps prevent the hot iron tip from coming in contact with flammable materials or causing accidental injury to your hand. Most soldering stations come with this built in and also include a sponge or brass sponge for cleaning the tip.

Solder

Solder is a metal alloy material that is melted to create a permanent bond between electrical parts. It comes in both lead and lead-free variations with diameters of .032″ and .062″ being the most common. Inside the solder core is a material known as flux which helps improve electrical contact and its mechanical strength.

For electronics soldering, the most commonly used type is lead-free rosin core solder. This type of solder is usually made up of a Tin/Copper alloy. You can also use leaded 60/40 (60% tin, 40% lead) rosin core solder but it's becoming less popular due to health concerns. If you do use lead solder, make sure you have proper ventilation and that you wash your hands after use.

Net Wt.: 8 oz. (226.7g)
64-009
60-40 ROSIN-CORE SOLDER
LF-99 Lead-Free Solder
99.3% tin 0.7% copper, 4.84g/
5 ft., 0.031" Dia. Rosin Core
7 56619 00705 2
WARNING: This product contains or produces a chemical known to the state of

When buying solder, make sure NOT to use acid core solder as this will damage your circuits and components. Acid core solder is sold at home improvement stores and is mainly used for plumbing and metal working.

As mentioned earlier, solder does come in a few different diameters. The thicker diameter solder (.062″) is good for soldering larger joints more quickly but it can make soldering smaller joints difficult. For this reason, it’s always a good idea to have both sizes on hand for your different projects.

Helping Hand (Third Hand)

A helping hand is a device that has 2 or more alligator clips and sometimes a magnifying glass/light attached. This clips will assist you by holding the items you are trying to solder while you use the soldering iron and solder. A very helpful tool to have in your makerspace.

Soldering Safety

Now that you know what tools and materials are required, it’s time to briefly discuss ways of staying safe while soldering.

Soldering irons can reach temperatures of 800′ F so it’s very important to know where your iron is at all times. We always recommend you use a soldering iron stand to help prevent accidental burns or damage.

Make sure you are soldering in a well ventilated area. When solder is heated, there are fumes released that are harmful to your eyes and lungs. It's recommended to use a fume extractor which is a fan with a charcoal filter that absorbs the harmful solder smoke.

It's always a good idea to wear protective eye wear in case of accidental splashes of hot solder. Lastly, make sure to wash your hands when done soldering especially if using lead solder.

Tinning The Tip

Before you can start soldering, you need to prep your soldering iron by tinning the tip with solder. This process will help improve the heat transfer from the iron to the item you're soldering. Tinning will also help to protect the tip and reduce wear.

Step 1: Begin by making sure the tip is attached to the iron and screwed tightly in place.

Step 3: Wipe the tip of the soldering iron on a damp wet sponge to clean it. Wait a few seconds to let the tip heat up again before proceeding to step 4.

Step 4: Hold the soldering iron in one hand and solder in the other. Touch the solder to the tip of the iron and make sure the solder flows evenly around the tip.

You should tin the tip of your iron before and after each soldering session to extend its life. Eventually, every tip will wear out and will need replacing when it becomes rough or pitted.

How To Solder

To better explain how to solder, we're going to demonstrate it with a real world application. In this example, we're going to solder an LED to a circuit board.

Step 1: Mount The Component – Begin by inserting the leads of the LED into the holes of the circuit board. Flip the board over and bend the leads outward at a 45′ angle. This will help the component make a better connection with the copper pad and prevent it from falling out while soldering.

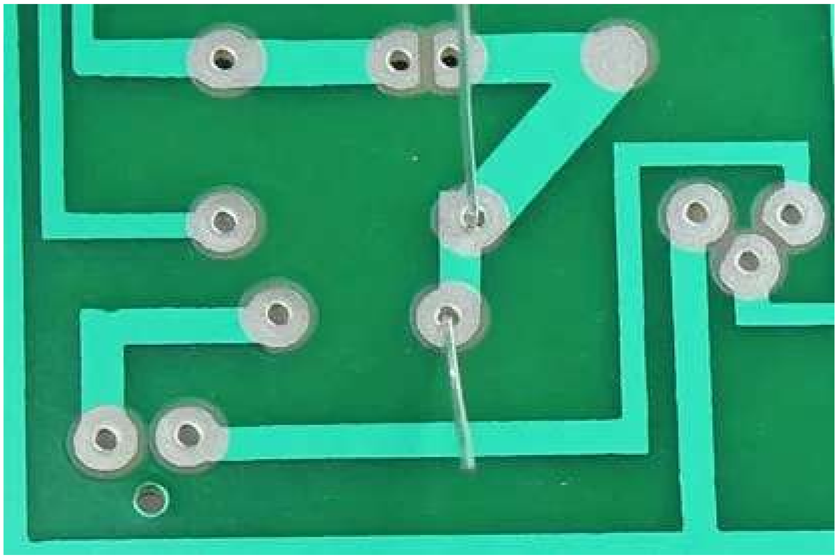

resistor lead at the same time. You need to hold the soldering iron in place for 3-4 seconds in order to heat the pad and the lead.

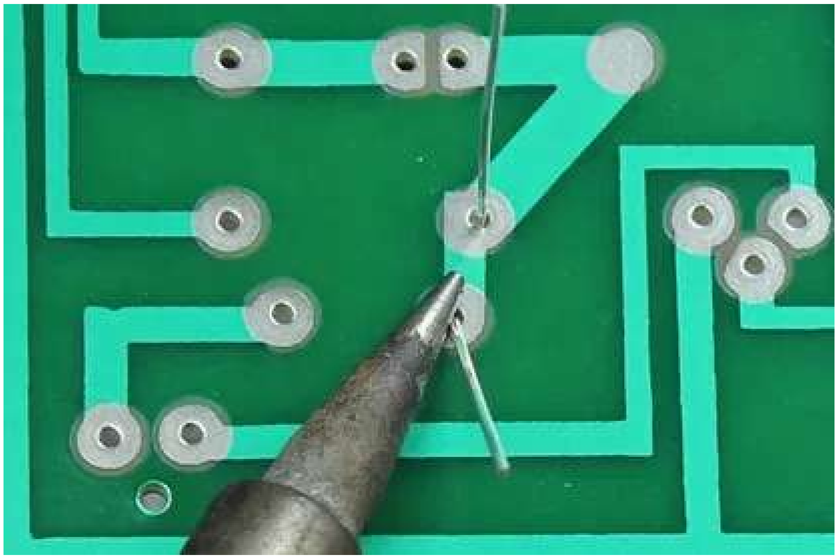

Step 3: **Apply Solder To Joint** – Continue holding the soldering iron on the copper pad and the lead and touch your solder to the joint. **IMPORTANT** – Don’t touch the solder directly to the tip of the iron. You want the joint to be hot enough to melt the solder when it’s touched. If the joint is too cold, it will form a bad connection.

Step 4: **Snip The Leads** – Remove the soldering iron and let the solder cool down naturally. Don't blow on the solder as this will cause a bad joint. Once cool, you can snip the extra wire from leads.

A proper solder joint is smooth, shiny and looks like a volcano or cone shape. You want just enough solder to cover the entire joint but not too much so it becomes a ball or spills to a nearby lead or joint.

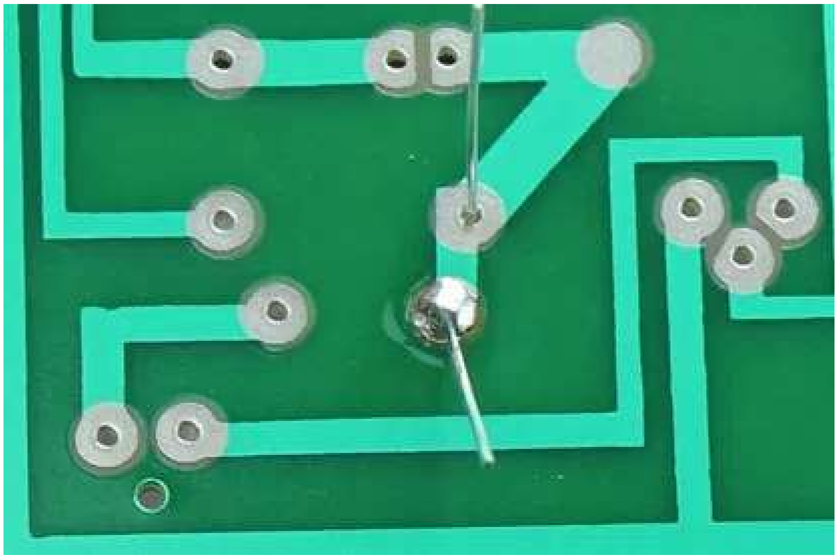

How To Solder Wires

Begin by removing the insulation from the ends of both wires you are soldering together. If the wire is stranded, twist the strands together with your fingers.

Make sure your soldering iron is fully heated and touch the tip to the end of one of the wires. Hold it on the wire for 3-4 seconds.

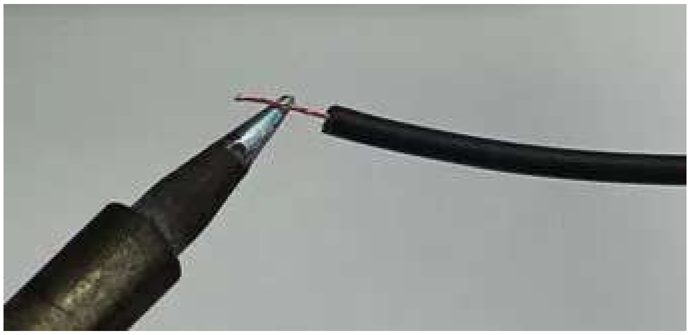

Keep the iron in place and touch the solder to the wire until it's fully coated. Repeat this process on the other wire.

Hold the two tinned wires on top of each other and touch the soldering iron to both wires. This process should melt the solder and coat both wires evenly.

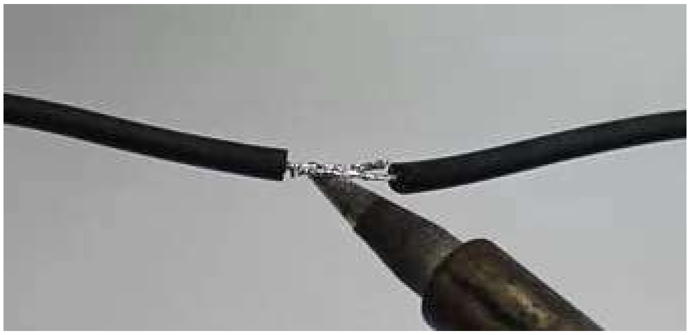

Remove the soldering iron and wait a few seconds to let the soldered connection cool and harden. Use heat shrink to cover the connection.

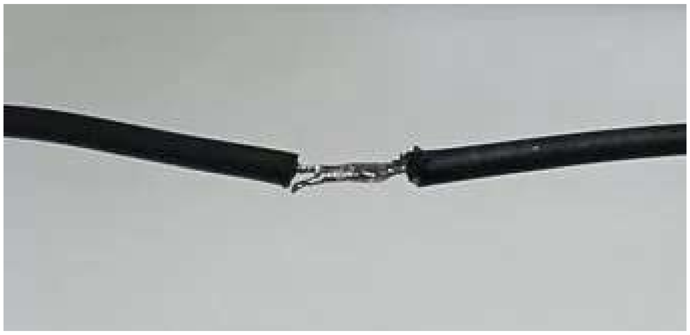

Desoldering

The good thing about using solder is the fact that it can be removed easily in a technique known as desoldering.

To desolder a joint, you will need solder wick which is also known as desoldering braid.

Desoldering Braid
RoHS
6402090
RadioShack

Step 1 – Place a piece of the desoldering braid on top of the joint/solder you want removed.

Step 2 – Heat your soldering iron and touch the tip to the top of the braid. This will heat the solder below which will then be absorbed into the desoldering braid. You can now remove the braid to see the solder has been extracted and removed. Be careful touching the braid when you are heating it because it will get hot.

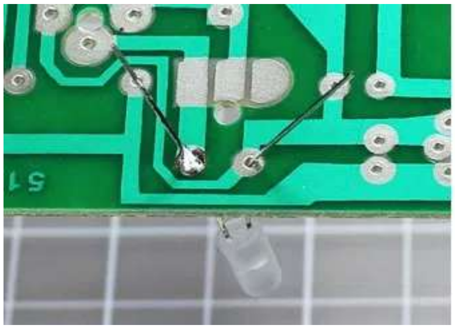

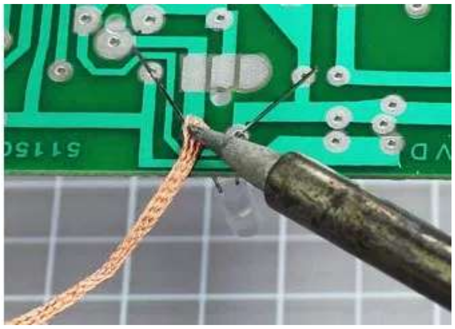

solder with a press of a button.

To use, press the plunger down at the end of the solder sucker. Heat the joint with your soldering iron and place the tip of the solder sucker over the hot solder. Press the release button to suck up the liquid solder. In order to empty the solder sucker, press down on the plunger.

Soldering & Electronics

- Digikey
- Sparkfun
- Adafruit
- Jameco
- Velleman
- Mouser

FREE – Learn To Solder Guide (17 pages)

Makerspaces.com

Makerspaces.com was built to help schools and libraries start and run acquire the skills needed for the 21st century.

Adafruit Guide To Excellent Soldering

Created by Bill Earl

Last updated on 2014-06-28 08:30:54 PM EDT

Guide Contents

Guide Contents 2
Tools 4
Building a Soldering Toolkit 4
Choosing a Soldering Iron 4
- Basic Irons 5
- Better Irons 5
- Best Irons 5

Irons to avoid 6
- For emergencies only: 6
- Not for circuit board use: 6

Essential Tools and Supplies: 6
- Stand 7
- Solder 7
- Diagonal Cutters 7

Other Handy Tools and Supplies 7
- Vise 8
- Third Hand 8
- Solder Sucker 9
- Solder Wick 9

Preparation 10
- Heat the Iron 10
- Clean the Iron 10
- Tin the Tip 11
- Make sure that the joint is clean 11
- Immobilize the Joint 11
- Steady the Board 12

Making a good solder joint 13
- Heat the joint 13
- Apply the solder 14
- Let It Flow 14

Let It Cool 14
Trim the Lead 14
Congratulations! 15
Problems? 16
Surface Mount Components 17
Immobilize the Joint 17
Heat the Joint 17
Apply the Solder 18
Let it Flow 18
Let it Cool 18
Problems? 18
Common Soldering Problems 20
The Ideal Solder Joint 20
Disturbed Joint 21
Cold Joint 22
Overheated Joint 24
Insufficient Wetting (Pad) 25
Insufficient Wetting (Pin) 27
Insufficient Wetting 28
(Surface Mount) 28
Solder Starved 29
Too Much Solder 30
Untrimmed Leads 31
Solder Bridge 32
Lifted Pad 33
Repairing a Lifted Pad 34
Stray Solder Spatters 34
All of the Above! 35

Tools

Building a Soldering Toolkit

If you are just getting started in Electronics, Ladyada's Electronics Toolkit (http://adafru.it/136) (pictured above) is a great kit full of quality tools - including everything you need to make great solder joints. If you would rather build your toolkit piece-by-piece, read on:

Choosing a Soldering Iron

There are many types of soldering irons. For most Adafruit kits and projects, you will want a pencil-style soldering iron with 25 watts or more.

An under-powered iron is a poor investment. It will end up costing you more in ruined kits and damaged components.

- It will take longer to heat the joint, allowing heat to spread to the component being soldered - potentially overheating and damaging the component.
- Longer heating times will also give more time for oxides to form on the surfaces being soldered. This will prevent the solder from flowing and result in a poor joint.
- Longer recovery times between joints can result in frustration, 'cold joints' or both.

You don't need to spend a fortune to get a good iron. Advanced features such as temperature control and interchangeable tips are nice to have, but not essential for hobbiest-level work.

Basic Irons

There are many basic pencil style irons that are suitable for hobbiest use. But you will need one that is capable of heating the joints quickly enough. Choose an iron with 25 watts at a minimum.

Better Irons

An adjustable temperature iron with a little more power will give you a bit more control and allow you to work faster. The Adjustable 30W 110v Soldering Iron (http://adafru.it/180) in the store is an excellent choice.

This iron is also available as part of Ladyada's Electronics Toolkit (http://adafru.it/136), which contains many other essential soldering tools.

Best Irons

A professional-style temperature-controlled iron with interchangeable tips and 50 watts or more of power is a joy to work with. Feedback control keeps the tip temperature at precisely the level you set. The extra watts speed recovery time so that you can work faster. Interchangeable tips let you select the ideal tip shape for specialized work.

The 65 watt Hakko FX-888 (http://adafru.it/303) is an excellent professional quality soldering iron. The Weller WES51 or WESD51 are also excellent choices for serious electronics work.

Irons to avoid

In addition to underpowered irons, there are several types of irons to avoid for general circuit-board work.

For emergencies only:

These irons are handy for occasions when you have no place to plug in a regular soldering iron. But they are not the best choice for a primary soldering tool:

- **Butane Powered Irons** have plenty of power but are difficult to control.
- **Battery Powered Irons** are generally underpowered for most work.

Not for circuit board use:

These tools are not suitable for circuit board work:

- **Torches** of any kind are not suitable for electronics work and will damage your circuit boards.
- **Soldering Guns** are OK for working with heavy gauge wires, but don't have the precision necessary for soldering delicate electronics components.
- ***Cold-Heat™*** Irons inject current into the joint to heat the tip. This current can be damaging to sensitive electonic components. Avoid these irons for electronics work.

Essential Tools and Supplies:

These tools are the bare-minimum essentials required for soldering:

Stand

If your soldering iron does not have a built-in stand, you will need a safe place to rest the hot iron between uses. A Soldering Iron Stand (http://adafru.it/150) will keep your iron from rolling around and protect both you and your work surface from burns.

Most stand holders come with a sponge and tray for cleaning your soldering iron.

ROSIN CORE
GUARANTEED VIRGIN METALS
60/40
GUARANTEED ALLOY
4 OZ.
.031 DIA

Solder

Standard 60/40 lead/tin Rosin Core Solder (http://adafru.it/145) is the easiest type to work with.

170M Xcelite® U.S.A.

Diagonal Cutters

You will also need a pair of Diagonal Cutters (http://adafru.it/152) for trimming component leads after soldering.

Other Handy Tools and Supplies

These are some other tools and supplies you might find useful when working on soldering projects.

Vise

A vise holds your work steady as you solder. This is important for both safety and sound joints. The Panavise Jr (http://adafru.it/151) is an ideal size for most Adafruit kits and projects.

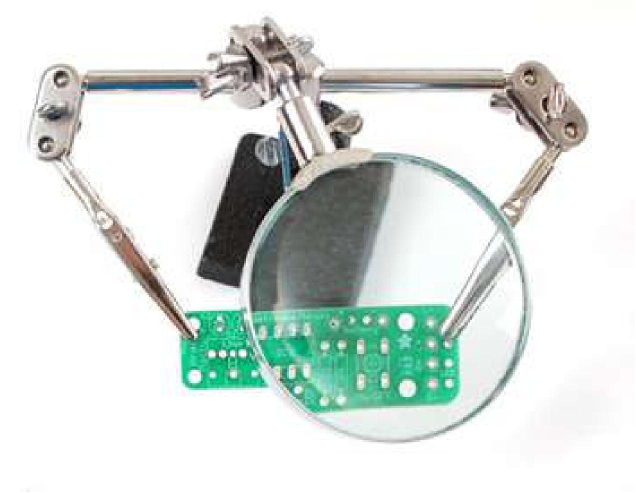

Third Hand

A Helping Third Hand (http://adafru.it/291) Tool is a good for smaller boards, or to hold things in place while terminating or splicing wires.

Solder Sucker

A Solder Sucker (http://adafru.it/148) is a very helpful tools for removing excess solder or when you need to de-solder a joint. As the name implies, this device literally sucks the solder out of the joint.

ROHS
35+ YEARS SPIRIG PATENTED VACUUMISATION WICK MANUFACTURE
no clean
size AB (3)
1,5 m / 5 ft
DESOLDERING BRAID
3S-WICK
part: LS/AB-1.5 m
TRESSE à DESSOUDER
LOTSAUGLITZE
antistatic
swiss

Solder Wick

Solder Wick (http://adafru.it/149) is another way to clean excess solder from a joint. Unlike the solder sucker, the wick soaks up the molten solder.

Preparation

Lockout
F/C
701
Weller
WESD51
Output for PES51
Iron Only.
Use Only Original
Weller Tips.
ESD SAFE

Heat the Iron

Plug an and/or turn on your soldering iron to warm up. If you are using a temperature controlled iron, set it to 700F/370C for 60/40 or 750F/400C for lead-free solder.

While the iron is heating dampen the sponge with a little bit of water.

Clean the Iron

Wipe the tip of the hot iron on the damp sponge to clean off any oxidation.

Do not use files or abrasives to clean the tip. It will damage the plating and ruin the tip.

Tin the Tip

Apply a small amount of solder to the tip and wipe again to tin the tip. You should have a thin, shiny layer of molten solder on the tip of your iron.

If the tip is badly oxidized and difficult to tin, it can usually be reconditioned with some tip-tinning paste.

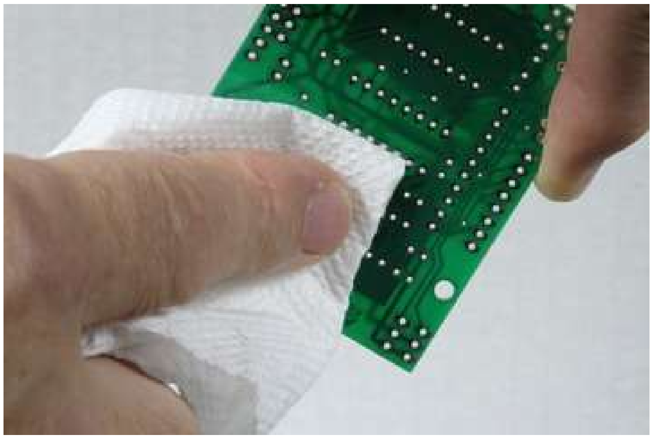

Make sure that the joint is clean

Dirt, oxidation and oily fingerprints can prevent the solder from wetting the solder-pad to create a solid joint. All Adafruit boards are plated to prevent oxidation, but if your board appears dirty from storage or handling, wipe it down with a little isopropyl alcohol.

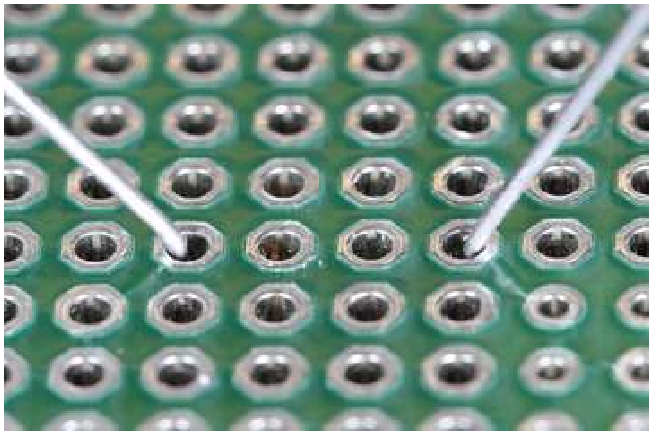

Immobilize the Joint

This is very important! The parts being joined must not move during the soldering process. If there is any movement as the molten solder is solidifying, you will end up with an unreliable 'cold joint'.

Most through-hole components can be immobilized by simply bending the leads on the solder-side of the hole.

Steady the Board

A vise is a good way to keep the board from moving around while you try to solder it.

Once the joint is clean and immobilized, you are ready to apply the solder.

Making a good solder joint

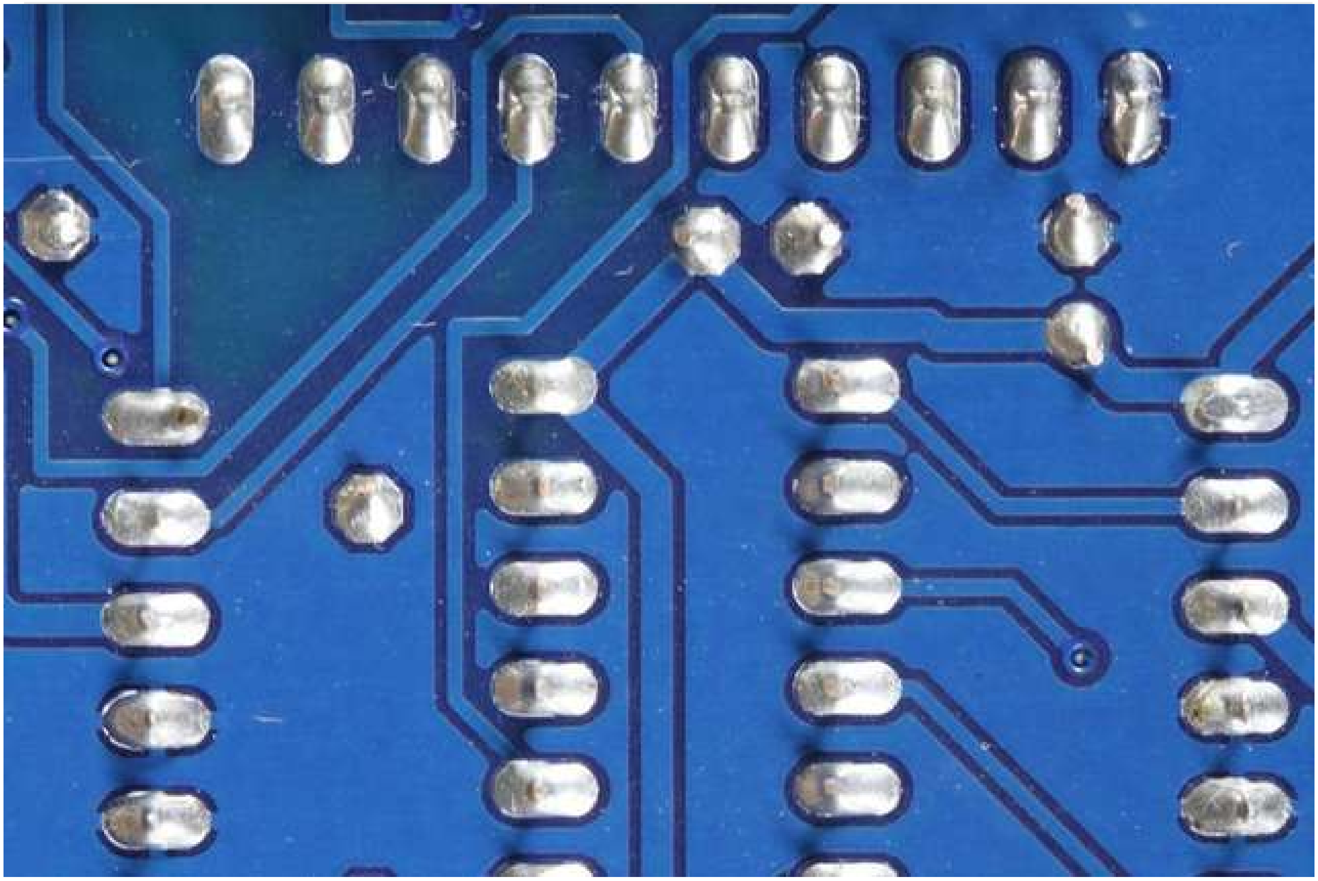

Once you have prepared the your tools and the joint to be soldered, making a good solder joint requires just a few simple steps.

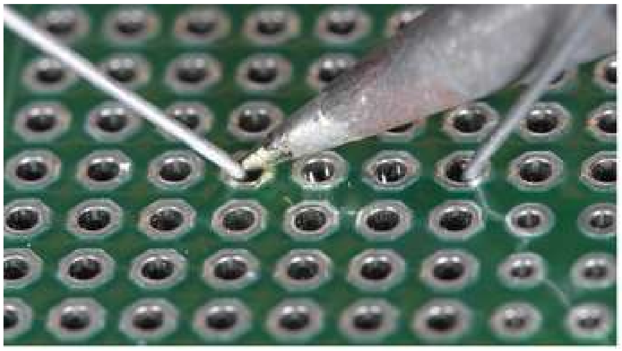

Heat the joint

Heat the joint with the tip of the iron. Be sure to heat both the solder pad and the component lead or pin. A small drop of solder on the tip will help to transfer the heat to the joint quickly.

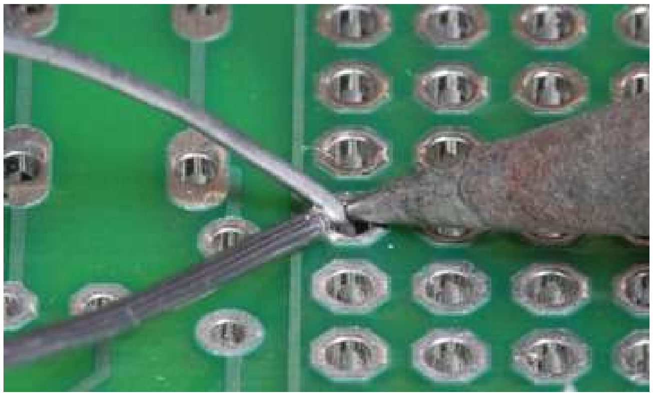

Apply the solder

Touch the end of the solder to the joint so that it contacts both the solder pad and the component lead or pin. It should melt and flow smoothly onto both the pin and the pad. If the solder does not flow, heat the joint for another second or two and try again.

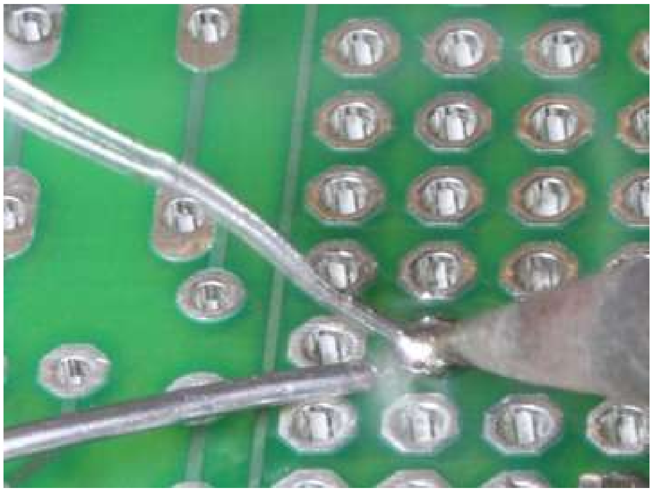

Let It Flow

Keep heating the solder and allow it to flow into the joint. It should fill the hole and flow smoothly onto both the solder pad and the pin or component lead.

Let It Cool

Once enough solder has been added to the joint and it has flowed well onto both the component lead and the solder pad, remove the iron from the joint and allow it to cool undisturbed.

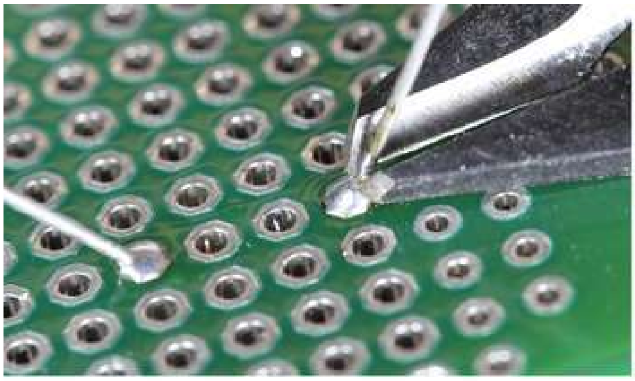

Trim the Lead

Use your diagonal cutters to trim the lead close to the board.

Note: This step applies only to components with wire leads. It is not necessary to trim the pins on Integrated circuit chips or sockets.

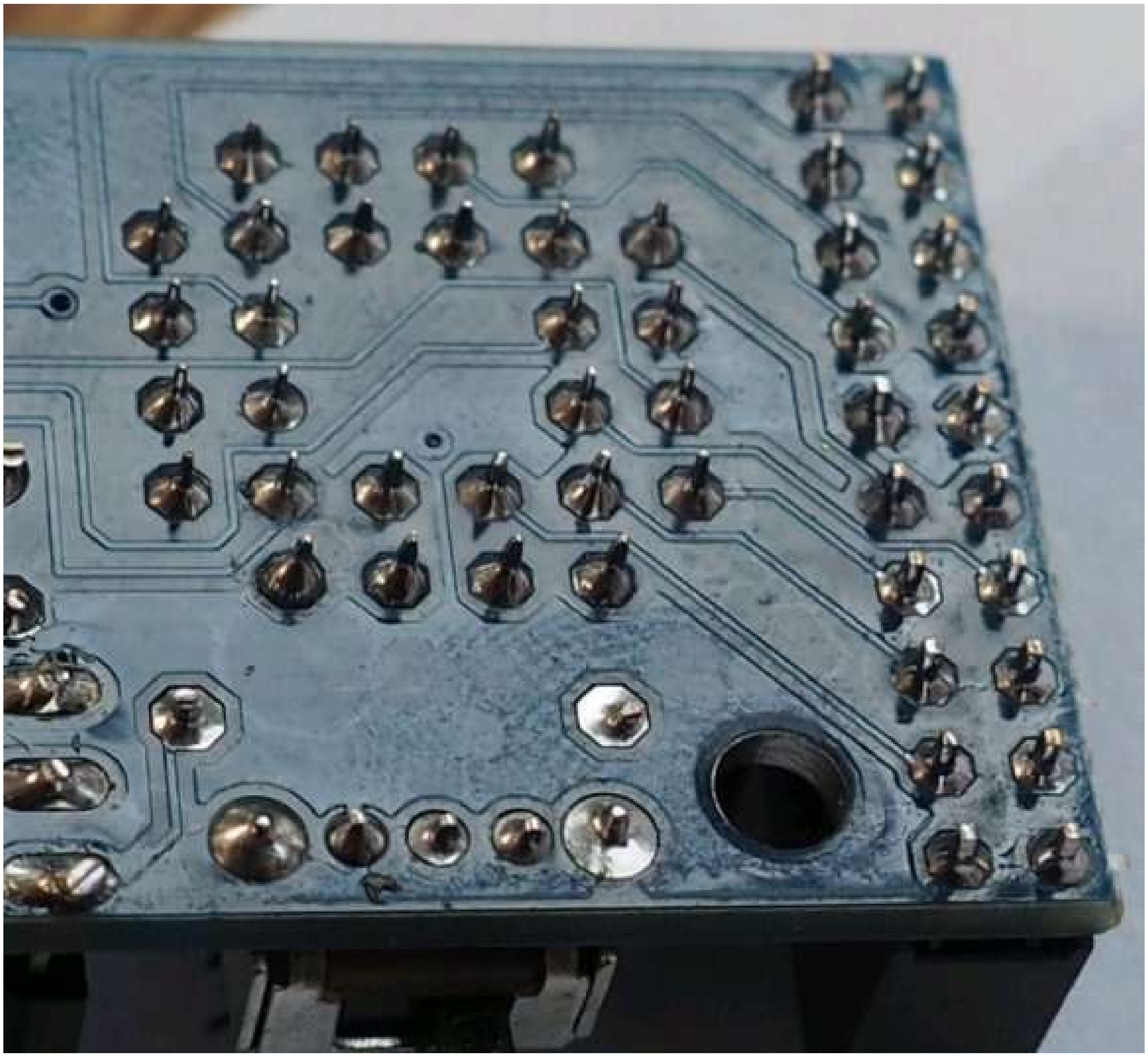

Congratulations!

Reward yourself with a Soldering Badge (http://adafru.it/465).

Problems?

The last page of this guide illustrates a number of common soldering problems with advice on prevention and repair.

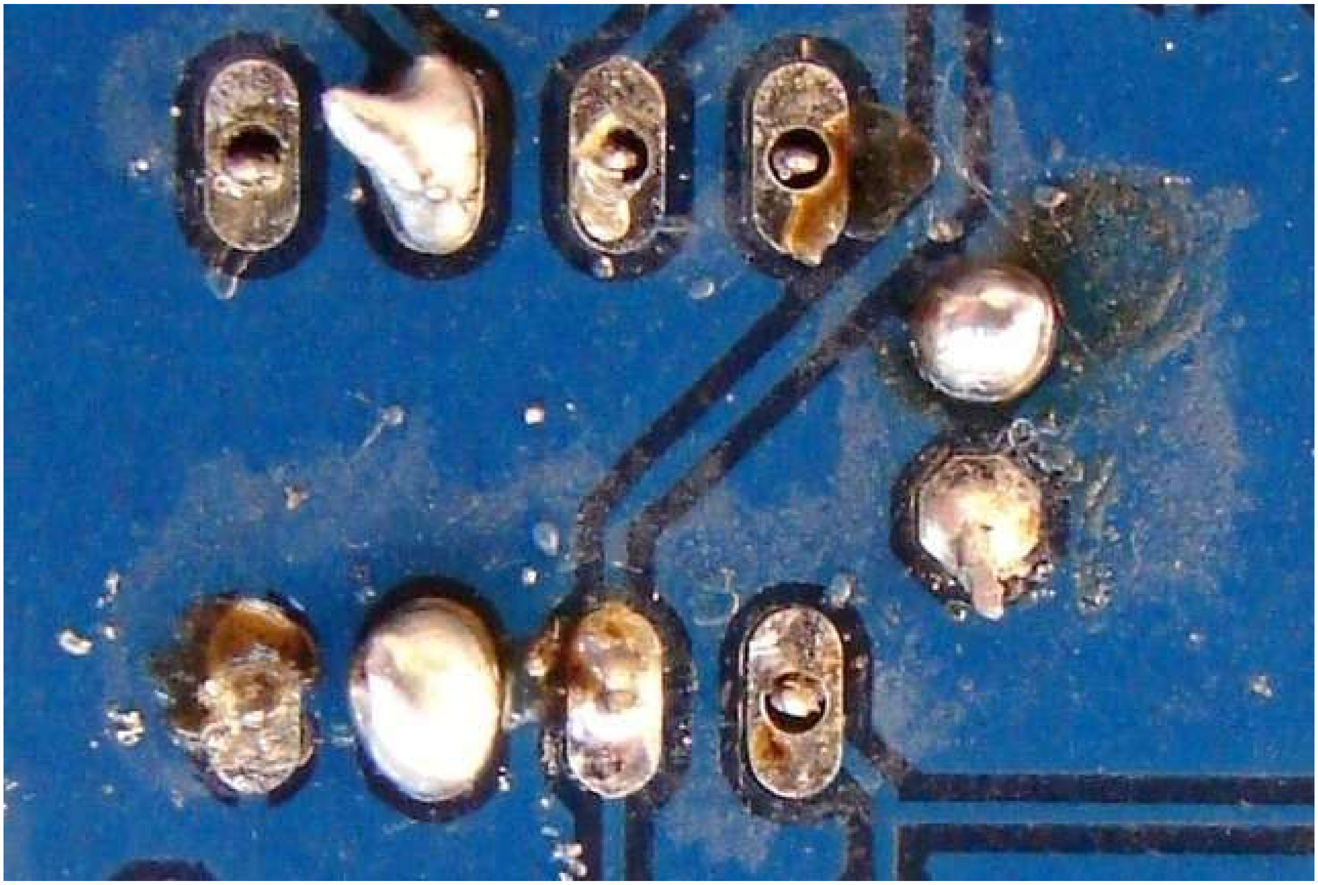

Surface Mount Components

The previous page showed how to make a good through-hole joint. But more and more components are only available in surface mount form these days. Not all surface mount packages are easily worked by hand, but there are plenty that can be managed with the same basic tools used for through-hole soldering.

Let's start with a surface-mount part common to several Adafruit kits: The SD Card Holder:

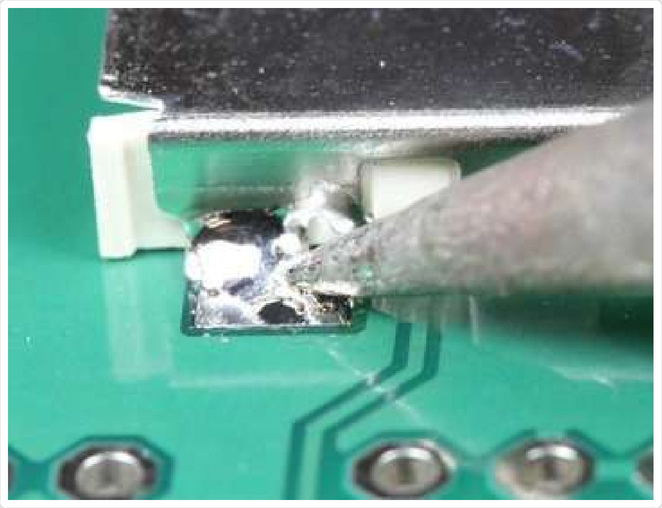

Immobilize the Joint

Unlike many surface mount components, immobilizing the SD card holder is relatively easy. There are small pegs on the back that fit into positioning holes in the board. Once it is in place, solder the four small corner tabs to make it permanent.

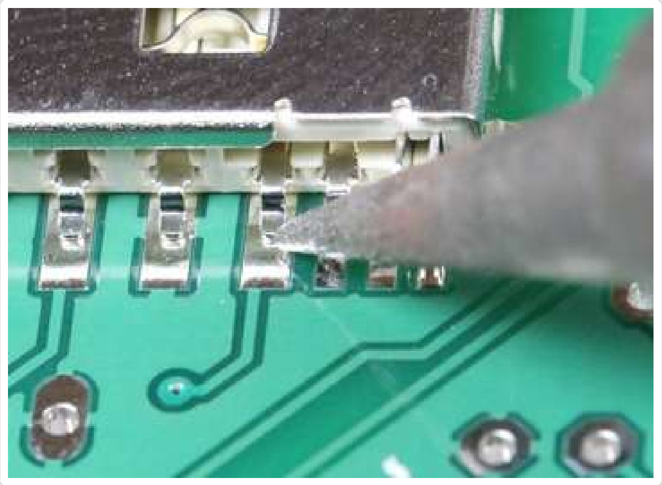

Heat the Joint

Start by putting the tip of the hot iron on the solder pad adjacent to the pin. The pad will take longer to heat, so we apply most of the heat to the pad to start.

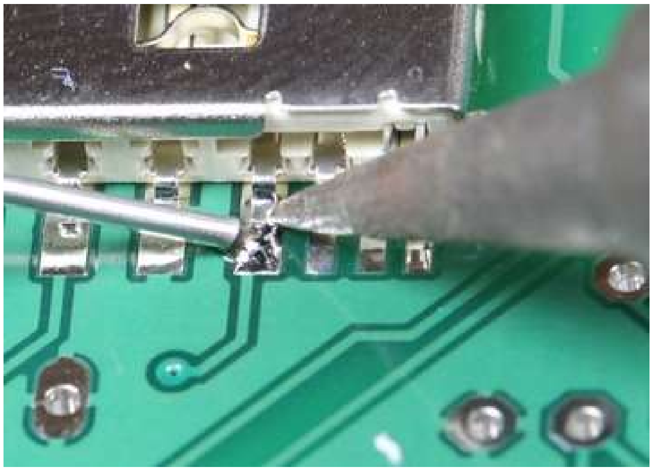

Apply the Solder

When the joint is hot, apply solder to the side opposite the iron. The solder should melt and start to flow into the joint.

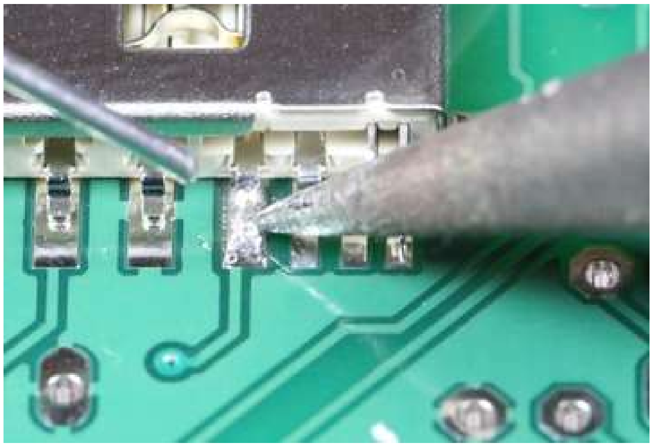

Let it Flow

Apply just enough solder to ensure a good joint, then keep the heat on while the solder wicks up between the pin and the pad to make a good electrical bond.

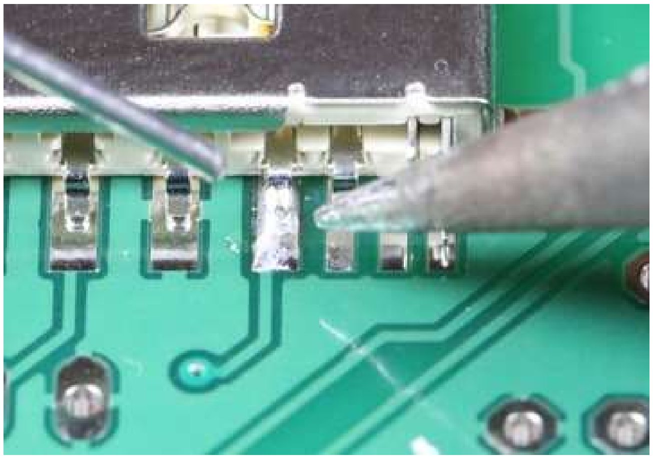

Let it Cool

Remove the iron and allow the joint to cool undisturbed.

Problems?

The last page of this guide illustrates a number of common soldering problems with advice on prevention and repair.

Common Soldering Problems

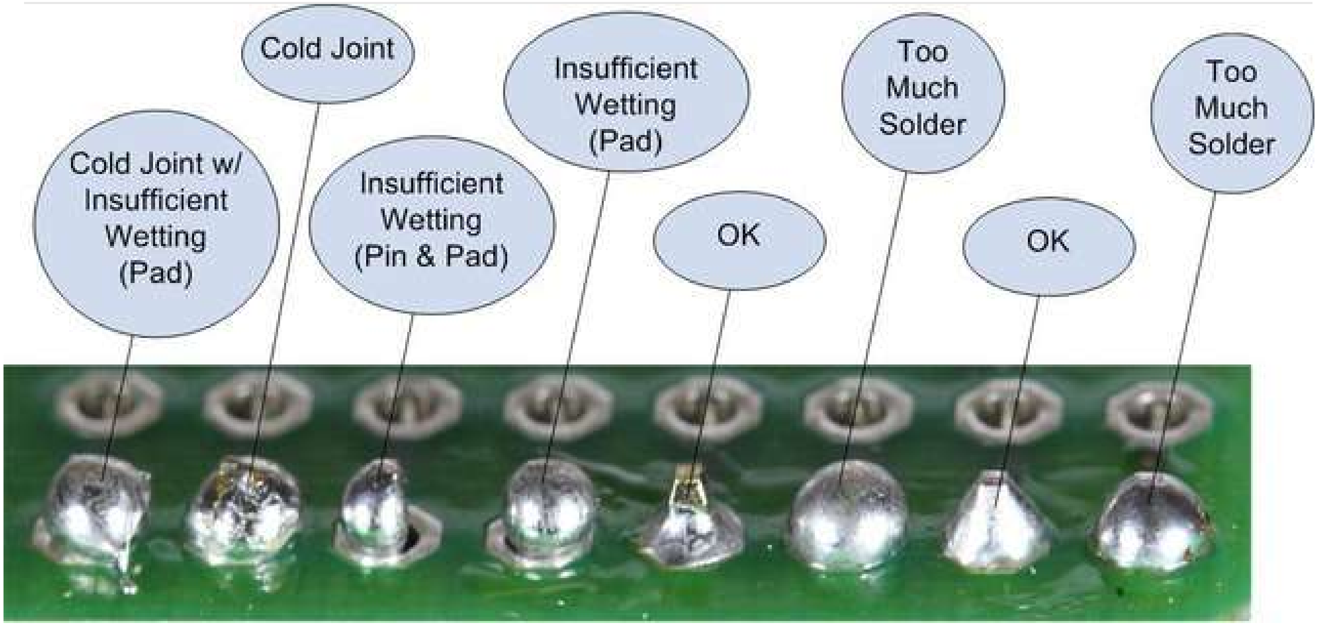
Cold Joint
Insufficient Wetting (Pad)
Too Much Solder
Too Much Solder
Cold Joint w/ Insufficient Wetting (Pad)
Insufficient Wetting (Pin & Pad)
OK
OK

The Ideal Solder Joint

The ideal solder joint for through-hole components should resemble the diagram below.

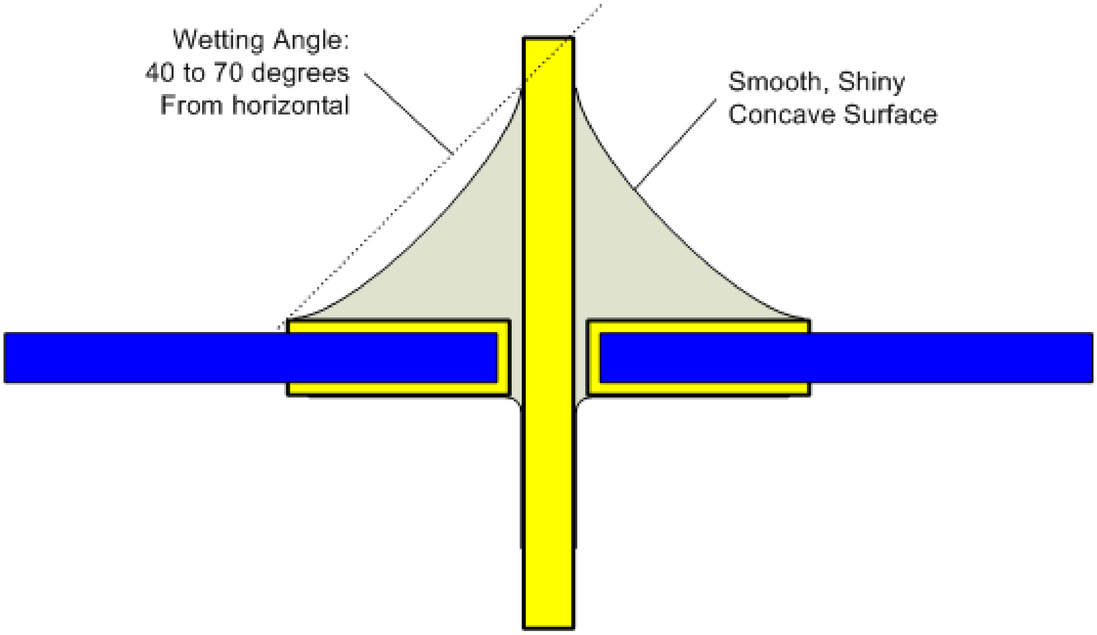
Wetting Angle: 40 to 70 degrees From horizontal
Smooth, Shiny Concave Surface

The photos that follow show some common soldering problems, with suggestions for repair and prevention:

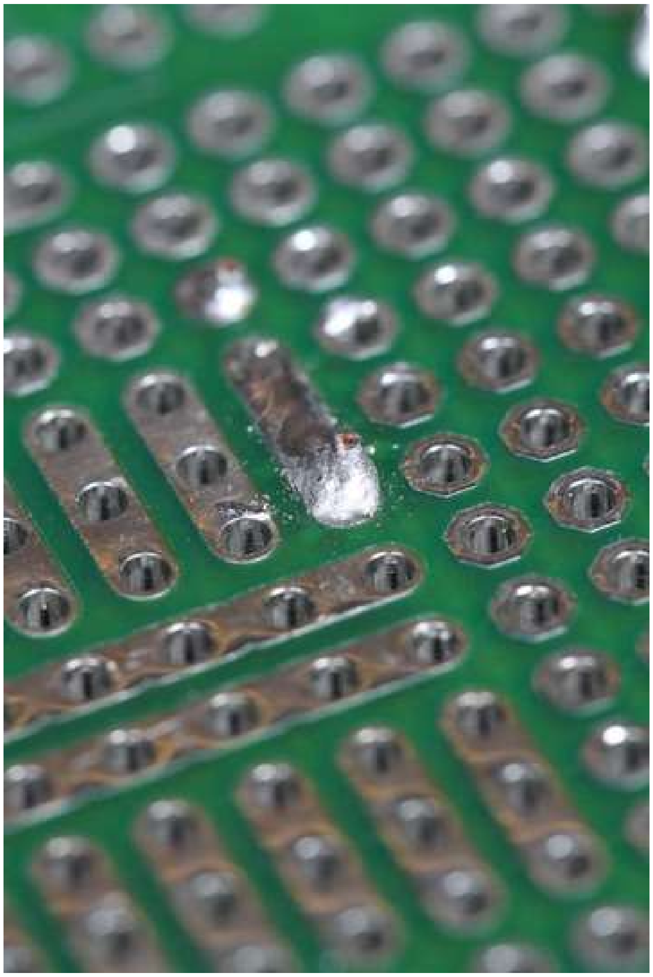

Disturbed Joint

A Disturbed joint is one that has been subjected to movement as the solder was solidifying. The surface of the joint may appear frosted, crystalline or rough.

Often called a 'Cold Joint'. They can look similar to a true cold joint, but the cause is different.

Repair: This joint can be repaired by reheating and allowing it to cool undisturbed.

Prevention: Proper preparation, including immobilizing the joint and stabilizing the work in a vise can prevent disturbed joints.

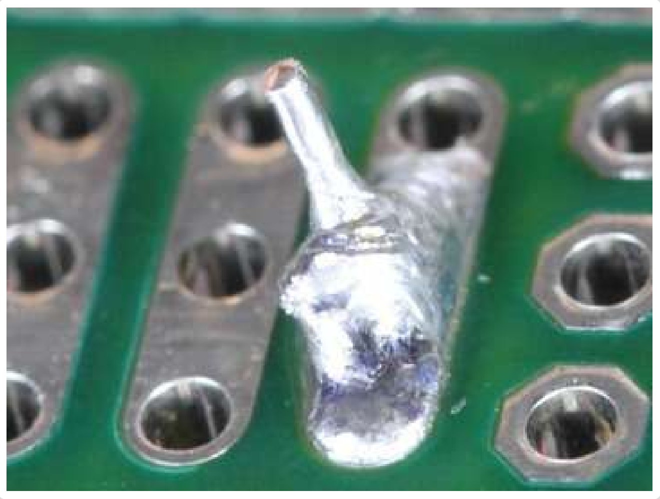

Cold Joint

A 'Cold Joint' is one where the solder did not melt completely. It is often characterized by a rough or lumpy surface. Cold joints are unreliable. The solder bond will be poor and the cracks may develop in the joint over time.

Repair: Cold joints can usually be repaired by simply re-heating the joint with a hot iron until the solder flows. Many cold joints (such as the one pictured) also suffer from too much solder. The excess solder can usually be drawn-off with the tip of the iron.

Prevention: A properly pre-heated soldering iron with sufficient power will help prevent cold joints.

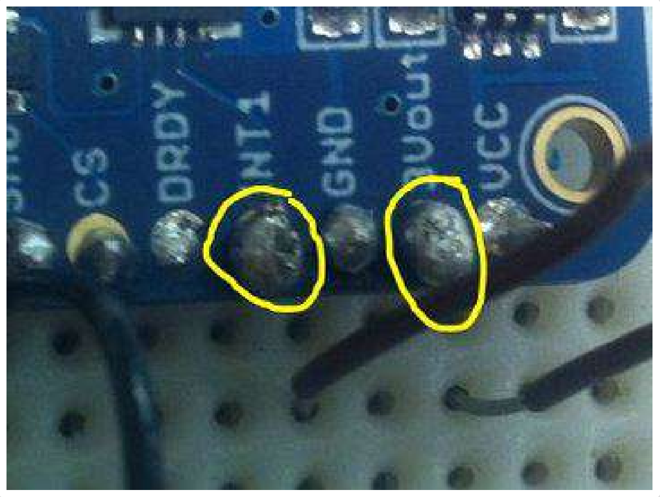

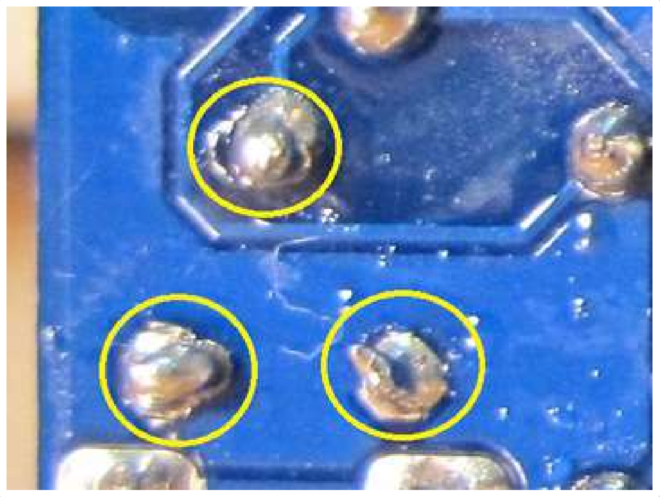

CE1
CE0
SCLK

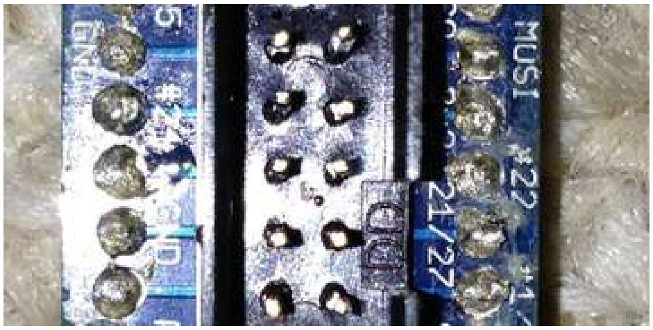
GND
#2
GND
MOSI
#22
21/27

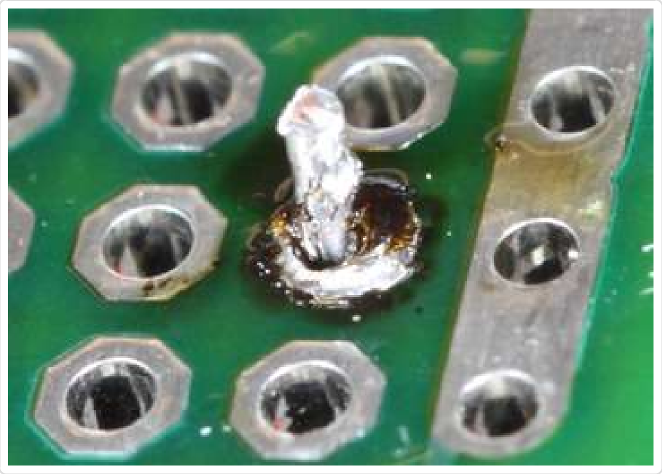

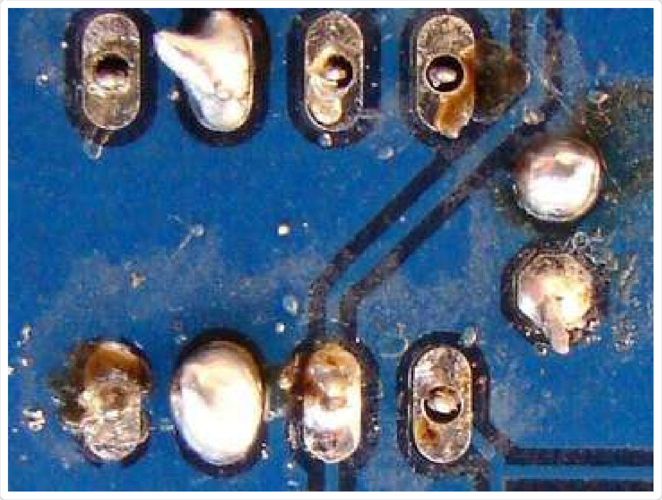

Overheated Joint

At the other extreme, we have the overheated joint. The solder has not yet flowed well and the residue of burnt flux will make fixing this joint difficult.

Repair: An overheated joint can usually be repaired after cleaning. Careful scraping with the tip of a knife, or little isopropyl alcohol & a toothbrush will remove the burnt flux.

Prevention: A clean, hot soldering iron, proper preparation and cleaning of the joint will help prevent overheated joints.

Insufficient Wetting (Pad)

These two joints both show signs of insufficient wetting of the solder pad. The solder has wetted the leads nicely, but it has not formed a good bond with the pad. This can be caused by a dirty circuit board, or by failing to apply heat to the pad as well as the pin.

Repair: This condition can usually be repaired by placing the tip of the hot iron at the base of the joint until the solder flows to cover the pad.

Prevention: Cleaning the board and even heating of both the pad and the pin will prevent this problem.

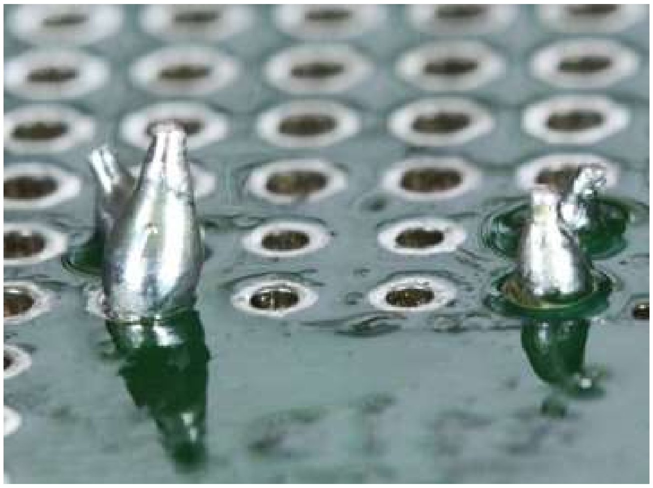

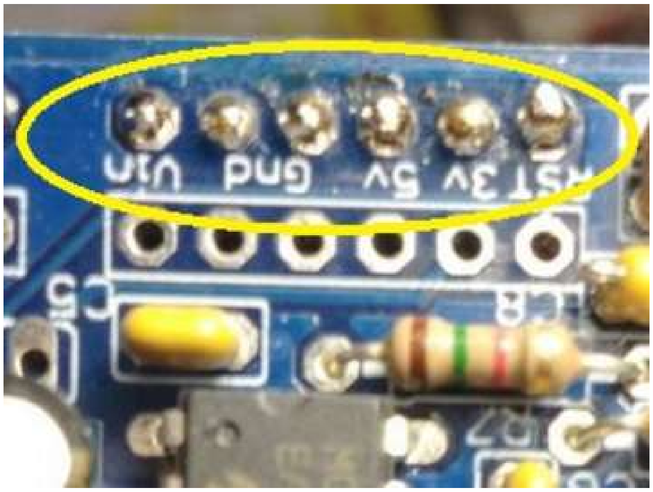

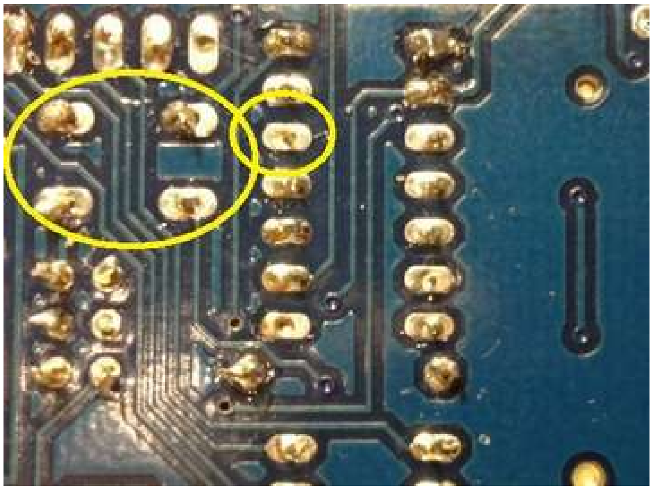

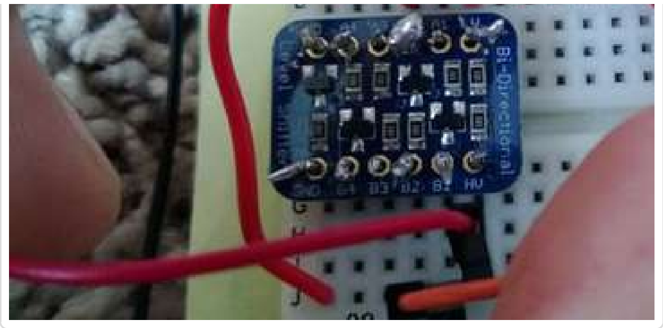
Bi-Directional
GND
B4
B3
B2
B1
HV

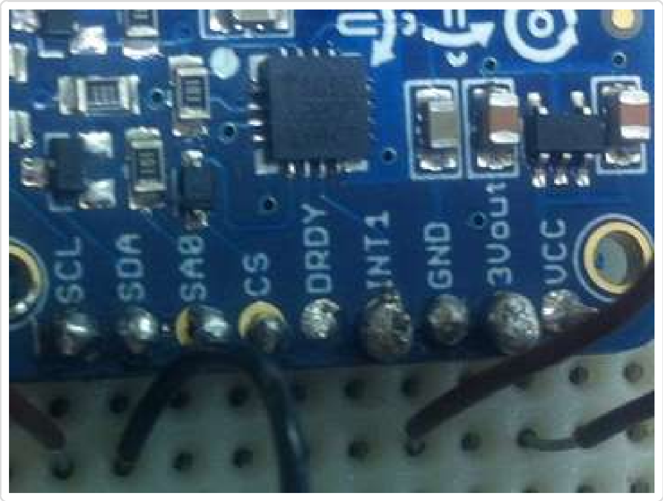
SCL
SDA
SA0
CS
DRDY
INT1
GND
3Vout
VCC

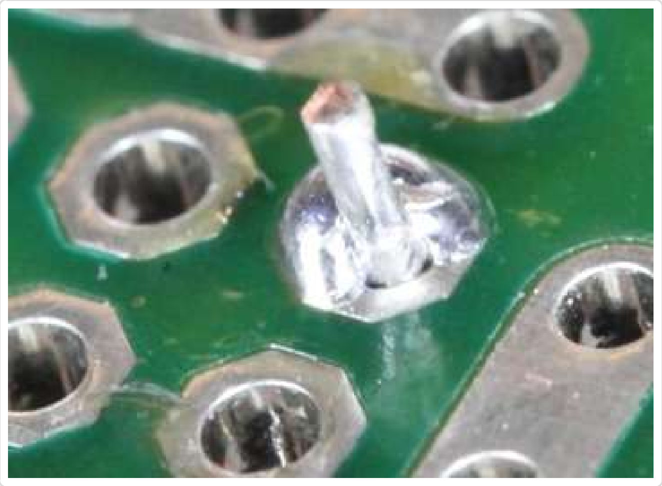

Insufficient Wetting (Pin)

This solder in this joint has not wetted the pin at all and has only partially wetted the pad. In this case, heat was not applied to the pin and the solder was not given adequate time to flow.

Repair: This joint can be repaired by re-heating and applying more solder. Be sure that the tip of the hot iron is touching both the pin and the pad.

Prevention: Even heating of both the pin and the pad will prevent this problem.

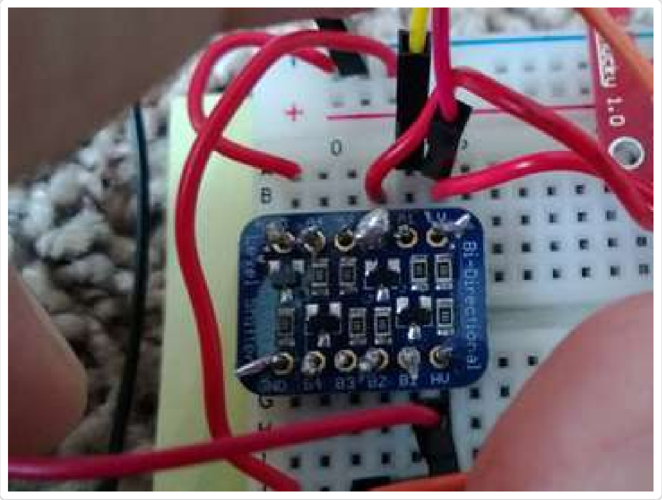

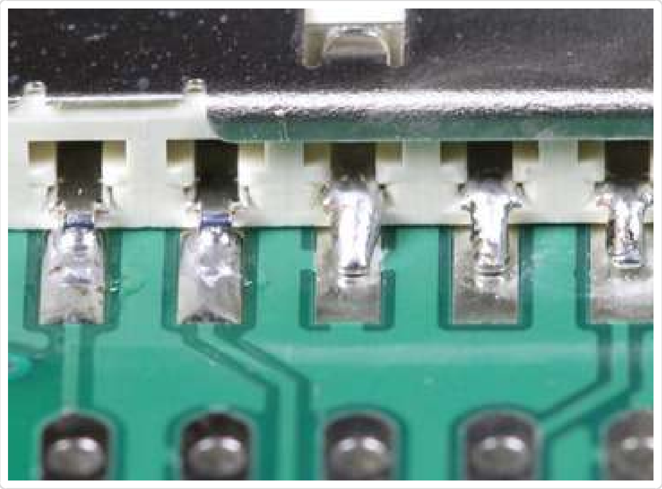

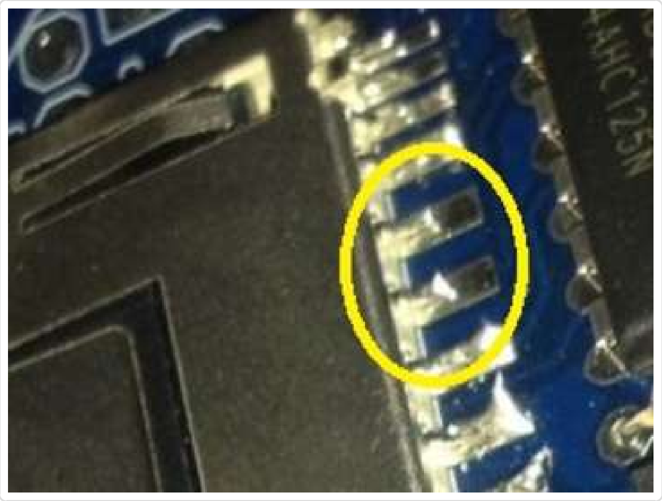

Insufficient Wetting (Surface Mount)

Here we have three pins of a surface mount component where the solder has not flowed onto the solder pad. This is caused by heating the pin instead of the pad.

Repair: This is easily repaired by heating the solder pad with the tip of the iron, then applying solder until it flows and melts together with the solder already on the pin.

Prevention: Heat the pad first.

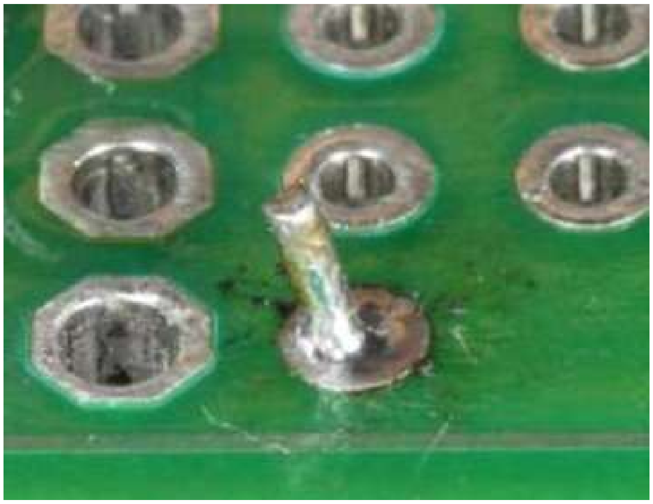

Solder Starved

A solder starved joint simply does not have enough solder. It may make good electrical contact, but it is hard to verify by inspection. In any case, it is not a strong joint and may develop stress cracks and fail over time.

Repair: Re-heat the joint and add more solder to make a good strong joint.

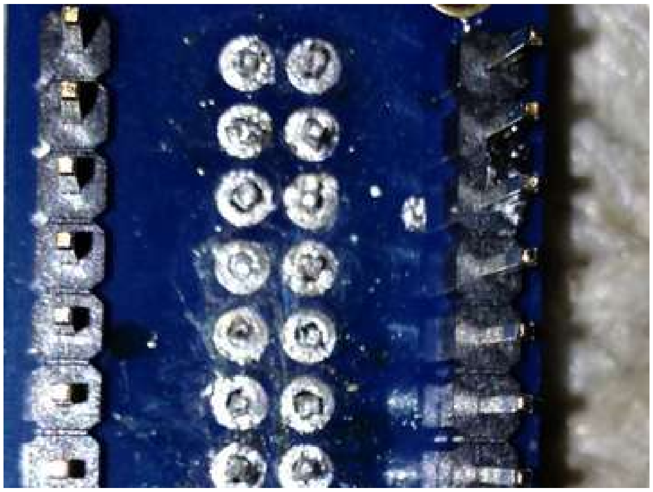

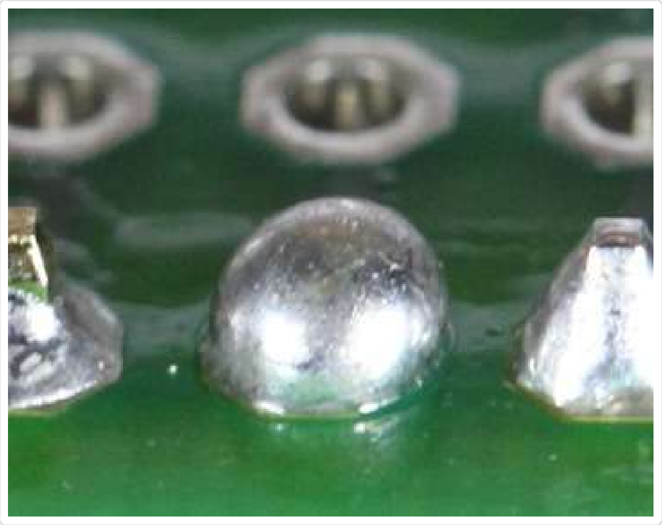

Too Much Solder

This might be a perfectly good joint, but we can't tell for sure. It is entirely possible that this blob of solder wets neither the pin nor the pad and is not a reliable electrical connection. The best evidence of proper wetting (and good electrical contact) is a nice concave surface as on the joint on the far left.

Repair: It is usually possible to draw off some of the excess solder with the tip of a hot iron. In extreme cases, a solder-sucker or some solder wick can be helpful as well.

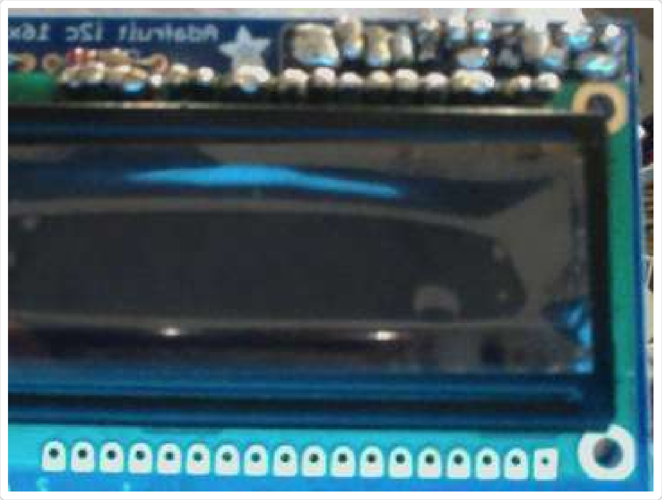

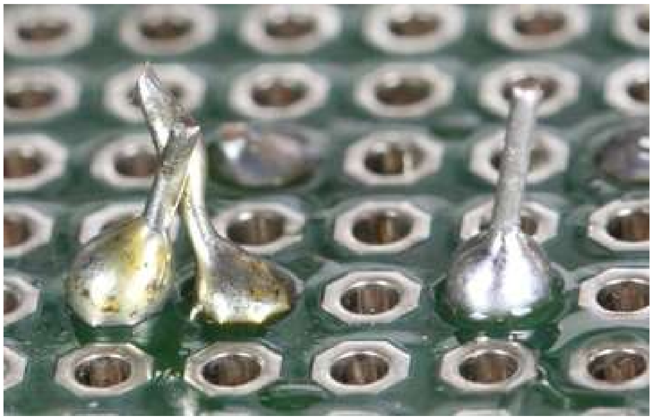

Untrimmed Leads

Leads that are too long are potential short circuits. The two joints on the left are an obvious danger of touching. But the one on the right is long enough to be dangerous as well. It would not take much force to bend that lead over to touch an adjacent trace.

Repair: Trim all leads just at the top of the solder joint.

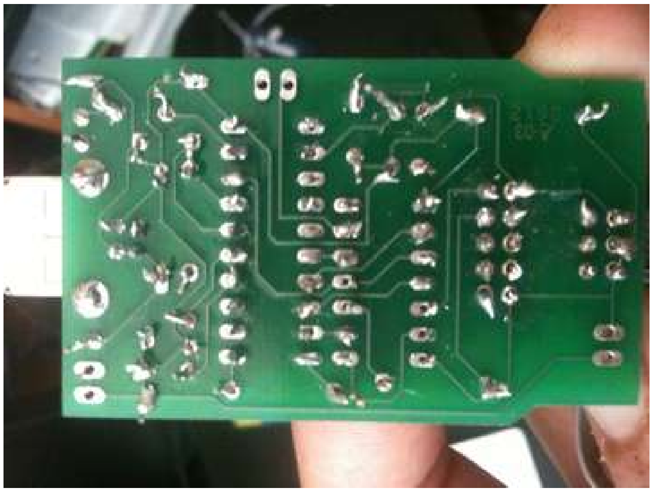

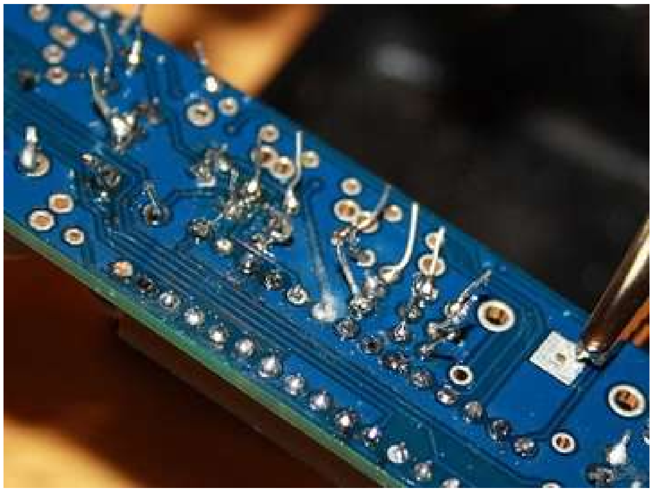

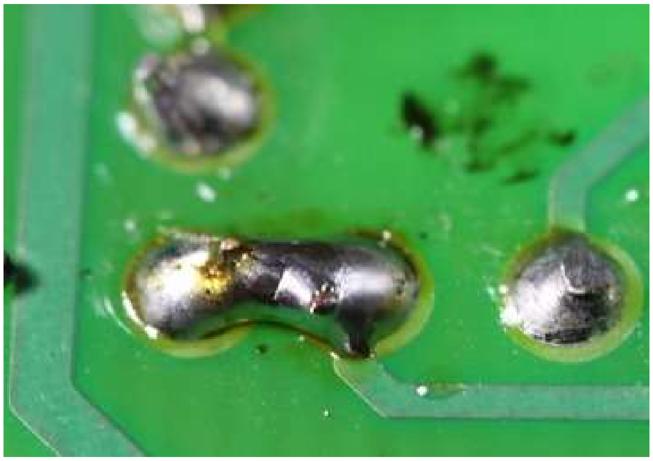

Solder Bridge

The left two solder joints have melted together, forming an unintended connection between the two.

Repair: Sometimes the excess solder can be drawn off by dragging the tip of a hot iron between the two solder joints. If there is too much solder, a solder sucker or solder wick can help get rid of the excess.

Prevention: Solder bridges most often happen between joints with too much solder to begin with. Use only enough solder to make a good joint.

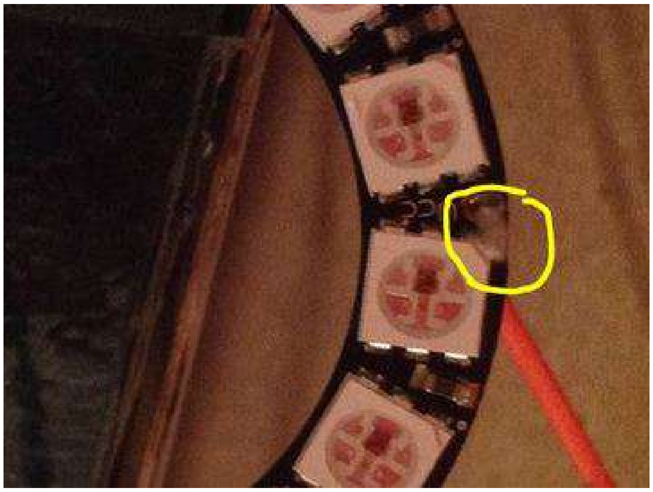

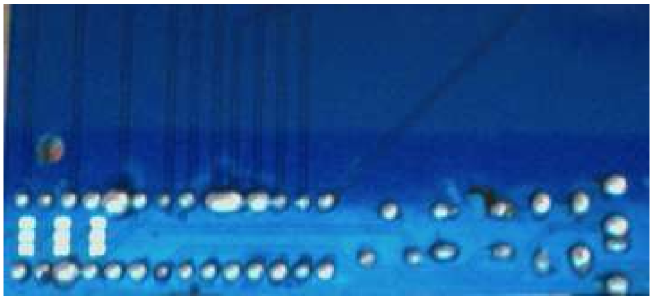

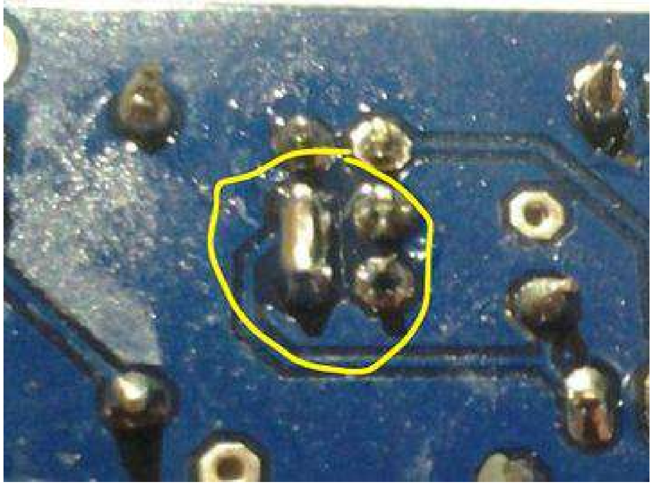

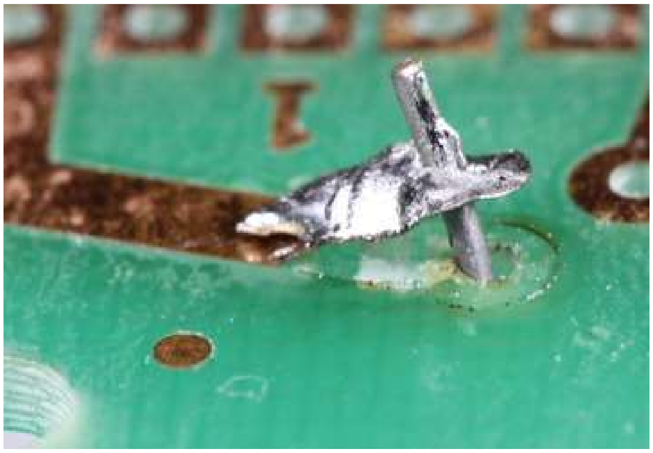

Lifted Pad

This photo shows a solder pad that has become detached from the surface of the circuit board. This most often occurs when trying to de-solder components from the board. But it can result simply from overworking the joint to the point where the adhesive bond between copper and the board is destroyed.

Lifted pads are especially common on boards with thin copper layers and/or no through-plating on the holes.

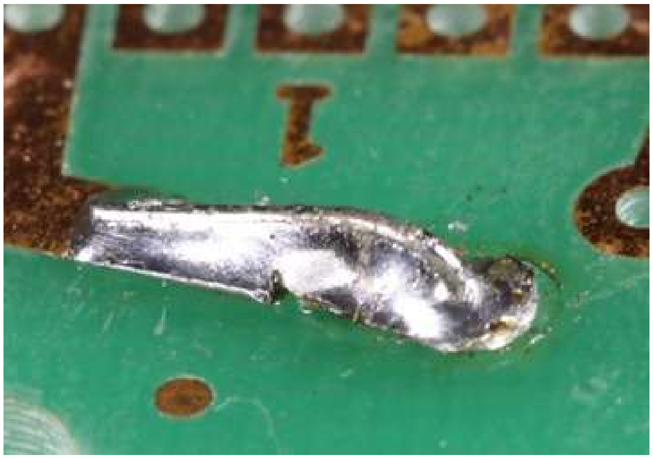

Repairing a Lifted Pad

It may not be pretty, but a lifted pad can usually be repaired. The simplest repair is to fold the lead over to a still-attached copper trace and solder it as shown to the left. If your board has a solder-mask, you will need to carefully scrape off enough to expose the bare copper.

Other alternatives are to follow the trace to the next via and run a jumper to there. Or, in the worst case, follow the trace to the nearest component and solder your jumper to the leg of that. Not exactly pretty, but functional.

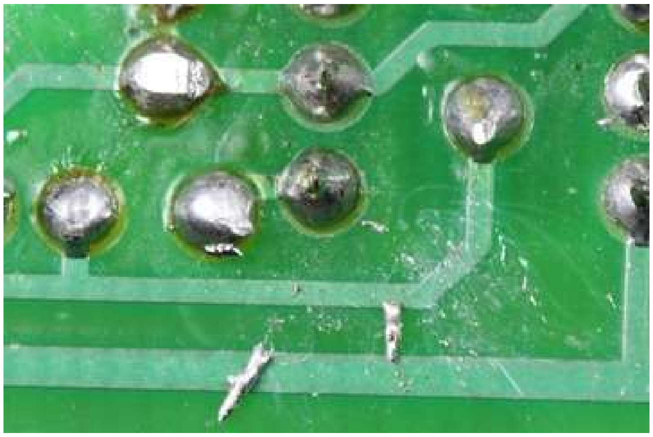

Stray Solder Spatters

These bits of solder are held to the board only by sticky flux residue. If they work loose, they can easily cause a short circuit on the board.

Repair: These are easy to remove with the tip of a knife or tweezers.

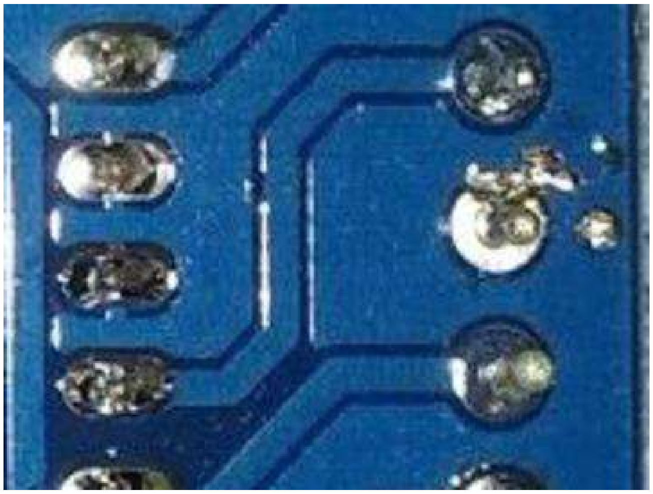

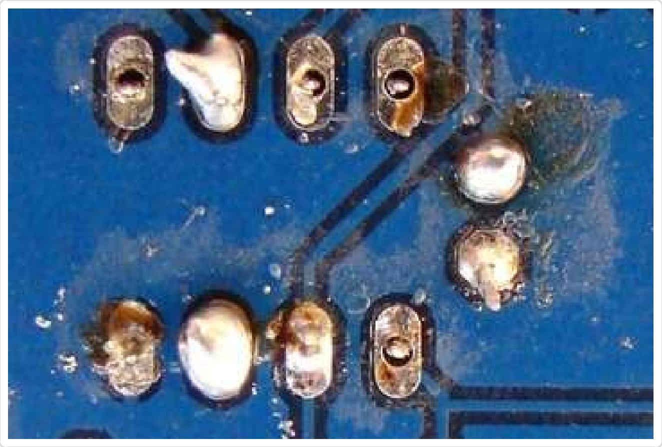

All of the Above!

Don't panic. Take your time. Most joints can be repaired with patience. If the solder refuses to flow the way you want it to:

1. Stop and let the joint cool.
2. Clean and tin your iron.
3. Clean off any burnt flux from the joint.
4. Let the iron come back up to temperature.
5. Then reheat the joint and try again.

AN114

HAND SOLDERING TUTORIAL FOR FINE PITCH QFP DEVICES

Scope

This document is intended to help designers create their initial prototype systems using Silicon Lab's TQFP and LQFP devices where surface mount assembly equipment is not readily available. This application note assumes that the reader has at least basic hand soldering skills for through-hole soldering. The example presented will be the removal, cleanup and replacement of a TQFP with 48 leads and 0.5 mm lead pitch.

Safety

Work should be done in a well-ventilated area. Prolonged exposure to solder fumes and solvents can be hazardous. There should be no presence of sparks or flames when solvents are in use.

Materials

The right materials are key to a good solder job. The list below are the recommendations from Silicon Labs. Other materials may work, so the user should feel free to substitute and experiment. The use of organic solder is highly recommended.

Required

1. Wire wrap wire (30 gauge) *
2. Wire strippers for wire wrap wire *
3. Soldering station - variable temperature, ESD-safe. Should support temperatures 800°F (425°C). This example uses a Weller model EC1201A. The soldering wand should have a fine tip no more than 1 mm wide.
4. Solder - 10/18 organic core; 0.2" (0.5 mm) diameter
5. Solder flux - liquid type in dispenser
6. Solder wick - size C 0.075" (1.9 mm)
7. Magnifier - 4X minimum. An inexpensive headset OptiVISOR by Donegan Optical Co. is used for this example.
8. ESD mat or tabletop and ESD wrist strap - both grounded
9. Tweezers with pointed (not flat) tips
10. Isopropyl Alcohol
11. Small stiff bristle brush for cleaning (nylon or other non-metallic material). Cut off hair to approximately 0.25" (6 mm)

* Required for device removal only.

Optional

1. Board vise to hold printed circuit board
2. Dental pick (90 degree bend)
3. Compressed dry air or nitrogen to dry boards

Copyright © 2003 by Silicon Laboratories

4. Optical inspection stereo microscope 30-40X

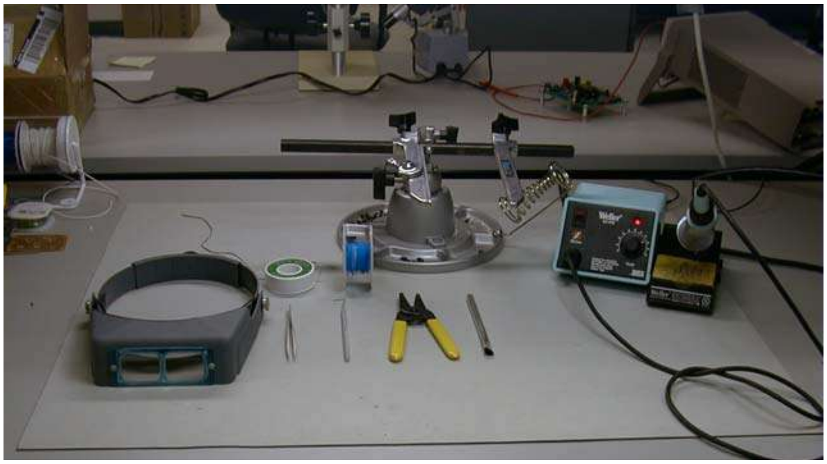

Figure 1. Some of the materials to get started...

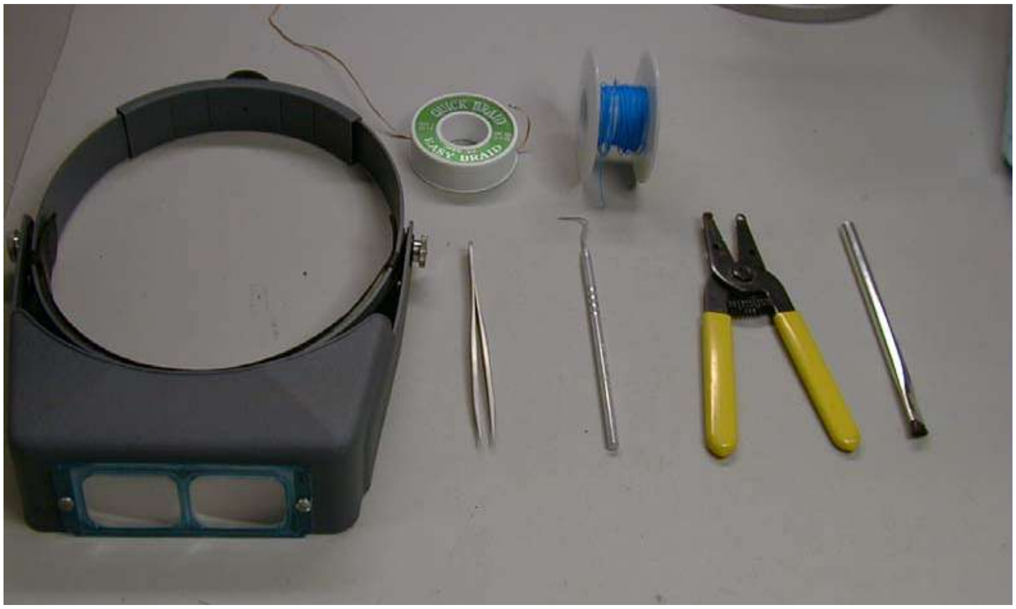

Figure 2. clock-wise from left: 4X magnifying headset, solder wick, wire wrap wire, stiff cleaning brush, wire strippers, and pointed tweezers

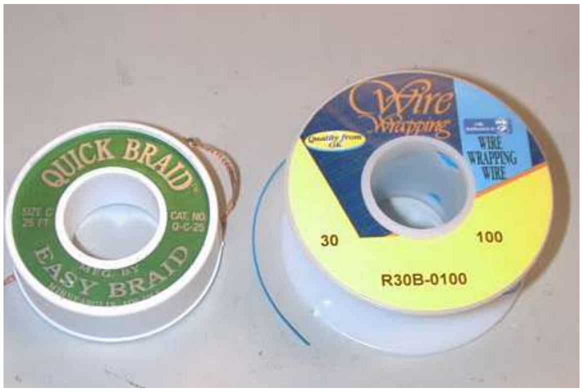

Figure 3. Solder wick and wire wrap wire

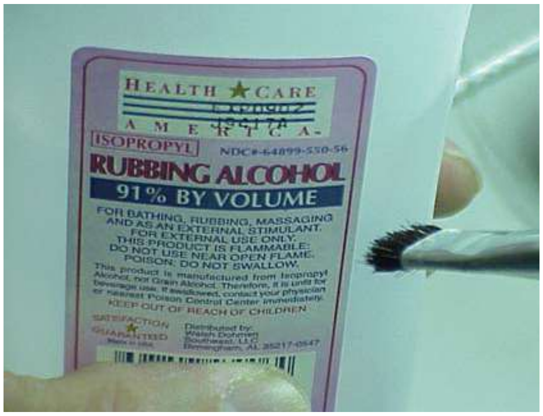

Figure 3. Isopropyl Alcohol

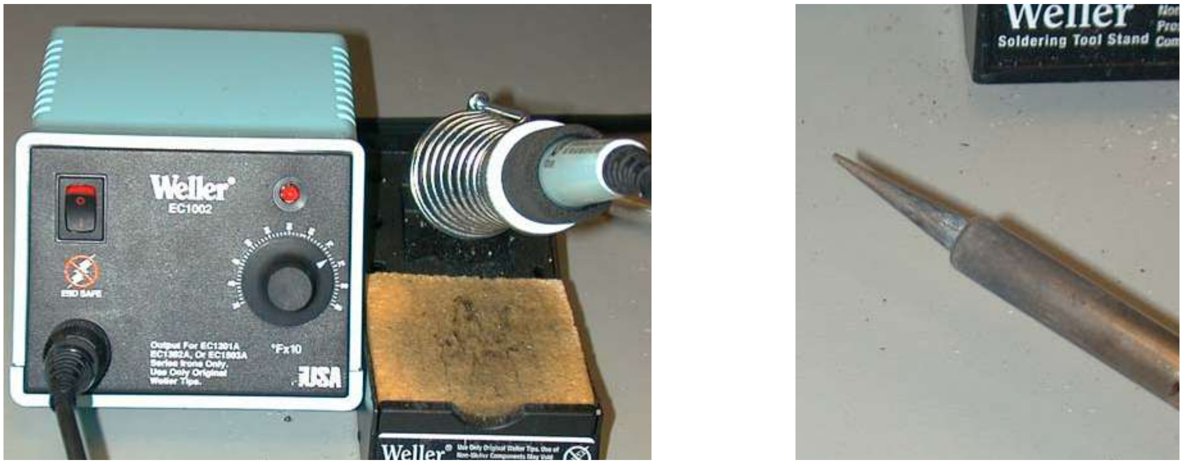

Figure 4. ESD safe soldering station with fine tip wand. This one is Weller model EC1201A

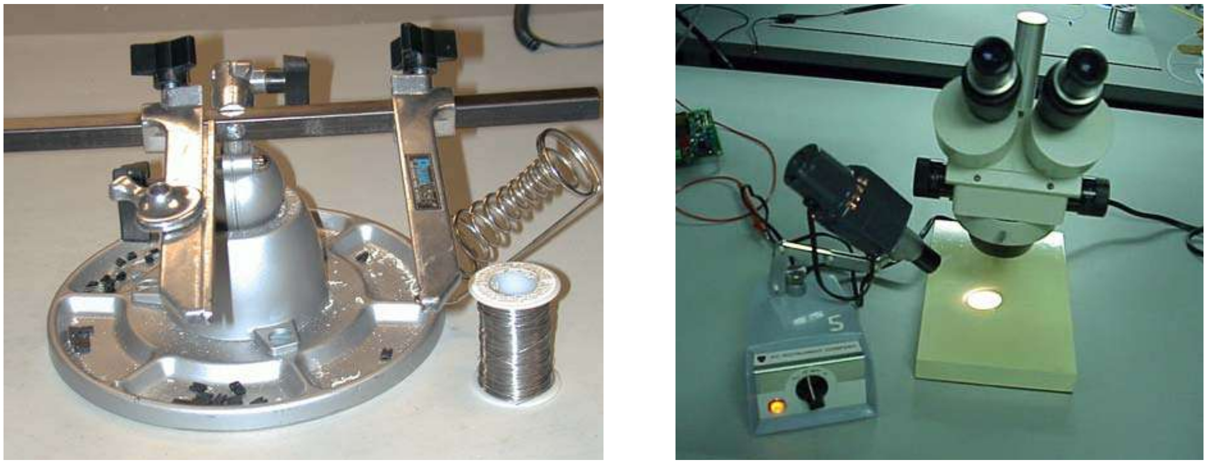

Figure 5. Optional equipment includes a PCB vise and an inspection microscope 7-40X

Procedure

The following procedures cover the replacement of a TQFP 48 pin device with 0.5 mm lead pitch. The lead shape is the standard gullwing associated with JEDEC standard QFPs. This procedure section is divided into three parts:

A. Part Removal

B. Board Cleanup

C. Soldering a new device.

If you are soldering parts to a new printed circuit board, skip part A and refer to the new board cleanup section in part B.

A. Part Removal

Preparation:

- Board with IC to be removed is mounted in a holder or vise. A PCB holder/vise is optional but it is required that the PCB is held steady for the part removal.
- The soldering station is warmed up to 800°F (422°C) and the solder tip is clean.
- ESD precautions have been taken.

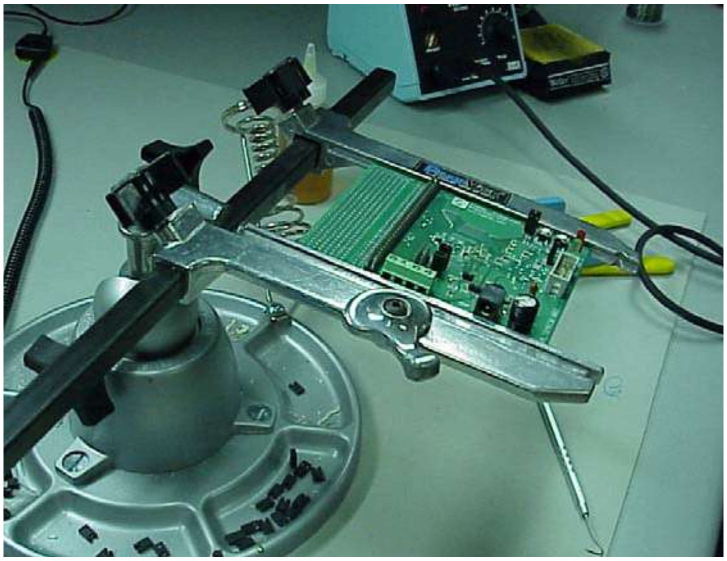

Figure 6. Ready to get started

Begin by wetting all the leads with flux to enhance the initial solder wicking cleanup. Wick up solder as much as possible from the QFP leads. Be careful not to scorch the PCB board with prolonged solder heat.

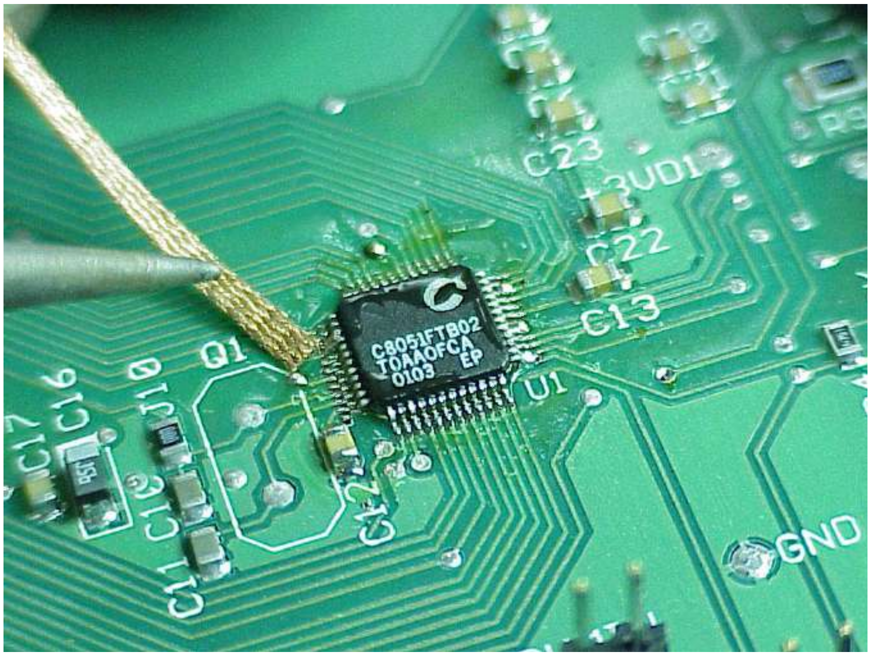
C23
3UD1
C22
C13
C8051FT802
T0AAOFCA
0103 EP
U1
Q1
C16
C17
J10
C12
C11
C1
GND
R9

Figure 7. Apply flux; wick excess solder from pins

Next, strip off approximately 3 inches of insulation from a piece of the 30 gauge wire wrap wire. Cut the wire at a comfortable 12 inch length or so.

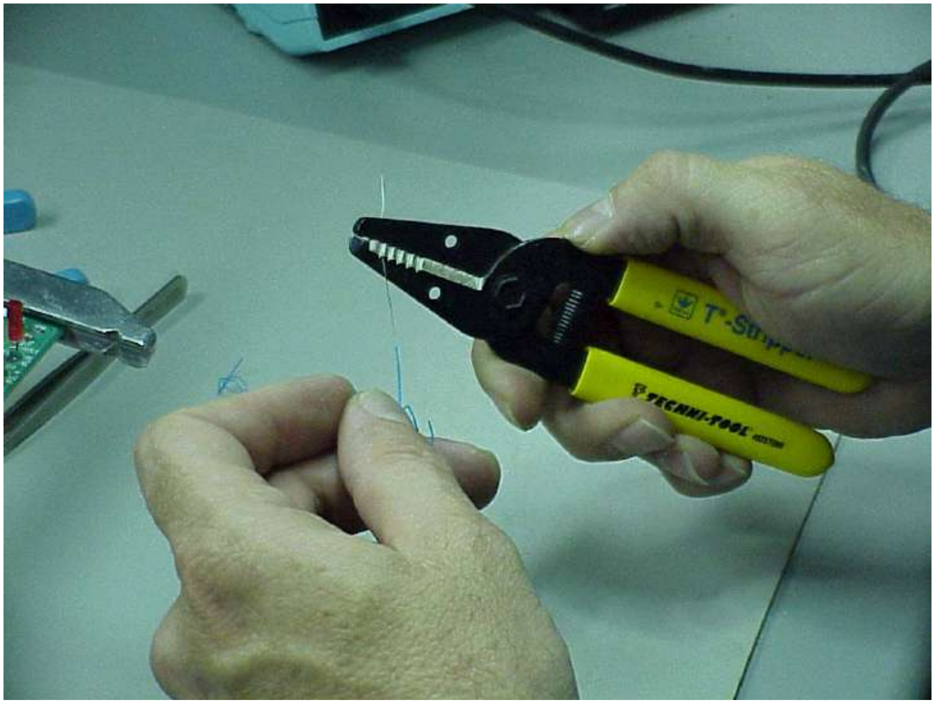
T'-Stripper
TECHNI-TOOL

Figure 8. Wire stripping

Feed the wire behind and under the leads on one side of the IC as shown in Figure 9.

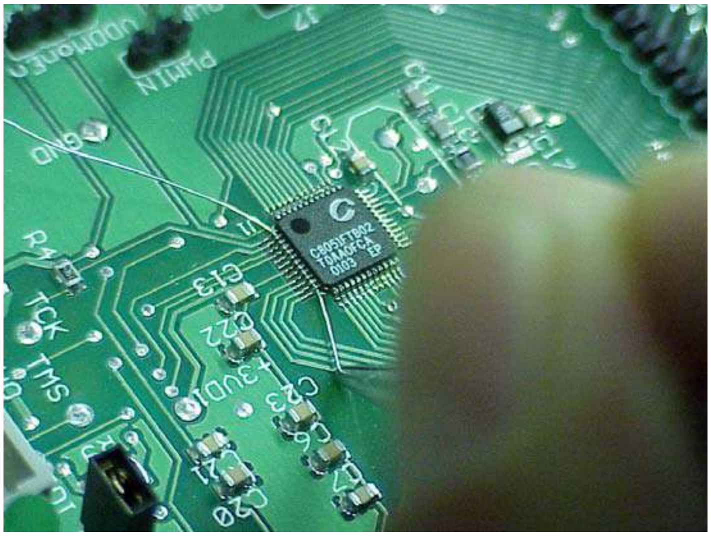

Figure 9. Feeding wire under QFP leads

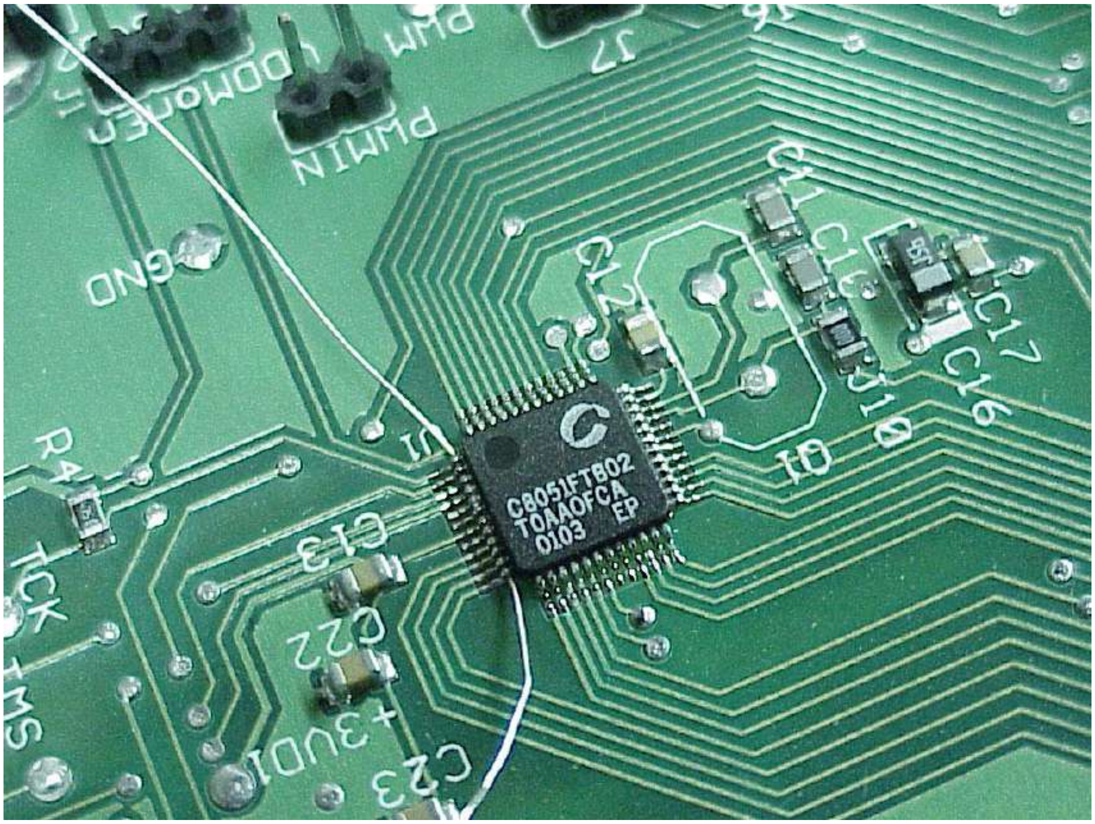

Figure 9. Wire with one side anchored to nearby component

Solder tack (anchor) one end of the 3 inch wire to a nearby via or component on the PCB. The anchor point should be in a location similar to that should in Figure 10.

Dispense a small amount of liquid flux across the leads.

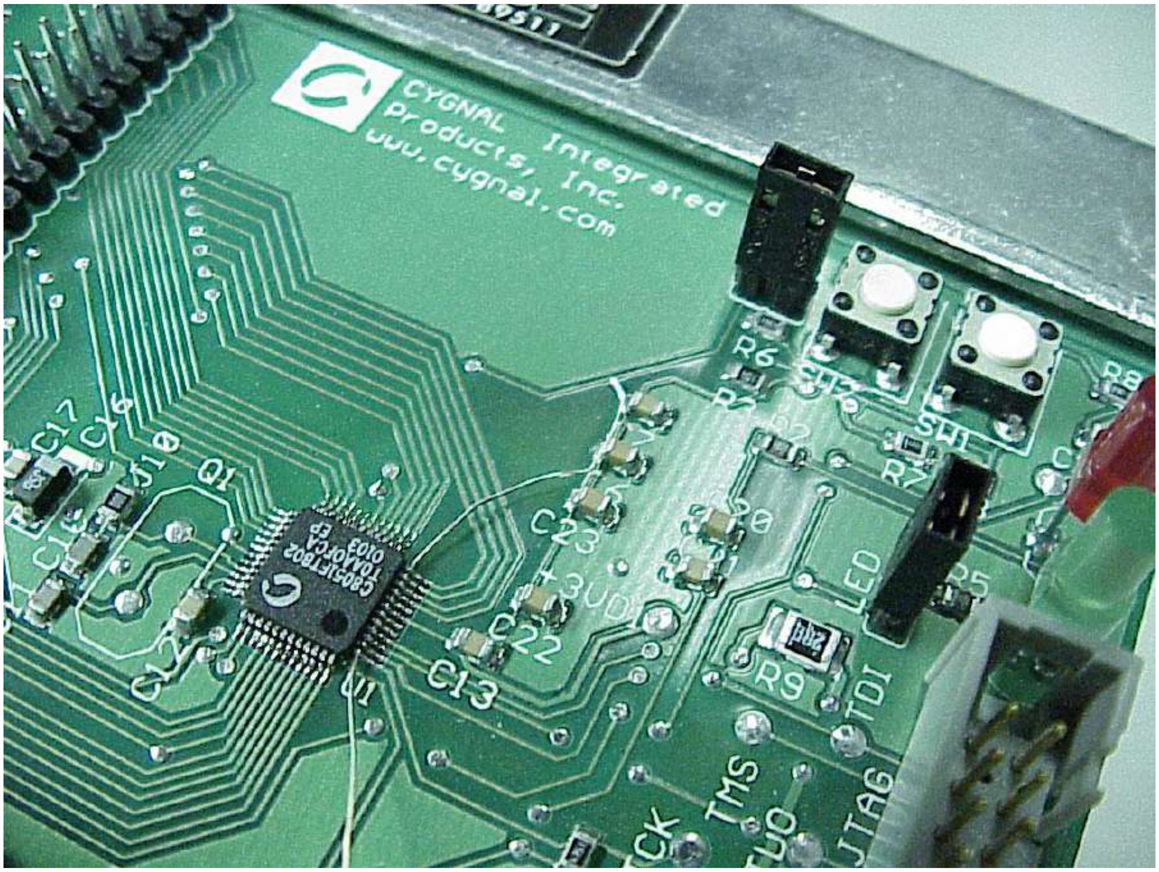

Figure 10. Bus wire is anchored on C6

Hold the loose (non-anchored) end of the wire with tweezers in close proximity to the device as shown in Figure 11.

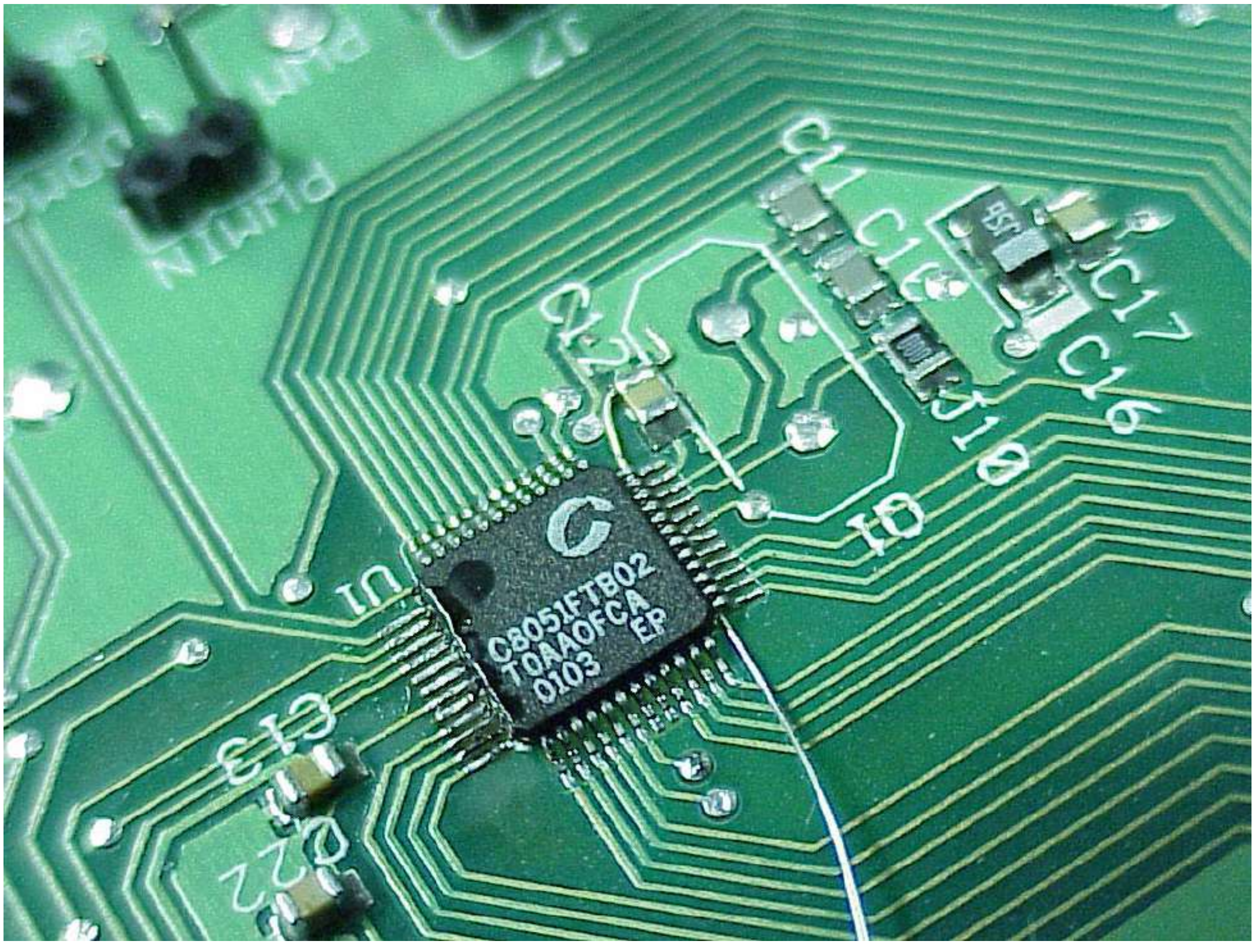

Figure 11. Side 2 anchored and ready for solder heat

You will now need to simultaneously apply solder heat and pull the wire away from the QFP, pulling at a slight upward angle from the board surface. Apply solder heat beginning at the lead closest to your tweezers. As the solder melts, gently pull the wire away from the QFP while continuing to move the solder heat from pin to pin to the right. You should not pull very hard. Pull as the solder melts. Do not leave the solder heat on any lead for more than necessary. The first lead will take the longest to heat, and after the wire gets hot, solder on the other leads will melt quickly. Excessive heat will damage the IC device and the PCB pad. The removal of 12 pins from a 48TQFP should take about 5 seconds total. Signs of excessive heat are:

- Melted plastic on the IC device
- PCB pads that lift off
- Brown scorch marks on the PCB

With one side of the QFP completed, repeat the same procedure on the other three sides of the QFP. Cut off the dirty part of the wire wrap wire or use a new piece for each side. Dispense flux again for each side.

Note that in the following pictures, the old IC device is not being saved. There is slightly more heat being applied here than necessary to speed up the process. This result is some melted plastic and missing gull-wing leads. These are visible in the pictures that follow. If you are trying to save the IC being removed, then you must be very careful to apply as little heat as possible during the removal process such that the QFP leads remain intact in the plastic QFP body. This will require some experimentation with solder heat settings and timing.

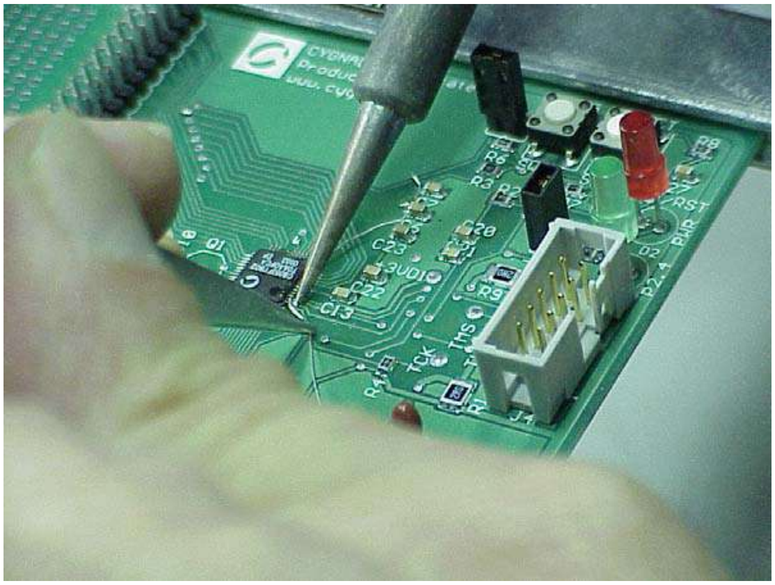

Figure 12. Hold tweezers close to device

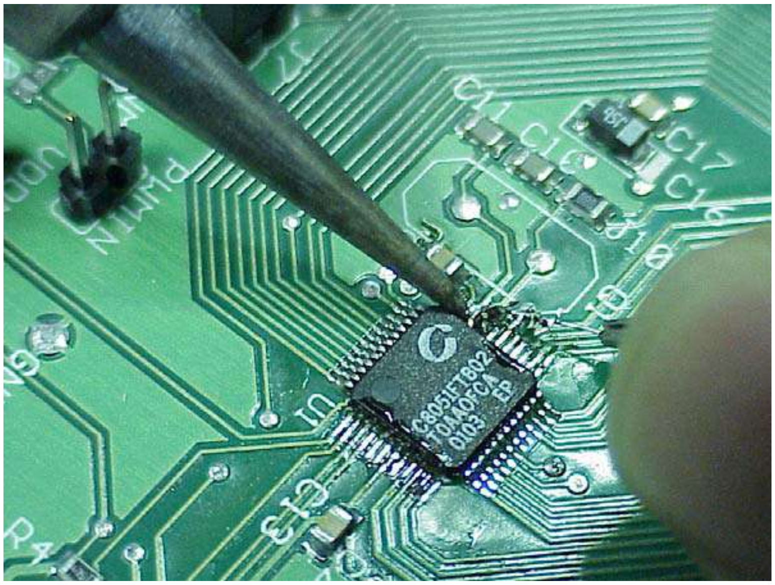

Figure 14. Side 2 almost complete

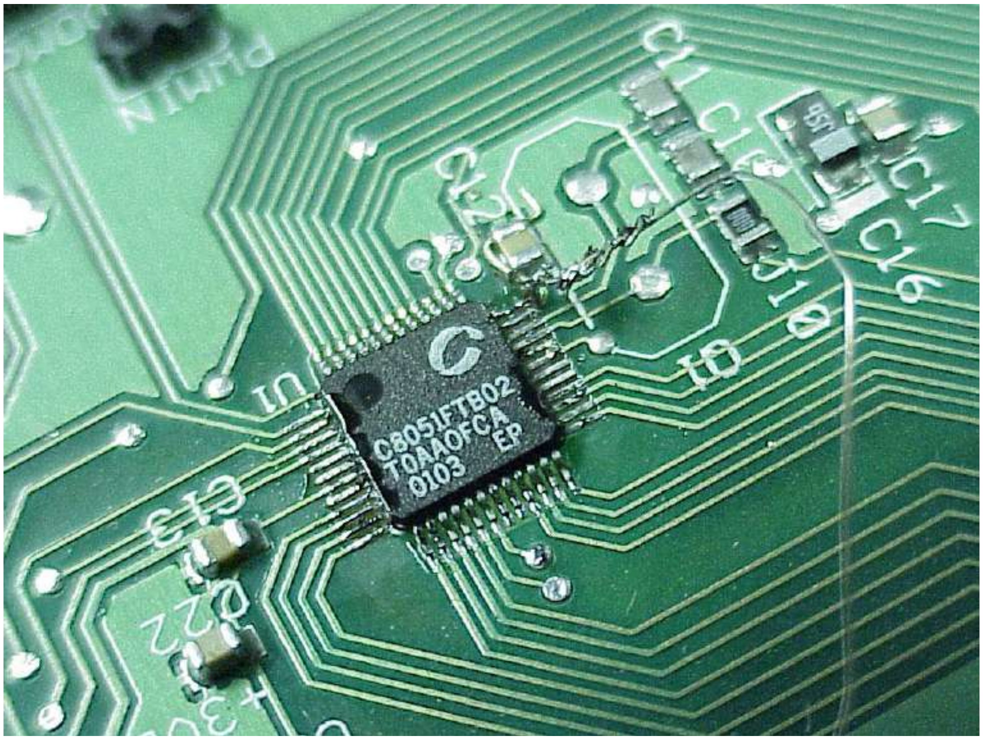

Figure 15. Side 2 is done

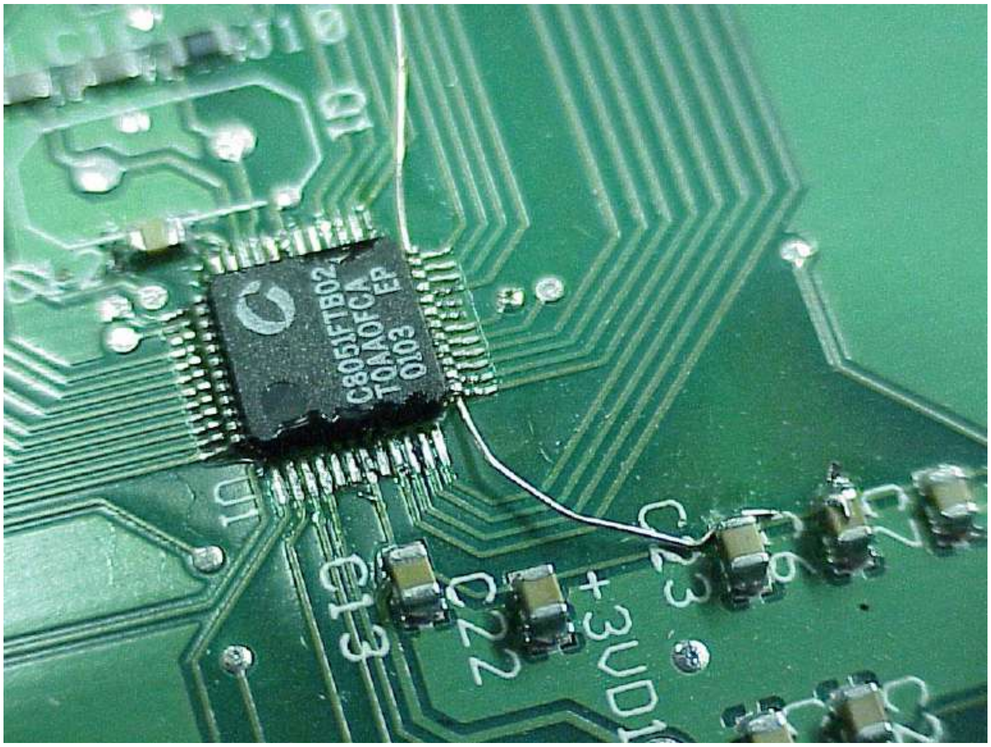

Figure 16. Side 3 anchored and ready

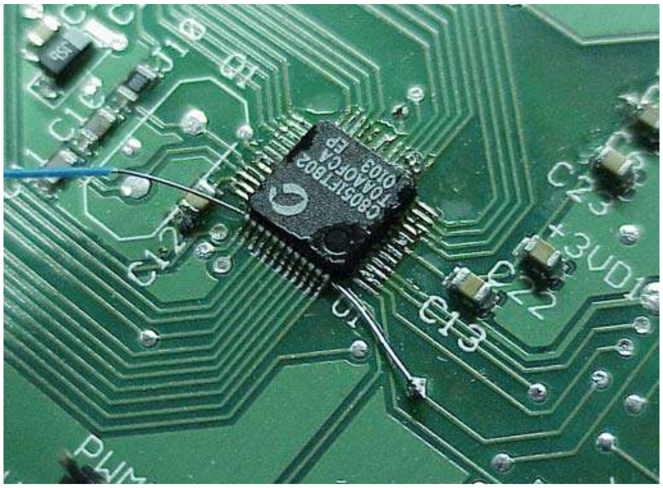

Figure 17. Side 4 anchored to a via

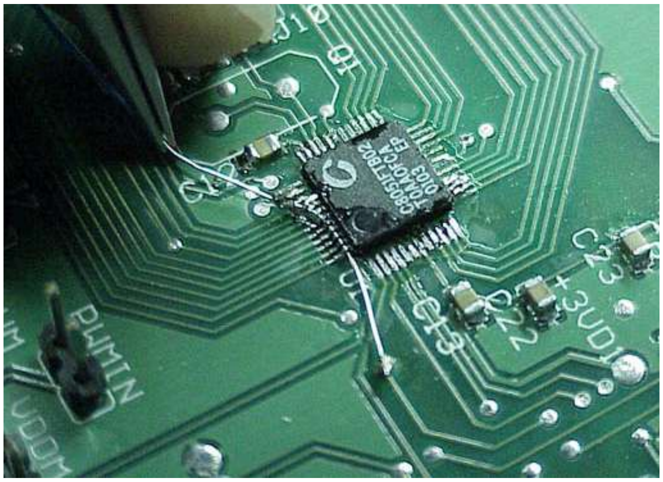

Figure 18. Side 4 removal started

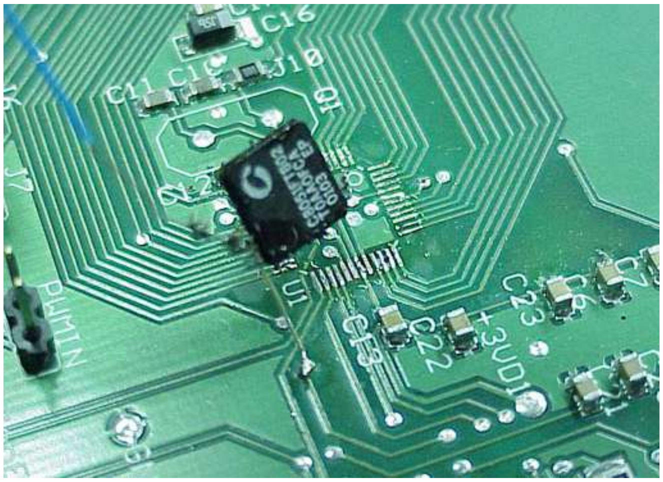

Figure 19. QFP removal a second before completion

B. Board Cleanup

New PCBs

For mounting a device to a new PCB, the amount of cleanup should be minimal. On a new PCB, there should be no solder on the pads. Brushing the pads with isopropyl alcohol (Figure 33) and drying the board should be enough preparation to begin the mounting procedure.

Reworked PCBs

The following section is the cleanup sequence that follows the QFP removal in the previous section. After removing the device, the solder pads will need cleanup. The idea is to clean the pads so that they are flat and free or solder and flux. Solder wick the pads until they are flat and dull. A clean pad appearance should be a dull silver color.

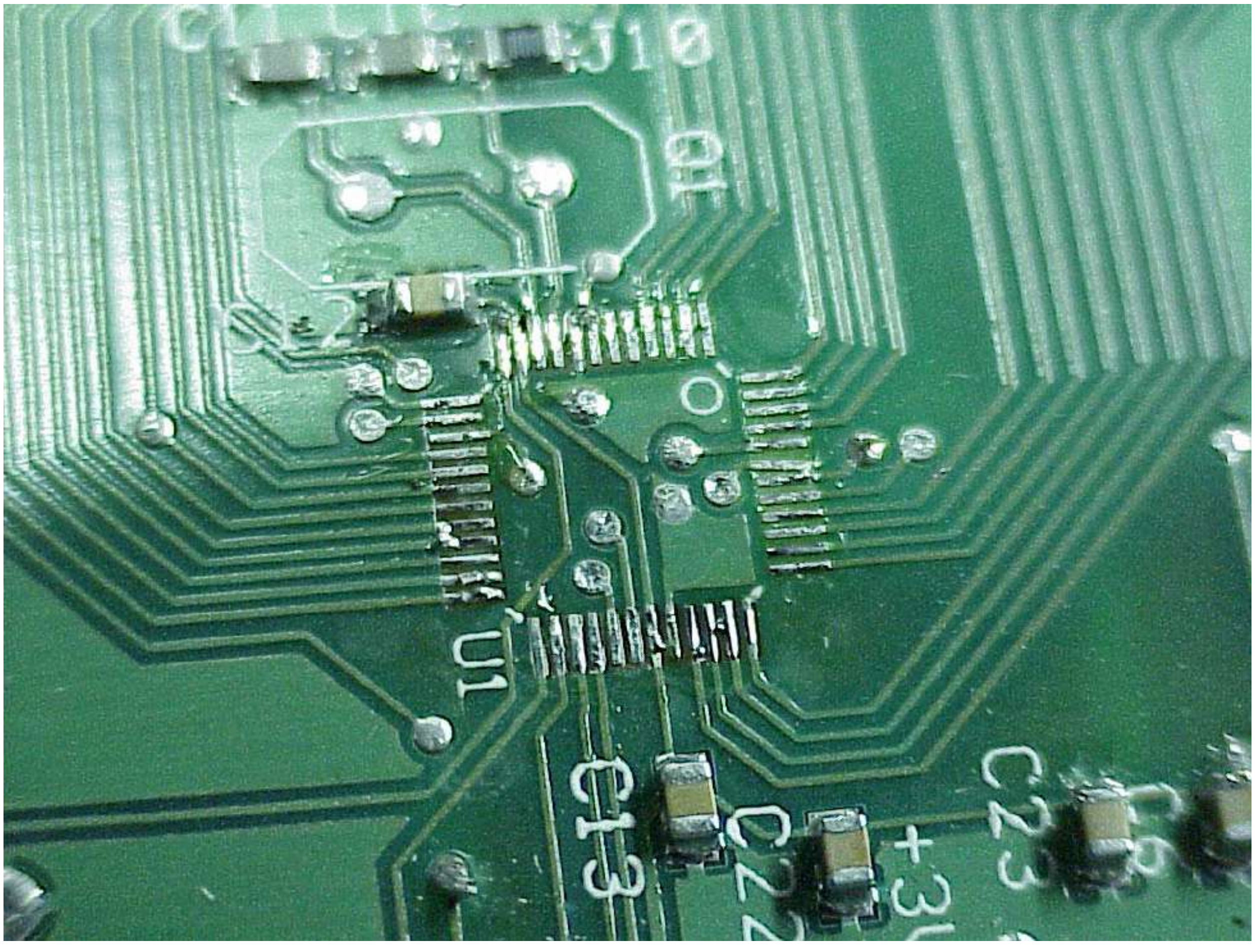

Figure 20. Pads after QFP removal procedure

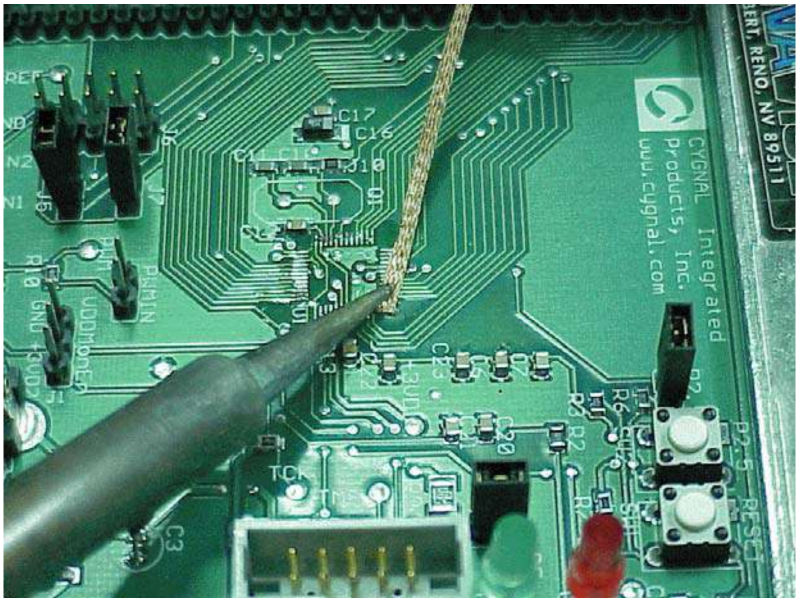

Figure 21. Solder wick to remove solder on the pads

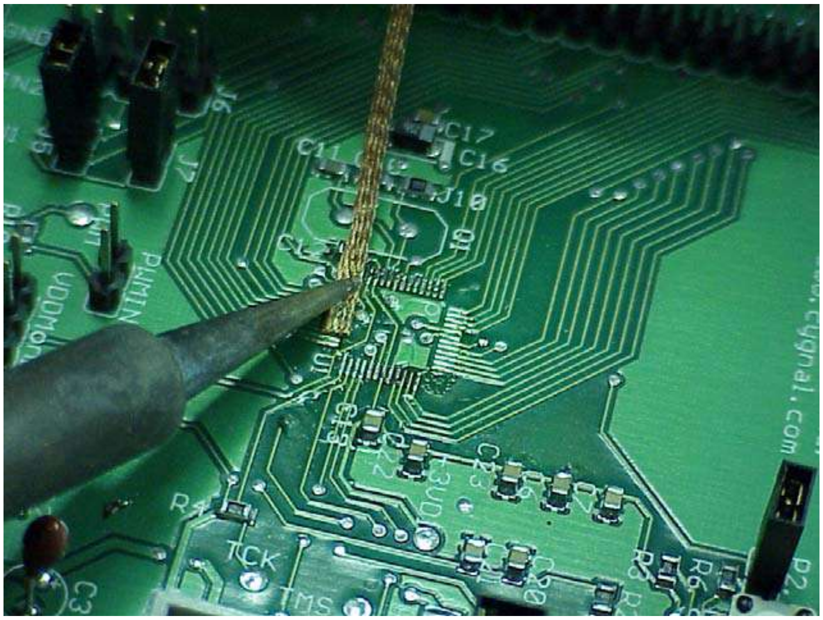

Figure 22. Repeat for all pads

If any pads become loose from the PCB, use a dental pick or other pointed object to re-align the pad (Figure 23 and Figure 24).

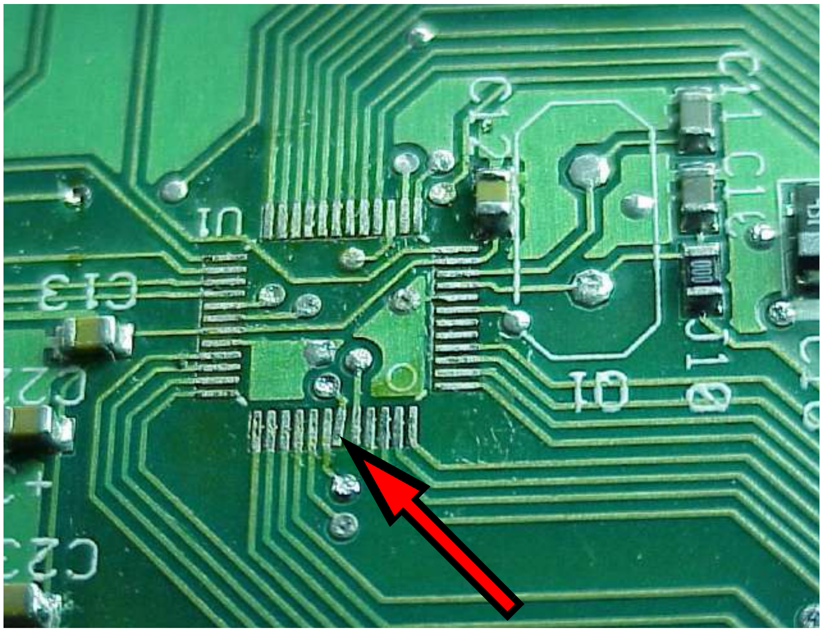
C12
C11
C16
U1
C13
C22
C23
Q1
J10

Figure 23. Clean pads but one pad is slightly crooked

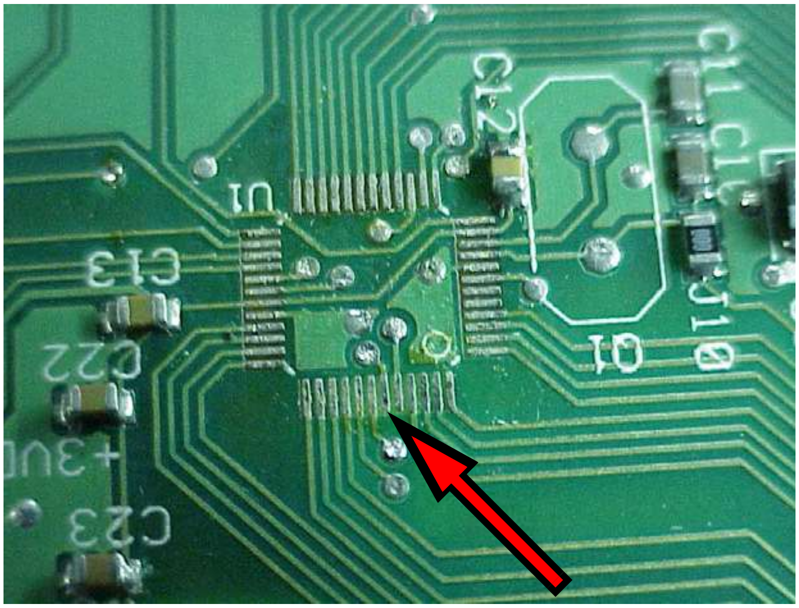
C12
C11
C16
U1
C13
C22
+3V
C23
Q1
J10

Figure 24. Pad straightened out

B. Soldering a new QFP

The pads on the PCB should be clean and free of any solder.

Carefully place the new QFP device on the PCB using tweezers or another safe method. Make sure the part is not dropped as the leads can be easily damaged.

Align the part over the pads using a small pick or similar tool to push the part. Get the alignment as accurate as you can. Also, make sure that the part is oriented correctly (pin 1 orientation).

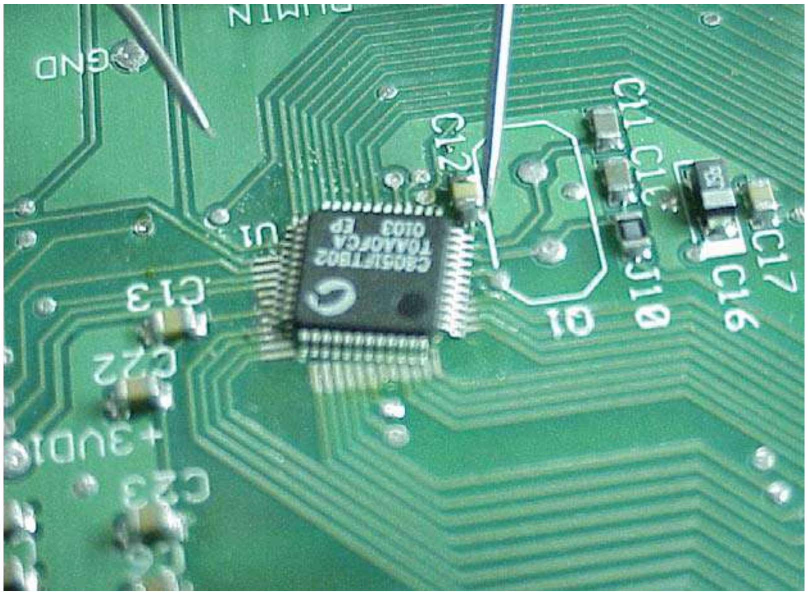

Figure 25. New QFP near pads, ready for alignment

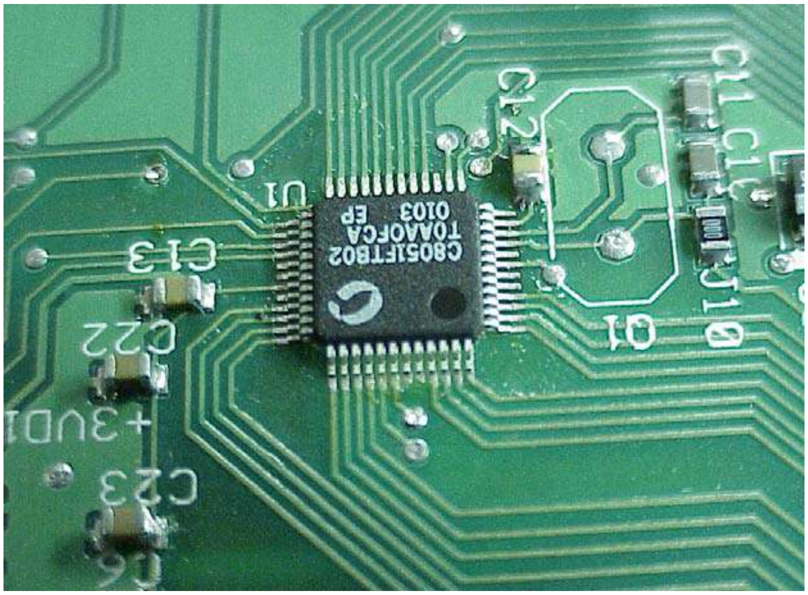

Figure 26. QFP aligned

Adjust the soldering station temperature to 725°F (385°C). Put a small amount of solder on the tip of the solder iron. While holding down the aligned QFP with a pick or other pointed tool, add a small amount of solder flux to the corner leads in two opposite corners. While still holding the part down with the pick, solder down two opposite corner leads on the QFP. Do not worry about excess solder or shorts between adjacent leads at this time. The idea is to anchor down the aligned QFP with solder so it does not move.

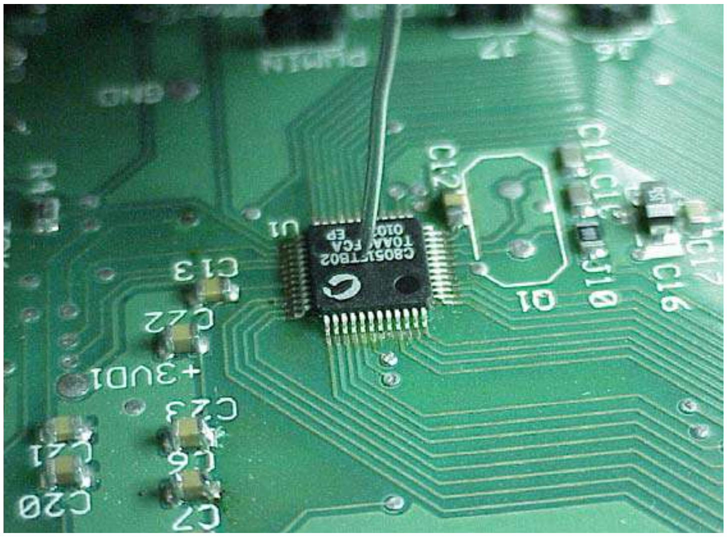

Figure 27. Aligned QFP ready for solder anchoring

Re-check the QFP alignment after soldering the corners. If necessary, make adjustments or remove and start over to get good QFP to PCB alignment.

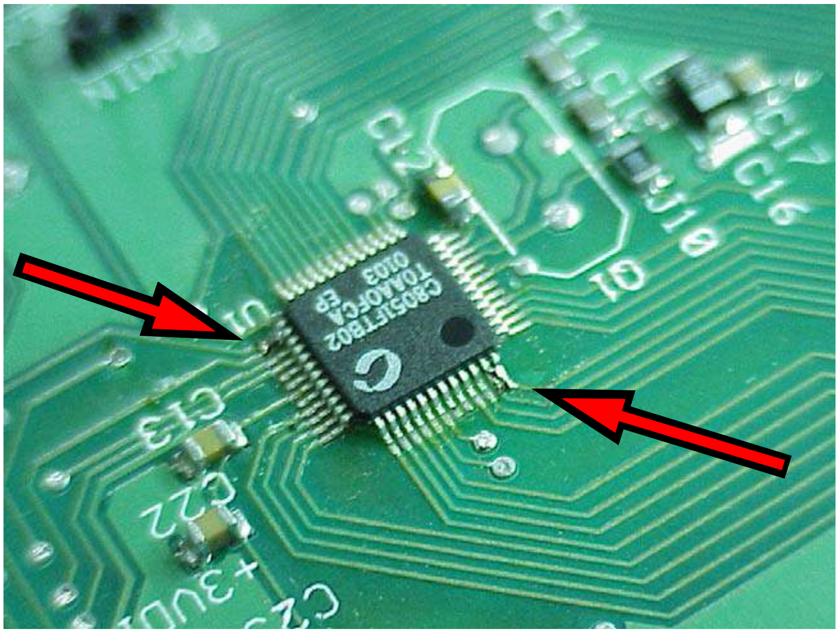

Figure 28. Aligned QFP with corners soldered down

Now you are ready to solder all the leads. Add solder to the tip of the soldering iron. Dispense flux over all the leads to keep them wet.

Touch the solder iron tip to the end of each QFP lead until the solder is seen running up the lead. Repeat for all the leads. Add small amounts of solder as needed to the soldering iron tip. Again, do not worry if you see some solder bridging as you will clean that up in the next step.

When soldering, keep the soldering iron tip parallel with the pins being soldered to prevent excessive solder shorts.

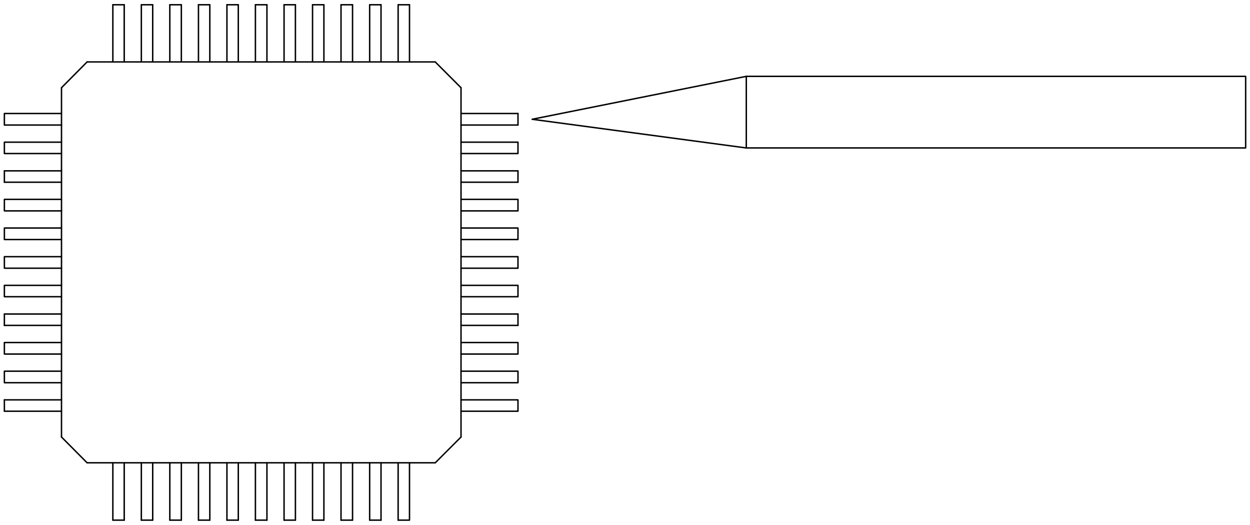

Figure 29. Keep iron tip parallel to pins being soldered

After soldering all the leads, wet all the leads with flux to enhance the solder wicking cleanup. Wick up solder where needed to eliminate any shorts/bridging.

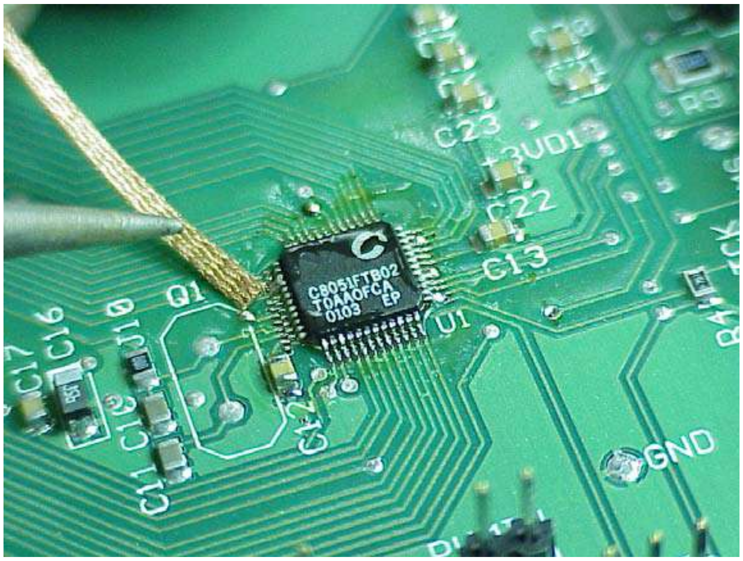

Figure 30. Wicking Solder #1

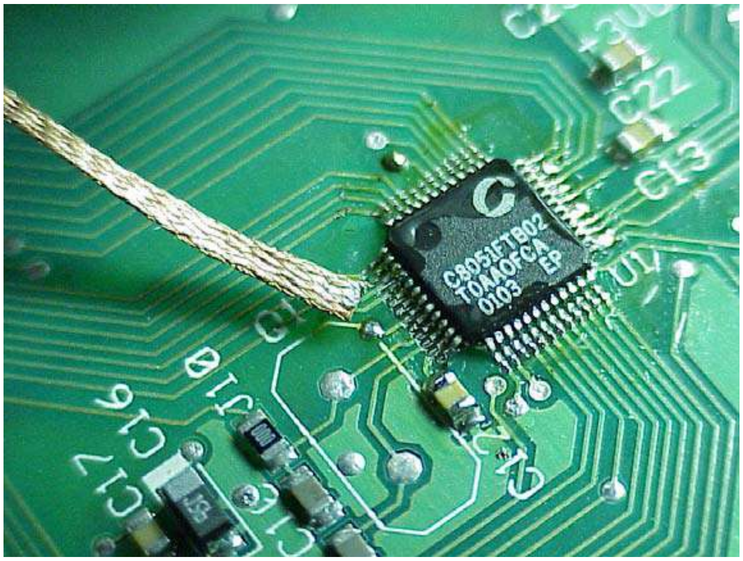

Figure 31. Wicking Solder #2

Hand inspect the board using 4X magnification (or higher) for shorts or marginal solder joints. Solder joints should have a smooth melt transition between each device pin and the PCB. Rework any pins as needed

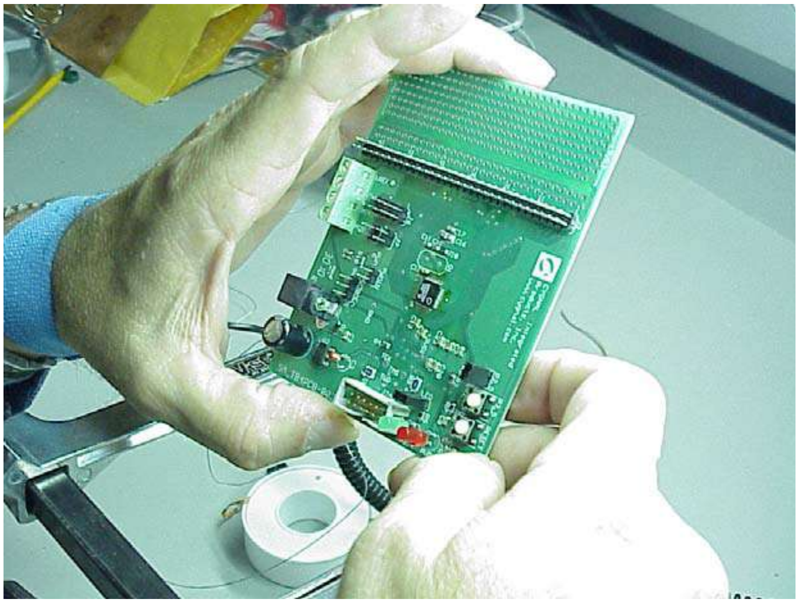

Figure 32. Visual Inspection

After the inspection passes, it is time to clean the flux off the board. Dip the stiff bristle brush into alcohol and wipe in the direction of the leads. Use moderate, but not excessive pressure. Use liberal amounts of alcohol and brush well between the QFP leads until the flux disappears.

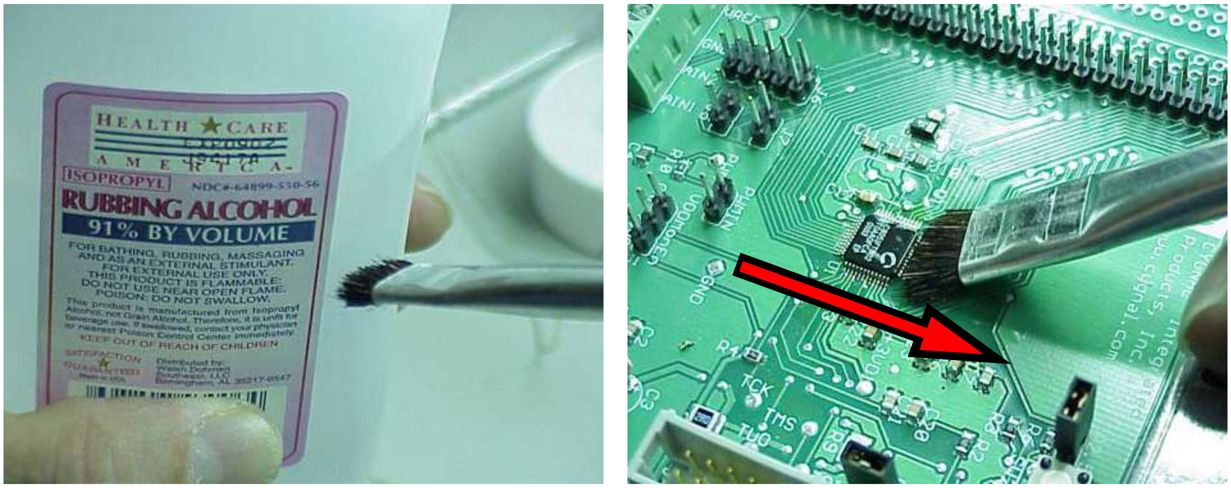

**Figure 33. Isopropyl Alcohol and stiff brush used for cleanup.
Brush only in direction of leads**

Dry the board with compressed dry air or nitrogen. If this is not available, let the board dry for 30 minutes or more to let the alcohol evaporate under the QFP. The QFP leads should look bright and there should be no flux residue.

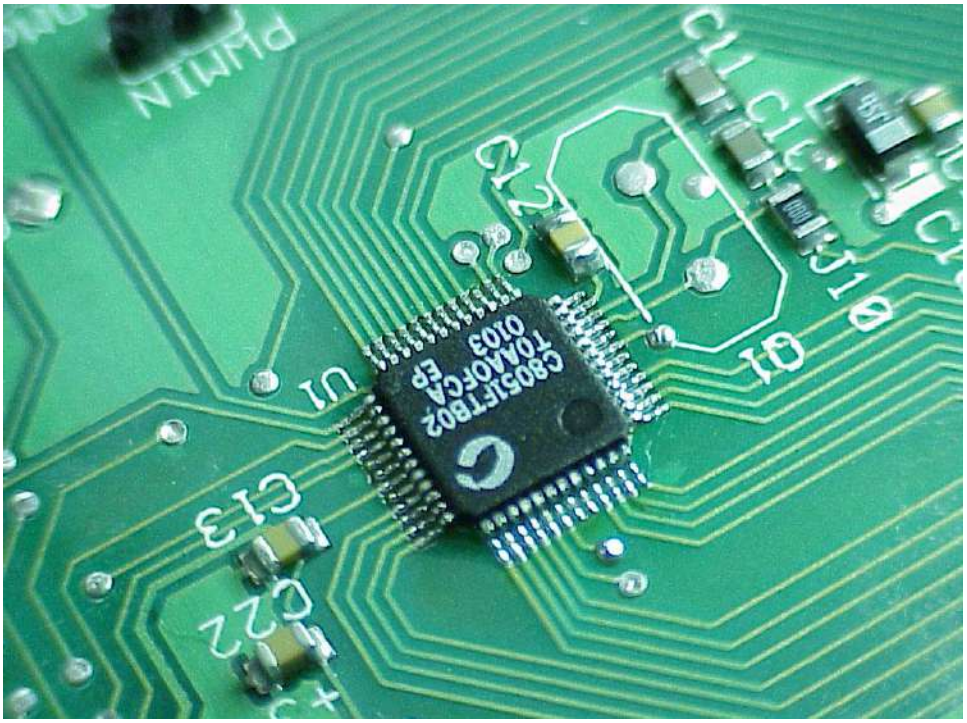

Figure 34. Clean and Pristine

Re-inspect the board for workmanship. Rework any leads if needed.

Figure 35. Stereo zoom inspection station (7X to 40X magnification) helps to inspect solder joint workmanship

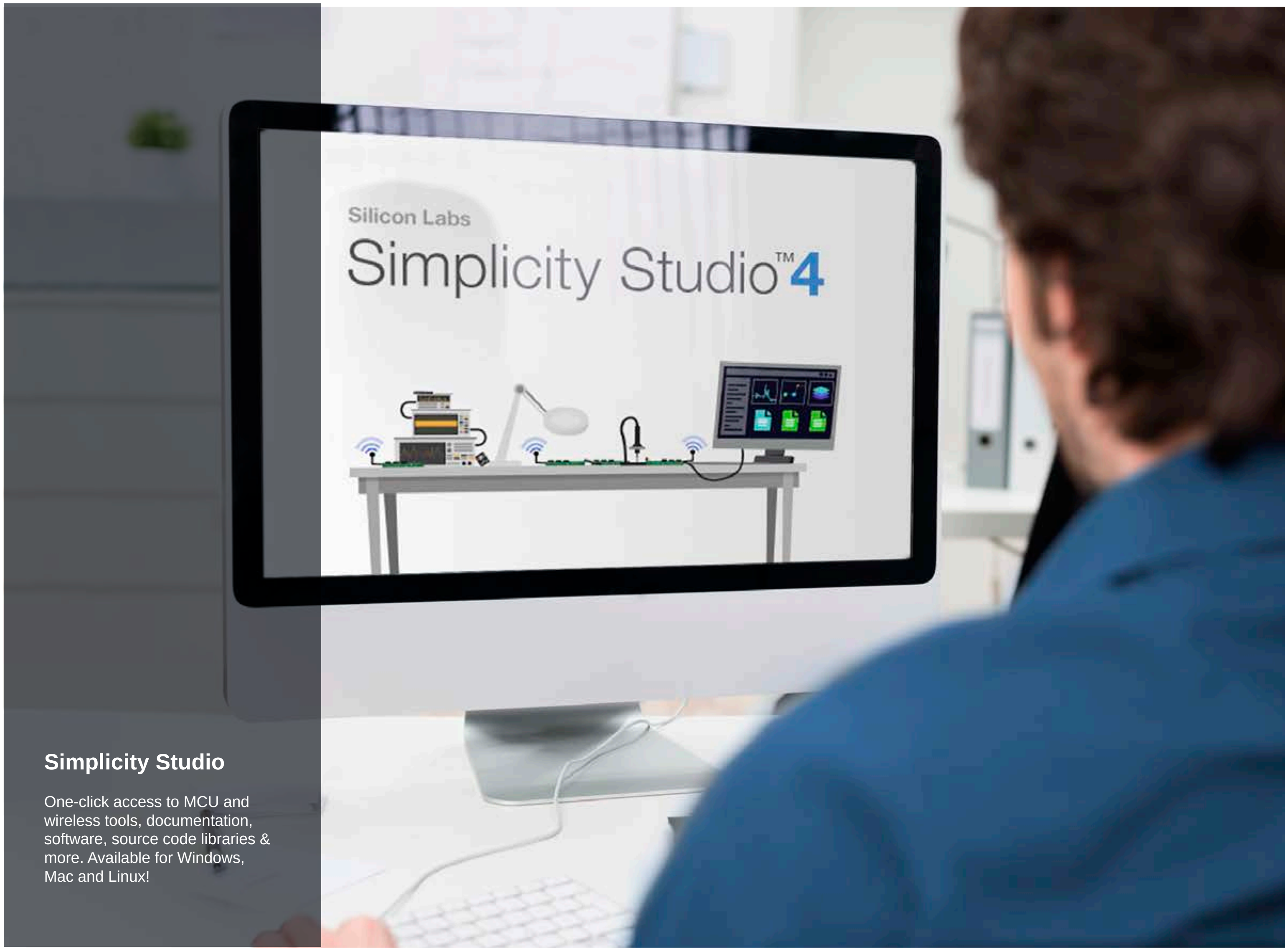

Simplicity Studio

One-click access to MCU and wireless tools, documentation, software, source code libraries & more. Available for Windows, Mac and Linux!

IoT Portfolio
www.silabs.com/IoT

SW/HW
www.silabs.com/simplicity

Quality
www.silabs.com/quality

Support and Community
community.silabs.com

Disclaimer
Silicon Labs intends to provide customers with the latest, accurate, and in-depth documentation of all peripherals and modules available for system and software implementers using or intending to use the Silicon Labs products. Characterization data, available modules and peripherals, memory sizes and memory addresses refer to each specific device, and "Typical" parameters provided can and do vary in different applications. Application examples described herein are for illustrative purposes only. Silicon Labs reserves the right to make changes without further notice and limitation to product information, specifications, and descriptions herein, and does not give warranties as to the accuracy or completeness of the included information. Silicon Labs shall have no liability for the consequences of use of the information supplied herein. This document does not imply or express copyright licenses granted hereunder to design or fabricate any integrated circuits. The products are not designed or authorized to be used within any Life Support System without the specific written consent of Silicon Labs. A "Life Support System" is any product or system intended to support or sustain life and/or health, which, if it fails, can be reasonably expected to result in significant personal injury or death. Silicon Labs products are not designed or authorized for military applications. Silicon Labs products shall under no circumstances be used in weapons of mass destruction including (but not limited to) nuclear, biological or chemical weapons, or missiles capable of delivering such weapons.

Trademark Information
Silicon Laboratories Inc.® , Silicon Laboratories®, Silicon Labs®, SiLabs® and the Silicon Labs logo®, Bluegiga®, Bluegiga Logo®, Clockbuilder®, CMEMS®, DSPLL®, EFM®, EFM32®, EFR, Ember®, Energy Micro, Energy Micro logo and combinations thereof, "the world’s most energy friendly microcontrollers", Ember®, EZLink®, EZRadio®, EZRadioPRO®, Gecko®, ISOmodem®, Precision32®, ProSLIC®, Simplicity Studio®, SiPHY®, Telegesis, the Telegesis Logo®, USBXpress® and others are trademarks or registered trademarks of Silicon Labs. ARM, CORTEX, Cortex-M3 and THUMB are trademarks or registered trademarks of ARM Holdings. Keil is a registered trademark of ARM Limited. All other products or brand names mentioned herein are trademarks of their respective holders.

Silicon Laboratories Inc.
400 West Cesar Chavez
Austin, TX 78701
USA

http://www.silabs.com

Soldering Guidelines

Solder0 R0.1

EE400/1
Design

Revision History

Rev.	Date	Contributor	
0.1	September 2005	L. Wyard-Scott	Updates to reflect changed Capstone Design Course format and new de/soldering equipment.
0.0	January 2004	V.J. Sieben, L. Wyard-Scott, E. Tiong	Creation.

Contents

1 Soldering Technique **1**

1.1 Required Tools . . . 1

1.2 Important Soldering Tips . . . 2

1.3 Precautions . . . 3

1.4 How to Solder Through-Hole Components . . . 4

1.5 How to Solder Surface-Mount Components . . . 8

1.6 Wire Connectors and Headers . . . 9

  1.6.1 Crimp Connectors and Pins . . . 9

  1.6.2 DB Connectors (Solder Cup Type) . . . 10

  1.6.3 Joining Two Wires . . . 10

1.7 Testing Connections . . . 11

1.8 Milled Board Soldering . . . 12

1.9 Post-Soldering Cleanup . . . 12

  1.9.1 PCB Cleanup . . . 12

  1.9.2 Work Area Cleanup . . . 12

2 Desoldering Technique **13**

2.1 Solder Pump/Sucker . . . 13

2.2 Desoldering Braid/Wick . . . 14

2.3 Desoldering Iron . . . 15

2.4 Removing Components Effectively . . . 16

3 Glossary **17**

Bibliography **18**

A Pictures of Tools **19**

Chapter 1

Soldering Technique

This guide is intended to give an introduction to basic soldering technique. If the reader desires to pursue advanced soldering techniques, there are several books as well as technical institutions that offer detailed courses in the area of soldering.

1.1 Required Tools

Before discussing soldering techniques, one should must set up a "soldering work station". The following is a list of tools for a typical soldering station. See Appendix A for pictures of the various parts listed below.

1. **Variable temperature soldering iron:** used for applying heat to joints during the soldering process.
2. **Damp sponge:** for cleaning soldering iron tip.
3. **Rosin-core solder:** to electrically and mechanically bond a component to the PCB.
4. **Wire cutters or side cutter:** for trimming component leads and stripping insulation from wires.
5. **Needle nose pliers:** for holding, placing and shaping components.
6. **Desoldering pump and/or desoldering braid:** for removing solder.
7. **Scotch tape and/or a "Third Hand":** for securing components.

8. **Safety glasses:** for eye protection. These are mandatory in the lab.
9. **Magnifier:** to provide more detail during intricate work. A magnifying glass is convenient, but an illuminated magnifier is better.
10. **Light source:** to prevent eye-strain.
11. **Ventilation:** to extract and dispel fumes generated during the soldering process.
12. **Flux:** to clean components and PCB pads.
13. **Acid brush:** to assist in the removal of flux residue.

Prior to soldering it is a good idea to have all components organized as it will make populating the PCB more efficient. Make a BOM (bill of materials) for the PCB, and ensure before soldering that the components have been collected. The schematic and PCB layout will also be referred to when populating the board.

1.2 Important Soldering Tips

The following tips provide a quick guideline on how to make proper joints.

Cleanliness: All parts, including the soldering iron tip, must be clean and free from grease, oxidation and contamination. Solder does not flow over contaminated areas; moreover, solder is repelled by dirt. Severe contamination is evident when solder begins to “bead”. A common source of contamination is oxidation. Old components and copper boards will often have an oxide layer that prevents a good solder joint. Ensure all components have shiny leads and the PCB has clean traces. An abrasive such as a blue or pink eraser, emery paper, or stell wool can be used to remove the oxidized layer from the PCB board and components.

Tinning: In addition to being clean, the soldering iron tip must also be tinned (coated with solder). Tinning the tip allows solder to flow on the components more quickly rather than the soldering iron tip itself. Tinning involves adding a few millimetres of solder to the tip and then wiping and rotating the tip on the damp sponge to reveal a shiny surface on the tip of the soldering iron: a thin layer of solder will coat or “tin” the tip of the soldering iron. **When done soldering,**

tinning the iron is required to protect the tip from oxidation thereby dramatically increasing its life.

Temperature: Ensure that both the component leads and the PCB's copper layer are heated at the same time. The soldering iron tip should contact both the component and the PCB pad. This will ensure that each surface is relatively close in temperature resulting in a good joint. If there is a temperature difference between the two surfaces, the solder will form a "dry" joint. Soldering irons are typically set around 650 Fahrenheit, depending on the lead-tin ratio of the solder being used. Too much heat causes excessive "sputtering" of flux, and too little doesn't melt the solder in a timely manner.

Duration: The duration that the iron is in contact with the component and PCB is dependent on the size of the joint and your soldering iron temperature. For the typical PCB through-hole joint, it should take a few seconds to heat the joint and apply the solder. This will require practice, so don't expect to be fast if you are a beginner. Excessive heat (several seconds in duration) will damage sensitive semiconductors. If this is a concern, use a heat sink attached to the component leads: sometimes as simple as an alligator clip. These concerns can sometimes be avoided by soldering sockets instead of the semiconductor itself.

Adequate solder coverage: If too little solder is applied, the joint will not make a secure connection and will cause erratic behaviour. However, if too much solder is applied, the joint may bridge with adjacent joints resulting in electrical shorts. How much solder to apply comes with experience. Figs. 1.4 through 1.7 show good and bad solder joints.

Handling: Most modern electronics systems contain static-sensitive devices. Use proper handling procedures to minimize the likelihood of damage: grounding wrist-straps, grounded soldering irons, grounding mats, etc.

1.3 Precautions

Soldering Irons get very hot (600-800^{0}F, 315-425^{0}C), please ensure you follow precautions during use. Basic safety precautions are listed below.

- Never leave your iron turned on while unattended.

- Turn the soldering iron off when it is not being used. If the iron is left on for long periods of idle time, the soldering iron tip will be destroyed through oxidation.
- Eye protection must always be worn when soldering. Hot flux can spit up and into an unprotected eye. **In the Capstone Design Lab, use of eye protection is mandatory.**
- If the cord of the soldering iron is damaged, inform the lab staff who will ensure it is replaced.
- Never set the soldering iron down on anything other than an iron stand.
- To prevent burning your fingers, use needle nose pliers, heat resistant gloves, or a third hand tool to hold small pieces.
- Familiarize yourself with the safe handling of all materials used during the soldering process. This includes solder, flux, alcohol, and desoldering braid. Each has a Material Safety Data Sheet (MSDS) and can be found in the lab or online. “Safe Operating Procedures” are found posted on the wall close to the soldering facilities.

1.4 How to Solder Through-Hole Components

Most of the soldering done in the Capstone Design Lab is through-hole. A through-hole joint is a type of soldering joint in which the component joins with the PCB pad through a physical hole in the board. The following steps will illustrate how to make a proper through hole solder joint on a PCB.

1. Ensure that the printed circuit board and all components are clean. Cleaning can be achieved with a mild abrasive and/or the application of flux.
2. Plug in the soldering iron, turn it on, and let it warm up for 2–3 minutes.
3. Wet the soldering station sponge with the water provided in the lab. Do not wet the sponge in the bathroom or the water fountain.
4. Clean the tip of the soldering iron and tin it with solder.

5. Insert the component into the holes. Ensure that the component is secure by taping the component or by using a third hand. Optionally, the component leads can be clinched as shown in Fig. 1.1. This technique, however, is not recommended for two-sided boards as the flow of solder to the component side is restricted.

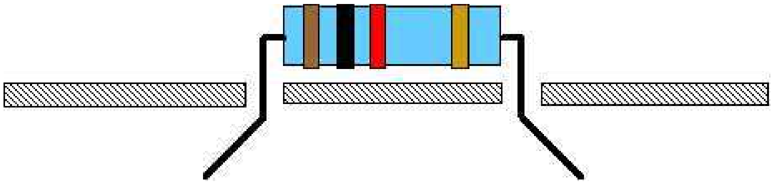

Figure 1.1: Clinched Component Leads

6. Apply the soldering iron tip to one side of joint making contact with the component lead and the board copper foil, ensuring that both are heated up to the same temperature as shown in Fig. 1.2 and Fig 1.3. Notice the tinned tip in Fig. 1.3.
7. Slowly add a few millimetres of solder to the other side of the joint. DO NOT apply solder to the soldering iron tip. If enough heat was applied to the PCB pad and component wire, the solder will flow freely onto the joint.

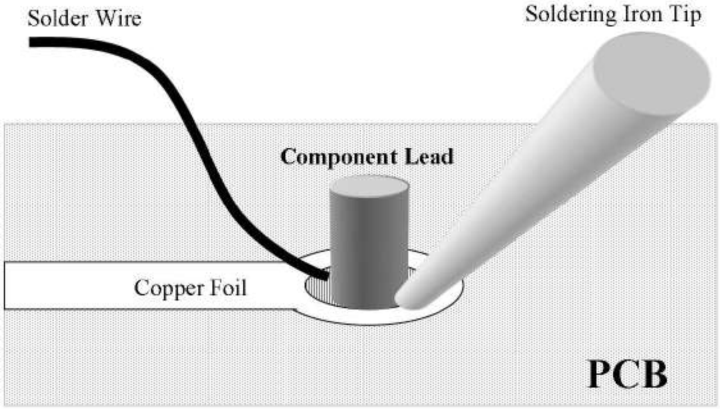

Figure 1.2: Method of Soldering Through Hole Joints

8. Remove the solder when the joint is suitably covered as shown in

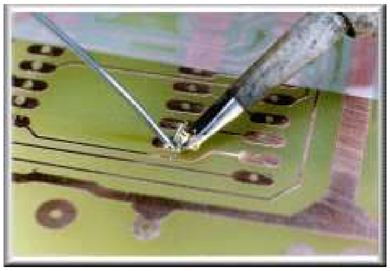

Figure 1.3: Method of Soldering Through Hole Joints, [8]

Fig. 1.4. The goal is to get the joint to be a "fillet": a curve as shown in Fig. 1.4.

9. If the PCB is double-sided, the solder should flow through the hole around the component lead and make a bond on the component side of the board (opposite to the side that the solder was applied). If this "wicking" does not occur, the hole may be undersized, clinching could be blocking the solder's path, or the component lead is not clean.
10. Remove the soldering iron and allow the joint to cool naturally.

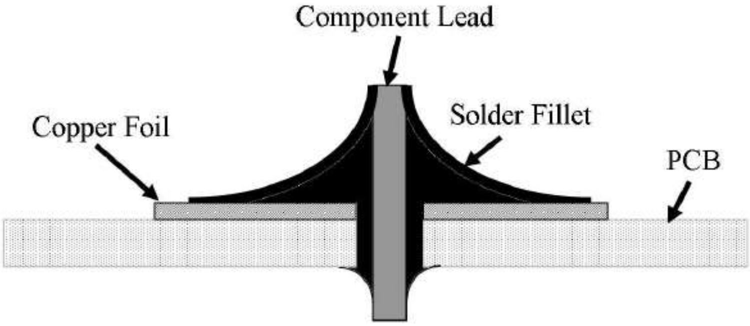

Figure 1.4: Good Solder Joint - Solder Fillet

11. Cut the lead of the component, if necessary.

Fig. 1.6 and Fig. 1.7 below are pictures showing good and bad solder joints.

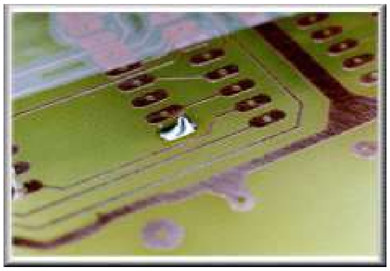

Figure 1.5: Good Solder Joint, [8]

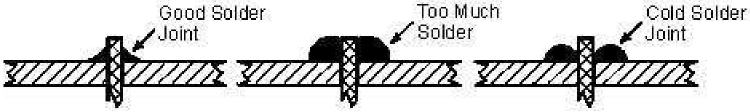
Good Solder
Joint
Too Much
Solder
Cold Solder
Joint

Figure 1.6: Joint Examples, [3]

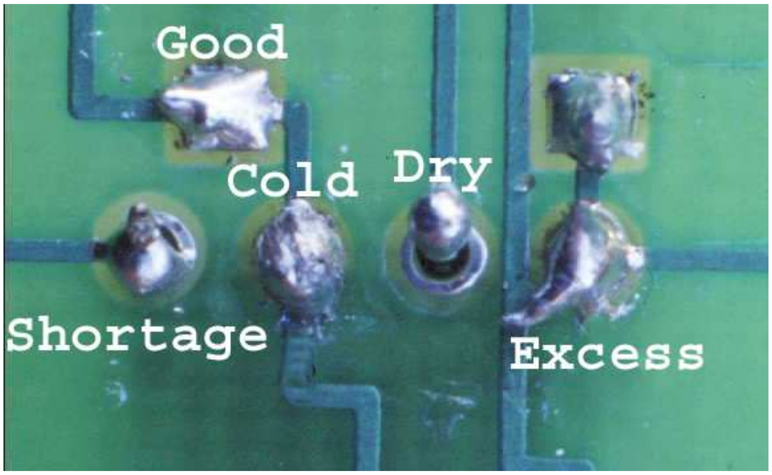
Good
Cold
Dry
Shortage
Excess

Figure 1.7: Joint Examples, [2]

1.5 How to Solder Surface-Mount Components

Surface mount soldering requires more experience and skill than through hole. It is recommended that one practices with through-hole prior to attempting any surface mount soldering. As the name suggests, surface mount involves soldering a component to either the top or bottom surface of a PCB. Depending on the footprint, the pads are usually a spaced closer together (finer pitch), making the soldering more susceptible to solder bridges, etc.

Figs. 1.8 and 1.9 show examples of good surface-mount solder joints.

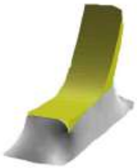

Figure 1.8: Surface Mount Good Solder Joint, [6]

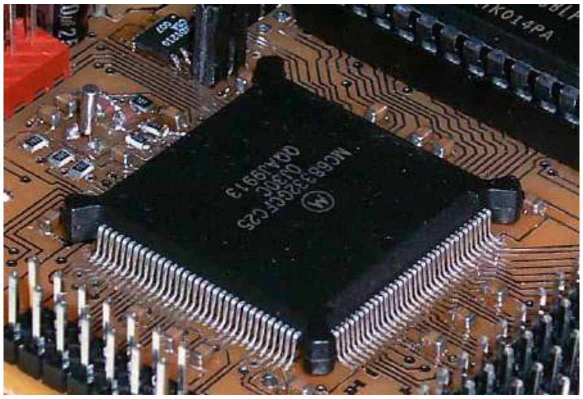

Figure 1.9: Surface Mount IC with Good Solder Joints, [5]

The actual soldering of the joints is similiar to the through-hole method. One difficulty, however, is maintaining the part's alignment on the PCB pads. A good technique is outlined here:

1. Align the component on the PCB pads. This can be aided with the use of tweezers and dental picks.
2. Secure the component to the PCB by applying a small amount of pressure onto the top of the component using a small slot screwdriver.

An index finger resting on the end of the screwdriver provides enough force to secure the device.

3. Solder one of the corner component leads to the PCB pad.
4. Align the remaining pads and solder the opposite corner PCB pad.
5. Solder the remaining pads in a pattern that does not build-up too much heat in the device.

1.6 Wire Connectors and Headers

When PCBs are manufactured they often have connectors to peripheral devices. These connectors are like other components in how they are soldered onto the PCB. However, the plug that matches the connector usually also requires some soldering. **Please note that with very few exceptions, the wire used for cabling is stranded (rather than solid-core) due to its higher strength and flexibility.** Below are procedures for a few of the more common plug types.

1.6.1 Crimp Connectors and Pins

1. Strip off about 1cm of the wire insulation.
2. Place the exposed wire into the crimp-style connector or pin. The wire should just barely show coming out the other side. Some crimp pins have two crimping areas – one for the stripped wire, and the other for the wire *with* the insulation.
3. Crimp the connector with the appropriate crimping tool. Note that the connectors are colour coded to settings on the crimping tool, try and match the correct colour.
4. As added insurance, apply solder to the joint to solidify the connection.

Before crimping all pins destined for a connector housing, start with one and ensure that it fits successfully into the housing. If the crimp is made too tightly, or otherwise mis-shaped, the pin may no longer fit.

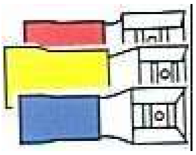

Figure 1.10: Colour Coded Crimp Connectors

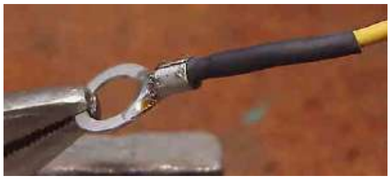

Figure 1.11: Crimp connector without shrink wrap, [1]

1.6.2 DB Connectors (Solder Cup Type)

1. Strip off about 5mm of the wire insulation.
2. Tin the wire.
3. Slide shrink wrap on the wire.
4. Place the wire into the DB - connector solder end.
5. Apply the soldering iron tip to the connector and wire.
6. Apply a few millimetres of solder until the joint is adequately covered.
7. Let the joint cool.
8. Slide the shrink wrap over the joint and heat the shrink wrap with a heat gun (or other heat source) until the wrap shrinks around the joint.

1.6.3 Joining Two Wires

1. Slide shrink wrap on one of the wires.
2. Tin or coat each wire with some solder.
3. Join the wires in a way that provides a good mechanical connection (such as twisting together) and apply the soldering iron tip and some solder.

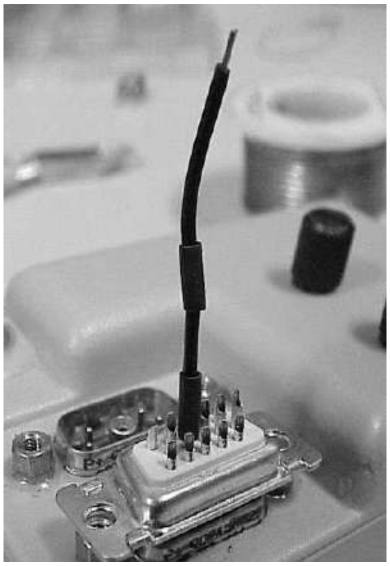

Figure 1.12: DB Connector with shrink wrap, [7]

4. Let the joint cool.
5. Slide shrink wrap over the joint and heat the shrink wrap with a heating gun or a heat source until it the the wrap shrinks around the joint.

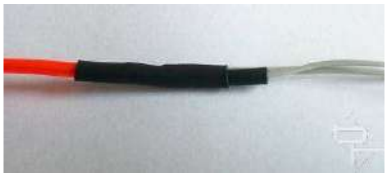

Figure 1.13: Two wire joint insulated by shrink wrap, [4]

1.7 Testing Connections

After completely soldering a component to a PCB, it is good practice to ensure connectivity between the component wire leads and the PCB pads they are soldered to. A DMM (Digital Multi-Meter) is sufficient to determine connectivity; many DMMs include an audible connectivity setting, but failing this, measure the joint resistance.

1.8 Milled Board Soldering

In-lab manufacture of PCBs uses a technique known as “milling”. The milling technique involves cutting out the PCB tracks and pads from the copper-clad board.

Due to its construction, a milled PCB is susceptible to solder-bridging across the milled grooves, particularly when an excess of solder is used. Finding a short caused by a bridge is a difficult task, particularly when many solder joints exist. To this end, make a limited number of solder connections and then test for bridges using a multimeter.

1.9 Post-Soldering Cleanup

1.9.1 PCB Cleanup

The flux left behind by rosin-core solders, or perhaps as part of the cleaning process, needs to be removed from the PCB. Due to the flux’s sticky nature, dirt gathers and contributes to short-circuiting problems. The most frustrating part of this situation is that a short does not necessarily occur immediately. Weeks, or even years later, a short can develop.

To remove flux, alcohol is used. Apply the alcohol liberally and then brush away with an acid brush, starting at the center of the PCB and working out toward the edges. This is a time-consuming task, but a clean board is well-worth the effort.

1.9.2 Work Area Cleanup

The importance of keeping your work area clean cannot be emphasized enough. When clear of obstructions and garbage, handling a hot soldering iron is safer. The soldering process itself involves chemicals and substances which are known to have ill-effects in humans. Wiping-down the work-area surfaces with a moist paper towel will help reduce some contamination. When you are done soldering, **wash your hands with soap and water** to get rid of contamination. The primary concern here is accidental ingestion of the chemicals. For more information, please refer to the Safe Operating Procedures, posted in the lab’s soldering area.

Chapter 2

Desoldering Technique

Desoldering may be required for several reasons:

- a component may have failed
- a wrong part was installed;
- a design modification necessitates a change; or
- if a board contains expensive components that can be salvaged.

Whatever the reason, there are three common techniques to remove solder from a joint: using a desoldering pump, a desoldering wick, or desoldering iron.

Regardless of the method used, if it is permissible to destroy the part during removal, then a lot of time and effort can be saved. For instance, using wire cutters to trim off all the pins of a through-hole IC so they can be removed individually makes the removal process much easier. In most cases, avoiding damage to the PCB is of paramount importance.

2.1 Solder Pump/Sucker

A desoldering pump is exactly what it sounds like: a pump that sucks up solder. Usually the pump is spring-loaded and provides a recoil when released. The four steps below outline how to desolder a joint using a desoldering pump.

1. Prime the desoldering pump. This involves depressing the desoldering pump spring. The pump will click when correctly depressed.

2. Heat the joint from one side with the soldering iron tip. Wait 1-2 seconds until the solder begins to melt from the soldering iron heat.
3. Put the pump tip on the other side of the joint. Don't be afraid to actually touch the joint.

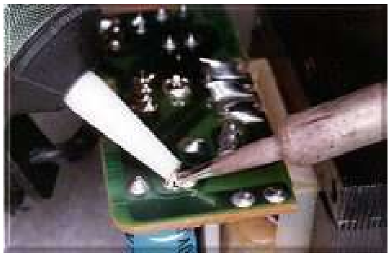

Figure 2.1: Desoldering with a Solder Pump, [8]

4. Press the desoldering pump spring-release button to suck up the solder.

This procedure should be repeated if the joint has a significant amount of solder. If done correctly the joint should eventually look as shown in Fig. 2.2 shown below.

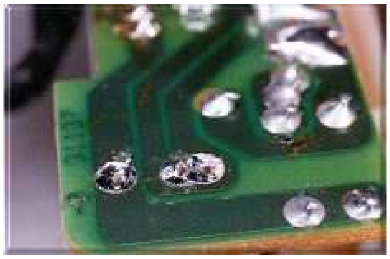

Figure 2.2: Clean Joint from Desoldering with a Solder Pump, [8]

2.2 Desoldering Braid/Wick

A desoldering braid removes solder from a joint using a technique known as “wicking”. Desolder braid is a piece of material that sponges up molten solder by capillary action that draws solder away from the joint the braid. The following procedure outlines the basic steps involved in desoldering a joint using desoldering braid.

1. Place the desoldering braid over the joint.
2. Press the desoldering braid onto the joint with the soldering iron tip. This will apply heat to the desoldering braid and the joint allowing the molten solder to flow.
3. Wait for the solder to melt. The solder should flow onto the braid and away from the joint.

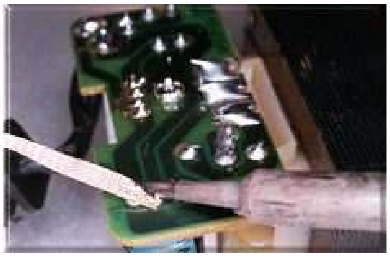

Figure 2.3: Desoldering with a Solder Wick, [8]

4. Cut off the solder coated portion of the desoldering braid. **There should be no copper visible in the portion that is removed: braid is very expensive.**

2.3 Desoldering Iron

Using a desoldering iron, available with higher-end soldering stations, is similar to using a solder sucker. The desoldering iron is essentially a soldering iron with a built-in vacuum.

1. Make contact between the iron and the joint to be desoldered, ensuring the vacuum opening is not blocked.
2. Once the solder on the joint has become molten, depress the button on the iron to activate the vacuum.

It is very important for the life of the desoldering iron that the vacuum assembly be cleaned after every session of use. Please ask the lab personnel for a description of the cleaning procedure.

As with a normal soldering iron, the desoldering iron tip should be tinned before and after every use.

2.4 Removing Components Effectively

To remove a component with little or no damage to the PCB or component takes practice and **patience**. Often when removing a component, PCB pads and tracks may be damaged by "lifting" off the board. Below are some tips that may be useful when removing components from a PCB.

- Desolder all joints pertaining to a component prior to removal. It may be necessary to go over the joints several times before all the solder is removed. Practice patience.

- Use pliers and *gently* pull on components while applying heat to the joints. The assistance of a friend can help with this.

- Do not pull with force, because the PCB pads and tracks will be damaged.

- Do not try and remove components by prying them.

- Do not try to push the component out of the holes with the soldering tip. This will certainly lift the copper pad off of the PCB and ruin the iron tip.

- In order to get better thermal conductivity between the iron and the solder joint, it is sometimes effective to *add* a small amount of solder.

Chapter 3

Glossary

BOM: Bill of Materials, which contains a list of all components and values contained on a particular PCB. Used as a “shopping list”.

Desoldering Braid: A material used to remove solder with capillary action.

Desoldering Pump: A device used to remove solder with a swift vacuum action.

Flux: Cleans the surfaces that are being heated by bringing contaminants to the surface. Most solders include flux in their core. For excessive contamination use a flux pen.

Footprint: The spacing pattern or layout of pads for a particular component or integrated chip, as used on a PCB.

PCB (Printed Circuit Board): A fibreglass board upon which copper traces are laminated to make connections between various components. The copper traces serve as flat wires connecting various components.

Shrink Wrap: A tube that fits over a wire, that when heated shrinks to provide insulation and support for the joints.

Bibliography

[1] R. Glass. Electrical wire soldering for beginners. Technical report. http://www.airheads.org/contrib/solder.html.

[2] D. Lauder. How to solder. Technical report, Oct 2001. http://dragon.herts.ac.uk/ eleqdml/teaching/general/soldering/.

[3] W. M. Leach. The leach amp. Technical report, 2000. See section on Assembly of Circuit Boards, http://users.ece.gatech.edu/ mleach/lowtim/part2.html.

[4] Macroman. Cutting neon string. Technical report. http://www.bit-tech.net/article/59/.

[5] K. Maxon. Have you seen my new soldering iron? Technical report. http://www.seattlerobotics.org/encoder/200006/oven-art.htm.

[6] NIST. Archive of surface evolver code. Technical report. http://www.ctcms.nist.gov/solder/archive.html.

[7] Serial Tester. Serial tester. Technical report. http://army-gps.robins.af.mil/tech/Serial-tester/proc06.jpg.

[8] A. Winstanley. The basic electronics soldering and desoldering guide. Technical report, Feb 1999. http://www.epemag.wimborne.co.uk/solderfaq.htm.

Appendix A

Pictures of Tools

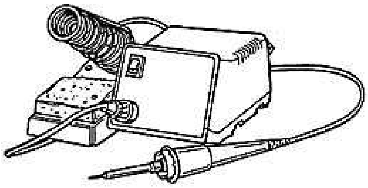

Figure A.1: Soldering Iron

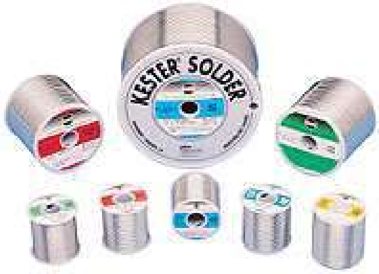
KESTER SOLDER

Figure A.2: Solder Wire

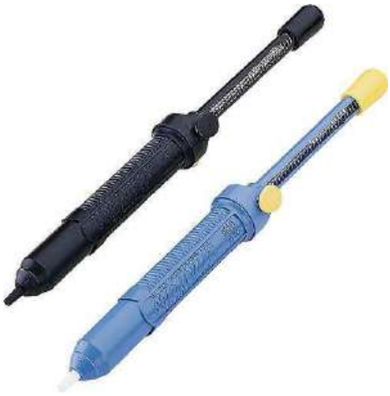

Figure A.3: Desolder Pump

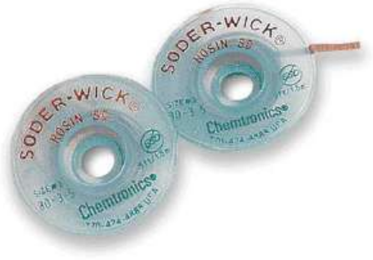
SODER-WICK
ROSIN SD
Chemtronics
SODER-WICK
ROSIN SD
Chemtronics

Figure A.4: Desolder Braid

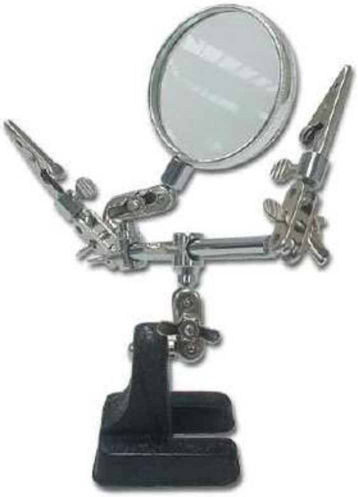

Figure A.5: Third Man

1

Soldering

Contents

Introduction 2

1 Soldering Printed Circuit Boards.. 3

2 Types of Solder.. 4

3 Types of PCB 5
  3.1 Through-Hole Components 5
  3.2 Surface Mount Components 5

4 Flux.. 6

5 Soldering Irons 7
  5.1 Tips 8
  5.2 Tip Contamination and Cleaning.. 9

6 Ovens and Hot Plates.. 10

7 Desoldering.. 11

8 Hazards involved in soldering 12
  8.1 Heat.. 12
  8.2 Toxic materials 12
  8.3 Spattering.. 13
  8.4 Electrical Safety... 13

2

Introduction

Soldering is a process used for joining metal parts to form a mechanical or electrical bond. It typically uses a low melting point metal alloy (solder) which is melted and applied to the metal parts to be joined and this bonds to the metal parts and forms a connection when the solder solidifies. It is different to welding in that the parts being joined are not melted and are usually not the same material as the solder.

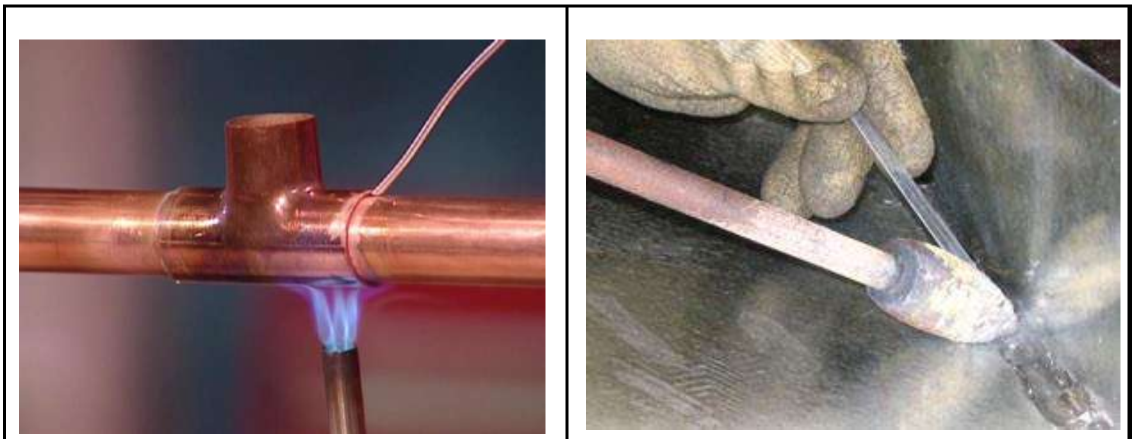

Soldering copper pipe

Sheet metal fabrication

Figure 1 – Different Types of Soldering

Soldering is a common practice for assembling electrical components and wiring. Although it can be used for plumbing, sheet metal fabrication or automotive radiator repair the techniques and materials used are different to those used for electrical work. This document is intended to provide guidance on the safe working methods and proper tools and techniques for soldering of electrical components.

1 Soldering Printed Circuit Boards

Soldering may be used to join wires or attached components to a printed circuit board (PCB). Wires, component leads and tracks on circuit boards are mostly made of copper. The copper is usually covered with a thin layer of tin to prevent oxidization and to promote better bonding to other parts with solder. When soldering bare copper wires they are often “tinned” by applying molten solder before making a joint.

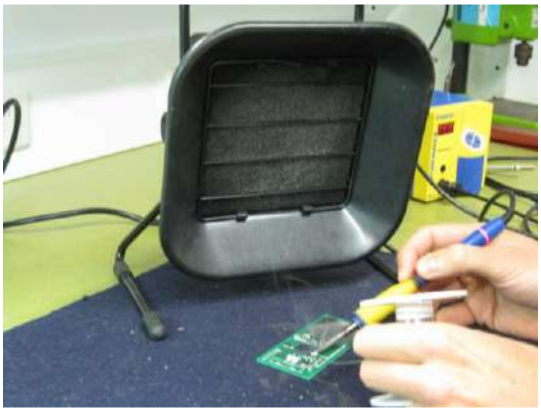	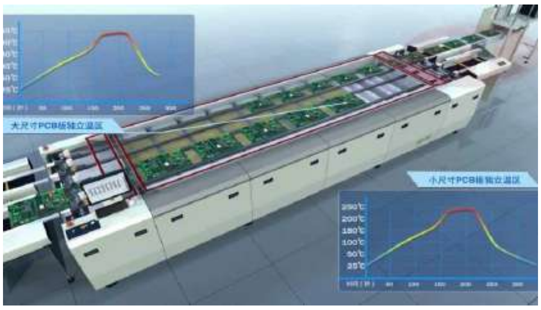
Hand soldering components on PCB	Industrial oven for surface mount

Figure 2 – Different Types of PCB Soldering

2 Types of Solder

There are different types of solder used for electrical work. They are broadly classified as tin/lead solders or lead free solders. Tin/lead solders have been used for many years because of their ease of use however they have been phased out of commercial use due to the harmful effects on humans and the environment. Tin/lead solder is still available and is used by "hobbyists" and other non-commercial users as it is still easier to use than lead free types. When using tin/lead (or leaded) solder there are additional safety precautions that must be observed.

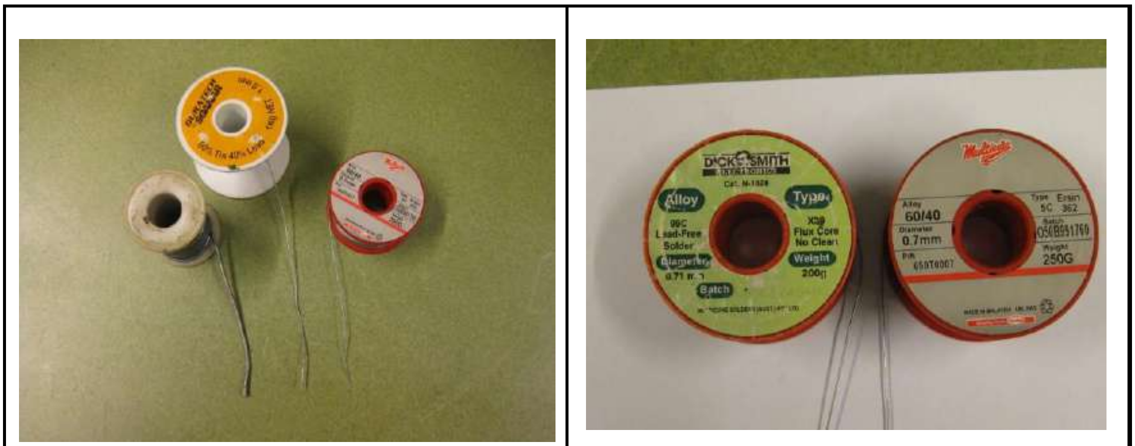
60% Tin 40% Lead
DICK SMITH ELECTRONICS
Alloy
99C Lead-Free Solder
Diameter
0.71 mm
Type
X39 Flux Core No Clean
Weight
200g
Batch
Alloy
60/40
Diameter
0.7mm
Type Ersin 5C 362
Weight
250G

Different gauges of solder wire

Solder composition is labelled
(Lead free on left)

Figure 3 – Different Types of Solder

3 Types of PCB

Printed circuit boards (PCBs) are populated by electronic components and these may be “surface mount” or “through-hole” types.

3.1 Through-Hole Components

As the description “through-hole” suggests, the leads of the component are passed through holes in the PCB and then soldered to a “pad” on the reverse side of the PCB. Soldering is accomplished by heating the component lead and PCB pad with a soldering iron and melting solder wire into the joint. This type of construction was common from the 1960’s until early 2000’s and is still used by hobbyists and in small scale production where manual assembly is preferred.

3.2 Surface Mount Components

Commercial circuits are mostly of the surface mount type as these are cheaper to make, more compact and easier to automate assembly. For surface mount construction the component’s pads are on the same side of the PCB as the component and the component connections sit onto these pads. Soldering is accomplished by applying solder paste onto component pads on the PCB, placing the component onto the paste and then heating the entire assembly to melt the solder. Commercial assembly uses ovens to heat the boards. Hobbyists can also use surface mount components and soldering can be accomplished by applying solder paste and melting with a hot plate, small oven or soldering iron. Some surface mount joints can be soldered using a soldering iron and solder wire.

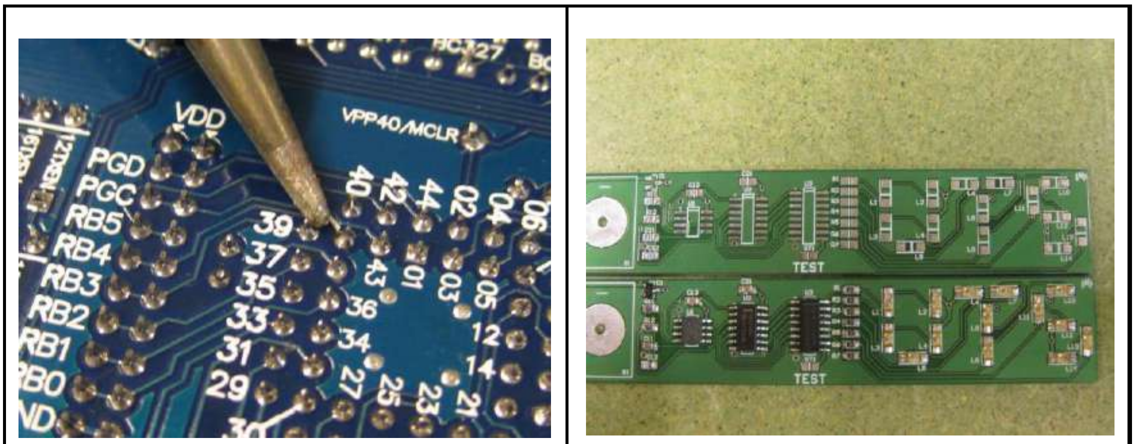

Through-hole PCB
(non-component side)

Surface mount PCB
(component side)

Figure 4 – Different Types of PCB

4 Flux

For electrical soldering both solder wire and solder paste contain flux. This helps to clean the surfaces being soldered and prevent oxidization of the hot solder. The composition of the flux will vary depending on whether it is in a paste or wire, leaded or unleaded solder. Solder wire usually contains a flux called "rosin". Most fluxes will produce fumes when the solder is heated and these fumes are likely harmful to your health. For occasional soldering it may be sufficient to have a well-ventilated workspace but for longer or repeated exposure a fume extractor should be used. Solder flux can also cause solder to spatter and eye protection should be worn when soldering.

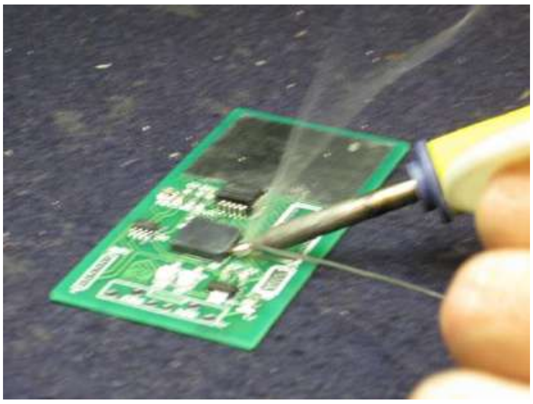	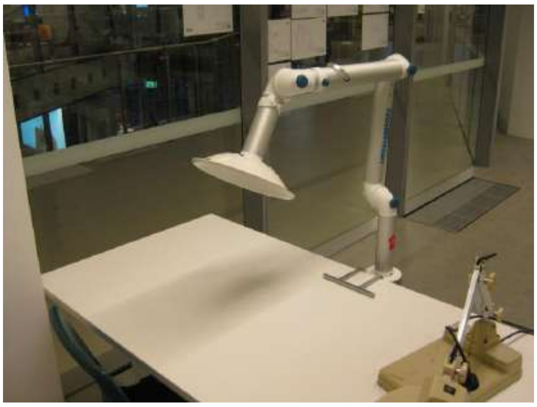
Fumes produced when soldering	Workstation with fume extraction system

Figure 5 – Fumes and Fume Extraction Systems

5 Soldering Irons

Soldering irons come in many varieties and sizes. Soldering irons may be electric, gas powered or externally heated. Most common types are electric. Simple electric soldering irons have no controls and you simply plug them in and wait for them to heat up. Their temperature is regulated by the power of the heating element and heat loss to the environment. Some soldering irons have temperature controls which allow the user to set a desired operating temperature for the soldering iron. This is useful if the soldering iron is being used for different types of solders which have different melting points or if the soldering iron is being used for other purposes such as heating heatshrink. It also introduces a problem if the user does not set an appropriate temperature for the work, solder can be overheated and decompose. Hotter is not better! A temperature of around 320 °C works well for 60/40 leaded solder. Some temperature controlled soldering irons use interchangeable tips to change the temperature at which they operate.

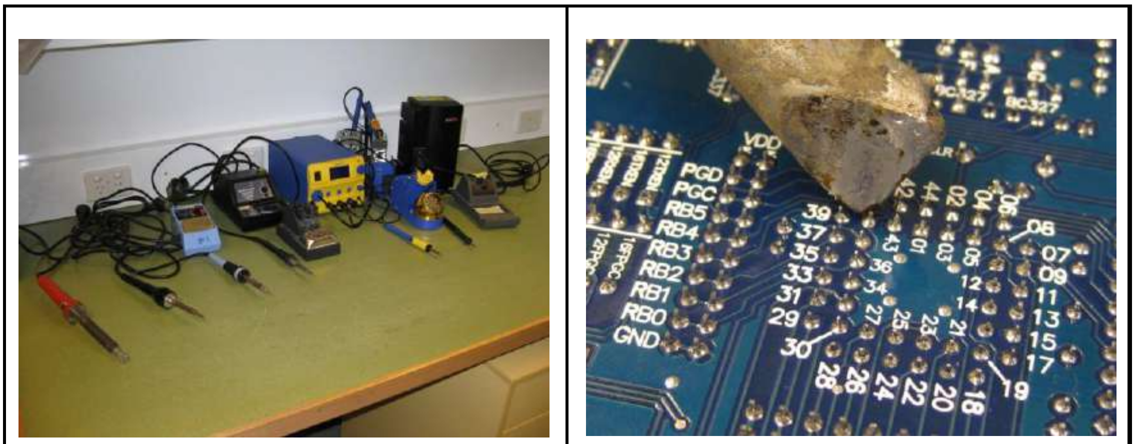

A range of electric soldering irons

Choose an iron appropriate for your task!

Figure 6 – Types of soldering irons

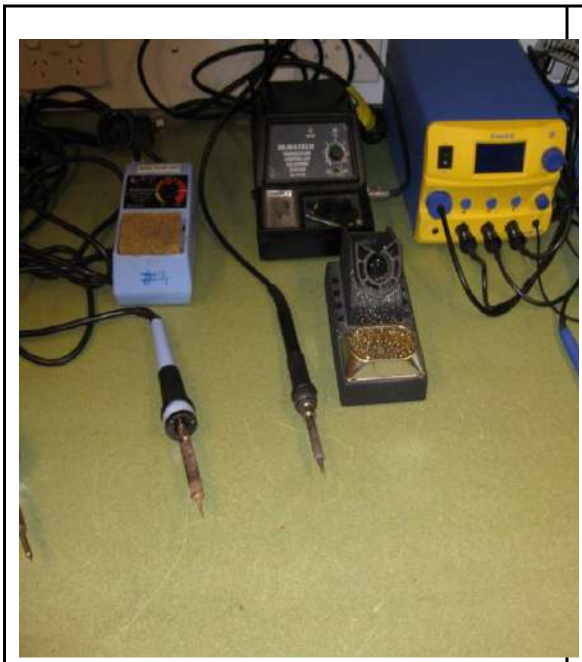	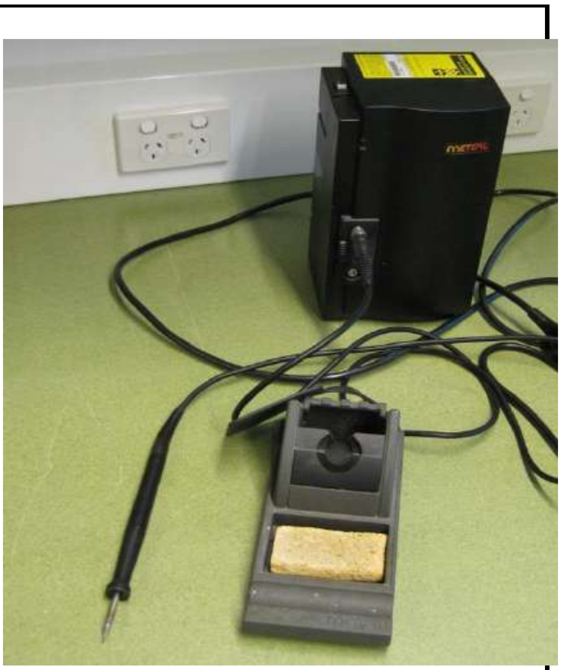
Adjustable temperature soldering stations	Different tips set the temperature for this station

Figure 7 – Types of soldering stations

5.1 Tips

Heat is transferred from the tip of the soldering iron to the joint by thermal conduction enabled by metal to metal contact between the tip and joint. The tips of soldering irons come in various shapes and sizes to enable the best contact to be made. Most tips are either conical or chisel shapes. The shape is largely a personal preference and you can use whichever type works best for you. The size of the tip should be selected to allow the tip to be placed against the joint being soldered without interfering with adjacent parts. The tip should be large enough to conduct sufficient heat into the joint to allow the solder to melt and flow properly. The choice of tip size is not a precise calculation and a “normal” size tip will work for most joints on a PCB.

5.2 Tip Contamination and Cleaning

The thermal conduction from the tip to the joint may be inhibited by contamination on the tip. This contamination can be formed by burnt solder flux or oxidized solder. To make best thermal contact the tip should be cleaned using a tip cleaner(!). Two types are a wet sponge or a brass wire wool. The wet sponge removes the contaminated material when the tip is wiped across it, the water in the sponge cools the solder and the mechanical abrasion removes the contamination leaving a thin coating of clean solder on the tip. This method can cause the tip temperature to dip momentarily. The brass wire wool type removes the contamination by mechanical abrasion and bonding contaminated solder to the brass. The tip is pushed into the brass wool and when it is withdrawn the tip is clean with a thin coating of solder. You must not “wipe” the tip on the brass wool type because the springiness of the brass wool may flick molten solder which may cause burns to people or objects. You should never “flick” excess solder from the soldering iron as this may also cause burns or damage.

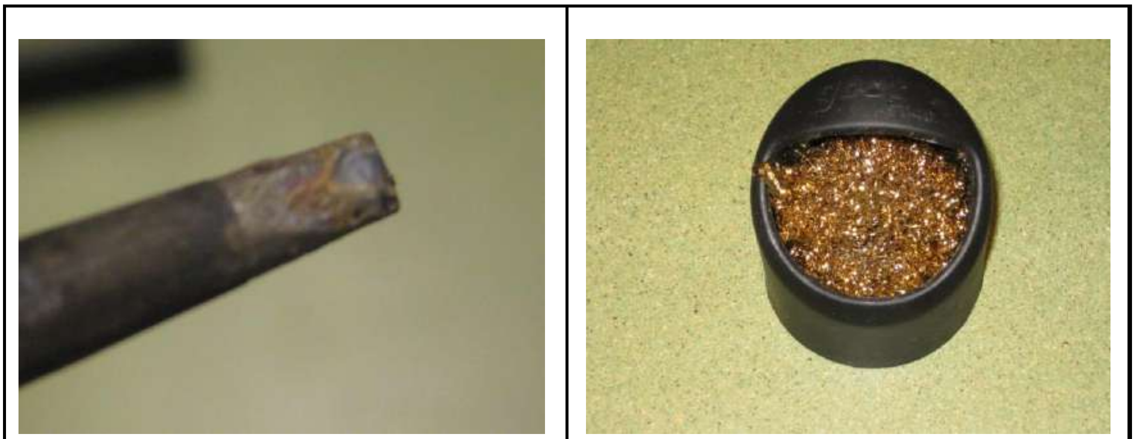

Tip contaminated with old solder and burnt flux

Brass wool type tip cleaner

Figure 8 – Dirty tip and brass wool cleaner

6 Ovens and Hot Plates

For surface mount soldering the heat is usually applied to the whole PCB and all components soldered at the same time. For commercial work this is done in large ovens, often with conveyors to move the boards through the oven. For small scale work simple infrared ovens are available. Another technique uses a hotplate. The same considerations for temperature apply, although because the heat source does not come into contact with the solder or flux, contamination is less likely. Time and temperature are considerations with these methods as the components are exposed to the high temperature for the period required for the solder paste to melt and flow.

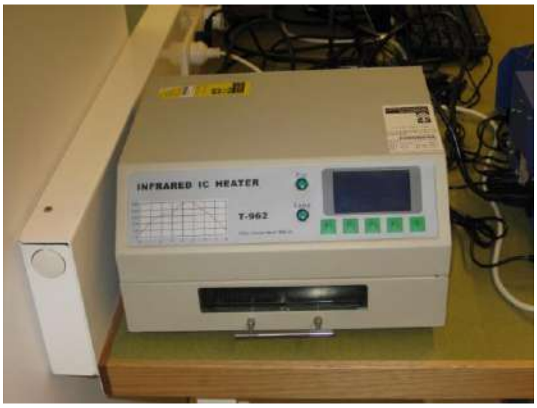	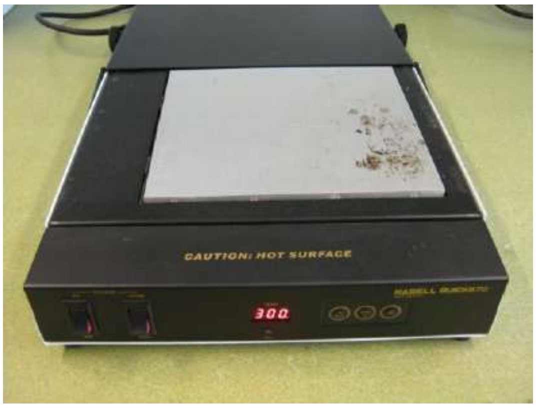
Infrared oven for surface mount soldering	Hotplate for surface mount soldering

Figure 9 – Oven and hot plate used for surface mount soldering

7 Desoldering

If a part that has been soldered needs to be replaced it needs to be “de-soldered”. Depending on the part and type of joint it may be possible to simply re-melt the solder and remove the part, or it may be necessary to remove the solder from the joint so the part can be freed. Some methods for removing solder are solder wick, solder sucker or de-soldering tool. Solder wick is a copper braid which is applied to the joint and heated with a soldering iron. As the solder in the joint is melted it is drawn into the solder wick like a sponge and is removed from the joint. A solder sucker is a spring loaded syringe or rubber bulb. The tip of the solder sucker is placed near the joint as the joint is melted by a soldering iron. When the sucker is operated a vacuum is created which draws the molten solder from the joint into the body of the sucker. A de-soldering tool is a type of soldering iron with a hollow tip and is connected to a pump or vacuum source. The tip of the de-soldering tool is placed onto the joint, typically over a component lead, and once the solder has melted the pump is operated to draw the molten solder away.

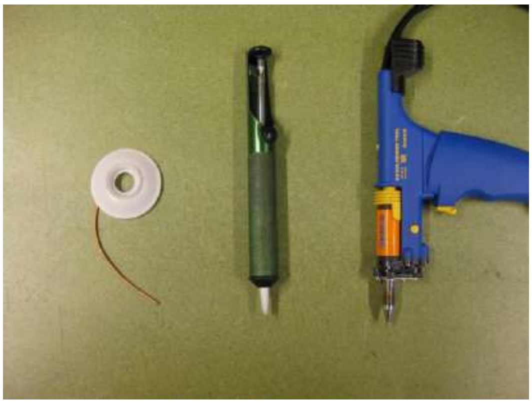	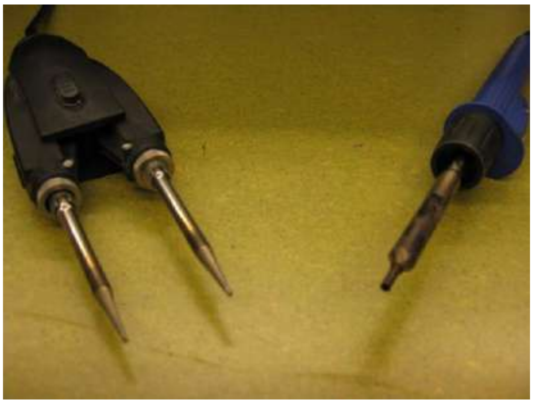
Solder wick, syringe type sucker, vacuum de-soldering tool	Hot tweezers or Hot air nozzle are options for SMDs

Figure 10 – Desoldering tools and tweezers for surface mount devices

8 Hazards involved in soldering

8.1 Heat

Although solder has a relatively low melting point this temperature is more than high enough to cause serious burns to people or objects. It is important to hold the soldering iron only by the insulated handle, never touch the heating element or tip when the soldering iron is on. The soldering iron will remain hot for some time after it is turned off so always check that it has cooled down before touching it, e.g. if changing the tip. When you are not soldering always keep the soldering iron in a proper holder so that you don’t touch it accidently and it doesn’t heat or burn other objects such as the benchtop. Don’t hold parts being soldered with your hands as these will also be heated when being soldered. Don’t flick molten solder from the soldering iron or wipe the tip on brass wool type tip cleaners.

If using a hot plate for surface mount soldering do not touch the hot plate. Use utensils such as pliers to place and remove PCB’s from the hotplate. If using hot air tools for soldering, de-soldering or rework, do not direct the hot air stream onto yourself or other people. If using an oven allow the PCB to cool before handling or use utensils. Don’t place hot PCB’s on temperature sensitive surfaces.

If burns occur they should be treated by holding under cold running water for several minutes and assistance sought if burns are severe. Incidents should be reported.

8.2 Toxic materials

Leaded solder contains lead which is a harmful material. Use of this type of solder will probably involve handling it and your skin may become contaminated by it. Although it is unlikely that the lead can be absorbed directly through your skin it may be ingested indirectly if it is transferred by handling food whilst your skin is contaminated. Always wash your hands thoroughly before eating or handling food.

Solder flux creates fumes when heated during soldering which may be harmful if inhaled. Use a fume extractor to avoid inhaling fumes.

8.3 Spattering

Solder and flux can spit or spatter when heated. Always wear eye protection (safety glasses) when soldering.

8.4 Electrical Safety

Electric soldering irons are plugin appliances and must have a current safety test tag. The test will confirm that the soldering iron conforms to electrical safety standards and has not been damaged at the time of the test. Before use you should visually check that the soldering iron does not have damage such as melted insulation on the lead, broken or cracked handle or exposed conductors. Don’t use damaged equipment and report the damage.

For electrical safety the exposed metal parts such as the tip and heating element are earthed. Don’t solder on any live equipment as contact with the earthed tip may cause damage to the equipment or soldering iron.

SMD Soldering

Information & Instructions

This is a walkthrough of a SMD Soldering Workshop. The following information and instructions can be used as a simple method for soldering SMD components. The recipient of this information should acknowledge the danger of soldering *(written on the back page of this pamphlet)* before attempting to solder.

Recommended Tools:

- Soldering Iron
- Solder
- Tweezers
- Solder Wick
- Flux Pen

① Add solder to one pad.

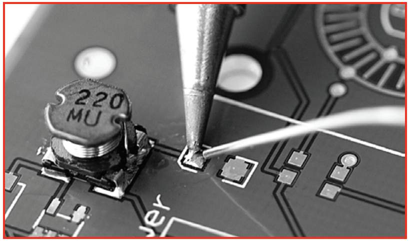

② While that pad is molten, slide the component into place. Do not push down from the top - slide the component into the blob of solder horizontally.

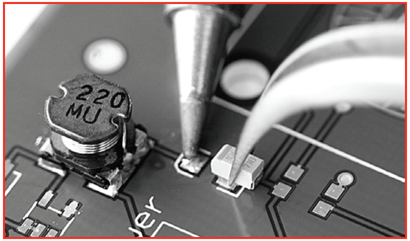

③ Align the component while connection is still molten.

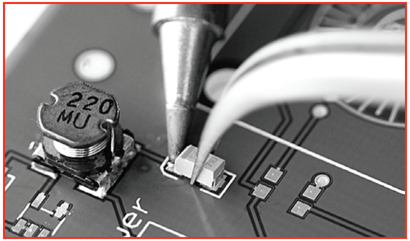

④ Once you have good alignment, continue to hold the component in place, and remove your iron. Continue to hold component for 1-2 seconds while the solder joint solidifies.

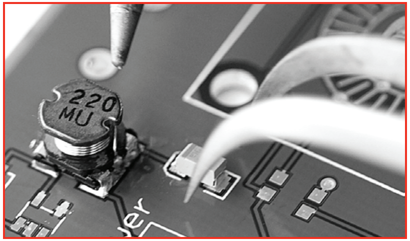

⑤ From above, the alignment looked good. From the side, you can see the rear pad is hovering slightly above the PCB. This can lead to problems on multi-pin components (open connections). Be sure the component is flush up against the PCB before soldering more connections. Re-grip the component, re-heat pad 1 and push the component flush against the PCB.

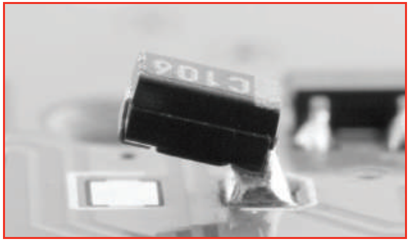

⑥ This is how a tantalum capacitor should look after making both solder connections.

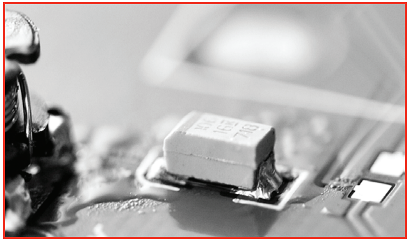

⑦ If alignment is not good, do not solder more than 1 pad! Re-heat the joint, re-adjust component until aligned correctly, then move on to soldering other connections.

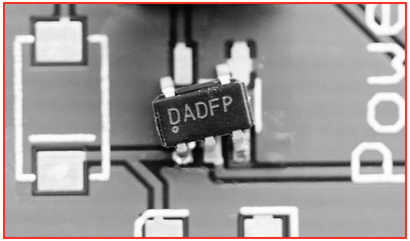

⑧ This is bad. It would be nearly impossible to finish the connections on this part. Make sure you have the component flush against the PCB.

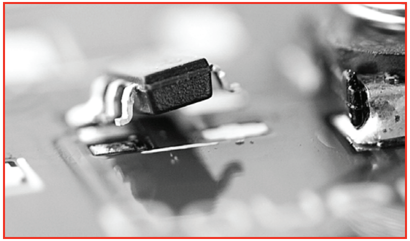

⑨ If you solder multiple pins together, don't worry about it! It can be easily fixed. Do not worry about jumpers! There are actually three pins under that blob.

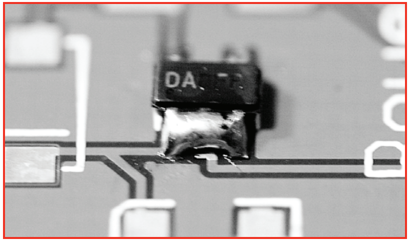

⑩ Pull out some solder wick. Put a small amount of solder on the end of your iron (this will transfer heat from iron to wick to the jumper). Sandwich the wick in between the iron and the solder jumper.

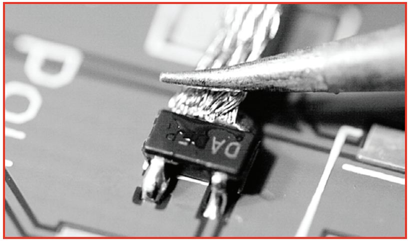

⑪ Hold still for 2-3 seconds. You will see solder start to flow up the wick. Once the excess solder has flowed into the wick, carefully lift up the wick and your iron in one fluid motion.

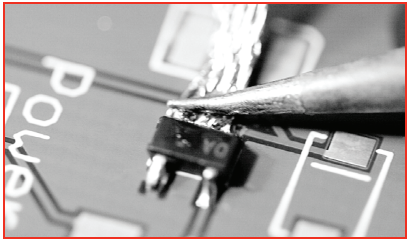

⑫ Nice and clean!

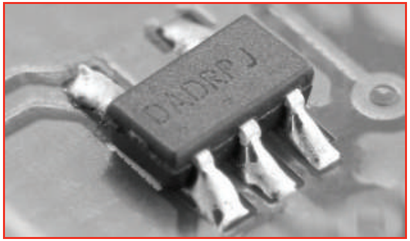

⑬ Bad bad bad. There was not enough solder for the connection on the left. Middle pin is lacking solder and should have been heated for longer. Right pin has had solder applied by an iron rather than applied to two metal contacts (the board and the pin).

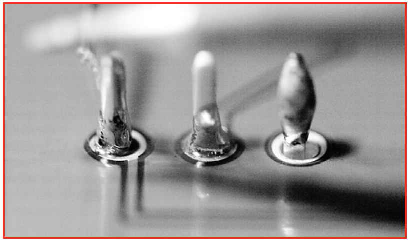

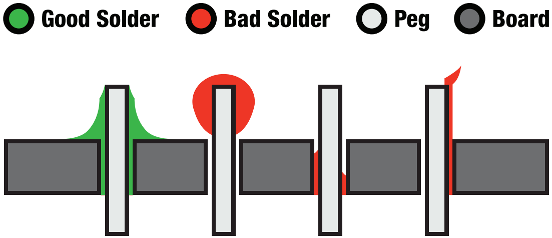

TIPS & TRICKS

	A	Solder flows smooth all the way to the bottom.
	B	Error: Solder balls up on top of pad, not connecting pin to pad. Solution: Flux then wick. If you wick too much add solder.
	C	Error: Too little solder makes for weak connection. Solution: Flux then add solder.
	D	Error: Bad Connection... and ugly... oh so ugly. Solution: Flux then add solder.

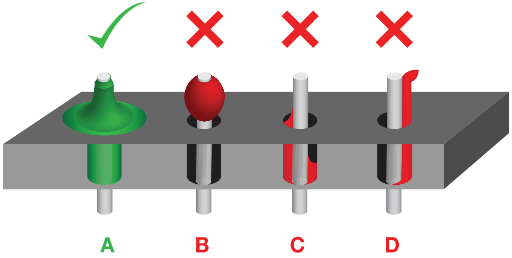

Safety Issues

Soldering

The tip of the iron is normally 700 °F, hot enough to melt metal. It is normal for the handle of the soldering iron to heat up a bit. Hold it like a pencil and move your hand further away from the tip if the heat is uncomfortable. The solder smokes because the rosin inside the solder is burning off - it’s not harmful.

sparkfun.com

RoHS COMPLIANT PB

© SparkFun Electronics, Inc. All Rights Reserved. All other trademarks contained herein are the property of their respective owners.

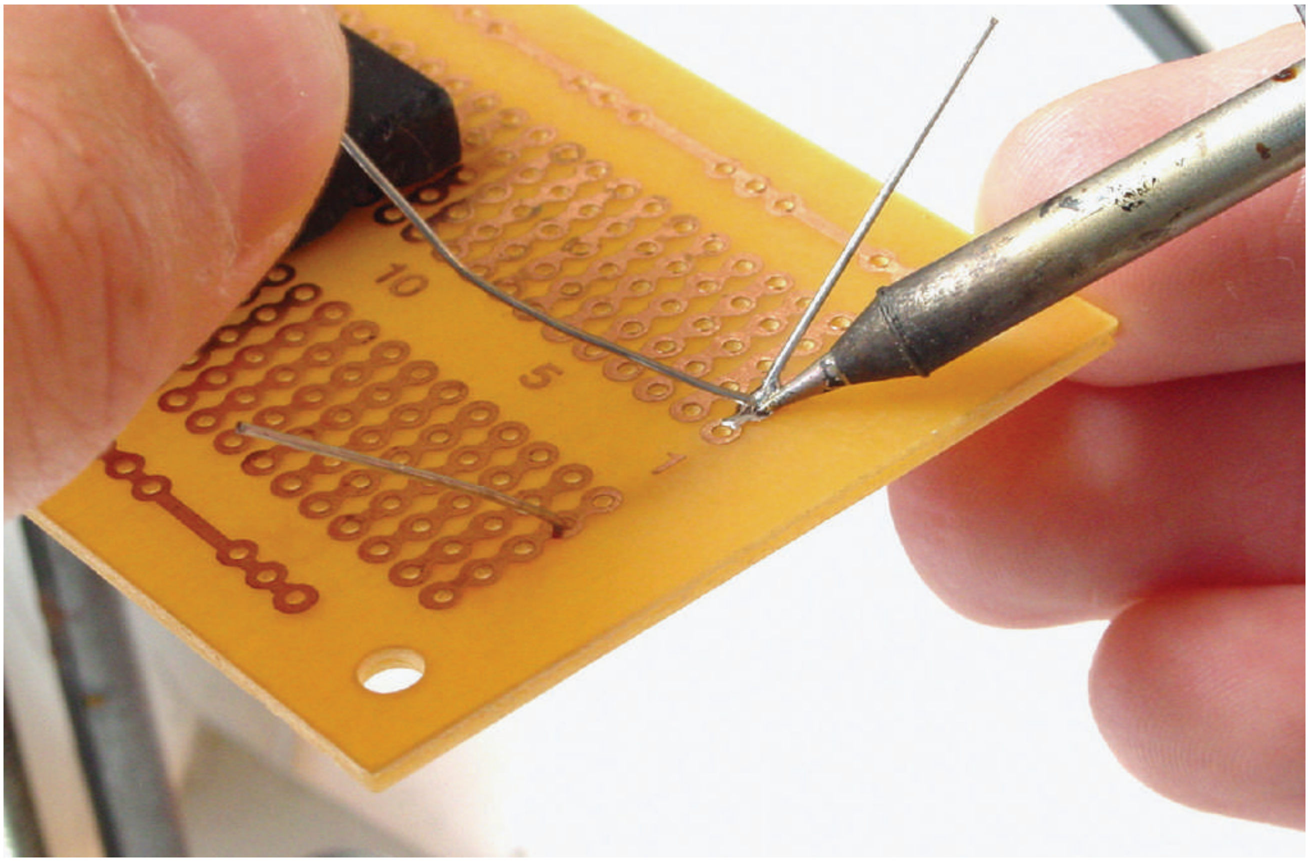

Soldering and Desoldering

Step-by-step instructions for making (and unmaking) the perfect solder joint.

By Joe Grand

The two key parts of soldering are good heat distribution and cleanliness of the soldering surface and component. With practice, you'll become comfortable and experienced with the process.

In this primer, I'll explain how to solder a component onto a printed circuit board (PCB). I'll also provide desoldering tips and show you how to remove a surface-mount component from a printed circuit board using a Chip Quik kit. And I'll show you how to remove a component by removing the solder in a way that won't damage the components or the circuit board.

Photography by Joe Grand

Reprinted with permission from *Hardware Hacking*, copyright 2004, Syngress Publishing, ISBN: 1-932266-83-6, pp. 34–40.

Tools of the Trade

Soldering iron You could pay as little as $10 or as much as $1,000 for a soldering iron. I recommend a fine-tip, 700°F, 50W soldering stick iron. A good general-purpose iron for hardware hacking is the Weller W60P Controlled-Output Soldering Iron, which sells for under $70.

Solder Should be thin gauge (0.032" or 0.025" diameter) 60/40 rosin core.

Desoldering tool (aka solder sucker) A manual vacuum device that pulls up hot solder, useful for removing components from circuit boards. I like the one RadioShack sells (#64-2098, $10).

IC extraction tool Helps lift integrated circuits from the board during removal/desoldering.

Chip Quik SMD Removal Kit Allows you to remove surface mount components quickly and easily. Chipquik.com offers the kit for $16.

Sandpaper A very fine-grit sandpaper is useful for removing oxidation from component and circuit board surfaces.

Desoldering braid Woven metal material used to wick up melted solder.

Small, flat-tip screwdriver Comes in handy for removing some types of components.

Needlenose pliers, wire cutters, and vise These common tools will make your job easier.

The Chip Quik SMD Removal Kit

The Chip Quik SMD Removal Kit allows you to quickly and easily remove surface-mount components such as PLCC, SOIC, TSOP, QFP, and discrete packages. The main component of the kit is a low-melting-temperature solder (requiring less than 300°F) that reduces the overall melting temperature of the solder on the SMD pads. Essentially, this enables you to just lift the part right off the PCB.

INCLUDES
- Alcohol pads for cleaning the circuit board after device removal
- A special low-melting-temperature alloy
- Standard no-clean flux
- Application syringe

CHIPQUIK
CURITY
Alcohol Prep
CHIPQUIK

Desoldering Tips

FOR STANDARD THROUGH-HOLE COMPONENTS
- First grasp the component with a pair of needlenose pliers. Heat the pad beneath the lead you intend to extract and pull gently. The lead should come out. Repeat for the other lead.
- If solder fills in behind the lead as you extract it, use a spring-loaded solder sucker to remove the excess solder.

FOR THROUGH-HOLE ICs OR MULTI-PIN PARTS
- Use a solder sucker or desoldering braid to remove excess from the hole before attempting to extract the part.
- You can use a small, flat-tip screwdriver or IC extraction tool to help loosen the device from the holes.
- Be careful to not overheat components, since they can become damaged and may fail during operation.

BEFORE YOU START

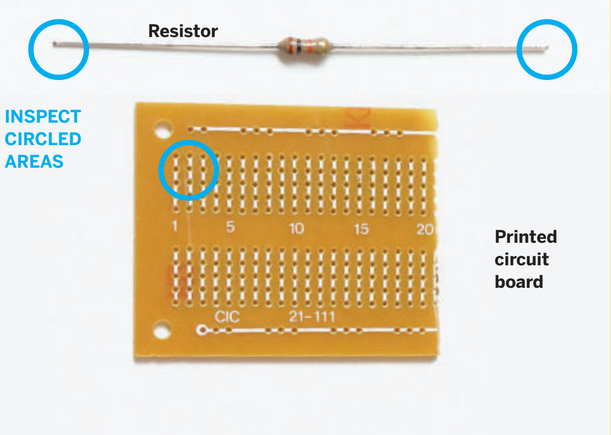

Joe Grand is the president of Grand Idea Studio, Inc. (joe@grandideastudio.com), a product-development and intellectual-property licensing firm. He specializes in embedded system design, computer security research, and inventing new concepts and technologies. He is also a host on The Discovery Channel's *Prototype This*.

SOLDERING

SOLDERING A RESISTOR TO A CIRCUIT BOARD

BEFORE YOU START

Inspect the leads or pins for oxidation. If the metal surface is dull, sand with fine sandpaper until shiny. In addition, use the sandpaper to clean the oxidation and excess solder from the soldering iron tip to ensure maximum heat transfer.

This simple example shows the step-by-step process to solder a through-hole component to a printed circuit board (PCB). I used a piece of prototype PCB and a single resistor.

⚠ DANGER: It's important to consider safety precautions. Improper handling of the soldering iron can lead to burns or other physical injuries. Wear safety goggles and other protective clothing when working with solder tools. With temperatures hovering around 700°F, the tip of the soldering iron, molten solder, and flux can quickly sear through clothing and skin. Keep all soldering equipment away from flammable materials and objects. Be sure to turn off the iron when it's not in use and store it properly in its stand.

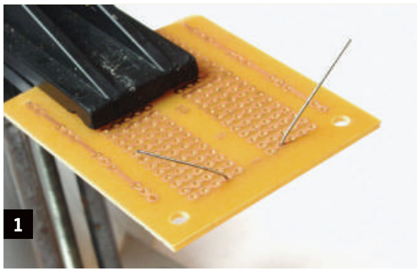
1

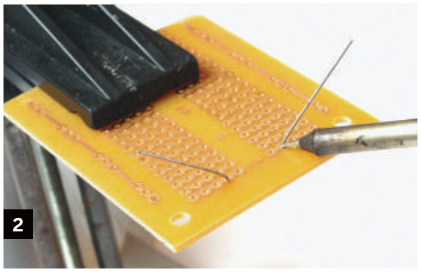
2

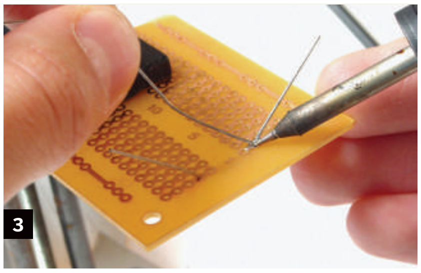
3

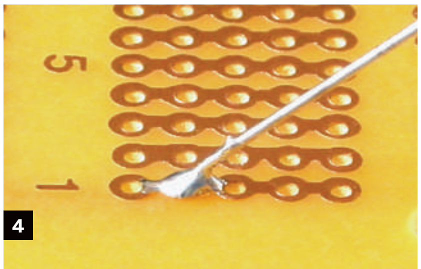
4

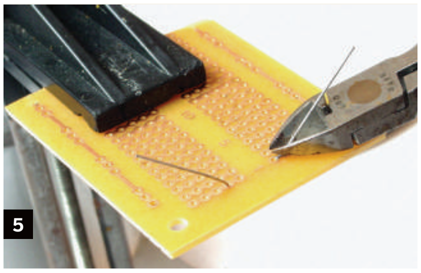
5

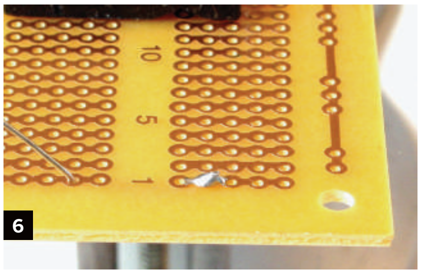
6

1. Bend and insert the component leads into the desired holes on the PCB. Flip the board to the other side. Slightly bend the lead you'll be soldering to prevent the component from falling out when the board is turned upside down.

2. To begin the actual soldering process, allow the tip of your iron to contact both the component lead and the pad on the circuit board for about 1 second before feeding solder to the connection. This will allow the surface to become hot enough for solder to flow smoothly.

3. Next, apply solder sparingly and hold the iron in place until solder has evenly coated the surface. Ensure that the solder flows all around the 2 pieces (component lead and PCB pad) that you're fastening together.

Don't put solder directly onto the hot iron tip before it has made contact with the lead or pad; doing so can cause a cold-solder joint (a common mistake that can prevent your hack from working properly). Soldering is a function of heat, and if the pieces aren't heated uniformly, solder may not spread as desired. A cold-solder joint will loosen over time and can build up corrosion.

4. When it appears that the solder has flowed properly, remove the iron from the area and wait a few seconds for the solder to cool and harden. Do not attempt to move the component during this time.

The solder joint should appear smooth and shiny, resembling the image above. If your solder joint has a dull finish, reheat the connection and add more solder.

5. Once the solder joint is in place, snip the lead to the desired length. Usually, you'll simply cut the remaining portion of the lead that isn't part of the actual solder joint. This prevents any risk of short circuits between leftover component leads on the board.

6. Here's a completed soldering example.

DESOLDERING

SMD REMOVAL WITH CHIP QUIK

BEFORE YOU START

Use a rubbing alcohol pad to remove any residue from the solder pads. Verify that the solder pads are clean and free of cuts or solder jumps before proceeding. Desoldering, or removing a soldered component from a circuit board, is typically trickier than soldering because you can easily damage the device, the circuit board, or surrounding components. For surface mount devices (SMDs) with more than a few pins, the easiest method to remove the part is the Chip Quik SMD Removal Kit, as shown in the following step-by-step example.

PLEASE READ through this example completely before attempting SMD removal on an actual device. When removing the device, be careful not to scratch or damage any of the surrounding components or pull up any PCB traces.

THREE PRIMARY FUNCTIONS OF FLUX

- **» Cleans metal surfaces to assist the flow of *filler metals* (solder) over *base metals* (device pins).**
- **» Assists with heat transfer from heat source (soldering iron) to metal surface (device pins).**
- **» Helps in the removal of *surface metal oxides* (created by oxygen in the air when the metal reaches high temperatures).**

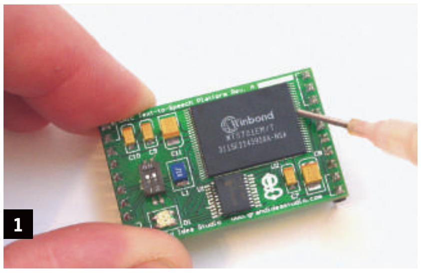
1

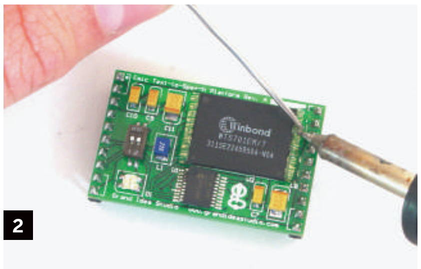
2

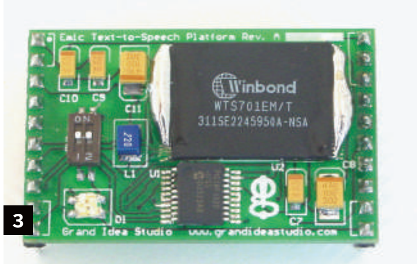
3

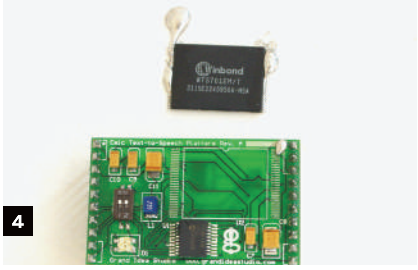
4

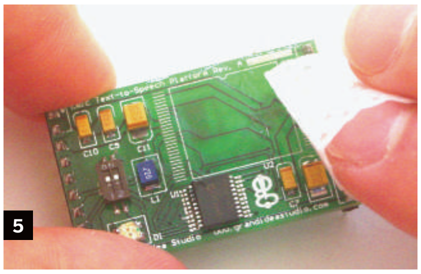
5

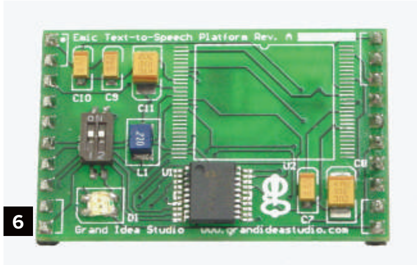
6

1. The first step is to assemble the syringe, which contains the no-clean flux. Simply insert the plunger into the syringe and push down to dispense the compound. The flux should be applied evenly across all the pins on the package you'll be removing. (Flux is a chemical compound used to assist in the soldering or removal of electronic components or other metals.)

2. Once the flux is evenly spread over the pins of the target device, the next step is to apply the special Chip Quik alloy to the device. This step is just like soldering: apply heat to the pins of the device and the alloy at the same time. The alloy has a melting point of approximately 300°F, which is quite low. You shouldn't have to heat the alloy with the soldering iron for very long before it begins to melt. The molten alloy should flow around and under the device pins. Starting at one end of the device, simply heat and apply the alloy. Repeat for the other side(s) of the device.

3. Flux will help ensure a nice flow of the alloy onto the device pins. Make sure the alloy has come in contact with every single pin by gently moving the soldering iron around the edges of the device. Avoid touching nearby components on the PCB with the soldering iron.

4. Now that the alloy has been properly applied to all pins of the device, it's time to remove the device from the board. After making sure that the alloy is still molten by reheating all of it with the soldering iron, gently slide the component off the board. You can use a small, jeweler's flat-tip screwdriver to help with the task. If the device is stuck, reheat the alloy and wiggle the part back and forth to help the alloy flow underneath the pads of the device and loosen the connections.

5. The final step in the desoldering process is to clean the circuit board. This step is important because it will remove any impurities left behind from the Chip Quik kit and get you ready for the next step.

First, use the soldering iron to remove any stray alloy left on the device pads or anywhere else on the circuit board. Next, apply a thin, even layer of flux to all of the pads that the device was just soldered to. Use the included alcohol swab or a flux-remover spray to remove the flux and clean the area.

6. The desoldering process is now complete. The surface-mount device has been removed and the circuit board cleaned.

TECHNOLOGY IN ACTION™

Building Your Own Electronics Lab

A Guide to Setting Up Your Own Gadget Workshop

Dale Wheat

For your convenience Apress has placed some of the front matter material after the index. Please use the Bookmarks and Contents at a Glance links to access them.

friendsof™

Apress®

Contents at a Glance

- About the Author xi
- About the Technical Reviewer xii
- Acknowledgments xiii
- Introduction xiv
- Chapter 1: Planning Your Electronics Workshop 1
- Chapter 2: Building Your Tool Chest 19
- Chapter 3: Components 57
- Chapter 4: A Portable Mini-Lab 87
- Chapter 5: The Cozy Corner Lab 105
- Chapter 6: The Small Group Lab and Classroom 135
- Appendix A: Getting Started with Tool Building 145
- Index 161

Introduction

We're surrounded by electronics these days. Did you ever wonder how it all worked? Do you know just enough about electricity to be dangerous? Would you like to learn more? Then this book is for you.

There are many excellent textbooks on how electricity works. This is not one of them. This book helps you get started learning about electricity and electronics by showing you how to set up your own electronics laboratory. You might have guessed that from the title.

Along the way, several simple exercises are presented to show you some basic concepts about how electricity works in the real world—not just in a lecture hall. By the time you get done reading this short book, you'll have a good idea of where to begin and what kinds of tools and components you might need, and you'll have some excellent advice about how to keep it all organized.

Once you've got your own space set up to perform your electronic experiments, the sky is the limit.

Here's what you have to look forward to in this book, one chapter at a time.

Chapter 1: Planning Your Electronics Workshop

Your very own electronics laboratory is a great place to build, repair, invent, and learn more about electronics. We seem to be surrounded by electronics more and more every day. We can easily become dependent upon them, without even understanding how they work. Having your own electronics lab can help you gain some control over your electronic minions. Here you can learn about their inner mysteries, including how to repair them when possible or, better yet, improve them with your own custom modifications. You can also turn your own ideas into reality by building electronic circuits from scratch.

What will you need to plan your lab? What if you already have a basic understanding of electronics and the beginnings of a workshop at your disposal? This chapter helps get you going in the right direction.

Chapter 2: Building Your Tool Chest

You're going to need some tools in your lab, as well as the skills to use them effectively. This chapter will get you started.

If you're starting from scratch, don't worry. There's not a whole lot you absolutely *must* have to get started. Some basic hand tools and a place to keep them organized is all you will need. You don't even have to spend a lot of money at first. As you progress in your hobby, you will most likely want to add to your tool chest and upgrade some of those tools. Electronics can be the perfect hobby because it can take up (1) all your spare time and (2) all your discretionary income. What more could you ask for?

Chapter 3: Components

You're going to need some components to play with, as well as the knowledge to use them effectively. This chapter will introduce you to some of the bits and pieces that make up modern electric and electronic circuits. You'll also learn a little bit about how to identify components from their appearance and markings, when available.

Once you've got an idea about what these parts do in a circuit, you'll learn a little more about how to measure their electrical properties and put them to use. You'll also be shown what *not* to do, in some select examples.

Chapter 4: A Portable Mini-Lab

Having a portable (or at least (*trans*portable) electronics lab comes in handy in several circumstances. Maybe you don't have a place (yet) for a permanent home for all that equipment. Maybe you need to bounce at a moment's notice, heading off on electronic adventures at the drop of a hat. Or maybe you like to keep everything where you can find it in a hurry, without having to rummage through shelves and boxes, looking for just the right tool or component.

Whatever your motivation for wanting a portable lab at your disposal, this chapter should help you get started.

Chapter 5: The Cozy Corner Lab

You don't need an entire garage or extra office for a functional laboratory for your electronic endeavors. It's certainly nice if these areas are available to you, but you might be surprised at how much you can do in just a small space, if you set it up properly and maintain it with determination.

Chapter 6: The Small Group Lab and Classroom

The first part of this book deals with how to get along with electrons and make them do your bidding. In this chapter, you'll explore a completely different topic: other people, and how to get along with them.

You might not ever figure out how to get *them* to do your bidding, but at least you can keep them from stealing your tools. Maybe.

Appendix A: Getting Started with Tool Building

Once you've become comfortable in your lab and had time to play with a few circuits, you might start to notice that some of the "tools" you've been using are just simple electronic circuits themselves. Good examples are power supplies and meters. These are great tool-building projects because you can see useful results early.

This appendix takes a look at building a couple of simple electronic tools that might be of use in your lab. Hopefully you will develop a better understanding of how these tools work. Ideally, you will progress from the basic question of "Does it work?" to the more involved questions of "How well does it work and what can I do to improve it?"

CHAPTER 1

Planning Your Electronics Workshop

Are you interested in electronics? Would you like to set up your own "electro lab" to conduct amazing experiments, build crazy gizmos, and repair or modify your existing electronics? Then this book is for you!

Your very own electronics laboratory is a great place to build, repair, invent, and learn more about electronics. We seem to be surrounded by electronics more and more every day. We can easily become dependent upon them, without even understanding how they work. Having your own electronics lab can help you gain some control over your electronic minions. Here you can learn about their inner mysteries, including how to repair them when possible or, better yet, improve them with your own custom modifications. You can also turn your own ideas into reality by building electronic circuits from scratch.

What will you need to plan your lab? What if you already have a basic understanding of electronics and the beginnings of a workshop at your disposal? This chapter helps get you going in the right direction.

What to Expect

This book can't teach you everything about electronics, and it doesn't try. That would take several lifetimes. What it can do is introduce you to the tools and the skills you will need to set up your own electronics lab. This will include a very basic introduction to electricity, a little bit of theory, some safety tips, and a whole lot of example projects. It's a great starting place. Where it will lead you is mostly determined by you and what you're wanting to do with electronics.

If you're already familiar with electricity and electronic concepts, there is still plenty of fun and interesting stuff waiting for you to explore. The field of electronics itself is constantly expanding. From the time this book was written to the time that you've finished reading this sentence, many advances will have occurred in both our understanding of electronics as well as the development of new applications and electronic devices. This is one of the many things that makes the study of electronics so interesting and exciting.

Every lab is different. Your lab will reflect not only your immediate electronics goals but also your personality. Feel free to pick and choose from all the ideas presented in this book and add in some of your own. The repair shop and the design studio are necessarily going to be arranged differently. There really is no right or wrong way to go.

Even the best laid plans, so the saying goes, end up somewhere unexpected. When planning your electronics lab for the first time, or for the tenth time, keep in mind that your interests, resources, and reasons for wanting to work on electronics are going to change over time. Please feel free to reinvent yourself and your lab as conditions permit. You might also be forced to reconsider your priorities when other factors in your life exert themselves. That being said, stay tuned for “A Cautionary Tale” at the end of this chapter.

Some configurations are more appropriate for certain endeavors, and nothing replaces the right tool for the job. This book should at least give you an idea of what is involved and get you pointed down the right road.

It Starts with a Plan

Most of the decisions you’ll make when planning your laboratory will depend on what *you’re* interested in doing there. Having a good plan at the beginning is like having a clear map when taking a long journey. It helps you from getting lost and also is handy for figuring out when you’ve arrived.

■ **Note** If you fail to plan, you plan to fail.

Remember, at this point in the game, you don’t have to make any final decisions about what you want to be able to do in your lab. This book is all about illustrating possibilities and giving you the information and advice necessary to branch out into the areas that hold the most interest for *you*.

A Broad Outline

Here’s what you need to get started: a little time, a small amount of space, some basic tools, and a few components. Those are the boring parts. The fun stuff includes your ideas, goals, and inspiration. Combine all that with some fundamental information about how electricity and electronics work, and you’re in business.

A Little Time

Like any good hobby, setting up your electronics lab and conducting experiments there is going to take up some of your time. For some, this is the *main reason* for getting started in electronics and building a lab: a place to escape, unwind, and tinker with ideas, prototypes, and complex systems. Maybe it’s your job. Maybe it’s your passion. Perhaps it’s a bit of both.

You’re going to need time to plan and build your lab. This book will spell out several specific arrangements for various-sized labs, but they are just examples and not hard-and-fast rules. More time will be required for obtaining the proper tools and materials that you will use there. Time will be spent not only *doing* things in your lab, but also in just *thinking* about doing things in your lab. For example, you might build a clever little gizmo that ought to do a simple task, only to find out that it refuses to work. You might need to spend a little bit of time thinking of ways of testing it to find out where the problem lies. This is sometimes called *troubleshooting* or *debugging*.

This is the basis of the scientific method. To help understand how something works, you first form a theory that would explain its behavior. Next you prepare one or more experiments that will demonstrate the accuracy of your theory. The experiments are conducted and their results are analyzed. If the results are in agreement with your theory’s predictions, then your theory is more than likely correct. If not, you may need to modify or extend your theory to accommodate the behavior observed. All of these things take time, which helps explain why we don’t have jet packs or time machines yet.

A Little Space

You don't have to start out with a cavernous underground bunker for your electronics lab. If you already have one, that's great. You might have an extra classroom available, some unused warehouse space, or an empty bench in a garage.

Then again, you might only have a corner of the dining-room table, and then only between mealtimes. That's plenty of room when you're just getting started! Once you learn more about electronics, and especially when you learn more about what it is that *you* want to do with electronics, then you can start looking for a more permanent home for your toys, tools, and spare parts.

The main points that will be stressed throughout this book when planning and working in your space are safety, good lighting, and organization, in that order.

- Your lab needs to be relatively safe for you, and *very* safe for your visitors and neighbors. The "Go Away!" sign on the door does not guarantee that you will be free from curious visitors, pesky and otherwise. People are naturally attracted to the creative and mysterious activities going on in your lab.
- Be prepared to buy, build, or borrow more lighting, because you're going to need it. Proper lighting improves the quality of your doodlings and tinkering enormously. You might think you already have enough light, and unless you were a photographer in a previous life, you don't. (Photographers understand that you can never have too much light.)
- An organized lab is a productive lab. Don't waste time looking for parts or tools when it would have only taken mere moments to put them in their proper places in the first place. A place for everything, and everything in its place. You get to decide where all these tools and bits get to live, so spend some time and do it wisely.

These three tenets will be repeated throughout the book, so you might as well get used to hearing about them.

Basic Tools

You really don't need a lot of tools to get started. There is a tendency to want to be a bit overprepared before tackling any new project, but it really isn't necessary when it comes to planning and building a new electronics lab. This is also true when sorting out a lab that's gone badly out of control.

You're going to be working with wire quite often. Claude Debussy has been quoted as saying that music is the space between the notes. It's the connections between the various electronic components that make the magic happen in an electronic circuit, in the same way that it's the relationships between people that make their lives so interesting.

The connections between all the various imaginable electronic components almost always start out as bits of wire. These connections can then be optimized into copper traces etched on a printed circuit board or even as metallization links in an integrated circuit. You're going to be working with a lot of wires in this hobby, and that usually boils down to two basic hand tools: *wire cutters* and *wire strippers*.

A wire cutter does what it says and says what it does: it cuts wire. A wire stripper is a more specialized tool that helps remove the outer layer of insulation from a wire. Ever more specialized versions of both tools abound. Some tools aim to perform both tasks at once, while never perfecting the art of either.

Resist the temptation to use your teeth for cutting and stripping wire. If you're caught in a situation without proper tools, use your fingernails. They grow back faster than teeth.

■ **Tip** Use the right tool for the job. Knowing which tool is the right tool is half the battle.

If you're performing detailed work on smaller circuits, you might benefit from having some small needle-nosed pliers, a small vise, clip, or helping-hands jig, and possibly a magnifying glass.

A popular and handy skill in the electronic arena is *soldering*. Soldering is the process of making electrical and mechanical connections. This is done by heating the items to be joined together and adding solder. The solder melts and fills in all the gaps between the items being joined. Once cooled, the solder forms a solid, conductive link.

■ **Note** Soldering is similar to welding, with the exception that welding actually melts the parts being joined.

A suitably sized soldering iron and the right kind of solder are indispensible for many projects, including repair, prototyping, and production. Picking up the skill to make quality connections is obtained in the same way as all worthwhile skills: with lots of practice.

Besides these basic fabrication tools, you will also want some *test and measurement equipment*. These are devices used to measure various aspects of electrical circuit behavior. These range from the very simple, such as voltage meters, current meters, and continuity testers, to the very exotic (and sometimes expensive), which include oscilloscopes and various logic, network, and signal analyzers. See the Appendix for information about building your own simple test equipment for your lab.

This is just a quick summary of the kinds of tools you might want in your electronics lab. Chapter 2 goes into much more detail about the tools you will probably need or want in your lab, as well as the skills needed to operate them safely and effectively.

Fundamental Components

As mentioned in the previous section, you're more than likely going to be wiring up various contraptions and rewiring others. This is probably going to involve a lot of wire.

For small voltages and currents, small-gauge insulated copper wire is often used. It's easy to handle, cut, and splice. This can be done with mechanical connectors such as wire-nuts or screw terminals, as well as with soldering.

Back in the day when computers were all connected using fat cables, lots of tiny wires were bundled up into cables and snaked all over a typical desktop computer installation. With the increasing popularity of wireless devices, those fat, juicy cables are lying around, discarded, unwanted, and just waiting to be harvested for their valuable conductors. You can also buy various kinds of wire in all colors and sizes from most hardware stores and from online suppliers.

Now consider the types of projects you'll be working on:

- If you're most interested in building your own lighting, you're going to want some lightbulbs, sockets, and LEDs, or whatever lighting technology appeals most to you. Don't forget that you're also going to need switches, knobs, dimmers, housings, plugs, and *more wire*.

- If you like to work with things that move, you’re going to need motors, relays, solenoids, servos, and fans. You’ll probably be needing some gears, pulleys, chains, belts, shafts, and couplings, along with connectors, switches, controllers, position sensors, and *more wire*.
- If you’re into making lots and lots of sound, you’ll be needing speakers, headphones, microphones, pickups, transducers, transformers, amplifiers, connectors, jacks, plugs, and, you guessed it, *more wire*.
- Are computers and digital devices more to your liking? Be thinking about getting yourself a wide variety of chips, microcontrollers, software, cables, sockets, connectors, and still *more wire*.

Now you’ve got an idea of some of the different components you might want to have on hand, depending on what kind of inspiration happens to strike. You’re going to need a place to keep all these different parts sorted. This will play a big part in the overall planning phase of your lab. Maybe your old fishing tackle box has enough room for your basic hand tools and enough secret compartments to hold an interesting variety of components. On the other end of the spectrum, maybe you need to start thinking about some industrial-grade shelving units or pallet racks. Now is the time to think about all these factors and start making the right kind of plans to help make your lab a successful one.

Again, this is just a quick summary of some of the types of components you might want to work on in your lab. Chapter 3 is going to take you much farther into understanding many of the more common electrical and electronic components that you are likely to encounter in your lab.

Ideas

Admit it: you’ve already got some ideas for cool projects that you’d like to be working on in your own electronics lab. That’s great! You’re well on your way to making those dreams a reality.

The best possible thing you could do with your ideas is to write them down somewhere. It can be as simple as a sketch on a piece of scrap paper or as formal as a complete set of project documentation. The important thing is that you actually get the idea across using either words or pictures in a form that you will be able to revisit in the future. This will be of immense benefit to you, as well as anyone else that might be interested in your ideas.

It’s also a great idea to keep a log as you work on your projects. That’s the perfect place to jot down those ideas as they come popping out of your brain. Be as detailed as you can be. You will thank yourself later.

Even if you don’t already have a backlog of ideas jostling for your attention in the lab, it’s still a good idea to keep a log and explore the things that interest you. Leonardo da Vinci kept detailed notebooks on almost every possible subject. Anything that interested him got written down, sketched, or diagrammed in his notes. Even today, people are still getting inspiration from his doodles (see Figure 1-1).

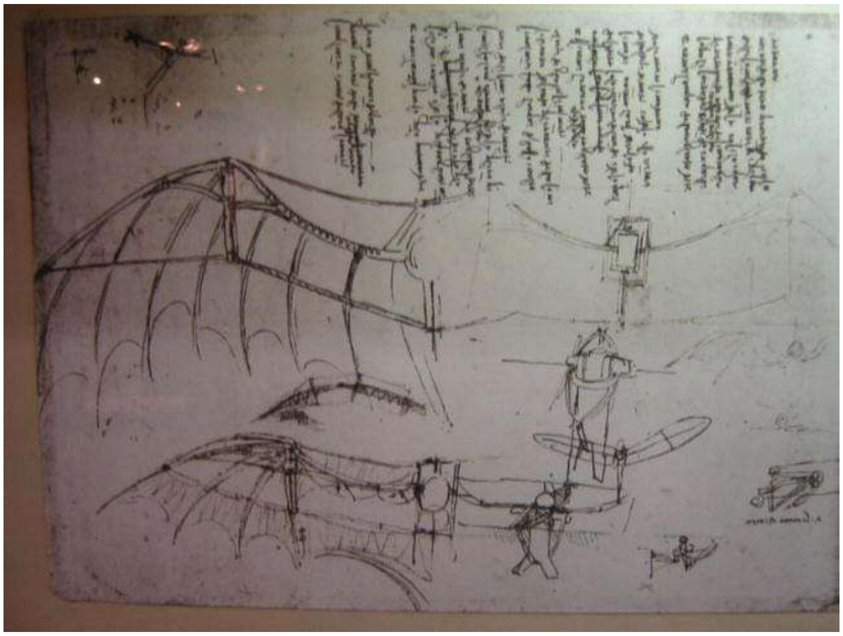

Figure 1-1. *Sketches and notes for a "flying machine" by Leonardo da Vinci. Photograph by Arnaud 25 (public domain), via Wikimedia Commons.*

Writing down your ideas also helps you pick which projects you want to work on first. It just isn't possible to work on all of them at the same time! Narrowing down the list helps get you focused and lets you concentrate all of your talents and energies on the most important projects.

When you're just beginning, the "most important projects" are the ones that provide the most positive reinforcement in the shortest amount of time. Plan a few simple, straightforward projects with a high probability of success at first. You can become more ambitious and take greater risks when you've built up a good reserve of self-confidence and techniques. Remember, the only experiments that fail are the ones from which you don't learn anything. Sometimes learning how *not* to do something is just as important as learning how to do it correctly.

Goals

The best way to turn great ideas into spectacular projects is to have a plan. Set some achievable goals for yourself. When you're first starting, don't be afraid to take those baby steps. You have to crawl before you can teleport.

Project planning is its own art form, and a dark and mysterious art it can be sometimes. It really helps to have clearly defined goals and expectations early in a project. Otherwise, how will you know when you're finished?

Any large, complex problem or project can be broken down into smaller, more manageable subprojects. Unfortunately, this can be carried too far. You don't need to reinvent the wheel every time you start a new project. There's a saying that suggests, "Don't build anything you can buy." The counterpoint to this wisdom is the equally ancient saying, "Don't buy what you can build." You'll have to be the judge as to which one applies best to you and your situation.

Do you have lots of time and very little money? This is good news! This helps simplify the decision-making process enormously. Design and build everything yourself, from scratch. This can include your own tools as well as custom-made components. The trick to this approach is not get lost in all the inevitable details that will arise.

No time to spare, but there's a bit of a budget available? This is likewise good news. With few exceptions, there's someone out there that's willing to part with just what you need for a reasonable price. The trick here is to find them and make that connection.

Inspiration!

Remind yourself what your inspiration is. What initially prompted you to start thinking about building your own electronics lab? Why did you want to get into electronics as a hobby or a career in the first place? Do you even remember?

It really helps to have a firm grasp on what your underlying motives are (or were) when faced with the challenges that regularly present themselves along the way. While it's true that some things just need doing, does it really have to be *you* that does them? Could you farm out some of the work and still feel a sense of satisfaction in the completion of the project? Or do you feel the need to take the do-it-yourself mind frame to its utmost limit?

Whatever your inspiration is, or wherever it came from, do your best to keep it in your sights while pursuing your dreams and goals. It makes it that much more satisfying when you eventually attain them.

How Electricity Works

There are two ways to go about learning how electricity actually works. The first involves just a whole bunch of *fascinating* theoretical information about atoms, electrons, fairies, and dragons. The second way is to just assume that it does indeed work, and spend some time actually putting it to work.

You don't need to know everything about electricity to have fun with it. A passing understanding is sufficient to get you started. As your technical requirements grow more elaborate, you will find yourself going back to revisit some of the basic concepts, until your comprehension is sufficient to carry you forward.

The Theoretical Approach

The prevailing theories about how electricity works center around the orbits of negatively charged subatomic particles called *electrons*. Electrons are quite small, even when compared to individual atoms. Electrons normally whiz around the *nucleus* of the atom, which is usually made up of other subatomic particles called *protons* and *neutrons*. The exception is hydrogen, whose most common isotope has only a single proton and no neutrons. Protons have a positive electrical charge and neutrons have no electrical charge.

When an atom has a balance of protons and electrons, it is said to be electrically neutral. If an atom has more electrons than protons, it has a negative electrical charge. Conversely, if there are more protons than electrons, the atom has a positive charge. An atom that has either a positive or negative electrical charge is called an *ion*.

While flying round and round the nucleus, electrons tend to cluster into well-defined orbits, or *shells*. The innermost shell of an atom can hold as many as 2 electrons. Once this orbit is "full," up to 8 more electrons can be accommodated in the next larger shell. The next shell holds up to 18 electrons, and the one after that can contain up to 32 electrons. After that, the shells get smaller again, stepping back down symmetrically from 18 to 8 to 2.

Each unique combination of electrons and protons represents one of the chemically distinct *elements*. Each element has its own particular chemical characteristics, and cannot be further subdivided by conventional chemical or mechanical means. Toward the end of the *periodic table*, which arranges all the known elements into meaningful rows and columns, you'll start to find some exceptions to the shell symmetry. These heavier elements, however, don't usually turn up in most electronics labs, so we'll not be spending a lot of time there.

Individual atoms can combine into more complex structures called *molecules*. One popular molecule is water, which is made up of two hydrogen atoms and one oxygen atom. Each of the hydrogen atoms has a single electron. The oxygen atom has eight electrons. The first two electrons in the oxygen atom occupy the first shell, leaving six electrons in the next outer shell. That leaves two "open" places in the second shell, which is where the two electrons from the two hydrogen atoms come into play. The electrons are shared by the three atoms, linking them all together and forming a *covalent bond*. Water has vastly different physical properties than either hydrogen or oxygen.

It's also possible to bump an electron out of orbit around a nucleus. When this happens, the atom becomes an ion and is electrically positive. This makes it especially attractive to any loose electrons that happen to be in the vicinity. This is because electrons are negatively charged, and in the case of electrical charges, opposites attract. This movement of electrons is how an electric charge flows through a substance.

Some materials are very good at slinging those electrons around, and as such are called good *conductors* of electrical energy. Examples of good conductors are copper and silver. Some materials resist the flow of electrical charge, and can be used as *insulators*.

An important concept to understand about the flow of electricity is that it is based on the movements of tiny, *negatively charged particles*. This is referred to as the flow of *electron current*, but this term is rarely used. For some reason, the more popular, or conventional, idea of electric current flow is from the more positive point to the more negative point in a circuit. This is referred to as *conventional current* and is the standard used throughout this book.

The fairies and the dragons come into the picture when you start to look any deeper into how the building blocks of matter are actually (or allegedly) constituted. The previous quick summary of how electricity works covers less than 1 percent of the physics needed to understand even how basic atomic principles operate. Even then there are various schools of thought as to how, *exactly*, all that tiny stuff works together to produce rainbows and cousins and moon rocks. A proton, once thought to be an indivisible *elementary particle*, is now thought to be built out of a "down" quark and two "up" quarks. There are also "strange" and "charmed" quarks flitting about, along with an entire pantheon of mythological particles. Go figure.

Now all of that *fascinating* information only scratched the surface of how *electricity* works. Electricity does a lot of work for us. We use it for lighting, heating, motion; just all kinds of things. *Electronics* is where we start to get into some really bizarre physics involving semiconductors, transistors, computer chips, and other miracles of the modern age. Explaining even the basic ideas about how *electronics* works is way beyond the scope of this little book. It's certainly possible to learn and understand it, but for the purposes of setting up a lab and building some example projects, we can just assume that it does indeed work. We can start from there and have some fun with it.

If your curious mind just can't stand *not knowing* how electronics works, the interwebs are your friends, and so are the helpful folks at your local library.

A Practical Approach

An easier way to think about electricity is to skip over *why* it flows and concentrate on *how* it flows. That's what we'll be doing in the remainder of this book, for the most part.

■ **Note** Electricity flows in a circle.

Any continuous flow of electrical current travels in a circle. There are exceptions to this broad generalization, but most of the example circuits encountered in this book will have a definite circular arrangement, if you know how to look for it.

Here are some technical terms that will be used throughout this book. You don't have to memorize any of them right now (there will *not* be a test), but you should at least take a glance at them so you know where to come back when you want to know more.

> *Circuit*: A circuit is anything that forms a continuous loop or circle. The term *circuit* is formed from the same Latin word that means *circle* or *wheel.*
>
> *Amps*: The *intensity* of an electrical current is measured in units called *amperes*, named after André-Marie Ampère. This is usually shortened to *amp* in popular usage. The symbol for amperage in a circuit is *I* (for intensity), while the unit is abbreviated *A*. For example, you could say that the intensity of the electrical current (I) in a circuit was measured to be 3.75A (amperes). Electrical current is measured with an *ammeter*.
>
> *Volts*: The voltage in a circuit is the measure of the difference in electrical *potential*, or how much work it *could* do. The unit of voltage is the volt (V), named after Alessandro Volta. This is sometimes referred to as pressure or tension. For example, a "high-tension wire" means a high-voltage line, not necessarily one that has been pulled really, really tight. Voltage is, not surprisingly, measured with a *voltmeter*.
>
> *Watts*: Electrical power is measured in watts (W), named after James Watt. The easy way to calculate watts, or the amount of energy that is converted into work (either useful or not) is to multiply the voltage across the circuit by the current flowing through the circuit. For example, if a 12V battery is connected to a lightbulb that draws 2A, the bulb converts the electrical energy into both light and heat (mostly heat) to the tune of 24W (12V × 2A = 24W).
>
> *Ohms*: The resistance to the flow of electrical current is measured in *ohms*, named after Georg Ohm. A *resistor* is a simple electronic component that exhibits a fixed resistance to the flow of electrical current. The symbol for resistance is *R* and the unit of measure is the capital Greek letter omega, Ω.

Safety

Make no mistake: electricity can kill you. It bears repeating, as well as emphasis.

■ **Warning** Electricity can kill you.

Working safely with electricity is like working safely with a sharp knife, for example. A good knife is a useful tool. Life would be cumbersome and crude (not to mention short) without good tools. However, when using *any* tools it's important to understand their proper use and operation. The same goes for electricity.

The point here is not to be overdramatic about the dangers of electricity. Billions of people are safely and peacefully coexisting with electricity, largely due to wise safety rules and well-designed electrical appliances. However, there are three major ways that you can be harmed when working around electricity, and it is important that you be aware of them:

- Your body uses tiny amounts of electricity to send nerve impulses to and from your brain. Sending relatively giant shocks of electricity through your body is a good way to stop your heart from beating or your lungs from breathing. You need to keep doing both of those things, so be careful.
- Since your body is not a very good conductor of electricity, it converts the electrical current into heat, which can produce very damaging burns, both on the inside and the outside of your body.
- Additionally, many electrically related injuries are sustained while trying to *get away* (i.e., your body's involuntary reflex when exposed to dangerous voltages).

Yet working safely with electricity is easy. Observe the following rules and you'll be relatively safe, both in your lab and outside of it:

- Always assume a circuit is "live." Electrical devices can retain lethal voltages even when completely disconnected from power.
- Don't work with high voltage (more than 36 volts) unless you're specifically trained. Reading this book does *not* mean you're "specifically trained," even if you read it *twice*.
- Don't short-circuit batteries (i.e., deliberately connect the positive and negative terminals together). Even very small batteries can hold dangerous amounts of power and can produce enough heat to burn you. Even worse, it can start a fire, which is yet another unpleasant way to die.
- Be careful when using metal (conductive) tools around battery terminals so that you don't accidentally short the terminals together or to other circuitry.
- Don't assume other people are aware of these dangers.

Lighten Up!

Now it's time to have a little fun with electricity, don't you think? While it's important to keep safety in mind, it's just as important to be curious and willing to try new things. Don't let the (important) safety rules immobilize your creativity. Use these rules, like any tool, to leverage your enjoyment of this fun and exciting hobby.

Let's start with a simple circuit that is almost guaranteed to lighten things up. Let's take apart a flashlight (or *torch* for you metric types) and see how it works.

First, find a typical household flashlight. The flashlight in Figure 1-2 was purchased for $3, and it included two new D-cell batteries. Put the batteries in the flashlight and make sure it lights up when you turn on the switch. If the bulb is burned out or the batteries are dead, then this experiment is not going to be much fun at all.

DORCY
Frost Brite
ClearView Design
TrueSpot Reflector
Modèle transparent
Réflecteur TrueSpot
DORCY
HEAVY DUTY

Figure 1-2. *A typical household flashlight will be the ~~victim~~ subject of our first electrical experiment.*

Once we've established that the flashlight works as intended, we can begin the experiment.

DISSECTING A FLASHLIGHT

We don't need most of the pieces of the flashlight body. Let's take it apart and have a look inside.

1. Open the flashlight be unscrewing the end with the lightbulb in it.
2. Remove the batteries. Set them aside for the moment. We'll use them in a bit.
3. Remove the plastic retainer that holds the bulb in the reflector.
4. Take out the bulb.
5. We'll need to replace the wiring that was contained within the flashlight's body. Find a short piece of wire 8" to 12" long. A specialized jumper wire with alligator clips on the end was used in the photos for this exercise, but almost any kind of wire will do. See Figure 1-3.

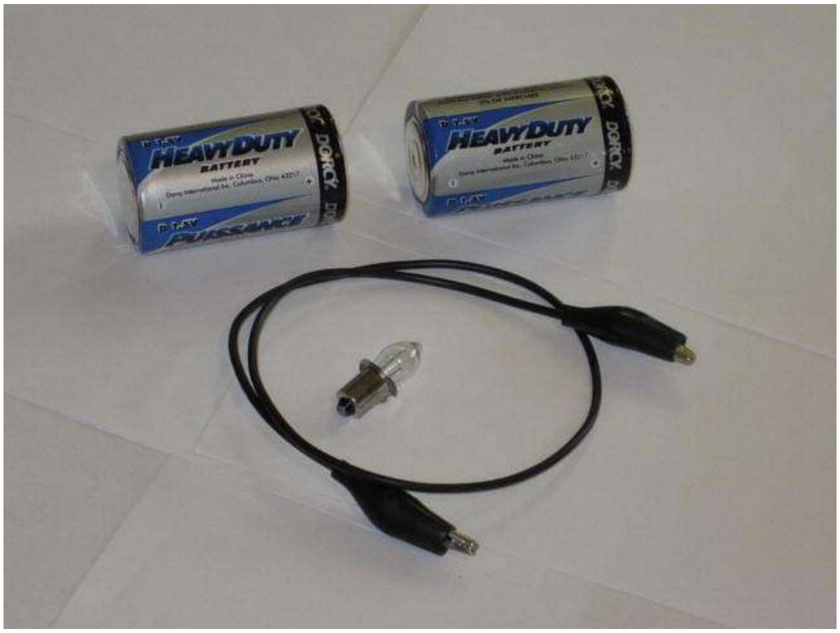

Figure 1-3. *All the pieces you'll need to light up the bulb.*

6. If you're using plain, insulated wire, you'll need to remove a small bit of it from both ends of the wire to expose the *conductor*, or metal wire part inside.
7. This step can be tricky. Try to hold both batteries in one hand, along with the bulb. The batteries need to be pointed in the same direction. If this is too hard to do, try it with just one battery.

8. The positive end of the battery has a small cap on it. Touch the bottom point of the bulb to the top of the battery.
9. Now, using your other hand, connect the wire from the bottom of the battery to the metal base of the bulb.
10. The lightbulb should start to glow. See Figure 1-4.

■ **Caution** If you have a super-bright bulb, it can get hot. Don't burn yourself!

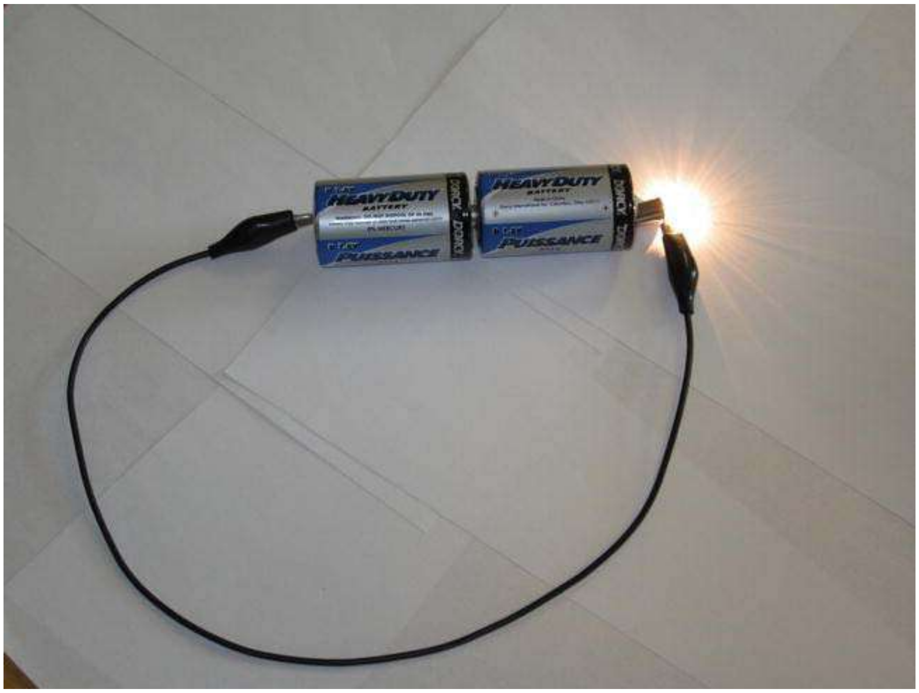

Figure 1-4. *The completed circuit, showing the glowing bulb. Electricity is flowing!*

Several important things are going on in this very simple circuit. Electrical current is flowing, producing useful work (in this case, light). Can you see the "circle"? When you *break* the circuit, the bulb stops glowing. This is exactly how the switch in the flashlight body works. Peek inside the flashlight and see if you can tell how it works. By connecting and disconnecting the wire, you can make the bulb flash. If you only use one battery, the bulb turns on, but is not as bright. Why do you think that happens?

Environmental Impact

You affect your environment every day, and your environment affects you, as well. As you learn more about electronics and all the interesting and wonderful things it can teach you, be mindful of what kind of trail you're leaving behind for others to clean up.

A good way to get a feel for your environmental impact is to take a look at your garbage. It doesn't sound that interesting or fun, but it *will* teach you a lot about yourself and your habits. What are you throwing away? Could you have found either another use for it, or perhaps another home? Is it going to be a nuisance to someone else, or even worse, a hazard?

A good example is the use of "disposable" batteries. We used batteries (technically *cells*, but that distinction will be spelled out in more detail in Chapter 3) in our very first experiment. We buy new batteries and plug them into our devices, and when they run out of power we take them out and throw them away. Where do they go? How do they then interact with the environment? They don't just "go away."

Disposable batteries are very convenient. They are generally inexpensive, compact, self-contained, and reliable. This *convenience*, however, comes with a price all its own. Consider using *rechargeable* batteries whenever possible in your lab. They cost a little more than their disposable counterparts, but can easily and cheaply be refilled with electrical goodness over and over again.

You should also be aware of any hazardous materials that you use or store in your lab. For example, lead is a very common material used in many electrical devices, even though its use is being reduced worldwide. It is found in older-technology rechargeable batteries, solder, and printed circuit boards.

Some of the chemicals used to etch printed circuit boards can be dangerous if not used, stored, and disposed of properly. Take the time to understand what, exactly, you're working with when you're conducting your experiments in your laboratory.

Just like the captain of a seagoing vessel, you're in charge of what happens in *your* lab. With this authority come important responsibilities that rest squarely upon your shoulders. Please take these responsibilities seriously.

Budgeting

How much is all this stuff going to cost? A lot of it depends on how creative you want to be in stocking your lab with tools and parts.

If your budget is small, then start small. If your budget is nonexistent, then you'll just have to start even smaller. You should generally approach acquisitions for your lab under one of these three headings:

- Buying new
- Finding used
- Harvesting or recycling

Buying New

Buying everything new is obviously the most expensive approach, at least from the *money* side of it. What a lot of people fail to take into account is that their *time* is also a very valuable commodity. Try to balance the one with the other, as befits your particular situation.

Except for the folks in the most *dreadful* of hurries, it always makes sense to shop around for the tools, equipment, and components that you are going to need in your lab. You'll find that a lot of "professional-grade" equipment carries an enormous price premium compared with "consumer-" or "hobby-grade" alternatives. This is generally to be blamed on the fact that there is a correspondingly enormous amount of money to be made in the field of electronics.

New components are almost always more desirable to use, especially when you're looking at a production environment and you need to make sure everything produced is of the highest possible quality and reliability. New, unused components are also more appropriate when conducting complex experiments, as this helps to reduce, if not eliminate, many variables in the testing and analysis phases. You don't have to worry that someone (probably you) might have "borrowed" some parts for a quick fix or some other temporary usage and then surreptitiously replaced them back in inventory, having possibly compromised them in some unseen way.

Finding Used

A prudent way to fill up a new lab, especially when first starting out, is to look for bargains on used tools and components. This is especially true of the fixtures, shelves, cabinets, benches, and other furniture that belongs in a lab.

Tools wear out. Tools are replaced. This is a fact of life when you're in the business of working with tools. The good news is that "one man's trash is another man's treasure." If we consider the possibility of women participating, we get three more potential combinations, all of which can end up in a win-win situation.

Many used tools can be retrofitted and placed back into service in your lab. On the other hand, there are a lot of tools that are simply used up and can not feasibly be brought back into useful service. Either their repair is cost prohibitive or they are no longer supported by their original manufacturers.

Another factor working in favor of the patient and observant collector of discarded tools is the inevitable march of progress. What was shiny and top-of-the-line a year ago might be intolerable or insufficient next year, from the standpoint of the original purchaser. Many bargain opportunities arise from this simple understanding.

This can sometimes be applied to the purchase of components that have never been used, but have been sitting somewhere waiting for their chance. A commonly used phrase to describe these parts, which are often sold at a discount, is *new old stock (NOS)*. Sometimes these parts are sold this way because of overstock or because a manufacturing run was cancelled or cut short.

Harvesting or Recycling

There's an absurdly large amount of perfectly good equipment, tooling, fixtures, and components sitting out there, looking for a new home. A lot of companies, as well as individuals, either outgrow their present fittings or turn toward other interests and endeavors. This can be an excellent opportunity to acquire the furnishings you need to populate your lab.

On the other hand, please don't go crazy. You don't have to pick up every discarded appliance or chair you see sitting by the roadside. Why pay rent for something you can easily and cheaply obtain when you actually need it? Are you really wanting to pay the heating and air-conditioning bill for that box full of odd-sized capacitors that are probably past their prime? Don't fall into the trap of the *false economy* when it comes to bargain-hunting for your lab. The most important ingredient in your lab is *you*. Leave a little room to move around a bit.

Another factor to keep in mind is *shelf life*. Not all parts (or tools) last forever, especially when not being used. Batteries are the worst. You can prolong the shelf life of nonrechargeable batteries by storing them in your refrigerator, as this slows down the electrochemical processes within them that produce the electricity. Even solar panels will wear out with either constant use or constant disuse, although this can take years to happen.

Some Example Workshops

Here's a little peek into some actual electronic laboratories, each with its own personality and history. Each one grew from something small into something, well, not so small.

A Fresh Start

Harold Timmis is an electronic engineer, software developer, and author. His curiosity about how things worked started as a small child, which led him from taking apart everything in his parents' house to studying engineering at the Florida Institute of Technology.

Harold keeps a lab in his home where he can work on gadgets in his spare time. Harold recently moved, and used the opportunity to "clean house" and set up everything in his lab exactly the way he wanted things.

He took this photo of his new lab after working on a couple of projects (see Figure 1-5). Note that everything is within easy reach when sitting at the desk. It's still quite tidy, as you can see.

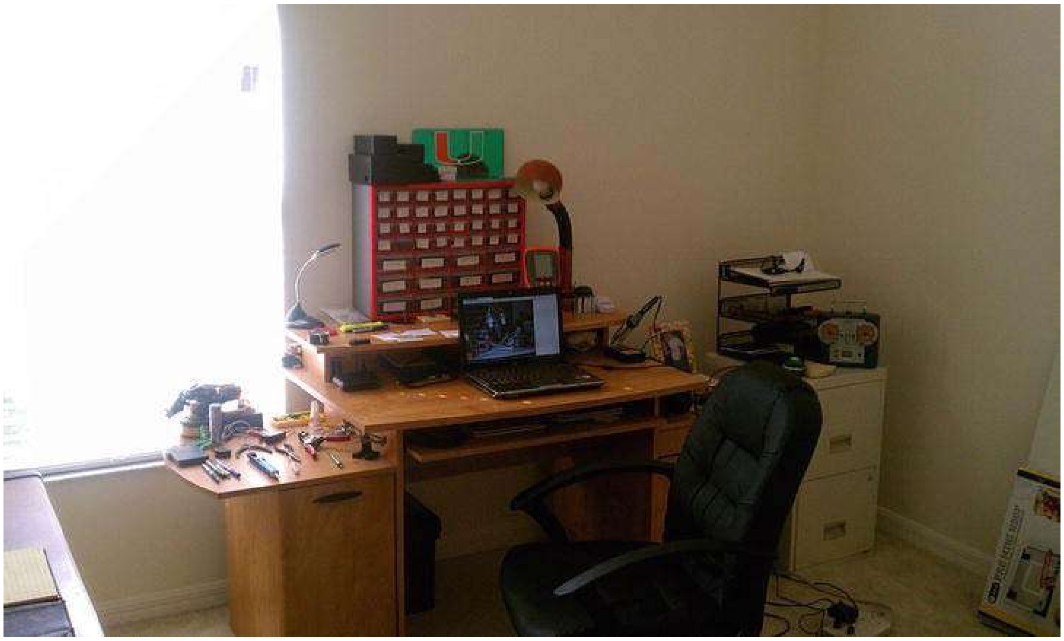

Figure 1-5. *The reorganized home lab of Harold Timmis (photo by Harold Timmis)*

The Robot Clubhouse

The Dallas Personal Robotics Group (`http://dprg.org`) is the world's oldest personal robotics group known to man. From 2002 to 2009, Mike Dodson of Modern Assemblies donated the use of a portion of one of his warehouses to DPRG. This gave DPRG a permanent home for meetings, contests, and social gatherings.

Robot builders enjoyed a spacious and comfortable area for working on robot projects, among other pursuits. Over the years, members contributed time, equipment, and lots of hard work to improving the space. The photo in Figure 1-6 shows one end of the available space, after an extensive remodeling. Several workstations along the wall are in the process of being set up. That good-looking fellow seated at the table is the author.

When Mike retired in 2009, DPRG began a search for another home.

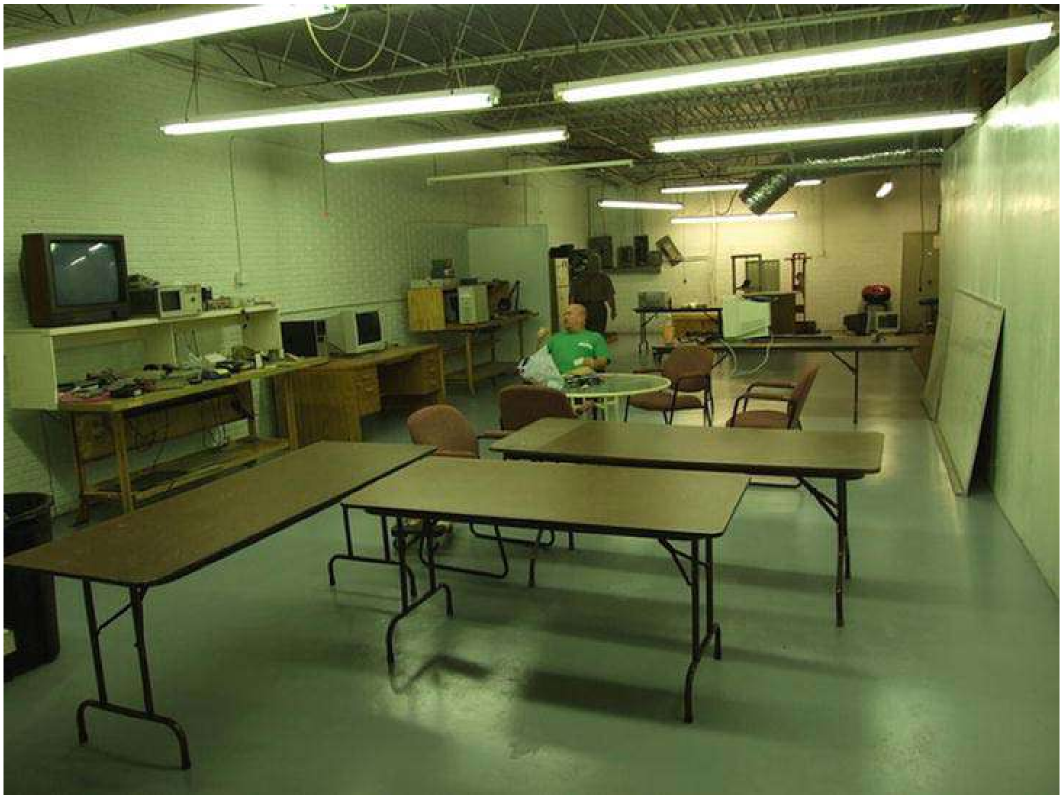

Figure 1-6. *DPRG's world headquarters from 2002 to 2009 (photo by R. Steve Rainwater)*

A Cautionary Tale

It's relatively easy to go from a well-designed and effective lab to something, let's say, less desirable. This doesn't happen overnight, but it can happen. Let this be a warning to you about how even the best of intentions can still produce embarrassing and unproductive confusion (see Figure 1-7).

Unfinished projects and leftover parts and tools compete for the limited space available in this photo. Not only is a lot of time wasted in searching for both components and tools, but the limited space restricts the scale of projects that can be entertained.

Don't let this happen to you.

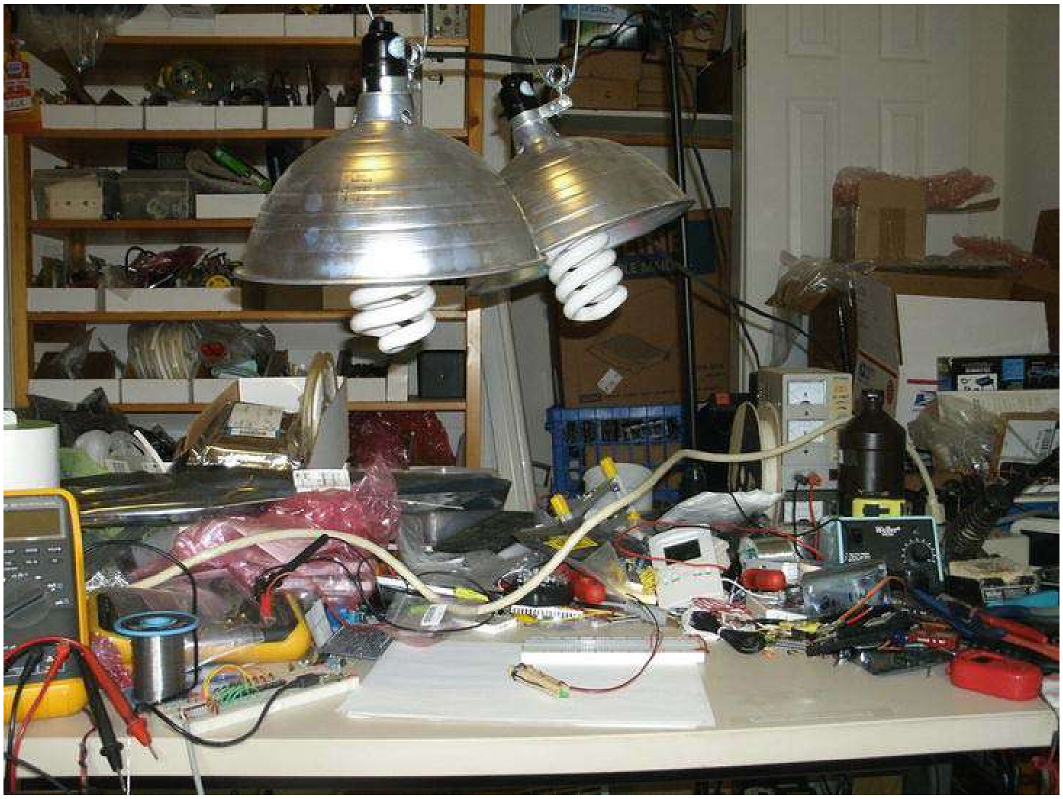

Figure 1-7. *It's not safe. It's not pretty. It's not even properly "eccentric." You have been warned.*

Summary

So now you know a little something about electricity and electronics. Hopefully this chapter has covered your very basic questions about how all this magic stuff actually works.

Now you're ready to start poking around some circuits. You're going to need the right tools for the job. Chapter 2 will get you started. You'll learn about some basic hand tools that are very useful in the lab, as well as some fairly high-tech machines that you might require in the future.

CHAPTER 2

Building Your Tool Chest

You're going to need some tools in your lab, as well as the skills to use them effectively. This chapter will get you started. If you're starting from scratch, don't worry. There's not a whole lot you absolutely *must* have to get started. Some basic hand tools and a place to keep them organized is all you will need. You don't even have to spend a lot of money at first. As you progress in your hobby, you will most likely want to add to your tool chest and upgrade some of those tools. Electronics can be the perfect hobby because it can take up (1) all your spare time and (2) all of your discretionary income. What more could you ask?

The very basic hand tools, such as screwdrivers and tweezers, are mostly self-evident in their usage. Try not to poke yourself in the eye. Where some helpful advice is appropriate, it will be given.

Here's an excellent example of a simple safety rule that will absolutely improve and extend your life: when using any kind of cutting tool, always direct the cutting motion away from yourself. For example, you might receive a package in the mail, and get all excited about opening it. You then discover it's sealed up with indestructible packing tape and can't simply be ripped open with your bare hands. Use a small knife or box cutter and make a single cut at a time, starting from the point closest to you on the package and cutting outward. Using this method, should the knife slip for any reason, the blade will travel away from you. This allows your blood and other bits to stay inside you, where they can do the most good. Never cut toward yourself.

If you've already got some tools (or even a *lot* of tools), it never hurts to review their proper usage. Also, if you're suffering from an *overabundance* of tools, to the point where you can no longer keep them organized, this chapter might help you to prioritize them, at least as far as permanent residency in your lab is concerned.

■ **Tip** Every tool in your shop must *earn* its place there. It must *deserve* to be there. *No exceptions*.

The Most Important Tool Ever

The most important tool you'll ever use is your mind. Every other tool you will ever work with depends on the correct and proper functioning of your mind. If you're careful with it, you won't misplace it, neglect it, damage it, or use it for the wrong job. Your mind is a wonderful asset when cared for properly, and a terrible liability when ill-treated.

Your mind is in many ways like a muscle. The more you use it, the stronger it gets. Learning new things does not "push out" the old things you've already learned. Your brain is capable of forging new neural pathways when challenged, and there's plenty of extra capacity in that old noggin of yours to handle whatever may arise.

■ **Tip** Try to learn something new every day.

Take care of that brain of yours. Just like you wouldn't use a dull drill bit or a rusty saw, don't try to work on electrical or electronic projects when your mind is not in the right place. Ideally, you should never try to get any work done in your lab when you are sleepy, distracted, angry, upset, medicated, or overly tired. Deal with whatever problems you're facing in a responsible manner and then get back to work in the lab.

- Are you sleepy? Take a nap or head off to bed for the night. Really bad mistakes are made when the brain does not get enough sleep. Sleep deprivation is notorious for being obvious (and obnoxious!) to everyone else except you. "Really bad mistakes" made while working with electrical or electronic projects can produce very hazardous conditions for you and everyone around you, such as fire, damaged equipment, injuries, and even death.

- Are you angry or upset? Then your mind is not going to be adequately focused on what you need to do in the lab. "Fools! I'll destroy them all!" are famous last words *for a reason*. Give yourself some time to calm down from whatever has angered or upset you. Also, it's OK for you to give yourself permission to admit that you are angry or upset about something in the first place. Ignoring it probably won't make it go away, and probably won't make your feelings about it go away, either. Walk away, count to ten, or whatever works for you.

- Feeling distracted? Then now is *not* the time to be conducting electronic experiments in the lab. You need to be *present* in the lab, both bodily and mentally. There are just too many ways to make simple blunders that have far-reaching and sometimes overly expensive consequences.

- Have your cognitive functions been compromised? This can happen from taking certain medicines or even recreational intoxicants. If the little bottle says, "Don't operate machinery or drive an automobile," then you probably shouldn't be making important decisions or working with dangerous circuits in your lab, either. Wait until you're clear-headed enough to proceed with your plan for world domination, adding blinking eyes to a plush doll, or what have you.

The worst thing about cognitive impairment is that the first thing that gets impaired is the ability to judge the level of your cognitive impairment!

Care and Feeding of Your Brain

Here are a few tips to keep your brain (and the rest of you) in good working order:

- Stay hydrated. Drink more water than you think you need. This might result in more trips to the bathroom, but this is much more desirable than the alternative. A good flow of water through your systems helps your body feed and repair itself, as well as aiding in the elimination of toxins. Drink more water.
- Avoid excess consumption of coffee, tea, sodas, and especially "energy" drinks. These contain ingredients that stress both your body and your mental capacity. They do not contain energy, but trick your body into going into overdrive by triggering a stress reaction.
- Take frequent breaks. Get up and walk around, if possible. Raise your arms over your head and hold them there as you count to ten. This redirects blood flow and helps restore a balance within your peripherals.

All Those Other Bits, Too

Just as your mind is the most important tool you'll ever use, the rest of your bodily bits and personal pieces are vital and important in your work in the lab. For the most part, there are no spare parts laying around for things like your eyes, your ears, your hands, or your skin.

Should you wear eye protection in the lab? Yes, you should. Are you always going to do so? Probably not. It's true, it can be inconvenient or perhaps a tad uncomfortable. Consider the cost, however. Even the tiniest scratch on your eyeball can take weeks to heal and leave you susceptible to infection or vision loss. It's just not worth it to be lazy. Ask the next one-eyed person you see. If you're a bit uncomfortable asking one-eyed people, "Hey, what's with the one eye?" then perhaps you can just try to imagine them advocating or encouraging you along the same path they took, whether intentionally or accidentally. They won't do it.

Invest in quality eye protection that is comfortable and properly sized for you. Provide eye protection for visitors to your lab as well, if they will be present when soldering, welding, machine tooling, or wire clipping is being performed. Then invest in the habit of wearing your eye protection whenever it is appropriate.

■ **Note** Don't pretend that your prescription glasses with polycarbonate lenses are "safety glasses." They aren't. They do not protect either your eyes or your face from impact or objects entering from the side, as proper safety glasses should. If you usually wear glasses, use safety glasses that comfortably accommodate your glasses when worn.

Will you be working with really loud noises? These could be from powerful audio amplifiers ("it goes to 11") or power tools. Hearing loss from exposure to excessively loud noises is cumulative. This means it slowly (or suddenly) chips away at your hearing, bit by bit. Why do you think old people turn up the TV so loud?

Hearing protection is both affordable and comfortable. For temporary protection, you can use disposable earplugs. If you need protection all day, consider earmuffs designed for hearing protection. Don't let "Hey, y'all, check this out" be the last thing you ever hear.

It's hard to work with hand tools with no hands. It's possible; just really hard. Take care of your hands. There are a hundred ways to burn, tear, cut, or otherwise mangle those mitten-fittings of yours. Have a variety of work gloves on hand to handle the different needs of your lab. The choice will be ultimately determined by the particular hazards present in your lab. Are you working with chemicals as well as electronics? How about sharp edges or glassware? Normally, you won't need to wear gloves for soldering small electronics, but if you start working with really large-scale projects, such as sculptures or plumbing, you might.

Make sure you are still "handy" when wearing the proper gloves for the job. Make sure the gloves fit you properly and allow you the adequate dexterity you need.

■ **Tip** Be safe. Don't be lazy.

Basic Hand Tools

Assuming that you're going to start small and build toward bigger things, you should also start small with some small hand tools. These tools will make it much easier to work on small parts and wires.

Good hand tools increase your effective strength and improve the accuracy and precision of your motions. They allow you to focus a great deal of your strength on exactly the part that needs it, without wasting a lot of effort. Good hand tools also insulate you from sharp, hot, or otherwise unpleasant objects when you work on them.

Take care of your hand tools. Put them up when you're done with them. Make sure they are clean and ready to be used again the next time you need them. Few things are more frustrating than attempting to start a quick repair job and finding out that you need to spend twice as much time preparing your tools for use.

Have some sort of methodical plan or reasoning involved with the storage of your tools. They need a safe and comfortable home, just like you do. You need to be able to find them quickly when you want them. Try out different organizational strategies until you find one that works for you and your lab. This is an easy task when you have just a few tools to organize. Once you've accumulated a large collection of hand tools, this is no longer optional. You *must* know what tools are available and where they are located.

Frequently used tools should be within easy reach of your primary workstation. If your tools need any sort of adjustment or setup before use, make sure you've got all the proper bits handy. For example, a drill press will often have a chuck key used to tighten and loosen the chuck that holds a drill bit in place. That key needs to be in the immediate vicinity of the drill press, as well as easy to locate. Don't spend a lot of time needlessly hunting for tools. Decide for yourself on some safe and practical policies for where you keep your tools.

■ **Tip** Don't waste time guessing where your tools are located. *Decide* where they are located.

Just like everything else, hand tools have a limited useful lifetime. While some tools can easily outlast you, it's up to you to keep track of exactly how much utility you can continue to expect from your tools. This doesn't mean that you have to trash your hand tools once they develop some tiny flaw. If it's beyond your capabilities to repair the tool, then consider replacing it and reassigning it to the second string or backup tool chest. Having a spare tool, even a slightly imperfect one, on hand when you need it can make the difference between success and failure in a project.

Bear in mind, however, that some tools are simply too dangerous to keep around once they've outlived their useful life span. A dull cutting tool that cannot be reasonably resharpened is a good example of a tool that has lost its honored place among your other tools. Thank it for its service, wipe away that little tear, and then figure out a good way to recycle any of its bits that are still functional.

Wire Cutters and Wire Strippers

As you might have picked up from Chapter 1, you can spend a lot of time working with wire in this hobby. Understanding a little bit about how wire is made and how it functions in a circuit will aid you in your experiments.

How Wire is Made—The Very Short Version

Most wire is made of copper or a copper alloy. Some wire used for electrical wiring in houses is made of aluminum. The wire itself starts out as a fairly thick strand and is stretched and pulled—a process called *drawing*—until the desired diameter is obtained. This strand is then wrapped up on spools or bobbins.

- Some wire is made of a single strand of copper (or aluminum). This is called *solid-core* or *single-strand* wire. It is a single strand of wire, usually circular in cross-section.
- Multiple, thinner strands of wire are often bundled or twisted together to make larger-diameter wire. Stranded wire is much more flexible than solid-core wire, but is more expensive to produce because of the extra manufacturing steps required. Because stranded wire has a larger surface area per unit of length, it is more susceptible to environmental hazards such as corrosion.

Most electrical wire is covered with an insulating jacket, usually made of plastic or rubber. This sheathing must be removed to make electrical connections. There is sometimes more than one layer of insulation on a wire, depending on what its final application is destined to be.

More information about wire as a *component* will be presented in Chapter 3. For now, the emphasis is on the characteristics of wire that help determine the best tools to use when working with wire.

Selecting the Correct Tool

Both copper and the aluminum alloys used for electrical wiring are fairly soft metals and can be easily cut or formed with simple hand tools. Wire *cutters* are specialized tools made to make a clean cut through the wire's strand or stands. Wire *strippers* are another kind of specialized tool used to remove the insulation from a wire.

Figure 2-1 shows two rows of commonly available wire cutters.

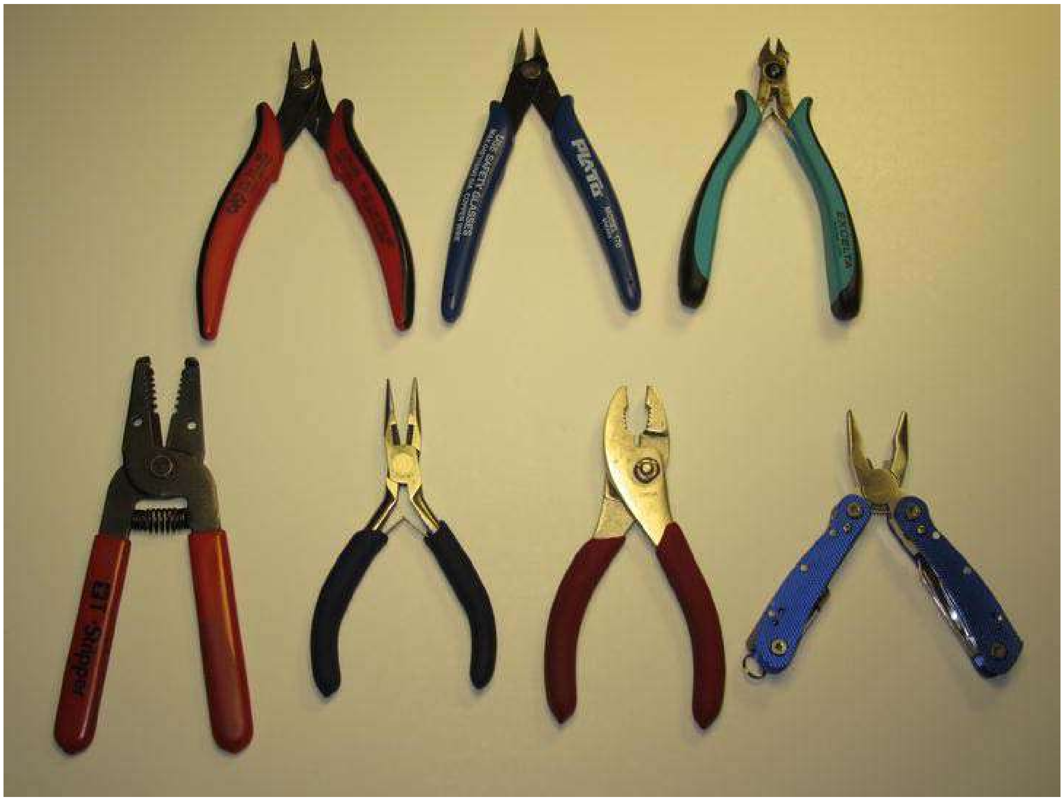

Figure 2-1. *Commonly available wire-cutting tools. The top row shows three styles of diagonal cutters, which are the preferred tools for cutting small-gauge copper wire. The bottom row shows some tools that offer wire cutting as a secondary facility, and should only be used when proper wire cutters are not available.*

The top row of cutting implements in Figure 2-1 are various forms of what are known as diagonal cutters, side cutters, or flush cutters. These cutters have beveled cutting edges that meet in the middle of the wire being cut. One side of the cutters is flat, or *flush*, allowing wires to be cut with some precision. These are the preferred cutters for small-gauge copper wire, as well as for clipping excess component leads from a printed circuit board (PCB). See Figure 2-2.

Compare the cutting action of the side cutters to the other, less optimized wire cutters in the second row of cutters in Figure 2-1. These cutters either use a shearing action to cut the wire or include a flush-cutting area as a secondary feature.

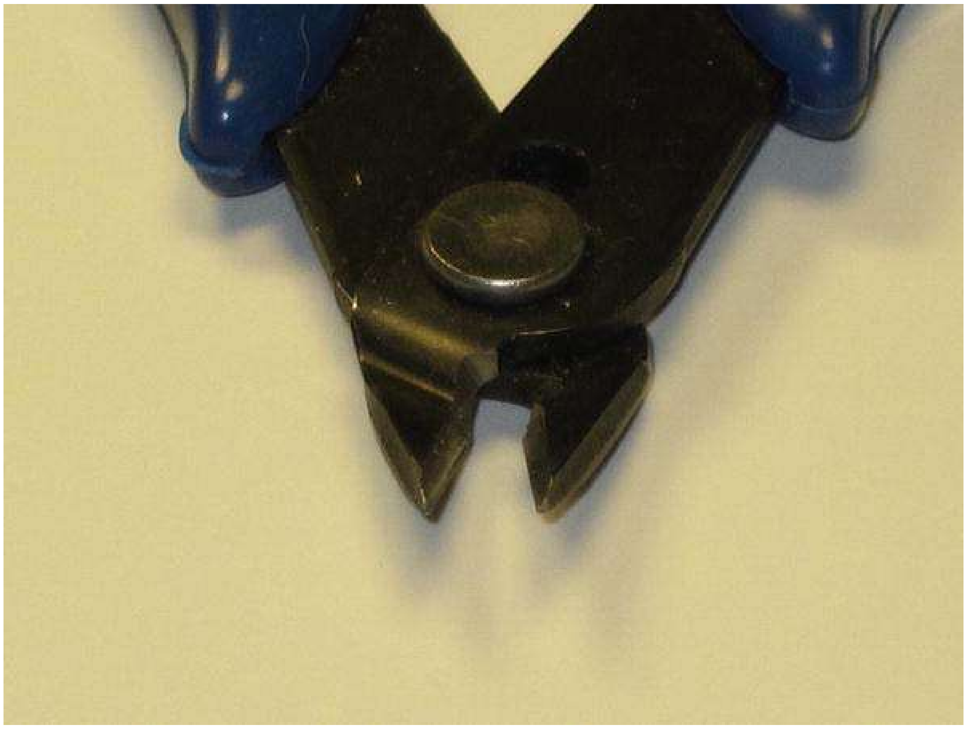

Figure 2-2. *A close-up look at the business end of a pair of side cutters. When the two cutting edges meet in the middle, the bottom side is flat, or flush, allowing the user to get quite close to where a wire or component lead emerges from a solder joint.*

Use caution when clipping component leads or small wires with any wire cutters. The cut lead can easily fly away at sufficient velocity to poke you right in the eye or other sensitive area. If possible, hold *both* ends of a wire being cut or direct the cut leads *away* from your face.

■ **Caution** Use eye protection when clipping leads. Don't say you weren't warned about this hazard.

Wire Strippers

Wire strippers are tools that allow you to easily strip the nonconductive insulation from wires. You can try chewing it off with your remaining teeth, but this is not recommended.

Wire strippers usually consist of a pair of hinged blades with some sort of machined notch. The notch allows the blades to cut the insulation around the wire without cutting or nicking the underlying conductor. A good example is shown in Figure 2-3.

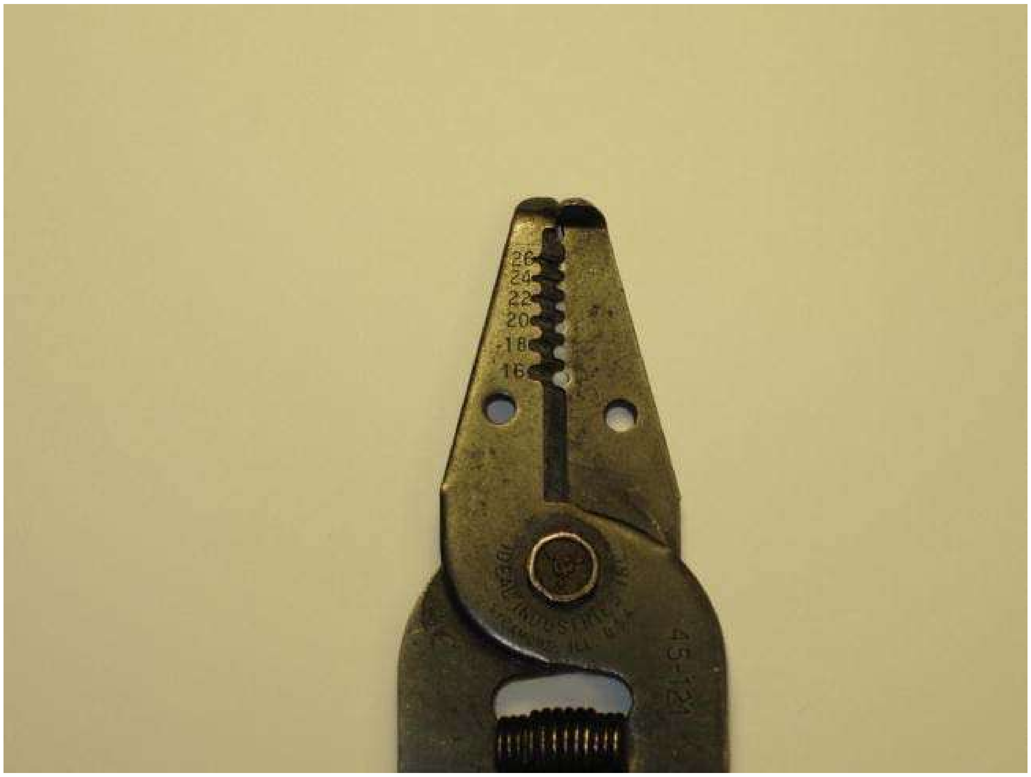

Figure 2-3. *The popular T-Stripper combination wire cutter and stripper tool from Ideal Industries has been around for over half a century. Note the wire gauge sizes stamped into the face of the cutter blade. New models feature laser-engraved text.*

The wire strippers in Figure 2-3 can easily strip the insulation from wires with a conductor diameter, or *gauge*, between 16 (0.0508" or 1.291 mm) and 26 (0.0159" or 0.405 mm). The higher-numbered wire gauges represent *smaller*-diameter wires. Once upon a time, the gauge number represented the number of *drawing* operations that were performed on a wire when being manufactured. The more times it was drawn through the processing equipment, the thinner the wire became. Today, the wire-manufacturing industry in the United States and Canada has adopted the American Wire Gauge standard (AWG). See Table 2-1.

Table 2-1. *The AWG. A more complete listing is available at http://en.wikipedia.org/wiki/American_wire_gauge, from which this partial list was obtained.*

AWG	Diameter Area			Copper Resistance	
	Inch mm		mm²	Ω/km mΩ/m	Ω/1000 ft mΩ/ft
0000 (4/0) 0.46	00	11.684 107		0.1608	0.04901
000 (3/0)	0.4096	10.404	85.0 0.20	28	0.06180
00 (2/0)	0.3648	9.266	67.4 0.25	57	0.07793
0 (1/0)	0.3249	8.252	53.5 0.32	24	0.09827
1 0.28	93	7.348	42.4 0.40	66	0.1239
2 0.25	76	6.544	33.6 0.51	27	0.1563
3 0.22	94	5.827	26.7 0.64	65	0.1970
4 0.20	43	5.189	21.2 0.81	52	0.2485
5 0.18	19	4.621	16.8 1.02	8	0.3133
6 0.16	20	4.115	13.3 1.29	6	0.3951
7 0.14	43	3.665	10.5 1.63	4	0.4982
8 0.12	85	3.264	8.37 2.06	1	0.6282
10 0.10	19	2.588	5.26 3.27	7	0.9989
12 0.08	08	2.053	3.31 5.21	1	1.588
14 0.06	41	1.628	2.08 8.28	6	2.525
16 0.05	08	1.291	1.31 13.1	7	4.016
18 0.04	03	1.024	0.823 20.9	5	6.385
20 0.03	20	0.812	0.518 33.3	1	10.15

22 0.02	53	0.644	0.326 52	.9 6	16.14
24 0.02	01	0.511	0.205 84	.2 2	25.67
26 0.01	59	0.405	0.129	133.9	40.81
28	0.0126 0	.32 1	0.0810 2	12. 9	64.90
30	0.0100 0	.25 5	0.0509 3	38. 6	103.2

We'll return to this table again in Chapter 3 when the electrical properties of wire are discussed. For now, note that as the wire gauge number goes *up*, the wire's physical diameter goes *down*. Similarly, as a conductor's physical size goes down, its ability to carry current also goes down, and its electrical *resistance* goes up.

■ **Note** Other standards exist throughout the world. Steel-wire manufacturers use a different scale, as their interest is more on the tensile strength of the wire, not its electrical properties.

Note that the AWG is for single-strand copper wire. Equivalent gauges in stranded wire are physically a bit larger, although they have the same effective cross-sectional area and can carry the same amount of current.

When using wire strippers, keep in mind that your goal is to cut and remove only the nonconductive insulation from the wire, leaving the internal conductor intact. Even a small nick can result in a weak point in the wire, both electrically and mechanically.

Practice Makes Perfect

The best way to attain proficiency with a new tool is with practice—lots and lots of practice. This is true for working with your hands as well as with that big brain of yours. Ideally, any task worth doing will involve at least a little bit of both manual dexterity and focused brain power.

Cutting and stripping wires is an excellent example. For your first practice assignment, spend some time turning long, beautiful lengths of wire into little bits of wire with stripped ends.

CUTTING AND STRIPPING WIRE

For this exercise, you're going to need the following:

- Wire cutters
- Wire strippers
- Much wire
- A comfortable, well-lit location

Cut the Wire into Segments

Don't watch TV or be otherwise distracted. Pay attention to what you are doing.

Begin by cutting the wire into various-sized segments from 2" to 6" in length. Accuracy in length is not critical at this point. You need to get a good feel for your wire cutters and how much force is needed to cut the wire, without cutting yourself or others.

Collect the cut wire segments into a bowl or small container for the next step.

Strip the Wire Segments

Using your wire strippers, remove about 1/4" (6 mm) of insulation from both ends of each wire. Try not to cut or nick the wire core. The insulation that you remove can either be discarded or collected for use in some sort of art or craft project.

Once you've completed the cutting and stripping portion of this exercise, sort the wires according to their length, or their color if you used different colors of wire. Compare the quality and consistency of your work by looking at the first few wires and the last ones you did. Do you see any improvements in your work?

If you spend about an hour working on your wire-cutting and wire-stripping technique, you will perhaps master a skill that will prove useful for the rest of your life, as well as prepare a big pile of wires that you can use in many of the sample projects illustrated in this book.

Pliers and Tweezers

Every day, the size of the "typical" electronic component gets smaller and smaller. This miniaturization process has brought about a world where you can carry a phenomenally powerful computer in your pocket, and sometimes even make phone calls with it. The individual transistors, resistors, capacitors, and other components inside such modern marvels are literally microscopic in size. You need a fairly powerful microscope to be able to see them at all.

In your typical home electronics lab, and even in many professional labs, you will not be dealing with these tiny things, except as large clumps all tied together inside a chip or other framework. Still, the typical components you will use are often a bit too small for all but the tiniest of fingers. That's where pliers and tweezers come into play. They are an extension of your own hands and fingers, increasing your precision and allowing you to perform delicate tasks with tiny parts.

There are several types of pliers and tweezers that you may use on a daily basis in your lab. See Figure 2-4.

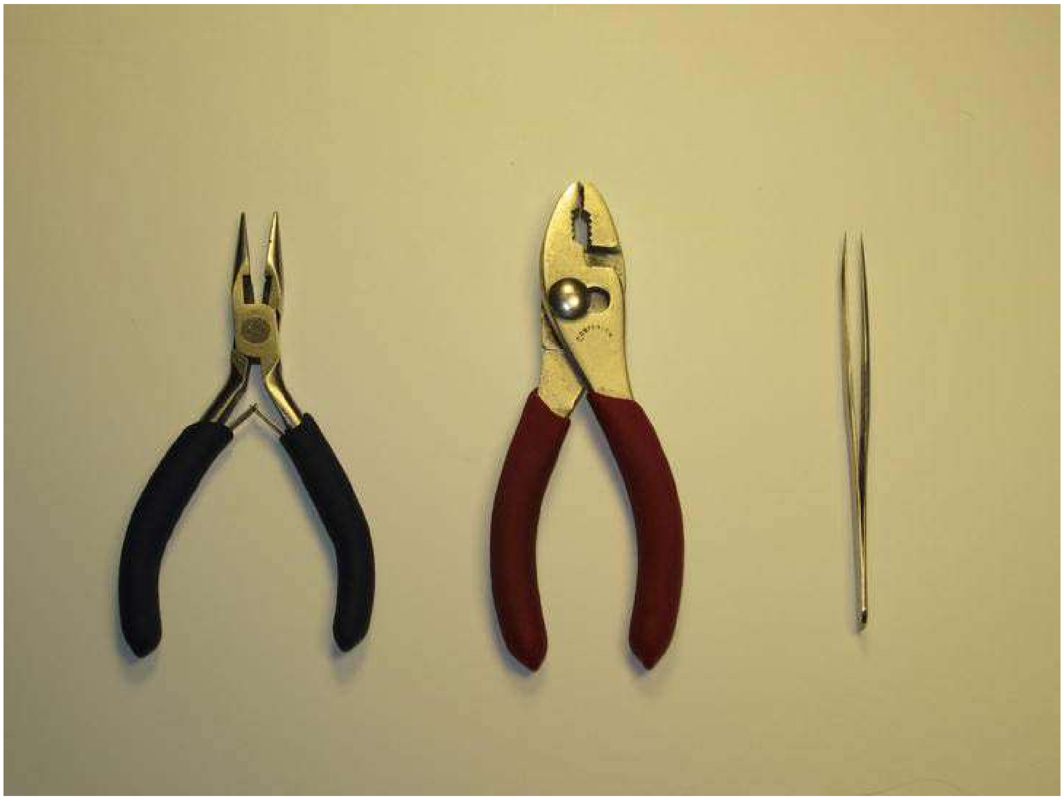

Figure 2-4. *Pliers and tweezers aid in working with small parts. On the left is a pair of needle-nosed pliers with cushioned grips and leaf springs. In the center is a pair of slip-joint pliers that can be adjusted to accommodate two different ranges of gripping action. On the right is a pair of needle-sharp tweezers, which can be used to pick up and place even the tiniest of electronic components—if your hand is steady enough!*

Screwdrivers, Nut Drivers, Wrenches, and Ratchets

A lot of electrical connections can be made using only a screwdriver. Figure 2-5 shows an electrical connection using a *barrier strip*. A specialized connector is crimped onto the end of the wire and held in place with a screw. A variety of crimp-on connectors are available and should be installed using the appropriate tool, as indicated by the connector's manufacturer. Just smashing the connector onto the wire with some pliers is not recommended.

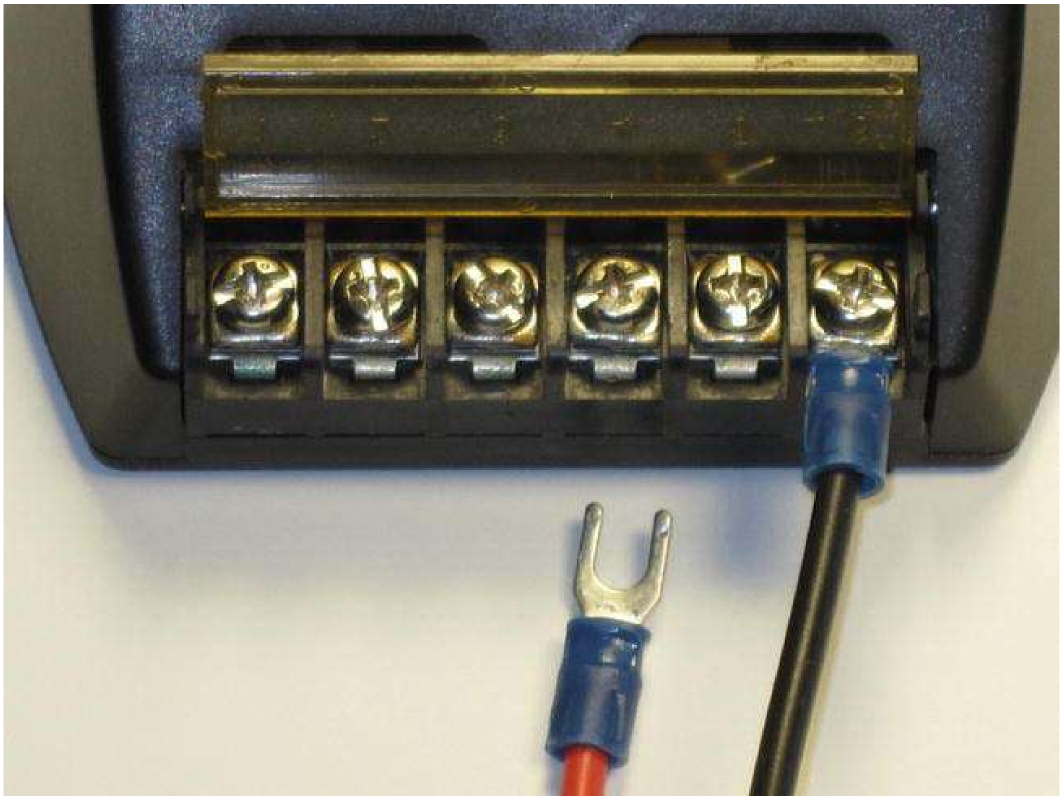

Figure 2-5. *A barrier strip on a low-cost solar-charge controller. No soldering required! The specialized connector is held in place with a screw. The connector is affixed to the end of the wire by deformation using a specially designed crimping tool. You can see the indentations in the insulating collar around the connector.*

Another type of connector that is tightened with a screw is called a *terminal block*. Unlike a barrier strip, no specialized connector is required on the wire. You can simply insert a stripped wire into the terminal block and tighten the connector using a screwdriver. Try to strip the wire only enough to allow the bare conductor to enter the terminal block, without leaving any noninsulated wire exposed that could accidentally short out against another wire. See Figure 2-6.

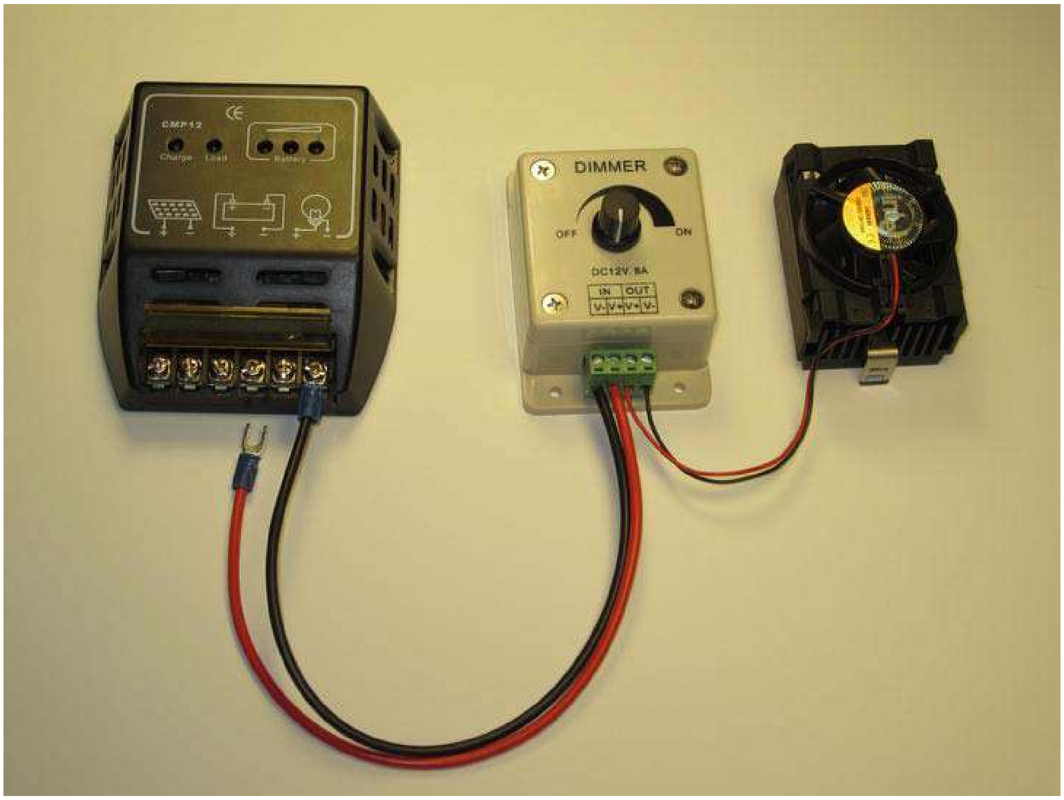
CE
CMP12
Charge
Load
Battery
DIMMER
OFF
ON
DC12V, 8A
IN
OUT
V-
V+
V+
V-

Figure 2-6. *A terminal block uses a screw for tightening the connection with a bare wire.*

Besides being useful for making electrical connections using barrier strips and terminal blocks, screwdrivers are also quite handy in their original function, which is to tighten and loosen mechanical fastening devices such as screws and bolts. See Figure 2-7.

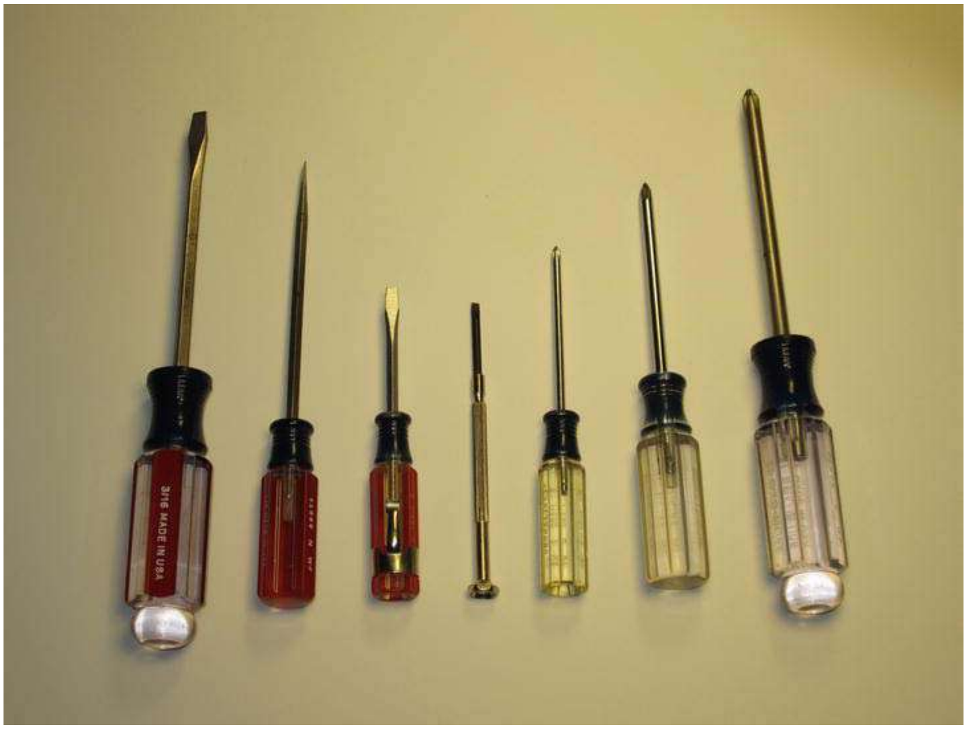

Figure 2-7. *A variety of screwdrivers. The four on the left are standard, or straight-bladed, screwdrivers. The three on the right are Phillips-style screwdrivers. Other specialty bits are also available.*

You should only use screwdrivers to drive screws and bolts. Don't try to use them as levers, pry bars, chisels, punches, or other tools. When you need a chisel, use a chisel. Keep your screwdrivers clean and dry. Put them back where they belong when you're done with them. You get to decide where they belong.

Screws, Bolts, and Nuts

There are a *very* large number of different kinds of screws and bolts. Some screws, such as wood screws, are made to be driven directly into wood and create their own threads in the material as they are installed. Other screws, largely referred to as machine screws, mate with a premachined or tapped hole, or a machine nut.

Nuts are often shaped with four or six sides that allow them to be held in place while the screw is turned. Alternately, the nuts themselves are turned while the screw is held in place. Some screws and bolts have a similarly shaped head that allows them to be turned with a wrench or nut driver instead of a traditional screwdriver.

Please resist the urge to overtighten screws and bolts. The mechanical advantage gained by the inclined plane of the screw thread multiplies the force used to turn the screw into a terrific amount of pressure on the screw itself. You can easily damage either the screw or the threads in the material to be fastened by overtightening a screw.

More Specialized Hand Tools

A good set of wire tools, pliers, and screwdrivers will go a long way when working with electronics. However, sometimes you need a more specialized tool for more specialized work.

Here are some of the more common tools that are a little more focused on a particular task than they are for general-purpose usage.

Wire Crimpers

As mentioned previously in this chapter, you can make solid electrical and mechanical connections using barrier strips and specially terminated wires. The optimized connectors used for this purpose require optimized tools for best performance.

Figure 2-8 depicts some examples of specialized crimp tools for electrical connections.

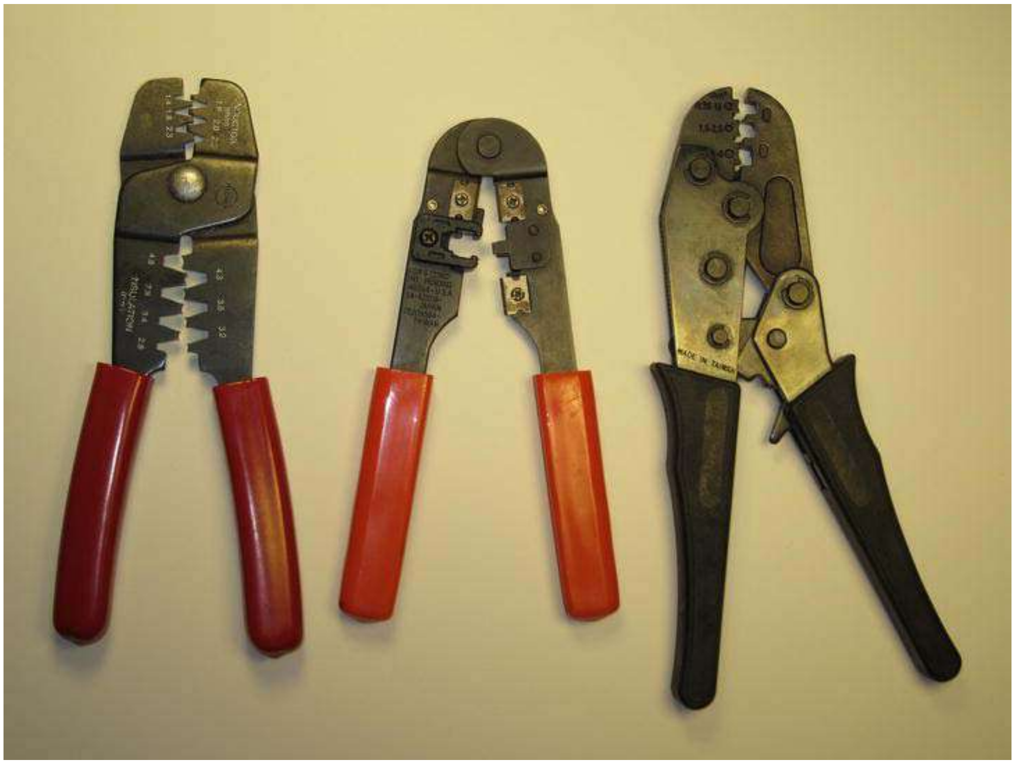

Figure 2-8. *Some common crimping tools for affixing electrical connectors to wire*

Referring again to Figure 2-8, the crimpers on the left are for crimping tiny pins to the ends of tiny wires that are then inserted into plastic connector bodies. The center crimp tool is for crimping modular jacks onto telephone cables. It also includes a wire-cutter blade (bottom) and a pair of controlled-depth stripper blades for removing only the outside layer of insulation from silver-satin flat telephone cord, without disturbing the individually insulated wires contained within. The crimp tool on the right features a ratcheting action that increases the crimping force applied to the connectors. This is the tool used to crimp the connectors show previously in Figure 2-5.

Computers

You might not think about computers as being "power tools" in your toolbox. As the size of computers, laptops, tablets, and smart phones decreases, and their computing power, networking ability, and storage capacities increase, they become much better candidates for use in the lab.

Traditional "desktop" personal computers (PCs), laptops, netbooks, and tablets can perform multiple roles in your laboratory setting. If you are lucky enough to have Internet access on your computer, it can become a conduit for all types of useful electronics information, in addition to its more accepted role as a source of funny cat videos. See Figure 2-9.

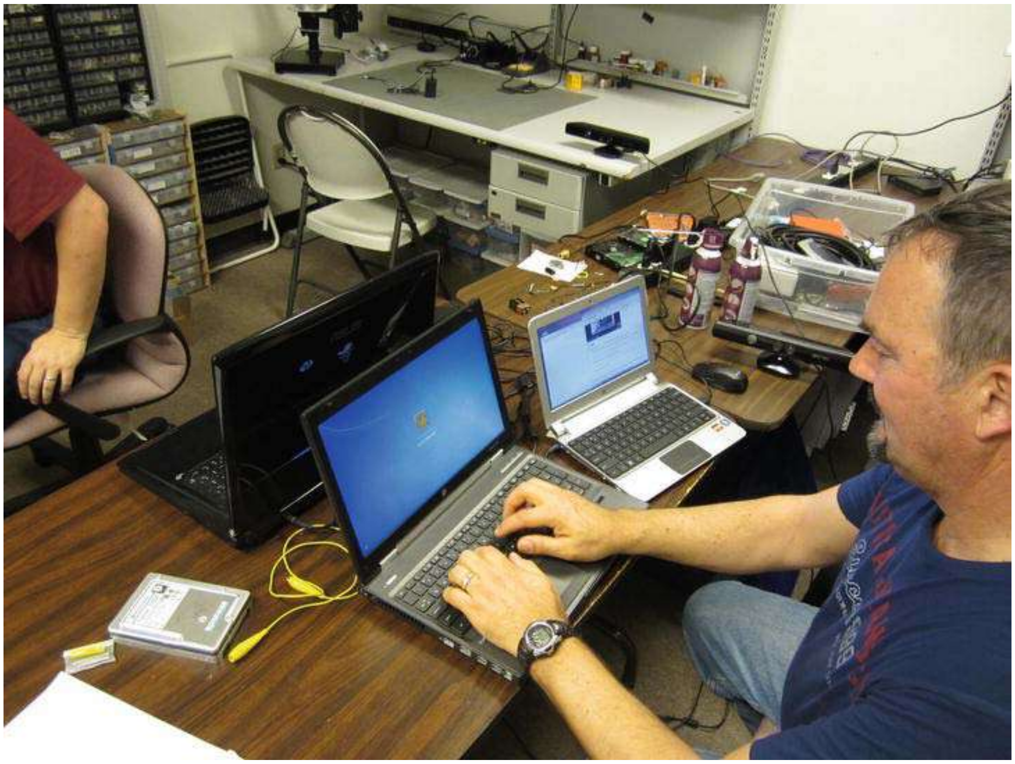

Figure 2-9. *Computers being used in the lab. Paul Bouchier of DPRG uses two laptops at once when testing new robot software with the Microsoft Kinect (just to the right of his mouse).*

Here are just a few of the many possible areas of practical knowledge, learning opportunities, and bits of interesting information that you can access using your computer:

> *Data sheets*: Almost all electronics manufacturers provide online access to a library of technical data sheets and application notes for their products. These should be your first choice for information about how and when to use their products. These sources also contain important product *errata*, which contains corrections and updates for products that have been made since their first publication.
>
> *Online tutorials*: The Web abounds with detailed and thought-provoking tutorials and background information on every topic imaginable, including, not surprisingly, mountains of information on computers and electronics.
>
> *User forums*: People tend to cluster around others with similar interests. Once upon a time when people where geographically limited and unable to communicate or interact with folks around the world, this severely restricted a lot of the more specialized areas of knowledge. Today, you can find user groups centered on every conceivable human endeavor (and some inconceivable ones, too) that are filled with helpful, knowledgeable members that are more than happy to answer questions and share experiences with you. Do try to be courteous, respectful, and fair when asking for help with your projects, and always try to repay in kind by helping those with similar questions that you've had to master.
>
> *Sharing/collaboration web sites*: As the cost and complexity of publishing your work on the Internet becomes more manageable, it becomes faster and easier to let others know about your progress, as well as your problems. You can easily set up your own web page for little or no cash, or join many of the available user groups that allow their members to post, browse, and otherwise benefit from the works of other similarly minded individuals.
>
> *Design tools*: You can find many free or reasonably priced electronic design and simulation tools online. Some manufacturers also provide free tools specifically for optimizing designs for their products.
>
> *Online suppliers*: Where you once had to walk (or drive) down to the store and see what was available in your area as far as electronic components and tools went, now you can browse through the catalogs of suppliers from all over the world, from the comfort of your own lab. Not only does this provide a much larger field from which to choose, it also levels the playing field, so to speak, forcing worldwide competition for your hard-earned cash. Bonus: Some manufacturers offer a limited numbers of free samples. Yip!

This list only scratches the surface of the kinds of information you can access using a computer. You can also use a computer as part of your electronic projects, and several examples of doing this are included later in the book.

Tools for Soldering

You don't have to learn how to solder to work with electronics, but it helps. It really does. It opens a whole new world of possible projects, repairs, and custom modifications to existing equipment. Just remember, as always, to keep your work area clean and safe. See Figure 2-10.

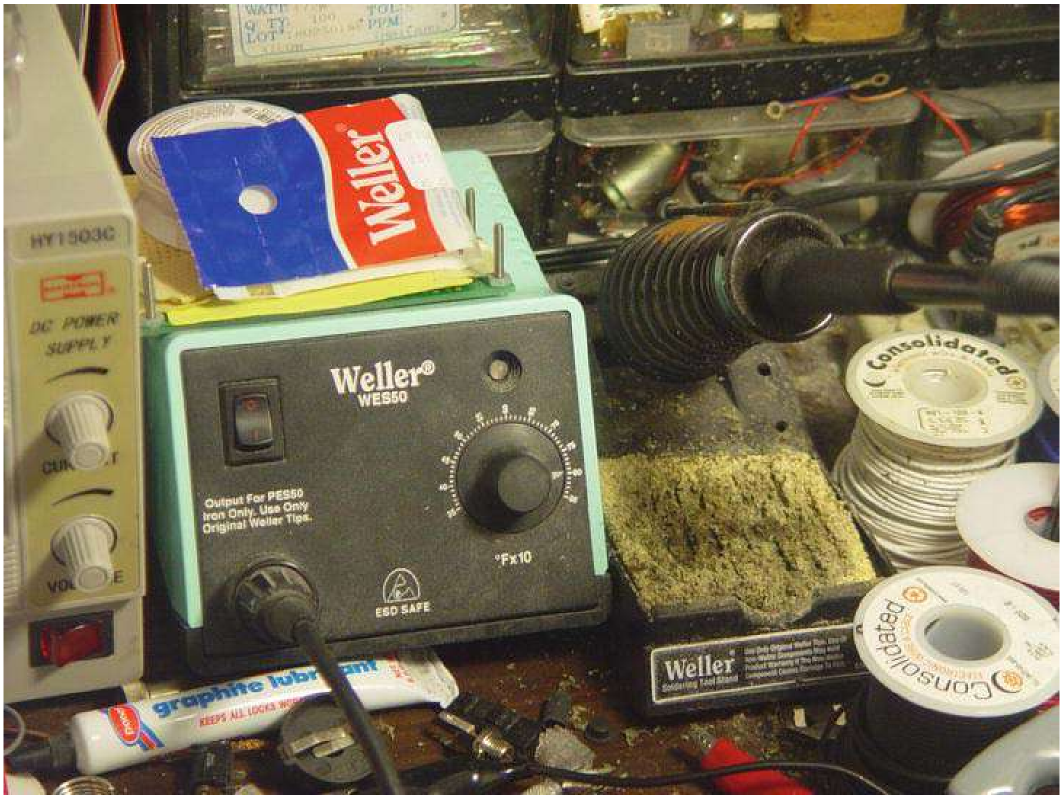
Weller
Weller
WES50
Output For PES50
Iron Only. Use Only
Original Weller Tips.
°Fx10
ESD SAFE
DC POWER
SUPPLY
HY1503C
Consolidated
Consolidated
graphite lubricant
Weller
Soldering Tool Stand

Figure 2-10. *A Weller controlled-temperature soldering station. What a mess!*

Like any worthwhile skill, soldering is best mastered with lots and lots of practice. The best way to learn this useful skill is to have someone show you how it's done. It's difficult, but not impossible, to learn proper soldering techniques from a book. That being said, have a look at Figure 2-11.

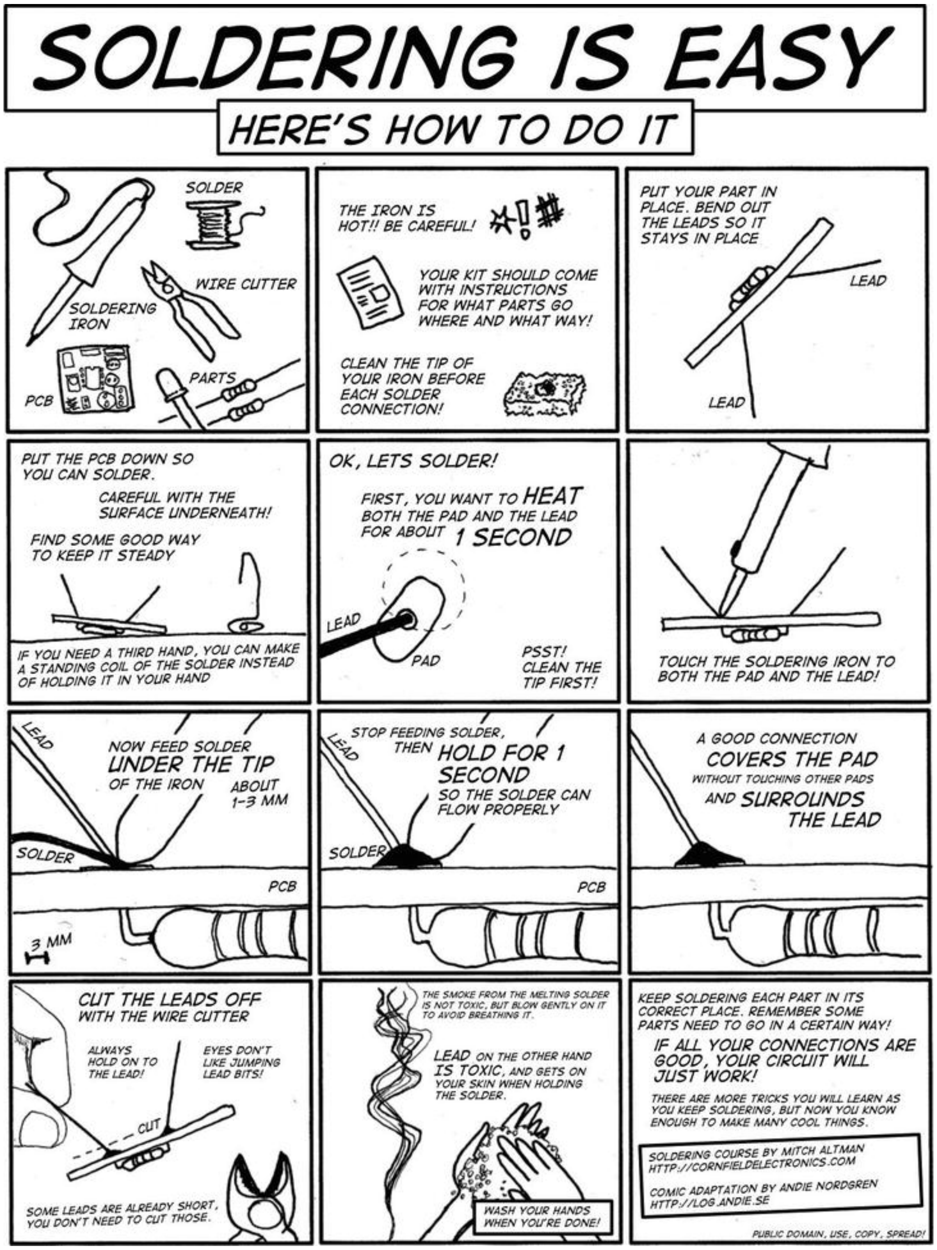
SOLDERING IS EASY
HERE'S HOW TO DO IT
SOLDER
WIRE CUTTER
SOLDERING IRON
PCB
PARTS
THE IRON IS HOT!! BE CAREFUL! *!#
YOUR KIT SHOULD COME WITH INSTRUCTIONS FOR WHAT PARTS GO WHERE AND WHAT WAY!
CLEAN THE TIP OF YOUR IRON BEFORE EACH SOLDER CONNECTION!
PUT YOUR PART IN PLACE. BEND OUT THE LEADS SO IT STAYS IN PLACE
LEAD
LEAD
PUT THE PCB DOWN SO YOU CAN SOLDER.
CAREFUL WITH THE SURFACE UNDERNEATH!
FIND SOME GOOD WAY TO KEEP IT STEADY
IF YOU NEED A THIRD HAND, YOU CAN MAKE A STANDING COIL OF THE SOLDER INSTEAD OF HOLDING IT IN YOUR HAND
OK, LETS SOLDER!
FIRST, YOU WANT TO HEAT BOTH THE PAD AND THE LEAD FOR ABOUT 1 SECOND
LEAD
PAD
PSST! CLEAN THE TIP FIRST!
TOUCH THE SOLDERING IRON TO BOTH THE PAD AND THE LEAD!
LEAD
NOW FEED SOLDER UNDER THE TIP OF THE IRON
ABOUT 1-3 MM
SOLDER
PCB
3 MM
LEAD
STOP FEEDING SOLDER, THEN HOLD FOR 1 SECOND SO THE SOLDER CAN FLOW PROPERLY
SOLDER
PCB
A GOOD CONNECTION COVERS THE PAD WITHOUT TOUCHING OTHER PADS AND SURROUNDS THE LEAD
CUT THE LEADS OFF WITH THE WIRE CUTTER
ALWAYS HOLD ON TO THE LEAD!
EYES DON'T LIKE JUMPING LEAD BITS!
CUT
SOME LEADS ARE ALREADY SHORT, YOU DON'T NEED TO CUT THOSE.
THE SMOKE FROM THE MELTING SOLDER IS NOT TOXIC, BUT BLOW GENTLY ON IT TO AVOID BREATHING IT.
LEAD ON THE OTHER HAND IS TOXIC, AND GETS ON YOUR SKIN WHEN HOLDING THE SOLDER.
WASH YOUR HANDS WHEN YOU'RE DONE!
KEEP SOLDERING EACH PART IN ITS CORRECT PLACE. REMEMBER SOME PARTS NEED TO GO IN A CERTAIN WAY!
IF ALL YOUR CONNECTIONS ARE GOOD, YOUR CIRCUIT WILL JUST WORK!
THERE ARE MORE TRICKS YOU WILL LEARN AS YOU KEEP SOLDERING, BUT NOW YOU KNOW ENOUGH TO MAKE MANY COOL THINGS.
SOLDERING COURSE BY MITCH ALTMAN HTTP://CORNFIELDELECTRONICS.COM
COMIC ADAPTATION BY ANDIE NORDGREN HTTP://LOG.ANDIE.SE
PUBLIC DOMAIN, USE, COPY, SPREAD!

***Figure 2-11.** Instructions for soldering. This delightful illustration by Andie Nordgren captures the basics of soldering as taught by soldering guru Mitch Altman (http://log.andie.se/post/397677855/soldering-is-easy).*

To solder, you will need a soldering iron. You will also need some solder and probably some wire cutters. That's just the tip of the iceberg, however. Like any good tool, your soldering station will soon accumulate many *accessories* to enhance your soldering prowess.

To begin your soldering adventure, you will need a *small* soldering iron, some solder, and a damp sponge. The big, pistol-grip style of soldering irons are for much larger projects, such as plumbing. Ideally you want something in the 15W to 50W range. See Figure 2-12.

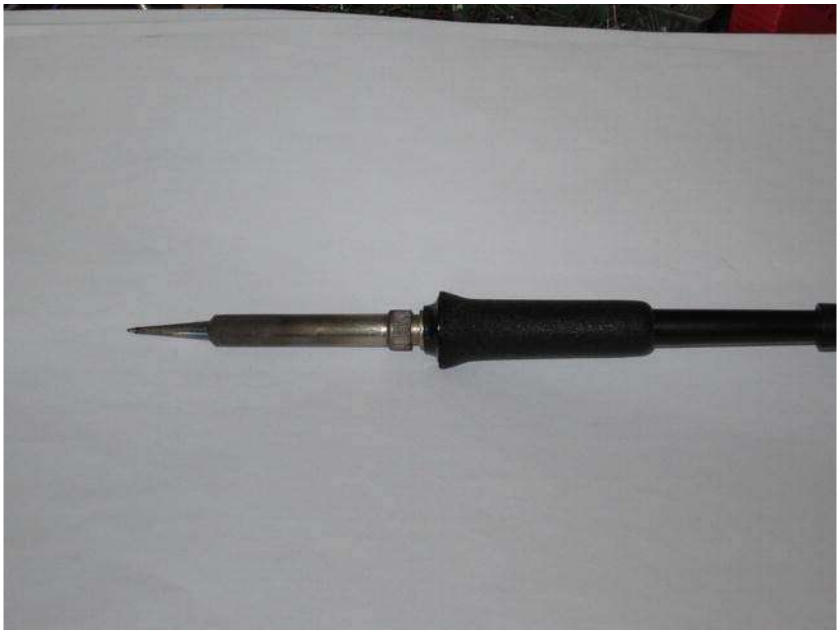

Figure 2-12. *The pencil-grip style of soldering irons is better suited for working on small electronics projects.*

The soldering iron in Figure 2-12 has a replaceable, conical tip. These tips are generally interchangeable for a particular model of soldering iron, but unfortunately not interchangeable between different models or manufacturers. Different tip shapes are also available, depending on what types of things you're wanting to solder.

The tip is usually made of copper, which is an excellent conductor of heat. Since copper really likes to combine with other metals at high temperatures (which is exactly what happens when soldering), most tips are coated in a thin layer of steel to give them much longer useful lifetimes. Tips still wear out, however, and should be replaced.

In this particular soldering iron, the next part of the iron is the removable barrel that holds the tip in place. This barrel is where the heating element is housed, so remember that it is just as hot as the tip! Try not to touch it. You will probably not want to make that mistake more than once.

The next section is the padded and insulated handle of the soldering iron, which is made of dense plastic foam. A quality soldering iron will be comfortable in your hand and won't produce excessive fatigue, even after hours of use. Lower-quality and lower-priced models will more rapidly take their toll on your hands.

Not shown in Figure 2-12 is the flexible power cord running back to the soldering iron's base. This cord supplies the electrical power that is converted to heat in the barrel, as well as some wires for a temperature sensor. This allows the unit to control the temperature of the tip with great accuracy. Note the temperature dial in Figure 2-10. Lower-cost units omit the temperature control features and run at a constant power output.

It will take some amount of time for the soldering iron tip to reach a working temperature. Likewise, it will also take some time to cool off once you've turned off the soldering iron.

■ **Caution** Soldering irons are very hot in use and can cause painful burns. Always treat them with respect!

The solder you use will be determined by what kind of soldering you want to do. For most small, electronics-related soldering projects, a small-gauge tin-lead or lead-free solder can be used. Bear in mind that lead is a poisonous material that slowly accumulates in your system, should you come into contact with it, as your body has no good way of eliminating it. Wash your hands before and after soldering, just to be safe.

Soldering works best when everything being soldered is sparklingly clean. To help facilitate this cleanliness, chemical compounds known as soldering rosin are used. These compounds chemically clean the surfaces being soldered when exposed to high levels of heat. Some types of solder have rosin incorporated into them. You can also obtain rosin as a paste into which you can dip wires and component leads before soldering. Other types can be applied with a dispensing pen.

■ **Caution** Don't breathe the smoke emitted during the soldering process. If possible, use a fan or fume extractor to redirect the soldering fumes away from you.

Once your soldering iron is up to its proper operating temperature, clean the tip by wiping it gently on a damp sponge. Your tips will last longer if you resist the urge to clean them before replacing the soldering iron in its holder. Even though it seems counterintuitive to "put it up dirty," this will more than double the life of your soldering iron tips. Just be sure to clean the tip *before* you solder anything.

Power Tools

Everybody loves power tools, don't they? Sometimes it's enough to just *imagine* that you might need one to justify a trip to the store and subsequent purchase. While this may be human nature (or perhaps only the nature of one particular gender), it doesn't fully justify having one of each in the lab, *just in case*.

Drills and Drill Bits

You will, however, find a few particular tools to be especially useful in your lab. A good hand drill or drill press is one of those tools. There seems to be no end of things electronical that need a few more holes drilled into them. Electronic enclosures and control panels need good, clean holes drilled into them. If you make your own PCBs at home, you will find that they need *lots* of holes drilled into them.

Like any power tool, an electric drill needs and deserves your respect. You'd be surprised how much unintentional damage you can cause with such a simple device. Observe these simple safety rules when operating a power drill:

- Keep your work area clean and well illuminated.
- Remove any watches or jewelry you might be wearing.
- If you have long hair, keep it pulled back and tied out of the way.
- Always clamp down any work to be drilled on a drill press.
- Always remove drill bits from the drill chuck after use.
- Discard dull or rusted drill bits that cannot be resharpened.
- Always wear safety glasses when operating power tools.

Hand drills come in all shapes, sizes, and price ranges. Some are powered by hand cranks and some are electric. Rechargeable batteries make for very portable drills that you can carry almost anywhere. See Figure 2-13. The only drawback is their limited charge time. That's why you should always have at least one spare battery charged up and ready to go. Electric drills with wires attached don't have this problem. Just don't trip over the cord!

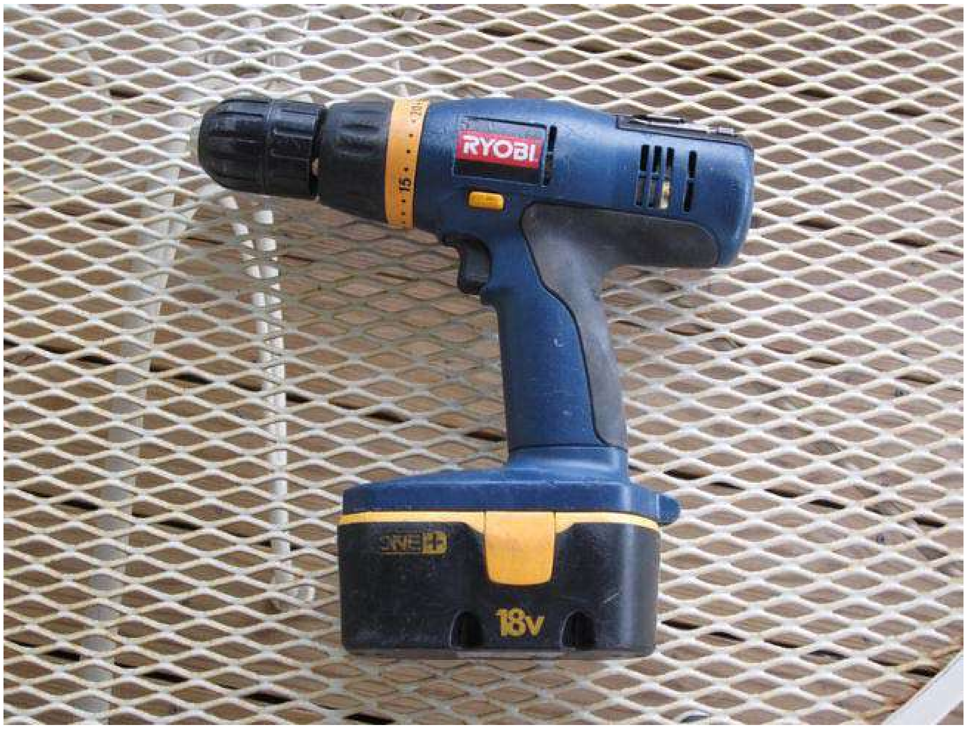

Figure 2-13. *A rechargeable hand drill. This model is both variable speed and reversible, which are two very desirable features in a hand drill.*

■ **Note** Power drills are also very handy when used with screwdriver bits. They can save you time and many blisters!

Drill bits likewise come in all sizes and shapes. You can buy sets with a variety of sizes, and these will usually come with a handy storage container that conveniently labels all the sizes for you. This only works, however, if you take the time to return the drill bits to their proper location once you've finished working with them.

A drill press has a large frame that facilitates drilling a precisely aligned hole, over and over again. You can make spindle-speed adjustments to drill presses (even small ones, such as the one pictured in Figure 2-14) either through an electronic speed control or by changing up the belts and pulleys that drive the spindle. There are even small frames for mounting handheld drills or rotary tools, such as the Dremel brand of tools. You can often, however, find actual, dedicated drill presses on sale for about the same price as the daintier Dremel tools.

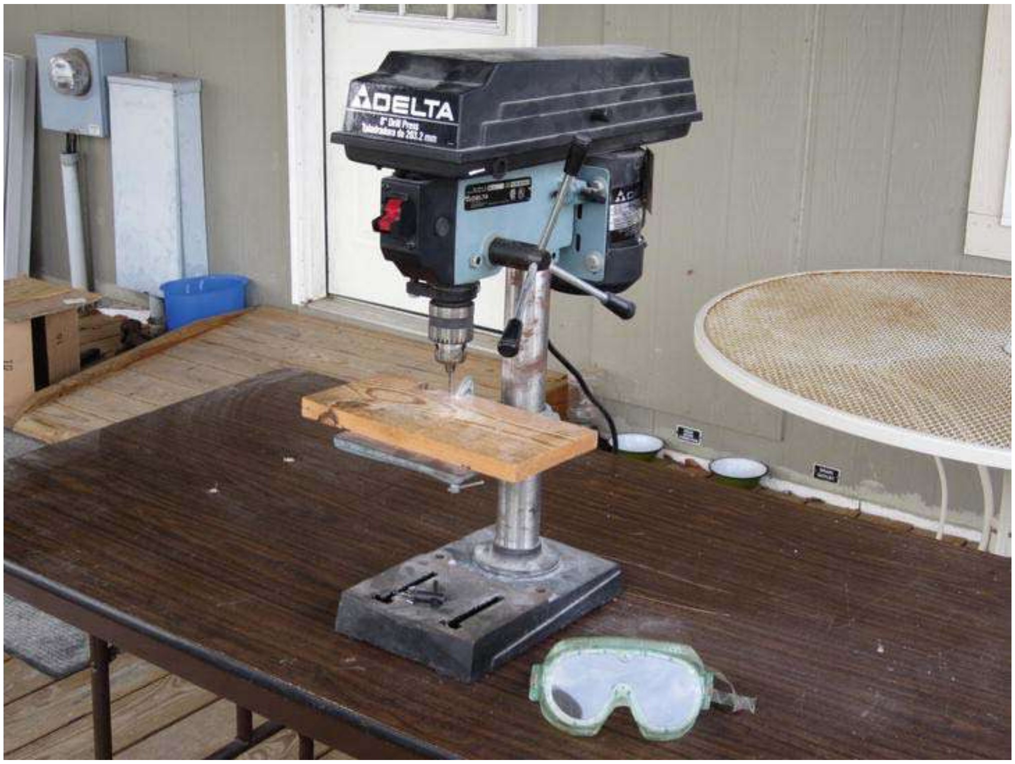

Figure 2-14. *A small tabletop drill press*

■ **Caution** When working with drills and presses, always keep your work area clear. Always wear safety glasses when using power tools. Try to keep the chuck key handy. Don't leave drill bits mounted in the chuck when not in use. Many chuck keys have built-in springs that prevent them from remaining in the chuck. Those that don't risk being accidentally left in the chuck when the drill its turned on, which can either fling the chuck across the workshop or make a big mess of your favorite hands.

Laser Cutters and 3D Printers

While not everyone can afford some of the more expensive and esoteric tools that are available, some of these tools are getting more popular than others; excellent examples are laser cutters and 3D printers. See Figures 2-15 and 2-16.

Laser cutters and engravers use a powerful carbon-dioxide laser to cut through wood and non-chlorine-based plastics. The laser beam is focused onto the work surface using mirrors that are mounted in an XY gantry, allowing the cutting area to be controlled by a computer. Typical installations require venting to the outside because of the various aromas produced in the process, as well as a water-cooling system for the laser tube itself.

There are also some newer laser cutters being built by hobbyists from scratch.

__Figure 2-15.__ A laser cutter. A powerful 40W carbon-dioxide laser is used to cut through wood and non-chlorine-based plastic, all under computer control. It can also be used as an engraver. This model is manufactured by Full Spectrum Engineering.

3D printers seem to be all the rage these days. We're all pretty accustomed to the idea of printing on paper with a computer, although it's still a pretty recent innovation, as far as the history of printing goes. The idea of printing a solid object, on the other hand, is considered to be something of a game-changer.

The simplest 3D printers use a plastic extrusion and a motorized gantry to operate as a kind of a robotic glue gun. A plastic filament is fed into a heated fixture that is moved around a print bed, leaving a trail of cooling, gooey plastic. Either the extruder or the print bed is moved using motors under computer control to build up a layer at a time of the desired object.

Higher-technology 3D printers use this and other techniques to produce tangible, solid objects. Some use layers of fine powder that is glued together with a binder dispensed from a print head, similar to an inkjet printer. As each layer is printed, the object is lowered and another layer is deposited. This can even be done with sugar, as in the case of the CandyFab 4000, designed by the clever folks at Evil Mad Scientist Laboratories (`www.evilmadscientist.com/article.php/candyfab`). A stream of heated air is used to fuse a layer of sugar into a single solid structure, and an object is built up of successive layers.

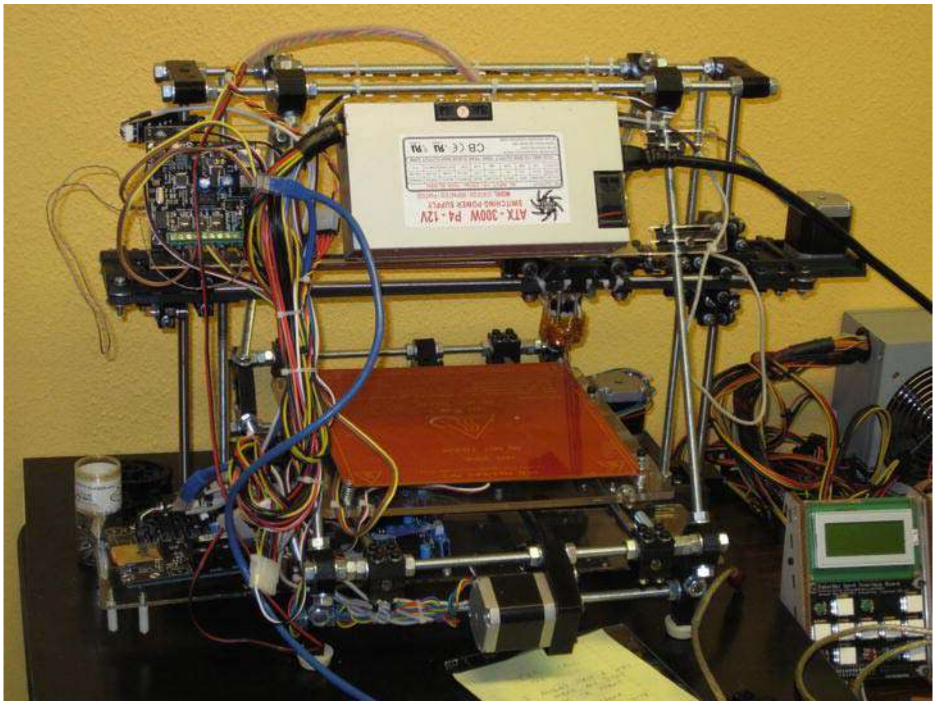

Figure 2-16. *A 3D printer at Dallas Makerspace. This machine uses a heated extruder to build 3D objects out of plastic.*

Basic Test and Measurement Equipment

When it is behaving itself, electricity is generally invisible. To "see" what it is doing, or not doing, you will need some specialized tools. These fall into the two general categories of *test* and *measurement* equipment.

The line between testing and measuring is sometimes blurry. One possible distinction might be that testing verifies proper operation within acceptable parameters, while measurement actually assigns values to those parameters. An example of testing is, Does that flashlight work? The answer is yes or no, arrived at by the simple expedient of turning it on and observing the results. An example of measurement is, How much current does that flashlight draw? The answer might be 0.78A, and this would be determined using the appropriate measurement device, which in this case could be an ammeter.

Inspection Equipment

Sometimes, however, measuring things isn't hard at all, especially when the things being measured are actual, tangible items that you can hold in your hand. What you will need are some basic tools, primarily to be able to *see* what it is that you are measuring, and secondarily to take accurate measurements of them, at least in regard to their physical dimension. This comes in very handy when trying to identify components during repair or equipment upgrades, when the original technical documentation is not at hand.

Even the youngest and finest of eyeballs will need help with the teeny-tiny components that are available nowadays. You should try to find at least a small magnifying glass to examine fine details. If your budget allows, you might want to consider a microscope of modest power to look at your handiwork. Careful shopping will result in finding some older microscopes at bargain prices. Expect to pay a premium price for the shiny, new stuff. See Figure 2-17.

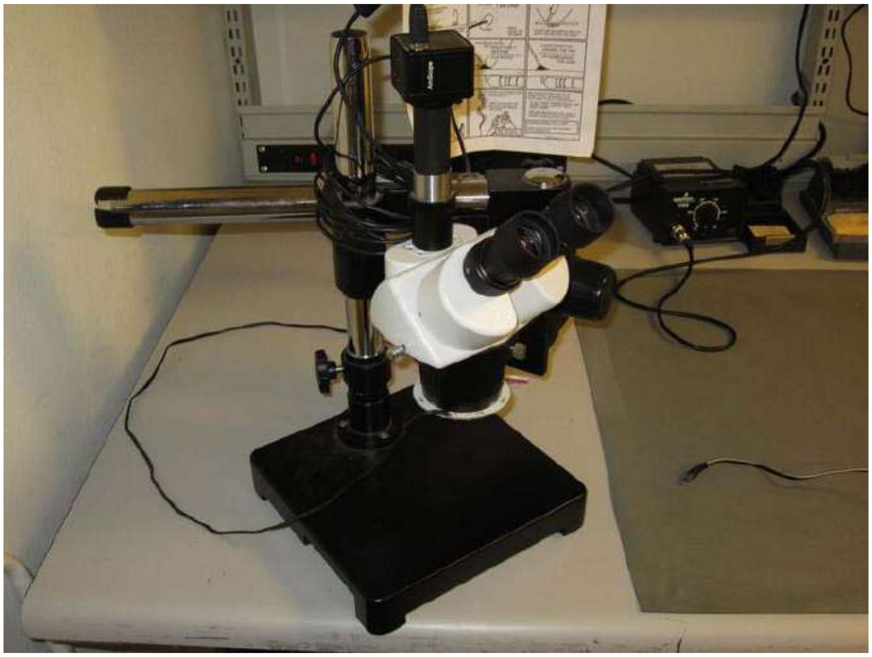

Figure 2-17. *A very nice stereo microscope. The power cord running to the white ring (partially hidden) supplies power to a ring of white LEDs that help illuminate the field of view. The base is very heavy to help stabilize the microscope.*

Even a humble magnifying lens can be very helpful in the lab, especially if it has a built-in light. See Figure 2-18.

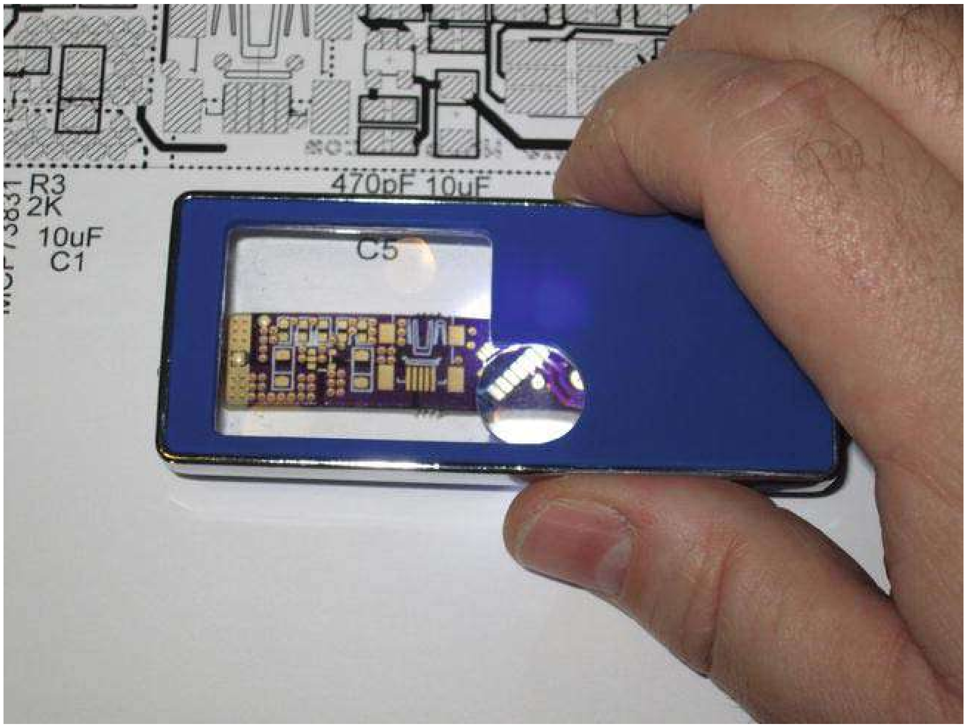

Figure 2-18. *A small, handheld magnifier with a built-in LED light is very handy for examining small features and details, such as are found on many PCBs.*

Physical Measurement Devices

Once you know what something looks like, at extreme close-ups, can you describe in absolute terms how big it is?

Take a look at Figure 2-19. Here we have, from the top to the bottom, a traditional tape measure, marked in both English and metric units (fractional inches and decimal centimeters), a digital caliper that can render measurements in either system at the touch of a button, and an architect's scale. While it may not be evident in the top-down view, the architect's scale is triangular in cross-section, resulting in not two but *six* different measurement options. This is mostly due to architects liking to draw things to scale, instead of full-sized.

The digital caliper is interesting in that it also provides a computer interface so that its readings can be pulled in by a computing device of some sort. These devices used to be quite expensive, but mass production has brought their prices down significantly.

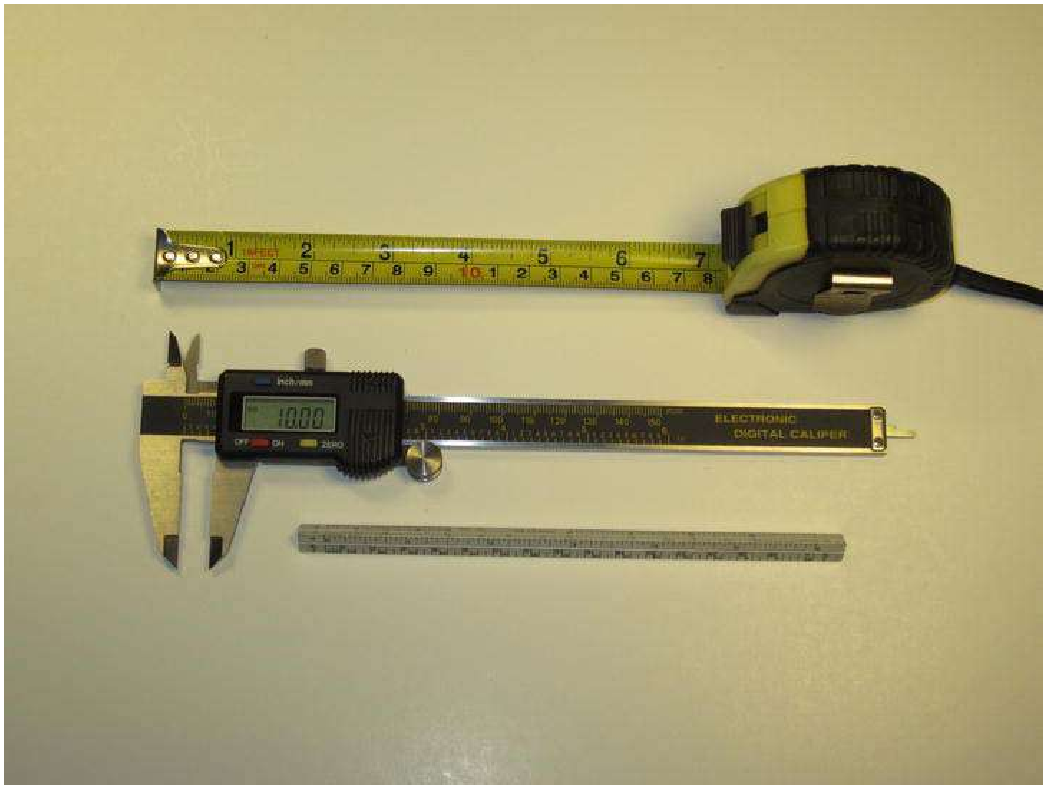

Figure 2-19. *From top to bottom, a tape measure, a set of digital calipers, and an architect's scale.*

Simple Electronic Measurement Tools

Now that you've got the tools to examine what these things look like and how big or small they are, let's move on to the more mysterious art of measuring their electrical properties.

Voltmeters

Voltage is a measure of electromotive force, and the standard unit of measure is the *volt*, named after Alessandro Volta, an Italian physicist. Voltage is sometimes compared to the pressure of water flowing through a garden hose, while amperage is compared to the amount of water flowing through the hose at any given time. It's not particularly useful to push this water analogy too far, but it's helpful in the beginning.

Voltage can be accurately measured with, you guessed it, a voltmeter. See Figure 2-20.

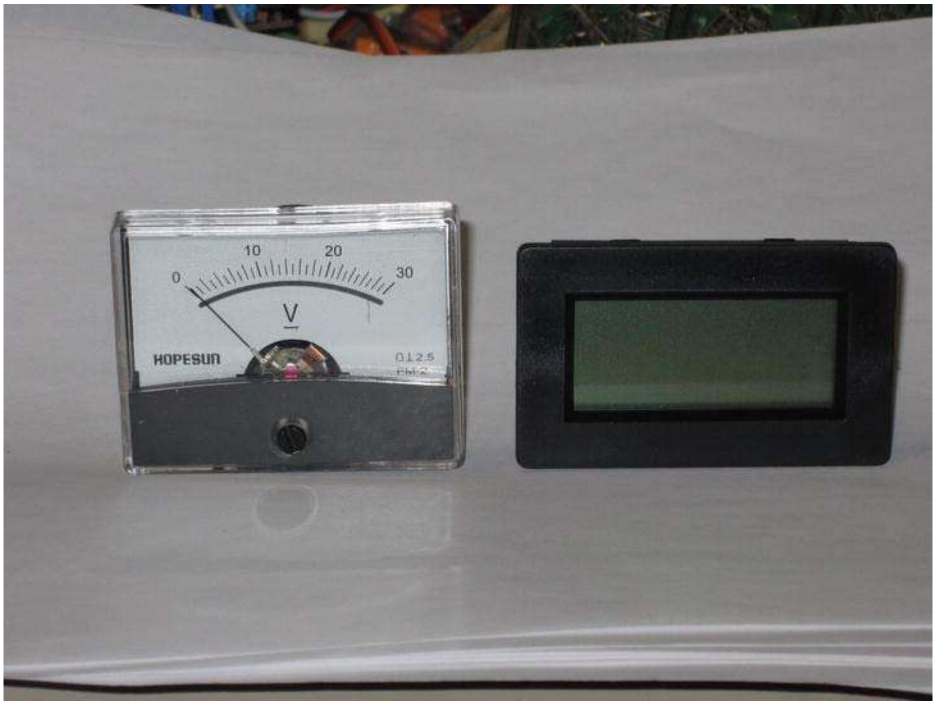

Figure 2-20. *A panel-mount analog voltmeter (left) and a digital version (right)*

A voltmeter measures the voltage, or *potential difference*, in a circuit across two points. Sometimes one of the points is a fixed reference point, such as the electrical *ground* mentioned in many circuits. This is often just a convenience for the purpose of measuring things, as the ground in an electrical circuit may or may not be connected to the ground under your feet.

A voltmeter will have two terminals, labeled positive (+) and negative (–). Looking at the analog voltmeter in Figure 2-20 (the one on the left), we see a needle that can swing from 0 on the left to 30 on the right. The prominent capital *V* on the dial indicates that the unit of measure is the volt. This indicates that you can use this meter to directly read from 0 to 30 volts, just by placing its terminals across a voltage. The meter itself is electromechanical in design, and creates a magnetic field that works against a spring and a fixed magnet. This action causes the needle to move to a point generally corresponding to the labeled voltage on the dial. These needles are not especially accurate, but can be corrected using the screwdriver adjustment at the bottom of the meter.

An analog panel meter such as this one does not require a separate power supply for proper operation. It draws all the power it needs from the signal being measured. This makes it unsuitable for measuring low-power circuits, where the current draw from the meter would unduly influence the reading.

The panel meter on the right is a digital circuit that measures the voltage and displays it on the built-in liquid crystal display (LCD). It requires a separate power supply for normal operation.

Ammeters

An *ammeter* looks and operates just like a voltmeter, except that it measures electrical current flow instead of voltage. The flow of electrical current is measured in units called *amperes* (A). The basic mechanism is almost identical to a voltmeter. To convert a current flow into a voltage, an electrical device known as a *shunt* is used. The shunt is simply a fixed-value resistor that produces a voltage across its terminals in relation to the amount of current that is flowing through it and its inherent resistance to the flow of current.

This relationship between voltage, current, and resistance is elegantly expressed in *Ohm's Law*. If you know two of the measurements in a circuit, you can calculate the third, based on their fixed and proportional relationship to each other. You will find this law to be exceedingly helpful to you when designing, repairing, and upgrading electronics.

To avoid using fractions, which seem to terrify so many people for no good reason, you can use Ohm's Law thusly:

- Voltage (volts) = current (amps) × resistance (ohms)
- Current (amps) = voltage (volts) ÷ resistance (ohms)
- Resistance (ohms) = voltage (volts) ÷ current (amps)

That's all there is to it, if you're just working with volts, amps, and ohms. Things can get weird, however, when you start to work with alternating current (AC), which is the kind of electricity you find in wall outlets and audio circuits.

Ohmmeters

An excellent application of Ohm's Law is found in building a resistance meter, also called an ohmmeter, because it measures resistance. Resistance is measured in units called ohms, often abbreviated with the capital Greek letter omega (Ω).

There are two ways to do measure resistance. One is to pass a known voltage though the unknown resistance and measure the current that is flowing in the circuit. The other is to pass a fixed current though an unknown resistance and measure the voltage across it. This second method is the one most commonly used.

You're most likely going to use a multimeter to measure resistance, as this is one of the more common functions of these handy devices. Mutlimeters are covered in the very next section.

Advanced Test and Measurement Equipment

It would seem that for simple circuits, having just a voltmeter, an ammeter, and an ohmmeter would be all you need. For the most part, this is correct. However, have you considered what would happen if you applied 100V to the 30V meter shown in Figure 2-20? What if you simply reversed the polarity of the signal by accident? Something bad for the meter, that's for sure. It might not actually catch on fire *right away*, but it would certainly damage the delicate, spring-loaded mechanism, possibly rendering it useless as a reliable measuring instrument.

The same can be said for an unprotected ammeter or ohmmeter. Luckily, the nice people that make good test equipment stay up late at night, thinking of the stupid things we might do with their products, and protect them accordingly. Let's look at some meters that measure multiple electrical characteristics, which strangely enough are called *multimeters.*

Multimeters

Multimeters will almost always allow you to measure various ranges of DC (direct current) volts, AC volts, current, and resistance. More advanced meters also measure things like frequency and duty cycle or even temperature. Some meters will even directly measure capacitors, inductors, and transistors using specialized connectors.

As with most tools, you generally get what you pay for with multimeters. The meter on the left in Figure 2-21 was purchased for under $3, including shipping from China. It arrived DOA (dead on arrival) and was not deemed important enough to return. It will make an interesting clock or fake lie detector someday. The unit on the right was purchased new for nearly $400 and has provided over a decade of useful service. The unit in the center is modeled after the unit on the right, it would seem, but can be bought new for between $10 and $20. The unit in the photo was bought wholesale for $4, because sometimes you get lucky. It's often handy to have a second meter in the lab—for example, when you need to measure the voltage *and* the current flowing in a circuit simultaneously, which you just can't do with a single meter.

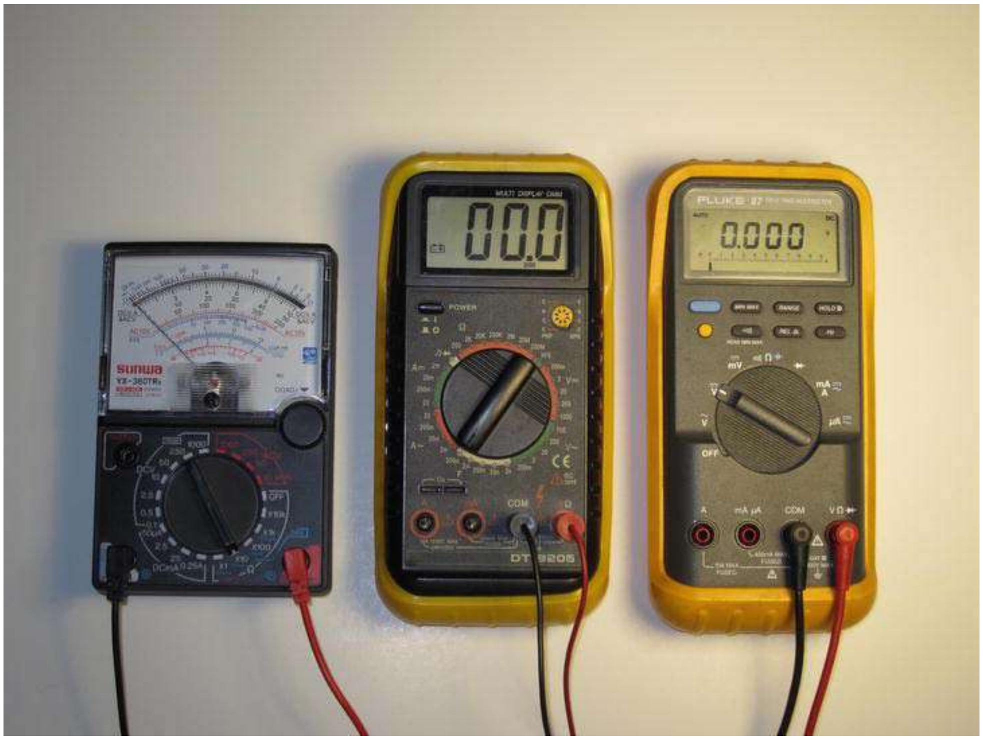

Figure 2-21. *Examples of some multimeters. The meter on the left features an analog display that is easy to read and acommodates several readout scales simultaneously. The remaining two meters have digital readouts and offer more accuracy. All the meters use a large rotary knob for selecting the proper function. The higher-cost unit on the right features auto-ranging, intellignetly deciding what range to use. The right two units also have high-impact boots to help protect them in the event of a fall.*

Your present budget will help determine what level of multimeter you might want to obtain. If you're wanting to sink less cash into a multimeter right now, consider looking for used equipment in good shape. As with any good tool, you can spend as much as you like on these. The sky really is the limit here.

Most all-in-one meters protect their inputs using fuses. You will have to disassemble your meter to replace these fuses, so think twice and measure once (to paraphrase the old carpenter's saying, "Measure twice and cut once").

■ **Tip** Be sure to set up your meter for the proper function and range *before* taking measurements.

Oscilloscopes

Once you've gone beyond measuring simple DC and AC voltages and move on to more complex signals and waveforms, you'll have to lay aside your trusty meter and fire up your favorite oscilloscope. This instrument is more complex than your traditional multimeter, and offers a different perspective on what's going on in your circuit.

While many modern oscilloscopes have many modes of operation, their basic function is to draw a picture of your signal on a small screen. In the olden days, this was done with a moving dot that was swept across a phosphorescent screen, leaving a residual, fading line of light. If this pattern was repeated over time, a visual representation of the signal could be seen on the screen. See Figure 2-22.

Modern oscilloscopes use a computer monitor or built-in screen display to draw the signal. This saves space, weight, and power requirements.

When purchasing an oscilloscope, the features you pay most for are the number of available channels and the bandwidth. The least expensive oscilloscopes will have a single channel. Having multiple channels allows you to "see" multiple signals simultaneously, which is often helpful in understanding what is going on in a circuit.

Bandwidth refers to the range of frequencies that can accurately be measured by the oscilloscope. Lower bandwidth scopes require much simpler electronics, and are thus cheaper to produce.

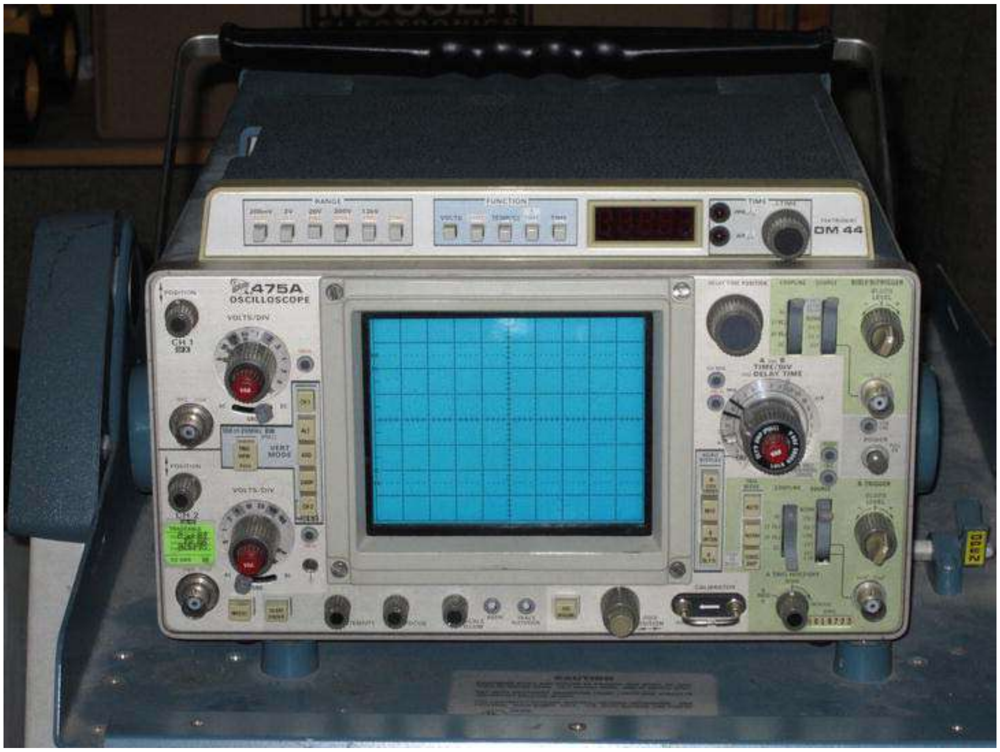

Figure 2-22. *A Tektronix 475A oscilloscope. This piece of vintage test equipment is considered by some to be the best analog oscilloscope ever made. It also has a digital multimeter built in along the top edge.*

The Golden Age

Once upon a time, not so long ago, a quality oscilloscope weighed 40 pounds and had a folding handle to make it, if not portable, at least *transportable*. It also had its own line of furniture to place it on, so that you could wheel it around the lab as you needed. Oh, and it also cost about three months' salary, and that's if you had a *good* job.

The good news? That same scope still works today. It can be repaired or calibrated to the exact same standards as when it was brand new, *plus* you can buy it on eBay for a few hundred dollars, sometimes less. That's because when these devices were designed back in the 1970s, they were made to be the best in the world, and cost was not a factor.

Modern Oscilloscopes

Today you can buy a quality oscilloscope for a few hundred to a few thousand dollars. These scopes not only display your signal on a large, colorful screen, but they can also analyze the signal and tell you all kinds of interesting things about it. For an example of such an oscilloscope, see Figure 2-23.

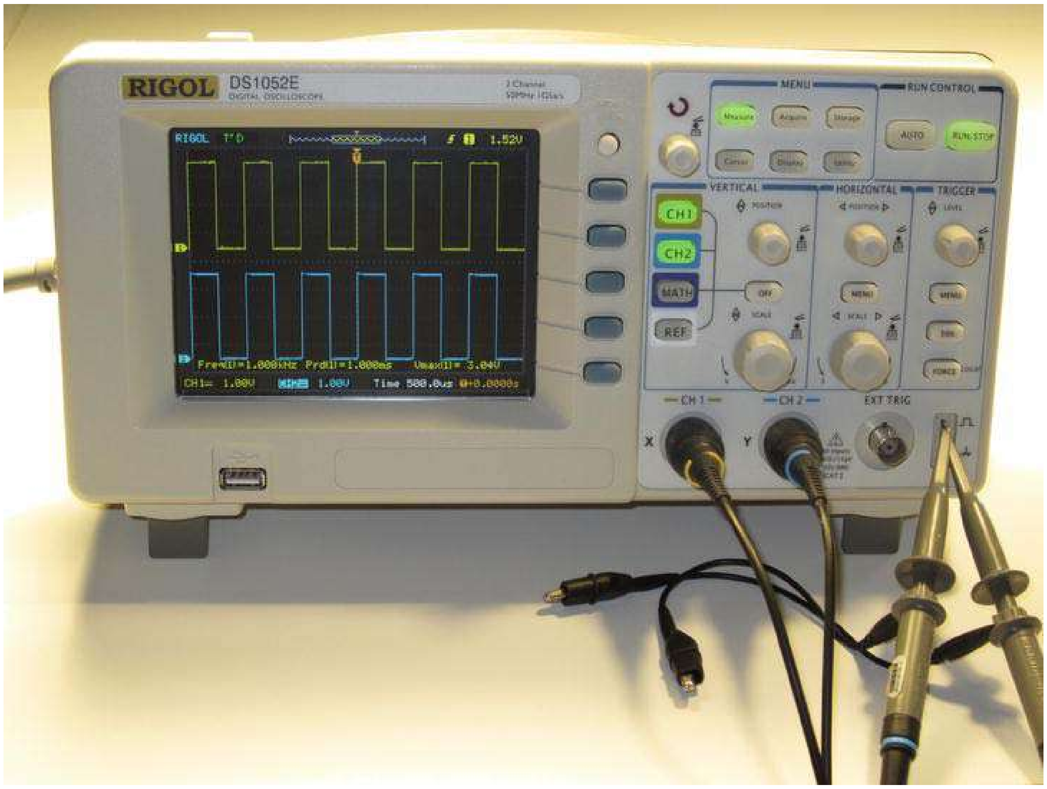
RIGOL DS1052E
DIGITAL OSCILLOSCOPE
MENU
RUN CONTROL
VERTICAL
HORIZONTAL
TRIGGER
CH1
CH2
MATH
REF
EXT TRIG

Figure 2-23. *A modern digital storage oscilloscope. The RIGOL DS1052E is an affordable scope that will handle most troubleshooting and design challenges that you might encounter in your lab.*

This new breed of digital storage oscilloscope (DSO) can even connect to your PC to be remotely operated and allow you to capture screen shots that can then be pasted into books about how to build your own electronics lab. It's true. See Figure 2-24.

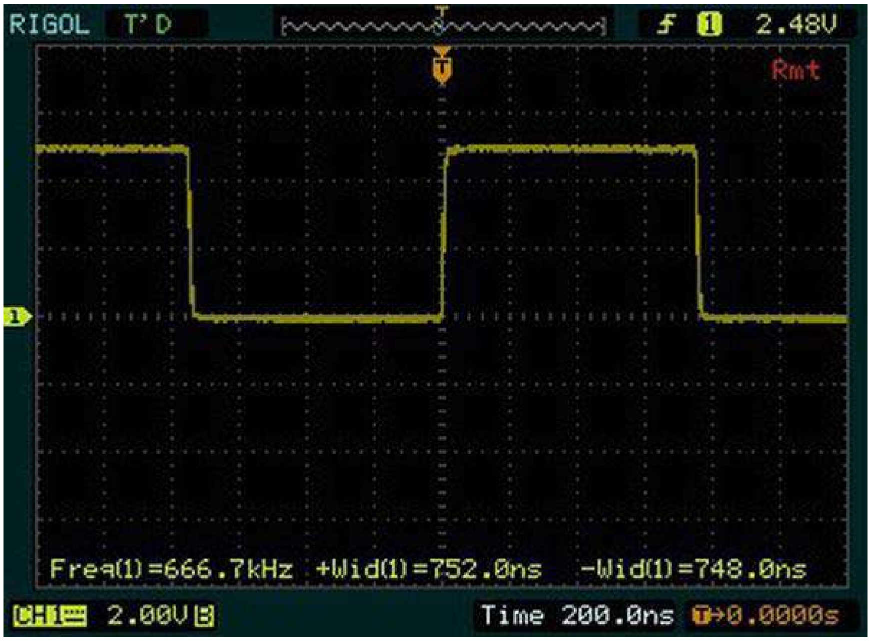

Figure 2-24. *A screen capture made with the RIGOL DS1052E oscilloscope.*

Tiny Oscilloscopes

The march of electronic miniaturization never stops, and so it is with oscilloscopes. If both your needs and your budget are modest enough, perhaps you can get by with something like the pocket-sized wonder shown in Figure 2-25.

While little scopes like this lack the accuracy and bandwidth of their larger relatives, they are definitely handy to have when you need to look at a signal. If you can keep your expectations in check when it comes to the user interface, you can get a lot of utility from such a small package.

Additionally, one of the benefits of a device such as this is that the hardware and internal software (yes, there's a little computer in there, somewhere) is all open source. This means that you have access to all the schematics and source code. And what does that mean? It means that if you have the skills and the right equipment, you can redesign the entire apparatus to fit your own requirements, assuming that you understand the basic limitations of the electronics. The sky is the limit. You are empowered.

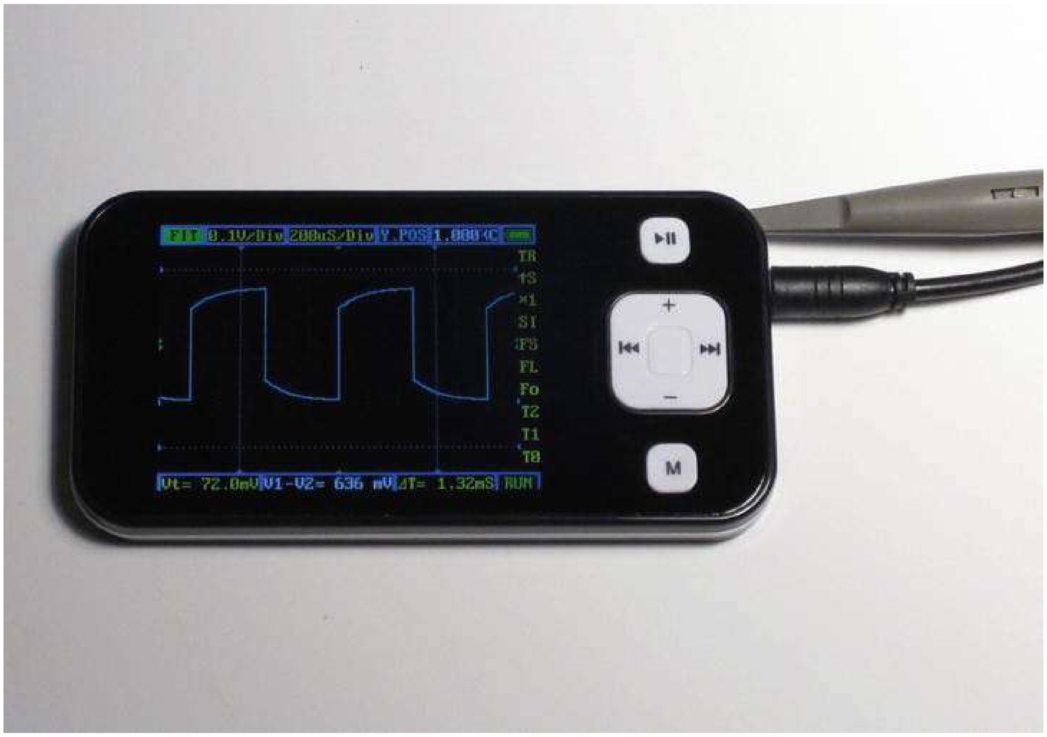

Figure 2-25. *A handheld oscilloscope, which can also fit in your pocket. This is the DSO201, which has been discontinued by the manufacturer, but can still be snapped up for a song from various online distributors, as well as eBay. This unit was purchased new for $60 and features a single channel with a bandwidth of 1 MHz.*

Power Supplies

One of the most important pieces of test equipment in your lab is a reliable power supply. Building a good bench power supply is an excellent first or second electronics project (see Chapter 5 for more on adding power supply to your lab and Appendix A for more on building your own). This is because the basic design is very simple and easy to build, yet allows you to add any number of additional features later.

Once you get tired of throwing away spent 9V batteries and grow weary of waiting for the red light on the battery charger to go out, it's time to invest (either time or money or both) in a decent bench power supply.

It doesn't have to be fancy. In fact, the simpler the better at first. What you're aiming for primarily at this point is reliability. You want to know that when you turn on your 5V supply, it's putting out 5 volts, not 2 or 11, no matter what you do to it. If it's overloaded, it needs to deal with it and not be catching on fire. If it's underloaded, or not loaded at all, it still needs to maintain a nice, clean output.

Want a cheap or free power supply? Do you have any dead or discarded electrical gizmos that came with a charger? These make excellent starter supplies for the new lab. Do you have any extra USB chargers? These are even better, because they supply up to 2.5W (and sometimes more) of *regulated* 5V. We'll use some of these in experiments in Chapter 3.

Organizing Your Tools

So now you have an idea about some of the tools you want (or already have) for your lab. Where are they all going to live? This is a big question, and deserves a big answer.

One way to look at this book is that the entire latter half is dedicated to answering this question. There are, obviously, many possible arrangements that will work for most people. But we're talking about *your* lab, with *your* tools, arranged and ordered to meet, specifically, *your* needs.

As you progress through Chapter 3 (the very next chapter in this book), you'll get a chance to play with both your tools and some interesting electronic components, all in the interest of getting you familiar with some basic concepts. These concepts and ideas are not new, but they may be new to you. The good news is that time invested today and in the very near future should pay off big dividends to you later in life. Who knows what you'll devise in that electronics lab of yours?

Summary

Starting with your brain and working outward, this chapter has covered some of the tools and devices you might use in your electronics lab. These include basic hand tools, computers, tools for soldering, some common power tools you might find handy, as well as some fancy-schmancy tools you might only find in a hackerspace, school, or electronics clubs, such as laser cutters and 3D printers.

You also learned about some of the basic test and measurement tools you'll be needing when working on electronic circuits in your lab. These include magnifying glasses, measuring tapes, and meters (oh my!), as well as oscilloscopes and power supplies.

Additionally, *lots* of advice on safety was presented. These topics were learned the hard way. Do try to learn from the mistakes of others, won't you?

Just a hint was given about how you might want to organize your tools within your lab. Take care of your tools and they will most certainly take care of you. The details about possible organizational scenarios is left for the second half of the book, where you will learn about setting up labs big and small, as befits your needs, budget, and interests.

Chapter 3 ties in much that you have learned here in Chapter 2 by introducing you to not only some of the basic electronics components and their uses and characteristics, but also how to use a good number of the tools described in this chapter. Hopefully this chapter has whetted your appetite and readied you to dive into some circuits and see what kind of mischief you can wreak on the world.

CHAPTER 3

Components

You're going to need some components to play with, as well as the knowledge to use them effectively. This chapter will introduce you to some of the bits and pieces that make up modern electric and electronic circuits. It will also teach you a little bit about how to identify components from their appearance and markings, when available.

Once you've got an idea about what these parts do in a circuit, you'll learn a little more about how to measure their electrical properties and put them to use. You'll also be shown what *not* to do, in some select examples.

There are an enormous number of both manufacturers and vendors of electronic components these days. Trying to list even a few of them here in print doesn't really make much sense, as a good percentage of that information will be incorrect or obsolete as soon as the ink dries on the page.

Having a variety of basic components readily available is convenient, until your collection grows beyond your ability to house or organize it. This happens far too often, in reality. Most common electronic parts can be obtained within a day or two from local sources, unless you live in a really secluded part of the planet. Also, most of the more common parts will typically cost only a few pennies apiece, so there aren't a lot of reasons to stockpile random parts. Keep a few of each on hand for experimentation, but hold off on pointless hoarding . . . if you can.

Conductors and Interconnects

In any electrical or electronic circuit, there are just as many connections between components as there are components. These include wires, cables, connectors, plugs, sockets, and all those other pieces that help the little electrons zip around the circuit.

First, let's take a look at wires and interconnects in a very abstract way. We can think of wires as connections between other components, without having to get into a lot of detail about the wire itself. We just want a way to indicate that there is an electrical connection among all the bits, and using an electrical schematic diagram is an easy and effective way to do this.

Electrical Schematics

A *schematic* diagram is any kind of stylized drawing that represents a simplified view of something. Usually some sorts of symbols are used to represent the parts that make up the whole, as well as the connections or relationships that exist between them and possibly the outside world. You can have schematics for plumbing, hiking trails, or tribbles, a well as electrical and electronic circuits. Let's talk about electronic schematics for a little bit.

The electrical symbol for a wire is almost always just a plain, solid line. If there's something really special about the wire that needs to be used, it is usually indicated in a note somewhere on the schematic. This can be used to indicate what color the wire needs to be, what size is appropriate, whether there are any special requirements for the insulation, and so forth. Most of the time it's just a plain, solid line. For an example of just how simple it can be to draw a wire in a schematic, see Figure 3-1.

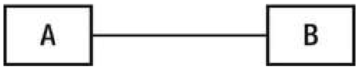

Figure 3-1. *A very simple electrical schematic diagram, showing how to get from point A to point B with a single wire. Drawing wires in schematics is easy!*

If a wire starts in just one place and ends in just one place, everything stays nice and easy. If multiple wires need to converge on a single point, or one signal wire needs to split up to go to more than one destination, it gets just a little more complicated, but not too much. See Figure 3-2.

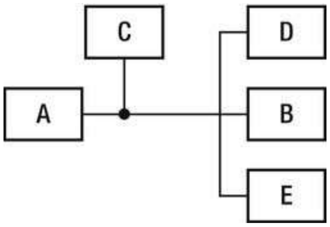

Figure 3-2. *Wires that are connected to each other usually have a dot at their intersection. If there is no dot, there is no connection between overlapping wires.*

In the figure, points A, B, and C are connected together. Points D and E are connected together, but are not connected to points A, B or C, even though the wires look like they are crossing. In a professional schematic, a careful draftsperson would route the wire from point D to point E on the right side of the components to avoid any possible confusion.

Some folks like to draw a little bump, or bridge, when two unconnected wires cross in a schematic, just to make sure that everyone knows that the wires aren't connected. This gets to be a bit tedious once you've got more than two or three components in your schematic, and it can look a bit unsightly. Feel free, however, to draw your schematics however you see fit. Just remember that others might not share your particular drafting sensibilities.

Wire and Cables

Chapter 2 took a quick look at wire and how it is made to help you pick out the best wire-cutting and wire-stripping tools. Now it is time to consider wire as an electrical component within your circuits. Knowing what a wire can and cannot do will help you pick the right kind of wire to use. Yes, wire has limits! This is important both for building new circuits from scratch as well as repairing, replacing, or upgrading wiring in existing circuits.

Wire can be used for all sorts of artsy and craftsy purposes. Copper wire is the most common wire used in electronics. Sometimes the copper conductor is *tinned* with another conductive metal with a low melting point to make it easier to solder. When you cut this wire, you can see the copper on the inside.

Pure copper wire is getting very expensive these days. Unscrupulous people try to pass off copper alloy with inferior electrical properties as the real deal. A quick way to test is with a magnet. Copper is not attracted to magnets.

As mentioned in Chapter 2, wire comes in two major categories: solid core and stranded. Solid core wire is just a single piece of copper wire, while stranded wire is made up of many smaller strands of copper wire, usually twisted together. Stranded wire is generally more flexible than solid core wire.

Copper wire is measured using a gauge system. Different systems exist around the world. A common system used in the United States and Canada is the American Wire Gauge (AWG) standard. It uses a series of single numbers to denote the relative size of the wire, with larger numbers representing smaller-diameter wires. These numbers go all the way down to zero (0 gauge), but they don't stop there. Larger wires than 0 gauge are called 00 gauge, double-ought, or sometimes 2/0. There are also three- and four-digit versions, but after that, those giant wires are specified with their actual cross-sectional diameter.

Using Wire: The Very Short Version

Fatter wires carry more current with less heat being generated due to resistance. The wire's insulation is rated for a maximum voltage and temperature. If this temperature is exceeded, the insulation will soften, melt, or catch on fire. Avoid this by using the right sized wire for the job. Exceeding the maximum voltage can cause more problems if the current arcs through the insulation. This can cause a short circuit and may even cause a fire. It is also likely to give off noxious fumes. Choose wisely!

For low voltages (under 36V) and small current loads (under 1A), you can use whatever you want. It doesn't matter. Pick something pretty. A thing of beauty is a joy forever.

For higher-voltage applications, make sure the wire's insulation is rated *far in excess* of what you anticipate. The rating should be visibly printed on the insulation itself. See Figure 3-3.

Figure 3-3. *The wire's insulation only works up to a point. That point is usually printed on the insulation itself, but you may need to put on your good reading glasses to see it. This wire states, "LL33911 CSA TEW 105°C 600V FTT 14AWG," which translates to the manufacturer's part number, Canadian Standards Association certified thermoplastic equipment wire, maximum temperature of 105°C, 600V insulation dialectric strength, 14 gauge.*

For higher-current applications, you really need to use the highest-gauge wire that is necessary to carry the required current. Hot wires cause fires.

Using Wire: The Longer Version

If you will spend some time learning about the properties of wire, you will be able to make *much* better decisions regarding its use and application. Assuming that the overall shape of the cross section of the wire is round, the gauge of the wire determines how much resistance to the flow of electrical current the wire will exhibit over a given distance. This resistance, in turn, helps determine the amount of electrical current that the wire can safely carry.

When electric current meets resistance, the electrical power is converted into some other form of energy. Usually this energy is given off as heat, or thermal energy. It can also produce magnetic fields, sound, light, and radio waves.

A good example of this is demonstrated by the heater coils on an electric stovetop. When the stove's switch is turned on, current flows through the coil. Most of the current is converted directly into heat. Electric heaters are very efficient this way. A small amount of the electrical power is converted into light. You can see this as a red glow coming from the heater coil when it is hot. The coil also creates a magnetic field.

If you know the gauge of a wire, you can look up the wire's characteristic resistance. These numbers will be close but perhaps not exact, as there will always be small variations in wire diameter, conductor purity, and other factors. Generally, however, the numbers are pretty close.

Let's take another look at the AWG table (see Table 3-1), and look at a few simple examples to help understand what these wires can do for us. (A more complete listing is available at `http://en.wikipedia.org/wiki/American wire gauge`, from which this partial list was obtained.)

Table 3-1. *The AWG*

AWG	Diameter Area			Copper Resistance	
	Inch	mm	mm²	Ω/km mΩ/m	Ω/1000 ft mΩ/ft
0000 (4/0)	0.46 00	11.6 84	107	0.1608	0.04901
000 (3/0)	0.4096	10.404	85.0	0.20 28	0.06180
00 (2/0)	0.3648	9.266	67.4	0.25 57	0.07793
0 (1/0)	0.3249	8.252	53.5	0.32 24	0.09827
1 0.28	93	7.348	42.4	0.40 66	0.1239
2 0.25	76	6.544	33.6	0.51 27	0.1563
3 0.22	94	5.827	26.7	0.64 65	0.1970
4 0.20	43	5.189	21.2	0.81 52	0.2485
5 0.18	19	4.621	16.8	1.02 8	0.3133
6 0.16	20	4.115	13.3	1.29 6	0.3951
7 0.14	43	3.665	10.5	1.63 4	0.4982

8 0.12	85	3.264	8.37 2.	06 1	0.6282
10 0.10	19	2.588	5.26 3.	27 7	0.9989
12 0.08	08	2.053	3.31 5.	21 1	1.588
14 0.06	41	1.628	2.08 8.	28 6	2.525
16 0.05	08	1.291	1.31 1	3.1 7	4.016
18 0.04	03	1.024	0.823 2	0.9 5	6.385
20 0.03	20	0.812	0.518 3	3.3 1	10.15
22 0.02	53	0.644	0.326 5	2.9 6	16.14
24 0.02	01	0.511	0.205 8	4.2 2	25.67
26 0.01	59	0.405	0.129	133.9	40.81
28	0.0126	0.32 1 0	08 10	212.9	64.90
30	0.0100	0.25 5 0	05 09	338.6	103.2

Conductivity

The first thing you need to know is that copper wire really, really wants to conduct electricity. It wants to conduct electricity more than it wants to *resist* the flow of electrical current. Even so, copper is not a perfect conductor, and it has a certain amount of inherent resistance. This resistance causes the energy carried by the electrical current to be sidetracked, or converted into heat and other forms of energy. The higher the resistance in the wire, the more heat will be produced.

Normally this heat is a very small thing and is easily dissipated into the surrounding air. You usually can't even tell by touching that a wire is a fraction of a degree warmer than its surroundings—but this is only true when the amount of energy is relatively small, like in our rewired flashlight example. When you start to push a *lot* of power through a wire, the heat buildup can outpace the cooling-off provided by the surrounding air, and the temperature of the wire increases.

Again, this is no big deal as long as that temperature stays within a comfortable range. But what happens when the temperature gets really high?

First, the insulation will get soft, melt, or even catch on fire. After the insulation, if any, has burned off, the wire will glow red hot and eventually melt. This is known as *catastrophic wire failure*, which can result in what electrical engineers refer to as an unauthorized thermal event, which we call a fire. Yikes!

But look around you. You're probably sitting right next to some wires right now, very few of which (hopefully) are on fire, even though they are carrying variously large loads. How do they keep from burning up? The answer is in the proper selection of the right *size* of wire for the job. Bigger wires have less resistance to the flow of electricity, so they are able to carry more power with less heat being produced.

Now, you don't want to have to wire up all of your electronic circuits using jumper cables from your car's trunk, as that gets to be cumbersome after a while. Let's look at some good guidelines for what size wire to use for your projects.

In Chapter 2 you learned about (or were reminded of) Ohm's Law, which illustrates the relationships between voltage, current, and resistance in any electrical circuit. In summary, $V = I \times R$, where V is volts, I is current in amps, and R is resistance in ohms. It's not just a good idea, it's the law!

So how much electricity will we lose going through a foot or so of 16-gauge wire? If we know that one foot of 16-gauge wire has a resistance of 4mΩ (and we know this from looking it up in the chart), and we build a circuit that pushes 1A of current through this wire, we can figure out a couple of interesting things using Ohm's Law and a calculator.

Applying Ohm's Law

If you know two of the three variables used in Ohm's Law, you can calculate the third variable. Since we know R and I, we can calculate V. Since V = I × R, we can substitute our values of 1A and 4mΩ (0.004Ω) into the equation, and get 0.004V, or 4 millivolts.

That's not very many volts at all! In this case, that's a good thing. The voltage we just calculated is the *voltage drop* across the length of the wire. Those millivolts never make it from one end of the wire to the other—they're lost. Poor millivolts, they never had a chance.

OK, let's not let their heroic sacrifice be in vain. If we know volts, and we know amps, we can also calculate power, which is the rate at which energy is converted to work. The unit of measure for electrical power is *watts*, and is also the same as Joules per second (J/s), but that's another story. We usually use the capital letter *W* to abbreviate wattage, but we use the capital letter *P* to represent the *value* in equations.

Calculating power is simple if you know the volts and the amps. Just multiply them together: P = V × I. Easy.

So in our example so far, we have 4mV, 1A, and 4mΩ. How much power is being dissipated (wasted) in this circuit? P = 0.004V × 1A = 0.004W, or 4 milliwatts, which is just four-thousandths of a watt. That's just about enough power to light up a dim LED, and not much else. So you're not going to be starting any fires with these numbers. You most likely would not even be able to measure the rise in temperature that is occurring in the wire (oh yes, it's occurring, but it's very, very small).

Let's put 10A through a foot of 30-gauge wire. On second thought, let's not, and just do the math instead. The voltage drop will be V = I × R, or 10A × 0.1032Ω = 1.032V. The power dissipated would be P = V × I, or 1.032V × 10A = 10.32W. That wire would be a crispy critter in no time. Ten watts doesn't sound like much when you're talking about a home stereo or a lightbulb or a hair dryer, but that skinny little wire has almost no surface area with which to dissipate the generated heat. This means that the temperature goes up, quickly. Bye-bye, little piece of wire.

As a side note, you just learned how most current-limiting fuses operate. The conductor within the fuse literally *fuses*, or evaporates, when the internally generated heat produced from an "overcurrent" situation rises beyond its melting point. Poof, the circuit is opened, and property and lives are saved.

■ **Warning** Never circumvent the protection afforded by a fuse.

Ideally what you want is the smallest wire that will carry the required amount of current with an acceptable rise in temperature. Remember that any rise in temperature is in addition to the *ambient* temperature. If you're sitting in a room that is "room temperature," which is assumed to be 25°C or 77°F (the same thing, if you want to check it), and you're expecting a 50°C rise in temperature due to waste heat, your wire is going to be 75°C, which is far too hot to touch. Remember that 100°C is the boiling point of (pure) water (at sea level). Hot!

One more thing that complicates the thermal calculations is that copper, as a conductor, has a *positive thermal coefficient*, meaning that the characteristic resistance of a piece of wire (which we can look up in the AWG table) goes *up* by a known amount as the temperature increases. You need to consider this when you build circuits that are *supposed* to get hot, such as electric heaters. In most other cases, it's best to keep the heat to a minimum, as it represents wasted power, plain and simple.

Testing Your Wires

Can something as simple as a wire go bad? Yes, it most certainly can. How can you tell? Usually, it's because something doesn't work the way you think it should, or *worse*, it works sometimes and not others. A great deal of time, the problem is right at the end of the wire, where it is trying to make a connection with something else. However, it's possible for a wire to develop a problem right in the middle. This can be caused by a tiny nick in the insulation that lets in nasty elements such as rain and dirt, allowing the inner conductor to corrode. This type of problem takes time to make itself apparent, but it can certainly happen, especially in equipment that is exposed to the outside world for any length of time.

A quick way to test a wire is with a piece of test equipment called a *continuity tester*. This simple device usually is battery powered and has two probes. When you connect the probes together, the tester makes a noise, lights up a light, or maybe does both. Make sure your continuity tester's batteries are fresh and that the probes, if they are removable, are firmly seated in their sockets. Even test equipment needs testing from time to time.

■ **Note** A continuity tester is one of the easiest pieces of test equipment to cobble together yourself. See Appendix A for more information about building your own tools.

As mentioned previously, the problem is almost always at the very end of the wire. This is a good thing to know, so you can start looking in the most likely places first, saving valuable troubleshooting time.

A much better way to test a wire is with an ohmmeter. Using an ohmmeter, you can determine the *exact* amount of resistance in a wire, which under ideal circumstances should be very low. One ohm or less is typical for most normal-sized wires that you are likely to encounter. Super-long wires can have higher resistance, and that's where you can consult the handy AWG table to find out what the resistance is *supposed* to be.

When testing wires, especially wires with attached connectors, it's a good idea to wiggle the wire about vigorously, as this will reveal some intermittent connections for you.

Storing Your Wire

If you're buying new wire, it will probably come on a reel. It's often handy to keep common sizes and colors on your bench, if you have room. A paper towel–holder makes an excellent wire dispenser rack, although you can also spend more time or money buying a prebuilt one or designing and fabricating your own.

A custom red-and-black wire dispenser is shown in Figure 3-4. You might use a fixture like this if many pairs of identical-length wires are needed for a project. Not having to chase the wire spools all over the bench really saves a lot of time. This structure was built using a MicroRax miniature slotted aluminum extrusion (*http://www.microrax.com*), which is perhaps a bit of overkill, but was a lot of fun to plan and build. A coat hanger, artfully bent, would have also done the trick.

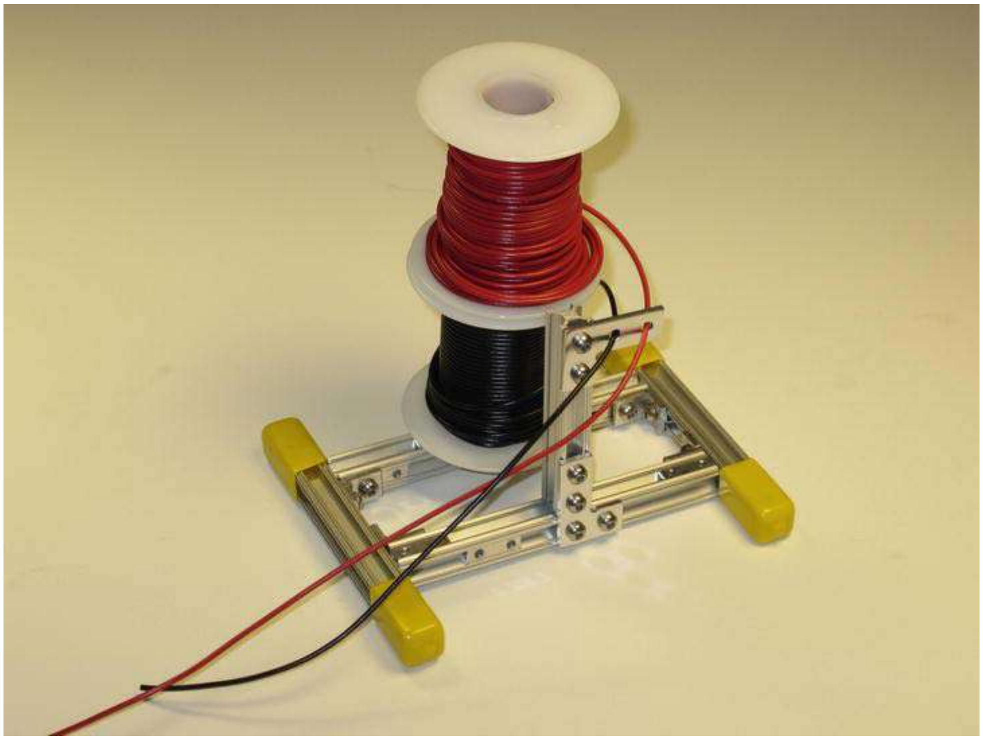

Figure 3-4. *A custom wire-dispensing frame made from modular aluminum extrusion. This kind of fixture makes it easier to cut both a red and a black wire to almost exactly the same length, which happens a lot when you use red and black wire for positive and negative power connections, respectively.*

Connectors

Connectors make life easier for the casual wiring enthusiast. The hardcore types solder everything together, but that is not always practical or even the right thing to do in every circumstance.

Keep in mind that every connector, just like every wire, has some sort of electrical limits. These limits must be respected.

Clip Leads

Also called *alligator clips* or *test leads*, you can buy ready-made wires or build your own from individual connectors. The flexible rubber or plastic boot extends the insulation of the wire. A spring-loaded jaw, often with serrated teeth, is used to clamp down on the wire, lead, or electrical terminal desired (see Figure 3-5).

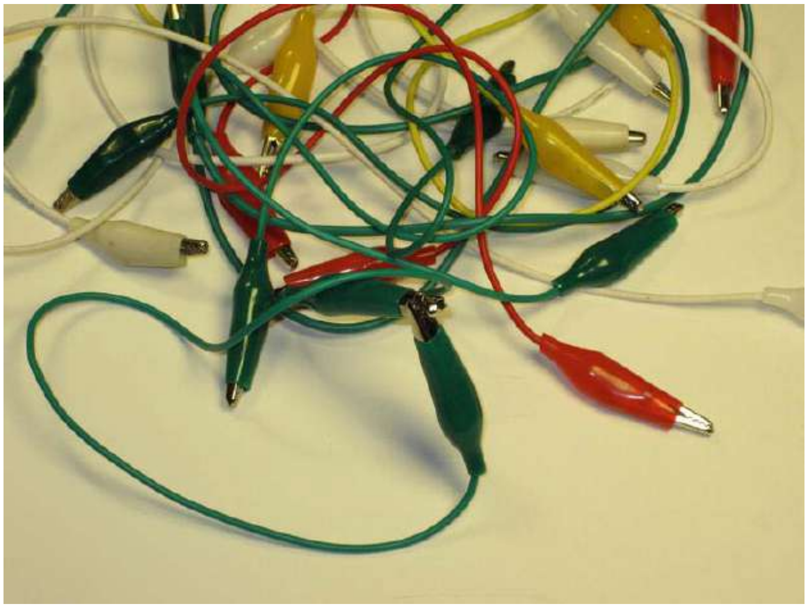

Figure 3-5. *Clip leads come in all sizes, gauges, colors, and terminal types. They are handy for making quick electrical connections.*

Clip leads are quite handy to have around your bench, if you've got the room for them. They tend to get tangled up, so some people hang them from a specially slotted rack, making them easier to pluck at a moment's notice.

Keep in mind that clip leads are generally for *temporary* electrical connections, such as when you're testing something or trying out a prototype circuit. They don't make very good long-term solutions to your electrical connection needs for a variety of reasons.

The first problem is that even when the spring-loaded jaws make a solid, low-impedance connection when first attached, this electrical connection will deteriorate over time as oxidation inevitably forms on the contacts.

Another problem is that many mass-produced clip leads are inexpertly smashed onto the ends of wires, without taking advantage of the alligator clips' built-in features, such as strain reliefs and soldering points. This leads to additional resistance in your circuit at both ends of the clip lead, where you would probably never think to look. The next time you buy some brand-new clip leads, hook them up to your ohmmeter and see just how conductive they really are. A good practice is to carefully remove the insulating boot and make a proper electrical connection yourself, so that you've got one less thing to worry about.

Solderless Breadboards

A very versatile prototyping platform is the solderless breadboard. Once upon a time, people would tack together simple electrical circuits using wires, small nails, and a hammer, utilizing a wooden plank or small board as the substrate. The kitchen breadboard was often just right in size and weight and would mysteriously disappear into the workshop. You can still cobble together simple circuits using this technique.

The solderless breadboard is a miracle of modern manufacturing techniques. A body made of precision-molded plastic has an array of small openings built in, and within each opening is a row of small spring clips that are just the right size to grasp and hold a small-gauge (22–26 AWG) wire. Adjacent points within the array are electrically connected together, which lets you easily connect small components into a circuit using your bare hands, with no soldering required. About all you would normally need would be some solid-core wire (preferably), some wire cutters, and some wire strippers. You can even buy precut jumper wires specifically made for use with solderless breadboards. They're terrifically handy! See Figure 3-6.

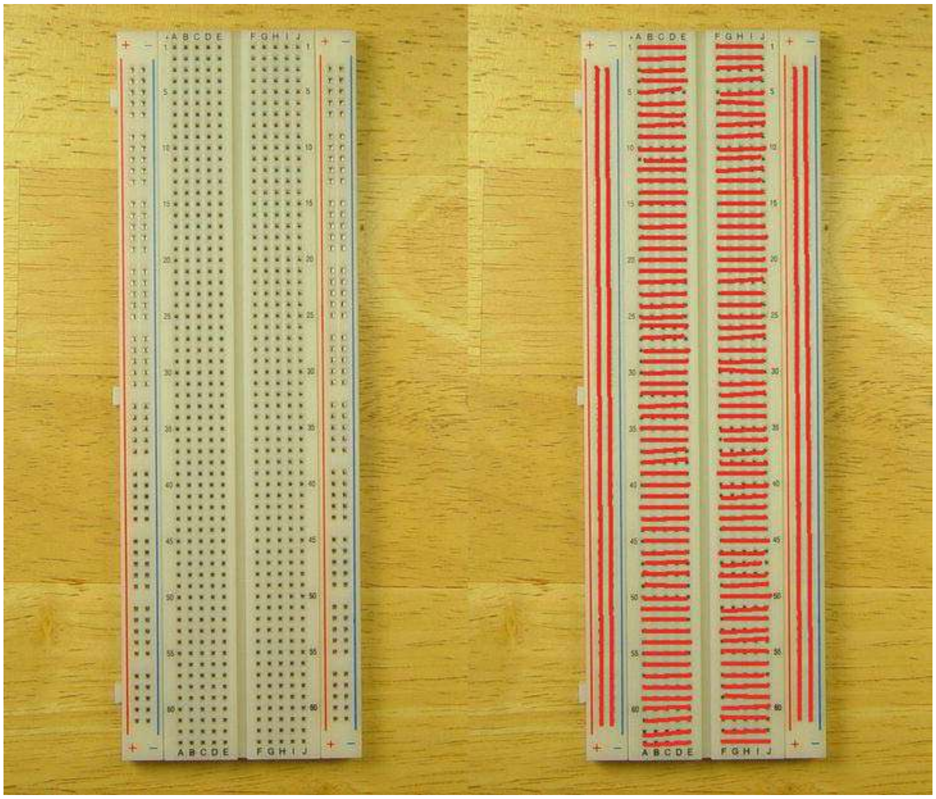

Figure 3-6. *On the left is a solderless breadboard with 830 tie points. Each hole has a spring clip inside that can hold a small wire or component lead. On the right, the underlying electrical connections between the tie points are illustrated.*

The spacing between adjacent tie points is exactly 0.1", which corresponds to many common "footprints" for electronic components. You can easily build simple circuits in just a few minutes. You can even build fairly complex circuits using integrated circuits and other components. Figure 3-7 shows a complete Arduino-compatible microcomputer that you can build with your bare hands and then program using your PC and the free Arduino software.

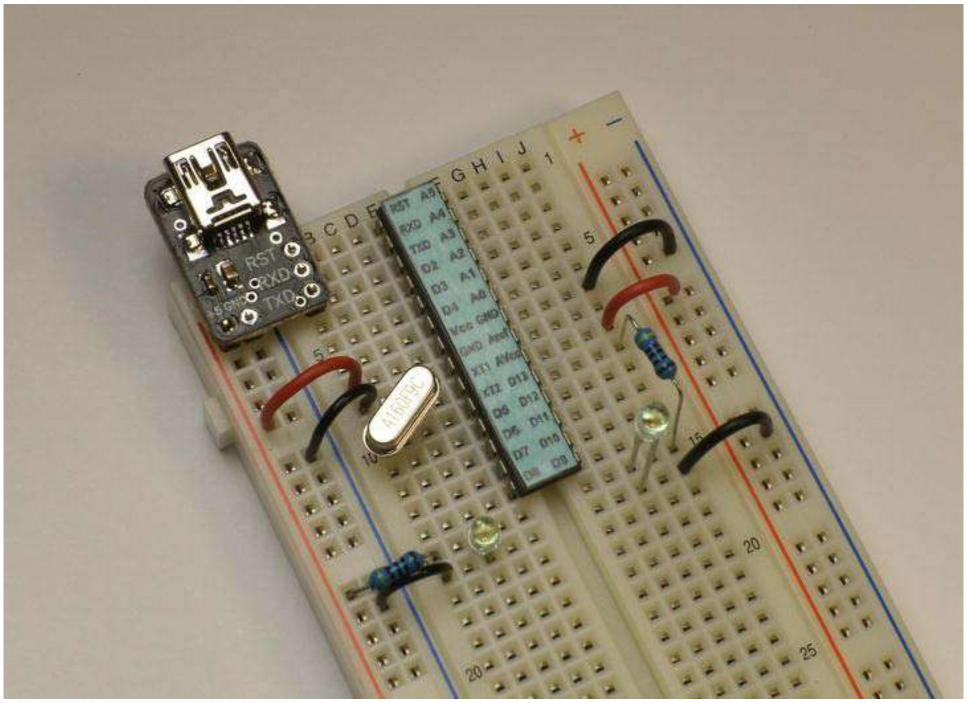

Figure 3-7. *A breadboard Arduino, which is compatible with the popular Arduino Duemilanove microcontroller, can be built using just a handful of components on a solderless breadboard, leaving plenty of room for more experimentation. The black, rectangular component in the upper-left corner is a miniature USB adapter that supplies power, and provides communication and control signals for the microcontroller.*

You'll find that projects tend to expand to fill the available area (and budget). See Figure 3-8.

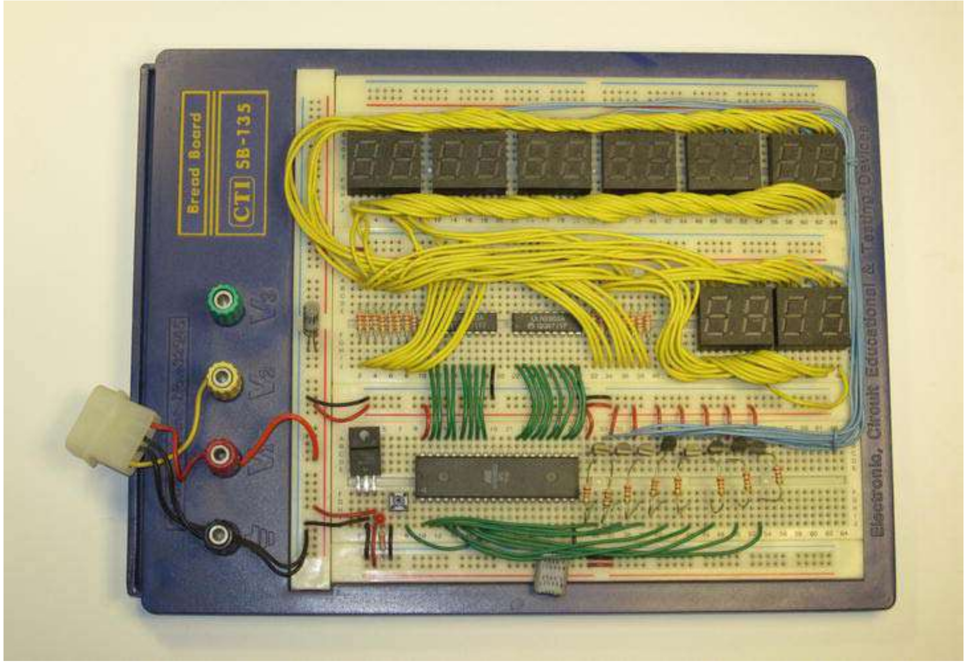

Figure 3-8. *A prototype LED display with computer control, built on a solderless breadboard. That's a lot of wires!*

Passive Components

Most electronic components fall into one of two different categories: *passive* and *active*. There are a few different ideas floating around as to what, exactly, is the difference between passive and active components. For the purposes of this book, we will use the simple definition that an *active* component requires a power source to work properly, while a *passive* component does not.

Resistors

A good example of a passive component is a *resistor*. You've already learned a lot about resistance in an electronic circuit. Resistance is measured in ohms, and it converts electricity into heat. When you're working with wires, cables, and other conductors, resistance is usually a *bad thing*. You want the least amount of resistance possible. However, there are times when you both want and need resistance in a circuit.

One example of when you would want resistance in a circuit is when you want to light up an LED. Unlike an incandescent lightbulb, an LED has no built-in current-limiting ability. Without a fixed amount of external resistance added to the circuit, an LED would self-destruct in a very short time, because it would overheat from excessive current flowing through it. A resistor would limit the amount of current flowing and control the brightness of the LED.

The exception to this dire prediction is when the power source that is lighting up the LED has insufficient gumption to blow up anything, such as with a small coin-cell battery. Anything bigger than that normally has no trouble convincing the LED to release its magic smoke, without which it can no longer function.

A typical resistor has a fixed resistance and a maximum power rating. The resistance is determined by the material from which the resistor is made. The value of a resistor, as you already know, is measured in ohms, and can range from just a fraction of an ohm (often expressed as milliohms) up to millions of ohms.

Through-Hole and Surface-Mount Resistors

Resistors come in all shapes and sizes. Generally speaking, the larger the surface area of the resistor, the more power it can safely dissipate. A typical resistor that you might work with on a solderless breadboard or on an older PCB is shown in Figure 3-9. These are called *through-hole* components because their extended leads are meant to be formed and threaded through holes that are drilled into PCBs. Also in Figure 3-9, for comparison, are some teensy-weensy *surface-mount* resistors. Surface-mount resistors have no leads and are meant to be soldered directly to pads laid out on the surface of the PCB.

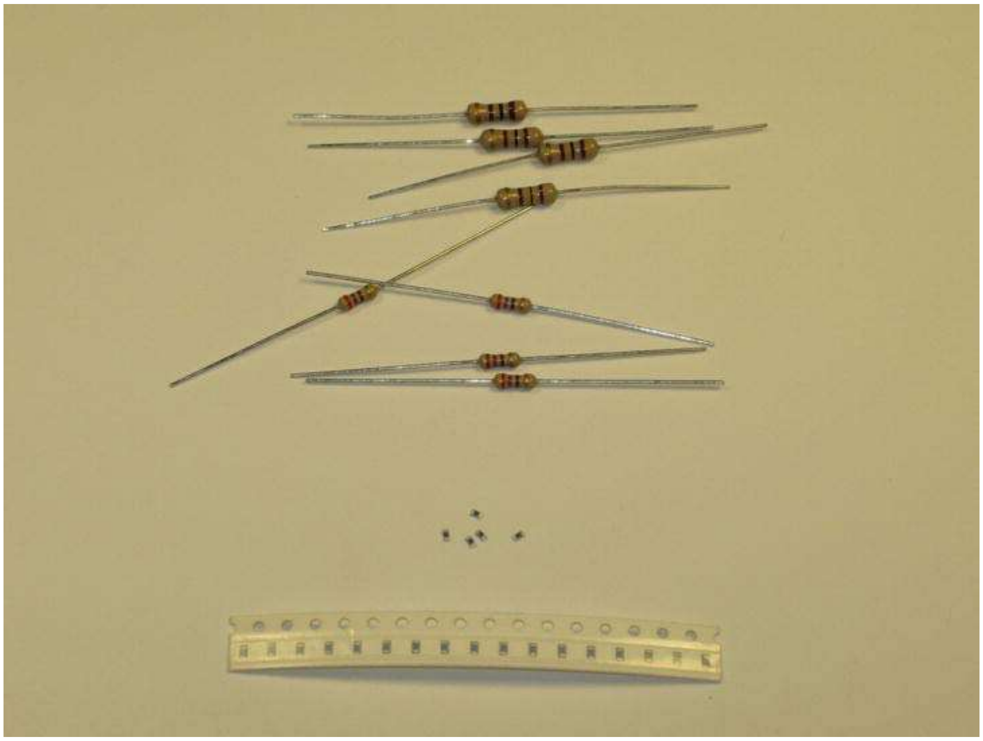

Figure 3-9. *Some through-hole resistors (top) and surface mount resistors (bottom). The surface mount resistors are so small that they are shipped packaged in tape, which is usually wound on reels for automated placement in PCB assembly.*

You can easily form the leads of through-hole resistors by hand or by using a *lead-bending jig*, a small tool with accurately spaced notches. Traditionally, these resistors have their leads formed so as to be threaded through holes in a PCB, and then soldered to the PCB. The extra lead length on the bottom of the PCB is then clipped off. See Figure 3-10.

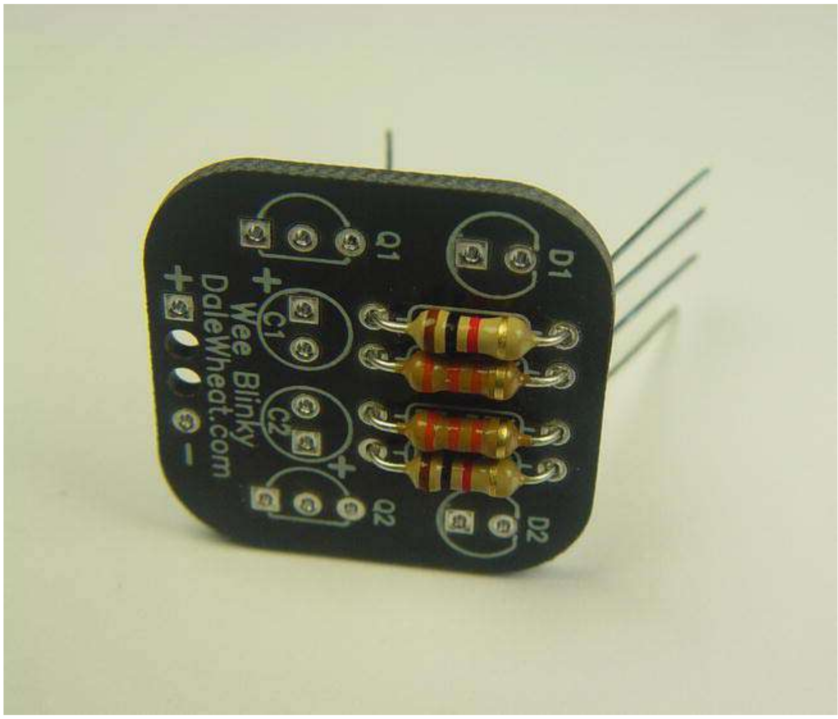

Figure 3-10. *Resistor leads are bent at right angles and inserted into holes in the PCB to be soldered.*

Schematic Symbols and Reference Designators

The schematic symbol for a fixed-value resistor is a zigzag line. See Figure 3-11. Every component in your circuit should have a name, which is also called a *reference designator*. Typically, most reference designators will have a short letter abbreviation to describe what kind of part it is (*R* is for *resistor*) and a sequential number assigned as more parts are added to the schematic. R1 would be the first resistor, R2 would be the second resistor, and so on.

If there is only one resistor in your circuit, then you don't have to number them, do you? You can just called it *R*, or *Galactic Emperor R the Mighty*, or whatever strikes your fancy. Most people stick to simple numbers.

Your numbering scheme depends a lot on the complexity of your schematic. If you only have a few resistors in your circuit, you can start at one and go up as needed. If you have a complex circuit with many subcircuits, you might allocate R101, R102, R103, and so on, to the first subcircuit (e.g., a power supply), and then assigned R201, R202, and so on, to a different section, such as a display. It's entirely up to you. Use a system that helps *you* keep track of things.

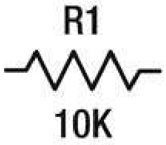

Figure 3-11. *The schematic symbol for a fixed-value resistor is a zigzag. The reference designator for this resistor is R1 and its value is 10,000 ohms, or 10KΩ. A simple rectangle symbol is also used to represent resistors, although many components are available in this physical form factor. Use the symbol you like best.*

Standard Resistor Values and Tolerances

Remember how wires have their own peculiar numbering scheme to describe their basic characteristics (i.e., the AWG, which uses *smaller* numbers to represent *larger* wires)? Well, resistors have their own peculiar numbering system as well to describe their electrical properties, or *value*. Even though the *value* of a resistor is measured in ohms and succinctly describes its characteristic resistance, not every possible value is readily available.

Some smart cookies at the Electronics Industries Alliance (EIA; see `http://www.ecaus.org/eia/site/index.html#`) have come up with some standard values that fill in most of the commonly used values for resistors and other components. Their scale is not linear but logarithmic, meaning that it takes larger steps as the values go up. This makes a little more sense once you understand how component value *tolerances* work.

You can buy, for example, a 1Ω resistor and expect it to be pretty close to being 1Ω at room temperature, but it's not going to be *exactly* 1Ω. It's going to be pretty close. How close? Well, that mostly depends on how much money you're willing to spend on this single resistor. If the answer is "not that much," then "pretty close" is going to mean within 5 percent of the stated value. That means your 1Ω resistor could actually measure as little as 0.95Ω or as much as 1.05Ω and still be "within tolerance." Five-percent resistors cost only a few cents even in small quantities.

Given that there's a little wiggle room for each resistor value, people who can tolerate their resistor values being off by as much as 5 percent don't need to be ordering 50 resistors of 0.99Ω and another 50 of 1.01Ω when they could just as easily (and certainly less expensively) order 100 resistors of 1Ω with a 5 percent tolerance.

This overlap in the values of the resistors allows us to skip from 1.0Ω to 1.1Ω without having to keep track of all the little resistors in between. This is because a 1.1Ω resistor could measure as little as 1.05Ω or as much as 1.15Ω and still be in tolerance.

With a standard resistor tolerance of 5 percent, we can use just 24 values to cover the range between 1Ω and 10Ω. The same multiples can then be used to cover the range between 10Ω and 100Ω, and so on. Aren't logarithms fun? See Table 3-2.

Table 3-2. *The E24 Standard Values for 5 Percent Resistor Value Tolerances (Higher and Lower Decades Are Also Available)*

1.0 10		100	1.0K	10K	100K	1.0M
1.1 11		110	1.1K	11K	110K	1.1M
1.2 12		120	1.2K	12K	120K	1.2M
1.3 13		130	1.3K	13K	130K	1.3M
1.5 15		150	1.5K	15K	150K	1.5M

1.6 16		160	1.6K	16K	160K	1.6M
1.8 18		180	1.8K	18K	180K	1.8M
2.0 20		200	2.0K	20K	200K	2.0M
2.2 22		220	2.2K	22K	220K	2.2M
2.4 24		240	2.4K	24K	240K	2.4M
2.7 27		270	2.7K	27K	270K	2.7M
3.0 30		300	3.0K	30K	300K	3.0M
3.3 33		330	3.3K	33K	330K	3.3M
3.6 36		360	3.6K	36K	360K	3.6M
3.9 39		390	3.9K	39K	390K	3.9M
4.3 43		430	4.3K	43K	430K	4.3M
4.7 47		470	4.7K	47K	470K	4.7M
5.1 51		510	5.1K	51K	510K	5.1M
5.6 56		560	5.6K	56K	560K	5.6M
6.2 62		620	6.2K	62K	620K	6.2M
6.8 68		680	6.8K	68K	680K	6.8M
7.5 75		750	7.5K	75K	750K	7.5M
8.2 82		820	8.2K	82K	820K	8.2M
9.1 91		910	9.1K	91K	910K	9.1M

Color Codes

Even the old-fashioned through-hole parts are a bit too small to legibly label with a numeric value and tolerance factor. A system of color codes was developed to designate resistor values, using colors to represent both single digits, a power-of-ten multiple, and the resistor's tolerance. These colors are painted on the resistors in bands. A little memorization and careful practice will have you reading resistor color codes in no time.

For 5 percent tolerance resistors, only four bands are needed to completely specify the resistor's value and tolerance. The first two bands represent the first two significant digits of the value. The next value is the multiplier. The final value is a special case to tell the resistor's tolerance.

The first two bands, as well as the multiplier value, use the color codes shown in Table 3-3 to represent the digits 0 through 9. (Unfortunately, these colors were chosen before the subtleties of colorblindness were widely understood.)

***Table 3-3.** The Resistor Color Codes for the Significant Digits and Multiplier Digit of the Resistor's Value*

Digit	Color
0	Black
1	Brown
2	Red
3	Orange
4	Yellow
5	Green
6	Blue
7	Violet
8	Gray
9	White

The last band indicates the tolerance of the resistor using a different set of colors, as shown in Table 3-4. The most common are 5 percent (gold) and 1 percent (brown). There is supposed to be a slightly larger gap between the digits and the tolerance bands, but this is sometimes hard to see.

Table 3-4. *Resistor Color Codes for the Final Color Band*

Tolerance Color	
0.05% Orange	
0.25% Blue	
0.5% Green	
1% Bro	wn
2% R	ed
5% Go	ld
10% Si	lver
20% No	ne

The multiplier digit can be thought of as either the number of 0s after the specified digits or as the exponent to which the number 10 is raised, as in scientific notation. Some examples make this easy to understand and remember.

A resistor marked with brown, black, red, and gold stripes has a resistance of 1KΩ with a tolerance of 5 percent. An easy way to decipher this code is to do a simple substitution for the first two digits (brown is 1 and black is 0, so we have 10 so far). Next decipher the multiplier band and write down that many 0s after the first two digits. Red is 2, so write down two 0s. That gives us 1000 so far. The final band is gold, and that means a 5 percent tolerance; 1,000 ohms is usually abbreviated as 1.0KΩ or simply 1KΩ.

■ **Note** *K* stands for *kilo*, the Greek prefix for 1,000. Likewise, *M* stands for *mega*, the prefix for 1 million.

This method works for most values from 10Ω to 10MΩ. A problem occurs, however, when we need to specify a value less than 10, as a negative multiplier value must be used to get an exponent of less than 0. For the value 1 and those less than 10Ω, a gold band is used for the multiplier. For values less than 1Ω, a silver band is used.

To save even more precious space on our schematics, another convention is to replace the decimal point with the multiplier. For example, 4.7KΩ would be written as just 4K7. 1.8MΩ would become 1M8 and 1.5Ω would be 1E5. 10KΩ would still be 10K using this method. If it is clearly understood that the unit of measure is the ohm, then even the Ω can be omitted.

Resistor Power Ratings

Any resistor has a maximum power-handling capability. This rating is usually expressed in watts or fractions of a watt. The larger through-hole resistors shown previously in Figure 3-9, for example, have a power rating of 1/2W (or 0.5W, if you prefer). This means that if you attempt to dissipate more than half a watt of power through the resistor, it is most likely going to quickly overheat and fail.

Thermal failure in a resistor starts with a thin curl of smoke, and then results in the discoloration of the resistor itself and sometimes actual flames. In any case, it gets dangerously hot, so do try to avoid this situation.

If you know the voltage across the resistor, as well as the current, you can easily calculate the power being dissipated using the simple formula P = V × I. Remember that P is the power value in watts, V is the voltage in volts, and I is the current in amps. One volt at one amp is one watt, which is way more than enough to smoke our little resistors. Be careful! Any time you calculate the correct value for a resistor in ohms, be sure to also calculate the required power-handling capability.

Variable Resistors

So far we've only looked at *fixed*-value resistors. They usually exhibit the same resistance day in and day out. That's what we love about them. But what if you need to vary the resistance in a circuit? We have resistors for that, too! A good example is a volume control on an amplifier. Turn it to the left (down), and the volume goes down. Turn it to the right (up), and the volume goes up. How does that work?

One kind of variable resistor is called a *potentiometer*. Potentiometers are sometimes abbreviated as *pots*. Pots usually have three terminals instead of two. Two of the terminals connect to the fixed resistor that is inside the potentiometer. Instead of being all sealed up like the fixed resistors you've been studying, the internal, resistive conductor is exposed along a track or groove. The third terminal, which is often physically in the center, between the two other terminals, is called the *wiper*, and can be moved from one end of the fixed resistor to the other end, making electrical connection to the inner part of the fixed resistor as it goes. As the wiper moves along the length of the resistive element, it taps into the resistor at a different point, varying the resistance. This is how a potentiometer can be used as a variable resistor.

Figure 3-12 shows a typical three-terminal potentiometer with a round body. The shaft can be turned to adjust the resistance. There are many variations on this basic design.

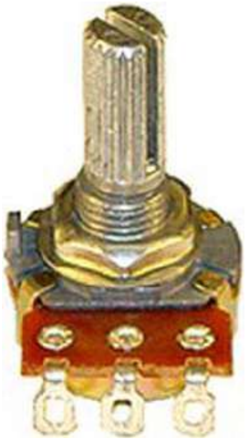

Figure 3-12. *A typical potentiometer, or variable resistor. Image by BG Micro (http://bgmicro.com).*

The schematic symbol for a potentiometer looks suspiciously like a resistor with an extra terminal, which happens to be exactly what it is. The extra terminal is the wiper, shown as the upward-pointing arrow. See Figure 3-13.

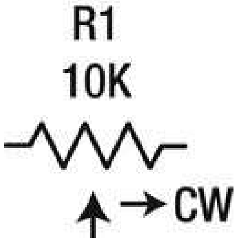

Figure 3-13. *The schematic symbol for a potentiometer is a fixed resistor with a wiper terminal. The "CW" arrow indicates which way the wiper goes when the shaft is rotated clockwise.*

Potentiometers generally come with two distinctive *response curves*, or *tapers*:

- A *linear* curve means that every equal step of rotation about the axis changes the resistance by the same amount. For example, if you were to center the shaft in the middle of its rotation, the resistance from the wiper contact to either of the end contacts would be nearly exactly half of the total resistance between the two end contacts. This is quite useful for when you need to adjust a voltage up or down in a relatively straight line. This is sometimes indicated by adding a capital *B* to the end of the resistance value (e.g., 10KΩB).
- In many audio circuits, however, it has been discovered that the human ear is anything *but* linear. A special response curve that more accurately describes this sensitivity is called an *audio* curve or taper. An *A* is used as a suffix to indicate that an audio taper is used for a potentiometer.

Note that linear and audio tapers are *not* interchangeable. If you find some random potentiometers in a parts bin, you can get a quick idea of what you've got by centering the shafts and seeing if the resistance is "symmetrical" or not, using an ohmmeter. Equal resistance to either end terminal indicates a linear taper, and anything else is mostly probably an audio taper.

Just like any other resistor, a potentiometer has a maximum power rating as well as a tolerance. These are usually either printed on the body of the potentiometer, if there is room, or hidden away in a product data sheet somewhere.

Specialty Resistors

There are many different kinds of resistors made for many special kinds of applications. High-power resistors are built out of sturdier materials that can take the thermal stress required to dissipate large amounts of heat. See Figure 13-14.

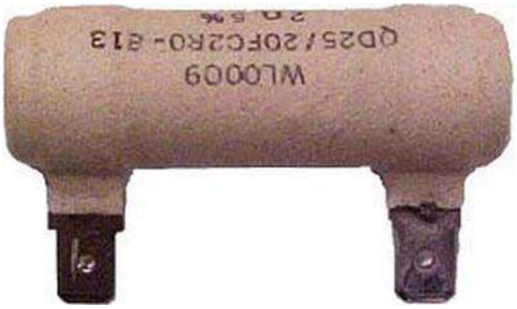

Figure 13-14. *A power resistor can handle much more heat dissipation than the smaller though-hole or surface-mount varieties. They are often so big that they need special mounting hardware. Image by BG Micro (http://bgmicro.com).*

Another popular option is the *multiturn* potentiometer. These potentiometers have an internal gear system that allows their shafts or wipers to be moved slowly across their internal resistive tracks by multiple turns of the control element. Increasing the number of turns allows the user much finer control of the resistance adjustment. It also takes the user longer to get from one point to another along the scale.

Capacitors

The basic definition of a *capacitor* is two conductors separated by an insulator. By itself, it doesn't sound like that would be a very interesting configuration at all. It turns out, however, that there's a lot more to it.

When a voltage is applied across a capacitor, no steady current flows through it. Remember, there is an *insulator* in the middle of that capacitor. Any insulator, just like the insulation on a piece of wire, *does not conduct electricity*.

What happens is that an *electric field* is created between the two conductors, or *plates*, as they are called when they are part of a capacitor. Now we get back to the fairies and dragons and molecules and so forth, way down deep at the heart of the problem. This electric field is a form of energy. Once it is built up by the difference in voltage, or *potential*, between the two conductors, it has nowhere to go, so there it stays . . . at least until something else in the circuit changes.

For this reason, capacitors are sometimes used as *reservoirs* of electrical energy in a circuit, holding a bit of spark here and there as needed. Just as it takes electrical energy to build up this electric field, a certain amount of electrical energy will be released when the field collapses. This will happen when the voltage between the two conductors drops.

Using capacitors as tiny electric batteries in a circuit is only one of many applications for this versatile part. While a capacitor will not conduct a *steady* flow of electric current between its terminals, it will often *appear* to conduct *alternating current* (as opposed to *direct current*), which is why you see them so often in audio circuits. As a charge builds up on one plate within the capacitor, it then attracts electrons from the other side, pulling them in from the other side. Even though they don't "jump the gap" of the central insulator, it looks like a tug-of-war is going on from both sides.

This behavior can also be used to send the higher-frequency components of a signal one way in a circuit and the lower-frequency components of a signal in another direction. This is one way that analog filters are built, and it has all kinds of useful and interesting applications.

Capacitors come in all shapes and sizes. They are made of many different materials, depending on what particular characteristics are needed. See Figure 3-15.

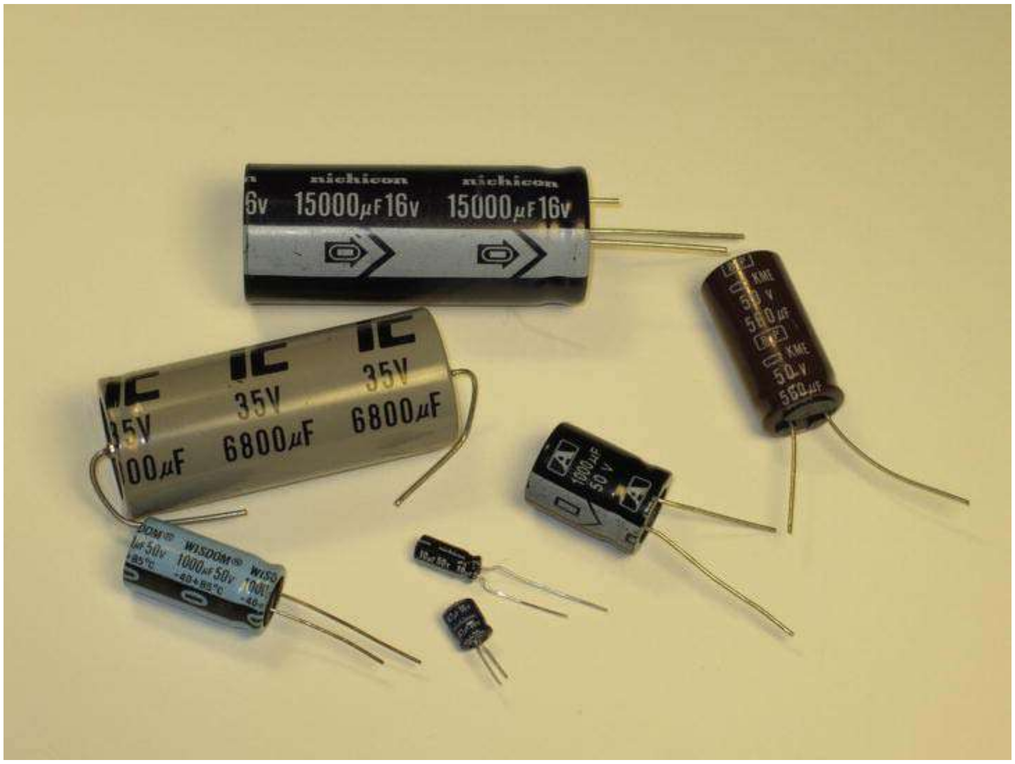

Figure 3-15. *Some electrolytic capacitors of various sizes and capacities. Both much larger and much smaller capacitors exist and are very common in electronics.*

Schematic Symbols

While there are a large number of special-purpose capacitors that are custom made for certain applications, a very large number of capacitors fall into two broad categories: polarized and nonpolarized. The schematic symbols for both types of capacitors display the fundamental construction of two conducting plates separated by an insulator. See Figure 3-16.

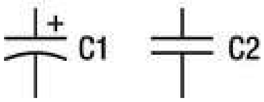

Figure 3-16. *The schematic symbols for polarized capacitors (left) and nonpolarized capacitors (right)*

C1 is a *polarized* capacitor, meaning one terminal is more positive than the other. In practice, this is indicated on the part itself in a variety of ways. One convention is that the *anode*, or more positive lead, is physically a little bit longer than the *cathode*, or negative lead. There is generally some sort of mark made on the capacitor body itself. Usually this is a stripe on the *negative* side of the capacitor.

What's really odd and potentially confusing is that one type of capacitor, specifically that made primarily of tantalum, has a stripe that indicates the *positive* lead of the polarized capacitor. Know your capacitors!

A polarized capacitor that is installed backward will experience terrible physical and thermal stresses and can sometimes explode, sending bits of itself all over the place.

C2 in Figure 3-16 illustrates a *nonpolarized* capacitor. It doesn't matter which way it is installed in a circuit. It is electrically symmetrical and works just as well in either orientation.

Inductors

An inductor is almost always made out of coils of wire. Sometimes inductors are quite large, as with transformers, while other times they are as small as a speck. An inductor would appear, electrically, as a simple short-circuit if it weren't configured to take advantage of certain principles of *electromagnetism*, or the tendency for the flow of electric current in a conductor to form a *magnetic field* around the conductor. The shape and orientation of the loops of wire that compose the inductor determine its electrical characteristics.

In many ways, an inductor is the opposite or complement of the capacitor. While a capacitor does not conduct direct current from one terminal to another, an inductor will happily do so, at least for a time. While a capacitor will resist a change in voltage across its terminal as it builds up its electric field, an inductor will resist a change in current flow across its terminal once it has established a magnetic field.

Because of their complementary natures, the two components can be used together to form a *tuned circuit*—one that resonates at a particular frequency, or allows the passage of only a certain range of frequencies, with the capacitor shuffling the high-frequency portions and blocking the low-frequency portions, while the inductor does the opposite, blocking the high-frequency components and passing the low-frequency ones. This is how many older radios and televisions worked before the advent of low-cost solid-state electronics.

Once an inductor has built up a magnetic field, any drop in the current flow across its terminals will result in the collapse of the field and the generation of a very-high-voltage spike. This is the basic electrical circuit that was used in automobile engines to create the high-voltage spark needed to ignite the fuel-and-air mixture in the combustion chamber.

Inductors can also be used to efficiently pass alternating current via a magnetic field from one inductor to another. This is how the primary windings of a power transformer *induce* a current in the neighboring secondary coil, which, depending on the ratio of windings and a few other factors, determines if the transformer *steps up* the voltage or *steps down* the voltage. It can even keep the voltage the same, for when electrical isolation is needed between two different systems.

Most electric motors employ some sort of inductor that alternately attracts and repels another electromagnet or permanent magnet. The most basic direct-current (DC) motors require a fairly complex mechanical arrangement to perform the necessary switch-swap to alternate the fields. This process is called *commutation*. In other types of electric motors, such as brushless motors and stepper motors, this commutation process is done electronically, externally to the motor itself, and requires more complex circuitry to control the speed and direction of the motors. See Figure 3-17.

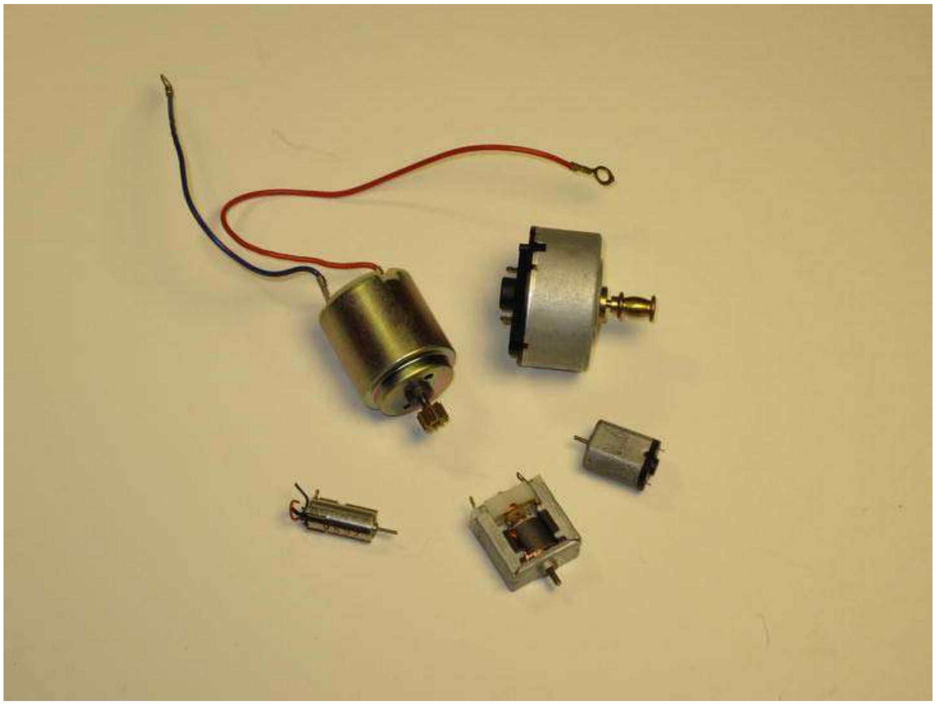

Figure 3-17. *Some simple DC motors. When direct current is applied to their terminals, they spin. The magic happens with an electrical and mechanical arrangement inside that performs the necessary commutation of the magnetic fields.*

Switches

The most basic way to control the flow of electric current in a circuit it to literally *break* the circuit. This means that instead of forming a conductive loop from power source to components and back again, a gap is introduced somewhere in the loop.

This gap can come anywhere in the loop. It can be right next to the battery, for example, or somewhere off in the middle—as long as the circle is broken, no current will flow.

If you perform the lightbulb experiment in Chapter 1, you can see how this works. When everything is lined up properly and the final connection is made, the lightbulb lights up. When any of the connections are broken, the light goes out. It's really as simple as that.

Most of the electrical switches that you operate every day are really mechanical in nature. Their main function is to physically move two conductors either together or apart.

In the lightbulb experiment, you could use the wire as a switch, by physically touching either the battery terminal or the lightbulb (depending on how you were holding everything together). With a little practice, you could probably send Morse code, or at least a variation on "one if by land, two if by sea."

There are many kinds of switches for many kinds of switching tasks. The most basic are *momentary action* switches. These are usually push buttons, and only conduct electricity when pressed. Alternately, they can be built so that they conduct electricity all the time, *except* when pressed:

- The first option is called *normally open*. When a circuit is open, it is not conducting electricity. When you *close* the circuit, by completing the loop, electricity can flow. A normally open switch would be welcome in a doorbell circuit. When no one is pressing the button, no bells are ringing. It would be annoying the other way around.
- A *normally closed* switch acts like a solid piece of wire (being a conductor of electricity) until it is activated, when it ceases to provide a path for the electricity to follow. This type of switch can be used as a *kill switch* on a gasoline-powered engine, such as a lawn mower, for example. By interrupting the flow of electricity generated by the magneto and powering the spark plug, you interrupt the internal combustion cycle, which will eventually cause the engine to slow down and then stop.

This type of switch is called a *single-throw* switch. It's either on or it's off—conducting electricity or not conducting electricity.

■ **Note** For almost every possible configuration of switches, there is a corresponding *opposite* configuration that is almost as popular.

Another useful configuration for a switch is to switch from one circuit to another circuit, and not just on and off. This is referred to as a *double-throw* circuit. A switch like this normally has three terminals. A *common* terminal is usually, but not always, in the center. At any rate, the common terminal is electrically connected to the moving part of the switch that is going to make contact with one or the other two remaining contacts. The *normally closed* contact is the one that will be making contact with the common terminal when the switch is not activated, or at rest. The *normally open* contact is simply the other terminal. This assumes that the switch has some sort of bias or mechanical spring action that returns it to a certain orientation when no one is activating it.

Some double-throw switches can remain in either state indefinitely. A typical light switch used in indoor lighting is a good example. It's just as happy to stay on as off.

Switches can have more than one circuit built into them. These switches are called *double pole*, *triple pole*, and so forth. They're like two or more single-pole switches physically taped together, so that when one is switched, they all are switched. Usually these poles are electrically isolated from each other.

We'll play with switches again in Chapter 5, where you'll learn some interesting ways to wire them up.

Active Components

While it could truthfully be said that *all* electronic components require some electricity to do anything, the distinction made in this book between *passive* and *active* components is that *active* components need a little more power to be really interesting.

First we'll look at sources of this magical electricity, and then we'll move on to some of the fascinating and complex components that will happily use it all up for you.

Power Supplies

You're going to need some sort of power source for any electric or electronic circuit you run across. If it doesn't use electricity in some structured form, it's not really electronics, now is it?

Batteries

You can use regular batteries to power a lot of your circuits, but they eventually wear out, and then you have no real recourse other than to throw them away or recycle them, if possible. If you're going to be using batteries, at least get some good ones. Many of the bargain brands are terribly inferior in their manufacture and are only providing you with a false economy. Spend a little more and get higher-quality parts, and you will be rewarded with longer life and more reliable operation.

A great way to work with batteries is to use the appropriate battery connectors or battery holders. Sometimes you can remove these from obsolete electronic devices that have outlived their useful lifetime. Always observe the proper orientation and polarity of batteries. Even smaller batteries contain an annoying amount of power, especially when it all decides to come out at once, such as when you accidentally plug something in backward or short something out. This can result in damaged parts, burns, or even fires. Be careful with batteries.

Having a fuse holder wired inline with your battery-powered supply is an excellent and cheap way to prevent such "unauthorized thermal events." Keep several spare fuses on hand as well. Fuses cost little and save much.

Rechargeable batteries are becoming more common, and they keep coming down in price while improving in longevity and power capacity, and decreasing in size and weight. Any rechargeable battery technology beats any kind of primary, or nonrechargeable, battery in the long run.

■ **Caution** The same warnings about battery polarity and fuses apply to rechargeable batteries as well.

Since there's no easy way to "turn off" a battery, it would be a good idea to include a power switch in your supply, if possible. Just make sure that whatever switch you select can stand up to the maximum current you will want to be extracting from your power supply. You should also make your switch large and easy to see, so that you can find it quickly when you need it. An obvious power-on indicator is also a great plus for a homemade power supply. This can be as simple as a green LED (with appropriate current-limiting resistor—see below) or as complex as a voltage meter indicating how much juice is left.

Transformers

Another great power supply that you probably already own is a power adapter or transformer from another electronic appliance. You should use caution, of course, with anything that plugs into the wall for power, especially if you're planning on monkeying around with the other end of it.

Almost all commercially produced power adapters have a slew of certifications plastered on them. Somewhere in there should be the basic data concerning what type of power input it expects, what it is going to try to produce, and what the maximum limits are on the output. Make sure that whatever it is that you're going to plug into your wall is specifically designed to plug into *your* wall, as electrical power standards vary from one location to another.

Make sure that any repurposed power supply you use does not get overly warm when in use. This is an indication that it is presently or has been overloaded in the past. Such a device is a fire-starter waiting for its magic moment.

You need to know how much voltage is coming out of your power supply. Even if it is clearly printed on the device itself, if you have access to a voltmeter, you need to verify the output voltage before connecting it to anything. Many lower-cost power adapters are *unregulated* and will produce an output voltage that will vary wildly depending on the load applied to it. A *regulated* voltage output, on the other hand, will provide consistent voltage levels on the output, even with varying loads.

USB adapters and chargers are becoming quite common and can provide *regulated* +5V at up to 1/2A and sometimes more. These are great devices to use for small electronic experiments. They are generally small, lightweight, and reliable.

A dedicated bench power supply is a very handy piece of test and measurement gear to have available in your lab. A good supply will have a variable voltage and a readout to show you what's going on (or coming out). A step up and you get to set the maximum current that it will supply (to keep things from getting out of control too fast). Additional desirable features include multiple outputs and remote control from your PC.

No matter what kind of power supply you use, use it responsibly. Don't leave equipment powered up if you're not using it. Don't attempt to bypass safety measures that are in place for *your* protection.

Diodes

A diode is a semiconductor device that will conduct current in one direction only. That's the simple story. Of course, the reality is much more complex. A typical diode is composed of two pieces of a semiconductor material, usually silicon, but sometimes germanium or some other exotic substance. The diode action occurs in the *junction* between the two types of semiconductors, which have been specially manufactured to contain the exact right balance of impurities required. This is generally not the kind of thing you can build in your lab, unless you're Jeri Ellsworth.

These two bits of semiconductor, called the *N region* and the *P region* (for negative and positive, respectively), are sealed up in either a glass or plastic enclosure. A standard diode has two terminals—an *anode* and a *cathode*—which represent the positive and negative leads, again respectively. Conventional current (i.e., current flowing from the more positive voltage to the more negative voltage) can flow through the diode from the anode to the cathode, but not the other way around. It's magic! Well, technically it's semiconductor physics, but either answer will serve for our purposes.

Diodes come in all sizes. Generally speaking, the more power a diode is expected to handle, the larger it is going to be. *Small-signal* diodes are the same size or smaller than typical resistors. Power diodes, often called *rectifiers*, depending on their application, can be so large they have to be mounted with large bolts.

The two most important electrical characteristics for the happy care and feeding of diodes are their maximum reverse voltage and forward current capability. You don't want to exceed either one.

Since a diode is, at best, only ever going to have a cryptic part number written on it, you'll need to consult the manufacturer's data sheet to find out all the necessary information. Many semiconductor producers offer "commodity" parts that conform to standard part numbering schemes. That way, you know that a 2N4148 diode from one manufacturer is going to be pretty close to a 2N4148 diode from any other manufacturer, but you should always check to make sure.

Diodes typically have a *forward voltage* that must be exceeded before current will flow. This voltage is essentially lost and is dissipated as waste heat from the device. Silicon diodes will have an average forward voltage around 0.7V and up. Germanium diodes will have a smaller forward voltage, in the area of 0.3V. Schottky diodes use a metal-to-semiconductor interface region and achieve substantially lower voltage drops, making them much more suitable for higher-power applications.

LEDs

One special kind of diode is the popular *light-emitting diode*. You see these little guys everywhere these days. In fact, some folks go around with black tape, just trying to cover them all up so they can get their darkness back at night.

The reason that they are everywhere is that they are cheap, reliable, and easy to use. They don't take up much room, don't take up much power, and don't generate that much heat. The exceptions to these admirable qualities, of course, are in the newer generation of LED lighting products that emit tremendous amounts of light and produce tremendous amounts of heat. Every month it seems that a new record is set for efficiency. Let's hope that trend continues.

Being essentially diodes, LEDs have both an anode and a cathode. LEDs only conduct current in a single direction. You'll need to make sure you limit the amount of current flowing through your LEDs, because the LEDs themselves sure aren't going to do it.

LEDs designed to be indicator lights can be powered with as little as 1mA. High-efficiency LEDs can be used for illumination purposes and still only draw 20–30mA each. These very modest power requirements make LEDs a favorite for microcontroller projects, as these output power levels are easily handled by most popular chips.

The forward voltage drop across LEDs is typically much higher than for a rectifier or signal diode. In fact, the voltage goes up as the emitted color of the LED goes from the infrared and red end of the spectrum (1.2–2V) up to the blue and ultraviolet end of the spectrum (3–3.6V).

Transistors

A true modern miracle, the semiconductor transistor ushered in the age of solid-state electronics. More than a million individual transistors can be used to build modern computer processors, all on a single chip.

The first types of transistors to be built and used were called *bipolar junction transistors (BJTs)*. This type of transistor is similar in construction to the diode, except that there are two semiconductor junction areas inside.

This arrangement of magic beans allows a small amount of electrical current to be applied to one of the three terminals of the transistor, called the *base*, which in turn allows a much greater amount of electrical current to flow between the remaining two terminals, known as the *collector* and *emitter*.

Two complementary types of BJTs are made, depending on the way that the positive and negative semiconductors are sandwiched together. NPN transistors will conduct conventional current from the collector to the emitter when current flows from the base to the emitter. The opposite variety, called the PNP transistor, does precisely the opposite, conducting conventional current from the emitter to the collector when the base is negatively biased.

Transistors can be used to make amplifiers, switches, current limiters, voltage regulators, oscillators, and any number of other useful circuits. They are generally very low cost, lightweight, and reliable.

As with any other electrical components, you need to be aware of the electrical limitations of your transistors. Even when a transistor is fully saturated and conducting very efficiently, it's still dissipating a certain amount of its energy as heat. Too much heat and say goodbye to your transistor.

A newer type of transistor is becoming quite popular in power circuit applications, and uses a different type of semiconductor property to accomplish its mission. The *MOSFET (metal-oxide–semiconductor field-effect transistor)* has three terminals, just like the BJT. Each terminal performs a similar job, but does it in a different way. The *gate* of a MOSFET is electrically isolated from the conduction *channel* between the *drain* and *source* terminals. However, when a proper voltage is applied to the gate, it creates an electrical field that influences the conductivity of the channel between the drain and the source.

The important differences between BJTs and MOSFETs lie in the type of signal needed to drive each of the transistors and how much power is wasted by being converted to heat. A BJT is a current-controlled device, where more current into (or out of) the base results in more current flow through the device. The MOSFET is a voltage-controlled device, where more voltage applied to the gate increases the current flowing through the channel.

Additionally, MOSFETs typically have much lower on-state resistance than BJTs, meaning they can more efficiently switch higher currents. Finally, MOSFETs have a *positive temperature coefficient*, meaning that they conduct less power as they heat up—which is a good thing. BJTs are subject to *thermal runaway*—conducting more current as they heat up, which in turn generates more heat, and so on, until something snaps, crackles, or pops.

Integrated Circuits

As parts get more specialized, they become less fun. There are a hundred ways to wire up some transistors, resistors, and capacitors, but there's generally only one way to hook up a thermocouple interface chip.

The good news is that it has become much easier for companies to design and build vastly complex chips that can do things for us that, honestly, most people cannot comprehend. And that's OK. We need airplanes that can, mostly, fly themselves. If we had to depend on humans to get everything done, you can imagine what kind of world we'd live in. Humans are notoriously sleepy, hungry, distracted, and forgetful. A voltage comparator chip does nothing but compare two voltages, all day, every day, perfectly, all the time.

Integrated circuits come in an unbelievably wide range of functions, packages, and capabilities. The first integrated circuits were amplifiers. Integrated circuits allowed manufacturers to make chips that have dozens or hundreds of transistors and other parts *all at once*. Just plop that little silicon chip into a plastic package with legs on it, and you're done.

What kind of chips do you need in your lab? Well, that's going to depend heavily on what kind of circuits you're wanting to design, build, or repair.

Many older chips had a standard 0.1" pin spacing that made them easy to plug into a solderless breadboard for prototyping purposes. Most newer chips have been considerably shrunk, because phones and tablet computers are all the rage these days. It's still possible to work with these newer chips, but it almost always requires mounting them to some sort of breakout board or expander that brings out their tiny leads to a more reasonable, and reachable, size.

Acquiring Components

Where do components come from? The same place as everything else—the store—just like food, just like money.

The electronic component manufacturing industry spans the globe. If someone's not putting something electronic together, they're taking something electronic apart.

Once upon a time it took some hunting to find sources for electronic parts. Swap meets and "ham fests" were great places to find components, assemblies, and entire appliances just ready to be hacked, studied, improved, or rebuilt. Now you have to spend time looking for places to *get away* from all the electronic hustle and bustle.

One of the great things about today's Internet is that it takes almost no effort to find both electronic parts and the information needed to use them. The real trick these days is finding the best deal.

If you're strictly a hobbyist, looking to spend some quality time alone or with your kids, learning about electronics and building fun projects, then it just about doesn't matter if you get the absolute, rock-bottom price on every component and tool. Even if you're a budding entrepreneur, you're still in a much better position today to leverage those brilliant ideas floating around in that noggin of yours than you would have been even ten years ago, all because of the ubiquitous nature of electronics and technology saturating our lives. Not only can you get the right tools and the best parts shipped right to your door for a song, you can easily network with knowledgeable folk from all over the globe, contributing ideas and brainstorming across time zones and borders.

Buying New Components

It would be more than a full-time job just keeping up with all of the new tools, parts, software, and hardware that are available today. Even if you could, it would take many lifetimes to make and grow the connections and relationships you need to successfully compete in the global economy.

That being said, a good place to start is in your own backyard. If at all possible, you should deal locally for the things you need. When what you want or need cannot be obtained locally, expand outward until you find it.

While it's certainly nice to be able to find all your needs met close to home, it's also good to know that the world is shrinking every day. The sort of deals and relationships that couldn't possibly have existed in your parents' time are happening right now.

If you are in a position to spend some money on quality tools and components, then by all means do so. You will not regret it. Buying questionable items is like buying your groceries out of some guy's van in a parking lot. Maybe he'll be there next week and maybe he won't.

Even the small fry can get treated like an honored guest when dealing with the right people. The trick is finding the right people. The best way to find them is to ask your friends and others with similar interest where they trade. Ask them how they feel about their relationships with their vendors. Are they just a number? Or are they a partner?

Recycling Used Components

It's really a bad idea to waste anything. In a more perfect world, there'd be no "other" category at the recyclers. Everything would have an exit strategy, especially the complex and fragile toys we're so addicted to these days.

Until that day, try to look at your trash with new eyes. Is it conceivable that someone else might be able to use this equipment, or is it truly past its useful life? Or are you just being lazy?

You can do your part, and you should believe that others are doing their part as well. Take the time to find them, find out what they can offer you, and find out what you have that they might be able to use. You won't know until you ask.

Be careful when harvesting components with which you are unfamiliar. Novice recyclers should give yesterday's cathode ray tube (CRT) TVs and monitors a wide berth. They can retain *lethal* voltages for considerable periods of time after being retired from service. Just leave them on the side of the road. The same goes for microwave ovens, or any equipment that has refrigerant in it. The stuff they use nowadays is quite toxic.

There is a growing surplus of both components and “last year’s models” available to the student of electronics these days. Sometimes it’s worth your time and money just to take something apart to see how it works. Then sometimes you find some useful parts inside that you can use in projects of your own.

Component Data Sheets

Every component ever manufactured has a data sheet. This data sheet contains everything you ever wanted or needed to know about this particular component. It’s probably available right now on the manufacturer’s web site, along with a dozen application notes. If not, then it’s probably available on a university’s computer system in a project folder.

This isn’t always the case, sadly. Past a certain point, it doesn’t make sense to allocate resources to information that no one wants or needs anymore. That point varies with each company, with some diligently archiving their information and others tossing it out the window at the end of the quarter.

When possible, collect and maintain your own private set of data sheets. When you acquire a new piece of equipment, get everything you can find in electronic form and stash it away somewhere where you can find it when you need it. If you’re experimenting with some new components, squirrel away those data sheets, just in case you strike gold.

Have you built something cool? Write up your own data sheet for it! Pretend that you’re still going to be interested in this project next year. Better yet, imagine that *someone else* could take an interest in your work. You’ll be pleasantly surprised at how useful you’ll find this information in the future, after you’ve slept a bit.

Summary

This chapter has covered a lot of different types of electronic components, but it has only scratched the surface of what is presently available.

It’s perfectly possible to spend a lifetime pleasantly tinkering with electronics and never have a clue how any of the underlying principles work. Sometimes all you need is a *practical* understanding of what the different components do in a circuit. Hopefully this chapter has given you a rough outline of what is involved. Ideally, your interest has been sparked enough to pursue more detailed study elsewhere. Good luck and happy discovering!

Now that you’re armed with some basic tools and some ideas about some of the available components, it’s time to get busy building your own lab, one experiment at a time.

The remainder of this book will take you on a journey of trial, error, and experimentation that will help you build the best possible lab for your needs, both for now and in the future.

CHAPTER 4

A Portable Mini-Lab

Having a portable (or at least (*trans*portable) electronics lab comes in handy in several circumstances. Maybe you don't have a place (yet) for a permanent home for all that equipment. Maybe you need to bounce at a moment's notice, heading off on electronic adventures at the drop of a hat. Or maybe you like to keep everything where you can find it in a hurry, without having to rummage through shelves and boxes, looking for just the right tool or component.

Whatever your motivation for wanting a portable lab at your disposal, this chapter should help you get started.

Have Lab, Will Travel

First, a note or two is in order concerning *travelling* with your electronics lab. Nothing makes security screeners at an airport or customs inspectors at border crossings more nervous than a bunch of mysterious devices, especially devices that are tangled up in a rat's nest of wires.

Do what you can to keep your collection of tools and components tidy and in working order. You will most likely be asked to at least turn on any questionable items and demonstrate that they are not hollowed-out shells for transporting questionable items. It also doesn't hurt to clearly label equipment as to its function and ownership, which makes sense in any case, not just travel scenarios.

The Compact Executive Model

"A full mongoose is a slow mongoose." This is an important lesson learned from Rudyard Kipling's *Rikki-Tikki-Tavi*. Rikki, a mongoose, has to stay fast on his feet to be able to combat Nag the cobra and hope to win. Therefore, Rikki learns to only take a few bites at a time when eating and never gorge himself.

Although hopefully not in a life-or-death situation like Rikki's, you can learn a thing or two from this simple philosophy. Only take what you absolutely need when travelling, and never take something you can obtain at your destination. These things will only slow you down. Don't be a slow mongoose.

■ **Tip** Travel light.

A slim and lightweight briefcase or attaché can hold many of the tools and components you might need for your mobile electronic experimentations. The main advantage is that you will only need one hand to carry it (leaving one hand free to open doors, bribe guards, or fend off ninja—assuming you have two hands to begin with). See Figure 4-1.

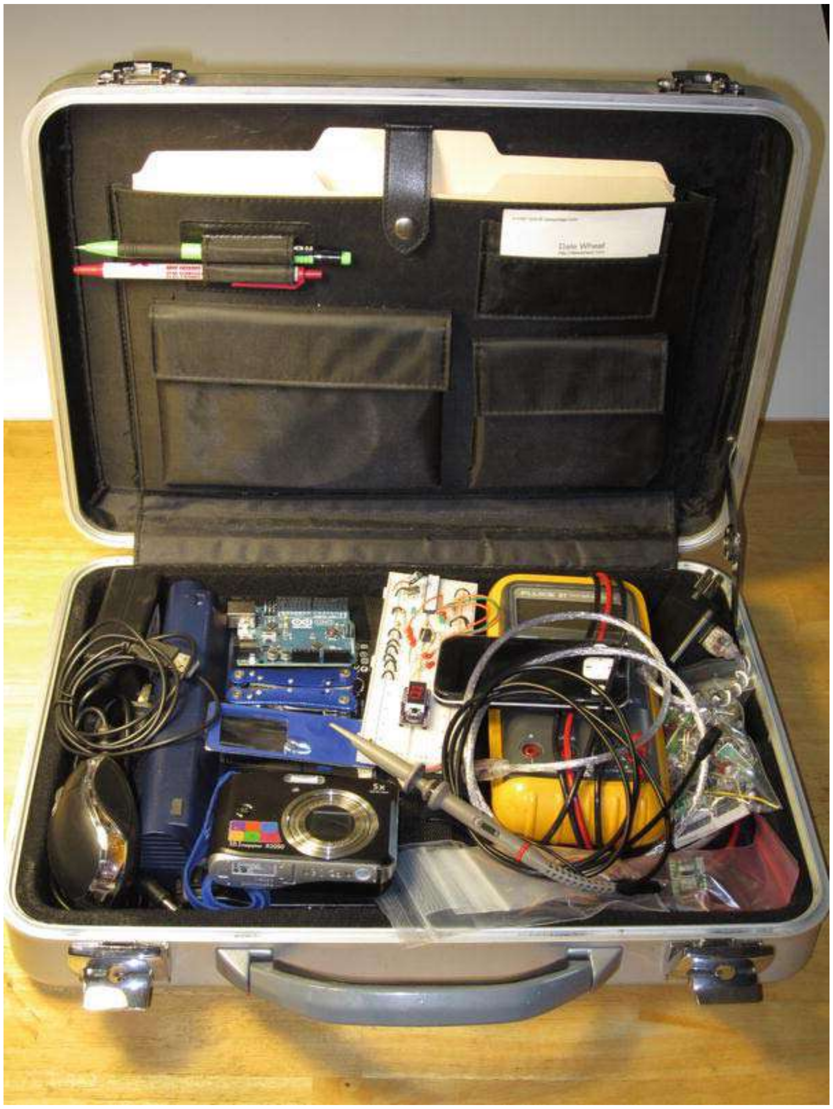

Figure 4-1. *The compact executive mobile electronics laboratory. There's plenty of room for lots of tiny things, but not much else. Also handy for transporting big stacks of cash, should the situation arise.*

Let's take a look at what all you can take with you without a lot of heavy lifting. Starting on the left and moving to the right, we have

- Netbook PC with AC charger and USB mouse (it might be hard to see in Figure 4-1 because it's upside down and covered by other stuff)
- Arduino Uno programmable microcontroller (or *physical computing platform*)
- Folding multitool
- Lighted magnifier
- Digital camera
- Solderless breadboard with lots of precut jumper wires
- Digital multimeter and probes
- Compact single-channel oscilloscope with probe
- AC-powered USB charger—can also be used as a stand-alone regulated 5V power supply
- USB cable
- A variety of LEDs in all shapes and sizes
- Additional passive components
- A variety of batteries, both rechargeable and otherwise, and various battery holders (not shown)
- Static-sensitive integrated circuits and modules stored in antistatic bags
- Notebooks, writing implements, blank media, and documentation
- Your contact information in case of loss

One of the interesting features of this lab is that it is completely self-contained, power-wise. While it's nice to have access to AC power, it's not strictly required, at least for a while. The Arduino and the attached breadboard are powered by the battery in the netbook, via the USB cable. The multimeter has its own 9V battery, while the mini-oscilloscope can be recharged from a spare USB connector. With an extended-life battery, you can expect four to five hours of on-location electronic experimentation.

If your netbook or PC has Wi-Fi or other connectivity to the Internet, you can browse tutorials, data sheets, user forums, and even this book, unless you've already downloaded these items to the internal hard drive.

Have a look at this particular lab deployed and ready for action in Figure 4-2.

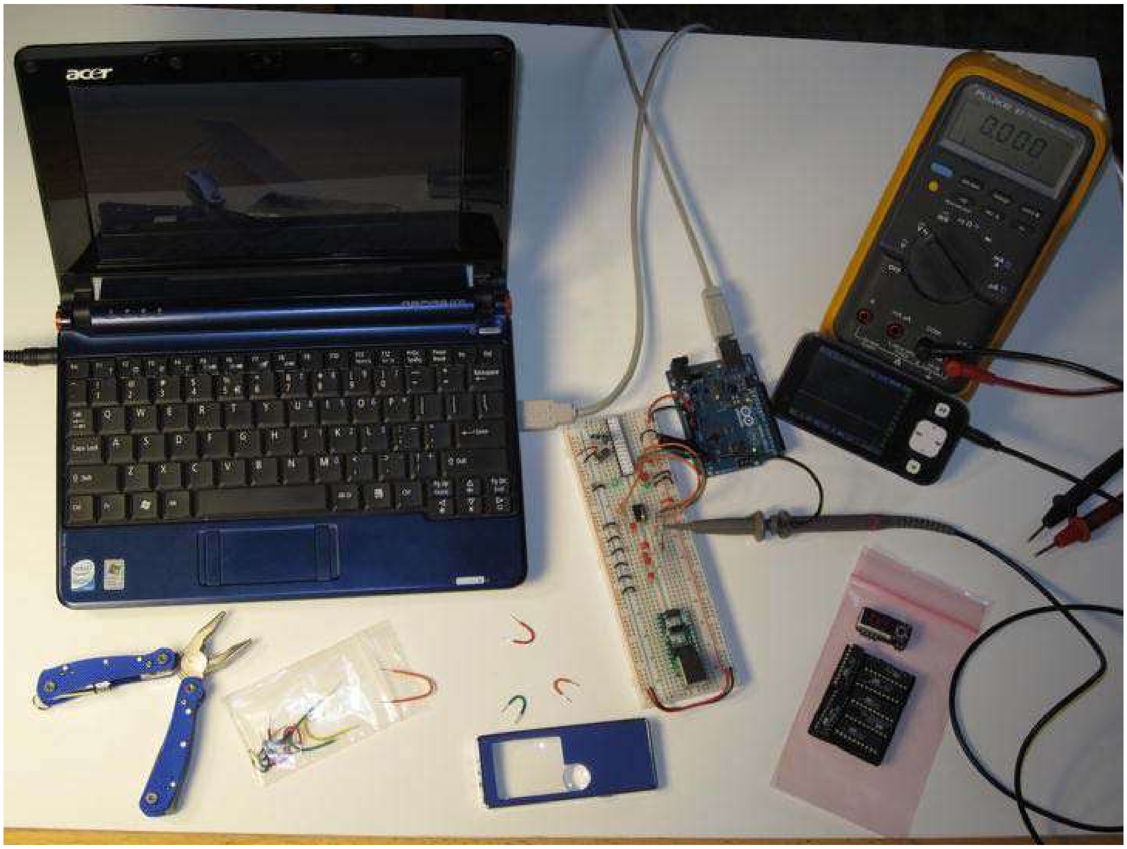

Figure 4-2. *The compact executive lab is deployed on a desktop. You have access to lots of electronic building and debugging capabilities in a small area. Using a solderless breadboard helps reduce the number of fabrication tools that you need to carry with you on your adventures.*

The Arduino was included for its small size, popularity, and ease of use. There's a whole world waiting for you inside that little board, waiting to be explored, but that's a story for another day.

You can use the USB cable to extract a regulated +5V supply from your PC for your low-voltage-experimenting pleasure. You can also use an inexpensive USB charger to do the same thing, if you don't feel like hauling a whole PC around with you. Just don't try to draw more than half an amp of current, which is the "official" USB limit for supplying power to USB devices. Technically, *unenumerated* devices (i.e., just a USB cable with nothing intelligent enough to identify itself to your computer) can only draw up to 100 mA (0.1A) of current. A USB charger, on the other hand, is not so picky.

No Soldering Required

One notable omission from this particular configuration is any sort of soldering tool. You could certainly add a small soldering iron, some solder, and a few more hand tools. You'll either need to boot out some of the bulkier items, such as the PC or the full-sized multimeter, or get a slightly bigger carrying case.

The Ever-Shrinking Computing Device

One option that will be gaining momentum in the very near future is the use of tablet computers or smart phones instead of the physically larger laptops or netbooks. They already have a lot of the computing capabilities of reasonably powered portable computers. Also, you can look forward to innovative hardware and software interfaces that will make them more useful in the lab as programmable test and measurement devices. While name-brand products still command a premium price, many no-name or OEM (original equipment manufacturer) products are being sold for a fraction of the cost.

Electrical Experiments You Can Perform

Here are some simple electronic experiments that you can perform using only the tools and components found in the compact executive mobile electronics lab. Start out with something simple and incrementally build on your success. Don't try the most complex problems first. What you want is to build up both your knowledge *and* your confidence at the same time.

Lighten Up, 20th Century Style

Here's an easy experiment that will teach you several interesting things about electronics, believe it or not. You can light up a typical LED using only a lithium coin cell. It needs to be a lithium coin cell that produces at least three volts (marked "3V"), not one of the smaller batteries that only supplies 1.5V, such as a watch or hearing-aid battery, as these will not produce enough voltage to light up a typical LED.

Slide the coin cell between the leads of your LED. There are four possible outcomes, three of which are "nothing interesting happens." Let's start with the least likely possibility, in which the LED lights up. Sometimes you just get lucky.

First Possible Outcome: The LED Lights Up When Attached to the Coin Cell

Figure 4-3 shows the results of this outcome. Notice that the longer lead of the sLED (the positive, or anode, lead) is touching the side of the coin cell marked with a +. No current-limiting resistor is required in this circuit because the battery's internal resistance is sufficient to keep from blowing up the LED. It just happens to work out that way, this time. Most other times, you will need to add a current-limiting resistor to the circuit.

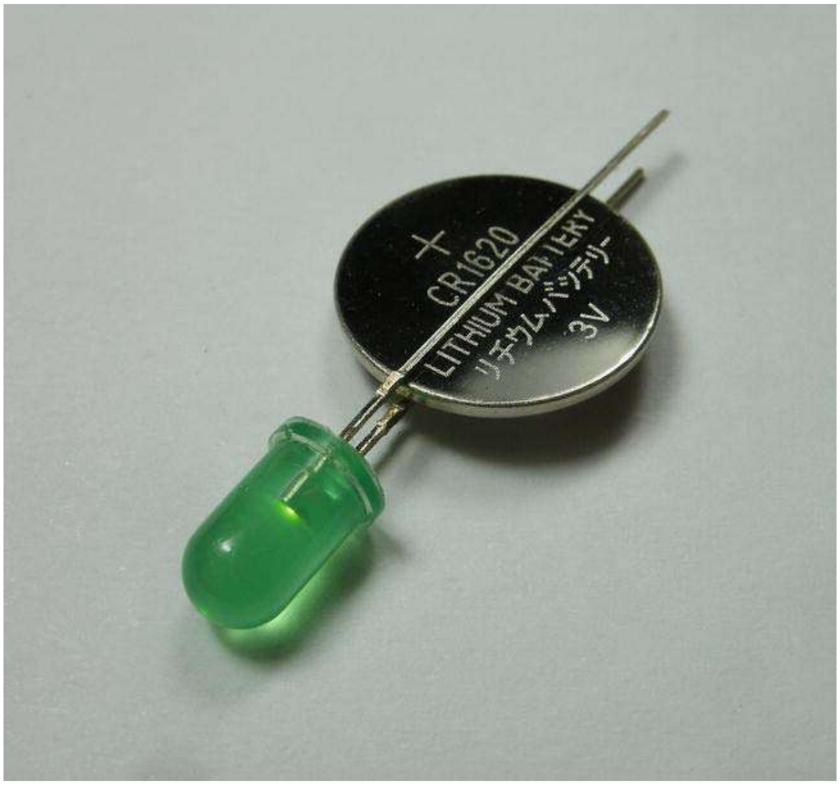

Figure 4-3. *A typical LED lights up when attached to a lithium coin cell. No extra wires needed!*

The other three scenarios? Let's look at them one at a time.

Second Possible Outcome: Nothing Interesting Happens

Why? Because the LED has been installed backward. Alternatively, the coin cell has been installed backward. Remember from Chapter 2 that LEDs are light-emitting diodes, and diodes only carry current in one direction. Therefore, when connected backward, they block the flow of electric current, preventing anything interesting from happening in the rest of the circuit.

Typical lithium coin cells have one broad, flat side that usually has something printed or etched on it, almost always with a + to indicate the positive polarity of the cell. The other side is often textured or rippled in some way to help make better electrical contact when installed in a device. This other side is the negative terminal of the battery.

The LED, being a diode, is *polarized*. It will only work when installed in the proper orientation inside the circuit. Most LEDs will offer one or more ways of easily identifying which lead is which (see Figure 4-4).

- Brand-new LEDs will have extended leads about an inch or so long. One lead will be slightly longer than the other one. The longer lead is the *anode*, or positive lead. The shorter lead is the *cathode*, or negative lead. This information is lost when the leads of the LED are trimmed—for example, after being installed in a PCB.
- Another industry standard for identifying the polarity of an LED is to look at the small ridge or shoulder at the base of the LED body. There will be a flat portion molded into the otherwise circular outline. This flat spot marks the cathode, or negative terminal of the LED.

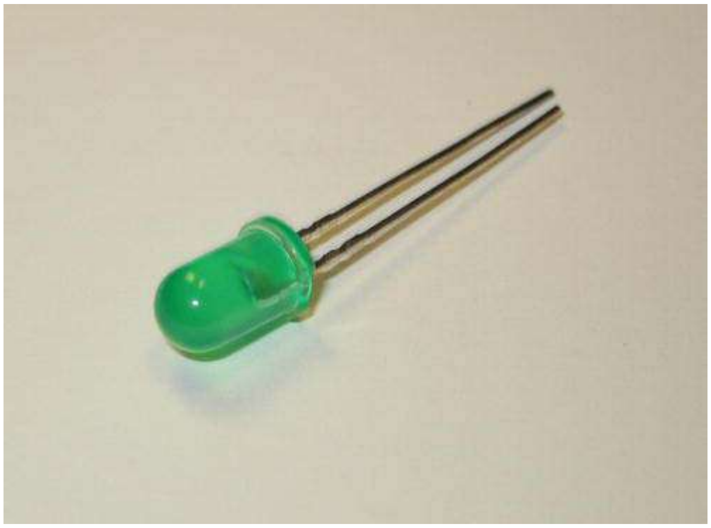

Figure 4-4. *Typical LEDs in the 3mm, 5mm, 8mm, and 10mm sizes almost always have two polarity indicators. The first is that the longer lead is the positive lead, or anode (top). The second is that the flat side on the base indicates the negative lead, or cathode (bottom).*

On really tiny LEDs, there may be an even tinier mark printed or etched onto the LED to indicate the cathode. You will sometimes need a magnifying glass just to see it!

The quick and easy fix for this problem is to swap the direction of either the LED or the battery. If this still produces no good results, we have to consider some more possibilities.

Third Possible Outcome: Nothing Interesting Happens

This time, after establishing that the LED has been properly oriented with respect to the polarity of the battery, we must entertain the possibility that the battery has been previously depleted. Coin cells have a limited amount of energy stored within them, and this can be easily drained in normal use over time.

A quick way to measure a battery's electrical goodness is to use a voltmeter. See Figure 4-5.

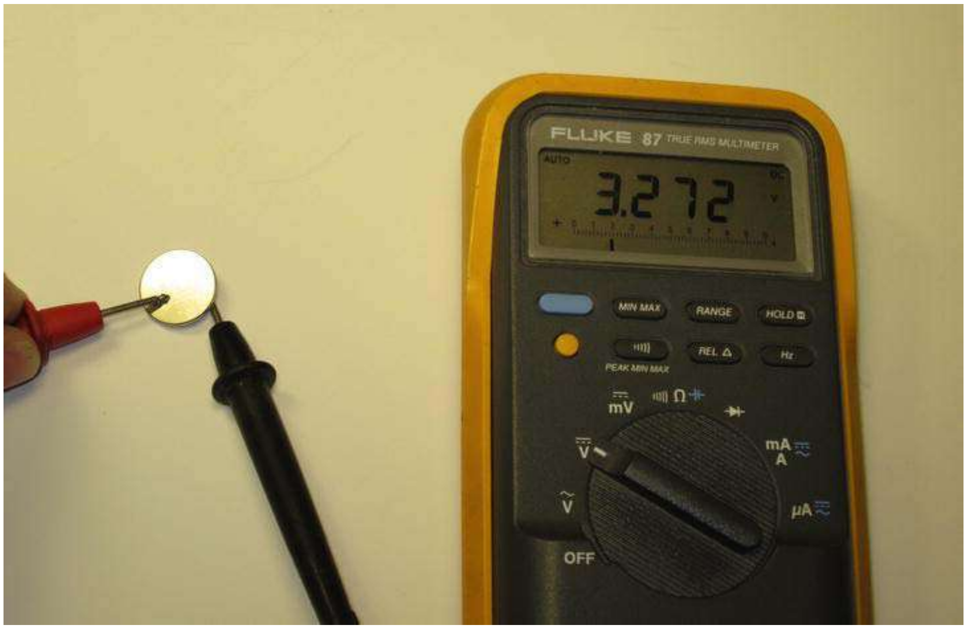

Figure 4-5. *Measuring a coin cell's electrical goodness using a digital multimeter. This one is chock-full of goodness!*

If you have a multimeter, be sure to set it to read DC voltage, in the appropriate voltage range (unless you have a fancy "autoranging" meter). Touch one of the meter's probes to one side of the battery and the other probe to the other side of the battery, and see how much voltage is measured. Anything below 3V should be considered depleted and of no further use, as far as electronic-type projects are concerned. The batteries are shiny and round, so they may have a second career in some sort of art project, but otherwise they should be disposed of in a responsible manner.

Also, perhaps because of their very shininess and roundness, they like to be eaten by small children. Don't let this happen! Take whatever steps are necessary to prevent small children from handling or even touching your electronic tools and components. Use extra care when your lab is mobilized.

■ **Caution** No unsupervised children, no matter how adorable, are allowed in the lab!

Fourth Possible Outcome: Nothing Interesting Happens Because the LED Is Defective

This can happen when an LED is subjected to too much current and the tiny, tiny wires within the plastic body overheat and fail. It can also occur when the LED has been exposed to the elements and moisture has penetrated the casing, allowing the internal connections to oxidize. This scenario, while within the realm of possibility, is certainly the least likely of the four, unless you are habitually cruel to your LEDs. It's also easy to test, if you have a known-good lithium coin cell handy.

Where's the Resistor?

You might be wondering where the current-limiting resistor can be found in this circuit. It's in there, all right, but it's hidden within the battery, so to speak. Coin cells are designed to provide tiny amounts of current over a long period of time. You're not going to arc-weld or start your car with one. As such, they have a much higher *internal resistance* than most any other kind of battery. We use this "feature" to our advantage when testing LEDs using just a coin cell, as the maximum amount of current that will normally flow out of the battery is below that necessary to destroy the LED.

A Slightly More Permanent Circuit

You can build a more permanent circuit for the coin cell and LED using a solderless breadboard. If you recall from the "Solderless Breadboards" section of Chapter 3 (see Figure 3-6), a solderless breadboard is an array of spring-loaded tie points, which serve to both physically hold electronic components and provide electrical connections between certain points. Let's build an LED circuit on a solderless breadboard using an LED, a coin cell, a resistor, a jumper wire, and a custom battery holder (see Figure 4-6). Later on, in Figure 4-7, you can see details of the custom battery holder.

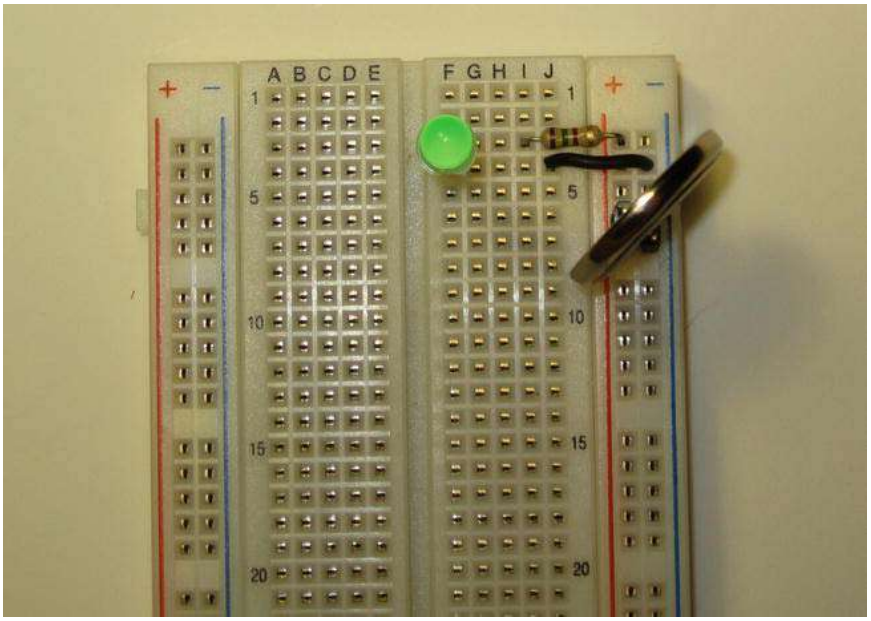

Figure 4-6. *A slightly more permanent LED circuit, built on a solderless breadboard. The tie points on the breadboard are labeled left to right with the letters A–E and F–J for the columns. The rows are numbered starting at 1 at the top. The LED's anode, for example, is installed in tie point F3. The power rails on either side are simply marked with + for positive and – for negative.*

Referring again to Figure 4-6, you can see the overall layout of the LEDcircuit. The power rails along both sides of the breadboard are just longer versions of the electrical connections that connect each of the groups of five tie points together. In this example, the right-hand power rails are used, but the left-hand side is not.

The schematic diagram of this circuit shows how simple it is, at least in principle (see Figure 4-7). Can you see the circle?

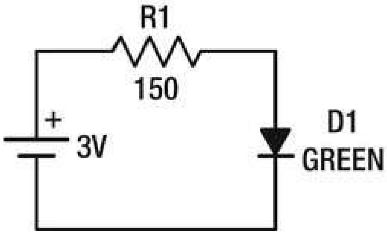

Figure 4-7. *The schematic diagram of the LED circuit as built on the solderless breadboard. Schematics are intentionally simpler than reality (most of the time) in order to help convey the very basic idea of how things are to be wired and the relationships between the components, without bogging you down in all the inevitable details that are involved.*

The Battery

The symbol for the battery actually looks like the battery in this case. The same symbol is used for most single-cell batteries, even when they are of completely different shapes or proportions. Multicell batteries are usually represented with repeated stacks of long and short lines.

The battery symbol's long line represents the positive terminal of the battery, and the short line represents the negative terminal. Again, it's just a coincidence that the wide side of the lithium coin cell is the positive terminal. The addition of the + sign in the schematic is redundant, but you'd be surprised how many people accidentally reverse this symbol. Better safe than sorry when it comes to electricity!

The Resistor

The battery's positive terminal is connected via a wire (drawn as a simple, solid line) to one side of the resistor. The schematic symbol for the resistor is a zigzag line.

The resistor is labeled "R1" (resistor 1) in this schematic, even though it's the only resistor and could just as well have been labeled "RESISTOR #1," or just plain "resistor." The numeric value underneath the resistor is the component's primary electrical characteristic, at least as far as we're concerned when looking at the schematic. In this case, it indicates a resistor whose *value* is 150Ω. In this simple example, the value of the resistor is quite flexible. A larger value, such as 1KΩ (1,000 ohms) or more, would also work, except that the LED would not shine as brightly. This is because *more* resistance results in *less* current flow, and it's the current flow that makes the LED shine. More current, more shiny.

We haven't specified any of the other important characteristics of the resistor, such as the tolerance or the power-handling capacity. Because this is such a low-power circuit, almost anything will do. When you start to work with really big LEDs that require correspondingly really big power supplies, this becomes a much more critical issue. For now, we start simply.

The LED

The other side of the resistor is connected, again, with a simple, solid line, to the anode of the LED, whose *reference designator* in this schematic is D1 (for diode 1). You could have called it "LED," and future generations would probably have understood your intentions, which should be your overriding goal. Its characteristic "value" in this case is "green." Of course, any color will do, as long as you have the eyes to see it.

The LED's schematic diagram effectively communicates its polarized nature. The arrow indicates the flow of *conventional current*. The flat bar indicates the cathode, or negative terminal of the LED. This corresponds to the flat side of the LED body in this example. The cathode is connected via a wire back to the negative terminal of the battery, completing the circuit.

Can you see the circle yet? Hint: It's disguised as a rectangle. The important point is that it forms a complete loop.

Building the Actual Circuit

Comparing the schematic to the photograph of the completed circuit might not immediately click in your head. Take a look at Figure 4-8 for a slightly different angle, where you can see some more details, such as how the battery is being mounted.

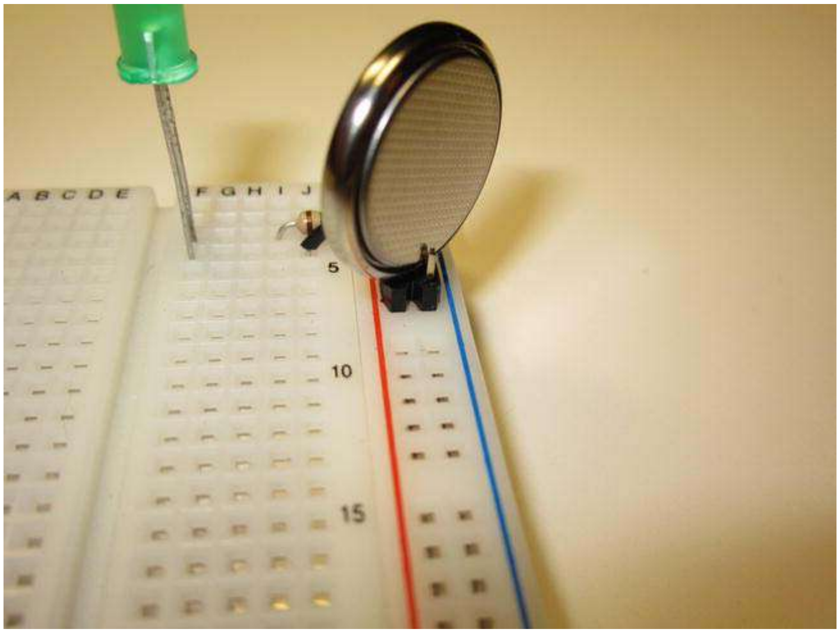

Figure 4-8. *Another view of the LED circuit as built on a solderless breadboard. The battery holder is a modified pin block that has had a couple of its pins removed, allowing the coin cell to fit in on the diagonal. A thinner coin cell could have fit without your having to remove the corner pins. A proper coin cell holder is inexpensive and would provide a much more robust solution to the problem.*

Looking at Figure 4-8, you can see how the positive side of the battery is making physical and electrical contact with the gold-plated pin of the pin block. The remainder of the pin emerges underneath the black plastic molding to make contact with the spring-loaded tie point within the body of the solderless breadboard, which in this case is the negative power rail. The positive terminal, in a similar fashion, makes contact with the positive power rail.

By attaching the battery directly to the right-side power rails, you ensure that your circuits are only a short jumper away from power, no matter where along the breadboard you decide to build them. Handy!

Note that the left-side and right-side power rails are not connected at this time. If you would like them to be, you can easily attach them using jumper wires. See Figure 4-9.

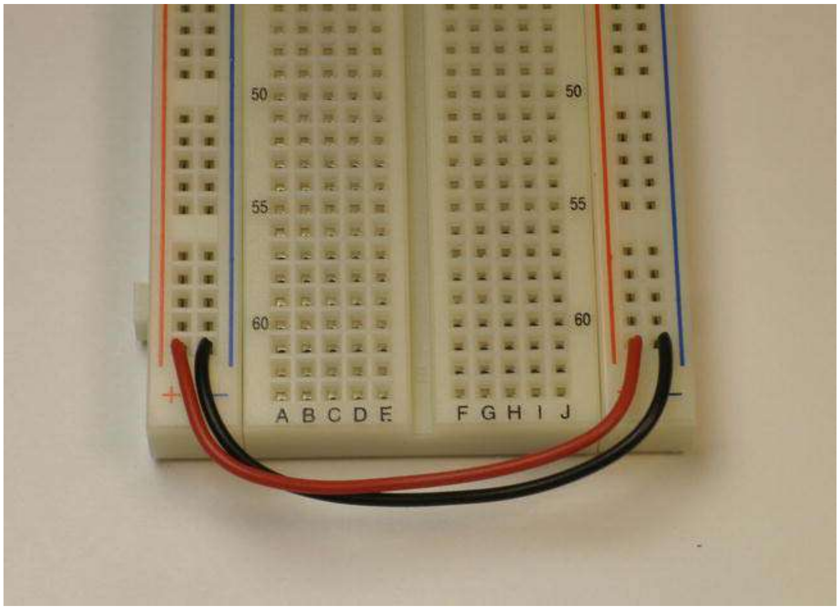

Figure 4-9. *Attaching the left- and right-side power rails using jumpers. By default, the power rails are not connected. There are some situations where you need them to not be connected, such as when using multiple supply voltages within a single circuit. These things happen.*

Now, referring *all the way back* to Figure 4-6 (sorry about that), you should be able to see how the positive terminal of the battery is indeed connected to one side of the resistor. The other side of the resistor, then, makes an electrical connection with the anode of the LED (the non-flat side, when seen from above). The cathode is connected to the negative power rail using a short jumper wire. The negative power rail then completes the circuit back to the battery. The circle is complete.

Note that the resistor could have been placed either before the LED in the circuit (as shown in the photographs) or after the LED. The only important points are that it is in the same circuit, or circle, and that the current must flow through *both* the LED and the resistor. It really doesn't matter which one comes first, especially in this simple circuit. It will only make a difference when you start to do fancy stuff, like adding switches and such to complicate things.

If you're using a different kind of solderless breadboard (and there are a few distinct varieties), you may or may not have power rails available. If so, fret not. You can still make point-to-point connections using more jumper wires. Now you're starting to understand why you were assigned to cut and strip so many jumper wires back in Chapter 2! They should start to be coming in handy just about now.

Multiple Uses for the LED

Now we've moved from the 19th century incandescent bulb to the 20th century LED. What kind of lighting devices will we encounter in the centuries ahead?

One of the nice features of LEDs is that they don't typically require a great deal of power to operate. This is certainly true of indicator-style LEDs, whose only purpose is to be seen as either on or off. We can leave this LED installed right where it is and use it as a handy power-on indicator for future projects. You might be surprised how many heads get scratched in puzzlement, wondering why a newly wired prototype isn't working, only to find out after exhaustive investigation that the power wasn't turned on! Then again, you might not be surprised by this.

Now let's take this perfectly working circuit and break it—on purpose. If we interrupt the circuit, for example, by removing any of the components or their connections, the circuit stops working. To that end, remove the short jumper wire and replace it with two longer jumper wires. Plug in each of the new, longer jumper wires where the two ends of the short jumper were installed. See Figure 4-10.

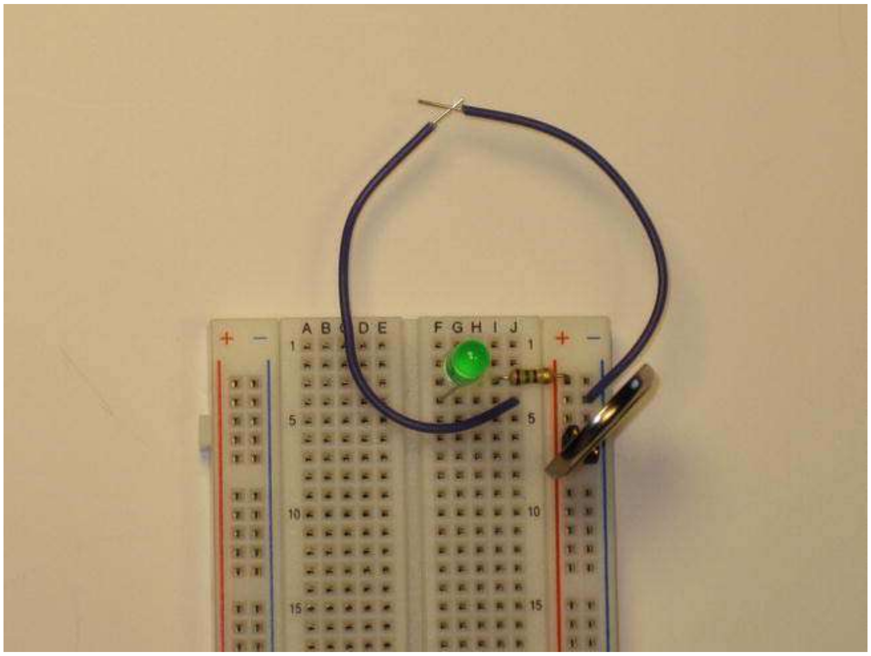

Figure 4-10. *By intentionally breaking the circuit, we stop the flow of electricity through the components, extinguishing the LED. By reconnecting the longer jumpers (or just touching them together), we reestablish the circuit, and the LED should light back up again. We can use this mechanism to test whether there is continuity in another circuit.*

Now, by touching the two probe wires together, you recomplete the circuit, and the LED should glow in happiness. This is exactly how a switch works. You can use this basic idea to build a continuity tester that lights up the LED when the current flows in the circuit. This is very handy for testing cables and wiring in a more complex assembly.

More electronic experiments, including how to both calculate and measure what's happening in a circuit, can be found in Chapter 5.

The Special Project Portable Laboratory

When you move past the simple joys of lighting LEDs and wiring up tutorial circuits, where do you go and what do you build? How about a robot?

Doug Paradis is an award-winning writer and robot builder. When Doug wants to work on one of his robots, he packs up a small kit of tools and heads over the local robot club, which in Doug's case is the Dallas Personal Robotics Group (DPRG). They meet every Tuesday at the Dallas Makerspace for an informal evening of robot building and socializing. See Figure 4-11.

Figure 4-11. *Doug Paradis of DPRG, arriving at the Dallas Makerspace for an evening of robot building. Doug uses the club's laser cutter and other specialized tools to give his robots that winning edge. He brings his own set of hand-picked tools along with him.*

Figure 4-12 shows what Doug brings with him to work on his robots.

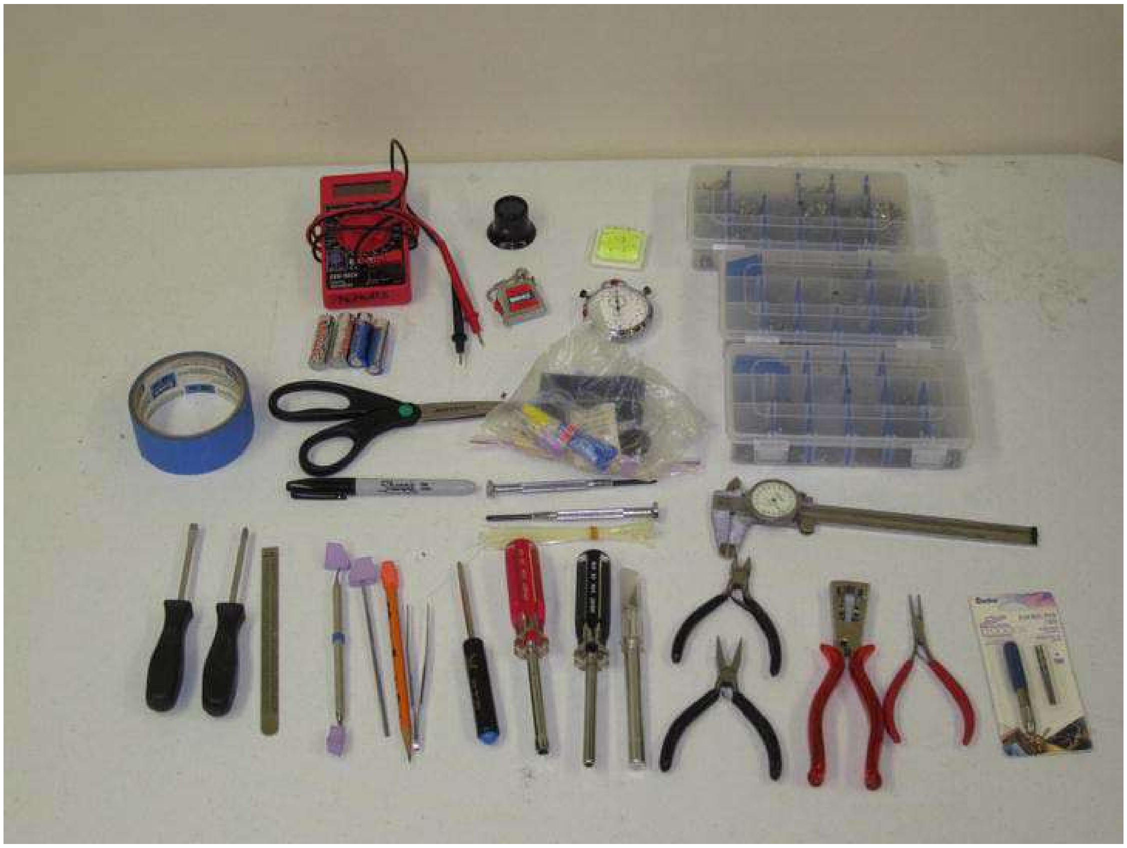

Figure 4-12. *The contents of Doug's portable workshop. A variety of hand tools, precison measurement devices, and fasteners (including tape, tie wraps, and super glue) are easily stored in a small tool bag for easy transport. The comparmentalized storage containers at the upper right contain a large assortment of screws, washers, and nuts.*

Doug has learned the lesson of travelling light. The electronic portions of his robotic creations evolve slowly over time, but the mechanical aspects get a lot of tweaking along the way.

Not everyone has perfected the art of light travelling. See Figure 4-13.

DRAIN
OUTLET

Figure 4-13. *A sort-of-portable robotics lab, with lots of tools and spare parts. One way of looking at this bounteous splendor is "three trips to the car."*

This caravan of tools, parts, and miscellaneous supplies forms the entourage of Kyle the Robot. Kyle is very much a work-in-progress. The basic framework is made of hand-tooled aluminum. The large wheels use O-rings for tires and are driven by 24V gear motors. The 24V supply is made up of two 12V, 5Ah rechargeable lead-acid batteries. Ah stands for amp hours, and represents the capacity of the battery. In this case, it can deliver an amp of current for five hours, or five amps for one hour. The custom PCB holds the electronics for the drive motors and will be controlled by a separate microcontroller that has yet to be added. See Figure 4-14.

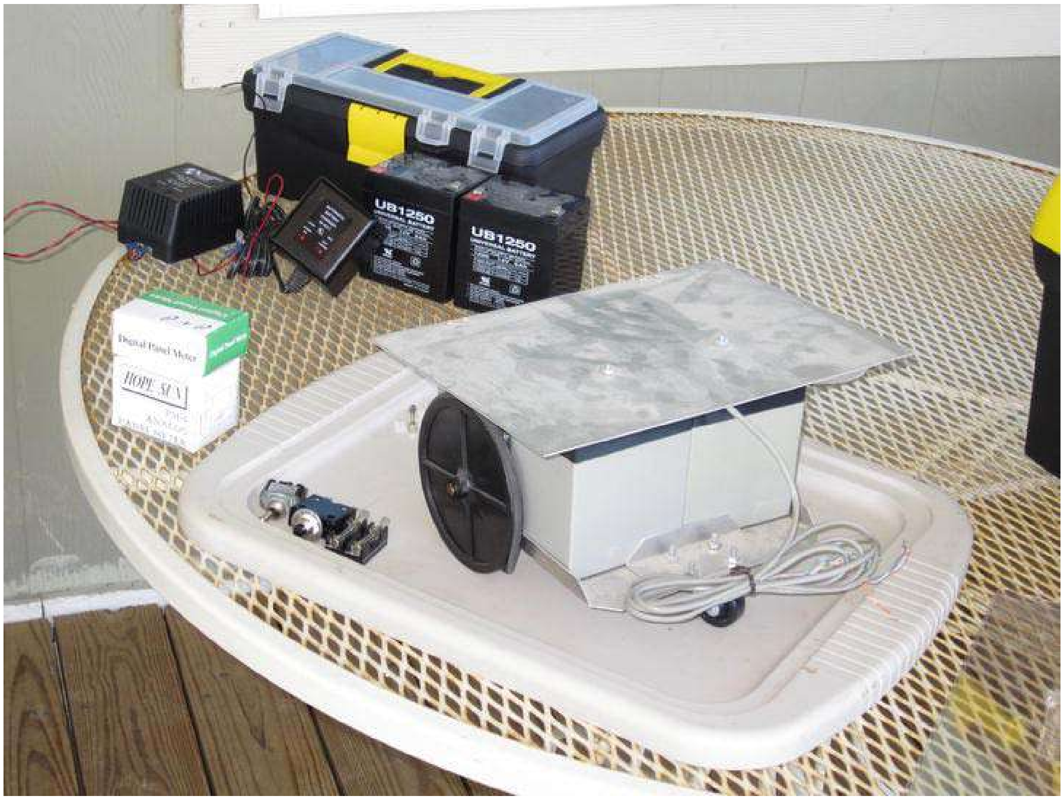

Figure 4-14. *Kyle the Robot is a work-in-progress who requires a wide range of specialized tools and components. Some new components are being fitted, including dedicated battery voltmeters, fuses, a circuit breaker, and a master power switch. Once the sharp corners are filed off, Kyle will be able to roll around the house, looking for chores to do.*

Summary

As we come to the end of this chapter, you should have some good ideas bouncing around in your head about how you want to pack your mobile electronics lab, what tools and components you *absolutely* must have with you, and how you want them organized. The best way to find out for sure what works for you is to make a few field trips, paying careful attention to what was actually used compared to what you anticipated needing. This will help you whittle down your cargo to the bare essentials, which is what you want. Remember, travel light. Excess baggage is a purely optional curse.

In the next chapter, we'll look at some suggestions for organizing and working in a more permanent setting. This could be a domain as small as the corner of a desk or as large as a warehouse. You will soon discover, if you haven't already, that this hobby, like many other projects, will quickly expand to fill the space available. Try to keep in mind that *you* are the one that ultimately makes the decisions. Effort spent planning ahead of time will help keep the sprawl to a minimum, while helping you stay focused on what you wanted to be able to do in the first place.

CHAPTER 5

The Cozy Corner Lab

Chapter 4 focused on the needs of the travelling electronic adept. Here in Chapter 5 you can come home to a lab that is always ready and waiting for some electrical fun and doodling. Nice!

You don't need an entire garage or extra office for a functional laboratory for your electronic endeavors. It's certainly nice if these areas are available to you, but you might be surprised at how much you can do in just a small space, if you set it up properly and maintain it with determination.

The first part of this chapter will deal with the task of rehabilitating a small, dedicated work area that has succumbed to years of poor organizational habits. These things happen. The second half of the chapter explores the possibilities of starting afresh with a nice, clean space—either something brand new or just newly cleaned. Some more projects help break in your new work area, as well as familiarize you with more interesting information about some of the basic electronic components reviewed in Chapter 3.

The Rehabilitation of a Small but Useful Area

Let's perform a little magic, in the form of a complete makeover for a small workstation. Have a look at Figure 5-1. Don't let this happen to you!

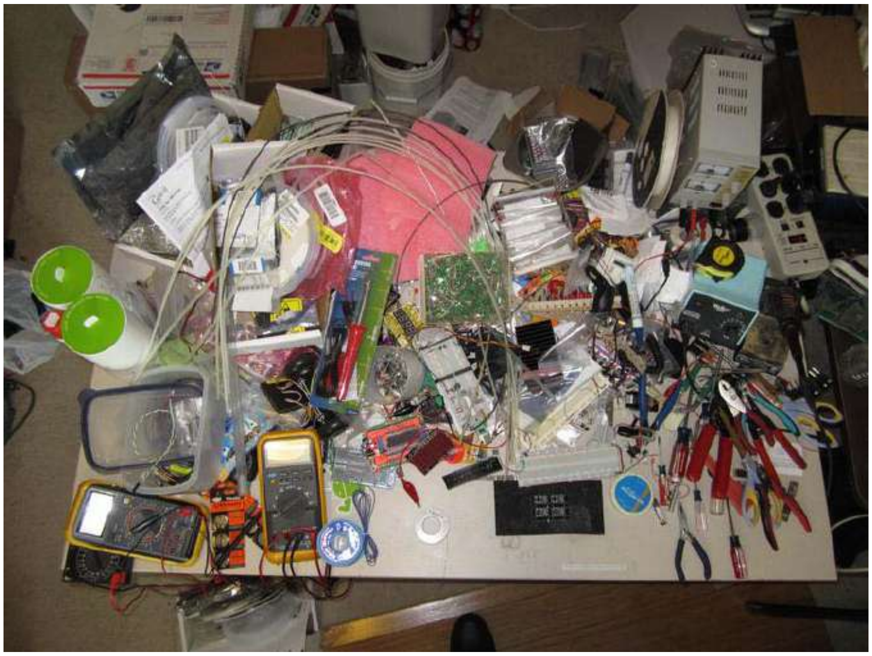

Figure 5-1. *A workstation that has been in continuous use for several years. Not much planning went into the original organization of this work area. Such a space can often accumulate flotsam and jetsam that might have at one time seemed appropriate, but now just competes for precious space. The effective work area is down to only a few dozen square inches! This limits the scope of projects that can be comfortably undertaken. It also causes unnecessary delays in rooting around for tools and parts. It's not pretty and it's not safe.*

Where to begin? Like any other project, this big project is made up of many smaller projects, none of which are especially difficult, if approached in the right order and with the right attitude.

If you're starting from scratch with your electronics lab and don't want to wade through this mess, jump on over to the "Adding a Power Source" section of this chapter.

Science to the Rescue!

Let's use the science of *taxonomy* to help break down this task into smaller, more manageable jobs. Taxonomy comes from the Greek words for "arrangement" (*taxis*) and "method" (*nomia*). That sound like *exactly* what is needed here! An "arrangement method" would assist in building up a plan for attacking this project.

■ **Tip** Method plus rigor equals success! Have a plan, pursue it with determination, and you will succeed.

Let's break down the overall category of "big mess on table" into two major *taxa* (singular: *taxon*), or categories. Ideally, what we will find here are *tools* and *components*. There will be other bits and pieces that don't easily fit into one category or the other (or will maybe fit into both), but it gives us a place to start.

Organization

Divide and *conquer*! These are the two words that will help us through the ordeal of rehabilitating this workspace. We'll get to do a lot of dividing here, and end up as conquerors. The spoils will be a useful (and tidy!) work area that will encourage you to begin new electronic adventures, instead of scaring small children.

The Great Divide

It would be possible to reorganize this workspace *in situ* (i.e., in place), but it would require a great deal of inner peace and lots of patience. Lacking these, a complete reboot is in order. Let's clear off the entire table and start afresh. This will provide the excellent opportunity to arrange things in exactly the best possible order.

First, get two big containers. These can be cardboard boxes or plastic storage bins—it really doesn't matter. We'll just be using them as temporary staging areas as we make like an embryo (and divide and divide again). See Figure 5-2.

***Figure 5-2.** Two large cardboard boxes are enlisted to help with the initial organization of the contents of the tabletop. One will be for tools and the other one for components. Further subdivisions will be made later.*

Now, the first thing you will want to do is remove any items that really don't belong in the lab at all, such as coffee cups, personal effects, and any obvious trash. Everything that is about to go into one of the two boxes should eventually find its way back to the work area, so now is the time to omit the things just don't belong here.

Is it a tool? Is it a component? How do we tell? Ask five people and you will get five different answers. For the purposes of this little endeavor, a *component* is something that will become a part of a project, while a *tool* is something that is used to modify or manipulate a component. Another way of looking at it is that components get used up quickly, while tools get used up slowly. If you build enough projects, you will be lucky enough to see exceptions to these guidelines.

Here are some examples. Wire is a component. Wire cutters are tools. That was easy. A flashlight is a "visual inspection aid," so it goes under the category of tools. Batteries for the flashlight are consumable, so they are components. There's no real need to be ultra-precise at this point. Use your best judgment and make up policies that make sense to *you*.

It might help, from a physical logistics standpoint, to use smaller subcontainers to hold the really small bits as you're lumping them into the two broad categories. Folding cardboard bin boxes are handy both on the tabletop as well as on shelves, if you have them. The cardboard bin boxes shown in Figure 5-3 are from Uline (`http://uline.com`), part numbers S-16268 (4" wide) and S-16269 (6" wide).

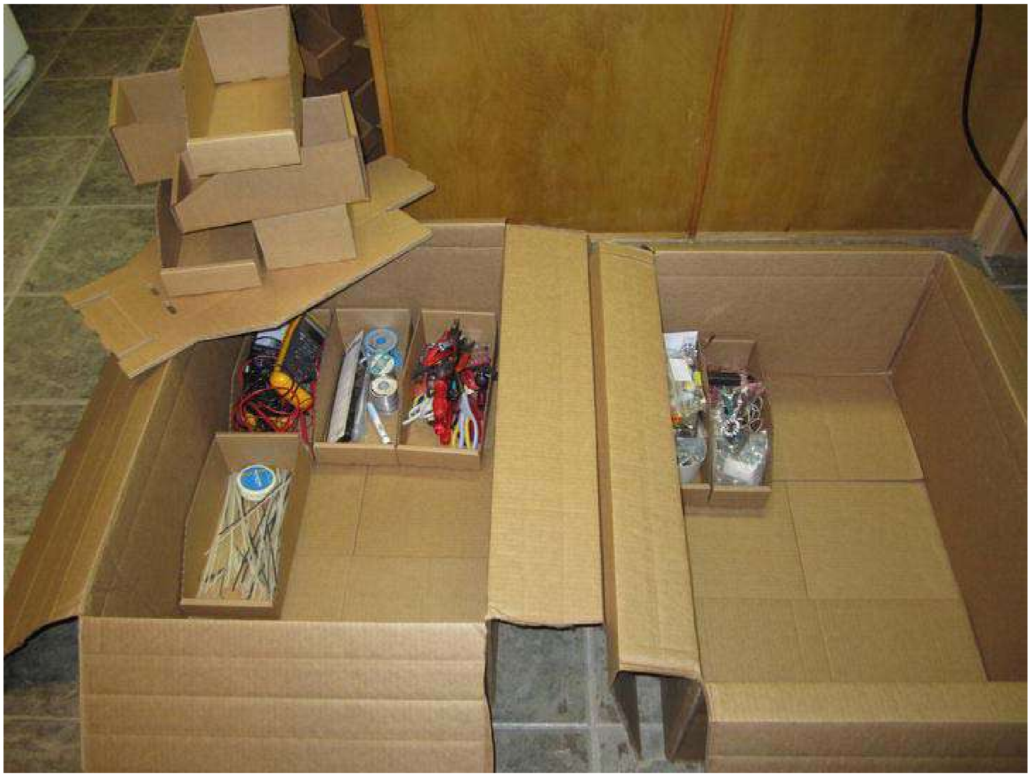

Figure 5-3. *Folding cardboard bin boxes help contain some of the smaller bits and pieces during the Big Sort. You can use whatever happens to be available. Some preliminary subsubdivision is going on here. For the most part, the tools are on the left and the components are on the right.*

After everything is removed from the table and the cardboard boxes are nearly full, the table itself should get a good washing. Behold the *tabula rasa* (blank slate) depicted in Figure 5-4.

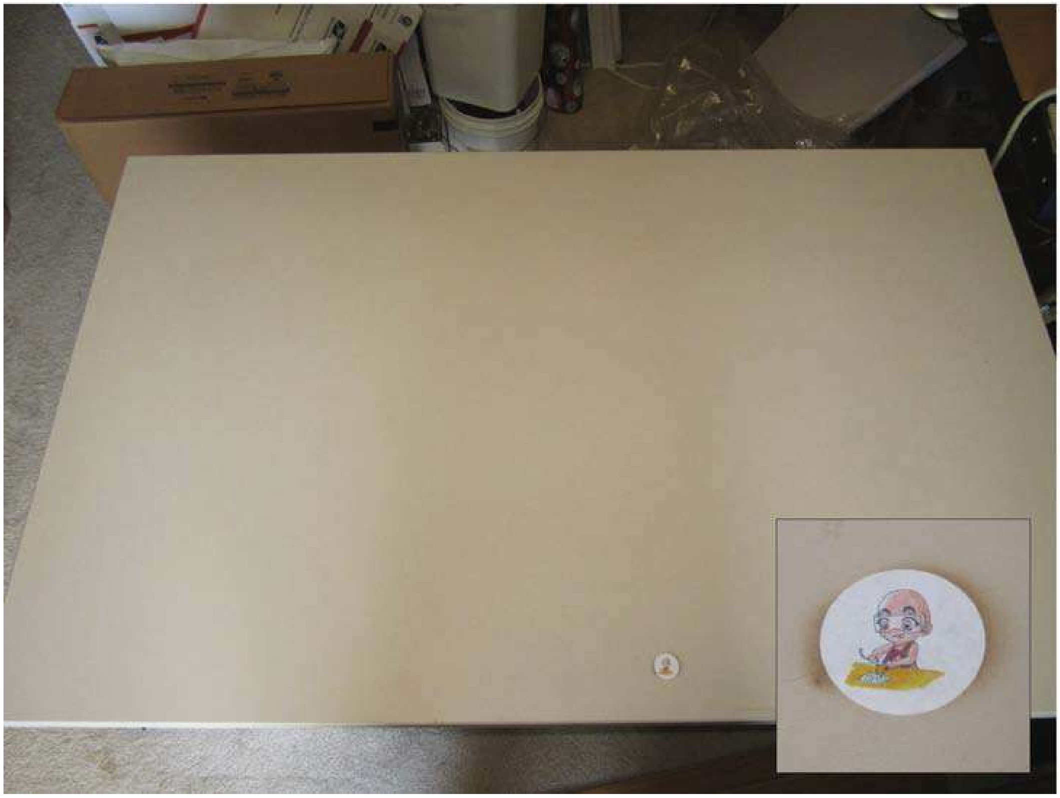

Figure 5-4. *The tabletop has been cleared of tools, components, and various bits of debris. A good washing with soap and water completes the cleansing. A burned spot (the result of a poorly supervised thermal test) is covered with a small sticker (see inset).*

Adding a Power Source

Here's where a bit of planning will save you a lot of headaches in the future. What you never seem to have enough of in any laboratory, especially electronics labs, are lots of conveniently located power outlets. Let's take care of that issue right away, before building anything else onto the tabletop. See Figure 5-5.

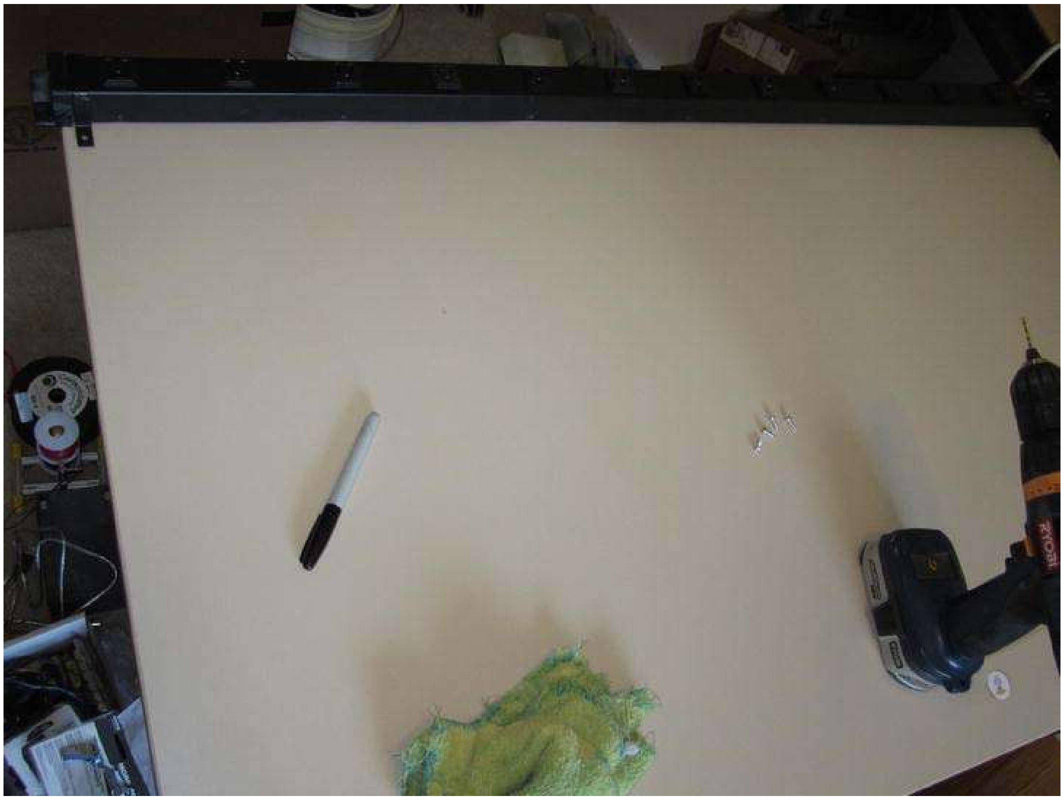

Figure 5-5. *A 12-outlet power strip is attached to the back edge of the tabletop. This will provide a safe and convenient source of power for all those gadgets you'll eventually want on your bench. Tools needed for the job include a drill motor to drill out some small pilot holes for the brackets, a small drill bit, and a felt-tipped marker. Before drilling the pilot holes, align the brackets where you want them to go and use the marker to put a dot where the pilot hole should be drilled. Make sure the power strip is not plugged in when you're drilling right next to it!*

Having lots of extra power outlets is a nice bonus, but is not absolutely required. What you want to avoid, however, is excessive daisy-chaining of power strips and extension cords. This is a fire hazard. Remember from Chapter 3 that every conductor has a certain amount of resistance. This resistance turns electricity into heat. Enough heat makes things catch on fire.

If you do have the luxury of being able to mount a permanent power strip (which, technically, remains a temporary power tap as far as the manufacturer is concerned), then be sure to install it securely using the hardware provided. See Figure 5-6.

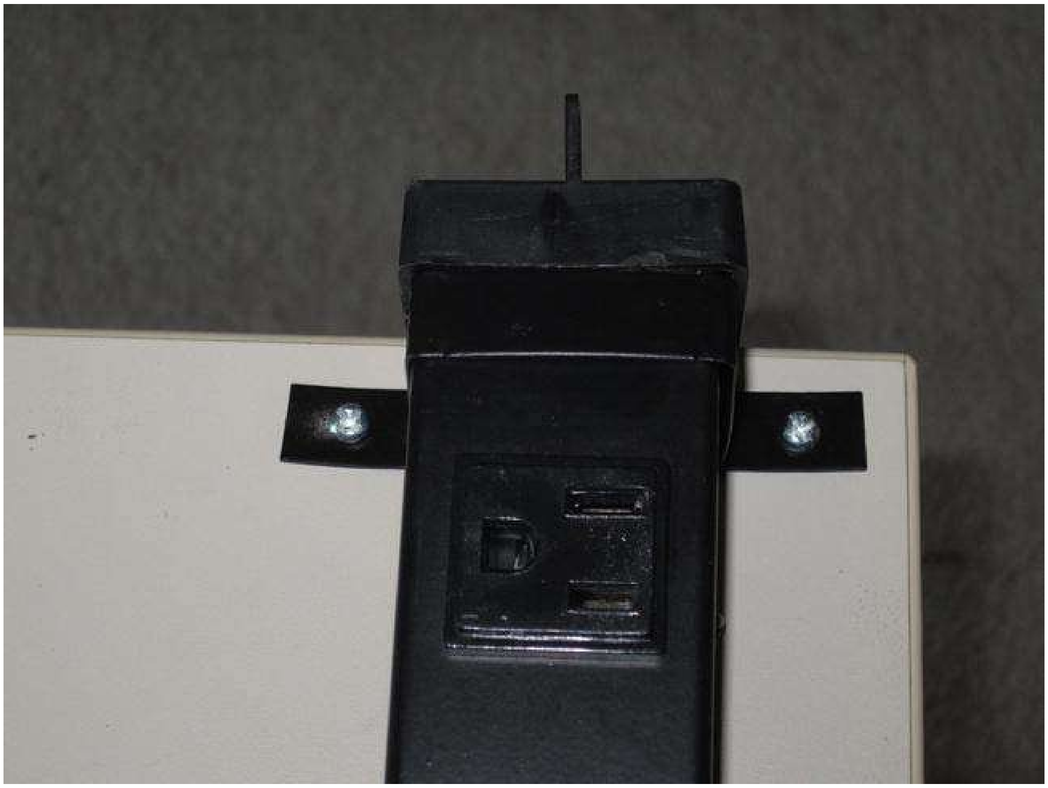

Figure 5-6. *The power strip is fastened securely to the table using the provided hardware. You do not want your power strip wiggling around, or worse yet, falling off the table, dragging all of your power tools with it!*

■ **Tip** Test every outlet of your power strip before permanently mounting it to your workbench.

A Clean, Well-Lighted Space

You can never have enough light, it seems. What passes for normal room illumination might be sufficient for such adventurous undertakings as finding your way to the door or looking for the remote control for the television. For anything more demanding, you're going to have to add some more lighting, and probably lots of it.

At a minimum, you are going to want some *task lighting*, which helps illuminate the immediate area of your focus. In this case, where a small table is being used as a workbench, an articulated fluorescent lamp with a built-in magnifier will do a great job, at least until more lighting can be strung up. See Figure 5-7.

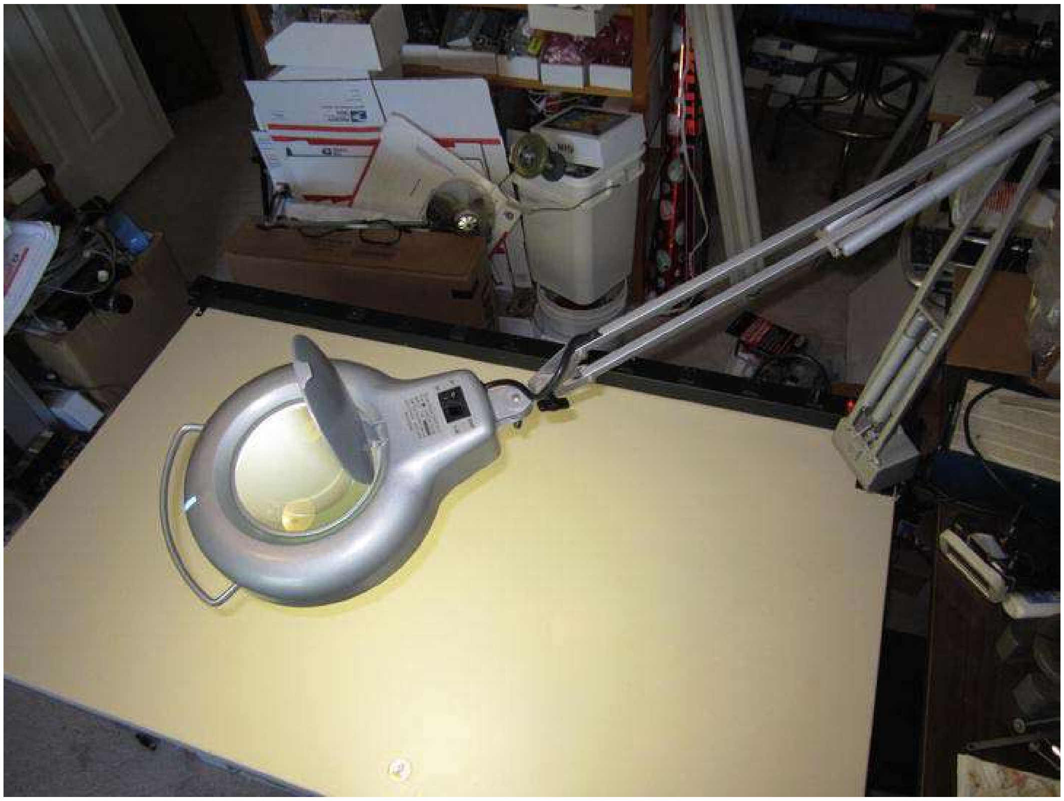

Figure 5-7. *You will need some task lighting for your bench. This desk lamp also includes a handy magnifier for really close work.*

Tools

Now we're ready to bring some of the tools back in to the lab. Here's where you get to decide which tools get a *permanent* home on the bench and which ones get put up and stored when not in use.

The permanently deployed tools on your workbench should be the ones that you find that you are using the most. The ones that get put up after every use are the ones that either only get used infrequently or pose some sort of hazard, such as sharp edges or hot surfaces. Again, you get to decide. It's your lab.

Tools for Soldering

An activity that is practiced frequently in many electronics labs is soldering. If you haven't learned how to solder yet, you are strongly encouraged to do so. Here's a handy setup for soldering, sitting in one corner of the newly refurbished workbench. See Figure 5-8.

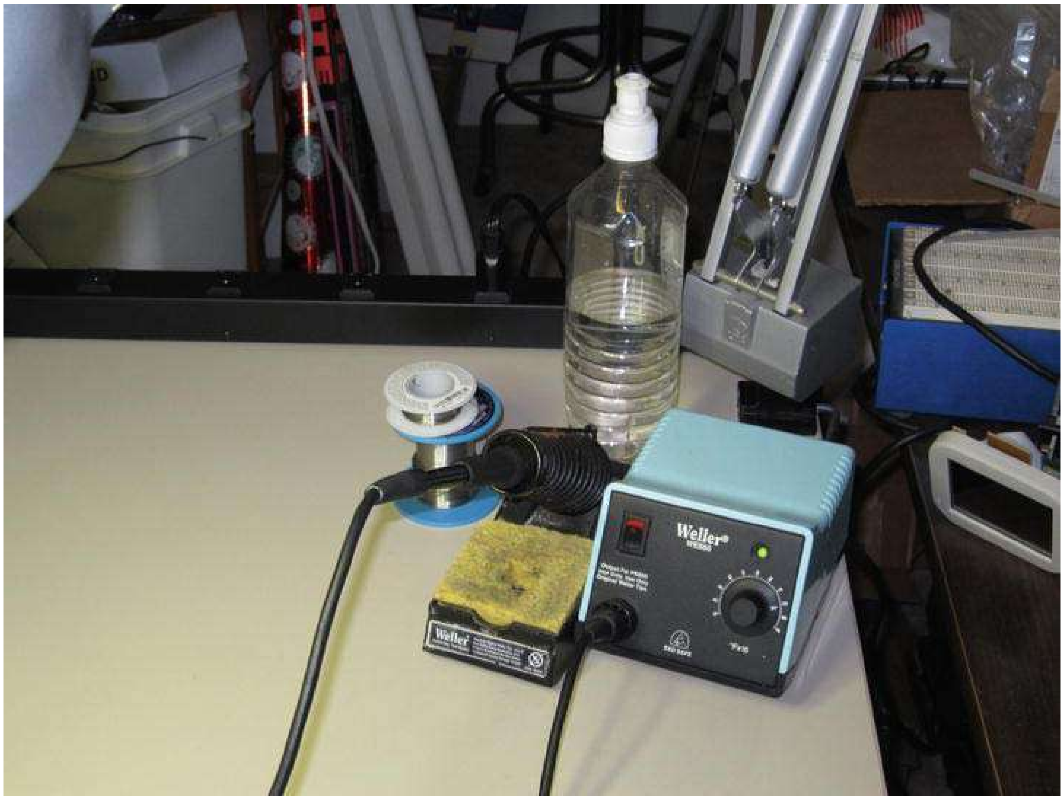

Figure 5-8. *A temperature-controlled soldering station with a stand for the soldering iron, a water bottle to wet the tip-wiping sponge, and some different kinds of solder. Some people prefer to have the soldering station nearer to the front edge of the table. Use whatever configuation makes the most sense in your lab.*

■ **Caution** Soldering irons are hot and can cause serious burns. Do not place a soldering iron where you might accidentally knock it off the work table or brush against it.

Hand Tools and Some Test Equipment

No matter what you work on at the bench, you will most likely discover the need for some common hand tools. When working with electronics, especially, you will often need wire cutters, wire strippers, and small pliers. These tools have been laid out conveniently (for a right-handed person) at the lower-right-hand side of the bench, while the less-often-used hand tools can reside in a box.

A regulated, adjustable bench power supply is also very handy when working with electronics of all kinds. A good, reliable multimeter can also claim a permanent home on the bench, as these will get a lot of use. See Figure 5-9.

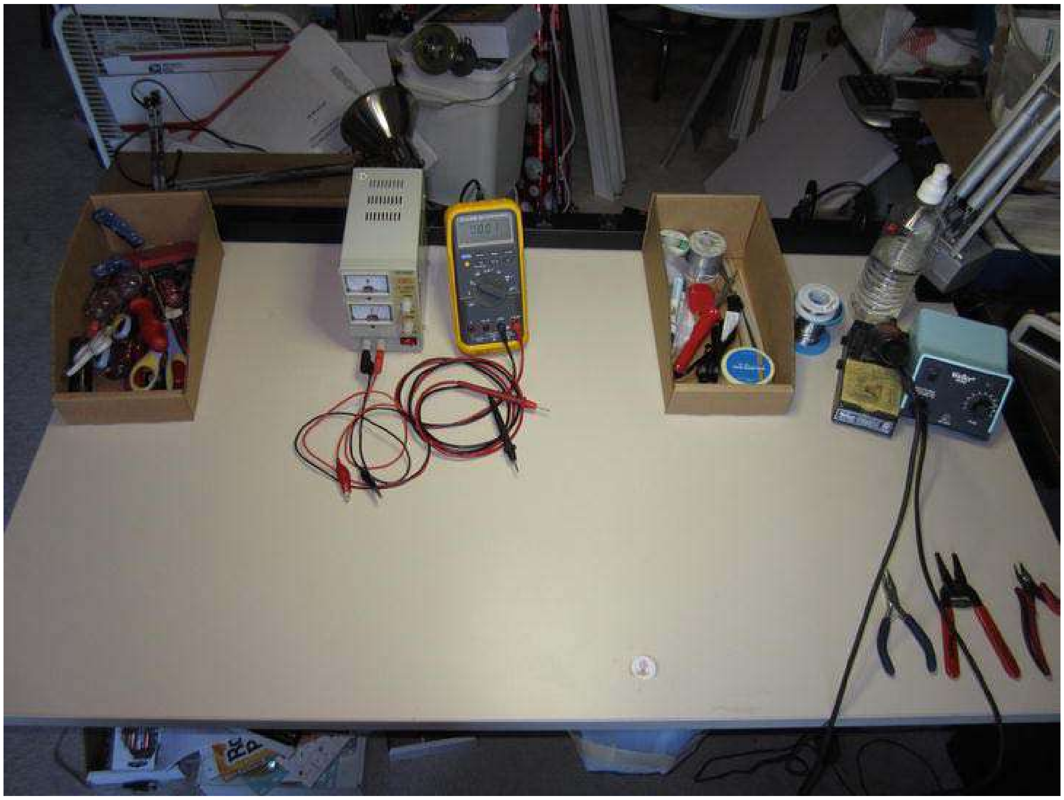

Figure 5-9. *A clean bench with just enough tools to have some fun. Extra soldering supplies, including more kinds of solder , extra soldering iron tips, soldering flux, and an emergency backup soldering iron, sit in a box adjacent to the soldering station. Hand tools and basic test-and-measurement equipment complete the set of "most-useful tools." There's still room for a few more, when you need them.*

Under the Desk

If your workbench has the luxury of drawers, by all means use them! Try to keep your under-desk environment as free of clutter as possible. It's not really a very convenient place to store unused tools or components. You're going to drop the occasional item from time to time, so having to dig out a bunch of junk just to find it is just going to be more frustrating for you. Keep it clean!

You *should* keep a medium-sized trash can under or near your work area, as you will inevitably generate some waste products in this hobby. By minimizing the amount of food-related garbage deposited in your laboratory trash receptacle, you can postpone the invasion of little critters looking for a free meal. Also, bits of wire and old components don't tend to stink after a few days, but a half-eaten sandwich might.

■ **Tip** Copper easily recycles at a premium price. Collect your copper wire snippings and cash them in someday.

A nice addition to the under-the-table accessories you might consider is some low-wattage strip lighting. LEDs are especially good for this. You will appreciate this more once you conduct your first search-and-rescue mission for that teensy-tiny component that dropped off the edge of your table and into the carpet.

Components

Most of our tools are now ready to go. The most frequently used tools are permanently set up, ready for action. The less frequently used tools are carefully stored where we can get to them when we need them, but are not in the way the rest of the time. This is a big improvement over the situation you saw back in Figure 5-1.

Now we need some components with which to work, using our scientifically arrayed tools. Have a look at Figure 5-10 to see one example of how to keep your parts organized yet within easy reach.

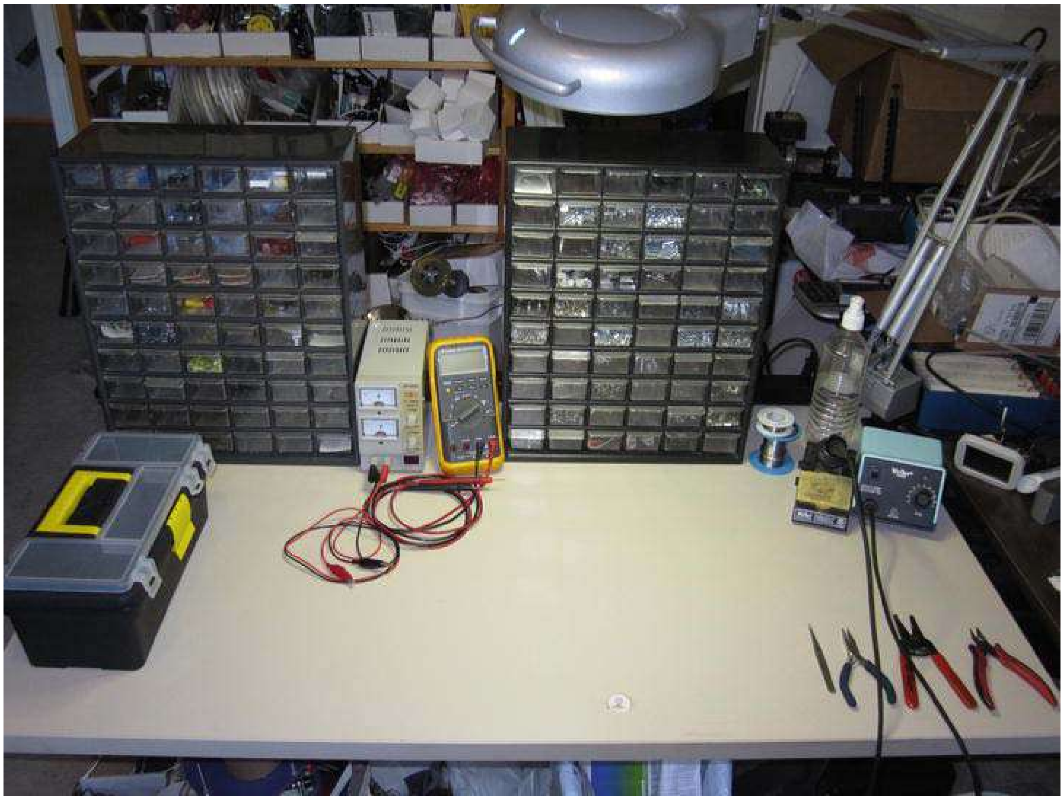

Figure 5-10. *Parts bins are used to organize components. This style of container is available in many sizes with different numbers and sizes of drawers. Even more parts are stored on a shelf behind the work area. The most commonly used hand tools are laid out on the right. The other hand tools are collected in a small toolbox on the left.*

Where to Go from Here

This work space could be further augmented with shelving, both above the work surface and underneath. More lighting wouldn't hurt, either. However, this is plenty to get started. Let's build some electronics projects on this squeaky-clean table and see if we can dirty it up a bit.

Projects

Let's look at a few smaller circuits you can easily build on your tidy, new work area using a solderless breadboard. Expanding on the LED-and-coin-cell experiments in Chapter 4, we'll add a few more components and hopefully start to see how they all interact.

There are any number of different types of solderless breadboards available today. A range of them is shown in Figure 5-11.

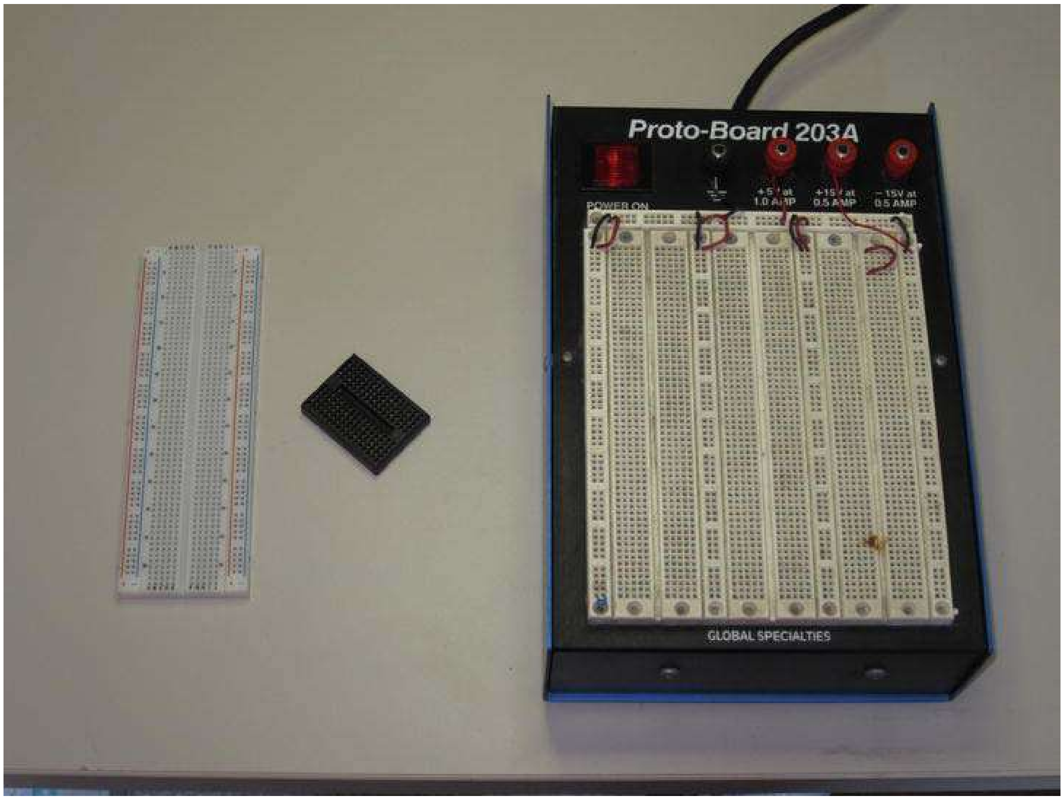

Figure 5-11. *Different sizes of solderless breadboards, all with the same lead pitch and internal connections. The small black unit in the center is perfect for little circuits with only a few small components. The white unit on the left has more tie points and offers power rails along each side. The deluxe, powered unit on the right adds even more space, as well as multiple regulated power supplies for all your tinkering requirements.*

Parallel and Series Circuits

The circuits presented so far have all illustrated the "loop" characteristic of electrical circuits. Electricity generally flows in a circle. It's possible to have more than one path through a circuit, however. What happens then?

Like most good questions, the answer is, "It depends." Let's hook up *two* LEDs at once and try to figure out exactly what's going on. Figure 5-12 shows how it's put together. See Figure 5-13 for the schematic diagram.

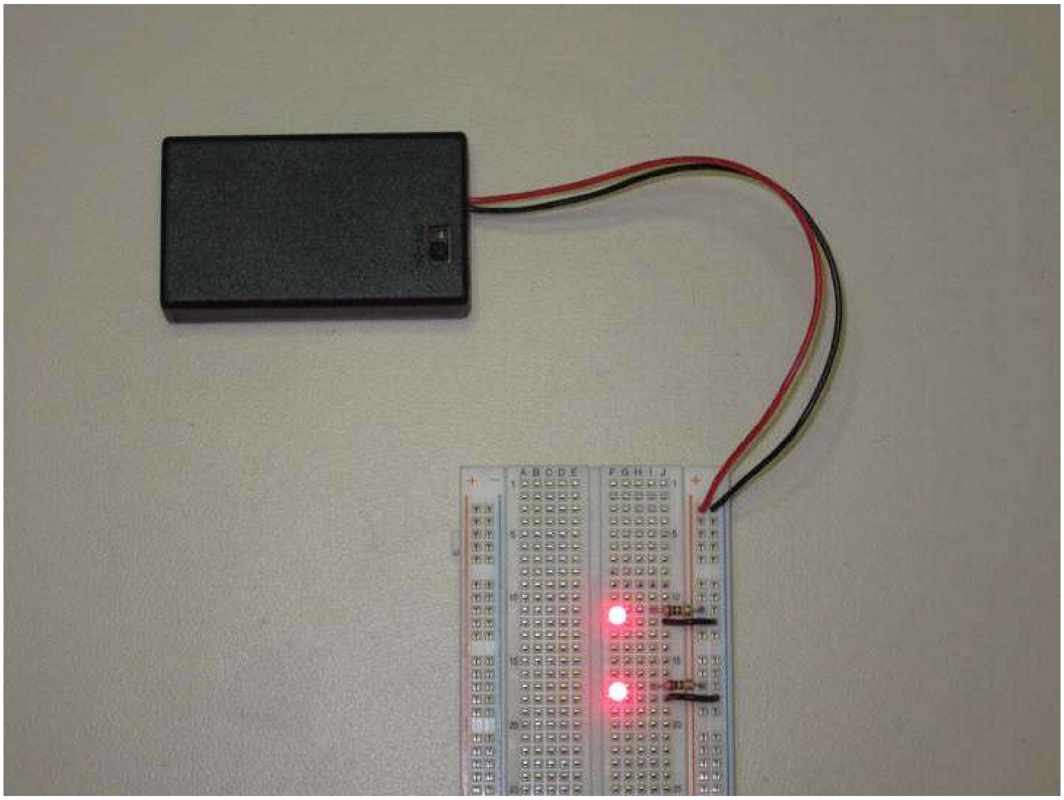

Figure 5-12. *Two LEDs are illuminated at the same time. This simple circuit demonstrates both parallel and series electrical circuits. A 3×AAA battery holder with an integrated power switch is used to provide power to the circuit.*

To build this circuit as shown, you will need the following components:

- A solderless breadboard with power rails
- A 3×AAA battery holder (which holds three AAA cells) and three fresh AAA cells—rechargeable cells are better
- Two LEDs (any visible color will do
- Two 100Ω 1/4W resistors
- Two short jumper wires

Using the letters printed on the breadboard to find the columns across the top and the numbers to find the rows, we can describe each of the tie points using a single letter and a number, such as A-1 or B-2. The following exercise shows how to put it all together.

ASSEMBLY INSTRUCTIONS

1. Install the top LED with its anode (longer lead) in tie point F-11 and the cathode in tie point F-12.
2. Insert one of the resistors from the red power rail on the right at row 11 across to tie point I-11.
3. Connect the right-hand blue power rail at row 12 to tie point J-12 using a short jumper wire.
4. Install another LED just below the top LED, with its anode in tie point F-17 and the cathode in F-18.
5. Install the other resistor from the right-hand power rail at row 17 across to tie point I-17.
6. Use another short jumper wire to connect the blue power rail on the right at row 18 to tie point J-18.
7. Install three fresh AAA cells into the battery holder.
8. Insert the red wire (positive) coming from the battery holder to the red power rail on the right at row 3 (the top).
9. Insert the black wire (negative) from the battery holder to the blue power rail on the right at row 3.
10. If your battery holder has a power switch, turn it on now.

The circuit is now complete. Please verify that both LEDs are illuminated, as shown in Figure 5-12.

Troubleshooting

If both LEDs are not shining at this point, you're going to have to do some troubleshooting.

- The first thing to do is double-check your wiring. Does your breadboard look like the one in Figure 5-12?
- Next, make sure you're using good batteries. If you have a voltmeter, you can test the batteries by measuring their voltage. New alkaline or carbon zinc cells should read 1.5V or more. Rechargeable nickel-cadmium (NiCd) or nickel–metal hydride (NiMH) batteries should read around 1.2V each when completely charged.
- Are the cells properly installed in the holder? Are they in the proper orientation and firmly seated?

- Is the power switch (if available) not in the On position? This happens more frequently than anyone would like to admit. Wiggle the switch back and forth and see if it's only making intermittent connection.
- Check the orientation of the LEDs. They will only work if installed in the correct orientation. It only takes a second to unplug them, swap them around, and try them the other way. At these low voltages, it won't hurt them to be installed backward.
- Are your resistors the right value? The 100Ω value is not critical. Anything from 100Ω to 1KΩ will do nicely. Don't use smaller values, as this will allow too much current to flow through the circuit, perhaps damaging the LEDs. How much current? We'll figure that out in just a bit. Larger-value resistors will decrease the amount of light emitted by the LEDs, down to the point where you can't see them anymore.
- Are you sure your LEDs emit visible light? Millions of infrared LEDs are manufactured every month for use in remote controls and night-vision gear. This light is not visible with the human eye, although some digital cameras can see it. Just because it looks like an LED doesn't mean it *is* an LED.

A Different View of the Same Thing

Your version of the circuit in Figure 5-12 might look a little different, and that's just fine. The important point is that both LEDs are lit up at the same time when you apply power.

Let's look at the traditional schematic for this circuit to get an idea of what's going on here. See Figure 5-13.

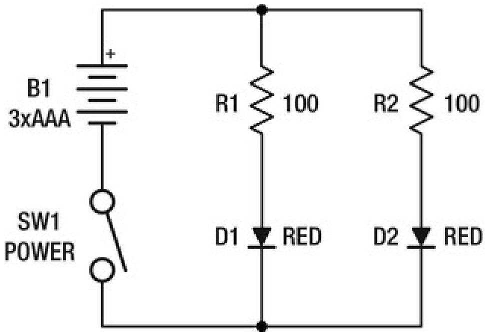

Figure 5-13. *The traditional schematic diagram of the two-LED circuit. This diagram should convey the very basic information needed to build the circuit, as well as understand the important aspects of how the circuit is supposed to work, without going into much detail at all about the actual implementation or assembly of the components.*

By now you should be able to look at the schematic and see the loop formed by the circuit. The only new twist is that there are *two* paths along which the electricity can flow, either through R1 and D1 (the top LED on the breadboard) or through R2 and D2 (the bottom LED, or the initials of a certain adorable robot from the movies).

The three AAA cells in the battery holder are shown as B1 in the diagram. The power switch, labeled SW1, is shown as a separate component to remind us that the circuit can easily be broken, halting the flow of electricity, which is something we want to happen from time to time. The connections between the two resistors and the two LEDs should be somewhat obvious. I'll discuss the details for the parallel and series connections momentarily.

Remember that an electronic schematic diagram is an abstract representation, boiled down to only the essential items involved and the relationships among them. As such, it doesn't need to resemble the final product, but it could. See Figure 5-14.

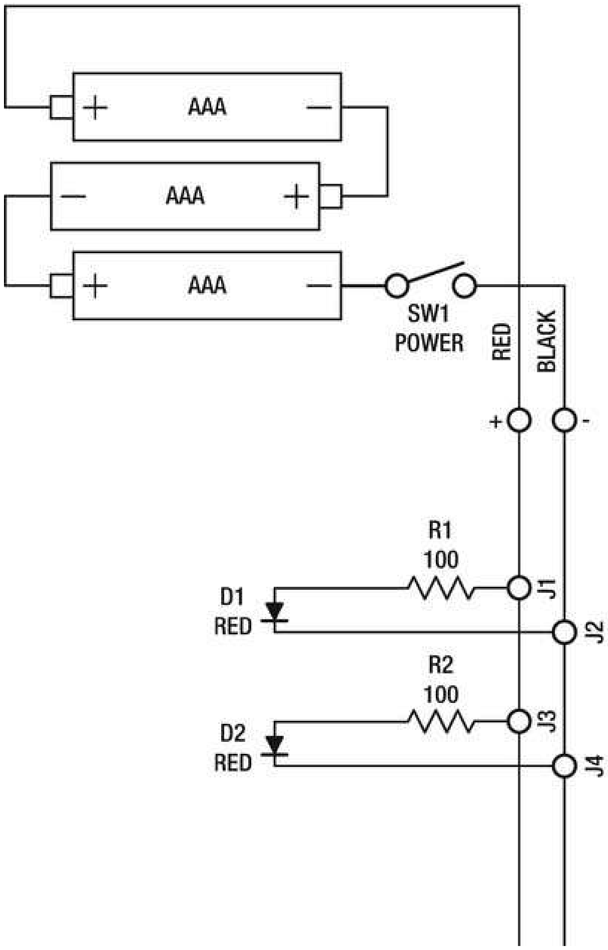

Figure 5-14. *Another schematic diagram of the two-LED circuit, showing the approximate physical layout of the components on the solderless breadboard. This style of schematic is often used when describing how to actually build a circuit, showing the constraints of the actual, physical components and the connections between them. It conveys more specific information than the previous example in Figure 5-13.*

Batteries in Series and Parallel

Let's look at the batteries first. Technically, as mentioned in Chapter 1, there is one *battery* in this circuit, and it happens to be composed of three *cells*. The term *battery* is very often used interchangeably with *cells*, and the distinction is minor. It's better to avoid confusion than it is to be exactly, precisely *correct*—usually.

We see three cells sitting side by side in both the schematic in Figure 5-14 and the actual battery holder shown in Figure 5-12. The cells are connected *in series*, meaning that one connects to another that connects to another, and so forth. They are like beads on a single wire. You should be able to trace the wiring (the solid lines) that connects the individual cells in the schematic.

Note the polarity of each of the cells within the battery. The positive terminal of the bottom-most cell is wired to the negative terminal of the center cell. The positive terminal of the center cell is connected to the negative terminal of the topmost cell.

The negative terminal of the bottom-most cell is connected to one side of an *SPST (single-pole, single-throw)* switch. The other side of the switch emerges from the battery holder's body as the black, or negative, lead from the battery pack. The positive terminal of the topmost cell connects directly to the red, or positive, lead coming out of the battery pack.

The switch could have just as easily been installed in the positive lead, and it would have worked exactly the same. The battery holder used in the photograph happens to switch the negative lead, as was revealed by examining the internal wiring using a multimeter as a continuity tester.

When the switch is *open* (i.e., in the Off position), the circle is broken and no electricity is going to flow. When the switch is *closed*, the circuit is complete and power is made available to the remainder of the LED circuit, assuming that it is wired up correctly.

The amount of *voltage* being supplied by the battery holder depends on the electrical characteristics of the individual cells. If alkaline batteries are used, each cell will provide 1.5V, assuming that the cells are relatively fresh. Rechargeable NiCd or NiMH cells will provide around 1.2V when fully charged.

When batteries like these are connected in series, the total voltage available is the sum of each of the cells. Just add all the voltages together. For example, using fresh alkaline batteries, the total voltage would be 3 × 1.5V = 4.5V. Rechargeable cells would provide 3 × 1.2V = 3.6V when fully charged.

If the cells were wired *in parallel*, with all of the positive terminals being connected together and all of the negative terminals wired together, the total output voltage would only be 1.5V for alkaline cells. The total output *current*, however, would be triple the capacity of each individual cell. See Figure 5-15 for a comparison of series, parallel, and combination circuits using batteries.

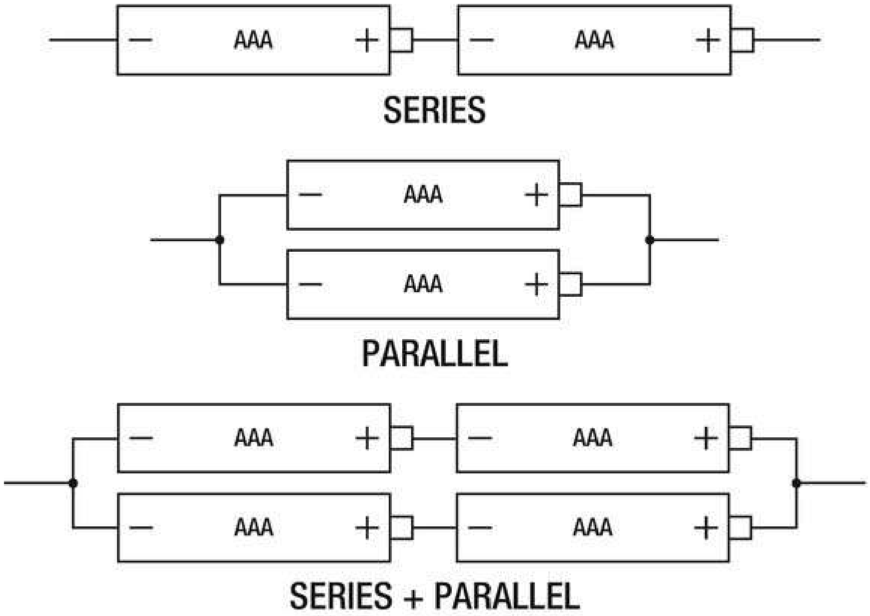

Figure 5-15. *Different arrangements of the wiring of the batteries produce different results. The top two cells are arranged in series. Their respective voltages are added together. The center circuit consists of two cells arranged in parallel. This configuration doubles the current that can be delivered. The bottom drawing shows a combination of both series and parallel circuits, doubling both the voltage and the current.*

When connected in series, the total current output is the same as the current capacity of each individual cell.

In summary, if you need more voltage, stack the cells end to end. If you need more current, stack them side by side. Even more complex arrangements are possible, using both parallel and series circuits to achieve the desired power output characteristics. This practice is very common in hobby model aircraft and battery-powered vehicles. A designation such as 3S2P means that there are two parallel circuits (2P) of three cells in series (3S).

Looking back at the circuit in Figures 5-13 and 5-14, we see that the “battery” used in this circuit consists of three cells in series. This means that if alkaline batteries of 1.5V each are used, the total voltage available to the circuit is 4.5V. If typical rechargeable cells (NiCd or NiMH) are used, we can expect 3.6V to flow through the circuit.

Switches in Series

The power switch, labeled SW1 in the schematic, is *in series* with the rest of the circuit. As mentioned in the previous section, when this switch is open, no current flows through the circuit at all. The switch is intentionally placed at this location in this circuit to act as a bottleneck. When the power switch is turned off, no power flows through the circuit. That seems like a good idea, considering the desired operational goal.

The schematic symbol for SW1 shows a pretty good likeness to the internal wiring of any simple SPST switch. There are two terminals, and a conductor that can be moved to either electrically connect them or electrically isolate them. As drawn, the switch is in the open position.

If we added another identical switch in series with the existing switch, what would happen? Take a look at Figure 5-16.

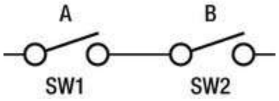

Figure 5-16. *Two switches in series. Both switches must be closed for electricity to flow through this part of the circuit. If either switch is open, the circuit is broken and no current will flow.*

Both switch A (SW1) and switch B (SW2) must be moved to the closed position for any electricity to flow through the circuit. If either switch is opened, then no current will flow. This is sometimes referred to as an *AND* circuit. Both A *and* B must be on for the circuit to operate.

A circuit just like this is often used in security systems. Switches are attached to doors and windows. All of the switches are wired in series, forming a large loop. The switches are all configured to be closed when the doors and windows are closed, and to open when doors or windows are opened.

If all the doors and windows are closed, then the loop of switches will allow the flow of electricity. Although it sounds exactly *backward*, when electricity is flowing in the circuit, this keeps the security alarm *turned off*. When any one of the switches is opened, the circuit is broken. The alarm then sounds.

Most real security systems are more complex than this, of course. The main advantage of this particular configuration is that it is easy to install and requires the least amount of wiring. The chief disadvantage is that when the alarm is triggered, you have no idea exactly *which switch* caused it.

Switches in Parallel

Just like with batteries, switches behave differently when configured in a different manner. Let's connect two switches *in parallel* this time and see what happens. See Figure 5-17.

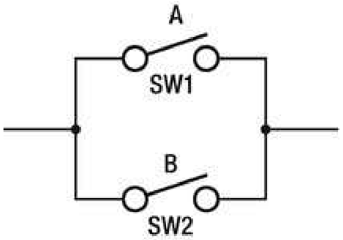

Figure 5-17. *Two SPST switches in parallel. If either one of the switches is closed, then power can flow through this part of the circuit. Both switches must be opened to stop the flow of current.*

The circuit in Figure 5-17 will conduct electricity if switch A (SW1) *or* switch B (SW2) are closed. To stop the flow of current, you have to open both switches. This can be used as a simple *OR* circuit. If switch A *or* switch B is closed, then power can flow through the circuit.

Looking back at the schematics in Figures 5-13 and 5-14, we see that SW1 is in series with the rest of the circuit. This makes it a very effective way to turn power on and off from a single location. Handy!

The Three-Way Switch

You might think that the circuit shown in Figure 5-17 could be used to turn lights on and off from two different locations. You would be partly correct. You could turn lights *on* from either switch location, since flipping either switch to On will cause current to flow and lights (or other appliances) to operate.

Unfortunately, you would not be able to turn *off* the light from either switch if the other switch were on. So this configuration is good for turning lights on but not off. A slightly more complex arrangement is required to do both.

We will need two switches, as before, but they will need to have a little more functionality to accomplish what we've set out to do. Instead of a switch that's either open or closed, we need a switch that alternates between two different routes, depending on how it is positioned. This type of switch is called a *double-throw* switch, as it connects a *common* terminal to either one of two different terminals. It can be *thrown* in two different directions. This should be much easier to understand when looking at a schematic symbol. See Figure 5-18.

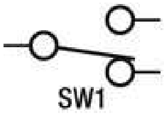

Figure 5-18. *A single-pole, double-throw (SPDT) switch. The common terminal on the left side is connected to either the top or bottom terminal on the right. Variations of this type of switch have a neutral position in the middle. Push-button (i.e., momentary contact) versions of this switch will label the two right-side terminals as either "normally open" or "normally closed," depending on the physical configuration of the switch.*

Now comes the clever bit. We'll connect a pair of these switches together by tying the two non-common terminals on one switch to their counterparts on the other switch. This results in four possible combinations of switch positions (both down, one up, the other up, and both up). Two possibilities turn the light on, and the other two turn the light off. The important point is that the "on" or "off" state of the entire circuit can be changed from either location, independently of the state of the other switch. Problem solved! Again, this should be easier to see when presented as a diagram. See Figure 5-19.

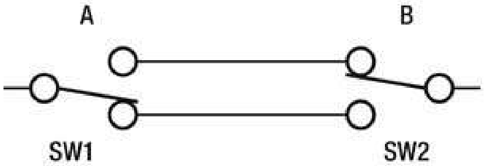

Figure 5-19. *A three-way switch circuit built using two SPDT switches. Power is either carried by the upper wire or the lower wire, as long as both switches are in the same position. If not, then no power flows across this part of the circuit. As shown, the three-way switch is effectively off. Flip either switch to turn it on.*

If you ask a real electrician how to wire a three-way switch, you'll get several answers, or more likely, you'll get, "It depends." That's because *it really does* depend on the actual wiring situation. Sometimes the power is routed through one or the other of the two switches, and sometimes the power is routed through the load. Each of these scenarios requires a different routing of wires and connections.

You are strongly advised to consult a real electrician concerning wiring in your house, where your family, pets, and all your stuff live. Less-than-perfect wiring causes fires or electrocution, which are both very unpleasant ways to die.

■ **Caution** This book does *not* teach you how to wire electricity in your home. Consult a professional.

Resistors in Series

Just like batteries and switches, resistors perform different functions in a circuit depending on their configuration. Let's look at using resistors *in series* first, and then consider using them in parallel later. Both combinations have their uses, as you will see.

Let's look at just a small part of the schematic that we've been working with so far in this chapter. See Figure 5-20.

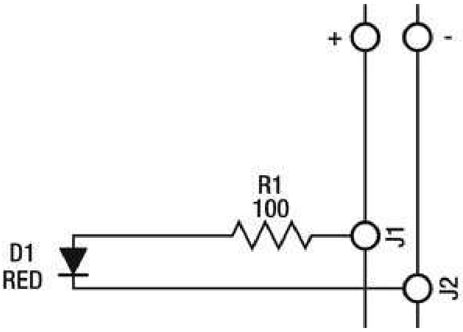

Figure 5-20. *One part of the circuit showing an LED, a resistor, and the power rails from our solderless breadboard. The resistor and LED are connected in series. The battery and power switch are not shown.*

LED Curiosities

D1 is an LED. It is a polarized component and will only work when installed in the proper orientation. The schematic indicates a red LED should be used, but feel free to use whatever color LED you like or happen to have available. The triangle shape of the schematic symbol helps remind us of the direction of *conventional current* through the device. The actual flow of electrons is the other way, but that's only because electrons are negatively charged particles and do everything backward. Apparently this was not well understood when electricity was first being tamed for our use.

LEDs are fascinating devices, emitting light in a wide range of colors, coming in an astounding array of shapes and sizes, and generally requiring very little power to operate. They are almost as easy to use as incandescent bulbs and are certainly more power efficient.

Being semiconductors, however (which leads to their being semimagical in their inner workings), LEDs generally only operate correctly within tightly bounded parameters. For example, a typical LED will only need between 2 and 3V to shine quite brightly. Incandescent lightbulbs, on the other hand, can operate from less than 1V to many hundreds of volts, depending on their construction and intended purpose.

While LEDs have a minimum *forward voltage* requirement that must be met for them to work at all, the brightness of an LED is determined solely by the amount of current flowing through it.

LEDs Have No Self-Control

In addition, LEDs have no internal, inherent mechanism to limit the amount of current flowing through them. Incandescent bulbs, when cold, have a low internal resistance, allowing much current to flow through them when connected to the right power supply. Remember that less resistance means more current can flow, and vice versa. This resistance almost instantly goes way up as the filament of the bulb glows white hot. The filament itself then becomes the limiting factor in how much current flows through the bulb, usually stabilizing at the proper power level.

LEDs have no such sense of self-control. They will conduct as much current though themselves as you are willing to provide, up to and beyond the point where the internal heat generated by the small amount of resistance they do possess causes them to literally burn up. In general, it's usually the tiny, almost microscopic internal bond wires that fizzle out, not the actual LED chip itself. However, dead is dead, and LEDs are generally considered to be "unrepairable" subassemblies. (Believe it or not, that's the correct technical term, even though it's not the correct English word—*irreparable* is the correct term when speaking of non-electrical things).

Care and Feeding of LEDs

The simplest possible solution to this ~~problem~~ opportunity for excellence is the use of a *current-limiting resistor* in series with the LED. Recall that a resistor will resist the flow of current in a circuit. The resistor used in this circuit, R1, is indicated to exhibit a resistance of 100Ω.

Let's say that D1, the red LED in our circuit, requires 2V in normal operation. This is pretty common for red LEDs. Generally speaking, the forward voltage requirement for LEDs goes up as you get farther up the color spectrum. Blue LEDs, for example, often require 3.0V to 3.6V to operate.

If we're using rechargeable cells in our prototype (and we really should be, you know), we can expect 3.6V to be made available to the resistor-plus-LED loop within our circuit. If 2V is needed for the LED, where are the rest of those volts going?

The answer is that they are being used to heat up the current-limiting resistor R1. Resistors turn current into heat. That's what they do. They do it, however, in a very precise and predictable way, which makes it easy for us to calculate how much current, voltage, and heat will be churning around in our circuits.

■ **Caution** Third-grade arithmetic ahead: Addition, subtraction, multiplication, and (shudder) division. You have been warned.

If we started with 3.6V, and two of those volts are being used by the LED, then our good friend Subtraction tells us that we've got 1.6V flowing through R1. For a thorough explanation of why this happens the way it does, feel free to look up Kirchhoff's Voltage Law. It basically says that all the voltages in a circuit end up being equal. The voltage across the resistor and LED will equal the battery voltage. The voltages within the LED-in-series-with-resistor circuit add up, just like battery voltages add up in series.

If we know the voltage across R1 (1.6V) and the resistance of R1 (100Ω), then we can use Ohm's Law to calculate the current flowing through R1. Reviewing what you learned in Chapter 3 about the predictable and inviolable relationships between voltage, resistance, and current, you can see that I = E / R (i.e., current equals voltage divided by resistance), where I is the current in amps, E is the voltage in volts. and R is the resistance in ohms.

It's not that hard! It sounds much worse than it actually is. The answer is 0.016A, or 16mA (1.6V/100Ω). That's the teeny-tiny amount of current needed to light up a typical red LED. Most small LEDs can handle up to 25mA safely, and some can handle much more than that, especially if they are intended for illumination purposes.

One quick side trip to multiplication land and we'll know everything we need to know about R1 in this circuit. While the schematic specified the resistance of R1 as 100Ω (the primary characteristic or component *value*), nothing was said about the power-handling capability of this component.

Power is calculated as voltage multiplied by current. That's all there is to it. We know the voltage: 1.6V. We know the current: 16mA. We multiply them together to get 0.0256W (watts), or almost 26mW. Remember that you have to enter the current as 0.016 because the formula calls for the number of *amps*, not milliamps.

Not quite 26mA is not quite much of anything. You will not be able to tell that the resistor is shedding all those extra volts as heat with your fingertip. It might not even register with a sensitive thermometer. It's just a really tiny amount of power. A typical small resistor (a 1/4W resistor) will handle 250mW of power, which is almost ten times what is needed in this application.

■ **Tip** It's always a good idea to calculate the power requirements of your components, just so you're not *too* surprised when the magic smoke leaves the circuit, rendering it crispy and nonfunctional.

OK, we're back from the frightening land of multiplication. It wasn't *that* bad was it? The remainder of understanding resistors in series depends only upon the science of addition. Counting on your fingers will be good enough.

We can change the brightness of the LED by changing the amount of current flowing through it. We can change the amount of current flowing through the LED by either raising the supply voltage or by changing the resistance of R1. Since we can easily substitute any other value of resistor for the 100Ω resistor with which we began, this is easy to do. Just don't go *too low* in resistor value, as this will allow *too much* current to flow through the LED, possibly damaging it permanently.

Adding resistors in series adds their resistances together. For example, if you placed two 100Ω resistors end to end, the total resistance across them would be 200Ω. That's really all there is to it: 100Ω + 100Ω = 200Ω.

If you perform this experiment with the LED circuit we built on the breadboard, you can see that the LED will not shine as brightly as it did before. This is because the resistance has doubled, which means, all other factors remaining the same, that the current through the LED has been cut in half. Half the current results in half the brightness, since LEDs are controlled by the amount of current flowing through them.

What's interesting about this arrangement is that the power-handling capacity of the two resistors is *also* double what it was when only one resistor was in the circuit. This is because each resistor is now dissipating half the power when compared with the previous scenario. If two 1/4W resistors are used in series, they can together safely dissipate up to 1/2W of power, although pushing any component to 100 percent of its capacity is just asking for trouble, especially when heat is concerned.

By connecting resistors end to end, you add their resistances together, as well as their power-handling capabilities. By keeping a small variety of basic values on hand (e.g., 1Ω, 10Ω, 100Ω, 1KΩ, 10KΩ, etc.), you can easily stack up the right combination needed for almost any electronic application. This is the method used in the very handy device known as a *decade resistor* or *decade box*, which has dials for each digit in the desired resistance, allowing you to dial up any value of resistor needed.

Extra credit: If you've followed along this far, consider researching voltage dividers built using only resistors.

Resistors in Parallel

What happens when resistors are connected side by side instead of end to end? That's where things get interesting, at least from a mathematical point of view.

An easy way to look at parallel resistors is to think about the amount of current flowing through each one, and then add them all up. See Figure 5-21.

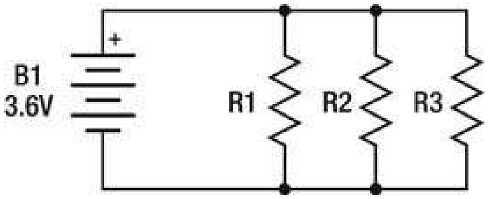

Figure 5-21. *Parallel resistors each conduct a portion of the total current flow through the circuit. Each resistor allows a certain amount of current to flow. The sum of their currents can be calculated if their individual resistances are known.*

Let's assume that R1, R2, and R3 have resistances of 100Ω, 200Ω, and 300Ω, respectively. The total resistance of the circuit across the battery terminals is *not* 600Ω, as would be the case if they were in series instead of parallel. The total resistance is less than 100Ω, which is the least of the resistors. How can this be?

First, let's think in terms of how much current is flowing in each of the three parallel pathways within this circuit. If B1 is putting out 3.6V, and R1 is 100Ω, then we know to use Ohm's Law to calculate the current, with I = E / R. In this case, I = 3.6V / 100Ω = 0.036A, or 36mA. *In addition*, the path through R2 is drawing half that (because R2 is double the value of R1—get it?), or 18 mA. *In addition*, the final pathway is drawing one-third as much current as the first one, or 12 mA. Just to verify: I = 3.6 / 300Ω = 0.012A, or 12mA. Everything checks out.

Now add all those currents together: 36mA + 18mA + 12mA = 66mA. That's the total current, using this method. Again, since we know both the voltage and the current, we can find the *effective* or *equivalent resistance* of all three resistors in parallel, as if they were one single resistor. Rearranging the formula for Ohm's Law, we find that R = E / I. In our case, R = 3.6 / 0.066A = ~54.55Ω. See? Smaller than 100Ω, somehow.

Thinking about the individual *currents* flowing through each resistor is the same as adding the current-supplying capacity of battery cells together when arrayed in parallel. Just add them together. However, we had to calculate each current separately and then add them together. There is a more direct way, depending upon how you look at things, but beware: it involves not only *fractions*, but *fractions upon fractions*. Oh, the horror.

■ **Note** Some people don't dig math. Weird, but true.

Here is the equation for calculating the total resistance (R_T) of three resistors in parallel (R_1, R_2, R_3):

$$R_T = \frac{1}{\left(\frac{1}{R_1}\right) + \left(\frac{1}{R_2}\right) + \left(\frac{1}{R_3}\right)}$$

The good thing about this equation is that you don't need to know what the supply voltage is to get the correct answer. All you need to know are the values of the resistors. If you need to use more resistors, just add more terms to the equation.

Calculating the value of parallel resistors occurs quite frequently when designing electronics. Some special cases are simple enough to do in your head. For example, if two identical-value resistors are placed in parallel, their combined resistance is half the value of each individual resistor. That is, if you placed two 100Ω ohm resistors in parallel, their effective resistance would be 50Ω. Similarly, if you have three identical resistors, the net resistance is one-third the value of each constituent resistor. Make sense? These are special cases of the general equation just shown.

This happens so much that an alternate notation method has been developed: $R_T = R_1|R_2|R_3$. This is not often used for hobby electronics, however.

The power-handling ratings of resistors are added together, so if you have two 10W resistors in parallel, the net power-handling capacity is now 20W. At least that part was easy.

Capacitors in Series and Parallel

Capacitors are completely different from resistors. When calculating their values in series and in parallel, you have to use *exactly the opposite* methods. Isn't electronics interesting?

The values of capacitors *in parallel* are added together. This makes them much more like batteries in parallel than resistors. However, you still have to use the fractions-upon-fractions method to calculate the value of capacitors in series.

Again, some shortcuts are available. Two identical capacitors in series have *half* the capacitance of each individual capacitor, but *double* the working voltage, which is the maximum voltage they were designed to withstand before failing.

Inductors in Series and Parallel

It gets more interesting, believe it or not. Inductors act like resistors as far as adding their values together in series and taking the reciprocal of the sum of the reciprocals of their values in parallel, but *only* when the magnetic fields of all of the inductors involved do not interfere with each other. If the magnetic fields generated when current flows though the inductors overlap, then you have to factor in their *mutual inductance*, which starts to get complicated.

Extra credit: Figure out how all that works and let me know.

LEDs in Series and Parallel

You learned (or were reminded of) a few of the basic characteristics of LEDs in the previous section, although it was really to help illustrate the operation of resistors in series and parallel. A couple of major oversimplifications were introduced to help concentrate on the resistor. We'll address a couple of these issues here.

Lighting up a single LED using a battery and a fixed current-limiting resistor is fun, but it's pretty basic stuff. Lighting up a *lot* of LEDs is where it gets to be a *lot* of fun. To do so, you need to learn a little more about the weirdness that is *semiconductor*.

LEDs are diodes (*LED* stands for *light-emitting diode*, after all). Diodes are semiconductors and are made by elves in a hollow tree. Semiconductors do not have nice linear voltage-to-current curves.

For the purposes of simplicity, in the previous section, it was suggested that LEDs have a certain minimum voltage that they need to operate. This is called their *forward voltage*, and is sometimes referred to as their *voltage drop* within a circuit. So far, this is all true. You can find an LED's forward voltage specification in the data sheet published by the manufacturer.

What you'll find, when you look, is that every manufacturer only publishes a *range* of voltages, such as 1.8V–2.2V, giving a minimum and a maximum voltage. This is what they are contractually obligated to deliver in their product. Anything outside of this range is considered *defective*, or at least out of tolerance. You might get a good deal on "rejects" that fall outside these parameters but still light up quite nicely. Such is the provenance of many of the penny LEDs available from overseas. Yes, they exist. And yes, you get what you pay for.

There are two reasons why LED manufacturers are being so deliberately vague in their specifications. The first is that process variations are going to produce a certain randomness to these values. One batch might be identical to the next batch, but then the third batch might be all over the place. Such things happen.

The second reason is where the real weirdness begins. Forward voltage is not a fixed value in any given device, but instead increases as the current increases. Again, this increase is *nonlinear* and cannot be calculated with a simple formula.

The best way to demonstrate this to yourself is to connect an LED to a power source via a variable resistor, such as a potentiometer. For the protection of your LED, add a fixed resistor in series with the potentiometer so that you establish a maximum current level, even when the potentiometer is turned all the way up (i.e., has almost zero resistance).

Now apply power and adjust the potentiometer until the LED just barely lights up. Now measure the voltage across the terminals of the LED, using a voltmeter. It should be in the neighborhood of 2V–3V. For extra credit, measure the current flowing *through* the LED, using an ammeter. This measurement will most likely be less than 1mA.

Compare these readings to what you find when the LED is at full brightness. The forward voltage across the LED goes up as the current goes up. To map the nonlinearity of this function, you would need to record many readings of both voltage and current, and then plot them in a graph.

Just as you can wire resistors, capacitors, and inductors in series and in parallel, you can do the same with LEDs. As with resistors, the forward voltage of the LEDs are added together when arranged in a series circuit. This raises the minimum voltage required to light up a string of LEDs in series. This means that you can't light up ten LEDs in a single string using a 9V battery. You have to first overcome the minimum forward voltage just to get any current at all flowing. Then you have to up the current to get the desired brightness, which simultaneously raises the total forward voltage through the circuit. You still need a current-limiting resistor in this circuit—but only one is needed. You don't have to have one for each LED.

The amount of current flowing through the string of LEDs is determined by the supply voltage, the cumulative forward voltage of all the LEDs, and the resistance of the current-limiting resistor. The same amount of current is flowing in every part of the circuit. If 25mA is being drawn from the power supply to light up your single string of LEDs, that means that each and every LED has 25mA of juice flowing through it—more than enough to light them up quite nicely.

If you have a 12V power supply (or a car battery), you could string up to five or maybe six red LEDs in series, with a small current-limiting resistor. That's because each red LED will typically have a forward voltage between 1.8V and 2.2V. Other colors higher up the rainbow might need more voltage per LED, meaning fewer can be lit at once using a single string of LEDs in series.

But don't fret. You can light up as many LEDs as you want by placing multiple strings of LEDs in parallel with each other. Each additional string of LEDs adds to the total current draw, so you will eventually find a limit along that dimension as well. For an example of a small LED array using both parallel and series circuits, see Figure 5-22. The schematic is shown in Figure 5-23.

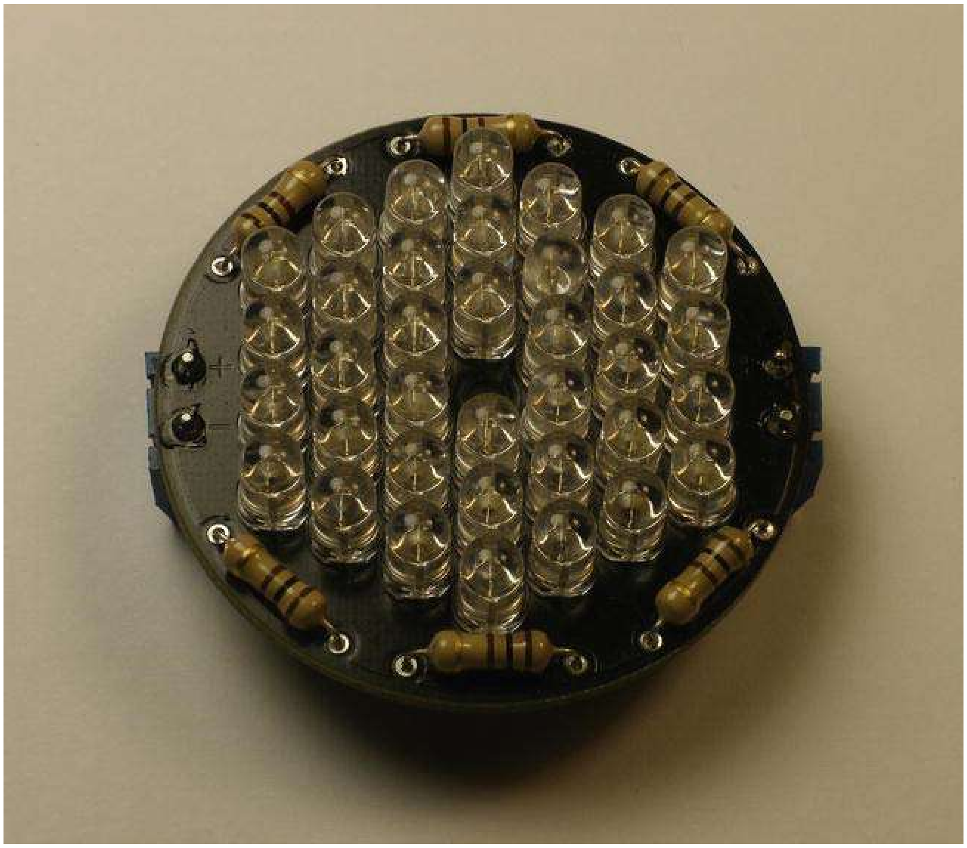

Figure 5-22. *An array of six parallel strings of six infrared LEDs in series. These infrared spotlights are used for covert surveillance and security applications.*

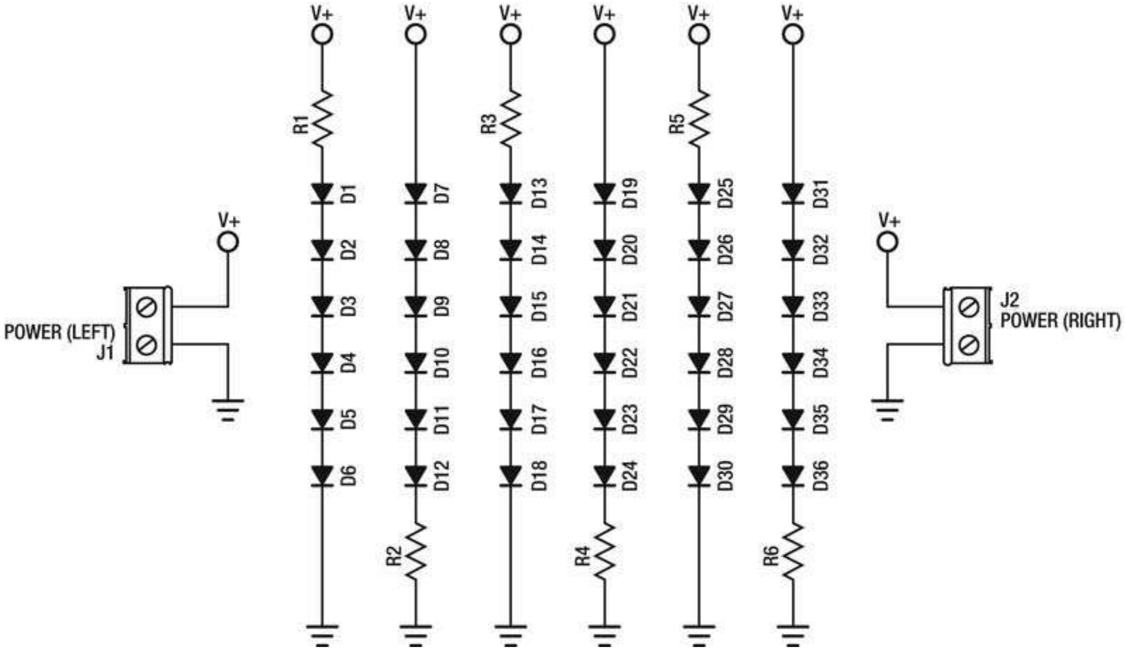

Figure 5-23. *The schematic of the LED array, showing both parallel and series circuits. The circuit offers two power connectors to make it easy to wire multiple LED arrays using the same power supply.*

Summary

Even if your workbench doesn't look as bad as the one presented at the beginning of this chapter (or even if it looks *worse*), hopefully you've learned a thing or two about the importance of they saying, "A place for everything and everything in its place." Since *you* get to decide where everything goes, you can either make it easy or hard on yourself as the years go by.

The workspace in this chapter extended to a total of 10 square feet. If this is what you have to work with, you can certainly set up a comfortable, safe, and productive area for your electronic projects. If you have even more space available (adding a shelf either above or below the desk is a great way to do this), then even better.

Hopefully some of the more detailed descriptions of how electronic components can be wired up have given you some ideas for projects of your own. The very best possible way to learn more about electronics is to think up some simple experiments and try to build them in your lab. Have a goal, take lots of notes, and have some fun—even your failures can be used as stepping stones to more ambitious and creative projects in the future.

Have a *lot* more room? Then perhaps you should be thinking about setting up a group lab or classroom. Chapter 6 will give you some helpful advice about what you are going to need to get started. See you there!

CHAPTER 6

The Small Group Lab and Classroom

The first part of this book covered what you need to know to build your own personal electronics lab. Now it's time to share the love with your fellow electronics enthusiasts. Where to begin?

Like most important questions, the answer remains, "It depends." What resources are available to you and your group? How many folks will likely want to be working together, and for how long? How well do you really know these people? And these are just a few of the questions you're going to encounter along the way.

Take the time to understand exactly who "you" (plural; in Texas, *y'all*) are and what your requirements and goals could be. By doing so, you map out the future of your successful facility. Likewise, *not* planning ahead or making reasonable assumptions about your needs is simply an invitation to problems, problems, and more problems. Who wants *more* problems?

■ **Tip** If you fail to plan, you plan to fail.

Know Yourself

The ancient Greek maxim "know thyself" has as many definitions as it has authors, it seems. Plato indicated in several of his dialogues that Socrates would use this phrase to admonish his students to learn more about themselves and their own characters and capabilities before attempting to figure out the rest of the universe. Wise advice then and now.

As far as knowing yourself in relation to setting up a shared electronics lab, you should first think about what role *you* are going to play in this endeavor. Yes, *you*. Here are some of the major ones, and you might end up being all of them at some point.

Facilitator or Coordinator

In the role of the facilitator, you plan on providing an environment for others to use for their projects and explorations in electronics. You may or may not partake in these activities yourself, but have been tasked with (or volunteered for) the organization and coordination of the lab.

It's critically important to understand the needs of your intended ~~lab monkeys~~ users. See the "Know Your Group" section later in the chapter for more information on this topic.

As the lab facilitator, lab workers will come to you with questions, requests, comments, and complaints. Be prepared to address their needs in a responsive and professional manner—even if they're just your cousin.

Your responsibilities will include

- Communicating the scope of the lab (i.e., deciding what *is* and what *is not* acceptable or expected)
- Publishing any important information or updates regarding location, available times, features, rules, and possibly ongoing projects
- Crafting policies involving safety, proper tool usage and training, and participant conduct
- Maintaining supplies of commonly used materials
- Inspecting and repairing tools as the need arises

To balance out your responsibilities, you should be granted adequate authority in the following areas:

- Characterizing the scope of the lab, based on present and anticipated future needs
- Enforcing the policies of the lab
- Budgetary discretion in acquiring, provisioning, and maintaining lab assets
- Granting or revoking access to the lab based on established parameters

If you're going to be the Keeper of the Lab, then you should have both a personal interest in keeping it in good shape as well as the wherewithal to decide how things should be done.

Good luck finding the right balance there.

Teacher or Presenter

Let's assume that your lab has already been set up (or can be set up in a jiffy), and your main goal is to present information or teach classes there. Again, it should be apparent that a good knowledge of your intended audience will play a vital part in your success. See the "Know Your Group" section for more information.

Once you understand the needs of your class and the goals of your presentation, the next step is to leverage the existing facilities to your best advantage. Will you need specialized presentation equipment or materials? If so, see what is already available, and then make a shopping list of everything else you will need. Examples include lighting, a projector, specialized power distribution, and test equipment.

Are you making an interactive presentation? If so, what materials will you need to provide for your students to be able to fully participate in your presentation?

What will be your optimum teacher-to-student ratio? If you have a simple message that's easily communicated and reinforced, a single speaker can effectively address a rather large crowd. Once you start adding layers of complexity to the subject matter, you're going to have to be prepared to backtrack and explain things in more detail from time to time. If you're leading a group through a technical exercise, it's often profitable to have assistants available to help just those individuals that need a little more help, without having to stop the class in its tracks. These assistants need to be up to speed on the subject matter as well as have good communication and troubleshooting skills.

Remember that students look to the teacher for leadership, not only as a subject-matter expert. Take the lead, set a good example, and share your enthusiasm for your topic with your students.

A successful teacher leads a class on a journey of discovery. Make sure you don't lose too many along the way.

Participant

Your goal as primarily a group lab participant is to make effective use of the facilities at hand, making the most of your time and energy. Having access to a well-stocked lab with a good array of tools can be a decided advantage for both your immediate project goals, and it can be an excellent opportunity to accelerate your knowledge of electronics and other mysterious arts.

Imagining yourself as an active participant is the best perspective when laying out the flow of your lab. Set up workstations that have almost everything you need, leaving room for the complex or specialized stuff elsewhere. There are more suggestions about specific workstation configurations in the "Workstations" section of this chapter.

While the active-participant model helps you squeeze out the maximum utilization of your lab, be sure to consider the sort-of-working-but-also-sort-of-goofing-off model. Not everyone wants to be working full-tilt all of the time. Leave room for a certain amount of idleness, contemplation, and whimsy. You might be surprised at the results.

Observer

You want to be in on the action, but not necessarily working on anything in particular. That's great! Don't underestimate the importance of the purely social aspects of technical gatherings.

Some people just like to hang out with other like-minded individuals. This is a good thing—when resources and circumstances allow it. Watching creative, talented, and driven folks working on projects with passion is very often the best kind of inspiration. Let your kids see you having fun and making cool stuff in your lab, and the next thing you know, they are going to want to be there whenever you are. Pretty soon, they are going to want to start working on projects of their own—and that is its own special reward.

Know Your Group

Now you know a little bit about who you are and what your expectations of your lab may be. Now it's time to get to know the group of people who will most benefit from access to the lab.

Identify the needs and requirements of your group by first identifying *precisely* who your intended audience is going to be. Knowing this one item, up front, before any big plans are laid, is the best way to ensure success down the road.

There are several broad categories for which you might need to prepare. The more you know about your particular group, the better. Let's start fairly close to home and expand outward.

Your Family

Once you've established your own electronics lab at home, it's great to be able to share some of the fun with those that are closest to you. This might be your children, parents, siblings, or extended family. A well-planned and well-maintained electronics lab is going to generate some interest, or at least some curiosity. Here's how you can carve out some room for them, while still being able to get things done yourself.

It's very natural to want to share your hobbies and interest with others, and a safe and well-stocked electronics lab is a great place to do so. This is an especially ripe opportunity for parents to pass on their knowledge, skills, values, and work ethic to the next generation.

The family lab can be a logical extension of the family home. Everything is shared, including responsibilities and privileges. In more agrarian societies, this is already well understood at an early age, where chores around the farm and in the fields and a respect for the tools being used help enhance a sense of group belonging and worth. Everyone has a stake in the prosperity of the family endeavor.

Do you want your kids to appreciate what you give them? Then give them chores before you give them presents. Let them clean up the lab, put away tools, inventory parts, and take out the trash before spoiling them with all the fun stuff you get to do in the lab. This will help foster a sense of ownership in the long-term success of the lab, whether you're there to supervise or not.

It bears repeating (from Chapter 4) that no unsupervised children, no matter how adorable, should be allowed in the lab. Would you allow unsupervised cooking in the kitchen? It's the same basic scenario. When everything goes well, everyone has a good time. It only takes a split-second distraction to go from a good time to a trip to the emergency room in either case. Have rules and enforce them, even if it's your elders that you are protecting by doing so.

Sometimes having a resource like a shared electronic laboratory is the first opportunity people have to assert some sort of control over their relationships. Use the opportunity to make friends with your relatives, while you still can.

Your Friends

Unlike your family, as the saying goes, you get to *pick* your friends. Now you have the opportunity to provide a cool place to hang out as well as build all manner of awesome gizmos. It's only natural to want to share some of the fun with your like-minded friends.

The basic framework of rules and accountability apply to the friendly lab as well as any other group endeavor. Let your friends know up front what your expectations of them are in regard to use of the lab. Can they use your tools? Can they store artwork-in-progress at your place? When is it cool to come over to tinker, and when is it not? These kinds of questions should be dealt with early in the game; otherwise, these issues can flare up and become trouble spots. There's no substitute for good communication.

Once you extend your laboratory's hospitality to those outside your immediate family, it's a good idea to clearly designate what is your personal property and what belongs to the lab and all its participants. This can be as simple as putting name tags on all your tools and marking shared resources appropriately.

This is where *trust* really factors into the success or failure of a group lab. Just like in math, you want to eliminate as many variables as possible to ensure that a good time is had by all, and to ensure the long-term safety and viability of the facility itself. Which of your friends do you *trust* in your lab? Think about these things at length before opening the doors to just anybody.

Coworkers and Fellow Students

So you've been tasked with organizing or supervising a lab for a school class or university department. You may or may not have any choice in who will be granted access, or even when they can avail themselves to the facilities.

The balance of responsibility and authority is crucial in this scenario, just like in most other situations in life. Hopefully you will be in a position to enforce whatever rules are already in place. Ideally, you will have some say in crafting new rules as they become necessary, or adjust existing rules when they need it. In either case, it's just not fair for someone to expect you to do a good job, yet not provide you the tools to get it done.

Positions like this become available from time to time in most schools. Sometimes there's an opportunity for a small paycheck, and other times all you are going to get are points with the administration. Hopefully, at least you will also enjoy lab privileges. That's got to be worth something, yes?

If you are helping to run an existing lab, there are probably others you can consult who have had this particular honor in the past. Try to find out who these folks might be and see if you can get any helpful advice from them about your lab and its care and feeding.

If you're starting from scratch, you will enjoy the double-edged sword of setting things up the way you want, as well as setting things up the way everyone else wants it. You *get* to establish whatever level of functionality you want, with the caveat that you then must maintain that level. It's a freedom, like most freedoms, that comes with a price.

Special-Interest Groups

Are you a member of club that has a special interest in electronics? Several popular interests, such as ham radio operators, remote-control model flyers or drivers, robot builders, and computer users, have a love-hate relationship with electronics. Having a dedicated electronics lab available to the membership is a great way to build up the love as well as work out the hate.

Given a highly specialized crowd, such as the members of a university or trade school, you might even find people organized into an electronics club, especially these days when almost every aspect of our lives is touched by this technology in one way or another.

The specialized nature of your group helps focus what areas of electronics are of most practical interest, at least from a lab point of view. This cascades into easier choices when it comes to stocking the lab with workstations, tools, and components.

The Dallas Personal Robotics Group (DPRG; see `http://dprg.org`), for example, had the use of an 1,800-square-foot warehouse for over six years, rent free, due to the generosity of one of the members, Mike Dodson of Modern Assemblies. Having access to an excellent industrial setting allowed the club members to dream of and then realize much more ambitious robot projects than could be attempted at home. After Mike's retirement, the DPRG went into a nomadic phase and finally spun off a dedicated workshop and meeting area, now called the Dallas Makerspace.

The General Public

A lot of people see the popularity of do-it-yourself (DIY) and *maker-movement* projects and envision themselves as entrepreneurs, seeing a need and being able to meet it. Unfortunately, most of these people are wrong.

Building a lab dedicated to public use sounds like a great and noble aim. The realities of such an undertaking, including the financial, logistical, and business related, are not to be underestimated. While it seems to work well in certain areas and at certain times, it's almost always the result of the right combination of people coming together at the right time.

Hackerspaces (a terribly unfortunately chosen label) are springing up all over the world. The term *hacker,* while understood within some technical communities to mean a person of respectable skills, has a decidedly *unrespectable* taint to it when viewed from the outside. To the vast majority of the English-speaking world, a hacker is a rogue computer jockey, a criminal intent on breaking and entering into computer systems for nefarious purposes.

The term *maker,* on the other hand, leaves a more positive first impression. The Dallas Makerspace members deliberately made this decision when naming themselves, while they were still just a subcommittee within the board of directors of the nomadic DPRG, looking for a new home. It was important to them to maintain the very positive image that the DPRG had established and built over its long history, while still looking to the future.

If you're going to set up a lab for use by the general public, whether you decide to call it a hackerspace, a makerspace, or something entirely different, you get to make almost *no* assumptions about the good intentions of your membership. It's a sad state of affairs, to be sure, but it seems to be the prevailing wisdom of the day.

Some popular and successful organizations—such as Noisebridge (`www.noisebridge.net`), in San Francisco, California, which was incorporated in the state as a "non-profit educational corporation for public benefit"—operate almost entirely on a "consensus basis," sometimes known as "anarchic chaos." Noisebridge's overriding motto is "Be excellent to each other," very often shortened to just "Be excellent." Its very real success and sense of community among members is based on what seems to be a very local tolerance that is not to be found in many parts of the globe. If you ever get a chance, do stop by and prepare to be impressed. It's an amazing place.

A good parallel would be to visit your local public library. In the areas that still maintain libraries (shockingly, they are on the decline), you will find a finely tuned machine that meshes well with the surrounding locality. Are there bars on the windows? If so, you're going to need some bars on your windows. Are there metal detectors at the entrance? Oh, dear. Do you see any uniformed security forces present? These are not good signs.

Don't be completely discouraged, however. Sometimes a community in decline needs a little boost of energy and an injection of esprit de corps. This can often be accomplished by introducing some new alternatives for people's spare time. An ancient Muslim saying reads, "Trust in Allah, but tie your camel."

A Shared Resource

Do you have a dedicated space available? If so, you are way ahead of the game. Finding a good location is half the battle. If your space is truly dedicated to your pursuits and activities (i.e., a private home, club, or makerspace), you have the advantage of being able to leave tools, components, and projects in place, returning to them (and finding them still there!) when the opportunity arises. "Truly dedicated" in this sense can be equated with having your own key to the door and reasonably free access. While a 24-hour facility sounds like a great idea, there are a lot of logistical problems associated, such as security, utility costs, scheduling, and parking, just to name a few.

If your space is not-so-dedicated, such as a school lab or friend's garage, you can still make it a home away from home if you are able to arrange for secure storage of all your shiny toys. This could take the form of individual lockers or a storage closet.

No matter what the logistical arrangements, the main challenge to providing a versatile and effective setting for electronic and other kinds of tinkers is basically a social one. Once it's more than just you, it can start to get complicated.

Luckily, there are some time-tested methods for making these types of things run smoothly.

Be a Hospitable Host

If you are taking the lead in arranging a shared work space, please remember that you are responsible for extending the appropriate hospitality to your present and future guests. This single factor will greatly determine the vibe, feeling, or ambiance of the facility, even more than having the latest and greatest tools and gizmos on hand.

This task may or may not fall entirely on your shoulders. Many hands make light work. If you are able to delegate some of the duties of a hospitable host to other members of your group, consider doing so.

Think of your task in the same light as planning a successful party, even if it's one that you anticipate will last for days on end. These basic ideas are as old as civilized living. Think about fabulous parties or memorable events that you have been privileged to attend in the past, and then plan on taking it up a notch.

Try to anticipate the human needs of your attendees as well as their technological expectations. People have what seems like an endless list of requirements, some of which you can address, some not.

Make your location known. Make this information easy to find to those interested, and easy to update when the venue changes. A web site is trivial to set up these days, and even a post to a popular forum will tend to be accessible via Internet search engines, in time. There's a new social media outlet popping up every third Tuesday, so avail yourself to what makes sense for your audience. Word of mouth only goes so far these days.

Once you've gone to the trouble of publishing your location, go the extra mile and put out some sort of actual sign or indicator that your arriving guests have found the right spot. If you have an identifiable logo or symbol associated with your group, be sure to proudly wave that flag. People don't like feeling lost, or unsure whether they're entering a welcoming space or about to trespass on private property.

Let everyone know when they are expected as well as when they are *not* expected. Post your regular meeting times, if you have them. Alternatively, make known the appropriate hours of operation. It's disappointing to be turned away at the door just because it's 1:00 a.m. on a school night.

When possible, welcome your guests. This makes a big impression on both newcomers as well as the regulars. Being part of a group is supposed to *enhance* a feeling of inclusiveness and belonging. We are social animals at heart, and even a simple hello is often the prelude to an interesting and productive time together.

Have rules. Make these rules known. A sign on the front door is not out of place, as long as it's not too negative or off-putting. Rules should *prevent* problems from occurring, and not be enacted as a spontaneous reaction when problems inevitably occur.

Ensure your guests' comfort when attending. Is there enough room for the largest (and the smallest) participant to safely enter and exit the facilities? Is there adequate lighting in the common areas, if they exist? Is there a place for everyone to sit? Who has dominion over the thermostat?

Do you plan on making refreshments available? Will this duty rotate among the membership? Keep track of such things, if only to keep from overextending yourself in all the myriad details of hosting a successful meeting. While massive banqueting and consumption of yummies make for a festive and enjoyable association, these activities are not always aligned with maintaining an effective laboratory. Separate the feeding trough from the work areas. Stay focused on what you're trying to accomplish here.

Will everyone have access to restrooms, and whose job is it to keep these areas tidy?

And the most important question of all: Who takes out the trash?

Be a Courteous Guest

Just as it's important to play the role of the hospitable host, it is likewise critical to be a courteous guest when visiting a facility, whether you are a paid-up member in good standing or an invited guest. Again, these guidelines will vary according to the intents and purposes of your lab, but the basics will always remain the same.

If possible, and when appropriate, let your host know if you are planning on attending. This won't apply in a causal, drop-in type of laboratory. Even so, by making your intentions of attending known ahead of time, you might influence (one way or the other) the attendance of other potential participants. Nothing succeeds like success. A well-attended party has a better potential for awesomeness than a sparsely peopled event.

Arrive on time or within the established time frames set forth by your host. Similarly, unless you have been granted special access, plan on leaving at a reasonable time as well. Benjamin Franklin wrote in his *Poor Richard's Almanack* (1736) that "Fish & Visitors stink in 3 days." Don't be a stinker.

Try to contribute something. Don't show up empty-handed, expecting to be fed, entertained, and enlightened.

Workstations

Now it's time to talk about your workstations. Even if it's just you in the lab, having a separate workstation for each type of activity is a great idea. For example, you might have one workstation set up for programming a computer, another for soldering electronics, and another for working with power tools, such as a drill press or a saw.

Workstations dedicated to a particular task should have everything within easy reach, including tooling, supplies, power, and lighting. Don't waste time by forcing people to wander around the lab looking for the right drill bit or magnifying glass. Keep everything close so you can find what you need, when you need it.

Having a place for everything ahead of time is the best way to follow the wise adage, "A place for everything and everything in its place." You get to decide where everything lives. It doesn't hurt to go ahead and label everything conspicuously so that others can find their way around the lab.

If you have several people wanting to work on the same kinds of projects at the same time, you're going to need multiple, nearly identical workstations. The exact number depends on your available space and the number of people you expect to want to participate simultaneously, as well as your budget.

Most indoor-oriented work can often be performed at a work table. Exceptions include giant-scale mechatronics, pyrotechnics, and underwater robotics. Some of their constituent bits and pieces, however, can be crafted on an adequately prepared work table.

Make sure each work area has plenty of power outlets available. You're almost guaranteed to run short at some point. Take care not to overload your circuits, however. If you anticipate needing a lot more power than is presently provisioned, you're going to have to consider the real possibility of having your facility expanded by qualified, *licensed* professionals. These requirements will vary from one location to the next. Check with your local building codes to be sure.

As repeated often in previous chapters, you can never have *too much* lighting available, especially for the more detail-oriented chores. Typical indoor room lighting is just about adequate for finding your way out of the room, and that's about all. You'll need area lighting for each workstation, and possible additional task lighting, depending on your particular requirements.

Providing network access to each workstation can be challenging. Wireless networking is a good way to accommodate a relatively small number of users at once. More than a handful of computers trying to access the Internet at the same time can bring many wireless routers to a slow crawl. Dedicated high-speed Ethernet connectivity is more expensive and takes more effort to deploy, but offers much greater bandwidth to your lab-folk.

If you're especially cramped (or *dimensionally constrained*) in your lab, you'll need to get inventive. For example, a table placed in the center of the room can be approached from multiple sides. There's no reason all the workbenches have to be pushed up against a wall, allowing access from only a single side. Get creative!

Safety Planning

While it's tempting to fill up all the available space in a lab with workstations, especially when the budget permits it, do try to leave a bit of common area here and there. You always want to be able to walk *away* from a workstation with the same (or sometimes even more) rapidity you approached it with. Fire is an excellent example of such a motivator.

■ **Tip** Keep a properly selected and maintained fire extinguisher at each workstation, if possible.

Post emergency procedures prominently, where everyone, including visitors, can see. Have an evacuation plan, complete with map and "You are here" indicator posted at each workstation. Hopefully you will never need it.

Presentations

Are you excited about your latest project? Have you just learned a great new skill? Setting up your lab for group presentations is an excellent way to communicate these ideas to your friends, family, and coworkers.

Make your ~~victims~~ guests comfortable. Have adequate seating for everyone, with plenty of room between the rows (if you have more than one row of chairs). Ensure that everyone has a good view of what you intend to present.

Will the attendees be expected to take notes? Then perhaps you should set up tables where they can comfortably jot down some of the more spectacular aspects of your presentation. If you will be providing a handout or other documentation packages, make sure you have enough for the anticipated demand. Will you be entertaining questions or comments either during the presentation or afterward? Then make sure that not only can your audience see *you*, but that *you* can see *them*.

If you're going to be using some sort of projector for your presentation, by all means set aside some time to get to know the equipment ahead of time. Nothing ever seems to work just exactly as you imagined that it should. This seems to be an invariant law of presentation physics. No one is really sure why.

Your audience's time is just as valuable as your own. In fact, if there are only two people attending, then their collective time is worth approximately *twice* as much as yours. Think about it. Have all your ducks in a row, so to speak. Wasting time figuring out which cable goes where is not a good use of your time together.

Summary

Wow! You can really go crazy setting up a shared electronics lab, can't you?

Before you even flip on the light in your shiny new group lab, you should already have an insightful idea of who is going to be using it and why. You should also have some honest expectations of what your own expectations are going to be, and how you would like to interact with both the lab itself and its population.

The details, of course, are always going to be specific to your particular situation. Be on guard, however, for the future. It's going to sneak up on you, someday. Not too long ago, for example, it took a fairly hefty desk to keep a modern personal computer up off the floor. That kind of computing power now fits in a pocket. Think of what better uses you could dream up for that space these days!

That kind of forward thinking is exactly what turns the dreams of today into the reality of tomorrow. It's forward-thinking people like you, dear reader, that help map out those realities. Take the opportunity you have right now to spend some quality time daydreaming about how things *could* be. Then take those steps necessary to make it happen. Your family, your society, and your world expect great things from you. You can do it!

APPENDIX A

Getting Started with Tool Building

Once you've become comfortable in your lab and have had time to play with a few circuits, you might start to notice that some of the "tools" you've been using are just simple electronic circuits themselves. Good examples are power supplies and meters. These are great tool-building projects because you can see useful results early.

Let's take a look at building a couple of simple electronic tools that might be of use in your lab. Hopefully you will develop a better understanding of how these tools work. Ideally, you will progress from the basic question of "Does it work?" to the more involved questions of "How well does it work and what can I do to improve it?"

Please bear in mind that these exercises, while hopefully useful, are to illustrate what needs to go into the building of practical, reliable tools. At every stage you should be thinking about what design choices you have and which choices will enhance the usefulness of your tool-building process.

A Very Simple Power Supply

Almost all electronic projects need electricity to work. Having a good, dependable power supply at your disposal on your bench will get you started faster on new projects. You won't have to worry about dead batteries and flaky power adapters. See Figure A-1.

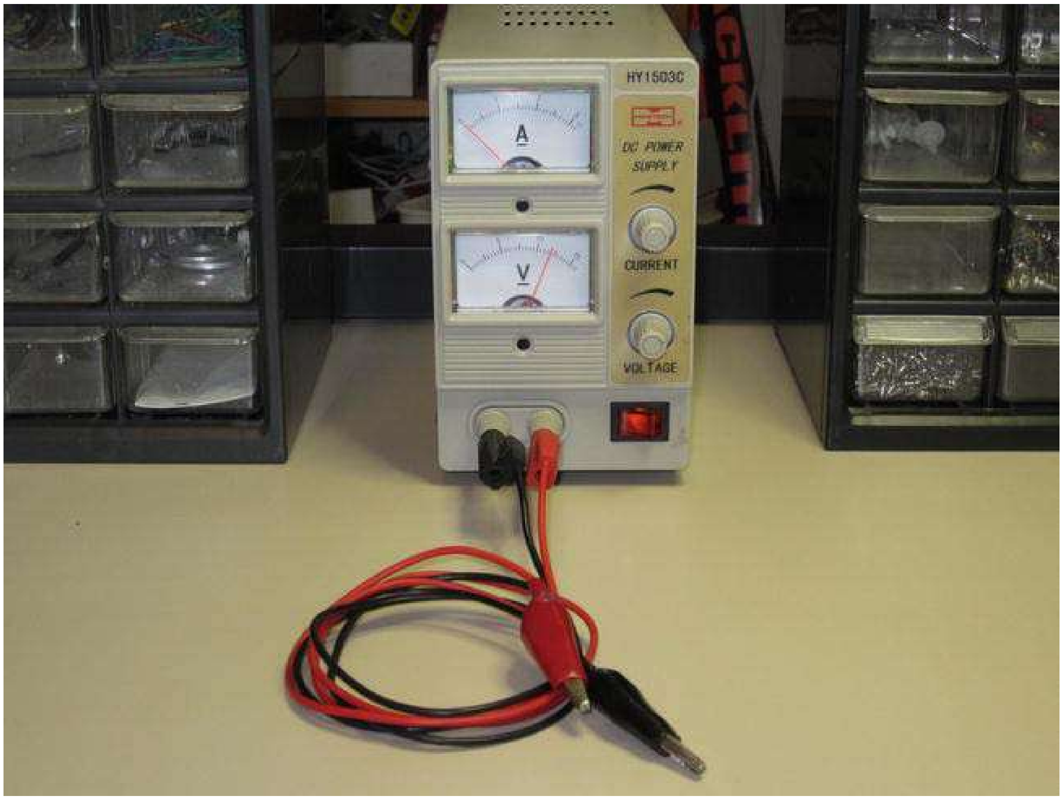

Figure A-1. *A professional bench power supply. This device allows both the voltage and the maximum current to be adjusted. It provides two dedicated meters to show how much power is flowing: one for voltage and the other for current. It also provides a modular set of plugs (banana plugs) for easily connecting and disconnecting power leads.*

A bench power supply such as the one shown in Figure A-1 has many features and includes extra circuitry to make it safe and easy to use. It will also set you back a good chunk of change (i.e., cost money). We're going to build something a bit simpler to start. Feel free to add more bells and whistles as your needs arise.

■ **Caution** Anything plugged into the wall has the potential to shock, burn, or kill you. Do not defeat their safety features!

An Easy 2.5 Watts

You probably have a spare USB charger lying around somewhere. If you don't already, you more than likely will at some point in the near future, as these are becoming more common with smaller electronic devices. If not, you can obtain one relatively cheaply if you do some smart shopping. The USB charger shown in Figure A-2 was obtained from overseas via eBay for less than $1. You will also need a USB cable about 3 or 4 feet long. Again, don't spend a lot of money on a new cable. If you don't already have one floating around in a drawer somewhere, you probably know someone that does.

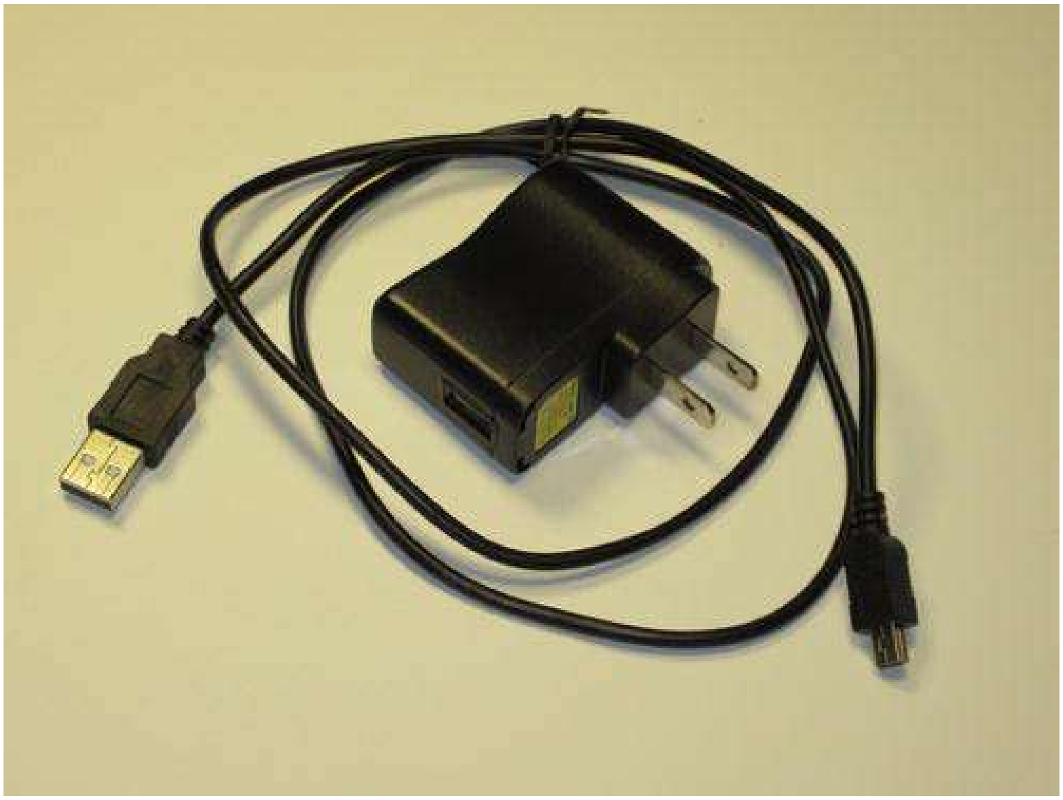

Figure A-2. *A USB charger that plugs into the wall and provides a regulated +5V supply. This plug is intended for US wiring, but will operate from 110AC to 240VAC at 50 or 60Hz, which pretty much covers the whole world, if you have the right adapter plug.*

It doesn't matter what type of USB cable you find, because we're going to cut off the smaller end and use the bare wires inside the cable. It just has to have the USB A–style plug shown on the left side of Figure A-2. The plug will fit into the socket of the charger. You do not need to open or modify the USB charger at all, and you are seriously advised not to do so for any reason.

Let's take a closer look at the label on the charger to find out exactly what it's supposed to be doing for us. See Figure A-3.

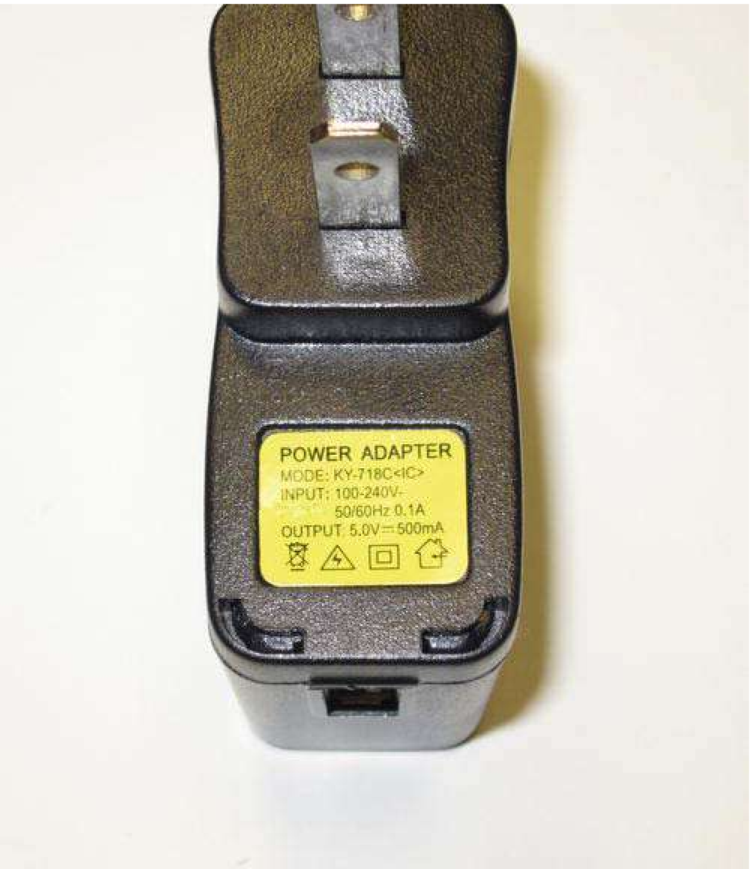

Figure A-3. *The label of the USB charger tells us that it accepts anything from 110VAC to 240VAC at either 50Hz or 60Hz. This wide range of input power allows the same internal electronic design to be used worldwide. Only the prongs that plug into the wall vary from location to location. The label also tells us that the USB charger will output +5V at up to 500mA. That's 2.5W!*

The "OUTPUT: 5.0V ⎓ 500mA" marking tells us what we need to know about the flavor of the power coming out of the adapter. The "5.0V" is the measure of the voltage. This is a very useful voltage level for small projects. Many digital chips require a 5V supply, although 3.3V is also becoming quite popular.

The ⎓ symbol represents *direct current (DC)* voltage. The symbol for *AC (alternating current)* voltage is the tilde, or wavy line symbol ~. The "500mA" is the maximum current that can be supplied by this charger (500 milliamps, or 0.5A), although you are cautioned not to expect a low-cost item such as this one to produce 100 percent of its rated power 100 percent of the time.

While this USB charger will most likely work for our intended purposes, it's best to keep those expectations in check. Note that "Model:" is misspelled ("Mode:"); and also note the absence of the magic letters *USB* or the USB logo. Also, there is a complete lack of any safety certifications present on this humble device, as befits its ~~cheap~~ economically optimized provenance. Compare this with the veritable constellation of safety certifications on the name-brand laptop charger shown in Figure A-4.

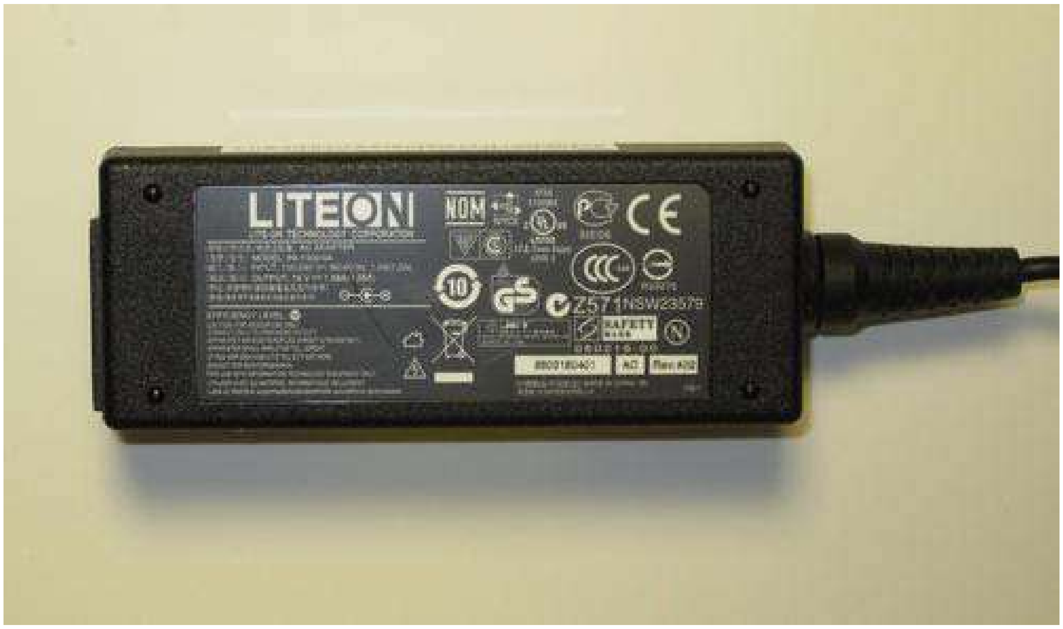

Figure A-4. *A name-brand laptop charger with several international safety certifications. Every one of these certifications costs money to obtain, and shows that the device was demonstrated to be safe for use in a typical home or business application.*

Next we adapt the USB cable to our needs. In the words of the Red Queen, “Off with his head!”

■ **Caution** Make sure the USB charger is not plugged into the wall or a power strip when making modifications to the cable.

First, plug the USB cable into the USB charger. Now cut off the *other* end of the cable. See Figure A-5.

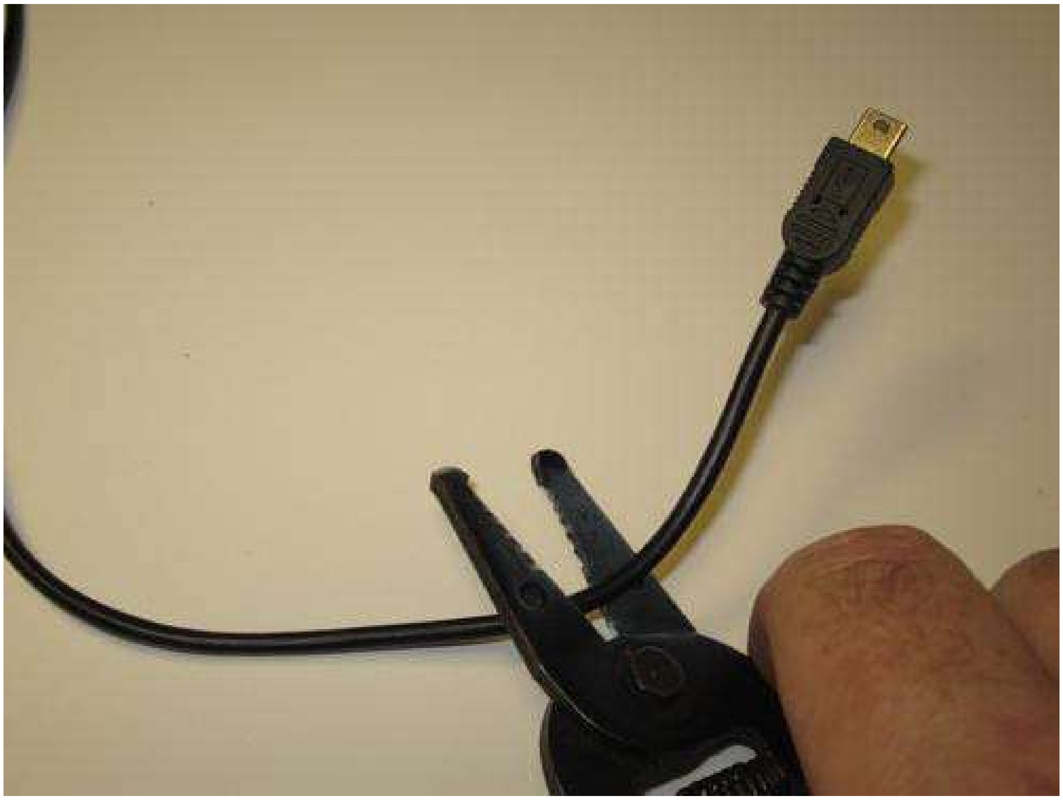

Figure A-5. *Cut off the device end of the USB cable. If you cut off the wrong end, you're not going to be able to plug it back into the charger, at least not without a considerable amount of effort. To ensure you cut the correct end, plug the host end of the cable into the USB charger before snipping, and then cut off the other end.*

Now carefully strip off about an inch of the cable's insulating jacket. Do your best to not nick the insulation on the wires inside. Within the jacket you should find four wires. Two of these wires carry the +5V and ground from the power supply. The other two wires carry the data on the USB connection. Since we're only interested in the power, we can ignore the other two wires for now. See Figure A-6.

s

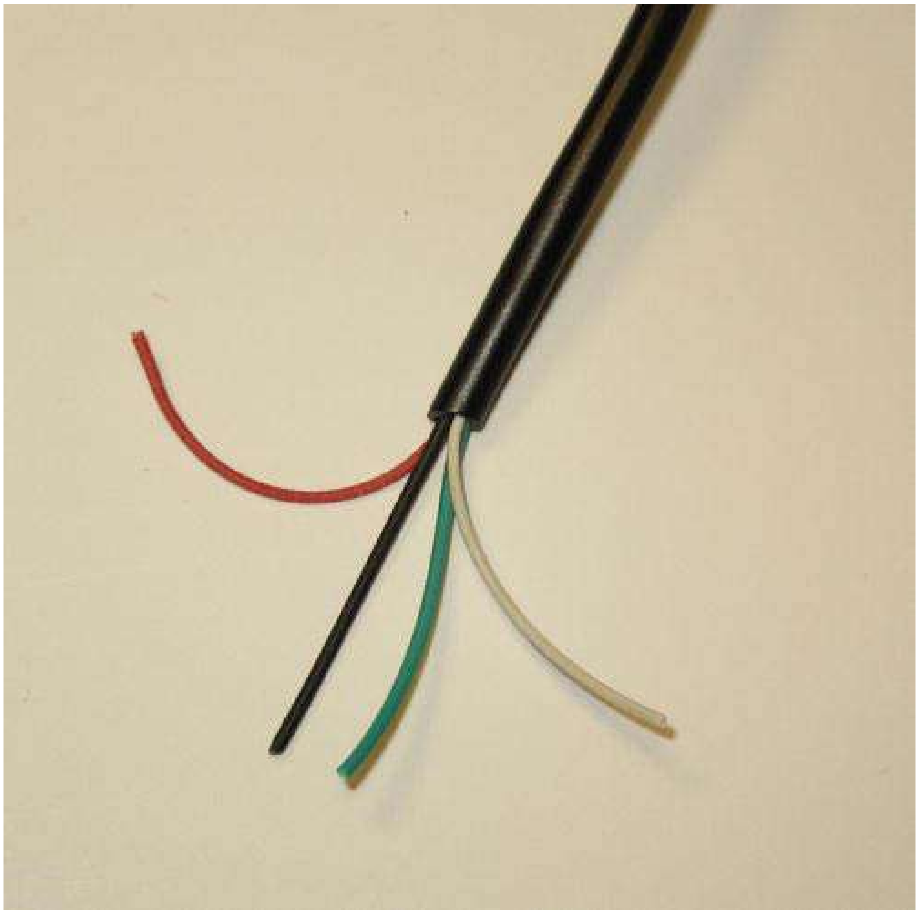

Figure A-6. *The four internal wires within the USB cable. From left to right they are red (VBUS or +5V), black (ground), green (USB D+), and white (USB D–). Use the red and black wires to pull a regulated +5V from the charger.*

The USB standard specifies every little detail about how to build and mark USB cables, even down to the colors of the internal wires in the cable that you're not ever supposed to see! That's dedication.

■ **Note** Apple violates the USB standard by not having a raised USB symbol molded into its cables.

However, sometimes cable manufacturers fudge a bit and use whatever four-conductor cable they happen to have lying around. A very common substitute cable uses red, blue, yellow, and brown wires internally. See Table A-1 for a list of which wires go where.

Table A-1. *Standard USB Wire Colors and Their Functions, Along with Popular Alternate Colors*

Function	Pin Number	Standard USB Color	Alternate Color
VBUS (+5V)	1	Red	Red
D– 2		White	Blue
D+ 3		Green	Yellow
GND 4		Black	Brown

Now is the time to verify the wiring before we get much further. Even if your cable has the official wire colors within the insulating jacket, you are strongly advised to double-check both the voltage and the polarity coming out of the charger before connecting it to your tender, sensitive circuits. Minimize your assumptions about the compliance of others. Testing proves testing works!

Strip about 1/4 inch of the insulation from the ends of the red and black wires. Get out your meter and set it up to read voltage at or above the 5V range. If your fancy meter has autoranging, then just set it to read DC voltage. Plug the USB cable into the charger, if it's not already installed. Make sure the red and black wires are not touching, and then plug the charger into the wall or a convenient power strip.

Now measure the voltage between the red and black wires. Hopefully you should see a reading at or near 5.0V. The USB standard specifies a tolerance of ±10 percent, so anything between 4.5V and 5.5V is considered OK. See Figure A-7.

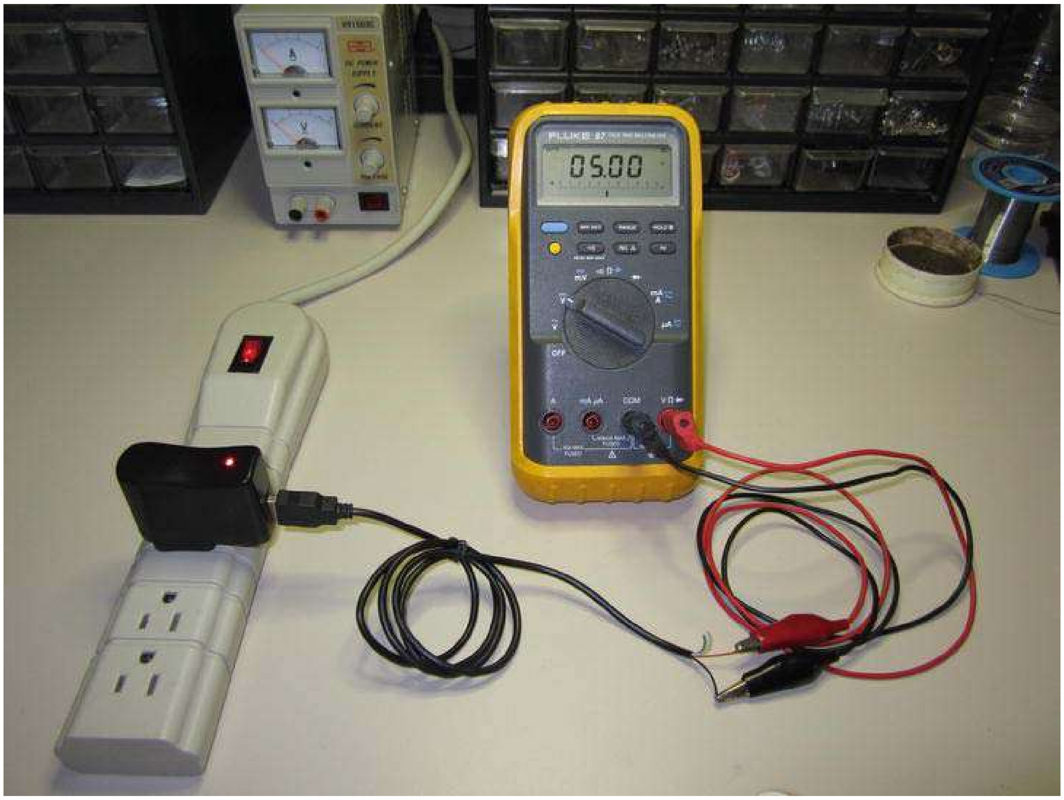

***Figure A-7.** Measure the voltage coming from the USB charger before connecting it to any of your circuits. This one reads 5V on the dot! This is a much safer method than trial by fire (i.e., plugging it into your circuit and looking for a fire).*

Now that you've confirmed that the red and black wires do indeed carry the regulated +5V from the USB charger, you can clip off the unused green and white wires. At this point you can either use the bare, stripped wires of the USB cable or solder on a more robust connector. Since this power supply is destined to be used with a solderless breadboard, a small, two-pin connector with the appropriate spacing will be attached. See Figure A-8.

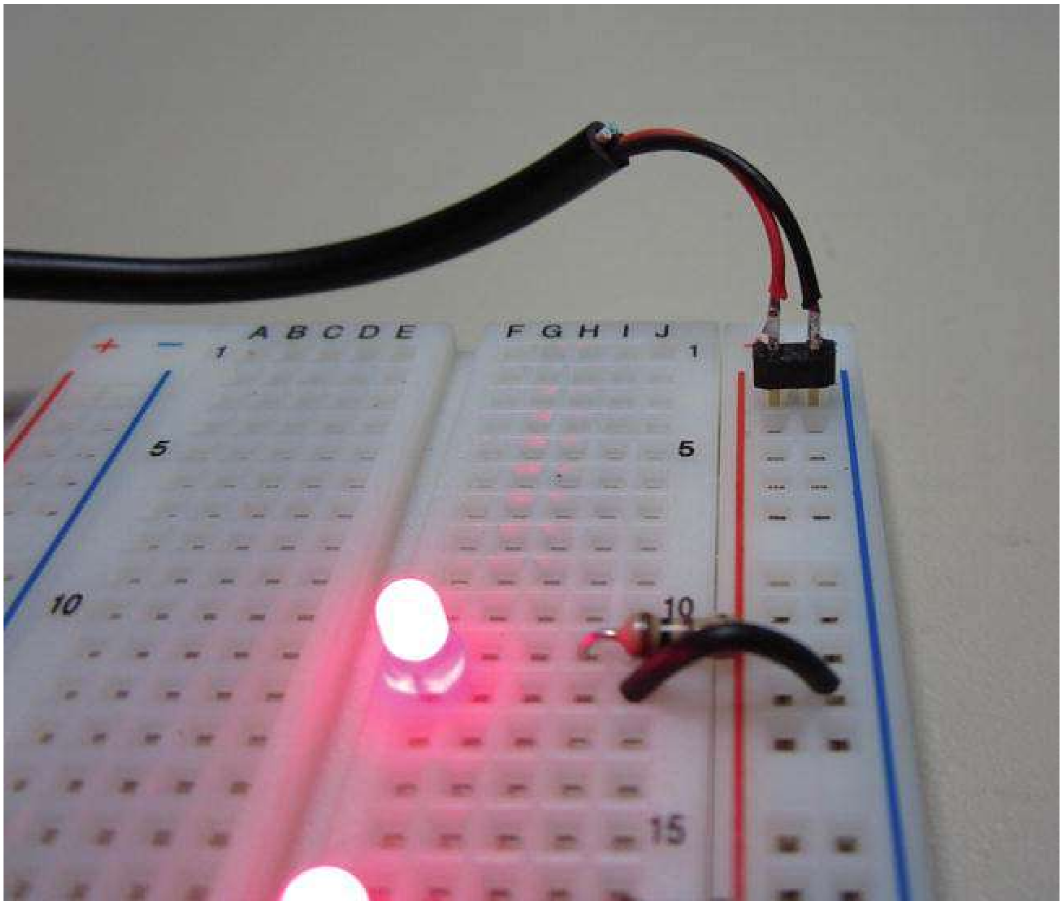

Figure A-8. *A gold-plated pin connector is soldered to the modified USB cable. The red and black insulation is left visible to be a polarity indicator when installing the cable to the solderless breadboard. Another happy LED!*

Adding Features

This power supply works just fine for most projects. If your requirements are, "Translate AC power from the wall into regulated DC power I can use for small projects, without spending a lot of money," then this contraption fits the bill precisely.

You can see for yourself in Figure A-8 that it does indeed work as advertised. Looking back at Figure A-7, you can see that the charger itself has a power-on indicator LED built into the housing. What it doesn't have is a power control or a switch. The power to the circuit is controlled by the switch on the power strip. This works just fine as long as you don't mind turning on and off everything else that happens to be plugged into the same power strip.

Various Indicators

Not all USB chargers have a power indicator, however. Some have both a power indicator LED as well as a *charge status indicator*, which can tell you at a glance if your USB device is fully charged or still charging. A charge status indicator is usually implemented by measuring the amount of current being drawn from the supply. If more than a preset current is flowing, the power supply assumes a battery is still being charged. When that current drops below a certain level, it assumes that the battery has been completely (or maybe just mostly) charged. Some chargers use a red or orange LED to indicate charging, and a green or blue LED to show that the battery is finished charging. Again, this type of charge status display is provided to give the user a quick status update with a single glance. You don't need to consult the product manual to interpret the status code, if you can remember that red means "charging" and green means "charged."

Let's rearrange some of the parts on the solderless breadboard to make room for some new features. See Figure A-9.

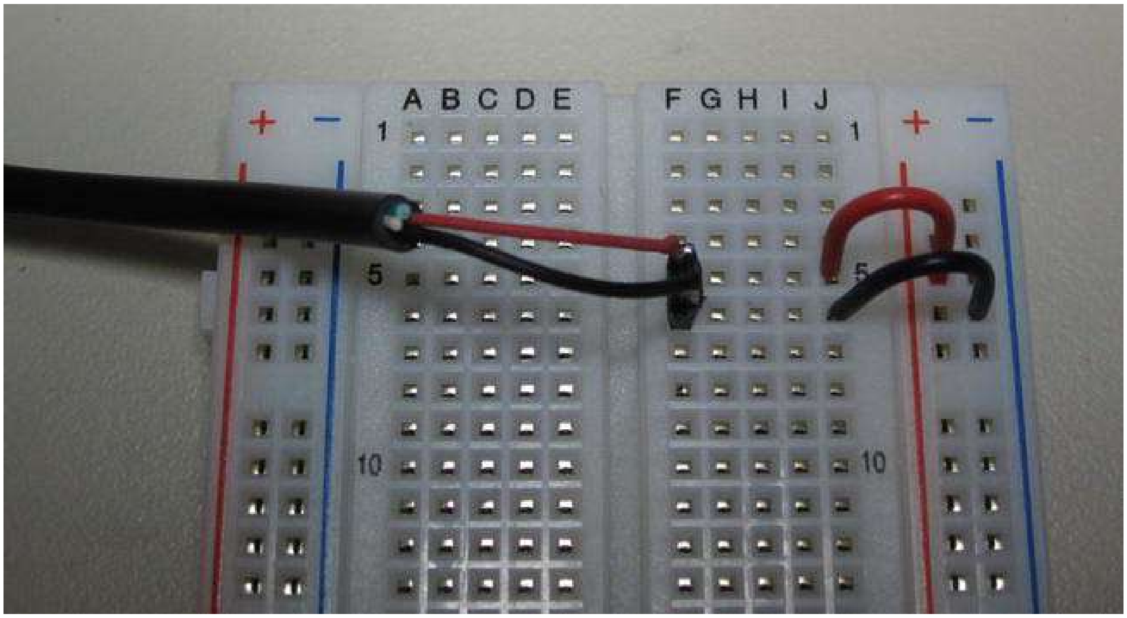

Figure A-9. *The regulated power coming from the USB charger is no longer directly connected to the breadboard's power rails. Instead, it is routed to some internal tie points and then connected to the power rails via two short jumpers.*

Power-On Indicator

First, we add a local power-on indicator by reinstalling the LED and current-limiting resistor used in the previous experiments. While this duplicates the function of any LEDs that might be present on the USB charger itself, it brings the indicator closer to the area of interest, which in this case is the circuitry onboard the solderless breadboard.

The anode of the LED goes into tie point F-3, with the cathode being installed at tie point F-4. The current-limiting resistor spans the gap between tie point I-3 and the right-hand positive (red) power rail. A short jumper connects tie point J-4 and the right-hand negative power rail. See Figure A-10.

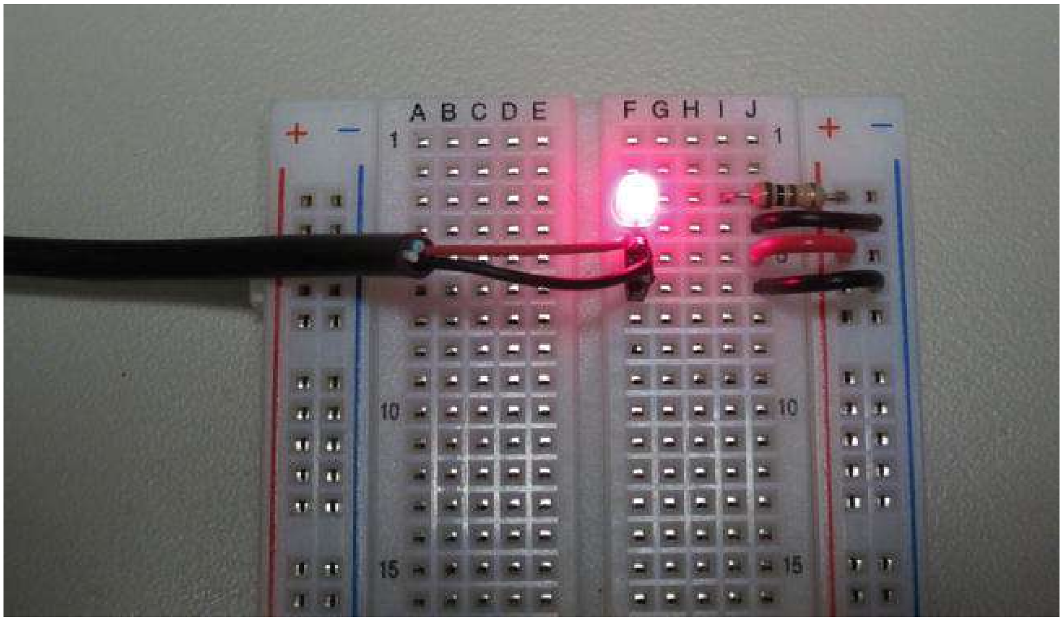

Figure A-10. *Add an LED and a current-limiting resistor across the power rails of the solderless breadboard to give an at-a-glance power indicator, right where the action is taking place.*

Now we have a good power-on indicator, right where we need it. It may or may not be redundant, depending on the features of your USB charger. This type of indicator is digital in the sense that it tells us that the circuit is either on or off, without giving us a lot of precise information about anything in between. For a lot of applications, however, this is more than enough information.

Power Supply Voltage

What if you *did* either want or need more precise information about the power supply? One method would be to permanently connect your multimeter to the circuit so that you can see how many volts are being delivered. That method assumes that you're not going to need your multimeter for anything else, ever.

A practical variation of that method, however, it to provide convenient *test points* in your circuit so that you can quickly and easily take a voltage reading when the need arises. Be sure to position the test points so that they are easy to access. Having a nice, big, readable label wouldn't hurt, either. Also, try to arrange things so that it's not likely that contact with a metal object, such as a tool or your wristwatch, will accidentally short out the power supply by touching both test points at the same time. Even simple "features" get complicated quickly, it seems!

If you want an accurate (and permanent) voltage reading on your circuit, consider adding a voltmeter. Small panel meters are relatively inexpensive and reasonably accurate. A DC voltmeter has two leads, positive and negative, that are attached across the voltage to be measured. See Figure A-11.

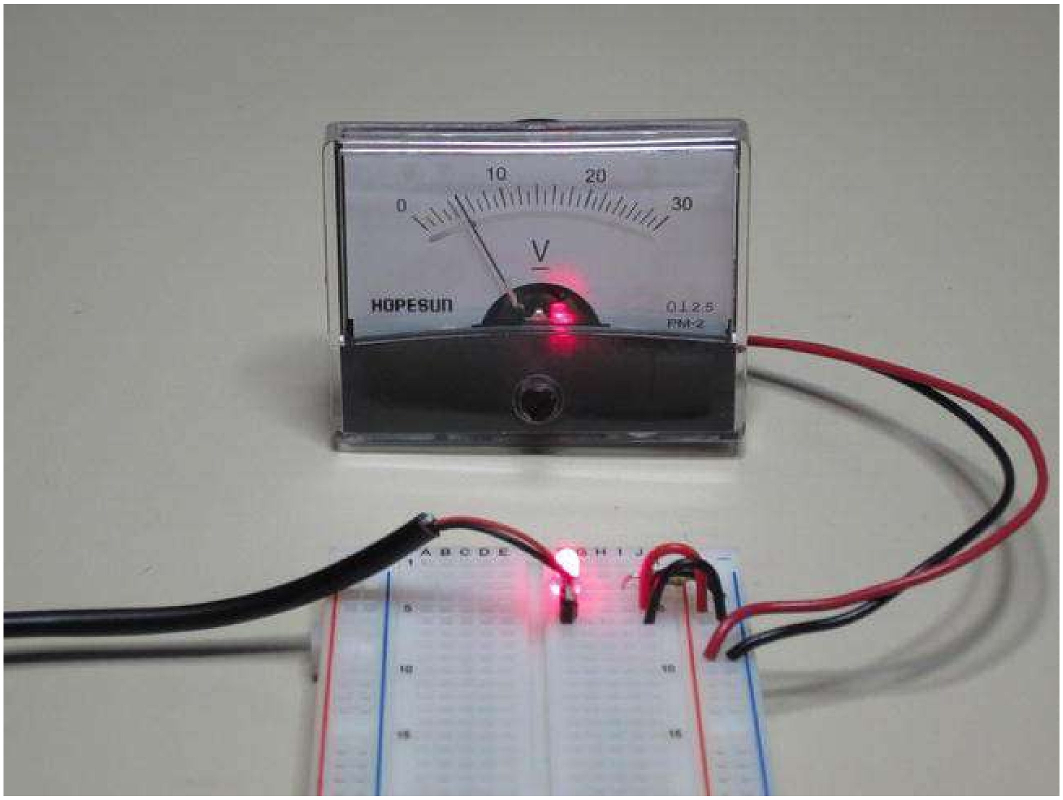

Figure A-11. *A DC voltmeter is attached to the circuit to accurately measure the supply voltage. The V on the meter's legend stands for voltage. The underline beneath the V indicates that it is designed to measure DC voltages, as opposed to AC voltages.*

The panel meter shown in Figure A-11 has a range of 0–30VDC. That's a bit more than we need, since our little power supply is only supposed to produce 5.0VDC ±10 percent. Since the meter's range does not accurately match the expected range of supply voltages, we lose a bit of precision when attempting to read the display. All the action is crowded together on the left side of the meter.

It might seem that a 0–5VDC meter would be preferable, but that would not allow for any overvoltage conditions at all. Remember, the USB standard allows for up to 5.5VDC to be delivered over the cable. A 0–6VDC meter, if available, would be more suitable. The meter introduced in Figure A-11, however, would be perfect when used with the laptop supply shown in Figure A-4, because its output voltage is 19.0V. That leaves a bit of room for overvoltage conditions and gives a wide enough swing of the needle to produce fine detail.

Power Supply Current

Measuring the *voltage* at any point in a circuit is easy. Just put a voltmeter between two points and voilà! In our previous example, we measured the VBUS power supply voltage with reference to GND, the ground potential of the circuit. Many voltage measurements are made with respect to ground.

To measure the current flowing through a circuit, we have two options. The first option is the most direct route. Install an ammeter (a meter that measures *amps* of current) inline with the circuit you want to measure. This requires that you *break* the circuit to make room for the meter. Once installed, the current flowing in the circuit actually flows through the ammeter as well.

The second option involves using a bit of trickery and some math. Instead of installing an ammeter directly inline with the current flow to be measured, we install a resistor of known resistance and measure the *voltage* that develops across the resistor.

Since we know that the values of volts, ohms, and amps are always related in any circuit, we can use Ohm's Law to determine how much current is flowing through a circuit by measuring the voltage across a known resistance.

To make things very easy, let's use a 1Ω resistor in series with the power supply, inserted in the circuit between where the positive lead of the power supply arrives at the circuit and where it is connected to the rest of the circuitry.

Unlike our previous use of Ohm's Law to calculate the missing value based on two known values, we only use *one* known value (the value of the resistor, called a *shunt resistor* when used in this manner). This one known value (resistance) then establishes the relationship between the two other values, voltage and current. As the current increases in the circuit, so does the voltage across the resistor. The two values will mirror each other—as one increases or decreases, the other will follow.

Since we used a 1Ω resistor in the circuit, this sets the relationship between current through the circuit and voltage across the resistor at a 1:1 ratio. For every ampere of current through the circuit, for example, 1V will be measured across the resistor.

Our power supply is only supposed to be able to deliver 0.5A of current, maximum, so we'll be dealing with voltages of less than 0.5V, unless something terribly unexpected happens.

Again, we could permanently attach a multimeter across the leads of a 1Ω resistor inserted into the circuit, or we could attach a voltmeter that is sensitive enough to give accurately readable information in the range we need. The 0–30VDC meter used before is much too large in range to do us any good here. We might see a tiny wiggle in the needle, but that's all.

Instead, we'll use a different kind of meter entirely. A digital voltmeter contains its own dedicated measurement circuitry and a numeric display to indicate the voltage reading. It also requires its own separate power supply. See Figure A-12.

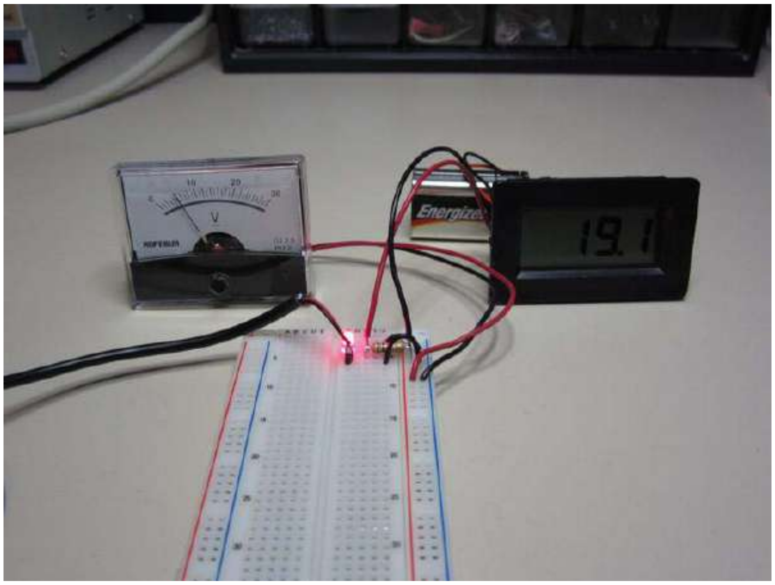

Figure A-12. *A digital voltage meter is used to read the voltage across the shunt resistor installed in the circuit. Because the shunt resistor has a 1.0Ω value, Ohm's Law dictates that the reading in millivolts on the digital meter correspond exactly to the number of milliamps flowing through the circuit. Math is useful, sometimes.*

The jumper wire connecting the incoming power supply directly to the power rail has been replaced with a 1Ω resistor. Leads from both ends of the resistor have been connected to the input terminals of the digital voltage meter on the right. The digital meter, unlike the analog meter on the left, requires an independent power supply to operate, in this case a 9V battery.

So now we've added some features to our simple power supply, including a power-on indicator, a supply voltage reading, and a supply current reading.

These are very useful features to have on any power supply. However, at this point in the prototyping stage, we've just barely covered what the added circuitry is supposed to do. Packaging all this up into an easy-to-use and reliable device takes as much if not more thought and effort.

First, you have to rebuild this entire circuit in a more permanent fashion. Next, you will have to wedge all that clever gadgetry into some sort of enclosure to protect the delicate bits while still permitting access to the inputs, outputs, and controls, if any. Then you get to label everything so that you can remember how to use it in the future.

Summary

So now you just might be developing a certain respect for how much work goes into building even the simplest tools—tools that are useful, accurate, reliable, and an asset to your toolbox. Remember, every tool on your workbench has to earn its place there. This applies to tools you design and build yourself, as well.

Good luck with your tool-building projects. Sometimes it's just as much fun to work on your tools as it is to work *with* your tools.

VOL. 1

Charles Platt

Encyclopedia of Electronic Components

Power Sources & Conversion

Resistors ▪ Capacitors ▪ Inductors
Switches ▪ Encoders ▪ Relays ▪ Transistors

O'REILLY®

Make:
makezine.com

How to Use This Book

1

To avoid misunderstandings regarding the purpose and method of this book, here is a quick guide regarding the way in which it has been conceived and organized.

Reference vs. Tutorial

As its title suggests, this is a reference book, not a tutorial. In other words, it does not begin with elementary concepts and build sequentially toward concepts that are more advanced.

You should be able to dip into the text at any point, locate the topic that interests you, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will not find concepts being introduced in a sequential, cumulative manner.

My book *Make:Electronics* follows the tutorial approach. Its range, however, is more circumscribed than that of this encyclopedia, because a tutorial inevitably allocates a lot of space to step-by-step explanations and instructions.

Theory and Practice

This book is oriented toward practicality rather than theory. I am assuming that the reader mostly wants to know how to use electronic components, rather than why they work the way they do. Consequently I have not included any proofs of formulae, any definitions rooted in electrical theory, or any historical background. Units are defined only to the extent that is necessary to avoid confusion.

Many books on electronics theory already exist, if theory is of interest to you.

Organization

The encyclopedia is divided into entries, each entry being devoted to one broad type of component. Two rules determine whether a component has an entry all to itself, or is subsumed into another entry:

1. A component merits its own entry if it is (a) widely used or (b) not-so-widely used but has a unique identity and maybe some historical status. A widely used component would be a **bipolar transistor**, while a not-so-widely-used component with a unique identity would be a **unijunction transistor**.
2. A component does not merit its own entry if it is (a) seldom used or (b) very similar in function to another component that is more widely used. For example, the *rheostat* is sub-

sumed into the **potentiometer** section, while *silicon diode*, *Zener diode*, and *germanium diode* are combined together in the **diode** entry.

Inevitably, these guidelines required judgment calls that in some cases may seem arbitrary. My ultimate decision was based on where I would expect to find a component if I was looking for it myself.

Subject Paths

Entries are not organized alphabetically. Instead they are grouped by subject, in much the same way that books in the nonfiction section of a library are organized by the Dewey Decimal System. This is convenient if you don't know exactly what you are looking for, or if you don't know all the options that may be available to perform a task that you have in mind.

Each primary category is divided into subcategories, and the subcategories are divided into component types. This hierarchy is shown in Figure 1-1. It is also apparent when you look at the top of the first page of each entry, where you will find the path that leads to it. The **capacitor** entry, for instance, is headed with this path:

power > moderation > capacitor

Any classification scheme tends to run into exceptions. You can buy a chip containing a *resistor array*, for instance. Technically, this is an *analog integrated circuit*, but should it really be included with solid-state relays and comparators? A decision was made to put it in the **resistor** section, because this seemed more useful.

Some components have hybrid functions. In Volume 2, in the *integrated circuit* subcategory, we will distinguish between those that are *analog* and those that are *digital*. So where should an **analog-digital converter** be listed? It will be found under *analog*, because that category seems better associated with its primary function, and people may be more likely to look for it there.

Primary Category	Secondary Category	Component Type
power	source	battery
	connection	jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electro-magnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
		stepper motor
discrete semi-conductor	single junction	diode
		unijunction transistor
	multi-junction	bipolar transistor
		field-effect transistor

Figure 1-1. *The subject-oriented organization of categories and entries in this encyclopedia.*

Inclusions and Exclusions

There is also the question of what is, and what is not, a component. Is wire a component? Not for the purposes of this encyclopedia. How about a **DC-DC converter**? Because converters are now sold in small packages by component suppliers, they have been included as components.

Many similar decisions had to be made on a case-by-case basis. Undoubtedly, some readers will disagree with the outcome, but reconciling all the disagreements would have been impossible. Speaking personally, the best I could do was create a book that is organized in the way that would suit me best if I were using it myself.

Typographical Conventions

Throughout this encyclopedia, the names of components that have their own entries are presented in **bold type.** Other important electronics terms or component names are presented in *italics* where they first appear in any one section.

The names of components, and the categories to which they belong, are all set in lower-case type, except where a term is normally capitalized because it is an acronym or a trademark. *Trimpot*, for instance, is trademarked by Bourns, but *trimmer* is not. **LED** is an acronym, but *cap* (abbreviation for **capacitor**) is not.

Where formulae are used, they are expressed in a format that will be familiar to computer programmers but may be unfamiliar to others. The * (asterisk) symbol is used in place of a multiplication sign, while the / (slash symbol) is used to indicate division. Where pairs of parentheses are nested, the most deeply nested pair identifies the operations that should be performed first.

Volume Contents

Practical considerations relating to book length influenced the decision to divide *The Encyclopedia of Electronic Components* into three volumes. Each volume deals with broad subject areas as follows.

Volume 1

Power, electromagnetism, and discrete semiconductors.

The *power* category includes sources of power and methods to distribute, store, interrupt, and modify power. The *electromagnetism* category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the main types of diodes and transistors.

Volume 2

Integrated circuits, light sources, sound sources, heat sources, and high-frequency sources.

Integrated circuits are divided into analog and digital components. *Light sources* range from incandescent bulbs to LEDs and small display screens; some reflective components, such as liquid-crystal displays and e-ink, are also included. *Sound sources* are primarily electromagnetic.

Volume 3

Sensing devices.

The field of sensors has become so extensive, they easily merit a volume to themselves. *Sensing devices* include those that detect light, sound, heat, motion, pressure, gas, humidity, orientation, electricity, proximity, force, and radiation.

At the time of writing, volumes 2 and 3 are still in preparation, but their contents are expected to be as described above.

Safari® Books Online

Safari Books Online is an on-demand digital library that lets you easily search over 7,500 technology and creative reference books and videos to find the answers you need quickly.

With a subscription, you can read any page and watch any video from our library online. Read books on your cell phone and mobile devices. Access new titles before they are available for print, and get exclusive access to manuscripts in development and post feedback for the authors.

Copy and paste code samples, organize your favorites, download chapters, bookmark key sections, create notes, print out pages, and benefit from tons of other time-saving features.

O'Reilly Media has uploaded this book to the Safari Books Online service. To have full digital access to this book and others on similar topics from O'Reilly and other publishers, sign up for free at *http://my.safaribooksonline.com*.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

MAKE
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

MAKE unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. MAKE celebrates your right to tweak, hack, and bend any technology to your will. The MAKE audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our educational system—our entire world. This is much more than an audience, it's a worldwide movement that Make is leading—we call it the Maker Movement.

For more information about MAKE, visit us online:

MAKE magazine: *http://makezine.com/magazine/*
Maker Faire: *http://makerfaire.com*
Makezine.com: *http://makezine.com*
Maker Shed: *http://makershed.com/*

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at:

http://oreil.ly/encyc_electronic_comp_v1

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

battery

2

This entry covers electrochemical power sources. Electricity is most often generated electromagnetically, but since these sources cannot be classified as components, they are outside the scope of the encyclopedia. Electrostatic sources are excluded for similar reasons.

A battery is sometimes referred to as a *cell* or *power cell*, but can actually contain multiple cells, as defined in this entry. It used to be called an *accumulator* or a *pile*, but those terms are now archaic.

OTHER RELATED COMPONENTS

- **capacitor** (see Chapter 12)

What It Does

A battery contains one or more *electrochemical cells* in which chemical reactions create an electrical potential between two immersed terminals. This potential can be discharged as *current* passing through a *load*.

An electrochemical cell should not be confused with an *electrolytic cell*, which is powered by an external source of electricity to promote *electrolysis*, whereby chemical compounds are broken down to their constituent elements. An electrolytic cell thus consumes electricity, while an electrochemical cell produces electricity.

Batteries range in size from *button cells* to large *lead-acid* units that store power generated by solar panels or windmills in locations that can be off the grid. Arrays of large batteries can provide bridging power for businesses or even small communities where conventional power is unreliable. Figure 2-1 shows a 60KW, 480VDC self-watering battery array installed in a corporate data center, supplementing wind and solar sources and providing time-of-day peak shaving of energy usage. Each lead-acid battery in this array measures approximately 28″ × 24″ × 12″ and weighs about 1,000 lb.

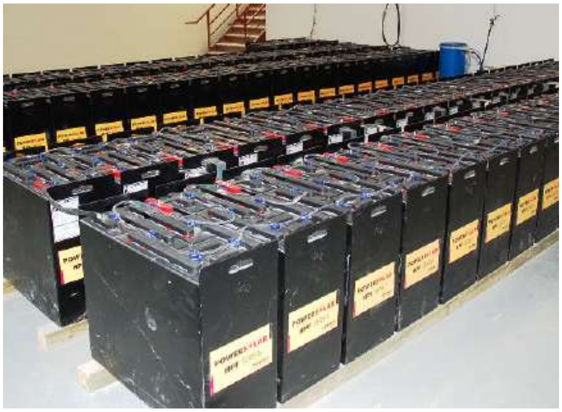

Figure 2-1. *A battery array providing 60KW at 480VDC as backup for a corporate data center. (Photo by permission of Hybridyne Power Systems, Canada, Inc., and the Hybridyne group of companies. Copyright by Hybridyne, an internationally registered trademark of Hybridyne Power Systems Canada Inc. No right of further reproduction unless specifically granted by Hybridyne.)*

Schematic symbols for a battery are shown in Figure 2-2. The longer of the two lines represents the positive side of the battery, in each case. One way to remember this is by imagining that the longer line can be snipped in half so that the two segments can combine to form a + sign. Traditionally, multiple connected battery symbols indicate multiple cells inside a battery; thus the center symbols in the figure could indicate a 3V battery, while those on the right would indicate a voltage greater than 3V. In practice, this convention is not followed conscientiously.

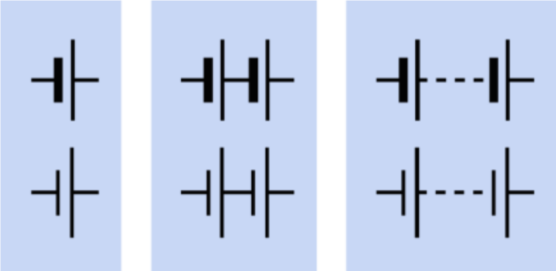

Figure 2-2. *Schematic symbols for a battery. Each pair of symbols within a blue rectangle is functionally identical.*

How It Works

In a basic battery design often used for demonstration purposes, a piece of copper serves as an *electrode*, partially immersed in a solution of copper sulfate, while a piece of zinc forms a second electrode, partially immersed in a solution of zinc sulfate. Each sulfate solution is known as an *electrolyte*, the complete battery may be referred to as a *cell*, and each half of it may be termed a *half-cell*.

A simplified cross-section view is shown in Figure 2-3. Blue arrows show the movement of electrons from the zinc terminal (the *anode*), through an external load, and into a copper terminal (the *cathode*). A *membrane separator* allows the electrons to circulate back through the battery, while preventing electrolyte mixing.

Orange arrows represent positive copper *ions*. White arrows represent positive zinc ions. (An ion is an atom with an excess or deficit of electrons.) The zinc ions are attracted into the zinc sulfate electrolyte, resulting in a net loss of mass from the zinc electrode.

Meanwhile, electrons passing into the copper electrode tend to attract positive copper ions, shown as orange arrows in the diagram. The copper ions are drawn out of the copper sulfate electrolyte, and result in a net accumulation of copper atoms on the copper electrode.

This process is energized partially by the fact that zinc tends to lose electrons more easily than copper.

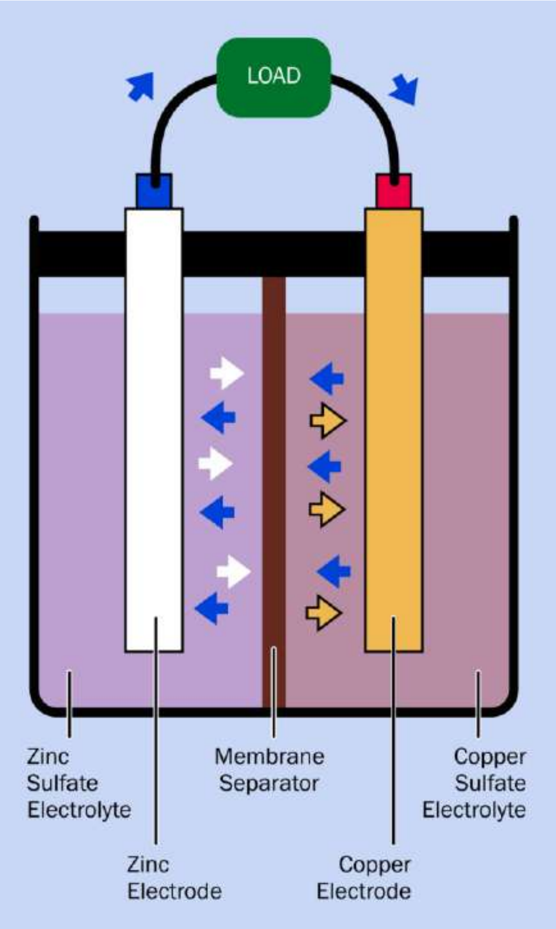

Figure 2-3. *A classically simple electrochemical cell. See text for additional details.*

Batteries for use in consumer electronics typically use a paste instead of a liquid as an electrolyte, and have been referred to as *dry cells*, although this term is becoming obsolete. The two half-cells may be combined concentrically, as in a typical 1.5-volt C, D, AA, or AAA alkaline battery (see Figure 2-4).

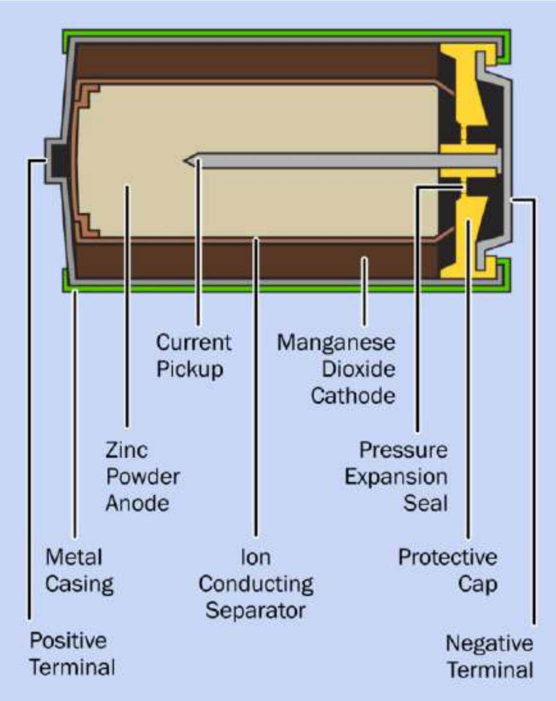

Figure 2-4. *Cross-section view of a typical 1.5-volt alkaline battery.*

A 1.5V battery contains one cell, while a 6V or 9V battery will contain multiple cells connected in series. The total voltage of the battery is the sum of the voltages of its cells.

Electrode Terminology

The electrodes of a cell are often referred to as the *anode* and the *cathode*. These terms are confusing because the electrons enter the anode inside the cell and leave it outside the cell, while electrons enter the cathode from outside the cell and leave it inside the cell. Thus, the anode is an electron emitter if you look at it externally, but the cathode is an electron emitter if you look at it internally.

Conventional current is imagined to flow in the opposite direction to electrons, and therefore, outside the cell, this current flows from the cathode to the anode, and from this perspective, the cathode can be thought of as being “more positive” than the anode. To remember this, think of the letter t in “cathode” as being a + sign, thus: ca+hode. In larger batteries, the cathode is often painted or tagged red, while the anode may be painted or tagged black or blue.

When a reusable battery is recharged, the flow of electrons reverses and the anode and the cathode effectively trade places. Recognizing this, the manufacturers of rechargeable batteries may refer to the more-positive terminal as the anode. This creates additional confusion, exacerbated further still by electronics manufacturers using the term “cathode” to identify the end of a **diode** which must be “more negative”(i.e., at a lower potential) than the opposite end.

To minimize the risk of errors, it is easiest to avoid the terms “anode” and “cathode” when referring to batteries, and speak instead of the negative and positive terminals. This encyclopedia uses the common convention of reserving the term “cathode” to identify the “more negative” end of any type of diode.

Variants

Three types of batteries exist.

1. *Disposable batteries*, properly (but infrequently) referred to as *primary cells*. They are not reliably rechargeable because their chemical reactions are not easily reversible.
2. *Rechargeable batteries*, properly (but infrequently) known as *secondary cells*. They can be recharged by applying a voltage between

the terminals from an external source such as a *battery charger*. The materials used in the battery, and the care with which the battery is maintained, will affect the rate at which chemical degradation of the electrodes gradually occurs as it is recharged repeatedly. Either way, the number of charge/discharge cycles is limited.

3. *Fuel Cells* require an inflow of a reactive gas such as hydrogen to maintain an electrochemical reaction over a long period. They are beyond the scope of this encyclopedia.

A large **capacitor** may be substituted for a battery for some applications, although it has a lower energy density and will be more expensive to manufacture than a battery of equivalent power storage. A capacitor charges and discharges much more rapidly than a battery because no chemical reactions are involved, but a battery sustains its voltage much more successfully during the discharge cycle. See Figure 2-5.

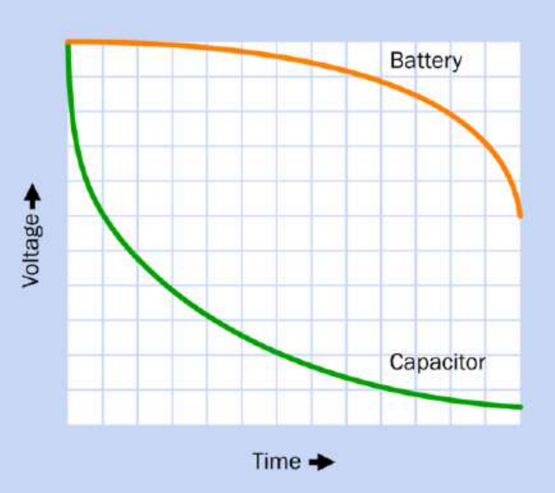

Figure 2-5. *The voltage drop of a discharging capacitor is much steeper initially than that of a battery, making capacitors unsuitable as a battery substitute in many applications. However, the ability of a capacitor to discharge very rapidly at high amperage can sometimes be a significant advantage.*

Capacitors that can store a very large amount of energy are often referred to as *supercapacitors.*

Disposable Batteries

The energy density of any disposable battery is higher than that of any type of rechargeable battery, and it will have a much longer shelf life because it loses its charge more slowly during storage (this is known as the *self-discharge rate*). Disposable batteries may have a useful life of five years or more, making them ideal for applications such as smoke detectors, handheld remotes for consumer electronics, or emergency flashlights.

Disposable batteries are not well suited to delivering high currents through loads below 75Ω. Rechargeable batteries are preferable for higher-current applications. The bar chart in Figure 2-6 shows the rated and actual capabilities of an alkaline battery relative to the three most commonly used rechargeable types, when the battery is connected with a resistance that is low enough to assure complete discharge in 1 hour.

The manufacturer's rating of watt hours per kilo is typically established by testing a battery with a relatively high-resistance load and slow rate of discharge. This rating will not apply in practice if a battery is discharged with a C-rate of 1, meaning complete discharge during 1 hour.

Common types of disposable batteries are *zinc-carbon cells* and *alkaline cells*. In a zinc-carbon cell, the negative electrode is made of zinc while the positive electrode is made of carbon. The limited power capacity of this type of battery has reduced its popularity, but because it is the cheapest to manufacture, it may still be found where a company sells a product with "batteries included." The electrolyte is usually ammonium chloride or zinc chloride. The 9V battery in Figure 2-7 is actually a zinc-carbon battery according to its supplier, while the smaller one beside it is a 12V alkaline battery designed for use in burglar alarms. These examples show that batteries cannot always be identified correctly by a casual assessment of their appearance.

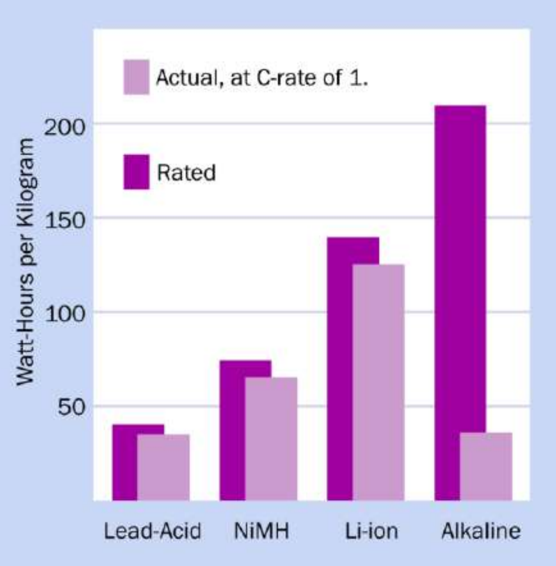

Figure 2-6. *Because of their relatively high internal resistance, alkaline batteries are especially unsuited to high discharge rates, and should be reserved for applications where a small current is required over a long period. (Chart derived from http://batteryuniversity.com.)*

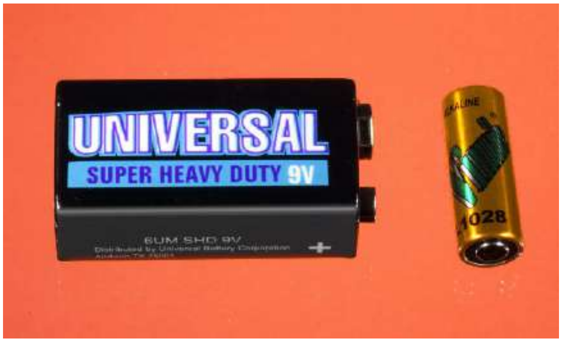

Figure 2-7. *At left, a cheap carbon-zinc battery; at right, a 12V alkaline burglar-alarm battery. See text for additional details.*

In an alkaline cell, the negative electrode is made of zinc powder, the positive electrode is manganese dioxide, and the electrolyte is potassium hydroxide. An alkaline cell may provide between three to five times the power capacity of an equal size of zinc-carbon cell and is less susceptible to voltage drop during the discharge cycle.

Extremely long shelf life is necessary in some military applications. This may be achieved by using a *reserve battery*, in which the internal chemical compounds are separated from each other but can be recombined prior to use.

Rechargeable Batteries

Commonly used types are *lead-acid*, *nickel cadmium* (abbreviated *NiCad* or *NiCd*), *nickel-metal hydride* (abbreviated *NiMH*), *lithium-ion* (abbreviated *Li-ion*), and *lithium-ion polymer*.

Lead-acid batteries have existed for more than a century and are still widely used in vehicles, burglar alarms, emergency lighting, and large power backup systems. The early design was described as *flooded*; it used a solution of sulfuric acid (generically referred to as *battery acid*) as its electrolyte, required the addition of distilled water periodically, and was vented to allow gas to escape. The venting also allowed acid to spill if the battery was tipped over.

The *valve-regulated lead-acid* battery (*VRLA*) has become widely used, requiring no addition of water to the cells. A pressure relief valve is included, but will not leak electrolyte, regardless of the position of the battery. VRLA batteries are preferred for *uninterruptible power supplies* for data-processing equipment, and are found in automobiles and in electric wheelchairs, as their low gas output and security from spillage increases their safety factor.

VRLA batteries can be divided into two types: absorbed glass mat (AGM) and gel batteries. The electrolyte in an AGM is absorbed in a fiber-glass mat separator. In a gel cell, the electrolyte is mixed with silica dust to form an immobilized gel.

The term *deep cycle battery* may be applied to a lead-acid battery and indicates that it should be more tolerant of discharge to a low level—perhaps 20 percent of its full charge (although manufacturers may claim a lower number). The plates in a standard lead-acid battery are composed of a lead *sponge*, which maximizes the surface area available to acid in the battery but can be physically abraded by deep discharge. In a deep cycle battery, the plates are solid. This means they are more robust, but are less able to supply high amperage. If a deep-discharge battery is used to start an internal combustion engine, the battery should be larger than a regular lead-acid battery used for this purpose.

A sealed lead-acid battery intended to power an external light activated by a motion detector is shown in Figure 2-8. This unit weighs several pounds and is trickle-charged during the daytime by a 6″ × 6″ solar panel.

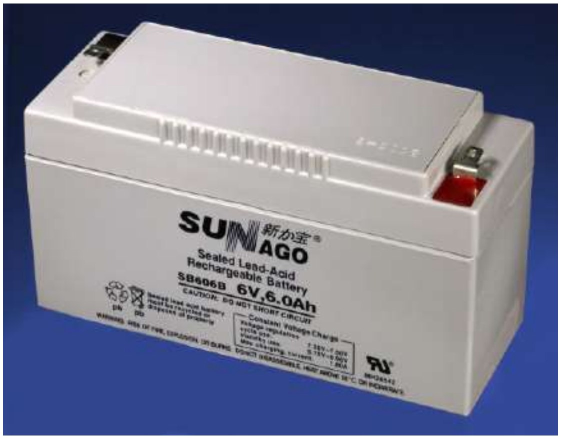

Figure 2-8. *A lead-acid battery from an external light activated by a motion sensor.*

Nickel-cadmium (*NiCad*) batteries can withstand extremely high currents, but have been banned in Europe because of the toxicity of metallic cadmium. They are being replaced in the United States by *nickel-metal hydride* (*NiMH*) types, which are free from the *memory effect* that can prevent a NiCad cell from fully recharging if it has been left for weeks or months in a partially discharged state.

Lithium-ion and lithium-ion polymer batteries have a better energy-to-mass ratio than NiMH batteries, and are widely used with electronic devices such as laptop computers, media players, digital cameras, and cellular phones. Large arrays of lithium batteries have also been used in some electric vehicles.

Various small rechargeable batteries are shown in Figure 2-9. The NiCad pack at top-left was manufactured for a cordless phone and is rapidly becoming obsolete. The 3V lithium battery at top-right was intended for a digital camera. The three batteries in the lower half of the photograph are all rechargeable NiMH substitutes for 9V, AA, and AAA batteries. The NiMH chemistry results in the AA and AAA single-cell batteries being rated for 1.2V rather than 1.5V, but the manufacturer claims they can be substituted for 1.5V alkaline cells because NiMH units sustain their rated voltage more consistently over time. Thus, the output from a fresh NiMH battery may be comparable to that of an alkaline battery that is part-way through its discharge cycle.

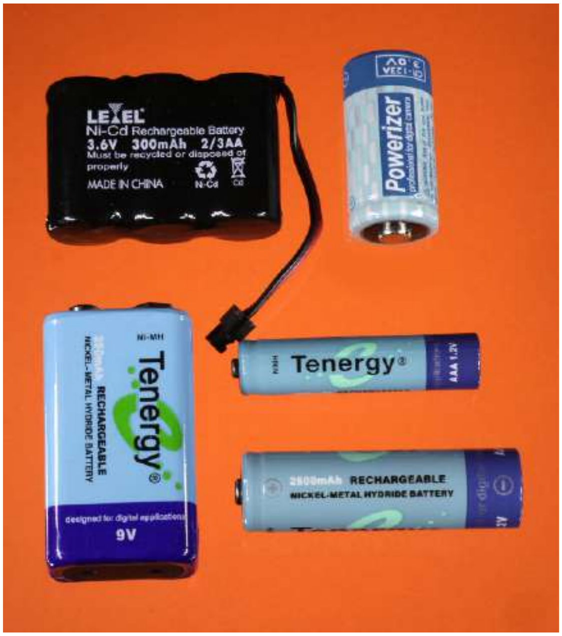

Figure 2-9. *Top left: NiCad battery pack for a cordless phone. Top right: Lithium battery for a digital camera. The other batteries are rechargeable NiMH substitutes for everyday alkaline cells.*

NiMH battery packs are available to deliver substantial power while being smaller and lighter than lead-acid equivalents. The NiMH package in Figure 2-10 is rated for 10Ah, and consists of ten

D-size NiMH batteries wired in series to deliver 12VDC. This type of battery pack is useful in robotics and other applications where a small motor-driven device must have free mobility.

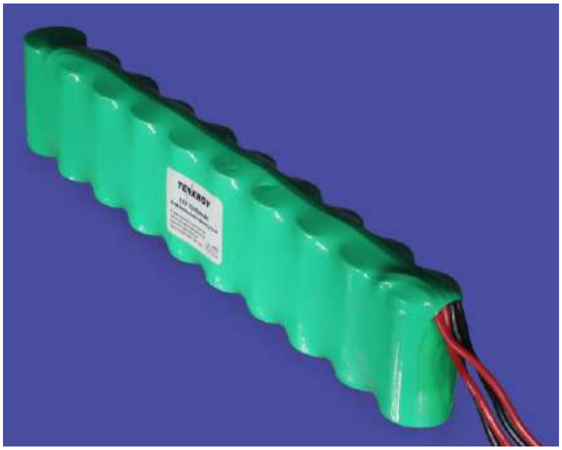

Figure 2-10. *This NiMH battery pack is rated at 10Ah and delivers 12 volts from ten D-size cells wired in series.*

Values

Amperage

The current delivered by a battery will be largely determined by the resistance of the external load placed between its terminals. However, because ion transfer must occur inside the battery to complete the circuit, the current will also be limited by the *internal resistance* of the battery. This should be thought of as an active part of the circuit.

Since a battery will deliver no current if there is no load, current must be measured while a load is attached, and cannot be measured by a meter alone. The meter will be immediately overloaded, with destructive results, if it is connected directly between the terminals of a battery, or in parallel with the load. Current must always be measured with the meter in series with the load, and the polarity of the meter must correspond with the polarity of the battery. See Figure 2-11.

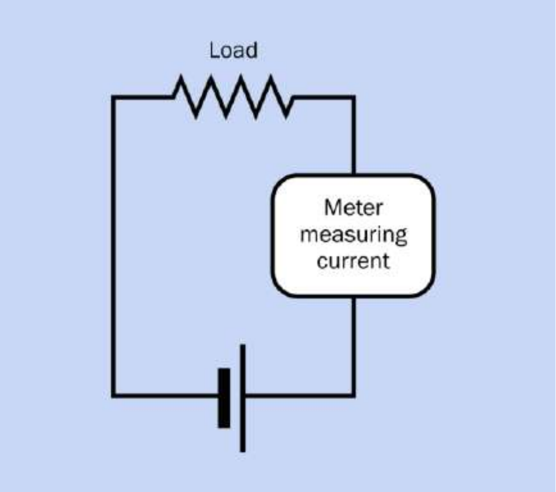

Figure 2-11. *When measuring current using an ammeter (or a multimeter configured to measure amps), the meter must be placed in series with the battery and a load. To avoid damaging the meter, it must never be applied directly across the terminals of the battery, or in parallel with a load. Be careful to observe the polarity of the meter.*

Capacity

The *electrical capacity* of a battery is measured in *amp-hours*, abbreviated *Ah*, *AH*, or (rarely) *A/H*. Smaller values are measured in *milliamp-hours*, usually abbreviated *mAh*. If I is the current being drawn from a battery (in amps) and T is the time for which the battery can deliver that current (in hours), the amp-hour capacity is given by the formula:

```
Ah = I * T
```

By turning the formula around, if we know the amp-hour rating that a manufacturer has determined for a battery, we can calculate the time in hours for which a battery can deliver a particular current:

```
T = Ah / I
```

Theoretically, Ah is a constant value for any given battery. Thus a battery rated for 4Ah should provide 1 amp for 4 hours, 4 amps for 1 hour, 5 amps for 0.8 hours (48 minutes), and so on.

In reality, this conveniently linear relationship does not exist. It quickly breaks down as the current rises, especially when using lead-acid batteries, which do not perform well when required to deliver high current. Some of the current is lost as heat, and the battery may be electrochemically incapable of keeping up with demand.

The *Peukert number* (named after its German originator in 1897) is a fudge factor to obtain a more realistic value for T at higher currents. If n is the Peukert number for a particular battery, then the previous formula can be modified thus:

$$T = Ah / I^{n}$$

Manufacturers usually (but not always) supply Peukert's number in their specification for a battery. So, if a battery has been rated at 4Ah, and its Peukert number is 1.2 (which is typical for lead-acid batteries), and I=5 (in other words, we want to know for how long a time, T, the battery can deliver 5 amps):

$$T = 4 / 5^{1.2} = \text{approximately } 4 / 6.9$$

This is about 0.58 hours, or 35 minutes—much less than the 48 minutes that the original formula suggested.

Unfortunately, there is a major problem with this calculation. In Peukert's era, the amp-hour rating for a battery was established by a manufacturer by drawing 1A and measuring the time during which the battery was capable of delivering that current. If it took 4 hours, the battery was rated at 4Ah.

Today, this measurement process is reversed. Instead of specifying the current to be drawn from the battery, a manufacturer specifies the time for which the test will run, then finds the maximum current the battery can deliver for that time. Often, the time period is 20 hours. Therefore, if a battery has a modern 4Ah rating, testing has probably determined that it delivered 0.2A for 20 hours, not 1A for 4 hours, which would have been the case in Peukert's era.

This is a significant distinction, because the same battery that can deliver 0.2A for 20 hours will not be able to satisfy the greater demand of 1A for 4 hours. Therefore the old amp-hour rating and the modern amp-hour rating mean different things and are incompatible. If the modern Ah rating is inserted into the old Peukert formula (as it was above), the answer will be misleadingly optimistic. Unfortunately, this fact is widely disregarded. Peukert's formula is still being used, and the performance of many batteries is being evaluated incorrectly.

The formula has been revised (initially by Chris Gibson of SmartGauge Electronics) to take into account the way in which Ah ratings are established today. Suppose that AhM is the modern rating for the battery's capacity in amp-hours, H is the duration in hours for which the battery was tested when the manufacturer calibrated it, n is Peukert's number (supplied by the manufacturer) as before, and I is the current you hope to draw from the battery. This is the revised formula to determine T:

$$T = H * (AhM / (I * H)^{n})$$

How do we know the value for H? Most (not all) manufacturers will supply this number in their battery specification. Alternatively, and confusingly, they may use the term *C-rate*, which can be defined as 1/H. This means you can easily get the value for H if you know the C-rate:

$$H = 1 / \text{C-rate}$$

We can now use the revised formula to rework the original calculation. Going back to the example, if the battery was rated for 4Ah using the modern system, in a discharge test that lasted 20 hours (which is the same as a C-rate of 0.05), and the manufacturer still states that it has a Peukert number of 1.2, and we want to know for how long we can draw 5A from it:

$$T = 20 * (4/(5 * 20)^{1.2}) = \text{approximately } 20 * 0.021$$

This is about 0.42 hours, or 25 minutes—quite different from the 35 minutes obtained with the old version of the formula, which should never be used when calculating the probable discharge time based on a modern Ah rating. These issues may seem arcane, but they are of great importance when assessing the likely performance of battery-powered equipment such as electric vehicles.

Figure 2-12 shows the probable actual performance of batteries with Peukert numbers of 1.1, 1.2, and 1.3. The curves were derived from the revised version of Peukert's formula and show how the number of amp-hours that you can expect diminishes for each battery as the current increases. For example, if a battery that the manufacturer has assigned a Peukert number of 1.2 is rated at 100Ah using the modern 20-hour test, but we draw 30A from it, the battery can actually deliver only 70Ah.

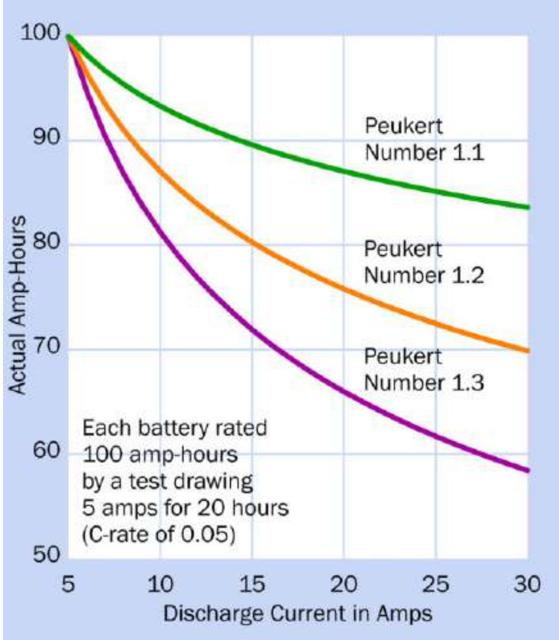

Figure 2-12. *Actual amp-hour performance that should be expected from three batteries of Peukert numbers 1.1, 1.2, and 1.3 when they discharge currents ranging from 5 to 30 amps, assuming that the manufacturer has rated each battery at 100Ah using the modern system, which usually entails a 20-hour test (a C-rate of 0.05).*

One additional factor: For any rechargeable battery, the Peukert number gradually increases with age, as the battery deteriorates chemically.

Voltage

The rated voltage of a fully charged battery is known as the *open circuit voltage* (abbreviated *OCV* or V_{oc}), defined as the potential that exists when no load is imposed between the terminals. Because the internal resistance of a volt meter (or a multimeter, when it is used to measure DC volts) is very high, it can be connected directly between the battery terminals with no other load present, and will show the OCV quite accurately, without risk of damage to the meter. A fully charged 12-volt car battery may have an OCV of about 12.6 volts, while a fresh 9-volt alkaline battery typically has an OCV of about 9.5 volts. Be extremely careful to set a multimeter to measure DC volts before connecting it across the battery. Usually this entails plugging the wire from the red probe into a socket separately reserved for measuring voltage, not amperage.

The voltage delivered by a battery will be pulled down significantly when a load is applied to it, and will decrease further as time passes during a discharge cycle. For these reasons, a **voltage regulator** is required when a battery powers components such as digital integrated circuit chips, which do not tolerate a wide variation in voltage.

To measure voltage while a load is applied to the battery, the meter must be connected in parallel with the load. See Figure 2-13. This type of measurement will give a reasonably accurate reading for the potential applied to the load, so long as the resistance of the load is relatively low compared with the internal resistance of the meter.

Figure 2-14 shows the performance of five commonly used sizes of alkaline batteries. The ratings in this chart were derived for alkaline batteries under favorable conditions, passing a small current through a relatively high-ohm load for long periods (40 to 400 hours, depending on battery type). The test continued until the final voltage for each 1.5V battery was 0.8V, and the final voltage for the 9V battery was a mere 4.8V. These voltages were considered acceptable when the

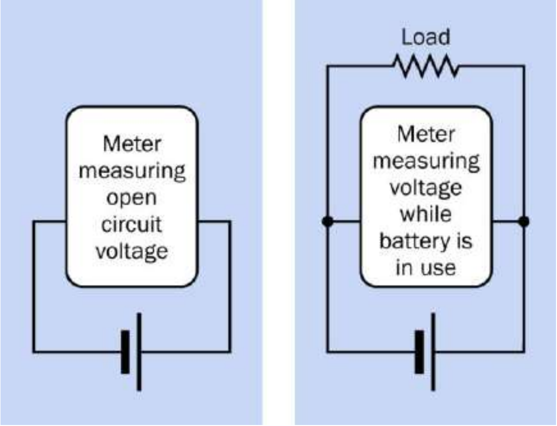

Figure 2-13. *When using a volt meter (or a multimeter configured to measure voltage), the meter can be applied directly between the battery terminals to determine the open-circuit voltage (OCV), or in parallel with a load to determine the voltage actually supplied during use. A multimeter must be set to measure DC volts before connecting it across a battery. Any other setting may damage the meter.*

Ah ratings for the batteries were calculated by the manufacturer, but in real-world situations, a final voltage of 4.8V from a 9V battery is likely to be unacceptable in many electronics applications.

Battery type	Rating (Ah)	Final voltage	Load (ohms)	Current (mA)
AAA	1.15	0.8	75	20
AA	2.87	0.8	75	20
C	7.8	0.8	39	40
D	17	0.8	39	40
9V	0.57	4.8	620	14

Figure 2-14. *The voltage delivered by a battery may drop to a low level while a manufacturer is establishing an amp-hour rating. Values for current, shown in the chart, were calculated subsequently as estimated averages, and should be considered approximate. (Derived from a chart published by Panasonic.)*

As a general rule of thumb, if an application does not tolerate a significant voltage drop, the manufacturer's amp-hour rating for a small battery may be divided by 2 to obtain a realistic number.

How to Use it

When choosing a battery to power a circuit, considerations will include the intended shelf life, maximum and typical current drain, and battery weight. The amp-hour rating of a battery can be used as a very approximate guide to determine its suitability. For 5V circuits that impose a drain of 100mA or less, it is common to use a 9V battery, or six 1.5V batteries in series, passing current through a **voltage regulator** such as the LM7805. Note that the voltage regulator requires energy to function, and thus it imposes a voltage drop that will be dissipated as heat. The minimum drop will vary depending on the type of regulator used.

Batteries or cells may be used in series or in parallel. In series, the total voltage of the chain of cells is found by summing their individual voltages, while their amp-hour rating remains the same as for a single cell, assuming that all the cells are identical. Wired in parallel, the total voltage of the cells remains the same as for a single cell, while the combined amp-hour value is found by summing their individual amp-hour ratings, assuming that all the batteries are identical. See Figure 2-15.

In addition to their obvious advantage of portability, batteries have an additional advantage of being generally free from power spikes and noise that can cause sensitive components to misbehave. Consequently, the need for smoothing will depend only on possible noise created by other components in the circuit.

Motors or other inductive loads draw an initial surge that can be many times the current that they use after they start running. A battery must be chosen that will tolerate this surge without damage.

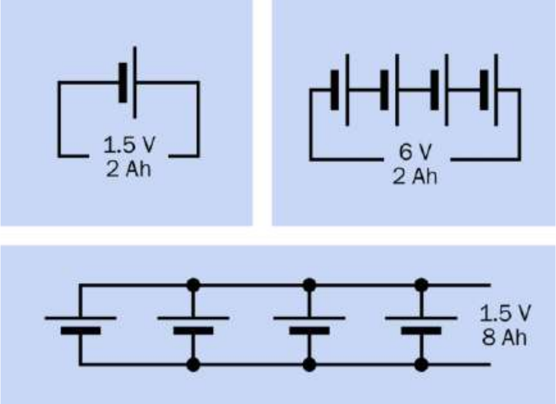

Figure 2-15. *Theoretical results of using 1.5V cells in series or in parallel, assuming a 2Ah rating for one cell.*

Because of the risk of fire, United States airline regulations limit the amp-hour capacity of lithium-ion batteries in any electronic device in carry-on or checked passenger baggage. If a device may be carried frequently as passenger baggage (for example, emergency medical equipment), NiMH batteries are preferred.

What Can Go Wrong

Short Circuits: Overheating and Fire

A battery capable of delivering significant current can overheat, catch fire, or even explode if it is short-circuited. Dropping a wrench across the terminals of a car battery will result in a bright flash, a loud noise, and some molten metal. Even a 1.5-volt alkaline AA battery can become too hot to touch if its terminals are shorted together. (Never try this with a rechargeable battery, which has a much lower internal resistance, allowing much higher flow of current.) Lithium-ion batteries are particularly dangerous, and almost always are packaged with a current-limiting component that should not be disabled. A short-circuited lithium battery can explode.

If a battery pack is used as a cheap and simple workbench DC power supply, a **fuse** or circuit breaker should be included. Any device that uses significant battery power should be fused.

Diminished Performance Caused by Improper Recharging

Many types of batteries require a precisely measured charging voltage and a cycle that ends automatically when the battery is fully charged. Failure to observe this protocol can result in chemical damage that may not be reversible. A charger should be used that is specifically intended for the type of battery. A detailed comparison of chargers and batteries is outside the scope of this encyclopedia.

Complete Discharge of Lead-Acid Battery

Complete or near-complete discharge of a lead-acid battery will significantly shorten its life (unless it is specifically designed for deep-cycle use —although even then, more than an 80% discharge is not generally recommended).

Inadequate Current

Chemical reactions inside a battery occur more slowly at low temperatures. Consequently, a cold battery cannot deliver as much current as a warm battery. For this reason, in winter weather, a car battery is less able to deliver high current. At the same time, because engine oil becomes more viscous as the temperature falls, the starter motor will demand more current to turn the engine. This combination of factors explains the tendency of car batteries to fail on cold winter mornings.

Incorrect Polarity

If a battery charger or generator is connected with a battery with incorrect polarity, the battery may experience permanent damage. The **fuse** or *circuit breaker* in a charger may prevent this from occurring and may also prevent damage to the charger, but this cannot be guaranteed.

If two high-capacity batteries are connected with opposite polarity (as may happen when a clumsy attempt is made to start a stalled car with jumper cables), the results may be explosive. Never lean over a car battery when attaching cables to it, and ideally, wear eye protection.

Reverse Charging

Reverse charging can occur when a battery becomes completely discharged while it is wired (correctly) in series with other batteries that are still delivering current. In the upper section of the schematic at Figure 2-16 two healthy 6V batteries, in series, are powering a resistive load. The battery on the left applies a potential of 6 volts to the battery on the right, which adds its own 6 volts to create a full 12 volts across the load. The red and blue lines indicate volt meter leads, and the numbers show the reading that should be observed on the meter.

In the second schematic, the battery on the left has become exhausted and is now a "dead weight" in the circuit, indicated by its gray color. The battery on the right still sustains a 6-volt potential. If the internal resistance of the dead battery is approximately 1 ohm and the resistance of the load is approximately 20 ohms, the potential across the dead battery will be about 0.3 volts, in the opposite direction to its normal charged voltage. Reverse charging will result and can damage the battery. To avoid this problem, a battery pack containing multiple cells should never be fully discharged.

Sulfurization

When a lead-acid battery is partially or completely discharged and is allowed to remain in that state, sulfur tends to build up on its metal plates. The sulfur gradually tends to harden, forming a barrier against the electrochemical reactions that are necessary to recharge the battery. For this reason, lead-acid batteries should not be allowed to sit for long periods in a discharged condition. Anecdotal evidence suggests that even a very small trickle-charging current can prevent sulfurization, which is why some people recommend attaching a small solar panel to a battery that is seldom used—for example, on a sail boat, where the sole function of the battery is to start an auxiliary engine when there is insufficient wind.

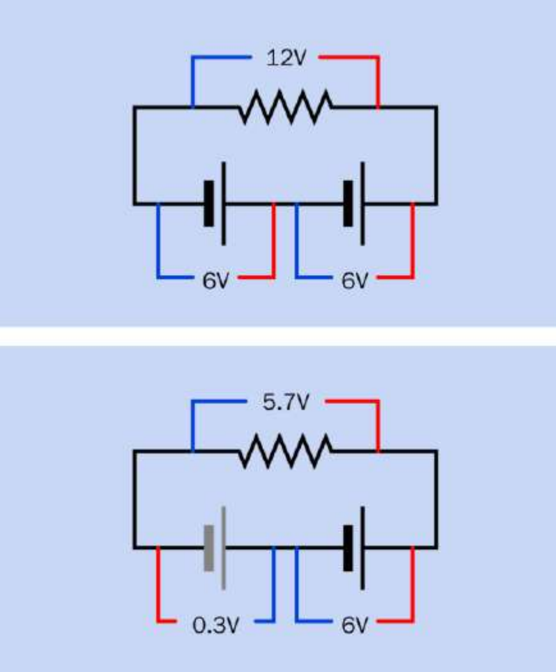

Figure 2-16. *When a pair of 6V batteries is placed in series to power a resistive load, if one of the batteries discharges completely, it becomes a load instead of a power source, and will be subjected to reverse charging, which may cause permanent damage.*

High Current Flow Between Parallel Batteries

If two batteries are connected in parallel, with correct polarity, but one of them is fully charged while the other is not, the charged battery will attempt to recharge its neighbor. Because the batteries are wired directly together, the current will be limited only by their internal resistance and the resistance of the cables connecting them. This may lead to overheating and possible damage. The risk becomes more significant when linking batteries that have high Ah ratings. Ideally they should be protected from one another by high-current **fuses.**

jumper

3

A jumper may also be referred to as a *jumper socket* or a *shunt*. A jumper should not be confused with *jumper wires*, which are not considered components for the purposes of this encyclopedia.

OTHER RELATED COMPONENTS

- **switch** (See Chapter 6)

What It Does

A jumper is a low-cost substitute for a **switch,** where a connection has to be made (or unmade) only a few times during the lifetime of a product. Typically it allows a function or feature on a circuit board to be set on a semipermanent basis, often at the time of manufacture. A *DIP switch* performs the same function more conveniently. See "DIP" (page 43).

There is no standardized schematic symbol to represent a jumper.

How It Works

A jumper is a very small rectangular plastic tab containing two (or sometimes more) metal sockets usually spaced either 0.1" or 2mm apart. The sockets are connected electrically inside the tab, so that when they are pushed over two (or more) pins that have been installed on a circuit board for this purpose, the jumper shorts the pins together. The pins are usually 0.025" square and are often part of a *header* that is soldered into the board. In a parts catalogue, jumpers may be found in a section titled "Headers and Wire Housings" or similar.

Three jumpers are shown in Figure 3-1. The blue one contains two sockets spaced 0.1" and is deep enough to enclose the pins completely. The red one contains two sockets spaced 2mm and may allow the tips of the pins to emerge from its opposite end. The black one contains four sockets, each pair spaced 0.1" apart.

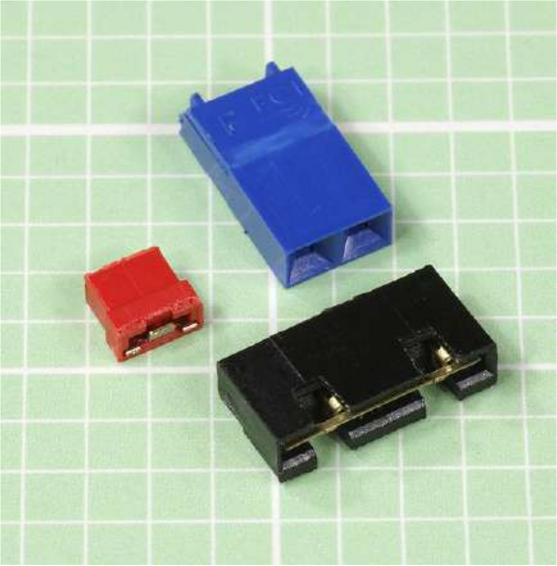

Figure 3-1. *Three jumpers containing two sockets spaced 2mm (left), two sockets spaced 0.1" (top right), and four sockets, each pair spaced 0.1" (bottom right).*

The set of pins with which a jumper is used is often referred to as a *header*. Headers are available with pins in single or dual rows. Some headers are designed to be snapped off to provide the desired number of pins. A dual 28-pin header is shown in Figure 3-2 with a black jumper pushed onto a pair of pins near the midpoint.

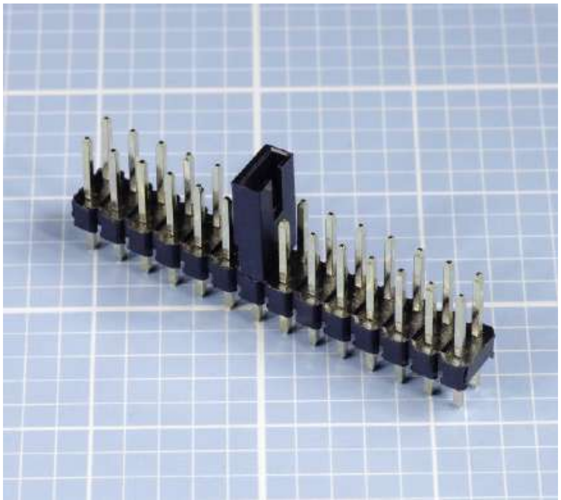

Figure 3-2. *A jumper pushed onto a pair of pins midway along a dual 28-pin header.*

Variants

A *jumper assembly* may be a kit containing not only the jumper but also the array of pins with which it is intended to be used. Check the manufacturer's datasheet to find out exactly what is included.

The most common types of jumpers have two sockets only, but variants are available with as many as 12 sockets, which may be arranged in one or two rows. *Header sockets* may be used as a substitute for purpose-made jumpers, with the advantage that they are often sold in long strips that can be snapped off to provide as many sockets as needed. However, the pins attached to header sockets must be manually connected by soldering small lengths of wire between them.

In some jumpers, the plastic tab extends upward for about half an inch and functions as a finger grip, making the jumper much easier to hold during insertion and removal. This is a desirable feature if there is room to accommodate it.

The sockets inside a jumper are often made from phosphor-bronze, copper-nickel alloy, tin alloy, or brass alloy. They are usually gold-plated, but in some instances are tin-plated.

Rarely, a jumper may consist of a metal strip with U-shaped connections suitable for being used in conjunction with screw terminals. Two jumpers of this type are shown in Figure 3-3. They should not be confused with high-amperage fuses that look superficially similar.

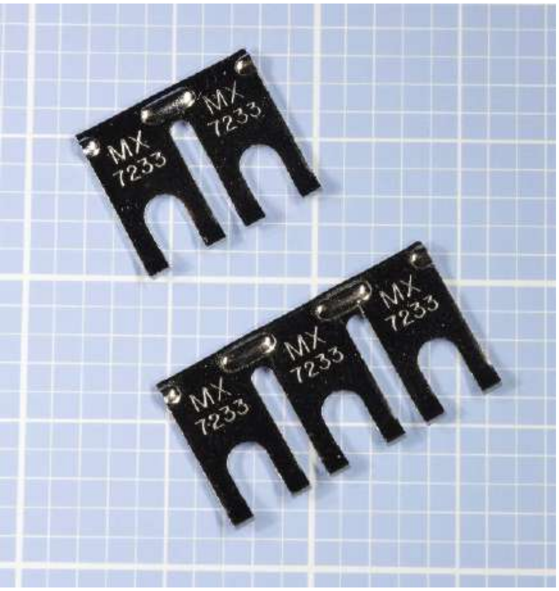

Figure 3-3. *These jumpers are designed to short together two or three screw terminals.*

Values

The spacing between the sockets in a jumper is referred to as its *pitch*. As previously noted, 0.1" and 2mm are the most popular values.

A typical maximum rating for a jumper of 0.1" pitch is 2A or 2.5A at 250V.

How to Use it

A jumper may activate a "set it and forget it" circuit function. An example would be the factory configuration of a product to work with 115VAC or 230VAC power input. End users were expected to set jumpers in some computer equipment sold during the 1980s, but this is no longer the case.

What Can Go Wrong

Jumpers are easily dropped, easily lost, and easily placed incorrectly. When purchasing jumpers, buy extras to compensate for their fragility and the ease of losing them.

Any location where a jumper may be used should be clearly labelled to define the function of each setting.

Cheap, poorly made jumpers may self-destruct from mechanical stresses when removed from their pins. The plastic casing can come away, leaving the sockets clinging naked to the pins protruding from the circuit board. This is another reason why it is a good idea to have a small stock of spare jumpers for emergencies.

Oxidation in jumpers where the contacts are not gold- or silver-plated can create electrical resistance or unreliable connections.

fuse

4

The alternate spelling *fuze* is seldom used.

OTHER RELATED COMPONENTS

- None

What It Does

A fuse protects an electrical circuit or device from excessive current when a metal element inside it melts to create an open circuit. With the exception of *resettable fuses* (discussed separately in "Resettable Fuses" (page 24)), a fuse must be discarded and replaced after it has fulfilled its function.

When high current melts a fuse, it is said to *blow* or *trip* the fuse. (In the case of a resettable fuse, only the word *trip* is used.)

A fuse can work with either AC or DC voltage, and can be designed for almost any current. In residential and commercial buildings, *circuit breakers* have become common, but a large cartridge fuse may still be used to protect the whole system from short-circuits or from overcurrent caused by lightning strikes on exposed power lines.

In electronic devices, the **power supply** is almost always fused.

Schematic symbols for a fuse are shown in Figure 4-1. Those at the right and second from right are most frequently used. The one in the center is approved by ANSI, IEC, and IEEE but is seldom seen. To the left of that is the fuse symbol understood by electrical contractors in architectural plans. The symbol at far left used to be common but has fallen into disuse.

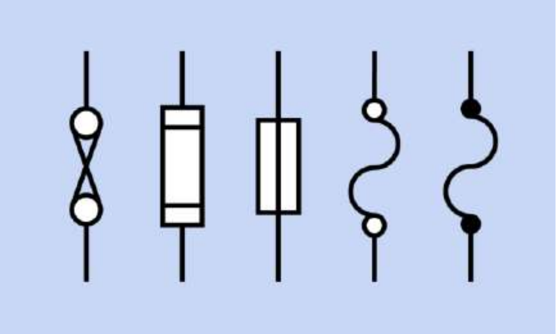

Figure 4-1. *Alternate schematic symbols for a fuse. See text for explanation.*

How It Works

The *element* in a fuse is usually a wire or thin metal strip mounted between two terminals. In a cartridge fuse, it is enclosed in a glass or ceramic cylinder with a contact at each end, or in a small metallic can. (Old-style, large, high-amperage fuses may be packaged in a paper or cardboard tube.) The traditional glass cartridge allows visual inspection to confirm that the fuse has blown.

A fuse responds only to current, not to voltage. When choosing a fuse that will be reliable in conditions of steady current consumption, a safe rule is to figure the maximum amperage when all components are functioning and add 50%. However, if current surges or spikes are likely, their duration will be relevant. If I is the current surge in amps and t is its duration in seconds, the surge sensitivity of a fuse—which is often referred to verbally or in printed format as I2t—is given by the formula:

```
I2t = I² * t
```

Some semiconductors also have an I2t rating, and should be protected with a similarly rated fuse.

Any fuse will present some resistance to the current flowing through it. Otherwise, the current would not generate the heat that blows the fuse. Manufacturer datasheets list the voltage drop that the internal resistance of a fuse is likely to introduce into a circuit.

Values

The *current rating* or *rated current* of a fuse is usually printed or stamped on its casing, and is the maximum flow that it should withstand on a continuous basis, at the ambient temperature specified by the manufacturer (usually 25 degrees Centigrade). The ambient temperature refers to the immediate environment of the fuse, not the larger area in which it may be located. Note that in an enclosure containing other components, the temperature is usually significantly higher than outside the enclosure.

Ideally a fuse should function reliably and indefinitely at its rated maximum amperage, but should blow just as reliably if the current rises by approximately 20% beyond the maximum. In reality, manufacturers recommend that continuous loading of a fuse should not exceed 75% of its rating at 25 degrees Centigrade.

The *voltage rating* or *rated voltage* of a fuse is the maximum voltage at which its element can be counted on to melt in a safe and predictable manner when it is overloaded by excess current. This is sometimes known as the *breaking capacity*. Above that rating, the remaining pieces of the fuse element may form an arc that sustains some electrical conduction.

A fuse can always be used at a lower voltage than its rating. If it has a breaking capacity of 250V, it will still provide the same protection if it is used at 5V.

Four differently rated glass cartridge fuses are shown in Figure 4-2. The one at the top is a slow-blowing type, rated at 15A. Its element is designed to absorb heat before melting. Below it is a 0.5A fuse with a correspondingly thinner element. The two smaller fuses are rated at 5A each. The center two fuses have a maximum voltage rating of 250V, while the one at the top is rated at 32V and the one at the bottom is rated at 350V. Clearly, the size of a fuse should never be used as a guide to its ratings.

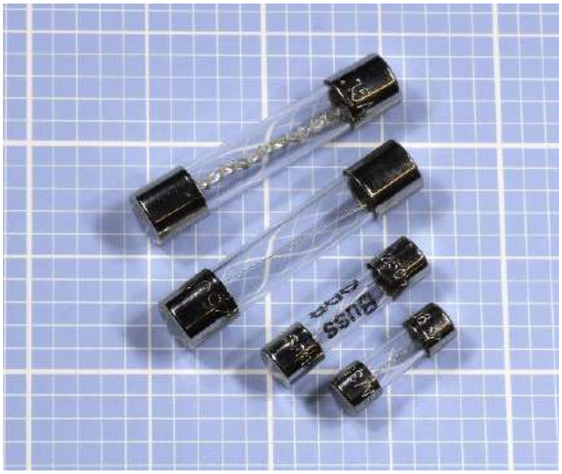

Figure 4-2. *Four glass cartridge fuses. See text for details.*

Variants

Early power fuses in residential buildings consisted of bare nichrome wire wrapped around a porcelain holder. In the 1890s, Edison developed *plug fuses* in which the fuse was contained in a porcelain module with a screw thread, compatible with the base of an incandescent bulb. This design persisted in some U.S. urban areas for more than 70 years, is still found in old buildings, and is still being manufactured.

Small Cartridge Fuses

Small cartridge fuses for appliances and electronics equipment—such as those shown in Figure 4-2--are available in sizes tabulated in Figure 4-3. With the exception of the 4.5mm diameter fuse (a European addition), these sizes were originally measured in inches; today, they are often described only with the equivalent metric measurement. Any cartridge fuse is usually available with the option of a lead attached to it at each end, so that it can be used as a through-hole component.

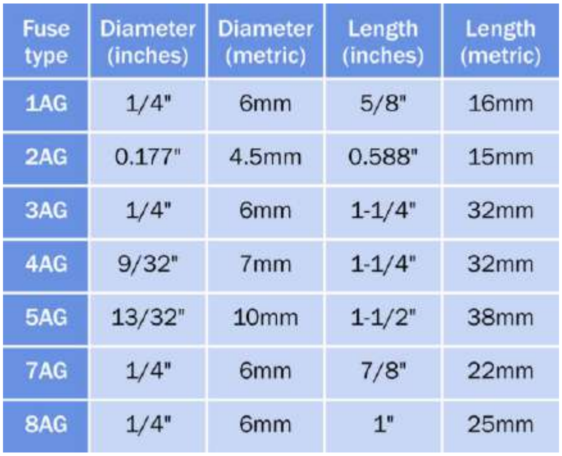

Fuse type	Diameter (inches)	Diameter (metric)	Length (inches)	Length (metric)
1AG	1/4"	6mm	5/8"	16mm
2AG	0.177"	4.5mm	0.588"	15mm
3AG	1/4"	6mm	1-1/4"	32mm
4AG	9/32"	7mm	1-1/4"	32mm
5AG	13/32"	10mm	1-1/2"	38mm
7AG	1/4"	6mm	7/8"	22mm
8AG	1/4"	6mm	1"	25mm

Figure 4-3. *The approximate physical sizes of commonly used small glass or ceramic cartridge fuses are shown here with the codes that are often used to identify them.*

Fuses may be fast acting, medium acting, or slow-blowing, the last of which may alternatively be referred to as *delay fuses*. Extra-fast–acting fuses are available from some manufacturers. The term *Slo-Blo* is often used but is actually a trademark of Littelfuse. None of the terms describing the speed of action of a fuse has been standardized with a specific time or time range.

Some cartridge fuses are available in a ceramic format as an alternative to the more common glass cylinder. If accidental application of extremely high current is possible (for example, in a multimeter that can be set to measure amps, and may be accidentally connected across a powerful battery), a ceramic cartridge is preferable because it contains a filler that will help to stop an arc from forming. Also, if a fuse is physically destroyed by application of very high current, ceramic fragments may be preferable to glass fragments.

Automotive Fuses

Automotive fuses are identifiable by their use of blades designed for insertion in flat sockets where the fuse is unlikely to loosen as a result of vibration or temperature changes. The fuses come in various sizes, and are uniformly color-coded for easy identification.

A selection of automotive fuses is shown in Figure 4-4. The type at the top is typically described as a “maxi-fuse” while the type at bottom-left is a “mini-fuse.” Here again, size is irrelevant to function, as all three of those pictured are rated 30A at 32V.

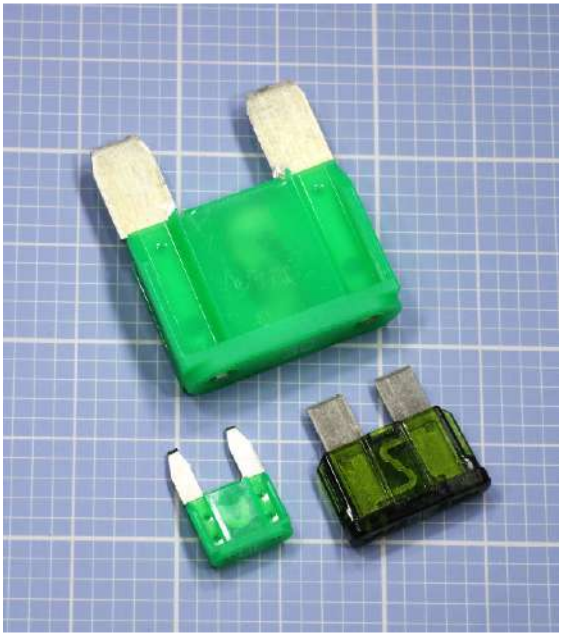

Figure 4-4. *Three automotive fuses. All have the same rating: 30A at 32V.*

In Figure 4-5, the largest of the fuses from Figure 4-4 has been cut open to reveal its element.

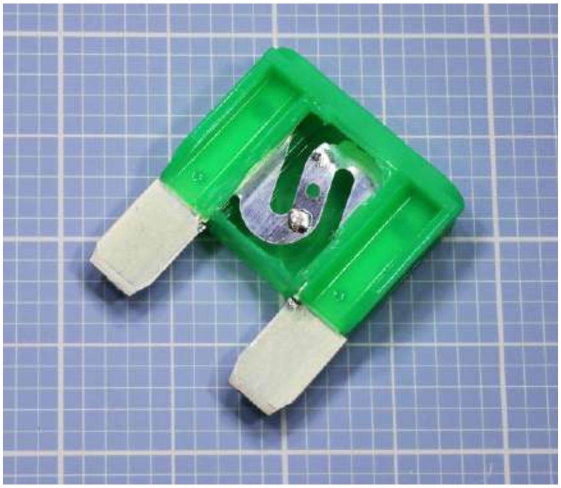

Figure 4-5. *The largest fuse from the previous figure, cut open to reveal its element.*

Usually automotive fuses are mounted together in a block, but if aftermarket accessory equipment is added, it may be protected by an inline fuse in a holder that terminates in two wires. This is shown with two sample fuses in Figure 4-6. Similar inline fuse holders are manufactured for other types of fuses.

Strip Fuses

High-amperage fuses for vehicles may be sold in "strip fuse" format, also known as a *fusible link*, designed to be clamped between two screw-down terminals. Since some **jumpers** may look very similar, it is important to keep them separate. A strip fuse is shown in Figure 4-7.

Through-Hole Fuses

Small fuses with radial leads, which seem appropriate for through-hole insertion in printed circuit boards, are actually often used in conjunction with appropriate sockets, so that they can be easily replaced. They are described in catalogues as "subminiature fuses" and are typically found in laptop computers and their power supplies, also televisions, battery chargers, and air conditioners. Three examples are shown in Figure 4-8. All have slow-blowing characteristics.

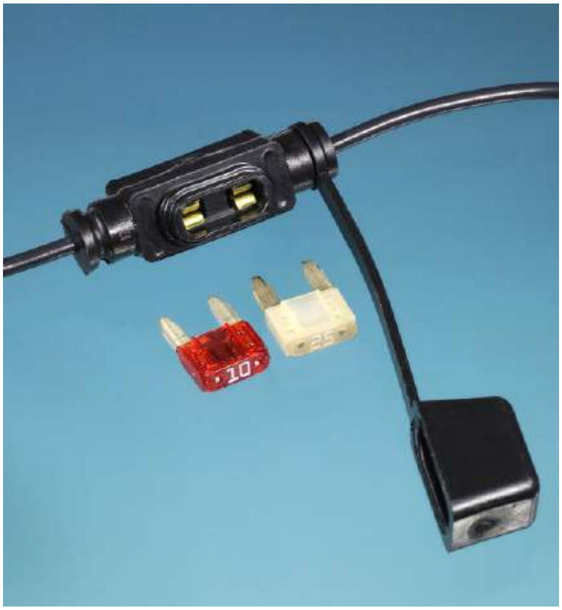

Figure 4-6. *Two blade-type fuses, commonly used for automotive applications, shown with an inline fuse holder. The plastic cap, at right, is closed over the holder when a fuse has been installed.*

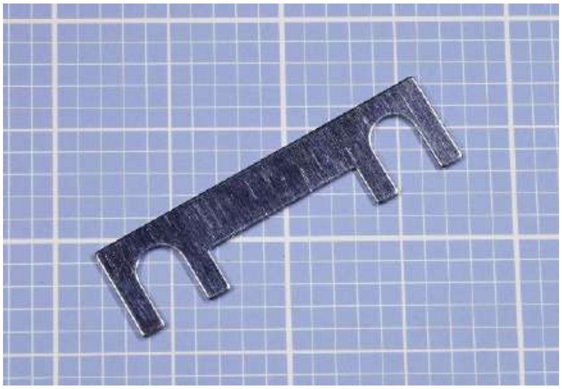

Figure 4-7. *This strip fuse is intended for use in diesel vehicles. The example shown is rated 100A at 36V.*

Resettable Fuses

Properly known as a *polymeric positive temperature coefficient* fuse (often abbreviated *PTC* or *PPTC*), a *resettable fuse* is a solid-state, encapsulated component that greatly increases its resistance in response to a current overload, but gradually returns to its original condition when the flow of current is discontinued. It can be thought

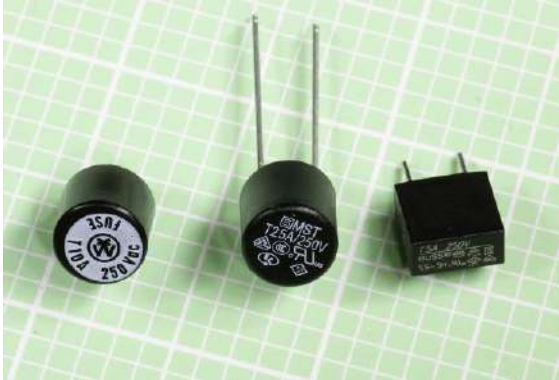

Figure 4-8. *Three subminiature fuses terminating in wire leads. From left to right: 10A at 250V, 2.5A at 250V, and 5A at 250V.*

of as a **thermistor** that has a nonlinear response. Three through-hole examples are shown in Figure 4-9. While different sizes of cartridge fuse may share the same ratings, differently rated resettable fuses may be identical in size. The one on the left is rated 40A at 30V, while the one on the right is rated 2.5A at 30V. (Note that the codes printed on the fuses are not the same as their manufacturer part numbers.) The fuse at the top is rated 1A at 135V.

When more than the maximum current passes through the fuse, its internal resistance increases suddenly from a few ohms to hundreds of thousands of ohms. This is known as *tripping* the fuse. This inevitably entails a small delay, but is comparable to the time taken for a slow-blowing fuse to respond.

A resettable fuse contains a polymer whose crystalline structure is loaded with graphite particles that conduct electricity. As current flowing through the fuse induces heat, the polymer transitions to an amorphous state, separating the graphite particles and interrupting the conductive pathways. A small current still passes through the component, sufficient to maintain its amorphous state until power is disconnected.

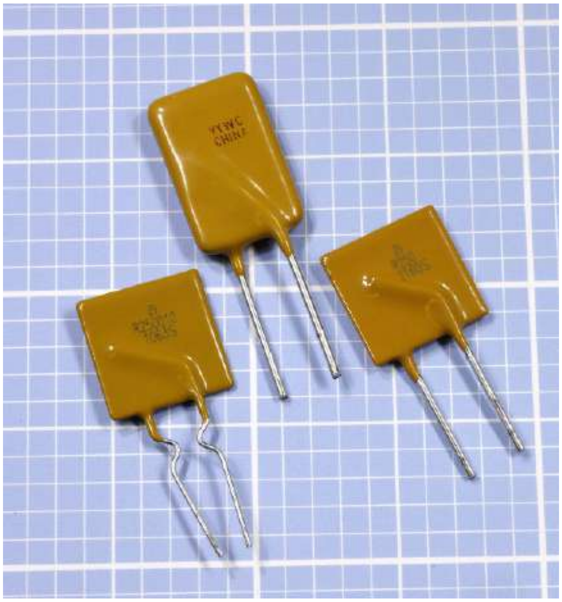

Figure 4-9. *Some through-hole resettable fuses. See text for details.*

After the resettable fuse cools, it gradually recrystallizes, although its resistance does not fall back completely to its original value for more than an hour.

The maximum safe level of current for a resettable fuse is known as the *hold current*, while the current that triggers its response is termed the *trip current*. Resettable fuses are available with trip-current ratings from 20mA to 100A. While conventional appliance and electronics fuses may be rated as high as 600V, resettable fuses are seldom rated above 100V.

Typical cartridge fuses are affected only to a minor extent by temperature, but the current rating of a resettable fuse may diminish to 75% of its normal value at 50 degrees Centigrade and may drop to 50% of its normal value at 80 degrees Centigrade. In other words, a fuse that is rated for 4A at 25 degrees may tolerate a maximum of only 3A when it operates at twice that temperature. See Figure 4-10.

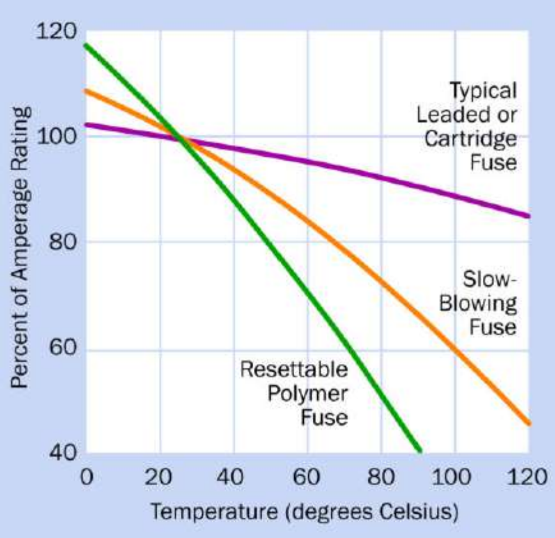

Figure 4-10. *The three curves give a very approximate idea of the temperature sensitivity of three types of fuses commonly used to protect electronic equipment. The left-hand scale provides an approximate value for the amperage which will trip the fuse.*

Conventional slow-blowing fuses are temperature-sensitive, but to a lesser degree than resettable fuses.

Resettable fuses are used in computer power supplies, USB power sources, and loudspeaker enclosures, where they protect the speaker coils from being overdriven. They are appropriate in situations where a fuse may be tripped relatively often, or where technically unsophisticated users may feel unable to replace a fuse or reset a circuit breaker.

Brand names for resettable fuses include PolySwitch, OptiReset, Everfuse, Polyfuse, and Multifuse. They are available in *surface-mount* (SMT) packages or as *through-hole* components, but not in cartridge format.

Surface Mount Fuses

Because surface-mount fuses are difficult or impossible to replace after they have been soldered onto the board, they are often resettable.

A surface-mount resettable fuse approximately 0.3" square is shown in Figure 4-11. It is rated for 230V and has an internal resistance of 50 ohms. Its hold current is 0.09A and its trip current is 0.19A.

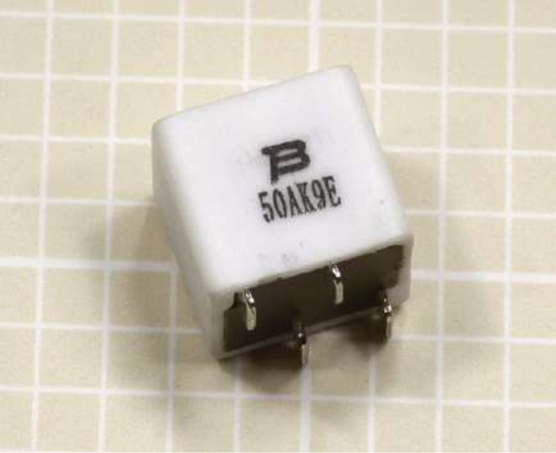

Figure 4-11. *A surface-mount resettable fuse. See text for details.*

How to Use it

Any equipment that may be plugged into a wall outlet should be fused, not only to protect its components but also to protect users who may open the box and start investigating with a screwdriver.

Equipment that contains powerful motors, pumps, or other inductive loads should be protected with slow-blowing fuses, as the initial surge of current when the equipment is switched on is likely to rise well above the rating of the fuse. A slow-blowing fuse will tolerate a surge for a couple of seconds. Other fuses will not.

Conversely, fast-acting fuses should be used with electronic equipment, especially integrated circuits that are quickly and easily damaged.

Any device using substantial **battery** power should be fused because of the unpredictable and generally bad behavior of batteries when they are short-circuited. Parallel connections between multiple large batteries should be fused to avoid the possibility that a highly charged battery may attempt to recharge its neighbor(s). Large "J size" fuses rated from 125A to 450A have become common in the solar power community, where banks of lead-acid batteries are often used. These fuses have a thick brass tab at each end, drilled so that they can be bolted into place. Alternatively, they will push-fit into an appropriate *fuseholder*.

For cartridge fuses up to 1/4" in diameter that don't have leads attached, appropriately sized fuseholders are available in several formats:

Panel mounted fuse enclosure is probably the most common, consisting of a plastic tube with a spring-contact at the bottom, and a plastic cap with a second contact inside. The cap either screws onto the tube of the fuse, or is pushed down and turned to hold it in place. A nut is provided to secure the fuseholder after it has been inserted into a hole drilled in the panel. The fuse is dropped into the tube, and the cap is applied. This type of holder is available in full-length or shorter, "low profile" formats. A low-profile holder is shown in Figure 4-12. It is shown assembled at right, with its component parts disassembled alongside.

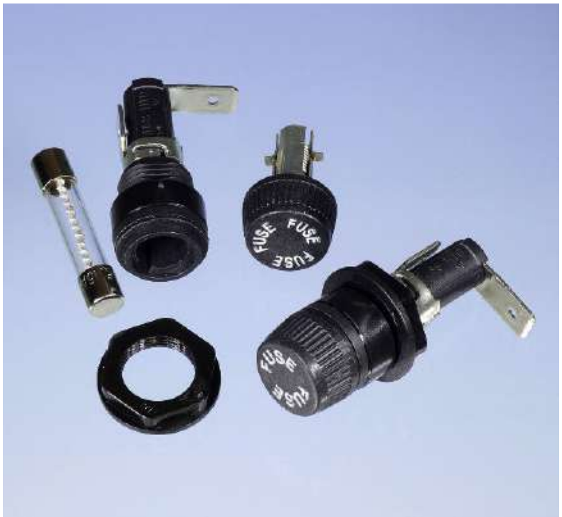

Figure 4-12. *A low-profile panel-mounted fuse holder shown disassembled (left) and assembled (right).*

Circuit board mounted fuse enclosure is basically the same as the panel-mounted version, but with through-hole solder pins attached.

Fuse block is a small plastic block with two clips on its upper surface for insertion of a cartridge fuse.

Fuse clips can be bought individually, with solder pins for through-hole mounting.

Inline fuse holder is designed to be inserted in a length of wire. Usually made of plastic, it will either terminate it, wires or will have metal contacts to crimp or solder at each end. See Figure 4-6.

Through-hole fuse holders are available for subminiature fuses.

What Can Go Wrong

Repeated Failure

When a fuse in a circuit blows frequently, this is known as *nuisance opening*. Often it can result from failure to take into account all the aspects of the circuit, such as a large filtering capacitor in a power supply that draws a major surge of current when the power supply is switched on. The formally correct procedure to address this problem is to measure the power surge, properly known as *peak inrush current*, with an oscilloscope, calculate the $I^2 * t$ of the wave form, and select a fuse with a rating at least 5 times that value.

A fuse should never be replaced with an equivalent length of wire or any other conductor.

Soldering Damage

When a through-hole or surface-mount fuse is soldered into place, heat from the soldering process can cause the soft metal element inside the fuse to melt partially and reflow. This is likely to change the rating of the fuse. Generally, fuses should be treated with the same caution as semiconductors when they are fixed in place with solder.

Placement

A fuse should be placed close to the power source or power input point in a circuit, so that it protects as much of the circuit as possible.

pushbutton

5

Often referred to as a *pushbutton switch* and sometimes as a *momentary switch*. In this encyclopedia, a pushbutton is considered separately from a **switch**, which generally uses a lever-shaped actuator rather than a button, and has at least one *pole* contact where a pushbutton generally has contacts that are not distinguishable from each other.

OTHER RELATED COMPONENTS

- **switch** (See Chapter 6)
- **rotary switch** (See Chapter 7)

What It Does

A pushbutton contains at least two contacts, which close or open when the button is pressed. Usually a spring restores the button to its original position when external pressure is released. Figure 5-1 shows schematic symbols for pushbuttons. The symbols that share each blue rectangle are functionally identical. At top is a normally-open single-throw pushbutton. At center is a normally-closed single-throw pushbutton. At bottom is a double-throw pushbutton.

Unlike a **switch**, a basic pushbutton does not have a primary contact that can be identified as the *pole*. However, a single pushbutton may close or open two separate pairs of contacts, in which case it can be referred to, a little misleadingly, as a double-pole pushbutton. See Figure 5-2. Different symbols are used for slider pushbuttons with multiple contact pairs; see "Slider" (page 31).

A generic full-size, two-contact pushbutton is shown in Figure 5-3.

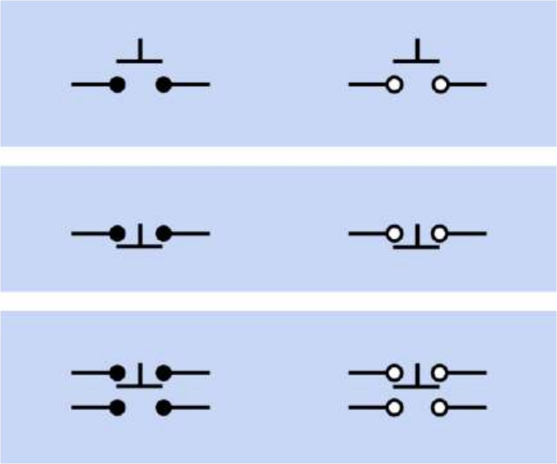

Figure 5-1. *Commonly used schematic symbols to represent a simple pushbutton. See text for details.*

How It Works

Figure 5-4 shows a cross-section of a pushbutton that has a single steel return spring, to create resistance to downward force on the button, and a pair of springs above a pair of contacts, to hold each contact in place and make a firm connection when the button is pressed. The two upper contacts are electrically linked, although this feature is not shown.

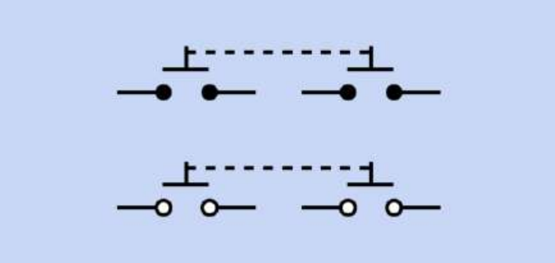

Figure 5-2. *Commonly used schematic symbols to represent a double-pole pushbutton.*

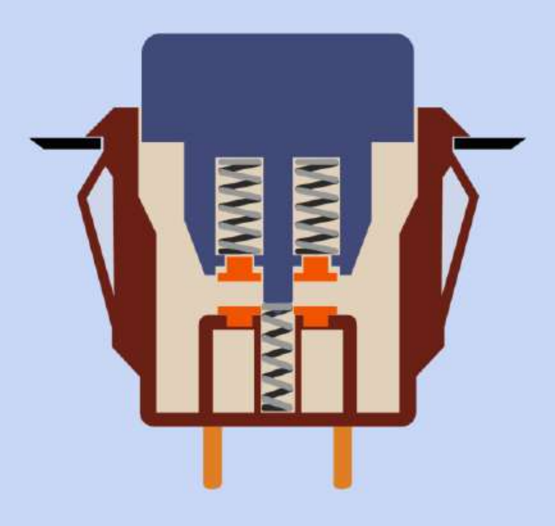

Figure 5-4. *Cross-section of a pushbutton showing two spring-loaded contacts and a single return spring.*

Variants

Poles and Throws

Abbreviations that identify the number of poles and contacts inside a pushbutton are the same as the abbreviations that identify those attributes in a **switch**. A few examples will make this clear:

SPST, also known as 1P1T
Single pole, single throw

DPST also known as 2P1T
Double pole, single throw

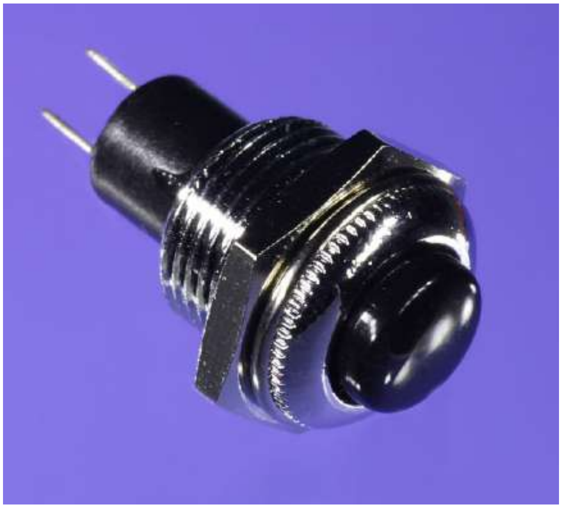

Figure 5-3. *The simplest, traditional form of pushbutton, in which pressing the button creates a connection between two contacts.*

SPDT also known as 1P2T
Single pole, double throw

3PST also known as 3P1T
Three pole, single throw

While a switch may have an additional center position, pushbuttons generally do not.

On-Off Behavior

Parentheses are used to indicate the momentary state of the pushbutton while it is pressed. It will return to the other state by default.

OFF-(ON) or (ON)-OFF
Contacts are normally open by default, and are closed only while the button is pressed. This is sometimes described as a *make-to-make* connection, or as a *Form A* pushbutton.

ON-(OFF) or (OFF)-ON
Contacts are normally closed by default, and are open only while the button is pressed. This is sometimes described as a *make-to-break* connection, or as a *Form B* pushbutton.

ON-(ON) or (ON)-ON
This is a double-throw pushbutton in which one set of contacts is normally closed. When the button is pressed, the first set of contacts

is opened and the other set of contacts is closed, until the button is released. This is sometimes described as a *Form C* pushbutton.

For a single-throw pushbutton, the terms NC or NO may be used to describe it as *normally closed* or *normally open*.

Slider

This type, also known as a *slide* pushbutton, contains a thin bar or rod that slides in and out of a long, narrow enclosure. Contacts on the rod rub across secondary contacts inside the enclosure. Closely resembling a slider switch, it is cheap, compact, and well adapted for multiple connections (up to 8 separate poles in some models). However, it can only tolerate low currents, has limited durability, and is vulnerable to contamination.

A four-pole, double-throw pushbutton is shown in Figure 5-5. A variety of plastic caps can be obtained to press-fit onto the end of the white nylon actuator.

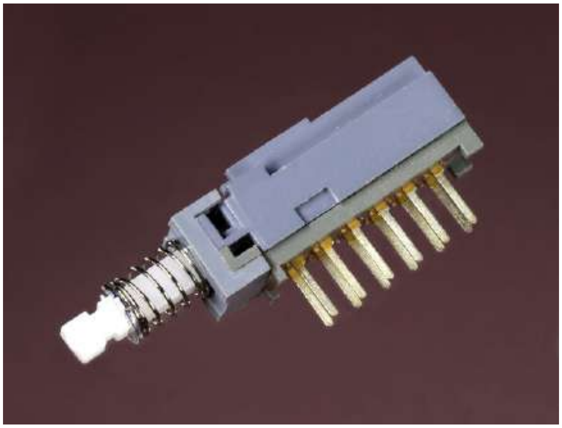

Figure 5-5. *A 4PDT slider pushbutton, shown without the cap that can be snapped onto the end of the actuator.*

Figure 5-6 shows schematic symbols for two possible slide pushbuttons, with a black rectangle indicating each sliding contact. The lead that functions as a pole is marked with a P in each case. Standardization for slide pushbutton schematic symbols does not really exist, but these examples are fairly typical. An insulating section that connects the sliding contacts internally is shown here as a gray rectangle, but in some datasheets may appear as a line or an open rectangle.

Since the symbols for a slide pushbutton may be identical to the symbols for a slide switch, care must be taken when examining a schematic, to determine which type of component is intended.

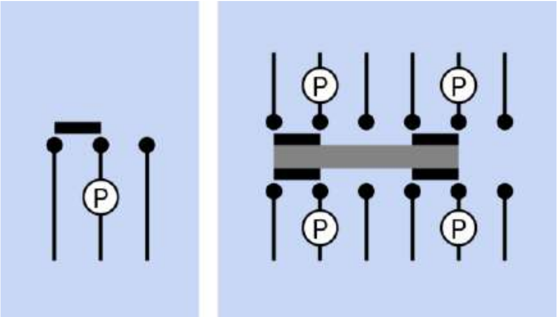

Figure 5-6. *Left: schematic symbol for a simple SPDT slide pushbutton, where a movable contact shorts together either the left pair or right pair of fixed contacts. Right: A 4PDT pushbutton in which the same principle has been extended. The movable contacts are attached to each other mechanically by an insulator. Each pole terminal is marked with a P.*

Styles

Many pushbutton switches are sold without caps attached. This allows the user to choose from a selection of styles and colors. Typically the cap is a push-fit onto the end of the rod or bar that activates the internal contacts. Some sample caps are shown in Figure 5-7, alongside a DPDT pushbutton. Any of the caps will snap-fit onto its actuator.

An *illuminated* pushbutton contains a small *incandescent bulb*, *neon bulb*, or **LED** (*light-emitting diode*). The light source almost always has its own two terminals, which are isolated from the other terminals on the button housing and can be wired to activate the light when the button is pressed, when it is released, or on some other basis. Pushbuttons containing LEDs usually

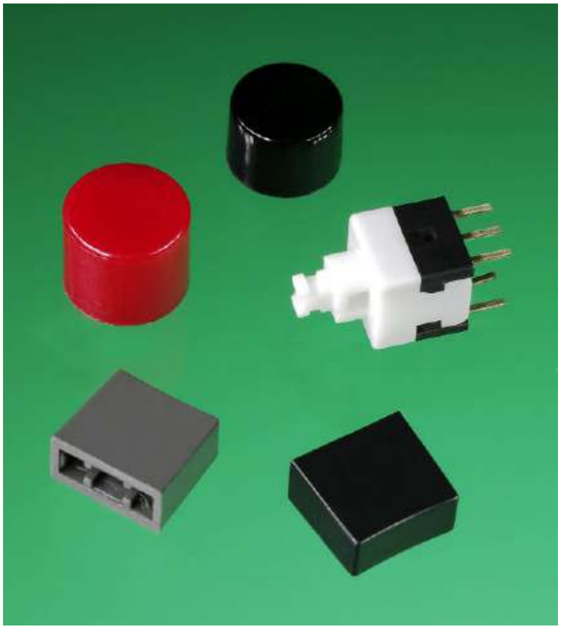

Figure 5-7. *Caps (buttons or knobs) that may be sold as separate accessories for some pushbuttons, shown here alongside a compatible pushbutton switch.*

require external series resistors, which should be chosen according to the voltage that will be used. See the **LED** entry in Volume 2 for additional commentary on appropriate series resistors. An example of an illuminated pushbutton is shown in Figure 5-8. This is a DPDT component, designed to be mounted on a printed circuit board, with an additional lead at each end connecting with an internal LED underneath the translucent white button.

Termination and Contact Plating

These options are the same as for a **switch** and are described in that entry.

Mounting Style

The traditional panel-mounted button is usually secured through a hole in the panel by tightening a nut that engages with a thread on the bushing of the pushbutton. Alternatively, a pushbutton housing can have flexible plastic protrusions on either side, allowing it to be snapped into place in an appropriate-sized panel cutout. This style is shown in Figure 5-4.

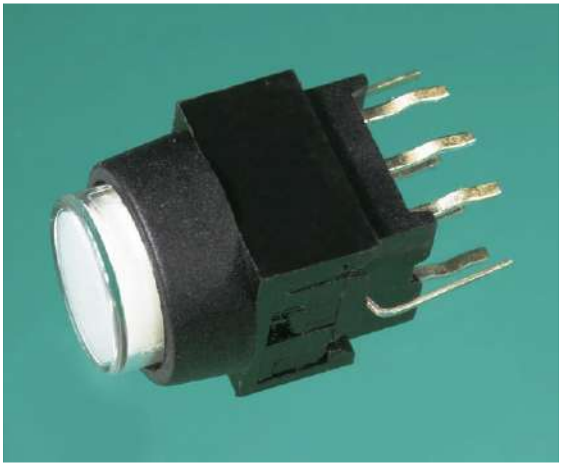

Figure 5-8. *This pushbutton contains an LED underneath the white translucent button.*

PC pushbuttons (pushbuttons mounted in a printed circuit board, or PCB) are a common variant. After the component has been installed in the circuit board, either the button must align with a cutout in the front panel and poke through it when the device is assembled, or an external (non-electrical) button that is part of the product enclosure must press on the actuator of the pushbutton after assembly.

Surface-mount pushbuttons that allow direct fingertip access are uncommon. However, about one-quarter of tactile switches are designed for surface mount at the time of writing. They are typically found beneath membranes that the user presses to activate the switch beneath—for example, in remotes that are used to operate electronic devices.

Sealed or Unsealed

A sealed pushbutton will include protection against water, dust, dirt, and other environmental hazards, at some additional cost.

Latching

This variant, also known as a *press-twice* pushbutton, contains a mechanical ratchet, which is rotated each time the button is pressed. The first press causes contacts to latch in the closed state. The second press returns the contacts to the open state, after which, the process repeats. This press-twice design is typically found on flashlights, audio equipment, and in automotive applications. While *latching* is the most commonly used term, it is also known as *push-push*, *locking*, *push-lock push-release*, *push-on push-off*, and *alternate*.

In a latching pushbutton with *lockdown*, the button is visibly lower in the latched state than in the unlatched state. However, buttons that behave this way are not always identified as doing so on their datasheets.

A six-pole double-throw pushbutton that latches and then unlatches each time it is pressed is shown in Figure 5-9.

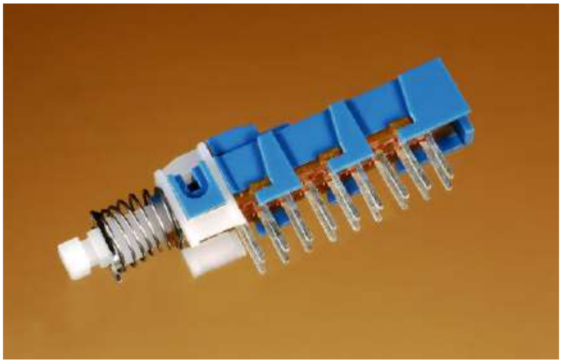

Figure 5-9. *This 6PDT pushbutton latches and then unlatches, each time it is pressed.*

Two more variants are shown in Figure 5-10. On the right is a simple DPDT latching pushbutton with lockdown. On the left is a latching pushbutton that cycles through four states, beginning with one "off" state, the remaining three connecting a different pair of its wires in turn.

A simple OFF-(ON) button may appear to have a latching output if it sends a pulse to a **microcontroller** in which software inside the microcontroller toggles an output between two states. The microcontroller can step through an unlimited number of options in response to each button press. Examples are found on cellular phones or portable media players.

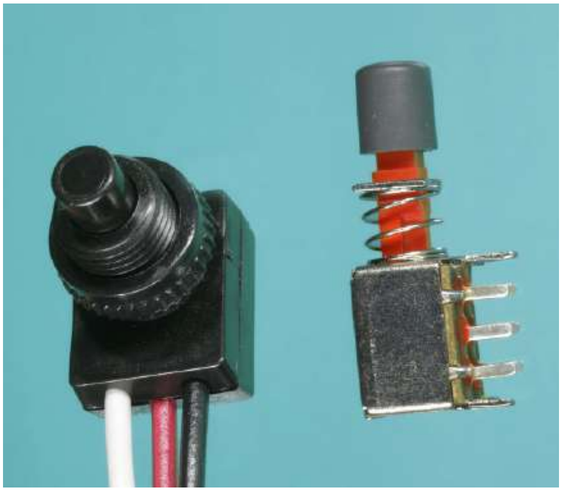

Figure 5-10. *At right, a simple DPDT latching pushbutton with lockdown. At left, this pushbutton cycles through four states, one of them an "off" state, the others connecting a different pair of its wires in turn.*

A mechanically latching pushbutton has a higher failure rate than a simple OFF-(ON) button, as a result of its internal mechanism, but has the advantage of requiring no additional microcontroller to create its output. Microcontrollers are discussed in Volume 2.

Foot Pedal

Foot pedal pushbuttons generally require more actuation force than those intended for manual use. They are ruggedly built and are commonly found in vacuum cleaners, audio-transcription foot pedals, and "stomp boxes" used by musicians.

Keypad

A keypad is a rectangular array of usually 12 or 16 OFF-(ON) buttons. Their contacts are accessed via a *header* suitable for connection with a ribbon cable or insertion into a printed circuit board. In some keypads, each button connects with a separate contact in the header, while all the buttons share a common ground. More often, the buttons are *matrix encoded*, meaning that each of them bridges a unique pair of conductors in a matrix. A 16-button matrix is shown in Figure 5-11. This configuration is suitable for polling by a **microcontroller**, which can be programmed to send an output pulse to each of the four horizontal wires in turn. During each pulse, it checks the remaining four vertical wires in sequence, to determine which one, if any, is carrying a signal. Pullup or pulldown resistors should be added to the input wires to prevent the inputs of the microcontroller from behaving unpredictably when no signal is present. The external appearance of two keypads is shown in Figure 5-12.

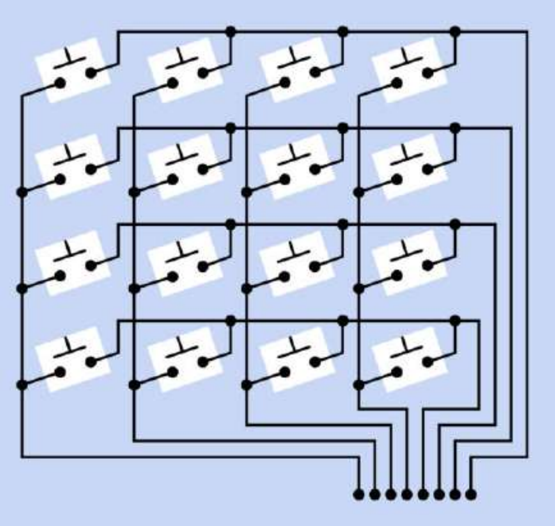

Figure 5-11. *Buttons in a numeric keypad are usually wired as a matrix, where each button makes a connection between a unique pair of wires. This system is suitable for being polled by a microcontroller.*

Tactile Switch

Despite being called a switch, this is a miniature pushbutton, less than 0.4" square, designed for insertion in a printed-circuit board or in a solderless breadboard. It is almost always a SPST device but may have four pins, one pair connected to each contact. Tactile switches may be PC-mounted behind membrane pads. An example is shown in Figure 5-13.

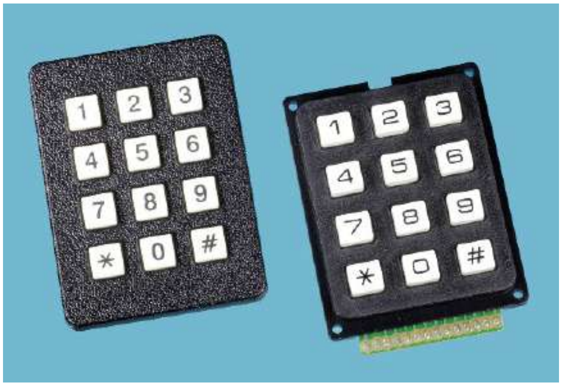

Figure 5-12. *The keypad on the left is matrix-encoded, and is polled via seven through-hole pins that protrude behind it. The keypad on the right assigns each button to a separate contact in its header. See the text for details about matrix encoding.*

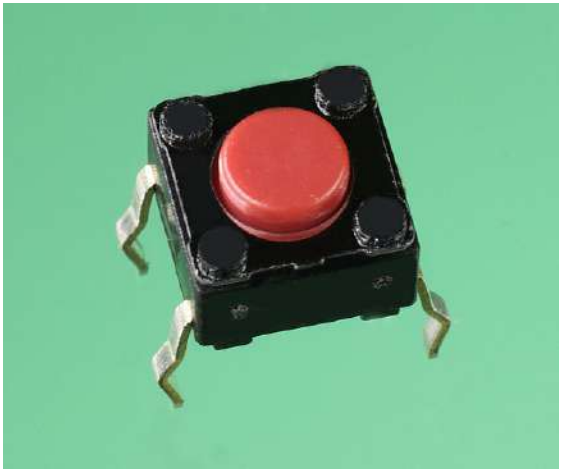

Figure 5-13. *A typical tactile switch.*

Membrane Pad

Typically found on devices such as microwave ovens where contacts must be sealed against particles and liquids. Finger pressure on a membrane pad closes hidden or internal pushbuttons. They are usually custom-designed for specific product applications and are not generally available as generic off-the-shelf components. Some surplus pads may be found for sale on auction websites.

Radio Buttons

The term *radio buttons* is sometimes used to identify a set of pushbuttons that are mechanically interlinked so that only one of them can make an electrical connection at a time. If one button is pressed, it latches. If a second button is pressed, it latches while unlatching the first button. The buttons can be pressed in any sequence. This system is useful for applications such as component selection in a stereo system, where only one input can be permitted at a time. However, its use is becoming less common.

Snap-Action Switches

A snap-action switch (described in detail in the **switch** section of this encyclopedia) can be fitted with a pushbutton, as shown in Figure 5-14. This provides a pleasingly precise action, high reliability, and capability of switching currents of around 5A. However, snap-action switches are almost always single-pole devices.

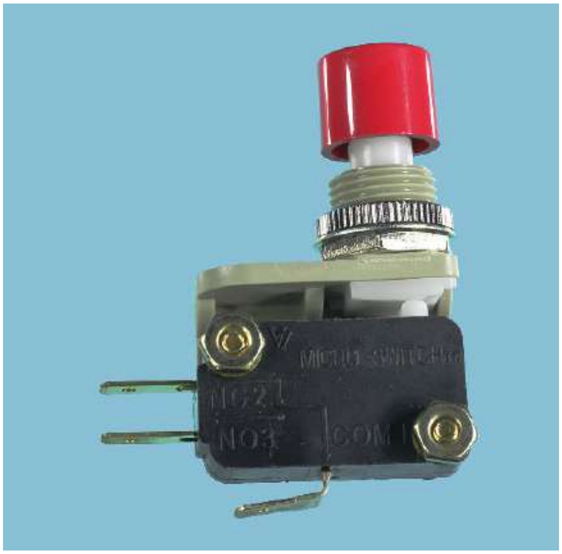

Figure 5-14. *A pushbutton mounted on top of a SPDT snap-action switch.*

Emergency Switch

An emergency switch is a normally-closed device, usually consisting of a large pushbutton that clicks firmly into its "off" position when pressed, and does not spring back. A flange around the button allows it to be grasped and pulled outward to restore it to its "on" position.

Values

Pushbutton current ratings range from a few mA to 20A or more. Many pushbuttons have their current ratings printed on them but some do not. Current ratings are usually specified for a particular voltage, and may differ for AC versus DC.

How to Use it

Issues such as appearance, tactile feel, physical size, and ease of product assembly tend to dictate the choice of a pushbutton, after the fundamental requirements of voltage, current, and durability have been satisfied. Like any electromechanical component, a pushbutton is vulnerable to dirt and moisture. The ways in which a device may be used or abused should be taken into account when deciding whether the extra expense of a sealed component is justified.

When a pushbutton controls a device that has a high inductive load, a *snubber* can be added to minimize arcing. See "Arcing" (page 47) in the **switch** entry of this encyclopedia, for additional information.

What Can Go Wrong

No Button

When ordering a pushbutton switch, read datasheets carefully to determine whether a cap is included. Caps are often sold separately and may not be interchangeable between switches from different manufacturers.

Mounting Problems

In a panel-mount pushbutton that is secured by turning a nut, the nut may loosen with use, allowing the component to fall inside its enclosure when the button is pressed. Conversely, overtightening the nut may strip the threads on the pushbutton bushing, especially in cheaper components where the threads are molded into plastic. Consider applying a drop of Loc-Tite or similar adhesive before completely tightening the nut. Nut sizes vary widely, and finding a replacement may be time-consuming.

LED Issues

When using a pushbutton containing an LED, be careful to distinguish the LED power terminals from the switched terminals. The manufacturer's datasheet should clarify this distinction, but the polarity of the LED terminals may not be clearly indicated. If a diode-testing meter function is unavailable, a sample of the switch should be tested with a source of 3 to 5VDC and a 2K series resistor. Briefly touching the power to the LED terminals, through the resistor, should cause the LED to flash dimly if the polarity is correct, but should not be sufficient to burn out the LED if the polarity is incorrect.

Other Problems

Problems such as arcing, overload, short circuits, wrong terminal type, and contact bounce are generally the same as those associated with a **switch**, and are summarized in that entry in this encyclopedia.

switch

The term **switch** refers here to a physically operated mechanical switch, controlled by flipping a lever or sliding a knob. Although there is some overlap of function, **rotary switches** and **pushbuttons** have their own separate entries. Solid-state switching components are described in entries for **bipolar transistor**, **unijunction transistor**, and **field-effect transistor**. Integrated-circuit switching devices will be found in Volume 2. *Coaxial switches* are used for high-frequency signals, and are not included in this encyclopedia. *Multidirectional switches* differentiate up, down, left, right, diagonal, rotational, and other finger inputs, and are not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **pushbutton** (See Chapter 5)
- **rotary switch** (See Chapter 7)

What It Does

A switch contains at least two contacts, which close or open when an external lever or knob is flipped or moved. Schematic symbols for the most basic type of on-off switch are shown in Figure 6-1.

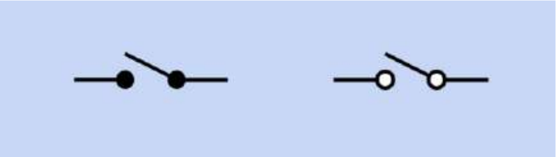

Figure 6-1. *The two most common schematic symbols for a SPST switch, also known as an on-off switch. The symbols are functionally identical.*

The most fundamental type of switch is a *knife switch*, illustrated in Figure 6-2. Although it was common in the earliest days of electrical discovery, today it is restricted to educational purposes in schools, and (in a more robust format) to AC electrical supply panels, where the large contact area makes it appropriate for conducting high amperages, and it can be used for "hot switching" a substantial load.

How It Works

The *pole* of a switch is generally connected with a movable contact that makes or breaks a connection with a secondary contact. If there is only one pole, this is a *single-pole* switch. If there is an additional pole, electrically isolated from the first, with its own contact or set of contacts, this is a *two-pole* switch, also known as a *double-pole* switch. Switches with more than 4 poles are uncommon.

If there is only one secondary contact per pole, this is a *single-throw* or *ST* switch, which may also be described as an *on-off* or *off-on* switch. If there is an additional secondary contact per pole, and the pole of the switch connects with the second contact while disconnecting from the first, this is a *double-throw* or *DT* switch, also known as a *two-way* switch.

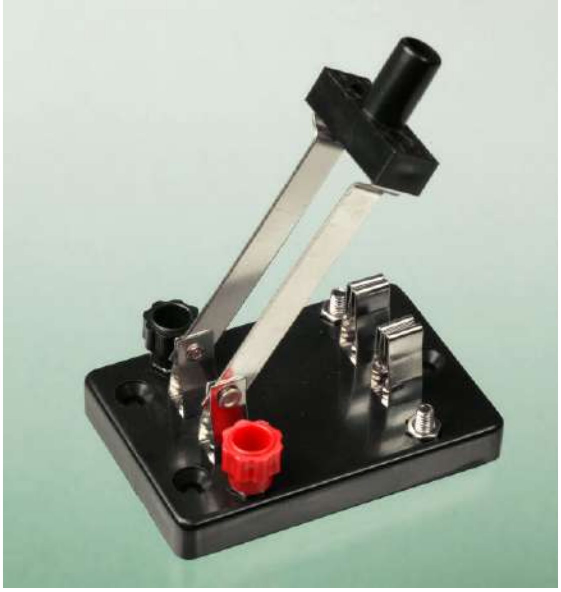

Figure 6-2. *A DPST knife switch intended for the educational environment.*

A double-throw switch may have an additional center position. This position may have no connection (it is an "off" position) or in some cases it connects with a third contact.

Where a switch is spring-loaded to return to one of its positions when manual pressure is released, it functions like a **pushbutton** even though its physical appearance may be indistinguishable from a switch.

Variants

Float switch, *mercury switch*, *reed switch*, *pressure switch*, and *Hall-effect switch* are considered as sensing devices, and will be found in Volume 3.

Terminology

Many different types of switches contain parts that serve the same common functions. The *actuator* is the lever, knob, or toggle that the user turns or pushes. A *bushing* surrounds the actuator on a toggle-type switch. The *common contact* inside a switch is connected with the *pole* of the switch. Usually a *movable contact* is attached to it internally, to touch the secondary contact, also known as a *stationary contact* when the movable contact is flipped to and fro.

Poles and Throws

Abbreviations identify the number of poles and contacts inside a switch. A few examples will make this clear:

SPST also known as 1P1T
Single pole, single throw

DPST also known as 2P1T
Double pole, single throw

SPDT also known as 1P2T
Single pole, double throw

3PST also known as 3P1T
Three pole, single throw

Other combinations are possible.

In Figure 6-3, schematic symbols are shown for double-throw switches with 1, 2, and 3 poles. The dashed lines indicate a mechanical connection, so that all sections of the switch move together when the switch is turned. No electrical connection exists between the poles.

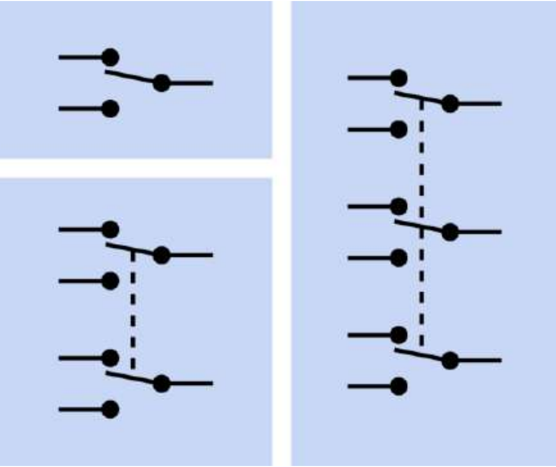

Figure 6-3. *Schematic symbols to represent three types of double-throw switch. Top left: Single-pole. Bottom left: Double-pole. Right: Triple-pole, more commonly known as 3-pole.*

On-Off Behavior

The words ON and OFF are used to indicate the possible states of a switch. The additional word NONE is used by some manufacturers to indicate that a switch does not have a center position. Some manufacturers don't bother with the word NONE, assuming that if the word is omitted, a center position does not exist.

ON-OFF or ON-NONE-OFF

A basic on-off SPST switch with no center position.

ON-ON or ON-NONE-ON

A basic SPDT switch with no center position.

ON-OFF-ON

A double-throw switch with center-off position (no connection when the switch is centered).

ON-ON-ON

A triple-throw switch where the center position connects with its own set of terminals.

Parentheses are used in descriptions of spring-loaded switches to indicate a momentary state that lasts only as long as pressure is applied to the actuator.

(ON)-OFF or OFF-(ON)

A spring-loaded switch that is normally off and returns to that position when pressure is released. Also known as *NO* (normally open), and sometimes described as *FORM A*. Its performance is similar to that of a push-button and is sometimes described as a *make-to-make* connection.

ON-(OFF) or (OFF)-ON

A spring-loaded switch that is normally on and returns to that position when pressure is released. This is sometimes described as a *make-to-break* connection. Also known as *NC* (normally closed), and sometimes described as *FORM B*.

(ON)-OFF-(ON)

A spring-loaded double-throw switch with a no-connection center position to which it returns when pressure on its actuator is released.

Other combinations of these terms are possible.

Most double-throw switches break the connection with one contact (or set of contacts) before making the connection with the second contact (or set of contacts). This is known as a *break before make* switch. Much less common is a *make before break* switch, also known as a *shorting switch*, which establishes the second connection a moment before the first connection is broken. Use of a shorting switch may cause unforeseen consequences in electronic components attached to it, as both sides of the switch will be briefly connected when the switch is turned.

Snap-Action

Also known as a *limit switch* and sometimes as a *microswitch* or *basic switch*. This utilitarian design is often intended to be triggered mechanically rather than with finger pressure, for example in 3D printers. It is generally cheap but reliable.

Two snap-action switches are shown in Figure 6-4. A sectional view of a snap-action ON-(ON) limit switch is shown in Figure 6-5. The pole contacts are mounted on a flexible strip which can move up and down in the center of the switch. The strip has a cutout which allows an inverted U-shaped spring to flip to and fro. It keeps the contacts pressed together in either of the switch states.

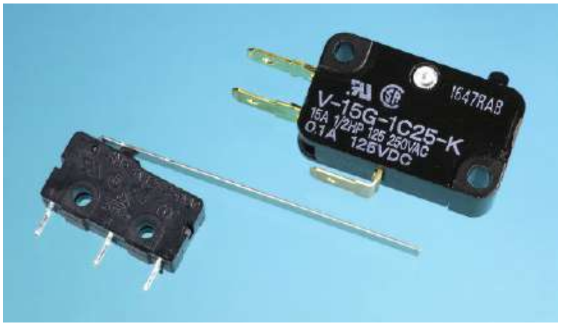

Figure 6-4. *Two SPDT snap-action switches, also known as limit switches. The one on the right is full-size. The one on the left is miniature, with an actuator arm to provide additional leverage. The arm may be trimmed to the required length.*

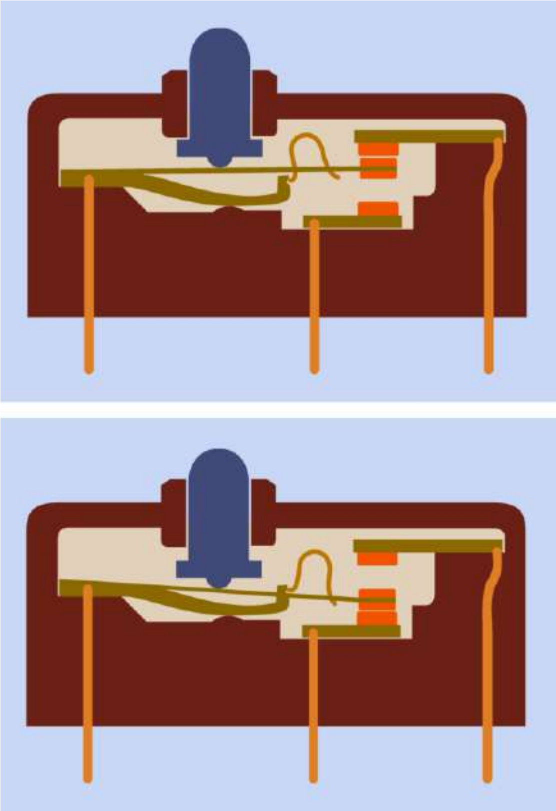

Figure 6-5. *Top: Two contacts inside this limit switch are touching by default. Bottom: When the external button is pressed, it pushes a flexible metal strip downward until it connects with the lower contact. The inverted-U–shaped component is a spring that rests inside a cutout in the flexible strip and resists motion through the central part of its travel.*

The term *snap action* refers to a spring-loaded internal mechanism which snaps to and fro between its two positions. This type of switch is usually SPDT and has a momentary action; in other words, it functions in ON-(ON) mode, although OFF-(ON) and (less often) ON-(OFF) versions are available. The body of the switch is sealed, with a small button protruding through a hole. A thin metal arm may provide additional leverage to press the button. A roller may be mounted at the end of the arm so that the switch can be activated as it slides against a moving mechanical component such as a cam or a wheel. The switch is commonly used to limit the travel or rotation of such a component. Literally thousands of variants are available, in different sizes, requiring different amounts of force for activation. Subminiature snap-action switches can often be actuated by a pressure of only a few grams.

Rocker

Three rocker switches are shown in Figure 6-6. A sectional view of a rocker switch is shown in Figure 6-7. In this design, a spring-loaded ball bearing rolls to either end of a central rocker arm when the switch is turned. Rocker switches are often used as power on-off switches.

Slider

Many types of slider switch (also known as *slide switch*) are widely used as a low-cost but versatile way to control small electronic devices, from clock-radios to stereos. The switch is usually mounted on a circuit board, and its knob or cap protrudes through a slot in the panel. This design is more vulnerable to dirt and moisture than other types of switch. It is usually cheaper than a toggle switch but is seldom designed for use with a high current.

Most slide switches have two positions, and function as SPDT or DPDT switches, but other configurations are less commonly available with more poles and/or positions. A subminiature slide switch is shown in Figure 6-8, while some schematic representations are shown in Figure 6-9,

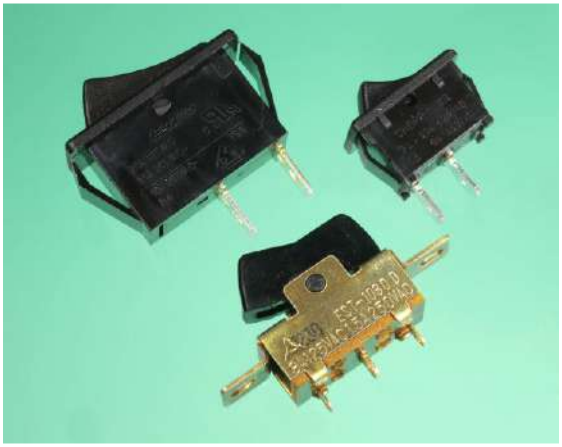

Figure 6-6. *Three rocker switches, the upper two designed for push-insertion into a suitably sized rectangular hole in a panel. The switch at front-center is intended to be screwed in place, and is more than 20 years old, showing that while the choice of materials has changed, the basic design has not.*

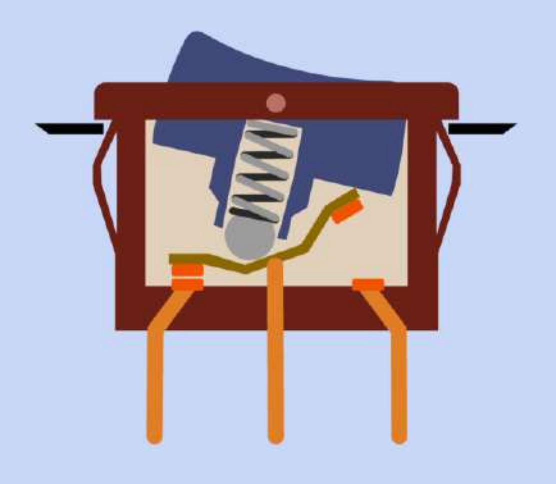

Figure 6-7. *This sectional view of a rocker switch shows a spring-loaded ball-bearing that rolls to and fro along a rocker arm, connecting either pair of contacts when the switch is turned.*

where a black rectangle indicates a sliding internal contact, and a terminal that functions as a pole is identified with letter P in each case. Top left: A SPDT switch using a two-position slider. Top right: A 4PDT slide switch. Bottom left: There are no poles in this switch, as such. The slider can short together any of four pairs of contacts. Bottom right: The slider shorts together three possible pairs of contacts out of four. Here again, there is no pole.

Note that the schematic representation of a slide switch may be identical to that of a slide **pushbutton**. A schematic should be inspected carefully to determine which type is intended.

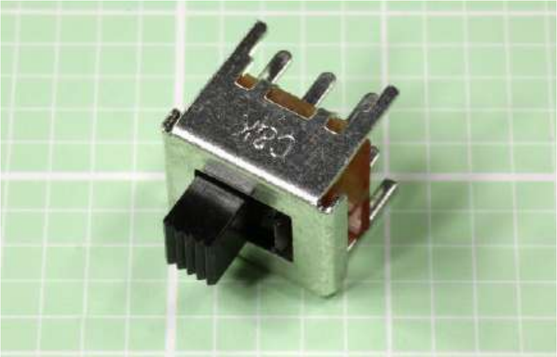

Figure 6-8. *This subminiature slide switch is less than half an inch long, rated 0.3A at 30VDC. Larger versions look almost identical, but can handle only slightly more current.*

The representation of sliders in schematics has not been standardized, but the samples shown are common.

Toggle

A *toggle switch* provides a firm and precise action via a lever (the *toggle*) that is usually tear-drop shaped and nickel plated, although plastic toggles are common in cheaper variants. Formerly used to control almost all electronic components (including early computers), the toggle has declined in popularity but is still used in applications such as automobile accessory kits, motorboat instrument panels, and industrial controls.

Three miniature DPDT toggle switches are shown in Figure 6-10. Two full-size, heavy-duty toggle switches are shown in Figure 6-11. A full-size, four-pole, double-throw heavy-duty toggle switch is shown in Figure 6-12. Toggle switches with more poles are extremely rare.

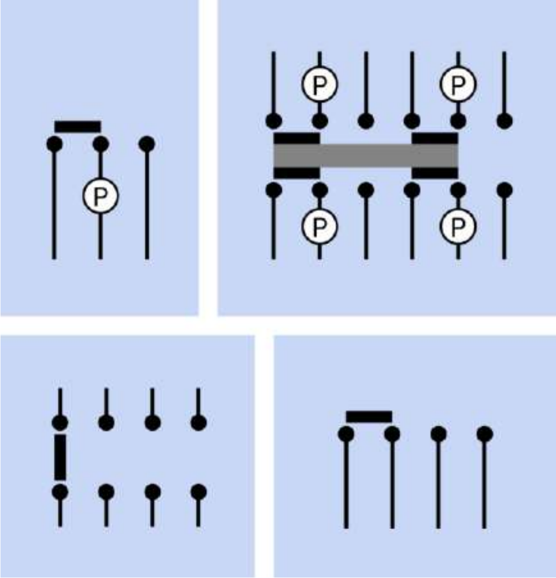

Figure 6-9. *Slide switch schematics. Each black rectangle represents a movable contact that connects two pairs of fixed contacts at a time. Detailed commentary on these variants will be found in the body of the text. Manufacturers may use variants of these symbols in their datasheets (for example, the gray rectangle indicating an insulating contact carrier, at top right, may be represented as a single line, or a black outline with a white center).*

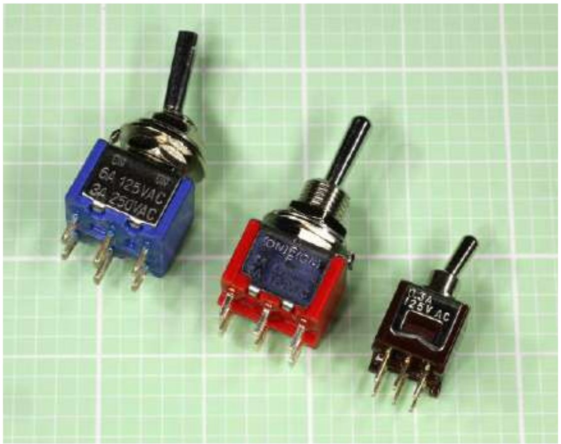

Figure 6-10. *Three miniature toggle switches with current ratings ranging from 0.3A to 6A at 125VAC. Each small square in the background grid measures 0.1" x 0.1".*

An automotive toggle switch is shown in Figure 6-13. Its plastic toggle is extended to minimize operating error.

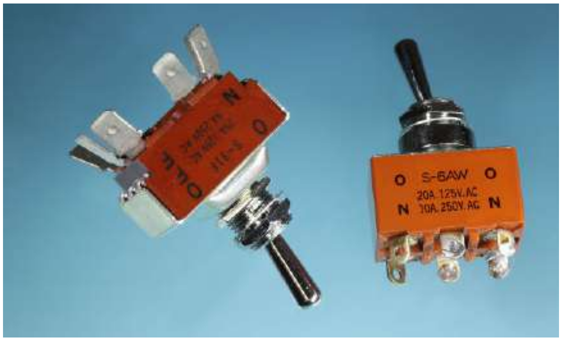

Figure 6-11. *Two full-size toggle switches capable of handling significant current. At left, the switch terminates in quick-connect terminals. At right, the switch has solder terminals (some of them containing residual traces of solder).*

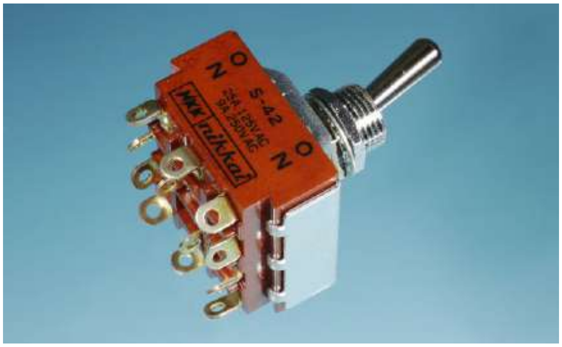

Figure 6-12. *A 4PDT full-size toggle switch with solder terminals, capable of switching 25A at 125VAC. Four-pole switches are relatively unusual.*

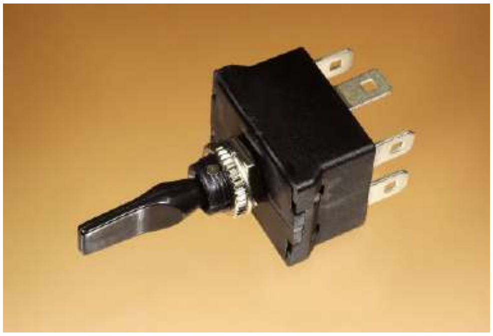

Figure 6-13. *A toggle switch intended for control of automotive accessories.*

High-end toggle switches are extremely durable and can be sealed from environmental contamination with a thin *boot* made from molded rubber or vinyl, which screws in place over the toggle, using the thread on the switch bushing. See Figure 6-14.

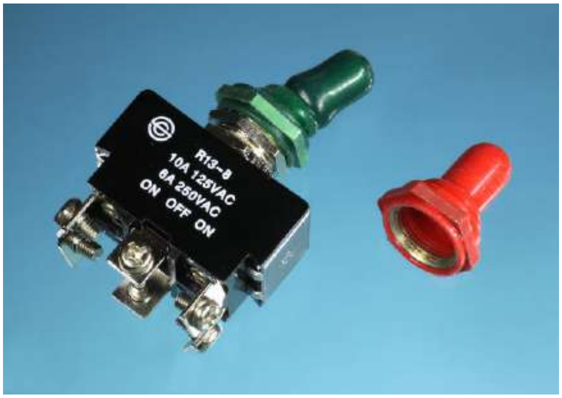

Figure 6-14. *A rubber or vinyl boot can be used to protect a toggle switch from contamination with dirt or water. Each boot contains a nut that screws onto the threads of a toggle switch, as shown at left.*

A *locking toggle switch* has a toggle that must be pulled out against the force of a retaining spring, before the toggle can be moved from one position to another. The toggle then snaps back into place, usually engaging in a small slot in the bushing of the switch.

DIP

A DIP switch is an array of very small, separate switches, designed for mounting directly on a circuit board, either in through-hole or surface-mount format. Through-hole DIP switches have two rows of pins with a 0.1" pitch, the rows being spaced 0.3" apart to fit a standard DIP (dual-inline package) socket or comparable configuration of holes in the board. Surface-mount DIP switches may have 0.1" or 0.05" pitch.

Most DIP arrays consist of SPST switches, each of which can close or open a connection between two pins on opposite sides of the switch body. The switch positions are usually labelled ON and OFF. Figure 6-15 shows a selection of DIP switches. Figure 6-16 shows the internal connections in a DIP switch.

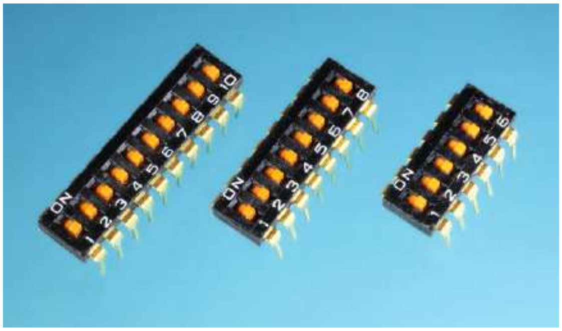

Figure 6-15. *As shown here, DIP switches are available with a variety of "positions," meaning the number of switches, not the number of switch states.*

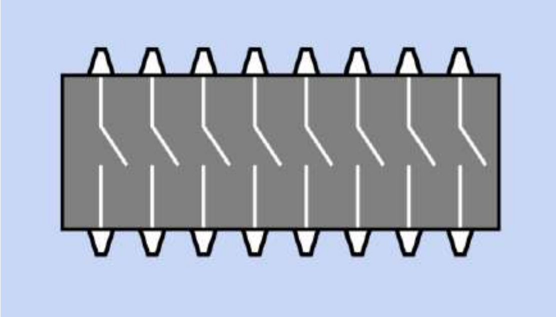

Figure 6-16. *The interior connections of a 16-pin DIP switch.*

The number of switches in a DIP array is usually referred to as its number of "positions." This should not be confused with the two positions of each physical switch lever. SPST DIP switches are made with 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, and 16 positions.

Early IBM-compatible desktop computers often required the user to set the position of an internal DIP switch when making routine upgrades such as installing an additional disk drive. While this feature is now obsolete, DIP switches are still used in scientific equipment where the user is expected to be sufficiently competent to open a cabinet and poke around inside it. Because of the 0.1" spacing, a small screwdriver or the tip of a pen is more appropriate than a finger to flip individual levers to and fro.

DIP switches may also be used during prototype development, as they allow a convenient way to test a circuit in numerous different modes of operation.

Most DIP switches have wire terminals which are just long enough for insertion into a standard breaboard.

DIP switch package options include standard, low-profile, right-angle (standing at 90 degrees relative to the circuit board), and piano (with switch levers designed to be pressed, like tiny rocker switches, instead of being flipped to and fro).

Some SPDT, DPST, DPDT, 3PST, and 4PST variants exist, but are uncommon. Multiple external pins connect with the additional internal switch contacts, and a manufacturer's datasheet should be consulted to confirm the pattern of internal connections. A surface-mount, 0.1" pitch, DPST DIP switch is shown in Figure 6-17, with a plastic cover to protect the switches from contamination during wave soldering (at left), and with the cover peeled off (at right).

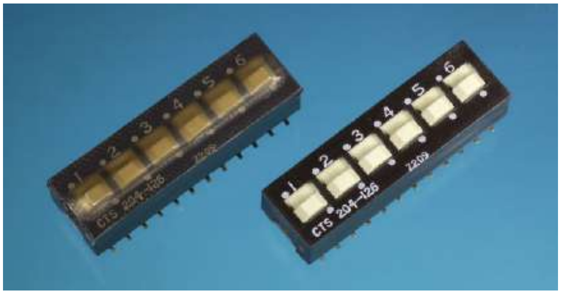

Figure 6-17. *A SPDT surface-mount double-throw DIP switch, sold with a plastic cover (shown at left) to protect it during wave soldering. The cover has been removed at right.*

SIP

A SIP switch is an array of small, separate switches, identical in concept to a DIP switch, but using only one row of pins instead of a double row. The applications for SIP switches are the same as DIP switches; the primary difference is simply that the SIP switch occupies a little less space, while being perhaps slightly less convenient to use.

One terminal of each switch usually shares a common bus. The internal connections in a typical 8-pin SIP array are shown in Figure 6-18. Pin spacing is 0.1", as in a typical DIP switch.

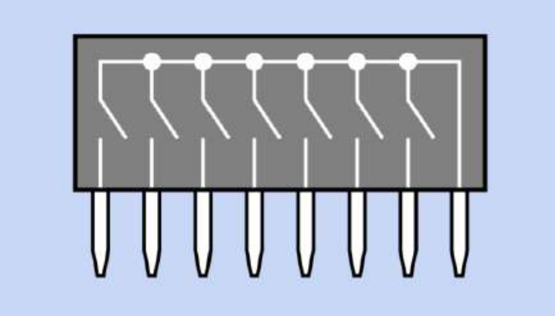

Figure 6-18. *The interior connections of an 8-pin SIP switch incorporating a common bus.*

Paddle

A paddle switch has a flat-sided tab-shaped plastic actuator, relatively large to allow a firm, error-free grip. Internally it is often comparable with a rocker switch, and is generally used with AC power. Some toggle-switch bodies are also sold with paddle-shaped actuators. A subminiature paddle switch is shown in Figure 6-19.

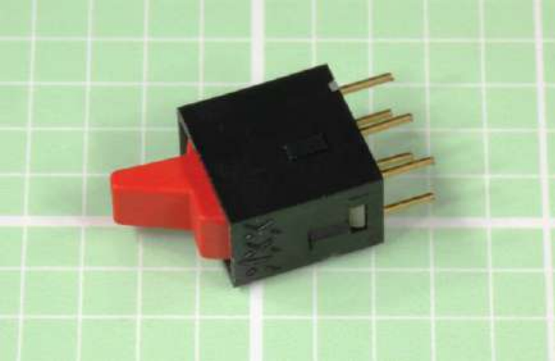

Figure 6-19. *A subminiature paddle switch. Full-size versions are often used as power switches.*

Vandal Resistant Switch

Typically fabricated from stainless steel, this is designed to withstand most types of abuse and is also weather-proof. The pushbuttons that allow pedestrians to trigger a traffic signal are a form of vandal-resistant switch.

Tactile Switch

This is considered to be a **pushbutton**, and is described in that entry. See "Tactile Switch" (page 34).

Mounting Options

A *panel mount* switch generally has a threaded bushing that is inserted from behind the front panel of a product, through a hole of appropriate size. It is supplied with a lockwasher and a nut (often, two nuts) that fit the thread on the switch bushing.

Front panel mount usually means that screws visible from the front of the panel are attached to a bracket on the switch behind the panel. The actuator of the switch is accessible through a cutout in the panel. This mounting style is mostly used for rocker switches and sometimes for slide switches.

Subpanel mount means that the switch is attached to a separate plate or chassis behind the control panel. The actuator of the switch is accessible through a cutout.

Snap-in mount requires a switch with flexible plastic or metal tabs each side, designed to push through a cutout in the panel, at which point the tabs spring out and retain the switch.

PC mount switches have pins that are soldered into a printed circuit board. They may have additional solderable lugs to provide mechanical support.

Surface mount switches are attached to a board in the same manner as other surface-mount components.

Termination

Switches (and pushbuttons) are available with a variety of terminals.

Solder lugs are small tabs, each usually perforated with a hole through which the end of a wire can be inserted prior to soldering.

PC terminals are pins that protrude from the bottom of the switch, suitable for insertion in a printed circuit board. This style is also known as a *through-hole*. The terminals may have a right-angle bend to allow the component to be mounted flat against the board, with the switch actuator sticking out at the side. This termination style is known as *right-angle PC.* Many manufacturers offer a choice of straight or bent pin terminals, but the component may be listed in a catalog under either of those options, with no indication that other options exist. Check manufacturer datasheets carefully.

Quick connect terminals are spade-shaped to accept push-on connectors, commonly used in automotive applications. Hybrid quick-connect terminals that can also function as solder lugs are sometimes offered as an option.

Screw terminals have screws premounted in flat terminals, for solderless attachment of wires.

Wire leads are flexible insulated wires, often with stripped and tinned ends, protruding at least an inch from the body of the component. This option is becoming uncommon.

Contact Plating Options

The internal electrical contacts of a switch are usually plated with silver or gold. Nickel, tin, and silver alloys are cheaper but less common. Other types are relatively rare.

Values

Switches designed for electronic devices vary widely in power capability, depending on their purpose. Rocker switches, paddle switches, and toggle switches are often used to turn power on and off, and are typically rated for 10A at 125VAC, although some toggle switches go as high as 30A. Snap-action or limit switches may be similarly rated, although miniature versions will have reduced capability. Slide switches cannot handle significant power, and are often rated around 0.5A (or less) at 30VDC. DIP and SIP switches have a typical maximum rating of 100mA at 50V and are not designed for frequent use. Generally they are used only when the power to the device is off.

How to Use it

Power Switches

When a simple SPST switch is used to turn DC power on and off, it conventionally switches the positive side of the power supply, also sometimes known as the *high side*. The primary reason for following this convention is that it is widely used; thus, following it will reduce confusion.

More importantly, an on-off switch that controls AC power must be used on the "live" side of the supply, not the "neutral" side. If you have any doubts about these concepts (which go beyond the scope of this book), consult a reference guide on this subject. Using a DPST component to switch both sides of an AC supply may be a worthwhile additional precaution in some applications. The ground wire of an AC supply should never be switched, because the device should always be grounded when it is plugged into an electric outlet.

Limit Switches

An application for two limit switches with a DC motor and two rectifier **diodes** is shown in Figure 6-20. This diagram assumes that the motor turns clockwise when its lower terminal is positive, and counter-clockwise when its upper terminal is positive. Only two terminals are used (and shown) in each limit switch; they are chosen to be normally-closed. Other terminals inside a switch may exist, may be normally-open, and can be ignored.

The motor is driven through a dual-coil, DPDT latching **relay**, which will remain in either position indefinitely without drawing power. When the upper coil of the relay receives a pulse from a pushbutton or some other source, the relay flips to its upper position, which conducts positive current through the lower limit switch, to the lower terminal of the motor. The motor turns clockwise until the arm attached to its shaft hits the lower limit switch and opens it. Positive current is blocked by the lower diode, so the motor stops.

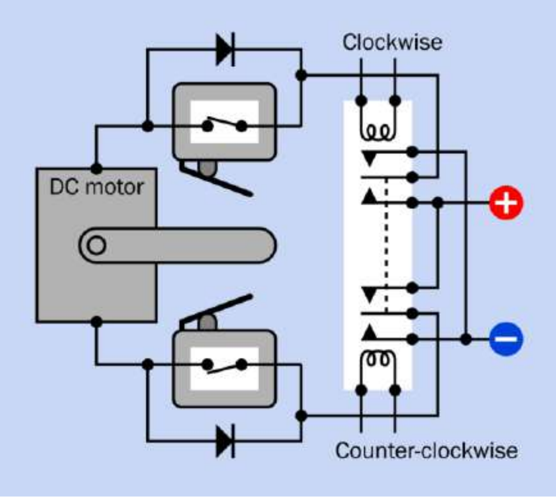

Figure 6-20. *In this schematic, normally-closed limit switches are opened by pressure from an arm attached to a motor, thus switching off its power at each end of its permitted travel and preventing overload and burnout. A two-coil latching relay activates the motor. Rectifier diodes allow power to reach the motor to reverse its rotation when a limit switch is open.*

When the lower coil of the relay is activated, the relay flips to its lower position. Positive current can now reach the upper side of the motor through the upper limit switch. The motor runs counter-clockwise until its arm opens the upper limit switch, at which point the motor stops again. This simple system allows a DC motor to be run in either direction by a button-press of any duration, without risk of burnout when the motor reaches the end of its travel. It has been used for applications such as raising and lowering powered windows in an automobile.

A DPDT pushbutton could be substituted for the latching relay if manual control, only, is acceptable. However, in this scenario, sustained pressure on the pushbutton would be necessary to move the motor arm all the way to the opposite end of its travel. A DPDT switch might be more appropriate than a pushbutton.

Logic Circuits

Logic circuits that depend purely on switches can be constructed (for example, to add binary numbers) but are rare and have no practical applications. The most familiar and simplest example of manually switched logic is a pair of SPDT switches in house wiring, one positioned at the top of a flight of stairs and the other at the bottom, as shown in Figure 6-21. Either switch will turn the light on if it is currently off, or off if it is currently on. To extend this circuit by incorporating a third switch that has the same function as the other two, a DPDT switch must be inserted. See Figure 6-22.

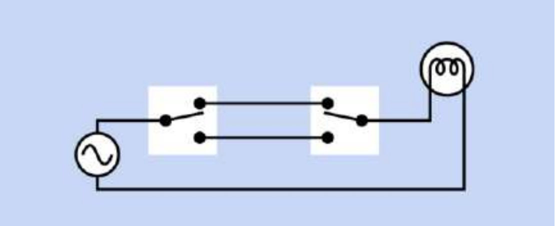

Figure 6-21. *SPDT switches are commonly used in house wiring so that either of them will turn a shared light on if it is off, or off it is on.*

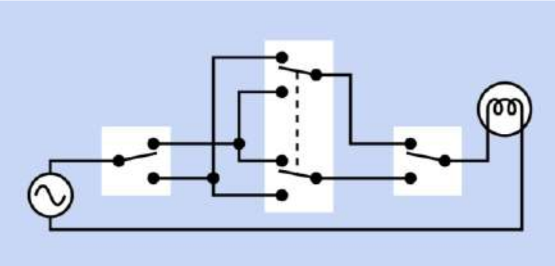

Figure 6-22. *A DPDT switch must be inserted if three switches must have identical function to control the on-off state of a single light bulb.*

Alternatives

As **microcontrollers** have become cheaper and more ubiquitous, they have taken over many functions in electronic products that used to be served by switches. A menuing system driven by a microcontroller can use one **rotational encoder** with a SPST pushbutton built into it to select and adjust numerous parameters in a device such as a car stereo, where functions were once selected and adjusted by individual switches and potentiometers. The rotational-encoder option takes up less space, is cheaper to build (assuming a microcontroller is going to be used in the device for other purposes anyway), and can be more reliable, as it reduces the number of electromechanical parts. Whether it is easier to use is a matter of taste. Cost and ergonomics may be the primary factors to consider when choosing where and how to use switches.

What Can Go Wrong

Arcing

The contacts inside a switch will be rapidly eroded if *arcing* (pronounced "arking") occurs. An electric arc is a spark that tends to form when a switch is opened while conducting a high current or high voltage (typically 10A or more and 100V or more). The most common cause is an inductive load that generates back-EMF when it is switched on and forward-EMF when it is switched off. The surge can be many times the amperage that the load draws during continuous operation. In DC circuits, arcing can be reduced by using a rectifier **diode** in parallel with the load (with its polarity blocking normal current flow). This is often referred to as a *flyback diode* or *freewheeling diode*. In AC circuits, where a diode cannot be used in this way, a *snubber* (a simple combination of capacitor and resistor) may be placed around the load. A snubber can also be used around the switch itself, in DC circuits. See "Snubber" (page 108).

When switching an inductive load, it is generally prudent to use switches rated for twice as much current as the circuit will normally draw.

Dry Joints

Switches that control significant current will have substantial terminals, and these terminals will be attached to heavy-gauge wire. When using solder to make this type of connection, the combined heat capacity of the wire and the terminal will sink a lot more heat than a small component on a circuit board. At least a 30W soldering iron should be used. Lower-wattage irons may be incapable of melting the solder completely (even though they seem to), and a "dry joint" will result, which can have a relatively high electrical resistance and will be mechanically weak, liable to break later. Any good solder joint should withstand flexing of the wire attached to it.

Short Circuits

Because many switches are still wired in with solder tabs, screw terminals, or quick-connect terminals, wires that become accidentally detached can be a significant hazard. Heat-shrink tubing should be applied to enclose wires and terminals at the rear of a power switch, as an additional precaution. Power switches should always be used in conjunction with appropriate **fuses**.

Contact Contamination

Sealed switches should be used in any environment where dirt or water may be present. Slide switches are especially vulnerable to contamination, and are difficult to seal. Switches used in audio components will create "scratchy" sounds if their contacts deteriorate.

Wrong Terminal Type

Because switches are available with a wide variety of terminal types, it's easy to order the wrong type. Switches may be supplied with pins for through-hole insertion in circuit boards; screw terminals; quick-disconnect terminals; or solder lugs. Variants may also be available for surface mount. If a project requires, for example, the insertion of pins in a printed circuit board, and a switch is supplied with solder lugs, it will be unusable.

Part numbers generally include codes to identify each terminal variant, and should be studied carefully.

Contact Bounce

Also known as *switch bounce*. When two contacts snap together, extremely rapid, microscopic vibrations occur that cause brief interruptions before the contacts settle. While this phenomenon is not perceptible to human senses, it can be perceived as a series of multiple pulses by a **logic chip**. For this reason, various strategies are used to *debounce* a switch that drives a logic input. This issue is explored in detail in the entry on logic chips in Volume 2 of the encyclopedia.

Mechanical Wear

Any toggle or rocker switch contains a mechanical pivot, which tend to deteriorate in harsh environments. Friction is also an issue inside these switches, as the design often entails the rounded tip of a lever rubbing to and fro across the center of a movable contact.

The spring inside a snap switch or limit switch may fail as a result of metal fatigue, although this is rare. A slide switch is far less durable, as its contacts rub across each other every time the switch changes position.

In any application that entails frequent switching, or where switch failure is a critical issue, the most sensible practice is to avoid using cheap switches.

Mounting Problems

In a panel-mount switch that is secured by turning a nut, the nut may loosen with use, allowing the component to fall inside its enclosure. Conversely, overtightening the nut may strip the threads on the switch body, especially in cheaper components where the threads are molded into plastic. Consider applying a drop of Loc-Tite or similar adhesive after moderately tightening the nut. Note that nut sizes vary widely, and finding a replacement may be time-consuming.

Cryptic Schematics

In some circuit schematics, the poles of a multi-pole switch may be visually separated from each other, even at opposite sides of the page, for convenience in drawing the schematic. Dotted lines usually, but not always, link the poles. In the absence of dotted lines, switch segments are often coded to indicate their commonality. For example, SW1(a) and SW1(b) are almost certainly different parts of the same switch, with linked poles.

7

rotary switch

Not to be confused with **rotational encoder**, which has its own entry in this encyclopedia.

OTHER RELATED COMPONENTS

- **switch** (See Chapter 6)
- **rotational encoder** (See Chapter 8)

What It Does

A rotary switch makes an electrical connection between a *rotor*, mounted on a shaft that is turned by a knob, and one of two or more stationary contacts. Traditionally, it was the component of choice to select wavebands on a radio receiver, broadcast channels on a television or inputs on a stereo preamplifier. Since the 1990s, it has been substantially superceded by the **rotational encoder**. However it still has applications in military equipment, field equipment, industrial control systems, and other applications requiring a rugged component that will withstand heavy use and a possibly harsh environment. Also, while the output from a rotational encoder must be decoded and interpreted by a device such as a microcontroller, a rotary switch is an entirely passive component that does not require any additional electronics for its functionality.

Two typical schematic symbols for a rotary switch are shown in Figure 7-1. They are functionally identical. A simplified rendering of the interior of a traditional-style rotary switch is shown in Figure 7-2. A separate contact (not shown) connects with the rotor, which connects with each of the stationary contacts in turn. The colors were chosen to differentiate the parts more clearly, and do not correspond with colors in an actual switch.

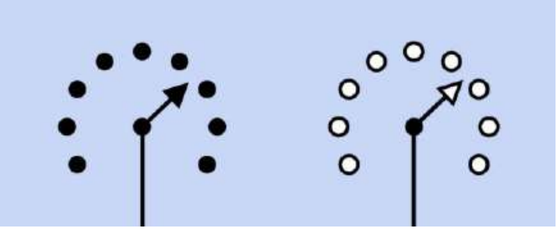

Figure 7-1. *Typical schematic symbols for a rotary switch. The two symbols are functionally identical. The number of contacts will vary depending on the switch.*

A selection of rotary switches is shown in Figure 7-3. At top-left is an *open frame* switch, providing no protection to its contacts from contaminants. This type of component is now rare. At top-right is a twelve-position, single-pole switch rated 2.5A at 125VAC. At front-left is a four-position, single-pole switch rated 0.3A at 16VDC or 100VAC. At front right is a two-position, two-pole switch with the same rating as the one beside it. All the sealed switches allow a choice of panel mounting or through-hole printed circuit board mounting.

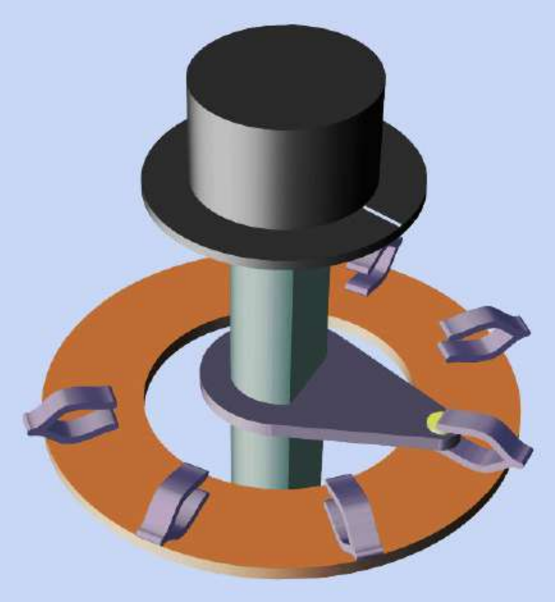

Figure 7-2. *A simplified rendering of interior parts in a basic SP6T rotary switch. Arbitrary colors have been added for clarity.*

How It Works

A switch may have multiple poles, each connecting with its own rotor. The rotors are likely to be on separate *decks* of the switch, but two, three, or four rotors, pointing in different directions, may be combined on a single deck if the switch has only a small number of positions.

Rotary switches are usually made with a maximum of twelve positions, but include provision for limiting the number of positions with a *stop*. This is typically a pin, which may be attached to a washer that fits around the bushing of the switch. The pin is inserted into a choice of holes to prevent the switch from turning past that point. For example, an eight-position rotary switch can be configured so that it has only seven (or as few as two) available positions.

A specification for a rotary switch usually includes the angle through which the switch turns between one position and the next. A twelve-position switch usually has a 30-degree turn angle.

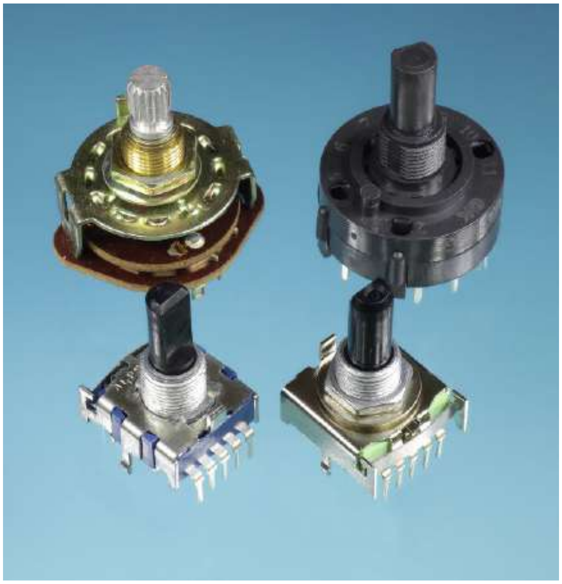

Figure 7-3. *A selection of rotary switches. See text for details.*

Variants

Conventional

The traditional style of rotary switch is designed to be panel-mounted, with a body that ranges from 1" to 1.5" in diameter. If there is more than one deck, they are spaced from each other by about 0.5". The switch makes an audible and tactile "click" as it is turned from one position to the next.

A rugged sealed five-deck rotary switch is shown in Figure 7-4. It has five poles (one per deck), and a maximum of 12 positions. The contacts are rated 0.5A at 28VDC. This type of heavy-duty component is becoming relatively rare.

If the rotor in a switch establishes a connection with the next contact a moment before breaking the connection with the previous contact, this is known as a *shorting* switch, which may also be described as a *make-before-break* switch. In a *nonshorting* or *break-before-make* switch, a tiny

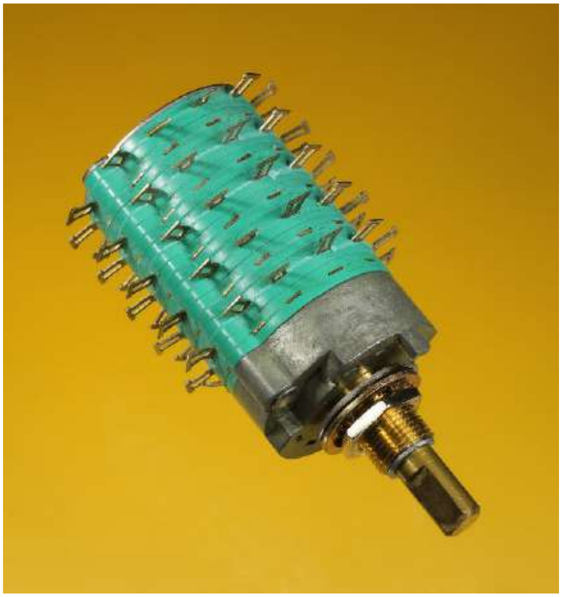

Figure 7-4. *A five-pole, twelve-position rotary switch.*

interval separates one connection from the next. This can be of significant importance, depending on the components that are connected with the switch.

The shaft may be round, splined, or D-shaped in section. A knob is seldom supplied with a switch and must be chosen to match the shaft. Some shaft dimensions are metric, while others are measured in inches, with 1/4" diameter being the oldest standard. Some switches with a splined shaft are supplied with an adapter for a knob of D-shaped internal section; the adapter can be slipped onto the shaft in any of 12 or more positions, to minimize the inconvenience of positioning the body of the switch itself so that the knob is correctly oriented in relation to positions printed on the face of the panel.

Miniature rotary switches may be as small as 0.5" diameter, and usually terminate in pins for through-hole mounting on a PC board. Miniature switches usually have lower current ratings than full-size switches.

Rotary switches must be securely anchored to resist the high turning forces that can be inflicted upon them by users. In a panel-mount design, a nut is tightened around a thread on the bushing of the switch. Through-hole versions can be secured to the PC board with the shaft protruding loosely through a cutout in the panel. To minimize mechanical stress on the circuit board, the detents in a PC-board switch are usually weaker than in a full-size switch, and the knob is usually smaller, allowing less leverage.

Rotary DIP

A conventional *DIP switch* is a linear array of miniature SPST switches designed to fit a standard DIP (dual-inline package) layout of holes in a circuit board. It is described in the **switch** entry of this encyclopedia. A *rotary DIP* switch (also known as an *encoded output rotary switch* or a *coded rotary switch*) does not conform with a DIP layout, despite its name. It is approximately 0.4" square and usually has five pins, one of which can be considered the input or common pin while the other four can function as outputs. The pins are spaced at 0.1" pitch from one another. Pin function and layout are not standardized.

A dial on top of the switch has either 10 positions (numbered 0 through 9) or 16 positions (0 through 9 followed by letters A through F). One switch of each type is shown in Figure 7-5.

Each position of the dial closes pairs of contacts inside the component to create a unique binary-coded decimal pattern (in a 10-position switch) or binary-coded hexadecimal pattern (in the 16-position switch) on the four output pins. The pin states are shown in Figure 7-6. A rotary DIP switch is a relatively flimsy device, and is not designed for frequent or heavy use. It is more likely to be a "set it and forget it" device whose state is established when it is installed in a circuit board.

Because each position of the switch is identified with a unique binary pattern, this is an example of *absolute encoding*. By contrast, a typical **rotational encoder** uses *relative encoding*, as it merely generates a series of undifferentiated pulses when the shaft is turned.

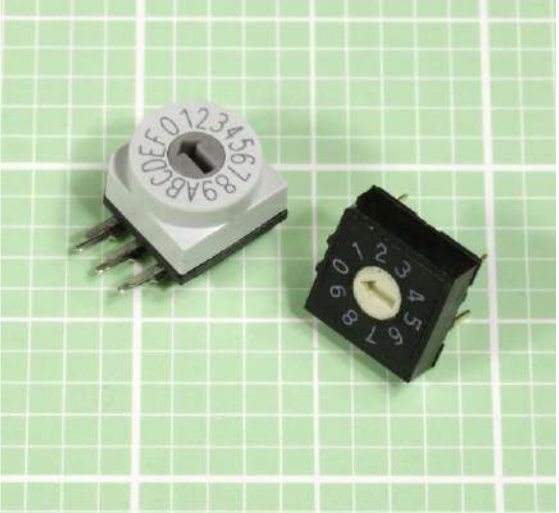

Figure 7-5. *A rotary DIP switch, also known as an encoded output rotary switch, may be used as a substitute for a DIP switch in some applications.*

A *real-coded* rotary DIP makes a connection between input and output pins wherever a binary 1 would exist. In the *complement-coded* version, the output is inverted. The switch is primarily intended for use with a **microcontroller**, enabling only four binary input pins on the microcontroller to sense up to sixteen different switch positions.

A six-pin rotary DIP variant is available from some manufacturers, with two rows of three pins, the two center pins in each row being tied together internally, and serving as the pole of the switch.

Rotary DIPs are available with a screw slot, small knurled knob, or larger knob. The screw-slot version minimizes the height of the component, which can be relevant where circuit boards will be stacked close together. A *right-angle PC* variant stands at 90 degrees to the circuit board, with pins occupying a narrower footprint. The switch on the left in Figure 7-5 is of this type.

While most rotary DIPs are through-hole components, surface-mount versions are available.

Most rotary DIPs are sealed to protect their internal components during wave-soldering of circuit boards.

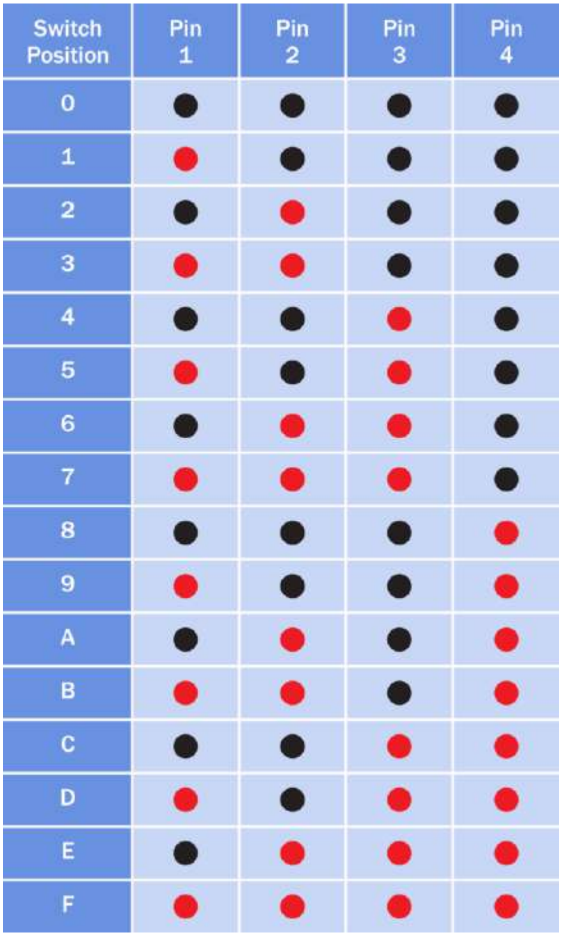

Switch Position	Pin 1	Pin 2	Pin 3	Pin 4
0	⚫	⚫	⚫	⚫
1	🔴	⚫	⚫	⚫
2	⚫	🔴	⚫	⚫
3	🔴	🔴	⚫	⚫
4	⚫	⚫	🔴	⚫
5	🔴	⚫	🔴	⚫
6	⚫	🔴	🔴	⚫
7	🔴	🔴	🔴	⚫
8	⚫	⚫	⚫	🔴
9	🔴	⚫	⚫	🔴
A	⚫	🔴	⚫	🔴
B	🔴	🔴	⚫	🔴
C	⚫	⚫	🔴	🔴
D	🔴	⚫	🔴	🔴
E	⚫	🔴	🔴	🔴
F	🔴	🔴	🔴	🔴

Figure 7-6. *Positive and negative states of the four output pins of a real-coded 16-position rotary DIP switch, assuming that the common pin of the switch is connected with a positive supply voltage. A ten-position rotary DIP switch would use only the states from 0 through 9. In a complement-coded switch, the positive and negative states would be reversed.*

Gray Code

A Gray code (named after its originator, Frank Gray) is a system of absolute encoding of a switch output, using a series of nonsequential binary numbers that are chosen in such a way that each number differs by only one digit from the preceding number. Such a series is useful because it eliminates the risk that when a switch turns, some bits in the output will change before others, creating the risk of erroneous interpretation. A minority of rotary switches or rotational encoders are available with Gray-coded outputs. Typically, a microcontroller must use a lookup table to convert each binary output to an angular switch position.

PC Board Rotary Switch

Miniature switches with a conventional, non-encoded output are available for printed-circuit board mounting, sometimes requiring a screwdriver or hex wrench to select a position. A single-pole eight-position switch of this type is shown in Figure 7-7. Its contacts are rated to carry 0.5A at 30VDC, but it is not designed to switch this current actively.

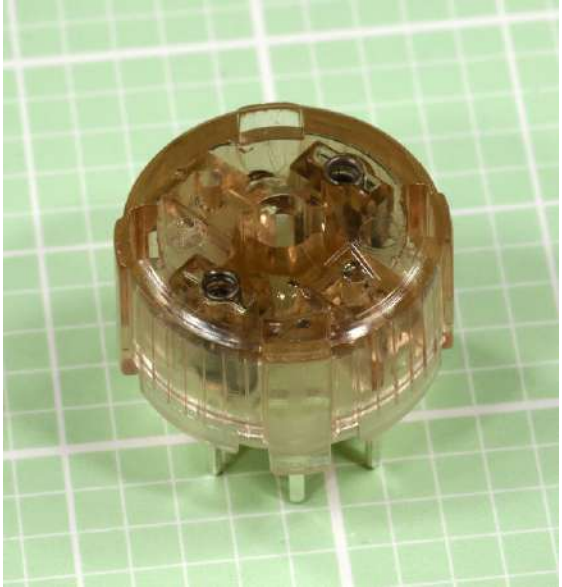

Figure 7-7. *This miniature switch is designed for insertion on a printed circuit board. It can be used to make a setting before a device is shipped to the end user.*

Mechanical Encoder

A *mechanical encoder* functions similarly to a rotary DIP switch but is intended for much heavier use. It outputs a binary-coded-decimal value corresponding with its shaft position, is typically the size of a miniature rotary switch, and is designed for panel mounting. The Grayhill Series 51 allows 12 positions, each generating a code among four terminals. The Bourns EAW provides 128 positions, each generating a code among 8 terminals.

Pushwheel and Thumbwheel

A *pushwheel switch* is a simple electromechanical device that enables an operator to provide a code number as input to data processing equipment, often in industrial process control. The decimal version contains a wheel on which numbers are printed, usually in white on black, from 0 through 9, visible one at a time through a window in the face of the switch. A button above the wheel, marked with a minus sign, rotates it to the next lower number, while a button below the wheel, marked with a plus sign, rotates it to the next higher number. A connector at the rear of the unit includes a common (input) pin and four output pins with values 1, 2, 4, and 8. An additional set of pins with values 1, 2, 3, and 4 is often provided. The states of the output pins sum to the value that is currently being displayed by the wheel. Often two, three, or four pushwheels (each with an independent set of connector pins) are combined in one unit, although individual pushwheels are available and can be stacked in a row.

A *thumbwheel switch* operates like a pushwheel switch, except that it uses a thumbwheel instead of two buttons. Miniaturized thumbwheel switches are available for through-hole mounting on PC boards.

Hexadecimal versions are also available, with numbers from 0 through 9 followed by letters A through F, although they are less common than decimal versions.

Keylock

A *keylock switch* is generally a two-position rotary switch that can be turned only after insertion of a key in a lock attached to the top of the shaft. This type of switch almost always has an OFF-(ON) configuration and is used to control power.

Keylock switches are found in locations such as elevators, for fire-department access; in cash registers; or on data-processing equipment where switching power on or off is reserved for a system administrator.

Values

A full-size rotary switch may be rated from 0.5A at 30VDC to 5A at 125VAC, depending on its purpose. A very few switches are rated 30A at 125VAC; these are high-quality, durable, expensive items.

A typical rotary DIP switch is rated 30mA at 30VDC and has a carrying current rating (continuous current when no switching occurs) of no more than 100mA at 50VDC.

How to Use it

In addition to its traditional purpose as a mode or option selector, a rotary switch provides a user-friendly way to input data values. Three ten-position switches, for instance, can allow user input of a decimal number ranging from 000 to 999.

When used with a **microcontroller**, a rotary switch can have a *resistor ladder* mounted around its contacts, like a multi-point *voltage divider*, so that each position of the rotor provides a unique potential ranging between the positive supply voltage and negative ground. This concept is illustrated in Figure 7-8, where all the resistors have the same value. The voltage can be used as an input to the microcontroller, so long as the microcontroller shares a common ground with the switch. An analog-digital converter inside the microcontroller translates the voltage into a digital value. The advantage of this scheme is that it allows very rapid control by the user, while requiring only one pin on the microcontroller to sense as many as twelve input states.

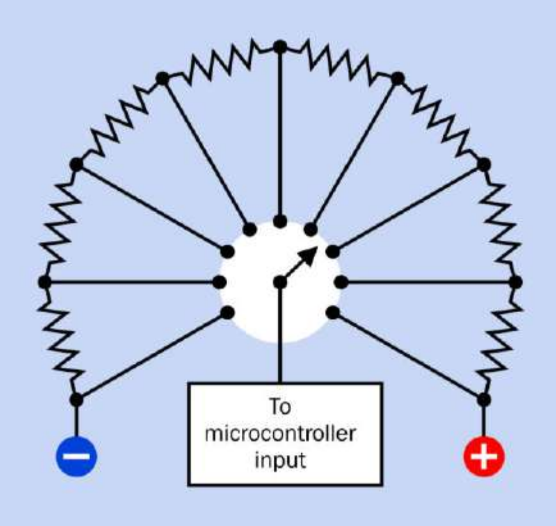

Figure 7-8. *A resistor ladder can be formed around the contacts of a rotary switch, with the pole of the switch connected to a microcontroller that has an analog-digital converter built in. The microcontroller converts the voltage input to an internal digital value. Thus, one pin can sense as many as twelve input states.*

For a ladder consisting of 8 resistors, as shown, each resistor could have a value of 250Ω. (The specifications for a particular microcontroller might require other values.) To avoid ambiguous inputs, a nonshorting rotary switch should be used in this scheme. A *pullup resistor* of perhaps 10K should be added to the microcontroller input, so that there is no risk of it “floating” when the switch rotor is moving from one contact to the next. The code that controls the microcontroller can also include a blanking interval during which the microcontroller is instructed to ignore the switch.

Because the rotary switch is an electromechanical device, it has typical vulnerabilities to dirt and moisture, in addition to being bulkier, heavier, and more expensive than a **rotational encoder**. Rotary switches have also been partially replaced by pushbuttons wired to a microcontroller. This option is found on devices ranging from digital alarm clocks to cellular phones. In addition to being cheaper, the pushbutton alternative is preferable where space on a control panel, and behind it, is limited.

What Can Go Wrong

Vulnerable Contacts

Most modern rotary switches are sealed, but some are not. Any switch with exposed contacts will be especially vulnerable to dirt and moisture, leading to unreliable connections. This was an issue in old-fashioned TV sets, where periodic contact cleaning of the channel selector switch was needed.

Exposed contacts are also more vulnerable to side-effects from temperature cycling (when a device warms up and then cools down).

Contact Overload

The contacts on a cheap rotary switch are especially vulnerable to arcing, as the user may turn the switch slowly, causing gradual engagement and disengagement of contacts instead of the snap-action that is characteristic of a well-made toggle switch. If a rotary switch may control significant currents or current surges, it must be appropriately rated, regardless of the extra expense. For more information on arc suppression in switches, see "Arcing" (page 47).

Misalignment

Most knobs for rotary switches consist of a pointer, or have a white line engraved to provide clear visual indication of the position of the switch. If this does not align precisely with indications printed on the panel, confusion will result. For hand-built equipment, the switch can be installed first, after which the control-panel indications can be glued or riveted in place on a separate piece of laminated card, plastic, or metal for precise alignment. If the switch is not secured tightly, its body may turn slightly under repeated stress, leading to erroneous interpretation of the knob position.

Misidentified Shorting Switch

If a shorting switch is used where a nonshorting switch was intended, the results can be disconcerting or even destructive, as one terminal will be briefly connected with the adjacent terminal while the switch is being turned. Multiple functions of a circuit may be activated simultaneously, and in a worst case scenario, adjacent terminals may be connected to opposite sides of the same power supply.

User Abuse

The turning force that must be applied to a full-size conventional rotary switch is significantly greater than the force that is applied to most other types of panel-mounted switches. This encourages aggressive treatment, and the turning motion is especially likely to loosen a nut holding the switch in place. The lighter action characteristic of miniature rotary switches does not necessarily solve this problem, as users who are accustomed to older-style switches may still apply the same force anyway.

Rotary switches should be mounted in expectation of rough use. It is prudent to use Loc-Tite or a similar compound to prevent nuts from loosening, and a switch should not be mounted in a thin or flimsy panel. When using a miniature rotary switch that has through-hole mounting in a circuit board, the board must be sufficiently robust and properly secured.

Wrong Shaft, Wrong Knobs, Nuts That Get Lost, Too Big to Fit

These problems are identical to those that can be encountered with a **potentiometer**, which are discussed in that entry in this encyclopedia.

rotational encoder

The term *rotational encoder* used to be reserved for high-quality components, often using optical methods to measure rotation with precision (more than 100 intervals in 360 degrees). Cheaper, simpler, electromechanical devices were properly referred to as *control shaft encoders*. However, the term *rotational encoder* is now applied to almost any device capable of converting rotational position to a digital output via opening and closing internal mechanical contacts; this is the sense in which the term is used here. It is sometimes distinguished from other types of encoder with the term *mechanical rotary encoder*. Magnetic and optical rotary encoders do not contain mechanical switches, are classified as *sensors* by this encyclopedia, and will appear in Volume 3. They are found in a device such as an *optical mouse*.

OTHER RELATED COMPONENTS

- **rotary switch** (See Chapter 7)

What It Does

A rotational encoder has a knob that a user can turn to display a series of prompts on an LCD screen, or to adjust the input or output on a product such as a stereo receiver. The component is almost always connected to inputs on a **microcontroller** and is usually fitted with *detents* that provide tactile feedback suggesting many closely spaced positions. The encoder often allows the user to make a selection by pushing the knob in, which closes an internal momentary switch. Thus, this type of encoder functions as a pushbutton as well as a switch.

A rotational encoder is an *incremental* or *relative* device, meaning that it merely creates and breaks internal switch connections when rotation occurs, without providing a unique code to identify each absolute rotational position. An *absolute* encoder is discussed in the **rotary switch** entry of this encyclopedia.

No schematic symbol exists to represent a rotational encoder.

How It Works

An encoder contains two pairs of contacts, which open and close out of phase with each other when the shaft rotates. In a clockwise direction, the A pair of contacts may be activated momentarily before the B pair; in a counter-clockwise direction, the B pair may be activated before the A pair. (Some encoders reverse this phase difference.) Thus if one contact from each pair is connected with two inputs of an appropriately programmed microcontroller, and if the other contact of each pair is connected with negative ground, the microcontroller can deduce which way the knob is turning by sensing which pair of contacts closes first. The microcontroller can then count the number of pulses from the contacts and interpret this to adjust an output or update a display.

A simplified schematic is shown in Figure 8-1. The two buttons inside the dashed line represent the two pairs of contacts inside the encoder, while the chip is a microcontroller. The knob and shaft that activate the internal switches are not shown. The schematic assumes that when a contact closes, it pulls the chip input to a low state. A pullup resistor is added to each input of the chip to prevent the pins from “floating” when either pair of contacts is open.

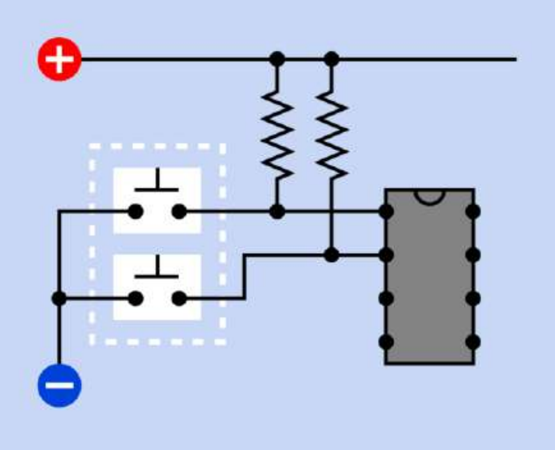

Figure 8-1. *Simplified schematic showing the typical setup for a rotational encoder. The pushbuttons inside the dashed line represent the contacts inside the encoder. The chip is a microcontroller.*

Figure 8-2 gives a conceptual view of the outputs of an encoder that is turned clockwise (top) and then counter-clockwise (bottom). Some encoders may reverse this phase sequence. Red and black colors have been assigned to the pin states on the assumption that the terminals that are common to both pairs of contacts are connected with negative ground. Thus a “high” pulse in the graphical representation actually indicates that the encoder is grounding its output.

Microcontrollers have become so ubiquitous, and rotational encoders are so cheap, they have displaced rotary switches in many applications where a low current is being switched. The combination of a rotational encoder and a microcontroller is very versatile, allowing display and control of an almost unlimited number of menus and options.

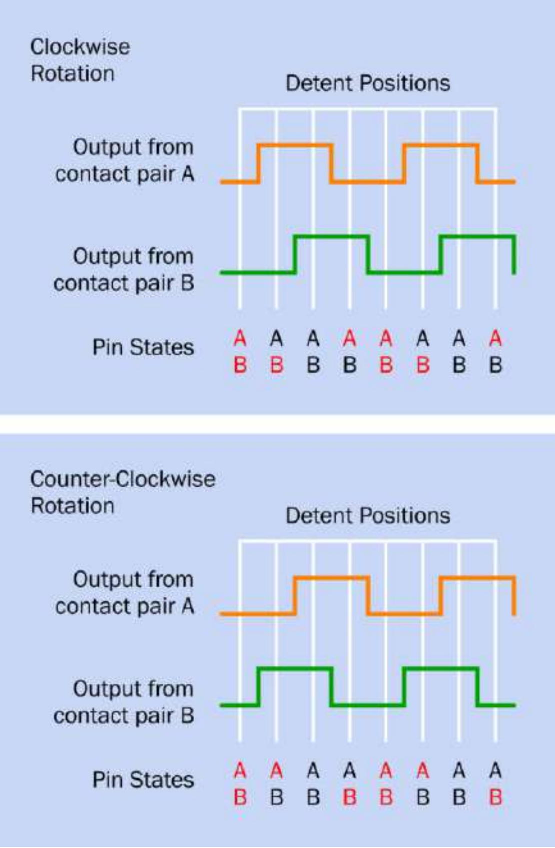
Clockwise Rotation
Detent Positions
Output from contact pair A
Output from contact pair B
Pin States
A A A A A A A A
B B B B B B B B
Counter-Clockwise Rotation
Detent Positions
Output from contact pair A
Output from contact pair B
Pin States
A A A A A A A A
B B B B B B B B

Figure 8-2. *Hypothetical outputs from a rotational encoder, assuming that the common terminals of the contact pairs are connected to negative ground. A high pulse in the graphical representation therefore indicates that the contact pair is grounded. The number of detents relative to the number of pulses per rotation varies from one type of encoder to another.*

Variants

There are two types of rotational encoders containing mechanical contacts: absolute and relative. An absolute encoder generates a code corresponding with each specific rotational position. The code is usually a binary output among four or more pins. It is discussed under *mechanical encoder* in the **rotary switch** section of this encyclopedia. The variants listed here are all relative encoders.

Pulses and Detents

Rotational encoders from different manufacturers may have as few as 4 or as many as 24 pulses per rotation (PPR), with 12 to 36 detents (or no detents at all, in a few models.) The relationship between pulses and detents shown in Figure 8-2 is typical but is far from being universal. The number of detents may be equal to, greater than, or less than the number of pulses per rotation.

Format

Rotational encoders are generally panel-mounted or through-hole devices. In the latter category, most are horizontally mounted, with a minority being at 90 degrees to the board.

Output

In an encoder containing two switches, four switch-state combinations are possible: OFF-OFF, ON-OFF, OFF-ON, and ON-ON. This is known as a *quadrature* output. All of the rotational encoders discussed here conform with that system.

Rotational Resistance

Rotational encoders vary widely in the resistance that they offer when the user turns the knob. This is largely a function of the detents, if they are included. Still, all rotational encoders generally offer less rotational resistance than a rotary switch, and do not have the kind of heavy-duty knobs that are typically used with rotary switches. Since an encoder creates only a stream of pulses without any absolute positional information, a knob with any kind of pointer on it is inappropriate.

Values

Virtually all rotational encoders are designed to work with a low-voltage supply, 12VDC or less. All of them are intended for low currents, reflecting their purpose to drive microcontroller inputs. Some sample rotational encoders are pictured in Figure 8-3. At rear: nine pulses per rotation (PPR), 36 detents, 10mA at 10VDC. Far left: 20PPR, 20 detents, with switch. Far right: 24PPR, no detents, 1mA at 5VDC. Center (blue): 16PPR, no detents, 1mA at 5VDC. Front: 12PPR, 24 detents, 1mA at 10VDC, requires Allen wrench or similar hexagonal shaft to engage with the rotor.

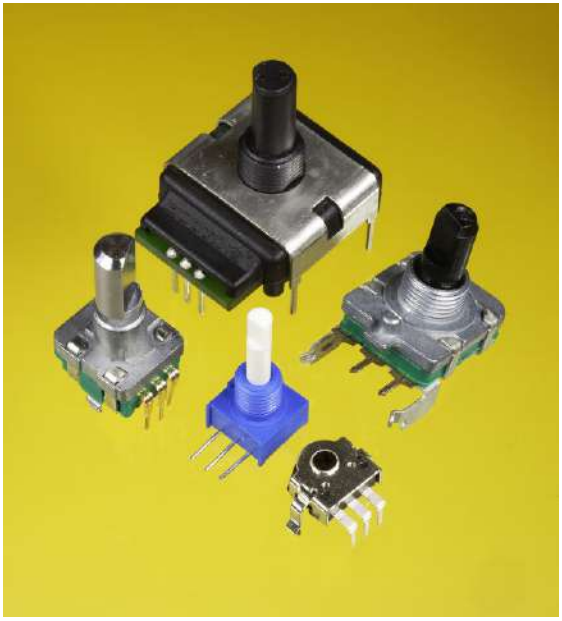

Figure 8-3. *Rotational encoders with a variety of specifications. See text for details.*

Contact Bounce

Any mechanical switch will suffer some degree of *contact bounce* when its contacts close. Datasheets for rotational encoders may include a specification for bounce duration ranging from around 2ms to 5ms, which is sometimes known as the *settling time*. Naturally, a lower value is preferred. The microcontroller that interprets the positional information from the encoder can include a debouncing routine that simply disregards any signals during the bounce period following switch closure.

Sliding Noise

Sliding noise is the opposite of contact bounce. When two contacts have made a connection and then rub across each other (as occurs inside a rotational encoder while the knob is being turned), the connection may suffer momentary lapses. Datasheets for rotational encoders generally do not supply ratings for this.

How to Use it

As noted above, a rotational encoder can only be used in conjunction with a microcontroller or similar device that is capable of interpreting the phase difference between the pairs of contacts, and is capable of counting the number of opening/closing events while the knob is being turned. (Some dedicated chips are designed for this specific purpose.)

It can be adapted to be driven by a **stepper motor**, to provide feedback regarding the rotation of the motor shaft, and its output can also be interpreted to calculate angular acceleration.

Programming the microcontroller is the most significant obstacle. Generally the program should follow a sequence suggested by this pseudocode:

Check:

- If the encoder contains a pushbutton switch, check it. If the pushbutton is being pressed, go to an appropriate subroutine.
- The status of contacts A.
- The status of contacts B.

Compare their status with previously saved states for A and B. If the status has not changed, repeat from Check.

Debounce:

- Recheck the contacts status rapidly and repeatedly for 50ms, and count the states for contacts A and B. (The 50ms duration may be adjusted for different encoders, as an encoder with a higher number of pulses per rotation will tend to create shorter pulses.)
- Compare the total number of changed states with unchanged states.

If the changed states are in a small minority, probably the signal was erroneous, caused by bounce or sliding noise. Go back to Check and start over.

Interpret:

- Deduce the rotational direction from these four possibilities:
  - — Contacts A were open and have closed.
  - — Contacts A were closed and have opened.
  - — Contacts B were open and have closed.
  - — Contacts B were closed and have opened. (The specific type of encoder will determine how these transitions are interpreted.)
- Revise the variable storing the direction of rotation if necessary.
- Depending on the direction of rotation, increment or decrement a variable that counts pulses.
- Take action that is appropriate to the direction of rotation and the cumulative number of pulses.
- Go back to Check again.

What Can Go Wrong

Switch Bounce

In addition to a debouncing algorithm in the microcontroller, a 0.1μF bypass capacitor can be used with each of the output terminals from the encoder, to help reduce the problem of switch bounce.

Contact Burnout

Rotational encoders are TTL-compatible. They are not generally designed to drive even a small output device, such as an LED. The contacts are extremely delicate and will be easily damaged by any attempt to switch a significant current.

relay

Properly known as an *electromagnetic armature relay* to distinguish it from a *solid-state relay*. However, the full term is very rarely used. It may also be described as an *electromechanical relay*, but the term **relay** is normally understood to mean a device that is not solid state.

OTHER RELATED COMPONENTS

- **solid state relay** (Volume 2)
- **switch** (See Chapter 6)

What It Does

A relay enables a signal or pulse of electricity to switch on (or switch off) a separate flow of electricity. Often, a relay uses a low voltage or low current to control a higher voltage and/or higher current. The low voltage/low current signal can be initiated by a relatively small, economical switch, and can be carried to the relay by relatively cheap, small-gauge wire, at which point the relay controls a larger current near to the load. In a car, for example, turning the ignition switch sends a signal to a relay positioned close to the starter motor.

While solid-state switching devices are faster and more reliable, relays retain some advantages. They can handle double-throw and/or multiple-pole switching and can be cheaper when high voltages or currents are involved. A comparison of their advantages relative to **solid state relays** and **transistors** is tabulated in the entry on **bipolar transistor** in Figure 28-15.

Common schematic symbols for single-throw relays are shown in Figure 9-1 and for double-throw relays in Figure 9-2. The appearance and orientation of the coil and contacts in the symbols may vary significantly, but the functionality remains the same.

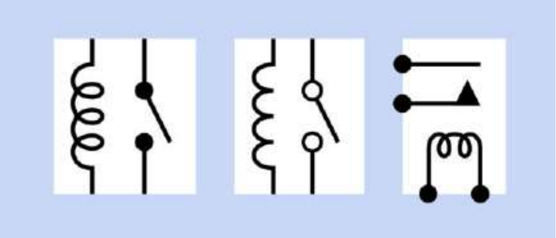

Figure 9-1. *Commonly used schematic symbols for a SPST relay. The symbols are functionally identical.*

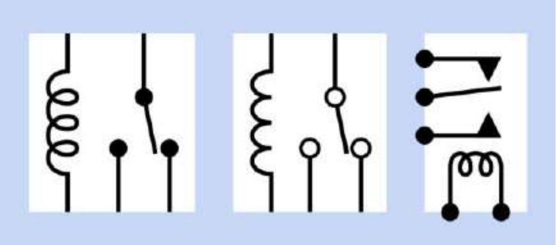

Figure 9-2. *Commonly used schematic symbols for a SPDT relay. The symbols are functionally identical.*

How It Works

A relay contains a *coil*, an *armature*, and at least one pair of *contacts*. Current flows through the coil, which functions as an **electromagnet** and generates a magnetic field. This pulls the armature, which is often shaped as a pivoting bracket that closes (or opens) the contacts. These parts are visible in the simplified rendering of a DPST relay in Figure 9-3. For purposes of identification, the armature is colored green, while the coil is red and the contacts are orange. The two blue blocks are made of an insulating material, the one on the left supporting the contact strips, the one on the right pressing the contacts together when the armature pivots in response to a magnetic field from the coil. Electrical connections to the contacts and the coil have been omitted for simplicity.

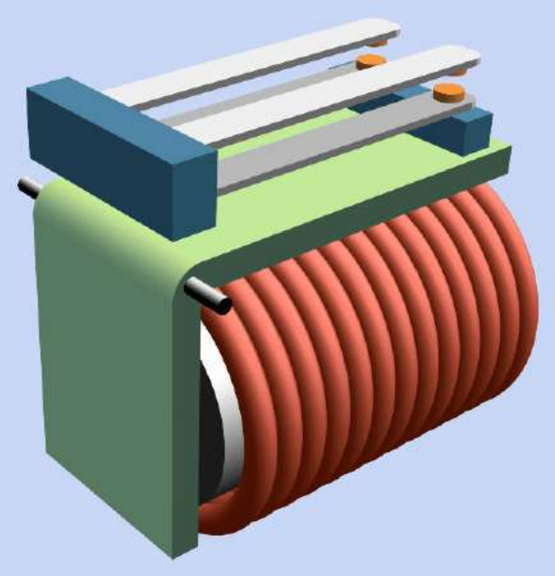

Figure 9-3. *This simplified rendering shows the primary parts of a DPST relay. See text for details.*

Various small relays, capable of handling a variety of voltages and currents, are pictured in Figure 9-4. At top-left is a 12VDC automotive relay, which plugs into a suitable socket shown immediately below it. At top-right is a 24VDC SPDT relay with exposed coil and contacts, making it suitable only for use in a very clean, dry environment. Continuing downward, the four sealed relays in colored plastic cases are designed to switch currents of 5A at 250VAC, 10A at 120VAC, 0.6A at 125VAC, and 2A at 30VDC, respectively. The two blue relays have 12VDC coils, while the red and yellow relays have 5V coils. All are nonlatching, except for the yellow relay, which is a latching type with two coils. At bottom-left is a 12VDC relay in a transparent case, rated to switch up to 5A at 240VAC or 30VDC.

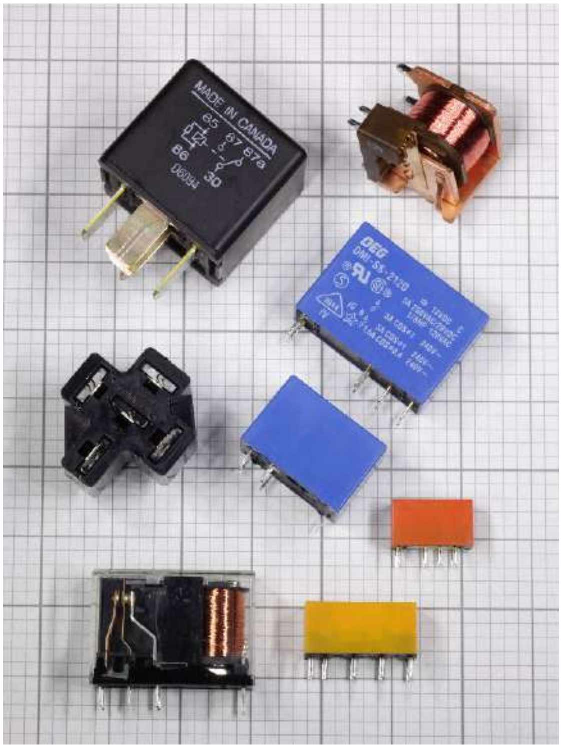

Figure 9-4. *An assortment of small DC-powered relays. See text for details.*

The configuration of a relay is specified using the same abbreviations that apply to a **switch**. SP, DP, 3P, and 4P indicate 1, 2, 3, or 4 poles (relays with more than 4 poles are rare). ST and DT indicate single-throw or double-throw switching. These abbreviations are usually concatenated, as in 3PST or SPDT. In addition, the terminology Form

A (meaning normally open), Form B (normally closed), and Form C (double-throw) may be used, preceded by a number that indicates the number of poles. Thus "2 Form C" means a DPDT relay.

Variants

Latching

There are two basic types of relay: *latching* and *nonlatching*. A nonlatching relay, also known as a *single side stable* type, is the most common, and resembles a *momentary switch* or **pushbutton** in that its contacts spring back to their default state when power to the relay is interrupted. This can be important in an application where the relay should return to a known state if power is lost. By contrast, a latching relay has no default state. Latching relays almost always have double-throw contacts, which remain in either position without drawing power. The relay only requires a short pulse to change its status. In semiconductor terms, its behavior is similar to that of a *flip-flop*.

In a *single-coil* latching relay, the polarity of voltage applied to the coil determines which pair of contacts will close. In a *dual-coil* latching relay, a second coil moves the armature between each of its two states.

Schematic symbols for a dual-coil latching relay are shown in Figure 9-5. Some symbol styles do not make it clear which switch position each coil induces. It may be necessary to read the manufacturer's datasheet or test the relay by applying its rated voltage to randomly selected terminal pairs while testing for continuity between other terminal pairs.

Polarity

There are three types of DC relay. In a *neutral relay*, polarity of DC current through the coil is irrelevant. The relay functions equally well either way. A *polarized relay* contains a diode in series with the coil to block current in one direction. A *biased relay* contains a permanent magnet near the armature, which boosts performance when current flows through the coil in one direction, but blocks a response when the current flows through the coil in the opposite direction. Manufacturers' datasheets may not use this terminology, but will state whether the relay coil is sensitive to the polarity of a DC voltage.

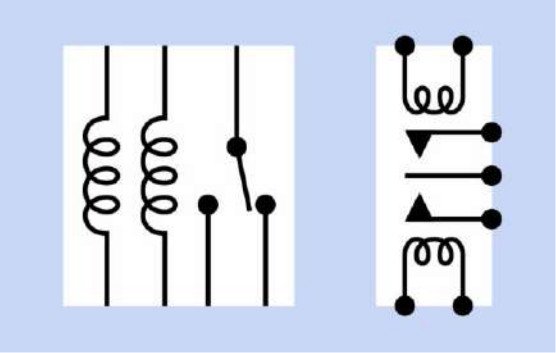

Figure 9-5. *Schematic symbols for a two-coil latching relay. The symbols are functionally identical.*

All relays can switch AC current, but only an *AC relay* is designed to use AC as its coil current.

Pinout Variations

The layout and function of relay pins or quick connects is not standardized among manufacturers. Often the component will have some indication of pin functions printed on it, but should always be checked against the manufacturer's datasheet and/or tested for continuity with a meter.

Figure 9-6 shows four sample pin configurations, adapted from a manufacturer's datasheet. These configurations are functionally quite different, although all of them happen to be for DPDT relays. In each schematic, the coil of the relay is shown as a rectangle, while the pins are circles, black indicating an energized state and white indicating a non-energized state. The bent lines show the possible connections between the poles and other contacts inside the relay. The contacts are shown as arrows. Thus, pole 4 can connect with either contact 3 or contact 5, while pole 9 can connect with either contact 8 or contact 10.

Top-left: Polarized nonlatching relay in its resting condition, with no power applied. Top right: Single-coil latching relay showing energized contacts (black circles) when the coil is powered with the polarity indicated. If the polarity is reversed, the relay flips to its opposite state. Some manufacturers indicate the option to reverse polarity by placing a minus sign alongside a plus sign, and a plus sign alongside a minus sign. Bottom-left and bottom-right: Polarized latching relays with two coils, with different pinouts.

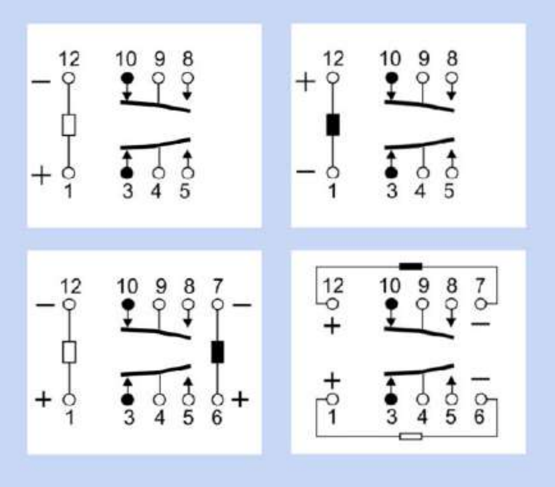

Figure 9-6. *Relay pinouts depicted in the style commonly found in manufacturers' datasheets, showing different relay types. Top-left: Single coil, nonlatching. Top-right: Single coil, latching. Bottom left: Two-coil, latching. Bottom right: Two-coil, latching, alternate pinouts. (Adapted from a Panasonic datasheet.)*

In these diagrams, the relay is seen from above. Some datasheets show the relay seen from below, and some show both views. Some manufacturers use slightly different symbols to indicate interior functions and features. When in doubt, use a meter for verification.

Reed Relay

A *reed relay* is the smallest type of electromechanical relay with applications primarily in test equipment and telecommunications. With a coil resistance ranging from 500 to 2000 ohms, these relays consume very little power. The design consists of a *reed switch* with a coil wrapped around it. Figure 9-7 shows a simplified rendering. The two black contacts are enclosed in a glass or plastic envelope and magnetized in such a way that a magnetic field from the surrounding coil bends them together, creating a connection. When power to the coil is disconnected, the magnetic field collapses and the contacts spring apart.

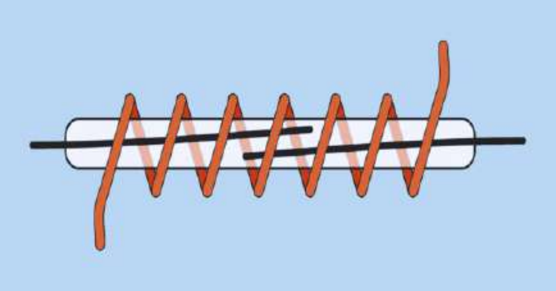

Figure 9-7. *This simplified rendering shows a reed relay, consisting of a magnetized reed switch inside a glass or plastic pod, activated by a coil wrapped around it.*

In Figure 9-8, two reed relays are shown, at top-left and center-right. At bottom-left, the type of relay on the right has been opened by a belt sander to reveal its copper coil and inside that, a capsule in which the relay contacts are visible.

Surface-mount reed relays can be smaller than 0.5″ × 0.2″. Through-hole versions are often around 0.7″ × 0.3″ with pins in two rows, though some are available in SIP packages.

Reed relays have limited current switching capacity and are not suitable to switch inductive loads.

Small Signal Relay

A *small signal relay* is also known as a *low signal relay*. This type may have a footprint as small as a reed relay but generally stands slightly taller, requires slightly more coil current, and is available in versions that can switch slightly higher

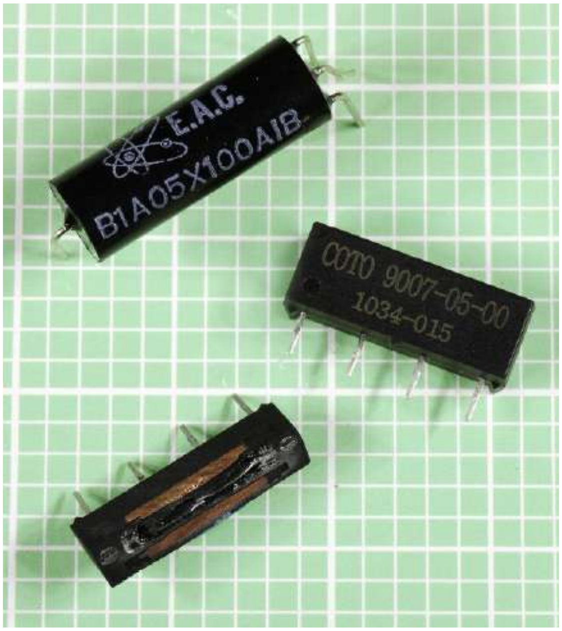

Figure 9-8. *Three reed relays, one of which has had its packaging partially removed by a belt sander to reveal its copper coil and internal contacts.*

voltages and currents. There are usually two rows of pins, spaced either 0.2″ or 0.3″ apart. The red and orange relays in Figure 9-4 are small signal relays.

Automotive Relays

An *automotive relay* is typically packaged in a cube-shaped black plastic case with quick-connect terminals at the bottom, typically plugged into a socket. Naturally they are designed to switch, and be switched by, a 12VDC supply.

General Purpose/Industrial

These relays cover a very wide range and are usually built without significant concern for size. They may be capable of switching high currents at high voltages. Typically they are designed to plug into a socket such as an *octal base* of the type that was once used for **vacuum tubes**. The base, in turn, terminates in solder tabs, screws, or quick connects and is designed to be screwed to a chassis. It allows the relay to be unplugged and swapped without resoldering.

Two industrial relays are shown in Figure 9-9. Both are DPDT type with 12VDC coils and rated to switch up to 10A at 240VAC. The one on the left has an octal base. An octal socket that fits an octal base is shown in Figure 9-10.

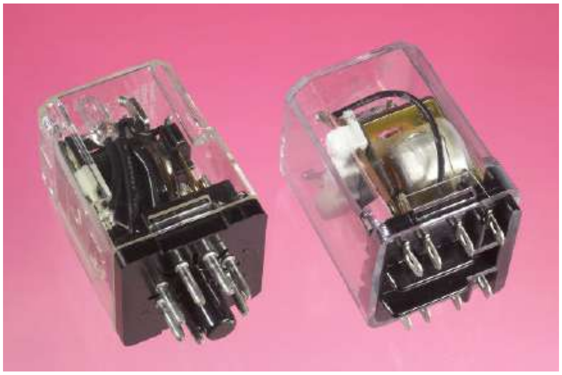

Figure 9-9. *Two relays powered by 12VDC, capable of switching up to 10A at 240VAC.*

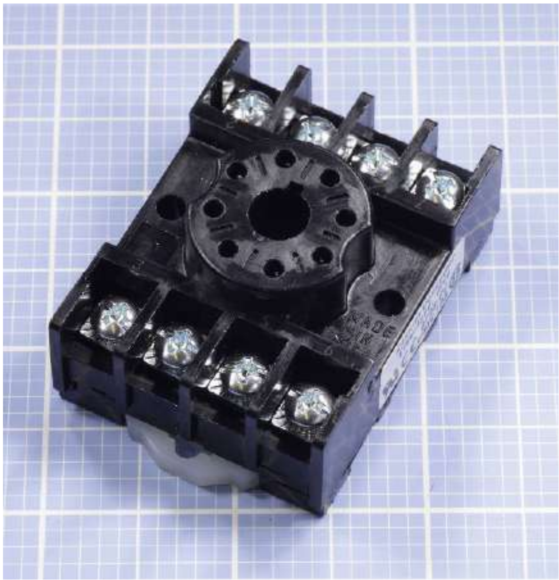

Figure 9-10. *An octal socket with screw terminals, designed to accept a relay with an octal base.*

Time Delay Relay

Generally used to control industrial processes, a *time delay relay* switches an output on and off at preset time intervals that can be programmed to repeat. The example in Figure 9-11 has a 12VDC coil and is rated to switch up at 10A at 240VAC. It has an octal base.

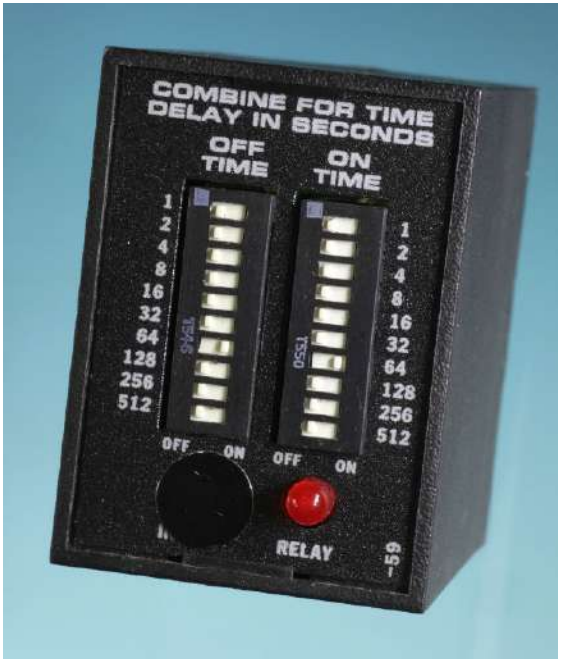

Figure 9-11. *The control switches on a time-delay relay, allowing separately configured "on" and "off" intervals.*

Contactor

A contactor functions just like a relay but is designed to switch higher currents (up to thousands of amperes) at higher voltages (up to many kilovolts). It may range from being palm-sized to measuring more than one foot in diameter, and may be used to control heavy loads such as very large motors, banks of high-wattage lights, and heavy-duty power supplies.

Values

Datasheets usually specify maximum voltage and current for the contacts, and nominal voltage and current for the coil, although in some cases the coil resistance is stated instead of nominal coil current. The approximate current consumption can be estimated, if necessary, by using Ohm's Law. The minimum voltage that the relay needs for activation is sometimes described as the *Must Operate By* voltage, while the *Must Release By* voltage is the maximum coil voltage that the relay will ignore. Relays are rated on the assumption that the coil may remain energized for long periods, unless otherwise stated.

While the contact rating may suggest that a relay can switch a large load, this is not necessarily true if the load has significant inductance.

Reed relays

Usually use a coil voltage of 5VDC and have a contact rating of up to 0.25A at 100V. Through-hole (PCB) versions may have a coil voltage of 5VDC, 6VDC, 12VDC, or 24VDC and in some cases claim to switch 0.5A to 1A at up to 100V, although this rating is strictly for a noninductive load.

Small signal/low signal relays

Usually use a coil voltage ranging from 5VDC to 24VDC, drawing about 20mA. Maximum switching current for noninductive loads ranges from 1A to 3A.

Industrial/general purpose relays

A very wide range of possible values, with coil voltages ranging up to 48VDC or 125VAC to 250VAC. Contact rating is typically 5A to 30A.

Automotive relays

Coil voltage of 12VDC, and contact rating often 5A at up to 24VDC.

Timer relays

Usually these specify a coil voltage of 12VDC, 24VDC, 24VAC, 125VAC, or 230VAC. The timed interval can range from 0.1 sec to 9999 hours in some cases. Common values for contact ratings are 5A up to 20A, with a voltage of 125V to 250V, AC or DC.

How to Use it

Relays are found in home appliances such as dishwashers, washing machines, refrigerators, air conditioners, photocopy machines, and other products where a substantial load (such as a motor or compressor) has to be switched on and off by a control switch, a thermostat, or an electronic circuit.

Figure 9-12 shows a common small-scale application in which a signal from a microcontroller (a few mA at 5VDC) is applied to the base of a transistor, which controls the relay. In this way, a logic output can switch 10A at 125VAC. Note the rectifier diode wired in parallel with the relay coil.

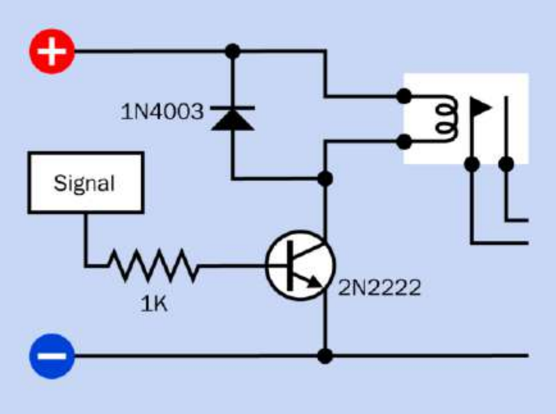

Figure 9-12. *A signal from a digital source such as a microcontroller can switch substantial voltage and current if it is applied to the base of a transistor that activates a relay.*

A latching relay is useful wherever a connection should persist when power is switched off or interrupted, or if power consumption must be minimized. Security devices are one common application. However, the circuit may require a "power reset" function to restore known default settings of latched relays.

A circuit including every possible protection against voltage spikes is shown in Figure 9-13, including a snubber to protect the relay contacts, a rectifier diode to suppress back-EMF generated by the relay coil, and another rectifier diode to protect the relay from EMF generated by a motor when the relay switches it on and off. The snubber can be omitted if the motor draws a relatively low current (below 5A) or if the relay is switching a noninductive load. The diode around the relay coil can be omitted if there are no semiconductors or other components in the circuit that are vulnerable to voltage spikes. However, a spike can affect components in adjacent circuits that appear to be electrically isolated. A severe spike can even be transmitted back into 125VAC house wiring. For information on using a resistor-capacitor combination to form a snubber, see "Snubber" (page 108).

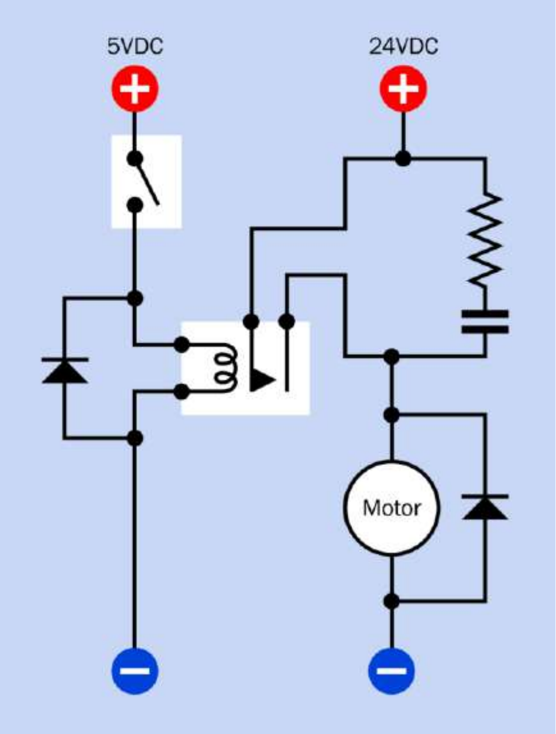

Figure 9-13. *This hypothetical schematic shows three types of protection against voltage spikes induced by an inductive load (a motor, in this instance) and the coil of the relay.*

What Can Go Wrong

Wrong Pinouts

The lack of standardization of relay pinouts can cause errors if one relay is replaced with another that appears to be the same, but isn't. In particular, the pins that connect with normally-closed contacts may trade places with pins connected with normally-open contacts, in relays from different manufacturers.

Pinouts are also confusing in that some datasheets depict them from above, some from below, and some from both perspectives.

Wrong Orientation

Small relays of through-hole type usually have pins spaced in multiples of 0.1". This allows them to be inserted the wrong way around in a perforated board. Almost all relays have an identification mark molded into one end or one corner of the plastic shell. Manufacturers do not standardize the position and meaning of these marks, but they are usually replicated in datasheets. When using a relay of a type that you have not used before, it is a sensible precaution to test it with a meter to verify the functions of its terminals before installing it.

Wrong Type

A latching relay may have exactly the same appearance as a nonlatching relay from the same manufacturer, and the same two pins may energize the coil. However, in a latching relay, the contacts won't spring back to their nonenergized position, causing functional errors that may be difficult to diagnose. The part numbers printed on latching and nonlatching versions of the same relay may differ by only one letter or numeral and should be checked carefully.

Wrong Polarity

A relay with a DC-energized coil may require power to be applied with correct polarity and may malfunction otherwise.

AC and DC

A relay coil designed to be powered by DC will not work from AC and vice-versa. The contact rating of a relay is likely to be different depending whether it is switching AC or DC.

Chatter

This is the noise created by relay contacts when they make rapid intermittent connection. Chatter is potentially damaging to relay contacts and should be avoided. It can also create electrical noise that interferes with other components. Likely correctible causes of chatter include insufficient voltage or power fluctuations.

Relay Coil Voltage Spike

A relay coil is an inductive device. Merely switching a large relay on and off can create voltage spikes. To address this problem, a rectifier diode should be placed across the coil terminals with polarity opposing the energizing voltage.

Arcing

This problem is discussed in the **switch** entry of this encyclopedia. See "Arcing" (page 47). Note that because the contacts inside a *reed relay* are so tiny, they are especially susceptible to arcing and may actually melt and weld themselves together if they are used to control excessive current or an inductive load.

Magnetic Fields

Relays generate magnetic fields during operation and should not be placed near components that are susceptible.

The reed switch inside a reed relay can be unexpectedly activated by an external magnetic field. This type of relay may be enclosed in a metal shell to provide some protection. The adequacy of this protection should be verified by testing the relay under real-world conditions.

Environmental Hazards

Dirt, oxidation, or moisture on relay contacts is a significant problem. Most relays are sealed and should remain sealed.

Relays are susceptible to vibration, which can affect the contacts and can accelerate wear on moving parts. Severe vibration can even damage a relay permanently. **Solid-state relays** (discussed in Volume 2) should be used in harsh environments.

resistor

10

OTHER RELATED COMPONENTS

- **potentiometer** (See Chapter 11)

What It Does

A resistor is one of the most fundamental components in electronics. Its purpose is to impede a flow of current and impose a voltage reduction. It consists of two wires or conductors attached at opposite ends or sides of a relatively poor electrical conductor, the resistance of which is measured in ohms, universally represented by the Greek omega symbol, Ω.

Schematic symbols that represent a resistor are shown in Figure 10-1 (Left: The traditional schematic symbol. Right: The more recent European equivalent). The US symbol is still sometimes used in European schematics, and the European symbol is sometimes used in US schematics. Letters K or M indicate that the value shown for the resistor is in thousands of ohms or millions of ohms, respectively. Where these letters are used in Europe, and sometimes in the US, they are substituted for a decimal point. Thus, a 4.7K resistor may be identified as 4K7, a 3.3M resistor may be identified as 3M3, and so on. (The numeric value in Figure 10-1 was chosen arbitrarily.)

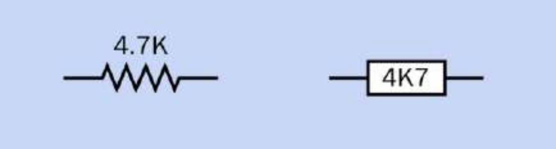

Figure 10-1. *Resistor symbols. The left one is more common in the United States, while the right one is widely used in Europe. The 4.7K value was chosen arbitrarily.*

A resistor is commonly used for purposes such as limiting the charging rate of a **capacitor**; providing appropriate control voltage to semiconductors such as **bipolar transistors**; protecting **LED**s or other semiconductors from excessive current; adjusting or limiting the frequency response in an audio circuit (in conjunction with other components); pulling up or pulling down the voltage at the input pin of a digital logic chip; or controlling a voltage at a point in a circuit. In this last application, two resistors may be placed in series to create a *voltage divider*.

A **potentiometer** may be used instead of a resistor where variable resistance is required.

Sample resistors of various values are shown in Figure 10-2. From top to bottom, their power dissipation ratings are 3W, 1W, 1/2W, 1/4W, 1/4W, 1/4W, and 1/8W. The accuracy (tolerance) of each resistor, from top to bottom, is plus-or-minus 5%, 5%, 5%, 1%, 1%, 5%, and 1%. The beige-colored body of a resistor is often an indication that its tolerance is 5%, while a blue-colored body often indicates a tolerance of 1% or 2%. The blue-bodied resistors and the dark brown resistor contain metal-oxide film elements, while the beige-bodied resistors and the green resistor contain carbon film. For more information on resistor values, see the upcoming *Values* section.

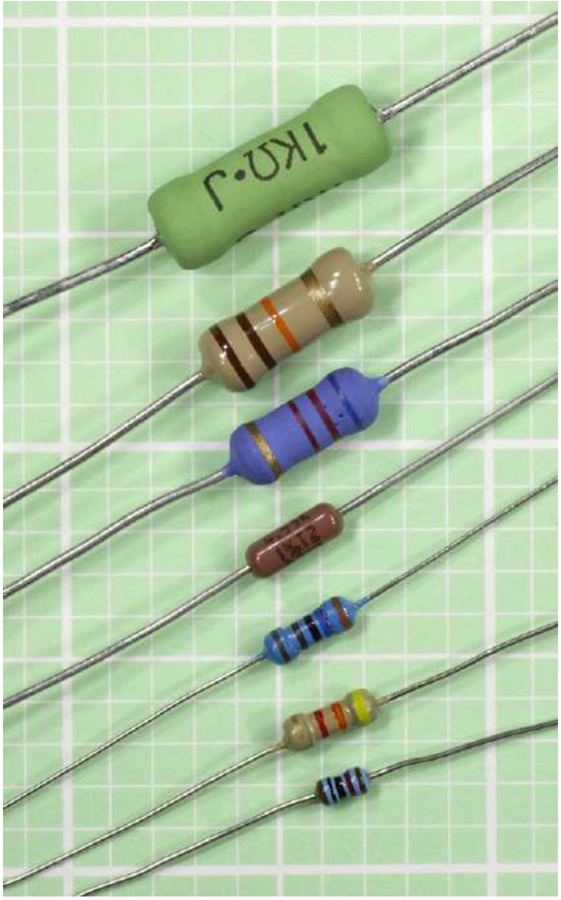

Figure 10-2. *A range of typical resistors. See text for details.*

How It Works

In the process of impeding the flow of current and reducing voltage, a resistor absorbs electrical energy, which it must dissipate as heat. In most modern electronic circuits, the heat dissipation is typically a fraction of a watt.

If R is the resistance in ohms, I is the current flowing through the resistor in amperes, and V is the voltage drop imposed by the resistor (the difference in electrical potential between the two contacts that are attached to it), Ohm's law states:

```
V = I / R
```

This is another way of saying that a resistor of 1Ω will allow a current of 1 amp when the potential difference between the ends of the resistor is 1 volt.

If W is the power in watts dissipated by the resistor, in a DC circuit:

```
W = V * I
```

By substitution in Ohm's law, we can express watts in terms of current and resistance:

$$W = I^2 / R$$

We can also express watts in terms of voltage and resistance:

$$W = V^2 * R$$

These alternates may be useful in situations where you do not know the voltage drop or the current, respectively.

Approximately similar relationships exist when using alternating current, although the power will be a more complex function.

Variants

Axial resistors have two leads that emerge from opposite ends of a usually cylindrical body. *Radial resistors* have parallel leads emerging from one side of the body and are unusual.

Precision resistors are generally defined as having a tolerance of no more than plus-or-minus 1%.

General-purpose resistors are less stable, and their value is less precise.

Power resistors are generally defined as dissipating 1 or 2 watts or more, particularly in power supplies or power amplifiers. They are physically larger and may require *heat sinks* or fan cooling.

Wire-wound resistors are used where the component must withstand substantial heat. A wire-wound resistor often consists of an insulating tube or core that is flat or cylindrical, with multiple turns of resistive wire wrapped around it. The wire is usually a nickel-chromium alloy known as *nichrome* (sometimes written as *Ni-chrome*) and is dipped in a protecting coating.

The heat created by current passing through resistive wire is a potential problem in electronic circuits where temperature must be limited. However, in household appliances such as hair dryers, toaster ovens, and fan heaters, a nichrome element is used specifically to generate heat. Wire-wound resistors are also used in 3D printers to melt plastic (or some other compound) that forms the solid output of the device.

Thick film resistors are sometimes manufactured in a flat, square format. A sample is shown in Figure 10-3, rated to dissipate 10W from its flat surface. The resistance of this component is 1K.

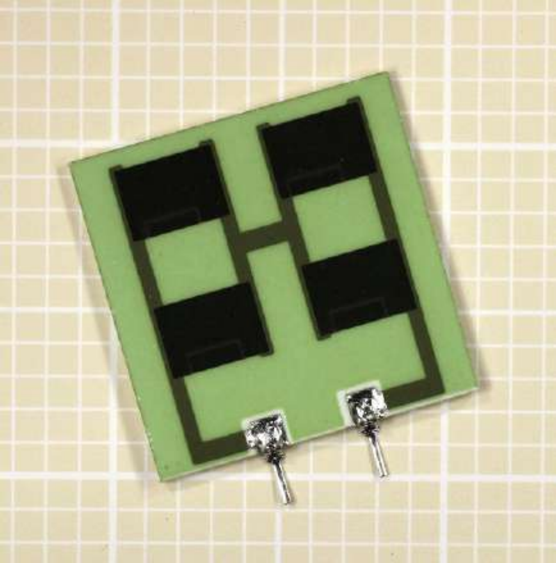

Figure 10-3. *A thick-film resistor measuring about 1" square and 0.03" thick.*

Surface-mount resistors generally consist of a resistive ink film printed on top of a tablet of aluminum oxide ceramic compound, often approximately 6mm long, known as a 2512 form factor. Each surface-mount resistor has two nickel-plated terminations coated in solder, which melts when the resistor is attached to the circuit board. The upper surface is coated, usually with black epoxy, to protect the resistive element.

Resistor Array

This is also known as a *resistor network* or *resistor ladder*, and consists of a chip containing multiple equal-valued resistors.

A resistor array in a *single-inline package* (or *SIP*) may have three possible internal configurations: isolated, common bus, and dual terminator. These options are shown at top, center, and bottom, respectively, in Figure 10-4. The isolated variant is commonly available in SIPs with 6, 8, or 10 pins. The common-bus and dual-terminator configurations generally have 8, 9, 10, or 11 pins.

In the isolated configuration, each resistor is electrically independent of the others and is accessed via its own pair of pins. On a common bus, one end of each resistor shares a bus accessed by a single pin, while the other ends of the resistors are accessed by their own separate pins. A dual-terminator configuration is more complex, consisting of pairs of resistors connected between ground and an internal bus, with the midpoint of each resistor pair accessible via a separate pin. The resistor pairs this function as voltage dividers and are commonly used in emitter-coupled logic circuits that require termination with -2 volts.

A *dual-inline package* (*DIP*) allows a similar range of internal configurations, as shown in Figure 10-5. At top, isolated resistors are commonly available in DIPs with 4, 7, 8, 9, or 10 pins. At center, the common bus configuration is available in DIPs with 8, 14, 16, 18, or 20 pins. At bottom, the dual-terminator configuration usually has 8, 14, 16, 18, or 20 pins.

The external appearance of SIP and DIP resistor arrays is shown in Figure 10-6. From left to right, the packages contain seven 120Ω resistors in

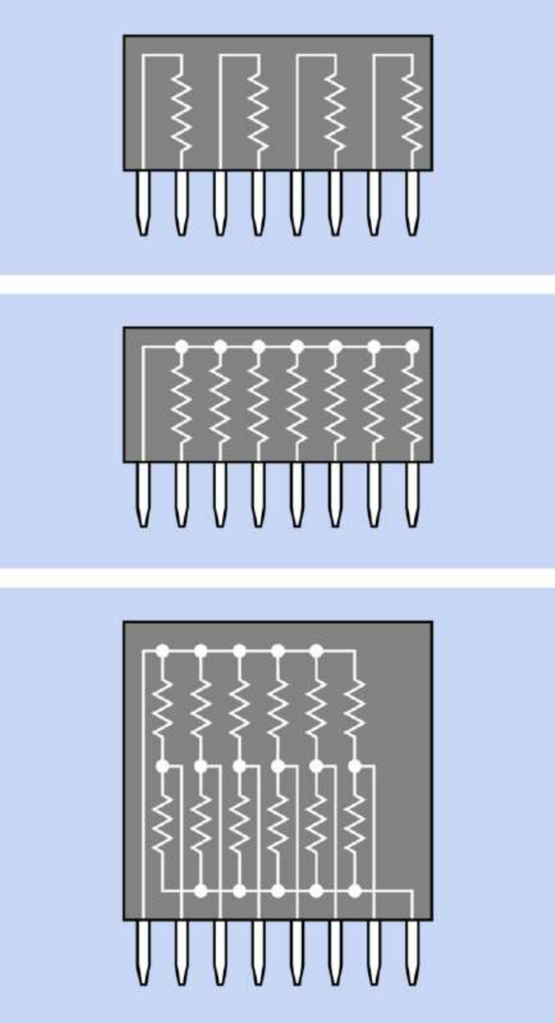

Figure 10-4. *Multiple resistors can be embedded in a single-inline package (SIP) in a variety of formats. See text for additional details.*

isolated configuration; thirteen 120Ω resistors in bussed configuration; seven 5.6K resistors in bussed configuration; and six 1K resistors in bussed configuration.

Resistor arrays with isolated or common-bus configurations are a convenient way to reduce the component count in circuits where pullup, pulldown, or terminating resistors are required for multiple chips. The common-bus configuration is also useful in conjunction with a 7-segment **LED display**, where each segment must be terminated by a series resistor and all the resistors share a common ground or common voltage source.

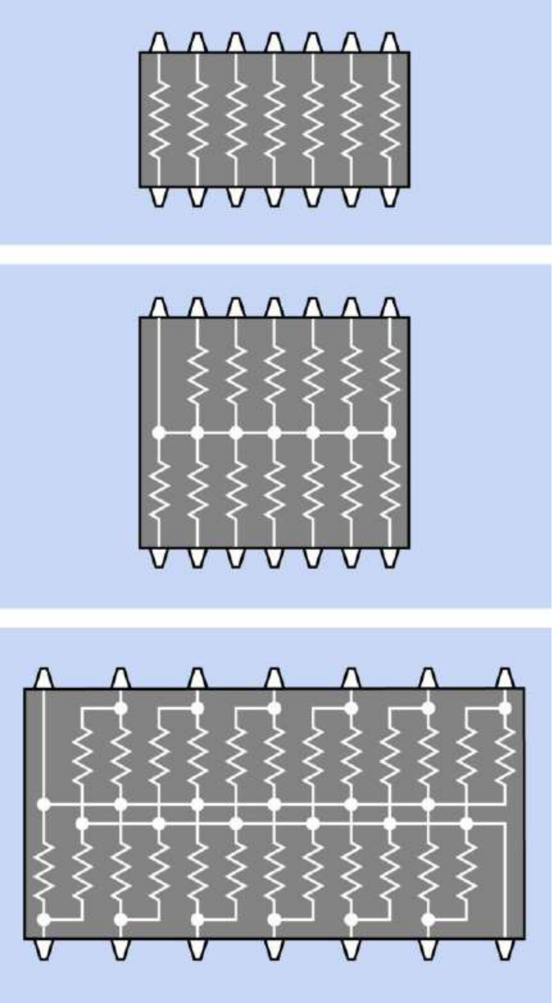

Figure 10-5. *Multiple resistors can be obtained embedded in a dual-inline package (DIP). See text for additional details.*

Surface-mount chips are available containing a pair of resistors configured as a single voltage divider.

Chips containing multiple *RC circuits* (each consisting of a capacitor and a resistor in series) are available, although uncommon. A package containing a single RC circuit may be sold as a

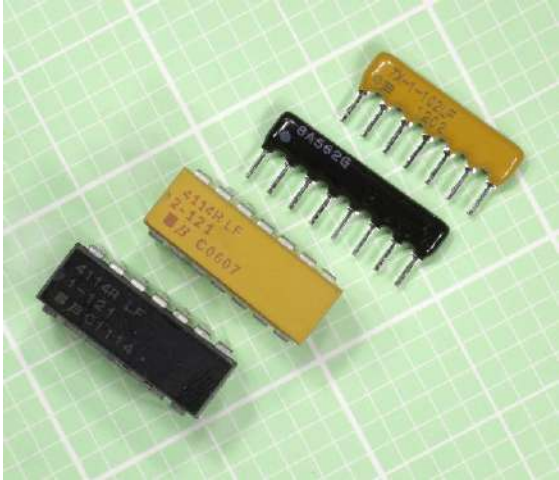

Figure 10-6. *Resistor arrays in DIP and SIP packages. See text for values.*

snubber to protect contacts in a **switch** or **relay** that switches a large inductive load. More information on snubber circuits is in the **capacitor** entry of this encyclopedia; see "Snubber" (page 108).

Values

1 *kilohm,* usually written as 1K, is 1,000Ω. 1 *megohm,* usually written as 1M or 1 meg, is 1,000K. 1 *gigaohm* is 1,000 megs, although the unit is rarely used. Resistances of less than 1Ω are uncommon and are usually expressed as a decimal number followed by the Ω symbol. The term *milliohms* (thousandths of an ohm) is used in special applications. Equivalent resistor values are shown in Figure 10-7.

A resistance value remains unchanged in DC and AC circuits, except where the AC reaches an extremely high frequency.

In common electronics applications, resistances usually range from 100Ω to 10M. Power ratings may vary from 1/16 watt to 1000 watts, but usually range from 1/8 watt to 1/2 watt in most electronic circuits (less in surface-mount applications).

Ohms	Kilohms	Megohms
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 10-7. *Equivalent values in ohms, kilohms, and megohms.*

Tolerance

The tolerance, or precision, of a resistor may range from plus-or-minus 0.001% up to plus-or-minus 20%, but is most commonly plus-or-minus 1%, 2%, 5%, or 10%.

The traditional range of resistor values was established when a tolerance of 20% was the norm. The values were spaced to allow minimum risk of a resistor at one end of its tolerance range having the same value as another resistor at the opposite end of its tolerance range. The values were rounded to 10, 15, 22, 33, 47, 68, and 100, as illustrated in Figure 10-8 where each blue diamond represents the possible range of actual values of a 20% resistor with a theoretical value shown by the white horizontal line at the center of the diamond.

Resistor factors repeat themselves in multiples of 10. Thus, for example, beginning with a resistor of 100Ω, subsequent increasing values will be 150, 220, 330, 470, 680, and 1K, whereas the range of resistors beginning with 1Ω will be 1.5, 2.2, 3.3, 4.7, 6.8, and 10Ω.

Resistance multiplication factors are now expressed as a list of preferred values by the International Electrotechnical Commission (IEC) in their 60063 standard. Intermediate factors have been added to the basic sequence to

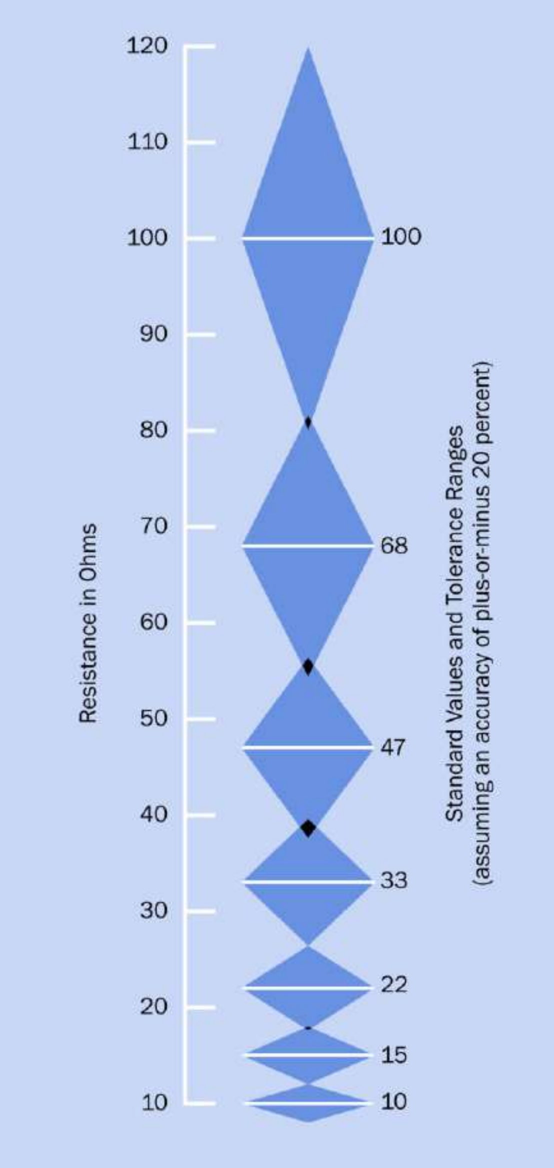

Figure 10-8. *Graphical representation of standard resistor values (white lines) established by the International Electrotechnical Commission, showing the acceptable range of actual values (dark blue areas) assuming precision of plus-or-minus 20%. The overlap, if any, between each range and the next is shown in black.*

accommodate better tolerances. A table showing resistor values for tolerances of plus-or-minus 20%, 10%, and 5% appears in Figure 10-9. Resistors with a tolerance of 5% have become increasingly common.

	Resistor Tolerances (plus or minus)		
	20%	10%	5%
Standard Resistor Values	100	100	100
			91
		82	82
			75
	68	68	68
			62
		56	56
			51
	47	47	47
			43
		39	39
			36
	33	33	33
			30
		27	27
			24
	22	22	22
			20
		18	18
			16
	15	15	15
			13
		12	12
			11
	10	10	10

Figure 10-9. *Standard values for resistors of different precisions. For resistors outside the range shown, values can be found by multiplying or dividing (repeatedly, if necessary) by a factor of 10.*

The IEC has established 3-digit preferred values for resistors with values accurate to plus-or-minus 0.5%.

Because many capacitors still have a tolerance no better than 20%, their values also conform with the old original set of resistance values, although the units are expressed in farads or fractions of a farad. See the **capacitors** entry in this encyclopedia for additional information.

Value Coding

Through-hole axial resistors are traditionally printed with a sequence of three colored bands to express the value of the component, each of the first two bands representing a digit from 0 through 9, while the third band indicates the decimal multiplier (the number of zeroes, from 0 to 9, which should be appended to the digits). A fourth band of silver or gold indicates 10% or 5% tolerance respectively. No fourth band would indicate 20% tolerance, although this has become very rare.

Many resistors now have five color bands, to enable the representation of intermediate or fractional values. In this scheme, the first three bands have numeric values (using the same color system as before) while the fourth band is the multiplier. A fifth band, at the opposite end of the resistor, indicates its tolerance.

In Figure 10-10 the numeric or multiplier value of each color is shown as a "spectrum" at the top of the figure. The tolerance, or precision of a resistor, expressed as a plus-or-minus percentage, is shown using silver, gold, and various colors, at the bottom of the figure.

Two sample resistors are shown. The upper one has a value of 1K, indicated by the brown and black bands on the left (representing numeral 1 followed by a numeral 0) and the third red band (indicating two additional zeroes). The gold band at right indicates a precision of 5%. The lower one has a value of 1.05K, indicated by the brown, black, and green bands on the left (representing numeral 1 followed by numeral 0 followed by a numeral 5) and the fourth band brown (indicating one additional zero). The brown band at right indicates a precision of 1%.

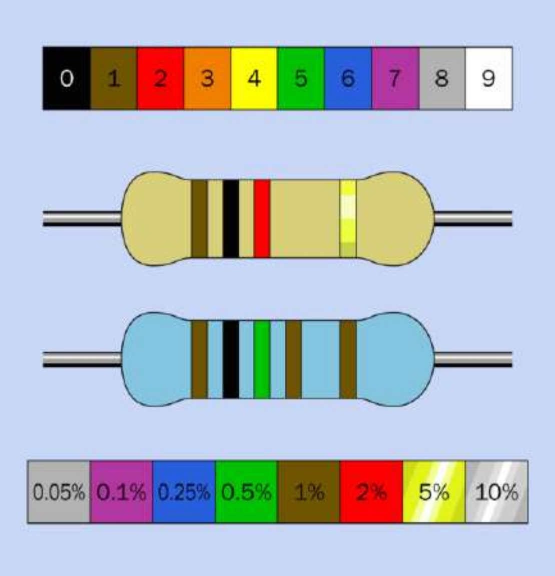

Figure 10-10. *Color coding of through-hole resistors. See text for details.*

In extremely old equipment, resistors may be coded with the body-tip-dot scheme, in which the body color represents the initial digit, the end color represents the second digit, and a dot represents the multiplier. The numeric identities of the colors is the same as in the current color scheme.

In all modern schemes, the three or four bands that show the resistance value are spaced close together, while a larger gap separates them from the band that shows the tolerance. The resistor value should be read while holding the resistor so that the group of closely-spaced numeric bands is on the left.

Confusingly, some resistors may be found where the first three bands define the value, using the old three-band convention; the fourth band indicates tolerance; and a fifth band at the opposite end of the component indicates reliability. However, this color scheme is uncommon.

Other color-coding conventions may be found in special applications, such as military equipment.

It is common for through-hole carbon-film resistors to have a beige body color, while through-hole metal-film resistors often have a blue body color. However, in relatively rare instances, a blue body color may also indicate a *fusible* resistor (designed to burn out harmlessly like a fuse, if it is overloaded) while a white body may indicate a *non-flammable* resistor. Use caution when replacing these special types.

Some modern resistors may have their values printed on them numerically. Surface-mount resistors also have digits printed on them, but they are a code, not a direct representation of resistance. The last digit indicates the number of zeroes in the resistor value, while the preceding two or three numbers define the value itself. Letter R is used to indicate a decimal point. Thus a 3R3 surface-mount resistor has a value of 3.3Ω, while 330 would indicate 33Ω, and 332 indicates 3,300Ω. A 2152 surface-mount resistor would have a value of 21,500Ω.

A surface-mount resistor with a single zero printed on it is a *zero ohm* component that has the same function as a **jumper** wire. It is used for convenience, as it is easily inserted by automated production-line equipment. It functions merely as a bridge between traces on the circuit board.

When resistor values are printed on paper in schematics, poor reproduction may result in omission of decimal points or introduction of specks that look like decimal points. Europeans have addressed this issue by using the letter as a substitute for a decimal point so that a 5.6K resistor will be shown as 5K6, or a 3.3M resistor will be shown as 3M3. This practice is followed infrequently in the United States.

Stability

This term describes the ability of a resistor to maintain an accurate value despite factors such as temperature, humidity, vibration, load cycling, current, and voltage. The *temperature coefficient* of a resistor (often referred to as T_{cr} or T_c, not to be confused with the time constant of a charging capacitor) is expressed in parts per million change in resistance for each degree centigrade deviation from room temperature (usually assumed to be 25 degrees Centigrade). T_c may be a positive or a negative value.

The *voltage coefficient* of resistance—often expressed as V_c—describes the change of a resistor's value that may occur as a function of changes in voltage. This is usually significant only where the resistive element is carbon-based. If V1 is the rated voltage of the resistor, R1 is its rated resistance at that voltage, V2 is 10% of the rated voltage, and R2 is the actual resistance at that voltage, the voltage coefficient, V_c, is given by this formula:

$$V_c = (100 * (R1 - R2)) / (R2 * (V1 - V2))$$

Materials

Resistors are formed from a variety of materials.

Carbon composite. Particles of carbon are mixed with a binder. The density of the carbon determines the end-to-end resistance, which typically ranges from 5Ω to 10M. The disadvantages of this system are low precision (a 10% tolerance is common), relatively high voltage coefficient of resistance, and introduction of noise in sensitive circuits. However, carbon-composite resistors have low inductance and are relatively tolerant of overload conditions.

Carbon film. A cheap and popular type, made by coating a ceramic substrate with a film of carbon compound. They are available in both through-hole and surface-mount formats. The range of resistor values is comparable with carbon-composite types, but the precision is increased, typically to 5%, by cutting a spiral groove in the carbon-compound coating during the manufacturing process. The carbon film suffers the same disadvantages of carbon composite resistors, but to a lesser extent. Carbon film resistors generally should not be substituted for metal film resistors in applications where accuracy is important.

Metal film. A metallic film is deposited on a ceramic substrate, and has generally superior characteristics to carbon-film resistors. During manufacture, a groove may be cut in the metal film to adjust the end-to-end resistance. This may cause the resistor to have higher inductance than carbon-composite types, though it has lower noise. Tolerances of 5%, 2%, and 1% are available. This type of resistor was originally more expensive than carbon-film equivalents, but the difference is now fractional. They are available in both through-hole and surface-mount formats. They are available in lower-wattage variants (1/8 watt is common).

Thick-film resistors are spray-coated, whereas thin-film resistors are sputtered nichrome. Thin-films enjoy a flatter temperature coefficient and are typically used in environments that have a huge operational temperature range, such as satellites.

Bulk metal foil. The type of foil used in metal film resistors is applied to a ceramic wafer and etched to achieve the desired overall resistance. Typically these resistors have axial leads. They can be extremely accurate and stable, but have a limited maximum resistance.

Precision wire-wound. Formerly used in applications requiring great accuracy, but now largely replaced by precision metal foil.

Power wire-wound. Generally used when 1 or 2 watts or more power dissipation is required. Resistive wire is wrapped around a core that is often ceramic. This can cause the resistor to be referred to, inaccurately, as "ceramic." The core may alternatively be fiberglass or some other electrically insulating compound that actively sinks heat. The component is either dipped (typically in vitreous enamel or cement) or is mounted in an aluminum shell that can be clamped to a heat sink. It almost always has the ohm value printed on it in plain numerals (not codes).

Two typical wire-wound resistors are shown in Figure 10-11. The upper resistor is rated at 12W and 180Ω while the lower resistor is rated at 13W and 15K.

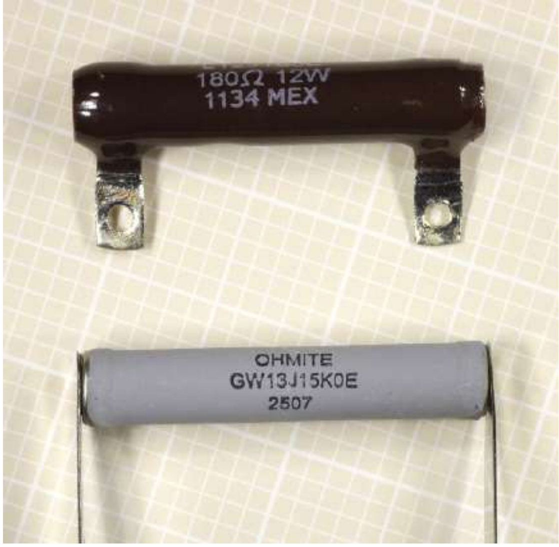

Figure 10-11. *Two wire-wound resistors of greatly differing resistance but similar power dissipation capability.*

A larger wire-wound resistor is shown in Figure 10-12, rated for 25W and 10Ω.

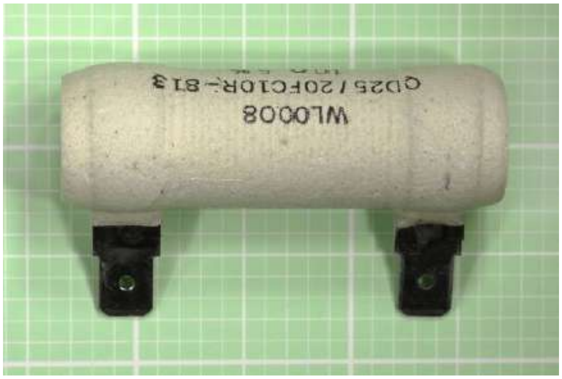

Figure 10-12. *A large wirewound resistor rated to dissipate 25W.*

In Figure 10-13, two resistors encapsulated in cement coatings are shown with the coatings removed to expose the elements. At left is a 1.5Ω 5W resistor, which uses a wire-wound element. At right is a very low-value 0.03Ω 10W resistor.

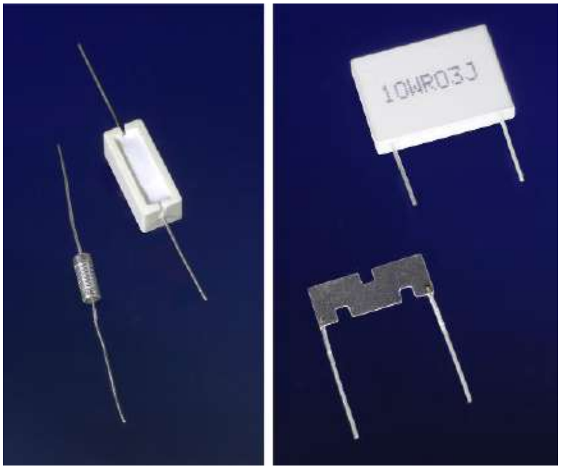

Figure 10-13. *Two low-value resistors with their cement coatings removed to show the resistive elements.*

In Figure 10-14, the resistor at right has an exposed 30Ω element while the resistor at left is rated 10W and 6.5Ω, enclosed in an anodized aluminum shell to promote heat dissipation.

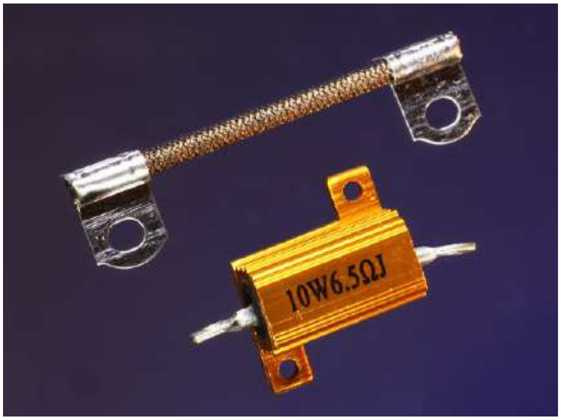

Figure 10-14. *A 30Ω resistor (right) and 6.5Ω resistor (left).*

In power resistors, heat dissipation becomes an important consideration. If other factors (such as voltage) remain the same, a lower-value resistor will tend to pass more current than a higher-value resistor, and heat dissipation is proportional to the square of the current. Therefore power wire-wound resistors are more likely to be needed where low resistance values are required. Their coiled-wire format creates significant inductance, making them unsuitable to pass high frequencies or pulses.

How to Use it

Some of the most common applications for a resistor are listed here.

In Series with LED

To protect an LED from damage caused by excessive current, a *series resistor* is chosen to allow a current that does not exceed the manufacturer's specification. In the case of a single through-hole LED (often referred to as an *indicator*), the forward current is often limited to around 20mA, and the value of the resistor will depend on the voltage being used. (See Figure 10-15.)

When using high-output LEDs (which may contain multiple elements in a single 5mm or 10mm package), or LED arrays that are now being used for domestic lighting, the acceptable current may be much greater, and the LED unit may contain its own current-limiting electronics. A datasheet should be consulted for details.

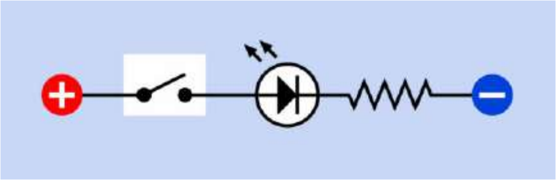

Figure 10-15. *A series resistor is necessary to limit the current that passes through an LED.*

Current Limiting with a Transistor

In Figure 10-16, a transistor is switching or amplifying current flowing from B to C. A resistor is used to protect the base of the transistor from excessive current flowing from point A. Resistors are also commonly used to prevent excessive current from flowing between B and C.

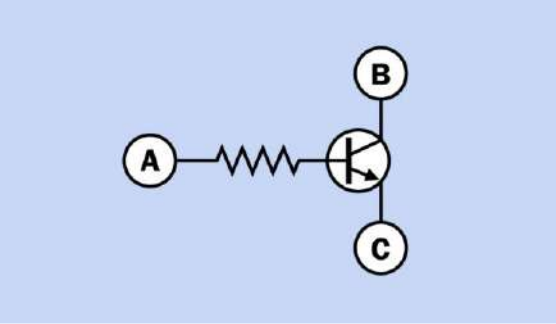

Figure 10-16. *A resistor is typically necessary to protect the base of a transistor from excessive current.*

Pullup and Pulldown Resistors

When a mechanical switch or pushbutton is attached to the input of a logic chip or **microcontroller**, a *pullup* or *pulldown* resistor is used, applying positive voltage or grounding the pin, respectively, to prevent it from "floating" in an indeterminate state when the switch is open. In Figure 10-17, the upper schematic shows a pulldown resistor, whereas the lower schematic shows a pullup resistor. A common value for either of them is 10K. When the pushbutton is pressed, its direct connection to positive voltage or to ground easily overwhelms the effect of the resistor. The choice of pullup or pulldown resistor may depend on the type of chip being used.

Audio Tone Control

A resistor-capacitor combination can limit the high-frequency in a simple audio tone-control circuit, as shown in Figure 10-18. Beneath a signal travelling from A to B, a resistor is placed in series with a capacitor that passes high frequencies to ground. This is known as a low-pass filter.

RC Network

A resistor will adjust the charge/discharge time when placed in series with a capacitor, as in Figure 10-19. When the switch closes, the resistor limits the rate at which the capacitor will charge itself from the power supply. Because a capacitor has an ideally infinite resistance to DC current, the voltage measured at point A will rise until it

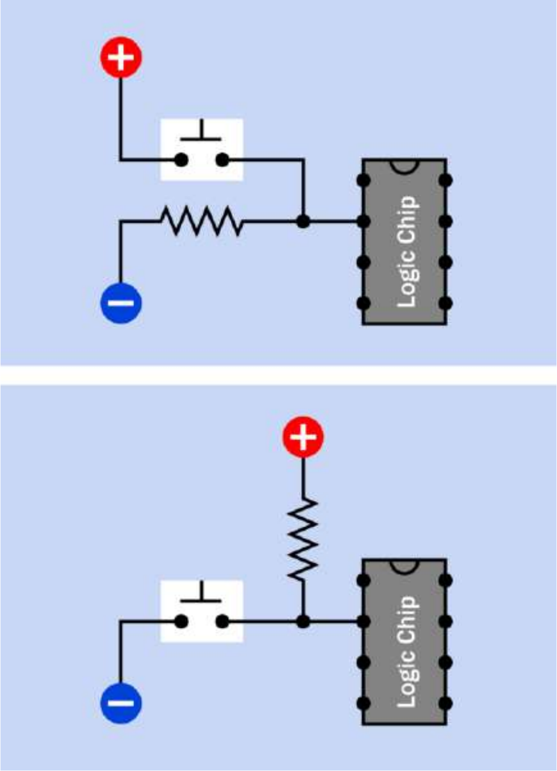

Figure 10-17. *A pulldown resistor (top) or pullup resistor (bottom) prevents an input pin on a logic chip or microcontroller from "floating" in an indeterminate state when the button is not being pressed.*

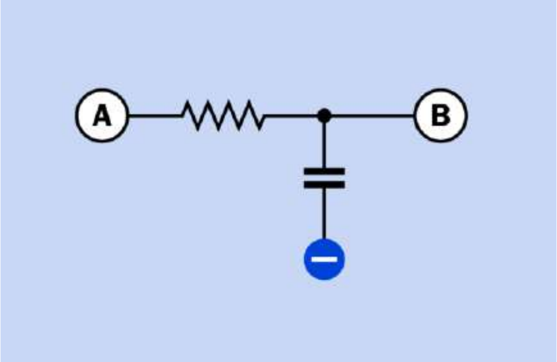

Figure 10-18. *This configuration may be used to remove high frequencies from an audio signal. It is known as a low-pass filter because low frequencies are passed from A to B.*

is close to the supply voltage. This is often referred to as an *RC* (*resistor-capacitor*) network and is discussed in greater detail in the **capacitor** section of this encyclopedia.

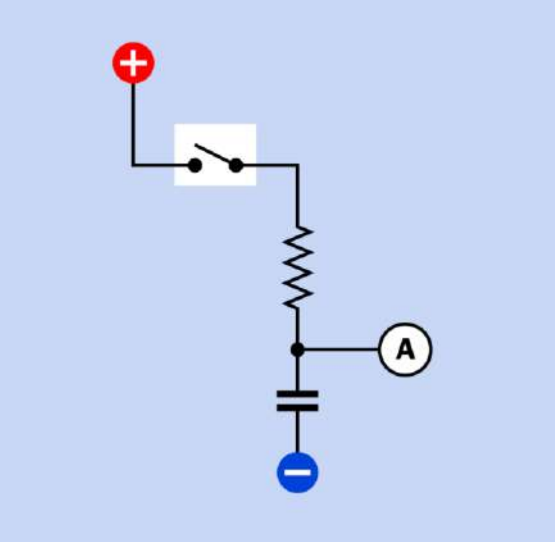

Figure 10-19. *In an RC (resistor-capacitor) network, a resistor limits the rate of increase in potential of the capacitor, measured at A, when the switch is closed.*

Voltage Divider

Two resistors may be used to create a voltage divider (see Figure 10-20). If V_{in} is the supply voltage, the output voltage, V_{out}, measured at point A, is found by the formula:

$$V_{out} = V_{in} * (R2 / (R1 + R2))$$

In reality, the actual value of V_{out} is likely to be affected by how heavily the output is loaded.

If the output node has a high impedance, such as the input to a logic chip or comparator, it will be more susceptible to electrical noise, and lower-value resistors may be needed in the voltage divider to maintain a higher current flow and maintain stability in the attached device.

Resistors in Series

If resistors in series have values R1, R2, R3 . . . the total resistance, R, is found by summing the individual resistances:

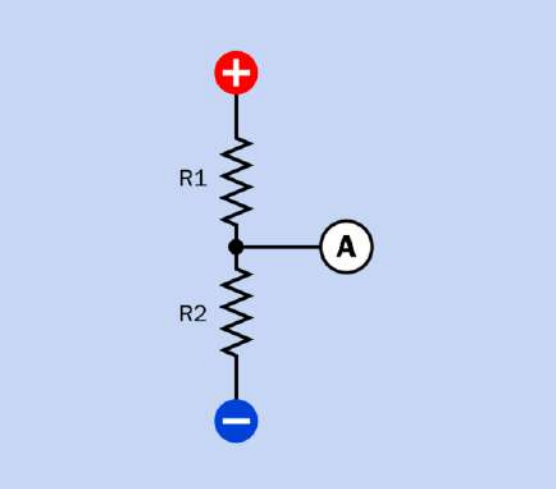

Figure 10-20. *In a DC circuit, a pair of resistors may be placed in series to function as a voltage divider. The voltage measured at A will be lower than the supply voltage, but above ground potential.*

```
R = R1 + R2 + R3. . .
```

The current through each of the resistors will be the same, whereas the voltage across each of them will vary proportionately with its resistance. If the supply voltage across the series of resistors is VS, and the total of all the resistor values is RT, and the resistance of one resistor is R1, the voltage across that resistor, V1, will be given by the formula:

```
V1 = VS * (R1 / RT)
```

Resistors in Parallel

Where two or more resistors (R1, R2, R3 . . .) are wired in parallel, their total resistance, R, is found from the formula:

```
1/R = ( 1/R1 ) + ( 1/R2 ) + ( 1/R3 ). . .
```

Suppose that R1, R2, R3 . . . all have the same individual resistance, represented by RI, and the number of resistors is N. Their total resistance, RT, when wired in parallel, will be:

```
RT = RI / N
```

If each resistor has an equal resistance and also has an equal individual rating in watts (represented by WI), the total wattage (WT) that they can handle when wired in parallel to share the power will be:

```
WT = WI * N
```

Therefore, if an application requires high-wattage resistors, multiple lower-wattage, higher-value resistors may be substituted if they are wired in parallel—and may even be cheaper than a single high-wattage wire-wound resistor. For example, if a 5W, 50Ω resistor is specified, 10 resistors can be substituted, each rated at 0.5W and 500Ω. Bear in mind that if they are tightly bundled, this will interfere with heat dissipation.

What Can Go Wrong

Heat

Resistors are probably the most robust of all electronic components, with high reliability and a long life. It is difficult to damage a resistor by overheating it with a soldering iron.

The wattage rating of a resistor does not necessarily mean that it should be used to dissipate that amount of power on a constant basis. Small resistors (1/4 watt or less) can overheat just as easily as big ones. Generally speaking, it is safe practice not to exceed 75% of a resistor’s power rating on a constant basis.

Overheating is predictably more of a problem for power resistors, where provision must be made for heat dissipation. Issues such as component crowding should be considered when deciding how big a heat sink to use and how much ventilation. Some power resistors may function reliably at temperatures as high as 250 degrees Centigrade, but components near them are likely to be less tolerant and plastic enclosures may soften or melt.

Noise

The electrical *noise* introduced by a resistor in a circuit will vary according to the composition of the resistor, but for any given component, it will be proportional to voltage and current. Low-noise circuits (such as those at the input stage of a high-gain amplifier) should use low-wattage resistors at a low voltage where possible.

Inductance

The coiled wire of a wire-wound resistor will be significantly inductive at low frequencies. This is known as *parasitic inductance*. It will also have a resonant frequency. This type of resistor is unsuitable for applications where frequency exceeds 50KHz.

Inaccuracy

When using resistors with 10% tolerance, imprecise values may cause greater problems in some applications than in others. In a voltage divider, for instance, if one resistor happens to be at the high end of its tolerance range while the other happens to be at the low end, the voltage obtained at the intersection of the resistors will vary from its expected value. Using the schematic shown in Figure 10-20, if R1 is rated for 1K and R2 is rated for 5K, and the power supply is rated at 12VDC, the voltage at point A should be:

```
V = 12 * (( 5 / (5 + 1)) = 10
```

However, if R1 has an actual value of 1.1K and R2 has an actual value of 4.5K, the actual voltage obtained at point A will be:

```
V = 12 * (( 4.5 / (4.5 + 1.1)) = 9.6
```

If the resistors are at opposite ends of their respective tolerance ranges, so that R1 has an actual value of 900Ω while the lower resistor has an actual value of 5.5K, the actual voltage obtained will be:

```
V = 12 * (( 5.5 / (5.5 + 0.9)) = 10.3
```

The situation becomes worse if the two resistors are chosen to be of equal value, to provide half of the supply voltage (6 volts, in this example) at their intersection. If two 5K resistors are used, and the upper one is actually 4.5K while the lower one is 5.5K, the actual voltage will be:

```
V = 12 * (( 5.5 / (4.5 + 5.5)) = 6.6
```

Whether this variation is significant will depend on the particular circuit in which the voltage divider is being used.

Common through-hole resistors may occasionally turn out to have values that are outside their specified tolerance range, as a result of poor manufacturing processes. Checking each resistor with a meter before placing it in a circuit should be a standard procedure.

When measuring the voltage drop introduced by a resistor in an active circuit, the meter has its own internal resistance that will take a proportion of the current. This is known as *meter loading* and will result in an artificially low reading for the potential difference between the ends of a resistor. This problem becomes significant only when dealing with resistors that have a high value (such as 1M), comparable with the internal resistance of the meter (likely to be 10M or more).

Wrong Values

When resistors are sorted into small bins by the user, errors may be made, and different values may be mixed together. This is another reason for checking the values of components before using them. Identification errors may be nontrivial and easily overlooked: the visible difference between a 1 megohm resistor and a 100Ω resistor is just one thin color band.

potentiometer

11

Also known as a *variable resistor*; may be substituted for a *rheostat*.

OTHER RELATED COMPONENTS

- **rotational encoder** (see Chapter 8)
- **resistor** (see Chapter 10)

What It Does

When a voltage is applied across a potentiometer, it can deliver a variable fraction of that voltage. It is often used to adjust sensitivity, balance, input, or output, especially in audio equipment and sensors such as motion detectors.

A potentiometer can also be used to insert a variable resistance in a circuit, in which case it should really be referred to as a *variable resistor,* although most people will still call it a potentiometer.

It can be used to adjust the power supplied to a circuit, in which case it is properly known as a *rheostat*, although this term is becoming obsolete. Massive rheostats were once used for purposes such as dimming theatrical lighting, but solid-state components have taken their place in most high-wattage applications.

A full-size, classic-style potentiometer is shown in Figure 11-1.

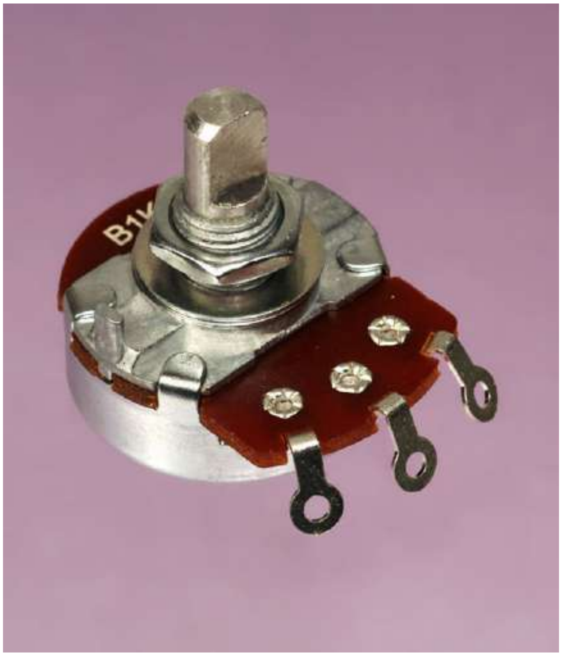

Figure 11-1. *A generic or classic-style potentiometer, approximately one inch in diameter.*

Schematic symbols for a potentiometer and other associated components are shown in Figure 11-2, with American versions on the left and European versions on the right in each case. The symbols for a potentiometer are at the top. The correct symbols for a variable resistor or rheostat are shown at center, although a potentiometer symbol may often be used instead. A preset variable resistor is shown at the bottom, often referred to as a *trimmer* or *Trimpot*. In these examples, each has an arbitrary rated resistance of 4,700Ω. Note the European substitution of K for a decimal point.

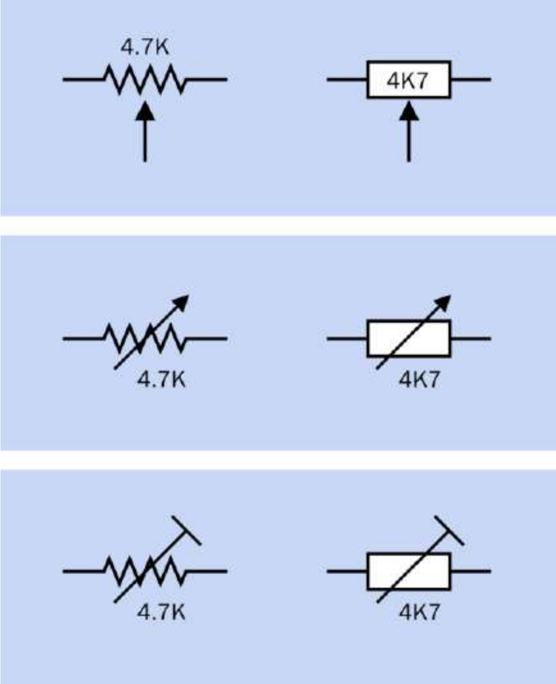

Figure 11-2. *American (left) and European (right) symbols for a potentiometer, a rheostat, and a trimmer potentiometer, reading from top to bottom. The 4.7K value was chosen arbitrarily.*

How It Works

A potentiometer has three terminals. The outer pair connect with the opposite ends of an internal resistive element, such as a strip of conductive plastic, sometimes known as the *track*. The third center terminal connects internally with a contact known as the *wiper* (or rarely, the *pick-off*), which touches the strip and can be moved from one end of it to the other by turning a shaft or screw, or by moving a slider.

If an electrical potential is applied between opposite ends of the resistive element, the voltage "picked off" by the wiper will vary as it moves. In this mode, the potentiometer works as a resistive voltage divider. For example, in a potentiometer with a linear taper (see "Variants," coming up), if you attach the negative side of a 12V battery to the right-hand end terminal and the positive side to the left-hand end terminal, you will find an 8V potential at the center terminal when the potentiometer has rotated clockwise through one-third of its range. In Figure 11-3, the base of the shaft (shown in black) is attached to an arm (shown in green) that moves a wiper (orange) along a resistive element (brown). The voltages shown assume that the resistive element has a linear taper and will vary slightly depending on wire resistance and other factors.

Because a potentiometer imposes a voltage reduction, it also reduces current flowing through it, and therefore creates waste heat which must be dissipated. In an application such as an audio circuit, small currents and low voltages generate negligible heat. If a potentiometer is used for heavier applications, it must be appropriately rated to handle the wattage and must be vented to allow heat to disperse.

To use a potentiometer as a *variable resistor* or *rheostat*, one of its end terminals may be tied to the center terminal. If the unused end terminal is left unconnected, this raises the risk of picking up stray voltages or "noise" in sensitive circuits. In Figure 11-4, a potentiometer is shown adjusting a series resistance for an *LED* for demonstration purposes. More typically, a trimmer would be used in this kind of application, since a user is unlikely to need to reset it.

Variants

Linear and Log Taper

If the resistive element in a potentiometer is of constant width and thickness, the electrical potential at the wiper will change in ratio with the rotation of the wiper and shaft (or with movement of a slider). This type of potentiometer is said to have a linear taper even though its element does not actually taper.

For audio applications, because human hearing responds nonlinearly to sound pressure, a potentiometer that has a linear taper may seem to have a very slow action at one end of its scale and an abrupt effect at the other. This problem used to be solved with a non-uniform or tapered re-

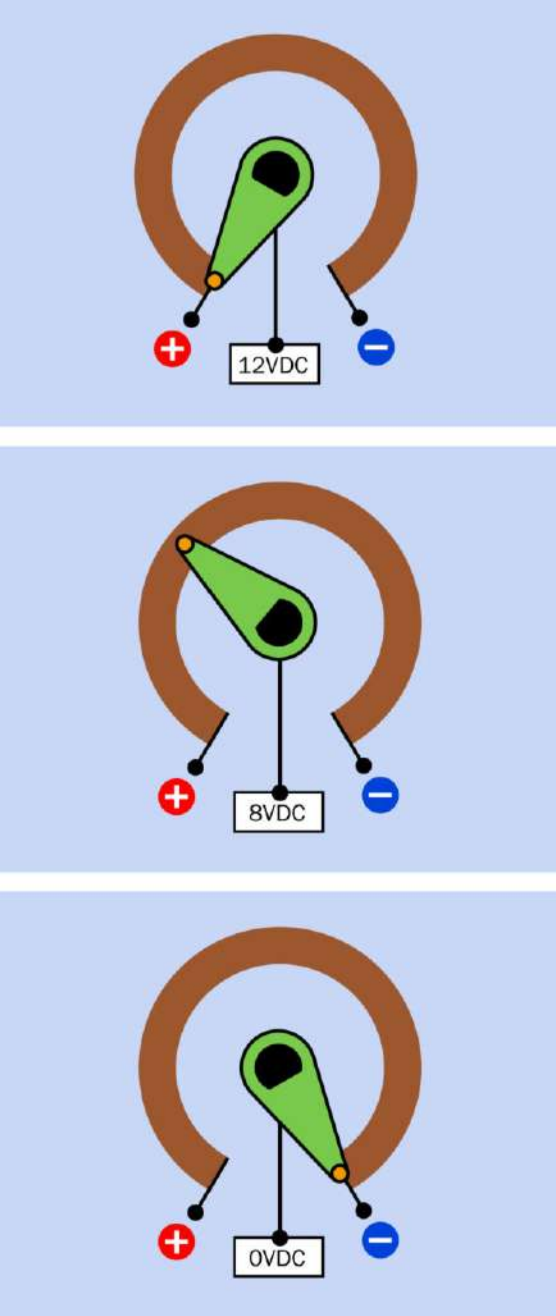

Figure 11-3. *Inside a potentiometer. See text for details.*

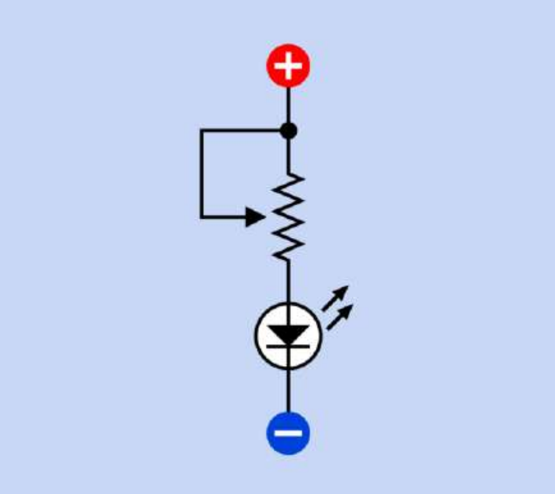

Figure 11-4. *A potentiometer can be used to adjust a series resistance, as shown in this schematic. Tying the wiper to one of the end terminals reduces the risk of picking up electrical noise.*

sistive element. More recently, a combination of resistive elements has been used as a cheaper option. Such a potentiometer is said to have an audio taper or a log taper (since the resistance may vary as a logarithm of the angle of rotation). A reverse audio taper or antilog taper varies in the opposite direction, but this type has become very uncommon.

Classic-style Potentiometer

This consists of a sealed circular can, usually between 0.5″ and 1″ in diameter, containing a resistive strip that is shaped as a segment of a circle. A typical example is shown in Figure 11-1, although miniaturized versions have become more common. A shaft mounted on the can turns the internal wiper that presses against the strip. For panel-mount applications, a threaded bushing at the base of the shaft is inserted through a hole in the front panel of the electronics enclosure, and a nut is tightened on the bushing to hold the potentiometer in place. Often there is also a small offset index pin that, when paired with a corresponding front panel hole, will keep the pot from spinning freely.

Many modern potentiometers are miniaturized, and may be packaged in a box-shaped plastic enclosure rather than a circular can. Their power ratings are likely to be lower, but their principle of operation is unchanged. Two variants are shown in Figure 11-5.

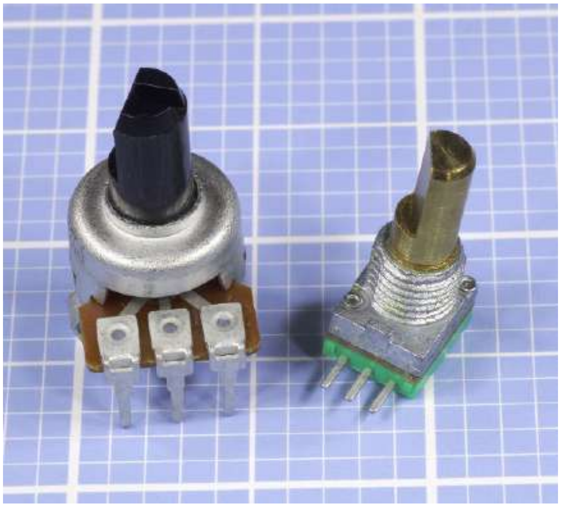

Figure 11-5. *Two modern miniaturized potentiometers. At left: 5K. At right: 10K. Both are rated to dissipate up to 50mW.*

The three terminals on the outside of a potentiometer may be solder lugs, screw terminals, or pins for direct mounting on a circuit board. The pins may be straight or angled at 90 degrees.

The resistive element may use carbon film, plastic, cermet (a ceramic-metal mixture), or resistive wire wound around an insulator. Carbon-film potentiometers are generally the cheapest, whereas wire-wound potentiometers are generally the most expensive.

Wire-wound potentiometers may handle more power than the other variants, but as the wiper makes a transition from one turn of the internal wire element to the next, the output will tend to change in discrete steps instead of varying more smoothly.

In a potentiometer with detents, typically a spring-loaded lever in contact with notched internal wheel causes the shaft to turn in discrete steps that create a stepped output even if the resistive element is continuous.

The shaft may be made of metal or plastic, with its length and width varying from one component to another. A control knob can be fitted to the end of the shaft. Some control knobs are push-on, others have a set screw to secure them. Shafts may be splined and split, or round and smooth, or round with a flat surface that matches the shape of a socket in a control knob and reduces the risk of a knob becoming loose and turning freely. Some shafts have a slotted tip to enable screwdriver adjustment.

Some shaft options for full-size potentiometers are shown in Figure 11-6.

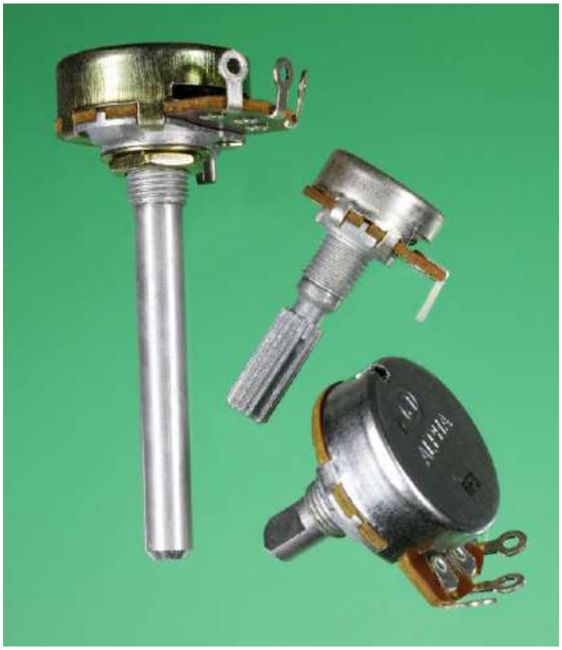

Figure 11-6. *Three shaft options for potentiometers.*

Multiple-Turn Potentiometer

To achieve greater precision, a track inside a potentiometer may be manufactured in the form of a helix, allowing the wiper to make multiple turns on its journey from one end of the track to the other. Such multiple-turn potentiometers typically allow 3, 5, or 10 turns to move the wiper from end to end. Other multiple-turn potentiometers may use a screw thread that advances a wiper along a linear or circular track. The latter is comparable with a trimmer where multiple turns of a screwdriver are used to rotate a worm gear that rotates a wiper between opposite ends of a circular track.

Ganged Potentiometer

Two (or rarely, more) potentiometers can be stacked or combined so that their resistive elements and wipers share the same shaft but can use different voltages or have different taper. Each resistance-wiper assembly is known as a *cup*, and the potentiometers are said to be *ganged*.

Flat ganged potentiometers combine two resistive elements in one enclosure. Some *dual ganged* potentiometers are concentric, meaning that the pots are controlled separately by two shafts, one inside the other. Suitable concentric knobs must be used. You are unlikely to find these potentiometers sold as components in limited quantities.

Switched Potentiometer

In this variant, when the shaft is turned clockwise from an initial position that is fully counterclockwise, it flips an internal switch connected to external terminals. This can be used to power-up associated components (for example, an audio amplifier). Alternatively, a switch inside a potentiometer may be configured so that it is activated by pulling or pushing the shaft.

Slider Potentiometer

Also known as a *slide potentiometer*. This uses a straight resistive strip and a wiper that is moved to and fro linearly by a tab or lug fitted with a plastic knob or finger-grip. Sliders are still found on some audio equipment. The principle of operation, and the number of terminals, are identical to the classic-style potentiometer. Sliders typically have solder tabs or PC pins. In Figure 11-7, the large one is about 3.5” long, designed for mounting behind a panel that has a slot to allow the sliding lug to poke through. Threaded holes at either end will accept screws to fix the slider behind the panel. A removable plastic finger-grip (sold separately, in a variety of styles) has been pushed into place. Solder tabs underneath the slider are hidden in this photo. The smaller slider is designed for through-hole mounting on a circuit board.

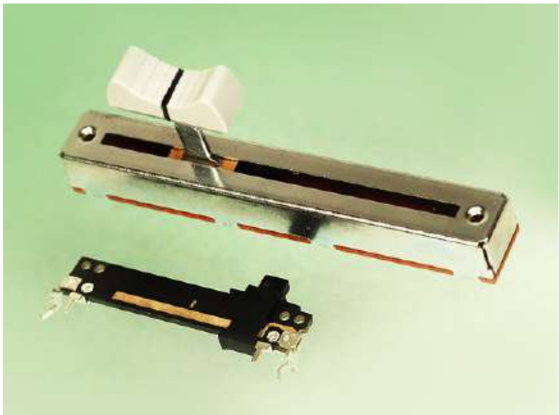

Figure 11-7. *Slider potentiometers.*

Trimmer Potentiometer

Often referred to as Trimpots, this is actually a proprietary brand name of Bourns. They are usually mounted directly on circuit boards to allow fine adjustment or trimming during manufacturing and testing to compensate for variations in other components. Trimmers may be single-turn or multi-turn, the latter containing a worm gear that engages with another gear to which the wiper is attached. Trimmers always have linear taper. They may be designed for screwdriver adjustment or may have a small knurled shaft, a thumb wheel, or a knob. They are not usually accessible by the end user of the equipment, and their setting may be sealed or fixed when the equipment is assembled. In Figure 11-8, the beige Spectrol trimmer is a single-turn design, whereas the blue trimmer is multi-turn. A worm gear inside the package, beneath the screw head, engages with an interior gear wheel that rotates the wiper.

In Figure 11-9, a 2K trimmer potentiometer has a knurled dial attached to allow easy finger adjustment, although the dial also contains a slot for a flat-blade screwdriver.

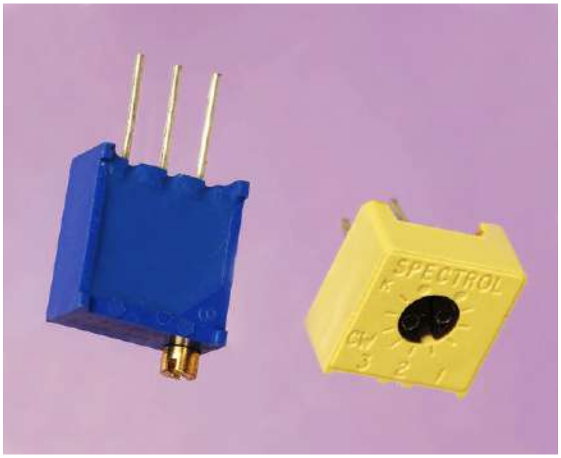

Figure 11-8. *Like most trimmers, these are designed for through-hole mounting on a circuit board.*

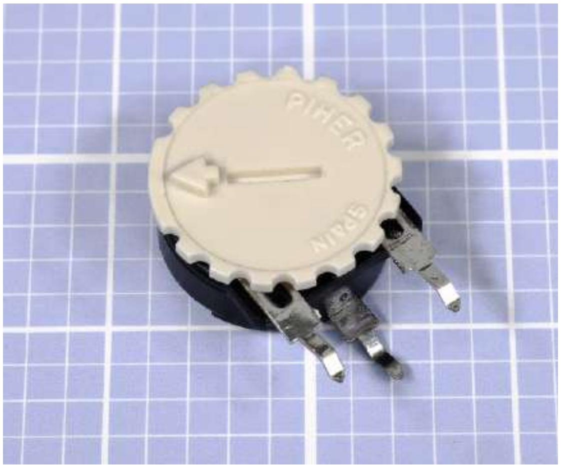

Figure 11-9. *A trimmer potentiometer with a knurled dial to facilitate finger adjustment.*

How to Use it

The classic-style potentiometer was once used universally to control volume, bass, and treble on audio equipment but has been replaced increasingly by digital input devices such as **tactile switches** (see "Tactile Switch" (page 34)) or **rotational encoders** (see Chapter 8), which are more reliable and may be cheaper, especially when assembly costs are considered.

Potentiometers are widely used in lamp dimmers and on cooking stoves (see Figure 11-10). In these applications, a solid-state switching device such as a **triac** (described in Volume 2) does the actual work of moderating the power to the lamp or the stove by interrupting it very rapidly. The potentiometer adjusts the duty cycle of the power interruptions. This system wastes far less power than if the potentiometer controlled the lighting or heating element directly as a rheostat. Since less power is involved, the potentiometer can be small and cheap, and will not generate significant heat.

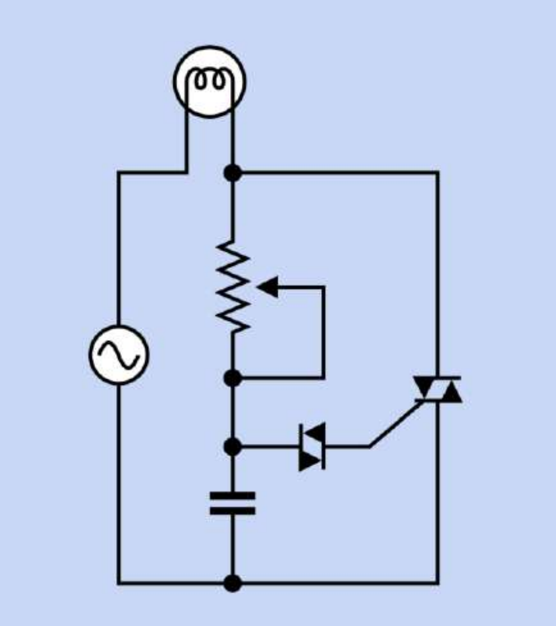

Figure 11-10. *Typical usage of a potentiometer in conjunction with a diac, triac, and capacitor to control the brightness of an incandescent bulb, using an AC power supply. Diacs and triacs are discussed in Volume 2.*

Because true logarithmic potentiometers have become decreasingly common, a linear potentiometer in conjunction with a fixed resistor can be used as a substitute, to control audio input. See Figure 11-11.

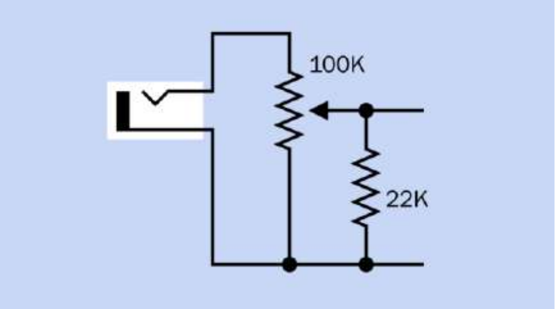

Figure 11-11. *In this circuit, a 100K linear potentiometer is used in conjunction with a 22K resistor to create an approximately logarithmic volume control for an audio system with input coming from a mono jack socket at left.*

A potentiometer may be used to match a sensor or analog input device to an analog-digital converter, or it can calibrate a device such as a temperature or motion sensor.

What Can Go Wrong

Wear and Tear

Since classic-style potentiometers are electromechanical devices, their performance will deteriorate as one part rubs against another. The long open slot of a slider potentiometer makes it especially vulnerable to contamination with dirt, water, or grease. Contact-cleaner solvent, lubricant-carrying sprays, or pressurized "duster" gas may be squirted into a potentiometer to try to extend its life. Carbon-film potentiometers are the least durable and in audio applications will eventually create a "scratchy" sound when they are turned, as the resistive element deteriorates.

If the wiper deteriorates to the point where it no longer makes electrical contact with the track, and if the potentiometer is being used as a variable resistor, two failure modes are possible, shown in Figure 11-12. Clearly the right-hand schematic is a better outcome. This is an argument for always tying the wiper to the "unused" end of the track.

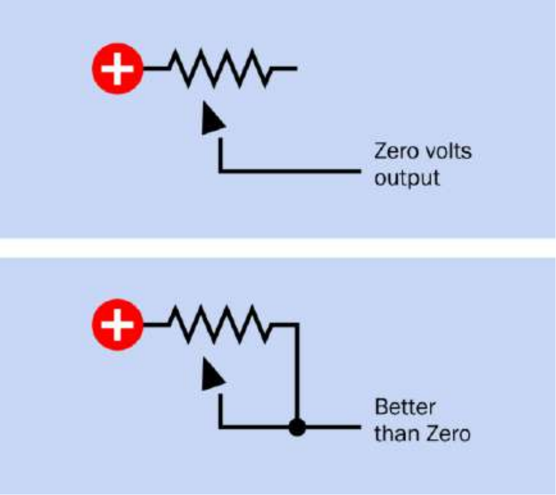

Figure 11-12. *If the wiper of a potentiometer breaks (indicated by the loose arrow head) as a result of wear and tear, and the potentiometer is being used as a variable resistor, the voltage from it will drop to zero (top schematic) unless the wiper has been tied to one end of the track (bottom schematic).*

If you are designing a circuit board that will go through a production process, temperature variations during wave soldering, and subsequent washing to remove flux residues, create hostile conditions for potentiometers, especially sliders where the internal parts are easily contaminated. It will be safer to hand-mount potentiometers after the automated process.

Knobs that Don't Fit

Control knobs are almost always sold separately from potentiometers. Make sure the shaft of the potentiometer (which may be round, round-with-flat, or knurled) matches the knob of your choice. Note that some shaft diameters are expressed in inches, while others are metric.

Nuts that Get Lost

For panel-mounted potentiometers, a nut that fits the thread on the bushing is almost always included with the potentiometer; an additional nut and lock washer may also be supplied. Because there is no standardization of threads on potentiometers, if you lose a nut, you may have some difficulty finding an exact replacement.

A Shaft that Isn't Long Enough

When choosing a shaft length, if in doubt, buy a potentiometer with a long shaft that you can cut to the desired length.

Sliders with No Finger Grip

Slider potentiometers are often sold without a knob or plastic finger-grip, which must be ordered separately and may be available in different styles. The finger-grip usually push-fits onto the metal or plastic tab or lug that moves the slider to and fro.

Too Big to Fit

Check the manufacturer's datasheet if you need to know the physical size of the potentiometer. Photographs may be misleading, as a traditional-style potentiometer that is 0.5" in diameter looks much the same as one that is 1" in diameter. High-wattage potentiometers will be more costly and physically large (2 to 3 inches in diameter). See Figure 11-13.

Overheating

Be sure to leave sufficient air space around a high-wattage potentiometer. Carefully calculate the maximum voltage drop and current that you may be using, and choose a component that is appropriately rated. Note that if you use the potentiometer as a rheostat, it will have to handle more current when its wiper moves to reduce its resistance. For example, if 12VDC are applied through a 10-ohm rheostat to a component that has a resistance of 20 ohms, current in the circuit will vary from 0.4 amps to 0.6 amps depending on the position of the rheostat. At its maximum setting, the rheostat will impose a 4V voltage drop and will therefore dissipate 1.6 watts from the full length of its resistive element. If the rheostat is reset to impose only a 4-ohm resistance, the voltage drop that it imposes will be 2V, the current in the circuit will be 0.5 amps, and the rheostat will therefore dissipate 1 watt from 4/10ths of the length of its resistive element. A wire-wound potentiometer will be better able to handle high dissipation from a short segment of its element than other types of rheostat. Add a fixed resistor in series with a rheostat if necessary to impose a limit on the current.

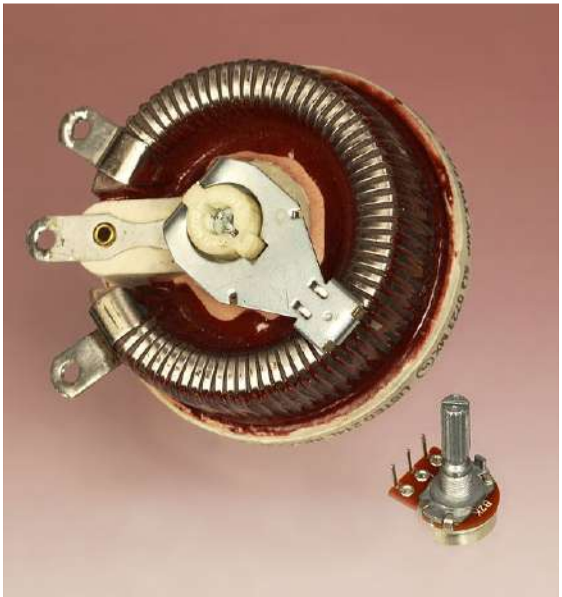

Figure 11-13. *The large potentiometer is approximately 3" in diameter, rated at 5 ohms, and able to handle more than 4 amps. The small potentiometer is 5/8" diameter, rated at 2K and 1/4 watt, with pins designed for through-hole insertion in a circuit board, and a grooved shaft that accepts a push-on knob. Despite the disparity in size, the principle of operation and the basic features are identical.*

When using a trimmer potentiometer, limit the current through the wiper to 100mA as an absolute maximum value.

The Wrong Taper

When buying a potentiometer, remember to check the specification to find out whether it has linear or audio/log taper. If necessary, attach a meter, with the potentiometer set to its center position, to verify which kind of taper you have. While holding the meter probes in place, rotate the potentiometer shaft to determine which way an audio/log taper is oriented.

capacitor

12

Quite often referred to as a *cap*. Formerly known (primarily in the United Kingdom) as a *condenser*, but that term has become obsolete.

OTHER RELATED COMPONENTS

- **variable capacitor** (See Chapter 13)
- **battery** (See Chapter 2)

What It Does

A capacitor connected across a DC power source will accumulate a charge, which then persists after the source is disconnected. In this way, the capacitor stores (and can then discharge) energy like a small rechargeable battery. The charge/discharge rate is extremely fast but can be limited by a series resistor, which enables the capacitor to be used as a timing component in many electronic circuits.

A capacitor can also be used to block DC current while it passes pulses, or electrical "noise," or alternating current, or audio signals, or other wave forms. This capability enables it to smooth the output voltage provided by **power supplies**; to remove spikes from signals that would otherwise tend to cause spurious triggering of components in digital circuits; to adjust the frequency response of an audio circuit; or to couple separate components or circuit elements that must be protected from transmission of DC current.

Schematic symbols for capacitors are shown in Figure 12-1. At top-left is a nonpolarized capacitor, while the other two indicate that a polarized capacitor must be used, and must be oriented as shown. The variant at the bottom is most commonly used in Europe. Confusingly, the nonpolarized symbol may also be used to identify a polarized capacitor, if a + sign is added. The polarized symbols are sometimes printed without + signs, but the symbols still indicate that polarity must be observed.

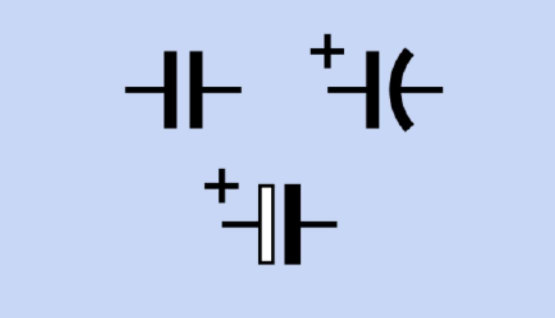

Figure 12-1. *Schematic symbols for polarized and nonpolarized capacitors. See text for details.*

How It Works

In its simplest form, a capacitor consists of two plates, each with a lead attached to it for connection with a DC power source. The plates are separated by a thin, insulating layer known as the *dielectric*, which is usually a solid or a paste but may be liquid, gel, gaseous, or vacuum.

The plates in most capacitors are made from thin metal film or metallized plastic film. To minimize the size of the component, the film may be rolled up to form a compact cylindrical package, or multiple flat sections may be interleaved.

Electrons from the power source will migrate onto the plate attached to the negative side of the source, and will tend to repel electrons from the other plate. This may be thought of as creating *electron holes* in the other plate or as attracting *positive charges*, as shown in Figure 12-2. When the capacitor is disconnected from the power source, the opposite charges on its plates will persist in equilibrium as a result of their mutual attraction, although the voltage will gradually dissipate as a result of *leakage,* either through the dielectric or via other pathways.

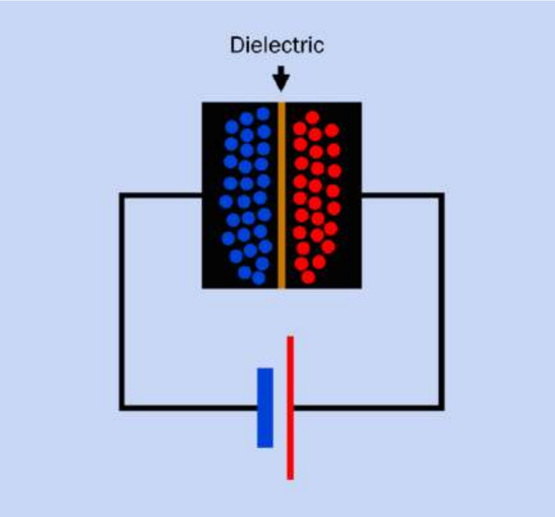

Figure 12-2. *Because the plates of a capacitor are electrically conductive, they will become populated with positive and negative charges when connected with a DC power source. As opposite charges attract each other, they will tend to congregate on either side of the dielectric, which is an insulating layer. The battery symbol is shown here colored for clarity.*

When a **resistor** is placed across the leads of a charged capacitor, the capacitor will discharge itself through the resistor at a rate limited by the resistance. Conversely, if a capacitor is charged through a resistor, the resistor will limit the charging rate.

A resistor in series with a capacitor is known as an *RC network* (Resistor-Capacitor network). In Figure 12-3, an RC circuit is shown with a SPDT switch that charges or discharges the capacitor via a series resistor. The voltage at point A increases nonlinearly (relative to the negative side of the power supply) while the capacitor is charging, and decreases nonlinearly while the capacitor is discharging, as suggested by the graphs. At any moment, the time that the capacitor takes to acquire 63% of the difference between its current charge and the voltage being supplied to it is known as the *time constant* for the circuit. See "The Time Constant" (page 105) for additional information.

When a capacitor is connected across an AC voltage source, each surge of electrons to one plate induces an equal and opposite positive surge to the other plate, and when polarity of the power supply reverses, the charges on the plates switch places. These surges may make it seem that the capacitor is conducting AC current, even though the dielectric that separates the plates is an insulator. See Figure 12-4. Often a capacitor is said to "pass" AC, even though this is not really happening. For convenience, and because the concept is widely established, this encyclopedia refers to capacitors as "passing" AC.

Depending on the size of the capacitor, it will block some AC frequencies while passing others. Generally speaking, a smaller capacitor will pass high frequencies relatively efficiently, as each little surge of current fills each plate. However, the situation is complicated by the *inductive reactance* (which creates the *effective series resistance*) of a capacitor, as discussed below. See "Alternating Current and Capacitive Reactance" (page 106).

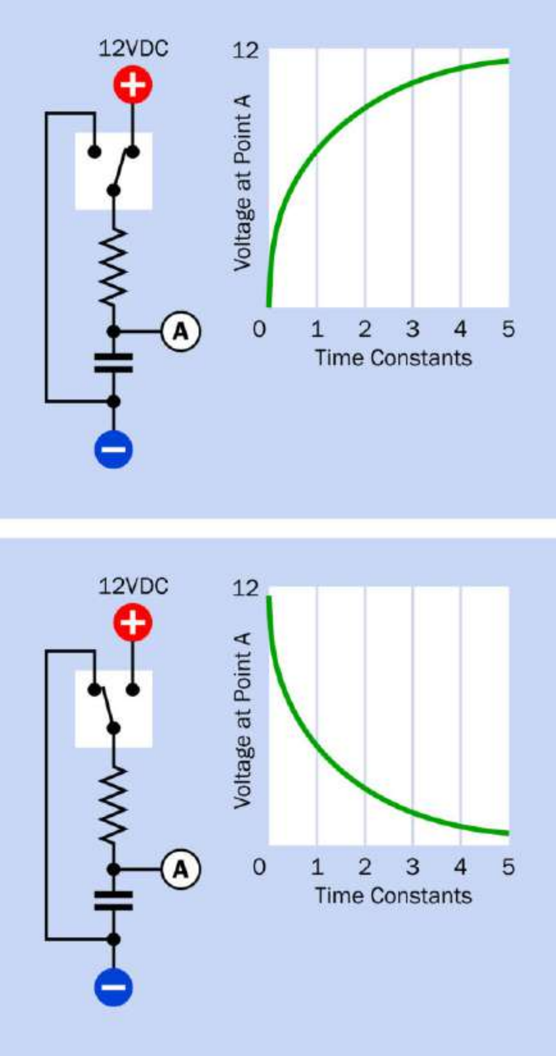

Figure 12-3. *An RC (Resistor-Capacitor) network with a switch to control charge and discharge of a capacitor. At top, the curve gives an approximate idea of the charging behavior of the capacitor. At bottom, the curve illustrates its discharging behavior.*

Variants

Format

The three most common packages for capacitors are cylindrical, disc, and rectangular tablet.

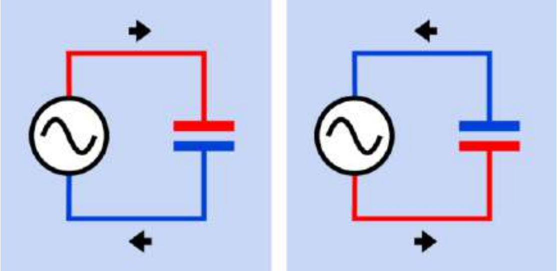

Figure 12-4. *In the left diagram, a source of alternating current charges the upper plate of a capacitor positively and the lower plate negatively. This process entails a flow of conventional current shown by the arrows. A moment later, when the AC current flow reverses, the flow also reverses, creating the impression that the capacitor "passes" AC current.*

A *cylindrical capacitor* may have *axial leads* (a wire attached to each end) or *radial leads* (both wires emerging from one end). Radial capacitors are more widely used as they allow easy insertion into a circuit board. The capacitor is usually packaged in a small aluminum can, closed at one end, capped with an insulating disc at the other end, and wrapped in a thin layer of insulating plastic. Some samples are shown in Figure 12-5 and Figure 12-6.

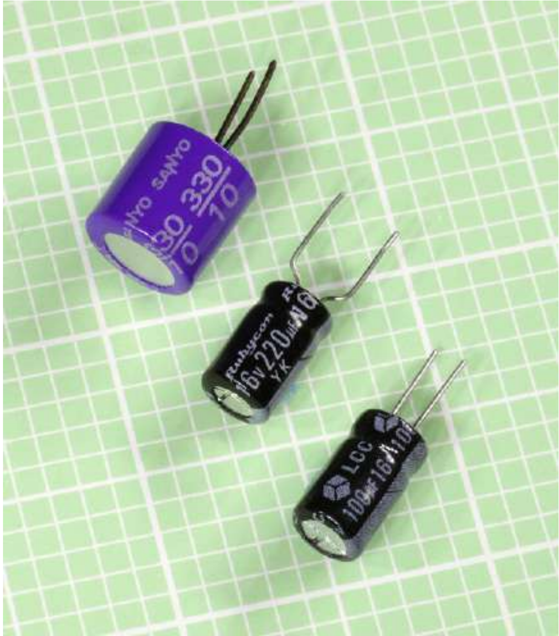

Figure 12-5. *Cylindrical capacitors with radial leads. All are electrolytic.*

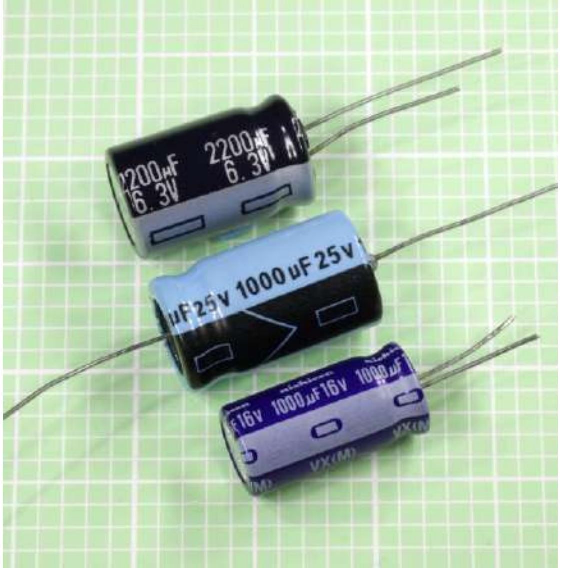

Figure 12-6. *Cylindrical capacitors with radial leads (top and bottom) and axial leads (center). All are electrolytic.*

A *disc capacitor* (sometimes referred to as a *button capacitor*) is usually encased in an insulating ceramic compound, and has radial leads. Modern small-value ceramic capacitors are more likely to be dipped in epoxy, or to be square tablets. Some samples are shown in Figure 12-7.

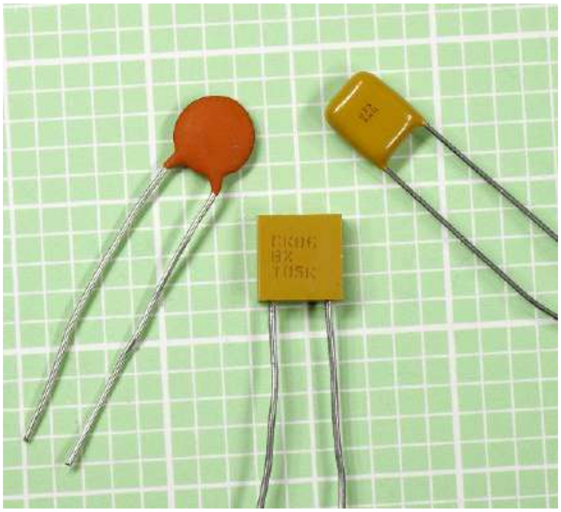

Figure 12-7. *Generic ceramic capacitors. Left: rated for 0.1µF at 50V. Center: 1µF at 50V. Right: 1µF at 50V.*

A *surface-mount capacitor* is square or rectangular, usually a few millimeters in each dimension, with two conductive pads or contacts at opposite ends. It appears almost identical to a surface-mount resistor. Larger-value capacitors are inevitably bigger but can still be designed for surface-mount applications. See Figure 12-8.

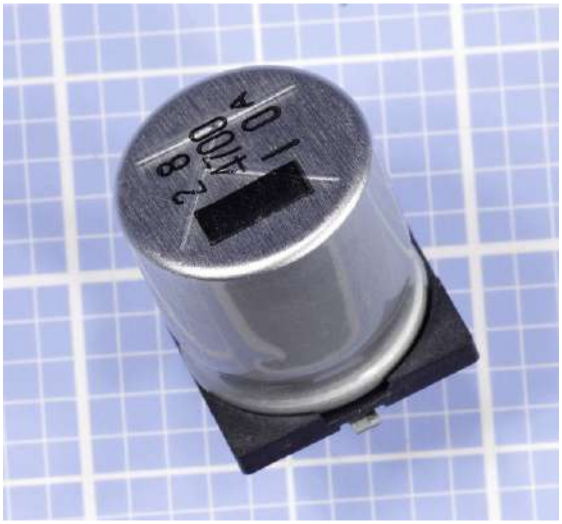

Figure 12-8. *Most surface-mount capacitors are as tiny as other surface-mount components, but this 4,700µF electrolytic (at 10V) has a base approximately 0.6" square. A solder tab is visible at the center of the nearest edge.*

Many capacitors are nonpolarized, meaning that they are insensitive to polarity. However, electrolytic and tantalum capacitors must be connected "the right way around" to any DC voltage source. If one lead is longer than the other, it must be the "more positive" lead. A mark or band at one end of the capacitor indicates the "more negative" end. Tantalum capacitors are likely to indicate the positive lead by using a + sign on the body of the component.

An arrow printed on the side of a capacitor usually points to the "more negative" terminal. In an aluminum can with axial leads, the lead at one end will have an insulating disc around it while the other lead will be integral with the rounded end of the can. The wire at the insulated end must be "more positive" than the wire at the other end.

A *capacitor array* contains two or more capacitors that are isolated from each other internally and accessed by external contacts. They are sold in surface-mount format and also in *through-hole* chips of DIP (dual-inline package) or SIP (single-inline package) format. The internal components may be connected in one of three configurations: isolated, common-bus, or dual-ended common bus. Technically the isolated configuration should be referred to as a capacitor array, but in practice, all three configurations are usually referred to as *capacitor networks*. See Figure 12-9 and Figure 12-10.

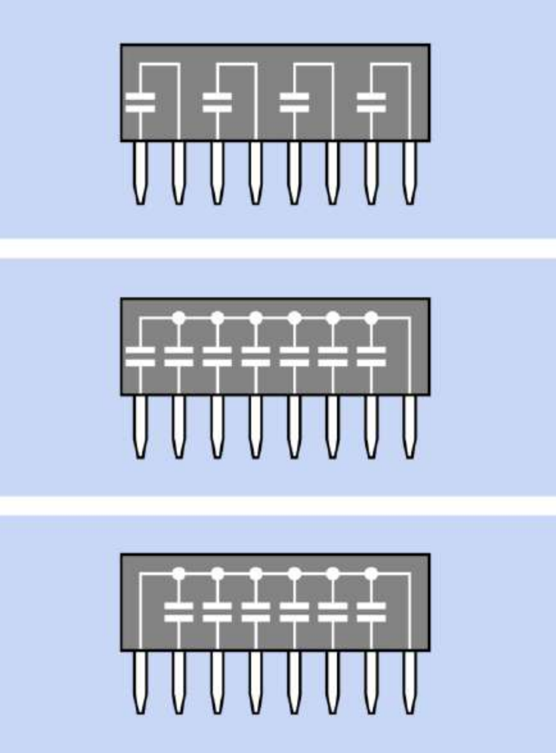

Figure 12-9. *A capacitor network most often consists of a single-inline package (SIP) chip containing multiple capacitors in one of three configurations shown here. Top: Isolated. Center: Common bus. Bottom: Dual-ended common bus. Individual capacitor values ranging from 0.001μF to 0.1μF are common.*

Capacitor networks can reduce the component count in circuits where digital logic chips require bypass capacitors. They are comparable in concept to *resistor arrays*.

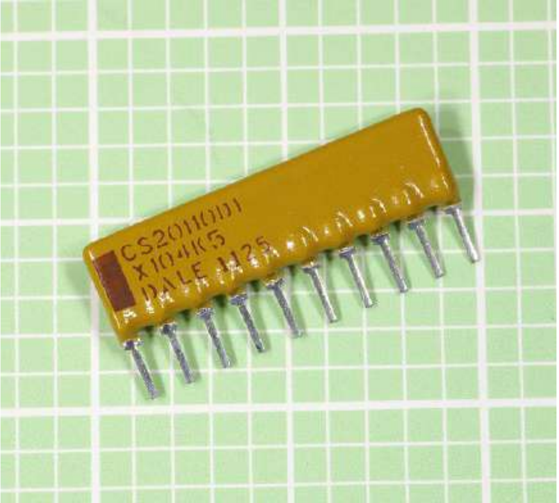

Figure 12-10. *A capacitor array in through-hole, SIP format.*

Chips containing *RC circuits* (multiple resistor-capacitor pairs) are available, although uncommon.

Principal Types

Electrolytic capacitors are relatively cheap, compact, and available in large values. These attributes have made them a popular choice in consumer electronics, especially for power supplies. The capacitive capability of an electrolytic is refreshed by periodic application of voltage. A moist paste inside the capacitor is intended to improve the dielectric performance when voltage is applied, but can dry out during a period of years. If an electrolytic is stored for 10 years or so, it may allow a short circuit between its leads when power is applied to it. The capacitors in Figure 12-5 and Figure 12-6 are all electrolytic. The capacitor in Figure 12-11 is at the high end of the scale.

A *bipolar electrolytic* is a single package containing two electrolytic capacitors in series, end-to-end, with opposed polarities, so that the combination can be used where the voltage of a signal fluctuates above and below 0VDC. See Figure 12-12 and Figure 12-13. This type of component is likely to have “BP” (bipolar) or “NP”

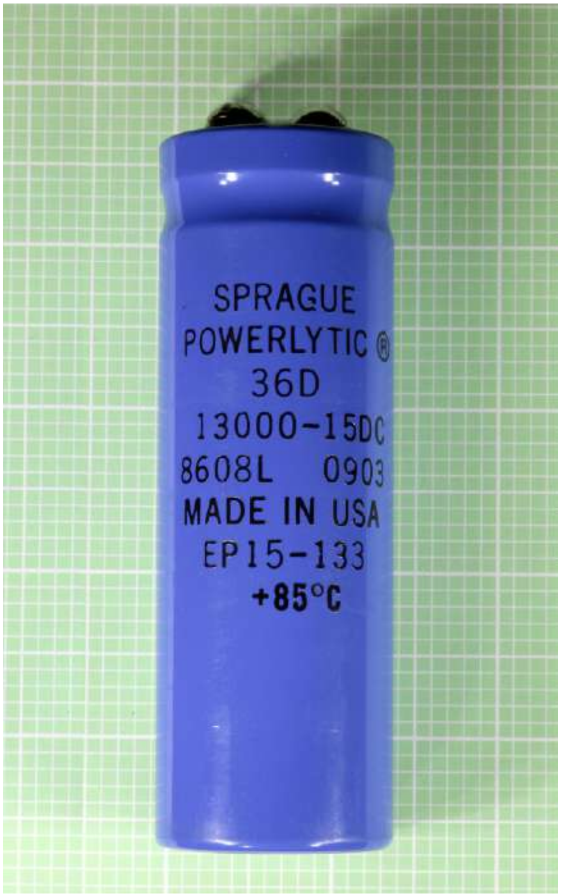

Figure 12-11. *This 13,000µF electrolytic capacitor is larger than would be required in most everyday applications.*

(nonpolarized) printed on its shell. It may be used in audio circuits where polarized capacitors are normally unsuitable, and is likely to be cheaper than non-electrolytic alternatives. However, it suffers from the same weaknesses as all electrolytics.

Tantalum capacitors are compact but relatively expensive, and can be vulnerable to voltage spikes. They are sensitive to application of the wrong polarity. Typically they are epoxy-dipped rather than mounted inside a small aluminum can like electrolytics, and consequently the electrolyte may be less likely to evaporate and dry out. In Figure 12-14, two tantalum capacitors (rated 330µF at 6.3V, left, and 100µF at 20V, right) are shown above a polyester film capacitor (rated

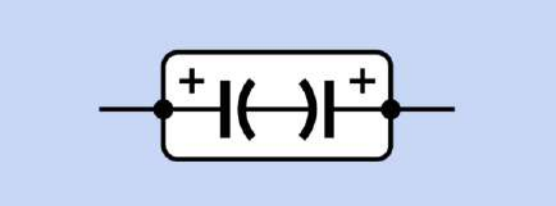

Figure 12-12. *Schematic view of the internal configuration of a bipolar electrolytic capacitor, also known as nonpolarized electrolytic capacitor. It consists of two electrolytics in series, with opposing polarities.*

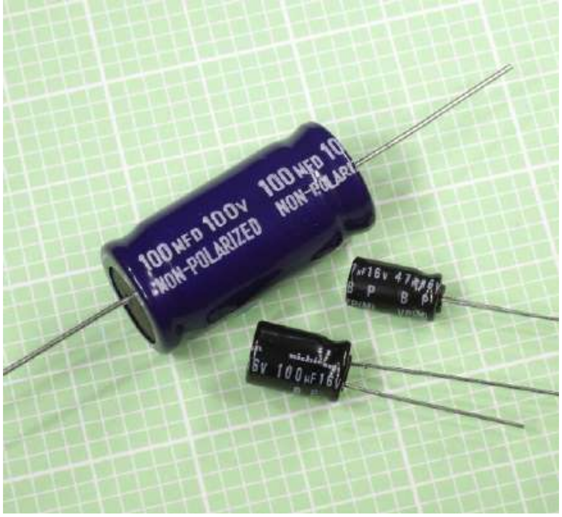

Figure 12-13. *Bipolar electrolytic capacitors. The larger size of the one at top-left is a consequence of its higher voltage rating. “BP” on the other two capacitors is an acronym for “bipolar,” meaning that they have no polarity, even though one lead may be shorter than the other.*

10µF at 100V). Surface-mount tantalum capacitors are decreasing in popularity as large-value ceramic capacitors are becoming available, with smaller dimensions and lower equivalent series resistance.

Plastic-film capacitors are discussed in the following section.

Single-layer *ceramic* capacitors are often used for bypass, and are suitable for high-frequency or audio applications. Their value is not very stable with temperature, although the “NPO” variants are more stable. Multilayer ceramic capacitors are more compact than single-layer ceramic, and

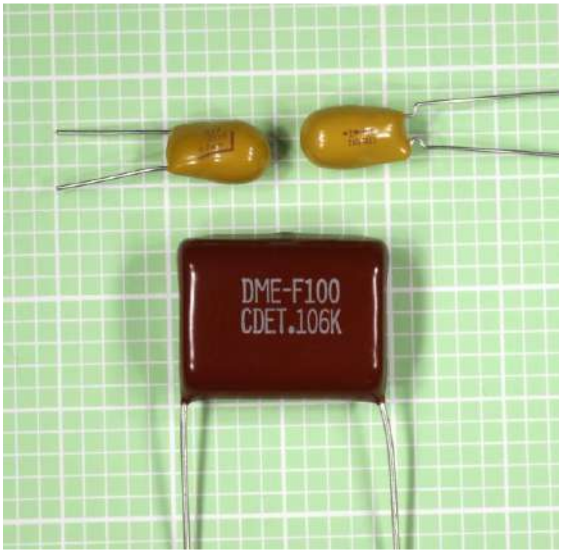

Figure 12-14. *Two tantalum capacitors are shown above a polyester film capacitor. The polarity of the tantalum capacitors is indicated by the plus signs adjacent to the longer lead, in each case. The polyester capacitor is nonpolarized.*

consequently are becoming increasingly popular. Three multilayer ceramic capacitors are shown in Figure 12-15. At bottom-right, even the largest (rated at 47μF at 16V) is only 0.2″ square.

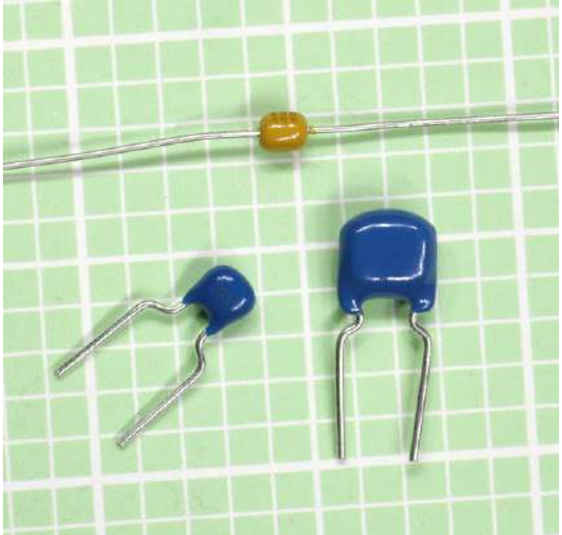

Figure 12-15. *Multilayer ceramic capacitors are extremely compact, and are nonpolarized. Top: 1,000pF (i.e. 1nF) at 100V. Bottom left: 1μF at 25V. Bottom right: 47μF at 16V.*

Dielectrics

The dielectric used in a capacitor most often consists of an electrolytic layer, a ceramic compound, a plastic film (polycarbonate, polypropylene, or polystyrene), or paper.

An *electrolytic layer* in an electrolytic capacitor traditionally consists of paper soaked in an electrolyte. It is interleaved with a thin film of aluminum on which is deposited a layer of aluminum oxide. The layers are rolled up to create a cylindrical component. The functioning dielectric is created when voltage is applied.

Polyester

This is the most common type of plastic film, with the highest dielectric constant, enabling highest capacitance per unit volume. Widely used in DC applications, but the rolled layers create parasitic inductance. Often used in decoupling, coupling, and bypass, but not so suitable for situations requiring stability and low leakage. May not be suitable for high current.

Polycarbonate

Thermally very stable, often specified for filters and timing circuits that require a fixed frequency. An excellent type of capacitor, compatible for mil-spec applications, but expensive.

Mylar, Polyester, and other plastic-film types are often used in audio circuits, where their voltage limitation (typically less than 100VDC) is not a problem, and their nonpolarized attribute is an advantage.

Polypropylene

Vulnerable to heat (a maximum of 85 degrees Centigrade is common), and less thermally stable than polycarbonate. A very low power dissipation factor allows it to handle higher power at higher frequencies. Available with tolerances down to 1%. These capacitors are a popular choice in crossover networks for loudspeaker combinations, and are used in switching power supplies. They tend to be physically larger than other capacitors using film dielectric.

Values

Farads

The electrical storage capacity of a capacitor is measured in *farads,* universally represented by the letter F. A capacitor that can be charged with a potential difference between its plates of 1 volt, in a time of 1 second, during which it draws 1 amp, has a capacitance of 1 farad.

Because the farad is a large unit, capacitors in electronic circuits almost always have fractional values: *microfarads* (µF), *nanofarads* (nF), and *picofarads* (pF). The Greek letter µ (mu) should be used in the µF abbreviation, but a lowercase letter u is often substituted. Thus, for example, 10uF means the same as 10µF.

1F = 1,000,000µF, and 1µF = 1,000,000pF. Therefore, 1 farad is equivalent to 1 trillion picofarads—a very wide range of possible values. See Figure 12-16 and Figure 12-17 for charts showing equivalent values in different units.

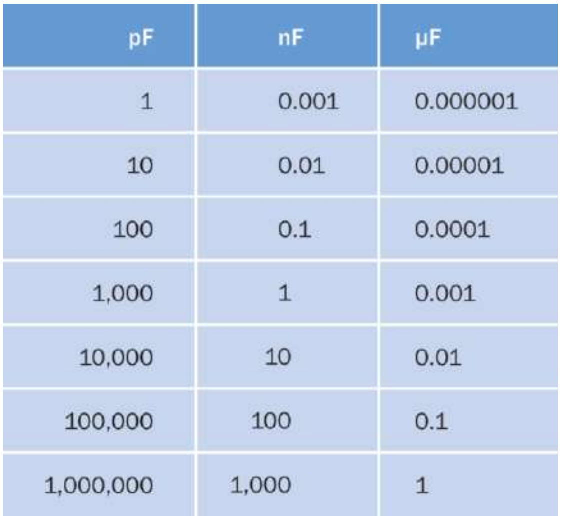

pF	nF	µF
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 12-16. *Equivalent values for picofards, nanofarads, and microfarads. The nF unit is used primarily in Europe.*

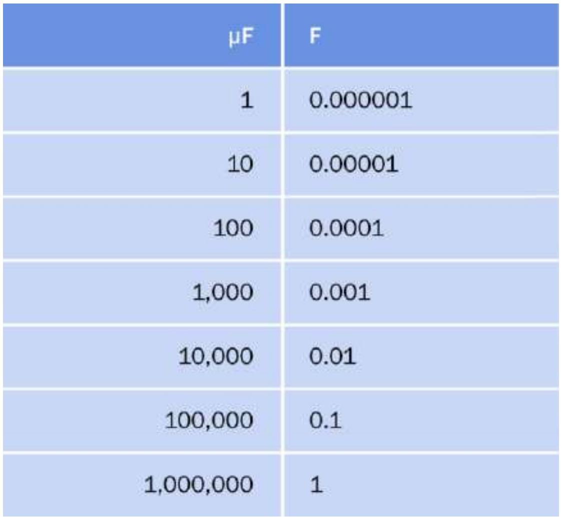

µF	F
1	0.000001
10	0.00001
100	0.0001
1,000	0.001
10,000	0.01
100,000	0.1
1,000,000	1

Figure 12-17. *Equivalent values for microfarads and farads. Because the farad is such a large unit, electronic circuits almost always use fractional values.*

The nF unit is more common in Europe than in the United States. A 1nF capacitance is often expressed in the US as 0.001µF or 1,000pF. Similarly, a 10nF capacitance is almost always expressed as a 0.01µF, and a 0.1nF capacitance is more likely to be expressed as 100pF.

European schematics may use value-symbols as a substitute for decimal points. For example, a 4.7pF capacitor may be shown as 4p7, a 6.8nF capacitor may be shown as 6n8, and a 3.3µF capacitor may be shown as 3µ3.

Commonly Used Values

The traditional range of capacitor values was established on the same basis as the traditional range of resistor values, by assuming an accuracy of plus-or-minus 20% and choosing factors that would minimize the possible overlap between adjacent tolerance ranges. The factors 1.0, 1.5, 2.2, 3.3, 4.7, 6.8, and 10 satisfy this requirement. See Chapter 10 for a more detailed explanation, including a graphical representation of values and overlaps in Figure 10-8. While many resistors are now manufactured with high precision, 20%

tolerance is still common for electrolytic capacitors. Other types of capacitors are available with an accuracy of 10% or 5%, but are more expensive.

While large-value capacitors are likely to have their actual value printed on them, smaller capacitors are identified by a variety of different codes. These codes are not standardized among manufacturers, and exist in various colors and abbreviations. A multimeter that can measure capacitance is a quicker, easier, and more reliable method of determining the value of a component than trying to interpret the codes.

In addition to capacitance, a large capacitor is likely to have its working voltage printed on it. Exceeding this value increases the risk of damaging the dielectric. In the case of electrolytic capacitors, a voltage that is much lower than the rated value should also be avoided, because these capacitors require an electrical potential to maintain their performance.

In common electronics applications, values larger than 4,700µF or smaller than 10pF are unusual.

Electrolytics are available at a moderate price in a wider range of values than other commonly used capacitors. They range from 1µF to 4,700µF and sometimes beyond. Working voltages typically range from 6.3VDC to 100VDC, but can be as high as 450VDC.

Tantalum capacitors are usually unavailable in sizes above 150µF or for voltages above 35VDC.

Single-layer ceramic capacitors have small values ranging from 0.01µF to 0.22µF, with working voltages usually not exceeding 50VDC, although very small-value capacitors may be rated much higher for special applications. Poor tolerances of +80% to -20% are common.

Some variants of multi-layer ceramic capacitors are capable of storing up to 47µF, although 10µF is a more common upper limit. They are seldom rated above 100VDC. Some are accurate to plus-or-minus 5%.

Dielectric Constant

If A is the area of each plate in a capacitor (measured in square centimeters), and T is the thickness of the dielectric (measured in centimeters), and K is the *dielectric constant* of the capacitor, the capacitance, C (measured in farads) will be obtained from the formula:

```
C = (0.0885 * K * A) / T
```

The dielectric constant of air is 1. Other dielectrics have different standard values. Polyethylene, for instance, has a constant of approximately 2.3. Thus a capacitor of 1 square centimeter plate area and polyethylene dielectric 0.01 centimeters thick would have a capacitance of about 20pF. A tantalum capacitor of equal plate area and dielectric thickness would have capacitance closer to 100pF, since the dielectric constant of tantalum oxide is much higher than that of polyethylene.

The Time Constant

When a capacitor is charged in series through a resistor (it is used in an RC network), and it begins with no charge on its plates, the *time constant* is the time, in seconds, required to charge the capacitor to 63% of the supply voltage. After an additional, identical interval of time, the capacitor will acquire 63% of the remaining difference between itself and the power supply. In theory the capacitor gets closer and closer to a full charge, but never quite reaches 100%. However, five time constants are sufficient for the capacitor to reach 99%, which is regarded as close enough to being fully charged for all practical purposes.

Refer to Figure 12-3 for a schematic of an RC network.

The time constant is a simple function of the resistance and the capacitance. If R is the value of the resistor (in ohms), and C is the value of the capacitor (in farads), the time constant, TC, will be obtained by the formula:

```
TC = R * C
```

If we multiply the R value by 1,000 while dividing the C value by 1,000, the time constant remains the same, and we can use the more convenient values of kilohms for the resistance and µF for the capacitance. In other words, the formula tells us that a 1K resistor in series with a 1,000µF capacitor has a time constant of 1 second.

The formula suggests that if the value of R diminishes to zero, the capacitor will charge instantly. In reality, the charging time will be rapid but finite, limited by factors such as the electrical resistance of the materials used.

Multiple Capacitors

When two or more capacitors are wired in parallel, their total capacitance is the sum of their separate capacitances. When two or more capacitors are wired in series, the relationship between their total capacitance © and their individual capacitances (C1, C2, C3 . . .) is given by this formula:

```
1 / C = (1/C1) + (1/C2) + (1/C3). . .
```

The formula to calculate the total capacitance of capacitors connected in series resembles the one used to calculate the total resistance of resistors connected in parallel. See Chapter 10.

Alternating Current and Capacitive Reactance

The apparent resistance of a capacitor to AC is properly known as *capacitive reactance*. In the following formula, capacitive reactance (X_C, in ohms) is derived as a function of capacitance (C, in farads) and AC frequency (f, measured in hertz):

$$X_C = 1 / (2 * \pi * f * C)$$

The formula shows that when frequency becomes zero, capacitive reactance becomes infinite; in other words, a capacitor has theoretically infinite resistance when DC current tries to flow through it. In reality, a dielectric has a finite resistance, and thus always allows some leakage.

The formula also shows that capacitive reactance diminishes when the size of the capacitor increases and/or the frequency being applied to it increases. From this it appears that an AC signal will be attenuated less at higher frequencies, especially if we use a small capacitor. However, a real-world capacitor also exhibits some degree of *inductive reactance*. This value will depend on its configuration (cylindrical vs. multiple flat plates), its physical length, the materials from which it is fabricated, the lengths of its leads, and other factors. Inductive reactance tends to *increase* with frequency, and since capacitive reactance tends to *decrease* with frequency, at some point the curves for the two functions intersect. This point represents the capacitor's *self-resonant frequency*, which is often referred to simply as its *resonant frequency*. See Figure 12-18.

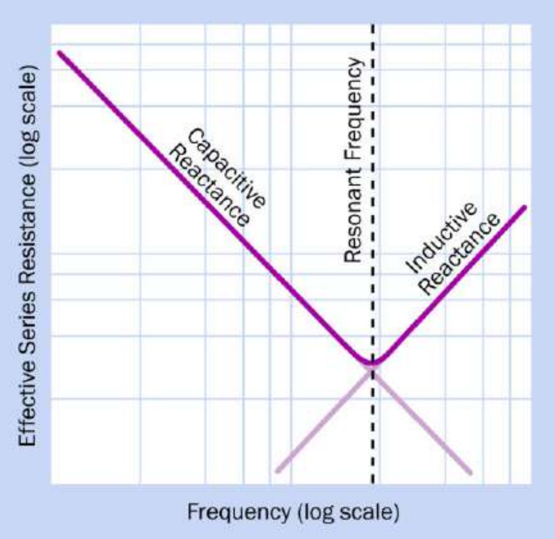

Figure 12-18. *As an AC current applied to a capacitor increases in frequency, the* capacitive reactance *of the component decreases, while its* inductive reactance *increases. The resonant frequency of the capacitor is found where the two functions intersect.*

Equivalent Series Resistance

A theoretically ideal capacitor would be purely reactive, without any resistance. In reality, capacitors are not ideal, and they have *equivalent series resistance,* or *ESR*. This is defined as the resistor that you would have to place in series with an ideal version of the capacitor, so that the combination would behave like the real version of the capacitor on its own.

If X_c is the reactance of the capacitor, then its Q factor (which means its *quality factor*) is given by the simple formula:

$$Q = X_C / ESR$$

Thus, the quality factor is higher if the ESR is relatively low. However, the reactance of the capacitor will vary significantly with frequency, and this simple formula is only an approximate guide.

The Q-factor for capacitors should not be confused with the Q-factor for inductors, which is calculated quite differently.

How to Use it

The figures illustrate some simplified schematics for common applications.

Bypass Capacitor

In Figure 12-19, a low-value capacitor (often 0.1µF) is placed near the power input pin of a sensitive digital chip to divert high-frequency spikes or noise to negative ground. This *bypass capacitor* may also be described as a *decoupling capacitor*.

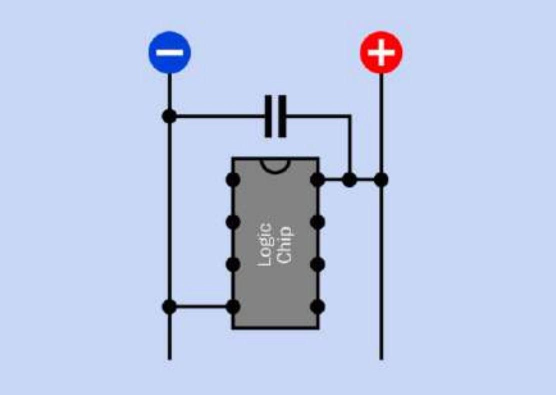

Figure 12-19. *A bypass capacitor (typically 0.1µF) configured to protect an integrated circuit logic chip from voltage spikes and noise in the power supply.*

Coupling Capacitor

In Figure 12-20, a 1µF *coupling capacitor* transmits a pulse from one section of a circuit to another, while blocking the DC voltage. Some reshaping of the waveform may occur.

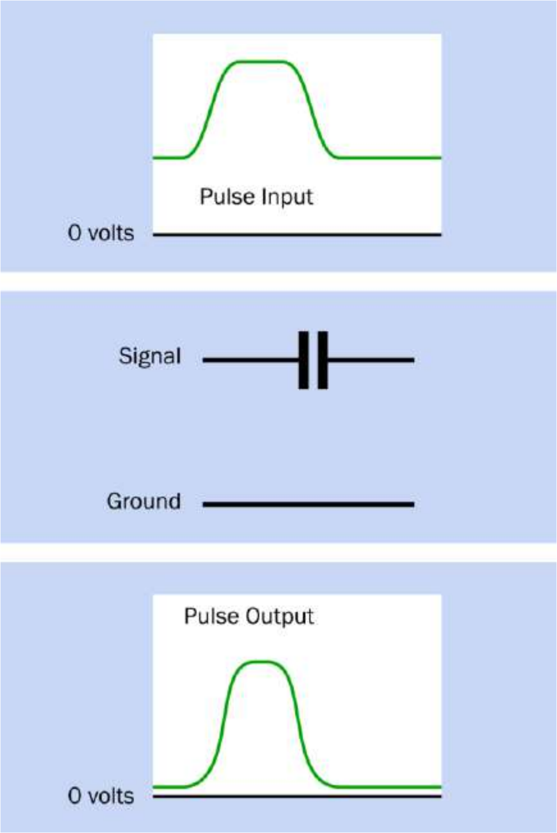

Figure 12-20. *A coupling capacitor (typically around 1µF) preserves DC isolation of one section of a circuit from another, while allowing a pulse to be transmitted.*

High-Pass Filter

In Figure 12-21, a 0.1µF capacitor blocks the low-frequency component of a complex waveform and transmits only the higher frequency that was superimposed on the low frequency.

Low-Pass Filter

In Figure 12-22, a 0.1µF decoupling capacitor diverts the higher frequency component of a complex waveform to negative ground, allowing only

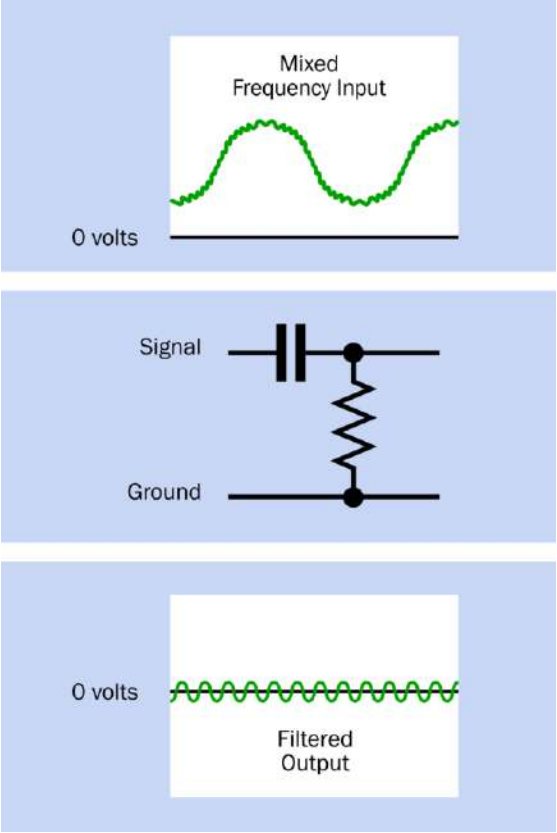

Figure 12-21. *A small capacitor (typically 0.1μF) can be used to create a high-pass filter, passing high frequencies while blocking low frequencies.*

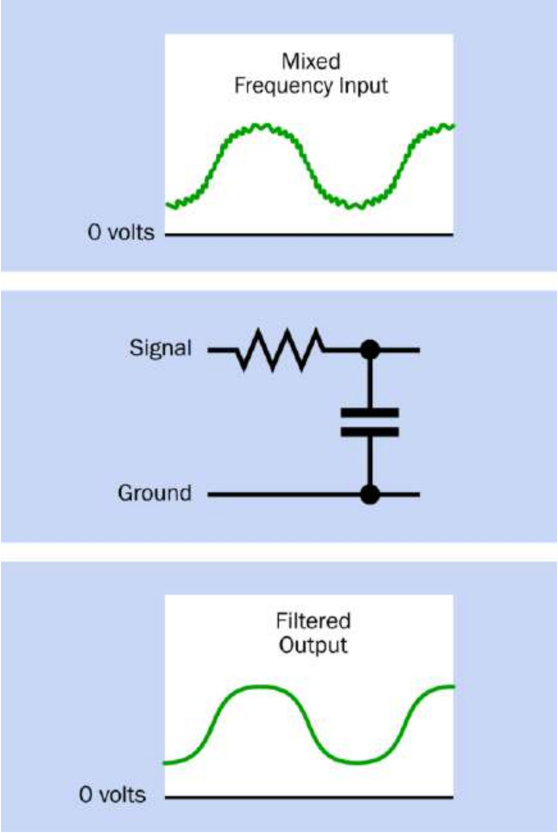

Figure 12-22. *A small capacitor (typically 0.1μF) in this configuration routes high frequencies to negative ground, filtering them out of an analog signal.*

the lower frequency to be preserved. A lower-value capacitor (such as 0.001μF) will bleed away high-frequency noise from an AM radio source without affecting audio frequencies.

Smoothing Capacitor

In Figure 12-23, a 100μF capacitor charges and discharges to smooth an AC signal after a diode has removed the negative portion.

Snubber

In Figure 12-24, an RC network (inside a white dashed line) is known as a *snubber* when used to protect a switch from the problem of *arcing* (pronounced "arking")—that is, a sustained spark that can quickly erode the switch contacts. Arcing may occur in switches, pushbuttons, or relays that control an inductive load, such as a large motor. This problem can become significant at high DC currents (10A or more) or relatively high AC or DC voltages (100V or more).

When the switch is opened, the magnetic field that has been sustained by the inductive load collapses, causing a surge of current, or *forward EMF*. The capacitor in the snubber absorbs this surge, thus protecting the switch contacts. When the switch is closed again, the capacitor discharges itself, but the resistor limits the outrush of current—again, protecting the switch.

A snubber placed around the switch in a DC circuit could typically use a 0.1μF capacitor (polypropylene or polyester) rated for 125VAC/200VDC, and a 100-ohm carbon resistor rated 0.5

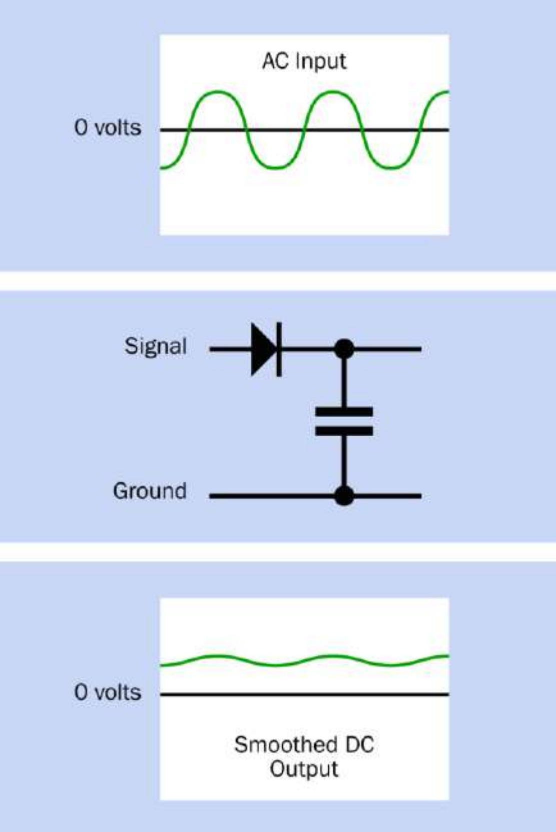

Figure 12-23. *A capacitor of 100µF or more smooths the upper half of an AC signal that has passed through a diode. The capacitor charges during each positive pulse and discharges to "fill the gaps" between them.*

watt or higher. Prepackaged snubbers containing appropriate capacitor-resistor pairs are available from some parts suppliers, primarily for industrial use.

In an AC circuit, a snubber can be placed around the inductive load itself. Although a **diode** is often used this way in a DC circuit, it cannot be used with AC.

Although solid-state switching devices such as a **solid state relay** contain no mechanical contacts, they may still be damaged by substantial pulses of back-EMF, and can be protected by a snubber where they are controlling inductive loads that take 10A or more at 100V or more.

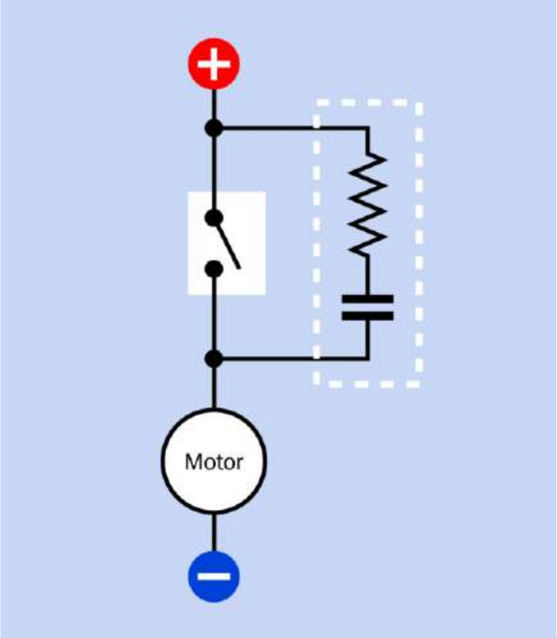

Figure 12-24. *An RC network (outlined with a white dashed line) protects a switch that controls a high inductive load. Used in this way, the RC network is known as a snubber.*

Capacitor as a Battery Substitute

A capacitor may be substituted for a battery for some applications, although it has a lower energy density and will be more expensive to manufacture. A capacitor charges and discharges much more rapidly than a battery because no chemical reactions are involved, but a battery sustains its voltage much more successfully during the discharge cycle.

Capacitors that can store a very large amount of energy are often referred to as *supercapacitors.*

What Can Go Wrong

Common problems associated with capacitors are age-related deterioration (especially in electrolytics), inductive reactance (especially in cylindrical formats), nonlinear response, resistivity, excessive current leakage, and dielectric memory. Some of these problems are discussed below. A manufacturer's datasheet should be consulted carefully in conjunction with the notes regarding compositions in the preceding Variants section before making a commitment to a particular type of capacitor.

Wrong Polarity

A polarized capacitor may offer virtually no resistance if it is connected the wrong way around to a DC power source. A very high current can result, damaging the capacitor and probably other components in the circuit. Failing to observe the polarity of a tantalum capacitor can have destructive or even explosive consequences, depending on the amperage.

Voltage Overload

If the DC working voltage of a capacitor is exceeded, there is a risk of breaking down the dielectric and allowing a spark, or arc, that will form a short circuit. Note that the DC rating of a capacitor does not mean that it can be used safely with an equivalent AC voltage. The maximum AC voltage should be no greater than approximately 0.7 times the DC rated voltage. If a DC-rated capacitor is used directly across an AC power line, it will create an effective short circuit.

If capacitors are connected in series or in parallel, ideally the voltage rating for each capacitor should be the same, and certainly no less than the supply voltage.

Tantalum capacitors are easily damaged by current spikes that exceed their maximum working voltage, and are unsuitable for high-frequency coupling because of their inductance.

Leakage

Charge leakage is a problem especially associated with electrolytic capacitors, which are not suitable for storing a charge over a significant interval. Polypropalene or polystyrene film capacitors are a better choice.

Dielectric Memory

Also known as *dielectric absorption*, this is a phenomenon in which a capacitor's electrolyte displays some percentage of its former voltage after the capacitor has been discharged and then disconnected from the circuit. Single-layer ceramic capacitors especially tend to suffer from this problem.

Specific Electrolytic Issues

Electrolytic capacitors have high inductive reactance, are not manufactured to close tolerances, and deteriorate significantly with age. While other components may be stockpiled and used over a period of years, this is not a sensible policy with electrolytics.

The "capacitor plague" affecting many of these capacitors manufactured from 1999 onward provided a salutary lesson regarding their potential weaknesses. Faulty composition of the dielectric allowed it to deteriorate, liberating hydrogen gas, which eventually caused the aluminum shells of the capacitors to bulge and burst. Circuit boards from major manufacturers were affected. Because the problem took two years to become apparent, literally millions of boards with faulty capacitors had been sold before the fault was diagnosed and eventually corrected.

Unfortunately electrolytics cannot be easily replaced with other types of capacitors in applications such as power supplies, because substitutes will be considerably larger and more expensive.

Heat

The equivalent series resistance (ESR) of a large capacitor inevitably means that it must dissipate some power as heat during use. Ripple current can also create heat. Capacitor performance will change as the temperature increases. A common maximum component temperature for electrolytic capacitors is 85 degrees Centigrade.

Vibration

In a high-vibration environment, electrolytics should be protected by clamping them mechanically in place, using a *capacitor clamp*, also known as a *c-clamp*.

Misleading Nomenclature

Rarely, in the United States, the term “mF” may be used as a probable alternative to μF. This can be a source of confusion and risk because mF is properly (but very rarely) used to mean “millifarads.” The term should always be avoided.

variable capacitor

13

Formerly known (primarily in the United Kingdom) as a *variable condenser*. The term is now obsolete.

OTHER RELATED COMPONENTS

- **capacitor** (See Chapter 12)

What It Does

A variable capacitor allows adjustment of capacitance in much the same way that a **potentiometer** allows adjustment of resistance.

Large variable capacitors were developed primarily to tune radio receivers, in which they were known as *tuning capacitors*. Cheaper, simpler, and more reliable substitutes gradually displaced them, beginning in the 1970s. Today, they are still used in semiconductor fabrication, in RF plastic welding equipment, in surgical and dental tools, and in ham radio equipment.

Small *trimmer capacitors* are widely available and are mostly used to adjust high-frequency circuits. Many of them look almost indistinguishable from *trimmer potentiometers*.

The schematic symbols commonly used to represent a variable capacitor and a trimmer capacitor are shown in Figure 13-1.

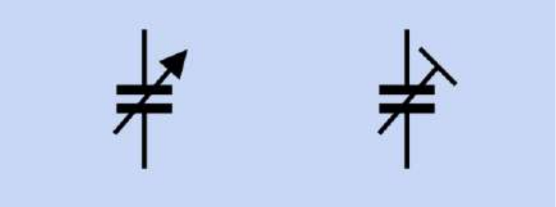

Figure 13-1. *Typical schematic symbols for variable capacitor (left) and trimmer capacitor (right).*

A *varactor* is a form of diode with variable capacitance, controlled by reverse voltage. See "Varactor Diode" (page 225) for this component.

How It Works

The traditional form of variable capacitor consists of two rigid semicircular plates separated by an air gap of 1mm to 2mm. To create more capacitance, additional interleaved plates are added to form a stack. One set of plates is known as the *rotor*, and is mounted on a shaft that can be turned, usually by an externally accessible knob. The other set of plates, known as the *stator*, is mounted on the frame of the unit with ceramic insulators. When the sets of plates completely overlap, the capacitance between them is maximized. As the rotor is turned, the sets of plates gradually disengage, and the capacitance diminishes to near zero. See Figure 13-2.

The air gaps between the sets of plates are the *dielectric*. Air has a dielectric constant of approximately 1, which does not vary significantly with temperature.

The most common shape of plate is a semicircle, which provides a linear relationship between capacitance and the angle of rotation. Other shapes have been used to create a nonlinear response.

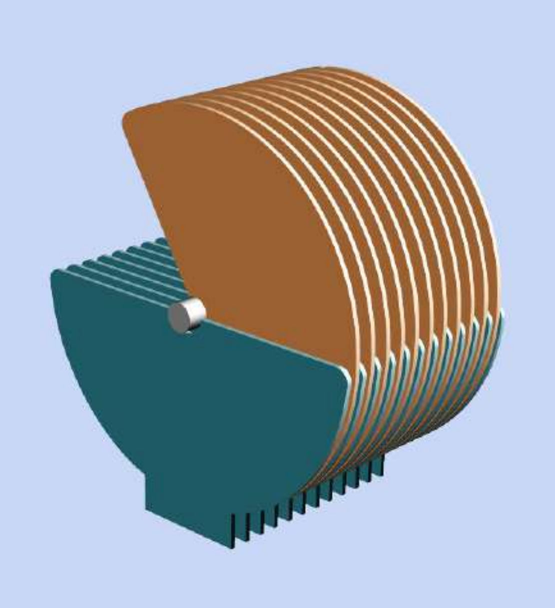

Figure 13-2. *In this simplified view of a variable capacitor, the brown plates constitute the rotor, attached to a central shaft, while the blue plates are the stator. The colors have no electrical significance and are added merely for clarity. The area of overlap between rotor and stator determines the capacitance.*

Reduction gears may be used to enable fine tuning of a variable capacitor, which means multiple turns of a knob can produce very small adjustments of the capacitor. At the peak of variable capacitor design, units were manufactured with high mechanical precision and included *anti-backlash gears*. These consisted of a pair of equal-sized gears mounted flat against each other with a spring between them that attempted to turn the gears in opposite directions from each other. The pair of gears meshed with a single pinion, eliminating the looseness, or backlash, that normally exists when gear teeth interlock. A vintage capacitor with a spring creating anti-backlash gearing (circled) is shown in Figure 13-3. This is a two-gang capacitor—it is divided into two sections, one rated 0 to 35pF, the other rated 0 to 160pF.

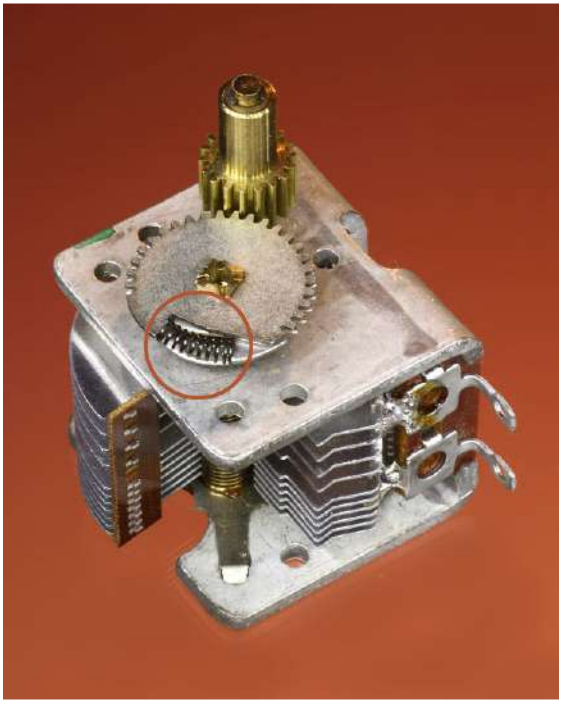

Figure 13-3. *A "traditional style" variable capacitor of the type designed to tune radio frequencies. The spring, circled, enables anti-backlash gearing.*

Variants

The traditional variable capacitor, with exposed, air-spaced, rigid, rotating vanes, is becoming hard to find. Small, modern variable capacitors are entirely enclosed, and their plates, or vanes, are not visible. Some capacitors use a pair of concentric cylinders instead of plates or vanes, with an external thumb screw that moves one cylinder up or down to adjust its overlap with the other. The overlap determines the capacitance.

Trimmer capacitors are available with a variety of dielectrics such as mica, thin slices of ceramic, or plastic.

Values

A large traditional capacitor can be adjusted down to a near-zero value; its maximum will be no greater than 500pF, limited by mechanical factors. (See Chapter 12 for an explanation of capacitance units.)

A maximum value for a trimmer capacitor is seldom greater than 150pF. Trimmers may have their values printed on them or may be color-coded, but there is no universal set of codes. Brown, for example, may indicate either a maximum value around 2pF or 40pF, depending on the manufacturer. Check datasheets for details.

The upper limit of a trimmer's rated capacitance is usually no less than the rated value, but can often be 50% higher.

Formats

All trimmer capacitors are designed for mounting on circuit boards. Many are surface-mount, with a minority being through-hole. Surface-mount units may be 4mm × 4mm or smaller. Through-hole are typically 5mm × 5mm or larger. Superficially, trimmer capacitors resemble single-turn trimmer potentiometers with a screw head in the center of a square package. A through-hole example is shown in Figure 13-4.

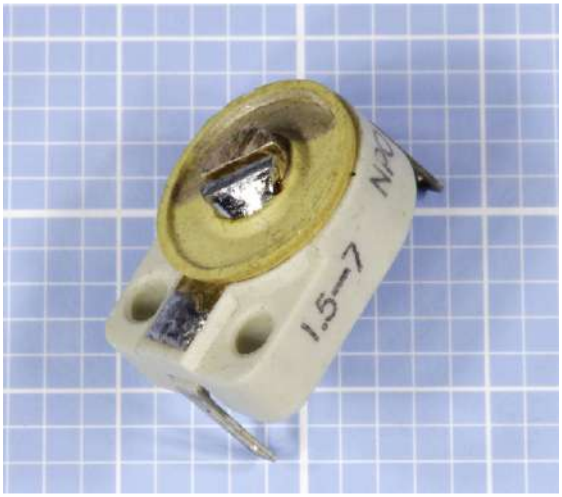

Figure 13-4. *A trimmer capacitor rated 1.5pF to 7.0pF.*

How to Use it

A variable capacitor is often used to tune an *LC circuit,* so called because a coil (with reactance customarily represented by letter L) is wired in parallel with a variable capacitor (represented by letter C). The schematic in Figure 13-5 shows an imaginary circuit to illustrate the principle. When the switch is flipped upward, it causes a large fixed-value capacitor to be charged from a DC power source. When the switch is flipped down, the capacitor tries to pass current through the coil—but the coil's *reactance* blocks the current and converts the energy into a magnetic field. After the capacitor discharges, the magnetic field collapses and converts its energy back into electricity. This flows back to the capacitor, but with inverted polarity. The cycle now repeats with current flowing in the opposite direction. A low-current **LED** across the circuit would flash as the voltage oscillates, until the energy is exhausted.

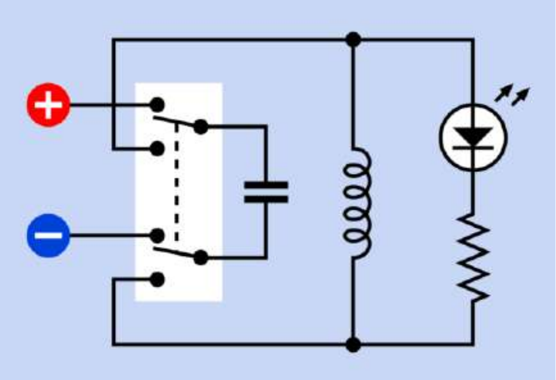

Figure 13-5. *In this imaginary circuit, the capacitor is charged through the double-pole switch in its upper position. When the switch is turned, the capacitor forms an LC (inductance-capacitance) circuit with the coil, and resonates at a frequency determined by their values. In reality, extremely high values would be needed to obtain a visible result from the LED.*

Because the oscillation resembles water sloshing from side to side in a tank, an LC circuit is sometimes referred to as a *tank circuit.*

In reality, unrealistically large values would be required to make the circuit function as described. This can be deduced from the following formula, where f is the frequency in Hz, L is inductance in Henrys, and C is capacitance in Farads:

```
f = 1 / (2π * √(L * C) )
```

For a frequency of 1Hz, a massive coil opposite a very large capacitor of at least 0.1F would be needed.

However, an LC circuit is well-suited to very high frequencies (up to 1,000MHz) by using a very small coil and variable capacitor. The schematic in Figure 13-6 shows a high-impedance earphone and a diode (right) substituted for the LED and the resistor in the imaginary circuit, while a variable capacitor takes the place of the fixed capacitor. With the addition of an antenna at the top and a ground wire at the bottom, this LC circuit is now capable of receiving a radio signal, using the signal itself as the source of power. The resonant frequency of the circuit is tuned by the variable capacitor. The impedance peaks at the resonant frequency, causing other frequences to be rejected by passing them to ground. With suitable refinement and amplification, the basic principle of an LC circuit is used in AM radios and transmitters.

Because variable capacitors are so limited in size, they are unsuitable for most timing circuits.

Trimmer capacitors are typically found in high-power transmitters, cable-TV transponders, cellular base stations, and similar industrial applications.

They can be used to fine-tune the resonant frequency of an oscillator circuit, as shown in Figure 13-7.

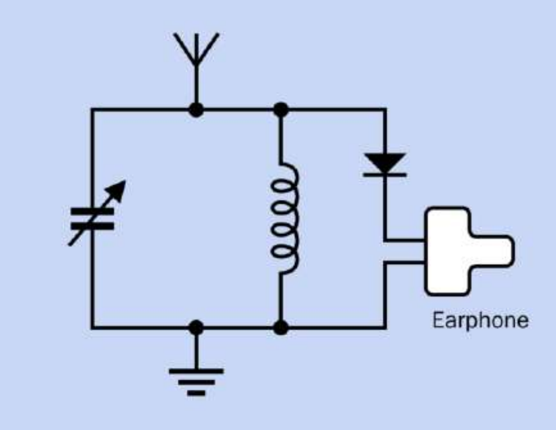

Figure 13-6. *The principle of an LC circuit is used here in a basic circuit that can tune in to a radio station and create barely audible sound through the earphone at right, using only the broadcast signal for power. The variable capacitor adjusts the frequency of the circuit to resonate with the carrier wave of the radio signal.*

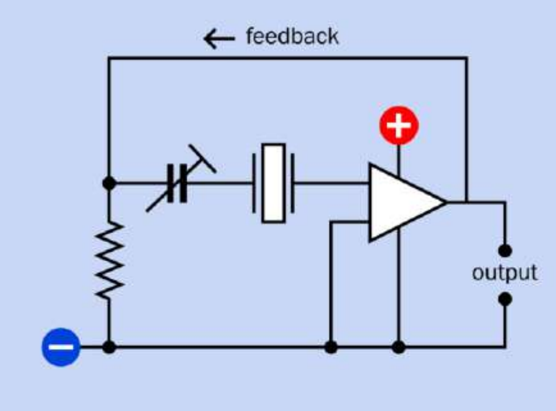

Figure 13-7. *A trimmer capacitor in series with a crystal fine-tunes the frequency of this basic circuit using an op-amp.*

In addition to tuning a circuit frequency, a trimmer capacitor can be used to compensate for changes in capacitance or inductance in a circuit that are caused by the relocation of wires or rerouting of traces during the development process. Readjusting a trimmer is easier than swapping fixed-value capacitors. A trimmer may also be used to compensate for capacitance in a circuit that gradually drifts with age.

What Can Go Wrong

Failure to Ground Trimmer Capacitor While Adjusting it

Although trimmer capacitors are not polarized, the manufacturer may mark one terminal with a plus sign and/or the other with a minus sign. If the capacitor is adjusted while its negative terminal is floating or ungrounded, a metal screwdriver blade will create erroneous readings. Always ground the appropriate side of a trimmer capacitor before fine-tuning it, and preferably use a plastic-bladed screwdriver.

Application of Overcoat Material or "Lock Paint"

Overcoat material is a rubbery adhesive that may be spread over assembled components to immunize them against moisture or vibration. *Lock paint* is a dab of paint that prevents a screw adjustment from turning after it has been set. Most manufacturers advise against applying these materials to a trimmer capacitor, because if penetration occurs, the capacitor can fail.

Lack of Shielding

Variable capacitors should be shielded during use, to protect them from external capacitive effects. Merely holding one's hand close to a variable capacitor will change its value.

inductor

14

The term **inductor** is used here to describe a coil that has the purpose of creating *self-inductance* in an electronic circuit, often while passing alternating current in combination with resistors and/or capacitors. A *choke* is a form of inductor. By comparison, the **electromagnet** entry in this encyclopedia describes a coil containing a center component of ferromagnetic material that does not move relative to the coil, and has the purpose of attracting or repelling other parts that respond to a magnetic field. A coil containing a center component of ferromagnetic material that moves as a result of current passing through the coil is considered to be a **solenoid** in this encyclopedia, even though that term is sometimes more broadly applied.

OTHER RELATED COMPONENTS

- **solenoid** (See Chapter 21)
- **electromagnet** (See Chapter 20)

What It Does

An inductor is a coil that induces a magnetic field in itself or in a *core* as a result of current passing through the coil. It may be used in circuits to block or reshape AC current or a range of AC frequencies, and in this role can "tune" a simple radio receiver or various types of oscillators. It can also protect sensitive equipment from destructive voltage spikes.

The schematic symbol for an inductor includes a coil that can be drawn in two basic styles, shown at the top and at the bottom of Figure 14-1. The style at the bottom has become more common. In each vertical section of the diagram, the functionality of the symbols is identical.

One or two parallel lines alongside the coil indicate that it is wound around a solid core of material that can be magnetized, while one or two dotted lines indicate that it is wound around a core containing metal particles, such as iron filings. Where no core is shown, this indicates an air core.

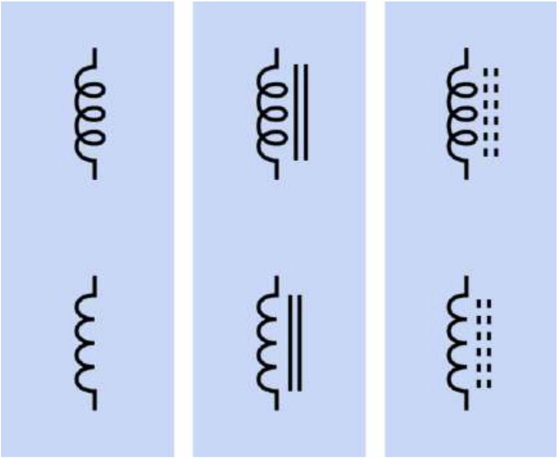

Figure 14-1. *The coil symbol for an inductor may be drawn in two styles which are functionally identical. Line(s) beside the coil indicate a solid core. Dotted line(s) indicate a core containing metal particles.*

A selection of inductors designed for through-hole mounting is shown in Figure 14-2.

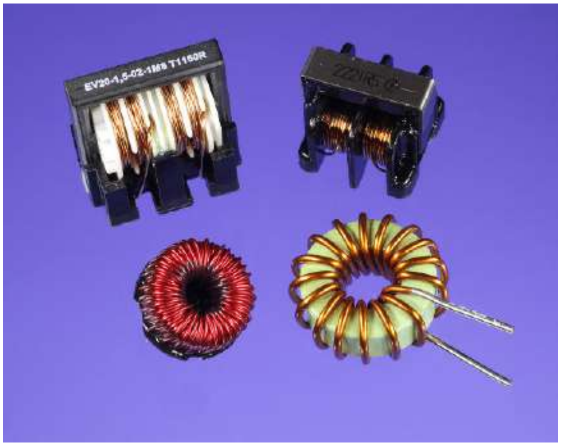

Figure 14-2. *Four inductors designed for through-hole insertion into printed circuit boards.*

How It Works

Direct current passing through an electrical conductor, such as a wire, creates a magnetic field around the conductor. In Figure 14-3, *conventional current* (flowing from positive to negative) is passing through a straight wire from left to right, as indicated by the red/blue arrow. The resulting magnetic field is indicated by the green arrows. If the wire is now bent into a curve, as shown in Figure 14-4, the magnetic field exerts an aggregate force downward through the curve. This magnetic force is conventionally said to flow from *south* to *north*.

If direct current could be induced to circulate through an unbroken circle of wire, the resulting magnetic field could exert a force through the circle as shown in Figure 14-5, assuming clockwise circulation of conventional current as suggested by the red/blue arrows.

Conversely, if a magnet was pushed through the center of the circle, it would induce a pulse of electric current in the circle. Thus, electricity passing through a wire can induce a magnetic field around the wire, and conversely, a magnet moving near a wire can induce an electric current in the wire. This principle is used in an electrical *generator*, and also in a **transformer**, where alternating current in the primary coil induces a fluctuating magnetic field in the core, and the field in the core is turned back into alternating current in the secondary coil.

Note that a static or unchanging magnetic field will not induce a flow of electricity.

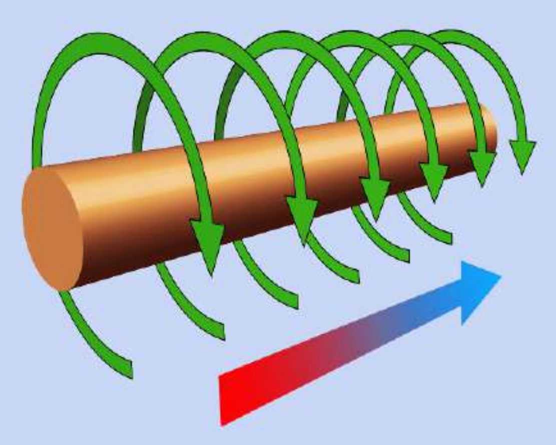

Figure 14-3. *Conventional current passing through a wire from left to right (as indicated by the red/blue arrow) induces a magnetic field around the wire (shown by the green arrows).*

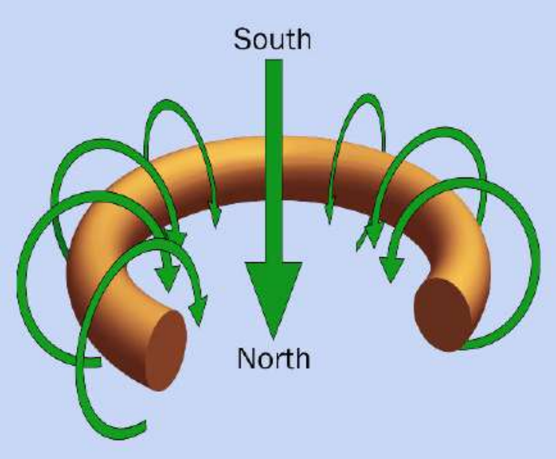

Figure 14-4. *If the wire is bent into a curve, the magnetic fields can create a net force shown by the large green arrow.*

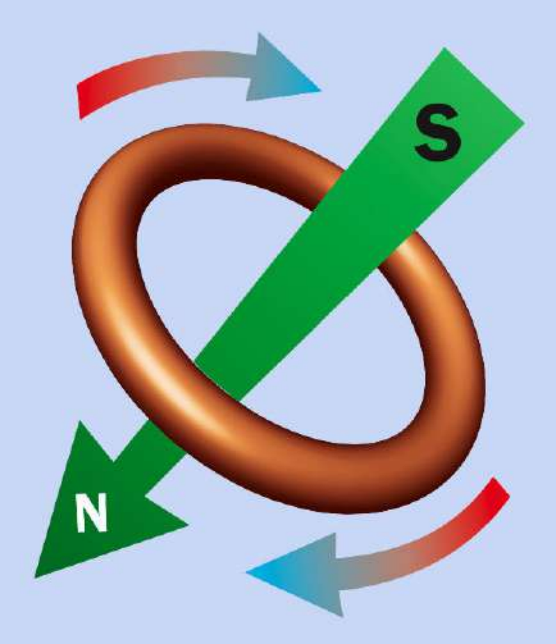

Figure 14-5. *Hypothetically, if conventional current flows around a circular conductor (as suggested by the red/blue arrows), it will create a magnetic field that can create a force as shown by the green arrow.*

DC Through a Coil

If the wire is formed into a helix (a series of approximate circles) as shown in Figure 14-6, and if DC current is passed through the wire, the aggregate of the magnetic fields can create a force in the direction of the green arrow in each example, depending whether the wire is wound clockwise or counter-clockwise, and depending on the direction of the current. The helix is usually referred to as a *coil* or a *winding*.

In actuality, a magnetic field is not open-ended, and its lines of force are completed by circling around outside the inductor, to complete a *magnetic circuit*. This completion of the field can be demonstrated by the traditional high-school experiment of positioning a compass or scattering iron filings on a sheet of paper above a magnet. A simplified depiction of lines of force completing a magnetic circuit is shown in Figure 14-7, where a coil is inducing the magnetic field. Note that throughout this encyclopedia, the color green is used to indicate the presence of magnetic force.

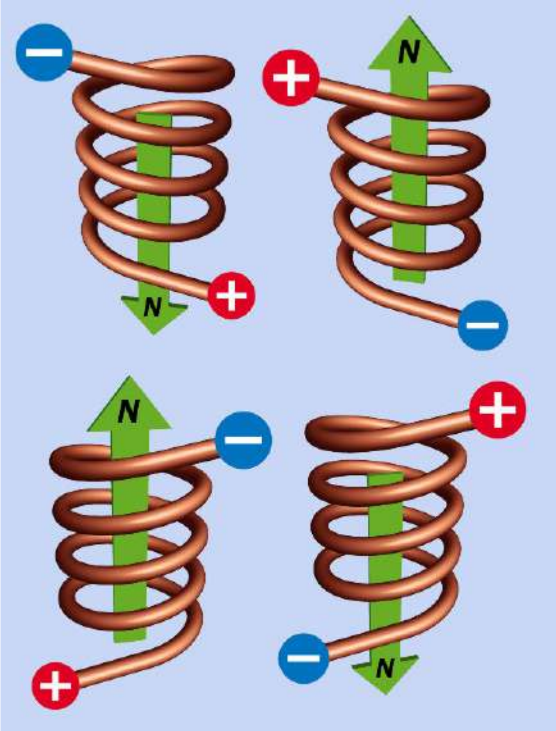

Figure 14-6. *When DC current flows through a coil, it creates magnetic fields that will exert a force whose direction depends on the direction of the current and on whether the coil is wound clockwise or counterclockwise. The force is shown by the green arrow in each case.*

The completion of a magnetic field is not relevant to the primary function of the inductor. In fact the external part of the magnetic field is mostly a source of trouble in electronics applications, since it can interact with other components, and may necessitate the use of magnetic shielding. In addition, the field is weakened by completing itself through air, as air presents much greater *reluctance* (the magnetic equivalent of resistance) than the core of an electromagnet.

The polarity of a magnetic field created by a coil can be demonstrated by moving a small permanent magnet toward the coil, as shown in

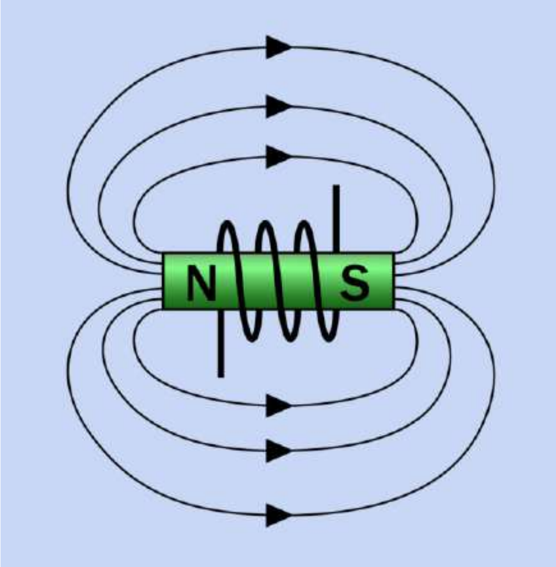

Figure 14-7. *A magnetic field in reality is not open-ended, and each line of force traveling through a rod-shaped magnet or electromagnet is completed outside of the magnet. The completion of magnetic fields has been omitted from other diagrams here for clarity.*

Figure 14-8. If the magnet has opposite polarity to the coil, it will tend to be repelled, as like poles repel. If it has the same polarity, it will tend to be attracted, because opposite poles attract. This principle may be used in **solenoids**.

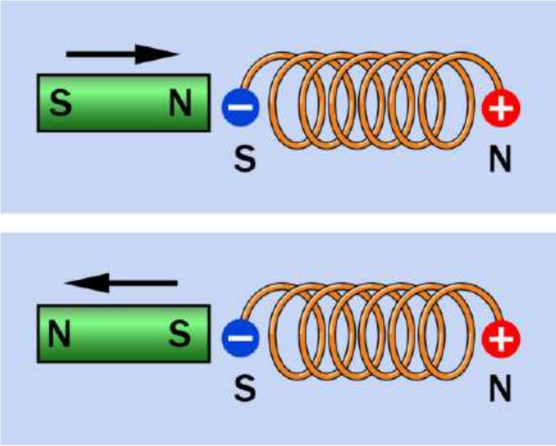

Figure 14-8. *A permanent magnet (left) will either be drawn toward a DC-energized coil or repelled from it, depending on the polarity of the two magnetic fields.*

Magnetic Core

The inductive power of a coil will be enhanced, and the saturation point will be reduced, by using a *magnetic core*. The term "magnetic" in this context does not mean that the core is a permanent magnet; it means that the core can be magnetized briefly by a transient pulse of electricity through the surrounding coil.

A core enhances the effectiveness of an inductor because it has a lower reluctance than that of air. In other words, magnetic flux will flow much more readily through the core than through air.

Roughly speaking, the *permeability* of a magnetic circuit is the opposite of reluctance; it is a measure of how easily a magnetic field can be induced, and is usually expressed relative to the permeability of air, which is approximately 1. The permeability of different core types is discussed in the following "Values" section.

The core of the coil contains *magnetic domains* that behave as tiny magnets, with north and south poles. In the absence of a polarizing magnetic field, the domains are randomly aligned. As a magnetic field is introduced around them and grows stronger, the domains align themselves with it, increasing the total magnetic force. When the domains are almost all uniformly aligned, the core approaches *magnetic saturation* and ceases adding to the net magnetic field. At this point the current in the inductor is said to be *continuous*.

When power to the coil is disconnected, the domains revert partially to their previous random orientation. Thus the core remains a weak permanent magnet. This effect is known as *hysteresis*, while the weak residual field is known as *remanent magnetism*.

EMF and Back-EMF

When DC current is connected through an inductor, the creation of a magnetic field takes a brief but measurable period of time. The field induces an *EMF* (*electro-motive force*) in the wire. Since this force opposes the supplied current, it is referred to as *back-EMF*. It lasts only so long as the field is increasing to its full strength. After the field reaches a steady state, current flows through the coil normally.

This transient resistive effect is caused by the *self-inductance* of the coil, and is opposite to the behavior of a capacitor, which encourages an initial inrush of direct current until it is fully charged, at which point it blocks subsequent current flow.

When high-frequency alternating current attempts to flow through an inductor, if each pulse is too brief to overcome the back-EMF, the coil will block the current. A coil can thus be designed to block some frequencies but not others.

Even a simple electrical circuit that does not contain a coil will still have some self-inductance, simply because the circuit consists of wires, and even a straight length of wire induces a magnetic field when the power is switched on. However, these inductive effects are so small, they can generally be ignored in practical applications.

The transient electrical resistance to alternating current caused by either an inductor or a capacitor is known as *reactance*, although it occurs under opposite electrical conditions, as the coil impedes an initial pulse of DC current and then gradually allows it to pass, while a capacitor allows an initial pulse of DC current and then impedes it.

When a flow of DC current through a coil is switched off, the magnetic field that was created by the coil collapses and releases its stored energy. This can cause a pulse of forward EMF, and like back-EMF, it can interfere with other components in a circuit. Devices such as motors and large relays that contain substantial coils can create problematic spikes of back-EMF and forward-EMF. The forward-EMF that occurs when power to the coil is interrupted is typically dealt with by putting a diode in parallel with the coil, allowing current to circulate through it. This is known as *clamping* the voltage transient. A diode-capacitor combination known as a *snubber* is also commonly used. For a schematic and additional information on this topic, see "Snubber" (page 108).

A schematic to demonstrate EMF and back-EMF is shown in Figure 14-9. The coil can be a 100-foot spool of 26-gauge (or smaller) hookup wire, or magnet wire. It will function more effectively if a piece of iron or steel, such as half-inch galvanized pipe, is inserted through its center. When the button is pressed, current is briefly impeded by the back-EMF created by the coil, and is diverted through D1, making it flash briefly. Then the coil's reactance diminishes, allowing the current to flow through the coil and bypass the LED. When the pushbutton is released, the coil's magnetic field collapses, and the consequent forward-EMF circulates through D2, causing it to flash briefly. Note that the polarity of back-EMF and forward-EMF are opposite, which is why the LEDs in the circuit are oriented with opposite polarities.

The 220Ω resistor should be rated at 1/4 watt minimum, and the button should not be held down for long, as the electrical resistance of the coil is relatively low. The LEDs ideally should be rated for a minimal forward current of no more than 5mA.

Electrical and Magnetic Polarity

Various mnemonics and images have been created to assist in memorizing the polarity or direction of the magnetic field that will be created by a flow of electricity. The *right-hand rule* suggests that if the fingers of the right hand are curled around a coil in the same direction in which the turns of the coil were wound, and if conventional DC current also flows in this direction, the extended thumb will point in the direction of the principal force that can be created by the magnetic field.

By convention, the magnetic field is oriented from *south* to *north*, which can be remembered since the north end of the magnetic field will be

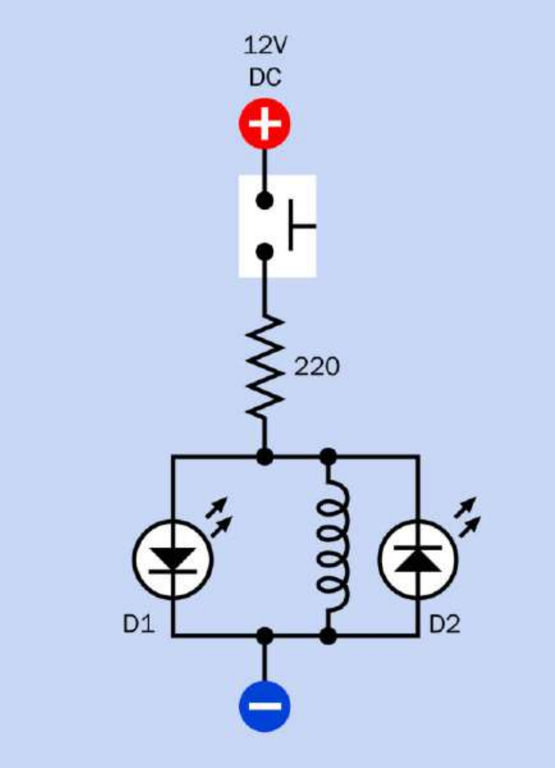

Figure 14-9. *A test circuit to demonstrate the EMF and back-EMF created when DC current starts and stops passing through a coil. See text for details.*

the negative end of the coil (north and negative both beginning with letter N). This mnemonic only works if conventional (positive) current flows through a coil that is wound clockwise.

Another model is the "corkscrew rule" in which we imagine conventional DC current flowing from the handle of a corkscrew, down through its metal section, toward the pointed end. If the corkscrew is turned clockwise, in the same direction as the electricity, the corkscrew will sink into the cork in the same as the direction as the resulting magnetic force.

Variants

Variants include core materials, core shapes, termination style (for through-hole mounting in perforated board, or for surface-mount), and external finish (some inductors are dipped in insulating material, while others allow their copper magnet wire to be exposed).

In addition there are two functional variants: variable inductors and ferrite beads. Their schematic symbols are shown in Figure 14-10.

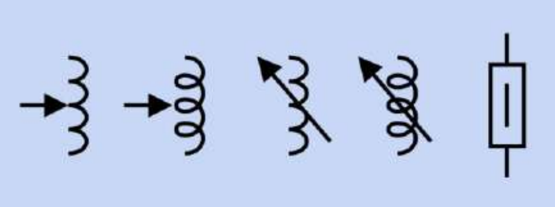

Figure 14-10. *Schematic symbols for a ferrite bead (farthest right) and variable inductors (all other symbols, which are functionally identical).*

Magnetic Cores

A magnetic core may be made from solid iron, plates of iron or steel separated by thin insulating material, powdered iron mixed with a binder, or a ferrite compound derived from nickel, zinc, manganese, or a combination. An iron core has at least 1,000 times the permeability of air, while some ferrites are 10,000 times as permeable.

One major disadvantage of a magnetic core is *hysteresis*, which in this context refers to the tendency of the core to retain some magnetic "memory" as a cycle of alternating current changes from positive to negative. This residual magnetism must be overcome by the next positive pulse of AC. The tendency of the core to retain magnetic polarity is known as its *retentivity*. Iron cores are especially retentive.

Another disadvantage of some magnetic cores is that they may host eddy currents induced by the magnetic field of the coil. These electrical currents tend to circulate through the core, reducing efficiency by generating waste heat, especially if coil currents are high. Forming a core from iron or steel plates, separated by thin layers of insulation, will inhibit these currents. Powdered iron inhibits eddy currents because the particles have limited contact. Ferrites are nonconductive, and are therefore immune to eddy currents. They are widely used.

Hysteresis and eddy currents both incur energy losses with each AC cycle. Therefore, the losses increase linearly as the AC frequency increases. Consequently, inductor cores that suffer either of these problems are not well-suited to high frequencies.

Nonmagnetic Cores

The problems associated with magnetic cores may be avoided by winding the coil around a nonmagnetic core that may be hollow, ceramic, or plastic. A hollow core is referred to as an *air core*. The permeability of ceramic and plastic cores is close to that of air.

An inductor with a nonmagnetic core will be immune to eddy currents and retentivity, but will have to be significantly larger than a magnetic-cored coil with comparable inductance. In the case of a very primitive radio receiver, such as a crystal set, the air-cored coil that selects a radio frequency may be several inches in diameter. A basic circuit diagram for a crystal set (so-called because it uses a diode containing a germanium crystal) is shown in Figure 14-11. The antenna, at top, receives signals broadcast from radio stations. The coil can be tapped (as indicated by the black dots) as a simple way to select different inductance values, blocking all but a narrow range of frequencies. The T-shaped white component at right is a high-impedance earphone. The diode blocks the lower half of the alternating current in a radio signal, and since the signal is *amplitude-modulated*, the earphone responds to variations in intensity in the signal and reproduces the sound encoded in it.

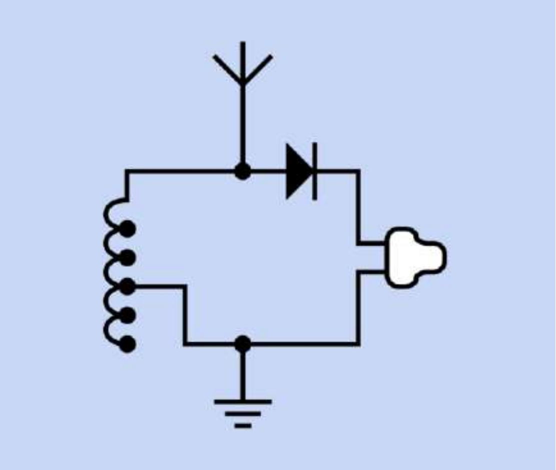

Figure 14-11. *An early and basic application for an inductor is to select radio-station frequencies, as in this schematic for a crystal set. See text for details.*

Variable Inductors

A *variable inductor*, also known as an *adjustable inductor*, is relatively uncommon but can be fabricated by using a magnetic core that penetrates the center of the inductor on an adjustable screw thread. The inductance of the assembly will increase as a larger proportion of the magnetic core penetrates into the open center of the coil. A photograph of a variable inductor is at Figure 14-12.

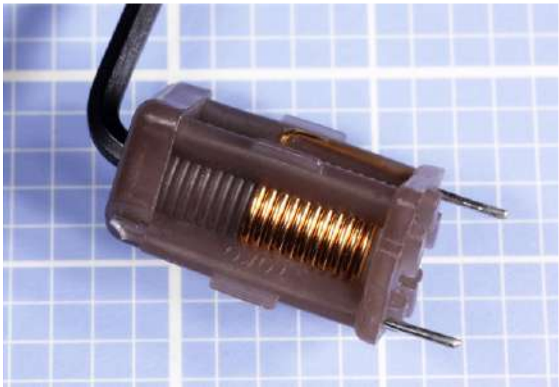

Figure 14-12. *A variable inductor. Its inductance is adjusted via a screw thread that varies the insertion of the core in the coil. In this component the core is turned by inserting a hex wrench, as shown. It is rated from 0.09µH to 0.12µH.*

Ferrite Beads

A *ferrite bead* inverts the design of a typical inductor by running a wire through a hole in the center of the bead, instead of coiling the wire around the core. Two ferrite beads are shown in Figure 14-13. At top, the bead is divided into two sections, each mounted in one-half of a plastic clam shell, which can be closed around a wire. At bottom, the bead must be threaded onto a wire. The purpose is either to limit radio-frequency radiation from a wire by absorbing it into the bead (where it is transformed into heat), or to protect a wire from external sources of radio-frequency radiation. Computer cabling to external devices; lamp dimmers; and some types of motors can be sources of radio frequency.

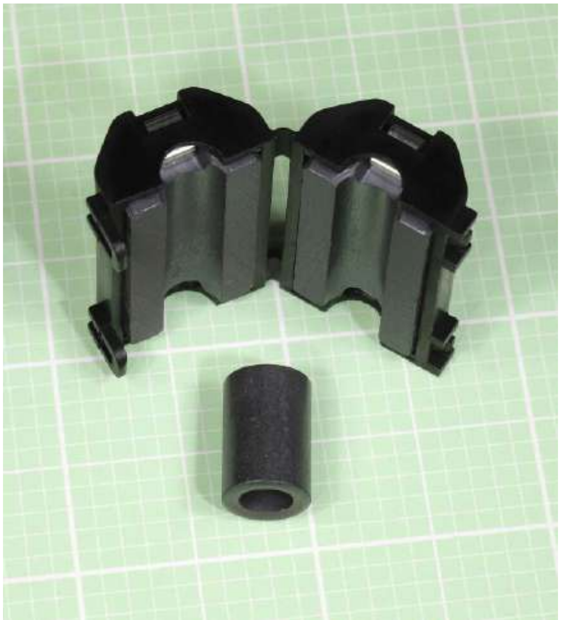

Figure 14-13. *Two examples of ferrite beads. They can inhibit radio-frequency radiation from a wire, or protect the wire from interference.*

Toroidal Cores

The magnetic circuit created by a rod-shaped core must be completed by the lines of force traveling back around from one end of the rod to the other, through the surrounding air. Since air has low permeability, this is a major source of inefficiency. By comparison, a torus (a geometrical shape resembling a donut) completes the entire magnetic circuit inside its core. This significantly increases its efficiency. Also, because its field is better contained, a toroidal inductor needs little or no shielding to protect other components from stray magnetic effects.

Two through-hole toroidal inductors are shown in Figure 14-2. Bottom left: Rated at 345μH. Bottom right: Rated at 15μH. The one at bottom-left has pins beneath it for insertion into a printed circuit board.

Surface-mount inductors often are toroidal to maximize the efficiency of a component that has to function on a very small scale. Examples are shown in Figure 14-14, Figure 14-15, and Figure 14-16.

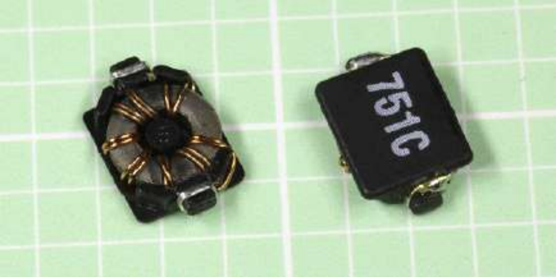

Figure 14-14. *In a typical toroidal inductor, the coil is wrapped around a magnetic core shaped as a torus. This surface-mount component (viewed from the bottom, at left, and from the top, at right) is at the low end of the range of component sizes. It is rated at 750nH.*

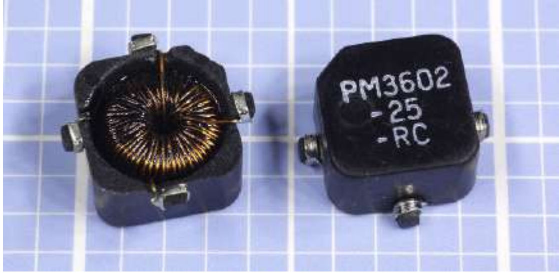

Figure 14-15. *A medium-sized surface-mount toroidal inductor (viewed from the bottom, at left, and from the top, at right). It is rated at 25μH.*

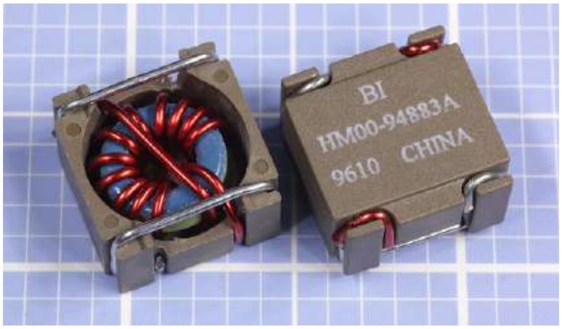

Figure 14-16. *A larger-sized surface-mount toroidal inductor (viewed from the bottom, at left, and from the top, at right). It is rated at 3.8µH.*

A chart showing some types of inductor cores, their commonly available inductances, and their maximum frequencies is shown in Figure 14-17.

Gyrator

A *gyrator* is a small network, sometimes encapsulated in a silicon chip, using **resistors**, a semiconductor, and a **capacitor** to simulate some but not all of the behavior of a coil-based inductor. The semiconductor may be a transistor or a capacitor, depending on the specific circuit. A sample schematic is shown in Figure 14-18. Because no magnetic effects are induced, the gyrator is completely free from the problems of saturation and hysteresis, which affect coils with cores, and also produces no back-EMF. It simply attenuates a signal initially, and then gradually lowers its reactance, thus imitating this aspect of an inductor.

A gyrator may be used where a coil may be unacceptably large (as in a cellular phone) or where signal quality is of paramount importance—for example, in a graphic equalizer or other audio components that perform signal processing at input stages, such as preamplifiers.

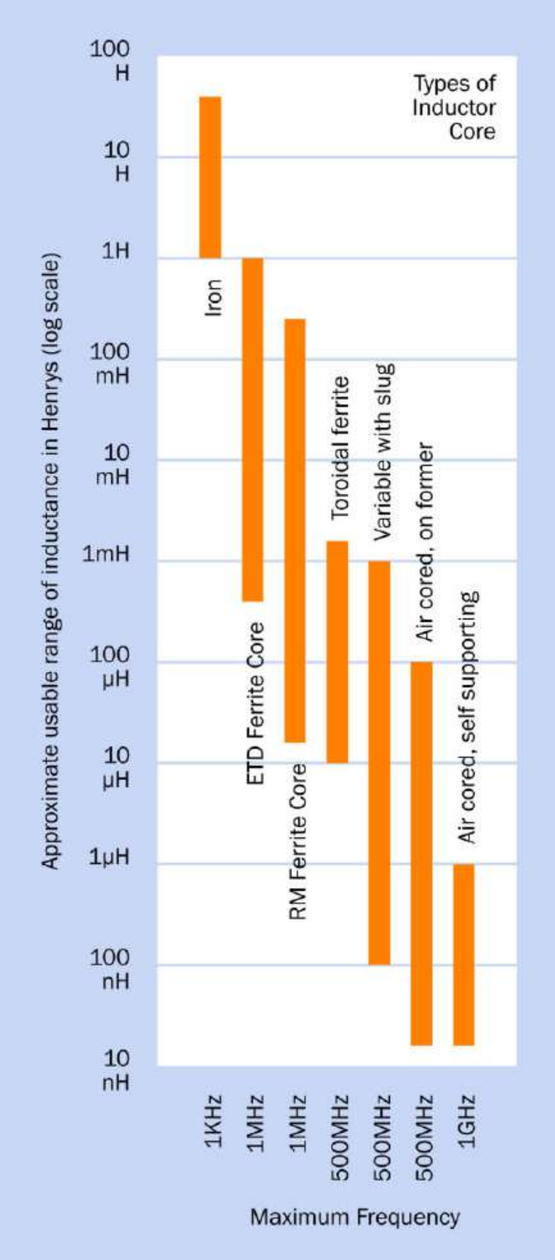

Figure 14-17. *Some commonly used inductor cores and their characteristics. Adapted from "Producing wound components" by R.Clark@surrey.ac.uk.*

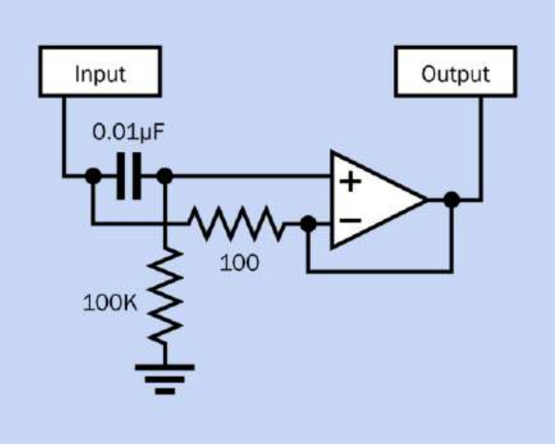

Figure 14-18. *A possible schematic for a coil substitute known as a gyrator, which may be used where a conventional coil would be unacceptably bulky.*

A gyrator does impose some limits on circuit design. While neither side of a real inductor needs to be at ground potential, a gyrator does require a ground connection. However, the performance advantages of gyrators are significant, as they can emulate high inductance without parasitic effects, can be more accurately calibrated (leading to more predictable performance), and do not create magnetic fields that can interfere with other components.

Values

Calculating Inductance

The magnetic inductance of a coil is measured with a unit known as the Henry, named after Joseph Henry, a pioneer in electromagnetism. It is defined by imagining a coil in which current is fluctuating, causing the creation of EMF. If the rate of fluctuation is 1 amp per second and the induced EMF is 1 volt, the inductance of the coil is 1 Henry.

The letter L is commonly used to represent inductance. To derive a useful formula, L will be expressed in microhenrys. If D is the diameter of a coil, N is the number of turns of wire, and W is the width of coil (when the windings are viewed from the side, as shown in Figure 14-19), the precise relationship of the variables is complex but can be reduced to an approximate formula:

$$L = \text{approx } (D^2 * N^2) / 18 * D) + (40 * W$$

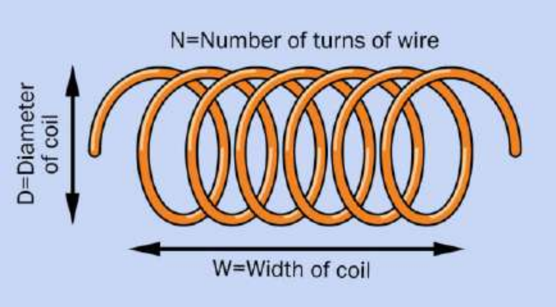

Figure 14-19. *Dimensions of a coil, referenced by a formula to calculate its approximate inductance. See text for details.*

From this, it is clear that inductance tends to increase with coil diameter, and also increases (more significantly) with the square of the number of turns. If the number of turns remains constant, inductance will be higher for a coil that is short and fat than for a coil that is narrow and long.

Because the Henry is a large unit, inductors in electronics circuits typically have their inductances measured in millihenrys (mH), microhenrys (μH), and nanohenrys (nH), where 1H = 1,000mH, 1 mH = 1,000μH, and 1μH = 1,000 nH. This relationship is shown in Figure 14-20.

Calculating Reactance

The reactance of an inductor (that is, its dynamic resistance to alternating current) varies with the frequency of the current. If f is the AC frequency (in Hertz), and L is the inductance (in Henrys), the reactance, X_L in ohms, is given by the formula:

$$X_L = 2 * \pi * f * L$$

From this equation, it's apparent that as the frequency tends toward zero (DC current), or if the inductance tends toward zero (a short piece of

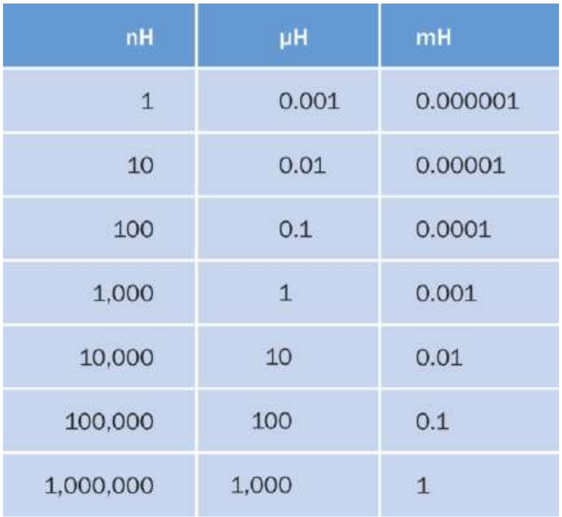

nH	µH	mH
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 14-20. *Inductance is typically measured in nanohenrys (nH), microhenrys (µH), and millihenrys (mH). Equivalent values in these units are shown here.*

straight wire), the reactance will tend toward zero. Conversely, the inductor will impede current increasingly as the frequency and/or the inductance increases.

Calculating Reluctance

The letter S is often used to represent reluctance, while Greek letter µ customarily represents permeability (not to be confused with the use of µ as a multiplication factor of 1/1,000,000, as in µF, meaning "microfarad"). If A is the area of cross-section of the magnetic circuit and L is its length:

```
S = L / µ * A
```

Datasheet Terminology

A typical manufacturer's datasheet should include an *inductance index* for an inductor, expressed in µH per 100 turns of wire (assuming the wire is in a single layer) for inductors with a powdered iron core, and mH per 1,000 turns of wire for inductors with ferrite cores.

The *DCR* is the DC resistance of an inductor, derived purely from the wire diameter and its length.

The *SRF* is the *self-resonant frequency*. An inductor should be chosen so that AC current passing through it will never get close to that frequency.

ISAT (or I_{sat}) is the saturation current, which results in a magnetic core losing its function as a result of magnetic saturation. When this occurs, inductance drops and the charge current rate increases drastically.

Series and Parallel Configurations

Because the inductance of a coil conducting DC current is proportional to the current, the calculations to derive the total inductance of coils in series or in parallel are identical to the calculations used for resistors.

In series, all the coils inevitably pass the same current, and the total inductance is therefore found by summing the individual inductances. When coils are wired in parallel, the current distributes itself according to the inductances; therefore, if L1 is the reluctance of the first coil, L2 is the reluctance of the second coil, and so on, the total reluctance L of the network is found from the formula:

```
1/L = 1/L1 + 1/L2 + 1/L3. . .
```

This is shown in Figure 14-21. In reality, differences between the coils (such as their electrical resistance), and magnetic interaction between the coils, will complicate this simple relationship.

Time Constant

Just as the *time constant* of a capacitor defines the rate at which it accumulates voltage when power is applied through a resistor, the time constant of an inductor defines the rate at which it gradually allows amperage to pass through it, overcoming the EMF generated by the coil. In both cases, the time constant is the number of seconds that the component requires to acquire approximately 63% of the difference between its current value and its maximum value. In the case of an inductor, suppose we assume zero internal resistance in the power source, zero resistance in

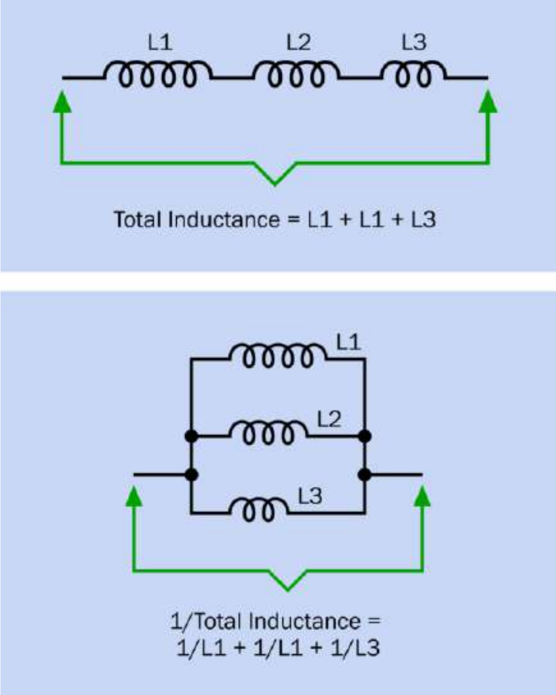

Figure 14-21. *Calculating the total inductance of inductors in parallel (top) and series (bottom).*

the coil windings, and an initial current of zero. If L is the inductance of the coil and R is the value of the series resistor, then the time constant—TC—is given in seconds by the formula

$$TC = L / R$$

Therefore a coil of 10 millihenrys (0.01 Henry) in series with a 100-ohm resistor will pass 63% of the full current in 0.0001 seconds, or 1/10 of a millisecond; it will take an equal additional amount of time for the current to rise by another 63% of the remaining difference between its charge and the maximum amperage of the circuit. In theory, the reactance of a coil can never diminish to zero, but in practice, five time constants are considered adequate to allow maximum current flow.

How to Use it

Because the inductance of an inductor peaks as current increases, and then gradually diminishes, an inductor can be used to block or attenuate high frequencies. A circuit that does this is often referred to as a *low-pass filter*. The schematic and a graph suggesting its performance are shown in Figure 14-22. A basic application could be the *crossover network* in a loudspeaker system, where high-frequency signals are blocked from a low-frequency driver and are diverted to a high-frequency driver.

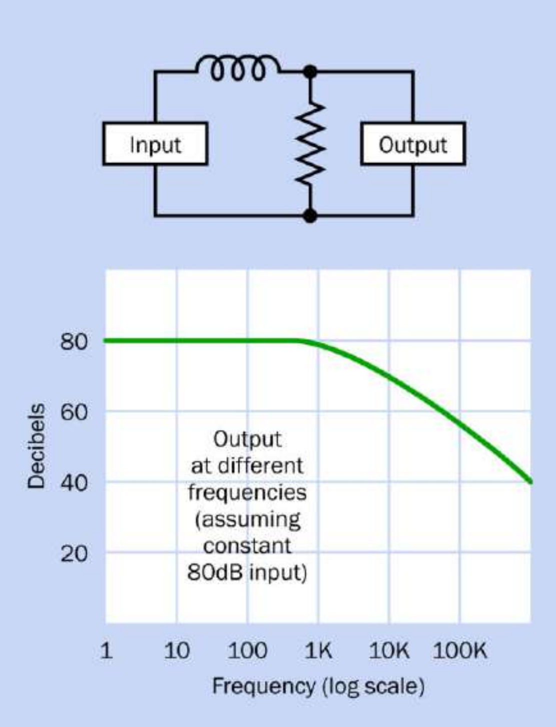

Figure 14-22. *By using the ability of an inductor to block a range of frequencies, a low-pass filter blocks higher frequencies.*

If the location of the inductor is shifted so that it shunts the signal away from the output, the results are reversed, and the circuit becomes a *high-pass filter*. The schematic and a graph suggesting its performance are shown in Figure 14-23.

Note that **capacitors** may also be used to create frequency filters, but because their function is roughly inverse to that of inductors, the place-

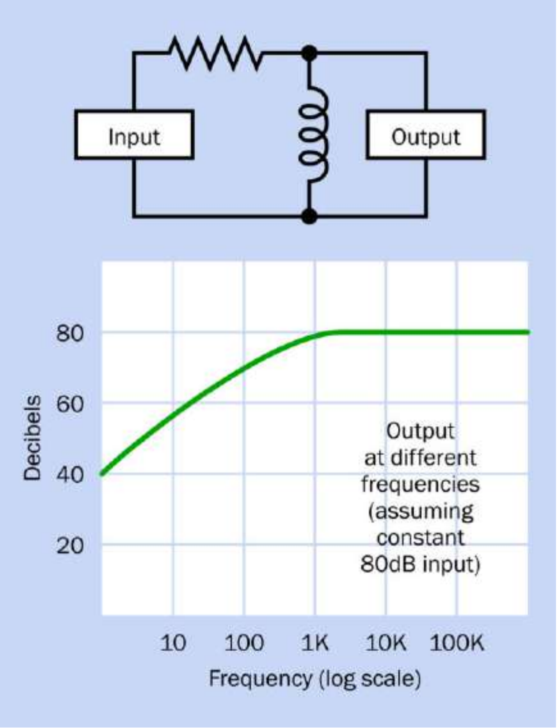
Input
Output
80
60
40
20
Decibels
Output at different frequencies (assuming constant 80dB input)
10
100
1K
10K
100K
Frequency (log scale)

Figure 14-23. *Here the inductor diverts low frequencies away from the output, allowing high frequencies to pass through.*

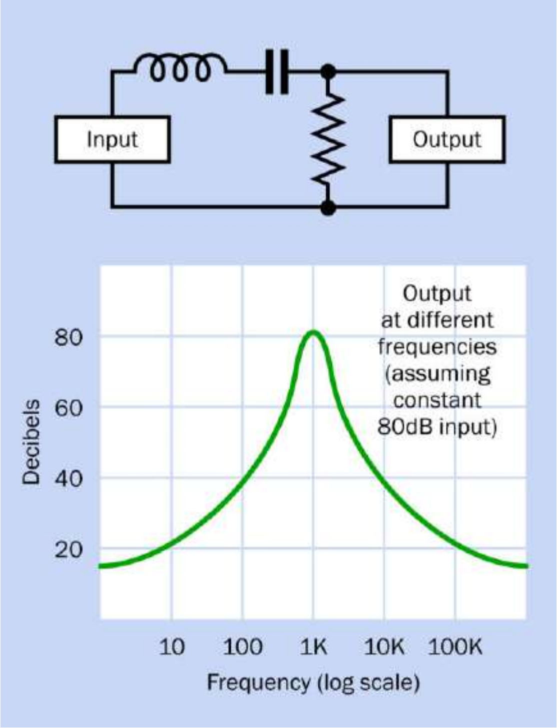
Input
Output
80
60
40
20
Decibels
Output at different frequencies (assuming constant 80dB input)
10
100
1K
10K
100K
Frequency (log scale)

Figure 14-24. *If the values of a capacitor and an inductor are correctly chosen, and the components are placed in series, the inductor blocks high frequencies while the capacitor blocks low frequencies, creating a bandpass filter, in which only a narrow band of frequencies can get through.*

ment of a capacitor in a circuit would be opposite to the placement of the inductor. Examples of filter circuits using capacitors are found in the entry for that component in this encyclopedia.

An inductor can be combined with a capacitor to form a *bandpass filter*, as shown in Figure 14-24. In this configuration, the inductor blocks the high frequencies while the capacitor blocks the low frequencies, allowing only a limited band of frequencies to get through.

Once again if the location of the components is shifted to shunt the signal away from the output, the results are reversed, as shown in Figure 14-25. This is known as a *notch filter*.

The performance of these filters will depend on the component values, and in most applications, additional components will be necessary. Sophisticated filter circuits are outside the scope of this encyclopedia.

Inductors are of great importance in **DC-DC converters** and **AC-DC power supplies** where voltage changes are enabled by rapid switching. See the relevant entries of this encyclopedia for additional details.

Generally, as electronic equipment has become increasingly miniaturized, the unavoidable bulk of inductors has limited their application. However they may still be used to tune oscillators, to block sudden spikes in power supplies, and to protect equipment from sudden voltage spikes (they are used, for example, in surge suppressors for computing equipment).

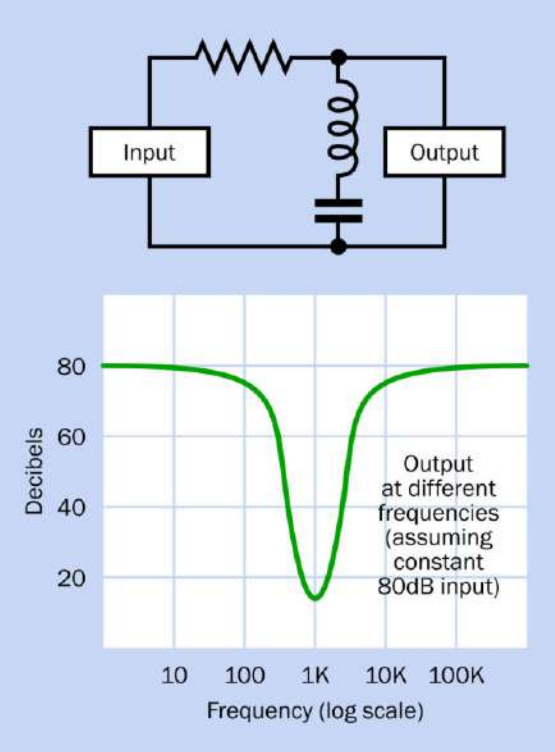
Input
Output
Decibels
80
60
40
20
Output at different frequencies (assuming constant 80dB input)
10
100
1K
10K
100K
Frequency (log scale)

Figure 14-25. *Here the capacitor and inductor block all frequencies except a narrow band, which they divert from the output. The result is a notch filter.*

Core Choices

Air-cored inductors have relatively low inductance, because of their low permeability. However, they can be operated at very high frequencies up to the gigahertz range, and can tolerate higher peak currents.

Inductors with an iron core suffer increasing power losses due to hysteresis and eddy currents as the AC frequency passing through the inductor increases. Consequently, iron-cored inductors are not suitable for frequencies much above 10KHz.

Miniaturization

A low-value inductor can be formed by etching a spiral onto a circuit board, in applications where size must be minimized. They may also be incorporated in integrated circuit chips. However, in small devices such as cellular phones, it is more common to use a coil substitute such as a *gyrator*, as described previously.

What Can Go Wrong

Real-World Defects

The theoretically ideal inductor has no resistance or capacitance and suffers no energy losses. In reality, an inductor possesses both resistance and capacitance, also creates electrical noise, and may pick up electrical noise. It tends to create stray magnetic fields, and generally is more troublesome to deal with than its two cousins, the **resistor** and the **capacitor**.

Parasitic capacitance occurs between adjacent turns of wire. This capacitance becomes more significant at higher frequencies, leading ultimately to a situation where the coil becomes *self resonant*.

The workarounds for these problems involve coil geometries and choices of core material that go beyond the scope of this encyclopedia.

A *gyrator* should be considered as a possible substitute where inductors are troublesome or excessively expensive.

Saturation

Inductance increases as the current passing through a coil increases, but if a magnetic core is used, its contribution to inductance will stop abruptly when the core becomes magnetically *saturated*. In other words, when all of the randomly distributed *magnetic domains* in the core have been induced to align themselves with the pervasive magnetic field, the core cannot become more highly magnetized, and ceases to contribute to the inductance. Note that as a core approaches saturation levels, its hysteresis increases because reversing its magnetization requires greater energy. Antidotes to saturation would include a larger core, a lower current, a smaller number of turns in the coil, and using a core with lower permeability (such as air).

RF Problems

Radio frequencies (*RF*) introduce various problems affecting the efficiency of inductors. The *skin effect* is the tendency of high-frequency AC current to flow primarily on the surface of a strand of wire. The *proximity effect* refers to the tendency of the magnetic fields caused by adjacent wires to introduce eddy currents in the coil. Both of these effects increase the effective resistance of the coil. Various coil geometries have been developed to minimize these effects, but are outside the scope of this encyclopedia. The fundamental lesson is that coils specifically designed for RF are the only ones that should be used with RF.

AC-AC transformer

15

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See Chapter 16)
- **DC-DC converter** (See Chapter 17)
- **DC-AC inverter** (See Chapter 18)

What It Does

A transformer requires an input of *alternating current* (AC). It transforms the input voltage to one or more output voltages that can be higher or lower.

Transformers range in size from tiny impedance-matching units in audio equipment such as microphones, to multi-ton behemoths that supply high voltage through the national power grids. Almost all electronic equipment that is designed to be powered by municipal AC in homes or businesses requires the inclusion of a transformer.

Two small power transformers are shown in Figure 15-1. The one at the rear is rated to provide 36VAC at 0.8A when connected with a source of 125VAC. At front, the miniature transformer is a Radio Shack product designed to provide approximately 12VAC at 300mA, although its voltage will be more than 16VAC when it is not passing current through a load.

Transformer schematic symbols are shown in Figure 15-2. The different coil styles at left and right are functionally identical. Top: A transformer with a magnetic core—a core that can be magnetized. Bottom: A transformer with an air core. (This type of transformer is rare, as it tends to be less efficient.) The input for the transformer is almost always assumed to be on the left, through the *primary* coil, while the output is on the right, through the *secondary* coil. Often the two coils will show differing numbers of turns to indicate whether the transformer is delivering a reduced voltage (in which case there will be fewer turns in the secondary coil) or an increased voltage (in which case there will be fewer turns in the primary coil).

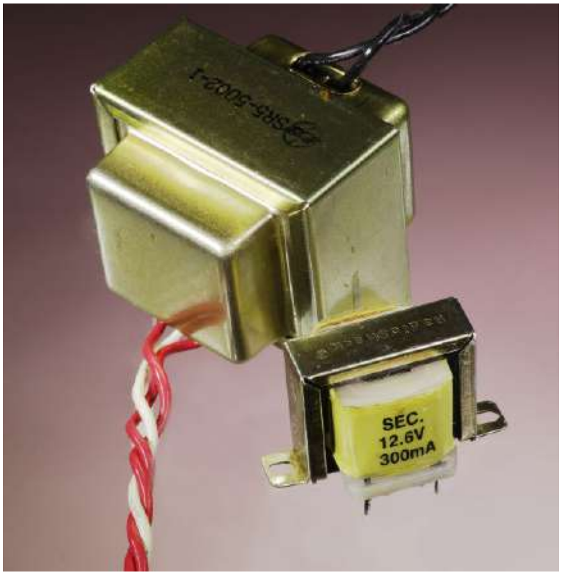

Figure 15-1. *Two small power transformers. The one at the rear measures approximately 1" × 2" × 2" and is rated to provide 36VAC at 0.8A. The term "sec" on the smaller unit is an abbreviation for "secondary," referring to the rating for its secondary winding.*

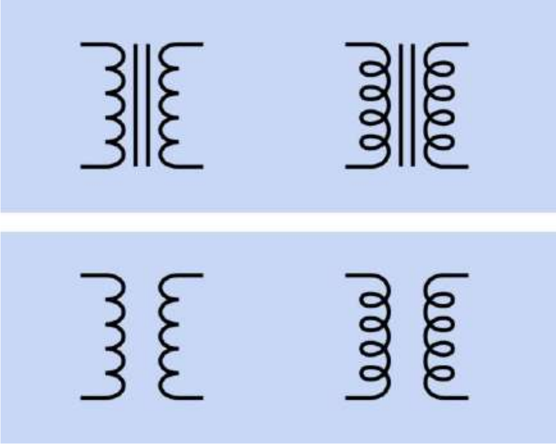

Figure 15-2. *Alternate symbols for a transformer with a ferromagnetic core (top) and air core (bottom). The differing coil symbols at left and right are functionally identical.*

How It Works

A simplified view of a transformer is shown in Figure 15-3. Alternating current flowing through the primary winding (orange) induces magnetic flux in a laminated core formed from multiple steel plates. The changing flux induces current in the secondary winding (green), which provides the output from the transformer. (In reality, the windings usually consist of thousands of turns of thin magnet wire, also known as enameled wire; and various different core configurations are used.)

The process is known as *mutual induction*. If a load is applied across the secondary winding, it will draw current from the primary winding, even though there is no electrical connection between them.

In an ideal, lossless transformer, the ratio of turns between the two windings determines whether the output voltage is higher, lower, or the same as the input voltage. If V_p and V_s are the voltages across the primary and secondary windings respectively, and N_p and N_s are the number of turns of wire in the primary and secondary windings, their relationship is given by this formula:

$$V_p \ / \ V_s = N_p \ / \ N_s$$

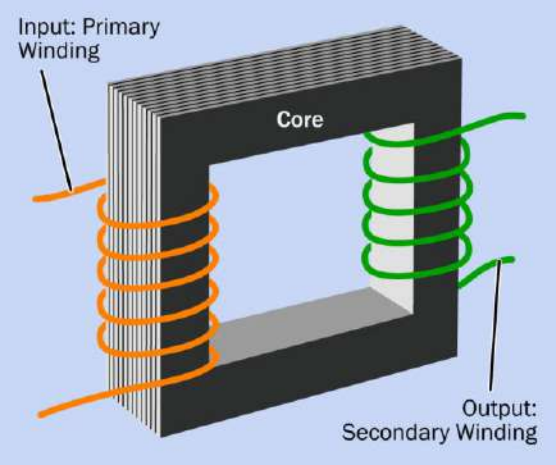

Figure 15-3. *Three basic parts of a transformer, shown in simplified form.*

A simple rule to remember is that fewer turns = lower voltage while more turns = higher voltage.

A *step-up transformer* has a higher voltage at its output than at its input, while a *step-down transformer* has a higher voltage at its input than at its output. See Figure 15-4.

In an ideal, lossless transformer, the power input would be equal to the power output. If V_{in} and V_{out} are the input and output voltages, and I_{in} and I_{out} are the input and output currents, their relationship is given by this formula:

$$V_{in} * I_{in} = V_{out} * I_{out}$$

Therefore, if the transformer doubles the voltage, it allows only half as much current to be drawn from the secondary winding; and if the voltage is cut in half, the available current will double.

Transformers are not 100% efficient, but they can be more than 98% efficient, and relationships between voltage, current, and the number of turns in the windings are reasonably realistic.

When the transformer is not loaded, the primary winding behaves like a simple inductor with reactance that inhibits the flow of current. Therefore a power transformer will consume relatively

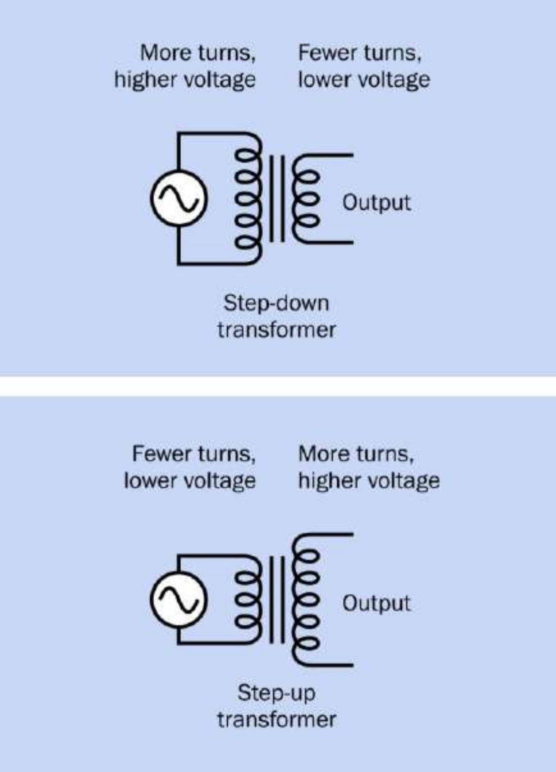
More turns, higher voltage
Fewer turns, lower voltage
Output
Step-down transformer
Fewer turns, lower voltage
More turns, higher voltage
Output
Step-up transformer

Figure 15-4. *The ratio of input voltage to output voltage is equal to the ratio of primary turns to secondary turns in the transformer windings, assuming a transformer of 100% efficiency.*

little electricity if it is left plugged in to an electrical outlet without any load connected to its output side. The power that it does consume will be wasted as heat.

The Core

The ferromagnetic core is often described as being made of iron, but in reality is more often fabricated from high permeability silicon steel. To reduce losses caused by eddy currents, the core is usually laminated—assembled from a stack of plates separated from each other by thin layers of varnish or a similar insulator. Eddy currents tend to be constrained within the thickness of each plate.

Because a DC voltage would cause magnetic saturation of the core, all transformers must operate with alternating current or pulses of current. The windings and geometry of a transformer are optimized for the frequency range, voltage, and current at which it is designed to operate. Deviating significantly from these values can damage the transformer.

Taps

A *tap* on a transformer is a connection part-way through the primary or (more often) the secondary coil. On the primary side, applying an input between the start of a coil and a tap part-way through the coil will reduce the number of turns to which the voltage is applied, therefore increasing the ratio of output turns to input turns, and increasing the output voltage. On the secondary side, taking an output between the start of a coil and a tap part-way through the coil will reduce the number of turns from which the voltage is taken, therefore decreasing the ratio of output turns to input turns, and decreasing the output voltage. This can be summarized:

- A tap on the primary side can increase output voltage.
- A tap on the secondary side can provide a decreased output voltage.

In international power adapters, a choice of input voltages may be allowed by using a double-throw switch to select either the whole primary winding, or a tapped subsection of the winding. See Figure 15-5. Modern electronics equipment often does not require a voltage adapter, because a **voltage regulator** or **DC-DC converter** inside the equipment will tolerate a wide range of input voltages while providing a relatively constant output voltage.

A transformer's secondary winding is often tapped to provide a choice of output voltages. In fact, most power transformers have at least two outputs, since the cost of adding taps to the secondary winding is relatively small. As an

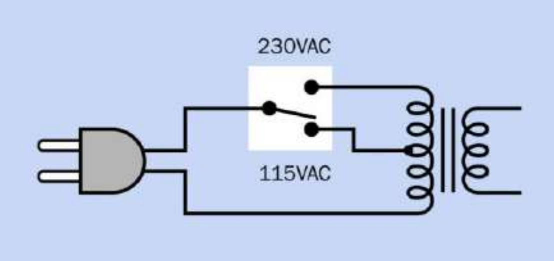

Figure 15-5. *An international power adapter can provide a fixed output voltage by using a double-throw switch to apply 230VAC voltage across a transformer's primary winding, or 115VAC to a tapped midpoint of the primary winding.*

alternative to tapped outputs, two or more separate secondary windings may be used, allowing the outputs to be electrically isolated from each other. See Figure 15-6.

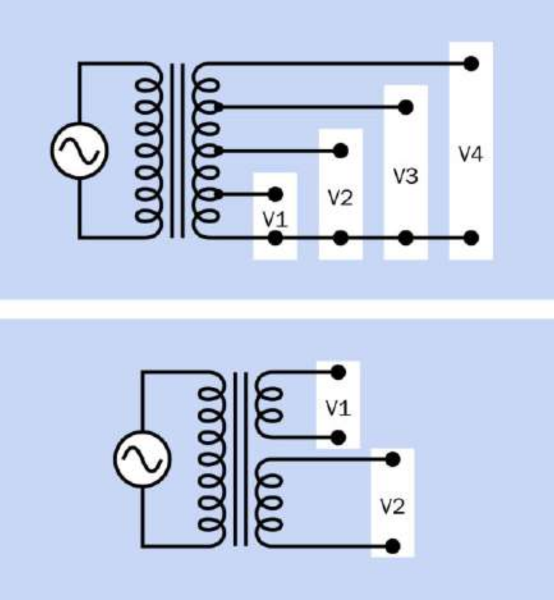

Figure 15-6. *Multiple output voltages may be obtained from a transformer by tapping into the secondary winding (top) or using two or more separate secondary windings (bottom), in which case the outputs will be electrically isolated from each other.*

If the winding on the primary side of a transformer is coiled in the same direction as the winding on the secondary side, the output voltage will be 180 degrees out of phase with the input voltage. In schematics, a dot is often placed at one end of a transformer coil to indicate where the coil begins. If the dots on the primary and secondary sides are at the same ends of the coils, there will be a 180 degree phase difference between input and output. For many applications (especially where the output from a power transformer is going to be converted to DC), this is immaterial.

If there is a center tap on the secondary winding, and it will be referenced as ground, the voltages relative to it, at opposite ends of the secondary winding, will be out of phase. See Figure 15-7.

Variants

Core Shapes

The *shell core* is a closed rectangle, as shown in Figure 15-3. This is the most efficient but most costly to manufacture. A C-shaped core is another option (three sides of the rectangle) and an E-I core is popular, consisting of a stack of E-shaped plates with two coils wound around the top and bottom legs of the E, or wound concentrically around the center leg of the E. An additional stack of straight plates is added to close the gaps in the E and form a magnetic circuit.

In Figure 15-8, the small transformer from Figure 15-1 has been sliced open with a band saw and a belt sander to reveal a cross-section of its windings. This clearly shows that its primary and secondary windings are concentric. It also reveals the configuration of its core, which is in the E-I format. In Figure 15-9, the E-I configuration is highlighted to show it more clearly.

Power Transformer

Typically designed to be bolted onto a chassis or secured inside the case or cabinet housing a piece of electrical equipment with solder tabs or connectors allowing wires to connect the transformer to the power cord, on one side, and a cir-

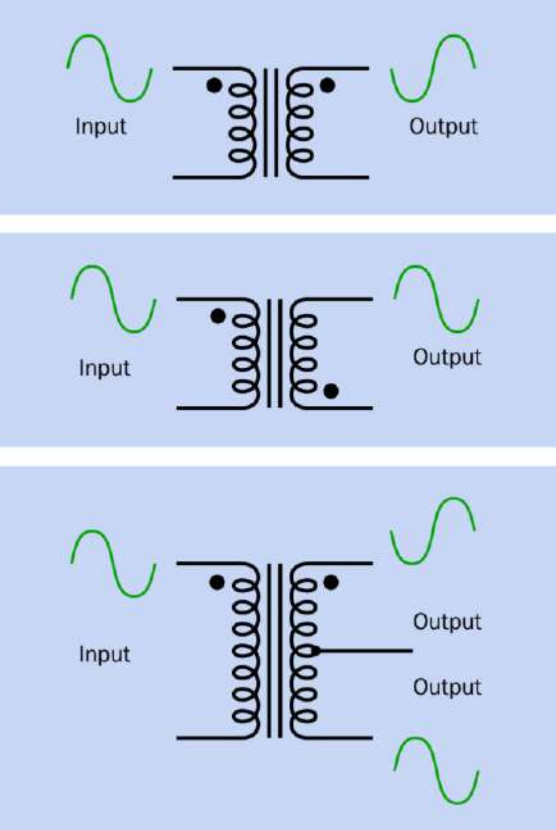

Figure 15-7. *A dot indicates the start of each winding. Where primary and secondary windings are in the same direction, the voltage output will be 180 degrees out of phase with the input. Where the dots indicate windings in opposite directions, the voltage output will have the same phase as the input. Where a center tap on the secondary winding serves as a common ground, the voltages at opposite ends of the secondary winding will be opposite to each other in phase.*

cuit board, on the other side. Smaller power transformers such as the one in Figure 15-1 have "through-hole" design with pins allowing them to be inserted directly onto circuit boards.

Plug-in Transformer

Usually sealed in a plastic housing that can be plugged directly into a wall power outlet. They are visually identical to **AC adapters** but have an AC output instead of a DC output.

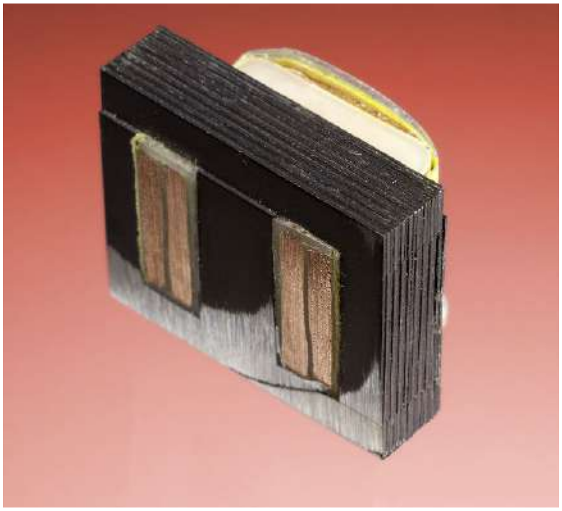

Figure 15-8. *The small transformer from the first figure in this entry is shown sliced open to reveal its internal configuration.*

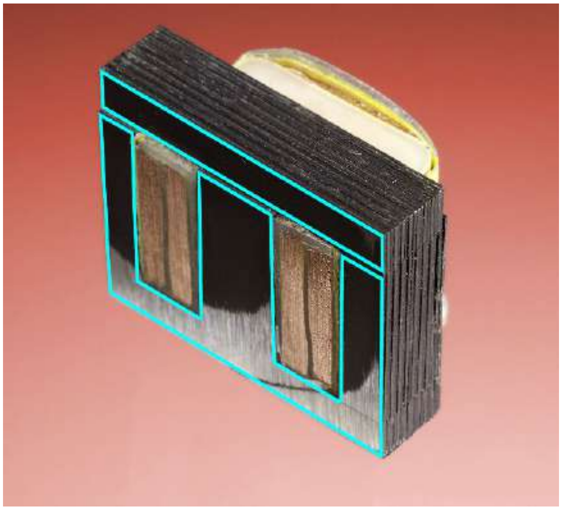

Figure 15-9. *The "EI" shaped plates that form the core of the transformer are outlined to show their edges.*

Isolation Transformer

Also known as a *1:1 transformer* because it has a 1:1 ratio between primary and secondary windings, so that the output voltage will be the same as the input voltage. When electrical equipment is plugged into the isolation transformer, it is separated from the electrical ground of AC power wiring. This reduces risk when working on "live" equipment, as there will be negligible electrical potential between itself and ground. Consequently, touching a grounded object while also touching a live wire in the equipment should not result in potentially lethal current passing through the body.

Autotransformer

This variant uses only one coil that is tapped to provide output voltage. Mutual induction occurs between the sections of the coil. An autotransformer entails a common connection between its input and output, unlike a two-coil transformer, which allows the output to be electrically isolated from the input. See Figure 15-10. Autotransformers are often used for impedance matching in audio circuits, and to provide output voltages that differ only slightly from input voltages.

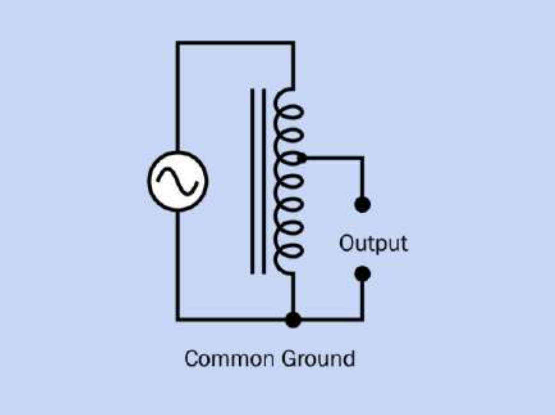

Figure 15-10. *An autotransformer contains only one coil and core. A reduced output voltage can be obtained by tapping into the coil. A common connection prevents the output from being electrically isolated from the input.*

Variable Transformer

A variable transformer, also known as a *variac*, resembles a wire-wound **potentiometer**. Only one winding is used. A *wiper* can be turned to contact the winding at any point, and serves as a movable tap. Like an autotransformer, a variable transformer entails a common connection between input and output.

Audio Transformer

When a signal is transmitted between two stages of a circuit that have different impedance, the signal may be partially reflected or attenuated. (Impedance is measured in ohms but is different from DC electrical resistance because it takes into account reactance and capacitance. It therefore varies with frequency.)

A device of low input impedance will try to draw significant current from a source, and if the source has high output impedance, its voltage will drop significantly as a result. Generally, the input impedance of a device should be at least 10 times the output impedance of the device that is trying to drive it. Passive components (resistors, and/or capacitors, and/or coils) can be used for impedance matching, but in some situations a small transformer is preferable.

If N_p and N_s are the number of turns of wire in the transformer primary and secondary windings, and Z_p is the impedance of a device (such as an audio amplifier) driving the transformer on its primary side, and Z_s is the impedance of a device (such as a loudspeaker) receiving power from the secondary side:

$$N_p / N_s = \sqrt{(Z_p / Z_s)}$$

Suppose that an audio amplifier with rated output impedance of 640Ω is driving a loudspeaker with 8Ω impedance. A matching transformer would be chosen with a ratio of primary turns to secondary turns give by:

```
√(640/8) = √80 = approximately 9:1
```

The two transformers in Figure 15-11 are through-hole components designed for telecommunications purposes, but are capable of passing audio frequencies and can be used for impedance matching in applications such as a preamplifier.

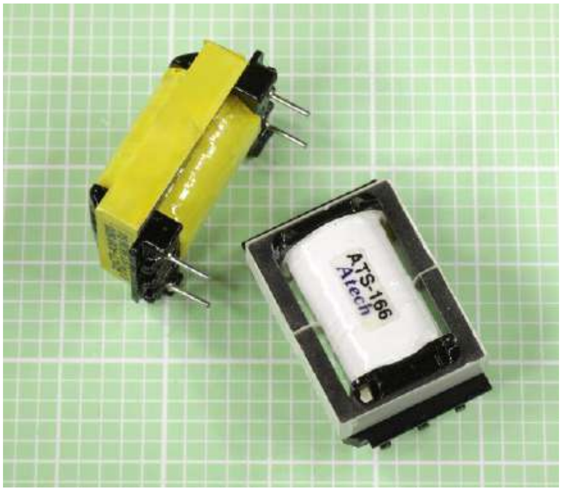

Figure 15-11. *Through-hole transformers. See text for details.*

In Figure 15-12, the transformers are designed for audio coupling. The one on the right has impedances of 500 ohms (primary) and 8 ohms (secondary). On the left is a fully encapsulated line matching transformer with a 1:1 turns ratio.

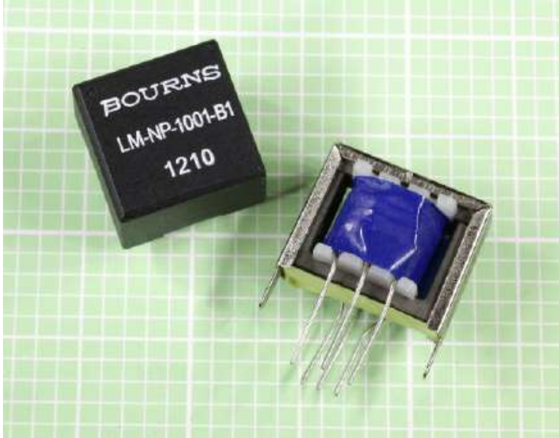

Figure 15-12. *Through-hole transformers. See text for details.*

Split-Bobbin Transformer

This variant has primary and secondary coils mounted side by side to minimize capacitive coupling.

Surface-Mount Transformer

May be less than 0.2" square and is used for impedance matching, line coupling, and filtering. Two surface-mount transformers are shown in Figure 15-13.

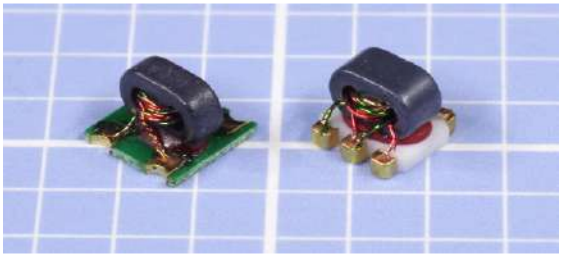

Figure 15-13. *Two surface-mount transformers, each measuring less than 0.2" square, typically used in communications equipment and suited for frequencies higher than 5 MHz.*

Values

When selecting a power transformer, its power handling capability is the value of primary interest. It is properly expressed by the term VA, derived from "volts times amps." VA should not be confused with watts because watts are measured instantaneously in a DC circuit, whereas in an AC circuit, voltage and current are fluctuating constantly. VA is actually the *apparent power*, taking reactance into account.

The relationship between VA and watts will vary depending on the device under consideration. In a worst-case scenario:

```
W = 0.65 VA (approximately)
```

In other words, the averaged power you can draw from a transformer should be no less than two-thirds of its VA value.

Transformer specifications often include input voltage, output voltage, and weight of the component, all of which are self-explanatory. Coupling transformers may also specify input and output impedances.

How to Use it

For most electronic circuits, a power transformer will be followed by a *rectifier* to convert AC to DC, and **capacitors** to smooth fluctuations in the supply. Using a prepackaged **power supply** or *AC adapter* that already contains all the necessary components will be more time-effective and probably more cost-effective than building a power supply from the ground up. See Chapter 16.

What Can Go Wrong

Reversal of Input and Output

Suppose a transformer is designed to provide an output voltage of 10 volts from domestic AC power of 115 volts. If the wrong side of the transformer is connected with 115VAC by mistake, the output will now be more than 1,000 volts—easily enough to cause death, quite apart from destroying components that are connected with it. Reversing the transformer in this way may also destroy it. Extreme caution is advisable when making connections with power transformers. A meter should be used to check output voltage. All devices containing transformers should be fused on the live side and grounded.

Shock Hazard from Common Ground

When working on equipment that uses an autotransformer, the chassis will be connected through the transformer to one side of 115VAC power. So long as a plug is used that prevents reversed polarity, the chassis should be “neutral.” However, if an inappropriate power cord is used, or if the power outlet has been wired incorrectly, the chassis can become live. For protection, before working on any device that uses 115VAC power with an autotransformer, plug the device into an isolation transformer, and plug the isolation transformer into the wall outlet.

Accidental DC Input

If DC current is applied to the input side of a transformer, the relatively low resistance of the primary coil will allow high current that can destroy the component. Transformers should only be used with alternating current.

Overload

If a transformer is overloaded, heat will be generated that may be sufficient to destroy the thin layers of insulation between coil windings. Consequently, input voltage can appear unexpectedly on the output side. Transformers with a toroidal (circular) core are especially hazardous in this respect, as their primary and secondary windings usually overlap.

Some (not all) power transformers contain a thermal fuse that melts when it exceeds a temperature threshold. If the fuse is destroyed, the transformer must be discarded.

The consequences of moderate overloading may not be obvious, and can be cumulative over time. Ventilation or heat sinkage should be taken into account when designing equipment around a power transformer.

Incorrect AC Frequency

Single-phase AC power in the United States fluctuates at 60Hz, but Great Britain and some other countries use AC power at 50Hz. Many power transformers are rated to be compatible with either frequency, but if a transformer is specifically designed for 60Hz, it may eventually fail by overheating if it is used with a 50Hz supply. (A 50Hz transformer can be used safely with 60Hz AC.)

AC-DC power supply

16

Also known as an *AC adapter*. When packaged as a palm-sized plastic package that plugs directly into a power outlet, it is occasionally known colloquially as a *wall-wart*.

OTHER RELATED COMPONENTS

- **transformer** (See Chapter 15)
- **DC-DC converter** (See Chapter 17)
- **DC-AC inverter** (See Chapter 18)

What It Does

An AC-DC power supply converts *alternating current* (*AC*) into the *direct current* (*DC*) that most electronic devices require, usually at a lower voltage. Thus, despite its name, a power supply actually requires an external supply of power to operate.

Larger products, such as computers or stereo equipment, generally have a power supply contained within the device, enabling it to plug directly into a wall outlet. Smaller battery-powered devices, such as cellular phones or media players, generally use an external power supply in the form of a small plastic pod or box that plugs into a wall outlet and delivers DC via a wire terminating in a miniature connector. The external type of power supply is often, but not always, referred to as an *AC adapter*.

Although an AC-DC power supply is not a single component, it is often sold as a preassembled modular unit from component suppliers.

Variants

The two primary variants are a *linear regulated power supply* and *switching power supply*.

Linear Regulated Power Supply

A linear regulated power supply converts AC to DC in three stages:

1. A power **transformer** reduces the AC input to lower-voltage AC.
2. A *rectifier* converts the AC to unsmoothed DC. Rectifiers are discussed in the entry on **diodes** in this encyclopedia.
3. A **voltage regulator**, in conjunction with one or more **capacitors**, controls the DC voltage, smooths it, and removes *transients*. The regulator is properly known as a *linear voltage regulator* because it contains one or more **transistors**, which are functioning in linear mode—that is, responding linearly to fluctuations in base current, at less than their saturation level. The linear voltage regulator gives the linear regulated power supply its name.

A simplified schematic of a linear regulated power supply is shown in Figure 16-1.

This type of power supply may be described as *transformer-based*, since its first stage consists of a transformer to drop the AC input voltage before it is rectified.

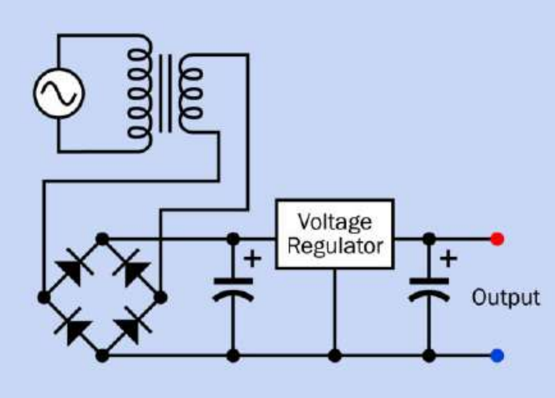

Figure 16-1. *A basic linear regulated power supply.*

Because the rectifier in a power supply generally passes each pulse of AC through a pair of silicon diodes, it will impose a voltage drop of about 1.2V at peak current. A smoothing capacitor will drop the voltage by about 3V as it removes ripple from the current, whereas a voltage regulator typically requires a difference of at least 2V between its input and its output. Bearing in mind also that the AC input voltage may fluctuate below its rated level, the output from the power transformer should be at least 8VAC higher than the ultimate desired DC output. This excess power will be dissipated as heat.

The basic principle of the linear regulated power supply originated in the early days of electronic devices such as radio receivers. A transistorized version of this type of power supply remained in widespread use through the 1990s. Switching power supplies then became an increasingly attractive option as the cost of semiconductors and their assembly decreased, and high-voltage transistors became available, allowing the circuit to run directly from rectified line voltage with no step-down power transformer required.

Some external AC adapters are still transformer-based, but are becoming a minority, easily identified by their relatively greater bulk and weight. An example is shown in Figure 16-2.

Figure 16-2. *A simple transformer-based power supply can be encapsulated in a plastic shell, ready to plug into a power outlet. However, today this format more typically contains a switching power supply, which is usually lighter, smaller, and cheaper.*

Figure 16-3 shows the handful of components inside a cheap, relatively old AC adapter. The output from a power transformer is connected directly to four diodes (the small black cylinders), which are wired as a full-wave rectifier. A single electrolytic capacitor provides some smoothing, but because there is no voltage regulator, the output will vary widely depending on the load. This type of AC adapter is not suitable for powering any sensitive electronic equipment.

Switching Power Supply

Also known as a *switched-mode power supply*, an *SMPS,* or *switcher*, it converts AC to DC in two stages.

1. A *rectifier* changes the AC input to unsmoothed DC, without a power transformer.
2. A DC-DC **converter** switches the DC on and off at a very high frequency using *pulse-width modulation* to reduce its average effective voltage. Often the converter will be the *flyback* type, containing a transformer, but the high-frequency switching allows the transformer to be much smaller than the power transformer required in a linear regulated power supply. See the DC-DC **converter** entry in this encyclopedia for an explanation of the working principles.

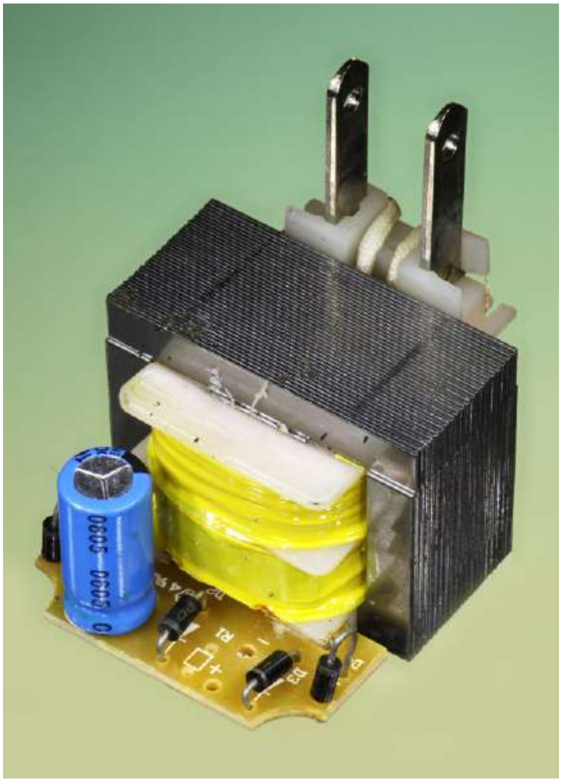

Figure 16-3. *A relatively old, cheap AC adapter contains only the most rudimentary set of components, and does not supply the kind of properly regulated DC power required by electronic equipment.*

A simplified schematic of a switching power supply is shown in Figure 16-4.

The interior of a relatively early switching power supply designed to deliver 12VDC at up to 4A is shown in Figure 16-5. This supply generated considerable waste heat, necessitating well-spaced components and a ventilated enclosure.

The type of small switching power supply that is now almost universally used to power laptop computers is shown in Figure 16-6. Note the smaller enclosure and the higher component count than in the older power supply shown in Figure 16-5. The modern unit also delivers considerably more power, and generates less waste heat. Although this example is rated at 5A, the transformer (hidden under the yellow wrapper at the center of the unit) is smaller than the power transformer that would have been found in an old-style AC adapter delivering just 500mA.

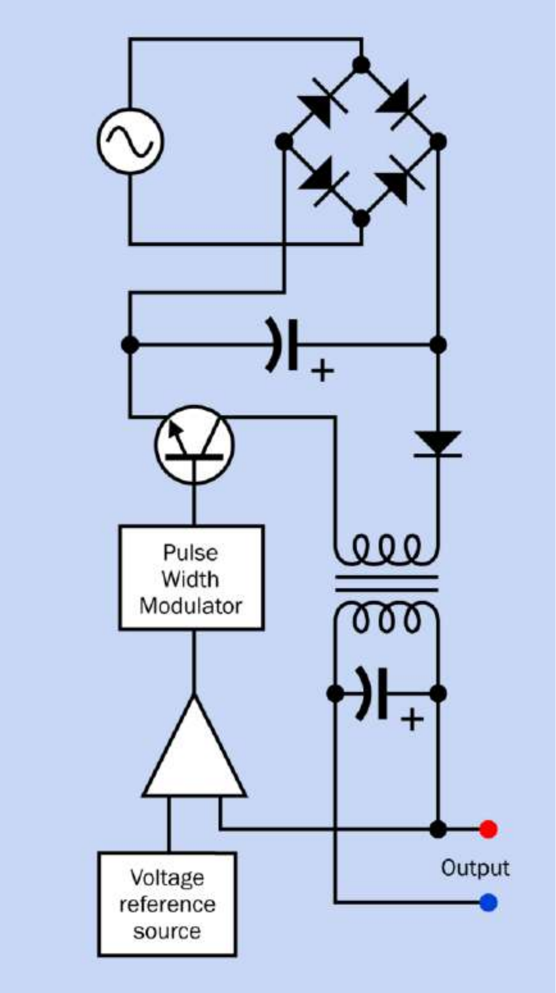

Figure 16-4. *Greatly simplified schematic showing the principal components of a switching power supply. Note the absence of a 115VAC power transformer. The transformer that is inserted subsequently in the circuit functions in conjunction with the high switching frequency, which allows it to be very much smaller, cheaper, and lighter.*

The modern power supply is completely sealed, where earlier versions required ventilation. On

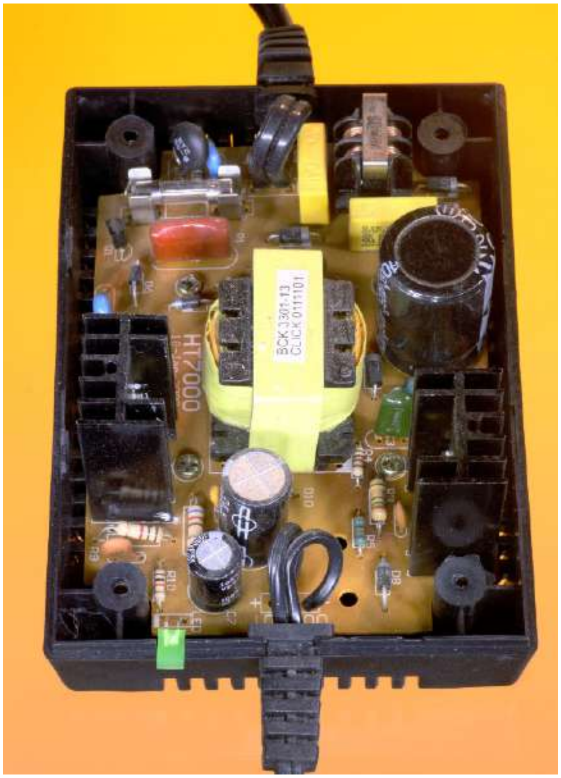

Figure 16-5. *The interior of an early switching power supply.*

the downside, the plastic case of the switching supply requires a metal liner (removed for this photograph) to contain high-frequency electromagnetic radiation.

Unregulated Power Supply

Typically this consists of a transformer and rectifying diodes with little or no smoothing or voltage control of the output.

Adjustable Power Supply

This is usually a linear power supply incorporating an adjustable **voltage regulator**. This type of supply has laboratory applications and is found as a benchtop item to power electronics design projects during their development.

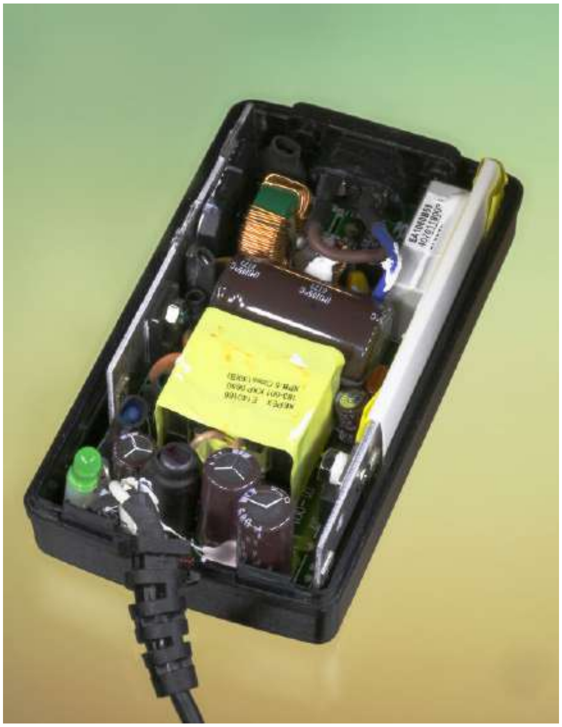

Figure 16-6. *The interior of the type of switching power supply that powers a laptop computer.*

Voltage Multiplier

Devices such as photocopiers and laser printers, televisions, cathode-ray tubes, and microwave ovens require voltages significantly higher than those supplied by domestic AC power outlets. A voltage multiplier usually contains a step-up transformer followed by DC conversion components, but detailed consideration is outside the scope of this encyclopedia.

Formats

An *open frame* power supply consists of components on a circuit board, usually mounted on a metal chassis, with no enclosure or fan cooling.

A *covered* power supply is enclosed in a protective perforated metal box with a cooling fan if needed. Power supplies sold for desktop computers are usually in this format.

Power supplies are also available in rack-mount and DIN-rail formats.

How to Use it

Because a switching power supply contains no power transformer, it is lighter and smaller, and may be cheaper than a linear power supply. It is also more efficient and generates less waste heat. These advantages have made switching power supplies the most popular option to provide DC power for electronics devices. However, the high-frequency switching tends to create *electromagnetic interference* (*EMI*), which must be filtered to protect the output of the device and also to minimize the risk of this interference feeding back into AC power wiring. The high-frequency switched power may also generate harmonics, which must be suppressed.

High-quality linear regulated power supplies still find application in laboratory equipment, low-noise signal processing, and other niches where excellent regulation and low-ripple output are necessary. They are relatively heavy, bulky, and inefficient.

See Figure 16-7 for a chart comparing the advantages and disadvantages of linear and switching power supplies.

What Can Go Wrong

High Voltage Shock

One or more capacitors in a power supply may retain a relatively high voltage for some time after the unit has been unplugged. If the power supply is opened for inspection or repairs, caution is necessary when touching components.

Capacitor Failure

If *electrolytic capacitors* fail in a switching power supply (as a result of manufacturing defects, disuse, or age), allowing straight-through conduction of alternating current, the high-frequency switching semiconductor can also fail, allowing input voltage to be coupled unexpectedly to the output. Capacitor failure is also a potential problem in linear power supplies. For additional information on capacitor failure modes, see Chapter 12.

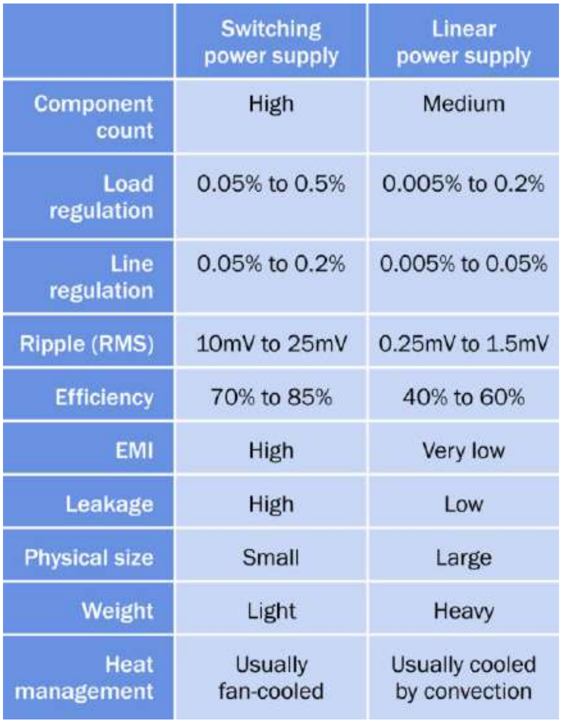

	Switching power supply	Linear power supply
Component count	High	Medium
Load regulation	0.05% to 0.5%	0.005% to 0.2%
Line regulation	0.05% to 0.2%	0.005% to 0.05%
Ripple (RMS)	10mV to 25mV	0.25mV to 1.5mV
Efficiency	70% to 85%	40% to 60%
EMI	High	Very low
Leakage	High	Low
Physical size	Small	Large
Weight	Light	Heavy
Heat management	Usually fan-cooled	Usually cooled by convection

Figure 16-7. *Comparison of attributes of linear regulated power supplies and switching power supplies. (Adapted from Acopian Technical Company.)*

Electrical Noise

If *electrolytic capacitors* are used, their gradual deterioration over time will permit more electrical noise associated with high-frequency switching in a switching power supply.

Peak Inrush

A switching power supply allows an initial inrush or surge of current as its capacitors accumulate their charge. This can affect other components in the circuit, and requires fusing that tolerates brief but large deviations from normal power consumption.

DC-DC converter

17

Often referred to as a *switching regulator*, and sometimes as a *switcher*, not to be confused with a *switching power supply*.

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See Chapter 16)
- **voltage regulator** (See Chapter 19)
- **DC-AC inverter** (See Chapter 18)

What It Does

A DC-DC converter, often referred to simply as a *converter*, receives a DC voltage as its input and converts it to a regulated DC voltage as its output. The output voltage may be higher or lower than the input voltage, may be user-adjustable by adding an external resistor, and may be completely electrically isolated from the input, depending on the type of converter that is used. The overall efficiency is not greatly affected by the difference between input and output voltage, and can exceed 90%, minimizing waste heat and enabling the unit to be extremely compact.

A DC-DC converter is an integrated circuit package that includes a high-speed switching device (almost always, a *MOSFET*) in conjunction with an oscillator circuit, an inductor, and a diode. By comparison, a **linear regulator** is usually based around bipolar transistors. Its input must always be at a higher voltage than its output, and its efficiency will be inversely proportional with the voltage drop that it imposes. See the **voltage regulator** entry in this encyclopedia for additional information.

There is no single symbol to represent a DC-DC converter. Some simplified schematics showing the principles of operation of commonly used converters are referenced under the following Variants section.

A DC-DC converter is also typically found in the output stage of a switching **AC-DC power supply**.

How It Works

An internal oscillator controls a *MOSFET* semiconductor that switches the DC input on and off at a high frequency, usually from 50KHz to 1MHz. Output voltage is adjusted by varying the *duty cycle* of the oscillator—the length of each “on” pulse relative to each “off” interval. This is known as *pulse-width modulation*, or *PWM*. The duty cycle is controlled by sampling the output of the converter and using a comparator to subtract the output voltage from a reference voltage, to establish an error value. This is passed to another comparator, which subtracts the error voltage from an oscillator ramp signal. If the error increases, the oscillator signal is more heavily clipped, thus changing the effective ratio of on/off pulse lengths. A simplified schematic of the PWM circuit is shown in Figure 17-1, which omits other components for clarity. The system of subtracting an error voltage from a ramp oscillator voltage to obtain a pulse-width modulated signal is illustrated in Figure 17-2.

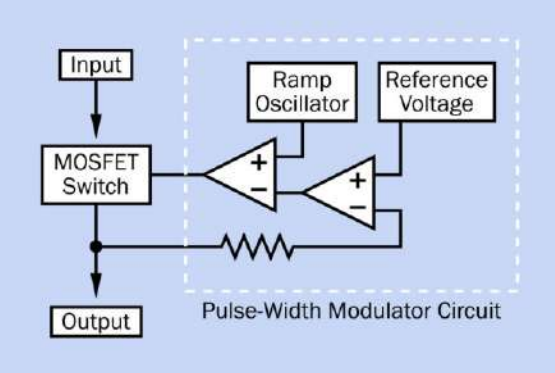

Figure 17-1. *The heart of a DC-DC converter is a MOSFET switch, which operates at a high frequency with pulse-width modulation used to create an adjustable DC output.*

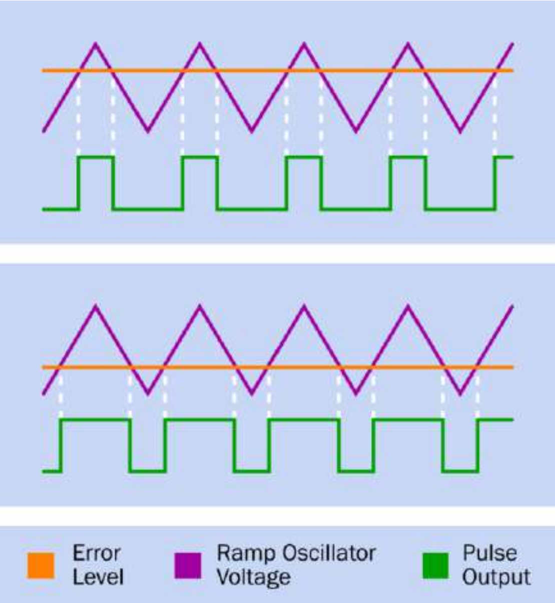

Figure 17-2. *To achieve pulse-width modulation, an error-level voltage is established by comparing the output from the converter with a reference voltage. The error level, shown as an orange line, is then subtracted from the output from a ramp oscillator. The pulse width varies accordingly.*

The key to the efficiency of a DC-DC converter is an inductor, which stores energy in its magnetic field during "on" pulse and releases it in the discharge phase. Thus, the inductor is used as a temporary reservoir and minimizes the ripple current. All converter variants use a coil for this purpose, although its placement varies in relation to the diode and capacitor that complete the basic circuit.

Variants

Four basic switching circuits are used in DC-DC converters and are defined in the coming sections, with a formula to determine the ratio between input voltage (V_{in}) and output voltage (V_{out}) in each case. In these formulae, variable D is the duty cycle in the pulse train generated through an internal MOSFET switch. The duty cycle is the fraction of the total on-off cycle that is occupied by each "on" pulse. In other words, if T_{on} is the duration of an "on" pulse and T_{off} is the "off" time:

$$D = T_{on} / (T_{on} + T_{off})$$

Buck Converter

See Figure 17-3. The output voltage is lower than the input voltage. The input and output share a common ground. For this circuit:

$$V_{out} = V_{in} * D$$

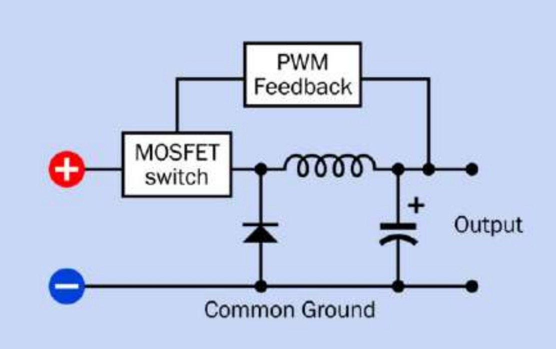

Figure 17-3. *Basic topology of a buck-type DC-DC converter.*

Boost Converter

See Figure 17-4. The output voltage is greater than the input voltage. The input and output share a common ground. For this circuit:

```
Vout = Vin / (1-D)
```

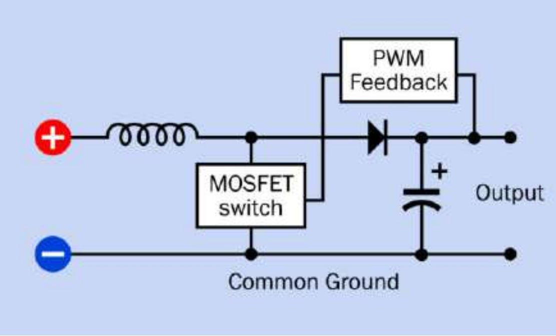

Figure 17-4. *Basic topology of a boost-type DC-DC converter.*

Flyback Converter with Inductor

Commonly known as a *buck-boost converter*. See Figure 17-5. The output voltage can be less than or greater than the input voltage. The input and output share a common ground. For this circuit:

```
Vout = Vin * (D / (1-D))
```

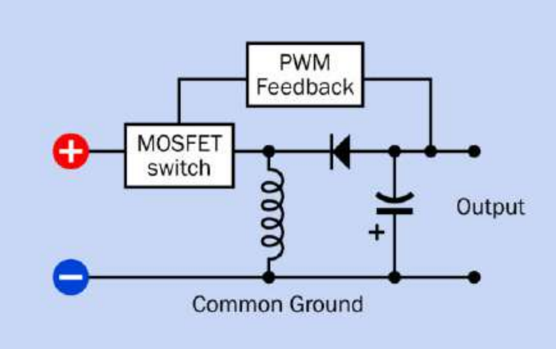

Figure 17-5. *Basic topology of a flyback-type DC-DC converter.*

Flyback Converter with Transformer

See Figure 17-6. The output voltage can be less than or greater than the input voltage. The input and output are isolated from one another. For this circuit:

```
Vout = Vin * (D / (1-D))
```

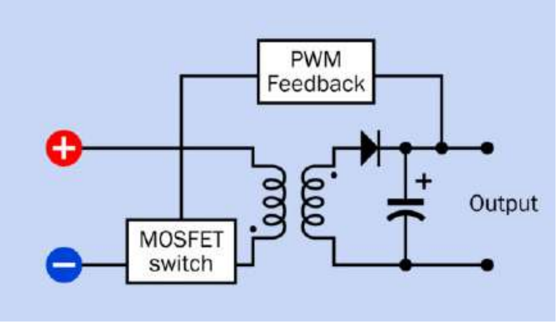

Figure 17-6. *Basic topology of a flyback-type DC-DC converter. (Buck, boost, and flyback topologies adapted from Maxim Integrated Products.)*

Using a transformer in the converter allows multiple outputs with different voltages, supplied through multiple transformer windings.

Formats

A converter may be packaged in a flat rectangular box that requires no additional heat sink and has pins for through-hole insertion into a PC board. Sizes usually range up to to 2″ × 2″. Power handling can range from 5 to 30 watts. Converters of this type are shown in Figure 17-7. (Top: Input range of 9 to 18VDC, fixed output of 5VDC at 3A completely isolated from the input. Typical efficiency of approximately 80%. The case is made of copper, providing good heat dissipation with electrical shielding. Center: Input range of 9 to 18VDC, fixed output of 5VDC at 500mA completely isolated from the input. Typical efficiency of approximately 75%. The manufacturer claims that external capacitors are only needed in critical applications. Bottom: SIP format, fixed input of 12VDC, fixed output of 5VDC at 600mA completely isolated from the input. Typical efficiency of approximately 75%. Requires external capacitors for ripple rejection.)

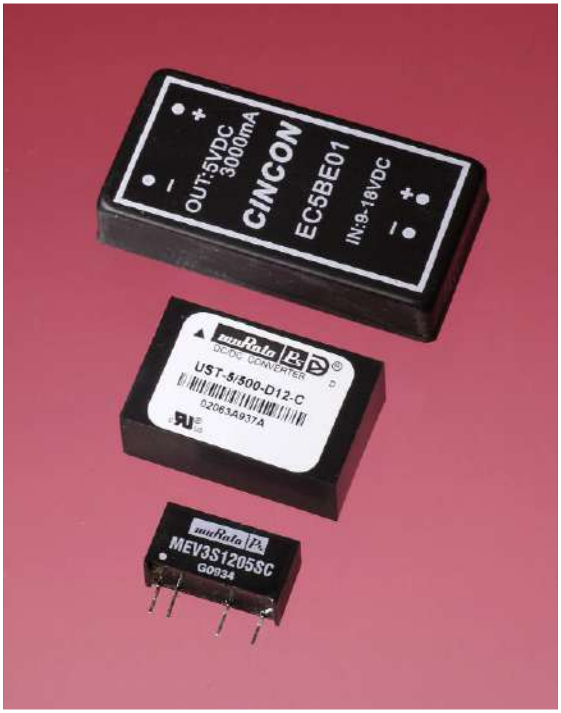

Figure 17-7. *A selection of sealed DC-DC converters.*

Lower-power converters are also available as surface-mount devices.

Some adjustable-output converters are supplied as multiple surface-mounted components preinstalled on a mini-board that has pins for through-hole insertion in a printed circuit board. Their high efficiency enables them to handle a lot of power for their size. In Figure 17-8, the converter accepts a 4.5 to 14VDC input range and has an adjustable output of 0.6 to 6VDC. It is rated at a surprising 10A or 50W and is more than 90% efficient. However, it draws 80mA in a no-load state, causing it to become quite hot. A thermal cutout or automatic shutdown may be used if the converter will not be driving a consistent load.

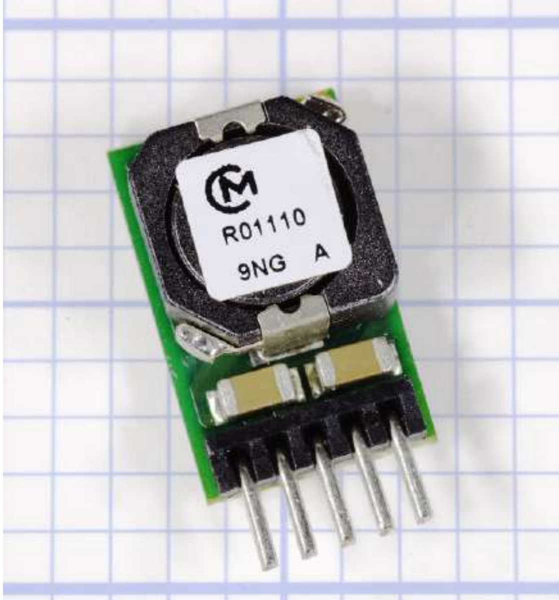

Figure 17-8. *An adjustable DC-DC converter rated for 10A or 50W. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.*

The miniboard in Figure 17-9 accepts an input voltage from 7 to 36VDC and has an adjustable output ranging from 2.5 to 12.6VDC, at up to 6A. It is non-isolated (has a common negative bus) and claims to be more than 95% efficient at full load.

The miniboard in Figure 17-10 accepts an input voltage from 4.5 to 14VDC and has an adjustable output ranging from 0.6 to 6VDC at up to 20A. It is non-isolated (has a common negative bus) and claims to be more than 90% efficient at full load.

Values

Relevant values include:

Nominal Input Voltage and Frequency

A wide range of input voltages is often acceptable, as the PWM can vary accordingly. Converters

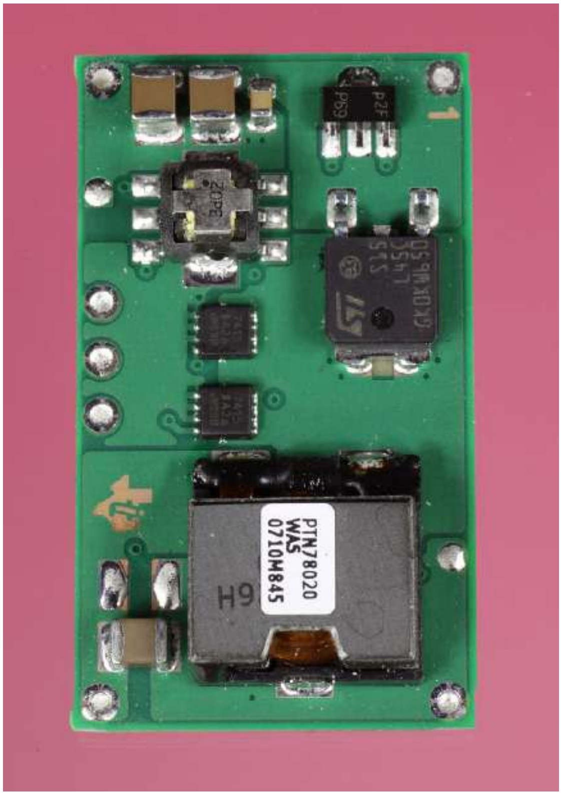

Figure 17-9. *Another adjustable DC-DC converter. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.*

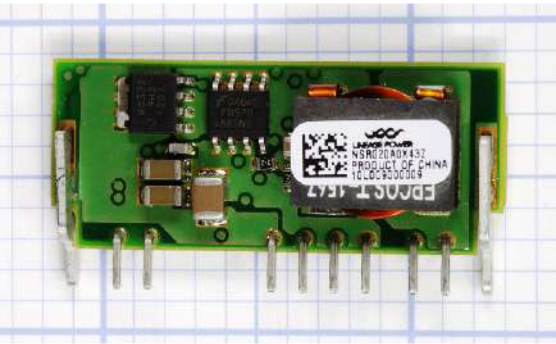

Figure 17-10. *Another adjustable DC-DC converter. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.*

often allow equipment to be usable internationally, on any voltage ranging from 100VAC to 250VAC, at a fequency of 50Hz or 60Hz, without any adaptation.

Output Voltage

As previously noted, many converters allow the output voltage to be adjusted by adding an external resistor or potentiometer. Alternatively, there may be multiple fixed output voltages, accessible via different pins on the package. They may also provide a positive voltage and equally opposite negative voltage relative to a ground pin.

Input Current and Output Current

Because input voltage and output voltage are likely to be different, the current alone is not a reliable guide to power handling.

A datasheet should specify input current with no load (open circuit on the output side). This current will have to be entirely dissipated as heat.

Load Regulation

This is usually expressed as a percentage and suggests the extent to which output voltage may be pulled down when the load on a DC-DC converter increases. If V_{nil} is the measured output voltage with no load, and V_{max} is the measured output voltage with the maximum rated load:

$$\texttt{Load regulation} = 100 * (V_{nil} - V_{max})/V_{max}$$

However, note that some converters are designed with the expectation that they will never be used with zero load across the output. In these cases, V_{nil} will be the voltage at minimum rated load.

Efficiency

This is a measure of how much input current must be dissipated as heat. A converter with a 12-volt input, drawing a maximum 300mA input current, will consume 3.6 watts (3,600mW). If it is 80% efficient, it will have to dissipate roughly 20% of its power as heat, or 720mW.

Ripple and Noise

Sometimes abbreviated R/N, this may be measured in mV or as a percentage. Check the specification carefully to determine whether the ripple-and-noise values require use of external smoothing capacitors. Often, this is the case.

Isolated or Non-Isolated

This crucial piece of information is often found near the top of a datasheet, not in the detailed specifications.

How to Use it

Because a converter creates electrical noise, it should be prevented from affecting other components by adding substantial *bypass capacitors* as close as possible to its input and output pins. For most converters, external capacitors are mandatory, and their *effective series resistance* (ESR) should be as low as possible (see the **capacitor** entry in this encyclopedia for an explanation of ESR). Tantalum capacitors are preferable to electrolytics for this reason, and are also more durable. Some manufacturers recommend placing a tantalum capacitor in parallel with an electrolytic. A small ceramic capacitor, typically 0.1µF, is often recommended in an addition to larger-value capacitors on the output side.

The voltage rating of each capacitor should be twice the voltage at the point in the circuit where it is used. The capacitance value will usually be higher for higher-current converters. Values of 100µF are common, but for high amperage, a value may be as high as 1,000µF.

While datasheets are often inadequate for some types of components, datasheets for DC-DC converters usually include detailed instructions regarding bypass capacitors. Following these instructions is essential. In the relatively rare instances that a datasheet makes no mention of bypass capacitors for a converter, this does not necessarily mean that the capacitors are unnecessary. The manufacturer may assume that they will be used as a matter of course.

Converters are used in a very wide range of devices, supplying power ranging from a few milliamps to tens of amps. At the lower end of the scale, devices such as cellular telephones, portable computers, and tablets contain subcircuits that require different voltages, some of which may be higher than the voltage of a single battery or battery pack that powers the device. A converter can satisfy this requirement. Because a converter can be designed to maintain a fixed output in response to a range of input voltages, it can also compensate for the gradual decline in voltage that occurs during battery usage.

A boost-type converter can be used to double the voltage from a single 1.5V battery in an LED flashlight where 3 volts are required to power the LED. Similarly, a boost-type converter can provide the necessary voltage to run a cold-cathode fluorescent tube that provides backlighting in an LCD computer display.

On a circuit board that is primarily populated with 5VDC components and is fed by a single 5VDC power supply, a converter can be used to supply 12VDC for one special purpose, such as an analog-digital converter or a serial data connection.

If electromechanical **relays** or other inductive loads share a common ground with components ,such as *logic chips* or **microcontrollers**, it may be difficult to protect the sensitive components from voltage spikes. A A flyback converter with a transformer separating the output from the input can allow the “noisy” section of the circuit to be segregated, so long as the converter itself does not introduce noise. Since the electromagnetic interference (EMI) introduced by converters varies widely from one model to another, specifications should be checked carefully.

Very low-power components can pick up EMI from the wires or traces leading into and out of a converter. In this type of circuit, adequate noise suppression may be impossible, and a converter may not be appropriate.

What Can Go Wrong

Electrical Noise in Output

Electrolytic capacitors may be inadequate to smooth the high frequencies used. Multilayer ceramic capacitors or tantalum capacitors may be necessary. Check the manufacturer's datasheet for minimum and maximum values. Also check the datasheet for advice regarding placement of capacitors on the input side as well as across the output.

Excess Heat with No Load

Some converters generate substantial heat while they are powered without a load. The manufacturer's datasheet may not discuss this potential problem very prominently or in any detail. Check the input rating, usually expressed in mA, specified for a no-load condition. All of this current will have to be dissipated as heat, and the very small size of many converters can result in high localized temperatures, especially since many of them allow no provision for a heat sink.

Inaccurate Voltage Output with Low Load

Some converters are designed to operate with at least 10% of full rated load across their output at all times. Below this threshold, output voltage can be grossly inaccurate. Read datasheets carefully for statements such as this: "Lower than 10% loading will result in an increase in output voltage, which may rise to typically double the specified output voltage if the output load falls to less than 5%." Always use a meter to verify the output voltage from a converter at a variety of different loads, and perform this test before installing the converter in a circuit.

DC-AC inverter

18

A power inverter must not be confused with a *logic inverter*, which functions as a digital component in *logic circuits* to invert the state of a low-voltage DC input from high to low or low to high. Logic inverters are discussed in Volume 2.

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See Chapter 16)
- **DC-DC converter** (See Chapter 17)

What It Does

A power inverter is included here as counterpoint to a **power supply** or *AC adapter*, since it has the opposite function. The inverter receives an input of *direct current* (typically 12VDC from a car battery) and delivers an output of *alternating current* (AC) in the range 110VAC-120VAC or 220VAC-240VAC, suitable to power many low-wattage appliances and devices. The interior of a low-cost inverter is shown in Figure 18-1.

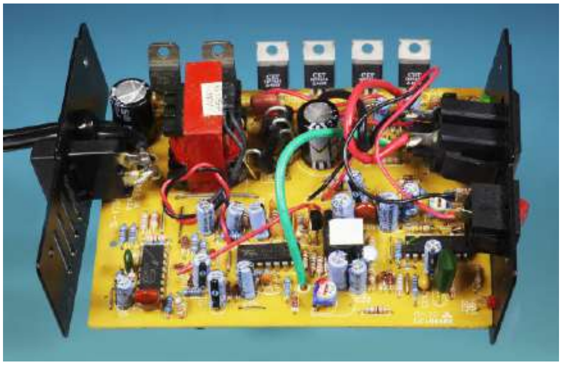

Figure 18-1. *The interior components in a 175-watt inverter.*

How It Works

The first stage of an inverter typically raises a 12VDC input to a higher DC voltage via an internal **DC-DC converter**, then uses a switching circuit to create an approximation of the sinusoidal profile that is characteristic of AC voltage.

Digital switching components naturally tend to create square waves, whose simple appearance conceals the presence of higher frequencies, or *harmonics*, that are ignored by some devices (especially those that convert electricity into heat) but can be troublesome in consumer electronics equipment. A primary objective of inverter design is to adapt or combine square waves to emulate a classic AC sine wave with reasonable fidelity. Generally speaking, the more accurately an inverter emulates a sine wave, the more expensive it tends to be.

The most primitive inverter would create a plain square wave such as that shown in red in Figure 18-2, superimposed on a comparable sine wave (in green). Note that alternating current rated at 115 volts actually peaks at around 163 volts because the number 115 is the approximate *root mean square* (*RMS*) of all the voltage values during a single positive cycle. In other words, if the voltage is sampled x times during a cycle, an

RMS value can be derived by squaring each sample, adding all the samples, dividing by x, and then taking the square root of the result. The RMS value is important as a means to calculate actual power delivered because it can be multiplied by the current to obtain an approximate value in watts.

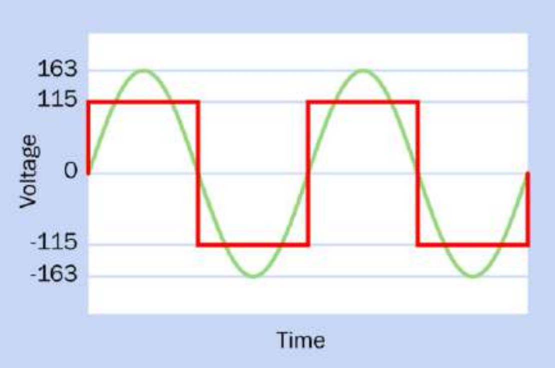

Figure 18-2. *Comparison of an AC voltage sine wave (green) and a square wave of the same frequency (red), both delivering a roughly similar amount of power.*

Variants

As a first step toward a better approximation of a sine wave, gaps of zero voltage can be inserted between square-wave pulses. This "gapped" square wave is shown in Figure 18-3.

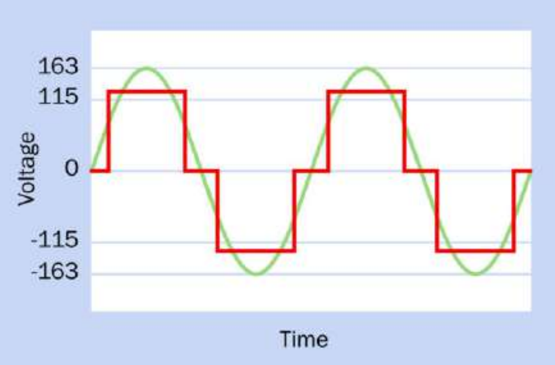

Figure 18-3. *Introducing pauses or gaps of zero voltage between square-wave pulses can produce slightly improved resemblance to a sine wave.*

A further improvement can be achieved if an additional, shorter pulse of higher voltage is added to each primary pulse, as shown in Figure 18-4. Outputs of this kind are referred to as *modified sine wave*, although they are actually square waves modified to emulate a sine wave. Their inaccuracy is expressed as *total harmonic distortion* (*THD*). Some authorities estimate that the THD of gapped square-wave output is around 25%, whereas the addition of shorter square waves reduces this to around 6.5%. This is a topic on which few people agree, but there is no doubt that a "stacked" sequence of square waves provides a closer emulation of a sine wave.

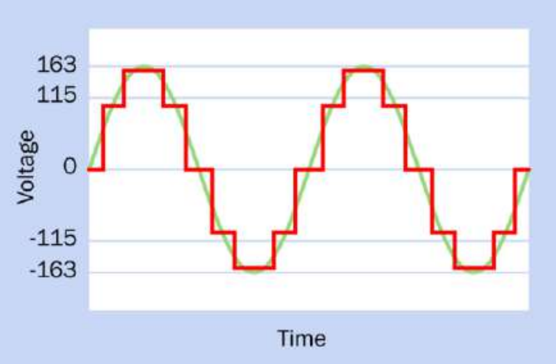

Figure 18-4. *A secondary stream of narrower square-wave pulses can improve the fidelity of an inverter's output.*

A *true sinewave inverter* typically uses *pulse-width modulation* (*PWM*) to achieve THD of less than 1%. It generates a stream of pulses much higher in frequency than that of the AC output, and varies their widths in such a way that their averaged voltage closely approximates the voltage variations in a sine wave. A simplified representation of this principle is shown in Figure 18-5.

Values

Small inverters are typically rated to deliver up to 100 watts and may be fitted with a 12VDC plug for insertion in a vehicle's cigarette lighter. Since a cheap inverter may be only 80% efficient, 100 watts at 135VAC will entail drawing as much as

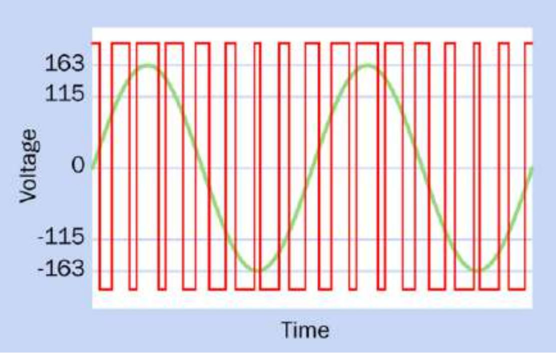

Figure 18-5. *Pulse-width modulation adjusts the widths of pulses delivered at a high frequency. The pulse widths can be averaged to generate voltage that follows a close approximation of a sine wave.*

10 amps at 12VDC. Cigarette lighters are usually fused at 15 or 20 amps, so 100 watts is a reasonable value. Inverters that are rated above 150 watts usually have cables terminating in oversize alligator clips for direct connection to the terminals of a 12V battery.

While the *cold cranking* rating of a car battery may be 100 amps or more, the battery is only designed to deliver that power for up to 30 seconds at a time. Inverters rated for as much as 500 watts will exceed the normal capacity of a single car battery, although if the battery is mounted in a vehicle, it can be supplemented by running the engine so that the alternator shares some of the load. A 500-watt inverter is better supplied by two or more 12-volt car batteries wired in parallel.

How to Use it

Small inverters are typically used in vehicles to run cellphone chargers, music players, or laptop computers. Large inverters are an integral part of off-the-grid solar and wind-powered systems, where battery power must be converted to AC house current. Uninterruptible power supplies contain batteries and inverters capable of running computer equipment for a brief period. Battery-driven electric vehicles with AC motors use inverters with an exceptionally high current rating.

There is a lack of consensus regarding possible harmful effects of powering electronics equipment with a low-cost modified sine wave inverter. Anecdotal evidence suggests that where the equipment uses its own *switching power supply* or uses an *AC adapter* (either mounted internally or as an external plug-in package), the filtering built into the power supply will block harmonics from the inverter.

Other evidence suggests that cheap inverters may have adverse effects on devices containing synchronous motors that run direct from AC. There are reports that fluorescent lighting and photographic electronic flash systems may be unsuitable for use even with modified sine wave inverters. However, differences in product design and component quality make it impossible to generalize. A cheaply made inverter may generate a wave form that is not even a close approximation of a square wave. See Figure 18-6.

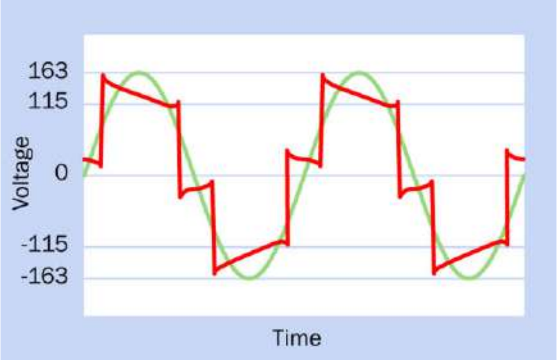

Figure 18-6. *A cheaply made inverter can generate a distorted wave form that is even higher in noise than a pure square wave. This sample is adapted from an actual oscilloscope trace.*

What Can Go Wrong

If multiple batteries are connected in parallel, using suitably heavy-gauge wire to power a large inverter, the batteries must be identical in specification and age, and must all be equally charged to prevent high and potentially dangerous flows of current among the batteries as they attempt to reach an equilibrium among themselves. Interconnections must be firmly clamped to clean battery terminals. For additional information, see the **battery** entry in this encyclopedia.

Problems associated with inverters are likely to be mundane. A 12V wiring to the inverter can overheat if items such as clothes or bedding are left on top of it; a high-wattage fan-cooled inverter can overheat if the fan is obstructed by poor placement or impaired by accumulated dirt; alligator clips may become dislodged from battery terminals; and power surges drawn by inductive loads such as motors may trigger the inverter's *breaker*, especially if they are used in conjunction with other equipment.

As always, high amperage should be treated with caution and respect, regardless of it being delivered at "only 12 volts."

voltage regulator

19

Correctly known as a *linear voltage regulator* to distinguish it from a *switching regulator* or **DC-DC converter**. However, the full term is not generally used, and "voltage regulator" is normally understood to mean a linear voltage regulator.

OTHER RELATED COMPONENTS

- **DC-DC converter** (See Chapter 17)
- **AC-DC power supply** (See Chapter 16)

What It Does

A linear voltage regulator provides a tightly controlled DC output, which it derives from an unregulated or poorly regulated DC input. The DC output remains constant regardless of the load on the regulator (within specified limits). It is a cheap, simple, and extremely robust component.

There is no single schematic symbol for a linear voltage regulator.

The general physical appearance of a commonly used type of regulator, rated for an output of around 1A DC, is shown at Figure 19-1. The LM7805, LM7806, LM7812, and similar regulators in the LM78xx series are encapsulated in this type of package, with pins that are spaced at 0.1" and have functions as shown. Other types of regulator may differ in appearance, or may look identical to this one but have different pin functions. Always check datasheets to be sure.

How It Works

All linear regulators function by taking some feedback from the output, deriving an error value by comparing the output with a reference voltage (most simply provided by a zener diode), and using the error value to control the base of a *pass transistor* that is placed between the input and the output of the regulator. Because the transistor operates below saturation level, its output current varies linearly with the current applied to its base, and this behavior gives the linear regulator is name. Figure 19-2 shows the relationship of these functions in simplified form; Figure 19-3 shows a little more detail, with a *Darlington pair* being used as the pass transistor. The base of the pair is controlled by two other transistors and a comparator that delivers the error voltage. This version of a voltage regulator is known as the *standard type*.

The voltage difference required between the base and emitter of an NPN transistor is a minimum of 0.6V. Because multiple transistors are used inside a standard-type voltage regulator, it requires a minimum total voltage difference, between its input and its output, of 2VDC. This voltage difference is known as the *dropout voltage*. If the voltage difference falls below this minimum, the regulator ceases to deliver a reliable output voltage until the input voltage rises again. *Low dropout regulators* allow a lower voltage difference, but are more expensive and less commonly used. They are described under the following Variants section.

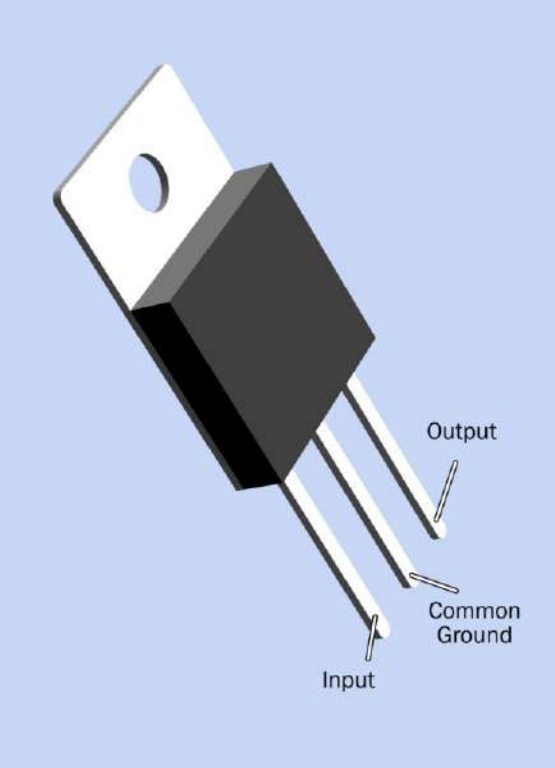

Figure 19-1. *The package design of a commonly used voltage regulator. Others may be significantly different, and the pin functions may vary. Check manufacturer datasheets for details.*

Discrete components could in theory be used to build a voltage regulator, but this ceased to be cost-effective several decades ago. The term is now understood to mean one small integrated package containing the basic circuit augmented with additional, desirable features, such as automatic protection against overload and excessive heat. Instead of burning out if it is overloaded, the component simply shuts down. Most voltage regulators also tolerate accidentally reversed power connection (as when batteries are inserted the wrong way around) and accidentally reversed insertion of the regulator in a circuit board.

Other components can satisfy the requirement to deliver power at a reduced voltage. Most simply, if two resistors in series are placed across a

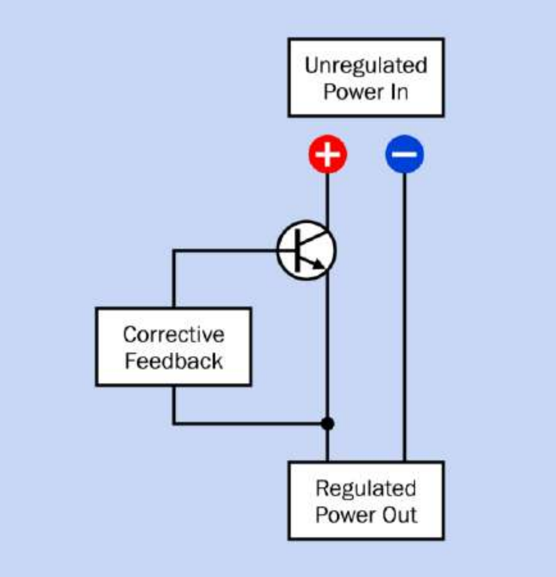

Figure 19-2. *A linear voltage regulator basically consists of a transistor whose base is controlled by corrective feedback derived from the output.*

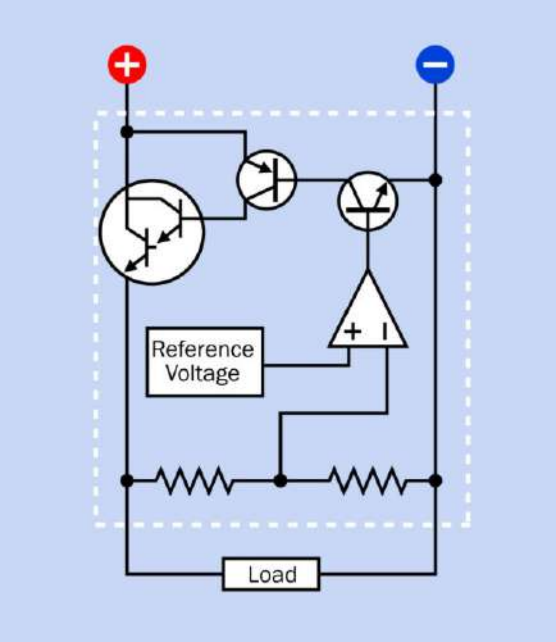

Figure 19-3. *The fundamental internal features of a standard-type voltage regulator, including a Darlington pair, two transistors, a voltage divider, comparator, and reference voltage source, shown inside the dashed white line.*

power source, they form a *voltage divider*, which provides an intermediate voltage at the connection between them. However, this voltage will vary depending on fluctuations in the input voltage and/or load impedance. A voltage regulator is the simplest way to supply a voltage that remains stable regardless of excursions in the input or fluctuations in power consumed by the load.

The disadvantage of a standard-type voltage regulator is that it is inefficient, especially when a relatively high input voltage is used to deliver a relatively low output voltage. If V_{in} is the input voltage, V_{out} is the output voltage, and I_{out} is the output current, the average power loss, P, is given by the formula:

$$P = I_{out} * (V_{in} - V_{out})$$

For example, if the output current is 1A, the input voltage is 9VDC, and the output is 5VDC, 44% of the input power will be wasted, and the component will be only 56% efficient. The wasted power (about 4 watts, in this case) will be dissipated as heat. Even when a standard-type regulator runs at its minimum 2VDC dropout voltage, it must dissipate 1W when delivering 0.5A.

Variants

Packaging

The package for the LM78xx series of regulators, shown in Figure 19-1, incorporates an aluminum plate drilled with a hole so that it can be bolted to a heat sink. Voltage regulators with a lower rated maximum output current (typically, 100mA) do not have the same need for a heat sink, and are available in a package that resembles a small transistor.

Some integrated circuits are available containing two voltage regulators, electrically isolated from each other.

Popular Varieties

In the LM78xx series, the last two digits in the part number specify the output voltage, which is fixed. Thus the LM7805 delivers 5VDC, the LM7806 delivers 6VDC, and so on. For regulators with a fractional voltage output (3.3VDC being common), an additional letter may be inserted in the part number, as in the 78M33.

Many copies of the LM78xx series are made by different manufacturers, the copies being functionally identical, regardless of additional letters that are added to the part number to identify its source or other attributes.

The LM78xx regulators are mostly rated to be accurate within 4%, although actual samples almost always deliver voltages that are more precise than this range suggests.

Adjustable Regulators

While the majority of regulators have a fixed output, some allow the user to set the output by adding one or more resistors. The LM317 is a popular example. Its output voltage can range from 1.25VDC to 37VDC and is set via a resistor and a trimmer potentiometer, as illustrated in Figure 19-4. If R1 is the fixed-value resistor and R2 is the trimmer, as shown in the schematic, the output voltage, V_{out}, is given by the formula

$$V_{out} = 1.25 * (1 + (R2 / R1))$$

Typical values for R1 and R2 would be 240Ω and 5K, respectively. With the trimmer at the middle of its range, V_{out} would be 1.25 * (1 + (2500 / 240)) = approximately 15VDC, requiring an input of at least 17VDC. However, if the trimmer is reduced to 720Ω, the output would be 5VDC. In practice, the value of a trimmer should be chosen so that a mid-range setting provides approximately the desired output. This will enable fine adjustment of the output voltage.

While the versatility of an adjustable regulator is desirable, its overall power dissipation is still proportional to the difference between the input voltage and the output voltage. To minimize heat loss, this difference should not exceed the dropout voltage by a larger amount than is absolutely necessary.

An adjustable regulator may require larger bypass capacitors than a regulator with a fixed output. A manufacturer's recommendations for the LM317 are shown in Figure 19-4.

Negative and Positive Regulators

While most linear voltage regulators are designed for "positive input" (conventional current flow from input to output), some are intended for "negative input." In this variant, the common terminal is positive, and the input and output are negative in relation to it.

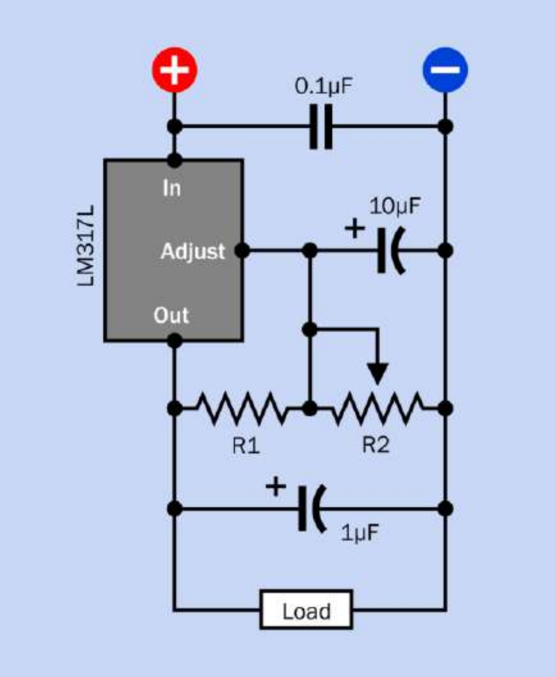

Figure 19-4. *Schematic for the LM317L adjustable voltage regulator, based on a circuit recommended by National Semiconductor, with bypass capacitors added for ripple rejection.*

Low-Dropout Linear Regulators

Low-dropout regulators (sometimes referred to as *LDO* regulators) allow a much lower dropout voltage by using a single PNP or MOSFET transistor. LDO regulators are popularly used in battery-powered devices where efficiency should be maximized and heat dissipation should be minimized. For example, the LM330 is a regulator with a 5VDC output, tolerating a dropout voltage of 0.6V, allowing it to be used with four AAA cells. In an LDO regulator the dropout voltage actually varies with load current and may diminish to as little as one-tenth of its rated value when the output current is minimal.

The majority of low-dropout regulators are sold in surface-mount packages, and are designed for maximum output of 100mA to 500mA. Only a few exceptions exist. They tend to be slightly more expensive than regulators with the typical 2V dropout rating.

Three voltage regulators are shown in Figure 19-5. From left to right, they are rated 5VDC at 1A, 12VDC at 1A, and 5VDC at 7.5A. The two smaller regulators are of the LM78xx series. The larger regulator claims a low maximum dropout voltage of 1.5VDC, and its output voltage can be adjusted with an external potentiometer and resistor.

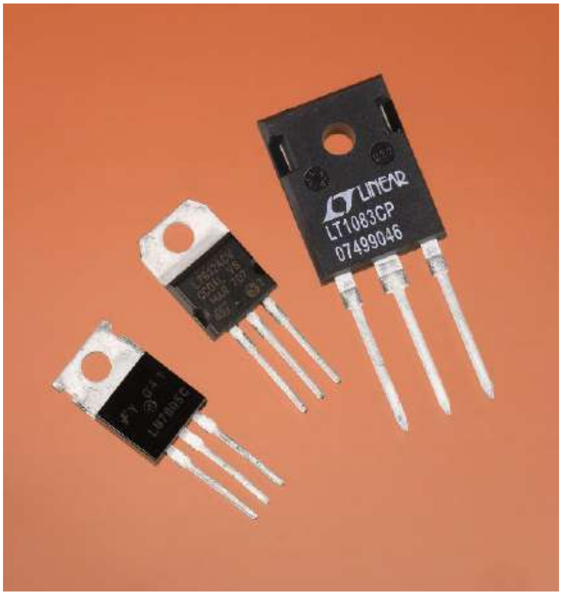

Figure 19-5. *Two voltage regulators from the LM78xx series, and a third high-current, low-dropout, adjustable regulator rated 5VDC (adjustable upward) at 7.5A.*

Quasi-Low-Dropout Linear Regulators

Where a standard regulator uses a Darlington pair as the pass transistor and an LDO uses a single PNP transistor, the so-called Quasi-LDO uses a combination of NPN and PNP transistors and has an intermediate dropout voltage, typically a maximum of 1.5VDC. However, the terms LDO and Quasi-LDO are not used uniformly in the industry. One manufacturer markets Quasi-LDO regulators as LDO regulators, and describes its LDO regulators as *Very Low Dropout* regulators. Consult datasheets to determine the actual specification of the product, regardless of its classification.

Additional Pin Functions

Some voltage regulators include an extra pin, typically known as an *enable pin*, which switches off the device in response to a signal from a microcontroller or logic gate.

Some regulators offer another option, an additional *status pin* that can signal a microcontroller that an error mode exists if the regulator output falls significantly below its rated value.

In battery-powered devices, a low-battery sensor is a desirable feature, since a regulator may simply shut down without warning if the input voltage is insufficient. A few regulators, such as the LP2953, provide a low-battery warning output via an extra pin.

Values

Linear voltage regulators with a single, fixed output are commonly available to supply DC outputs of 3.3, 5, 6, 8, 9, 10, 12, 15, 18, and 24 volts, with a few variants offering fractional values in between. The most commonly used values are 5, 6, 9, 12, and 15 volts. The input voltage may be as high as 35VDC.

Maximum output current is typically 1A or 1.5A, in the traditional three-pin, through-hole, TO-220 format. A surface-mount version is available. Other surface-mount formats have lower power limits.

Accuracy may be expressed as a percentage or as a figure for load regulation in mV. A typical load regulation value would be 50mV, while voltage regulation accuracy ranges from 1% to 4%, depending on the manufacturer and the component. While low-dropout regulators are generally more efficient, they do require more ground-pin current. This is not usually a significant factor.

How to Use it

Some components, such as many old-design CMOS chips or the traditional TTL version of the 555 timer, allow a wide range of acceptable input voltages, but most modern logic chips and microcontrollers must have a properly controlled power supply. Regulators such as the LM7805 are traditionally used to provide this, especially in small and relatively simple devices that draw a moderate amount of current, have a low component count, and are powered via a battery or an **AC adapter**. A fully fledged switching power supply is overkill in this kind of application.

A linear voltage regulator cannot respond instantly to changes in input voltage. Therefore, if the input supply contains voltage spikes, these spikes may pass through the regulator. Bypass capacitors should be applied preventively. A sample schematic showing an LM7805 regulator with bypass capacitors recommended by a manufacturer is shown in Figure 19-6.

In a battery-powered device where standby power is required for long periods and full power is only needed intermittently, the *quiescent current* drawn by a minimally loaded voltage regulator is important. Modern LDO regulators may draw as little as 100µA when they are very lightly loaded. Other types may consume significantly more. Check datasheets to find the most appropriate component for a particular application.

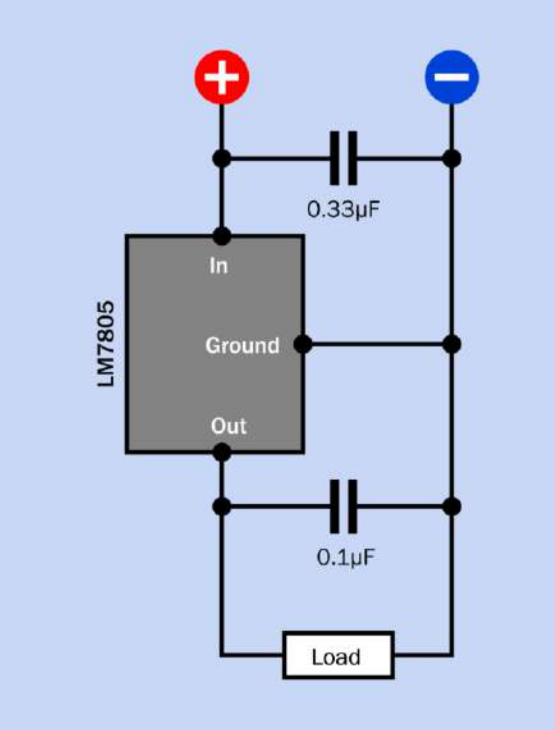

Figure 19-6. *Typical schematic for use of an LM7805 regulator, with capacitor values based on recommendations from Fairchild Semiconductor.*

Note that DC-DC power **converters** may draw a lot of current when they are lightly loaded, and will dissipate large amounts of heat as a result. An LDO is therefore preferable in this situation.

What Can Go Wrong

Inadequate Heat Management

The ability to "dial up" a wide range of voltages from an adjustable regulator such as the LM317 can be a temptation to use it on a "one size fits all" basis, to deliver any output ranging from 5VDC to 18VDC from a uniform 24VDC input. Assuming 1A output current, the worst-case power dissipation in this scenario would be almost 20W. To achieve reasonable efficiency and maintain waste heat at a manageable level, the input voltage should not exceed the output voltage by much more than the dropout voltage.

Even when a voltage regulator is used correctly, it can generate more heat than was expected if the requirements of a circuit are altered during development. An initial handful of components may draw only 100mA, but as more capabilities are requested and more parts are added (especially relays or LED displays) the power consumption can quickly add up, generating an unexpected amount of waste heat and raising the possibility of a sudden (and mysterious) shutdown if the regulator does not have an adequate heat sink.

Transient Response

When there is a major fluctuation in the demand by the load (for example, if an inductive device is switched on elsewhere in the circuit), the voltage regulator requires a finite time to adjust itself and maintain its specified output voltage. This time lag is known as its *transient response*. If a momentary fluctuation is likely, and other components may be sensitive to it, a larger capacitor should be used between the output of the voltage regulator and ground.

The transient response time may also be insufficient to block sudden, brief spikes in input voltage. This may occur, for example, when a low-cost AC adapter that does not have a properly smoothed output is used as the power source. Additional 1µF bypass capacitors may be added at the input and output of a regulator to provide better protection from power fluctuations.

Misidentified Parts

Many types of linear voltage regulators appear physically identical. Care is needed to distinguish those which have fixed output from those that allow a variable output. When using the LM78xx series, double-check the last pair of digits in the part number, which provide the only guide regarding the output. Using an LM7808 instead of an LM7805 may be sufficient to destroy all the 5VDC chips in a logic circuit. It is advisable to use a meter to check the output of any power supply before connecting it with a circuit.

Misidentified Pins

The LM78xx series of voltage regulators uses a very intuitively obvious and consistent scheme for the functions of its pins: input on the left, ground in the center, and output on the right, when looking at the regulator from the front, with its pins facing downward. Unfortunately the consistency of this scheme can encourage an unthinking habit for making connections. The LM79xx series of negative voltage regulators swaps the identity of the input and ground pins, whereas adjustable regulators use yet another different scheme. Good practice suggests checking a component against the manufacturer's datasheet before connecting it.

Dropout Caused by Low Battery

If a regulator rated to deliver 6VDC has a 2VDC dropout voltage and is powered from a 9V battery, the battery can easily drop below the minimum acceptable 8VDC if it becomes old or depleted. When this happens, the output from the regulator will tend to fall, or may oscillate.

Inaccurate Delivered Voltage

A voltage regulator maintains its output voltage between its output pin and ground pin. Thin traces on a circuit board, or a long run of very small-gauge wiring, can impose some electrical resistance, reducing the actual voltage delivered to a component. Ohm's Law tells us that the voltage drop imposed by a trace (or thin wire) will be proportional to the current flowing through it. For example, if the resistance between the output pin of a voltage regulator and a component is 0.5Ω and the current is 0.1A, the voltage drop will be only 0.05V. But if the current increases to 1A, the voltage drop is now 0.5V. Bearing this in mind, a linear voltage regulator should be positioned close to voltage-sensitive components. In printed circuit designs, the traces that deliver power should not have significant resistance.

When using linear voltage regulators with adjustable output, there may be a temptation to connect adjustment resistor R1 to the positive end of the load, to obtain a "more accurate" delivered voltage. This configuration will not produce the desired result. R1 should always be connected as closely as possible between the output pin and the adjustment pin of the voltage regulator, while R2 should connect between the adjustment pin and the negative end of the load. This is illustrated in Figure 19-7, where the gray wire in each schematic indicates that it possesses significant resistance.

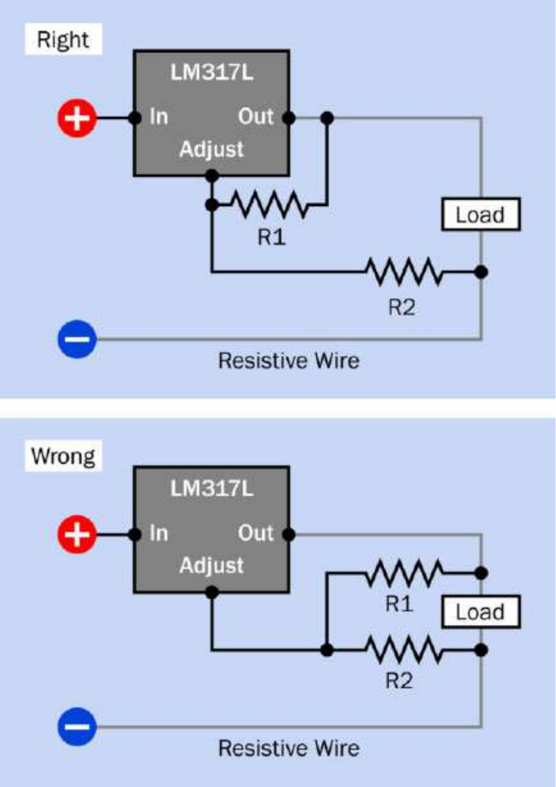

Figure 19-7. *When the connection between an adjustable-output voltage regulator and load components has a significant resistance (shown here as a gray "resistive wire"), R1 should always be connected as closely as possible to the pins of the regulator, as shown in the upper schematic. (Derived from schematics prepared by National Semiconductor.)*

electromagnet

20

The term **electromagnet** is used here to mean a coil containing a core of ferromagnetic material that does not move relative to the coil. The core is used solely to create a magnetic field that attracts or repels other parts that have appropriate inherent magnetic properties. Where a center component moves in response to the magnetic force created by current through a coil, this is discussed in the **solenoid** entry. By comparison, the **inductor** entry describes a coil that may or may not have a ferromagnetic core, and is used for the specific purpose of creating reactance, or self-inductance, in an electronic circuit, often in association with alternating current and in combination with resistors and/or capacitors. The **inductor** entry contains a basic discussion and explanation of magnetic force created by electricity.

OTHER RELATED COMPONENTS

- **solenoid** (See Chapter 21)
- **electromagnet** (See Chapter 20)
- **DC motor** (See Chapter 22)
- **AC motor** (See Chapter 23)

What It Does

An electromagnet consists of a coil that creates a magnetic field in response to an electric current. The field is channeled and reinforced by a core of magnetic material (that is, material that can be magnetized). Electromagnets are incorporated in motors, generators, loudspeakers, microphones, and industrial-sized applications such as mag-lev trains. On their own, they provide a means for electric current to hold, lift, or move objects in which a magnetic field can be induced.

A very small, basic electromagnet about 1 inch in diameter is shown in Figure 20-1. No specific schematic symbol for an electromagnet exists, and the symbol for an induction coil with a solid core is often used instead, as shown in Figure 14-1 (the center variant of each of the three) in the **inductor** entry of this encyclopedia.

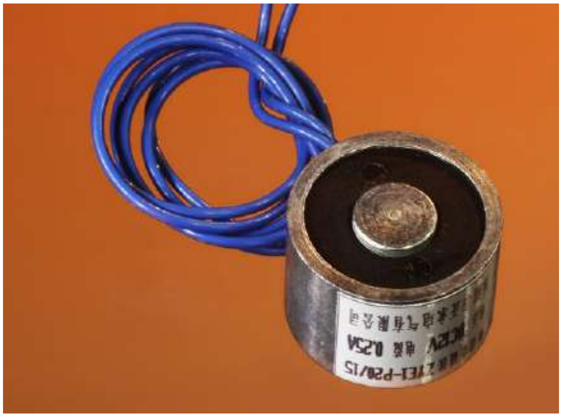

Figure 20-1. *An electromagnet approximately 1 inch in diameter, rated to draw 0.25A at 12VDC.*

How It Works

Electric current flowing through a circle of wire (or a series of connected loops that form a helix or *coil*) will induce a magnetic field through the center. This is illustrated in the **inductor** entry of this encyclopedia, specifically in diagrams Figure 14-3, Figure 14-4, Figure 14-5, and Figure 14-6.

If a stationary piece of ferromagnetic material is placed in the center of the circle or coil, it enhances the magnetic force because the *reluctance* (magnetic resistance) of the material is much lower than the reluctance of air. The combination of the coil and the core is an electromagnet. This is illustrated in Figure 20-2. For a lengthier discussion of this effect, see "Magnetic Core" (page 122).

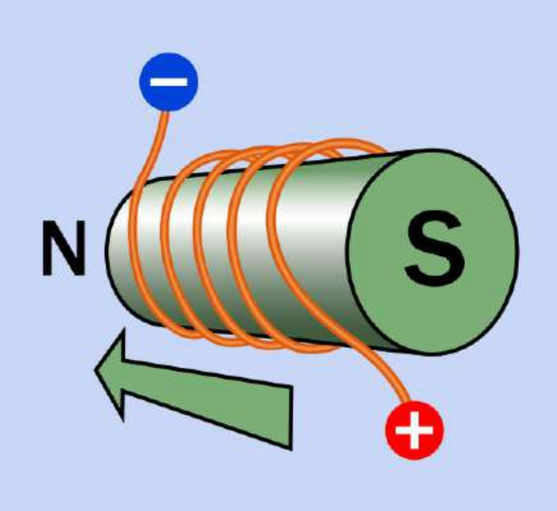

Figure 20-2. *Direct conventional current flowing through a wire coiled around a ferromagnetic rod induces a magnetic force in the rod, conventionally considered to flow from south to north.*

The magnitude of the electromagnetic *flux density* will be proportional to the current flowing through the coil, assuming a DC power source.

Variants

Electromagnet designs vary according to their application. The simplest design consists of a single coil wound around a rod which may terminate in a plate for applications such as lifting scrap metal. This design is relatively inefficient because the magnetic circuit is completed through air surrounding the electromagnet.

A more efficient, traditional design consists of a U-shaped core around which are wound one or two coils. If the U-shaped core is smoothly curved, it resembles a *horse-shoe magnet*, as shown in Figure 20-3. This design has become relatively uncommon, as it is cheaper to make windings across two separate, straight vertical cores and bridge them. However, the horseshoe configuration is extremely efficient, as the coils induce north and south magnetic polarities in the open ends of the U-shaped core, and the magnetic circuit is completed through any object that is attracted toward the open ends and links them. The attracted object is shown as a rectangular plate in Figure 20-3. Because a magnetic circuit will naturally attempt to limit its extent, and because this goal will be achieved when the circuit is completed, the attractive force of the U-shaped magnet is maximized.

An electromagnet powered by direct current naturally produces a consistently polarized, stable magnetic field. When AC current is applied, an electromagnet may still be used to exert an attractive force on a passive object that is not magnetized but is capable of being magnetized. The electromagnet will change its polarity at almost the same frequency as the AC, and will induce equal and opposite fluctuating polarity in the target, causing mutual attraction. The core of the magnet will be composed of plates separated by thin layers of insulation to inhibit the eddy currents induced by the AC, but still an AC-powered electromagnet will be less efficient than a comparable DC-powered electromagnet because it will also suffer from *hysteresis* as power is consumed by repeatedly reversing the polarity of the *magnetic domains* in the core.

Some electromagnets that are described as suitable for AC power actually contain *rectifiers* that convert the AC to DC.

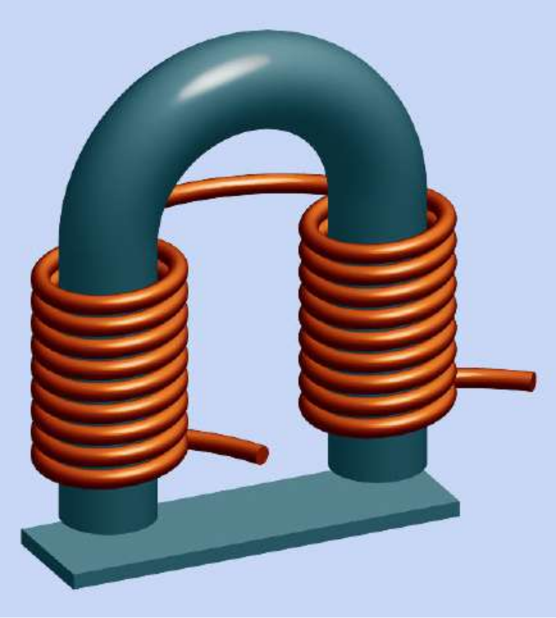

Figure 20-3. *This traditional design for an electromagnet has a pedigree stretching back for more than a century. It maximizes efficiency by completing a magnetic circuit through any object that the magnet attracts.*

Values

Electromagnets are typically calibrated in terms of their power consumption and retaining force (the weight of an iron target that they can support). The retaining force is usually measured in grams or kilograms.

How to Use it

Electromagnets are used mostly as subassemblies in other components, such as motors and generators, relays, loudspeakers, and disk drives. They have also been used in audio (and video) tape recorders to magnetize ferric oxide on tape, using a magnetic field of varying strength to record an audio signal. In this application, a form of horseshoe magnet with an extremely narrow gap is used, the width of the gap determining the highest frequency that the electromagnet can record, in conjunction with the speed of the tape moving past the head.

The tape recording process can be reversed when the electromagnet "reads" the tape and turns the signal back into a weak alternating current that can be amplified and reproduced through a loudspeaker.

A simple application for an electromagnet is in a traditional-style doorbell, where one or two coils attract a spring-loaded lever, at the top of which is a knob that hits a bell. When the lever is pulled toward the bell, it breaks a contact that supplies power to the electromagnet. This allows the lever to spring back to its original position, which reestablishes the circuit, repeating the process for as long as power is applied to the bell. The bulk and weight of the component parts in this type of doorbell are making it obsolete, as electronic versions containing small loudspeakers become relatively cheaper. However, a **solenoid** may still be used in the type of bell that creates a single chime or pair of chimes.

In any device using a *cathode-ray tube*, electromagnetic coils are used to form a *yoke* around the neck of the tube, to deflect the beam of electrons on its way to the screen. A similar principle is used in electron microscopes. In some cases, electrostatically charged plates are used to achieve the same purpose.

An electromagnet may be used to activate a *reed switch* (the diagram in Figure 9-7 shows such a switch). In this application, the combination of the electromagnet and the switch are functioning as a **relay**.

When an electromagnet is energized by alternating current, it can be used to *degauss* (in other words, to demagnetize) other objects. The AC is either applied with diminishing current, so that the alternating magnetic polarities gradually subside to zero, or the electromagnet is gradually moved away from the target, again reducing the magnetic influence to (virtually) zero. This latter procedure may be used periodically to demagnetize record and replay heads on tape recorders, which otherwise tend to acquire residual magnetism, inducing background hiss on the tape.

Traditional large-scale applications for electromagnets tend to involve lifting and moving heavy objects or scrap metal, such as junked cars. A more modern application is in magnetic resonance imaging (MRI), which has revolutionized some areas of medicine.

Very large-scale applications for electromagnets include particle accelerators, in which multiple magnetic coils are energized sequentially, and fusion-power generators, where high-temperature plasma is contained by a magnetic field.

What Can Go Wrong

Because an electromagnet requires constant power to maintain its magnetic force, yet it is not doing any actual work so long as its target remains stationary (in contact with the core of the magnet), the current running through the coil of the magnet must be dissipated entirely as heat. Further discussion of this issue will be found at "Heat" (page 177) in the **solenoid** section of this encyclopedia.

solenoid

21

The term **solenoid** was historically used to describe any coil without a magnetic core. More recently and more commonly it describes a coil inside of which a cylindrical plunger moves in response to the magnetic field generated by the coil. In this encyclopedia, the term **electromagnet** has its own entry, and describes a coil with a center component of ferromagnetic material that does not move relative to the coil. It is used solely to attract or repel other parts that have inherent magnetic properties. By comparison, the **inductor** entry describes a coil that is used for the specific purpose of creating reactance, or self-inductance, in an electronic circuit, often in association with alternating current and in combination with resistors and/or capacitors. The **inductor** entry contains a basic discussion and explanation of magnetic force created by electricity.

OTHER RELATED COMPONENTS

- **inductor** (See Chapter 14)
- **solenoid** (See Chapter 21)

What It Does

A typical solenoid consists of a hollow coil inside a *frame*, which may be a sealed cylinder or box-shaped with open sides. In the case of a cylinder, its opposite ends may be referred to as *pole faces*.

At least one of the pole faces has a hole through which a *plunger* (also known as an *armature*) is pulled or pushed by the solenoid. Thus, the solenoid is a device for applying a linear mechanical force in response to current passing through it. In most solenoids, current must be maintained in order to maintain the mechanical force.

A small open-frame solenoid is pictured in Figure 21-1. The upper section of the figure shows the three basic parts: frame, compression spring, and plunger. The lower part of the figure shows the parts assembled.

A larger, closed, cynlindrical solenoid is shown in Figure 21-2, with the plunger and spring removed.

A 3D rendering showing a simplified, imaginary, cylindrical solenoid cut in half appears in Figure 21-3. The diagram includes a gray cylindrical shell, often described as the frame; the coil, shown in orange; the plunger, which is pulled into the coil by its magnetic field; and the triangular stop, which limits the plunger's upward travel. The frame of the solenoid exists not merely to protect the coil, but to provide a magnetic circuit, which is completed through the plunger.

The lower end of the plunger is often fitted with a nonmagnetic yoke or perforated plate for connection with other components. Stainless steel can be used for this purpose. The stop may be fitted with a thrust rod (also fabricated from stainless steel) if the solenoid is intended to "push" as well as "pull." Springs to adjust the force of the plunger, or to return it to its initial position when the current through the coil is interrupted, are not shown in the rendering.

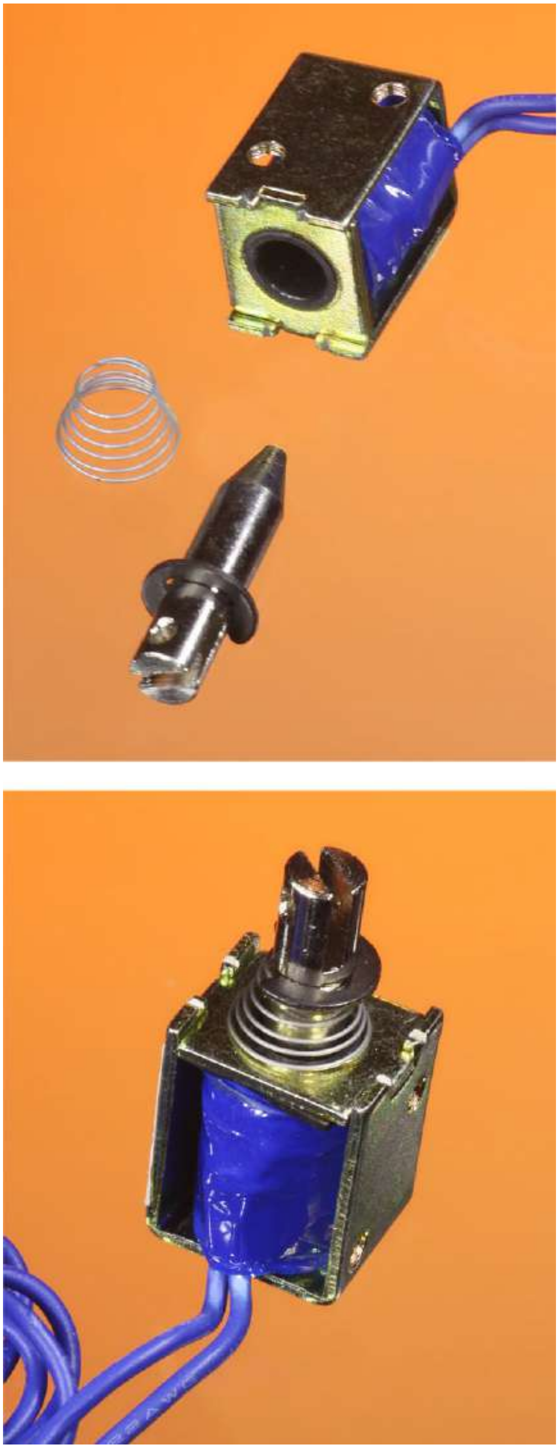

Figure 21-1. *A small 12VDC solenoid.*

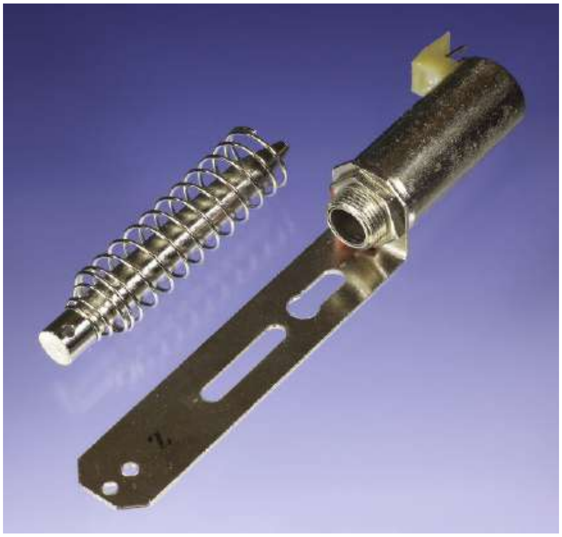

Figure 21-2. *A larger solenoid rated for 24VDC.*

Because there is no standardized schematic symbol for a solenoid, and because this type of component is so widely used in conjunction with valves, any diagram involving solenoids is more likely to emphasize fluid or gas flow with symbols that have been developed for that purpose. In such circuits, a solenoid may be represented simply by a rectangle. However, the symbols shown in Figure 21-4 may occasionally be found.

How It Works

Current flowing through the coil creates a magnetic force. This is explained in the **inductor** entry of this encyclopedia, using diagrams in Figure 14-3, Figure 14-4, Figure 14-5, and Figure 14-6.

If the plunger is fabricated from a material such as soft iron, the coil will induce an equal and opposite magnetic polarity in the plunger. Consequently the plunger will attempt to occupy a position inside the coil where the ends of the plunger are equal distances from the ends of the coil. If a collar is added to the free end of the plunger, this can increase the pulling force on the plunger when it is near the end of its throw because of the additional magnetic pull distributed between the collar and the frame of the solenoid.

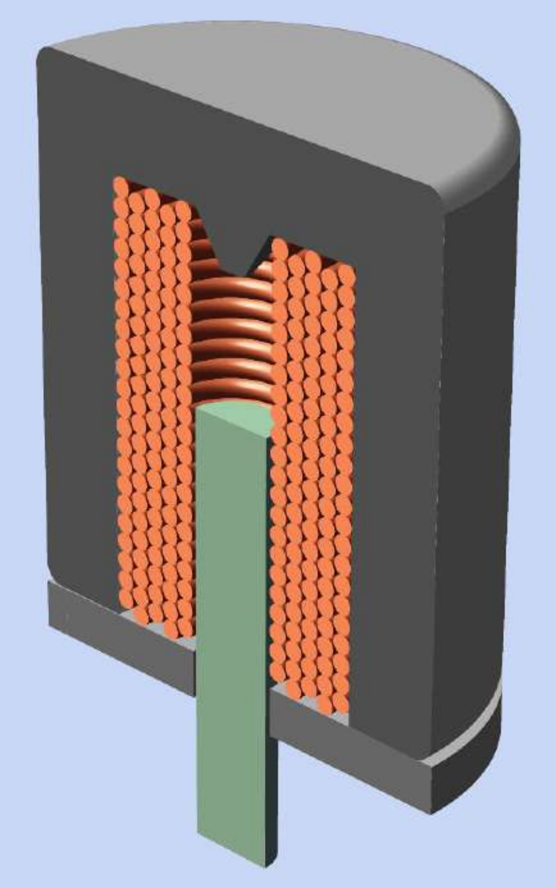

Figure 21-3. *A simplified view of a solenoid cut in half, showing the primary parts.*

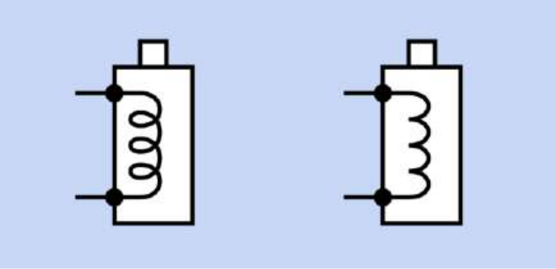

Figure 21-4. *Although no standard schematic symbol for a solenoid exists, these symbols may sometimes be found.*

A spring can be inserted to apply some resistive force to compensate for the increase in pulling force that occurs as a larger proportion of the plunger enters the coil. A spring may also be used to eject the plunger, partially at least, when current to the coil is interrupted.

If the plunger is a permanent magnet, reversing DC current to the coil will reverse the action of the plunger.

A solenoid with a nonmagnetized plunger may be energized by AC current, since polarity reversals in the magnetic field generated by the coil will induce equal and opposite reversals in the polarity of the plunger. However, the force curve of an AC-powered solenoid will be different from the force curve of a DC-powered solenoid. See Figure 21-5. The alternating current is likely to induce humming, buzzing, and vibration.

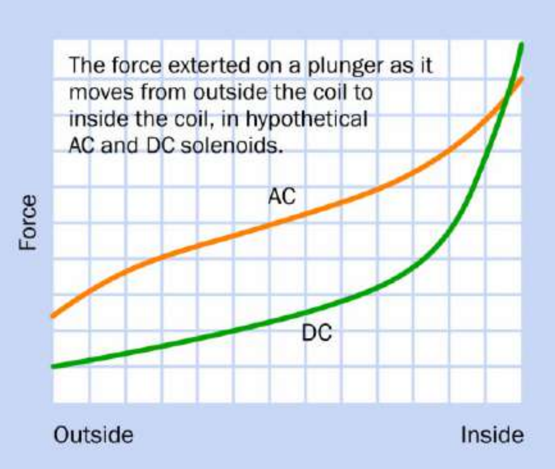

Figure 21-5. *A comparison of the force exerted on a plunger, relative to its position as it enters the coil, in hypothetical AC and DC solenoids.*

The frame of the solenoid increases the magnetic power that the coil can exert by providing a magnetic circuit of much lower *reluctance* than that of air (reluctance being the magnetic equivalent of electrical resistance). For a lengthier discussion of this effect, see "Magnetic Core" (page 122) in the **inductor** entry of this encyclopedia. If current flowing through the coil increases to the point where the frame becomes magnetically *saturated*, the pulling power of the solenoid will level off abruptly.

The heat generated by a solenoid when it is maintained in its energized state may be reduced if the manufacturer includes a series resistor and a switch that functions as a *bypass switch*. The switch is normally closed, but is opened mechanically when the plunger reaches the end of its throw, thus diverting electricity through the series resistor. This itself will generate some heat as a result of the current flowing through it, but by increasing the total resistance of the system, the total heat output will be reduced. The resistor value is chosen to provide the minimum power needed to retain the plunger at the end of its throw.

Variants

The most common variant is tubular, with open-frame as a secondary option. A tubular solenoid has been shown in Figure 21-2.

Additional variants include:

Low Profile

A shorter, fatter solenoid which may be used if a short throw is acceptable.

Latching

A permanent magnet holds the plunger when it reaches the end of its travel, and continues to hold it after power to the solenoid is disconnected. The plunger itself is also a permanent magnet, and is released by running current of reverse polarity through the coil.

Rotary

This variant is similar in principle to a brushless DC motor and causes the armature to rotate through a fixed angle (typically ranging from 25 to 90 degrees) instead of moving linearly. It is used as a mechanical indicator in control panels, although it is being displaced by purely electronic indicators.

Hinged Clapper

Instead of a plunger, a small hinged panel (the "clapper") moves in when the solenoid is active, and springs back when the power is interrupted.

Values

The stroke length, duty cycle, and holding force are the most significant values found in solenoid datasheets.

Holding forces for DC solenoids can range from a few grams to hundreds of kilograms. The holding force will be inversely proportional to the length of the solenoid, if all other variables are equal. The force that the solenoid can exert on its plunger also varies depending on the position of the plunger in the length of its throw.

Duty cycle is of special importance because the solenoid continues to draw power and create heat so long as it is holding the plunger at the end of its throw (assuming the solenoid is not the latching type). The initial current surge in an AC solenoid generates additional heat.

The duty cycle is simply calculated. If T1 is the time for which the solenoid is on and T2 is the time for which the solenoid is off, the duty cycle, D, is derived as a percentage from the formula

```
D = 100 * (T1 / (T1 + T2))
```

Some solenoids are designed to withstand a 100% duty cycle, but many are not, and in those cases, there is a maximum value not only for D but for the peak "on" time, regardless of the duty cycle. Suppose a solenoid is rated for a 25% duty cycle. If the solenoid is appropriately switched on for one second and off for three seconds, the heat will be allowed to dissipate before it has time to reach overload levels. If the solenoid is switched on for one minute and off for three minutes, the duty cycle is still 25%, but the heat that may accumulate during a one-minute "on" cycle may overload the component before the "off" cycle can allow it to dissipate.

Coil Size vs. Power

Because additional windings in a coil will induce a greater magnetic force, a larger solenoid tends to be more powerful than a smaller solenoid. However this means that if a larger and a smaller solenoid are both designed to generate the same force over the same distance, the smaller solenoid will probably draw more current (and will therefore generate more heat) because of its fewer coil windings.

How to Use it

Solenoids are primarily used to operate valves in fluid and gas circuits. Such circuits are found in laboratory and industrial process control, fuel injectors, aircraft systems, military applications, medical devices, and space vehicles. Solenoids may also be used in some electronic locks, in pinball machines, and in robotics.

What Can Go Wrong

Heat

Overheating is the principal concern when using solenoids, especially if the maximum "on" time is exceeded, or the duty cycle is exceeded. If the plunger is prevented from reaching the end of its throw, this can be another cause of overheating.

Because coil resistance increases with heat, a hot solenoid passes less current and therefore develops less power. This effect is more pronounced in a DC solenoid than an AC solenoid. A manufacturer's force curve should show the solenoid performance at its maximum rated temperature, which is typically around 75 degrees Centigrade, in a hypothetical ambient temperature of 25 degrees Centigrade. Exceeding these values may result in the solenoid failing to perform. As in all coils using magnet wire, there is the risk of excessive heat melting the insulation separating the coil windings, effectively shortening the coil, which will then pass more current, generating more heat.

AC Inrush

When an AC solenoid reaches the end of its travel, the sudden stop of the plunger results in forward *EMF* that generates additional heat. Generally speaking, a longer stroke creates a greater surge. Rapid cycling will therefore exacerbate coil heating.

Unwanted EMF

Like any device containing a coil, a solenoid creates back EMF when power is connected, and forward EMF when the power is disconnected. A protection diode may be necessary to suppress power spikes that can affect other components.

Loose Plunger

The plunger in many solenoids is not anchored or retained inside the frame and may fall out if the solenoid is tilted or subjected to extreme vibration.

DC motor

22

In this section, the term "traditional DC motor" is used to describe the oldest and simplest design, which consists of two *brushes* delivering power via a rotating, sectioned *commutator* to two or more electromagnetic coils mounted on the motor shaft. *Brushless* DC motors (in which DC is actually converted to a pulse train) are also described here because "brushless DC" has become a commonly used phrase, and the motor is powered by direct current, even though this is modified internally via *pulse-width modulation*.

OTHER RELATED COMPONENTS

- **AC motor** (See Chapter 23)
- **stepper motor** (See Chapter 25)
- **servo motor** (See Chapter 24)

What It Does

A traditional DC motor uses direct current to create magnetic force, which turns an output shaft. When the polarity of the DC voltage is reversed, the motor reverses its direction of rotation. Usually, the force created by the motor is equal in either direction.

How It Works

Current passes through two or more coils that are mounted on the motor shaft and rotate with it. This assembly is referred to as the *rotor*. The magnetic force produced by the current is concentrated via cores or poles of soft iron or high-silicon steel, and interacts with fields created by permanent magnets arrayed around the rotor in a fixed assembly known as the *stator*.

Power to the coils is delivered through a pair of *brushes*, often made from a graphite compound. Springs press the brushes against a sleeve that rotates with the shaft and is divided into sections, connected with the coils. The sleeve assembly is known as the *commutator*. As the commutator rotates, its sections apply power from the brushes to the motor coils sequentially, in a simple mechanical switching action.

The most elementary configuration for a traditional DC motor is shown in Figure 22-1.

In reality, small DC motors typically have three or more coils in the rotor, to provide smoother operation. The operation of a three-coil motor is shown in Figure 22-2. The three panels in this figure should be seen as successive snapshots of one motor in which the rotor turns progressively counter-clockwise. The brushes are colored red and blue to indicate positive and negative voltage supply, respectively. The coils are wired in series, with power being applied through the commutator to points between each pair of coils. The direction of current through each coil determines its magnetic polarity, shown as N for north or S for south. When two coils are energized in series without any power applied to their midpoint, each develops a smaller magnetic field than an individually energized coil. This is

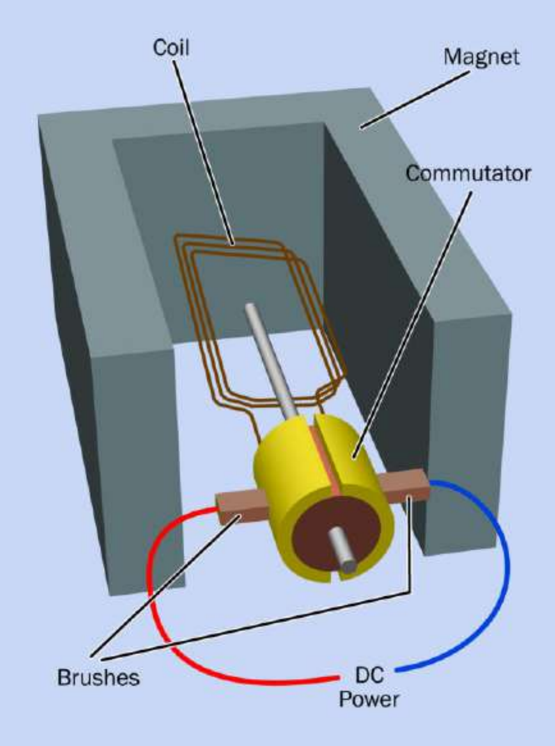

Figure 22-1. *The simplest traditional DC motor contains these parts. The combination of coil, shaft, and commutator is the rotor. The fixed magnetic structure in which it rotates is the stator.*

indicated in the diagram with a smaller white lowercase n and s. When two ends of a coil are at equal potential, the coil produces no magnetic field at all.

The stator consists of a cylindrical permanent magnet, which has two poles—shown in the figure as two black semicircles separated by a vertical gap for clarity—although in practice the magnet may be made in one piece. Opposite magnetic poles on the rotor and stator attract each other, whereas the same magnetic poles repel each other.

DC motors may be quite compact, as shown in Figure 22-3, where the frame of the motor measures about 0.7″ square. They can also be very powerful for their size; the motor that is shown disassembled in Figure 22-4 is from a 12VDC bilge pump rated at 500 gallons per hour. Its output was delivered by the small impeller attached to the rotor at right, and was achieved by using

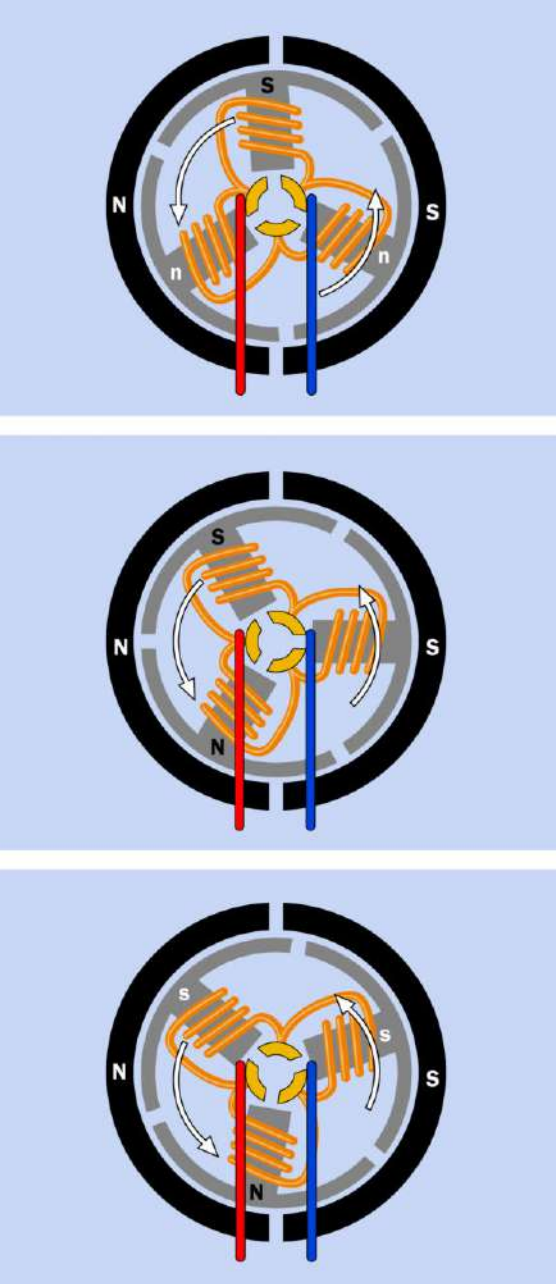

Figure 22-2. *Three sequential views of a typical three-coil DC motor viewed from the end of its shaft (the shaft itself is not shown). Magnetic effects cause the rotor to turn, which switches the current to the coils via the commutator at the center.*

two extremely powerful neodymium magnets, just visible on the inside of the motor's casing (at top-left) in conjunction with five coils on the rotor.

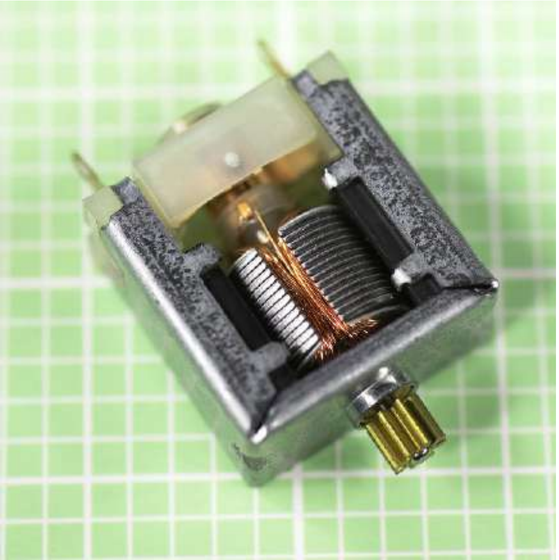

Figure 22-3. *A miniature 1.5VDC motor measuring about 0.7" square.*

Variants

Coil Configurations

The series connection of coils used in Figure 22-2 is known as the *delta configuration*. The alternative is the *wye configuration* (or *Y configuration*, or *star configuration*). Simplified schematics are shown in Figure 22-5. Generally speaking, the delta configuration is best suited to high-speed applications, but provides relatively low torque at low speed. The wye configuration provides higher torque at low speed, but its top speed is limited.

Gearhead Motor

A *gearhead motor* (also often known as a *gear motor*) incorporates a set of reduction gears that increase the torque available from the output shaft while reducing its speed of rotation. This is often desirable as an efficient speed for a traditional DC motor may range from 3,000 to 8,000 RPM, which is too fast for most applications. The gears and the motor are often contained in a single sealed cylindrical package. Two examples are shown in Figure 22-6. A disassembled motor, revealing half of its gear train under the cap and the other half still attached to a separate circular plate, appears in Figure 22-7. When the motor is assembled, the gears engage. As in the case of the bilge-pump motor, the stator magnets are mounted inside the cylindrical casing. Note that the brushes, inside the circular plate of white plastic, have a resistor and capacitor wired to suppress voltage spikes.

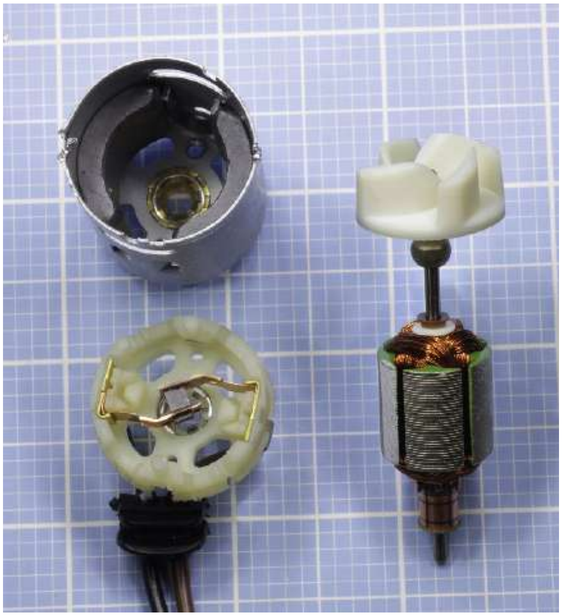

Figure 22-4. *A traditional DC motor removed from its cylindrical casing. The brushes of the motor are attached to the white plastic end piece at bottom-left. Large squares on the graph paper in the background are 1" × 1", divided in 0.1" increments. The motor was used in a small bilge pump.*

Spur gears are widely used for speed reduction. *Planetary gears* (also known as *epicyclic gears*) are a slightly more expensive option. Spur gears such as those in Figure 22-8 may require three or more pairs in series. The total speed reduction is

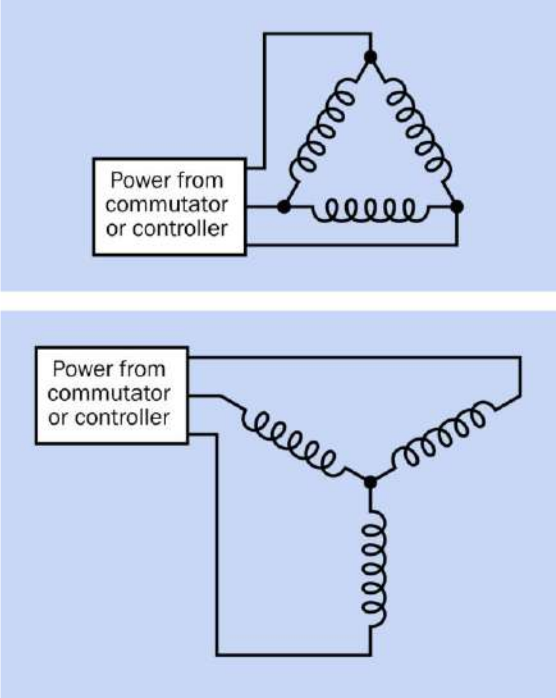

Figure 22-5. *Coils on the rotor of a traditional DC motor may be connected in delta configuration (top) or wye configuration (bottom).*

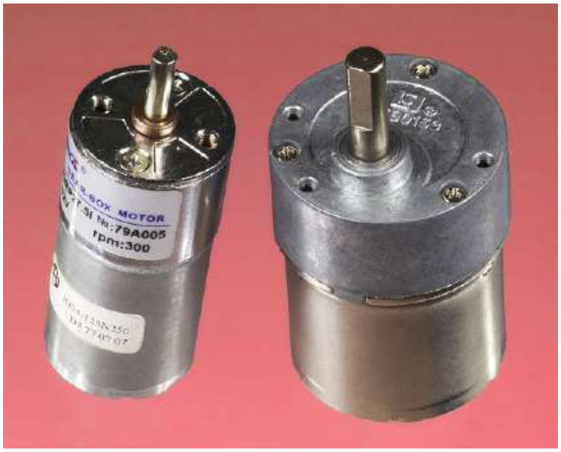

Figure 22-6. *Two typical small gearhead motors.*

found by multiplying the individual ratios. Thus, if three pairs of gears have ratios of 37 : 13, 31 : 15, and 39 : 17, the total speed reduction ® is obtained by:

```
R = ( 37 * 31 * 39 ) / ( 13 * 15 * 17 )
```

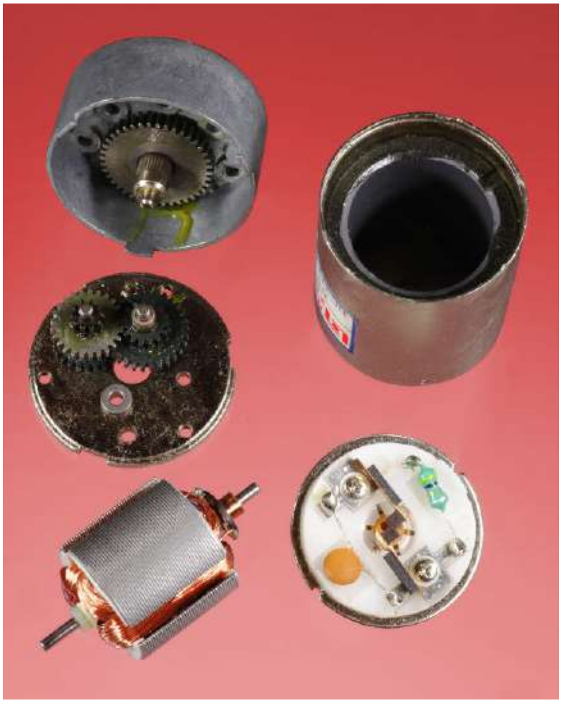

Figure 22-7. *Spur gears from a gearhead motor provide speed reduction and increased torque.*

Therefore:

```
R = 44733 / 3315 = about 13.5 : 1
```

Datasheets almost always express R as an integer. For example, the gear train shown in Figure 22-7 is rated by the manufacturer as having an overall reduction of 50:1. In reality, the reduction can be expected to have a fractional component. This is because if two gears have an integer ratio, their operating life will be shortened, as a manufacturing defect in a tooth in the smaller gear will hit the same spots in the larger gear each time it rotates. For this reason, the numbers of teeth in two spur gears usually do not have any common factors (as in the example above), and if a motor rotates at 500 RPM, a gear ratio stated as 50:1 is very unlikely to produce an output of exactly 10

RPM. Since traditional DC motors are seldom used for applications requiring high precision, this is not usually a significant issue, but it should be kept in mind.

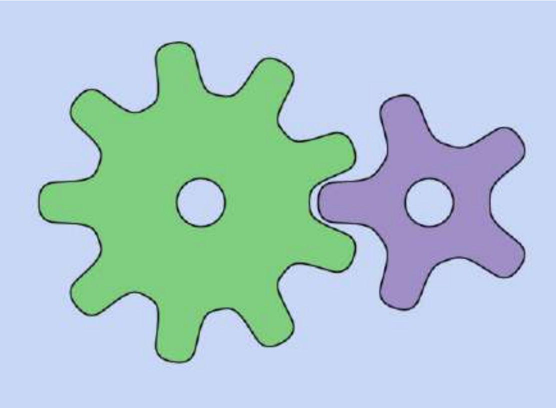

Figure 22-8. *A pair of spur gears.*

Figure 22-9 shows *planetary gears*, also known as *epicyclic gears*. The outer ring gear is properly referred to as the *annulus*, while the *sun gear* is at the center, and the intermediate *planet gears* may be mounted on a *carrier*. The greatest speed reduction will be achieved by driving the sun gear while the annulus is kept in a stationary position and the output is taken from the carrier of the planet gears. If A is the number of teeth in the annulus and S is the number of teeth in the sun gear, the total speed reduction, R, is given by the following formula:

```
R = (S + A) / S
```

Note that in this drive configuration, the number of teeth in each planet gear is irrelevant to the speed reduction. In Figure 22-9, the sun gear has 27 teeth whereas the annulus has 45 teeth. Therefore, the reduction is found by:

```
R = (27 + 45) / 27 = about 2.7 : 1
```

Successive reductions can be achieved by stacking planetary gear sets, using the carrier of one set to drive the sun gear in the next set.

Planetary gears are used primarily if a motor drives a heavy load, as the force is divided among more gear pairs, reducing wear and tear on gear teeth and minimizing the breakdown of lubrication. A planetary gear train may also be more compact than a train of spur gears. These advantages must be evaluated against the higher price and slightly increased friction resulting from the larger number of gears interacting with each other.

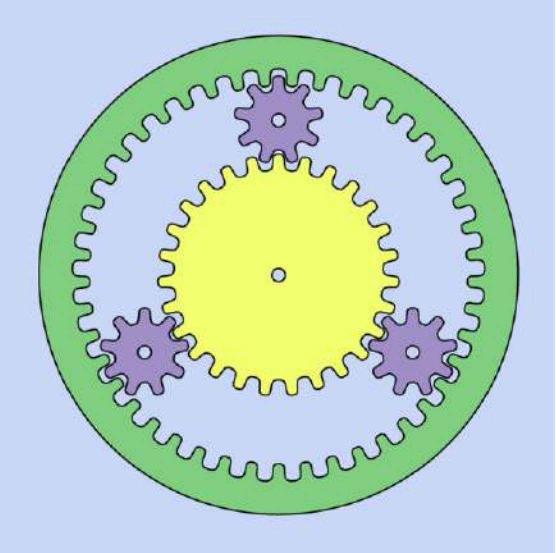

Figure 22-9. *Planetary gears, also known as epicyclic gears, share the torque from a motor among more teeth than simple spur gears.*

Brushless DC Motor

In a *brushless* DC motor, sometimes referred to as a *BLDC motor*, the coils are located in the stator and the permanent magnets are relocated in the rotor. The great advantage of this design is that power can be applied directly to the coils, eliminating the need for brushes, which are the primary source of failure in DC motors as a result of wear and tear. However, since there is no rotating commutator to switch the DC current to the coils, the current must be switched by electronic components, which add to the cost of the motor.

In the *inrunner* configuration the stator surrounds the rotor, whereas in the *outrunner* configuration the stator is located in the center of the motor while the rotor takes the form of a ring or cup that spins around the stator. This is a common design for small cooling fans, where the blades are attached to the outer circumference of a cup that is lined with permanent magnets. An example is shown in Figure 22-10. In this picture, the stator coils are normally hidden from view, being fixed to the fan housing (shown at the top of the picture). Power is controlled by the surface-mount components on the green circular circuit board. The cup attached to the fan blades contains permanent magnets.

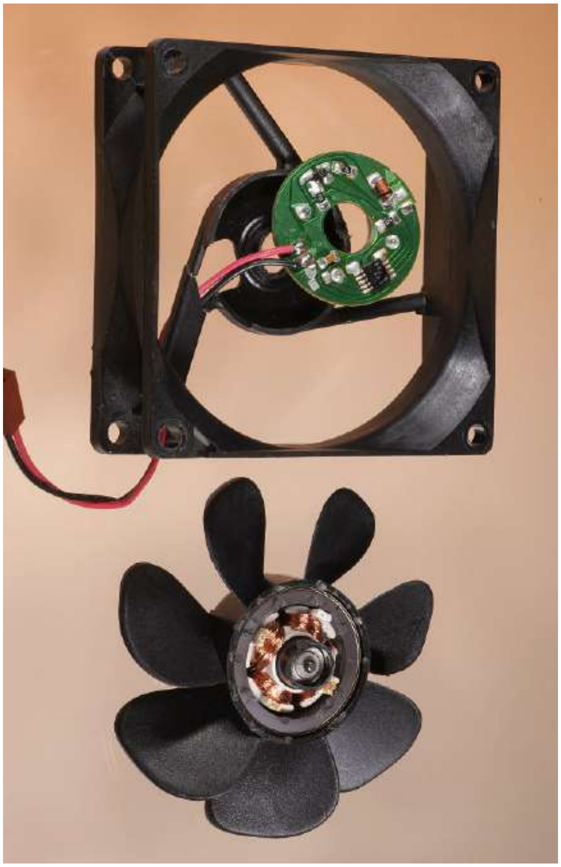

Figure 22-10. *A typical brushless DC cooling fan uses stationary coils, with permanent magnets rotating around them.*

The use of a solid-state switching system to energize the coils sequentially is known as *electronic commutation*. *Hall effect* sensors may be used to detect the position of the rotor and feed this information back to the frequency control circuit, so that it stays "one step ahead" of the rotor (when bringing it up to speed) or is synchronized with the rotor (for a constant running speed). The system is comparable to a *reluctance motor* or *synchronous motor.* These variants are described in the **AC motor** section of this encyclopedia.

While traditional DC motors have been commercially available since the late 1800s, brushless DC motors were not introduced until the 1960s, when the availability of solid-state control electronics began to make the motor design economically viable.

Linear Actuator

Linear actuator is a generic term for any device that can exert a pushing or pulling force in a straight line. In industrial applications, actuators may be powered pneumatically or hydraulically, but smaller-scale units are usually driven by a traditional DC motor. These are more properly (but not often) referred to as *electromechanical linear actuators*.

The rotational force of the motor is typically converted to linear motion by using a threaded motor shaft in conjunction with a nut or collar. The unit is often mounted in an enclosure containing *limit switches* that stop the motor automatically at the limits of travel. For an explanation of limit switches, see "Limit Switches" (page 46) in the **switch** entry in this encyclopedia.

Values

A manufacturer's datasheet should list the maximum operating voltage and typical current consumption when a motor is moderately loaded, along with the *stall current* that a motor draws when it is so heavily loaded that it stops turning. If stall current is not listed, it can be determined empirically by inserting an ammeter (or multimeter set to measure amperes) in series with the motor and applying a braking force until the motor stops. Motors should generally be protected with slow-blowing **fuses** to allow for the power fluctuations that occur when the motor starts running or experiences a change in load.

In addition, the *torque* that a motor can deliver should be specified. In the United States, torque is often expressed in *pound-feet* (or *ounce-inches* for smaller motors). Torque can be visualized by imagining an arm pivoted at one end with a weight hung on the other end. The torque exerted at the pivot is found by multiplying the weight by the length of the arm.

In the metric system, torque can be expressed as gram-centimeters, Newton-meters, or dyne-meters. A Newton is 100,000 dynes. A dyne is defined as the force required to accelerate a mass of 1 gram, increasing its velocity by 1 centimeter per second each second. 1 Newton-meter is equivalent to approximately 0.738 pound-feet.

The speed of a traditional DC motor can be adjusted by varying the voltage to it. However, if the voltage drops below 50% of the rated value, the motor may simply stop.

The power delivered by a motor is defined as its speed multiplied by its torque at that speed. The greatest power will be delivered when the motor is running at half its *unloaded speed* while delivering half the *stall torque*. However, running a motor under these conditions will usually create unacceptable amounts of heat, and will shorten its life.

Small DC motors should be run at 70% to 90% of their unloaded speed, and at 10% to 30% of the stall torque. This is also the range at which the motor is most efficient.

Ideally, DC motors that are used with reduction gearing should be driven with less than their rated voltage. This will prolong the life of the motor.

When choosing a motor, it is also important to consider the *axial loading* (the weight or force that will be imposed along the axis or shaft of the motor) and *radial loading* (the weight or force that will be imposed perpendicularly to the axis). Maximum values should be found in motor datasheets.

In the hobby field, motors for model aircraft are typically rated in watts-per-pound of motor weight (abbreviated w/lb). Values range from 50 to 250 w/lb, with higher values enabling better performance.

Relationships between torque, speed, voltage, and amperage in a traditional DC motor can be described easily, assuming a hypothetical motor that is 100% efficient:

If the amperage is constant, the torque will also be constant, regardless of the motor speed.

If the load applied to the motor remains constant (thus forcing the motor to apply a constant torque), the speed of the motor will be determined by the voltage applied to it.

If the voltage to the motor remains constant, the torque will be inversely proportional with the speed.

How to Use it

A traditional DC motor has the advantages of cheapness and simplicity, but is only suitable for intermittent use, as its brushes and commutator will tend to limit its lifetime. Its running speed will be approximate, making it unsuitable for precise applications.

As the cost of control electronics has diminished, brushless DC motors have replaced traditional DC motors. Their longevity and controllability provide obvious advantages in applications such as hard disk drives, variable-speed computer fans, CD players, and some workshop tools. Their wide variety of available sizes, and good power-to-weight ratio, have encouraged their adoption in toys and small vehicles, ranging from remote-controlled model cars, airplanes, and helicopters to personal transportation devices such as the Segway. They are also used in direct-drive audio turntables.

Where an application requires the rotation of a motor shaft to be converted to linear motion, a prepackaged *linear actuator* is usually more reliable and simpler than building a crank and connecting rod, or cam follower, from scratch. Large linear actuators are used in industrial automation, while smaller units are popular with robotics hobbyists and can also be used to control small systems in the home, such as a remote-controlled access door to a home entertainment center.

Speed Control

A *rheostat* or **potentiometer** may be placed in series with a traditional DC motor to adjust its speed, but will be inefficient, as it will achieve a voltage drop by generating heat. Any rheostat must be rated appropriately, and should probably be wire-wound. The voltage drop between the wiper and the input terminal of the rheostat should be measured under a variety of operating conditions, along with the amperage in the circuit, to verify that the wattage rating is appropriate.

Pulse-width modulation (*PWM*) is preferable as a means of speed control for a traditional DC motor. A circuit that serves this purpose is sometimes referred to as a *chopper*, as it chops a steady flow of current into discrete pulses. Usually the pulses have constant frequency while varying in duration. The pulse width determines the average delivered power, and the frequency is sufficiently high that it does not affect smoothness of operation of the motor.

A **programmable unijunction transistor** or PUT can be used to generate a train of pulses, adjustable with a potentiometer attached to its emitter. Output from the transistor goes to a *silicon-controlled rectifier* (*SCR*), which is placed in series with the motor, or can be connected directly to the motor if the motor is small. See Figure 27-7.

Alternatively, a *555 timer* can be used to create the pulse train, controlling a *MOSFET* in series with the motor.

A **microcontroller** can also be used as a pulse source. Many microcontrollers have PWM capability built in. The microcontroller will require its own regulated power supply (typically 5VDC, 3.3VDC, or sometimes less) and a switching component such as an *insulated-gate bipolar transistor* (*IGBT*) to deliver sufficient power to the motor and to handle the flyback voltage. These components will all add to the cost of the system, but many modern devices incorporate microcontrollers anyway, merely to process user input. Another advantage of using a microcontroller is that its output can be varied by rewriting software, for example if a motor is replaced with a new version that has different characteristics, or if requirements change for other reasons. Additionally, a microcontroller enables sophisticated features such as pre-programmed speed sequences, stored memory of user preferences, and/or responses to conditions such as excessive current consumption or heat in the motor.

A PWM schematic using a microcontroller and IGBT is shown in Figure 22-11.

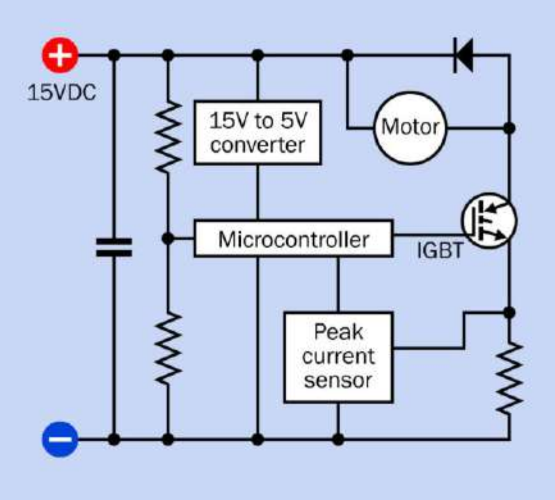

Figure 22-11. *A sample schematic for control of a DC motor via pulse-width modulation, using a microcontroller and an insulated-gate bipolar transistor.*

Direction Control

The *H bridge* is a very early system for reversing the direction of a DC motor simply by swapping the polarity of its power supply. This is shown in Figure 22-12. The switches diagonally opposite each other are closed, leaving the other two switches open; and then to reverse the motor, the switch states are reversed. This is obviously a primitive scheme, but the term "H bridge" is still used when prepackaged in a single chip such as the LMD18200 H bridge motor controller from National Semiconductor.

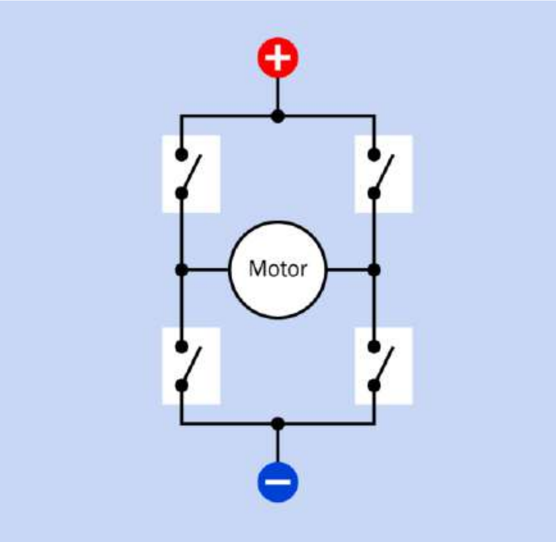

Figure 22-12. *A DC motor can be reversed by this very basic circuit, known as an H bridge, by opening and closing pairs of switches that are diagonally opposite each other.*

A double-throw, double-pole **switch** or **relay** can achieve the same purpose, as shown in Figure 22-13.

Limit Switches

When a traditional DC motor is used reversibly within a restricted range of motion, it can be fitted with *limit switches* to prevent the motor from stalling and burning out at either end of its permitted travel. Limit switches are explained in "Limit Switches" (page 46) in the **switch** entry in this encyclopedia.

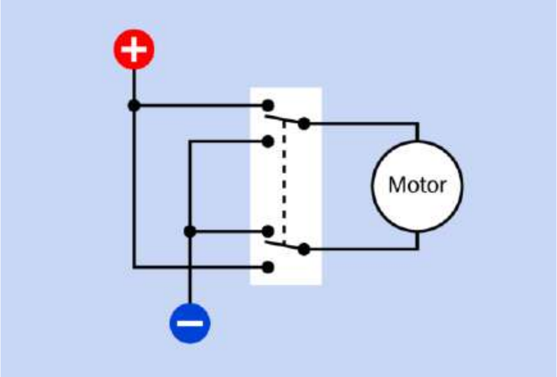

Figure 22-13. *A DPDT switch or relay can reverse the direction of a traditional DC motor simply by swapping the polarity of the power supply.*

What Can Go Wrong

Brushes and Commutator

The primary cause of failure in DC motors is abrasion of the brushes and wear and tear, oxidation, and/or accumulation of dirt on the commutator. Some motors are designed to allow replacement of the brushes; sealed motors and gearhead motors generally are not. High current and high speed will both tend to accelerate wear in the areas where the brushes meet the commutator.

Electrical Noise

The intermittent contact between the brushes and sections of the commutator of a traditional DC motor can induce voltage spikes that may travel back into the power supply for the motor and cause seemingly random effects in other components. Sparking in the commutator can be a significant source of *electromagnetic interference* (*EMI*), especially where cheap or poorly fitted brushes are used. Even if the commutator is running cleanly, the rapid creation of a magnetic field in a motor winding, following by collapse of the field, can create spikes that feed back into the power supply.

Wires that power a motor should be in twisted-pair configuration, so that their radiated EMI tends to cancel itself out. They should be routed away from data lines or encoder outputs, and may be shielded if necessary. Data lines from sensors in brushless motors may also be shielded.

Installing a capacitor across the motor terminals can significantly reduce EMI. Some motors have capacitors preinstalled. If the motor is in a sealed casing, it may have to be disassembled to reveal whether a capacitor is present.

Heat effects

Since all motors in the real world are less than 100% efficient, some power is lost by the motor during normal operation, and will be dissipated as heat. The resistance of motor windings, and consequently the magnetic force that they generate, will decrease as the temperature rises. The motor becomes less efficient, and will try to draw more current, worsening the situation. A manufacturer's rating for maximum temperature should be taken seriously.

The insulation of the coil windings is usually the most vulnerable part of a motor if excess heat persists. Short circuits between adjacent coils as a result of insulation breakdown will degrade the performance of the motor while increasing its power consumption, which will create even more heat.

Where motor casings have protruding ridges, these cooling fins should be exposed to ambient air.

Frequent starting, stopping, and reversing will tend to generate heat as a result of power surges, and will reduce the lifetime of the motor.

Ambient Conditions

A warm, dry environment will tend to dry out bearing lubricants and graphite brushes. Conversely, a very cold environment will tend to thicken the bearing lubricants. If a motor will be used in unusual environmental conditions, the manufacturer should be consulted.

Wrong Shaft Type or Diameter

Motors have a variety of possible output shaft diameters, some measured in inches and others in millimeters, and shafts may be long or short, or may have a D-shaped cross section or splines to mate with appropriate accessories such as gears, pulleys, or couplings. Careful examination of datasheets and part numbers is necessary to determine compatibility. In the hobby-electronics world, retailers may offer purpose-built discs or arms for specific motor shafts.

Incompatible Motor Mounts

Mounting lugs or flanges may or may not be provided, and may be incompatible with the application for which the motor is intended. The same motor may be available with a variety of mount options, differentiated only by one letter or digit in the motor's part number. A mount option that was available in the past may become obsolete or may simply be out of stock. Again, examination of datasheets is necessary.

Backlash

Backlash is the looseness or "slack" in a gear train that results from small gaps between meshing gear teeth. Because backlash is cumulative when gears are assembled in series, it can become significant in a slow-output gearhead motor. When measured at the output shaft, it is generally in the range of 1 to 7 degrees, tending to increase as the load increases. If a geared motor is used as a positioning device, and is fitted with an encoder to count rotations of the motor shaft, control electronics may cause the motor to hunt to and fro in an attempt to overcome the hysteresis allowed by the backlash. A **stepper motor** or **servo motor** is probably better suited to this kind of application.

Bearings

When using a motor that is not rated for significant axial loading, the bearings may be damaged by applying excessive force to push-fit an output gear or pulley onto the motor shaft. Even minor damage to bearings can cause significant noise (see the following section) and a reduced lifetime for the component.

In brushless DC motors, the most common cause of failure is the deterioration of bearings. Attempting to revive the bearings by unsealing them and adding lubricant is usually not worth the trouble.

Audible Noise

While electric motors are not generally thought of as being noisy devices, an enclosure can act as a sounding board, and bearing noise is likely to increase over time. Ball bearings become noisy over time, and gears are inherently noisy.

If a device will contain multiple motors, or will be used in close proximity to people who are likely to be sensitive to noise (for example, in a medical environment), care should be taken to insure that motor shafts are properly balanced, while the motors may be mounted on rubber bushings or in sleeves that will absorb vibration.

AC motor

23

The distinction between AC and DC motors has become blurred as controllers for DC motors make increasing use of *pulse-width modulation*, which can be viewed as a form of alternating current. All motors that consume DC power are referenced in the **DC motor** section of this encyclopedia, regardless of whether they modulate the power internally. **Stepper motors** and **servo motors** are considered as special cases, each with its own entry. AC motors, described here, are those that consume alternating current, usually in the form of a sine wave with a fixed frequency.

OTHER RELATED COMPONENTS

- **DC motor** (see Chapter 22)
- **stepper motor** (see Chapter 25)
- **servo motor** (see Chapter 24)

What It Does

An AC motor uses a power supply of alternating current to generate a fluctuating magnetic field that turns a shaft.

How It Works

The motor consists primarily of two parts: the *stator*, which remains stationary, and the *rotor*, which rotates inside the stator. Alternating current energizes one or more coils in the stator, creating fluctuating magnetic fields that interact with the rotor. A simplified representation is shown in Figure 23-1, where the coils create magnetic forces indicated by the green arrows, N representing North and S representing South.

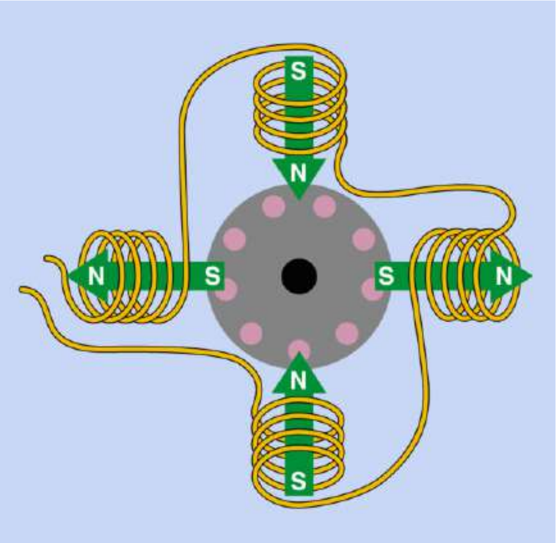

Figure 23-1. *A simplified representation of a basic AC motor. The green arrows indicate magnetic force.*

Stator Design

Plug-in electric fans typically use AC motors. The stator from a large electric fan is shown in Figure 23-2, where the large diameter of each coil maximizes its magnetic effect. The stator from a smaller electric fan is shown in Figure 23-3, in which only one coil is used. (The coil is wrapped in black electrical tape.)

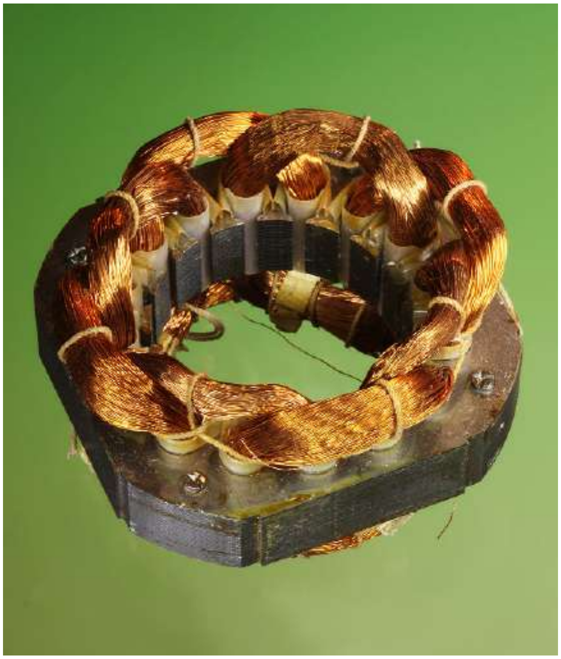

Figure 23-2. *The stator from a large electric fan. Each coil of copper wire is centered on a lug pointing inward to the hole at the center, where the rotor would normally be mounted. The coils overlap because their diameter is maximized to increase their magnetic effect. Each coil is tapped to allow speed selection in steps via an external rotary switch.*

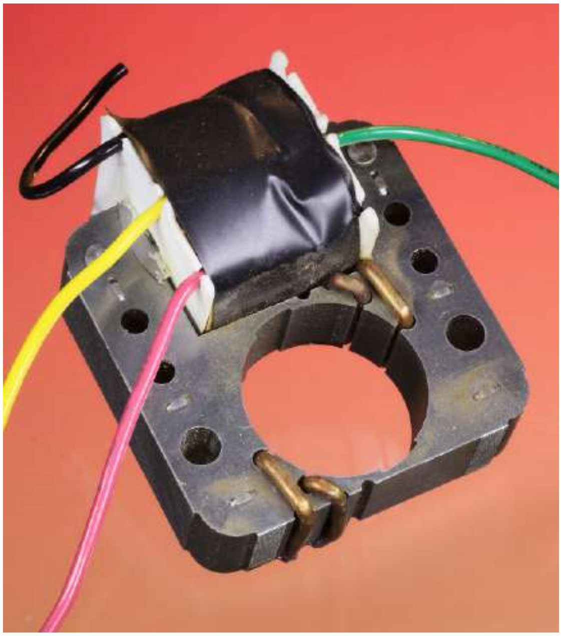

Figure 23-3. *This stator from a smaller electric fan uses only a single coil, wrapped in black tape. It is sufficient to induce a magnetic field but is generally less efficient than a motor using multiple coils.*

The core of a stator resembles the core of a **transformer** in that it usually consists of a stack of wafers of high-silicon steel (or sometimes aluminum or cast iron). The layers are insulated from one another by thin layers of shellac (or a similar compound) to prevent eddy currents that would otherwise circulate through the entire thickness of the stator, reducing its efficiency.

The coil(s) wound around the stator are often referred to as *field windings*, as they create the magnetic field that runs the motor.

Rotor Design

In most AC motors, the rotor does not contain any coils and does not make any electrical connection with the rest of the motor. It is powered entirely by induced magnetic effects, causing this type of motor to be known generally as an *induction motor*.

As the AC voltage changes from positive to negative, the magnetic force induced in the stator collapses and a new field of opposite polarity is created. Because the stator is designed to create an asymmetrical field, it induces a rotating magnetic field in the rotor. The concept of a rotating magnetic field is fundamental in AC motors.

Like the stator, the rotor is fabricated from wafers of high-silicon steel; embedded in the wafers are nonmagnetic rods, usually fabricated from aluminum but sometimes from copper, oriented approximately parallel to the axis of rotation. The rods are shorted together by a ring at each end of the rotor, forming a conductive “cage,” which explains why this device is often referred to colloquially as a *squirrel cage* motor.

Figure 23-4 shows the configuration of a rotor cage with the surrounding steel wafers removed for clarity. In reality, the rods in the cage are almost always angled slightly, as shown in Figure 23-5, to promote smooth running and reduce *cogging*, or fluctuations in torque, which would otherwise occur.

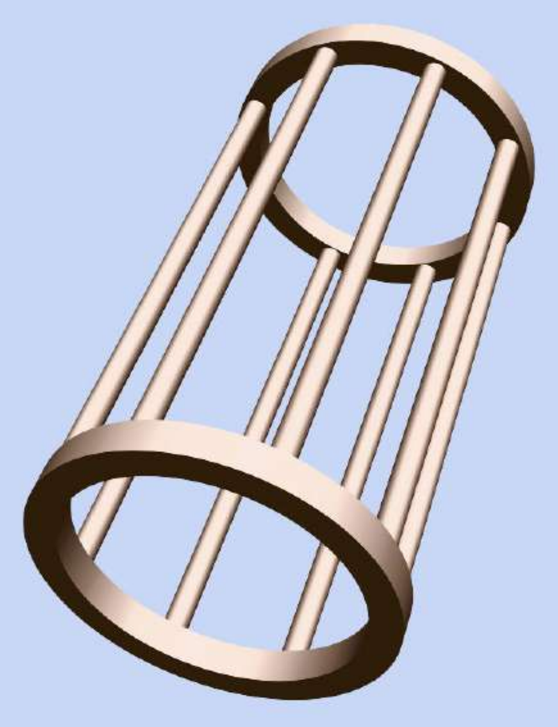

Figure 23-4. *The rotor of a typical AC motor contains a cage of aluminum (or sometimes copper) in which eddy currents occur, as a result of the rotating magnetic field inside the steel body of the rotor (which is omitted here for clarity). These currents cause their own magnetic fields, which interact with the fields generated by coils in the stator.*

In Figure 23-6, the steel wafers of a rotor are shown, with channels to accommodate an angled aluminum cage. Figure 23-7 shows a cross-section of a rotor with the cage elements in pale red and the steel wafer in gray.

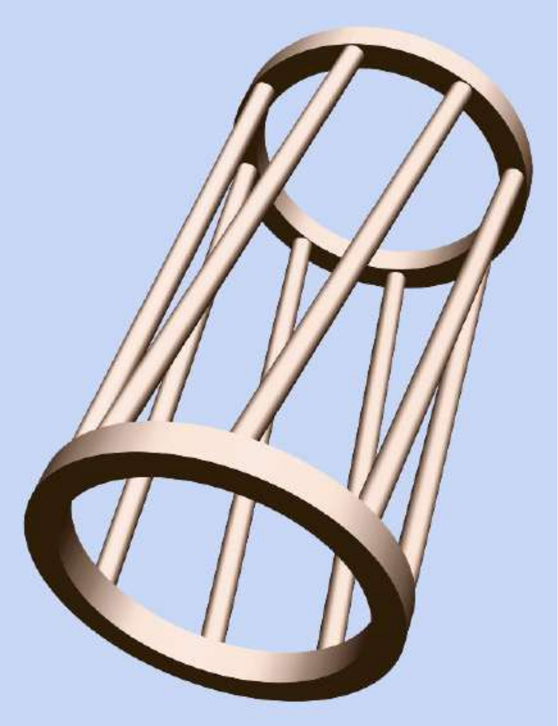

Figure 23-5. *To promote smooth running of the motor, the longitudinal elements of the cage are typically angled, as suggested in this rendering.*

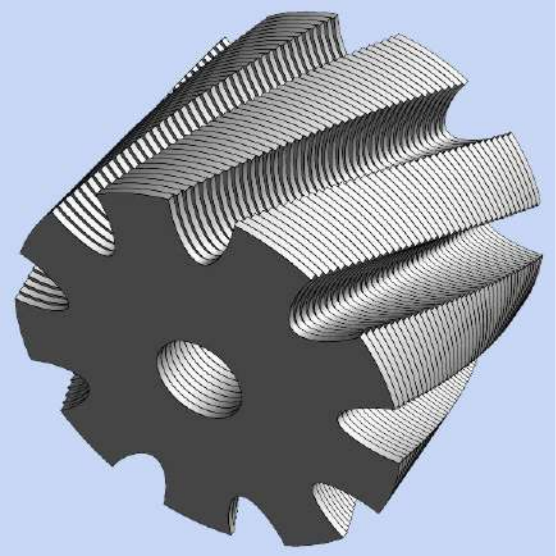

Figure 23-6. *The steel wafers in the rotor of an AC motor are typically offset as shown here. The channels are to accommodate a cage of aluminum or copper conductors.*

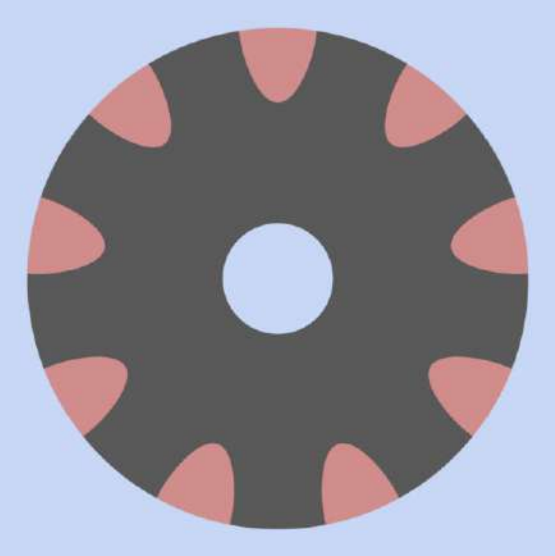

Figure 23-7. *Cross-section of a rotor with steel shown in gray and embedded elements of an aluminum cage shown in pale red.*

The actual rotor from an induction motor is shown in Figure 23-8. This rotor was removed from the stator shown in Figure 23-3. The bearings at either end of the rotor were bolted to the stator until disassembly.

Although the cage is nonmagnetic, it is electrically conductive. Therefore the rotating magnetic field that is induced in the steel part of the rotor generates substantial secondary electric current in the cage, so long as the magnetic field inside the rotor is turning faster than the rotor itself. The current in the longitudinal elements of the cage creates its own magnetic field, which interacts with the fields created by coils in the stator. Attraction and repulsion between these fields causes the rotor to turn.

Note that if the turning speed of the rotor rises to match the frequency of the alternating current powering the coils in the stator, the cage in the rotor is no longer turning through magnetic lines of force, and ceases to derive any power from them. In an ideal, frictionless motor, its unloaded operating speed would be in equilibrium with the AC frequency. In reality, an induction motor never quite attains that speed.

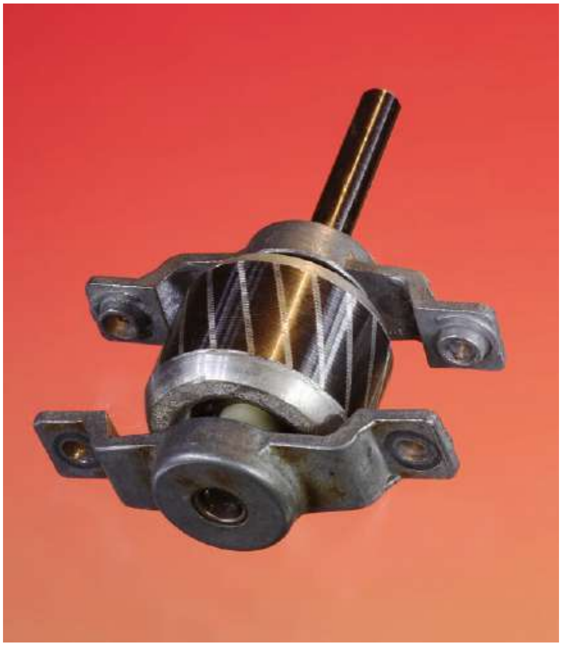

Figure 23-8. *The rotor from a small fan motor. The aluminum cage and its end pieces are the pale gray sections, steel plates are the darker sections.*

When power is applied while the rotor is at rest, the induction motor draws a heavy surge of current, much like a short-circuited **transformer**. Electrically, the coils in the stator are comparable to the primary winding of a transformer, while the cage in the rotor resembles the secondary winding. The turning force induced in the stationary rotor is known as *locked-rotor torque*. As the motor picks up speed, its power consumption diminishes. See Figure 23-9.

When the motor is running and a mechanical load is applied to it, the motor speed will drop. As the speed diminishes, the cage of conductors embedded in the rotor will derive more power, as they are turning more slowly than the rotating magnetic field. The speed of rotation of the field is determined by the frequency of the AC power, and is therefore constant. The difference in rota-

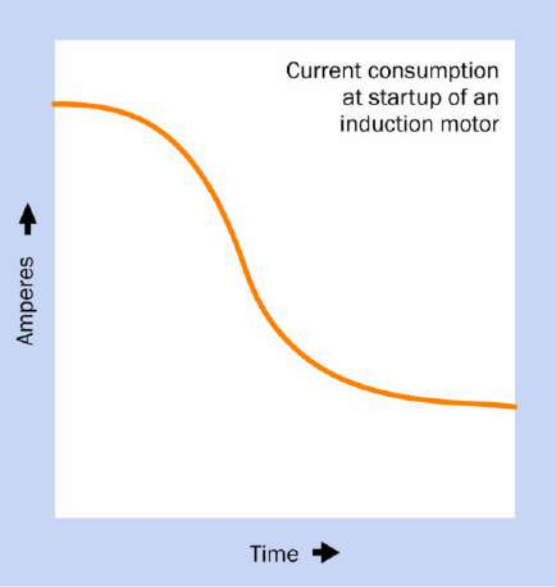

Figure 23-9. *An approximated graph showing the typical current consumption of an AC induction motor as it starts from rest and gains speed over a period of time.*

tional speed between the magnetic field and the rotor is known as *slip*. Higher levels of slip induce greater power, and therefore the induction motor will automatically find an equilibrium for any load within its designed range.

When running under full load, a small induction motor may have a slip value from 4 to 6 percent. In larger motors, this value will be lower.

Variants

Variants of the generic induction motor described above are generally designed to take advantage of either *single-phase* or *three-phase* alternating current.

A *synchronous motor* is a variant in which the rotor maintains a constant speed of rotation regardless of small fluctuations in load.

Some AC motors incorporate a *commutator*, which allows an external connection to coils mounted on the rotor, and can enable variable speed control.

A *linear motor* may consist of two rows of coils, energized by a sequence of pulses that can move a permanent magnet or electromagnet between the coils. Alternatively, the linear motor's coils may move as a result of magnetic interaction with a segmented fixed rail. Detailed description of linear motors is outside the scope of this encyclopedia.

Single-Phase Induction Motor

The majority of induction motors run on single-phase alternating current (typically, from domestic wall outlets). This type of motor is not innately self-starting because the stator coils and rotor are symmetrical. This tends to result in vibration rather than rotation.

To initiate rotation, the stator design is modified so that it induces an asymmetrical magnetic field, which is more powerful in one direction than the other. The simplest way to achieve this is by adding one or more *shorting coils* to the stator. Each shorting coil is often just a circle of heavy-gauge copper wire. This ploy reduces the efficiency of the motor and impairs its starting torque, and is generally used in small devices such as electric fans, where low-end torque is unimportant. Because the shorting coil obstructs some of the magnetic field, this configuration is often known as a *shaded pole* motor.

Copper shorting coils are visible in the fan motor shown in Figure 23-3.

A **capacitor** is a higher-cost but more efficient alternative to a shorting coil. If power is supplied to one or more of the stator coils through a capacitor, it will create a phase difference between these coils and the others in the motor, inducing an asymmetrical magnetic field. When the motor reaches approximately 80% of its designed running speed, a centrifugal switch may be included to take the capacitor out of the circuit, since it is no longer necessary. Switching out the capacitor and substituting a direct connection to the stator coils will improve the efficiency of the motor.

A third option to initiate rotation is to add a second winding in the stator, using fewer turns of smaller-gauge wire, which have a higher resistance than the main winding. Consequently the magnetic field will be angled to encourage the motor to start turning. This configuration is known as a *split-phase* induction motor, in which the starter winding is often referred to as the *auxilliary winding* and consists of about 30% of the total stator windings in the motor. Here again, a centrifugal switch can be incorporated, to eliminate the secondary winding from the circuit when the motor has reached 75 to 80 percent of its designed running speed.

The relationship between motor speed and torque of the three types of motors described above is shown in Figure 23-10. These curves are simplified and do not show the effect that would be produced by introducing a centrifugal switch.

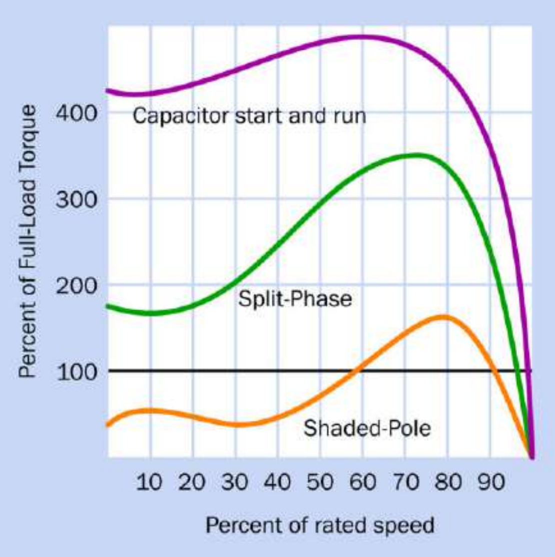

Figure 23-10. *Approximate curves showing the relationship between speed and torque for three types of single-phase induction motor. (Graph derived from AC Induction Motor Fundamentals published by Microchip Technology Inc.)*

Three-Phase Induction Motor

Larger induction motors are often *three-phase* devices. Three-phase AC (which is by far the most common form of *polyphase* AC) is delivered by a power utility company or generator via three wires, each of which carries alternating current with a phase difference of 120 degrees relative to the other two, usually for industrial applications. A common configuration of stator coils for a three-phase motor is shown in Figure 23-11. Since the three wires take it in turns to deliver their peak voltage, they are ideally suited to turn the stator of a motor via induction, and no shorting coil or capacitor is needed for startup. Heavy-duty 3-phase induction motors are extremely reliable, being brushless and generally maintenance-free.

Synchronous Motor

A synchronous motor is a form of induction motor that is designed to reach and maintain equilibrium when the rotor is turning in perfect synchronization with the AC power supply. The speed of the motor will depend on the number of poles (magnetic coils) in the stator, and the number of phases in the power supply. If R is the RPM of a synchronous motor, f is the frequency of the AC current in Hz, and p is the number of poles per phase:

```
R = (120 * f) / p
```

This formula assumes 60Hz AC current. In nations where 50Hz AC is used, the number 120 should be replaced with the number 100.

Two basic types of synchronous motors exist: *direct current excited*, which require external power to start turning, and *non-excited*, which are self-starting. Since non-excited synchronous motors are more common in electronic applications, this encyclopedia will not deal with direct current excited variants.

A *hysteresis motor* is a synchronous motor containing a solid rotor cast from cobalt steel, which has high coercivity, meaning that once it is magnetized, a substantial field is required to reverse the magnetic polarity. Consequently the polarity of the rotor lags behind the constantly changing

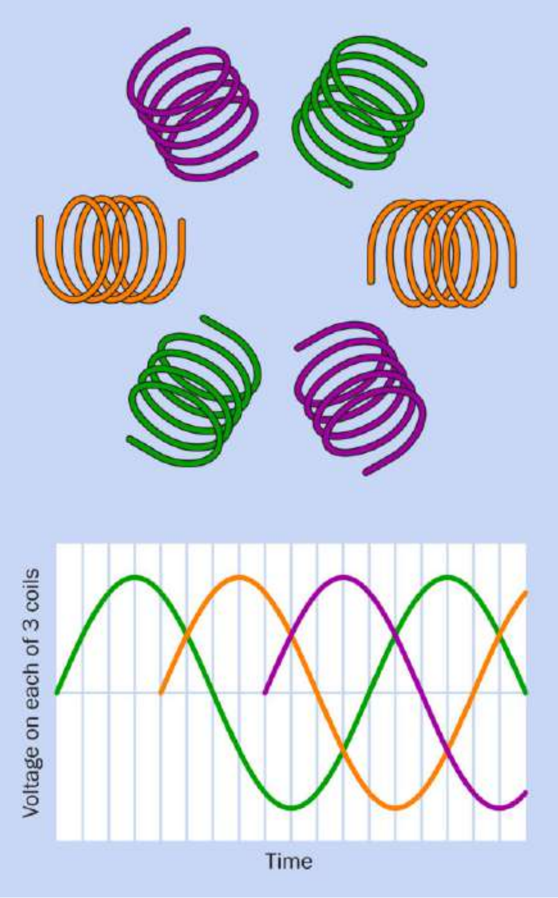

Figure 23-11. *The graph shows voltage delivered via three wires constituting a three-phase power supply. (The curve colors are arbitrary.) A three-phase motor contains a multiple of three coils—often six, as shown here diagrammatically. The three wires of the power supply are connected directly to the coils, which induce a rotating magnetic field.*

polarity of the stator, creating an attracting force that turns the rotor. Because the lag angle is independent of motor speed, this motor delivers constant torque from startup.

Reluctance Motor

Reluctance is the magnetic equivalent to electrical resistance. If a piece of iron is free to move in a magnetic field, it will tend to align itself with the field to reduce the reluctance of the magnetic circuit. This principle was used in very early reluctance motors designed to work from AC and has been revived as electronics to control variable frequency drives have become cheaper.

The simplest reluctance motor consists of a soft iron rotor with projecting lugs, rotating within a stator that is magnetically energized with its own set of inwardly projecting poles. The rotor tends to turn until its lugs are aligned with the poles of the stator, thus minimizing the reluctance.

A basic reluctance motor design is shown in Figure 25-2. It is located in the **stepper motor** section of this encyclopedia, as stepper motors are a primary application of the reluctance principle.

Although a reluctance motor can be used with polyphase fixed-frequency AC power, a variable frequency drive greatly enhances its usefulness. The timing of the frequency is adjusted by the speed of the motor, which is detected by a sensor. Thus the energizing pulses can remain “one step ahead” of the rotor. Since the rotor is not a magnet, it generates no back-EMF, allowing it to reach very high speeds.

The simplicity of the motor itself is a compensating factor for the cost of the electronics, as it requires no commutator, brushes, permanent magnets, or rotor windings. Characteristics of reluctance motors include:

- Cheap parts, easily manufactured, and high reliability.
- Compact size and low weight.
- Efficiencies greater than 90% possible.
- Capable of high start-up torque and high speed operation.

Disadvantages include noise, cogging, and tight manufacturing tolerances, as the air gap between the rotor and stator must be minimized.

A reluctance motor can function synchronously, if it is designed for that purpose.

Variable Frequency Drive

A basic induction motor suffers from significant problems. The surge of power that it draws when starting from rest can pull down the supply voltage enough to affect other devices that share the AC power supply. (Hence, the brief dimming of lights that may occur when the compressor in an air conditioner or refrigerator starts running.) While the motor is turning, it can introduce electrical noise, which feeds back into the power supply, once again causing potential problems for other devices. In addition, the narrow range of speed of an AC induction motor is a great disadvantage.

The advent of cheap solid-state technology encouraged the development of *variable-frequency power supplies* for induction motors. Because the impedance of the motor will diminish as the frequency diminishes, the current drawn by the motor will tend to increase. To prevent this, a variable frequency supply also varies the voltage that it delivers.

Wound-Rotor AC Induction Motor

The stator of this variant is basically the same as that of a single-phase induction motor, but the rotor contains its own set of coils. These are electrically accessible via a *commutator* and *brushes*, as in a **DC motor**. Because the maximum torque (also known as *pull-out* torque) will be proportional to the electrical resistance of the coils in the rotor, the characteristics of the motor can be adjusted by adding or removing resistance externally, via the commutator. A higher resistance will enable greater torque at low speed when the slip between the rotor speed and rotation of the magnetic field induced by the stator is greatest. This is especially useful in corded power tools such as electric drills, where high torque at low speed is desirable, yet the motor can accelerate to full speed quickly when the external resistance is reduced. Typically, the resistance is adjusted via the trigger of the drill.

Figure 23-12 shows a wound-rotor AC induction motor. The disadvantage of this configuration is the brushes that supply power to the rotor will eventually require maintenance. Much larger wound-rotor motors are also used in industrial applications such as printing presses and elevators, where the need for variable speed makes a simple three-phase motor unsuitable.

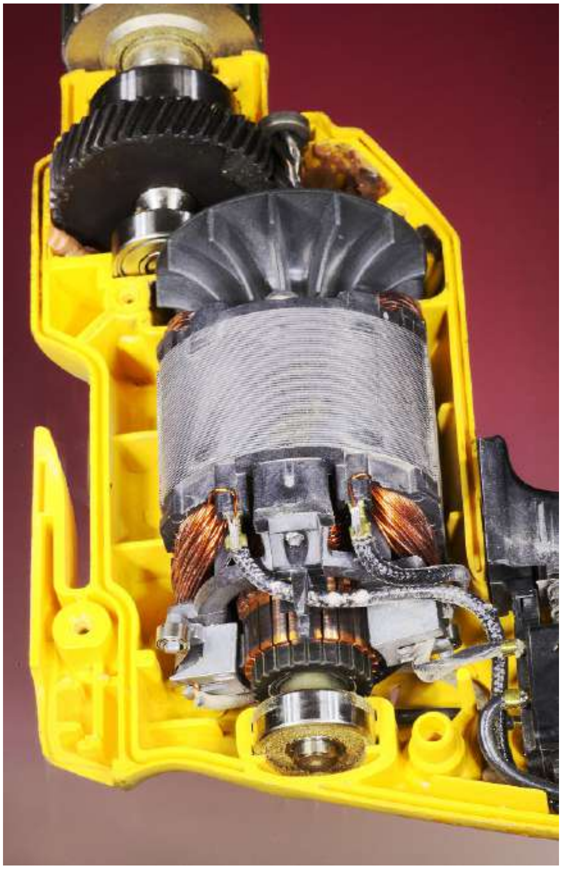

Figure 23-12. *A motor in a corded electric drill uses coils in a brushed rotor to enable variable speed output. In most AC motors, the speed is not adjustable and the rotor does not make any electrical connection with the rest of the motor.*

Universal Motor

A wound-rotor motor may also be described as a *universal motor* if its rotor and stator coils are connected in series. This configuration allows it to be powered by either AC or DC.

DC supplied to the rotor and the stator will cause mutual magnetic repulsion. When the rotor turns, the brushes touching the split commutator reverse the polarity of voltage in the rotor coils, and the process repeats. This configuration is very similar to that of a conventional DC motor, except that the stator in a universal motor uses electromagnets instead of the permanent magnets that are characteristic of a **DC motor**.

When powered by AC, the series connection between stator and rotor coils insures that each pulse to the stator will be duplicated in the rotor, causing mutual repulsion. The addition of a shorting coil in the stator provides the necessary asymmetry in the magnetic field to make the motor start turning.

Universal motors are not limited by AC frequency, and are capable of extremely high-speed operation. They have high starting torque, are compact, and are cheap to manufacture. Applications include food blenders, vacuum cleaners, and hair dryers. In a workshop, they are found in routers and miniature power tools such as the Dremel series.

Because a universal motor requires commutator and brushes, it is only suitable for intermittent use.

Inverted AC Motors

Some modern domestic appliances may seem to contain an AC motor, but in fact the AC current is *rectified* to DC and is then processed with pulse-width modulation to allow variable speed control. The motor is really a **DC motor**; see the entry on this type of motor for additional information.

Values

Because a basic AC induction motor is governed by the frequency of the power supply, the speed of a typical four-pole motor is limited to less than 1,800 RPM (1,500 RPM in nations where 50Hz AC is the norm).

Variable-frequency, universal, and wound-rotor motors overcome this limitation, and can reach speeds of 10,000 to 30,000 RPM. Synchronous motors typically run at 1,800 or 1,200 RPM, depending on the number of poles in the motor. (They run at 1,500 or 1,000 RPM in locations where the frequency of AC is 50Hz rather than 60Hz).

For a discussion of the torque that can be created by a motor, see "Values" (page 184) in the **DC motor** entry in this encyclopedia.

How to Use It

Old-fashioned record players (where a turntable supports a vinyl disc that must rotate at a fixed speed) and electric clocks (of the analogue type) were major applications for synchronous motors, which used the frequency of the AC power supply to control motor speed. These applications have been superceded by CD players (usually powered by brushless DC motors) and digital clocks (which use crystal oscillators).

Many home appliances continue to use AC-powered induction motors. Small cooling fans for use in electronic equipment are sometimes AC-powered, reducing the current that must be provided by the DC power supply. An induction motor generally tends to be heavier and less efficient than other types, and its speed limit imposed by the frequency of the AC power supply is a significant disadvantage.

A simple induction motor cannot provide the sophisticated control that is necessary in modern devices such as CD or DVD players, ink-jet printers, and scanners. A **stepper motor**, **servo motor**, and **DC motors** controlled with *pulse-width modulation* are preferable in these applications.

A *reluctance motor* may find applications in high-speed, high-end equipment including vacuum cleaners, fans, and pumps. Large variable reluctance motors, with high amperage ratings, may be used to power vehicles. Smaller variants are being used for power steering systems and windshield wipers in some automobiles.

What Can Go Wrong

Compared with other devices that have moving parts, the brushless induction motor is one of the most reliable and efficient devices ever invented. However, there are many ways it can be damaged. General problems affecting all types of motors are listed at "Heat effects" (page 188). Issues relating specifically to AC motors are listed below.

Premature Restart

Large industrial three-phase induction motors can be damaged if power is reapplied before the motor has stopped rotating.

Frequent Restart

If a motor is stopped and started repeatedly, the heat that is generated during the initial surge of current is likely to be cumulative.

Undervoltage or Voltage Imbalance

A voltage drop can cause the motor to draw more current than it is rated to handle. If this situation persists, overheating will result. Problems also are caused in a three-phase motor where one phase is not voltage-balanced with the others. The most common cause of this problem is an open circuit-breaker, wiring fault, or blown fuse affecting just one of the three conductors. The motor will try to run using the two conductors that are still providing power, but the result is likely to be destructive.

Stalled Motor

When power is applied to an induction motor, if the motor is prevented from turning, the conductors in the rotor will carry a large current that is entirely dissipated as heat. This current surge will either burn out the motor or blow a fuse or circuit breaker. Care should be taken, in equipment design, to minimize the risk that an induction motor may stall or jam.

Protective Relays

Sophisticated protective relays are available for industrial 3-phase motors, and can guard against all of the faults itemized above. Their details are outside the scope of this encyclopedia.

Excess Torque

As has been previously noted, the torque of an induction motor increases with the slip (speed difference) between the rotation of the magnetic field and the rotation of the rotor. Consequently, if the motor is overloaded and forced to run more slowly, it can deliver more rotational force. This can destroy other parts attached to the motor, such as drive belts.

Internal Breakage

An overloaded induction motor may suffer some cracking or breakage of its rotor. This may be obvious because of reduced power output or vibration, but can also be detected if the motor's power consumption changes significantly.

servo motor

24

Should be referred to as an *RC servo* if it is intended for use in small devices that are remote-controlled and battery powered. However, in practice, the RC acronym is often omitted.

OTHER RELATED COMPONENTS

- **AC motor** (See Chapter 23)
- **DC motor** (See Chapter 22)
- **stepper motor** (See Chapter 25)

What It Does

A servo motor is actually a combination of a motor, reduction gearing, and miniaturized control electronics, usually packaged together inside a very compact sealed plastic case. The motor itself may be AC or DC, and if DC, it may be brushed or brushless. What distinguishes a servo from other types of motor is that it is not designed for continuous rotation. It is a position-seeking device. Its rotational range may be more than 180 degrees but will be significantly less than 360 degrees. Two typical RC servos are shown in Figure 24-1. A side view of a motor is shown in Figure 24-2.

The electronics inside the motor enclosure interpret commands from an external controller. The command code specifies the desired turn angle measured as an offset either side of the center position of the motor's range. The motor turns quickly to the specified position and stops there. So long as the command signal continues and power to the motor is sustained, the motor will hold its position and "push back" against any external turning force. In the absence of such a force, while the motor is stationary, it will use very little current.

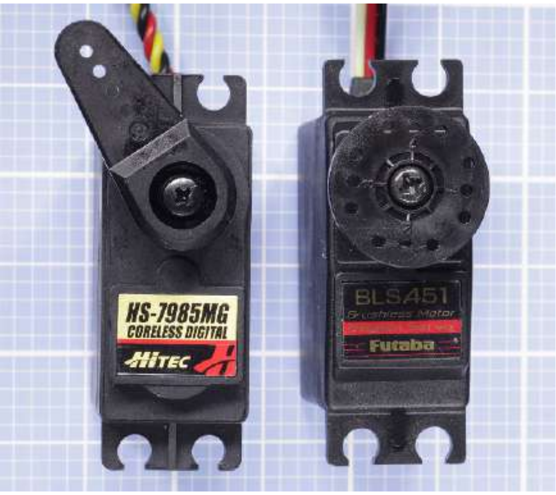

Figure 24-1. *A typical RC servo motor is capable of more than 50 inch-ounces of torque yet can be driven by three or four AA alkaline cells in series, and weighs under 2 ounces.*

The electronics inside a typical RC servo motor are shown in Figure 24-3.

How It Works

Servo motors are generally controlled via *pulse-width modulation* (*PWM*).

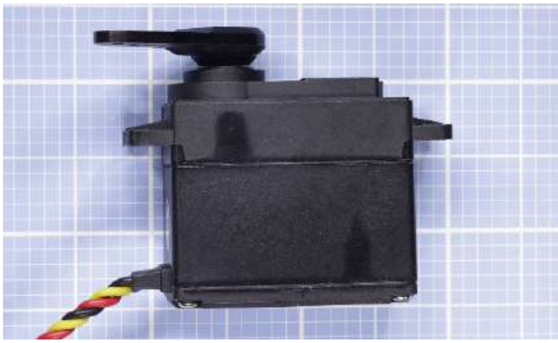

Figure 24-2. *RC servo motors are mostly similar in size. This is a typical side view.*

An industrial servo typically requires a controller that is an off-the-shelf item sold by the manufacturer of the motor. The encoding scheme of the control signals may be proprietary. A heavy-duty servo may be designed to run from three-phase power at a relatively high voltage, and may be used in applications such as production-line automation.

The remainder of this encyclopedia entry will focus primarily on small RC servos rather than industrial servos.

For small RC servos, the stream of control pulses is at a constant frequency of 20ms, with the positive durations of each pulse being interpreted as a positioning command to the motor, and the gaps between the pulses being disregarded. A typical range of pulse widths for a small motor is 1ms to 2ms, specifying a range of -90 to +90 degrees either side of a center location. Many modern motors are capable of excursions beyond these limits, and can be calibrated to establish the precise relationship between pulse width and turn angle. The motor can then be controlled by a lookup table in microcontroller software, or by using a conversion factor between degree-angle and pulse width.

Figure 24-4 shows the typical range of pulse widths within the fixed 20ms period (a frequency of 50Hz) between the start of one pulse and the start of the next, and the meaning of each pulse width to the servo motor. Intermediate pulse widths are interpreted as instructions to rotate to intermediate positions.

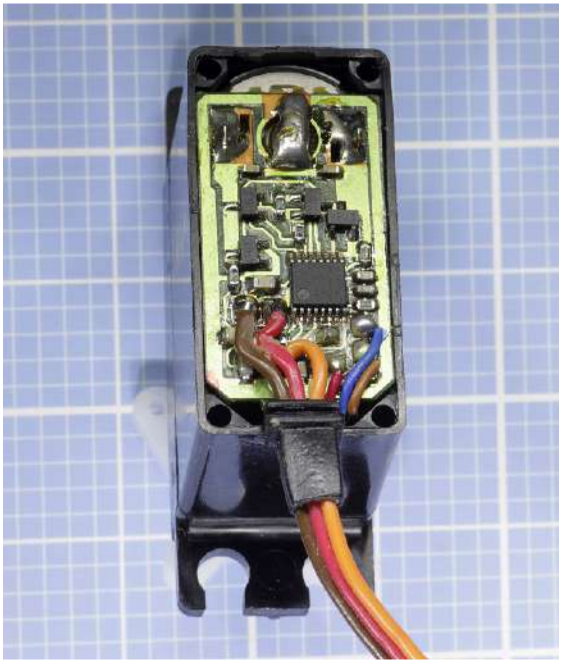

Figure 24-3. *The electronics inside a servo motor decode a stream of pulses that specify the turn angle of the motor.*

Small servo motors require the user to provide a controller that will conform with the above specification. This is often achieved by programming a *microcontroller*, and some microcontroller chips make this especially easy by providing a PWM output specifically tailored to the requirements of an RC servo. Either way, the microcontroller can be directly connected to the servo, enabling an extremely simple and flexible way to manage a positioning device.

Alternatively, a simple pulse generator such as a *555 timer chip* can be used, or controller boards are available from hobbyist supply sources. Some controller boards have USB connections enabling a servo to be governed by computer software.

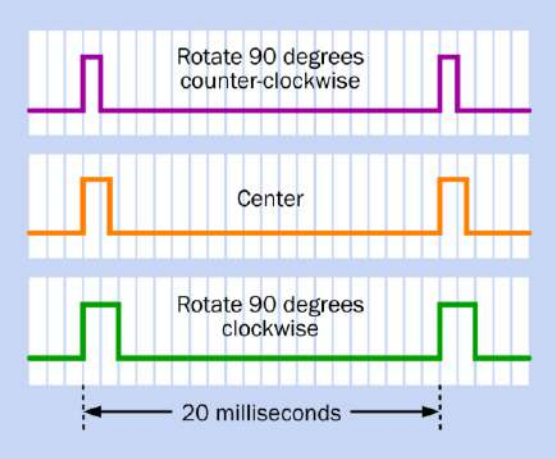

Figure 24-4. *The turn angle of a small RC servo motor is determined by a pulse width from a controller ranging from 1ms to 2ms in duration. The frequency of the pulses is constant at 50Hz.*

In Figure 24-5, a schematic illustrates the connection of a 555 timer with an RC servo, with component values to create a constant frequency of about 48Hz (slightly more than 20ms from peak to peak). The 1μF capacitor in the circuit charges through the 2.2K resistor in series with the diode, which bypasses the 28K resistor. This charging time represents the "on" cycle of the timer chip. The capacitor discharges through the 28K resistor, representing the "off" cycle. The 1K potentiometer, in series with the 5K resistors, acts as a voltage divider applied to the control pin of the timer, adjusting the timer's charge and discharge thresholds. Turning the potentiometer will lengthen or shorten the "on" time of each cycle, without changing the frequency. In practice, because capacitors are manufactured with wide tolerances, the frequency of the timer output cannot be guaranteed. Fortunately most servos will tolerate some inaccuracy.

Since the motor shares the power supply of the timer in this circuit, a protection diode and capacitor have been added between the power supply to the motor and negative ground, to suppress noise and back-EMF.

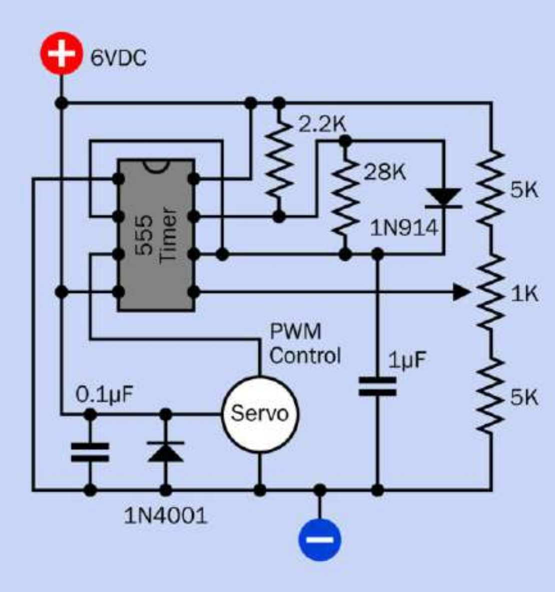

Figure 24-5. *An RC servo can be controlled via a 555 timer with appropriate component values. The potentiometer determines the angular position of the servo.*

Inside a servo motor's casing, the electronics include a **potentiometer** that turns with the output shaft, to provide feedback confirming the motor's position. The limited turning range of the potentiometer determines the turn limits of the motor output shaft.

Variants

Small servos may contain brushed or brushless DC motors. Naturally the brushless motors have greater longevity and create less electrical noise. See the **DC motor** entry in this encyclopedia for a discussion of brushed versus brushless motors.

Servos may use nylon, "Karbonite," or metal reduction gearing. The nylon gears inside a cheaper RC servo are shown in Figure 24-6.

Brushless motors and metal gears add slightly to the price of the motor. Metal gears are stronger than nylon (which can crack under load) but may wear faster, leading to *backlash* and inaccuracy in the gear train. The friction between nylon-to-nylon surfaces is very low, and nylon is certainly

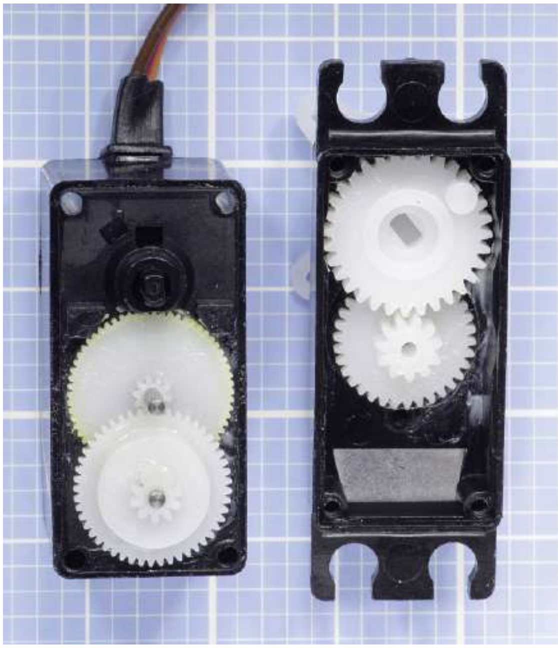

Figure 24-6. *Nylon gearing inside a servo motor.*

adequate and may be preferable if a servo will not be heavily loaded. "Karbonite" is claimed to be five times stronger than nylon and may be a satisfactory compromise. If a gear set experiences a failure (for example, teeth can be stripped as a result of excessive load), manufacturers usually will sell a replacement set to be installed by the user. Installation requires manual dexterity and patience, and some skill.

Servos may have roller bearings or plain sintered bearings, the latter being cheaper but much less durable under side loading.

So-called *digital servos* use faster internal electronics than the older, so-called *analog servos*, and because they sample the incoming pulse stream at a higher frequency, they are more responsive to small, rapid commands from the controller. For this reason they are preferred by hobbyists using servos to control the flight of model airplanes. Externally, the control protocol for digital and analog servos is the same, although a digital servo can be reprogrammed with new code values establishing the limits to its range. A standalone programming unit must be purchased to achieve this.

The most popular manufacturers of small servo motors are Futaba and Hitec. While their control protocols are virtually identical, the motor output shafts differ. The shaft is typically known as the *spline*, and is grooved to fit push-on attachments. The spline of a Futaba motor has 25 grooves, while Hitec uses 24 grooves. Attachments must be appropriate for the brand of motor that has been chosen.

Values

A small servo typically weighs 1 to 2 ounces, has a rotation time of 1 to 2 seconds from one end of its travel to the other, and can exert a surprisingly robust torque of 50 ounce-inches or greater.

Voltage

Small servos were originally designed to run from 4.8V rechargeable batteries in model aircraft. They can be driven with 5VDC to 6VDC on a routine basis. A few servos are designed for higher voltages.

Amperage

The datasheets provided by most manufacturers often fail to specify the power that a servo will draw when it is exerting maximum torque (or indeed, any torque at all). Since small servos are often driven by three or four AA alkaline batteries in series, the maximum current draw is unlikely to be much greater than 1 amp. When the motor is energized but not turning, and is not resisting a turning force, its power consumption is negligible. This feature makes servos especially desirable for remote-controlled battery-powered devices.

Some motors that have a turning range exceeding 180 degrees will respond to pulses of less than 1ms or greater than 2ms. A newly acquired motor should be tested with a microcontroller that steps through a wide range of pulse durations, to determine the limits empirically. Pulses that are outside the motor's designed range will generally be ignored and will not cause damage.

The *turn rate* or *transit time* specified in a datasheet is the time a servo takes to rotate through 60 degrees, with no load on the output shaft. A high-torque servo generally achieves its greater turning force by using a higher reduction gear ratio, which tends to result in a longer transit time.

How to Use it

Typical applications for a small servo include rotating the flaps or rudder of a model aircraft, steering a model boat, model car, or wheeled robot, and turning robotic arms.

A servo generally has three wires, colored red (power supply), black or brown (ground), and orange, yellow, or white (for the pulse train from the controller). The ground wire to the motor must be common with the ground of the controller, and consequently a ceramic bypass **capacitor** of 0.1µF or 0.01µF should be placed between the (red) power wire to the motor and ground. A protection **diode** should also be used. Neither a diode nor a capacitor should be attached to the wire carrying control signals, as it will interfere with the pulse train.

When powering the motor, an AC adapter should only be used with some caution, as its power output may be inadequately smoothed. A voltage regulator is not necessary, but bypass **capacitors** are mandatory. Figure 24-7 shows two hypothetical schematics. The upper section of the figure shows a battery-driven system, possibly using four 1.2V NiMH rechargeable batteries. Since batteries do not generally create voltage spikes, no capacitors are used, but a diode is included to protect the microcontroller from EMF when the servo stops and starts. The lower section of the figure shows the additional precautions that may be necessary when using DC power from an **AC adapter**. The DC-DC **converter**, which derives 6VDC for the motor requires smoothing capacitors (this should be specified in its datasheet), and so does the **voltage regulator**, which delivers regulated 5VDC power to the microcontroller. Once again, the protection diode is included. In both diagrams, the orange wire represents the control wire transmitting pulses to the servo motor.

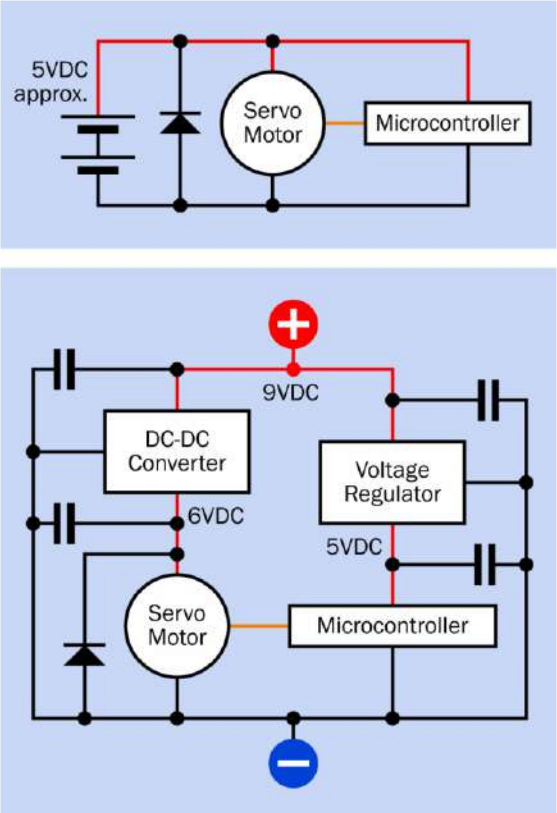

Figure 24-7. *Two possible schematics to run a small servo motor, the upper example using battery power (for example, from four 1.2V NiMH cells) and the lower example using a 9VDC AC adapter. See text for additional explanation.*

Various shaft attachments are available from the same online hobby-electronics suppliers that sell servos. The attachments include discs, single arms, double arms, and four arms in a cross-shaped configuration. A single-arm attachment is often known as a *horn*, and this term may be applied loosely to any kind of attachment. The horn is usually perforated so that other components can be fixed to it by using small screws or nuts and bolts. Figure 24-8 shows a variety of horns.

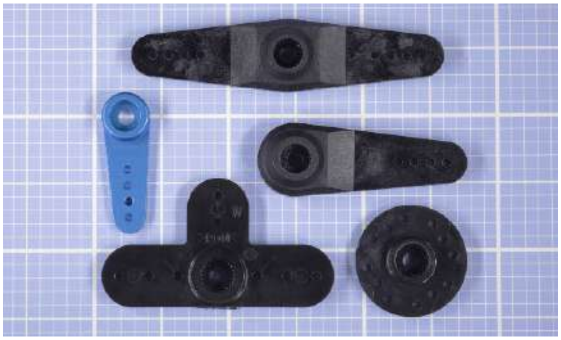

Figure 24-8. *Various shaft attachments, known as horns, are available from motor manufacturers. The blue one is metallic; the others are plastic.*

After the horn is pushed onto the spline (the motor shaft), it is held in place with one central screw. As previously noted, the two major manufacturers of small servos, Futaba and Hitec, have incompatible splines.

Modification for Continuous Rotation

It is possible to modify a small servo motor so that it will rotate continuously.

First the motor case must be opened, and the potentiometer must be centered by using a controller to send some 1.5ms pulses. The potentiometer must then be glued or otherwise secured with its wiper in this precise center position, after which the potentiometer is disconnected from the drive train.

Mechanical stops that would limit the rotation of the motor shaft must be cut away, after which the motor is reassembled. Because the potentiometer has been immobilized, the motor's internal electronics will now "see" the shaft as being in its center position at all times. If the controller sends a pulse instructing the motor to seek a position clockwise or counter-clockwise from center, the motor will rotate in an effort to reach that position. Because the potentiometer will not provide feedback to signal that the motor has achieved its goal, the shaft will continue to rotate indefinitely.

In this mode, the primary distinguishing characteristic of the servo has been disabled, in that it can no longer turn to a specific angle. Also, stopping the servo may be problematic, as it must receive a command that precisely matches the fixed position of the potentiometer. Since the potentiometer may have moved fractionally during the process in which it was immobilized, some trial and error may be needed to determine the pulse width that corresponds with the potentiometer position.

The purpose of modifying a servo for continuous rotation is to take advantage of its high torque, small size, light weight, and the ease of controlling it with a microcontroller.

In response to the interest shown by hobbyists in modifying servos for continuous rotation, some manufacturers now market servos with continuous rotation as a built-in feature. Typically they include a trimmer potentiometer to calibrate the motor, to establish its center-off position.

What Can Go Wrong

Incorrect Wiring

The manufacturer's datasheet should be checked to confirm the color coding of the wires. While a simple DC motor can be reversed by inverting the polarity of its power supply, this is totally inappropriate for a servo motor.

Shaft/Horn Mismatch

Attachments for the spline of one brand of motor may not fit the spline of another brand, and cannot be forced to fit.

Unrealistically Rapid Software Commands

Microcontroller software that positions a servo must allow sufficient time for the servo to respond before the software specifies a new position. It may be necessary to insert delay loops or other wait times in the software.

Jitter

A servo arm that twitches unpredictably usually indicates that the pulse train is being corrupted by external electrical noise. The control wire to the servo should be as short as possible, and should not run closely adjacent to conductors carrying AC or high frequency current switching, or control wires for other servo motors.

Motor Overload

A servo capable of delivering 2 lbs of force 1 inch from its shaft can easily generate enough torque, when it stalls, to break itself free from its mounts, or bend or break any arm or linkage attached to its shaft. Ideally, a relatively "weak link" should be included so that if breakage occurs, it will be predictable and will be relatively easy and cheap to repair.

Unrealistic Duty Cycle

Small servos are designed for intermittent use. Constant cycling will cause wear and tear, especially if the motor has a brushed commutator or metal reduction gears.

Electrical Noise

Brushed motors are always a source of electrical interference, and any servo will also tend to create a voltage dip or surge when it starts and stops. A protection diode may be insufficient to isolate sensitive microcontrollers and other integrated circuit chips. To minimize problems, the servo can be driven by a source of positive voltage that is separate from the regulated power supply used by the chips, and larger filter capacitors may be added to the voltage supply of the microcontroller. A common ground between the motor and the chips is unfortunately unavoidable.

stepper motor

25

Also often referred to as a *stepping motor*, and sometimes known as a *step motor*. It is a type of *induction motor* but merits its own entry in this encyclopedia as it has acquired significant and unique importance in electronics equipment where precise positioning of a moving part is needed and digital control is available.

OTHER RELATED COMPONENTS

- **DC motor** (See Chapter 22)
- **AC motor** (See Chapter 23)
- **servo motor** (See Chapter 24)

What It Does

A stepper motor rotates its drive shaft in precise steps in response to a timed sequence of pulses (usually one step per pulse). The pulses are delivered to a series of coils or *windings* in the *stator*, which is the stationary section of the motor, usually forming a ring around the *rotor*, which is the part of the motor that rotates. Steps may also be referred to as *phases*, and a motor that rotates in small steps may be referred to as having a high *phase count*.

A stepper motor theoretically draws power for its stator coils at a constant level that does not vary with speed. Consequently the torque tends to decrease as the speed increases, and conversely, it is greatest when the motor is stationary or locked.

The motor requires a suitable control system to provide the sequence of pulses. The control system may consist of a small dedicated circuit, or a microcontroller or computer with the addition of suitable driver transistors capable of handling the necessary current. The torque curve of a motor can be extended by using a controller that increases the voltage as the speed of the control pulses increases.

Because the behavior of the motor is controlled by external electronics, and its interior is usually symmetrical, a stepper motor can be driven backward and forward with equal torque, and can also be held in a stationary position, although the stator coils will continue to consume power in this mode.

How It Works

The stator has multiple poles made from soft iron or other magnetic material. Each pole is either energized by its own coil, or more commonly, several poles share a single, large coil. In all types of stepper motor, sets of stator poles are magnetized sequentially to turn the rotor and can remain energized in one configuration to hold the rotor stationary.

The rotor may contain one or more permanent magnets, which interact with the magnetic fields generated in the stator. Note that this is different from a *squirrel-cage* **AC motor** in which a "cage" is embedded in the rotor and interacts with a rotating magnetic field, but does not consist of permanent magnets.

Three small stepper motors are shown in Figure 25-1. Clockwise from the top-left, they are four-wire, five-wire, and six-wire types (this distinction is explained in the following section). The motor at top-left has a threaded shaft that can engage with a collar, so that as the motor shaft rotates counter-clockwise and clockwise, the collar will be moved down and up.

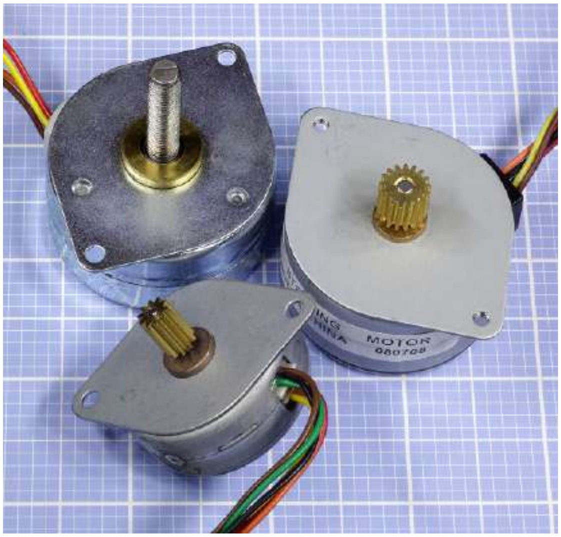

Figure 25-1. *Three small stepper motors.*

Reluctance Stepper Motors

The simplest form of stepper motor uses a rotor that does not contain permanent magnets. It relies on the principle of *variable reluctance*, reluctance being the magnetic equivalent of electrical *resistance*. The rotor will tend to align its protruding parts with the exterior source(s) of the magnetic field, as this will reduce the reluctance in the system. Additional information about variable reluctance is included in "Reluctance Motor" (page 197) in the section of this encyclopedia dealing with the **AC motor**.

A variable reluctance motor requires an external controller that simply energizes the stator coils sequentially. This is shown in Figure 25-2, where six poles (energized in pairs) are arrayed symmetrically around a rotor with four protrusions, usually referred to as *teeth*. Six stator poles and four teeth are the minimum numbers for reliable performance of a reluctance stepper motor.

In the diagram, the core of each pole is tinted green when it is magnetized, and is gray when it is not magnetized. In each section of this diagram, the stator coils are shown when they have just been energized, and the rotor has not yet had time to respond. External switching to energize the coils has been omitted for simplicity. In a real motor, the rotor would have numerous ridges, and the clearance between them and the stator would be extremely narrow to maximize the magnetic effect.

In a 6-pole reluctance motor where the rotor has four teeth, each time the controller energizes a new pair of poles, the rotor turns by 30 degrees counter-clockwise. This is known as the *step angle*, and means that the motor makes 12 steps in each full 360-degree rotation of its shaft. This configuration is very similar to that of a 3-phase AC *induction motor*, as shown in Figure 23-11 in the **AC motor** section of this encyclopedia. However, the AC motor is designed to be plugged into a power source with a constant frequency, and is intended to run smoothly and continuously, not in discrete steps.

Generally, reluctance motors tend to be larger than those with magnetized rotors, and often require feedback from a sensor that monitors shaft angle and provides this information to control electronics. This is known as a *closed loop* system. Most smaller stepper motors operate in an *open loop* system, where positional feedback is considered unnecessary if the number of pulses to the motor is counted as a means of tracking its position.

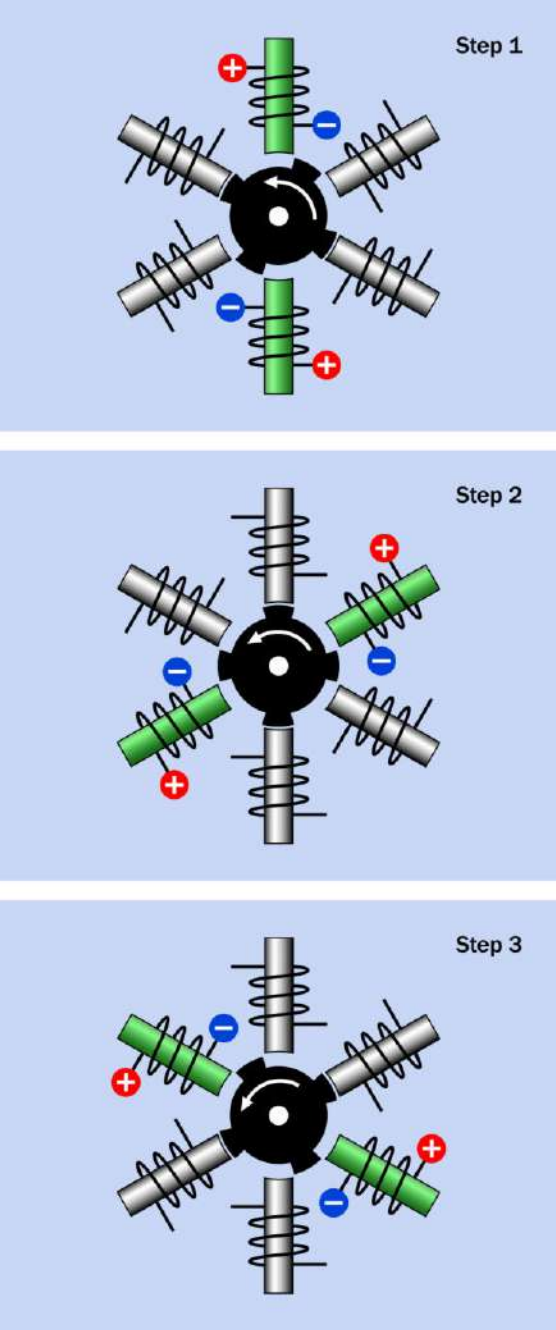

Figure 25-2. *In a variable reluctance stepper motor, the rotor moves to minimize magnetic reluctance each time the next pair of coils is energized. At each step, the coils have been energized a moment before the rotor has had time to respond.*

Permanent Magnet Stepper Motors

More commonly, the rotor of a stepper motor contains permanent magnets, which require the controller to be capable of reversing the magnetic field created by each of the stator coils, so that they alternately attract and repel the rotor magnets.

In a *bipolar* motor, the magnetic field generated by a coil is reversed simply by reversing the current through it. This is shown diagrammatically in Figure 25-3. In a *unipolar* motor, the magnetic field is reversed by applying positive voltage to the center tap of a coil, and grounding one end or the other. This is shown diagrammatically in Figure 25-4.

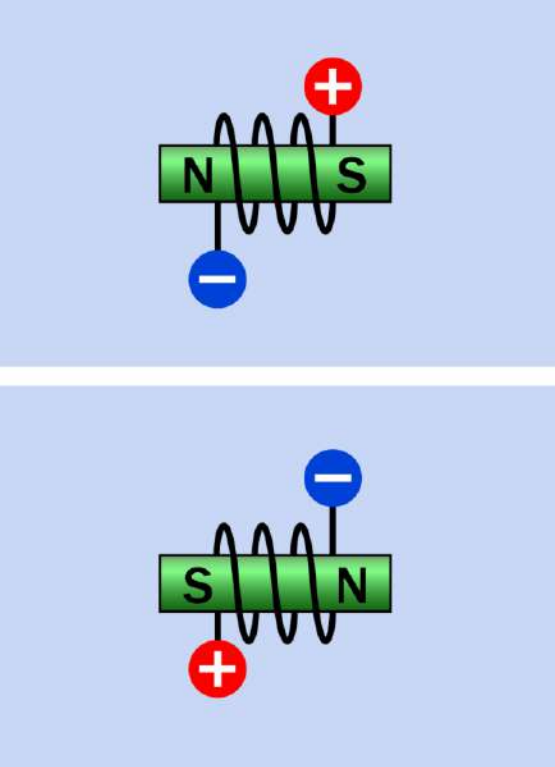

Figure 25-3. *In a bipolar motor, the magnetic field generated by each stator coil is reversed simply by reversing the current through the coil.*

Either type of motor is often designed with an upper and lower deck surrounding a single rotor,

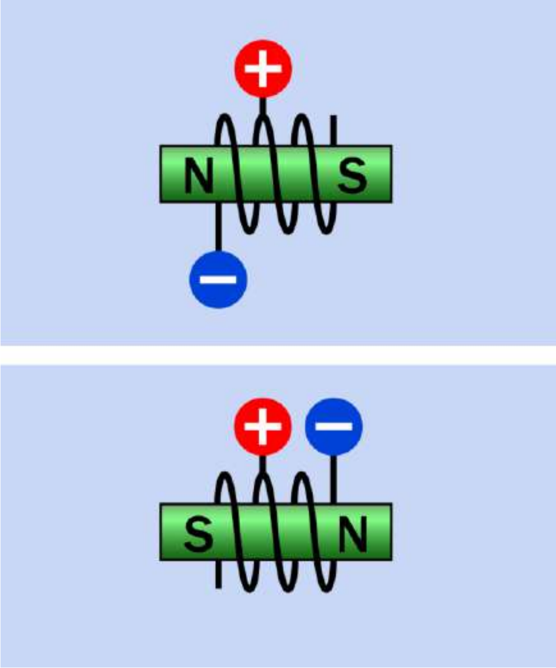

Figure 25-4. *The magnetic field of this coil is reversed by applying positive voltage constantly to a center tap and grounding one end of the coil or the other.*

as suggested in Figure 25-5. A large single coil, or center-tapped coil, induces a magnetic field in multiple poles in the top deck, out of phase by one step with a second set of poles, energized by their own coil, in the bottom deck. (All three motors shown in Figure 25-1 are of this type.) The rotor of the motor is tall enough to span both decks, and is rotated by each of them in turn.

In Figure 25-6, the decks of a two-deck four-wire motor have been split apart. The rotor remains in the left-hand section. It is enclosed within a black cylinder that is a permanent magnet divided into multiple poles. In the right-hand section, a coil is visible surrounding metal "teeth" that function as stator poles when the coil is energized.

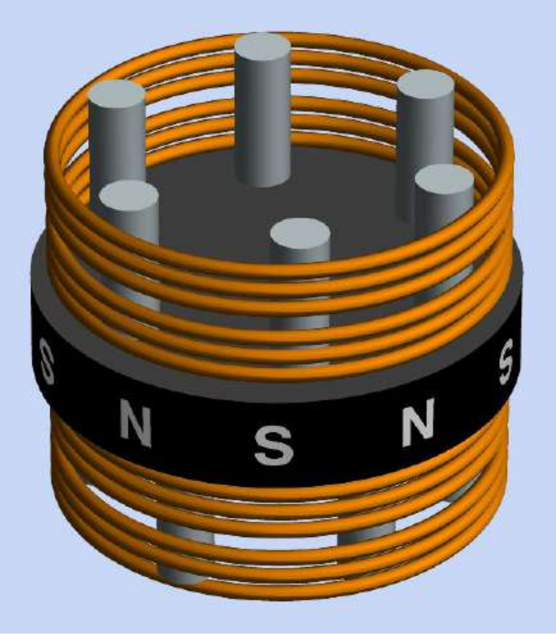

Figure 25-5. *A simplified rendering of the common "two deck" type of motor. See text for details.*

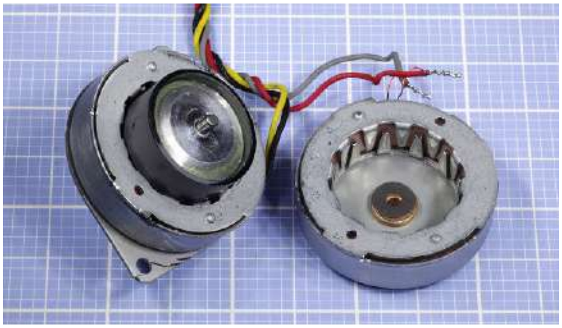

Figure 25-6. *A two-deck stepper motor split open to reveal its rotor (left) and one of the stators (right) encircled by a coil.*

In Figure 25-7, the same motor has been further disassembled. The coil was secured with a length of tape around its periphery, which has been removed to make the coil visible. The remaining half of the motor, at top-right, contains a second, concealed but identical coil with its own set of poles, one step out of phase with those in the first deck.

Because the field effects in a two-deck stepper motor are difficult to visualize, the remaining diagrams show simplified configurations with a minimum number of stator poles, each with its own coil.

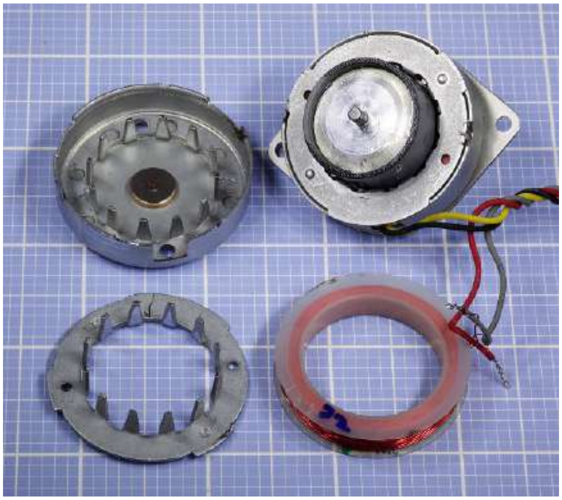

Figure 25-7. *The stepper motor from the previous figure, further disassembled.*

Bipolar Stepper Motors

The most basic way to reverse the current in a coil is by using an *H-bridge* configuration of switches, as shown in Figure 25-8, where the green arrow indicates the direction of the magnetic field. In actual applications, the switches are solid-state. Integrated circuits are available containing all the necessary components to control a bipolar stepper motor.

Four sequential steps of a bipolar motor are shown in Figure 25-9, Figure 25-10, Figure 25-11, and Figure 25-12. The H-bridge control electronics for each coil are omitted for clarity. As before, energized coils are shown with the pole inside the coil tinted green, while non-energized coils are gray, and the rotor is shown before it has had time to respond to the magnetic field in each step.

Unipolar Motors

The control electronics for a unipolar motor can be simpler than those for a bipolar motor, as off-the-shelf switching transistors can ground one end of the coil or the other. The classic five-wire unipolar stepper motor, often sold to hobbyists and used in robotics projects and similar applications, can be driven by nothing more elaborate

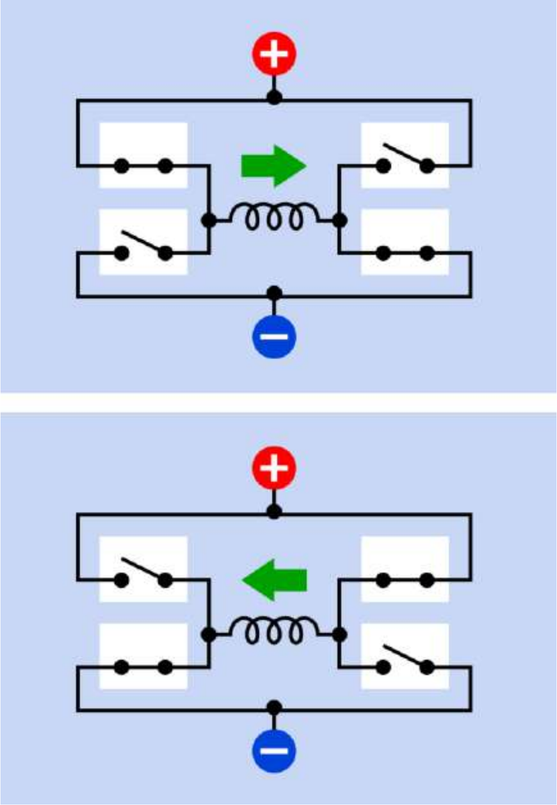

Figure 25-8. *The simplest and most basic way to reverse the current through a coil is via an H-bridge circuit. In practice, the switches are replaced by solid-state components.*

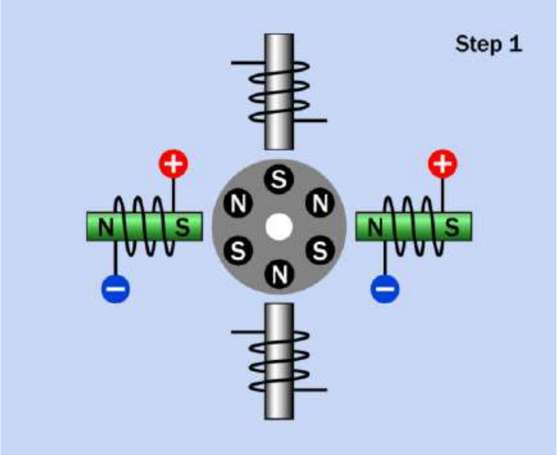

Figure 25-9. *A bipolar stepper motor depicted a moment before the rotor has had time to make its first step in response to magnetic fields created by the stator coils.*

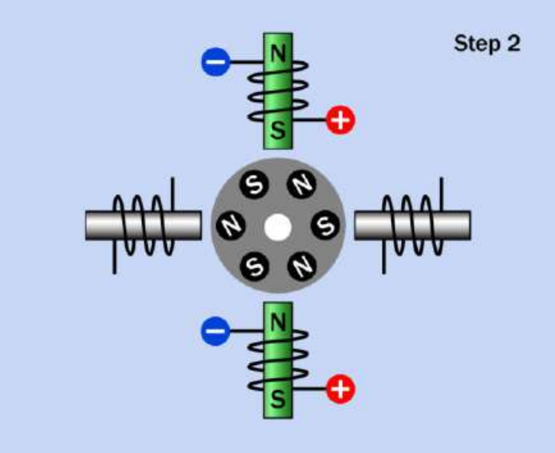

Figure 25-10. *The bipolar stepper motor from the previous figure is shown with its rotor having advanced by one step, and coil polarity changed to induce it to make a second step.*

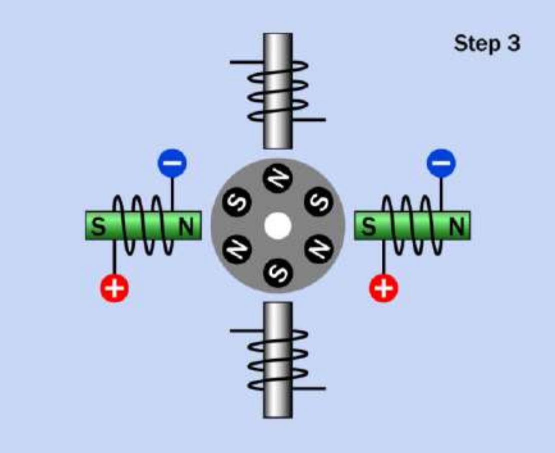

Figure 25-11. *The bipolar stepper motor after taking its second step, immediately before making its third step.*

than a set of *555 timer* chips. However, this type of motor is less powerful for its size and weight because only half of each coil is energized at a time.

In Figure 25-13, Figure 25-14, Figure 25-15, and Figure 25-16, the simplest configuration of a unipolar system is shown in diagrammatic form using four stator coils and a rotor containing six magnetic poles. Each figure shows the stator coils when they have just been energized, a movement before the rotor has had time to move in response to them. Coils that are energized are shown with the metal cores tinted green. Wires that are not conducting current are shown in gray. The open and closed positions of switches a, b, c, and d suggest the path that current is taking along the wires that are colored black.

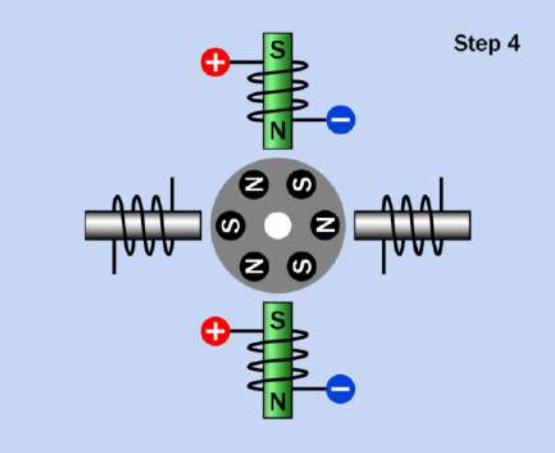

Figure 25-12. *The bipolar stepper motor after taking its third step. When the rotor responds to the new pattern of magnetic fields, its orientation will be functionally identical with that shown in the first step.*

Note that coils on opposite sides of the motor are energized simultaneously, while the other pair of coils is de-energized. Adjusting the controller so that it overlaps the "on" cycles of the coils can generate more torque, while consuming more power.

A motor containing more stator poles can advance in smaller steps, if the poles are separately energized. However, if the coils have individual windings, this will increase the cost of the motor.

Variants

In addition to bipolar and unipolar variants, previously described, three others are available.

High Phase Count

This term describes any type of stepper motor in which additional poles reduce the step size. The advantages of a high phase count include

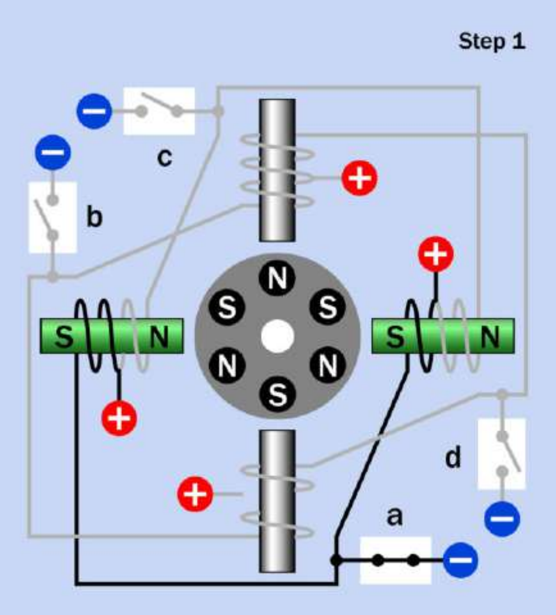

Figure 25-13. *The coils of this unipolar stepper motor are shown an instant after they have been energized, before the rotor has had time to respond by making its first step.*

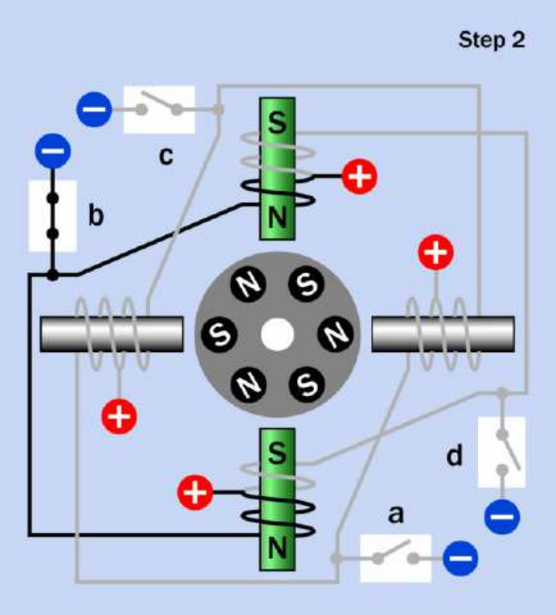

Figure 25-14. *The same motor from the previous figure is shown with coils energized to induce the rotor to make its second step.*

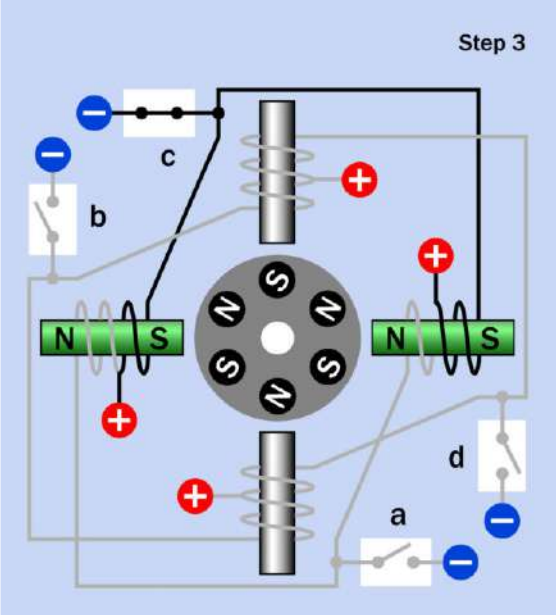

Figure 25-15. *The same motor from the previous figure is shown with coils energized to induce the rotor to make its third step.*

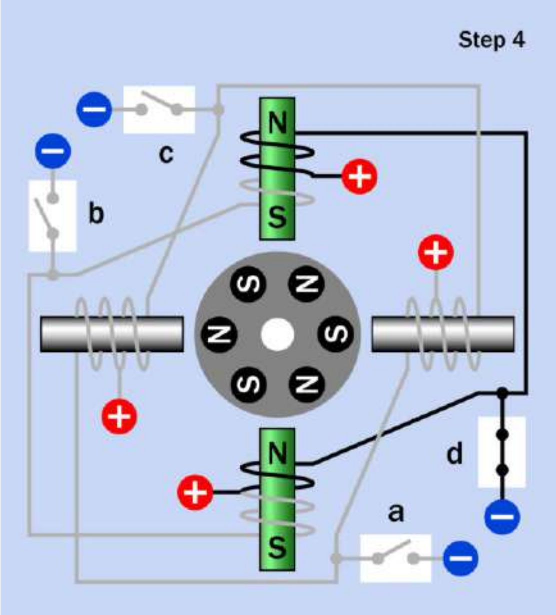

Figure 25-16. *When the rotor makes its fourth step, it will be back in an orientation that is functionally identical with the first figure in this series.*

smoother running at high speed and greater precision when selecting a desired motor position. The additional coils also enable higher power density, but naturally tend to add to the cost of the motor.

Hybrid

This type of motor uses a toothed rotor that provides variable reluctance while also containing permanent magnets. It has become relatively common, as the addition of teeth to the rotor enables greater precision and efficiency. From a control point of view, the motor behaves like a regular permanent-magnet stepper motor.

Bifilar

In this type of motor, also sometimes known as a *universal* stepper motor, two coils are wound in parallel for each stator pole. If there are two poles or sets of poles, and both ends of each winding are accessible via wires that are run out of the motor, there will be eight wires in total. Consequently this type is often referred to as an *8-wire motor*.

The advantage of this scheme is that it allows three possible configurations for the internal coils. By shorting together the wires selectively, the motor can be made to function either in unipolar or bipolar mode.

In Figure 25-17, the upper pair of simplified diagrams depicts one end of one coil connected to the beginning of the other, while positive voltage is applied at the midpoint, as in a unipolar motor. The magnetic polarity of the coil is determined by grounding either end of the coil. The section of each coil that is not conducting current is shown in gray.

The center pair of diagrams shows the adjacent ends of the coils tied together, so that they are now energized in parallel, with the magnetic polarity being determined by the polarity of the voltage, as in a bipolar motor.

The coils may also be connected in series, as shown in the lower pair of diagrams. This will provide greater torque at low speed and lower torque at high speed, while enabling higher-voltage, lower-current operation.

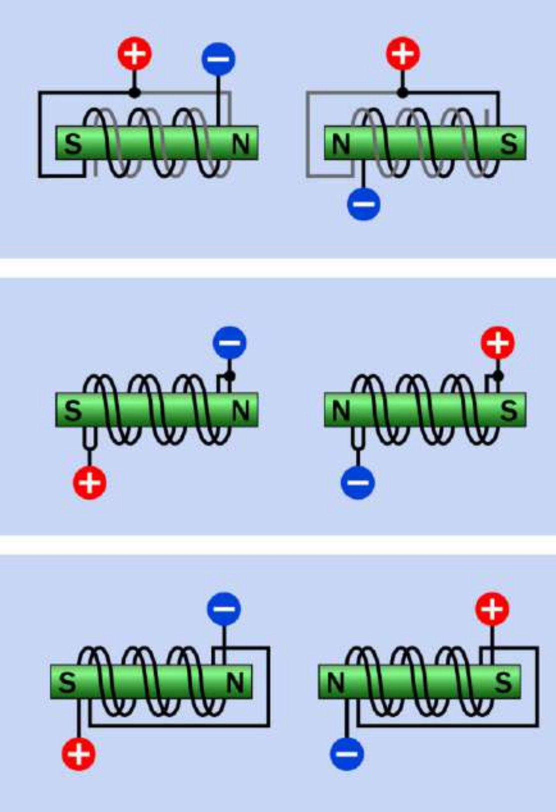

Figure 25-17. *In a bifilar motor, two coils are wound in parallel around each stator pole and can be connected with a center tap to emulate a unipolar motor (upper diagrams), or can be energized in parallel (middle diagrams) or series (lower diagrams) to emulate a bipolar motor.*

Multiphase

In a multiphase motor, multiple stator coils are usually connected in series, with a center tap applied between each pair. A possible configuration is shown in Figure 25-18, where the two diagrams show two consecutive steps in rotation, although the step angle could be halved by changing the voltage polarity in only one location at a time. The way in which the motor is wired enables only one stator coil to be unpowered during any step, because its two ends are at equal potential. Therefore this type of motor is capable of high torque in a relatively small format.

In some multiphase motors, additional wires allow access to both ends of each coil, and the coils are not connected internally. This allows control of the motor to be customized.

Microstepping

An appropriately designed stepper motor can be induced to make very small, intermediate steps if the control voltage is modulated to intermediate levels. Step angles as low as 0.007 degrees are claimed by some manufacturers. However, a motor running in this mode is less able to generate torque.

The simplest form of microstepping is half-stepping. To achieve this in a unipolar motor, each coil passes through an "off" state before its magnetic polarity is reversed.

Sensing and Feedback

So long as the series of pulses to the motor allows the rotor ample time to respond, no feedback mechanism from the rotor is necessary to confirm its position, and an open-loop system is sufficient. If sudden acceleration, deceleration, load fluctuations, and/or rotation reversal will occur, or if high speeds are involved, a *closed loop system*, in which a sensor provides positional feedback, may be necessary.

Voltage Control

Rapid stepping of a motor requires rapid creation and collapse of magnetic fields in the stator windings. Therefore, self-inductance of the windings can limit the motor speed. One way to overcome this is to use a higher voltage. A more sophisticated solution is to use a controller that provides a high initial voltage, which is reduced or briefly interrupted when a sensor indicates that coil current has increased sufficiently to overcome the self-inductance of the windings and has reached its imposed limit. This type of controller may be referred to as a *chopper drive* as the voltage is "chopped," usually by power transistors. It is a form of *pulse width modulation*.

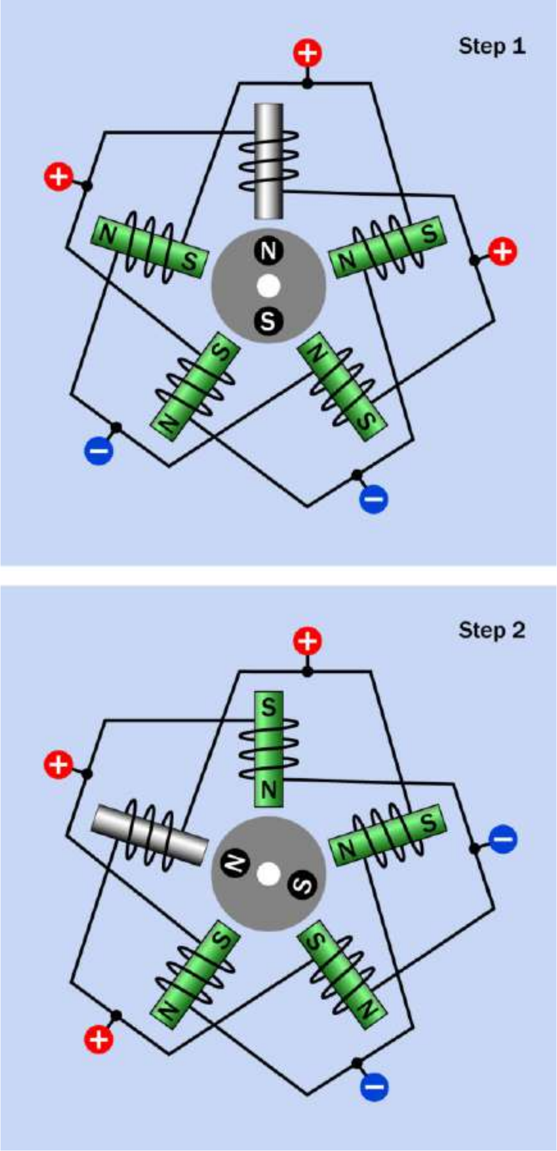

Figure 25-18. *A multiphase stepper motor. By applying voltage in the pattern shown, only one coil is not energized during each step. This enables high torque compared with the size of the motor.*

Values

The *step angle* of a stepper motor is the angular rotation of its shaft, in degrees, for each full step. This will be determined by the physical construction of the motor. The coarsest step angle is 90 degrees, while sophisticated motors may be capable of 1.8 degrees (without microstepping).

The maximum torque that a motor can deliver is discussed in "Values" (page 184) in the **DC motor** entry of this encyclopedia.

Motor weight and size, shaft length, and shaft diameter are the principal passive values of a stepper motor, which should be checked before it is selected for use.

How to Use it

Stepper motors are used to control the seek action in disk drives, the print-head movement and paper advance in computer printers, and the scanning motion in document scanners and copiers.

Industrial and laboratory applications include the adjustment of optical devices (modern telescopes are often oriented with stepper motors), and valve control in fluid systems.

A stepper motor may be used to power a *linear actuator*, usually via a screw thread (properly known as a lead screw) or worm gear. For more on linear actuators, see "Linear Actuator" (page 184). While the stepper motor will enable greater accuracy than a traditional DC motor, the gearing inevitably will introduce some imprecision.

Advantages of stepper motors include:

- Precise positioning, typically within 3 percent to 5 percent per step. The percentage step error does not accumulate as the motor rotates
- Able to run at a wide range of speeds, including very slow speeds without reduction gearing
- Trouble-free start, stop, and reverse action
- Cheap controller hardware where open-loop applications are acceptable
- High reliability, since no brushes or commutator are involved

Disadvantages include:

- Noise and vibration
- Resonance at low speeds
- Progressive loss of torque at high speeds

Protection Diodes

While a small stepper motor may be driven directly from power transistors, *darlington pairs*, or even *555 timers*, larger motors will create *back-EMF* when the magnetic field of each stator coil is induced or forward EMF when the field is allowed to collapse, and bipolar motors will also induce voltage spikes when the current reverses. In a unipolar motor, while only one-half of the coil is actually energized via its center tap, the other half will have an induced voltage, as the coil acts like a *linear transformer*.

A simplified schematic illustrating diode placement for a bipolar motor is shown in Figure 25-19.

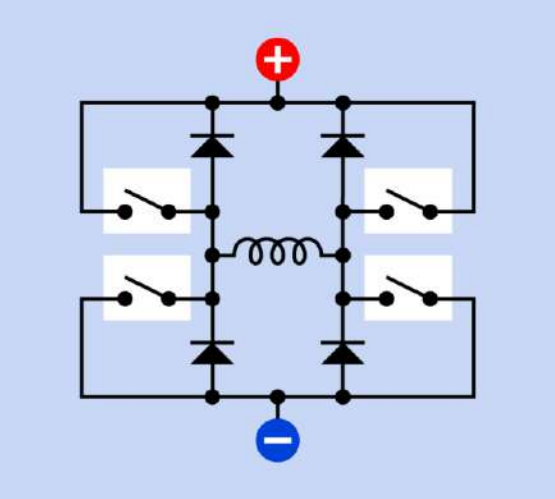

Figure 25-19. *The H-bridge circuit must be augmented with protection diodes to guard against back-EMF created by fluctuating current in the stator coil.*

Integrated circuit chips are available taht incorporate protection diodes, in addition to the necessary power transistors. Stepper motors may also have protection diodes built in. Consult the manufacturer's datasheet for details before attaching a motor to a power source.

Positional Control

The built-in control electronics of a **servo motor** typically turn the shaft to a precisely known position in response to pulse-width modulation from an exterior source such as a **microcontroller**, whereas the angle of rotation of a stepper motor in an open-loop system must be calculated by counting the number of steps from an initial, home position. This limitation of a stepper motor can be overcome by using a closed-loop system, but that will require monitoring the motor, adding complexity to the external controller. The choice between stepper and servo motors should be evaluated on a case-by-case basis.

What Can Go Wrong

General problems affecting all types of motors are listed in "Heat effects" (page 188). Issues relating more specifically to stepper motors are listed in the following sections.

Incorrect Wiring

Because a stepper motor is driven via multiple conductors, there is a significant risk of wiring errors, especially since many motors are not identified with part numbers. The first challenge, then, may be to determine what type of motor it is. When the motor is disconnected from any power, and the shaft is rotated with finger and thumb, a magnetized-rotor motor will not spin as freely as a reluctance motor, because the magnets in the rotor will provide intermittent turning resistance.

If a unipolar motor is relatively small and is fitted with five wires, almost certainly the motor contains two coils, each with a center tap, and their function can be determined by applying positive voltage to the red wire and grounding each of the other wires in turn. Attaching a small piece of tape to the motor shaft will assist in viewing its orientation.

A multimeter set to measure ohms can also be useful in deducing the internal coil connections of the motor, since the end-to-end resistance of a coil should be approximately twice the resistance between the center tap and either end of the coil.

A multiphase motor may have five wires, but in this case, the resistance between any two nonadjacent wires will be 1.5 times the resistance between any two adjacent wires.

Step Loss

In an open-loop system, if the motor skips or misses pulses from the controller, the controller no longer has an accurate assessment of the shaft angle. This is known as *step loss*. Since this can be caused by sudden changes in control frequency, the frequency should be increased (or decreased) gradually. This is known as *ramping* the motor speed. Stepper motors cannot respond instantly to changes in speed, because of inertia in the rotor or in the device that the motor is driving.

Where the motor turns one or more steps beyond its commanded stopping point, this is known as *overshoot*.

Step loss may also occur if the motor continues turning after power has been interrupted (either intentionally or because of an external fault). In an open-loop system, the controller should be designed to reset the motor position when power is initiated.

Excessive Torque

When the motor is stationary and not powered, *detent torque* is the maximum turning force that can be applied without causing the shaft to turn. When the motor is stationary and the controller does deliver power to it, *holding torque* is the maximum turning force that can be applied without causing the shaft to turn, and *pull-in torque* is the maximum torque which the motor can apply to overcome resistance and reach full speed. When the motor is running, *pull-out torque* is the maximum torque the motor can deliver without suffering step loss (pulling it out of sync with its controller). Some or all of these values should be specified on the motor's datasheet. Exceeding any of them will result in step loss.

Hysteresis

When a controller directs a stepper motor to seek a specified position, the term *hysteresis* is often used to mean the total error between the actual position it reaches when turning clockwise, and the actual position it reaches when turning counter-clockwise. This difference may occur because a stepper motor tends to stop a fraction short of its intended position, especially under significant load. Any design that requires precision should be tested under real-world conditions to assess the hysteresis of the motor.

Resonance

A motor has a natural resonant frequency. If it is stepped near that frequency, vibration will tend to be amplified, which can cause positional errors, gear wear (if gears are attached), bearing wear, noise, and other issues. A good datasheet should specify the resonant frequency of the motor, and the motor should run above that frequency if possible. The problem can be addressed by rubber motor mounts or by using a resilient component, such as a drive belt, in conjunction with the drive shaft. *Damping* the vibration may be attempted by adding weight to the motor mount.

Note that if the motor has any significant weight attached directly to its shaft, this will lower its resonant frequency, and should be taken into account.

Resonance may also cause *step loss* (see preceding sections).

Hunting

In a closed-loop system, a sensor on the motor reports its rotational position to the controller, and if necessary, the controller responds by adjusting the position of the motor. Like any feedback system, this entails some lag time, and at certain speeds the motor may start *hunting* or *oscillating* as the controller over-corrects and must then correct its correction. Some closed-loop controllers avoid this issue by running mostly in open-loop mode, using correction only when the motor experiences conditions (such as sudden speed changes), which are likely to cause step loss.

Saturation

While it may be tempting to increase the torque from a stepper motor by upping the voltage (which will increase the current through the stator coils), in practice motors are usually designed so that the cores of the coils will be close to saturation at the rated voltage. Therefore, increasing the voltage may achieve very little increase in power, while causing a significant increase in heat.

Rotor Demagnetization

The permanent magnets in a rotor can be partially demagnetized by excessive heat. Demagnetization can also occur if the magnets are exposed to high-frequency alternating current when the rotor is stationary. Therefore, attempting to run a stepper motor at high speed when the rotor is stalled can cause irrevocable loss of performance.

diode

26

The term **diode** almost always means a semiconductor device, properly known as a *PN junction diode*, although the full term is not often used. It was formerly known as a *crystal diode*. Before that, **diode** usually meant a type of *vacuum tube*, which is now rarely used outside of high-wattage RF transmitters and some high-end audio equipment.

OTHER RELATED COMPONENTS

- **rectifier** (See "Rectification" (page 227))
- **unijunction transistor** (See Chapter 27)
- **LED** (light-emitting diode) (Volume 2)

What It Does

A diode is a two-terminal device that allows current to flow in one direction, known as the *forward direction,* when the *anode* of the diode has a higher positive potential than the *cathode*. In this state, the diode is said to be *forward biased*. If the polarity of the voltage is reversed, the diode is now *reverse biased*, and it will attempt to block current flow, within its rated limits.

Diodes are often used as *rectifiers* to convert alternating current into direct current. They may also be used to suppress voltage spikes or protect components that would be vulnerable to reversed voltage, and they have specialized applications in high-frequency circuits.

A *Zener diode* can regulate voltage, a *varactor diode* can control a high-frequency oscillator, and *tunnel diodes*, *Gunn diodes*, and *PIN diodes* have high-frequency applications appropriate to their rapid switching capability. An **LED** (*light-emitting diode*) is a highly efficient light source, which is discussed in Volume 2 of this encyclopedia. A **photosensitive diode** will adjust its ability to pass current depending on the light that falls upon it, and is included as a *sensor* in Volume 3.

See Figure 26-1 for schematic symbols representing a generic diode.

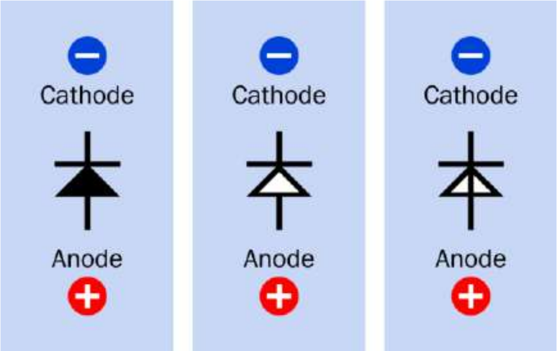

Figure 26-1. *Commonly used schematic symbols for a generic diode. All the symbols are functionally identical. The direction of the arrow formed by the triangle indicates the direction of conventional current (from positive to negative) when the diode is forward-biased.*

The basic diode symbol is modified in various ways to represent variants, as shown in Figure 26-2.

At top

Each symbol in the group of six indicates a Zener diode. All are functionally identical.

Bottom-left
Tunnel diode.

Bottom-center
Schottky diode.

Bottom-right
Varactor.

A triangle with an open center does not indicate any different function from a triangle with a solid center. The direction of the arrow always indicates the direction of conventional current, from positive to negative, when the diode is forward-biased, although the functionality of Zener diodes and varactors depends on them being reverse-biased, and thus they are used with current flowing opposite to the arrow symbol. The bent line used in the Zener symbol can be thought of as an opened letter Z, while the curled line used in the Schottky diode symbol can be thought of as a letter S, although these lines are sometimes drawn flipped left-to-right.

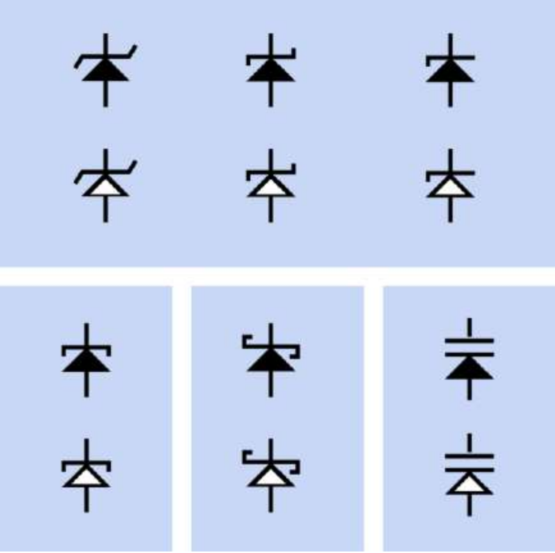

Figure 26-2. *Commonly used schematic symbols for specialized types of diodes. See text for details.*

A range of rectifier and signal diodes is shown in Figure 26-3. (Top: Rectifier diode rated 7.5A at 35VDC. Second from top: Rectifier diode rated 5A at 35VDC. Center: Rectifier diode rated 3A at 35VDC. Second from bottom: 1N4001 Rectifier diode rated 1A at 35VDC. Bottom: 1N4148 signal switching diode rated at 300mA.) All values are for forward continuous current and RMS voltage. Each cylindrical diode is marked with a silver stripe (a black stripe on the 1N4148) to identify its *cathode*, or the end of the diode that should be "more negative" when the component is forward biased. Peak current can greatly exceed continuous current without damaging the component. Datasheets will provide additional information.

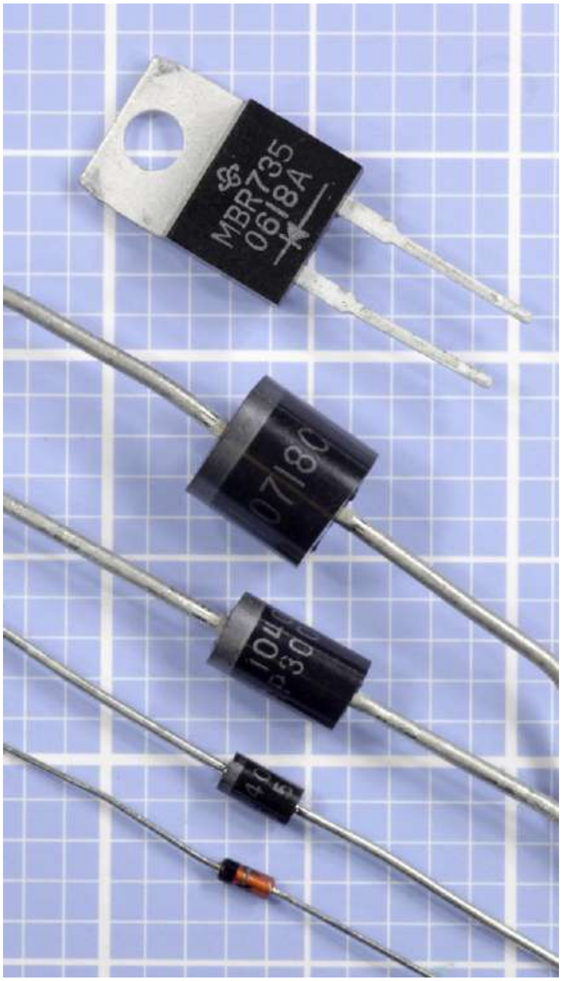

Figure 26-3. *Diodes ranging in continuous forward-current capability from 7.5A (top) to 300mA (bottom). See text for additional details.*

How It Works

A PN diode is a two-layer *semiconductor*, usually fabricated from silicon, sometimes from germanium, and rarely from other materials. The layers are doped with impurities to adjust their electrical characteristics (this concept is explained in more detail in Chapter 28). The *N layer* (on the negative, cathode side) has a surplus of electrons, creating a net negative charge. The *P layer* (on the positive, anode side) has a deficit of electrons, creating a net positive charge. The deficit of electrons can also be thought of as a surplus of "positive charges," or more accurately, a surplus of *electron holes*, which can be considered as spaces that electrons can fill.

When the negative side of an external voltage source is connected with the cathode of a diode, and the positive side is connected with the anode, the diode is forward-biased, and electrons and electron holes are forced by mutual repulsion toward the junction between the n and p layers (see Figure 26-4). In a silicon diode, if the potential difference is greater than approximately 0.6 volts, this is known as the *junction threshold voltage*, and the charges start to pass through the junction. The threshold is only about 0.2 volts in a germanium diode, while in a Schottky diode it is about 0.4 volts.

If the negative side of an external voltage source is connected with the anode of a diode and positive side is connected with the cathode, the diode is now reverse-biased, and electrons and electron holes are attracted away from the junction between the n and p layers. The junction is now a *depletion region*, which blocks current.

Like any electronic component, a diode is not 100% efficient. When it is forward-biased and is passing current, it imposes a small voltage drop of around 0.7V for a silicon-based diode (Schottky diodes can impose a drop of as little as 0.2V, germanium diodes 0.3V, and some LEDs between 1.4V and 4V). This energy is dissipated as heat. When the diode is reverse-biased, it is still not 100% efficient, this time in its task of blocking current. The very small amount of current that manages to get through is known as *leakage*. This is almost always less than 1mA and may be just a few µA, depending on the type of diode.

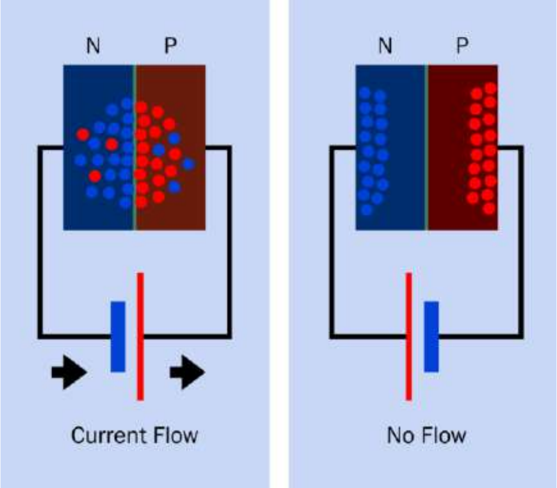

Figure 26-4. *Inside a PN junction diode. Left: in forward-biased mode, voltage from a battery (bottom, with plates colored for clarity) forces charges in the N and P layers toward the central junction of the diode. Current begins to flow. Right: in reverse-biased mode, charges in the N and P layers are attracted away from the central junction, which becomes a depletion region, unable to pass significant current.*

The performance of a theoretical generic PN diode is illustrated in Figure 26-5. The right-hand side of the graph shows that if a diode is forward-biased with a gradually increasing potential, no current passes until the diode reaches its junction threshold voltage, after which the current rises very steeply, as the *dynamic resistance* of the diode diminishes to near zero. The left-hand side of the graph shows that when the diode is reverse-biased with a gradually increasing potential, initially a very small amount of current passes as leakage (the graph exaggerates this for clarity). Eventually, if the potential is high enough, the diode reaches its intrinsic *breakdown voltage*, and once again its effective resistance diminishes to near zero. At either end of the curve, the diode will be easily and permanently damaged by excessive current. With the exception of Zener diodes and varactors, reverse bias on a diode should not be allowed to reach the breakdown voltage level.

The graph in Figure 26-5 does not have a consistent scale on its Y axis, and in many diodes the magnitude of the (reverse-biased) breakdown voltage will be as much as 100 times the magnitude of the (forward-biased) threshold voltage. The graph has been simplified for clarity.

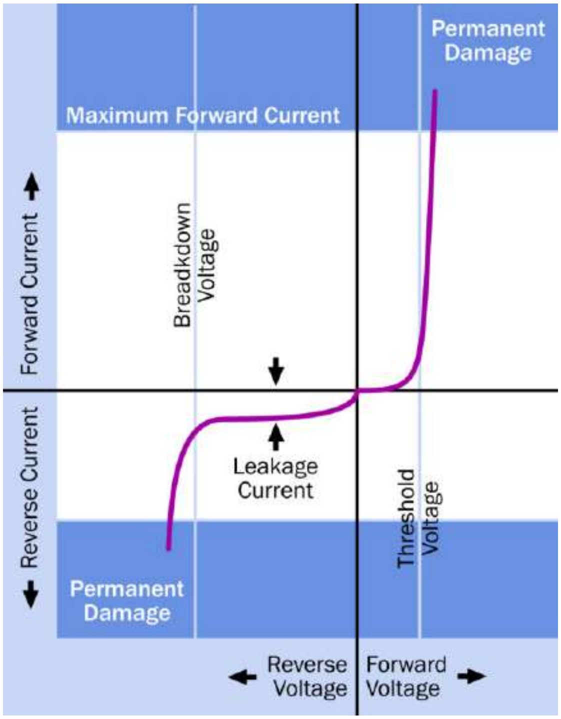

Figure 26-5. *As the forward voltage across a diode reaches the junction threshold, the diode begins passing current. If the voltage across the diode is reversed, initially a small amount of current leakeage occurs. Excessive forward or reverse voltage will create sufficient current to destroy the component.*

Variants

Packaging

Some diodes have no information at all printed on them, while others may have a part number. Any additional information is rare. No convention exists for indicating the electrical characteristics of the component by colors or abbreviations. If one terminal is marked in any way, almost certainly it is the cathode. One way to remember the meaning of a stripe on the cathode end of a rectifier diode or signal diode is by thinking of it as resembling the line in the diode schematic symbol.

Signal Diodes

Also known as *switching diodes* and *high-speed diodes*, their small size provides a low junction capacitance, enabling fast response times. They are not designed to withstand high currents. Signal diodes traditionally were packaged with axial leads for through-hole installation (like traditional-style resistors). Although this format still exists, signal diodes are now more commonly available in surface-mount formats.

Rectifier Diodes

Physically larger than signal diodes, and capable of handling higher currents. Their higher junction capacitance makes them unsuitable for fast switching. Rectifier diodes often have axial leads, although different package formats are used where higher currents are involved, and may include a heat sink, or may have provision for being attached to a heat sink.

There are no generally agreed maximum or minimum ratings to distinguish signal diodes from rectifier diodes.

Zener Diode

A Zener diode generally behaves very similarly to a signal or rectifier diode, except that its breakdown voltage is lower.

The Zener is intended to be reverse-biased; that is, conventional current is applied through it "in the wrong direction" compared with conventional diodes. As the current increases, the *dynamic resistance* of the Zener diode decreases. This relationship is shown in Figure 26-6, where the two colored curves represent the performance of different possible Zener diodes. (The curves are adapted from a manufacturer's datasheet.) This behavior allows the Zener to be used in simple voltage-regulator circuits, as it can allow a reverse current to flow at a voltage limited by the diode's breakdown voltage. Other applications for Zener diodes are described in "DC Voltage Regulation and Noise Suppression" (page 230). A typical Zener diode is shown in Figure 26-7.

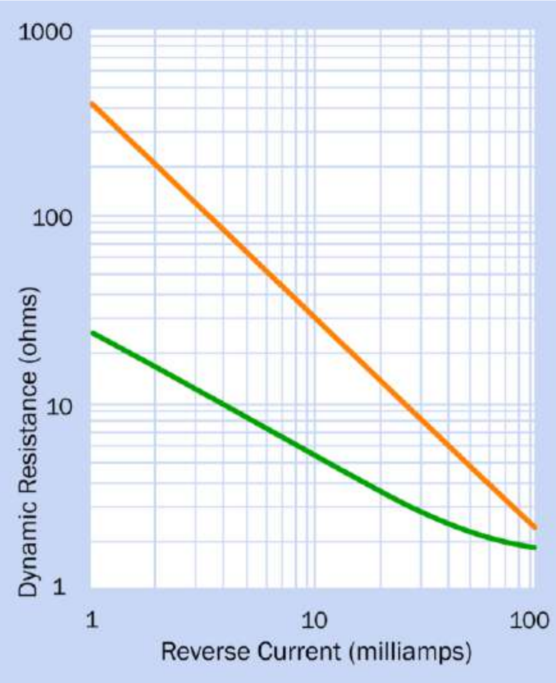

Figure 26-6. *A manufacturer's datasheet may include graphs of this kind, showing the variation in dynamic resistance of two reverse-biased Zener diodes in response to changes in current.*

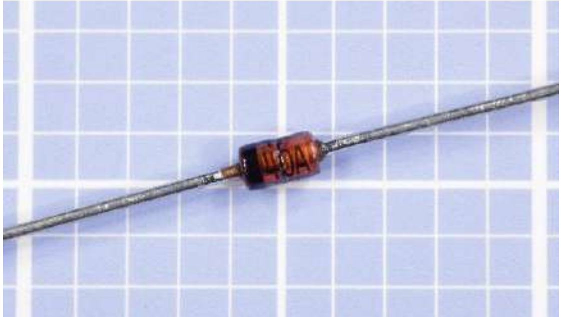

Figure 26-7. *A 1N4740 Zener diode.*

Transient Voltage Suppressor (TVS)

A form of Zener diode designed to protect sensitive devices from transient voltage spikes by clamping them—in other words, diverting the energy to ground. A TVS can absorb as much as 30,000 volts from a lightning strike or static discharge. Typically the Zener diode is incorporated in a network of other diodes in a surface-mount integrated circuit chip.

Zener diodes can also be used in circuits to handle electrostatic discharge (ESD), which can occur when a person unknowingly accumulates an electrostatic potential and then grounds it by touching an electronic device.

Schottky Diode

This type has a low junction capacitance, enabling faster switching than comparable generic silicon diodes. It also imposes a lower forward voltage drop, which can be desirable in low-voltage applications, and allows less power dissipation when a diode is necessary to control current flow. The Schottky diode is fabricated with a semiconductor-to-metal junction, and tends to be slightly more expensive than generic silicon diodes with similar voltage and current specifications.

Varactor Diode

Also known as a *varicap*, this type of diode has variable capacitance controlled by reverse voltage. While other diodes may exhibit this same phenomenon, the varactor is specifically designed to exploit it at very high frequencies. The voltage expands or contracts the depletion region in the junction between the P and N regions, which can be thought of as analogous to moving the plates of a capacitor nearer together or farther apart.

Because the capacitance of a varactor has a low maximum of about 100pF, its uses are limited. It is used extensively in RF applications where its voltage-controlled variable capacitance provides a uniquely useful way to control the frequency of an oscillator circuit. In almost all radio, cellular, and wireless receivers, a varactor controls a phase-locked loop oscillator. In ham radio receivers, it can be used to adjust the tuning of a filter that tracks an incoming radio frequency.

A varactor is always reverse-biased below its breakdown voltage, so that there is no direct conduction. The voltage that controls a varactor must be absolutely free from random fluctuations that would affect its resonant frequency.

Tunnel Diode, Gunn Diode, PIN Diode

Mostly used in very high frequency or microwave applications, where ordinary diodes are unacceptable because they have insufficiently high switching speeds.

Diode Array

Two or more diodes may be encapsulated in a single DIP or (more commonly) surface-mount integrated circuit chip. The internal configuration and the pinouts of the chip will vary from one device to another. Diode arrays may be used for termination of data lines to reduce reflection noise.

Bridge Rectifier

Although this is a diode array, it is commonly indexed in parts catalogues under the term *bridge rectifier*. Numerous through-hole versions are available with ratings as high as 25A, some designed for single-phase input while others process three-phase AC. Screw-terminal components can rectify more than 1,000 volts at 1,000 amps. The package does not usually include any provision for smoothing or filtering the output. See “Rectification” (page 227) for more information on the behavior of a bridge rectifier.

Values

A manufacturer’s datasheet for a typical generic diode should define the following values, using abbreviations that may include those in the following list.

- Maximum sustained forward current: I_f or I_o or I_{Omax}
- Forward voltage (the voltage drop imposed by the diode): V_f
- Peak inverse DC voltage (may be referred to as maximum blocking voltage or breakdown voltage): P_{iv} or V_{dc} or V_{br}
- Maximum reverse current (also referred to as leakage): I_r

Datasheets may include additional values when the diode is used with alternating current, and will also include information on peak forward surge current and acceptable operating temperatures.

A typical signal diode is the 1N4148 (included at the bottom of Figure 26-3), which is limited to about 300mA forward current while imposing a voltage drop of about 1V. The component can tolerate a 75V peak inverse voltage. These values may vary slightly among different manufacturers.

Rectifier diodes in the 1N4001/1N4002/1N4003 series have a maximum forward current of 1A and will impose a voltage drop of slightly more than 1V. They can withstand 50V to 1,000V of inverse voltage, depending on the component. Here again, the values may vary slightly among different manufacturers.

Zener diodes have a different specification, as they are used with reverse bias as voltage-regulating devices rather than rectification devices. Manufacturers’ data sheets are likely to contain the following terminology:

- Zener voltage (the potential at which the diode begins to allow reverse current flow when it is reverse-biased, similar to breakdown voltage): V_z
- Zener impedance or dynamic resistance (the effective resistance of the diode, specified when it is reverse-biased at the Zener voltage): Z_z

- Maximum or admissible Zener current (or reverse current): I_z or I_{zm}
- Maximum or total power dissipation: P_d or P_{tot}

Zener voltage may be defined within a minimum and maximum range, or as a simple maximum value.

Limits on forward current are often not specified, as the component is not intended to be forward-biased.

How to Use it

Rectification

A *rectifier diode*, as its name implies, is commonly used to rectify alternating current—that is, to turn AC into DC. A half-wave rectifier uses a single diode to block one-half of the AC sinewave. The basic circuit for a half-wave rectifier is shown in Figure 26-8. At top, the diode allows current to circulate counter-clockwise through the load. At bottom, the diode blocks current that attempts to circulate clockwise. Although the output has "gaps" between the pulses, it is usable for simple tasks such as lighting an LED, and with the addition of a smoothing capacitor, can power the coil of a DC relay.

A full-wave bridge rectifier employs four diodes to provide a more efficient output, usually filtered and smoothed with appropriate **capacitors**. The basic circuit is shown in Figure 26-9. A comparison of input and output waveforms for half-wave and full-wave rectifiers appears in Figure 26-10.

Discrete components are seldom used for this purpose, as off-the-shelf bridge rectifiers are available in a single integrated package. Rectifier diodes as discrete components are more likely to be used to suppress *back-EMF* pulses, as described below.

An old but widely used design for a full-wave bridge rectifier is shown in Figure 26-11. This unit measured approximately 2″ × 2″ × 1.5″ and was divided into four sections (as indicated by the solder terminals on the right-hand side), each

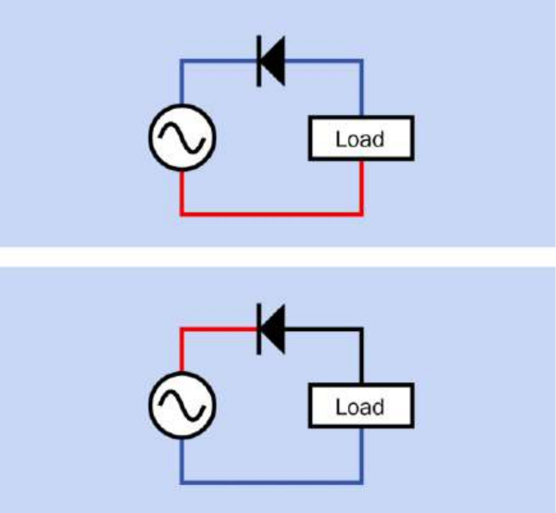

Figure 26-8. *A half-wave rectifier. In this configuration the diode allows AC current to circulate counter-clockwise but blocks it clockwise.*

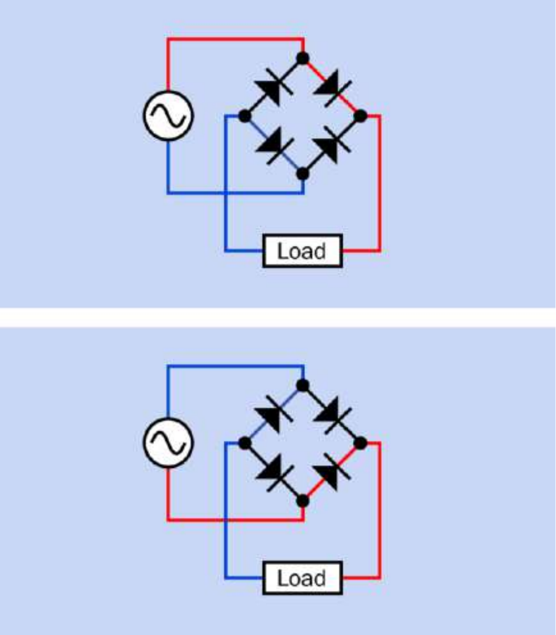

Figure 26-9. *The basic circuit commonly used to form a bridge rectifier, with color added to indicate polarity. Wires shown in black are not passing current because diodes are blocking it. Note that the polarity at the load remains constant.*

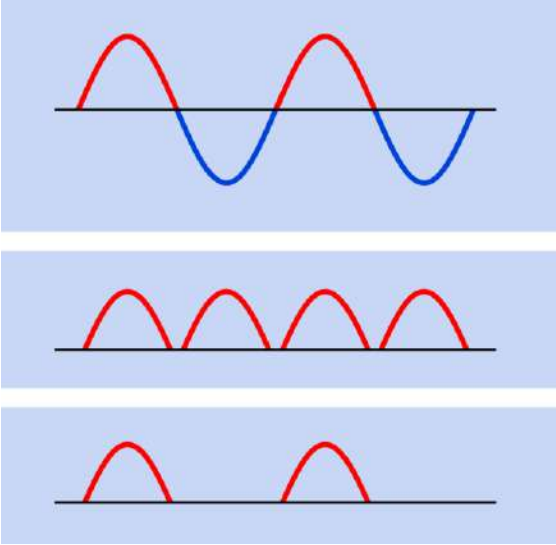

Figure 26-10. *Top: The voltage-amplitude sinewave of an alternating current source that fluctuates between positive voltage (shown red) and negative voltage (shown blue) relative to a neutral (black) baseline. Center: AC current converted by a full-wave rectifier. Because the diodes do not conduct below their threshold voltage, small gaps appear between pulses. Bottom: Output from a half-wave rectifier.*

section corresponding with the functionality of one modern diode. Figure 26-12 shows relatively modern rectifier packages, the one on the left rated at 20A continuous at 800V RMS, the one on the right rated 4A continuous at 200V RMS. In Figure 26-13, the one on the left is rated 4A continuous at 50V RMS, whereas the one on the right is rated 1.5A at 200V RMS.

DC output from rectifier packages is usually supplied via the outermost pins, while the two pins near the center receive AC current. The positive DC pin may be longer than the other three, and is usually marked with a + symbol.

Full-wave bridge rectifiers are also available in surface-mount format. The one in Figure 26-14 is rated for half an amp continuous current.

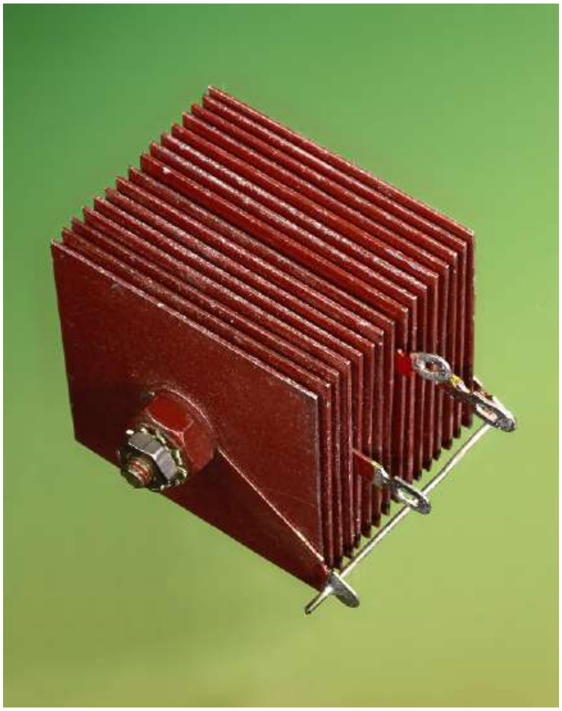

Figure 26-11. *Prior to the perfection of chip fabrication in the late 1960s, it was common to find silicon rectifiers of this type, measuring about 2" square.*

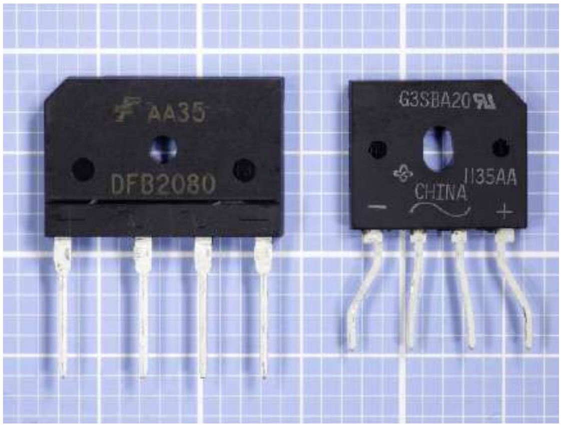

Figure 26-12. *Full-wave bridge rectifies are commonly available in packages such as these. See text for details.*

Back-EMF Suppression

A relay coil, motor, or other device with significant inductance will typically create a spike of voltage when it is turned on or off. This *EMF* can be shunted through a rectifier diode to safeguard other components in the circuit. A diode in this configuration may be referred to as a *protection diode*, a *clamp diode*, or *transient suppressor*. See Figure 26-15.

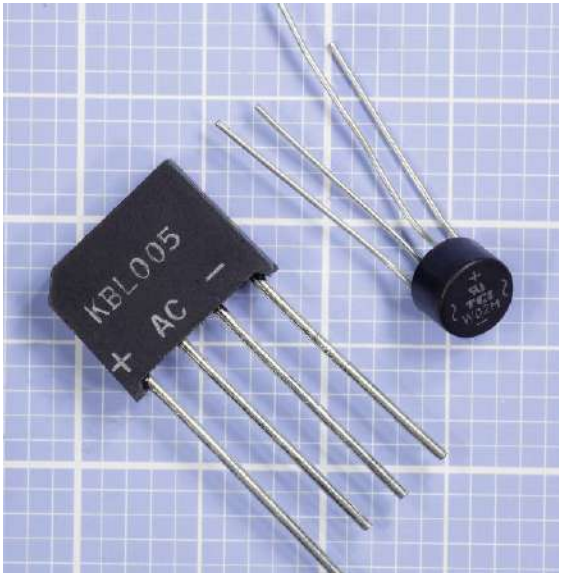

Figure 26-13. *Smaller full-wave bridge rectifiers capable of 1.5A to 4A continuous current.*

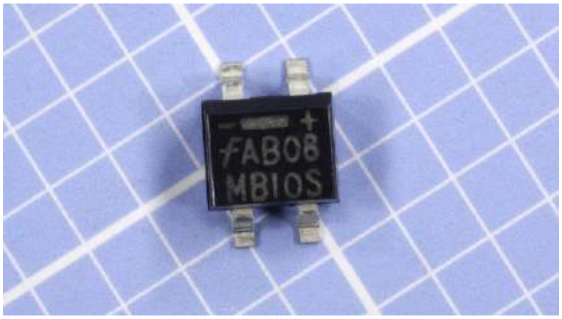

Figure 26-14. *This surface-mount component contains four diodes forming a full-wave bridge rectifier circuit, and can pass 0.5A continuous current. It measures approximately 0.2" square.*

Voltage Selection

A diode is sensitive to the relative voltage between its anode and cathode terminals. In other words, if the cathode is at 9V relative to the ground in the circuit, and the anode is at 12V, the 3V difference will easily exceed the threshold voltage, and the diode will pass current. (Actual tolerable values will depend on the forward voltage capability of the diode.) If the voltages are reversed, the diode will block the current.

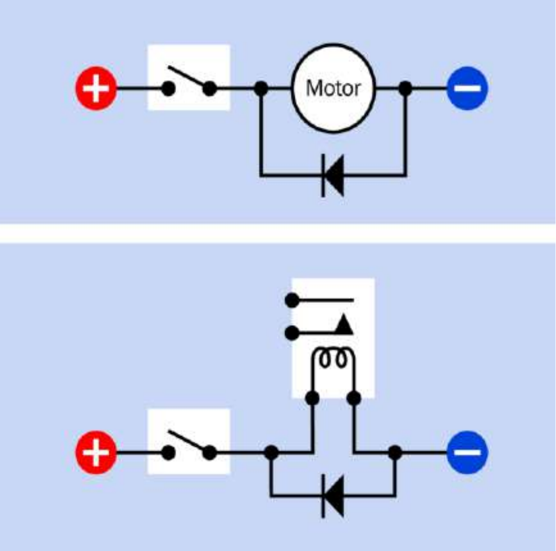

Figure 26-15. *A rectifier diode is very often placed across a motor (top), relay (bottom), or other device with significant inductance that creates a spike of reverse voltage when switched on or off. The surge is shunted through the diode, protecting other components in the circuit.*

This attribute can be used to make a device choose automatically between an AC adapter and a 9V battery. The schematic is shown in Figure 26-16. When an AC adapter that delivers 12VDC is plugged into a wall outlet, the adapter competes with the battery to provide power to a voltage regulator. The battery delivers 9VDC through the lower diode to the cathode side of the upper diode, but the AC adapter trumps it with 12VDC through the upper diode. Consequently, the battery ceases to power the circuit until the AC adapter is unplugged, at which point the battery takes over, and the upper diode now prevents the battery from trying to pass any current back through the AC adapter.

The voltage regulator in this schematic accepts either 12VDC or 9VDC and converts it to 5VDC. (In the case of 12VDC, the regulator will waste more power, which will be dissipated as waste heat.)

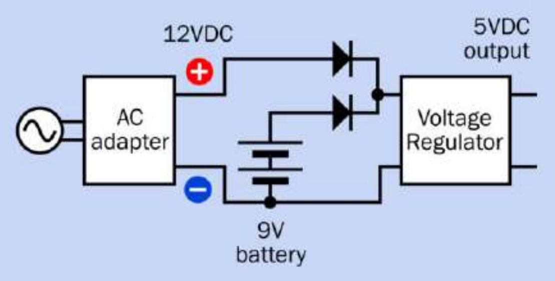

Figure 26-16. *Two diodes with their cathodes tied together will choose automatically between an AC adapter that delivers 12VDC and an internal 9V battery.*

Voltage Clamping

A diode can be used to *clamp* a voltage to a desired value. If an input to a 5V CMOS semiconductor or similarly sensitive device must be prevented from rising out of range, the anode of a diode can be connected to the input and the cathode to a 5V voltage source. If the input rises much above 5.6V, the potential difference exceeds the diode's junction threshold, and the diode diverts the excess energy. See Figure 26-17.

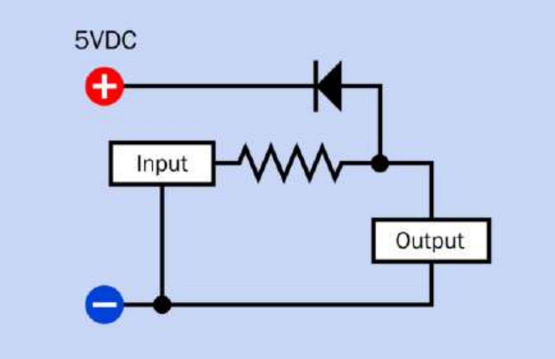

Figure 26-17. *A clamping diode can limit output voltage—in this example, to about 5.6V. If the input rises above that value relative to the common ground, the potential difference across the diode feeds the excess voltage back through it to the 5V source.*

Logic Gate

A signal diode is less than ideal as a logic gate, because it imposes a typical 0.6V voltage reduction, which can be significant in a 5V circuit and is probably unacceptable in a 3.3V circuit. Still, it can be useful on the output side—for example, if two or more outputs from a logic chip or microcontroller are intended to drive, or share, another device such as a single **LED**, as shown in Figure 26-18. In this role, the diodes wired in parallel behave similarly to an OR gate, while preventing either output from the chip from feeding current back into the other output.

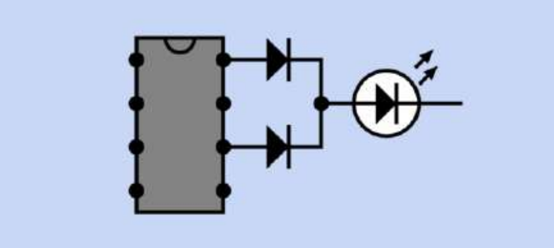

Figure 26-18. *Two or more outputs from a logic chip or microcontroller may be coupled with diodes to power another device, such as an LED, while protecting the chip from backflow of current. The diodes form a logical OR gate.*

DC Voltage Regulation and Noise Suppression

As previously noted, the dynamic resistance of a reverse-biased Zener diode will diminish as the current increases. This relationship begins at the point where breakdown in the diode begins—at its *Zener voltage*--and is approximately linear over a limited range.

The unique behavior of the Zener makes it usable as a very simple voltage controller when placed in series with a resistor as shown in Figure 26-19. It is helpful to imagine the diode and the resistor as forming a kind of voltage divider, with power being taken out at point A in the schematic. If a supply fluctuation increases the input voltage, this will tend to increase the current flowing through the Zener, and its dynamic resistance will diminish accordingly. A lower resistance in its position in the voltage divider will reduce the output voltage at point A, thus tending to compensate for the surge in input voltage.

Conversely, if the load in the circuit increases, and tends to pull down the input voltage, the current

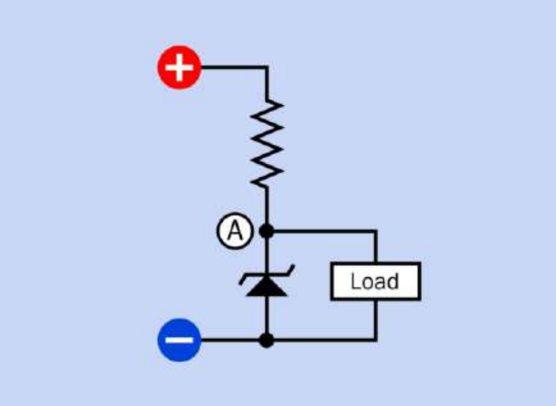

Figure 26-19. *A simplified, basic circuit illustrating the ability of a Zener diode to compensate for variations in the power supply or load in a circuit, creating an approximately constant voltage at point A.*

flowing through the Zener will diminish, and the voltage at point A will tend to increase, once again compensating for the fluctuation in the circuit.

As the series resistor would be a source of heat, a transistor could be added to drive the load, as shown in Figure 26-20.

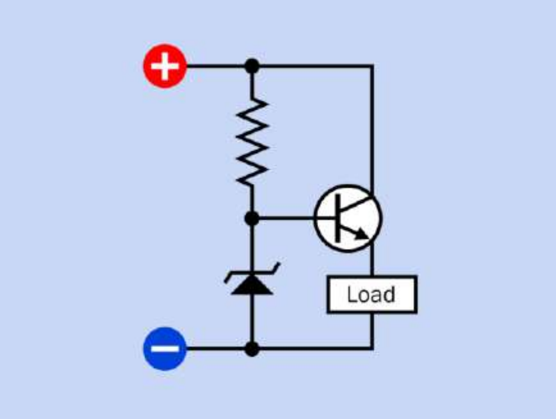

Figure 26-20. *A transistor could be added to the circuit in the previous figure to reduce power waste through the resistor.*

A manufacturer's datasheet may provide guidance regarding the dynamic resistance of a Zener diode in response to current, as previously shown in Figure 26-6. In practice, a packaged **voltage regulator** such as the LM7805 would most likely be used instead of discrete components, since it includes self-calibrating features, requires no series resistor, and is relatively unaffected by temperature. However, the LM7805 contains its own Zener diode, and the principle of operation is still the same.

AC Voltage Control and Signal Clipping

A more practical Zener application would be to limit AC voltage and/or impose clipping on an AC sinewave, using two diodes wired in series with opposed polarities. The basic schematic is shown in Figure 26-21, while clipping of the AC sinewave is illustrated in Figure 26-22. In this application, when one diode is reverse-biased, the other is forward-biased. A forward-biased Zener diode works like any other diode: it allows current to pass relatively freely, so long as the voltage exceeds its threshold. When the AC current reverses, the Zeners trade their functions, so that the first one merely passes current while the second one limits the voltage. Thus, the diodes divert peak voltage away from the load. The Zener voltage of each diode would be chosen to be a small margin above the AC voltage for voltage control, and below the AC voltage for signal clipping.

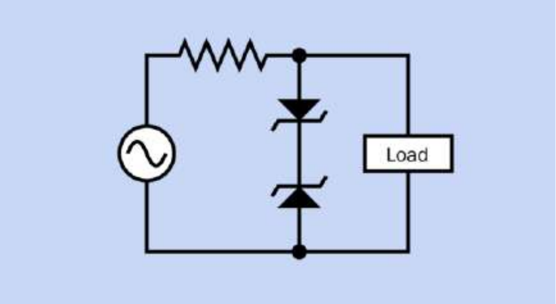

Figure 26-21. *Two Zener diodes placed in series, with opposite polarities, can clip or limit the voltage sinewave of an AC signal.*

Voltage Sensing

A Zener diode can be used to sense a small shift in voltage and provide a switched output in response.

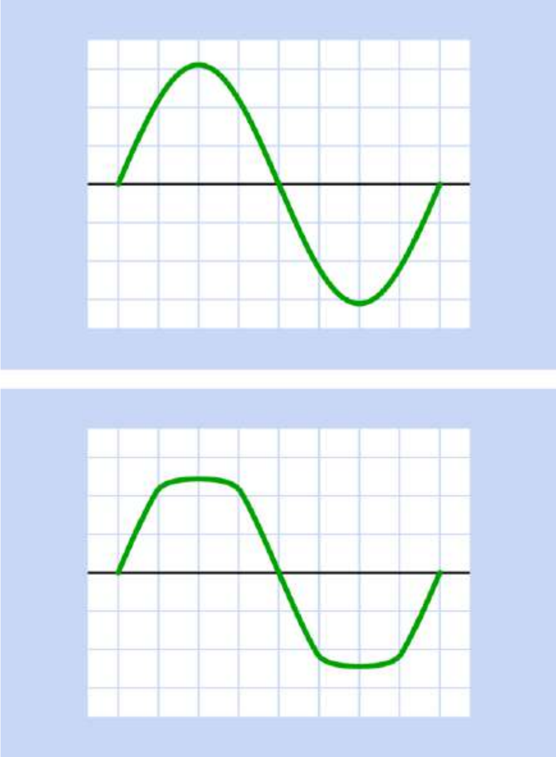

Figure 26-22. *AC input showing a pure sinewave (left) and a clipped version (right) created by Zener diodes wired in series, as in the previous figure.*

In Figure 26-23, the upper schematic shows a Zener diode preventing voltage from reaching the emitter of a PNP transistor while the divided input signal is below the Zener (breakdown) voltage of the diode. In this mode, the transistor is relatively non-conductive, very little current flows through it, and the output is now at near-zero voltage. As soon as the input signal rises above the Zener voltage, the transistor switches on and power is supplied to the output. The input is thus replicated in the output, as shown in the upper portion of Figure 26-24.

In Figure 26-23, the lower schematic shows a Zener diode preventing voltage from reaching the base of an NPN transistor while the input signal is below the Zener (breakdown) voltage of the diode. In this mode, the transistor is relatively non-conductive, and power is supplied to the output. As soon as the input signal rises above the Zener voltage, the transistor is activated, diverting the current to ground and bypassing the output, which is now at near-zero voltage. The input is thus inverted, as shown in the lower portion of Figure 26-24 (provided there is enough current to drive the transistor into saturation).

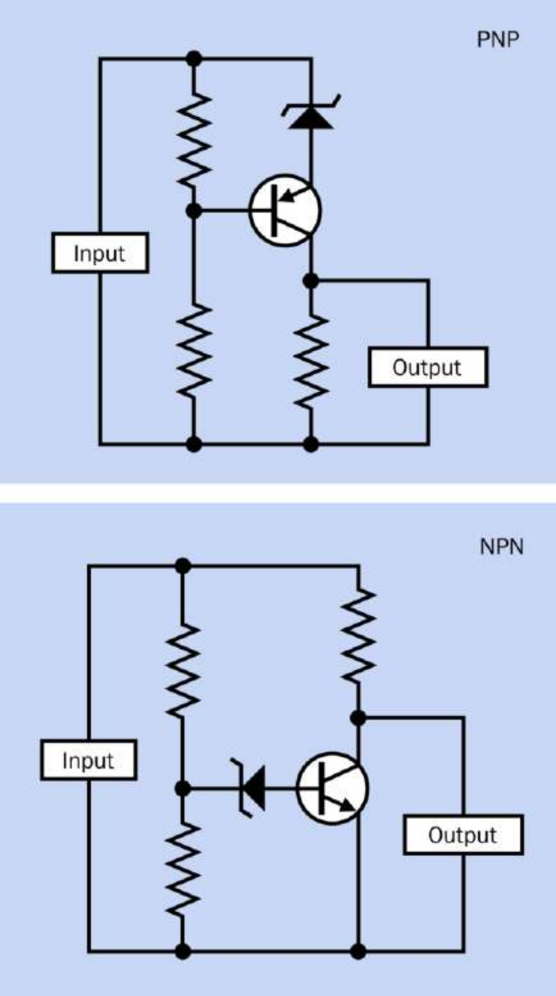

Figure 26-23. *A Zener diode can be used in conjunction with a PNP transistor. See text for details.*

What Can Go Wrong

Overload

If maximum forward current is exceeded, the heat generated is likely to destroy the diode. If the diode is reverse-biased beyond its peak in-

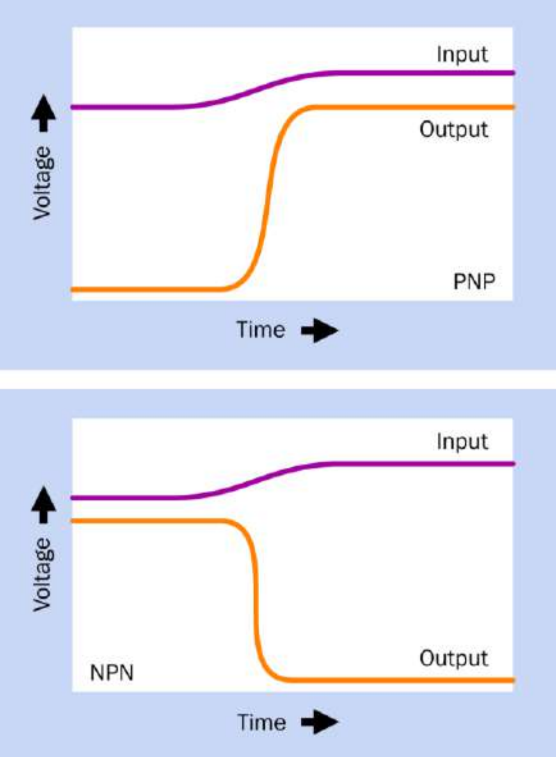

Figure 26-24. *Theoretical output from the transistors in the two previous schematics.*

verse voltage limit, the current will overwhelm the diode's ability to block it, and an *avalanche breakdown* will occur, once again probably destroying the component. The graph in Figure 26-5 illustrates the performance range of a hypothetical generic diode.

Reversed Polarity

Zener diodes look almost identical to other types, and all diodes share the same convention of marking the cathode for identification. Yet Zeners must be reverse-biased while others are forward-biased. This creates a significant risk of installing a diode "the wrong way around," with potentially destructive or at least confusing results, especially when used in a power supply. The very low resistance of a diode to forward current makes it especially vulnerable to burnout if installed incorrectly.

Wrong Type of Diode

If a Zener diode is used accidentally where a signal or rectifier diode is appropriate, the circuit will malfunction, as the Zener will probably have a much lower breakdown voltage, and therefore will not block reverse current. Conversely, if a signal or rectifier diode is used where the circuit calls for a Zener diode, reverse voltage will be clamped (or regulated at the diode's forward voltage value). Since diodes are often poorly marked, a sensible precaution is to store Zener diodes separately from all other types.

unijunction transistor

27

The unijunction transistor (UJT) and programmable unijunction transistor (PUT) are different internally, but are sufficiently similar in function to be combined in this entry.

OTHER RELATED COMPONENTS

- **diode** (See Chapter 26)
- **bipolar transistor** (See Chapter 28)
- **field-effect transistor** (See Chapter 29)

What It Does

Despite their names, the unijunction transistor (UJT) and programmable unijunction transistor (PUT) are not current-amplification devices like bipolar transistors. They are switching components that are more similar to diodes than to transistors.

The UJT can be used to build low- to mid-frequency oscillator circuits, while the PUT provides similar capability with the addition of more sophisticated control, and is capable of functioning at lower currents. The UJT declined in popularity during the 1980s after introduction of components such as the *555 timer*, which offered more flexibility and a more stable output frequency, eventually at a competitive price. UJTs are now uncommon, but PUTs are still available in quantity as through-hole discrete components. Whereas an integrated circuit such as a 555 timer generates a square wave, unijunction transistors in oscillator circuits generate a series of voltage spikes.

The PUT is often used to trigger a **thyristor** (described in Volume 2) and has applications in low-power circuits, where it can draw as little as a few microamps.

Schematic symbols for the two components are shown in Figure 27-1 and Figure 27-2. Although the symbol for the UJT is very similar to the symbol for a field-effect transistor (FET), its behavior is quite different. The bent arrow identifies the UJT, while a straight arrow identifies the FET. This difference is of significant importance.

The schematic symbol for a PUT indicates its function, as it resembles a diode with the addition of a gate connection.

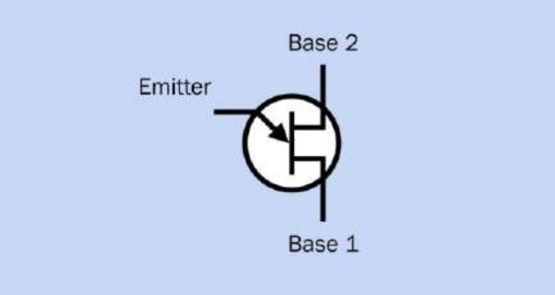

Figure 27-1. *Schematic symbol for a unijunction transistor (UJT). Note the bent arrow. The symbol for a field-effect transistor looks similar, but has a straight arrow. The functionality of the two components is very different.*

In Figure 27-3, the transistors at left and center are old-original unijunction transistors, while the one at right is a programmable unijunction transistor. (Left: Maximum 300mW, 35V interbase voltage. Center: 450mW, 35V interbase voltage. Right: 300mW, 40V gate-cathode forward voltage, 40V anode-cathode voltage.)

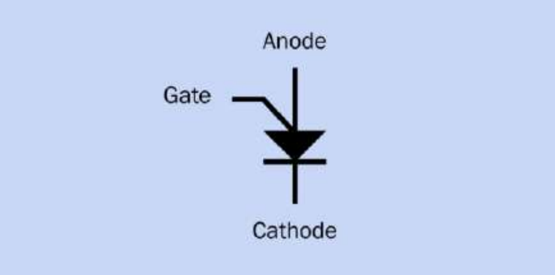

Figure 27-2. *Schematic symbol for a programmable unijunction transistor (PUT). The symbol accurately suggests the similarity in function to a diode, with the addition of a gate to adjust the threshold voltage.*

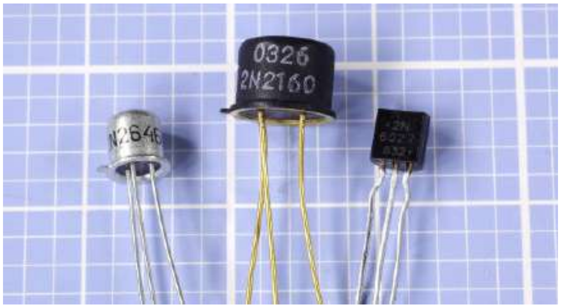

Figure 27-3. *The unijunction transistors at left and center are becoming obsolete; the one at the right is a programmable unijunction transistor (PUT), still readily available and widely used as a thyristor trigger.*

How It Works

The UJT is a three-terminal semiconductor device, but contains only two sections sharing a single junction—hence its name. Leads attached to opposite ends of a single channel of N-type semiconductor are referred to as base 1 and base 2, with base 2 requiring a slightly higher potential than base 1. A smaller P-type insert, midway between base 1 and base 2, is known as the emitter.

The diagram in Figure 27-4 gives an approximate idea of internal function.

When no voltage is applied to the emitter, a relatively high resistance (usually more than 5K) prevents significant current flow from base 2 to base 1. When the positive potential at the emitter increases to a *triggering voltage* (similar to the *junction threshold voltage* of a forward-biased diode), the internal resistance of the UJT drops very rapidly, allowing current to enter the component via both the emitter and base 2, exiting at base 1. (The term "current" refers, here, to conventional current; electron flow is opposite.) Current flowing from base 2 to base 1 is significantly greater than current flowing from the emitter to base 1.

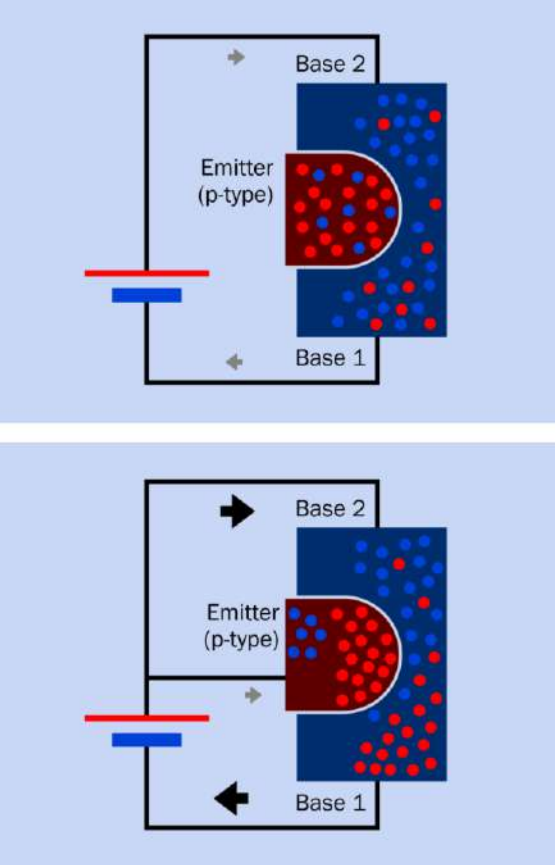

Figure 27-4. *Internal workings of a unijunction transistor.*

The graph in Figure 27-5 outlines the behavior of a UJT. As the voltage applied to the emitter increases, current flowing into the component from the emitter increases slightly, until the triggering voltage is reached. The component's internal resistance now drops rapidly. This pulls down the voltage at the emitter, while the current continues to increase significantly. Because of the drop in resistance, this is referred to as a *negative resistance* region. The resistance actually cannot fall below zero, but its change is negative. After emitter voltage drops to a minimum known as the *valley voltage*, the current continues to increase with a small increase in voltage. On datasheets, the peak current is often referred to as I_p while valley current is I_v.

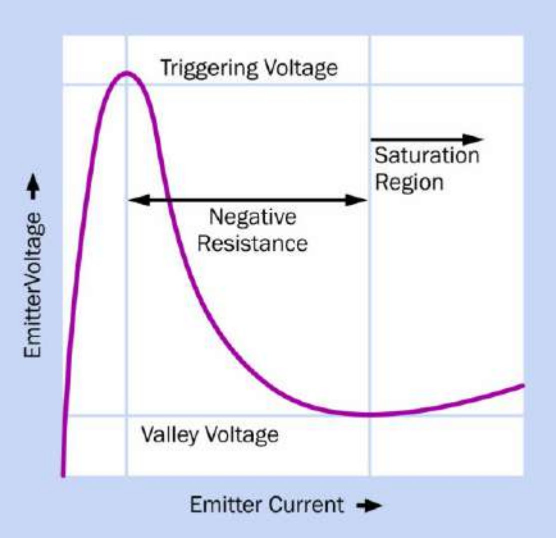

Figure 27-5. *Response curve of a unijunction transistor (UJT). When positive potential at the emitter reaches the triggering voltage, internal resistance drops radically and the component goes through a phase known as "negative resistance" as current increases.*

Figure 27-6 shows a test circuit to demonstrate the function of a UJT, with a volt meter indicating its status. A typical supply voltage would range from 9VDC to 20VDC.

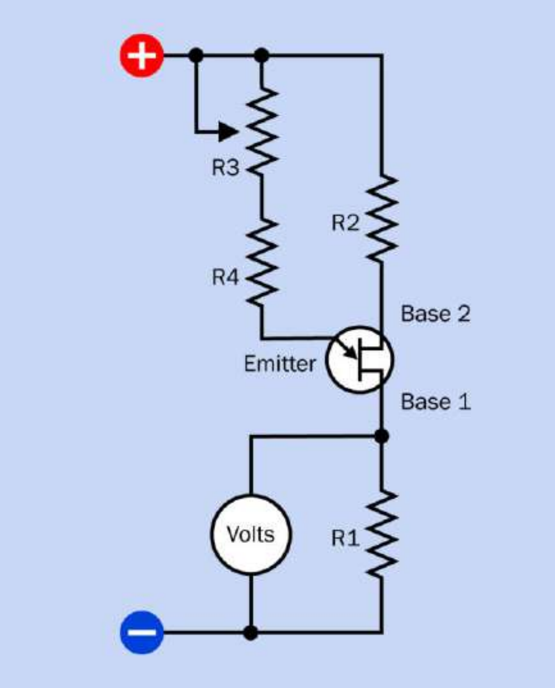

Figure 27-6. *A test circuit for a unijunction transistor (UJT) using a volt meter to show its response as a potentiometer increases the voltage applied to its emitter.*

A PUT behaves similarly in many ways to a UJT but is internally quite different, consisting of four semiconducting layers and functioning similarly to a thyristor.

The PUT is triggered by increasing the voltage on the anode. Figure 27-7 shows a test circuit for a PUT. This component is triggered when the voltage at its anode exceeds a threshold level, while the gate sets the threshold where this occurs. When the PUT is triggered, its internal resistance drops, and current flows from anode to cathode, with a smaller amount of current entering through the gate. This behavior is almost identical to that of a forward-biased diode, except that the threshold level can be controlled, or "programmed," according to the value of the positive potential applied at the gate, with R1 and R2 establishing that potential by functioning as a voltage divider.

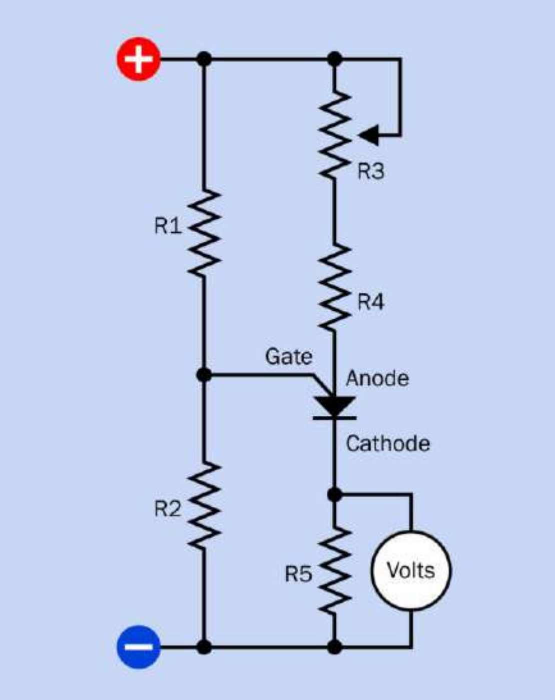

Figure 27-7. *A test circuit for a programmable unijunction transistor (PUT) using a volt meter to show its response as a potentiometer increases the voltage applied to its anode.*

The voltage output of a PUT follows a curve that is very similar to that shown in Figure 27-5, although current and voltage would be measured at the cathode.

Variants

PUTs and UJTs are not made as surface-mount components.

UJTs are usually packaged in black plastic, although older variants were manufactured in cans. PUTs are almost all packaged in black plastic. With the leads pointing downward and the flat side facing toward the viewer, the lead functions of a PUT are usually anode, gate, and cathode, reading from left to right.

Values

The triggering voltage of a UJT can be calculated from the values of R1 and R2 in Figure 27-7 and the voltage at base 1. The term R_{bb} is often used to represent the sum of R1 + R2, with V_{bb} representing the total voltage across the two resistors (this is the same as the supply voltage in Figure 27-6). V_t, the triggering voltage, is given by:

$$V_t = V_{bb} * (R1 / R_{bb})$$

The term $(R1/R_{bb})$ is known as the *standoff ratio*, often represented by the Greek letter ϱ.

Typically the standoff ratio in a UJT is at least 0.7, as R1 is chosen to be larger than R2. Typical values for R1 and R2 could be 180Ω and 100Ω, respectively. If R4 is 50K and a 100K linear potentiometer is used for R3, the PUT should be triggered when the potentiometer is near the center of its range. The emitter saturation voltage is typically from 2V to 4V.

If using a PUT, typical values in the test circuit could be supply voltage ranging from 9VDC to 20VDC, with resistances 28K for R1 and 16K for R2, 20Ω for R5, 280K for R4, and a 500K linear potentiometer for R3. The PUT should be triggered when the potentiometer is near the center of its range.

Sustained forward current from anode to cathode is usually a maximum of 150mA, while from gate to cathode the maximum is usually 50mA. Power dissipation should not exceed 300mW. These values should be lower at temperatures above 25 degrees Centigrade.

Depending on the PUT being used, power consumption can be radically decreased by upping the resistor values by a multiple of 100, while supply voltage can be decreased to 5V. The cathode output from the PUT would then be connected with the base of an NPN transistor for amplification.

How to Use it

Figure 27-8 shows a simple oscillator circuit built around a UJT, Figure 27-9 shows a comparable circuit for a PUT. Initially the supply voltage charges the capacitor, until the potential at the emitter of the UJT or the gate of the PUT reaches the threshold voltage, at which point the capacitor discharges through the emitter and the cycle repeats. Resistor values would be similar to those used in the test circuits previously described, while a capacitor value of 2.2µF would provide a visible pulse of the LED. Smaller capacitor values would enable faster oscillation. In the PUT circuit, adjusting the values of R1 and R2 would allow fine control of triggering the semiconductor.

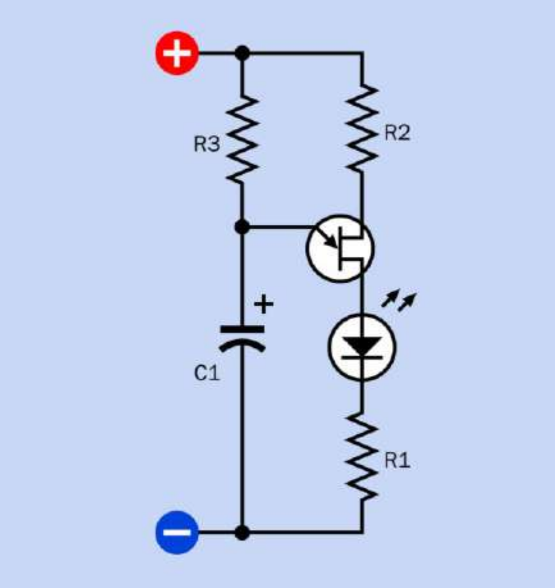

Figure 27-8. *A basic oscillator circuit using a unijunction transistor (UJT). As the capacitor accumulates charge, the voltage on the emitter increases until it triggers the UJT, at which point the capacitor discharges through the emitter.*

Probably the most common use for a PUT at this time is to trigger a **thyristor**.

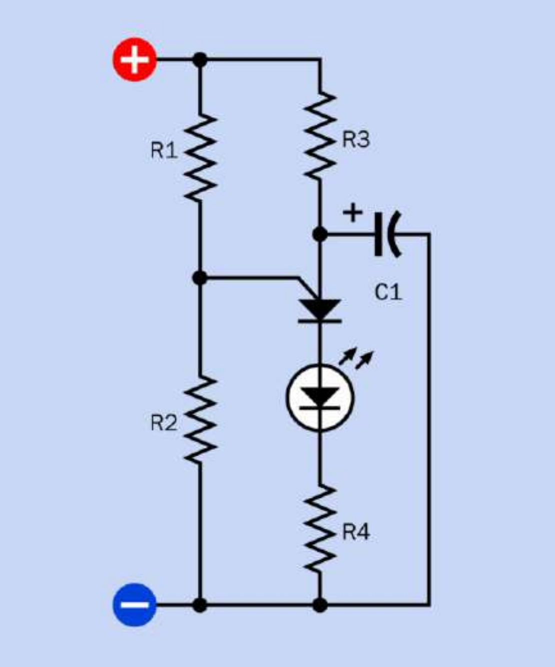

Figure 27-9. *A basic oscillator circuit using a programmable unijunction transistor (PUT). As the capacitor accumulates charge, the voltage on the anode increases until it triggers the PUT, at which point the capacitor discharges through the anode. The gate voltage is preset by R1 and R2 to adjust the triggering voltage.*

What Can Go Wrong

Name Confusion

A programmable unijunction transistor (PUT) is sometimes referred to simply as a "unijunction transistor" (UJT). Bearing in mind the totally different modes of operation of UJT and PUT, the PUT should always be identified by its acronym or by its full name. A circuit will not function if a UJT is substituted for a PUT, or a PUT is substituted for a UJT.

Incorrect Bias

Neither the UJT nor the PUT is designed to operate with reverse bias. In the UJT, a small forward bias should be applied from base 2 to base 1 (that is, base 2 should be at a higher potential relative to base 1) regardless of the voltage on the emitter. The emitter voltage may vary from 0 volts upward. The PUT must be forward biased between its anode and cathode (the anode must have a higher potential relative to the cathode), with an intermediate positive voltage at the gate established by resistors R1 and R2 functioning as a voltage divider (see Figure 27-7). Failure to observe correct biasing will result in unpredictable behavior and possible damage to the component.

Overload

Like any semiconductor, the UJT and the PUT must be protected from excessive current, which can burn out the component. Never connect either of these components directly across a power source without appropriate resistances to limit current flow. Maximum continuous power dissipation for UJTs and PUTs is usually 300mW.

bipolar transistor

28

The word *transistor*, on its own, is often used to mean *bipolar transistor*, as this was the type that became most widely used in the field of discrete semiconductors. However, *bipolar transistor* is the correct term. It is sometimes referred to as a *bipolar junction transistor* or *BJT*.

OTHER RELATED COMPONENTS

- **unijunction transistor** (See Chapter 27)
- **field-effect transistor** (See Chapter 29)
- **diac** (Volume 2)
- **triac** (Volume 2)
- **relay** (Volume 2)
- **solid-state relay** (Volume 2)

What It Does

A bipolar transistor amplifies fluctuations in current or can be used to switch current on and off. In its amplifying mode, it replaced the *vacuum tubes* that were formerly used in the amplification of audio signals and many other applications. In its switching mode it resembles a **relay**, although in its "off" state the transistor still allows a very small amount of current flow, known as *leakage*.

A bipolar transistor is described as a *discrete* semiconductor device when it is individually packaged, with three leads or contacts. A package containing multiple transistors is an *integrated circuit*. A Darlington pair actually contains two transistors, but is included here as a discrete component because it is packaged similarly and functions like a single transistor. Most integrated circuits will be found in Volume 2 of this encyclopedia.

How It Works

Although the earliest transistors were fabricated from germanium, silicon has become the most commonly used material. Silicon behaves like an insulator, in its pure state at room temperature, but can be "doped" (carefully contaminated) with impurities that introduce a surplus of electrons unbonded from individual atoms. The result is an *N-type semiconductor* that can be induced to allow the movement of electrons through it, if it is *biased* with an external voltage. *Forward bias* means the application of a positive voltage, while *reverse bias* means reversing that voltage.

Other dopants can create a deficit of electrons, which can be thought of as a surplus of "holes" that can be filled by electrons. The result is a *P-type semiconductor*.

A bipolar NPN transistor consists of a thin central P-type layer sandwiched between two thicker N-type layers. The three layers are referred to as *collector*, *base*, and *emitter*, with a wire or contact attached to each of them. When a negative charge is applied to the emitter, electrons are forced by mutual repulsion toward the central base layer. If a forward bias (positive potential) is applied to the base, electrons will tend to be attracted out through the base. However, because the base layer is so thin, the electrons are now close to the collector. If the base voltage increases, the additional energy encourages the electrons to jump into the collector, from which they will make their way to the positive current source, which can be thought of as having an even greater deficit of electrons.

Thus, the emitter of an NPN bipolar transistor emits electrons into the transistor, while the collector collects them from the base and moves them out of the transistor. It is important to remember that since electrons carry a negative charge, the flow of electrons moves from negative to positive. The concept of positive-to-negative current is a fiction that exists only for historical reasons. Nevertheless, the arrow in a transistor schematic symbol points in the direction of conventional (positive-to-negative) current.

In a PNP transistor, a thin N-type layer is sandwiched between two thicker P-type layers, the base is negatively biased relative to the emitter, and the function of an NPN transistor is reversed, as the terms "emitter" and "collector" now refer to the movement of electron-holes rather than electrons. The collector is negative relative to the base, and the resulting positive-to-negative current flow moves from emitter to base to collector. The arrow in the schematic symbol for a PNP transistor still indicates the direction of positive current flow.

Symbols for NPN and PNP transistors are shown in Figure 28-1. The most common symbol for an NPN transistor is shown at top-left, with letters C, B, and E identifying collector, base, and emitter. In some schematics the circle in the symbols is omitted, as at top-right.

A PNP transistor is shown in the center. This is the most common orientation of the symbol, since its collector must be at a lower potential than its emitter, and ground (negative) is usually at the bottom of a schematic. At bottom, the PNP symbol is inverted, allowing the positions of emitter and collector to remain the same as in the symbol for the NPN transistor at the top. Other orientations of transistor symbols are often found, merely to facilitate simpler schematics with fewer conductor crossovers. The direction of the arrow in the symbol (pointing out or pointing in) always differentiates NPN from PNP transistors, respectively, and indicates current flowing from positive to negative.

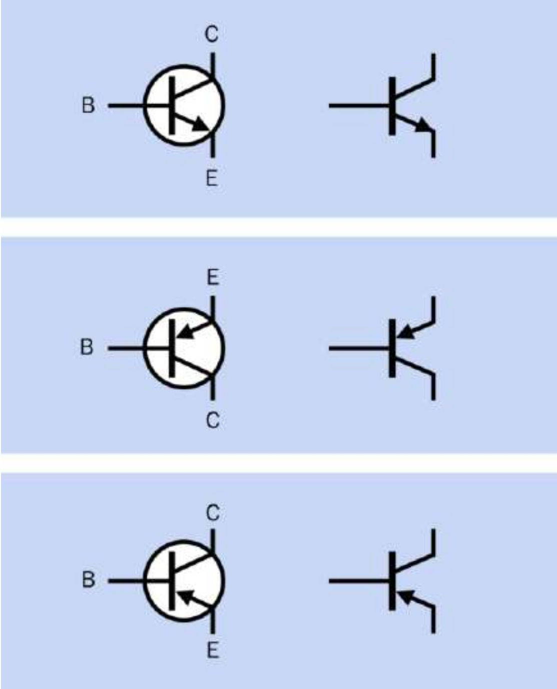

Figure 28-1. *Symbols for an NPN transistor (top) and a PNP transistor (center and bottom). Depending on the schematic in which the symbol appears, it may be rotated or inverted. The circle may be omitted, but the function of the component remains the same.*

NPN transistors are much more commonly used than PNP transistors. The PNP type was more difficult and expensive to manufacture initially, and circuit design evolved around the NPN type. In addition, NPN transistors enable faster switching, because electrons have greater mobility than electron-holes.

To remember the functions of the collector and the emitter in an NPN transistor, you may prefer to think in terms of the collector collecting positive current *into* the transistor, and the emitter emitting positive current *out of* the transistor. To remember that the emitter is always the terminal with an arrow attached to it (both in NPN and PNP schematic symbols), consider that "emitter" and "arrow" both begin with a vowel, while "base" and "collector" begin with consonants. To remember that an NPN transistor symbol has its arrow pointing outward, you can use the mnemonic "N/ever P/ointing i/N."

Current flow for NPN and PNP transistors is illustrated in Figure 28-2. At top-left, an NPN transistor passes no current (other than a small amount of leakage) from its collector to its emitter so long as its base is held at, or near, the potential of its emitter, which in this case is tied to negative or ground. At bottom-left, the purple positive symbol indicates that the base is now being held at a *relatively* positive voltage, at least 0.6 volts higher than the emitter (for a silicon-based transistor). This enables electrons to move from the emitter to the collector, in the direction of the blue arrows, while the red arrows indicate the conventional concept of current flowing from positive to negative. The smaller arrows indicate a smaller flow of current. A resistor is included to protect the transistor from excessive current, and can be thought of as the load in these circuits.

At top-right, a PNP transistor passes no current (other than a small amount of leakage) from its emitter to its collector so long as its base is held at, or near, the potential of the emitter, which in this case is tied to the positive power supply. At bottom-right, the purple negative symbol indicates that the base is now being held at a *relatively* negative voltage, at least 0.6 volts lower than the emitter. This enables electrons and current to flow as shown. Note that current flows into the base in the NPN transistor, but out from the base in the PNP transistor, to enable conductivity. In both diagrams, the resistor that would normally be included to protect the base has been omitted for the sake of simplicity.

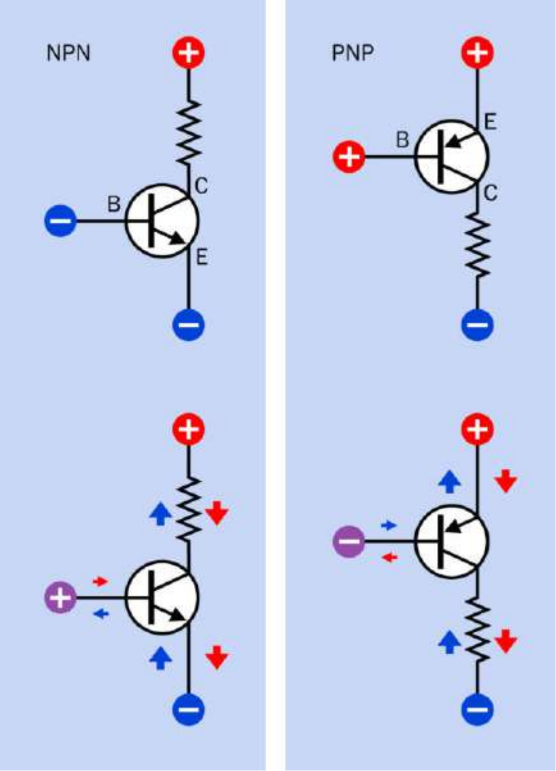

Figure 28-2. *Current flow through NPN and PNP transistors. See text for details.*

An NPN transistor amplifies its base current only so long as the positive potential applied to the collector is greater than the potential applied to the base, and the potential at the base must be greater than the potential at the emitter by at least 0.6 volts. So long as the transistor is biased in this way, and so long as the current values remain within the manufacturer's specified limits, a small fluctuation in current applied to the base will induce a much larger fluctuation in current between the collector and the emitter. This is why a transistor may be described as a *current amplifier*.

A *voltage divider* is often used to control the base potential and ensure that it remains less than the potential on the collector and greater than the potential at the emitter (in an NPN transistor). See Figure 28-3.

See Chapter 10 for additional information about the function of a voltage divider.

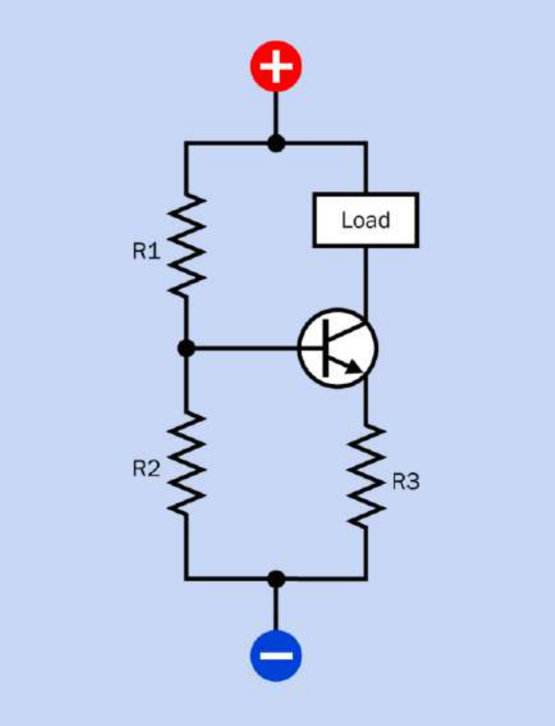

Figure 28-3. *Resistors R1 and R2 establish a voltage divider to apply acceptable bias to the base of an NPN transistor.*

Current Gain

The amplification of current by a transistor is known as its *current gain* or *beta value*, which can be expressed as the ratio of an increase in collector current divided by the increase in base current that enables it. Greek letter β is customarily used to represent this ratio. The formula looks like this:

$$\beta = \Delta I_c \ / \ \Delta I_b$$

where I_c is the collector current and I_b is the base current, and the Δ symbol represents a small change in the value of the variable that follows it.

Current gain is also represented by the term h_{FE}, where E is for the common Emitter, F is for Forward current, and lowercase letter h refers to the transistor as a “hybrid” device.

The beta value will always be greater than 1 and is often around 100, although it will vary from one type of transistor to another. It will also be affected by temperature, voltage applied to the transistor, collector current, and manufacturing inaccuracies. When the transistor is used outside of its design parameters, the formula to determine the beta value no longer directly applies.

There are only two connections at which current can enter an NPN transistor and one connection where it can leave. Therefore, if I_e is the current from the emitter, I_c is the current entering the collector, and I_b is the current entering the base:

$$I_e = I_c + I_b$$

If the potential applied to the base of an NPN transistor diminishes to the point where it is less than 0.6V above the potential at the emitter, the transistor will not conduct, and is in an “off” state, although a very small amount of leakage from collector to emitter will still occur.

When the current flowing into the base of the transistor rises to the point where the transistor cannot amplify the current any further, it becomes *saturated*, at which point its internal impedance has fallen to a minimal value. Theoretically this will allow a large flow of current; in practice, the transistor should be protected by resistors from being damaged by high current resulting from saturation.

Any transistor has maximum values for the collector current, base current, and the potential difference between collector and emitter. These values should be provided in a datasheet. Exceeding them is likely to damage the component.

Terminology

In its *saturated mode*, a transistor's base is saturated with electrons (with no room for more) and the internal impedance between collector and emitter drops as low as it can go.

The *cutoff mode* of an NPN transistor is the state where a low base voltage eliminates all current flow from collector to emitter other than a small amount of leakage.

The *active mode*, or *linear mode*, is the intermediate condition between cutoff and saturated, where the beta value or h_{FE} (ratio of collector current to base current) remains approximately constant. That is, the collector current is almost linearly proportional to the base current. This linear relationship breaks down when the transistor nears its saturation point.

Variants

Small signal transistors are defined as having a maximum collector current of 500 mA and maximum collector power dissipation of 1 watt. They can be used for audio amplification of low-level inputs and for switching of small currents. When determining whether a small-signal transistor can control an inductive load such as a motor or relay coil, bear in mind that the initial current surge will be greater than the rated current draw during sustained operation.

Small switching transistors have some overlap in specification with small signal transistors, but generally have a faster response time, lower beta value, and may be more limited in their tolerance for collector current. Check the manufacturer's datasheet for details.

High frequency transistors are primarily used in video amplifiers and oscillators, are physically small, and have a maximum frequency rating as high as 2,000 MHz.

Power transistors are defined as being capable of handling at least 1 watt, with upper limits that can be as high as 500 watts and 150 amps. They are physically larger than the other types, and may be used in the output stages of audio amplifiers, and in switching **power supplies** (see Chapter 16). Typically they have a much lower current gain than smaller transistors (20 or 30 as opposed to 100 or more).

Sample transistors are shown in Figure 28-4. Top: A 2N3055 NPN power transistor. This type was originally introduced in the late 1960s, and versions are still being manufactured. It is often found in power supplies and in push-pull power amplifiers, and has a total power dissipation rating of 115W. Second row, far left: general purpose switching-amplification PNP power transistor rated for up to 50W power dissipation. Second row, far right: A high-frequency switching transistor for use in lighting ballast, converters, inverters, switching regulators, and motor control systems. It tolerates relatively high voltages (up to 700V collector-emitter peak) and is rated for up to 80W total power dissipation. Second row, center-left and center-right: Two variants of the 2N2222 NPN small signal switching transistor, first introduced in the 1960s, and still very widely used. The metal can is the TO-19 package, capable of slightly higher power dissipation than the cheaper plastic TO-92 package (1.8W vs. 1.5W with a collector temperature no greater than 25 degrees Centigrade).

Packaging

Traditionally, small-signal transistors were packaged in small aluminum "cans" about 1/4" in diameter, and are still sometimes found in this form. More commonly they are embedded in buds of black plastic. Power transistors are packaged either in a rectangular module of black

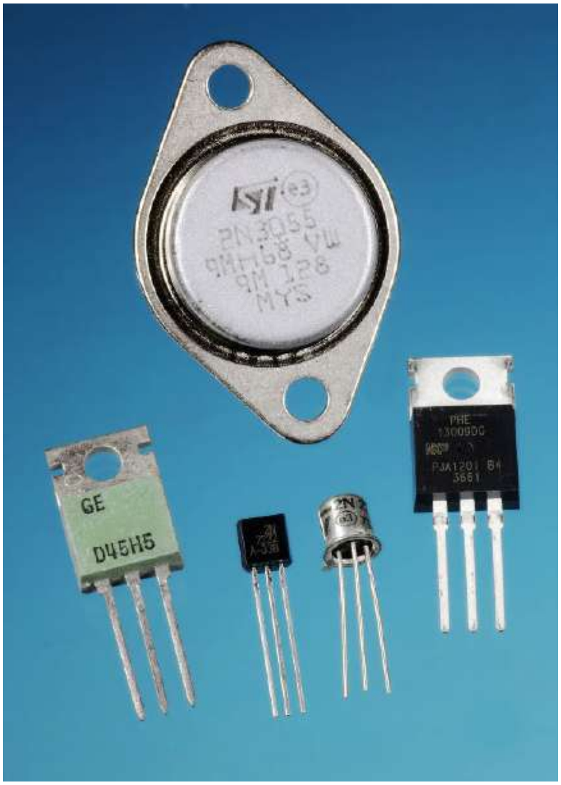

Figure 28-4. *Samples of commonly used transistors. See text for details.*

plastic with a metal back, or in a round metal "button." Both of these forms are designed to dissipate heat by being screw-clamped to a *heat sink*.

Connections

Often a transistor package provides no clue as to which lead is the emitter, which lead is the base, and which lead is the collector. Old can-style packaging includes a protruding tab that usually points toward the emitter, but not always. Where power transistors are packaged in a metal enclosure, it is typically connected internally with the collector. In the case of surface-mount transistors, look for a dot or marker that should identify the base of a bipolar transistor or the gate of a field-effect transistor.

A through-hole transistor usually has its part number printed or engraved on its package, although a magnifying glass may be necessary to see this. The component's datasheet may then be checked online. If a datasheet is unavailable, meter-testing will be necessary to confirm the functions of the three transistor leads. Some multimeters include a transistor-test function, which may validate the functionality of a transistor while also displaying its beta value. Otherwise, a meter can be put in diode-testing mode, and an unpowered NPN transistor should behave as if diodes are connected between its leads as shown in Figure 28-5. Where the identities of the transistor's leads are unknown, this test will be sufficient to identify the base, after which the collector and emitter may be determined empirically by testing the transistor in a simple low-voltage circuit such as that shown in Figure 28-6.

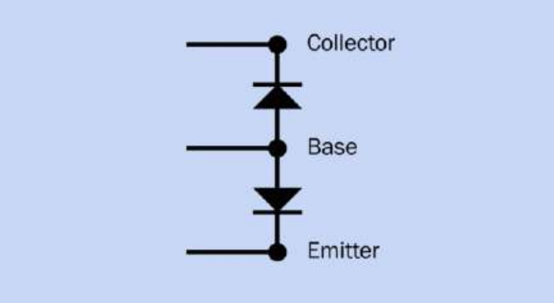

Figure 28-5. *An NPN transistor can behave as if it contains two diodes connected as shown. Where the functions of the leads of the transistor are unknown, the base can be identified by testing for conductivity.*

How to Use it

The following abbreviations and acronyms are common in transistor datasheets. Some or all of the letters following the initial letter are usually, but not always, formatted as subscripts:

h_{FE}
Forward current gain

β
Same as hFE

V_{CEO}
Voltage between collector and emitter (no connection at base)

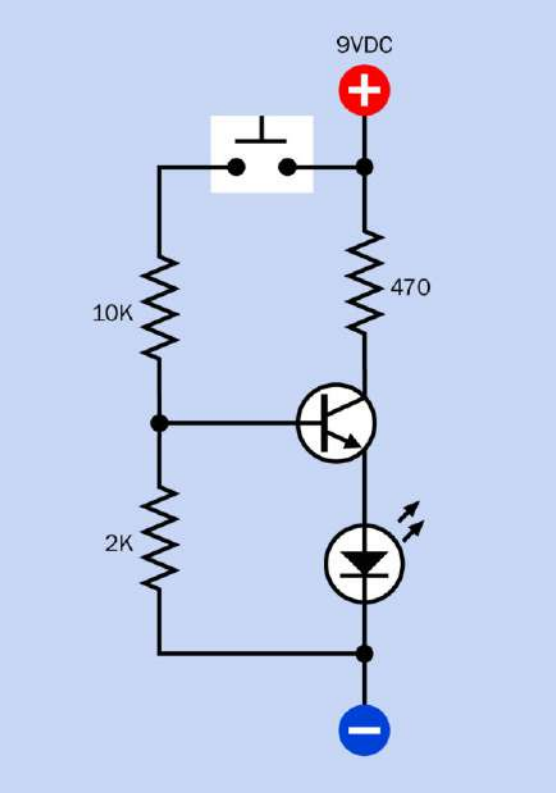

Figure 28-6. *This simple schematic can be used to breadboard-test a transistor empirically, determining its functionality and the identities of its collector and emitter leads.*

V_{CBO}

Voltage between collector and base (no connection at emitter)

V_{EBO}

Voltage between emitter and base (no connection at collector)

V_{CEsat}

Saturation voltage between collector and emitter

V_{BEsat}

Saturation voltage between base and emitter

I_c

Current measured at collector

I_{CM}

Maximum peak current at collector

I_{BM}

Maximum peak current at base

P_{TOT}

Total maximum power dissipation at room temperature

T_J

Maximum junction temperature to avoid damage

Often these terms are used to define "absolute maximum values" for a component. If these maximums are exceeded, damage may occur.

A manufacturer's datasheet may include a graph showing the *safe operating area* (SOA) for a transistor. This is more common where power transistors are involved, as heat becomes more of an issue. The graph in Figure 28-7 has been adapted from a datasheet for a silicon diffused power transistor manufactured by Philips. The safe operating area is bounded at the top by a horizontal segment representing the maximum safe current, and at the right by a vertical segment representing the maximum safe voltage. However, the rectangular area enclosed by these lines is reduced by two diagonal segments representing the *total power limit* and the *second breakdown limit*. The latter refers to the tendency of a transistor to develop internal localized "hot spots" that tend to conduct more current, which makes them hotter, and able to conduct better—ultimately melting the silicon and causing a short circuit. The total power limit and the second breakdown limit reduce the safe operating area, which would otherwise be defined purely by maximum safe current and maximum safe voltage.

Uses for discrete transistors began to diminish when integrated circuits became cheaper and started to subsume multi-transistor circuits. For instance, an entire 5-watt audio amplifier, which used to be constructed from multiple components can now be bought on a chip, requiring just

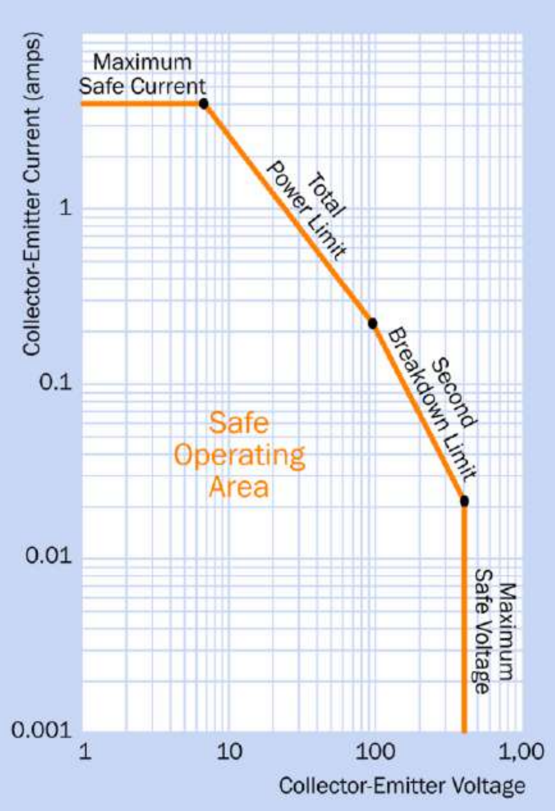

Figure 28-7. *Adapted from a Philips datasheet for a power transistor, this graph defines a safe operating area (SOA) for the component. See text for details.*

a few external capacitors. More powerful audio equipment typically uses integrated circuits to process inputs, but will use individual power transistors to handle high-wattage output.

Darlington Pairs

Discrete transistors are useful in situations where current amplification or switching is required at just one location in a circuit. An example would be where one output pin from a **microcontroller** must switch a small motor on and off. The motor may run on the same voltage as the microcontroller, but requires considerably more current than the typical 20mA maximum available from a microcontroller output. A *Darlington pair* of transistors may be used in this application. The overall gain of the pair can be 100,000 or more. See Figure 28-8. If a power source feeding through a potentiometer is substituted for the microcontroller chip, the circuit can function as a motor speed control (assuming that a generic DC motor is being used).

In the application shown here, the microcontroller chip must share a common ground (not shown) with the transistors. The optional resistor may be necessary to prevent leakage from the first transistor (when in its “off” state) from triggering the second. The diode protects the transistors from voltage transients that are likely when the motor stops and starts.

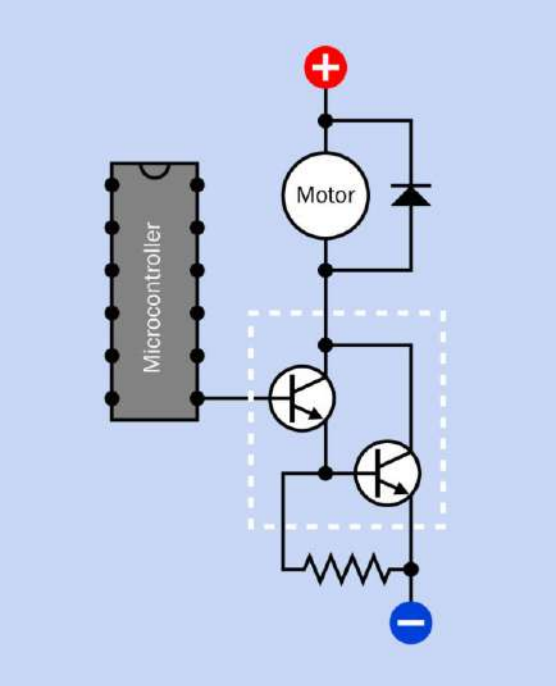

Figure 28-8. *Where the emitter of one NPN transistor is coupled to the base of another, they form a Darlington pair (identified by the dashed rectangle in this schematic). Multiplying the gain of the first transistor by the gain of the second gives the total gain of the pair.*

A Darlington pair can be obtained in a single transistor-like package, and may be represented by the schematic symbol shown in Figure 28-9.

Various through-hole Darlington packages are shown in Figure 28-10.

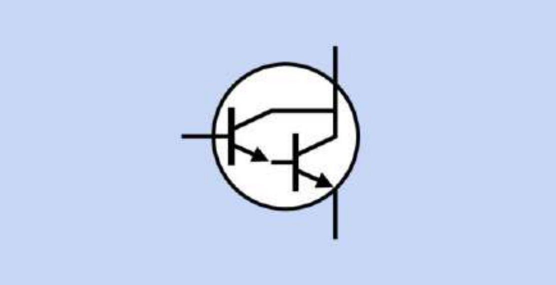

Figure 28-9. *When a Darlington pair is embedded in a single transistor-like package, it may be represented by this schematic symbol. The leads attached to the package can be used as if they are the emitter, base, and collector of a single NPN transistor.*

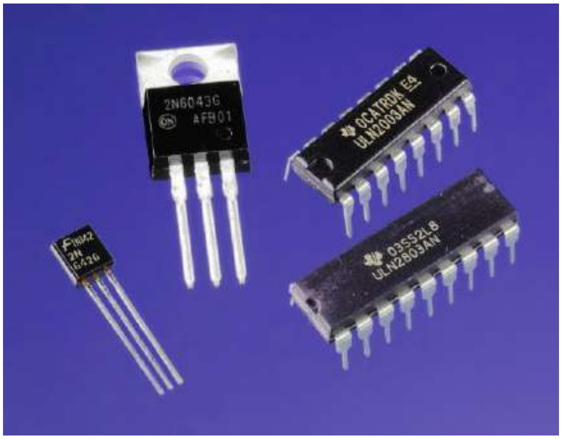

Figure 28-10. *Various packaging options for Darlington pairs. From left to right: The 2N6426 contains a Darlington pair rated to pass up to 500mA continuous collector current. The 2N6043 is rated for 8A continuous. The ULN2003 and ULN2083 chips contain seven and eight Darlington pairs, respectively.*

Seven or eight Darlington pairs can be obtained in a single integrated chip. Each transistor pair in these chips is typically rated at 500mA, but they can be connected in parallel to allow higher currents. The chip usually includes protection diodes to allow it to drive inductive loads directly.

A typical schematic is shown in Figure 28-11. In this figure, the microcontroller connections are hypothetical and do not correspond with any actual chip. The Darlington chip is a ULN2003 or similar, containing seven transistor pairs, each with an "input" pin on the left and an "output" pin opposite it on the right. Any of pins 1 through 7 down the left side of the chip can be used to control a device connected to a pin on the opposite side.

A high input can be thought of as creating a negative output, although in reality the transistors inside the chip are sinking current via an external device—a motor, in this example. The device can have its own positive supply voltage, shown here as 12VDC, but must share a common ground with the microcontroller, or with any other component which is being used on the input side. The lower-right pin of the chip shares the 12VDC supply because this pin is attached internally to clamp diodes (one for each Darlington pair), which protect against surges caused by inductive loads. For this reason, the motor does not have a clamp diode around it in the schematic.

The Darlington chip does not have a separate pin for connection with positive supply voltage, because the transistors inside it are sinking power from the devices attached to it.

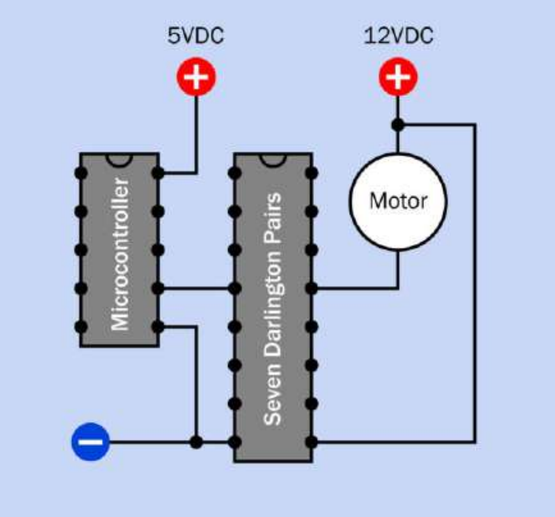

Figure 28-11. *A chip such as the ULN2003 contains seven Darlington pairs. It will sink current from the device it is driving. See text for details.*

A surface-mount Darlington pair is shown in Figure 28-12. This measures just slightly more than 0.1" long but is still rated for up to 500mA collector current or 250mW total power dissipation (at a component temperature no higher than 25 degrees Centigrade).

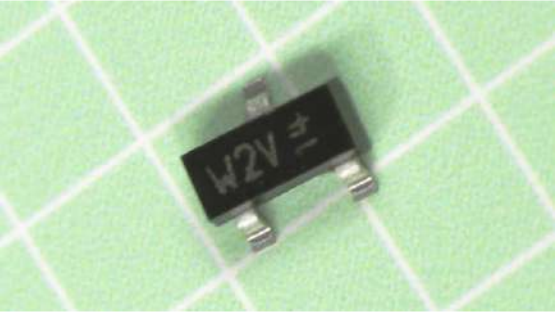

Figure 28-12. *A surface-mount package for a Darlington pair. Each square in the background grid measures 0.1". See text for additional details.*

Amplifiers

Two basic types of transistor amplifiers are shown in Figure 28-13 and Figure 28-14. The *common-collector* configuration has current gain but no voltage gain. The capacitor on the input side blocks DC current from entering the amplifier circuit, and the two resistors forming a voltage divider on the base of the transistor establish a voltage midpoint (known as the *quiescent point* or *operating point*) from which the signal to be amplified may deviate above and below.

The *common-emitter* amplifier provides voltage gain instead of current gain, but inverts the phase of the input signal. Additional discussion of amplifier design is outside the scope of this encyclopedia.

In switching applications, modern transistors have been developed to handle a lot of current compared with earlier versions, but still have some limitations. Few power transistors can handle more than 50A flowing from collector to emitter, and 1,000V is typically a maximum value. Electromechanical relays continue to exist because they retain some advantages, as shown in the table in Figure 28-15, which compares switching capabilities of transistors, **solid-state relays**, and electromechanical **relays**.

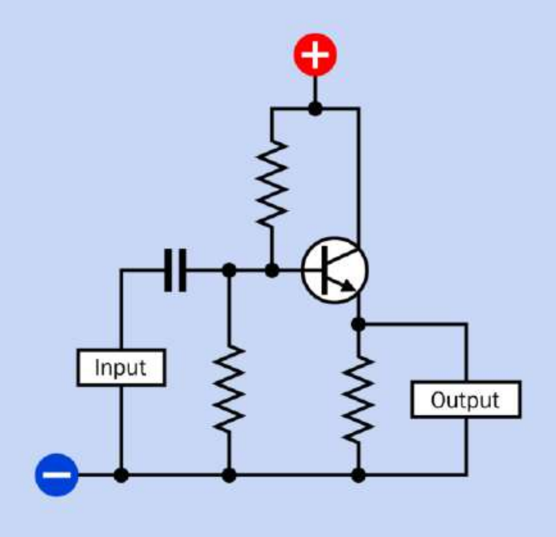

Figure 28-13. *The basic schematic for a common-collector amplifier. See text for details.*

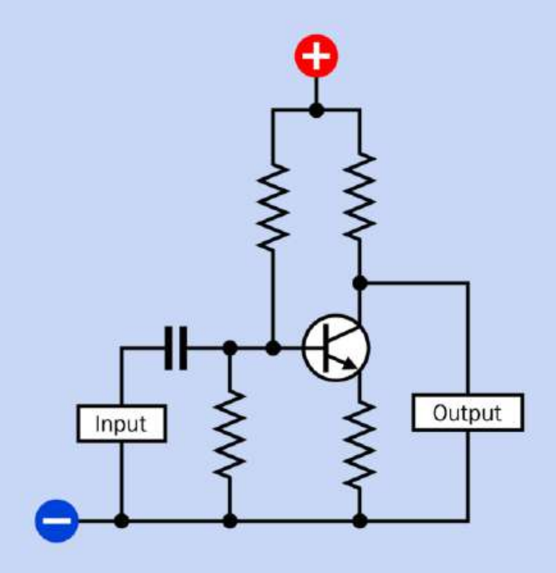

Figure 28-14. *The basic schematic for a common-emitter amplifier. See text for details.*

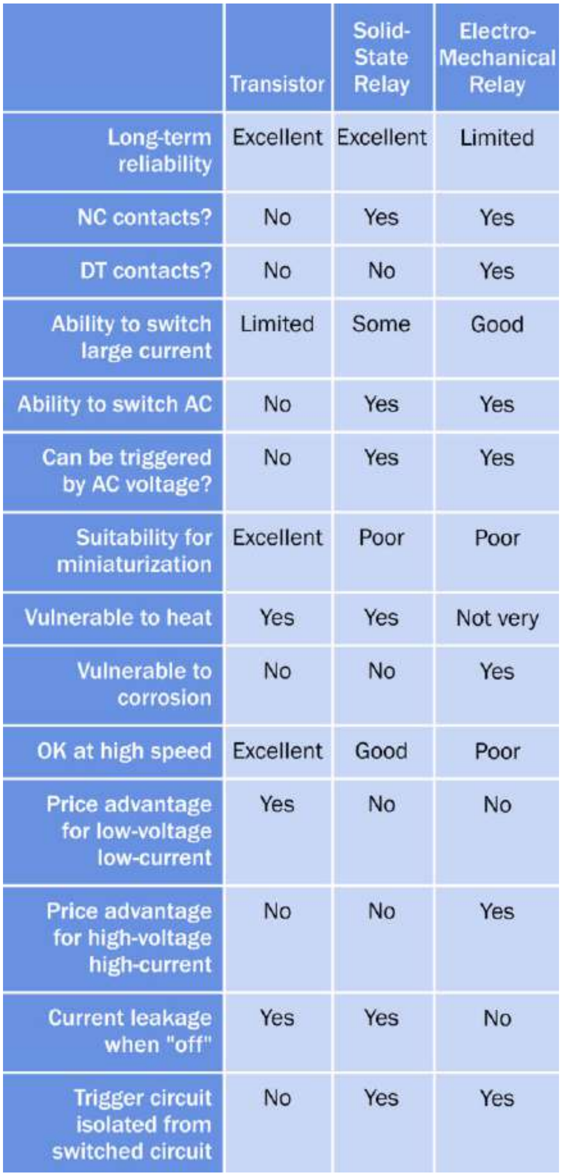

	Transistor	Solid-State Relay	Electro-Mechanical Relay
Long-term reliability	Excellent	Excellent	Limited
NC contacts?	No	Yes	Yes
DT contacts?	No	No	Yes
Ability to switch large current	Limited	Some	Good
Ability to switch AC	No	Yes	Yes
Can be triggered by AC voltage?	No	Yes	Yes
Suitability for miniaturization	Excellent	Poor	Poor
Vulnerable to heat	Yes	Yes	Not very
Vulnerable to corrosion	No	No	Yes
OK at high speed	Excellent	Good	Poor
Price advantage for low-voltage low-current	Yes	No	No
Price advantage for high-voltage high-current	No	No	Yes
Current leakage when "off"	Yes	Yes	No
Trigger circuit isolated from switched circuit	No	Yes	Yes

Figure 28-15. *A comparison of characteristics of switching devices.*

What Can Go Wrong

Wrong Connections on a Bipolar Transistor

Failing to identify a transistor's leads or contacts correctly can obviously be a potential source of damage, but swapping the collector and emitter accidentally will not necessarily destroy the transistor. Because of the inherent symmetry of the device, it will in fact function with collector and emitter connections reversed. Rohm, a large semiconductor manufacturer, has included this scenario in its general information pages and concludes that the primary indicator of transposed connections is that the β value, or h_{FE} , drops to about 1/10th of specification. If you are using a transistor that works but provides much less amplification than you expect, check that the emitter and collector leads are not transposed.

Wrong Connections on a Darlington Pair Chip

While a single-component package for a Darlington pair functions almost indistinguishably from a single transistor, multiple Darlington pairs in a DIP package may create confusion because the component behaves differently from most other chips, such as logic chips.

A frequent error is to ground the output device instead of applying positive power to it. See Figure 28-11 and imagine an erroneous connection of negative power instead of the 12VDC positive power.

Additional confusion may be caused by reading a manufacturer's datasheet for a Darlington pair chip such as the ULN2003. The datasheet depicts the internal function of the chip as if it contains logic inverters. While the chip can be imagined as behaving this way, in fact it contains bipolar transistors that amplify the current applied to the base of each pair. The datasheet also typically will not show the positive connection that should be made to the common-diode pin (usually at bottom-right), to provide protection from surges caused by inductive loads. This pin must be distinguished carefully from the common-ground pin (usually at bottom-left). The positive connection to the common-diode pin is optional; the common-ground connection is mandatory.

Soldering Damage

Like any semiconductor, transistors are vulnerable to heat and can be damaged while soldering, although this seldom happens if a low-wattage iron is used. A copper alligator clip can be applied as a heat sink to each lead before it is soldered.

Excessive Current or Voltage

During use, a transistor will be damaged if it is subjected to current or voltage outside of its rated range. Passing current through a transistor without any series resistance to protect it will almost certainly burn it out, and the same thing can happen if incorrect resistor values are used.

The maximum wattage that a transistor can dissipate will be shown in its datasheet. Suppose, for example, this figure is 200mW, and you are using a 12VDC supply. Ignoring the base current, the maximum collector current will be 200 / 12 = approximately 15mA. If the transistor's emitter is connected to ground, and the load applied to the transistor output has a high impedance, and if we ignore the transresistance, Ohm's Law suggests that a resistor that you place between the collector and the supply voltage should have a resistance of at least 12 / 0.015 = 800 ohms.

When transistors are used in switching applications, it is customary for the base current to be 1/5th of the collector current. In the example discussed here, a 4.7K resistor might be appropriate. A meter should be used to verify actual current and voltage values.

Excessive Leakage

In a Darlington pair, or any other configuration where the output from one transistor is connected with the base of another, leakage from the first transistor while in its "off" state can be amplified by the second transistor. If this is unacceptable, a bypass resistor can be used to divert some of the leakage from the base of the second transistor to ground. Of course the resistor will also steal some of the base current when the first transistor is active, but the resistor value is typically chosen so that it takes no more than 10% of the active current. See Figure 28-8 for an example of a bypass resistor added to a Darlington pair.

field effect transistor

28

The term **field-effect transistor** encompasses a family primarily consisting of the *junction field-effect transistor* (or *JFET*, which is the simplest generic type) and the *metal-oxide semiconductor field-effect transistor* (or *MOSFET*, also sometimes known as an *insulated-gate field-effect transistor*, or *IGFET*). Because the principles of operation overlap considerably, the entire -FET family is grouped in this entry.

OTHER RELATED COMPONENTS

- **bipolar transitor** (See Chapter 28)
- **unijunction transistor** (See Chapter 27)
- **diode** (See Chapter 26)

What It Does

A field-effect transistor creates an electric field to control current flowing through a channel in a semiconductor. MOSFETs of microscopic size form the basis of complementary metal oxide semiconductor (CMOS) *integrated circuit chips*, while large discrete MOSFETs are capable of switching substantial currents, in lamp dimmers, audio amplifiers, and motor controllers. FETs have become indispensable in computer electronics.

A **bipolar transistor** is generally thought of as a *current amplifier* because the current passing through it is controlled by a smaller amount of current passing through the base. By contrast, all FETs are considered to be *voltage amplifiers*, as the control voltage establishes field intensity, which requires little or no current. The negligible leakage through the gate of an FET makes it ideal for use in low-power applications such as portable hand-held devices.

How It Works

This section is divided into two subsections, describing the most widely used FETs: JFETs and MOSFETs.

JFETs

A *junction field-effect transistor* (or *JFET*) is the simplest form of FET. Just as a bipolar transistor can be of NPN or PNP type, a JFET can have an *N-channel* or *P-channel*, depending whether the channel that transmits current through the device is negatively or positively *doped*. A detailed explanation of semiconductor doping will be found in the **bipolar transistor** entry.

Because negative charges have greater mobility, the N-channel type allows faster switching and is more commonly used than the P-channel type. A schematic symbol for it is shown in Figure 29-1 alongside the schematic for an NPN transistor. These symbols suggest the similarity of the devices as amplifiers or switches, but it is important to remember that the FET is a primarily a voltage amplifier while the bipolar transistor is a current amplifier.

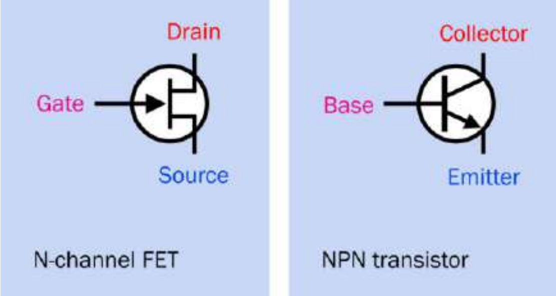

Figure 29-1. *A comparison between schematic symbols for N-channel JFET (left) and NPN bipolar transistor (right) suggests their functional similarity as switches or amplifiers, although their behavior is markedly different.*

Three JFETs are shown in Figure 29-2. The N-channel J112 type is supplied by several manufacturers, the figure showing two samples, one from Fairchild Semiconductor (left) and the other from On Semiconductor (right). Although the full part numbers are different, the specifications are almost identical, including a drain-gate voltage of 35V, a drain-source voltage of 35V, and a gate current of 50mA. The metal-clad 2N4392 in the center has similar values but is three times the price, with a much higher power dissipation of 1.8W, compared with 300mW and 350mW for the other two transistors respectively.

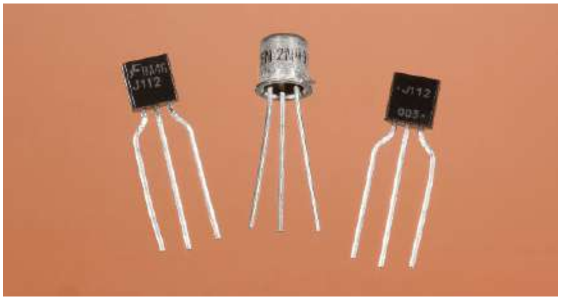

Figure 29-2. *Junction Field Effect Transistors (JFETs). See text for details.*

Schematic symbols for N-channel and P-channel JFETs are shown in Figure 29-3, N-channel being on the left while P-channel is on the right. The upper-left and lower-left symbol variants are both widely used and are functionally identical. The upper-right and lower-right variants likewise mean the same thing. Because the upper variants are symmetrical, an S should be added to clarify which terminal is the source. In practice, the S is often omitted, allowing some ambiguity. While the source and drain of some JFETs are in fact interchangeable, this does not apply to all types.

The circle around each symbol is occasionally omitted when representing discrete components, and is almost always omitted when multiple FETs are shown connected to form an integrated circuit.

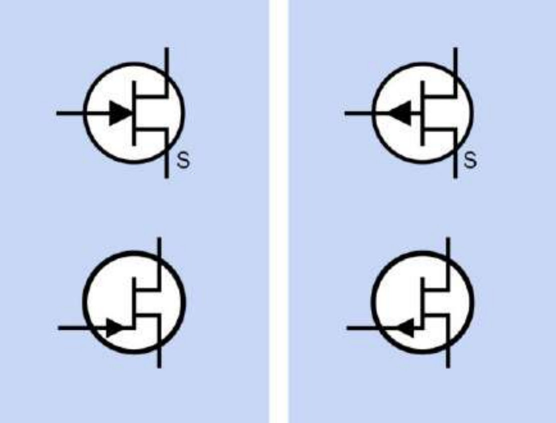

Figure 29-3. *Schematic symbols for junction field-effect transistors (JFETs). Left: N-channel. Right: P-channel. The symbols at top and bottom on each side are functionally identical. Circles may be omitted. Letter S may be omitted from the symmetrical symbol variants, even though this creates some ambiguity.*

The internal function of an N-channel JFET is shown diagrammatically in Figure 29-4. In this component, the source terminal is a source of electrons, flowing relatively freely through the N-doped channel and emerging through the drain. Thus, conventional current flows nonintuitively from the drain terminal, to the source terminal, which will be of lower potential.

The JFET is like a normally-closed switch. It has a low resistance so long as the gate is at the same potential as the source. However, if the potential of the gate is reduced below the potential of the source—that is, the gate acquires a more relatively negative voltage than the source—the current flow is *pinched off* as a result of the field created by the gate. This is suggested by the lower diagram in Figure 29-4.

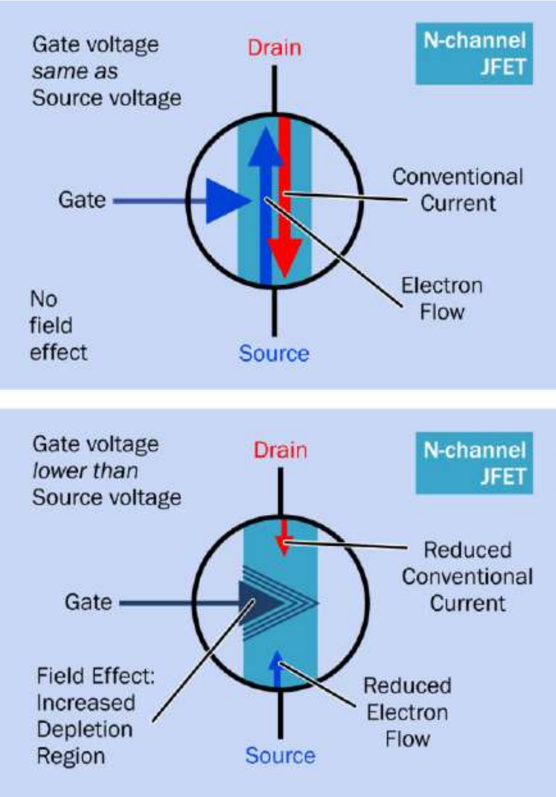
Gate voltage
same as
Source voltage
Drain
N-channel
JFET
Gate
Conventional
Current
Electron
Flow
No
field
effect
Source
Gate voltage
lower than
Source voltage
Drain
N-channel
JFET
Reduced
Conventional
Current
Gate
Field Effect:
Increased
Depletion
Region
Reduced
Electron
Flow
Source

Figure 29-4. *At top, conventional current flows freely from drain to source through the channel of an N-doped JFET. At bottom, the lowered voltage of the gate relative to the source creates a field effect that pinches off the flow of current.*

The situation for a P-channel JFET is reversed, as shown in Figure 29-5. The source is now positive (but is still referred to as the source), while the drain can be grounded. Conventional current now flows freely from source to drain, so long as the gate is at the same positive potential as the source. If the gate voltage rises above the source voltage, the flow is pinched off.

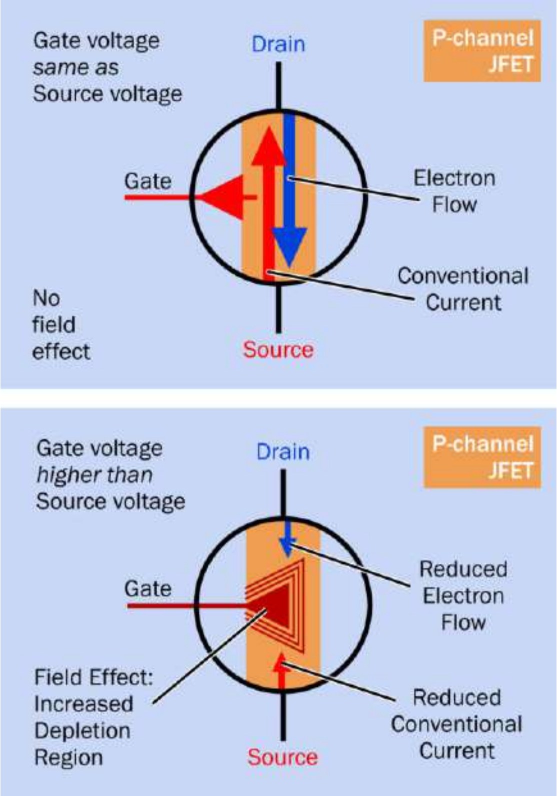
Gate voltage
same as
Source voltage
Drain
P-channel
JFET
Gate
Electron
Flow
Conventional
Current
No
field
effect
Source
Gate voltage
higher than
Source voltage
Drain
P-channel
JFET
Gate
Reduced
Electron
Flow
Field Effect:
Increased
Depletion
Region
Reduced
Conventional
Current
Source

Figure 29-5. *At top, conventional current flows freely from source to drain through the channel of a P-doped JFET. At bottom, the higher voltage of the gate relative to the source creates a field effect that pinches off the flow of current.*

A bipolar transistor tends to block current flow by default, but becomes less resistive when its base is forward-biased. Therefore it can be referred to as an *enhancement device*. By contrast, an N-channel JFET allows current to flow by default, and becomes more resistive when its base is reverse-biased, which widens the *depletion layer* at the base junction. Consequently it can be referred to as a *depletion device*.

The primary characteristics of a junction field-effect transistor relative to an NPN bipolar transistor are summarized in the table in Figure 29-6.

JFET Behavior

The voltage difference between gate and source of a JFET is usually referred to as V_{gs} while the voltage difference between drain and source is referred to as V_{ds}.

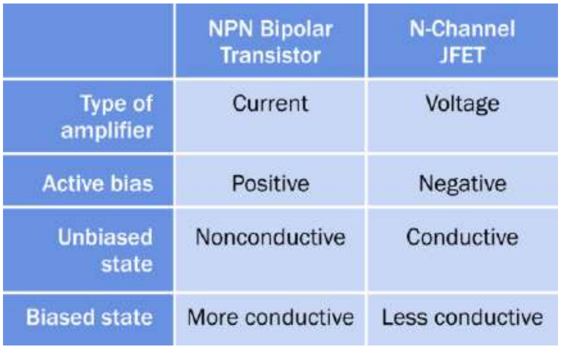

	NPN Bipolar Transistor	N-Channel JFET
Type of amplifier	Current	Voltage
Active bias	Positive	Negative
Unbiased state	Nonconductive	Conductive
Biased state	More conductive	Less conductive

Figure 29-6. *This table contrasts the characteristics of an N-channel JFET with those of an NPN bipolar transistor.*

Suppose the gate of an N-channel JFET is connected with the source, so that $V_{gs}=0$. Now if V_{ds} increases, the current flowing through the channel of the JFET also increases, approximately linearly with V_{ds}. In other words, initially the JFET behaves like a low-value resistor in which the voltage across it, divided by the amperage flowing through it, is approximately constant. This phase of the JFET's behavior is known as its *ohmic region*. While the unbiased resistance of the channel in a JFET depends on the component type, it is generally somewhere between 10Ω and 1K.

If V_{ds} increases still further, eventually no additional flow of current occurs. At this point the channel has become *saturated*, and this plateau zone is referred to as the *saturation region*, often abbreviated I_{dss}, meaning "the saturated drain current at zero bias." Although this is a nearly constant value for any particular JFET, it may vary somewhat from one sample of a component to another, as a result of manufacturing variations.

If V_{ds} continues to increase, the component finally enters a *breakdown* state, sometimes referred to by its full formal terminology as *drain-source breakdown*. The current passing through the JFET will now be limited only by capabilities of the external power source. This breakdown state can be destructive to the component, and is comparable to the breakdown state of a typical **diode**.

What if the voltage at the gate is reduced below the voltage at the source—such as V_{gs} becomes negative? In its ohmic region, the component now behaves as if it has a higher resistance, and it will reach its saturation region at a lower current value (although around the same value for V_{ds}). Therefore, by reducing the voltage on the gate relative to the voltage at the source, the effective resistance of the component increases, and in fact it can behave as a *voltage-controlled resistor*.

The upper diagram in Figure 29-7 shows this graphically. Below it, the corresponding graph for a P-channel JFET looks almost identical, except that the current flow is reversed and is pinched off as the gate voltage rises above the source voltage. Also, the breakdown region is reached more quickly with a P-channel JFET than with an N-channel JFET.

MOSFETs

MOSFETs have become one of the most widely used components in electronics, everywhere from computer memory to high-amperage switching power supplies. The name is an acronym for *metal-oxide semiconductor field-effect transistor*. A simplified cross-section of an N-channel MOSFET is shown in Figure 29-8.

Two MOSFETs are shown in Figure 29-9.

Like a JFET, a MOSFET has three terminals, identified as drain, gate, and source, and it functions by creating a field effect that controls current flowing through a channel. (Some MOSFETS have a fourth terminal, described later). However, it has a metal source and drain making contact with each end of the channel (hence the term "metal" in its acronym) and also has a thin layer of silicon dioxide (hence the term "oxide" in its acronym) separating the gate from the channel, thus raising the impedance at the gate to at least

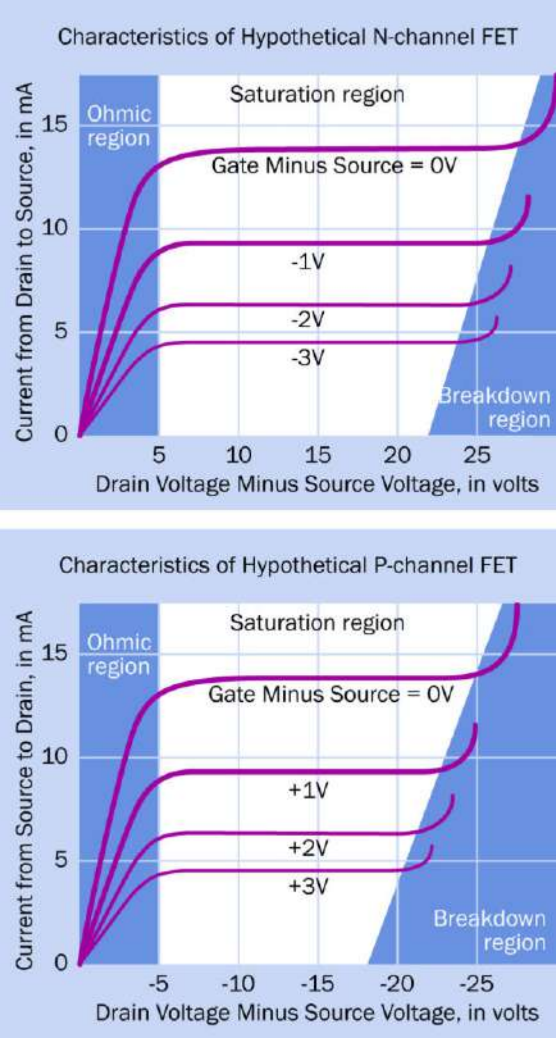

Figure 29-7. *The top graph shows current passing through the channel of an N-channel JFET, depending on gate voltage and source voltage. The lower graph is for a P-channel JFET.*

100,000 gigaohms and reducing gate current essentially to zero. The high gate impedance of a MOSFET allows it to be connected directly to the output of a digital integrated circuit. The layer of silicon dioxide is a *dielectric*, meaning that a field appled to one side creates an opposite field on the other side. The gate attached to the surface of the layer functions in the same way as one plate of a capacitor.

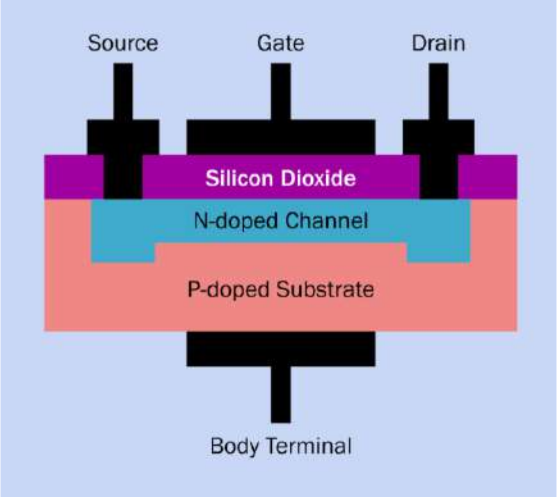

Figure 29-8. *Simplified diagram of an N-channel MOSFET. The thickness of the silicon dioxide layer has been greatly exaggerated for clarity. The black terminals are metallic.*

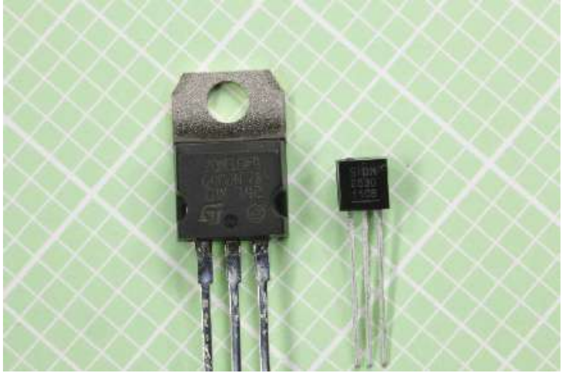

Figure 29-9. *Two MOSFETs. At left, the TO-220 package claims a drain current of up to 65A continuous, and a drain-to-source breakdown voltage 100V. At right, the smaller package offers a drain current of 175mA continuous, and a drain-to-source breakdown voltage of 300V.*

The silicon dioxide also has the highly desirable property of insulating the gate from the channel, thus preventing unwanted reverse current. In a JFET, which lacks a dielectric layer, if source voltage is allowed to rise more than about 0.6V higher than gate voltage, the direct internal connection between gate and channel allows negative charges to flow freely from source to gate, and as the internal resistance will be very low, the resulting current can be destructive. This is why the JFET must always be reverse-biased.

A MOSFET is freed from these restrictions, and the gate voltage can be higher or lower than the source voltage. This property enables an N-channel MOSFET to be designed not only as a depletion device, but alternatively as an *enhancement device*, which is "normally off" and can be switched on by being forward-biased. The primary difference is the extent to which the channel in the MOSFET is N-doped with charge carriers, and therefore will or will not conduct without some help from the gate bias.

In a depletion device, the channel conducts, but applying negative voltage to the gate can pinch off the current.

In an enhancement device, the channel does not conduct, but applying positive voltage to the gate can make it start to do so.

In either case, a shift of bias from negative to positive encourages channel conduction; the depletion and enhancement versions simply start from different points.

This is clarified in Figure 29-10. The vertical (logarithmic) scale suggests the current being conducted through the channel of the MOSFET, while the green curve describes the behavior of a depletion version of the device. Where this curve crosses the center line representing 0 volts bias, the channel is naturally conductive, like a JFET. Moving left down the curve, as reverse bias is applied (shown on the horizontal axis), the component becomes less conductive until finally its conductivity reaches zero.

Meanwhile on the same graph, the orange curve represents an enhancement MOSFET, which is nonconductive at 0 volts bias. As forward bias increases, the current also increases—similar to a bipolar transistor.

To make things more confusing, a MOSFET, like a JFET, can have a P-doped channel; and once again it can function in depletion or enhancement mode. The behavior of this variant is shown in Figure 29-11. As before, the green curve shows the behavior of a depletion MOSFET, while the orange curve refers to the enhancement version. The horizontal axis now shows the voltage difference between the gate and the drain terminal. The depletion component is naturally conductive at zero bias, until the gate voltage increases above the drain voltage, pinching off the current flow. The enhancement component is not conductive until reverse bias is applied.

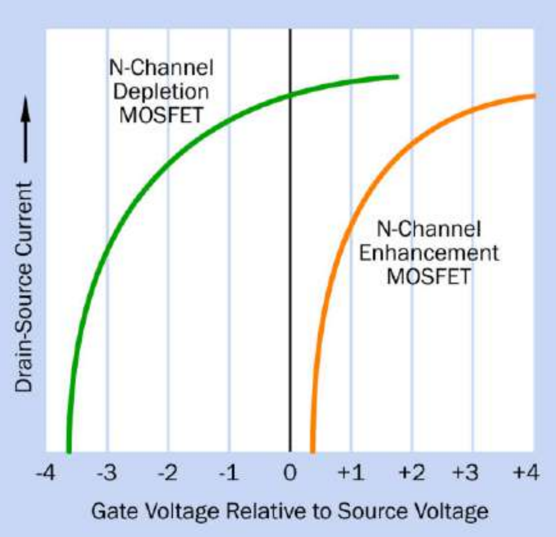

Figure 29-10. *The current conduction of depletion and enhancement N-channel MOSFETs. See text for details. (Influenced by The Art of Electronics by Horowitz and Hill.)*

Figure 29-12 shows schematic symbols that represent depletion MOSFETs. The two symbols on the left are functionally identical, representing N-channel versions, while the two symbols on the right represent P-channel versions. As in the case of JFETs, the letter "S" should be (but often is not) added to the symmetrical versions of the symbols, to clarify which is the source terminal. The left-pointing arrow identifies the components as N-channel, while in the symbols on the right, the right-pointing arrows indicate P-channel MOS-

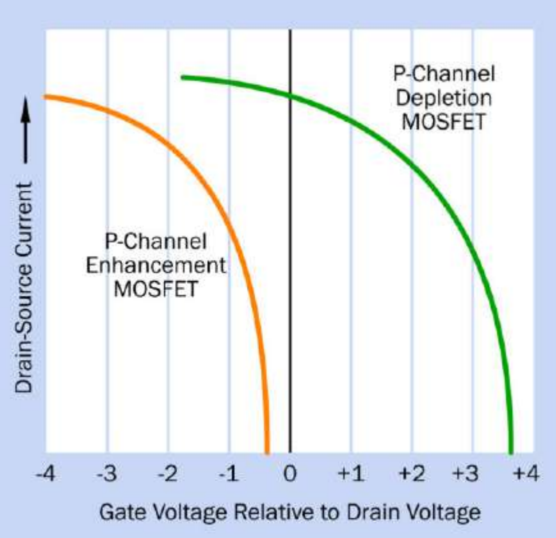

Figure 29-11. *The current conduction of depletion and enhancement P-channel MOSFETs. See text for details.*

FETs. The gap between the two vertical lines in each symbol suggests the silicon dioxide dielectric. The right-hand vertical line represents the channel.

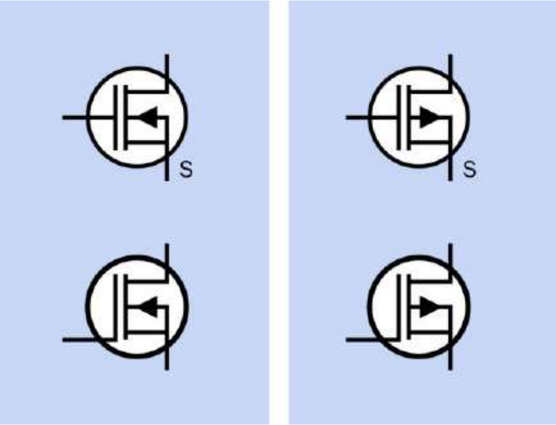

Figure 29-12. *Schematic symbols for depletion MOSFETs. These function similarly to JFETs. The two symbols on the left are functionally identical, and represent N-channel depletion MOSFETs. The two symbols on the right are both widely used to represent P-channel depletion MOSFETs.*

For enhancement MOSFETs, a slightly different symbol uses a broken line between the source and drain (as shown in Figure 29-13) to remind us that these components are "normally off" when zero-biased, instead of "normally on." Here again a left-pointing arrow represents an N-channel MOSFET, while a right-pointing arrow represents a P-channel MOSFET.

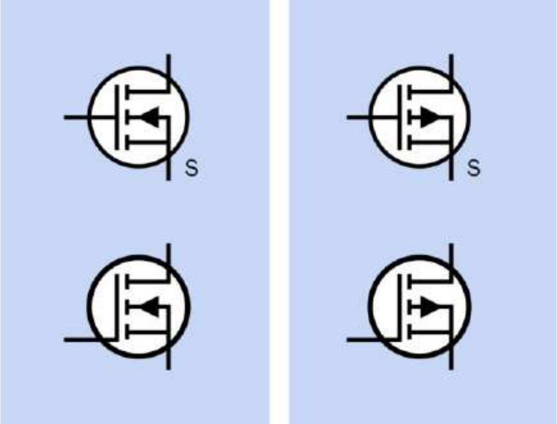

Figure 29-13. *Schematic symbols for enhancement MOSFETs. The two on the left are functionally identical, and represent N-channel enhancement MOSFETs. The two on the right represent P-channel enhancement MOSFETs.*

Because there is so much room for confusion regarding MOSFETs, a summary is presented in Figure 29-14 and Figure 29-15. In these figures, the relevant parts of each schematic symbol are shown disassembled alongside text explaining their meaning. Either of the symbols in Figure 29-14 can be superimposed on either of the symbols in Figure 29-15, to combine their functions. So, for instance, if the upper symbol in Figure 29-14 is superimposed on the lower symbol in Figure 29-15, we get an N-channel MOSFET of the enhancement type.

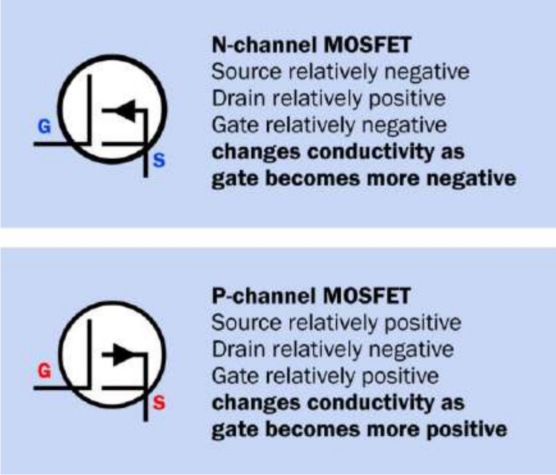
G
S
N-channel MOSFET
Source relatively negative
Drain relatively positive
Gate relatively negative
changes conductivity as gate becomes more negative
G
S
P-channel MOSFET
Source relatively positive
Drain relatively negative
Gate relatively positive
changes conductivity as gate becomes more positive

Figure 29-14. *Either of the two symbols can be combined with either of the two symbols in the next figure, to create one of the four symbols for a MOSFET. See text for details.*

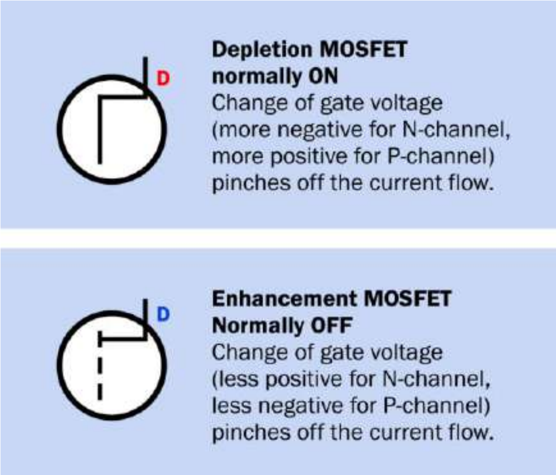
D
Depletion MOSFET normally ON
Change of gate voltage (more negative for N-channel, more positive for P-channel) pinches off the current flow.
D
Enhancement MOSFET Normally OFF
Change of gate voltage (less positive for N-channel, less negative for P-channel) pinches off the current flow.

Figure 29-15. *Either of the two symbols can be combined with either of the two symbols from the previous figure, to create one of the four symbols for a MOSFET. See text for details.*

In an additional attempt to clarify MOSFET behavior, four graphs are provided in Figure 29-16, Figure 29-17, Figure 29-18, and Figure 29-19. Like JFETs, MOSFETs have an initial ohmic region, followed by a saturation region where current flows relatively freely through the device. The gate-to-source voltage will determine how much flow is permitted. However, it is important to pay close attention to the graph scales, which differ for each of the four types of MOSFET.

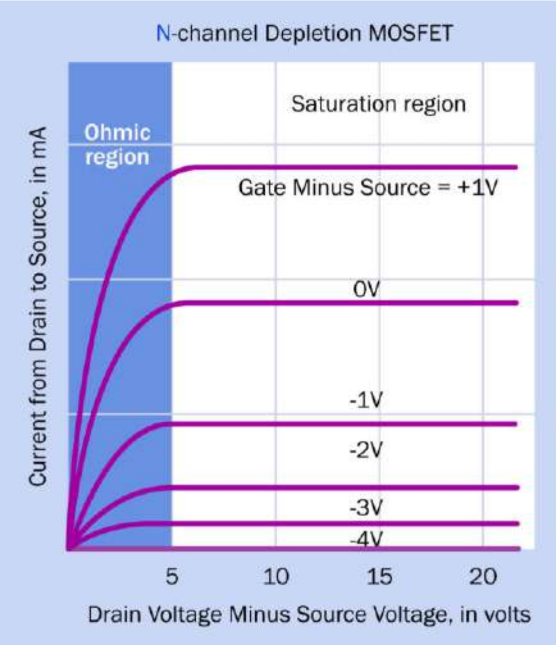

Figure 29-16. *Current flow through a depletion-type, N-channel MOSFET.*

In all of these graphs, a bias voltage exists, which allows zero current to flow (represented by the graph line superimposed on the horizontal axis). In other words, the MOSFET can operate as a switch. The actual voltages where this occurs will vary with the particular component under consideration.

The N-channel, enhancement-type MOSFET is especially useful as a switch because in its normally-off state (with zero bias) it presents a very high resistance to current flow. It requires a relatively low positive voltage at the gate, and effectively no gate current, to begin conducting conventional current from its drain terminal to its source terminal. Thus it can be driven directly by typical 5-volt logic chips.

Depletion-type MOSFETs are now less commonly used than the enhancement-type.

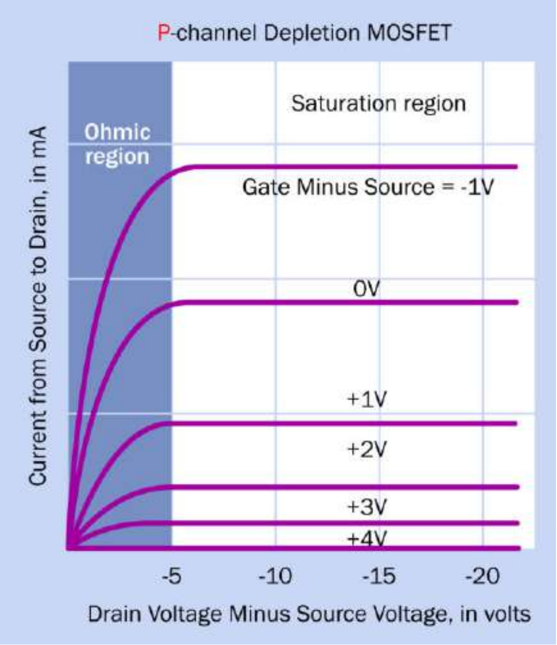

Figure 29-17. *Current flow through a depletion-type, P-channel MOSFET.*

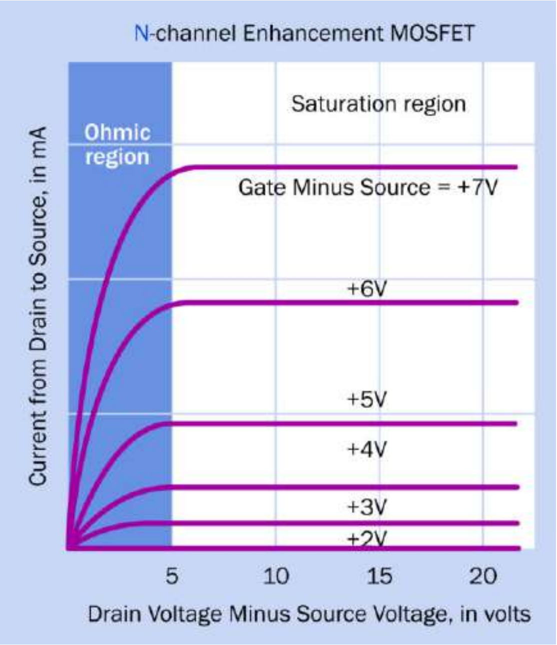

Figure 29-18. *Current flow through an enhancement-type, N-channel MOSFET.*

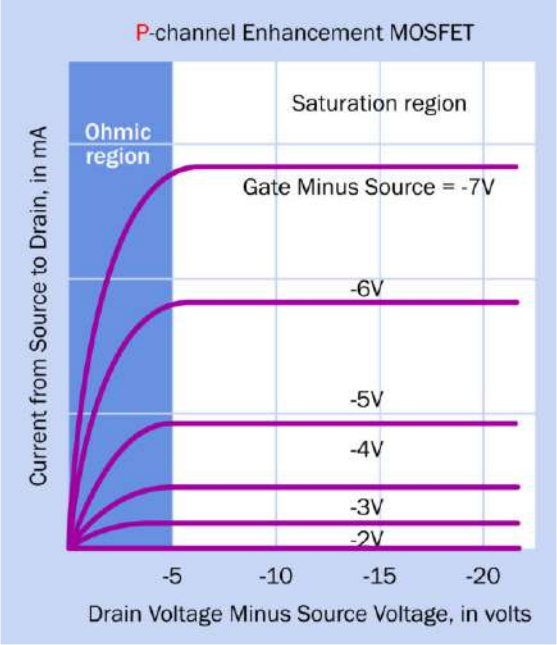

Figure 29-19. *Current flow through an enhancement-type, P-channel MOSFET.*

The Substrate Connection

Up to this point, nothing has been said about a fourth connection available on many MOSFETs, known as the *body terminal*. This is connected to the substrate on which the rest of the component is mounted, and acts as a diode junction with the channel. It is typically shorted to the source terminal, and in fact this is indicated by the schematic symbols that have been used so far. It is possible, however, to use the body terminal to shift the threshold gate voltage of the MOSFET, either by making the body terminal more negative than the source terminal (in an N-channel MOSFET) or more positive (in a P-channel MOSFET). Variants of the MOSFET schematic symbols showing the body terminal are shown in Figure 29-20 (for depletion MOSFETS) and Figure 29-21 (for enhancement MOSFETS).

A detailed discussion of the use of the body terminal to adjust characteristics of the gate is beyond the scope of this encyclopedia.

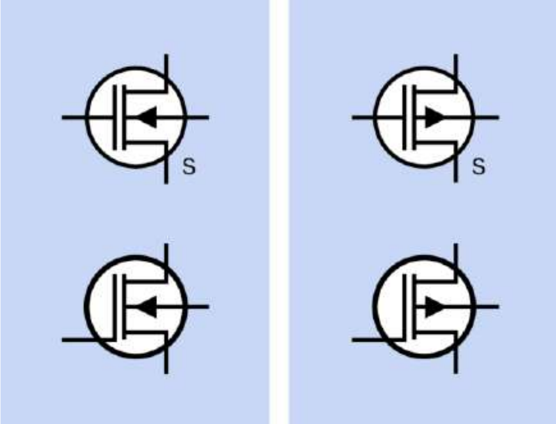

Figure 29-20. *Schematic symbol variants for depletion MOSFETs, showing the body terminal separately accessible instead of being tied to the source terminal.*

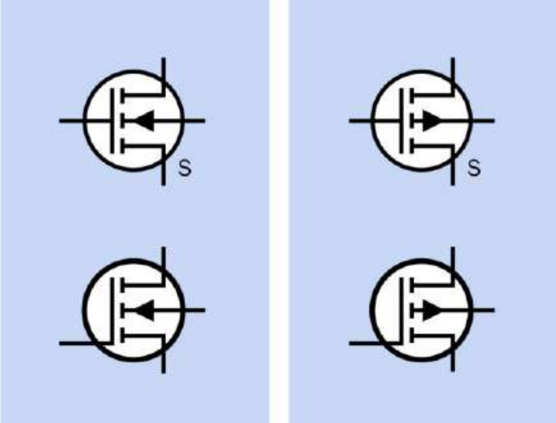

Figure 29-21. *Schematic symbol variants for enhancement MOSFETs, showing the body terminal separately accessible instead of being tied to the source terminal.*

Variants

A few FET variants exist in addition to the two previously discussed.

MESFET

The acronym stands for *MEtal-Semiconductor Field Effect Transistor*. This FET variant is fabricated from gallium arsenide and is used primarily in radio frequency amplification, which is outside the scope of this encyclopedia.

V-Channel MOSFET

Whereas most FET devices are capable of handling only small currents, the *V-channel MOSFET* (which is often abbreviated as a *VMOS FET* and has a V-shaped channel as its name implies) is capable of sustained currents of at least 50A and voltages as high as 1,000V. It is able to pass the high current because its channel resistance is well under 1Ω. These devices, commonly referred to as *power MOSFETs*, are available from all primary semiconductor manufacturers and are commonly used in switching power supplies.

Trench MOS

The TrenchMOS or Trenchgate MOS is a MOSFET variant that encourages current to flow vertically rather than horizontally, and includes other innovations that enable an even lower channel resistance, allowing high currents with minimal heat generation. This device is finding applications in the automobile industry as a replacement for electromechanical **relays**.

Values

The maximum values for JFETs, commonly found listed in datasheets, will specify V_{ds} (the drain-source voltage, meaning the potential difference between drain and source); V_{dg} (the drain-gate voltage, meaning the potential difference between drain and gate); V_{gsr} (the reverse gate-source voltage); gate current; and total device dissipation in mW. Note that the voltage differences are relative, not absolute. Thus a voltage of 50V on the drain and 25V on the source might be acceptable in a component with a V_{ds} of 25V. Similarly, while a JFET's "pinch-off" effect begins as the gate becomes "more negative" than the source, this can be achieved if, for example, the source has a potential of 6V and the gate has a potential of 3V.

JFETs and MOSFETs designed for low-current switching applications have a typical channel resistance of just a few ohms, and a maximum switching speed around 10Mhz.

The datasheet for a MOSFET will typically include values such as gate threshold voltage, which may be abbreviated V_{gs} (or V_{th}) and establishes the relative voltage at which the gate starts to play an active role; and the maximum on-state drain current, which may be abbreviated $I_{d(on)}$ and establishes the maximum limiting current (usually at 25 degree Centigrade) between source and gate.

How to Use it

The combination of a very high gate impedance, very low noise, very low quiescent power consumption in its off state, and very fast switching capability makes the MOSFET suitable for many applications.

P-Channel Disadvantage

P-channel MOSFETs are generally less popular than N-channel MOSFETS because of the higher resistivity of P-type silicon, resulting from its lower carrier mobility, putting it at a relative disadvantage.

Bipolar Substitution

In many instances, an appropriate enhancement-type MOSFET can be substituted for a bipolar transistor with better results (lower noise, faster action, much higher impedance, and probably less power dissipation).

Amplifier Front Ends

While MOSFETs are well-suited for use in the front end of an audio amplifier, chips containing MOSFETs are now available for this specific purpose.

Voltage-Controlled Resistor

A simple voltage-controlled resistor can be built around a JFET or MOSFET, so long as its performance remains limited to the linear or ohmic region.

Compatibility with Digital Devices

A JFET may commonly use power supplies in the range of 25VDC. However, it can accept the high/low output from a 5V digital device to control its gate. A 4.7K pullup resistor is an appropriate value to be used if the FET is to be used in conjunction with a TTL digital chip that may have a voltage swing of only approximately 2.5V between its low and high thresholds.

What Can Go Wrong

Static Electricity

Because the gate of a MOSFET is insulated from the rest of the component, and functions much like a plate of a capacitor, it is especially likely to accumulate static electricity. This static charge may then discharge itself into the body of the component, destroying it. A MOSFET is particularly vulnerable to electrostatic discharge because its oxide layer is so thin. Special care should be taken either when handling the component, or when it is in use. Always touch a grounded object or wear a grounded wrist band when handling MOSFETs, and be sure that any circuit using MOSFETs includes appropriate protection from static and voltage spikes.

A MOSFET should not be inserted or removed while the circuit in which it performs is switched on or contains residual voltage from undischarged capacitors.

Heat

Failure because of overheating is of special concern when using power MOSFETs. A Vishay Application Note ("Current Power Rating of Power Semiconductors") suggests that this kind of component is unlikely to operate at less than 90 degrees Centigrade in real-world conditions, yet the power handling capability listed in a datasheet usually assumes an industry standard of 25 degrees Centigrade.

On the other hand, ratings for continuous power are of little relevance to switching devices that have duty cycles well below 100%. Other factors also play a part, such as the possibility of power surges, the switching frequency, and the integrity of the connection between the component and its heat sink. The heat sink itself creates uncertainty by tending to average the temperature of the component, and of course there is no simple way to know the actual junction temperature, moment by moment, inside a MOSFET.

Bearing in mind the accumulation of unknown factors, power MOSFETs should be chosen on an extremely conservative basis. According to a tutorial in the *EE Times*, actual current switched by a MOSFET should be less than half of its rated current at 25 degrees, while one-fourth to one-third are common. Figure 29-22 shows the real-world recommended maximum drain current at various temperatures. Exceeding this recommendation can create additional heat, which cannot be dissipated, leading to further accumulation of heat, and a thermal runaway condition, causing eventual failure of the component.

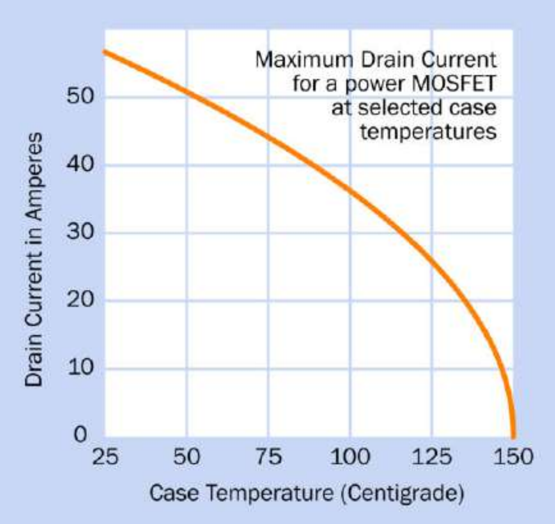

Figure 29-22. *Maximum advised drain current through a power MOSFET, related to case temperature of the component. Derived from EE Times Power MOSFET Tutorial.*

Wrong Bias

As previously noted, applying forward bias to a JFET can result in the junction between the gate and the source starting to behave like a forward-biased diode, when the voltage at the gate is greater than the voltage at the source by approximately 0.6V or more (in an N-channel JFET). The junction will present relatively little resistance, encouraging excessive current and destructive consequences. It is important to design devices that allow user input in such a way that user error can never result in this eventuality.

Schematic Symbols

A

This section contains a compilation of schematic symbols for components that have been described in this volume. They are sequenced primarily in alphabetical order, as this section is intended for use as an index. However, symbols that have a strong similarity are grouped together; thus **potentiometer** is found adjacent to **resistor**, and all types of transistors are in the same group.

The symbol variants shown in each blue rectangle are functionally identical.

Where a component has mandatory polarity or is commonly used with a certain polarity, a red plus sign (+) has been added for guidance. This sign is not part of the symbol. In the case of polarized capacitors, where a plus sign is normally shown (or should be shown) with the symbol, the plus sign is a part of the symbol and appears in black.

This is not intended to be an exhaustive compilation of symbol variants. Some uncommon ones may not be here. However, the list should be sufficient to enable identification of components in this volume.

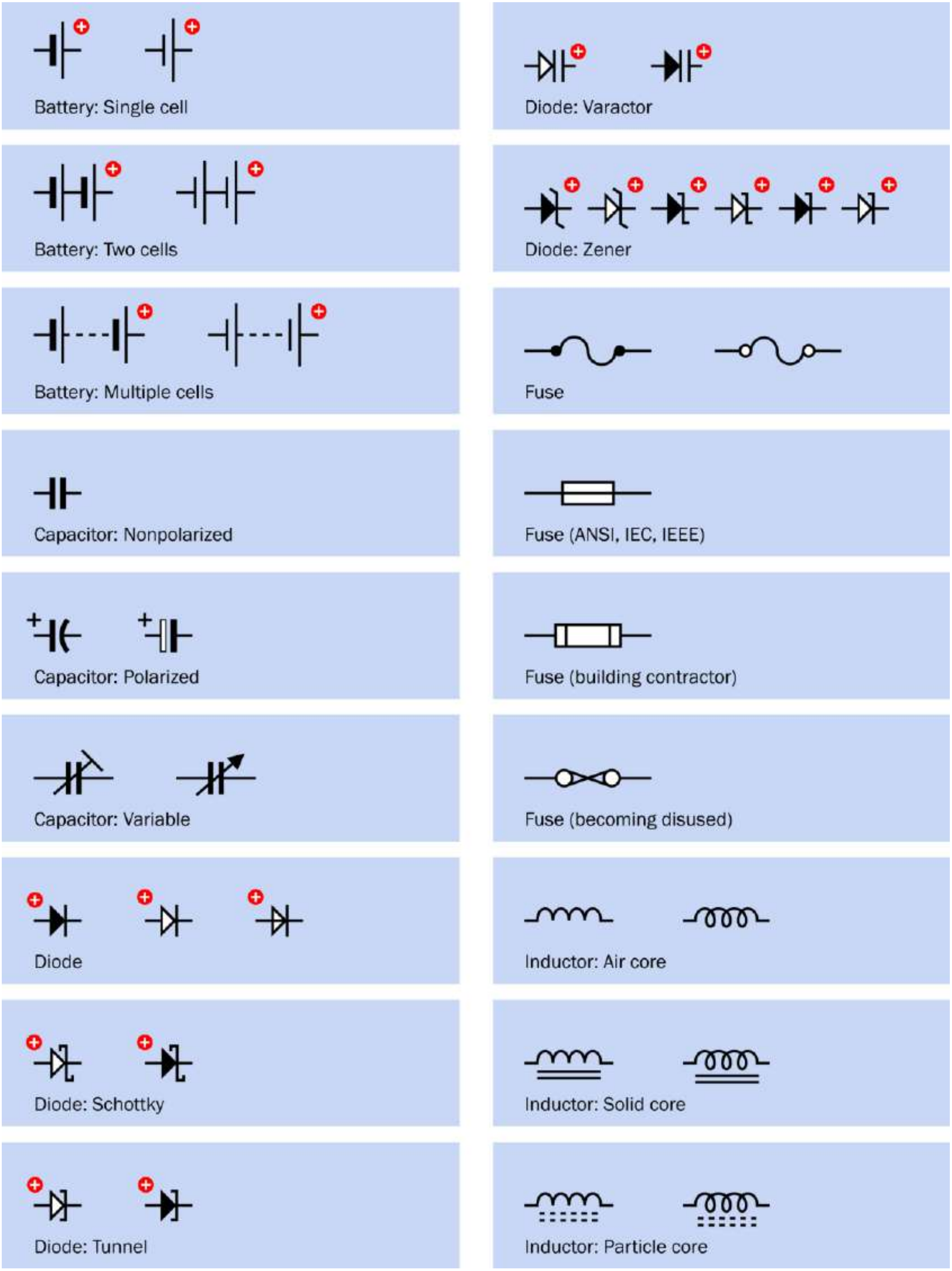
Battery: Single cell
Battery: Two cells
Battery: Multiple cells
Capacitor: Nonpolarized
+
+
Capacitor: Polarized
Capacitor: Variable
Diode
Diode: Schottky
Diode: Tunnel
Diode: Varactor
Diode: Zener
Fuse
Fuse (ANSI, IEC, IEEE)
Fuse (building contractor)
Fuse (becoming disused)
Inductor: Air core
Inductor: Solid core
Inductor: Particle core

Figure A-1. *Schematic symbols*

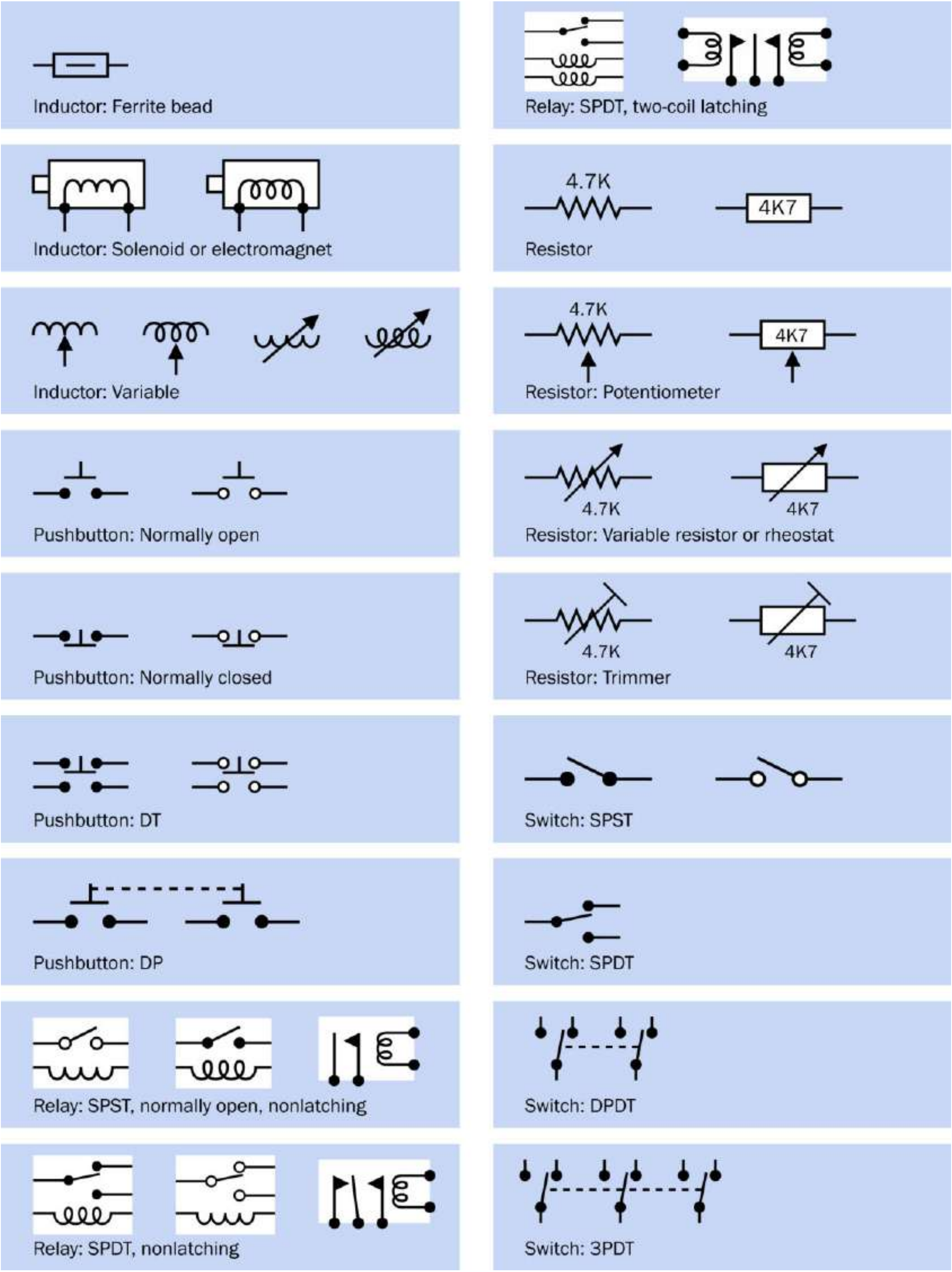
Inductor: Ferrite bead
Inductor: Solenoid or electromagnet
Inductor: Variable
Pushbutton: Normally open
Pushbutton: Normally closed
Pushbutton: DT
Pushbutton: DP
Relay: SPST, normally open, nonlatching
Relay: SPDT, nonlatching
Relay: SPDT, two-coil latching
4.7K
4K7
Resistor
4.7K
4K7
Resistor: Potentiometer
4.7K
4K7
Resistor: Variable resistor or rheostat
4.7K
4K7
Resistor: Trimmer
Switch: SPST
Switch: SPDT
Switch: DPDT
Switch: 3PDT

Figure A-2. *Schematic symbols, continued*

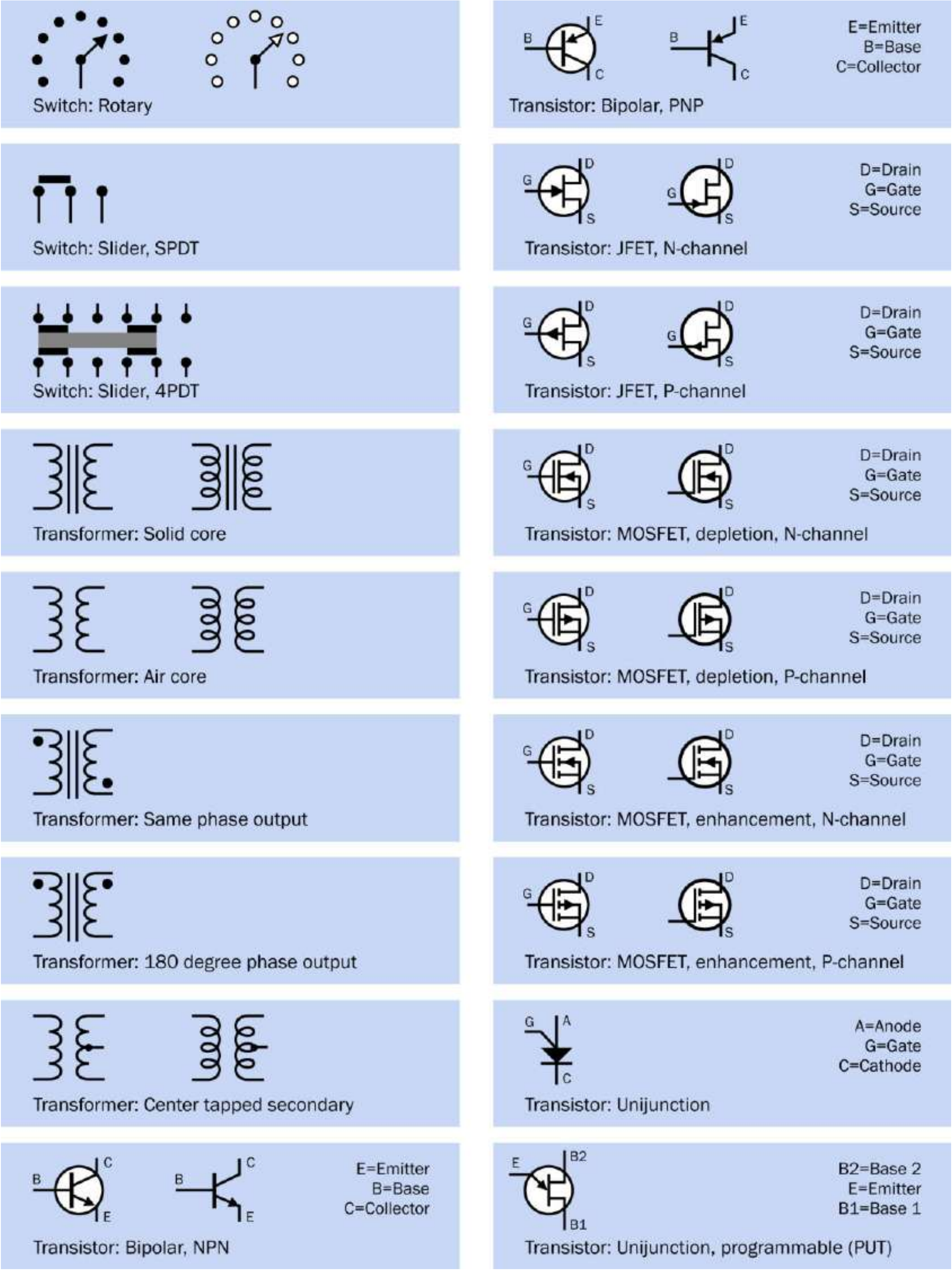
E
B
C
E
B
C
E=Emitter
B=Base
C=Collector
D
G
S
D=Drain
G=Gate
S=Source
G
A
C
A=Anode
G=Gate
C=Cathode
C
B
E
E=Emitter
B=Base
C=Collector
E
B2
B1
B2=Base 2
E=Emitter
B1=Base 1

Switch: Rotary

Transistor: Bipolar, PNP

Switch: Slider, SPDT

Transistor: JFET, N-channel

Switch: Slider, 4PDT

Transistor: JFET, P-channel

Transformer: Solid core

Transistor: MOSFET, depletion, N-channel

Transformer: Air core

Transistor: MOSFET, depletion, P-channel

Transformer: Same phase output

Transistor: MOSFET, enhancement, N-channel

Transformer: 180 degree phase output

Transistor: MOSFET, enhancement, P-channel

Transformer: Center tapped secondary

Transistor: Unijunction

Transistor: Bipolar, NPN

Transistor: Unijunction, programmable (PUT)

Figure A-3. *Schematic symbols, continued*

How to Use This Book

This is the second of three volumes. Its purpose is to provide an overview of the most commonly used electronic components, for reference by students, engineers, hobbyists, and instructors. While you can find much of this information dispersed among datasheets, introductory books, websites, and technical resources maintained by manufacturers, the *Encyclopedia of Electronic Components* gathers all the relevant facts in one place, properly organized and verified, including details that may be hard to find elsewhere. Each entry includes typical applications, possible substitutions, cross-references to similar devices, sample schematics, and a list of common problems and errors.

You can find a more detailed rationale for this encyclopedia in the Preface to Volume 1.

Volume Contents

Practical considerations influenced the decision to divide this encyclopedia into three volumes. Each deals with broad subject areas as follows.

Volume 1

Power; electromagnetic devices; discrete semiconductors

The *power* category includes sources of electricity and methods to distribute, store, interrupt, convert, and regulate power. The *electromagnetic devices* category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the primary types of diodes and transistors. A contents listing for Volume 1 appears in Figure P-1.

Volume 2

Thyristors (SCRs, diacs, and triacs); integrated circuits; light sources, indicators, and displays; and sound sources

Integrated circuits are divided into analog and digital components. *Light sources, indicators, and displays* are divided into reflective displays, single sources of light, and displays that emit light. *Sound sources* are divided into those that create sound, and those that reproduce sound. A contents listing for Volume 2 appears in Figure P-2.

Volume 3

Sensing devices

The field of sensors has become so extensive, they easily merit a volume to themselves. *Sensing devices* include those that detect light, sound, heat, motion, pressure, gas, humidity, orientation, electricity, proximity, force, and radiation.

At the time of writing, Volume 3 is still in preparation, while Volume 1 is complete and is available in a variety of formats.

Primary Category	Secondary Category	Component Type
power	source	battery
	connection	jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electromagnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
		stepper motor
discrete semiconductor	single junction	diode
		unijunction transistor
	multi-junction	bipolar transistor
		field-effect transistor

Figure P-1. *The subject-oriented organization of categories and entries in Volume 1.*

Primary Category	Secondary Category	Component Type
discrete semiconductor	thyristor	SCR
		diac
		triac
integrated circuit	analog	solid-state relay
		optocoupler
		comparator
		op-amp
		digital potentiometer
		timer
	digital	logic gate
		flip-flop
		shift register
		counter
		encoder
		decoder
		multiplexer
light source, indicator or display	reflective	LCD
	single source	incandescent lamp
		neon bulb
		fluorescent light
		laser
		LED indicator
		LED area lighting
	multi-source or panel	LED display
		vacuum-fluorescent
		electroluminescence
sound source	audio alert	transducer
		audio indicator
	reproducer	headphone
		speaker

Figure P-2. *The subject-oriented organization of categories and entries in Volume 2.*

Organization

Reference versus Tutorial

As its title suggests, this is a reference book, not a tutorial. A tutorial begins with elementary concepts and builds sequentially toward concepts that are more advanced. A reference book assumes that you may dip into the text at any point, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will find some repetition, as each entry is intended to be self-sufficient, requiring minimal reference to other entries.

My books *Make: Electronics* and *Make: More Electronics* follow a tutorial approach. They don't go into as much depth as this Encyclopedia, because a tutorial inevitably allocates a lot of space to step-by-step explanations and instructions.

Theory and Practice

This book is oriented toward practicality rather than theory. I assume that the reader mostly wants to know how to use electronic components, rather than why they work the way they do. Consequently, I have not included proofs of formulae or definitions rooted in electrical theory. Units are defined only to the extent necessary to avoid confusion.

Many books on electronics theory already exist, if theory is of interest to you.

Entries

This encyclopedia is divided into entries, each entry being devoted to one broad type of component. Two rules determine whether a component has an entry all to itself, or is subsumed into another entry:

Rule 1

> A component merits its own entry if it is (a) widely used, or (b) not so widely used but has a unique identity and maybe some historical status. The **bipolar transistor** entry is an example of a widely used component, whereas the **unijunction transistor** entry is an example of a not so widely used component with a unique identity.

Rule 2

> A component does not merit its own entry if it is (a) seldom used, or (b) very similar in function to another component that is more widely used. For example, a *rheostat* is subsumed into the **potentiometer** section, while *silicon diode*, *Zener diode*, and *germanium diode* are combined together in the **diode** entry.

Inevitably, these guidelines required judgment calls which in some cases may seem arbitrary. My ultimate decision was based on where I would expect to find a component if I was looking for it myself.

Subject Paths

Entries are not organized alphabetically. They are grouped by subject, in much the same way that books in the nonfiction section of some libraries are organized by the Dewey Decimal System. This is convenient if you don't know exactly what you are looking for, or if you don't know all the options that may be available to perform a task that you have in mind.

Each primary category is divided into subcategories, and the subcategories are divided into component types. This hierarchy is shown in Figure P-2. It is also apparent when you look at the top of the first page of each entry, where you will find the path that leads to it. The **diac** entry, for instance, is headed with this path:

discrete semiconductor > thyristor > diac

Any classification scheme will run into exceptions. You can buy a chip containing a *resistor array*, for instance. Technically, this is an *analog integrated circuit*, but a decision was made to put it in the **resistor** section of Volume 1, because it can be directly substituted for a group of resistors.

Some components have hybrid functions. A **multiplexer**, for instance, may pass analog signals and may have "analog" in its name. However, it is digitally controlled and is mostly used in conjunction with other digital integrated circuits. This seemed to justify placing it in the digital category.

Inclusions and Exclusions

There is also the question of what is, and is not, a component. Is wire a component? Not for the purposes of this encyclopedia. How about a **DC-DC converter**? Because converters are now sold in small packages by component suppliers, they are included in Volume 1 as components.

Many similar decisions had to be made on a case-by-case basis. Some readers will disagree with the outcome, but reconciling all the disagreements would have been impossible. The best I could do was to create a book which is organized in the way that would suit me best if I were using it myself.

Typographical Conventions

Within each entry, **bold type** is used for the first occurrence of the name of a component that has its own entry elsewhere. Other important electronics terms or component names may be presented in *italics*.

The names of components, and the categories to which they belong, are all set in lowercase type, except where a term is normally capitalized because it is an acronym or a trademark. The term *Trimpot*, for instance, is trademarked by Bourns, but *trimmer* is not. **LED** is an acronym, but *cap* (abbreviation for **capacitor**) is not.

The European convention for representing fractional component values eliminates decimal points. Thus, values such as 3.3K and 4.7K are expressed as 3K3 and 4K7. This style has not been adopted to a significant degree in the United States, and is not used in this encyclopedia.

In mathematical formulae, I have used the style that is common in programming languages. The * (asterisk) is used as a multiplication symbol, while the / (forward slash) is used as a division symbol. Where some terms are in parentheses, they must be dealt with first. Where parentheses are inside parentheses, the innermost ones must be dealt with first. So, in this example:

```
A = 30 / (7 + (4 * 2) )
```

You would begin by multiplying 4 times 2, to get 8; then add 7, to get 15; then divide that into 30, to get the value for A, which is 2.

Visual Conventions

Figure P-3 shows the conventions that are used in the schematics in this book. A black dot always indicates a connection, except that to minimize ambiguity, the configuration at top right is avoided, and the configuration at top center is used instead. Conductors that cross each other without a black dot do not make a connection. The styles at bottom right are sometimes seen elsewhere, but are not used here.

All the schematics are formatted with pale blue backgrounds. This enables components such as switches, transistors, and LEDs to be highlighted in white, drawing attention to them and clarifying the boundary of the component. The white areas have no other meaning.

Photographic Backgrounds

All photographs of components include a background grid that is divided into squares measuring 0.1". Although the grid is virtual, it is equivalent in scale to physical graph paper placed immediately behind the component. If the component is photographed at an angle, the grid may be reproduced at a similar angle, creating perspective on the squares.

Background colors in photographs were chosen for contrast with the colors of the components, or for visual variety. They have no other significance.

Component Availability

Because there is no way of knowing if a component may have a long production run, this encyclopedia is cautious about listing specific part numbers. To find a specific part that has a narrow function, searching the websites maintained by suppliers will be necessary. The following suppliers were checked frequently during the preparation of the book:

- Mouser Electronics
- Jameco Electronics

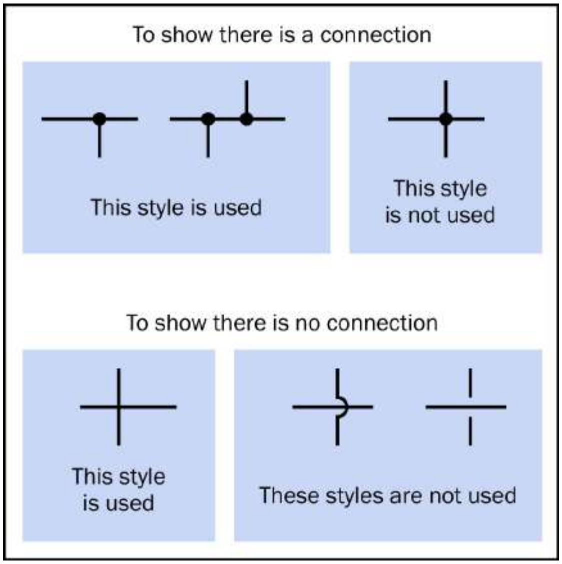
To show there is a connection
This style is used
This style is not used
To show there is no connection
This style is used
These styles are not used

Figure P-3. *Visual conventions that are used in the schematics in this book.*

When seeking obsolete parts, or those that are nearing the end of their commercial life, eBay can be very useful.

Issues and Errata

If you believe you have found an error in this book, you will find guidance on how to report it here: *http://bit.ly/eec_v2_errata*.

Before posting your own erratum, please check those that have been submitted previously, to see if someone else already reported it.

I value and encourage reader feedback. However, before you post feedback publicly to a site such as Amazon, I have a request. Please be aware of the power that you have as a reader, and use it fairly. A single negative review can create a bigger effect than you may realize. It can certainly outweigh half-a-dozen positive reviews. If you feel you have not received a prompt or adequate response from the O'Reilly errata site mentioned here, you can email me personally at:

make.electronics@gmail.com

I check that address irregularly—sometimes only once in a couple of weeks. But I do answer all messages.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of plans and pricing for enterprise, government, education, and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like Maker Media, O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and hundreds more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

Make:
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

Make: unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. Make: celebrates your right to tweak, hack, and bend any technology to your will. The Make: audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our educational system—our entire world. This is much more than an audience, it's a worldwide movement that Make: is leading—we call it the Maker Movement.

For more information about Make:, visit us online:

Make: magazine: *http://makezine.com/magazine/*
Maker Faire: *http://makerfaire.com*
Makezine.com: *http://makezine.com*
Maker Shed: *http://makershed.com/*

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at: *http://bit.ly/encyclopedia_of_electronic_components_v2*.

Acknowledgments

Any reference work draws inspiration from many sources. Datasheets and tutorials maintained by component manufacturers were considered the most trustworthy sources of information online. In addition, component retailers, college texts, crowd-sourced reference works, and hobbyist sites were used. The following books provided useful information:

Boylestad, Robert L. and Nashelsky, Louis: *Electronic Devices and Circuit Theory*, 9th edition. Pearson Education, 2006.

Braga, Newton C.: *CMOS Sourcebook*. Sams Technical Publishing, 2001.

Hoenig, Stuart A.: *How to Build and Use Electronic Devices Without Frustration, Panic, Mountains of Money, or an Engineering Degree*, 2nd edition. Little, Brown, 1980.

Horn, Delton T.: *Electronic Components*. Tab Books, 1992.

Horn, Delton T.: *Electronics Theory*, 4th edition. Tab Books, 1994.

Horowitz, Paul and Hill, Winfield: *The Art of Electronics*, 2nd edition. Cambridge University Press, 1989.

Ibrahim, Dogan: *Using LEDs, LCDs, and GLCDs in Microcontroller Projects*. John Wiley & Sons, 2012.

Kumar, A. Anand: *Fundamentals of Digital Circuits*, 2nd edition. PHI Learning, 2009.

Lancaster, Don: *TTL Cookbook*. Howard W. Sams & Co, 1974.

Lenk, Ron and Lenk, Carol: *Practical Lighting Design with LEDs*. John Wiley & Sons, 2011.

Lowe, Doug: *Electronics All-in-One for Dummies*. John Wiley & Sons, 2012.

Mims III, Forrest M.: *Getting Started in Electronics*. Master Publishing, 2000.

Mims III, Forrest M.: *Electronic Sensor Circuits & Projects*. Master Publishing, 2007.

Mims III, Forrest M.: *Timer, Op Amp, & Optelectronic Circuits and Projects*. Master Publishing, 2007.

Predko, Mike: *123 Robotics Experiments for the Evil Genius*. McGraw-Hill, 2004.

Scherz, Paul: *Practical Electronics for Inventors*, 2nd edition. McGraw-Hill, 2007.

Williams, Tim: *The Circuit Designer's Companion*, 2nd edition. Newnes, 2005.

I also made extensive use of information on vendor sites, especially:

- Mouser Electronics
- Jameco Electronics
- All Electronics
- sparkfun
- Electronic Goldmine
- Adafruit
- Parallax, Inc.

In addition, some individuals provided special assistance. My editor, Brian Jepson, was immensely helpful in the development of this book. Philipp Marek and Steve Conklin reviewed the text for errors. My publisher demonstrated faith in my work. Kevin Kelly unwittingly influenced me with his legendary interest in "access to tools." It was Mark Frauenfelder who originally brought me back to the pleasures of building things, and Gareth Branwyn who revived my interest in electronics.

Lastly, I should mention my school friends from decades ago: Patrick Fagg, Hugh Levinson, Graham Rogers, William Edmondson, and John Witty, who helped me to feel that it was OK to be a nerd building my own audio equipment, long before the word "nerd" actually existed.

—Charles Platt, 2014

SCR

1

The acronym **SCR** is derived from *silicon-controlled rectifier*, which is a gate-triggered type of *thyristor*. A thyristor is defined here as a semiconductor having four or more alternating layers of p-type and n-type silicon. Because it predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, a thyristor is considered to be a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**), it is considered to be an integrated circuit.

Other types of thyristor are the **diac** and **triac**, each of which has its own entry.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **diac** (see Chapter 2)
- **triac** (see Chapter 3)

What It Does

In the 1920s, the *thyratron* was a gas-filled tube that functioned as a switch and a rectifier. In 1956, General Electric introduced a solid-state version of it under the name *thyristor*. In both cases, the names were derived from the thyroid gland in the human body, which controls the rate of consumption of energy. The thyratron and, subsequently, the thyristor enabled control of large flows of current.

The **SCR** (silicon-controlled rectifier) is a type of thyristor, although the two terms are often used as if they are synonymous. Text that refers loosely to a thyristor may actually be discussing an SCR, and vice versa. In this encyclopedia, the **SCR**, **diac**, and **triac** are all considered to be variant types of thyristor.

An SCR is a solid-state switch that in many instances can pass high currents at high voltages. Like a **bipolar transistor**, it is triggered by voltage applied to a gate. Unlike the transistor, it allows the flow of current to continue even when the gate voltage diminishes to zero.

How It Works

This component is designed to pass current in one direction only. It can be forced to conduct in the opposite direction if the reversed potential exceeds its *breakdown voltage*, but this mistreatment is likely to cause damage.

By comparison, the diac and triac are designed to be bidirectional.

The SCR has three leads, identified as anode, cathode, and gate. Two functionally identical versions of the schematic symbol are shown in Figure 1-1. Early versions sometimes included a circle drawn around them, but this style has become obsolete. Care must be taken to distinguish between the SCR symbol and the symbol that represents a **programmable unijunction transistor** (PUT), shown in Figure 1-2.

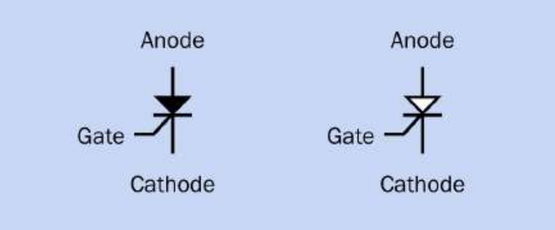

Figure 1-1. *Two functionally identical schematic symbols for an SCR (silicon-controlled rectifier). The symbol on the left is more common.*

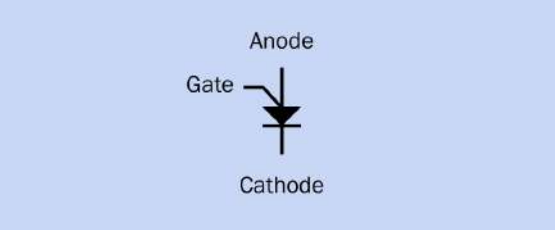

Figure 1-2. *The symbol shown here is for a programmable unijunction transistor (PUT). Care must be taken to distinguish it from the symbol for an SCR.*

Switching Behavior

When the SCR is in its passive or nonconductive state, it will block current in either direction between anode and cathode, although a very small amount of *leakage* typically occurs. When the SCR is activated by a positive voltage at the gate, current can now flow from anode to cathode, although it is still blocked from cathode to anode. When the flow reaches a level known as the *latching current*, the flow will continue even after the triggering voltage drops to zero. This behavior causes it to be known as a *regenerative* device.

If the current between anode and cathode starts to diminish while the gate voltage remains zero, the current flow will continue below the latching level until it falls below the value known as the *holding current*. The flow now ceases. Thus, the only way to end a flow of current that has been initiated through an SCR is by reducing the flow or attempting to reverse it.

Note that the self-sustaining flow is a function of current rather than voltage.

Unlike a transistor, an SCR is either “on” or “off” and does not function as a *current amplifier*. Like a diode, it is designed to conduct current in one direction; hence the term *rectifier* in its full name. When it has been triggered, the impedance between its anode and cathode is sufficiently low that heat dissipation can be managed even at high power levels.

The ability of SCRs to pass relatively large amounts of current makes them suitable for controlling the power supplied to motors and resistive heating elements. The fast switching response also enables an SCR to interrupt and abbreviate each positive phase of an AC waveform, to reduce the average power supplied. This is known as *phase control*.

SCRs are also used to provide *overvoltage protection*.

SCR packages reflect their design for a wide range of voltages and currents. Figure 1-3 shows an SCR designed for on-state current of 4A RMS (i.e., measured as the root mean square of the alternating current). Among its applications are small-engine ignition and *crowbar* overvoltage protection, so named because it shorts a power supply directly to ground, much like a crowbar being dropped across the terminals of a car battery (but hopefully with a less dramatic outcome). See Figure 1-15.

In Figure 1-4, the SCR can handle up to 800V repetitive peak off-state voltage and 55A RMS. Possible applications include AC rectification, crowbar protection, welding, and battery charging. The component in Figure 1-5 is rated for 25A and 50V repetitive peak off-state voltage. To assess the component sizes, bear in mind that the graph line spacing is 0.1”.

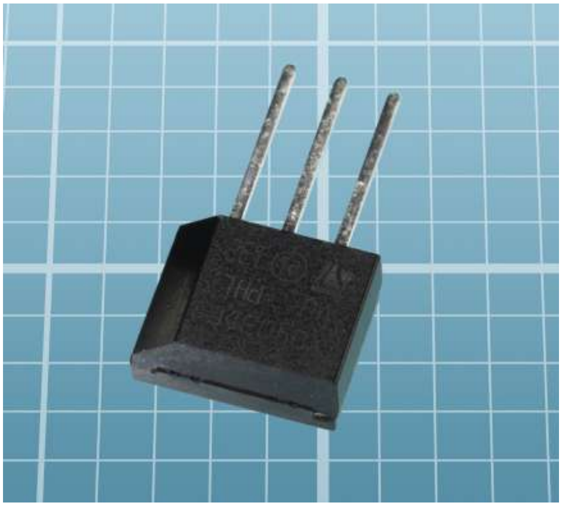

Figure 1-3. *SCR rated for 400V repetitive off-state voltage, no greater than 4A RMS.*

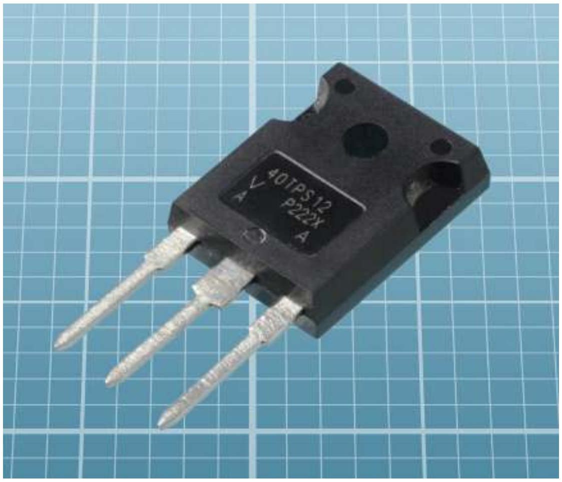

Figure 1-4. *SCR rated for 800V repetitive off-state voltage, no greater than 55A RMS.*

Internal Configuration

The function of an SCR can be imagined as being similar to that of a PNP transistor paired with an NPN transistor, as shown in Figure 1-6. In this simplified schematic, so long as zero voltage is applied to the "gate" wire, the lower (NPN) transistor remains nonconductive. Consequently, the upper (PNP) transistor cannot sink current, and this transistor also remains nonconductive.

When voltage is applied to the "gate," the lower transistor starts to sink current from the upper transistor. This switches it on. The two transistors now continue to conduct even if power to the "gate" is disconnected, because they have created a positive feedback loop.

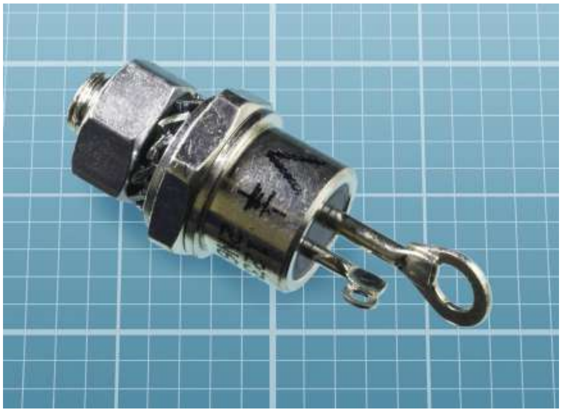

Figure 1-5. *Stud-packaged SCR rated for 50V repetitive off-state voltage, no greater than 25A RMS.*

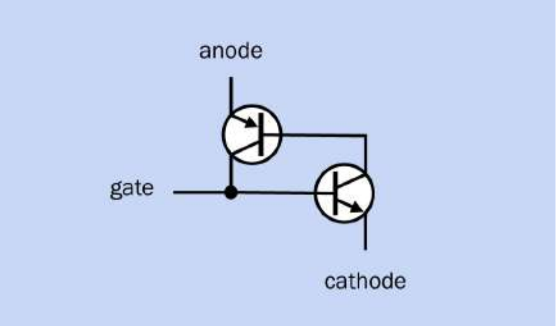

Figure 1-6. *An SCR behaves similarly to an NPN and a PNP transistor coupled together.*

Figure 1-7 shows the same two transistors in simplified form as sandwiches of p-type and n-type silicon layers (on the left), and their combination in an SCR (on the right). Although the actual configuration of silicon segments is not as simple or as linear as this diagram suggests, the SCR can be described correctly as a *PNPN device*.

An SCR is comparable with an electromagnetic **latching relay**, except that it works much faster and more reliably.

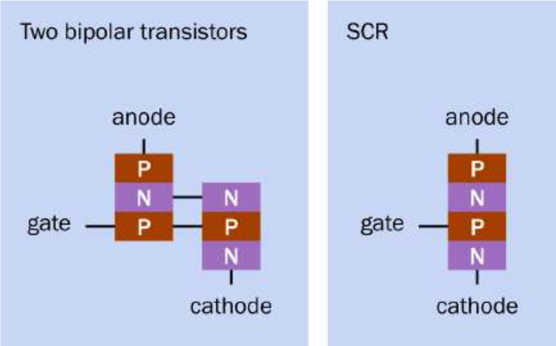

Figure 1-7. *The two transistors from the previous figure are shown here in simplified form as two stacks of p-type and n-type silicon layers. These layers are combined in an SCR, on the right.*

Breakdown and Breakover Voltage

The curves in Figure 1-8 illustrate the behavior of a hypothetical SCR, and can be compared with the curves shown for a diac in Figure 2-5 and a triac in Figure 3-10. Beginning with zero voltage applied between anode and cathode, and zero current flowing (i.e., at the center origin of the graph), if we apply a voltage at the anode that is increasingly negative relative to the cathode (i.e., we attempt to force the SCR to allow negative current flow), we see a small amount of leakage, indicated by the darker blue area (which is not drawn to scale). Finally the *breakdown voltage* is reached, at which point the negative potential overcomes the SCR and its impedance drops rapidly, allowing a surge of current to flow, probably damaging it.

Alternatively, starting once again from the center, if we apply a voltage at the anode that is increasingly positive relative to the cathode, two consequences are possible. The dashed curve assumes that there is zero voltage at the gate, and shows that some leakage occurs until the applied potential at the anode reaches the *breakover voltage*, at which point the SCR allows a large current flow, which continues even when the voltage decreases.

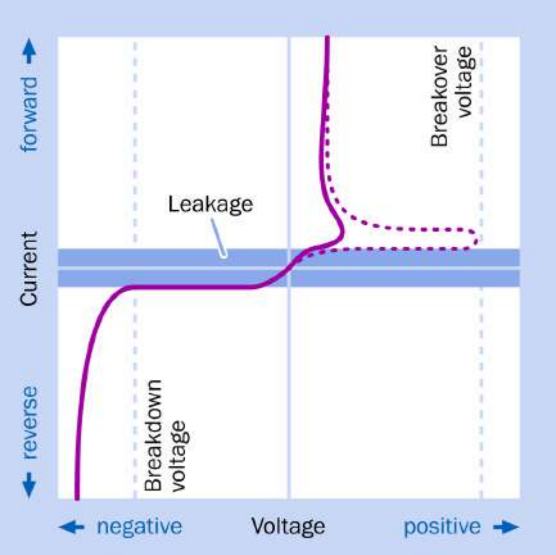

Figure 1-8. *The solid curve shows current passing between the anode and cathode of a hypothetical SCR for varying voltages, while a triggering voltage is applied to the gate. The dashed curve assumes that no triggering voltage is applied to the gate.*

In practice, the SCR is intended to respond to a positive gate voltage. Under these circumstances, its behavior is shown by the solid curve in the top-right quadrant in Figure 1-8. The SCR begins to conduct current without having to reach the breakover voltage at the anode.

- When used as it is intended, the SCR should not reach breakdown or breakover voltage levels.

SCR Concept Demo

In Figure 1-9, pushbutton S1 applies voltage to the gate of the SCR, which puts the SCR in self-sustaining conductive mode. When S1 is released, the meter will show that current continues to pass between the anode and the cathode. The X0403DF SCR suggested for this circuit has a holding current of 5mA, which a 5VDC supply should be able to provide with the 1K resistor in the circuit. If necessary, this resistor can be reduced to 680Ω.

Now if pushbutton S2 is pressed, the flow is interrupted. When S2 is released, the flow will not resume. Alternatively, if pushbutton S3 is pressed while the SCR is conducting current, the flow is diverted around the SCR, and when the pushbutton is released, the flow through the SCR will not resume. Thus, the SCR can be shut down either by a normally closed pushbutton in series with it (which will interrupt the current), or a normally open pushbutton in parallel with it (which will divert the current).

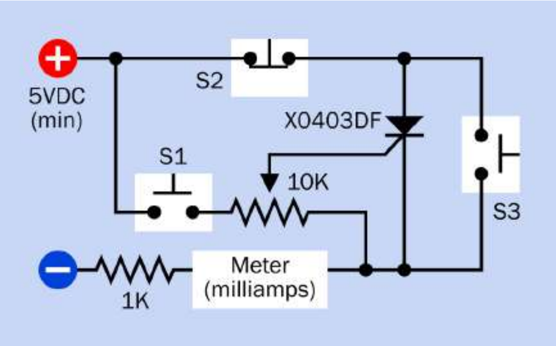

Figure 1-9. *In this test circuit, S1 triggers the SCR, while S2 or S3 will stop it. See text for additional details.*

The test circuit is shown installed on a breadboard in Figure 1-10. In this photograph, the red and blue wires supply a minimum of 5VDC. The two red buttons are tactile switches, the one at top left being S1 in the schematic while the one at bottom right is S3. The large switch with a rectangular button is S2; this is normally closed, and opens when pressed. The X0403DF SCR is just below it and to the right. The square blue trimmer is set to the midpoint of its range.

AC Current Applications

If the SCR is used with alternating current, it stops conducting during each negative cycle, and is retriggered in each positive cycle. This suggests one of its primary applications, as a controllable rectifier that can switch rapidly enough to limit the amount of current that passes through it during each cycle.

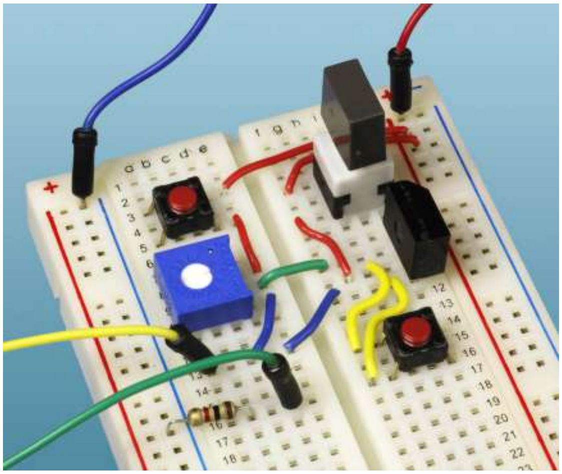

Figure 1-10. *A breadboarded version of the SCR test circuit. The two red buttons correspond with S1 and S3 in the schematic, while the large rectangular button at top right opens S2. See text for details.*

Variants

SCRs are available in surface-mount, through-hole, and stud packages, to handle increasing currents and voltages. Some special-purpose SCRs can control currents of hundreds of amps, while high-power SCRs are used to switch thousands of amps at more than 10,000V in power distribution systems. They are too specialized for inclusion in this encyclopedia.

Typical power ratings for SCRs in general use are summarized in the next section.

Values

Any SCR will impose a forward voltage drop, which typically ranges from around 1V to 2V, depending on the component.

Because SCRs are often used to modify AC waveforms, the current that the component can pass is usually expressed as the root mean square (RMS) of its peak value.

Commonly Used Abbreviations

- V_{DRM} Maximum repetitive forward voltage that may be applied to the anode while no voltage is applied to the gate (i.e., when the SCR is not in conductive mode).
- V_{RRM} Maximum repetitive reverse voltage that may be applied to the anode while no voltage is applied to the gate (i.e., when the SCR is not in conductive mode).
- V_{TM} Maximum on-state voltage while the SCR is in conductive mode. T indicates that this value changes with temperature.
- V_{GM} Forward maximum gate voltage.
- V_{GT} Minimum gate voltage required to trigger.
- V_{GD} Maximum gate voltage that will not trigger.
- I_{DRM} Peak repetitive forward blocking current (i.e., maximum leakage).
- I_{RRM} Peak repetitive reverse blocking current (i.e., leakage in the off state).
- I_{GM} Maximum forward gate current.
- $I_{T(RMS)}$ Maximum RMS current between anode and cathode while the SCR is in conductive mode. T indicates that this value changes with temperature.
- $I_{T(AV)}$ Maximum average current between anode and cathode while the SCR is in conductive mode. T indicates that this value changes with temperature.
- I_{GT} Maximum gate current required to trigger.
- I_H Typical holding current.
- I_L Maximum latching current.
- T_C Case temperature, usually expressed as an acceptable range.
- T_J Operating junction temperature, usually expressed as an acceptable range.

Surface-mount variants may tolerate maximum anode-cathode currents that typically range from 1A to 10A. Maximum voltages as high as 500V are allowed in some cases. Leakage in the “off” state may be as high as 0.5mA or as low as 5μA. Gate trigger voltage is likely to range from 0.8V to 1.5V, and trigger current of 0.2mA to 15mA is typical.

Through-hole variants may be packaged in TO-92 format (like discrete transistors) or, more commonly, in TO-220 format (like a typical 1A voltage regulator). They may be rated for a maximum of 5A up to 50A, depending on the component, with maximum voltages ranging from 50V to 500V. Leakage is similar to surface-mount variants. The gate trigger voltage is typically around 1.5V, and trigger current ranges from 25–50mA.

A stud-type SCR may have a maximum 50A to 500A current rating, although some components are capable of tolerating even higher values. Maximum voltages of 50V to 500V are possible. Leakage is likely to be higher than in other formats, with 5mA to 30mA being common. The gate trigger voltage is typically 1.5 to 3V, and trigger current may range from around 50mA to 200mA.

How to Use It

Although other applications are possible, in practice SCRs have two main applications:

- Phase control, which interrupts each positive phase of an AC power supply. It can moderate the speed of a motor or the heat generated by a resistive load.
- Overvoltage protection. This can safeguard sensitive components in a circuit where there is a DC power supply.

SCRs are often incorporated in ground-fault circuit interruptors (although not usually as discrete components) and in automotive ignition systems.

Phase Control

Phase control is a convenient way to control or limit the AC power delivered to a load by abbreviating each pulse in the AC waveform. This is done by adjusting the gate voltage so that the SCR blocks the first part of each positive phase, then conducts the remainder, and then stops conducting below its holding level. The SCR will then block the reversed flow in the negative phase of the AC waveform, but an additional SCR with opposite polarity can be added.

This is a form of *pulse-width modulation*. It is highly efficient, as the effective internal resistance of the SCR is either very high or very low, and the component does not waste significant energy in the form of heat.

On a graph showing the fluctuating voltage of an AC waveform, a single cycle is customarily divided into four stages: (1) zero voltage, (2) maximum positive voltage, (3) zero voltage, (4) minimum negative voltage, all measurements being made between the live side of the supply and the neutral side of the supply.

The cycle then repeats. Its transitions are often referred to as *phase angles* of 0 degrees, 90 degrees, 180 degrees, and 270 degrees, as shown in Figure 1-11.

The fluctuating voltage in an AC power supply is proportional with the sine of the phase angle. This concept is illustrated in Figure 1-12. If an imaginary point (shown as a purple dot) is moving in a circular path counterclockwise at a constant speed, its vertical distance (shown in green) above or below the X axis (horizontal centerline) can represent an AC voltage corresponding with the angle (shown as purple arcs) of the circle radius to the point, each angle being measured from at the center relative to a start position at right on the X axis.

When an SCR is used for phase control, the point at which it starts to conduct may be anywhere from 0 to almost 180 degrees. This is achieved by diverting a small amount of the AC power into an RC network attached to the gate of the SCR, as shown in Figure 1-13. The capacitor in this schematic introduces a delay that can be varied by the potentiometer. This enables the SCR to be triggered even after the peak of the AC power signal. In Figure 1-14, the AC power is shown by the center (green) curve, and the slightly delayed, reduced voltage at the gate is shown by the upper, purple curve. When the gate voltage rises to the trigger level, it causes the SCR to begin conducting current, creating an abbreviated output shown in the bottom curve. In this way, triggering from an AC phase angle of 0 degrees to almost 180 degrees is possible. The phase angle where the SCR begins to allow conduction is known as the *conduction angle*.

If two SCRs with opposite polarity are placed in parallel with each other, they can be used to provide phase control on both the positive-going and negative-going sections of an AC cycle. This configuration is used in high-powered devices. A **triac** is used for the same purpose with lower current.

Six SCRs may be used to control three-phase power.

Overvoltage Protection

The tolerance of an SCR for high current makes it suitable for use in a crowbar voltage limiting circuit.

In Figure 1-15, the SCR does not conduct current (other than a small amount of leakage) until the Zener diode senses a voltage above the maximum level considered safe. The diode then allows power to reach the gate of the SCR. Its impedance drops immediately, and the resulting surge of current trips the fuse. After the cause of the overvoltage condition is corrected, the fuse can be replaced and the circuit may resume functioning.

A capacitor is included so brief spikes in the power supply will be passed to ground without triggering the SCR. A resistor of around 100Ω ensures that the gate voltage of the SCR remains near zero during normal operation. When the Zener diode starts to conduct current, the resistor acts as a voltage divider with the diode, so that sufficient voltage reaches the SCR to activate it.

This circuit may be unsuitable for low-voltage power supplies, because the Zener diode has to be chosen with a high enough rating to prevent small power fluctuations from tripping it. Bearing in mind that the real triggering voltage of the diode may be at least plus-or-minus 5% of its rated voltage, the diode may have to be chosen with at least a 6V rating in a 5V circuit, and it may not be activated until the voltage is actually 6.5V. This may be insufficient to protect the components being used with the power supply.

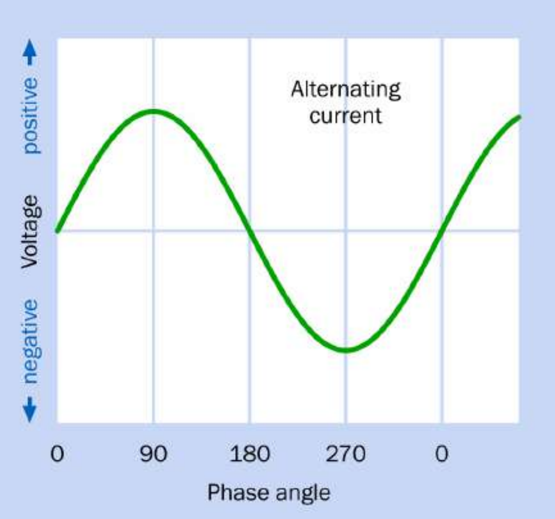

Figure 1-11. *An AC waveform is customarily measured in degrees of phase angle.*

What Can Go Wrong

Like other semiconductors, an SCR can be adversely affected by excessive heat. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, the values for triggering current and holding current are valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Unexpected Triggering Caused by Voltage

A very rapid increase in forward voltage at the anode can induce a triggering voltage in the gate by capacitive coupling. As a result, the SCR can trigger itself without any external application of gate voltage. This is sometimes known as *dv/dt triggering*. If necessary, a snubber circuit can be added across the anode input to prevent sudden voltage transitions.

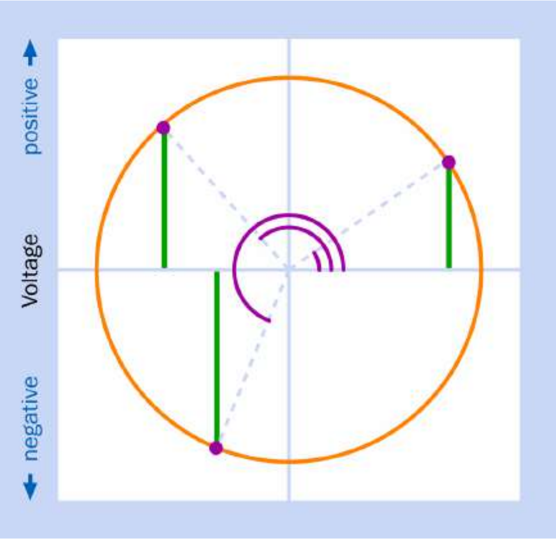

Figure 1-12. *The fluctuating voltage of an AC power supply (shown as vertical green lines) is proportional with the sines of the angles (purple arcs) in this diagram. The angles are referred to as phase angles.*

Confusion of AC and DC Ratings

The on-state current for an SCR is averaged only over the width of each pulse that the SCR actually conducts. It is not time-averaged over an entire AC cycle, and it will be different again from a DC rating. Care must be taken to match the current rating with the way in which the component will actually be used.

Maximum Current versus Conduction Angle

Current-carrying capability will be very significantly affected by the length of the duty cycle when the SCR is being used to abbreviate each positive AC pulse. When the SCR imposes a 120-degree conduction angle, it may be able to handle twice the average on-state current as when it is imposing a 30-degree conduction angle. The manufacturer's datasheet should include a graphical illustration of this relationship. If an SCR is chosen for a high conduction angle, and the angle is later reduced, overheating will result, and damage is likely.

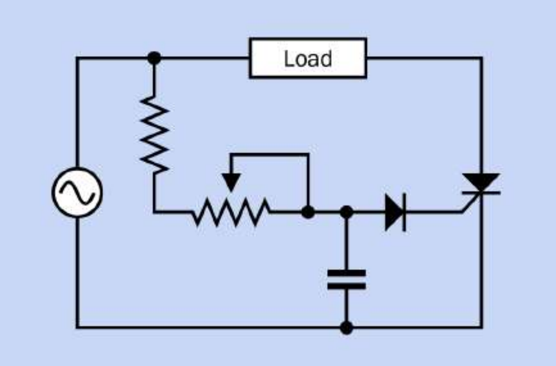

Figure 1-13. *In this schematic, an SCR is used to apply phase control, adjusting the power that passes through a load.*

Confusing Symbols

When reading a schematic, unfortunate errors can result from failure to distinguish between the symbol for a **programmable unijunction transistor (PUT)** and the symbol for an SCR. The characteristics of a PUT are described in Volume 1 of this encyclopedia.

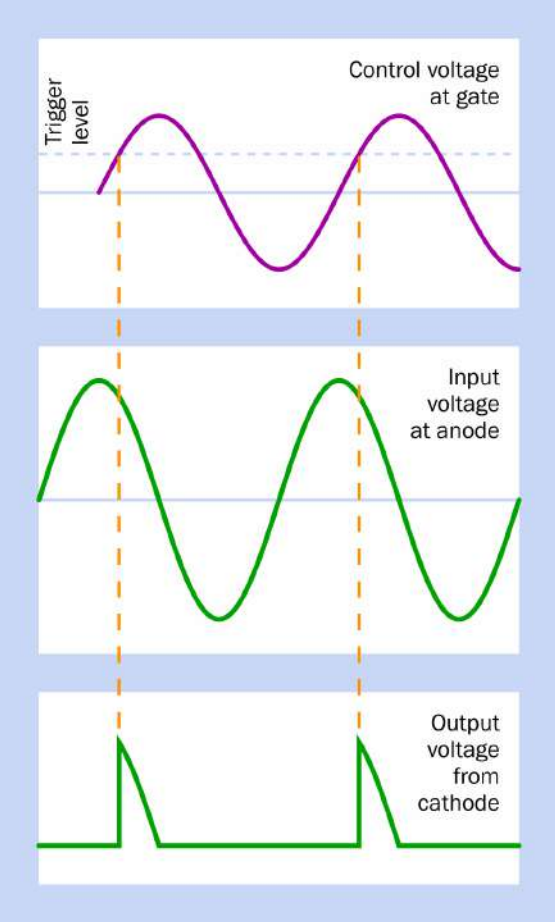

Figure 1-14. *If the AC power applied to the anode of an SCR (center) is reduced in voltage and delayed slightly by an RC network, it can trigger the SCR (top), causing it to pass only an abbreviated segment of each positive AC pulse (bottom).*

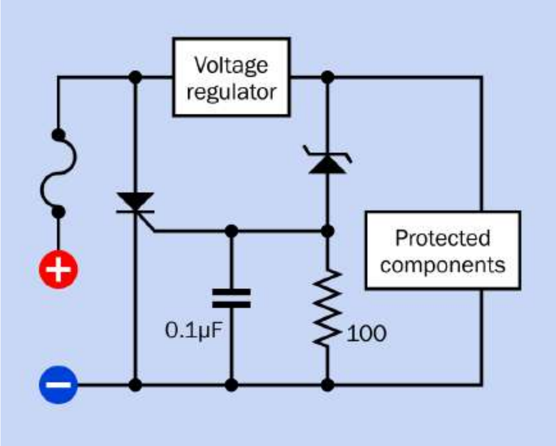
Voltage regulator
Protected components
0.1µF
100

Figure 1-15. *In this schematic, an SCR is used to provide crowbar overvoltage protection for sensitive components.*

diac

2

A **diac** is a self-triggering type of *thyristor*. Its name is said to be derived from the phrase "diode for AC," and because it is not an acronym, it is not usually capitalized.

A thyristor is defined here as a semiconductor having four or more layers of p-type and n-type silicon. Because the thyristor predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, it is categorized as a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**) it is considered to be an integrated circuit.

Other types of thyristor are the **SCR** (silicon-controlled rectifier) and the **triac**, each of which has its own entry in this encyclopedia.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **SCR** (see Chapter 1)
- **triac** (see Chapter 3)

What It Does

The **diac** is a bidirectional thyristor with only two terminals. It blocks current until it is subjected to sufficient voltage, at which point its impedance drops very rapidly. It is primarily used to trigger a **triac** for purposes of moderating AC power to an **incandescent lamp**, a resistive heating element, or an AC **motor**. The two leads on a diac have identical function and are interchangeable.

By comparison, a **triac** and an **SCR** are thyristors with three leads, one of them being referred to as the gate, which determines whether the component becomes conductive. A triac and a diac allow current to flow in either direction, while an SCR always blocks current in one direction.

Symbol Variants

The schematic symbol for a diac, shown in Figure 2-1, resembles two diodes joined together, one of them inverted relative to the other. Functionally, the diac is comparable with a pair of Zener diodes, as it is intended to be driven beyond the point where it becomes saturated. Because its two leads are functionally identical, they do not require names to differentiate them. They are sometimes referred to as A1 and A2, in recognition that either of them may function as an anode; or they may be identified as MT1 and MT2, MT being an acronym for "main terminal."

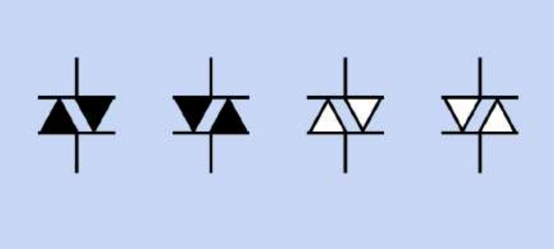

Figure 2-1. *Symbol variants to represent a diac. All four are functionally identical.*

The symbol may be reflected left to right, and the black triangles may have open centers. All of these variants mean the same thing. Occasionally the symbol has a circle around it, but this style is now rare.

When only a moderate voltage is applied (usually less than 30V) the diac remains in a passive state and will block current in either direction, although a very small amount of *leakage* typically occurs. When the voltage exceeds a threshold known as its *breakover level*, current can flow, and the diac will continue to conduct until the current falls below its *holding level*.

A sample diac is shown in Figure 2-2.

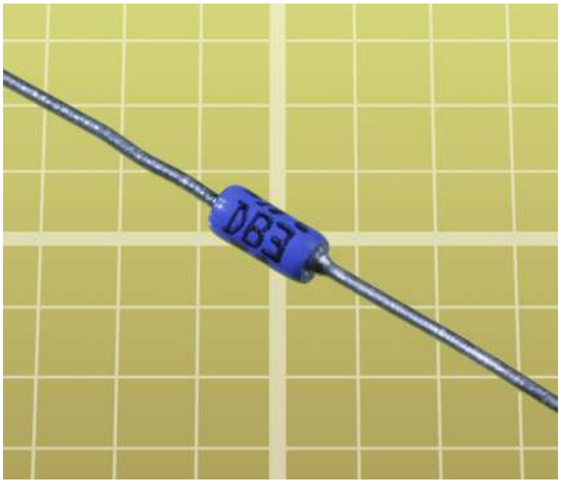

Figure 2-2. *Because a diac is not intended to pass significant current, it is typically packaged in a small format. The graph squares in the photograph each measure 0.1".*

How It Works

Figure 2-3 shows a circuit that demonstrates the conductive behavior of a diac.

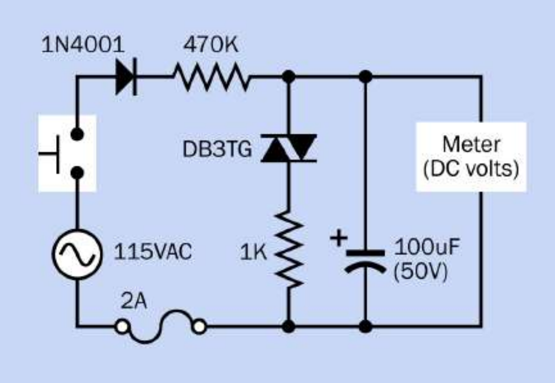

Figure 2-3. *A test circuit to demonstrate the behavior of a diac. See text for details.*

When the pushbutton is held down, current from the positive side of the AC supply flows through the diode and the 470K resistor to the capacitor. The diac is not yet conductive, so the capacitor accumulates a potential that can be monitored with the volt meter. After about 30 seconds, the charge on the capacitor reaches 32V. This is the breakover voltage for this particular diac, so it becomes conductive. The positive side of the capacitor can now discharge through the diac and the 1K series resistor to ground.

If the pushbutton is released at this moment, the meter will show that the capacitor discharges to a potential below the holding level of the diac. The capacitor now stops discharging because the diac has ceased being conductive.

If the pushbutton is held down constantly, the meter will show the capacitor charging and then discharging through the diac repeatedly, so that the circuit behaves as a *relaxation oscillator*. The 1K series resistor is included to protect the diac from excessive current. If a standard quarter-watt resistor is used, it should not become unduly warm because current passes through it only intermittently.

- Because this circuit uses 115VAC, basic precautions should be taken. The fuse should not be omitted, the capacitor should be rated for at least 50V, and the circuit should not be touched while it is connected to the power source. Breadboarding a circuit using this voltage requires caution and experience, as wires can easily come loose, and components can be touched accidentally while they are live.

Figure 2-4 shows the test circuit on a breadboard. The red and blue leads at the top of the photograph are from a fused 115VAC power supply. The live side of the supply passes through a diode to a pushbutton switch that has a rectangular black cap. A 470K resistor connects the other side of the switch to the positive side of a 100µF electrolytic capacitor, and also to the diac (small blue component). A 1K resistor connects the other end of the diac back to the negative side of the capacitor, which is grounded. The yellow and blue wires leaving the photograph at the left are connected with a volt meter, which is not shown.

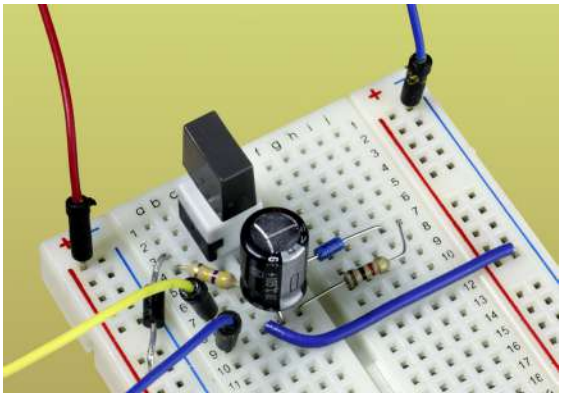

Figure 2-4. *A breadboarded version of the diac test circuit. See text for details.*

The behavior of a diac is also illustrated in Figure 2-5, which can be compared with the curves in Figures 3-10 and 1-8, depicting the behavior of a triac and an SCR respectively.

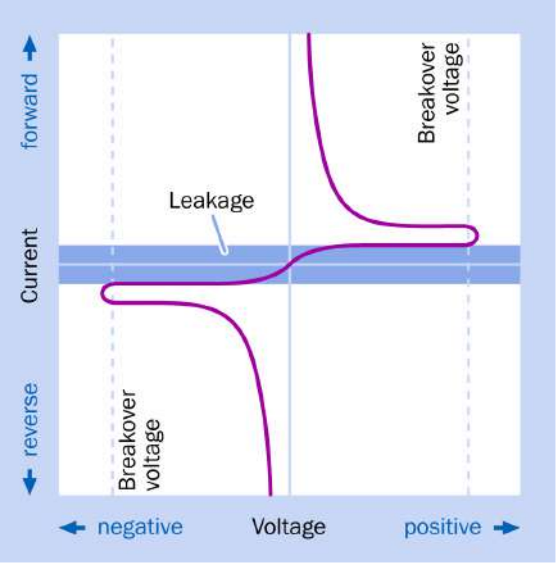

Figure 2-5. *The curve shows current passing through a diac when various voltages are applied.*

Switching AC

The diac cannot function as a switch, because it lacks the third terminal which is found in a triac, an SCR, or a bipolar transistor. However, it is well suited to drive the gate of a triac, because the behavior of a diac is symmetrical in response to opposite voltages, while the triac is not. If an AC voltage applied to a diac is adjusted with a potentiometer in an RC circuit, the diac will pass along a portion of each positive or negative pulse, and will delay it by a brief amount of time determined by the value of the capacitor in the RC circuit and the setting of the potentiometer. This is known as *phase control*, as it controls the *phase angle* at which the diac allows current to flow.

See Figure 3-13 for a schematic showing a diac driving a triac. See Figures 1-14 and 3-11 for graphs illustrating phase control. See "Phase Control" for a discussion of phase in AC waveforms.

Variants

Diacs are available in through-hole and surface-mount formats. Because they are not intended to handle significant current, no heat sink is included.

A *sidac* behaves very similarly to a diac, its name being derived from "silicon diode for alternating current." Its primary difference from generic diacs is that it is designed to reach its breakover voltage at a higher value, typically 120VAC or 240VAC.

Values

When performing its function to trigger a triac, a diac is unlikely to pass more than 100mA.

The breakover voltage of a diac is usually between 30V and 40V, with a few versions designed for up to 70V. When the diac starts to conduct, its on-state impedance is sufficient to reduce the voltage significantly, with 5V being a typical minimum output voltage.

Although the rise time when a diac responds is very brief (around 1μs), the component is not expected to run at a high frequency. It will normally be used with 50Hz or 60Hz AC to trigger a triac. For this reason, its repetitive peak on-state current is usually specified at no more than 120Hz.

Abbreviations in datasheets are likely to include:

- V_{BO} Breakover voltage (sometimes may be specified as latching voltage, which for a diac is the same thing).
- V_{BO1} - V_{BO2} Breakover voltage symmetry. The hyphen is intended as a minus sign, so that this value is the maximum difference between breakover voltage in each direction.
- V_O Minimum output voltage.
- I_{TRM} Repetitive peak on-state current.
- I_{BO} Breakover current, usually the maximum required, and less than 20μA.
- I_R Maximum leakage current, usually less than 20μA.
- T_J Operating junction temperature, usually expressed as an acceptable range.

What Can Go Wrong

Like other semiconductors, a diac is heat sensitive. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, a value for breakover current is valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Low-Temperature Effects

A higher breakover voltage will be required by a diac operating at low temperatures, although the variation is unlikely to be greater than plus-or-minus 2% within a normal operating range. Temperature has a much more significant effect on a triac.

Manufacturing Tolerances

The breakover voltage for a diac is not adjustable, and may vary significantly between samples of the component that are supposed to be identical. The diac is not intended to be used as a precision component. In addition, while its breakover voltage should be the same in either direction, a difference of plus-or-minus 2% is possible (1% in some components).

triac

3

A **triac** is a gate-triggered type of *thyristor*. Its name was probably derived from the phrase "triode for AC," and because it is not an acronym, it is not usually capitalized.

A thyristor is defined here as a semiconductor having four or more layers of p-type and n-type silicon. Because the thyristor predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, it is categorized as a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**) it is considered to be an integrated circuit.

Other types of thyristor are the **SCR** (silicon-controlled rectifier) and the **diac**, each of which has its own entry in this encyclopedia.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **SCR** (see Chapter 1)
- **diac** (see Chapter 2)

What It Does

The **triac** is ubiquitous in AC dimmers for **incandescent lamps**. It is also used to control the speed of AC motors and the output of resistive heating elements. It is a type of *thyristor* which contains five segments of p-type and n-type silicon and has three leads, one of them attached to a gate that can switch a bidirectional flow of current between the other two. Its name was originally a trademark, generally thought to be derived from the phrase "triode for AC." A triode was a common type of vacuum tube when thyristors were first introduced in the 1950s.

By comparison, a **diac** is a thyristor with only two leads, allowing current to flow in either direction when the component reaches a *breakover voltage*. Its name was probably derived from the phrase "diode for AC." It is often used in conjunction with a triac.

An **SCR** (silicon-controlled rectifier) is a thyristor that resembles a triac, as it has three leads, one of them a gate. However, it only allows current to flow in one direction.

Symbol Variants

The schematic symbol for a triac, shown in Figure 3-1, resembles two diodes joined together, one of them inverted relative to the other. While a triac does not actually consist of two diodes, it is functionally similar, and can pass current in either direction.

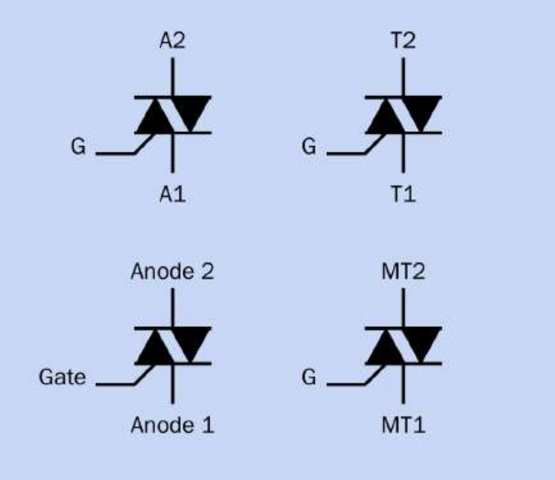

Figure 3-1. *The schematic symbol for a triac, with four naming conventions that are used for its leads. The different conventions do not indicate any functional difference.*

An appended bent line represents the gate. The labels for the other two leads are not standardized, and can be referred to as A1 and A2 (for Anode 1 and Anode 2), or T1 and T2 (for Terminal 1 and Terminal 2), or MT1 and MT2 (for Main Terminal 1 and Main Terminal 2). The choice of terms does not indicate any functional difference. In this encyclopedia entry, A1 and A2 are used.

The A1 terminal (or T1, or MT1) is always shown closer to the gate than A2 (or T2, or MT2). This distinction is important because although the triac can pass current in either direction, its behavior is somewhat asymmetrical.

- Voltages are expressed relative to terminal A1 (or T1, or MT1, if those terms are used).

The schematic symbol may be reflected or rotated, the black triangles may have open centers, and the placement of the bent line representing the gate may vary. However, terminal A1 is always nearer to the gate than terminal A2.

Figure 3-2 shows 12 of the 16 theoretical possibilities. All of these variants are functionally identical. Occasionally the symbol has a circle around it, but this style is now rare.

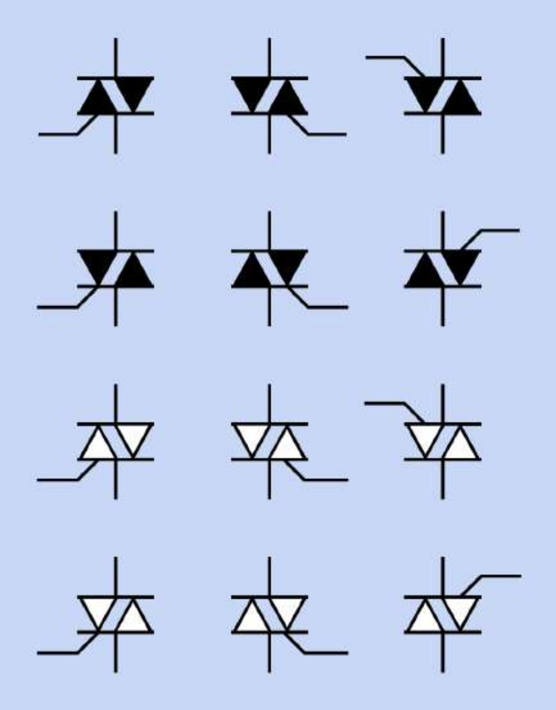

Figure 3-2. *Interchangeable variants of the schematic symbol for a triac.*

Triacs with various characteristics are shown in Figures 3-3, 3-4, and 3-5.

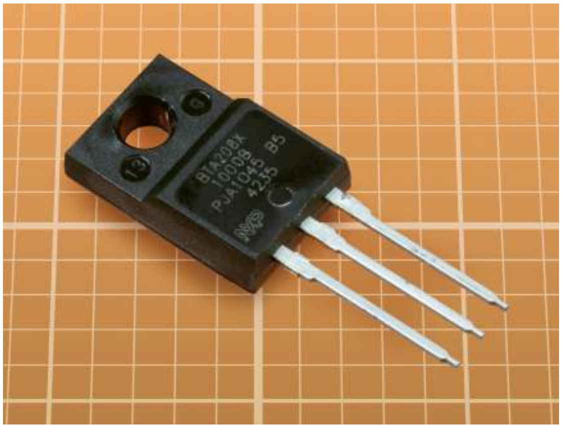

Figure 3-3. *The BTA208X-1000B triac can conduct 8A continuous on-state current RMS, and withstands peak off-state voltage of up to 1,000V. This is a "snubberless" triac.*

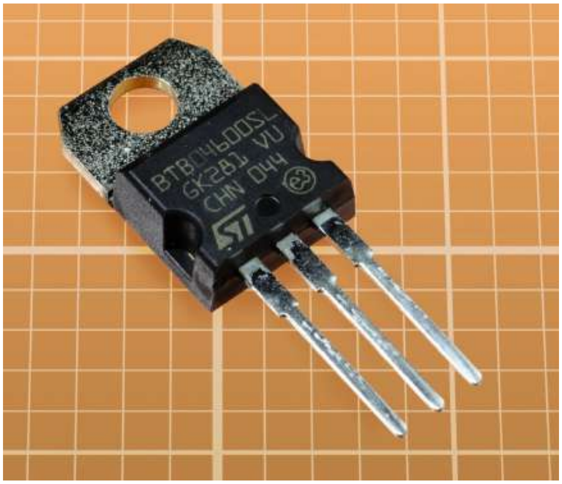

Figure 3-4. *The BTB04-600SL triac can conduct 4A continuous on-state current RMS, and withstands peak off-state voltage of up to 600V.*

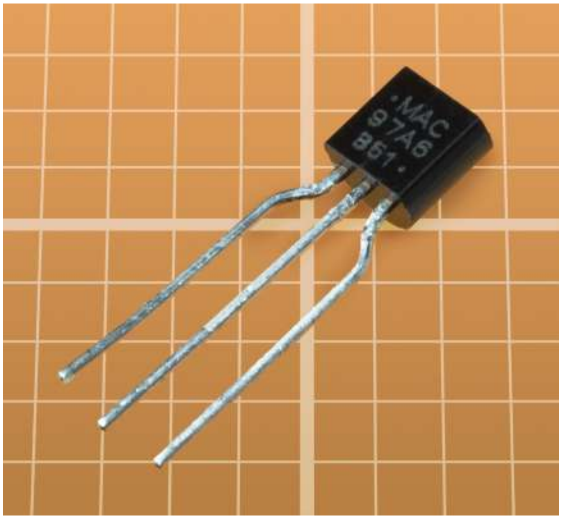

Figure 3-5. *The MAC97A6 triac can conduct 0.8A continuous on-state current RMS, and withstands peak off-state voltage of up to 400V.*

How It Works

When no gate voltage is applied, the triac remains in a passive state and will block current in either direction between A1 and A2, although a very small amount of *leakage* typically occurs. If the gate potential becomes sufficiently positive *or* negative relative to terminal A1, current can begin to flow between A1 and A2 in *either* direction. This makes the triac ideal for controlling AC.

Quadrants

While a gate voltage is applied, four operating modes are possible. In each case, A1 is the reference (which can be thought of as being held at a neutral ground value). Because the triac is conducting AC, voltages above and below ground will occur. The four modes of operation are often referred to as four *quadrants*, and are typically arranged as shown in Figure 3-6.

In some reference sources (especially educational text books), current is shown with an arrow indicating a flow of electrons moving from negative to positive. Because the type of current flow is often undefined, diagrams should be interpreted carefully. In this encyclopedia, current is always shown flowing from a more-positive location to a more-negative location.

Quadrant 1 (upper right)

A2 is more positive than A1, and the gate is more positive than A1. Conventional current (positive to negative) will flow from A2 to A1. (This behavior is very similar to that of an **SCR**.)

Quadrant 2 (upper left)

A2 is more positive than A1, and the gate is more negative than A1. Once again, conventional current (positive to negative) will flow from A2 to A1.

Quadrant 3 (lower left)

A2 is more negative than A1, and the gate is more negative than A1. Conventional current is reversed from A1 to A2.

Quadrant 4 (lower right)

A2 is more negative than A1, but the gate is more positive than A1. Conventional current is reversed from A1 to A2.

- Note that two positive symbols or two negative symbols in Figure 3-6 do not mean that both locations are of equal voltage. They simply mean that these locations are at potentials that are significantly different from A1.

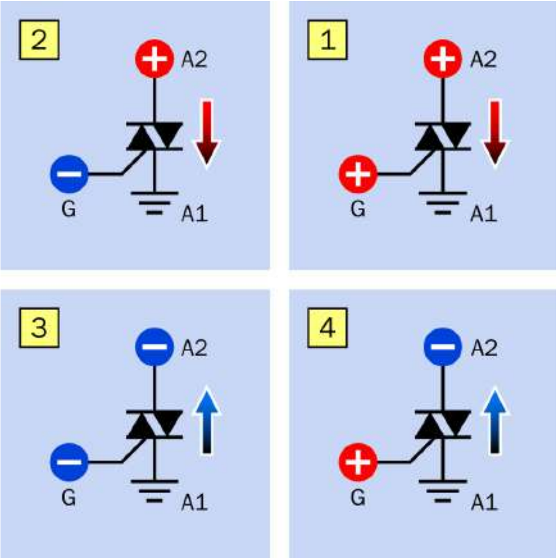

Figure 3-6. *The "quadrants" of triac behavior. Positive and negative symbols indicate which terminal is "more positive" or "more negative" than A1. The ground symbol represents a potential midway between positive and negative. See text for details.*

Suppose that gate current increases gradually. When it reaches the *gate threshold current* of the triac, the component starts conducting between A1 and A2. If the current between A1 and A2 rises above the value known as the *latching current*, it will continue to flow, even if gate current disappears completely.

If the self-sustaining current through the triac gradually diminishes, while there is no voltage applied to the gate, conduction between the main terminals will stop spontaneously when it falls below a level known as the *holding current*. This behavior is similar to that of an SCR. The triac now returns to its original state, blocking current until the gate triggers it again.

The triac is sufficiently sensitive to respond to rapid fluctuations, as in 50Hz or 60Hz AC.

Threshold, Latching, and Holding Current

Figure 3-7 shows the relationship between the gate threshold current, the latching current, and holding current. In the upper half of the figure, gate current is shown fluctuating until it crosses the threshold level. This establishes current flow between the main terminals, shown in the lower half of the figure. Prior to this moment, a very small amount of leakage current occurred (shown in the figure, but not to scale).

In this hypothetical scenario, the triac starts passing current between external components—and the current exceeds the latching level. Consequently, gate current can diminish to zero, and the triac remains conductive. However, when external factors cause the current between the main terminals to diminish below the holding level, the triac abruptly ceases to be conductive, and current falls back to the leakage level.

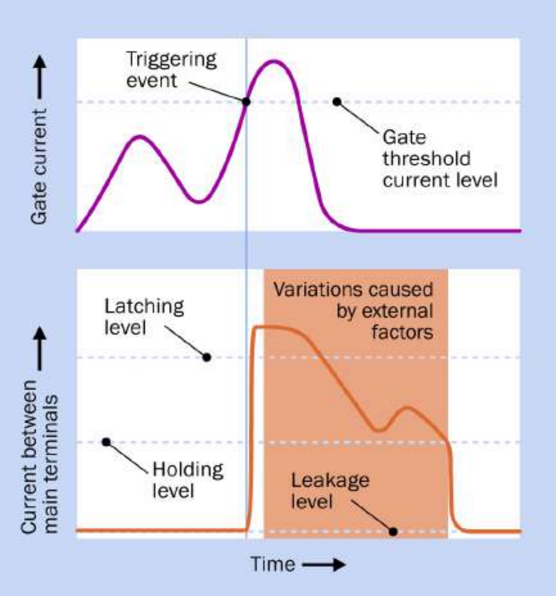

Figure 3-7. *The relationship between gate current of a triac and the current between its main terminals. See text for details.*

Unlike a **bipolar transistor**, a triac is either "on" or "off" and does not function as a *current ampli-*

fier. When it has been triggered, the impedance between A1 and A2 is low enough for heat dissipation to be manageable even at relatively high power levels.

Triac Testing

Figure 3-8 shows a circuit which can demonstrate the conductive behavior of a triac. For simplicity, this circuit is DC powered. In a real application, the triac is almost always used with AC.

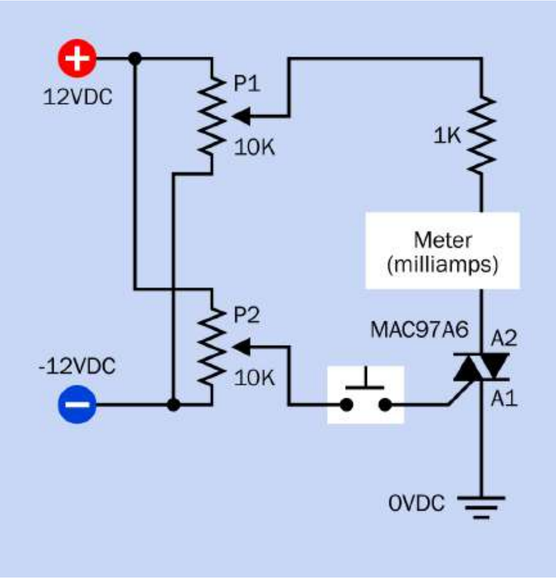

Figure 3-8. *A test circuit to show the behavior of a triac when varying positive and negative potentials are applied to the gate and to the A2 terminal, relative to A1.*

Note that this circuit requires at least a +12VDC and -12VDC power supply (higher values may also be used). The ground symbol represents a midpoint voltage of 0VDC, applied to terminal A1 of the triac, which is an MAC97A6 or similar. If a dual-voltage power supply is unavailable, the gate of the triac can be connected directly to +12VDC, omitting potentiometer P2; but in this case, only two operating modes of the triac can be demonstrated by turning potentiometer P1.

Each potentiometer functions as a voltage divider between the positive and negative sides of the power supply. P1 applies a positive or negative voltage to A2, relative to A1. P2 applies a positive or negative voltage to the gate, relative to A1.

If the test begins with both potentiometers at the top ends of their range, A1 and G both have a positive potential relative to A1, so that the triac is now in quadrant 1 of its operating modes. Pressing the pushbutton should cause it to start conducting current limited by the 1K resistor, and the meter should change from measuring 0mA to around 12mA. If the pushbutton is released, the triac should continue to conduct current, because 12mA is above this triac's latching level. If P1 is slowly moved toward the center of its range, the current diminishes, ceasing when it falls below the holding level. If P1 is now moved back to the top of its range, the current will not resume until the triac is retriggered with the pushbutton.

The test can be repeated with P1 at the top of its range and P2 at the bottom of its range, to operate the triac in quadrant 2; P1 at the bottom of its range and P2 at the bottom of its range, to operate the triac in quadrant 3; and P1 at the bottom of its range and P2 at the top of its range, to operate the triac in quadrant 4. The functionality should be the same in each case. The pushbutton will initiate a flow of current, which will diminish when P1 is turned toward the center of its range.

In any of these quadrants, P2 can be turned slowly toward the center of its range while the pushbutton is pressed repeatedly. This will allow empirical determination of the gate threshold current for this triac. The meter, measuring milliamps, will measure the current if it is inserted between the wiper of the potentiometer and the gate of the triac.

The test circuit is shown installed on a breadboard in Figure 3-9. In this photograph, the red and blue wires at left supply +12VDC and -12VDC relative to the black ground wire at top right. The yellow and green wires connect with a meter set to measure milliamps. The red button is a tactile switch, while the MAC97A6 triac is just above it and to the left. The square blue 10K trimmers are set to opposite ends of their scales, so that the meter will show current flowing when the tactile switch is pressed.

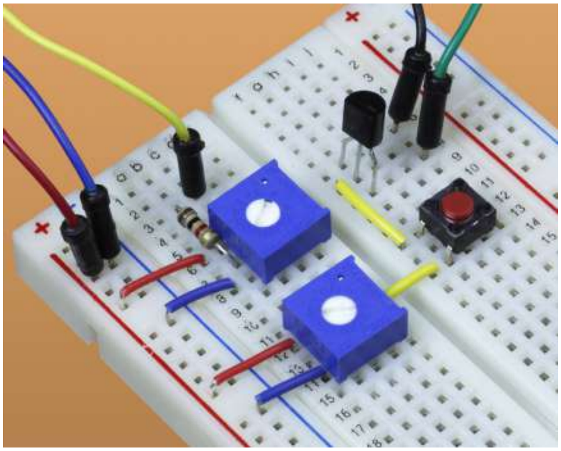

Figure 3-9. *A breadboarded triac test circuit.*

Breakover Voltage

If a much higher voltage is applied to A2, the triac can be forced to conduct current without any triggering voltage being applied to the gate. This occurs when the potential between A1 and A2 reaches the triac's *breakover voltage*, although the component is not designed to be used this way. The concept is illustrated in Figure 3-10, which can be compared with the behavior of an SCR illustrated in Figure 1-8 and the behavior of a diac shown in Figure 2-5. While the term *breakdown voltage* defines the minimum reverse voltage required to force a diode to conduct, *breakover voltage* refers to the minimum forward voltage that has this effect. Because a triac is designed to conduct in both directions, it can be thought of as having a breakover voltage in both directions.

In Figure 3-10, the numbers in yellow squares are the quadrants referred to in Figure 3-6. The solid curve represents current flow if a triggering voltage is applied to the gate while a positive or negative potential is applied to A2, relative to A1. If the gate is not triggered while the voltage between A1 and A2 gradually increases, the dashed section of the curve illustrates the outcome when the component reaches breakover voltage. Although this may not damage the triac, the component becomes uncontrollable.

- In normal usage, the voltage between A1 and A2 should not be allowed to reach breakover level.

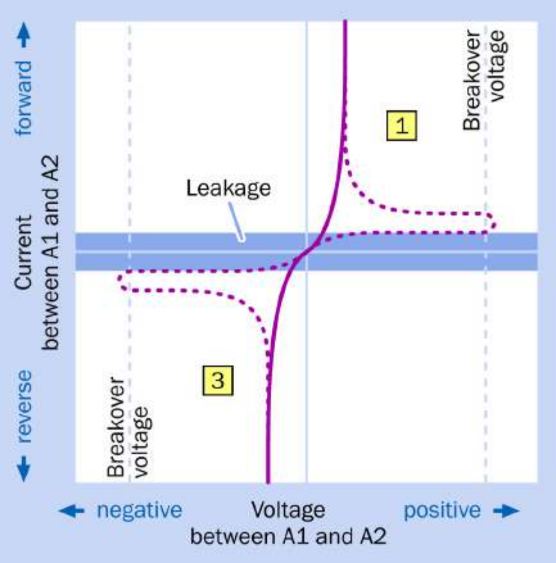

Figure 3-10. *The solid curve shows current passing between A1 and A2 in a hypothetical triac, for varying voltages, while triggering voltage is applied to the gate. The dashed curve assumes that no triggering voltage is applied to the gate. The numbers in yellow squares are the quadrants of triac operation.*

Switching AC

"Switching" AC with a triac means interrupting each pulse of current so that only a portion of it is conducted through to the load. Usually this is done with the triac functioning in quadrants 1 and 3. In quadrant 3, the polarity of the flow between A1 and A2 is opposite to that in quadrant 1, and the gate voltage is also reversed. This enables a relatively simple circuit to control the duration of each half-cycle passing through the triac. The theory of this circuit is shown in Figure 3-11.

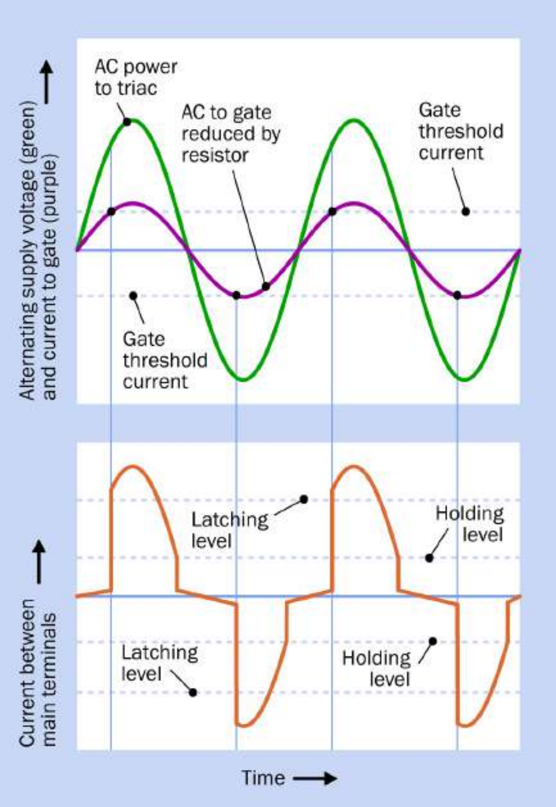
Alternating supply voltage (green) and current to gate (purple)
AC power to triac
AC to gate reduced by resistor
Gate threshold current
Gate threshold current
Current between main terminals
Latching level
Holding level
Latching level
Holding level
Time

Figure 3-11. *To moderate the power of AC current, a triac can block a section of each AC pulse.*

The upper section of Figure 3-11 shows alternating voltage to the triac in green. The purple curve represents the gate current of the triac, reduced by a variable resistor. (The figure is for conceptual purposes only; the alternating power supply voltage and the fluctuating gate current cannot actually share the same vertical scale of a graph.)

Figure 3-11 can be compared with Figure 3-7, except that the negative threshold level for the gate is now shown as well as the positive threshold level. Remember, either a positive or negative voltage can activate the gate.

In Figure 3-11, initially the triac is nonconductive. As time passes, the gate current reaches the threshold level, and this triggering event enables current to flow between the main terminals of the triac, as shown in the lower part of the figure.

This current exceeds the latching level, so it continues to flow, even though the gate current diminishes below its threshold level. Finally the current between the main terminals falls below the holding level, at which point the triac stops conducting. It waits for the next triggering event, which occurs as the power supply swings to negative.

This simple system blocks a section of each AC pulse, which will vary in length depending how much current is allowed to flow through the gate. Because the blocking process occurs rapidly, we notice only the reduced overall power passing through the triac (in terms of the brightness of a light, the heat emitted by a resistive element, or the speed of a motor).

Unfortunately, there is a problem in this scenario: the triac does not quite behave symmetrically. Its gate threshold level for positive current is not exactly equal and opposite to its gate threshold level for negative current. The upper part of Figure 3-11 shows this flaw in the differing vertical offsets of the positive and negative thresholds from the central zero line.

The result is that negative AC pulses through the triac are shorter than positive pulses. This asymmetry produces harmonics and noise that can feed back into power supply wiring, interfering with other electronic equipment. The actual disparities in gate response, in each quadrant of operation for two triacs, are shown in Figure 3-12.

Current capacity of triac	Ratio of gate current required to conduct, relative to quadrant 1		
	Quadrant 2	Quadrant 3	Quadrant 4
4 amp	1.6	2.5	2.7
10 amp	1.5	1.4	3.1

Figure 3-12. *Because the internal structure of a triac is asymmetrical, it requires a different trigger current in each of its operating quadrants. This table, derived from a Littelfuse technical briefing document, shows the ratio of the minimum trigger current in quadrants 2, 3, and 4 relative to quadrant 1.*

See Figure 1-14 for a graph illustrating phase control in the SCR. See "Phase Control" for a discussion of phase in AC waveforms generally.

Triac Triggered by a Diac

The problem of asymmetrical triggering can be overcome if the triac is triggered with a voltage pulse generated by another component that does behave symmetrically. The other component is almost always a diac, which is another type of thyristor. Unlike an SCR or a triac, it has no gate. It is designed to be pushed beyond its breakover voltage, at which point it latches and will continue to conduct until current flowing through it diminishes below its holding level. See Chapter 2 for more information about the diac.

In Figure 3-13, the diac is shown to the right of the triac, and is driven by a simple RC network consisting of a fixed resistor, a potentiometer, and a capacitor. (In an actual application, the RC network may be slightly more complex.) The capacitor takes a small amount of time to charge during each half-cycle of AC. The length of this delay is adjusted by the potentiometer, and determines the point in each AC half-cycle when the voltage to the diac reaches breakover level. Because the delay affects the phase of the AC, this adjustment is known as *phase control*.

As the voltage exceeds breakover level, the diac starts to pass current through to the gate of the triac, and triggers it. The holding level of the diac is lower than its latching level, so it continues to pass current while the capacitor discharges and the voltage diminishes. When the current falls below the holding level, the diac stops conducting, ready for the next cycle. Meanwhile, the triac continues to pass current until the AC voltage dips below its holding level. At this point, the triac becomes nonconductive until it is triggered again.

This chopped waveform will still create some harmonics, which are suppressed by the coil and capacitor at the left side of the circuit in Figure 3-13.

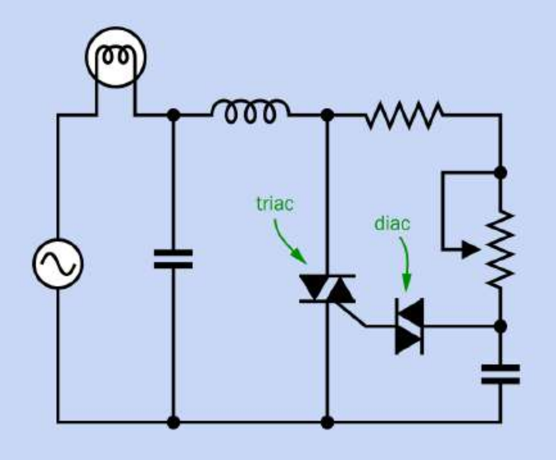

Figure 3-13. *A minimal schematic showing typical operation of a triac, with a diac supplying pulses to the triac gate. The potentiometer adjusts the delay created by the capacitor.*

Other Triac Drivers

It is possible, although unusual, to drive a triac from a source other than a diac.

Simple on-off control can be achieved by using a special optocoupler such as the MOC3162 by Fairchild Semiconductor. This emits a switching signal to a triac only when the AC voltage passes through zero. A *zero cross* circuit is desirable because it creates much less interference. The use of an optocoupler helps to isolate the triac from other components.(((“zero cross circuit”))

Phase control can be achieved using an optocoupler such as the H11L1, which can be driven by rectified but unsmoothed AC after it passes through a Zener diode to limit the voltage. The output from the optocoupler is logic-compatible and can be connected with the input to a timer such as the 555, set to one-shot mode. Each pulse from the timer passes through another optocoupler such as the MOC3023, which uses an internal LED to trigger the gate of a triac.

Yet another possibility is to use the programmed output from a microcontroller, through an optocoupler, to control the gate of a triac. An online search for the terms "microcontroller" and "triac" will provide some additional suggestions.

Charge Storage

While switching AC, the internal charge between A1 and A2 inside the triac requires time to dissipate before the reverse voltage is applied; otherwise, *charge storage* occurs, and the component may start to conduct continuously. For this reason, the triac is normally restricted to relatively low frequencies such as domestic 60Hz AC power.

When a triac controls a motor, the phase lag between voltage and current associated with an inductive load can interfere with the triac's need for a transitional moment between a positive and negative voltage cycle. In a datasheet, the term *commutating dv/dt* defines the rate of rise of opposite polarity voltage that the triac can withstand without locking into a continuous-on state.

An RC *snubber network* is often wired in parallel with A1 and A2 to control the rise time of voltage to the triac, as shown within the darker blue rectangle in Figure 3-14, where a resistor and capacitor have been added just to the left of the triac. The highest resistance and lowest capacitance, consistent with trouble-free operation, should be chosen. Typical values are 47Ω to 100Ω for the resistor, and 0.01µF to 0.1µF for the capacitor.

Variants

Triacs are available in through-hole and surface-mount packages.

Some components that are referred to as triacs actually contain two SCR components of opposite polarity. The "alternistor" range from Littelfuse is an example. The SCR will tolerate faster voltage rise times than a conventional triac, and is more suitable for driving inductive loads such as large motors.

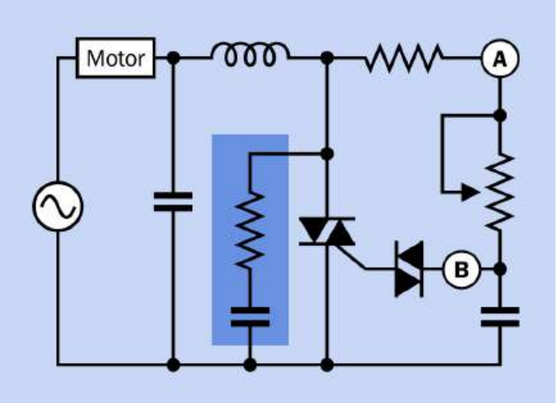

Figure 3-14. *To prevent a triac from locking itself into a continuous-on state while driving an inductive load such as a motor, a snubber circuit can be added (shown here as a resistor and capacitor in the darker blue rectangle to the left of the triac).*

A *snubberless triac*, as its name implies, is designed to drive an inductive load without need for a snubber circuit. An example is the STMicroelectronics BTA24. Datasheets for this type of component impose some limits that may be stricter than for a generic triac.

Values

Surface-mount triacs are typically rated between 2A to 25A of switched AC current (RMS), the higher-current versions being as large as 10mm square. The necessary gate trigger voltage may range from 0.7V to 1.5V. Through-hole packages may be capable of slightly higher currents (up to 40A), with gate trigger voltages of 1V to 2.5V being common.

As noted previously, the majority of triacs are restricted to relatively low frequency switching, 60Hz being very common.

Abbreviations in datasheets are likely to include:

- V_{DRM} or V_{RRM} Peak repetitive reverse off-state voltage. The maximum reverse voltage that the component will withstand in its "off" state without experiencing damage or allowing current to pass.

- V_{TM} The maximum voltage difference between A1 and A2, measured with a short pulse width and low duty cycle.
- V_{GT} Gate trigger voltage necessary to produce the gate trigger current.
- I_{DRM} Peak repetitive blocking current (i.e., maximum leakage).
- I_{GM} Maximum gate current.
- I_{GT} Minimum gate trigger current.
- I_H Holding current.
- I_L Latching current.
- $I_{T(RMS)}$ On-state RMS current. The maximum value passing through the component on a continuous basis.
- I_{TSM} Maximum non-repetitive surge current. Specified at a stated pulse width, usually 60 Hz.
- T_C Case temperature, usually expressed as an acceptable range.
- T_J Operating junction temperature, usually expressed as an acceptable range.

What Can Go Wrong

Like other semiconductors, a triac is heat sensitive. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, a value for triggering current is valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Low-Temperature Effects

Significantly higher gate current will be required by a triac operating at low temperatures. It is quite possible that the component will need twice as much current at 25° C compared with 100° C, junction temperature. If the triac receives insufficient current, it will not turn on.

Wrong Type of Load

If an **incandescent lamp** is replaced with a **fluorescent light** or **LED area lighting**, a preexisting triac may no longer work as a dimmer. Fluorescent lamps will have some inductance, and may also provide a capacitive load, either of which will interfere with the normal behavior of a triac.

The light output of an LED varies very differently compared with the light output of an incandescent bulb, in response to reduction in power. Therefore an LED should be dimmed using pulse-width modulation that is appropriate for its output characteristics. A triac is generally not suitable.

Wrongly Identified Terminals

A triac is often thought of as a symmetrical device, because it is designed to switch AC current using either positive or negative voltage at the gate. In reality, its behavior is asymmetrical, and if it is installed “the wrong way around” it may function erratically or not at all.

Failure to Switch Off

As already noted (see “Charge Storage” on page 23), a triac will tend to suffer from *charge storage* if there is insufficient time between the end of one half-cycle and the beginning of the next. A component that works with a resistive load may cease to function if it is used, instead, to power an inductive load.

solid-state relay

4

A **solid-state relay** is less-commonly referred to by its acronym, *SSR*. It is sometimes regarded as an **optocoupler**, but in this encyclopedia the two components have separate entries. An optocoupler is a relatively simple device consisting of a light source (usually an **LED**) and a light sensor, in one package. It is used primarily for isolation rather than to switch a high current. A solid-state relay can be thought of as a substitute for an electromagnetic **relay**, usually has additional components in its package, and is intended to switch currents of at least 1A.

A component that works like a solid-state relay but only switches a 5V (or lower) logic signal may be referred to as a *switch*, even though it is entirely solid-state. This type of component is included in this entry because it functions so similarly to a solid-state relay.

OTHER RELATED COMPONENTS

- **electromagnetic relay** (see Volume 1)
- **optocoupler** (see Chapter 5)

What It Does

A **solid-state relay** (SSR) is a semiconductor package that emulates an electromagnetic **relay** (see Volume 1). It switches power on or off between its output terminals in response to a smaller current and voltage between its input terminals. Variants can switch AC or DC and may be controlled by AC or DC. An SSR functions as a *SPST switch*, and is available in normally open or normally closed versions. SSRs that function as an SPDT switch are relatively unusual and actually contain more than one SSR.

No single schematic symbol has been adopted to represent a solid-state relay, but some alternatives are shown in Figure 4-1:

Top

An unusually detailed depiction of an SSR that switches DC current using MOSFETS. Symbols for this device often omit the diodes on the output side and may simplify the MOSFET symbols.

Bottom left

An SSR that uses an internal **triac** to switch AC. The box labeled 0x indicates that this is a *zero-crossing* relay, meaning that it switches when alternating voltage crosses the 0V level from positive to negative or negative to positive.

Bottom right

A generic SSR, showing a symbol for a normally open relay, although whether it is designed for AC or DC is unclear.

Advantages

- Great reliability and long life.
- No physical contacts that are vulnerable to arcing and erosion or (under extreme conditions) that could weld themselves together.

- Very fast response, typically 1µs on and 0.5µs off.
- Very low power consumption on the input side, as low as 5mA at 5VDC. Many solid-state relays can be driven directly from logic chips.
- Lack of mechanical noise.
- No contact bounce; a clean output signal.
- No coil that would introduce back EMF into the circuit.
- Safe with flammable vapors, as there is no sparking of contacts.
- Often smaller than a comparable electromagnetic relay.
- Insensitive to vibration.
- Safer for switching high voltages, as there is complete internal separation between input and output.
- Some variants work with input control voltages as low as 1.5VDC. Electromagnetic relays typically require at least 3VDC (or more, where larger relays are required to switch higher currents).

Disadvantages

- Less efficient; its internal impedance introduces a fixed-value voltage drop on the output side (although this may be negligible when switching higher voltages).
- Generates waste heat in its "on" mode, in accordance with the voltage drop.
- Passes some *leakage* current (usually measured in microamps) on the output side when the relay is supposed to be "off."
- A DC solid-state relay usually requires observation of polarity on the output side. An electromagnetic relay does not.
- Brief voltage spikes on the input side, which would be ignored by a slower electromagnetic relay, may trigger a solid-state relay.
- More vulnerable than an electromagnetic relay to surges and spikes in the current that is switched on the output side.

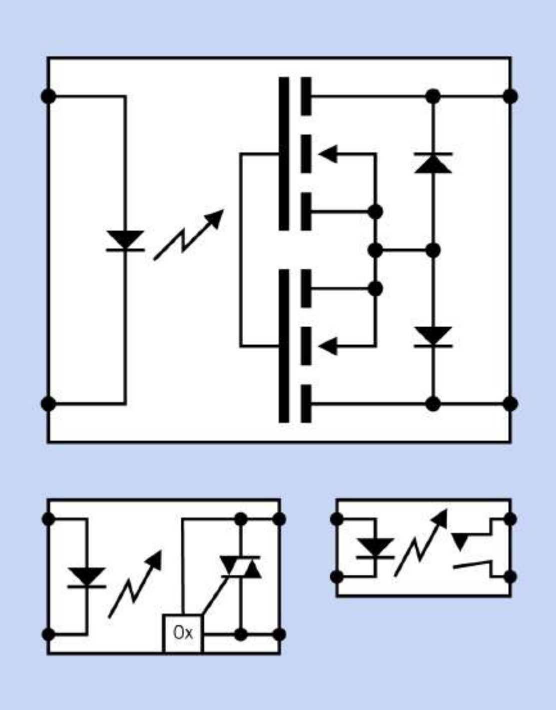

Figure 4-1. *Schematic symbols for solid-state relays have not been standardized. See text for details.*

How It Works

Almost all modern SSRs contain an internal **LED** (light-emitting diode, see Chapter 22) which is switched on by the control input. Infrared light from the LED is detected by a sensor consisting of one or more **phototransistors** or **photodiodes**. In a relay that controls DC current, the sensor usually switches a **MOSFET** (see Volume 1) or an **SCR** (silicon-controlled rectifier—see Chapter 1). In relays that control AC current, a **triac** (see Chapter 3) controls the output. Because the input side and the output side of the SSR are linked only by a light signal, they are electrically isolated from each other.

The MOSFETs require so little power, it can be provided entirely by light falling on an array of 20 or more photodiodes inside the SSR package.

Typical solid-state relays are shown in Figures 4-2 and 4-3.

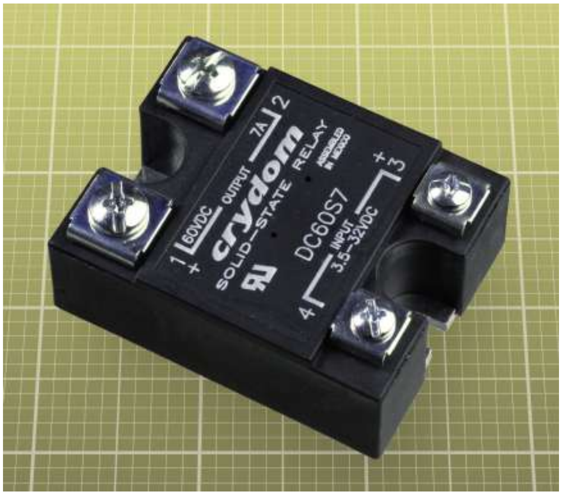

Figure 4-2. *A solid-state relay capable of switching up to 7A DC. See text for a detailed description.*

The Crydom DC60S7 accepts a control voltage ranging from 3.5VDC to 32VDC, with a typical input current of less than 3mA. Maximum turn-on time is 0.1ms and maximum turn-off time is 0.3ms. This relay can switch up to 7A and tolerates a surge of up to twice that current. It imposes a voltage drop of as much as 1.7VDC, which can become a drawback when switching voltages that are significantly lower than its maximum 60VDC. The electronics are sealed in thermally conductive epoxy, mounted on a metal plate approximately 1/8" thick which can be screwed down onto an additional heat sink.

The Crydom CMX60D10 tolerates a more limited range of control voltages (3VDC to 10VDC) and requires a higher input current of 15mA at 5VDC. However, its very low maximum on-state resistance of 0.018Ω imposes a much smaller voltage drop of less than 0.2 volts when passing 10A. This results in less waste heat and enables a *single-inline package* (SIP) without a heat sink. The CMX60D10 weighs 0.4 ounces, as opposed to the 3 ounces of the DC60S7. Relays from other manufacturers use similar packaging and have similar specifications.

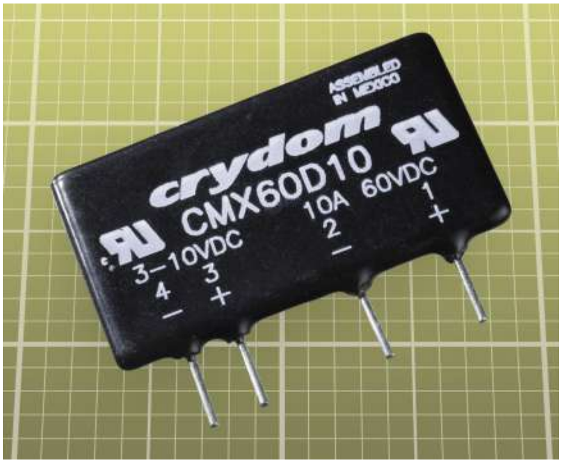

Figure 4-3. *A solid-state relay capable of switching up to 10A. Its lower internal resistance results in less waste heat and enables a smaller package. See text for a detailed description.*

Variants

Many solid-state relays have protective components built into the package, such as a varistor on the output side to absorb transients. Check datasheets carefully to determine how much protection from external components may be necessary when switching an inductive load.

Instantaneous versus Zero Crossing

A *zero crossing* SSR is one that (a) switches AC current and (b) will not switch "on" until the instant when the AC voltage crosses through 0V. The advantages of this type are that it does not have to be built to switch such a high current, and creates minimal voltage spike when the switching occurs.

All SSRs that are designed to switch AC will wait for the next voltage zero crossing before switching to their "off" state.

NC and NO Modes

Solid-state relays are SPST devices, but different models may have a normally closed or normally open output. If you require double-throw operation, two relays can be combined, one normally closed, the other normally open. See Figure 4-4. A few manufacturers combine a normally closed relay and a normally open relay in one package, to emulate a SPDT relay.

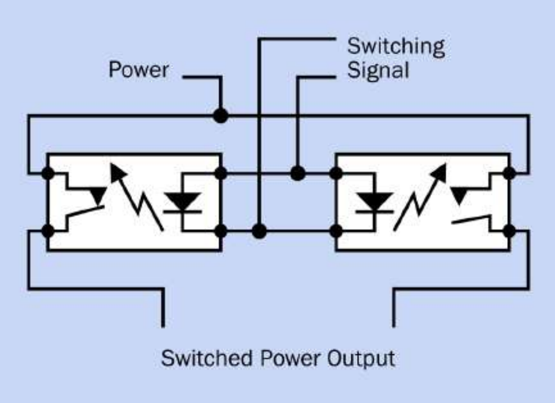

Figure 4-4. *A normally closed solid-state relay can be paired with a normally open solid-state relay to emulate a SPDT switch. This combination is available in a single package from some manufacturers.*

Packaging

High-current solid-state relays are often packaged with screw terminals and a metal base that is appropriate for mating with a *heat sink*. Some are sold with heat sinks integrated. Spade terminals and crimp terminals may be optional. The Crydom DC60S7 shown in Figure 4-2 is an example. This type of package may be referred to as *industrial mount*.

Lower-current solid-state relays (5A or less), and those with a very low output resistance, may be packaged with single-inline pins for through-hole mounting in circuit boards.

Solid-State Analog Switch

DIP packaging may be used for solid-state relays that are designed for compatibility with the low voltages and currents of logic chips. This type of component may be referred to simply as a *switch*. The 74HC4316 is an example, pictured in Figure 4-5.

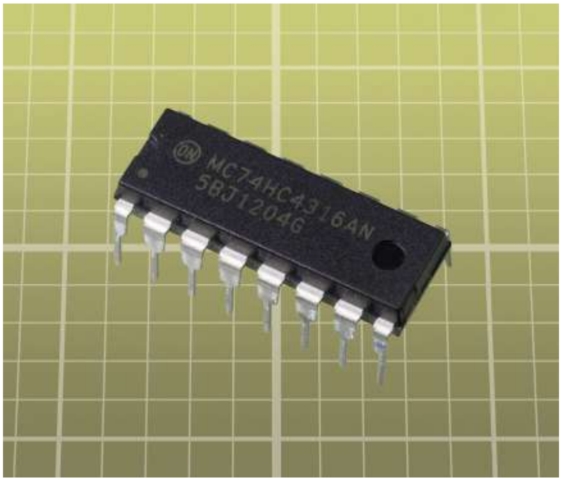

Figure 4-5. *This DIP package contains four "switches" that function as solid-state relays but are restricted to low voltages and currents, compatible with logic chips. See text for details.*

Typically the control voltage and the switched voltage are limited between +7V and −7V, with a maximum output current of 25mA. Each internal switch has its own Control pin, while an additional Enable pin forces all switches into an "off" state if its logic state is high. The simplified functionality of this component is illustrated in Figure 4-6, without showing internal optical isolation.

The "on" resistance of each internal pathway will be approximately 200Ω when the component is powered with +5VDC on the positive side and 0VDC on the negative side. This resistance drops to 100Ω if the negative power supply is -5VDC.

If all of the outputs from the chip are shorted together, it functions as a **multiplexer** (see Chapter 16). In fact, this type of switch component is often listed in catalogs as a multiplexer, even though it has other applications.

Because the component tolerates equal and opposite input voltages, it is capable of switching AC.

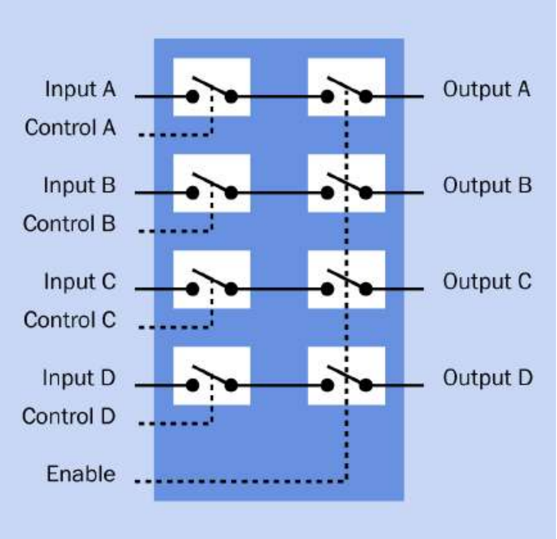

Figure 4-6. *The functionality of a chip containing four solid-state analog switches. A high state on a Control pin closes its associated switch. The Enable pin must be held low for normal operation; a high Enable state forces all the switches into the "off" position. If the outputs are tied together, this component can function as a multiplexer.*

Values

Industrial-mount solid-state relays typically can switch currents ranging from 5A to 500A, with 50A being very common. The higher-current relays mostly require DC control voltage; 4V to 32V are typical, although some versions can go much higher. They contain an SCR or triac to switch AC.

Smaller solid-state relays in SIP, DIP, or surface-mount packages often use MOSFETs on the output side, and are often capable of switching up to 2A or 3A. Some can switch either AC or DC, depending on the way the output is wired. The LED on the input side may require as little as 3mA to 5mA for triggering.

How to Use It

Solid-state relays find their primary uses in telecommunications equipment, industrial control systems and signalling, and security systems.

The component is very simple externally. Power on the input side can come from any source capable of delivering the voltage and current specified by the manufacturer, and any device that doesn't exceed maximum current rating can be connected to the output side, so long as provision is made for suppressing back-EMF from an inductive load, as shown in Figure 4-7. Often a solid-state relay can be substituted directly for an electromagnetic relay, without modifying the circuit.

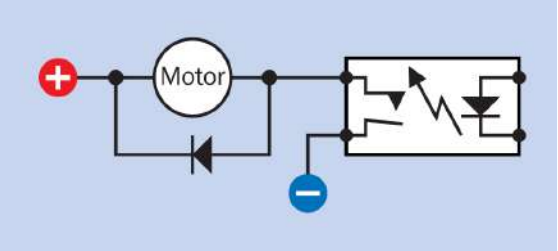

Figure 4-7. *Use of a diode around an inductive load, to protect a solid-state relay from back-EMF.*

Solid-state relays are heat sensitive, and their rating for switching current will diminish as their temperature increases. Manufacturer datasheets will provide specific guidance. Using a heat sink will greatly improve the performance. Bear in mind that the relay generates heat continuously while it is in its "on" mode—about 1 watt per ampere.

Because it requires so little current on the input side (typically no more than 15mA), a solid-state relay can usually be driven directly by chips such as *microcontrollers* that would not be able to activate an equivalent electromagnetic relay.

Applications may take advantage of the solid-state relay's reliability, immunity to vibration, lack of contact sparking, freedom from coil-induced surges on the input side, and lack of contact bounce on the output side. A solid-state relay is ideal within digital equipment that is sensitive to power spikes. It may switch a fuel pump that handles volatile, flammable liquids, or a wastewater pump in a basement subject to flooding (where long-term zero-maintenance reliability is necessary, and contact corrosion could be a risk in electromagnetic relays). Small solid-state relays can switch motors in robots or appliances where vibration is common, and are often used in arcade games.

What Can Go Wrong

Overheating Caused by Overloading

Relays must be *derated* when used at operating temperatures above the typical 20 or 25º C for which their specification applies. In other words, the sustained operating current must be reduced, usually by an amount such as 20% to 30% for each 10-degree increase in ambient temperature. Failure to observe this rule may result in failure of the component. Burnout may also occur if a high-current solid-state relay is used without a heat sink, or the heat sink isn't big enough, or *thermal compound* is not applied between the solid-state relay and the heat sink.

Overheating Caused by Bad Terminal Contact

If the screw terminals on the output side of a high-current solid-state relay are not tightened sufficiently, or if there is a loose spade terminal, or if a crimped connection isn't crimped tightly enough, the poor contact will create electrical resistance, and at high currents, the resistance will create heat, which can cause the solid-state relay to overheat and burn out.

Overheating Caused by Changing Duty Cycle

If a high-current solid-state relay is chosen for an application where it is in its "on" state only half the time, but the application changes during product development so that the solid-state relay is in its "on" state almost all the time, it will have to dissipate almost twice as much heat. Any time the duty cycle is changed, heat should be considered. The possibility of the relay being used in an unconventional or unexpected manner should also be considered.

Overheating Caused by Component Crowding

Overheating increases dramatically when components are tightly crowded. At least 2cm (3/4") should be allowed between components.

Overheating in Dual Packaging

When a package contains two solid-state relays, the additive effects of the heat created by each of them must be considered.

Reverse-Voltage Burnout

Because a solid-state relay is more sensitive to *back-EMF* than an electromagnetic relay, greater care should be used to protect it from reverse voltage when switching inductive loads. A *protection diode* should be used, and a *snubber* can be added between its output terminals, if it is not included inside the relay package.

Low Voltage Output Current May Not Work

Unlike electromagnetic relays, solid-state relays require some voltage on the output side to enable their internal operation. If there is no voltage, or only a very low voltage, the SSR may not respond to an input. The minimum voltage required on the output side is usually specified in a datasheet.

To test a solid-state relay, apply actual voltages on input and output sides and use a load such as an incandescent light bulb. Merely applying a meter on the output side, set to measure continuity, may not provide sufficient voltage to enable the relay to function, creating the erroneous impression that it has failed.

Inability to Measure AC Output

When a multimeter is used to test continuity across the output of an AC-switching solid-state relay of zero-crossing specification, the meter will generate enough voltage to prevent the solid-state relay from finding zero voltage across its output terminals, and consequently the solid-state relay won't switch its output.

Relay Turns On but Won't Turn Off

When a solid-state relay controls a relatively high-impedance load such as a small **solenoid** (see Volume 1) or a **neon bulb** (see Chapter 19), the relay may switch the device on but will seem unable to switch it off. This is because the leakage current of the solid-state relay, in its "off" state, may be just enough to maintain the load in its "on" state.

If an SSR containing a triac is used erroneously to switch DC, it will not be able to switch off the current.

Relays in Parallel Won't Work

Two solid-state relays usually cannot be used in parallel to switch twice as much current. Because of small manufacturing variances, one relay will switch on a moment before the other. When the first relay is on, it will divert the load current away from the second relay. The second relay needs a small amount of current on its output side, to function. Without any current, it will not switch on. This means the first relay will pass the total current without any help from the second relay, and will probably burn out, while the second relay does nothing.

Output Device Doesn't Run at Full Power

A solid-state relay imposes a voltage reduction on its output side. This will be a fixed amount, not a percentage. When switching 110V, this difference may be negligible; when switching 12V, it may deliver only 10.5V, which represents enough of a drop to cause a motor or a pump to run noticeably more slowly. The internal switching device inside the relay (MOSFET, triac, SSR, or bipolar transistor) will largely determine the voltage drop. Check the manufacturer's datasheet before using the relay.

Solid-State Relays and Safety Disconnects

Because a solid-state relay always allows some leakage in its "off" state, it can still deliver a shock when used to switch high voltages. For this reason, it may not be suitable in a safety disconnect.

optocoupler

5

Sometimes known as an *optoelectronic coupler*, *opto-isolator*, *photocoupler*, or *optical isolator*.

A **solid-state relay** is sometimes referred to as an **optocoupler**, but in this encyclopedia it has a separate entry. An optocoupler is a relatively simple device consisting of a light source (usually an LED) and a light sensor, both embedded in one package. It is used primarily for isolation rather than to switch a high current. A solid-state relay can be thought of as a substitute for an electromagnetic **relay**, usually has additional components in its package, and is intended to switch currents of at least 1A.

OTHER RELATED COMPONENTS

- **electromagnetic relay** (see Volume 1)
- **solid-state relay** (see Chapter 4)

What It Does

An optocoupler allows one section of a circuit to be electrically isolated from another. It protects sensitive components, such as logic chips or a microcontroller, from voltage spikes or incompatible voltages in other sections of a circuit. Optocouplers are also used in medical devices where a patient has to be protected from any risk of electric shock, and are used in devices which conform with the MIDI standard for digital control of music components.

In Figure 5-1, three possible applications for an optocoupler are suggested:

Top
: The output from a logic chip passes through an optocoupler to an inductive load such as a relay coil, which may create voltage spikes that would be hazardous to the chip.

Center
: The noisy signal from an electromagnetic switch passes through an optocoupler to the input of a logic chip.

Bottom
: The low-voltage output from a sensing device on a human patient passes through an optocoupler to some medical equipment, such as an EEG machine, where higher voltages are used.

Internally, an optocoupler works on the same principle as a **solid-state relay**. An LED is embedded on the input side, shining light through an interior channel or transparent window to a sensing component that is embedded on the output side. Because the only internal connection is a light beam, the input and output of the optocoupler are isolated from each other.

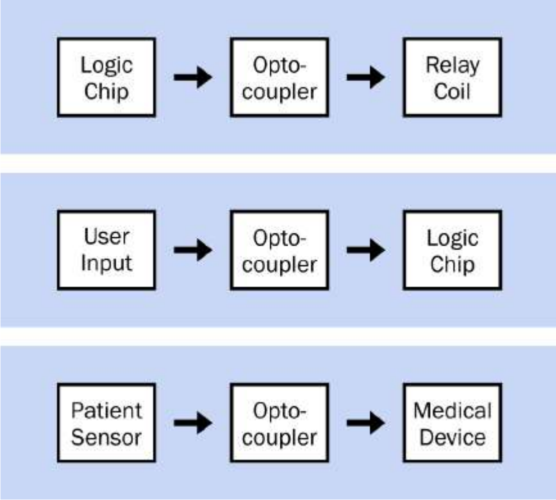

Figure 5-1. *Possible applications for a photocoupler. See text for details.*

Isolation transformers were traditionally used for this purpose prior to the 1970s, when optocouplers became competitive. In addition to being smaller and cheaper, an optocoupler can also pass slow-changing signals or on-off DC states which a transformer would ignore.

More recently, inductive and capacitive coupling components have become available in surface-mount packages that are competitive with optocouplers for high-speed data transfer. They also claim to be more durable. Because of the gradual reduction in output from an LED, the performance of an optocoupler degrades over time, and is typically rated for up to 10 years.

How It Works

The LED in an optocoupler almost always emits light in the near-infrared part of the spectrum, and is matched to the sensitivity of a **phototransistor**, or a **photodiode**, or (less often) a **photoresistor** that provides the output. Photosensitive **triacs** and **SCRs** are also sometimes used.

The most common type of optocoupler uses a bipolar phototransistor with an open-collector output. Schematic symbols for this type are shown in Figure 5-2:

Top left
The most common generic form.

Top right
Two diodes on the input side allow the use of alternating current.

Center left
An additional terminal allows addition of bias to the photosensitive base of the output transistor, to reduce its sensitivity.

Center right
An Enable signal can be used as the input to the NAND, suppressing or enabling the output.

Bottom left
A *photodarlington* allows higher emitter current.

Bottom right
Relatively uncommon, and is also used for a **solid-state relay**.

In each symbol, the diode is an LED, and the zigzag arrow indicates light that is emitted from it. A pair of straight arrows, or wavy arrows, may alternatively be used.

An optocoupler in through-hole DIP format is shown in Figure 5-3.

An *optical switch* can be thought of as a form of optocoupler, as it contains an LED opposite a sensor. However, the LED and the sensor are separated by an open slot, to allow a thin moving object to pass through, interrupting the light beam as a means of detecting the event. It is categorized as a *sensor* in this encyclopedia, and will be found in Volume 3.

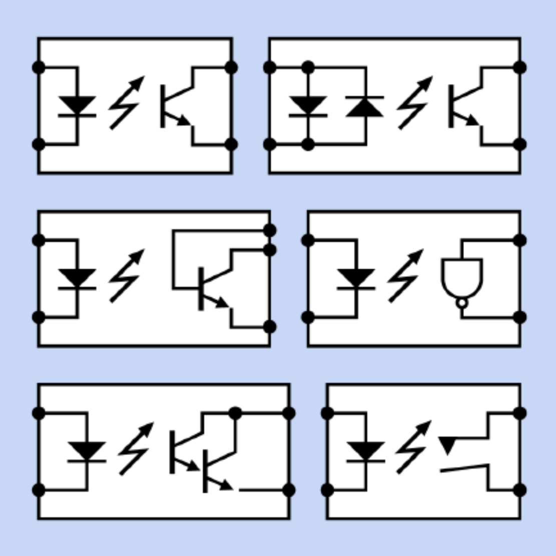

Figure 5-2. *Six variants of schematic symbols that may be used to represent an optocoupler. See text for details.*

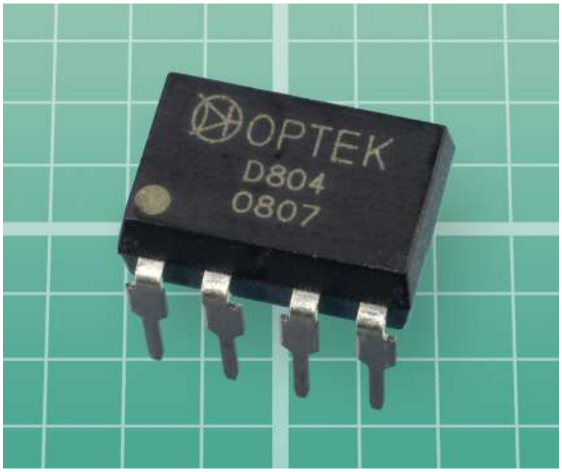

Figure 5-3. *An optocoupler in through-hole 8-pin DIP format.*

Variants

Internal Sensors

Historically, a photoresistor (often referred to as a *photocell*) was the first type of sensor to be used. It has a more linear response than other sensor types, but its response is much slower. It is still found in audio applications. "Stomp box" pedals used by guitarists typically contain an optocoupler that employs a photoresistor, and are valued for their linearity and their immunity from the mechanical wear, contamination, and "scratchiness" that builds up over time in a **potentiometer**. Optocouplers also eliminate *ground loops*, which tend to be induced by small differences in ground potential, introducing hum or buzz in audio applications when two or more power supplies are tied together.

The type of optocoupler that contains a photoresistor and is commonly used by musicians was initially trademarked as a *Vactrol*, and that term is still used generically. Vactrols have also been used to provide audio compression in telephone voice networks, and were used in photocopiers and photographic exposure meters, but these applications are now obsolete.

Photoresistors are becoming uncommon because of their cadmium content, which is unlawful in many countries (especially in Europe) because of its environmental toxicity.

A photodiode provides the fastest response time in an optocoupler, limited primarily by the characteristics of the LED that shines light upon it. A *PIN diode* can respond in less than a nanosecond; its acronym is derived from its fabrication from p-type and n-type semiconductor layers with an *intrinsic* layer connecting them. This additional layer can be responsive to light. When the diode is slightly reverse-biased, a photon entering the intrinsic layer can dislodge an electron, enabling current to flow. The reverse bias enlarges the active area and enhances the effect. In this mode, the PIN acts like a photoresistor, appearing to reduce its resistance in response to light.

When the PIN is used in *photovoltaic* mode, no bias is applied, and the component actually generates a small voltage (less than 1VDC), like a solar cell, in response to incoming light. Where an optocoupler uses a MOSFET on its output side, as many as 30 photodiodes may be connected in series to develop the necessary threshold voltage to trigger the transistor. This arrangement is common in **solid-state relays**.

A bipolar phototransistor is a slower-speed device but is still usually capable of a 5µs response time or better. Its open collector requires external voltage and a *pull-up resistor* to deliver a positive output so long as the phototransistor is nonconductive. When the LED turns on, the phototransistor sinks current, effectively creating a low output. In this way, the optocoupler functions like an *inverter*, although some variants include a noninverting output.

Basic Optocoupler Types

An optocoupler with *high linearity* will respond more proportionally to variations in current to its LED. *High Speed* optocouplers are used for high-frequency data transfer. *Logic-output* optocouplers have a clean high/low output transition, rather than an *analog output*, which varies with fluctuations in the input. Linearity is of importance only where an optocoupler is being used to transmit an analog signal with some fidelity. Some logic-output optocouplers provide the function of a Schmitt trigger on their output side.

While optocouplers are available in various package formats, the DIP style with six or eight pins remains popular, providing sufficient physical space for the LED, the sensor, and a light channel, while providing good electrical isolation.

Variants may have two or four optocouplers combined in one package. A *bidirectional* optocoupler may consist of two optocouplers in parallel, inverted with respect to each other.

Values

In a datasheet, the characteristics of primary importance in an optocoupler are:

- CTR is the Current Transfer Ratio, the ratio of maximum output current to input current, expressed as a percentage. With a bipolar phototransistor output, 20% is a typical minimum CTR. With a photodarlington output, the CTR may be 1,000% but the bandwidth is much lower—the response time may be measured in microseconds rather than nanoseconds. Optocouplers with a photodiode output have a very low CTR, and their output is in microamps. However, they provide the most linear response.
- $V_{CE(MAX)}$ is the maximum collector-emitter voltage difference (in an optocoupler with a bipolar phototransistor output). Values from 20 to 80 volts are common.
- V_{ISO} is the maximum potential difference, in VDC, between the two sides of the optocoupler.
- I_{MAX} is the maximum current the transistor can handle, generally in mA.
- Bandwidth is the maximum transmittable signal frequency, often in the range of 20kHz to 500kHz.

The LED in an optocoupler typically requires 5mA at a forward voltage of 1.5V to 1.6V.

The maximum collector current on the output side of an optocoupler is unlikely to be higher than 200mA. For higher output currents, a solid-state relay should be considered. It provides photo-isolation on the same basis as an optocoupler, but high-current versions tend to be considerably more expensive.

How to Use It

The primary purpose of an optocoupler is to provide protection against excessive voltage—from transients, incompatible power supplies, or equipment with unknown characteristics. If a device is designed to be plugged into a USB port on a computer, for instance, the computer may be isolated via an optocoupler.

A series resistor for the LED is not built into most optocouplers, because the value of the resistor will depend on the input voltage that is used. Care must be taken to determine what the maximum voltage on the input side will be, and a series resistor should be chosen to reduce current appropriately. Allowance should be made for some degradation in the performance of the LED over time.

For an optocoupler with an open-collector output, a pull-up resistor is necessary in most applications. The voltage from the optocoupler must be matched to the input requirements of other components, and the collector current must remain within the specified limits. Some trial and error in resistor selection may be necessary.

In Figure 5-4, a schematic shows typical component values in a test circuit using a pushbutton as input. The separation of the two power supplies is emphasized by the different color shades used for the positive and negative symbols. Although the input side and the output side of an optocoupler may be used with a common ground, this defeats its purpose in providing complete isolation between the sections of the circuit.

The pinouts for an optocoupler must be checked carefully in the manufacturer's datasheet. While the input for an 8-pin DIP chip is usually applied to pins 2 and 3, the output pin functions are not standardized and will vary depending on the internal configuration of the chip. An optocoupler such as the Optek D804, with an enable function using an internal NAND gate, requires its own power supply.

Where an optocoupler allows an external connection to the base of its internal bipolar output phototransistor, reverse bias applied to this pin will decrease the sensitivity of the optocoupler but can increase its immunity to noise on the input side.

What Can Go Wrong

Overload conditions on the input or the output side of an optocoupler will be the most likely cause of failure.

Age

Because optocouplers are typically rated for only 10 years of average use, the age of a component may cause it to fail.

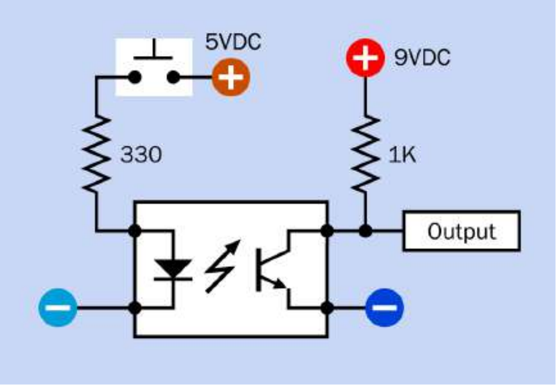

Figure 5-4. *Typical values for a series resistor (to protect the LED) and pull-up resistor (to control current and voltage on the output side) in an optocoupler test circuit.*

LED Burnout

Because the LED is hidden inside the component, there is no immediate indication of its performance. A meter can be inserted into the circuit on the input side to determine if current is passing through the LED. A meter set to measure volts can be used to discover whether the LED is imposing a normal voltage drop. While significant overload will cause immediate burnout, slightly exceeding the current rating of the LED may have more pernicious consequences, as the LED may not fail until days or weeks have passed without any sign of trouble. The failure of the optocoupler will be unexpected and difficult to determine.

Transistor Burnout

Here again the damage caused by excessive current may be progressive, occurring over a prolonged period. The easiest way to test an optocoupler that may have failed is by removing it from the circuit. A socketed DIP package is preferable for this purpose.

6

comparator

Although a **comparator** has the same schematic symbol as an **op-amp**, their applications differ and they are described in separate sections of this encyclopedia.

This entry describes an analog comparator. A *digital comparator* is very different, being a logic chip that compares two binary numbers that can be referred to as A and B. Outputs from the chip indicate whether A>B or A<B or A=B. The digital comparator does not have an entry in this encyclopedia.

OTHER RELATED COMPONENTS

- **op-amp** (see Chapter 7)

What It Does

A **comparator** is an integrated circuit chip that compares a variable voltage on one input pin with a fixed, reference voltage on a second input pin. Depending which voltage is higher, the output from the comparator will be high or low.

The output will make a clean transition between two fixed values, even if the input is infinitely variable. Thus the comparator can function as an *analog-digital converter*, as shown in Figure 6-1.

Because the output voltage range can be adjusted up or down independently of the input range, a comparator can also function as a *voltage converter*.

Hysteresis

If positive feedback is added through external resistors, *hysteresis* can be introduced. We may imagine a *hysteresis zone* extending above and below the reference voltage level. Small input variations that occur within the zone will be ignored. The comparator only reacts when the input signal emerges above or below the hysteresis zone. When the input signal returns into the hysteresis zone, this event also will be ignored. The concept is illustrated graphically in Figure 6-2. A circuit to create hysteresis is shown in Figure 6-10.

How It Works

The schematic symbol for a comparator is shown in Figure 6-3. This seems identical to the symbol for an op-amp, described in Chapter 7, but an op-amp is traditionally a *dual-voltage* device using positive and negative power sources that are equal and opposite, in addition to a zero value midway between the two. Modern comparators mostly use a conventional single voltage, and therefore the negative symbol used in comparator schematics throughout this section of the encyclopedia represents 0 volts. It has the same meaning as the ground symbol found in many schematics elsewhere.

The two inputs to a comparator are described as *inverting* and *noninverting* (for reasons explained later). Confusingly, these are identified with plus and minus symbols inside the triangle that represents the component. These plain black-and-white symbols have nothing to do with the power supply.

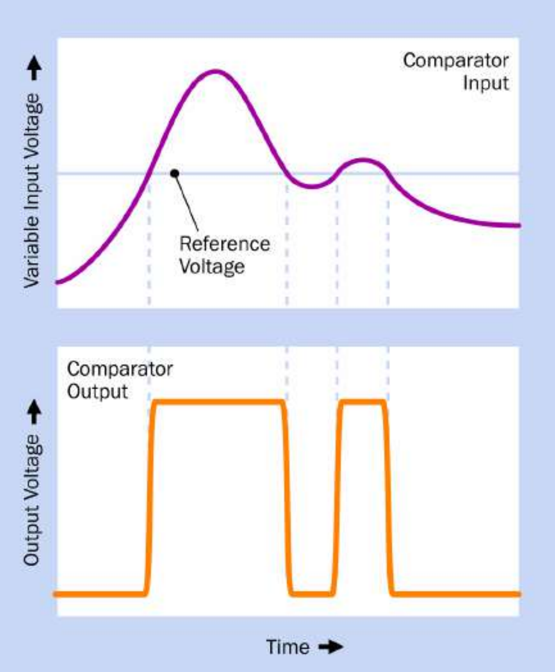

Figure 6-1. *The basic behavior of a simple comparator is shown here.*

Often, in schematics, the power supply is not shown, because it is assumed to be present. However, all comparators require a power supply in order to function.

The basic internal and external connections used in conjunction with a typical comparator are shown in Figure 6-4.

The potentiometer at top left is often a *trimmer*, to fine-tune a reference voltage. The variable input can come from a sensor or any other device capable of delivering a voltage up to the limit set by V1.

The output is often an *open collector* from an internal bipolar transistor, as shown in the figure.

Note that as many as three different voltages can be used, as indicated by the different colors associated with V1, V2, and V3. However, they must share a common ground to enable the comparator to make valid comparisons.

When the noninverting input exceeds the voltage of the inverting input, the output transistor goes into its "off" state, and blocks current from an external *pullup resistor*. Because the current from the resistor now has nowhere else to go, it is available to drive other devices attached to the comparator output, and the output appears to be high.

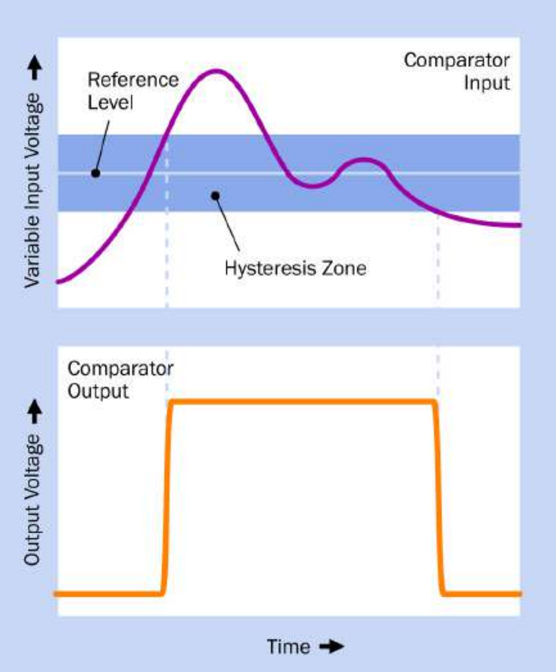

Figure 6-2. *The performance of a comparator shown in the previous figure can be modified by the addition of hysteresis. Small variations that occur within the hysteresis zone are ignored.*

When the noninverting input falls below the voltage of the inverting input, the transistor becomes conductive, and sinks almost all the current from the pullup resistor, assuming other devices attached to the output have a relatively high impedance. The output from the comparator now appears to be low.

This can be summed up as follows:

- When a variable voltage is applied to the noninverting input, and it rises *above* the reference voltage applied to the *inverting input*, the output transistor turns off, and the comparator delivers a *high* output.
- When a variable voltage is applied to the noninverting input, and it falls *below* the reference voltage applied to the *inverting input*, the output transistor turns on, and the comparator delivers a *low* output.

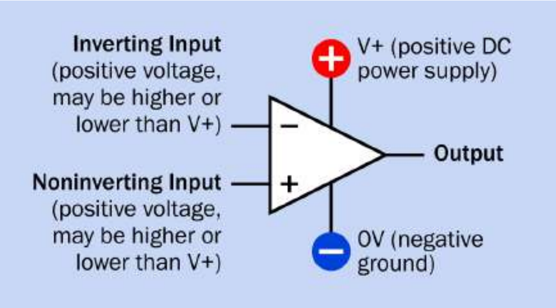
Inverting Input (positive voltage, may be higher or lower than V+)
Noninverting Input (positive voltage, may be higher or lower than V+)
V+ (positive DC power supply)
Output
0V (negative ground)
−
+

Figure 6-3. *The symbol for a comparator is the same as the symbol for an op-amp, even though they often require different types of power supply and their functions are significantly different.*

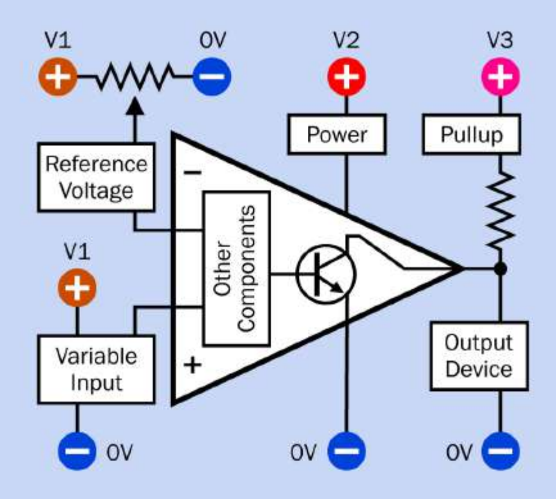
V1
0V
V2
V3
Reference Voltage
Power
Pullup
V1
Other Components
Variable Input
Output Device
0V
0V
0V
−
+

Figure 6-4. *Connections to a comparator, and their functions.*

If the reference voltage and the variable voltage are swapped between the input pins, the behavior of the comparator is reversed. This relationship is illustrated in Figure 6-5. When a voltage transition is applied to the *inverting* input, the transition is *inverted* at the output.

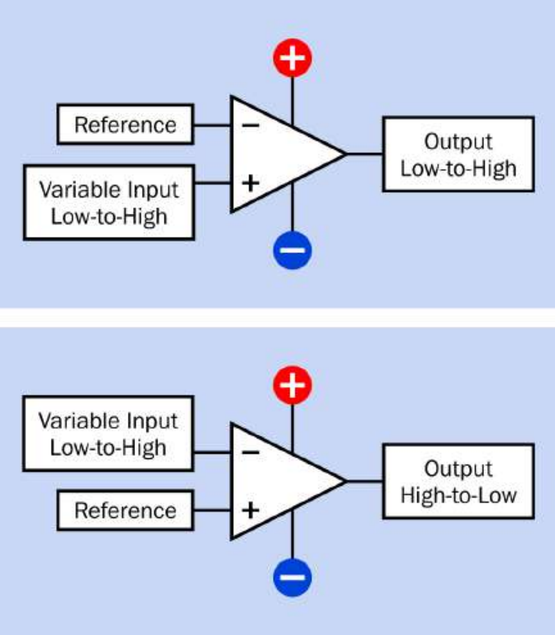
Reference
Variable Input Low-to-High
Output Low-to-High
Variable Input Low-to-High
Reference
Output High-to-Low
−
+

Figure 6-5. *Depending which input pin is used for the reference voltage, and which input carries a variable voltage, the comparator output either follows the variable voltage or inverts it.*

Placement of the plus and minus signs inside the comparator symbol may vary. Most often, the minus sign is above the plus sign, as shown in all the schematics here. Sometimes, however, for convenience in drawing a schematic, the plus sign may be shown above the minus sign. Regardless of their placement, the plus sign always identifies the noninverting input, and the minus sign always identifies the inverting input. To avoid misinterpretations, schematics should be inspected carefully.

Where a power supply for the comparator is shown, the positive side is always attached to the upper edge of the symbol, while 0V ground is always attached to the lower edge.

Differences from an Op-Amp

Saturation versus linearity

The output of a comparator is optimized for saturation (high or low, without intermediate levels, using positive feedback). The output of an op-amp is optimized for linearity (faithful reproduction of nuances in the input, using negative feedback).

Output mode

The majority of comparators have open-collector outputs (or open-drain outputs in CMOS devices) where the voltage is established by a pullup resistor. This can be adjusted for compatibility with other components, especially 5VDC logic. Only a minority have push-pull amplifier outputs that require no pullup resistor. By comparison, among op-amps, a push-pull output that functions as a voltage source is the traditional default.

Faster response

A comparator responds more quickly than an op-amp to changes in input voltage, if the op-amp is used in the role of a comparator. The comparator is primarily a switching device, not an amplifier.

Hysteresis

This is generally desirable in a comparator, for reasons already explained, and some components are designed with hysteresis built in. This feature is undesirable in an op-amp, as it degrades sensitivity.

Open-loop operation

(i.e., without feedback) this can be used with a comparator. An op-amp is intended for use in closed-loop circuits (i.e., with feedback), and manufacturers will not specify its performance in an open loop.

As previously noted, a comparator usually requires a single-voltage power supply, while an op-amp often requires a dual-voltage power supply.

Variants

Where a comparator uses a MOSFET output transistor, it may have an open-drain output, which requires a pullup resistor, as with an open-collector output.

Some comparators have a *push-pull* output, capable of supplying output current (usually a small amount). In these instances, no pullup resistor is necessary or desirable. The output voltage range will be closest to *rail-to-rail* values (i.e., the range of the power supply) where MOSFETs are used for the output, as MOSFETs impose a smaller voltage drop than bipolar transistors.

The advantage of an open collector (or open drain) relative to a push-pull output is that it allows the output voltage to be set independently of the power supply voltage. Another advantage is that multiple outputs can be connected in parallel, as in a *window comparator* circuit (described below).

Some comparators incorporate a reference voltage on the chip, based on the power supply to the chip. In this case, a separate reference voltage does not have to be supplied, and the component will draw less current.

Many chips are available containing two or more comparators. This is often expressed as the number of *channels* in the component. A *dual comparator* typically allows two different voltage sources for the outputs of the comparators. They will share the same 0V ground, however. Chips such as the LM139 and LM339 contain four comparators, and are available in through-hole or surface-mount formats. They have become a generic choice, costing less than $1 apiece.

An LM339 comparator chip is shown in Figure 6-6. This is a quad chip, meaning that it contains four comparators. They share a common power supply. The chip is TTL and CMOS compatible, is typically powered by 5VDC, but can be driven by up to 36VDC. The input differential voltage range also extends up to 36V.

6AALTEM
LM339N

Figure 6-6. *The LM339 quad comparator chip, shown here, was introduced long ago but remains widely used.*

Some comparators have an internal *latch* function that is accessed by a dedicated pin. The latch-enable signal forces the comparator to assess its inputs and hold an appropriate output which can then be checked by other components.

Values

In a datasheet, V_{IO} (also referred to as V_{OS}) is the *input offset voltage*. This is a small voltage, in addition to the reference voltage, which the comparator will require to *toggle* its output in either direction, up or down. Figure 6-7 shows this graphically. V_{IO} sets the limit of *resolution* of the comparator, which will not respond unless the input voltage exceeds the reference voltage by this amount. A smaller value for V_{IO} is better than a larger value. Common values for V_{IO} range from 1mV to 15mV. The actual offset voltage tends to vary between one sample of a component and another. V_{IO} is the maximum allowed value for a component.

Because the comparator will not respond until the reference voltage is exceeded by V_{IO}, the output pulse width will be narrower than if the comparator reacted at the point where the variable voltage input was precisely the same as the reference voltage.

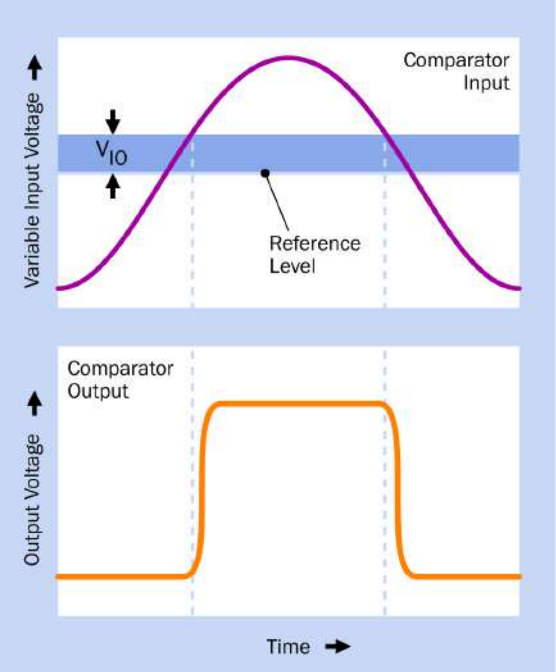

Figure 6-7. *The input offset voltage is the very small voltage that a comparator requires, additional to the reference input voltage, before it will toggle its output from low to high or high to low.*

V_{TRIP+} and V_{TRIP-} are the *rising* and *falling voltages*, respectively, that will trip the comparator output where the comparator exhibits some innate hysteresis without an external feedback loop. They are also referred to as *Lower State Transition Voltage* (LSTV) and *Upper State Transition Voltage* (USTV).

V_{HYST} is the *hysteresis range* defined as V_{TRIP+} minus V_{TRIP-}. The relationship is shown graphically in Figure 6-8.

A_{VD} is the *voltage gain* of a comparator, in which the letter "A" can be thought of as meaning "amplification." The gain is measured as a maximum ratio of output voltage to input voltage. Typically it ranges from 40 to 200.

Supply voltage for modern comparators is often low, as the components are used in surface-mount format for battery-powered devices where low power consumption is a primary concern. Thus, 3VDC is common as a power requirement, and 1.5VDC comparators are available. Still, older chips can use as much as 35VDC.

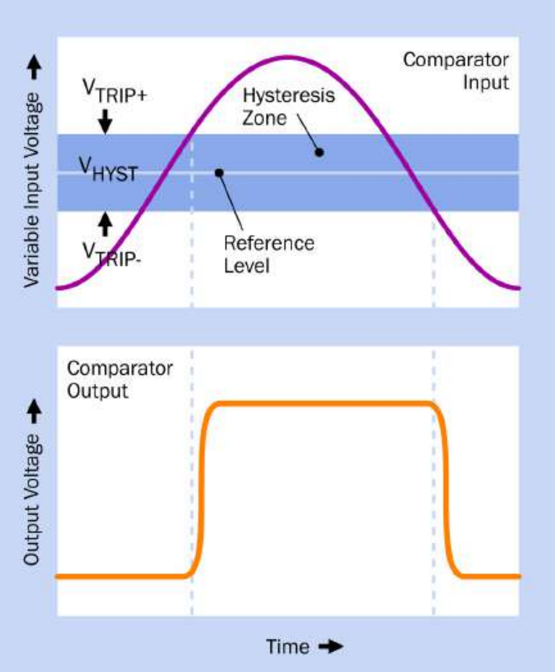

Figure 6-8. *The value of V_{TRIP} shows the hysteresis in a comparator—the range of input voltages, relative to the reference voltage, in which it will not respond.*

Supply current can range from 7mA down to below 1μA.

I_{SINK} is the recommended typical or maximum *sink current* that the component will tolerate, if it has an open-collector output. This value should be considered in relation to the power dissipation, P_D.

The *propagation delay* in a comparator is measured from the moment when an input (usually a square wave) reaches the triggering value, to the time when the consequent output reaches 50% of its final value.

When a comparator is driving CMOS logic using a 5VDC power supply, a typical value for a pullup resistor is 100K. It does not have to be lower, because CMOS has such a high input impedance.

How to Use It

In Figure 6-1, a hypothetical comparator responds immediately when the input voltage equals the reference voltage value. However, this is an idealized scenario. A magnified view, in Figure 6-9, suggests that the comparator is likely to respond with *jitter* when the input signal is very close to the reference voltage, because of tiny variations in heat, current, and other variables. This jitter will cause significant problems if the comparator is driving a device such as a relay, directly or indirectly.

Hysteresis eliminates this uncertainty around the transition level of the input, by telling the comparator to ignore small irregularities in the input voltage. Hysteresis is also useful in many situations where larger variations in a sensor input should be ignored. In Figure 6-2, for instance, suppose that the input voltage comes from a temperature sensor. The small bump in the right-hand section of the curve is probably unimportant; it could be caused by someone opening a door, or a person's body heat in brief proximity to the sensor. There's no point in responding to every little event of this type. In this application, the larger, longer-term temperature trend is what matters, and significant hysteresis is appropriate.

Also, if a comparator is being used as a thermostat, to switch a heating system on and off, we do not want the comparator to respond as soon as the temperature rises just a small amount. The heating system should run for a while before it elevates the temperature beyond the hysteresis zone.

The usual way to create hysteresis is with *positive feedback*. In Figure 6-10, a connection from the output of the comparator runs back through a 1M potentiometer to the variable (noninverting) input. The effect that this has is to reinforce the input voltage with the output voltage, as soon as the comparator input goes high. Now the input can diminish slightly without switching off the comparator. But if the input declines significantly, even the feedback from the output voltage won't be sufficient to maintain the variable input at a higher level than the reference voltage. (Remember, the "high" output voltage from the comparator is a fixed value; it does not change in proportion with the input voltage.) Consequently, the output toggles to low. Now the variable input is deprived of help from the comparator output, so it will be low enough that it has to rise considerably to toggle the comparator back on again. During that period, once again, small variations will be ignored.

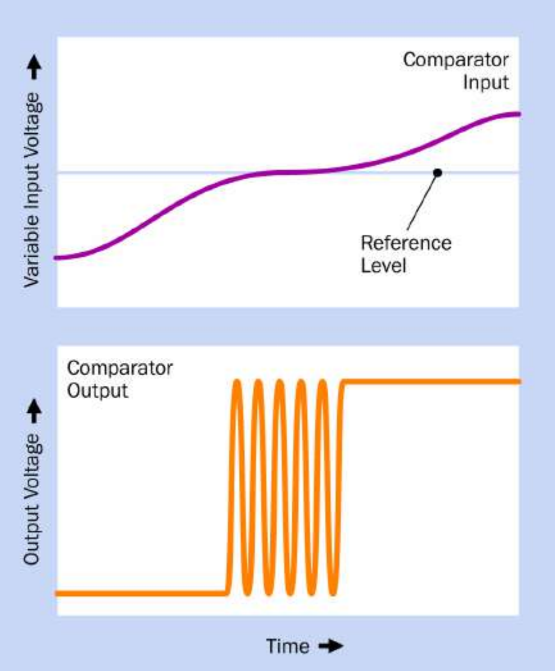

Figure 6-9. *In real-world applications, tiny variations where the variable input voltage crosses the reference voltage can induce jitter in the output from a comparator that has no hysteresis.*

In the schematic, a **phototransistor** (PT1, at left) is in series with a 3.3K resistor to adjust its voltage output to a suitable range. A 1M potentiometer at upper-left is wired as a voltage divider, so that it can establish a reference level that matches the light level that we wish to detect with the phototransistor.

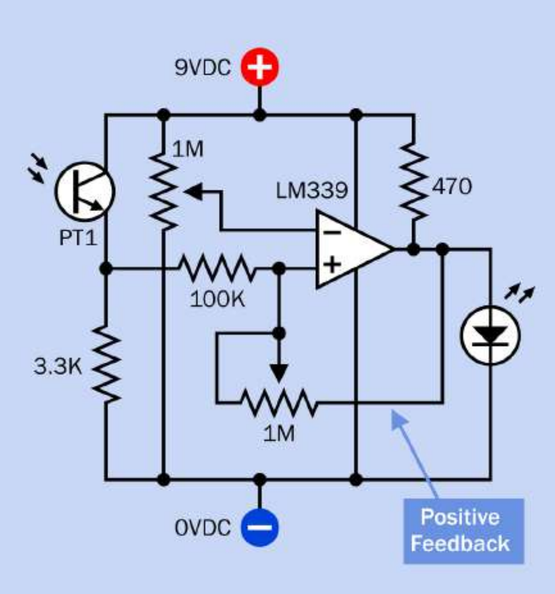

Figure 6-10. *A simple circuit to achieve hysteresis with positive feedback to the variable input of a comparator.*

The 470Ω resistor is the pullup resistor, which protects the LED from excessive current. The lower 1M resistor adjusts the amount of positive feedback, which determines the width of the hysteresis zone.

Values for components may have to be adjusted depending on the supply voltage, the variable input voltage, and other factors. But the principle will remain the same. Note that in the example shown, all the positive voltage sources are identical. In practice, different voltages could be used, so long as they share a common ground.

AND gate

A set of open-collector comparators can function jointly as an AND gate, when their outputs are tied together with one pullup resistor. So long as all the output transistors are nonconductive, the output will be high. If just one comparator toggles into conductive mode, the output will be low. See Figure 6-11.

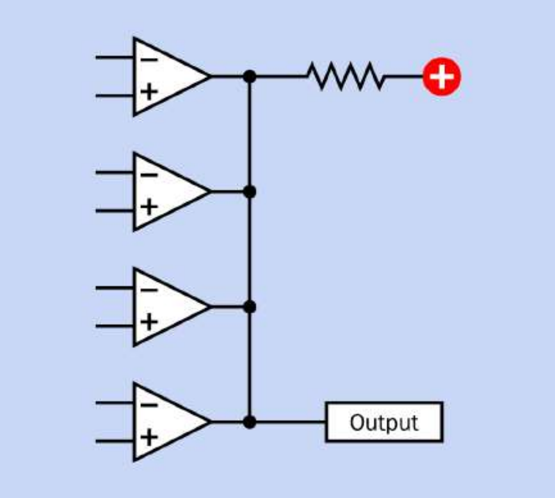

Figure 6-11. *If the outputs of multiple open-collector comparators are tied together with a suitable pullup resistor, they will function as an AND gate.*

Bistable Multivibrator

If positive feedback to the noninverting input of the comparator is sufficiently high, a voltage almost at 0V ground will be required to counter the high output from the comparator—after which, a voltage almost equal to the supply voltage will be needed to turn it back on. In other words, the comparator is behaving like a *bistable multivibrator*, or **flip-flop**.

Relaxation Oscillator

A *relaxation oscillator*, which is a form of *astable multivibrator,* can be created using direct positive feedback in combination with delayed negative feedback. In Figure 6-12, positive feedback goes to the noninverting input, as before, but negative feedback also passes through a 220K resistor to the inverting input of the comparator. A 0.47µF capacitor initially holds the inverting input low, while the capacitor charges. Gradually the capacitor reaches and exceeds the charge on the noninverting input, so the output from the comparator toggles to its low state. This means that its internal transistor is now sinking current, and it discharges the capacitor. Because the noninverting input is being held at a voltage midway between supply and ground by the two 100K resistors forming a voltage divider, eventually the voltage on the inverting input controlled by the capacitor falls below the noninverting voltage, so the cycle begins again.

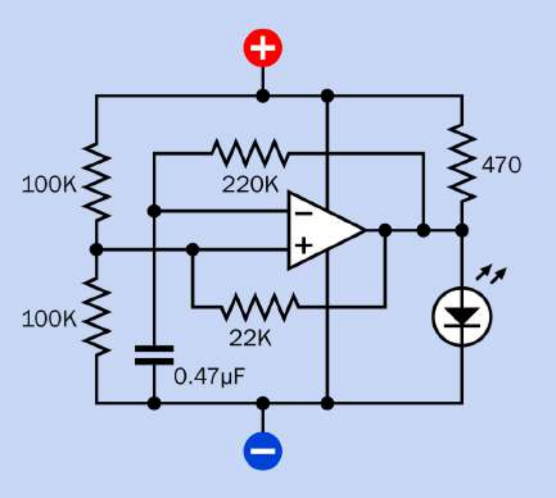

Figure 6-12. *A comparator can be used to create a relaxation oscillator.*

Level Shifter

Where a comparator is used simply to change the level of an input voltage, it can be referred to as a *level shifter*. An example of a level shifter is shown in Figure 6-13, in which a high/low 3VDC logic input is converted to a high/low logic output at 5VDC.

Window Comparator

A *window comparator* is a circuit (not a single component) that will respond to input voltages that deviate outside an acceptable “window” of values. In other words, the circuit responds anytime the variable input is either unacceptably low or unacceptably high.

An example could be an alarm that will sound if a temperature is either too low or too high. In Figure 6-14, two comparators are used to create a window comparator circuit, both sharing a variable voltage input from a sensor. A voltage divider is necessary to establish a higher voltage limit at the noninverting input of the upper comparator, while a separate voltage divider would establish a lower limit at the inverting input of the lower comparator. If an alarm has an appropriate resistance, it can be used instead of a pull-up resistor. The alarm will sound when the output from either comparator is low, which happens if the inverting input has a higher voltage than the non-inverting input.

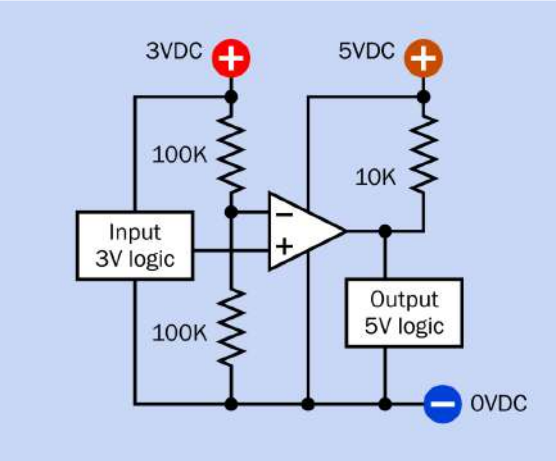

Figure 6-13. *A comparator can be used to convert high and low 3V logic inputs into high and low 5V logic outputs.*

Other Applications

As previously noted, a comparator can be used as a simple *analog-digital converter*. It has "one bit" accuracy (i.e., its output is either high or low).

A comparator can be used as a *zero point finder* when its variable voltage input is attached to an AC signal. The output from the comparator will be toggled whenever the AC signal passes through zero volts. The output will be a square wave (approximately) instead of a sine wave.

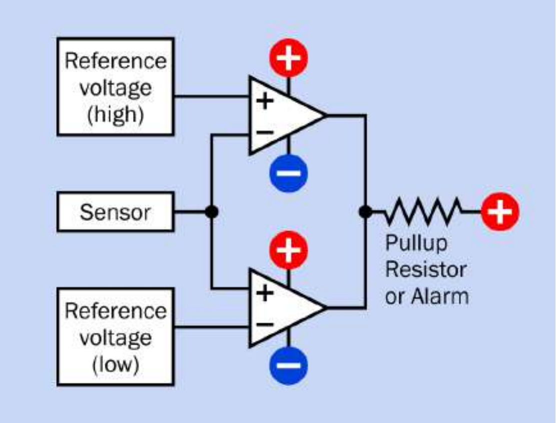

Figure 6-14. *A basic, simplified circuit for a window comparator. See text for details.*

A *continuous converter* changes its output promptly in response to a change in input. This requires continuous current consumption. Because many applications only need to check the output from a comparator at intervals, power can be saved by using a *clocked* or *latched* comparator.

What Can Go Wrong

Oscillating Output

The high input impedance of a comparator is vulnerable to stray electromagnetic fields. If the conductors leading to and from the comparator are relatively long, the output can couple capacitively with the input during voltage transitions, causing unwanted oscillations.

The commonly recommended solution to this problem is to add 1µF bypass capacitors to the power supply on either side of the comparator. However, some manufacturers recommend alternatives such as introducing just a small amount of hysteresis, or reducing the value of input resistors to below 10K.

If a chip contains multiple comparators, and one of them is unused, one of its input pins should be tied to the positive side of the voltage supply while the other should be tied to 0V ground, to eliminate the possibility of an oscillating output.

Confused Inputs

A comparator will function if its two inputs are swapped accidentally, but its high and low output states will be the inverse of what is expected. Also, if positive feedback is used, transposed inputs can create oscillations. Because the comparator symbol may appear in a schematic with the noninverting input either below or above the inverting input, the inputs are easily transposed by accident.

One way to remember which way the inputs should be connected is to use the mnemonic: "plus, high, positive." The *plus* input creates a *high* output when the input becomes *more positive* than the reference voltage on the other input. The opposite is less intuitively obvious: the *minus* input creates a *high* output when the input becomes *more negative* than the reference voltage on the other input.

Wrong Chip Type

Different comparators offer different outputs: open collector, open drain, and push-pull. While open collector and open drain function similarly, the pullup resistor value is likely to be different in each case. If a push-pull output is mistakenly connected as if it is open collector or open drain, it will not work correctly, if at all. Different types of comparators must be sorted and stored in clearly labeled bins.

Omitted Pullup Resistor

It is relatively easy to forget to include the pullup resistor on an open-collector output. In this case, when the transistor inside the comparator is in its nonconductive state, the output pin will be floating, with an indeterminate voltage that will create confusing or random results.

CMOS Issues

As usual when using CMOS chips, it is bad practice to allow unconnected, floating inputs. This is an issue where a chip contains multiple comparators, some of which are not being used. The solution recommended by some manufacturers is to tie one input of an unused comparator to the supply voltage, and the other input of the same comparator to ground.

Erratic Output

If positive feedback is insufficient, the comparator output may show signs of *jitter*. Conversely, if the positive feedback is excessive, the comparator may get stuck in an on state or an off state. Feedback must be chosen carefully.

Swapped Voltages

A comparator is often capable of controlling an output voltage that is much higher than that of its power supply. Because both voltages are applied to different pins on the same chip, mistakes can be made quite easily. The chip is likely to be damaged if the voltages are swapped accidentally between the relevant pins.

Heat-Dependent Hysteresis

Remember that the voltages at which the comparator turns on and off will vary slightly with the temperature of the component. This *drift* should be tested by running the comparator at higher temperatures.

op-amp

7

Although a comparator has the same schematic symbol as an **op-amp**, their applications differ and they are described in separate sections of this encyclopedia.

The unabbreviated name for an op-amp is an *operational amplifier*, but this term is seldom used.

OTHER RELATED COMPONENTS

- **comparator** (see Chapter 6)

What It Does

An op-amp is an *operational amplifier* consisting of multiple transistors packaged in an integrated circuit chip. It senses the fluctuating voltage difference between two inputs, uses power from an external supply to amplify that difference, and uses *negative feedback* to ensure that the output is an accurate replica of the input. Its amplification can be adjusted by changing the values of two external resistors.

Op-amps were developed originally using vacuum tubes, for use in analog computers, before the era of digital computing. Their implementation in integrated circuits dates from the late 1960s, when chips such as the LM741 were introduced (lower-noise versions of it still being widely used today). Multiple op-amps in a single package were introduced in the 1970s.

An LM741 is shown in Figure 7-1. Inside the 8-pin, DIP package is a single op-amp.

How It Works

In alternating current, voltages deviate above and below a zero value, which is sometimes referred to as the *neutral* value. This occurs in domestic power supplies and in audio signals, to name two very common examples. A *voltage amplifier* multiplies the positive and negative voltage excursions, using an external power source to achieve this. Most op-amps are voltage amplifiers.

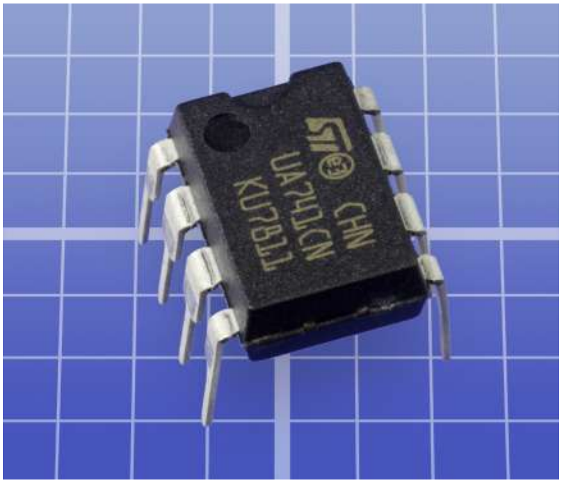

Figure 7-1. *The LM741, shown here, is still one of the most widely used op-amps.*

An ideal amplifier maintains a *linear relationship* between its input and its output, meaning that the output voltage values are a constant multiple of the input voltages over a wide range. This is illustrated in Figure 7-2, where the lower curve is a duplicate of the upper curve, the only difference being that its amplitude is multiplied by a fixed amount (usually much greater than shown here). The ratio is properly known as the *gain* of the amplifier, usually represented with letter A (for amplification).

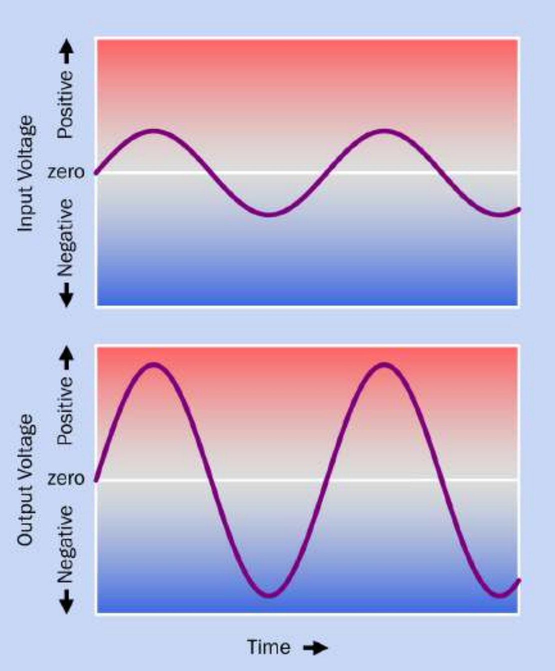

Figure 7-2. *In an ideal voltage amplifier, the output voltage will be a duplicate of the fluctuating input voltage, the only difference being that the amplitude of the output is multiplied by a fixed amount. This ratio is known as the gain of the amplifier.*

Figure 7-3 shows the triangular symbol for a generic single-input amplifier (not an op-amp). It may contain any number of components. The triangle almost always points from left to right, with its input on the left and its output on the right, and power attached above and below. This is often a *dual voltage* power supply, which is convenient for amplifying a signal that fluctuates above and below 0V. In some schematics, the power supply connections may not be shown, as they are assumed to exist.

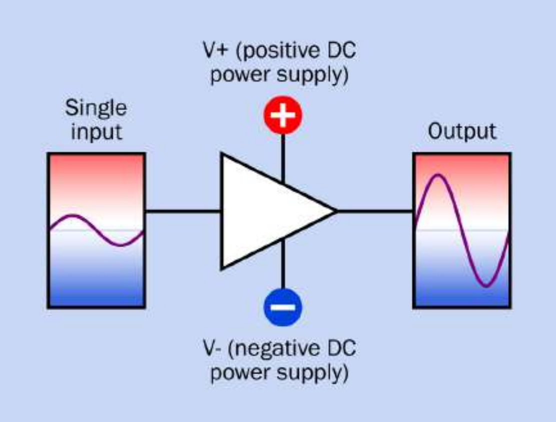

Figure 7-3. *The generic symbol for a single-input amplifier (not an op-amp), with the positive side of its power supply being equal and opposite in value to the negative side, and 0V being at the midpoint between them.*

- While the blue negative symbol is generally used throughout this encyclopedia to indicate 0V ground, it represents a voltage identified as V– in a dual voltage power supply, being equal in value but opposite in polarity to the positive side of the supply, V+. (Sometimes these voltages are indicated as V– and V+.)

The output from this imaginary generic amplifier is shown in the figure as a linear amplification of the input.

Dual Inputs

An op-amp has two inputs instead of one, and amplifies the voltage difference between them. Its symbol is shown in Figure 7-4. The upper input in this figure is held at 0V, midway between V+ and V–. Because the op-amp has so much gain, an accurate reproduction of its input would create an output exceeding the voltage of the power supply. Because this is not possible, the output tends to become saturated and consequently is *clipped* when it reaches its maximum value, as shown in the figure. The thumbnail graphs give only an approximate impression, as they are not drawn to the same scale.

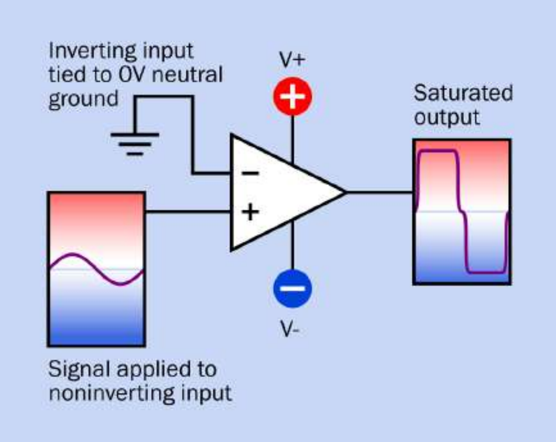
Inverting input tied to 0V neutral ground
V+
Saturated output
−
+
V-
Signal applied to noninverting input

Figure 7-4. *An op-amp has so much gain, its output will tend to saturate, producing a square wave regardless of the shape of the input.*

The small black plus and minus signs alongside the two inputs to the op-amp have nothing to do with the voltage supplied to the component. The “minus” input is properly referred to as the *inverting input* while the “plus” input is the *noninverting input*, in recognition of their functions.

The inputs are sometimes arranged with the minus above the plus, and sometimes with the plus above the minus. Schematics should be inspected carefully to note which arrangement is being used.

The positive and negative power connections to the op-amp may be omitted, but if shown, they always place V+ at the top, regardless of which way around the inputs are presented.

If a signal is applied to the noninverting input, while the inverting input is held at 0V ground, the op-amp provides an output in which the voltage is not inverted relative to the input.

If the input connections are swapped, so that the inverting input receives the incoming signal while the noninverting input is tied to 0V ground, the output from the op-amp is inverted (the gain remains the same). See Figure 7-5.

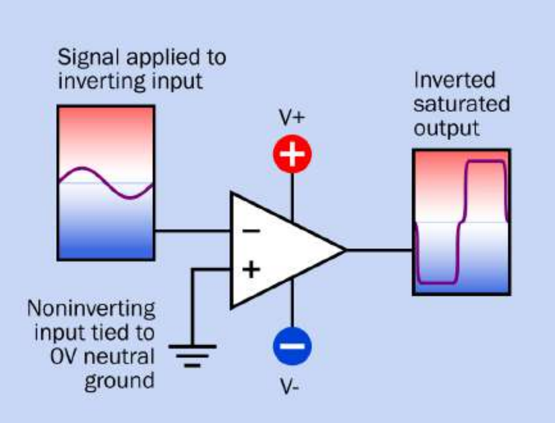
Signal applied to inverting input
V+
Inverted saturated output
−
+
Noninverting input tied to 0V neutral ground
V-

Figure 7-5. *When the incoming signal is applied to the inverting input of an op-amp while its noninverting input is held at 0V ground, the output is inverted.*

- An op-amp that is being used without any other components to moderate its output is functioning in *open loop* mode.

Negative Feedback

To create an output that is an accurate replica of the input, the op-amp must be brought under control with *negative feedback* to the input signal. This is illustrated in Figure 7-6. A resistor connects the output back to the inverting input, so that the input is automatically reduced to the point where the output is no longer saturated. The values of R1 and R2 will determine the gain of the op-amp, as explained in “How to Use It” on page 53. The op-amp is now functioning in its intended *closed loop* mode, meaning that the output is being tapped for feedback.

To obtain a linear output that is noninverted, connections are made as shown in Figure 7-7. The resistors form a voltage divider between the output and 0V ground, effectively increasing the comparison value on the inverting input.

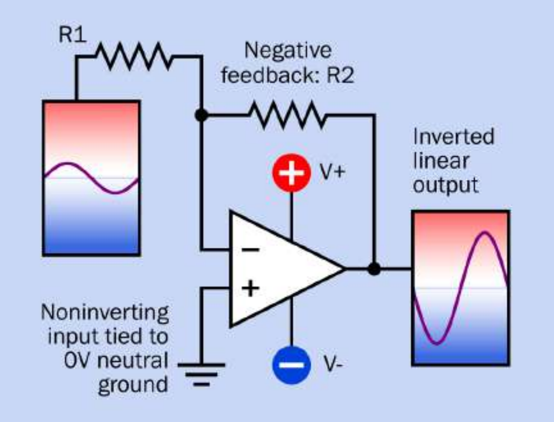

Figure 7-6. *A resistor applies negative feedback to the inverting input of an op-amp, and creates a linear output.*

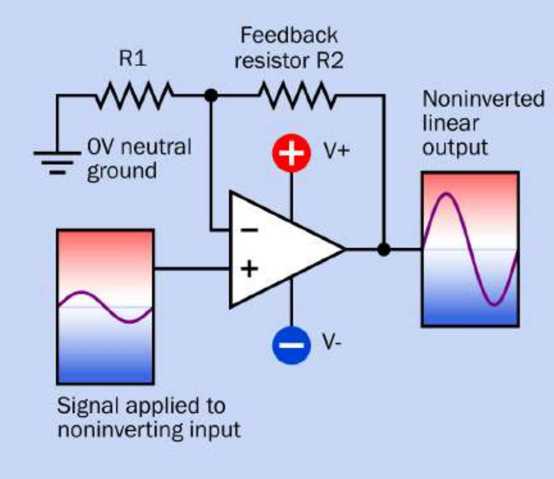
R1
Feedback resistor R2
Noninverted linear output
0V neutral ground
V+
−
+
V-
Signal applied to noninverting input

Figure 7-7. *Where the incoming signal is applied to the noninverting input, negative feedback is created by using a pair of resistors forming a voltage divider between the output and 0V ground.*

- Note that the gain of an op-amp is specific to a particular frequency range of AC signal. This is discussed in "How to Use It" on page 53.

Op-Amps and Comparators

A **comparator** can be regarded as a type of op-amp, and in fact an op-amp can be used as a comparator, comparing a variable DC voltage on one input with a reference voltage on another input. However, the two types of components have diverged in design to the point where they should be considered separately. The distinction is sufficiently important to have prompted Texas Instruments to issue an Application Report in 2001 titled "Op Amp and Comparators—Don't Confuse Them!".

Differences in function are summarized in the previous entry discussing comparators (see "Differences from an Op-Amp" on page 42).

Variants

Because op-amps are mostly low-current devices, they are widely available in very small surface-mount formats, in addition to the through-hole DIP packages which used to be more common.

Many chips are available containing two or more op-amps. This is often expressed as the number of *channels* in the component. A dual chip contains two op-amps, while a quad chip contains four op-amps. Usually all the op-amps in a chip share the same power supply. Bipolar or CMOS transistors may be used.

Because op-amps are widely available in dual and quad packages, it's quite common for a circuit designer to have one op-amp in a chip "left over." The designer may be tempted to use that spare unit as a comparator instead of installing an additional chip. To address this situation, some manufacturers offer hybrid op-amp chips containing an additional comparator. The Texas Instruments TLV2303 and TLV2304 are examples.

Values

The op-amps derived from 1970s designs often tolerate a wide range of power-supply voltages. Plus-or-minus 5VDC to plus-or-minus 15VDC is a common range. Modern op-amps are available that run from as little as 1VDC to as much as 1,000VDC.

Op-amps are available for frequencies ranging from 5KHz all the way up to 1GHz.

A “classic” op-amp such as the LM741, which is still widely used, will operate with a power supply ranging from plus-or-minus 5VDC to plus-or-minus 22VDC. Its output is rated for up to 25mA, and its input impedance is at least 2MΩ. The most current it will draw from an input is around 0.5µA.

V_{IO} is the *input offset voltage*. In an ideal component, the output from an op-amp should be 0V when its inputs have a voltage difference of 0V. In practice, the output will be 0V when the inputs differ by the offset voltage. V_{IO} is likely to be no greater than a couple of mV, and negative feedback can compensate for the offset.

V_{ICR} is the *common mode voltage range*. This is the range of input voltages that the op-amp will tolerate. This can never be more than the positive power supply voltage and will often be less, depending on the types of transistors that are used on the input side. If an input voltage goes outside the common mode voltage range, the op-amp will stop functioning.

V_{IDR} is the *input differential voltage range*—the maximum permissible difference between peak positive and peak negative input voltages. This is often expressed as plus-or-minus the power supply voltage, or slightly less. Exceeding the range can have destructive consequences.

I_B is the *input bias current*, averaged over the two inputs. Most op-amps have extremely high input impedance and consequently use very low input currents.

Slew rate at unity gain is the rate of change of the output voltage caused by an instantaneous change on the input side, when the output of the op-amp is connected directly back to the inverting input (during operation in noninverting mode).

How to Use It

In addition to being an amplifier for AC signals, an op-amp can serve as an oscillator, filter, signal conditioner, actuator driver, current source, and voltage source. Many applications require some understanding of the complexities of mathematics describing alternating current, which are not included this encyclopedia. Almost all the applications have a common starting point, however, which is to establish and control the gain of the feedback circuit.

Controlling the Gain

"A_{VOL} is the *open-loop voltage gain*, defined as the maximum voltage amplification that can be achieved when no feedback is applied from the output to an input. This remains constant until the AC frequency rises to a point known as the *breakover frequency*. If the frequency continues to rise, the maximum gain diminishes quite rapidly, until finally it terminates in 1:1 amplification at the *unity gain frequency*. This transition is shown by the orange line in Figure 7-8. The length of each purple line shows the frequencies which can be tolerated when the op-amp is used in closed-loop mode, and a negative feedback loop limits the gain. For example, where the gain is just 10:1, it can remain constant to just above 10KHz.

Note that both of the scales in this graph are logarithmic.

Calculating Amplification

So long as an op-amp is used within the boundaries of the graph, its voltage amplification can be controlled by choosing appropriate feedback and input resistors. If the op-amp is being used in noninverting mode, and R1 and R2 are placed as shown in Figure 7-7, the amplification ratio, A, is found approximately by the formula:

```
A = (approximately) 1 + (R2 / R1)
```

From this it can be seen that when R1 is very large compared with R2, the gain diminishes to near unity. If R1 becomes infinite and R2 is zero, the gain is exactly 1:1. This can be achieved by replacing R2 with a section of wire (theoretically of zero resistance) and omitting R1 entirely, as in Figure 7-9. In this configuration, the output from the op-amp should be identical with its input.

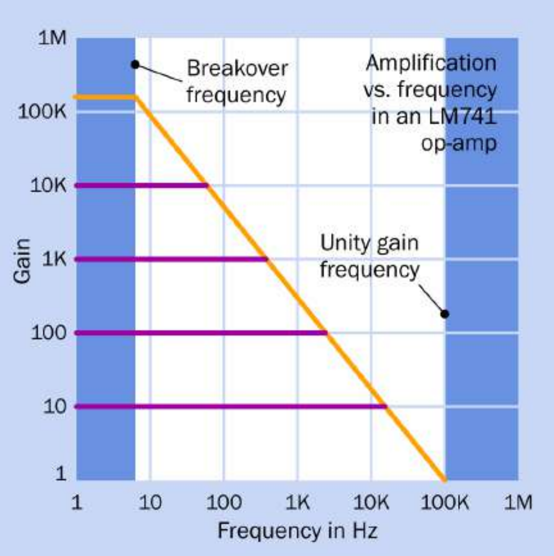

Figure 7-8. *Where each horizontal purple line meets the diagonal orange line, this is the maximum frequency that can be used without reduction in the maximum gain of an op-amp.*

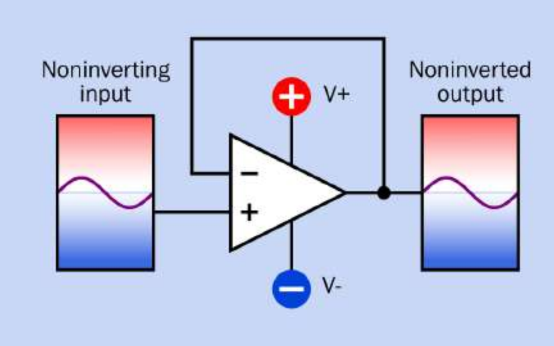

Figure 7-9. *While an op-amp is in noninverting mode, if the feedback resistor is replaced with a section of wire and the 0V ground connection is omitted entirely, the gain of the op-amp diminishes theoretically to 1:1.*

If the op-amp is being used in inverting mode, and R1 and R2 are placed as shown in Figure 7-6, then the voltage amplification ratio, A, is found approximately by the formula:

```
A = (approximately) -(R2 / R1)
```

- Note the minus sign. In inverting mode, gain is expressed as a negative number.
- In a practical circuit, at the expected frequency, the amplification factor established by choice of resistors should be no more than 20.
- An inverting circuit has a relatively low input impedance. For this reason, in most applications, a noninverting circuit is preferred.

Unintentional DC Voltage Amplification

Although the op-amp is intended primarily as an AC signal voltage amplifier, it will also amplify a DC difference between the voltages on its inputs. In the upper section of Figure 7-10, a positive DC offset is inverted and amplified to the point where the output is forced to its negative limit, and the signal is lost, because its fluctuations have been overwhelmed by the positive offset. A coupling capacitor (shown in the lower section of the figure) removes the DC voltage while passing the AC signal. The appropriate capacitor value will depend on the frequency of the signal.

Low-Pass Filter

An op-amp can facilitate a very simple low-pass audio filter, just by adding a capacitor to the basic inverting circuit previously shown in Figure 7-6. The filter schematic is shown in Figure 7-11. Capacitor C1 is chosen with a value that passes higher audio frequencies and blocks lower audio frequencies. Because the gain of the basic inverting circuit is approximately –(R2 / R1), the op-amp functions normally when the impedance of C1 is blocking the low frequencies, forcing them to pass through R2. Higher frequencies, however, are able to bypass R2 through C1, lowering the effective resistance of the feedback section of the circuit, thus reducing its gain. This way, the power of the op-amp is greatly reduced for higher frequencies compared with lower frequencies. A passive RC circuit could achieve the same effect, but would attenuate the signal, while the op-amp circuit boosts part of it.

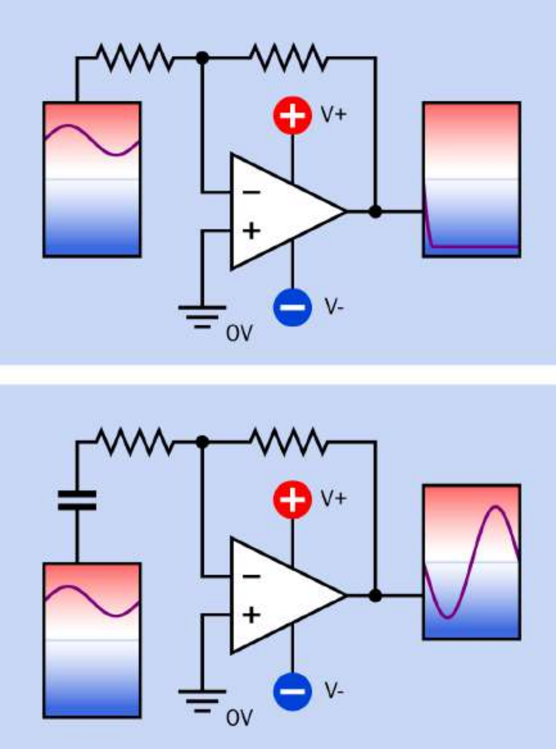

Figure 7-10. *The addition of a capacitor at the input of an op-amp is often necessary to prevent any DC voltage offset being amplified. In the upper section of this figure, a DC offset is large enough to force the inverted output to its negative limit, and the signal is completely lost.*

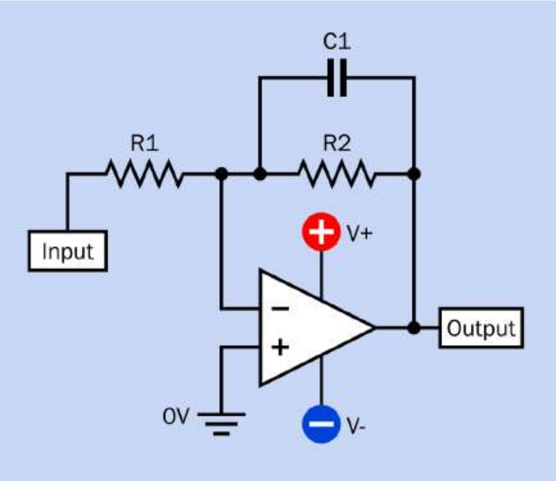

Figure 7-11. *A very basic low-pass filter, which works by allowing capacitor C1 to bypass resistor R2 at higher audio frequencies.*

High-Pass Filter

A simple high-pass filter can be created by adding a capacitor to the basic noninverting circuit previously shown in Figure 7-7. The filter schematic is shown in Figure 7-12. Once again capacitor C1 is chosen with a value that passes higher audio frequencies and blocks lower audio frequencies. Because the gain of the basic noninverting circuit is approximately 1 + (R2 / R1), the op-amp functions normally when the impedance of C1 is blocking the low frequencies, forcing them to pass through R1. Higher frequencies, however, are able to bypass R1 through C1, lowering the effective resistance of that section of the circuit, thus reducing the negative feedback and increasing the gain. This way, the power of the op-amp is increased for higher frequencies compared with lower frequencies. A passive RC circuit could achieve the same effect, but would attenuate the signal, while the op-amp circuit boosts part of it.

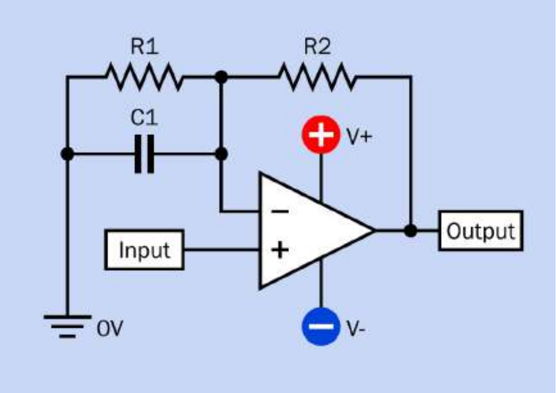

Figure 7-12. *A very basic high-pass filter, which works by allowing capacitor C1 to bypass resistor R1 at higher audio frequencies.*

Relaxation Oscillator

The schematic in Figure 7-13 is similar to the circuit shown in Figure 6-12 using a comparator. It functions as a *relaxation oscillator*, which is a form of *astable multivibrator*. The lower half of the circuit is a positive feedback loop that reinforces the output while the upper half of the circuit is charging the capacitor. Eventually the charge on the capacitor exceeds the voltage on the noninverting input of the op-amp, creating negative feedback that exceeds the positive feedback. The capacitor discharges and the cycle repeats. The component values in the figure should generate an output that runs at around 2Hz. Reducing the value of the capacitor will increase the frequency.

Single Power Source

A few op-amps are designed to work from single voltages, but they are a relatively small minority, and will clip the output signal if the input goes negative. Power supplies are readily available that provide multiple voltages such as +15VDC, 0V, and -15VDC. They are ideal for driving an op-amp—but may not be useful for any other components in the circuit. Can an op-amp that is designed for dual voltages be made to run from a single supply, such as 30VDC?

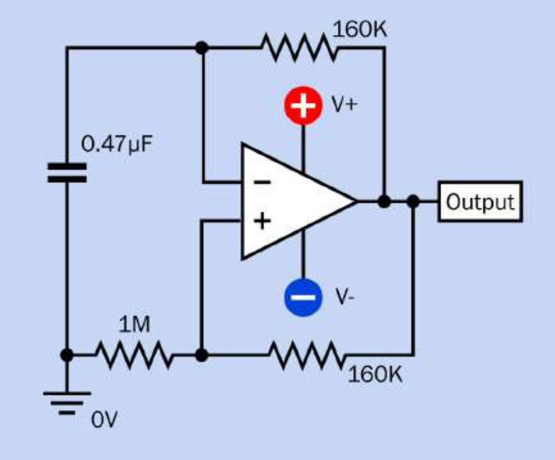

Figure 7-13. *A relaxation oscillator.*

This is relatively easy to do. The op-amp simply needs a potential difference to power its internal transistors, and 30VDC on the V+ pin with 0VDC on the V– pin will work just as well as +15VDC and -15VDC. However, referring back to Figure 7-6, if the op-amp is used in inverting mode, an intermediate voltage must be supplied to the noninverting input. Likewise, in noninverting mode, an intermediate voltage is necessary for one of the inputs, and must be half-way between the extremes of the power supply. If the supply is +15VDC and -15VDC, the midpoint is 0V. If the supply is 30VDC and 0V, the midpoint is 15VDC.

Because the inputs of an op-amp have a very high impedance and draw negligible current, the intermediate voltage can be provided with a simple voltage divider, as shown in Figure 7-14, where R3 and R4 should be no greater than 100K each. Their exact values are not important, so long as they are equal.

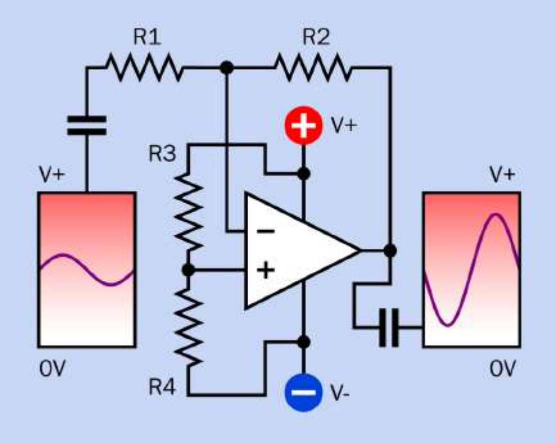

Figure 7-14. *A voltage divider, comprised of R3 and R4 in this schematic, can provide a voltage halfway between V+ and negative ground, enabling the op-amp to use just one power supply instead of two.*

A coupling capacitor should still be used on the input side, as shown, because there is no guarantee that the input signal will be centered precisely on 15V, and any offset will be amplified, potentially causing clipping of the signal. For similar reasons, a coupling capacitor is also added on the output side.

Offset Null Adjustment

Some op-amps provide two pins for *offset null adjustment*, which is a setup process to ensure that identical voltage on the two inputs will produce a null output. This is a way of compensating for any internal inconsistencies introduced during the manufacturing process.

To perform offset null adjustment, both input pins are connected directly to 0V ground, and the ends of a trimmer potentiometer (typically, 10K) are connected with the offset null pins, while the wiper of the potentiometer is centered and then connected with the negative power supply. The probes of a meter that is set to measure DC volts are placed between the output of the op-amp and 0V ground. The potentiometer is then adjusted until the meter shows a reading of 0VDC. A schematic is shown in Figure 7-15.

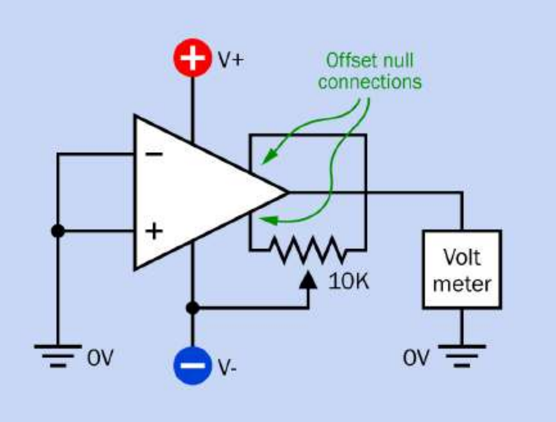

Figure 7-15. *Connections for making an offset null adjustment to an op-amp that allows this procedure.*

What Can Go Wrong

Power Supply Problems

Op-amps are especially vulnerable to reversed polarity in a power supply. If there is even a remote possibility of this occurring, a diode in series with one side of the supply can provide protection.

A more realistic concern is the destructive consequence of an input signal that exceeds the power supply voltage(s) of the op-amp. Even if the input is within the acceptable range, it can still cause permanent damage if it is applied before the op-amp powers up.

Bad Connection of Unused Sections

Multiple op-amps are often combined in a single package. If some of these "sections" remain unused, they will still receive power from the shared supply, and will attempt to function. If the inputs are left unconnected, they will pick up small stray voltages by capacitance or induction, and in the absence of negative feedback, the op-amp will create unpredictable outputs, consuming power and possibly interacting with other sections of the same chip. Figure 7-16 shows three incorrect options for addressing this problem, and one recommended option (derived from Texas Instruments Application Report SLOA067).

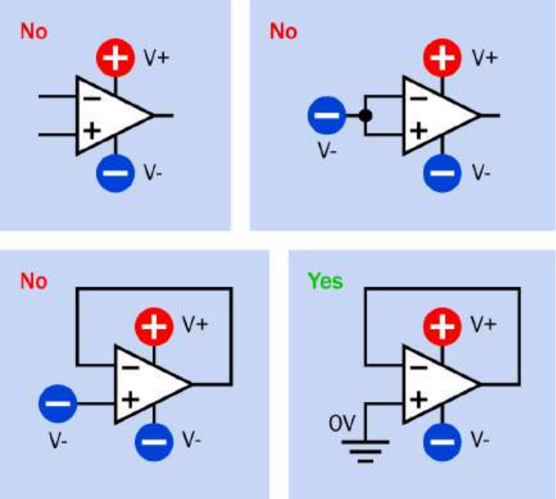

Figure 7-16. *When multiple op-amps share a chip, one that is unused will still receive power from the shared supply. Its inputs must not be allowed to float, and must be connected to minimize activity and power consumption. Three common errors are shown here, with one recommended option. Note the distinction between 0V ground (0V) and negative power (V−).*

Oscillating Output

The inputs of an op-amp are vulnerable to stray electromagnetic fields. If the conductors leading to and from the op-amp are relatively long, the output can couple capacitively with the input during voltage transitions, causing unwanted oscillations.

The commonly recommended solution to this problem is to add a 1μF bypass capacitor between the power supply and 0V ground. However, some manufacturers recommend alternatives such as introducing a very small amount of hysteresis, or reducing the value of input resistors to below 10K.

Confused Inputs

A schematic may show an op-amp with the noninverting input above the inverting input, or vice versa. The only indication of this will be the plus and minus signs inside the chip, which can be extremely small and easily overlooked. For convenience in drawing a diagram, two op-amps in the same circuit may have their inputs shown in opposite configurations. Special care must be taken to verify that the inverting and noninverting input pins on a chip are correctly assigned.

digital potentiometer

A **digital potentiometer** is also known as a *digitally adjustable potentiometer*, a *digitally controlled potentiometer*, a *digitally programmed potentiometer* (with acronym *DPP*), a *digpot*, or a *digipot*. The terms are functionally interchangeable. Because the abbreviation *pot* is often used to describe an analog potentiometer, some people refer to digital potentiometers colloquially as *digital pots*. In printed documentation, the letters in *pot* may be capitalized. Because it is an abbreviation, not an acronym, it is not capitalized here.

Because this component enables digital control of a variable voltage, it is a *mixed signal device*. It is classified here as an analog chip because it primarily emulates the function of an analog device. It may be thought of as a form of *digital-analog converter*, although this encyclopedia does not have a section devoted to that type of component or to *analog-digital converters*, as their application is relatively specialized.

OTHER RELATED COMPONENTS

- **potentiometer** (see Volume 1)

What It Does

This component is an integrated circuit chip that emulates the function of an analog **potentiometer**. It is often described as being *programmable*, meaning that its internal resistance can be changed via a control input.

Digital potentiometers are particularly suited for use in conjunction with a *microcontroller*, which can control the internal resistance of the component. Possible applications include adjustment of the pulse width of an oscillator or multistable multivibrator (e.g., using the Control pin of a 555 **timer** chip); adjustment of the gain in an **op-amp**; specification of voltage delivered by a *voltage regulator*; and adjustment of a *bandpass filter*.

A digital potentiometer in combination with a microcontroller may also be used in conjunction with a pair of external buttons or a rotational encoder, to adjust the gain of an audio amplifier and for similar applications.

Advantages

A digital potentiometer offers significant advantages over an analog potentiometer:

- Reliability. The digital component may be rated for as many as a million cycles (each storing the wiper position in an internal memory location). An analog component may be capable of just a few thousand adjustment cycles.
- Digital interface.
- Elimination of long signal paths or cable runs. The digital potentiometer can be placed close to other chips, whereas an analog potentiometer often has to be some distance away to enable control by the end user. Reduction in the length of signal paths can reduce capacitive effects, while elimina-

tion of cable runs will reduce manufacturing costs.

- Reduction in size and weight compared with a manual potentiometer.

A digital potentiometer also has some disadvantages:

- Its internal resistance is somewhat affected by temperature.
- It is not usually capable of passing significant current. Few chips can sink or source more than 20mA at the output, and 1mA is common. The output is primarily intended for connection with other solid-state components that have high impedance.
- Users may prefer the immediacy and tactile feel of a knob attached to an analog potentiometer, rather than a pair of buttons or a rotational encoder.

How It Works

A digital potentiometer changes the point at which a connection is made along a *ladder* of many fixed resistors connected in series inside the chip. Each end of the ladder, and each intersection between two adjacent resistors, is known as a *tap*. The pin that can connect with any of the taps is referred to as the *wiper*, because it emulates the function of a wiper in an analog potentiometer. In reality, a digital potentiometer does not contain a wiper or any other moving parts.

A fully featured digital potentiometer allows access to each end of the ladder through two pins that are often labeled "high" and "low," even though they are functionally interchangeable (except in the case of a component that simulates a logarithmic taper, as described later). The "low" end of the ladder is sometimes numbered 0. In this case, if there are *n* resistors, the "high" end of the ladder will be numbered *n*. Alternatively, if the "low" end of the ladder is numbered 1, and there are *n* resistors, the "high" end will be numbered *n+1*. This principle is illustrated in Figure 8-1.

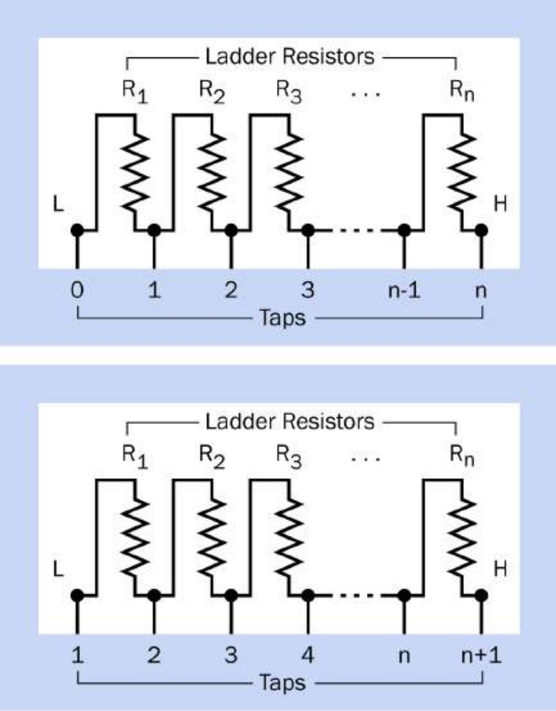

Figure 8-1. *Available wiper connections to a resistor ladder inside a digital potentiometer, showing two numbering systems that may be used.*

The "low" pin on a digital potentiometer may be identified as L, or A, or R_L, or P_A in a datasheet, while the "high" pin may be identified as H, or B, or R_H, or P_B and the pin that accesses the wiper is typically identified as W, or R_W, or P_W. Letters L, H, and W are used below. Although the L and H pins are functionally interchangeable, their labels are useful to identify which direction the W connection will move in response to an external signal.

Digital potentiometers are available with as few as 4 or as many as 1,024 taps, but common values are 32, 64, 128, or 256 taps, with 256 being the most common.

No specific schematic symbol represents a digital potentiometer. Often the component is shown as an analog potentiometer symbol inside a box that has a part number, as suggested in

Figure 8-2. Control pins and the power supply may be omitted if the schematic is just intended to show logical connections. Alternatively, if the digital potentiometer is depicted in a schematic where it is connected with other components such as a microcontroller, multiple pins and functions may be included, as shown in Figure 8-3. The pins additional to L, H, and W are explained below.

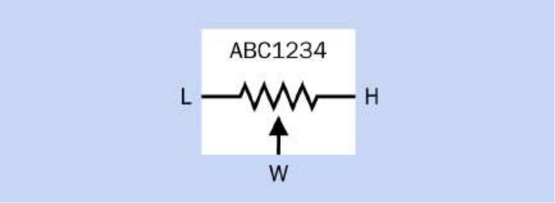

Figure 8-2. *There is no single specific symbol to represent a digital potentiometer. It may be shown using an analog potentiometer symbol in a box with a part number, as suggested here, where power connections and additional pins are omitted for clarity.*

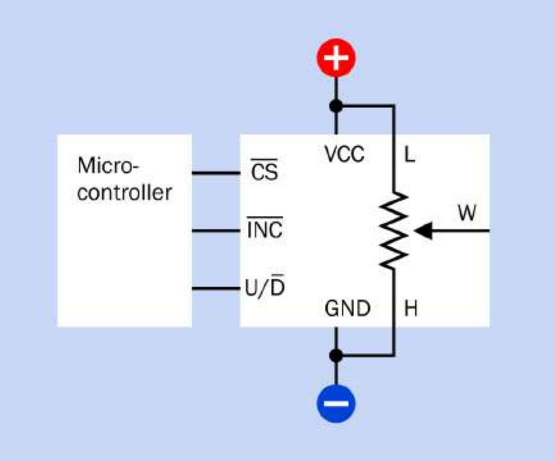

Figure 8-3. *If a digital potentiometer appears in a schematic where it is connected with other components such as a microcontroller, additional pins and functions may be indicated. This generic representation of a digital potentiometer shows some of the functions that can be included.*

Variants

A *dual* digital potentiometer contains two complete units, while a *quad* contains four. Triples exist but are relatively uncommon. A few chips contain six potentiometers. Multiple digital potentiometers in a chip can be used as the digital equivalent of *ganged* analog potentiometers, for simultaneous synchronized adjustment of multiple inputs in an audio system (two channels in a stereo amplifier, or more in a surround-sound system).

The pinouts of a sophisticated quad digital potentiometer chip are shown in Figure 8-4. Other quad chips have different pinouts and capabilities; there is no standardized format as there is with digital logic chips. In this example, the high/low states of Address 0 and Address 1 select one of the four internal resistor ladders, numbered 0 through 3. The Chip Select pin makes the whole chip either active or inactive. The Write Protect pin disables writing to the internal wiper memory. The Serial Clock pin inputs a reference pulse stream to which the serial input data must be synchronized. The Hold pin pauses the chip while data is being transmitted, allowing the data transmission to be resumed subsequently. The NC pins have no connection.

Volatile and Nonvolatile Memory

Any type of digital potentiometer requires memory to store its current wiper position, and this memory may be volatile or nonvolatile. Nonvolatile memory may be indicated in a datasheet by the term *NV*.

A digital potentiometer with volatile memory will typically reset its wiper to a center-tap position if power is disconnected and then restored. A digital potentiometer with nonvolatile memory will usually restore the most recently used wiper position, provided the chip is fully powered down and then fully powered up without glitches in the supply. If a microcontroller is being used to control the digital potentiometer, it can store the most recent resistance value in its own nonvolatile memory, in which case the type of memory in the potentiometer becomes irrelevant.

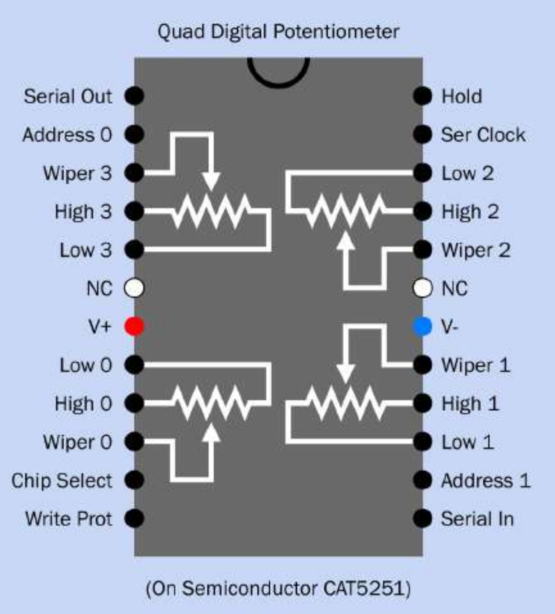

Figure 8-4. *Pinouts of a sophisticated quad digital potentiometer chip. Other chips will have different pinouts and capabilities. This example is available in surface-mount formats only. See text for details.*

Taper

Digital potentiometers are available with *linear taper* or *logarithmic taper*. In the former, each resistor in the ladder has the same value. In the latter, values are chosen so that the cumulative resistance between the wiper and the L end of the ladder increases geometrically as the wiper steps toward the H end of the ladder. This is useful in audio applications where sound intensity that increases exponentially may seem to increase linearly when perceived by the human ear.

A microcontroller can emulate logarithmic steps by skipping some taps in the ladder in a digital potentiometer, but this will result in fewer increments and lower precision.

Data Transfer

Digital potentiometers are mostly designed to use one of three serial protocols:

- *SPI*. This acronym is derived from *serial peripheral interface*, a term trademarked by Motorola but now used generically. The standard is adapted in various radically different ways among digital potentiometers.
- *I2C*. More correctly printed as *I^2C* and properly pronounced "I squared C," this acronym is derived from the term *inter-integrated circuit*. Developed by Philips in the 1990s, it is a relatively slow-speed bus-communication protocol (up to 400kbps or 1Mbps in its basic form). It is built into some microcontrollers. The standard is more uniformly and rigorously defined than SPI.
- *Up/down*, also sometimes known as *push-button* or *increment/decrement* protocol.

Both SPI and I2C are supported by many microcontrollers, including the Atmel AVR at the heart of the Arduino.

These three systems for controlling a digital potentiometer are described in more detail in the following sections.

SPI

This is the most widely used protocol, but when reading datasheets, a lot of care must be taken to determine how it varies in each case.

The Microchip 4131-503, shown in Figure 8-5, uses SPI protocol. It contains 128 resistors and can be powered by 1.8VDC to 5.5VDC.

The one feature that all versions of SPI have in common is that a series of high/low pulses is interpreted by the chip as a set of bits whose value defines a tap point in the resistor ladder. In computer terminology, every tap point has an *address*. The incoming bits define the address, after which the status of an additional input pin can tell the chip to move the wiper to that location.

Typically, there will be a chip select pin, identified as CS; a serial data input pin, identified as SDA, SI, DIN, or a similar acronym; and a serial clock pin, identified as SCL, SCLK, or SCK, which must receive a stream of pulses to which the high/low data input pulses must be synchronized. In addition, the SPI protocol allows bidirectional (duplex) serial communication. Only a minority of digital potentiometers make use of this capability, but where it exists, a serial data output pin may be labeled SDO. Alternatively, one pin may be multiplexed to enable both input and output, in which case it may be labeled SDI/SDO.

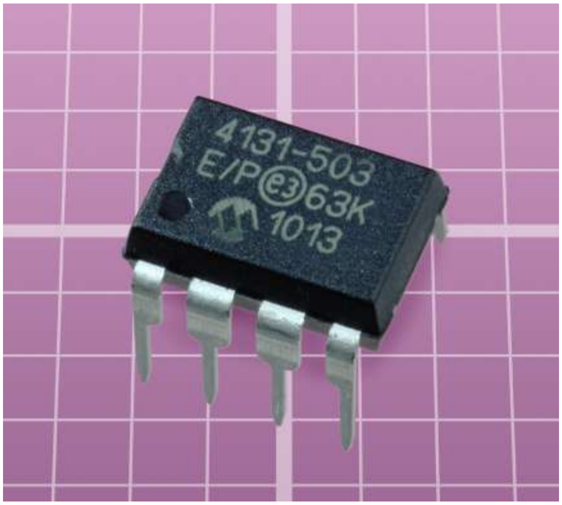

Figure 8-5. *This digital potentiometer uses SPI protocol. See text for details.*

If a pin is active-low, a bar (a horizontal line) will be printed above its acronym.

The most common type of digital potentiometer has 255 resistors and therefore 256 tap points, which have addresses numbered 0 through 255, each of which can be specified by a sequence of eight data bits constituting one byte. However, a different coding system will be applied in chips that have a different number of taps. In a 32-tap component, for instance, data is still sent in groups of eight bits, but only the first five bits define a *tap address*, while the remaining three are interpreted as commands to the chip.

Most 256-tap chips use an SPI protocol in which two eight-bit bytes are sent, the first being interpreted by the chip as a command, while the second specifies a tap address. Each manufacturer may use a different set of command codes, and these will vary among chips even from the same manufacturer.

Most commonly, three wires are used for data transmission and control (causing these chips to be described as *3-wire* programmable potentiometers).

CS is usually, but not always, pulled low to activate the digital potentiometer for input. A series of low or high states is then applied to the data-input pin. Each time the clock input changes state (usually on the rising edge of the clock pulse) the state of the data input is copied to a **shift register** inside the chip. After all the bits have been clocked in, CS can change from low to high, causing the contents of the shift register to be copied into a decoder section of the chip. The first bit received becomes the most significant bit in the decoder. The value of the eight bits is decoded, and the chip connects the W pin directly to the corresponding tap along the ladder of 255 internal resistors.

I2C Protocol

The I2C specification is controlled by NXP Semiconductors (formerly Philips), but can be used in commercial products without paying licensing fees. Only two transmission lines are required: one carrying a clock signal, the other allowing bidirectional data transfer synchronized with the clock (although many digital potentiometers use the I2C connection only to receive data). The pins are likely to be identified by the same acronyms as the pins on a chip that uses SPI protocol.

As in SPI, a command byte is followed by a data byte, although the command set differs from that of SPI and will also differ among various I2C chips. Full implementation of I2C allows multiple devices to share a single bus, but this capability may remain unused.

Up/Down Protocol

This simpler, asynchronous protocol does not require a clock input. The chip will respond to data pulses that are received at any speed (up to its maximum speed), and the pulse widths can be inconsistent.

Each pulse moves the wiper connection one step up or down the ladder. While this has the advantage of simplicity, the taps are not addressable, and consequently the wiper cannot skip to any tap without passing through intervening taps incrementally. This is not an inconvenience when the potentiometer controls audio gain, which is a primary application.

In some chips, an increment pin, usually labeled INC, receives pulses while the high or low state of a second pin, usually labeled U/D, determines whether each pulse will step the wiper up the ladder or down the ladder.

In other chips, pulses to an Up pin will step the wiper up the ladder, while pulses to a Down pin will step the wiper down the ladder.

Either of these chip designs can be referred to as a *two-wire* type. If an additional chip-select pin is included (labeled CS on datasheets), this type of digital potentiometer can be referred to as a *three-wire* type. The chip select pin is likely to be active-low, meaning that so long as it has a high state, the chip will ignore incoming signals.

The CAT5114 shown in Figure 8-6 uses an U/D pin. It contains 31 resistors, is available in 8-pin DIP or surface-mount formats, and can be powered by 2.5VDC to 6VDC. Each of its logic inputs draws only 10μA.

In six-pin chips the INC pin is omitted, one of the H, L, or W pins will be omitted, and the U/D pin will function differently. When CS is pulled low, the chip checks the state of the U/D pin. If it is high, the chip goes into increment mode; if it is low, the chip goes into decrement mode. So long as CS remains low, each transition of the U/D pin from low to high will either increment or decrement the wiper position, depending on the mode that was sensed initially. When CS goes high, further transitions on the U/D pin will be ignored until CS goes low again, at which point the procedure repeats.

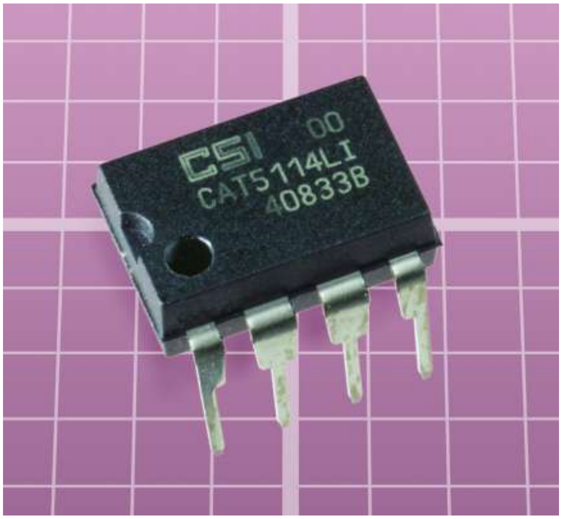

Figure 8-6. *This digital potentiometer contains 31 resistors and uses the simplest up/down protocol to step from one tap to the next.*

The chip does not provide any feedback regarding the position of its wiper, and consequently a control device such as a microcontroller cannot know the current wiper position. If the chip has nonvolatile memory (as is the case in many up/down digital potentiometers), it will resume its previous wiper location at power-up, but here again a control device will have difficulty determining what that position is. Therefore, in its basic form, an up/down chip is only appropriate for simple tasks, especially in response to up/down pushbuttons.

Other Control Systems

A few digital potentiometers use a parallel interface. Because they are relatively uncommon, they are not included here.

Connections and Modes

Some variants of digital potentiometers minimize the chip size and number of connections by limiting accessibility to the internal resistor ladder. In a chip designed to function in *rheostat mode*, the W pin is eliminated and the chip moves an internal connection point to change the resistance between the H and L pins.

In some variants, the low end of the ladder is permanently, internally connected with ground, and the L pin is omitted. In other variants, one end of the ladder is unconnected inside the chip.

A chip designed to function in *voltage divider mode* will include all three pins—H, L, and W—except in some instances where the low end of the ladder is grounded internally.

Variants are shown in Figures 8-7, 8-8, 8-9, and 8-10. Because some pins may be omitted, and there is no standardization of function among the pins that do exist, circuits and chips must be examined carefully prior to use.

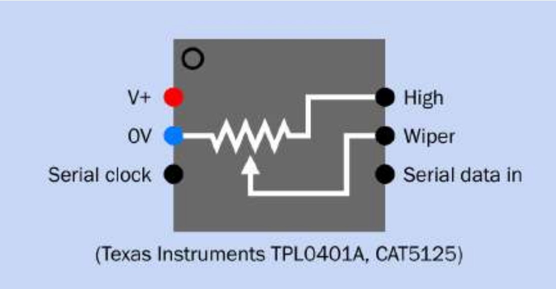
V+
0V
Serial clock
High
Wiper
Serial data in
(Texas Instruments TPL0401A, CAT5125)

Figure 8-7. *Some digital potentiometers minimize chip size and provide specialized functionality by eliminating pins. In the variant shown here, the W pin provides a voltage between H and an internal ground connection. The chip is controlled via I2C serial protocol.*

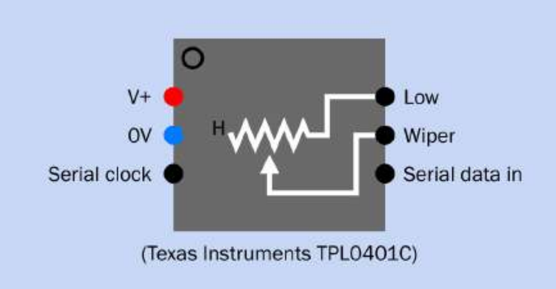
V+
0V
Serial clock
H
Low
Wiper
Serial data in
(Texas Instruments TPL0401C)

Figure 8-8. *In this variant, the H end of the internal resistor ladder is allowed to float inside the chip, and the digital potentiometer functions as a rheostat. The chip is controlled via I2C serial protocol.*

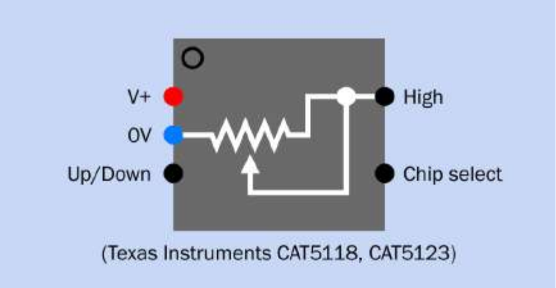
V+
0V
Up/Down
High
Chip select
(Texas Instruments CAT5118, CAT5123)

Figure 8-9. *This variant provides a variable resistance between the H pin and an internal connection with negative ground. Pin 5 is omitted. The chip is controlled by up/down pulses.*

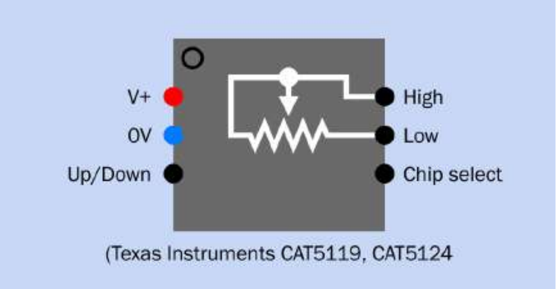
V+
0V
Up/Down
High
Low
Chip select
(Texas Instruments CAT5119, CAT5124

Figure 8-10. *This variant provides a variable resistance between H and L pins, without allowing either end of the resistor ladder to float. The W pin is omitted, as the wiper is tied internally to the H pin. The chips listed are controlled by up/down pulses.*

Values

A primary limitation of digital potentiometers is that they cannot withstand significant current. This may prevent them from being substituted for an analog potentiometer unless changes are made in the circuit. H, L, and W pins are usually unable to source or sink continuous, sustained current exceeding 20mA.

Wiper resistance is the resistance that is added internally by the wiper. This is nontrivial; it is often around 100Ω, and can be as high as 200Ω.

Typical end-to-end resistance of the ladder of internal resistors may range from 1K to 100K. Values of 1K, 10K, and 100K are common.

While the number of taps is likely to be a power of two in chips where the taps are addressable, up/down chips are not so constrained and may contain, for example, 100 taps.

The end-to-end resistance of a whole ladder may vary by as much as 20% from one sample of a chip to the next. Among resistor ladders in digital potentiometers sharing the same chip (i.e., in dual or quad chips) the variation will be much smaller.

Almost all digital potentiometers are designed for a supply voltage of 5V or less. The H and L pins are not sensitive to polarity, but the voltage applied to either of them must not exceed the supply voltage.

How to Use It

While most microcontrollers contain one or more analog-digital converters that change an analog *input* to an internal numeric value, a microcontroller cannot create an analog *output*. A digital potentiometer adds this functionality, although applications will be restricted by its limitation on current.

An up/down digital potentiometer can be controlled directly by a pair of **pushbuttons**, one of which will increase the resistance value while the other will reduce it. The pushbuttons must be *debounced* when used in this way. An alternative to pushbuttons is a **rotational encoder**, which emits a stream of pulses when its shaft is turned. In this case, an intermediate component (probably a microcontroller) will usually interpret the pulse stream and change it to a format that the digital potentiometer can understand.

Where a digital potentiometer is used in audio applications, it should be of the type that moves the wiper connection from one tap to another during a *zero crossing* of the audio signal (i.e., that is, at the moment when the AC input signal passes through 0V on its way from positive to negative or negative to positive). This suppresses the "click" that otherwise tends to occur during switching. Potentiometers with this feature may include phrases such as "glitch free" in their datasheets.

Digital potentiometers that are intended primarily for audio applications often have 32 taps spaced at intervals of 2dB. This will be sufficient to satisfy most listeners.

Achieving Higher Resolution

For sensitive applications where a resolution with more than 1,024 steps is required, multiple digital potentiometers with different step values can be combined. One way of doing this is shown in Figure 8-11. In this circuit, the wipers of P2 and P3 must be moved in identical steps, so that the total resistance between the positive power supply and negative ground remains constant. These two potentiometers could be contained in a dual chip, and would receive identical up/down commands. P1 is at the center of the voltage divider formed by P2 and P3, and is adjusted separately to "fine tune" the output voltage that is sensed at point A.

If all three of the potentiometers in this circuit contain 100 taps, a combined total of 10,000 resistance steps will be possible.

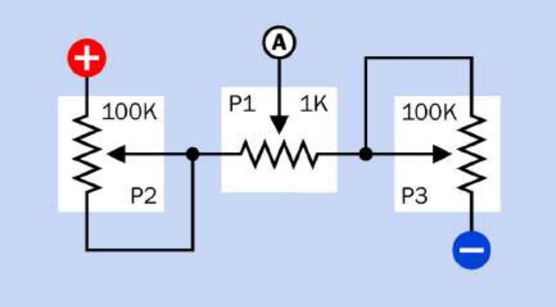

Figure 8-11. *If all three digital potentiometers in this schematic have 100 taps, and the wipers of P2 and P3 are moved in synchronization, the voltage measured at point A can have a high resolution of up to 10,000 steps.*

What Can Go Wrong

Noise and Bad Inputs

Because a digital potentiometer is capable of receiving data at speeds as high as 1MHz, it is sensitive to brief input or power fluctuations, and can misinterpret them as instructions to move the wiper—or can misinterpret them as command codes, in a component using SPI or I2C serial protocol.

To minimize noise in the power supply, some manufacturers recommend installing a 0.1µF capacitor as close as possible to the power supply pin of the component. In addition, it is obviously important to provide clean input signals. This means thorough debouncing of any electromechanical switch or pushbutton inputs.

Wrong Chip

The wide diversity of input protocols and pinouts creates many opportunities for installation error.

Up/down, SPI, and I2C protocol require totally different pulse streams. Many manufacturers offer components that are distinguished from each other by just one or two digits in their part numbers, yet have radically different functionality.

If more than one specific type of digital potentiometer may be used during circuit development, they should be stored carefully to avoid inadvertant substitutions. Using the wrong chip may be particularly confusing in that an inappropriate input protocol will still produce some results, although not those which were intended.

Controller and Chip Out of Sync

As noted in the discussion of data transmission protocols, most digital potentiometers are not capable of providing feedback to confirm the position of the internal wiper. A designer may wish to include a power-up routine which establishes the state of the digital potentiometer by resetting it to a known position, at one end of its scale or the other.

Nonlinear Effects

While the end-to-end resistance of the resistor ladder inside a digital potentiometer is not likely to be affected significantly by changes in temperature, the resistance at the wiper is more heat sensitive.

In an up/down chip, there can be differential errors between incremental and decremental modes. In other words, if a tap is reached by stepping up to it incrementally, the resistance between the W pin and H or L may not be quite the same as if the same tap is reached by stepping down to it decrementally. The difference may not be significant, but may be puzzling for those who are unfamiliar with this phenomenon.

Some differences may be found among resistors in a ladder. That is, in a supposedly linear digital potentiometer, each resistor may differ in value slightly from the next.

Data Transfer Too Fast

When using a microcontroller to send data to a digital potentiometer, a small delay may be necessary between pulses, depending on the microcontroller's clock speed. A digital potentiometer may require a minimum pulse duration of 500ns. Check the manufacturer's datasheet for details.

8

timer

A device that creates a single timed pulse, or a series of timed pulses with timed intervals between them, is properly known as a *multivibrator*, although the generic term **timer** has become much more common and is used here.

Three types of multivibrator exist: *astable*, *monostable*, and *bistable*. The behavior of *astable multivibrators* and *monostable multivibrators* is described in detail in this entry. A timer chip can also be made to function as a bistable multivibrator. This is described briefly below, but it is not a designed function of a timer. The primary discussion of bistable multivibrators will be found in the entry of this encyclopedia dealing with **flip-flops**.

OTHER RELATED COMPONENTS

- **flip-flop** (See Chapter 11)

What It Does

A *monostable* timer emits a single timed pulse of fixed length in response to a triggering input that is usually of shorter duration. Many monostable timers are also capable of running in *astable* mode, in which the timer spontaneously emits an ongoing stream of timed pulses with timed gaps between them. A dual-mode timer can run in either mode, determined either by external components attached to it, or (less commonly) by changing the status of a mode selection pin.

Monostable Mode

In monostable mode, the timer emits a pulse in response to a change from high to low voltage (or, less commonly, from low to high voltage) at a *trigger pin*. Most timers respond to a *voltage level* at the trigger pin, but some are insensitive to any persistent pin state and only respond to a *voltage transition*. This is known as *edge triggering*.

The pulse generated by the timer may consist of a change from low to high (or, less commonly, from high to low) at an output pin. The length of the pulse will be determined by external components, and is independent of the duration of the triggering event, although in some cases, an output pulse may be prolonged by *retriggering* the timer prematurely. This is discussed below.

At the end of the output pulse, the timer reverts to its quiescent state, and remains inactive until it is triggered again.

A monostable timer can control the duration of an event, such as the time for which a light remains on after it has been triggered by a motion sensor. Alternatively, the timer can impose a delay, such as the interval during which a paper towel dispenser refuses to respond after a towel has been dispensed. A timer can also be useful to generate a clean pulse in response to an unstable or noisy input, as from a manually operated pushbutton.

Astable Mode

In astable mode, a timer will generally trigger itself as soon as power is connected, without any need for an external stimulus. However, the output can be suppressed by applying an appropriate voltage to a *reset pin*.

External components will determine the duration of each pulse and the gap between it and the next pulse. The pulse stream can be slow enough to control the flashing of a turn signal in a 1980s automobile, or fast enough to control the bit rate in a data stream from a computer.

Modern timer circuits are often incorporated in chips that have other purposes. The flashing of a turn signal in a modern car, for instance, is now likely to be timed by a microcontroller that manages many other functions. Still, chips that are purely designed as timers remain widely used and are very commonly available in numerous through-hole and surface-mount formats.

How It Works

The duration of a single pulse in monostable mode, or the frequency of pulses in astable mode, is most commonly determined by an external *RC network* consisting of a resistor in series with a capacitor. The charging time of the capacitor is determined by its own size and by the value of the resistor. The discharge time will be determined in the same way. A **comparator** inside the timer is often used to detect when the potential on the capacitor reaches a *reference voltage* that is established by a voltage divider inside the chip.

Variants

The 555 Timer

An eight-pin integrated circuit manufactured by Signetics under part number 555 was the world's first fully functioned timer chip, introduced in 1972. It combined two comparators with a **flip-flop** (see Chapter 11) to enable great versatility while maintaining excellent stability over a wide range of supply voltages and operating temperatures. Subsequent timers have been heavily influenced by this design. A typical 555 timer chip is shown in Figure 9-1.

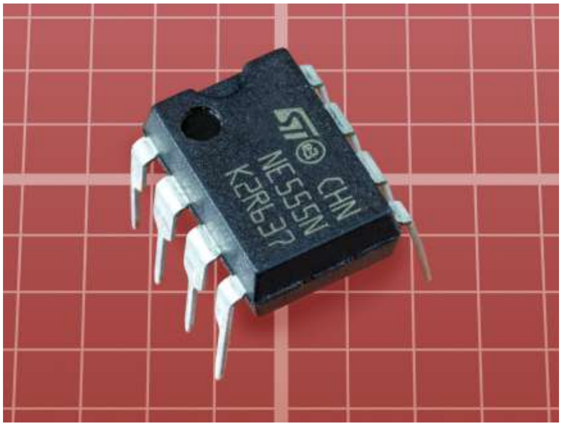

Figure 9-1. *A typical 555 timer chip. Functionally identical versions in which the "555" identifier is preceded or followed by different letter combinations are available from many different manufacturers.*

The 555 was designed by one individual, Hans Camenzind, working as an independent consultant for Signetics. According to a transcript of an interview with Camenzind maintained online at the Transistor Museum, "There was nothing like it at the time. You had to use quite a few discrete components—a comparator, a Zener diode or even two. It was not a simple circuit."

The 555 timer quickly became the most widely used chip in the world, and was still selling an annual estimated 1 billion units three decades after its introduction. It has been used in spacecraft, in intermittent windshield wiper controllers, in the early Apple II (to flash the cursor), and in children's toys. Like many chips of its era, its design was unprotected by patents, allowing it to be copied by numerous manufacturers.

The initial version was built around **bipolar transistors**, and consequently is referred to as the *bipolar version* or (more often) the *TTL version*, this being a reference to *transistor-transistor logic* protocol. Within a few years, CMOS versions based around **MOSFETs** were developed. They reduced the ability of the chip to sink or source current at its output pin, but consumed far less power, making them better suited to battery-operated products. The CMOS versions were and still are pin-compatible with the original bipolar version, both in through-hole and surface-mount formats. Their timing parameters are usually the same.

555 Monostable Operation

The internal functionality of a 555 timer wired to run in monostable mode is illustrated in Figure 9-2 with the chip seen from above. The pins are identified in datasheets by the names shown. To assist in visualizing the behavior of the chip, this figure represents the internal flip-flop as a switch which can be moved by either of two internal comparators, or by an input from the Reset pin.

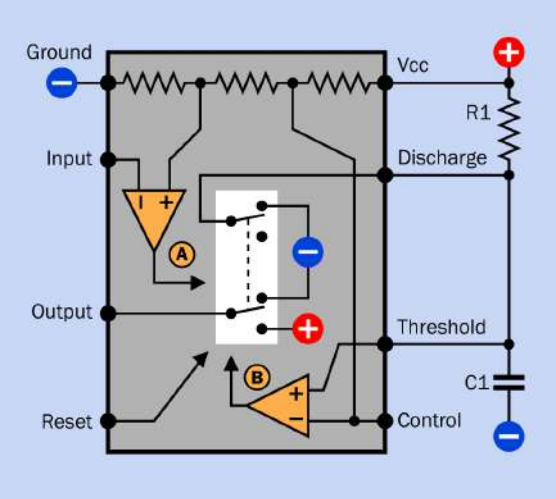

Figure 9-2. *The internal functions of a 555 chip, with its flip-flop represented as a switch that can be moved by either of two comparators, or by a low voltage on the Reset pin. An external resistor and capacitor, shown as R1 and C1, cause the timer to run in monostable (one-shot) mode, generating a single high pulse when the state of the Input pin is pulled from high to low.*

Inside the chip, three resistances of 5K each are connected between V+ (positive supply voltage) and negative ground. It has been suggested that the part number of the 555 chip was derived from these three 5K resistors, but Hans Camenzind has pointed out that Signetics was already using three-digit part numbers beginning with the number 5, and probably chose the 555 part number because the sales department had high expectations for the chip and wanted its number to be easily memorable. (A similar rationale explains the part number of the 2N2222 transistor.)

The resistances inside the timer function as a *voltage divider*, providing a reference of 1/3 of V+ to the noninverting pin of Comparator A and 2/3 of V+ to the inverting pin of Comparator B. (See Chapter 6 for an explanation of the functioning of comparators.)

When power is initially supplied to the timer, if the Input pin is at a high state, Comparator A has a low output, and the flip-flop remains in its “up” position, allowing the Output pin to remain in a low state. The flip-flop also grounds the lower end of R1, which prevents any charge from accumulating on capacitor C1.

If the state of the Input pin is pulled down externally to a voltage less than 1/3 of V+, Comparator A now creates a high output that changes the flip-flop to its “down” position, sending a high signal out through the Output pin. At the same time, C1 is no longer grounded, and begins to charge at a rate determined by its own size and by the value of R1. When the charge on the capacitor exceeds 2/3 of V+, it activates Comparator B, which forces the flip-flop into its “up” position. The Output pin goes low, C1 discharges itself into the Discharge pin, and the timer’s cycle is at an end.

The low voltage on the Input pin of the timer must end before the end of the output cycle. If the voltage on the Input pin remains low, it will re-trigger the timer, prolonging the output pulse.

A pullup resistor may be used on the Input pin to avoid false triggering, especially if an external electromechanical **switch** or **pushbutton** is used to pull down the Input pin voltage.

The Reset pin should normally be held high, either by being connected directly to positive supply voltage (if the reset function will not be needed) or by using a pullup resistor. If the Reset pin is pulled low, this will always interrupt an output pulse regardless of the timer's current status.

If a voltage higher or lower than 2/3 of V+ is applied to the Control pin, this will change the reference voltage on Comparator B, which determines when the charging cycle of C1 ends and the discharge cycle begins. A lower reference voltage will shorten each output pulse by allowing a lower charge limit for C1. If the control voltage drops to 1/3 of V+ (or less), the capacitor will not charge at all, and the pulse length will diminish to zero. If the control voltage rises to become equal to V+, the capacitor will never quite reach that level, and the pulse length will become infinite. A workable range for the control voltage is therefore 40% to 90% of V+.

Because the Control pin is an input to the chip, it should be grounded through a 0.01µF ceramic capacitor if it will not be used. This is especially important in CMOS versions of the timer.

A defect of the bipolar 555 is that it creates a *voltage spike* when its Output pin changes state. If it will be sharing a circuit with sensitive components, a 0.01µF bypass capacitor should be added as closely as possible between the V+ pin and negative ground. The voltage-spike problem was largely resolved by the CMOS 555.

555 Astable Operation

In Figure 9-3, the 555 timer chip is shown with external components and connections to run it in astable mode. The pin names remain the same but have been omitted from this diagram because of limited space. The labeling of the two external resistors and capacitor as R1, R2, and C1 is universal in datasheets and manufacturers' documentation.

When the timer is powered up initially, capacitor C1 has not yet accumulated any charge. Consequently, the state of the Threshold pin is low. But the Threshold pin is connected externally with the Input pin, for astable operation. Consequently, the Input pin is low, which forces the flip-flop into its "down" state, creating a high output. This happens almost instantaneously.

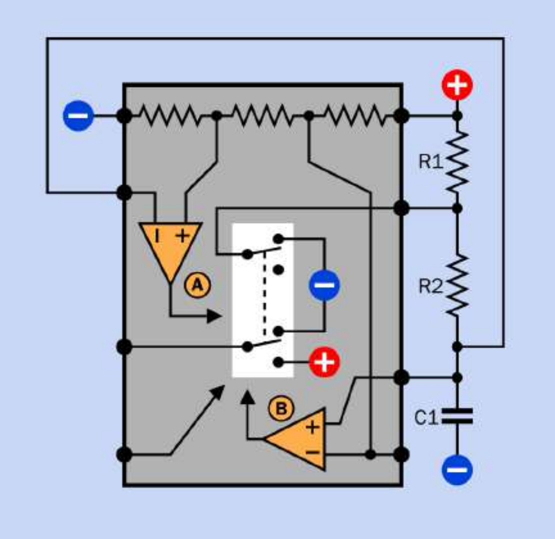

Figure 9-3. *The internal functions of a 555 chip, with two external resistors and a capacitor wired to run the timer in astable (free-running) mode.*

While the flip-flop is "down," the Discharge pin is not grounded, and current flowing through R1 and R2 begins to charge the capacitor. When the charge exceeds 2/3 of positive supply voltage, Comparator B forces the flip-flop into its "up" position. This ends the high pulse on the Output pin, and starts to drain the charge from the capacitor through R2, into the Discharge pin. However, the voltage on the capacitor is still being shared by the Input pin, and when it diminishes to 1/3 of V+, the Input pin reactivates Comparator A, starting the cycle over again.

The functions of the Reset and Control pins are the same as in monostable mode. Because voltage applied to the Control pin changes the length of each pulse and the gaps between pulses, it has the effect of adjusting the frequency of the output in astable mode.

When power is first connected to the timer, C1 must initially charge from an assumed state of zero potential to 2/3 V+. Because subsequent cycles will begin when the capacitor is at 1/3 V+, the first high output pulse from the timer will be slightly longer than subsequent output pulses. This is unimportant in most applications, especially because the rate at which a capacitor accumulates charge is greater when beginning from 0V than when it has reached 1/3 V+. Still, the longer initial pulse can be noticeable when the timer is running slowly.

Because the capacitor charges through R1 and R2 in series, but discharges only through R2, the length of each positive output pulse in astable mode is always greater than the gap between pulses. Two strategies have been used to overcome this limitation. See "Separate Control of High and Low Output Times" on page 80.

556 Timer

The 556 consists of two 555 bipolar-type timers in one package. An example of the chip is shown in Figure 9-4. The pinouts are shown in Figure 9-5. Although 556 timers have become relatively uncommon compared with the 555, they are still being manufactured in through-hole and surface-mount versions by companies such as Texas Instruments and STMicroelectronics, under part numbers such as NA556, NE556, SA556, and SE556 (with various letters or letter pairs appended). Each timer in the chip has its own set of inputs and outputs, but the timers share the same V+ and ground voltages.

558 Timer

This 16-pin chip is now uncommon, and many versions have become obsolete. It has been identified by a part number such as NE558 although different prefix letters may be used. The NTE926, shown in Figure 9-6, is actually a 558 timer.

The chip contains four 555 timers sharing a common power supply, common ground, and common control-pin input. For each internal timer, the Threshold and Discharge functions are connected internally, so that the timers can only be used in one-shot mode. However, one timer can trigger another at the end of its cycle, and the second timer can then retrigger the first, to create an astable effect.

Each timer is edge-triggered by a voltage transition (from high to low), instead of being sensitive to a voltage level, as is the case with a 555 timer. Consequently the timers in the 558 chip are insensitive to a constant (DC) voltage.

NE556N
GK8300DQ
VW GK CHN

Figure 9-4. *An example of the 556 timer chip.*

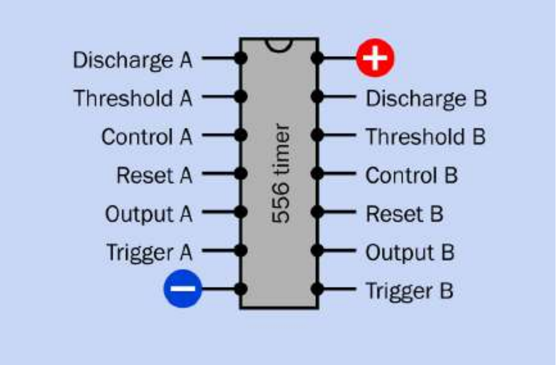

Figure 9-5. *The 556 timer contains two separate 555 timers sharing the same power supply and ground. The pin functions for timer A and timer B are shown here.*

The output from each timer is an open collector, and therefore requires an external pullup resistor. Each output is capable of sinking up to 100mA.

NTE 074K
NTE926

Figure 9-6. *The NTE926 is a 558 timer chip.*

CMOS 555 Timer

While the part numbers of many CMOS versions are significantly different from part numbers of the bipolar versions, in some instances the CMOS numbers are only distinguished by a couple of initial letters. The ST Microelectronics TS555 series and Texas Instruments TLC555 series, for example, use MOSFETs internally. The ST Microelectronics SE555 series and Texas Instruments SA555 series use bipolar transistors internally.

One way to distinguish between the two types, when searching a website maintained by a parts supplier, is to begin by looking generically for a "555 timer" and then add a search filter to show chips either with a minimum power supply of 3VDC (which will be CMOS) or with a minimum power suply of 4.5VDC (which will be bipolar).

CMOS versions of the 555 timer do not create the power spike that is characteristic of the bipolar versions during output transitions. The CMOS chips can also be powered by a lower voltage (3VDC, or 2VDC in some cases), and will draw significantly less current in their quiescent state. They also require very little current for threshold, trigger, and reset functions.

The wiring of external resistors and capacitors to the CMOS version of the chip, and the internal voltage levels as a fraction of V+, are identical to the original 555 timer. Pin functions are likewise identical. The only disadvantages of CMOS versions are their greater vulnerability to static discharge, and their lower output currents. The TLC555, for instance, will source only 15mA (although it can sink 10 times that amount). Other manufacturers have different specifications, and datasheets should be checked carefully.

5555 Timer

The 5555 contains a digital **counter** that enables it to time very long periods. Its full part number is 74HC5555 or 74HCT5555, although these numbers may be preceded or followed by letter combinations identifying the manufacturer. It is not pin-compatible with a 555 timer.

Two input pins are provided, one to trigger the timer on a rising edge, the other on a falling edge, of the input pulse. The inputs are Schmitt-triggered.

The timer is rated for 1Hz to 1MHz (using an external resistor and external capacitor). The counter section can divide the pulse frequency by values ranging from 2 to 256. For longer timed periods, different settings on the digital control pins will divide the frequency by values ranging from 2^{17} through 2^{24} (131,072 through 16,777,216). This enables the timer to achieve a theoretical pulse length lasting for more than 190 days. The timer will accept a clock input from an external oscillator to achieve better accuracy than is available with a resistor-capacitor timing circuit.

7555 Timer

This 8-pin chip is a CMOS version of the 555 timer, manufactured by companies such as Maxim Integrated Products and Advanced Linear Devices. Its characteristics are similar to those of CMOS 555 timers listed above, and the pinouts are the same.

7556 Timer

This 14-pin chip contains two 7555 timers, sharing common power supply and ground connections. Pinouts are the same as for the original 556 timer, as shown in Figure 9-5.

4047B Timer

This 14-pin CMOS chip was introduced in an effort to address some of the quirks of the 555 timer while also providing additional features. It runs in either monostable or astable mode, selectable by holding one input pin high or another input pin low. In astable mode, its duty cycle is fixed at approximately 50%, a single resistor being used for both charging and discharging the timing capacitor. An additional "oscillator" output runs twice as fast as the regular output.

In monostable mode, the 4047B can be triggered by a positive or negative transition (depending on which of two input pins is used). It ignores steady input states and will also ignore additional trigger pulses that occur during the output pulse. However, a retrigger pin is provided to extend the output pulse if desired.

Complementary output pins are provided, one being active-high while the other is active-low.

To time very long periods, the 4047B was designed to facilitate connection with an external counter.

The power supply for the 4047B can be as low as 3VDC. Its maximum source or sink output current is only 1mA when powered at 5VDC, but up to 6.8mA when powered at 15VDC.

The chip is still available from manufacturers such as Texas Instruments (which markets it as the CD4047B) in through-hole and surface-mount formats. However, despite its versatility, the 4047B is less popular than dual monostable timers, described in the next section.

Dual Monostable Timers

Various timers that run only in monostable mode are available in dual format (i.e., two timers in one chip). This format became popular partly because two monostable timers can trigger each other to create an astable output, in which the pulse width, and the gap between pulses, can be set by a separate resistor-capacitor pair on each timer. This allows greater flexibility than is available when using a 555 timer.

Most dual monostable timer chips are edge-triggered by a change in input voltage, and will ignore a steady DC voltage. Consequently, the output from one timer can be connected directly to the input of another, and no coupling capacitors are necessary.

As in the 4047B, the user has a choice of two input pins for each timer, one triggered by a transition from low to high, the other triggered by a transition from high to low. Similarly, each timer has two outputs, one shifting from low to high at the start of the output cycle, the other shifting from high to low.

The values of a single resistor and capacitor determine the pulse duration of each timer.

Dual monostable timers often have the numeric sequence 4528 or 4538 in their part numbers. Examples include the HEF4528B from NXP, the M74HC4538 from STMicroelectronics, and the MC14538B from On Semiconductor. The 74123 numeric sequence identifies chips that have a very similar specification, with chip-family identifiers such as HC or LS inserted, as in the 74HC123 and 74LS123, and additional letters added as prefix or suffix. The pinouts of almost all these chips are identical, as shown in Figure 9-7. However, Texas Instruments uses its own numbering system, and datasheets should always be consulted for verification before any connections are made.

Many chips of this type are described as "retriggerable," meaning that if an additional trigger pulse is applied to the input before an output pulse has ended, the current output pulse will be extended in duration. Check datasheets carefully to determine whether a chip is "retriggerable" or will ignore new inputs during the output pulse.

The 74HC221 dual monostable vibrator (pictured in Figure 9-8) functions very similarly to the components cited above, but has slightly different pinouts.

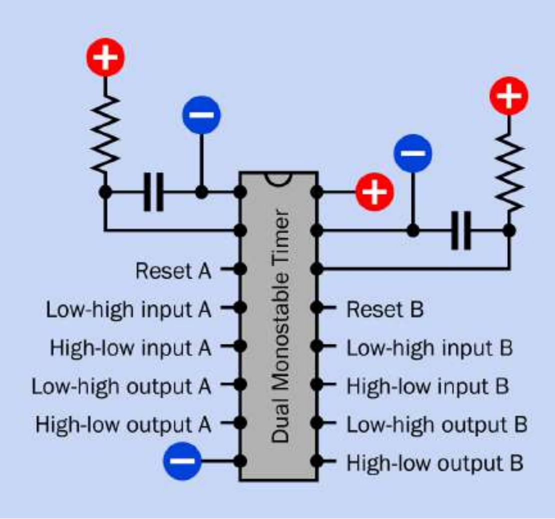

Figure 9-7. *Pin functions for most 4528, 4538, and 74123 series of dual monostable timer chips. An RC network is shown connected for each timer. Note that Texas Instruments uses different pinouts on its versions.*

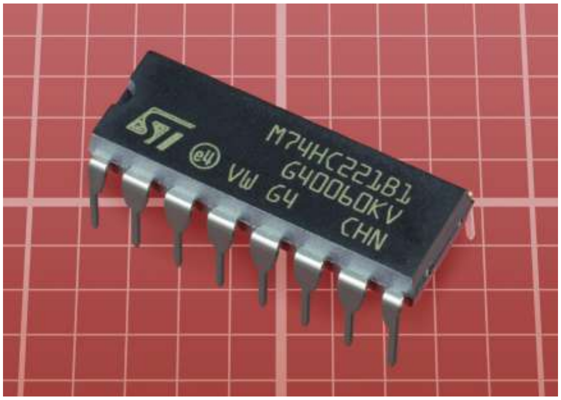

Figure 9-8. *A dual-timer chip containing two monostable multivibrators that can function in astable mode if they are connected externally to trigger each other.*

Values

555 Timer Values

The original bipolar version of the 555 timer was designed to operate using a wide range of positive supply voltages, from 4.5VDC to 16VDC. CMOS versions vary in their recommended V+ values, and datasheets must be consulted for verification.

The output of a bipolar 555 is rated to source or sink up to 200mA. In practice, the maximum current will be lower when the timer is powered at the low end of its range, around 5VDC. Attempting to source more than 50mA will pull down the voltage internally, affecting operation of the timer.

CMOS versions all impose restrictions on output current, allowing higher values for sinking than sourcing. Again, datasheets must be consulted for the values, which vary widely from one component to another.

The voltage measured on the output pin, when it is used for sourcing current, will always be lower than the power supply voltage, and a 1.7V drop is commonly specified for bipolar versions. In practice, the voltage drop that is actually measured may be less, and will vary according to the load on the output.

The voltage drop does not increase significantly with a higher supply voltage, and because it is a relatively constant value, it becomes less significant when a higher value for V+ is used.

CMOS versions of the 555 timer achieve a claimed output source voltage that is only 0.2V less than the power supply.

When choosing values for R1 and R2, a minimum for each resistor is 5K, although 10K is preferred. Lower values will increase power consumption, and may also allow overload of the internal electronics when the chip sinks current from C1. A typical maximum value for each resistor is 10M.

A high-value capacitor may cause the timer to function less accurately and predictably, because large capacitors generally allow more *leakage*. This means that the capacitor will be losing charge at the same time that it is being charged through R1 + R2. If these resistors have high values, and the capacitor has a value of 100µF or more, the rate of charge may be so low that it is comparable with the rate of leakage. For this reason, a 555 timer is not a good choice for timing intervals much greater than a minute. If a large-value capacitor is used, tantalum is preferable to electrolytic.

The minimum practical value for a timing capacitor is around 100pF. Below this, performance may not be reliable.

Although some CMOS versions may enable fast switching, the shortest practical output pulse for a 555 timer is around 10 microseconds. On the input pin, a triggering pulse of at least 1 microsecond should be used.

Time Calculation in Monostable Mode

If R1 is measured in kilohms and C1 is measured in microfarads, the pulse duration, T, in seconds, of a 555 timer running in monostable mode can be found from this simple formula:

```
T = 0.0011 * R1 * C1
```

This relationship is the same in all versions of the 555. Figure 9-9 provides a quick and convenient way to find the pulse value using some common values for R1 and C1. Resistors can be obtained with tolerances below plus-or-minus 1%, but capacitors are often rated with an accuracy of only plus-or-minus 20%. This will limit the accuracy of the pulse values shown in the chart.

Time Calculation in Astable Mode

If R1 and R2 are measured in kilohms and C1 is measured in microfarads, the frequency of pulses, F (measured in Hz) of a 555 timer running in astable mode can be found from this simple formula:

```
F = 1440 / ( ( R1 + (2 * R2) ) * C1)
```

This relationship is found in all versions of the 555. Figure 9-10 shows the frequency for common values of R2 and C1, assuming that the value of R1 is 10K. In Figure 9-11, a value of 100K is assumed for R1.

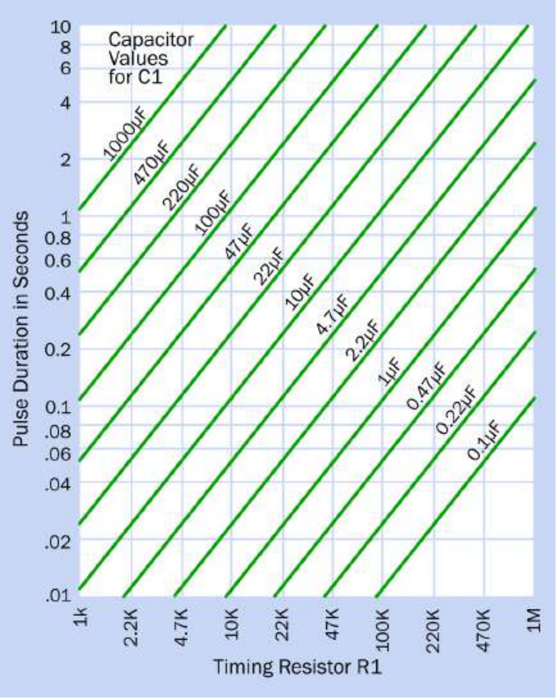

Figure 9-9. *To determine the pulse duration of a 555 timer running in monostable mode: find the value of R1 on the horizontal scale, follow its vertical grid line upward to the intersection with a green line which corresponds with the value of capacitor C1, and read across to the vertical scale providing the duration in seconds. Both axes are logarithmic.*

Dual Monostable Timers

Dual chips such as the HEF4528B from NXP, the M74HC4538 from STMicroelectronics, the MC14538B from On Semiconductor, and the 74HC123 from Texas Instruments have widely varying requirements for power supply. Some accept a limited range from 3VDC to 6VDC, while others tolerate a range of 3VDC to 20VDC. When powered with 5VDC their required input and output states are compatible with those of 5V logic chips.

Output pins of these chips source and sink no more than 25mA (much less in some instances). Because there are so many variants, they cannot be summarized here, and datasheets must be consulted for details.

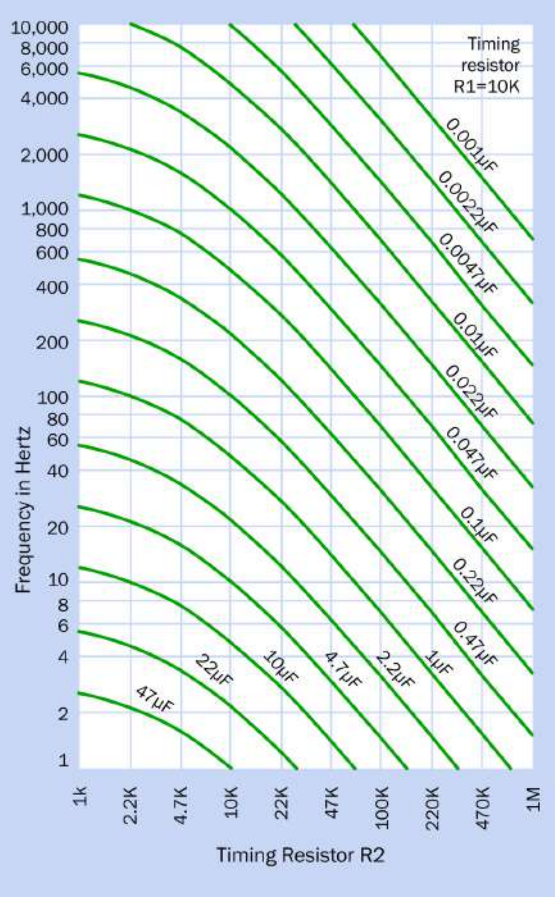

Figure 9-10. *To obtain the frequency of a 555 timer running in astable mode, when R1 has a value of 10K: find the value of R2 on the horizontal scale, follow its vertical grid line upward to the intersection with a green curve which corresponds with the value of capacitor C1, and read across to the vertical scale providing the frequency in Hertz. Both axes are logarithmic.*

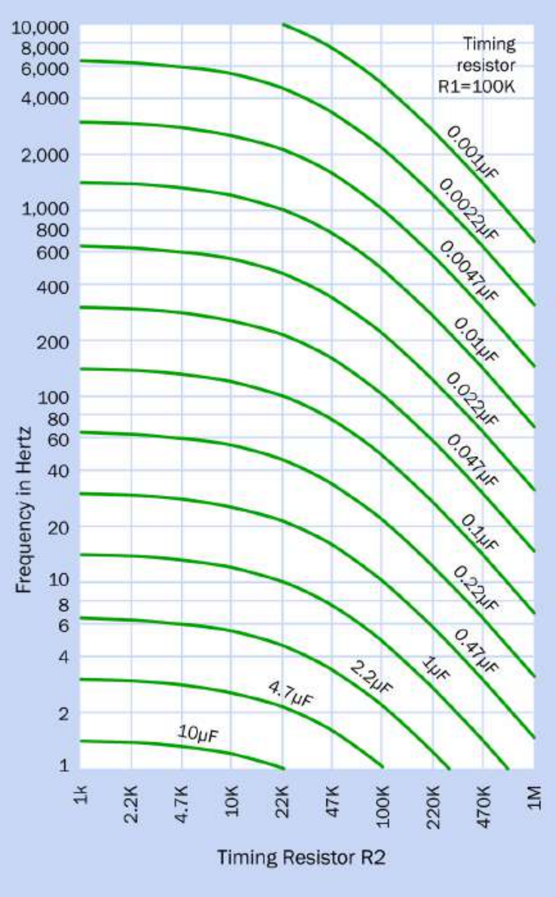

Figure 9-11. *To obtain the frequency of a 555 timer running in astable mode, when R1 has a value of 100K: find the value of R2 on the horizontal scale, follow its vertical grid line upward to the intersection with a green curve which corresponds with the value of capacitor C1, and read across to the vertical scale providing the frequency in Hertz. Both axes are logarithmic.*

As these timers are all monostable, and each timer uses just one resistor and one capacitor, the only formula required is to give the pulse time as a function of these two variables. If R is the resistor value in ohms, and F is the capacitor value in farads, and K is a constant supplied by the manufacturer, the pulse time T, in seconds, is found from the formula:

```
T = R * F * K
```

K ranges between 0.3 and 0.7 depending on the manufacturer and also on the voltage being used. Its value should be found in the manufacturer's datasheet. If R is measured in megohms and F is measured in microfarads, the formula is still valid, as the multipliers cancel each other out.

Generally speaking, these dual monostable CMOS timers are not intended for pulse duration exceeding 1 minute.

The timing capacitor should be no larger than 10µF, as it discharges directly and rapidly through the chip.

How to Use It

Where a timer is required to drive a load such as a relay coil or small motor directly, the original TTL version of the 555 timer will be the only choice. Even in this instance, a protection diode must be used across the inductive device.

For smaller loads and applications in chip-to-chip circuits, CMOS versions of the 555, including the 7555, use less power, cause less electrical interference, and are pin-compatible while using the same formulae to calculate frequency in astable mode or pulse duration in monostable mode. They are of course more vulnerable to static discharge, and care must be taken to make a connection to every pin (the capacitor that grounds the Control pin, if Control is not going to be used, is mandatory).

In dual monostable timers, unused rising-edge trigger inputs should be tied to V+ while unused falling-edge trigger inputs should be tied to ground. A Reset pin that will not be used should be tied to V+, unless that entire timer section of the chip will be unused, in which case the pin should be grounded.

To measure durations longer than a few minutes, a timer which incorporates a programmable counter to divide the clock frequency is the sensible choice. See the description of the 5555 timer that was included earlier in this entry.

The original bipolar version of the 555 remains a robust choice in hobby applications such as robotics, and its design allows some versatile variations which may even be used in logic circuits. A variety of configurations are shown in the schematics below.

555 Monostable Mode

The basic schematic for a 555 timer running in monostable mode is shown in Figure 9-12. In this particular example, a pushbutton that is liable to suffer from *switch bounce* is connected to the Input pin of the timer, which responds to the very first connection made by the pushbutton and ignores the subsequent "bounces," thus producing a "clean" output. To avoid retriggering, which results in a prolonged output pulse, the timer's output should exceed the time for which the button is likely to be pressed. The output should also exceed the duration of any possible switch bounce, which can otherwise create multiple output pulses. In the schematic, an LED is attached to the timer output for demonstration purposes.

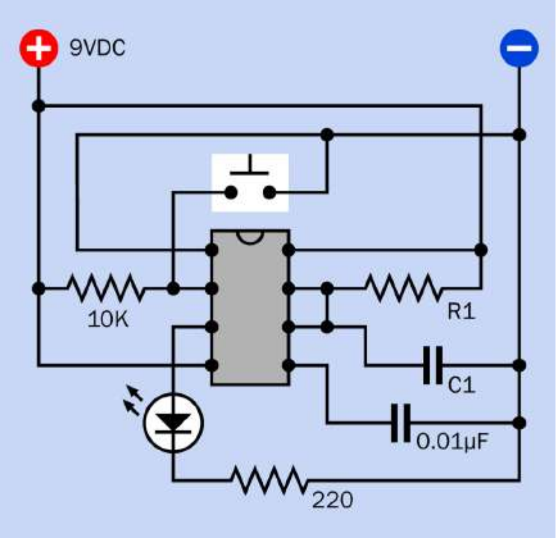

Figure 9-12. *The basic monostable configuration of a 555 timer. This particular circuit debounces an input from a pushbutton switch and converts it to a clean pulse of fixed duration, powering an LED for demonstration purposes.*

This circuit is shown on a breadboard in Figure 9-13. The red and blue wires, at the top of the photograph, supply 9VDC to the board. R1 is 1M, while C1 is 1µF, creating a pulse of just over 1 second. A tactile switch, just above the timer, provides the input.

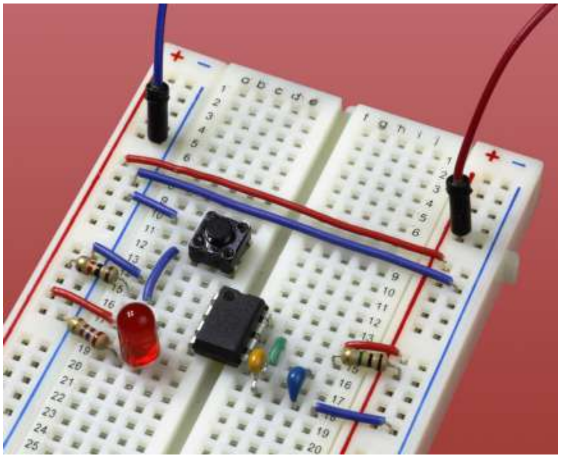

Figure 9-13. *The basic monostable configuration for a 555 timer, mounted on a breadboard.*

555 Astable Mode

A basic schematic for a 555 timer running in astable mode is shown in Figure 9-14. Once again, an LED is attached to the output for demonstration purposes. If the pulse rate exceeds the persistence of vision, a small loudspeaker can be used instead, in series with a 47Ω resistor and a 100µF capacitor.

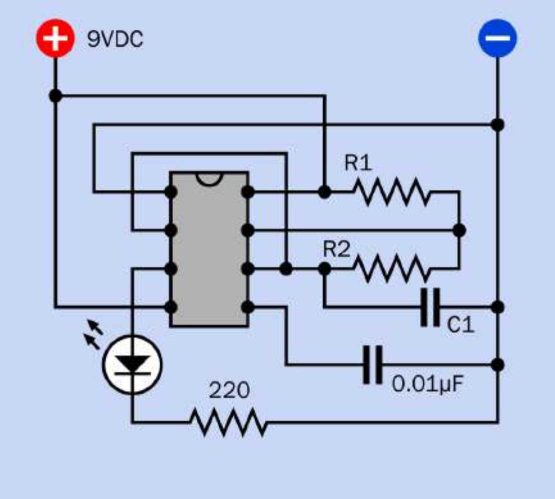

Figure 9-14. *A 555 timer with external connections and components causing it to run in astable (free-running) mode.*

Separate Control of High and Low Output Times

In Figure 9-15, a bypass diode has been added around R2. The capacitor now charges primarily through R1, as the diode has a much lower effective resistance than R2. It discharges only through R2, as the diode blocks current in that direction. Consequently, the length of the high output pulse can be adjusted with the value of R1 only, while the length of the low output pulse can be adjusted with the value of R2 only. The duration of the high output can be lower than, or equal to, the duration of the low output, which is not possible with the basic configuration of components in Figure 9-14.

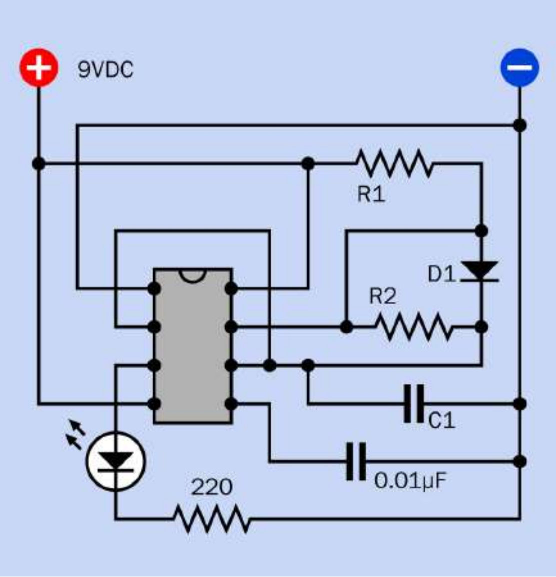

Figure 9-15. *In this circuit, a diode bypasses R2, so that the "on" time and the "off" time of the 555 timer can be set independently of each other, with R1 and R2, respectively.*

555 Fifty Percent Astable Duty Cycle: 1

In Figure 9-16, the circuit enables a fixed astable output duration of approximately 50% high and 50% low. Initially, C1 has no charge, pulling the Input of the timer low, and causing it to begin a cycle with a high pulse from the Output pin, as usual. In this demonstration circuit, the output illuminates an LED. At the same time, resistor R1 is attached to the output and charges C1. When the voltage on C1 reaches 2/3 of V+, this is communicated to the timer Input pin, which ends the "high" cycle and initiates low status on the Output pin. This starts to sink the charge from C1, through R1. When the voltage drops to 1/3 V+, this initiates a new cycle. Because only one resistor is used to charge and discharge the capacitor, we may imagine that the charge and discharge times should be identical. However, a higher load on the output will probably pull down the output voltage to some extent, lengthening the charge time. Conversely, a load on the Output pin that has low resistance will probably sink at least some of the charge from the capacitor, shortening the discharge cycle.

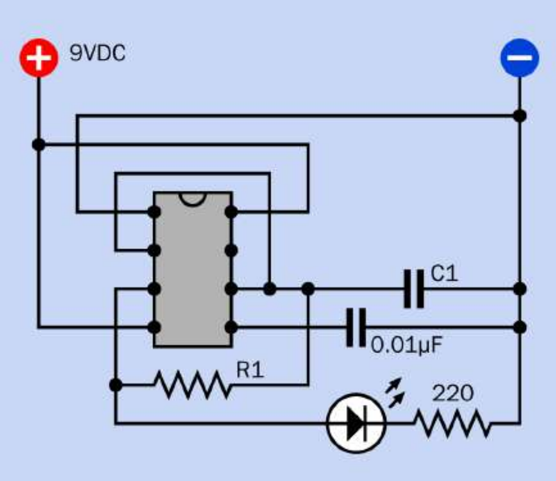

Figure 9-16. *This configuration provides an approximate 50-50 on-off duty cycle at the output pin, although the precise duration will depend on the load.*

555 Fifty Percent Astable Duty Cycle: 2

In Figure 9-17, a small modification of the basic astable circuit shown in Figure 9-14 provides another way to enable a 50% duty cycle. Compare the two schematics, and you will see that just the connection between R1 and R2 has been altered so that C1 now charges only through R1, and discharges only through R2. However, in this configuration the capacitor is discharging into a voltage divider created by the two resistors. Empirical adjustment of the resistor values may be necessary before the duty cycle is precisely 50%.

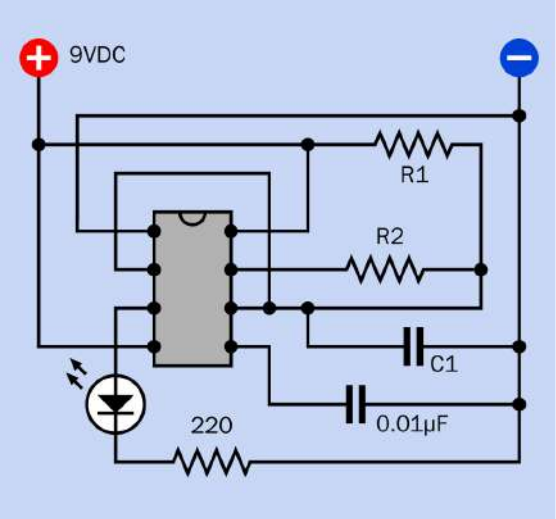

Figure 9-17. *An alternative configuration to provide an approximate 50-50 on-off duty cycle in a 555 timer.*

Use of the 555 Control Pin

In Figure 9-18, a potentiometer and two series resistors allow a varying voltage to be applied to the Control Pin. This will lengthen or reduce both the charge and the discharge times of the timing capacitor. If values for the capacitor and its associated resistors are chosen to create a frequency of approximately 700Hz, a 10K potentiometer should demonstrate more than an octave of audible tones through the loudspeaker. Other components can be substituted for a potentiometer, creating possibilities for producing pulse-width modulation. Alternatively, if a large capacitor is added between the Control pin and ground while a second 555 timer, running slowly in astable mode, applies its output to the Control pin, the charging and discharging of the capacitor will apply a smoothly rising and falling voltage. If the first 555 timer is running at an audio frequency, the output will have a "wailing siren" effect.

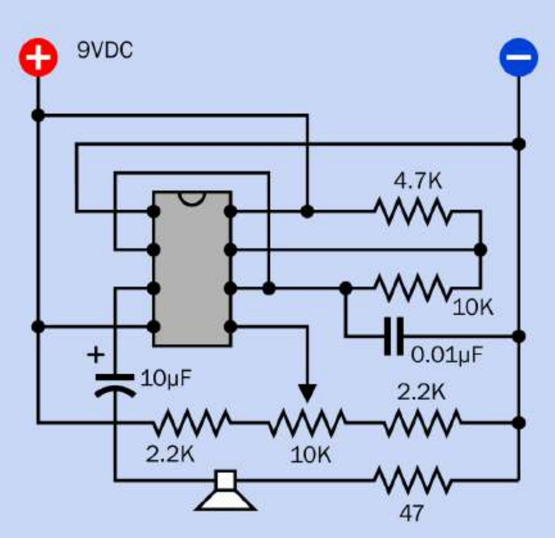

Figure 9-18. *A circuit that allows adjustment of the astable 555 frequency by increasing or lowering the voltage on the Control pin.*

Figure 9-19 shows the components specified in Figure 9-18 mounted on a breadboard.

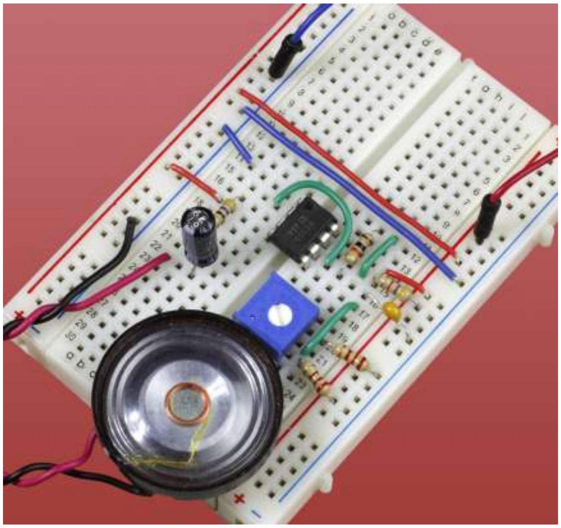

Figure 9-19. *The components in the previous schematic are shown here mounted on a breadboard. They will generate an audio output ranging between approximately 425Hz and 1,075Hz. A lower value for the timing capacitor will shift the audio range higher in frequency.*

555 Flip-Flop Emulation

The flip-flop inside a 555 timer can be accessed to control the timer's outputs. In Figure 9-20, pushbutton switch S1 applies a negative pulse to the Input pin, creating a high output from the timer, which illuminates LED D1. Normally the pulse length would be limited by the charge time of a capacitor attached to the Threshold pin, but in this circuit there is no capacitor, and the Threshold pin is hard-wired to negative ground. So, it never rises to 2/3 of positive power, and the output from the timer remains high indefinitely.

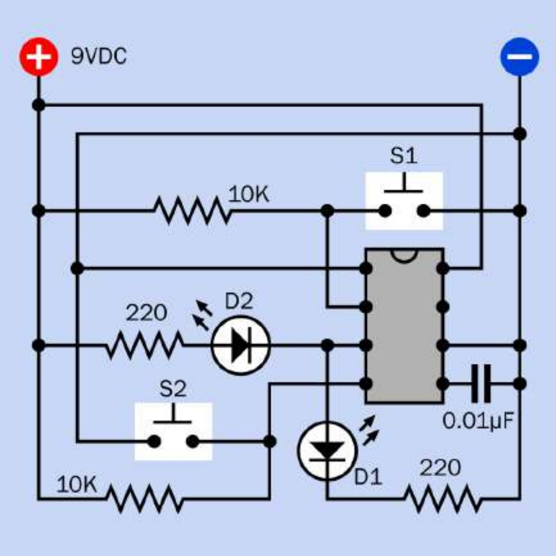

Figure 9-20. *A 555 timer can have its timing features disabled so that it functions as a flip-flop.*

However, if S2 is pressed, it grounds the Reset pin of the timer, which ends the high output and pulls the Output pin down to a low state. D1 goes out and D2 lights up, as the timer is now sinking current through it. When S2 is released, the timer output remains low and D2 remains illuminated, because the Input pin is held high by a pullup resistor. Therefore, the timer now functions in bistable mode, as a flip-flop. While this may be seen as an inappropriate use of the chip, because its full functionality is being disabled, its ability to deliver substantial current and to tolerate a wide range of supply voltages may make it more convenient to use than a digital flip-flop. See Chapter 11 for more information about flip-flops.

A 555 timer emulating a flip-flop is shown on a breadboard in Figure 9-21.

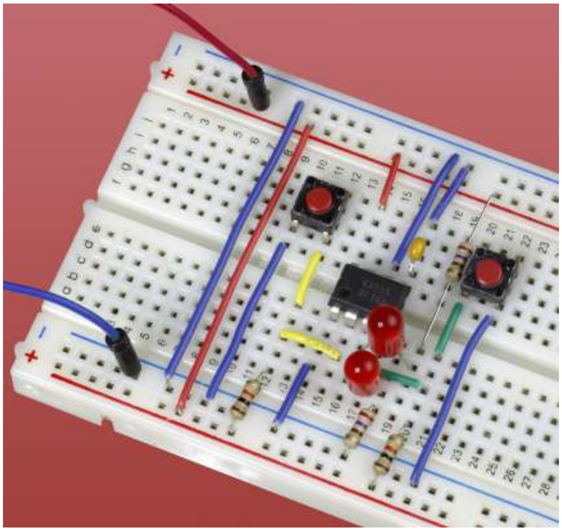

Figure 9-21. *The schematic in which a 555 timer acts as a flip-flop is shown here adapted for a breadboard.*

555 Hysteresis

The comparators inside a 555 timer enable the chip to produce hysteresis. In Figure 9-22, the Input pin and the Threshold pin are shorted together, and C1, the timing capacitor, is omitted. A 10K potentiometer, wired as a voltage divider, delivers a voltage to the Input pin ranging smoothly from V+ to negative ground. As the input dips below 1/3 V+, the Output pin goes high, lighting LED D1. Now if the input voltage gradually rises, the output remains high, even as the input rises above 1/3 V+. The output state is "sticky" because the timer does not end an output pulse until the Threshold pin tells it to, by reaching 2/3 V+. When this finally occurs, the Output pin goes low, D1 goes out, and D2 comes on, sinking current into the Output pin.

Suppose, now, the input voltage starts to go down again. Once again the output state is "sticky" because it remains low until the Input pin drops below the 1/3 level. When that happens, the output finally flips back to a high state, D2 goes out, and D1 comes on.

In the "dead zone" between 1/3 and 2/3 of supply voltage, the timer remains in its current mode, waiting for the input to stray outside of those limits. This behavior is known as *hysteresis*, and is of special importance when processing a varying signal, such as the voltage from a temperature sensor, to control an on/off device such as a thermostat. In fact the 10K potentiometer in this demo could be replaced with a thermistor or a phototransistor, wired in series with a resistor to create a voltage divider which will have an input range compatible with the 555 timer. The hysteresis can then be adjusted by varying the supply voltage that powers the timer, as this will change the values of 1/3 V+ and 2/3 V+. Alternatively, varying the voltage on the Control pin will affect the hysteresis.

A **comparator** can provide much more versatile control of hysteresis by using positive feedback (see Chapter 6 for additional details). But the 555 timer provides a quick-and-simple substitute, and its greater ability to source or sink current enables it to be connected with a wider range of other components.

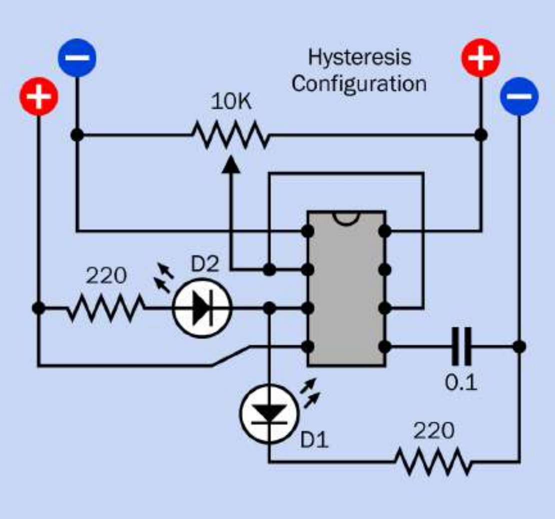

Figure 9-22. *A 555 timer wired so that it creates hysteresis, when supplied by a variable input voltage.*

555 and Coupling Capacitors

As previously noted, when a basic bipolar 555 timer (and some of its variants) is wired in monostable mode, it will retrigger itself indefinitely if its input remains low. One way to avoid this is by using a coupling capacitor. This will pass a transition from high to low, but will then block a steady subsequent voltage. In Figure 9-23, a phototransistor in series with a resistor provides a variable voltage to the noninverting input of a comparator. The reference voltage of the comparator is adjusted with a potentiometer, and resistor R3 provides positive feedback, ensuring that the output from the comparator will be quick and clean. The output from the 555 timer goes through a transistor to the relay, shown at the bottom.

It is important to see the function of the coupling capacitor, C3, with the pullup resistor, R2, which holds the Input pin of the 555 timer high by default. When the output from the comparator drops from high to low, C3 passes this transition to the Input pin of the timer, momentarily overcoming the positive potential, and triggering the timer. After the timer responds, however, C3 blocks any continuation of low voltage from the comparator. Pullup resistor R2 resumes its function of holding the input high, and prevents the timer from being retriggered.

555 Loudspeaker Connection

A small 8-ohm loudspeaker can be driven from the output of a bipolar 555 timer running in astable mode, but should be isolated from it with a 10µF to 100µF capacitor. A series resistor of 47Ω (minimum) should be used. See Figure 9-24.

Burst Mode

It is sometimes useful to create a short beep of fixed length in response to a button-press. The beep should terminate even if the button is held down. This "burst mode" can be achieved with the circuit in Figure 9-25, where the button connects power to a bipolar-type 555 timer running in astable mode, and an RC network applies a decreasing potential to a 47µF capacitor, which is wired to the Reset pin of the timer. The resistor in series with the capacitor will vary the length of the beep. When voltage to the pin drops below approximately 0.3V, output from the timer stops and cannot restart until the button is released.

A resistor of greater than 1.5K may not allow the input value at the reset pin to fall below the voltage, which is necessary to enable a reset. If a lower power supply voltage than 9VDC is used, the resistor value should be higher—for example, a 5VDC power supply works well with a 1.5K to 2K resistor.

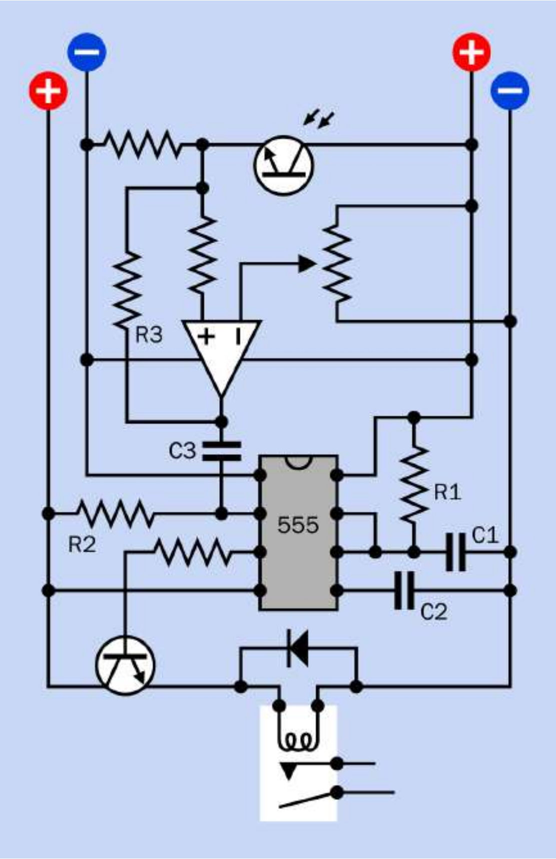

Figure 9-23. *A coupling capacitor (C3) is used in this circuit to isolate the 555 timer from a sustained low input from the comparator. The capacitor only passes a transition from high to low. The rest of the time, the pullup resistor (R3) holds the input high.*

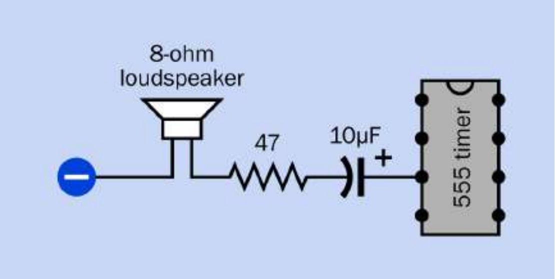

Figure 9-24. *A small 8-ohm loudspeaker can be attached through a capacitor and a resistor to the output of a bipolar 555 timer.*

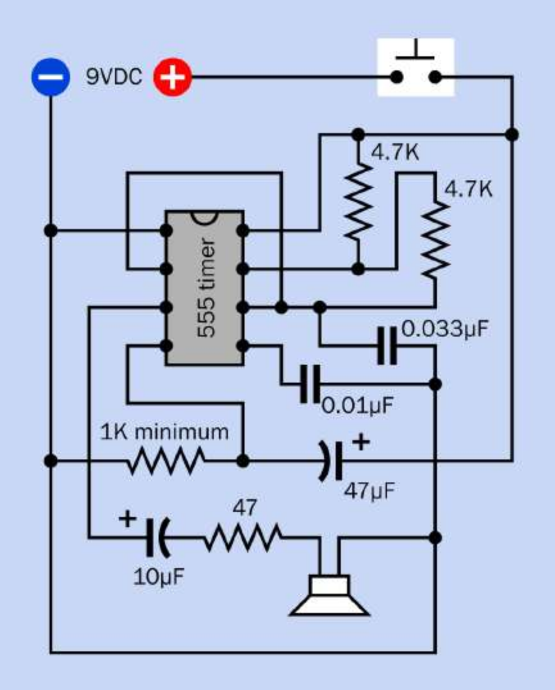

Figure 9-25. *An RC circuit, wired to apply a decreasing voltage on the Reset pin of a bipolar 555 timer, will shut off the timer shortly after it is powered up. This can be used to create a fixed-length beep in response to a button press of any duration.*

Figure 9-26 shows the components installed on a breadboard.

"You Lose" Game Sound

A timer is a simple, cheap way to create a variety of simple game sounds. The schematic in Figure 9-27 makes a groaning sound as the 100µF capacitor wired to the Control pin of a bipolar-type 555 gradually charges through the 1K resistor. Note that if a larger resistor is paired with a smaller capacitor, the effect will differ. The 150K resistor is included to discharge the capacitor reasonably quickly in time for the next cycle.

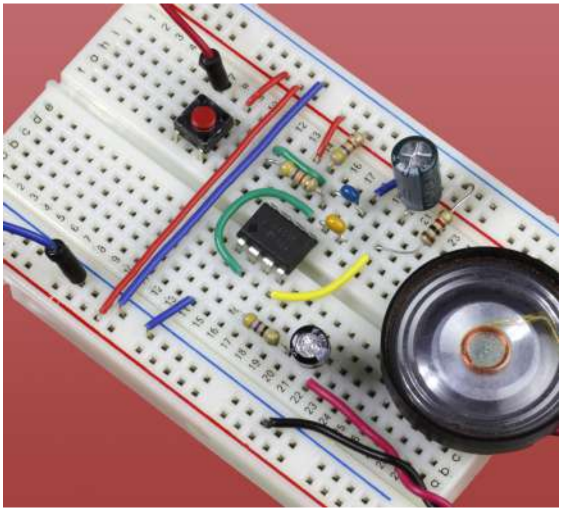

Figure 9-26. *The "burst mode" circuit installed on a breadboard with a miniature loudspeaker.*

What Can Go Wrong

Dead Timer

Like any chip, the 555 can be damaged by overvoltage, excessive source current or sinking current, static electricity, incorrectly applied polarity of power supply, and other forms of abuse. The TTL version of the timer is fairly robust, but the CMOS type much less so. Check for obvious errors such as lack of supply voltage, incorrect or ambiguous input voltages, and unusual current draw (too high, or none at all, at the V+ pin). Use the meter probes on the actual pins of the chip, in case there is a break in the wiring that feeds them. Because timer chips are cheap, a reserve supply of them should be maintained.

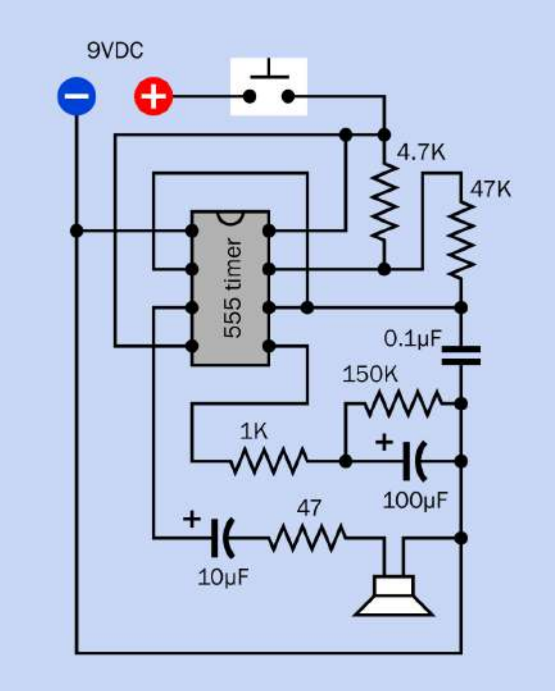

Figure 9-27. *An RC circuit, wired to apply an increasingly positive voltage to the Control pin of a bipolar timer running in astable mode, will gradually pull down the frequency at the Output pin, creating a sound that may be useful in simple game applications.*

CMOS Confused with Bipolar

The part numbers of some bipolar chips are very similar to those of some CMOS versions, and the chips look physically identical. But the CMOS version is easily overloaded, as it may source only 10mA to 20mA maximum while the TTL version is capable of 200mA. Make sure that your chips are carefully labeled when they are stored.

The Pulse that Never Ends

If a 555 timer responds correctly to a high-to-low transition on the input pin, but the output pulse continues indefinitely, check the voltage on pin 6 to see if the timing capacitor is charging above 2/3 of V+. While a 555 can run from 5VDC, a high-current device on the Output pin can pull down the voltage inside the chip to the point where the capacitor never charges sufficiently to end the cycle.

Also check that the input transition from high to low lasts for a shorter time than the pulse. A persistent low input can retrigger the timer.

Erratic Chip Behavior

Possible causes include:

- Floating pins. The Input pin, in particular, should always be connected with a defined voltage (via a 10K pullup resistor, if necessary), and must not be allowed to float at an indeterminate potential.
- Voltage spikes. A timer can be triggered by transients from other components, especially inductive loads. If the input to a monostable timer dips for even a fraction of a second, the timer will initiate a new cycle. A *protection diode* should be used in conjunction with an inductive load.
- Voltage spikes can also introduce variations in the pulse train from an astable timer.
- TTL versions of the 555 timer will tolerate a wide range of supply voltage, but if a voltage regulator is not used, fluctuations in voltage can have unpredictable consequences.

Interference with Other Components

Because the bipolar version of a 555 timer creates a voltage spike when its output changes state, it can interfere with the normal function of other components, especially CMOS chips. A 0.1µF bypass capacitor can be applied between the timer's V+ pin and ground.

Erratic Behavior of Output Devices

If a 555 timer powers an output device such as a relay, and the relay is not opening or closing in a reliable manner, first check that it is receiving sufficient voltage. If the 555 timer is powered with 5VDC, its output will be only around 4VDC.

This problem can be avoided by using the output from the timer to control the voltage on the base of a transistor which switches a separate source of power to the relay coil.

Fatal Damage Caused by Inductive Loads

While it is possible to drive an inductive load such as a small motor or relay directly from a TTL 555 timer, two precautions should be taken. First, the motor or the coil of the relay should have a clamping diode added around it, as is standard practice. Second, because the output of the timer is capable of sinking current as well as sourcing current, it can be protected from sinking back-EMF by inserting a diode in series with the load. This is illustrated in Figure 9-28.

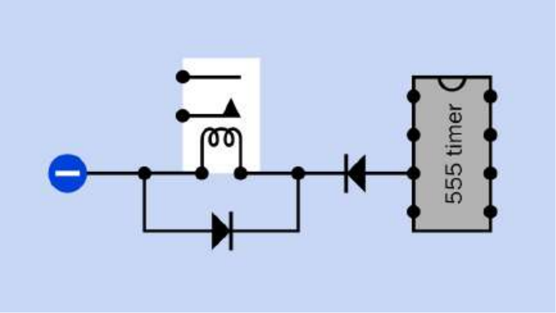

Figure 9-28. *In addition to a standard protection diode clamped around an inductive load such as a relay coil, the 555 timer can be protected from back-EMF by adding a diode in series. The series diode must of course be rated to carry sufficient current through the coil. When choosing a relay, allowance must be made for the voltage drop that will be imposed by the series diode.*

logic gate

10

Only basic **logic gates** are included in this entry—that is, components that perform a Boolean logic operation on two to eight inputs (or one input, in the case of an inverter) to create a single high or low logical output.

OTHER RELATED COMPONENTS

- **flip-flop** (see Chapter 11)

What It Does

A **logic gate** is a circuit that delivers an output, either high or low, depending on the states of its two inputs, either or both of which can be high or low.

Some gates may have more than two inputs, and an *inverter* only has one input, but the basic gates all conform with the two-input, one-output model. The components that constitute a logic gate are almost always etched into a wafer inside a silicon chip.

In a digital computer, a *high logic state* is traditionally close to 5VDC and represents a value of 1 in *binary arithmetic*, while a *low logic state* is traditionally close to 0VDC and represents a binary 0. Modern devices may use a lower voltage for the high state, but the principle is still the same.

A small network of logic gates can perform binary addition, and all other operations in a digital computer are built upon this foundation.

Origins

The concept of digital logic originated in 1894, when English mathematician George Boole announced his invention of a form of algebra (now referred to as *Boolean algebra*) to analyze combinations of two logical states that could be interpreted as "true" and "false." This concept had few practical applications until the 1930s, when Claude Shannon saw that because a basic switch has two states, Boolean algebra could enable analysis of complex networks of switches that were being used in telephone systems.

Because the state of a switch could also be used to represent the values 0 or 1 in binary arithmetic, and because a transistor could function as a switch, Boolean algebra was implemented in solid-state digital computing equipment.

How It Works

While conventional arithmetic uses *arithmetical operators* to represent procedures such as addition or multiplication, Boolean algebra uses *Boolean operators*. The operators of special interest in digital electronics are named AND, NAND, OR, NOR, XOR, and XNOR.

Although each gate actually contains multiple transistors, it is represented by a single logic symbol, as shown in Figure 10-1. The names of the Boolean operators are customarily printed all in caps. A gate requires a power supply and a connection with negative ground, separate from its inputs, but these connections are omitted from gate schematics because they are assumed to exist.

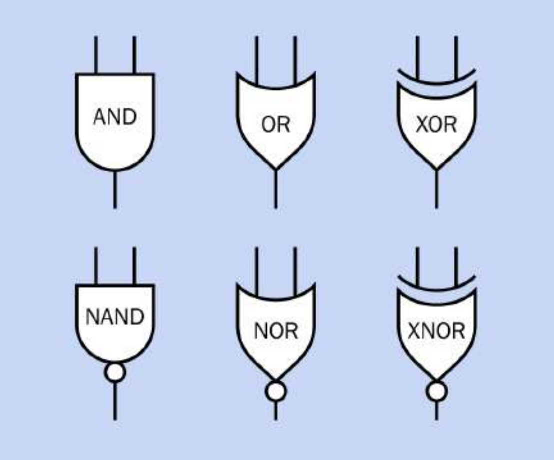

Figure 10-1. *Six types of two-input logic gates are used in digital electronics, although the XNOR gate is rare, as it has few applications. The names are customarily printed in uppercase letters.*

The functions of the gates with two inputs can be defined in electrical terms. In Figure 10-2, the four possible combinations of inputs are tabulated in the left column, with red indicating a high input and black indicating a low input. The corresponding output from each gate is shown beneath its name. This kind of tabulation is known as a *truth table*, as it is derived from Boolean algebra which originally concerned itself with "true" and "false" states.

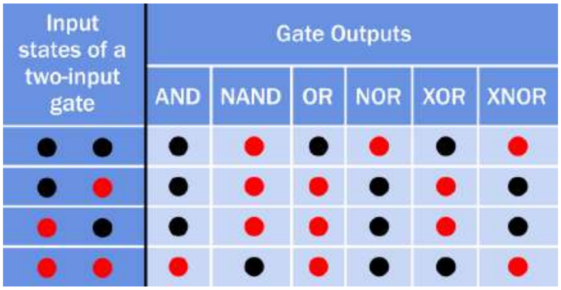

Input states of a two-input gate		Gate Outputs					
		AND	NAND	OR	NOR	XOR	XNOR
●	●	●	🔴	●	🔴	●	🔴
●	🔴	●	🔴	🔴	●	🔴	●
🔴	●	●	🔴	🔴	●	🔴	●
🔴	🔴	🔴	●	🔴	●	●	🔴

Figure 10-2. *The four possible combinations of input states in a 2-input logic gate are shown at left. The corresponding output from each gate is shown beneath its name. Red indicates a high state, while black indicates a low state.*

The truth table assumes that *positive logic* is being used. Negative logic is very uncommon, but if it were used, the red dots in the truth table would correspond with low inputs and outputs, while the black dots would correspond with high inputs and outputs.

Inversion

The small circles appended to the outputs of NAND, NOR, and XNOR gates mean that the output of each gate is inverted compared with the AND, OR, and XOR gates. This can be seen by inspection of the output states shown in Figure 10-2. The circles are known as *bubbles*.

Sometimes logic symbols are shown with a bubble applied to one input, as in Figure 10-3. In these cases, the circle indicates that an input must be inverted. More than one gate may be needed to achieve this logic function in an actual circuit. The style is often used to show the inner workings of an IC, using a minimum number of logic symbols.

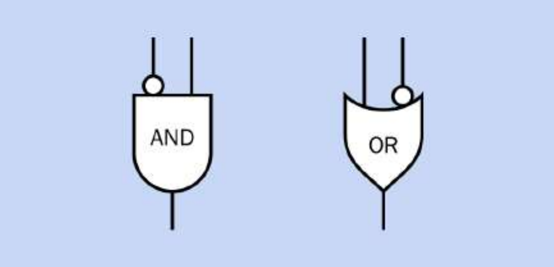

Figure 10-3. *The circle in a logic gate symbol indicates that a signal is being inverted. Circles can be inserted at gate inputs, but in a real circuit a separate inverter is likely to be needed to create this effect.*

Single-Input Gates

Two gates exist that have a single input and a single output, shown in Figure 10-4. The *buffer* should not be confused with the symbol for an **op-amp** or a **comparator**. (Those components always have two inputs.) The output state of a buffer is the same as its input state, but the component may be useful to deliver more current or to isolate one section of a circuit from another.

When a bubble is appended to a buffer, it becomes a NOT gate, more commonly known as an *inverter*. Its function is to create an output state that is opposite to its input state.

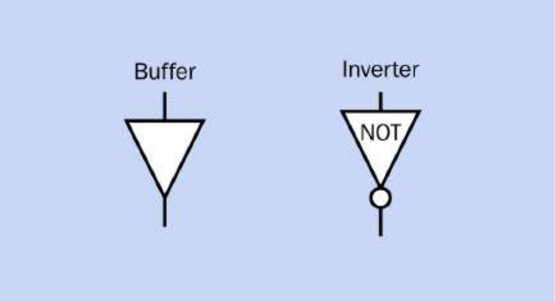

Figure 10-4. *The two logic gates that have only one input and one output. Note that in some schematics showing internal logic of ICs, the bubble on an inverter may be found on the input side instead of the output side.*

Gates with More than Two Inputs

AND, NAND, OR, and NOR gates can have any number of inputs, as suggested in Figure 10-5, although practical factors usually limit the inputs to a maximum of eight.

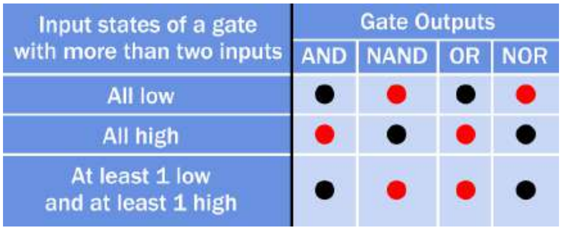

Input states of a gate with more than two inputs	Gate Outputs			
	AND	NAND	OR	NOR
All low	●	🔴	●	🔴
All high	🔴	●	🔴	●
At least 1 low and at least 1 high	●	🔴	🔴	●

Figure 10-5. *The previous table has been modified to show the outputs from logic gates that have more than two inputs. XOR and XNOR gates are not included in the table, because a strict interpretation of their logic requires that a unique output state exists if one input is high while the other is low.*

The rules can be summarized like this:

- Output from an AND gate: Low if any of its inputs is low, high if all of its inputs are high.
- Output from a NAND gate: High if any of its inputs is low, low if all of its inputs are high.
- Output from an OR gate: High if any of its inputs is high, low if all of its inputs are low.
- Output from a NOR gate: Low if any of its inputs is high, high if all of its inputs are low.

In the case of XOR and XNOR gates, their logic requires that a unique output state must exist if one input is high while the other input is low.

In fact, so-called three-input XOR gates do exist, an example being the 74LVC1G386 chip, in which the output is high if all three inputs are high, or if one input is high, but not if two inputs are high or no inputs are high. Further discussion of more-than-two-input XORs is outside the scope of this encyclopedia.

Boolean Notation

For reference, the original written notation for Boolean operators is shown in Figure 10-6. Unfortunately, the notation for these operators was never properly standardized, and in more than one instance, multiple symbols acquired the same meaning. The letters P and Q are often, but not always, used to represent two input states that can be true or false.

- The use of a horizontal line above a symbol, to indicate that its state has been reversed, has carried over to datasheets where this notation can show that an output state from any digital chip is inverted. The line is known as a *bar*.

Arithmetical Operations

Suppose we wish to sum two binary numbers, each containing two digits. There are four digits altogether, and depending on their values, there are 16 different possible addition sums, as shown in Figure 10-7.

If A0 and B0 represent the rightmost digits of the two numbers being added, and S0 is the sum of those two digits, inspection of the figure shows that the sum can be derived using just three rules:

1. If A0 = 0 and B0 = 0, then S = 0.
2. If A0 and B0 have opposite states, then S0 = 1.
3. If A0 = 1 and B0 = 1, then S0 = 0, and carry 1 to the next place left.

If A0 and B0 are the two inputs to an XOR logic gate, the output of the gate satisfies all three rules, except the need to carry 1 to the next place left. This last function can be satisfied with an AND gate. The function of two gates is known as a *half adder*, and is shown in the top section of Figure 10-8.

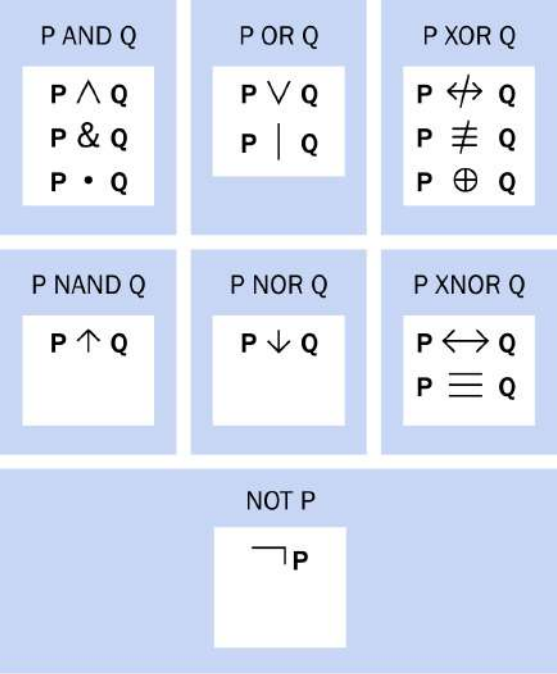
P AND Q
P ∧ Q
P & Q
P • Q
P OR Q
P ∨ Q
P | Q
P XOR Q
P ↮ Q
P ≢ Q
P ⊕ Q
P NAND Q
P ↑ Q
P NOR Q
P ↓ Q
P XNOR Q
P ↔ Q
P ≡ Q
NOT P
¬P

Figure 10-6. *Boolean operators as they have been expressed in written notation. Lack of standardization has resulted in more than one symbol representing some of the operators.*

When we consider the next pair of binary digits to the left, the situation now becomes more complicated, because we may be carrying 1 into this addition sum from the previous stage, and we still need to be able to to carry 1 out (if necessary) to the next stage. An assembly of five logic gates can deal with this, and their combination is known as a *full adder*. This is shown in the bottom section of Figure 10-8.

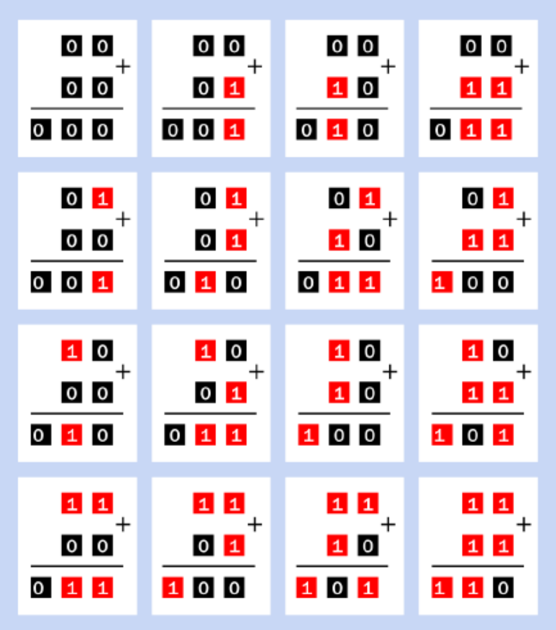

Figure 10-7. *Sixteen different addition sums are possible, when summing two binary numbers of two digits each.*

The combination of XOR and AND gates shown in Figure 10-8 is not the only one that works to add binary numbers. However, it may be the most intuitively obvious.

Other Operations

Binary arithmetic remains the most important application of logic gates, but individually packaged gates are seldom used for that purpose anymore. They were long since subsumed into large multifunction computing chips.

Single gates still have application in small systems, or to modify the inputs and outputs of microcontrollers, or to convert the output from one complex digital chip to make it compatible with the input of another. This last application is often referred to as *glue logic*.

Applications for single gates are discussed in "How to Use It" on page 103.

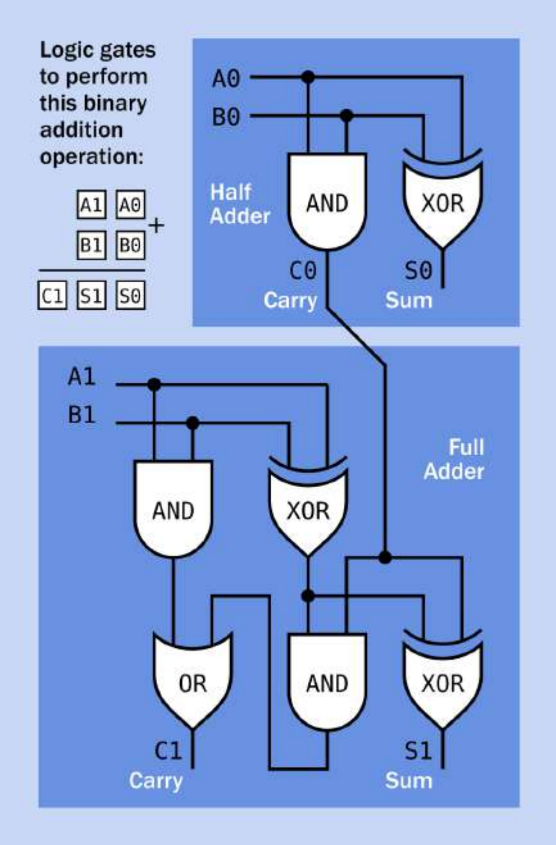

Figure 10-8. *Logic gates can be used to add binary numbers, using a high input or output to represent a binary 1 and a low input or output to represent a binary 0. This schematic shows one possible way for gates to add two two-digit binary numbers.*

Variants

Chips containing logic gates were introduced in the 1960s. The 7400 NAND chip, from Texas Instruments, was the first of a series that became so influential, the same basic part numbers (with letters added before, after, and among the digits) are still used today. An example of a currently available through-hole 7400 chip is shown in Figure 10-9.

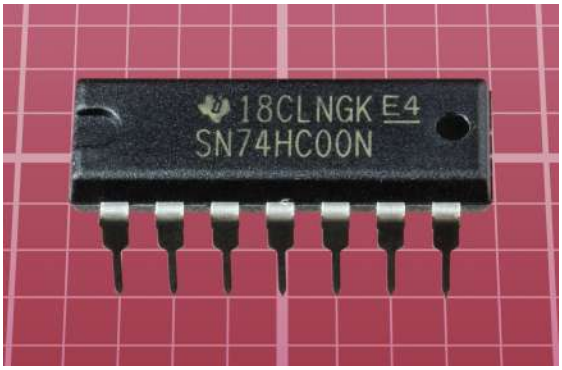

Figure 10-9. *A modern version of a 7400 chip containing four NAND gates.*

Initially, these chips conformed with a *transistor-transistor logic* (*TTL*) standard that had been invented at TRW in 1961 and introduced in commercial products by Sylvania in 1963. It established the now-familiar standard of 5VDC for the power supply. Many logic chips now use other voltages, but the term "high" still means an input or output that is near to the supply voltage, while "low" means an input or output that is near to negative ground. The exact definition of "near" will be found in datasheets for the chips.

The *7400 series* was successful partly because it was engineered for compatibility. The output from one gate could be connected directly to the input of another gate, with a few bypass capacitors added on a circuit board to suppress voltage spikes caused by rapid switching. Earlier components had not been so easy to interface with each other. The new standard dominated the industry to the point where dozens of manufacturers started making chips that conformed with it, and a single board could mix-and-match chips from multiple vendors.

Because many logic chips acquired part numbers that began with 74, they are often referred to as the 74xx series, where other digits (sometimes more than two) can be substituted for xx. This avoids ambiguity, as the very first chip in this format was a NAND gate that had 7400 as its actual part number. In the text below, 7400 will refer to that specific chip, whereas 74xx will refer to the whole series of chips.

RCA introduced a competing family of logic gates in 1968, using CMOS transistors. As each part number began with a 4 and contained four digits, this was referred to as the *4000 series*. The CMOS chips were slower and more expensive, but tolerated a wider range of power supply voltages (3V to 12V, initially). Their biggest advantage was that they used much less current. This was important, as TTL chips created a lot of waste heat. The lower power consumption of CMOS also enabled one chip to control the inputs of many others, which simplified circuit design. This one-to-many relationship is known as *fanout*.

Ultimately, CMOS chips transcended their early limitations. While they were reserved initially for battery-powered devices in which very low power consumption was more important than speed, CMOS is now used almost everywhere, still maintaining its advantage of low current (almost zero, in fact, while a chip is quiescent) while equalling the speed of TTL. However, CMOS logic chips are very often pin-compatible with the old TTL components, and modern CMOS part numbers are often derived from the old 74xx series.

Most CMOS logic chips in the old 4000 series are still available, and may be used in situations where a power supply greater than 5VDC is convenient.

Part Numbers

As the performance of semiconductors gradually improved, successive *families* of logic chips were introduced, identified by one-letter, two-letter, or three-letter acronyms. The acronym was inserted into the part number, so that a 7400 NAND gate in the HC (high-speed CMOS) family became a 74HC00 NAND gate.

Because these chips were available from multiple sources, the part number was also preceded with one or more letters indicating the manufacturer. And because each chip was manufactured in different versions (for example, some complied with military specifications, while others didn't), letters were also appended to the end of the part number. Today, the appended letters may indicate whether the chip is of the old through-hole format, or conforms with a more recent surface-mount format.

Summing up:

- Prefix: manufacturer ID.
- Numerals, omitting any letters in the middle: Chip functionality.
- Middle letters: Chip family.
- Suffix letters: Package format.

Thus, for example, the actual part number for a 74HC00 NAND chip could be SN74HC00N, where the SN prefix indicates that it is manufactured by Texas Instruments and the N suffix means that it is in plastic dual-inline-pin (DIP) format. (The SN prefix was introduced by Texas Instruments in the earliest days of integrated circuits as an acronym for "semiconductor network," meaning that multiple transistors were "networked" on a wafer of silicon. Other manufacturers used their own schemes for part numbering, and so SN became exclusively identified with Texas Instruments.)

The system of augmenting part numbers has been further extended by inserting 1G, 2G, or 3G immediately after the family identifier, to indicate surface-mount chips that contain one, two, or three logic gates. If the "G" identifier is missing, the chip usually has four logic gates, which was the standard used in the original 74xx series. This rule applies even in surface-mount formats, where the surface-mount pads of four-gate chips have the same functions as the pinouts of the original TTL versions (except in the case of square-format surface-mount chips, which are not discussed here).

When searching a catalog to find a chip by its part number, it helps to remember that searching for a 7400 chip may not find any hits, but searching for a 74HC00 (or any other valid number containing a family identifier) is much more likely to be understood.

A key to understanding part numbers is shown in Figure 10-10. The upper part of the figure is a guide to interpreting numbers on a generic basis, while the lower part interprets the specific part number shown.

Families

As of 2013, the HC family in the 74xx series has become so widely used, it can be considered the default in the traditional DIP 14-pin format. Incremental improvements are still being made, and new families are being introduced, primarily in surface-mount formats which use lower power-supply voltages (down to around 1VDC).

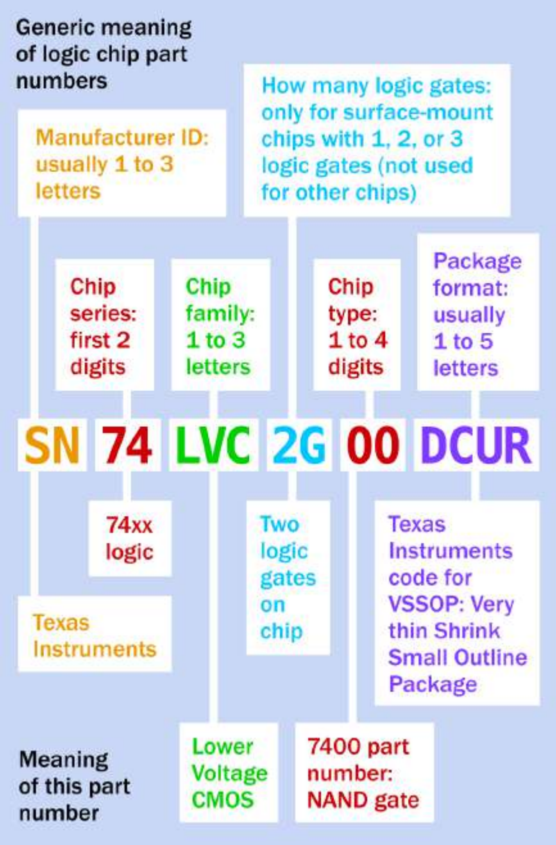
Generic meaning of logic chip part numbers
Manufacturer ID: usually 1 to 3 letters
How many logic gates: only for surface-mount chips with 1, 2, or 3 logic gates (not used for other chips)
Chip series: first 2 digits
Chip family: 1 to 3 letters
Chip type: 1 to 4 digits
Package format: usually 1 to 5 letters
SN 74 LVC 2G 00 DCUR
74xx logic
Two logic gates on chip
Texas Instruments code for VSSOP: Very thin Shrink Small Outline Package
Texas Instruments
Meaning of this part number
Lower Voltage CMOS
7400 part number: NAND gate

Figure 10-10. *How to interpret the segments of a logic chip part number in the 74xx family (in this case, a 7400 NAND gate).*

Here is an historical summary of the most important chip families.

- 74xx: Original series of bipolar TTL chips.
- 74Hxx: Bipolar TTL, high speed, about twice as fast as the original 74xx chips, but twice the power consumption.
- 74Lxx: Bipolar TTL, lower power consumption than the original TTL, but also much lower speed.
- 74LSxx: Bipolar TTL, lower power with Schottky input stages, faster than original TTL. Some LS chips are still being manufactured.
- 74ASxx: Bipolar TTL, Advanced Schottky, intended to supercede the 74Lxx.
- 74ALSxx: Bipolar TTL, Advanced Low Power Schottky, intended to supercede the 74LSxx.
- 74Fxx: Bipolar TTL, Faster.
- 74HCxx: CMOS high-speed emulation of 74LSxx.
- 74HCTxx: CMOS but with similar high-state input voltage threshold to bipolar TTL chips, for compatibility.
- 74ACxx: Advanced CMOS.
- 74ACTxx: Advanced CMOS emulation of TTL with similar high-state input voltage threshold to TTL, for compatibility.
- 74AHCxx: Advanced Higher-Speed CMOS, three times as fast as HC.
- 74VHCxx: Very High Speed CMOS.
- 74AUCxx, 74FCxx, 74LCXxx, 74LVCxx, 74ALVCxx, 74LVQxx, 74LVXxx: Various specifications, many using power supply voltages of 3.3V or below.

In the 4000 series, an early significant improvement was the 4000B family, which allowed a higher power supply limit (18V instead of 12V) and was much less susceptible to damage by static discharge. The 4000B family almost totally replaced the old 4000 family, and most 4000B chips are still available, as they are useful in situations where a power supply delivers more than 5VDC.

- When it is referenced casually, the B at the end of a chip number in the 4000 series may be omitted. When the number is listed in a catalog, the B is included.

Chips with 45 as their first two digits were introduced as a new generation, but were not widely adopted. After that, the 4000 series ceased to evolve, and CMOS chips adopted 74xx part numbers, distinguishing themselves by the insertion of letter groups in the center of the number.

To add to the confusion, some 4000 series part numbers were appended to 74xx part numbers, so that, for example, the 74HCT4060 is designed to be compatible with the old 4060B chip.

Family Interoperability

One of the most important issues relating to chip families is their differing specifications for a low-state voltage and a high-state voltage in inputs and outputs.

The original 74xx TTL series, using a 5VDC power supply, used these approximate specifications:

- Output: 74xx voltage representing a low state (at most 0.4V to 0.5V)
- Input: 74xx input voltage interpreted as a low state (maximum 0.8V)
- Output: 74xx voltage representing a high state (at least 2.4V to 2.7V)
- Input: 74xx input voltage interpreted as a high state (minimum 2V)

This provided a safe margin of error of at least 0.4V when chips were communicating with each other.

In the CMOS 4000 family, however, logic chips required a minimum input of 3V to 3.5V to be interpreted as a high state. The minimum acceptable output from a TTL chip was below this level, creating problems if anyone should try to use the output from a TTL chip to communicate with an input on a CMOS chip.

One solution is to add a 4.7K pullup resistor to the TTL output, guaranteeing that it won't fall too low. But this wastes power, and the need for the resistor is easily forgotten. Another option is to use the HCT or ACT family of CMOS logic. The "T" in these family names indicates that they have been engineered to share the input standards of the old TTL chips. They still deliver the same high output as other types of CMOS, making them seem to be the best possible solution. Unfortunately, it does entail a compromise: the "T" chips are more sensitive to noise, among other factors.

- Ideally, chip families should not be intermingled.

Gates per Chip

Each of the original 74xx chips contained multiple gates within the limits of a uniform 14-pin through-hole format. The gates that were most commonly used had two inputs, and there were four of these gates per chip.

However, the desire for miniaturization, and the use of automatic chip-placement and soldering equipment, made one-gate and two-gate logic chips desirable and practical in surface-mount format. (Three-gate surface-mount chips exist, but are sufficiently unusual that they are not described in this encyclopedia.)

Two Inputs, Single Gate

Where a chip contains just one logic gate, it is almost always a surface-mount component, and the part number has 1G in the middle to indicate "one gate." Pad functions are shown in Figure 10-11. The layout is standardized for all logic gates, with the exception of XNOR gates, which are not manufactured in surface-mount format.

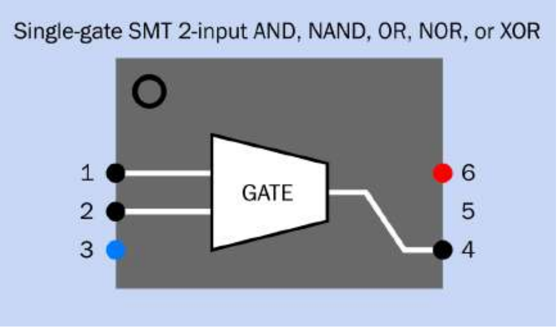

Figure 10-11. *Internal configuration and solder-pad functions for a two-input surface-mount single-gate logic chip that can contain an AND, NAND, OR, NOR, or XOR gate. XNOR gates are not manufactured in this format.*

In the figure, a gate is shown in generic form, indicating that it may be an AND, NAND, OR, NOR, or XOR gate, depending on the part number of the chip. Inputs are on the left of the gate, while its output is on the right. The chip does not have a solder pad in position 5, but the pad at top right is identified as pin 6 for consistency with the numbering pattern in other surface-mount components where six pads are common.

The generic part numbers for single-gate surface-mount two-input logic chips are shown here, with letter x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND gate: x74x1G08x
- OR gate: x74x1G32x
- NAND gate: x74x1G00x
- NOR gate: x74x1G02x
- XOR gate: x74x1G86x

Three Inputs, Single Gate

AND, NAND, OR, and NOR single gates are available with more than two inputs. Their output is determined by rules shown in Figure 10-5. XOR and XNOR gates are not included in the table, because a strict interpretation of their logic requires that a unique output state exists if one input is high while the other is low.

The pad functions for a surface-mount single-gate logic chip with three inputs are shown in Figure 10-12. The generic part numbers for these chips are shown below. Again, each x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

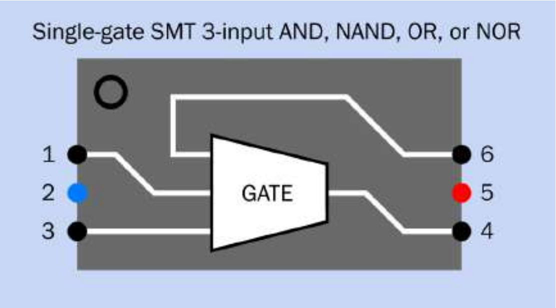

Figure 10-12. *Internal configuration and solder-pad functions for a three-input surface-mount single-gate logic chip that can contain an AND, NAND, OR, or NOR gate.*

The generic part numbers for single-gate surface-mount three-input logic chips are shown here, with each x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND: x74x1G11x
- NAND: x74x1G10x
- OR: x74x1G32x
- NOR: x74x1G27x

Single Gate, Selectable Function

A few surface-mount chips can emulate a variety of two-input gates, by using appropriate external connections. The internal logic of one example, with generic part number x74x1G97x (an actual example would be Texas Instruments SN74LVC1G97), is shown in Figure 10-13. Depending which pin is grounded and which other pins are used as inputs, the chip can emulate all five of the most commonly used gates. To achieve this, however, some inputs have to be inverted.

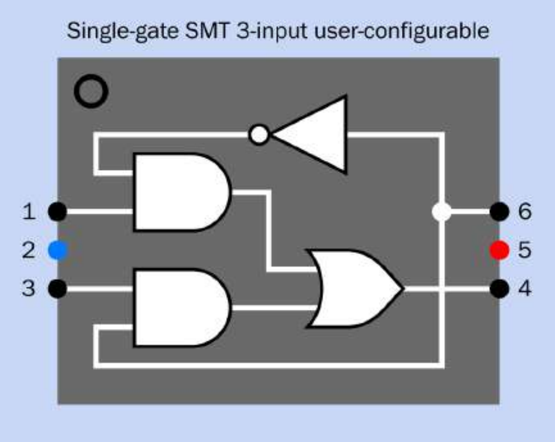

Figure 10-13. *Internal configuration for a configurable surface-mount chip that can emulate various two-input logic gates, depending which inputs are used and which are grounded. Some inputs have to be inverted to emulate some gates.*

Two Inputs, Dual Gate

Two-input surface-mount AND, NAND, OR, NOR, and XOR gates are available in dual layout (two gates per chip). The internal logic and pad functions are shown in Figure 10-14. The generic part numbers for these chips are shown here. Again, each x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

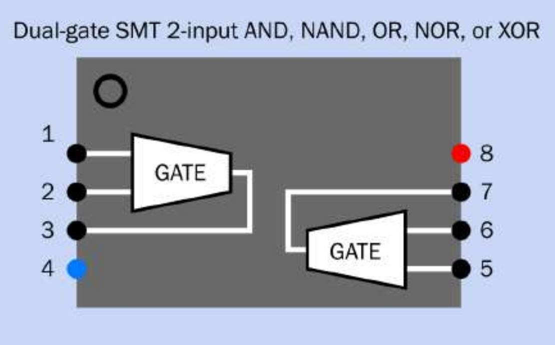

Figure 10-14. *Internal configuration and solder-pad functions for a two-input surface-mount dual-gate logic chip that can contain two AND, NAND, OR, NOR, or XOR gates. XNOR chips are not manufactured in this format.*

The generic part numbers for dual-gate surface-mount two-input logic chips are shown below, with each x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND: x74x2G08x
- NAND: x74x2G00x
- OR: x74x2G32x
- NOR: x74x2G02x
- XOR: x74x2G86x

Original 74xx 14-Pin Format

Each of the original 74xx TTL chips contained multiple gates within the limits of a uniform 14-pin chip format. The available options were, and still are:

- Quad 2-input: Four gates with two inputs each
- Triple 3-input: Three gates with three inputs each
- Dual 4-input: Two gates with four inputs each
- Dual 5-input: Two gates with five inputs each
- Single 8-input: One gate with eight inputs

The five-input chips have become so uncommon that they are not described in this encyclopedia.

Quad Two-Input 74xx Pinouts

14-pin DIP 74xx quad two-input logic chips are available in AND, NAND, NOR, XOR, or XNOR versions, all of which have an internal layout shown in Figure 10-15. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The four connections leading to a gate are its inputs, while the single connection from a gate is its output.

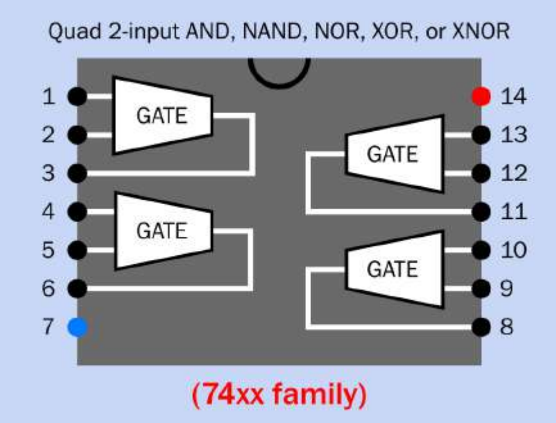
Quad 2-input AND, NAND, NOR, XOR, or XNOR
1
2
3
4
5
6
7
GATE
GATE
GATE
GATE
14
13
12
11
10
9
8
(74xx family)

Figure 10-15. *In a 14-pin quad two-input 74xx logic chip, the AND, NAND, NOR, XOR, and XNOR versions all share this generic layout.*

- The 14-pin quad two-input OR chip has different pinouts from all the other 74xx logic chips. It is shown in Figure 10-16.

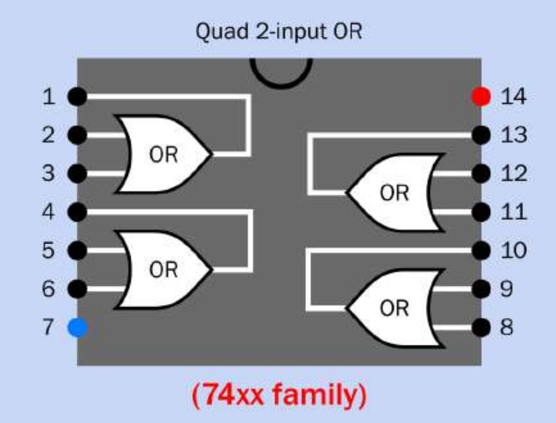
Quad 2-input OR
1
2
3
4
5
6
7
OR
OR
OR
OR
14
13
12
11
10
9
8
(74xx family)

Figure 10-16. *In a quad two-input 74xx OR chip, this layout is used, which is different from that used in all the other quad two-input logic gates.*

Triple Three-Input 74xx Pinouts

The AND, NAND, and NOR versions of a 14-pin DIP 74xx triple three-input logic chip all have an internal layout shown in Figure 10-17. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. Three connections leading to a gate are its inputs, while a single connection from a gate is its output.

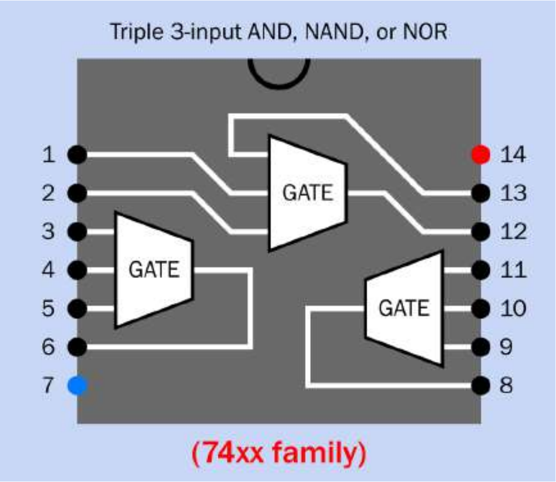
Triple 3-input AND, NAND, or NOR
1
2
3
4
5
6
7
GATE
GATE
GATE
14
13
12
11
10
9
8
(74xx family)

Figure 10-17. *In a 14-pin triple three-input 74xx logic chip, the AND, NAND, and NOR versions all share this generic layout.*

- The 14-pin triple three-input OR chip has different pinouts from all the other 74xx logic chips. It is shown in Figure 10-18.

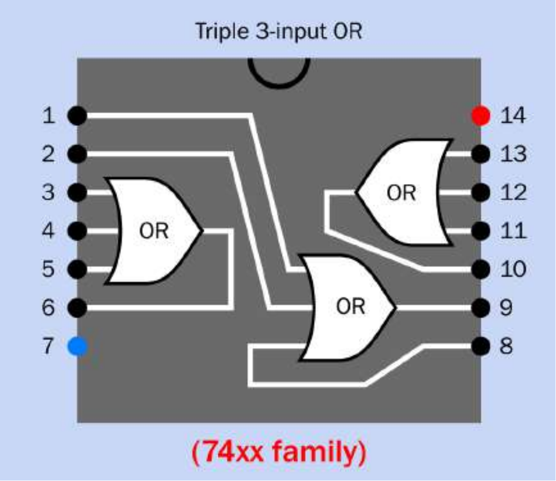
Triple 3-input OR
1
2
3
4
5
6
7
OR
OR
OR
14
13
12
11
10
9
8
(74xx family)

Figure 10-18. *In a triple three-input 74xx OR chip, this layout is used, which is different from that used for all the other triple three-input logic gates.*

Dual Four-Input 74xx Pinouts

A 14-pin DIP 74xx dual four-input logic chip contains two four-input gates. The AND, NAND, and NOR versions all have an internal layout shown in Figure 10-19. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type.

- There is no OR chip of the 14-pin dual four-input type.

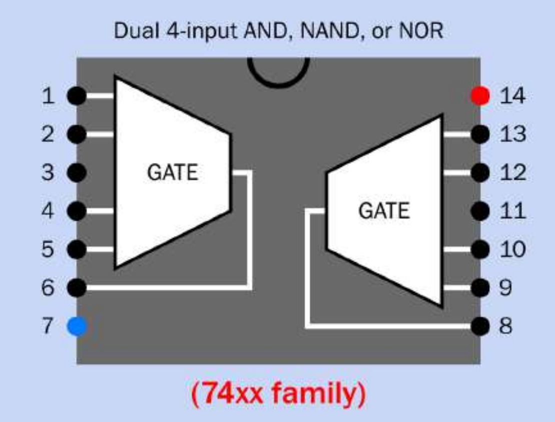

Figure 10-19. *In a 14-pin dual four-input 74xx logic chip, the AND, NAND, and NOR versions all share this generic layout. There is no 74xx OR chip with four inputs per gate.*

Single Eight-Input 74xx Pinouts

A 14-pin DIP 74xx single eight-input NAND chip contains one eight-input gate, as shown in Figure 10-20. The layout is unchanged in surface-mount format.

- There is no AND chip of the 14-pin single eight-input type.

A 14-pin eight-input logic chip in the 74xx series, able to function as both an OR and a NOR, is shown in Figure 10-21. The output from the NOR gate is connected with pin 13, but also passes through an inverter to create an OR output at pin 1. (Because a NOR gate is equivalent to an inverted-OR, when its output is inverted again, it returns to being an OR.)

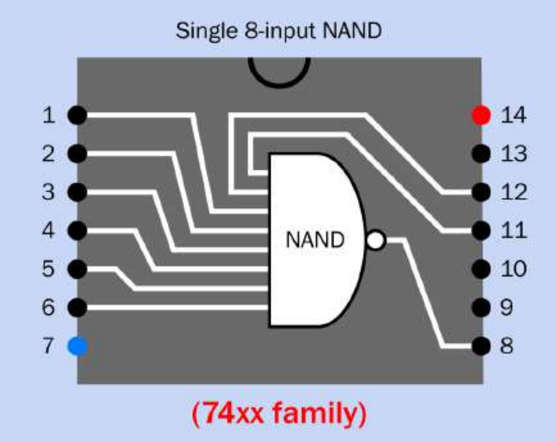

Figure 10-20. *The internal layout of single eight-input NAND chip in the 14-pin 74xx series. There is no 74xx AND chip with eight inputs per gate.*

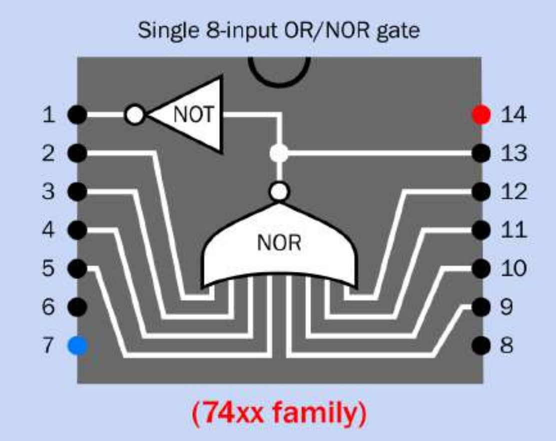

Figure 10-21. *The internal layout of single eight-input OR/NOR chip in the 14-pin 74xx series. Pin 13 has the NOR output, while pin 1 has the OR output.*

The following list shows the generic part numbers for DIP and surface-mount versions of 14-pin logic chips in the 74xx series that have two or more inputs per gate. As before, an x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

- Quad 2-input AND: x74x08x
- Quad 2-input NAND: x74x00x
- Quad 2-input OR: x74x32x
- Quad 2-input NOR: x74x02x
- Quad 2-input XOR: x74x86x
- Quad 2-input XNOR: x74x266x
- Triple 3-input AND: x74x11x
- Triple 3-input NAND: x74x10x
- Triple 3-input OR: x74x4075x
- Triple 3-input NOR: x74x27x
- Dual 4-input AND: x74x21x
- Dual 4-input NAND: x74x20x
- Dual 4-input NOR: x74x4002x
- Single 8-input NAND: x74x30x
- Single 8-input OR/NOR: x74x4078x

74xx Inverters

Single, dual, and triple inverter packages in the 74xx series are available in surface-mount format only. Their internal arrangement is shown in Figures 10-22, 10-23, and 10-24.

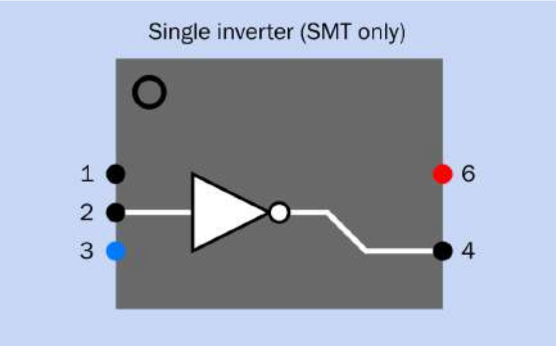

Figure 10-22. *The internal layout of a 74xx series logic chip containing one inverter. This is available in surface-mount format only. Pin 5 is absent. Pin 1 is not connected.*

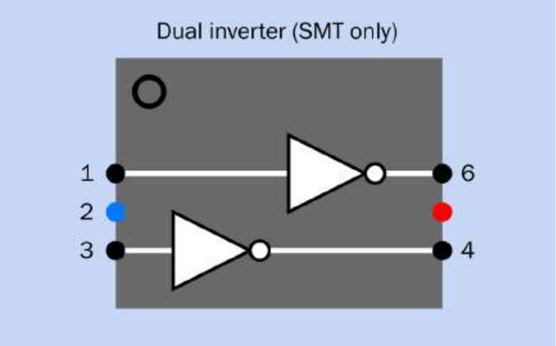

Figure 10-23. *The internal layout of a 74xx series logic chip containing two inverters. This is available in surface-mount format only.*

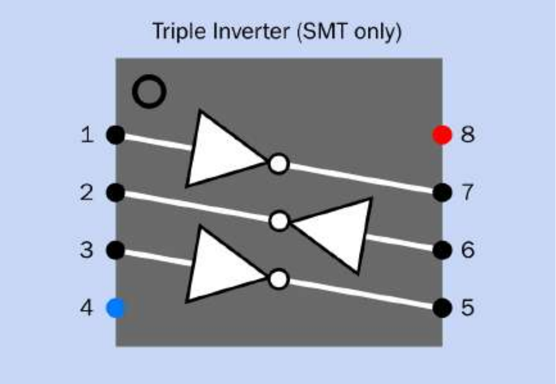

Figure 10-24. *The internal layout of a 74xx series logic chip containing three inverters. This is available in surface-mount format only.*

In the 14-pin format, a hex inverter chip (containing six inverters) is available, as shown in Figure 10-25. The layout is the same for DIP and surface-mount formats.

Generic part numbers for inverter chips are as follows:

- Single inverter: x74x1G04x
- Dual inverter: x74x2G04x
- Triple inverter: x74x3G14x
- Hex inverter: x74x04x

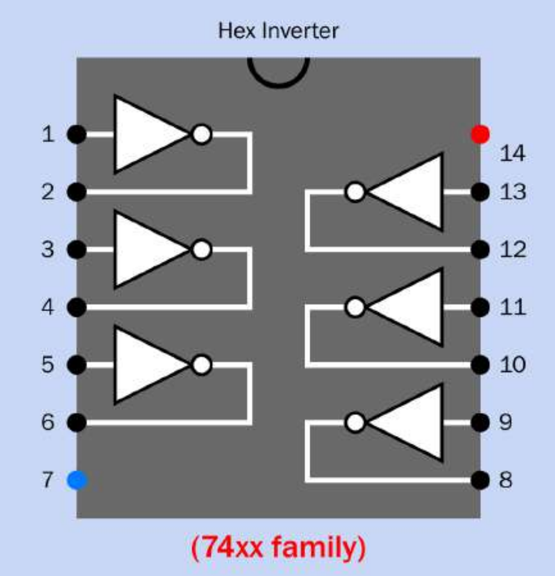

Figure 10-25. *The internal layout of a 14-pin 74xx hex inverter logic chip, containing six inverters. This layout is the same for DIP and surface-mount versions.*

Additional Variations

Some chips in the 74xx series (both DIP and surface mount versions) have variants with open drain or open collector outputs, while others have inputs that are configured as Schmitt triggers. These variants will be found as hits when searching supplier websites for logic chips by gate name and number of inputs.

Pinouts in the Original 4000 Series

Each of the original 4000 CMOS chips contained multiple gates within the limits of a uniform 14-pin chip format. The available options were, and still are:

- Quad 2-input: Four gates of two inputs each
- Triple 3-input: Three gates of three inputs each
- Dual 4-input: Two gates of four inputs each
- Single 8-input: One gate of eight inputs

In the 4000 family, 14-pin quad two-input logic chips are available in AND, OR, NAND, NOR, XOR, or XNOR versions, all of which have an internal layout shown in Figure 10-26. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The four connections leading to a gate are its inputs, while the single connection from a gate is its output.

Unlike the 74xx family, the quad two-input OR chip in the 4000 family has the same pinouts as the other types of quad two-input logic chips.

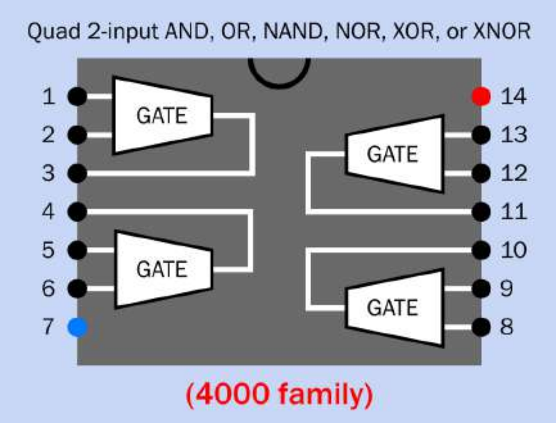

Figure 10-26. *In the 4000 family, the AND, OR, NAND, NOR, XOR, and XNOR versions of a quad two-input logic chip all share this generic layout.*

In the 4000 family, a 14-pin triple three-input logic chip contains three three-input gates. The AND, OR, NAND, and NOR versions all have an internal layout shown in Figure 10-27. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The three connections leading to a gate are its inputs, while a single connection from a gate is its output.

Unlike the 74xx family, the triple three-input OR chip in the 4000 family has the same pinouts as the other types of triple three-input logic chips.

In the 4000 family, a 14-pin dual four-input logic chip contains two four-input gates. The AND, NAND, OR, and NOR versions all have an internal layout shown in Figure 10-28. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. Each pair of connections leading to a gate are its inputs, while the single connection from a gate is its output.

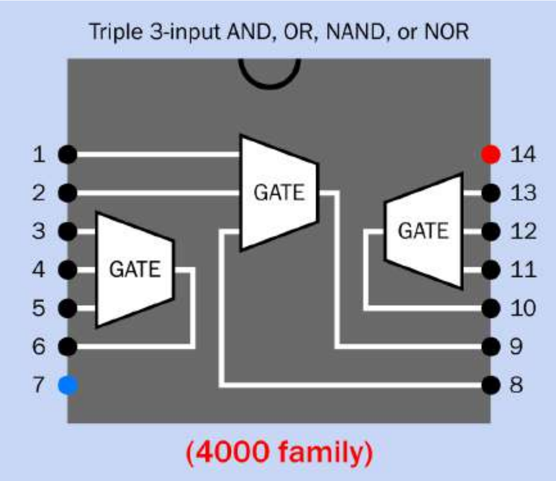

Figure 10-27. *In the 4000 family, the AND, OR, NAND, and NOR versions of a triple three-input logic chip all share this generic layout.*

Note that the 4000 family does have a dual four-input OR chip, whereas the 74xx family does not.

In the 4000 family, a 14-pin eight-input logic chip with AND and NAND outputs is available, as shown in Figure 10-29.

The following list shows the generic part numbers for 14-pin logic chips in the 4000 family that have two or more inputs per gate (in actual part numbers, letters will be substituted where an x appears):

- Quad 2-input AND: x4081x
- Quad 2-input NAND: x4011x
- Quad 2-input OR: x4071x
- Quad 2-input NOR: x4001x
- Quad 2-input XOR: x4070x
- Quad 2-input XNOR: x4077x
- Triple 3-input AND: x4073x
- Triple 3-input NAND: x4023x
- Triple 3-input OR: x4075x
- Triple 3-input NOR: x4025x
- Dual 4-input AND: x4082x
- Dual 4-input NAND: x4012x
- Dual 4-input OR: x4072x
- Dual 4-input NOR: x4002x
- Single 8-input AND/NAND: x4068x

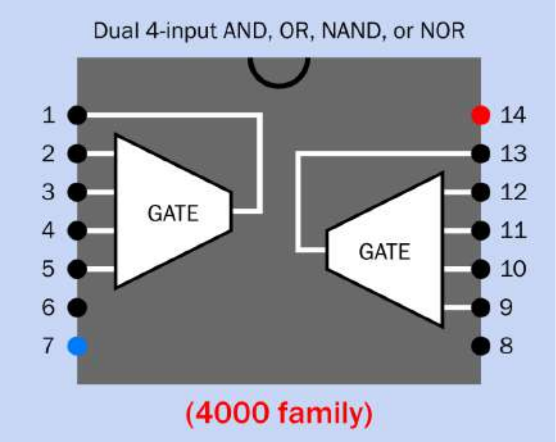

Figure 10-28. *In the 4000 family, the AND, OR, NAND, and NOR versions of a dual four-input logic chip all share this generic layout.*

4000 Series Inverters

In the 4000 family, the 4069B is a 14-pin hex inverter chip (containing six inverters), as shown in Figure 10-30. This has the same pinouts as the x74x04x chip.

How to Use It

Which Family

In DIP format, the HC family has existed for more than 30 years, and has become established as a widely used default choice.

In surface-mount formats, the choice of family will largely be determined by the choice of supply voltage.

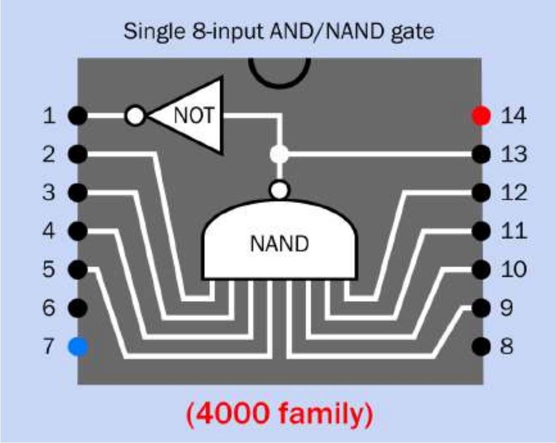

Figure 10-29. *In the 4000 family, a single eight-input AND/NAND chip has this internal layout. The inverted output from the NAND gate becomes an AND output from pin 1 of the chip.*

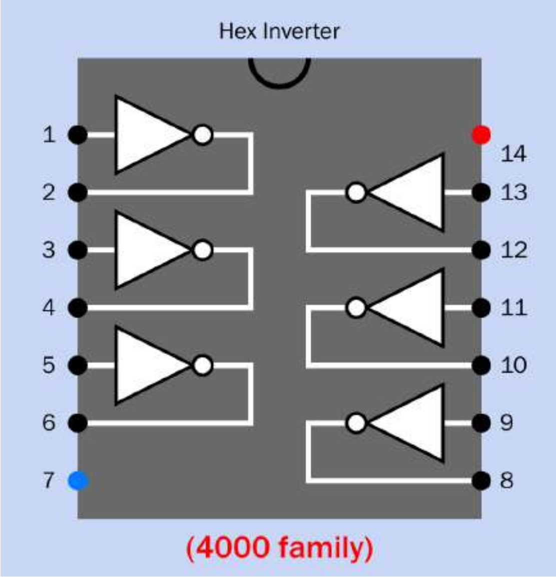

Figure 10-30. *The internal layout of a 14-pin 4069B hex inverter logic chip, containing six inverters. This layout is the same as for the x74x04x chip.*

Although the 4000 series is now more than 40 years old, it may still be useful where a 5VDC power supply is not required for other reasons in a circuit and would be added purely to power a 74xx series logic gate or other digital chip. If a circuit contains a 9VDC or 12VDC relay, for instance, a Darlington pair may be used with that voltage to drive the relay, and an old-school 4000 series logic chip could share the same supply. The relay coil would need a clamping diode to protect the logic chip from transients.

Applications

The output from a logic chip may be used as an input for a microcontroller, to enable multiple inputs to share one pin. An eight-input NAND gate, for instance, could combine the inputs from eight normally on motion sensors. If just one sensor responds to an intrusion, the gate output would change from high to low.

Logic gates may be useful in any simple device that has to respond to a single, specific combination of inputs. A digital combination lock is one example; games of chance are another. Most simple dice simulations use logic gates to convert the output from a counter to drive a dice-pattern of LEDs.

A logic gate may be used as an interface between an electromechanical **switch** and a circuit containing digital chips. A 10K pullup or pulldown resistor prevents the gate input from floating when the switch is open. A buffer can be used for this purpose, or an inverter, or any “spare” gate on a logic chip that is already in the circuit. One input of the chip can be tied to the positive power supply or negative ground, to create an appropriate input from the chip when the switch, attached to the other input, is opened or closed.

A jam-type **flip-flop** can be used to *debounce* a switch input. See Chapter 11 for details. If two NOR or two NAND gates are unused in a circuit, they can form a flip-flop.

In the original CMOS 4000 family, a positive output may be capable of driving an LED if the current does not exceed 5mA with a power supply of 5VDC or 10mA with a power supply of 10VDC. Note that the output voltage will be pulled down significantly by these loads. In the 74HCxx family, chips can source or sink as much as 20mA, but here again the output voltage will be pulled down. Note that the total limit for all outputs from a 74HCxx chip is around 70mA.

The output from a logic chip can be passed through a buffer such as the 7407, which has an open-collector output capable of sinking as much as 200mA. This enables direct drive of modest loads, so long as they are not inductive.

Solid-state relays and **optocouplers** can be driven directly from logic chips, as they draw very little current. A solid-state relay can switch 50A or more.

What Can Go Wrong

Two problems are common when using CMOS digital chips: damage from *static electricity*, and erratic behavior caused by *floating pins*.

Static

The early 4000 series CMOS chips were especially vulnerable, but more recent CMOS designs generally include diodes at the inputs, which reduces the risk. Still, logic chips should be protected by inserting them into anti-static foam or enclosing them in conductive wrappers until they are installed in a board. While handling chips, it is good practice to be grounded, ideally using a wrist-mounted ground wire.

Floating Pins

Any pin which is unconnected in a logic chip is considered to be “floating,” and can pick up signals by capacitive coupling, possibly disrupting the behavior of the chip and also causing power consumption, as the ambiguous pin state will tend to prevent that gate in the chip from entering quiescent mode.

Generally speaking, input pins in a TTL logic chip that are not being used for any purpose should be tied to the positive voltage supply, while unused CMOS pins should be tied to negative ground.

Family Incompatibilities

As previously noted, older TTL logic chips may deliver a “high” output voltage that is lower than the minimum expected by newer CMOS logic chips. The best option is not to mix families, but if chips are stored carelessly, some intermingling can occur. Part numbers should be checked if one chip appears to be ignoring output from another.

Overloaded Outputs

If a circuit calls for a logic chip with an open-collector output, and a regular chip is used by mistake, it will almost certainly be damaged.

Output Pulled Down

If the output from one logic chip is connected with the input of another logic chip, and if the output from the first chip is also connected to an **LED**, the LED may pull down the output voltage so that the second chip will not recognize it as a high state. As a general rule, a logic output can drive an LED, or can drive another logic chip, but not both. Very-low-current LEDs, which draw as little as 2mA, may be acceptable.

Incorrect Polarity and Voltages

Logic chips can be knocked out by applying incorrect polarity, or voltage to the wrong pin, or the wrong voltage. Modern logic chips tolerate a very limited voltage range, and a 74xx series chip will be irrevocably damaged if it is used where a 4000 series chip was specified for a power supply higher than 6VDC.

If a chip is inserted upside-down, it will probably be damaged when voltage is applied.

Bent Pins

Like all through-hole chips, DIP logic chips can be inserted accidentally with one or more pins bent underneath the chip. This error is very easy to miss. The bent pins will not make contact with any socket that is used, and the chip will behave unpredictably. Check for proper pin insertion with a magnifying glass if necessary.

Unclean Input

Logic chips expect a clean input without voltage spikes. A 555 timer of TTL type generates spikes in its output which can be misinterpreted as multiple pulses by the input of a logic chip. A CMOS-type 555 timer is more suitable for connection with logic chips.

If a pushbutton, rotational encoder, or electromechanical switch provides a high or low input, the input must be debounced. In hardware, this is traditionally done with a **flip-flop**. It can also be done with code in a microcontroller.

Analog Input

The input of a logic chip can be connected directly with a thermistor, phototransistor, or similar analog component, but only if there is some certainty that the voltage at the input pin will remain within the range that is acceptable to the chip. In the case of a phototransistor, for example, it should be exposed to a limited, known range of light intensity.

In general, it is best to avoid applying intermediate-voltage signals to a digital logic input, as they can create unpredictable output, or output of an intermediate voltage. A **comparator** can be placed between the analog source and the digital logic chip, or a logic chip with a Schmitt-trigger input can be used.

flip-flop

11

The term **flip-flop** is sometimes printed with a space instead of a hyphen, but the hyphenated form seems to predominate in the United States. Therefore, the hyphen is included here. The term *flipflop* (with neither a hyphen nor a space) is sometimes seen, but is unusual. The acronym *FF* is confined mostly to logic diagrams or schematics.

The term *latch* is sometimes used interchangeably with *flip-flop* but is assumed here to describe a minimal asynchronous circuit that responds immediately and transparently to an input. A flip-flop can function as a latch and also as a synchronous device which is opaque, meaning that the input does not flow directly through to the output.

OTHER RELATED COMPONENTS

- **counter** (see Chapter 13)
- **shift register** (see Chapter 12)

What It Does

Transistors enable logic gates; logic gates enable flip-flops; and flip-flops enable many mathematical, storage, and retrieval functions in digital computing. Most flip-flops today are embedded in much larger integrated circuits that have complex functions. However, they are still available as separate components in chip form, and will be discussed on that basis here.

A flip-flop is the smallest possible unit of memory. It can store a single bit of data, represented by either a high or low *logic state*. (A full explanation of logic states is included in the **logic gate** entry. See Chapter 10.) Flip-flops are especially useful in **counters**, **shift registers**, and *serial-to-parallel converters*.

A flip-flop circuit can be classified as a form of *bistable multivibrator*, as each of its outputs is stable in one of two states until an external trigger stimulates it to "flip" from one state and "flop" into the other. (For a comprehensive discussion of monostable and astable multivibrators, see the **timer** entry in Chapter 9.)

An *asynchronous* flip-flop will respond immediately to a change of input, and can be used for applications such as debouncing the signal from an electromechanical **switch** or building a *ripple* **counter**. More often, a flip-flop is *synchronous*, meaning that a change in input state will be unrecognized until it is enabled by a low-to-high or high-to-low transition in a stream of pulses from an external *clock*.

How It Works

Every flip-flop has two outputs, each of which may have a high or low state. When the flip-flop is functioning normally, the outputs will be in opposite logical states, one being high while the other is low. These outputs are typically identified as Q and NOT-Q (the latter term meaning a letter Q with bar printed above it, sometimes referred to verbally as "Q-bar"). In datasheets and other documents where a bar symbol cannot be represented easily above a letter, the NOT-Q output may be represented as letter Q with an apostrophe after it, as in Q'.

Almost always, in a schematic diagram, a flip-flop is represented by a simple rectangle, with inputs and outputs identified by letters and other marks. Because a description of the inner workings is necessary before the different types of flip-flop can be understood, schematic symbols for various flip-flops will not be introduced until "Variants" on page 116.

The simplest flip-flop contains two logic gates whose function can be most easily understood if the inputs are controlled by a SPDT switch. It can be created from two NAND gates or two NOR gates, as described next. This type of component can be described as:

- *asynchronous*: Will accept data on an impromptu basis, as it is not synchronized with a clock.
- *jam-type*: Colloquial equivalent of *asynchronous*. The input is jammed in at any time, forcing an immediate change of output.
- *transparent*: The input state flows straight through to the output.

NAND-Based SR Flip-Flop

Figure 11-1 shows two NAND gates attached to a SPDT switch, with two pullup resistors. When either of the NANDs has a floating input from the switch, the pullup resistor attached to that input will maintain it in a high state. The data inputs for the NAND gates are labeled S and R, meaning Set and Reset, giving this component its name as an *SR* flip-flop:

- In a NAND-based SR flip-flop, a low state is considered an active logic input, as indicated by the bar placed above each letter.
- A high state is considered an inactive logic output.

The schematic style in this figure, with diagonally crossing conductors, is universally used and easily recognizable. The equivalent schematic in Figure 11-2, which might be created by circuit-drawing software, has the same functionality but would not be immediately recognizable as a flip-flop. The "classic" crossed-conductor representation is preferable.

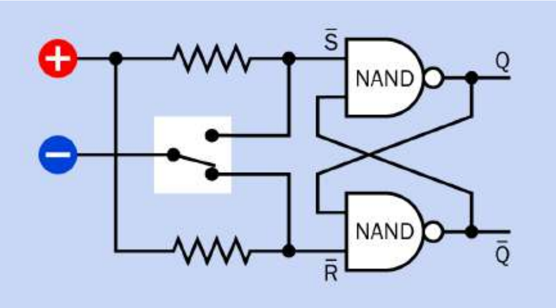

Figure 11-1. *The schematic for a simple NAND-based SR-type flip-flop, with a switch and pullup resistors added to control two inputs.*

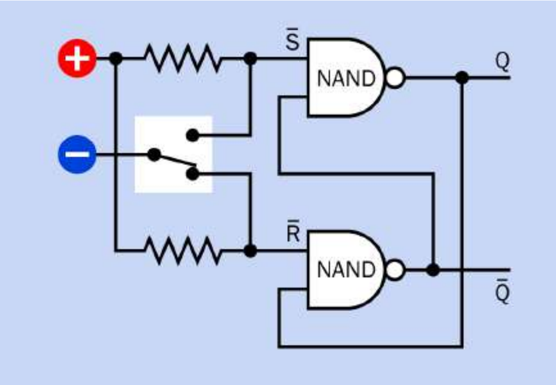

Figure 11-2. *An alternative component layout for an SR flip-flop, functionally identical to the previous schematic, but not so easily recognizable. The layout with a pair of diagonally crossing conductors has become so ubiquitous, it should be considered to be a standard.*

The first step toward understanding the behavior of flip-flops is to recall the relationship between the two inputs and the output of NAND or NOR gates. This is shown in Figure 11-3, where red indicates a high logic state and black indicates a low logic state.

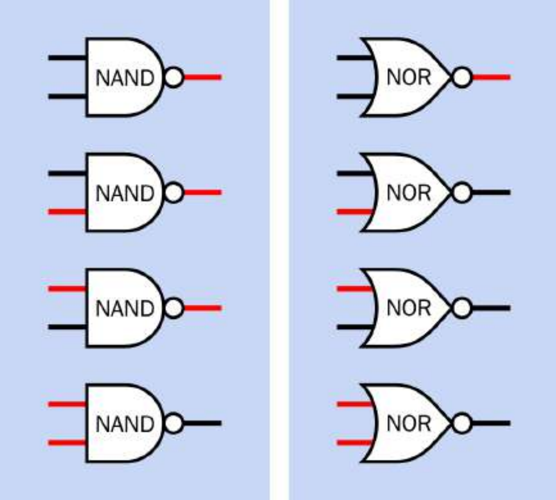

Figure 11-3. *The four possible input combinations for a NAND gate and a NOR gate, with the corresponding logical outputs. A flip-flop can be built from two NANDs or two NORs.*

The behavior of a NAND gate can be summarized:

- Both inputs high: Output low.
- Other input combinations: Output high.

Figure 11-4 shows a series of snapshots of the SR flip-flop circuit as the switch moves from one position, through an intermediate state where it makes no connection, to the other position. Remember that in this circuit, the active logical input state is low, and the active logical output state is high.

In the top panel, the pullup resistor of the lower NAND is overwhelmed by the direct connection to negative ground, which holds the R input in a low state. The other input of this gate is irrelevant, because the output from a NAND will be high if either of its inputs is low. So, the lower NAND has a high output, which feeds back to the secondary input of the upper NAND gate. The S input of this gate is high, because of the pullup resistor. Because both inputs of this gate are now high, its output is low, which feeds back to the lower gate. The lower gate doesn't change its output, because either of its low inputs is enough to keep its output high. So, the circuit is in equilibrium. A high state on the NOT-Q output is known as the *Reset* state for a NAND-based flip-flop.

The second panel shows what happens if the switch now moves up into a neutral, disconnected position. The R input of the lower NAND now becomes high, because of the pullup resistor. But this NAND still has one low input, supplied by the output of the upper gate, so its output remains high, and the circuit is still in equilibrium. This is known as the *Hold* state for the NAND circuit.

Suppose the switch bounces to and fro between the states shown in the first two panels. The output from the circuit won't change. This shows that the circuit provides a method for eliminating *switch bounce*—the very fast, momentary spikes that occur when the mechanical contacts of a switch open and close.

The third panel shows what happens if the switch now moves to its upper position. The top input of the upper NAND gate is now pulled low. Consequently, its output goes high. This feeds back to the lower gate. Its other input is high because of the pullup resistor. With both of its inputs high, its output goes low. The gate outputs have flipped and swapped values. A high state on the Q output is known as the *Set* state for a NAND-based flip-flop.

The circuit still remains in equilibrium even if the switch returns to its central, disconnected position shown in the bottom panel. Therefore, the debouncing capability of the circuit works equally well for both positions of the switch.

NOR-Based SR Flip-Flop

Figure 11-5 shows a similar circuit using two NOR gates attached to an SPDT switch. Because the NOR gates function differently, this circuit uses active-high input logic, and pulldown resistors are needed instead of pullup resistors. The output from the circuit still uses active-high logic, and is identical with the NAND-based circuit in this respect, although the relative positions of the Q and NOT-Q outputs have been swapped.

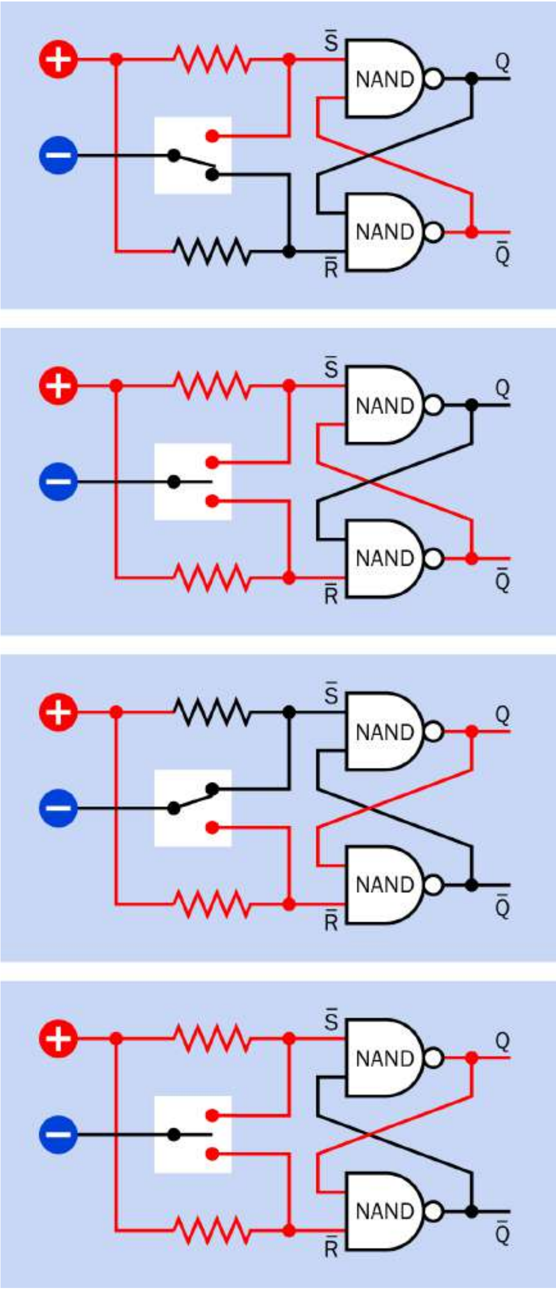

Figure 11-4. *Four snapshots of transitions in a NAND-based SR flip-flop as the switch moves down and up through an intermediate no-connection zone. See text for details.*

- In a NOR-based SR flip-flop, a high state is considered an active logic input, as indicated by the absence of a bar placed above the letters S and R.
- A high state is considered an active logic output.

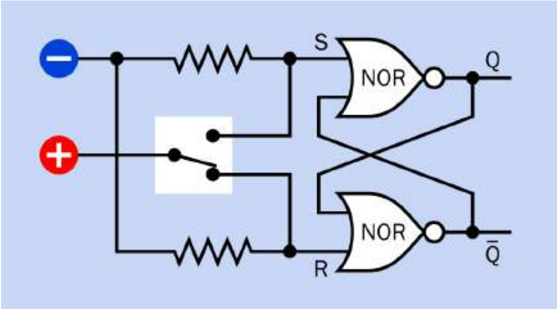

Figure 11-5. *The schematic for a simple SR flip-flop using NOR gates instead of NANDs.*

In the case of the NOR gate:

- Both inputs low: Output high.
- Other input combinations: Output low.

Figure 11-6 shows a series of snapshots as the switch moves from one position to the other, through intermediate states where it makes no connection. Remember that in this circuit, the active logical state is high at both the inputs and the outputs.

In this circuit, as in the previous circuit using NAND gates, it will ignore switch bounce, allowing the gate outputs to remain stable.

Forbidden States

Either of the circuits described so far depicts an SR flip-flop, regardless of whether it is NAND-based or NOR-based. Its input and output states are summarized in Figure 11-7. However, as this table suggests, there are some input states that create problems.

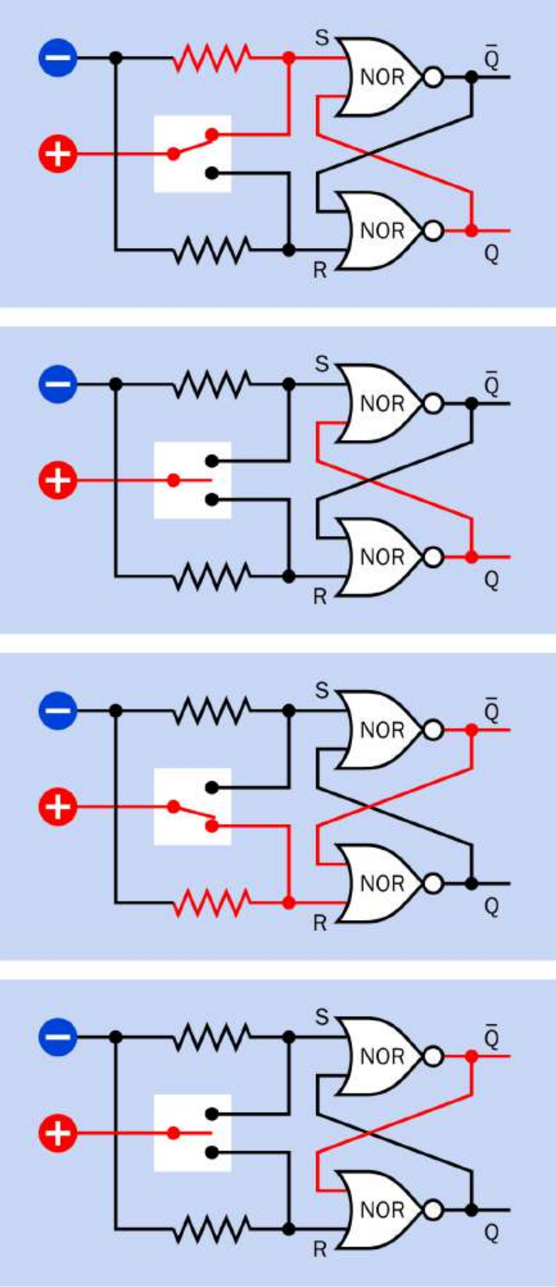

Figure 11-6. *Four snapshots of a NOR-based SR flip-flop, showing the consequence of changing switch positions, comparable with the NAND-based flip-flop.*

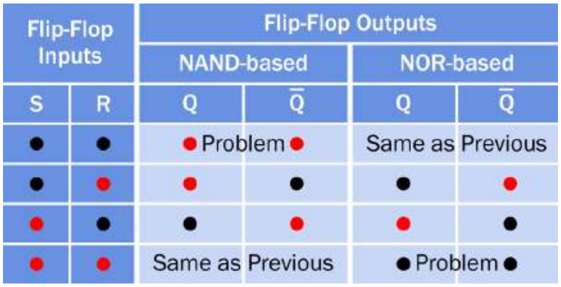

Flip-Flop Inputs		Flip-Flop Outputs			
		NAND-based		NOR-based	
S	R	Q	$\bar{Q}$	Q	$\bar{Q}$
●	●	🔴 Problem 🔴		Same as Previous	
●	🔴	🔴	●	●	🔴
🔴	●	●	🔴	🔴	●
🔴	🔴	Same as Previous		● Problem ●	

Figure 11-7. *A table of input states and the consequent output states for NAND-based and NOR-based SR flip-flops.*

In either the NAND-based flip-flop or the NOR-based flip-flop, the output when the switch is in its unconnected center position will remain the same as when the switch was in its previous position. This is the usefulness of the flip-flop: it remembers the previous state. These situations are identified as “Same as Previous” in the table.

The pullup resistors (in a NAND-based flip-flop) and pulldown resistors (in a NOR-based flip-flop) are intended to guarantee that both inputs will be high (NAND) or both inputs will be low (NOR) even when the switch makes no connection. Therefore, it should be impossible for both inputs to be low (NAND) or high (NOR).

But what happens if the circuit is powered up with the switch in the unconnected position? One input of each gate is controlled by the output of the other gate. But what will those outputs be?

In the NAND-based version, the outputs from the NANDs will be low while the chips are powering up. As soon as the NAND chips are functioning, each of them will sense that it has one input high and one input low, so it will change its output to high.

But now that each chip has a high output, it will feed back to the secondary input of the other chip. Now both chips have both inputs high. This will cause them both to change their outputs to low—but this will feed back again, flipping the outputs back to high again. In fact, if the gates are absolutely identical, the circuit will oscillate very rapidly. This is sometimes known as *ringing*.

In real life, the gates will not be absolutely identical, and eventually one of them will respond fractionally ahead of the other, tipping the circuit into the state shown either in the second panel or the fourth panel in Figure 11-4. But which chip will win? There is no way of knowing. This is known as a *race condition*, and the winner is unpredictable.

A similar but opposite situation occurs in the NOR-based flip-flop if it is powered up with the switch in the disconnected position, and the S and R outputs are both low, because of the pull-down resistors. Here again it will be a race condition.

We can address the problem by making a rule that the switch must always be in one position or the other when the flip-flop is powered up. But what if there is a faulty switch? Or what if a power interruption occurs while the switch is changing position?

Another problem occurs if the switch makes one contact a fraction before it breaks the other contact. This would result in both S and R inputs being low, in a NAND flip-flop. The same state could occur if a separate logic circuit is driving the S and R inputs, and an error causes it to make S and R both low. This is shown in Figure 11-8. Because the output from a NAND gate is always high if at least one of its inputs is low, both gates now have a high output, and the circuit is stable.

The problem is, the states of the outputs from a flip-flop should always be opposite to each other. If both of them are high, this can create logic problems in the rest of the circuit attached to the flip-flop.

- In a NAND-based SR flip-flop, if S and R are both low, this is known as a *forbidden state* or a *restricted combination*.

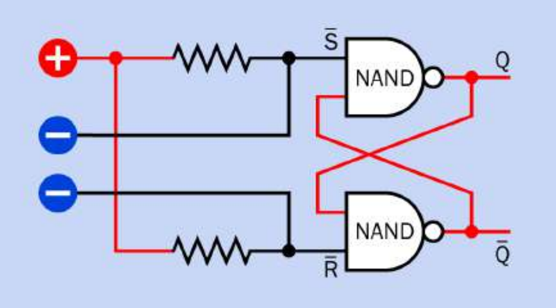

Figure 11-8. *What happens when both S and R inputs to a NAND flip-flop are low as a result of an error in a separate control circuit.*

A similar problem afflicts a NOR-based SR flip-flop, except that the forbidden state will occur when the S and R inputs are both high.

- In a NOR-based SR flip-flop, if S and R are both high, this is a *forbidden state* or a *restricted combination*.

The SR flip-flop is useful as a switch debouncer, but for computing applications, it is vulnerable to errors.

The JK Flip-Flop

Because the name of the JK flip-flop shares the initials of Jack Kilby, who won a Nobel prize for his fabrication of the world's first integrated circuit, some people speculate that this type of flip-flop was named after him. The attribution seems implausible, and may have gained currency simply because a flip-flop was the first device that Kilby happened to build when he was developing an integrated circuit.

Regardless of how it came to be named, the JK design is shown in Figure 11-9. This is commonly referred to as a JK *latch*. The electromechanical switches that were shown driving the SR flip-flop, along with pullup or pulldown resistors, are no longer included, because the inputs at positions J and K are assumed to come from other devices that have properly defined high and low states. Their behavior may be unpredictable, but neither of them will ever have a floating state.

This is a *gated* circuit, meaning that an additional input stage blocks direct access to the output stage, and it is also a *synchronous* circuit, as it uses a train of pulses at a clock input. Two three-input NAND gates are placed in front of a NAND-based SR flip-flop, and they address the problem of simultaneous identical inputs by using crossover feedback from the second stage to the first stage (via the conductors at top and bottom of the schematic).

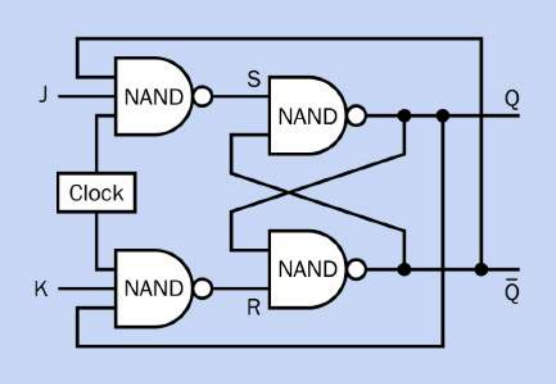

Figure 11-9. *The basic circuit for a clocked JK flip-flop, using two additional NAND gates prepended to an SR flip-flop.*

Versions of a JK flip-flop are possible using NOR gates, but are less common. Only the NAND-based version will be considered here.

In the case of a three-input NAND gate:

- All three inputs high: Output low.
- Other input combinations: Output high.

Because of the additional pair of NANDs, the circuit now recognizes a high input as logically active, instead of the low-active input in the previous SR flip-flop using NAND gates. Consequently, two simultaneous high inputs might be expected to create the type of forbidden state that was caused by two simultaneous low inputs previously. However, in Figure 11-10, the top and bottom panels show that simultaneous high inputs at J and K will support two possible valid outputs, where the state at Q is always opposite to the state at NOT-Q. In fact, when both inputs are high, a positive pulse on the clock input will *toggle* the outputs (i.e., they will switch places). In fact, the toggling will continue so long as the clock input is high. Consequently, this type of flip-flop is intended for use with short clock pulses.

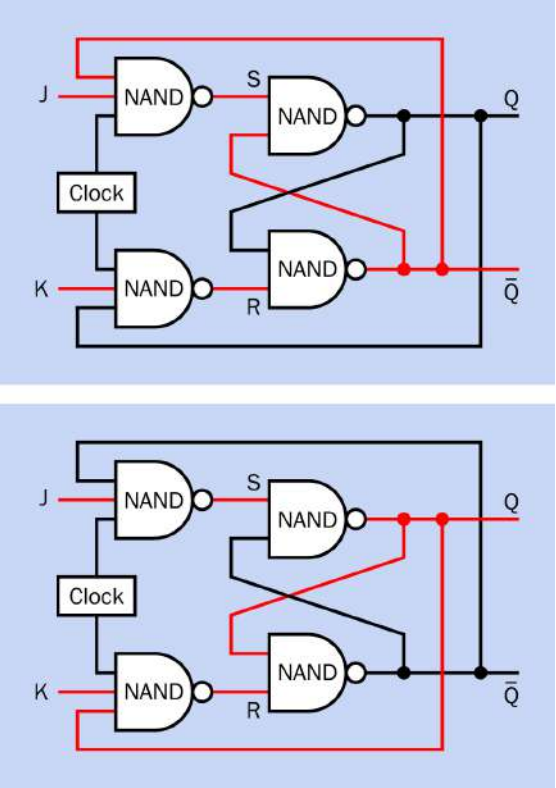

Figure 11-10. *When the J input and the K input are both high, this no longer causes a forbidden state. The combination will toggle the outputs of the flip-flop between the two states shown here.*

Master-Slave Flip-Flop

A more stable form is shown in Figure 11-11 where yet another stage has been added, this one being a "master" to the first. In fact, this configuration is known as a *master-slave* flip-flop, the slave stage being driven by the master stage but remaining inactive until a low clock input at the master stage passes through an *inverter* to become a high clock input at the slave stage. The master and slave stages thus take turns, one being activated by a high clock pulse while the other is activated by the low part of the pulse cycle. The output from the slave stage cannot feed back to the master stage while the clock pulse is still high, and thus the timing issue in the single-stage JK latch is eliminated. Because the master-slave version of the JK configuration is not *transparent*, it is correctly known as a flip-flop rather than a latch.

In addition, Preset and Clear inputs may be added to override the clock to Set or Reset the outputs. These inputs are active-low.

Figure 11-12 summarizes the behavior of a JK master-slave flip-flop that is triggered by the falling edge of each clock pulse (shown as a downward-pointing arrow in the Clock column of the table). Note that the output will be delayed while the slave stage waits for the second part of each clock cycle.

The letter X in the table indicates that the state in that cell is irrelevant.

When J and K are both low, the states of Q and NOT-Q will remain the same as in the previous cycle, and this is still referred to as a *Hold* state. When J and K are both high, the outputs toggle, meaning that their new states will be opposite to the previous states.

D-Type Flip-Flops

A D-type flip-flop places an inverter between two inputs to guarantee that they will always be in opposite states, and uses a clock signal to copy their states to a pair of logic gates.

When an inverter is added between the inputs in this way, either an SR flip-flop or a JK flip-flop can become a D-type flip-flop. Figure 11-13 shows the simplest possible D-type circuit, added to a basic SR flip-flop. Only one data input is now required (customarily labeled D), because it drives the other through the inverter.

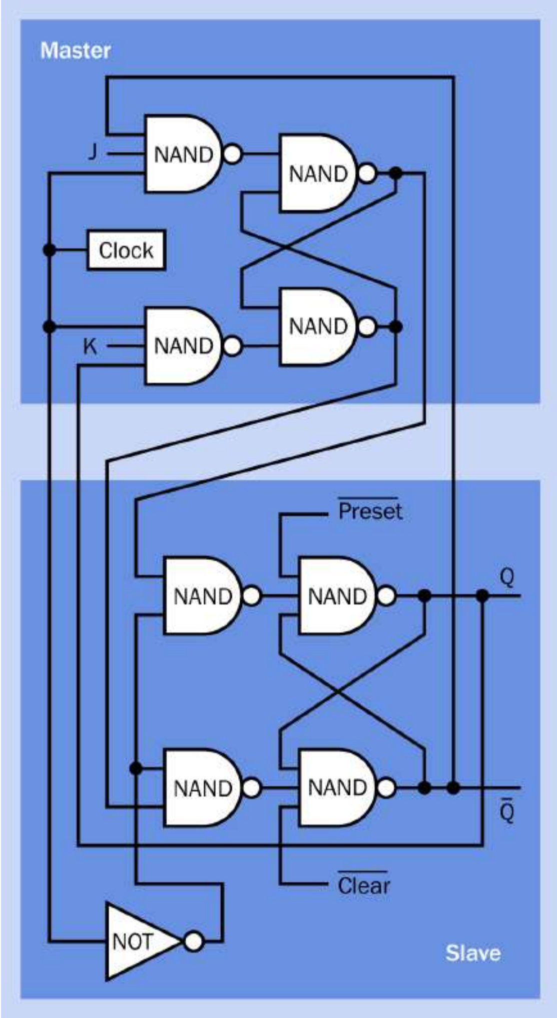

Figure 11-11. *A master-slave circuit that drives one flip-flop with another.*

Flip-Flop Inputs					Outputs	
Clock	Preset	Clear	J	K	Q	$\overline{Q}$
X	●	🔴	X	X	🔴	●
X	🔴	●	X	X	●	🔴
↓	🔴	🔴	●	●	Same as Prev	
↓	🔴	🔴	●	🔴	●	🔴
↓	🔴	🔴	🔴	●	🔴	●
↓	🔴	🔴	🔴	🔴	Toggle	

Figure 11-12. *A table showing inputs and outputs for a JK master-slave flip-flop.*

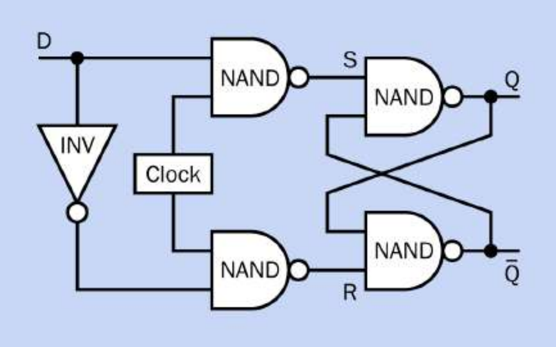

Figure 11-13. *A simple D-type flip-flop. The inverter guarantees that the state of one input will always be opposite to the state of the other.*

Figure 11-14 uses snapshots to show how the circuit responds to changing input and clock conditions. Initially, in the top panel, the data input is high, the clock input is high, and the Q output is high. In the second panel, the clock goes low, causing the output from the upper NAND gate in the input stage to change from low to high. But the upper NAND gate in the output stage still has one low input, so its state remains unchanged. In fact, the S and R inputs of the output NANDs are now both high, which creates the hold condition.

In the third panel, the D input changes from high to low, but this has no effect so long as the clock is low. The D input can fluctuate repeatedly, and nothing will happen until the clock goes high, as shown in the fourth panel. Now the clock copies the new D input state through to the output.

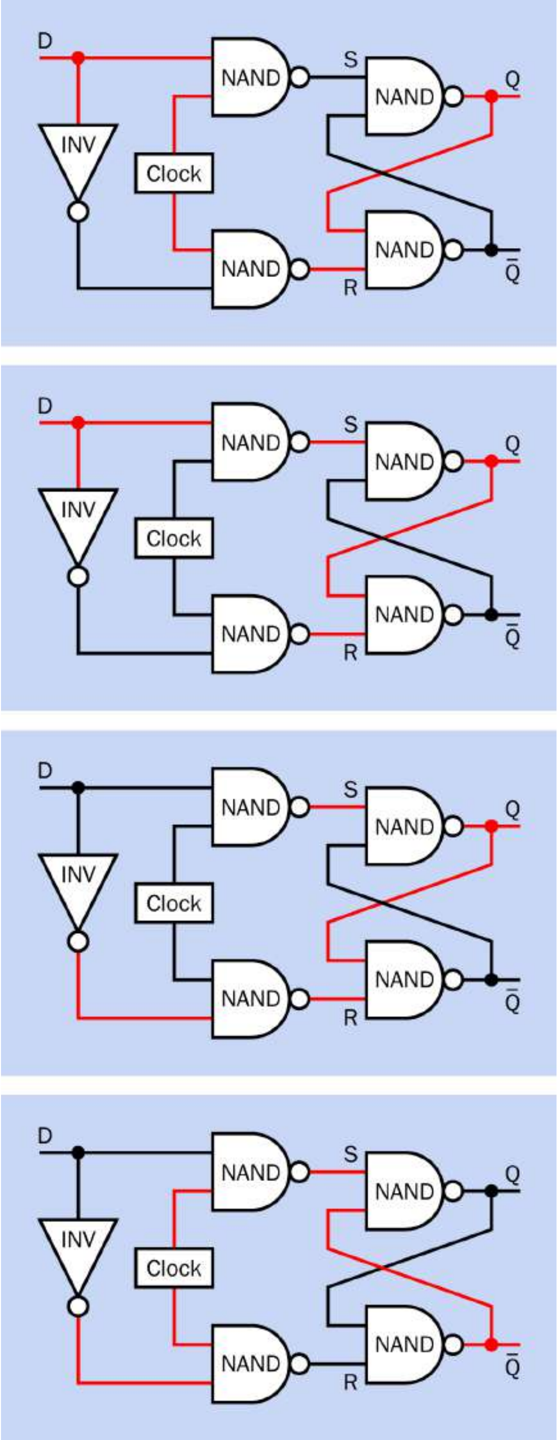

Figure 11-14. *Four snapshots showing the behavior of a D-type flip-flop.*

Summary

- An SR flip-flop can be used for switch debouncing, but in other applications it can enter an unacceptable race condition if its inputs and power supply are not carefully controlled.
- A JK flip-flop is gated, meaning that an SR circuit is preceded with an input stage and a clock input. This eliminates the race condition, adds the ability to toggle the outputs, but requires a very brief clock input. The circuit is edge-triggered.
- A master-slave flip-flop consists of two flip-flops, one driving the other. They can be JK type or SR type. The first flip-flop is activated by a positive clock state, while the second is activated by the subsequent negative clock state. Timing issues are resolved.
- A D-type flip-flop is gated with an inverter between the inputs, so that they cannot be simultaneously high or low. Consequently, only one input, labeled D, is needed. A high state on the D input causes a Set condition, while a low input causes a Reset condition, but only when the clock copies the status of the inputs through to the outputs. The status of the outputs remains stable (the flip-flop enters a hold condition) after the clock goes low.
- The JK circuit used to be widely used, because of its versatility. The D-type circuit now predominates.
- A T-type (toggling) flip-flop exists but is uncommon, and is not included in this encyclopedia.
- Many flip-flop circuits exist in addition to the ones that have been illustrated here. Only the most commonly cited circuits have been included.

A chip containing two positive-edge triggered D-type flip-flops is shown in Figure 11-15. Each flip-flop in this component has its own data, set, and reset input and complementary outputs.

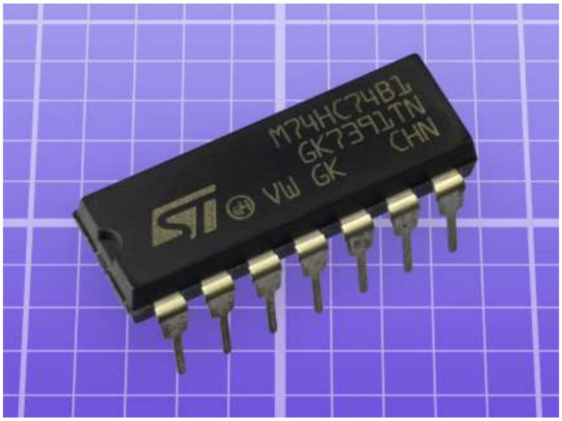

Figure 11-15. *This chip contains two positive-edge triggered D-type flip-flops.*

Variants

A selection of schematic symbols representing flip-flops is shown in Figure 11-16. Letters S, R, J, K, or D define the type of flip-flop. Q and NOT-Q are the outputs. CLK is the clock input but may alternatively be identified with letter E, meaning *Enable*. SRCK or SCLK may also identify it, the abbreviations being intended to mean “serial clock.

A triangle preceding CLK indicates that the flip-flop is positive-edge triggered. A circle, properly termed a *bubble*, preceding the triangle, indicates that the flip-flop is negative-edge triggered. In other locations, the bubble indicates that the input (or output) is inverted; it means the same thing as a bar printed above the text abbreviation, and indicates active-low logic. Synchronous inputs are shown on the left side of the flip-flop with the CLK input, while asynchronous inputs (if any) are shown above and below the flip-flop rectangle.

Using these guidelines, the examples in Figure 11-16 can be decoded:

1. An unclocked SR flip-flop with active-low inputs (probably NAND-based).
2. An unclocked SR flip-flop with active-high inputs (probably NOR-based).

3. An SR flip-flop with active-high inputs, pulse-triggered by an active-high clock input.
4. A JK flip-flop with active-high inputs, edge-triggered by a rising-edge clock input. The bubble on the lower Q output means the same thing as a letter Q with a bar printed above it.
5. A D-type flip-flop pulse-triggered by an active-low clock input.
6. A D-type flip-flop edge-triggered by a falling-edge clock input.
7. A JK flip-flop with active-high inputs, edge-triggered by a rising-edge clock input, with asynchronous active-low Preset and Clear inputs.

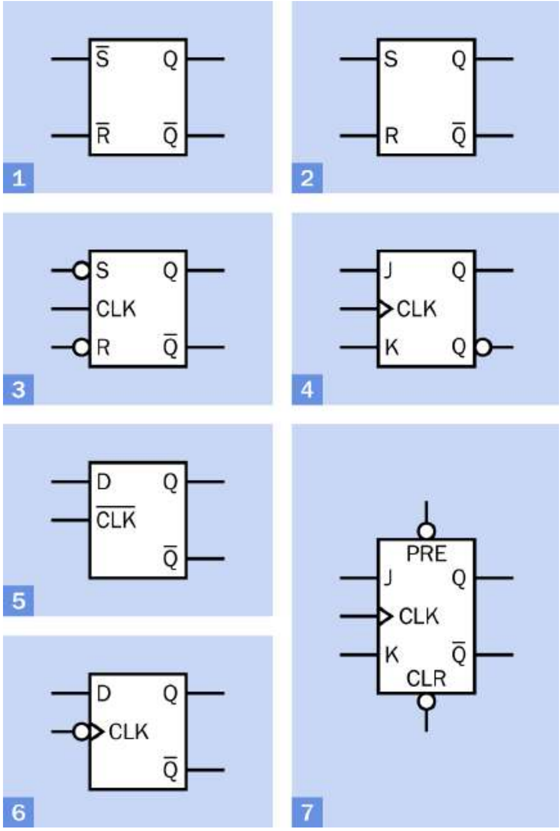

Figure 11-16. *The schematic symbol for a flip-flop is an annotated rectangle. See text for an explanation of the letters and marks.*

Packaging

Only about 10% of the flip-flops listed by a typical parts warehouse are through-hole chips. The rest are now surface-mount. Still, even if the search is narrowed further to through-hole packages in the 74xx and 4000 series, at least 100 types still exist. They provide opportunities in education and prototyping work, even though they are less often used as standalone components.

A package often contains more than one flip-flop. Dual and quad arrays are common. The flip-flops may be independently clocked, or may share a single clock input; datasheets should be checked carefully for details. Octal flip-flops, such as the D-type 74x273, are intended for use as eight-bit registers.

Many of the older flip-flops are numbered in the 74xx series of logic chips. See Chapter 10 for a detailed guide to this numbering system and the various logic families. D-type flip-flops include 74x74, 74x75, 74x174, and 74x175, where an acronym for the logic family is substituted for the x. Old-style CMOS flip-flops include the 4042B D-type latch, the 4043B quad NOR SR flip-flop, and the 4044B quad NAND SR flip-flop. The last two are synchronous, and both allow two Set inputs, labeled S1 and S2 in the datasheet.

Examples of JK flip-flops include the 74x73, 74x76, and 74x109.

Values

As is the case with other logic chips, most flip-flops in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See "Variants" on page 93 for a discussion of acceptable high and low logic input voltages. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can manage around 20mA.

If a flip-flop is used for high-speed operation, the following values must be considered:

- t_S Setup time: The minimum time in nanoseconds for an input to be constant before the next clock pulse can process it.
- t_H Hold time: The minimum time in nanoseconds for an input to persist after the active edge of a clock pulse that has processed it. The interaction between a clock pulse and an input state takes a brief but measurable amount of time; errors may occur if the clock is given less than that amount of time to do its job.
- t_{CO} Clock-to-output: The elapsed time after an active clock edge, before the output changes. This is a function of the internal workings of the chip, and may be broken down into low-to-high and high-to-low output transitions, as follows.
- T_{PLH} Propagation to Low-to-High: The elapsed time after an active clock edge, before a low-to-high swing occurs at an output. This may not be identical to T_{PHL}.
- T_{PHL} Propagation to High-to-Low: The elapsed time after an active clock edge, before a high-to-low swing occurs at an output. This may not be identical to T_{PLH}.
- f_{MAX} Maximum clock frequency for reliable operation.
- $t_{W(H)}$ The minimum high clock pulse width in nanoseconds.
- $t_{W(L)}$ The minimum low clock pulse width in nanoseconds.

In a shift register or counter, where multiple flip-flops are cascaded but they share the same clock, the t_{CO} of one flip-flop must be shorter than the hold time of the next flip-flop, to allow the input of data to be complete before the window of opportunity is over.

How to Use It

The asynchronous SR flip-flop is of primary use in debouncing switches. Examples are the single MAX6816, dual MAX6817, and octal MAX6818.

D-type flip-flops are widely incorporated in frequency dividers, which are used to count pulses and display a binary output. If the NOT-Q output is wired back to the D input, the pulse stream to the clock input will have the following effect:

1. Suppose the initial D state is low and the initial state of the NOT-Q output is low.
2. The first high clock pulse propagates the low D state into the flip-flop.
3. The next low clock state forces the NOT-Q output high. This feeds back and creates a high D input.
4. The second clock pulse propagates the high D state into the flip-flop.
5. The next low clock state pulls the NOT-Q output low. This feeds back and creates a low D input.

The sequence then repeats. Only one high output is generated at NOT-Q (or at Q) for every two clock pulses; thus the circuit can become a divide-by-two **counter**. If the Q output is tapped to serve as the clock input for another flip-flop, that circuit now has a divide-by-four output. A series of many flip-flops can be chained together, so long as the propagation of signals along the chain is fast enough to occur before the next clock pulse. This is known as an *asynchronous counter*.

For more information on the use of counters, see Chapter 13.

While flip-flops have tended to be integrated with other components in digital computing, they are still used as registers where 8 or 16 bits of serial data must be assembled at a time, prior to being disseminated as parallel data.

What Can Go Wrong

Ambiguous Documentation

For reasons which are unclear, instructional texts and tutorials can be erratic when describing flip-flops:

- A truth table may fail to clarify whether the circuit uses active-high or active-low logic.
- Truth tables from different sources are often inconsistent in their representation of current and future output states, and may even fail to include the clock status in a clocked flip-flop.
- Tutorials may include logic diagrams for some types of circuit, but not others.
- NOR gates may be used, without any mention that NAND gates can also be used (and may be more common or convenient).
- The active-low or active-high status of inputs in an SR flip-flop may not be shown.

Bearing this in mind, manufacturer datasheets should be consulted whenever possible as the primary source of information.

Faulty Triggering

In many cases, a flip-flop designed for edge triggering can give erroneous results if it is level-triggered, and vice versa. Rising-edge-triggered flip-flops must be distinguished from falling-edge-triggered flip-flops. As always, it is important for similar parts that have similar functions to be stored separately.

Metastability

The behavior of flip-flops has been described in this entry under ideal conditions, where they are operating well within parameters established by the manufacturer. In reality, non-ideal scenarios may occur, especially where inputs such as data and clock, or clock and reset, are almost simultaneous. This may be difficult to avoid if a signal is received from an external source such as a sensor, with no way to control its arrival time. If the input occurs within the setup time or the hold time of a clock pulse, the flip-flop may be unable to determine whether the input precedes or follows the clock.

This may lead to *metastability*, meaning an unpredictable output and/or oscillations that take several clock cycles to settle into a stable state. If the output from a flip-flop may be used by two separate components with slightly different response times, one may interpret the oscillating output as a high state while the other interprets it as low. In a computing circuit, metastability can lead to calculation errors or a system crash. To avoid these issues, limits in datasheets should be observed. Attention should be paid to the manufacturer specifications for minimum setup time and hold time, so that the circuit has sufficient opportunity to recognize a signal and respond.

One solution to metastability is to connect multiple flip-flops in series, all sharing a common clock signal. This will tend to filter out irregularities, at the expense of requiring additional clock cycles if the flip-flops are not transparent.

Metastable-hardened flip-flops minimize metastability but cannot eliminate it completely.

Other Issues

Problems that tend to affect digital chips generally are listed in the section of the entry on logic gates (see “What Can Go Wrong” on page 105).

shift register

12

The term **shift register** is rarely hyphenated. In this encyclopedia, no hyphen is used.

A shift register can function as a *queue*, but this term is more usually applied to software. When the output from the last stage of a shift register is connected back to its input, it can function as a *ring counter*, but that application is described in the **counter** entry of this encyclopedia.

Component catalogs sometimes list shift registers as binary ripple **counters**, instead of giving them their own section. In this encyclopedia, a binary counter is considered to have binary-weighted outputs (with values 1, 2, 4, 8…. in decimal notation) and is described in the **counter** entry. A shift register has outputs that are not necessarily binary-weighted.

OTHER RELATED COMPONENTS

- **flip-flop** (see Chapter 11)
- **counter** (see Chapter 13)
- **multiplexer** (see Chapter 16)

What It Does

A *register* is a component (or a small section of computer memory) that stores information. The smallest unit of information is one *bit* (i.e., one *binary digit*) with a value 1 or 0 that can be represented by a high or low logic state. A **shift register** most commonly is designed to store eight bits, although some store four.

Each bit is memorized by the status of a **flip-flop** inside the register. For a detailed description of flip-flops, see Chapter 11. When a pulse from an external *clock* is received by the shift register, all of the bits in storage are moved along one step, from each flip-flop to the next. The high or low status of an input pin at that moment is *clocked in* to the first flip-flop, while the bit in the last flip-flop is overwritten by the bit preceding it. A diagram representing the function of a basic four-bit shift register is shown in Figure 12-1. Note that the status of the input pin is ignored until the moment when a clock pulse copies it into the first flip-flop. In the figure, when the input pin has a brief high state that ends immediately before clock pulse three, the high state is ignored.

A shift-register chip is shown in Figure 12-2.

Because the functionality of a shift register is now often incorporated in much larger logic chips, it is less widely used as a stand-alone component than it used to be. It is still useful for purposes of serial-parallel or parallel-serial conversion, and for small tasks such as scanning a matrix-encoded keyboard or keypad. It also has educational applications and can be used in conjunction with a microcontroller.

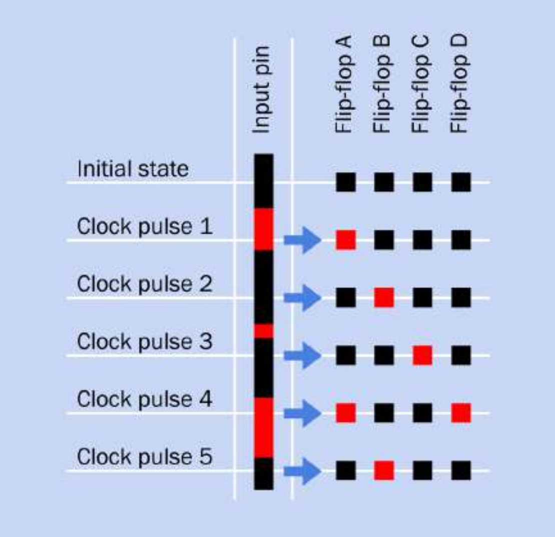
Input pin
Flip-flop A
Flip-flop B
Flip-flop C
Flip-flop D
Initial state
Clock pulse 1
Clock pulse 2
Clock pulse 3
Clock pulse 4
Clock pulse 5

Figure 12-1. *The function of a four-bit shift register in which each flip-flop may be set to a high or a low state, represented here with red or black squares. After a high bit is clocked into the chip, it is moved one space along by each subsequent clock pulse.*

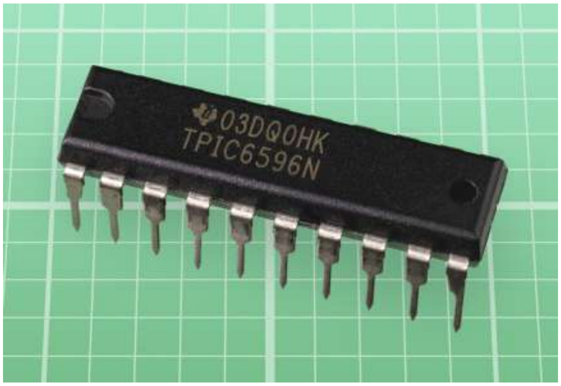
03DQ0HK
TPIC6596N

Figure 12-2. *This 8-bit shift-register chip is unusual in that it uses "power logic," in which open-drain outputs enable it to drive relatively high-current devices. It can sink up to 250mA at each of its output pins, at up to 45VDC.*

Schematic Representation

No specific symbol exists for a shift register. It is represented in a schematic by a simple rectangle, often (but not always) with control inputs on the left, data inputs arrayed along the upper edge, and data outputs along the lower edge. An example is shown in Figure 12-3, along with a diagram showing the physical chip and its pinouts. The meaning of the abbreviations identifying the inputs, outputs, and control functions will be described in "How It Works" on page 122.

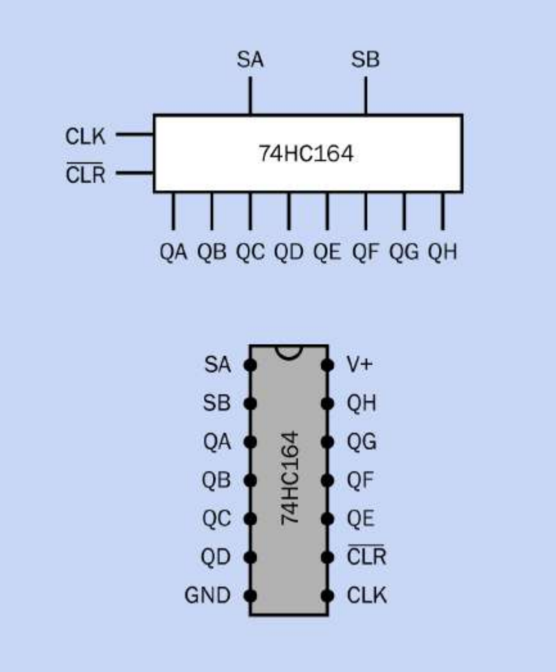
SA
SB
CLK
$\overline{\text{CLR}}$
74HC164
QA QB QC QD QE QF QG QH
SA
SB
QA
QB
QC
QD
GND
V+
QH
QG
QF
QE
$\overline{\text{CLR}}$
CLK
74HC164

Figure 12-3. *Typical schematic representation of a shift register, compared with the pinouts of the actual component.*

The schematic symbol representing a shift register may appear superficially similar to the physical form of the chip which contains it, but the physical layout of the pins is unlikely to be the same.

How It Works

A shift register generally consists of a chain of D-type flip-flops. See the entry describing **flip-flops** in Chapter 11 for a detailed explanation of this component.

The simplest shift register functions as a *serial-in, serial-out* device, abbreviated with the acronym *SISO*. Because the first bit that enters it will be the first to leave at the opposite end, it can also be described as a *first-in, first-out* data storage device, using the acronym *FIFO*.

The basic connections between flip-flops in a four-bit SISO shift-register are shown in Figure 12-4. The D input in each section refers to the fact that it is a D-type flip-flop. The primary output from each flip-flop is identified with letter Q.

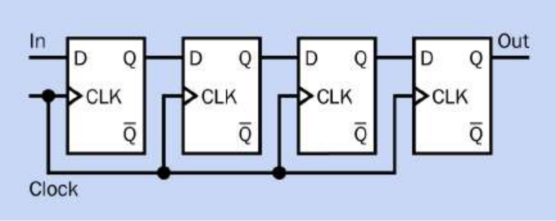

Figure 12-4. *The simplest shift register is a serial-in, serial-out (SISO) device. This example contains four D-type flip-flops.*

Each clock input is labeled CLK. When the output of each flip-flop is coupled to the input of the next flip-flop, and both share the same clock signal, the clock signal will cause the state of the third flip-flop to be sent to the fourth, the output of the second to be copied to the third, the output of the first to the second, and the input state will be copied to the first.

Abbreviations and Acronyms

The shift register will usually have an additional input that forces an immediate "clear" of all the registers, regardless of the clock state at that moment. This input is usually labeled CLR and will have a bar printed above it if it is active-low (which is the usual convention). If there is a pin labeled MR (meaning "Master Reset"), it will have the same function as CLR.

Because its effect is independent of the clock state, the clear signal is described as an *asynchronous* input.

While the abbreviation CLK is frequently used to identify the clock input, SCLK is also used (meaning "serial clock"), and occasionally the abbreviation CP may be found, meaning "clock pulse" input. If the shift register contains two stages, one to clock data in and the second to clock data out, they may be separately clocked, in which case they will be identified with different abbreviations. These are not standardized, but should be explained in the manufacturer's datasheet. No matter which abbreviation is used for a clock input, it will have a bar printed above it if the input is active-low.

Shift registers are generally *edge triggered*, meaning that the rising or falling edge of a clock pulse triggers the bit-shifting operation. If the component responds to a clock transition from low to high, it is *rising-edge triggered*. If it responds to a transition from high to low, it is *falling-edge triggered*, and this may be indicated in the schematic by a small circle, properly known as a *bubble*, preceding the triangle which indicates that this is an edge-triggered device.

Most shift registers are positive-edge triggered.

Parallel Outputs and Inputs

In many shift registers, data may be read out in parallel (from all flip-flops simultaneously), using pins provided for this purpose. In this mode, the shift register can function as a *serial-parallel converter* (*serial in, parallel out*, represented by the acronym *SIPO*). A simplified schematic of the internal connections is shown in Figure 12-5.

Where parallel outputs are provided, they are often identified as QA, QB, QC, and so on (moving from left to right) but may alternatively be described as Q1, Q2, Q3, Q4, and so on.

In a schematic, the input pin is conventionally shown as being at the left end of the component. Often two inputs are provided, connected internally as inputs to a NAND gate. The inputs are likely to be labeled A and B, but may alternatively be named SA and SB, indicating that they are serial inputs. S1 and S2 are alternative classifications. If parallel inputs exist, they may be identified as PA, PB, PC, and so on.

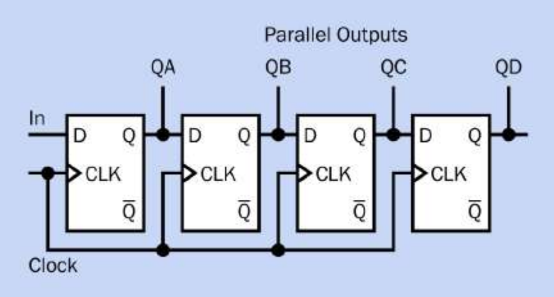

Figure 12-5. *Many shift registers have pins connected to points along the chain of flip-flops. These connections enable data to be read from the shift register in parallel.*

If serial data is supplied asynchronously, as in the illustration in Figure 12-1, it will be ignored until the shift register is triggered by the next clock pulse. The input state at that moment will then be copied into the first flip-flop, while the data that is already being stored in the shift register will be moved along the chain. In datasheets, this is customarily represented by a diagram such as the one in Figure 12-6. This diagram assumes that the shift-register is rising-edge triggered. Note that a brief fluctuation in the input which does not coincide with a clock-trigger event will be ignored.

Variants

Serial In, Serial Out

A basic SISO shift register allows only for serial input (at one end of the chain of flip-flops) and serial output (at the other end of the chain). No pins are available for parallel output of data.

This type of component usually permits 64-bit storage, where parallel output is simply impractical, as too many pins would be required. An example is the 4031B chip. This includes provision for recirculation of bits, so that it will also function as a ring counter (see Chapter 13 for a discussion of this function). As is always the case with logic chips, the part number will be preceded by letter(s) identifying the manufacturer, and a suffix will distinguish variants of the chip.

Another type of SISO shift register is programmable. It will store any number of bits from 1 through 64, determined by a binary number applied in the form of high/low states to five pins reserved for this purpose. An example is the 4557B.

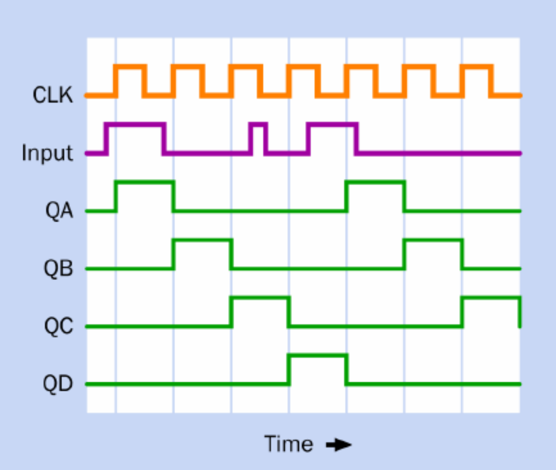

Figure 12-6. *In a rising-edge-triggered shift register, the high or low state of an asynchronous input (purple line) is copied into the first flip-flop of a shift register by each clock pulse (orange line). Brief fluctuations that do not coincide with a rising clock pulse are ignored. Existing data in the register is shifted from one flip-flop to the next.*

Serial In, Parallel Out

The majority of serial-input shift registers allow parallel output from points along the chain in addition to serial output at the end of the chain. These chips almost all are 8-bit registers. Typically two inputs are provided, one of which can be used to receive bits that recirculate from the end of the chain, back to the beginning. Widely used examples are the 4094B and the 74x164, where an acronym identifying the logic family will be substituted for the x.

Parallel In, Serial Out

A minority of shift registers are able to function as *parallel-serial converters* (*parallel in, serial out*, represented by the acronym *PISO*). Typically this type of chip allows *jam-type* parallel data input, meaning that the data is forced into the flip-flops via a separate pin for each of them. Parallel input is enabled by the status of a serial/parallel control pin. When the control pin reverts to its opposite status, each clock pulse will now shift data along the chain of flip-flops, allowing it to be read from the final output one bit at a time. Thus, data can be entered into the chip in parallel and read out of it serially. Examples are the 4014B and 4021B. Both are 8-bit.

Parallel In, Parallel Out

Shift registers that permit parallel output in addition to parallel input are almost all of the *universal* type, described in the next section.

Universal

A *universal* shift register is capable of all four modes of operation: SISO, SIPO, PISO, and PIPO. The four modes of the component are selected by the high or low status of two *mode select* pins. In addition, this component may have the ability to shift the register states either left or right. A *bidirectional* shift register has this same capability, and may also have PISO and PIPO capability, depending on the chip. Examples are the 74x195 and 74x299, where an acronym identifying the logic family will be substituted for the x in the number.

Universal shift registers are almost all 4-bit or 8-bit. They often have relatively complicated features, such as access to internal JK flip-flops, or pins that are multiplexed to provide different functionality depending whether an enable pin is held high or low. Datasheets must be checked carefully to ensure correct use.

Dedicated shift registers of SIPO or PISO type will be easier to use.

Values

As is the case with other logic chips, most flip-flops in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC. See the section on logic gates in Chapter 10 for a discussion of acceptable high and low logic-input states. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can usually manage around 20mA.

If a shift register is used for high-speed operation, the following values must be considered (identical notation, and similar values, are found in specifications for flip-flops):

- t_S Setup time: The input state of a shift register must exist for a very brief period before the clock trigger that processes it. This period is known as the *setup time*. In the 4000 series of chips, recommended setup may be as long as 120ns. The value will be much lower in 74xx chips.
- t_H Hold time: The minimum time in nanoseconds for an input to persist after the active edge of a clock pulse that has processed it. In many shift registers, no hold time is necessary, as the chip has already been activated by the rising edge of the clock pulse.
- t_{CO} Clock-to-output: The elapsed time after a clock trigger, before the output changes. This is a function of the internal workings of the chip, and may be broken down into low-to-high and high-to-low output transitions, as follows.
- T_{PLH} Propagation to Low-to-High: The elapsed time after an active clock trigger, before a low-to-high swing occurs at an output. This may not be identical to T_{PHL}.
- T_{PHL} Propagation to High-to-Low: The elapsed time after an active clock trigger, before a high-to-low swing occurs at an output. This may not be identical to T_{PLH}.
- f_{MAX} Maximum clock frequency for reliable operation. In the older design of 4000 series chips, 3MHz may be recommended with a power supply of 5VDC. Higher frequencies are possible with a higher voltage power

supply. Frequencies as high as 20MHz are possible in the 5VDC 74HC00 series.

- $t_{W(H)}$ The minimum high clock pulse width in nanoseconds. In the older design of 4000 series chips, 180ns may be recommended with a power supply of 5VDC. Shorter pulses are possible with higher voltage power supply. Pulses as short as 20ns are possible in the 5VDC 74HC00 series.
- $t_{W(L)}$ The minimum low clock pulse width in nanoseconds. This is likely to be the same as $t_{W(H)}$.

Power Considerations

Shift registers conform with the usual power-supply requirements for logic families. These are described in detail in the **logic gate** entry in Chapter 10. Likewise, the ability of a shift register to source or sink current is usually determined by its logic family. In a few cases, however, shift registers have an additional *open-drain* output stage for each register, capable of sinking currents as high as 250mA. The Texas Instruments TPIC6596 shown in Figure 12-2 is an example. When an open-drain output is connected with a logic device whose input cannot be allowed to float indeterminately, a *pullup resistor* must be added.

Three-State Output

A chip may have a *three-state output* (also known as *tri-state output*, a term which was trademarked but is now used generically). This means it will be capable of changing its outputs so that instead of sourcing or sinking current in a logical high or low state, they can have a high impedance. The chip then becomes "invisible" to others which may be sharing the same output bus. The high-impedance state is usually applied to all outputs from the shift register simultaneously when enabled by a separate pin, often identified as *OE*, meaning *output-enable*. Examples of three-state shift registers are the 74x595 or the 4094B chip.

The high-impedance state can be thought of as being almost equivalent to switching the outputs of the shift register out of the circuit. Consequently, if other components sharing the bus are also in high-impedance output mode, the state of the bus will "float," with uncertain results. To avoid this, a pullup resistor of 10K to 100K can be used on each bus-line.

Where the internal components of a shift register are shown in a datasheet, a three-state output is usually represented with a buffer symbol or inverter symbol that has an additional control input located on its upper edge, as shown in Figure 12-7. This should not be confused with the similar placement of a positive power supply input to an amplifier or op-amp. (Schematics showing the interior elements of a logic chip almost never include power-supply connections.)

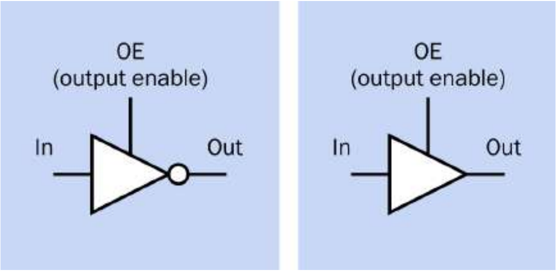

Figure 12-7. *A shift register may be capable of a three-state output, where high impedance is an option additional to the usual high or low logic state. An output enable pin allows this. It is typically shown as an additional input to an inverter (left) or buffer (right) inside the shift-register chip.*

How to Use It

The SISO application of a shift register can be used simply to delay the transmission of data by storing it and moving it from one flip-flop to the next before it is read out of the end of the chain.

The SIPO application of a shift register (serial in, parallel out) may be useful where a microcontroller has insufficient outputs to control multiple devices. Serial data can be sent from a single microcontroller output to the input of a shift register. The chip can then drive a separate device from each of its parallel output pins. If there are eight devices, the microcontroller can send a sequence of eight bits, each of which will control the on/off status of one device when the bits are read out of the shift register in parallel. If more devices are used, an additional shift-register can be daisy-chained to the output from the first.

Clock signals can be supplied from the microcontroller, along with a signal to the clear input of the shift register, if desired. Alternatively, the old bit states in the shift register can simply be "clocked out" and replaced with a new set of serial data. During the process of "bit banging," the parallel outputs of the shift register can remain connected directly with the output devices if the clock speed from the microcontroller is fast enough, as devices such as relays will not respond to extremely brief pulses.

For generic shift registers that do not have open-drain outputs, a buffer will be needed to provide sufficient current for any device using more current than an LED.

If a shift register is configured for PISO mode (parallel in, serial out) it can be placed on the input side of a microcontroller, polling a variety of devices and feeding their states into the microcontroller serially.

Dual Inputs

Where a shift register has two serial inputs (as is often the case), they are almost always linked as inputs to an internal NAND gate. This allows the output from the end of the chain of flip-flops to be connected back to the beginning of the chain, if the shift register is to function as a ring counter. However, if this function is not used and a single input is required, the two inputs to the shift register can be tied together. In this configuration, the input becomes active-low. The two possible configurations are shown in Figure 12-8. It is important never to leave an input *floating*, or unconnected.

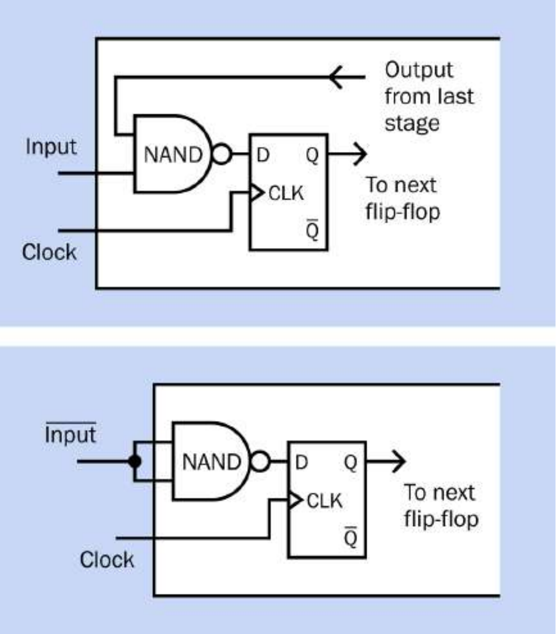

Figure 12-8. *Two possible configurations where a shift register allows two inputs linked with an internal NAND gate.*

Preloading the Shift Register

Where a shift register will be used to output a single recirculating bit (or in similar applications), the component must be preloaded with a high state in its first register. This may require a one-shot (monostable) timer which is activated only when the circuit is powered up.

Polling a Keyboard

Two shift registers can be used to scan the data lines in a matrix-encoded keyboard or keypad. This is often known as *polling* the keyboard. Provided the scan rate is sufficiently fast, the user experiences a seemingly immediate response to a key-press.

While the full schematic is too large and complex to be included here, many examples can be found online.

Arithmetical Operations

Shift registers were traditionally used to perform arithmetical operations on binary numbers. If the number is represented by eight bits (i.e., one byte) with the most significant digit on the left, shifting the bits one space to the right will have the effect of dividing the byte value by 2. If the bits are shifted one space to the left, the byte value will be multiplied by 2 (provided an additional register is available to store the most significant bit after it has been shifted). This concept is illustrated in Figure 12-9.

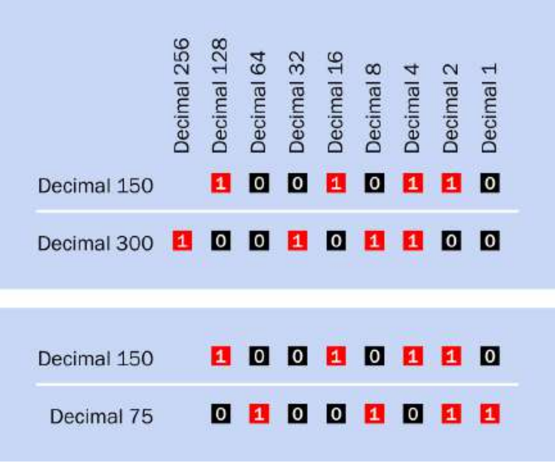

Figure 12-9. *In the upper section of this diagram, a binary number represented by eight bits in a shift register is multiplied by 2 by shifting all the bits one space to the left. In the lower section, the same binary number is divided by 2 by shifting all the bits one space to the right. The binary values are shown in decimal notation to the left.*

In the upper section of the figure, the binary number 10010110 (chosen arbitrarily) is represented in the eight flip-flops of a shift register. The decimal place value of each digit is indicated. Adding up the place values of the 1s in the number, the total is 128 + 16 + 4 + 2 = 150. Below the white line, the digits have been shifted one place to the left, with the leftmost digit carried over to an additional location, while a zero is inserted in the rightmost location. Assuming that the additional place at the leftmost location has a place value of 256, the total is now 256 + 32 + 8 + 4 = 300.

In the lower section of the figure, the same binary number has been shifted one space to the right, with a 0 introduced in the leftmost location. The decimal value is now 64 + 8 + 2 + 1 = 75.

While this application for shift registers was common during development of digital computing in the 1960s and 1970s, the shift register as a separate component became less common subsequently, as its functionality was acquired by modern CPU chips.

Buffering

A shift register may also be used as a *buffer* between two circuits where the clock speeds are different. Digits are clocked in by the first circuit, then clocked out by the second. Some shift registers allow two clock inputs and can be used for this purpose.

What Can Go Wrong

Problems that tend to affect digital chips generally are listed in the entry on logic gates (see "What Can Go Wrong" on page 105).

Confusing Classification

Because of the functional similarity to a binary ripple **counter**, a shift register is sometimes listed by component suppliers as a counter. In fact, a binary counter will almost always have outputs that have place values 1, 2, 4, 8.... and upward, while the outputs from a shift register will not have place values.

When searching for a shift register, it can be found by specifying a "counting sequence" of serial to parallel, serial to serial, parallel to serial, or parallel to parallel. If the "counting sequence" is simply up or down, the component is a counter, not a shift register.

Inadequate Setup Time

Each flip-flop in a shift register must have a stable input state before the next triggering event shifts the data. If this setup time is reduced below the minimum specified in the datasheet, results will be unpredictable.

Unconnected Input

Because many shift registers have a choice of two inputs to the same chain of internal flip-flops, it is easy to leave one of them unconnected by accident. A floating input will be vulnerable to any stray electromagnetic fields, and is almost certain to create random effects.

Output Enable Issues

The output enable pin on a shift register that has three-state logic outputs is usually active-low. Consequently, if the pin is left unconnected, the logic outputs may go into high-impedance mode, or will fluctuate unpredictably. Where three-state outputs are not required, a safe course of action will be to avoid using three-state chips.

Floating Output Bus

If a pullup resistor is omitted from a bus that is shared by three-state chips, the results will be unpredictable. Even if the circuit design seems to guarantee that at least one chip will have a high or low output on the bus, a pullup resistor should still be included.

counter

13

The term **counter** is used here to mean a digital-logic chip. A counter could be built from discrete transistors, but this approach is obsolete. Counters may also be devised from parts such as multiple relays, or a solenoid advancing a ratchet wheel, but such electromechanical devices are not included in this encyclopedia.

In this encyclopedia, a counter by definition has binary-weighted outputs with values 1, 2, 4, 8…. in decimal notation. The exception to this rule is a *ring counter*, which does not have binary-weighted outputs but is included here because its name identifies it as a counter. A **shift register** may be used as a ring counter, but is more versatile and has many other functions; hence it has a separate entry.

Gray code counters, in which successive outputs differ by only one binary digit, are not described in this encyclopedia.

OTHER RELATED COMPONENTS

- **flip-flop** (see Chapter 11)
- **shift register** (see Chapter 12)

What It Does

A counter can be used to count events, or can measure time in convenient intervals if it is connected with a component such as a *quartz crystal* that operates at a precise and reliable frequency. The counter receives input pulses (usually referred to as a *clock input*) and counts a predetermined number of them before restarting from the beginning. It will repeat in this fashion so long as power is connected, and the clock pulses continue, and a reset signal is not supplied.

Almost all counters create some form of output during the count. Most commonly, the output is a pattern of high and low states expressing the number of clock pulses in binary code. Where a counter counts to a very high number before recycling, some intermediate binary digits may be omitted.

While standalone counter chips are not as widely used now as in the early days of computing, they still find application in industrial processes, small devices, and education, and can be used to control incremental devices such as stepper motors. They can be used in conjunction with microcontrollers.

Schematic Representation

No specific logic symbol exists for a counter. It is most often shown in a schematic as a rectangle with clock input(s) and clear input(s) on the left and outputs on the right. An example appears in Figure 13-1, above a representation of the physical chip and its pinouts. The meaning of the abbreviations identifying the inputs, outputs, and control functions will be found in "How It Works" on page 132. Because the two MR inputs for this particular counter are ANDed inside the chip, the AND symbol is included with the counter symbol.

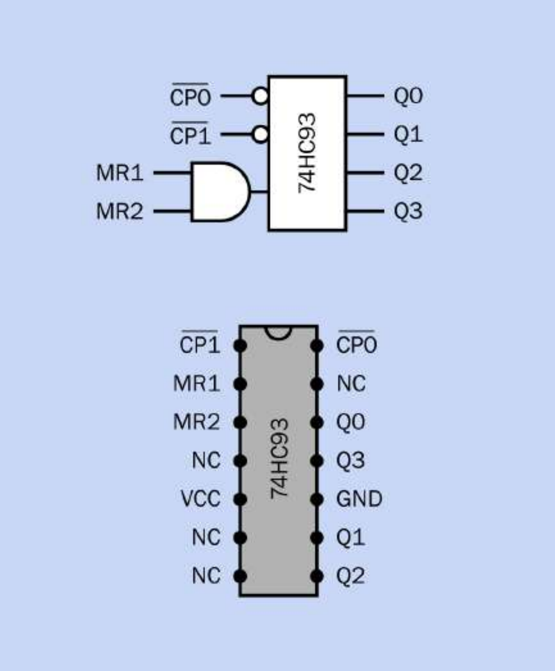

Figure 13-1. *Typical schematic representation of a counter, compared with the pinouts of the actual component.*

A counter chip is shown in Figure 13-2.

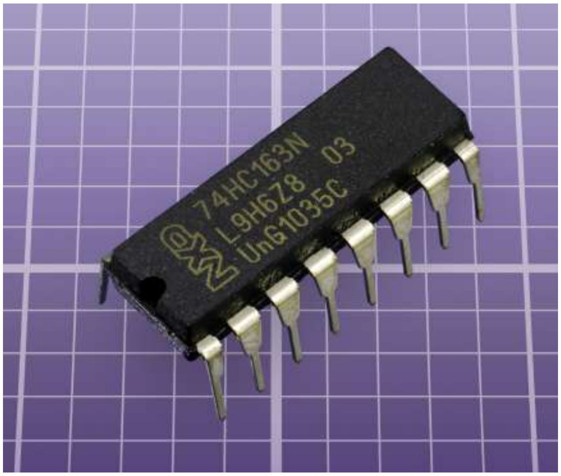

Figure 13-2. *The 74HC163 shown in this photograph is a 4-bit synchronous counter capable of being preloaded with a starting value, and able to do a synchronous reset.*

How It Works

A counter is built from a chain of **flip-flops**, with each one triggering the next. JK, T-type, or D-type flip-flops may be used. For a thorough description of a flip-flop, see Chapter 11. In Figure 13-3, a D-type flip-flop is shown, triggered by each rising clock pulse.

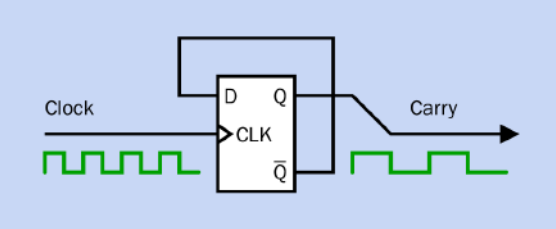

Figure 13-3. *When the complementary (NOT-Q) output from a D-type flip-flop is connected back to its input, the Q output frequency is half of the clock input frequency. See text for additional details.*

Initially the Q output of the flip-flop is low, so its NOT-Q output (identified by Q with a bar above it) is high. This feeds back to the D input, but has no effect until the rising edge of the next clock pulse copies the high D input to the Q output. The Q output is now latched high while the NOT-Q output is latched low and feeds back to the D input. The triggering event has passed, so the low D input does not have any immediate effect. The rising edge of the next clock pulse copies the low state of the D input to the output, and changes the NOT-Q output to high, causing the cycle to repeat. As a result, the output frequency of the flip-flop is one-half of the input frequency from the clock. If the output is carried to the next flip-flop to become its clock input, once again the frequency will be divided by two.

Modulus and Modulo

The *modulus* of a counter is the number it will count up to, before it repeats. This is sometimes written incorrectly as *modulo*.

In fact, *modulo* is the name of an arithmetical operation, often abbreviated as *MOD* (usually capitalized, even though it is not an acronym). This operation consists of a division sum in which the *remainder* from the division is the result of the operation. Thus 100 modulo 5 gives a result of 0, because there is no remainder when 100 is divided by 5. But 100 modulo 7 gives a result of 2, because 2 is the remainder of the division operation.

To exacerbate the confusion, *MOD* is also used as an adjective referring to the modulus of a counter. Thus a MOD-4 counter has a modulus of 4, and a MOD-16 counter has a modulus of 16. As a general rule, when a counter is being described, *modulo* and *MOD* will mean the same thing as *modulus*. This may be confusing to people such as computer programmers who are already familiar with the correct usage of MOD as an arithmetical operator.

In a counter, to obtain a modulus that is not a power of two, logic gates inside the chip can intercept a particular value (such as 1010 binary, which is 10 decimal) and use this as a signal to restart the count at zero. External connections to the chip can achieve the same purpose.

Pin Identifiers

Abbreviations and acronyms are used in datasheets to identify pin functions. These identifiers have not been standardized, and many variants exist.

CLK is the abbreviation most commonly used for the clock input, sometimes alternatively shown as CK or CP. If it is active-low, or if its falling edge will be active, a bar will be printed above it. Where a printed font does not permit an underscore to be placed in this way, CLK' (the abbreviation followed by an apostrophe) may be used instead. Often two or more stages of a counter can be separately clocked, and the input pins will be identified with abbreviations such as CLK1 and CLK2, or 1CLK and 2CLK, or CKA and CKB, or CP1 and CP2, or similar.

Where a clock input is edge-triggered, this is indicated by a small triangle. The triangle can be seen in Figure 13-3.

CLR designates a pin which will clear the count and reset it back to zero. This signal is often active-low, indicated by a bar printed above the abbreviation.

In a schematic, a small circle, properly known as a *bubble*, may be placed at any input which is active-low. On a clock input, the circle indicates that it is falling-edge triggered. See Figure 11-16 for various implementations of symbols with flip-flop schematics.

The CLR operation may be *synchronous* (the pin state will not be recognized until the next clock pulse) or *asynchronous* (in which case the pin state overrides the clock and resets the counter immediately). MR stands for "Master Reset," and has the same function as CLR.

Where two or more counters (or multiple stages within one counter) can be reset separately, more than one clear input will be shown, and may be identified with abbreviations such as CLR1 and CLR2, or MR1 and MR2.

Output pins will almost always be identified as Q0, Q1, Q2 … or QA, QB, QC … up to the maximum necessary to express the modulus. If two or more counters are included in one chip, outputs may be prefixed with an appropriate number. Thus 2Q3 would be the third output in the second counter. Multiple counters in one chip are identified with numbers starting from 1.

Where internal flip-flops are shown, they will have identifiers such as FF1 or FF2. Each flip-flop will have its own clear function, identified as C or CD (the latter meaning "clear data"), and may have inputs labeled D1, D2, D3… in a D-type flip-flop or J and K in a JK flip-flop. See Chapter 11 for an explanation of flip-flop inputs and outputs.

The input to a counter is always imagined to begin from the left, and consequently the leftmost flip-flop shown in an internal schematic will express the least significant bit of the current value of the count, even though a binary number is written with the least significant bit in the rightmost place.

If a counter is capable of receiving parallel data as an input (explained below) it will have a pin labeled PE for parallel enable. It may also have a CE or CET pin, for count enable.

As is generally the case in logic chips, VCC or V+ are typically used to identify the positive power supply pin, while GND or V- will identify the negative-ground pin. NC means that a pin has no internal connection at all, and consequently requires no external connection.

Variants

All counter chips use binary code internally, and the number of *bits* (*binary digits*) in the counter's modulus will be the same as the number of internal flip-flops. A 4-bit counter (the usual minimum) will have a modulus of 2^4 which is 16. A 21-bit counter (the maximum typically available) will have a modulus of 2^{21} which is 2,097,152. For higher moduli, counters can be chained together, each sending a carry signal to the next. This is known as a *cascade*.

Multiple counters, with different moduli, may be chained in a single chip. For example, in a digital clock that displays hours and minutes using the 60Hz frequency of an American domestic power supply as its timebase, the initial counting stage will have a modulus of 60, to count individual seconds. The next counting stages will have moduli of 60, 10, and 6, so that they can count from 00 to 59 minutes. Additional stages in the chip will tally hours.

A counter with a *parallel input* can be preloaded with an initial value (in binary code) from which it may count up or down. A *parallel-enable* pin may put the counter into a mode where the number can be *jam loaded*, meaning that it is jammed into the counter regardless of the clock state. Other types of counter are loaded synchronously.

Ripple versus Synchronous

In a *ripple counter* each internal flip-flop triggers the clock input of the next, so that their states change in a rapid but incremental sequence from the input to the output. This is also known as an *asynchronous* counter. Because the final state will not be valid until the clock pulse has rippled all the way through the counter (and through additional counters if they are cascaded together), a ripple counter will tend to suffer from a *propagation delay* of up to a microsecond. Ripple counters may also create output spikes or momentary transient count values that are invalid. Therefore, they are more suitable for applications such as driving a numeric display than for interfacing at high speed with other logic chips.

In a *synchronous counter*, all the flip-flops are clocked simultaneously. A synchronous counter is better suited to operation at high speed.

Of the counter chips available today, about half are synchronous and half are asynchronous.

Ring, Binary, and BCD

A counter that activates output pins one at a time sequentially is said to have a *decoded output*. It is often referred to as a *ring counter*. It has the same number of output pins as its modulus. An example is the 4017B chip.

A *binary counter* is more common and has an *encoded output*, meaning that it will express the running total of the count in binary code through *weighted outputs* that typically have (decimal) values of 1, 2, 4, 8, and so on. A modulus-8 counter (often referred to as an *octal counter*) will require three outputs which represent the binary numbers 000, 001, 010, 011, 100, 101, 110, and 111 (decimal 0 through 7) before going back to 000 and repeating.

A *modulus-16* counter, also known as a *hexadecimal counter* or a *divide-by-16 counter*, will have a binary output represented by four output pins, counting from 0000 through 1111 (decimal 0 through 15). Four-digit binary counters are very common, and their outputs are compatible with other components such as a **decoder**, which converts a binary-number input into a ring-counter-style output.

A *decade counter* is a modulus-10 binary counter. It is described as having a *binary-coded decimal* output (often expressed with the acronym *BCD*), using four weighted output pins which represent the numbers 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, and 1001 (decimal 0 through 9) before repeating. Because this counter skips binary outputs from 1010 through 1111 (decimal 10 through 15), it is said to have a *shortened modulus*.

Figure 13-4 shows a schematic diagram of JK flip-flops in a decade ripple counter. The J and K inputs are all tied to the positive power supply, as this causes the clock input to toggle the output high and low. Note that because the primary input is always shown at the left end of the component, the least significant output bit (Q0) is in the leftmost position.

To intercept binary 1010 (decimal 10), an internal NAND gate is used. Its output goes low when its two inputs, from Q1 and Q3, go high. The output from the NAND immediately activates the CLR function on all the flip-flops, so that as soon as the decade counter reaches 1010 (decimal 10), it resets itself to 0.

In this particular chip, the preload for each flip-flop is tied to the positive power supply, so that it is always inactive. In some counters, the preload feature of each flip-flop is accessible via pins outside the chip. This creates the potential hazard of preloading the counter with one of the numbers that it normally skips (for instance, 11 decimal in a decade counter). This is referred to as an *invalid number* or *disallowed state*. (This use of the term "state" refers only to the binary number stored in the counter's flip-flops. It has nothing to do with the high-state or low-state voltages used to represent binary 0 or 1.)

The counter's datasheet should include a *state diagram* showing how the counter will deal with this situation. It may reset itself to a valid value after a maximum of two steps, but this can still cause confusion, depending on the application.

The state diagram for a 74HC192 counter is shown in Figure 13-5.

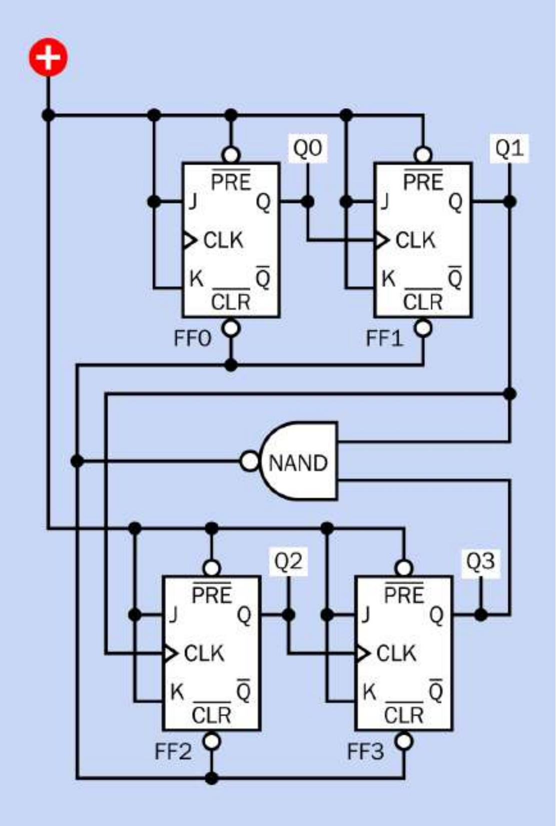

Figure 13-4. *The internal logic of a synchronous decade counter that uses JK flip-flops.*

Clock Sources

The clock input may be provided by a **timer** chip or by an *RC network*, which has the advantage of being able to run at a relatively low speed for purposes where this is desirable. It may alternatively be provided by a *quartz crystal* oscillating at a much higher frequency such as 1MHz. Successive counters may be necessary to reduce this value, depending on the application.

In some counters, the clock is built into the chip. More commonly, a resistor and capacitor may be used externally to establish a clock rate in conjunction with logic gates inside the chip. The datasheet for this type of component will include a formula for calculating the clock frequency from the resistor and capacitor values. The 4060B chip is an example.

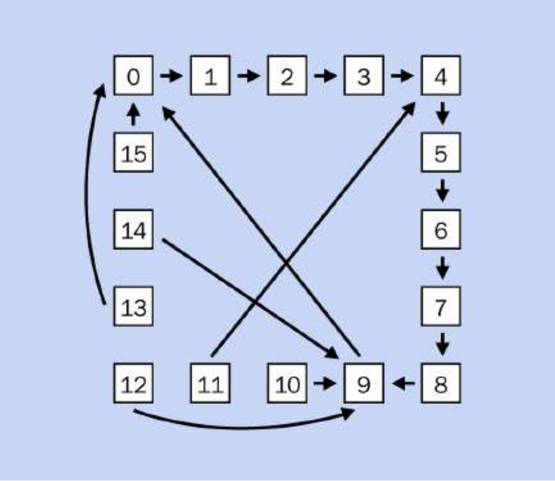

Figure 13-5. *A state diagram shows the transitions that a counter will make from each number to the next (in decimal notation), including the transitions which enable it to exit from disallowed states. This example is for a 74HC192 chip.*

Rising Edge and Falling Edge

A counter may be designed to be triggered either by the rising edge or the falling edge of the clock input, or by its high or low logic state. Generally speaking, ripple counters use the falling edge, so that the final output from one counter can become the clock input of the next. In other words, when the most significant digit of the first counter changes from a high to low logic state, this transition toggles the least significant bit of the second counter.

Synchronous counters generally use the rising edge of the clock input. If multiple synchronous counters are cascaded, they must all share the same clock signal, and will all change their flip-flop states simultaneously.

Multiple Stages

It is common for a counter chip to contain two or more *stages* with differing moduli. To take a common example, a divide-by-2 stage and a divide-by-5 stage that are both present in a single chip can be used to create a decade counter by connecting external pins. The extra stages provide a choice of moduli if they are used individually.

Single and Dual

Counter chips may contain two counters of the same modulus. This is known as a *dual* counter. Dual 4-bit counter chips are common. Each counter can be used separately, or they can be cascaded, in which case the total modulus will be found by multiplying the individual moduli together.

High-State, Low-State, and Three-State

Almost all counters use positive logic where a 1 is represented by a high state and 0 by a low state. Some counters allow an additional output state which has a high impedance and is equivalent to an open circuit. This feature is useful when two or more chips share the same output bus. It is discussed in the entry for **shift registers** in "Three-State Output" on page 126.

Descending Output

Most components only create an ascending count. The output can be converted to a descending count by passing each binary state through an inverter, but this will only work properly if the modulus is equal to the number of states. In a BCD counter, its inverted outputs will count from decimal 15 to decimal 6, not from decimal 9 to decimal 0.

A few counters are available which are designed to create a correct descending count. Other counters are available which allow the user to set the mode to ascending or descending. Examples are the 74x190 or 74x192 (where an acronym for the chip family will be substituted for the letter x).

A descending output is useful in combination with a parallel input, where a user may set an initial value from which the counter will descend to 0. With suitable logic, this can enable a user-specified delay period.

Programmable Counters

A programmable counter can usually allow a modulus ranging from 2 to more than 10,000. The counter counts down by dividing an initial number repeatedly with a value that is preset with binary inputs. An example is the 4059B chip.

Examples

Many counter specifications date back to the 4000 family of logic chips. Versions of them subsequently became available in the 74xx series, often with the old 4000 part number preceded by 74x (where x is replaced by a designation of the logic family). For example, a version of the 4518B dual BCD chip can be obtained as the 74HC4518. As is the case with all logic chips, this part number will be preceded by letter(s) designating the particular manufacturer, with a suffix added to differentiate minor variants of the chip. The 74xx series has the advantage of higher speed and a greater ability to source or sink current at its output pins.

Most of the original CMOS chips, such as the 4518B, are still available, even in surface-mount versions. These offer the possible advantage of being able to use a higher power supply voltage.

Many counters offer multiple options such as different modulus values that can be selected by external pin connections. Some chips are tolerant of slow clock frequencies; others are not. Most are edge-triggered, but a few are level triggered. Some, such as the 4518B mentioned above, allow a choice of a rising-edge clock input and falling-edge clock input on different pins. For a specific application, it is really necessary to read a variety of datasheets to select the chip that is most suitable.

Values

As is the case with other logic chips, most counters in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount 74xx versions may use voltages as low as 2VDC.

See the section on logic gates in Chapter 10 for a discussion of acceptable high and low logic-input states. On the output side, the 4000 series chips are able to source or sink less than 1mA at 5VDC, but the 74HCxx series can usually manage around 20mA.

A few counters are capable of delivering more power through additional output stages that can drive LEDs. The 4026B decade counter is still being manufactured, capable of powering modest 7-segment displays. The 4033B has the additional option of blanking any leading zeros in a multidigit display. Other chips that were designed for direct connection to LED numerals have become obsolete as the need for this application has diminished. The 74C925, 74C926, 74C927, and 74C928 are examples. They may still be found from surplus outlets, but should not be specified in new circuit designs.

What Can Go Wrong

The entry that deals with problems affecting shift registers (see "What Can Go Wrong" on page 128), describes issues which also affect counters. The entry that deals with logic chips (see "What Can Go Wrong" on page 105), describes problems affecting all types of logic chips. In addition, the potential problems listed below are specific to counters.

Lock-Out

This is the condition which occurs if a counter with a shortened modulus is loaded with a binary state that is out of its range. Consult the datasheet and study its state diagram to determine the most likely outcome if this problem occurs.

Asynchronous Artifacts

Because the flip-flops in an asynchronous (ripple) counter do not change simultaneously, they create very brief false outputs while the ripple process is taking place. In a 4-bit counter, the binary number 0111 (decimal 7) should be followed by 1000 (decimal 8). However, the rightmost digit (i.e., the least significant bit) will change to a 0 initially, creating 0110 as a momentary binary output (decimal 6). The carry operation will then change the next digit to a 0, creating 0100 (decimal 4). The carry operation continues, changing the next digit to a 0, creating 0000. Finally the operation completes by creating 1000 as the correct output.

These intermediate states on the output pins are often referred to as *glitches*. Because they are extremely brief, they will be indetectable when a counter is used to drive a display. They can cause significant issues, however, if the outputs of the counter are connected with other logic chips.

Another type of asynchronous problem will occur if the clock speed is sufficiently high that a new pulse is received at the first flip-flop before the ripple of changing states has passed all the way through to the final flip-flop. This will result in a different brief invalid value on the output pins.

Noise

Old TTL-type counters, such as the 74LSxx series, are especially noise-sensitive. Adding a 0.1μF or 0.047μF bypass capacitor as close to the power supply pin as possible is recommended. Breadboarding counters of this type may result in errors if a high-frequency clock is used, because conductors such as patch-cords are liable to pick up noise. Modern 74HCxx counters are preferable.

encoder

14

In this encyclopedia, an **encoder** is a digital chip that converts a decimal-valued input into a binary-coded output.

The term “encoder” may alternatively refer to a **rotational encoder** (also known as a *rotary encoder*) which has a separate entry in Volume 1 of this encyclopedia. The term may also describe a *code hopping encoder*, which is an encryption device used in keyless entry systems for automobiles.

OTHER RELATED COMPONENTS

- **decoder** (see Chapter 15)
- **multiplexer** (see Chapter 16)

What It Does

An encoder is a *logic chip* that receives an input consisting of an active logical state on one of at least four input pins, which have decimal values from 0 upward in increments of 1. The encoder converts the active pin number into a binary value represented by logic states on at least two output pins. This behavior is opposite to that of a **decoder**.

Encoders are identified in terms of their inputs and outputs. For example:

- 4-to-2 encoder (four input pins, two output pins)
- 8-to-3 encoder (eight input pins, three output pins)
- 16-to-4 encoder (sixteen input pins, four output pins)

In the early days of computing, encoders processed interrupts. This application is now rare, and relatively few encoder chips are still being manufactured. However, they are still useful in small devices—for example, if a large number of inputs must be handled by a microcontroller that has insufficient pins to receive data from each individually.

Schematic Symbol

Like other logic-based components, the encoder does not have a specific schematic symbol and can be represented by a plain rectangle as in Figure 14-1, with inputs on the left and outputs on the right. The bars printed above some of the abbreviations indicate that an input or output is active-low. In this chip, the 74LS148, all inputs and outputs are active-low.

Generally speaking, inputs labeled D0, D1, D2… are used for data input, although they may simply be numbered, with no identifying letter. The encoded outputs are typically identified as Q0, Q1, Q2… or A0, A1, A2… with Q0 or A0 designating the least significant bit in the binary number.

Pins labeled E and GS are explained in the following section.

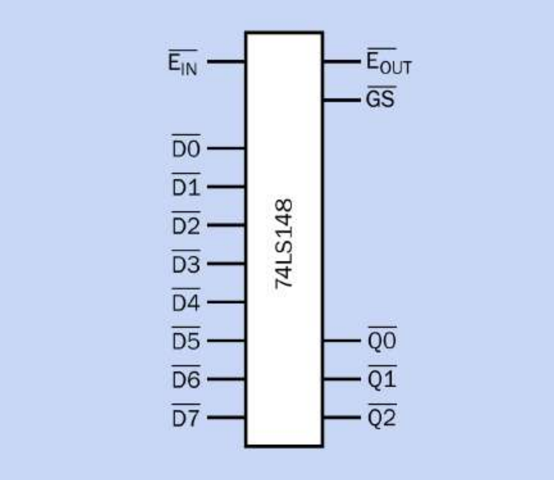

Figure 14-1. *While no specific schematic symbol exists for an encoder chip, this style is commonly used. Shown here is a 16–to–4 encoder with active-low inputs and outputs.*

Similar Devices

The similarities and differences between encoder, decoder, multiplexer, and demultiplexer can cause confusion.

- In an **encoder**, an active logic state is applied to one of four or more input pins, while the rest remain in an inactive logic state. The input pin number is converted to a binary code which is expressed as a pattern of logic states on two or more output pins.
- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive logic state.
- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.
- An analog multiplexer may allow its inputs and outputs to be swapped, in which case it becomes a **demultiplexer**. It can connect a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

How It Works

An encoder contains **logic gates**. The internal logic of an 8-to-3 encoder is shown in Figure 14-2, where the darker blue rectangle represents the chip. The switches in this figure are external and are included only to clarify the concept. An open switch is imagined to provide an inactive logic input, while a single closed switch provides an active logic input. (Multiple active inputs can be handled by a *priority encoder*, described below).

Each input switch has a numeric status from 1 to 7. The switch with value 0 does not make an internal connection, because the output from the OR gates is 000 by default.

The logic state of each OR output represents a binary number, weighted with decimal values 1, 2, and 4, as shown at the bottom of the figure. Thus, if switch 5 is pressed, by tracing the connections it is clear that the outputs of OR gates 4 and 1 become active, while the output from gate 2 remains inactive. The values of the active outputs thus sum to 5 decimal.

Figure 14-3 shows the outputs for all possible input states of a 4-to-2 encoder. Figure 14-4 shows the outputs for all possible input states of an 8-to-3 encoder. These diagrams assume that a high logic state is an active logic state, on input or output. This is usually the case.

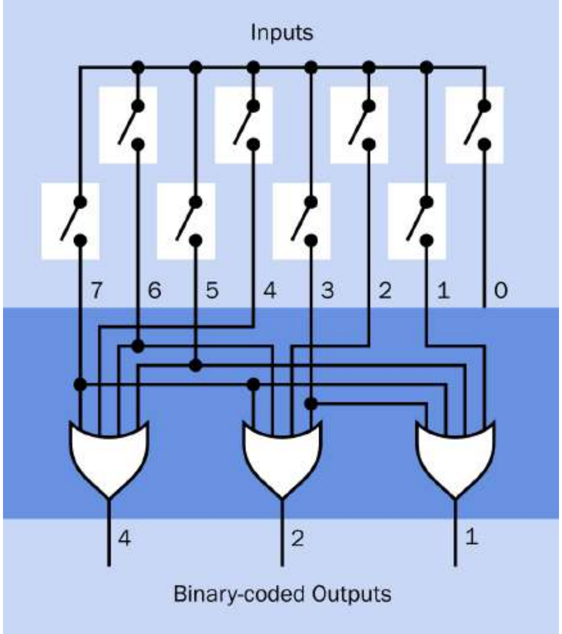
Inputs
7
6
5
4
3
2
1
0
4
2
1
Binary-coded Outputs

Figure 14-2. *A simplified simulation of the internal logic of an 8-to-3 encoder. The dark blue rectangle indicates the space inside the chip. The external switches are included only to clarify the concept. An encoder chip would have an Enable line to create an active output.*

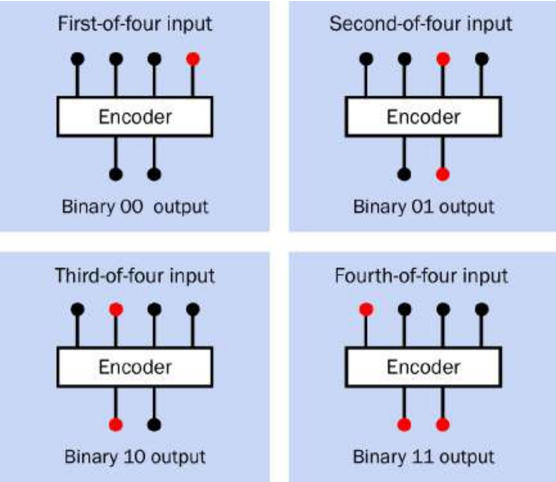
First-of-four input
Encoder
Binary 00 output
Second-of-four input
Encoder
Binary 01 output
Third-of-four input
Encoder
Binary 10 output
Fourth-of-four input
Encoder
Binary 11 output

Figure 14-3. *The four possible inputs of a 4-to-2 encoder (top of each panel) and the encoded outputs (below).*

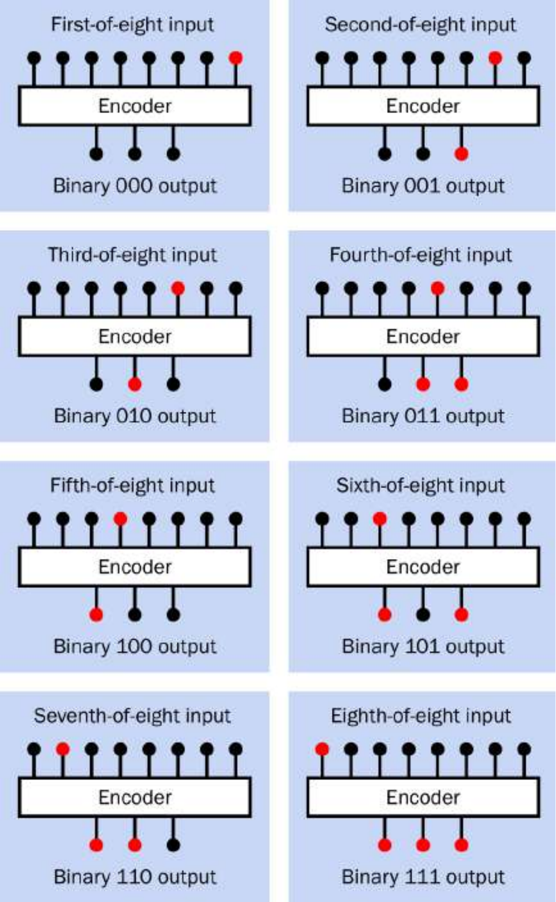
First-of-eight input
Encoder
Binary 000 output
Second-of-eight input
Encoder
Binary 001 output
Third-of-eight input
Encoder
Binary 010 output
Fourth-of-eight input
Encoder
Binary 011 output
Fifth-of-eight input
Encoder
Binary 100 output
Sixth-of-eight input
Encoder
Binary 101 output
Seventh-of-eight input
Encoder
Binary 110 output
Eighth-of-eight input
Encoder
Binary 111 output

Figure 14-4. *The eight possible inputs of an 8-to-3 encoder (at the top of each panel), and the encoded outputs (below). Note that one input of an encoder must always be logic-high. All logic-low inputs are not a valid state.*

Unlike ripple counters, where propagation delays can reduce the overall response time of the component, decoders respond within two or three nanoseconds.

Variants

A *simple encoder* assumes that only one input pin can be logically active at a time. A *priority encoder* assigns priority to the highest-value input pin if more than one happens to receive an active input. It ignores any lower-value inputs. An example is the 74LS148, which is an 8-to-3 chip.

A few encoders feature *three-state* outputs (also known as *tri-state*), in which a high-impedance or "floating" output state is available in addition to the usual high and low logic states. The high-impedance state allows multiple chips to share an output bus, as those that are in high-impedance mode appear to be disconnected. This is useful if two or more encoders are cascaded to handle a larger number of inputs.

Values

As is the case with other logic chips, most encoders in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See the section on logic gates in Chapter 10 for a discussion of acceptable high and low input states. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can manage around 20mA.

How to Use It

Suppose that a microcontroller should respond to an eight-position rotary switch. Because the switch cannot be turned to more than one position at a time, all of its eight contacts can be connected with the inputs on an encoder, which will deliver a 3-bit binary number to three inputs of the microcontroller. Code inside the microcontroller then interprets the pin states.

This is shown in Figure 14-5. Pulldown resistors would be needed on the input pins of the encoder, to prevent them from floating when they are not connected by the rotary switch. They have been omitted from this diagram for simplicity. Debouncing the switch would be handled by the microcontroller.

Other forms of input may be used instead of a rotary switch. For example, the outputs from eight comparators or eight phototransistors could be passed through an encoder.

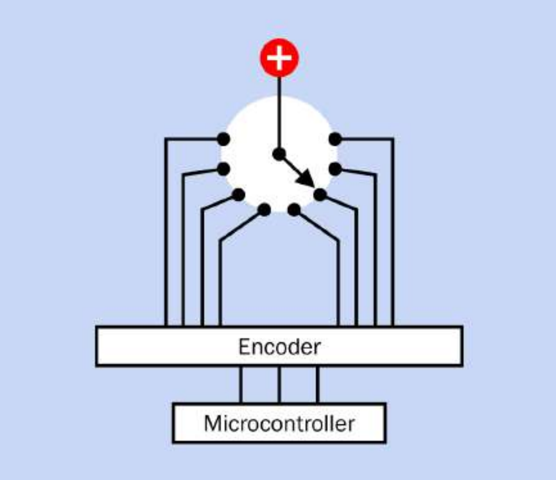

Figure 14-5. *Output from an eight-input rotary switch could be connected through an 8-to-3 encoder to provide input to a microcontroller using a reduced number of pins. Pulldown resistors have been omitted for simplicity.*

Cascaded Encoders

Encoders are often provided with features to facilitate handling additional inputs via multiple chips. Typically, a second Enable pin is provided, as an output that connects with the Enable input of the preceding chip. This preserves the priority function, so that an input on the second chip prevents any additional input to the first chip from affecting the output. In a datasheet, the enable pins may be labeled E_{IN} and E_{OUT}, or E_I and E_O.

In addition, a GS pin will be included, meaning "Group Select." It is logically active only when the encoder is enabled and at least one input is active. The GS pin of the most-significant encoder provides an additional binary digit.

The outputs from two encoders can be linked via OR gates, as shown in Figure 14-6, where the lower chip's GS output provides the most significant bit of a four-bit binary number.

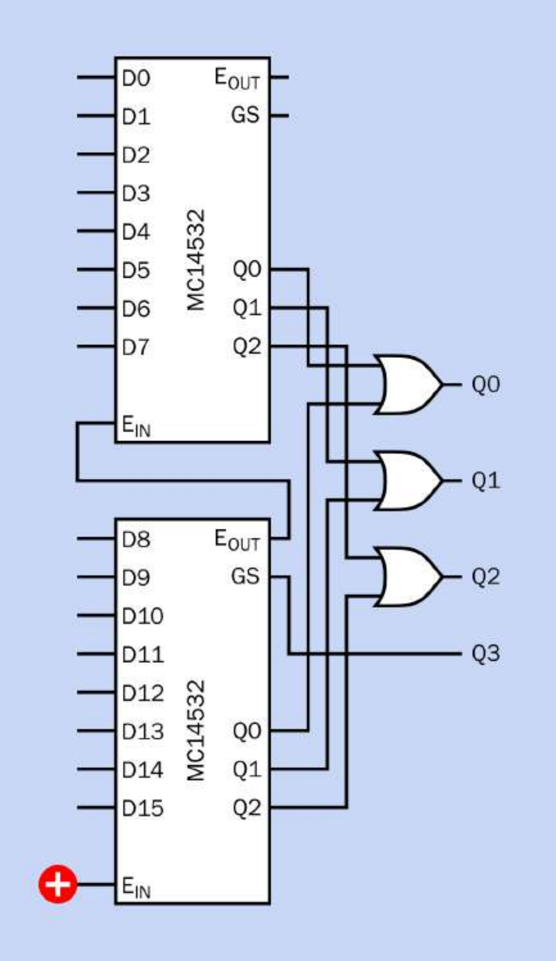
D0
D1
D2
D3
D4
D5
D6
D7
EOUT
GS
MC14532
Q0
Q1
Q2
EIN
D8
D9
D10
D11
D12
D13
D14
D15
EOUT
GS
MC14532
Q0
Q1
Q2
EIN
Q0
Q1
Q2
Q3

Figure 14-6. *Two eight-bit encoders can be cascaded to handle 16 separate inputs. In this example, the encoders use active-high logic.*

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates in "What Can Go Wrong" on page 105.

See "What Can Go Wrong" on page 149 in the entry describing **decoders** for a list of more specific problems that also afflict encoders.

decoder

15

In this Encyclopedia, a **decoder** is a digital chip that receives a binary-coded input and converts it to a decimal output by applying a logic state to one of a sequence of pins, each of which is assigned an integer value from 0 upward.

The term "decoder" also refers to components and devices that have other functions, such as decoding audio or video formats. These functions are not included here.

OTHER RELATED COMPONENTS

- **encoder** (see Chapter 14)
- **multiplexer** (see Chapter 16)

What it Does

A decoder receives a binary-coded number on two or more input pins. It *decodes* that number and expresses it by activating one of at least four output pins.

The behavior of a decoder with a two-bit binary input is shown in four sequential snapshots in Figure 15-1, where the least significant bit of the input is on the right in each diagram, and the output moves from right to left.

Figure 15-2 shows a similar sequence in a decoder where various values of a three-bit input are decoded to create an eight-pin output.

One sample state of a four-bit decoder is shown in Figure 15-3.

All of these figures assume that a high state represents an active input or output. In a few chips, a low state is used to represent an active output.

Decoders with 2, 3, or 4 input pins are common. To handle a binary input greater than 1111 (decimal 15), decoders can be chained together, as described below.

Manufacturers' datasheets often describe decoders in terms of their inputs and outputs. Typical examples would include:

- 2-to-4 decoder (two input pins, four output pins)
- 3-to-8 decoder (three input pins, eight output pins)
- 4-to-10 decoder (for converting *binary-coded decimal* to decimal output)
- 4-to-16 decoder (also known as a *hex decoder*).

Input Devices

The input pins of a decoder can be driven by a counter that has a binary-coded output. A decoder can also be driven by a microcontroller, which may have an insufficient number of output pins to control a variety of devices. Two, three, or four of the outputs can be used to represent a binary number which is passed through the decoder to activate the devices one at a time, perhaps with transistors or Darlington arrays introduced to handle the load. This is suggested in Figure 15-4.

A **shift register** can be used for a similar purpose, but often has only one pin for input. This pin must be supplied sequentially with a serial pattern of bits that will match the desired high/low states of the output pins. The relative advantage of this system is that a shift register can generate any pattern of output states. A one-of-many decoder can activate only one output at a time.

LED Driver

A special case is a *seven-segment decoder* designed to drive a seven-segment **LED display** numeral. A binary-coded decimal number on four input pins is converted to a pattern of outputs appropriate for lighting the segments of the display that will form a number from decimal 0 through 9.

Schematic Symbol

Like other logic-based components, the decoder does not have a specific schematic symbol and is represented by a plain rectangle as in Figure 15-5, with inputs on the left and outputs on the right. The bars printed above the E and LE abbreviations (which stand for Enable and Latch Enable, respectively) indicate that they are active-low. In this chip, the 74HC4514, all outputs are active-high, but in a related 4-to-16 decoder, the 74HC4515, all outputs are active-low. In both of these chips, the Enable pin is held low to activate the outputs. The Latch Enable pin freezes the current state of the outputs (i.e., it latches them) when it is held low.

Generally speaking, pins labeled A0, A1, A2… in a datasheet are often the binary inputs (although A, B, C… may be used), with A0 designating the least significant bit. Outputs are usually labeled Y, and are activated in sequence from Y0 when the binary input starts counting upward.

Similar Devices

The similarities and differences between encoder, decoder, multiplexer, and demultiplexer can cause confusion.

- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive logic state.
- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.
- An analog multiplexer may allow its inputs and outputs to be reversed, in which case it becomes a **demultiplexer**. It can connect a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

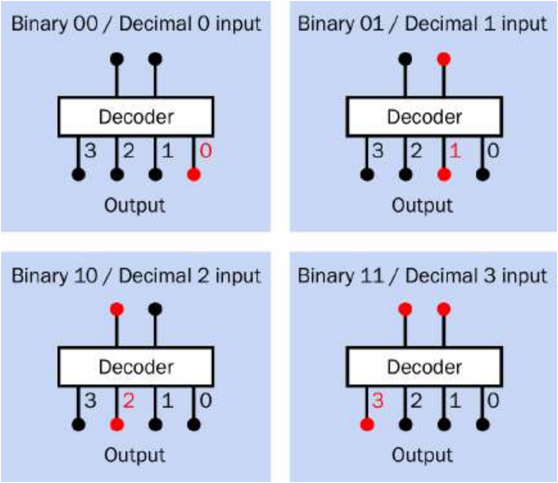
Binary 00 / Decimal 0 input
Decoder
3
2
1
0
Output
Binary 01 / Decimal 1 input
Decoder
3
2
1
0
Output
Binary 10 / Decimal 2 input
Decoder
3
2
1
0
Output
Binary 11 / Decimal 3 input
Decoder
3
2
1
0
Output

Figure 15-1. *A decoder with two input pins can interpret their binary-number representation to create an active logic state on one of four output pins.*

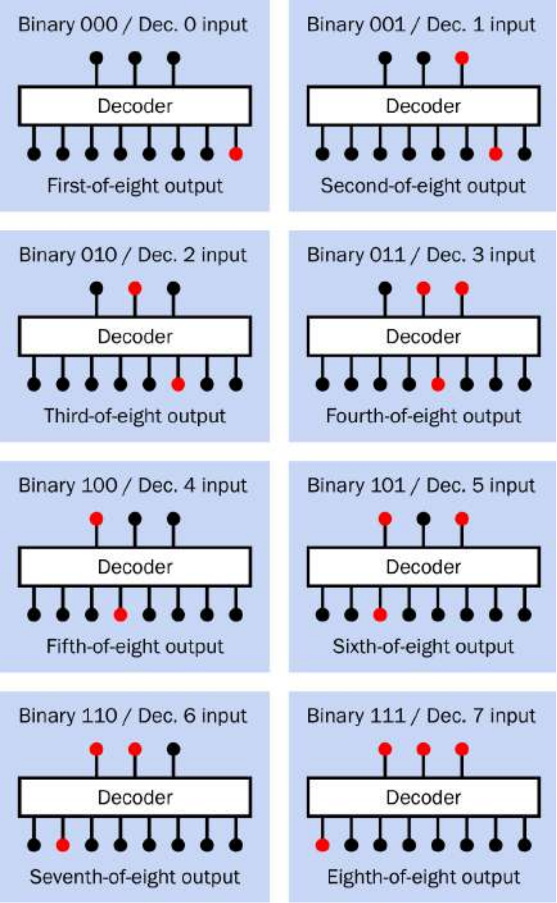
Binary 000 / Dec. 0 input
Decoder
First-of-eight output
Binary 001 / Dec. 1 input
Decoder
Second-of-eight output
Binary 010 / Dec. 2 input
Decoder
Third-of-eight output
Binary 011 / Dec. 3 input
Decoder
Fourth-of-eight output
Binary 100 / Dec. 4 input
Decoder
Fifth-of-eight output
Binary 101 / Dec. 5 input
Decoder
Sixth-of-eight output
Binary 110 / Dec. 6 input
Decoder
Seventh-of-eight output
Binary 111 / Dec. 7 input
Decoder
Eighth-of-eight output

Figure 15-2. *A decoder with three input pins can interpret their binary-number representation to create a high logic state on one of eight output pins.*

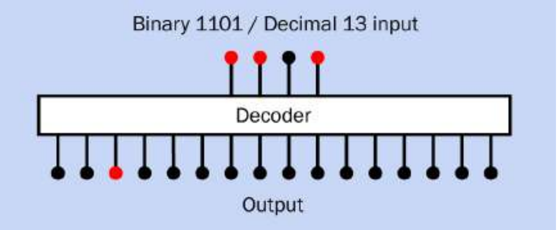
Binary 1101 / Decimal 13 input
Decoder
Output

Figure 15-3. *A decoder with four input pins can interpret their binary-number representation to create a high logic state on one of 16 output pins. Only one of the 16 possible states is shown here.*

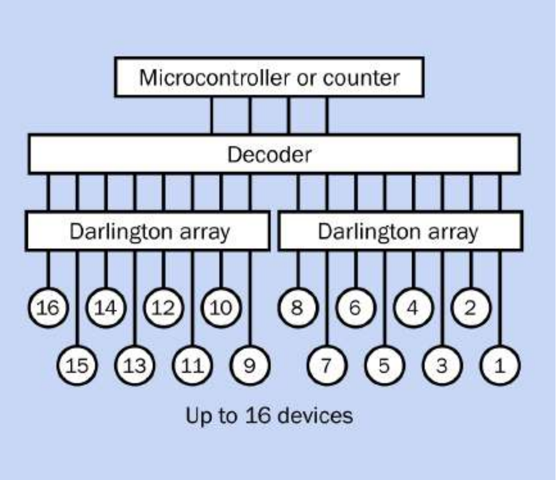
Microcontroller or counter
Decoder
Darlington array
Darlington array
16
14
12
10
8
6
4
2
15
13
11
9
7
5
3
1
Up to 16 devices

Figure 15-4. *Four outputs from a binary counter or microcontroller can be used by a decoder to activate one of up to 16 output devices.*

A photograph of a 74HC4514 decoder chip appears in Figure 15-6.

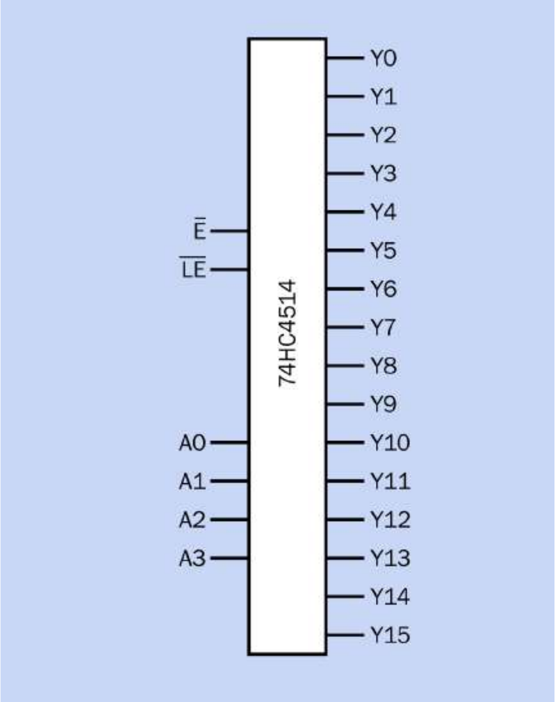
74HC4514
$\overline{E}$
$\overline{LE}$
A0
A1
A2
A3
Y0
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
Y10
Y11
Y12
Y13
Y14
Y15

Figure 15-5. *While no specific schematic symbol exists for a decoder chip, this style is commonly used. Shown here is a 4-to-16 decoder.*

93A1XQ8
CD74HC4514EN

Figure 15-6. *The 24-pin 74HC4514 decoder chip processes a 4-bit input and represents it by making one of its 16 output pins active-high.*

How It Works

A decoder contains logic gates, each of which is wired to respond to a unique binary pattern of inputs. (In the case of a seven-segment decoder, the internal logic is more complicated.) Figure 15-7 shows the logic of a 2-to-4 decoder. The darker blue area contains the components inside the chip. The external switches are included only to clarify the function of the decoder. An open switch is imagined to provide a low logic input, while each closed switch provides a high logic input.

Unlike ripple counters, where propagation delays can reduce the overall response time of the component, decoders function within two or three nanoseconds.

Variants

Decoder variants have not proliferated with time, and relatively few are available. Most are 3-to-8, 4-to-16, and binary-coded-decimal types.

The 7447 and 74LS47 are seven-segment decoders that have an open-collector output capable of driving a 7-segment display directly. The 7448 is similar but also contains built-in resistors and a capability to blank out leading zeros in a display. However, some suppliers now list the 74LS48 as obsolete. It may be still available from old stock, but should not be specified in new circuits.

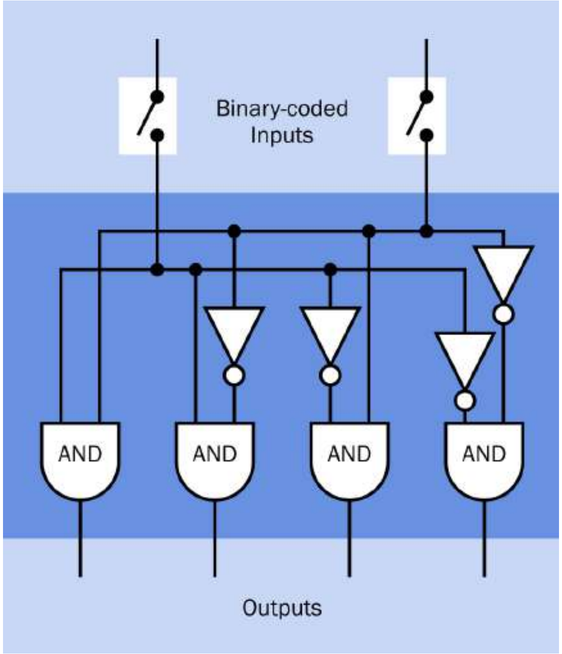

Figure 15-7. *A simplified simulation of the logic in a decoder. An actual chip would have an Enable line to activate the output. The dark blue rectangle indicates the space inside the chip.*

Although 74LS47 is still being manufactured, and is available in surface-mount as well as through-hole format, a version is not available in the widely used HC family of 74xx chips. Care must be taken to satisfy the input voltage requirements of the 74LS47 when driving it with 74HCxx chips.

Values

As is the case with other logic chips, most decoders in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See the section on logic gates in Chapter 10 for a discussion of acceptable high and low input states. On the output side, the 4000 series chips are able to source or sink less than 1mA at 5VDC, but the 74HCxx series can manage around 20mA.

How to Use It

The original applications for decoders in computer circuits have become uncommon, but the chips can still be useful in small appliances and gadgets where multiple outputs are controlled by a **counter** or microcontroller.

Although 16 is usually the maximum number of outputs, some chips are designed to allow expansion. The 74x138 (where a chip family identifier such as LS or HC can be substituted for the letter x) is a 3-to-8 decoder with two logic-low Enable pins and one logic-high Enable. If a value-8 binary line is applied to the low-enable of one chip and the high-enable of another, the first chip will be disabled when the line goes high to indicate that the binary number 1000 has been reached, and the second chip can continue upward from there by sharing the same three less-significant-bit inputs. As many as four chips can be chained in this way.

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates in "What Can Go Wrong" on page 105.

Glitches

Although a decoder typically functions faster than a ripple counter, it suffers the same tendency to introduce brief *glitches* in its output. These are momentary invalid states which occur while processes inside the chip that are slightly slower are catching up with other processes that reach completion slightly faster. A brief *settling time* is necessary to ensure that the output is stable and valid. This will be irrelevant when powering a device such as an **LED indicator**, which will not display such brief transients. The problem may be more important if the output from the decoder is used as an input to other logic chips.

If the input to a decoder is derived from a ripple counter, the input may also contain glitches, which can cause erroneous outputs from the decoder. It is better to use a synchronous counter on the input side of a decoder.

Unhelpful Classification

Online parts suppliers tend to list decoders under the same category heading as encoders, multiplexers, and demultiplexers, making it difficult to find what you want. Under this broad subject heading (which will include thousands of chips), if you search by selecting the number of inputs relative to the number of outputs that you have in mind, this will narrow the search considerably.

Active-Low and Active-High

Chips with identical appearance and similar part numbers may have outputs that are either active-low or active-high. Some may offer a latch-enable pin, while others have enable pins that must be pulled low or forced high to produce an output. Accidental chip substitution is a common cause of confusion.

multiplexer

16

May be abbreviated as a *mux* (this term is sometimes printed all in caps), and may be referred to alternatively as a *data selector*. Some sources maintain that a multiplexer has no more than two channels, whereas a data selector has more, but there is no consensus on this, and datasheets continue to use the term "multiplexer" predominantly.

Analog multiplexers are usually bidirectional, and thus will function equally well as *demultiplexers*. Consequently, this encyclopedia does not contain a separate entry for demultiplexers.

OTHER RELATED COMPONENTS

- **encoder** (see Chapter 14)
- **decoder** (see Chapter 15)

What It Does

A multiplexer can select one of two or more input pins, and connect it internally with an output pin. Although it is an entirely solid-state device, it behaves as if it contains a **rotary switch** in series with a SPST switch, as shown in Figure 16-1. A binary code applied to one or more Select pins chooses the input, and an Enable pin establishes the connection with the output. The Select and Enable functions are processed via an internal section referred to as a *decoder*, not to be confused with a **decoder** chip, which has its own entry in this encyclopedia.

All multiplexers are digitally controlled devices, but may be described as either *digital* or *analog* depending how they process the input signal. A digital multiplexer creates an output that is adjusted to logic-high or logic-low within the limits of its logic family. An analog multiplexer does not impose any processing on the voltage, and passes along any fluctuations. Thus, it can be used with alternating current.

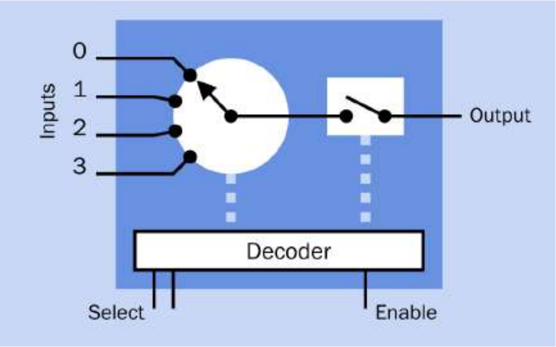

Figure 16-1. *A multiplexer functions as if it contains a rotary switch. The switch position is determined by a binary number applied to external Select pins. The internal connection is completed by applying a signal to an Enable pin.*

Because an analog multiplexer merely switches a flow of current, it can be *bidirectional*; in other words, it can function as a *demultiplexer*, in which case the input is applied to the pole of the (imaginary) internal switch and outputs are taken from the terminals.

Differential Multiplexer

A *differential* multiplexer contains multiple switches that are differentiated from one another (i.e., they are electrically isolated, although they are controlled by the same set of select pins). A differential multiplexer is conceptually similar to a rotary switch with two or more decks controlled by a single shaft. See Figure 16-2.

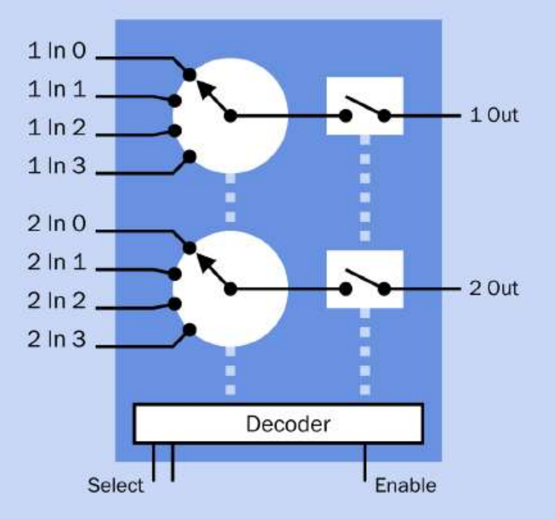

Figure 16-2. *A differential multiplexer contains two or more electronic switches that are differentiated from one another, similarly to the decks on a rotary switch. Although the channels into each switch are typically numbered from 0 upward, the switches are numbered from 1 upward.*

A bidirectional dual 4-channel differential analog multiplexer is shown in Figure 16-3.

Modern multiplexers are often found switching high-frequency data streams in audio, telecommunications, or video applications.

Similar Devices

The similarities and differences between multiplexer, demultiplexer, encoder, and decoder can cause confusion:

- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.
- An *analog* multiplexer may allow its inputs and outputs to be reversed, allowing it to become a **demultiplexer**, connecting a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

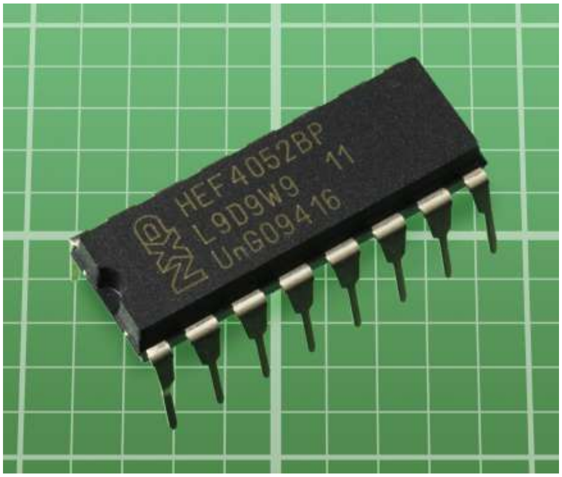

Figure 16-3. *This CMOS chip contains two four-channel differential analog multiplexers.*

- In an **encoder**, an active logic state is applied to one of four or more input pins, while the rest remain in an inactive logic state. The input pin number is converted to a binary code which is expressed as a pattern of logic states on two or more output pins.
- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive

logic state. A digital multiplexer does not allow reversal of its inputs and outputs, but a decoder functions as if it were a digital demultiplexer.

How It Works

The multiple inputs to a multiplexer are referred to as *channels*. Almost always, the number of channels is 1, 2, 4, 8, or 16. A 1-channel component is only capable of "on" or "off" modes and functions similarly to a SPST switch.

If there are more than two channels, a binary number will determine which channel is connected internally. The number of channels is usually the maximum that can be identified by the number of select pins, so that 2 pins will control 4 channels, 3 pins will control 8 channels, and 4 pins (the usual maximum) will control 16 channels.

In multiplexers with three or more channels, an enable pin is usually still present to activate or deactivate all the channels simultaneously. The enable feature may be described alternatively as a *strobe*, or may have an inverse function as an *inhibit* pin.

Although a rotary switch is helpful in conceptualizing the function of a multiplexer, a more common representation (sometimes in datasheets) is an array of SPST switches, each of which can be opened or closed by the decoder circuit. A typical example, depicting a dual differential multiplexer, is shown in Figure 16-4. Note that the internal decoder can only close one switch in each channel at a time.

The switch analogy is appropriate in that when an output from a multiplexer is not connected internally (i.e., its switch is "open") it is effectively an open circuit. However, some multiplexers contain *pullup resistors* to give each output a defined state. This can be an important factor in determining whether the multiplexer is suitable for a particular application.

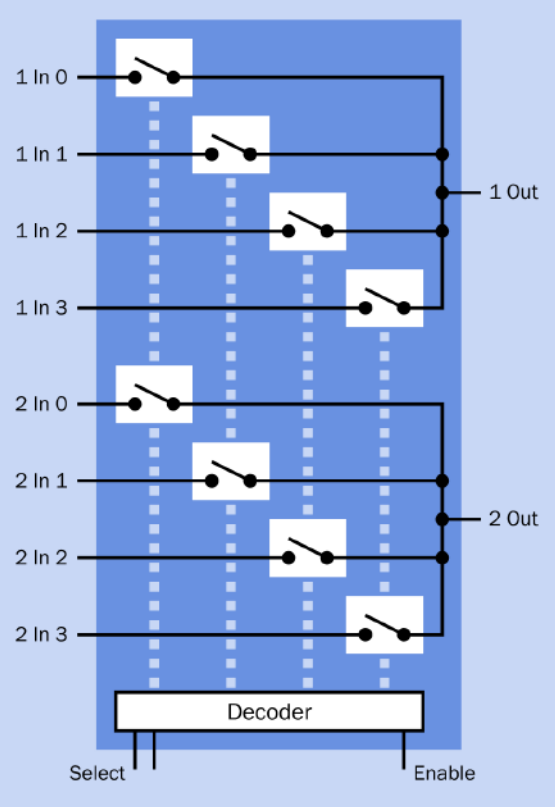

Figure 16-4. *The internal function of a dual multiplexer is commonly represented as a network of SPST switches, each of which is controlled by decoder logic.*

A digital multiplexer actually contains a network of logic gates, shown in simplified form in Figure 16-5.

A demultiplexer has internal logic shown in simplifier form in Figure 16-6.

Schematic Symbol

In a schematic, a multiplexer and demultiplexer may be represented by a trapezoid with its longer vertical side oriented toward the larger number of connections. This is shown in Figure 16-7. However, this symbol is falling into disuse.

More often, as is the case with most logic components, a multiplexer or demultiplexer is represented by a rectangle with inputs on the left and outputs on the right, as shown in Figure 16-8. The distinction between inputs and outputs is problematic, however, in an analog multiplexer which will allow data flow to be reversed.

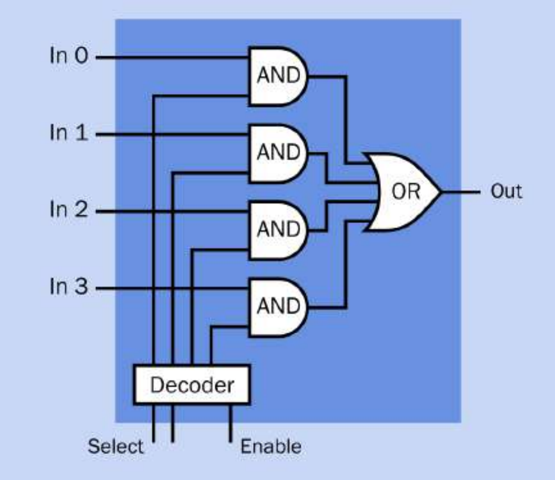
In 0
In 1
In 2
In 3
AND
AND
AND
AND
OR
Out
Decoder
Select
Enable

Figure 16-5. *A simplified representation of the logic gates in a digital multiplexer.*

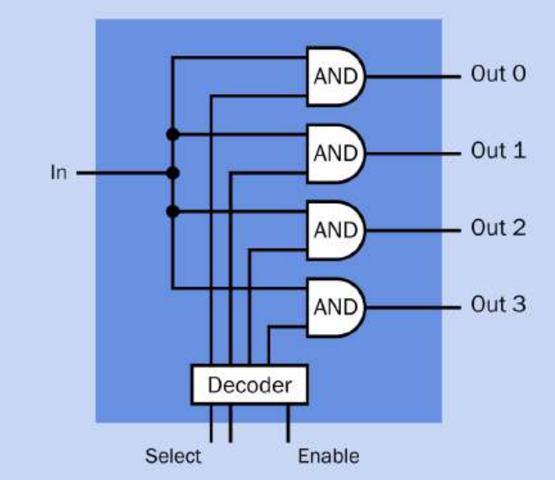
In
AND
AND
AND
AND
Out 0
Out 1
Out 2
Out 3
Decoder
Select
Enable

Figure 16-6. *A simplifier representation of the logic gates in a digital demultiplexer.*

Pin Identifiers

The lack of standardization in the identification of pin functions is perhaps more extreme in the case of multiplexers than for other types of logic chips.

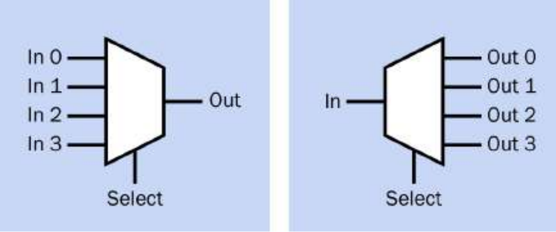
In 0
In 1
In 2
In 3
Out
Select
In
Out 0
Out 1
Out 2
Out 3
Select

Figure 16-7. *The traditional symbol for a multiplexer (left) and demultiplexer (right). The trapezoid is oriented with its longer vertical side facing the larger number of connections. This symbol is falling into disuse.*

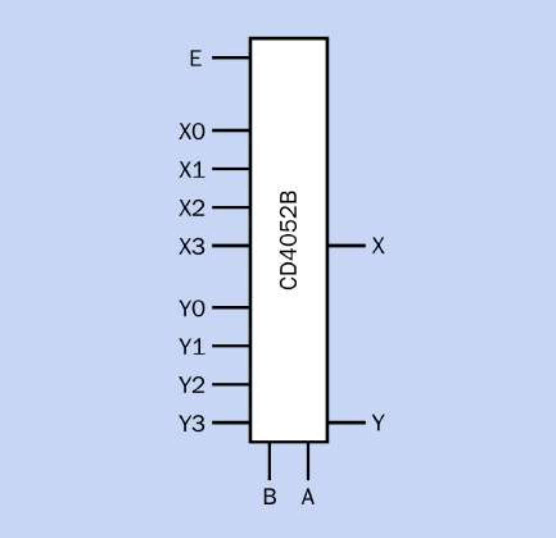
E
X0
X1
X2
X3
Y0
Y1
Y2
Y3
CD4052B
X
Y
B
A

Figure 16-8. *A simple rectangle is most often used as a schematic symbol for a multiplexer, but the abbreviations assigned to pin functions are not standardized. See text for details.*

An output enable pin will be shown as E or EN, or occasionally OE. It may alternatively be described as an inhibit pin, labeled INH, or sometimes will be called a strobe. The function is the same in each case: one of its logic states will enable the internal switches, while its other logic state will prevent any internal switches from closing.

Switch inputs may be labeled S0, S1, S2… or X0, X1, X2… or may simply be numbered, almost always counting up from 0. Where two or more sets of switches coexist in one package, each set of inputs may be distinguished from the others by preceding each identifier with a numeral or letter to designate the switch, as in 1S0, 1S1, 1S2… or 1X0, 1X1, 1X2… (Switches are generally numbered from 1 upward, even though their inputs are numbered from 0 upward.) Outputs may be identified using the same coding scheme as inputs, bearing in mind that the inputs and outputs of an analog multiplexer usually are interchangeable. Some manufacturers, however, prefer to identify each multiplexer output by preceding it with letter Y. Alternatively, Z1, Z2, Z3… may identify the outputs from switches 1, 2, 3… Fortunately, datasheets usually include some kind of key to this grab-bag of abbreviations.

Control pins are often identified as A, B, C… with letter A representing the least significant bit in the binary number that is applied to the pins.

Voltages can be confusing in multiplexers. Components intended for use with digital inputs are straightforward enough, as the supply voltage will be identified as V_{CC} and is typically 5VDC for through-hole packages (often lower for surface-mount), while negative ground is assumed to be 0VDC. However, where a multiplexer may be used with AC inputs in which the voltage varies above and below 0V, supply voltages above and below 0VDC are also possible—such as +7.5VDC and −7.5VDC, to take a random example. Three power-supply pins may be provided for this purpose. The positive supply will usually be identified as V_{DD} (the D refers to the Drain in the internal MOSFETs). A V_{EE} pin may be at 0VDC or at a negative value equal and opposite to V_{DD}. The E in this abbreviation is derived from Emitter voltage, even though the component may not contain a bipolar transistor with an emitter. Customarily, a V_{SS} pin (the S being derived from the Source in the internal MOSFETs) will be at 0VDC, and other voltages will be measured above and below this baseline. This ground pin may alternatively be labeled GND.

As is customary in logic chips, low-active control pins will have a bar printed above their identifiers, or an apostrophe will be placed after an identifier if the font does not permit printing the bar. Alternatively, low-active pins may be represented by showing a small circle, properly referred to as a *bubble*, at the input or output point of the symbol for the multiplexer. Note that analog inputs and outputs are neither high-active nor low-active; they merely pass voltages through.

Variants

Most multiplexers are "break before make" devices, where one input is disconnected before the next input is connected. However, some exceptions exist, and datasheets should be checked for this. It can be a significant issue, because make-before-break switching will briefly connect external devices with each other, through the chip.

Many multiplexers can tolerate control voltages above the usual high value in a logic circuit—as high as 15VDC in some cases. The voltage that is switched by the multiplexer may be the same as the control voltage, or may be higher.

Some analog multiplexers have *overvoltage protection* that allows them to withstand input voltages that are twice or three times the recommended maximum.

Datasheets may mention "internal address decoding," meaning that the binary number input, specifying a channel to be switched, is decoded inside the chip. In fact, virtually all multiplexers now have on-chip address decoding, and this feature should be assumed to exist, regardless of whether it is mentioned.

Values

The voltage to be switched will usually be referred to as the *input voltage*, V_{IN}.

An analog multiplexer should not be subjected to current exceeding the value that it is designed to switch. This is known as the maximum *channel current*. A typical value would be 10mA, although many modern surface-mount components are designed for currents in the microamp range.

The *on-resistance* is the resistance imposed by the analog multiplexer on the signal flowing through it. While modern, specialized analog multiplexers may have an on-resistance as low as 5Ω, these are relatively unusual. An on-resistance of 100Ω to 200Ω is more common. This value will vary within a component depending on the power supply voltage and the voltage being switched. It will increase slightly as V_{IN} deviates above (or below) 0V, will increase substantially for lower values of supply voltage, and will increase significantly with temperature.

The curves in Figure 16-9 show on-resistance of an analog multiplexer varying with input voltage, with three different power supplies: plus-and-minus 2.5VDC (described in the graph as a "spread" of 5VDC), plus-and-minus 5VDC (a "spread" of 10VDC), and plus-and-minus 7.5VDC (a "spread" of 15VDC). These curves were derived from a datasheet for the MC14067B analog multiplexer; curves for other chips will be different, although the basic principles remain the same.

Switching time is an important consideration in high-speed applications. The "on" and "off" times specified in a datasheet (often as t_{ON} and t_{OFF}) are a function of the propagation delay from the control input to the toggling of the switch, and are generally measured from the halfway point of the rising or falling edge of the control input, to the 90% point of the output signal level.

Leakage current is the small amount of current (often measured in picoamperes) that the solid-state switch will pass when it is in its "off" state. This should be insignificant except when very high-impedance loads are used.

Separate switches inside a multiplexer may have characteristics that differ slightly from one another. Differences in on-resistance between adjacent switches can be important when switching parallel analog signals. A datasheet should mention the extent to which switches have matched characteristics, and may define the maximum deviation from one another using the abbreviation R_{ON} even though this same term may be used, confusingly, to denote the on-resistance of each individual switch.

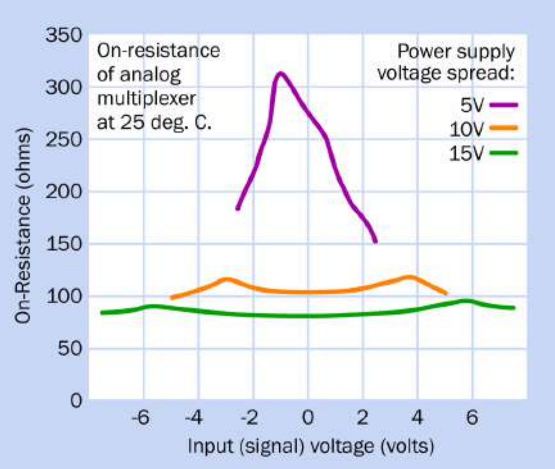

Figure 16-9. *Variations in on-resistance in an analog multiplexer. Each voltage "spread" is the difference between positive supply voltage and an equal-and-opposite negative ground voltage. Thus a "spread" of 10VDC means plus and minus voltages of 5VDC. (Curves derived from On Semiconductor datasheet for MC14067B analog multiplexer.)*

How to Use It

A multiplexer may be used as a simple switch to choose one of multiple inputs, such as a choice of input jacks on a stereo system. A dual differential multiplexer is useful in this application, as it can use a single select signal to switch two signal paths simultaneously.

A multiplexer can also be used as a digital volume control by switching an audio signal among a variety of resistances, similar to a **digital potentiometer**. In this application, the possible presence of pullup resistors inside the multiplexer must be considered.

Where a microcontroller must monitor a large number of inputs (for example, a range of temperature sensors or motion sensors), a multiplexer can reduce the number of input pins required. Its data-select pins will be cycled through all the possible binary states by the microcontroller, to select each data input in turn, while its single-wire output will carry the analog data to a separate pin on the microcontroller which performs an analog-digital conversion.

Conversely, a demultiplexer (i.e., an analog multiplexer such as the 4067B chip which can be used in demultiplexer mode) can be used by a microcontroller to switch multiple components on and off. Four outputs from the microcontroller can connect with the control pins of a 16-channel demultiplexer, counting from binary 0000 through binary 1111 to select output pins 0 through 15. After selecting each pin, the microcontroller can send a high or low pulse through it. The process then repeats. (A **decoder** can be used in the same way.)

Other Application Notes

Multiplexers may be *cascaded* to increase the inputs-to-outputs ratio.

Modern multiplexers are found on computer boards where they choose among video output ports, or as PCI express channel switches.

A multiplexer may be used as a parallel-to-serial converter, as it samples multiple channels and converts them into a serial data stream.

In telecommunications, a multiplexer can sample voice signals from multiple separate inputs and combine them into a digital stream that can be transmitted at a faster bit rate over a single channel. However, this application goes far beyond the simple uses for multiplexers described here.

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates (see "What Can Go Wrong" on page 105).

Pullup Resistors

While they are often necessary to prevent connections from floating, pullup resistors built into a multiplexer may have unexpected consequences if the user is unaware of them.

Break Before Make

For most applications, it is desirable for each internal solid-state switch to break one connection before making a new one. This avoids the possibility of separate external components being briefly connected with each other through the multiplexer. Datasheets should be checked to verify that a multiplexer functions in break-before-make mode. If it doesn't, the enable pin can be used momentarily to disable all connections before a new connection is established.

Signal Distortion

Where a multiplexer is passing analog signals, signal distortion can result if the on-resistance of multiple internal switches varies significantly at different voltages. A datasheet for an analog multiplexer should usually include a graph showing on-resistance over the full signal range. The flatter the graph is, the less distortion the component will create. This is often described in datasheets as R_{ON} Flatness.

Limits of CMOS Switching

Although most multiplexers are built around CMOS transistors, their switching speed may be insufficient for video signals, and their on-resistance may vary enough to introduce distortion. Multiplexers are available with complementary bipolar switching for very high-speed applications. They impose some penalties in cost and power consumption.

Transients

Switch capacitance inside a multiplexer can cause transients in the output when the switch changes state. An allowance for *settling time* may be necessary. This will be additional to the switching speed claimed by the datasheet.

LCD

17

The full term *liquid-crystal display* is seldom used. Its acronym, **LCD**, is much more common. Sometimes the redundant combination **LCD display** is found. All three terms refer to the same device. In this encyclopedia, the first two words in *liquid-crystal display* are hyphenated because they are an adjectival phrase. Other sources often omit the hyphen.

The acronym **LED** (for **light-emitting diode**) is easily confused with **LCD**. While both devices display information, their mode of action is completely different.

OTHER RELATED COMPONENTS

- **LED display** (see Chapter 24)

What It Does

An LCD presents information on a small display panel or screen by using one or more segments that change their appearance in response to an AC voltage. The display may contain alphanumeric characters and/or symbols, icons, dots, or pixels in a *bitmap*.

Because of its very low power consumption, a basic monochrome LCD is often used to display numerals in battery-powered devices such as digital watches and calculators. A small liquid-crystal display of this type is shown in Figure 17-1.

Color-enabled, backlit LCDs are now frequently used in almost all forms of video displays, including those in cellular telephones, computer monitors, game-playing devices, TV screens, and aircraft cockpit displays.

How It Works

Light consists of electromagnetic waves that possess an electric field and a magnetic field. The fields are perpendicular to each other and to the direction in which the light is traveling, but the field polarities are randomly mixed in most visible radiation. This type of light is referred to as *incoherent*.

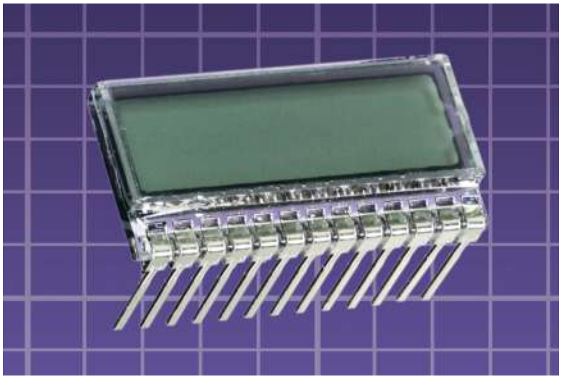

Figure 17-1. *A small, basic monochrome LCD.*

Figure 17-2 shows a simplified view of an LCD that uses a backlight. Incoherent light emerges from the backlight panel (A) and enters a vertical polarizing filter (B) that limits the electric field vector. The polarized light then enters a liquid crystal (C) which is a liquid composed of molecules organized in a regular helical structure that rotates the polarity by 90 degrees when no voltage is applied to it. The light now passes through a horizontal polarizing filter (D) and is visible to the user.

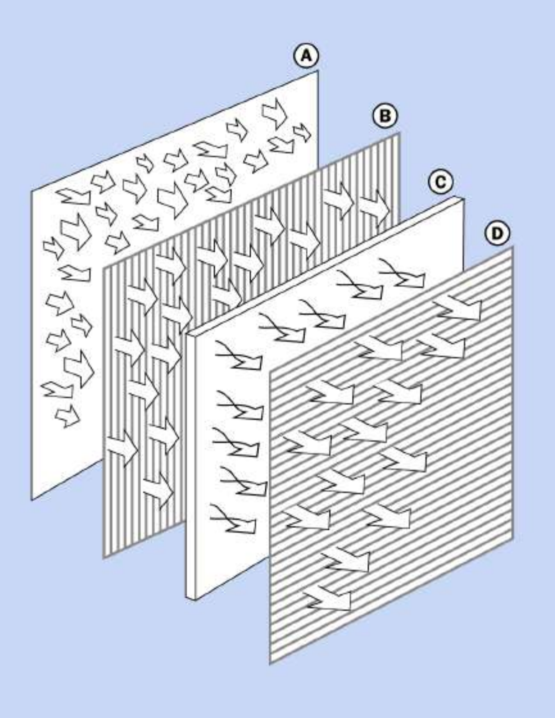

Figure 17-2. *The combination of two polarizers and a liquid crystal appears transparent when voltage is not applied. See text for details.*

- A liquid crystal itself does not emit light. It can only modify light that passes through it.

Figure 17-3 shows what happens when voltage is applied to the liquid crystal via transparent electrodes (not included in the figure). The molecules reorganize themselves in response to the electric potential and allow light to pass without changing its polarity. Consequently, the vertically polarized light is now blocked by the front, horizontally polarized filter, and the display becomes dark.

A liquid crystal contains ionic compounds that will be attracted to the electrodes if a DC voltage is applied for a significant period of time. This can degrade the display permanently. Therefore, AC voltage must be used. An AC frequency of 50Hz to 100Hz is common.

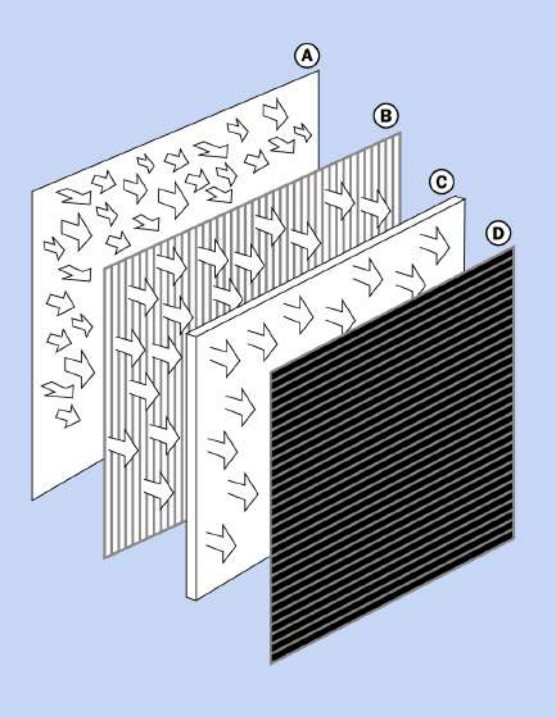

Figure 17-3. *The LCD appears dark when voltage is applied. See text for details.*

Variants

A *transmissive LCD* requires a backlight to be visible, and is the type illustrated in Figure 17-2. In its simplest form, it is a monochrome device, but is often enhanced to display full color by adding red, green, and blue filters. Alternatively, instead of a white backlight, an array of pixel-sized red, green, and blue **LEDs** may be used, in which case filters are unnecessary.

Backlit color LCDs have displaced *cathode-ray tubes*, which used to be the default system in almost all video monitors and TVs. LCDs are not only cheaper but can be fabricated in larger sizes. They do not suffer from *burn in*, where a persistent unchanging image creates a permanent scar in the phosphors on the inside of a tube. However, large LCDs may suffer from dead pixels or stuck pixels as manufacturing defects. Different manufacturers and vendors have varying policies regarding the maximum acceptable number of pixel defects.

In a *reflective LCD*, the structure is basically the same as that shown in Figure 17-2 except that a reflective surface is substituted for the backlight. Ambient light enters from the front of the display, and is either blocked by the liquid crystal in combination with the polarizing filters, or is allowed to reach the reflective surface at the rear, from which it reflects back through the liquid crystal to the eye of the user. This type of display is very easily readable in a bright environment, but will be difficult to see in dim conditions and will be invisible in darkness. Therefore, it may be augmented with a user-activated light source mounted at the side of the display.

A *transreflective LCD* contains a translucent rear polarizer that will reflect some ambient light, and is also transparent to enable a backlight. While this type of LCD is not as bright as a reflective LCD and has less contrast, it is more versatile and can be more energy efficient, as the backlight can be switched off automatically when ambient light is bright enough to make the display visible.

Active and Passive Types

An *active matrix* LCD adds a matrix of thin-film transistors to the basic liquid-crystal array, to store the state of each segment or pixel actively while the energizing AC voltage transitions from positive to negative. This enables a brighter, sharper display as *crosstalk* between adjacent pixels is reduced. Because *thin-film transistors* are used, this is often described as a *TFT* display; but the term is interchangable with *active matrix*.

A *passive matrix* LCD is cheaper to fabricate but responds sluggishly in large displays and is not so well suited to fine gradations in intensity. This type of component is used primarily in simple monochrome displays lacking intermediate shades of gray.

Crystal Types

Twisted Nematic (TN) are the cheapest, simplest type of LCD, allowing only a small viewing angle and average contrast. The appearance is limited to black on gray. The response rate is relatively slow.

Super Twisted Nematic (STN) displays were developed in the 1980s for passive LCDs, enabling better detail, wider view angle, and a faster response. The natural appearance is dark violet or black on green, or dark blue on silver-gray.

Film-compensated Super Twisted Nematic (FSTN) uses an extra coating of film that enables a pure black on white display.

Double Super Twisted Nematic provides further enhancement of contrast and response times, and automatic contrast compensation in response to ambient temperature. The appearance is black on white. This display requires backlighting.

Color Super Twisted Nematic (CSTN) is an STN display with filters added for full color reproduction.

Seven-Segment Displays

The earliest monochrome LCDs in devices such as watches and calculators used seven segments to display each numeral from 0 through 9. This type of LCD is still used in low-cost applications. A separate control line, or electrode, connects to each segment, while a *backplane* is shared by all the segments, connecting with a *common* pin to complete the circuit.

Figure 17-4 shows a typical seven-segment display. The lowercase letters *a* through *g* that identify each segment are universally used in datasheets. The decimal point, customarily referred to as “dp,” may be omitted from some displays. The array of segments is slanted forward to enable more acceptable representation of the diagonal stroke in numeral 7.

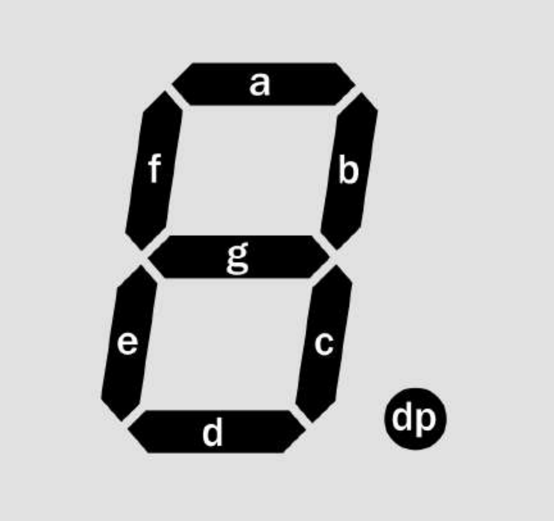

Figure 17-4. *Basic numeric display format for LCD numeric displays (the same layout is used with LEDs). To identify each segment, lowercase letters are universally used.*

Seven-segment displays are not elegant but are functional and are reasonably easy to read, as shown in Figure 17-5. Letters A, B, C, D, E, and F (displayed as A, b, c, d, E, F because of the restrictions imposed by the small number of segments) may be added to enable display of hexadecimal values.

In appliances such as microwave ovens, very basic text messages can be displayed to the user within the limitations of 7-segment displays, as suggested in Figure 17-6.

The advantage of this system is low cost, as 7-segment displays are cheap to fabricate, entail the fewest connections, and require minimal decoding to create each alphanumeric character. However, numbers 0, 1, and 5 cannot be distinguished from letters O, I, and S, while letters containing diagonal strokes, such as K, M, N, W, X, and Z, cannot be displayed at all.

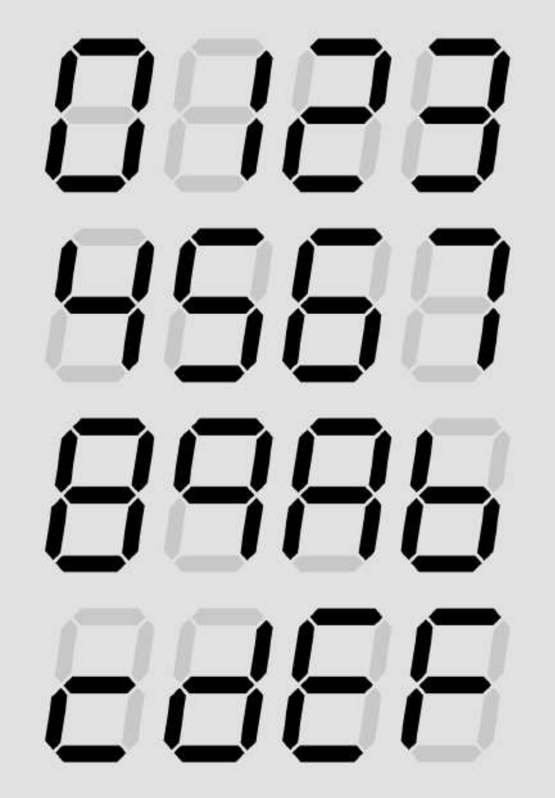

Figure 17-5. *Numerals and the first six letters of the alphabet created with a 7-segment display.*

Additional Segments

Alphanumeric LCDs were developed using 14 or 16 segments to enable better representation of letters of the alphabet. Sometimes these displays were slanted forward, like the 7-segment displays, perhaps because the style had become familiar, even though the addition of diagonal segments made it unnecessary. In other cases, the 14 or 16 segments were arrayed in a rectangle. See Figure 17-7.

The same words represented in Figure 17-6 are shown in Figure 17-8, using 16-segment LCDs. Clearly, the advantage gained by enabling diagonal strokes entailed the disadvantage of larger gaps in the letters, making made them ugly and difficult to read.

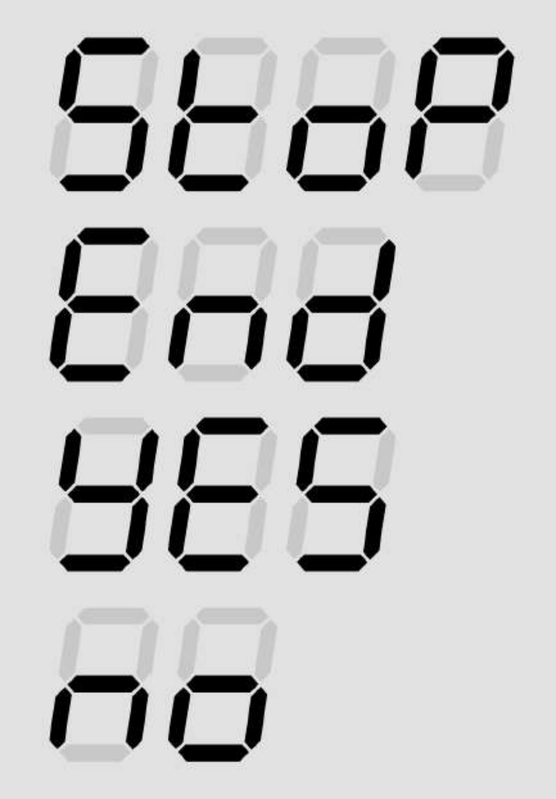

Figure 17-6. *Basic text messages can be generated with 7-segment displays, although they cannot contain alphabetical letters that use diagonal strokes.*

A full character set using 16-segment LCDs is shown in Figure 17-9. This conforms partially with the ASCII coding system, in which each character has an identifying numeric code ranging from 20 hexadecimal for a letter-space to 7A hexadecimal for letter z (although this character set does not attempt to represent lowercase letters differently from uppercase). The ASCII acronym stands for *American Standard Code for Information Interchange*.

Because backlit LCDs had become common by the time 16-segment displays were introduced, the characters were often displayed in light-on-dark or "negative" format, as suggested in this figure. LEDs, of course, have always used the light-on-dark format, as an LED is a light-emitting component.

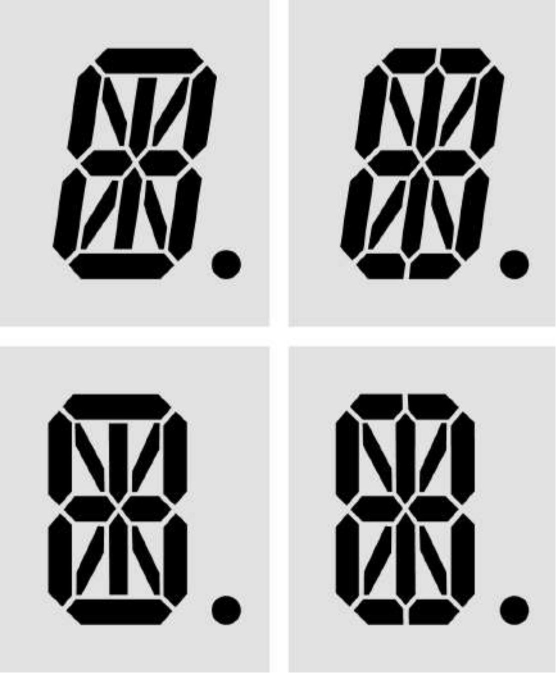

Figure 17-7. *LCDs using 14 segments (left) and 16 segments (right) were introduced to represent a full alphabet in addition to numerals. Sometimes these displays were slanted forward, like the previous 7-segment type, even though this was no longer necessary to represent the number 7.*

Dot-Matrix Displays

The 16-segment displays were never widely popular, and the declining cost of microprocessors, LCD fabrication, and ROM storage made it economic to produce displays using the more easily legible 5x7 dot-matrix alphabet that had been common among early microcomputers. Figure 17-10 shows a dot-matrix character set that is typical of many LCDs.

Because the original ASCII codes were not standardized below 20 hexadecimal or above 7A hexadecimal, manufacturers have represented a variety of foreign-language characters, Greek letters, Japanese characters, accented letters, or symbols using codes 00 through 1F and 7B through FF. The lower codes are often left blank, allowing user installation of custom symbols. Codes 00 through 0F are often reserved for control functions, such as a command to start a new line of text. There is no standardization in this area, and the user must examine a datasheet for guidance.

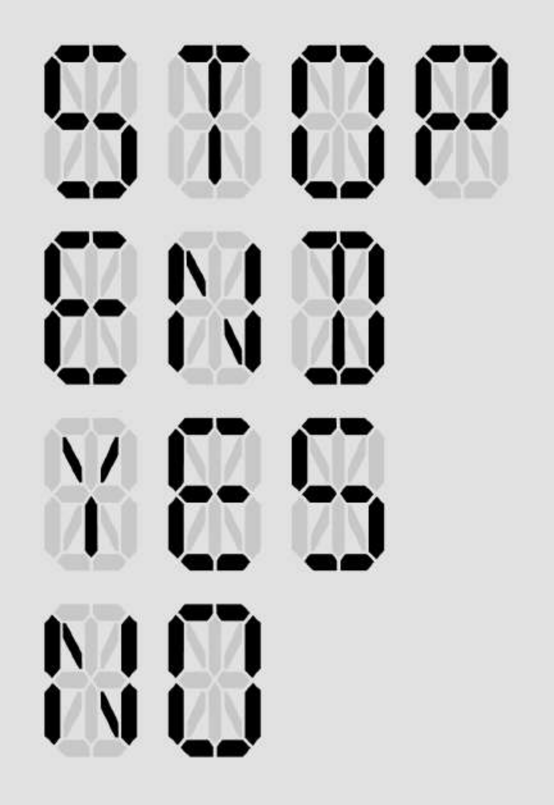

Figure 17-8. *The same text messages shown previously using 7-segment LCDs are shown here using 16-segment displays.*

Dot-matrix LCDs are usually packaged in arrays consisting of eight or more columns and two or more rows of characters. The number of columns is always stated before the number of rows, so that a typical 8 x 2 display contains eight alphanumeric characters in two horizontal rows. An array of characters is properly referred to as a *display module*, but may be described, confusingly, as a *display*, even though a single seven-segment LCD is itself a display. A 16x2 display module is shown from the front in Figure 17-11 and from the rear in Figure 17-12.

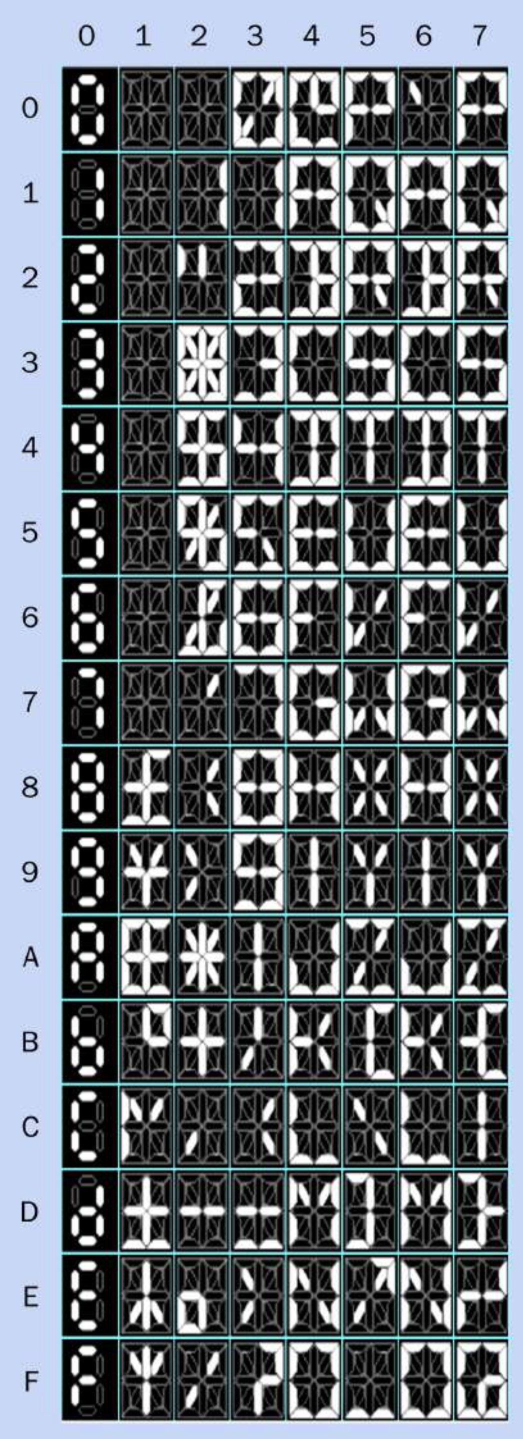

Figure 17-9. *A full character set using 16-segment LCDs.*

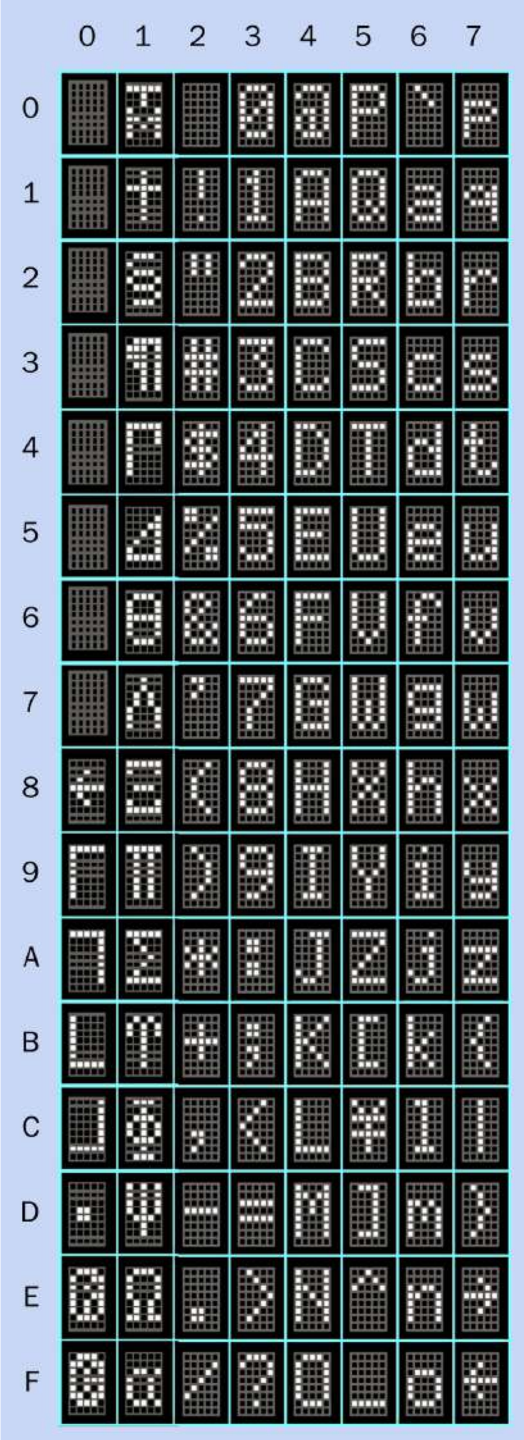

Figure 17-10. *A dot-matrix character set typical of LCDs capable of displaying a matrix of 5×7 dots.*

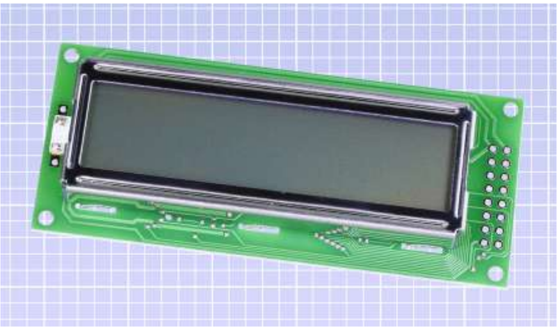

Figure 17-11. *A 16x2 LCD display module seen from the front.*

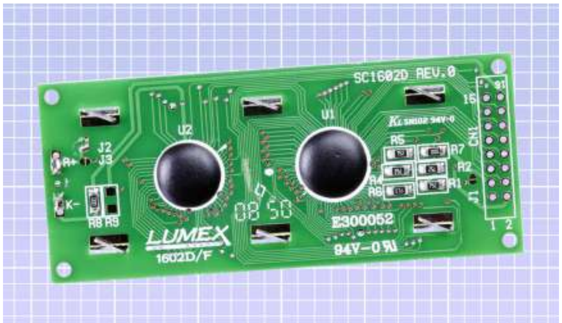

Figure 17-12. *The same 16x2 LCD display module from the previous figure, seen from the rear.*

Multiple-character display modules have been widely used in consumer electronics products such as audio components and automobiles where simple status messages and prompts are necessary—for example, to show the volume setting or broadcast frequency on a stereo receiver. Backlighting is almost always used.

Because the cost of small, full-color, high-resolution LCD screens has been driven down rapidly by the mass production of cellular phones, color displays are likely to displace monochrome dot-matrix LCD display modules in many applications. Similarly, touchscreens will tend to displace pushbuttons and tactile switches. Touchscreens are outside the scope of this encyclopedia.

Color

The addition of filters to create a full color display is shown in simplified form in Figure 17-13.

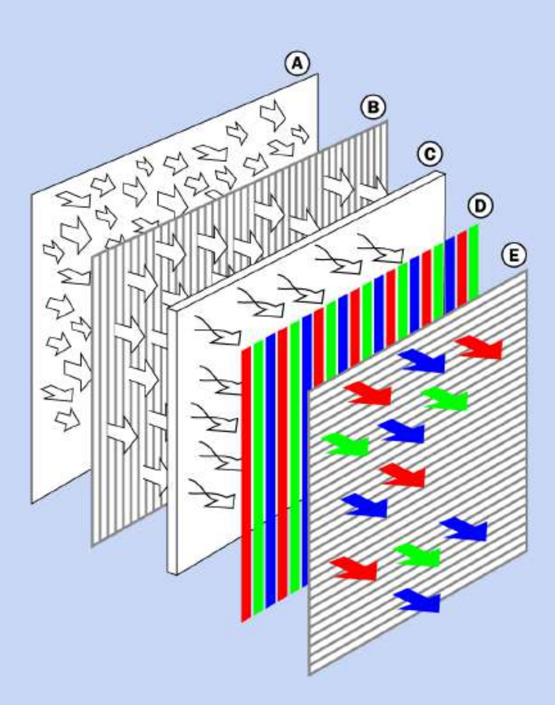

Figure 17-13. *The addition of red, green, and blue color filters, in conjunction with variable density liquid crystal pixels, enables an LCD full-color display.*

Red, green, and blue are almost always used as primary colors for transmitted light, because the combination of different intensities of these *RGB* primaries can create the appearance of many colors throughout the visible spectrum. They are said to be *additive* primaries, as they create brighter colors when they are combined. The principle is illustrated in Figure 17-14.

The use of the word "primaries" to refer to red, green, and blue can cause confusion, as full-color printed materials use a different set of *reflective* primaries, typically cyan, magenta, and yellow, often with the addition of black. In this *CMYK* system, additional layers of pigment will absorb, or subtract, more visible frequences. See Figure 17-15.

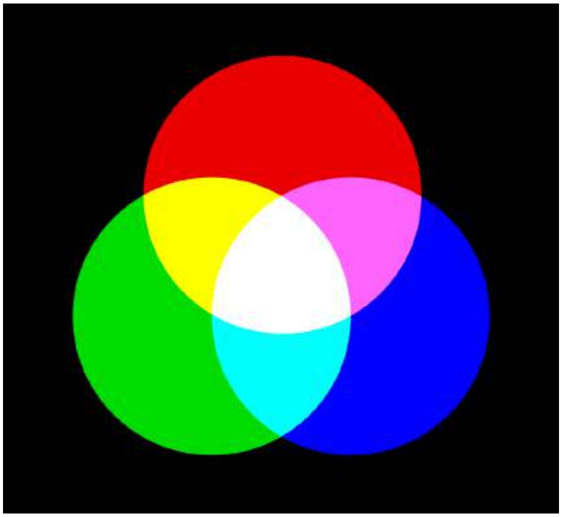

Figure 17-14. *When colors red, green, and blue are transmitted directly to the eye, pairs of these additive primaries create secondary colors cyan, magenta, and yellow. Combining all three additive primaries creates an approximation of white light. This can be verified by viewing a color monitor with a magnifying glass.*

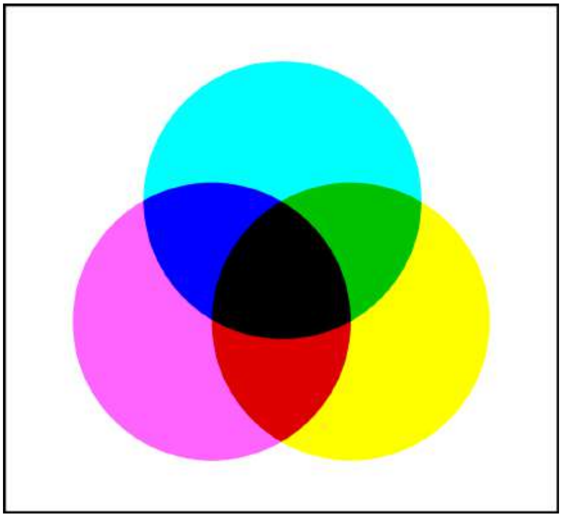

Figure 17-15. *When ink colors cyan, magenta, and yellow are superimposed on white paper and are viewed in white light, pairs of these subtractive primaries create secondary colors red, green, and blue. Overprinting all three subtractive primaries creates an approximation of black, limited by the reflective properties of available pigments. Black ink is usually added to provide additional contrast.*

The complete range of colors that can be created as a combination of primaries is known as the *gamut*. Many different RGB color standards have been developed, the two most widely used being *sRGB* (almost universal in web applications) and *Adobe 1998* (introduced by Adobe Systems for Photoshop, providing a wider gamut). None of the available systems for color reproduction comes close to creating the full gamut that can be perceived by the human eye.

Backlighting Options

For monochrome LCDs, **electroluminescent** backlighting may be used. It requires very low current, generates very little heat, and has a uniform output. However, its brightness is severely limited, and it requires an inverter that adds significantly to the current consumption.

For full-color LCDs, **fluorescent** lights were originally used. They have a long lifetime, generate little heat, and have low power consumption. However, they require a relatively high voltage, and do not work well at low temperatures. Early flat screens for laptop computers and desktop monitors used *cold-cathode fluorescent panels.*

Subsequently, white light-emitting diodes (**LED**s) were refined to the point where they generated a range of frequencies that was considered acceptable. Light from the LEDs passes through a diffuser to provide reasonably consistent illumination across the entire screen. LEDs are cheaper than fluorescent panels, and allow a thinner screen.

High-end video monitors use individual red, green, and blue LEDs instead of a white backlight. This eliminates the need for colored filters and produces a wider gamut. So-called *RGB* LCD monitors are more expensive but are preferred for professional applications in video and print media where accurate color reproduction is essential.

Zero-Power Displays

Some techniques exist to create LCDs that require power only to flip them to and fro between their transparent and opaque states. These are also known as *bistable* displays, but have not become as widely used. They are similar in concept to *e-ink* or *electronic paper* displays, but the principle of operation is different.

How to Use It

So long as an LCD consists of just one numeral, it can be driven by just one decoder chip that translates a binary-coded input into the outputs required to activate the appropriate segments of the LCD. The evolution of multi-digit displays, alphanumeric displays, dot-matrix displays, and graphical displays has complicated this situation.

Numeric Display Modules

An LCD consisting of a single digit is now a rare item, as few circuits require only one numeral for output. More commonly, two to eight numerals are mounted together in a small rectangular panel, three or four numerals being most common. A typical digital alarm clock uses a four-digit numeric display module, incorporating a colon and indicators showing AM/PM and alarm on/off. Other numeric display modules may include a minus sign.

Modules that are described as having 3.5 or 4.5 digits contain three full digits preceded by a numeral 1 composed of two segments. Thus, a 3-digit module can display numbers from 000 through 999, while a 3.5-digit display can display numbers from 000 through 1999, approximately doubling the range.

Numeric display modules of the type described here do not contain any decoder logic or drivers. An external device, such as a microcontroller, must contain a lookup table to translate a numeric value into outputs that will activate the appropriate segments in the numbers in a display, with or without decimal points and a minus sign. To avoid reinventing the wheel, a programmer may download code libraries for microcontrollers to drive commonly used numeric display modules. It is important to remember, though, that segments in monochrome LCDs must be activated by AC, typically a square wave with a frequency of 30Hz to 90Hz.

An alternative is to use a decoder chip such as the 4543B or 4056B, which receives a binary-coded decimal input (i.e., 0000 through 1001 binary, on four input puts) and translates it into an output on seven pins suitable for connection with the seven segments of a 7-segment display. The 4543B requires a square-wave input to its "phase" pin. The square-wave must also be applied simultaneously to the backplane of the LCD, often identified as the "common" pin on datasheets. Pinouts for the 4543B are shown in Figure 17-16.

The 4543B includes provision for "display blanking," which can be used to suppress leading zeros in a multidigit number. However, the lack of outputs to control a minus sign or decimal point limits the decoder to displaying positive integers.

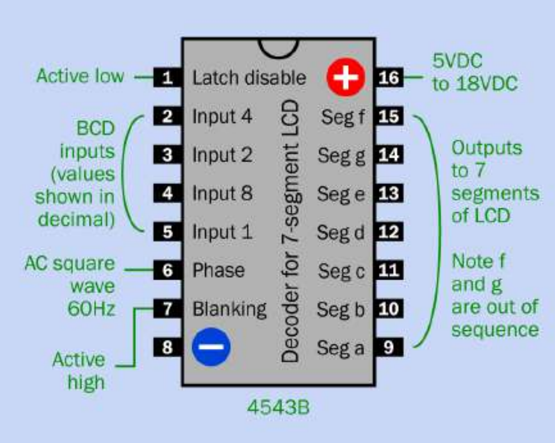

Figure 17-16. *Pinouts for the 4543B decoder chip, which is designed to drive a seven-segment numeric LCD.*

The power supply for a 4543B can range from 5VDC to 18VDC, but because the logic-high output voltage will be almost the same as that of the power supply, it must be chosen to match the power requirements of the LCD (very often 5VAC).(((

To drive a three-digit numeric display module, a separate decoder chip can be used to control each digit. The disadvantage of this system is that each decoder requires three inputs, so that a three-digit display will require nine outputs from the microcontroller.

To deal with this issue, it is common to *multiplex* a multi-digit display. This means that each output from the decoder is shared among the same segments of all the LCD numerals. Each LCD numeral is then activated in sequence by applying AC voltage to its common pin. Simultaneously, the decoder sends the data appropriate to that LCD. This process must be fast enough so that all the digits appear to be active simultaneously, and is best managed with a microcontroller. A simplified schematic is shown in Figure 17-17. It can be compared with a similar circuit to drive LED displays, shown in Figure 24-13.

Alphanumeric Display Module

Arrays of dot-matrix LCDs that can display alphabetical characters as well as numerals require preset character patterns (usually stored in ROM) and a command interpreter to process instructions that are embedded in the data stream. These capabilities are often built into the LCD module itself.

While there is no formal or de facto standard, the command set used by the Hitachi HD44780 controller is installed in many displays, and code libraries for this set are available for download from sites dedicated to the Arduino and other microcontrollers. Writing code from scratch to control all aspects of an alphanumeric display is not a trivial chore. The Hamtronix HDM08216L-3-L30S is a display that incorporates the HD44780.

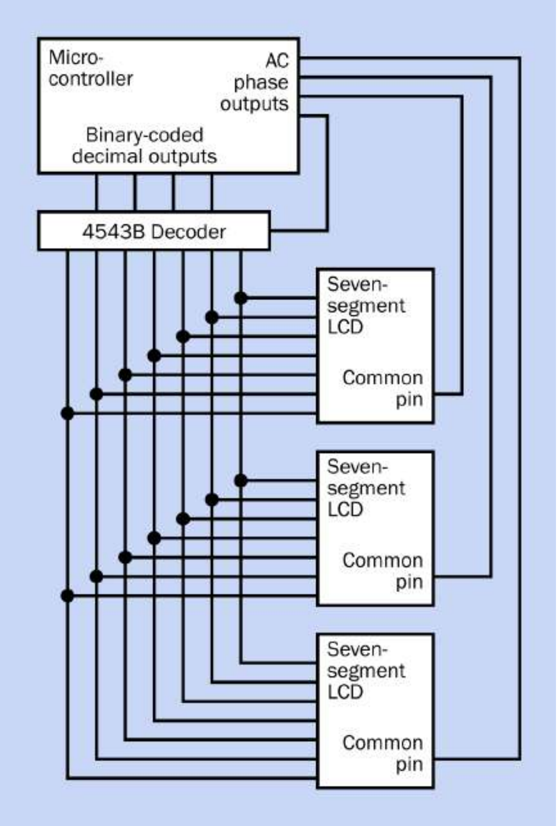

Figure 17-17. *When two or more numeric displays are multiplexed, a control device (typically, a microcontroller) activates each of them in turn via its backplane (common terminal) while sending appropriate data over a shared bus.*

Regardless of which standard is used, some features of alphanumeric display modules are almost universal:

- Register select pin. Tells the display whether the incoming data is an instruction, or a code identifying a displayable character.
- Read/write pin. Tells the display whether to receive characters from a microcontroller or send them to a microcontroller.
- Enable/disable pin.
- Character data input pins. There will be eight pins to receive the 8-bit ASCII code for each displayable character in parallel. Often there is an option to use only four of these pins, to reduce the number of microcontroller outputs necessary to drive the display. Where four pins are used, each 8-bit character is sent in two segments.
- LED backlight pin. Two may be provided, one connected to the anode(s) of the LED backlight, the other to the cathode(s).
- Reset pin.

Embedded instruction codes can be complex, including commands to reposition the cursor at a specific screen location, backspace-and-erase, scroll the display, and erase all characters on the screen. Codes may be included to adjust screen brightness and to switch the display between light-on-dark (negative) and dark-on-light (positive) characters.

Some display modules also have graphics capability, allowing the user to address any individual pixel on the screen.

Because of the lack of standardization in control codes, manufacturer's datasheets must be consulted to learn the usage of a particular alphanumeric display module. In addition to datasheets, online user forums are a valuable source of information regarding quirks and undocumented features.

What Can Go Wrong

Temperature Sensitivity

Liquid crystals vary in their tolerance for low and high temperatures, but generally speaking, a higher voltage may be necessary to create a sufficiently dense image at a low temperature. Conversely, a lower voltage may be necessary to avoid "ghosting" at a high temperature. An absolutely safe operating temperature range is likely to be 0 through 50 degrees Celsius, but check the manufacturer's datasheet for confirmation. Special-purpose LCDs are available for extreme temperatures.

Excessive Multiplexing

A twisted nematic display is likely to perform poorly if its duty cycle is greater than 1:4. In other words, more than four displays should not be multiplexed by the same controller.

DC Damage

An LCD can be damaged quickly and permanently if it is subjected to DC current. This can occur by accident if, for example, a timer chip is being used to generate the AC pulse stream, and the timer is accidentally disconnected, or has an incorrect connection in its RC network. Check timer output with a meter set to measure AC volts before allowing any connection to the common pin of an LCD.

Bad Communications Protocol

Many alphanumeric display modules do not use a formal communications protocol. Duplex serial or I2C connection may not be available. Care must be taken to allow pauses of a few milliseconds after execution of embedded commands, to give the display sufficient time to complete the instruction. This is especially likely where a command to clear all characters from the screen has to be executed. If garbage characters appear on the screen, incorrect data transfer speed or lack of pause times may be to blame.

Wiring Errors

This is often cited by manufacturers as the most common cause of failure to display characters correctly, or lack of any screen image at all.

incandescent lamp

18

The terms *incandescent light*, *incandescent bulb*, and *incandescent light bulb* are often used interchangeably with **incandescent lamp**. Because the term "lamp" seems to be most common, it is used here. A *panel-mounted indicator lamp* is considered to be an assembly containing an incandescent lamp.

A *carbon arc*, which generates light as a self-sustaining spark between two carbon electrodes, can be thought of as a form of incandescent lamp, but is now rare and is not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **LED area lighting** (see Chapter 23)
- **LED indicator** (see Chapter 22)
- **neon bulb** (see Chapter 19)
- **fluorescent light** (see Chapter 20)

What It Does

The term *incandescent* describes an object that emits visible light purely as a consequence of being hot. This principle is used in an **incandescent lamp** where a wire *filament* glows as a result of electric current passing through it and raising it to a high temperature. To prevent oxidation of the filament, it is contained within a sealed bulb or tube containing an inert gas under low pressure or (less often) a vacuum.

Because incandescent lamps are relatively inefficient, they are not considered a wise environmental choice for area lighting and have been prohibited for that purpose in some areas. However, small, low-voltage, panel-mount versions are still widely available. For a summary of advantages of miniature incandescent lamps relative to **light-emitting diodes** (LEDs) see "Relative Advantages" on page 179.

Schematic symbols representing an incandescent lamp are shown in Figure 18-1. The symbols are all functionally identical except that the one at bottom right is more likely to be used to represent small panel-mounted indicators.

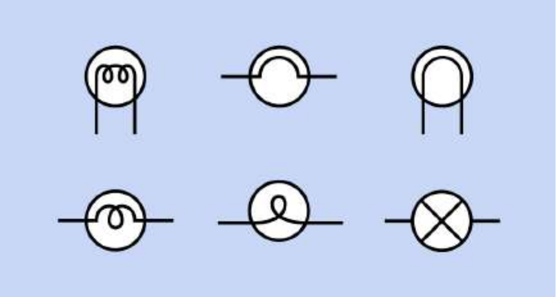

Figure 18-1. *A variety of symbols can represent an incandescent lamp. The one at bottom right may be more commonly used for small panel-mounted indicators.*

The parts of a generic incandescent light bulb are identified in Figure 18-2:

A: Glass bulb.

B: Inert gas at low pressure.

C: Tungsten filament.

D: Contact wires (connecting internally with brass base and center contact, below).

E: Wires to support the filament.

F: Internal glass stem.

G: Brass base or cap.

H: Vitreous insulation.

I: Center contact.

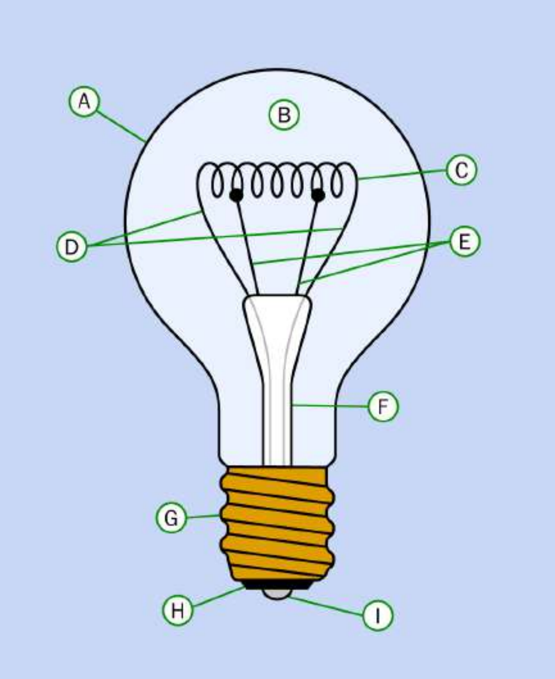

Figure 18-2. *The parts of a typical incandescent lamp (see text for details).*

History

The concept of generating light by using electricity to heat a metal originated with Englishman Humphrey Davy, who demonstrated it with a large battery and a strip of platinum in 1802. Platinum was thought to be suitable because it has a relatively high melting point. The lamp worked but was not practical, being insufficiently bright and having a short lifespan. In addition, the platinum was prohibitively expensive.

The first patent for an incandescent lamp was issued in England in 1841, but it still used platinum. Subsequently, British physicist and chemist Joseph Swan spent many years attempting to develop practical carbon filaments, and obtained a patent in 1880 for *parchmentized thread*. His house was the first in the world to be illuminated by light bulbs.

Thomas Edison began work to refine the electric lamp in 1878, and achieved a successful test with a carbonized filament in October 1879. The bulb lasted slightly more than 13 hours. Lawsuits over patent rights ensued. Carbonized filaments were used until a tungsten filament was patented in 1904 by the German/Hungarian inventor Just Sándor Frigyes and the Croatian inventor Franjo Hanaman. This type of bulb was filled with an inert gas, instead of using a vacuum.

Many other pioneers participated in the effort to develop electric light on a practical basis. Thus it is incorrect to state that "Thomas Edison invented the light bulb." The device went through a very lengthy process of gradual refinement, and one of Edison's most significant achievements was the development of a power distribution system that could run multiple lamps in parallel, using filaments that had a relatively high resistance. His error was insisting on using *direct current* (DC) while his rival Westinghouse pioneered *alternating currrent* (AC), enabling power transmission over longer distances through the use of **transformers**. The use of AC also enabled Tesla's brushless induction motor.

By the mid-1900s, most incandescent bulbs used tungsten filaments.

How It Works

All objects emit electromagnetic radiation as a function of their temperature. This is known as *black body radiation*, based on the concept of an object that absorbs all incoming light, and thus does not reflect any sources from outside itself. As its temperature increases, the intensity of the radiation increases while the wavelength of the radiation tends to decrease.

If the temperature is high enough, the wavelength of the radiation enters the visible spectrum, between 380 and 740 nanometers. (A nanometer is one-billionth of a meter.)

The melting point of tungsten is 3,442 degrees Celsius, but a lamp filament typically operates between 2,000 and 3,000 degrees. At the higher end of this scale, evaporation of metal from the filament tends to cause deposition of a dark residue on the inside of the bulb, and erodes the filament more rapidly, to the point where it eventually breaks. At the lower end of this scale, the light will be yellow and the intensity will be reduced.

Spectrum

The color of black-body radiation is measured using the Kelvin temperature scale. The increment of 1 degree Kelvin is the same as 1 degree Celsius, but the Kelvin scale has a zero value at *absolute zero*. This is the theoretical lowest conceivable temperature, at which there is complete absence of heat. It is approximately –273 degrees Celsius.

From this it is evident that if K is a temperature in degrees Kelvin and C is a temperature in degrees Celsius:

`K = C + 273` (approximately)

Calibration of light sources in degrees Kelvin is common in photography. Many digital cameras allow the user to specify the *color temperature* of lights that are illuminating an indoor scene, and the camera will compensate so that the light source appears to be pure white with all colors in the visible spectrum being represented equally.

Some computer monitors also allow the user to specify a white value in degrees Kelvin.

Color temperature is used in astronomy, because the spectrum of many stars is comparable with that of a theoretical black body.

A color temperature of 1,000 degrees K will have a dark orange hue, while 15,000 degrees K or higher will have a blue hue comparable to that of a pale blue sky. The color temperature of the sun is approximately 5,800 K. Interior lighting is often around 3,000 K, which many people find acceptable because it creates pleasant flesh tones. An incandescent bulb described by the manufacturer as "soft white" or "warm" will have a lower color temperature than one which is sold as "pure white" or "paper white."

Graphs showing the emission of wavelengths at various color temperatures are shown in Figure 18-3. The rainbow section indicates the approximate range of visible wavelengths between ultraviolet, on the left, and infrared, on the right. For purposes of clarity, the peak intensity for each color temperature has been equalized. In reality, increasing the temperature also increases the light output.

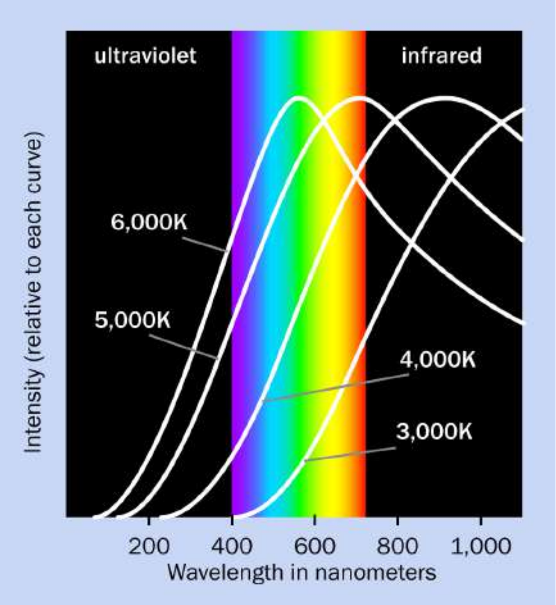

Figure 18-3. *Approximate peak wavelengths for black-body radiation at various color temperatures in degrees Kelvin. The curves have been adjusted so that their peak values are equalized. Adapted from an illustration in the reference book Light Emitting Diodes by E. Fred Schubert.*

Non-Incandescent Sources

So long as light is generated by heating a filament, plotting the intensity against wavelength will result in a smooth curve without irregularities. A higher Kelvin value will simply displace and compress the curve laterally without changing its basic shape to a significant degree.

The introduction of **fluorescent** sources and, subsequently, light-emitting diodes (**LEDs**) has complicated this scenario. Because they are *luminescent* rather than *incandescent*, they do not generate an evenly weighted, continuous range of wavelengths.

LEDs tend to emit monochromatic light, meaning that it is tightly centered around just one color. A "white" LED is really a blue LED in which a phosphor coating on the semiconductor die is excited to create light over a broader range. A fluorescent light tends to create *spectral lines* which show up as sharp peaks at a few wavelengths determined by the mercury inside the bulb. Figure 18-4 illustrates these problems.

The human eye tends to compensate for the yellow emphasis of incandescent lamps and for the irregularities in spectra emitted by other light sources. Also, the eye is often unable to distinguish between "white" light created as a mix of all the visible wavelengths, and light that appears white even though it is dominated by a few isolated wavelengths from a fluorescent source.

However, when the eye views colors that are illuminated by a source that has gaps in its spectrum, some of the colors will appear unnaturally dull or dark. This is true also if an imperfect source is used as a backlight to create colors on a video monitor. Colors rendered by different light sources are shown in Figure 23-7 and subsequent figures.

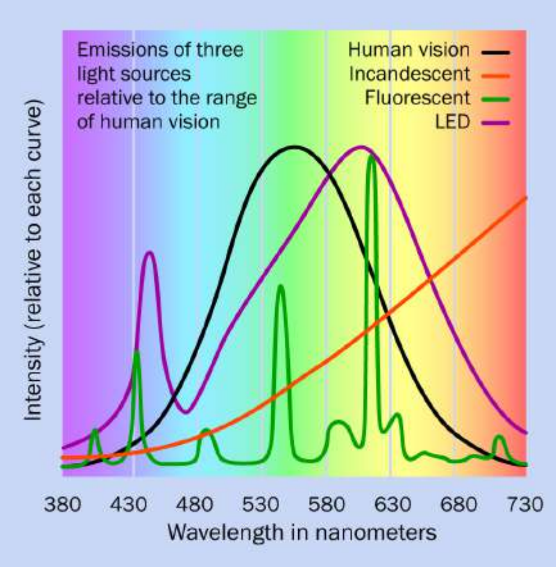

Figure 18-4. *The relative performance of three light sources compared with sensitivity of the human eye to the visible spectrum. Note that the range of wavelengths on the horizontal scale in this figure is not the same as the range in the previous figure. The color assigned to each curve is arbitrary. Adapted from VU1 Corporation.*

Photography is adversely affected by the use of LEDs or fluorescents as a light source. Reds, for example, can seem dark when lit by white LEDs, while blues can be inappropriately intense. Because the source does not have an emission curve comparable to that of an incandescent light, the auto-white balance feature of a digital camera may be unable to address this problem, and it cannot be resolved by entering a different Kelvin number manually.

The fidelity with which a light source is capable of displaying the full visible spectrum is known as the *color rendering index* (CRI), ranging from a perfect score of 100 down to 0 or even lower (sodium-vapor street lighting has a negative value). Computing the index requires standard reference color samples and has been criticized for generating scores that do not correlate well with subjective assessments.

Incandescent bulbs can have a CRI of 100, while an uncorrected "white" LED may score as low as 80.

Power Consumption

Approximately 95% of the power consumed by an incandescent lamp generates heat instead of visible light. This wastage of power in room lighting is compounded by the power consumption of air conditioning to remove the heat from enclosed spaces in hot climates. While the heat from incandescent lamps does reduce the need for space heating in cold environments, heat is delivered more efficiently by using systems designed for that purpose. Consequently, greater energy efficiency can be achieved with a light source that generates less heat, regardless of ambient air temperature.

Variants

Miniature Lamps

Prior to the development of **LED**s, all light-emitting panel-mounted indicators were either **neon bulbs** or **incandescent lamps**. The use of neon is limited by its need for a relatively high voltage.

Miniature incandescents were the traditional choice for battery-powered light sources, and at the time of writing are still used in cheap flashlights. Variants are available that are as small as a 5mm LED, with a claimed life expectancy that is comparable, although they draw more current to generate an equivalent light intensity, because much of their power is wasted in infrared wavelengths.

The photograph in Figure 18-5 is of a miniature lamp terminating in pins spaced 0.05" apart. The total height of the lamp, including its ceramic base, is less than 0.4," while its diameter is just over 0.1". It draws 60mA at 5V and is rated for 25,000 hours.

The photograph in Figure 18-6 is of a lamp of similar size and power consumption, but terminating in wire leads and rated for 100,000 hours. It emits 0.63 lumens.

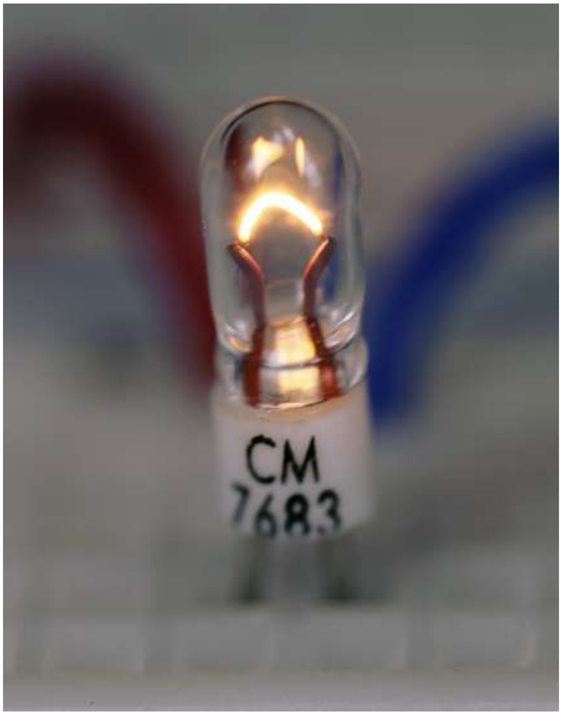

Figure 18-5. *A miniature lamp less than 0.4" high, terminating in pins spaced 0.05" apart.*

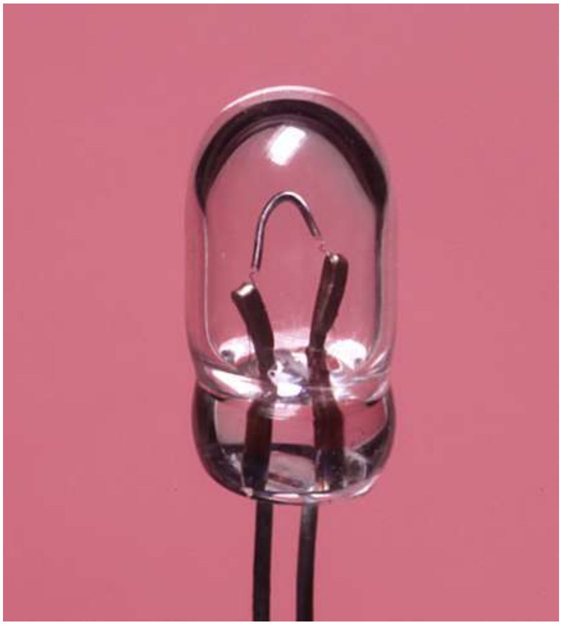

Figure 18-6. *This lamp is 0.25" high and terminates in wire leads.*

The lamp pictured in Figure 18-7 is slightly larger, with a glass envelope about 0.25" diameter. It is rated for less than half the lifetime of the lamp in Figure 18-6 but emits three times as much light —a typical tradeoff. Various base styles are available.

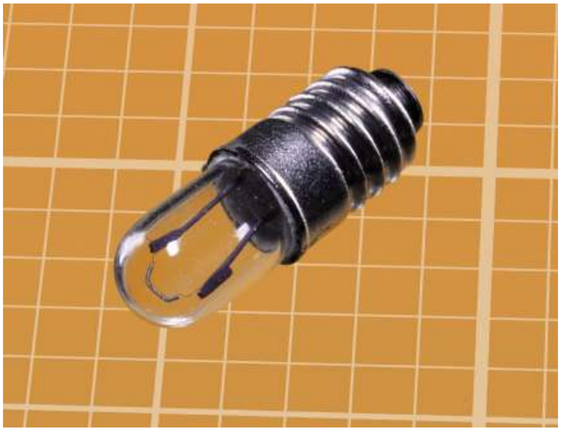

Figure 18-7. *This lamp has a glass envelope about 0.35" high. Its screw-in base makes it easier to replace than an LED.*

In the United States, the light output from miniature incandescent lamps may be measured in lumens, but is more often rated in *mean spherical candlepower* (MSCP). An explanation of light measurement is included in "MSCP" on page 178.

Lamp lenses provide a quick and simple way to add color to a miniature incandescent lamp. Usually the lens is cylindrical with a hemispherical end cap, and is designed to push-fit or snap-fit over a small lamp. Even when the cap is translucent, it may still be referred to as a lens.

Panel-Mount Indicator Lamps

This term often refers to a tubular assembly containing a miniature lamp, ready for installation.

The enclosure is often designed to snap-fit into a hole drilled in the panel. If the incandescent bulb inside the enclosure cannot be replaced, the component is said to be "non-relampable." Figure 18-8 shows a 12-volt panel-mount indicator lamp.

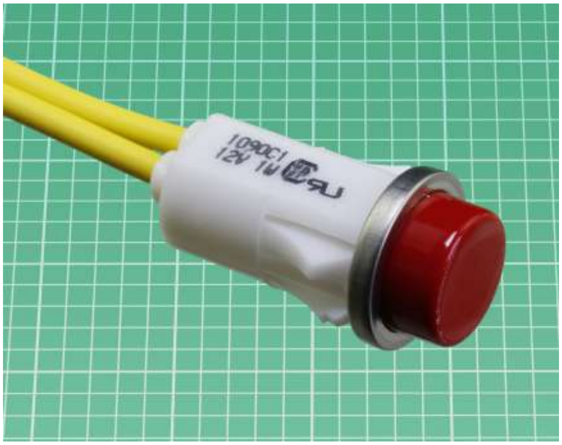

Figure 18-8. *This panel-mount indicator lamp is designed to push-fit into a hole 1/2" in diameter. The bulb inside it is not replacable, causing the assembly to be classified as "non-relampable."*

Halogen or Quartz-Halogen

This is a type of incandescent lamp containing gases under pressure in which halogens such as iodine or bromine cause evaporated tungsten atoms to be redeposited on the filament. A halogen lamp can therefore operate at a higher temperature, creating a light that is less yellow and brighter than that from a comparable incandescent lamp. It also enables a smaller bulb, but requires an envelope of borosilicate-halide glass (often termed *fused quartz*) instead of regular glass. A halogen lamp will be slightly more efficient than an incandescent bulb of the same wattage, and will last longer.

Halogens are available in a variety of formats. The small bulb pictured in Figure 18-9 consumes 75W, emitting 1,500 lumens at 3,000 degrees Kelvin. The light intensity is claimed to be equivalent to that of a 100W incandescent bulb. It has a mini-candelabra base.

Oven Lamps

Oven lamps are designed to withstand the high temperature in an oven. Typically they are usable with ambient temperatures up to 300 degrees C. A common power rating is 15W.

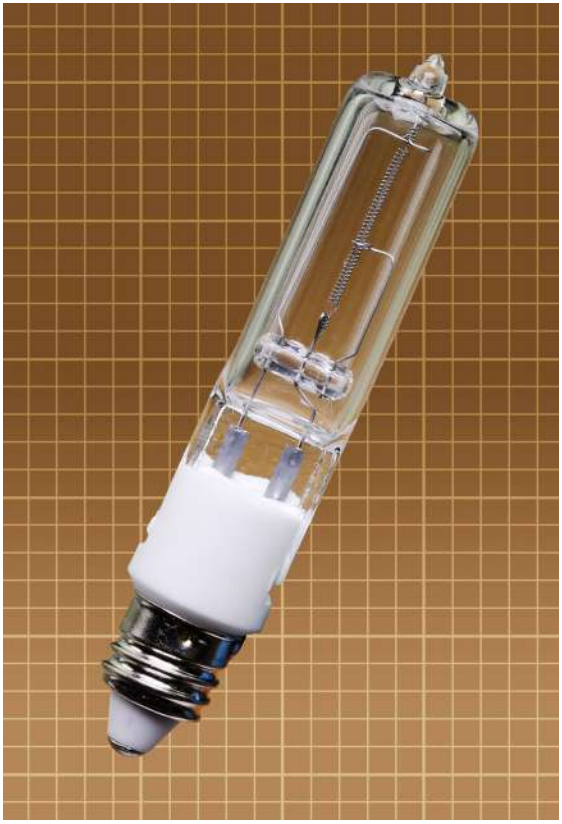

Figure 18-9. *A halogen bulb slightly more than 2" in length, designed for 115VAC.*

Base Variants

Miniature lamps are available with a wide variety of connection options, including wire terminals, single-contact bayonet, double-contact bayonet, miniature screw base, and fuse style. Most of these options require a matching socket.

Screw-in lamps for room illumination are common in household lighting in the United States and many other countries (but not in the UK, where bayonet fittings are used). The US socket size is designated by letter E followed by a number that gives the socket diameter in millimeters. Common sizes are E10, E14 and E27.

A *bayonet base* is fitted with two small lugs protruding on opposite sides. The lamp is secured by pushing it in and twisting it to engage the lugs in slots in the socket. The advantage of a bayonet base is that the bulb is less likely to become loose as a result of vibration.

A *pin base* consists simply of a pair of pins that will push-fit into small holes in a socket.

A *flange base* has a flange that engages in a socket where flexible segments will retain it.

A *wedge base* is forced between two contacts which retain the bulb by friction.

Some indicator lamps terminate simply in long, thin leads that can be soldered.

Values

While the power consumption of full-size incandescent lamps is rated in watts, small indicator lamps are rated in milliamps at the voltage for which they are designed. Miniature lamps may require specific voltages ranging from as low as 2V to 24V. A higher voltage generally necessitates a longer filament, which may entail a larger bulb.

The light that a lamp will emit can be measured in two ways: either as the power of the lamp (not its power consumption, but its radiating power), or as the light delivered to a specific area at a specific distance. These two measurements may differ because a lamp may concentrate its light in a beam, as in the case of a reflector bulb or an LED.

Power

Flux, in watts, is a measurement of energy flow in joules per second. The total radiating power of a lamp, in all wavelengths, in all directions, is known as its *radiant flux*. Because invisible wavelengths are of little interest when assessing the brightness of a lamp, the term *luminous flux* is used to describe the apparent brightness of the lamp in the visible spectrum. The unit for luminous flux is the *lumen*.

The human eye is most responsive to yellow-green hues in the center of the spectrum. Consequently, the measurement of luminous flux is weighted toward green at a wavelength of 555 nanometers. Red and violet are considered to have low luminous flux, while infrared and ultraviolet have a zero value.

When considering a value expressed in lumens, remember:

- Lumens are a measure of the total radiated power output of a light source, in all directions, in the visible spectrum only, weighted toward the characteristics of the human eye.
- The number of lumens of a light source does not define the direction in which the light is shining, or its uniformity.
- The abbreviation for lumen is *lm*.

A conventional incandescent lamp that consumes 100W of electricity is likely to have a light output of about 1,500 lumens. A 40W fluorescent tube can have a light output of about 2,600 lumens.

Illuminance

The *illuminance* of a light source is defined as the luminous flux per unit of area. This can be thought of as the brightness of a surface illuminated by the source.

Illuminance is measured in *lux*, where 1 lux = 1 lumen per square meter. For accurate calibration, the illuminated surface should be spherical in shape, and must be located 1 meter from the light source, with the source at the geometrical center of the sphere.

Illuminance used to be measured in *foot-candles*, where 1 foot-candle was 1 lumen per square foot.

- The number of lumens per square meter (lux) does not define the size of the illuminated area, only the brightness per unit of area.
- A lamp that has a tightly focused beam can achieve a high lux rating. When selecting a lamp for an application, the angle of dispersion of the beam must be considered in conjunction with its lux rating.

Intensity

A *candela* measures the luminous flux within an angle of dispersion. The angle is three-dimensional, and can be imagined as the sharpness of a point of a cone, where the light source is at the point and the cone represents the dispersion of light.

The three-dimensional angle of dispersion is measured in *steradians*. If a light source is at the center of a sphere that has a radius of 1 meter, and is illuminating one square meter of the surface of the sphere, the angle of dispersion is 1 steradian.

- A source of 1 lumen which projects all its light through a dispersion angle of 1 steradian is rated at 1 candela.
- The number of candelas does not define the angle of dispersion, only the intensity within that angle.
- A light source rated for 1,000 candelas could have a power of 10 lumens concentrated within an angle of 0.01 steradians, or could have a power of only 1 lumen concentrated within an angle of 0.001 steradians.
- There are 1,000 millicandelas in 1 candela. The abbreviation for candela is *cd* while the abbreviation for a millicandela is *mcd*.
- LEDs are often rated in mcd. The number describes the intensity of light within its angle of dispersion.

MSCP

Although the term *candlepower* is obsolete, it has been redefined as being equal to 1 candela. *Mean spherical candlepower* (MSCP) is a measurement of all the light emitted from a lamp in all directions. Because the light is assumed to be omnidirectional, it fills 4 * π (about 12.57) steradians. Therefore 1 MSCP = approximately 12.57

lumens. In the United States, MSCP is still the most common method of rating the total light output of a miniature lamp.

Efficacy

The *radiant luminous efficacy* (abbreviated *LER*) assesses how effective a lamp is at channeling its output within the visible spectrum, instead of wasting it in other wavelengths, especially infrared. LER is calculated by dividing the power emitted in the visible spectrum (the *luminous flux*) by the power emitted over all wavelengths.

Thus, if VP is the power emitted in the visible spectrum, and AP is the power emitted in all wavelengths:

```
LER = VP / AP
```

LER is expressed in *lumens per watt*. It can range from a low value of around 12 lm/W for a 40W incandescent bulb to 24 lm/W for a quartz halogen lamp. Fluorescent lamps may average 50 lm/W. LEDs vary, but can achieve 100 lm/W.

Efficiency

The *radiant luminous efficiency* (abbreviated *LFR*) of a lamp measures how good its radiant luminous efficacy is, compared with an imaginary ideal lamp. (Note the difference between the words "efficiency" and "efficacy.") LFR is determined by dividing the radiant luminous efficacy (LER) by the maximum theoretical LER value of 683 lm/W, and multiplying by 100 to express the result as a percentage. Thus:

```
LFR = 100 * ( LER / 683 )
```

The LFR ranges from around 2% for a 40W bulb to 3.5% for a quartz halogen lamp. LEDs may be around 15% while fluorescents are closer to 10%.

How to Use It

When first introduced, LEDs were limited by their higher price, lower maximum light output, and inability to display blue or white. The price difference has disappeared for small indicators, while gaps in the color range have been filled (although the color rendering index of LEDs is still inferior).

Brightness remains an advantage for large incandescents relative to LEDs, as they are more upwardly scalable. However, fluorescents and vapor lamps have an advantage for very high light output, as in the lighting of big-box stores or parking lots. Thus the range of applications for incandescent bulbs is diminishing, especially because common types are now illegal for domestic light fixtures in many parts of the world.

Relative Advantages

When choosing whether to use an incandescent lamp or an LED, these advantages of an incandescent lamp should be considered:

- The intensity can be adjusted with a **triac**-based dimmer. Regular fluorescents cannot be dimmed, while LEDs often require different dimmer circuitry.
- The intensity can also be adjusted with a rheostat. The output from fluorescents cannot.
- Easy white-balance correction. LEDs and fluorescents do not naturally produce a consistent output over the visible spectrum.
- Can be designed to operate directly from a wide range of voltages (down to around 2V and up to around 300V). A higher voltage entails a longer filament wire, which may require a larger bulb. LEDs require additional components and circuitry to use higher voltages.
- Incandescent bulbs are more tolerant of voltage fluctuations than LEDs. With battery operation, the incandescent will still provide some reduced light output when the voltage has diminished radically. LEDs will not perform at all at currents lower than their threshold.
- An incandescent is nonpolarized and may be socketed, which simplifies user replacement.

LEDs are polarized and are usually soldered in.

- Can be powered by AC or DC without any modification or additional circuitry. LEDs require DC, which must be provided through a transformer and rectifier, or similar electronics, if AC power is the primary source.
- Can be equally visible from a wide range of viewing angles. LEDs have restricted viewing angles.
- The heat output from an incandescent bulb may occasionally be useful (for example in a terrarium, or in incubators for poultry).
- Trouble-free switching. Fluorescents tend to hesitate and blink when power is applied, and they require a *ballast* to energize them. The lifespan of fluorescents is reduced by frequent switching.
- No low-temperature problems. Incandescent lamps are not significantly affected by low temperatures. Fluorescents may not start easily in a cold environment, and may flicker or glow dimly for 10 minutes (or more) until they are warm enough to function properly.
- Easy disposal. Fluorescent lights contain small quantities of mercury that are an environmental hazard. They should not be mixed with ordinary trash. Compact fluorescent lamps (CFLs) and LEDs used for room lighting will be packaged with electronics that should ideally be recycled, although this is not very practical. Incandescent bulbs impose the least burden on the environment when they are thrown away.

However, the incandescent lamp has some obvious disadvantages:

- Relatively inefficient.
- More susceptible to vibration.
- More fragile.
- Likely to have a shorter natural life expectancy than LEDs, fluorescents, or neon bulbs, although the lifetime of a small panel indicator can be equal to that of an LED if a low color temperature is acceptable.
- Requires a filter or tinted glass envelope to generate colored light. This further reduces the lamp's efficiency.
- Cannot be miniaturized to the same degree as an LED indicator.

Derating

The lifespan of a lamp can be greatly extended by choosing one with a higher current rating or using it at a lower voltage. The light output will be reduced, and the color temperature will be at a lower Kelvin number, but in some situations this tradeoff may be acceptable.

The graphs in Figure 18-10 suggest that if the voltage of a hypothetical miniature lamp is reduced to 80% of the manufacturer's recommended value, this can make the lamp last 20 times as long. Note, however, that this will cut the light intensity to 50% of its normal value.

Conversely, using 130% normal voltage will give 250% of the normal light output, while shortening the life of the lamp to 1/20 of its normal value. Naturally these figures are approximations that may not apply precisely to a specific lamp.

What Can Go Wrong

High Temperature Environment

If an incandescent lamp is used in an environment hotter than 100 degrees Celsius, the life of the lamp is likely to be reduced by the "water cycle." Any water molecules inside the glass envelope will break down, allowing oxygen to combine with the tungsten filament to form tungsten oxide. The tungsten is deposited on the inside of the glass while the oxygen is liberated and begins a new cycle.

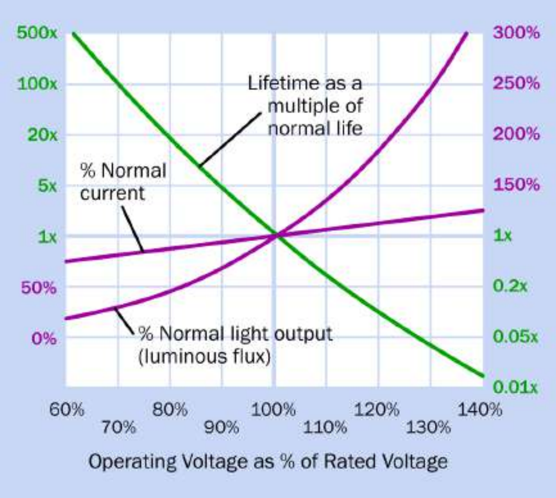

Figure 18-10. *The life expectancy of a hypothetical miniature lamp is very strongly influenced by voltage. Applying only 60% of the rated voltage can make a lamp last 500 times its normal lifespan, although it will greatly reduce light output. Note that the vertical axes apply to curves of the same color. Adapted from "Characteristics of Miniature Lamps" from Toshiba Lighting and Technology Corporation.*

Fire Risk

The partially evacuated bulb of an incandescent lamp provides some separation and protection from the heat in the filament, but if the bulb cannot disperse heat by radiation or convection, its temperature can rise to the point where it ignites flammable materials.

Halogen lamps have an elevated fire risk because they operate at a higher temperature and are smaller, providing less surface area to disperse the heat. They also contain gases under seven to eight atmospheres of pressure. Thermal stress can cause a halogen bulb to shatter, and fingerprints on the glass can increase this risk.

Current Inrush

When an incandescent lamp is first switched on, its filament has one-tenth the resistance that it will exhibit when it becomes hot. Consequently, the lamp will take a large initial surge of current, which stabilizes after about 50 milliseconds. This should be considered if one or more small lamps shares a DC power supply with components such as logic chips that may be sensitive to voltage fluctuations.

Replacement Problems

Because of the limited life of incandescent lamps, they should be installed in such a way that they are easy to replace. This can be an issue with panel indicators, where disassembly of a device may be necessary to reach the lamp.

The range of small incandescent lamps is diminishing, and may continue to diminish in the future. Future availability of replacement lamps should be considered when designing a circuit. When building equipment in small quantities, spare lamps should be purchased for future use.

neon bulb

19

The terms **neon bulb**, *neon indicator*, and *neon lamp* tend to be used interchangeably. In this encyclopedia, a neon *bulb* is defined as a glass capsule containing two electrodes in neon gas (or a combination of gases in which neon is present). A neon *lamp* is an assembly containing a neon bulb, usually using a plastic tube with a tinted transparent cap at one end. A neon *indicator* is a miniature neon lamp that is usually panel-mounted.

Large-scale neon tubes used in signage are not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **incandescent lamp** (see Chapter 18)
- **fluorescent light** (see Chapter 20)
- **LED indicator** (see Chapter 22)

What It Does

When voltage is applied between two electrodes inside a neon bulb, the inert gas inside the bulb emits a soft red or orange glow. This color may be modified by using a tinted transparent plastic cap, known as a *lens*, in a *neon lamp* assembly.

A neon bulb is usually designed for a power supply of 110V or higher. It functions equally well with alternating or direct current.

The schematic symbols in Figure 19-1 are commonly used to represent either a neon bulb or a neon lamp. They are all functionally identical. The black dot that appears inside two of the symbols indicates that the component is gas filled. The position of the dot inside the circle is arbitrary. Even though all neon bulbs are gas filled, the dot is often omitted.

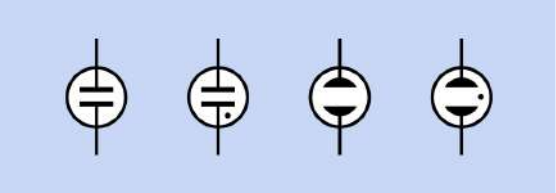

Figure 19-1. *Any of these symbols may represent a neon bulb or a neon lamp. The dot in two of the symbols indicates that the component is gas filled. All neon bulbs are gas filled, but the dot is often omitted.*

The photograph in Figure 19-2 shows a neon bulb with a series resistor preattached to one lead. Many bulbs are sold in this configuration, because a resistor must be used to limit current through the bulb. The bulb has no polarity and can be used on an AC or DC power supply. The same bulb is shown in its energized state in Figure 19-3.

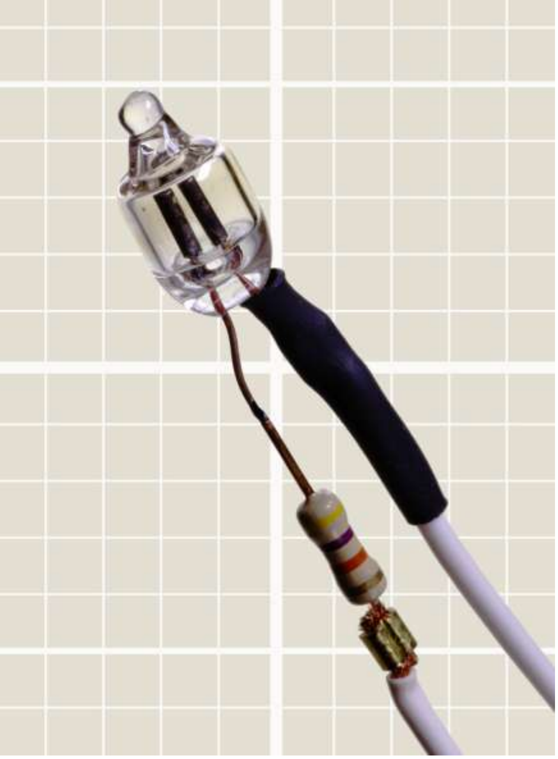

Figure 19-2. *A typical neon bulb with series resistor attached to one lead.*

How It Works

Construction

The parts of a neon bulb are illustrated in Figure 19-4. When the bulb is fabricated, it begins as a glass tube. The leads are made of *dumet*, consisting of a copper sheath around a nickel iron core. This has the same coefficient of expansion as glass, so that when the glass is heated and melted around the leads, it forms a seal that should be unaffected by subsequent temperature fluctuations. This area is known as the *pinch* in the tube.

Nickel electrodes are welded onto the leads before the leads are inserted into the tube. The electrodes have an emissive coating that reduces the minimum operating voltage. The glass tube is filled with a combination of neon and argon gases, or pure neon for higher light output (which reduces the life of the component). The top end of the glass tube is heated until it melts, and is pinched off. This creates a distinctive protrusion known as the *pip*.

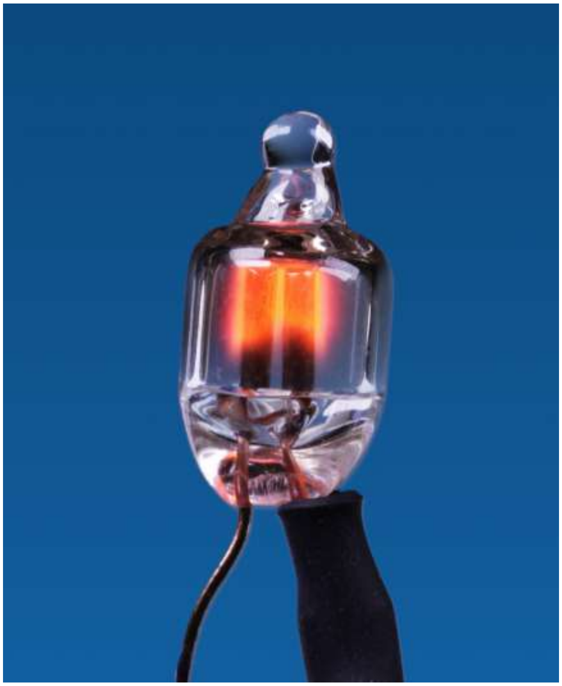

Figure 19-3. *The same bulb from the previous photograph, energized with 115VAC.*

Ionization

When a voltage is applied between the leads to the bulb, the gas becomes ionized, and electrons and ions are accelerated by the electric field. When they hit other atoms, these are ionized as well, maintaining the ionization level. Atoms are excited by collisions, moving their electrons to higher energy levels. When an electron returns from a higher level to a ground state, a photon is emitted.

This process begins at the *starting voltage* (also known as the *striking voltage*, the *ignition voltage*, or the *breakdown voltage*) usually between 45V and 65V for standard types of bulb, or between 70V and 95V for high-brightness types.

When the bulb is operating, it emits a soft radiance known as a *glow discharge* with a wavelength ranging from 600 to 700 nanometers.

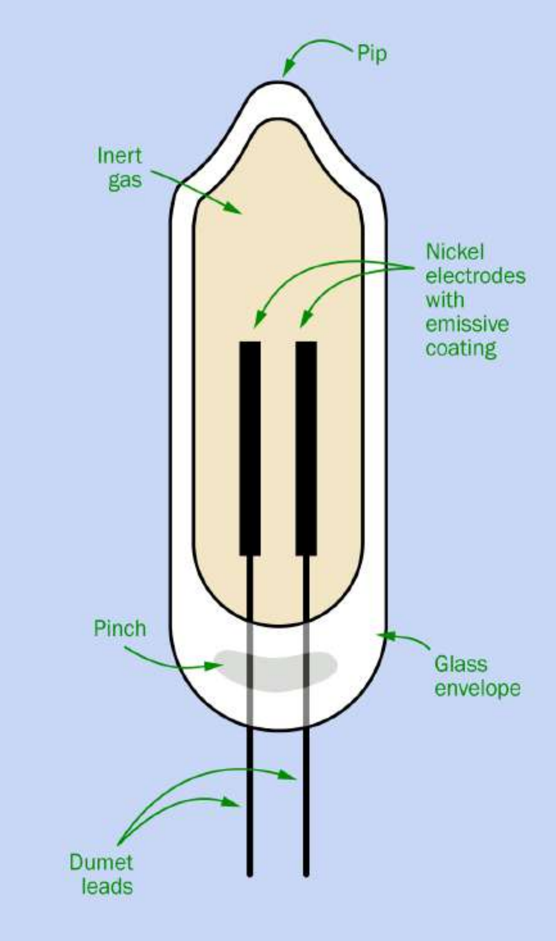

Figure 19-4. *The parts of a neon bulb. See text for details.*

The ionization of the gas allows current to flow through it. This will continue even if the power supply is reduced by 10 to 20 volts to a level known as the *maintaining voltage*.

Negative Resistance

When the glow discharge persists below the starting voltage, this is a form of *hysteresis*, meaning that the neon bulb tends to "stick" in its on state. It remains on while its power supply decreases to the maintaining voltage, but once it switches off, it will "stick" in its off state until the power supply increases again above the maintaining voltage to the starting voltage. The concept of hysteresis is discussed in the entry on **comparators**. See Figure 6-2.

A neon bulb is said to have *negative resistance*. If the current is allowed to increase without restraint, the resistance eventually decreases while the current increases further. If this runaway behavior is not controlled, the bulb will destroy itself.

This behavior is characteristic of gas-discharge tubes generally. A graph showing this appears in Figure 19-5. Note that both scales are logarithmic. Also note that the curve shows how current will be measured in response to voltage. If the voltage is reduced after it has increased, the transitional events shown by the graph will not recur in reverse order. This is especially true if arcing is allowed to begin, as it will almost certainly destroy the component.

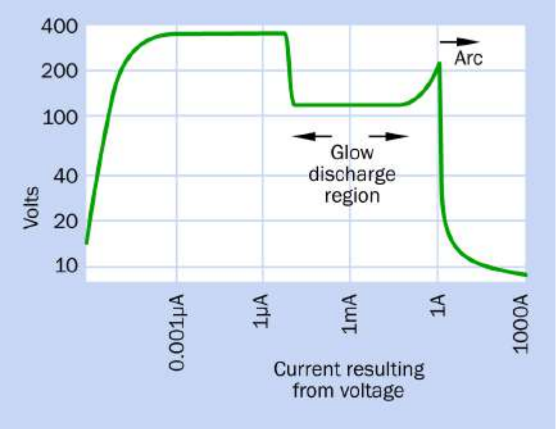

Figure 19-5. *A gas discharge tube, such as a neon bulb, is said to have a negative resistance, as current passing through it tends to increase uncontrollably after the gas is ionized and becomes conductive. (Derived from measurements made by David Knight, on a web page named after his radio ham call sign, G3YNH.)*

A neon bulb can be controlled very simply with a series resistor that maintains it in gas-discharge mode. To understand the operation of the resistor, consider the combination of the lamp and the resistor as a voltage divider, as shown in Figure 19-6. Before the lamp begins to pass current, it has an almost infinite resistance. Therefore, the voltage on both sides of the resistor will be approximately equal, the bulb passes almost no current, and it remains dark.

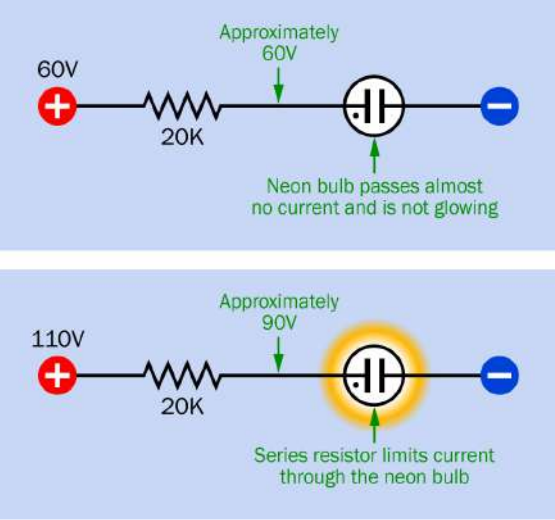

Figure 19-6. *A series resistor is essential to limit the current through a neon bulb.*

After the lamp begins to pass current, the requirement now is for the series resistor to reduce the voltage from the supply level (probably around 110V) to the maintaining level (probably around 90V). This means that the desired voltage drop is 20V, and if the manufacturer's specification tells us that the lamp should pass 1mA (i.e., 0.001 amps), R, the value of the series resistor, is given by Ohm's Law:

```
R = 20 / 0.001
```

Thus, the value for R is 20K. In fact, the value of a resistor supplied with a neon bulb may range from 10K to 220K, depending on the characteristics of the bulb and the supply voltage that will be used.

Now if the bulb's effective internal resistance falls radically, the resistor still limits the currrent. In a hypothetical worst-case scenario, if the bulb's resistance drops all the way to zero, the resistor must now impose the full voltage drop of 110V, and the current, I, will be found by Ohm's law:

```
I = 110 / 20,000
```

That is, about 5mA, or 0.005A.

Neon tubes used in signage require a more sophisticated voltage control circuit which is not included in this encyclopedia.

How to Use It

The use of a neon bulb for an indicator lamp is primarily limited to situations where domestic supply voltage (115VAC or 220VAC) is readily available. "Power on" lights are the obvious application, especially as neon indicators can accept AC. The switch shown in Figure 19-7 is illuminated by an internal neon bulb. The rectangular indicator in Figure 19-8 is designed to run on domestic supply voltage, and its internal bulb and resistor can be clearly seen through the green plastic. The assembly in Figure 19-9 is about 0.5" in diameter, which is the lower limit for neon indicators.

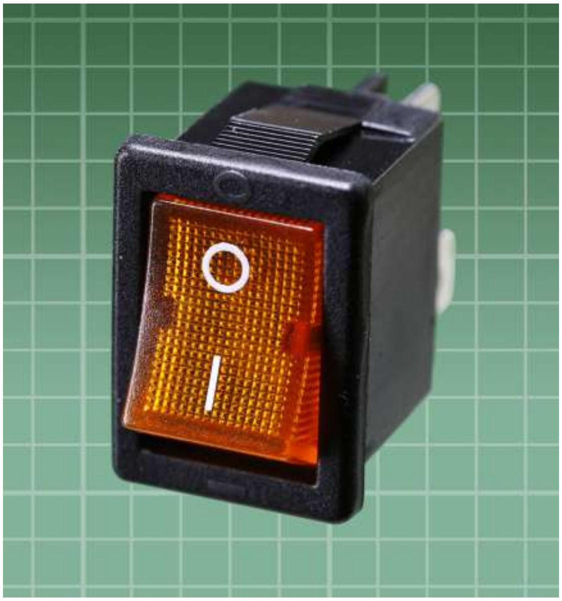

Figure 19-7. *This power switch is illuminated by an internal neon bulb.*

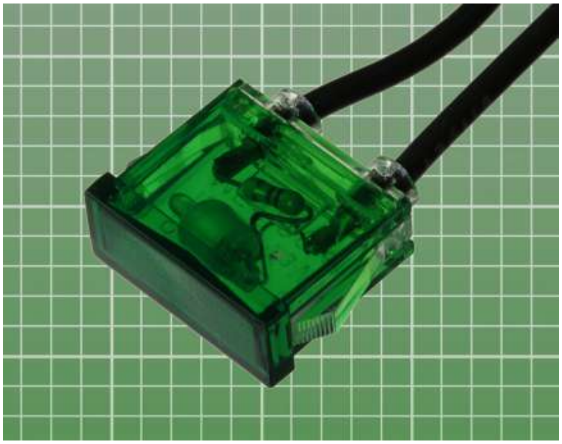

Figure 19-8. *The neon bulb and its series resistor are visible inside this indicator.*

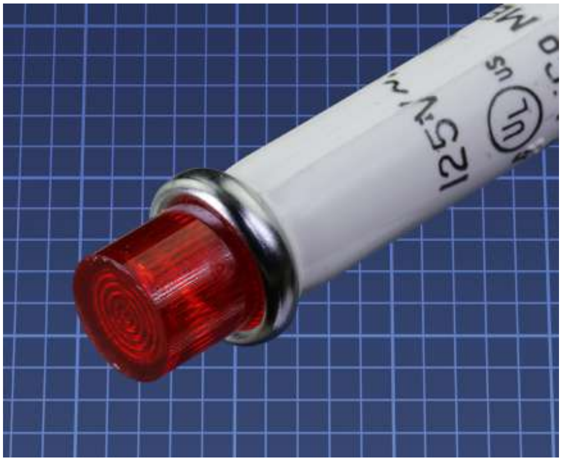

Figure 19-9. *A relatively small neon indicator lamp, designed for insertion in a hole 0.5" diameter.*

Limited Light Output

Neon bulbs have a light output of around 0.06 lumens per milliamp of consumed power (standard brightness type) or 0.15 lumens per milliamp of consumed power (high brightness type).

Comparing this value with the intensity of **LED** indicators is difficult. Their light output is customarily measured in *millicandelas* (mcd), because LED indicators almost always include a lens that focuses the light, and the candela is a measurement of luminous flux within an angle of dispersion. Moreover, because the intensity of neon indicators is not of great interest in most applications, datasheets usually do not supply an intensity value.

One way around the problem of comparisons is to use the standard of *radiant luminous efficacy* (LER), which is defined in the entry on incandescent lamps (see "Efficacy" on page 179). A standard-brightness neon bulb has an LER of about 50 lumens per emitted watt of luminous flux. A light-emitting diode may reach an LER of 100 lm/W. However, a neon bulb operates typically around 1 mA while an LED indicator may use 20mA. Therefore, a typical LED indicator may appear to be 30 to 50 times brighter than a typical neon bulb.

Consequently, neon may be an inferior choice in a location where there is a high level of ambient light. Direct sunlight may render the glow of a neon indicator completely invisible.

Efficiency

Because a neon bulb does not use a lot of power and generates negligible heat, it is a good choice where current consumption is a consideration (for example, if an indicator is likely to be on for long periods). The durability and low wattage of neon bulbs, and their convenient compatibility with domestic power-supply voltage, made them a favorite for night-lights and novelty lamps in the past. Figure 19-10 shows an antique bulb containing an ornamental electrode, while Figure 19-11 is a piece of folk art, approximately 1" in diameter, mounted on a plug-in plastic capsule containing a neon bulb.

Ruggedness

Neon bulbs are a good choice in difficult environments, as they are not affected by vibration, sudden mechanical shock, voltage transients, or frequent power cycling. Their operating temperature range is typically from -40 to +150 degrees Celsius, although temperatures above 100 degrees will reduce the life of the lamp.

Figure 19-10. *In bygone decades, ornamental neon bulbs with specially shaped electrodes were popular.*

Figure 19-11. *Neon folk art survives in this hand-painted night-light sold in a Florida tourist shop.*

Power-Supply Testing

When driven by DC current, only the negative electrode (the *cathode*) of a neon bulb will glow. When AC current passes through the bulb, both terminals will glow.

If a bulb (with series resistor) is placed between the "hot" side of a domestic AC power supply and ground, the bulb will glow. If it is placed between the neutral side of the supply and ground, it will not glow.

These features enable a neon bulb to be used for simple power-supply testing.

Life Expectancy

The metal of the electrodes gradually vaporizes during everyday use of a neon bulb. This is known as *sputtering* and can be observed as the glass capsule becomes darkened by deposition of vaporized metal. The electrodes will have a more limited life in a lamp used with DC voltage, where sputtering affects only the cathode. Using AC, the electrodes take turns functioning as the cathode, and vaporization is distributed between both of them.

Failure of a neon lamp can occur as sputtering erodes the electrodes to the point where the maintaining voltage will increase until it almost reaches the level of the power supply. At this point, the bulb will flicker erratically.

Failure can also be defined as a gradual reduction in brightness to 50% of rated light output, caused by accumulated deposition in the glass capsule. Because deposition occurs more heavily on the sides of the bulb, a longer apparent life is possible if the bulb is mounted so that it is viewed from the end.

Typically, neon bulbs are rated for 15,000 to 25,000 hours (two to three years of constant operation). However, the life can be greatly increased by a slight reduction in voltage, which may be achieved by substituting a series resistor with a slightly higher value.

The relationship between operating life and resistor value is shown below. If LA is the normal operating life, LB is the extended operating life, RA is the normal resistor value, and RB is a higher resistor value:

$$LB = LA * (RB / RA)^{3.3}$$

For example, if a normal resistor value is 20K, and it is increased to 22K, the life of the lamp should increase by a factor of slightly more than 1.4.

Variants

A typical neon bulb terminates in leads, and a lamp assembly often has solder tabs, although it may have a base with a screw thread, flange, or bayonet pins for insertion into a compatible socket. A lamp assembly that does not use a base will either snap-fit into a hole of appropriate size and shape, or may be retained with a nut that engages with a plastic thread on the cylinder of the lamp.

Some neon bulbs or lamp assemblies terminate in pins for direct insertion into a printed circuit board.

Almost all neon bulbs operate either in the 100V to 120V range or in the 220V to 240V range.

Light intensity is expressed either as “standard” or “extra-bright,” although datasheets usually do not define those terms.

Nixie Tubes

Nixie tubes, first marketed in 1955, were used to display numerals from 0 through 9 in the days before LEDs took over this capability. They are no longer being manufactured.

Each numeral was physically formed from metal and functioned as an electrode inside a tube filled with a neon-based gas mixture. The typographical elegance of the digits and their aesthetically pleasing glow made Nixies enduringly popular. With a long lifespan, vintage tubes are still usable and can be purchased cheaply from sources such as eBay. Many originate in Russia, where Nixie-type displays were manufactured into the 1980s. The Russian tubes can be identified by their use of a numeral 5 that is a numeral 2 turned upside-down.

Nixie tubes typically require 170VDC. This creates a challenge for a power supply and switching, and can be a safety hazard.

Figure 19-12 shows six Nixie-type tubes repurposed for use as a 24-hour digital clock.

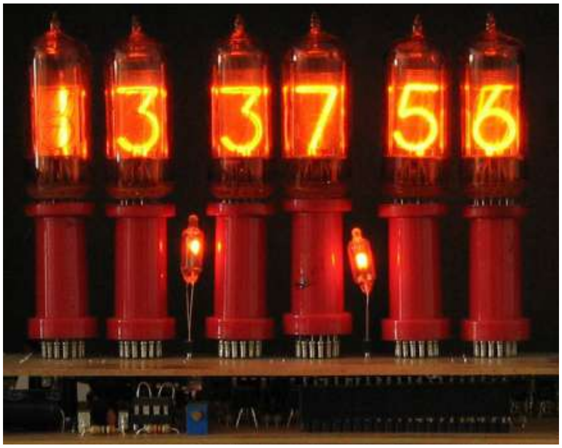

Figure 19-12. *A 24-hour clock using Nixie-type tubes. Source: Wikipedia, public domain.*

What Can Go Wrong

False Indication

Because a neon bulb requires so little power, it may be energized by induced voltages from elsewhere in a circuit, especially if inductive components such as transformers are used. To prevent this, a high-value resistor can be placed in parallel with the bulb, in addition to the series resistor that must always be used.

Failure in a Dark Environment

Because a neon bulb requires a minimal amount of light to initiate its own photon emissions, it may take time to start glowing in a very dim environment, and may not light at all in total darkness. A few bulbs include a small amount of radioactive material that enables them to self-start in complete absence of ambient light.

Premature Failure with DC

The life expectancy quoted in datasheets for neon bulbs usually assumes that they are powered by AC. Because DC results in faster vaporization of the electrodes, the expected lifetime should be reduced by 50% if DC power will be used.

Premature Failure through Voltage Fluctuations

Because the deterioration of a neon bulb accelerates rapidly with current, a sustained voltage that passes slightly more current can radically reduce the expected lifespan.

Replacement

Replacement can be an issue with panel indicators, where disassembly of a device may be necessary to reach the bulb. Bear in mind, however, that an easily removable bulb becomes vulnerable to tampering.

light source, indicator, or display > single source > fluorescent light

20

fluorescent light

This entry deals primarily with *fluorescent tubes* (infrequently but sometimes described as *fluorescent lamps*), and *compact fluorescent lamps* (CFLs) that are marketed as a substitute for **incandescent lamps**. *Cold-cathode fluorescent lamps* (CCFLs) are also mentioned.

Vacuum fluorescent devices have a separate entry in this encyclopedia. A fluorescent tube or CFL does not contain a vacuum.

Although the diode(s) in a white **LED area lighting** unit are coated with a layer of fluorescent phosphors, they are not categorized here as fluorescent lights, and have their own entry.

A **neon bulb** resembles a fluorescent light in that it is a *gas-discharge* device, but the interior of its glass envelope is usually not coated with fluorescent phosphors, and therefore it has its own entry.

OTHER RELATED COMPONENTS

- **incandescent lamp** (see Chapter 18)
- **LED area lighting** (see Chapter 23)
- **vacuum fluorescent** devices (see Chapter 25)
- **neon bulb** (see Chapter 19)

What It Does

Fluorescent tubes or *compact fluorescent lamps* (CFLs) are primarily used for area lighting. A partially disassembled CFL appears in Figure 20-1, showing the control electronics that are normally hidden inside the base.

There is no standardized schematic symbol to represent a fluorescent light. Figure 20-2 shows three commonly used symbols for a fluorescent tube on the left, and three symbols for a CFL on the right. Note that two of the symbols for a CFL are the same as those for an **incandescent lamp**, shown in Figure 18-1.

How It Works

Luminescence is the emission of light as a result of a process that does not require heat. (The opposite phenomenon is *incandescence*, in which heating causes an object to emit light; see Chapter 18 for a description of **incandescent lamps**.)

Fluorescence is a form of luminescence. It occurs when electrons in a material are energized and then make a transition back to ground level, at which point they radiate their energy as visible light. The incoming energy can consist of other light at a higher frequency. Some creatures, including species of arachnids and fish, will fluoresce when they are lit with ultraviolet light.

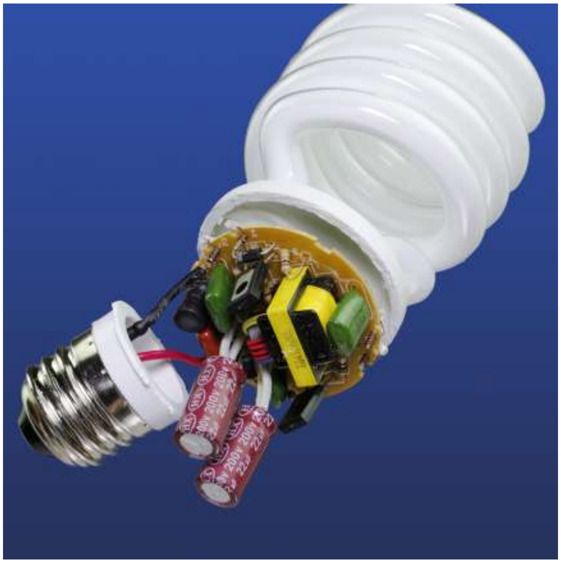

Figure 20-1. *A compact fluorescent lamp with its base cut away to reveal the control electronics.*

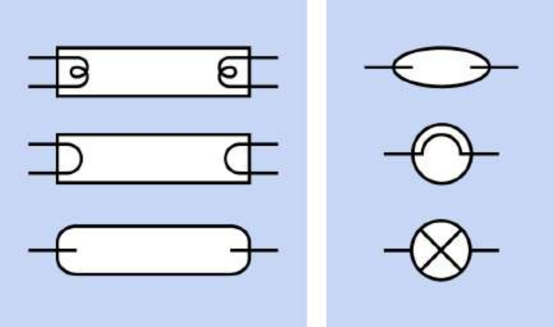

Figure 20-2. *Schematic symbols to represent fluorescent tubes and bulbs are not standardized. See text for details.*

A *fluorescent tube or lamp* contains a very small amount of mercury vapor that can be stimulated to emit ultraviolet light. This encounters a thin layer of *phosphors* coating the inner surface of the glass enclosure. The light causes the phosphors to *fluoresce*, emitting a diffuse radiance in the visible spectrum.

The tube or lamp also contains one or more inert gases such as argon, xenon, neon, or krypton at about 0.3% of normal atmospheric pressure. Two electrodes inside the glass enclosure are made primarily from tungsten, which can be preheated to initiate ionization of the gas. Confusingly, both electrodes are often referred to as *cathodes*.

The function of the gas is not to emit light, but to conduct electric current, so that free electrons may encounter mercury atoms, raising their electrons briefly to a higher energy level. When one of these electrons reverts from its unstable energized state to its previous energy level, it emits a photon at an ultraviolet wavelength.

Figure 20-3 provides a diagram showing the interior of a fluorescent tube.

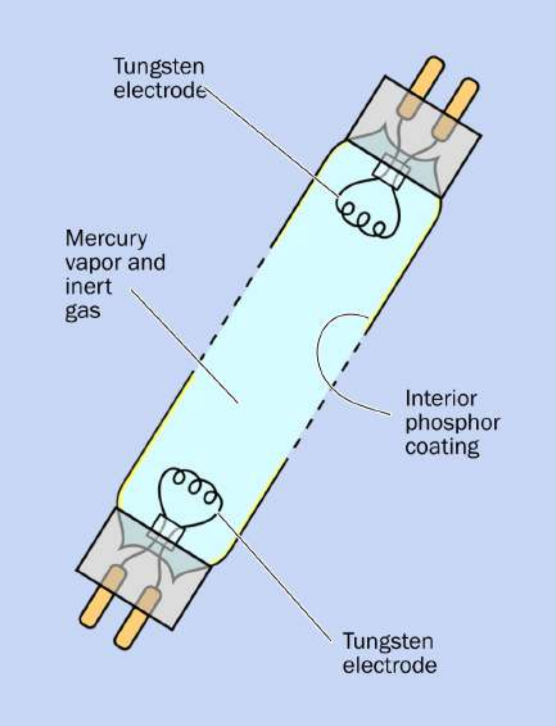

Figure 20-3. *The basic parts of a fluorescent tube.*

Ballast and Starter

Heating the tungsten electrodes is necessary but not sufficient to trigger ionization. A high-voltage pulse is also needed when the light is switched on. In a typical 48" tube, the pulse may range from 200V to 300V.

After current flow has been established, the gas, which is now a plasma, enters a phase of negative resistance. Current passing through it will tend to increase even if the voltage decreases. This process must be controlled to prevent the formation of an *arc*, which will destroy the electrodes. (A similar process occurs in any gas discharge tube, such as a **neon bulb**, and is described in a graph in Figure 19-5.)

To heat the electrodes, ionize the gas, and then control the current, the fixture for a fluorescent tube contains components that are separate from the tube. In their simplest, traditional form, these components consist of a *starter* and a *ballast*. The starter is a neon bulb that contains a bimetallic strip serving as a normally closed switch. It allows current to flow through the electrodes in series, to heat them. The basic circuit is shown in Figure 20-4.

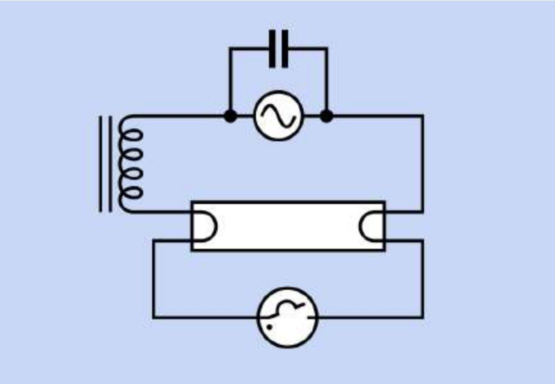

Figure 20-4. *The traditional circuit to trigger ionization of the gas in a fluorescent tube uses a starter (shown at the bottom as a neon tube containing a bimetallic strip, which serves as a switch) and ballast (an inductive load, shown at left).*

The starting process may not be immediately successful, in which case the starter may repeat several times in succession, causing the tube to flicker before its discharge becomes stable. In a cold environment, the tube will have more difficulty starting.

After the tube becomes conductive, current between the electrodes bypasses the starter. At this point, the *ballast* limits the current to prevent an arc from forming. The simplest form of ballast is a coil that functions as an **inductor**.

In a more modern system, an *electronic ballast* replaces the starter-ballast combination. It not only applies the initial surge of high current but also raises the 50Hz or 60Hz frequency of the power supply to 10KHz or more. This increases the efficiency of the tube and eliminates any visible flickering of the light.

All compact fluorescent bulbs (CFLs) contain electronic ballasts. The small components visible in Figure 20-1 are the ballast.

Flicker

When a fluorescent tube uses a conventional ballast and is illuminated with 50Hz or 60Hz AC, the glow discharge stops each time the current flow passes through the zero point in its cycle. In fact, the ionized gas in the tube cannot conduct until it is close to the maximum voltage, and stops conducting when the voltage rolls off. Consequently, the voltage across the tube fluctuates in an approximate square wave, and the light output begins and ends very abruptly. Although this occurs 100 times per second on a 50Hz supply and 120 times per second on a 60Hz supply, some people complain that the flicker is noticeable and can induce headaches.

The rapid on-off discharge is hazardous when it illuminates rotating parts in machinery, as a stroboscopic effect can make the parts seem to be stationary. To mitigate this effect, adjacent tubes in a fixture are powered by separate supplies that are out of phase. This is done either by using a three-phase power supply or by adding an LC circuit to the supply for one of the tubes.

Variants

The traditional type of ballast is also known as a *rapid-start ballast*. By preheating the electrodes, it reduces damage to them that otherwise tends to occur during the starting process. A tube designed for use with a rapid-start ballast has two contacts at each end, and is referred to as a *bi-pin* tube.

An electronic ballast is also known as an *instant-start ballast*. It does not preheat the electrodes, and a tube designed to work with it has only one pin at each end.

CCFLs

A *cold cathode fluorescent lamp* (CCFL) may resemble a miniature fluorescent tube, typically measuring 2mm to 5mm in diameter. The tube may be straight or bent into a variety of shapes. It works on the same principle as a full-size fluorescent tube, containing mercury vapor and one or more inert gases, with an interior coating of phosphors to enable fluorescence. CCFLs are available in many colors and many shades of white.

As its name implies, the electrodes in a CCFL are not heated to establish ionization. Instead, a very high voltage (1,000VAC or more) is applied, dropping to 500VAC to 600VAC after the flow of current has been established. Because CCFLs have been often used to backlight laptop computer screens, inverter circuits are commonly available that create a high-frequency output at a high voltage from an input that can range from 3VDC to 20VDC. The inverter also includes provision to dim the CCFL by using *pulse-width modulation*.

Some CCFLs are designed for illumination of small spaces—for example, the interior of a display case. A few CCFLs look exactly like CFLs and can be used in light fixtures. Some may be compatible with the type of dimmer designed for incandescent lamps.

A CCFL usually has a limited light output compared with that of a conventional fluorescent tube, but has the advantage of working better at low temperatures. Some are designed for signage and exterior lighting in cold-weather locations.

They have a relatively long lifetime of up to 60,000 hours. A hot-cathode fluorescent lamp may fail between 3,000 and 15,000 hours.

Any tube or bulb that uses unheated electrodes to ionize a gas is technically a cold-cathode device, but will not be identified as a CCFL unless it also has an inner layer of phosphors to achieve fluorescence.

It is important to match a tube with the type of ballast installed in a fixture. This is not an issue with CFLs, as they have the appropriate ballast built in.

Sizes

Straight bi-pin tubes are sold in the United States in the following standard sizes:

- T5: 5/8″ diameter. A more modern tube, but still with tungsten electrodes that serve to heat it.
- T8: 1″ diameter. Very often 24″ or 48″ in length, consuming 18W or 36W respectively.
- T12: 1-1/2″ diameter.
- T17: 2-1/8″ diameter.

CFLs are sold in a very wide variety of configurations.

Comparisons

Fluorescent lights have significant advantages and disadvantages. On the plus side:

- After the fixture containing the ballast has been paid for, a tube is relatively cheap. A CFL or an LED light does not have this advantage, as the electronics are built in and will be discarded when the light fails.
- Fluorescent lights have a longer life than incandescent bulbs.
- Fluorescent lights are available in a wide range of shades of white.
- Fluorescent tubes create a diffuse radiance that is ideal for general lighting using ceiling-mounted fixtures. They do not cast harsh shadows.

On the minus side:

- Fluorescents were traditionally more energy-efficient than any other light source, but **LED area lighting** is now more efficient in some designs. LEDs are expected to become more efficient in the future.
- A fluorescent tube with a traditional type of ballast may cause complaints of flickering. By comparison, an LED light uses DC, and an incandescent bulb retains sufficient heat between power cycles so that it does not appear to flicker.
- Fluorescent flicker creates problems when shooting video.
- The fluorescent emission spectrum has sharp peaks that give the lighting an unnatural look.
- In applications that require a defined beam of light, a fluorescent source cannot be used.
- Conventional ballasts can create radio interference, especially in the AM band.
- Because fluorescent tubes and bulbs contain mercury, they require proper disposal, which can incur fees.
- Even an instant-on fluorescent light tends to hesitate briefly when it is switched on.
- The lifespan of a fluorescent light is greatly reduced if it is cycled on and off frequently. An incandescent bulb is less severely affected by cycling, and an LED light is not affected at all.
- Fluorescent lights have difficulty starting at low temperatures.

Values

Brightness

The intensity of a fluorescent light is measured in *lumens per watt*. Because invisible wavelengths are of little interest when assessing brightness, *luminous flux* is used to describe apparent brightness in the visible spectrum. The unit for luminous flux is the *lumen*. Additional information about light measurement is included in the entry describing incandescent lamps (see "Power" on page 177).

Spectrum

The spectrum of photons emitted from mercury vapor in a fluorescent light has wavelengths that peak at 253.7 nanometers and 185 nanometers. (A *nanometer*, customarily abbreviated as nm, is one-billionth of a meter.) These wavelengths are invisible, being in the ultraviolet range, but when the light is transposed into the visible spectrum by the layer of phosphors, "spikes" in the range of wavelengths are still present. For a comparison of output curves for incandescent, fluorescent, and LED lights, see the graph in Figure 18-4.

Various formulations for the phosphors in a tube or CFL attempt to modify the character of the light to suit the human eye, but none of them looks as "natural" as the radiance from an incandescent bulb, probably because the characteristics of incandescent light are very similar to those of sunlight.

What Can Go Wrong

Unreliable Starting

At a low temperature, the mercury inside a fluorescent tube may be slow to vaporize. At very low temperatures, vaporization may not be possible at all. Until the mercury vaporizes, fluorescence will not occur.

Terminal Flicker

As a tube ages, it may start to conduct current only in one direction, causing it to flicker visibly. As it ages more, the gas discharge becomes even less reliable, and the flicker becomes erratic. Eventually, the gas discharge fails completely. In this state, a tube may show only a dim light at each end, in proximity to the tungsten electrodes.

Cannot Dim

Neither the older style of “conventional” ballast nor a modern electronic ballast will respond appropriately to a dimmer of the type designed for incandescent bulbs. This may be an important factor when an incandescent bulb is swapped out for a CFL.

Burned Out Electrodes

Like the tungsten filament in an **incandescent lamp**, the tungsten electrodes in a fluorescent tube suffer progressive erosion. This is evident when a black tungsten deposit forms on the inside of the tube at one or both ends.

Ultraviolet Hazard

Some critics of CFLs maintain that the complex shape of a coiled or zig-zag tube tends to permit small imperfections in the internal phosphor coating, potentially allowing ultraviolet light to escape. If this occurs, and if a CFL is used in a desk fixture in close proximity to the user, ultraviolet light could elevate the risk of skin cancer.

laser

21

The term *maser* was coined in the 1950s to describe a device that used stimulated emission to amplify microwaves. When a device using similar principles amplified visible light in 1960, it was termed an *optical maser*. However, that term is now obsolete, having been replaced with **laser**. This term is always printed in lowercase letters, even though it is an acronym for Light Amplification by Stimulated Emission of Radiation.

The invented verb *to lase* is derived from *laser* and is used to describe the process of generating laser light, with the past participle *lased* and present participle *lasing* sometimes being used.

Thousands of laser variants exist. Because of space limitations, this entry will concentrate primarily on *laser diodes*, which are the smallest, most common, and most affordable type.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)

What It Does

A **laser** generally emits a thin beam of intense light, often in the visible spectrum, and usually in such a narrow range of wavelengths, it can be considered *monochromatic*. The light is also *coherent*, as explained below.

Light output from a laser has three important attributes:

- Intensity. A high-powered laser can deliver energy to a very small, well-defined area, where it may be capable of burning, cutting, welding, or drilling. Large lasers may also be used as weapons, or for power transmission.
- Collimation. This term describes a beam of light that has parallel boundaries, and therefore does not disperse significantly when passing through a transparent medium such as air, glass, or a vacuum. A laser beam can have such excellent collimation, it can be used in precision measuring devices, and has been transmitted over very long distances, even from the Earth to the Moon, where astronauts placed reflectors during the Apollo missions.
- Controllability. Because the beam can be generated with eletrical power, its intensity can be modulated rapidly with relatively simple electronic circuits, enabling applications such as burning microscopic pits in the plastic of a CD-ROM or DVD.

Laser diodes are now more common than all other forms of lasers. They are found in pointers, printers, barcode readers, scanners, computer mice, fiber-optic communications, surveying tools, weapon sights, and directional lighting sources. They are also used as a light source to trigger more powerful lasers.

No generic symbol is used for a laser, but a laser diode is often represented with the same symbol that is used for a light-emitting diode. See Figure 22-2 in the entry for **LED indicators**.

How It Works

A laser is built around a *gain medium*, which is a material that can amplify light. The medium can be a solid, liquid, gas, or plasma, depending on the type of laser.

Initially, an input of energy provides stimulation for some atoms in the gain medium. This is known as *pumping* the laser. The energy input can come from a powerful external light source, or from an electric current.

Stimulation of an atom raises the quantum energy level of an electron associated with the atom. When the electron collapses back to its former energy state, it releases a photon. This is known as *spontaneous emission*.

If one of the photons encounters an atom that has just been excited by the external energy source, the atom may release two photons. This is known as *stimulated emission*. Beyond a threshold level, the number of released photons can increase at an exponential rate.

If two parallel reflectors are mounted at opposite ends of the gain medium, they form a *resonant cavity*. Light bounces to and fro between the reflectors, while pumping and stimulated emission amplifies the light during each pass. If one of the mirrors is partially transparent, some of the light will escape through it in the form of a laser beam. The partially transparent mirror is known as the *output coupler*.

Laser Diode

A laser diode contains an **LED**. (See "How It Works" on page 207 for a more detailed description of the function of an LED.) The p-n junction of the diode functions as the resonant cavity of the laser. Forward bias injects charges into the junction, causing spontaneous emission of photons. The photons, in turn, cause other electrons and electron-holes to combine, creating more photons in the process of stimulated emission. When this process crosses a threshold level, current passing through the diode causes it to *lase*.

The original patent for a laser diode was filed by Robert N. Hall of General Electric in 1962, and the diagram in Figure 21-1 is derived from the drawing in that patent, with color added for clarity.

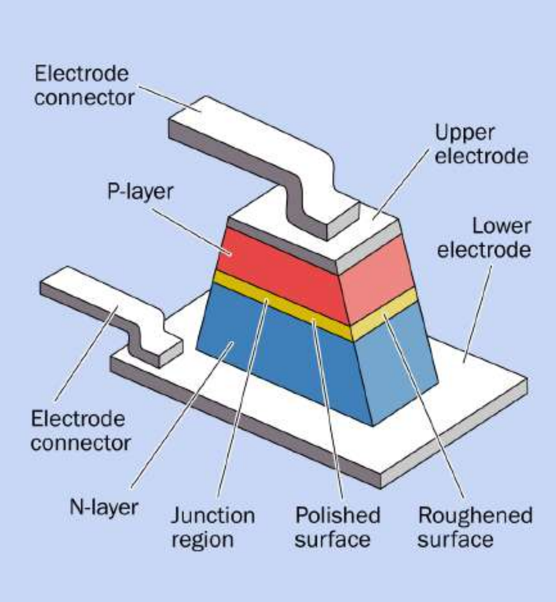

Figure 21-1. *The original design for a laser diode, from the patent filed in 1962.*

In the figure, the junction shown as a yellow layer forms the resonant cavity in which lasing occurs. It measures only 0.1 microns thick (the diagram is not drawn to scale). Its vertical front side is highly polished, and is parallel to the back side, which is also highly polished. Thus, photons reflect between these two vertical sides. The slanted face visible in the figure, and the other slanted face opposite it, are oriented and roughened to minimize internal reflection between them.

Figure 21-2 shows a simplified cross-section of the laser diode.

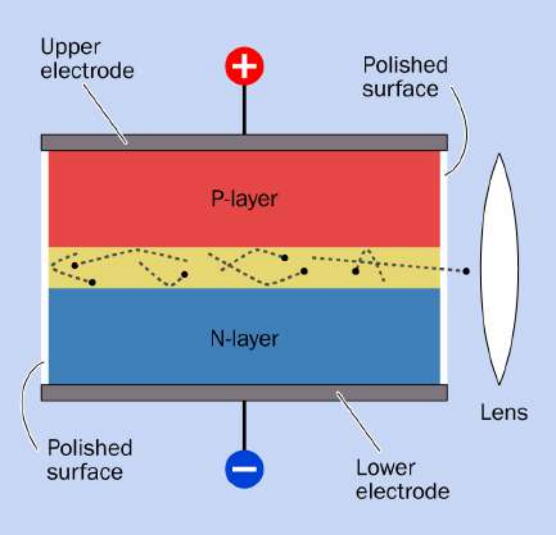

Figure 21-2. *Simplified cross-section of a laser diode.*

Figure 21-3 shows a cross-section of the diode installed in a component sold as a laser. It includes a photodiode to sense the intensity of light emerging through the polished rear end of the laser diode. External electronics are necessary to control the intensity of the laser, using feedback from the photodiode.

The component has three pins (shown pale yellow in the figure), one connecting to the photodiode, another connecting to the p-type layer of the laser diode, and the third being common to the n-type layer of the laser diode and the ground side of the photodiode.

A photograph of a laser diode is shown in Figure 21-4. Note the three pins, comparable to the pins shown in Figure 21-3, indicating that this component requires external control electronics.

In Figure 21-5, a laser is shown with a surface-mount chip adjacent to the solder pad connecting the blue wire. The presence of this chip, with only two wires, indicates that this component has its own control electronics and requires only a DC power supply.

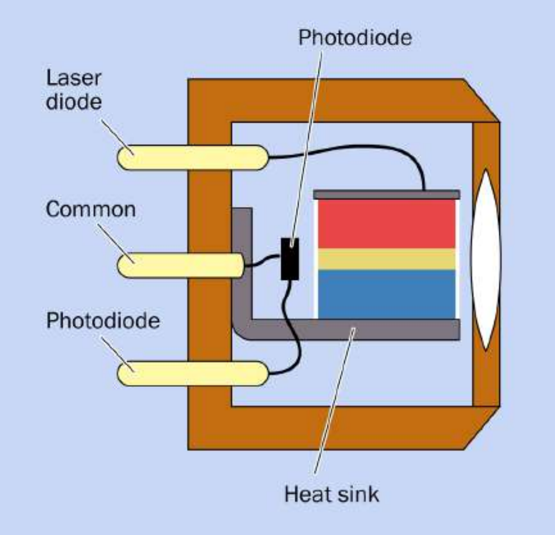

Figure 21-3. *A laser diode is typically mounted with a photodiode to provide feedback for a driver circuit, to control the current consumed by the laser.*

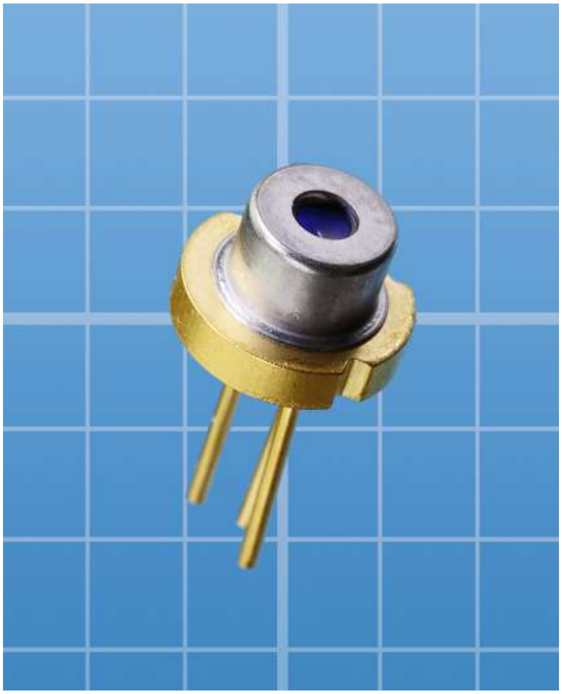

Figure 21-4. *Lite-On 505T laser diode that emits light at 650nm. Power consumption 5mW at 2.6VDC. As indicated by the graph squares, this component is only about 0.2" in diameter.*

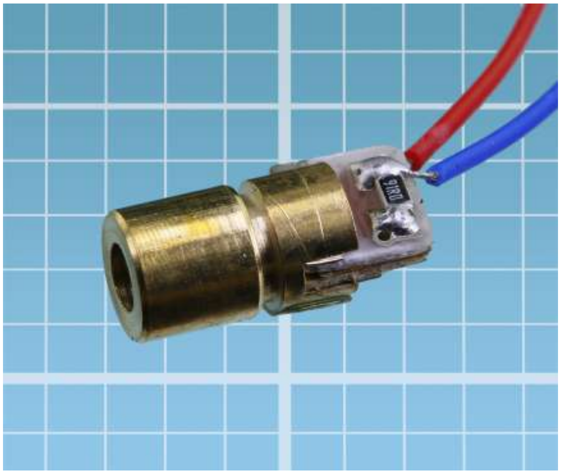

Figure 21-5. *This laser incorporates its own control electronics and requires only a 5VDC power supply. It draws 30mA and generates an output up to 5mW.*

Coherent Light

The emission of *coherent light* by a laser is often explained by suggesting that wavelengths are synchronized with each other. In fact, there are two forms of coherence that can be described approximately as *spatial coherence* and *wavelength coherence*.

If an observer looks up at a cloudy sky, the eye will perceive light radiating chaotically from many distances and directions. Thus, the light is not spatially coherent. The light also consists of many wavelengths, and thus it is not wavelength-coherent.

The filament of an incandescent lamp is a much smaller source of light, but still large enough to generate a profusion of light emissions that are spatially incoherent. The light also includes many different wavelengths.

Suppose a barrier containing a very tiny hole is placed in front of the **incandescent lamp**. If the aperture is very small, an observer on the far side will see the light as a point source. Consequently, the light that emerges from it is now spatially coherent, and will not have chaotically overlapping waves. If the light then passes through a filter, its wavelengths also will become coherent. This is suggested in Figure 21-6 where the light source is an incandescent lamp emitting a wide range of wavelengths.

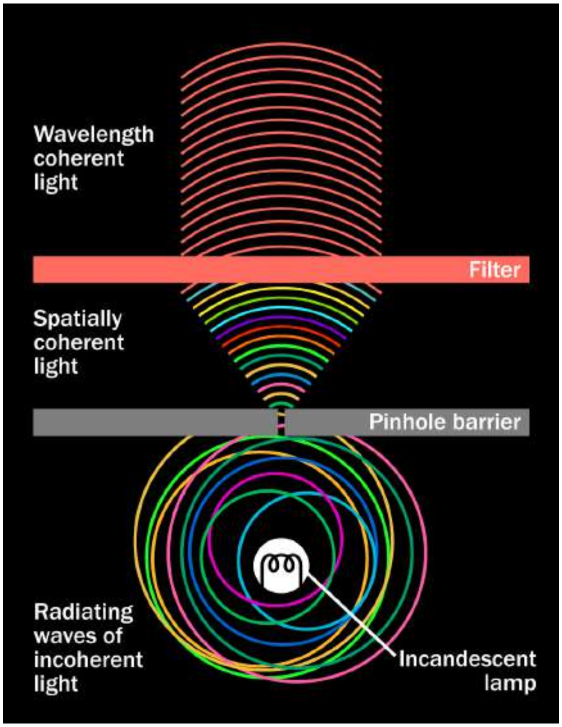

Figure 21-6. *An incandescent lamp, at the bottom of the figure, emits incoherent light at many wavelengths (exaggerated here for clarity). When it passes through a pinhole, it becomes spatially coherent. When it then passes through a colored filter, it becomes wavelength coherent.*

The small amount of light emerging through a pinhole is inevitably much dimmer than light from the original source. A laser, however, amplifies its light output, as well as tending to behave like a point source. The "hall of mirrors" effect of the parallel reflective surfaces in the resonant cavity causes much of the light to shuttle to and fro over a long distance before it emerges through the output coupler. Any light that deviates significantly from the axis of the laser will not escape at all, because the deviations will be cumulative with each reflection. Thus, the light from a laser appears to come from a point source at an almost infinite distance.

Because of the particular geometry of a light-emitting diode, the output from a laser diode is not naturally collimated, and tends to spread by an angle of around 20 degrees. A lens must be used to focus the beam.

Variants

Lasers are sold generally as fully assembled tools for a specific purpose. A very brief summary of CO2, fiber, and crystal lasers is included here.

CO2 Lasers

The gain medium is primarily carbon dioxide but also contains helium and nitrogen, with sometimes hydrogen, water vapor, and/or xenon. The laser is electrically pumped, causing a gas discharge. Nitrogen molecules are excited by the discharge and transfer their energy to the CO2 molecules when colliding with them. Helium helps to return the nitrogen to base energy state and transfer heat from the gas mixture.

CO2 lasers are infrared, and are commonly used in surgical procedures, including ophthalmology. Higher powered versions have industrial applications in cutting a very wide range of materials.

Fiber Lasers

Light is pumped via diodes and amplified in purpose-built glass fibers. The resulting beam has a very small diameter, providing a greater intensity than CO2 lasers. It can be used for metal engraving and annealing, and also for working with plastics.

Crystal Lasers

Like fiber lasers, they are pumped by diodes. These compact lasers are available in a very wide variety of wavelengths, covering the whole visible spectrum, infrared, and ultraviolet. They find applications in holography, biomedicine, interferometry, semiconductor inspection, and material processing.

Values

The output power of a laser is measured in watts (or milliwatts). This should not be confused with the power consumed by the device.

In the United States, any device sold as a laser pointer is limited to a power output of 5mW. However, laser diodes packaged similarly to laser pointers can be mail-ordered with an output of 200mW or more. The legal status of these lasers may be affected by regulations that vary state by state.

In a CD-RW drive that is capable of burning a disc, the diode may have a power of around 30mW. A laser mounted in a CD-ROM assembly is shown in Figure 21-7.

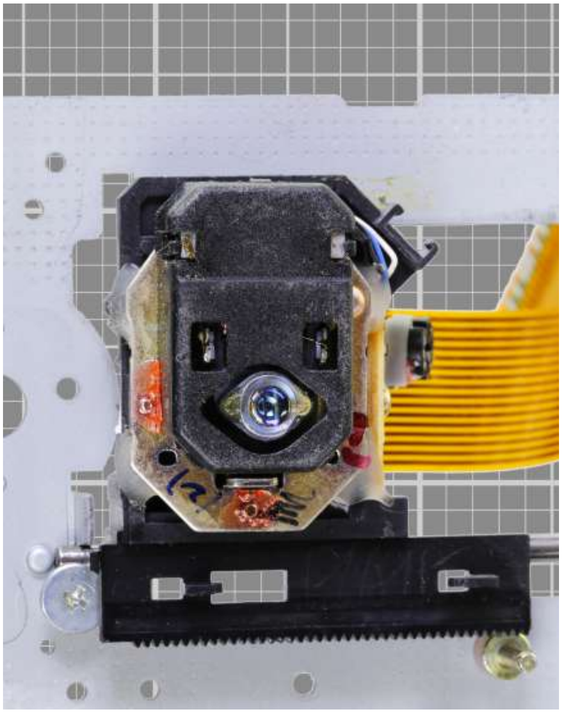

Figure 21-7. *An assembly incorporating a laser for reading a CD-ROM.*

Lasers have such a narrow range of wavelengths, they are given specific output values in nanometers. A laser in an optical mouse may have a wavelength of 848nm; in a CD drive, 785nm; in a bar-code reader, 670nm; in a modern laser pointer, 640nm; in a Blu-ray disc player, 405nm.

How to Use It

While powerful lasers in a laboratory setting have exotic applications, a typical low-power laser diode has become so affordable (costing less than $5 in some instances, at the time of writing) it can be considered merely as a useful source of a clearly defined light beam, ideal for detecting the position of a movable mechanical component or the presence of an intruder.

Generic light-emitting diodes are made with a view angle (i.e., a dispersion angle) as low as 3 degrees, but the beam is soft-edged compared with the precise boundary of a laser beam, and cannot be used reliably in conjunction with sensors at a distance of more than a few inches.

Laser diodes that are sold as components may or may not have current-limiting control electronics built in. Applying power to the laser diode directly will result in thermal runaway and rapid destruction of the component. Drivers for laser diodes are available separately as small, preassembled circuits on breakout boards.

For many applications, it may be simpler and cheaper to buy a laser diode as an off-the-shelf product. A laser pointer provides an easy way to get a source of laser light, and if it would normally be driven by two 1.5V batteries, it can be adapted to run off a 5V supply by using a 3.3V voltage regulator.

Common Applications

In addition to being used with PowerPoint presentations and in conjunction with position sensors, laser pointers have other applications:

- Astronomy. A high-powered laser beam is visible even in clear air as a result of interaction with air molecules. This is known as *Rayleigh scattering*. The phenomenon allows one person to point out a star (or planet) for another person. Because celestial objects are so far away, parallax error is not detectable by two people viewing the beam while standing next to each other. A laser pointer may also be mounted on a telescope to assist in aiming the telescope at an object of interest. This is easier than searching for an object through an eyepiece.
- Target acquisition. Lasers are commonly used on firearms to assist in targeting, especially in low-light conditions. Infrared lasers can be used in conjunction with infrared viewing goggles.
- Survival. A small laser can be included in emergency supplies to signal search teams. A laser can also be used to repel predatory animals.

What Can Go Wrong

Risk of Injury

Lasers are potentially dangerous. Those that have an infrared or ultraviolet output are more dangerous than those with a visible beam, as there is no visual warning that the laser is active. A laser is capable of scarring the retina, although controversy exists regarding the power output that should be considered a high risk.

If a project incorporates a laser, it should be switched off while building or testing the device. It may be advisable to wear protective glasses that block laser light even when an experimenter feels confident that a laser is switched off.

Active lasers should never be pointed at people, vehicles, animals (other than dangerous animals), or oneself.

Inadequate Heat Sink

Lasers may be designed and rated for intermittent use. The burner assembly for a CD-ROM drive, for instance, will be rated for pulsed power, not continuous power. Read datasheets carefully, and provide an adequate heat sink.

Uncontrolled Power Supply

A diode laser that does not have a feedback system in place to control the flow of current can self-destruct.

Polarity

Both the light-emitting diode and the photodiode in a three-pin laser package can be damaged by incorrect polarity of applied power. Pin functions should be checked carefully against datasheets.

LED indicator

22

In this encyclopedia, an **LED indicator** is defined as a component usually 10mm or smaller in diameter, made of transparent or translucent epoxy or silicone, most often containing one *light-emitting diode*. It is purposed as a status indicator in a device, rather than as a source of illumination, and is sometimes referred to as a *standard LED*.

LED indicators that emit infrared and ultraviolet light are included in this entry. LEDs that are designed to illuminate large interior or exterior areas are discussed in a separate entry as **LED area lighting**. They are sometimes described as *high-brightness LEDs* and almost always emit white light.

The term **light-emitting diode** is becoming less common, as the acronym **LED** has become ubiquitous. The acronym does not usually include periods between the letters.

The words "light emitting" are hyphenated here, as they form an adjectival phrase, but in everyday usage the hyphen is often omitted, and no definitive rule seems to exist.

Originally, a standard LED contained only one diode, but may now include multiple diodes, either to emit additional light or to provide a range of colors. In this encyclopedia, a single epoxy or silicone capsule is still considered to be an **LED indicator** regardless of how many diodes it contains. By contrast, any component consisting of multiple separately discernible light-emitting diodes, as in a seven-segment numeral, a 14- or 16-segment alphanumeric character, a dot-matrix character, or a display of multiple characters, is listed in a separate entry as an **LED display**.

OTHER RELATED COMPONENTS

- **LED area lighting** (see Chapter 23)
- **LED display** (see Chapter 24)
- **incandescent lamp** (see Chapter 18)
- **neon bulb** (see Chapter 19)
- **laser** (see Chapter 21)

What It Does

An **LED indicator** emits light in response to a small current, typically around 20mA (but sometimes much less), at a voltage lower than 5VDC. It is usually molded from epoxy or silicone that may be colorless and transparent (often referred to as *water clear*), or colorless but translucent, or tinted and transparent, or tinted and translucent.

The color of the light is initially determined by the chemical compounds used internally, and by their dopants; therefore, a water-clear LED may emit colored light.

Ultraviolet LEDs are usually water-clear. Infrared LEDs often appear to be black, because they are opaque to the visible spectrum while being transparent to infrared.

When an LED indicator is described as being *through hole,* it has leads for insertion into holes in a circuit board. The term does not mean that the indicator itself is meant to be pushed through a hole in a panel, although this may also be done. The LED is cylindrical with a hemispherical top that acts as a lens. The leads are relatively thick, to conduct heat away from the component. A dimensioned diagram of a typical LED measuring 5mm in diameter is shown in Figure 22-1.

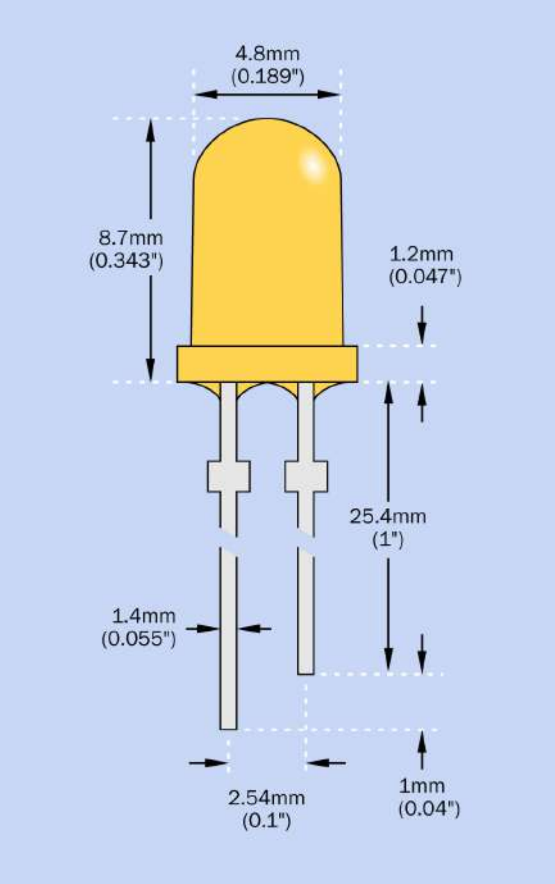

Figure 22-1. *Dimensions of a typical 5mm LED. The longer wire connects with the anode, while the shorter wire connects with the cathode. Adapted from a datasheet published by Lite-On Technology Corporation.*

An LED indicator that is not the through-hole type is usually a surface-mount component. LEDs for surface mounting are mostly rectangular and can be as small as 1mm x 0.5mm. They may require a *heat sink*.

Schematic Symbols

Figure 22-2 shows a variety of symbols that are commonly used to represent an LED. The triangle at the center of each symbol points in the direction of conventional (positive-to-negative) current flow—from the anode to the cathode. Each pair of arrows radiating away from the diode indicates emitted light. Wavy arrows are sometimes used to represent infrared (thermal) radiation. Often, however, an infrared LED is represented in exactly the same style as an LED that emits visible light. With the exception of the wavy arrows, the various styles of schematic symbol are functionally identical and do not identify different attributes of the component such as size or color.

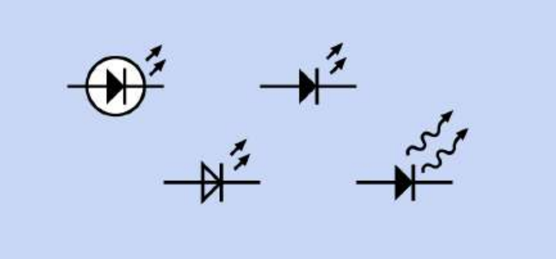

Figure 22-2. *Various symbols may be used to represent an LED. See text for details.*

Common Usage

LED indicators have mostly displaced **neon bulbs** and miniature **incandescent lamps** for the purpose of showing the status of a device. They are found in industrial control panels, home audio systems, battery chargers, washer/dryers, and many other consumer-electronics products. Higher output variants are used in flashlights, traffic signals, taillights on automobiles, and for illumination of subjects that are being photographed. LED indicators may be assembled in large numbers in attention-getting displays such as Christmas lights.

Red, orange, yellow, green, and blue are the basic standard colors. LEDs that appear to generate white light are common, but they do not emit an evenly weighted spectrum of wavelengths. See "Non-Incandescent Sources" on page 174 for a discussion of this topic.

How It Works

An LED, like any **diode**, contains a semiconductor *PN junction* that conducts current only in the forward direction (i.e., from the more-positive side of a power supply to the more-negative side). The diode becomes conductive above a *threshold voltage* sufficient to force electrons in the n-type region and holes in the p-type region to combine with each other. Each time this occurs, energy is released. The energy liberated by one electron-hole combination creates a *photon*, or one quantum of light.

The amount of energy released depends on the *band gap*, which is a property of the semiconductor material. The band gap is the smallest energy that can create an electron-hole pair. The energy determines the light's wavelength, and thus the color.

The band gap also determines the threshold voltage of the LED. For this reason, LEDs of different colors have widely different threshold voltages.

Because an LED will often be used in devices where the DC power supply exceeds the maximum forward voltage, a *series resistor* is customarily used as a simple way to restrict current through the diode.

The light emitted by a colored LED indicator tends to include only a narrow range of wavelengths. However, the addition of a phosphor coating to the diode can broaden the output. This technique is used to make the light from a blue LED appear white, as shown in Figure 22-3. Most white LEDs are actually blue LEDs with a colored phosphor layer added. See the section on **LED area lighting** in Chapter 23 for a more detailed discussion of this topic.

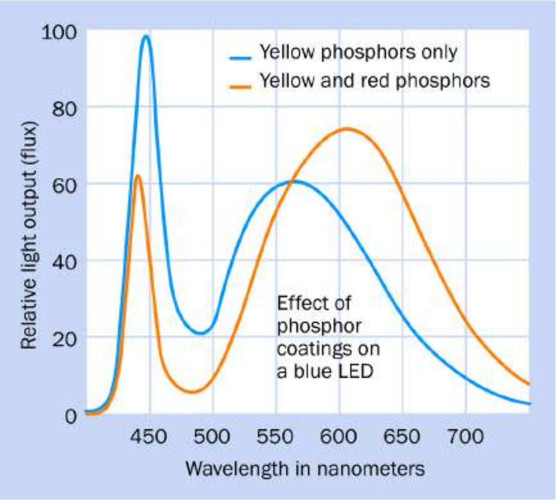

Figure 22-3. *Increasing the range of emitted wavelengths by adding phosphors to a blue LED. Source: Philips Gardco Lighting.*

Multicolor LEDs and Color Mixing

If red, green, and blue light sources are mounted extremely close together, the eye perceives them as a single source, of a color determined by their combined relative intensities. This system of additive color mixing is shown in Figure 17-14 in the entry dealing with LCDs. It is used in LED indicators that contain red, green, and blue light-emitting diodes in a single epoxy or silicone capsule.

While most video monitors use white LEDs or **fluorescent lights** to form a backlight for an **LCD** video screen, some high-end monitors use a matrix of very tiny red, green, and blue LEDs, because the combination of these separate colors generates a wider *gamut* of color wavelengths. The concept of gamut is discussed in "Color" in the **LCD** entry. The tiny LEDs in a backlight cannot be considered as indicators, but indicators are used for this purpose in billboard-sized video displays.

Variants

LED indicators vary widely in size, shape, intensity, view angle, diffusion of light, wavelength of light, minimum and maximum forward voltage, and minimum and maximum forward current.

Size and Shape

The original sizes for round LED indicators were 3mm, 5mm, or (more rarely) 10mm in diameter. Today, through-hole LEDs are sold in many intermediate sizes, although 3mm and 5mm are still most widely used.

The traditional round LED indicator is now augmented with square and rectangular shapes. In a parts catalog, a pair of dimensions such as 1mm × 5mm suggests that the LED is rectangular.

Intensity

The light intensity of an LED is usually expressed in *millicandelas*, abbreviated *mcd*. There are 1,000 mcd in a candela. For more information about units for measurement of light, see "Intensity" on page 178.

The candela measures the *luminous flux*, or visible radiant power, contained within a specified angle of dispersion, usually referred to as the *view angle*. This can be imagined as the rotated angle at the apex of a cone, where the cone defines the "spread" of the light, and the source is at the apex.

If a diode is emitting a fixed amount of luminous flux, the rating in mcd will increase with the inverse square of the view angle. This is because the light delivered to an area in front of the LED will become more intense as the angle becomes smaller. The use of mcd to rate the brightness of an LED can be misleading if it is not considered in comparison with the view angle.

For example, suppose an LED is rated at 1,000 mcd and has a view angle of 20 degrees. Now suppose the same diode is embedded in a different epoxy or silicone capsule with a lens that creates a view angle of only 10 degrees. The LED will now be rated at 4,000 mcd, even though its total power output is unchanged.

- To compare the brightness of two LED indicators meaningfully, they should share the same view angle.

Four through-hole LED indicators with a wide range of specifications are shown in Figure 22-4. From left to right: water-clear white generic, 10mm; Vishay TLCR5800 5mm (emitting red, even though the capsule is water-clear), rated for 35,000mcd with 4 degrees view angle; Everlight HLMPK150 5mm red diffused, rated for 2mcd with 60 degrees view angle; and Chicago 4302F5-5V 3mm green, rated for 8mcd at 60 degrees view angle, containing its own series resistor to allow direct connection with a 5VDC power supply.

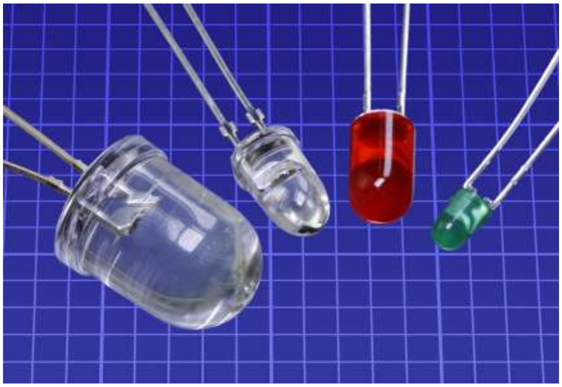

Figure 22-4. *Four assorted LED indicators with very different specifications. See text for details.*

Efficacy

The *radiant luminous efficacy* (LER) of an incandescent light source compares how effective it is at channeling its output within the visible spectrum, instead of wasting it in other wavelengths, especially infrared. Note that the word *efficacy* has a different meaning from *efficiency*. The LER acronym may help to avoid confusion.

LER is expressed in lumens per watt, and in an incandescent bulb it is calculated by dividing the power emitted in the visible spectrum (the *lumi-*

nous flux) by the power emitted over all wavelengths. This is described in detail on "Efficacy" on page 179 in the entry dealing with incandescent lamps.

In an LED indicator, almost all the radiation can be within the visible spectrum, but some power is wasted by generating heat internally. The efficacy varies depending on the type of LED; thus a red-orange indicator can have an efficacy of 98% while a blue LED will be probably below 40%.

Diffusion

Some LED indicators use epoxy or silicone that is formulated to be translucent or "cloudy" instead of transparent. They diffuse the light so that it is not projected in a defined beam, has a softer look, and has an approximately equal intensity when viewed from a wider range of angles.

"Clear" and "diffused" are options that must be taken into account when choosing LEDs from an online catalog, unless the user is willing to turn a clear LED into a diffuse LED by applying some sandpaper.

Wavelength and Color Temperature

The wavelength of light is measured in *nanometers* (abbreviated *nm*), a nanometer being 1 billionth of a meter. The visible spectrum extends from approximately 380nm to 740nm. Longer wavelengths are at the red end of the spectrum, while shorter wavelengths are at the blue end.

A typical LED emits a very narrow range of wavelengths. For example, Figure 22-5 shows the emission from a standard red LED indicator manufactured by Lite-On. Graphs of this type are typically included in manufacturers' datasheets.

Because a red LED stimulates the cones in the eye that respond to red light, it "looks red" even though the color is not comparable with the natural red that is seen, for instance, in a sunset. That natural color actually contains an additional spread of wavelengths.

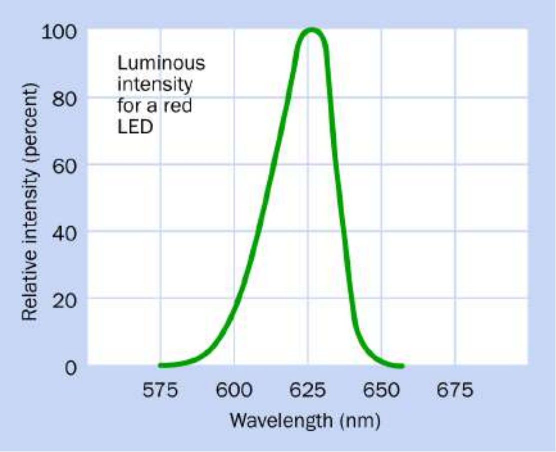

Figure 22-5. *The narrow range of wavelengths emitted by a typical red LED indicator.*

The following list shows the ranges of peak output values, in nanometers, for the most commonly available basic LED indicators (LEDs that emit other wavelengths are available, but they are less common):

- Infrared LED: 850 to 950
- Red LED: 621 to 700
- Orange LED: 605 to 620
- Amber LED: 590 to 591
- Yellow LED: 585 to 590
- Green LED: 527 to 570
- Blue LED: 470 to 475
- Ultraviolet LED: 385 to 405

Figure 22-6 shows this list graphically, omitting infrared and ultraviolet LEDs.

For almost 30 years, blue LEDs were a laboratory curiosity of little practical value, as efficiencies were stuck around 0.03%. An efficiency of more than 10% was finally achieved in 1995. Blue LEDs were marketed soon afterward.

However, when yellow phosphors are added to create the impression of white light by spreading the output over the whole visible spectrum, the wavelengths around 500nm are still not well represented, as suggested in Figure 22-3.

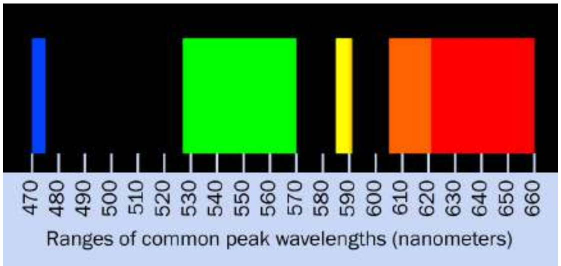

Figure 22-6. *Ranges for peak wavelengths of the most commonly used LEDs. (Source: Survey of approximately 6,000 through-hole LEDs stocked at www.mouser.com.)*

Fluorescent lights perform even more poorly than white LEDs, as can be seen in Figure 18-4 in the entry describing **incandescent lamps**.

Because white LEDs do not emit a single peak of wavelengths, their color is expressed in color temperature rather than nanometers. The concept of color temperature is explained in "Spectrum" on page 173. White LEDs are available rated from 2,800 to 9,000 degrees Kelvin, and are discussed in more detail in the **LED area lighting** entry in this encyclopedia.

Internal Resistor

To eliminate the chore of adding a series resistor to limit current through an LED, some indicators are sold with a series resistor built in. They may be rated for use with 5VDC or 12VDC, but are externally indistinguishable from each other. They are also externally indistinguishable from LEDs that do not contain series resistors. Figure 22-7 shows two 3mm LEDs, the one on the right containing its own series resistor, the one on the left being a generic LED without a series resistor.

Because of the nonlinear response of a diode, LEDs with or without internal resistors cannot be distinguished from each other reliably with a multimeter. If the meter is set to measure ohms, typically it will give an "out of range" error to all types of LED. If it is set to identify a diode, the reading will not tell you if the LED contains a resistor.

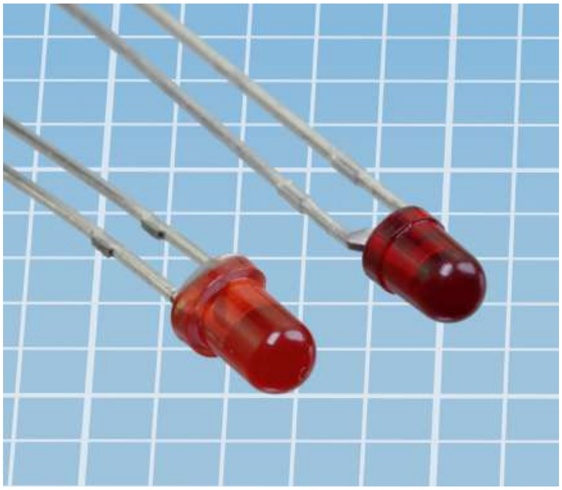

Figure 22-7. *An LED (left) that does not contain its own series resistor is usually indistinguishable from one that does (right).*

One way to determine whether an LED contains an internal series resistor would be to connect it with a variable power supply through a multimeter set to measure mA. Carefully increase the supply voltage from zero until the meter shows a current of 20mA. If the LED does not contain a series resistor, the supply voltage will be close to the recommended forward voltage for that type of LED (no lower than 1.6V for a red LED, and no higher than 3.6V for a white LED). If the LED does contain a series resistor, the supply voltage will be higher. This procedure is time consuming, but may be worthwhile to evaluate multiple LEDs that are known to be identical.

Multicolored

The leads for an LED indicator containing two or more diodes can be configured in several ways:

- Two leads, two colors. Two diodes are mounted internally in parallel, but with opposite polarity.
- Three leads, two colors. Two diodes share a common anode or common cathode.
- Four leads, three colors (RGB). Three diodes share a common anode or common cathode.

- Six leads, three colors. Three diodes, each with its own pair of leads, separate from the others.

Infrared

Most infrared emitters are LEDs that generate wavelengths longer than 800nm. They are found in handheld remotes to control consumer-electronics devices such as televisions and stereo systems, and are also used in some security systems, although *passive infrared motion detectors*, which assess infrared radiation from sources such as people or vehicles, are more commonly used for this purpose.

In conjunction with an infrared emitter, an infrared sensor is necessary, and must be sensitive to the same wavelength. To prevent false positives, the emitter modulates its output, typically with a carrier frequency between 10 and 100kHz. Remotes often use carrier frequencies of 30 to 56kHz. At the receiving end, the signal is processed with a band-pass filter matching the modulation frequency. Many different pulse-coding schemes are used, and no particular standard is dominant.

Ultraviolet

Because ultraviolet radiation can damage the eyes, LED indicators that emit ultraviolet light are potentially dangerous and should be used with caution. A yellow eyeshield can be worn to block the short wavelengths.

Ultraviolet light can be used to cure some adhesives and dental filling material. It can also kill bacteria, and can detect fluorescent print on bank notes, to check for counterfeiting. Ultraviolet flashlights are sold to detect some species of pests, such as scorpions, which fluoresce in response to ultraviolet light.

Values

The specification for an LED will include the wavelength of emitted light, luminous intensity, maximum forward voltage and current, maximum reverse voltage and current, and working values for voltage and current. All these values are important when choosing an indicator for a specific function.

White LEDs for room lighting or external use are calibrated differently. See the entry for **LED area lighting** in Chapter 23.

Forward Current

About half of all the thousands of available types of LED indicators are rated for a typical forward current of 20mA to 25mA. Absolute maximum ratings may be twice as high, but should not generally be applied.

The light intensity of a typical 5mm red LED indicator is plotted against its forward current in Figure 22-8. Note that current and light intensity have an approximately linear relationship up to the typical working current of 20mA. Even above this point, to the absolute maximum of 50mA, the light intensity rolls off only a very small amount.

Although an LED indicator can be dimmed by controlling the current passing through it, the current does not have a linear relationship with the applied voltage, and the indicator will stop functioning completely when the voltage drops below the threshold required by the diode. Consequently, LEDs are commonly dimmed by using *pulse-width modulation*.

Because of the nonlinear response of a diode, LEDs with or without internal resistors cannot be distinguished from each other reliably with a multimeter. If the meter is set to measure ohms, typically it will give an “out of range” error to all types of LED. If it is set to identify a diode, the reading will not tell you if the LED contains a resistor.

Low-Current LEDs

Indicators that require a very low forward current are convenient for direct connection to output pins of logic chips and other integrated circuits. Although a single output from an HC family chip is capable of supplying 20mA without damaging the chip, the current will pull down the output voltage, so that it cannot be used reliably as an input to another chip while also lighting the LED.

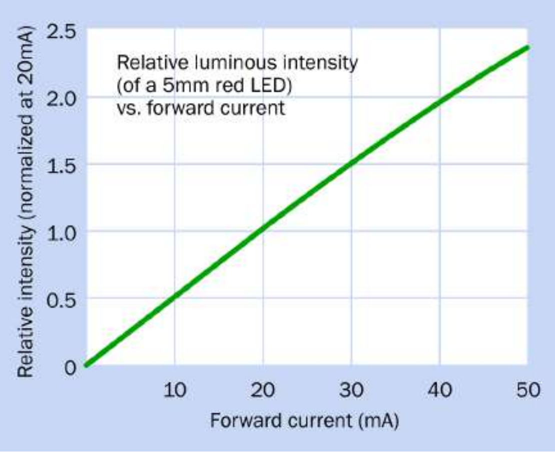

Figure 22-8. *The relationship between forward current and light intensity of a typical 5mm LED indicator is approximately linear up to the 20mA operating current, and almost linear up to the absolute maximum of 50mA.*

Various LED indicators drawing 2mA or 1mA are available, with intensities typically ranging between 1.5mcd and 2.5mcd. This low light output is still bright enough for viewing in a lab-bench environment. Low-current blue LEDs are not available. The only LEDs that draw as little as 1mA are red, as this is the most efficient type.

Using a higher value series resistor with a generic LED will of course reduce its current consumption, and some light will be visible so long as the forward voltage across the LED remains at its minimum level or above.

Forward Voltage

Red is the color that requires not only the least forward current, but the lowest forward voltage. In the range of 1.6VDC to 1.7VDC, all the LEDs are red. Typical forward voltages for various colors are shown here:

- Infrared LED: 1.6V to 2V
- Red LED: 1.6V to 2.1V
- Orange LED: 1.9V to 2.1V
- Amber LED: 2V to 2.1V
- Yellow LED: 2V to 2.4V
- Green LED: 2.4V to 3.4V
- Blue LED: 3.2V to 3.4V
- Ultraviolet LED: 3.3V to 3.7V
- White LED: 3.2V to 3.6V

Color Rendering Index

The *color rendering index* (CRI) evaluates the fidelity with which a light source is capable of displaying the full visible spectrum. It ranges from a perfect score of 100 down to 0 or even lower (sodium-vapor street lighting has a negative value). Computing the index requires standard reference color samples and has been criticized for generating scores that do not correlate well with subjective assessments.

Incandescent bulbs can have a CRI of 100, while an uncorrected white LED may score as low as 80.

Life Expectancy

Because the light output from an LED tends to decrease very gradually with time, the life expectancy is often defined as the number of hours required for the output to diminish to 70% of its output when new. Life expectancy is commonly stated on datasheets for high-brightness white LEDs, but is often omitted from datasheets for LED indicators.

Unlike incandescent lamps and fluorescent lights, LEDs do not have a shorter lifespan if they are frequently cycled on and off.

Light Output and Heat

The light intensity of an LED, measured in mcd, can vary from a few mcd to a maximum of 40,000mcd. Intensities above 30,000mcd generally are achieved by limiting the view angle to 15 degrees or less. Because the candela is weighted toward the central, green segment of the visible spectrum, green LEDs are likely to have a relatively high mcd rating. LEDs rated between 20,000mcd and 30,000mcd, with a view angle of 30 degrees, are almost all green.

Datasheets may often include a *derating curve* showing the lower limit that should be placed on forward current through an LED indicator when its temperature increases. In Figure 22-9, the LED should be operated only within the boundary established by the green line.

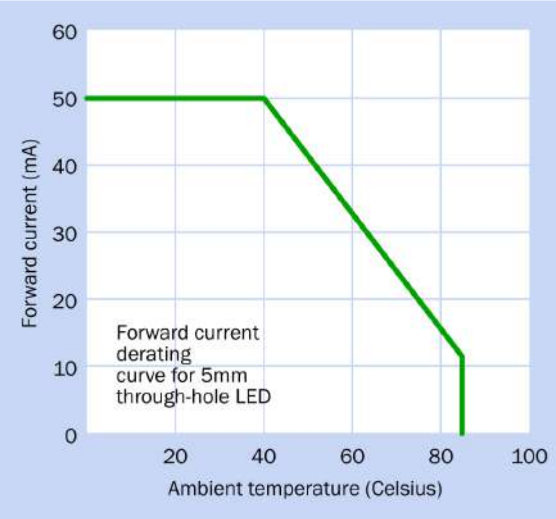

Figure 22-9. *Safe operation of an LED entails limiting the forward current if the temperature increases. The green line shows the boundary for operation of this particular component.*

View Angle

An LED formed from transparent epoxy or silicone (either water-clear or tinted) will create a well-defined beam with a view angle as narrow as 4 degrees or as wide as 160 degrees (in a few instances). The most common view angles for LED indicators are 30 degrees and 60 degrees.

Datasheets for LED indicators often include a *spatial distribution* graph showing the relative intensity of the light when viewed at various angles from the axis of the LED. The spatial distribution graph in Figure 22-10 is for an LED with a view angle defined as 40 degrees. This is the angle at which the relative luminous intensity diminishes to 50%.

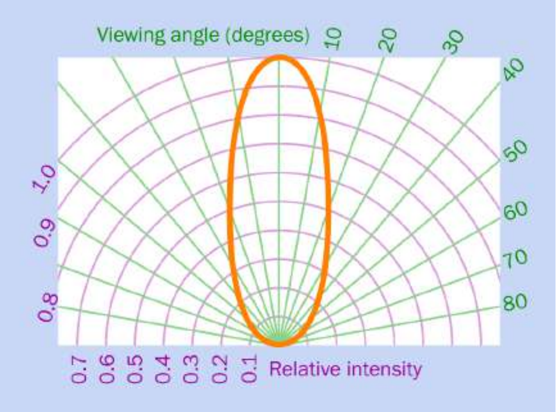

Figure 22-10. *A spatial distribution graph shows the relative intensity of light from an LED at various view angles.*

The view angle is of special concern in devices such as flashlights, where the spread of the beam affects the functionality.

How to Use It

Like all semiconductor devices, LEDs can be impaired by excess forward current and may break down irreversibly if subjected to excess reverse voltage. Their limits for reverse voltage are much lower than those of a rectifier diode. They are also vulnerable to heat, but are not particularly vulnerable to static electricity.

Polarity

A through-hole LED will have two leads of unequal length. The longer lead connects internally with the *anode* of the diode, and should be wired externally to the “more positive” side of a power source. The shorter lead connects internally with the *cathode* of the diode, and should be wired externally to the “more negative” side of a power source.

To remember the functions of the leads, consider that the plus sign would be twice as long as a minus sign if its horizontal and vertical stroke were disassembled and placed end to end.

If a round LED has a flange around its base, a flat spot in the flange will be closest to the cathode side of the component.

Series Resistor Value

Because the effective internal resistance of a diode is not a constant value at different voltages, a trial-and-error approach may be necessary to determine the ideal value for a series resistor with an LED indicator. For this purpose, a trimmer potentiometer can be used with a sample LED while measuring the current through it and the voltage drop across it. A fixed-value resistor can then be substituted. If the choice is between a resistor value that is a little too high and a value that is a little too low, the higher value resistor should be used.

An approximate value can be found using a very simple formula in which R is the resistor value, V_{CC} is the supply voltage, V_F is the forward voltage specified for the LED, and I is the desired current:

$$R = (V_{CC} - V_F) / I$$

Normally a series resistor rated at 1/4 watt will be acceptable, and 1/8 watt may be used in 5VDC circuits. However, care should be taken with a power supply of 9V or higher. Suppose an LED is rated for 1.8V forward voltage at 20mA. In a 5V circuit, the voltage drop across the series resistor will be:

$$V = 5 - 1.8 = 3.2$$

Therefore, the resistor must dissipate 3.2V * 20mA = 64mW. This is comfortably below the 125mW rating of a 1/8 watt resistor. However, with a 9V power supply, the voltage drop across the series resistor will be:

$$V = 9 - 1.8 = 7.2$$

Now the resistor must dissipate 7.2V * 20mA = 144mW. This exceeds the 125mW limit for a 1/8 watt resistor.

LEDs in Parallel

If multiple LEDs are to be driven in parallel, and none of them has to be switched individually, it is naturally tempting to save time by using a single series resistor for all of them. In these circumstances, assess the maximum current carefully and multiply by the voltage drop imposed by each of the LEDs, to determine the wattage of a series resistor.

Linking dissimilar LEDs in parallel is not recommended, because the threshold voltage decreases with increasing temperature. The hottest LED will therefore receive the largest current, and thus become even hotter. Thermal runaway can result.

LEDs containing their own series resistors can safely be wired in parallel.

Multiple Series LEDs

A series resistor wastes current by dissipating it as heat. In an application where two or more LED indicators will be illuminated simultaneously, the LEDs can be connected in series with a lower-value resistor, and three LEDs in series may eliminate the need for a resistor completely, depending on the voltage of the power supply. Here again a trimmer potentiometer should be used to determine an ideal value for any series resistor that may be necessary.

Comparisons with Other Light Emitters

Because LED indicators have largely replaced neon bulbs and miniature incandescent lamps, comparisons are of limited importance at this point. The situation regarding **LED area lighting** is different in that it is still competing actively with fluorescent lights and, in some instances, halogen. A list of advantages and disadvantages for high-intensity white LEDs is given in "Comparisons" on page 223. The advantages of incandescent lamps are listed in "Relative Advantages" on page 179.

Other Applications

LEDs are used in **optocouplers** and in **solid-state relays**. Usually an infrared LED is embedded inside a chip or a plastic module, and emits light through an interior channel to activate a phototransistor. This arrangement provides electrical isolation between the switching signal and the switched current.

Some sensors use an LED paired with a phototransistor at opposite sides of a U-shaped plastic mount. A sensor of this type can monitor industrial processes or may be found inside a photocopy machine, to detect the presence of a sheet of paper.

What Can Go Wrong

Excessive Forward Voltage

Like any diode, the LED has a *threshold voltage* in the forward direction. If this threshold is exceeded, the effective internal resistance of the LED falls very rapidly. Current rises equally rapidly, and quickly damages the component, unless it is protected by an appropriate series resistor.

Excessive Current and Heat

Exceeding the recommended value for forward current, or allowing an LED to overheat, will shorten its lifetime and cause a premature dimming of light output. LEDs generally require some current limiting or regulation (most commonly with a series resistor). They should not be connected directly to a voltage source such as a battery, even if the battery voltage matches the voltage of the diode. The exception to this rule is if the internal resistance of the battery is high enough to limit the current, as in the case of button-cell batteries.

Storage Issues

LEDs of different types are often indistinguishable from each other. They can also be indistinguishable from photodiodes and phototransistors. Careful storage is mandatory, and reusing LEDs that have been breadboarded may cause future problems if they are wrongly identified.

Polarity

If the leads on an LED indicator are trimmed, and if the indicator lacks a flange in which a flat spot will identify the cathode, the component is easily misused with reversed polarity. If it is connected with a component that has limited current sourcing capability (for instance, the output pin of a digital chip), the LED will probably survive this treatment. However, maximum reverse voltage is often as low as 5VDC. To minimize the risk of errors, the anode lead can be left slightly longer than the cathode lead when they are trimmed for insertion in a breadboard or perforated board.

Internal Resistors

As previously noted, it is difficult to distinguish an LED that contains its own series resistor from another LED that does not. The two types should be stored separately, and should be reused circumspectly.

LED area lighting

23

The term **LED area lighting** is used in this encyclopedia to describe a white LED source that is bright enough to illuminate rooms, offices, or outdoor areas. It may also be used in desk lamps or table lamps as *task lighting*. LEDs for these purposes may be categorized as *high-brightness*, *high-power*, *high-output*, or *high-intensity*. A complete fixture containing at least one light source is properly known as a *luminaire*, although the term is not uniformly applied and is sometimes written incorrectly as a *luminary*.

The full term *light-emitting diode* is not normally applied to an LED used for area lighting. For this purpose, the LED acronym has become universal. Periods are not normally placed between the letters.

While an LED area-lighting package may contain more than one diode, it is still categorized here as a single source. By contrast, any component consisting of multiple separately discernible light-emitting diodes, as in a seven-segment numeral, a 14- and 16-segment alphanumeric character, a dot-matrix character, or a display of multiple characters, is listed in a separate entry as an **LED display**.

The term *OLED* is an acronym for *Organic Light-Emitting Diode*, a thin panel in which an organic compound is contained between two flat electrodes. Despite its functionality as a form of LED, its design is similar to that of thin-film electroluminiscent light sources. Therefore it is discussed in the entry on **electroluminescence**.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)
- **incandescent lamp** (see Chapter 18)
- **fluorescent light** (see Chapter 20)
- **neon bulb** (see Chapter 19)
- **electroluminescence** (see Chapter 26)

What It Does

High-brightness white LEDs provide a plug-compatible alternative to **incandescent lamps**, *halogen lighting*, and **fluorescent lights** for work spaces and the home.

At the time of writing, products are still evolving rapidly in the field of LED area lighting. A shared goal of manufacturers is to increase efficiency while reducing retail price to the point where high-brightness LEDs will displace fluorescent tubes for most low-cost lighting applications.

A wall-mounted LED reflector-bulb that emulates a halogen fixture is shown in Figure 23-1. A small LED floodlight for exterior use is shown in Figure 23-2. An early attempt to package an LED area light in a traditional-style bulb is shown in Figure 23-3. Within a decade, as LED area lighting continues to evolve, some of these examples may look quaint. Configurations are evolving, with final results that remain to be seen.

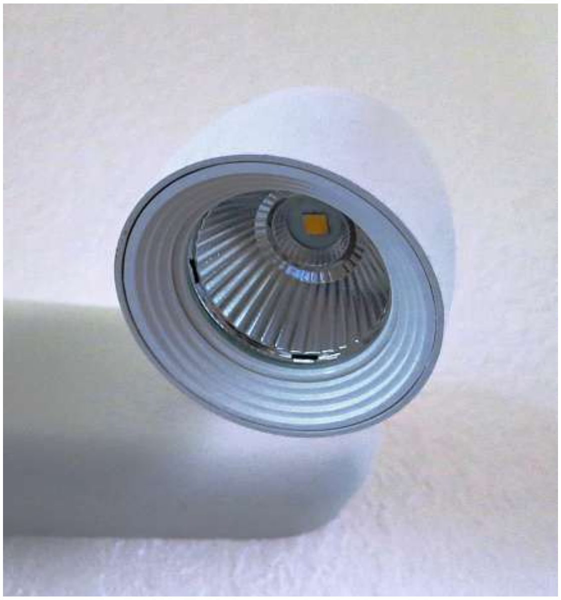

Figure 23-1. *A small LED reflector-light emulating a halogen fixture. Note the square of yellow phosphors mounted on the diode.*

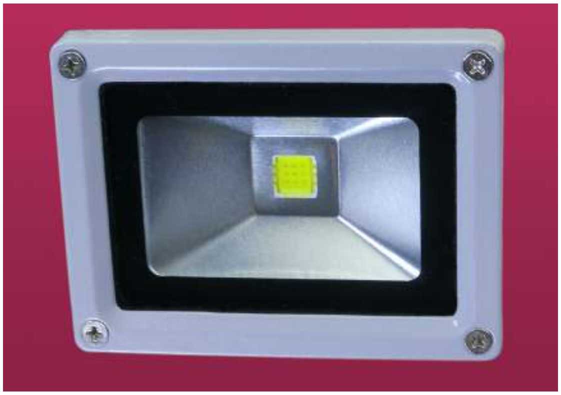

Figure 23-2. *A floodlight for exterior use. Nine LEDs are assembled behind the yellow phosphors. The steel frame measures about 4" by 3".*

Trends in Cost and Efficiency

The *luminous flux* of a source is the total power that it radiates in all directions, within the visible spectrum. The unit that measures luminous flux is the *lumen*. For a more detailed discussion of this topic, see "Power" on page 177. Since 1965, the cost per lumen of light from a given color of LED has decreased by about a factor of 10, while the maximum number of lumens emitted by an LED package has increased by a factor of 20, during each decade. This is known as Haitz's Law, named after Dr. Roland Haitz of Agilent Technologies. Figure 23-4 illustrates it graphically.

Figure 23-3. *An LED light bulb. Unlike an incandescent bulb, it focuses the illumination in one direction, like a reflector-light. Consuming only 6W, it is claimed to be equivalent to a 40W incandescent bulb.*

Schematic Symbol

Schematic symbols that are commonly used to represent an LED are shown in Figure 23-5. The symbol remains the same regardless of the size or power of the component, but architectural plans may represent any type of light using the circle-and-X symbol at bottom right.

How It Works

A high-brighteness LED functions on the same basis explained in the entry describing **LED indicators**. Photons are emitted when electrons are sufficiently energized to cross a PN junction and combine with electron-holes.

An LED that appears white, or off-white, actually emits blue light that is re-radiated over a wide range of wavelengths by adding a layer of yellow phosphors to the chip. A cutaway diagram of an

LED chip (properly known as a *die*), mounted under a silicone lens, is shown in Figure 23-6.

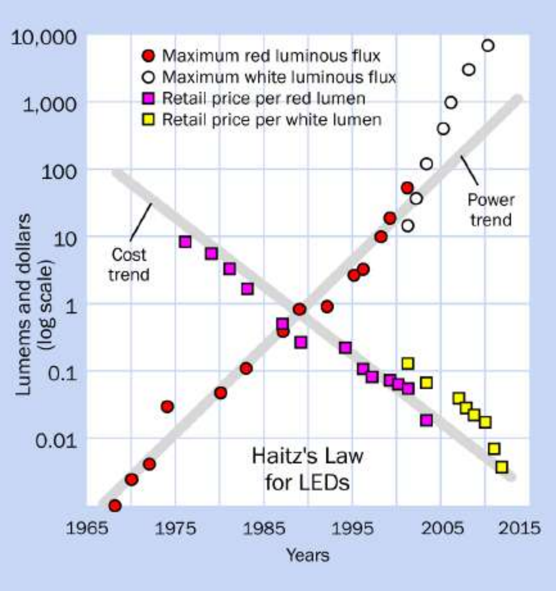

Figure 23-4. *The increase in light output (luminous flux, in lumens) of a single LED compared with the decrease in cost-per-lumen during the years since 1965. The vertical logarithmic scale measures both dollars and lumens. Source: Philips Gardco site-lighting fact sheet with additional data from a "Strategies in Light Report" published by Semiconductor Equipment and Materials International in 2013.*

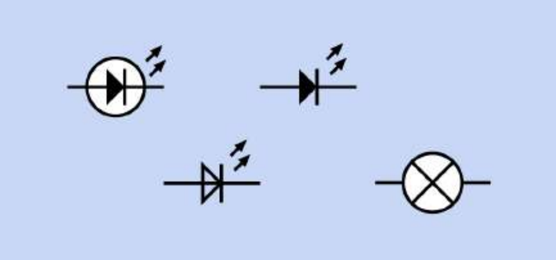

Figure 23-5. *The symbol for an LED remains the same regardless of its size and power, but architectural plans may use the circle-and-X symbol at bottom right for any type of light fixture.*

LEDs are mass-produced by etching them into crystals that are then cut into *wafers* before being subdivided into *dies*, like silicon chips. Most of the blue LEDs that form the basis of white lighting use sapphire crystals as their substrate. The crystal may range in diameter from two inches to six inches. Large sapphire wafers are also finding potential applications in camera lens covers and scratch-resistant cover plates for cellular phones.

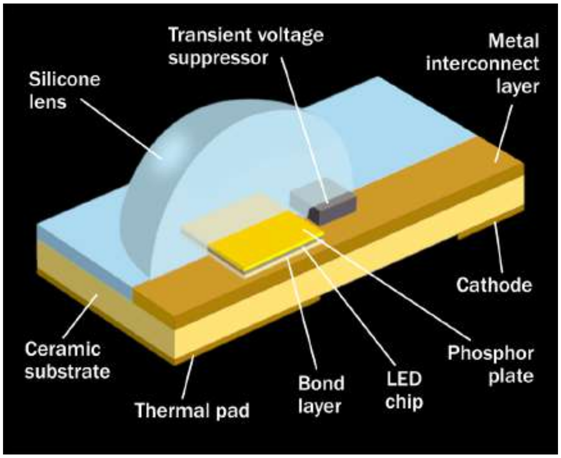

Figure 23-6. *Cutaway diagram of a high-brightness white LED. Adapted from Philips Lumileds Technical Reference document.*

While a die for an LED indicator may be 0.3mm x 0.3mm, a die in a high-brightness LED is often 1mm x 1mm. The size is limited by technical issues involving total internal reflection of the generated light.

The precise color of the light is adjusted by adding red phosphors to the yellow phosphors. This reduces the overall efficiency of the LED by around 10% but creates a "warmer" radiance. A graphical illustration of this principle is shown in Figure 22-3, in the entry on LED indicators.

The *color temperature* of white or offwhite light is measured in degrees Kelvin, typically ranging from 2,500K to 6,500K, where a lower number represents a light with more red in it and a higher number represents a light with more blue in it. This system of measurement was originally used with **incandescent bulbs** to define the temperature of the filament, which determined its color. See "Spectrum" on page 173 for a detailed explanation.

Visible Differences

The effects of different types of illumination are compared in Figure 23-7. To create this figure, first a color chart was prepared in Photoshop and printed on high-gloss photo-grade paper with a Canon Pro9000 Mark II inkjet printer, which has separate colors for red and green in addition to cyan, pale cyan, magenta, pale magenta, yellow, and black.

The color chart was then photographed twice with a Canon 5D Mark II, using a fixed white balance of 4000K. The first exposure was made with "daylight spectrum" LED lighting (claimed color temperature of 6500K) while the second was made with halogen lighting (claimed color temperature of 2900K). The photographs were adjusted in Photoshop for levels only, to fill the available range of 256 values. The two exposures show how the same chart would appear when viewed under the different lights, if the human eye did not adjust itself at all. Note the large area of the LED exposure which is rendered in shades of blue or purple. Also note the dullness of the reds. This confirms the everyday belief that "daylight spectrum" LEDs tend to have a cold, purplish cast while incandescents have a warmer, yellow look.

The same camera was then used to make two more exposures, this time with the white balance set to 6500 for LED lighting and 2900 for halogen lighting, which would be the recommended standard procedure, suggesting the kind of compensation that the human eye also tends to make for different ambient lighting. The result is shown at Figure 23-8. The LED version has improved, but the reds and yellows are still muted. The halogen version also looks better than before, but the magenta end of the spectrum has too much yellow in it. These images show the limits of white-balance correction for indoor photography.

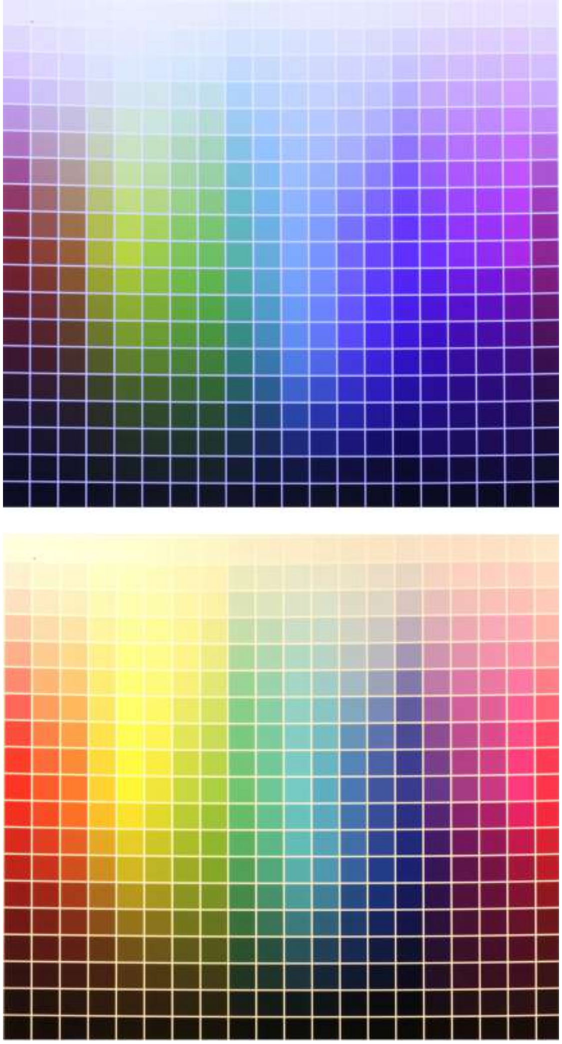

Figure 23-7. *The same printed color chart viewed with "daylight white" LEDs (top) and halogen lighting (bottom), without any compensation to allow for the different light spectra. A fixed white balance of 4000K was used for both pictures.*

Side-by-Side Comparison

Because the human eye is much better at comparing colors when they are adjacent to each other, another color chart was prepared using just six color bars of fully saturated red, yellow, green, cyan, blue, and magenta, with paler and darker versions above and below. The bars were separated with large white gaps. The chart was photographed first with the white balance set to

6500 under "daylight white" LED lighting and then again with the white balance at 2900 for halogen lighting. In Photoshop, the bars from the halogen version were copied and pasted beside the bars from the LED version to facilitate an A-B comparison. The result is shown in Figure 23-9.

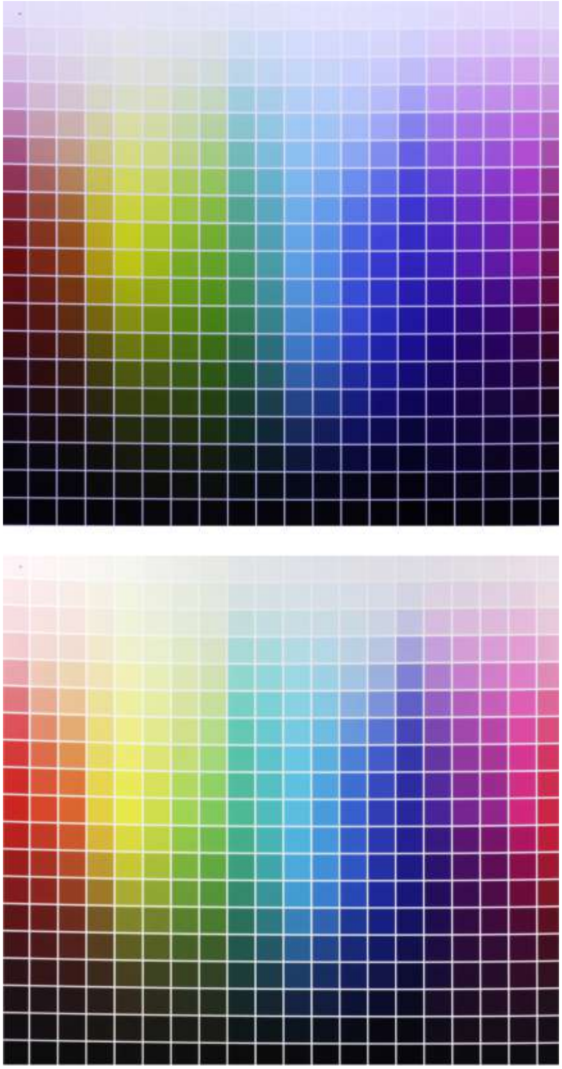

Figure 23-8. *The same color chart as before, photographed with appropriate color temperature settings of 6500K (top, using "daylight-spectrum" LEDs) and 2900K (bottom, using halogen).*

In each pair of colors, LEDs lit the one on the left, and halogen lights lit the one on the right. This figure shows the dramatic difference at the red end of the spectrum, and the poor reproduction of yellows by LED lighting. However, the LED rendering of green is better, and likewise the rendering of magenta, except where its darker version is concerned. Among the pale versions of the colors, the LED lights produce much less density (i.e., they have a brighter look) in the blues, greens, and cyans. The low densities will show up as pale highlights in a photograph of an object, and the picture will tend to have excessive contrast. This will also contribute to the "harsh" look of "daylight white" LED lighting which may be perceived by the eye.

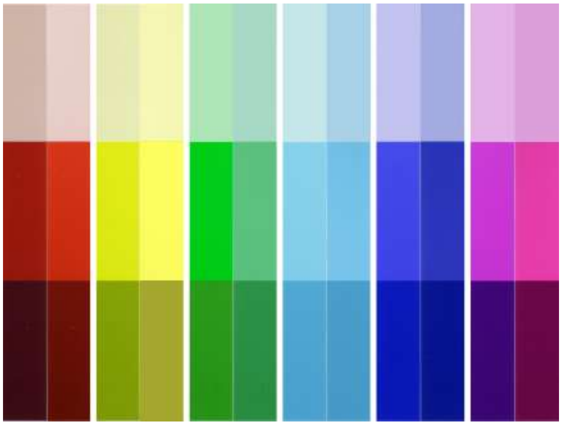

Figure 23-9. *A range of six fully saturated colors, with lighter and darker shades added above and below, photographed first with "daylight spectrum" LEDs and then with halogen lights, after which the two sets of colors were paired for easy comparison. The LED version is on the left in each pair.*

Halogen is deficient at the blue-violet end of the spectrum, even when the camera has an appropriate white-balance setting. Photographers can correct this using image-editing software. LED "daylight spectrum" lights are more difficult to correct. LEDs classified as "warm" should reproduce reds better, but may not do so well with blues.

Diffuse light from a uniformly cloudy sky may be the most ideal form of lighting for photographing objects, but this is of little help for people who work (or take photographs) under artificial lights.

LED lights that contain separate red, green, and blue emitters may perform better, but create a different problem, in that shadows will tend to have color fringes caused by the small offsets between the colored emitters.

Heat Dissipation

An LED is less than 100% efficient because not all electrons mate with electron holes. Some manage to bypass the semiconductor junction; some recombine without generating light; and some transfer their energy to other atoms. In each instance, waste heat is created. While the heat in an incandescent bulb is mostly dissipated by radiation, an LED must get rid of the heat almost entirely by conduction, typically through a *heat sink*. This complicates the design of a fixture, because the integrity of the pathway to dispose of the heat must be retained when the LED bulb or tube is replaced.

Efficacy

The *radiant luminous efficacy* (LER) of an incandescent light source measures how effective it is at channeling its output within the visible spectrum instead of wasting it in infrared radiation. LER is expressed in lumens per watt, and in an incandescent bulb, it is calculated by dividing the power emitted in the visible spectrum (the *luminous flux*) by the power emitted over all wavelengths. This is described in detail in "Efficacy" on page 179 in the entry dealing with incandescent lamps.

In an LED indicator, almost all the radiation can be contained within the visible spectrum, which suggests that its efficacy should be 100%. However, because some waste heat is still created internally, the efficacy is calculated by dividing the light output, in lumens, by the power input, in watts, at the voltage required by the LED. (Lumens can be converted directly to watts, and therefore this division makes a comparison between similar units).

In an LED lighting fixture that contains its own electronics to convert higher voltage AC to lower voltage DC, the power consumption of the fixture is measured not at the diode, but at the input side of the electronics. Therefore, the inefficiency of the electronics reduces the efficacy value of the lighting unit.

Dimming

An incandescent bulb is very sensitive to reduction in power. It becomes radically inefficient, emitting perhaps 1% of its normal light output if the power is reduced to 40%.

LEDs have an almost linear response to the supplied power. Usually a triac-based dimmer will not work well with LED area lighting, and a dimmer designed for LEDs must be substituted, using pulse-width modulation.

Ultraviolet Output

The gas plasma in a fluorescent light generates ultraviolet wavelengths that are shifted to the visible spectrum by the phosphor coating inside the glass envelope. Imperfections in the phosphor coating can potentially allow leakage of ultraviolet light, causing some researchers to claim that the use of CFLs (compact fluorescent lighting) for close-up work with desk lamps can increase the risk of developing some forms of skin cancer. (This claim remains controversial.)

LED manufacturers are quick to point out that white LEDs do not emit any ultraviolet radiation. Figure 23-10 shows spectral power distribution curves derived from measurements of three high-brightness Color Kinetics LEDs manufactured by Philips. The manufacturer states categorically that "The LED-based color and white light products made by Color Kinetics do not emit outside the visible spectrum." Infrared radiation is also negligible.

Color Variation

The *correlated color temperature* (CCT) is determined by finding the conventional incandescent color temperature which looks most similar to the light from a white LED. Unfortunately, because the CCT standard is insufficiently precise, and because small manufacturing inaccuracies can occur, two LED sources with the same CCT number may still appear different when they are side by side. While the human eye adjusts itself to overall color temperature, it is sensitive to differences between adjacent sources. If two or more white LEDs in a lighting fixture do not have identical spectra, the difference will be noticeable.

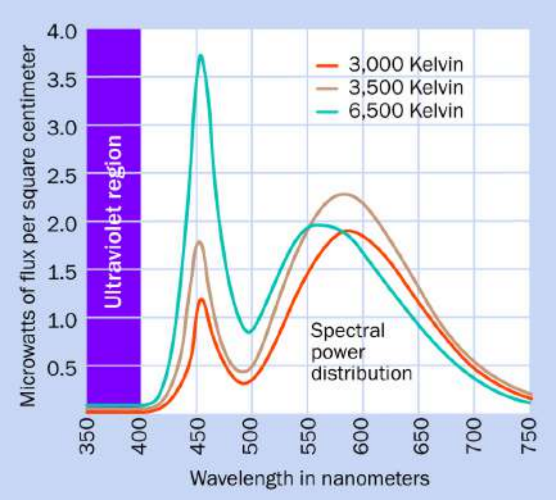

Figure 23-10. *Spectral power distribution curves for three high-brightness white LED lamps showing no ultraviolet emissions. (Adapted from a graph in a white paper published by Color Kinetics Incorporated.)*

To address the issue, manufacturers introduced the concept of "binning," in which lights are subclassified to tighter specifications and are assigned bin numbers based on their measured characteristics. The Philips Optibin system, for instance, assesses the light from various angles, as well as perpendicularly to the source. This is especially important where a large area such as a building lobby is painted white and is lit by high-brightness LEDs that must appear uniform in color temperature.

Variants

LED area lighting products are often designed to emulate the form factors of incandescent bulbs, halogen reflector bulbs, or fluorescent tubes. The standard screw-in base of an LED light bulb, the pin-base of a miniature 12V LED reflector bulb, and the pins on an LED tube enable easy migration to the newer technology.

Strip lights are unique to LED lighting systems. They are thick, flexible plastic ribbons in which are embedded a series of LEDs. For area lighting, the LEDs are white, and the strips can contain necessary control electronics for conversion of AC power. The strips can be placed behind ledges or moldings to provide soft, even illumination of the ceiling above.

Strip lights are also available for 12VDC power, to create lighting effects in customized automobiles and trucks. These strip lights are available in various colors in addition to white. Many have multicolor capability and can be controlled with a handheld remote.

Comparisons

The advantages of an **incandescent lamp** are listed in "Relative Advantages" on page 179, while advantages of **fluorescent lights** are listed in "Comparisons" on page 194. These lists can be compared with the following advantages for LED area lighting:

- While the life of an incandescent lamp for room lighting can be as little as 1,000 hours, LED area lighting typically claims up to 50,000 hours.
- The lifetime of an incandescent bulb is the average time it can emit light before catastrophic failure. The lifetime of an LED is the average time it can emit light before gradually dimming to 70% of its rated output. This is a much gentler, less inconvenient failure mode that does not require immediate replacement.
- Unlike a fluorescent light or incandescent bulb, the LED does not contain hot tungsten that fails as a result of erosion.

- Unlike a fluorescent light, an LED does not contain mercury, and therefore does not require special recycling arrangements that entail associated fees.
- While fluorescents can have difficulty starting in low temperatures, an LED is not sensitive to a cold environment.
- Bright LEDs are available in a wide range of colors that do not require filtering. Filters greatly reduce the efficiency of incandescent bulbs when they are used in applications such as traffic signals or rear lights on automobiles.
- High-brightness LEDs can be dimmable. Fluorescent lights are usually not dimmable, or perform poorly in this role.
- LEDs are inherently directional, because the die radiates light at an angle of 90 degrees to its plane. This makes it ideal for ceiling mounting, where as much light as possible should be directed downward. A fluorescent tube or incandescent bulb often requires a reflector which reduces the overall efficiency.
- LEDs are insensitive to cycling. The life expectancy of an incandescent bulb or (especially) a fluorescent tube is reduced by cycling it on and off.
- No flickering. Fluorescent tubes may start to flicker as they age.
- No electrical interference. Fluorescent tubes can interfere with AM radio reception and some audio devices.
- Safe from breakage. LED area lighting does not necessarily use any glass.

However, high-brightness LEDs still have some barriers to overcome:

- Cost. In the United States, before 60W incandescent bulbs were legislated out of existence, they could be sold profitably for less than $1 each. A T8 fluorescent tube, measuring 1” diameter and 48” long, currently costs between $5 and $6 (retail) but has a life expectancy in the region of 25,000 hours, and uses only 20% of the power of an incandescent bulb to generate two to three times as much light. Clearly the fluorescent tube is a more economical choice, despite the price of the electronics that must be included in the fixture to start the tube. By comparison, currently the purchase price of an LED tube is three times that of a fluorescent tube. It may last twice as long, but is not significantly more efficient, generating perhaps 100 lumens per watt while a fluorescent is typically capable of 90 lumens per watt. Prototype high-brightness LEDs have exceeded 200 lumens per watt, and should be competitive with fluorescents by 2020, but even then, migration will take time.
- Heat sensitivity. Heat reduces the light output and the lifespan of LED fixtures.
- Placement issues. Because LEDs are heat sensitive, they must be installed in locations that do not become excessively hot, their heat sinks must be correctly oriented, and they must have adequate ventilation.
- Color shift. Heat and age may cause the color temperature of an LED to shift slightly, as the color is usually derived from two types of phosphors.
- Nonuniformity. Manufacturing inconsistencies can cause LEDs of the same type to display slightly different color temperatures. Fluorescents and incandescents are more uniform.
- Lower heat output than incandescents. While this is an advantage from the point of view of efficiency, it can be a disadvantage in applications such as traffic signals or airport runway lighting where waste heat can help to keep the lights free from snow or ice.

Values

Although the output from an LED area light is directional, while the output from an incandescent bulb or a fluorescent light is omnidirectional, the intensity is measured the same way in each instance, using *lumens*. This unit expresses the total light emission, without taking directionality into account. (The intensity of **LED indicators** is calibrated in candelas, which measure the power within an angle of dispersion; but candelas are not used for area lighting.)

Typical values for incandescent bulbs are 450 lumens for a power consumption of 40 watts, 800 lumens for a consumption of 60 watts, 1,100 lumens for a consumption of 75 watts, and 1,600 lumens for a consumption of 100 watts. Because much of the output from an incandescent bulb may be wasted by using inefficient reflectors or allowing the light to shine in directions where it is not needed, a high-brightness LED rated at 1,000 lumens may actually appear brighter than a 75-watt incandescent bulb.

A T8 fluorescent tube measuring 48″ long by 1″ in diameter consumes only 32 watts but emits almost 3,000 lumens—when it is new. This output gradually diminishes by as much as 40% over the lifetime of the tube.

Incandescent bulbs deliver between 10 and 15 lumens per watt, approximately. A new fluorescent tube produces around 80 to 90 lumens per watt, and LED area lighting at the time of writing can provide 100 lumens per watt, under real-world conditions.

What Can Go Wrong

Wrong Voltage

Many high-brightness LED lighting units can be used with either 115VAC or 230VAC. There are exceptions, however. Check the specifications to make sure. Also, it is important to avoid applying domestic supply voltage to 12V LED miniature reflector-bulbs that are intended to replace 12VAC halogen bulbs of the same size.

Overheating

If a high-brightness LED fixture is equipped with a heat sink, this must be exposed to freely flowing air. Any vanes on the heat sink should be oriented vertically to encourage convection, and the fixture must not be placed in an enclosure. Overheating will radically shorten LED life.

Fluorescent Ballast Issues

A fluorescent fixture contains a *ballast* to limit the tendency of the tube to draw excessive current. The ballast is contained in a plastic box attached to the back of the frame in which the tube is mounted.

A *magnetic ballast* contains a coil, and is bypassed by an additional *starter* that applies unlimited current for one second when the power is switched on, preheating the tube to initiate plasma discharge.

An *electronic ballast* performs the same function without a separate starter.

Some LED tubes designed as substitutes for fluorescent tubes may allow a magnetic ballast to remain in the circuit, but may not tolerate an electronic ballast. Other LED tubes require any type of ballast to be unwired from the circuit. The unwiring operation will require disconnection of a couple of wires by removing wire nuts (assuming that the fixture has been designed to comply with U.S. building codes). The wires are then reconnected to apply power directly to the tube, and the wire nuts are reapplied to complete the new connection. The ballast can remain passively in the fixture.

Failing to remove the ballast and/or the starter from a fluorescent fixture before installing an LED tube that requires direct connection to the power supply can damage the tube. Connecting the power incorrectly to the LED tube may result in it failing to light up. Documentation supplied with the LED tube should provide guidance for disconnecting the ballast and connecting the tube. Note that the pin functions on LED tubes are not standardized at this time.

Misleading Color Representation

Because the spectrum of a white LED is not evenly weighted across all wavelengths, it will fail to represent some colors accurately, as shown previously. This can be important if LEDs are used to illuminate full-color printing or artwork, or if they are installed in stores selling merchandise such as clothes, furnishings, or food.

LED display

24

In this encyclopedia, a component consisting of multiple separately discernible light-emitting diodes, such as a seven-segment numeral, 14- or 16-segment alphanumeric character, a dot-matrix character, or a display module containing multiple characters, is categorized as an **LED display**. The term *light-emitting diode* is hardly ever used to describe an LED display, as the LED acronym has become ubiquitous. The acronym does not usually include periods between the letters.

An **LED indicator** is defined here as a component usually 5mm or smaller in diameter, made of transparent or translucent epoxy or silicone, most often containing one *light-emitting diode*. It is purposed as a status indicator in a device, rather than as a source of illumination, and is sometimes referred to as a *standard LED.*

LEDs that are designed to illuminate large living or working areas are discussed in a separate entry as **LED area lighting**. They are sometimes referred to as *high-brightness LEDs* and almost always emit white light.

The term *OLED* is an acronym for *Organic Light-Emitting Diode*, a thin panel in which an organic compound is contained between two flat electrodes. Despite its functionality as a form of LED, its design is similar to that of thin-film electroluminiscent light sources. Therefore it is discussed in the entry on **electroluminescence**.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)
- **LED area lighting** (see Chapter 23)
- **vacuum-fluorescent** (see Chapter 25)
- **electroluminescence** (see Chapter 26)
- **LCD** (see Chapter 17)

What It Does

An LED display presents information on a panel or screen by using multiple segments that emit light in response to a DC current, almost always at a voltage ranging between 2VDC and 5VDC. The display may contain alphanumeric characters and/or symbols; simple geometrical shapes; dots; or pixels that constitute a *bitmap*.

A *liquid-crystal display*, or **LCD**, serves the same purpose as an LED display and may appear very similar, except that a liquid crystal reflects incident light while an LED emits light. The increasing use of backlighting with LCDs has made them appear more similar to LED displays.

There is no schematic symbol to represent an LED display. Where a segmented display is used, often the segments are represented with drawn outlines.

The simplest, most basic, and probably the best-known example of an LED display is the *seven-*

segment numeral, one of which is shown in Figure 24-1. This is a Kingbright HDSP-313E with a character height of 0.4".

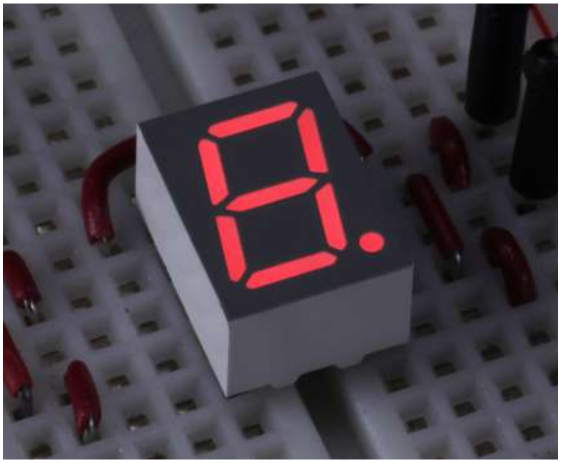

Figure 24-1. *The most basic LED display, able to create numerals from 0 through 9 using seven light-emitting segments that can be illuminated individually. An eighth segment forms the decimal point.*

How It Works

The process by which an LED generates light is explained in "How It Works" on page 207, in the entry dealing with **LED indicators**. Each light-emitting diode in an LED display is functionally the same as the diode in an LED indicator.

LEDs must be driven with DC. This is a primary distinction between an LED display and an LCD, which requires AC.

Variants

LCD comparisons

LCDs and LED displays can look very similar. This raises the obvious question: which is appropriate for a particular application?

LCDs (without backlighting) are more appropriate for applications such as digital watches and solar-powered calculators where power consumption must be minimized. They are capable of running for years from a single button cell.

LCDs are easily visible in bright ambient light, where LED displays are not. LCDs can also be designed to display complex pictographic shapes and symbols, while the segments of an LED display are more constrained to be simple in shape.

An LCD is more likely to be affected by temperature than an LED, and powering it entails some slight inconvenience, because it requires an AC source that is unlikely to be useful elsewhere in a circuit. If the LCD uses LED backlighting, it will also require a low-voltage DC power source for the backlight. An LED display is easier to use in that it can be driven directly from a microcontroller or logic chip, with only some series resistors to limit the current, and the addition of transistors to provide additional power where necessary.

Seven-Segment Displays

Early seven-segment LED displays were used in digital calculators, before LCDs became an affordable, practical alternative that greatly extended battery life. Initially, the size of the diodes was limited, sometimes requiring magnifying lenses to make them legible.

Seven-segment displays are still used in some low-cost applications, although LCDs have become more common.

Figure 24-2 shows how the segments are identified with letters *a* through *g*. This scheme is used universally in datasheets, and is also used for LCDs. The decimal point, customarily referred to as "dp," is omitted from some displays. The segments are slanted forward to enable more acceptable reproduction of the diagonal stroke in numeral 7.

Although seven-segment displays are not elegant in appearance, they are functional and are reasonably easy to read. They also enable the representation of hexadecimal numbers using letters A, B, C, D, E, and F (displayed as A, b, c, d, E, F because of the restrictions imposed by the small number of segments), as shown in Figure 24-3.

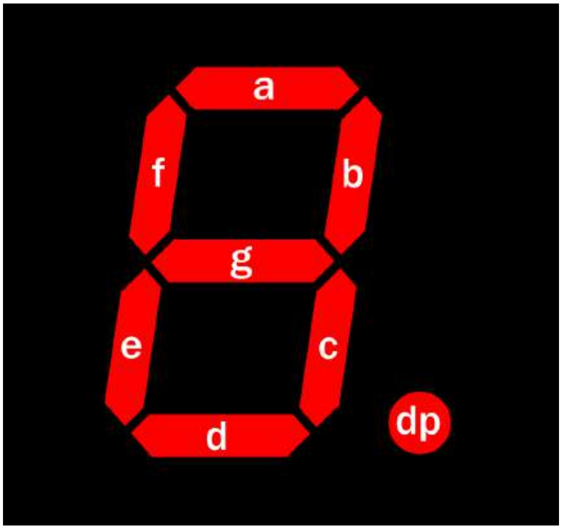

Figure 24-2. *A seven-segment LED display. The lowercase identifying letters are universally used in datasheets.*

In appliances such as microwave ovens, very basic text messages can be displayed to the user within the limitations of seven-segment displays, as suggested in Figure 24-4.

Numbers 0, 1, and 5 cannot be distinguished from letters O, I, and S, while letters containing diagonal strokes, such as K, M, N, V, W, X, and Z, cannot be displayed at all.

Multiple Numerals

Displays consisting of a single numeral are now rare, as few applications require only one digit. Displays of two, three, and four digits are more common, as shown in Figure 24-5.

Additional Segments

Displays with 14 or 16 segments were introduced in an effort to enable the representation of all the letters of the alphabet. The segment layout of these LED displays is identical to that of comparable LCDs. The differences between 14-segment and 16-segment displays are shown in Figure 24-6. Some are angled forward like seven-segment displays, even though the addition of diagonal segments makes this unnecessary for display of characters such as numeral 7.

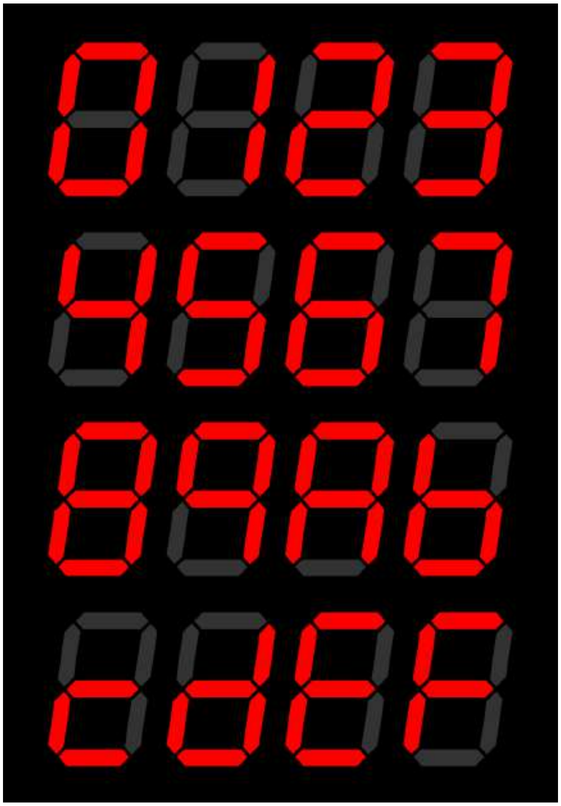

Figure 24-3. *Numerals and the first six letters of the alphabet created with seven-segment displays.*

Figure 24-7 shows the scheme for identifying the segments of a 16-segment display. This naming convention is used in all datasheets. The lowercase letters that were customary with seven-segment displays are usually abandoned in favor of uppercase, perhaps to avoid confusion with the letter L. Note that letter I is omitted from the sequence.

For a complete alphanumeric character set enabled by a 16-segment display, see Figure 17-9 in the entry discussing LCDs.

An example of a 16-segment alphanumeric LED display is shown in Figure 24-8, mounted on a breadboard and wired to show the letter N. This is a Lumex LDS-F8002RI with a character height of 0.8”. The component is still available at the time of writing, but in limited quantities.

Generally speaking, 16-segment displays were never very popular, because the gaps between adjacent segments impaired legibility. LED versions remain more readily available than LCD versions, but dot-matrix displays allow a better-looking, more easily legible alphabet, with the added possibility of simple graphics.

Dot-Matrix Displays

In the 1980s, some personal computers used a video character set in which each letter, numeral, punctuation mark, and special character was formed on a video screen from a fixed-size matrix of dots. A similar alphabet is now used in LED dot-matrix displays (and LCDs, as shown in Figure 17-10).

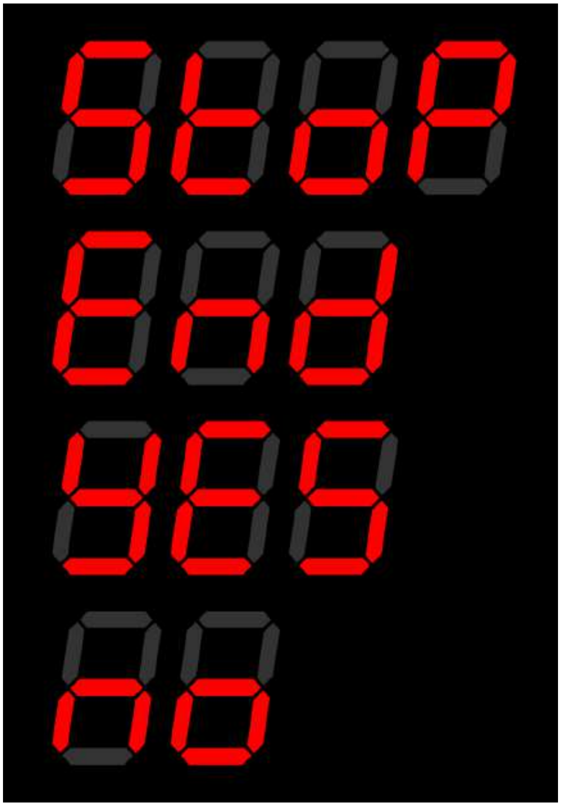

Figure 24-4. *Basic text messages can be generated with seven-segment displays, although they cannot represent alphabetical letters containing diagonal strokes.*

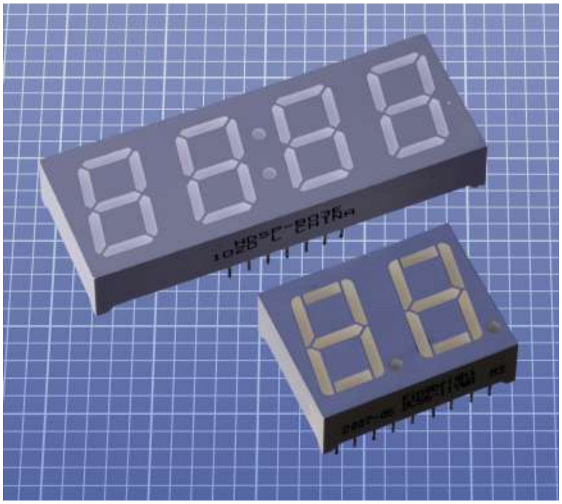

Figure 24-5. *Multiple seven-segment LED displays are often combined in a single component. Top: An Avago 2.05VDC 20mA display designed for a clock. Bottom: A Kingbright two-digit display which draws 20mA at 2.1VDC. The unlit outlines of the numerals would normally be hidden behind panels that are tinted to the same colors emitted by the LED segments when lit.*

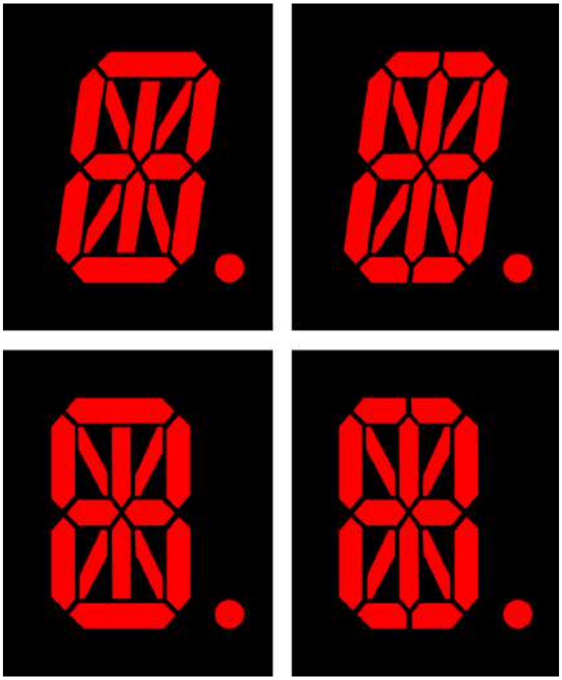

Figure 24-6. *Layouts for 14-segment and 16-segment alphanumeric LEDs are identical to those of LCDs.*

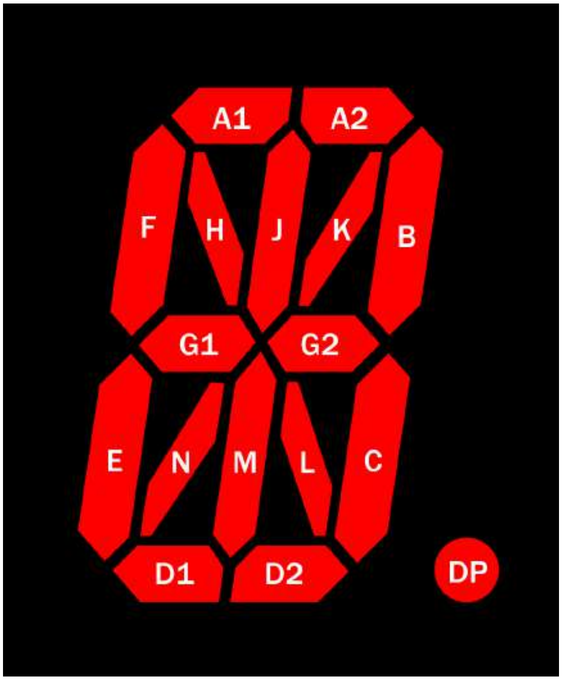

Figure 24-7. *The scheme for identification of segments in a 16-segment alphanumeric LED display.*

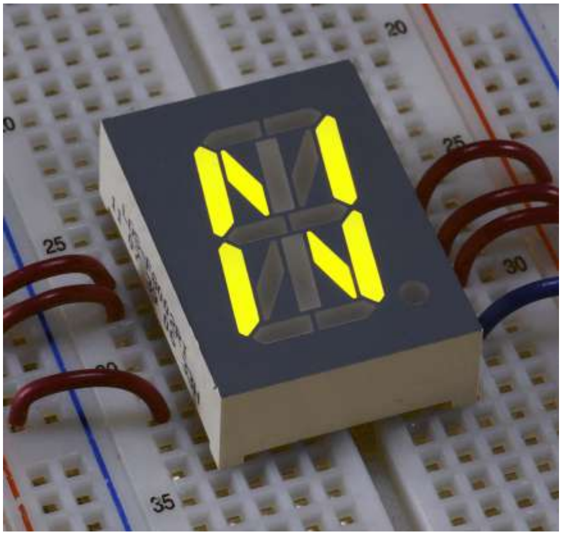

Figure 24-8. *A 16-segment alphanumeric LED display showing the letter N.*

Alphanumeric dot-matrix characters are often grouped in two or more rows with eight or more characters per row. The number of characters is always listed before the number of rows, so that an 8x2 display would contain eight alphanumeric characters in two horizontal rows. This type of component is properly described as a *display module*.

Display modules are used in consumer electronics products such as a stereo receiver where simple status messages and prompts are necessary —for example, to show the tone control settings or the frequency of a radio station. Because the cost of small, full-color, high-resolution LCD screens has been driven down rapidly by the mass production of cellular phones, and because these high-resolution screens are much more versatile, they have already displaced dot-matrix display modules in many automobiles and are likely to follow a similar path in other devices.

Pixel Arrays

The 8x8 pixel array of LED dots shown in Figure 24-9 measures 60mm square (slightly more than 2″) and contains 64 LEDs, each approximately 5mm in diameter. Similar arrays are available in other sizes and with different numbers of dots. Displays of the same type may be assembled edge-to-edge to enable scrolling text or simple graphics.

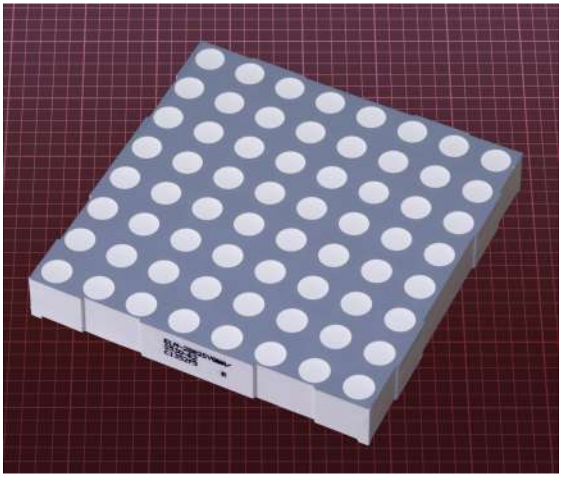

Figure 24-9. *An 8x8 matrix of LED dots measuring more than two inches square.*

Multiple Bar Display

A bar display is a row of small rectangular LEDs in a single component. It may be used for digital representation of an analog signal. The higher the voltage of the signal, the more bars will be illuminated. A typical application would be to show the signal strength of an input to an audio recorder. Ten bars are often used, as in the display shown in Figure 24-10, but multiple components can be combined end to end.

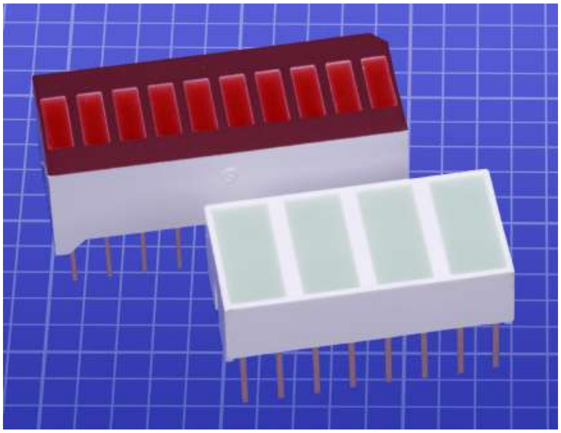

Figure 24-10. *Two LED bar displays in which segments can be lit individually.*

Single Light Bar

A light bar can be thought of as a single-source LED, as it is configured as a single square or rectangle. It is mentioned here, rather than in the entry for single-source LED indicators, because variants may be subdivided into two, three, four, or (sometimes) more discrete sections. These variants are often included in the same datasheet as the monolithic version.

A light bar contains multiple LEDs (often, four in number) behind a translucent panel that provides evenly diffused radiance.

Values

The values for most LED displays are basically the same as for LED indicators, in terms of color, brightness, current consumption, and voltage. See "Values" on page 211 for information.

Multiple-character dot-matrix LED display modules may have different requirements for forward voltage and forward current, depending on drivers that are incorporated in the module. Because there is no standardization for these modules, it will be necessary to consult the manufacturer's datasheet.

How to Use It

Seven-Segment Basics

The diodes in a seven-segment LED display share either a *common anode* or a *common cathode*, the latter being more frequently used. The two types of internal wiring are provided for convenience only. Externally, the displays function identically.

A schematic suggesting the internal wiring and pinouts of a typical ten-pin common-cathode display is shown in Figure 24-11. The pins are numbered as seen from above. Appended to each number is the identity of the segment to which it is connected. Pins 3 and 8 are connected with the cathodes of all the internal LEDs. Both of these pins should be used, to serve as heat sinks for the display.

Note that series resistors are not included inside the display and must be added externally. Their value will be determined by the power supply, to limit the forward current and forward voltage through the LEDs to the extent specified by the manufacturer.

An encapsulated *resistor array* containing either seven or eight resistors in an SIP or DIP chip can be used instead of individual resistors. A seven-segment LED display would require the type of resistor array in which both ends of each resistor are accessible.

Where two or more numerals are combined in a single component, this type of display is likely to have two horizontal lines of pins. In this case, pin 1 will be at the bottom-left corner, seen from above. As always, the pins are numbered counterclockwise, seen from above.

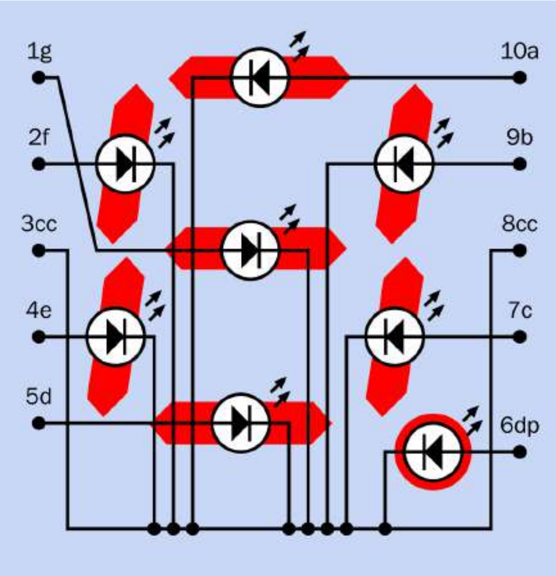
1g
2f
3cc
4e
5d
10a
9b
8cc
7c
6dp

Figure 24-11. *A schematic view of internal connections and pinouts of a seven-segment common-cathode LED display. The numbers refer to the pins of the component, seen from above. The 1 pin may also have a mark beside it on the component, for identification. The orientation of the display can be deduced if there is a decimal point, as this should be at the bottom-right corner.*

Where three or more numerals are combined in a single component, the pinouts may be designed for multiplexing rather than individual access to every segment of each numeral. A four-digit clock display, for instance, may have seven pins that connect in parallel to respective segments in all of the numerals, and four additional pins that can ground each numeral in turn, so that they can be selected sequentially.

Driver Chips and Multiplexing

Illuminating the appropriate segments in a single numeral can be done directly from a microcontroller, or through a driver chip such as the well-known and widely used 4543B that converts a binary-coded decimal input into appropriate segment output patterns. The chip can source sufficient current to drive each segment through a series resistor. Its pinouts are shown at Figure 24-12.

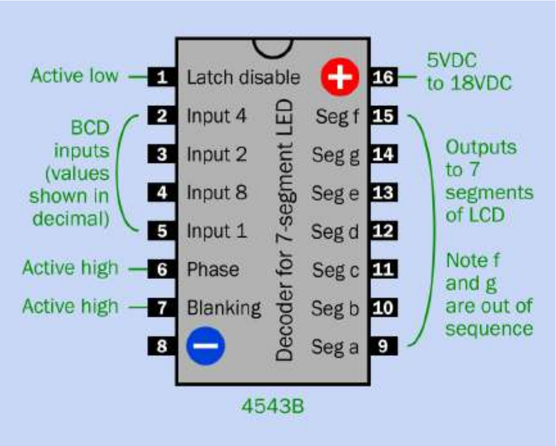
Active low
1
Latch disable
16
5VDC to 18VDC
BCD inputs (values shown in decimal)
2
Input 4
3
Input 2
4
Input 8
5
Input 1
Decoder for 7-segment LED
Seg f
15
Seg g
14
Seg e
13
Seg d
12
Seg c
11
Seg b
10
Seg a
9
Outputs to 7 segments of LCD
Note f and g are out of sequence
Active high
6
Phase
Active high
7
Blanking
8
4543B

Figure 24-12. *Pinouts of the 4543B seven-segment LED driver chip.*

When used in conjunction with a microcontroller, the 4543B can drive several seven-segment displays by multiplexing them. The basic schematic to achieve this is shown in Figure 24-13, omitting optional features such as leading-zero blanking or connections for a decimal point. The microcontroller sends the binary code for the first numeral and simultaneously grounds the common cathode of that numeral through a transistor, which is needed because as many as seven segments of the numeral may be passing current in parallel. The microcontroller then sends the binary code for the second numeral, and grounds it; then sends the binary code for the third numeral, and grounds it; and the cycle repeats. So long as this process is performed at sufficient speed (at least 50Hz), persistence of vision will create the illusion that all the numerals are active simultaneously. The circuit can be compared with a similar circuit to drive LCDs, shown in Figure 17-17.

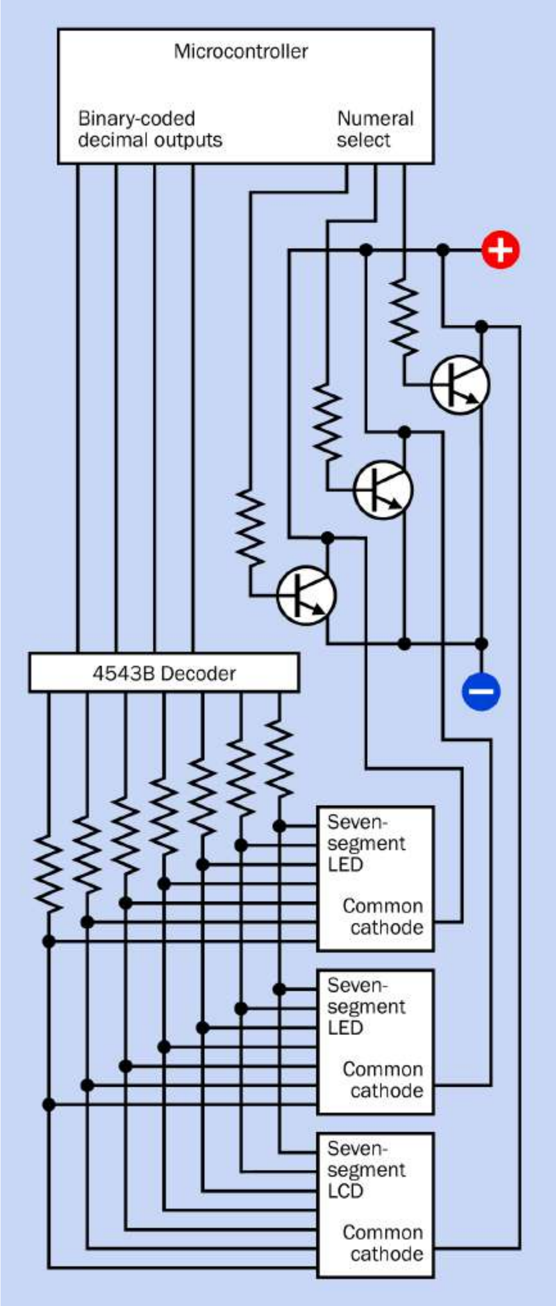

Figure 24-13. *A basic, simplified circuit for driving multiple seven-segment LED displays by multiplexing them.*

The disadvantage of this system is that the microcontroller must update the numerals constantly while performing other duties. To reduce this burden, a "smarter" driver such as the MC14489 can be used, controlling up to five 7-segment digits, or the Intersil ICM7218, which can control up to eight 7-segment digits.

The MC14489 controller receives data serially, using SPI protocol, and handles the details of addressing the LEDs. Because it contains latches to sustain the displayed data, a microcontroller only needs to communicate with the driver when the displayed information needs to be updated.

The ICM7218 is a more sophisticated chip, available in several variants, one of which can receive data on an 8-bit bus and run the seven-segment displays in hexadecimal mode.

Sixteen-Segment Driver Chip

The MAX6954 by Maxim can drive up to eight 16-segment alphanumeric LED displays using a scheme known as Charlieplexing, named after a Maxim employee named Charlie Allen who came up with the concept as a way of reducing the pin count required for multiplexing. Other Maxim controllers use this same protocol, which is transparent to the user.

A microcontroller sends data serially via I2C protocol to the MAX6954, which contains a variety of features. It can drive 14-segment and 7-segment displays as well as 16-segment displays, and contains a 104-character alphabet for each of them. Setting up a microcontroller to send the various necessary command codes to the MAX6954 is not a trivial matter, and bearing in mind the probably impending end-of-life of 16-segment displays, a better option may be to use dot-matrix LED display modules that have controller logic built in.

Dot-Matrix LED Display Modules

A dot-matrix LED display module requires data to define a character set, and a command interpreter to process instructions that will be embedded in a serial data stream. These capabilities are provided either by separate chips or (more often) are incorporated into the LED display module itself.

The SSD1306 is a monochrome graphical controller capable of I2C or SPI serial communication, or parallel communication. When this capability is built into a display module, only one of these types of communication may be activated.

The SSD1331 is a color graphical controller with similar communication capabilities.

The WS0010 is a monochrome controller, compatible with HD44780, which is designed to control LCDs.

Typical controller functions are summarized in "Alphanumeric Display Module" on page 168. Because there is no standardization in this field, precise details must be found in manufacturers' datasheets.

Pixel Arrays

The connections inside an 8x8 pixel array are shown in Figure 24-14, where the schematic symbols for LEDs have been replaced by gray circles for space reasons. To illuminate one LED, power is supplied to the intersection where it resides. In the figure, each vertical conductor (identified as A1, A2 . . . A8) can power the anodes of a column of eight LEDs, while each horizontal conductor (identified as C1, C2 . . . C8) can ground the cathodes of a row of eight LEDs. If only one vertical conductor is connected with positive power while one horizontal conductor is grounded, only one LED will light up, at the intersection of the active conductors.

A problem occurs if we wish to illuminate two LEDs. Suppose they are located at (A3,C2) and (A6,C5). Unfortunately, providing power to them will also result in activating LEDs at (A3,C5) and (A6,C2), as shown in Figure 24-15, where the yellow circles represent LEDs that have been switched on.

The answer to this problem is to rasterize the process. In other words, data is supplied on the array one line at a time, as in the process by which a TV picture is generated. If this is done quickly enough, persistence of vision will create the illusion that the LEDs are illuminated simultaneously.

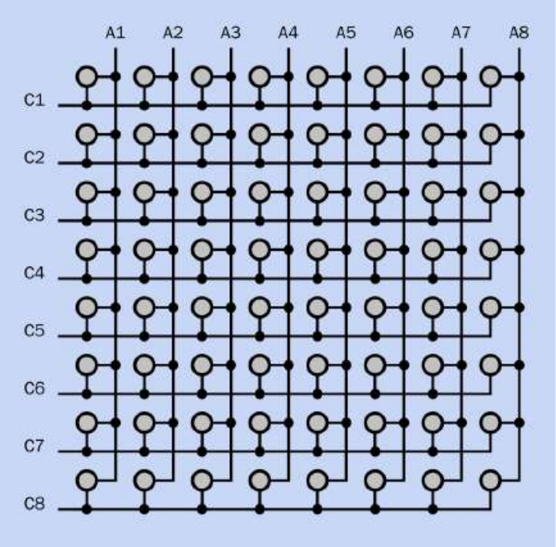

Figure 24-14. *Internal connections in the 8x8 matrix. Each gray circle represents an LED.*

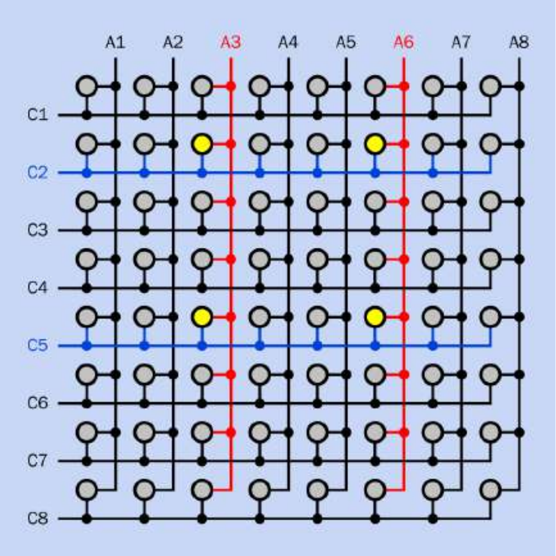

Figure 24-15. *An attempt to illuminate LEDs at (A3,C2) and (A6,C5) will also activate the LEDs at (A3,C5) and (A6,C2).*

A form of multiplexing is used to achieve this. One row of LEDs is connected to negative ground for a brief interval. During this interval, the anodes of selected LEDs are powered momentarily. Then the next row is grounded, and selected LEDs along that row are powered momentarily. This process is repeated for all eight rows before being repeated.

If several 8x8 matrices are assembled edge to edge, their horizontal conductors can be common to all of them. A horizontally scrollable display (sometimes referred to by the archaic term, *electric newspaper*) would then be possible, although the circuit design would be nontrivial.

Multiple Bar Display Driver

The LM3914 is a driver for a bar display that compares an analog input with a reference voltage and provides power to the segments of a multiple bar display, ranging from 2mA to 30mA, adjustable to match the specification of the display that is being used. The chip can generate either a "thermometer" effect, as more outputs are activated when the analog input increases, or a "moving dot" effect, in which only one output is on at a time.

One-Digit Hexadecimal Dot Matrix

While multi-character dot-matrix LED display modules are a versatile way to display prompts and numbers, a simpler component is sometimes sufficient. The Texas Instruments TIL311 is a minimal dot-matrix LED display that receives a binary value from 0000 through 1111 on its four input pins and generates the output in hexadecimal form, using numerals 0 through 9 and letters A through F. The sixteen possible outputs in the dot-matrix display are shown in Figure 24-16. Although this component is no longer being manufactured, it is widely available from many sources, especially in Asia. It eliminates the series resistors and controller chip that are customary for a seven-segment display, and has a better-looking output.

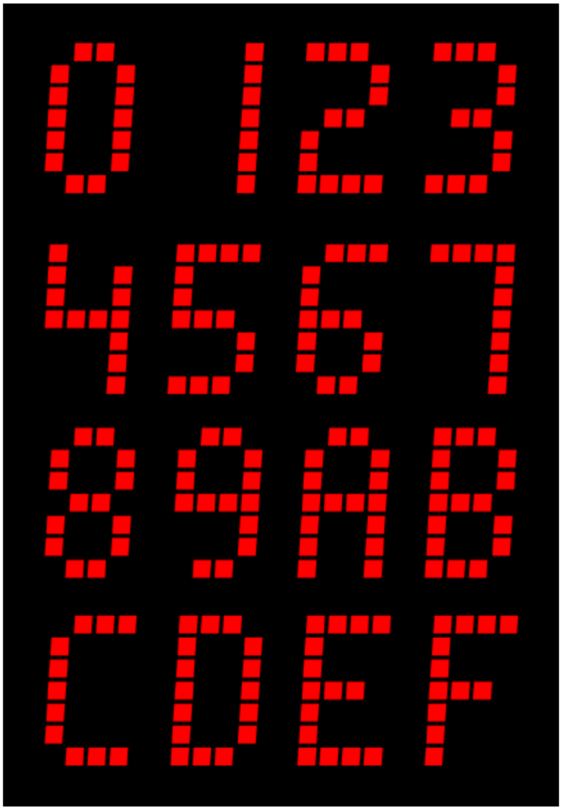

Figure 24-16. *Sixteen possible outputs that can be displayed by the Texas Instruments TIL311 to show a hexadecimal value in response to a four-digit binary input.*

A sample of the TIL311 is shown in Figure 24-17 displaying the number 2.

If two or more of these chips are put together, they can be multiplexed to display multi-digit decimal or hexadecimal integers.

The chip features two decimal points, one to the left of the displayed numeral, and one to the right. If they are activated, they require their own series resistors to limit the current.

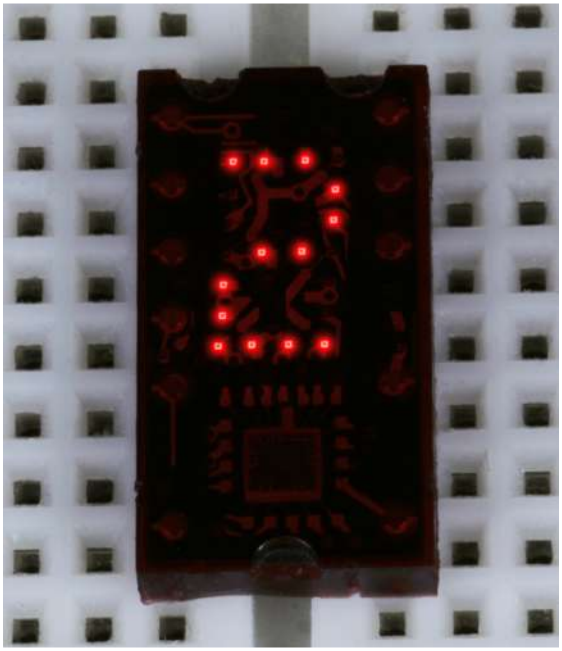

Figure 24-17. *The Texas Instruments TIL311 can be driven directly by a microcontroller or counter chip, with no series resistors necessary. It can generate a hexadecimal output.*

What Can Go Wrong

Common Anode versus Common Cathode

An LED display containing a common cathode is usually identical in appearance to a display containing a common anode, and the two versions will be distinguished by only one digit or letter in their part numbers. Because LED displays have a limited tolerance for reverse voltage, part numbers should be double-checked before applying power.

Incorrect Series Resistance

A common error is to assume that only one series resistor is necessary for a seven-segment LED display, either between the common cathode pin and ground, or, if there is a common anode, between that and the positive power supply. The problem is that if the resistor is suitable for a single LED, its value will be too high when several segments of the display are sinking current or drawing current through it. If its value is reduced, it will be too low when only two segments are using it (as when generating the number 1).

To provide equal illumination of all the segments, each must have its own series resistor.

Multiplexing Issues

When several displays are multiplexed, they naturally appear dimmer, creating a temptation to compensate by upping the current. Because current is only being applied to each display intermittently, a natural assumption is that a higher current can be safely used.

This may or may not be true. When running an LED device with pulsed current, the peak junction temperature, not the average junction temperature, determines the performance. At refresh rates below 1kHz, the peak junction temperature is higher than the average junction temperature, and the average current must therefore be reduced.

Datasheets must be checked to determine whether a device is designed with multiplexing in mind and, if so, what the recommended peak current is. Very often this value will be accompanied by a maximum duration in milliseconds, and a calculation may be necessary to determine the refresh rate, bearing in mind how many other LED displays are being multiplexed in the same circuit at the same time.

Irresponsible multiplexing will shorten the life of an LED display or burn it out.

vacuum-fluorescent display

25

The term **vacuum-fluorescent display** is seldom hyphenated, but the first two words are hyphenated here as they constitute an adjectival phrase. The acronym *VFD* is becoming increasingly popular, although it is ambiguous, being also used to identify a *variable frequency drive*. In both instances, the acronym is printed without periods between the letters.

The entry in this encyclopedia dealing with **fluorescent lights** does not include VFDs, because their purpose and design are very different. A VFD is an informational display, often showing numerals and letters, while a fluorescent light merely illuminates a room or work area. Although a VFD does use fluorescent phosphors, they are printed onto light-emitting segments of the display instead of being applied to the inside surfaces of a glass envelope.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)
- **LCD display** (see Chapter 17)
- **electroluminescence** (see Chapter 26)

What It Does

A **vacuum-fluorescent display** or *VFD* superficially resembles a backlit monochrome **LCD** or an **LED display**, as it can represent alphanumeric characters by using *segments* or a *dot matrix*, and can also display simple shapes. It is often brighter than the other information display systems, and can emit an intense green phosphorescent glow that some people find aesthetically pleasing, even though a grid of very fine wires is superimposed internally over the displayed image.

There is no specific schematic symbol to represent a vacuum-fluorescent display.

How It Works

The display is mounted inside a sealed capsule containing a high vacuum. A widely spaced series of very fine wires, primarily made of tungsten, functions as a *cathode*, moderately heated to encourage electron emission. The wires are often referred to as *filaments*.

A **fluorescent light** uses AC, and both of its electrodes are often confusingly referred to as cathodes. A VFD uses DC, and its cathode array has the function that one would expect, being connected with the negative side of the DC power supply.

Opposite the cathode, just a few millimeters away, is an anode that is subdivided into visible alphanumeric segments, symbols, or dots in a matrix. Each segment of the anode is coated with *phosphors*, and individual segments can be separately energized via a substrate. When electrons strike a positively charged anode segment, it emits visible light in a process of *fluorescence*. This behavior can be compared with that of a *cathode-ray tube*. However, the cathodes in a VFD are efficient electron emitters at a relatively low temperature, while the cathodes in a cathode-ray tube require substantial heaters.

Anode, Cathode, and Grid

A *grid* consisting of a mesh of very fine wires is mounted in the thin gap between the filaments of the cathode and the segments of the anode. A simplified view of this arrangement is shown in Figure 25-1.

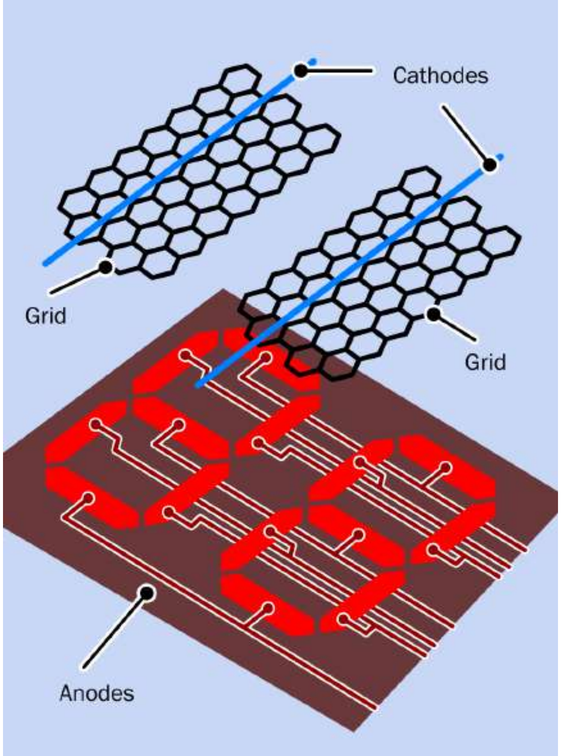

Figure 25-1. *The basic elements of a vacuum-fluorescent display.*

The polarity of the charge on the grid controls and diffuses electrons emitted by the cathode. If a grid section is negatively charged, it repels electrons and prevents them from reaching the sections of the anode beneath it. If the grid section is positively charged, it encourages electrons to reach the anode. Thus, the grid functions in the same way as the grid in a triode vacuum tube, but its conductors are so thin, they are barely visible.

How to Use It

Electronic calculators used vacuum-fluorescent displays during the 1970s, before LED displays and LCDs became more competitive. Purely numeric VFD modules are still available as strings of digits, although they are becoming uncommon and have been replaced by alphanumeric dot-matrix modules where each VFD character is mounted in its own glass module on a separate substrate.

Figure 25-2 shows the interior of a Commodore calculator from the 1970s, with its nine-digit vacuum-fluorescent display enclosed in one glass capsule.

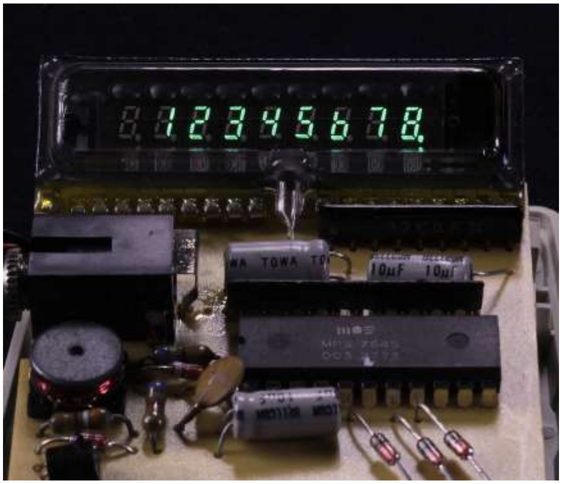

Figure 25-2. *The vacuum-fluorescent display from a 1970s Commodore calculator.*

A closeup of three digits from the previous figure appears in Figure 25-3, showing the grid superimposed above each numeral.

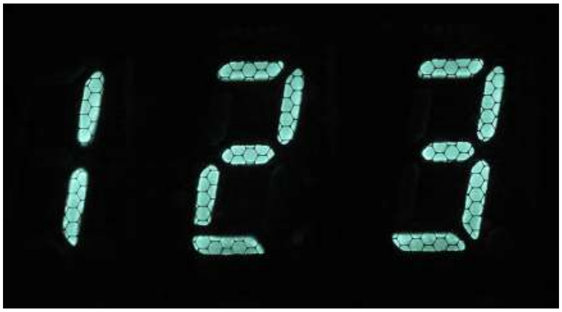

Figure 25-3. *Three digits from the previous figure, showing the grid that controls their illumination.*

A tinted filter of the same color as the display is usually placed in front of it, to conceal some of its workings. Thus, the Commodore calculator used a green filter in front of its green numerals. Figure 25-4 shows two seven-segment numerals from another device, with the filter removed. This reveals not only the grid but also the horizontal wires that function as the cathode. Connections between the segments of the numerals and a backplane are also visible.

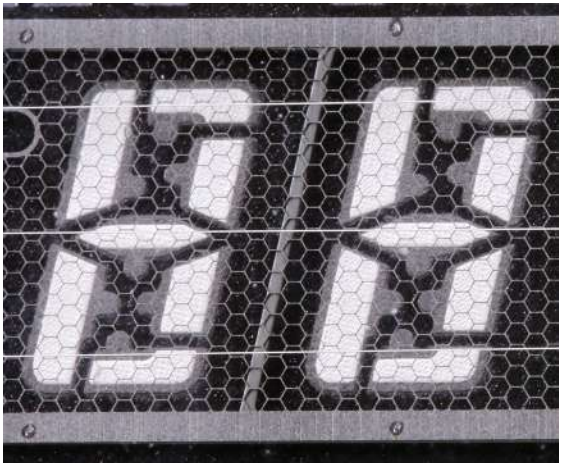

Figure 25-4. *Seven-segment numerals viewed without a colored filter, revealing the cathode (horizontal wires) and the grid (wire mesh).*

Modern Application

A modern VFD module is likely to be mated with a driver that converts 5VDC to the higher voltage (typically 50VDC to 60VDC) required for the display. Built-in logic may offer the option to receive data via an 8-bit parallel bus or with SPI serial protocol, and will contain a character set. A typical display resolution is 128 x 64 pixels.

The combination of a grid and a segmented anode enables a VFD to be controlled by *multiplexing*. For instance, in a display of four seven-segment numerals, the same equivalent segments in all four numerals can be connected in parallel while a separate grid covers each numeral. When each grid is positively energized, it selects the corresponding numeral, and the on-off segment patterns appropriate to that numeral are supplied. This procedure is repeated for each numeral in turn. Persistence of vision makes it appear that they are all active simultaneously.

Variants

Color

Although a VFD cannot provide a full-color display, selected anode segments can be coated with different phosphor colors, which can fluoresce simultaneously. Two or three individual colors are typically used, as in the display for a CD player where color helps to distinguish a variety of different functions. A closeup of a portion of the display from a CD player (with color filter removed) appears in Figure 25-5.

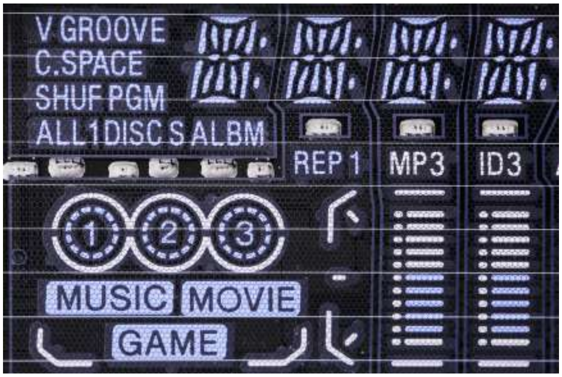

Figure 25-5. *The lefthand section of a vacuum-fluorescent display from a CD player.*

Character Sets and Pictorial Design

In the past, VFDs have combined seven-segment numerals in the same display as custom-shaped anodes. Solid-state gain meters in an audio amplifier, for instance, have used numerals beside a pictorial representation of gain levels resembling analog meters. The look and layout of a display of this type has been unique to a particular product.

Modern VFDs tend to use a generic dot-matrix display in which a character set in firmware dictates how patterns of dots are grouped to form numbers, letters, symbols, or icons.

The appearance of character sets generated with generic segments and dot-matrix arrays is thoroughly discussed and illustrated in the entry describing liquid-crystal displays in Chapter 17. VFD alphanumeric modules are identical in visual design to LCD modules, even though the internal electronics are different.

Comparisons

Two advantages of a VFD are that it functions well at low temperatures (unlike an LCD) and has sufficient brightness and contrast to be usable in sunlight (unlike many LED displays). It can be viewed from almost any angle.

Typical applications have included digital instrumentation in automobiles, information displays in audio and video consumer-electronics equipment, and numerical readouts in vending machines, medical devices, and some digital clocks.

Because a VFD requires a relatively high voltage, has significant power consumption, can show only a limited range of fixed colors, and is more expensive than LED displays or LCDs, its popularity has declined since the end of the 1990s.

What Can Go Wrong

Fading

VFDs gradually fade with age, as a result of reduced electron emission from the electrodes or diminishing performance of the phosphor coatings. Increasing the working voltage can prolong the life of a display.

electroluminescence

26

The field of **electroluminescent** devices is sometimes referred to as *EL*. The same acronym can also be applied as an adjective to an individual electroluminescent device, as in, for example, "an EL panel."

An *organic light-emitting diode*, more commonly known by its acronym *OLED*, is included in this entry because it is technically an electroluminescent device and its design concept is similar to that of an electroluminescent panel. Generic LEDs are also technically electroluminescent, but are not commonly described as such, and have their own entries in this encyclopedia under the subject categories **LED indicator**, **LED area lighting**, and **LED display**.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)
- **LCD display** (see Chapter 17)
- **fluorescent light** (see Chapter 20)
- **vacuum-fluorescent** display (see Chapter 25)

What It Does

An electroluminescent device configured as a *panel*, *ribbon*, or *rope-light* contains phosphors that emit light in response to a flow of electricity.

Panels can be used as backlights for LCD displays or, more often, as always-on low-power devices such as exit signs and night lights. Ribbons and rope lights (the latter being also known, more accurately, as *light wires*) are used mainly as recreational novelties. They can be battery powered through a suitable voltage converter. A battery-powered rope light can be wearable.

Thin-film OLED electroluminescent panels are used in small video screens in handheld devices. At the time of writing, OLED TV screens measuring 50" or more have been demonstrated, but are not yet economic for mass production.

No specific schematic symbol exists to represent any electroluminescent device or component.

How It Works

Luminescence is the emission of light as a result of a process that does not require heat. (The opposite phenomenon is *incandescence*, in which heating causes an object to emit light; see Chapter 18 for a description of **incandescent lamps**.)

Electroluminescence is luminescence resulting from stimulation by electricity. This very broad definition really includes devices such as **LED**s, although they are hardly ever described in those terms. Electroluminescence generally refers to panels, films, or wires where electrodes are in direct contact with light emitters such as phosphors.

The exception is an *organic LED*, usually known by its acronym *OLED*, which is frequently described as an electroluminescent device, perhaps because its configuration as a sandwich of thin, flat layers resembles an electroluminescent panel. Two of the layers are semiconductors, and they interact as light-emitting diodes.

Phosphors

A *phosphor* is a compound such as zinc sulfide that will emit light when it receives an energy input from another light source or from electricity. Typically the compound must be mixed with an *activator* such as copper or silver.

For many decades, TV sets and video monitors were built around *cathode-ray tubes* in which the interior of the screen, at the front of the tube, was coated with phosphors. A beam of electrons that fluctuated in intensity generated a picture on the screen by drawing it as a series of lines.

Derivation

The term *phosphor* is derived from *phosphorescence*, which in turn comes from the name of the element *phosphorous*, which will glow when it oxidizes in moist air. (These terms were established before other forms of luminescence were discovered and understood. The behavior of phosphorous is really an example of *chemiluminscence*.)

For our purposes, a phosphor is a compound that is capable of fluorescence or electroluminescence.

Variants

Panels

Electroluminescent panels using phosphor powder, sometimes referred to as *thick phosphor*, are a popular choice where a constant, uniform, low light output is acceptable.

An electric potential is established between two films that act as electrodes, separated by a layer of phosphor crystals. Some manufacturers refer to this configuration as a *light-emitting capacitor* because the structure resembles a capacitor, even though that is not its purpose. The front film is transparent, allowing light to escape.

An electroluminescent panel can be powered by AC or DC but requires at least 75V. Its power consumption is self-limiting, so that no control electronics are required other than a voltage converter if battery power is used.

The phosphors generate a constant, evenly distributed luminescence over the entire area, although the output is not very intense. Applications include night-lights, exit signs, and backlighting for wristwatches.

Panelescent electroluminescent lighting by Sylvania was used for instrument panel displays in some car models such as the Chrysler Saratoga (1960 through 1963) and Dodge Charger (1966 through 1967). It is still used for night-lights. *Indiglo* electroluminiscent displays are still widely used in wristwatches.

The interior components of a disassembled electroluminescent night-light are shown in Figure 26-1. The panel emits a natural pale green glow. A separate blue or green filter passes the glow while blocking other colors of incident light that would otherwise reflect off the panel.

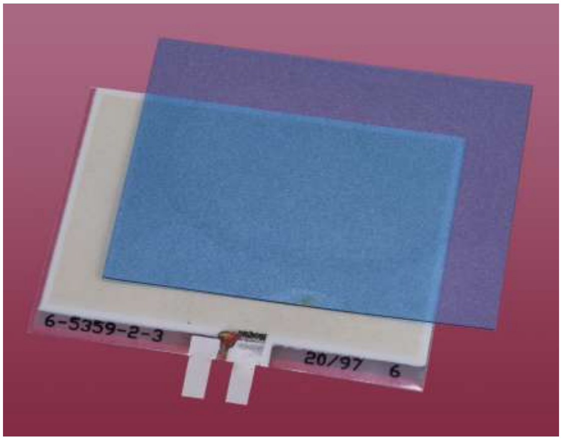

Figure 26-1. *The two interior components of an electroluminiscent night-light: the luminescent panel, and a separate translucent filter.*

Electroluminescent night-lights were popular in the 1970s and 1980s, often featuring cartoon characters to appeal to children. Figures 26-2 and 26-3 show the same night-light in its daytime off-state and its night-time on-state, respectively.

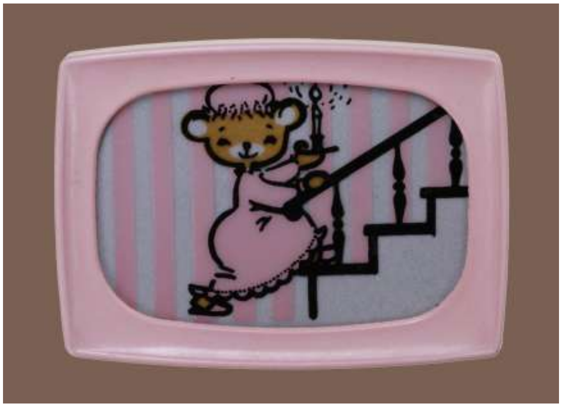

Figure 26-2. *A vintage Panelescent brand night-light, several decades old, in its off-state.*

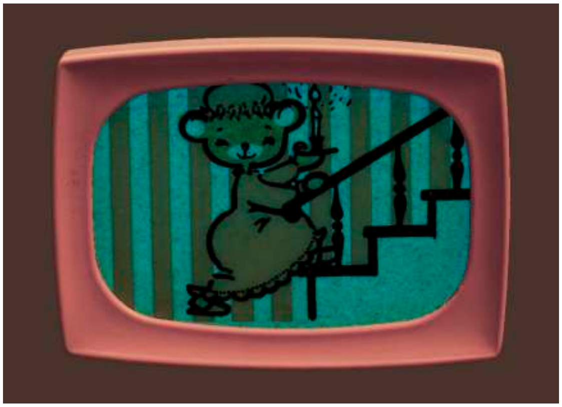

Figure 26-3. *The same night-light, with its green radiance visible under conditions of low ambient light.*

Advantages of electroluminescent panels include the following:

- Low current consumption. One US manufacturer claims that a single exit sign will use electricity costing less than 20 cents per year, while the annual cost of a night-light will be less than 3 cents per year.
- Long life, up to 50,000 hours.
- Self-regulating; no control circuitry required.
- Omnidirectional light output.
- Very wide operational temperature range, between approximately -60 and +90 degrees Celsius.
- Can be plugged directly into a wall outlet.

Disadvantages include:

- Limited light output.
- Very limited choice of colors.
- Not very efficient, 2 to 6 lumens per watt (although the low light output naturally entails low power consumption).
- Gradual reduction in phosphor performance over time.
- High voltage required: 60V to 600V. Ideal for plugging into a wall outlet, but requires a converter when used with battery-powered devices.

Flexible Ribbons

The light-emitting layers inside a night-light are somewhat flexible, and can be made more flexible by reducing their thickness. The result is an electroluminescent ribbon that has some novelty value, and may be used for customizing automobiles. Figure 26-4 shows a ribbon about 1.5" wide and 12" long, designed for 12VDC power applied through an inverter.

Rope Light

A rope light or wire light may resemble a *glowstick*. However, a glowstick generates light from chemiluminescence (chemical reactions that release photons), while a rope light uses electricity.

Figure 26-5 shows a rope light powered by two AA batteries connected through an inverter.

At the center of the rope light is a conductor that serves as one electrode. It is coated in phosphors, and the layer of phosphors is protected by a transparent sheath. One or more thin wires is wrapped around the sheath in a spiral, with large gaps between one turn and the next. These wires serve as the second electrode. The wires are enclosed in transparent insulation that forms an outer sheath.

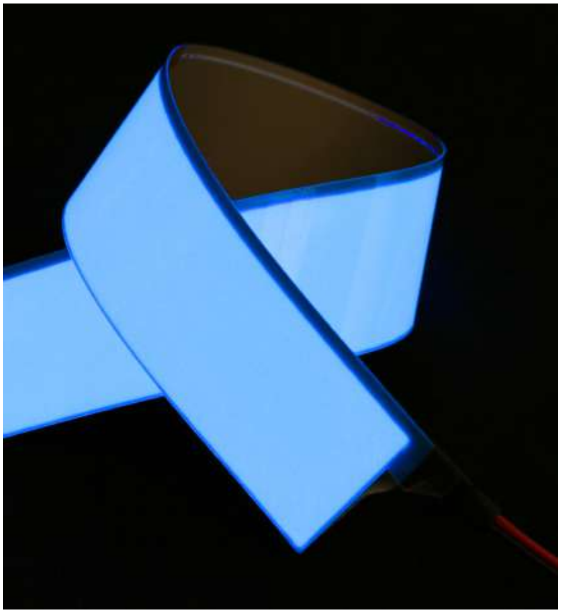

Figure 26-4. *A 12" length of electroluminescent ribbon.*

When AC is applied between the electrodes, the layer of phosphors emits light that radiates out in the gaps between the thin wires. The color of the light can be modified by using tinted outer insulation.

OLED

An OLED uses two thin, flat electrodes, somewhat like a thick-phosphor electroluminescent panel, except that it contains more layers and is capable of generating more light. The layers in an OLED are "organic" in that they consist of chemically organic molecules containing carbon and hydrogen atoms and generally do not contain heavy metals.

While an LCD video monitor or TV screen must have a separate backlight, an OLED generates its own light. This reduces the thickness of the display to a few millimeters and makes it potentially more efficient.

The semiconductor layers are subdivided into pixels, each functioning as a light-emitting diode, while additional layers carry a matrix of conductors for pixel addressing. In an AMOLED, the conductors form an active matrix, while in a PMOLED, they form a passive matrix.

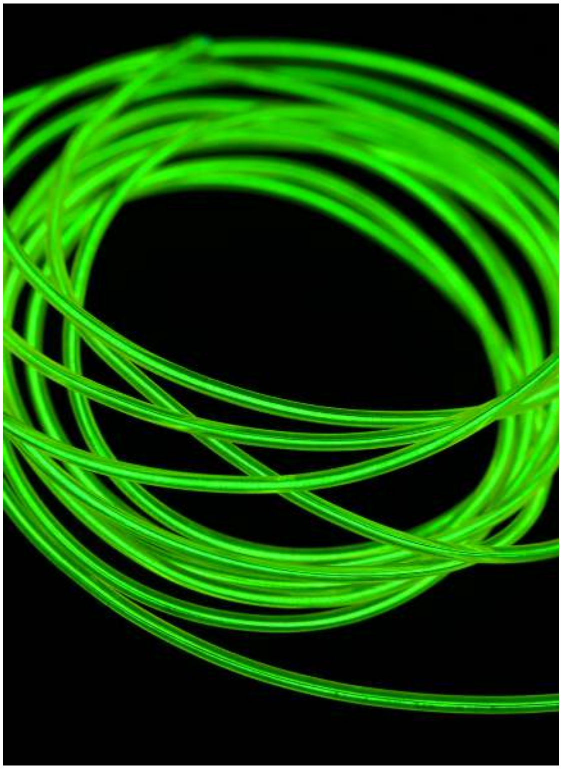

Figure 26-5. *A length of glowing rope light, also known as a light wire.*

In an active matrix, each pixel is backed with a *thin-film transistor* to store its state while the energizing voltage transitions. This is often described as a *TFT* display; but the term is interchangable with "active matrix."

In a passive matrix, each pair of conductors simply supplies current to a pixel. This is cheaper and easier to fabricate but is less responsive.

The terms "active matrix" and "passive matrix" have the same meaning as when used to describe a **liquid-crystal display**.

Monochrome OLED display modules with dot-matrix characters are currently available from China for just a few dollars. Although they appear superficially similar to LCD modules, they generate pure white-on-black characters.

Small full-color OLED screens are used in smartphones and on camera backs, but at the time of writing, large OLED screens are not a mature technology, partly because of production costs. A great variety of chemicals and layer configurations have been tried, and the application of pixels to a substrate has been attempted with vacuum deposition through a shadow mask and with a system similar to inkjet printing. Pixels that emit red, green, and blue light have also been used. Filtered pixels have been used. One dominant process has not yet emerged.

Longevity and brightness have been problems. Where red, green, and blue diodes have been used, the different colors deteriorate at different rates. While the human eye tolerates an overall reduction in brightness, it does not tolerate a slight color shift caused by blue pixels, for instance, losing brightness more rapidly than red pixels.

Because OLED screens promise to be thinner, lighter, and brighter, and may eliminate the need for a fragile glass substrate, there is a strong incentive to develop this technology, which seems likely to gain dominance in the future.

OLED panels may also become a source of diffuse, shadowless room lighting or office lighting when practical problems have been solved and costs have fallen significantly.

transducer

27

The term **transducer** is used here to describe a *noise-creating device* that is driven by external electronics. By comparison, an **audio indicator** (discussed in the next entry) contains its own internal electronics and only requires a DC power supply. Either of these components is often described as a *beeper* or *buzzer*.

A **speaker**, more properly termed a *loudspeaker*, is an electromagnetic transducer but is seldom described in those terms. It has a separate entry in this encyclopedia and is defined here as a *sound reproduction device* that is larger and more powerful than a typical transducer and has a more linear frequency response.

While piezoelectric transducers formerly used crystals, only the more modern piezoelectric type that uses a ceramic wafer will be considered here.

Some transducers convert sound into electricity, but these are categorized as *sensors*, and will be discussed in Volume 3. The only transducers discussed in this entry are those that convert electricity into sound.

OTHER RELATED COMPONENTS

- **audio indicator** (see Chapter 28)
- **headphone** (see Chapter 29)
- **speaker** (see Chapter 30)

What It Does

An audio **transducer** is a device that can create an *alert*. It requires an AC signal that is supplied by external electronics, and in its simplest form may be referred to as a *buzzer* or a *beeper*.

Audio alerts are used in microwave ovens, washer/dryers, automobiles, gasoline pumps, security devices, toys, phones, and many other consumer products. They are often used in conjunction with touch pads, to provide audio confirmation that a tactile switch has been pressed.

The schematic symbols in Figure 27-1 can be used to represent any kind of audio alert, including **indicators**, which contain their own electronics to generate a simple tone or series of tones. Type A is probably the most popular symbol. Types B and C often appear with the word "buzzer" printed beside them for clarification. D and E are really symbols for a **speaker**, but are often used for an alert. F is the symbol for a *crystal*, now sometimes used to indicate a piezoelectric noise maker. G specifically represents an electromagnetic transducer, but is seldom used.

How It Works

A circular diaphragm is glued at its edges inside a cylindrical plastic enclosure, usually measuring from around 0.5″ to 1.5″ in diameter. The enclosure is sealed at the bottom but has an opening at the top, so that sound can emerge from the upper side of the diaphragm without being partially cancelled by sound of opposite phase that is emitted from the underside of the diaphragm. The enclosure also amplifies the sound by resonating with it, in the same way that the body of a guitar or violin amplifies a note being played on the strings.

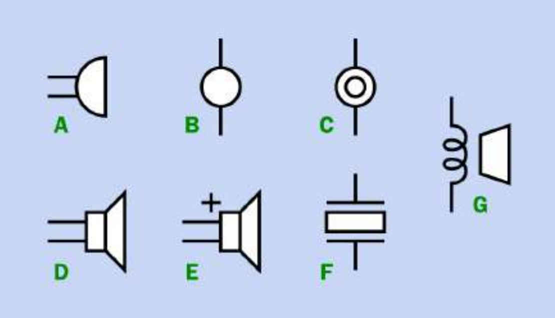

Figure 27-1. *An assortment of symbols which can represent a transducer or an indicator. See text for details.*

The diaphragm is activated either electromagnetically or piezoelectrically, as described next.

Externally, a transducer may be indistinguishable from an audio indicator such as the one pictured at Figure 28-1.

Variants

Electromagnetic

An electromagnetic transducer contains a diaphragm that is usually made of plastic. Mounted on it is a smaller *ferromagnetic disc* that responds to the fluctuating field from AC passing through a coil. When the diaphragm vibrates, it creates pressure waves that are perceived by the human ear as sound.

A car horn is a particularly loud form of electromagnetic transducer.

Piezoelectric

A piezoelectric transducer contains a diaphragm consisting of a thin brass disc on which is mounted a *ceramic wafer*. When an AC signal is applied between the piezoelectric wafer and the disc, the disc flexes at that frequency.

The term *piezo* is derived from the Greek *piezein*, which means "to squeeze or press."

Ultrasonic Transducer

The diaphragm in an *ultrasonic transducer* vibrates at a frequency above the range of human hearing. This component may be electromagnetic, piezoelectric, or crystal-based. Often it is used in conjunction with an ultrasonic receiver as a distance measuring device. The two components can be sold pre-mounted on a breakout board. An output from the board can consist of a pulse train where the pulse duration is proportional to the distance between the transducer and the nearest sound-reflecting object.

An ultrasonic transducer is pictured in Figure 27-2. Its internal components are shown in Figure 27-3.

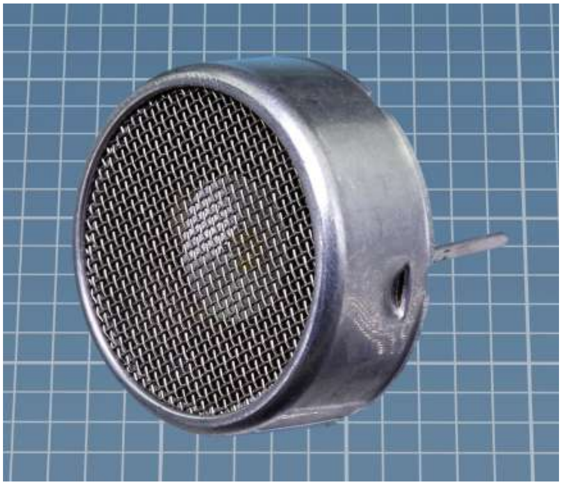

Figure 27-2. *The exterior of an ultrasonic transducer.*

Submersible ultrasonic transducers may be used in cleaning systems, where they agitate a liquid that dislodges dirt or debris. Ultrasonic transducers are also used in echo-sounding and sonar equipment with marine applications.

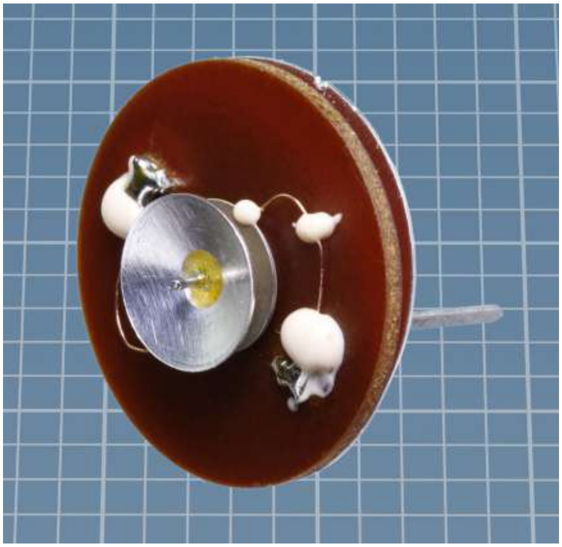

Figure 27-3. *Inside an ultrasonic transducer, a small aluminum cone is the sound radiating element. The white blobs are adhesive to secure the thin wires.*

Formats

Some transducers are available in surface-mount format, measuring about 0.5″ square or less. Because the resonant frequency is related to the size of the component, surface-mount transducers usually generate a high-pitched beep.

Values

Frequency Range

Audio frequency is measured in Hertz, abbreviated Hz, named after Heinrich Rudolf Hertz, the first scientist to prove the existence of electromagnetic waves. The H in Hz is capitalized because it refers to a real name. One thousand Hertz can be written as 1 kiloHertz, almost always abbreviated as 1kHz (note that the k is lowercase).

The human ear is often described as being able to detect sounds between 20Hz and 20kHz, although the ability to hear sounds above 15kHz is relatively unusual and diminishes naturally with age. Sensitivity to all frequencies can be impaired by long-term exposure to loud noise.

The most common frequencies applied to audio transducers range between 3kHz and 3.5kHz. Piezoelectric elements are inefficient for generating sounds below 1kHz, but electromagnetic transducers are better able to generate lower frequencies. Their response curve can be approximately flat to frequencies as low as 100Hz.

Sound Pressure

Sound pressure can be measured in Newtons per square meter, often abbreviated as Pa. Newtons are units of force, while Pa is an abbreviation of Pascals.

The *sound pressure level* (SPL) of a sound is not the same as its sound pressure. SPL is a logarithmic value, to base 10, in units of decibels (dB), derived from the pressure of a sound wave relative to an arbitrary reference value, which is 20 micro-Pascals (20µPa). This is the agreed minimum threshold of human hearing, comparable to a mosquito at a distance of three meters. It is assigned the level of 0dB.

Because the decibel scale is logarithmic, a linear increase in the decibel level of a sound does not correspond with a linear increase in actual sound pressure:

- For each additional 6dB in the SPL, the actual sound pressure approximately doubles.
- For each additional 20dB in the SPL, the actual sound pressure is multipled by 10.

Bearing in mind that 0dB corresponds with the reference sound pressure of 20µPa, an SPL of 20dB represents a sound pressure of 200µPa (that is 0.0002Pa), and so on.

Many tables show an estimated decibel level for various noise sources. Unfortunately, these tables may contradict each other, or may fail to mention the distance at which a sound is measured. Figure 27-4 shows estimates derived by averaging eight similar tables. It should be viewed as an approximate guide.

Decibels	Noise Example
140	Jet engine at 50 meters
130	Threshold of pain
120	Loud rock concert
110	Automobile horn at 1 meter
100	Jackhammer at 1 meter
90	Propeller plane 300 meters above
80	Freight train at 15 meters
70	Vacuum cleaner
60	Business office
50	Conversation
40	Library
30	Quiet bedroom
20	Leaves rustling
10	Calm breathing at 1 meter
0	Auditory threshold

Figure 27-4. *Approximate decibel values for some sound sources (averaged from a selection of eight similar charts).*

Sometimes the claim is made that an increase of +10 on the decibel scale will correspond with a subjective experience that the noise is "twice as loud." Unfortunately, this statement cannot be quantified.

Weighted Sound Values

Subjective assessment of sound is complicated by the nonlinear frequency response of the human ear, which causes some frequencies to seem "louder" than others, even though their sound pressure is the same. The frequency weighting of the ear can be determined by playing a reference tone of 1kHz at 20dB and then doing an A-B comparison with a secondary tone at another frequency, asking the subject to adjust the gain of the secondary tone up or down until the two tones seem equally loud.

This procedure is performed for a range of frequencies. The test is then repeated with a louder 1kHz reference tone, at 30dB. Repetitions continue to a final reference tone of 90dB.

The resulting curves are known as *equal-loudness contours*. An averaged set, from multiple sources, has become an international standard with ISO number 226:2003. The curves shown in Figure 27-5 are derived from that standard. The curves show that the sound pressure of lower frequencies must be boosted by a significant amount to sound as loud as a 1kHz frequency, while a frequency around 3kHz must be reduced slightly, because it tends to sound louder than all others.

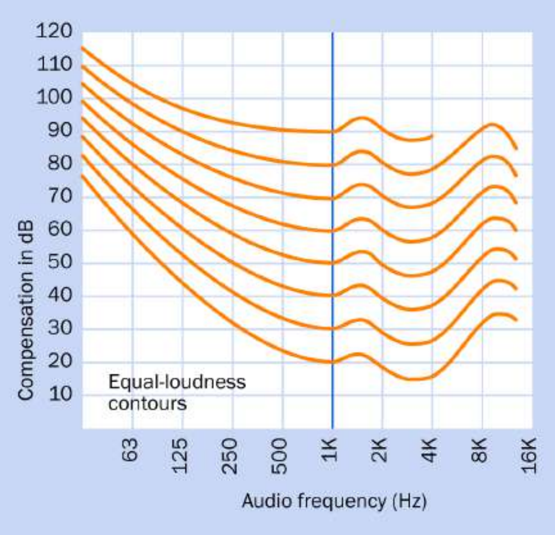

Figure 27-5. *Equal loudness contours derived from ISO 226:2003. See text for details.*

Although the accuracy of equal-loudness contours is controversial, they have been the basis of a widely used weighting system to adjust dB values to represent subjective perceptions of loudness. This *A-weighting* system remains the best-known and most widely applied audio standard in the United States, even though it has been criticized for assigning too little value to sounds that are brief in duration. If a sound level is expressed in dBA, it is A-weighted, meaning that the sounds to which the ear is least sensitive are assigned a value that is lower than their measured value. Thus, a tone of 100Hz has a dBA value about 20dB lower than its dB value, because the human ear is relatively insensitive to low-pitched sounds. dBA values are used in regulations that limit noise in the work place and other environments.

Unweighted Values

If sound intensity is expressed in dBSPL, it is a measurement of the actual Sound Pressure Level and has not been adjusted with the A-weighting system. A graph of unadjusted dbSPL values will display low frequencies as if they are more intense than the ear will perceive. In practical terms, subjective perception of low-end rolloff will be even more severe than the graph makes it appear.

If sound intensity is expressed merely in dB, probably it is unweighted and should be considered as dBSPL.

From a practical point of view, when choosing a tone for a transducer, a 500Hz tone may sound relatively mellow and not subjectively loud. A 3.5kHz tone can be a good attention-getting signal, as the ear is most sensitive in that range.

Transducers generally have a sound pressure rating in dBSPL ranging between 65dBSPL to 95dBSPL, with just a few products that can make more or less noise.

Measurement Location

The sound pressure from an audio alert will naturally diminish if the measurement point moves farther away. Therefore, any rating in decibels should be expressed with reference to the distance at which the measurement is made.

Measurement locations may be expressed in centimeters or inches, and may vary from 10cm to 1 meter, even in datasheets for different devices from the same manufacturer. If the measurement distance doubles, the SPL diminishes by approximately 6dB.

Limitations

A piezoelectric transducer is not intended as a sound reproducer, and does not have a smooth or flat frequency response. The curve for the Mallory PT-2040PQ is not unusual, reproduced in Figure 27-6. This component measures about 3/4” in diameter, is rated for 5VDC, and uses only 1.5mA to generate 90dB (measured at a distance of 10cm). Like many piezoelectric audio devices, its response peaks around 3500kHz and diminishes above and below that value, especially toward the low end. While it is perfectly adequate as a “beeper,” it will not reproduce music successfully.

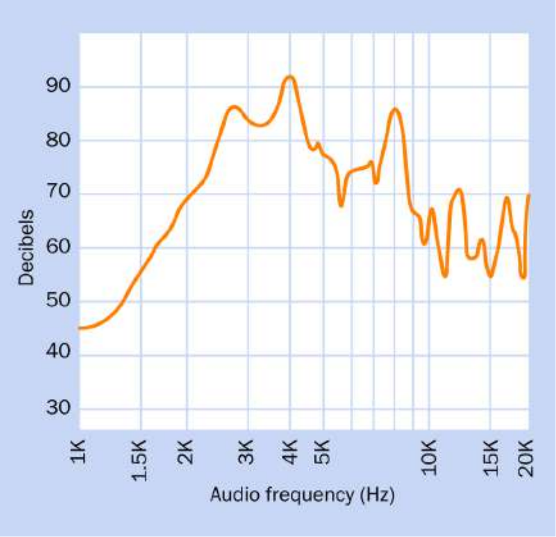

Figure 27-6. *The frequency response from a typical small piezoelectric transducer.*

An electromagnetic transducer is better able to generate low frequencies than a piezoelectric transducer. It has a low impedance that may be suitable in some circuits. However, it is slightly heavier than a comparable piezoelectric transducer, uses much more power, and as an AC device containing a coil, it can create electromagetic interference or may cause fluctuations in the circuit as an inductive load. It is also vulnerable to magnetic interference from elsewhere, while a piezoelectric transducer is not.

While an electromagnetic transducer can be used to reproduce speech or music, and will do a better job than a piezoelectric transducer, its performance will still be dissatisfying. A miniature speaker is more appropriate for the task.

Voltage

Transducers are typically designed to work with voltages ranging from 5VAC to 24VAC. The ceramic wafer in a piezoelectric transducer usually cannot withstand voltages much above 40VAC, and its sound output will not increase significantly above 30VAC.

Current

Typical piezoelectric transducers use less than 10mA and generate negligible heat. An electromagnetic transducer may draw as much as 60mA.

How to Use It

Appropriate Sound Intensity

An alert should be chosen with reference to the environment in which it will be used. To be easily heard, it should be at least 10 dB louder than ambient background noise.

Volume Control

Sound pressure can be lowered by reducing the voltage. Because a transducer does not consume much current, a *trimmer* can serve as a volume control. Alternatively, a rotary switch with a set of fixed-value resistors can select preset sound values.

AC Supply

Although a transducer is an AC device, it is unlikely to be designed for voltage that fluctuates positively and negatively either side of a neutral value. Typically it is intended for voltage that fluctuates between 0V (ground) and the rated positive value of the power supply, and its pins, wires, or terminals are usually marked accordingly. If it has wire leads, the red lead should be connected to the more-positive side of the supply. If it has pins, the longer pin should be more positive.

The alternating signal for a transducer can be supplied by any simple oscillator or astable multivibrator circuit. For a given peak voltage, a square wave will generate a louder signal than a sinusoidal wave. A simple 555 timer circuit can be used, with a second monostable timer to limit the duration of the beep if necessary. An astable 555 can be used to test the transducer and select the audio frequency that sounds best.

Self-Drive Transducer Circuit

If a transducer has three wires or pins, it is probably a *self-drive* type. The datasheet may identify its inputs as M, G, and F, meaning Main, Ground, and Feedback. The Feedback terminal is connected with a section of the diaphragm which vibrates 180 degrees out of phase with the Main terminal. This facilitates a very simple external drive circuit, such as that in Figure 27-7, where the frequency is determined by the transducer's resonant frequency.

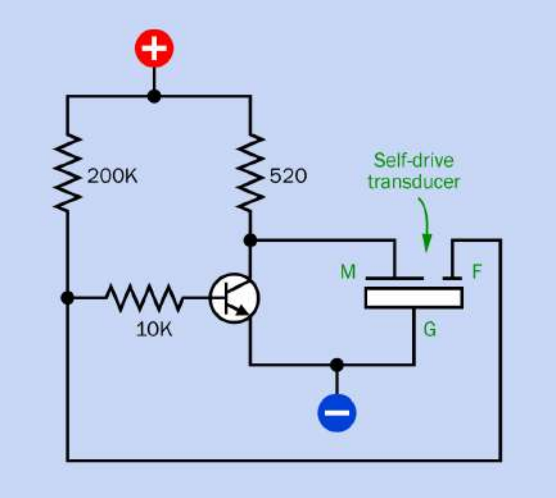

Figure 27-7. *A circuit to control a self-drive type of piezoelectric transducer.*

What Can Go Wrong

Overvoltage

Mallory Sonalerts, one of the largest producer of piezoelectric alerts, states that in the "vast majority" of returned products, the failure was caused by excessive voltage, often in the form of a transient voltage spike.

Leakage

If an alert makes a small, low-volume sound when it is supposed to be off, some current is leaking through it. Less than 1mA may be enough to cause this problem. According to one manufacturer, it can be fixed by placing a 30V Zener transient voltage suppressor diode in series with the alert, or by wiring a small **incandescent lamp** in parallel with the alert.

Note that when the alert is activated, the full supply voltage will be seen at the lamp.

Component Mounting Problems

Some alerts are packaged with mounting holes, but many are not. Those with pins can be soldered into a board, but those without must be glued in place or inserted into a cavity from which they cannot shake loose. Silicone adhesive is recommended, but care must be taken to avoid any of it dripping into the alert before it sets.

Moisture

If an alert will be used in a location where it is vulnerable to moisture, it should be of a type that is sealed against the environment. Even a sealed unit should ideally be oriented so that it faces slightly downward.

Transducer-Indicator Confusion

Externally, a transducer and an indicator often look identical, and some of them are not marked with a manufacturer's part number. Damage can be caused by applying DC to a transducer or AC to an indicator. If both types of parts are kept in inventory, they should be carefully labeled.

Connection with a Microcontroller

A piezoelectric transducer can be driven by a microcontroller, but an electromagnetic transducer is not appropriate in that role, because of its relatively higher current consumption and its behavior as an inductive load.

audio indicator

28

An **audio indicator** is defined here as a noise-creation device that generates a simple tone or series of tones. Unlike a **transducer**, which requires an external source of AC to determine its audio frequency, an indicator contains its own electronics and requires only a DC power supply. Either of these components is often described as a *beeper* or *buzzer*.

While piezoelectric alerts formerly used crystals, only the more modern piezoelectric type that uses a ceramic wafer will be considered here.

OTHER RELATED COMPONENTS

- **transducer** (see Chapter 27)
- **headphone** (see Chapter 29)
- **speaker** (see Chapter 30)

What It Does

When DC power is applied to an **audio indicator**, in its simplest form it creates a continuous or intermittent tone of a fixed frequency. This is called an *alert*.

Audio alerts are used in microwave ovens, washer/dryers, automobiles, gasoline pumps, security devices, toys, phones, and many other consumer products. They are often applied with touch pads, to provide audio confirmation that a tactile switch has been pressed.

A few indicators are programmed to create a two-tone sound, or multiple-tone sequences.

See Figure 27-1 in the previous entry for an assortment of schematic symbols that may be used to represent either an alert or a transducer.

How It Works

A circular diaphragm is glued at its edges inside a cylindrical plastic enclosure, usually measuring from around 0.5″ to 1.5″ in diameter. The enclosure is sealed at the bottom but has a small hole at the top, so that sound can emerge from the upper side of the diaphragm without being partially cancelled by sound of opposite phase that is emitted from the underside of the diaphragm. The enclosure also contains electronics to generate one or more audio tones, and amplifies the sound by resonating with it, in the same way that the body of a guitar or violin amplifies a note being played on the strings.

A PUI XL453 piezoelectric audio indicator is pictured in Figure 28-1, fully assembled on the right, and with its circuit board and diaphragm removed on the left. This indicator creates a pulsed tone at 3.5kHz with a sound pressure of 96dB. It draws 6mA at 12VDC and measures approximately 1″ in diameter.

For more information about the measurement of sound frequency and pressure, see “Frequency Range” on page 251 and “Sound Pressure” on page 251 in the previous entry.

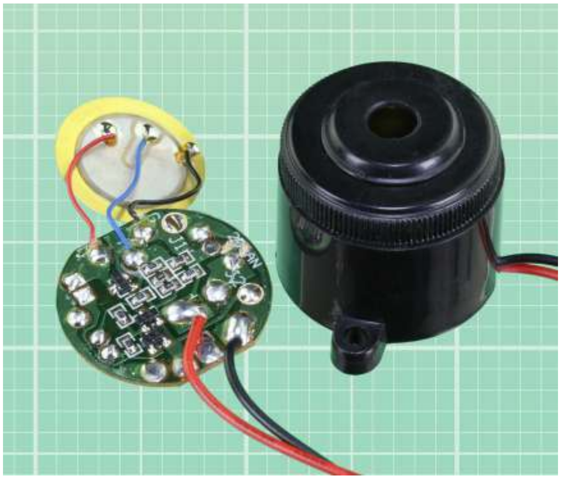

Figure 28-1. *A typical piezoelectric audio indicator.*

Externally, an audio indicator may be indistinguishable from a transducer. However, internally, an indicator is almost always a piezoelectric device, in which a *ceramic wafer* is mounted on a thin brass diaphragm. The term *piezo* is derived from the Greek *piezein*, which means to squeeze or press.

A transducer (described in the previous entry) is a piezoelectric or electromagnetic alert that does not usually contain its own circuitry and must be driven by an external source of AC, which establishes the audio frequency.

The distinction between an indicator and a transducer is often unclear in parts catalogs, where all alerts may be identified as *buzzers*, even though they mostly beep rather than buzz.

Audio Frequency

For a discussion of audio frequency, see "Frequency Range" on page 251 in the previous entry.

History

Probably the earliest form of electrically activated alert was the door bell, in which a 6VDC battery-powered solenoid pulled a spring-loaded lever terminating in a small hammer. The hammer struck the bell, but the motion of the lever also opened a pair of contacts, cutting off power to the solenoid. The lever sprang back to its rest position, which closed the contacts and repeated the cycle so long as power was supplied through an external pushbutton.

Subsequent systems used a small loudspeaker powered by AC house current through a step-down transformer. This created a buzzing sound and may have been the origin of the term "buzzer."

Small components that made a beeping sound only became common when digital equipment required a simple, cheap way to confirm user input or attract attention to the status of a device.

Variants

Sound Patterns

Because an audio indicator contains its own electronics, the manufacturer has the freedom to create various patterns of sound output.

The default is a steady tone. Other common variants include an intermittent tone and a dual tone that fluctuates rapidly between two frequencies. This is sometimes referred to as a *siren*. A few variants can generate an output pattern consisting of several tones in sequence, or effects such as warbling or whooping sounds, which are used mainly in alarm systems.

Formats

Some audio indicators are available in surface-mount format, measuring 1/2" square or less. Because the resonant frequency is related to the size of the component, surface-mount alerts usually make a high-pitched beep.

Panel-mount and board-mount formats range from about 1/2" to 1.5" in diameter. A small audio alert designed to be mounted on a circuit board is shown in Figure 28-2, with its top removed on the right to expose the brass diaphragm glued around the edges. The same component is shown with its plastic enclosure removed completely in Figure 28-3.

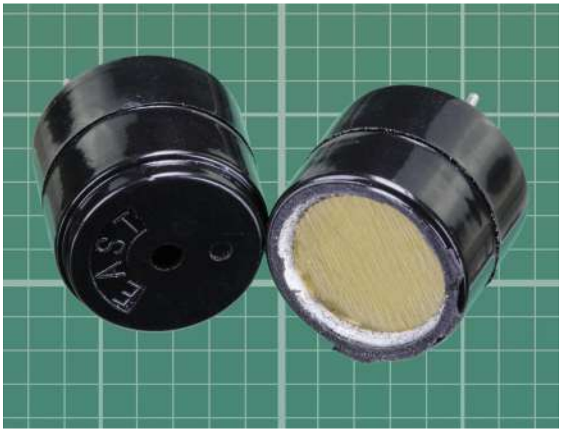

Figure 28-2. *An audio indicator approximately 0.5" in diameter, partially disassembled on the right, revealing its brass diaphragm.*

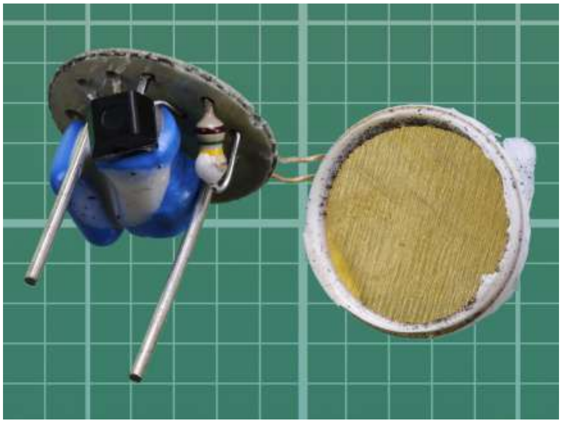

Figure 28-3. *The same indicator from the previous photograph, with its enclosure completely removed.*

Values

For an explanation and discussion of sound pressure and its measurement in decibels, see "Sound Pressure" on page 251 in the previous entry.

Audio indicators generally have a sound pressure rating in dBSPL ranging between 65dBSPL to 95dBSPL, with just a few products that make more or less noise. At 120dB and above, most products are packaged as alarm sirens ready for installation, often with a small horn attached. Their power consumption can be 200mA or more, and they are many times the price of a simple indicator designed for circuit-board mounting.

Voltage

An audio indicator containing its own electronics will almost always be rated somewhere in the range from 5VDC to 24VDC. Sirens intended for use with burglar alarms are often designed for 12VDC or 24VDC, as these are popular values for security systems with battery backup. However, in addition to a rated voltage, a datasheet may specify a wide range of acceptable operating voltages. For example, an indicator with a *rated* voltage of 12VDC may have an *operating* voltage of 3VDC to 24VDC. Naturally, the sound intensity will vary with the voltage, but not as much as one might assume. The graph at Figure 28-4 shows that the sound output from an alarm, measured in decibels, increases by only 8dB when voltage increases by almost a factor of five. Of course, the decibel scale is not linear, but human perception of sound is not linear, either.

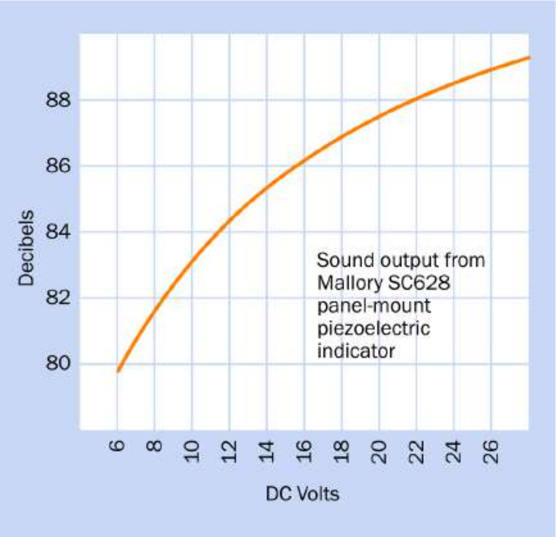

Figure 28-4. *Variation of sound output relative to voltage, in a commonly used piezoelectric indicator.*

Current

Typical piezoelectric indicators use less than 10mA (often as little as 5mA) and generate negligible heat.

Frequency

The most common frequencies for indicators range between 3kHz and 3.5kHz. Piezoelectric elements are inefficient for generating sounds below 1kHz.

Duty Cycle

Piezoelectric alerts generate very little heat and can be run on a 100% duty cycle.

If an alert will be pulsed briefly, the minimum pulse time is 50ms. A shorter duration will merely generate a clicking sound.

How to Use It

Appropriate Sound Intensity

An indicator should be chosen with reference to the environment in which it will be used. To be easily heard, it should be at least 10 dB louder than the ambient background noise.

Volume Control

Sound intensity can be reduced by reducing the voltage. Because an indicator does not consume much current, a trimmer can serve as a volume control. Alternatively, a rotary switch with a set of fixed-value resistors can select preset sound values.

However, in many indicators, variations in voltage may have relatively little effect on sound output, as shown in Figure 28-4.

Wiring

An indicator requires DC voltage. Because the indicator contains a transistor, polarity of the power supply is important. If the indicator has leads attached, the one intended for connection to the positive side of the power supply will be red. If it has pins, the longer pin will be for the positive connection.

What Can Go Wrong

The potential problems in an indicator are the same as those for a transducer. See "What Can Go Wrong" on page 254 in the previous entry.

headphone

29

The term **headphone** is used here to include almost any device that fits into or over the ear for the purpose of sound reproduction. (*Hearing aids* are not included.) Because headphones are used in pairs, the term is usually pluralized.

The term *phones* is fairly common as a colloquial diminution of *headphones* but is not used here.

An *earphone* used to be a single sound reproduction device designed for insertion into the ear, but has become rare. Pairs of *earbuds* are now common.

Because this encyclopedia assigns more emphasis to electronic components than to consumer products, this entry provides only a superficial overview of fully assembled headphones, and deals more with the drivers inside them, their principles of operation, and the general topic of sound reproduction.

OTHER RELATED COMPONENTS

- **transducer** (see Chapter 27)
- **speaker** (see Chapter 30)

What It Does

A headphone converts fluctuations of an electric signal into pressure waves that the human ear perceives as sound. It can be used for reproduction of music for entertainment purposes, or for speech in telecommunications, broadcasting, and audio recording.

Two symbols for headphones are shown in Figure 29-1. The symbol on the left shows a single headphone or earphone; when this symbol is flipped horizontally, it can represent a microphone. The pictographic symbol on the right has been used for many decades, but is still often found in schematics.

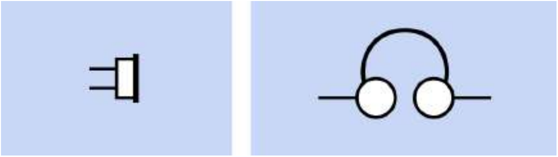

Figure 29-1. *Schematic symbols for a single earphone or headphone (left) and a pair of headphones (right).*

How It Works

Audio Basics

Sound is transmitted as pressure waves through a medium that is usually air but can be a gas, fluid, or solid. The speed of transmission will vary with the density and other attributes of the medium. Small hairs in the inner ears, known as *cilia*, vibrate in sympathy with pressure waves and transmit nerve impulses to the brain, which interprets the impulses as sound.

Three quantities describe the propagation of any type of wave, including a sound wave: its frequency (customarily represented with letter f), its speed of propagation (represented with letter v, for velocity), and its wavelength from peak-to-peak (represented by the Greek letter lambda, which appears as this λ symbol).

The relationship is defined by a very simple equation:

```
v = λ * f
```

Velocity is usually measured in meters per second, wavelength in meters, and frequency in Hertz, abbreviated Hz. One cycle per second is 1Hz. The H is always capitalized, as it refers to the name of Heinrich Rudolf Hertz, the first scientist to prove the existence of electromagnetic waves. One thousand Hertz can be written as 1 kiloHertz, almost always abbreviated as 1kHz (note that the k is lowercase).

The human ear is often described as being able to detect sounds between 20Hz and 20kHz, although the ability to hear sounds above 15kHz is relatively unusual and diminishes naturally with age. Sensitivity to all frequencies can be impaired by long-term exposure to loud noise.

Naturally occurring sounds can be converted to fluctuations in voltage by a **microphone**, which will be found listed as a *sensor* in Volume 3 of this encyclopedia. Artificial sounds can be generated as voltage fluctuations by oscillators and other electronic circuits. In either case, the output fluctuations can range between an upper limit set by a positive supply voltage and a lower limit established by electrical ground (which is assumed to be 0 volts). Alternatively, the fluctuations can range between the positive supply voltage and an equal and opposite negative supply voltage, with 0V lying midway between the two. This option can be less convenient electrically but is a more direct representation of sound, because sound waves fluctuate above and below ambient air pressure, which can be considered analogous to a ground state.

The concept of positive and negative sound waves is illustrated in Figure 29-2 (originally published in the book *Make: More Electronics*).

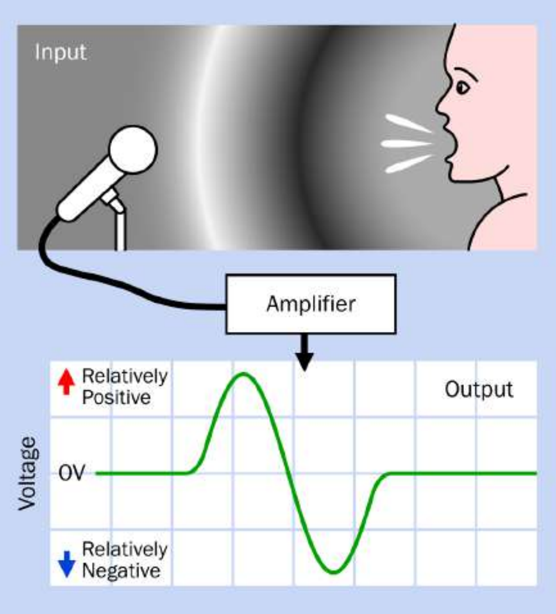

Figure 29-2. *The use of positive and negative voltages to represent a wave of high sound pressure followed by a trough of lower pressure.*

The topic of sound amplification is explored in detail in the entry on *op-amps* in Chapter 7.

A headphone inverts the function of a microphone by converting electricity back into air-pressure waves. This is done electromagnetically (moving a diaphragm in response to an electromagnet) or electrostatically (moving a membrane in response to electrostatic force between two charged electrodes).

Variants

Moving Coil

The most enduringly popular type of headphone uses a coil attached to a *diaphragm*. This is known as a *moving-coil* headphone, as the coil moves with the diaphragm. It can also be referred to as having a *dynamic driver* or *dynamic transducer*, “dynamic” referring to the movement of the coil.

The moving-coil concept is illustrated in Figure 29-3. The coil slides into a deep, narrow, circular slot in a magnet that is attached to the plastic frame of the headphone. The diaphragm is supported at its edges by a flexible rim. Variations in current passing through the coil create a fluctuating magnetic field that interacts with the field of the fixed magnet, causing the diaphragm to move in and out. A very similar configuration is used in many **loudspeakers**. Detail modifications may be made to increase efficiency, reduce production costs, or enhance sound quality, but the principle remains the same.

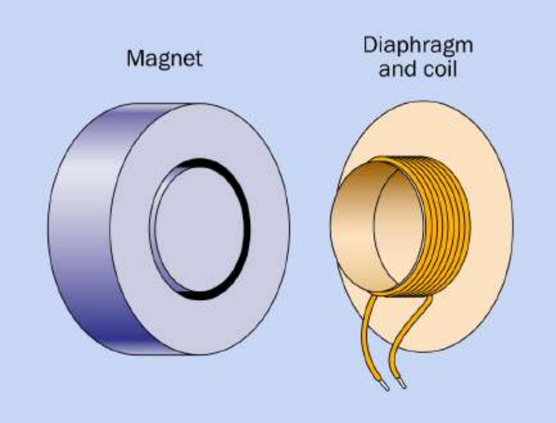

Figure 29-3. *The basic elements of a moving-coil headphone.*

The internal element of a headphone is shown in Figure 29-4. A plastic diaphragm is visible, measuring slightly less than 2" in diameter. The magnet and coil are concealed underneath.

The element in the previous figure is normally enclosed in an assembly such as the one in Figure 29-5, which incorporates a soft padded rim to rest upon the ear.

In an effort to achieve a more balanced frequency response, some designs use two moving-coil drivers in each headphone, optimized for low frequencies and high frequencies, respectively.

Earbuds, described after the next section, often use a miniaturized version of the moving-coil design.

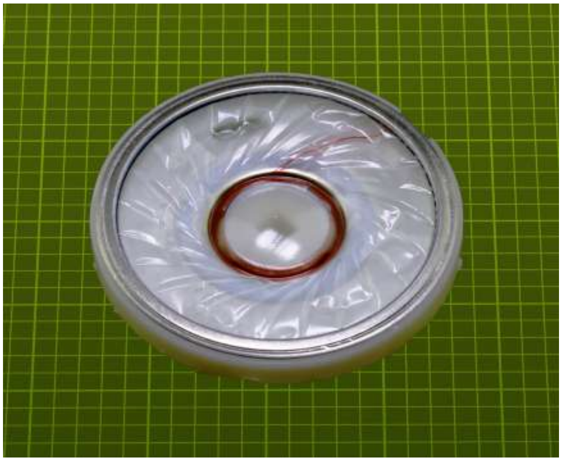

Figure 29-4. *The sound reproducing element removed from a headphone.*

Figure 29-5. *The sound reproducing element from the previous figure is normally packaged inside this enclosure.*

Other Types

Electrostatic headphones use a thin, flat diaphragm suspended between two grids that function as electrodes. A fluctuating potential between the grids, coupled with a reverse-phase voltage on the diaphragm, will vibrate the diaphragm, generating pressure waves. A relatively high voltage is necessary to achieve this, anywhere from 100V to 1,000V, supplied through a conversion unit between the headphones and an amplifier. Electrostatic headphones are known for low distortion and an excellent high frequency response, at some extra cost.

Electret headphones work on a similar principle, except that the membrane is permanently charged, and a high voltage is not required. Electret headphones tend to be small, inexpensive, and not of high sound quality.

A *balanced armature* design, often referred to by the acronym *BA*, uses a pivoting magnet that is claimed to increase efficiency while reducing stress on a diaphragm. BA drivers can be extremely compact, contained within a sealed metal enclosure measuring less than 10mm x 10mm x 5mm. They are commonly used in conjunction with *in-ear* earphones, described in the next section.

Mechanical Design

Circumaural headphones use large soft pads to encircle the ear and block external noise. Their size tends to make them heavy, requiring a well-designed headband to provide comfortable support. *Supra-aural* headphones are smaller and lighter, resting on the ears instead of enclosing them. They cannot exclude ambient noise, and may have inferior bass response compared with the circumaural type.

Open-back headphones, also known as *acoustically transparent*, are favored by some audiophiles because their vented outer surfaces are thought to create a more natural sound, similar to that of a **speaker**. The open backs naturally allow ambient noise to intrude, but also allow the sound generated by the headphones to be heard by others in a room. *Closed-back* headphones contain their sound and provide more insulation against ambient noise.

Earbuds rest just within the outer folds of the ear, facing inward like a pair of tiny **speakers**. They are easily dislodged and provide very little insulation against ambient noise. Their use became common after the introduction of Apple's iPod. A pair of earbuds, one of them with its plastic cover removed, is shown in Figure 29-6.

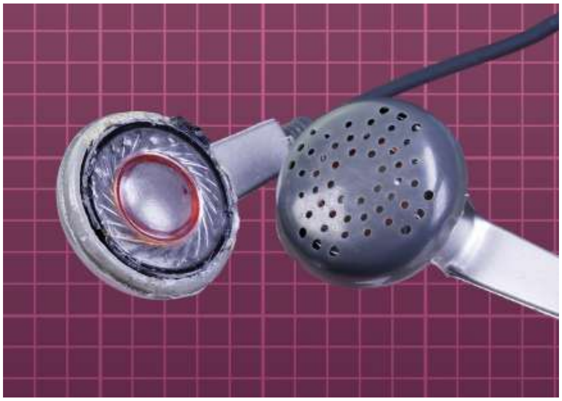

Figure 29-6. *A pair of earbuds, one with its cover removed to show the sound reproducing element, which closely resembles the diaphragm in a full-sized headphone.*

In-ear headphones are designed for insertion into the ear canal, often using a soft sheath that conforms with the ear like an earplug. This sheath is disposable for hygienic reasons, and because it may lose some of its plasticity with use. It excludes most ambient noise, and by minimizing the air gap between the driver of the headphone and the ear drum enables a high quality of sound reproduction.

In-ear headphones are also known as *in-ear monitors*, *IEMs*, *ear canal headphones*, *earphones*, and *canalphones*. A pair of in-ear headphones is shown in Figure 29-7, one of them with its foam sheath removed. The rectangular silver-colored object in the headphone on the left contains a transducer to create sound pressure.

A *headset* consists of one or two headphones plus a flexible microphone that extends to the promiximity of the mouth of the user.

Noise-cancelling headphones, popularized by Bose, monitor external noise with a built-in microphone and generate sound of opposite phase, to provide some cancellation. They are particularly effective on jet aircraft, where background noise tends to be consistent.

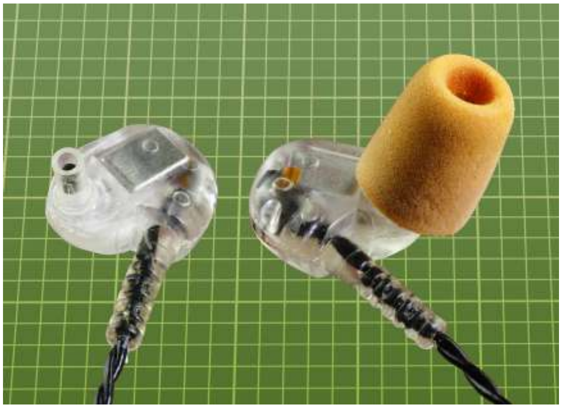

Figure 29-7. *A pair of in-ear headphones, supplied with disposable foam plugs that conform flexibly with the ear canal. The headphone on the left is shown with its plug removed.*

Although an *earphone* is almost obsolete, it is still obtainable from specialty suppliers. It has a high impedance, making it suitable for use with a *crystal-set radio*. An earphone is pictured in Figure 29-8.

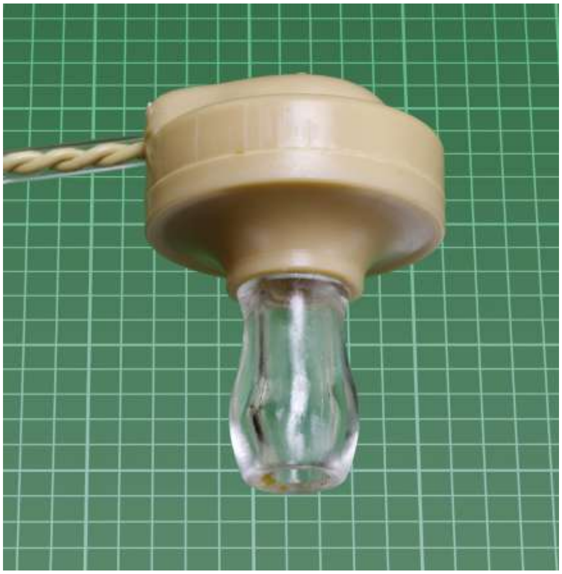

Figure 29-8. *A vintage earphone of the type suitable for use with a crystal-set radio.*

Values

Intensity

Sound pressure is measured in decibels. For a complete explanation and discussion of weighted and unweighted decibel scales, see "Sound Pressure" on page 251 in the **transducer** entry.

Frequency Response

A plot of sound pressure as a function of frequency shows the *frequency response* of a headphone. Measuring the sound pressure meaningfully is a challenge, because the ear canal will add coloration to the sound and can amplify some frequencies while masking others. Ideally, measurement should be done at the ear drum, but this is not feasible. Consequently, high-end headphones are evaluated by making sound measurements inside simulated ear canals in a dummy human head.

A comparison between a high-quality $500 audio product and a transducer that is sold as a component for less than $1 illustrates the difference in frequency response; see Figure 29-9. The Sennheiser headphones have a smooth response that rises toward the low end, compensating for the lack of bass response that tends to be a problem in headphones, and the relative insensitivity of the human ear to low frequencies. The fluctuations at the high end are within about 5dB.

By comparison, the Kobitone emphasizes the range between 3kHz and 4kHz because its primary task is to be heard, and these are the frequencies where human hearing is most sensitive. Its low-frequency response trails off (although is still much better than that of a piezoelectric transducer, where the low response typically diminishes by 40dB to 50dB). The low-frequency output of the Kobitone is actually impressive bearing in mind that the component is only 9mm in diameter. It draws 60mA at 5VAC.

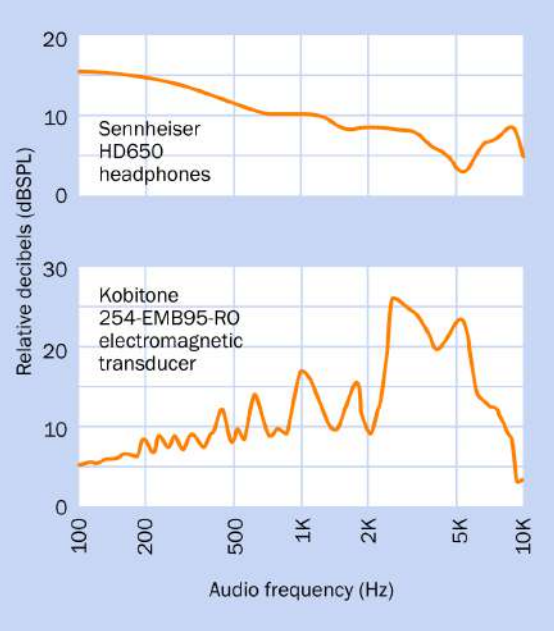

Figure 29-9. *Comparison between frequency responses of a $1 electromagnetic transducer intended as an audio alert and a $500 pair of headphones intended for sound reproduction. The upper graph is derived from a review online at headroom.com; the lower graph is from the manufacturer's datasheet.*

Some manufacturers of audio reproduction equipment prefer not to supply frequency response curves. Instead they may claim, for example, that the frequency response of a product ranges from 100Hz to 20kHz. This claim means very little unless it is accompanied by a range of sound pressure levels. If the frequency response is consistent within a range of, say, plus-or-minus 5dB, this may be acceptable. If the range is plus-or-minus 20dB, it is not acceptable. The ability to reproduce a high note or a low note is not useful if the sound is too faint to hear.

Distortion

The *total harmonic distortion* (THD) of any audio equipment measures its tendency to add spurious *harmonics* of a single frequency. If headphones are required to reproduce a pure 1kHz sinewave, they will also tend to create an additional 3kHz tone that is an artifact. This can be caused by the mechanical behavior of a vibrating diaphragm. The human ear recognizes distortion as a fuzzy or rasping sound. A square wave theoretically contains all the harmonics that are odd multiples of its fundamental frequency, and sounds extremely distorted.

THD should be less than 1% in good-quality audio devices.

Impedance

The electrical impedance of headphones is relevant in that it should match the output specification of the amplifier that drives them.

What Can Go Wrong

Overdriving

Headphones can be damaged by overdriving them. Because a low frequency requires larger excursions of a diaphragm to transmit the same energy as a high frequency, headphones are especially vulnerable to being damaged by bass at high volume.

Hearing Damage

Human hearing can be damaged by prolonged listening to headphones at a high volume. Some controversy remains regarding an acceptable limit for sound pressure.

Mismatched Impedance

If the impedance of headphones does not match the output of the amplifier driving them, distortion or a skewed frequency response can result. This is known as *mismatching*.

Incorrect Wiring

In most consumer products, a pair of headphones will share a common ground. While the connections in a typical three-layer jack plug have been standardized, hand-wired repairs or extensions should be tested carefully. Incorrect wiring will cause unpredictable results.

speaker

30

The term **speaker** is a diminution of *loudspeaker*. The full word is now so rarely used, some catalogs do not recognize it as a search term. This encyclopedia acknowledges contemporary usage by using **speaker** rather than *loudspeaker*.

A fully assembled consumer product can be referred to as a "speaker," but it also contains one or more individual components that are described as "speakers." To resolve this ambiguity, referring to the components as *drivers* is helpful, but this practice can create more ambiguity because other types of components are also called "drivers." The only real guide to the meaning of *speaker* is the context in which it is used.

For the purposes of this entry, a speaker is a *sound reproduction device*, distinguished from a typical electromagnetic **transducer** by being larger and more powerful, with a more linear frequency response. A transducer may be used as a *noise-creating device* to provide an alert, informing the user of the status of a piece of equipment. Because some speakers have been miniaturized for use in handheld products, they may be used as transducers, allowing some overlap between the two categories.

Because this encyclopedia assigns more emphasis to electronic components than to consumer products, this entry provides only a superficial overview of fully assembled speakers, and deals more with the drivers inside them, their principles of operation, and the general topic of sound reproduction.

OTHER RELATED COMPONENTS

- **headphone** (see Chapter 29)
- **transducer** (see Chapter 27)

What It Does

A speaker converts fluctuations of an electric signal into pressure waves that the human ear perceives as sound. It can be used for entertainment purposes or to provide information in the form of spoken words or distinctive sounds (as in the case of a miniature speaker in a cellular phone, playing a ring tone).

The internationally accepted schematic symbol for a speaker is shown in Figure 30-1.

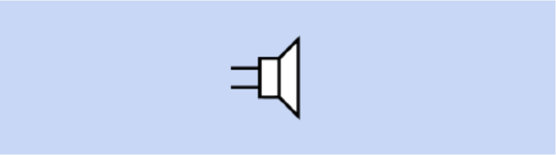

Figure 30-1. *Only one symbol exists to represent a speaker. This is it.*

How It Works

For a summary of basic concepts and terminology relating to sound and its reproduction, see "Audio Basics" on page 261 in the previous entry.

Construction

A speaker contains a *diaphragm* or *cone* with a coil attached to it. Fluctuations of current through the coil interact with a permanent magnet, causing the speaker to emit pressure waves that are proportionate with the current. The design is similar in concept to that of a headphone, shown diagrammatically in Figure 29-3. The primary difference is that a speaker of around 2″ or more will use a cone rather than a flat diaphragm. The cone shape is more rigid and creates a more directional sound.

A 2″ speaker rated for 1/4W with a 63Ω coil is shown in Figure 30-2, undamaged on the left but with its cone cut out on the right. The neck of the cone, which is normally inserted in the circular groove in the speaker magnet, is shown with the inductive coil wrapped around it.

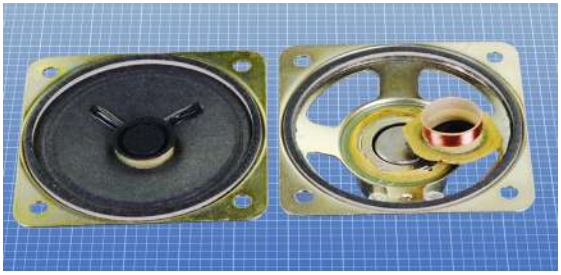

Figure 30-2. *On the left is a 2″ speaker. On the right, its cone has been cut away to reveal the magnet, with a circular groove in it. The neck of the cone, which normally slides into the groove, is shown removed.*

A speaker with a cone 4″ in diameter is shown from the rear in Figure 30-3.

A miniature surface-mount speaker is shown from the front and from the rear in Figures 30-4 and 30-5. It measures just under 0.4″ diameter and was made for Motorola. Its power rating is 50mW.

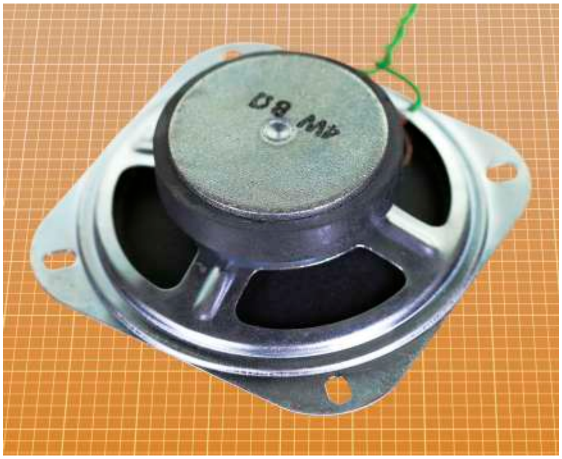

Figure 30-3. *The back side of a speaker with a cone measuring approximately 4″ in diameter. Its magnet is the large round section that is uppermost. This unit is rated for 4W and has an impedance of 8Ω.*

A speaker designed for a cellular phone is shown in Figure 30-6. Note the close resemblance in design to the driver used in an earbud, shown in Figure 29-6 in the previous entry.

In the past, speaker cones were made from tough, fibrous paper. Modern cones are more likely to be plastic, especially in small sizes.

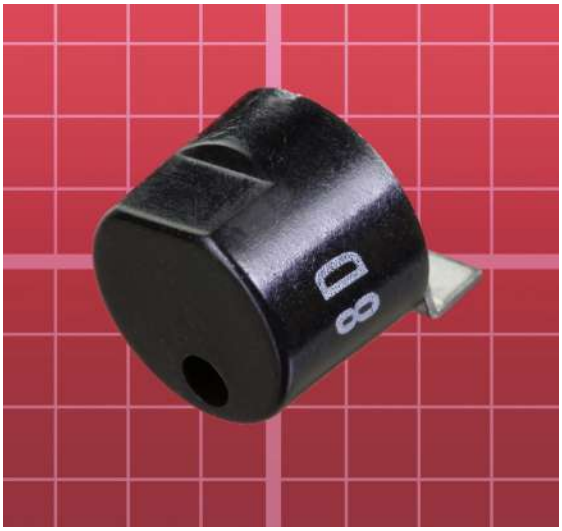

Figure 30-4. *Front view of a surface-mount speaker measuring less than 0.4″ diameter.*

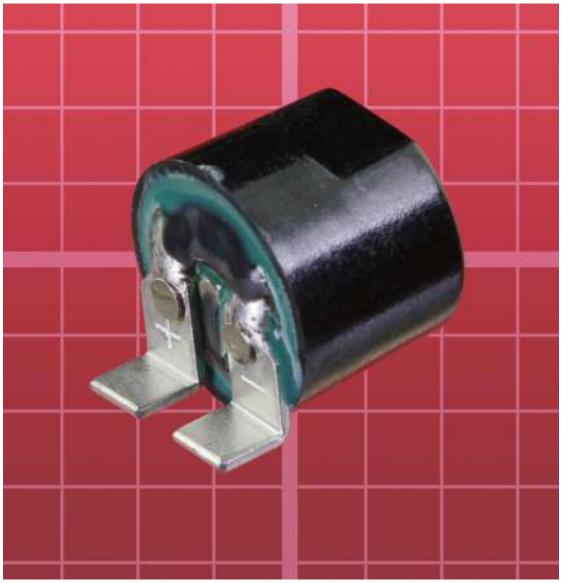

Figure 30-5. *Rear view of the speaker shown in the previous figure.*

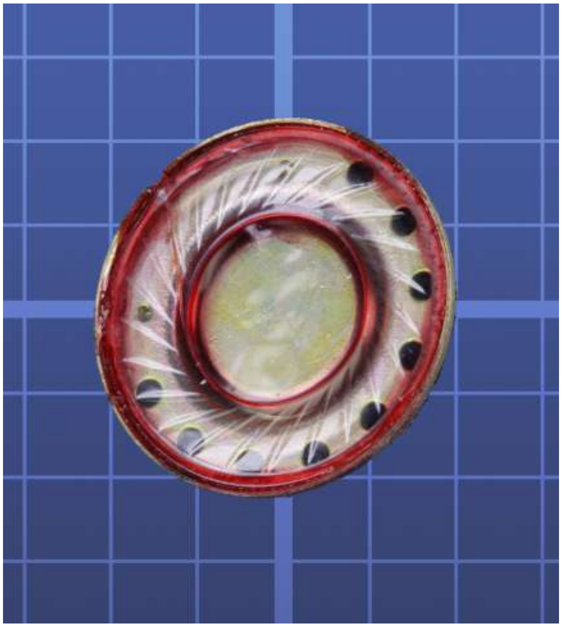

Figure 30-6. *A miniature speaker measuring slightly more than 1/2" diameter and only 0.13" thick, designed for use in a cellular phone. It has an impedance of 150Ω.*

Multiple Drivers

Generally speaking, a large-diameter speaker cone is more effective than a small cone at moving the greater volumes of air associated with reproduction of bass notes. However, the inertia of a large cone impairs its ability to vibrate at high frequencies.

To address this problem, a large speaker and a small speaker often share a single enclosure. A *crossover network* using coils and capacitors prevents low frequencies from reaching the small speaker and high frequencies from reaching the large speaker. The basic principle is shown in the simplified schematic in Figure 30-7.

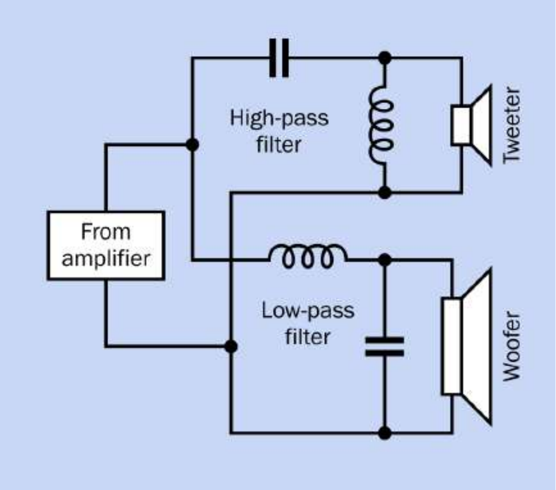

Figure 30-7. *The basic principle of a crossover network.*

Because the crossover network must be "tuned" to match the characteristics of the speakers, and because the combined sound pressure of the speakers must be relatively consistent over a wide range of frequencies, an actual network usually includes additional components.

Because the audio output from an amplifier consists of alternating current, polarized capacitors cannot be used. Polyester capacitors are common.

The small speaker in a pair is known as a *tweeter* while the large speaker is a *woofer*. Although these appellations were whimsically coined, they have endured.

More than two speakers may be combined in an enclosure, in a wide variety of configurations.

Venting

A speaker radiates pressure waves from the back side of its cone as well as from the front, and because the waves from front and back are opposite in phase, they will tend to cancel each other out.

In a miniature speaker, this problem can be addressed simply by sealing the section of the enclosure at the rear. For larger components, a more efficient enclosure can be designed with a *vent* or *reflex port* at the front. Pressure waves from the back of the speaker are diverted over a sufficient distance inside the cabinet so that by the time they emerge through the port, they are approximately in phase with low frequencies from the front of the speaker, although the back wave will lag the front wave by one wavelength.

This design is referred to as a *bass-reflex enclosure*, and was almost universal in high-fidelity components until amplifiers became increasingly powerful during the 1960s. At that point, Acoustic Research, located in Massachusetts, marketed a product line in which speaker enclosures were sealed, the argument being that when an amplifier can deliver 100W per channel, efficiency is no longer an issue, and a sealed enclosure can eliminate compromises associated with a bass-reflex design.

Acoustic Research referred to their concept as "air suspension," as the cushion of air in the sealed cavity helped to protect the speaker by limiting its excursion. This configuration is now often referred to as a *closed-box speaker*. Some audiophiles argue that it must always be inherently superior to a bass-reflex design, partly because of the one-wavelength lag time associated with a reflex port. However, as in many aspects of sound reproduction, the debate is inconclusive.

Resonance

The enclosure for a speaker will tend to have a dominant resonant frequency. This should be lower than the lowest frequency that the speaker will reproduce; otherwise, the resonance will emphasize some frequencies relative to others, creating unwanted peaks in the response.

One reason why high-quality speakers tend to be physically heavy is to reduce their resonant frequency. A modern Thiel speaker assembly, for instance, uses a front panel of particle board that is a full 2" thick. However, heavy enclosures are expensive to transport and inconvenient to locate or relocate in the home.

To address this problem, a tweeter and a woofer can be mounted in separate boxes. The enclosure for the tweeter can be very small, lightweight, and suitable for placement on a shelf, while the heavy box for the woofer can go on the floor. Human senses have difficulty locating the source of low-frequency sound, so the woofer can be located almost anywhere in a room. In fact, its single speaker can serve both stereo channels.

This configuration has become the default for computer speakers. It is also used in home-theater systems, where the woofer has now become a subwoofer capable of very low frequency reproduction.

Miniature Speakers

If an electronics project has an audio output, and the circuit board will be sharing an enclosure with a small speaker, the size of the box and the material from which it is fabricated will affect the sound quality significantly. A box made from thin hardwood may add resonance that sounds pleasing, if the speaker is being used just for simple electronic tones. By comparison, a metal box may sound "tinny." A box fabricated from a plastic such as ABS will be relatively neutral, provided the plastic is reasonably thick (1/4" being preferable).

Variants

Electrostatic Speaker

The principle of an electrostatic speaker is the same as that of an electrostatic headphone. A charged membrane is stretched between two grids in front of it and behind it that act as electrodes. Because the membrane is so light, it responds with very little latency, and its large surface area creates a diffused sound that many audiophiles find pleasing. However, a high voltage is required to drive electrostatic speakers, and they are not cheap.

Powered Speakers

A unit containing its own driver electronics is referred to as a *powered speaker*, and is used almost universally with desktop computers, because the computer itself does not contain a power amplifier. Powered speakers may also enable a more versatile crossover network.

A subwoofer may have its own amplifier allowing control of the cutoff frequency above which the speaker will not attempt to reproduce sound. The electronics can include protection for a speaker against being overdriven.

Wireless Speakers

A wireless link between a stereo receiver and its speakers will eliminate the speaker wires that are normally necessary. However, the speakers themselves must be powered, and will have to be wired to electric outlets.

Innovative Designs

The need for small speakers in consumer products such as laptop computers has encouraged innovative designs. The speaker in Figure 30-8 is just 1″ square, and its shape is easier to accommodate in a small product than the traditional circular speaker. In Figure 30-9, the interior of the same speaker shows that inductive coils are applied to a square plastic diaphragm.

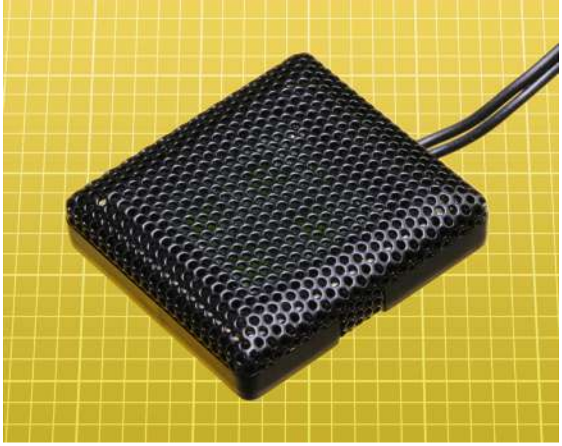

Figure 30-8. *A speaker 1″ square, suitable for use with a small electronic device.*

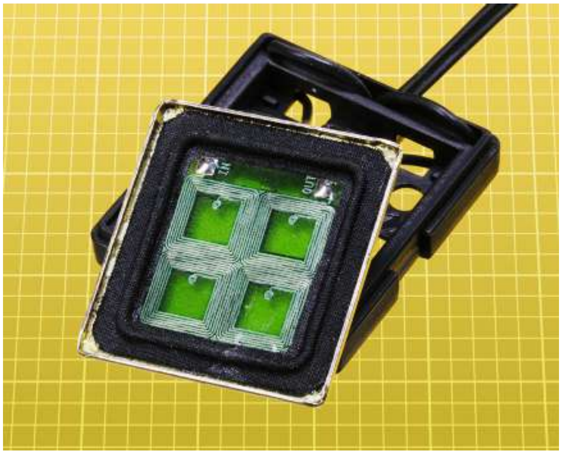

Figure 30-9. *The speaker from the previous figure, opened to show its inductive coils applied to a square plastic diaphragm.*

Values

The typical *impedance* for speakers in audio systems is 8Ω. Small speakers may have a higher impedance, which can be useful when driving them from devices that have limited power, such as a TTL-type 555 timer.

In the United States, the diameter of a circular speaker is usually expressed in inches. Speakers larger than 12″ are rare for domestic use. A 4″ speaker used to be considered minimal because of its limited low-end frequency response, but much smaller speakers have become common in portable devices.

The low-end frequency response of a miniature loudspeaker designed to be surface-mounted on a circuit board will be very poor. The graph in Figure 30-10 was derived from data supplied by the manufacturer.

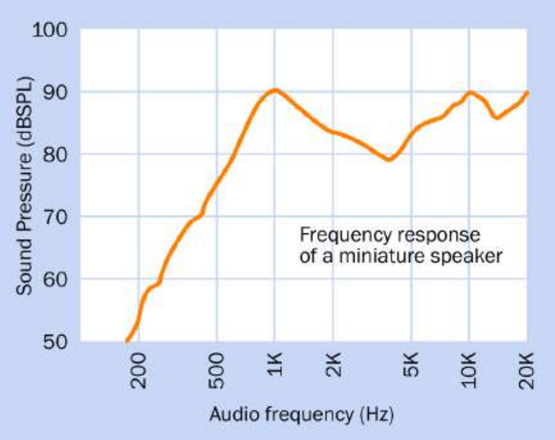

Figure 30-10. *Frequency response for a speaker measuring 15mm x 15mm x 5mm. Small dimensions and lack of an enclosure result in a negligible low-end response.*

Power rating for speakers is in watts, measured on a root-mean-square (RMS) basis.

Sensitivity is measured in decibels, at a distance of one meter, while a speaker is reproducing a single constant tone with a power input of 1W. A speaker assembly designed for undemanding home use may be rated at 85dB to 95dB.

Efficiency is a measure of sound power output divided by electrical power input. A value of 1% is typical.

What Can Go Wrong

Damage

As is the case with headphones, the most common problem affecting a speaker is damage caused by overdriving it. Because a low frequency requires larger excursions of a speaker cone to transmit an amount of energy comparable to that of a high frequency, loud bass notes can be hazardous to the cone. On the other hand, if an amplifier generates distortion (perhaps because it, too, is being overdriven), the harmonics created by the distortion can damage high-frequency speakers.

Magnetic Field

Even a small speaker (measuring 2" or less) contains a magnet that is sufficiently powerful to cause problems if it is in close proximity to other components, especially if sensors such as reed switches or Hall-effect switches are being used. Initial circuit testing should be done with the speaker as far away as possible, to eliminate it as a source of interference.

Vibration

Solder joints will be stressed if they are subjected to low-frequency vibration from a speaker. Loose parts may rattle, and parts that are bolted into place may become unbolted. The speaker itself may become loose. Thread-locker such as Loc-Tite should be applied to nuts before they are tightened.

Preface

This third and final volume of the *Encyclopedia of Electronic Components* is devoted entirely to sensors.

Two factors have caused very significant changes in the field of sensors since the 1980s. First, features such as antilock braking, airbags, and emissions controls stimulated the development of low-priced sensors for automotive applications. Many of these sensors were fabricated in silicon as MEMS (microelectromechanical) devices.

The second wave began in 2007 when MEMS sensors were installed in the iPhone. A modern phone may contain almost a dozen different types of sensors, and their size and price have been driven down to a point that would have been unimaginable 20 years previously.

Many MEMS sensors are now as cheap as basic semiconductor components such as a voltage regulator or a logic chip, and they are easy to use in conjunction with microcontrollers. In this Encyclopedia, we have allocated significant space to this segment of the market, hoping that the specific products that we have chosen will remain popular and available for at least the next decade.

In addition, we have devoted space to older components where durability has been proven.

Purpose

While much of the information in this volume can be found dispersed among datasheets, introductory texts, Internet sites, and technical briefings published by manufacturers, we believe there is a real need for a durable resource that assembles all the relevant data in one place, properly organized and verified, including details that may be hard to find elsewhere.

This volume may also serve a useful purpose by attempting to categorize and classify components in a field that is remarkably chaotic. For example, is an *object presence sensor* different from a *proximity sensor*? Some manufacturers seem to think so; others disagree. Understanding the distinctions and the underlying principles can be important if you are trying to decide which sensor to use.

Sensor terminology can also be confusing. To take another example, what is the difference between a *reflective interrupter*, a *reflective object sensor*, a *reflective optical sensor*, a *reflective photointerrupter*, and an *opt-pass sensor*? These terms are used in various datasheets to describe components that are all *retroreflective sensors*. Understanding the proliferating variety of terminology can be essential if you simply want to find something in a product index.

Organization

As in volumes 1 and 2, this volume is organized by subject. For example, if you want to measure temperature, you'll find the entries for a thermistor and a thermocouple next to each other, in an entire section devoted to the sensing of heat. This will help you to compare capabilities and choose the component that best suits your application.

The subject path leading to each sensor is shown at the top of the first page of each entry. For gas flow rate, for instance, you would follow this path:

```
fluid > gas > flow rate
```

Note that the word "fluid" is properly used to include gases as well as liquids.

Exceptions and Conflicts

Unfortunately, some sensors are not easily categorized. There are four problems in this area.

1. What Does a Sensor Really Sense?

A GPS chip is a radio receiver, picking up transmissions from satellites. Does this mean it should be categorized as a sensor of radio waves? No, its purpose is to tell you your location. Therefore, it is categorized as a location sensor. This leads to the first general rule: sensors are categorized by their primary purpose. Secondary purposes may be found in the index.

2. How Many Sensors Are in a Sensor?

Many surface-mount chips perform more than one sensing function. For example, an inertial measurement unit (often identified by its acronym, *IMU*) can contain three gyroscope sensors and three accelerometers—and may contain three magnetometers, too. How should it be categorized?

The answer is that an IMU will be mentioned in more than one entry in the Encyclopedia, because it performs more than one function; but it will not have its own separate entry, because each entry in the Encyclopedia is for a single primary sensing function.

The names of multisensor chips are, of course, included in the index.

3. How Many Stimuli Can One Sensor Sense?

A single sensing element may be used in multiple different types of sensors. The most notable example is the Hall-effect sensor, which can be found in magnetometers, object presence sensors, speed sensors, current sensors, and dozens more. Modern automobiles can contain Hall-effect sensors everywhere from the ignition system to the trunk-locking mechanism. If you are using a hard drive with rotating platters, it probably contains a Hall-effect sensor to monitor the speed of rotation. If you have a generic computer keyboard, each keypress is probably detected with a Hall-effect sensor.

Bearing this in mind, how should a Hall-effect sensor be classified? And where should you expect to find an explanation of how it works?

The answer is that where different types of components contain the same type of sensing element, the entry for each component will include a cross-reference to one location where the sensing element is explained in detail.

This location will be chosen for its relevance. Thus, Hall-effect sensors are explained in the entry for **object presence** sensors, because this is their primary function. While it is true that a Hall-effect sensor works by detecting a magnetic field, that is not its most common application.

4. Too Many Sensors!

Wikipedia lists more than 100 general types of sensors, and even that list is probably not complete. Consequently, we had to pick and choose. Some of the decisions may seem arbitrary, but all of them were made on the grounds of practicality. There were three principles for deciding what to include and what to leave out.

1. Is it a component? We are more interested in board-mounted components than in packaged products that happen to contain sensors. For instance, a thermocouple is often enclosed in a tubular steel probe, and its wire is often plugged in to a specially designed meter that displays temperature. While we do include a photograph of a probe, we are primarily interested in the welded wires of the thermocouple inside it.
2. How much does it cost? An industrial ultrasonic sensor to check items on a factory conveyor belt will be sealed into a module with a waterproof grommet around a shielded cable—which is all very nice, but will not be very affordable. This Encyclopedia is more interested in board-mountable components for one-tenth of the price.
3. How many people are likely to want it? The stock of each type of sensor was checked on component vendor sites. If a sensor wasn't in the inventory, or if only a couple of variants were stocked, we concluded that the limited demand probably didn't justify including it here. For example, a Ferraris acceleration sensor responds to eddy currents in a rotating motor shaft, as a way of measuring vibration in the shaft. This is a really interesting device, but is unlikely to be on most people's shopping lists.

Volume Contents

Having explained the organization of this book and our decisions to include or omit various components, we now present a summary of the contents of all three Encyclopedia volumes:

Volume 1

Power; electromagnetic devices; discrete semiconductors.

The *power* category includes sources of electricity and methods to distribute, store, interrupt, convert, and regulate power. The *electromagnetism* category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the primary types of diodes and transistors. See Figure P-1 for a contents listing.

Primary Category	Secondary Category	Component Type
power	source	battery
	connection	jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electro-magnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
		stepper motor
discrete semi-conductor	single junction	diode
		unijunction transistor
	multi-junction	bipolar transistor
		field-effect transistor

Figure P-1 *The subject-oriented organization of categories and entries in Volume 1 of this Encyclopedia.*

Volume 2

Thyristors (SCRs, diacs, and triacs); integrated circuits; light sources, indicators, and displays; and sound sources.

Integrated circuits are divided into analog and digital components. *Light sources, indicators, and displays* are divided into reflective displays, single sources of light, and displays that emit light. *Sound sources* are divided into those that create sound, and those that reproduce sound. A contents listing for Volume 2 appears in Figure P-2.

Volume 3

All the most common types of sensing devices, including those that detect location, presence, proximity, orientation, oscillation, force, load, human input, liquid properties, gas types and concentrations, pressure, flow rate, light, heat, sound, and electricity. A contents listing for Volume 3 appears in Figure P-3.

Method

Reference Versus Tutorial

As its title suggests, this is a reference book, not a tutorial. A tutorial such as *Make: Electronics* begins with elementary concepts and builds sequentially toward concepts that are more advanced. A reference book assumes that you may dip into the text at any point, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will find some repetition, as each entry is intended to be self-sufficient, requiring minimal reference to other entries.

Theory and Practice

This book is oriented toward practicality rather than theory. We assume that the reader mostly wants to know how to use electronic components, rather than why they work the way they do. Consequently we do not include detailed proofs of formulae or definitions rooted in electrical theory.

Primary Category	Secondary Category	Component Type
discrete semi-conductor	thyristor	SCR
		diac
		triac
integrated circuit	analog	solid-state relay
		optocoupler
		comparator
		op-amp
		digital potentiometer
		timer
	digital	logic gate
		flip-flop
		shift register
		counter
		encoder
		decoder
		multiplexer
light source, indicator or display	reflective	LCD
	single source	incandescent lamp
		neon bulb
		fluorescent light
		laser
		LED indicator
		LED area lighting
	multi-source or panel	LED display
		vacuum-fluorescent
		electroluminescence
sound source	audio alert	transducer
		audio indicator
	reproducer	headphone
		speaker

Figure P-2 *The subject-oriented organization of categories and entries in Volume 2.*

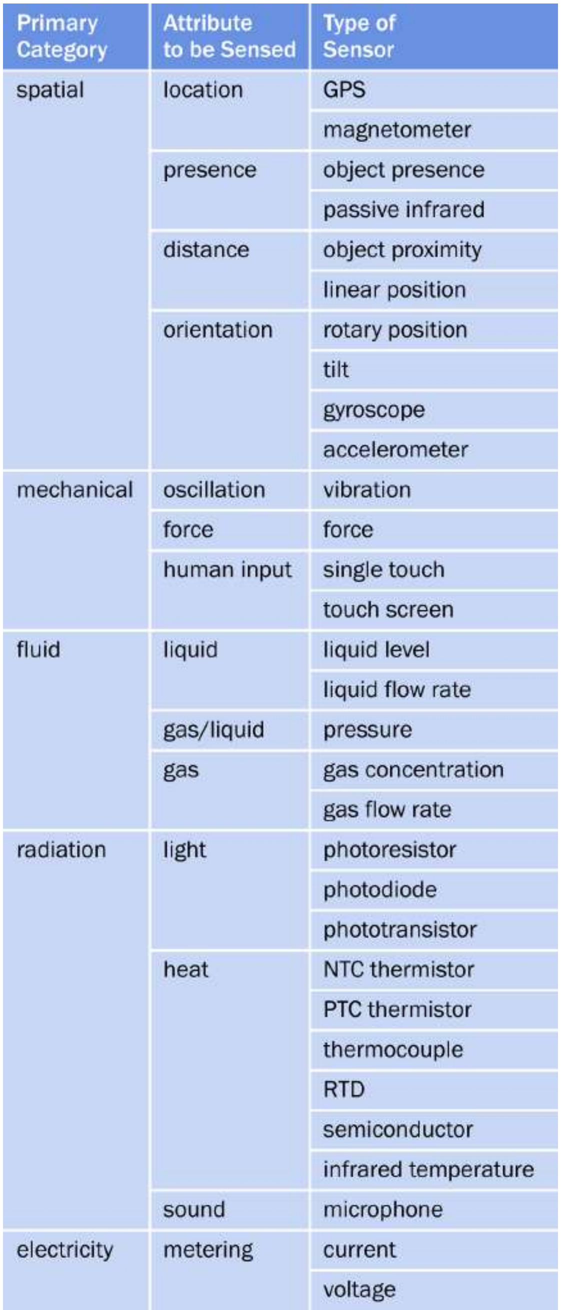

Primary Category	Attribute to be Sensed	Type of Sensor
spatial	location	GPS
		magnetometer
	presence	object presence
		passive infrared
	distance	object proximity
		linear position
	orientation	rotary position
		tilt
		gyroscope
		accelerometer
mechanical	oscillation	vibration
	force	force
	human input	single touch
		touch screen
fluid	liquid	liquid level
		liquid flow rate
	gas/liquid	pressure
	gas	gas concentration
		gas flow rate
radiation	light	photoresistor
		photodiode
		phototransistor
	heat	NTC thermistor
		PTC thermistor
		thermocouple
		RTD
		semiconductor
		infrared temperature
	sound	microphone
electricity	metering	current
		voltage

Figure P-3 *The subject-oriented organization of categories and entries in Volume 3.*

Sensor Output

In Volumes 1 and 2 of the *Encyclopedia*, each entry included hints on how to use a component. However, many sensors have identical forms of output, which are processed in a similar way. To avoid repetition, general guidance for using nine principal types of sensor outputs has been placed in Appendix A at the back of this volume.

For example, many sensors provide an analog voltage output that varies with the phenomenon that is being sensed. In Appendix A, you will find suggestions on how to adjust the range of the output, if necessary, or how to digitize it with an analog-to-digital converter.

You will also find a comparison between serial protocols such as I2C and SPI, both of which are commonly used when a microcontroller communicates with a digital sensor via a bus.

Glossary

In the world of sensors, many terms tend to recur. *Hysteresis* is one; *MEMS* is another. Rather than define these terms repeatedly, some quick definitions are gathered in a Glossary. Please remember the existence of the glossary if you encounter a term that is unfamiliar. See Glossary.

In many instances, terms that are italicized in the text are defined in the glossary.

Typographical Conventions

Within each entry, **bold type** is used for the first occurrence in each entry of the name of a component that has its own entry elsewhere. Other important electronics terms or component names may be presented in *italics*.

The names of components, and the categories to which they belong, are all set in lowercase type, except where a term is normally capitalized because it is an acronym or a trademark, or contains a proper noun. The term *Hall effect*, for instance, has an initial cap because it is named after a person named Hall. The term *GPS* is all in caps, because it is an acronym; but *psi* (meaning pounds per square inch) remains in lowercase, because even though it is an acronym, the lowercase form is more common.

The situation is different when specifying units that are named after electrical pioneers. All of these units should be lowercased when spelled out. Thus, when referring to the SI unit of force, it is "the newton." However, where a unit named after a person is abbreviated, the abbreviation is capitalized, as in N for newtons, Hz for hertz, Pa for pascals, and A for amperes.

Mathematical Syntax

In mathematical formulae, we have used the style that is common in programming languages. The * (asterisk) is used as a multiplication symbol, while the / (forward slash) is used as a division symbol. Where some terms are in parentheses, they must be dealt with first. Where parentheses are inside parentheses, the innermost ones must be dealt with first. Consider this example:

```
A = 30 / (7 + (4 * 2) )
```

You would begin by multiplying 4 times 2, to get 8; then add 7, to get 15; then divide that into 30, to get the value for A, which is 2.

Visual Conventions

Figure P-4 shows the conventions that are used in the schematics in this book. A black dot always indicates a connection, except that to minimize ambiguity, the configuration at top-right is avoided, and the configuration at top-center is used instead. Conductors that cross each other without a black dot do not make a connection. The styles at bottom right are sometimes seen elsewhere, but are not used here.

All the schematics are formatted with pale blue backgrounds. This enables components such as switches, transistors, and LEDs to be highlighted in white, drawing attention to them and clarifying the boundary of the component. The white areas have no other meaning.

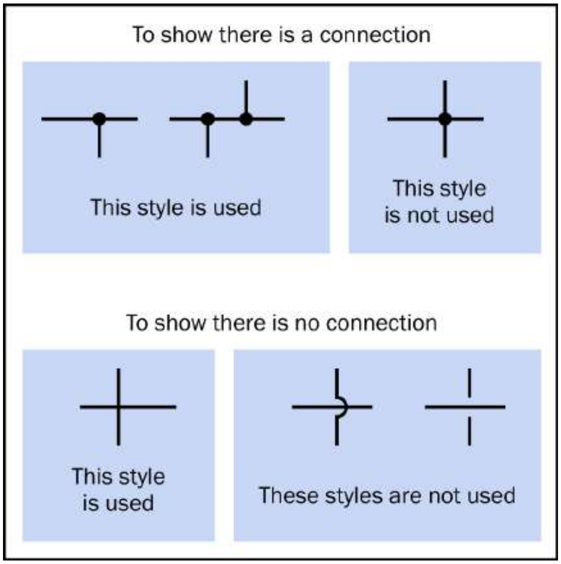
To show there is a connection
This style is used
This style is not used
To show there is no connection
This style is used
These styles are not used

Figure P-4 *Visual conventions that are used in the schematics in this book.*

Units and Backgrounds

So long as the United States clings stubbornly to the habit of expressing dimensions in inches, there is a good argument to follow this custom in books intended for an American audience. With this in mind, Volumes 1 and 2 mostly avoided metric units of length. However, as time passed, the books were translated for use in many parts of the world where the inch is regarded as an anachronism.

Recognizing that we now have an international audience, we have used the metric system throughout this volume (with very few exceptions, such as a photograph of an American plumbing fixture that is designed to fit 3/4" pipe). For readers who are metrically impaired, here are some units of length, and their abbreviations:

- 1 nanometer (nm)
- 1 micrometer (μm) = 1,000nm
- 1 millimeter (mm) = 1,000μm
- 1 centimeter (cm) = 10mm
- 1 meter (m) = 100cm = 1,000mm

A micrometer is also known as a *micron*.

The basic conversion factor from meters to inches is 0.0254. Thus:

- 1 inch = 2.54cm = 25.4mm
- 1/1000 inch = 25.4µm

Sometimes 1/1000 inch is called a *mil*.

In many of the component photographs, a graph-paper background is included. Each square in these backgrounds is 1mm.

To avoid confusion, please remember that a few of these same component photographs appeared in books such as *Make: More Electronics*, where the background grid was in tenths of an inch. Captions to photographs in this volume will remind you that millimeters are now used.

Background colors in the photographs were chosen for contrast with the colors of the components, or for visual variety. They have no other significance.

Component Availability

The world of sensors is changing rapidly, and we have no way of knowing if a component will enjoy a long production run. We recommend checking availability at the following suppliers, which we used frequently during the preparation of the book:

- *http://www.mouser.com*
- *http://www.jameco.com*
- *http://www.sparkfun.com*
- *http://www.adafruit.com*

For obsolete parts, or those that are nearing the end of their commercial life, eBay can be very useful. Alternatively, new substitutions for old parts are often listed at *http://www.mouser.com*.

Issues and Errata

There are three situations where the reader and the writer may want to communicate with each other.

- We may want to tell you if the book contains a mistake of some significance. This is *us-informing-you* feedback.
- You may want to tell us if you think you found an error in the book. This is *you-informing-us* feedback.
- You may be having trouble making something work, and you don't know whether we made a mistake or you made a mistake. You would like some help. This is *you-asking-us* feedback.

Here's how you can deal with each of these situations.

Us Informing You

If you already registered your contact information in connection with *Make: Electronics* (second edition) or *Make: More Electronics*, you don't need to register again for updates relating to the Encyclopedia. If you have not already registered, here's how it works.

The only way you can be notified if there's an error in the book is if you supply your contact information. If we have your email address:

- You will be notified of any significant errors that are found in this book, and you will receive a correction.
- You will be notified if there is a completely new edition of this book, or of *Make: Electronics*, or any other books by Charles Platt. These notifications will be very rare.

Your contact information will not be used for any other purpose.

Simply send a blank email (or include some comments in it, if you like) to:

make.electronics@gmail.com

Please put REGISTER in the subject line.

You Informing Us

If you only want to report an error that you have found, it's really better to use the "errata" system maintained by our publisher. The publisher uses the "errata" information to fix the error in updates of the book.

If you feel sure that you found an error, please visit:

http://bit.ly/encyclopedia_electronic_components_v3

The web page will tell you how to submit errata.

You Asking Us

Our time is obviously limited, but if you have a question, a quick answer may be available. You can send email to *make.electronics@gmail.com* for this purpose. Please put the word HELP in the subject line.

Going Public

There are dozens of forums online where you can discuss this book and mention any problems you are having, but please be aware of the power that you have as a reader, and use it fairly. A single negative review can create a bigger effect than you may realize. It can certainly outweigh half-a-dozen positive reviews.

Responses in the past have been generally positive, but in a couple of cases people have been annoyed over small issues such as being unable to find a part online. Help is available on this kind of topic, if you need it. All you have to do is send a request to *make.electronics@gmail.com*.

Safari® Books Online

Safari

Safari Books Online is an on-demand digital library that delivers expert content *in both book and video form from the world's leading authors in technology and business.*

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of plans and pricing for enterprise, government, education, and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and hundreds more. For more information about Safari Books Online, please visit us online.

You can access the errata page at *http://bit.ly/encyclopedia-electronic-components-v3*.

Make: unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. Make: celebrates your right to tweak, hack, and bend any technology to your will. The Make: audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our educational system—our entire world. This is much more than an audience, it's a worldwide movement that Make is leading. We call it the Maker Movement.

For more information about Make:, visit us online:

Make: magazine: *http://makezine.com/magazine*
Maker Faire: *http://makerfaire.com*
Makezine.com: *http://makezine.com*
Maker Shed: *http://makershed.com*

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com.

Acknowledgments

Datasheets and tutorials maintained by component manufacturers were considered the most trustworthy sources of information online. In addition, component retailers, college texts, crowd-sourced reference works, and hobbyist sites were used. The following books provided useful information:

Boylestad, Robert L. and Nashelsky, Louis: *Electronic Devices and Circuit Theory*, 9th edition. Pearson Education, 2006.

Braga, Newton C.: *CMOS Sourcebook*. Sams Technical Publishing, 2001.

Hoenig, Stuart A.: *How to Build and Use Electronic Devices Without Frustration, Panic, Mountains of Money, or an Engineering Degree*, 2nd edition. Little, Brown, 1980.

Horn, Delton T.: *Electronic Components*. Tab Books, 1992.

Horn, Delton T.: *Electronics Theory*, 4th edition. Tab Books, 1994.

Horowitz, Paul and Hill, Winfield: *The Art of Electronics*, 2nd edition. Cambridge University Press, 1989.

Ibrahim, Dogan: *Using LEDs, LCDs, and GLCDs in Microcontroller Projects*. John Wiley & Sons, 2012.

Kumar, A. Anand: *Fundamentals of Digital Circuits*, 2nd edition. PHI Learning, 2009.

Lancaster, Don: *TTL Cookbook*. Howard W. Sams & Co, 1974.

Lenk, Ron and Lenk, Carol: *Practical Lighting Design with LEDs*. John Wiley & Sons, 2011.

Lowe, Doug: *Electronics All-in-One for Dummies*. John Wiley & Sons, 2012.

Mims III, Forrest M.: *Getting Started in Electronics*. Master Publishing, 2000.

Mims III, Forrest M.: *Electronic Sensor Circuits & Projects*. Master Publishing, 2007.

Mims III, Forrest M.: *Timer, Op Amp, & Optelectronic Circuits and Projects*. Master Publishing, 2007.

Predko, Mike: *123 Robotics Experiments for the Evil Genius*. McGraw-Hill, 2004.

Scherz, Paul: *Practical Electronics for Inventors*, 2nd edition. McGraw-Hill, 2007.

Williams, Tim: *The Circuit Designer's Companion*, 2nd edition. Newnes, 2005.

In addition, three individuals provided special assistance. Our editor, Brian Jepson, was immensely helpful in the development of this book. Philipp Marek reviewed the text for errors, and Erico Narita collaborated on the Photoshop work.

GPS

1

The acronym **GPS** properly refers to the entire Global Positioning System, including satellites and ground-based control installations. However, a *GPS sensor* consists of a surface-mount chip that processes signals from GPS satellites using a small rectangular antenna, often mounted on top of a *GPS chip*.

A *GPS module* is usually a small board on which a GPS sensor is mounted with additional components. A *GPS receiver* is a device including a data display and other features, such as memory, in addition to a GPS module. In casual colloquial usage, someone who refers to "a GPS" usually means a GPS receiver.

GPS is almost always capitalized without periods.

OTHER RELATED COMPONENTS

- **magnetometer** (see Chapter 2)

What It Does

The Global Positioning System is a navigational aid jointly funded by the U.S. Department of Defense and the U.S. Department of Transportation, while being maintained by the U.S. Air Force. Signals from GPS satellites can be received and processed by modules in a wide variety of equipment ranging from aircraft to wristwatches. The signals provide location data, and may also be used as an accurate time reference.

Schematic Symbol

There is no specific schematic symbol for a GPS chip. It is likely to be shown as a box containing abbreviations that define pin functions, similar to any integrated circuit chip.

GPS Segments

The Global Positioning System consists of three segments:

The space segment

This originally required 24 communications satellites, but was revised in 2011 to require 27, to provide better global coverage. As of August 2015, there were actually 31 satellites in service, with additional "spares" that can be activated if necessary. The satellites occupy orbits 12,500 miles above the Earth, allowing each of them to circle the planet twice in 24 hours. Specifications are maintained online.

The control segment

This includes a master ground-based control station, an alternate master control station, 12 command and control antennas, and 16 monitoring sites, all maintained by the U.S. Air Force.

The user segment
This consists of receiving devices, including those that are government-owned and those that are privately owned.

How It Works

Each satellite carries multiple atomic clocks that maintain precise time, and a pseudo-random number generator in the form of a linear-feedback **shift register** (see Volume 2).

A GPS receiver can distinguish the signals from at least four satellites by comparing their received pseudo-random bit sequences, and can compute the receiver's distance to each satellite by comparing the arrival times of satellite signals.

When a satellite appears above the horizon, it approaches a receiver. After passing overhead, it moves away. This relative motion causes a *Doppler shift* in the received frequency, which the receiver circuit must take into account.

GPS satellites transmit on several frequencies simultaneously. The one for civilian use is 1575.42 MHz, called L1. Another one, 1227.6 MHz, called L2, is reserved for military use.

Variants

A GPS chip generally processes input from an antenna and provides output through solder pads. The antenna is often integrated as a ceramic square or rectangle mounted above the chip, but many chips can also process input from an external antenna. Figure 1-1 shows a GPS chip with a metal shield that is easily mistaken for an antenna. In Figure 1-2, the GPS sensor does incorporate a ceramic antenna.

Some GPS chips contain flash memory for internal data logging, although this is not a standard feature.

Suppliers such as Adafruit and Sparkfun offer GPS modules mounted on breakout boards for easier connection with other components, as shown in Figure 1-2. Some breakout boards also include provision for battery backup with a button cell.

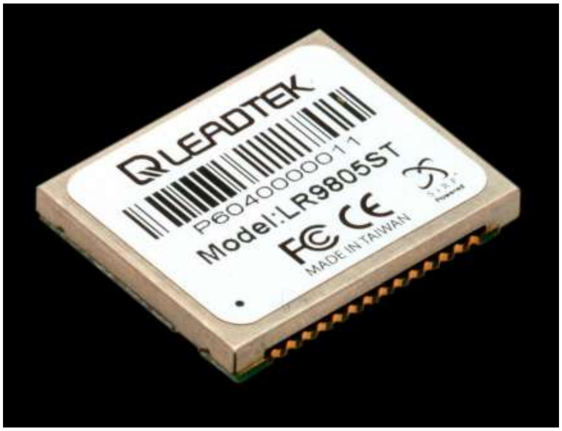

Figure 1-1 *A GPS sensor. This surface-mount chip is hidden beneath a metal shield.*

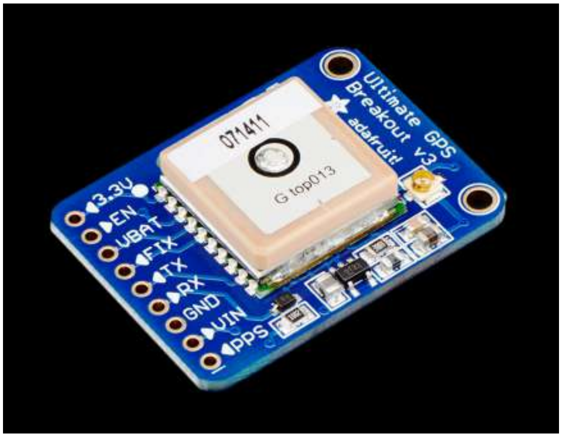

Figure 1-2 *A breakout board offered by Adafruit, incorporating a GPS sensor.*

GPS capability is almost always included in modern smartphones and tablets. It is used in handheld devices for navigation on foot, and in devices designed to be mounted in motor vehicles. Cars offer GPS capability as an option accessed via a built-in screen.

A *GPS tracker* is a device that may lack a display and simply logs its position to internal memory, from which data can be downloaded to a computer later. Many (older) handheld GPS receivers have a connector giving access to a serial or USB port, and provide data in the same NMEA format as the GPS modules described below.

After the Global Positioning System became widely used, other competing systems were introduced. These include the European Galileo, the Russian GLONASS (an acronym for Global Navigation Satellite System), and the Chinese Beidou. As of 2015, GLONASS had become fully operational. Some receivers, including those in cellular phones, compare signals from GPS and GLONASS satellites to achieve higher accuracy.

Values

Sensitivity is expressed usually in dBm, meaning the power ratio in decibels (dB) of the measured power referenced to one milliwatt (mW).

Time to First Fix (TTFF) is the time required to obtain an initial fix from a satellite.

Number of channels is the number of satellites that a GPS receiver can track simultaneously. Early GPS receivers could sense only four channels. Modern units may be able to deal with 22.

Power consumption may be measured in milliwatts. For example, the FGPMMOPA6H GPS standalone module by G.top claims power consumption of 82mW during acquisition of satellite signals and 66mW while tracking subsequently. At a typical voltage of 4VDC, the chipset consumes about 20mA and approximately 17mA, respectively.

Form factor. This is the size of a chip, often determined by the dimensions of the ceramic antenna on top of it. Dimensions may be 15mm x 15mm or larger.

Update rate. The number of position measurements per second. While 1 update per second is often sufficient, some chips generate updates more rapidly. The frequency of updates is expressed in hertz.

Output type. This is often TTL-level serial providing NMEA data. The baud rate can vary and is often selectable.

Supply voltage. Often below 5VDC.

Current consumption. Higher during satellite acquisition.

How to Use It

A GPS module requires only a DC power supply, and will start outputting data as soon as it has identified satellites that are currently within range.

Data provided by a GPS module uses a rather slow and primitive plain-ASCII protocol known as NMEA, developed by the National Marine Electronics Association. Each block of data is known as a *sentence*, and can be parsed independently of previous and subsequent sentences. The default transmission rate is 4,800 bits per second, using 8 bits to identify an ASCII character, no parity, and 1 stop bit. However, some GPS modules use a serial rate of 9,600 bps or faster.

A sentence begins with a two-letter abbreviation defining the type of device employing the sentence. For a GPS device, the abbreviation is GP. The sentence continues with another abbreviation of three characters or more, describing the type of data being transmitted, so that numeric values in the sentence can be interpreted correctly.

The remainder of a sentence consists of letters and numerals in plain ASCII, with values separated by commas. A sentence cannot contain more than 80 characters. A sentence will specify the latitude, longitude, and altitude of the GPS, and a value defining the time when the readings were derived from satellite signals. Some sentence data structures are proprietary, established by the device manufacturer, and will begin with the letter P.

A GPS device may send a variety of different sentence types in succession, to overcome the 80-character sentence length limit. Each sentence will be preceded with an identifier. Sentence types and data contents will be defined in a manufacturer's datasheet.

Output from GPS chips may be compatible with a microcontroller. Output from a GPS breakout board will almost certainly be compatible, and the board is likely to have its own voltage regulator. The microcontroller can receive serial data from the GPS chip, can stop the GPS chip via its enable pin, and can start and stop internal logging of data in flash memory on the GPS chip if this feature is available.

Code libraries are available online for a microcontroller, to enable it to receive and interpret serial data from a GPS device.

Pulse per Second Output

As the GPS positioning depends on calculating distances from the time it takes for a radio signal to travel, accurate timekeeping is needed. When a GPS receiver obtains a fix for its position, it will get a value for the current time as well. This makes GPS receivers usable to supply time and frequency standards. Most receiver modules report the time together with the position. Many also provide a special PPS output, which produces one pulse per second, precisely synchronized with the satellites in view.

The accurate time provided by the GPS receiver can be used to *discipline* a crystal oscillator. This means measuring the crystal oscillator frequency relative to a GPS-provided reference, and continuously adjusting the crystal frequency to keep it stable.

What Can Go Wrong

Problems generally affect chips and modules rather than devices.

Electrostatic Discharge

The patch antenna on a GPS chip connects with the chip via an RF input. If the antenna is subjected to an electrostatic discharge, the chip can be permanently damaged. Likewise, damage will result if a discharge is applied to the RF input, for example through a soldering iron. The chip must be grounded before any work is done involving the RF input.

Failure to Ground Properly

The ground solder pad on a chip, or the ground pin on a breakout board, must make contact before voltage is applied to other pads or pins.

Cold Joints

The patch antenna on a GPS chip can function as a heat sink, increasing the risk of cold solder joints when the chip is being mounted on a board.

Restricted Availability

U.S. regulations limit the export of some GPS devices capable of rapid positional updates that could be usable in military aircraft or missiles. Other restrictions also apply. Suppliers may be inconsistent in their refusal to allow purchase of restricted items outside of the United States.

Inability to Detect Satellites

Any GPS device may fail to detect satellite signals if its view of the sky is obstructed. Reception is usually possible through window glass, but may not be available through walls, a roof, heavy tree cover, and other natural obstructions.

Exceeding Maximum Velocity or Altitude

Security regulations limit the capability of GPS devices to function above 60,000 feet or at greater than 1,200 miles per hour. Outside of these limits, the GPS device will not provide data. This may affect applications in amateur rocketry or high-altitude balloons.

2

magnetometer

This entry deals only with magnetic sensors that respond to the Earth's magnetic field. Small magnetic sensors such as the ubiquitous *Hall sensor* may be used for many other purposes, such as determining the position or rotational speed of mechanical components. These applications involve **object presence** sensing; see Chapter 3. In that entry, Hall sensors are discussed at "Hall-Effect Sensor".

In the past, a **magnetometer** was a bulky measurement device incorporating knobs or other controls and some form of display. While that use of the word is still common, this entry deals only with chip-based magnetic sensors.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **gyroscope** (see Chapter 9)
- **GPS** (see Chapter 1)

What It Does

A traditional compass consists of a thin magnetized strip of metal balanced on a pivot. It aligns itself with the Earth's magnetic field.

A *scalar magnetometer* measures the total strength of a magnetic field. A *vector magnetometer* can measure the strength in a specified direction. In particular, it can provide a numeric output describing the angle between the orientation of the measuring device and the Earth's magnetic poles.

Chip-based magnetometers are usually vector-type, containing three sensors mounted orthogonally—that is, each of them at 90 degrees to the other two. Suitable software can interpret the analog readings from the sensors to calculate magnetic north or south regardless of the angle at which the instrument is being held, relative to the ground.

Schematic Symbol

There is no specific schematic symbol for a magnetometer.

IMU

A **gyroscope** measures the rate of rotation of the enclosure in which it is mounted. This is properly known as the *angular velocity*. A gyroscope will also respond to changes in the rate of rotation. It does not measure linear motion or a static angle of orientation.

An **accelerometer** measures variations in linear motion and will also measure its own static orientation relative to the force of gravity. If an accelerometer rotates around its own axis, it will not measure angular velocity.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an *IMU* (inertial measurement unit), which can provide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Applications

Magnetometers are found in handheld equipment such as digital compasses, cameras, and cellular phones. They are usually surface-mount chips that are manufactured in large quantities and may be used in conjunction with microcontrollers. For the hobby-electronics community, or for experimental product development, a magnetometer may be mounted on a breakout board for ease of use. A board using a Honeywell HMC5883L is shown in Figure 2-1.

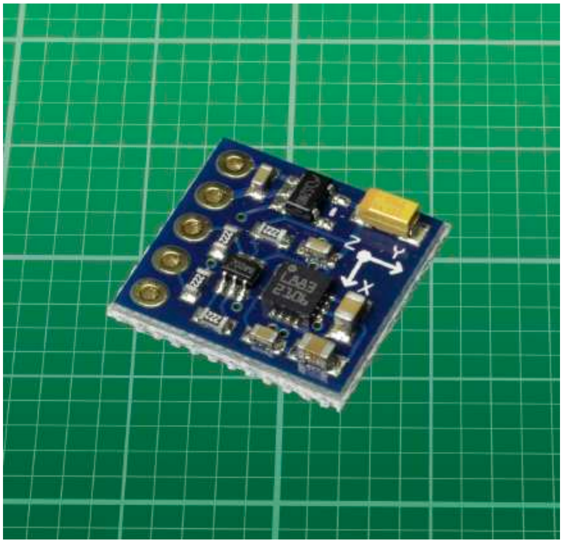

Figure 2-1 *The Honeywell HMC5883L 3-axis magnetometer mounted on a breakout board. The background grid is in millimeters.*

How It Works

An explanation of magnetometers requires an understanding of the fundamentals of magnetism.

Magnetic Fields

A magnetic field is often represented by *field lines* that show the strength and vector of the field. Field lines associated with a simple permanent magnet are shown in Figure 2-2, where the strength, or *flux density*, at any point, is indicated by the spacing the lines, while the angle tangential to a line indicates its vector. (For an extended discussion of magnetism, see **electromagnet** in Volume 1.)

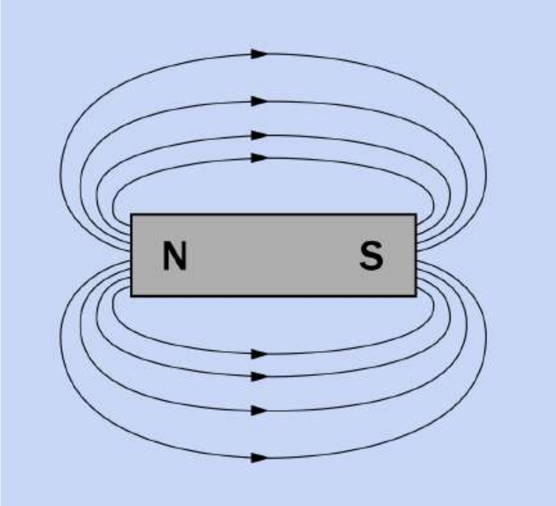

Figure 2-2 *Field lines representing magnetic flux created by a bar magnet. Space between the field lines is inversely proportional to flux density. In reality this is a three-dimensional effect, and a more accurate representation would show the magnet and field lines revolved around the axis of the magnet.*

The flux density of a magnetic field is usually represented by the letter B, and is measured in *newton-meters per ampere*, more commonly referred to as *teslas* (T). An older unit of measurement was gauss (G), 1 tesla being equivalent to 10,000 gauss. Some datasheets still refer to gauss.

The Earth's magnetic field is believed to result from convection currents in the outer liquid of the Earth's core. The strength of this field varies from 25 to 65 microteslas (0.25 to 0.65 gauss) depending on the location where it is measured. To a very rough approximation, the Earth behaves as if a giant bar magnet connects the magnetic north pole and the magnetic south pole. See Figure 2-3.

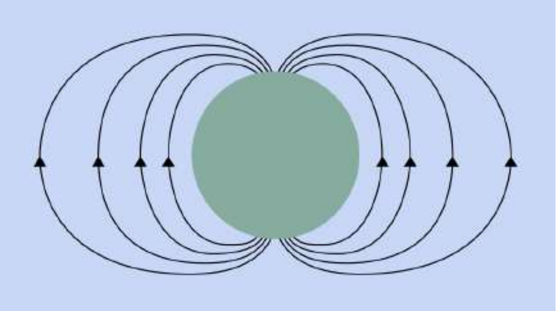

Figure 2-3 *The magnetic field of the Earth approximately resembles the field around a bar magnet.*

Note that in northern and southern latitudes, the field lines angle steeply down toward the surface, while the lines are approximately parallel with the surface near the equator. Consequently a magnetometer held tangentially to the surface of the Earth will tend to measure a stronger horizontal field near the equator than near the poles.

The varying angle of the field lines tangential to the Earth is known as their *inclination*. Variations in field strength can be used to determine an approximate location, although the **GPS** (Global Positioning System, using satellites) enables this much more accurately.

Confusion results from the fact that the nothern magnetic pole of the Earth actually behaves as a south pole, while the southern magnetic pole behaves as a north pole. That is, when a permanent magnet is free to rotate, its north pole will orient itself toward the so-called northern magnetic pole of the Earth, even though opposite poles attract. The northern pole of the Earth should therefore be thought of as the pole that attracts the north end of a compass.

Earth's Axes

Planet Earth revolves around an imaginary line referred to as the *axis of rotation*. This line is close to, but not quite coincident with, the *magnetic axis* connecting the magnetic north pole and magnetic south pole, as shown in Figure 2-4.

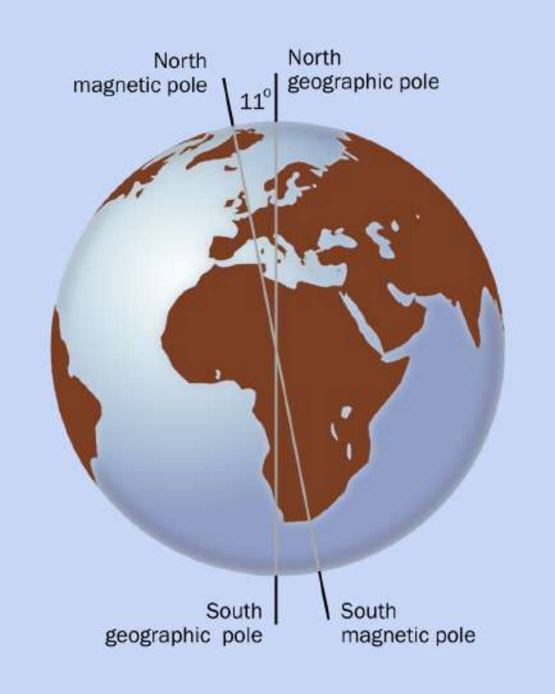

Figure 2-4 *The angle between the magnetic axis and the rotational axis of the Earth is approximately 11 degrees.*

Magnetic declination is the angle between the magnetic north pole and the geographical north pole, as perceived by an observer. This angle will vary depending on the observer's location on the surface.

Because of declination, the direction of magnetic force at points on the surface of the planet will vary with latitude and longitude, as shown in Figure 2-5. In this figure, the *magnetic meridians* are shown in red, superimposed on the *geographical meridians* shown in green. Magnetic meridians show the direction of magnetic force, while geographic meridians are drawn between the ends of the axis of rotation of the Earth. While there is an approximate correlation between the two, in some areas, especially near the north and south poles, the discrepancy can be more than 40 degrees.

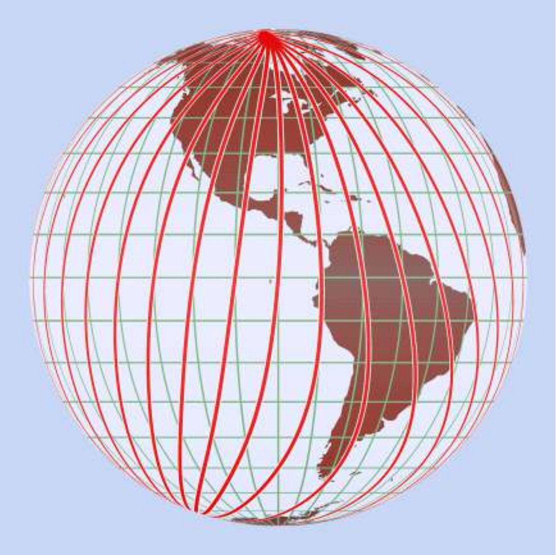

Figure 2-5 *The red lines indicate the direction in which a compass would be likely to point to magnetic north. The green lines connect the geographical poles of the planet. (From Wikimedia Commons.)*

Standard declination tables for locations on the Earth are available, and these values must be added to, or subtracted from, the reading from a compass or magnetometer to determine the direction of geographical north. Navigational systems customarily express the heading of a vehicle or vessel relative to geographical north, as shown in Figure 2-6.

Coil Magnetometer

Current flowing through a wire creates a magnetic field with flux density that is directly proportional to the current in amperes. Conversely, a changing magnetic field will induce current in a wire. This principle is used in a *coil magnetometer*, which can detect buried objects when the coil moves above them. A *rotating coil magnetometer* can determine magnetic field strength while remaining in a stationary position. However, the coil in a coil magnetometer must be relatively large.

Hall Effect and Magnetoresistance

In modern handheld devices, a magnetometer will generally use either the *Hall effect* (see "Hall-Effect Sensor") or *magnetoresistance*, described here.

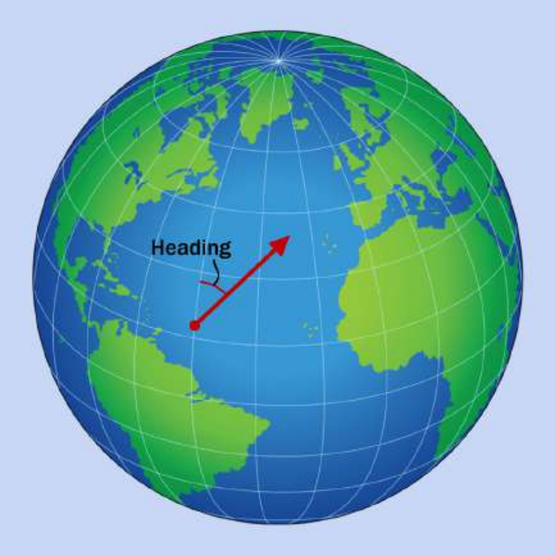

Figure 2-6 *Heading is an angle usually calculated relative to geographical north, sometimes referred to as "true north."*

Magnetoresistance is a phenomenon where the resistance of a material changes fractionally when it is exposed to a magnetic field. This is capable of yielding greater accuracy than Hall-effect sensors, but has tended to be more expensive.

Orthogonally oriented sensors contained in a surface-mount chip are aligned with axes identified with letters X, Y, and Z. These sensors are analog devices whose values are converted to digital values by an onboard *analog-to-digital converter* (ADC). The values are stored in *registers* that are available to other devices, often via the *I2C* protocol, which is widely used by microcontrollers.

Typically there will be two eight-bit registers for each axis, one defining the high byte and the other defining the low byte for the digital value. In reality the ADC is likely to use 10 to 13 bits, with the remaining 6 to 3 bits being unused.

Variants

The Freescale Semiconductor FXMS3110 is a typical low-cost chip containing a 3-axis magnetometer sensor. Many chips now additionally include a 3-axis accelerometer sensor. An example is the LSM303 by STMicroelectronics, which is sold on a breakout board by Adafruit. See Figure 2-7.

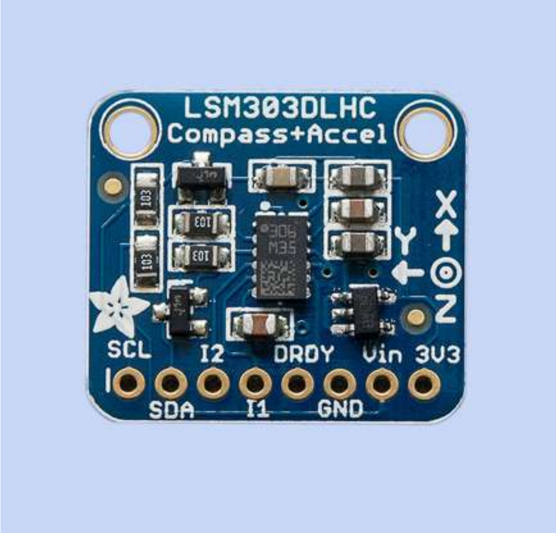

Figure 2-7 *The LSM303 is a chip manufactured by STMicroelectronics. It is shown here on a breakout board from Adafruit.*

See Chapter 10 for an explanation of the function of accelerometers.

The InvenSense MPU-9250 is a highly sophisticated IMU, including a 3-axis **gyroscope** in addition to a 3-axis magnetometer and 3-axis accelerometer.

A processing unit in the MPU-9250 reconciles the nine variables, and the digital output can be accessed via I2C or SPI protocol at speeds up to 1MHz. All functions of the chip are contained within a package measuring less than 3mm square.

See Chapter 9 for additional details about the functioning of a gyroscope.

How to Use It

A basic 3-axis magnetometer sensor such as the HMC5883L can be tested with a microcontroller that can receive data from its registers via I2C protocol. This is relatively easy with an Arduino, which was designed to be I2C-capable.

Several breakout boards are available with the HMC5883L mounted on them. Many of these boards contain a voltage regulator, allowing a 5VDC power supply to be used even though the chip is designed for a typical supply of 2.5VDC.

In addition to the power supply, the breakout board only requires two connections for I2C communication, to its SCL pin (serial clock input) and SDA pin (serial data input/output). If basic I2C software is installed on the microcontroller, it will read digital values from the magnetometer registers. If additional software is used, it will convert the values to magnetic flux densities in microteslas for the X, Y, and Z axes. Code libraries to achieve these goals are widely available online.

A more sophisticated chip such as the InvenSense MPU-9250 can be used similarly, but yields additional data for conversion. Here again, code libraries can be found online. The slightly older MPU-9150 is sold on a breakout board with downloadable code from Sparkfun.

What Can Go Wrong

Bias

Magnetometers are sensitive to their environment, which can induce *magnetometer bias* of two types.

Hard-iron bias is primarily caused by magnetized material inside the device containing the magnetometer. Because this effect is usually unvarying, compensating for it is relatively easy. *Soft-iron bias* is caused by interaction between variations in the Earth's magnetic field and materials inside a magnetometer that can be magnetized.

A common example of soft-iron bias would be power lines, generating a magnetic field that can affect model aircraft and drones using magnetometers for navigation.

Mounting Errors

Placement of a chip-based magnetometer on a circuit board is critical. The field effects of transformers or relays must be taken into account, and even the low voltage and low current in a circuit trace can create a magnetic field sufficient to disturb a chip. No traces on any layer of the board should pass across the footprint of the chip. A magnetometer should not be mounted within a ferromagnetic case.

spatial > presence > object presence

object presence sensor

3

An **object presence** sensor may also be described as an *object detector* or *detection sensor*.

The term **proximity** sensor may be applied to this component. However, in this Encyclopedia a proximity sensor has the capability of estimating the distance to a target. See Chapter 5. An object presence sensor merely detects whether an object is within a preset range, and does not provide additional information.

Optical and *magnetic* object-presence sensors are described and compared in this entry. *Ultrasonic sensors* are described in the entry dealing with proximity sensors, because they tend to be used for distance measurement rather than just for detection. Other methods of object presence sensing, including *capacitive*, *doppler*, *inductive*, *radar*, and *sonar*, are not included in this Encyclopedia.

A sensor that detects an object by receiving light reflected from it is categorized as a *reflective* sensor, and is included in this entry. (If a module includes a light source as well as a light sensor, it is properly categorized as *retroreflective*, although that term is not always used.)

A sensor that detects an object when it interrupts a beam of light is a *transmissive* sensor, and is included in this entry. It may also be described as a *through-beam* sensor or as an *optical switch*.

A sensor that responds to the motion of an object that emits infrared radiation is a **passive infrared** motion sensor, also described by its acronym, PIR, and often referred to simply as a *motion sensor*. It has its own separate entry. See Chapter 4.

Phototransistors and **photodiodes** may be used as sensing elements in presence sensors. These components have their own entries as light sensors. See Chapter 22 and Chapter 21.

OTHER RELATED COMPONENTS

- **proximity** sensor (see Chapter 5)
- **passive infrared** sensor (see Chapter 4)

What It Does

An **object presence** sensor verifies the presence or absence of an object within a predetermined range, without necessarily measuring how far away it is or how fast it is moving. The object may be described as a *target*.

Object presence sensing is often used to verify the correct function of an automated system—for example, the placement of objects on a conveyor belt. It can also be used to count objects as they pass a sensor.

Some types of security systems use presence sensors to sound an alarm if an intruder interrupts a beam of light. They can verify that a door or window is closed. They may also function as a limit switch to control the operation of a motor.

Schematic Symbol

In schematics, an optical presence sensor may be indicated with the symbol for an LED, plus the symbol for a **phototransistor**, with one or two arrows connecting them, as shown at top-left in Figure 3-1. Wavy-line arrows may indicate an infrared connection.

A **photodiode** may be substituted for a phototransistor, as shown at top-right in the figure.

A magnetic sensor may be shown with the symbol for a *Hall-effect sensor*, as in the lower half of Figure 3-1.

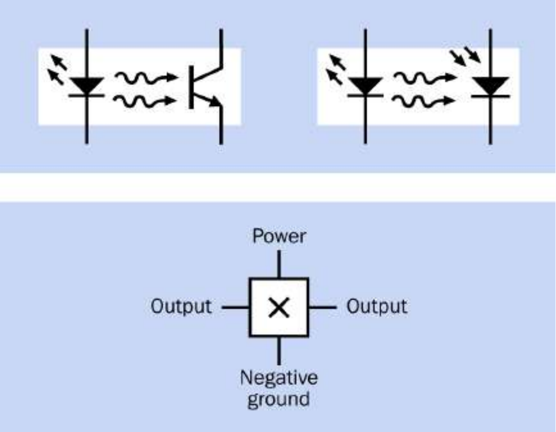
Power
Output
Output
Negative ground

Figure 3-1 *Top: two possible schematic symbols for an optical presence sensor, using a phototransistor (left) or photodiode (right). Other variants are possible. Bottom: the schematic symbol for a Hall-effect sensor, commonly used in a magnetic sensor.*

Variants

To assist the reader in comparing different options for detecting the presence of an object, this entry includes two primary variants: *optical* and *magnetic*.

The optical sensors are divided into *transmissive* and *reflective* (including *retroreflective*). The magnetic sensors are *reed switches* and *Hall-effect sensors*. A chart showing the categories and subcategories is shown in Figure 3-2.

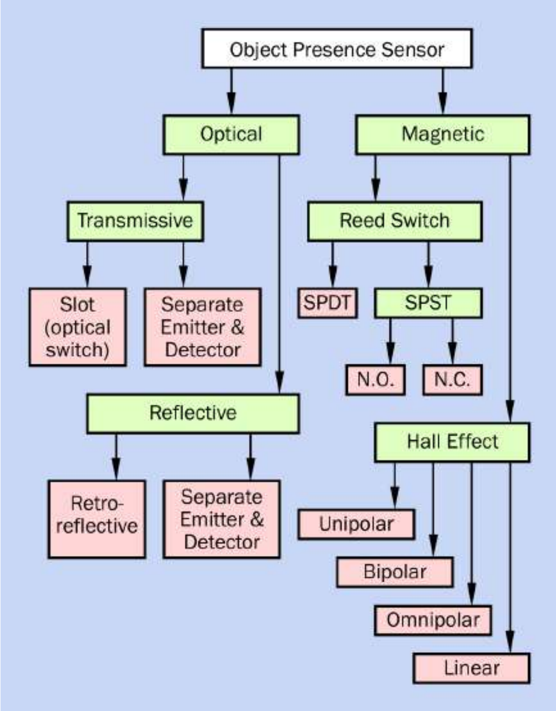
Object Presence Sensor
Optical
Magnetic
Transmissive
Reed Switch
Slot (optical switch)
Separate Emitter & Detector
SPDT
SPST
N.O.
N.C.
Reflective
Hall Effect
Retro-reflective
Separate Emitter & Detector
Unipolar
Bipolar
Omnipolar
Linear

Figure 3-2 *Categories of object presence sensors discussed in this entry. Other types of object presence sensors exist, but are less common, and are not included here.*

Optical Detection

A *transmissive optical sensor*, also known as a *through-beam sensor*, is really a pair of components, one emitting light and the other receiving it. The sensor is triggered if an object interrupts or reflects the light beam, as shown in Figure 3-3.

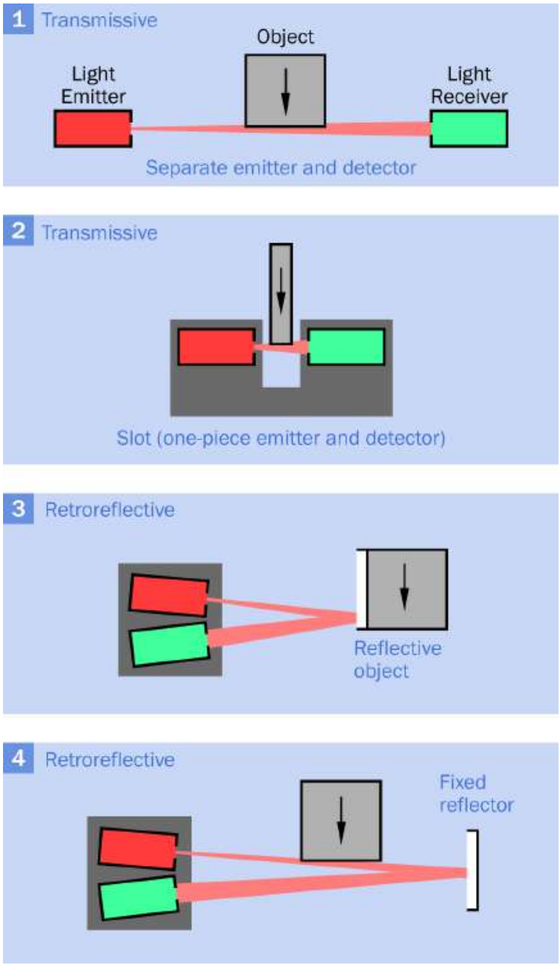

Figure 3-3 *Various configurations of optically activated object presence sensors. See text for details.*

If the light emitter and the light detector face each other across a small gap, they may be contained in a single module (usually with a slot in it) as shown in section 2 of Figure 3-3. This is often referred to as an *optical switch* (not to be confused with the solid-state switching devices used in telecommunications). It is sometimes described as a *photointerruptor* or *photointerrupter*.

A *reflective optical sensor* also consists of a light emitter and a light detector, but they are placed adjacent to each other, facing in the same general direction. When they are mounted in one module, as is often the case, the arrangement is properly known as a *retroreflective sensor*. Either way, the combination may be triggered in one of two ways:

- An object passing in front of the light beam reflects it back to the detector. The object must be naturally reflective (for example, glass containers or white boxes on a conveyor), or must have a reflective patch applied to it, or the light source must be bright enough to reflect from an object that is not highly reflective. This configuration is shown in section 3 of Figure 3-3.
- A stationary reflector can be mounted opposite the light emitter, in which case a detector beside the light emitter is triggered when an object interrupts the reflected light beam. This configuration is shown in section 4 of Figure 3-3.

Transmissive Optical Sensors

A light source and light detector may be sold as separate components in a matched pair. An example is the Vishay TCZT8020, shown in Figure 3-4. These components are small, each measuring no more than 5mm x 3mm. They are designed to be placed just a few millimeters apart. The light source is an infrared LED, while the detector is a **phototransistor** (see Chapter 22 for information about phototransistors).

The source and detector are both designed to use 5VDC. Output from the phototransistor is an open collector. Current through the collector must not exceed 50mA, and must be limited by a pullup resistor of 100 ohms or more. Current through the source must not exceed 60mA, and must be limited by an appropriate series resistor.

Details on using an open-collector output will be found in the Appendix. See Figure A-4.

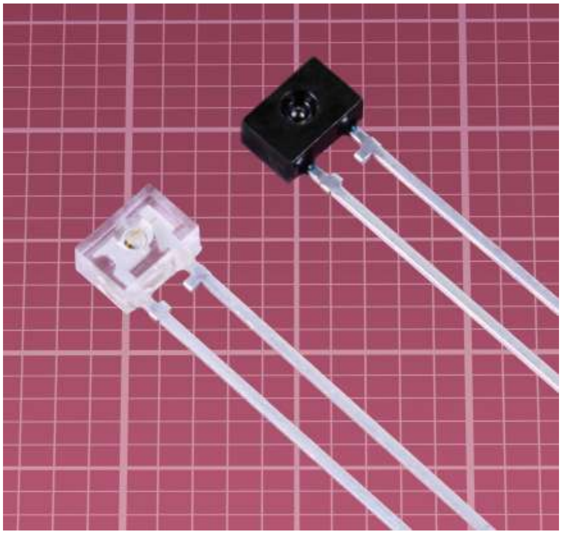

Figure 3-4 *A matched light source and light detector, for use as a transmissive optical sensor. The background grid is in millimeters.*

The Omron EE-SX range includes a variety of configurations of light source and detector separated by a 5mm slot in one module. The source is an infrared LED, and the detector is a phototransistor.

The Omron components tolerate a wide range of supply voltages, from 5VDC to 24VDC, with no series resistor necessary for the LED. Open-collector output from the phototransistor can tolerate 50mA to 100mA, depending on the particular version of the sensor. A red indicator LED is illuminated when an object blocks the slot in the sensor. Some versions have a high output when the slot is open, while others have a high output when the slot is closed. Because of their various features, these sensors are relatively expensive.

A much cheaper optical switch is the Everlight ITR9606 (described by the manufacturer as an "opto interrupter"). It is pictured in Figure 3-5. This is intended as a 5V device and has an open-collector output. It requires a series resistor for the LED and a pullup resistor for the open-collector output. Many similar detectors are available.

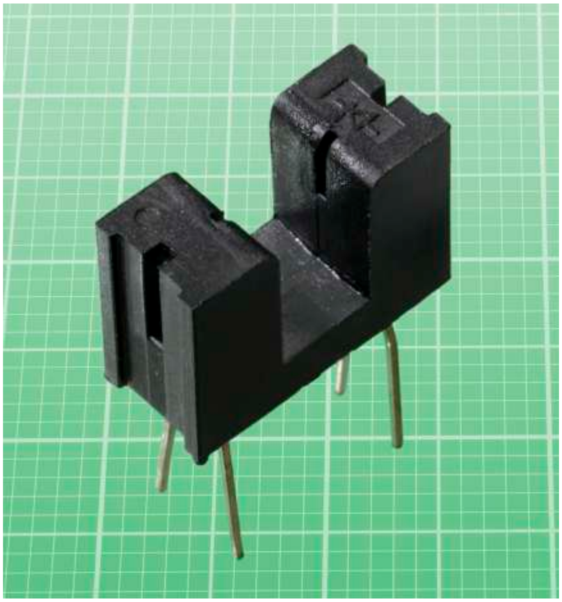

Figure 3-5 *A low-cost transmissive sensor, more commonly known as an optical switch. The background grid is in millimeters. The slot in the sensor is about 5mm wide.*

For longer-distance detection, an infrared receiver can be mounted separately from an infrared LED. The TSSP77038 from Vishay detects infrared light from as far away as 50cm, and pulls its open collector output low in response. The light must be modulated at 38kHz.

Polulo Robotics and Electronics sells a very affordable breakout board containing a TSSP77038 receiver paired with an infrared LED modulated by a 555 timer. It is shown in Figure 3-6. Because this board contains a light source as well as a light detector, it is really a retroreflective sensor.

Where distances exceeding 1 meter are involved, a laser coupled with a phototransistor that is shielded from ambient light may be necessary.

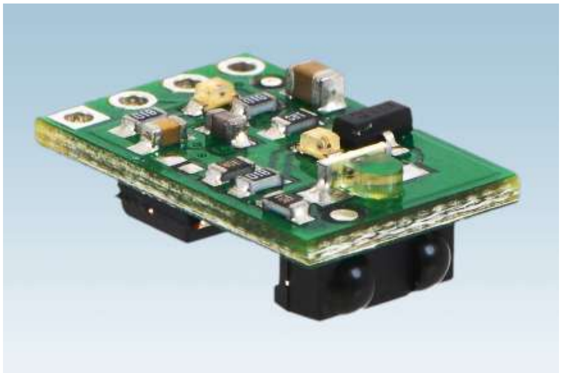

Figure 3-6 *The Vishay TSSP77038 installed with an appropriate light source on a breakout board from Polulo Robotics and Electronics.*

Retroreflective Optical Sensors

As is the case with a transmissive object detector, the retroreflective type may be listed by vendors as an *optical switch*. Other terms used in datasheets include *reflective interrupter*, *reflective object sensor*, *reflective optical sensor*, *reflective photointerrupter*, *opt-pass sensor*, and *photomicrosensor (reflective)*. The remarkable lack of standardization in terminology creates a problem when searching online for these devices. Why so many different names evolved for the exact same component is unclear.

Many retroreflective object-detection sensors are available in packages ranging from 5mm x 5mm to 10mm x 10mm in size. Almost all of these modules are analog devices using an infrared LED as the light source and a phototransistor as the sensor, with an open-collector output. (For more information about phototransistors, see Chapter 22.)

With a suitable pullup resistor, the output voltage will be proportional with the inverse of the distance. If V is the voltage, d is the distance, and k is a conversion factor:

```
V = k * ( 1 / d)
```

While many of the smaller modules are surface-mount, some have leads, as shown in Figure 3-7. A major limitation of these small components is that they have a very limited range, typically less than 5mm. They can only be used in applications such as process control where the position of a target is controlled and predictable.

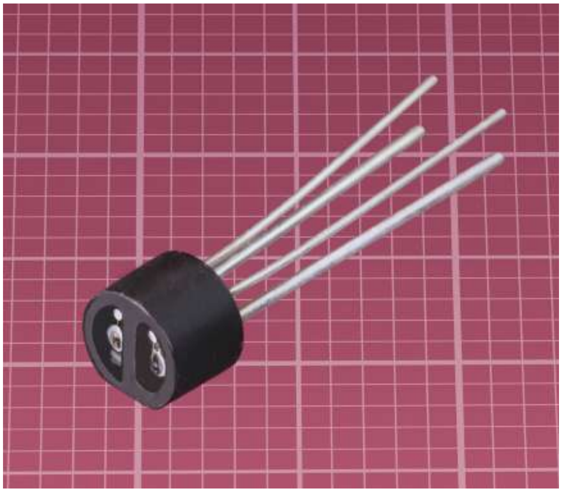

Figure 3-7 *The Rodan RT-530 is a small object-presence sensor with a limited range typical of this type of retroreflective component. The background grid is in millimeters.*

Another example of a retroreflective sensor in a small package is the Optek OPB606A shown in Figure 3-8. The background grid is in millimeters.

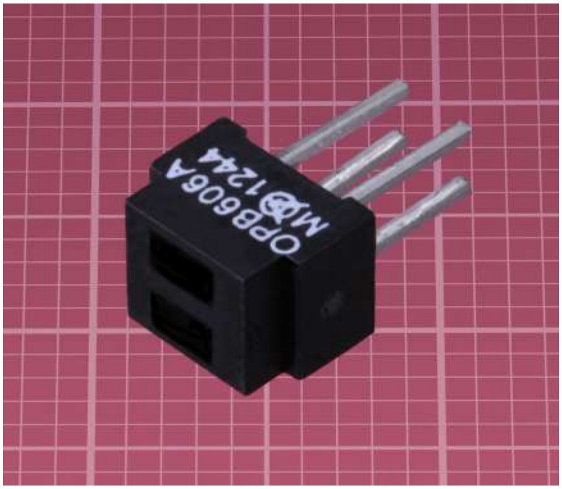

Figure 3-8 *The Optek OPB606A. The background scale is in millimeters.*

A retroreflective module, with a lensed LED and lensed phototransistor to focus the outgoing and reflected beams, is the Vishay TCRT5000 shown in Figure 3-9.

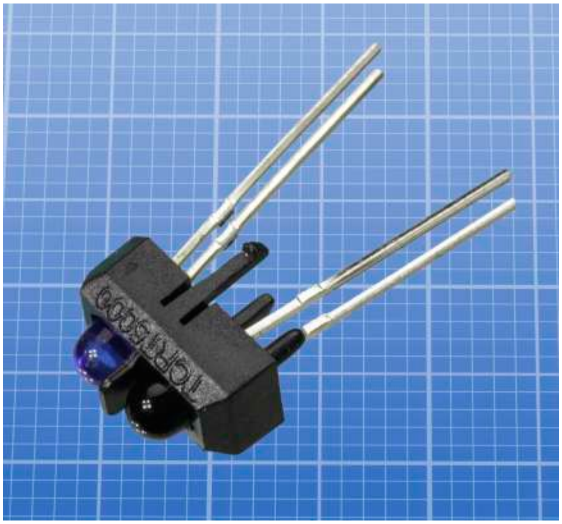

Figure 3-9 *The Vishay TCRT5000 retroreflective sensor. The background grid is in millimeters.*

Retroreflective modules that have a greater range tend to be physically larger, less common, and more expensive. Sharp makes a popular series. Some part numbers and distance sensing limits are: GP2Y0D805Z0F (5mm to 5cm), GP2Y0D810Z0F (2cm to 10cm), and GP2Y0D815Z0F (5mm to 15cm). Figure 3-10 shows the GP2Y0D810Z0F mounted on a small breakout board from Polulo Robotics and Electronics. The board is useful, as the pins on the sensor are spaced only 1.5mm apart. The size of the board is approximately 8mm x 20mm.

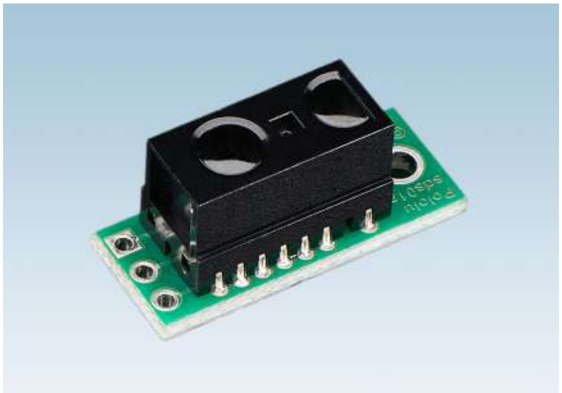

Figure 3-10 *The Sharp GP2Y0D810Z0F detection sensor mounted on a board from Polulo Robotics and Electronics. From a photograph by Adafruit Industries.*

Each of the detection sensors in this series is described in Sharp datasheets as a “Distance Measuring Sensor Unit,” but in fact they do not measure distance. A single output is normally logic-high and drops to logic-low when a target is within range. Sharp refers to this as a “digital” output, but in fact it is binary, and should not be confused with the digital buffer on a more sophisticated proximity sensor that contains an analog-to-digital converter and provides a numeric output.

It is important to distinguish the Sharp object-presence sensors listed above from the range of Sharp **proximity** sensors described in the entry in Chapter 5. The proximity sensors are physically larger, and most have an analog output that varies with the distance of a target.

Magnetic Sensors

Prepackaged magnetic sensing units are sold in many configurations for industrial and military applications. Although they may be referred to as “magnetic sensors,” they are outside the scope of this Encyclopedia. Here we discuss board-mounted components. Almost always, they use a *reed switch* or a *Hall-effect* sensor as their sensing elements.

Reed Switch

A reed switch is a magnetically activated mechanical switch. It consists of two metallic contacts in a small enclosure that is usually a glass capsule. The contacts are magnetic, and move in response to a magnetic field. A permanent magnet is used to activate the switch. Two reed switches are shown in Figure 3-11.

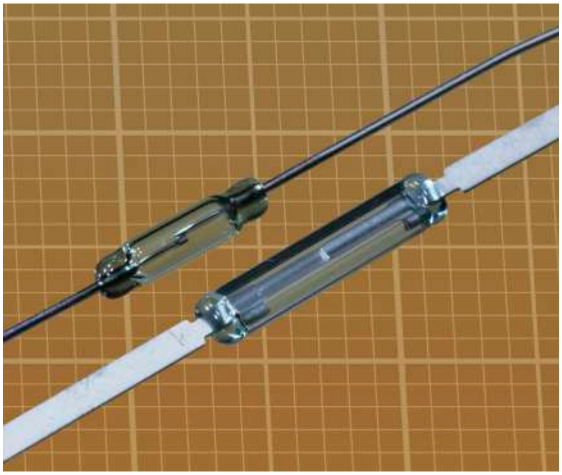

Figure 3-11 *Two SPST reed switches. Although the contacts may seem to be touching, in fact they are separated by a tiny gap, and these switches are the normally open type. The background grid is in millimeters.*

A reed switch exhibits a small amount of *hysteresis*, because the magnetic field strength required to overcome the mechanical resistance of the springy contacts is greater than the field strength required to keep them closed.

Very small electromagnetic relays that only switch a very low current may use a coil-activated reed switch. For purposes of this Encyclopedia, such a component is considered to be a relay, not a sensor. For more information about **relays**, see the entry in Volume 1.

The most common everyday application for a reed switch is in an alarm system that is triggered by an intruder entering a building. A magnet in a sealed plastic enclosure is attached to a door or window, while a reed switch, in another sealed plastic enclosure, is attached to the frame, very close to the magnet. Typical components of this type are shown in Figure 3-12. A diagram illustrating the mode of operation appears in Figure 3-13.

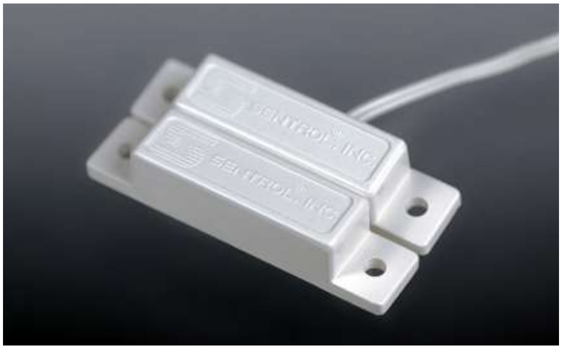

Figure 3-12 *A typical alarm sensor to detect the opening of a door or window. The nearer module contains a magnet, while the module behind it contains a reed switch.*

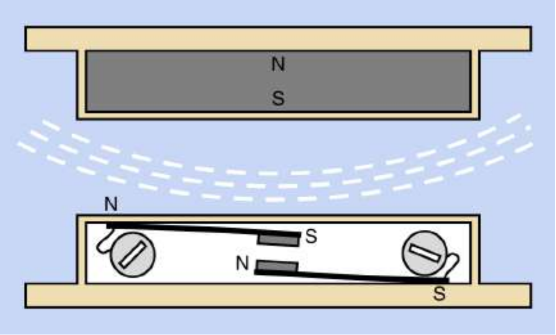

Figure 3-13 *The dashed white lines suggest the magnetic field that closes the contacts of a reed switch.*

So long as the alarmed door or window remains closed, the magnet activates the reed switch. If the door or window is opened, the magnet moves away from the reed switch, and its contacts relax. Usually in this application the reed switch is the normally open type, and is held in its closed state by the magnet. This allows multiple switches to be wired in series, completing a circuit. If a single switch opens, the circuit is broken, and an alarm is triggered.

Reed Switch Variants

Most reed switches are SPST, either normally open or normally closed, although normally open variants are more common. Some switches are SPDT, although this variant is relatively rare. An example of a SPDT reed switch is shown in Figure 3-14.

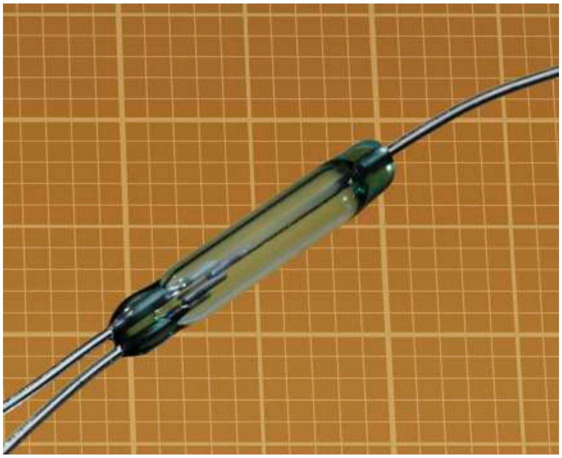

Figure 3-14 *A SPDT reed switch. The background grid is in millimeters.*

The physical size of a reed switch tends to vary roughly in proportion with its current-switching capability.

Reed switches are most commonly supplied with axial leads. A minority are available for surface-mount applications.

Some reed switches are packaged in a plastic pod to provide physical protection for the glass capsule.

Reed Switch Values

The datasheet for a reed switch is likely to contain the following values:

Pull-in: The minimum magnetic field strength required to activate the switch, often measured in ampere-turns.

Drop-out: The maximum magnetic field strength that allows the contacts of the switch to relax, often measured in ampere-turns. Pull-in will be a higher value than drop-out.

Maximum switching current: While a few industrial reed switches can switch as much as 100A, they are uncommon and expensive. A common value for a reed switch about 15mm long is 500mA.

Maximum carry current: If specified, it will be higher than the switching current.

Maximum switched power: Because a reed switch can be used with alternating current, its switching capability may be expressed in watts, or as VA (volts multiplied by amps). 10VA is a common value.

Maximum voltage: While reed switches are most often used with low voltages, some are rated to switch up to 200V.

How to Use a Reed Switch

While an optical object presence sensor may be supplied in a single package containing a light emitter and a light detector, a reed switch always requires an activating magnet that is mounted separately. For reliable operation, the maximum distance between the switch and the magnet is usually restricted to a few millimeters.

The orientation of a magnet that activates a reed switch is not crucial, but will affect the sensitivity of the switch. A manufacturer's datasheet should be consulted for information about optimal magnetic polarity.

As in other mechanical switches, the contacts in reed switches vibrate very briefly when opening or closing. This is known as *contact bounce*, and can be misintepreted as a series of separate signals by digital logic or a microcontroller. *Debouncing* with hardware designed for that purpose, or software (in a microcontroller), may be necessary. See the entry on **switches** in Volume 1.

Hall-Effect Sensor

A Hall-effect sensor responds to a magnetic field by generating a small voltage, usually amplified by a transistor included in the package with the sensor.

When a Hall sensor is in its "off" state (i.e., not triggered by a magnetic field), the sensor creates high resistance between the collector of an internal NPN transistor and negative ground. Consequently, if voltage through a pullup resistor is applied to the collector, the voltage on the collector will be high.

When the sensor is in its "on" state, the resistance drops, the voltage supplied to the collector from the pullup resistor is shunted to ground, and the output voltage appears to go low. As a general rule:

- Activated Hall sensors appear to have a low output.
- Inactive Hall sensors appear to have a high output.

Information about using open-collector outputs is contained in the Appendix. See Figure A-4.

The clean output, reliability, small size, and cheapness of Hall sensors have encouraged their use in devices as different such as hard drives, cameras, keyboards, and automobiles. They are useful in almost any situation where a sensor must detect a mechanical operation at close range. Three through-hole Hall sensors are shown in Figure 3-15. Surface-mount versions are much smaller.

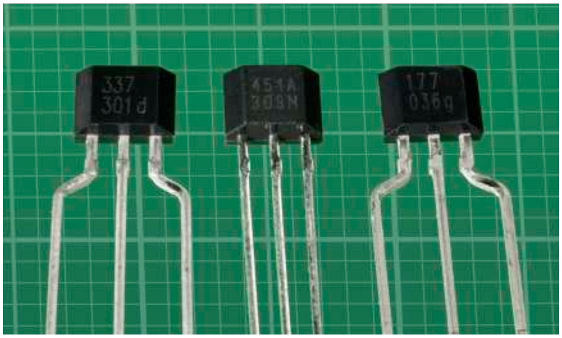

Figure 3-15 *Three through-hole Hall-effect sensors. The background grid is in millimeters.*

How a Hall-Effect Sensor Works

When current is flowing through the length of a conductor, and a magnetic field is applied across the width of the conductor, the field generates a force causing electrons and electron holes to accumulate asymmetrically on opposite sides. This is known as the *Hall effect*.

The voltage difference between the electron-rich and electron-depleted zones is known as the *Hall voltage*. It is proportional to the magnetic field, and inversely proportional to the density of free electrons in the material. For this reason, the Hall effect is easiest to observe in semiconductors, where the density of electrons or electron holes is low.

Hall sensor components contain amplifier circuitry in addition to the sensing element itself. Typically there is an open-collector output, and a comparator or Schmitt trigger to provide some hysteresis.

Hall-Effect Sensor Variants

Four primary variants of Hall sensors are widely used.

Unipolar Hall Sensor

This is activated when an external magnetic field exceeds a threshold value. When the field diminishes, the sensor switches off. Unipolar sensors are available in versions activated by the north magnetic pole or south magnetic pole.

Bipolar Hall Sensor

Proximity to one magnetic pole will switch it on. Proximity to the opposite magnetic pole will switch it off. The sensor remains in its current state (on or off) in the absence of a magnetic field.

Omnipolar Hall Sensor

Proximity to a strong magnetic field of either polarity will switch it on. Removal of the magnetic field will switch it off. An omnipolar sensor can be thought of as a pair of unipolar sensors mounted in opposite directions and with their (open-collector) outputs wired together. This component functions similarly to a reed switch, although of course it still requires a power supply.

Linear Hall Sensor

Also known as an *analog* Hall sensor, its output voltage varies in proportion to an external magnetic field instead of switching cleanly between high and low states. When no magnetic field is detected, the output is half of the sensor's supply voltage. In response to one magnetic polarity, the output can drop almost to zero. The opposite polarity can increase the output almost to supply voltage.

The output from a linear sensor usually is supplied from the emitter of an internal NPN transistor, not the collector. A minimum 2.2K resistor should be connected between output and ground.

The variable output can be interpreted as a measurement of distance between the sensor and a magnet. In this mode, a Hall sensor functions as a **proximity** sensor. However, it is not usually capable of measuring a distance of more than 10mm.

Other Applications

Hall sensors are incorporated in other types of components. A **magnetometer**, for example, may contain Hall sensors.

Additional discussion of Hall-effect sensors, with test circuits, will be found in the book *Make: More Electronics*, from which some of the figures here have been excerpted.

Values

Magnetic field at operating point is the minimum field necessary for the output to switch on. It is measured in tesla or gauss, and the abbreviation B_{OP} is used.

Magnetic field at release point is the maximum field that allows the output to switch off. It is measured in tesla or gauss, and the abbreviation B_{RP} is used.

Magnetic field range may be specified for linear (analog) Hall sensors.

Supply voltage may range as widely as 3VDC to 20VDC, or may be restricted between 3VDC to 5.5VDC. Check datasheets carefully.

Sourcing or sinking capability for the open-collector output is typically 20mA.

How to Use a Hall-Effect Sensor

Hall sensors are often made in 3-pin packages. Through-hole variants are usually made of black plastic and look like TO-92 transistors, but are slightly smaller.

Surface-mounted variants are common.

A typical through-hole Hall sensor has one bevelled face and a flat face on the opposite side. The bevelled face may be referred to as the "front" of the component in a datasheet. The sensor responds when an appropriate magnetic pole is brought close to the front face of the sensor.

The part number printed on the front of the sensor may be abbreviated as three digits. The code below this usually refers to the date of manufacture.

A simple circuit for a Hall sensor resembles a typical circuit for a phototransistor. Positive supply voltage and negative ground are applied to two of the three leads. Positive voltage is also applied, through a pullup resistor, to the third lead, which is the open-collector output (except in the case of a linear Hall sensor, described previously). The output pin is then tapped as the output from the sensor, to be applied to a component that will not draw more than 20mA.

Configuration of Object Presence Sensors

While most of the following suggestions relate specifically to Hall-effect sensors, some general principles may be applied to optical sensors.

Linear Motion

A presence sensor can be activated when the triggering source (such as light or a magnet) approaches it directly. This is sometimes referred to as *head-on mode*. Alternatively, triggering can be arranged when the source moves past the sensor. This is sometimes referred to as *slide-by mode*. The two modes are illustrated in sections 1 and 2 of Figure 3-16.

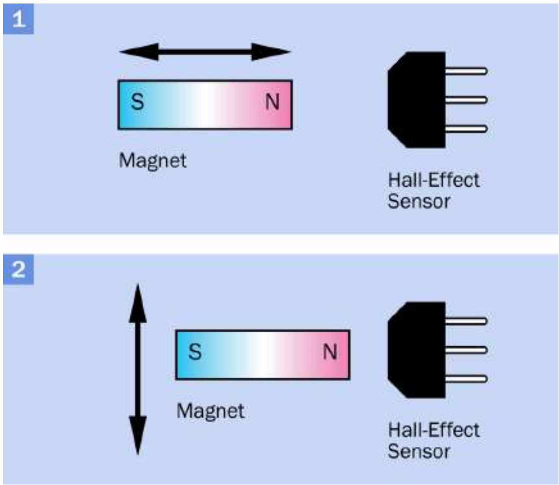

Figure 3-16 *Section 1 of this figure illustrates a presence sensor being used in head-on mode, while section 2 illustrates slide-by mode.*

Slide-by mode may be preferred because it eliminates the risk of damage to the sensor if overshoot occurs in head-on mode.

In slide-by mode, using a bipolar Hall-effect sensor, two magnets can be placed together with opposite polarity, creating a very steep transition in the overall magnetic field. This minimizes the risk of imprecise triggering. Using neodymium magnets, the triggering point can be adjusted with a precision of 0.01mm or better. See Figure 3-17.

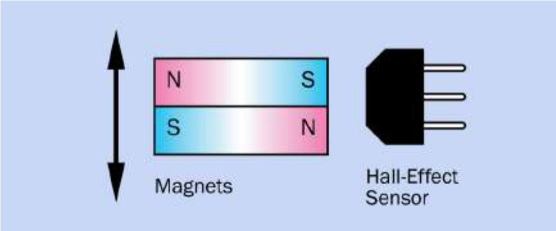

Figure 3-17 *In slide-by mode, two magnets with opposite polarity can be put together to create a very precise transition in a bipolar Hall-effect sensor.*

Sensing by Interruption

An *optointerrupter* is sensitive to an object passing between the light source and the light sensor. A comparable arrangement can be used with a Hall-effect sensor or a reed switch, but only if the interrupting object is thin and ferrous. This configuration is known as a *ferrous vane interruptor*.

Note that the magnet will exert significant force on the ferrous vane. This becomes an issue in sensors where the mechanical force is limited—for example, in paper-path sensors in a photocopy machine.

Additional information about the detection and measurement of moving objects will be found in the entry describing **linear position** sensors. See Chapter 6.

Angular Motion

One or more magnets can be used with a Hall-effect sensor to detect the angular motion, relative angular position, or absolute angular position of a rotating part. This data can be used to determine its speed of rotation. Some techniques to achieve this are described in the entry describing **rotary position** sensors. See Chapter 7.

Sensor Comparisons

Advantages of Optical Presence Sensors

- Not affected significantly by magnetic fields that can interfere with operation of a Hall-effect sensor or a reed switch.
- May be contained in a small, all-in-one package.
- Some optical sensors can operate over a distance of 50cm.
- Very well suited to sensing an object that blocks the light source (the *optointerrupter* configuration).

Disadvantages of Optical Presence Sensors

- Must have clear line-of-sight with the object and/or a reflector.
- Performance will be degraded by accumulation of dust or dirt.
- Limited lifetime of LED light source, if used continuously.
- May be accidentally triggered or impaired by some types of ambient light.
- Often requires a load resistor for the LED as well as a pullup resistor for the open-collector output.
- Range of acceptable voltage is usually narrow.

Advantages of a Reed Switch

- No polarity.
- No additional components required, other than a magnet.
- Able to switch AC or DC.
- Able to switch to voltages as high as 200V in some cases.
- Can be maintained by a magnet in open or closed state indefinitely without any power consumption.
- Many variants are capable of switching 500mA, and higher-current types are available.
- Can be activated through nonmagnetic materials (plastic, paper).
- Not significantly affected by dust or dirt that can degrade performance of an optical switch.

Disadvantages of a Reed Switch

- Requires a separate magnet (which must be placed carefully to avoid affecting other components).
- Cannot be miniaturized to the same extent as a surface-mount chip.
- Glass envelope is easily damaged.
- Arcing may occur between contacts.
- Will not work reliably when the activating magnet is more than a few millimeters from the switch.
- Can be activated accidentally by other magnetic fields.
- When sensing an object that comes between the switch and a magnet, only a ferrous object can be used.
- Must be debounced when connected with a logic chip or controller.

Advantages of a Hall Effect Sensor

- Robust solid-state component.
- Can be miniaturized for surface-mount applications.
- Very low cost.
- Fast response.
- No contact bounce.

- Extremely durable, with an almost unlimited lifetime.
- Not significantly affected by dust or dirt that can degrade performance of an optical switch.
- Low cost.

Disadvantages of a Hall Effect Sensor

- Requires a separate magnet (which must be placed carefully to avoid affecting other components).
- Open-collector output typically limited to 20mA or less.
- May be vulnerable to magnetic fields.
- When sensing an object that comes between the switch and a magnet, only a ferrous object can be used.

What Can Go Wrong

Optical Sensor Issues

Deterioration of LEDs

Most object presence sensors depend on infrared LEDs as a light source. An LED has many virtues (discussed in Volume 2), but does tend to suffer from gradual reduction in light output over a period of years. In a device such as a photocopy machine, which is used intermittently and may have an "economy" or "sleep" mode in which most of its components are powered down, LED-based detectors should last almost indefinitely. In other applications where an LED is powered continuously, its light output will diminish signicantly in 3 to 5 years. Bearing this in mind, an optical sensor should be chosen to operate well below its light-detection limits.

Object Too Close

Some optical and ultrasonic detectors triangulate on an object; that is, the light emitter and light sensor are angled slightly toward each other (as shown in sections 3 and 4 of Figure 3-3). The signal output from the sensor will peak at the distance where the emitter and sensor focus at a point. Consequently the output voltage will diminish when the object comes closer, which may create the misleading impression that the object is moving further away. To avoid spurious readings, detectors should not be used with objects that are closer than the minimum distance specified by the manufacturer.

Reed Switch Issues

Mechanical Damage

Bending the axial leads on a reed switch can easily fracture the glass envelope. Reed switches must be handled with care.

Contact Bounce

If the switch is wired to the input of a logic chip or microcontroller, contact bounce when the switch opens or closes is likely to be misinterpreted as multiple switching events. Debouncing will be necessary, either by additional components or by insertion of a momentary delay in the code embedded in a microcontroller.

Arcing

When switching high voltages or currents, an arc may be created briefly between the switch contacts, most often when they are moving from the closed to the open state. The arc erodes the contacts of the switch. Inductive loads make the arcing problem worse. If the switched voltage is kept below 5V arcing generally does not occur, extending the life of the switch.

passive infrared sensor

4

In coloquial speech, the term *motion sensor* is usually understood to mean a **passive infrared** motion sensor.

The acronym *PIR* is often used for a passive infrared sensor. It is always capitalized, without periods.

Object presence sensors and **proximity** sensors require an active source of a magnetic field, ultrasound, or infrared radiation. A passive infrared sensor does not require any such source, and responds passively to heat radiated from the object that is being detected.

OTHER RELATED COMPONENTS

- **object presence** sensor (see Chapter 3)
- **proximity** sensor (see Chapter 5)

What It Does

A **passive infrared** motion sensor, often described as a *PIR*, detects *black-body radiation*, which all objects emit as a function of their temperature relative to absolute zero. The sensor responds to infrared radiation centered around a wavelength of 10µm (10 microns, or 10,000nm). This is the approximate body temperature of people and animals.

The word "passive" in the term "passive infrared" refers to the behavior of the detector, which receives infrared radiation passively. **Proximity** sensors must generate their own infrared radiation actively, which is interrupted or reflected by nearby objects. See Chapter 5.

Schematic Symbols

Schematic symbols that are sometimes used to represent a passive infrared motion sensor are shown in Figure 4-1.

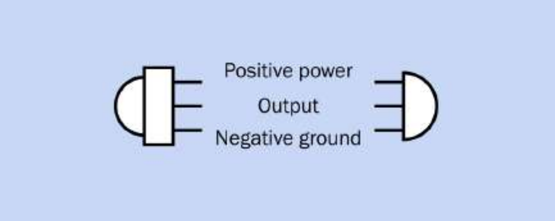

Figure 4-1 *Schematic symbols that may represent a passive infrared motion sensor. The orientation (pointing left or right) is arbitary. The pin sequence may vary.*

Applications

Motion-sensitive outdoor lighting almost always is based around a PIR. Similarly, a security system may sound an alarm or activate a video camera when a PIR indicates human activity.

Wildlife monitoring systems use PIRs to start a video camera that can then run for a preset interval.

Warning systems for automobiles have been developed that use a rear-facing PIR to detect pedestrians.

Industrial indoor lighting may use PIRs that switch the lights on automatically when people enter a room, and then switch the lights off (after a timed delay) when people are no longer detected in the room. The goal is to prevent wastage of energy as a result of employees forgetting to switch the lights off.

How It Works

A PIR module contains multiple components. Most visible is an array of at least 15 small lenses that focus infrared light from zones in the environment onto a *pyroelectric detector*, also known as a *pyroelectric sensor*. The response of the detector is processed by an amplifier, so that the signal can trigger an electromechanical **relay** or **solid-state relay** (see Volume 2). The relay operates an external device such as a light or an alarm.

Additional circuitry may allow the user to control the sensitivity of the PIR module and the length of time that the relay remains closed. The user may also be able to set the time of day when the PIR is active, or an additional **phototransistor** can shut down the PIR during daylight hours. If a phototransistor is included, its sensitivity is adjustable. For more information about phototransistors, see Chapter 22.

Pyroelectric Detector

The pyroelectric detector is actually a type of *piezoelectric* device. It is based around a wafer of lithium tantalate, which generates a small voltage in response to incident thermal radiation. However, like other piezoelectric components, it does not respond to a steady-state input, and must be activated by a transition. This distinguishes it from other types of light sensors, such as an infrared **photodiode**, in which the response is consistently related to a temperature input.

The response of a pyroelectric detector is suggested by the graphs in Figure 4-2.

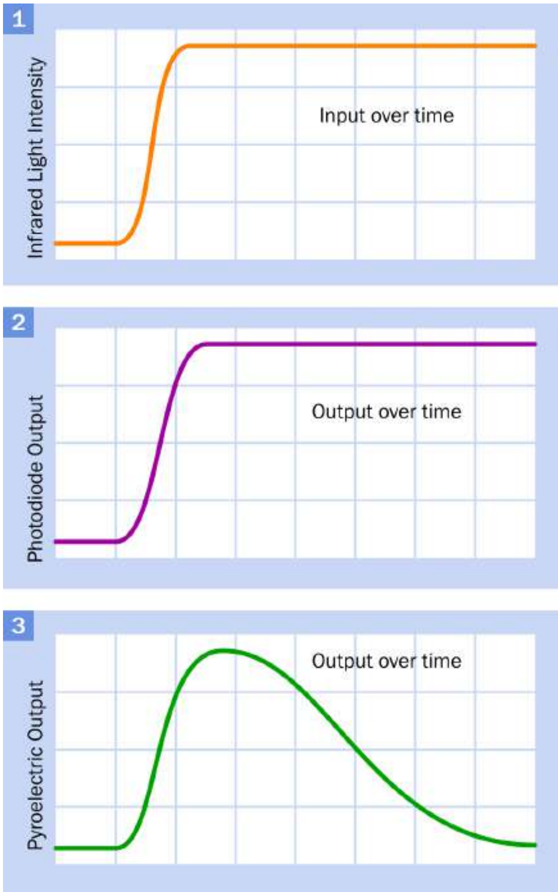

Figure 4-2 *Top: incident infrared light intensity. Center: the voltage output from a hypothetical photodiode. Bottom: the voltage output from a hypothetical pyroelectric detector.*

A pyroelectric detector in a PIR module is mounted in a sealed metallic container, as shown in Figure 4-3. The rectangular window in the detector is usually made of silicon, which is opaque to visible wavelengths but transparent to long-wave infrared radiation.

Elements

The pyroelectric detector used in a PIR contains at least two *elements* with opposite polarities, connected in series. If a sudden change in temperature affects both elements equally, their responses will cancel each other out. Thus, the detector ignores changes in ambient temperature. However, if a source of infrared radiation in the appropriate waveband affects one element before the other, the detector will emit two pulses of opposite polarity. See Figure 4-4.

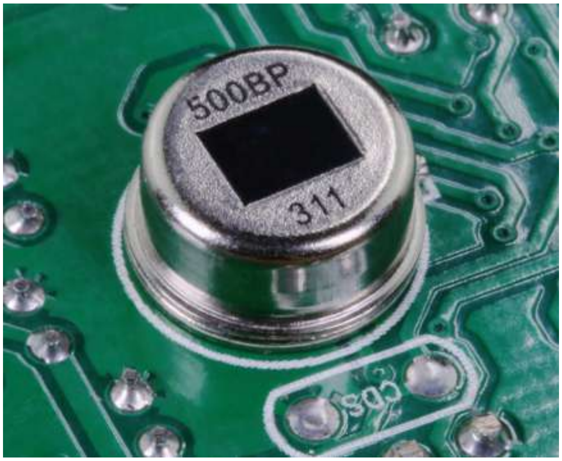

Figure 4-3 *A pyroelectric detector mounted on a small circuit board in a passive infrared radiation sensor.*

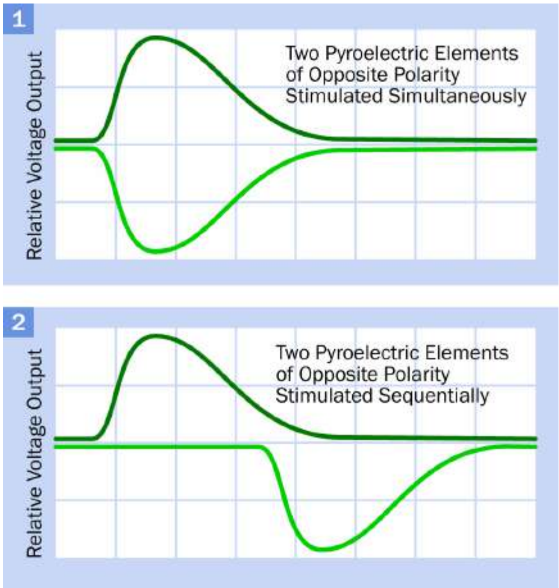

Figure 4-4 *Top: in a pyroelectric detector, if a change in temperature affects two elements of opposite polarity simultaneously, their voltages cancel each other out. Bottom: if one element is triggered before the other, the detector emits a signal.*

Lenses

Lenses are used to trigger the elements sequentially. Each lens faces toward one visible zone in the target area. When a source of infrared radiation moves from one zone to the next, it energizes the elements alternately, creating an output.

In some PIRs, four pyroelectric elements are used instead of two, to provide better coverage. The pairs of sensors may be wired in series or in parallel, but the principle remains the same.

The lenses are molded into a polyethylene dome that is often white and covers the pyroelectric detector. The dome appears smooth on the outside, but fine patterns of concentric ridges are molded on the inside. These are *fresnel lenses*, which are much cheaper, smaller, lighter, and easier to fabricate than conventional optical lenses. A fresnel lens introduces some distortion and aberration, but these defects are unimportant in a PIR.

Figure 4-5 illustrates the principle of a simple fresnel lens. The first section of this figure shows a conventional optical lens with one flat side and one curved side. A distant object emits almost-parallel beams of infrared light that are focused by the lens. Section 2 shows the same lens divided into segments that are stacked with no space between them. They behave in exactly the same way as the original lens. In section 3, each segment has been reduced in width, but because the optical faces still have the same geometry, they will still have the same function, although a small amount of distortion will be added by the reduction in width. This is a fresnel lens, which found an early application in lighthouses, where it greatly reduced the weight of very large glass lenses that focused the beam.

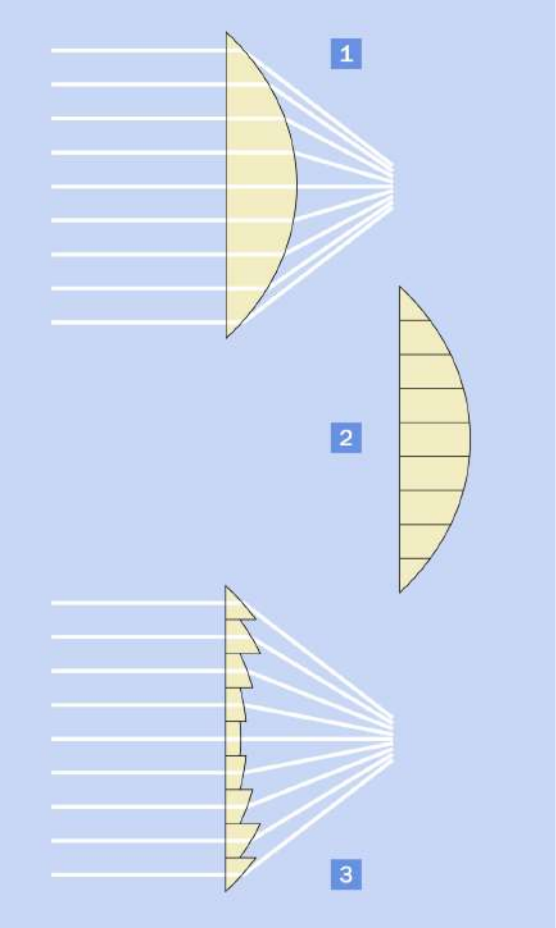

Figure 4-5 *Principle of a fresnel lens. See text for details.*

The same principle can be applied to a lens in which both surfaces are curved, as shown in Figure 4-6. In practice this will introduce more imperfections in the image, although some compensation is possible by adjusting the exact shape of the lens.

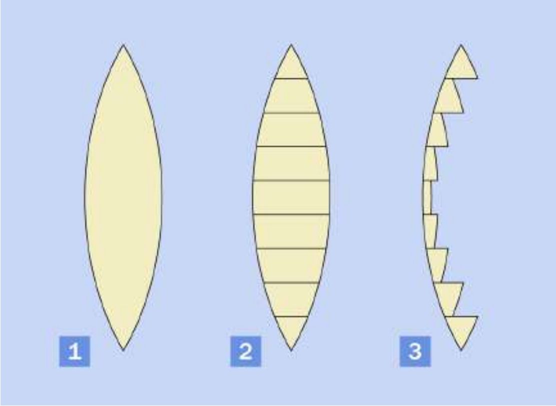

Figure 4-6 *The principle of a fresnel lens applied to a conventional lens in which both surfaces are curved.*

Figure 4-7 shows three curved fresnel lenses placed edge-to-edge, seen from above. In section 1 of this figure, infrared rays from a distant source are focused by the first lens onto the right-hand element of a pyroelectric sensor. In section 2, the external source has moved laterally, and the rays now focus on the left-hand element. In section 3, the source has moved into the zone covered by the centrally located fresnel lens, which focuses on the right-hand element again. The fluctuating inputs will trigger the sensor.

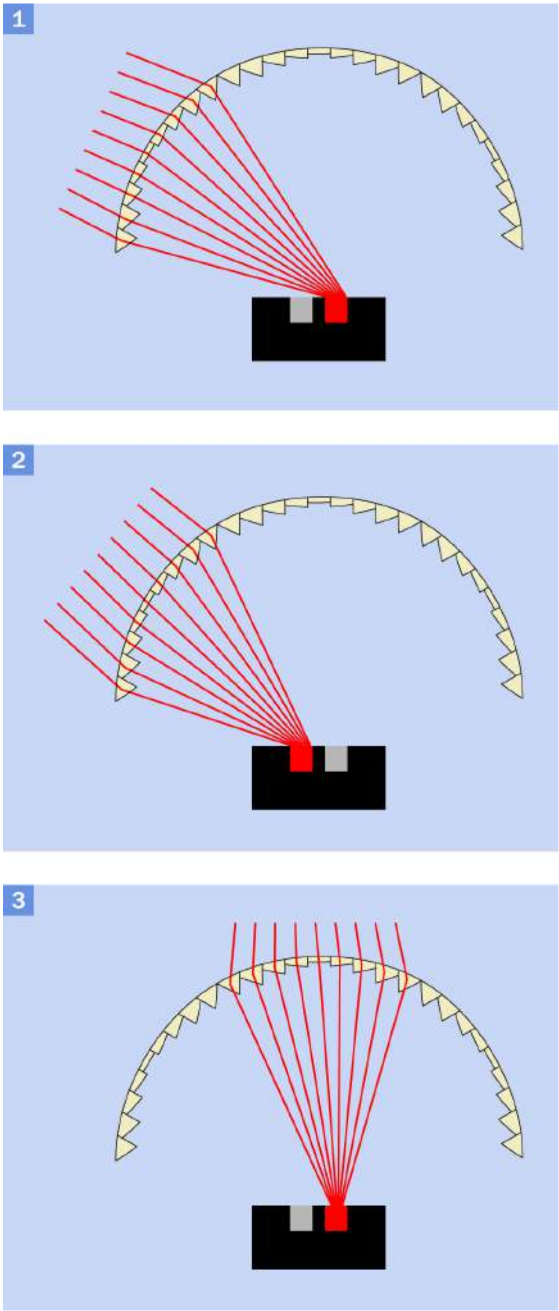

Figure 4-7 *Sections 1, 2, and 3 show fresnel lenses focusing an external source of infrared radiation on individual elements of a pyroelectric sensor.*

PIRs may combine fresnel lenses in a variety of patterns. Figure 4-8 shows an evenly weighted mosaic that would be suitable for a motion detector mounted on a ceiling, facing directly downward. Figure 4-9 shows a pattern weighted toward lateral motion, providing less sensitivity for motion above and below the primary band.

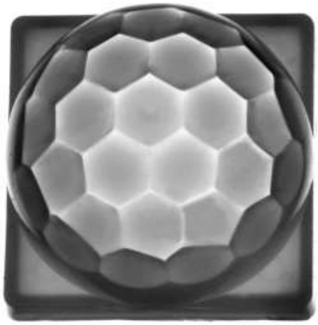

Figure 4-8 *An evenly weighted mosaic of fresnel lenses.*

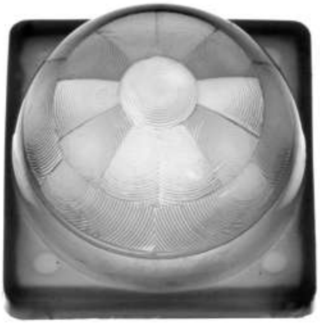

Figure 4-9 *A mosaic weighted more toward sensing lateral motion. The grooves of the fresnel lenses are visible.*

Variants

PIR sensor modules are available mounted on a small board such as the one shown in Figure 4-10 from Parallax, Inc. The detection range is 5 to 10 meters, selected by a jumper on the board. The three pins visible in the photograph are for power supply (3VDC to 6VDC), ground, and output. The output can source up to 23mA with a 5VDC power supply. Power consumption of the module is only 130μA when it is idle, or 3mA when it is active but has no load.

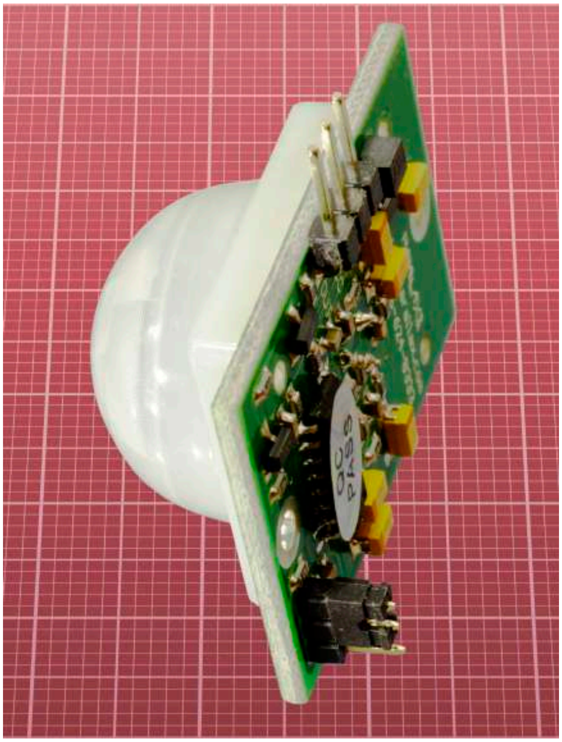

Figure 4-10 *A passive infrared detector mounted with basic necessary components on a small board.*

A board of this type still requires additional components to set the "on" time for a light or alarm, and to deactivate the PIR during daylight hours.

Various lens patterns are available, sold separately.

A PIR can be bought as a single component containing two elements and FET transistors to amplify the signal. Surface-mount and through-hole versions are available, requiring a typical power supply of 3VDC to 15VDC.

However, a PIR bought as a "bare" component requires significant external circuitry, using comparators or op-amps. Circuit design is nontrivial, entailing practical problems such as op-amps being sensitive to voltage spikes caused by activating a relay that shares the same power supply.

An easier alternative is an all-in-one detector, lens, and control circuit such as the Panasonic AMN31111, which is ready for board mounting. Its small output current of 100µA would be capable of activating a solid-state relay. Similar Panasonic PIRs offer a variety of ranges, sensitivities, and supply voltages.

The AMN31111 is in Panasonic's AMN series. There are many type numbers for combinations of analog or digital output, lens shape, and black or white lens. A selection of lens shapes, derived from the manufacturer's datasheet, is shown in Figure 4-11.

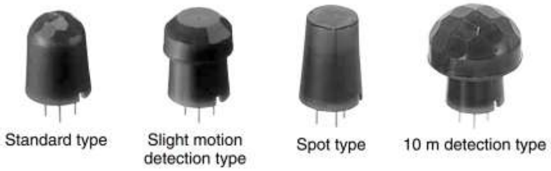

Figure 4-11 *Four lenses from the Panasonic AMN series of passive infrared sensors.*

What Can Go Wrong

Temperature Sensitivity

In warmer weather, objects in the field of view of a PIR will tend to be warmer, and the temperature difference between them and human skin will diminish. This can degrade the performance of a PIR.

Detector Window Vulnerability

The silicon window on a detector is vulnerable to dirt or grease. Avoid touching the component if it is not protected by lenses.

Moisture Vulnerability

Water absorbs far-infrared light. Consequently, condensation on the lens or detector can degrade performance, and a PIR may not function in heavy rain or snow.

proximity sensor

5

Proximity sensors that use infrared, ultrasound, and capacitance are described here. This Encyclopedia does not include proximity sensors that use magnetism, inductance, or other methods of determining distance.

Sometimes proximity sensors are referred to as *distance sensors*. An ultrasonic proximity sensor may be described as a *range finder* or a *ranger*.

High-end ultrasound proximity sensors sold as large, sealed modules with cabling can monitor the status of industrial processes. These commercial units are outside the scope of the Encyclopedia.

A sensor that detects whether an object is present, but does not measure the distance to it, is considered to be an **object presence** sensor and has its own separate entry. See Chapter 3. Many devices that are sold as proximity sensors or distance sensors actually do not provide meaningful distance data, and therefore in this Encyclopedia they are included with presence sensors.

Phototransistors and **photodiodes** may be used as sensing elements in proximity sensors. These components have their own entries as light sensors. See Chapter 22 and Chapter 21.

OTHER RELATED COMPONENTS

- **object presence** sensor (see Chapter 3)
- **passive infrared** sensor (see Chapter 4)

What It Does

A proximity sensor measures the distance between itself and a physical object that is often described as the *target*. The output from the sensor may be analog (voltage), serial data, or pulse-width modulation. It may be transmitted via a serial protocol such as SPI, TTL, or I2C, and may be stored as digital data in a register that is accessed by a microcontroller, using I2C. For additional details about protocol, see Appendix A.

Schematic Symbols

Either of the schematic symbols in Figure 5-1 may represent a proximity sensor, but are not used consistently. The sensor may also be represented as a rectangle containing text describing its function.

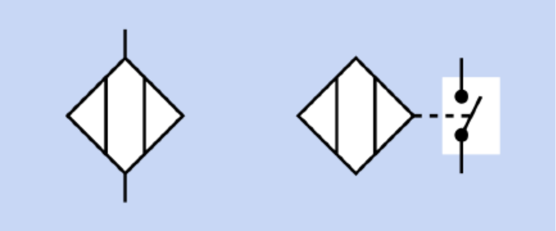

Figure 5-1 *Two ways to represent a proximity sensor in a schematic.*

Applications

In robotics, a proximity sensor can prevent a collision between a robot and an object or barrier in front of it. Some sophisticated proximity sensors can provide sufficient data for software to map the entire environment, but are outside the scope of this entry.

Proximity sensors can also be used in alarm systems, or for liquid-level sensing in storage tanks, or in automobiles to sound a warning if the driver is backing into an obstacle (although these devices are being augmented with rear-view video monitors).

In handheld devices, proximity sensors are used to sense the presence of the user's hand or face —for example, to shut down the display when a person raises a phone to talk into it.

Variants

This entry is subdivided to describe proximity sensors that use ultrasound, infrared light, and capacitance.

Ultrasound

An ultrasonic proximity sensor functions by emitting a short burst of sound and then listening for echoes from objects in front of it.

The sound is created by a **piezoelectric transducer** (see Volume 2) at a frequency between 30kHz and 50kHz, well above the range that can be detected by the human ear. The transducer may do dual duty as a microphone, sending and receiving sound on an alternating basis; or a second transducer, serving as a microphone, may be mounted beside the emitter on a small circuit board. The economically priced HC-SR04 is an ultrasound proximity detector popular in the robotics community, working reliably in the range of 2cm to 5m.

A board on which the sensor is mounted may include a microcontroller to measure the delay between propagation of a pulse and reception of the echo. Distance to the reflecting object is then calculated using the speed of sound in air at sea level, which is approximately 340 meters per second.

Infrared

An infrared proximity sensor requires a beam of infrared light from an LED that may be incorporated in the sensing module or mounted separately. Light reflects from the target and is detected by a **phototransistor** or **photodiode**. From the angle of the reflected light, onboard electronics can calculate the distance to the target by a process known as *triangulation*. See Figure 5-2. (This diagram is simplified. An actual sensor may use a linear array of photodiodes to assess the angle of the returning light beam.)

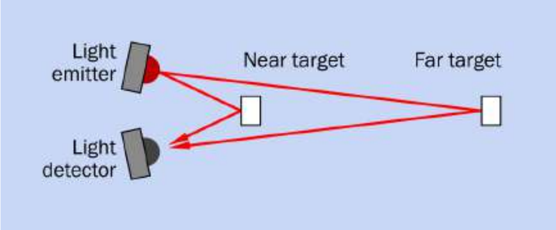

Figure 5-2 *An infrared proximity sensor determines the distance to an object by assessing the angle of reflected light.*

To reduce the risk of false positives, light from the LED contains only a narrow range of infrared wavelengths. Also, it is modulated at a frequency recognized by sensing circuitry in the module.

Relative Advantages

Ultrasound Devices

- Generally more suitable for detecting objects that are more than 1 meter away.
- Unaffected by direct sunlight, fluorescent tubes, and other light sources that can interfere with infrared devices.
- More accurate, often capable of placing objects within 5mm.
- Able to measure the distance to liquids and transparent objects, which may not be easily detected with infrared.

Infrared Devices

- Physically smaller—especially surface-mount versions.
- Able to measure the distance to soft objects, which may not be easily detected by ultrasound.
- More appropriate for targets that are closer than 10mm.
- More affordable.

Ultrasonic Examples

The proximity sensor in Figure 5-3 is a low-priced model, popular in the robotics community. Manufactured by MaxSonar, it uses a single element to send and receive. The manufacturer claims that it can detect a 6mm (quarter-inch) dowel 1.8 meters directly in front of it, and a 9cm dowel at 3.3 meters. This performance is beyond the capability of almost all infrared sensors.

MaxSonar offers a variety of sensors that appear physically very similar but have different range capabilities. Each has three outputs that can be accessed simultaneously: serial data at 9,600bps, analog voltage, and pulse-width modulation.

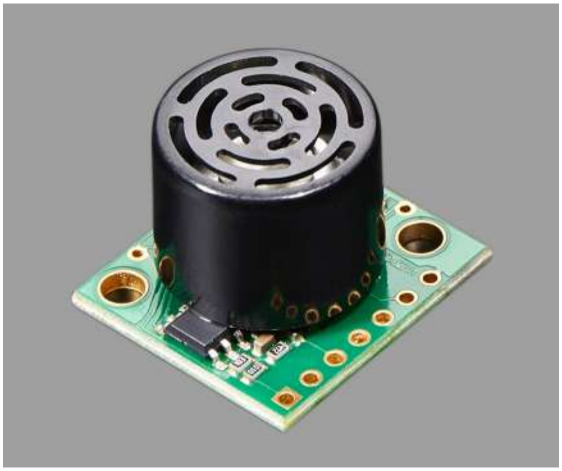

Figure 5-3 *The MaxSonar MB1003 can detect large solid objects as far away as 5 meters. Photo derived from an image by Adafruit.*

The serial output uses the RS232 protocol and consists of the letter R followed by four ASCII-coded numerals representing the measured range in millimeters. Thus, R1000 would indicate an object at 1 meter distance.

The analog voltage ranges linearly from 293mV when sensing an object at 300mm to 4,885mV for an object at 5,000mm.

The pulse-width output sends pulses ranging from 300µs indicating an object 300mm away to 5,000µs indicating an object 5 meters away.

The unit incorporates a temperature sensor that compensates for the lower density of air when its temperature increases. Weatherproofed versions are available. The power supply is 5VDC and must be smoothed.

Imports

Some international sources offer minimal paired ultrasound components at a very low price, such as the HC-SR04 from Cytron Technologies in Malaysia. See Figure 5-4.

Sound dispersion of the transducer is claimed to be plus-or-minus 15 degrees, and the claimed range is up to 4 meters. The module requires a triggering input pulse that must last at least 10µs. This prompts it to emit eight rapid ultrasound signals at 40kHz. The module measures the response time and applies a high state to its echo pin for a duration that is proportional with the distance measured. An external microcontroller must time the duration, then divide by a factor of 58 to obtain distance in centimeters.

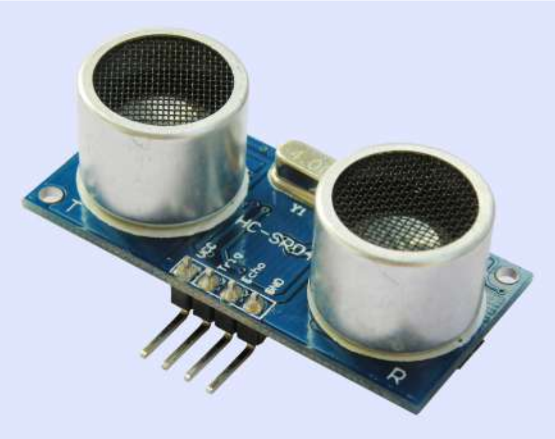

Figure 5-4 *The HC-SR04 is a very low-cost import that can provide acceptable performance when used with an external microcontroller.*

Many online sources offer simple code libraries for Arduino or PICAXE microcontrollers for use in conjunction with the HC-SR04.

Individual Elements

Individual ultrasonic components such as those shown in Figure 5-5 are available from many vendors. The user must add circuitry to generate a high frequency for a short duration, amplify the microphone signal, measure the time difference, and calculate distance.

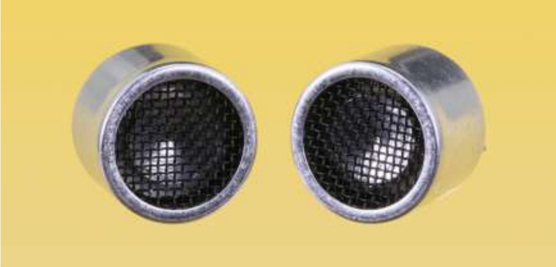

Figure 5-5 *These components, listed as the "40TR12B-R ultrasonic sensor kit" by the online supplier Jameco, could form the basis of a DIY ultrasonic proximity sensor.*

Infrared Examples

Sharp manufactures four infrared proximity sensors that are widely regarded as accurate and easy to use. They are popular in the robotics community. Their part numbers and ranges are:

- GP2Y0A51SK0F (20mm to 150mm)
- GP2Y0A21YK0F (10cm to 80cm)
- GP2Y0A02YK0F (20cm to 150cm)
- GP2Y0A60SZLF (10cm to 150cm).

The GP2Y0A60SZLF is the most recent product, with the most impressive specification. The GP2Y0A21YK0F is shown in Figure 5-6.

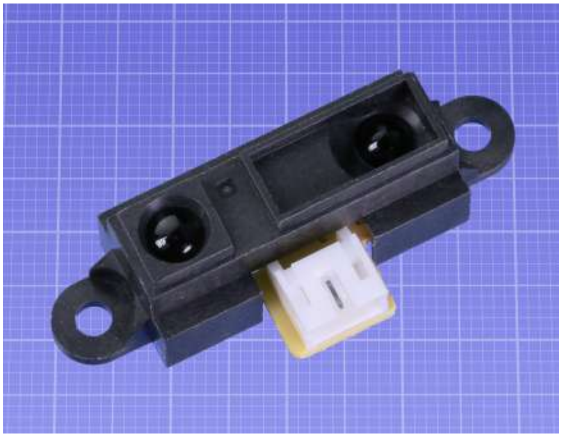

Figure 5-6 *The Sharp GP2Y0A21YK0F infrared proximity sensor. The background grid is in millimeters.*

Sharp describes these sensors as having an *analog output*. The voltage on an output pin varies in inverse proportion with the distance being measured. The relationship is illustrated by the graph in Figure 5-7, derived from the manufacturer's datasheet for model GP2Y0A02YK0F.

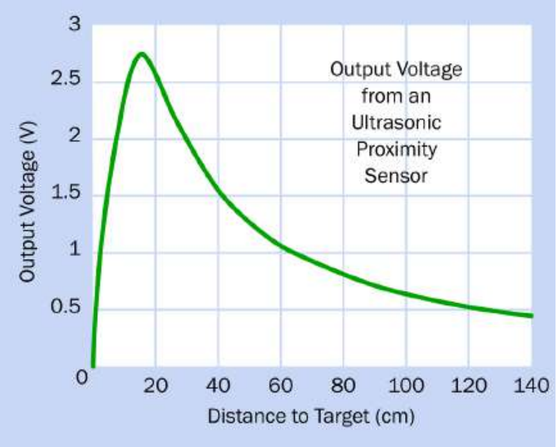

Figure 5-7 *Relationship between output voltage and target distance for a Sharp infrared proximity sensor, derived from the manufacturer's datasheet.*

The Sharp sensors can work with 5VDC. They consume around 30mA, except for the GP2Y0A60SZLF, which uses less current. Because the infrared LED functions in bursts, the manufacturer recommends protecting other components sharing the power supply by placing a 10μF capacitor across the sensor supply pins.

Trends in Infrared Proximity Sensing

Like many other types of sensors, proximity sensors have been affected by the huge market for handheld devices.

Handheld applications have had four consequences:

Miniaturization

Infrared proximity sensors are now commonly found in surface-mount chips measuring 5mm × 3mm or smaller.

Onboard processing

The status of a photodiode can be processed by a microcontroller on the same chip, to determine what the sensor is really "seeing." Input from an included ambient light sensor is factored into the evaluation.

Cost reduction

While chip-based proximity sensors have become increasingly sophisticated, their unit cost has plummeted, so that they are now actually much cheaper than simpler devices such as the Sharp analog sensors described above.

Complexity

Modern sensors must be programmed with a complex variety of instructions, and their coded output must be interpreted with a separate microcontroller running its own program. Whether this compensates for the low price and added functionality is something for the individual developer or experimenter to decide.

The Silicon Labs Si1145/46/47 series are chips with the kind of sophisticated capabilities required for handheld devices. An external microcontroller communicates with the sensor via an I2C connection, and can instruct it to adjust its distance range (from 1cm to more than 50cm), its analog-to-digital conversion sensitivity, and its current-sinking capability for up to 3 external LEDs. The chip incorporates ultraviolet sensing and ambient-light sensing capability. Its I2C connection can run at up to 3.4Mbps. Because its light output is pulsed for only 25.6μs every 800ms at 180mA, the average power consumption is only 9μA, assuming a supply voltage of 3.3VDC.

For additional details about protocols such as I2C, see Appendix A.

In addition to its use in handheld devices, the manufacturer suggests applications including heart-rate monitoring, pulse oximetry, and display backlighting control. These applications require only a subset of the features built into the sensor, but its price is so low, it may be cost-effective even when many of its capabilities are unused.

Proximity sensors with comparable specifications are available from many manufacturers. Examples are the Vishay VCNL4040 and the Avago HSDL-9100. Figure 5-8 shows the Silicon Labs SI1145 on the left, and the Avago HSDL-9100 on the right.

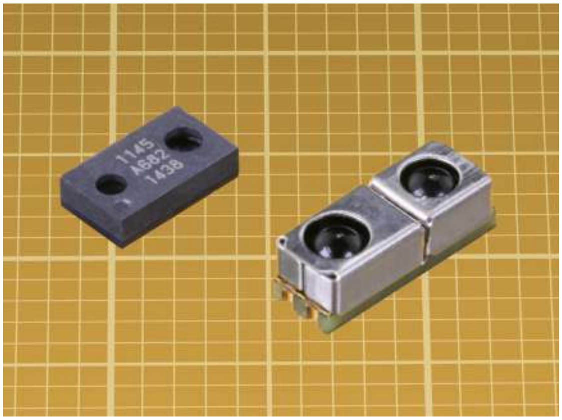

Figure 5-8 *Two sophisticated surface-mount proximity sensors with digital output. Left: Silicon Labs SI1145. Right: Avago HSDL-9100. The background grid is in millimeters.*

Several guidelines must be observed when using this kind of sensor. First and most obviously, if an external LED is used, it must have a peak wavelength compatible with the photodiode in the sensor. The LED must be placed as near as possible to the photodiode, as reducing the separation increases the sensitivity; but crosstalk between the photodiode and LED must be minimized, usually by placing a thin barrier between them that is barely taller than the higher component.

If the light emitter and/or light sensor are protected behind a transparent glass or plastic panel, it must have minimal resistance to infrared wavelengths, and its thickness must be chosen in compliance with guidelines in the sensor datasheet. To prevent light from reaching the sensor by reflection from the rear of the panel, a thin, opaque tube can be installed between the LED and the panel.

A sensor of this type may be configurable with a "detection scheme," meaning high and low sensing threshold levels appropriate to the object that is likely to be detected. Determining this may be a process of trial and error.

Capacitive Displacement Sensor

This is also known as a *capacitive linear displacement sensor*. It should not be confused with a capacitive **single touch** sensor, a human-input device that has its own entry. See Chapter 13.

A **capacitive displacement** sensor measures the distance between itself and a target that must be electrically conductive. Unlike optical or ultrasound position sensors, no additional source of light, sound, or other radiation is required. Unlike a magnetic position sensor, it does not require a separate permanent magnet. It simply measures its electrical capacitance with the target.

High-precision capacitive displacement sensors are used mainly for industrial process management. Lower-precision variants are much less expensive and can be used as **object presence** sensors to determine if an object is anywhere within a specified range.

A typical maximum range would be 10mm. For larger distances, optical and ultrasound sensors are more appropriate.

A selection of cylindrical sensor probes is shown in Figure 5-9.

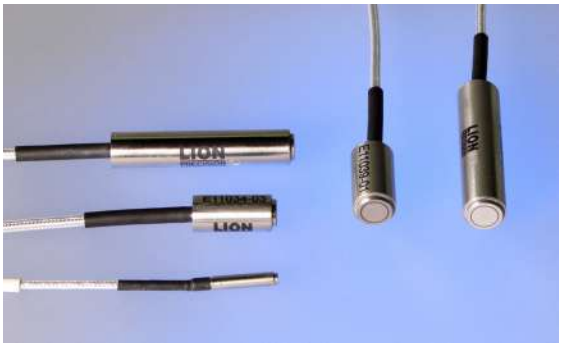

Figure 5-9 *Some high-precision capacitive displacement sensors from Lion, which builds them into cylindrical probes ranging from 3mm to 18mm in diameter.*

Applications

The high-precision variant of this type of sensor is commonly used during the production of small devices, such as disk drives. It can also measure vibration of a rotating metal part, such as a motor shaft, and may maintain automatic focus of a microscope.

Lower-precision variants can be used for applications such as counting objects on a conveyor.

When used to measure materials thickness, the sensor finds applications in checking automobile brake rotor fabrication and the thickness of silicon wafers.

How It Works

When two plates of electrically conductive material are placed opposite each other, they possess capacitance. This means that they have an electrical storage capacity enabled by accumulation of opposite charges on each plate.

The capacity is directly proportional to the plate area, inversely proportional to the distance between the plates, and is affected by the medium separating the plates, known as the *dielectric*.

If the plate area and dielectric remain constant, the distance between the plates will be the only factor affecting the capacitance. Therefore, the distance between the plates can be calculated by measuring the capacitance.

Measurement can be performed by evaluating the *displacement current* that passes through the dielectric from one plate to another when a pulse of voltage is applied. (Hence the term "capacitive displacement" in the name of this type of sensor.)

A detailed explanation of displacement current is included in the second edition of *Make: Electronics*.

The sensor itself functions as one plate of the capacitor while the target functions as the opposite plate. Alternating current is applied as a rapid series of pulses, and the current that passes between the plates is proportional to the distance between them.

Ideally, the target should be grounded to the current source. However, since AC is being used, capacitive coupling of the target to the current source is also possible, so long as the additional capacitor has a value of 0.1µF or higher.

Sources of Error

To obtain a meaningful measurement, the electric field from the sensor is focused on the target. Some dispersion will still occur, and the combination of sensor and target must be compatible. A target typically should be flat and should have a larger surface area than the sensor.

Humidity can affect sensor performance, as it changes the value of the dielectric. Temperature can affect performance, partly because it causes small dimensional variations in the sensor and the target.

The surface of the sensor and the surface of the target should be precisely parallel, because the spot where the field hits the target will be elongated if the target is tilted relative to the sensor. Elongation increases the capacitive area and affects the accuracy of measurement.

This type of sensor can also measure the thickness of a nonconductive material, if the material is a thin sheet that can be interposed between two sensors. In this mode the material functions as a dielectric, and its thickness will affect the AC current passing through it.

While lower-precision capacitive displacement sensors are relatively unusual, they can potentially serve as a relatively affordable and simple **object presence** sensor, so long as the target is conductive and will not be damaged by passing a small alternating current at a relatively low voltage.

Values

A high-precision capacitive displacement sensor can measure distances usually ranging from 0.25mm to 10mm with an accuracy that can be better than 0.05mm. High voltages are not required, with a supply of plus-or-minus 15V being common.

The sensing element, often referred to as a *probe*, is usually plugged into a custom-designed control unit that converts the capacitance measurement into a variable output voltage. The performance is then expressed in millimeters per volt. Thus if the voltage varies by 5V over a distance of 1mm, the sensor is rated as providing 0.2mm per volt.

What Can Go Wrong with Optical and Ultrasound Proximity Sensors

Object Too Close

Because both types of proximity sensors (ultrasonic and infrared) may include emitters that are angled toward the distance range for which they are designed, they may fail to “see” an object that is closer. Consequently the sensor will provide no response, or may detect a different object that is further away. In either case, if the sensor is being used on a moving device, the device may collide with the undetected near object.

Multiple Signals

If two or more sensors and emitters are used concurrently, their combined signals can interfere with each other and create inaccurate readings.

Inappropriate Surfaces

Ultrasonic proximity sensors are intended to identify a single object that is closer to the sensor than other objects, within a narrow beam dispersion angle. Multiple objects, complex surfaces, soft surfaces such as clothing or furnishings, or an unusual configuration of interior walls can create inaccurate readings.

Infrared sensors may be unable to “see” liquids or transparent objects, and may give different assessments of distance depending on the properties of a surface. Human skin, for example, is a poor reflector, as it absorbs some infrared radiation.

Environmental Factors

An ultrasonic transducer uses a very small moving diaphragm to generate sound. Like any system containing moving parts, it will be vulnerable to moisture or excessive humidity, and may need to be protected.

After a device is built and tested indoors, in a controlled environment, it is likely to behave differently if it is moved outside where the temperature is significantly higher or lower.

Deterioration of LEDs

As noted in the section on presence sensors (see Chapter 3), LEDs tend to suffer from gradual reduction in light output over time. The performance of an infrared proximity sensor may deteriorate over a period of years, depending on how intensively it is used.

linear position sensor

A **linear position** sensor may also be described as a *linear displacement sensor* or a *linear position transducer*. It is sometimes categorized as a type of **proximity** sensor, but in this Encyclopedia a proximity sensor emits a signal and receives an echo to measure the distance from itself to an object. By comparison, a linear position sensor measures the position of a sliding object within a stationary enclosure.

An **object presence** sensor can be considered as a form of linear position sensor, but it only responds to the presence of an object, without measuring its position.

OTHER RELATED COMPONENTS

- **proximity** sensor (see Chapter 5)
- **object presence** sensor (see Chapter 3)
- **rotary position** sensor (see Chapter 7)

What It Does

Control of a mechanical device may require accurate and timely information about the position of a movable part in the device. A **linear position** sensor can be used for this purpose.

Three attributes are likely to be of interest:

- Position
- Direction of motion
- Speed of motion

Typically a linear position sensor measures only the first attribute. Additional electronics may calculate the second and third attributes by taking multiple position readings. Thus a *speed sensor* is very likely to be built around a position sensor, and therefore this Encyclopedia does not have a separate entry for speed sensors.

Applications

Robotic arm positioning, wing flap and rudder positions on an aircraft, computer-controlled machine tools, 3D printers, and automobile seat position sensors are some of the many applications.

Schematic Symbol

In a schematic, a linear position sensor may be represented by symbols for the sensing elements that are inside it (potentiometer, LED, phototransistor, or others).

How It Works

Linear potentiometers, *magnetic linear encoders*, *optical linear encoders*, and *linear variable differential transformers (LVDTs)* are described here. Other options are available, but they tend to be more specialized, and are not included.

Linear Potentiometer

For a full description of **potentiometers**, see Volume 1.

A linear potentiometer, often referred to as a *slider potentiometer*, contains an electrical resistance in the form of a straight section of *track.* The track may be a strip of *resistive polymer* or (less often) may consist of an insulator with a coil of nichrome wire wrapped around it.

For sensing purposes, the potentiometer is wired as a voltage divider, and a fixed potential is applied across its full length, as shown in Figure 6-1. A *wiper* slides along the track, sensing a voltage that varies linearly with the wiper's position. Output from the wiper can be used directly to control an analog indicator such as a meter, or can be processed by an analog-to-digital converter.

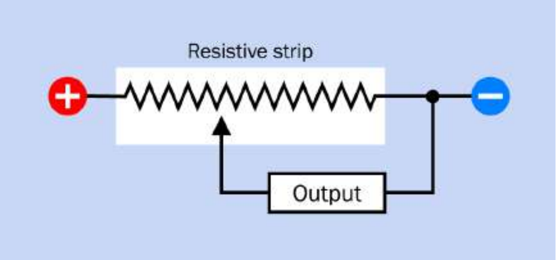

Figure 6-1 *A linear potentiometer can consist of a fixed track with a known voltage applied at each end, and a movable wiper.*

For audio applications, a slider potentiometer may have resistance that varies logarithmically with the position of the wiper. However, this type of component is not generally used as a position sensor.

For sensing purposes, the potentiometer is usually protected by a long, narrow box or tube through which a rod slides on sealed bearings. An example is shown in Figure 6-2.

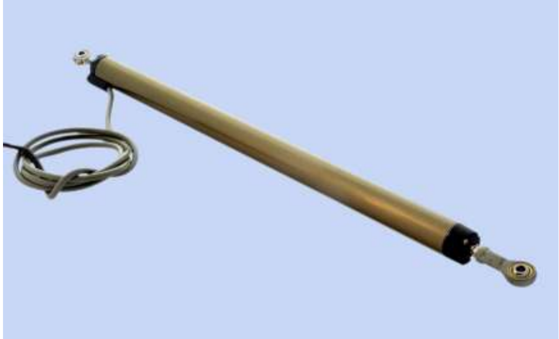

Figure 6-2 *A linear potentiometer contained in a tube with sealed bearings.*

Small linear potentiometers for position sensing are available from companies such as Bourns. The one shown in Figure 6-3 is about 20mm long, and the rod that slides through it has a travel of approximately 10mm. The component is available with resistance values ranging from 1K to 50K, and its power rating is 1/8 watt. The manufacturer claims a life expectancy of 500,000 cycles.

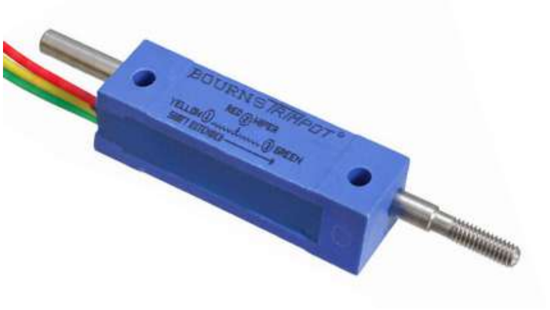

Figure 6-3 *A miniature linear potentiometer. The body is about 20mm long.*

A linear potentiometer is simple, inexpensive, compact, and requires few additional components. The track contains a lubricant, but inevitably some wear occurs as a result of motion of the wiper. Life expectancy will be reduced by vibration or by contamination with dirt or moisture.

A linear potentiometer may rarely be described as a *linear potentiometric sensor*.

Magnetic Linear Encoders

A ferrous rod or strip can be magnetized with alternating north and south poles. When it slides past a single bipolar Hall-effect sensor, it generates a pulse train from the sensor that can be interpreted to provide positional information. The principle of the component is shown in Figure 6-4. (Magnetic rotary encoders also exist; see "Rotary Encoders".)

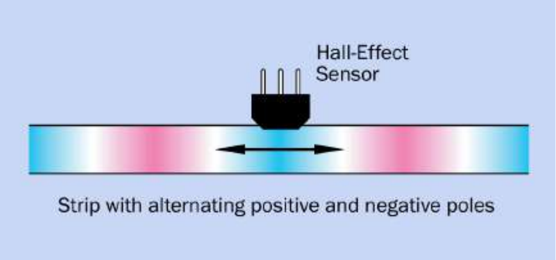

Figure 6-4 *When a strip magnetized with alternating north and south poles slides past a sensor, the pulse train from the sensor can be decoded to indicate the relative position of the strip.*

The sensing element may be described as a *read head*. If two are used, with a spacing equal to half the interval between the north and south poles on the strip, the phase difference between the pulse trains from the sensors will indicate the direction in which the magnetized strip is moving. This is shown in Figure 6-5.

The combination of pulse trains is known as *quadrature*, because there are four possible combinations: A and B both high, A and B both low, A high and B low, or A low and B high. The same principle is used in an optical rotary position sensor; see Figure 7-6.

This type of linear position sensor is often described as a *magnetic encoder*, meaning that the position of the sliding part is encoded in a series of pulses. Relatively high resolution is possible, as the north and south poles on a strip of ferrous material can be as close as 2mm. Optical encoders may use the same principle; see "Optical Linear Encoders".

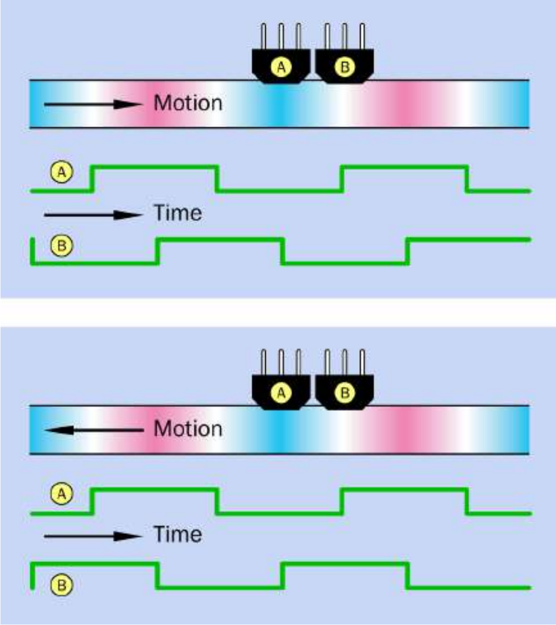

Figure 6-5 *Pulse trains from Hall-effect sensors A and B are shown in green. The phase difference between them can be interpreted to show the direction in which the magnetized strip is moving.*

The sensor can be built into a module containing an analog-to-digital converter that provides numeric output defining the position of the read head.

In an *absolute* magnetic encoder, nonvolatile memory can store the digitized position when the device is switched off. An *incremental* magnetic encoder does not store this information, and will require at least one additional *home sensor* to detect when the encoder is at either end of its travel. At power-up, an initialization routine moves the magnetized strip until the home sensor is triggered.

For details of Hall-effect sensors see "Hall-Effect Sensor".

Optical Linear Encoders

The operation of an optical linear encoder is identical to that of the magnetic linear encoder described immediately above, except that a sliding *optical grating* is used in conjunction with a light source and a detection device such as a **phototransistor** or **photodiode** that functions as the read head. The principle is shown in Figure 6-6. The grating may be described as a *codestrip*.

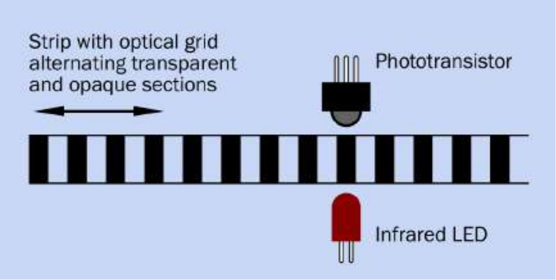
Strip with optical grid alternating transparent and opaque sections
Phototransistor
Infrared LED

Figure 6-6 *An optical linear encoder uses the same general principle as a magnetic linear encoder.*

For details of phototransistors, see Chapter 22. For details of photodiodes, see Chapter 21.

An example of a low-cost optical linear encoder is the Avago HEDS-9 series, which consists of a horseshoe-shaped module with an LED in one arm and an array of photodiodes in the opposite arm. When a codestrip passes between the arms, the module emits two pulse trains from internal comparators. The pulse trains are 90 degrees out of phase, and can be interpreted to show which way the codestrip is moving.

These sensors have a body measuring approximately 10mm wide and are designed to read opaque/transparent intervals ranging from 1.5 to 7.87 cycles per millimeter. The output rate can be as high as 20kHz. No pullup resistor is required on the output, as a 2.5K resistor is integrated.

A *codewheel* may be used, in which case the module senses rotation instead of linear motion. This is described in detail in the entry on **rotary position** sensors. See Chapter 7.

Linear Encoder Applications

Optical or magnetic linear encoders are found in some laboratory equipment, machine tools, and industrial robots. The mean time between failures can range from 100,000 to 1,000,000 hours. Optical encoders must be sealed to provide good protection from dust and dirt.

Linear Variable Differential Transformers

This type of sensor, often referred to by the acronym LVDT, tends to be used in industrial environments where great reliability is required under severe conditions. Examples are high-temperature steam valves and nuclear reactor control mechanisms. However, the robust, frictionless design suggests other possible applications, and custom fabrication is a possibility.

Figure 6-7 illustrates the general principle. Three coils are wound sequentially around a (nonmagnetic) stainless-steel tube, enclosed in a second tube also made of stainless steel. The coils act as transformers with the variable voltage ratio determined by the position of a solid iron armature that slides through them.

The center coil is the primary winding, to which AC is applied between 2kHz to 50kHz, depending on the application. (The frequency must be at least ten times the maximum rate of motion of the armature.) The iron armature is attached to a nonmagnetic rod. The voltages on the secondary windings provide the output from the sensor.

While more than one wiring arrangement is possible, the most common schematic is shown in Figure 6-8. The secondary coils are in series, with one of them reversed, so that the phase of the output is inverted as the armature moves from one end of its travel to the other. The phase detector responds to the phase difference by creating a DC output that varies with movement of the armature. All the functions shown in the schematic are available on a single integrated circuit chip.

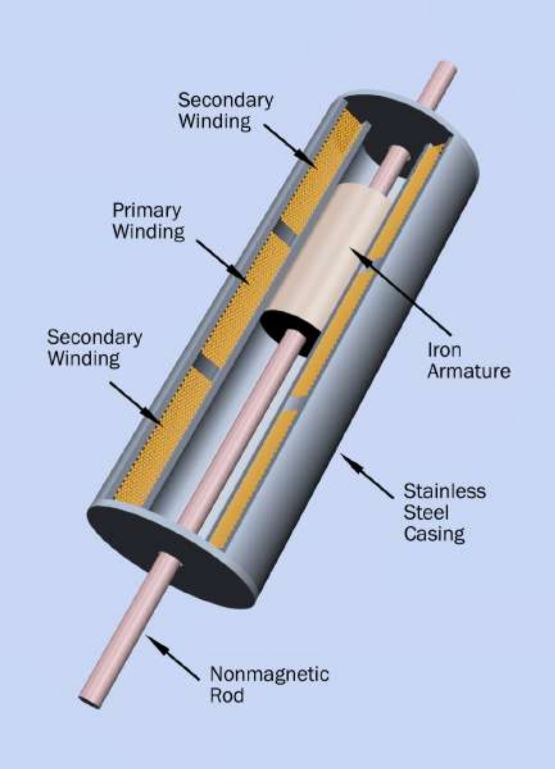
Secondary Winding
Primary Winding
Secondary Winding
Iron Armature
Stainless Steel Casing
Nonmagnetic Rod

Figure 6-7 *Cutaway drawing showing the internal design of a linear variable differential transformer, in which the position of a sliding iron armature determines the voltage induced in the secondary windings.*

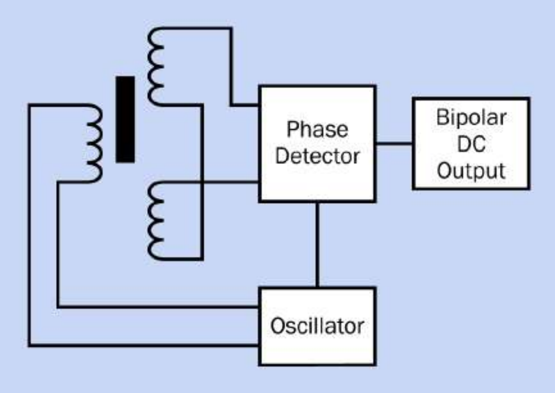
Phase Detector
Bipolar DC Output
Oscillator

Figure 6-8 *Typical schematic for using a linear variable differential transformer.*

An example of a linear variable differential transformer is shown in Figure 6-9.

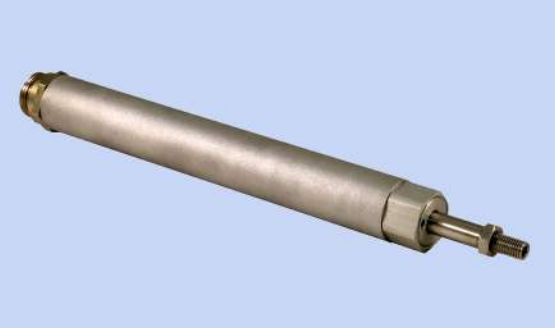

Figure 6-9 *External view of a linear variable differential transformer.*

What Can Go Wrong

Mechanical Issues

Any sliding mechanism will involve friction, and will be vulnerable to wear and tear, resulting in looseness that will degrade its accuracy. In addition, optical systems are vulnerable to dust and dirt.

LED Longevity

Light output of an LED diminishes over a period of years if the LED is "always on." This will limit the life of the sensor.

rotary position sensor

7

Alternative terms for **rotary position** sensor are *rotary sensor*, *rotational position sensor*, *angular position sensor*, and *angle sensor*.

The terms *rotary* and *rotational* are used interchangeably. In this entry, an attempt has been made to use the term that is most common for each specific application. For example, "rotary position sensor" versus "rotational encoder."

A few sensors are available that specifically measure *rotary speed*, but generally a rotary speed sensor uses information from a rotary position sensor. Therefore, this Encyclopedia does not include a separate entry for rotary speed sensing.

A **rotational encoder** can be used as a rotary position sensor. It is mentioned briefly here but is described in more detail in Volume 1 of this Encyclopedia in the entry describing **switches**.

OTHER RELATED COMPONENTS

- **linear position** sensor (see Chapter 6)

What It Does

Control of a mechanical device may require accurate and timely information about the orientation of a rotating part in the device. A **rotary position** sensor can be used for this purpose.

A sensor may be capable of measuring three attributes:

- Angular orientation
- Direction of rotation
- Speed of rotation

Typically a rotary position sensor measures only the first attribute. Additional electronics are required to calculate the second and third attributes by taking multiple position readings. Thus a *speed sensor* is very likely to be built around a position sensor, and therefore, this Encyclopedia does not have a separate entry for speed sensors.

Applications

In robotics, a rotary position sensor is commonly used to show the orientation of a pivoting arm or strut. It can also be used as a limit switch on a motor.

Specific applications include solar array positioning, remotely piloted vehicles, guidance and navigation, antenna positioning, and wind turbine pitch control.

Pulses from a rotary position sensor are used to measure speed of rotation in vehicles, industrial processes, and aviation. Small rotary speed sensors are built into devices such as cooling fans and computer hard drives.

Schematic Symbol

In a schematic, a rotary position sensor may be represented by symbols for the sensing elements that are inside it (potentiometer, LED, phototransistor, and others).

Potentiometers

Single-turn and multiturn potentiometers may be used as rotary position sensors. For basic information about **potentiometers**, see the entry describing this component in Volume 1.

Arc-Segment Rotary Potentiometer

An *arc-segment rotary potentiometer* is often referred to simply as a "potentiometer," as this type is more common than the multiturn type or the linear type. When used as a sensor, it can measure a turn angle that is less than 360 degrees.

This component contains a resistor in the shape of an arc, referred to as the *track.* It may be a strip of *resistive polymer* or (less often) may consist of an insulator with a coil of nichrome wire wrapped around it.

For sensing purposes, the potentiometer is wired as a voltage divider, and a fixed potential is applied along the full length of the track, as shown in Figure 7-1. A *wiper* slides along the track, sensing a voltage that varies linearly with the wiper's angular position. Output from the wiper can be used directly to control an analog indicator such as a meter, or can be processed by an analog-to-digital converter.

For audio applications, an arc-segment potentiometer may have resistance that varies logarithmically with the position of the wiper. However, this type of component is not generally used as a position sensor.

Low-cost potentiometers were traditionally used as volume or tone controls in stereo systems. When a potentiometer is designed for use as a sensor, it tends to be more ruggedly built and better protected against dust, dirt, and moisture. Its advantages are that it is simple, inexpensive, compact, and requires few additional components.

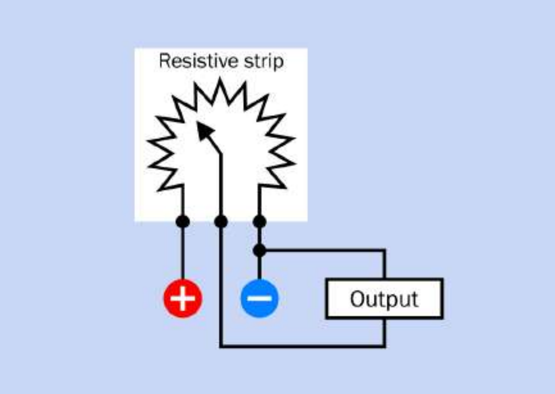

Figure 7-1 *An arc-segment potentiometer consists of a fixed track in the shape of an arc with a known voltage applied at each end, and a movable wiper.*

The major disadvantage is that although the track contains a lubricant, some wear will gradually result from friction with the wiper. Life expectancy will be further reduced if there is vibration or contamination with dirt or moisture.

End Stops

An arc-segment potentiometer usually has *end stops* to prevent the wiper from running off either end of the track. Typically, these stops limit rotation to around 300 degrees.

A few arc-segment potentiometers allow unrestricted rotation. The Bourns 6639 series is an example, although the wiper still passes through a "dead zone" of 20 degrees between the start and end of its track. An application for this type of potentiometer could be as a direction sensor for a weather vane.

Multiturn Rotary Potentiometer

A spiral track, which resembles a coil spring, enables a rotary potentiometer to make multiple turns. The wiper rotates inside the track and follows its contour. A potentiometer of this type will still be calibrated in degrees; thus, a 10-turn component may be listed as allowing an *electrical travel* of approximately 3,600 degrees.

The exterior of a multiturn rotary potentiometer is shown in Figure 7-2.

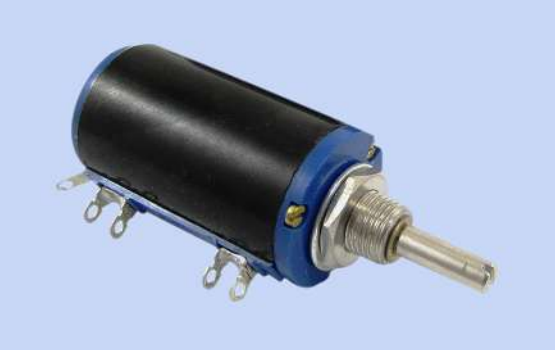

Figure 7-2 *A multiturn rotary potentiometer. Each pair of solder tags is attached internally to one end of the internal coiled track. The smaller tag at the far end connects with the wiper.*

A simpler type of multiturn potentiometer is intended for use as a trimmer—a small potentiometer that can be mounted on a circuit board to allow adjustment or calibration, often during the manufacturing process. This type of trimmer contains a worm gear that engages with a spur gear internally. The wiper is mounted on the spur gear. It has no applications as a sensor, but is mentioned here to avoid ambiguity, as it is probably the component that is most commonly referred to as a "multiturn potentiometer."

Magnetic Rotary Position Sensor

Externally, a modern magnetic rotary position sensor may look very much like an arc-segment rotary potentiometer. Internally, a permanent magnet is attached to the base of the shaft, and one or more **Hall-effect** sensors are mounted on a small circuit board immediately below the magnet, in the bottom of the enclosure. A simplified diagram appears in Figure 7-3.

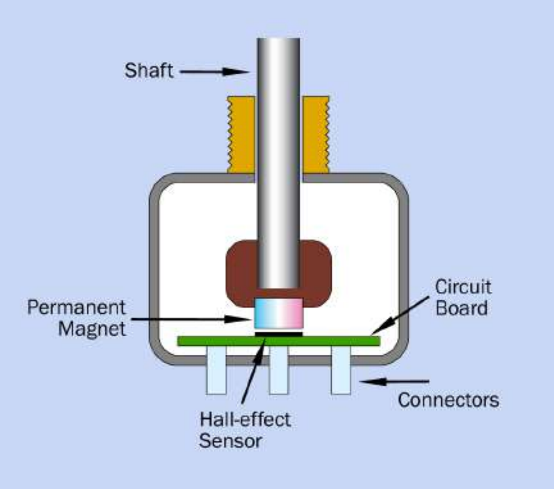

Figure 7-3 *Simplified interior view of a magnetic rotary position sensor.*

For more information about Hall-effect sensors, see "Hall-Effect Sensor".

A magnetic rotary position sensor may be described as a *noncontacting* sensor. Two views of an example are shown in Figure 7-4.

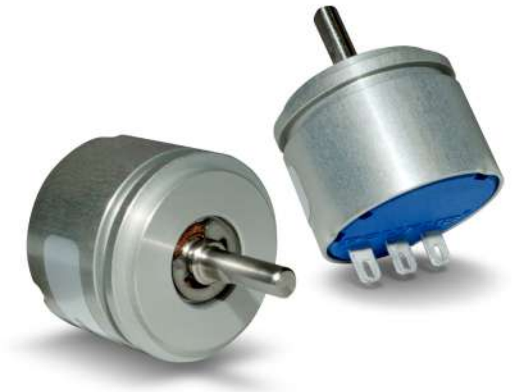

Figure 7-4 *Two views of a Bourns AMS22 magnetic rotary position sensor.*

The AMS22 sensor shown in Figure 7-4 has an analog output ranging from 0.1VDC to 4.9VDC when powered with 5VDC. A noncontacting sensor of this type will cost three to four times as much as a conventional potentiometer, but has the great advantage of extreme durability, with the manufacturer claiming a life of 50 million shaft rotations. A disadvantage is its low-current output, limited to 10mA. The maximum rotation speed of its shaft is 120rpm.

Rotary Position Sensing Chips

Chips of the type used in a magnetic rotary position sensor are available as individual components, many of which have advanced features. For example, the AM8192B angular magnetic encoder by RLS is a 44-pin surface-mount chip containing Hall sensors that detect the orientation of a permanent magnet above or below the chip. Various outputs provide information such as sine or cosine of the turn angle, incremental pulses, and digitized output via an SPI interface.

Rotary Encoders

Many rotating position sensors communicate their angle of rotation with an output that consists of a pulse train or some other coded signal. This type of rotary sensor is known as a *rotary encoder* or *rotational encoder*. (Linear encoders also exist; see "Magnetic Linear Encoders".)

The simplest version of this component is a *mechanical encoder* containing two small electromechanical switches that are activated, out of phase, by a toothed wheel attached to the rotating shaft. This component is described in detail in Volume 1 of the Encyclopedia, where it is categorized as being a form of **switch**. Its low cost and simplicity has made it popular for rotary controls on car radios and small stereo systems, but its switches have a limited life expectancy and create a "noisy" output that must be *debounced* if connected with a logic chip or microcontroller. Typically the microcontroller will include a pause of up to 50ms in its program code to allow time for the switch contacts to settle (although some manufacturers claim only 5ms).

Confusingly, a mechanical encoder is often identified only as a "rotational encoder," even though optical and magnetic rotational encoders also exist, as described immediately below. As a general rule, if a component is described simply as a rotational encoder, it probably contains electromechanical switches.

Optical Rotary Encoders

This type of component works on the same principle as an *optical linear encoder* of the type described in "Optical Linear Encoders". The difference is that a *codewheel* is used instead of a *codestrip*. Typically the codewheel is supplied by the manufacturer of the component that is designed to read it.

A *transmissive* codewheel is shown in Figure 7-5. The distance between the light emitter and the light detector has been exaggerated in this diagram for clarity.

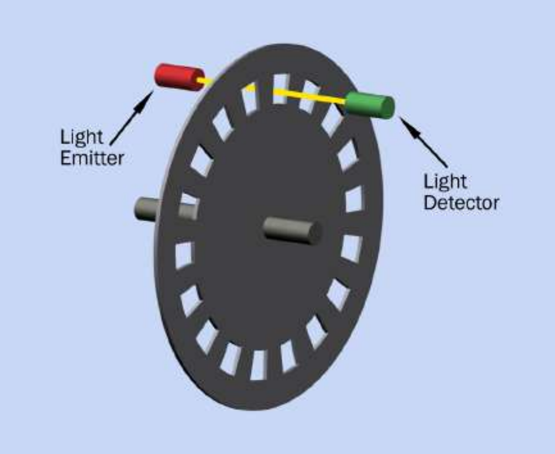

Figure 7-5 *A light-transmissive codewheel for use in an optical rotary encoder.*

Some optical rotary encoders use a *reflective* codewheel, in which case alternating sections of the wheel are light-absorptive and light-reflective, and the emitter and detector are both on the same side of the wheel.

If only one light emitter and light detector are used, the pulse train from the sensor reveals how many increments the wheel has turned relative to its previous position.

A single sensor cannot indicate the direction of rotation, but if a second emitter-sensor pair is added, 90 degrees out of phase with the first, a microcontroller can assess the phase difference between pulse streams to determine which way the wheel is turning. This principle is illustrated in Figure 7-6 where one transmitter-detector pair is located in position A and another is in position B. The resulting pulse trains, for clockwise and counterclockwise rotation of the wheel, are shown in green. The combination of pulse trains is known as *quadrature*, because there are four possible combinations: A and B both high, A and B both low, A high and B low, or A low and B high.

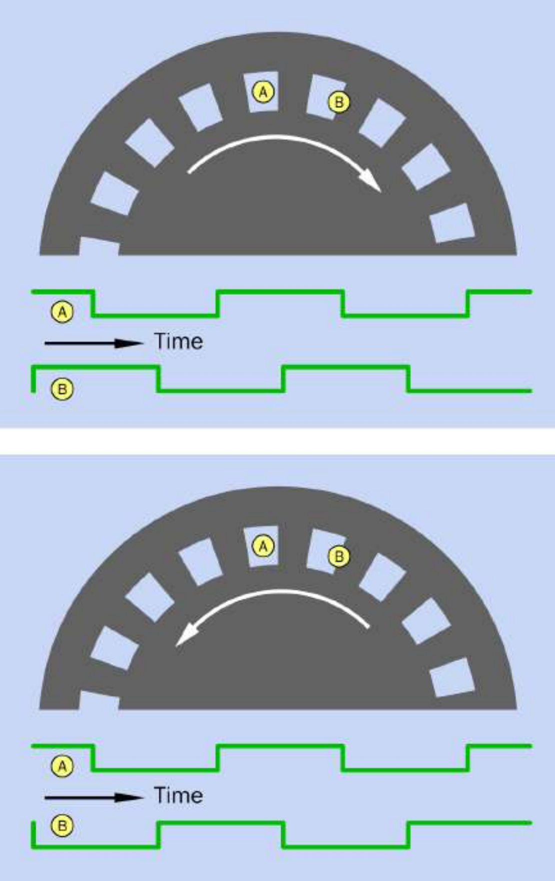

Figure 7-6 *With transmitter-detector pairs mounted at positions A and B, the phase difference between their pulse trains can show the direction of rotation of the wheel.*

This is the same principle as illustrated for a magnetic linear position sensor in Figure 6-5.

An alternative system of quadrature for an optical rotary encoder uses two separate tracks on the disc, each having an equal number of opaque and transparent sections, but half an interval out of phase.

Any system that reveals the relative motion of the wheel, but cannot determine its absolute angular position, is an *incremental sensor*.

Optical Products

High-end optical rotary encoders may use discs with as many as 600 sequential opaque and transparent segments to provide extremely high resolution. They are outside the scope of this Encyclopedia.

Moderately priced optical encoders are available as shaft-driven assemblies very similar in external appearance to potentiometers. Typically they have four terminals, one pair for power and ground connections and another pair for quadrature output from the two internal sensors, usually identified as A and B in datasheets. Some encoders also contain an on-off switch that is activated by pressing the shaft, in which case two additional terminals will be provided.

Bourns is a leading manufacturer of this type of encoder, an example being the EM14, which is mounted in a box-shaped body measuring 14mm square. It uses a 5VDC power supply and provides pulses of 4VDC minimum with intervals of 0.8VDC maximum. Variants are available with 8 to 64 pulses per revolution. Intended for audio applications, this type of encoder has a maximum rotation speed of 120rpm.

An example of an optical rotary encoder from a German manufacturer is shown in Figure 7-7. Its resolution is 25 pulses per rotation. The rectangular package measures approximately 19mm × 25mm. The power supply can be 3.3VDC or 5VDC.

Optical rotary encoders of this type cost about five times as much as mechanical rotary encoders at the time of writing, but their longevity and their clean output signals make them an attractive alternative, and the price difference may diminish over time.

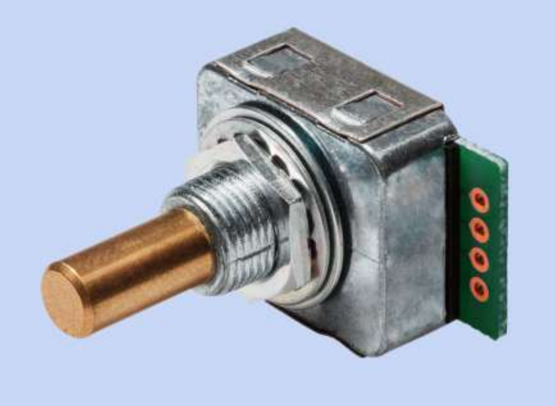

Figure 7-7 *A compact incremental optical rotary encoder in the MRB25 series from Megatron Elektronik AG & Co.*

Encoders such as the Avago HEDS-9 series are not protectively enclosed, and require assembly with a codewheel supplied by the manufacturer. See Figure 6-6 for additional information.

A more basic optical rotary encoder, shown in Figure 7-8, is sold by Cytron Technologies as an *optical switch* mounted on a small board, with a separate codewheel consisting of a slotted disc. This low-cost kit is intended for use in DIY robotics. More information about optical switches will be found in the entry dealing with **optical presence** sensors. See Figure 3-3.

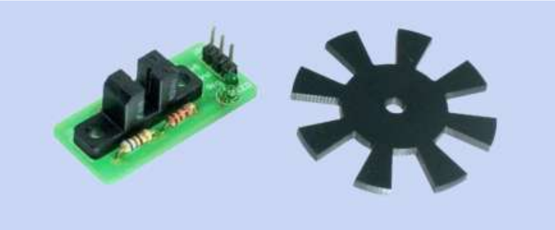

Figure 7-8 *A bare-bones optical encoder for DIY robotics.*

Computer Mouse Principles

The original design of a computer mouse, with a hard rubber ball, contained two optical rotary encoders oriented at right angles to each other. Each of them used a transmissive codewheel. The rolling ball turned the codewheels as the mouse was moved across a desktop, and electronics in the mouse converted the outputs from the encoders into a pulse train that could be interpreted by a computer. Figure 7-9 shows the primary components.

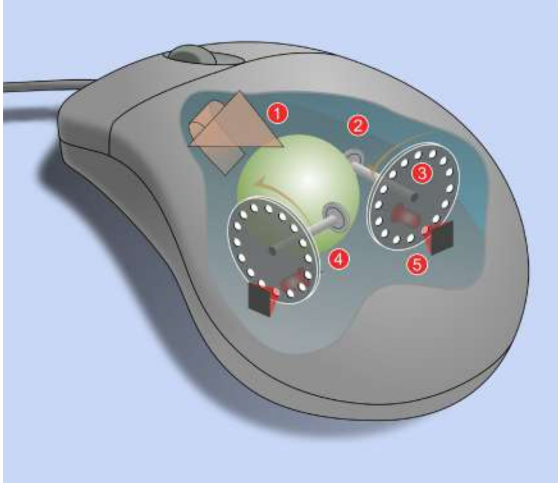

Figure 7-9 *From Wikimedia Commons, this rendering by Jeremy Kemp shows (1) rotation of the ball, (2) a roller touching the ball, (3) a transmissive optical codewheel, (4) an infrared LED that shines through a second codewheel, and (5) a sensor detecting the pulses of light.*

An *optical mouse* works on a different principle, maintaining a monochrome image of the desk surface on an optical array that functions like a very low-resolution camera sensor. Electronics in the mouse detect displacement of the image as the mouse is moved.

Rotational Speed

Incremental rotary encoders, which only supply relative data, are adequate in many applications, especially speed measurement, where a microcontroller can compare a pulse stream from a sensor with a target frequency, and can also provide feedback to control motor speed appropriately. This is convenient when a stepper motor is being used, or a DC motor that is controlled with pulse-width modulation. (See the entry on **motors** in Volume 1.)

Toothed wheels or magnetized wheels are commonly used to measure speed of rotation in applications ranging from automobile transmissions to computer disk drives.

While a second sensor can be added to determine the direction of rotation, this still does not provide information about the absolute orientation of a part.

Absolute Position

If associated electronics are equipped with nonvolatile memory, it may be used to store a sensor-wheel position from one session to the next. This may be sufficient in noncritical applications such as volume control in a car radio or stereo system.

Alternatively, an additional single window in an optical wheel can activate a *home sensor*. When the device is powered up, the wheel is turned until the home sensor is triggered, at which point the orientation of the wheel is known, and subsequent pulses from rotation sensor(s) will add or subtract angular information.

The pulse generated by a home sensor may be described in datasheets as a *reference signal* or *index signal*. In the early days of desktop computers, each 5.25-inch floppy diskette was perforated with an index hole for this purpose.

The Gray Code

For greater reliability in determining absolute position, an optical wheel can be divided into several concentric tracks, each of which contains a different coded sequence and is assigned its own light emitter and light sensor. The sensors are arranged in a radial line to scan the disc as it rotates. Since each detector will provide either a signal or no signal, output from the set of sensors can be combined to create a binary number.

Figure 7-10 shows binary codes from 0000 through 1111, corresponding with decimal numbers 0 through 15, where a white square is equivalent to a numeral 1 and a black square is equivalent to a numeral 0.

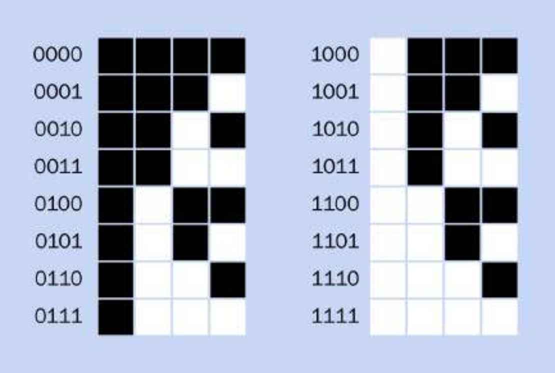

Figure 7-10 *Binary codes from 0000 through 1111, using white to represent 1 and black to represent 0.*

Figure 7-11 shows this system mapped onto a codewheel. The red circles indicate the locations of four stationary light detectors, which begin by providing a reading of 0000, since they coincide with four opaque areas of the wheel. Now if the wheel makes 1/16th of a full rotation in the direction of the arrow, the detectors will register 0001. If the wheel continues to rotate, the detectors will count in binary up to 1111 before the sequence repeats.

The problem with this design is that small manufacturing inaccuracies and other imperfections will result in some light detectors responding fractionally more quickly than others as the wheel rotates. This will occur in transitions where two or more adjacent segments of the wheel change between transparency and opacity. For instance, where 0011 is followed by 0100, transient values of 0010, 0001, 0111, 0110, or 0101 are possible. Although they will be brief, these values may trigger associated electronics.

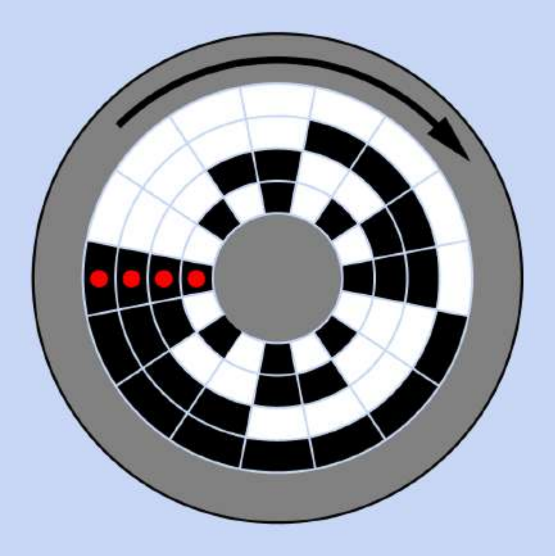

Figure 7-11 *The binary sequence mapped onto a code-wheel as areas of opacity and transparency. Red circles indicate light detectors.*

To eliminate this problem, a different code sequence can be used in which only one of the four sensors is allowed to make a transition from each value to the next. This is called a *Gray code*, and eliminates the issue of simultaneous transitions. A commonly used Gray code is shown in Figure 7-12.

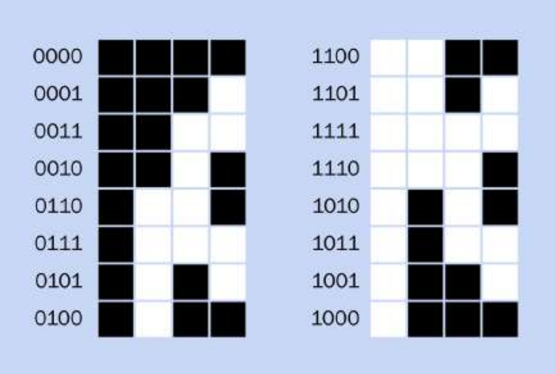

Figure 7-12 *A Gray code that allows only one binary digit to change in each transition from one value to the next.*

Magnetic Rotary Encoders

If a ferrous wheel is magnetically polarized in multiple domains, its rotation can be detected by a Hall-effect sensor in much the same way that a wheel divided into transparent and opaque sections can be assessed by light detectors and emitters. This is illustrated in Figure 7-13, where magenta and cyan bands indicate north and south magnetic poles.

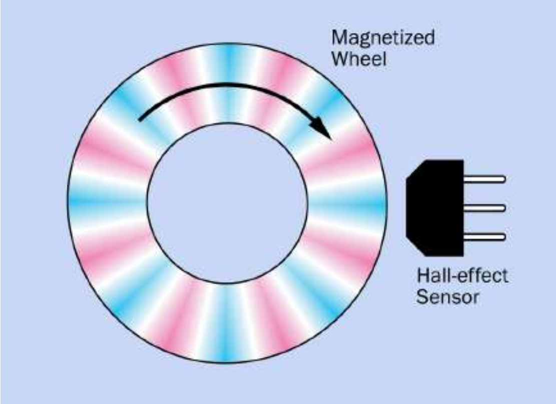

Figure 7-13 *A Hall-effect sensor can detect the rotation of a wheel that is divided into multiple north and south magnetic poles.*

See "Hall-Effect Sensor" for information about Hall-effect sensors.

An additional Hall sensor can be added, offset from the first, as in Figure 7-14. Once again the phase difference between the pulse trains can be used to determine the direction of rotation.

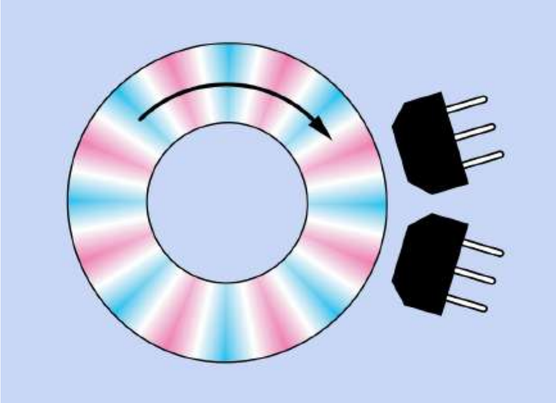

Figure 7-14 *Direction of rotation of the wheel can be deduced from the phase difference between the pulse trains from two sensors.*

Alternatively, a toothed wheel can be used, as suggested in Figure 7-15.

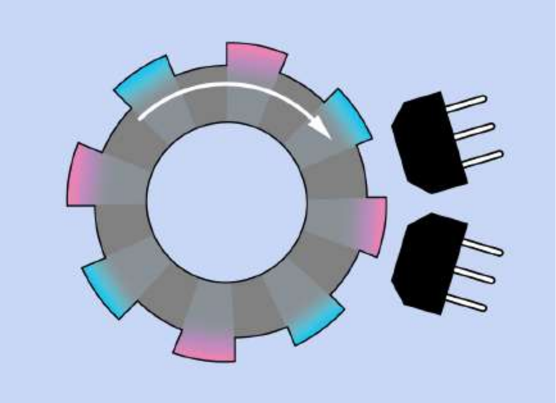

Figure 7-15 *A toothed wheel can trigger a Hall sensor if the teeth are magnetized.*

Another option is to mount a pair of magnets on a nonmagnetic wheel, with a Hall-effect sensor in the center. If it is a sensor with an analog output, the voltage will fluctuate smoothly between positive and negative, relative to the power supply for the sensor. Alternatively, a bipolar Hall-effect sensor can be used to provide a binary output. The concept is illustrated in Figure 7-16. An advantage of this configuration is that it provides approximate information about the absolute position of the wheel.

The two magnets will provide usable linear sensor outputs over a span of plus-or-minus 30 degrees of rotation, approximately. Additional magnets or sensors can produce a more complex output that would be decoded by a microcontroller.

How to Use It

An optical or rotary encoder is well suited for use with a microcontroller program that can count pulses, compare pulse trains, or interpret a Gray code. The microcontroller then takes appropriate action. For example, if a rotational encoder is used to control the gain of an audio amplifier, the microcontroller determines the direction and angle through which the encoder turns, and can respond by changing the value of a **digital potentiometer** (see Volume 1).

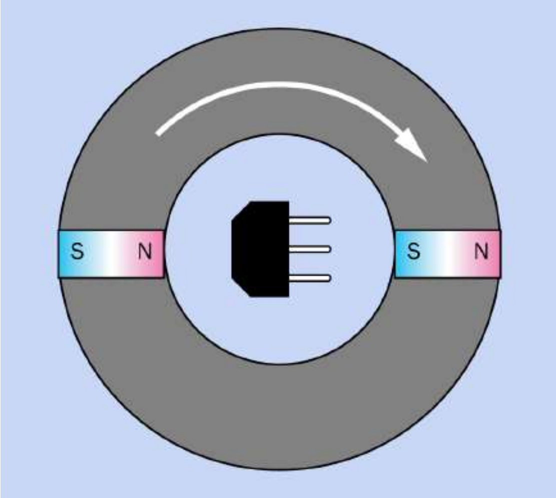

Figure 7-16 *A Hall-effect sensor can detect the angle of rotation of a ring on which two magnets are mounted as shown.*

Integrated circuit chips are available to convert a sequence of quadrature signals into an up-pulse or a down-pulse, thus eliminating the chore of achieving this with a microcontroller program. The LS7183 by LSI Computer Systems is an example.

What Can Go Wrong

Wiring Errors

If two sensors are used to detect the direction of rotation of a rotary encoder, and outputs from the sensors are accidentally swapped, the component will seem to work normally except that the indicated direction will be inverted.

Coding Errors

Microcontroller code that interprets quadrature signals must be fast enough to keep up with the pulse streams. If a microcontroller is performing other tasks, an interrupt may be necessary for processing the rotational data. This should not be a significant problem when interpreting human input from a knob or dial, but for a motor-driven encoder, pulse counting in hardware may be a better alternative.

Ambiguous Terminology

Rotational encoders are often referred to simply as "encoders," because they are more common than linear encoders. When searching for an encoder, check each datasheet carefully to determine which type you are dealing with.

tilt sensor

8

A *tilt switch* is defined here as an electromechanical switch, while a *tilt sensor* uses electronics. Both types are included in this entry.

A *tipover switch* is very similar to a tilt switch and uses the same principles. Therefore, it is included in this entry.

Some manufacturers' datasheets refer to tilt sensors as *tip sensors*. Some Asian supply catalogues refer to tilt switches as *breakover switches*.

An **accelerometer** can measure the angle at which it is held relative to the downward force of gravity, but it has additional capabilities. Therefore, it has its own separate entry in this Encyclopedia.

An *inclinometer* measures the incline, or positive slope, from an observation point to the top of an object such as a building or tree. The height of the object can be calculated from the angle. A *clinometer* can additionally measure a decline, or negative slope. These measurement devices are fully featured products as opposed to sensors, and are not included in this Encyclopedia.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **vibration** sensor (see Chapter 11)

What It Does

Three principal types of tilt sensor exist.

1. Single axis, single output. The sensor responds to being tilted around one horizontal axis, relative to the downward force of gravity.
2. Dual axis, dual output. The sensor contains two sensing elements at 90 degrees to each other. Each has an output determined by its angle of tilt from vertical around one axis.
3. Dual axis, single output. A single sensor responds to an angle of tilt from vertical around any horizontal axis.

A *tilt switch* is usually of the third type, and is defined here as containing an electromechanical or electronic switch that opens or closes a connection. Most tilt switches are SPST (normally open) or SPST (normally closed). A minority are DPDT.

A *tipover switch* is a type of high-current tilt switch that cuts power to a device such as an electric heater when it is tipped over.

This Enclopedia defines a *tilt sensor* as being an electronic component, as opposed to an electromechanical component. The distinction is often observed in datasheets, but not always.

Schematic Symbol

No specific schematic symbol is generally used for any variant of a tilt sensor. It can be represented with an annotated switch symbol.

How It Works

Because a tilt switch is a simpler device than a tilt sensor, it will be described first.

The most common type of tilt switch consists of a cylindrical metal or plastic enclosure, often measuring about 5mm by 15mm, containing two spherical steel balls that may be nickel-plated or gold-plated. When the switch is tilted, the balls eventually run downhill, and the lower ball completes an electrical connection between two contacts or between a single contact and the metal enclosure of the switch. The second ball is included to add weight and suppress vibration in the first.

Figure 8-1 shows a switch manufactured by Comus Global, rated for 0.25A at 60VAC or 60VDC, maximum. The body of the switch measures approximately 10mm x 5mm, and the switch is activated when tilting -10 degrees from horizontal. It is deactivated when tilting +10 degrees. A scale drawing of the interior is shown in Figure 8-2.

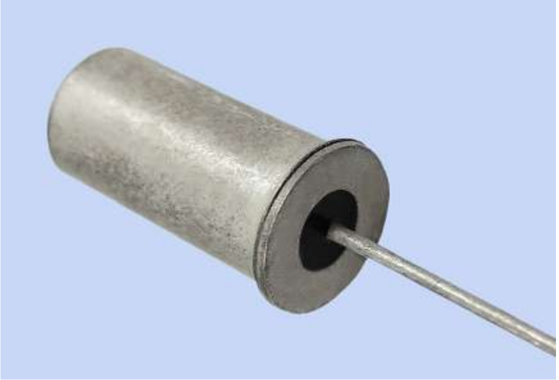

Figure 8-1 *The CW1300 tilt switch manufactured by Comus Global.*

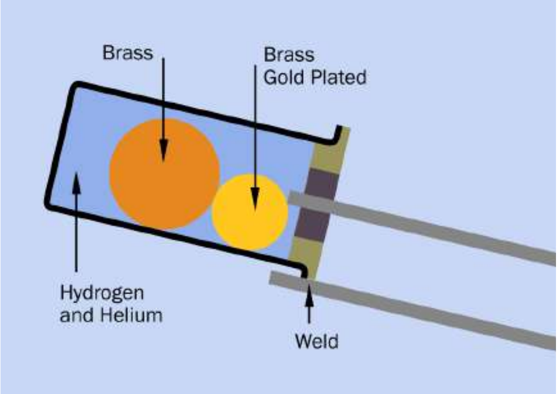

Figure 8-2 *Interior of the CW1300 tilt switch, from a scale drawing supplied by the manufacturer. The lower lead is welded to the shell of the sensor. The leads may be inserted in a circuit board.*

Figure 8-3 shows three common internal configurations of a generic tilt switch. The top version has axial leads and uses the metal shell of the switch to complete the circuit. The center version has radial leads, with a plastic shell. The bottom version has radial leads, one of which is attached to the metal shell to complete the circuit.

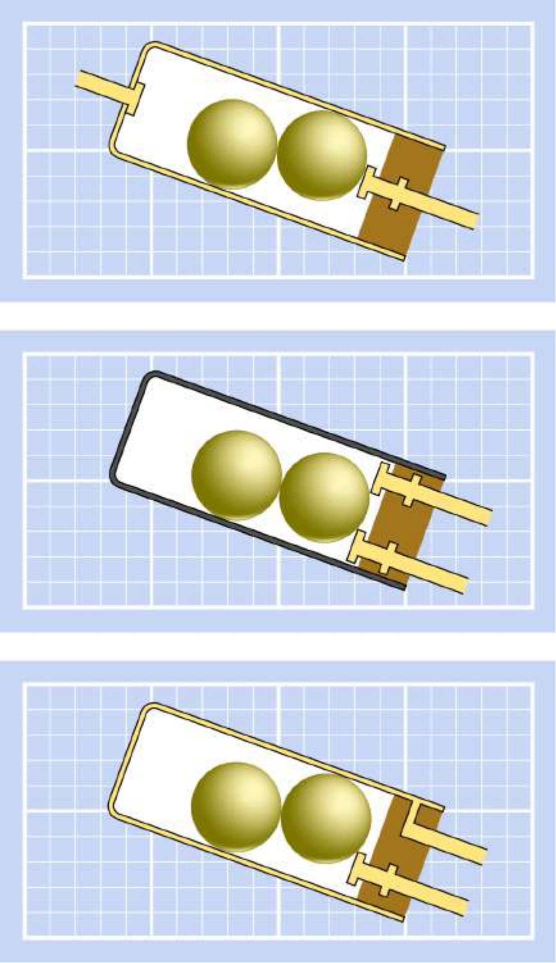

Figure 8-3 *Three variants of a generic ball-operated tilt switch. The graph-paper scale is in millimeters.*

Figure 8-4 shows the parts of a disassembled tilt switch.

Simplified Version

Where size is not a significant consideration, a tilt switch can be created by attaching a pivoting, weighted arm to a small snap-action switch.

Applications

An old-style (nonelectronic) thermostat may contain a tilt switch attached to the end of a bimetallic strip coiled into a spiral. When the strip bends in response to a drop in temperature, the switch closes its contacts, activating a relay that starts a heating unit. If the temperature rises, an additional set of contacts in the same relay may activate an air-conditioning unit. In old thermostats, the tilt switch may contain mercury in a glass tube, which should be handled with caution.

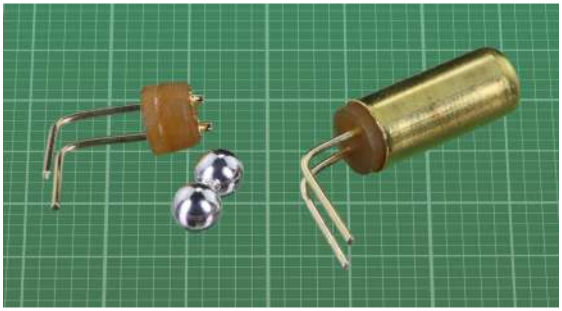

Figure 8-4 *At right, a tilt switch. At left, the cap removed, and the two balls that make internal contact. The background grid is in millimeters.*

A tilt switch may detect the opening of a door or window in a simple alarm system.

Tilt switches have been used in automobiles to switch on the interior light in the trunk when its lid is opened.

A normally closed tilt switch is often used to stop the inflow of granular material to a bin when it is almost full. This is colloquially known as a *bin switch*. In industrial applications of this kind, the switch is activated by a long lever that has a ball mounted at the end. The switch assembly is physically large. See Figure 8-5.

A normally open tilt switch may operate a valve or start a pump when the liquid in a tank drops below a certain point. If the switch uses a float to sense the level of the liquid, it is often known as a *float switch*. This is described in the entry on liquid level sensors. See Chapter 15.

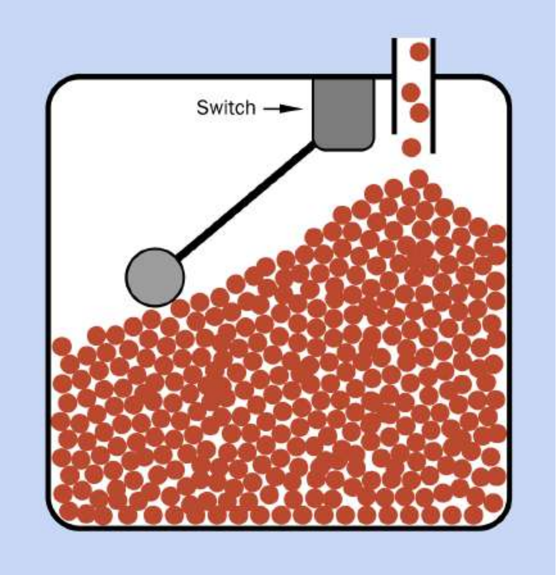

Figure 8-5 *The flow of granular material into a bin can be sensed and stopped by a tilt switch of this type, often known as a bin switch.*

A *tipover switch* may use the simplified system of a weighted arm that activates a snap-action switch. When used in conjunction with a room heater, the switch must handle substantial current.

A motorcycle may contain a tipover switch to stop the electric fuel pump if the motorcycle falls on its side.

Four tilt switches placed in a cross-shaped pattern on a flexible mount can be used as the basis of a very simple game controller, with a joystick mounted in the center.

Variants

The three configurations of ball-type tilt switches shown in Figure 8-3 are functionally interchangeable and can be chosen for convenience of their leads in fitting the circuit.

Mercury Switches

Early tilt switches contained a blob of mercury in a glass bulb. When the bulb was tilted, the mercury rolled to the end and made an electric connection between two metal contacts that penetrated the bulb.

A small mercury switch is shown in Figure 8-6. This type of sensor became less common after many countries classified mercury as an environmental hazard and established regulations restricting its use.

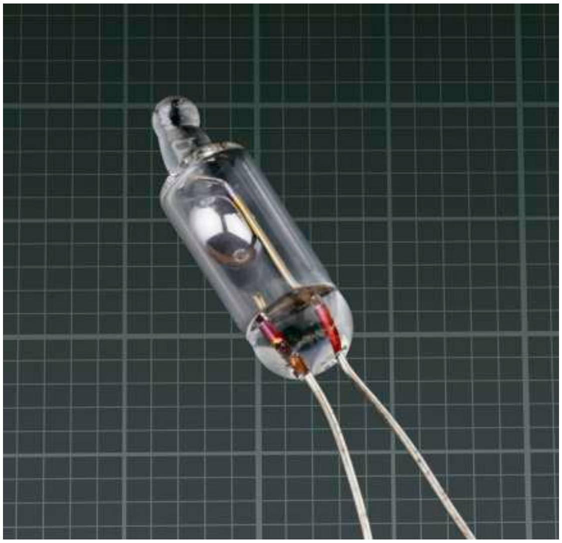

Figure 8-6 *Small mercury switch rated for 0.3A at 24VDC or 24VAC. Larger mercury switches can switch more power; 1A at 230V is common. The background grid is in millimeters.*

Mercury is an excellent electrical conductor. It remains in a liquid state between about -38 degrees Celsius and +356 degrees Celsius, and has very high surface tension, encouraging it to form a single blob instead of breaking up into small droplets. Because free space in the bulb is filled with an inert gas to prevent oxidation of the electrodes, a mercury switch can have a very long operating life. In the United States in the 1970s, some light switches were sold containing mercury switches with a claimed life expectancy of 100 years.

Pendulum Switch

This type of switch, now relatively rare, was found in vintage pinball machines. It consists of a pendulum about 5cm long, suspended inside a steel ring about 1cm in internal diameter. If the machine was rocked sufficiently to bring the pendulum in contact with the ring, the game was cancelled and the word "Tilt" appeared on the display. Consequently this was referred to as a tilt switch, although really it was a form of **vibration** sensor with a long period of oscillation.

Magnetization

Some tilt switches use a steel ball that is weakly magnetized, so that it will seat itself more firmly when it rolls into a circular depression or ring. This type of switch must be tilted back through a larger angle to dislodge the ball. Therefore it will exhibit greater *hysteresis*.

Tilt Sensors

Unlike a tilt switch, a tilt sensor is not built around an electromechanical switch.

The principle of a rolling ball has been miniaturized and encapsulated in a small enclosure (10mm square or smaller), in which the ball rolls to interrupt a beam from an internal LED shining on a phototransistor. Examples are found in the Panasonic AHF series. Internal circuitry ensures a clean on-off signal, free from the switch bounce that is a problem in basic ball-type tilt switches. However, the switch requires a power supply, and the open-collector output must be used with a pullup resistor. By comparison, a simple electromechanical tilt switch can be wired directly to the device that it controls.

Diagrams from the Panasonic datasheet show the three types of AHF sensor available for vertical, horizontal, and reverse mounting. In each case, the ball (dotted circle) rests in a shallow cup (dotted curve) where it obstructs the beam from an internal LED (not shown). See Figure 8-7. An exterior view of the AHF22 is shown in Figure 8-8.

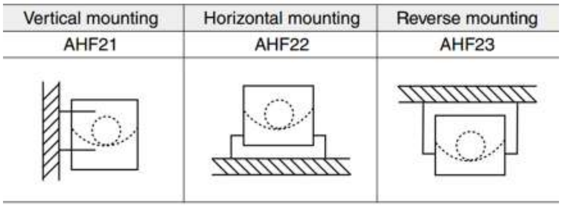

Figure 8-7 *Three types of Panasonic tilt sensor, from the manufacturer's datasheet. See text for details.*

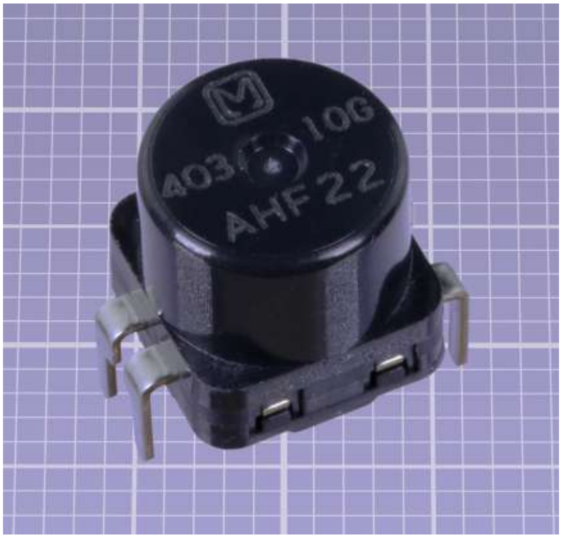

Figure 8-8 *Exterior view of the Panasonic AHF22 tilt sensor. The background grid is in millimeters.*

Two-Axis Tilt Sensors

The Rohm RPI-1035 is a surface-mount tilt sensor about 4mm square, with two phototransistor outputs that indicate which axis the sensor is tilting around. The outputs can be interpreted as a 2-bit binary number, with its four possible states indicating the rotation of the switch around two axes at 90 degrees to each other. Switches of this type were developed to indicate the orientation of consumer-electronic devices such as digital cameras, but more sophisticated sensors containing accelerometers are becoming price-competitive.

Surface-mount 2-axis tilt sensors have been made available on small breakout boards that are easy to use experimentally. An example is the Parallax 28036 shown in Figure 8-9.

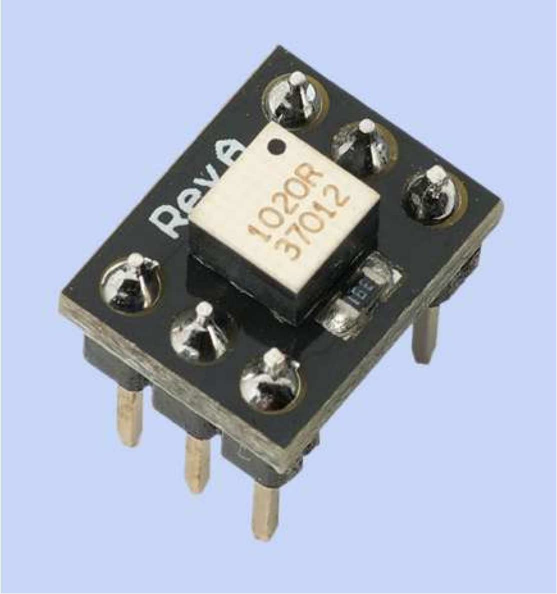

Figure 8-9 *A 2-axis tilt sensor mounted on a breakout board available from Parallax.*

The behavior of the rolling-ball sensor at the heart of this board is shown in Figure 8-10. The sensor contains a square cavity in which the ball is depicted as a blue circle. At one corner of the cavity is a red LED, while two phototransistors, labeled A and B, are at the corners on the left and right. When the sensor is held as in section 1 of the figure, with the LED at the top and the ball resting at the bottom, both phototransistors have a high output as they receive light from the LED.

In section 2 of Figure 8-10, the sensor has been turned through 90 degrees. The ball now prevents light from reaching sensor B, while phototransistor A is still active. In section 3 of the figure, the sensor has been turned through another 90 degrees, so that the ball prevents light from escaping from the LED, and both phototransistors are now dark. In section 4, the ball obstructs phototransistor A but not phototransistor B.

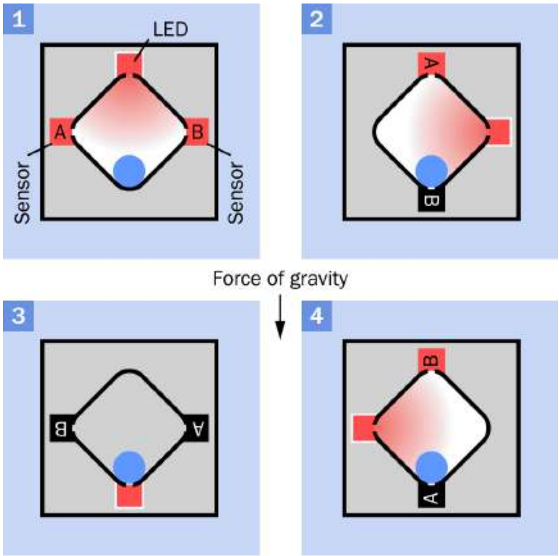

Figure 8-10 *The rolling-ball sensor inside the Parallax 28036. See text for details.*

Suppose that the sensor is placed flat on a horizontal surface. The sensor will now respond to being tipped either way around two horizontal axes. This justifies its description as a 4-directional tilt sensor, although its design suggests that it may have been intended for use as described above, rotating through four positions around one horizontal axis.

Values

A heavy-duty tilt switch can be rated as highly as 10A at 240VAC. More commonly, a tilt switch about 15mm long can be expected to switch about 0.3A at 24VAC or 24VDC.

The *operating angle* is the angle through which the switch must be turned to activate it, relative to its normal rest position.

The *return angle* is the angle to which the switch must be returned to deactivate it. Hysteresis results from the return angle being smaller than the angle that activates the sensor.

Tilt sensors with an open-collector output will specify maximum forward current for the internally mounted LED (usually no greater than

50mA), the maximum collector-emitter voltage at the output (typically 30V), and the maximum collector current (often 30mA). An explanation of open-collector outputs is given in Appendix A (see "3. Analog: Open Collector").

How to Use It

An electromechanical tilt switch can be connected directly between a power supply and a device, so long as the device does not draw more current than the switch is rated to handle. Note that inductive loads such as motors draw an initial surge that can be at least twice the operational rating, while relays may be more likely to create a voltage spike when disconnecting. Switches should be chosen accordingly. For more discussion of this topic, see **switch** in Volume 1.

A small tilt switch can be used in conjunction with a relay or transistor to amplify its signal sufficiently to drive a larger load.

If an electromechanical tilt switch is connected to an electronic device such as a microcontroller or logic chip, output from the switch will have to be *debounced* to prevent a series of brief voltage spikes that can cause false triggering when the switch turns on or off. A debouncing logic circuit or chip can be used, or the program code in a microcontroller can introduce a wait period of up to 50 milliseconds to allow the contacts to settle.

Mercury switches are much less likely to create a noisy output than rolling-ball switches, and may require little or no debouncing.

For an application that must sense rotation around two or three axes, multiple single-axis tilt switches can be combined. A microcontroller or logic gates will be necessary to evaluate signals from the switches, to determine the orientation.

What Can Go Wrong

Contact Erosion

If a ball-type tilt switch is subjected to current that exceeds its specification, arcing may erode its contacts, and they will become less reliable, especially if the contacts are plated with a thin metallic film that is eroded. For additional information on arcing in switches, see the **switch** entry in Volume 1.

Random Signals

During the brief time when a ball-type tilt switch is turning from one position to the other, vibration of the ball(s) inside it is likely to create erratic, random signals. If the output from the switch is being evaluated by a microcontroller, a debouncing routine may be insufficient to prevent the random signals from being sensed, and some programming will be necessary to ignore the signals during this transitional phase. If the switch is connected directly to a relay, the intermittent signals may occur sufficiently rapidly that the relay will ignore them.

Environmental Hazard

A device that incorporates a mercury switch may have to be re-engineered in the future if the availability of mercury switches becomes unreliable as a result of tighter environmental regulations. For the same reason, the end user may have difficulty replacing a mercury switch if it fails. Therefore, a ball-type tilt switch should be used instead of a mercury switch in any newly designed device.

Requirement for Gravity

Because a tilt switch depends on gravity to roll a ball or move a blob of mercury, it will not work in low-gravity, reversed-gravity, or zero-gravity conditions—for example, in a rocket during the unpowered phase of ascent and descent, or in an aircraft that performs aerobatic maneuvers. Performance of a tilt switch in a vehicle that accelerates or decelerates suddenly may also be unreliable. Likewise, it cannot be used on a small boat.

Requirement for Stability

A tilt switch will tend to give erroneous results in a location where there is significant vibration or where the object containing the switch may be turned or repositioned unpredictably by the user.

gyroscope

Historically, a **gyroscope** always contained a spinning disc. Some devices for navigation still depend on rotating elements, but they are outside the scope of this Encyclopedia. This entry deals primarily with *vibrating gyroscopes*, also known as *resonator gyroscopes*, that are MEMS devices contained within silicon chips.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **GPS** (see Chapter 1)
- **magnetometer** (see Chapter 2)

What It Does

A **gyroscope** resists rotation around any axis at right angles to its own axis of rotation or vibration. Consequently, if the gyroscope is allowed to move freely on gimbals in a sealed enclosure, the gyroscope will tend to maintain its orientation while the enclosure can rotate freely around it.

Taking this concept a step further, if the enclosure is mounted in an aircraft, the aircraft's rotation around two axes can be determined by referring to the gyroscope. If additional gyroscopes are added orthogonally to the first, the aircraft's rotation around all three axes can be determined.

A gyroscope does not measure linear motion in any direction, or any static angle of orientation.

Schematic Symbol

A chip-based gyroscope, magnetomer, or accelerometer may be represented in a schematic as a rectangular box containing abbreviations to identify pin functions (as in any integrated circuit chip).

IMU

An **accelerometer** measures variations in linear motion and will also measure its own static orientation relative to the force of gravity. If an accelerometer rotates around its own axis, it will not measure angular velocity.

A **magnetometer** measures the magnetic field surrounding it, and may be sufficiently sensitive to determine its orientation relative to the Earth's magnetic field.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an *IMU* (inertial measurement unit), which can provide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Applications

The first chip-based gyroscope was used in automobiles in 1998 as a yaw sensor in a skid-control system. Subsequent automotive applications include active suspension control, air bag sensors, rollover detection and prevention, and navigation systems.

Gyroscopes may be installed in military ordnance to provide backup in case an onboard **GPS** system fails, possibly as a result of radio jamming.

Handheld 3D game controllers and headsets may use gyroscopes to control images displayed to the viewer. A digital camera may employ a gyroscope to provide image stabilization. Gyroscopes are usually found in quadcopters or drones, are used to stabilize two-wheeled vehicles such as the Segway, and are used in robotics.

How It Works

The traditional form of gyroscope is a rotating wheel, which will resist turning forces perpendicular to its own axis of rotation. In Figure 9-1, three directions at right angles to each other are defined in the bottom-right corner of the diagram as X, Y, and Z. The wheel is rotating around the X axis, as shown by the green arrow. It will resist any turning force around the Y axis (red arrows) or the Z axis (yellow arrows).

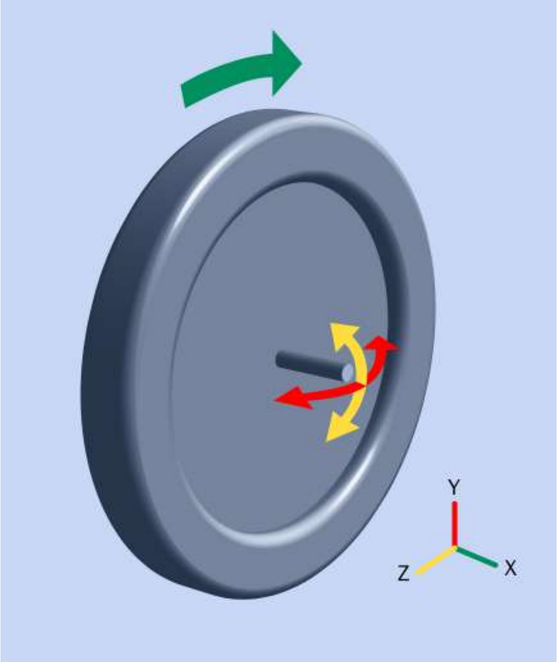

Figure 9-1 *In a traditional gyroscope, a wheel that is rotating (shown by the green arrow) will resist a turning force on either of the axes (shown by red and yellow arrows) perpendicular to its axis of rotation.*

Vibrating Gyroscope

A vibrating fork can be substituted for a wheel. In Figure 9-2 a fork is secured at its base while its tines are induced to vibrate toward each other and away from each other, as suggested by the double-ended arrow. In a chip-based gyroscope, this vibration is induced piezoelectrically or by static electricity.

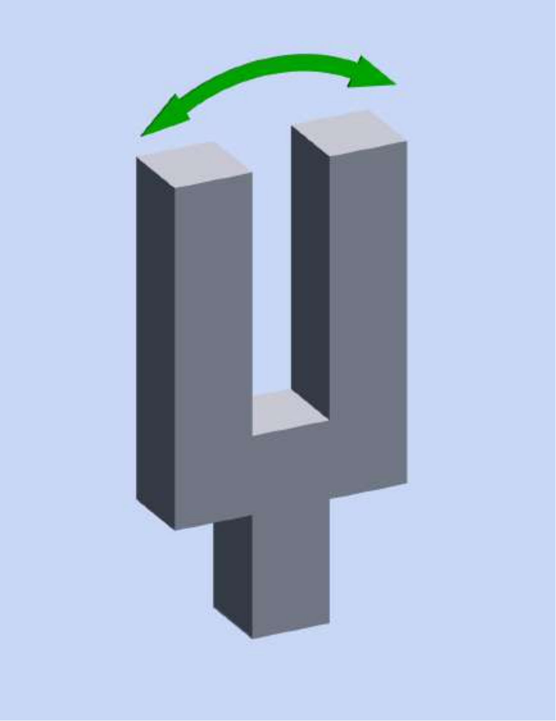

Figure 9-2 *A vibrating fork (green arrows) can be substituted for a rotating wheel in a gyroscope.*

Now suppose a turning force is applied around the vertical axis at the base of the fork, as suggested by the lower arrow in Figure 9-3.

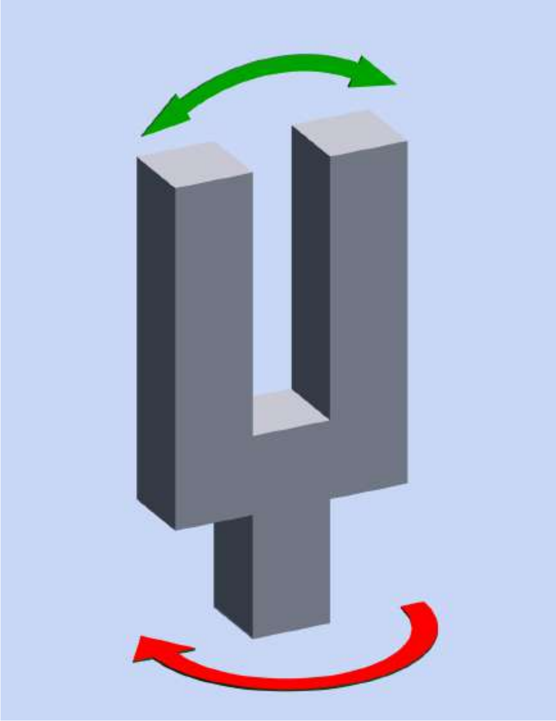

Figure 9-3 *A turning force is applied to the base of the fork around a vertical axis, as shown by the lower arrow.*

The angular momentum of the vibrating tines causes them to resist this turning force, and consequently they will tend to bend, as shown by the yellow arrows in Figure 9-4. The amount of their deflection can be measured capacitively. This system is used in many chip-based gyroscope systems, and may be referred to as a *vibrating gyroscope* or a *resonator gyroscope*.

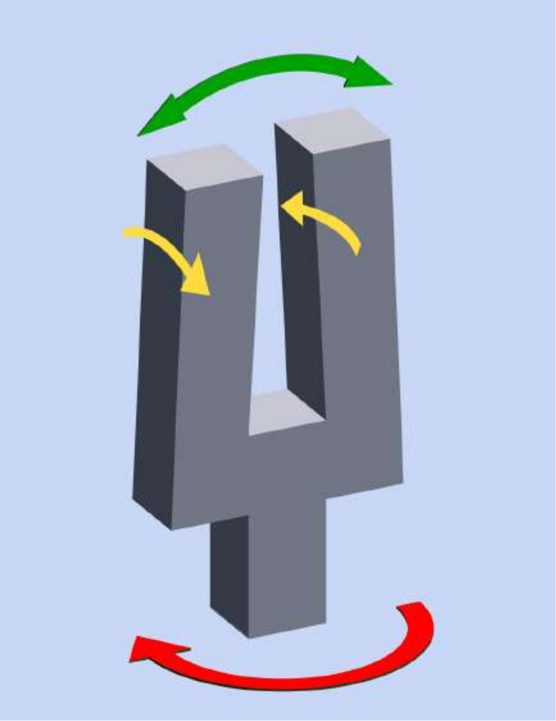

Figure 9-4 *The angular velocity of the rotating fork causes deflection of its vibrating tines, shown by the yellow arrows.*

An assembly of microscopic forks can be etched into a silicon chip. In Figure 9-5, an electron micrograph shows the interior of this type of chip. It contains three rotational sensors, responding to the X, Y, and Z axes of motion. These sensors respond to *pitch* (rotation around the X axis), *roll* (rotation around the Y axis), and *yaw* (rotation around the Z axis).

Fork-based sensors are analog devices whose values are converted to digital values by an onboard *analog-to-digital converter* (ADC). The values are stored in registers that are available to other devices, often via the I2C protocol, which is widely used by microcontrollers.

For additional details about I2C, see Appendix A.

Typically there will be two 8-bit registers for each axis. Each register stores the binary equivalent of a signed integer, where the positive or negative value represents the direction and magnitude of deflection, usually in degrees per second (dps).

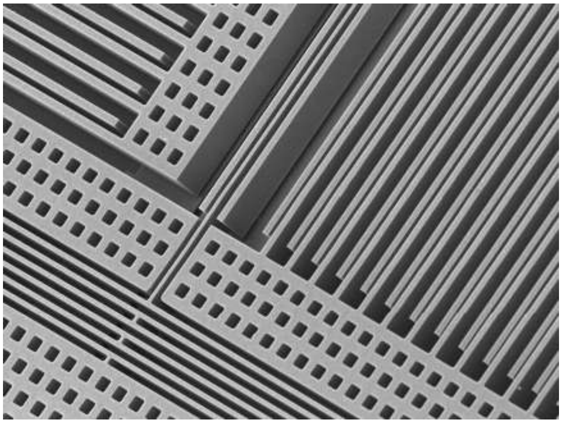

Figure 9-5 *Electron micrograph of the STMicroelectronics LIS331DLH vibrating gyroscope installed in the Apple iPhone 4. The parallel plates at bottom-left function as a spring, while the elements at top-left and at right measure capacitance as their orientation varies according to rotational velocity. Photo courtesy of MEMS Journal published by Chipworks*

Variants

The L3G420D by STMicroelectronics is a 3-axis gyroscope-only chip. It communicates via the SPI or I2C protocol, is approximately 4mm square, and can measure rotational rates up to plus-or-minus 2,000 degrees per second.

The Freescale FXAS21002C has a similar specification. Gyroscope-only chips of this type have fallen in price to the point where they are comparable to the retail cost of everyday components such as a small-signal relay or an audio amplifier on a chip.

IMUs

Chips that only contain gyroscopes are becoming less common as the cost of adding accelerometers decreases.

Gyroscopes and accelerometers are complementary, as gyroscopes are insensitive to linear motion or the Earth's gravity, but accelerometers can measure the rate of change of linear motion and orientation of the chip relative to the Earth. Software can combine this data to calculate the shape of the path being described by a device containing the chip, in addition to the changing velocity of the chip along that path.

The InvenSense MPU-6050 is a common 3-gyroscope, 3-accelerometer chip. It also includes an interface for connecting an external 3-axis magnetometer. The SPI and I2C communications protocols are supported. The MPU-6050 has been a popular choice in the hobby-electronics community, so that Arduino-compatible code to interpret its data is available from many sources. Breakout boards are available with the MPU-6050 installed. An example is the Sparkfun SEN-11028 shown in Figure 9-6.

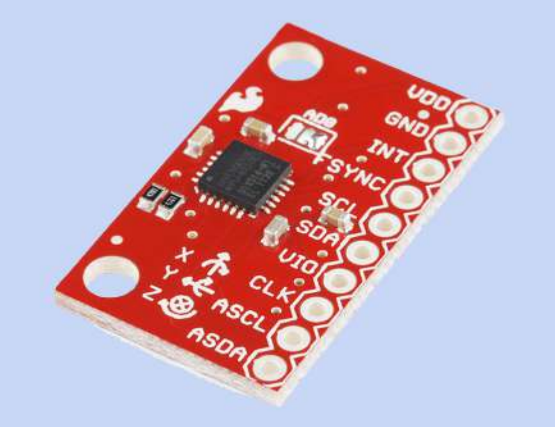

Figure 9-6 *A breakout board from Sparkfun, providing easy access to the InvenSense MPU-6050 chip combining three gyroscopes and three accelerometers.*

Values

The *rotational velocity* of a gyroscope element is usually expressed in degrees of rotation per second (dps), and sometimes in rotations per minute (RPM).

A datasheet will specify the number of sensor axes (usually 3), supply voltage (3.3VDC is common), maximum digital-low and minimum digital-high output voltages, and power consumption in normal mode and sleep mode. Power consumption is usually less than 10mA.

The *dynamic range* is the maximum forward and reverse rotational velocity, which usually will not exceed plus-or-minus 2,000 degrees per second. Lower ranges may be user-selectable. The advantage of selecting a lower maximum rate of change is that it can be converted to a digital value with higher precision.

The *sensor resonant frequency* will be several kilohertz, and must be higher than the frequency of any vibration that is applied to the sensor during use.

The *communications protocol* is usually I2C and SPI may be offered as an option, with a selectable digital output data rate.

Bias temperature coefficient describes the effect of temperature on the gyroscope.

The *resolution* of the gyroscope relates to the number of bits used in the digital output from the onboard ADC. A 16-bit resolution is common.

How to Use It

Using a smart chip such as the MPU-6050, the circuit designer can take advantage of its onboard digital motion processor (DMP). Still, obtaining orientation information from the contents of registers on the MPU-6050 is nontrivial. Online sources and code libraries are necessary. The book *Make: Sensors* contains code listings for the Raspberry Pi as well as the Arduino.

What Can Go Wrong

Temperature Drift

Vibrating materials at the heart of a chip-based gyroscope are likely to change their behavior with temperature. Typically the chip will contain a temperature sensor, the value of which can be used to adjust the output value of the gyroscope.

Mechanical Stress

Stress can be induced when a surface-mount chip is soldered to a board. The vibrating parts of a chip-based gyroscope may be adversely affected. Datasheets will supply information regarding maximum acceptable temperature during the soldering process.

Vibration

Because a chip-based gyroscope depends on the consistent behavior of internal vibrating parts, external vibration can degrade its accuracy. Sensor design can minimize the effects of vibration, but the datasheet should be consulted for details.

Placement

A gyroscope should be placed on a circuit board near a hard mounting point where deflection of the board will be minimized.

accelerometer

10

OTHER RELATED COMPONENTS

- **GPS** (see Chapter 1)
- **gyroscope** (see Chapter 9)
- **tilt** sensor (see Chapter 8)
- **vibration** sensor (see Chapter 11)

What It Does

Acceleration is the rate at which velocity changes over time. If a car takes 10 seconds to increase its speed from 30kph to 40kph relative to the road on which it is traveling, it is accelerating at an average of 1kph each second. If it then reduces its speed back to 30kph during another 10-second interval, it is decelerating at the same rate—although deceleration is really just negative acceleration.

While a car is accelerating, passengers will feel a lateral force exerted on them. Similarly, astronauts in a rocket that blasts off will feel a downward force. According to Einstein's theory of equivalence, forces resulting from acceleration are indistinguishable from the force of gravity.

Consequently, a sensor that measures acceleration can also measure gravity. This sensor is an **accelerometer**. Its output may be measured in *gravities*, abbreviated with the letter g (not to be confused with the usage of G to measure the strength of a magnetic field in gauss).

If three accelerometers are mounted orthogonally (at 90 degrees to each other), their readings can reveal:

- The direction of acceleration of a moving object.
- If an object has been dropped and is falling freely.
- Which way up it is being held in a stationary position.
- The severity of an impact when a moving object collides with some other object.

IMU

A **gyroscope** measures the rate of rotation of the enclosure in which it is mounted. This is properly known as the *angular velocity*. A gyroscope will also respond to changes in the rate of rotation. It does not measure linear motion or a static angle of orientation.

A **magnetometer** measures the magnetic field surrounding it, and may be sufficiently sensitive to determine its orientation relative to the Earth's magnetic field.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an *IMU* (inertial measurement unit), which can provide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Schematic Symbols

A chip-based accelerometer may be represented in a schematic as a rectangular box containing abbreviations to identify pin functions, as in any integrated circuit chip. No specific symbol is used for any of these components.

Applications

In the past, accelerometers were laboratory devices that calibrated the performance of cars, airplanes, and other types of vehicles. Measuring the ability of car tires to withstand cornering forces was an application where an accelerometer was used.

Miniaturization of accelerometer elements, coupled with a radical reduction in their cost, has enabled them to be installed in small electronic devices ranging from smartphones to hard drives.

In a phone or a camera, accelerometers can determine which way up the user is holding the device. The camera can rotate the picture appropriately, and the orientation of the picture can be saved with its image data.

In an external hard drive containing rotating platters, accelerometers can protect the read-write heads by rapidly parking them during the fraction of the second that elapses between someone dropping the hard drive and its impact with the floor.

Accelerometers can be installed in a 3D mouse or virtual-reality headset to determine its orientation and motion. This enables a video image to be updated appropriately. For example, the Nintendo Wii Remote has been marketed with an ADXL330 accelerometer.

In an automobile, an accelerometer can trigger the deployment of an air bag when the deceleration caused by an impact exceeds a threshold level.

How It Works

The simplest conceptual model of an accelerometer consists of a mass attached to one end of a coiled compression spring. The other end of the spring is anchored in an object whose acceleration is being measured. The mass can only move along the same axis as the spring.

Figure 10-1 shows three views of a simplified accelerometer, which is a sealed tube shown in dark red. In the center image, the accelerometer is in its rest state. The top image shows the mass (a dark blue square) responding when the tube accelerates from left to right. The third image shows it decelerating (that is, undergoing negative acceleration, or acceleration from right to left). Using an ideal spring, the displacement of the mass will be proportional to the rate of acceleration, within reasonable limits. The displacement can be measured optically or capacitively.

Note that the rest state will resume when the acceleration stops, regardless of constant motion in any direction. An accelerometer only measures a *change* in velocity. It does not measure a constant velocity.

An accelerometer cannot measure rotation around its own axis of movement. Therefore, it may be used in conjunction with a **gyroscope**, which measures angular velocity.

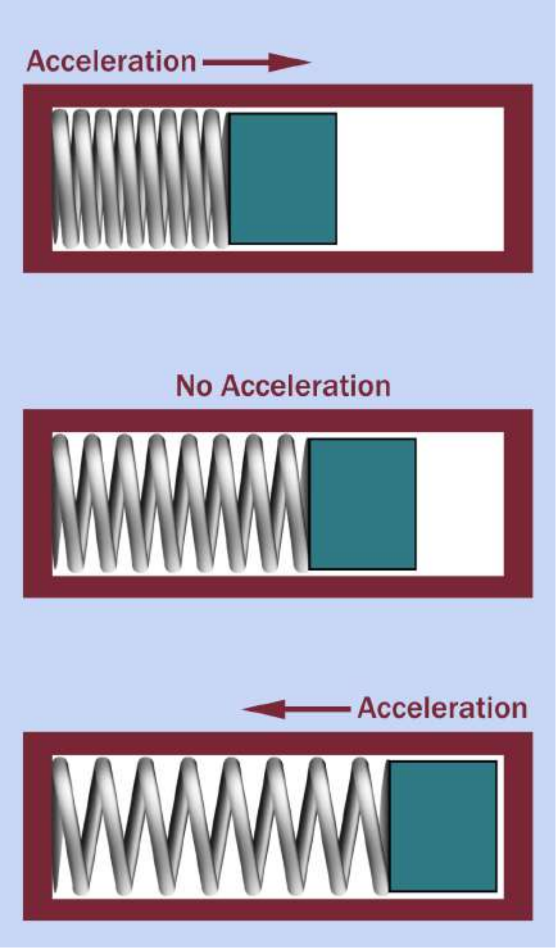

Figure 10-1 *Simplified view of an accelerometer consisting of a sealed tube (dark red outline) with a spring anchored to it at left. A small mass (dark blue) is attached to the right end of the spring. The mass responds to acceleration of the tube.*

Gravity and Free Fall

If a device containing the simplified accelerometer described previously rests on the ground, and is mounted vertically as in the left section of Figure 10-2, gravity acting on the mass will apply a force to one end of the spring, while the other end is restrained. The accelerometer will now measure 1g as a downward force.

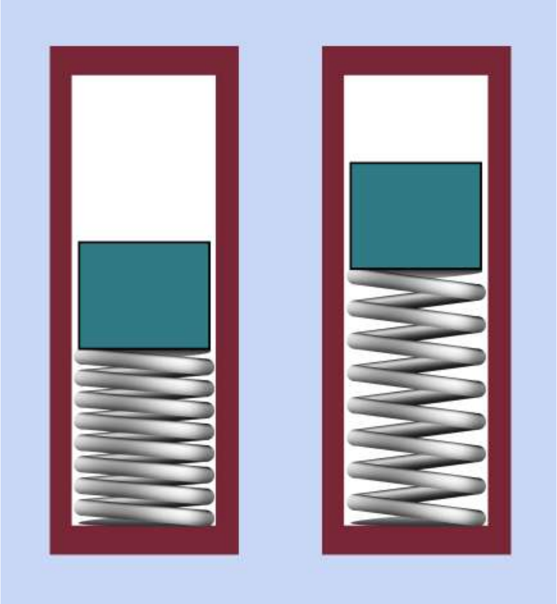

Figure 10-2 *Left: an accelerometer resting on the ground measures 1g, as the force of gravity pulls the mass downward. Right: in free fall, the accelerometer measures 0g.*

If the device is dropped so that it falls freely under the influence of gravity, it is in *free fall* and will accelerate at approximately 9.8 meters per second each second. This is usually written as 9.8m/sec^2, or can be described as 1 gravity, often expressed as 1g.

An accelerometer in free fall will measure 0g, as shown on the right in Figure 10-2, because all parts of the accelerometer are now accelerating equally under the force of gravity.

If three accelerometers are assembled orthogonally, and if one accelerometer is vertical, and the device is held motionless relative to the Earth, the vertical sensor will show 1g while the other two accelerometers will show 0g. If the device is dropped, all sensors will show 0g.

Rotation

If a device containing three accelerometers is turned over, and is not in free fall, accelerometers mounted orthogonally will show values that vary depending on their alignment with the force of gravity.

Calculation

The force acting upon an object will cause a rate of acceleration that can be calculated by Newton's Second Law of Motion, provided that the object can move freely and is not being subjected to any additional forces, such as the force of gravity. If F is the force, m is the mass of the object, and a is the acceleration:

```
F = m * a
```

And therefore:

```
a = F / m
```

If the mass is restrained by a spring whose compression or extension has an approximately linear relationship to the force applied to it, acceleration can be calculated as a function of the linear displacement of the mass.

These statements ignore relativistic effects that are insignificant unless ultra precise time and motion measurement may be involved.

In a real-world accelerometer, movement of the mass will require some form of damping to prevent it from oscillating.

Variants

Accelerometer prices dropped radically after 2010. In an effort to maintain profitability, manufacturers have loaded more features onto chips. While a 2-axis accelerometer such as the Memsic 2125 seemed a good option when first introduced, it is now facing obsolescence as 3-axis accelerometer chips that also contain 3-axis gyroscopes have become ubiquitous—and no more expensive.

Early chip-based accelerometers provided analog outputs where voltage was proportional with acceleration and could be processed by a comparator. On some breakout boards, such as the Dimension Engineering DE-ACCM6G, which used the STMicroelectronics LIS244ALH chip, a comparator was included (see Figure 10-3).

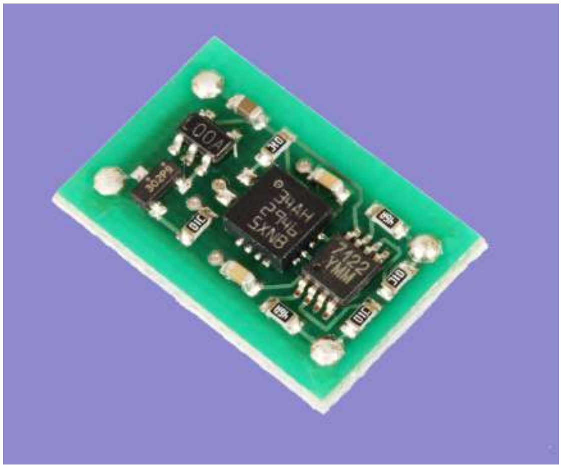

Figure 10-3 *A relatively early, relatively expensive breakout board using the 2-axis accelerometer LIS244ALH chip by STMicroelectronics. It has been superceded by chips that combine accelerometers, gyroscopes, and processors to provide a digital output.*

Because this board allowed a maximum output of only 0.83mA, it was only suitable for high-impedance logic chips or a microcontroller. However, because the output was analog, it could be passed through another comparator for direct connection to a piezo beeper to create a device that would sound an alarm when tilted. This is shown in Figure 10-4.

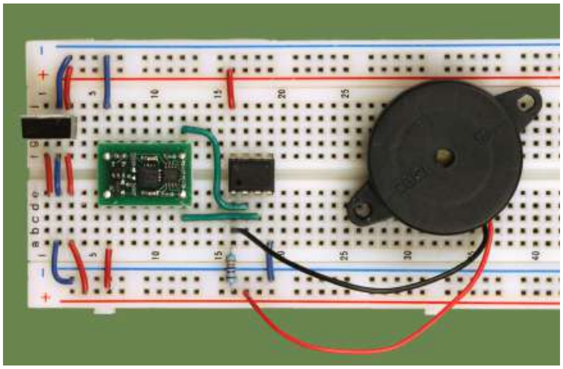

Figure 10-4 *Connecting the analog output from the DE-ACCM6G 2-axis comparator breakout board through a comparator to a piezo beeper.*

Many chips now contain gyroscopes as well as accelerometers. The electron micrograph in Figure 10-5 shows the interior of a chip of this kind, where the zig-zag shapes are “springs” etched into the die, the large areas patterned with dots are masses that can respond to various forms of motion, and the parallel plates are capacitive sensors.

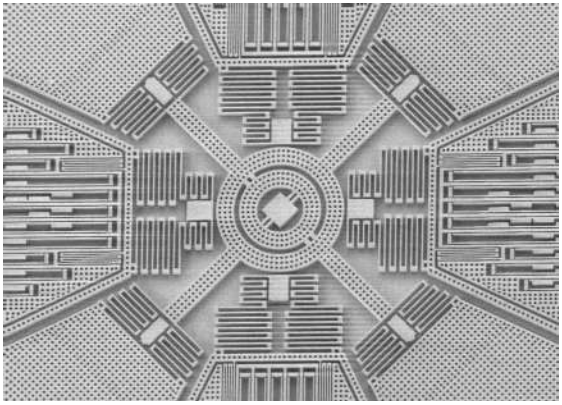

Figure 10-5 *Interior of a chip combining gyroscopes with accelerometers.*

The complexity of a chip combining two types of sensors creates a need for more complicated code to process the six outputs. Consequently most accelerometer chips now have their own onboard ADCs, and digital registers for communication with microcontrollers via the I2C communication protocol. Some chips also have onboard processing to interpret the mix of data. However, microcontrollers still need code to make sense of the data.

The hobby-electronics community has responded. A breakout board such as the LSM9DS0 from Adafruit, using a chip from STMicroelectronics that shares that same part number, tries to make this extremely complex chip accessible to experimenters. Shown in Figure 10-6, the LSM9DS0 contains a 3-axis magnetometer in addition to a 3-axis gyroscope and a 3-axis accelerometer.

The datasheet for the LSM9DS0 runs to more than 70 pages. At the time of writing, Adafruit is still refining code for the Arduino to make the features of this chip accessible.

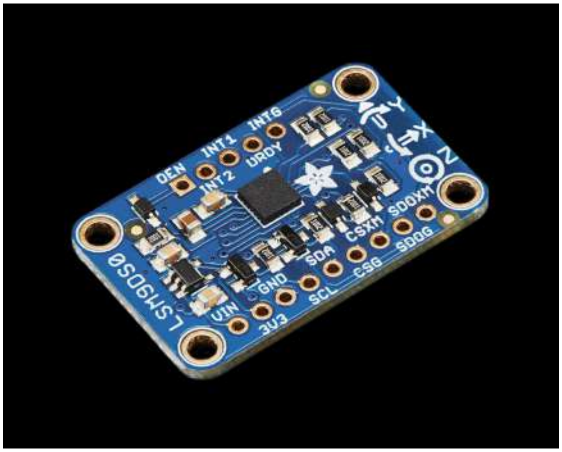

Figure 10-6 *This breakout board from Adafruit is built around an LSM9DS0 chip that combines accelerometers with gyroscopes and magnetometers.*

Despite the complexity of the LSM9DS0 chip, the breakout board sells for approximately the same price, at the time of writing, as the DE-ACCM6G 2-axis analog-output magnetometer four years ago. In the future we may expect IMU prices to fall still further, so that this type of multifunction chip becomes the default, and users simply ignore the functions that they don’t need.

Values

Current consumption will vary depending on the activity of the chip, and may be broken down for different types of sensors on multisensor chips. A modern accelerometer may typically draw less than 1mA. Gyroscope power consumption will be greater, as a segment of the chip will be maintained in a state of vibration.

Linear acceleration measurable by an accelerometer is customarily expressed in gravities, abbreviated g. Current generations of chips may be able to measure as much as plus-or-minus 16g, but because the value is converted internally to a digital quantity, smaller accelerations will not be expressed so accurately. Therefore the measurement range for acceleration is usually user-selectable by sending an appropriate code to the chip. Ranges may include plus-or-minus 2g, 4g, 6g, 8g, and 16g.

Sensitivity defines the smallest increment of acceleration measurable by the least significant bit (LSB) in an output register, for each of the acceleration ranges. In a range of plus-or-minus 2g, the internal 16-bit ADC may be capable of measuring 0.06 milligravities. When the range is reset to plus-or-minus 16g, the smallest increment may be around 0.7 milligravities.

The measurable range of gravities should not be confused with the maximum acceleration that the chip can tolerate without suffering internal mechanical damage, either while it is powered or unpowered. This will be more than 1,000g provided the duration is brief, and would only be experienced during an impact.

Linear acceleration sensitivity change versus temperature is usually expressed as a percentage, such as plus-or-minus 1.5%.

Output type will be analog or digital. If digital, the data protocol will be I2C or SPI. If I2C, the address of the device should be configurable. The data transfer rate is typically at least 100kHz, and this also may be configurable to different rates.

For additional details about protocols such as I2C, see Appendix A.

What Can Go Wrong

Mechanical Stress

Stress can be induced when a surface-mount chip is soldered to a board. The moving parts in a chip-based accelerometer may be adversely affected. Datasheets will supply information regarding maximum acceptable temperature during the soldering process.

Other Problems

If the accelerometer function is combined with other sensing functions such as magnetometer or gyroscope, see Chapter 2 or Chapter 9 for additional cautions about potential problems affecting these kinds of sensors.

vibration sensor

11

An **accelerometer** can measure some aspects of vibration. However, this entry deals primarily with mechanical devices (often described as *vibration switches*) and piezoelectric devices (often described as *vibration sensors*) that are solely intended to measure vibration.

A *vibrometer* measures vibration using a laser beam aimed at a reflective spot applied to a surface. It is usually a laboratory instrument, beyond the scope of this Encyclopedia.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **tilt** sensor (see Chapter 8)
- **force** sensor (see Chapter 12)

What It Does

A vibration sensor responds to repetitive mechanical motion. Most versions contain two switch contacts that are normally open and will close if the sensor vibrates in its designed frequency range. In some sensors, the frequency range and sensitivity are manually adjustable.

Large sensors are used as automatic shutdown switches responding to excessive vibration in machinery, and may be capable of switching substantial currents (10A or higher). Smaller versions can shut down domestic appliances such as a washing machine that is seriously out of balance during a spin cycle.

A vibration sensor can be used as a simple user-input device in toys and games.

A shock sensor can detect abuse of a sensitive device, for example, by including the sensor and a data logger when the device is transported.

Schematic Symbols

Either of the symbols in Figure 11-1 may represent a piezoelectric or piezoresistive vibration sensor, but they also represent other piezo-based devices.

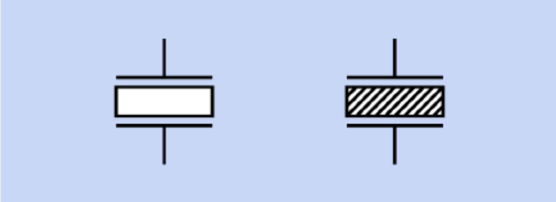

Figure 11-1 *Either of these symbols may represent a piezoelectric or piezoresistive device, including (but not limited to) vibration sensors that operate on this principle. The symbol on the left is more common.*

Variants

Vibration sensors use a wide variety of detection methods.

Pin-and-Spring

Probably the simplest type of sensor consists of a small, thin pin in the center of a miniature coil spring. The spring is anchored at its base, but its other end is free to vibrate. If the vibration reaches a sufficient amplitude, the spring touches the pin, completing a circuit between the two leads of the device.

An example is shown in Figure 11-2, where two identical sensors are shown, one of them cut open to reveal the gold-plated rod and spring inside. This sensor is rated for 10mA at up to 12VDC.

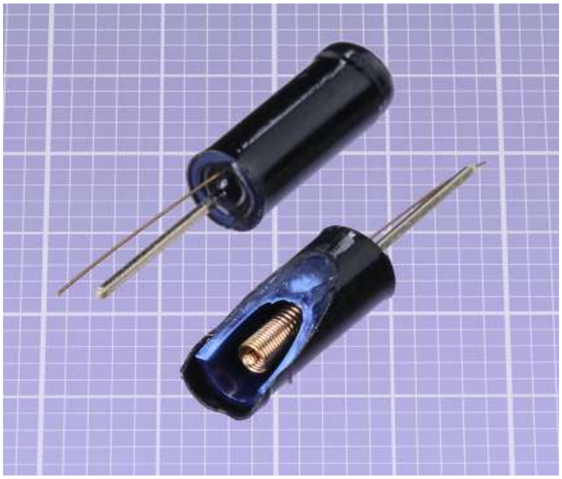

Figure 11-2 *When the spring in this sensor vibrates, it touches the pin that is centered in it. The background grid is in millimeters.*

Advantages of this system are low cost, ability to respond along two out of three axes, no power supply requirement, and ability to switch AC or DC. However, because the internal contact is extremely brief, it must be connected with a latching component of some type. A flip-flop could be used, or a 555 timer. The switch may also be connected with an input pin on a microcontroller, provided a pullup or pulldown resistor is used to prevent the input from floating when the switch is open.

Externally, the packaging of a pin-and-spring vibration sensor is almost indistinguishable from a small **tilt** sensor that contains one or two spherical metal balls. The tilt sensor may respond to vibration, but only of a large amplitude and low frequency.

A miniature board containing a pin-and-spring vibration sensor, comparator, and trimmer potentiometer for sensitivity control is sold cheaply by the Chinese supplier Elecrow as their product SW-18015P, shown in Figure 11-3. Elecrow also offers a wide range of other low-cost sensors.

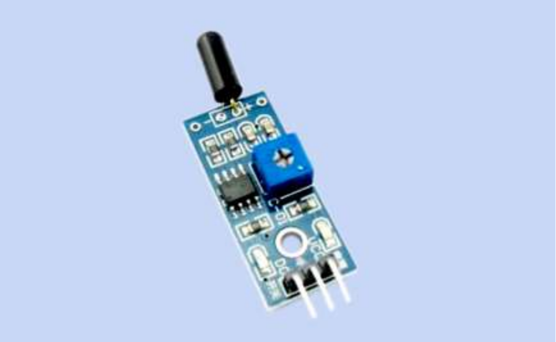

Figure 11-3 *A pin-and-spring vibration sensor mounted on a board with sensitivity control.*

Piezoelectric Strip

The LDT0-028K by Measurement Specialties is a section of piezoelectric polymer film laminated to a polyester substrate. The film is designed to be anchored at one end, allowing the other end to vibrate. An unweighted version and a weighted version are shown in Figure 11-4, each measuring about 13mm x 25mm. Addition of the weight alters the resonant frequency of the sensor.

Deflection of about 2mm is sufficient to generate a surprising 7VDC between the two leads. Larger deflections will generate higher voltages. The manufacturer suggests that direct connection to a CMOS component is possible. An op-amp may be used for signal conditioning.

A piezoelectric device of this type only generates voltage during the process of deflection. If the strip is held in a curved position, its output diminishes to zero.

This sensor has a resonant frequency around 170Hz when there is no weight attached to its free end.

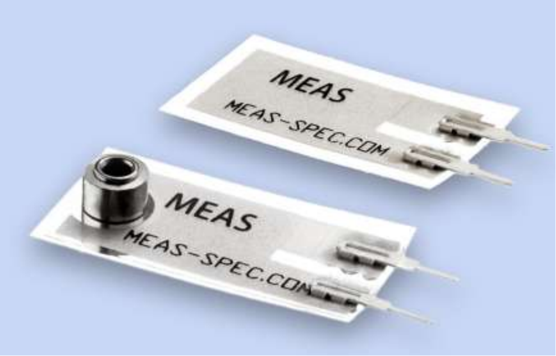

Figure 11-4 *Two versions of the LDT0-028K vibration sensor from Measurement Specialties, one with an added weight to lower its resonant frequency.*

Chip-Based Piezoelectric

The Murata PKGS series is an example of a surface-mount piezoelectric shock sensor. Measuring only about 1mm x 2mm x 4mm it has an analog output designed for connection through an op-amp. The manufacturer suggests application in a hard-disk drive to block read-write operations when vibration occurs. Similarly, these shock sensors may be used in CD-ROM or DVD drives. They may also be installed in cash dispensing machines to sound an alarm if vandalism occurs.

Toshiba's TB6078FUG is a similar product. Note that because these devices contain electronic components, they require a power supply (usually 3.3VDC to 5VDC) for operation.

"Mousetrap" Type

Some vibration switches rely on a simple system of leverage that is comparable to a mousetrap, in that a relatively small stimulus releases a strong spring. In Figure 11-5 the upper part of the figure shows the switch at rest, held in its position by a powerful spring and by the weight of a mass attached to a pivoted arm. In the lower part of the figure, severe vertical vibration has caused the assembly to move up and down with sufficient energy to overcome the tension in the spring while the inertia of the mass has resisted the motion. Consequently the arm has moved past the position where the spring is aligned with the pivot, and the spring now acts to hold the arm against a snap-action switch. A system of this type is used in sensors on some power-station cooling towers, where the loss of a large fan blade can result in major vibration.

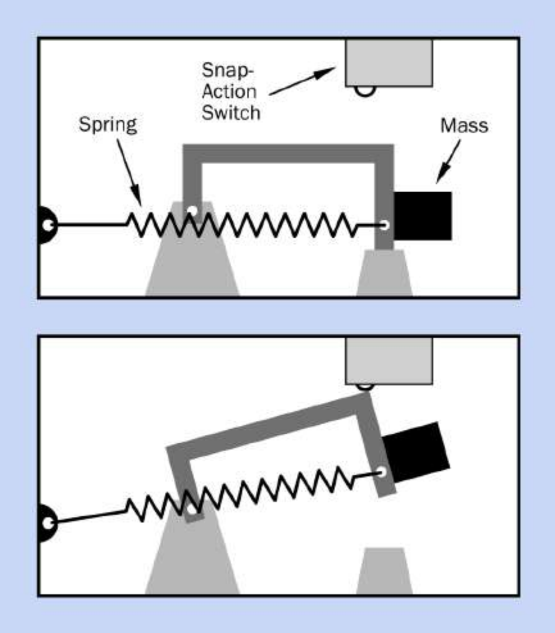

Figure 11-5 *A spring-loaded vibration sensor.*

Magnetic

Typically this method is used to detect excessive vibration in machines or other devices containing heavy rotating mechanical components. The sensor may be physically large, moving switch contacts that are designed to handle currents of 1A or significantly more.

In one system, a steel ball is retained by a permanent magnet that is barely powerful enough to prevent the ball from falling. Excessive low-frequency vibration will dislodge the ball, which falls and completes a circuit between two contacts. This activates a relay that powers down the piece of machinery that is vibrating.

The ball may be spaced a small distance from the magnet by a beveled nonmagnetic seat, and the seat may be movable with an external screw. This will adjust the sensitivity of the switch. After the switch has been triggered, it must be reset, which may entail using an external lever to raise the ball back to its location near the magnet.

Another magnetic system is shown in Figure 11-6, where a magnet on a vertical arm can be displaced by horizontal vibration, and an inertial mass on a spring-loaded horizontal rod can also dislodge the magnet in response to vibration along the other two axes.

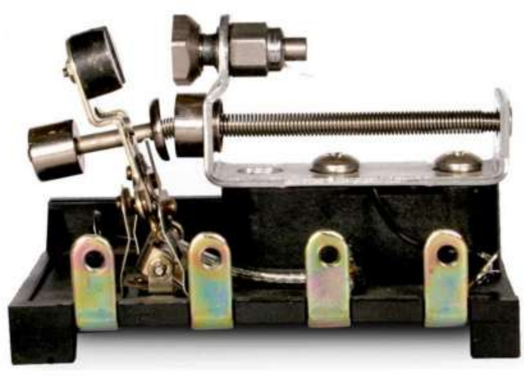

Figure 11-6 *A magnetic vibration switch. See text for details.*

Mercury

A small *mercury switch* may be used as a vibration sensor, although this application is uncommon. See Chapter 8 for more information about mercury switches.

Values

Measurement of vibration is a complex science of special interest in mechanical design, especially in areas such as the powertrain and suspension geometry in an automobile. Only a few fundamentals will be summarized here.

Primary Variables

The four primary variables in vibration are frequency, displacement, velocity, and acceleration. Frequency describes how rapidly the vibration occurs; displacement describes how far the vibrating object moves in each direction; velocity describes how fast it moves during each cycle; and acceleration describes how rapidly the velocity changes during each cycle. Different types of sensors can be chosen for their responsiveness to each attribute.

Figure 11-7 shows the theoretical relationships between displacement, velocity, and acceleration plotted against the frequency of vibration. The y axis of this graph (the vertical axis) is often labeled “amplitude,” but in reality it is being used to measure three different units, as shown. The curves indicate that if the velocity of vibration remains constant while frequency increases, acceleration must increase as a function of the frequency while displacement decreases. The acronym “rms” denotes that the values are measured as the “root mean square” of their fluctuations.

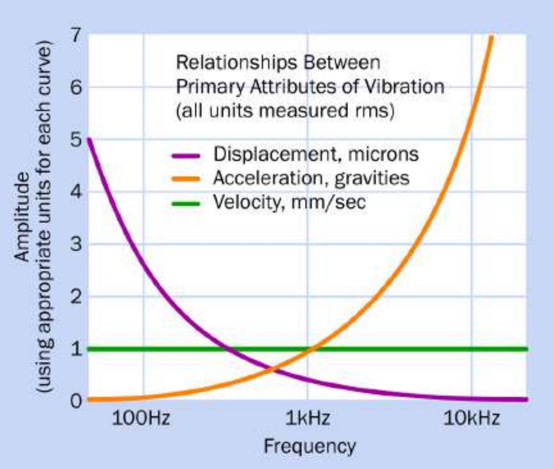

Figure 11-7 *Theoretical relationships among the primary attributes of vibration.*

Mechanical sensors or switches that respond to displacement are best suited to low frequencies, while piezoelectric sensors that are sensitive to acceleration are best suited to higher frequencies. Serious mechanical problems tend to result in low-frequency vibrations, while wear on bearings and in gear trains will be more likely to create high-frequency vibration.

Dynamic Attributes

Datasheets for small piezoelectric sensors may only show basic values such as the voltage range that is likely to be created.

Small mechanical sensors of the pin-and-spring type will be rated for maximum voltage and switching current. Typical values may be around 12V (AC or DC) and 10mA, indicating that the output from this type of sensor should be used with an op-amp, microcontroller, logic chip, solid-state relay, or other semiconductor with a high-impedance input.

Sensors that are manufactured for industrial applications will be rated for attributes such as measurable acceleration range (in gravities), temperature sensitivity, frequency response, resonant frequency, capacitance, and power requirement.

The sensitivity of piezoelectric sensors is usually expressed in mV/g. This type of sensor will require a comparator to process the tiny amount of current that it creates.

How to Use It

If a vibration sensor has an analog output requiring a comparator, the output from the comparator is likely to be of the open-collector type. This will require a pullup resistor of a value that provides an appropriate voltage for the next stage in the circuit. For more information about comparators, see Volume 2. For more information about the use of an open-collector output, see "3. Analog: Open Collector".

A *coupling capacitor* can remove the DC component from the comparator output, allowing only the frequency of vibration to pass through. The choice of capacitor value will depend on the frequency.

When using a piezoelectric sensor with analog output, a 10M resistor may be installed across its two terminals to reduce voltage drift.

The primary challenge in getting a vibration sensor to work successfully will be matching it to the source of vibration. Manufacturers' datasheets for chip-sized sensors often provide very little information about the optimal values for their products. Peak performance will occur when the natural resonant frequency of a sensor is close to the frequency of vibration that it must detect. Trial and error may be necessary.

A sensor must be mounted appropriately. Most sensors are directional, at least to some extent, and many will not respond significantly to vibration at 90 degrees to their primary axis of sensitivity. Their performance will also diminish if they are placed too far from the vibration source, or if they are mounted on a flexible or yielding surface that will tend to absorb vibration.

While industrial vibration switches may be adjusted manually, the response of small devices designed for circuit-board mounting can only be tweaked using external components to filter out unwanted signals from the sensor.

What Can Go Wrong

Long Cable Runs

The output from a piezoelectric vibration sensor is primarily an AC signal, fluctuating at the frequency of the vibration. Long cabling, or inadequately shielded cabling, can potentially introduce capacitive effects that can degrade the sensor signal. This issue will only affect higher frequencies.

Interference

Sensor signals can also be affected by electromagnetic interference from power lines, transformers, and large motors. This is a significant issue, as a primary application for industrial vibration sensors is to measure vibration created by motors.

Correct Grounding

For large sensing equipment, grounding may be important to shield cables that transmit data. In an industrial environment, however, grounding is primarily motivated by safety considerations, and an electrical ground can carry unwanted interference. Ground loops may be created if there are multiple ground points. Ideally, a "ground tree" should be used, where there is only one primary grounding point, and grounds to equipment branch out from it.

Fatigue Failure

In installations where some vibration normally exists, cables should be anchored properly to minimize the risk of fatigue failures.

force sensor

12

A *load cell* or *load sensor* is generally intended to measure a static load, while a *force sensor* can respond dynamically. However, this semantic distinction is not always observed. This entry differentiates between load cells and force sensors but includes them both.

Traditional types of *hydraulic* and *pneumatic* load sensors are not electronic devices. They are outside the scope of this Encyclopedia.

Capacitive, *ultrasonic*, *magnetic*, *optical*, and *electrochemical* force sensors are relatively unusual, and are not included in this Encyclopedia.

A force sensor is occasionally described as a **pressure** sensor, but that term is ambiguous, as it is more often used in conjunction with fluids. This Encyclopedia assumes that a pressure sensor only measures gas or liquid pressure. See Chapter 17.

A **vibration** sensor reacts to rapidly changing forces, but usually cannot measure them accurately, and is simply triggered when vibration exceeds a threshold. See Chapter 11.

Impact sensors that measure the force of a collision are outside the scope of this Encyclopedia.

A sensor designed to respond to a single touch from a fingertip is considered a human-input device and is discussed in the **single touch** sensor entry. See Chapter 13.

OTHER RELATED COMPONENTS

- **vibration** sensor (see Chapter 11)
- **single touch** sensor (see Chapter 13)

What It Does

A **force** sensor measures physical force that is applied to it, either by a person or by an object. Many force sensors respond rapidly and can measure fluctuating forces.

A *load cell* or *load sensor* is usually intended to measure the static weight of an object.

Applications

In robots, force sensors can provide feedback to limit the grip of a mechanical hand. They can also provide haptic feedback for a surgeon performing robotic surgery. In the future, force sensors may find increasing application in agriculture, as mechanized handling of fruit and other foods requires a carefully controlled gripping force.

In medicine, the use of force sensors to evaluate muscle strength in hands or limbs can be important as an indicator of neurological problems or to monitor progress in occupational therapy. Thin force sensors can be installed in shoes to check the weight distribution of each foot. They can also be used for entertainment purposes, to light LEDs in sneakers.

A force sensor may respond to a single-touch user input. See Chapter 13 for more information about **single touch** sensors. Some video game controllers use resistive force sensors to measure the amount of pressure applied to a button. The PlayStation is an example. (Old PlayStation DualShock 2 controllers are a salvageable source of pressure-sensitive buttons.)

Load sensors are used to weigh industrial products, and are also used domestically in kitchen and bathroom scales.

A load sensor can also detect human presence —for example, in the passenger seat of an automobile, where an air bag must not deploy if a young child is present, or in a hospital, to monitor the number of times the patient gets out of bed.

Schematic Symbol

No specific schematic symbol is used for either a force sensor or a load sensor. If a force sensor uses a piezoelectric or piezoresistive element, it may be represented by the symbol shown in Figure 11-1, which is used for many piezo devices.

How It Works

Two methods of force measurement are commonly used: resistive and piezoelectric.

A *piezoelectric* force sensor uses a piezoelectric element, often consisting of a quartz crystal, to convert force to a small voltage that can be amplified. However, this type of sensor only responds to changes in force. If a constant load is applied, the output peaks quickly and then gradually diminishes to zero.

Resistive force sensors change their electrical resistance when force is applied. They include *metallic strain gauges* and *plastic-film* sensors in which two layers of conductive ink are pressed together.

In SI (standard international) units, the force needed to activate a sensor is measured in newtons, abbreviated with a capital letter N. A newton is defined as the force that would accelerate a mass of 1 kilogram at 1 meter per second each second. More practically, in the gravitational field at the surface of the Earth, 1N = about 100 grams of weight. There are about 28 grams in an ounce; thus 1N is slightly less than 4 ounces.

Strain Gauge

A strain gauge is often made from metallic foil applied to an insulating flexible backing. The backing is glued to a shaped piece of metal, usually steel or aluminum, which is designed to flex slightly under pressure and may be referred to as a *spring*, even though it is often one solid object. Its deflection will be related to the force imposed on it.

The maximum deformation of the spring under a strain gauge is usually 500 to 2,000 parts per million (ppm) when subjected to the maximum force that it is designed to measure. A change of 1ppm is referred to as a *microstrain* (abbreviated με).

The strain gauge has no polarity, and functions like a force-controlled **potentiometer** (see Volume 1). The ratio of the change in its resistance to the change in the strain that it experiences is called the *gauge factor*. For metal foil gauges, the gauge factor is usually around 2.0. This is an approximately linear relationship.

The most common type of foil pattern is shown in Figure 12-1. In the figure, if a stretching force is applied horizontally, the multiple thin sections of foil are slightly elongated, and their resistance increases within the limits of elasticity of the foil. This effect is multiplied by the number of sections. If the stretching force is applied vertically, the sections are merely separated slightly, and the result is negligible.

Wheatstone Bridge Circuits

The very small changes in resistance in a strain gauge must be amplified to be usable, and the first step is to use a Wheatstone bridge circuit. The simplest form of this circuit is shown in Figure 12-2.

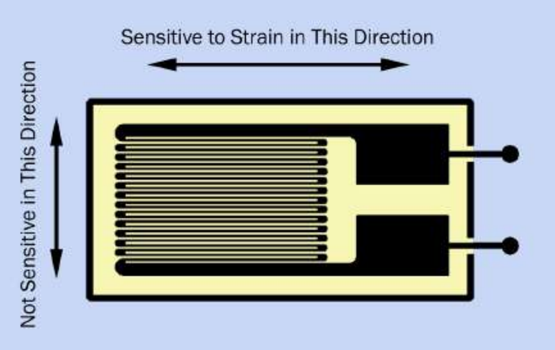
Sensitive to Strain in This Direction
Not Sensitive in This Direction

Figure 12-1 *The pattern of metallic foil used in a typical strain gauge.*

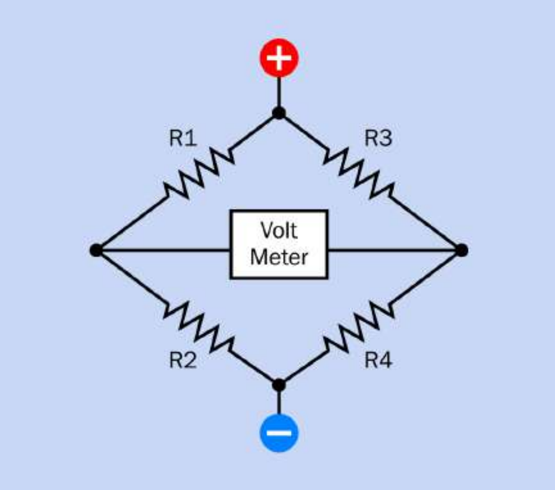

Figure 12-2 *A demonstration version of the basic Wheatstone bridge circuit that is often used to detect small changes in a resistance.*

Each pair of resistors (R1 + R2, and R3 + R4) functions as a voltage divider. If all the resistors have an exactly equal value, the voltage at the midpoint of each pair will be identical, and the volt meter at the center will have a zero reading. However, if the value of one resistor changes slightly, the meter will register the imbalance. This circuit is commonly used because of its sensitivity to small variations.

In Figure 12-3 two strain gauges have been substituted for resistors R3 and R4. The upper strain gauge has been mounted so that it experiences an increase in force at the same time that the lower strain gauge experiences a decrease, usually because one gauge is mounted on the top side of a flexing element while the other gauge is mounted on the underside.

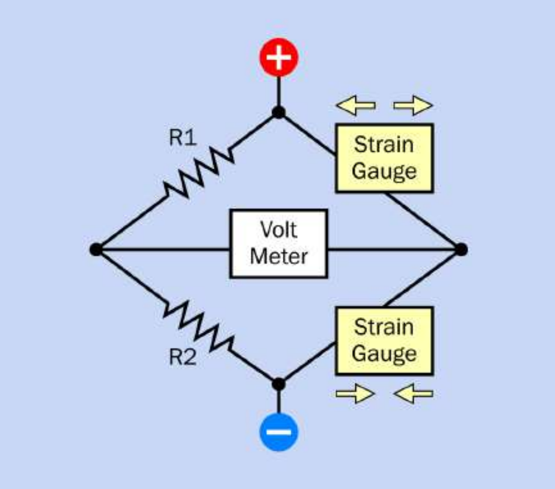

Figure 12-3 *Two strain gauges, oppositely oriented, can be used as resistances in the Wheatstone bridge circuit.*

Using two strain gauges in this way doubles the sensitivity of the Wheatstone bridge circuit. The configuration is known as a "half Wheatstone" force sensor, and will have three connecting wires. One will be black, one will be red, and the third will be a different color. The red and black wires are for connection to the power supply, as shown in the schematic, while the third wire is common and should be considered as an output.

If an additional two strain gauges are inserted in the Wheatstone bridge circuit, this is now a "full Wheatstone" force sensor (see Figure 12-4). Note, however, the diagonally symmetrical orientations of the strain gauges, necessary to multiply their effect.

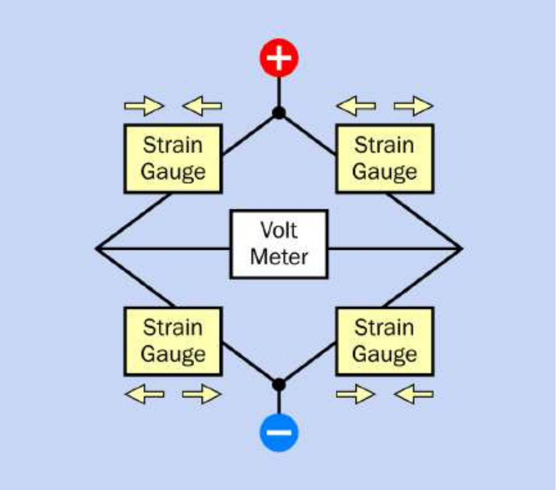

Figure 12-4 *Two additional strain gauges create a "full Wheatstone" circuit.*

A typical digital bathroom scale contains two half-Wheatstone force sensors, wired to create a full-Wheatstone configuration.

The force sensors shown in Figure 12-5 are rated for up to 50kg each, and can thus weigh up to 100kg if they are combined in a scale. In this figure, one sensor has been turned over to show its underside. The strain gauges are hidden in each sensor where the steel sections overlap.

Wheatstone Bridge Errors

Where more than one strain gauge is used in a Wheatstone bridge, they should ideally have identical performance. Since this is impossible as a result of manufacturing tolerances, some error correction is built into devices using multiple strain gauges.

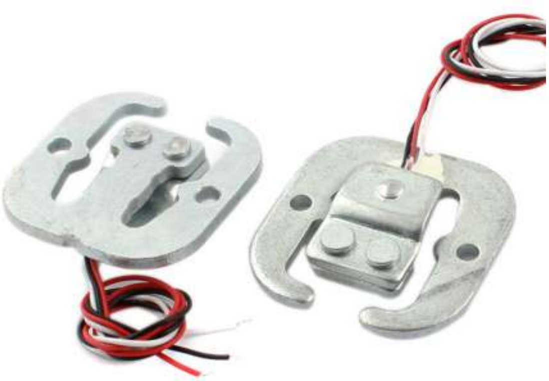

Figure 12-5 *Two force sensors suitable for a bathroom scale, each containing a pair of strain gauges that will flex oppositely under load.*

Strain-Gauge Amplification

The voltage output from a Wheatstone bridge circuit is given by the following formula, where V_{IN} is the supply voltage, V_{OUT} is the output, and R1 through R4 are the resistance values that were used in Figure 12-2.

$$V_{OUT} = [(R3/(R3+R4)) - (R2/(R1+R2))] * V_{IN}$$

The good news is that the output when a strain gauge is used will vary linearly with the load applied. The bad news is that it will be very small.

To amplify it, an op-amp such as the AD620 is often recommended. Using external resistors, its amplification factor can be adjusted from 1:1 to 10,000:1. Alternatively a chip such as the HX711 by Avia Semiconductor contains a 10-bit analog-to-digital converter and is specifically designed for use in weighing scales. Its digital output uses a very simple serial format. Sparkfun sells a breakout board incorporating this chip. See Figure 12-6.

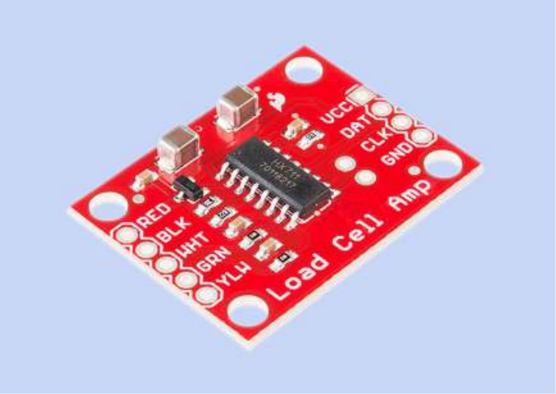

Figure 12-6 *Sparkfun HX711 breakout board for the HX711 amplifier chip, specifically designed for a full-Wheatstone array of strain gauges in a force sensor.*

Other Strain-Gauge Modules

Strain gauges are built into a variety of sensor modules. Figure 12-7 shows a selection, all of which are available from Sparkfun. Many more can be found online.

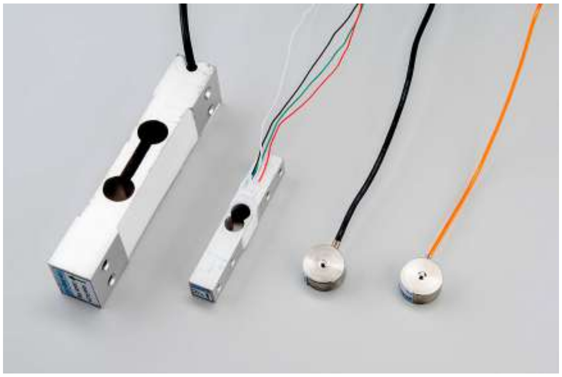

Figure 12-7 *A selection of load cells containing strain gauges, available from Sparkfun.*

Plastic-Film Force Sensors

Plastic-film resistive sensors contain two layers of conductive ink, sealed between two layers of thin, transparent plastic film. The resistance between the ink layers diminishes when they are pressed together. It may vary from as little as 30K when fully loaded to more than 1M when unloaded.

Like a strain gauge, this sensor has no polarity and requires no power supply.

Examples of this type of sensor are shown in Figures 12-8 and 12-9. Other shapes and sizes are available. Manufacturers include Tekscan, whose product is named FlexiForce; Alpha Electronic, in Taiwan; and Interlink Electronics.

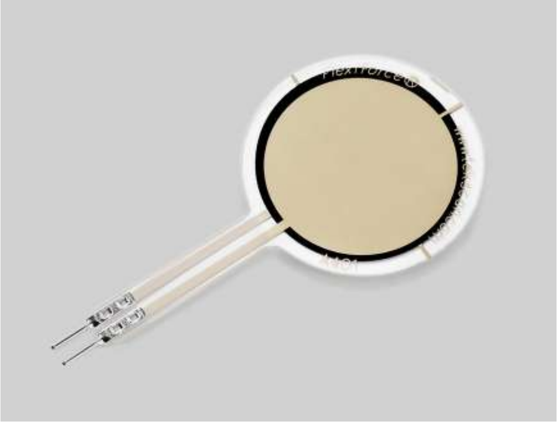

Figure 12-8 *A FlexiForce A401 resistive film sensor made by Tekscan, Inc. Its sensing area is 25.4mm in diameter (1 inch) and is rated to measure up to 111N (25lbs).*

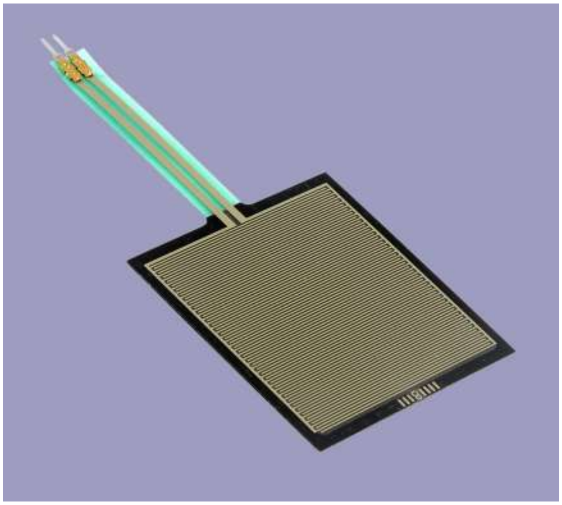

Figure 12-9 *An Interlink FSR406 resistive film sensor. Its sensing area is slightly less than 40mm square, and is rated to measure up to 20N (4.5lbs).*

Plastic-film sensors should not be confused with film-based piezoelectric vibration sensors, which are described in the **vibration** entry (see "Piezoelectric Strip"). Those sensors provide a transient output when they flex rapidly. The resistive sensors provide a stable output in response to a steady load.

Deformative Force Sensors

A sheet of natural or silicone-based rubber can be impregnated with conductive particles. The conductivity of the sheet may not change significantly when it is compressed, but if it is separated from a metal plate by a mesh of thin nylon fibers, compression will result in greater conductivity by pushing the rubber into the gaps between the fibers. See Figure 12-10.

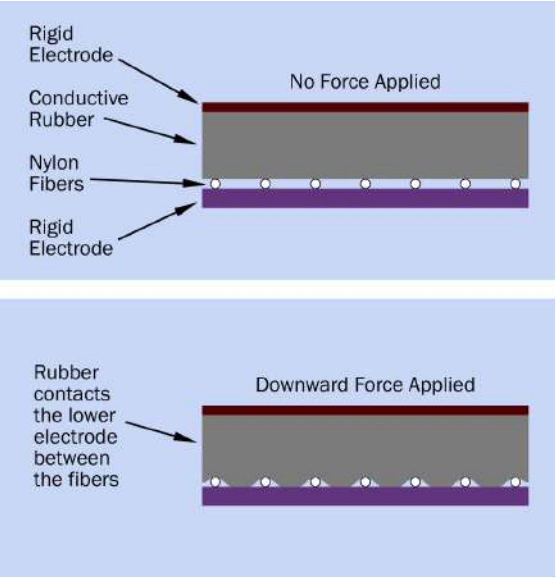

Figure 12-10 *In the top image, a flexible conductive layer is separated from an underlying rigid electrode by nylon mesh. In the lower image, a load applied to the flexible layer has forced more of it into contact with the electrode, reducing the resistance between them.*

Improvised Resistive Sensors

Polyethylene film impregnated with carbon particles is available under the brand name Velostat, owned by 3M. Although it was developed as an antistatic packaging material for semiconductors, it can be used to make a DIY force sensor. When the material is stretched, the embedded particles are more dispersed, and electrical resistance increases. When the material is compressed, its resistance is reduced.

Antistatic foam of the type used to package CMOS components can be used in the same way, although some types behave like memory foam, being slow to recover from pressure. The foam can be sandwiched between a pair of copper-plated circuit boards as electrodes.

How to Use It

Plastic-Film Resistive Force Sensors

The conductivity of a plastic-film sensor has an almost linear relationship with the force applied. In other words, if F is the force and I is the current:

```
I = k * F
```

where k is a constant determined by the characteristics of the materials used.

By Ohm's Law, R = V / I where V is the voltage drop across a resistance of value R. By substitution, using k * F instead of I:

```
R = V / (k * F)
```

Therefore, if a constant voltage is applied across the force sensor, the resistance of the sensor will be proportional to the reciprocal of the force (i.e., 1 / F). These relationships are shown in Figure 12-11.

For convenient measurement, it will be helpful if the resistance of the sensor can be converted to a voltage output that varies linearly with the force applied. To achieve this, a resistive force sensor of this type is customary amplified with an **op-amp** (see Volume 2). Using the schematic shown in Figure 12-12, the amplification ratio, A, is found from this formula:

```
A = 1 + (R2 / R1)
```

where R2 if the potentiometer, and R1 is the resistive force sensor. Therefore, the output from the op-amp should have an approximately linear relationship with the force applied to the sensor.

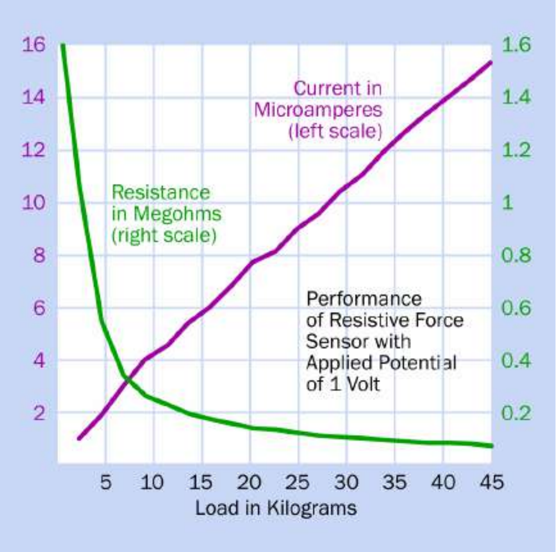

Figure 12-11 *The relationship between force, current, and resistance in a flexible resistive sensor, assuming a constant voltage of 1V is applied across it (derived from FlexiForce Sensors User Manual).*

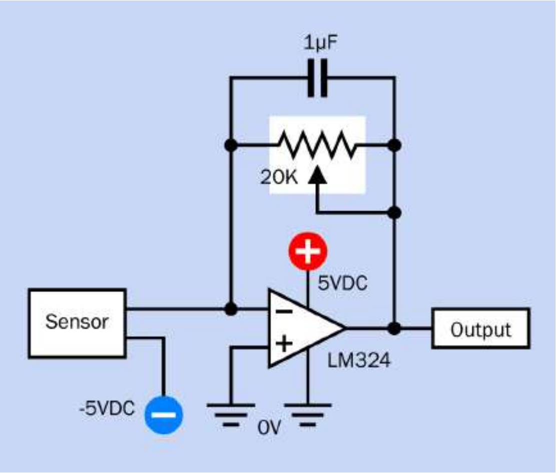

Figure 12-12 *An amplification circuit with an output that is approximately linear with force applied to a resistive film sensor.*

The capacitor in this circuit is added to suppress noise that the circuit might otherwise pick up.

An alternative is to wire a resistive sensor in series with a capacitor, and connect the other side of the capacitor to a comparator that has adjustable feedback. The resistance of the sensor will determine how quickly the capacitor charges. However, because of the mathematics describing the charge rate, the output from the comparator would not be linear with the force on the sensor. Also, provision would be necessary to discharge the capacitor intermittently.

Values

Film-Based Force Sensors for User Input

A very light pressure with a fingertip could be around 50g. A more defined finger-press would be 250g, and a heavy push with a finger would be around 1kg.

The Interlink range of flexible force sensors requires a minimum pressure of 0.2N, or about 20g. Similarly, Alpha products range from a minimum of 10g to 30g.

These specifications suggest that film-based force sensors may be used for one-touch user input, but the no-load resistance will be at least 1M, and more than 10M in some instances. A small amount of pressure is unlikely to reduce this electrical resistance much below 500K. Using an op-amp or comparator to detect that difference and convert it into a reliable on-off output may be vulnerable to noise and power-supply disturbance.

Another consideration is that film-based sensors provide no tactile feedback. For these reasons, and because film-based products are described by their manufacturers as "force sensors" rather than as "touch sensors," they are included in this entry rather than in Chapter 13. That said, they may be considered as an option for one-touch user input where they are appropriate—for instance, in games where players are likely to slap or hit a sensor vigorously.

Specifications for Film-Based Force Sensors

Durability of film-based sensors is excellent, with manufacturers claiming that performance is not degraded after 1 million applications of a 20kg load. Sensors have no polarity, will function using voltages from 1V to 15V in most cases, and have a response time of less than 5ms. They do not generate, and are not vulnerable to, electromagnetic interference.

Important attributes when evaluating film-type sensors are their limits of tolerance for applied force, and resistance values at each end of the range. Unfortunately these values are poorly documented for many sensors, or may be unspecified.

Maximum force may range from 20N to 440N, depending on the brand and model of sensor.

Electrical resistance is at least 1M when unloaded, and may be as high as 20M.

Accuracy ranges from plus-or-minus 2% to plus-or-minus 5% from one application of force to the next, depending on the model of sensor and the manufacturer. If force is not applied each time in exactly the same area of the sensor, results will vary. If a sensor is swapped with another sensor of the same type, sensor-to-sensor consistency will be uncertain. Therefore, film-based force sensors are not a good choice for applications where accuracy is important.

Sensors may be rated for 5% to 10% hysteresis.

The area of active detection may range from about 4mm wide (FlexiForce sensors) to more than 40mm wide (Interlink FSR-406).

Strain Gauges

Strain gauges are not sold as individual components by most electronics suppliers. They are sometimes available as surplus parts or from sites such as eBay, where specifications range from 100 ohms to 1K as the resistance when no load is applied.

A load sensor on which a strain gauge has been preinstalled is much easier to use, and will be plug-compatible with an appropriate amplifier chip as previously described.

What Can Go Wrong

Soldering Damage

The pins on plastic-film resistive sensors are embedded in thin plastic. Heat from a soldering iron can easily damage this plastic. Heat-sink alligator clips should be used while soldering, or the pins can be socketed instead of soldered.

Bad Load Distribution

Film-based sensors will not provide accurate readings if a load is imposed unevenly or inconsistently, or extends outside the detection area. A *puck* consisting of a small, rigid disc may be interposed between the source of the force and the sensor, to distribute the load within the maximum area. A puck may also be referred to as a *shim*.

Similarly, the sensor must be mounted on a flat, smooth surface, and if this is not available, a rigid plate should be interposed.

Water Damage

Although film-based sensors are enclosed in plastic, they are not waterproof. Immersion may cause the layers to delaminate.

Temperature Sensitivity

Because electrical resistance tends to vary with temperature, readings from resistive force sensors will vary with temperature.

Ambient temperatures of 70 degrees Celsius and above may damage a film-based sensor.

Leads Too Long

Although film-based sensors are supplied with a variety of lead lengths enclosed in the laminated layers of flexible plastic, the leads may be too long for a particular application. If they are trimmed, wires cannot be attached with solder, as it will melt the plastic. Conductive epoxy should be used.

single touch sensor

13

This entry only describes *capacitive* touch sensors. A *conductive* sensor, which uses the fingertip to complete a circuit between two exposed contacts, is not very common, and is not included here.

The type of touch sensor described in this entry requires no physical pressure for activation. It should not be confused with a resistive or piezoelectric **force** sensor that requires pressure. See Chapter 12.

An integrated circuit chip that processes a signal from a *touch pad* is often described as a *touch sensor*, even though it does not contain a sensing element. This entry describes it as a "touch sensor chip" to eliminate ambiguity, and refers to touch input elements as "touch pads."

Touch pads that contain *tactile switches* or *membrane switches* are described in the entry discussing **switches** in Volume 1. All types of switches are described in that volume, with the exception of a *reed switch*, which is magnetically activated and is therefore categorized as a sensor.

Capacitive touch sensors are sometimes referred to as *capacitive proximity sensors*, because they sense the proximity of a human fingertip. In this Encyclopedia, and in most other sources, a **proximity** sensor measures distance, not touch. See Chapter 5.

A *capacitive displacement sensor* employs the same principle as a capacitive touch sensor, but is used to detect the position of an object, not for human input.

OTHER RELATED COMPONENTS

- **force** sensor (see Chapter 12)
- **touch screen** (see Chapter 14)

What It Does

A *touch pad* detects the presence of a human fingertip (or other part of the body) and signals an integrated circuit chip, which is very often termed a *touch sensor*, even though it does not contain a sensing element itself. The chip creates an output to signify that human touch has been recognized.

A keypad of the type found on microwave ovens may appear similar to an array of touch pads, but is more likely to contain *membrane switches* or *tactile switches*, which are described with other forms of switches in Volume 1. The type of touch pad described in this entry requires no physical force and contains no parts that move or flex when pressed.

A modern **touch screen** is usually a capacitive device and can be thought of as an array of touch pads. See Chapter 14.

Applications

Capacitive touch sensors have become common as their cost has fallen relative to simpler components that respond to being pressed.

A touch sensor can be used to start or stop a process, or to power up or power down a device. Multiple sensors may be found wherever user input of a few alphanumeric characters is required. Because touch pads can be completely sealed, they are useful where hygiene is important.

Specific applications include the activation of a backlight in a handheld device, wake-up from standby, ear detection in a cellular telephone, control of medical devices, and activation of interior lighting in some automobiles.

The absence of moving parts or electrical contacts means that a touch pad is more reliable than any type of electromechanical switch. A disadvantage is that it provides no tactile feedback, and therefore will require a visual or audible confirmation when it responds to input. Lack of tactile feedback makes touch pads unsuitable for computer keyboards and other key-entry devices where rapid typing is required.

When the capacitive elements of a touch pad are transparent, it can be mounted in front of a screen.

Schematic Symbols

Either of the schematic symbols in Figure 13-1 may sometimes be used to represent a touch sensor, but not on a consistent basis.

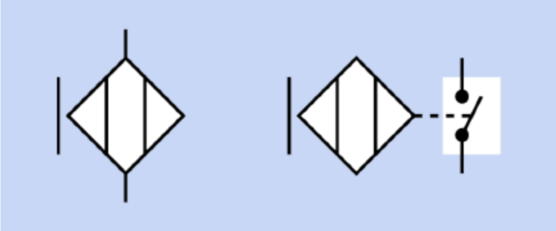

Figure 13-1 *Two possible schematic symbols that may represent a touch sensor.*

How It Works

The two plates in a capacitor are separated by an insulator known as the *dielectric*. Paper, plastic, glass, air, and other insulators can serve this purpose. Although no electrical connection exists between the plates, AC passes through the dielectric as a field effect.

Electrical capacitance exists between any pair of electrical conductors. The human body has high electrical resistance but still is electrically conductive, and therefore has capacitance with other conductive objects.

A touch pad can function as one side of a capacitor, with a fingertip functioning as the other side. In this mode, AC can pass from the touch pad and through the human body to ground. The current is very small, but fluctuations can be detected by an appropriately designed integrated circuit chip, or by a microcontroller.

The precise characteristics of a dielectric will affect the performance of a capacitor to some extent, but will not prevent it from working. Therefore a touch pad can function even when it is shielded behind a protective layer of glass or plastic, as is often the case.

A touch sensor chip generates pulses of low-voltage AC and sends them to a touch pad. The chip detects any variation in the current through the touch pad, indicating the presence of a fingertip. Where an input occurs, the chip changes its output, which usually requires a microcontroller for processing.

How to Use It

Touch sensor chips are available in as many as 40 different formats and configurations. All of them are surface-mount. For breadboarding, an experimenter can use breakout boards that have sensor chips installed. Figure 13-2 shows a product from Adafruit that is capable of addressing 12 touch pads. Its output is accessible from a microcontroller via the I2C protocol. For additional details about protocols such as I2C, see Appendix A.

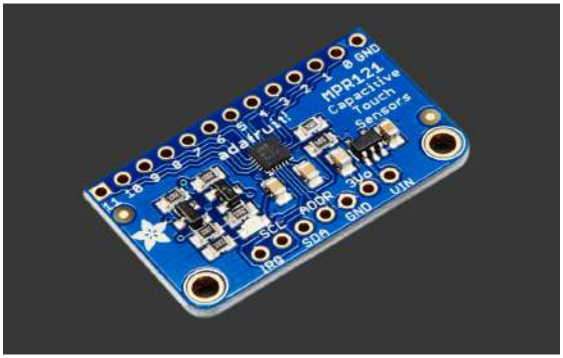

Figure 13-2 *A capacitive touch sensor chip on a breakout board from Adafruit.*

Similar breakout boards are available from Sparkfun, and from large online vendors such as *http://www.mouser.com* (where they are categorized as *development tools*).

While most touch sensor chips require a microcontroller, a few are available with the same number of output pins as touch-pad input pins, and each output pin will transition between logic-high and logic-low when an input on the corresponding pin is detected. Another breakout board from Adafruit, the AT42QT1070, uses this simple system.

A library exists for Arduino that enables two pins to sense touch on a piece of aluminum foil.

It can also work with conductive ink or paint.

Obtaining Touch Pads

Sensor chips are widely available as components, and are very inexpensive. On the other hand, touch pads are not common as components, probably because a touch pad is usually created as a pattern of copper traces etched onto a circuit board by a device manufacturer.

Touch pads from hobby-electronics sources usually include touch sensor chips. Sparkfun offers a 12-key keypad on this basis, and also a 9-key keypad designed as a “touch shield” for use with the Arduino. Both of the Sparkfun products are shown in Figure 13-3. They include the same MPR121 touch sensor chip as the breakout board from Adafruit shown in Figure 13-2, and require an I2C connection with a microcontroller.

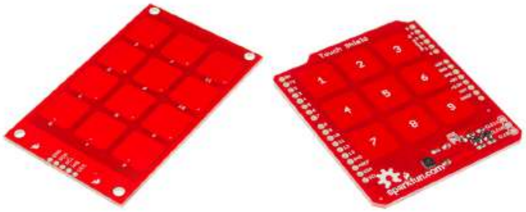

Figure 13-3 *Two capacitive keypads from Sparkfun, the one on the right designed as an Arduino shield.*

Because a capacitive touch pad is usually mounted inside an enclosure, the appearance of the bare touch pad as a component is unimportant. The outside of the enclosure can be printed with a design showing key outlines.

Individual Touch Pad

Adafruit sells the AT42QT1010 touch pad that emulates a momentary switch. Its output transitions from logic-low to logic-high when a finger presence is detected, and transitions back to logic-low when the finger is removed.

An alternative, shown in Figure 13-4, toggles between a logic-high and logic-low output and latches in each state with each single key press.

Both of these keypads contain sensor chips to generate the output.

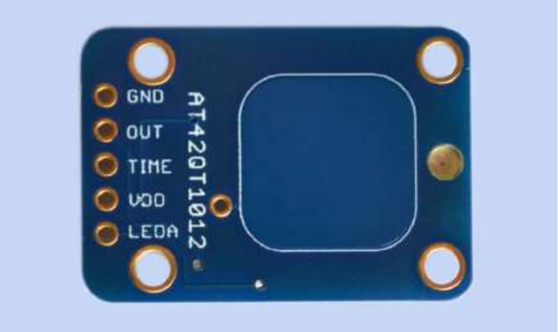

Figure 13-4 *The output from the AT42QT1012 sensor chip on this touch pad from Adafruit toggles between logic-high and logic-low each time the pad is touched.*

Wheels and Strips

A touch wheel uses a circular pattern of conductive traces, often referred to as *electrodes*, to receive finger input. A simple configuration is shown in Figure 13-5. The traces interlock without making contact with each other, so that moving the finger in a circular motion creates a capacitive input that rises and falls sequentially on each element. In the figure, three sections are used, each colored differently here for purposes of clarity. Other touch wheels may contain more sections.

Firmware that is designed to interact with wheel-shaped touch pads will generally assume that two electrodes are receiving input at one time. The firmware attempts to calculate the position and motion of the finger by assessing the relative capacitance of adjacent segments. Ideally the capacitance values should vary linearly and complementarily; that is, from a 50-50 value at a midpoint between two segments, the progression should change to 60-40, 70-30, and so on, as the finger moves around the wheel.

A *touch potentiometer* consists of multiple touch pads, usually arrayed in a straight line. This may be described as a *touch strip*. An example is made by GHI Electronics and sold by Robot Shop as their L12 Capacitive Touch Module, shown in Figure 13-6.

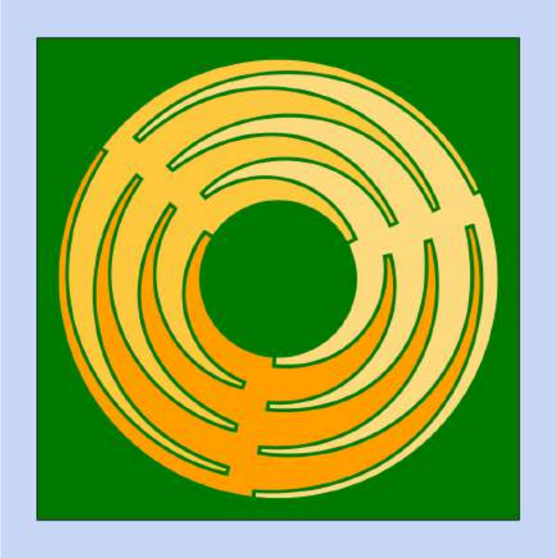

Figure 13-5 *A touch wheel created by copper traces on a circuit board (green). Each section is colored differently, for purposes of clarity.*

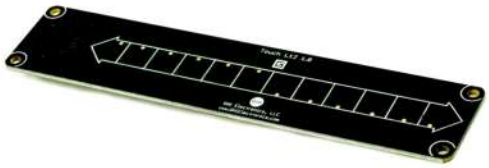

Figure 13-6 *Multiple capacitive touch pads arrayed as a strip, made by GHI Electronics.*

Design Considerations

A simple touch pad is often surrounded by additional copper that is grounded and may be described as a *shield* or a *guard*. The capacitance of a fingertip (spaced above the electrode by a layer of plastic or glass that functions as the dielectric) interferes with the field between the electrode and the shield.

The underside of the circuit board is often plated and grounded to protect the touch pad from electromagnetic interference. This, too, can be referred to as a shield or a guard. The ground plating can be in a hatched pattern to reduce its capacitance with the electrode above.

Layout of a circuit involving a single touch sensor chip can affect its performance significantly. A touch sensing device may have to be “tuned” to detect finger presence reliably.

A long trace between the touch sensor and an electrode will tend to pick up noise and will increase capacitance.

The distance between adjacent traces from multiple touch pads must be maximized to reduce capacitance between them. If output from a sensor chip uses the I2C or SPI digital protocol, any trace carrying that digital signal should be at least 4mm from input traces. If they cross, they should be at 90 degrees to each other.

Electrodes should not be shaped to resemble numbers or other characters printed above them. A single basic electrode should be circular.

What Can Go Wrong

Insensitive to Gloves

Gloves are a challenge for touch-sensor design, as they alter the dielectric and the distance between the electrode and the finger. Capacitive touch sensors may not work at all with some types of gloves. However, special gloves containing metallic threads are available.

Stylus Issues

A nonconductive stylus cannot activate a touch pad.

Conductive Ink

Ink that prints the shapes of touch pads on the exterior of a device should be nonconductive.

touch screen

14

The term **touch screen** is written as two words in this Encyclopedia. In many sources, the words are concatenated as “touchscreen.”

The two words are hyphenated here when they function as an adjective, but not otherwise. In manufacturers’ datasheets where the two words are used, they are not usually hyphenated.

OTHER RELATED COMPONENTS

- **single touch** sensor (see Chapter 13)
- **force** sensor (see Chapter 12)

What It Does

A touch screen is a video display with embedded touch sensing. The screen reports the position of the touch, and is used as a pointing device as an alternative to a mouse or trackpad. Some touch screens report pressure as well as position.

Touch screens are widely used in smartphones and tablets, and also in some laptop computers. Smaller, simpler touch screens may be found in office equipment such as photocopiers.

Schematic Symbol

No specific schematic symbol is used to represent a touch screen.

Variants

Early designs used infrared LEDs recessed into the edges of a frame around a screen. A matching photodiode picked up the focused beam from each LED. The presence of a fingertip was detected when it interrupted one or more of the beams. This system was not capable of high resolution, but was adequate for detecting user input at predefined locations.

Most touch screens currently are either resistive or capacitive.

Resistive Sensing

A resistive touch screen consists of two transparent layers that can be installed over a separate video display.

Each of the layers has uniform electrical resistance. Pressure from a fingertip on the outer layer (which we can refer to as layer 1) forces it to make contact at a point with the inner layer (referred to here as layer 2).

Two vertical electrodes connect with layer 1 along its left and right edges. Two horizontal electrodes connect with layer 2 along its top and bottom edges. When voltage is applied between the vertical electrodes on layer 1, the layer acts as a horizontal voltage divider. The voltage at the point where the layer is being pressed is applied to layer 2, and can be read from either of the electrodes on layer 2, so long as the metering has a much higher impedance than that of the layer. The voltage is decoded as a value for the horizontal position on layer 1.

An external switching device now repeats the procedure, except that it applies voltage to layer 2, and reads it from layer 1 to supply a value for the vertical position on layer 2. The sequence is illustrated by the top and bottom sections of Figure 14-1.

Because only four connections are necessary, this is referred to as a *four-wire* resistive touch screen. Five-wire variants exist, but are less common, and are not included in this entry.

Advantages of a resistive screen include:

- Simplicity. Only four connections are necessary, and the layers of the screen do not have to be subdivided into separate conductors.
- Low cost, relative to capacitive touch screens.
- Will respond equally well if a user wears gloves or uses a stylus.

Disadvantages of a resistive screen include:

- Some resistive versions require a stylus input instead of finger pressure.
- Resistive screens only respond to a one-location input. Two-finger gestures are not supported.
- Contact bounce occurs when the flexible layer is pressed against the underlying layer, and voltage spikes may be associated with switching power to the screen. To address this issue, firmware in a microcontroller may have to take a median value from several rapid readings.
- The flexible membrane is vulnerable to damage from sharp objects.

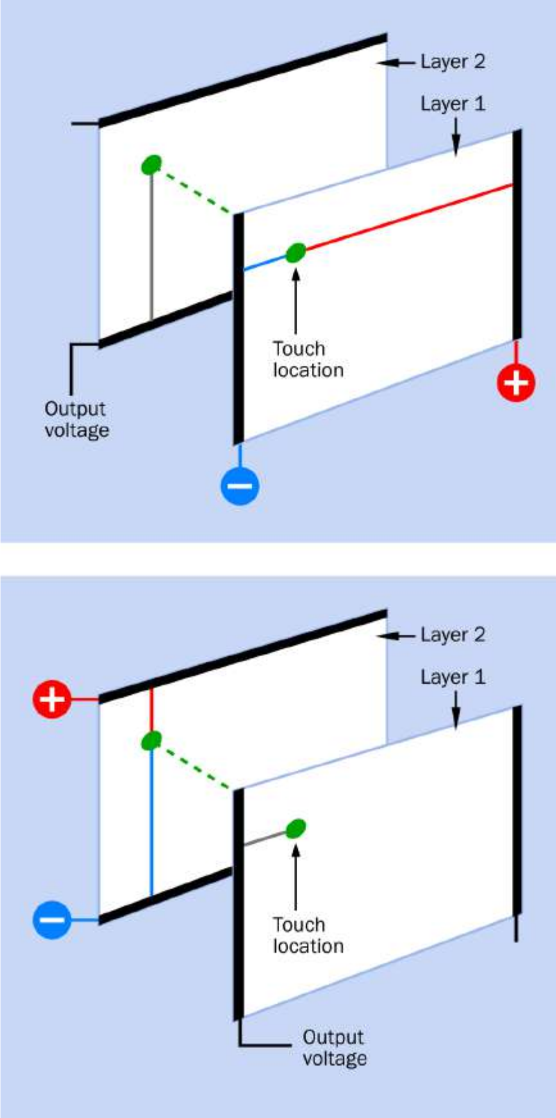

Figure 14-1 *Sections of a resistive touch screen are shown displaced for clarity. In reality they would be separated by a very small gap, allowing one section to make contact with the other in response to finger pressure.*

Capacitive Sensing

A capacitive touch screen can consist of an array of **single touch** sensors printed onto a glass panel as vertical and horizontal lines of transparent conductive ink.

Alternatively, a small capacitive screen can measure the tiny amount of current drained by a fingertip from four sources located at corners of the screen.

For more information about capacitive touch sensing, see the entry on touch sensors in Chapter 13.

Screens Available as Components

A wide variety of screens can be found in diagonal sizes ranging from 2 inches upward. Onboard electronics can be suitable for connection with a microcontroller using the I2C and SPI protocols, or USB. Different screen resolutions are available.

An example of a 3.5-inch touch screen with 320x200 resolution, mounted on a breakout board that can be used with a breadboard and an Arduino, is shown in Figure 14-2.

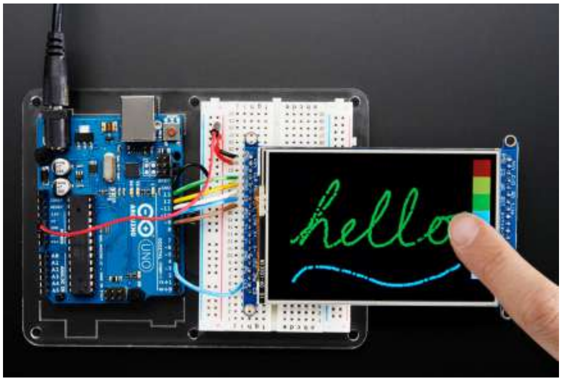

Figure 14-2 *An Arduino-compatible touch screen mounted on a breakout board, available from Adafruit.*

A 7-inch resistive touch screen that can be mounted on a separate video display is shown in Figure 14-3. It can be used with the STMPE610 controller chip, which converts resistive screen values into digital coordinates and can be accessed by a microcontroller over both SPI and I2C. This surface-mount chip is available on a breakout board.

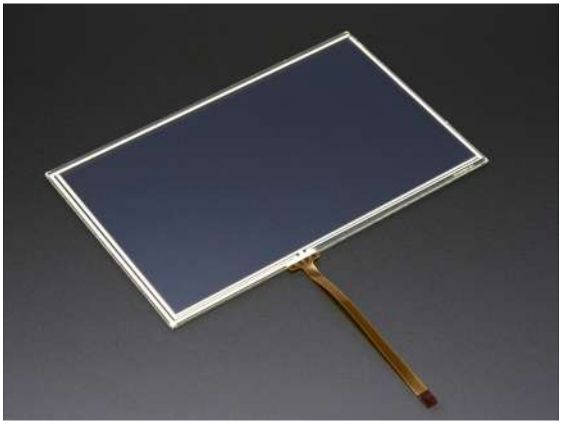

Figure 14-3 *This resistive touch screen is intended for use as a layer applied to a 7-inch video display. The screen is available from Adafruit.*

When choosing a touch screen as a component for a DIY project, the availability of microcontroller code libraries for reading and refreshing the display is an important consideration.

liquid level sensor

15

Level indicators that contain no electronic components are not included in this entry.

Specialized industrial level-sensing equipment is generally outside the scope of this Encyclopedia. This entry discusses small-scale, lower-cost sensors.

Liquid volume can be assessed by measuring liquid pressure at the bottom of a reservoir. Sensors for this purpose are discussed in the entry describing **gas/liquid pressure** sensors. See Chapter 17.

OTHER RELATED COMPONENTS

- **liquid flow rate** sensor (see Chapter 16)
- **gas/liquid pressure** sensor (see Chapter 17)

What It Does

Measuring the volume of liquid in a storage vessel or reservoir is such a fundamental task, countless methods have been devised, of which only the simplest and most common will be discussed here.

A liquid level sensor can have a binary output, meaning that it signals when the volume rises above or falls below a level that can be preset or reset. Often the sensor will be connected to a pump or valve that maintains a relatively constant volume in a container.

Alternatively a sensor can indicate the actual stored volume, either with an analog output or in digital increments.

Schematic Symbols

Three variants of a schematic symbol for a simple liquid level sensor are shown in Figure 15-1. They are not always used, however, and a sensor may be shown simply as an annotated switch.

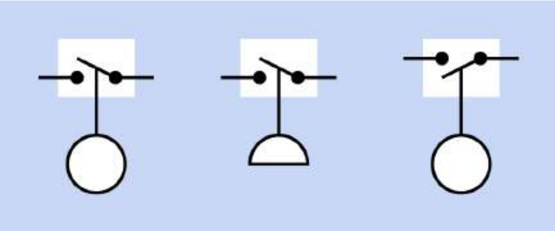

Figure 15-1 *Three variants of a schematic symbol to represent a liquid level sensor. The rightmost symbol indicates that a rising level closes, rather than opens, a switch.*

Applications

The fuel gauge in a vehicle is one of the most commonly encountered applications of a liquid level sensor. The water tank in a recreational vehicle or boat may use similar electronics. In industry, the choice of a sensor will be influenced by the type of liquid that is being stored, the desired accuracy, the temperature range, and whether the storage tank is sealed or open to atmospheric pressure.

How It Works

Desirable attributes of any liquid-level sensor include resistance to vibration, some damping to average out fluctuations caused by turbulence or sloshing in the liquid, resistance to chemical reactions with the liquid, and few moving parts that may require maintenance if the sensor is inside a sealed tank. Desirable attributes of an analog float sensor include a linear response, and some hysteresis if the application requires it.

This entry compares a variety of sensing strategies.

Binary-Output Float Sensor

The term "binary output" is used here to describe an output that only has two states (on and off, or logic-high and logic-low). The simplest type of liquid-level sensor with a binary output consists of a donut-shaped float that contains a permanent magnet and is free to slide vertically up and down a sealed tube containing a reed switch. The tube is supported on a bracket that can be mounted on the wall or the lid of the vessel containing the liquid.

The tube and float must be nonmagnetic, and the float must have a significantly lower specific gravity than the liquid that is being used ("significantly" because the float requires sufficient buoyancy to carry the weight of the magnet and overcome any friction between itself and the tube). A diagram illustrating this configuration is shown in Figure 15-2.

To change the level setting of the sensor, the bracket may be mounted on a screw thread to adjust its vertical position.

The reed switch can be normally open or normally closed, as needed to respond to a rising or falling liquid level. For basic information about reed switches, see "Reed Switch". Additional, detailed information about reed switches is included in the book *Make: More Electronics*.

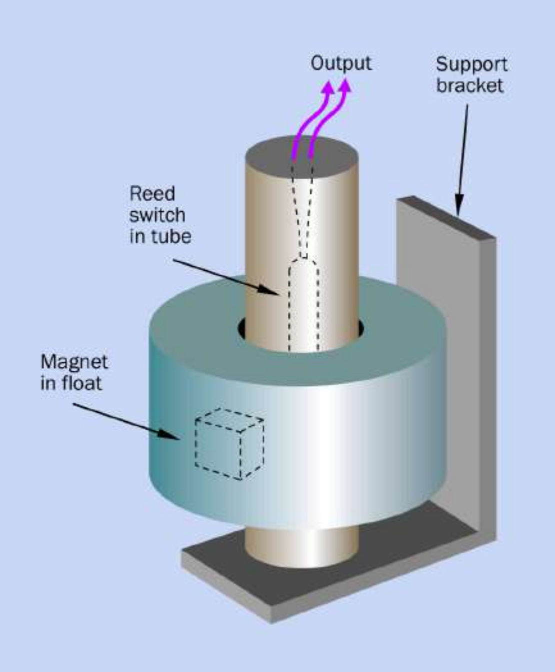

Figure 15-2 *The principal parts of a basic binary-output float sensor.*

For increased reliability, a Hall-effect sensor could be substituted for a reed switch. See "Hall-Effect Sensor" for general information about Hall-effect sensors.

Another binary-output float sensor is shown in Figure 15-3. This is a sealed plastic capsule containing a snap-action switch and a steel ball. The cable is attached to the underside of the top of a tank, and the capsule dangles into liquid in the tank. A separate weight (not shown) has a hole in the middle, and is threaded over the wire. The weight keeps the wire approximately in a vertical position as it dangles into the tank.

Figure 15-3 *An air-filled float that switches external power depending on its orientation.*

Figure 15-4 shows the components inside the float. When the liquid level in the tank falls, the float adopts the position shown on the left in the figure. The ball drops against a lever that closes a snap-action switch, which starts an external pump to replenish the tank. As the liquid level rises, the buoyancy of the air-filled float changes its orientation to that shown on the right in Figure 15-4. The ball drops and the switch opens, stopping the pump.

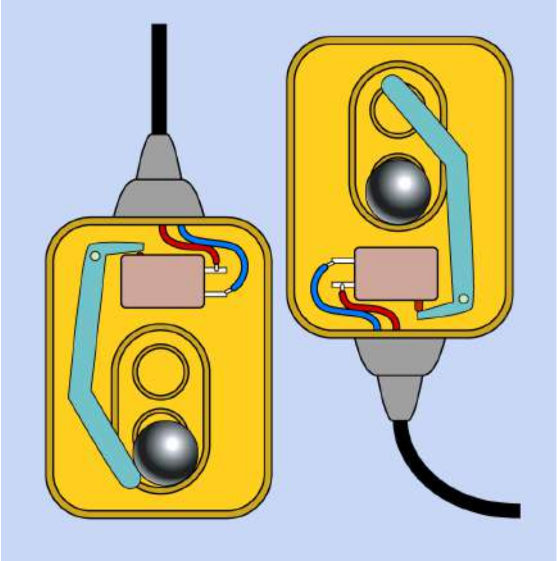

Figure 15-4 *Internal components of the float shown in the previous figure.*

Two circular indents on the inner surface of the plastic capsule prevent the ball from rolling erratically if there is turbulence in the liquid. They also provide some hysteresis.

Analog-Output Float Sensor

The simplest type of liquid-level sensor with an analog output consists of a float on an arm attached to a potentiometer, as shown in Figure 15-5. This very basic design was used in fuel tanks in vehicles for many decades. Disadvantages include a nonlinear response and the limited life expectancy characteristic of potentiometers. Some compensation for the nonlinear response can be made by using an analog fuel gauge with a nonlinear scale.

For more information about potentiometers, see "Arc-Segment Rotary Potentiometer".

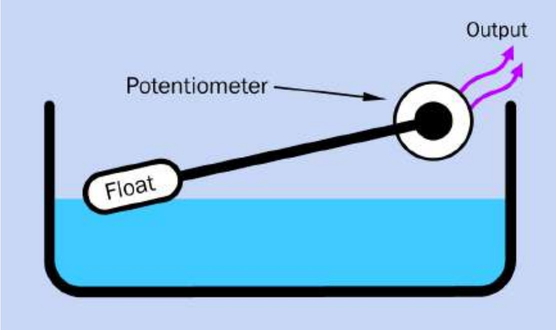

Figure 15-5 *A basic float sensor with an analog output.*

Incremental-Output Float Sensor

A schematic diagram for a simple float sensor with incremental output is shown in Figure 15-6. A magnet embedded in a donut-shaped float, similar to that shown in Figure 15-2, interacts with a sequence of reed switches installed in the central tube. The switches are spaced at equal intervals and apply power between resistors of equal value wired in series. This system has been used in motorcycle and automobile fuel tanks, where the switches may be enclosed in a (nonmagnetic) stainless-steel tube. The accuracy is limited by the number of reed switches.

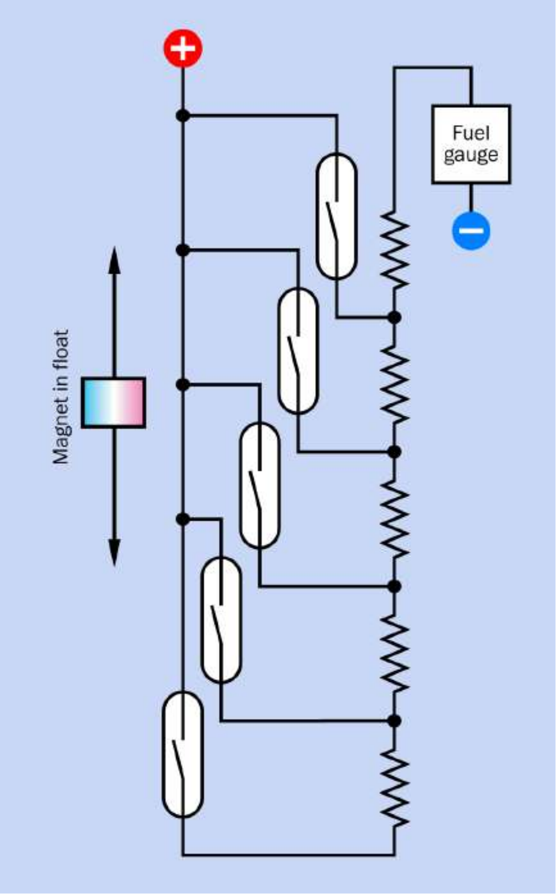

Figure 15-6 *A float sensor with incremental output.*

Displacement Level Sensors

If a heavy object, described as the *displacer*, is suspended in liquid, the effective weight of the object diminishes as the liquid rises up around it. This occurs because according to Archimedes' Principle, the upward buoyant force is equal to the weight of liquid that the object displaces. The displacer is suspended from a load sensor that measures its weight. Analog output from the sensor will be approximately linear with liquid level.

For a simplified diagram showing a displacement level sensor using this concept, see Figure 15-7.

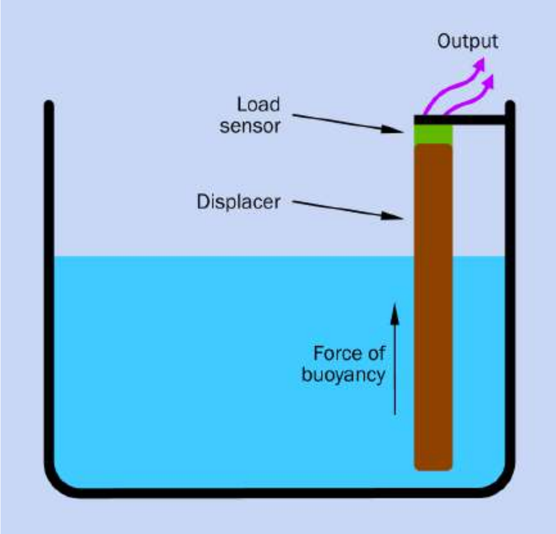

Figure 15-7 *A displacement sensor. The displacer is heavier than the liquid around it, but its effective weight diminishes as the liquid rises.*

For more information on load sensors, see Chapter 12.

Ultrasonic Level Sensors

An ultrasonic sensor can be used to measure the level of liquid in a reservoir, as shown in Figure 15-8. For more information about this type of sensor, see the entry discussing proximity sensors in Chapter 5. A disadvantage of using ultrasound for liquid level sensing is that the speed of sound will be affected by any vapor given off by a volatile liquid.

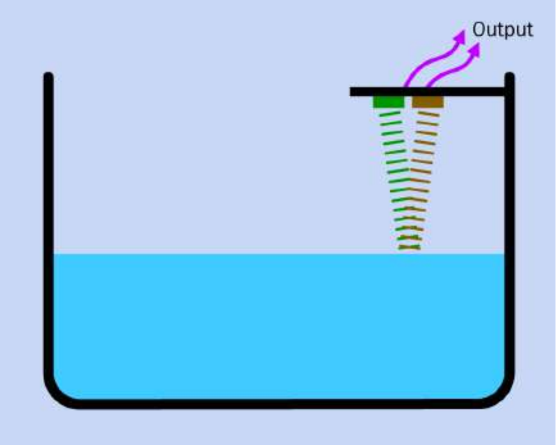

Figure 15-8 *An ultrasonic proximity sensor can measure the level of liquid in a reservoir.*

Reservoir Weight

The weight of a reservoir can be measured to assess the volume of liquid in it. This can be done by mounting the reservoir on load sensors. However, pipes leading to and from the reservoir must be designed so that they do not add or subtract any significant weight. Figure 15-9 suggests an arrangement, although the outflow will still change the weight to some extent depending on the amount of suction that is applied. For more information on load sensors, see Chapter 12.

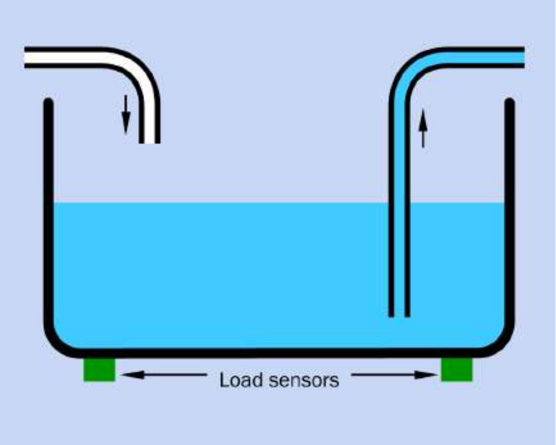

Figure 15-9 *Using load sensors to assess the weight of liquid in a reservoir requires that the weight of plumbing should not be imposed on the structure of the reservoir.*

Pressure Sensing

A differential pressure sensor can be added to a pipe near the bottom of a reservoir. The sensor measures the difference between the liquid pressure and ambient air pressure. See Figure 15-10.

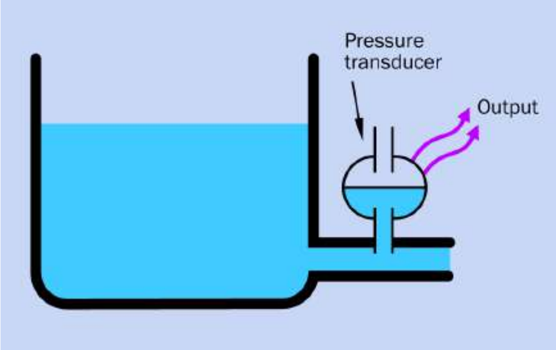

Figure 15-10 *A pressure sensor can assess the volume of liquid in a reservoir.*

This arrangement assumes that the reservoir is vented so that air pressure above the liquid level is equal to air pressure at the sensor. If the reservoir is not vented, a pipe must connect the reference port on the sensor with the space above the liquid.

The reservoir must have straight, vertical sides for the pressure to be directly proportional to the liquid volume.

Liquid volume can also be assessed by measuring the pressure inside a container, near the bottom. A submersible pressure sensor can be used, typically consisting of a watertight capsule fitted with a diaphragm that connects with an internal strain gauge. The sensor is lowered on a cable that also contains an air line. Because pressure in a liquid is affected by atmospheric pressure above the surface of the liquid, the sensor requires an air line so that its measurements are relative to the outside air.

Submersible pressure sensors are useful where access is limited—for example, when measuring fluctuations in an open-air municipal water reservoir.

What Can Go Wrong

Turbulence

Surface turbulence consisting of ripples, waves, or sloshing of liquid will tend to occur when a reservoir is refilled rapidly or is subjected to lateral movements as a result of being mounted in a moving vehicle. To minimize output fluctuations, some *damping* is desirable.

Baffles consisting of perforated plates inside the reservoir are a common strategy, as shown in Figure 15-11. In the upper section of the figure, lateral acceleration causes submersion of a float sensor in the reservoir. In the lower section, perforated baffles minimize the problem.

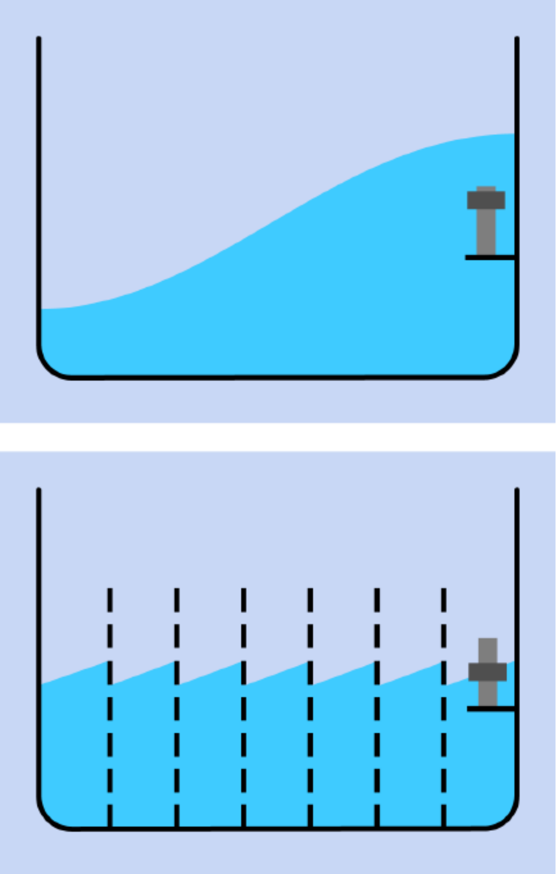

Figure 15-11 *Insertion of perforated baffles in a reservoir minimizes the sloshing that otherwise tends to occur when the reservoir is subject to lateral motion.*

Sensors that measure the weight or pressure of a liquid are less susceptible to turbulence. In a displacement sensor, the weight of the displacer provides a damping effect.

Tilting

All level sensors will tend to give inaccurate readings when a reservoir is tilted. A float sensor will be affected less if it is mounted centrally in a reservoir, because the reservoir will tilt around the sensor, as shown in Figure 15-12.

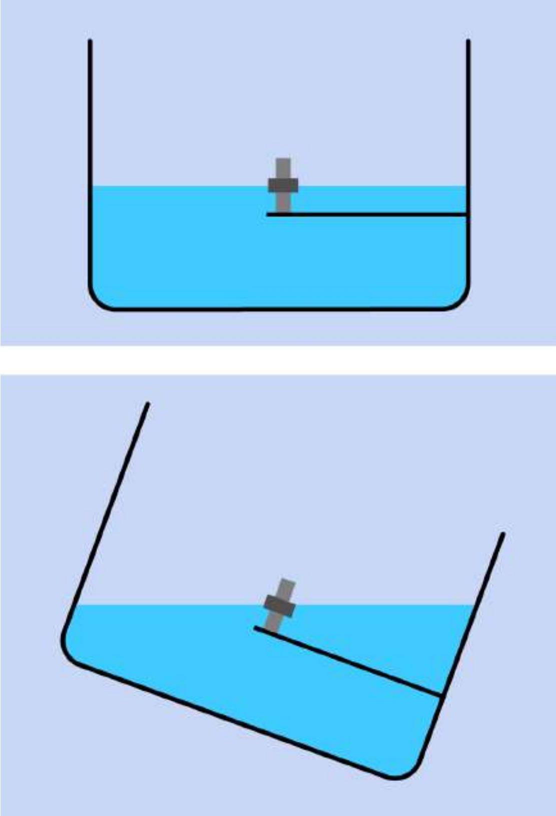

Figure 15-12 *If a float sensor is mounted centrally in a reservoir, it will be significantly less affected if the reservoir tilts.*

liquid flow rate sensor

16

Flow rate sensors that contain no electronic components are not included in this entry.

Some methods of liquid flow sensing can also be applied to gases, but sensors are usually designed for one application or the other. Therefore, gas flow sensors have their own entry. See Chapter 19.

Many liquid flow rate sensors are large devices designed for industrial applications. This entry focuses on lower-cost solid-state sensors.

OTHER RELATED COMPONENTS

- **liquid level** sensor (see Chapter 15)
- **gas/liquid pressure** sensor (see Chapter 17)
- **gas flow rate** sensor (see Chapter 19)

What It Does

A liquid flow rate sensor measures the rate at which liquid flows past or through the device. A water meter is an example of a flow rate sensor.

A sensor may have a binary output, meaning that it signals when flow stops or starts, or if its rate rises above or falls below a level that can be preset or reset. However, most flow meters have an analog output that varies with the volume per unit of time.

Measuring the flow rate of a liquid can be challenging if the viscosity is very high, the liquid is chemically reactive, or the rate is very low. Such factors may require specialized equipment that is outside the scope of this Encyclopedia.

This entry compares the most popular sensing strategies.

Schematic Symbols

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors. Typically they involve a single letter or an X in a circle. These symbols are not generally found in electronic schematics, and therefore they are not included here.

Paddlewheel Liquid Flow Rate Sensors

The simplest and most common liquid flow rate sensor uses a *paddlewheel*, also referred to as a *rotor*, that is mounted with its axis of rotation at 90 degrees to the direction of liquid flow. An example is the Koolance INS-FM16 shown in Figure 16-1. This sensor is intended for use in an aftermarket cooling system for the CPU in an overclocked desktop computer, but can be used in any system where the rate of flow ranges from 0.5 to 15 liters per minute. The paddlewheel has a pair of magnets mounted in it, activating a *reed switch* that is mounted in a sealed enclosure beneath the wheel. (For more information regarding reed switches, see "Reed Switch" in the entry describing **object presence** sensors.)

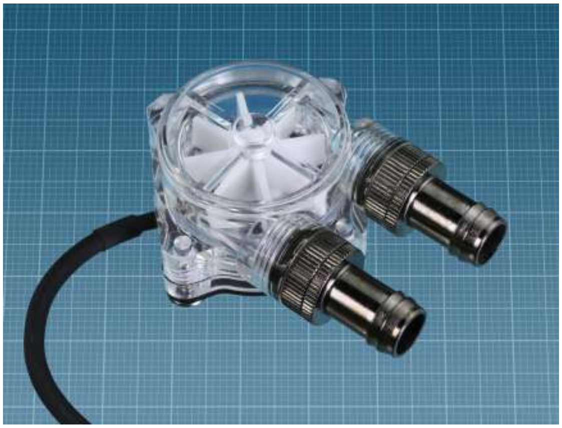

Figure 16-1 *A low-cost, simple paddlewheel sensor designed for rates of 0.5 to 15 liters per minute. The background grid is in millimeters.*

While the reed switch in the Koolance flow sensor inevitably suffers from contact bounce, it has the advantage of simplicity and can be used in conjunction with appropriate hardware or microcontroller code to debounce the pulse stream.

As in any device with a rotating part, friction and wear will afflict a paddlewheel sensor, especially because the bearings are often in the chamber through which the liquid passes. This eliminates the possibility for roller bearings. Typically a plain bearing is used, consisting of a pin that engages in a hole in the casing. Friction wears the bearing surface, creating a larger gap, which allows the rotor to vibrate or bounce instead of spinning smoothly. Flow resistance increases and accuracy is degraded.

In modern designs, the mass of the rotor is minimized to reduce the friction. Also, if the shaft is horizontal (but still at 90 degrees to the direction of the flow), buoyancy of the rotor in the liquid can reduce friction still further, as suggested in Figure 16-2. Ideally, the density of the rotor and the density of the liquid will be the same.

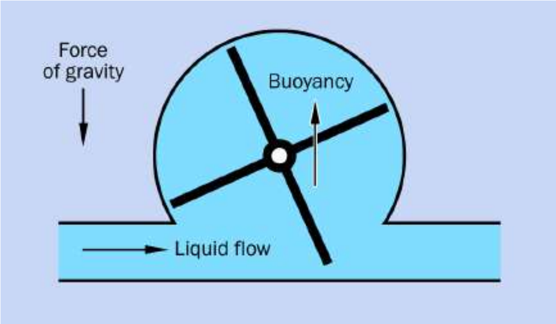

Figure 16-2 *In this configuration, friction on the rotor bearings is mitigated by taking advantage of the buoyancy of a low-density rotor in the liquid that passes through.*

The U-shaped flow path in the Koolance sensor in Figure 16-1 maximizes the responsiveness of the rotor, but an inline path is more common. An example is shown in Figure 16-3.

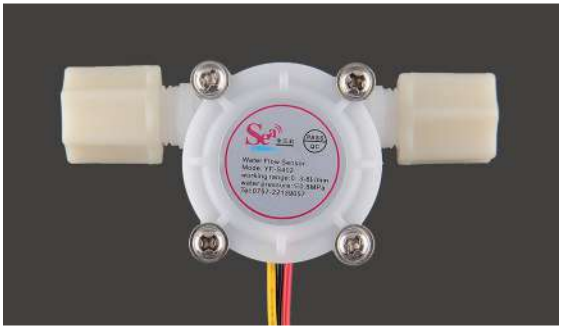

Figure 16-3 *An inline flow sensor rated for 3 to 6 liters per minute.*

Turbine Flow Rate Sensors

In a turbine-type sensor, two or more spiral blades are attached to a hub that rotates around an axis in line with the liquid flow, as shown in Figure 16-4. A magnet in each spiral blade triggers a reed switch or Hall-effect sensor mounted in a bracket that suspends the turbine from the interior walls of the tube.

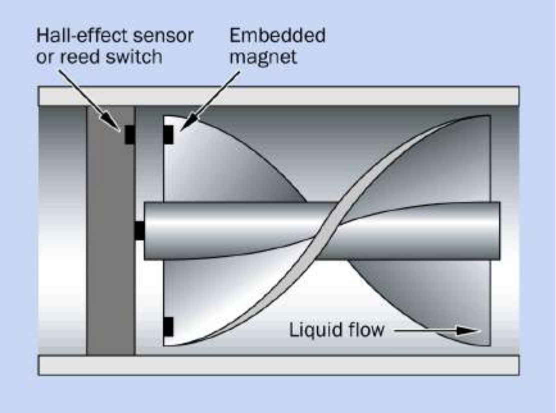

Figure 16-4 *Simplified view of a turbine flow rate sensor mounted inside a tube.*

The support bracket often consists of four struts, which impose resistance to liquid flow. The bearings suffer from the same kinds of problems as the bearings in a paddlewheel sensor, and must withstand additional load resulting from the inline force exerted by flow. Overall, while the turbine type of sensor is popular in laboratory equipment, it has disadvantages that are not shared by the bulkier paddlewheel type.

Limitations of Paddlewheels and Turbines

Both the paddlewheel and turbine types of sensors require a minimum flow to overcome the friction in their bearings. Below this minimum, liquid will find its way around the rotor without turning it. Even when the flow exceeds the minimum, response of the rotor is likely to be nonlinear as a result of turbulence and other factors.

Above a limit stated by the manufacturer, turbulence increases to the point where output from the sensor is no longer meaningful. Wear on the bearing will also increase with flow rate.

A significant problem for these types of sensors is that they do not respond well to sudden variations in flow. The paddlewheel, in particular, has inertia as a function of the diameter of the rotor, and will take some time to spin up in response to an increase in flow. Conversely, when the flow diminishes, the paddlewheel will tend to overrun.

The viscosity of the liquid passing through a paddlewheel or turbine sensor will have a very significant effect on its performance.

Thermal Mass Liquid Flow Rate Sensor

The thermal-mass system is commonly used when volumes are extremely low. The system is illustrated in Figure 16-5. A tube containing liquid is fabricated from a heat-conductive metal such as aluminum. It is enclosed in a larger tube, and the gap between them is filled with thermal insulation. A temperature sensor such as a thermistor measures the temperature of liquid entering the system. A second sensor, combined with a small resistive heater in the form of a coil around the tube, is placed downstream. Liquid passing through the tube will tend to remove heat more effectively at higher flow rates, and the difference in temperature between the two sensors is a logarithmic function of the flow rate.

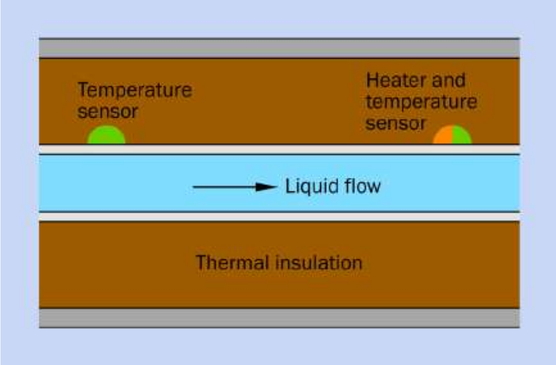

Figure 16-5 *In this type of low-flow sensor, the temperature differential between the two sensors is a logarithmic function of the flow rate.*

Variants of this system use slightly different tube configurations and sensor placement, but the principle is the same. Its advantages include the lack of any moving parts, and the avoidance of any probes intruding into the liquid, which is desirable in biochemical and medical applications.

The same principle is applied in many **gas flow rate** sensors. See "Mass Flow Rate Sensing".

Sliding Sleeve Liquid Flow Switch

This sensor is used in some domestic systems, where flow-activated water heating is required. A vertical section of brass (nonmagnetic) water pipe contains a sliding inner sleeve that incorporates a magnet. An external reed switch is activated when the sleeve is moved by water flowing through it. When the flow stops, the sleeve is returned to its rest position by the force of gravity.

Sliding Plunger Liquid Flow Switch

Figure 16-6 shows an exploded view of a similar device, using a plastic plunger that slides inside a nylon plumbing fixture designed for 3/4-inch pipe. The plunger contains a magnet and is restrained by a compression spring and a circular perforated plate. When water flow is sufficient to overcome the resistance of the spring, the plunger slides far enough for the magnet to activate a reed switch sealed into the external housing.

Ultrasonic Liquid Flow Rate Sensor

This type of sensor passes ultrasound through a liquid in a pipe. The speed of sound through the liquid is affected by the flow rate, and external electronics translate this lag time into a value for volume-per-minute. The system adjusts for variations in temperature that also affect the speed of sound.

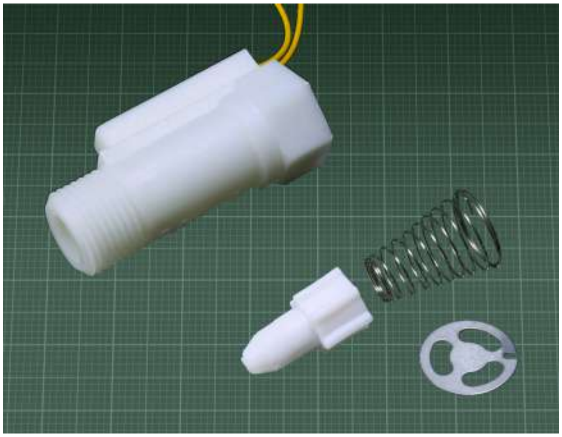

Figure 16-6 *Parts of a flow switch. The small plunger is inserted into the pipe and retained with the perforated plate and compression spring.*

Various configurations are available, some allowing ultrasound sources and detectors to be clamped to the outside of a pipe, as shown in Figure 16-7. To eliminate other variables, one ultrasound pulse is transmitted in the same direction as the flow, followed by another pulse contrary to the flow, and the difference between the two transmission times is used as an indicator of the flow rate.

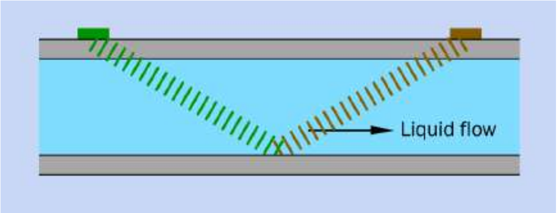

Figure 16-7 *Some ultrasound flow sensors are designed to be clamped externally to a pipe.*

Magnetic Liquid Flow Sensor

A magnetic field is induced in a metal pipe by a coil generating its field perpendicular to flow. The inside of the pipe is lined with nonconductive material in which two electrodes are mounted. Because water containing ions is conductive, the flow of water through the magnetic field induces a small potential difference between the electrodes. This voltage can be used as an indication of the flow rate.

To eliminate external factors that would also affect the potential difference, the polarity of current through the coil around the pipe alternates rapidly. The induced field remains the same regardless of the direction of the current.

A magnetic flow sensor should not be confused with a magnetic flow *switch.* Various types of switches are made, including large, heavy-duty devices where flow moves a magnet that triggers a shutoff valve. This type of industrial device is outside the scope of the Encyclopedia.

Differential Pressure Liquid Flow Meter

In this system, a pipe contains a perforated plate or some similar constrictor that partially obstructs the flow of liquid. Pressure is measured by a pair of pressure transducers placed before and after the constrictor. The pressure difference is an indicator of flow, because it increases as the flow rate increases.

This system was developed originally for large industrial applications but has been miniaturized and etched into silicon to measure very small flow rates. The Omron D6F-PH is an example, measuring less than 3cm square. It contains digital correction to enable a close-to-linear output. Because of its small size, it can be used only for slow flow rates, or as a bypass sensor. The general concept of a bypass sensor is illustrated in Figure 19-7, in the entry discussing gas flow sensors.

What Can Go Wrong

Vulnerability to Dirt and Corrosive Materials

MEMS liquid flow rate sensors containing very delicate, very small sensing elements are very vulnerable to contamination with dirt. Liquids should be filtered to minimize this risk. A manufacturer’s datasheet should provide information about the use of corrosive or chemically active liquids.

gas/liquid pressure sensor

17

Pressure measurement devices that do not contain any electronics, such as a nondigital tire-pressure gauge or a mercury manometer, are outside the scope of this Encyclopedia.

Many pressure sensing methods can be used with both gases and liquids. To avoid duplication, gas pressure sensors and liquid pressure sensors do not have separate entries. Both types are described here.

This entry deals almost exclusively with MEMS components. It does not include pressure measurement devices sold as industrial products.

Some manufacturers and vendors use the term *pressure sensor* to describe a component that measures mechanical load or force. In this Encyclopedia, the term only describes components that measure liquid or gas pressure. Mechanical load cells and force sensors will be found in the section on **force** sensors. See Chapter 12.

OTHER RELATED COMPONENTS

- **liquid level** sensor (see Chapter 15)
- **liquid flow rate** sensor (see Chapter 16)
- **gas flow rate** sensor (see Chapter 19)

What It Does

A pressure sensor measures the force exerted by a gas or liquid, often in a container or pipe.

Static pressure is measured under conditions that change slowly or not at all. *Dynamic pressure* is subject to fluctuations. Pressure sensors tend to be designed for one condition or the other.

Schematic Symbols

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors, including pressure sensors. Typically they involve a single letter or an X in a circle. These symbols are not generally found in electronic schematics, and therefore they are not included here.

Applications

Barometric sensors are found in barometers and weather stations. Altimeters are really a specialized form of barometric sensor, used in airborne vehicles. Gas pressure sensors have many industrial applications, and are used to monitor tire inflation in vehicles and the output from air compressors. They may also measure liquid pressure indirectly, as in a blood-pressure cuff.

Liquid pressure sensors are widely used to measure oil pressure in automobile engines and hydraulic braking systems. They have medical applications, and monitor water pressure in municipal systems and pumped supplies.

Design Considerations

As a result of random molecular movement, gases will tend to disperse to fill a container. After a gas reaches equilibrium, the pressure will be almost equal in all directions, affected only slightly by gravity. Where a gas is in a sealed, rigid container, pressure will vary linearly with temperature.

A liquid will tend to accumulate at the bottom of any container under the force of gravity. Liquid pressure in a container will be highest at the bottom, because of the weight of liquid above it. However, because almost all liquids are not easily compressible, they transmit force from any point in a container to any other point, including the sides of a container.

Units

Pressure is measured as force per unit area, which can be expressed in a confusing variety of units.

In the United States, gas and liquid pressure are still often expressed in pounds per square inch, abbreviated as PSI, or more often as psi, or sometimes as lb/in^2.

In standard international (SI) units, 1 bar of pressure is approximately equal to atmospheric pressure at sea level. Millibars are popular among meteorologists, 1 millibar being 1/1000 of a bar, equivalent to 100 pascals, where 1 pascal = 1 newton per square meter.

A bar is equivalent to 14.504 psi.

Blood pressure is measured in millimeters of mercury, because mercury manometers were used for this purpose originally. Atmospheric pressure may also be measured in millimeters of mercury, because the earliest barometers used a tube containing mercury. In the United States, some sources still refer to inches of mercury.

How It Works

Pressure sensing usually entails three stages.

1. A *sensing element* translates pressure into the mechanical displacement of a flexible part.
2. A *transducer* converts mechanical displacement into an electrical effect, either modifying resistance or creating a small voltage or current.
3. Electronics are used for *signal conditioning*. This may entail modifying a nonlinear signal, or may convert an analog output to a digital output.

As pressure sensors are increasingly taking the form of MEMS devices, all three stages may be combined in one silicon chip.

Basic Sensing Elements

Figure 17-1 shows four types of sensing elements that are used, or have been used, to convert pressure into mechanical motion. In each case, a green arrow shows where gas or liquid is introduced under pressure. 1: A Bourdon tube flexes under pressure, increasing its radius. The tube is hollow, open at one end and sealed at the other. 2: A coiled Bourdon tube uncoils partially under pressure, causing the top end to rotate. 3: A simple flat diaphragm. 4: A ribbed diaphragm.

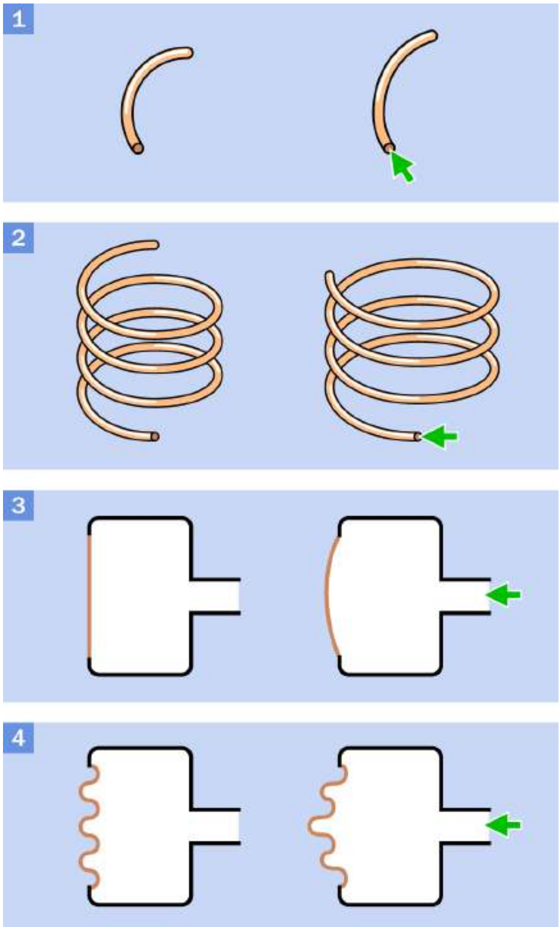

Figure 17-1 *Sensing elements. See text for details.*

The sensing elements numbered 1 and 2 are less likely to be used in modern systems, although a coiled Bourdon tube may be used to turn a potentiometer in low-cost oil-pressure sensing designs. Number 3 is well suited to MEMS devices, as it can be etched into silicon using the principle illustrated in Figure 17-2.

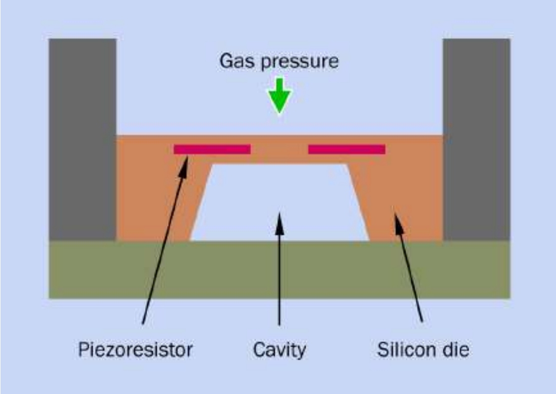

Figure 17-2 *A section of a flat-diaphragm pressure sensor, viewed from the side. An aperture in the top face of the chip allows air to enter. The sensor is etched into silicon, and the deflection of the wafer is measured using embedded piezoresistors.*

Relative Measurement

Pressure is a relative measurement. It is expressed relative to a *reference pressure* of some kind. Three types of measurement are commonly used:

1. *Absolute pressure*, relative to the zero value of a vacuum.
2. *Gauge pressure*, relative to ambient pressure (that is, pressure in the environment around the sensor). Air pressure is the reference source for gauge pressure, and a *vent* is incorporated into the sensing system.
3. *Differential pressure*. In this case, the pressure being measured is relative to some other pressure—for example, the pressure differential between two sealed tanks.

Figure 17-3 illustrates these three measurement types.

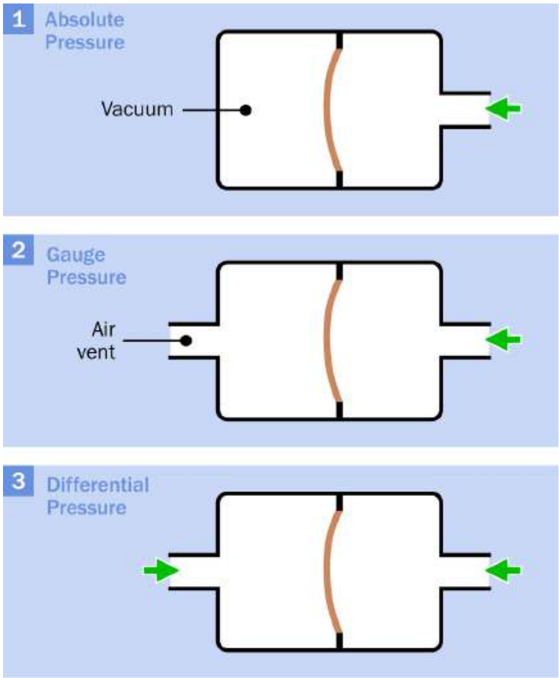

Figure 17-3 *Three ways in which pressure is measured.*

Variants

Ambient Air Pressure

A *barometric* sensor measures the pressure of air around it. This is an absolute value, relative to a vacuum.

The earliest barometers consisted of a tube sealed at one end, containing mercury. No air was allowed to enter the tube. It was inverted so that its open end was immersed in a small reservoir that was open to the air, the pressure of which supported the column of mercury in the tube. Thus, the height of the column was ratiometrically equivalent to atmospheric pressure. Because low air pressure is often an indicator of inclement weather, a barometer was a very simple tool for weather forecasting.

A modern barometric sensor consists of a chip with a vent hole in its upper surface, allowing ambient air pressure to reach a sensor inside the chip. A popular example has been the Bosch BMP085, shown soldered onto a breakout board in Figure 17-4. The BMP180 has a very similar specification, the primary difference being that it is designed for a SPI bus instead of I2C. For additional details about protocols such as SPI and I2C, see Appendix A.

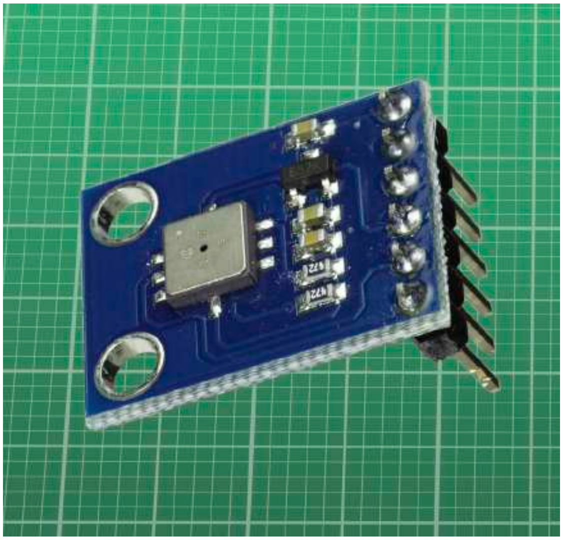

Figure 17-4 *A Bosch barometric sensor on a breakout board. The background grid is in millimeters.*

This sensor uses a supply voltage ranging from approximately 2VDC to 3.5VDC, but a breakout board includes a voltage regulator that tolerates a 5VDC supply. The output is digitized for access by a microcontroller, but the format is quite complex and consists of raw data that must be converted to air pressure by applying a formula. The manufacturer offers free code in the C language for this purpose. A thermometer that is built into the chip enables temperature compensation.

Subsequent products from Bosch include the BME280, which adds a humidity sensor. Breakout boards using this and other barometric chips are available from sources such as Sparkfun and Adafruit, where Arduino code libraries to interpret the data are available for download.

Altitude

A barometric sensor can be used to determine altitude. At sea level, air pressure is approximately 101kPa (kilopascals), or 760 millimeters of mercury. At 5,000 meters altitude, air pressure drops to 56kPa. The transition is nonlinear, changing most rapidly at lower altitudes.

Air pressure will be affected by temperature and weather conditions. Weather is unlikely to affect the value by more than plus-or-minus 1%, and temperature has a smaller effect; but the barometric sensor can be reset to a zero point based on current temperature and weather data.

Gas Pressure

Thousands of board-mountable gas pressure sensors are available from retail component suppliers. The need to make connections via tubing imposes a minimum size limit on these parts. Many have a through-hole format, but even the surface-mount variants can be relatively large (e.g., 10mm × 10mm). Most tend to have barbed tubing ports, the "barbs" being ridges to retain push-on flexible tubing. Output may be analog or digital, the digital protocol being designed for I2C, SPI, or plain TTL serial buses.

These components are designed for pressures ranging up to 500kPa (kilopascals). Some manufacturers use psi in their specifications, while others use bars and millibars, and a few state gas pressures equivalent to inches of water.

Dual-ported gas sensors for differential pressure measurement can be used for gauge measurement (i.e., relative to ambient air) if one port is left open. Absolute pressure sensors have only one port, as they measure pressure relative to a vacuum maintained inside the chip, behind the diaphragm.

- Some gas pressure sensors can do dual duty as liquid pressure sensors, but others cannot, and datasheets must be consulted carefully, as online vendors may not include this information prominently.

The sensor in Figure 17-5 by All Sensors Corporation is designed for flexible tubing with 1/16" internal diameter (1.6mm). Its pin spacing is 0.1" (2.54mm), making it breadboardable. This sensor is intended for gases only, but has a "10-inch" water-equivalent rating, meaning that it can tolerate a maximum pressure equivalent to that of a 10-inch (25.4cm) column of water.

When supplied with up to 16VDC, an internal Wheatstone bridge circuit provides an analog output that varies ratiometrically by a few millivolts over the full range of measurable pressures (the exact specification depending on the power supply). The output may be amplified with an external op-amp. Although this is a differential sensor, it can provide gauge pressure if one of the ports is left open.

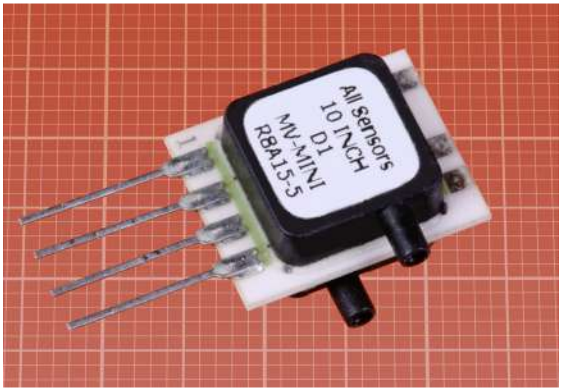

Figure 17-5 *A gas pressure sensor for use with 1/16" (1.6mm) internal diameter flexible tubing. The background grid is in millimeters.*

The larger sensor in Figure 17-6 is in the ADCA range from All Sensors Corporation, designed for flexible tubing with 1/8" internal diameter (3.2mm). This is intended for gases only, and has a "5-inch" water-equivalent rating. It requires a 5VDC power supply and has an internal op-amp that provides an output that varies ratiometrically by 0.2VDC (centered on 4VDC) over the range of measurable differential pressures. Like its smaller cousin, it can provide gauge pressure if one of the ports is left open.

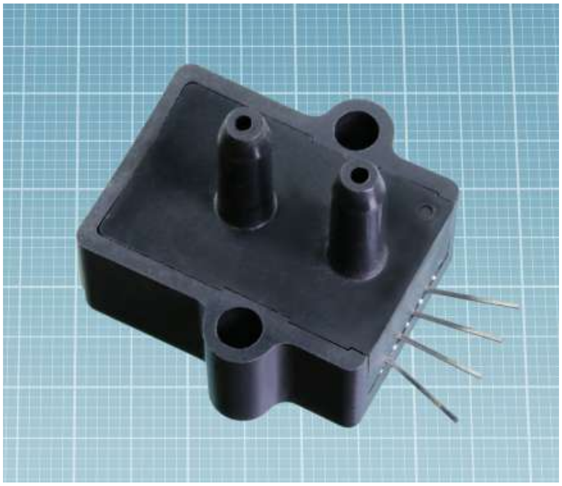

Figure 17-6 *A gas pressure sensor for use with 1/8" (3.2mm) internal diameter flexible tubing. The background grid is in millimeters.*

What Can Go Wrong

Vulnerability to Dirt, Moisture, and Corrosive Materials

MEMS pressure sensors containing very delicate, very small sensing elements must inevitably come into direct contact with gases and, in some cases, liquids. Datasheets will provide warnings regarding humidity and corrosive fluids, but the risk of contamination with dirt remains. Liquids should be filtered to minimize this risk.

Barometric sensors have a hole in the package, exposing the sensor element to the ambient air. This vent must be protected from direct contact with the environment.

Light Sensitivity

If light is allowed into the vent hole in a barometric sensor, it can create photocurrents in the chip. This will affect accuracy.

gas concentration sensor

18

Humidity sensors and *vapor sensors* are included in this entry, as they measure concentrations of a liquid while it is in its gas phase.

Sophisticated *industrial sensors* are available for accurate gas sensing, but this entry deals almost entirely with lower-cost solid-state sensors that are often classified as *board-mount* components.

OTHER RELATED COMPONENTS

- **gas/liquid pressure** sensor (see Chapter 17)

What It Does

Small semiconductor-based gas sensors provide a low-cost method for detecting specific gases in ambient air. The sensors vary their electrical resistance or capacitance in response to the concentration of the gas. They may be used in conjunction with an alarm system that will sound if gases such as *propane* or *carbon monoxide* exceed a preset level, or if the concentration of *oxygen* drops below a preset level.

Because vapor consists of a liquid in its gas phase, gas sensors can respond to vapor, as in the case of *alcohol sensors*.

A *humidity sensor* measures the amount of water vapor in the air, which can be important in refrigeration, HVAC (heating, venting, and air conditioning), medical equipment, meteorology, and storage rooms where art objects, antiques, or paper archives must be preserved in an environment that is neither too dry nor too moist, and is held at a constant temperature. Humidity control can be important also in climates where growth of mold is a concern. Humidity sensors are also used in automobile climate control and windshield defogging, and in the storage of food, fabrics, wood products, and medications.

High humidity coupled with high temperature contributes to the decomposition of many substances, while low humidity can cause desiccation. High humidity can also create damage if it causes materials to expand as a result of taking up moisture.

Many methods have been devised to detect individual gases, but semiconductor sensors are now dominant in applications where accuracy and gas-specificity are not crucial.

Schematic Symbol

No specific schematic symbol exists to represent a semiconductor gas sensor.

Semiconductor Gas Sensors

During the development of transistors in the 1950s, engineers noticed that semiconductor p-n junctions were sensitive to the presence of some gases in the atmosphere. This was regarded as a problem, which was solved by encapsulating transistors to prevent their exposure.

During the 1980s, Japanese law required installation of sensors in every home to detect hazardous concentrations of propane gas. This encouraged the development of cheap, long-lasting components that took advantage of semiconductor sensitivity.

Tin oxide is widely used in a variety of solid-state gas sensors. A sintered layer of the compound is deposited on a ceramic substrate in combination with other compounds such as antimony oxide. The granular layer functions as an n-type semiconductor in which electron transfer increases when certain gases are adsorbed among the grains. When the gas concentration diminishes, oxygen atoms displace the gas molecules, and the sensor returns to its original state. It is unimpaired by being activated, and has a life expectancy of at least 5 years during active (powered) use.

Each sensor includes a tiny resistive heater that is necessary for the chemical reaction to occur. Voltage must be applied to two pins that are connected with the heater. Two other pins connect internally with the sensing element. The resistance between these pins will vary with presence of the target gas; thus, this type of component has a resistive output.

It is a feature of semiconductor gas sensors that they tend to suffer from *cross-sensitivity*. That is, one sensor may respond to more than one gas. Manufacturers control this problem to some extent by adding filtering material around the semiconductor element, or by adjusting the proportions of dopants used in the semiconductor. Datasheets should be consulted carefully to see if a sensor may give false positives as a result of other gases that are likely to be present in the area where it will be used.

Figure 18-1 shows an MQ-5 propane sensor from Hanwei Electronics. The component also responds to methane, hydrogen, alcohol, and carbon monoxide, but with much less sensitivity. The manufacturer recommends calibrating the sensor by using a resistor ranging from 10K to 47K in series with the output resistance.

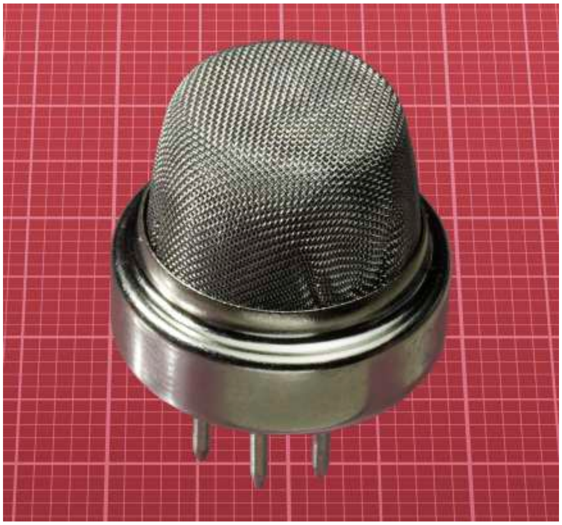

Figure 18-1 *A propane gas sensor. The background grid is in millimeters.*

Figure 18-2 shows an MQ-3 alcohol sensor, also manufactured by Hanwei. It has some sensitivity to benzene, but this is unlikely to be a problem, as benzene is seldom present in significant concentrations in ambient air. However, the response of the sensor to alcohol varies with temperature and humidity. Consequently this component can only be used as a "breathalyzer" if the user can be satisfied with an approximate response.

Semiconductor gas sensors are also available for detecting methane, carbon monoxide, hydrogen, ozone, and other gases.

These components are not low-current devices. Typically the internal heater in the Hanwei range has a resistance of slightly more than 30 ohms, and will draw 150mA to 160mA at 5V (i.e., slightly less than 1W). Because the heater is a simple resistive device, it can be used with AC or DC. The output resistance of the sensor can also be assessed with an AC or DC signal.

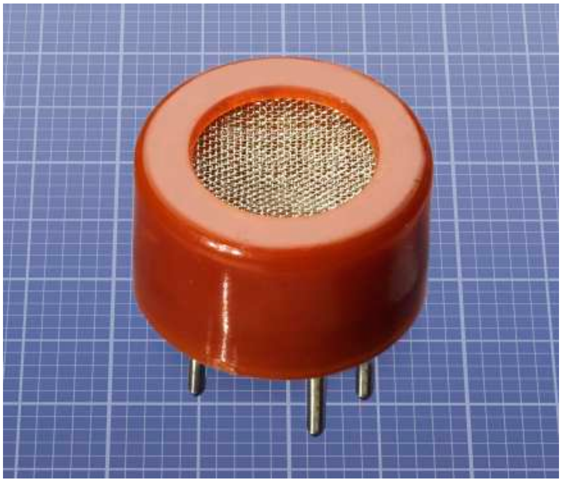

Figure 18-2 *An alcohol sensor. The background grid is in millimeters.*

A breakout board from Parallax simplifies the use of Hanwei gas sensors. The board is shown in Figure 18-3. It is compatible with carbon monoxide, propane, methane, and alcohol sensors, each of which can be plugged into a socket on the board. Two trimmers establish the sensitivity and the trip point for the sensing element, and the TTL output is then logic-high when a gas is detected, and logic-low otherwise.

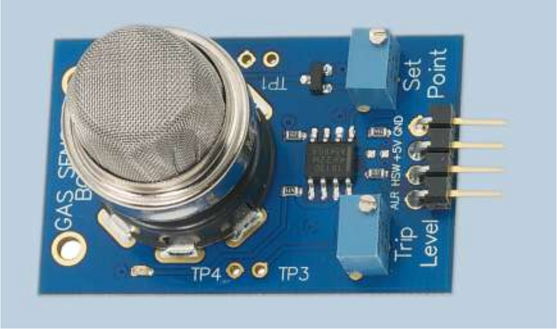

Figure 18-3 *A Parallax breakout board to simplify the use of Hanwei gas sensors.*

Oxygen Sensors

Oxygen sensors are often built using a membrane made of zirconium dioxide. This material has the property that it can transport oxygen ions when heated. One setup is to have a zirconium membrane separating the gas to be measured from ambient air. This is a type of fuel cell, called a concentration cell or Nernst cell. If the oxygen concentration differs between the two sides of the membrane, oxygen ions will flow through it. Only oxygen ions can move—not neutral oxygen atoms or molecules. The ions are negatively charged, so the transport will lead to a potential difference over the cell, which can be measured with platinum electrodes.

To comply with emissions regulations, automobiles sense the oxygen level of exhaust gases. This data controls the fuel-air ratio in the fuel injection system. Too much air results in the formation of nitrogen oxides, while too little air results in excessive carbon monoxide.

Humidity Sensors

Moisture content of the air is expressed in three different ways:

Absolute humidity

This is the weight of water vapor in a fixed volume of air. It is measured in grams per cubic meter in the metric system. An absolute humidity sensor is properly called a *hygrometer*.

Dew point

If a sample of air is cooled without a change of pressure, the dew point is the temperature at which moisture will start to condense. The dew point is a way of describing how humid the ambient air is currently, as water will condense more readily in humid conditions.

Relative humidity

This is often referred to by its acronym, RH. If temperature, pressure, and volume of a sample of air remain constant, relative humidity is the ratio between the current value of absolute humidity and the hypothetical level where the addition of water vapor would result in condensation. The ratio is expressed as a percentage. Thus, if moisture is already condensing in an air sample, relative humidity is 100%. If an air sample contains half the weight of moisture required for condensation to begin, relative humidity is 50%. If there is no moisture at all in the air, relative humidity is 0%.

Colloquial usage of the term “humidity” usually means relative humidity, and sensor output is usually convertible to this value. However, some absolute humidity sensors do exist.

Dew-Point Sensor

Historically, meteorologists used a *chilled mirror hygrometer*, in which a metal mirror was exposed to the atmosphere and cooled until the surface was seen to become misty with condensation. The temperature where this occurred was the dew point.

This system is still used in conjunction with an LED and a **phototransistor**. The LED is positioned so that its light reflects from a mirror, directly onto the phototransistor. The mirror is cooled until mist starts to form, which diffuses the reflected light and causes a well-defined change in output from the phototransistor.

Although the formula linking dew point with relative humidity is complex, a simplified approximation is available that is reasonably accurate so long as the relative humidity is 50% or greater. If RH is the relative humidity, t is the current temperature, and t_D is the dew-point temperature at which mist forms:

$$RH = 100 - (5 * (t - t_D)) \text{ approx.}$$

Although a chilled-mirror dew-point sensor has a reputation for being accurate, it is heavy, expensive, and impractical for most applications outside of meteorology.

Absolute Humidity Sensors

An absolute humidity sensor may use two NTC (negative temperature coefficient) **thermistors** in a Wheatstone bridge circuit. One thermistor is sealed in a compartment containing dry nitrogen, with zero humidity. The other is exposed to the atmosphere. Current passing through the thermistors raises their temperature to at least 200 degrees Celsius. Because heat radiates less efficiently when there is moisture in the air, the exposed thermistor will run hotter, and its resistance will be higher, at higher levels of humidity. This type of sensor is used in clothes dryers and wood kilns, among other applications.

See Chapters 23 and 24 for more information about thermistors.

Relative Humidity Sensors

Two main types of sensing elements are used to measure relative humidity: resistive and capacitive.

In a *resistive* sensing element, a thin layer of polymer, salt, or other hygroscopic substance is deposited on a substrate consisting of ceramic or other unreactive material. When the substance absorbs water, its electrical conductivity increases. Voltage is applied to the sensing element, and AC is used to avoid polarizing it. The current flow is processed externally, with conversion to DC followed by linearization, meaning that the current is processed to establish an almost linear relationship with gas concentration, with temperature compensation factored in. Alternatively, these functions can be performed by hardware built into the sensor, and a digital value for relative humidity can be accessed by an external microcontroller.

In a *capacitive sensor*, once again a thin film of polymer or metal oxide is deposited on a ceramic or glass substrate, but the film functions as a *dielectric* between two metal electrodes that serve as the plates of a capacitor. The dielectric value changes as the film absorbs moisture, and this causes the capacitance to vary, typically by 0.2pF to 0.5pF for each 1% change in relative humidity. This is almost a linear relationship extending over the entire range from 0% to 100% relative humidity.

The actual capacitance value at 50% relative humidity is likely to be between 100pF and 500pF. The sensor may be excited with AC from an external source, or can be incorporated in a chip that derives AC from a DC power supply and provides a digital output.

To determine the dew point or absolute humidity from a value for relative humidity, ambient temperature must also be measured. A chip such as the Si7005 from Silicon Labs includes a temperature sensor with a relative-humidity sensor based around a capacitor in which polyimide film forms the dielectric. If condensation forms, an on-chip heater will cause it to evaporate so that normal operation can resume. Data from the chip is supplied via an I2C interface.

Humidity Sensor Output

When the output is analog, resistance or capacitance of the internal sensing element is available via two pins or solder pads on the sensor. The analog value must be converted to a value for relative humidity by performing a calculation that takes the current temperature into account. The component may or may not include a temperature sensor.

With a digital output, an internal analog-to-digital converter can be accessed by a microcontroller via serial, I2C, or SPI interface. Alternatively, the sensor may communicate values using pulse-width modulation. Either way, the output is a value for relative humidity, calculated on the chip with reference to an onboard temperature sensor.

Analog Humidity Sensor

The Humirel HS1101 is a low-cost analog-output humidity sensor that varies its internal capacitance between approximately 160pF and 200pF as relative humidity increases from 0% to 100%. The response is almost linear, with the curve steepening slightly when humidity exceeds 80%. The component is shown in Figure 18-4.

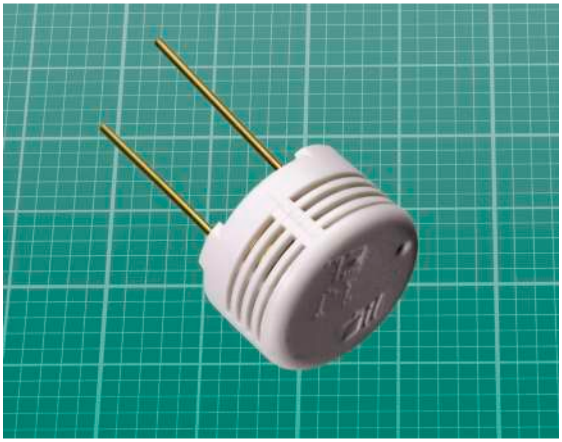

Figure 18-4 *A Humirel HS1101 humidity sensor. The background grid is in millimeters.*

The manufacturer claims a recovery time of 10 seconds after 150 hours of condensation. In other words, the performance of the sensor should be restored to its original specification.

A microcontroller can evaluate the output by measuring the charge time of the internal capacitor in the sensor. If the sensing element is wired in parallel with a 10M resistor, this will allow the capacitor to discharge before the microcontroller charges it again. A 220-ohm series resistor should be used between the microcontroller and the sensor, to limit the charge current. This circuit is illustrated in Figure 18-5.

Alternatively, the sensor manufacturer suggests using the capacitance value to control the output frequency of a 555 timer. A counter or microcontroller may be used to count the number of pulses per unit of time.

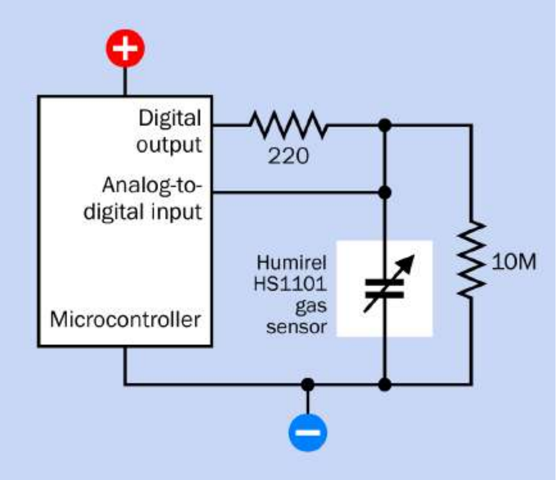

Figure 18-5 *Using a microcontroller to measure the charge time of a gas sensor with capacitive output.*

The Honeywell HIH4030 is a surface-mount humidity sensor with a more convenient analog voltage output that increases almost linearly from approximately 0.8VDC at 0% humidity to 3.8VDC at 100% humidity, assuming a 5VDC power supply. The sensor is available on a miniature breakout board from Sparkfun, shown in Figure 18-6.

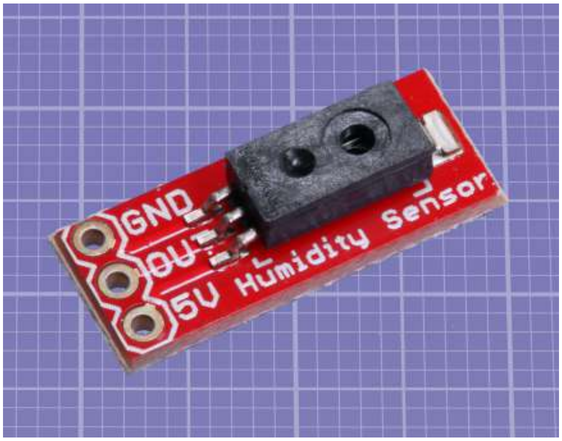

Figure 18-6 *The Honeywell HIH4030 on a breakout board from Sparkfun.*

Design Considerations

As relative humidity is temperature-dependent, a relative-humidity sensor must have the same temperature as the air it measures. Many datasheets recommend that when a sensor is mounted on a printed circuit board, slots should be milled around it to minimize heat transfer to it. The sensor should also be mounted as far away as possible from heat-generating components.

Where a capacitive humidity sensor with an analog output is located some distance from the electronics that will process its output, shielded cables or twisted pair cables should be used to minimize capacitance in the wiring. A decoupling capacitor between the supply voltage and ground close to the sensor can help to keep the supply voltage stable.

Digital Humidity Sensor

The AM2302, available from Adafruit, is a capacitive humidity sensor with a digital output accessible by a microcontroller via the I2C protocol. Onboard electronics calculate relative humidity with reference to an included temperature sensor. This component is pictured in Figure 18-7.

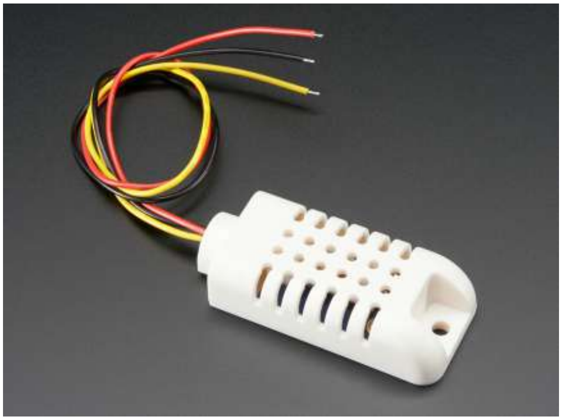

Figure 18-7 *A low-cost humidity sensor with temperature-compensated digital output. From a photograph by Adafruit.*

What Can Go Wrong

Contamination

A semiconductor gas sensor can be damaged by exposure to volatile chemical vapors. This would be unusual in a domestic environment, but is still an important consideration, as there will be no obvious indication that damage has occurred.

Recalibration

If a humidity sensor is exposed to very high humidity where condensation occurs, some datasheets advise a baking procedure, where the sensor is placed in warm, dry air for several hours and then allowed to rehumidify.

Soldering

Semiconductor gas sensors should be soldered quickly and at a controlled temperature, to minimize heat transfer.

gas flow rate sensor

19

Sensors that contain no electronic components are not included in this entry.

A **gas flow rate** sensor may often be described as a *mass flow* sensor or *mass flow rate* sensor. Although it measures volume rather than mass, the mass of gas can be calculated so long as its temperature and pressure are controlled.

Some methods of gas flow sensing can also be applied to liquids, but sensors are usually designed for one application or the other. Therefore, **liquid flow rate** sensors have their own entry. See Chapter 16.

An *anemometer* is a gas flow rate sensor that measures air speed. It is included in this entry.

Many gas flow rate sensors are large devices designed for industrial applications. This entry focuses mostly on lower-cost solid-state sensors that are often classified as *board-mount* components.

OTHER RELATED COMPONENTS

- **liquid flow rate** sensor (see Chapter 16)

What It Does

A **gas flow rate** sensor measures the volume of gas flowing past or through the device, usually inside a pipe. In most applications, users wish to know the mass of the gas that is passing per unit of time. Consequently a gas flow rate sensor is very often identified as a *mass flow rate* sensor. If it functions by heating the gas and measuring the heat dissipation, it is a *thermal mass flow rate sensor*.

A sensor that measures the flow of open air is often described as an *anemometer*. Its output will be expressed as a velocity, not as a volume or mass.

Applications

Mass flow rate sensors are used frequently for laboratory and medical applications, although the reliability and affordability of thermal mass flow rate sensors is now making them attractive for metering municipal gas supplies.

Anemometers are used mostly in meteorology, aviation, and on boats.

Schematic Symbol

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors, but they are not schematic symbols of the type generally found in electronic circuit diagrams, and are not included here.

How It Works

Because an anemometer functions so differently from a gas flow rate sensor that is enclosed in a pipe, the two types of sensors will be discussed separately.

Anemometer

The name of this device is derived from the Greek word *anemos*, meaning "wind." The *cup anemometer* was invented in 1846, using four hemispherical cups attached to horizontal arms rotating on a vertical shaft. The rate of rotation was proportional to wind speed over a substantial range, but the conversion factor between wind speed and RPM varied depending on the size of the cups and their distance from the shaft.

Anemometer design was simplified to three cups in 1926 and was modified in 1991 by adding a tag to one cup. This causes the speed of rotation to fluctuate as the tag rotates through the wind stream, and the direction of the wind can be calculated from the speed fluctuations. Not all anemometers rely on this principle, however, and a separate wind vane can be used to determine wind direction.

The basic design of a modern anemometer is illustrated in Figure 19-1.

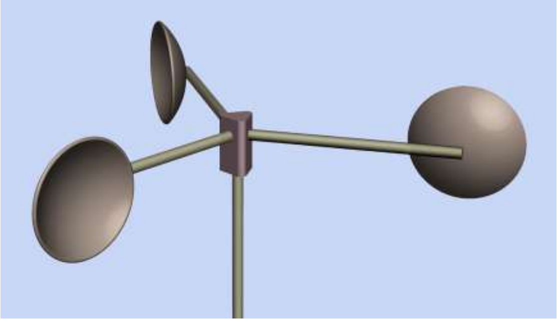

Figure 19-1 *The basic design of a metereological cup anemometer.*

Anemometers traditionally used a mechanical counter to log the number of rotations, which were checked at fixed intervals to derive the wind speed. The output of a modern anemometer may be achieved by generating AC or DC power, or from *Hall-effect sensors* (see "Hall-Effect Sensor" for a full discussion of Hall-effect sensors).

Handheld Anemometer

A digital handheld anemometer for personal use is shown in Figure 19-2. A cup anemometer made by Vaavud for use with a mobile phone (with appropriate software) is shown in Figure 19-3.

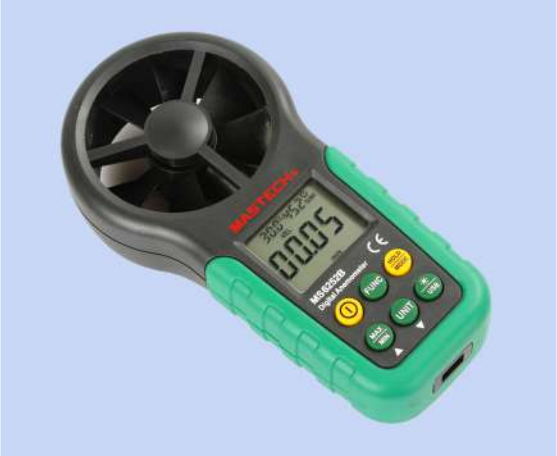

Figure 19-2 *A handheld digital anemometer.*

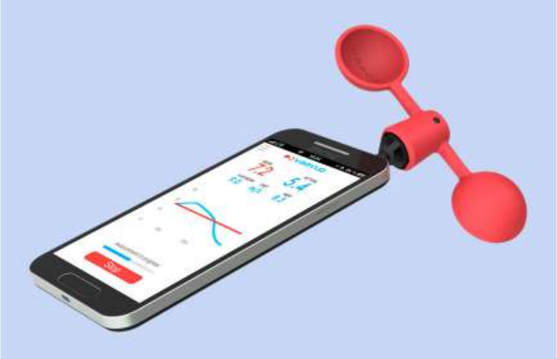

Figure 19-3 *A cup anemometer sold as an accessory for smartphones.*

Ultrasound Anemometer

The movement of air affects the speed of sound, enabling calculation of both wind direction and wind velocity by using an array of ultrasound emitters and receivers. The lack of rotating parts promises greater reliability, and in Figure 19-4 an ultrasound anemometer manufactured by Biral Metereological Sensors also includes heaters so that it will be immune to snow or ice accumulation during freezing conditions.

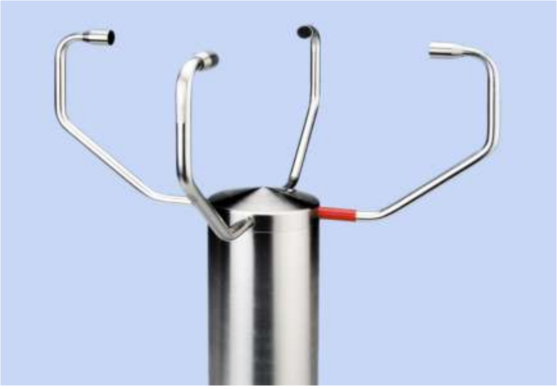

Figure 19-4 *Made by Biral Metereological Sensors, this anemometer uses ultrasound to determine wind speed and direction.*

Ultrasound anemometers have been built on a DIY basis by hobbyists, typically using off-the-shelf ultrasound emitters and an Arduino to decode the signals. Several of these projects are documented online.

Hot Wire Anemometer

A hot wire anemometer measures the cooling effect of the air. It heats a thin wire by passing current through it, and measures the heating power needed to keep the temperature constant.

It is also possible to keep either the voltage over the wire or the current constant, and to assess the wire temperature. The temperature can be measured directly or calculated from the wire's resistance, which increases as the wire gets hotter.

Mass Flow Rate Sensing

A mass flow rate sensor measures the flow speed of a gas. When this is multiplied by the density, the mass flow rate can be calculated.

Most sensors of this type heat the gas and are categorized as *thermal mass flow rate* sensors. The gas passes over a *thermopile* (consisting of several thermocouples wired in series), then a heater, and then another thermopile. These components are miniaturized and can be etched into a chip measuring 2mm square or less.

The temperature difference between the two thermopiles increases as the flow of gas becomes faster and transports more heat to the second thermopile. This is known as the *heat transfer principle*, illustrated in Figure 19-5.

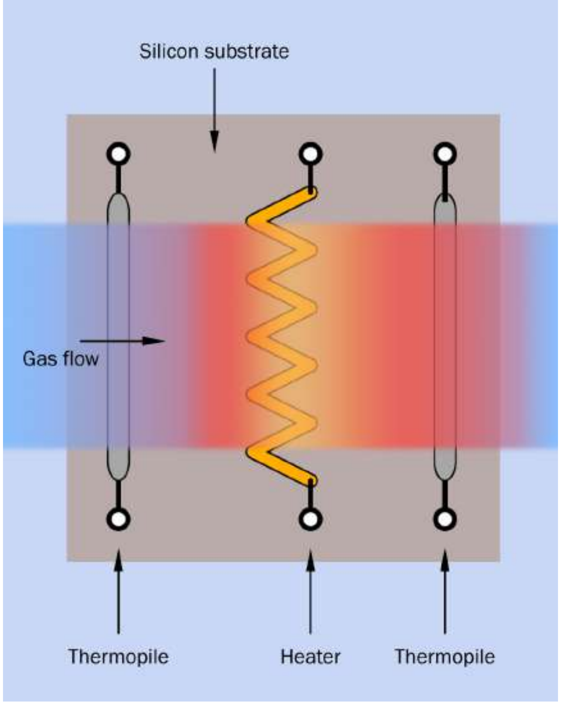

Figure 19-5 *In a thermal mass flow sensor, the flow of gas heats the thermopiles disproportionately, and the temperature difference can be used to evaluate the rate of flow.*

This principle is used also in a liquid thermal mass flow rate sensor, described in "Thermal Mass Liquid Flow Rate Sensor".

An example of a thermal mass flow rate sensor is shown in Figure 19-6.

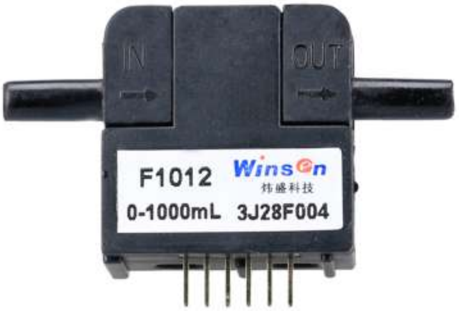

Figure 19-6 *A thermal mass flow rate sensor made by the Chinese manufacturer Zhengzhou Winsen Electronic Technology Co. Ltd.*

Applications

In medicine, mass flow rate sensors are used in anesthesia delivery, respiratory monitoring, sleep apnea machines, ventilators, and other devices. Industrial uses include air-to-fuel ratio measurement, gas leak detection, and gas metering.

While the metering of municipal gas supplies was traditionally done with all-mechanical devices, MEMS-based meters are replacing many of the 400 million mechanical gas meters estimated to exist worldwide.

Units

Mass flow rate sensors are often rated in SLM, meaning standard liters per minute. A "standard liter" has a temperature of 0 degrees Celsius and a pressure of 101.325 kPa (kilopascals). This pressure is equivalent to that of air at sea level. Because temperature and pressure are specified, the mass of a gas in a standard liter can be calculated by knowing the density of the gas. Thus SLM is a way of measuring mass flow, even though it is a measurement of volume.

The acronym SLPM is also used, but it means the same thing as SLM. SLs and SLPs are measurements of standard liters per second, while SCCM refers to standard cubic centimeters per minute.

Measuring Higher Volumes

MEMS sensors are typically equipped with inflow and outflow nozzles suitable for flexible tubing of 3mm or 5mm internal diameter. The nozzles are "barbed," meaning that they have ridges to retain the tubing.

Small tubing can only deal with low flow rates. A few sensors are threaded for standard plumbing pipes, and can accept volumes of up to 10 liters per minute. They are a minority. A low-rate sensor can still be used to measure higher volumes if it is supplied with just a percentage of the primary flow. This principle is illustrated in Figure 19-7, where an adjustable vane in the main pipe creates a pressure differential. A narrower, constricting section of pipe could have a similar effect, but it would not be adjustable.

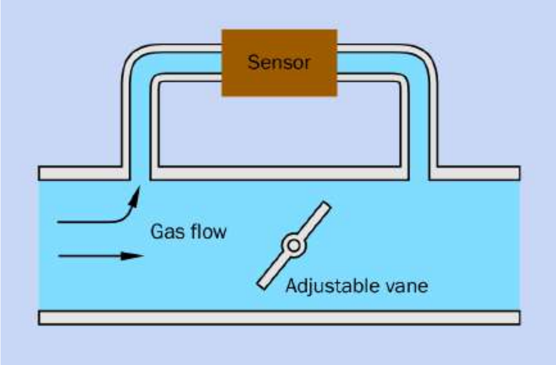

Figure 19-7 *A vane in a primary pipe can divert a percentage of the flow to a sensor.*

Output

Many mass flow sensors have an analog output consisting of a voltage that varies ratiometrically with gas flow. With a typical 5VDC power supply, output voltage may vary from around 1VDC to 4VDC.

Some sensors now incorporate analog-to-digital converters and data processing to provide SLM digital values, accessible from a microcontroller via an I2C interface.

What Can Go Wrong

The primary risk for gas flow rate sensors is damage caused by particles and contaminants in the gas stream. A 5-micron filter is recommended. A *dust segregation system* can also be used, consisting of a small compartment containing semicircular centrifugal chambers. Dust tends to follow the outer edge of the flow path, while the flow sensor is placed on the inner side of the path.

photoresistor

20

A **photoresistor** has a function similar to that of a **phototransistor**, but as its name implies, it is a purely passive component that varies its resistance in response to light.

The term *photocell* was used formerly, but has been displaced by the term **photoresistor**, which describes its function more accurately. The term *photoconductive cell* is sometimes used, or *light-dependent resistor* (or its acronym, *LDR*).

OTHER RELATED COMPONENTS

- **resistor** (see Volume 1)
- **photodiode** (see Chapter 21)
- **phototransistor** (see Chapter 22)

What It Does

Formerly known as a *photocell*, a **photoresistor** is a disc-shaped component with two leads. When light falls on the surface of the disc, resistance between the leads will diminish. Some photoresistors have a resistance in darkness as high as 10 megohms. A few have a resistance in bright light as low as 500 ohms, although several kilohms would be more common.

A photoresistor is less sensitive to light than a **phototransistor** or **photodiode**, and unlike them it is a passive component with no polarity. It presents equal resistance to current in either direction, and may be used with DC or AC.

Because cadmium sulfide is commonly used in this component and is regarded as hazardous to the environment, photoresistors are now unavailable in some regions (notably, Europe). However, at the time of writing, they are still available from many Asian sources, and from some importers in the United States.

Schematic Symbol

Six schematic symbols for a photoresistor are shown in Figure 20-1. They are functionally identical, regardless of whether the single slanting arrow across the resistor symbol is omitted.

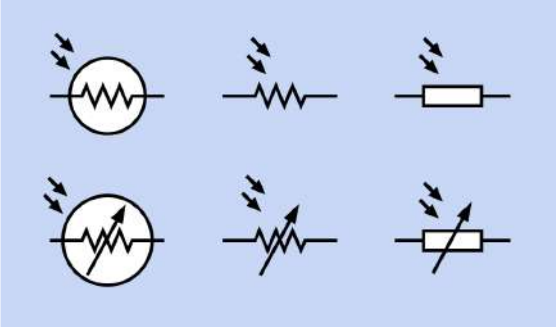

Figure 20-1 *All six symbols for a photoresistor are functionally identical.*

How It Works

Cadmium sulfide is a semiconductor. When exposed to light, more charge carriers are excited into states where they are mobile and can participate in conduction. As a result, electrical resistance decreases.

Construction

A closeup of a photoresistor appears in Figure 20-2. The brown material is a layer of cadmium sulfide deposited onto a ceramic base. The silver material is a conductive compound evaporated onto the cadmium sulfide to form two interlocking electrodes, in a pattern that maximizes the length of the boundary between each of them and the semiconductor. The electrodes connect with leads projecting from the back of the component.

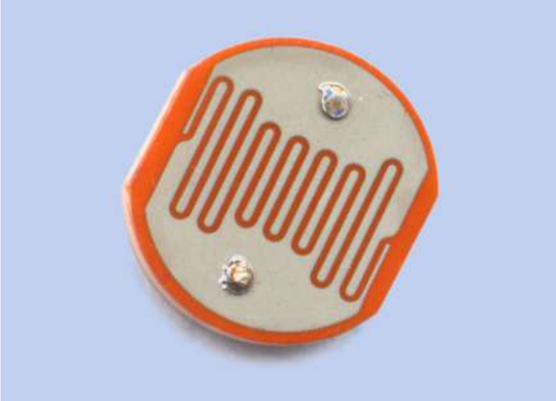

Figure 20-2 *Closeup of a photoresistor. Two interlocking electrodes are mounted on a brown semiconductor layer.*

Variants

Figure 20-3 shows a variety of photoresistors, illustrating the range of sizes available. Small photoresistors may be less than 5mm in diameter; large ones may be 25mm in diameter. The size generally suggests the ability of the component to pass current.

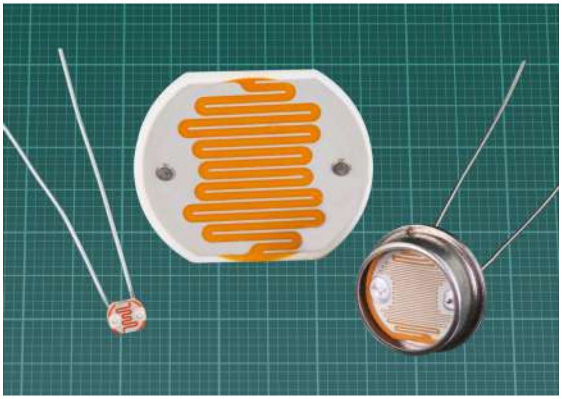

Figure 20-3 *Photoresistors are available in a wide range of sizes. The component in the center is shown at the same scale as the others. Generally speaking, larger components are able to conduct higher currents. The background grid is in millimeters.*

Photoresistors in Optical Isolators

An *optical isolator*, popularly known as an *optocoupler*, contains an LED opposite to a photoresistor, in a sealed package. This is discussed in Volume 2. A *Vactrol* is a similar component, except that the LED is placed opposite to a photoresistor. An example is shown in Figure 20-4.

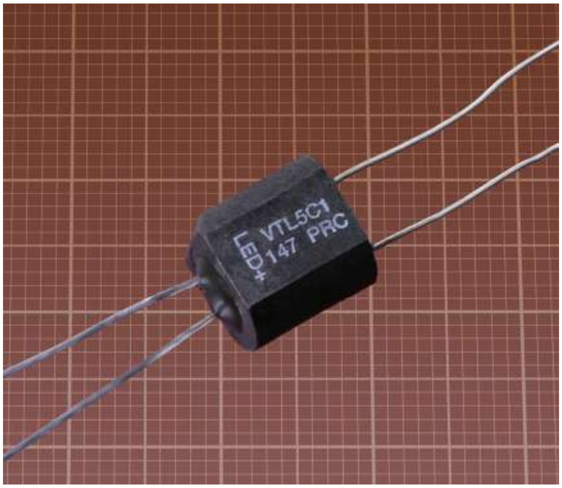

Figure 20-4 *A Vactrol, containing an LED opposite to a photoresistor. The background grid is in millimeters.*

Vactrol is a brand name owned by its initial manufacturer, Vactec. It was developed to control a vacuum tube, hence its name. In the 1950s, Vactrols were used in guitar amplifiers to control tremolo.

Because of the relatively slow response time of a photoresistor, and its sensitivity to temperature, optical isolators based on photoresistors are not used in digital devices. They still retain some utility in audio components and music equipment, where the ability of the photoresistor to pass AC is an advantage and its response time is adequate.

Values

Datasheets for a few photoresistors can still be found from some suppliers, such as Digi-Key, but are mostly unobtainable, as major semiconductor companies have stopped making these components. Vendors may quote basic values, but in the absence of part numbers or a manufacturer name, a buyer cannot verify the information.

The resistance range can be determined by testing a sample component. A typical range would be from 50K in a light intensity of 10 lux, up to 1M in darkness. Maximum power dissipation is likely to range from 100mW for a small photoresistor to 500mW for a large one.

Maximum voltage may be as high as 200V, but photoresistors will work just as well at low voltages.

Comparisons with a Phototransistor

Slower response

A photoresistor typically takes several milliseconds to respond to bright light, and can require more than 1 second to regain its dark resistance. A phototransistor is much more responsive, and a photodiode is faster still.

Narrower range of resistance

The maximum resistance of a photoresistor is almost always significantly lower than the effective maximum resistance of a phototransistor in darkness, and the minimum resistance is likely to be significantly higher than that of a phototransistor in bright light.

Greater current-carrying capacity

Often a photoresistor is rated for twice as much current as the output from a phototransistor.

Not directional

Because a photoresistor is not packaged with a lens, it is sensitive to incident light from anywhere in front of it. If an application requires that light sensitivity should be confined within a narrow angle while the component is insensitive to incident light from other directions, a phototransistor or photodiode should be used.

Temperature dependent

The resistance of a photoresistor varies more with temperature than the effective resistance of a phototransistor.

Cost

At this time, photoresistors are likely to be more expensive than phototransistors.

Lack of information

Photoresistors are often sold without means of checking their specification in a datasheet.

How to Use It

While the effective resistance of phototransistors and photodiodes varies with the applied voltage, photoresistors present the same resistance for a particular light intensity regardless of the voltage applied. This property has made them suitable for use in “stomp box” guitar-effects pedals.

Because the minimum resistance of a photoresistor in bright light tends to be relatively high, and because its response time is quite slow, it is suitable primarily as an analog component rather than as a switch.

In principle, any resistor in a circuit can be replaced by a photoresistor, and its smooth response to light variations would make it suitable, for example, as the resistance in an RC oscillator circuit, where it would determine the charge time of a capacitor. The frequency of the circuit would thus become light-dependent.

Cadmium sulfide photoresistors respond most actively to wavelengths of light ranging from 400nm to 800nm. This is especially important where an **LED indicator** is used as a light source, as LEDs often emit an extremely narrow range of wavelengths. (See the entry for LED indicators in Volume 2.)

Choosing a Series Resistor

To convert light intensity into a voltage, a photoresistor can be connected in series with a regular resistor, to form a voltage divider. There are two ways to do this, as shown in Figure 20-5. On the left side of this figure, light falling on the photoresistor will cause the output voltage to rise. On the right side, light will cause the output to drop.

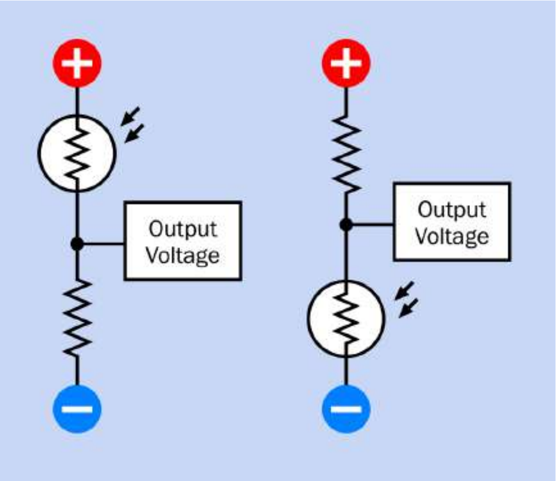

Figure 20-5 *Using a photoresistor to create a variable voltage. See text for details.*

If R_{MIN} is the minimum photoresistor value in the brightest light that will be used, and R_{MAX} is the maximum value in the dimmest light that will be used, a simple formula can be used to find R_S, the optimum value for a series resistor, which will provide the widest range of voltage values at the center of the voltage divider:

$$R_s = \sqrt{R_{min} * R_{max}}$$

What Can Go Wrong

Overload

Because a datasheet may be unavailable, a photoresistor may have to be used on a trial-and-error basis. A test-to-destruction approach may be necessary to determine the limits of the component.

Excessive Voltage

Exceeding the maximum rated voltage for a photoresistor, even for a short time, can cause irreversible damage. Depending on the component, overvoltage can range from 100V to 300V.

Confusion Among Components

Because photoresistors often have no part numbers printed on them and may be sold in miscellaneous assortments, one component may be easily confused with another that has different characteristics. Where two or more photoresistors are used in one device, their characteristics should be measured to determine whether they are functionally identical.

photodiode

21

OTHER RELATED COMPONENTS

- **diode** (see Volume 1)
- **photoresistor** (see Chapter 20)
- **phototransistor** (see Chapter 22)

What It Does

Light falling on a **photodiode** causes it to generate a very small current. This is often called the *photovoltaic effect*. The component functions like a solar panel; in fact, a solar panel can be thought of as being an array of very large photodiodes.

Often, a DC power source is used to apply *reverse bias* to a photodiode. This enables the component to deliver more current. It is now operating in *photoconductive* mode.

Schematic Symbols

Schematic symbols for a photodiode are shown in Figure 21-1. Wavy arrows are often (but not always) used to indicate infrared light.

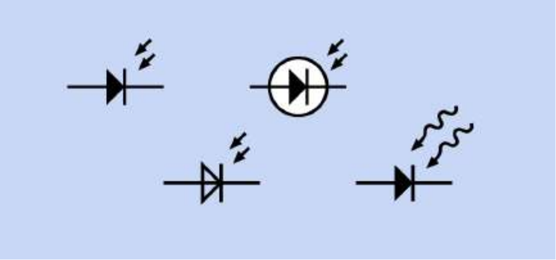

Figure 21-1 *Symbols representing a photodiode.*

Applications

The rapid response of photodiodes makes them suitable for use in optical disc drives, telecommunications, infrared data transfer, digital cameras, and optical switches. Many sensors in this Encyclopedia use a photodiode. **Proximity** sensors, optical encoders, and light meters are examples.

How It Works

When light falls on a semiconductor, it can excite an electron to a higher energy state. The electron then becomes mobile, and leaves behind an *electron hole* (See the entry discussing **diodes** in Volume 1.)

In photovoltaic mode, incident light creates pairs of electrons and electron holes in the semiconductor material. The electrons move to the cathode of the diode while the holes move to the anode, creating voltage between the two. Note that this happens to some extent even in the absence of visible light, as the photodiode may respond to infrared radiation. The tiny amount of current created without visible light is called *dark current*.

In photoconductive mode, light falling on the photodiode creates pairs of electrons and electron holes in the semiconductor material. These will move in opposite directions due to the bias voltage, which means a small current flows through the diode.

Photoconductive mode enables a faster response than photovoltaic mode, because reverse-bias voltage makes the depletion layer wider, which in turn decreases the capacitance. (The same effect is used in *capacitance diodes*.)

Simplified circuits showing the component in photovoltaic mode and photoconductive mode appear in Figure 21-2.

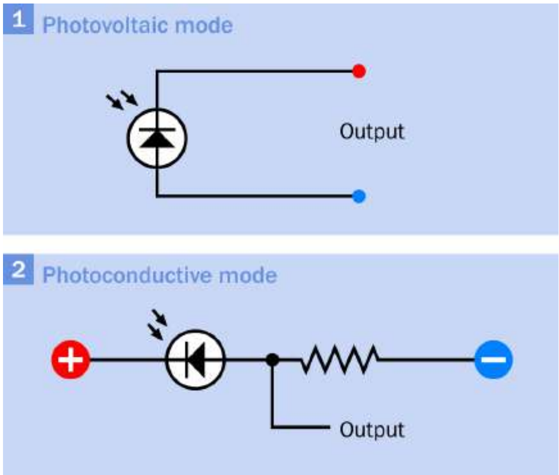

Figure 21-2 *Two modes in which a photodiode may be used. In photovoltaic mode, voltage between the two pins must be measured.*

Variants

PIN Photodiodes

Like *PIN diodes*, PIN photodiodes incorporate an undoped (intrinsic) semiconductor layer between the p- and n-doped layers. They are more sensitive and have faster responses than regular PN photodiodes. Many of the photodiodes available are of the PIN type.

Avalanche Diodes

When light enters the undoped region of the avalanche photodiode, it triggers the creation of electron-hole pairs. When electrons migrate toward the *avalanche region* of the diode, the cumulative field strength increases their velocity to the point where collisions with the crystal lattice create further electron-hole pairs.

This behavior causes the avalanche diode to be more sensitive than a PIN photodiode. However, the sensitivity also makes it vulnerable to electrical noise, and it is significantly affected by heat. A guard ring is added around the p-n junction, and a heat sink is often used.

Packages

Many surface-mount and through-hole versions are available. A selection is shown in Figure 21-3 and Figure 21-4. While a through-hole photodiode may look indistinguishable from a through-hole 3mm or 5mm LED, versions are available without lenses, and some are sensitive to incident light coming from the side (they are referred to as *side-looking* or *side-view* variants).

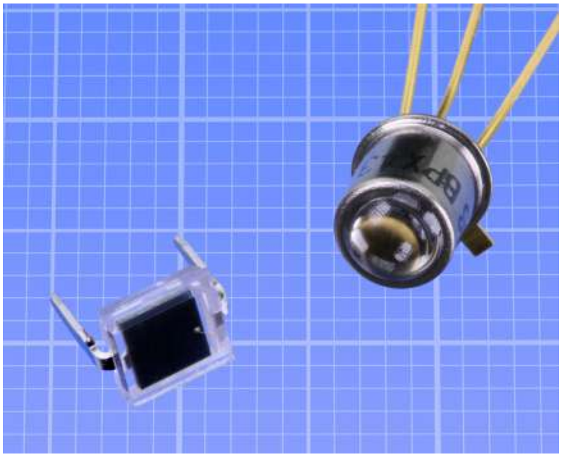

Figure 21-3 *Two sample photodiodes. Left: top-view unfiltered Vishay BPW34. Right: Osram BPX43 in metal can suitable for temperatures up to 125 degrees Celsius.*

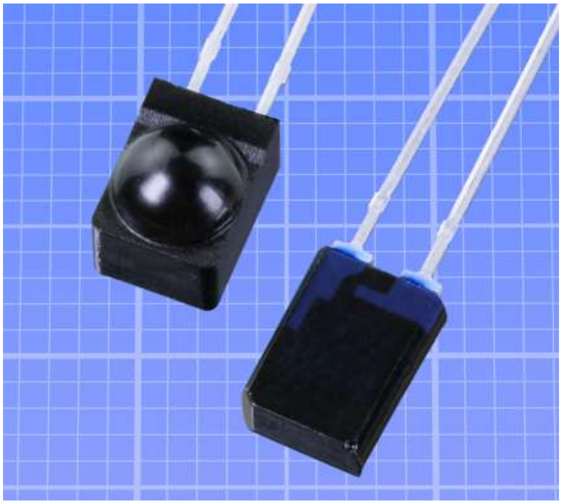

Figure 21-4 *Two side-looking photodiodes. Left: Vishay BPV22NF with lens. Right: Vishay BPW83 without a lens. Both have daylight-blocking filters.*

Wavelength Range

In order for a photon to be detected by the photodiode, it must carry enough energy to be able to create an electron-hole pair. This energy is a property specific to the semiconduct material, and is known as its *band gap*. Additionally, the epoxy packaging of the photodiode may be designed to block some wavelengths of light. Often, an application will require that the component should only respond to infrared light, not visible light.

Photodiode Arrays

A *photodiode array* has several photodiodes mounted in a row or in a grid, for imaging or measurement applications. A row of photodiodes may be used in a flatbed scanner, where it is moved relative to the reflective object being scanned.

In some arrays, photodiodes are available with color filters preinstalled, to facilitate full-color scanning using the transmitted primary colors.

Output Options

Because the output from a photodiode must be processed to be usable, options exist to convert it to a more convenient form, such as a wider range of voltages, a square wave signal where the frequency is proportional with light input, or a binary value accessible by a microcontroller via a serial bus such as I2C. For a discussion of sensor outputs see Appendix A.

Specific Variants

Light to Frequency

The Taos TSL235R is a 3-pin, through-hole chip. It combines a photodiode with logic that creates a square-wave pulse train in which frequency is proportional to the light intensity.

Logarithmic Light Meter

The Sharp GA1A1S202WP light sensor has an output voltage that changes logarithmically with the light level. This gives the sensor a large dynamic range, from 3 lux to 55,000 lux, without requiring a high-resolution analog-to-digital converter. (Human perception of light and sound levels is approximately logarithmic.) This is a surface-mount chip, but is available on a breakout board from Adafruit.

Ultraviolet to Analog

The ML8511 from Lapis Semiconductor combines an ultraviolet-sensitive photodiode with an op-amp that provides an output voltage from approximately 1V to 3V, varying with ultraviolet intensity. A breakout board containing this surface-mount chip is available from Sparkfun and many other sources.

Ultraviolet to Digital

The SI1145 from SiLabs combines ultraviolet sensing with data processing to create a UV index, readable from a microcontroller with I2C. Adafruit offers it on a breakout board.

Color to Digital

The Taos TCS3414FN module contains photodiodes sensitive to red, green, blue, and clear (no filtering). Four 16-bit analog-to-digital converters, one for each channel, provide a digital output accessible over an I2C bus. This module can be used to determine the color of ambient light with some accuracy.

Color to Analog

The Taos TCS3200 also uses red, green, blue, and clear photodiodes but encodes the output from each as a square wave in which the frequency corresponds with the light intensity. The surface-mount chip is available on a breakout board from Robot Shop, shown in Figure 21-5.

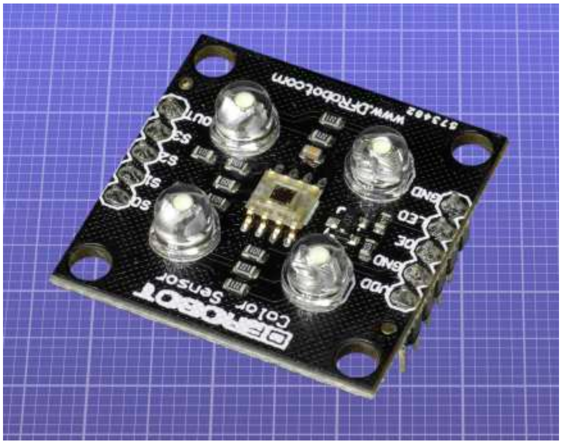

Figure 21-5 *This breakout board from Robot Shop uses the Taos TCS3200 chip to analyze the color of incident light.*

Values

Abbreviations found in datasheets are included in the list below, with values in parentheses for an Osram SFH229FA infrared photodiode, which resembles a 3mm through-hole LED. It has a peak sensitivity of 880nm and appears black to the human eye, being opaque to wavelengths of light shorter than 700nm, the red end of the visible spectrum.

In Figure 21-6 the SFH229FA is shown beside the SFH229, which has the same peak sensitivity of 880nm but is encapsulated in an untinted module, allowing a sensitivity that tapers gradually to below 400nm, in the green part of the visible spectrum. With the exception of their spectral range, the two photodiodes have identical specifications.

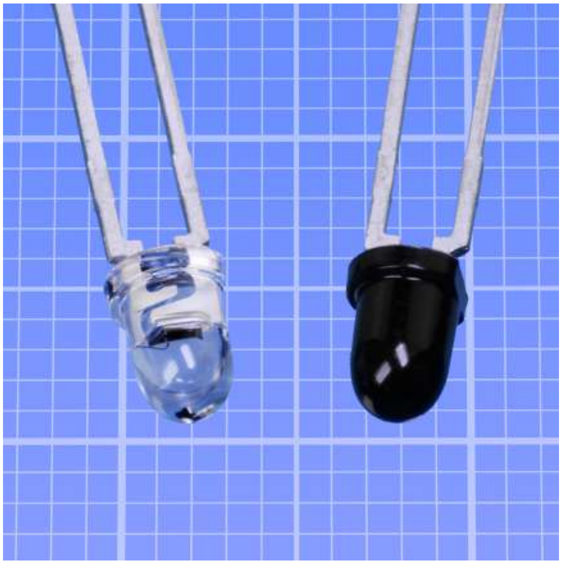

Figure 21-6 *Osram infrared photodiodes SFH229 (left) and SFH229FA (right). The background grid is in millimeters.*

- Typical forward voltage: V_F (1.3V)
- Typical photocurrent: I_P (20µA)
- Maximum power dissipation: P_{TOT} (150mW)
- The *half angle* is measured from the axis of the photodiode to the angle at which the sensitivity has dropped by 50% (plus or minus 17 degrees)
- Dark current: I_R (50pA)
- Wavelength of maximum sensitivity: $\lambda_{S\ MAX}$ (880nm)
- Response speed is the rise and fall time of photocurrent: t_r and t_f (10ns)

Infrared photodiodes exist with a variety of peak wavelengths. They are designed to function in conjunction with an LED that has a matching wavelength.

The angle of sensitivity depends on the geometry of the package.

Rise and fall speeds are important for high-speed measurement, signaling, or data transfer. The rise and fall time of a typical photodiode can be 1,000 times faster than that of a phototransistor. See "Values" in Chapter 22 for a comparison. Note also that the dark current of a photodiode is much lower than that for a phototransistor.

How to Use It

In photoconductive mode, the photodiode can be connected in series with a suitable resistor so that a voltage divider is formed, as shown in Figure 21-2. The voltage at the output will then vary almost linearly with the intensity of the light.

In the photoconductive mode of operation, the output signal is generally measured in millivolts and microamperes. This signal needs to be amplified, usually with an **op-amp** (described in Volume 2).

Figure 21-7 shows simplified schematics for a standard voltage amplifier, in section 1, and a transimpedance amplifier, in section 2.

A transimpedance amplifier measures the current through the photodiode and converts it into a voltage, without the need for a voltage divider. Advantages include less noise, and no need to determine the value of a voltage-divider resistor.

The output voltage of this amplifier is calculated by using this simple formula:

$$V = R * I_P$$

R is the value of the feedback resistor, which determines the gain of the amplifier.

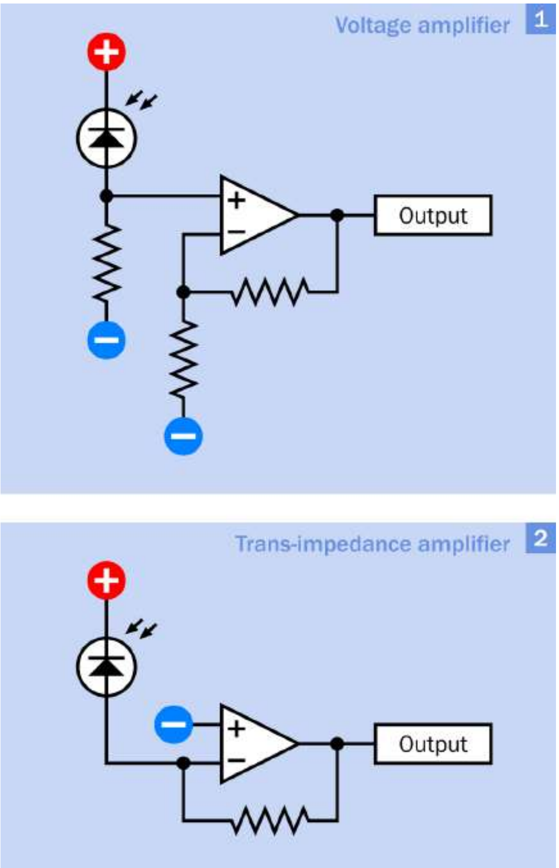

Figure 21-7 *Simplified op-amp circuits for use with photodiodes.*

What Can Go Wrong

Photodiodes can be hard to distinguish from phototransistors and LEDs, in particular infrared ones. They are typically not marked with type numbers. Using a meter may not be helpful, as regular LEDs behave similarly to photodiodes.

Decision procedure:

1. Does the component conduct in either direction while shielded from light? If so, it is a diode and not a phototransistor or photodiode.
2. Pass a weak current (e.g., 4mA) in the forward direction. If it emits visible or infrared light, it's an LED. (Infrared light may be visible when using a digital camera, or can be detected with a known phototransistor or photodiode.) If the package is clear but cloudy so that the die cannot be seen, it is probably a white LED in which the cloudiness is the fluorescent pigment converting blue light into white.

phototransistor

22

OTHER RELATED COMPONENTS

- **photodiode** (see Chapter 21)
- **photoresistor** (see Chapter 20)
- **passive infrared** sensor (see Chapter 4)
- **transistor** (see Volume 1)

What It Does

A **phototransistor** is a transistor controlled by exposure to light. (**Transistors** are described in Volume 1.) It can be either a bipolar transistor or field-effect transistor (FET), and its body is often superficially similar in appearance to a 3mm or 5mm **photodiode** or **LED indicator** encased in resin or plastic. (LED indicators are described in Volume 2.) However, some phototransistors are encased in a metal shell with a window in it.

The window or the plastic body is either transparent to visible light, or may appear black if the component is intended for use only with infrared while blocking visible wavelengths. A selection of phototransistors is shown in Figure 22-1 (left: Optek/TT Electronics OP506A with a broad spectral response centered around 850nm; center: Vishay TEKT5400S with a side-view lens; right: Vishay BPW17N).

Typically a phototransistor has two leads that connect internally with its collector and emitter (or source and drain, in the case of an FET). The base of the transistor (or the gate of an FET) responds to light and controls the flow of current between the leads.

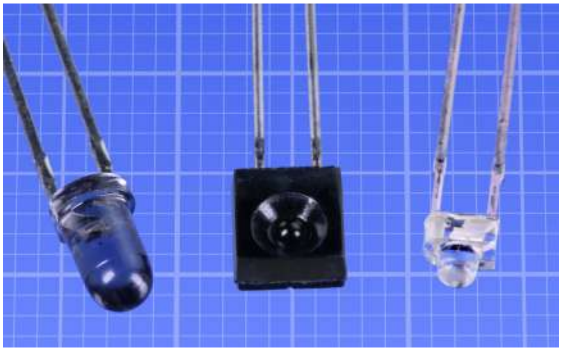

Figure 22-1 *A variety of phototransistors. The background grid is in millimeters. See text for details.*

In the absence of light, a bipolar phototransistor permits leakage between collector and emitter of 100nA or less. When exposed to light, it conducts up to 50mA. This alone differentiates it from a photodiode, which cannot pass much current.

Schematic Symbols

Symbols for a phototransistor are shown in Figure 22-2. They are functionally identical, with the exception of type C, where an additional connection to the base is included. Often (but not always), wavy arrows indicate infrared light.

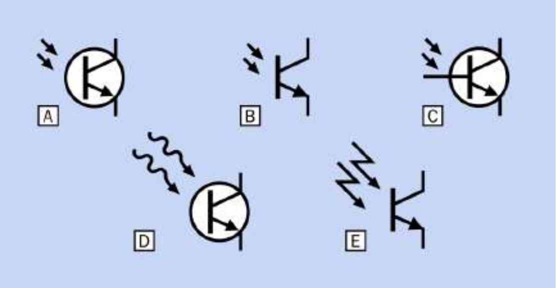

Figure 22-2 *Schematic symbols for a phototransistor.*

Applications

A phototransistor may be used for light measurement or as a light-sensitive switch.

Often the output from a phototransistor is used in conjunction with a *microcontroller* containing an analog-to-digital converter.

Where a clean on-or-off signal is required, a phototransistor can drive the input of a logic chip that contains a *Schmitt trigger*, or it can be processed by a **comparator**.

An *optocoupler* or **solid-state relay** (described in Volume 1) usually contains a phototransistor that is activated by an internal LED. Its purpose is to switch current while electrically isolating one section of a circuit from another.

How It Works

Like a photodiode, the phototransistor responds to light when the light generates electron-hole pairs in the semiconductor material. For a bipolar NPN phototransistor (the most common variant), the important region for pair generation is the reverse-biased collector-base interface. Photocurrent generated here acts as current injected into the base of an ordinary transistor, and permits a larger current to pass from the collector to the emitter.

The behavior of a phototransistor can be visualized as being similar to a photodiode controlling an ordinary bipolar transistor, as shown in Figure 22-3.

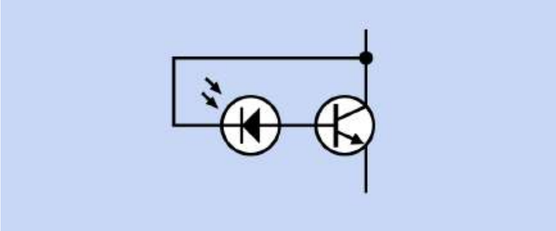

Figure 22-3 *A phototransistor is functionally similar to a photodiode controlling an ordinary transistor.*

Variants

While surface-mount variants are very widely available, through-hole packaging also remains common. When encapsulated like an LED indicator, a phototransistor gathers light from a relatively narrow angle. Variants with a flat surface are sensitive to light from almost any direction in front of the component.

Optional Base Connection

The base of a phototransistor is usually not accessible. However, some variants provide a base connection (or gate connection, in an FET) in addition to the collector and emitter (or the source and drain, in an FET). This third connection allows the application of a bias current, which can prevent low light levels from triggering the transistor.

Photodarlington

A photodarlington is a pair of bipolar transistors, the first being sensitive to light while the second acts as an amplifier for the first. This configuration is very similar to that of a *Darlington transistor* (described in Volume 1). The two-stage design makes it more sensitive to light than a regular phototransistor, but makes the response slower and less linear.

PhotoFET

A field-effect phototransistor is sometimes identified as a *photoFET*. They are relatively uncommon as separate components, but are used in optocouplers, because they have some interesting properties:

- Provided the applied voltage is low enough (less than 0.1V) the photoFET works as a controllable resistance—in contrast to bipolar transistors, where the current is controlled, and is relatively independent of the applied voltage.
- The FET transistor is symmetrical, functioning similarly for signals of either polarity. This makes an FET optocoupler suitable for AC signals.

Values

Abbreviations found in datasheets are included in the list below, with values in parentheses for an Osram SFH300FA infrared photodiode, which resembles a 5mm through-hole LED. It has a peak sensitivity of 880nm and appears black to the human eye, being opaque to wavelengths of light shorter than 700nm, the red end of the visible spectrum.

In Figure 22-4 the SFH300FA is shown beside the SFH300, which has the same peak sensitivity of 880nm but is encapsulated in an untinted module, allowing a sensitivity that tapers gradually to 450nm, in the green part of the visible spectrum. With the exception of their spectral range, the two phototransistors have identical specifications.

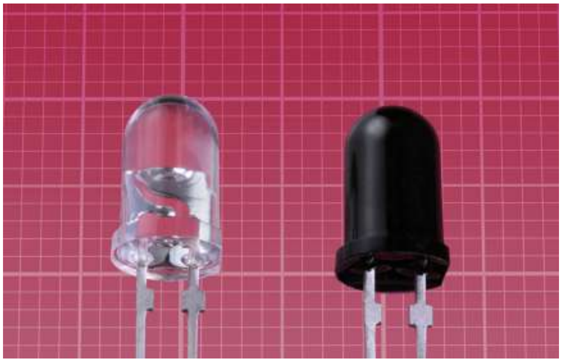

Figure 22-4 *Osram infrared phototransistors SFH300 (left) and SFH300FA (right). The background grid is in millimeters.*

- Maximum collector-emitter voltage: V_{CE} (35V)
- Maximum collector current: I_C (50mA)
- Maximum power dissipation: P_{TOT} (200mW)
- The *half angle* is measured from the axis of the photodiode to the angle at which sensitivity has dropped by 50% (plus or minus 25 degrees)

The angle of sensitivity depends on the geometry of the package. For phototransistors that resemble an LED indicator, with a rounded end that acts as a lens, typical values are plus-or-minus 20 degrees.

- Dark current (when the phototransistor receives no incident light): I_{CE0} (1nA)
- Wavelength of maximum sensitivity: $\lambda_{S\ MAX}$ (880nm)

Infrared phototransistors exist with a variety of peak wavelengths. They are designed to function in conjunction with an LED that has a matching wavelength.

- Response speed is the rise and fall time of photocurrent: t_r and t_f (10 µs)

Behavior Compared to Other Light Sensors

An extended list of comparisons between a photoresistor and a phototransistor will be found in the entry on **photoresistors**. See "Comparisons with a Phototransistor".

A photodiode has a close-to-linear electrical response over a much wider range of intensities of light than a phototransistor. Consequently, photodiodes tend to be the component of choice in applications where measurement of light must be wide-ranging and precise.

Photodiodes can pass less current than a phototransistor, but also tend to draw less current, making them appropriate in battery-powered devices that must draw as little current as possible.

Rise and fall speeds are important for high-speed measurement, signaling, or data transfer. The rise and fall time of a typical phototransistor can be 1,000 times slower than that of a photodiode. See "Values" in Chapter 21 for a comparison. Note also that the dark current of a phototransistor is much higher than that for a photodiode.

The ability of a phototransistor to sink 20mA to 50mA at its output is useful where it will be connected to a component that has relatively low impedance. For instance, a phototransistor can drive a piezoelectric audio transducer directly, or an LED indicator.

Unlike a photodiode, a phototransistor is a solid-state switch. Its saturation voltage (listed in datasheets as $V_{SE(SAT)}$ as described above) is the voltage drop between collector and emitter, and seldom exceeds 0.5V.

Binning

Small variations that occur during the fabrication process can cause inconsistency in the performance of phototransistors that share the same part number. To provide a more consistent response, manufacturers use *binning*, meaning that units sharing the same bin number are likely to share a tighter tolerance. (The same concept is used to minimize variations in **LED area lighting**, described in Volume 1.)

Datasheets will provide information on the availability and meaning of bin numbers, if available.

Bins with higher photocurrents typically have longer response times.

How to Use It

Most phototransistors are bipolar devices with an *open collector output*. That is, the collector of the transistor is accessible from one of the two leads, making it "open" to being used. See Figure A-4 for general instructions on using an open-collector output. A summary relating to phototransistors is included here.

The schematic in Figure 22-5 shows the basic concept. The resistor is referred to as a *pullup resistor*. When the phototransistor is receiving very little light, its effective resistance is high. Consequently almost all the current flowing through the pullup resistor will go to any device attached to the output, and the output voltage will seem to be "high."

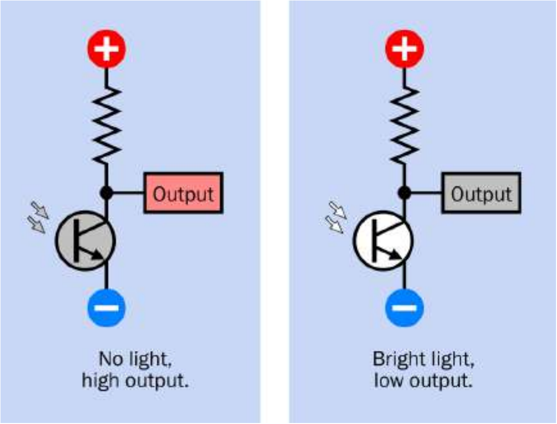

Figure 22-5 *How an open-collector output works.*

If the phototransistor is exposed to a significant light source, the effective resistance between the collector and emitter drops dramatically, and the phototransistor will sink current to ground. Consequently, the output will seem to be "low."

The pullup resistor is necessary between the power source and the collector pin to protect the phototransistor from sinking excessive current when it is exposed to light. The ideal value of the resistor will depend partly on the impedance of any device attached to the output.

In this scenario, exposure to light causes low output whereas darkness causes high output. What if we wish to have it the other way around?

The protective resistor can be moved to the emitter pin, where it becomes a pulldown resistor. It will still protect the phototransistor from passing excessive current, so long as the output is connected to a high impedance. The output is taken from the emitter pin, and will transition from low to high when the component is exposed to light. This is illustrated in Figure 22-6.

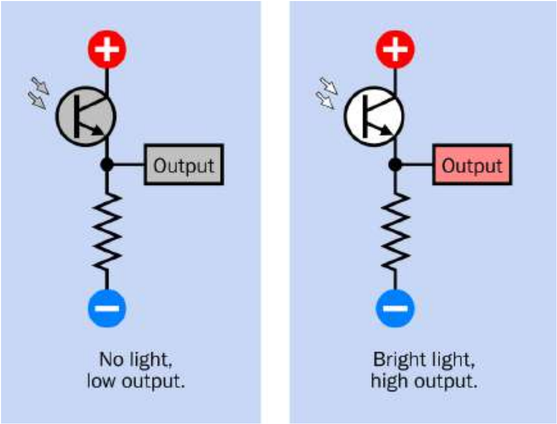

Figure 22-6 *Moving the resistor and taking an output from the emitter pin inverts the behavior of a phototransistor. Any device attached to the output must have a relatively high impedance to protect the phototransistor from excessive current.*

Output Calculation

Using an open-collector output, the photocurrent is almost independent of the applied voltage V_{CE}, provided that the voltage is higher than the saturation voltage $V_{CE(SAT)}$, which is typically between 0.4V and 0.5V.

If the pullup resistor has value R, the voltage across it is

$$U = R * I_P$$

where I_P is the photocurrent passed by the phototransistor.

When choosing R, one should consider the range of currents expected in the light conditions that will be measured, and the voltage range suitable for the next stage of the circuit. 10K is a reasonable starting point (e.g., when measuring light intensity with a microcontroller's analog input). The resistor value can be reduced from there, if necessary.

The value of the pullup resistor must also be chosen to restrict current within the limits specified by the phototransistor datasheet. A value for the resistor guaranteed to be safe is

```
R = V / I~MAX~
```

where V is the supply voltage and I_{MAX} is the maximum allowed current. With this value, the resistor limits the current to the highest allowed value, even if the phototransistor is brightly illuminated and assumed to conduct perfectly.

When V = 5V and I_{MAX} = 15mA, R should be at least 330 ohms.

What Can Go Wrong

Visual Classification Errors

Phototransistors can be visually similar to LEDs and photodiodes. They are easily confused, as neither type of component is usually marked with any identification number. The entry for **photodiodes** describes a system for distinguishing the three types of components. See "What Can Go Wrong" in Chapter 21.

Output Out of Range

The output voltage from a phototransistor will depend on the intensity of incident light, the value of any pullup resistor that is used, and the supply voltage. While a circuit is being developed, the light range may seem predictable, but in actual use the output range may not fall within expectations.

NTC thermistor

23

PTC thermistors, in which the resistance increases as the temperature increases, have a separate entry. See Chapter 24.

A *resistance temperature detector* or **RTD** has a resistance that increases as its temperature increases, but it is not usually classified as a thermistor, because its sensing element is fabricated differently. Its entry will be found at Chapter 26.

Semiconductor temperature sensors and **thermocouples** each have their own entries.

Infrared temperature sensors and **passive infrared** motion sensors have their own entries. They are *noncontact* temperature sensors that respond to infrared radiation.

OTHER RELATED COMPONENTS

- **PTC thermistor** (see Chapter 24)
- **infrared temperature** sensor (see Chapter 28)
- **passive infrared** motion sensor (see Chapter 4)
- **semiconductor** temperature sensor (see Chapter 27)
- **thermocouple** (see Chapter 25)
- **RTD** (resistance temperature detector) (see Chapter 26)

What It Does

An NTC thermistor is the most common type of discrete-component temperature sensor, and is usually the most affordable. Its resistance diminishes as its temperature increases. This behavior is referred to as a *negative temperature coefficient*, which is the source of the acronym NTC.

This is a simple, passive component that is not polarized. It requires no separate power supply, but an external device must pass a small AC or DC current through it to determine its resistance. This is known as an *excitation current*.

Schematic Symbols

Schematic symbols for a thermistor are shown in Figure 23-1. Those in the top row may still be found in the United States, but are being replaced by the European variants in the second row. The addition of -t° to the symbol indicates an NTC type of thermistor, while +t° indicates that it is the PTC type, with a positive coefficient (see Chapter 24). If no indication is shown, the thermistor is likely to be the NTC type.

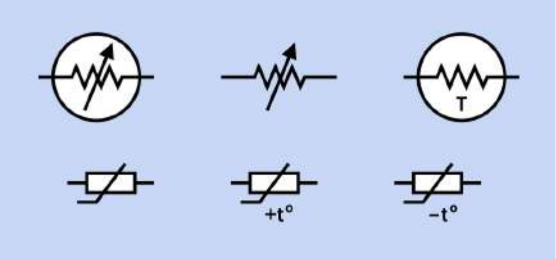

Figure 23-1 *Schematic symbols representing a thermistor. Letter t preceded by a plus or minus sign indicates whether the thermistor is the PTC or NTC type, respectively.*

Applications

Thermistors monitor temperature in air-conditioning systems, clothes washers, refrigerators, pool and spa controls, dishwashers, toasters, and other domestic devices. They are used in laser printers, 3D printers, industrial process controls, and medical equipment.

As many as 20 thermistors may be found in a modern automobile, measuring temperature in locations ranging from the transmission to the ambient air in the passenger compartment.

Comparison of Temperature Sensors

In this Encyclopedia, contact temperature sensors, which measure temperature by making contact with the source, are divided into five main categories, each of which has a separate entry. For convenience, these categories are listed in a comparative summary at the end of this entry. See "Addendum: Comparison of Temperature Sensors".

How an NTC Thermistor Works

Although the term *thermistor* suggests that it is a thermally sensitive resistor, in fact an NTC thermistor is a semiconductor.

Some metal oxides, such as ferric oxide or nickel oxide, become n-type semiconductors when they are treated with dopants. The exact mix is a proprietary secret of each manufacturer. Raising the temperature of this kind of material increases the number of charge carriers in it, promoting electron mobility and thus lowering its effective resistance.

To create a thermistor, the metal oxide mix is heated until it melts and turns into a ceramic. Typically a thin sheet is cut into small pieces for individual sensors. After two leads are connected, the assembly is dipped into epoxy or encapsulated in glass. The most common packages consist of a glass bead, surface-mount chip, or ceramic disc.

Figure 23-2 shows three NTC thermistors. At left is a Murata NXFT15XH103FA2B100 approximately 1mm in diameter, with a reference resistance of 10K and a tolerance of plus-or-minus 1%. At center is a Vishay NTCALUG03A103GC rated 10K at 2%, fitted with a mini-lug. At right is a TDK B57164K153K rated 15K and 3%.

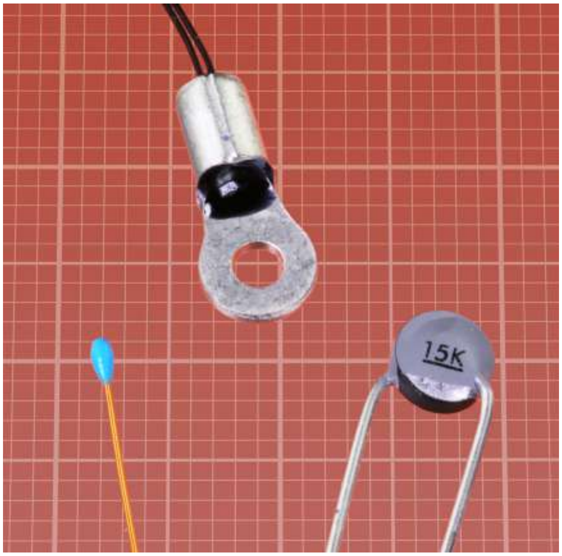

Figure 23-2 *Sample NTC thermistors. See text for details. The background grid is in millimeters.*

Output Conversion for Temperature Sensing

Ideally, the electrical behavior of a temperature sensor should be a linear function of temperature. Thermistors fail in this respect, as their resistance is an approximately inverse exponential function. This is illustrated in Figure 23-3, where the measured resistance of a thermistor rated for 5K at 25 degrees is plotted against temperatures from 0 degrees to 120 degrees Celsius.

- In many datasheets, graphs of this kind may appear flatter because they are customarily plotted against a vertical logarithmic scale.

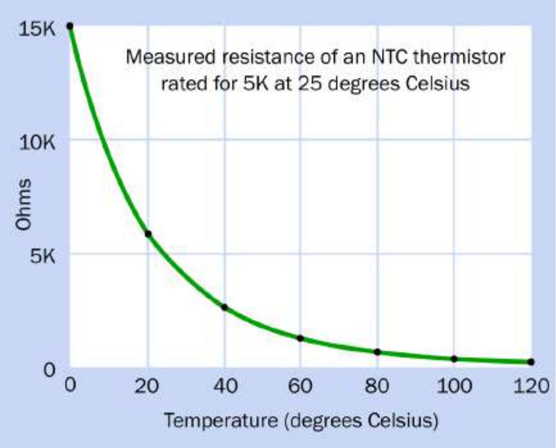

Figure 23-3 *Resistance of a thermistor from 0 to 120 degrees Celsius.*

To monitor the resistance of a thermistor, it can be placed in a simple voltage divider as shown in Figure 23-4, where the fluctuating resistance of the component creates a fluctuating voltage at point A.

- The voltage can be used as an input to a microcontroller that contains an analog-to-digital converter. Alternatively it can be connected directly to a solid-state relay, or amplified with an op-amp, or can be passed through a comparator to create an adjustable switching threshold.

Although this circuit is a voltage divider, it is also known as a *half bridge*, as it is half of a Wheatstone bridge.

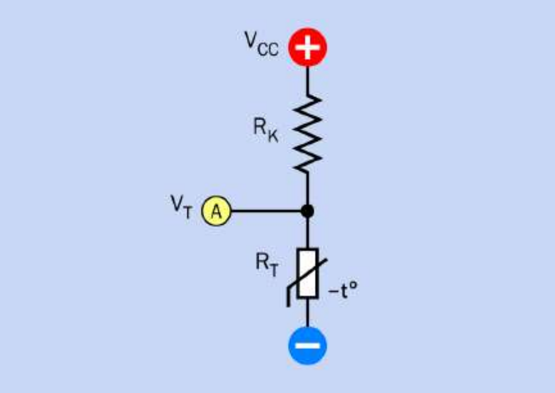

Figure 23-4 *A half-bridge circuit for determining the resistance of a thermistor.*

If V_{CC} is the supply voltage, V_T is the measured voltage at point A, R_T is the resistance of the thermistor, and R_K is the constant value of the series resistor, the basic formula for a voltage divider looks like this:

$$V_T = V_{CC} * (R_T / (R_T + R_K))$$

By transposing terms, a formula can be derived to obtain a value for R_T from the measured voltage and the value of R_K:

$$R_T = (R_K * V_T) / (V_{CC} - V_T)$$

Choosing a Series Resistor

The value for R_K in the formula should be chosen to provide a reasonably wide response over the range of temperatures for which the thermistor is likely to be used. To calculate R_K, another formula must be applied. If R_{MIN} is the resistance of the thermistor at the lowest likely temperature, and R_{MAX} is its resistance at the highest likely temperature:

$$R_k = \sqrt{R_{min} * R_{max}}$$

(This is the same formula as suggested in Figure 20-5 to find the value of a series resistor for use with a photoresistor.)

Wheatstone Bridge Circuit

The half-bridge circuit has the disadvantage that it does not compensate for nonlinearity of a thermistor. Voltage values will change rapidly at the low end of the temperature range, but will change more slowly at the high end, requiring an analog-to-digital converter with a high degree of accuracy to distinguish one voltage value from the next.

A full Wheatstone bridge circuit has a nonlinear output that compensates, somewhat, for the inverse nonlinearity of the thermistor. Referring to the circuit shown in Figure 23-5, the three resistors R_K are chosen using the formula above.

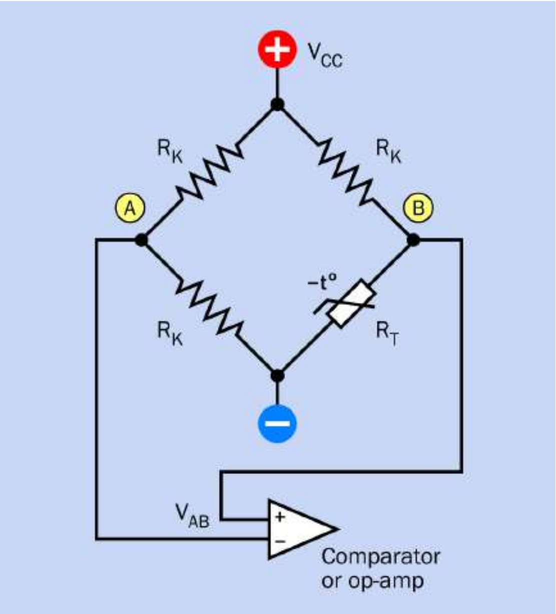

Figure 23-5 *A thermistor may be placed in a full Wheatstone bridge circuit. Outputs A and B are often connected with the two inputs of an op-amp or comparator.*

A standard formula provides the relationship between R_T, the resistance of the thermocouple; V_{CC}, the supply voltage; R_K, the fixed resistances; and V_{AB}, the output voltage measured between points A and B:

$$V_{AB} = (V_{CC} / 2) * (R_T - R_K) / (R_T + R_K)$$

From this formula, a version can be derived to calculate R_T by measuring the output voltage, V_A:

$$R_T = R_K * (V_{CC} + (2*V_{AB})) / (V_{CC} - (2*V_{AB}))$$

- The polarity of V_{AB} is reversible, depending on whether R_T is greater or less than R_K. To accommodate this, A and B can be connected to the two inputs of a comparator or op-amp.

Deriving the Temperature Value

After the resistance of the thermistor has been calculated, it can be converted to a temperature value. The datasheet for a thermistor will usually provide a table showing temperature values tabulated against resistance values, so that a lookup table can be created in a microcontroller program.

Alternatively, a datasheet usually includes constants that can be inserted in a resistance-to-temperature conversion equation, but this is nontrivial and requires natural logarithms, which may not be available in a language implemented on a particular microcontroller.

Inrush Current Limiter

NTC thermistors with appropriate characteristics can be used to limit the inrush of current that tends to occur when a circuit is switched on and large capacitors in the power supply charge very quickly.

An *inrush current limiter* is also known as a *surge limiter*, or may be referred to by its acronym, *ICL*. It is an NTC thermistor whose initial resistance diminishes rapidly as its temperature increases.

While NTC thermistors are the type most often used for inrush limiting, PTC thermistors can serve this purpose if wired differently. See "PTC Inrush Current Limiting". The remainder of this discussion refers only to NTC current limiters.

A suitable NTC thermistor can be placed as shown in the simplified schematic in Figure 23-6, where a rectified AC source is connected with a DC-to-DC converter, and a large smoothing capacitor is used. Initially the thermistor has resistance that is sufficient to limit current and generate heat. But the rise in temperature causes the resistance of the thermistor to fall. Eventually it reaches a steady state where it remains sufficiently warm to maintain a low resistance that imposes a negligible load on the circuit.

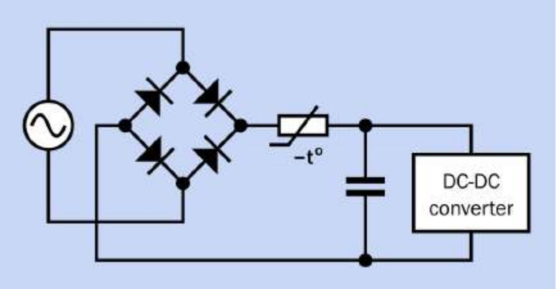

Figure 23-6 *Placement of an NTC thermistor that is designed for inrush current limiting.*

In thermistors that are used for temperature measurement, self-heating is an undesirable attribute. By contrast, an inrush current limiter depends on self-heating to perform its function.

The TDK B57237S509M inrush limiter, shown in Figure 23-7, is rated for 5A and has an initial resistance of 5 ohms at 25 degrees Celsius while not passing current. When tested with a 2,800μF capacitor at 110VAC, its resistance drops to a minimum of 0.125 ohms at 5A. The relationship between current and resistance is shown in Figure 23-8.

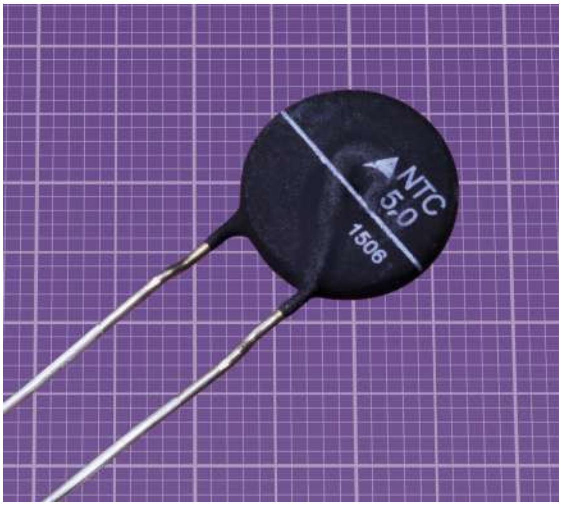

Figure 23-7 *A TDK B57237S509M NTC thermistor designed as an inrush current limiter, rated 5 ohms at 5 amps. The background grid is in millimeters.*

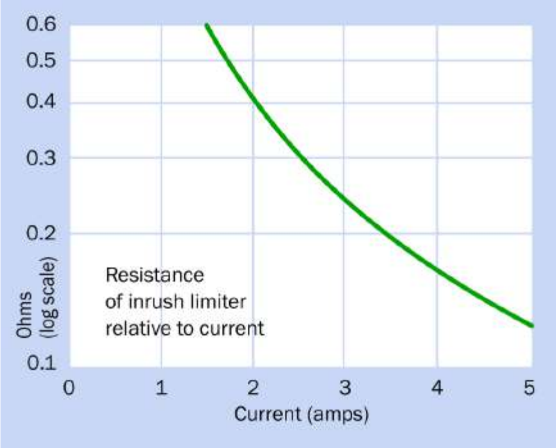

Figure 23-8 *This graph shows the relationship between resistance and current in a TDK B57237S509M inrush limiter.*

Restart

If a protected device is switched off momentarily and is then switched on again, the thermistor cannot provide protection, as it has not had time to cool down and regain its resistance. However, during the 30 seconds to 2 minutes required for heat in the thermistor to dissipate, smoothing capacitors are unlikely to lose much of their charge. Therefore, if the device is restarted, an inrush of current should not occur.

Thermistor Values

Datasheets for thermistors may be more complex and cryptic than for many components.

When examining a datasheet, first check to see if the thermistor is described as being suitable for temperature measurement or inrush current limiting. A component designed for temperature measurement will not survive inrush current, while one designed for inrush current limiting will have a very low resistance making it unsuitable for temperature measurement.

Time and Temperature

In most datasheets, lowercase letter t is used for values relating to time, whereas an uppercase T is used for values relating to temperature. Unfortunately, T may also be used as an abbreviation for "thermistor."

Resistance and Response

Letter R often means resistance, but may indicate response time, depending on the context in which it is used. For example, R_T is the resistance of a thermistor, and t_R is a response time.

Kilohms and Kelvin

Letter K may be used to represent temperature in degrees Kelvin, 0 degrees on the Celsius scale being approximately 273 degrees Kelvin. However, letter K is also used to represent thousands of ohms, sometimes in the same datasheet. In both instances, K is capitalized.

Reference Temperature

This is the temperature at which many attributes of the component are measured, such as its temperature coefficient and resistance. Usually the reference temperature is 25 degrees Celsius, but in some cases it may be 0 degrees, and other values are occasionally used. The term is abbreviated T_{REF}.

Reference Resistance

The reference resistance for a thermistor (sometimes described as its nominal resistance) may be referred to as R_R, and is the resistance at the reference temperature. It may be referred to as the "R value," but in thumbnail product descriptions it can be cited simply as "Resistance."

In datasheets, R25 or R_{25} is the resistance at 25 degrees Celsius. If this is the reference temperature, R_R and R_{25} will be the same.

Dissipation Constant

DC is the *power dissipation constant*, a ratio normally expressed as milliwatts per degree Celsius (written as mW/°C). This is a measurement of how much thermal power the thermistor can transfer to the environment for a 1 degree increase in temperature.

Temperature Coefficient

TC may be used as an acronym for the *temperature coefficient*, which represents the sensitivity of the thermistor. (Sometimes TCR is used instead of TC, the letter R denoting resistance. The two acronyms both mean the same thing.) The value is the percentage change in resistance for each change in temperature of 1 degree Celsius. Thus, if the resistance of a thermistor drops from 800 ohms to 768 ohms when the temperature increases from 28 to 29 degrees Celsius, TC = -4%. For NTC thermistors, which have a resistance that decreases when temperature increases, the temperature coefficient is negative. However, the minus sign may be omitted.

The coefficient may be expressed in parts per million (abbreviated ppm) instead of as a percentage. To convert parts per million to a percentage, divide by 10,000. Thus, a figure of 50,000ppm is equivalent to 5%.

Thermal Time Constant

Unfortunately TC is also used to represent the thermal *time constant*. If T_D is the temperature difference between the thermistor's initial temperature and a new, higher ambient temperature in which it finds itself, TC is the time it takes for the thermistor to add 63.2% of T_D to its current temperature. TC is expressed in seconds, and is defined without power being applied to the thermistor. A low thermal time constant is characteristic of a physically small thermistor that acquires heat rapidly. (TC is very similar to the concept of a time constant for a capacitor acquiring charge. See the entry on **capacitors** in Volume 1.)

Tolerance

The *tolerance* of a thermistor is a measure of its accuracy, usually at 25 degrees Celsius, unless a range of temperatures is stated. A thermistor rated for 5K, with a tolerance of plus-or-minus 1% at 25 degrees, may be found to have an actual resistance ranging from 4,950 to 5,050 ohms at that temperature. Some thermistors have a tolerance of plus-or-minus 20%. A tolerance better than plus-or-minus 1% is relatively rare.

Temperature Range

The *working temperature range* of any thermistor that uses silicon dioxide is usually between about -50 and +150 degrees Celsius (slightly wider for versions encapsulated in glass, and slightly narrower if accuracy is important).

Switching Current

For a thermistor with a nonlinear response, the switching current is the approximate current that forces a sharp transition in resistance. It is represented by I_S.

Power Limitations

Operating current is the maximum current recommended to avoid self-heating. The *power rating* is the maximum allowed power (usually 100mW to 200mW).

Interchangeability

To measure temperature reliably, two thermistors of the same type, from the same manufacturer, should display the same characteristics. This is known as *interchangeability*. A value of plus-or-minus 0.2 degrees Celsius is common for a modern thermistor, but is not often mentioned in a datasheet.

What Can Go Wrong

Self-Heating

Self-heating can affect the accuracy of an NTC thermistor that is used for temperature measurement. To get accurate temperature readings, keep the current as small as possible. When the resistance of a thermistor is at the high end of its range, brief pulses of current can be used.

Heat Dissipation

Where a thermistor is used for inrush current limiting, it will create some heat during the whole time that a device is switched on. If insufficient air space is allowed between the thermistor and other components, they may be affected.

Lack of Heat

An NTC thermistor will sometimes fail as an inrush current limiter. In very cold climates, it may never become warm enough for its resistance to drop to an acceptable level. Conversely, in a very hot location (such as close proximity to a hot-water pump) it may not get cool enough to provide adequate initial protection.

Addendum: Comparison of Temperature Sensors

A chart illustrating the five main types of contact sensors, and their variants, is shown in Figure 23-9.

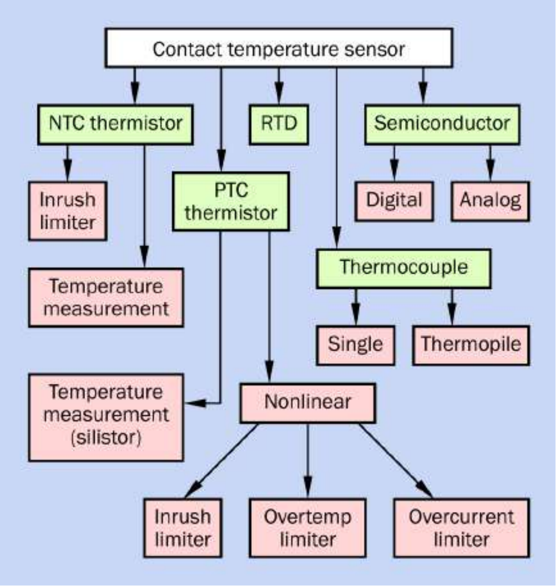

Figure 23-9 *Five types of contact temperature sensors (green boxes) and the variants (red).*

NTC Thermistor

The electrical resistance of an **NTC thermistor** diminishes as its temperature increases. Thus, it has a *negative temperature coefficient*, which is the source of the acronym NTC.

An NTC thermistor is traditionally used where low cost and simplicity are desirable and a relatively limited temperature range is acceptable (often -50 to +150 degrees Celsius). It has the advantage of familiarity, having existed in its present form for many decades. It remains the lowest-cost option among the various types of temperature sensors, and can be connected directly with an external device such as a solid-state relay, in which case no microcontroller is necessary.

PTC Thermistor

The sensing element for a positive-coefficient thermistor is a polycrystalline compound that increases in resistance very rapidly above a threshold temperature. This makes it suitable for blocking a high current to prevent circuit overload.

A *silicon temperature sensor*, sometimes called a *silistor*, can be considered as a PTC thermistor, in that it is a resistive component with a positive temperature coefficient. Its sensing element is etched into silicon.

PTC thermistors are passive, nonpolarized components with two leads or solder pads. For more information about them, see Chapter 24.

Thermocouple

This sensor consists of two wires made from different metals, joined at one end. The differing thermoelectric properties of the wires creates a very small voltage between their free ends. Thermocouples have the widest range of any contact sensor. They are simple, robust, and free from self-heating effects, as they consume no power. Their response is rapid, but very nonlinear, and their sensitivity is limited. They are used in industry and in laboratories, often plugged into a panel meter that combines a digital temperature display with hardware to decode the signal from the type of thermocouple being used.

For more information about thermocuples, see Chapter 25.

Resistance Temperature Detector

Often known by its acronym **RTD**, and sometimes referred to as a *Resistive Temperature Device*, it commonly uses a sensing element fabricated from pure platinum, nickel, or copper. The element may consist of wire wound around a core, or a very thin film deposited on an insulating substrate.

An RTD has a positive temperature coefficient, as its resistance increases while its temperature increases. It is very accurate and stable, providing an almost linear output, especially near the center of its range. However, its sensitivity is often one-tenth of that of an NTC thermistor.

Like a thermistor or a thermocouple, an RTD is a passive device, able to operate at a wide range of voltages and requiring no power supply. It is nonpolarized, with two leads or solder pads.

For more information about resistance temperature detectors, see Chapter 26.

Semiconductor Temperature Sensor

This is a chip-based sensor that requires no additional components to linearize its output, as linearization is performed in the chip.

The temperature range is similar to that of an NTC thermistor, but the output is a variable voltage with a positive temperature coefficient of about 20mV per degree Celsius, supplied by a built-in op-amp. Response time is 4 to 60 seconds.

This type of sensor requires a power supply, typically of 5VDC or less. It does not have to be calibrated before use, as it is trimmed during the production process to achieve accuracy that can be better than that of a thermistor. Manufacturers may claim plus-or-minus 0.15 degrees over the whole temperature range, which is usually -50 degrees Celsius to +150 degrees, but may be less for variants in which accuracy is more important.

The linear analog output is very convenient for use with a microntroller that has an analog-to-digital converter, and the relatively low cost makes this type of sensor increasingly competitive with thermistors.

An analog-to-digital converter may be included on the sensor chip, in which case it is often described as a *digital temperature sensor* or *digital thermometer*, with an output in degrees Celsius (or, sometimes, Fahrenheit) accessible via I2C or SPI bus. For additional details about protocols such as I2C and SPI, see Appendix A.

A *digital thermostat* or *thermostatic switch* is a semiconductor temperature sensor with a binary output that transitions from logic-high to logic-low (or vice-versa) if the temperature goes above a maximum or below a minimum level. The level can be programmed into the chip.

Semiconductor temperature sensors are identified with a variety of other names. For more information, see Chapter 27.

PTC thermistor

24

A *silistor*, or *silicon-based thermistor*, is included in this entry as a form of **PTC thermistor**.

A *resettable fuse* is not quite the same as a PTC thermistor. For more information, see the entry on **fuses** in Volume 1.

NTC thermistors, in which the resistance decreases as the temperature increases, have a separate entry. See Chapter 23.

A *resistance temperature detector* or **RTD** has a resistance that increases as its temperature increases, but it is not usually classified as a thermistor, because its sensing element is fabricated differently. Its entry will be found at Chapter 26.

Infrared temperature sensors, **semiconductor temperature** sensors, and **thermocouples** each have their own entries.

OTHER RELATED COMPONENTS

- **infrared temperature** sensor (see Chapter 28)
- **semiconductor** temperature sensor (see Chapter 27)
- **thermocouple** (see Chapter 25)
- **NTC thermistor** (see Chapter 23)
- **RTD** (resistance temperature detector) (see Chapter 26)

What It Does

The electrical resistance of a PTC thermistor increases as its temperature increases. Variants can measure temperature or can protect circuits by detecting excessive heat or current.

Because a PTC thermistor is a resistive sensor, it has no polarity. Current may flow through it in either direction, or AC may be used.

Schematic Symbols

The schematic symbol for a PTC thermistor is very similar to the symbol for an NTC thermistor. See Figure 23-1.

Comparison of Temperature Sensors

In this Encyclopedia, contact temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for

NTC thermistors. See "Addendum: Comparison of Temperature Sensors". Also see Figure 23-9.

PTC Overview

PTC thermistors can be subdivided into two groups:

- *Linear*, with a chip-sized silicon-based sensing element. They are sometimes referred to as *silistors*. The component has a very linear response and is used for temperature measurement. It may be connected directly to a microcontroller.

- *Nonlinear*, mostly using a sensing element containing barium titanate in a polycrystalline compound that increases in resistance very sharply above a threshold temperature. This type of sensor may be described as a *switching thermistor*, because its nonlinear output can activate a switching device.

The sensing elements in positive-coefficient thermistors are different in principle from the element in an NTC thermistor.

Nonlinear thermistors are used in two different ways:

Externally heated

> The thermistor responds to ambient heat or to the temperature of a device to which it is attached. It can be used to protect a circuit or a motor from overheating. Current through the thermistor is minimized to avoid self-heating.

Internally heated

> The thermistor responds to its own temperature caused by current passing through it. It can activate a warning signal or shut down equipment in the event of a short circuit. It can also control current for starting a motor or a fluorescent tube, and is sometimes used as a source to create localized heat.

Silistor for Temperature Measurement

A silicon-based PTC thermistor, sometimes known as a *silistor*, provides a highly desirable, almost linear relationship between temperature and resistance. A popular example is the KTY81 series from NXP, a sample of which is shown in Figure 24-1.

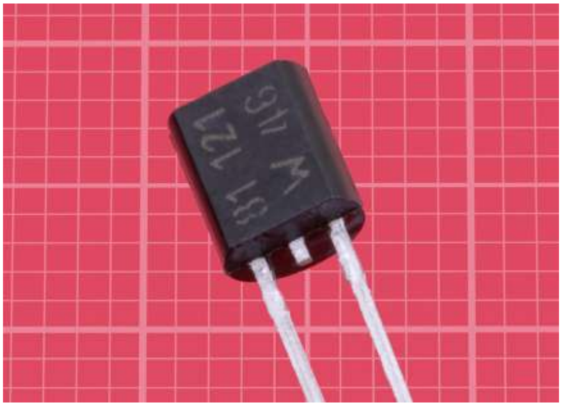

Figure 24-1 *A KTY81 thermistor from NXP. The background grid is in millimeters. Note the amputated center lead.*

The response of this thermistor is shown in Figure 24-2.

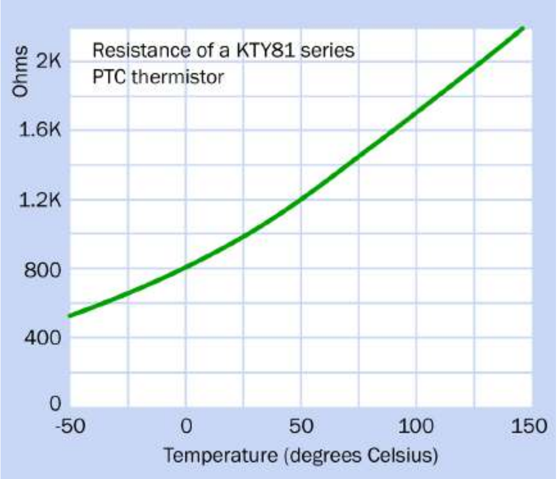

Figure 24-2 *Resistance of the KTY81 thermistor in response to temperature.*

Note that this graph is plotted with a linear vertical scale, unlike the performance curves for many thermistors that are plotted with a log scale. The log scale tends to make a response curve look flatter.

The sensor is a silicon chip designed on the “spreading resistance principle,” in which current fans out from a metal contact through a thin layer of silicon to a metallized bottom plane. This effect progresses less actively as the temperature increases. Although the result is partly dependent on polarity, a second metal contact is biased in the opposite direction, and when the two active regions of the chip are wired in series, the result is a component that has no polarity.

- The almost-linear output of this type of sensor makes it easy to use with a microcontroller that has a built in analog-to-digital converter.

Tolerance ranges from plus-or-minus 1% to 5%, depending on the temperature. Variants have a typical reference resistance of 1K or 2K. The temperature coefficient is commonly about 1%, which is considerably lower than that of a typical NTC thermistor, where 4% is common.

- Guidance on reading thermistor datasheets will be found in the entry describing NTC thermistors. See “Thermistor Values”.

For correct operation, a typical silistor requires a current ranging from around 0.1mA to 1mA.

- The lower sensitivity and slightly higher price of a PTC temperature-measurement thermistor, compared with an NTC thermistor, may explain why the NTC type seems to remain more popular, with many more variants available. In addition, the NTC type is much more tolerant of variations in current.
- Silistors continue to find some automotive applications, measuring oil temperature, transmission temperature, and climate control, among other parameters.

As a simple strategy to determine its resistance, a series resistor can be used with a PTC sensor to create a voltage divider. The circuit is identical to that used for NTC thermistors. See “Output Conversion for Temperature Sensing”.

RTDs

A *resistance temperature detector* or **RTD** is sometimes classified as a PTC thermistor. However, it has a different type of pure-metal sensing element, much lower sensitivity, and is discussed in a separate section of this Encyclopedia. See Chapter 26.

Nonlinear PTC Thermistors

Over-Temperature Protection

This type of nonlinear thermistor is externally heated, but has a switching function. If it is incorporated among other components on a circuit board, its output can be used to activate a warning signal, or can trigger a relay to shut down the circuit until the temperature subsides. This is of special interest for battery chargers where excessive heat can often be a problem, but is also useful in electronic devices generally.

To avoid the possibility of self-heating, current passing through the thermistor must be minimized to a few milliamps.

Some thermistors in the Vishay PTCSL series will make a transition at a temperature as low as 70 degrees Celsius. Others will be triggered by temperatures above 100 degrees. A typical response curve is shown in Figure 24-3, where resistance rises from 100 ohms at 25 degrees to around 1K at the transitional reference temperature of 90 degrees, and reaches at least 4K at 105 degrees.

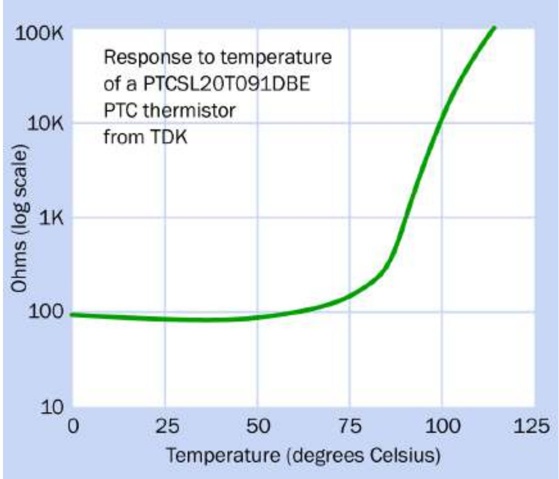

Figure 24-3 *The relationship of resistance to heat in an over-temperature protection thermistor.*

To respond to this transition, the manufacturer recommends a Wheatstone bridge circuit with its outputs connected to a comparator, as suggested for an NTC thermistor in Figure 23-5. The comparator can then activate a signal or a relay.

A picture of the PTCSL20T091DBE thermistor appears in Figure 24-4.

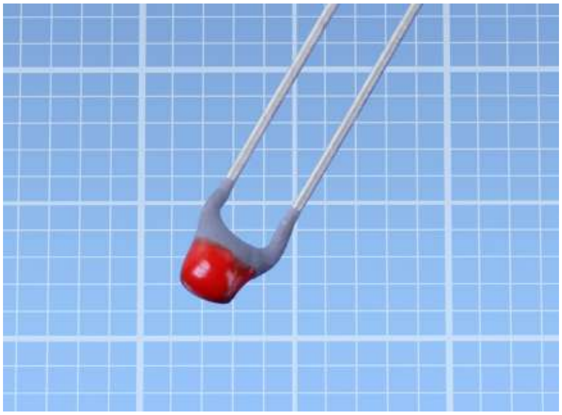

Figure 24-4 *A thermistor in the PTCSL range from TDK. It is color coded using a proprietary scheme by the manufacturer to indicate a reference temperature of 90 degrees Celsius. The background grid is in millimeters.*

This type of thermistor can tolerate a maximum of 30V (AC or DC).

Over-Current Protection

This type of nonlinear thermistor is a substitute for a fuse, as it responds to internal heat created by current passing through it. If the flow of current is excessive, the resistance of the thermistor increases, throttling the flow. When the over-current problem is resolved, the thermistor returns to its normal state. Whereas a fuse must be placed in a location allowing replacement, the thermistor is unharmed by its transition and does not have to be replaced.

Over-current may occur as a result of failure of other components, such as a rectifier diode or a capacitor, or can occur in situations such as a DC motor locking up.

The B598 series from TDK can tolerate voltages over 240V, AC or DC. They typically respond when currents exceed 100mA to 1A, depending on the specific component (a few fall outside that range), and many can withstand 1A to 7A. The B59810C0130A070 pictured in Figure 24-5 is switched by 980mA, can tolerate as much as 7A, and has a reference resistance of 3.5 ohms, rising above 10K when excessive current causes sufficient heat.

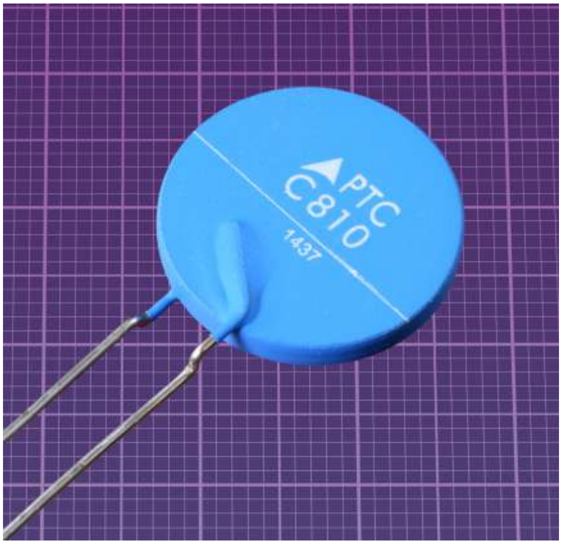

Figure 24-5 *A large over-current protection PTC thermistor. The background grid is in millimeters.*

An over-current thermistor of this type remains wired into the power supply for a device on a permanent basis. Its reference resistance will generate some heat, which is why this type of component is usually restricted to applications where the triggering current is below 1A.

The Murata PTGL07BD220N3B51B0 pictured in Figure 24-6 provides over-current protection with a reference resistance of 22 ohms, has a trip current of 200mA, and tolerates a maximum of 1.5A.

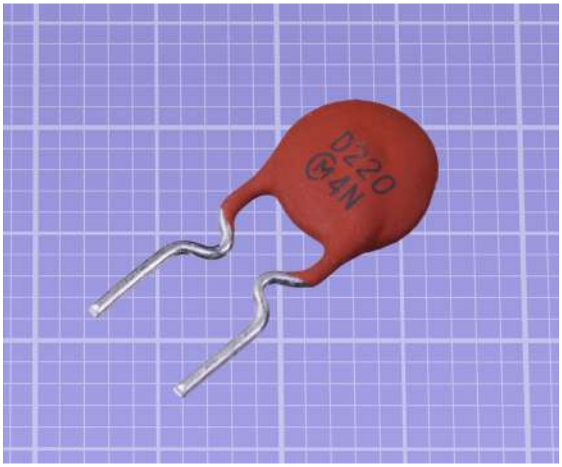

Figure 24-6 *An over-current protection PTC thermistor with a trip current of 200mA. The background grid is in millimeters.*

PTC Inrush Current Limiting

This nonlinear thermistor responds to internal heat caused by an inrush of current when power to a device is switched on. The inrush occurs when current flows rapidly into smoothing capacitors, charging them very rapidly. This can overload a power supply and shorten its life expectancy.

NTC thermistors are traditionally used as inrush limiters. The initially high resistance of this type of component blocks the surge in current, but as heating occurs, the resistance of the NTC thermistor drops rapidly. It remains in the circuit, imposing a relatively small load while the device functions normally. For more details of this application, see "Inrush Current Limiter" in the entry discussing NTC thermistors.

However, an NTC thermistor used in this way will waste some power. Suppose a supply of 120VAC is being used. If the power consumption of a device is 1,000W, the current will be approximately 8A. An NTC thermistor that has a resistance of 0.2 ohms while running warm will introduce a voltage drop of approximately 1.6V, consuming about 13W. This power loss will be greater in applications where the current is even higher—for example, in an electric vehicle recharging station.

To eliminate the loss, a timed bypass relay can be added around the thermistor. The relay closes automatically after a short interval, eliminating the power loss. This is known as *active inrush current limiting*.

However, in this arrangement, an ordinary resistor could be used instead. But in that case, why not use a PTC thermistor that has a reference (cold) resistance of 50 ohms or more? This not only limits the inrush current, but provides additional protection. If a smoothing capacitor in the circuit suffers a short circuit, or if the bypass relay fails to close, excess current passing through the PTC thermistor quickly raises its resistance, protecting the rest of the circuit.

The B5910 series of PTC thermistors from TDK is designed for inrush current limiting. They are packaged in a flame-retardant phenolic resin plastic case, as shown in Figure 24-7. The B59105J0130A020 has a reference resistance of 22 ohms that rises quickly beyond 10K when the temperature exceeds 120 degrees Celsius, as shown in Figure 24-8. This type of component is robust enough to withstand a complete short circuit across a 220-volt supply.

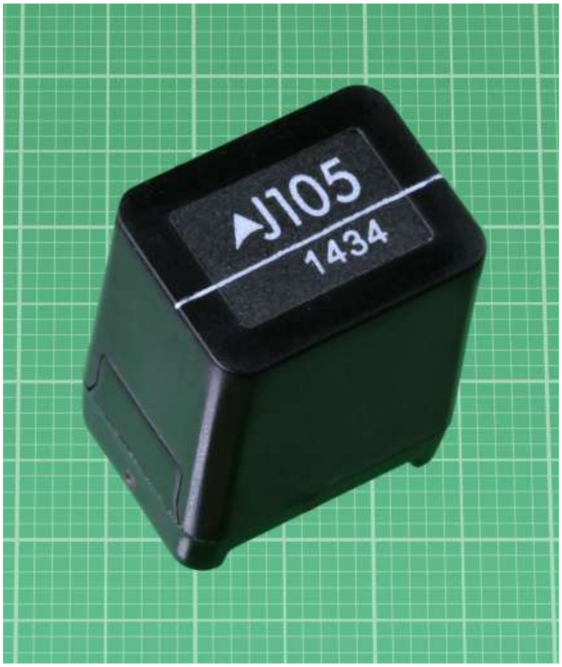

Figure 24-7 *This inrush current-limiting PTC thermistor by TDK is packaged in a flame-retardant case. The background grid is in millimeters.*

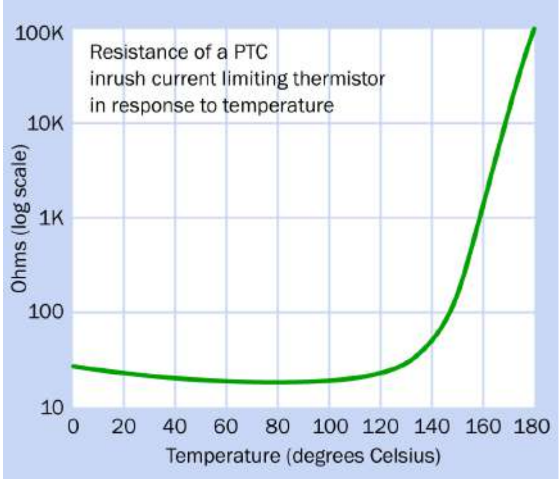

Figure 24-8 *Relationship of resistance to temperature in a PTC inrush current limiting thermistor. Note that the vertical axis has a logarithmic scale.*

PTC Thermistor for Starting Current

In some applications an initial inrush of current is actually necessary and desirable. An air conditioning compressor, for example, requires a surge of high current for "torque assist" when it is starting from a rest state.

High-current PTC thermistors may be used in this kind of situation. The Vishay PTC305C series is an example. These are heavy-duty components that have a switching time of about half a second, a maximum voltage rating of 410VAC or more, and a current rating from 6 to 36 amps.

The PTC thermistor has a relatively high temperature while the motor is running, and must be allowed time to cool before a restart is possible after shutdown. A waiting time of 3 to 5 minutes is imposed by a thermostat or separate time-delay relay.

PTC Thermistor for Lighting Ballast

The starting sequence for a fluorescent lamp requires that current should flow through the cathode heater initially. The thermistor allows this by bypassing a capacitor. Within less than a second, the resistance of the thermistor rises to block current. By this time, the heater has done its job and the lamp runs from high-frequency AC.

PTC Thermistor as a Heating Element

For small applications, a heating element can be made from a PTC thermistor, using its internal resistance to create heat. It has the advantage of being self-limiting, as its resistance rises with temperature. The TDK 5906 series is an example, shown in Figure 24-9. The component is approximately 12mm in diameter, and is designed to be clamped in place, not soldered. It has automotive applications for diesel fuel preheating and spray nozzle defrosting. Residential applications include vaporizers for air fresheners.

The initial resistance is as low as 3 or 4 ohms, rising very quickly at a transition temperature ranging from 70 to 200 degrees Celsius, depending on the specific component.

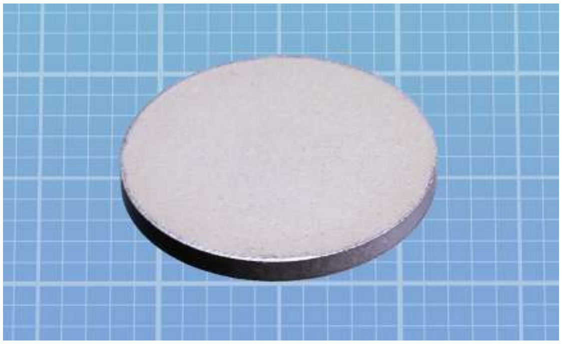

Figure 24-9 *This TDK B59060A0060A010 heating element is a PTC thermistor whose resistance rises rapidly around 80 degrees Celsius. Rated for 12VDC, it is intended for automotive applications. The background grid is in millimeters.*

What Can Go Wrong

Self-Heating

Self-heating may affect the accuracy of a temperature sensor. To get accurate readings, keep the current small. When the resistance of a thermistor is at the high end of its range, brief pulses of current can be used.

Heating Other Components

In cases where the self-heating of thermistors serves a useful purpose, as in surge protectors and when used for delays, the heat can damage nearby components or materials.

thermocouple

25

Because a *thermopile* is an assembly of thermocouples, it is included at the end of this entry. Other types of temperature sensors have their own entries.

OTHER RELATED COMPONENTS

- **NTC thermistor** (see Chapter 23)
- **PTC thermistor** (see Chapter 24)
- **semiconductor** temperature sensor (see Chapter 27)
- **RTD** (resistance temperature detector) (see Chapter 26)
- **infrared temperature** sensor (see Chapter 28)

What It Does

A **thermocouple** measures temperature by using a pair of wires made from dissimilar metals. At one end of each wire, they are joined together, often by welding them. The differing thermoelectric characteristics of the wires generates a very small voltage between their free ends, from which the temperature of the joined ends can be derived.

No power supply is needed for a thermocouple, but the voltage that it generates is extremely small (measured not just in millivolts, but microvolts) and very nonlinear, requiring hardware and/or software to convert it to a temperature value. Laboratory equipment or integrated circuit chips are available for this purpose.

Different types of thermocouples are available to measure different temperature ranges, and each type has its own characteristics, requiring appropriate conversion.

A “raw” thermocouple looks very unimpressive, as it merely consists of two wires welded together at one end. This is illustrated in Figure 25-1. The full length of the photocouple is shown in Figure 25-2.

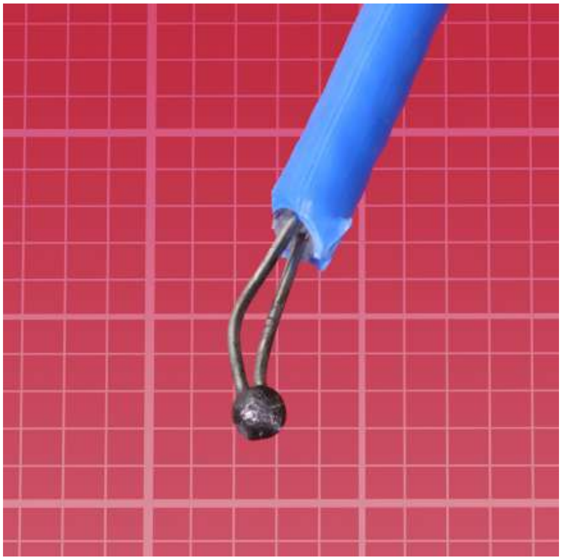

Figure 25-1 *Closeup of the welded wires in a K-type thermocouple. The background grid is in millimeters.*

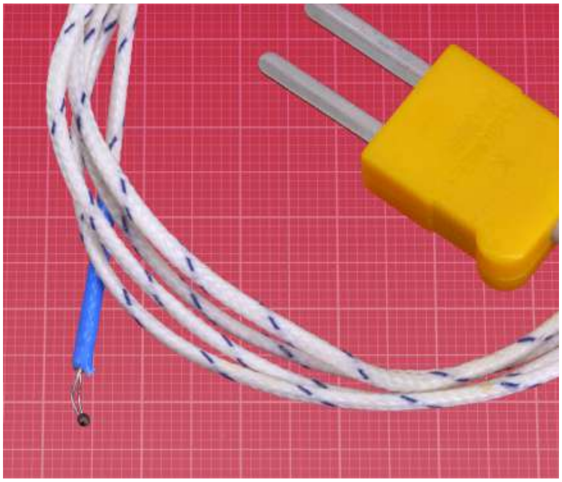

Figure 25-2 *Overview of the thermocouple in the previous photograph.*

A thermocouple sold as a commercial product is likely to be enclosed in a probe, as shown in Figure 25-3.

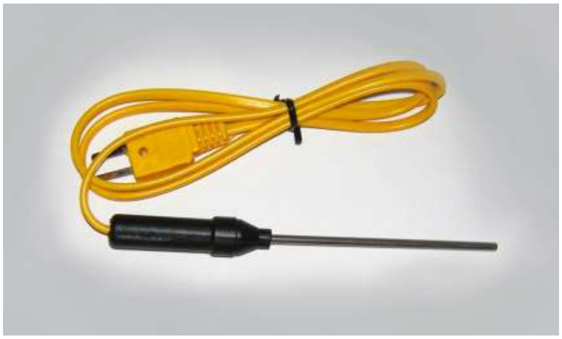

Figure 25-3 *A probe that contains a thermocouple.*

Schematic Symbol

A schematic symbol that is often used to represent a thermocouple is shown in Figure 25-4. Because this component does not consume current, the plus and minus signs do not mean that power should be applied to the wires. The positive sign indicates which wire will generate a higher voltage than the wire with the negative sign.

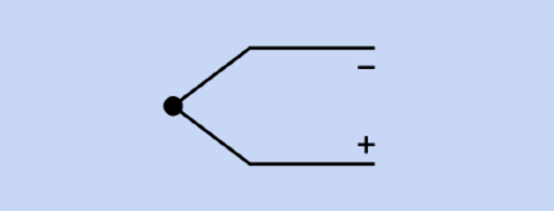

Figure 25-4 *A symbol that is often used for a thermocouple.*

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See "Addendum: Comparison of Temperature Sensors" for an overview. Also see Figure 23-9.

Thermocouple Applications

Thermocouples have a wider range than any other form of contact temperature sensor, some types being capable of measuring up to 1,800 degrees Celsius. The main limitation is the ability of the joint between the wires to withstand the heat. Appropriate insulation must be used, but segments of ceramic tube are marketed to serve this purpose if necessary.

The very small thermal mass of a thermocouple enables a rapid response to temperature fluctuations. No self-heating occurs, because the thermocouple consumes no power. It is simple and robust. However, its response is very nonlinear, and the tiny voltages involved are vulnerable to corruption by electrical noise. Accuracy is usually not better than plus-or-minus 0.5 degrees Celsius, and may be less at low temperatures.

Thermocouples are commonly found in laboratories and in some industrial applications, such as monitoring the temperature in a blast furnace or inside an internal combustion engine.

They may also measure temperatures as low as -200 degrees, but at temperatures below -100 degrees the temperature coefficient diminishes to the point where voltage increments are less than 30µV per degree Celsius.

How a Thermocouple Works

When one end of a piece of wire is maintained at a temperature that is different from the other end, the temperature gradient along the wire creates a small electromotive force that manifests itself as a difference in electrical potential between one end of the wire and the other. This is known as the *Seebeck effect*, named after the man who discovered it. The magnitude of the potential will depend on two factors: the temperature difference between the ends of the wire, and the type of wire that is used.

Figure 25-5 illustrates the concept. Part 1 of this figure shows two wires, named A and B. The left ends of the wires are heated to the same temperature, T_X, while the right ends remain at a cooler temperature, T_Y. Because the wires are composed of different metals, the voltage drop across each wire will be different.

To make this model useful, some factors must be eliminated. In part 2 of Figure 25-5, the hot ends of the wires have been welded together. This now guarantees that they share the same temperature and the same voltage, V_X. We do not yet know what these X values are.

In part 3 of the figure, the cold ends of the wires are clamped in an isothermal block, which keeps them at an equal temperature, still represented as T_Y. The block is not electrically conductive, so the cold ends of the wires still have different voltages, V_A and V_B. We cannot measure these voltages directly, because they are relative to V_X, which is unknown. However, a volt meter can measure V_A and V_B relative to each other.

The volt meter will have its own voltage gradient on its wires, and possibly a temperature gradient too, but both of these wires are made of the same metal (probably copper) and share the same temperature gradient. Therefore, their effects will be equal.

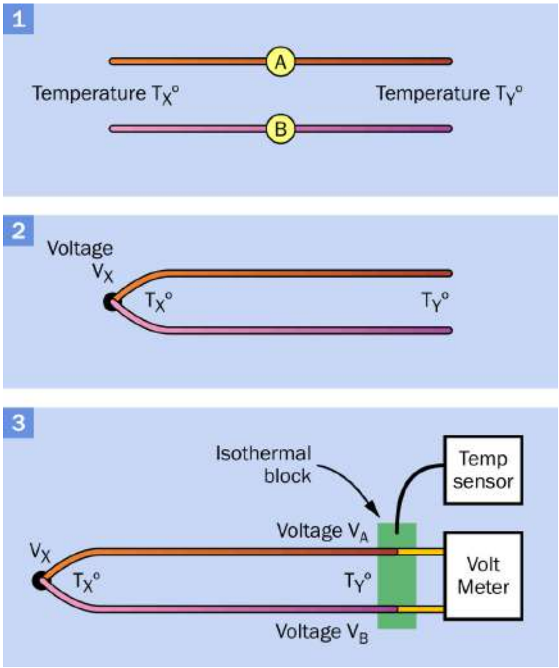

Figure 25-5 *Basic principles of a thermocouple. See text for details.*

A mathematical relationship exists between the temperature gradient and the voltage difference in each thermocouple wire. Suppose K_A is a constant or function that enables the voltage difference in wire A to be determined from its temperature gradient, and K_B serves the same function for wire B. Suppose T_{DIF} is the difference in temperature between T_X and T_Y. We may state:

$$K_A * (T_{DIF}) = V_X - V_A$$

$$K_B * (T_{DIF}) = V_X - V_B$$

By subtracting the second equation from the first and rearranging the terms, we get:

$$T_{DIF} * (K_A - K_B) = V_X - V_A - V_X + V_B$$

The two V_X terms cancel out, leaving $V_B - V_A$ on the right. We can recognize $V_B - V_A$ as the voltage difference measured by the meter. Call it V_M. So:

$$T_{DIF} = V_M / (K_A - K_B)$$

This enables calculation of the temperature difference between the ends of the wires, based on the meter reading and the conversion factor for each wire, which can be found experimentally. Because T_Y is being held at a known, constant value, we can determine the value of T_X:

$$T_X = T_Y + T_{DIF}$$

Thermocouple Details

When the thermocouple was first invented, the cold ends of the wires were placed in a bath of ice and water, forcing them to acquire and maintain a known temperature of 0 degrees Celsius.

The advent of accurately calibrated thermistors made it possible simply to measure the temperature of the cold ends. In this way, a thermistor enables a thermocouple to work. This prompts the question: why not just use the thermistor to measure T_X, and throw away the thermocouple? The reason is that a thermistor has a more limited range, seldom being used for temperatures above 150 degrees Celsius.

Note that the "hot end" of the thermocouple wires does not actually have to be hotter than the "cold end," even though those terms are commonly used. The equation to find T_X works just as well if T_Y is higher than T_X. The temperature difference will simply have a negative value instead of a positive value.

Because "hot" and "cold" are misleading terms, modern documents generally refer to the "measurement junction" and the "reference junction" of the wires. Note, however, that the wires are not actually joined with each other at the reference junction.

A common misconception is that voltage is generated where the wires are joined at the measurement junction. This is not correct. The voltage is a function of the temperature gradient between the measurement junction and the reference junction in each wire. Therefore, the way in which the wires are joined is irrelevant, provided there is an electrical connection between them. They can be welded, soldered, brazed, or crimped together.

How to Use It

Where a thermocouple is used in a laboratory, typically each wire is insulated, and they terminate in a plug that is inserted in a meter. The reference junction is hidden inside the meter, along with some electronics to decode the temperature data. The meter must have a setting that is appropriate to the specific type of thermocouple being used, so that the conversion factors are correct.

Because the type of metal in each wire must be consistent all the way from the measurement junction to the reference junction, other types of wires cannot be used to extend the reach of a thermocouple. Any extension must use wires made from the same metals. Connectors, also, must have pins and sockets that match the types of metals in the wires.

An extension wire for a thermocouple is shown in Figure 25-6.

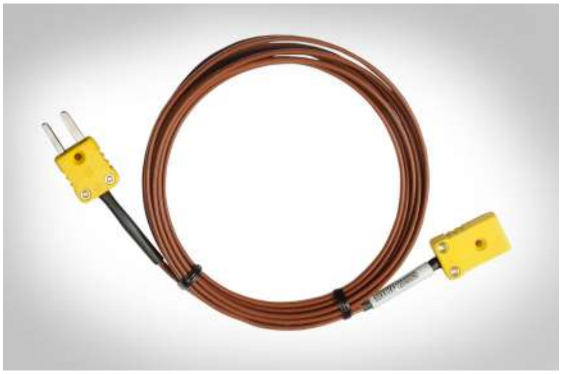

Figure 25-6 *Extension wire for a type K thermocouple. Note the polarized plug.*

Types of Thermocouples

Thermocouples are identified by ANSI standard codes consisting of single letters of the alphabet, shown below. Temperature ranges are approximate, in Celsius, with minimums rounded up and maximums rounded down to the nearest 50 degrees. Some data sources recommend narrower temperature ranges for practical use.

K type

-250 to +1,350 degrees. Most popular type of thermocouple. Positive wire is a nickel-chromium alloy, negative wire is a nickel-aluminum alloy. Commonly used in 3D printers.

J type

-200 to +1,200 degrees. Positive wire is iron, negative wire is a copper-nickel alloy. The iron wire is magnetic and vulnerable to corrosion. This thermocouple is not recommended for low temperatures, even though it is theoretically capable of measuring them.

T type

-250 to +400 degrees. Recommended for cryogenic applications. Positive wire is copper, negative wire is a copper-nickel alloy.

E type

-250 to +1,000 degrees. Most sensitive type, with the highest temperature coefficient. Positive wire is a nickel-chromium alloy, negative wire is a copper-nickel alloy.

N type

-250 to +1,300 degrees. An alternative to the K type, more stable at higher temperatures. Positive wire is a nickel-chromium-silicon alloy, negative wire is a nickel-silicon-magnesium alloy.

R type

-50 to +1,750 degrees. For high-temperature applications. Positive wire is a platinum-rhodium alloy, negative wire is platinum. Very low temperature coefficient.

S type

-50 to +1,750 degrees. For high-temperature applications. Positive wire is a platinum-rhodium alloy, negative wire is platinum. Very low temperature coefficient.

Seebeck Coefficients

Datasheets for thermocouples list the Seebeck coefficient, which is the temperature coefficient caused by the Seebeck effect, measured in microvolts per degree. In other words, the value provided by a Seebeck coefficient is the number of additional microvolts that a thermocouple will generate for an increase in 1 degree Celsius.

Each type of thermocouple has a different coefficient, and because thermocouples tend to have a very nonlinear response, the coefficients will vary with temperature. Figure 25-7 shows the variations for six types of thermocouples, over a range from -400 to +1,400 degrees Celsius. It is important to understand that the vertical scale shows the coefficient for each type of thermocouple—that is, the *change* in voltage, not the *actual* voltage, at temperature values along the horizontal axis.

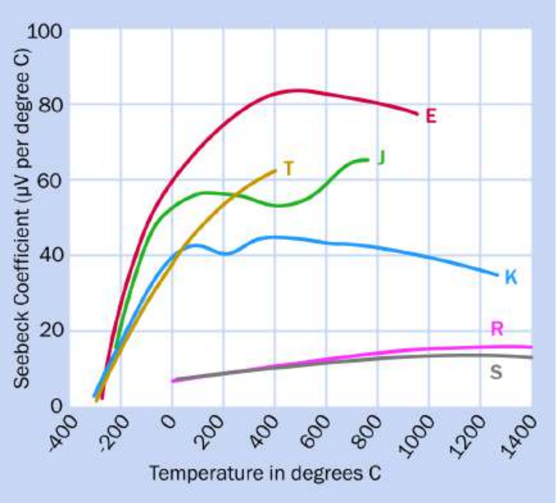

Figure 25-7 *The Seebeck (temperature) coefficient for six types of thermocouples. Partially derived from a datasheet published by Analog Devices.*

R and S type thermocouples have a relatively consistent response, but cannot achieve great accuracy, because the voltage increment is so small for each change in temperature. The K type thermocouple does relatively well between 0 and 1,200 degrees, but the J type only performs adequately between 0 and 800 degrees, while the T and E types are quite inconsistent.

For lower voltage readings, electrical noise becomes an issue. Thermocouple wires are often twisted together, and may also be shielded, to reduce sensitivity to noise. The electronics that decode the thermocouple voltage should include a filter to reject 50Hz or 60Hz interference from wiring in the vicinity.

Chips for Output Conversion

Meters that are designed to interpret the output from a thermocouple and display a temperature tend to be expensive, and may not be convenient for a custom-built application. Fortunately integrated circuit chips are now available to amplify the thermocouple output and apply signal conditioning to create an almost linear response.

The AD8494 and AD8496 from Analog Devices are precalibrated by laser wafer trimming to match the characteristics of J type thermocouples, while the AD8495 and AD8497 match the characteristics of K type. The chips require a power supply of at least 3VDC and have an analog output of 5mV per degree Celsius, enabling measurement over a range of almost 1,000 degrees. They require a very low supply current of 180µA. The manufacturer claims an accuracy of plus-or-minus 2 degrees Celsius.

The chip contains a temperature sensor which should be at the same temperature as the reference junction of the thermocouple. This means that the reference junction (typically, at the socket where a thermocouple plug is inserted) should be as close to the chip as possible, and the chip should be protected from heat created by other components. Any difference in temperature between the reference junction and the chip will create a temperature measurement error.

The MAX31855K chip from Maxim Integrated Products is another thermocouple-to-digital converter. It linearizes the output from the thermocouple and digitizes it as a temperature that is accessible by a microcontroller via a serial SPI bus. Breakout boards with this chip are available. The last letter of the chip number specifies the thermocouple type. Variants for J, K, N, T, S, and R are available.

The AD8495 is mounted on a breakout board from Adafruit, and the MAX31855K is mounted on a breakout board from Sparkfun. These boards are pictured in Figure 25-8.

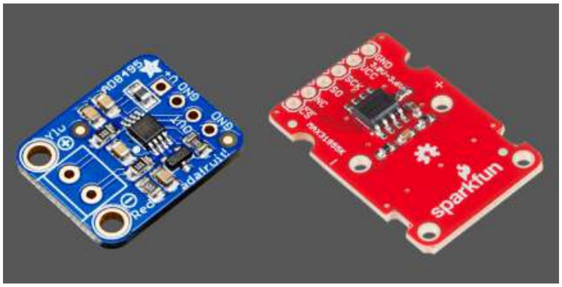

Figure 25-8 *Thermocouple amplifier/conversion chips from Adafruit (left) and Sparkfun (right).*

Thermopile

A thermopile is an assembly of thermocouples connected in series, as suggested in Figure 25-9, where a hot area is shown on the left and a cooler area is shown on the right. The figure assumes that the orange-colored wires have a voltage difference of 5mV from left to right, as a result of the temperature difference, while the purple-colored wires only have a voltage difference of 1mV from left to right. Therefore, the voltage difference between each reference junction and the next is 4mV, as shown on the right side of the figure. So long as the temperature difference exists, the voltage differences will be cumulative, totaling 16mV between top and bottom in this example.

Note that the multiple thermocouple junctions are not electrically connected with each other in each temperature zone.

In reality, more thermocouples than this will be added, and the voltage differences may be lower.

Generally a thermopile is not sold as a separate component from retail suppliers, but is built into other devices. It may be used to generate small amounts of current from a heat difference, as in an infrared thermometer. It can also be used as a safety device to shut down a gas supply if a burner is not lit. See Chapter 28.

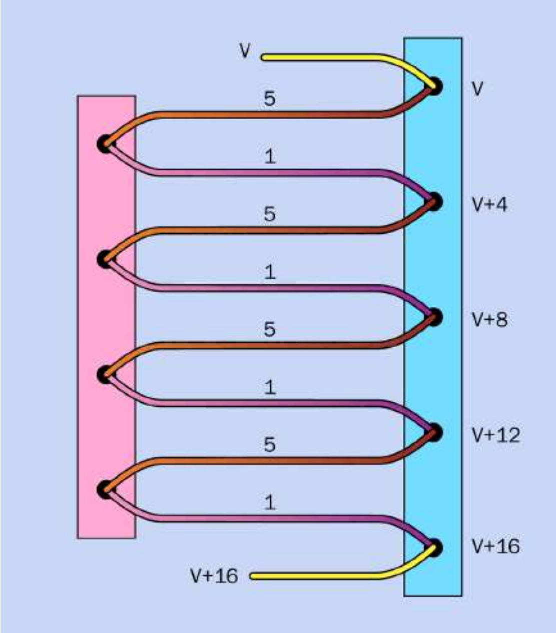

Figure 25-9 *The operating principle of a thermopile. The numbers represent mV, but are chosen arbitrarily as an example.*

What Can Go Wrong

Polarity

The output from a thermocouple has polarity. If this is not observed, an error will result.

Electrical Interference

Thermocouple wires are vulnerable to electrical interference, and should be a twisted pair or, ideally, shielded.

Metal Fatigue and Oxidation

The wire used in some thermocouples tends to be relatively brittle, and cannot withstand much flexing. Also, some metals or alloys are vulnerable to oxidation.

Using the Wrong Type

Different types of thermocouples have totally different characteristics. The electronics to decode the signal from a thermocouple must be matched to the type of thermocouple being used. The plugs on the ends of the thermocouple wires are often retained with screws. A detached plug should be replaced immediately, to avoid the error of attaching it to the wrong type of thermocouple.

Heat Damage from Creating a Thermocouple

If a thermocouple is made from two wires on a DIY basis by welding the tips of the wires together, minimal heat must be used to avoid degrading the alloys in the wires.

RTD (resistance temperature detector)

26

RTD is an acronym, either for *resistance temperature detector* or *resistive temperature device*. No definitive information seems to exist regarding which term is correct, but *resistance temperature detector* is more common.

Occasionally an RTD may be described as a *PTC thermistor*, but its sensing element is different, consisting of pure metal wire or film.

OTHER RELATED COMPONENTS

- **thermocouple** (see Chapter 25)
- **NTC thermistor** (see Chapter 23)
- **PTC thermistor** (see Chapter 24)
- **semiconductor** temperature sensor (see Chapter 27)
- **infrared temperature** sensor (see Chapter 28)

What It Does

A *resistance temperature detector*, also known as a *resistive temperature device*, is usually referred to as an **RTD**. It has a positive temperature coefficient (that is, its resistance increases as its temperature increases) but differs from a **PTC thermistor** in that its sensing element is pure metal instead of a semiconductor.

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See "Addendum: Comparison of Temperature Sensors" for an overview. Also see Figure 23-9.

RTD Attributes

Positive attributes of RTDs include:

- Accuracy, often plus-or-minus 0.01 degrees Celsius. This very small tolerance allows excellent interchangeability.
- Stability, with a response that drifts by as little as 0.01 degree per year.
- The output is an almost linear function of temperature, making them easily used with a microcontroller.
- Immunity to electrical noise.
- Reasonably rapid response to temperature changes (about 1 to 10 seconds).

Undesirable attributes include:

- Temperature coefficient about one-tenth of that of an NTC thermistor.
- To measure the resistance, some current must pass through the sensor, raising the possibility of self-heating (as in the case of other temperature sensors, with the exception of a thermocouple).
- Relatively high cost, especially of the wire-wound type.

The resistance curves for three generic NTC thermistors are shown in Figure 26-1, plotted against the resistance for a generic platinum RTD that has a reference resistance of 100 ohms at 0 degrees Celsius. Note that this graph is unlike many that illustrate the response of NTC thermistors, in that its vertical scale is not logarithmic.

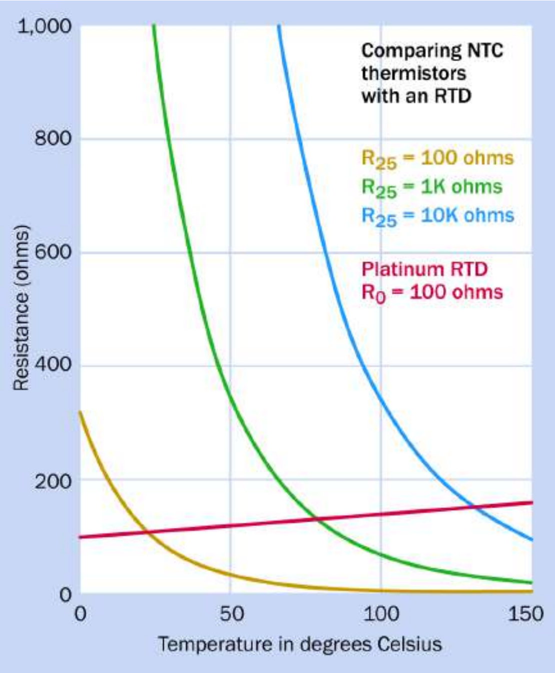

Figure 26-1 *The brown, green, and blue curves show the resistance of three generic NTC thermistors varying with temperature. The red line shows the resistance of a platinum RTD. Derived from a diagram created by Texas Instruments.*

Schematic Symbol

There is no specific schematic symbol for an RTD. Often the symbol for a thermistor may be used. See Figure 23-1.

Applications

Because of its high accuracy, an RTD may be used where precision is important. It can calibrate other temperature sensors, and may measure the temperature of a reference junction of a thermistor. However, it requires sensitive electronics for signal conditioning, because of its low temperature coefficient.

How It Works

An RTD exploits the fractional increase in electrical resistance of a metal film, metal filament, or (in some cases) a carbon film, when the temperature of the metal rises. In its simplest form, an RTD is a 2-wire device with no polarity.

The sensing element is often made from platinum, as this has a linear response to temperature over a wide range. High-quality RTD sensors with a wide temperature range usually consist of platinum wire that is wound around a glass or ceramic core. Smaller sensors may be fabricated from a thin layer of platinum evaporated onto an insulating substrate. Nickel may be substituted for platinum, and has a more sensitive but less linear response.

The wire-wound type is usable at temperatures as high as 500 Celsius (up to 1,000 degrees for some platinum-element types). Some variants are able to measure temperatures as low as -250 degrees.

DIN 60751 is an international standard defining the performance of platinum RTDs. It specifies a reference resistance of 100.00 degrees at 0 degrees, and a temperature coefficient between 0 degrees and 100 degrees. Outside of this range, a formula defines the response.

The response is almost precisely linear, ranging from 100 ohms at 0 degrees to approximately

138 ohms at 100 degrees. The temperature deviation from a straight line, between 0 and 100 degrees, is no more than plus-or-minus 0.8 degrees.

However, current through the RTD must be restricted to avoid self-heating. A range of 0.5mA to 1mA is recommended.

Variants

Some RTDs are potted in glass or resin, as shown in Figure 26-2. This shows an RTD in the TFPTL range from Vishay, containing a nickel thin-film sensing element with a temperature coefficient of about 0.4% and a tolerance of 0.01%. It is available with a very wide range of reference resistances, from 100 ohms to 5K (measured at 25 degrees Celsius). The temperature range is -55 to +70 or -55 to +150 degrees, and the maximum voltage is 30V to 40V, depending on the specific component.

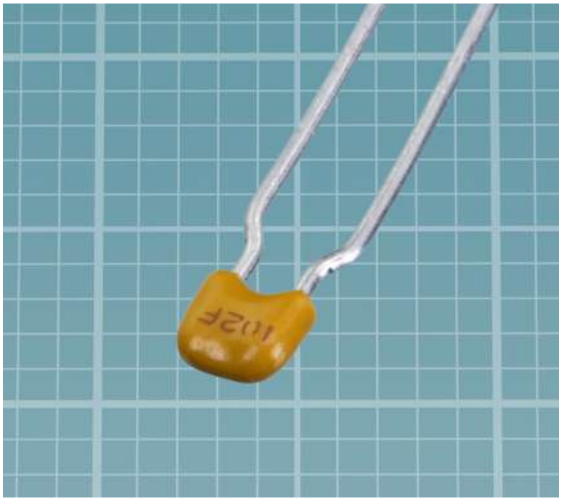

Figure 26-2 *An RTD in the TFPTL series from Vishay. The background grid is in millimeters.*

The flat package shown in Figure 26-3 may be encased in a protective sheath of plastic or silicone rubber, and can be used for surface temperature sensing where the component is glued to the exterior of a flat-sided container. This figure shows an RTD in the L420 range from Heraeus Sensor Technology, containing a platinum thin-film sensing element with a temperature coefficient of 0.385%. It is available in reference resistances of 100, 500, and 1,000 ohms (measured at 25 degrees Celsius). The temperature range is -50 degrees to +400 degrees.

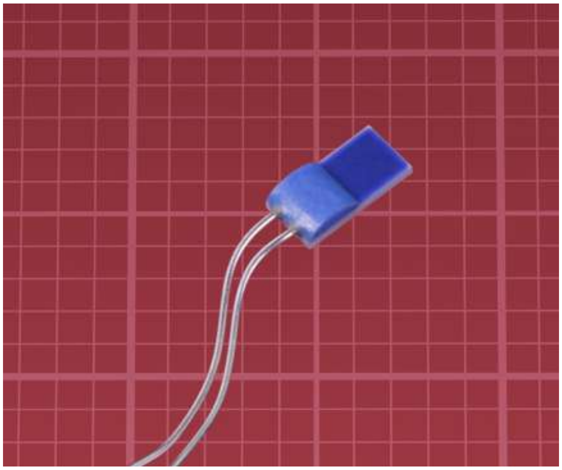

Figure 26-3 *An RTD in the L420 series from Heraeus. The background grid is in millimeters.*

Wiring

The leads to an RTD can be a source of error. If a simple 2-wire configuration is used, the leads will have an unknown resistance that will be affected by temperature, just as the element inside the RTD will have a temperature-sensitive resistance.

To enable temperature compensation, a three-wire design can be used. Figure 26-4 illustrates the principle. In the first section of this figure, resistances R_A and R_B remain unknown. In the second section, the resistances of R_B and R_C can be found by passing a test current through one wire and back through the other, bypassing the component. Assuming that all of the leads are identical in length and composition, the resistance of $R_A + R_B$ will be equal to that of $R_B + R_C$.

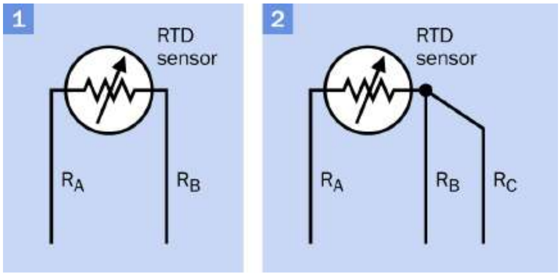

Figure 26-4 *A three-wire configuration enables temperature compensation for the leads to an RTD. See text for details.*

RTD Probe

For practical applications, an RTD sensor is often packaged inside a probe that can be indistinguishable from the type of probe used with a **thermocouple**. However, a thermocouple always uses two wires, as the wires themselves create the voltage. An RTD often uses three wires, as shown in Figure 26-5. This particular sensor is sold for use in a "Brew-Magic" system for brewing craft beer on a commercial basis.

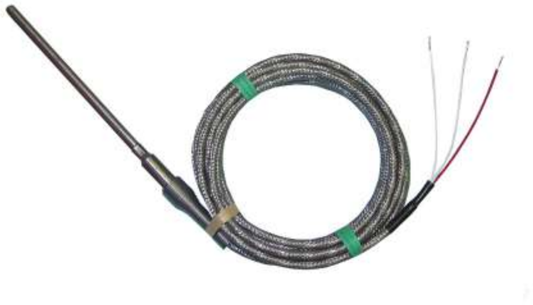

Figure 26-5 *A three-wire RTD is packaged inside this steel probe.*

Signal Conditioning

To process the signal of an RTD, a chip such as the LM75 from National Semiconductor can be used. This is calibrated for connection with a platinum RTD. Internally it converts the resistance of an RTD to a value of 5mV per degree Celsius. This then passes through an analog-to-digital converter on the chip, creating a digital value that can be read via an I2C bus.

What Can Go Wrong

Self-Heating

Self-heating is an issue for RTDs, just as it is for thermistors. Current through an RTD should be limited to 1mA, especially when measuring low temperatures.

Insulation Affected by Heat

The resistance of insulation on the wires leading to a sensor can change with temperature, leading to incorrect resistance readings. This is a more likely source of problems for RTDs than for thermistors, as RTDs are often used at higher temperatures and have a lower temperature coefficient.

Incompatible Sensing Element

If signal conditioning is applied to an RTD that has an incompatible sensing element, temperature readings will be incorrect. For example, an RTD with a nickel element should not be used with signal conditioning designed for a platinum element.

semiconductor temperature sensor

27

This type of sensor may also be referred to as a *bandgap temperature sensor*, a *diode temperature sensor*, a *chip-based temperature sensor*, or an *IC temperature sensor*.

The unfortunate term *integrated silicon-based sensor* is sometimes used, which can be confused with *silicon temperature sensor* (also known as a *silistor*), which is a type of **PTC thermistor**. See Chapter 24.

Some vendors do not divide temperature sensors into clear categories. Semiconductor temperature sensors may be classified as *board-mount* temperature sensors, even though many of them have leads and are not specifically designed to be mounted on circuit boards.

A semiconductor temperature sensor with a digital output is sometimes described as a *digital temperature sensor* or *digital thermometer*. This can be misleading, as the outputs from other types of temperature sensors may be digitized with appropriate components.

OTHER RELATED COMPONENTS

- **thermocouple** (see Chapter 25)
- **NTC thermistor** (see Chapter 23)
- **PTC thermistor** (see Chapter 24)
- **infrared temperature** sensor (see Chapter 28)
- **RTD** (resistance temperature detector) (see Chapter 26)

What It Does

A **semiconductor** temperature sensor is an integrated circuit chip incorporating a sensing element composed of transistor junctions. It has an approximately linear response and is easy to use, in some cases being designed for direct connection with a microcontroller, requiring no additional components.

In analog variants, the output consists either of voltage or current that varies with temperature. These components have a positive temperature coefficient, except for a few CMOS variants where a voltage output diminishes as temperature increases.

Digital variants are becoming more common, creating a numeric output accessible by a microcontroller.

In almost all semiconductor temperature sensors, the characteristics of silicon dioxide limit the temperature range to approximately -50 to +150 degrees Celsius (sometimes less).

This type of component is not (yet) as low in cost as a thermistor, but can include its own amplification, signal processing, and (optionally) analog-to-digital conversion on one chip.

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See "Addendum: Comparison of Temperature Sensors" for an overview. Also see Figure 23-9.

Semiconductor Temperature Sensor Applications

When a semiconductor temperature sensor is used in surface-mount format, it can measure the temperature of the board on which it is mounted. This enables protection from overheating, often in power supplies.

Because the sensing elements and signal processing circuits are all chip-based, they can be transplanted into other types of sensors. For example, a gas pressure sensor or a proximity sensor can have onboard compensation using a semiconductor temperature sensor. They have also been built into computer CPUs such as the Pentium series from Intel.

Some variants are manufactured in a three-lead TO-92 package, appearing superficially similar to bipolar transistors. They are suitable for remote temperature sensing, and have automotive applications such as measuring the temperature of the transmission, engine oil, or cabin interior. They may also be found in some heating and air-conditioning systems, and some kitchen equipment.

Schematic Symbol

No unique schematic symbol has been developed for a semiconductor temperature sensor. It may be represented by a rectangle containing text abbreviations to represent pin functions, similar to other types of integrated circuit chips.

In the case of a sensor with an output consisting of current that varies with temperature, the sensor may be shown as a *current source*, using the symbol in Figure 27-1. However, this symbol is not specific to temperature sensors; it is used for any component that is a current source.

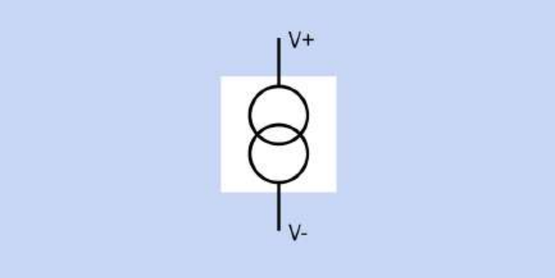

Figure 27-1 *A temperature sensor whose output consists of current varying with temperature may be shown in a schematic as a current source, using this symbol.*

Attributes

Desirable attributes of a semiconductor temperature sensor include:

- Easy to use. Few or no external components may be required, and little or no signal processing.
- Factory-calibrated, with an almost linear response.
- Versions with a digital output are easy to add to any system that already has an I2C bus. For additional details about protocols such as I2C, see Appendix A.

Undesirable attributes of semiconductor temperature sensors include:

- Limited temperature range, the same as thermistors.
- Self-heating issues, especially in versions where signal-processing functions are built into the same chip.

- Not as rugged as some types of temperature sensors.

For an explanation of terminology used for temperature sensors in datasheets, see "Thermistor Values".

How It Works

When a constant current is flowing through a p-n junction in a diode, the voltage across the diode will change by about 2mV for each change in temperature of 1 degree Celsius. This can be demonstrated by the simple circuit shown in section 1 of Figure 27-2.

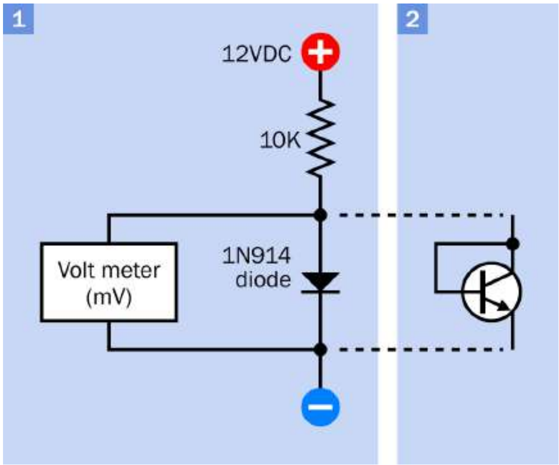

Figure 27-2 *Left, a basic circuit for demonstrating the temperature sensitivity of a diode. Right, an NPN transistor can be substituted to emulate the diode.*

Similarly, the voltage across the p-n junction in an NPN transistor varies with temperature, if the current is constant. A transistor can be substituted for a diode as suggested in section 2 of Figure 27-2. Integrated circuit chips that contain transistors can measure temperature by exploiting this phenomenon.

CMOS Sensors

Some semiconductor temperature sensors use CMOS instead of bipolar transistors. The general concept is similar, but they are described separately, below. See "CMOS Semiconductor Temperature Sensors".

Multiple Transistors

The heat sensitivity of a bipolar transistor can be defined with an equation. If the base-emitter voltage is V_{BE}, q is the charge of an electron, k is a constant (known as Boltzmann's constant), T is the temperature in degrees Kelvin (relative to absolute zero), I_C is the collector current, and I_S is the saturation current (which is less than I_C):

$$V_{BE} = ((k*T) / q) * \log_e (I_C / I_S)$$

The term $\log_e$ means, "the logarithm to base e of the expression in parentheses."

Because k and q have known values, the base-emitter voltage turns out to be proportional to the logarithm of the collector current divided by the saturation current. However, the saturation current depends on the geometry of a transistor, and varies with temperature in a nonlinear way.

To eliminate the factor of saturation current, one transistor can be compared with another transistor that has a larger emitter area. This enables the derivation of a new equation that specifies temperature while getting rid of the troublesome saturation currents, with their nonlinear behavior.

However, it may not be easy to fabricate two transistors, in the same silicon chip, that have the same characteristics except that the emitter area of one is bigger than that of the other. It is much easier to add multiple transistors in parallel, each of them identical to the first. The total emitter area will then be equal to the area in one transistor multiplied by the number of transistors.

In Figure 27-3, assuming all the transistors are at the same temperature in addition to being of identical specification, we can now write two equations. The figure allows room to show only three transistors, but suppose there are N of them. If V_{BE0} is the base-emitter voltage of transistor Q0, on the left, and V_{BEN} is the aggregate base-emitter voltage of the N transistors on the right:

$$V_{BE0} = ((k*T) / q) * \log_e (I_C / I_S)$$

$$V_{BEN} = ((k*T) / q) * \log_e (I_C / N*I_S)$$

From these, an equation can be derived that gets rid of I_C and I_S:

$$V_{BE0} - V_{BEN} = ((k*T) / q) * \log_e (N)$$

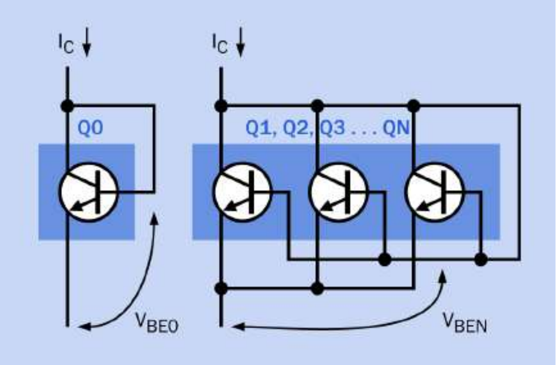

Figure 27-3 *Comparing the base-emitter voltage of one transistor with a set of identical transistors can enable measurement of temperature regardless of the collector current and saturation currents, so long as all the transistors are at the same temperature. See text for details.*

PTAT and the Brokaw Cell

Now if a comparator is added to control the current, a circuit known as the Brokaw Cell is created, shown in Figure 27-4. This is also known generically as a *bandgap temperature sensor*. (A couple of resistors have been omitted for the sake of simplicity.)

Typically, N = 8. That is, there is a set of eight transistors in addition to Q0 (only three being shown here). The voltage difference in the previous equation, $V_{BE0} - V_{BEN}$, now appears across R2 in the figure, and the voltage across R1 is *proportional to absolute temperature*, often referred to by its acronym, *PTAT*. This voltage can be found from this equation:

$$V_{PTAT} = ((k*T)/q) * \log_e(N) * (2*R1/R2)$$

The Brokaw Cell was the basis of the AD580 chip introduced in 1974 by Analog Devices, and the principle is now used very widely in semiconductor temperature sensors.

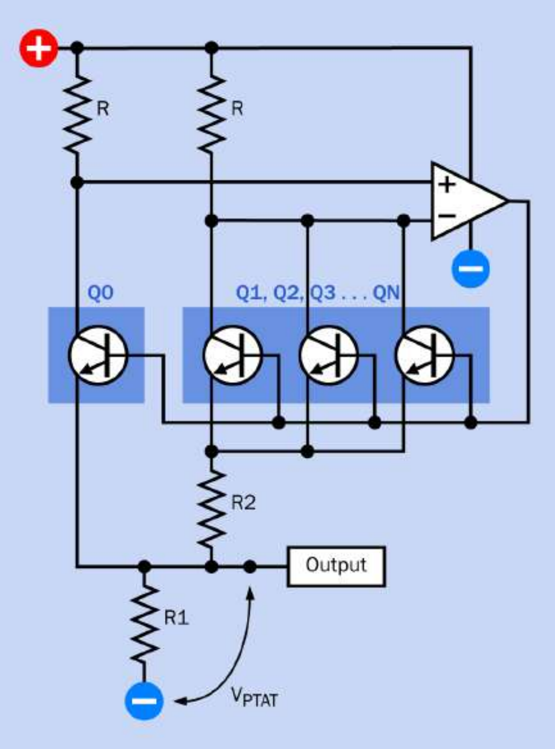

Figure 27-4 *The Brokaw Cell. See text for details.*

Variants

Three output types are used:

- Analog voltage output (voltage varies with temperature).
- Analog current output (current varies with temperature).
- Digital output.

A fourth type creates an output in the form of a square wave, either with a frequency or wavelength proportional to temperature. The Maxim MAX6576 and MAX6577 are examples. However, this type of output is so rare, it is not described in detail here.

Some semiconductor temperature sensors are CMOS-based, and have a voltage output with a negative temperature coefficient. They are described separately. See “CMOS Semiconductor Temperature Sensors”.

Analog Voltage Output

LM35 Series

The LM35 is a typical, widely used semiconductor temperature sensor, available from Analog Devices, Texas Instruments, and other manufacturers. Its output voltage changes by 10mV per degree Celsius over a range of approximately -50 degrees to +150 degrees. Accuracy is stated to be plus-or-minus 0.25 degrees at room temperature and plus-or-minus 0.75 degrees over the whole range.

The sensor can be obtained packaged like a transistor, in a TO-92 plastic capsule or metal can. It is also available as a surface-mount component, or in a TO-220 package, like a 5V voltage regulator, as shown in Figure 27-5.

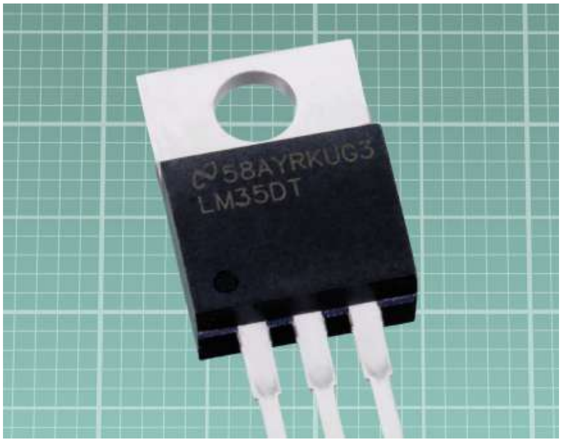

Figure 27-5 *This version of the LM35 can be used to measure surface temperature when secured with a bolt. The background grid is in millimeters.*

This is a three-wire component, two pins or solder pads being used for the power supply while the third serves as the sensor output. The supply voltage typically ranges from 4V to 30V. Necessary current consumption is only 60μA, which minimizes self-heating.

Because this device is specifically designed for the Celsius temperature range, its output is scaled to 0mV at 0 degrees. A pulldown resistor can be added to measure temperatures below zero.

A bypass resistor of 200 ohms between the output and ground is recommended as a precaution against capacitive effects in the cable run.

The LM34 is almost identical to the LM35, except that its output changes by 10mV per degree Fahrenheit instead of 10mV per degree Celsius.

LM135 Series

Although this sensor contains multiple NPN junctions, the manufacturer describes it as behaving like a zener diode in which the breakdown voltage is directly proportional to absolute temperature. The output increases by 10mV per degree over a range from -55 to +150 degrees Celsius.

For the LM135, the manufacturer claims an error of less than plus-or-minus 1 degree Celsius between 0 and 100 degrees. For the LM235 and LM335, in the same product series, the temperature range is narrower, the accuracy is lower, and the price, also, is lower. An LM335 sensor is shown in Figure 27-6.

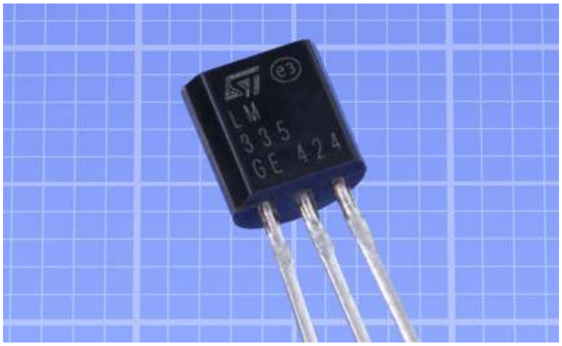

Figure 27-6 *A sample of the LM335 temperature sensor in a TO-92 package. The background grid is in millimeters.*

The sensor is available in a TO-92 package (plastic, like a transistor) or a TO-46 (metal can). It is also manufactured in a surface-mount format. The negative terminal is connected directly to ground, while the positive terminal is connected through a series resistor to the positive side of a power supply that can range from 5V to 40V. The third terminal, labeled "ADJ" on datasheets, allows for the output to be adjusted. Figure 27-7 shows the basic circuit. The value of R1 can be chosen to establish an optimal current of 1mA through the sensor, although a range of 400µA to 5mA is tolerable.

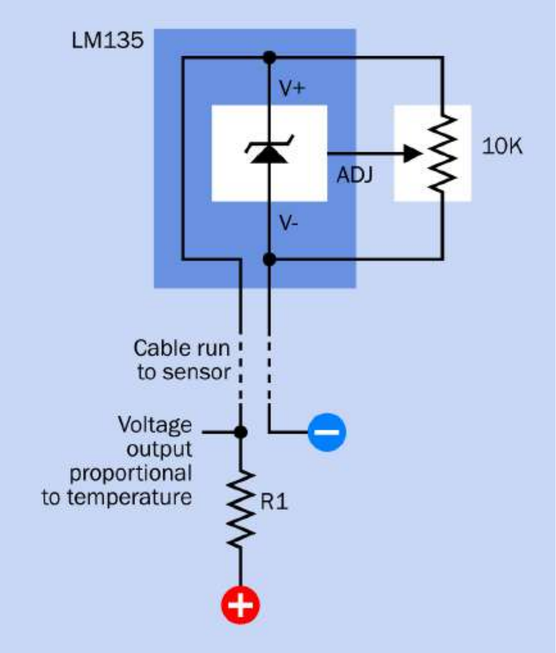

Figure 27-7 *Basic schematic for using an LM135, including output adjustment. Because the sensing element behaves like a zener diode, it is represented with the zener symbol.*

Analog Current Output

Fewer components exist using output current to measure temperature. The output is applied to a grounded resistor, and the voltage across the resistor then changes with current from the sensor.

The useful aspect of a current output is that its accuracy is unaffected by a cable run as long as 200 or 300 feet. Therefore, this type of component is appropriate as a remote sensor.

LM234-3 Series

This is a three-wire sensor, two wires being used for bias power supply and ground (labeled V+ and V- on the datasheet) and a third (labeled R) that delivers current proportional to temperature. The current from pin R passes through an external resistor to ground, and the voltage across this resistor varies by 214µV per degree Kelvin. A bias voltage ranging from 1V to 40V is required.

A sample of the LM234 is shown in Figure 27-8.

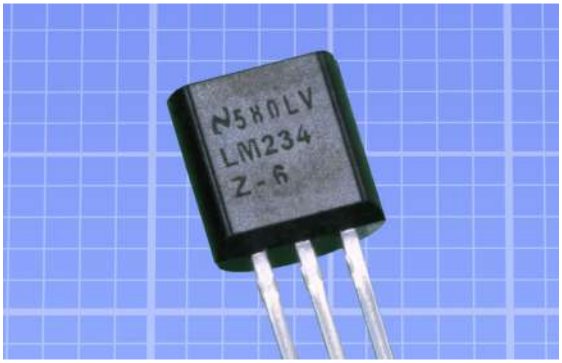

Figure 27-8 *The LM234Z temperature sensor in a TO-92 package. The background grid is in millimeters.*

If the component is used for remote sensing, the resistor should be 230 ohms and can be connected directly between pin R and pin V- of the sensor at the far end of a wire run. At the "home" end of the wire run, temperature output is taken from above a 10K resistor that is placed between the return wire and ground, as shown in Figure 27-9. With these component values, the output voltage will change by 10mV per degree Kelvin.

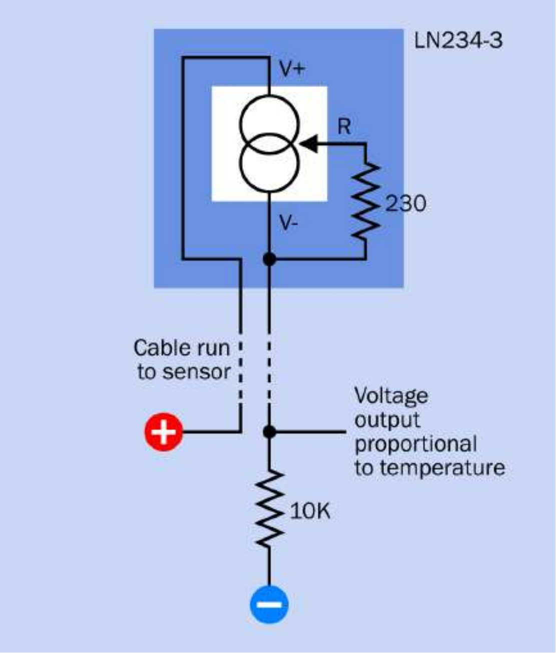

Figure 27-9 *Using an LM243-3 sensor with a current output that varies with temperature.*

The LM234-3 can be encapsulated in a plastic TO-92 package or a TO-46 metal can. A surface-mount version is also available.

The claimed accuracy is plus-or-minus 3 degrees. The temperature range is -25 degrees to +100 degrees.

AD590 Series

The Analog Devices AD590 (successor to the original AD580) is a current-output sensor that uses only two wires. Like the LN234-3 it is available in a TO-46 metal can, but with one lead making no internal connection. It also can be bought in a two-wire “flatpack,” or as a surface-mount chip (with eight solder pads, only two of which are connected).

Using a supply voltage of 4V to 30V, the sensor’s high-impedance output changes by 1μA per degree Kelvin. Voltage supply variations produce very small errors in the output current; substituting 10V for 5V creates a deviation of only 1μA.

Figure 27-10 shows an application for the AD590 using a resistor and a trimmer to adjust the scale factor. When properly set up, this circuit provides an output that changes by 1mV per degree Kelvin.

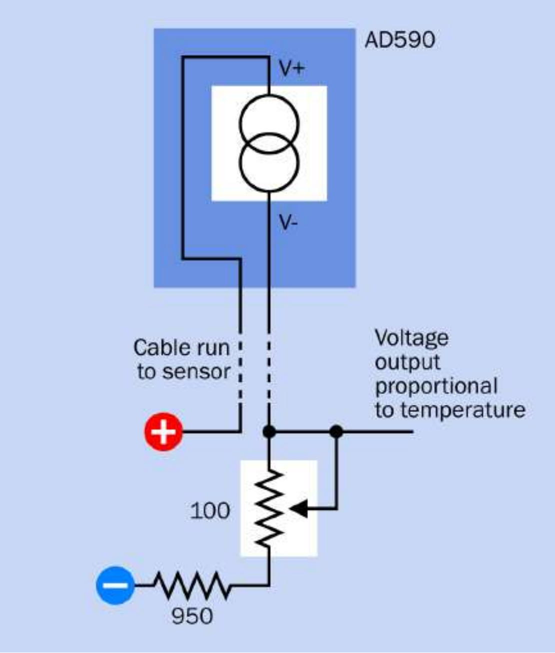

Figure 27-10 *The simplest circuit allowing fine adjustment for use with an AD590 sensor.*

Digital Output

Some of the most popular examples of semiconductor temperature sensors with digital output include the TMP102 series from Texas Instruments, MCP9808 series from Microchip, LM73 series from Texas Instruments and National Semiconductor, and DS18B20 series from Maxim. All of these components can measure a range typical of semiconductor temperature sensors, from approximately -50 to +150 degrees Celsius. Most of them claim an accuracy in the region of plus-or-minus 1 degree over the full range or 0.5 degree in the 0 to 100 range. With the exception of the Maxim DS18B20, which uses its own unique protocol, the components communicate via either the I2C or SMBus protocol.

TMP102 Series

This is available only in a surface-mount format. It has fewer features than the other sensors listed here, and is less accurate, claiming plus-or-minus 3 degrees Celsius over its maximum range of -40 to +125 degrees Celsius. However, it is less expensive. (For greater accuracy, the TMP112 is available.) This is a low-voltage chip requiring 1.4V to 3.6V as its power supply, drawing a quiescent current of 10µA. Temperatures are stored in a 12-bit or 13-bit format that requires some conversion, as a single bit represents 0.0625 degrees Celsius. An alert pin is activated when the measured temperature deviates above or below limits that are preset by the user. No hysteresis adjustment is available for the alerts. The TMP102 is available on a breakout board from Sparkfun, as shown in Figure 27-11.

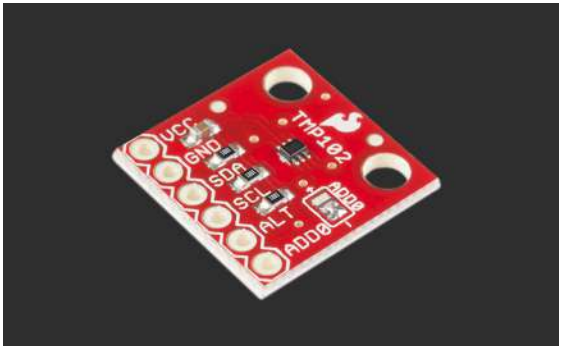

Figure 27-11 *The Texas Instruments TMP102 on a breakout board from Sparkfun.*

MCP9808 Series

This multifunction sensor is available either as a regular surface-mount component, or as a surface-mount with an exposed "thermal pad." It conforms with the I2C bus standard at up to 400kHz, allowing up to 16 sensors to share the same bus. The chip has a variety of temperature alert features, including high and low limits that can activate a dedicated "alert" pin, and a hysteresis value that can be set for the limits, to ignore brief temperature excursions. The chip can be put into "comparator mode," where it simply provides logic-high or logic-low output if the temperature is above or below a user-specified value. This feature makes the chip operate as a thermostat. Temperature resolution is user-selectable. The temperature storage format requires some conversion to obtain a Celsius value, to deal with negative values and fractional values. However the chip is available on a breakout board from Adafruit, as shown in Figure 27-12, and an Arduino code library is available.

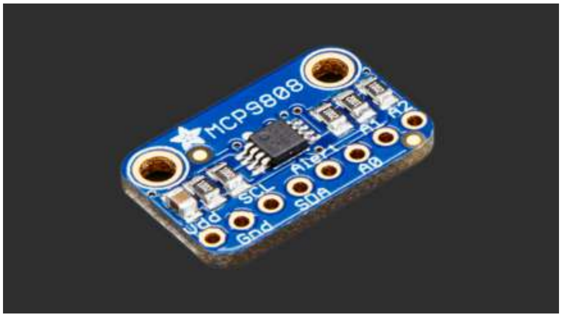

Figure 27-12 *The Microchip MCP9808 on a breakout board from Adafruit.*

LM73 Series

This sensor is only available in surface-mount format. It conforms with the I2C bus standard at up to 400kHz. Its temperature resolution can be set to 11, 12, 13, or 14 bits. An "alert" pin becomes active if the temperature exceeds a preprogrammed limit. An "address" pin can select one of three device addresses by being held in logic-high, ground, or disconnected status. The chip can be put into shutdown mode when power conservation is necessary.

DS18B20 Series

Unlike most digital sensors, this is a three-wire component, because it uses Maxim's proprietary "1-wire bus" with a unique protocol. The bus allows access to a 2-byte register storing digital output from the temperature sensor, but also allows the user to perform other functions, such as setting the resolution of an onboard analog-to-digital converter (which has a maximum resolution of 12 bits), setting a high-temperature and low-temperature alarm, and allowing the sensor to be identified, as each component has a unique 48-bit serial number in ROM.

The chip can draw sufficient power from the data bus to operate, so long as the bus is held high by a 4.7K pullup resistor. (Maxim describes this as "parasite power.") An internal capacitor sustains the chip briefly while the bus is used for its normal purpose of transferring data, but if the bus has low voltage for more than 480µs the chip will reset itself. The "parasite power" feature also will not work above 100 degrees Celsius. Perhaps recognizing that this system may create more problems than it solves, Maxim has also given the component a normal power input pin.

The DS18B20 is available in the TO-92 package and two sizes of surface-mount chip. Its lack of a standard I2C bus, and its use of complicated proprietary codes, create a steep learning curve. Still, it remains a popular sensor, and an Arduino code library for it is available online.

CMOS Semiconductor Temperature Sensors

CMOS variants of semiconductor temperature sensors have appeared relatively recently compared with the bipolar variants. They draw a very low quiescent current (typically, a few microamps) and can work with a power supply from 5.5VDC down to 2.2VDC, making them suitable for handheld battery-powered devices. An analog output is common. Popular examples are the LM20 and the LMT86 series.

Like bipolar sensors, the LMT86 sensors have a limited temperature range, between approximately -50 and +150 degrees Celsius. Again, like the bipolar sensors, they are available optionally in TO-92 and surface-mount packages. A significant difference is that the output has a negative temperature coefficient, diminishing by 10mV per degree Kelvin, because of the characteristics of CMOS semiconductors. The claimed accuracy is plus-or-minus 0.25 degrees Celsius. The output voltage covers a range of about 2V, diminishing from 0.5V below the supply voltage at -50 degrees Celsius.

A sample of the LMT86 is shown in Figure 27-13.

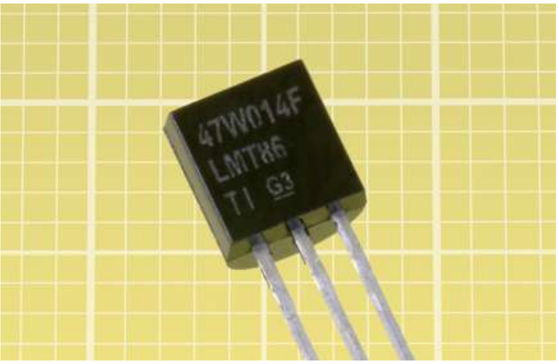

Figure 27-13 *A sample of the LMT86 CMOS temperature sensor. The background grid is in millimeters.*

What Can Go Wrong

Different Temperature Scales

Some voltage-output sensors create an output convertible to degrees Kelvin, while others use degrees Celsius. While the temperature degrees are the same in each scale, a component with an output in millivolts may assign 0mV either to 0 degrees Celsius or 0 degrees Kelvin (equivalent to -273.15 degrees Celsius). The advantage of using a Kelvin scale is that it avoids the problem of negative temperature values.

Rarely, a sensor may use degrees Fahrenheit.

Interference in Cable Runs

Sensors with a voltage output are susceptible to electrical interference. Twisted-pair or shielded cable runs are recommended when sensors are placed remotely.

For the Maxim DS18B20, which uses a 1-wire bus, multiple sensors should be connected along one run of wire (linear topology) instead of each sensor being connected to a central point (star topology). If the cable lengths are longer than a few meters, the topology starts to matter.

Latency

The packaging of semiconductor temperature sensors can create latency in their response time. While a thermocouple consists only of a pair of wires joined by a small dot of melted metal, a TO-92 semiconductor package adds thermal mass that will slow the response considerably. Moreover, copper leads will conduct heat from a circuit board if the board is warmer than its environment.

Surface-mount chips have a very low mass, but must be soldered to some kind of board, even if it is a very small one. Generally, other types of sensors may be appropriate where rapid response is necessary.

Processing Time

In a sensor with a digital output, the onboard analog-to-digital converter will add a small delay before the data becomes available, and during that delay, the component cannot respond to a new temperature. The output from an analog device may be more suitable for rapid detection of temperature variations.

infrared temperature sensor

28

An **infrared temperature** sensor is sometimes described as a *thermopile*. In reality the sensor module *contains* a thermopile. In this Encyclopedia, a thermopile is considered to be a separate component, described in the entry discussing **thermocouples**. See "Thermopile" in Chapter 25.

Other terms that are sometimes used for an infrared temperature sensor are *contactless thermometer* or *infrared thermometer*. This Encyclopedia classifies a thermometer as a commercially marketed product, not a component.

Devices such as a *radiation pyrometer*, *IR pyrometer*, *optical pyrometer*, or *thermal imager* provide ways of measuring infrared radiation, but are outside the scope of this Encyclopedia.

A **passive infrared** motion sensor (PIR) can detect infrared radiation, but only responds to fluctuations in intensity. An infrared temperature sensor measures the steady-state value of incident radiation.

OTHER RELATED COMPONENTS

- **passive infrared** motion sensor (see Chapter 4)
- **thermocouple** (see Chapter 25)

What It Does

Most temperature sensors discussed in other entries in this Encyclopedia are *contact sensors*, meaning that to measure the temperature of an object, liquid, or gas, they must make contact with it. In situations where contact is not possible or desirable, an **infrared temperature** sensor can be used. It responds to the *black-body radiation* (sometimes known as *characteristic radiation*) that is emitted by all materials above absolute zero (0 degrees Kelvin). This varies with temperature as a result of the movements of molecules.

Situations where noncontact sensors may be preferable to contact sensors include:

- An object is inconveniently located or too far away.
- The temperature of a large area must be measured.
- Contact with a small object would change the temperature of the object. The act of measurement would change the value being measured.
- The object is corrosive, abrasive, or otherwise liable to damage a sensor.
- The object is moving or vibrating.
- The surface of the object must not be contaminated (for example, unprotected foods).

- The temperature of the object is lower than around -50 degrees Celsius or higher than 1,300 degrees Celsius.

However, noncontact sensors have some limitations:

- Normally, only the surface temperature of a target can be measured.
- The optics of the sensor must be protected from dust, dirt, and liquid.
- The target must be clearly visible, in line-of-sight.
- Air pollution will degrade the temperature measurement. Some gases, such as carbon dioxide, will tend to absorb infrared radiation.
- The sensor will be affected by other sources, including reflected, transmitted, and convective heat.
- While an infrared sensor can theoretically respond to a very wide range of temperatures, in practice separate sensors of differing sensitivity are needed to cover a full range.
- Different types of materials emit differing intensities of black-body radiation, even if they are at the same temperature. Some compensation is necessary, or the surface of the object may have to be painted.

Applications

Handheld contactless thermometers were an early application for noncontact sensors.

In astronomy, the thermal radiation from the Sun and other stars is of interest to astronomers.

More recently, the declining cost of an infrared temperature sensor, and the ease of deploying it, have made it appropriate in consumer products. A significant area of adoption is in notebook computers and handheld devices, where processor performance must be balanced against the need to prevent the case from becoming too hot to hold comfortably. In this kind of application, gluing a sensor to the interior of the case would be a manual operation during the production process and would require a wired connection. An infrared temperature sensor mounted on the circuit board, viewing the underside of the case, can achieve the same objective more simply.

A contactless sensor is also very useful for measuring the temperature of rotating objects, such as heating rollers in a laser printer.

Schematic Symbol

No specific schematic symbol exists for an infrared temperature sensor.

How It Works

While nanometers (abbreviated nm) are generally used to measure visible wavelengths, the longer wavelengths of far-infrared are often measured in micrometers (abbreviated μm). The measurable infrared values are defined as ranging from 0.7μm to 14μm, corresponding to peak emissions from a black body ranging in temperature from 200 degrees Kelvin to 6,000 degrees Kelvin (about -70 to +5,700 Celsius).

Unfortunately an object does not emit just one wavelength of black-body radiation for each temperature value. It emits a spread of wavelengths that becomes wider as the temperature increases. However, the peak intensity also increases with temperature, when measured as *spectral radiance*, which is defined as watts per steradian, per micrometer of wavelength. (A steradian is the solid angle at the top of a cone, in this case the cone being of emitted radiation.) Because the intensity increases, it can be used to calculate the temperature.

Figure 28-1 illustrates this concept. Note that both of the axes have logarithmic scales.

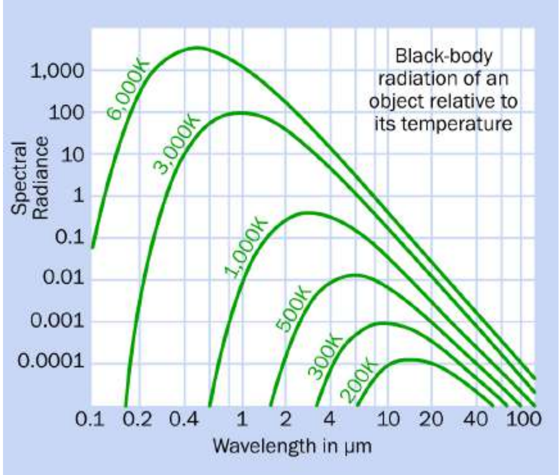

Figure 28-1 *The increase in intensity, and the widening spread of wavelengths, of black-body radiation emitted by one object at six different temperatures (in degrees Kelvin).*

All the curves are for one object. Each curve is specific to one temperature in degrees Kelvin, showing how the intensity of emitted radiation varies with wavelength. Note that radiation at wavelengths shorter than 0.7µm is within the visible spectrum; thus, objects at 1,000 degrees Kelvin, or hotter, may be seen to glow visibly.

Because of the wide variety of intensities and temperatures, an infrared sensor that is ideal for measuring a temperature of 1,000 degrees Kelvin will not provide an accurate result at 200 degrees Kelvin. The peak spectral radiance at 1,000 degrees is more than 10,000 times the peak at 200 degrees. Also, the curves in the figure are for an "ideal" object emitting pure black-body radiation. In reality, glass, plastic, and many other materials have a much lower *emissivity*, meaning that they emit less radiation, and are classified as *gray bodies*. A metallic object with a polished surface may emit one-tenth of true black-body radiation.

These issues cannot be ignored, but may be dealt with by relatively simple strategies. Infrared sensors will be rated according to their suitability for different temperature ranges, and the emissivity of the object being measured can be determined by consulting standard tables. Alternatively, the object can be spray-painted with special black paint (such as "Senotherm" or "3-M Black") that has a known emissivity of about 0.95 of pure black-body radiation. Alternatively, a specially formulated black sticker may be applied to the object that is being measured, so long as its temperature is within reasonable limits.

However, a basic infrared temperature sensor will not function reliably if it is pointed randomly at a variety of objects that vary widely in temperature. Specialized, expensive industrial devices incorporate compensation to deal with these issues, but they are outside the scope of this Encyclopedia.

Thermopile

A typical low-cost, chip-based infrared temperature sensor contains a *thermopile* consisting of multiple thermocouples etched into silicon and connected in series. The concept of a thermopile is illustrated in Figure 25-9, where a brief explanation is included.

The configuration of the thermopile is arranged so that the hot junctions of its thermocouples are all clustered in a small central area, where they receive incoming radiation through a window (often made of silicon) that is transparent to infrared wavelengths. The cold junctions are dispersed around the periphery, where they are shielded from incoming radiation. One way to visualize this is shown in Figure 28-2, although this is not a literal depiction of an actual sensor.

Instead of using alternating types of wire, as in a **thermocouple**, a chip-based thermopile often uses alternating segments of n-type and p-type silicon. The hot junctions are mounted on a thin film that has very little heat capacity, while the cold junctions are mounted on a thicker substrate that acts as a heat sink.

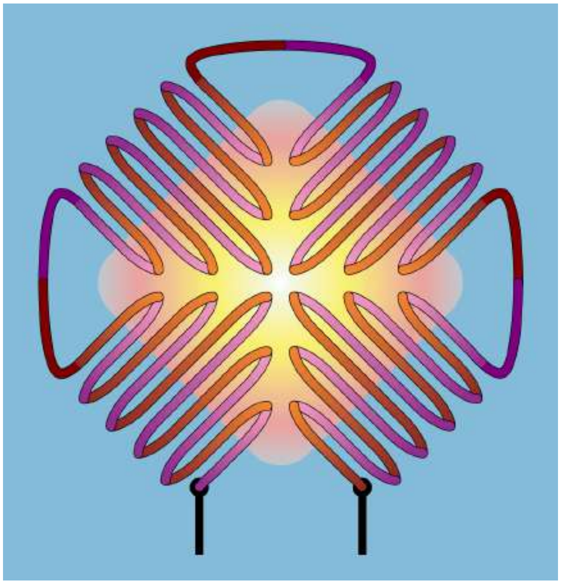

Figure 28-2 *Simplified diagram of the thermopile configuration inside an infrared sensor chip. Radiation arriving through a window in the chip affects the thermocouple junctions in the central region, while junctions around the edges remain at a lower temperature.*

Temperature Measurement

The voltage generated by the thermopile is related to the difference in temperature between the hot and cold thermocouple junctions. Thus there are three interrelated variables: hot temperature, cold temperature, and voltage. To calculate one variable, we must know the other two.

The hot temperature is what we wish to know. Therefore we must establish the voltage (which can easily be measured) and the cold temperature. The cold temperature can be determined by adding a thermistor inside the chip.

Typically an infrared temperature sensor with an analog output will have two pins that provide access to the internal thermistor, so that its temperature can be calculated from its resistance. Another two pins provide the voltage between the ends of the thermopile.

Interpreting and reconciling these values is not a trivial matter, especially bearing in mind that the thermistor has a negative temperature coefficient and a nonlinear output, and the thermopile will also have some nonlinearity. To simplify this situation, some infrared temperature sensors incorporate electronics to perform the necessary calculations and provide a digital output. This output can be converted to degrees of temperature by some fairly simple mathematical operations in an external microprocessor.

Variants

Two types of sensors are popular. One is surface-mounted, such as the TMP006, shown in Figure 28-3. This type generally has a digital output. The other type is a discrete component with four leads, such as the Amphenol ZTP135, shown in Figure 28-4. Discrete components may have either an analog or a digital output.

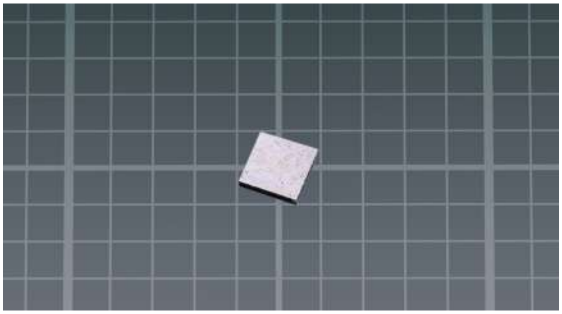

Figure 28-3 *A surface-mount infrared temperature sensor with digital output. Eight tiny solder pads are located on the underside. The background grid is in millimeters.*

Both types of sensors allow infrared light to enter through an area that is opaque to the visible spectrum but transparent to the appropriate range of wavelengths.

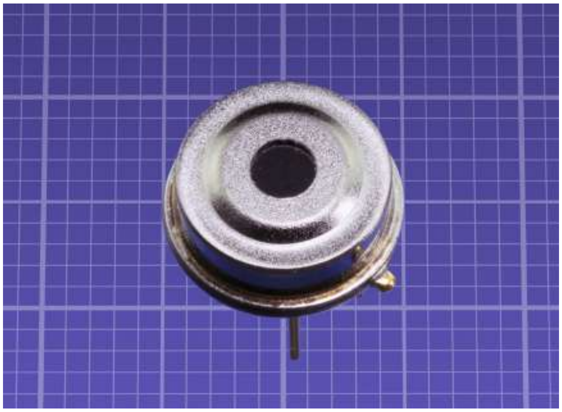

Figure 28-4 *A through-hole infrared temperature sensor with analog output. The background grid is in millimeters.*

The ZTP135 has an analog output shown in Figure 28-5.

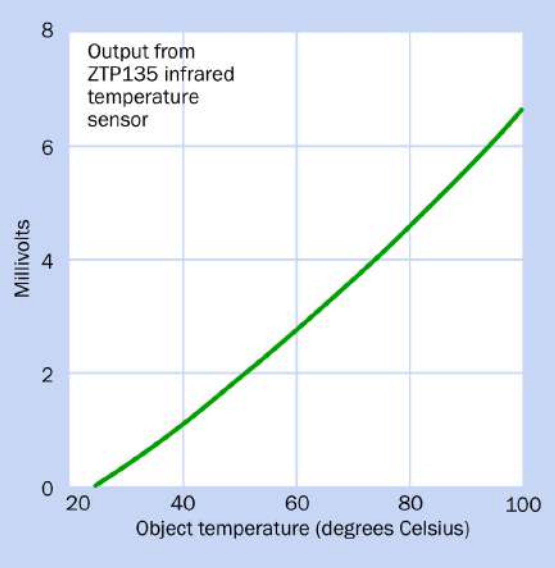

Figure 28-5 *Analog output from an infrared temperature sensor.*

The TMP006 is only about 1.5mm square, but is available on a breakout board from Sparkfun. Its successor, the TMP007, is available on a breakout board from Adafruit, shown in Figure 28-6.

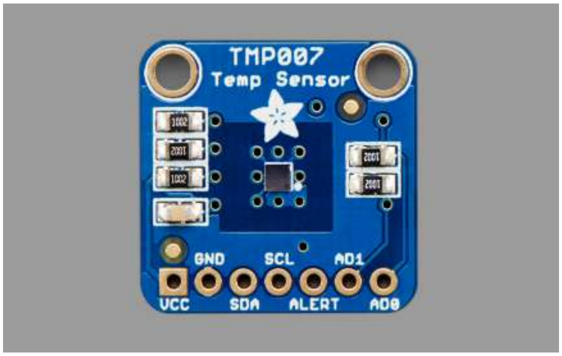

Figure 28-6 *The TMP007 sensor mounted on a breakout board from Adafruit.*

Surface-Mount Specifications

The TMP006 and TMP007 require a supply voltage that can range from 3.3V to 5V. These chips support the SMBus and I2C bus protocols, using a bus address that can be user-selected. An internal analog-to-digital converter uses 1 least-significant bit to represent 1/32nd of a degree Celsius, and data is saved as a 14-bit signed integer. Up to 16 temperature samples can be averaged internally.

Measurable temperature range is -40 to +125 degrees Celsius. Some hysteresis is built in. The TMP007 supports an alert mode if temperature falls above or below a user-specified threshold.

Sensor Arrays

Using multiple thermopile sensors arrayed in a line or a grid, with an array of lenses, it is possible to capture an image of temperature variations over a surface or a scene. This is known as *thermal imaging*. It can detect heat leakage from buildings, indicating poor insulation, or can locate hot spots in electronic circuits. Heimann Sensor has pioneered the miniaturization of a 31 x 31 grid of thermopile sensors in a single TO-8 or TO-39 package.

Values

Temperature Range

Chip-based infrared temperature sensors are typically designed for a range from about -20 degrees Celsius to about +125 degrees Celsius. Their peak sensitivity is to wavelengths between 4µ and 16µm.

Other types of infrared temperature sensors can have a much wider temperature range, but are more costly.

Field of View

Often referred to by its acronym, *FOV*, the field of view is the angle at the apex of an imaginary cone extending outward from the sensor, defining a boundary where the sensitivity diminishes below 50% of the value directly in front of the sensor. Greek letter ϕ may be used to represent the angle between the surface of the cone and the center line, while θ represents the angle between the opposite surfaces of the cone (i.e., 2 * ϕ). This is shown in Figure 28-7. θ is usually the angle defined as field of view.

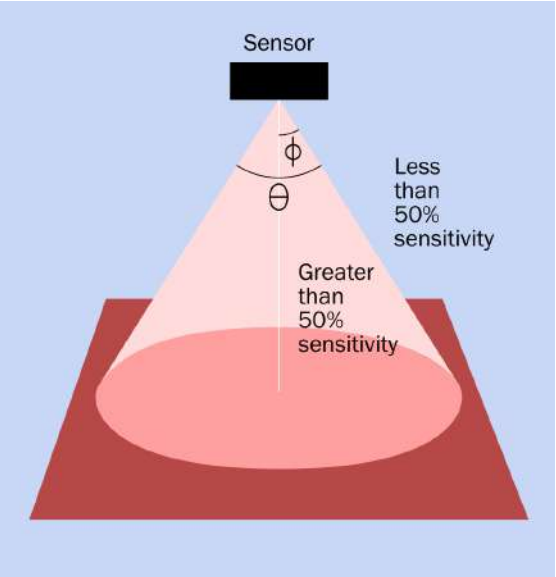

Figure 28-7 *Measurement of the field of view from a sensor, defined as the boundary of an imaginary cone where sensitivity drops below 50%.*

Because most infrared temperature sensing components do not have a lens, they are sensitive over a wide angle. The field of view is typically 90 degrees.

What Can Go Wrong

Inappropriate Field of View

The object being evaluated must fill the field of view of the sensor, to avoid measuring other objects around it.

Reflective Objects

A reflective object has lower infrared emissivity, and also may provide misleading output if the sensor is actually measuring thermal radiation reflected in the surface of the object in front of it. In a permanent installation, such as inside a device, the surface to be measured may have to be painted to reduce its reflectivity.

Glass Obstruction

Because glass is opaque to the spectrum of infrared that is of interest, temperature cannot be measured through a glass window. Silicon is opaque to visible wavelengths but is transparent to wavelengths longer than 2µm.

Multiple Heat Sources

Heat is transferred by convection, conduction, and radiation. While an infrared temperature sensor is designed to be sensitive to radiation, it will also respond to other heat sources. Warm or cold air currents will affect its response, and so will heat conducted through the material on which it is mounted. Careful placement of the sensor is important. A shield around the sensor, with a small hole in the center, can prevent convection, while correct location on a circuit board can minimize conduction.

Thermal Gradients

An infrared temperature sensor should be mounted in a stable environment where it will not be exposed to thermal gradients (one side being hotter than another). This asymmetry can cause inaccurate readings.

microphone

29

OTHER RELATED COMPONENTS

- **speaker** (see Volume 2)
- **headphone** (see Volume 2)

What It Does

The sensation of sound is created by rapid waves of air pressure impinging upon the eardrum. A microphone can convert these pressure waves into an alternating electrical signal that can be amplified, recorded, broadcast, transmitted through wires, and reproduced as sound by a headphone or speaker. The principle is illustrated in Figure 29-1. (For more information about sound reproduction, see the entries on **headphone** and **speaker** in Vol. 2.)

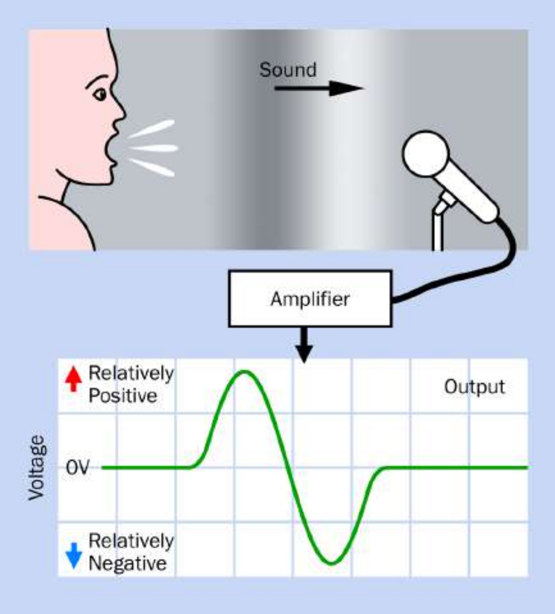

Figure 29-1 *The principle of converting pressure waves into an alternating electrical signal (adapted from an illustration in Make: More Electronics).*

Schematic Symbol

Various schematic symbols for a microphone have been used during the decades since its invention. A selection is shown in Figure 29-2. Each symbol assumes that sound is traveling from left to right. This is important when interpreting the symbol at top right, which can represent an earphone when it points in the opposite direction. Unfortunately, some schematics do not conform with this rule.

The two symbols at the bottom, showing a capacitor inside the microphone, should be reserved for condenser or electret microphones.

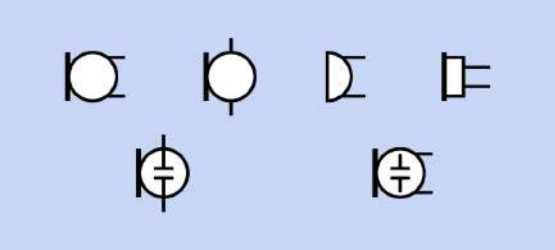

Figure 29-2 *A selection of schematic symbols that represent a microphone.*

How It Works

Some types of microphones generate a small voltage, while others have a fluctuating resistance that modulates a DC current.

Carbon Microphone

This was a very early attempt to reproduce sound. It contained carbon granules whose packing density increased and decreased in response to air pressure waves. When the density increased, the resistance between the granules diminished, and vice-versa. The principle is illustrated in Figure 29-3, and was patented by Thomas Edison in 1877 for use in telephones. As late as the 1950s (and even later in some countries), wired telephone handsets contained carbon microphones. Their bandwidth was extremely limited.

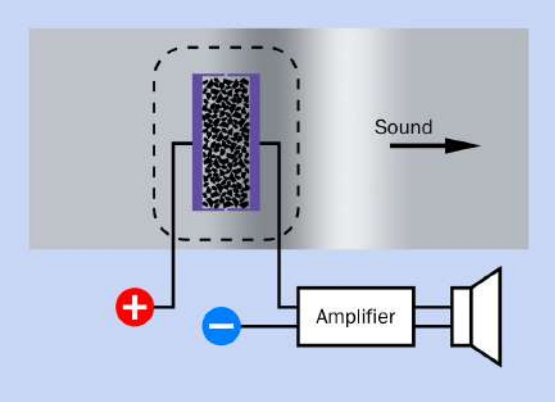

Figure 29-3 *The principle of a carbon microphone.*

Moving-Coil Microphone

Also known as a *dynamic* microphone, this consists of a very small, light coil of thin wire on a cylindrical tube that can vibrate along the axis of a permanent magnet. This principle is illustrated in Figure 29-4. A diaphragm is attached to the front of the tube, and responds to air pressure waves that penetrate the perforated enclosure of the microphone. Movements of the coil around the magnet create small alternating currents in the wire. The inertia of the coil, tube, and diaphragm, and the force needed to overcome the interaction between the coil and the magnet, impose a limit on the high-frequency response of this design.

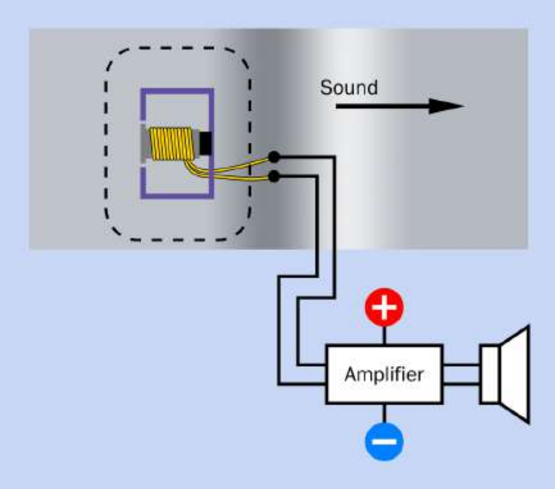

Figure 29-4 *The principle of a moving-coil microphone.*

Condenser Microphone

This type of microphone contains two thin discs or plates that form a capacitor. (In the early days of electricity, a capacitor used to be known as a condenser. The terminology has persisted for microphones.) An equal and opposite charge is applied to the plates. One plate is flexible, and as it responds to pressure waves, the capacitance between it and the other, rigid plate fluctuates. If the charge on the plates is kept approximately constant while the capacitance fluctuates, the voltage across the capacitor fluctuates also. These fluctuations can be amplified, as suggested in Figure 29-5.

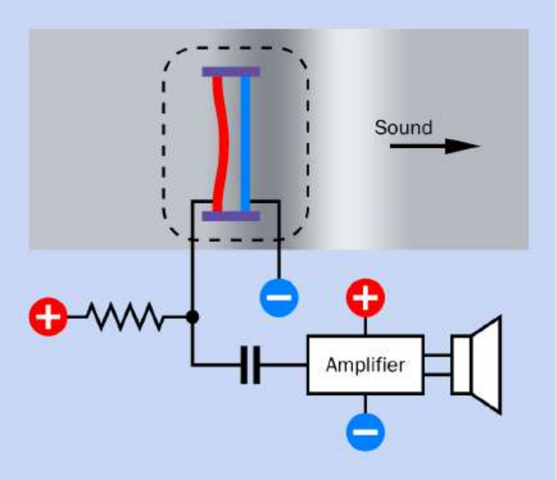

Figure 29-5 *The principle of a condenser microphone.*

Electret Microphone

This works on the same principle as a condenser microphone, except that its plates are made from a ferroelectric material that retains an electrical charge, just as iron will retain a magnetic polarization. The name of the microphone is derived from "electrostatic" and "magnet." While early electret microphones were of poor quality, they have evolved to rival condenser microphones, and are extremely affordable. Because the electret creates very small currents, it usually includes a transistor or op-amp in its package to boost the signal, and has an open-collector output. The basic circuit for an electret is shown in Figure 29-6. For more information about using an open-collector output, see Figure A-4.

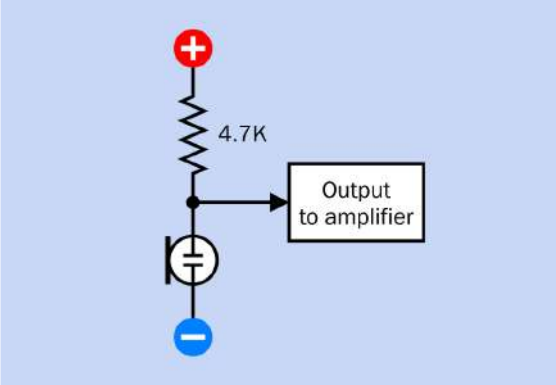

Figure 29-6 *The basic circuit for using an electret microphone.*

An example of a low-cost electret microphone is shown in Figure 29-7. This type of component is sold either with leads attached, or with solder pads.

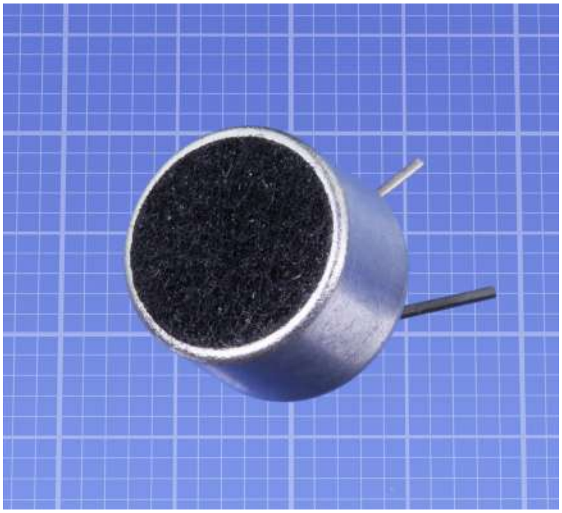

Figure 29-7 *A generic electret microphone. The background grid is in millimeters.*

MEMS Microphone

This type, often used in mobile phones, is a capacitive device that works on the same principle as a condenser microphone, although the component is etched in silicon and has a diaphragm that measures only about 1mm square. Many MEMS microphones have an analog output that is amplified in the same chip. Others have a digital output, using *PDM encoding*. This reduces the analog signal to a very fast bit stream, in which the density of the bits represents the amplitude of each fluctuation in a sound wave. PDM is an acronym for *pulse density modulation*. It requires an external clock signal to time the bit stream.

A breakout board from Sparkfun, on which is mounted an Analog Devices ADMP401 MEMS microphone with a preamplifier, is shown in Figure 29-8.

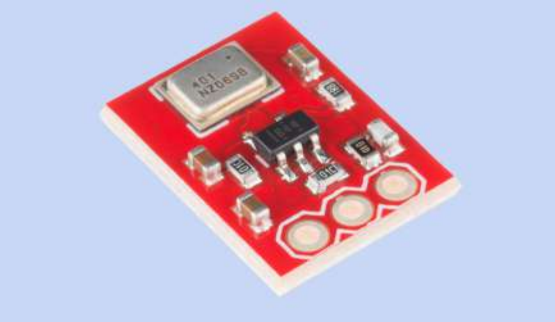

Figure 29-8 *A breakout board for a MEMS microphone (the metal-clad rectangular component at the far end).*

Piezoelectric Microphone

This has also been known as a *crystal* microphone. It contains a diaphragm that functions as a transducer. When it flexes in response to pressure waves, the mechanical energy is transformed into a small amount of electrical energy. Piezoelectric microphones were replaced by the moving-coil type in domestic audio devices when vacuum tubes were replaced with transistors, but may still be used as contact microphones to amplify acoustic musical instruments, or to trigger the playback of digitally sampled musical sounds.

Other variants include *ribbon* microphones (which were common in recording studios in the 1950s and 1960s, but have become rare), *laser* microphones, and *fiber-optic* microphones. They are not sufficiently common to be included in this entry.

Values

Sensitivity

Sound pressure is a complicated topic, explored in detail in the **transducer** entry in Volume 2. It can be measured in pascals, where 1 pascal = 1 newton per square meter.

The *sound pressure level* is a different concept. It measures the *relative* intensity of a sound, in a logarithmic scale calibrated in *decibels* (abbreviated dB). The reference value for this relative scale is 20 micropascals, considered to be the threshold of human hearing, comparable to a mosquito three meters away. This is assigned the value of 0dB.

From this point upward, the actual sound pressure doubles for each additional 6dB. A table of noise sources and their approximate decibel values is shown in Figure 29-9. This is derived from averaging eight similar tables, which are not always consistent in their estimates. It is an approximate guide only.

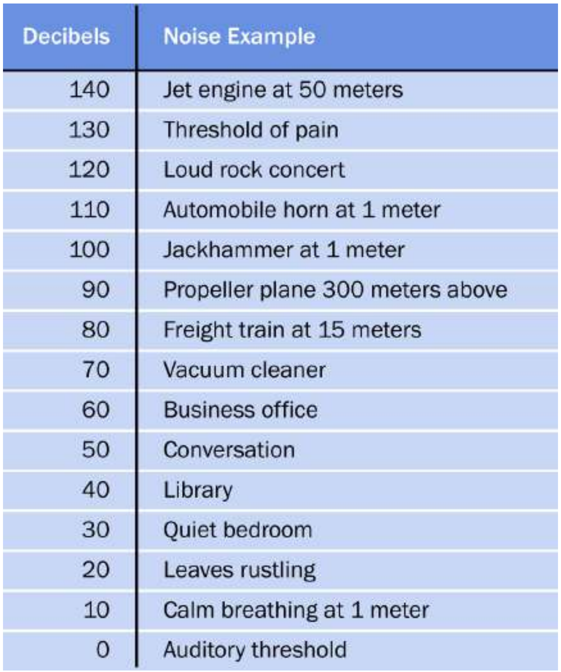

Decibels	Noise Example
140	Jet engine at 50 meters
130	Threshold of pain
120	Loud rock concert
110	Automobile horn at 1 meter
100	Jackhammer at 1 meter
90	Propeller plane 300 meters above
80	Freight train at 15 meters
70	Vacuum cleaner
60	Business office
50	Conversation
40	Library
30	Quiet bedroom
20	Leaves rustling
10	Calm breathing at 1 meter
0	Auditory threshold

Figure 29-9 *Decibel values for some common sound sources. Reproduced from Volume 2 of this Encyclopedia.*

The decibel unit is important when understanding the specifications of microphones, because it is used to measure their response. Microphone sensitivity is established with a standard input sine wave of 1kHz in frequency and 94dB in intensity (equivalent to 1 pascal in actual sound pressure), measured at the microphone. The sensitivity of an analog microphone is then defined as the number of decibels in an output signal of 1V. Because the output is an AC signal, voltage is measured as a root-mean-square (RMS) value.

For digital microphones, sensitivity is measured as the decibels that can be reproduced by a full-scale digital output. This value is abbreviated as dBFS.

Directionality

A microphone that has a directional response is desirable in many situations. Often, for example, sounds in front of the microphone are more important than sounds from behind the microphone. The directionality of a microphone (sometimes referred to as its *directivity*) is usually represented with a *polar graph* in which the microphone is seen from above, and its sensitivity to sounds from various directions is shown with a curve such as those in Figure 29-10. The circles are drawn at intervals of 5dB, with 0dB at the periphery and -30dB at the center. The precise response for an individual microphone should be shown in its documentation.

Frequency Response

Every microphone tends to be more sensitive to some sound frequencies than others. A manufacturer will provide a graph showing this sensitivity, in decibels, plotted against sound frequency on a logarithmic horizontal axis. Theoretically, human hearing extends from around 20Hz to 20kHz, but few people are actually capable of hearing the high end of that range, and 15kHz may be a more realistic limit for a young person, diminishing to 10kHz with middle age.

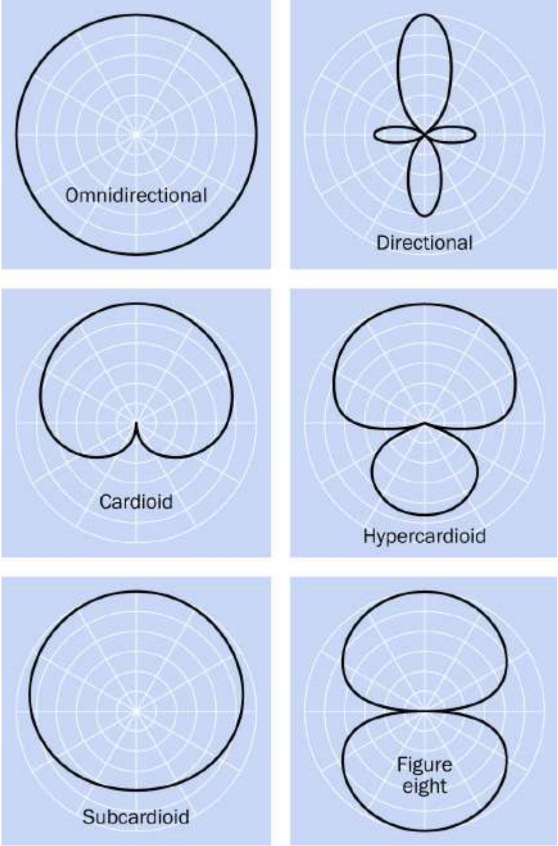

Figure 29-10 *Six generic sensitivity patterns. Individual microphones will deviate somewhat from these generic curves.*

An ideal flat response would show that a microphone is equally sensitive to all frequencies. In reality, *rolloff* usually occurs at low frequencies, and will eventually occur at high frequencies, although it may be preceded by a rise in response. If the central section of a curve is flat within plus-or-minus 1dB, this is a level of performance that was attained only by expensive studio microphones in the past. Electrets and MEMS microphones can now provide equivalent frequency response for $1 or $2 apiece, as opposed to the hundreds or even thousands of dollars that used to be necessary for professional equipment.

The response curve shown in Figure 29-11 is for the eMerging i436, an electret microphone sold in a module as an accessory to enable high-quality recordings on handheld devices. The rise around 15kHz may have been introduced deliberately by the manufacturer to compensate for reduced sensitivity of the human ear in that range.

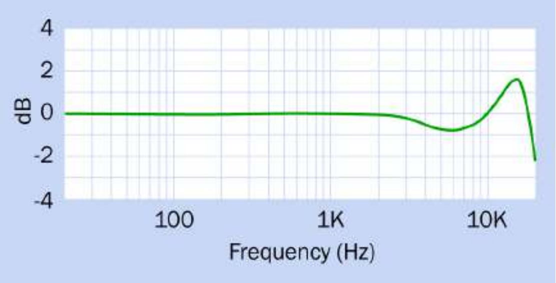

Figure 29-11 *Frequency response for an electret microphone.*

Impedance

The impedance value for a microphone is a function of its resistance, capacitance, and inductance. An amplifier input will also have an impedance rating, and for ideal power transfer between microphone and amplifier, the impedance values should be identical. However, a more important consideration in audio equipment is to avoid voltage loss between the output device (in this case, the microphone) and the input device (the amplifier). To achieve this, the output device should have a low impedance, while the input device should have a high impedance. Most microphones are rated at 150 to 200 ohms, while an amplifier may be rated at 1.5K to 3K.

Total Harmonic Distortion

When an audible sine wave is converted to an electrical output by a microphone (and by its preamplifier, if one is included in the module), the output may become corrupted by some multiples of the basic frequency. These are known as *harmonics*, and they are considered as a distortion of the signal. Total harmonic distortion, measured by a spectrum analyzer over the entire frequency range, should ideally be less than 0.01%.

Signal-to-Noise Ratio

Often abbreviated as S/N or SNR, the signal-to-noise ratio in a microphone is measured in decibels, and should be 60dB or higher.

What Can Go Wrong

Cable Sensitivity

Audio amplification is always vulnerable to electrical noise, which tends to be amplified along with the signal. Small signals from microphones require the use of shielded cables to reduce hum and other types of interference.

Noisy Power Supply

For similar reasons, power supplies must be as free as possible from voltage spikes and other fluctuations.

current sensor

30

The entry describes components that can be installed to monitor current on an indefinite basis. It does not include test equipment, such as test meters or multimeters.

A *current transformer* may be used to measure current, but is not included in this Encyclopedia.

OTHER RELATED COMPONENTS

- **voltage** sensor (see Chapter 31)

What It Does

A current sensor measures the flow of electricity through a wire or a device, and supplies an output that can be interpreted either visually or by a microcontroller to provide a reading in amperes or fractions of an ampere.

Applications

Current sensing is important in industrial applications such as the control of high-powered motors. It can be used to monitor the performance of an inverter, or for everyday purposes such as monitoring the long-term power consumption of an appliance. During product development, a current sensor can indicate the power consumption of a circuit as it changes with modifications.

This entry describes three methods to measure current: an ammeter, series resistor, and Hall sensor. While other methods exist, they are outside the scope of this Encyclopedia.

Ammeter

An ammeter that is sold as a standalone device with leads for circuit testing is often described as a *test meter*. Its functionality is usually built into a *multimeter*. Test meters and multimeters are outside the scope of this Encyclopedia.

An ammeter designed for permanent installation in a device or prototype is a type of *panel meter*, such as the one shown in Figure 30-1. This traditional-style analog meter may be less expensive than the many digital types that are available. It uses a magnetic field created by current flowing through a coil to pull a needle across a scale, against the force of a spring.

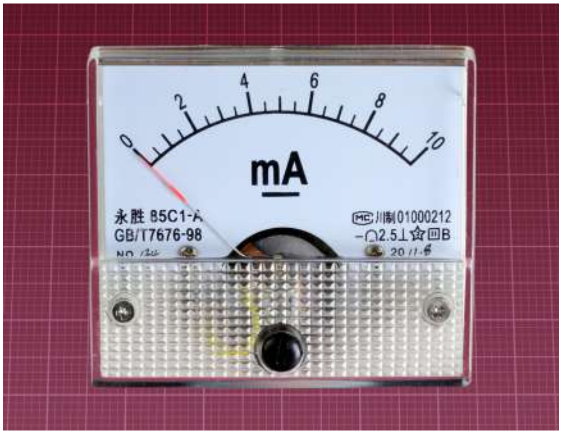

Figure 30-1 *A traditional-style analog ammeter.*

A digital ammeter allows a wider range of values to be viewed more easily. The meter from Adafruit in Figure 30-2 has a range of 0A to 9.99A, at voltages from 4.5VDC to 30VDC. The meter can be powered parasitically from the currents that it is measuring, or can use a separate isolated 5VDC supply.

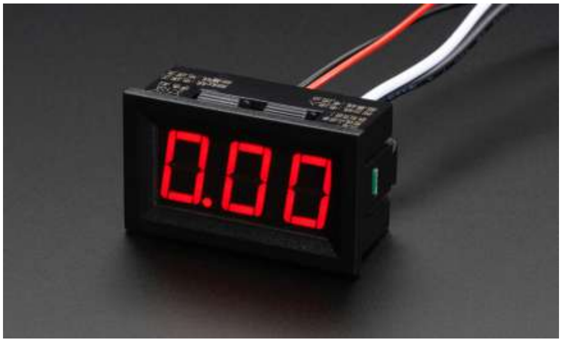

Figure 30-2 *A panel-mount digital ammeter that can measure up to 9.99A.*

Schematic Symbol

An ammeter may be represented in a schematic with the letter A inside a circle, as shown in Figure 30-3.

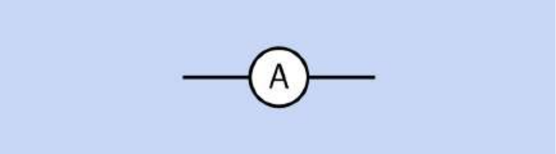

Figure 30-3 *An ammeter may be represented like this in a schematic.*

Ammeter Wiring

Two ways to use an ammeter in a circuit are illustrated in Figure 30-4, where the load may be any equipment, device, or component that provides some electrical resistance. Because current is the same at all points in a simple circuit, the current that the meter measures flowing through itself will be the same as the current flowing through the load, and the sequence of components is immaterial.

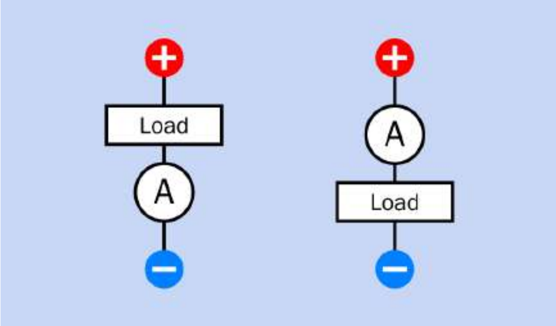

Figure 30-4 *Two options for placement of an ammeter in a circuit.*

However, regardless of the placement of the meter, the process of measuring current will inevitably change the value of the current being measured. This is because the ammeter imposes some internal resistance of its own. The resistance is extremely low, and may be considered negligible for loads of more than a few ohms.

- The low internal resistance of an ammeter means that it must never be connected in parallel with a load, or directly across a power source.

A disadvantage shared by analog and digital meters is that they are not usually interchangeable between AC and DC.

Series Resistor

The current flowing through a load can be calculated by measuring the voltage across a series resistor that is inserted between the load and its ground connection. The concept is illustrated in Figure 30-5.

Using Ohm's law, if U is the voltage drop, I is the current, and R is the value of the resistor:

```
I = U / R
```

This tells us that for a fixed value of R, current is proportional to voltage. Therefore, measuring the voltage enables calculation of the current, so long as the value of R is known.

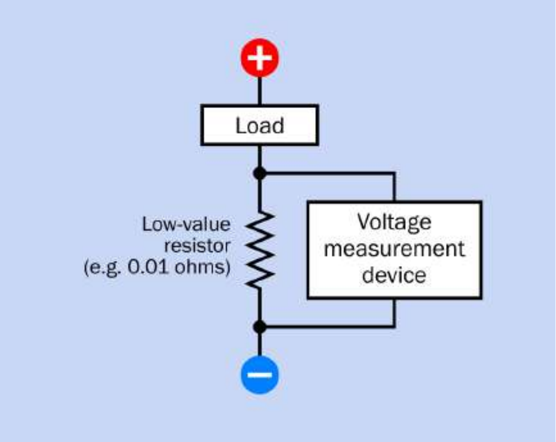
Load
Low-value resistor (e.g. 0.01 ohms)
Voltage measurement device

Figure 30-5 *A basic circuit for measuring current as a function of the voltage drop across a small-value resistor. The load in this figure is any circuit or device with a relatively higher resistance. The voltage measurement device could be a microcontroller or analog-to-digital converter.*

Suppose that R has a very small value, which is trivial compared with the resistance of the load. Consequently, the current in Figure 30-5 will be mainly determined by the load, and we may consider the value of the current to be almost the same with or without the addition of R. In that case, the voltage drop across the resistor will be smaller if R is smaller. A smaller voltage drop will not be as easy to measure, but a lower resistance will result in less wastage of power.

If P is the power:

$$P = R * I^2$$

An example may help to make this clear. Suppose the resistor has a value of 0.5 ohms, and the voltage drop across it is measured to be 1V. Ohm’s law shows that current flow is 1 / 0.5 = 2A. The power formula shows that P = 0.5 * 4 = 2W.

To waste less power, the value of the resistor should be reduced further. Suppose a resistor of 0.01 ohms is used, and the voltage drop across it is measured as 0.02 volts. The current is 0.02 / 0.01 = 2A, as in the preceding example, but the power dissipation is now only 0.01 * 4 = 0.04W, which is negligible.

But are resistors available that have values measured in fractions of an ohm?

Current-Sense Resistors

In fact, many *current sense* resistors are available, with values of 0.1 ohms, 0.001 ohms, 0.0001 ohms, and many in between. Some resistors have values measured in micro-ohms. Examples are shown in Figures 30-6, 30-7, and 30-8.

Measuring a small voltage drop is easily done by using a microcontroller. However, the connection to the microcontroller must be made as near to the resistor as possible, to eliminate the additional resistance of wires or circuit-board traces. For this reason, precision current-sense resistors may be equipped with four terminals. Two are wider, and are meant for connection to the flow of current. The other two are narrow, for measuring the voltage over the resistor. With this *4-point* configuration, voltage drop over the resistor can be measured as close to it as possible. The 0.001-ohm surface-mount resistor in Figure 30-8 is designed for 4-point measurement.

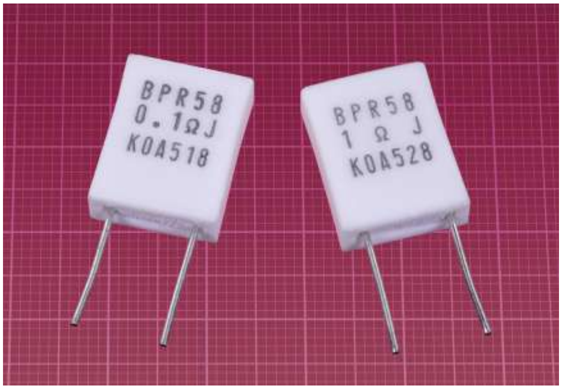

Figure 30-6 *Two current resistors manufactured by KOA Speer. Left: 0.1 ohms, 5W, 5%. Right: 1 ohm, 5W, 5%. The background grid is in millimeters.*

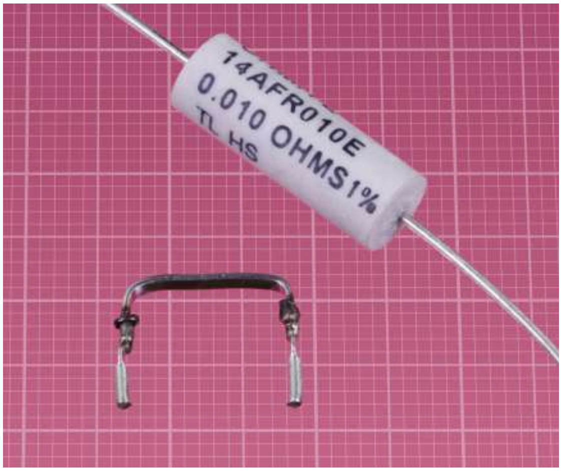

Figure 30-7 *Two styles of current resistors rated for 0.01 ohms. Bottom left: A plug-in version from TT Electronics, rated 1W and 5%. Top right: Ohmite, 4W, 1%. The background grid is in millimeters.*

Resistors that have the lowest values and are intended to tolerate high current may consist of just a metallic strip welded to solderable pins. This type of component is sometimes called an *open-air resistor*. It is commonly used in multimeters, for measuring currents up to 10A or above.

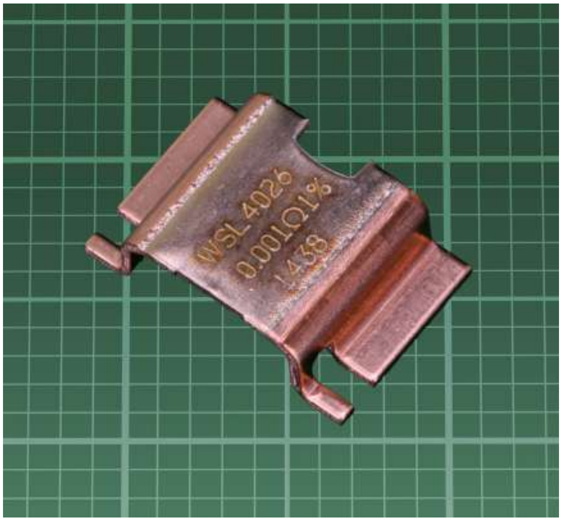

Figure 30-8 *This Vishay 4-terminal surface-mount current resistor is rated for 0.001 ohms, 3W, 1%. The background grid is in millimeters.*

Voltage Measurement

Some chips are designed for amplifying the voltage drop across a current-sensing resistor. An example is the Texas Instruments INA169.

A few chips contain an analog-to-digital converter in addition to the amplifier. The INA219 by Texas Instruments is designed to measure voltage as well as current, on the "high side" of a circuit—that is, between the positive power supply and the power input of the circuit. It makes its digital data available over an I2C bus.

For additional details about protocols such as I2C, see Appendix A.

Measuring current from the voltage drop across a series resistor offers the advantages of simplicity, ability to work with AC or DC, and low cost (although some resistors of extremely low value can be relatively expensive). A possible disadvantage is that the measurement circuit is not isolated from the circuit whose current is being measured.

Hall-Effect Current Sensing

The principle of a Hall-effect sensor is explained in the entry for **object presence** sensors. See "Hall-Effect Sensor". Normally this type of sensor is activated by an external permanent magnet, but it can also react to the magnetic field generated by current flowing through a wire.

Because the field generated around a wire is proportional with the current, the analog output voltage generated by a linear Hall-effect sensor can also be proportional with the current.

Hall-effect sensors for this specific purpose are available in 8-pad surface-mount packaging. The current to be measured passes through a copper conductor that is embedded in the chip. An example is Allegro's ACS712, for AC or DC currents up to 30A. The internal resistance of the current path through this chip is stated as 1.2 milliohms, and the path is isolated from the sensing circuitry.

Three variants of the chip are available, for currents of plus-or-minus 5A, 20A, and 30A. Depending on which variant is used, the output will range from 66mV to 185mV for each increase of 1A in the current path. Because the current path is isolated, the chip requires a separate power supply of 5VDC.

The 5A version of the AVS712 can be bought on breakout boards from Sparkfun. In Figure 30-9, the board at top-left contains only the ACS712, while the board at lower-right adds an op-amp with sensitivity control, to amplify the voltage output when measuring small currents.

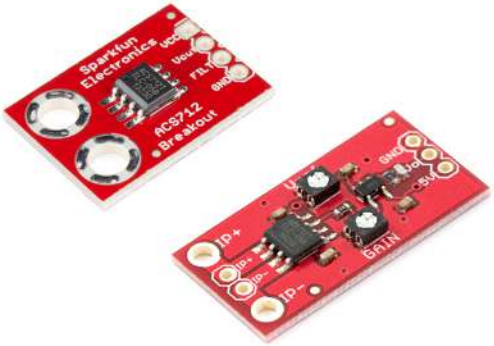

Figure 30-9 *Breakout boards using the AVS712 Hall-effect current sensor. The one at lower-right includes an op-amp to amplify small signals.*

What Can Go Wrong

Confusing AC with DC

A panel meter that is designed only to measure DC should not be used with AC, and vice-versa. Erroneous readings or damage to the meter may result.

Magnetic Interference

A disadvantage of Hall-effect current sensing is that the sensor can be affected by stray magnetic fields. Because a Hall-effect chip is responding to very small magnetic effects, it is vulnerable to interference. Consult the manufacturer's datasheet carefully regarding correct placement of a chip on a circuit board.

Incorrect Meter Wiring

The correct wiring of an ammeter is in series with a load, not in parallel with a load. This may seem an elementary error, but is easy to make if an ammeter and a volt meter are both being used, and they look similar.

Because some panel meters are not fused, applying an ammeter directly across a power supply, without any series resistance, may result in immediate and hazardous destruction of the meter.

Incorrect wiring can also occur when using a digital meter that has four wires emerging from it: two for current testing and two for a separate power supply. This issue is especially important when measuring substantial currents (above 1A).

Current Out of Range

Attempting to measure a current that exceeds the range of an ammeter may damage the meter or blow its internal fuse, if it has one.

voltage sensor

31

The entry describes the type of component that can be installed to monitor voltage on an indefinite basis. It does not include test equipment, such as test meters or multimeters.

OTHER RELATED COMPONENTS

- **current** sensor (see Chapter 30)

What It Does

A voltage sensor measures the electrical potential between any two points in a circuit, or the voltage supplied by a power source, and provides data in volts or fractions of a volt. It should not be confused with an *analog-to-digital converter*, which is not a sensor in itself, but can process the voltage output from a sensor by digitizing it. More information about analog-to-digital conversion is in the Appendix. See Appendix A.

Applications

Voltage measurement is important in conjunction with all types of power supplies, to verify their performance. A volt meter may also be used to show the output from various types of analog sensors that have voltage output.

A graphical display may be used in audio equipment to indicate signal level, which can be proportional with voltage.

Volt Meter

A volt meter that is sold as a standalone device with leads for circuit testing is often described as a *test meter*. Its functionality is usually built into a *multimeter*. Test meters and multimeters are outside the scope of this Encyclopedia.

A volt meter designed for permanent installation in a device or prototype is a type of *panel meter*, which is described here.

An antique analog panel meter is shown in Figure 31-1.

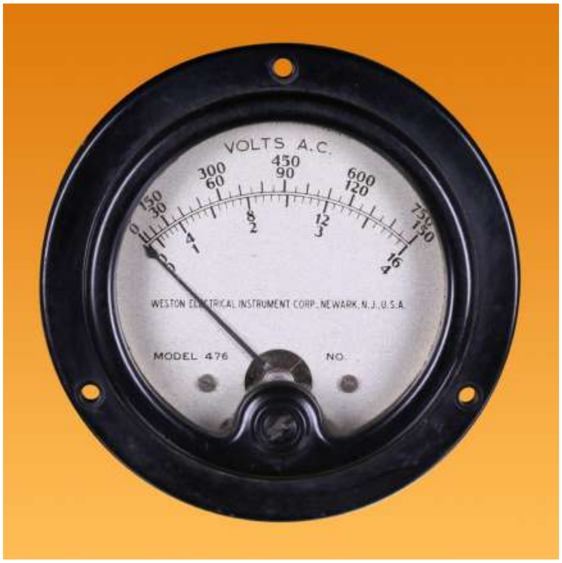

Figure 31-1 *An antique analog volt meter.*

The four scales on the dial correspond with separate input terminals at the rear of the unit. The unequally divided scales are a simple way to compensate for the nonlinear response of the mechanical movement inside the meter.

Modern analog volt meters are still manufactured, and may be cheaper than their digital equivalents. However, a digital volt meter allows a wider range of values to be viewed more easily. The meter in Figure 31-2, sold as a low-cost battery tester, measures voltages from 4VDC to 13VDC to an accuracy of two decimal places. It needs no separate power supply.

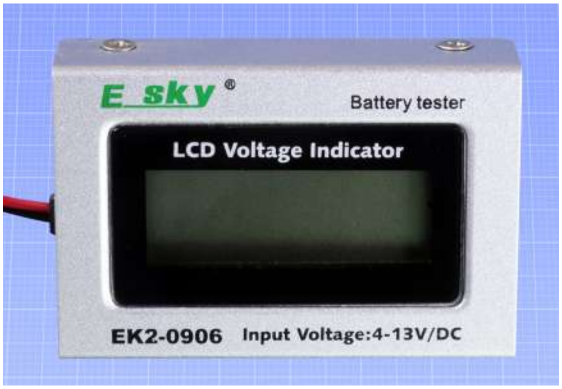

Figure 31-2 *A panel-mount digital volt meter.*

Sometimes the number of digits in a volt meter is specified as ending with one-half, as in 3.5 or 3-1/2 digits. This means that the most significant (leftmost) digit can only be a 1 or a blank. The extra “half digit” may seem inconsequential but doubles the range of displayable values. For example, a 2-digit display can only show 100 values, from 0 to 99. A 2-1/2-digit display can show 200 values, from 0 to 199.

Schematic Symbol

A volt meter may be represented in a schematic with the letter V inside a circle, as shown in Figure 31-3.

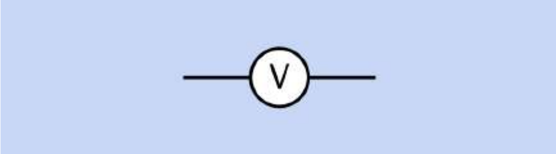

Figure 31-3 *A volt meter may be represented like this in a schematic.*

Volt Meter Wiring

Two ways to use a volt meter in a circuit are illustrated in Figure 31-4, where the load may be any equipment, device, or component that provides some electrical resistance.

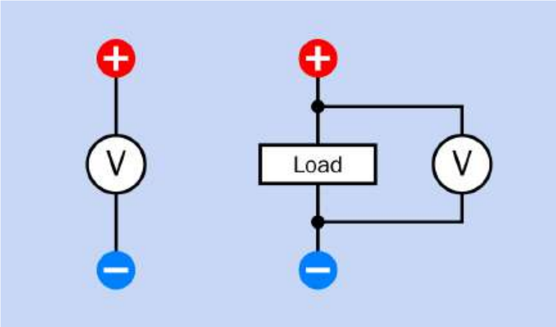

Figure 31-4 *Two options for placement of a volt meter.*

On the left side of the figure, a volt meter may be connected directly across a power source, because the internal resistance of the meter is so high, it will draw very little current. In this configuration, the meter measures the voltage of a source when the source is virtually unloaded.

On the right, the meter can measure the voltage drop across a load, or can measure the voltage across any component, or group of components, that are a subset of the load.

A disadvantage shared by dedicated analog and digital volt meters is that they are not usually interchangeable between AC and DC.

How It Works

An analog volt meter usually contains a built-in, high-value fixed resistance, and an ammeter measuring the current passing through it. The current sensed by the ammeter is then converted to a reading in volts.

Suppose that R represents the fixed resistance, U is the voltage drop across the resistance, and I is the current flowing through it. Ohm's Law tells us:

```
U = I * R
```

The formula shows that if the resistance is fixed, the voltage will vary in proportion with the current, and therefore can be calculated from it.

Load-Related Inaccuracy

When the meter is measuring the voltage drop created by a load that has high resistance comparable to the internal resistance of the meter itself, the meter will not give an accurate reading. Figure 31-5 illustrates this problem.

On the left side of the figure, two resistors 10M each are wired in series between a 9VDC power source and negative ground. Because the resistors are equal, each of them imposes an equal voltage drop of 4.5V.

On the right side of the figure, an additional 10M resistance has been added in parallel with the lower resistor. When two resistors of values R1 and R2 are wired in parallel, their total resistance, R, is given by this formula:

```
1 / R = ( 1 / R1 ) + ( 1 / R2 )
```

Therefore, the total resistance in the bottom half of the circuit is now 5M instead of 10M, and the voltage drop across the upper resistor becomes twice the voltage drop across the two lower resistors.

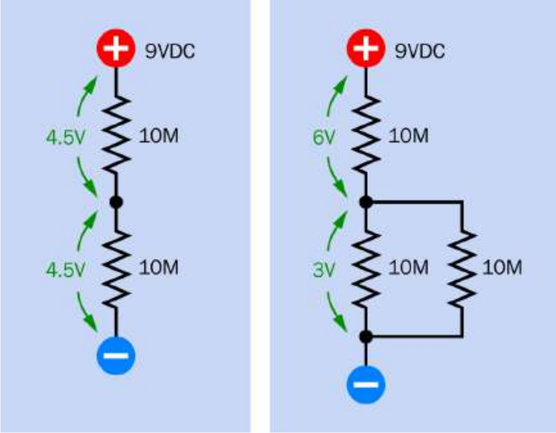

Figure 31-5 *If a meter measuring the voltage drop across a load has an internal resistance comparable to that of a load, the meter will not give an accurate reading. See text for details.*

Now suppose that the additional 10M resistor is actually the fixed resistance inside a volt meter. In fact, many volt meters do have an internal resistance of around 10M. Because the meter has reduced the resistance in the bottom half of the circuit by a factor of 2, it measures the voltage drop as 3V. If the meter had an ideal, infinite resistance, it would give the correct value of 4.5V. In the real world, that is impossible.

Bar Graph

Sometimes it is useful to represent voltage with a graphical display. A *bar graph* component makes this possible, and is often used in audio equipment.

The bar graph can consist of a row of *LEDs*. In a component designed for this purpose, there are usually 10 or more. To represent 0 volts, all the LEDs remain dark. More LEDs are illuminated as the voltage increases.

Because the bar graph itself only contains LEDs, a driver is necessary to convert voltage into the "rising thermometer" effect. Examples are the LM3914 (linear), the LM3915 (logarithmic, 3 dB per step), and the LM3916 (using VU or Volume Units, for audio). Each of these components has 10 LED outputs, and is fabricated using a chain of resistors as a multitap voltage divider, with 10 comparators.

A driver can be set to show only one LED corresponding with the current voltage level, or (more commonly) a cumulative string of LEDs from zero upward. Some bar graphs also contain different colored LEDs, such as green for the first 7, yellow for the next 2, and red for the final one.

A microcontroller containing an analog-to-digital converter can be used to illuminate a bar graph, instead of a driver chip.

An example of a bar graph is the Avago HDSP-4830, shown in Figure 31-6.

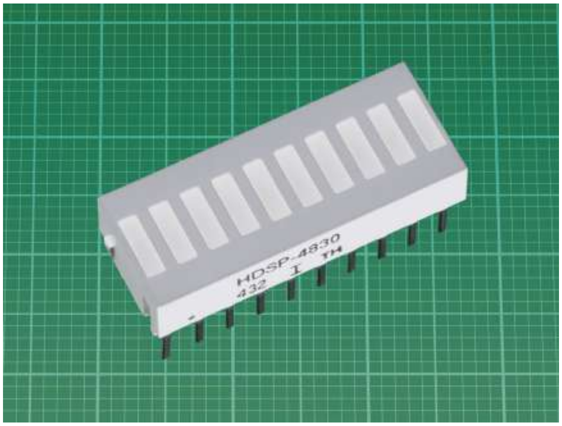

Figure 31-6 *A bar graph LED display, which can be used to represent voltage. The background grid is in millimeters.*

What Can Go Wrong

Confusing AC with DC

A panel meter that is designed only to measure DC should not be used with AC, and vice-versa. Erroneous readings or damage to the meter may result.

High Circuit Impedance

When a meter is measuring the voltage drop across a high-value resistance, it will give an inaccurate reading, as it is diverting a significant proportion of the current.

Voltage Out of Range

Attempting to measure a voltage that exceeds the limited range of a meter can damage the meter or blow its internal fuse, if it has one. Attempting to measure a voltage that is at the very low end of a meter's range will result in an inaccurate value.

Voltage Relative to Ground

If one input to a digital meter is connected with ground, the meter will only be able to measure voltage in a circuit relative to ground.

Sensor Output

A

This appendix provides some basic information about nine forms of sensor outputs, and the ways in which they can be processed. Other types of encoded output exist, but those examined here are the ones most likely to be found.

Figure A-1 provides an overview. Every sensor initially creates an analog output, which is sometimes connected directly to an output pin. In thermistors and photoresistors, for example, the internal resistance of the component constitutes its output. In many sensors, however, the behavior of the sensing element is processed internally to create voltage, open collector, encoded pulse stream, or current output.

If the sensor is chip-based, it may process the analog sensor response internally to create a binary output or a digital output.

In this Encyclopedia, the term "binary output" means "an output that has two states," usually logic-low or logic-high. The states may be accessible via an output pin, or may be processed internally to create a pulse stream. In that case the stream fluctuates between the two states as a way of encoding an analog value with pulse-width modulation (PWM) or frequency. Other types of encoding are also possible, but are unusual.

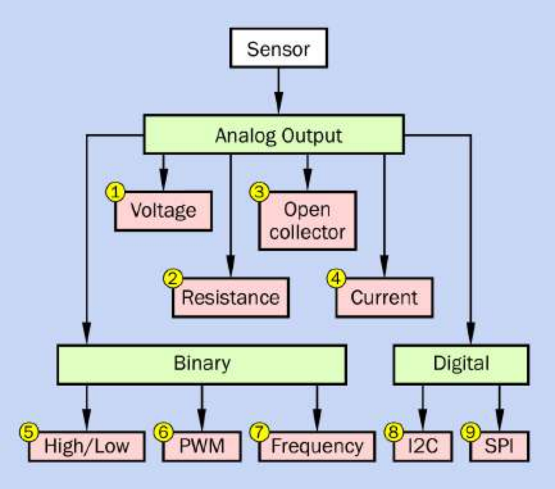

Figure A-1 *Nine possible types of output from sensors. The primary categories are highlighted in green.*

The term *digital output* is used here to mean one or two bytes of data that are stored in a register (memory location) in the sensor chip. While other forms of output are "always on" and can be accessed at any time from an output pin, a digital output is usually not available until an external device, such as a microcontroller, sends an instruction to the sensor chip, telling it to return the data. This two-way communication usually is handled by the I2C or SPI protocols (other protocols exist, but are less common).

Analog Outputs

Voltage is by far the most common form of analog output. Other forms of analog output can easily be converted to a voltage value, using the simple techniques described here.

1. Analog: Voltage

Direct Connection: Analog-to-Analog

An analog voltage output can be connected directly to an analog input, so long as the range is compatible and the sensor can provide sufficient current. Examples of external analog devices would be an analog volt meter, a light source or sound source that changes intensity, or a transistor or op-amp that will amplify the output for other audio/visual purposes.

If the voltage output from the sensor rises above a usable range, it can be converted to a lower value by applying it across two resistors connected in series to form a voltage divider. This is shown in Figure A-2.

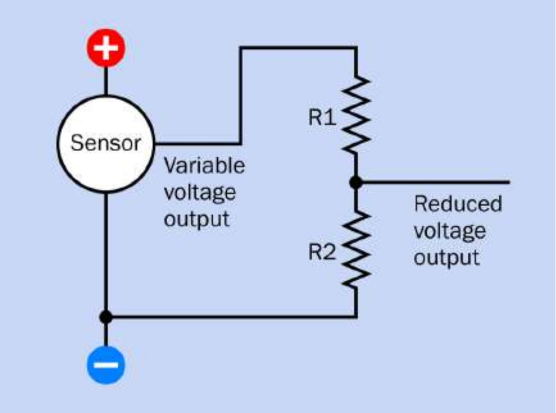

Figure A-2 *Using a voltage divider to reduce the range from a sensor.*

The values of R1 and R2 can be derived from this basic formula, where V_{SEN} is the voltage from the sensor and V_{OUT} is the output from the voltage divider:

Example A-1

$$V_{OUT} = V_{SEN} * (R2 / (R1 + R2))$$

The impedance of the device used for voltage sensing must be high relative to the values of R1 and R2. Note also that if the analog voltage out from the sensor varies linearly with the phenomenon being sensed, this relationship is likely to be disturbed by the voltage divider.

Analog-to-Binary Conversion

The term "binary" is used here to mean an output that can be in one of two states, such as logic-high and logic-low.

A varying analog voltage output can be simplified by passing it through a component that transforms the signal into binary form. This may be done by using a logic chip with a *Schmitt trigger* input, a *zener diode*, or a **comparator**. (For a description of comparators, see the entry in Volume 2.) A comparator provides desirable features such as adjustable positive feedback to create hysteresis. It may be used, for example, to convert a slowly changing signal from a phototransistor while the sun is setting, to a high/low output that can activate a relay to switch on a light.

Analog-to-Digital Conversion

The analog voltage output from a sensor can be digitized by an external analog-to-digital converter (ADC), either inside a microcontroller or by a separate ADC chip.

If a microcontroller is used, a sensor may often be wired directly to an input pin that connects internally with an ADC. A program in the microcontroller can then assess the integer output from the ADC and either execute a conditional statement or convert the value to a format appropriate for another device, such as a digital display.

If an ADC chip is used, there are thousands to choose from. A few basics:

- A *flash* converter contains a row of comparators with different reference voltages generated with a chain of equal resistors. The comparator outputs are fed into a priority **encoder** that outputs a binary number. This system is very fast but has limited resolution.

- A *successive approximation* converter uses a single comparator, comparing the input voltage with the output from a DAC. The binary number that is supplied to the DAC is determined one bit at a time, from the most significant to the least significant bit, using the comparator's result to determine if the bit should be 0 or 1. These bits are stored in a register, called a successive approximation register (SAR). When the process finishes, the SAR contains a binary representation of the input voltage. This type of ADC can achieve high resolution (many bits) at the cost of lower conversion speed.
- In a *dual slope* converter, a capacitor is charged for a fixed time, with a rate proportional to the input voltage, then discharged at a known rate while measuring time by counting clock pulses. The resulting count is the ADC output.

  The name of the converter is derived from the voltage on the capacitor.
- A *voltage-to-frequency* converter uses a voltage-controlled oscillator to produce pulses with a frequency proportional to the input voltage. If the pulses are counted over a fixed time interval, the count is proportional to the signal level.

The number of the bits in the output from an ADC must be sufficient to digitize the input voltage range with desired accuracy. Because the voltage range may contain unexpected peaks, a cautious strategy is to use many more bits than are necessary. However, this means that for most of the time, only a few bits will be used to represent the low end of the voltage range, and accuracy will suffer.

For example, suppose a voltage input normally ranges from 0V to 2V, with occasional brief excursions to 8V. An 8-bit ADC can provide 256 digital values to represent input voltages. If the values are spread uniformly over the full 8V input range, the least significant bit can measure 1/32nd of a volt, or about 31mV. Smaller voltage fluctuations will be ignored. On the other hand, if the 256 values are used to measure a range of just 2 volts, the least significant bit can measure 1/128th of a volt, or slightly less than 8mV—but voltages higher than 2V will be clipped.

An ADC will typically require a reference voltage, and will digitize the range from 0V to that voltage. The reference voltage must be chosen with the issues of accuracy and range in mind.

A microcontroller may provide a feature in its language to perform automatic scaling of an analog input within limits set by a variable in program code. This is done by comparing the input with a selectable voltage level, such as the voltage of the power supply, an externally supplied voltage, or a fixed built in reference. While the ADC in the microcontroller normally digitizes values from 0V to the supply voltage for the chip, the conversion routine can instruct the microcontroller to use its full number of bits (often, 10) to digitize an input range from 0V to 1V.

For a higher sample rate, an ADC chip may be connected with the microcontroller over an I2C or SPI bus.

2. Analog: Resistance

Resistance-to-Voltage Conversion

A sensor that changes its resistance as it responds to its environment can be placed in a *voltage divider* to provide an analog voltage output. This is illustrated in Figure A-3.

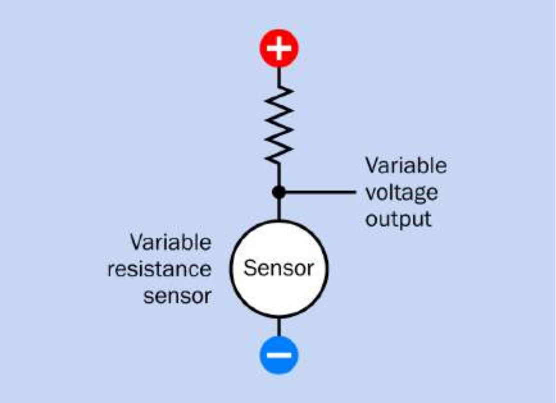
Variable
voltage
output
Variable
resistance
sensor
Sensor

Figure A-3 *The basic principle of putting a variable-resistance sensor in series with a fixed resistance, to create a voltage divider.*

To choose the value for the series resistor, if R_{MIN} and R_{MAX} are the minimum and maximum resistance values for the sensor, the optimum value R_S for the series resistor, to produce the widest variation in voltages, will be found from this formula:

$$R_s = \sqrt{(R_{min} * R_{max})}$$

When the sensor has been set up in this way, the output can now be processed in the same way as any analog voltage output from a sensor.

The datasheet of the sensor should be consulted to make sure that a series resistor value does not allow any chance of the sensor being damaged by excessive current.

3. Analog: Open Collector

Many sensor packages or modules include a bipolar transistor that has an *open-collector* output (or *open drain*, if a CMOS transistor is used). The transistor may or may not be incorporated in an internal **op-amp** (described in Volume 2). Either way, the principle is the same.

Figure A-4 shows a sensor, in darker blue, that has one connection for positive power, another connection for negative ground, and a third connection to the collector of the internal transistor.

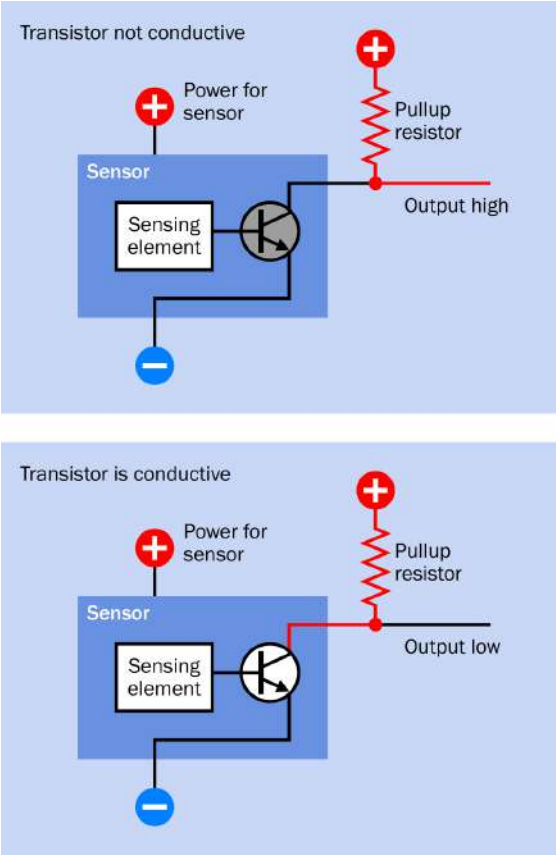
Transistor not conductive
Power for
sensor
Pullup
resistor
Sensor
Output high
Sensing
element
Transistor is conductive
Power for
sensor
Pullup
resistor
Sensor
Output low
Sensing
element

Figure A-4 *How to use an open-collector output from an internal transistor.*

In the upper section of the figure, the sensing element is not applying voltage to the base of the internal transistor, and the transistor only conducts a tiny amount of leakage current. Power applied to it through an external *pullup resistor* cannot reach negative ground in any significant quantity, and therefore it can provide a voltage input to a high-impedance device such as a microcontroller, or can power a component such as an LED, which draws relatively little current (20mA or less).

In the lower section of the figure, the sensing element is now applying voltage to the base of the transistor, drastically lowering its effective resistance. The transistor diverts current from the pullup resistor to ground, and the output appears to go low.

The type of sensing element will determine whether the transistor becomes conductive or nonconductive when the element detects a stimulus.

The value that should be chosen for the pullup resistor will depend on the impedance of any device attached to the open-collector output. A 10K resistor may be appropriate for use with a device such as a microcontroller, which has very high impedance. At the other extreme, if the device is an LED, a 330-ohm resistor may be necessary. The value of the pullup resistor must be sufficiently low to enable reliable operation, but sufficiently high to prevent excessive current from passing through the internal transistor when it becomes conductive (20mA is a common maximum value).

The voltage from the open collector can be processed in the same way as any analog voltage output from a sensor.

An open-collector output may be used when the outputs from multiple devices share the same bus. One device can drive the bus without the problem of other devices attempting to hold the bus voltage high.

4. Analog: Current

Relatively few sensors provide an output consisting of variations in current. Some semiconductor temperature sensors function in this way. The output current can be converted to a voltage output simply by placing a fixed series resistor, as shown in Figure A-5.

The voltage at the point shown, relative to negative ground, will vary linearly with the current. The value of the resistor should be defined in a datasheet for the sensor.

The voltage can now be processed in the same way as any analog voltage output from a sensor.

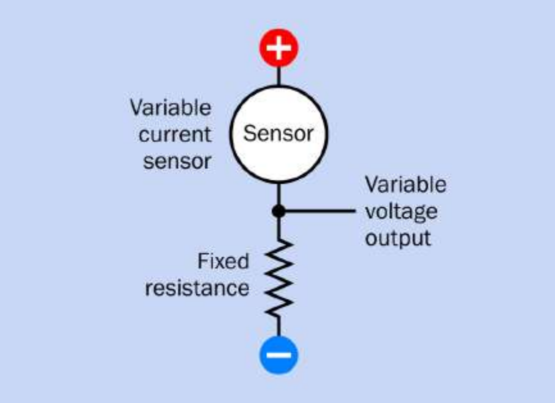

Figure A-5 *How to convert the output from a variable-current sensor.*

5. Binary: High/Low

A sensor that provides a binary output (that is, an output that is either logic-high or logic-low) can be connected directly with a microcontroller, if the voltage range is compatible. Program code in the microcontroller can then test the pin to establish its state. Note that some microcontrollers require a 3.3VDC power supply, while sensor chips may use 5VDC.

A binary output can also be used to control a solid-state relay, or an electromagnetic relay if transistor amplification is used. The output may be sufficient to power an LED indicator.

6. Binary: PWM

PWM is an abbreviation for *pulse-width modulation*. The sensor emits a stream of square-wave pulses with a fixed frequency, but the width of each pulse varies with the stimulus to which the sensor is responding. The width of each high pulse, relative to the wavelength between the start of one pulse and the start of the next, is called the *duty cycle*. A duty cycle of 0% means that there are no pulses at all. With a duty cycle of 100%, there are no gaps between pulses, so the output is high all the time. With a duty cycle of 50%, the duration of each high pulse is the same as the duration of the gap between pulses.

Various microcontrollers offer different ways to decode a PWM pulse stream. The most basic way is for a program to check an input pin repeatedly, as fast as possible, until a high state is detected. The microcontroller copies the value of its internal clock into a variable, then continues to check the input repeatedly until the pulse ends. The pulse duration that has been measured can be converted to a sensor value with a formula or lookup table.

This system is not recommended, as the microcontroller may miss the next pulse while it is converting the value of the previous pulse. To address this problem, a microcontroller language may offer a function that blocks execution of code while waiting for a pulse. The pulseIn() function on the Arduino is an example of this feature. However, the microprocessor must now spend most of its time waiting for a pulse instead of doing useful work.

A better solution is to write a program that is interrupt-driven.

Another option for decoding PWM is to use a low-pass filter that converts the pulse stream into an analog voltage, although some ripples will tend to remain.

Finally PWM can be used directly to power an LED or a DC motor, with transistor amplification as required. The speed of the motor or the brightness of the LED will vary with the duty cycle.

7. Binary: Frequency

Here again, the pulseIn() function on the Arduino may be used, so long as the frequency is a square wave with a known duty cycle.

8. Digital: I2C

In digital electronics, a *bus* is a communal pathway for sharing data among components or devices. The *I2C bus* is an abbreviation for *inter-integrated circuit* bus, developed originally by Philips in 1982. (Philips has since been subsumed into NXP Semiconductors.) The correct notation for I2C is I^2C, and it is spoken as "I-squared-C." However, the term is very commonly written as I2C.

The I2C standard defines a data sharing protocol that is limited to 400kHz (with some exceptions) and designed to work on a small scale, almost always within one device, and usually on one circuit board. It is a low-cost, simple design. Data is transmitted serially over two wires, and the devices sharing the bus are connected in parallel.

Typically there is one *master* device on the bus, and a number of *slave* devices. Masters and slaves can both transmit information, but the master normally initiates communication. It also emits a clock signal for synchronization of data.

A sensor is a slave device that can be interrogated by a microcontroller in its role as the master device. Because multiple slave devices can share a bus, the microcontroller needs a way to identify the slave that it wishes to talk to, and each slave is assigned a unique address for this purpose. Often a slave will allow the user to modify the last two bits of the address, so that up to four identical devices can share a bus.

Code libraries to support the I2C protocol are available for most microcontrollers, and communication with a sensor that uses I2C should simply require knowledge of the sensor's I2C address. However, the data registers in a sensor can be quite elaborate, requiring careful study of the manufacturer's datasheet. Multiple procedures may be required to set functions on a device (such as the sensitivity range of an accelerometer, or the threshold for a temperature alarm). Multiple procedures may also be required to read data out of a sensor (such as two bytes to define a temperature, and several bytes to obtain time as well as location readings from a GPS module).

9. Digital: SPI

SPI is an acronym for serial peripherals interface, a standard introduced by Motorola that serves a similar function to the I2C bus, described immediately above. The SPI standard is slightly more sophisticated, enabling duplex communication and higher data transfer speeds. However, SPI requires a minimum of three wires shared by all devices on the bus, and an additional device-selection line for each slave device. The benefit of the extra device lines is that devices are easier to select and address than on an I2C bus, where more program statements are required. As is the case for I2C, code libraries for microcontrollers are widely available to support SPI. However, the requirement for SPI to use three pins on a microcontroller, plus an additional pin for each slave device, is a disadvantage.

More sensors have I2C capability than SPI capability. The SPI protocol is potentially much faster than I2C.

A sensor that is SPI-enabled will almost certainly be available in a similar version that uses I2C. An increasing number of chip-based sensors support both protocols.

Glossary

This glossary is not comprehensive. It contains only the technical terms that have been used most frequently in this book in conjunction with sensor attributes.

ADC Analog-to-digital converter, which accepts a varying signal (usually, a voltage) as an input, and converts it to a digital value in the form of a binary number. This number is likely to have a high limit ranging from 255 decimal to 65535 decimal. Many microcontrollers contain their own ADC, which is multiplexed to assess the inputs on several pins. On an Arduino Uno, the ADC creates a digital value ranging from 0 through 1,023.

analog output When a sensor creates a voltage or varies its resistance without steps or increments, as a function of the phenomenon that it is measuring, this is an analog output.

binary output In this Encyclopedia, the term "binary output" describes a sensor output that only has two states: logic-high and logic-low, or on and off. The term is used in some datasheets, but more often a binary output is described, misleadingly, as an analog output.

breakout board A small printed-circuit board containing one or more integrated circuit chips, usually surface-mounted. The board makes the features of the chips easier to access, because it has pins or connectors with 2.54mm (0.1") spacing for convenient experimental use with a breadboard. Additional features may be included, such as a voltage regulator.

chip-based sensor This term is used in the Encyclopedia to describe a sensor that is etched into a silicon chip and usually has signal conditioning components and circuitry built in.

contact bounce The tiny and rapid vibrations of mechanical switch contacts, when the switch opens or closes. If the switch is connected with a digital device such as a logic chip, debouncing hardware may be necessary, to allow the contacts time to settle. If the switch is connected with a microcontroller, a delay from 5ms to 50ms may be written into program code. Different switches have widely different settling times.

decibel A unit that expresses relative power or intensity, often (but not exclusively) applied to audible sound. The decibel is one-tenth of a bel, and is abbreviated dB, with the B capitalized because it is derived from the name of Alexander Graham Bell. Because dB is a logarithmic unit, the scale has no zero origin. However, 0dB may be assigned arbitrarily to any intensity, in which case lower intensities will have a negative value. An increase of 3dB corresponds to a doubling of sound intensity (acoustic energy). However, when a sound is sensed by the human ear and evaluated by the brain, its subjective loudness doubles when the intensity increases by 10dB.

dielectric The insulating layer separating two plates in a capacitor.

hysteresis The difference between thresholds for switching an output on and off. When a sensor exhibits hysteresis, it may be unresponsive to a stimulus slightly above or below its equivalent current value. This may be useful to eliminate numerous responses to very small stimuli —for example, in a room thermostat.

I2C Interintegrated circuit bus. Sometimes written as I^2C, and often referred to verbally as "I-squared-C." A communications protocol that is often used between a microcontroller and other components on a circuit board. For a description, see "8. Digital: I2C" in the Appendix.

IMU Inertial measurement unit, consisting of three accelerometers and three gyroscopes, sometimes with the addition of three magnetometers. It can be used as a navigational aid. It may also be used in handheld user input devices, such as game controllers.

Kelvin A temperature scale, often abbreviated with the letter K, in which each degree is equivalent to a Celsius degree, but 0 degrees is at absolute zero—the temperature at which materials have no heat energy at all. 273 degrees K is approximately equal to 0 degrees C.

MEMS Microelectromechanical system, i.e., an integrated circuit chip that also contains tiny moving parts. For example, a MEMS accelerometer is built around microscopic springs that respond to accelerative forces.

newton A unit of force named after Isaac Newton, abbreviated with the capital letter N. A force of 1N will accelerate a mass of 1kg at a rate of 1 meter per second each second.

open collector output Many sensors have an open collector output, or contain an op-amp that has an open collector output. The output pin is attached to the collector of an internal transistor, with its emitter connected to negative ground. Positive voltage applied through a pullup resistor to the open collector will be grounded when the internal transistor is conducting current, but will be available for other devices when the transistor is off. See "3. Analog: Open Collector".

orthogonal Angled at 90 degrees. Three orthogonal elements in a sensor will all be angled at 90 degrees to each other.

pascal A unit of pressure equivalent to 1 newton of force per square meter.

PIR Passive infrared sensor. See Chapter 4.

pullup resistor A resistor that pulls up an output or input voltage in the absence of a signal. May be used in conjunction with an **open collector output**.

quadrature An encoding system for output from a pair of sensors. If the sensors are identified as A and B, four output combinations are possible: A high and B low; A high and B high; A low and B high; A low and B low. A common application is to show the direction of movement of a magnetic or optical pattern past the sensor pair.

reference temperature The temperature at which the output signal of a temperature sensor is measured. This is often listed in a datasheet.

register A section of memory that stores a digital value (usually 1 or 2 bytes in a sensor).

target Any object that is being detected by a motion sensor, proximity sensor, or presence sensor.

temperature coefficient The percentage increase or decrease in the value of a sensor as a result of unit change in temperature (usually 1 degree Celsius). Often abbreviated as TC. The value may be resistance, voltage, or current, depending on the sensor. The temperature coefficient should be negative if the value of the sensor diminishes when its temperature increases. If it is expressed in parts per million, abbreviated ppm, it can be converted to a percentage by dividing by 10,000.

Wheatstone bridge A network of four resistors. At least one of the resistors has an unknown value, while the others have precisely known reference values. The network enables calculation of the unknown value. See Figure 12-2.

Index

A

absolute humidity, 119, 120
absolute magnetic encoder, 41
absolute pressure, 113
accelerometer
  and gyroscope, 72
  applications, 70
  free fall, 71
  gravity, 71
  how it works, 70
  IMU, 5, 63, 66, 69
  Newton's Second Law, 72
  rotation, 71
  schematic symbol, 70
  spring, 70
  values, 73
  variants, 72
  what can go wrong, 74
  what it does, 69
ADC, 210, 217
air pressure, 114
alcohol vapor sensor (see gas concentration sensor, alcohol)
altimeter, 114
ammeter (see current sensor)
analog-to-digital conversion, 210, 217
anemometer, 125, 126
angle sensor (see rotary position sensor)
angular motion, 21
angular position sensor (see rotary position sensor)
Arduino-compatible
  gas flow rate sensor, 127
  gyroscope, 67
  IMU, 67
  magnetometer, 9
  pressure sensor, 114
  proximity sensor, 34
  temperature sensor, 184, 185
  touch screen, 97
  touch sensor, 91
avalanche diode, 136
axis of the earth, 7

B

bandgap temperature sensor (see semiconductor temperature sensor)
barometer, 114
barometric pressure, 114
Beidou positioning system, 3
bin switch, 57
black-body radiation, 25
Bourdon tube, 112
breakout board, 217
breakover switch (see tilt sensor)
Brokaw cell, 180

C

capacitive displacement sensor, 36
capacitive proximity sensor (see touch sensor)
capacitive touch screen, 96
capacitive touch sensor, 90
carbon microphone, 194
chip-based temperature sensor (see semiconductor temperature sensor)
clinometer, 55
coil magnetometer, 8
communications satellites, 1
compass, 6, 8
condenser microphone, 194
contact bounce, 18, 23
contents
  volume 1, xxi
  volume 2, xxii
  volume 3, xxii
crystal microphone, 196
current sensor
  ammeter, 199-200
  applications, 199
  Hall effect, 202
  how it works, 199
  microcontroller, 201
  panel meter, 199
  schematic, 200
  schematic symbol, 200

- series resistor, 200
- voltage drop, 200
- what can go wrong, 203
- what it does, 199

current-sense resistor, 201

D

debouncing, 18, 48, 61, 217
decibels, 196, 217
declination, 7
detection sensor (see presence sensor)
dew point, 119
dielectric, 90, 218
differential pressure, 113
diode temperature sensor (see semiconductor temperature sensor)
distance sensor (see proximity sensor)
dual slope converter, 211
dynamic pressure, 111

E

electret microphone, 195
encoder
- linear, 41
- magnetic, 41
- rotary, 48

errata, xxv

F

feedback to and from authors, xxv
flash converter, 211
float sensor, 100
float switch, 57
flux density, 6
flux, magnetic, 6
force sensor
- applications, 81
- deformative, 86
- how it works, 82
- how to use it, 86
- piezoelectric, 82
- plastic film, 82, 85-87
- resistive, 82, 86
- schematic symbol, 82
- strain gauge, 82, 83, 85, 88
- units of measurement, 82
- values, 87
- what can go wrong, 88
- what it does, 81
- wheatstone bridge, 83

four-wire touch screen, 96
frequency response, 197
fresnel lens, 27

G

Galileo positioning system, 3
gas concentration sensor
- alcohol, 118
- breakout board, 119
- cross-sensitivity, 118
- how it works, 117
- humidity, 117, 119
- oxygen, 119
- propane, 118
- schematic symbol, 117
- semiconductor-based, 117
- what can go wrong, 122
- what it does, 117

gas flow rate sensor
- anemometer, 125, 126
- applications, 128
- data output, 128
- schematic symbol, 125
- thermal mass, 127
- thermopile, 127
- units of measurement, 128
- what can go wrong, 129
- what it does, 125

gas pressure sensor (see pressure sensor)
gas sensor (see gas concentration sensor)
gauge pressure, 113
geographical meridians, 7
Global Positioning System (see GPS)
GLONASS positioning system, 3
GPS
- antenna, 2
- definitions, 1
- frequencies, 2
- how to use it, 3
- NMEA protocol, 3
- satellites, 1
- schematic symbol, 1
- time-keeping, 4
- tracker, 2
- values, 3
- what can go wrong, 4
- what it does, 1

gray code, 51
gyroscope
- and accelerometer, 72
- applications, 63
- axes, 64
- how it works, 64
- how to use it, 67
- IMU, 5, 63, 66, 69
- resonator (see gyroscope, vibrating)
- schematic symbol, 63
- values, 67
- variants, 66
- vibrating, 64
- what can go wrong, 67
- what it does, 63

H

Hall effect
- current sensor, 202

Hall effect sensor
- advantages, 22
- applications, xx, 52
- float sensor, 100
- how it works, 18
- how to use, 20
- magnetometer, 8
- rotary position sensor, 48
- variants, 19

hard-iron bias, 9
heading, navigational, 5
humidity sensor (see gas concentration sensor, humidity)
hygrometer, 120
hysteresis

  float sensor, 101, 218
  reed switch, 17, 218

I

I2C, 214, 218
IC temperature sensor (see semiconductor temperature sensor)
IMU, 5, 63, 66, 69, 218
inclination, 7
inclinometer, 55
inertial measurement unit (see IMU)
infrared proximity sensor, 32-34
infrared radiation, 188
infrared temperature sensor
  advantages, 187
  applications, 188
  array, 191
  compared with other temperature sensors, 153
  how it works, 188
  limitations, 188
  schematic symbol, 188
  thermopile, 189
  values, 191
  variants, 190
  what can go wrong, 192
  what it does, 187
inrush current limiter, 150, 161
interintegrated circuit bus (see I2C)

K

Kelvin temperature scale, 218

L

light meter, 137
light-dependent resistor (see photoresistor)
linear displacement sensor (see linear position sensor)
linear feedback shift register, 2
linear position sensor
  absolute, 41
  applications, 39, 42
  how it works, 39
  incremental, 41
  linear potentiometer, 40
  magnetic linear encoder, 41
  optical linear encoder, 41
  quadrature, 41, 218
  schematic symbols, 39
  what can go wrong, 43
  what it does, 39
linear potentiometer, 40
linear variable differential transformer, 42
liquid displacement sensor, 102
liquid flow rate sensor
  differential pressure, 109
  magnetic, 108
  paddlewheel, 105
  reed switch, 105
  rotor, 105
  schematic symbols, 105
  thermal mass, 107
  turbine, 106
  ultrasonic, 108
  what can go wrong, 109
  what it does, 105
liquid flow switch, 108
liquid level sensor
  displacement, 102
  float, 100
  float switch, 57
  how it works, 100
  incremental output, 101
  potentiometer, 101
  pressure of liquid, 103
  schematic symbols, 99
  tilting, 104
  turbulence, 104
  ultrasonic, 102
  weight of reservor, 103
  what can go wrong, 104
  what it does, 99
liquid pressure sensor (see pressure sensor)
load cell (see load sensor)
load sensor, 81
  (see also force sensor)
  applications, 82
  deformative, 86
  liquid level measurement, 103
  schematic symbol, 82
  strain gauge, 83, 85, 88
  values, 87
  what can go wrong, 88
  what it does, 81
LVDT (see linear variable differential transformer)

M

magnetic
  axis, 7
  ball in tilt sensor, 59
  declination, 7
  encoder, 41, 52
  field, 6
  flux, 6
  liquid flow rate sensor, 108
  meridians, 7
  presence sensors, 16
  rotary position sensor, 47
magnetic, linear encoder, 41
magnetometer
  breakout board, 6, 9
  coil, 8
  hard-iron bias, 9
  heading, 5
  how it works, 6
  how to use it, 9
  IMU, 5, 63, 69
  magnetoresistance, 8
  meridians, 7
  placement, 10
  scalar, 5
  schematic symbol, 5
  soft-iron bias, 9
  vector, 5
  what can go wrong, 9
  what it does, 5
magnetoresistance, 8
mass flow rate sensor (see gas flow rate sensor)
mathematical syntax, xxiv
membrane switch, 89
MEMS, xix, 218
MEMS microphone, 195
mercury switch, 58, 61, 78

meridians, 7
metric system, xxiv
microelectromechanical devices (see MEMS)
microphone
  carbon, 194
  condenser, 194
  crystal, 196
  directionality, 197
  dynamic, 194
  electret, 195
  frequency response, 197
  how it works, 194
  impedance, 198
  MEMS, 195
  moving coil, 194
  piezoelectric, 196
  rolloff, 197
  schematic symbols, 193
  sensitivity, 196
  signal-to-noise ratio, 198
  total harmonic distortion, 198
  units of measurement, 196
  what can go wrong, 198
  what it does, 193
motion sensor (see PIR)
mouse, 50
moving coil microphone, 194

N

navigational heading, 5
newton unit of force, 218
Newton's Second Law, 72
NMEA protocol, 3
north pole, 7
NTC thermistor (see thermistor, NTC)

O

object detector (see presence sensor)
object presence sensor (see presence sensor)
open collector output, 218
opt-pass sensor, 15
optical
  linear encoder, 41
optical isolator, 132
optical presence sensor, 12
optical rotary encoder, 48, 53
optical switch, 13, 15, 50
optocoupler, 142
optointerrupter, 21
orthogonal, 218
output
  analog current conversion, 213
  analog voltage conversion, 210, 217
  analog-to-binary conversion, 210
  analog-to-digital conversion, 210
  binary, 209, 210, 213, 217
  digital, 209
  dual slope converter, 211
  flash converter, 211
  frequency conversion, 214
  I2C, 214, 218
  open collector, 212, 218
  options, 209
  pullup resistor, 213, 218
  PWM conversion, 209, 213
  resistance-to-voltage conversion, 211
  scaling, 211
  SPI, 215
  successive approximation converter, 211
  voltage divider, 210
  voltage-to-frequency converter, 211
over-current protection, 160
over-temperature protection, 159
oxygen sensor (see gas concentration sensor, oxygen)

P

paddlewheel flow rate sensor, 105
panel meter, 199, 205
pascal unit of pressure, 218
passive infrared motion sensor (see PIR)
pendulum switch, 58
photocell (see photoresistor)
photoconductive cell (see photoresistor)
photodarlington, 142
photodiode
  applications, 135
  array, 137
  avalanche diode, 136
  band gap, 137
  color sensing, 137
  dark current, 135
  how it works, 135
  how to use it, 139
  light meter, 137
  photoconductive mode, 135
  phototransistor comparison, 143
  photovoltaic mode, 135
  PIN type, 136
  reverse bias, 135
  schematic symbols, 135
  ultraviolet, 137
  values, 138
  variants, 136-137
  wavelength range, 137
  what can go wrong, 140
  what it does, 135
photointerruptor, 13
photomicrosensor, 15
photoresistor
  availability, 131
  how it works, 132
  how to use it, 133
  optical isolator, 132
  phototransistor comparison, 133
  schematic symbols, 131
  sizes, 132
  voltage divider, 134
  wavelengths of sensitivity, 134
  what can go wrong, 134
  what it does, 131
phototransistor
  applications, 142
  base connection, 142
  binning, 144
  FET, 142
  how it works, 142
  how to use it, 144

  - optocoupler, 142
  - output current calculation, 145
  - photodarlington, 142
  - photodiode comparison, 143
  - photoresistor comparison, 133
  - schematic symbols, 141
  - values, 143
  - what can go wrong, 145
  - what it does, 141
- photovoltaic effect, 135
- piezoelectric
  - force sensor, 82
  - heat detector, 26
  - microphone, 196
  - ultrasonic transducer, 32
  - vibration sensor, 76
- PIN photodiode, 136
- pin-and-spring vibration sensor, 76
- PIR
  - applications, 25
  - fresnel lens, 27
  - how it works, 26
  - schematic symbols, 25
  - variants, 29
  - what can go wrong, 30
  - what it does, 25
- potentiometer
  - arc segment, 46
  - end stops, 46
  - float sensor, 101
  - logarithmic, 46
  - multiturn, 46
  - rotary position, 46
- presence sensor
  - angular motion, 21
  - configuration, 21
  - linear motion, 21
  - magnetic, 16
  - optical, 12, 22, 23
  - optointerrupter, 21
  - reflective, 12
  - retroreflective, 12, 15
  - schematic symbol, 12
  - transmissive, 12-13
  - what can go wrong, 23
  - what it does, 11
- pressure sensor
  - absolute pressure, 113
  - altitude, 114
  - ambient air, 114
  - applications, 111
  - barometric, 114
  - Bourdon tube, 112
  - diaphragm, 112
  - differential pressure, 113
  - dynamic, 111
  - gas and liquid behavior, 112
  - gases, 115
  - gauge pressure, 113
  - how it works, 112
  - schematic symbols, 111
  - sensing elements, 112
  - static, 111
  - units of measurement, 112
  - what can go wrong, 116
  - what it does, 111
- propane sensor (see gas concentration sensor, propane)
- proximity sensor
  - applications, 32
  - capacitive displacement, 36
  - infrared, 32-34
  - schematic symbols, 31
  - triangulation, 32
  - ultrasound, 32-33
  - what can go wrong, 38
  - what it does, 31
- PTC thermistor (see thermistor, PTC)
- pullup resistor, 213, 218
- pulse-width modulation (see PWM)
- PWM, 209, 213
- pyroelectric detector, 26

Q

- quadrature, 41, 49, 218

R

- range finder (see proximity sensor)
- reed switch
  - advantages, 22
  - alarm system, 17
  - float sensor, 100
  - how it works, 17
  - how to use it, 18
  - hysteresis, 17
  - liquid flow rate sensor, 105
  - values, 18
  - variants, 17
  - what can go wrong, 23
- reference sources, xxvii
- reference temperature, 218
- reflective interrupter, 15
- reflective object sensor, 15
- reflective optical sensor, 15
- reflective photointerrupter, 15
- register, 218
- relative humidity, 119, 120
- resistance temperature detector (see RTD temperature sensor)
- resistance-to-voltage conversion, 211
- resistive force sensor, 82, 86
- resistive temperature device (see RTD temperature sensor)
- resistive touch screen, 95
- retroreflective presence sensor, 12, 15
- rotary encoder, 48, 52-53
- rotary position sensor
  - absolute, 51
  - accelerometer, 71
  - applications, 45
  - gray code, 51
  - Hall Effect sensor, 48
  - incremental, 49
  - magnetic, 47
  - mouse, 50
  - optical encoder, 48
  - potentiometer, 46
  - quadrature, 49, 218
  - reflective optical, 48
  - rotary encoder, 48
  - schematic symbols, 46
  - sensing chips, 48
  - transmissive optical, 48
  - what can go wrong, 53
  - what it does, 45
- rotary sensor (see rotary position sensor)
- rotating coil magnetometer, 8

rotation sensor (see rotary position sensor)
rotational encoder (see rotary encoder)
rotational position sensor (see rotary position sensor)
RTD temperature sensor
  applications, 174
  attributes, 173
  compared with other temperature sensors, 153
  how it works, 174
  output conversion, 176
  probe enclosure, 176
  schematic symbol, 174
  three-wire variant, 175
  variants, 175
  what can go wrong, 176
  what it does, 173

S

Safari Books Online, xxvi
satellites, 1
scalar magnetometer, 5
schematic conventions, xxiv
schematic symbol
  accelerometer, 70
  current sensor, 200
  force sensor, 82
  gas concentration sensor, 117
  gas flow rate sensor, 125
  GPS, 1
  gyroscope, 63
  infrared temperature sensor, 188
  linear position sensor, 39
  liquid flow rate sensor, 105
  liquid level sensor, 99
  load sensor, 82
  magnetometer, 5
  microphone, 193
  photodiode, 135
  photoresistor, 131
  phototransistor, 141
  PIR, 25
  presence sensor, 12
  pressure sensor, 111
  proximity sensor, 31
  rotary position sensor, 46
  RTD temperature sensor, 174
  semiconductor temperature sensor, 178
  thermistor, NTC, 147
  thermistor, PTC, 157
  thermocouple, 166
  tilt sensor, 56
  touch screen, 95
  touch sensor, 90
  vibration sensor, 75
  voltage sensor, 206
Seebeck effect, 167
semiconductor temperature sensor
  analog current output, 182
  analog voltage output, 181
  applications, 178
  attributes, 178
  Brokaw cell, 180
  CMOS type, 185
  compared with other temperature sensors, 153
  digital output, 183
  how it works, 179
  multiple transistors, 179
  schematic symbol, 178
  variants, 180
  what can go wrong, 185
  what it does, 177
sensor
  categories, xx
  chip-based, 217
  multifunction, xx
  output options, 209
  suppliers, xxv
serial peripherals interface (see SPI)
shift register, 2
silistor, 158
slider potentiometer, 40
soft-iron bias, 9
solar panel, 135
sound pressure, 196
sources, xxvii
south pole, 7
speed sensor, 45
SPI, 215
static pressure, 111
stomp box, 133
strain gauge, 82-83, 85, 88
successive approximation converter, 211
suppliers, xxv
surge limiter (see inrush current limiter)

T

tactile switch, 89
target, 218
temperature coefficient, 147, 152, 218
temperature sensor
  comparison of types, 153
  infrared (see infrared temperature sensor)
  noncontact type, 187
  RTD (see RTD temperature sensor)
  semiconductor (see semiconductor temperature sensor)
  thermistor, NTC (see thermistor, NTC)
  thermistor, PTC (see thermistor, PTC)
  thermocouple (see thermocouple)
thermal mass flow rate sensor (see gas flow rate sensor)
thermal mass liquid flow rate sensor, 107
thermistor, NTC
  applications, 148
  compared with other temperature sensors, 153
  datasheet terminology, 152
  how it works, 148
  inrush current limiter, 150
  output conversion, 149
  schematic symbols, 147
  values, 152
  what can go wrong, 153
  what it does, 147
  wheatstone bridge, 150
thermistor, PTC

  - compared with other temperature sensors, 153
  - externally heated, 159
  - heating element, 162
  - inrush current limiter, 161
  - lighting ballast, 162
  - linear, 158
  - nonlinear, 158, 159
  - over-current protection, 160
  - over-temperature protection, 159
  - schematic symbols, 157
  - silistor, 158
  - starting current, 162
  - what can go wrong, 163
  - what it does, 157
- thermocouple
  - applications, 166
  - compared with other temperature sensors, 153
  - how it works, 167
  - how to use it, 168
  - identifiers, 169
  - measurement junction, 168
  - output conversion, 170
  - reference junction, 168
  - schematic symbol, 166
  - Seebeck coefficients, 169
  - Seebeck effect, 167
  - thermopile, 171
  - what can go wrong, 171
  - what it does, 165
- thermopile
  - gas flow rate sensor, 127
  - infrared temperature sensor, 189
  - overview, 171
- tilt sensor
  - applications, 57
  - bin switch, 57
  - breakout board, 59
  - float switch, 57
  - how it works, 56
  - how to use it, 61
  - magnetic ball, 59
  - mercury switch, 58, 61
  - miniaturized, 59
  - pendulum switch, 58
  - schematic symbol, 56
  - two-axis, 59
  - values, 60
  - what can go wrong, 61
  - what it does, 55
- tilt switch (see tilt sensor)
- tip sensor (see tilt sensor)
- tipover switch (see tilt sensor)
- touch pad (see touch sensor)
- touch screen
  - Arduino-compatible, 97
  - capacitive, 96
  - four-wire, 96
  - resistive, 95
  - schematic symbol, 95
  - variants, 95, 97
  - what it does, 95
- touch sensor
  - applications, 90
  - breakout boards, 91
  - capacitive, 90
  - design considerations, 92
  - how it works, 90
  - how to use it, 91
  - schematic symbols, 90
  - tactile feedback, 90
  - touch pad availability, 91
  - what can go wrong, 93
  - what it does, 89
  - wheels and strips, 92
- touch strip, 92
- touch wheel, 92
- triangulation, 32
- turbine flow rate sensor, 106

U

- ultrasonic anemometer, 126
- ultrasonic flow rate sensor, 108
- ultrasonic liquid level sensor, 102
- ultrasonic proximity sensor, 32-33
- units of measurement, xxiv

V

- Vactrol, 132
- vapor sensor (see gas concentration sensor)
- vector magnetometer, 5
- vibrating gyroscope, 64
- vibration sensor
  - how to use it, 79
  - magnetic, 77
  - mercury switch, 78
  - mousetrap type, 77
  - piezoelectric, 76
  - pin-and-spring, 76
  - schematic symbols, 75
  - values, 78
  - variables, 78
  - variants, 75
  - what can go wrong, 79
  - what it does, 75
- vibration switch (see vibration sensor)
- video display, 95
- volt meter (see voltage sensor)
- voltage divider, 210
- voltage sensor
  - accuracy, 207
  - bar graph, 207
  - how it works, 206
  - panel meter, 205
  - schematic, 206
  - schematic symbol, 206
  - volt meter, 205
  - what can go wrong, 208
  - what it does, 205
- voltage-to-frequency converter, 211

W

- wheatstone bridge, 83, 150, 160, 219

About the Authors

Charles Platt is the author of *Make: Electronics* and *Make: More Electronics*. He is a former senior writer for *Wired* magazine, and is a contributing editor to *Make:* magazine, for which he writes a column on electronics.

Fredrik Jansson is a physicist from Finland, with a PhD from Åbo Akademi University. He is currently living in The Netherlands, where he works on swarm robotics and simulates sea animals in the Computational Science group at the University of Amsterdam. Fredrik has always loved scavenging discarded household electronics for parts, and is a somewhat inactive radio amateur with the call sign OH1HSN. He also fact-checked Charles Platt's previous book, *Make: More Electronics*.

Colophon

The cover and body fonts are Myriad Pro, the heading font is Benton Sans, and the code font is Dalton Maag's Ubuntu Mono.

US 20030197093A1

(19) **United States**

(12) **Patent Application Publication**

St.Clair

(10) **Pub. No.: US 2003/0197093 A1**

(43) **Pub. Date: Oct. 23, 2003**

(54) MAGNETIC VORTEX WORMHOLE GENERATOR

(76) Inventor: **John Quincy St.Clair**, San Juan, PR (US)

Correspondence Address:
John St.Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/127,098**

(22) Filed: **Apr. 20, 2002**

Publication Classification

(51) **Int. Cl.^{7}** **B64C 27/00**; B64C 39/00
(52) **U.S. Cl.** **244/62**

(57) ABSTRACT

This invention relates to a magnetic vortex generator which has the ability to generate negative mass and a negative spring constant which, according to Einstein's General Theory of Relativity, is required in order to create a stable wormhole between our space and hyperspace. Two separate, but electrically connected, toroidal coils of differing radii, carry magnetic flux in opposite directions about their common centerline. According to Maxwell's equation, this produces bucking electric fields along said centerline. Because the two solenoids have different radii, the parallel spring constant of both coils is negative. The negative mass together with the negative spring constant produce a real resonant frequency which can distort the spacetime curvature due to the creation of powerful spikes of negative mass. This phenomenon, similar to the common electrical thunderstorm, opens up a wormhole into hyperspace through which low-density hyperspace energy can enter into our dimension. This energy finds many application in new types of power supplies, inertia-less and mass-less spacecraft, vehicles that can travel light-years by moving out of dimension through hyperspace, surgery-less medical tables, cranes for lifting heavy objects, cold-welded crystals for crystal rotors, folding space waveguides, and electromagnetic field propulsion vehicles using highly relativistic fields.

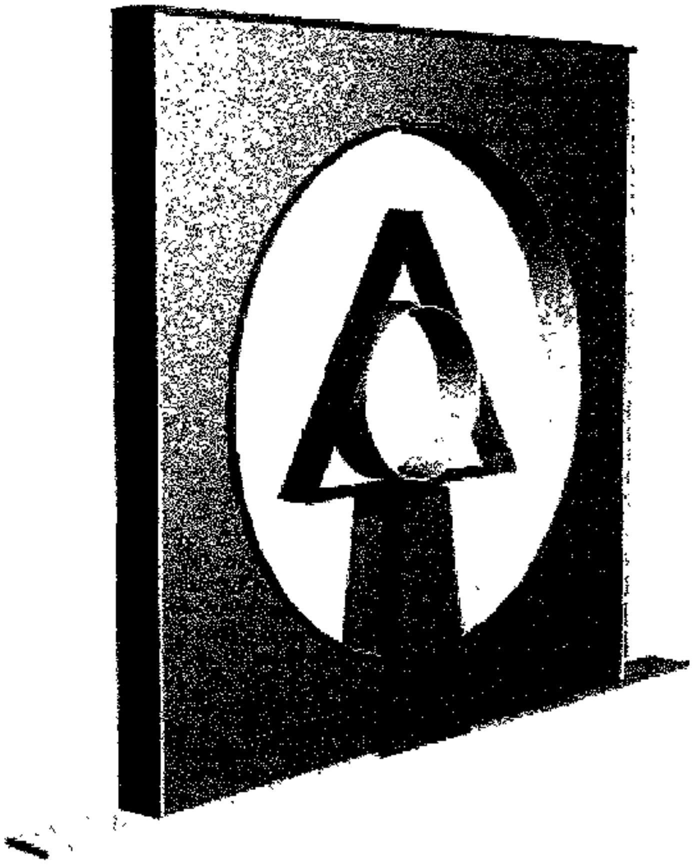

Figure 1

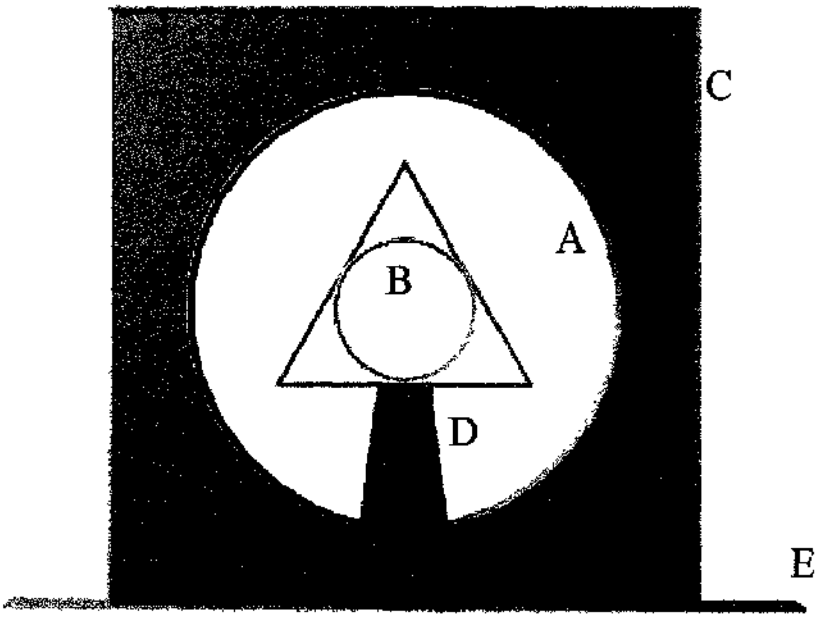

Figure 2

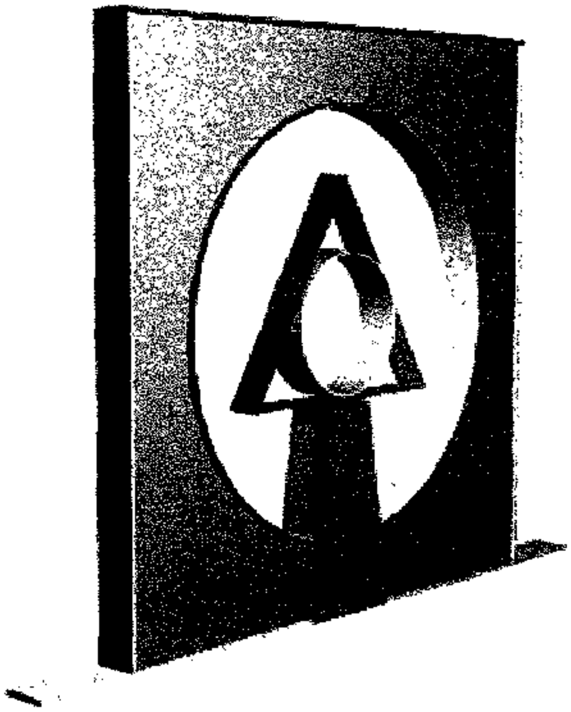

Figure 3

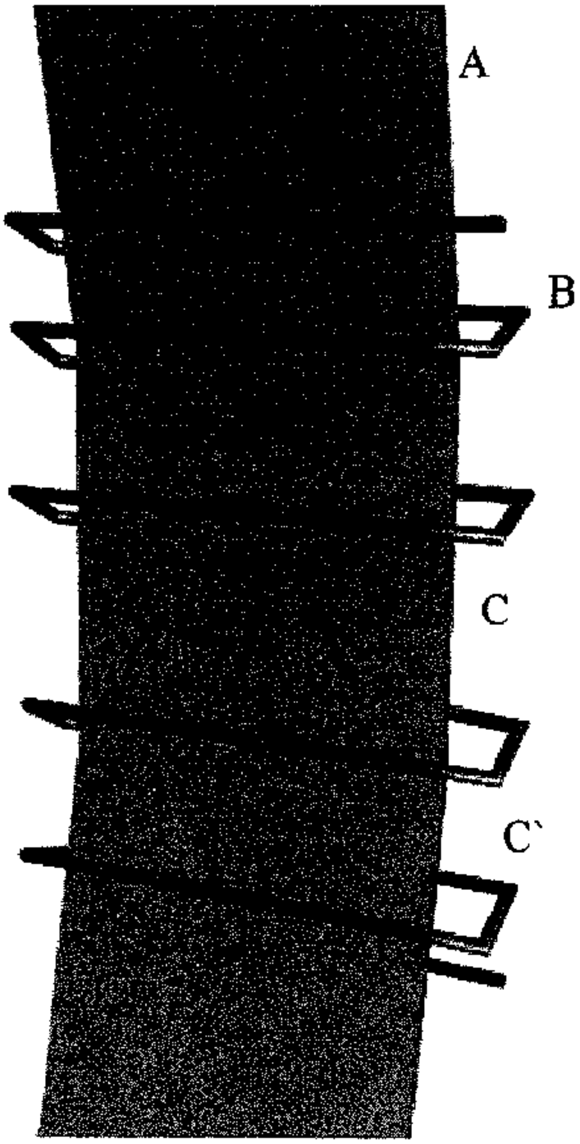

Figure 4

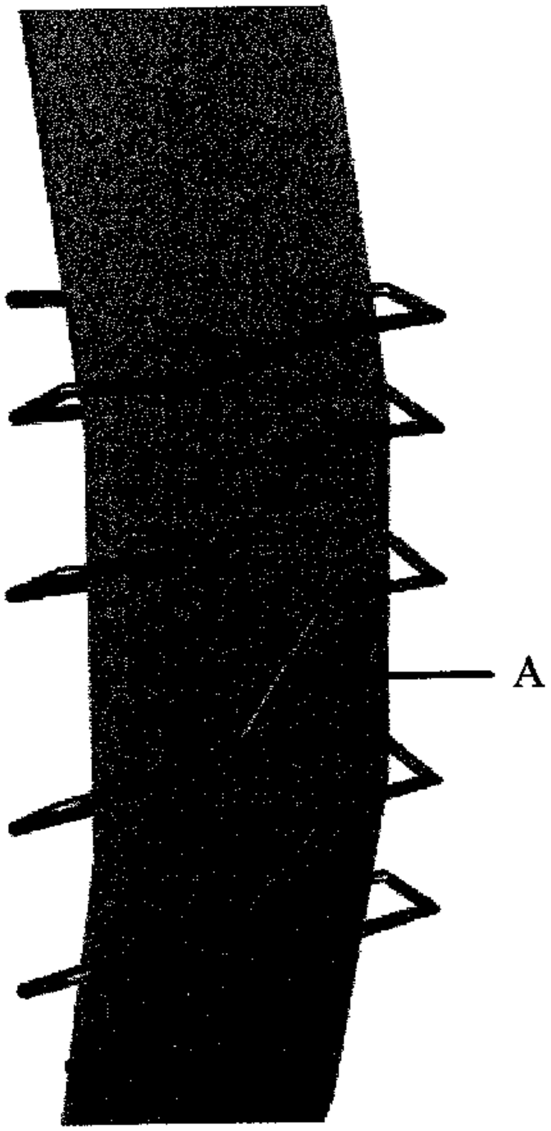

Figure 5

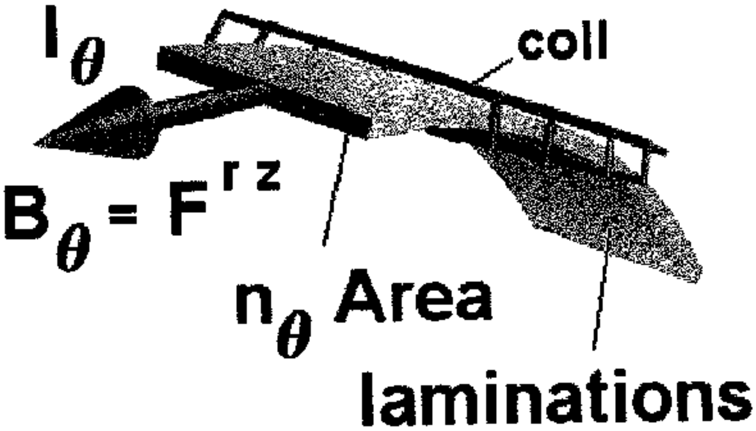
I_θ
coil
$B_\theta = F^{r z}$
n_θ Area
laminations

Figure 6

$$F^{\mu\nu} = \begin{array}{c|cccc|} & t & r & \theta & z \\ t & 0 & 0 & 0 & 0 \\ r & 0 & 0 & 0 & B_\theta \\ \theta & 0 & 0 & 0 & 0 \\ z & 0 & -B_\theta & 0 & 0 \end{array}$$

Figure 7

$$k^z = \frac{\text{coul}}{\text{sec}} \cdot m^2 \cdot \frac{1}{m^2} \cdot \frac{\text{kg}}{\text{sec} \cdot \text{coul}} = \frac{\text{newton}}{m} = \frac{\text{force}}{m}$$

Figure 8

$$k^z = \frac{1}{(r+a)^2} - \frac{1}{r^2} = -a \frac{(2 \cdot r + a)}{\left[(r+a)^2 \cdot r^2\right]}$$

Figure 9

$$g_{\alpha\beta} = \begin{array}{c|cccc|} & \mathrm{t} & \mathrm{r} & \theta & \mathrm{z} \\ t & -1 & & B(\cos\theta - \sin\theta) & 0 \\ r & 0 & 1 & 0 & 0 \\ \theta & B(\cos\theta - \sin\theta) & 0 & r^2 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

Figure 10

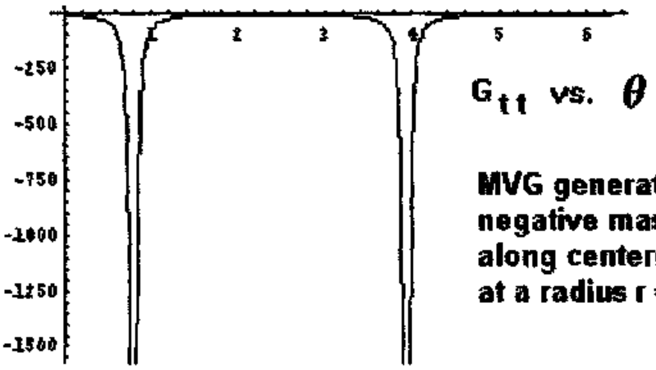

MVG generates two negative mass spikes along centerline of coils at a radius r = 0.

Figure 11

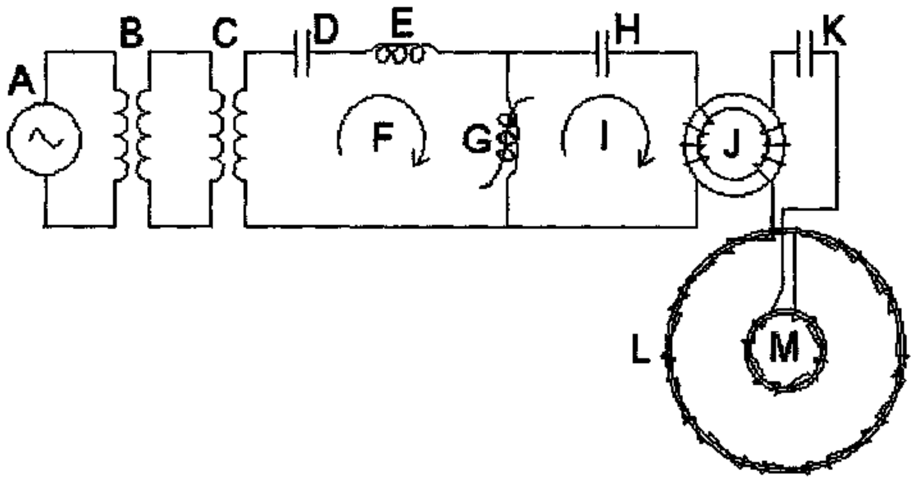
A
B
C
D
E
H
K
F
G
I
J
L
M

Figure 12

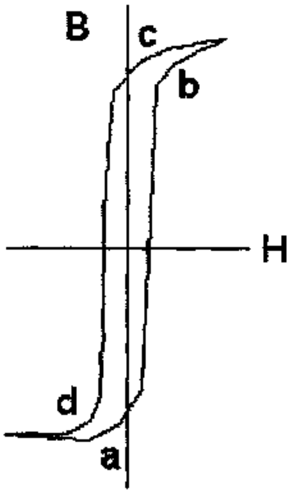

MAGNETIC VORTEX WORMHOLE GENERATOR

BRIEF SUMMARY OF THE INVENTION

[0001] This invention, which is the subject of my present application, is comprised of two solenoids wound with a common wire in opposite directions on two separate toroidal curved-sheet transformer laminates of differing radii. The smaller solenoid is mounted along the centerline of the larger solenoid. This circular magnetic geometry creates linear bucking electric fields along the centerline of the coils. Because the magnetic flux in the laminates travels in opposite directions along arcs of differing radii in the two coils, a negative mass and a negative spring constant are generated by the system. From the theory of gravitational physics, a negative mass is prerequisite to producing a wormhole because it allows the throat of the wormhole to remain open and stable. The creation of the wormhole is facilitated by the appearance of a negative spring constant which allows the spacetime curvature to resonate to such a degree that said wormhole develops between our dimension and another co-dimension of hyperspace. Because the physics constants of hyperspace are different from ours, the wormhole allows hyperspace energy having a low speed of light to enter our dimension. Because electromagnetic fields obey the Lorentz transformation, it is now possible with this lower velocity of light to create huge relativistic fields which can drive the new electromagnetic field propulsion vehicles.

BACKGROUND OF THE INVENTION

[0002] The idea for this coil configuration comes from the observation of powerful thunderstorms, as described by physicist Dr. Richard Feynman in his *Lectures on Physics,* a copy of which is enclosed as a reference. Upon reading his explanation, I realized that the thunderstorm is actually a hyperspace physics phenomenon.

[0003] After the passage of a large lightning storm, people have observed that a car tire rim has merged with the trunk of a tree growing in the ground. Due to the large branches of the tree, there is no possible way that it could slide down the branches and around the trunk. It was observed also that a straw of wheat became embedded in the hard wood of a telephone pole. It turns out that the thunderstorm offers an explanation as to how this can occur.

[0004] After reading Feynman's explanation, it can be seen that the key to this phenomenon is that there is a downward and an upward lightning bolt, sometimes occurring together if the leader branches into two paths. Bolts of lightning also like to strike tall objects such as telephone poles or trees. Now an electrical current moving downward produces a clockwise magnetic B field, as seen from above. On the return stroke, the current is moving upward which produces a counterclockwise magnetic B field. Thus the thunderstorm produces two huge bucking magnetic B fields which is the magnetic geometry of this magnetic vortex generator. Using my tetrahedron diagram, I will then show that the low density hyperspace energy with its low speed of light is able to pull the rim out of dimension so that it can merge with the tree at the moment the lightning strikes.

SUMMARY OF THE INVENTION

[0005] It is the object of this invention to produce two toroidal oppositely-directed magnetic flux fields in two separate yet electrically connected solenoids of differing radii. Because the lines of flux are traveling in toroidal, curved-sheet transformer laminates along arcs of different curvature, the fields produce what is known in gravitational physics as negative mass. Negative mass has the property that if you project it onto a hot surface, the surface will cool rather than heat up. The negative mass, together with the resonance of the spacetime curvature involving a negative spring constant, create a wormhole into hyperspace. This interdimensional connection allows low speed of light hyperspace energy into our dimension which can be used to decrease the weight of an object, or produce huge relativistic electromagnetic fields that can be used to drive the new electromagnetic field propulsion spacecraft.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0006] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0007] **FIG. 1**. Front view of magnetic vortex generator.

[0008] **FIG. 2**. Perspective view of the magnetic vortex generator.

[0009] **FIG. 3**. The non-linear coil winding on the interior face of the laminations.

[0010] **FIG. 4**. The non-linear coil winding on the exterior face of the laminations showing that there is an electrical current component in the theta direction.

[0011] **FIG. 5**. The coil variables for the tensor calculation of the negative spring constant.

[0012] **FIG. 6**. The Faraday electromagnetic tensor showing the position of the magnetic field in the $\{r,z\}$ slots.

[0013] **FIG. 7**. The units of the spring constant.

[0014] **FIG. 8**. The parallel spacetime spring constant.

[0015] **FIG. 9**. The cylindrical g metric tensor including the magnetic fields.

[0016] **FIG. 10**. A plot of the mass term G_n for Einstein's G curvature tensor showing that two negative mass spikes are created along the centerline of the generator. Negative mass is required to keep open the throat of the wormhole.

[0017] **FIG. 11**. Circuit diagram for the magnetic vortex generator.

[0018] **FIG. 12**. BH curve for non-linear SuperMalloy toroidal core. From the equation, $B=\mu H$, the slope of the line is equal to the permeability $\partial B/\partial H=\mu$.

DETAILED DESCRIPTION OF THE INVENTION

[0019] 1. The fact that negative mass is required to keep the throat of a wormhole open has been shown to be valid by physicist Dr. Kip Thorne in the enclosed reference physics paper. The key point of this invention is that if you have a negative mass, you also require a negative spring constant in order to get a real frequency and vice versa. The angular frequency of vibration is equal to the square root of the spring constant K divided by the mass M, or $\omega=\sqrt{K/M}$. If the mass is negative, and the spring constant is positive, the frequency is imaginary. Therefore, in order to get a real frequency, the spring constant has to be negative also.

[0020] 2. Referring to **FIG. 1**, the magnetic vortex generator consists of a large toroidal solenoid (A) and its support structure (C), a smaller toroidal coil (B) with its support structure (D) which holds the coil along the centerline of the larger coil. The entire structure sits on a base (E).

[0021] 3. The two solenoids are wound with a common wire in opposite directions on two separate toroidal curved-sheet transformer laminations of differing radii. From the right hand rule, a changing circular magnetic field in the direction of the fingers produces a linear electric field in the direction of the thumb. Because there are two coils producing two magnetic fields in opposite directions, there are two bucking electric fields down the centerline of the coils. This duplicates the thunderstorm conditions. A perspective view of the generator is shown in **FIG. 2**.

[0022] 4. A enlarged view of the non-linear coil winding is shown in **FIG. 3**. The wire (B) is wound around the thin transformer laminations (A) with a non-linear coil spacing as shown by the difference in spacing between length (C) and (C'). On this inner side of the coil, the wire is wound straight across where it can be seen that the wire is normal to the edge of the laminations. Not shown are two strips of Velcro which keep the wire aligned and offset from the tape-covered metal laminations.

[0023] 5. Since the wire is straight across on this side, the opposite side has to have the wire run in a diagonal manner as seen in **FIG. 4A**. In terms of cylindrical coordinates, the components of the electrical current in the coil flow in the z-direction across the breadth of the lamination and in the O-direction around the lamination.

[0024] 6. The reason for the non-linear spacing is to preserve the vector potential of the coil. The vector potential is a more important field than the magnetic field because it can extend out past the windings of a long solenoid. If you look at the units, it is the field momentum per charge or kilogram meter per second coulomb. Notice that the derivative of the vector potential with respect to time is an electric field, while a derivative with respect to length is the magnetic field. The inductance of the coil times the current density is equal to the vector potential. Thus the inductance gradient of the coil times the current per meter is the magnetic field. So the non-linear coil picks up an additional magnetic field around the coil besides the one created in the laminations by the winding itself

[0025] 7. Taking a closer look at the coil in **FIG. 5**, it is constructed of thin laminations taped together to form a cylindrical shape with a wire coil wrapped around it. This creates a magnetic field in the theta θ direction within the laminations. Because the coil is wrapped as a helix, there is a current component I_θ in the theta direction. The cross-sectional area Area through which the magnetic flux flows times a normal vector n in the theta direction is the tensor area n_θ Area. If the radius of the coil is r, then the curvature K is $1/r^2$ pointing in the radial direction. Because there are two coils with differing radii, the generator has two curvatures associated with it.

[0026] 8. In the geometry of electromagnetism, the magnetic field is part of the electromagnetic Faraday tensor $F^{\mu\nu}$ which is a 4 by 4 spacetime matrix having rows and columns of time t, radius r, angle θ and height z in cylindrical coordinates. The first index μ refers to the row, and the second index v refers to the column. The diagonal of the matrix is zero. The first row and column belong to the electric field. All the other slots are filled by the components of the magnetic field. The B_θ component is located in the complementary slots of r and z as shown in **FIG. 6**. Now a tensor product can be written with the available variables I^θ no Area $K_r F^{rz}=k^z$. This says that the current around the loop in the theta direction times the lamination area vector in the theta direction times the curvature K of the coil in the radial direction times the magnetic field in the theta direction is equal to a spring constant in the z-direction, or normal to the plane of the coil. All the tensor components cancel out except for the z-direction. That is, the coil produces a spacetime spring constant through the center of the coil where there are resonant bucking electric fields. So the coil is creating a magnetic spring. The units of the spring constant are shown in **FIG. 7** as force per meter.

[0027] 9. Because there are two coils operating in opposite directions in regions of differing curvature, there are two parallel spring constants generated along the centerline. Two springs in parallel sum, but the inner coil is negative due to the triple product of negative signs of current, field and area. Looking at the front view of the coils, the right hand rule shows the field going counterclockwise with the thumb pointing along the electric field in the positive z-direction. The inner coil with the field going in the clockwise direction has the electric field in the negative z-direction. Therefore the outer coil's positive spring constant sums with the inner coil's negative spring constant to produce an overall negative spring constant as shown in **FIG. 8**. The inner coil has a radius r, and the spacing between the outer and inner coil is a. In this design, the inner radius is 1, and the outer radius is 3 which is the magic ratio in physics of 1/3. That makes length a equal to 2. Substituting r=1 and a=2 into the spring constant equation shows that the ratio is negative 8/9. That is, the spring constant is negative as previously asserted. If the spring constant is negative, it must produce negative mass in order to have a real frequency of resonance. Since it produces negative mass, then it can produce a wormhole as shown by Dr. Kip Thorne.

[0028] 10. The geometry of hyperspace physics is based on the geometry of the tetrahedron which is circumscribed by the sphere. The corners of the tetrahedron touch the sphere at a latitude of 19.47122063° which turns out to be, in terms of planet cosmology, the location where all the large volcanoes and vortices occur on Earth, Mars, Jupiter, Uranus and Neptune. Furthermore, the cosine squared of this angle is 8/9, which is the spring constant ratio for the magnetic vortex generator. That is, the coil is interacting with the geometry of spacetime which is why it is such an effective wormhole generator. As will be demonstrated later, the tetrahedral geometry of hyperspace shows that the electron and proton are one and the same particle. This is a new discovery in science coming out of this research. The diagram also shows that when the speed of light is reduced, due to the low density of energy coming into our dimension through the wormhole, Planck's constant divided by the speed of light puts the electron at the boundary between space and hyperspace. That is, the electron and the proton go out of dimension which is the reason that the car tire rim can become merged with the tree. The ramifications of this magnetic vortex generator are enormous. It means developing new types of crystals through interdimensional merging which will be needed in the development of these spacecraft and their crystal rotors. It means the possibility of teleporting spacecraft through hyperspace over distances of light years using the new folding space waveguides. It means light-weight or inertia-less spacecraft which can be accelerated at hundreds of thousands of g's. It means the ability to lift extraordinary loads in construction work. It means surgery-less medical tables where tumors can be pulled directly out of the patient's body.

[**0029**] 11. Because we have a magnetic spring, then there is some resonant frequency at which the coil can be operated in order to create a large spacetime distortion. It is this distortion that creates the wormhole into hyperspace. From this it follows that the wormhole attaches to a low pressure region with a magnetic monopole. Because pressure is linear mass times the speed of light squared divided by area, a low pressure signifies a low speed of light and less dense matter. What this means is that opposite polarity spacetime curvature at two ends of an electrode will produce a voltage along the electrode, effectively creating a power supply. The reason for this is that spacetime curvature, as shown by Einstein's General Theory of Relativity equation $G_{\alpha\beta}=8\pi T_{\alpha\beta}$, is equal to the square of the electromagnetic fields in the stress-energy tensor T. So the curvature on the ends of the electrode appear as voltages and the system acts as a battery supply to power the spacecraft.

[**0030**] 12. The other characteristic of the magnetic vortex generator is that it can pull objects out of dimension allowing the object to apparently move through solid walls. The object doesn't actually move through the wall because the object is not in the same dimension as the wall; it just appears that way. Moving out of dimension and then back again on the other side of the wall would effectively move the object through the wall for all intents and purposes.

[**0031**] 13. The tetrahedron diagram is a physics diagram which plots the natural logarithm of mass to the natural logarithm of wavelength. The product of mass times wavelength is equal to Planck's constant divided by the speed of light. So the product of the electron mass times its wavelength is equal to the proton mass times its wavelength is equal to the Planck mass times the Planck wavelength. The Planck wavelength is the bottom dimensional limit of the universe. We live in the Planck box which is bounded by the Planck wavelength and Planck mass. Outside this box is hyperspace.

[**0032**] 14. Because logarithms sum, the sum of the log of the mass plus the log of the wavelength is a constant sum. Thus the electron and proton and Planck mass slide on a 45° line known as the base constant which is equal at the axes to the log of Planck's constant divided by the speed of light. Planck's constant is measured in joule-sec so that multiplying it by the frequency 1/second of light gives the energy of the photon particle.

[**0033**] 15. Planck's constant is equal to the Planck wavelength times the Planck mass times the speed of light. The 45° base constant is Planck's constant divided by the speed of light, which means that the speed of light cancels out top and bottom, leaving the area of the Planck box as the value of the base constant. Hyperspace has a low linear mass compared to our dimension. Therefore, Planck's constant is reduced when this energy enters our dimension through the wormhole created by the generator. And the base constant is also so reduced. In terms of logs, this means that the 45° base line becomes more negative and moves to the right on the diagram. As it does so, the base line intersects the electron at the Planck wavelength which is the separation point between space and hyperspace. That is, the electron moves out of dimension. Because the electron and proton are one and the same particle, as shown in reference tetrahedron diagram tet0565, the proton and hence the entire atom is taken out of dimension as well.

[**0034**] 16. All of this can be seen more easily graphically on the tetrahedron diagram itself, referring to reference tetrahedron diagram tet3025. The 45° line which intersects the electron at point (b) is the base constant for our dimension. As you can see, this base line intersects the horizontal axis at a value of −95.91546344 which is the log of Planck's constant h divided by the speed of light. Because Planck's constant is proportional to the linear mass, it is reduced in value by the low density hyperspace energy and, in terms of logs, becomes more negative. This moves the base line to the right at a value of around minus **105**. The new base line intersects the electron at point (a) which is located on the Planck wavelength that is the boundary between space and hyperspace. Thus the electron at point (a) goes out of dimension.

[**0035**] 17. It was inferred previously that a negative spring constant meant a negative mass was produced by the wormhole generator. This can actually be calculated using Einstein's General Theory of Relativity. The calculation starts with the g metric tensor which is a spacetime measurement of distance in terms of time t, radius r, horizontal angle θ and length z. This 4 by 4 matrix is shown in **FIG. 9** where the diagonal line has a signature in cylindrical coordinates of values equal to $\{-1, 1, r^2, 1\}$. All the other terms of the matrix are zero except for the magnetic fields in the two coils. Because the field is changing sinusoidally with time in the theta direction, the field has to go into the $\{t, \theta\}$ and $\{\theta, t\}$ slots of the matrix. Because the field in the inner coil is in the negative direction compared to the outer coil, and including a 90° phase shift between the two fields, a suitable magnetic field function would be $B(\cos(\theta)-\sin(\theta))$.

[**0036**] 18. Using a general relativity software package, Einstein's G curvature tensor can be calculated for this particular metric. The first term G t in the upper left hand slot in the corner is the mass term for the tensor. All the other terms are either electromagnetic fluxes or pressure terms involving the squares of the fields.

[**0037**] 19. Referring to **FIG. 10, a** plot of the mass as a function around a small circle shows that two negative mass spikes occur around the circumference of the circle. Because the radius is almost zero, the two spikes are actually coincident.

[**0038**] 20. This next section shows the electrical system used to drive the magnetic vortex generator. Referring to **FIG. 11**, the system is driven by a sinusoidal voltage source (A) into a 1:1 turns isolation transformer (B). The voltage is stepped up into the range of thousands of volts using a step-up hi-pot transformer (C). The first loop consists of a direct current blocking capacitor (D), a transformer choke (E), a variable inductance (G) and the current in the loop (F). The variable inductance is a coil winding on a toroidal core wound with thousandth-inch thick SuperMalloy tape. This coil acts as a magnetic switch due to its variable permeability.

[0039] 21. Referring to the accompanying drawing in **FIG. 12**, the slope of the BH curve is actually the permeability of the core. At point (a) on the curve, the slope is very low and therefore the permeability is very small. The resistance of the coil is the frequency of the current times the inductance of the coil. If the permeability is small, then the inductance is low, which means that the coil resistance is low initially. With a low resistance in the coil, current (F) flows through the winding rather easily. Then the coil goes from point (a) to point (b) where the permeability and resistance increase. This change in resistance from a low to a high value dumps the magnetic energy into capacitor (H). Then the non-linear coil saturates between points (b) and (c) where again the coil has a small slope and the coil resistance switches to a low value. Capacitor (D) then dumps its charge (I) through coil (G) producing a large voltage spike in the input and output winding of transformer toroidal coil (J). The magnetic flux in coil (J) then produces a voltage spike in coils (L) and (M) of the magnetic vortex generator. The frequency of oscillation of the generator is determined by capacitor (K) and the overall inductance of the two coils. The diagram shows that the winding is non-linear and in opposite directions going from outer coil to the inner coil.

[0040] 22. Frequency of oscillation has to be kept under 20 MHz in order to create a soft wormhole that connects to low pressure regions of hyperspace. The pressure regions of hyperspace are similar to the pressure produced by a dam holding water. The upper surface of the water, where there is no water pressure, is analogous to the black void into which our universe is expanding. In the middle of the dam, there is a region just above us having a lower water pressure which corresponds to the low density hyperspace energy. This analogy is not perfect because the many frequencies of hyperspace are quantized, as we know from quantum physics, rather than being a continuous spectrum of lower and lower frequencies down to the zero frequency of the black void.

What I claim as my invention is

1. A magnetic vortex generator which can generate negative mass according to Einstein's General Theory of Relativity, which is a tested and proven theory. As a result of this theory, it can be shown that negative mass is required to create a stable wormhole between space and hyperspace. Without the negative mass, the throat would close.

2. The generator consists of two solenoids wound with a common wire in opposite directions on two separate toroidal curved-sheet transformer laminates of differing radii. The smaller solenoid is mounted along the centerline of the larger solenoid.

3. According to Maxwell's equations, said geometrical and magnetic arrangement produces linear bucking electric fields along the centerline of said coils. Since curvature and the square of the electromagnetic fields are one and the same phenomenon according to Einstein's spacetime curvature tensor $G=8\pi\ T$, this resonance of the electric fields causes a resonance of the spacetime curvature and the opening of the wormhole.

4. Due to the fact that the magnetic flux travels within the laminations at a curvature equal to the inverse of the radius squared, each coil produces a spring constant which depends on the current, lamination area, magnetic field strength and the individual curvature of each coil.

5. Due to the fact that the flux travels in opposite directions in each solenoid, the spring constant of the outer coil is positive, and the spring constant of the inner coil is negative. These two spring constants, one positive and one negative, add in parallel to create a negative spring constant for the generator. Because the frequency of resonance is equal to the square root of the negative spring constant divided by the negative mass, the resonant frequency is positive real.

6. This process of creating and keeping open the wormhole allows low density hyperspace energy to enter our dimension. Because the linear mass is lower, Planck's constant, equal to the Planck mass times the Planck wavelength times the speed of light, is reduced to such an extent that the electron is moved out of dimension. Because the electron and proton are one and the same particle, when considering a path through space and hyperspace, the proton is also moved out of dimension. Thus this process of moving the atom in and out of dimension using the magnetic vortex generator has the ability to create cold-welded crystals, new types of materials and new ways to cold solder one disparate material to another.

7. As a consequence of said process, highly relativistic electromagnetic fields can be created because the velocity of light has been reduced considerably. These fields than can be used to produce life in the new class of electromagnetic propulsion vehicles.

8. As a consequence of said process, refrigeration systems can be created because negative energy cools rather than heats.

9. As a consequence of said process, power supplies can be made such that a differential spacetime curvature on the ends of a carbon electrode can create a differential voltage similar to a regular chemical battery.

10. As a consequence of this process, surgery-less medical tables can be fabricated whereby tumors can be pulled directly from the patient's body.

11. As a consequence of this process, the negative mass produced by the generator can offset the mass of a spacecraft to create a mass-less, inertia-less vehicle which can accelerate at hundreds of thousands of g's.

12. As a consequence of this process, the folding space waveguide becomes a reality whereby hyperspace energy, which has a low spring constant, can be easily folded and curved using powerful, relativistic electromagnetic fields. Spacecraft will be able to teleport themselves out-of-dimension over huge distances measured in terms of light-years.

13. Because the wormhole opens up an interdimensional connection to hyperspace having a magnetic monopole, a radial magnetic field is created. As a consequence of this process, a changing magnetic monopole field crossed with a changing electric field can produce a toroidal electromagnetic flux around the circular hull of the spacecraft. As a consequence of the merger of these two fields, a spacetime curvature G_{zz} is produced over the hull of the spacecraft which creates a tension or lift force which enables the spacecraft to ascend, hover or descend.

14. An electrical system, comprised of a variable frequency generator and amplifier, an isolation transformer and voltage step-up transformer, drives a non-linear inductance which switches on and off the current in the output circuit in such a manner as to produce large voltage spikes through magnetic vortex generator. The generator responds by producing two large negative mass spikes close to the centerline of the two coils.

15. The winding on each coil of the generator has a non-linear spacing to enhance the magnetic field and to reduce the interwinding coil capacitance.

16. The ratio of the radius of the small coil to that of the larger coil is 1/3, which is the magic ratio in physics. This creates a spring constant that is proportional to 8/9. The square of the cosine of the tetrahedral angle of 19.47° is equal to this ratio. Also the ratio of the area-to-volume ratio of the circumscribing sphere of a tetrahedron to the area-to-volume ratio of the tetrahedron is also 1/3. And the corners of the tetrahedron touch the circumscribing sphere at 19.47°. All the large volcanoes and vortices on Earth, Mars, Jupiter, Uranus and Neptune are located at this latitude. Thus this invention is more effective in developing a wormhole because it is tuned geometrically to the tetrahedral geometry of space.

* * * * *

US 20060071122A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0071122 A1**

St. Clair (43) **Pub. Date: Apr. 6, 2006**

(54) FULL BODY TELEPORTATION SYSTEM

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
4A
52 KINGS COURT
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/953,212**

(22) Filed: **Sep. 29, 2004**

Publication Classification

(51) **Int. Cl.**
B64C 39/00 (2006.01)

(52) **U.S. Cl.** **244/62**

(57) ABSTRACT

A pulsed gravitational wave wormhole generator system that teleports a human being through hyperspace from one location to another.

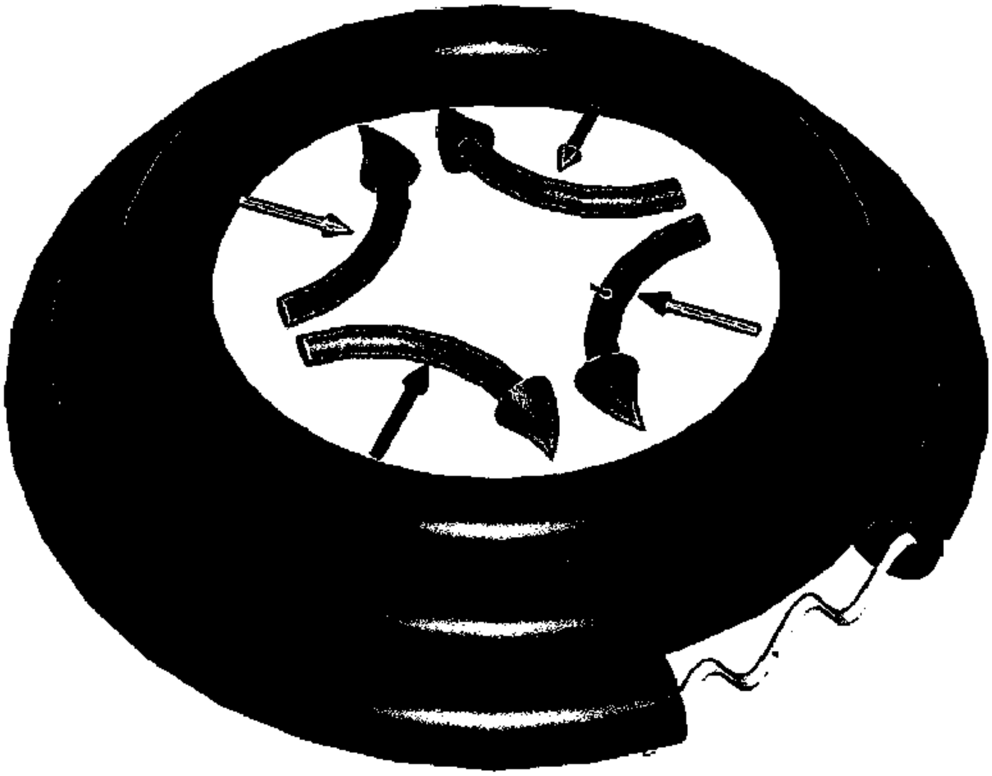

Figure 1

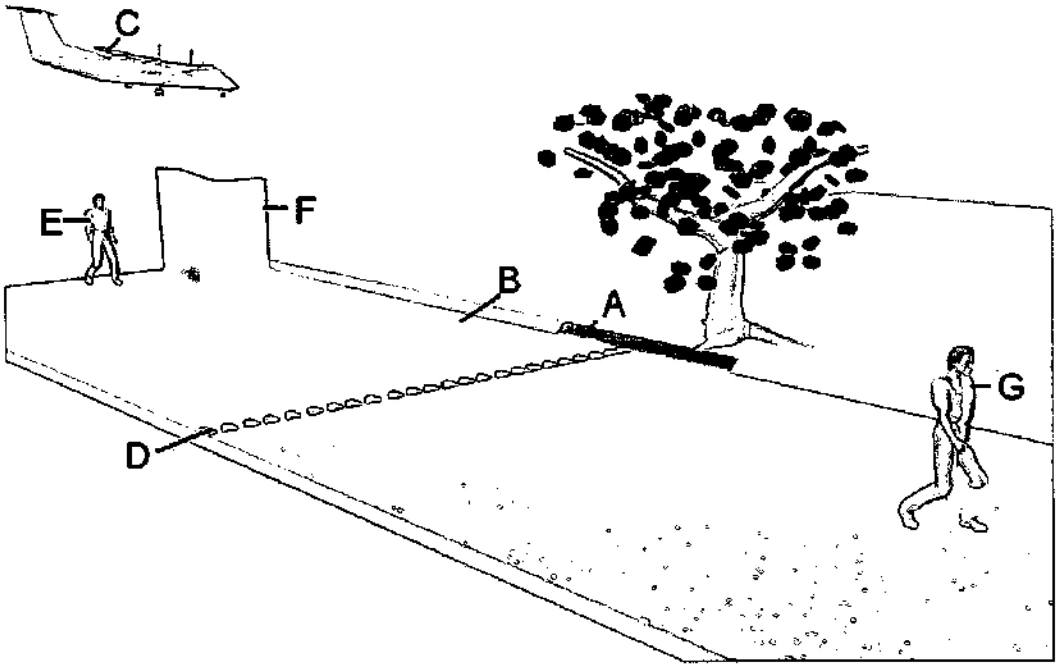

Figure 2

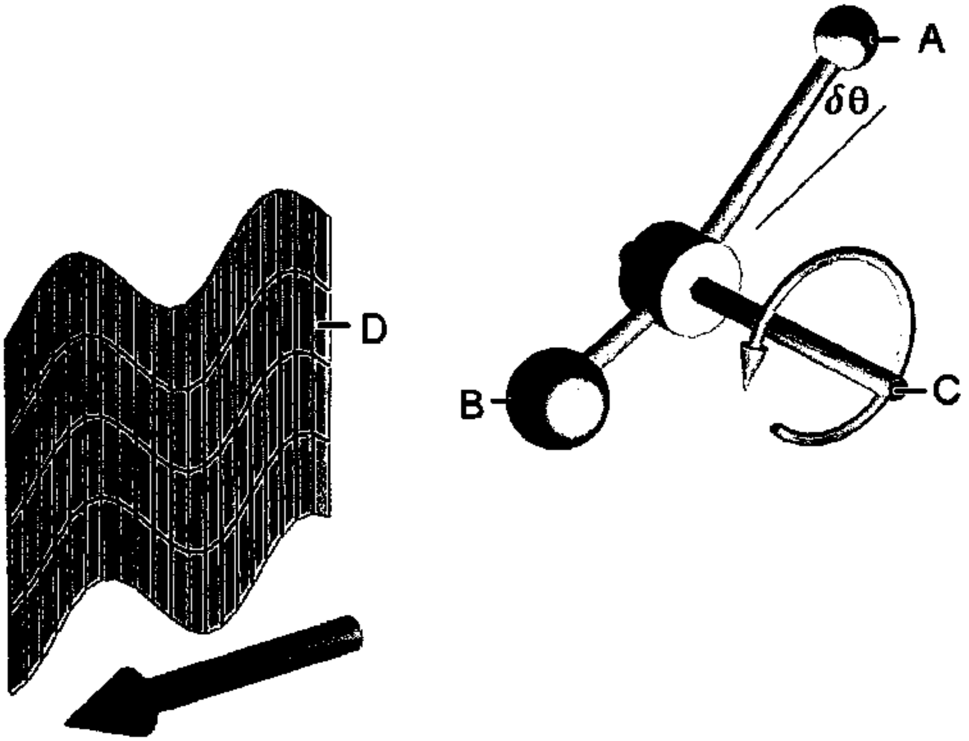

Figure 3

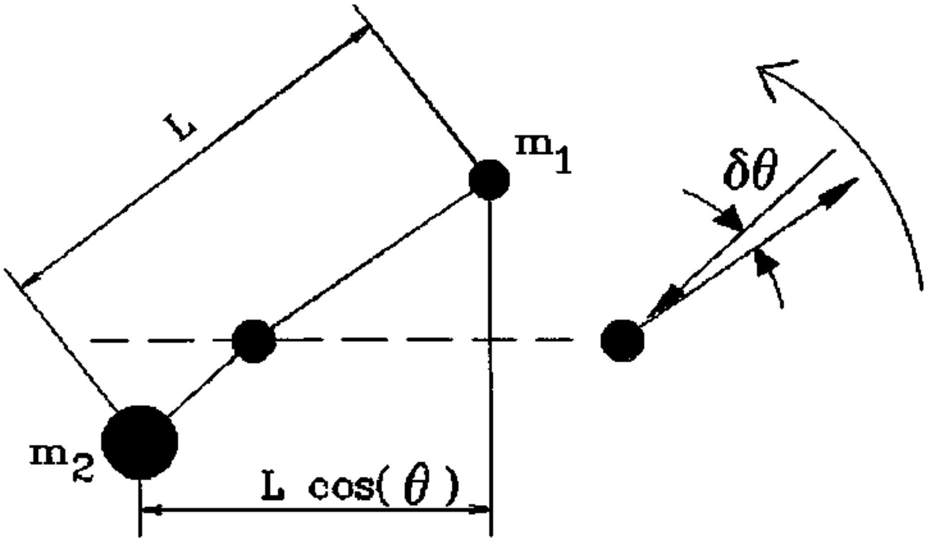
L
m_1
$\delta\theta$
m_2
L cos(θ)

Figure 4

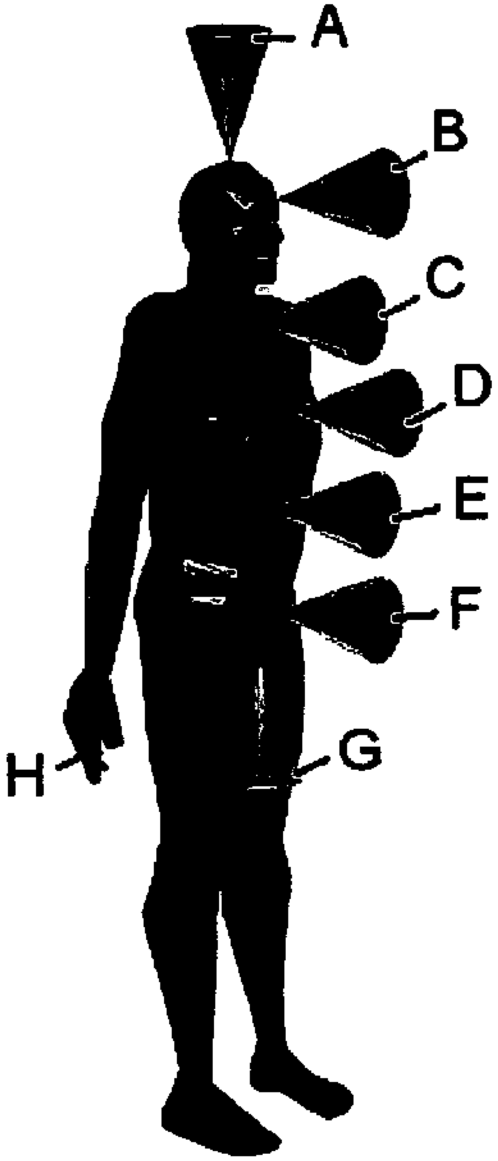

Figure 5

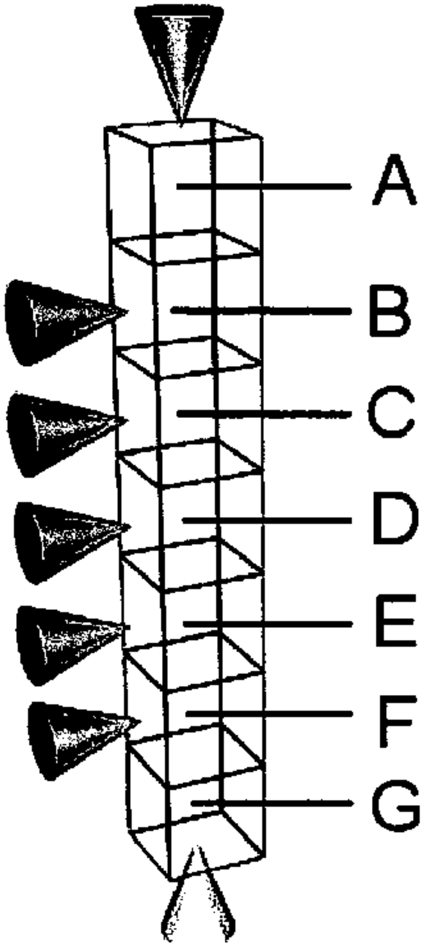

Figure 6

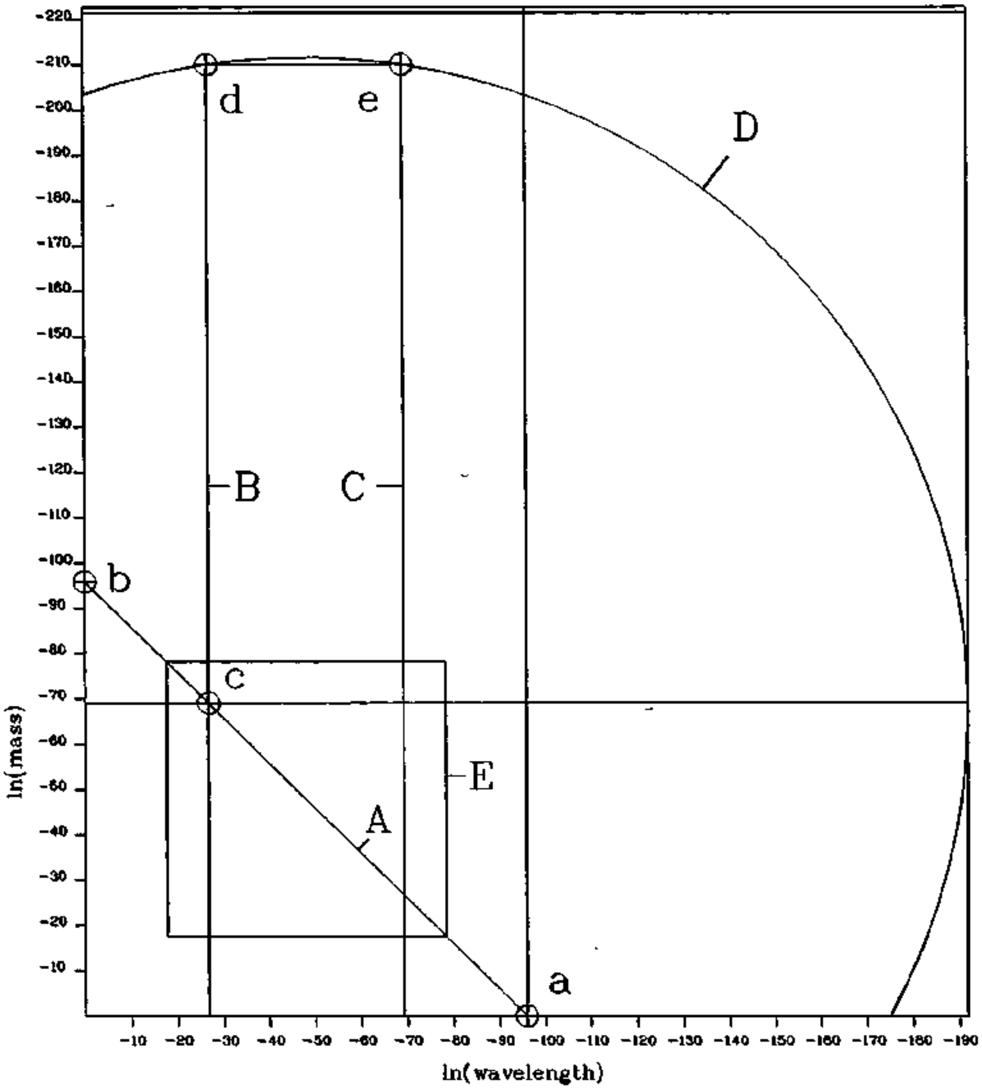
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
a
b
c
d
e
A
B
C
D
E

Figure 7

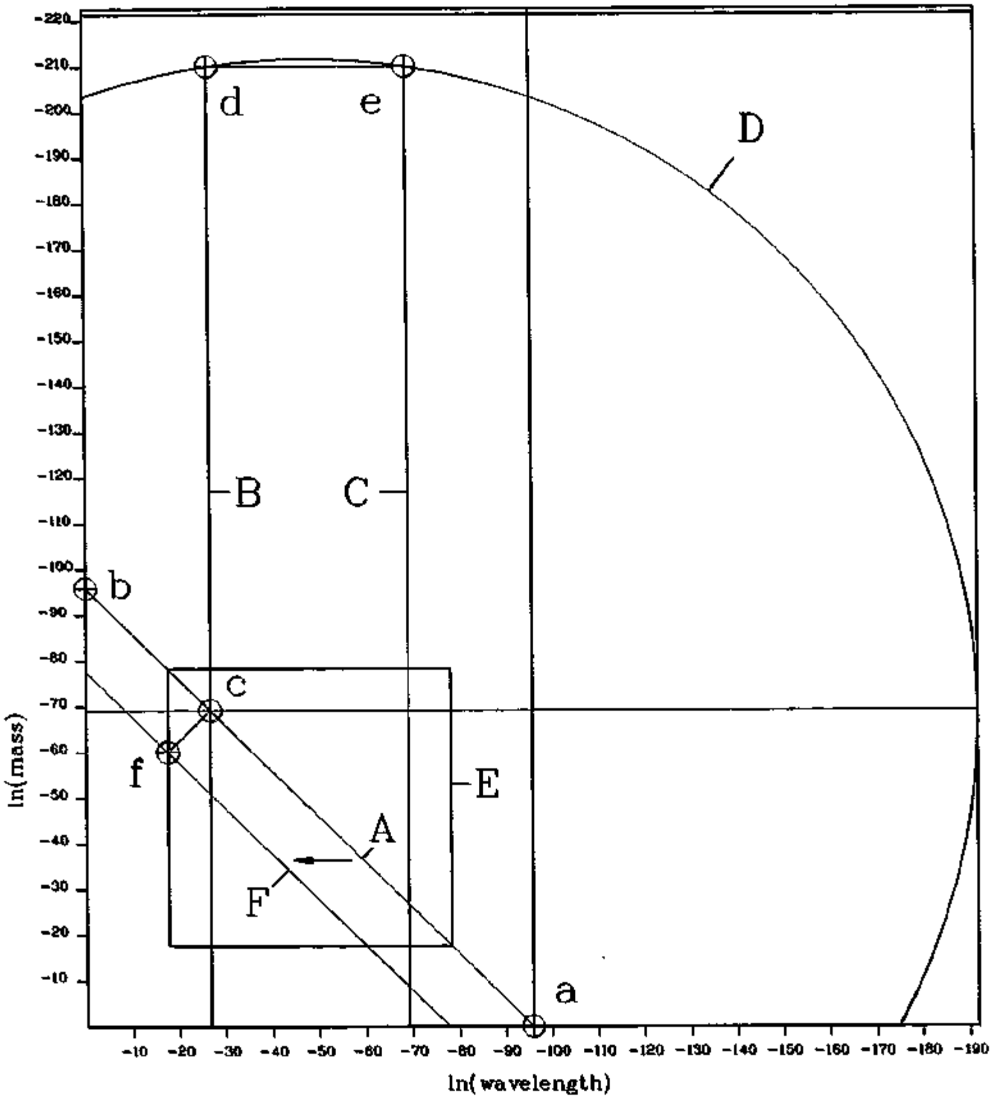
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
d
e
D
B
C
b
c
f
E
A
F
a

Figure 8

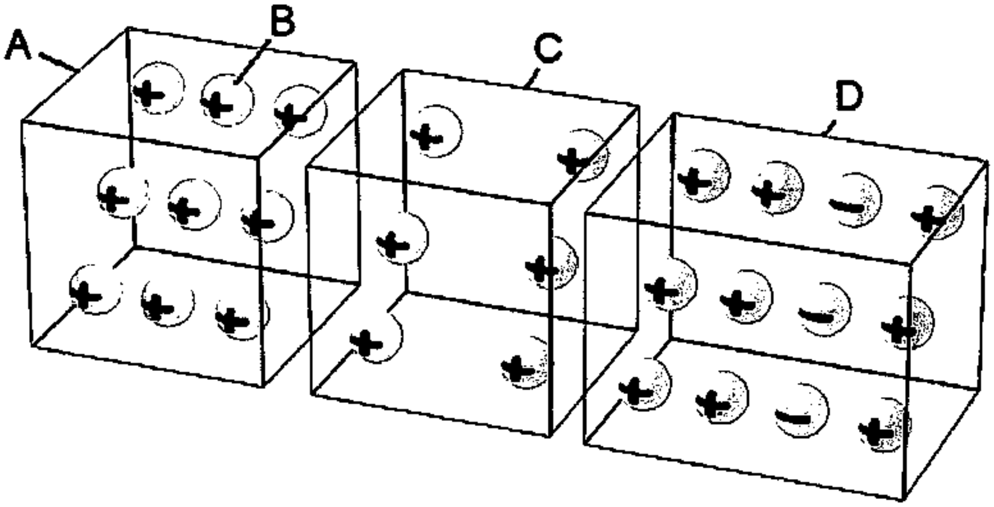

Figure 9

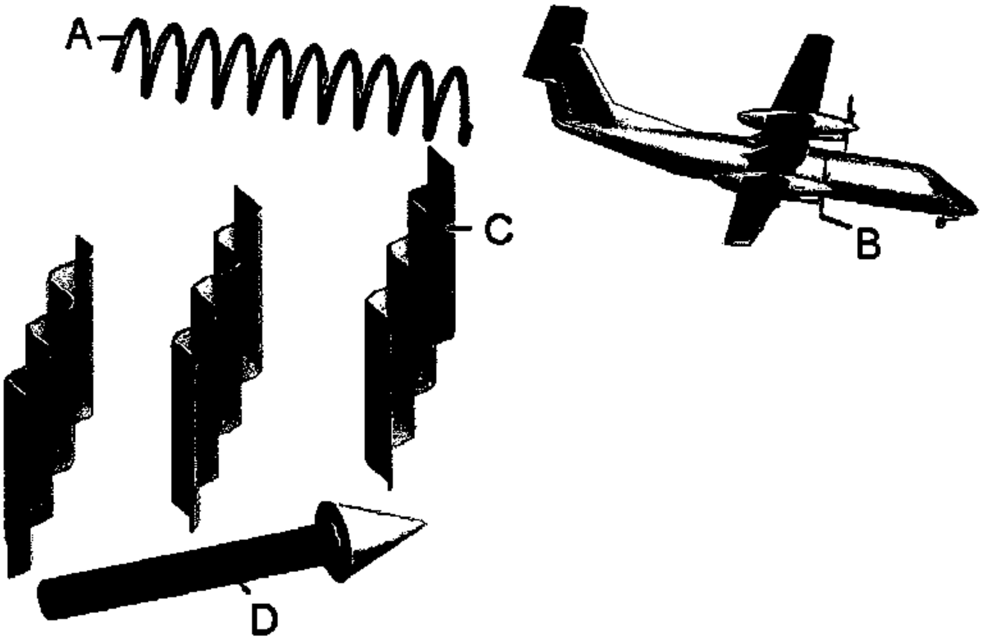

Figure 10

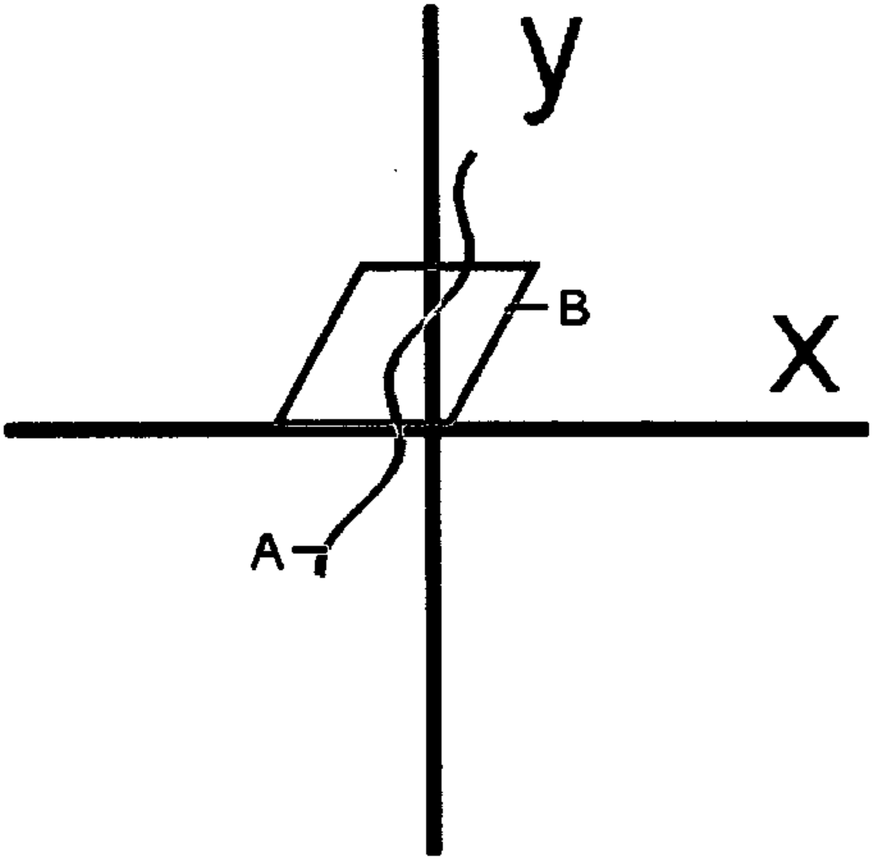

Figure 11

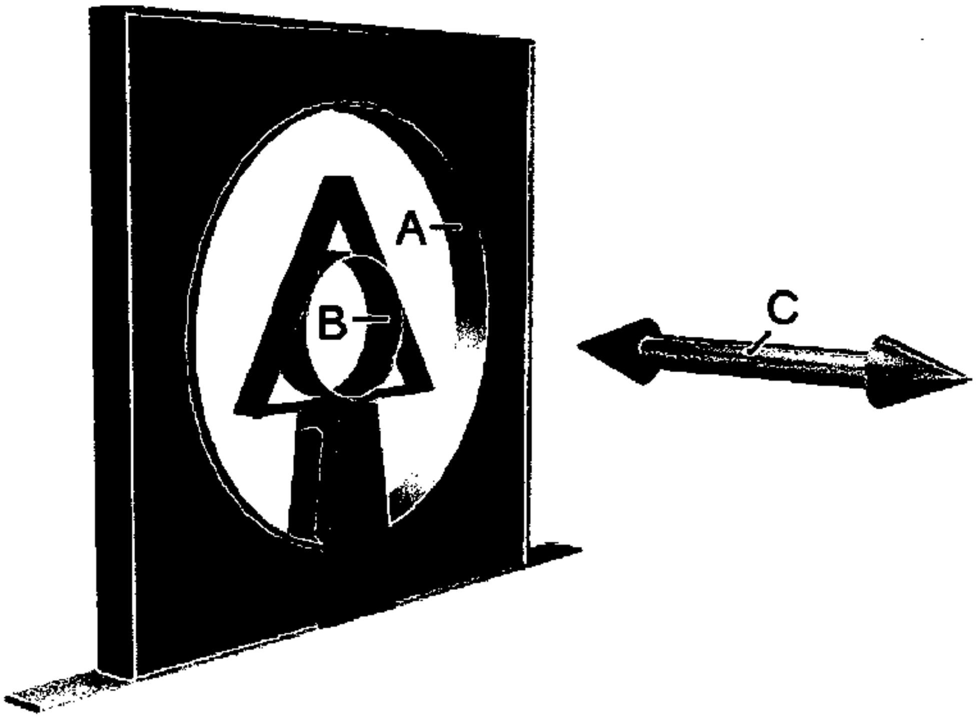

Figure 12

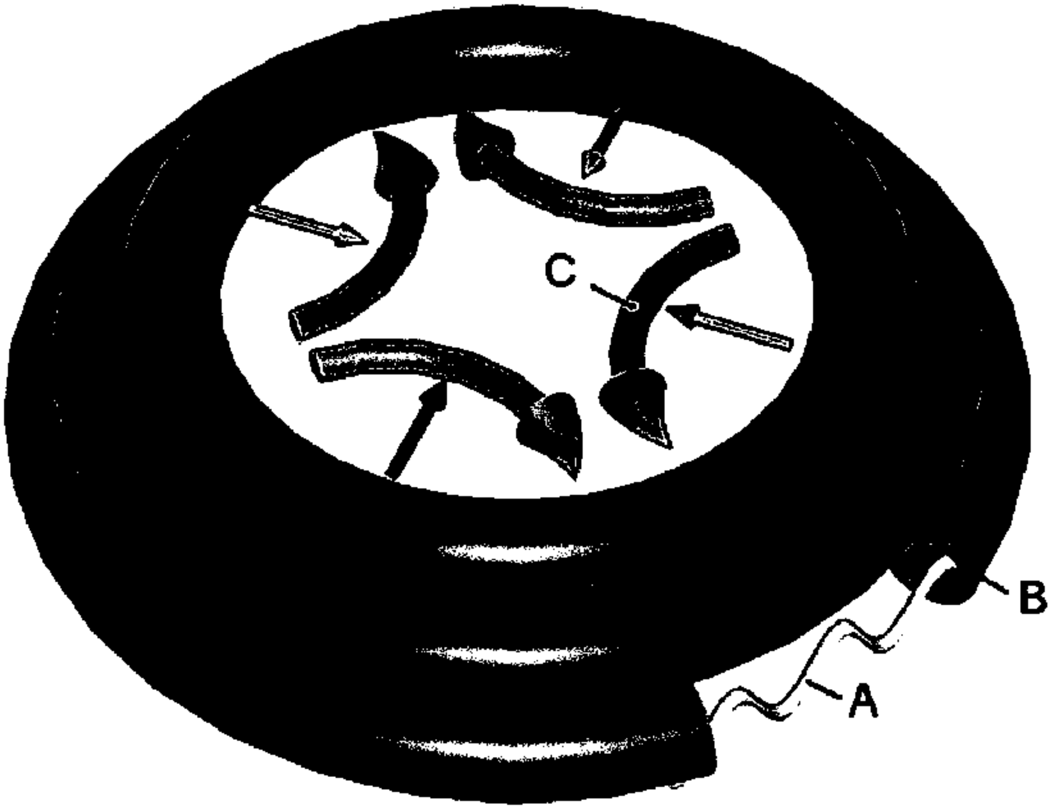

Figure 13

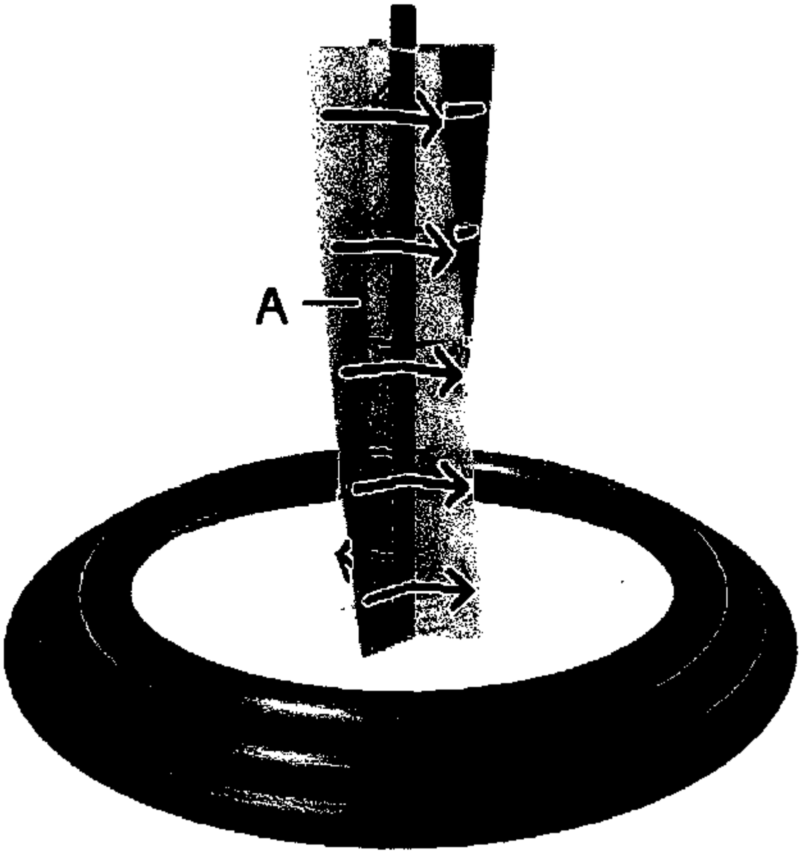

Figure 14

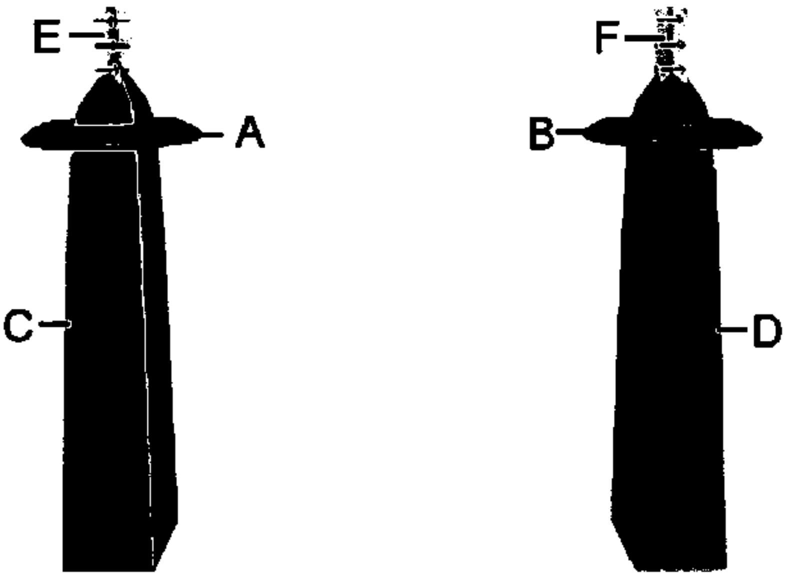

Figure 15

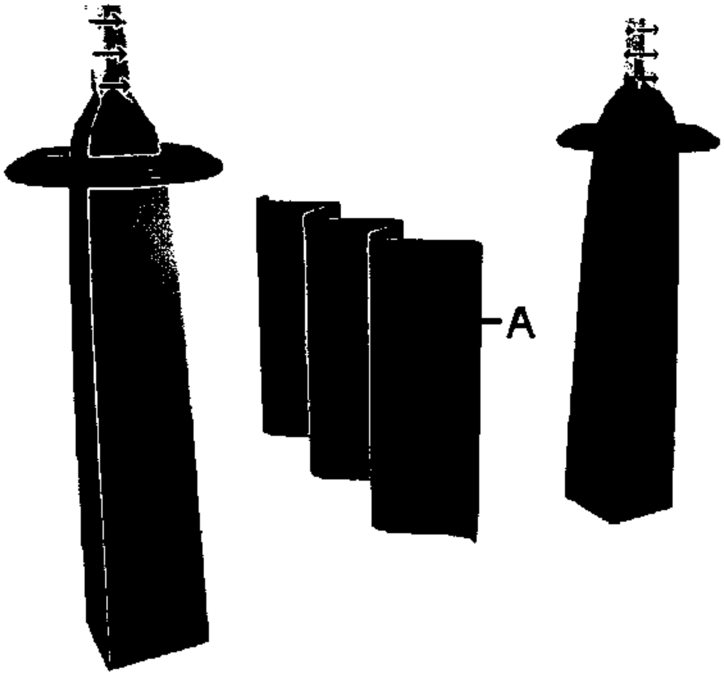

Figure 16

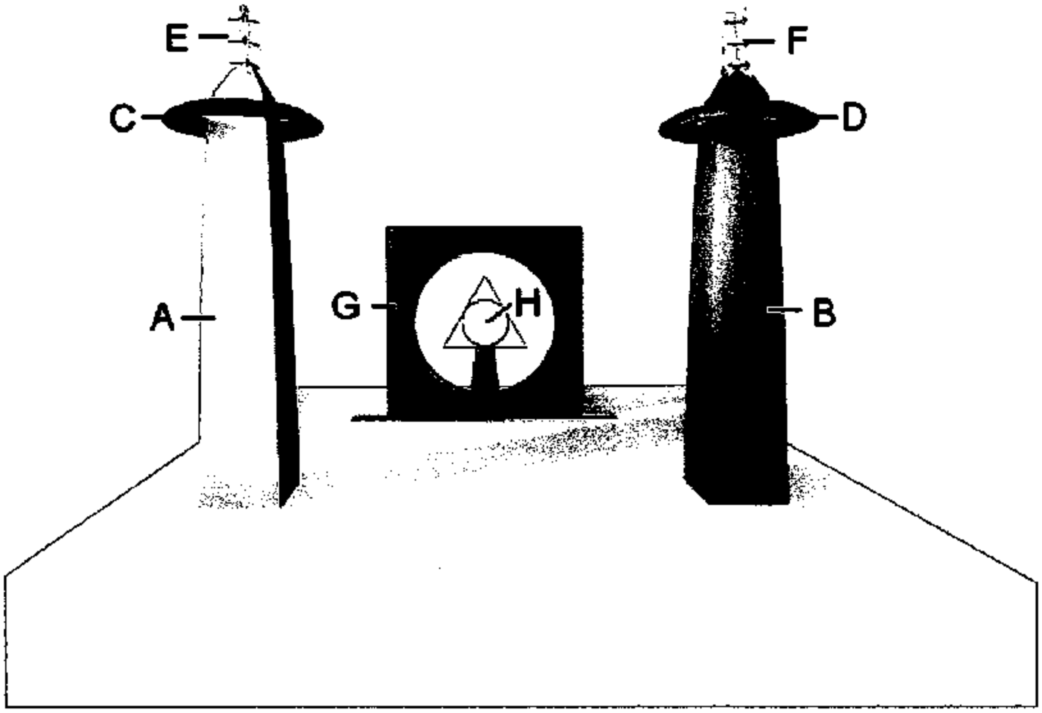

FULL BODY TELEPORTATION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a system that teleports a human being through hyperspace from one location to another using a pulsed gravitational wave traveling through hyperspace.

BACKGROUND OF THE INVENTION

[0002] The basis for this invention is an event, referring to **FIG. 1**, occurring on May 2, 2004, in which the inventor ("he") personally experienced a full-body teleportation while walking to the bus stop (A) along a road (B) that runs perpendicular to the nearby commercial airport runways where planes are landing. There is a wide iron grating (D) for water drainage that crosses the road at the center of the bus stop. The grating width is such that one has to make a concerted effort to jump across it in order to get from one side to the other. Approximately 50 meters from the iron grating, he (E) felt a vertical wave (F), similar to a flag waving in the breeze, traveling down the street toward the bus stop. The wave velocity was about 1 meter per second, which was slightly faster than his walking speed. In the next instance, he (G) found himself down the street near the corner of the next block. Realizing that he had passed the bus stop, he turned around to see the iron grating approximately 50 meters up the street in back of him. Because there was no recollection of having jumped across the iron grating nor of having passed the bus stop's yellow marker line, he realized that he had been teleported a distance of 100 meters while moving along with the traveling wave. It was obvious that the wave was pulsed because the front edge overtook the inventor, moved with him momentarily, and then the back edge of wave left him as it moved on down the street. While contemplating this sequence of events, he then looked up and saw in a span of a few seconds a twin-turboprop airplane (C) in the distance crossing above the road while making a shallow descent in order to land at the airport.

[0003] It took a number of days in order to understand this sequence of events. The explanation involves knowledge of a wide range of subjects such as gravitation physics, hyperspace physics, wormhole electromagnetic theory and experimentation, quantum physics, and the nature of the human energy field.

[0004] It is obvious from the above scenario that the airplane momentarily crossing perpendicular to the road generates the aforementioned pulse. Because the airplane has an engine on each wing, there are two propellers which conceivably are rotating out-of-phase with each other. That is, the blade of one propeller could be pointing up and the equivalent blade on the other engine could be pointing in a slightly different direction. Notice that the tip of the blade traces out a helix as the plane is landing.

[0005] In gravitation physics, referring to **FIG. 2**, it is known that two masses of mass m1 and m2 (A,B) attached by lever arms slightly offset by an angle δθ along the radial direction to the rotating shaft (C), will produce a gravitational wave (D) traveling perpendicular to the shaft. The mass and wave are referred to as the source and receptor respectively. Referring to a side view looking along the shaft **FIG. 3**, the product of the mass m times the angular acceleration a is a constant such that **m1*a*1** is equal to **m2*a*2**. The distance between the masses is length L, which makes an angle θ with the horizontal axis. The difference in time of travel to the receptor gives rise to a difference in phase δθ equal to the angular velocity ω of the rotating shaft times the length L times the cosine of the angle θ

$$\delta\theta=\omega L\cos(\theta)$$

[0006] At the receptor, the amplitude of the wave is equal to the mass times the acceleration times the phase difference divided by the radius r to the receptor

$$A=\frac{m_1a_1}{r}\delta\theta\approx\left(\frac{m\omega L\sin(\theta)}{r}\right)(\omega L\cos(\theta))=\frac{m\omega^2L^2\sin(2\theta)}{r}$$

Even though the turboprop airplane engines have a high rotational speed and a large separation distance between masses, the gravitational wave which is produced is small and not noticed. The problem is that the gravitational constant G in this dimension has such a small value equal to the speed of light c squared divided by the linear mass Ω of the universe

$$G=\frac{c^2}{\Omega}=\frac{(299792458\text{ m/s})^2}{1.346812891\cdot10^{27}\text{ kg/m}}=6.673200002\cdot10^{-11}\ \frac{\text{m}^3}{\text{kgs}^2}$$

[0007] On the other hand, a gravitational wave traveling in hyperspace would be magnified enormously due to the face that the linear mass is so small. The magnitude of the gravitational constant in hyperspace can be estimated in the following manner. At the beginning of the 20th century, a man's parents were dying of tuberculosis. With their permission, he placed them and their beds on weighing scales. When each one passed away, each scale registered a drop in mass equal to 0.071 kilograms. This is the mass of the hyperspace energy being which resides in the physical body. Because hyperspace is co-dimensional with our dimension, the energy being interpenetrates the body and controls its movement.

[0008] Referring to **FIG. 4**, a human being has seven vortices (A through G) which are aligned along the centerline of the body. Each vortex is actually a co-gravitational field K which causes a pendulum placed in the field to spin in circles. For this reason, the K field has units of inverse seconds similar to an angular velocity. The vortex transports energy from our dimension to the energy being located in hyperspace. The gravitational field g and the co-gravitational field K are equivalent gravitationally to the electric E field and the magnetic B field found in electromagnetism. The equivalent gravitational solution to an electromagnetic problem can be obtained by substituting the following gravitational constants for the electromagnetic constants

Electromagnetic	Gravitational
q (charge)	m (mass)
ρ (volume charge density)	ρ (volume mass density)
σ (surface charge density)	σ (surface mass density)
λ (line charge density)	λ (line mass density)
J (convection current density)	J (mass current density)

-continued

Electromagnetic	Gravitational
E (electric field)	g (gravitational field)
B (magnetic field)	K (co-gravitational field)
ϵ_0 (permittivity of space)	$-\frac{1}{4}\pi G$
μ_0 (permeability of space)	$-4\pi G/c^2$
$-\frac{1}{4}\pi\epsilon_0$ or $-\mu_0 c^2/4\pi$	G (gravitational constant)

[0009] Referring to **FIG. 5**, each vortex is connected through the pineal gland by light cords to a separate hyperspace quantum well having its own frequency and dimension. The reason for this separation is that the conical spiritual eye, attached to vortex B, has to have its own energy structure which is different from vortex (A) which is connected to the quantum energy field in which the mental processes are developed. Due to the high speed of light in our dimension, the quantum wells are the size appropriate to molecules and atoms. In hyperspace, where the speed of light is one meter per second, the quantum wells are huge and can be manipulated. This manipulation has shown that the quantum wells are in the shape of a cube about a meter on a side. This makes the whole structure about seven meters tall for a volume of seven cubic meters. Thus the mass density ρ of hyperspace is about

$$\rho_h = \frac{.071\ \mathrm{kg}}{7\ \mathrm{m}^3} = .01\ \frac{\mathrm{kg}}{\mathrm{m}^3}$$

which per unit area is the same value. Therefore the hyperspace gravitational constant is equal to

$$G_h = \frac{c^2}{\Omega_h} = \frac{\left(1\ \frac{\mathrm{m}}{\mathrm{s}}\right)^2}{.01\ \mathrm{kg/m}} = 100\ \frac{\mathrm{m}^3}{\mathrm{kgs}^2}$$

The enormous magnification of the gravitational constant is therefore of the order of

$$\frac{G_h}{G} = \frac{100}{6.6732 \cdot 10^{-11}} \approx 1.5 \cdot 10^{12}$$

The question is how does this amplified gravitational wave created by the rotating propellers and turbines get into hyperspace from our dimension?

[0010] The answer comes from experiments done using the ancient Chinese form of breathing known as Chi Kung. Using this breathing technique, we have been able to levitate the human body over six feet in the air. The internal temperature of the stomach is around 200 degrees Fahrenheit. By simultaneously squeezing the diaphragm to bring hot air up through the lungs, and breathing through the nose to bring cold air down, rotating vortices are generated in the lung passages when these two air masses meet and twist around each other as depicted in the famous Yin-Yang diagram. Because the lung has variable diameter passages from the large diameter at the throat to the final small air sacs, there is a spectrum of rotating frequencies.

[0011] From quantum physics it is known that if there is a temperature fluctuation occurring among a group of harmonic oscillators in the environment, then Planck's reduced constant $\hbar$ is increased by the cotangent of the constant times the frequency ω of the oscillator divided by twice Boltzmann's constant k times the temperature T

$$\hbar = \hbar\coth\left(\frac{\hbar\omega_n}{2kT}\right)$$

[0012] The effect of increasing Planck's constant, referring to **FIG. 6**, can be seen in the tetrahedron diagram. This diagram, of which there are now over 4000, plots the natural logarithm of mass on the vertical axis versus the natural logarithm of wavelength on the horizontal axis. In terms of mathematics, it is a subspace logarithmic manifold which projects geometrically the physics constants into our **4**D spacetime dimension. That is, it is the geometry of the tetrahedron circumscribed by the sphere that determines the mass of the proton and electron. The mass of the electron times its wavelength is equal to the mass of the proton times its wavelength which in turn is equal to Planck's constant h divided by the speed of light c

$$m_e\lambda_e = m_p\lambda_p = \frac{h}{c}$$

Taking the natural logarithm of the above equation shows that the mass plus the wavelength is equal to what is termed the base constant

$$\ln(m_e) + \ln(\lambda_e) = \ln\left(\frac{h}{c}\right) = -95.91546344$$

which is represented in **FIG. 6** by the 45 degree line (A) from point (a) on the horizontal axis to the vertical axis at point (b). The electron is located at point (c) which is the intersection of the electron wavelength (B) with line (A). The electron wavelength (B) reflects off the sphere (D) at points (d) and (e) and returns along line (C) as the electron mass. As shown in tetrahedron diagram tet0565, stored in the Library of Congress, the clockwise path of the electron transitions into the counter-clockwise path of the proton showing that the electron and proton are one and the same particle. Because the electron and proton travel in opposite directions along the path, they have the same charge but of opposite sign.

[0013] Our dimension is represented by Planck box (E) which is bounded by the Planck mass and the Planck wavelength. The Planck mass is equal to the linear mass of the universe times the Planck length which is the bottom limit of our dimension. The Planck wavelength is 2π times the Planck length. Notice that the electron is located within the Planck box.

[0014] Referring to **FIG. 7**, if there is an increase in Planck's constant due to the temperature fluctuations among the harmonic oscillators, the 45 degree base line (A) moves to the left on the tetrahedron diagram as shown by line (F).

Because of the increase in the base constant, there is a corresponding increase in the electron mass and wavelength. The electron moves from point (c) to point (f) which places it at the edge of the Planck box (F) which is the boundary between space and hyperspace. At point (f), the electron is essentially no longer in our dimension.

[0015] Referring to **FIG. 8**, imagine a box (A) filled with nine electron oscillators (B). If Planck's constant is increased near the three oscillators in the middle, these electrons will leave this dimension. This leaves six oscillators as shown in the box (C). However, box (C) is the equivalent of box (D) in which there are still nine positive mass oscillators together with 3 negative mass oscillators. Thus there is an accumulation of negative energy ($-\rho$) when information is lost from the environment to another dimension.

[0016] Dr. Kip Thorne, who co-authored the book *Gravitation* with Dr. Archibald Wheeler of Princeton University, has shown in a General Relativity spacetime curvature calculation that negative energy is required to open and stabilize the throat of a wormhole between space and hyperspace. The accumulation of negative energy in the aforementioned example generates wormholes between into hyperspace. Hyperspace has a low energy density because of the reduced speed of light in that dimension. Ordinarily, energy would not flow from hyperspace to space because space has a higher potential than the potential of hyperspace. This, of course, is the reason that the body vortices can flow energy into the energy field of the human being who is located in hyperspace. By creating negative energy, the potential becomes reversed such that low density hyperspace energy flows into our dimension as seen by the positive head

$$PE=\rho_{hyperspace}-(-\rho_{space})=+2\rho$$

The low-density energy fills the body which allows a human being to float upwards like a helium balloon as verified by Chi Kung breathing as well as spinning on a motorized platform known as the Chakra Vortex Accelerator. The latter device resulted in the first mechanical means to produce anti-gravity.

[0017] The process of creating spinning thermal fluctuations is the same as found in the hot air vortices created by the jet airplanes landing at the airport near the road where the full-body teleportation occurred. Large vortices are created over the wing of the airplane at the same time that the turbine engines are spinning hot vortices into relatively cold air. These conditions produce wormholes between space and hyperspace. It takes a twin turboprop airplane landing behind the jet to generate the gravitational wave in the region where the wormholes have formed. The gravitational wave then traverses the wormholes into hyperspace, becoming highly amplified due to the change in linear mass and speed of light. Because the propeller blades are co-linear with the road, the gravitational wave travels in the direction along the road where it was encountered by the inventor.

[0018] From experiments with cavitating bubbles (see patent application Cavitating Oil Hyperspace Energy Generator), it was found that it is possible to produce a wormhole if the surfaces of the bubble collapse asymmetrically. A symmetric collapse of a spherical bubble produces enormous spacetime curvature distortions. An asymmetric collapse, using a magnetic field to distort the collapse, produces, in addition to the same severe spacetime distortions, negative energy as the bubble collapses. Due to some General Relativity considerations, the wormhole that is created starts rotating in a manner similar to the beacon light produced by a lighthouse.

[0019] Referring to **FIG. 9**, due to the forward helical motion (A) of the propellers (B) as the airplane crosses the road, the pulsed gravitational wave (C) is skewed backward at an angle (D). Due to the wormholes created by the presence of thermal vorticity fluctuations generated by the wing and turbines of the airplane, this skewed wave moves into hyperspace where it is highly magnified and detected by the inventor.

[0020] Referring to top view **FIG. 10**, the gravitational wave (A) causes a skewed compression and expansion of the hyperspace quantum wells (B) which constitute the human energy being. Due to this asymmetric distortion in the xy-plane, the quantum wells take the physical body out of dimension as long as the wave pulse is traveling with the human energy field. Once the back edge of the gravitational wave moves on past the quantum wells, the body is then brought back into dimension.

SUMMARY OF THE INVENTION

[0021] It is the object of this invention to teleport a human being from one location to another by creating a pulsed gravitational wave traveling through hyperspace that asymmetrically compresses and expands the quantum wells of the human energy being. This spacetime curvature distortion of the hyperspace quantum wells pulls the physical body out of dimension such that the human being is teleported along with the wave. As the pulsed wave moves on past the quantum wells, the human is brought back into dimension at some distant location. The invention requires (1) a device that will generate a wormhole between space and hyperspace, and (2) a device that will generate a gravitational wave which can be inserted through the wormhole.

[0022] Referring to **FIG. 11**, a magnetic vortex wormhole generator has already been developed which generates a wormhole between space and hyperspace as described in a previous patent application entitled Magnetic Vortex Wormhole Generator. Using this generator, it was found that smoke blown through one side of the coil does not appear on the other side of cylindrical coil. The smoke flows through the wormhole and appears in a hyperspace co-dimension. It was this experiment that resulted in making first contact with the androids of the Grey aliens who told me, in a remote viewing session, that "We saw you blowing smoke into hyperspace."

[0023] The wormhole generator consists of two concentric cylindrical coils (A,B), one of larger radius than the other, made of thin transformer iron laminate wrapped in opposite directions with one continuous wire driven by a sinusoidal current. The solenoidal coil generates a magnetic field through the laminate. Because the electrical current flows in opposite directions at different radii through the two windings, bucking electric fields (C) are created along the centerline of the generator. These radially-offset magnetic fields and bucking electric fields, as shown by a calculation using Einstein's General Theory of Relativity, generate both an enormous spiking spacetime curvature and negative energy at small radius along the centerline where the wormhole is formed. The gravitational wave generator is then coupled to this wormhole generator.

[0024] Referring to **FIG. 12**, it is known from gravitation physics that injecting an electromagnetic wave (A) into a hollow toroidal waveguide (B) produces a hyperbolic spacetime curvature stress (C) in the plane of the waveguide. The tips of the arrows indicate compression and the tail of the arrows indicate expansion or stretching of spacetime. The reason for this spacetime curvature is because the waveguide forces the electromagnetic wave to curve around and travel in a circle. Spacetime has to compensate for this toroidal-generated stress by creating hyperbolic lines of stress in the inner plane of the toroid so that the overall spacetime curvature is zero. For a greater gravitational effect, three toroidal waveguides, phased 120 degrees apart, are used to seal off the curvature.

[0025] Referring to **FIG. 13**, the three toroids create a rotating, twisting, vertical propagating gravitational field (A) through the centerline of the toroids provided that the period of the electromagnetic wave is twice the period of the gravitational wave. This phase relationship is adjusted by selecting the correct radius for the frequency of the monochromatic wave.

[0026] In order to effectively use this gravitational wave, referring to **FIG. 14**, three phased toroidal waveguides (A,B) are mounted at the top of each of two identical square granite obelisks (C,D). The two obelisks are offset by a short distance between them. As the vertical gravitational wave rotates around along the vertical axis inside the obelisk, the edges of the square obelisks are compressed and expanded such as to create two cylindrical asymmetric gravitational waves traveling radially outward.

[0027] Referring to **FIG. 15**, these waves meet to form a plane gravitational wave (A) which travels down the centerline between the two obelisks.

[0028] Referring to **FIG. 16**, the full body teleportation system consists of the twin granite obelisks (A,B) on which are mounted near the top of each the toroidal waveguides (C,D) which produce the pulsed gravitational waves (E,F) that run the length of the obelisks. Because the gravitational wave is rotating inside the obelisk, the granite stone undergoes a very small asymmetrical compression and expansion. A cylindrical gravitational wave propagates out from each obelisk such that along the centerline between the two there is generated a plane gravitational wave. This wave enters the wormhole (H) created by the magnetic vortex generator which is located a short distance from and parallel to the obelisks. The wave is amplified by a factor of almost 10^{13} when it enters the hyperspace co-dimension.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0029] **FIG. 1**. Perspective view of site where full-body teleportation occurred.

[0030] **FIG. 2**. Perspective view of gravitational wave generator.

[0031] **FIG. 3**. Planar view of gravitational wave generator.

[0032] **FIG. 4**. Perspective view of seven vortices of human energy being.

[0033] **FIG. 5**. Perspective view of seven large quantum wells of human energy being.

[0034] **FIG. 6**. Tetrahedron diagram showing Planck's constant and electron.

[0035] **FIG. 7**. Tetrahedron diagram showing electron moving out of dimension.

[0036] **FIG. 8**. Perspective view showing production of negative energy.

[0037] **FIG. 9**. Perspective view of skewed gravitational wave produced by propellers.

[0038] **FIG. 10**. Planar view of skewed quantum wells deformed by gravitational wave.

[0039] **FIG. 11**. Perspective view of magnetic vortex wormhole generator.

[0040] **FIG. 12**. Perspective view of hyperbolic lines of stress generated by toroidal waveguide.

[0041] **FIG. 13**. Perspective view of rotating, twisting, propagating gravitational wave generated by toroidal waveguides.

[0042] **FIG. 14**. Perspective view of toroidal waveguides attached to obelisks.

[0043] **FIG. 15**. Perspective view of gravitational wave generated by obelisks.

[0044] **FIG. 16**. Perspective view of magnetic vortex wormhole generator and obelisk gravitational wave generator.

DETAILED DESCRIPTION OF THE INVENTION

[0045] 1. The obelisks are quarried out of granite stone and cut with a large-diameter diamond saw that is used in highway construction. The beveled piece at the top is cut separately and cemented in place. A tapered aluminum bracket holds the toroids in place.

[0046] 2. The electronics for the magnetic vortex generator are similar to that used in the patent application Magnetic Vortex Wormhole Generator.

[0047] 3. The electronics for the toroidal waveguides is the familiar stub and coaxial cable driven by an amplifier and pulsed variable-frequency generator.

I claim:

1. A full body teleportation system consisting of:

generating a pulsed gravitational wave which propagates through a magnetic vortex wormhole generator; and

generating a wormhole with the magnetic vortex generator whereby the pulsed gravitational wave traverses through the wormhole and enters into hyperspace where the wave is enormously magnified due to the lower speed of light in that dimension.

2. The method of claim 1, wherein the step of generating the pulsed gravitational wave comprises:

using two granite stone obelisks;

mounting monochromatic-wave toroidal waveguides on top of each obelisk to create a rotating, twisting, propagating gravitational wave through the vertical axis of each obelisk; and

creating a cylindrical compression and expansion in each obelisk to produce a plane gravitational wave traveling down the centerline between the two obelisks.

3. The method of claim 1, wherein the step of generating a wormhole into hyperspace comprises:

using two concentric cylindrical solenoidal coils of different radii connected by a single wire wrapped in opposite directions on thin iron transformer laminate;

generating bucking electric fields down the centerline of the vortex generator which creates a spacetime curvature distortion with negative energy in accordance with Einstein's General Theory of Relativity.

4. A teleportation system comprising:

generating a gravitational wave traveling through hyperspace which interacts with the human energy being; and

pulling the human energy being and physical body out of dimension when interacting with the pulsed gravitational wave such that the person is teleported from one location to another through hyperspace and back again into our 4D spacetime dimension.

* * * * *

US 20060145019A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2006/0145019 A1**

(43) **Pub. Date: Jul. 6, 2006**

(54) **TRIANGULAR SPACECRAFT**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/017,093**

(22) Filed: **Dec. 20, 2004**

Publication Classification

(51) **Int. Cl.**
B64G 1/40 (2006.01)

(52) **U.S. Cl.** **244/171.5**

(57) **ABSTRACT**

A spacecraft having a triangular hull with vertical electrostatic line charges on each corner that produce a horizontal electric field parallel to the sides of the hull. This field, interacting with a plane wave emitted by antennas on the side of the hull, generates a force per volume combining both lift and propulsion.

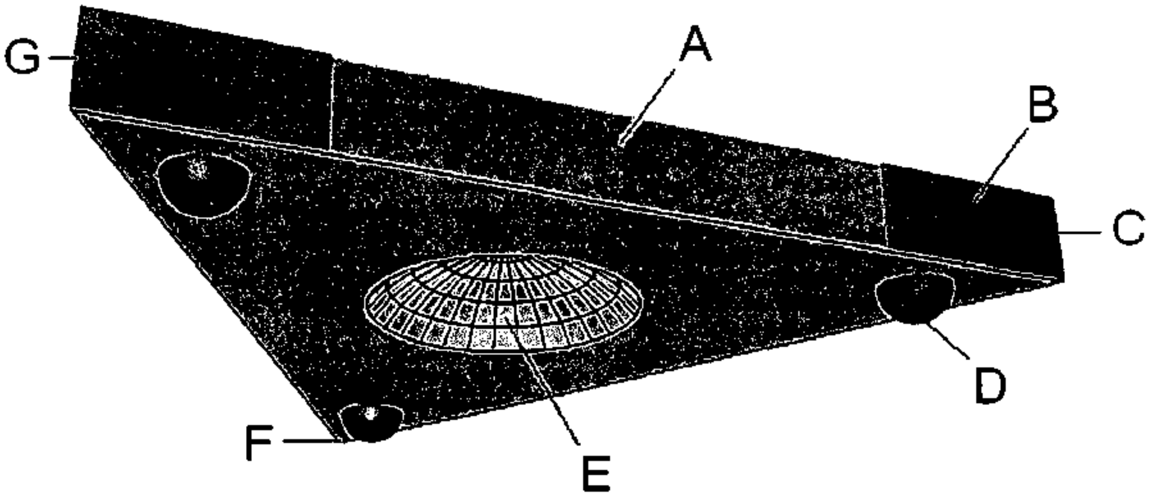

Figure 1

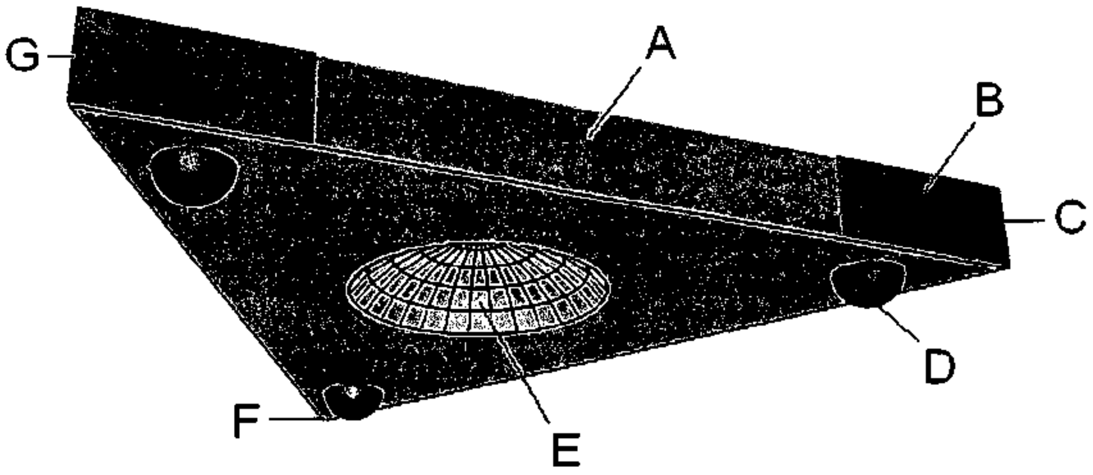

Figure 2

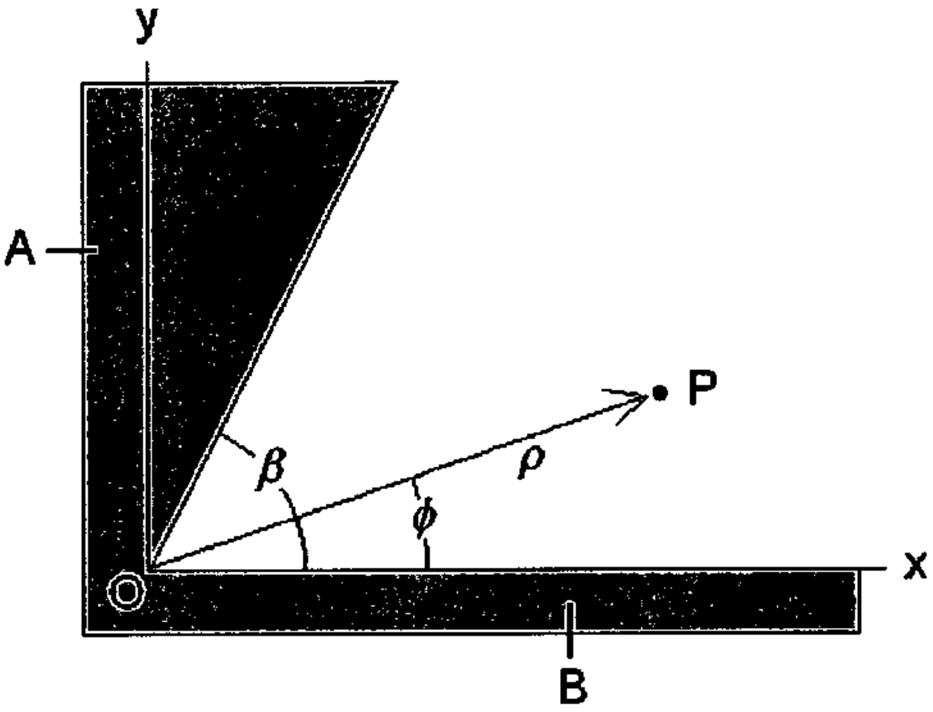
y
A
P
β
ρ
ϕ
x
O
B

Figure 3

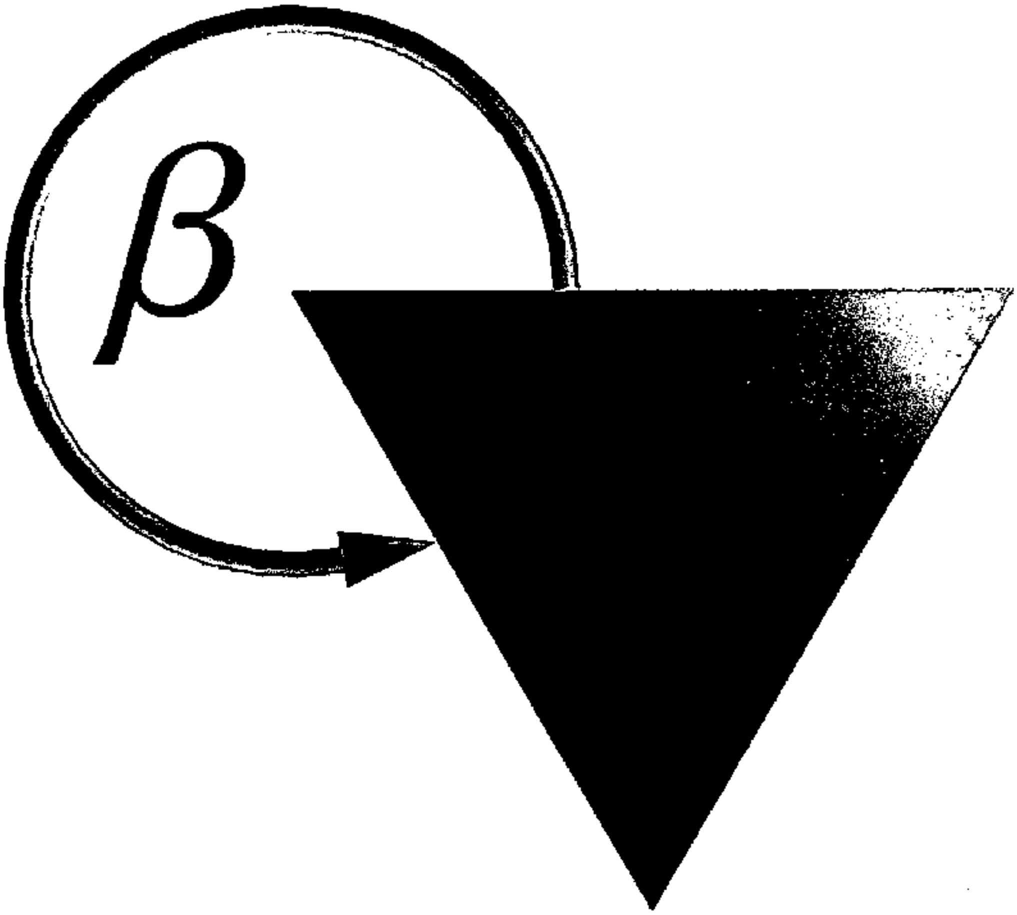

Figure 4

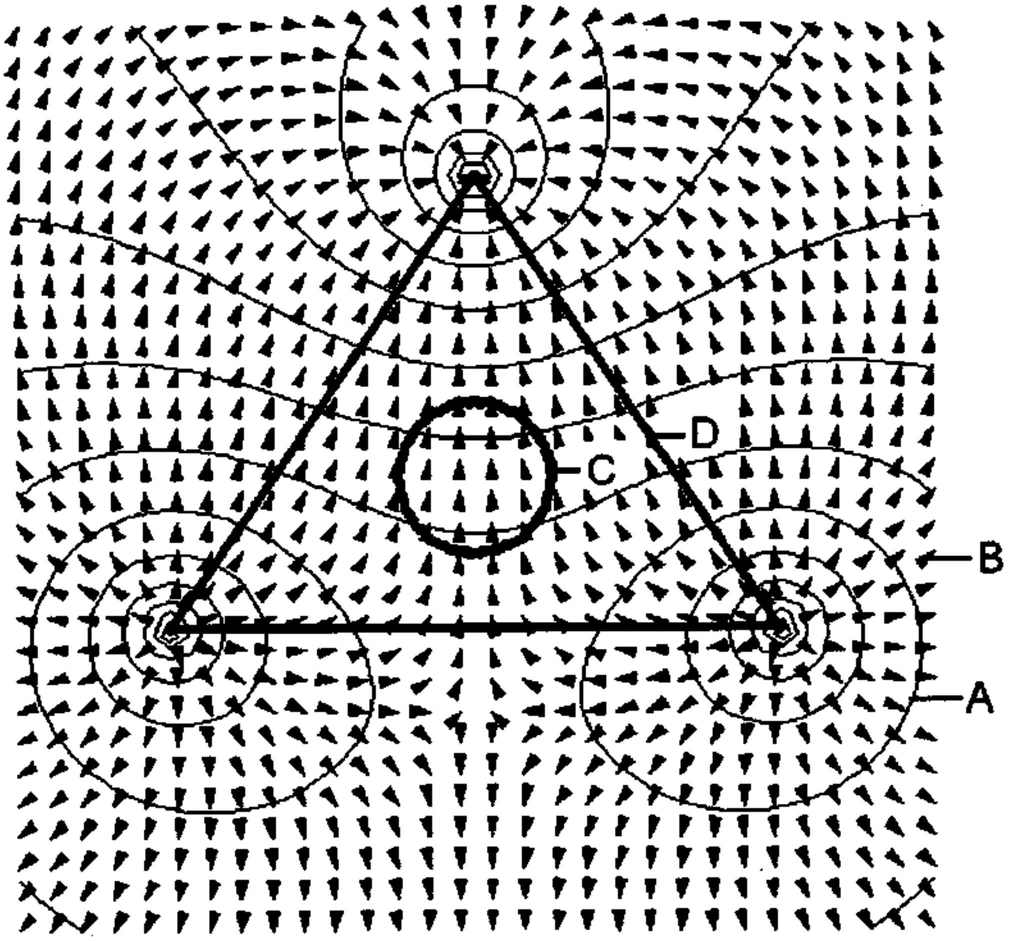

Figure 5

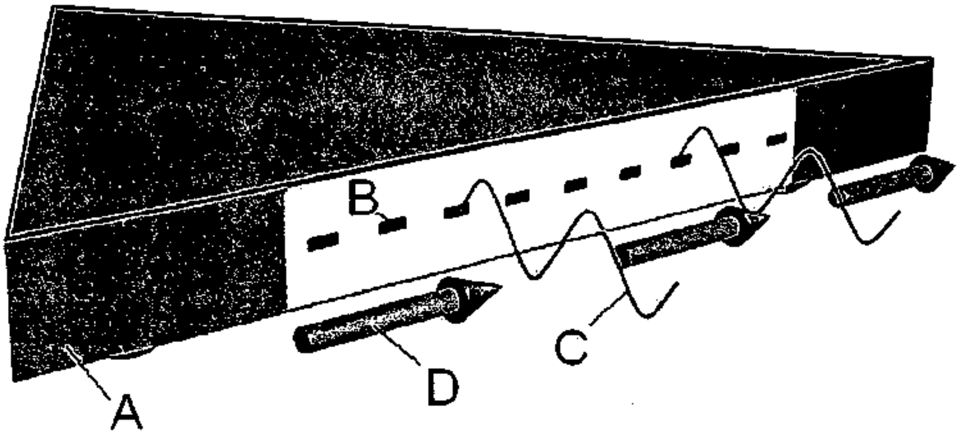

Figure 6

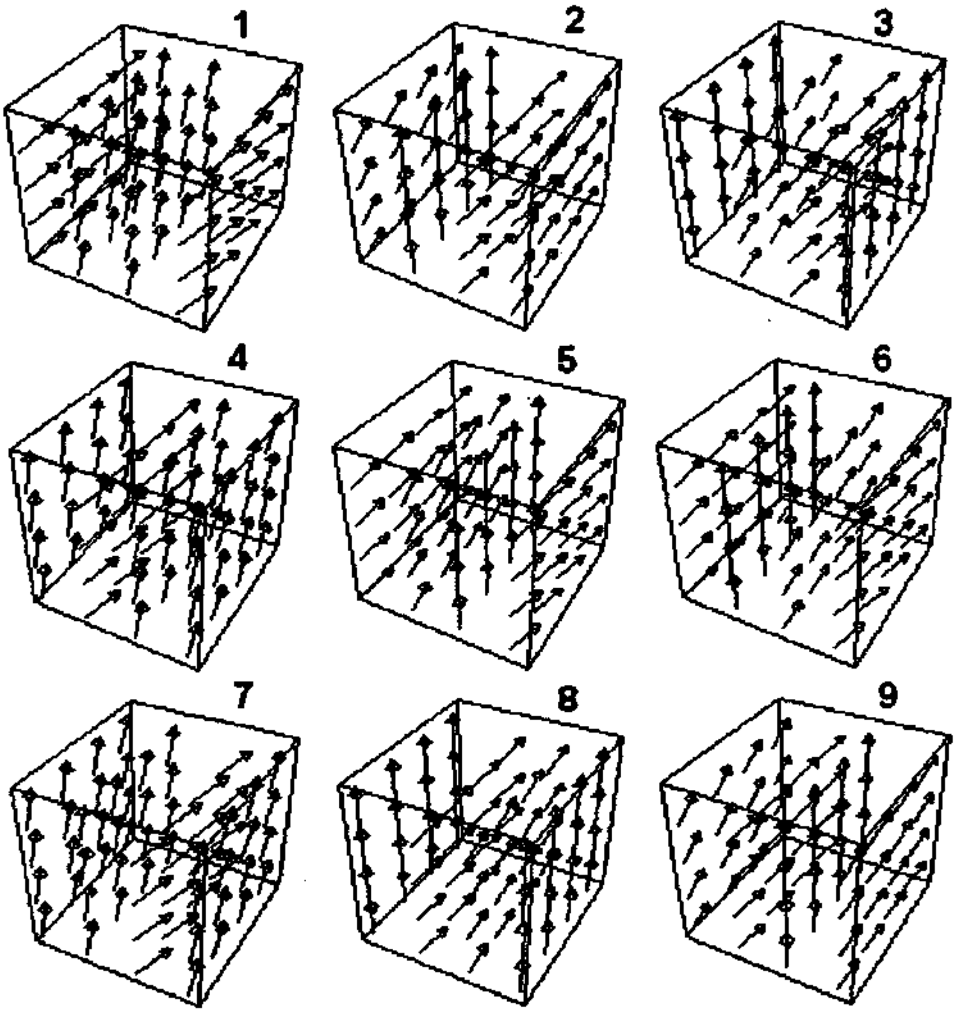

Figure 7

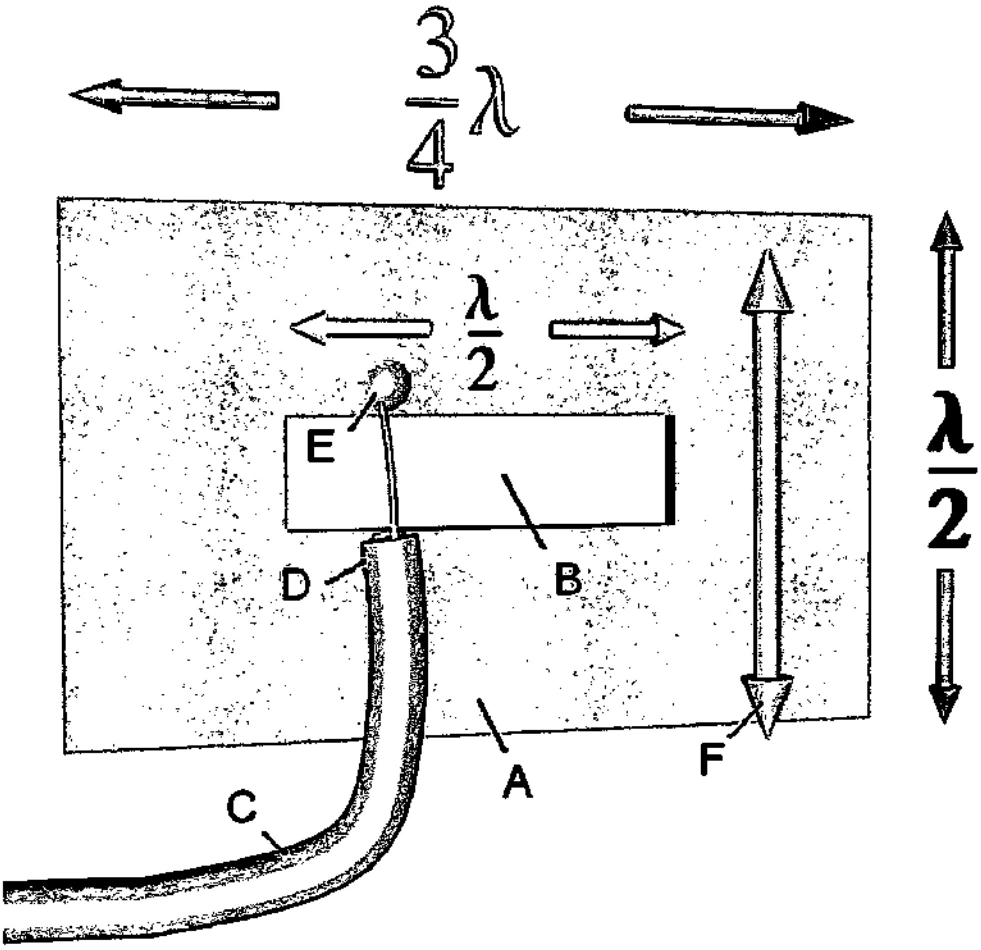
$\frac{3}{4}\lambda$
$\frac{\lambda}{2}$
$\frac{\lambda}{2}$
E
D
B
C
A
F

TRIANGULAR SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a spacecraft having a triangular hull with vertical electrostatic line charges on each corner. The line charges create a horizontal electric field that, together with a plane wave emitted by antennas on the side of the hull, generates a force per volume providing a unique combination of both lift and propulsion.

BACKGROUND OF THE INVENTION

[0002] Referring to **FIG. 1**, the spacecraft has a hull in the shape of an equilateral triangle. A parabolic antenna (E) is centrally located in the bottom of the hull. An array of horizontal slot antennas is located along the side of the hull (A). Each back corner (F,G) has a corner conducting plate which is charged to a positive voltage +V. The forward corner (C) has a conducting plate charged to a negative voltage −V. A motion control hemisphere (D) is located on the bottom surface in each of the three corners.

[0003] Referring to **FIG. 2**, two planes (A,B) intersect at the origin O at an opening angle β. Each plane (x,y) is charged to a voltage V. The potential at point P is determined in polar coordinates {ρϕ}. The Laplace equation for the potential Φ in polar coordinates is given by:

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\Phi}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2\Phi}{\partial\phi^2}=0$$

Using a separation of variables solution, the potential is given as the product of two functions:

$$\Phi(\rho,\phi)=R(\rho)\Psi(\phi)$$

which when substituted into the Laplace equation becomes:

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right)+\frac{1}{\Psi}\frac{d^2\Psi}{d\phi^2}=0$$

Since the two terms are separately functions of ρ and ϕ respectively, each one has to be constant with the sum of the constants equal to zero:

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right)=\nu^2 \quad \frac{1}{\Psi}\frac{d^2\Psi}{d\phi^2}=-\nu^2$$

These two equations have solutions:

$$R(\rho)=a\rho^{\nu+b\rho-\nu}$$

$$\psi(\phi)=A\cos(\nu\phi)+B\sin(\nu\phi)$$

The azimuthal angle ϕ is restricted to a value in the range $0\leqq\phi\leqq\beta$. The boundary condition is that the potential Φ is equal to V for any radius ρ when $\phi=0$ and $\phi=\beta$. This means that v has to be an integer value of π so that the sine function is zero:

$$\sin(\nu\beta)=\sin\left(\frac{m\pi}{\beta}\beta\right)=\sin(m\pi)=0 \quad m=1,2\ldots$$

which in turn means that the coefficient A of the cosine term has to be zero in the solution above. Choosing $b=0$ makes the general solution for the potential equal to:

$$\Phi(\rho,\phi)=V+\sum_{m=1}^{\infty}a_m\rho^{m\pi/\beta}\sin(m\pi\phi/\beta)$$

which shows that when the angle is zero, the sine is zero and the potential is V. If the angle is β, then there is a multiple of π such that the sine is zero again.

[0004] Because the series involves positive powers of the radius, for small enough ρ, only the first term $m=1$ in the series is important. Thus around $\rho=0$, the potential is approximately

$$\phi(\rho,\phi)=V+a_1\rho^{\pi/\beta}\sin(\pi\phi/\beta)$$

[0005] The electric field component is the negative gradient of the potential:

$$E_\phi(\rho,\phi)=-\frac{1}{\rho}\frac{\partial\Phi}{\partial\phi}=-\frac{\pi a_1}{\beta}\rho^{(\pi/\beta)-1}\cos(\pi\phi/\beta)$$

The surface charge distribution σ at $\phi=0$ and $\phi=\beta$ is equal to the electric field perpendicular to the surface times the permittivity of space ϵ_0:

$$\sigma(\rho)=\varepsilon_0E_\phi(\rho,0)=-\frac{\varepsilon_0\pi a_1}{\beta}\rho^{\frac{\pi}{\beta}-1}$$

Notice that if angle of intersection β is less than π, then the equation says that there is a very small radius to a positive power which means little charge density accumulation.

[0006] Referring to **FIG. 3**, the value of β, in the case of the triangular hull, is equal to 360° less 60° for a total of 300° or:

$$\beta=\frac{300}{180}\pi=\frac{5}{3}\pi$$

$$\rho^{\frac{\pi}{\frac{5}{3}\pi}-1}=\frac{1}{\rho^{\frac{2}{5}}}$$

which says that there is a charge density singularity to the two fifths power for small radius. Thus, the corner plates on the hull create a huge line charge density along the sharp vertical corner edge. The equation for the potential of a line charge density is given as:

$$\Phi(x,y)=-\frac{\lambda}{2\pi\varepsilon_0}\mathrm{Ln}((x-x_0)^2+(y-y_0)^2)$$

where λ is the charge per unit length in the vertical z-direction, and x_0 and y_0 are the location of the line charge in the xy-plane.

[0007] Referring to **FIG. 4**, the triangular hull (D) is plotted together with the potential contours (A) and the electric field arrows (B) created by the three corner line charges. The line charges are perpendicular to the paper. Notice that the electric field arrows are parallel crossing the center parabolic antenna (C). The electric field is also parallel to the sides (D) of the triangle.

[0008] Referring to **FIG. 5**, along the side of the triangle (A), an array (B) of horizontal slot antennas emit electromagnetic waves that have a vertically polarized electric E field (C). These traveling waves interact with the electric field (D) produced by the line charges on the corners of the triangle.

[0009] Using differential forms mathematics, this combination of fields is represented by the Hodge star of the differential of the wedge product of the two fields. The antenna electromagnetic field is a combination of a traveling magnetic field B_w, and electric field E_w. The stationary field E created by the line charges is perpendicular to the traveling wave.

$$*d(E\wedge(B_w+E_w\wedge dt))\frac{\varepsilon}{c}=\frac{\text{force}}{\text{volume}}$$

where ϵ is the linear capacitance of space and c is the speed of light. Thus there is a force per volume around the hull.

[0010] This combination of fields produces a spacetime curvature as determined by Einstein's General Theory of Relativity. The traveling electric field has an amplitude in the vertical z-direction and travels in the x-direction

$$E_{w=E_z}\cos(x-t)$$

The Faraday electromagnetic tensor contains all the electric and magnetic fields in all the {x,y,z} directions. The first row and first column contain the two electric fields

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} t & 0 & E_x & 0 & E_z\cos(x-t) \\ x & E_x & 0 & 0 & 0 \\ y & 0 & 0 & 0 & 0 \\ z & E_z\cos(x-t) & 0 & 0 & 0 \end{array}$$

The stress exerted on spacetime occurs in the xx, yy and zz-direction as calculated from the stress-energy tensor T of gravitational physics

$$4\pi T^{\mu\nu}=F^{\mu\alpha}F^{\mu}_{\alpha}-\frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

where g is the metric tensor for Cartesian space

$$g_{\alpha\beta}=\begin{array}{c|cccc|} t & -1 & 0 & 0 & 0 \\ x & 0 & 1 & 0 & 0 \\ y & 0 & 0 & 1 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

where the diagonal components are the coefficients of the elementary spacetime length ds squared

$$(ds)^2=-(dt)^2+(dx)^2+(dy)^2+(dz)^2$$

The calculation produces three stresses T^{xx},T^{yy} and T^{zz} in their respective {x,y,z} directions.

[0011] Referring to **FIG. 6**, these three stresses are plotted together as a 3D vector field animated over time in nine frames. The graphs show that there is a lift force as depicted by the vertical arrows as well as a force of propulsion as shown by the interspersed horizontal arrows. With the passage of time, these vectors exchange places with each other so that the lift becomes the propulsion and vice versa, creating a wavy stress-energy field around the hull.

SUMMARY OF THE INVENTION

[0012] This invention is a spacecraft with a triangular hull having charged flat plates on the vertical corners of the three sides. The two rear corners are charged to a potential V. The forward corner is charged to a potential −V. The 60° angle on the corner creates a line charge density singularity that produces a huge horizontal electric field pointing from the back to the front of the craft which is also parallel to the sides of the triangle. An array of horizontal slot antennas located on the sides of the triangular hull produce an electromagnetic wave with the electric field polarized in the vertical direction. This combination of fields produces a spacetime force in both the vertical and horizontal directions such that the spacecraft receives a lift force and a force of propulsion.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0013] **FIG. 1**. Perspective view of triangular spacecraft.

[0014] **FIG. 2**. Drawing of the intersection of two charged plates in order to calculate the charge density in the corner.

[0015] **FIG. 3**. Perspective view of the corner angle β for the equilateral triangle.

[0016] **FIG. 4**. Planar 2D graph showing the electric field produced by three line charges on the corners of the triangular hull.

[0017] **FIG. 5**. Perspective view of electric field produced by the linear charge interacting with the traveling electromagnetic wave produced by the slot antenna.

[0018] **FIG. 6**. 3D vector animation of the lift and thrust force generated by the fields.

[0019] **FIG. 7**. Perspective view of slot antenna.

DETAILED DESCRIPTION OF THE INVENTION

[0020] Referring to **FIG. 7**, the antenna (A) is made out of sheet copper in which a rectangular horizontal slot (B) has been notched out using a die press and sheet metal fixture. A coaxial cable from the amplifier and frequency generator is attached across the slot by soldering the outer cable (D) to one side of the slot and the inner cable (E) to the other side of the slot. This creates the positive and negative charges across the gap which forms the vertical electric field (F) which radiates out perpendicularly to the copper sheet.

[0021] Although the invention has been described with reference to specific embodiments, such as a particular antenna system, those skilled in the art will appreciate that many modifications and variations are possible without departing from the teachings of the invention. All such modifications and variations are intended to be encompassed within the scope of the following claims.

1. A spacecraft comprised of the following components:

(a) a triangular hull in the form of an equilateral triangle;

(b) two copper plates attached on opposite vertical sides at each of the three corners of the hull (**1***a*) such that a sharp vertical edge is formed where they come together;

(c) an electrostatic generator used to charge the back two copper-cladded corners (**1***b*) to a high positive voltage, and the third forward copper-cladded corner to a high negative voltage;

(d) a horizontal slot antenna array mounted-on the sides of the hull; and

(e) a frequency generator, antenna and coaxial cables to drive the antenna array (**1***d*).

2. To create, by claims (**1***a*, **1***b*, **1***c*), an intense vertical line charge at the corners (**1***b*) and a horizontal electric field that that is parallel to the sides of the hull (**1***a*);

3. To create, by claims (**1***d*,**1***e*), an electromagnetic wave with a vertically polarized electric field traveling outward from the side of the hull (**1***a*); and

4. To create, by claims (**2**,**3**), an interaction of the electrostatic field (**2**) with the electromagnetic wave (**3**) such that a combined spacetime curvature pressure is generated on the hull in the upward and forward direction to produce lift and propulsion respectively.

* * * * *

March 25, 1969 W. C. BROWN ET AL 3,434,678

MICROWAVE TO DC CONVERTER

Filed May 5, 1965

Sheet 1 of 3

FIG. 1

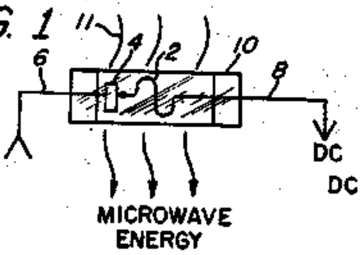
11
4
2
10
6
8
DC
MICROWAVE ENERGY

FIG. 2

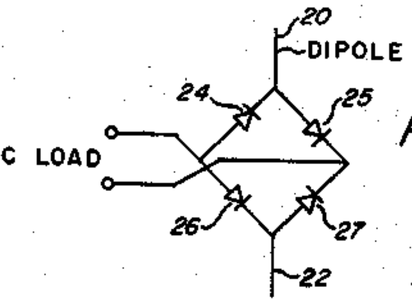
20
DIPOLE
24
25
DC LOAD
26
27
22

FIG. 3

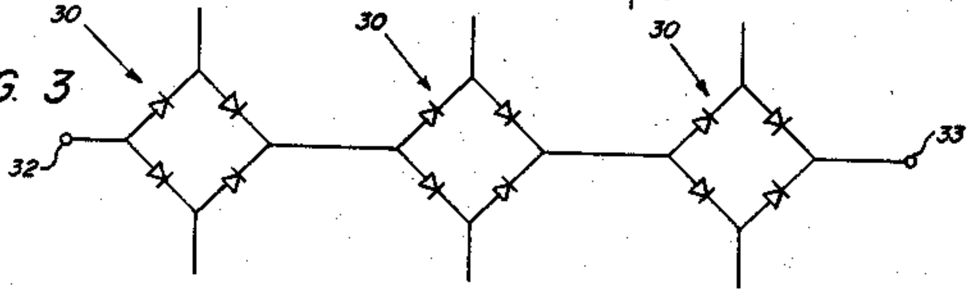
30
30
30
32
33

FIG. 4

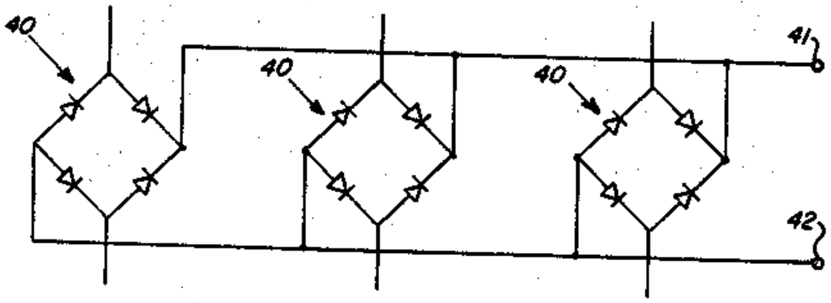
40
40
40
41
42

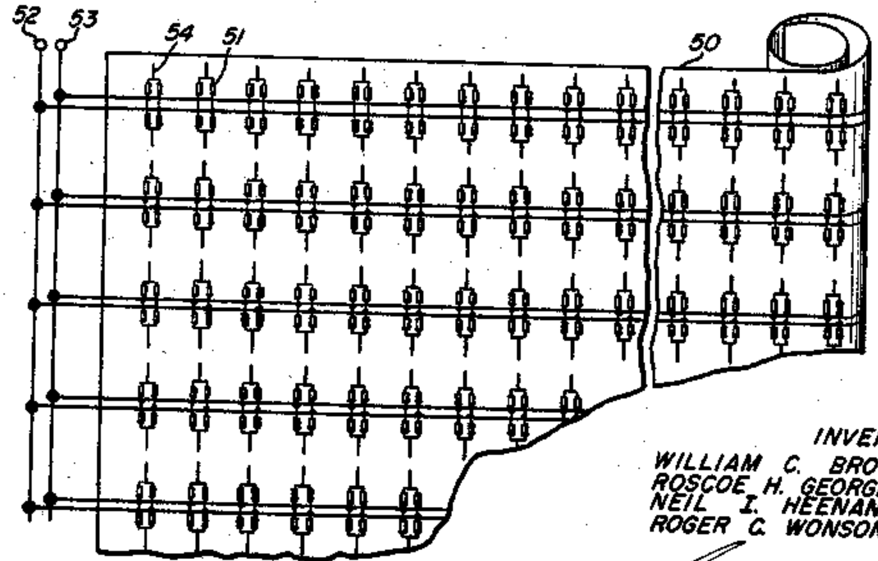
52
53
54
51
50

FIG. 5

INVENTORS
WILLIAM C. BROWN
ROSCOE H. GEORGE
NEIL I. HEENAN
ROGER C. WONSON

BY Edgar O. Rost
ATTORNEY

March 25, 1969 W. C. BROWN ET AL 3,434,678

MICROWAVE TO DC CONVERTER

Filed May 5, 1965 Sheet 2 of 3

FIG. 7

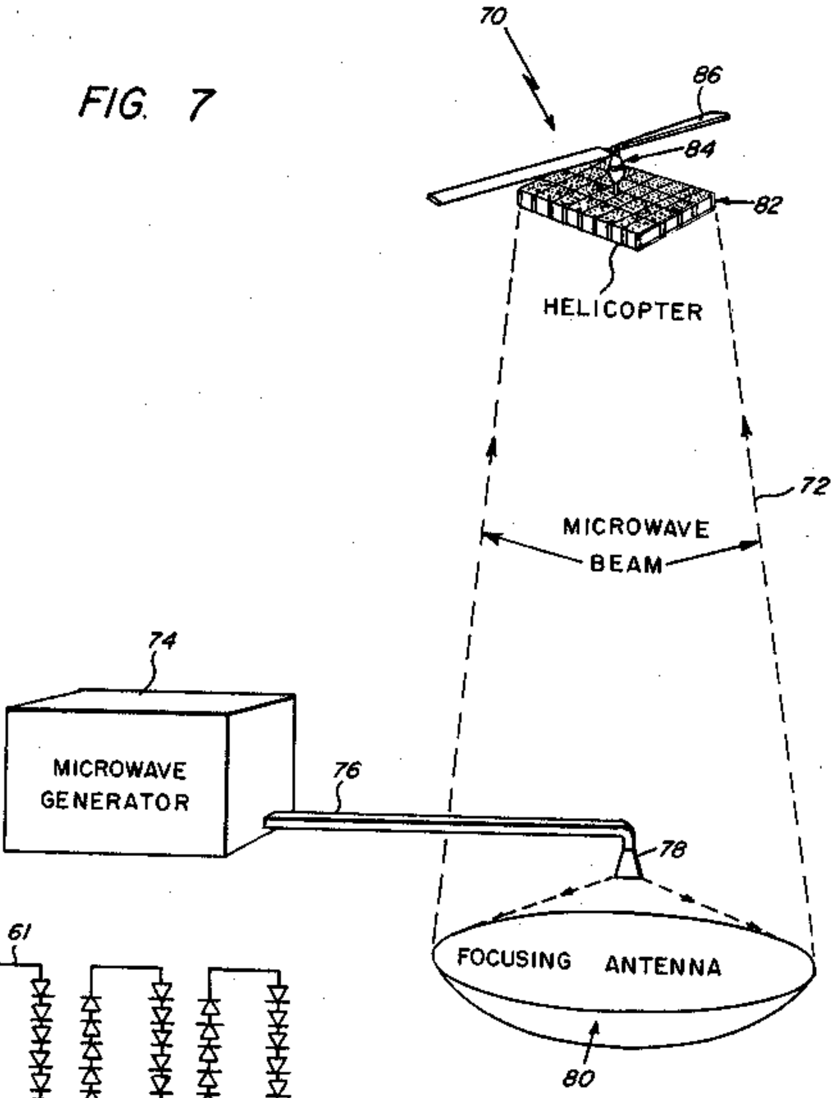
70
86
84
82
HELICOPTER
72
MICROWAVE
BEAM
74
MICROWAVE
GENERATOR
76
78
FOCUSING ANTENNA
80

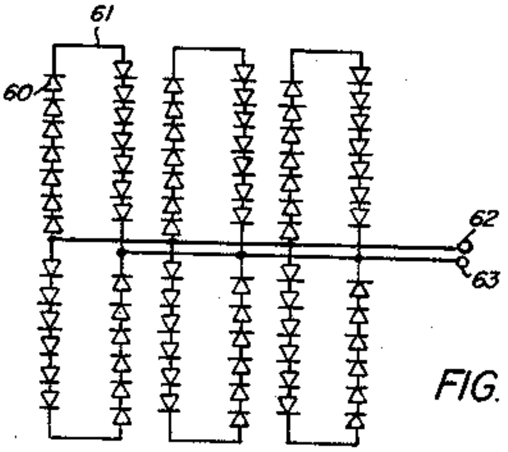

FIG. 6

INVENTORS
WILLIAM C. BROWN
ROSCOE H. GEORGE
NEIL I. HEENAN
ROGER C. WONSON

BY Edgar O. Rost
ATTORNEY

March 25, 1969 **W. C. BROWN ET AL** 3,434,678

MICROWAVE TO DC CONVERTER

Filed May 5, 1965 Sheet 3 of 3

FIG. 8

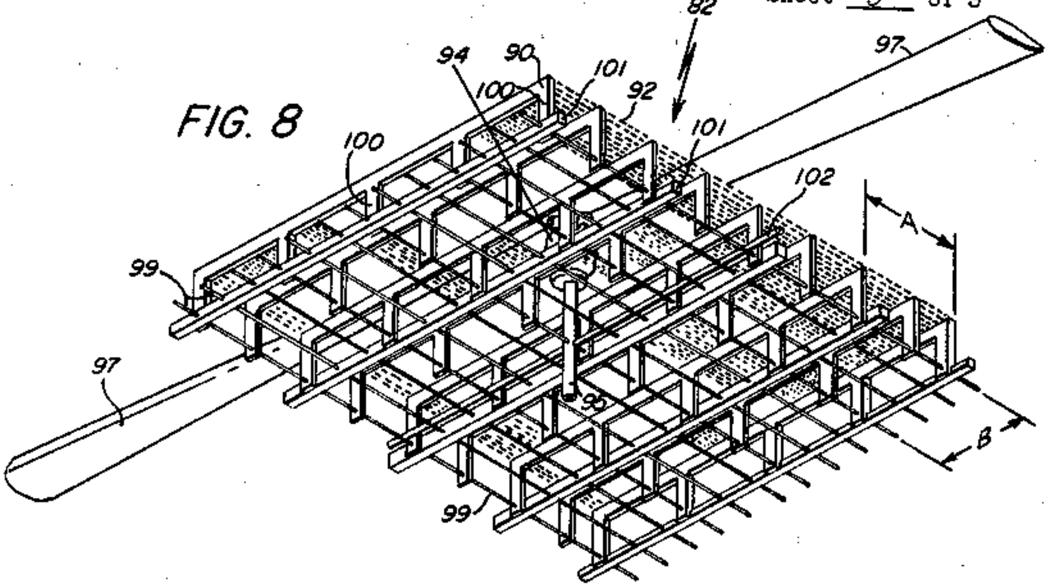
82
94
90
101
100
92
101
100
102
A
97
99
97
95
B
99

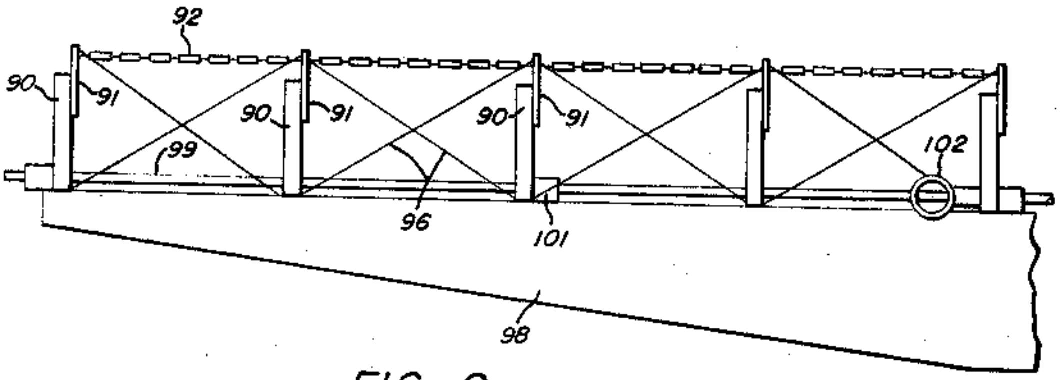
92
90
91
99
90
91
90
91
102
96
101
98

FIG. 9

INVENTORS
WILLIAM C. BROWN
ROSCOE H. GEORGE
NEIL I. HEENAN
ROGER C. WONSON

BY Edgar O. Rost
ATTORNEY

United States Patent Office

3,434,678
Patented Mar. 25, 1969

3,434,678

MICROWAVE TO DC CONVERTER

William C. Brown, Weston, Mass., Roscoe H. George, West Lafayette, Ind., and Neil I. Heenan, Needham, and Roger C. Wonson, Beverly, Mass., assignors to Raytheon Company, Lexington, Mass., a corporation of Delaware

Filed May 5, 1965, Ser. No. 453,415
Int. Cl. B64g *1/00*
U.S. Cl. 244—1
3 Claims

ABSTRACT OF THE DISCLOSURE

A combined antenna and conversion mechanism for reception of beamed high frequency electromagnetic energy in space including a large array of unidirectional current semiconductor rectifier devices. A self-supporting space vehicle utilizing the rectified DC electrical energy for propulsion is disclosed in an illustrative embodiment.

The present invention relates in general to the transfer of energy by means of an electromagnetic wave beam and more particularly to interception and rectification of such energy into low frequency electrical DC energy with a high degree of efficiency.

Improved technology in the field of microwave energy generation at superpower levels has resulted in the realization of electrical energy transmission over considerable distances for remote energization of devices or vehicles without the aid of wires. The transmission of microwave electromagnetic energy into space has been commonly employed in the radar pulse echo systems for the detection and orientation of desired objects within a predetermined scanning range of a transmitting antenna. Beams of a similar nature may now be employed for other useful purposes and the advantages attendant the utilization of electromagnetic energy in the microwave region in contrast with other wavelengths may now be enumerated.

Microwaves have been generally defined as high frequency radio waves whose wavelength is less than 30 centimeters, with a lower wavelength limit on the order of 1 millimeter sometimes being applied to what is commonly referred to as the "microwave region." The superiority of high frequency microwaves is due in part to the fact that it is generally desirable to focus the transmitted energy so as to achieve a high power density at a remote point or area with respect to a given power source. In accordance with the laws of optics, the sharpness of the microwave beam produced by a transmitting antenna varies as the ratio of antenna dimensions to the wavelength of the transmitted energy. Therefore, for a given or desired power density or beam sharpness, a decrease in the wavelength of the transmitted energy permits a corresponding decrease in the dimensions of the antenna. From the standpoint of mechanical considerations, it is desirable to employ small antennas and other components, and it is therefore advantageous to employ high frequency energy of very short wavelength. In addition, the difficulties encountered in long wave transmission as a result of natural and man-made interference or noise do not occur with any appreciable significance at microwave frequencies. Further, in aerospace applications with considerable distances separating the transmitter at an earth or mother planet location and the employment of shorter wavelength beamed energy is preferred since longer wave signals will generallly be reflected at certain altitudes by reflecting layers in the atmosphere.

In view of certain losses due to absorption which may occur in the atmosphere, microwaves in the region having the approximate bounds of 2 and 30 centimeters are readily adaptable to the convenient radiation of power to remote points without the utilization of wires. The preferred wavelengths are of the order of 5 or 10 centimeters to provide efficient focusing with existing transmitting antenna systems which may be maintained at a reasonable size. An illustrative device of the superpower high frequency microwave generators operative in the desired band is the so-called Amplitron which is an amplifier having a broad bandwidth and excellent performance characteristics for the focusing of the beam. Such devices are capable of producing 15 or 20 kilowatts of average continuous wave power in the neighborhood of 10 centimeters in wavelength with capabilities expected in the region of 500 kilowatts or more average power with 50 megawatts peak power. A complete description of such devices may be had by referring to Patent No. 2,933,723 issued Apr. 19, 1960 to William C. Brown and assigned to the assignee of the present invention.

With microwave energy capable of being generated and directed over longer distances conversion of such high frequency electromagnetic energy is of paramount concern. One conversion mechanism in the prior art involves direct conversion of such energy into heat which may then be utilized directly or indirectly for propulsion or generation of flight-producing forces. Examples of such devices for heat energy exchange as well as space vehicles utilizing such energy may be noted in Patent No. 3,174,705, issued Mar. 23, 1965, to D. Schiff et al., as well as U.S. Letters Patent No. 3,083,528, issued Apr. 2, 1963 and No. 3,114,517, issued Dec. 17, 1963, to William C. Brown. The heat exchanger method of conversion of electromagnetic energy into useful power is limited by the overall efficiencies of approximately 25 percent in the conversion of heat into mechanical or electrical work. Desirable, therefore, would be the direct rectification of the high frequency electromagnetic energy into low frequency electrical energy for the operation of many useful aerospace devices as well as systems.

The present invention has for its primary object the conversion of high frequency electromagnetic energy in the microwave region directly into low frequency electrical energy.

A further object of the present invention is the provision of a combined nondirectional receiving antenna and microwave electromagnetic energy to low frequency electrical energy conversion means in a unitary structure.

A still further object of the present invention is a provision of a new and novel combined nondirectional receiving antenna and microwave to DC energy converter for aerospace applications.

Another object of the present invention is the provision of a new and novel nondirectional receiving antenna and microwave to DC energy converter having a high degree of efficiency.

Still another object of the present invention is the provision of a new and novel aerospace vehicle with nondirectional receiving antenna and microwave to DC energy converter means with said vehicle being capable of being supported by its own energy generation means at a distance spaced apart from the power generation means.

In accordance with the teachings of the present invention, the above and other objects are achieved by the employment of efficient unidirectional microwave power rectifiers and dipole antenna means. Such rectifying devices, while being individually limited in power-handling capabilities, normally in the order of fractions of watts, have been found to be highly efficient means for the rectification of microwave power when assembled in large numbers in various arrays. It is interesting to note that the observed collective efficiency was on the order of 40 to 70 percent. In an illustrative embodiment, point-contact semiconductor diodes were arranged in four arm bridge connected networks with the networks interconnected in various configurations such as series, parallel and series-parallel.

In discussing aerospace applications, an additional problem is encountered in the beaming of microwave energy to a remote point and the interception and utilization of such electrical energy. In such applications the advantages of a vehicle which may be maintained in space for indeterminate periods of time without employing a local fuel source are readily apparent. Such devices could readily provide communication networks, surveillance functions using radar techniques along with numerous other functions. The acapture of the beamed high frequency electromagnetic energy raises the need for an efficient antenna means capable of intersecting the beam at high altitudes. Conventional techniques employed in microwave radar usage such as receiving antenna horns are capable of intersecting only a small portion of the beam energy and add considerable weight in applications involving heavier-than-air vehicles. In an exemplary embodiment of the invention a space vehicle, namely a helicopter, is disclosed for either moving flight or a stationary location with self-supporting electrically operative propulsion means. The semi-conductor diode rectifier arrays have been demonstrated to fulfill the receiving antenna functions as well as the electrical energy rectification means in a highly efficient manner. Such combined antenna and rectifier means has also assisted in reduction of the weight problem in airborne devices. Further, it has provided a nondirectional means for the interception of the microwave energy to thereby reduce the problems of focusing inherent in prior art directional horn type receiving antennas.

With the above features, advantages and objects in mind the invention will now be described by reference to the following detailed description together with the accompanying drawings in which:

FIG. 1 is a perspective view of an illustrative diode rectifier;

FIG. 2 is a schematic circuit diagram of a bridge connected diode network with dipole antenna means;

FIG. 3 is a schematic circuit diagram of a plurality of bridge connected networks arranged in series;

FIG. 4 is a schematic circuit diagram of a parallel bridge connected network array;

FIG. 5 is a perspective view of an illustrative embodiment of a combined antenna and rectifier array in a folded or rolled up configuration;

FIG. 6 is a schematic circuit diagram illustrating the bridge connected diode array incorporated in the aerospace vehicle shown in FIG. 7;

FIG. 7 is a schematic representation in elevation illustrative of a heavier-than-air aerospace vehicle incorparating the structure of the present invention;

FIG. 8 is a perspective view of the aerospace vehicle embodiment as viewed from the under portion thereof; and

FIG. 9 is an enlarged partial view in elevation of a portion of the illustrative embodiment shown in FIG. 8.

FIG. 1 illustrates a point-contact semiconductor diode rectifier of the type employed in radar microwave receiver apparatus to rectify returned radar pulses. Any of the high burnout semiconductor diodes having high rectification characteristics are preferred and are commercially available, such as the 1N82 or 1N830. The rectifying junction is formed by whisker element **2** contacting the semiconductor element **4** respectively connected to leads **6** and **8**. Silicon is preferred over germanium for element **4** because of its ability to operate at higher temperatures and thereby handle higher powers. Envelope **10** houses the rectifying elements and may be of a hermetically sealed dielectric material or combination metal and ceramic composition. The inherent characteristic of such diode rectifiers is that the microwave energy is intercepted and rectified in a unidirectional manner and the line **11** indicate pictorially the rays of the beamed electromagnetic microwave energy in a plane normal to the envelope. In FIG. 2 a full-wave bridge connected diode network is illustrated with the forward direction of the rectified DC electric current indicated by the direction of the arrow symbols. The network shown consists of half-wave dipoles **20** and **22** each terminated with a diode rectifier element **24** to **27** in an arm of the bridge connected network. The dipole elements **20** and **22** are of the half-wave configuration and may be spaced apart from each other a one-half wavelength at the frequency of the beamed electromagnetic energy.

Referring now to FIG. 3, an array of bridge connected diode networks each with the half-wave dipoles are shown connected in series. Each network is referred to by the numeral **30** and is similar in the bridge connections to the single element network shown in FIG. 2. The DC output of the collective rectified energy is coupled by means of terminals **32** and **33**. In FIG. 4, a similar number of individual bridge connected diode-dipole networks are shown connected in a parallel array. Each network is indicated by the numeral **40**, and the output terminals are indicated as **41** and **42**.

Any number of diode-dipole networks may be provided and in FIG. 5 such a multi-element array is illustrated by mounting on a flexible material **50** which may be rolled or folded into any desired package or enclosed within a capsule to be launched and released at a predetermined point in space. Any flexible material which is pervious to electromagnetic energy is preferred. The total power desired would be the determining factor in a number of individual diode-dipole elements required. In this embodiment, the bridge connected networks **51** are connected in parallel to the output load indicated by terminals **52** and **53**, and representative measurements of electrical characteristics have shown that approximately five watts of DC electrical energy is realizable for each square foot of area of the combined antenna-rectifier. While the dipole elements **54** have been indicated in a particular array, it is within the scope of the invention to stagger the placement of such dipoles to increase the overall efficiency of the antenna-rectifier.

To further increase the DC powder output, the full-wave bridge connected networks are preferably arranged with a plurality of diodes in series in each arm of the bridge. An illustrative schematic circuit diagram of such a configuration is shown in FIG. 6 wherein seven diodes **60** are shown in each arm of the bridge circuit and are connected in series for a total of twenty-eight diodes in each bridge network. The dipole members will then be the substantially U-shaped end portions **61** at the ends of each brace of seven diodes. In the illustration three such twenty-eight diode bridge networks are shown connected in parallel to terminals **62** and **63**. This closer spacing and compact arrangement has been shown to be a source of improved power output and is capable of a high degree of reliability through the redundant nature of the parallel series connections within each bridge network. If one of the diode rectifiers fails to function the over-all voltage drop across this element would be divided among the six remaining diode rectifiers. If any of the connecting wires between the diode elements should break, the adjacent arms of the other bridge assemblies would take the additional load due to the close proximity of the respective arms to each other. In addition, it is possible to have a number of open connections or inoperative diodes dispersed throughout the array without any serious impairment in performance.

In relation to the array concept to be hereinafter described it may be stated that within a six inch square area ten such individual bridge networks each containing twenty-eight diode rectifiers for a total of 280 diode rectifiers may be deployed in such a manner as to provide maximum exposed area for each diode as well as the connecting leads. Such an arrangement will be hereinafter referred to as a "module" and a DC output in excess of fourteen watts has been measured for such a module. Any number of such modules could be connected provided for a desired power yield and this module concept readily lends itself to use in certain aerospace applications now to be described.

In FIG. 7 a propelled type of space vehicle **70** is shown wholly supported by means of the transfer and rectification of continuous wave electromagnetic energy via a microwave beam **72**. The source of the microwave energy which may be of the Amplitron type device as described in the aforementioned issued Patent No. 2,933,723 is indicated as **74**. This energy is fed by waveguide means **76** to a transmitting horn **78** to illuminate an ellipsoidal beam forming focusing antenna **80** for the transmission of the microwave beam **72**. It will be appreciated by those skilled in the art that the representations of the microwave generation and transmitting antenna means are pictorial representations to illustrate the usage of the invention in diagrammatic form and the present invention is not limited to any particular source of microwave energy or transmitting antenna assembly. It may be stated the reflector of the antenna assembly is considerably larger than most of the reflectors of the prior art in order to focus a large amount of the microwave power at high altitudes for use in the transfer of energy to space vehicles. Such antenna assemblies may be partially supported in a large hollowed area on the earth's surface or other convenient means of support.

The space vehicle or helicopter **70** can be described as a main body member supporting antenna-rectifier means **82** including a large number of the so-called modules connected together and rigidly supported in a planar parallel array. A motor **84** is supported by the combined body member and the receiving antenna-rectifier means and actuates the rotor **86** of conventional design employed in such self-propelled hovering vehicles. The disclosed vehicle provides for the illumination of the planar array of the semiconductor diode dipole elements by the microwave beam and the direct conversion of the microwave power transmitted by the beam into usable electrical energy for the self-propulsion of the device without any local fuel supply being required.

FIGS. 8 and 9 illustrate a space vehicle **82** comprising a plurality of the combined receiving antenna-rectifier module means for interception and rectification of the electromagnetic microwave energy beam emanating from an earth or mother planet source. A planar array of the antenna-rectifier modules is mechanically supported by means of structural members **90** of any lightweight wood or metal. Insulators **91** positioned coextensive with the members **90** support the diode rectifier array and avoid interference with the receiving and electrical performance characteristics by the structural support members. Carrying forward the module concept of 280 diode rectifiers to provide an approximate power output of 14 watts, it was noted that any number of such modules may be coupled together since the individual module outputs are relatively insensitive to a wide range of load resistances connected to the common output terminals. To achieve the desired electrical output of approximately 120 volts and 250 watts of power, subgroups of four modules each were assembled and parallel connected with an approximate 30 volts available for each subgroup. Four such subgroups were series-connected to result in a total of 4,480 diode rectifiers or 16 modules assembled in a two foot square self-supporting planar parallel array structure. The individual diode rectifiers connected in each arm of the bridge network are indicated by the numeral **92**. An exemplary module configuration would extend within the area delineated by the dotted lines and reference letter A on one side and similar dotted lines and reference letter B on the other side.

A motor **94** is connected to the DC side of the overall array and may be additionally supported by tubular member **95**. A shaft and propulsion means consisting of rotor blades **97** provide for the upward lift of the overall vehicle for the self-supporting of same in space applications. Additional structural support such as interlaced rigging **96** of a high tensile strength material such as nylon or steel wire, as well as bracing member **98**, may be employed for strengthening of the body means to withstand the vibrational forces and downwash from the propulsion means.

In accordance with the well known technology of microwave transmission the combined array of diode rectifiers and propulsion means presents a specific load impedance which must be suitably matched to the transmitted microwave energy beam to result in maximum efficiency. In aerospace applications a mismatch of approximately ten to one may be evident. Matching of the load impedance to a value of approximately 377 ohms as the free space value will be provided by a plurality of coplanar parallel metallic rod members **99** disposed in a grating array in front of the diode rectifiers a predetermined distance. Rod members **99** are linearly disposed and extend in a similar direction as the assembled diode rectifiers. A selected frontal spacing of one-quarter of the wavelength of the microwave frequency being transmitted has been experimentally determined to be suitable for impedance matching purposes. An approximate spacing of two inches between the respective members was preferred for a selected microwave frequency of 2,450 megacycles. Each of the members **90** are provided with lateral sections **100** to support the elongated bar members **101** which in turn maintain the rod members **99** in the desired position. A tubular member **102** of a lightweight metal may also be provided to combine with the motor support member **95** for structural support.

The combined antenna-rectifier array provides a source of electrical energy to render any space vehicle self-supporting. The diode rectifier elements when assembled in the antenna array have been found to be nondirectional with respect to interception of the beamed microwave energy. This represents a large step forward in the utilization of high power microwave energy over the prior art horn-type receiving antennas which must be accurately focused and pointed in a particular direction for the reception of any energy. The connections between the respective members of the diode rectifier array and deployment in the parallel configuration serves to provide maximum exposed area. Such connections and in particular the end loop portions adjacent the terminus of each arm of the bridge networks serves as an efficient dipole for the interception of the microwave energy.

Although it is not intended as a full explanation of the high degree of efficiency attained with the disclosed antenna-rectifier array, it is believed that the whisker elements within the semiconductor diodes themselves are a contributing factor and may function as additional dipole elements. The disclosed embodiment functioned efficiently when illuminated by microwave energy generating a vertically polarized beam. Hence, an efficient and light weight energy conversion apparatus is disclosed which may be self-supporting without the requirement of a large local fuel supply payload.

It may be within the purview of the invention to use the available rectified electrical energy for performing many functions in addition to the actuation of the propulsion means. Hence, communications' payloads may be maintained at predetermined positions in space in a hovering attitude utilizing a portion of the electrical energy available. Relay signals to other such vehicles or return signals to ground stations would then be within the realm of possibility. Such available energy may also be employed for servomechanisms, stabilizing and countertorque systems for the navigation of such vehicles.

The electrical efficiencies realized with the combined receiving antenna and rectifier means have also provided certain weight advantages over other energy converters in aerospace applications. Examples of such converters would be heat exchangers or solar cells. In comparison to the present invention where five to eight pounds per kilowatt of energy realized is a normal characteristic, other energy conversion means weigh in the vicinity of 150 pounds per kilowatt of realizable energy. The inherent advantages of the present invention are therefore apparent. While the technology in the diode rectifier art is being continually advanced, new diode power rectifiers as well as integrated circuit techniques are readily available to future configurations of the present invention. The so-called Schottky barrier diodes could be employed to produce combined antenna-rectifier means weighing even less than two pounds per kilowatt of available energy.

Although the foregoing detailed description has referred to DC power rectification it will be evident that with suitable circuit components low frequency AC energy may also be made available. In addition, other propulsion means may be readily substituted using electrical energy. The embodiments disclosed herein are illustrative only and other modifications or alterations will be apparent to those skilled in the art which do not depart from the scope of the broadest aspects of the present invention as defined in the appended claims.

What is claimed is:

1. A space vehicle comprising:

body means;

said body means including spaced structural support members;

combined antenna and DC electrical energy rectification means for the interception and rectification of incident high frequency electromagnetic microwave energy carried by said support members in a planar parallel array;

said rectification means comprising a plurality of four

arm full-wave bridge connected rectifier circuit networks each having a plurality of unidirectional semiconductors in each arm;

said networks being electrically interconnected to common output terminals;

electrically operable propulsion means comprising a motor and rotor members carried by said body means and connected to said terminals for the utilization of said rectified DC energy; and

means for matching the load impedance of said combined antenna and electrical energy rectification means to the incident microwave energy.

2. A space vehicle according to claim 1 wherein said load impedance matching means are arranged in a coplanar array coextensive with said antenna and energy rectification means array, and spaced therefrom a distance of approximately one-quarter of a wavelength at the frequency of the microwave energy.

3. A space vehicle according to claim 2 wherein said load impedance matching means comprise a plurality of parallel disposed elongated metallic members.

References Cited

UNITED STATES PATENTS

1,217,149	2/1917	Caldwell	321—27 X
2,165,055	7/1939	Kafka	321—27
2,444,458	7/1948	Master	321—8
2,927,321	3/1960	Harris	343—6.8 X
3,098,971	7/1963	Richardson	325—592 X
3,174,705	3/1965	Schiff et al.	244—1

RODNEY D. BENNETT, *Primary Examiner.*

MALCOMB F. HUBLER, *Assistant Examiner.*

U.S. Cl. X.R.

307—151; 318—16; 321—27; 325—494; 343—100

Teleportation of Entangled Coherent States through Visual Pathways in the Human Brain

[1,2,3]**V. Salari**[*] , [1,2]**M. Rahnama**[†] and [4]**J. A. Tuszynski**[‡]

[1]*Department of Physics, Shahid Bahonar University of Kerman, P. O. Box 76175, Kerman, Iran*
[2]*Afzal Research Institute, Kerman, Iran*
[3]*Neuroscience Research Center, Kerman, Iran*
[4]*Department of Physics, University of Alberta, Edmonton, T6G 2J1, Canada*

Abstract:

In this paper we have proposed a new model for visual consciousness based on the premise there exists a quantum teleportation mechanism between the eye and the brain. In our model, some open questions will be answered.

1) Introduction:

Schrödinger's book "What is life?" has had an enormous influence on the development of molecular biology [1]. His insight has since then inspired many researchers to investigate the molecular basis of a living organism [2,3,4]. Several researchers have noticed the sweeping consequences that would follow from the discovery that living organisms might process information quantum mechanically, either at the bio-molecular level, or the cellular/neuronal level [5,6,7,8,9]. Computational models of the behavior of networks of neurons are still mainly based on the integrate-and-fire model of neural function. Mainstream cognitive neuroscience typically ignores the role of quantum physical effects in the neural processes underlying cognition and consciousness. However, many unsolved problems remain, suggesting the need to consider new approaches. Recent contributions to the investigations of quantum effects in the human brain are due to Jibu and Yasue, Pribram, Lockwood, Mavromatos and Nanopoulous, Hameroff and Penrose, and Stapp [10]. These authors have proposed models in which the operation of consciousness is associated with some sort of explicit wave function collapse. There have been numerous suggestions that consciousness is a macroscopic quantum effect possibly involving superconductivity, superfluidity, electromagnetic fields, Bose-Einstein condensation or some other physical mechanism. Perhaps the most specific model developed thus far is that of Penrose and Hameroff and it asserts that quantum information processing takes place in microtubules (MTs). It has been argued that MTs can process information similarly to a cellular automaton, and hence Hameroff and Penrose suggest that neuronal MTs may operate as a quantum computer.

In this paper, we investigate visual information transfer from a quantum information point of view. Theoretically we consider the conscious observer looking at a quantum system and propose that the state of this system is reported via superposed photons. We address the question whether the observer can receive the exact same state of this system quantum mechanically in his/her brain or this quantum state collapses before reaching the brain. Below we investigate the problem in detail.

2) Evolution of Information from the Eye to the Brain

We assume that a conscious observer directs his/her attention to a quantum system. For simplicity we consider this system to be a manifestation of the celebrated *Schrödinger's cat*. This system can exist in two states: Live$|L\rangle$, or Dead$|D\rangle$.

---

[*] *vahidsalari@graduate.uk.ac.ir*
[†] *Majid.rahnama@mail.uk.ac.ir*
[‡] *jtus@phys.ualberta.ca*

$$\psi_{sys} = \frac{1}{\sqrt{2}}(|L\rangle + |D\rangle) \qquad (2\text{-}1)$$

The state of this system is then reported via superposed photons. 4% of these photons are reflected from the cornea. 50% of remaining photons are dissipated through ocular media absorption. Other photons enter the 200-250 µm thick retina. There, photons interact with the photoreceptors in the rods and cons layer after 80% loss due to retinal transmission [11],[12]. In this case, we consider just a few remaining photons which are in a superposed quantum state. The key question here is whether this quantum state of photons can be reported to the brain.

When this state interacts with the last layer of retina, according to Thaheld, this superposed photon undergoes a wavefunction collapse. On the other hand, photons can be absorbed and then transformed into classical signals. Here, we use the symbols introduced by Tegmark [13] for the observer. The symbol $|\ddot{\_}\rangle$ denotes the state for which the information on photons is not received by the brain and thus the observer is amphoteric. The symbol $|\ddot{\smile}\rangle$ stands for the state in which the information received in the brain reports that the cat is alive (and the observer is happy). Finally, the symbol $|\ddot{\frown}\rangle$ corresponds to the state in which the information received in the brain indicates that the cat is dead (and the observer will be sad). It means that:

$$U|\ddot{\_}D\rangle = |\ddot{\frown}D\rangle \qquad (2\text{-}2\text{-}a)$$

$$U|\ddot{\_}L\rangle = |\ddot{\smile}L\rangle \qquad (2\text{-}2\text{-}b)$$

Where $U = \exp\left[-\frac{i}{\hbar}\int H_{photon-brain}\, dt\right]$.

Now, we consider another state in which the brain interacts with itself. Penrose and Hameroff have proposed a model of consciousness involving quantum computation with objective reduction in *microtubules* (or MTs) within the brain's neurons (see Figure 1) [14],[15],[16],[17]. MTs are cylindrical polymers comprised of the protein tubulin which organize numerous cellular activities including neuronal motor transport. According to Hameroff and Penrose, switching of tubulin conformational states is governed by quantum mechanical forces within each tubulin interior, and an essential feature of the Orch OR model is that tubulins may exist in quantum superpositions of two or more conformations. Therefore, these states could function as quantum bits, or "qubits" by interacting non-locally (through their entanglement) with other tubulin qubits so that MTs may act as quantum computers. When sufficiently many entangled tubulins are superposed for long enough to reach Penrose's OR threshold given by E=h/T, an objective reduction (OR) "conscious event" occurs as stated in the Orch-OR model.

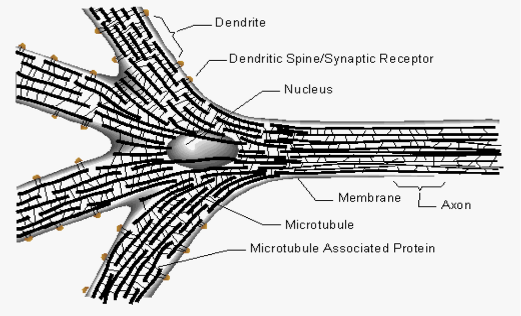
Dendrite
Dendritic Spine/Synaptic Receptor
Nucleus
Membrane
Axon
Microtubule
Microtubule Associated Protein

Figure 1 Representation of MTs inter a brain neuron. The Orch OR model suggests that the main information processing is implemented in these structures.

If previous evolution is described by Penrose's self-collapse in the brain (i.e. Orch-OR), MTs in the neurons of the brain collectively evolve and then collapse (i.e. conscious event) to one of the happy or sad states. It mathematically means that:

$$U|\overset{\cdot\cdot}{\_}\rangle = \frac{1}{\sqrt{2}}(|\overset{\cdot\cdot}{\smile}\rangle + |\overset{\cdot\cdot}{\frown}\rangle \qquad (2\text{-}3)$$

where $U = \exp[-\frac{i}{\hbar}\int H_{brain}\, dt]$.

If we compare this state (2-3) and the state which has the information about the photon (2-2), we can say that there's a great amount of correlation between the retina and the visual cortex, because their result should be identical. We know that in accordance with the Einstein-Podolsky-Rosen (EPR) pair, when two entities originate from a common source they can be entangled with each other. Retina has a similar layered structure as the gray-matter top layers of the cerebral cortex of the brain. In fact, retina is an extension of the central nervous system (the brain and spinal cord) that forms during embryonic development. One reason why scientists are interested in retinal processing is that retina is an accessible part of the brain that can be easily stimulated with light [18]. Thus, we can say that retina and visual cortex are entangled with each other. The other argument for this entanglement is that the image of a system is inverted on the retina but it is perceived in the brain directly. As we know, in two entangled entities when one has a property in some direction the other exhibits that property but in the opposite direction. Thus retina receives an inverted image, while it is received in the visual cortex upright. This property is special to entanglement.

The human body is made up of many organs, which themselves are made up of many millions of cells. How can such a system, with millions, or even billions, of parts function effectively and coherently? This is a no small feat as even small-size human societies often undergo periods of turbulence and trouble due to conflict and poor organization. Now some scientists think that quantum coherence in MTs is a major factor responsible for our bodies, and especially our brains, being so efficiency. Here, we intend to describe visual processing on the basis of quantum information theory.

3) Generation of Entangled Coherent States

Centrioles and cilia, which are microtubular structures, are involved in photoreceptor functions in single cell organisms and primitive visual systems. Cilia are also found in all retinal rod and cone cells. The dimensions of centrioles and cilia are comparable to the wavelengths of visible and infrared light (see Figure 2) [50]. Moreover, cytoskeletons are found mostly among the retina and the visual cortex in the cells of the optic nerve. Cytoskeletal structures of the centrioles can be expected to vibrate like a harmonic oscillator in its ground state. Vibrational dynamics of MT's has been the subject of a recent paper where typical frequency ranges have been discussed [19].

When a photon interacts with a centriole, its electric field can displace the potential of the harmonic oscillator and then releases it, thus generating coherent states [20]. It means that:

$$D(\alpha)|0\rangle = |\alpha\rangle \qquad (3\text{-}1)$$

Where $D(\alpha)$ is the displacement operator; $D(\alpha) = e^{\alpha a^{\dagger} - \alpha^{*} a}$, and $|0\rangle$ is the ground state of harmonic oscillator, and $|\alpha\rangle$ is the coherent state: $|\alpha\rangle = e^{-\frac{|\alpha|^2}{2}} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle$.

Hameroff [21] and Penrose [22] have suggested that MTs inside cells permit long-range quantum coherence, enabling quantum information processing to take place at the sub-cellular level. They use this hypothesis to develop their theory of consciousness. Cells interconnected by gap junctions form networks which fire synchronously, behaving like one giant neuron [23], and possibly accounting for synchronized neural activity such as coherent 40 Hz waves [24]. Marshall [25] suggested that coherent quantum states known as Bose-Einstein condensation occurred among neural proteins [26], [27], [28]. Other issues, such as preconscious-to-conscious transitions were identified by Stapp [29] with the collapse of a quantum wave function in presynaptic axon terminals [7]. The other reason for coherence of these structures is that light is an electromagnetic wave and thus is vibrational, and according to Froehlich's theory [44],[45],[46] it can take these cytoskeletal structures (i.e. nonlinear structures composed of electric dipoles) into one mode of frequency and coherence.

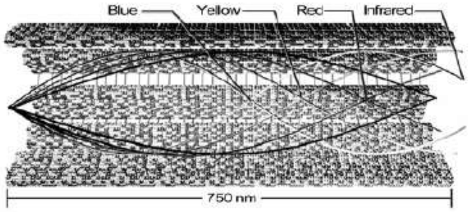

Figure 2 Representation of one centriole, the dimensions of centrioles are comparable to wavelengths of visible and infrared light.

Centrioles are two cylinders composed of MTs which are perpendicular to each other and in accordance with the Hameroff theory of the origin of cancer [30], centrioles are entangled with each other. Because of this entanglement, when a coherent state $|\alpha\rangle$ is generated in one centriole, in the other it will generate state $|-\alpha\rangle$. Now, we can say that after interaction of photons with centrioles, they generate "entangled coherent states" in these structures in the retina, i.e.:

$$|\varphi\rangle_{12} = (A|\alpha\rangle_1|\alpha\rangle_2 - B|-\alpha\rangle_1|-\alpha\rangle_2\ ) \qquad (3\text{-}2)$$

where $|\varphi\rangle_{12}$ is an entangled coherent state in centrioles with two modes 1 and 2. QED-cavity model of MTs [31] describes that coherent modes of electromagnetic radiation can be sustained in the interior of the MTs. These modes are provided by the interaction of the electric dipole moments of the ordered-water molecules in the interior of MTs with quantized electromagnetic radiation [32],[33]. Jibu, et. al.[49], have proposed that the quantum dynamical system of water molecules and the quantized electromagnetic field confined inside the hollow MT core can manifest a specific collective dynamical effect called superradiance [34] by which the MT can transform any incoherent, thermal and disordered molecular, atomic or electromagnetic energy into coherent photons inside the MT. Furthermore, they have also shown [49] that such coherent photons created by superradiance penetrate perfectly along the internal hollow core of the MT as if the optical medium inside it were made "transparent" by the propagating photons themselves. This is referred to as the quantum phenomenon of self-induced transparency [35]. Superradiance and self-induced transparency in cytoskeletal MTs can lead to "optical" neural holography [36]. Neurons (and maybe also other cells) may contain microscopic coherent optical supercomputers with enormous capacity. Thus Jibu, et. al.[49], suggest that MTs can behave as optical waveguides which result in coherent photons. They estimate that this quantum coherence is capable of superposition of states among MT spatially distributed over hundreds of microns. These in turn are in superposition with other MTs hundreds of microns away in other directions and so on. With the above conclusions $|\varphi\rangle_{12}$ can produce those photons which produced itself, thus if the state $|\varphi\rangle_{12}$ can be restored in the brain, it will reproduce the photons which were absorbed in the retina.

Additional arguments in favor of the feasibility of production of photons in the brain can be found in the conclusions of the paper [37], which also asserts that there exists a neural activity-dependent ultra-weak photon (biophoton) emission in the brain. Thus there is the possibility to restore the initial state of the photon in the brain after absorption in the eye. This process can be implemented through teleportation mechanism between retina and the visual cortex as will be discussed in the following sections.

4) The Decoherence Problem

The important question about the quantum processing in the living systems is: how is it possible for MTs to process information quantum mechanically while the environment surrounding them is relatively hot, wet and noisy?

According to the Orch-OR model, microtubular structures in the neurons of the brain process information quantum mechanically and to avoid decoherence, like lasers maintain quantum coherence against thermal noise. Water within cells is itself not truly liquid, but has been shown to be, to a large extent, ordered [38]. Most of the ordered water in the cell in fact surrounds the cytoskeleton [39]. MTs and other cytoskeletal components are embedded in cytoplasm which exists in alternating phases of (1) “sol” (solution, liquid); and (2) “gel” (gelatinous, solid). Among the most primitive of biological activities, “sol-gel transformations” within neurons and other living cells are caused by assembly and disassembly of cytoskeletal actin (e.g. regulated by calcium ions through the protein calmodulin, in turn regulated by MTs). Sol-gel transformations are essential in basic cellular activities such as (“amoeboid”) movement, growth and synaptic formation and neurotransmitter vesicle release [40], [41]. Transitions can occur rapidly (e.g. 40 sol-gel cycles per second), and some actin gels can be quite solid, and withstand deformation without transmitted response [42]. Cyclical encasement of MTs by actin gels may thus be an ideal quantum isolation mechanism. In the gel phase of cytoplasm, the water ordering surfaces of a MT are within a few nanometers of actin surfaces which also order water. Thus bundles of MTs encased in actin gel may be effectively isolated extending over the radius of the bundle, on the order of hundreds of nanometers. There are many mechanisms which can protect these structures against decohering factors. In general, Quantum states of tubulin/MTs are protected from environmental decoherence by biological mechanisms which include phases of actin gelatin, plasma-like Debye layering, coherent pumping and topological quantum error correction [54]. Moreover, quantum spin transfer between quantum dots connected by benzene rings (the same structures found in aromatic hydrophobic amino acids) is more efficient at warm temperature than absolute zero[43]. It is conjectured that the “flexibility” of the resonant benzene electrons is advantageous to quantum processes by harnessing ambient thermal energy. MTs may possibly utilize nonspecific thermal energy for "laser-like" coherent pumping, for example in the GHz range by a mechanism of "pumped phonons" suggested by Froehlich [44,45,46].

5) What is the mechanism of Teleportation?

According to the definition of teleportation as stated in [47], in the process of quantum teleportation, one can construct an exact replica of the original unknown quantum state at a cost of destroying the original state. Therefore, to call a quantum state transfer operation- quantum teleportation, the process should not only generate output states with better qualities than what can be done classically but also obey the no-cloning theorem [48].The quantum state of a system can be transmitted from a location to a distant one using only classical information provided that a quantum channel exists between the sender and the receiver. Sharing entangled states between the two parties opens the necessary quantum channel. Research in quantum state transfer, especially the quantum teleportation, has emerged as one of the major research areas of theoretical and experimental quantum mechanics. Assume that Alice wants to send Bob an unknown quantum state but when she receives this state she does not know anything about that unless she affects it and collapses it to classical state, or in other hand she destroys that quantum state. She just can send classical signals to Bob through classical channel, but if there is a shared entangled channel between Alice and Bob, Bob can construct initial quantum state with the help of classical signal which is sent by Alice and quantum channel between them. This operation is implemented by operating special unitary operators. For more details see [47]. Here, we want to simulate visual information transfer with teleportation mechanism. We know that when photon penetrates the retina, it changes to action potentials or electrical signals and these classical signals are sent to the brain for interpretation. It means that *retina* (Alice) wants to send the *brain* (Bob) a *photon state* (unknown quantum state), but *retina* (Alice) absorbs it (collapses the quantum state) and changes it to *action potential* (classical state) and send it through *membranes of axons of the brain neurons* (classical channel). *Brain* (Bob) can reconstruct the initial *state of photon* (unknown quantum state) to process it to emerge consciousness. Our arguments for this simulation are below conclusions:

> *Orch OR* [14], [15]: There is quantum processing in the neurons of the brain (there is a quantum channel between retina and brain)
>
> *Tegmark* [13]: Displacement of ions through membranes of brain neurons is a classical phenomenon (action potentials are classical signals and membranes of neurons are classical channels).
>
> *Thaheld* [11],[12]: Superposed photons do collapse in the retina (the quantum state is collapsed by the sender [Alice]).
>
> We assert that they describe different aspects of the teleportation mechanism. We simulate visual consciousness with the teleportation mechanism as shown in Table 5-1.

Table 5- 1 Simulation of the transferring visual information from the eye to the brain to the Teleportation mechanism.

Human Brain	***Quantum Teleportation Mechanism***
Retina	Alice
Membrane of axons in neurons	Classical channel
Cytoskeletal structures	Entangled channel (quantum channel)
Visual cortex	Bob
Action potentials	Classical signals

Now we want to investigate this teleportation mechanism via teleporting of entangled coherent states through visual pathways in more details. We will show how photon states are constructed in the visual cortex.

6) Teleportation of Entangled Coherent States through Visual Pathways

Super radiance and self induced transparency [49] besides Froehlich oscillations can cause the coupling of MT dynamics over wide areas and take them into a superposition and a coherent state. While in superposition, tubulins communicate with entangled tubulins in the same manner, and in MTs in neighboring neurons, and through macroscopic regions of the brain via tunneling through gap junctions and possibly tunneling nanotubes [50]. As we have already said, retina and the visual cortex are entangled with each other. Thus there is a quantum channel between retina and the visual cortex which is composed of microtubular structures. MTs interact with membrane structures mechanically by linking proteins, chemically by ions and second messenger signals, and electrically by voltage fields. The transduction of light into electrical signals takes place in the photoreceptors. Except for the ganglion cells, none of the retinal cells display action potentials [51].

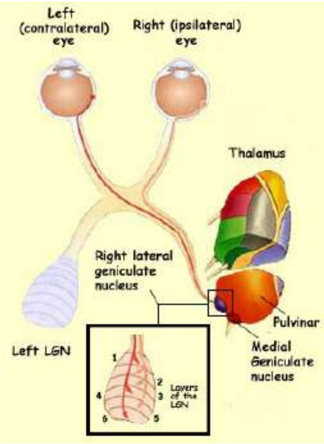

Figure 3 The LGN is a laminated structure, having 6 layers. Contralateral fibers and ipsilateral fibers couple in the LGN. The ipsilateral fibers of the optic nerve terminate in laminae 2,3 and 5 of LGN, while the contralateral fibers terminate in laminae 1, 4 and 6.[52]

Axons leaving the temporal half of the retina traverse the optic nerve to the optic chiasm, where they join the optic tract and project to ipsilateral structures. Axons leaving the nasal half of the retina cross the midline at the chiasm and terminate in contralateral structures. This arrangement means that all the axons in the optic tract carry information about the contralateral visual field. Axons of the optic tract terminate in three areas of the central nervous system, the lateral geniculate nucleus (i.e. LGN), the superior colliculus and the pretectal area. The trajectory through the LGN is the largest most direct and clinically most important pathway by which visual information reaches the cerebral cortex. About 80% of the optic tract axons synapse in the LGN. The LGN is a laminated structure, having 6 layers. Contralateral fibers and ipsilateral fibers couple in the LGN. The ipsilateral fibers of the optic nerve terminate in laminae 2,3 and 5 of LGN, while the contralateral fibers terminate in laminae 1, 4 and 6 of LGN (see Figure 3). There are about 10^6 neurons in each LGN, all of which project to the ipsilateral occipital cortex (area 17) as the optic radiations. The portion of the cerebral cortex that receives LGN axons is called the striate cortex and is usually labeled V1 to designate it as the primary visual cortical area (Figure 3). Virtually all information in the visual system is recognized as being processed by V1 first, and then passed out to higher order systems [53]. The upper visual cortex receives signals from the lower visual field and similarly, lower visual cortex process information from the upper visual field. The right visual cortex processes the left field of view and vice versa (see figure 4).

Now, we investigate the information transfer through visual pathways. The important model among quantum models about the processing in the human brain is the Orch-OR model which is based on the structure of the cytoskeleton. As already said this model asserts that the main processing in the neurons of the brain is performed in the MTs and the nature of the processing is mainly quantum mechanical. The processing unit in this model is *tubulin* which can be in a superposed state. Tubulins act like qubits in quantum computers. *Max Tegmark* believes that there is no quantum processing in the human brain. He has calculated decoherence times for every superposition state which can be possible in the neurons of the brain [13]. In his opinion, superposition states include ions such as Na^+ which are “in” and “out” of the membrane of axon. On the other hand, Na^+ ions are in the superposition of “in” and “out” with a separation distance comparable to the membrane thickness. He has considered three factors which can destroy this superposition state in neurons. *Collisions with the neighboring ions, Collisions with the water molecules* and *Interaction with distant ions* are the factors which Tegmark investigated for decoherence. He has estimated 10^{-19} s and 10^{-20} s for decoherence times. It is clear that above times are very small for the brain processes such as seeing, thinking, speaking and the other cognitive processes. Typically, dynamical timescales for neuron firing and cognitive processes are in the range of 10^{-4} to 1 seconds, whereas decoherence timescales are many orders of magnitude shorter. Thus, action potentials should be regarded as classical signals or the displacement of ions through the membrane of axons should be investigated classically. It is remarkable that Tegmark has also calculated decoherence time for MTs, but he has used wrong assumptions about these structures (for more details see [54]) and we just accept his calculations about action potentials. Thaheld [11], [12] believes that the wave function of any superposed photon state or states is always objectively changed within the complex architecture of the eye, and any incident photons have to run a very daunting gauntlet before they are even converted or transduced to retinal ganglion cell spike trains (To learn more about Thaheld arguments, the reader is referred to refs. [55,56,57]). According to Thaheld, the quantum state of photons does collapse in the retina and it does not reach the brain.

Is Thaheld right? Is not there any mechanism to rebuild the quantum state of photons in the brain? Here we accept that the states of photons collapse in the retina but we believe that they can be restored in the visual cortex via the teleportation mechanism.

Now, the question is how can it be possible to restore the exact state of photons in the brain while its state is collapsed in the retina. The other question which one may ask is: if this state is reported through action potentials how is this information reported to the brain and how can it interpret action potentials to obtain the exact state of the photons? Our solution to the above problems involves the teleportation of entangled coherent states through visual pathways. The state of the photon is teleported from the eye to the brain. On the other hand, the state of the photon is transferred via some "cut-and-paste" mechanism from the eye to the brain. But how is it possible?

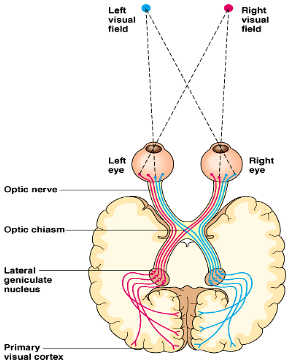

Figure 4 Visual pathways from the eye to the brain. See crossing of pathways.

We concluded before that retina and the visual cortex are entangled. Also we explained how the entangled coherent state is generated in the retina. Now, we want to formulate the process of information transfer from the retina to V1. The state (3-2) with two modes 1 and 2 should be teleported to V1. After the interaction of light with retina, modes 3, 4 and 5, 6 are generated through microtubular structures between retina and V1, and thus they produce entangled coherent channels between retina and V1. It means that the channels are:

$$|\psi\rangle_{35} = \frac{1}{\sqrt{N_\alpha}}\left(|\alpha\rangle_3|\alpha\rangle_5 - |-\alpha\rangle_3|-\alpha\rangle_5\right) \qquad (6\text{-}1)$$

$$|\phi\rangle_{46} = \frac{1}{\sqrt{N_\alpha}}\left(|\alpha\rangle_4|\alpha\rangle_6 - |-\alpha\rangle_4|-\alpha\rangle_6\right) \qquad (6\text{-}2)$$

Where N_α is the number of tubulins in each channel. Each mode is reported via a special fiber through visual pathways. All of the neurons which are collected in the LGN are divided into two major pathways: ipsilateral fibers and contralateral fibers. Information transfer in the contralateral fibers takes longer than information transfer in ipsilateral fibers because contralateral fibers have crossing relative to ipsilateral fibers and then they have longer lengths than ipsilateral fibers. On the other hand, contralateral fibers have a retarded phase relative to ipsilateral fibers. Now we want to answer the following questions. What is this phase difference? What is the role of this crossing? And how does crossing restore the initial state in the retina?

7) The Role of Phase Shift to Restore Information in LGN

When the information is collapsed in the retina, action potentials are produced. The shape of action potentials is the same for each neuron, but the main problem is which neurons are fired, or in other words which neurons carry action potentials and information. Consider two fibers selected from ipsilateral fibers and two fibers selected from contralateral fibers. The two ipsilateral fibers are called 1 and 2, and the two contralateral fibers are called 3 and 4 while the two fibers from the LGN to V1 are called 5 and 6 which are selected from the group of magnocellular and parvocellular fibers. Now, we start from the retina. The state of centrioles and channels is:

$$|\psi^{'}\rangle = |\varphi\rangle_{12} \otimes |\psi\rangle_{35} \otimes |\phi\rangle_{46}$$

$$= \frac{1}{N_{\propto}^{'}}(A|\alpha\rangle_1|\alpha\rangle_2|\alpha\rangle_3|\alpha\rangle_4|\alpha\rangle_5|\alpha\rangle_6$$

$$-A|\alpha\rangle_1|\alpha\rangle_2|\alpha\rangle_3|-\alpha\rangle_4|\alpha\rangle_5|-\alpha\rangle_6$$

$$-A|\alpha\rangle_1|\alpha\rangle_2|-\alpha\rangle_3|\alpha\rangle_4|-\alpha\rangle_5|\alpha\rangle_6$$

$$+A|\alpha\rangle_1|\alpha\rangle_2|-\alpha\rangle_3|-\alpha\rangle_4|-\alpha\rangle_5|-\alpha\rangle_6$$

$$-B|-\alpha\rangle_1|-\alpha\rangle_2|\alpha\rangle_3|\alpha\rangle_4|\alpha\rangle_5|\alpha\rangle_6$$

$$+B|-\alpha\rangle_1|-\alpha\rangle_2|\alpha\rangle_3|-\alpha\rangle_4|\alpha\rangle_5|-\alpha\rangle_6$$

$$+B|-\alpha\rangle_1|-\alpha\rangle_2|-\alpha\rangle_3|\alpha\rangle_4|-\alpha\rangle_5|\alpha\rangle_6$$

$$-B|-\alpha\rangle_1|-\alpha\rangle_2|-\alpha\rangle_3|-\alpha\rangle_4|-\alpha\rangle_5|-\alpha\rangle_6) \qquad (7\text{-}1)$$

All of the above states are collected in the LGN. But here the role of action potentials is very important. They determine which fibers are fired. If fibers 1 and 2 carry action potentials, then it shows that information passes through ipsilateral fibers. Thus to select information from the LGN to send it via fibers 5 and 6 to V1 there is no need for phase difference (or to apply the phase shift operator on the states) and thus the state of (3-2) can be transferred like its first state through fibers 5 and 6. In the formulation (7-1) we see that if fibers x and y are fired, the expression with $|\alpha\rangle_x|\alpha\rangle_y$ should be selected from the terms with coefficients +A besides the expression $|-\alpha\rangle_x|-\alpha\rangle_y$ from the terms with coefficients $-B$. Hence,

$$1,2\, firing \quad \xrightarrow{yields} \quad |\phi^{'}\rangle_{56} = A|\alpha\rangle_5|\alpha\rangle_6 - B|-\alpha\rangle_5|-\alpha\rangle_6 \qquad (7\text{-}2)$$

In another state, if fibers 1 and 4 are fired it means that one fiber is selected from ipsilateral fibers and the other is from contralateral fibers, thus they have phase difference with respect to each other. Hence,

$$1,4\, firing \quad \xrightarrow{yields} \quad |\phi^{'}\rangle_{56} = A|\alpha\rangle_5|-\alpha\rangle_6 - B|-\alpha\rangle_5|\alpha\rangle_6 \qquad (7\text{-}3)$$

To restore initial information, the operator

$$R(\varphi) = e^{-i\pi a^{\dagger}{}_{6}a_{6}}$$

should operate on the state in LGN in which fibers 1 and 4 have conveyed action potentials. This operator changes the ket $|\alpha\rangle_6$ to $|-\alpha\rangle_6$ and vice versa. It means that fiber 4 has a π radian phase difference with respect to fiber 1, and this phase difference can restore the exact state of the photon. If fibers 2 and 3 are fired, this yields:

$$2,3\ firing \quad \xrightarrow{yields} \quad |\phi'\rangle_{56} = A|-\alpha\rangle_5|\alpha\rangle_6 - B|\alpha\rangle_5|-\alpha\rangle_6 \qquad (7\text{-}4)$$

In this case the operator

$$R(\varphi) = e^{-i\pi a^{\dagger}{}_{5}a_{5}}$$

should be involved. For the case of 3 and 4 firing, this yields,

$$3,4\ firing \quad \xrightarrow{yields} \quad |\phi'\rangle_{56} = A|-\alpha\rangle_5|-\alpha\rangle_6 - B|\alpha\rangle_5|\alpha\rangle_6 \qquad (7\text{-}5)$$

in which case the operator

$$R(\varphi) = e^{-i\pi(a^{\dagger}{}_{5}a_{5}+a^{\dagger}{}_{6}a_{6})}$$

should be involved. In this case we see that the main path is that of ipsilateral fibers which are direct to each eye and fibers 3 and 4 both have a π radian phase difference with it. We also know that there are two LGNs and the left and right V1. Now, another question emerges. How do these two left and right parts in V1 can instantaneously receive information? To answer this question, we can say that the synaptic β-neurexin/neuroligin-1 adhesive protein complex is claimed to be a device mediating entanglement between the cytoskeletons of the cortical neurons. Thus the macroscopic coherent quantum state can extend through large brain cortical areas [58].

We see that crossing or rotation of neurons in the visual pathways has an important role in restoring information in the brain. Maybe rotations or crossings of neurons throughout the body are there for this very reason.

8) Discussion

In general, we can briefly summarize our approach by listing the following properties:

1- It can combine the Orch-OR model with Tegmark's approaches and the Thaheld conclusion in a compact physical model which is called Teleportation.
2- It investigates visual pathways from atomic scales to macroscopic scales. This approach includes classical descriptions as well as new answers to open questions.
3- It explains why the shape of action potentials stays the same. Classical models state that "sensations" are action potentials that reach the brain via sensory neurons, and "perception" is the awareness and interpretation of the sensation. It is reasonable to assume that the constant shape of action potentials cannot result in different profiles of information. Thus the shape of information should be due to neurons. In this approach MTs are the representatives of information carriers. In our approach action potentials just determine which neurons fire and which do not.
4- It describes why neurons cross at some point. This crossing causes a phase shift relative to a special pathway. In teleportation of entangled coherent states the phase shift operators can rebuild initial information.

5- It explains that how the inverted image on the retina is perceived in the brain as upright.
6- It can describe how different information can be simultaneously perceived as a binding nature of conscious experience. This can be done via quantum parallel processing.
7- It explains how the brain of the observer can receive quantum information from the environment. We can see that there still exists this possibility that the mind can play the main role in the measurement problem, and this is in accord with what London, Bauer, von Neumann and Wigner (initially) asserted.

9) Conclusions

In this paper we have theoretically demonstrated the plausibility of a quantum teleportation mechanism between the eye and the brain which can describe different aspects of visual processing through visual pathways. Our model covers both quantum and classical aspects of neuroscience. This mechanism can combine some features of the Orch-OR model with Tegmark's conclusions and Thaheld's belief in one general model.

Acknowledgement

V. Salari and M. Rahnama thank Dr. Vahid Sheybani, the head of Neuroscience Research Center, for his good instructions and helps.

References:

---

[1] Schrödinger. E, What is life? Cambridge University Press, London, 1944.
[2] Davies. P, Does quantum mechanics play a non-trivial role in life?, Biosystems, 78**,** 69, 2004.
[3] Matsuno. K, *Quantum and biological computation*, Biosystems, 35, 209-212, 1995.
[4] Penrose. R, *Shadows of the mind*, Oxford University Press, 1994.
[5] Matsuno. K and Paton. R. C, *Is there biology of quantum information?,* Biosystems, **55**, 39-46 2000.
[6] Vedral. V, *Entanglement hits the big time*, Nature, **425**, 28-9, 2003.
[7] Beck. F and Eccles. J.C, *Quantum aspects of brain activity and the role of consciousness*, Proc. Nat. Acad. Sci. USA, **89**, 11357-11362, 1992.
[8] Penrose. R, The Emperor's New Mind (Oxford University Press, Oxford, 1989); S.R., *Quantum coherence in MTs: a neural basis for emergent consciousness?*, Journal of Consciousness Studies, **1**, 98–118, 1994.
[9] Davies. P.C.W.s, *The Origin of Life* (Penguin, London, 2003); The Fifth Miracle (Simon & Schuster, New York, 1998)
[10] Tarlaci. S., NeuroQuantology 2003; 3: 292-294

[11] Thaheld. F. 2005 *Preprint* quant-ph/0509042
[12] Thaheld. F. 2006 *Preprint* quant-ph/060418

[13] Tegmark. M., Phys. Rev. E61, 4194-4206 (2000). It's preprint exists in [arXiv:quant-ph/9907009 v2].

[14] Penrose. R. and Hameroff. S.R., J. Conscious. Stud. **2**, 98(1995).
[15] Hameroff. S.R. and Penrose. R., J. Conscious. Stud. **3**, 36(1996).
[16] Hameroff. S.R., Philos. Trans. R. Soc. London Ser. A **356**, 1869(1998).
[17] Hameroff. S.R. ,NeuroQuantology , March 2007 , Vol 5 , Issue 1 , Page 1-8

[18] Dowling. J. E., *The Retina: An Approachable Part of the Brain*. Belknap Press of Hrward University Press, Cambridge. (1987).

[19] Portet. S., Tuszynski. J. A., Hogue. C. M. and Dixon. J. M., *Elastic vibrations in seamless microtubules*, European Biophysics Journal, 34, Number7, October 2005, 912-920.

[20] Gerry, C., Knight, P. "*Introductory Quantum Optics*", Cmbridge University Press,(2005).

---

[21] Hameroff, S.R., 1994. *Quantum coherence in MTs: a neural basis for emergent consciousness?* J. Conscious. Stud. 1, 98–118.

[22] Penrose, R., 1989. *The Emperor's New Mind*. Oxford UniversityPress, Oxford.

[23] Kandel, E. R., Siegelbaum, S. A. & Schwartz, J. H. 1991 *Synaptic transmission.* In Principles of neural science (ed. E. R. Kandel, J. H. Schwartz & T. M. Jessell), 3rd edn, pp. 121-134. New York: Elsevier. Karplus, M. & McCammon

[24] Jibu, M. *On a heuristic model of the coherent mechanism of the global reaction process of a group of cells. Bussei Kenkyuu (Mater. Phys. Res.)* **53**, 431-436. (In Japanese.) (1990)

[25] Marshall, I. N. *Consciousness and Bose-Einstein condensates*. New Ideas Psychol. **7**, 73, 83. (1989)

[26] Penrose, R. *Newton, quantum theory and reality*. In 300 years of gravity (ed. S. W. Hawking& W. Israel). Cambridge University Press. (1987).

[27] Bohm, D. & Hiley, B. J. 1993 *The undivided universe*. New York: Routledge.

[28] Jibu, M. & Yasue, K. 1995 *Quantum brain dynamics: an introduction*. Amsterdam: John Benjamins.

[29] Stapp, H. P. 1993 *Mind, matter and quantum mechanics*. Berlin: Springer.

[30] Hameroff. S. R., BioSystems **77** 119-136 . (2004).

[31] Mavromatos, N. E., Mershin, A. and Nanopoulos, D. V. (2002), "*QED-Cavity model of MTs implies dissipationless energy transfer and biological quantum teleportation*," http://arxiv.org/abs/quant-ph/0204021

[32] Del Giudice. E., Doglia. S., Milani. M. and Vitiello. G., Nucl. Phys. B275 (FS 17) (1986), 185.

[33] Del Giudice. E., Preparata. G. and Vitiello. G., Phys. Rev. Lett. 61 (1988), 1085.

[34] Dicke. R. H., Phys. Rev. 93 (1954) 99.

[35] McCall. S. L. and Hahn. E. L., Phys. Rev. Lett. 18 (1967) 908

[36] Hameroff. S. R., *Ultimate Computing: Biomolecular Consciousness and Nanotechnology* (North-Holland, Amsterdam, 1987), and Journal of Consciousness Studies, 1 (1994) 91.

[37] Bokkon. I., *Phosphene phenomenon: A new concept*, BioSystems 92(2008) 168-174.

[38] Clegg. J. S. Amer. J. Physiol. **246**, R133 (1984).

[39] Clegg. J. Coll. Phenomena **3**, 289 (1981).

[40] Miyamoto, S. "*Changes in mobility of synaptic vesicles with assembly and disassembly of actin network.*" *Biochim. Biophys. Acta* **1244**, 85-91. (1995).

[41] Muallem, S., Kwiatkowska, K., Xu, X. & Yin, H. L. *Actin filament disassembly is a sufficient final trigger for exocytosis in nonexcitable cells*. *J. Cell Biol.* **128**, 589-598. (1995).

[42] Wachsstock, D. H., Schwarz, W. H. & Pollard, T. D. *Cross-linker dynamics determine the mechanical properties of actin gels*. *Biophys. Jl* **66**, 801-809. (1994).

[43] Ouyang. M., Awschalom. D. D., " *coherent spin transfer between moleculary bridged quantum dots*". Science, 301, 1074-1078, (2003).

[44] Fröhlich. H., Int, J. Quantum Chem. **2**, 641 (1968).

[45] Fröhlich. H., Nature (London) **228**, 1093 (1970).

[46] Fröhlich. H., Proc. Natl. Acad. Sci. U.S.A. **72**, 4211 (1975).

[47] Bennett. C. H., Brassard. G., Crepeau. C., Jozsa. R., Peses. A., and Wootters. W. K., Phys. Rev. Lett. **70**, 1895 (1993).

[48] Wootters. W. K. and Zurek. W. H., Nature (London) **299**, 802 (1982); Dieks .D., Phys. Lett. A **92**, 271 (1982).

[49] Jibu, M., Hagan, S., Hameroff, S.R., Pribram, K.H., Yasue, K. (1994). "*Quantum optical coherenc in cytoskeletal MTs: implications for brain function*", Biosystems 32, 95-209.

[50] Hameroff. S., Tuszynski. J., "*Quantum states in proteins and protein assemblies: The essence of life?*", Proceedings of SPIE Conference on Fluctuationa and Noise, Canary Islands, June 2004

[51] D. D. Georgiev, arXiv: quant-ph/ 0208053

[52] Web address: *http://thebrain.mcgill.ca/.../i_09_cl_dev.html*

[53] Zeki, S. " *The visual image in mind and brain*", Scientific American, 267:42-50 (1993).

[54] Hagan. S., Hameroff. S. R., and Tuszynski. J. A., *Quantum computation in brain MTs: Decoherence and biological feasibility.* Physical Review E, Volume **65**, 061901. (2002).

[55] Thaheld. F. H., *A method to explore the possibility of nonlocal correlations between brain electrical activities of two spatially separated animal subjects.* BioSystems **73**, 205-216.(2004).

[56] Thaheld. F. H., *Can we determine if the linear nature of quantum mechanics is violated by the perceptual process?* BioSystems **71**, 305-309.(2003).

[57] Thaheld. F. H., Comment on *Quantum superpositions and definite perceptions: envisaging new feasible experimental tests.* Phys. Lett. A **273**,232-234.(2000).

[58] Georgiev. D. D., Glazebrook. J. F., Informatica **30,** 221–232 (2006).

DTIC Copy

AFRL-PR-ED-TR-2003-0034 AFRL-PR-ED-TR-2003-0034

Teleportation Physics Study

Eric W. Davis
Warp Drive Metrics
4849 San Rafael Ave.
Las Vegas, NV 89120

August 2004

Special Report

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
EDWARDS AIR FORCE BASE CA 93524-7048

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 25-11-2003	**2. REPORT TYPE** Special	**3. DATES COVERED** *(From - To)* 30 Jan 2001 – 28 Jul 2003

4. TITLE AND SUBTITLE

Teleportation Physics Study

5a. CONTRACT NUMBER
F04611-99-C-0025

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62500F

6. AUTHOR(S)

Eric W. Davis

5d. PROJECT NUMBER
4847

5e. TASK NUMBER
0159

5f. WORK UNIT NUMBER
549907

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Warp Drive Metrics
4849 San Rafael Ave.
Las Vegas, NV 89120

8. PERFORMING ORGANIZATION REPORT NO.

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory (AFMC)
AFRL/PRSP
10 E. Saturn Blvd.
Edwards AFB CA 93524-7680

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-PR-ED-TR-2003-0034

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This study was tasked with the purpose of collecting information describing the teleportation of material objects, providing a description of teleportation as it occurs in physics, its theoretical and experimental status, and a projection of potential applications. The study also consisted of a search for teleportation phenomena occurring naturally or under laboratory conditions that can be assembled into a model describing the conditions required to accomplish the transfer of objects. This included a review and documentation of quantum teleportation, its theoretical basis, technological development, and its potential applications. The characteristics of teleportation were defined and physical theories were evaluated in terms of their ability to completely describe the phenomena. Contemporary physics, as well as theories that presently challenge the current physics paradigm were investigated. The author identified and proposed two unique physics models for teleportation that are based on the manipulation of either the general relativistic spacetime metric or the spacetime vacuum electromagnetic (zero-point fluctuations) parameters. Naturally occurring anomalous teleportation phenomena that were previously studied by the United States and foreign governments were also documented in the study and are reviewed in the report. The author proposes an additional model for teleportation that is based on a combination of the experimental results from the previous government studies and advanced physics concepts. Numerous recommendations outlining proposals for further theoretical and experimental studies are given in the report. The report also includes an extensive teleportation bibliography.

15. SUBJECT TERMS
teleportation; physics, quantum teleportation; teleportation phenomena; anomalous teleportation; teleportation theories; teleportation experiments; teleportation bibliography

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Franklin B. Mead, Jr.
a. REPORT **Unclassified**	**b. ABSTRACT** **Unclassified**	**c. THIS PAGE** **Unclassified**	A	88	**19b. TELEPHONE NO** *(include area code)* (661) 275-5929

Standard Form 298
(Rev. 8-98)
Prescribed by ANSI Std. 239.18

NOTICE

USING GOVERNMENT DRAWINGS, SPECIFICATIONS, OR OTHER DATA INCLUDED IN THIS DOCUMENT FOR ANY PURPOSE OTHER THAN GOVERNMENT PROCUREMENT DOES NOT IN ANY WAY OBLIGATE THE US GOVERNMENT. THE FACT THAT THE GOVERNMENT FORMULATED OR SUPPLIED THE DRAWINGS, SPECIFICATIONS, OR OTHER DATA DOES NOT LICENSE THE HOLDER OR ANY OTHER PERSON OR CORPORATION; OR CONVEY ANY RIGHTS OR PERMISSION TO MANUFACTURE, USE, OR SELL ANY PATENTED INVENTION THAT MAY RELATE TO THEM.

FOREWORD

This Special Technical Report presents the results of a subcontracted study performed by Warp Drive Metrics, Las Vegas, NV, under Contract No. F04611-99-C-0025, for the Air Force Research Laboratory (AFRL)/Space and Missile Propulsion Division, Propellant Branch (PRSP), Edwards AFB, CA. The Project Manager for AFRL/PRSP was Dr. Franklin B. Mead, Jr.

This report has been reviewed and is approved for release and distribution in accordance with the distribution statement on the cover and on the SF Form 298. This report is published in the interest of scientific and technical information exchange and does not constitute approval or disapproval of its ideas or findings.

//Signed//

FRANKLIN B. MEAD, JR.
Project Manager

//Signed//

RONALD E. CHANNELL
Chief, Propellants Branch

//Signed//

PHILIP A. KESSEL
Technical Advisor, Space and Missile
Propulsion Division

//Signed// AFRL-ERS-PAS-04-155

RANNEY G. ADAMS III
Public Affairs Director

This Page Intentionally Left Blank

Table of Contents

Section **Page**

1.0 INTRODUCTION……1
  1.1 Introduction……1
  1.2 The Definitions of Teleportation……1
2.0 vm -TELEPORTATION……3
  2.1 Engineering the Spacetime Metric……3
    2.1.1 Wormhole Thin Shell Formalism……3
    2.1.2 "Exotic" Matter-Energy Requirements……11
  2.2 Engineering the Vacuum……11
    2.2.1 The Polarizable-Vacuum Representation of General Relativity……20
  2.3 Conclusion and Recommendations……26
3.0 q-TELEPORTATION……30
  3.1 Teleportation Scenario……30
  3.2 Quantum Teleportation……32
    3.2.1 Description of the q-Teleportation Process……34
    3.2.2 Decoherence Fundamentally Limits q-Teleportation……38
    3.2.3 Recent Developments in Entanglement and q-Teleportation Physics……38
  3.3 Conclusion and Recommendations……46
4.0 e-TELEPORTATION……50
  4.1 Extra Space Dimensions and Parallel Universes/Spaces……50
  4.2 Vacuum Hole Teleportation……52
  4.3 Conclusion and Recommendations……53
5.0 p-TELEPORTATION……55
  5.1 PK Phenomenon……55
    5.1.1 Hypothesis Based on Mathematical Geometry……60
  5.2 Conclusion and Recommendations……61
6.0 REFERENCES……63
APPENDIX A – A Few Words About Negative Energy……73
  A.1 A General Relativistic Definition of Negative or Exotic Energy……73
  A.2 Squeezed Quantum States and Negative Energy……73
APPENDIX B – THεμ Methodology……75

Approved for public release; distribution unlimited.

List of Figures

Figure **Page**

Figure 1. Diagram of a Simultaneous View of Two Remote Compact Regions (Ω_1 and Ω_2) of Minkowski Space Used to Create the Wormhole Throat $\partial\Omega$, Where Time is Suppressed in This Representation……………..…………………………..……………………………………5

Figure 2. The Same Diagram as in Figure 1 Except as Viewed by an Observer Sitting in Region Ω_1 Who Looks Through the Wormhole Throat $\partial\Omega$ and Sees Remote Region Ω_2 (Dotted Area Inside the Circle) on the Other Side………………………………………………………. 6

Figure 3. A Thin Shell of (Localized) Matter-Energy, or Rather the Two-Dimensional Spacelike Hypersurface $\partial\Omega$ (via (2.3)), Possessing the Two Principal Radii of Curvature ρ_1 and ρ_2……..8

Figure 4. A Schematic of Vacuum Quantum Field Fluctuations (a.k.a. Vacuum Zero Point Field Fluctuations) Involved in the "Light-by-Light" Scattering Process That Affects the Speed of Light……………………………………………………………………………………13

Figure 5. A Schematic of the Casimir Effect Cavity/Waveguide……………………………………...15

Figure 6. Classical Facsimile Transmission (Modified IBM Press Image)………………………………35

Figure 7. Quantum Teleportation (Modified IBM Press Image)…………………………………………36

Figure 8. Quantum Teleportation (From www.aip.org)..43

Approved for public release; distribution unlimited.

List of Tables

Table **Page**

Table 1. Metric Effects in the PV-GR Model When K > 1 (Compared With Reference Frames at Asymptotic Infinity Where K = 1)………………………………………………………..…21

Table 2. Metric Effects in the PV-GR Model When K < 1 (Compared With Reference Frames at Asymptotic Infinity Where K = 1)………………………………………………………...22

Table 3. Substantial Gravitational Squeezing Occurs When $\lambda \geq 8\pi r_s$ (For Electromagnetic ZPF)............28

Approved for public release; distribution unlimited.

Glossary

AEC	Average Energy Condition
AFRL	Air Force Research Laboratory
AU	Astronomical Unit
BBO	Beta (β)-Barium Borate
CGS	Centimeter-Gram-Second
CIA	Central Intelligence Agency
DARPA	Defense Advanced Research Projects Agency
DEC	Dominant Energy Condition
DIA	Defense Intelligence Agency
DNA	Deoxyribo Nucleic Acid
DoD	Department of Defense
EPR	Einstein, Podolsky and Rosen
ESP	Extrasensory Perception
eV	Electron Volt
FRW	Friedmann-Robertson-Walker
FTL	Faster-Than-Light
IBM	International Business Machines
INSCOM	Intelligence and Security Command
IR	Infrared
MeV	Mega-Electron Volt
MKS	Meter-Kilogram-Second
NEC	Null Energy Condition
NLP	Neuro-Linguistic Programming
NMR	Nuclear Magnetic Resonance
NSA	National Security Agency
PK	Psychokinesis
PPN	Parameterized Post-Newtonian
PRC	Peoples Republic of China
PV-GR	Polarizable-Vacuum Representation of General Relativity
QED	Quantum Electrodynamics
QISP	Quantum Information Science Program
R&D	Research and Development
SAIC	Science Applications International Corporation
SEC	Strong Energy Condition
SRI	Stanford Research Institute
USSR	Union of Soviet Socialist Republics
UV	Ultraviolet
WEC	Weak Energy Condition
ZPE	Zero-Point Energy
ZPF	Zero-Point Fluctuations

Acknowledgements

This study would not have been possible without the very generous support of Dr. Frank Mead, Senior Scientist at the Advanced Concepts Office of the U.S. Air Force Research Laboratory (AFRL) Propulsion Directorate at Edwards AFB, CA. Dr. Mead's collegial collaboration, ready assistance, and constant encouragement were invaluable to me. Dr. Mead's professionalism and excellent rapport with "out-of-the-box" thinkers excites and motivates serious exploration into advanced concepts that push the envelope of knowledge and discovery. The author owes a very large debt of gratitude and appreciation to both Dr. David Campbell, Program Manager, ERC, Inc. at AFRL, Edwards AFB, CA, and the ERC, Inc. staff, for supporting the project contract and for making all the paperwork fuss totally painless. Dr. Campbell and his staff provided timely assistance when the author needed it, which helped make this contract project run smoothly.

There are two colleagues who provided important contributions to this study that I wish to acknowledge. First, I would like to express my sincere thanks and deepest appreciation to my first longtime mentor and role model, the late Dr. Robert L. Forward. Bob Forward was the first to influence my interests in interstellar flight and advanced breakthrough physics concepts (i.e., "Future Magic") when I first met him at an AIAA Joint Propulsion Conference in Las Vegas while I was in high school (ca. 1978). The direction I took in life from that point forward followed the trail of exploration and discovery that was blazed by Bob. I will miss him, but I will never forget him. Second, I would like to express my sincere thanks and appreciation to my longtime friend, colleague and present mentor, Dr. Hal Puthoff, Institute for Advanced Studies-Austin, for our many discussions on applying his Polarizable Vacuum-General Relativity model to a quasi-classical teleportation concept. Hal taught me to expand my mind, and he encourages me to think outside the box. He also gave me a great deal of valuable insight and personal knowledge about the Remote Viewing Program. Last, I would like to offer my debt of gratitude and thanks to my business manager (and spouse), Lindsay K. Davis, for all the hard work she does to make the business end of Warp Drive Metrics run smoothly.

Eric W. Davis, Ph.D., FBIS
Warp Drive Metrics
Las Vegas, NV

Preface

The Teleportation Physics Study is divided into four phases. Phase I is a review and documentation of quantum teleportation, its theoretical basis, technological development, and its potential application. Phase II developed a textbook description of teleportation as it occurs in classical physics, explored its theoretical and experimental status, and projected its potential applications. Phase III consisted of a search for teleportation phenomena occurring naturally or under laboratory conditions that can be assembled into a model describing the conditions required to accomplish the disembodied conveyance of objects. The characteristics of teleportation were defined, and physical theories were evaluated in terms of their ability to completely describe the phenomenon. Presently accepted physics theories, as well as theories that challenge the current physics paradigm were investigated for completeness. The theories that provide the best chance of explaining teleportation were selected, and experiments with a high chance of accomplishing teleportation were identified. Phase IV is the final report.

The report contains five chapters. Chapter 1 is an overview of the textbook descriptions for the various teleportation phenomena that are found in nature, in theoretical physics concepts, and in experimental laboratory work. Chapter 2 proposes two quasi-classical physics concepts for teleportation: the first is based on engineering the spacetime metric to induce a traversable wormhole; the second is based on the polarizable-vacuum-general relativity approach that treats spacetime metric changes in terms of equivalent changes in the vacuum permittivity and permeability constants. These concepts are theoretically developed and presented. Promising laboratory experiments were identified and recommended for further research. Chapter 3 presents the current state-of-art of quantum teleportation physics, its theoretical basis, technological development, and its applications. Key theoretical, experimental, and applications breakthroughs were identified, and a series of theoretical and experimental research programs are proposed to solve technical problems and advance quantum teleportation physics. Chapter 4 gives an overview of alternative teleportation concepts that challenge the present physics paradigm. These concepts are based on the existence of parallel universes/spaces and/or extra space dimensions. The theoretical and experimental work that has been done to develop these concepts is reviewed, and a recommendation for further research is made. Last, Chapter 5 gives an in-depth overview of unusual teleportation phenomena that occur naturally and under laboratory conditions. The teleportation phenomenon discussed in the chapter is based on psychokinesis (PK), which is a category of psychotronics. The U.S. military-intelligence literature is reviewed, which relates the historical scientific research performed on PK-teleportation in the U.S., China and the former Soviet Union. The material discussed in the chapter largely challenges the current physics paradigm; however, extensive controlled and repeatable laboratory data exists to suggest that PK-teleportation is quite real and that it is controllable. The report ends with a combined list of references.

1.0 INTRODUCTION

1.1 Introduction

The concept of teleportation was originally developed during the Golden Age of 20th century science fiction literature by writers in need of a form of instantaneous disembodied transportation technology to support the plots of their stories. Teleportation has appeared in such SciFi literature classics as Algis Budry's *Rogue Moon* (Gold Medal Books, 1960), A. E. van Vogt's *World of Null-A* (Astounding Science Fiction, August 1945), and George Langelaan's *The Fly* (Playboy Magazine, June 1957). The Playboy Magazine short story led to a cottage industry of popular films decrying the horrors of scientific technology that exceeded mankind's wisdom: *The Fly* (1958), *Return of the Fly* (1959), *Curse of the Fly* (1965), *The Fly* (a 1986 remake), and *The Fly II* (1989). The teleportation concept has also appeared in episodes of popular television SciFi anthology series such as *The Twilight Zone* and *The Outer Limits*. But the most widely recognized pop-culture awareness of the teleportation concept began with the numerous *Star Trek* television and theatrical movie series of the past 39 years (beginning in 1964 with the first TV series pilot episode, *The Cage*), which are now an international entertainment and product franchise that was originally spawned by the late genius television writer-producer Gene Roddenberry. Because of *Star Trek* everyone in the world is familiar with the "transporter" device, which is used to teleport personnel and material from starship to starship or from ship to planet and vice versa at the speed of light. People or inanimate objects would be positioned on the transporter pad and become completely disintegrated by a beam with their atoms being patterned in a computer buffer and later converted into a beam that is directed toward the destination, and then reintegrated back into their original form (all without error!). "*Beam me up, Scotty*" is a familiar automobile bumper sticker or cry of exasperation that were popularly adopted from the series.

However, the late Dr. Robert L. Forward (2001) stated that modern hard-core SciFi literature, with the exception of the ongoing *Star Trek* franchise, has abandoned using the teleportation concept because writers believe that it has more to do with the realms of parapsychology/paranormal (a.k.a. psychic) and imaginative fantasy than with any realm of science. Beginning in the 1980s developments in quantum theory and general relativity physics have succeeded in pushing the envelope in exploring the reality of teleportation. A crescendo of scientific and popular literature appearing in the 1990s and as recently as 2003 has raised public awareness of the new technological possibilities offered by teleportation. As for the psychic aspect of teleportation, it became known to Dr. Forward and myself, along with several colleagues both inside and outside of government, that anomalous teleportation has been scientifically investigated and separately documented by the Department of Defense.

It has been recognized that extending the present research in quantum teleportation and developing alternative forms of teleportation physics would have a high payoff impact on communications and transportation technologies in the civilian and military sectors. It is the purpose of this study to explore the physics of teleportation and delineate its characteristics and performances, and to make recommendations for further studies in support of Air Force Advanced Concepts programs.

1.2 The Definitions of Teleportation

Before proceeding, it is necessary to give a definition for each of the teleportation concepts I have identified during the course of this study:

- *Teleportation* – SciFi: the disembodied transport of persons or inanimate objects across space by advanced (futuristic) technological means (adapted from Vaidman, 2001). We will call this *sf-Teleportation*, which will not be considered further in this study.

- *Teleportation* – psychic: the conveyance of persons or inanimate objects by psychic means. We will call this *p-Teleportation*.

- *Teleportation* – engineering the vacuum or spacetime metric: the conveyance of persons or inanimate objects across space by altering the properties of the spacetime vacuum, or by altering the spacetime metric (geometry). We will call this *vm-Teleportation*.

- *Teleportation* – quantum entanglement: the disembodied transport of the quantum state of a system and its correlations across space to another system, where *system* refers to any single or collective particles of matter or energy such as baryons (protons, neutrons, etc.), leptons (electrons, etc.), photons, atoms, ions, etc. We will call this *q-Teleportation*.

- *Teleportation* – exotic: the conveyance of persons or inanimate objects by transport through extra space dimensions or parallel universes. We will call this *e-Teleportation*.

We will examine each of these in detail in the following chapters and determine whether any of the above teleportation concepts encompass the instantaneous and or disembodied conveyance of objects through space.

2.0 vm-TELEPORTATION

2.1 Engineering the Spacetime Metric

A comprehensive literature search for vm-Teleportation within the genre of spacetime metric engineering yielded no results. No one in the general relativity community has thought to apply the Einstein field equation to determine whether there are solutions compatible with the concept of teleportation. Therefore, I will offer two solutions that I believe will satisfy the definition of vm-Teleportation. The first solution can be found from the class of traversable wormholes giving rise to what I call a true "stargate." A stargate is essentially a wormhole with a flat-face shape for the throat as opposed to the spherical-shaped throat of the Morris and Thorne (1988) traversable wormhole, which was derived from a spherically symmetric Lorentzian spacetime metric that prescribes the wormhole geometry (see also, Visser, 1995 for a complete review of traversable Lorentzian wormholes):

$$ds^2 = -e^{2\phi(r)}c^2dt^2 + [1 - b(r)/r]^{-1}dr^2 + r^2 d\Omega^2 \qquad (2.1),$$

where by inspection we can write the traversable wormhole metric tensor in the form

$$g_{\alpha\beta} = \begin{pmatrix} -e^{2\phi(r)} & 0 & 0 & 0 \\ 0 & [1-b(r)/r]^{-1} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2\sin^2\theta \end{pmatrix} \qquad (2.2)$$

using standard spherical coordinates, where c is the speed of light, $\alpha,\beta \equiv (0 = t, 1 = r, 2 = \theta, 3 = \varphi)$ are the time and space coordinate indices ($-\infty < t < \infty$; r: $2\pi r$ = circumference; $0 \leq \theta \leq \pi$; $0 \leq \varphi \leq 2\pi$), $d\Omega^2 = d\theta^2 + sin^2\theta d\varphi^2$, $\phi(r)$ is the freely specifiable redshift function that defines the proper time lapse through the wormhole throat, and $b(r)$ is the freely specifiable shape function that defines the wormhole throat's spatial (hypersurface) geometry. Such spacetimes are asymptotically flat. The Einstein field equation requires that a localized source of matter-energy be specified in order to determine the geometry that the source induces on the local spacetime. We can also work the Einstein equation backwards by specifying the local geometry in advance and then calculate the matter-energy source required to induce the desired geometry. The Einstein field equation thus relates the spacetime geometry terms comprised of the components of the metric tensor and their derivatives (a.k.a. the Einstein tensor) to the local matter-energy source terms comprised of the energy and stress-tension densities (a.k.a. the stress-energy tensor). The flat-face wormhole or stargate is derived in the following section.

2.1.1 Wormhole Thin Shell Formalism

The flat-face traversable wormhole solution is derived from the thin shell (a.k.a. junction condition or surface layer) formalism of the Einstein equations (Visser, 1989; see also, Misner, Thorne and Wheeler, 1973). We adapt Visser's (1989) development in the following discussion. The procedure is to take two copies of flat Minkowski space and remove from each identical regions of the form $\Omega \times \Re$, where Ω is a three-dimensional compact spacelike hypersurface and $\Re$ is a timelike straight line (time axis). Then identify these two incomplete spacetimes along the timelike boundaries $\partial\Omega \times \Re$. The resulting spacetime

Approved for public release; distribution unlimited.

is geodesically complete and possesses two asymptotically flat regions connected by a wormhole. The throat of the wormhole is just the junction $\partial\Omega$ (a two-dimensional space-like hypersurface) at which the two original Minkowski spaces are identified (see Figures 1 and 2).

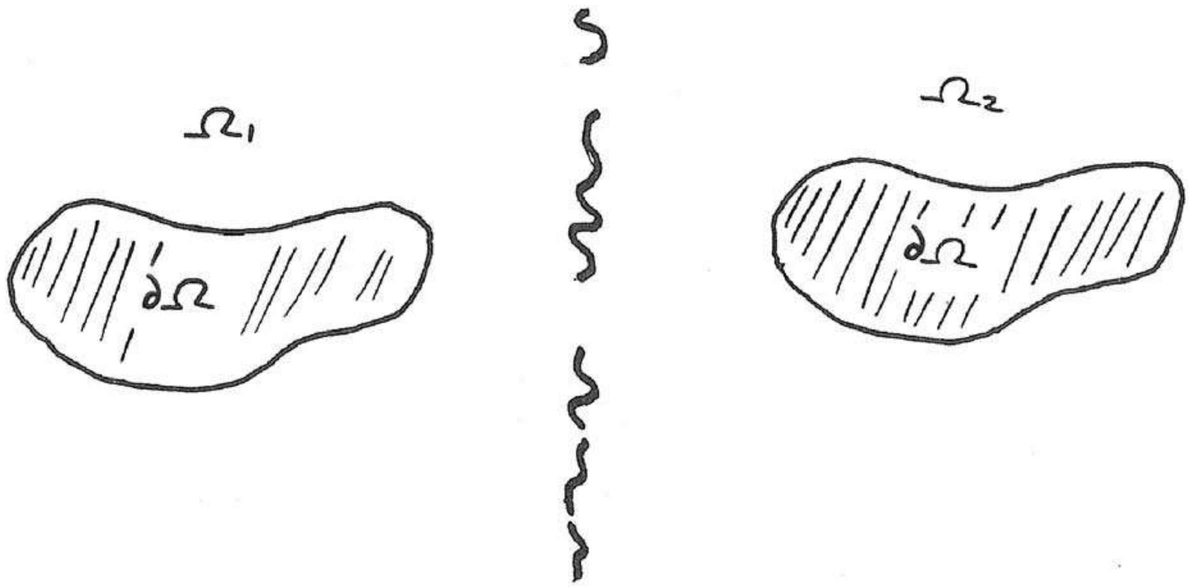
Ω_1
$\partial\Omega$
Ω_2
$\partial\Omega$

Figure 1. Diagram of a Simultaneous View of Two Remote Compact Regions (Ω_1 and Ω_2) of Minkowski Space Used to Create the Wormhole Throat $\partial\Omega$, Where Time is Suppressed in This Representation (adapted from Bennett et al., 1995)

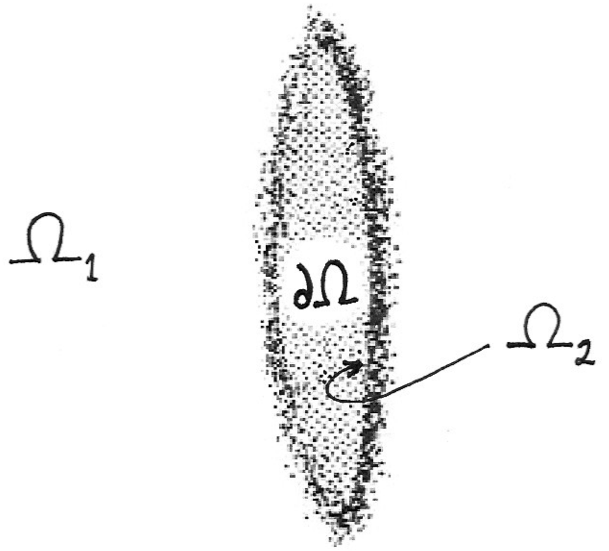

Figure 2. The Same Diagram as in Figure 1 Except as Viewed by an Observer Sitting in Region Ω_1 Who Looks Through the Wormhole Throat $\partial\Omega$ and Sees Remote Region Ω_2 (Dotted Area Inside the Circle) on the Other Side

The resulting spacetime is everywhere Riemann-flat except possibly at the throat. Also, the stress-energy tensor in this spacetime is concentrated at the throat with a δ-function singularity there. This is a consequence of the fact that the spacetime metric at the throat is continuous but not differentiable, while the connection is discontinuous; thus causing the Riemann curvature to possess a δ-function singularity (causing undesirable gravitational tidal forces) there. The magnitude of this δ-function singularity can be calculated in terms of the second fundamental form on both sides of the throat, which we presume to be generated by a localized thin shell of matter-energy. The second fundamental form represents the extrinsic curvature of the ∂Ω hypersurface (i.e., the wormhole throat), telling how it is curved with respect to the enveloping four-dimensional spacetime. The form of the geometry is simple, so the second fundamental form at the throat is calculated to be (McConnell, 1957):

$$\begin{aligned} K^{i}{}_{j}{}^{\pm} &= \pm \begin{pmatrix} \kappa_0 & 0 & 0 \\ 0 & \kappa_1 & 0 \\ 0 & 0 & \kappa_2 \end{pmatrix} \\ &= \pm \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1/\rho_1 & 0 \\ 0 & 0 & 1/\rho_2 \end{pmatrix} \end{aligned} \qquad (2.3),$$

where $i,j = 0,1,2$ and $K^{i\,\pm}_{j}$ is the second fundamental form. The full 4×4 matrix $K^{\alpha}{}_{\beta}$ has been reduced to 3×3 form, as above, for computational convenience because the thin shell (or hypersurface) is essentially a two-surface embedded in three-space. The overall ± sign in equation (2.3) comes from the fact that a unit normal points outward from one side of the surface and points inward on the other side. We hereafter drop the ± sign for the sake of brevity in notation. The quantities κ_0, κ_1, and κ_2 measure the extrinsic curvature of the thin shell of local matter-energy (i.e., the stuff that induces the wormhole throat geometry). Since the wormhole throat is a space-like hypersurface, we can exclude time-like hypersurfaces and their components in the calculations. Therefore we set $\kappa_0 = 0$ in equation (2.3) because it is the time-like extrinsic curvature for the time-like hypersurface of the thin shell of matter-energy. As seen in equation (2.3) κ_1 and κ_2 are simply related to the two principal radii of curvature ρ_1 and ρ_2 (defined to be the eigenvalues of K^{i}_{j}) of the two-dimensional spacelike hypersurface ∂Ω (see Figure 3). It should be noted that a convex surface has positive radii of curvature, while a concave surface has negative radii of curvature.

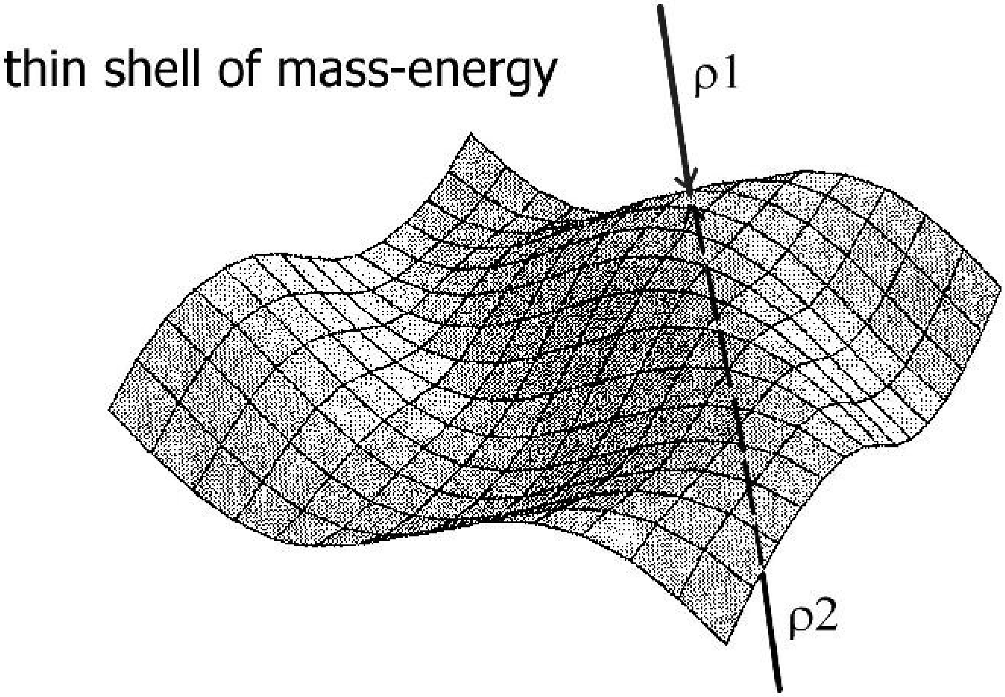
thin shell of mass-energy
ρ1
ρ2

Figure 3. A Thin Shell of (Localized) Matter-Energy, or Rather the Two-Dimensional Spacelike Hypersurface $\partial\Omega$ (via (2.3)), Possessing the Two Principal Radii of Curvature ρ_1 and ρ_2

It is a standard result of the thin shell or junction condition formalism that the Einstein field equation may be cast in terms of the surface stress-energy tensor S^i_j of the thin matter-energy shell localized in $\partial\Omega$ (note: we are exploiting the symmetry of the wormhole with respect to interchange of the two flat regions Ω_1 and Ω_2):

$$S^i{}_j = -\frac{c^4}{4\pi G}\left(K^i{}_j - \delta^i{}_j K^k{}_k\right) \quad (2.4),$$

where G is Newton's gravitational constant and δ^i_j is the (three-dimensional) unit matrix. K^k_k is the trace of equation (2.3):

$$\begin{aligned} K^k{}_k &= Tr\, K^i{}_j \\ &= \frac{1}{\rho_1} + \frac{1}{\rho_2} \end{aligned} \qquad (2.5)$$

and

$$\delta^i{}_j K^k{}_k = \begin{pmatrix} \frac{1}{\rho_1} + \frac{1}{\rho_2} & 0 & 0 \\ 0 & \frac{1}{\rho_1} + \frac{1}{\rho_2} & 0 \\ 0 & 0 & \frac{1}{\rho_1} + \frac{1}{\rho_2} \end{pmatrix} \quad (2.6).$$

Substituting (2.3) and (2.6) into (2.4) gives (after simplification):

$$S^i{}_j = \frac{c^4}{4\pi G}\begin{pmatrix} \frac{1}{\rho_1} + \frac{1}{\rho_2} & 0 & 0 \\ 0 & 1/\rho_2 & 0 \\ 0 & 0 & 1/\rho_1 \end{pmatrix} \quad (2.7).$$

The thin matter-energy shell's surface stress-energy tensor may be interpreted in terms of the surface energy density σ and principal surface tensions ϑ_1 and ϑ_2:

$$S^i{}_j = \begin{pmatrix} -\sigma & 0 & 0 \\ 0 & -\vartheta_1 & 0 \\ 0 & 0 & -\vartheta_2 \end{pmatrix} \quad (2.8).$$

Thus we arrive at the Einstein field equation by equating (2.8) and (2.7) and multiplying both sides by –1:

Approved for public release; distribution unlimited.

$$\begin{pmatrix} \sigma & 0 & 0 \\ 0 & \vartheta_1 & 0 \\ 0 & 0 & \vartheta_2 \end{pmatrix} = -\frac{c^4}{4\pi G} \begin{pmatrix} \dfrac{1}{\rho_1} + \dfrac{1}{\rho_2} & 0 & 0 \\ 0 & 1/\rho_2 & 0 \\ 0 & 0 & 1/\rho_1 \end{pmatrix} \qquad (2.9),$$

which gives the final result

$$\sigma = -\frac{c^4}{4\pi G}\left(\frac{1}{\rho_1} + \frac{1}{\rho_2}\right) \qquad (2.10a)$$

$$\vartheta_1 = -\frac{c^4}{4\pi G}\frac{1}{\rho_2} \qquad (2.10b)$$

$$\vartheta_2 = -\frac{c^4}{4\pi G}\frac{1}{\rho_1} \qquad (2.10c).$$

These are the Einstein equations. Equations (2.10a-c) imply that (for $\partial\Omega$ convex) we are dealing with ***negative*** surface energy density and ***negative*** surface tensions. This result is in fact the primary matter-energy requirement for traversable wormholes, as was proved by Morris and Thorne (1988), and later by Visser (1995), within the paradigm of classical Einstein general relativity. The negative surface tension (= *positive outward pressure*, a.k.a. gravitational repulsion or antigravity) is needed to keep the throat open and stable against collapse. The reader should **not** be alarmed at this result. Negative energies and negative stress-tensions are an acceptable result both mathematically and physically, and they manifest gravitational repulsion (antigravity!) in and around the wormhole throat. One only needs to understand what it means for stress-energy to be negative within the proper context. In general relativity the term "*exotic*" is used in place of "*negative*." The effects of negative energy have been produced in the laboratory (the Casimir Effect is one example). In short, negative energy arises from Heisenberg's quantum uncertainty principle, which requires that the energy density of any electromagnetic, magnetic, electric or other fields must fluctuate randomly. Even in a vacuum, where the average energy density is zero, the energy density fluctuates. This means that the quantum vacuum can never remain truly empty in the classical sense of the term. The quantum picture of the vacuum is that of a turbulent plenum of virtual (i.e., energy non-conserving) particle pairs that spontaneously pop in and out of existence. The notion of "zero energy" in quantum theory corresponds to the vacuum being filled with such fluctuations going on. This issue is further elaborated on and clarified in greater detail in Appendix A. We will also revisit this in Section 2.2. Finally, it should be noted that for the analysis in this section we assumed an ultrastatic wormhole [i.e., $g_{00} \equiv 1 \Rightarrow \phi(r) = 0$ in equation (2.1)] with the "exotic" matter-energy confined to a thin layer, and we dispensed with the assumption of spherical symmetry.

We can now build a wormhole-stargate and affect *vm-Teleportation* such that a traveler stepping into the throat encounters **no** exotic matter-energy there. This will require that our wormhole be flat shaped. To make the wormhole flat requires that we choose the throat $\partial\Omega$ to have at least one flat face (picture the thin shell in Figure 3 becoming a flat shell). On that face the two principal radii of curvature become $\rho_1 = \rho_2 = \infty$ as required by standard geometry. Substituting this into equations (2.10a-c) gives

$$\sigma = \vartheta_1 = \vartheta_2 = 0 \qquad (2.11),$$

Approved for public release; distribution unlimited.

which is a remarkable result. A further consequence of this is that now $K^i_j = 0$, thus making the Riemann curvature and stress-energy tensors (Riemann: $R^\alpha{}_\beta \sim K^\alpha{}_\beta$; stress-energy: $T^\alpha{}_\beta \sim K^\alpha{}_\beta$) at the throat become zero such that the associated δ-function singularities disappear there. This means that a traveler encountering and going through such a wormhole will feel no tidal gravitational forces and see no exotic matter-energy (that threads the throat). A traveler stepping through the throat will simply be ***teleported*** into the other remote spacetime region or another universe (note: the Einstein equation does not fix the spacetime topology, so it is possible that wormholes are inter-universe as well as intra-universe tunnels). We construct such a *teleportation* stargate by generating a thin shell or surface layer of "exotic" matter-energy much like a thin film of soap stretched across a loop of wire.

2.1.2 "Exotic" Matter-Energy Requirements

Now we have to estimate the amount of negative (or exotic) mass-energy that will be needed to generate and hold open a *vm-Teleportation* wormhole. A simple formula originally due to Visser (1995) for short-throat wormholes using the thin shell formalism gives:

$$\begin{aligned} M_{wh} &= -\frac{r_{throat}c^2}{G} \\ &= -(1.3469x10^{27}\,kg)\frac{r_{throat}}{1\,meter} \qquad (2.12), \\ &= -(0.709\,M_{Jupiter})\frac{r_{throat}}{1\,meter} \end{aligned}$$

where M_{wh} is the mass required to build the wormhole, r_{throat} is a suitable measure of the linear dimension (radius) of the throat, and $M_{Jupiter}$ is the mass of the planet Jupiter (1.90×10^{27} kg). Equation (2.12) demonstrates that a mass of $-0.709\ M_{Jupiter}$ (or -1.3469×10^{27} kg) will be required to build a wormhole 1 meter in size. As the wormhole size increases the mass requirement grows negative-large, and vice versa as the wormhole size decreases. After being alarmed by the magnitude of this, one should note that M_{wh} is **not** the total mass of the wormhole as seen by observers at remote distances. The non-linearity of the Einstein field equations dictates that the total mass is zero (actually, the total net mass being positive, negative or zero in the Newtonian approximation depending on the details of the negative energy configuration constituting the wormhole system). And finally, Visser et al. (2003) have demonstrated the existence of spacetime geometries containing traversable wormholes that are supported by ***arbitrarily small quantities*** of exotic matter-energy, and they proved that this was a general result. In Section 2.3 we will discuss how or whether we can create such a wormhole in the laboratory.

2.2 Engineering the Vacuum

Engineering the spacetime vacuum provides a second solution that also satisfies the definition of vm-Teleportation. The concept of "*engineering the vacuum*" was first introduced to the physics community by Lee (1988). Lee stated:

"*The experimental method to alter the properties of the vacuum may be called vacuum engineering…If indeed we are able to alter the vacuum, then we may encounter some new phenomena, totally unexpected.*"

This new concept is based on the now-accepted fact that the vacuum is characterized by physical parameters and structure that constitutes an energetic medium which pervades the entire extent of the universe. We note here the two most important defining properties of the vacuum in this regard (Puthoff et al., 2002):

- Within the context of quantum field theory the vacuum is the seat of all energetic particle and field fluctuations.

- Within the context of general relativity theory the vacuum is the seat of a spacetime structure (or metric) that encodes the distribution of matter and energy.

We begin our look into this concept by examining the propagation of light through space. We know from quantum field theory that light propagating through space interacts with the vacuum quantum fields (a.k.a. vacuum quantum field fluctuations). The observable properties of light, including the speed of light, are determined by these interactions. Vacuum quantum interactions with light lead to an effect on the speed of light that is due to the absorption of photons (by the vacuum) to form virtual electron-positron pairs followed by the quick re-emission (from the vacuum) of the photon (see Figure 4). The virtual particle pairs are very short lived because of the large mismatch between the energy of a photon and the rest mass-energy of the particle pair. A key point is that this process makes a contribution to the observed vacuum permittivity ε_0 (and permeability μ_0) constant and, therefore, to the speed of light c [$c = (\varepsilon_0\mu_0)^{-1/2}$].

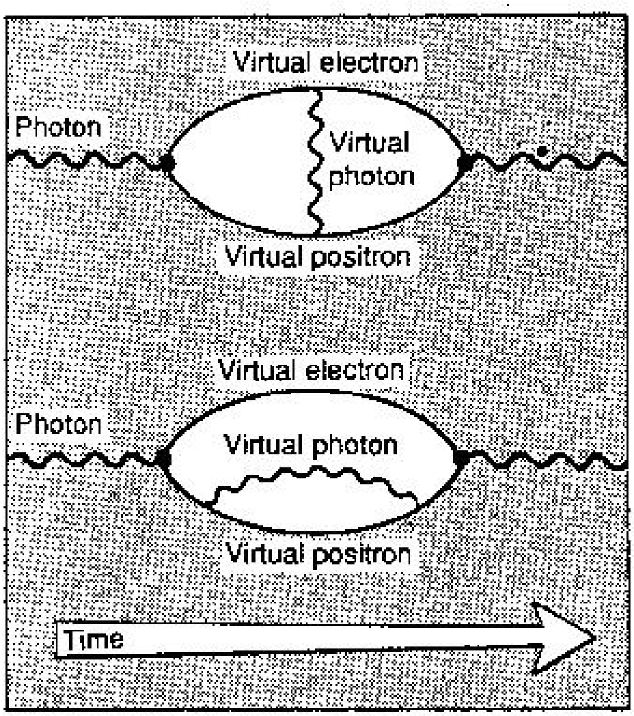
Virtual electron
Photon
Virtual photon
Virtual positron
Virtual electron
Photon
Virtual photon
Virtual positron
Time

Photons can change into a variety of short-lived, or "virtual", particles as they fly through the vacuum. The processes above, represented by Feynman diagrams, are "two-loop'" processes. They affect light's speed

Figure 4. A Schematic of Vacuum Quantum Field Fluctuations (a.k.a. Vacuum Zero Point Field Fluctuations) Involved in the "Light-by-Light" Scattering Process That Affects the Speed of Light (from Chown, 1990)

The role of virtual particle pairs in determining the ε_0 (μ_0) of the vacuum is analogous to that of atoms/molecules in determining the relative permittivity ε (and μ) of a dielectric material. We know that the absorption/re-emission of photons by atoms/molecules in a transparent medium (note: there are no strongly absorbing resonances, so the atoms/molecules remain in their excited states for a very short time before re-emitting photons) is responsible for the refractive index of the medium, which results in the reduction of the speed of light for photons propagating through the medium. This absorption/re-emission process is also known in physics as a scattering process. We know from experiment that a change in the medium leads to a change in ε (μ), thus resulting in a change of the refractive index. The key point arising from this analogy is that a modification of the vacuum produces a change in ε_0 (μ_0) resulting in a subsequent change in c, and hence, a corresponding change in the vacuum refraction index.

Scharnhorst (1990) and Latorre et al. (1995) have since proved that the suppression of light scattering by virtual particle pairs (a.k.a. coherent light-by-light scattering) in the vacuum causes an ***increase*** in the speed of light accompanied by a decrease in the vacuum refraction index. This very unique effect is accomplished in a Casimir Effect capacitor cavity (or waveguide) whereby the vacuum quantum field fluctuations (a.k.a. zero-point fluctuations or ZPF) inside have been modified (becoming anisotropic and non-translational invariant) to satisfy the electromagnetic boundary conditions imposed by the presence of the capacitor plates (or waveguide walls). The principal result of this modification is the removal of the electromagnetic zero-point energy (ZPE) due to the suppression of vacuum ZPE modes with wavelengths longer than the cavity/waveguide cutoff ($\lambda_0 = 2d$, where d = plate separation; see Figure 5). This removal of free space vacuum ZPE modes suppresses the scattering of light by virtual particle pairs, thus producing the speed of light increase (and corresponding decrease in the vacuum refraction index). We know from standard optical physics and quantum electrodynamics (QED) that the optical phase and group velocities can exceed c under certain physical conditions, but dispersion always ensures that the signal velocity is $\leq c$. But recent QED calculations (see, Scharnhorst, 1990 and Latorre et al., 1995) have proved that in the Casimir Effect system, the dispersive effects are *much weaker* still than those associated with the increase in c so that the phase, group and signal velocities will therefore *all increase* by the same amount. Note that, in general, no dispersion shows up in all of the modified vacuum effects examined by investigators.

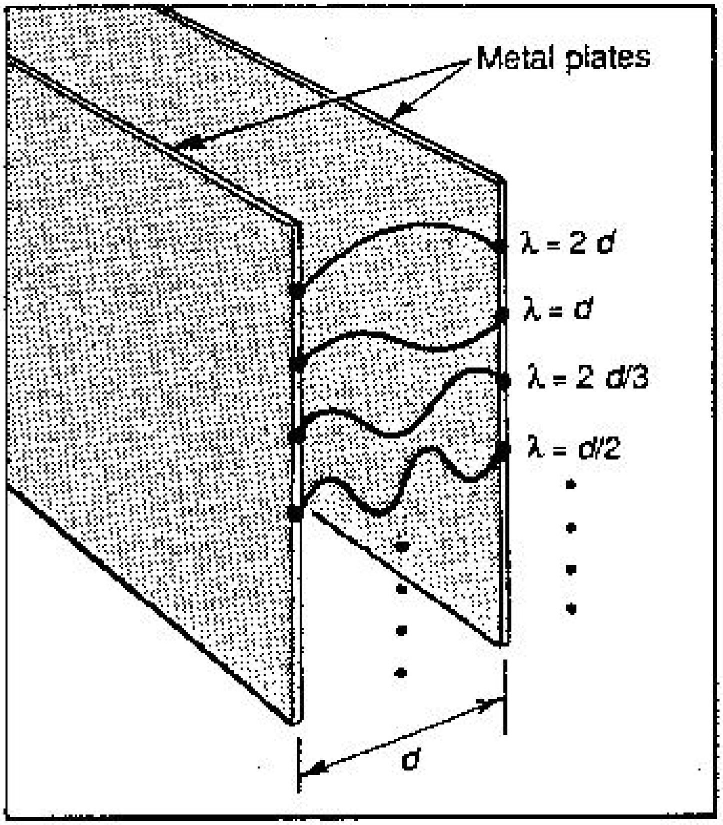
Metal plates
$\lambda = 2d$
$\lambda = d$
$\lambda = 2d/3$
$\lambda = d/2$
d

Casimir effect: the vacuum is full of virtual photons, but photons with wavelengths (λ), more than twice the separation of the plates, are excluded from the space between them. The imbalance pushes the plates together

Figure 5. A Schematic of the Casimir Effect Cavity/Waveguide (from Chown, 1990)

Examples demonstrating the increase in light speed (decrease in vacuum refraction index) via the Casimir Effect vacuum and other modified vacuum effects, as well as those effects producing a decrease in light speed (increase in vacuum refraction index), are described as follows. The vacuum modification effect on the speed of light described in the previous paragraph is (Scharnhorst, 1990):

$$\begin{aligned} \frac{c_\perp^{\;*}}{c_0} &= \left(1 + \frac{11}{2^6 \bullet (45)^2} \frac{e^4}{(m_e a)^4}\right) \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \\ &= \left(1 + \frac{11\pi^2}{8100} \alpha^2 \frac{1}{(m_e a)^4}\right) > 1 \end{aligned} \qquad (2.13),$$

where $c_\perp^{\;*}$ is the (modified) speed of light propagation perpendicular to the Casimir Effect capacitor plates, c_0 is the speed of light in free space (3×10^8 m/s in MKS units), m_e is the electron mass, α is the fine structure constant ($\approx 1/137$), e is the electron charge ($e^2 = 4\pi\alpha$ in quantum field theory natural units), a is the plate separation, $\hbar$ is Planck's reduced constant, and ε_0 is the vacuum permittivity constant. The condition $\hbar = c_0 = \varepsilon_0 = \mu_0 = 1$ stresses that (2.13), and all the equations that follow, are in quantum field theory natural units. The speed of light and vacuum refraction index measured parallel to the plates is unchanged from their free space values ($c_{||} = c_0$, $n_{||} = n_0 = 1$). The modified vacuum refraction index measured perpendicular to the plates is (Scharnhorst, 1990):

$$n_\perp = \left(1 - \frac{11}{2^6 \bullet (45)^2} \frac{e^4}{(m_e a)^4}\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \qquad (2.14).$$

Equations (2.13) and (2.14) show that in general $n_\perp < 1$ and $c_\perp^{\;*} > c_0$. But $c_\perp^{\;*} \to c_0$ and $n_\perp \to 1$ when $a \to \infty$ as expected, since we are now allowing all of the vacuum ZPE modes to re-enter the Casimir cavity in this case.

We now survey the additional examples of modified vacuums which increase/decrease light speed (from Latorre et al., 1995):

- For light (photons) propagating in a Friedmann-Robertson-Walker (FRW) vacuum (i.e., a homogeneous and isotropic Robertson-Walker gravitational background with Friedmann cosmology):

$$\frac{c^*}{c_0} = \left(1 + \frac{11}{45} \alpha G \frac{\rho_r + p}{m_e^{\;2}}\right) > 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \qquad (2.15),$$

where c^* is the modified vacuum speed of light, G is Newton's constant, ρ_r is the energy density and p is the pressure of a radiation-dominated universe ($p = \rho_r/3$). Here the speed of light is increased.

- For light (photons) propagating in a homogeneous and isotropic thermal vacuum:

$$\frac{c^*}{c_0} = \left(1 - \frac{44\pi^2}{2025} \alpha^2 \frac{T^4}{m_e^{\;4}}\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = k_B = 1) \qquad (2.16),$$

where T is the temperature of the vacuum and k_B is the Boltzmann constant. Here the speed of light is decreased.

- ❑ For light (photons) propagating in an anisotropic vacuum given by an external constant uniform magnetic field ***B***:

$$\frac{c_{\parallel}^{*}}{c_0} = \left(1 - \frac{8}{45}\alpha^2 \frac{\mathbf{B}^2}{m_e^4}\sin^2\theta\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1)$$

$$\frac{c_{\perp}^{*}}{c_0} = \left(1 - \frac{14}{45}\alpha^2 \frac{\mathbf{B}^2}{m_e^4}\sin^2\theta\right) < 1 \qquad (2.17),$$

where the speed of light is decreased in this vacuum for polarizations coplanar (||) with and perpendicular (⊥) to the plane defined by ***B*** and the direction of propagation, and θ is the angle between ***B*** and the direction of propagation. Latorre et al. (1995) calculated the polarization-average of (2.17) to give the averaged (modified) speed of light in the B-field:

$$\frac{c^{*}}{c_0} = \left(1 - \frac{22}{135}\alpha^2 \frac{\mathbf{B}^2}{m_e^4}\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \quad (2.18).$$

- ❑ For light (photons) propagating in an anisotropic vacuum given by an external constant uniform electric field ***E***, the polarization-averaged modified speed of light is:

$$\frac{c^{*}}{c_0} = \left(1 - \frac{22}{135}\alpha^2 \frac{\mathbf{E}^2}{m_e^4}\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \qquad (2.19).$$

Here the speed of light is decreased.

Equations (2.16) – (2.19) are the result of vacuum modifications that populate the vacuum with virtual or real particles that induce coherent (light-by-light) scattering, which reduces the speed of massless particles. By examining the form of equations (2.13) and (2.15) – (2.19) Latorre et al. (1995) discovered that the low energy modification of the speed of light is proportional to the ratio of the modified vacuum energy density (as compared to the standard vacuum energy density, $\rho_{vac} = 0$) over m_e^4, with a universal numerical coefficient and the corresponding coupling constants. And a general rule became apparent from their analysis that is applicable to modified vacua for massive and massless quantum field theories, for low energy:

$c^* > c_0$ (vacuum refraction index < 1) when the modified vacuum has a lower energy density
$c^* < c_0$ (vacuum refraction index > 1) when the modified vacuum has a higher energy density
$c^* = c_0$ (vacuum refraction index = 1) when the vacuum is free (or un-modified) with $\rho_{vac} = 0$

The first two rules explain the sign of the change of the speed of light. From this rule and the mathematical commonality between the form of (2.13) and (2.15) – (2.19) Latorre et al. (1995) found a single unifying expression to replace these equations:

$$\frac{c^{*}}{c_0} = 1 - \frac{44}{135}\alpha^2 \frac{\rho}{m_e^4} \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \qquad (2.20),$$

where ρ is the energy density of the modified vacua under consideration such that $\rho \rightarrow \rho_E \sim \boldsymbol{E}^2$ for the electric field vacuum, $\rho \rightarrow \rho_B \sim \boldsymbol{B}^2$ for the magnetic field vacuum, and $\rho \rightarrow \rho_T \sim \pi^2 T^4$ for the thermal vacuum. If the vacuum is a FRW gravitational vacuum, then one has to substitute one factor of α in (2.20) by $-m_e^2 G$ and $\rho \rightarrow \rho_r$. Equation (2.13) for the Casimir Effect vacuum studied earlier is recovered when $\rho \rightarrow \rho_{\text{Casimir}} = -(\pi^2/240)a^{-4}$.

Let us recast (2.20) into a more useful form. We subtract one from both sides of (2.20), do some algebra, and thus define the ratio of the change in the speed of light Δc in a modified vacuum to the speed of light in free space c_0:

$$\frac{c^*}{c_0} - 1 = \frac{c^* - c_0}{c_0} \equiv \frac{\Delta c}{c_0}$$

$$\frac{\Delta c}{c_0} = -\frac{44}{135}\alpha^2 \frac{\rho}{m_e^4} \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \quad (2.21).$$

Equations (2.20) and (2.21) are in quantum field theory natural units, which is completely undesirable for estimating physically measurable values of $\Delta c/c_0$. We thus transform or "unwrap" (2.20) and (2.21) back into MKS or CGS units by making the following substitutions (Puthoff, 2003)

$$\rho \text{ (natural units)} \rightarrow \frac{\rho}{\hbar c} \text{ (MKS or CGS units)}$$

$$m_e \text{ (natural units)} \rightarrow \frac{m_e c}{\hbar} \text{ (MKS or CGS units)},$$

and after some algebra and rearranging we arrive at the final result:

$$\frac{c^*}{c_0} = 1 - \frac{44}{135}\alpha^2 \frac{\rho}{m_e c_0^2}\left(\frac{\hbar}{m_e c_0}\right)^3 \qquad (2.22)$$

and

$$\frac{\Delta c}{c_0} = -\frac{44}{135}\alpha^2 \frac{\rho}{m_e c_0^2}\left(\frac{\hbar}{m_e c_0}\right)^3 \qquad (2.23),$$

where all quantities are now in MKS or CGS units. We chose the former units so that $c_0 = 3\times10^8$ m/s, $\hbar = 1.055\times10^{-34}$ J-s, $m_e = 9.11\times10^{-31}$ kg, and $\alpha = 1/137$. Note that the ratio of the modified vacuum energy density to the electron rest-mass energy has the dimension of (*volume*)$^{-1}$ while the quantity in the bracket is the cubed Compton wavelength of the electron having the dimension of (*volume*), and the product of these is dimensionless.

An excellent example for estimating the magnitude of the change in the speed of light (in a modified vacuum) is the Casimir Effect vacuum, since Casimir Effect experiments are common and widespread such that this would be ideal to experimentally test (2.23). We substitute the Casimir vacuum energy density $\rho_{\text{Casimir}} = -(\pi^2\hbar c_0/240)a^{-4}$ (in MKS units) into (2.23), do the algebra, insert the MKS values for the physical constants, and make further simplifications to get:

$$\begin{aligned}\frac{\Delta c}{c_0} &= -\frac{44}{135}\alpha^2\left(-\frac{\pi^2}{240}\frac{\hbar c_0}{a^4}\right)\frac{1}{m_e {c_0}^2}\left(\frac{\hbar}{m_e c_0}\right)^3 \\ &= \frac{11}{8100}\alpha^2\pi^2\left(\frac{\hbar}{m_e c_0 a}\right)^4 \qquad (2.24), \\ &\approx \left(1.59\times10^{-56}\right)a^{-4}\end{aligned}$$

where a (the plate separation) is in meters. Another useful equation is:

$$c^* = \left(1+\frac{\Delta c}{c_0}\right)c_0 \qquad (2.25),$$

where we make the substitution $c^* \to c_\perp^*$ for the present case. H. E. Puthoff and the author (Puthoff, 2003) compared the third line in (2.24) with equation (26) in Scharnhorst (1990) and discovered that the result cited there is in error, because the numerical coefficient is four orders of magnitude too small (Scharnhorst originally pointed out this error to Forward, 1996).

We now set $a = 10^{-6}$ m (1 μm) and we get $\Delta c/c_0 \approx 10^{-32}$ and $c_\perp^* \approx c_0$, which is a horrifically small 1 part in 10^{32} change that we cannot hope to measure at present. But for $a = 10^{-10}$ m (1 Å) we get $\Delta c/c_0 \approx 10^{-16}$ and $c_\perp^* \approx c_0$, which is a 1 part in 10^{16} change that could be measurable at present or in the very near future using high precision laser technology. Last, for $a = 1.1229\times10^{-14}$ m (11.229 fm or ≈ 11 times the nuclear diameter; 1 fm = 10^{-15} m) we find that $\Delta c/c_0 \approx 1$ and $c_\perp^* \approx 2c_0$. We are not able to do technical work at nuclear distances at this time; however, that could change as ultrahigh precision measurement technology continues to evolve. The threshold for the onset of significant changes in light speed occurs when $a < 10^{-12}$ m. This result is generally true for the other modified vacua surveyed in (2.15) – (2.19), since accessible (everyday) values for electric and magnetic field strengths, thermal temperatures and radiation densities are not large enough to overcome the size of the electron mass to create a measurable effect. However, there is a class of ultrahigh intensity tabletop lasers that have achieved such extreme electric and magnetic field strengths and temperatures that it may now be possible to consider using them to explore vacuum modification effects in the lab. We will return to this theme in a later section.

•**Key Point**: As disappointing as the Casimir Effect vacuum (and other modified vacua) results are, it should be strongly pointed out that special relativity theory says that if in one inertial reference frame an object travels only one part in 10^{16} (or even one part in 10^{32}) times faster than c_0, then one can find another reference frame where departure and arrival times of the object are simultaneous, and thus the velocity is ***infinite***. This is what motivates us to look at a teleportation mechanism based on engineering of the vacuum.

•**Technical Notes**:

- Equation (2.15) is interpreted as an increase in the speed of light due to a decrease in the number of vacuum ZPE modes. However, this effect is totally unrelated to light-by-light scattering in the vacuum because the gravitational background "squeezes" (as in squeezed quantum optics states; see Davis, 1999a) the ZPE modes, therefore reducing the vacuum energy density. We further note that the coefficient of 11 is the same for the gravitational vacuum as for the other modified vacua examples based on QED. This factor also appears in the coefficient of the Euler-Poincare characteristic spin-½ contribution to the gravitational trace anomaly (Birrell and Davies, 1982). It is beyond the scope of this study to consider the deep connections between quantum field theory and gravitation.

- We have excluded from our survey the Latorre et al. (1995) results pertaining to all other (high or low energy) modifications of the speed of massless particles. That is because the other examples invoked different QED theories possessing massless ($m_e = 0$), massive and intrinsic mass scales that introduced complex correction terms (beyond the leading low energy terms surveyed above) which are mass-related or running mass-related, and they introduced no new speed modification effects (beyond the low energy electron-positron virtual pair contributions); or no genuine speed modification was possible (especially for the massless Quantum Chromodynamic sector involving pseudo-Goldstone particles).

- There is ongoing (very noisy) controversy within the physics community over the effects of $c^* > c_0$ on causality. As this topic is beyond the scope of this study, I will make three points in this regard: 1) There are no grounds for microcausality violations in accordance with Drummond and Hathrell (1980). 2) A new definition of causality is in order for FTL (faster-than-light) phenomena. 3) Investigators have found that time machines (a.k.a. closed timelike curves) do not affect Gauss's theorem, and thus do not affect the derivation of global conservation laws from differential ones (Friedman et al., 1990). The standard conservation laws remain globally valid while retaining a natural quasi-local interpretation for spacetimes possessing time machines (for example, asymptotically flat wormhole spacetimes). Thorne (1993) states that it may turn out that causality is violated at the macroscopic scale. Even if causality is obeyed macroscopically, then quantum gravity might offer finite probability amplitudes for microscopic spacetime histories possessing time machines. Li and Gott (1998) found a self-consistent vacuum for quantum fields in Misner space (a simple flat space with closed timelike curves), for which the renormalized stress-energy tensor is regular (in fact zero) everywhere. This implies that closed timelike curves could exist at least at the level of semi-classical quantum gravity theory. Therefore, FTL causality paradoxes are just a reflection of our ignorance or inadequate comprehension of the physics of chronology and causality.

In this section we have shown how "vacuum engineering" can modify the speed of light, and how this can, in principle, lead to vm-Teleportation. The vacuum modification concepts summarized above lead us to a formal theory that implements the concept of vacuum engineering within a framework that parallels general relativity theory. This theory is called the Polarizable-Vacuum Representation of General Relativity. In the next section we will introduce and summarize this theory.

2.2.1 The Polarizable-Vacuum Representation of General Relativity

The polarizable-vacuum representation of general relativity (a.k.a. PV-GR) treats the vacuum as a polarizable medium of variable refractive index (Puthoff, 1999a, 2002a, b; Puthoff et al., 2002) exemplifying the concept of the vacuum modification (or vacuum engineering) effects surveyed and discussed in the previous section. The PV-GR approach treats spacetime metric changes in terms of equivalent changes in the vacuum permittivity and permeability constants (ε_0 and μ_0), essentially along the lines of the "$TH\varepsilon\mu$" methodology (see Appendix B for a brief description of this) used in comparative studies of alternative metric theories of gravity (Lightman and Lee, 1973; Will, 1974, 1989, 1993; Haugan and Will, 1977). Such an approach, relying as it does on parameters familiar to engineers, can be considered a "metric engineering" approach. Maxwell's equations in curved space are treated in the isomorphism of a polarizable medium of variable refractive index in flat space (Volkov et al., 1971); the bending of a light ray near a massive body is modeled as due to an induced spatial variation in the refractive index of the vacuum near the body; the reduction in the velocity of light in a gravitational potential is represented by an effective increase in the refractive index of the vacuum, and so forth. This optical-engineering approach has been shown to be quite general (de Felice, 1971; Evans et al., 1996a, b).

As recently elaborated by Puthoff (1999a, 2002a, b; Puthoff et al., 2002) the PV-GR approach, which was first introduced by Wilson (1921) and then developed by Dicke (1957, 1961), can be carried out in a self-consistent way so as to reproduce to appropriate order both the equations of general relativity and the match to the standard astrophysics weak-field experimental (PPN parameters and other) tests of those equations while posing testable modifications for strong-field conditions. It is in application that the PV-GR approach demonstrates its intuitive appeal and provides additional insight into what is meant by a curved spacetime metric.

Specifically, the PV-GR approach treats such measures as the speed of light, the length of rulers (atomic bond lengths), the frequency of clocks, particle masses, and so forth, in terms of a variable vacuum dielectric constant K in which the vacuum permittivity ε_0 transforms as $\varepsilon_0 \rightarrow K\varepsilon_0$ and the vacuum permeability transforms as $\mu_0 \rightarrow K\mu_0$ (see also, Rucker, 1977). In a planetary or solar gravitational potential $K = exp(2GM/rc_0^2) > 1$ (M is a local mass distribution, r is the radial distance from the center of M) while $K = 1$ in "empty" or free asymptotic space (Puthoff, 1999a, 2002a, b; Puthoff et al., 2002). In the former case, the speed of light is reduced, light emitted from an atom is redshifted as compared with a remote static atom (where $K = 1$), clocks run slower, objects/rulers shrink, etc. See Table 1.

Table 1. Metric Effects in the PV-GR Model When $K > 1$ (Compared With Reference Frames at Asymptotic Infinity Where $K = 1$; adapted from Puthoff et al., 2002**)**

Variable	**Determining Equation** (subscript 0 is asymptotic value where $K = 1$)	$K > 1$ *(typical mass distribution, M)*
modified speed of light $c^*(K)$	$c^* = c_0/K$	speed of light < c_0
Modified mass $m(K)$	$m = m_0K^{3/2}$	effective mass increases
modified frequency $\omega(K)$	$\omega = \omega_0K^{-1/2}$	redshift toward lower frequencies
modified time interval $\Delta t(K)$	$\Delta t = \Delta t_0K^{1/2}$	clocks run slower
modified energy $E(K)$	$E = E_0K^{-1/2}$	lower energy states
Modified length $L(K)$	$L = L_0K^{-1/2}$	objects/rulers shrink
dielectric-vacuum "gravitational" forces $F(K)$	$F(K) \propto \nabla K$	attractive gravitational force

When $K = 1$ we have the condition that $c^* = c_0$ (vacuum refraction index = 1), because the vacuum is free (or un-modified, and $\rho_{vac} = 0$) in this case. When $K > 1$, as occurs in a region of space possessing a gravitational potential, then we have the condition that $c^* < c_0$ (vacuum refraction index > 1), because the modified vacuum has a higher energy density in the presence of the local mass distribution that generates the local gravitational field. This fact allows us to make a direct correspondence between the speed of light modification physics discussion in Section 2.2 and the underlying basis for the physics of the PV-GR model. Under certain conditions the spacetime metric can in principle be modified to reduce the value of K to below unity, thus allowing for faster-than-light (FTL) motion to be physically realized. In this case, the local speed of light (as measured by remote static observers) is increased, light emitted from an atom is blueshifted as compared with a remote static atom, objects/rulers expand, clocks run faster, etc. See Table 2. We therefore have the condition that $c^* > c_0$ (vacuum refraction index < 1) because the modified vacuum has a lower energy density. In fact, Puthoff (1999a, 2002a) has analyzed certain special black hole metrics and found $K < 1$ from the model. We will return to this theme later. In what follows we briefly review and summarize the key points and equations from the development of the PV-GR model, and we refer the reader to Puthoff (1999a, 2002a, b) for more extensive discussion and derivations.

Table 2. Metric Effects in the PV-GR Model When $K < 1$ (Compared With Reference Frames at Asymptotic Infinity Where $K = 1$; adapted from Puthoff et al., 2002**)**

Variable	**Determining Equation** (subscript 0 is asymptotic value where $K = 1$)	$K < 1$ *(typical mass distribution, M)*
modified speed of light $c^*(K)$	$c^* = c_0/K$	speed of light > c_0
modified mass $m(K)$	$m = m_0 K^{3/2}$	effective mass decreases
modified frequency $\omega(K)$	$\omega = \omega_0 K^{-1/2}$	blueshift toward higher frequencies
modified time interval $\Delta t(K)$	$\Delta t = \Delta t_0 K^{1/2}$	clocks run faster
modified energy $E(K)$	$E = E_0 K^{-1/2}$	higher energy states
modified length $L(K)$	$L = L_0 K^{-1/2}$	objects/rulers expand
dielectric-vacuum "gravitational" forces $F(K)$	$F(K) \propto \nabla K$	repulsive gravitational force

We begin by recalling that in flat space electrodynamics, the electric flux vector ***D*** in a linear, homogeneous medium can be written

$$\begin{aligned}\mathbf{D} &= \varepsilon\mathbf{E} \\ &= \varepsilon_0\mathbf{E} + \mathbf{P} \qquad (2.26), \\ &= \varepsilon_0\mathbf{E} + \alpha_V\mathbf{E}\end{aligned}$$

where ε is the permittivity of the medium, the polarization ***P*** corresponds to the induced dipole moment per unit volume in the medium whose polarizability per unit volume is α_V, and ***E*** is the electric field. The identical form of the last two terms naturally leads to the interpretation of ε_0 as the polarizability per unit volume of the vacuum. The quantum picture of the vacuum, where it has been shown that the vacuum acts as a polarizable medium by virtue of induced dipole moments resulting from the excitation of virtual electron-positron particle pairs (Heitler, 1954), completely justifies the interpretation that the vacuum is a medium. Note that there are other virtual particle pairs in the vacuum that also contribute to this picture; however, it is the electron-positron pairs that dominate the others, as shown in Section 2.2. The basic postulate of the PV-GR model for curved space conditions is that the polarizability of the vacuum in the vicinity of localized mass-energy distributions differs from its asymptotic free space value by virtue of vacuum polarization effects induced by the presence of the local mass-energy. Thus the postulate for the vacuum itself is

$$\begin{aligned}\mathbf{D} &= \varepsilon\mathbf{E} \\ &\equiv K\varepsilon_0\mathbf{E}\end{aligned} \qquad (2.27),$$

where K (a function of position) is the modified dielectric constant of the vacuum due to the induced vacuum polarizability changes under consideration. Equation (2.27) defines the transformation $\varepsilon = K\varepsilon_0$.

Table 1 shows the various quantitative effects a polarizable vacuum (in the presence of positive mass-energy distributions) has on the various measurement processes important to general relativity. The effects demonstrated in the middle and right columns demonstrate the basis of the polarizable vacuum approach to general relativity. Table 2 shows what effects are manifested when negative mass-energy distributions induce vacuum polarizability changes that lead to FTL phenomenon. Experimental observations impose constraints on the model causing key physical constants to remain constant even with variable polarizability present in the local space. Puthoff (1999a, 2002a, b) has shown that the fine structure constant is constrained by observational data to remain constant within a variable polarizable vacuum, and this constraint actually defines the transformation $\mu = K\mu_0$. The elementary particle charge e is also taken to be constant in a variable polarizable vacuum because of charge conservation. And $\hbar$ remains a constant by conservation of angular momentum for circularly polarized photons propagating through the (variable polarizability) vacuum. The remaining constant of nature is the speed of light, and although the tables showed how this was modified in variable polarizability vacuums, it is interesting to see how this modification comes about. In a modified (variable polarizability) vacuum the speed of light is defined, as it is in standard electrodynamics, in terms of the permittivity and permeability by:

$$\begin{aligned} c^* &\equiv \left(\varepsilon\mu\right)^{-1/2} \\ &= \left(K\varepsilon_0 \bullet K\mu_0\right)^{-1/2} \\ &= \left(K^2\varepsilon_0\mu_0\right)^{-1/2} \\ &= \frac{1}{K}\left(\varepsilon_0\mu_0\right)^{-1/2} \\ &= \frac{c_0}{K} \end{aligned} \qquad (2.28),$$

where the permittivity/permeability transformations and the free space (un-modified vacuum) definition for c_0 were inserted. Note that (2.28) can be re-written as $c^*/c_0 = 1/K$, and this is to be compared with (2.22). Thus we see from (2.28), and by comparison with (2.22), that K plays the role of a variable refractive index under conditions in which the vacuum polarizability is assumed to change in response to general relativistic-type influences. One further note of interest is that the permittivity/permeability transformations also maintains constant the ratio

$$\sqrt{\frac{\mu}{\varepsilon}} = \sqrt{\frac{\mu_0}{\varepsilon_0}},$$

which is the impedance of free space. This constant ratio is required to keep electric-to-magnetic energy ratios constant during adiabatic movement of atoms from one position in space to another of differing vacuum polarizability (Dicke, 1957, 1961). And this constant ratio is also a necessary condition in the $TH\varepsilon\mu$ formalism for an electromagnetic test particle to fall in a gravitational field with a composition-independent acceleration (Lightman and Lee, 1973; Will, 1974, 1989, 1993; Haugan and Will, 1977).

Now we make the "crossover connection" to the standard spacetime metric tensor concept that characterizes conventional general relativity theory, as originally shown by Puthoff (1999a, 2002a, b). In flat (un-modified or free) space the standard four-dimensional infinitesimal spacetime interval ds^2 is given (in Cartesian coordinates with subscript 0) by

$$ds^2 = -c_0{}^2 dt_0{}^2 + \sum_{i=1}^{3} dx_{i0}{}^2 \qquad (2.29),$$

where $i \equiv (1 = x,\ 2 = y,\ 3 = z)$. This metric means that measuring rods and clocks are non-varying wherever one goes in spacetime to make measurements. However, this has been shown to be incorrect in general relativity theory, so the length and time transformations (between proper and coordinate values) given in the tables (middle columns) indicate that measuring rods and clocks do vary when placed in regions where $K \neq 1$. Therefore, we replace the time and space differentials in (2.29) with the length and time transformations in the tables into (2.29), and derive the general relativistic spacetime interval

$$ds^2 = -\frac{1}{K} c_0{}^2 dt^2 + K\left(\sum_{i=1}^{3} dx_i{}^2\right) \qquad (2.30).$$

Note that observers within a $K \neq 1$ region will always measure the speed of light to be c_0. Equation (2.30) defines an isotropic coordinate system, which is a common and useful way to represent spacetime metrics in general relativity studies. By inspection the metric tensor is written

$$g_{\alpha\beta} = \begin{pmatrix} -1/K & 0 & 0 & 0 \\ 0 & K & 0 & 0 \\ 0 & 0 & K & 0 \\ 0 & 0 & 0 & K \end{pmatrix} \qquad (2.31).$$

The Lagrangian density for matter-field interactions in a vacuum of variable K is given by Puthoff (1999a, 2002a, b) as

$$\begin{aligned} L_d = &-\left[\frac{m_0 c_0{}^2}{\sqrt{K}}\sqrt{1-\left(\frac{v}{c_0/K}\right)^2} + q\Phi - qA_i v^i\right]\delta^3(\mathbf{r}-\mathbf{r}_0) \\ &-\frac{1}{2}\left(\frac{\mathbf{B}^2}{K\mu_0} - K\varepsilon_0\mathbf{E}^2\right) - \frac{c_0{}^4}{32\pi G}\frac{1}{K^2}\left[(\nabla K)^2 - \frac{1}{(c_0/K)^2}\left(\frac{\partial K}{\partial t}\right)^2\right] \end{aligned} \qquad (2.32),$$

where the first term is the Lagrangian density for a free particle of mass m_0, charge q and 3-vector velocity $\boldsymbol{v}$ ($v = |\boldsymbol{v}|$, 3-vector components are labeled by i) interacting with electromagnetic fields via the electromagnetic field 4-vector potential $\boldsymbol{A}_\mu = (\Phi, A_i)$ (note that $\delta^3(\mathbf{r} - \mathbf{r}_0)$ is the delta function that locates the point particle at position $\mathbf{r} = \mathbf{r}_0$); the second term is the Lagrangian density for the electromagnetic fields themselves, and the last term is the Lagrangian density for K (treated here as a scalar variable). This last term emulates the Lagrangian density for the gravitational field. Equation (2.32) does not include any quantum gauge field interaction terms because it is beyond the scope of the present incarnation of the PV-GR approach to include them. We can obtain the equations of particle motion in a variable dielectric vacuum by performing the standard variations of the Lagrangian density $\delta(\int L_d\, dx\, dy\, dz\, dt)$ with respect to the particle variables. However, we are more interested in obtaining the "master equation" for K by varying the Lagrangian density with respect to K, and Puthoff (1999a, 2002a, b) gives the result:

$$\nabla^2\sqrt{K}-\frac{1}{\left(c_0/K\right)^2}\frac{\partial^2\sqrt{K}}{\partial t^2}$$

$$=-\frac{8\pi G}{c_0{}^4}\sqrt{K}\left\{\frac{\left(m_0c_0{}^2/\sqrt{K}\right)}{\sqrt{1-\left(\dfrac{v}{c_0/K}\right)^2}}\frac{1}{2}\left[1+\left(\frac{v}{c_0/K}\right)^2\right]\delta^3(\mathbf{r}-\mathbf{r}_0)\right. \qquad (2.33).$$

$$\left.+\frac{1}{2}\left(\frac{\mathbf{B}^2}{K\mu_0}+K\varepsilon_0\mathbf{E}^2\right)-\frac{c_0{}^4}{32\pi G}\frac{1}{K^2}\left[(\nabla K)^2+\frac{1}{\left(c_0/K\right)^2}\left(\frac{\partial K}{\partial t}\right)^2\right]\right\}$$

This equation describes the generation of general relativistic vacuum polarization effects due to the presence of matter and fields. By inspecting the right-hand side of the equation, we observe that changes in K are driven by the mass density (1st term), electromagnetic energy density (2nd term), and the vacuum polarization energy density itself (3rd term). In fact, the 3rd term emulates the gravitational field self-energy density. Note that the 2nd and 3rd terms in (2.33) appear with opposite signs with the result that electromagnetic field effects can counteract the gravitational field effects. Puthoff found that (2.33) gives the solution $K = exp(2GM/rc_0{}^2)$ in the vicinity of a static spherically symmetric (uncharged) mass M (in the low velocity limit $v << c_0$, $\partial K/\partial t = 0$, $\boldsymbol{E} = \boldsymbol{B} = 0$, $q = 0$), which reproduces to appropriate order the standard general relativistic Schwarzschild spacetime metric for the weak gravitational field conditions prevailing in the solar system. This solution guarantees that $K > 1$ near mass concentrations.

Of major importance to the present study are solutions giving $K < 1$ so that teleportation can be realized. Puthoff has found one such solution by studying the case of a static spherically symmetric mass M with charge Q familiar from the study of the Reissner-Nordstrϕm spacetime metric. In this case Puthoff found the result

$$K=\left[\cos\left(\frac{\sqrt{b^2-a^2}}{r}\right)+\frac{a}{\sqrt{b^2-a^2}}\sin\left(\frac{\sqrt{b^2-a^2}}{r}\right)\right]^2 \qquad \left(b^2>a^2\right) \qquad (2.34),$$

where $a^2 = (GM/c_0{}^2)^2$, $b^2 = Q^2G/4\pi\varepsilon_0 c_0{}^4$, and r is the radial distance from the center of M. And in this case (2.34) gives $K < 1$, which shows that FTL solutions are available in the PV-GR approach (as they are also in the Einstein theory). (For $a^2 > b^2$ the solution is hyperbolic-trigonometric and describes the standard Reissner-Nordstrϕm metric where $K > 1$.)

Generally speaking, in Einstein general relativity the Reissner-Nordstrϕm metric can be manipulated along with two shells of electrically charged matter to form a traversable wormhole (Schein and Aichelburg, 1996). But there are two drawbacks to this. The first is that the scheme involves dealing with the collapsed state of the stellar matter that generates the metric (a.k.a. Reissner-Nordstrϕm black hole) along with the unpleasant side effects that are encountered, such as the crushing singularities and multiple (unstable) event horizons. Second, the traversable wormhole is an eternal time machine connecting remote regions of the same universe together. Now there are no black hole solutions found in the PV-GR model because in that approach stellar matter collapses smoothly to an ultra-dense state and without the creation of singularities and event horizons (Puthoff, 1999b).

In either case, the Reissner-Nordstrϕm metric does not offer a viable mechanism for vm-Teleportation. We are more interested in examining other PV-GR cases (where $K < 1$ or even $K << 1$)

that emulate the effects of traversable wormhole metrics that do obey the vm-Teleportation definition, such as the example presented in Section 2.1. Equation (2.33) suggests that we search for a vacuum engineering concept that exploits electromagnetic fields to alter the vacuum dielectric constant *K* to induce the desired vm-Teleportation effect in the modified vacuum. (However, we can insert other source terms that will lead to the desired result.) We envision this particular teleportation concept to resemble Figure 2. [Note: Before this report went to press H. E. Puthoff, C. Maccone and the author discovered a number of $K < 1$ solutions to equation (2.33) that uniquely meet the definition of vm-Teleportation and FTL motion. We discovered that the generic energy density required to generate $K < 1$ solutions must be negative, and that the total energy density of the system as seen by remote observes is approximately zero. This unique result compares very well with the traversable wormhole mass-energy density requirements discussed in Section 2.1.2. This discovery will be the subject of a forthcoming paper.]

2.3 Conclusion and Recommendations

The concept we envision for vm-Teleportation is that animate or inanimate objects would be placed inside an environmentally enclosed vessel that would simply be moved into the teleportation device. The "teleporter" would be activated, and the vessel would almost immediately disappear and then reappear at the remote destination as if it were briefly moving through a portal or "stargate." The teleportation device might be required to operate in the vacuum of space outside of the Earth's atmosphere. We have shown two practically equivalent ways to implement vm-Teleportation. There is the manipulation of spacetime geometry via exploiting negative (i.e., quantum vacuum zero point) energy as shown by Einstein's general relativity theory, and there is the modification of the vacuum dielectric constant as shown by the PV-GR model. Both have a great deal of theoretical foundation to begin exploring experimentally. The PV-GR model needs additional theoretical work for the present application, but it is now mature enough for experimental exploration.

There already is extensive theoretical, and more importantly, experimental research proving that the vacuum can be engineered (or physically modified) so that the vacuum ZPE can be exploited (via the Casimir Effect, for example) to extract electrical energy or actuate microelectromechanical devices (see for example, Ambjϕrn and Wolfram, 1983; Forward, 1984, 1996, 1998; Puthoff, 1990, 1993; Cole and Puthoff, 1993; Milonni, 1994; Mead and Nachamkin, 1996; Lamoreaux, 1997; Chan et al., 2001, and the references cited therein). But most of this research involves very low energy density regimes, which are much too low for our purposes. The Mead and Nachamkin (1996) device is actually designed to extract electrical energy from the higher frequency/higher energy density ZPE modes. However, new ultrahigh-intensity lasers became available in the 1990s that have achieved extreme physical conditions in the lab that are comparable to the extreme astrophysical conditions expected to be found in stellar cores and on black hole event horizons (Perry, 1996; Mourou et al., 1998; Perry, 2000). The power intensity of these lasers has reached the point to where they actually probe QED vacuum physics and general relativistic physics, and they have even modified the vacuum itself. The lasers were originally called petaWatt lasers (operating range of 10^{14} – 10^{18} Watts/cm^2 at femtosecond pulses), but they have now reached power intensity levels in the 10^{25} – 10^{30} Watts/cm^2 range. The lasers were made possible by a novel breakthrough called "chirped pulse amplification" whereby the initial low energy/low power intensity laser beam is stretched, amplified and then compressed without experiencing any beam distortions or amplifier damage. This laser system was initially designed as a large-optics beam-line power booster for the NOVA laser fusion experiment at Lawrence Livermore National Laboratory. But researchers found a way to shrink the optics down to tabletop scale, and one can now own and operate a tabletop ultrahigh-intensity laser for ≈ \$500,000. The dimensions of the optical bench used by the University of California-San Diego is ≈ 5 m × 12 m (or ≈ 60 m^2; see Mourou et al., 1998). In tabletop lab experiments ultrahigh-intensity lasers have generated >> gigagauss magnetic fields, ≥ 10^{16} Volt/cm electric field strengths, >> terabar light pressures and >> 10^{22} m/sec^2 subatomic particle accelerations. These ultrahigh-intensity tabletop lasers are thus the ideal instrument with which to explore the fundamental physics underlying the two possible concepts for vm-Teleportation.

There are several ideas on how to generate negative energy in the lab that could potentially be extracted and concentrated in the proper fashion to induce the traversable flat-face wormhole outlined in Section 2.1.1 or induce the $K < 1$ condition (in the PV-GR model) outlined in Section 2.2.1. The schemes for generating negative energy are:

- ☐ Casimir Effect (described in Section 2.2): This is the easiest and most well known way to generate negative energy in the lab. The energy density $\rho_{\text{Casimir}} = -(\pi^2 \hbar c_0/240)a^{-4}$ within a Casimir capacitor cavity is negative and manifests itself by producing a force of attraction between the capacitor plates. This has been measured in the lab (see above references). Forward (1998) proposes a mechanism for the endless extraction of energy from the vacuum in a Casimir cavity by cyclic manipulation of the cavity dimensions.

- ☐ Moving Mirror: Negative quantum vacuum energy can be created by a single moving reflecting surface (a moving mirror). If a mirror moves with increasing acceleration, then a flux of negative energy emanates from its surface and flows out into the space ahead of the mirror (Birrell and Davies, 1982). However, this effect is known to be exceedingly small, and it is not the most effective way to generate negative energy.

- ☐ Optically Squeezed Laser Light: Negative quantum vacuum energy can also be generated by an array of ultrahigh intensity lasers with an ultrafast rotating mirror system. In this scheme a laser beam is passed through an optical cavity resonator made of lithium niobate crystal that is shaped like a cylinder with rounded silvered ends to reflect light. The resonator will act to produce a secondary lower frequency light beam in which the pattern of photons is rearranged into pairs. This is the quantum optical "squeezing" of light effect. (See Section A.2 in Appendix A for a complete definition and description of squeezed quantum states.) Therefore, the squeezed light beam emerging from the resonator will contain pulses of negative energy interspersed with pulses of positive energy. Another way to squeeze light would be to manufacture extremely reliable light pulses containing precisely one, two, three, etc. photons apiece and combine them together to create squeezed states to order. Superimposing many such states could theoretically produce bursts of intense negative energy. For the laser beam resonator example we find that both negative and positive energy pulses are of $\approx 10^{-15}$ second duration. We could arrange a set of rapidly rotating mirrors to separate the positive and negative energy pulses from each other. The light beam is to strike each mirror surface at a very shallow angle while the rotation ensures that the negative energy pulses are reflected at a slightly different angle from the positive energy pulses. A small spatial separation of the two different energy pulses will occur at some distance from the rotating mirror. Another system of mirrors will be needed to redirect the negative energy pulses to an isolated location and concentrate them there.

- ☐ Gravitationally Squeezed Vacuum Energy: A natural source of negative quantum vacuum energy comes from the effect that gravitational fields (of astronomical bodies) in space have upon the surrounding vacuum. For example, the gravitational field of the Earth produces a zone of negative energy around it by dragging some of the virtual particle pairs (a.k.a. virtual photons or vacuum ZPF) downward. This concept was initially developed in the 1970s as a byproduct of studies on quantum field theory in curved space (Birrell and Davies, 1982). However, Hochberg and Kephart (1991) derived an important application of this concept to the problem of creating and stabilizing traversable wormholes, and their work was corrected and extended by Davis (1999a). They proved that one can utilize the negative vacuum energy densities, which arise from *distortion of the electromagnetic zero point fluctuations* due to the interaction with a prescribed gravitational background, for providing a violation of the energy conditions (see

Approved for public release; distribution unlimited.

Section A.1 in Appendix A). Hochberg and Kephart (1991) showed that the squeezed quantum states of quantum optics provide a natural form of matter having negative energy density. And since the vacuum is defined to have vanishing energy density, anything possessing less energy density than the vacuum must have a negative energy density. The analysis, via quantum optics, shows that gravitation itself provides the mechanism for generating the squeezed vacuum states needed to support stable traversable wormholes. The production of negative energy densities via a squeezed vacuum is a necessary and unavoidable consequence of the interaction or coupling between ordinary matter and gravity, and this defines what is meant by gravitationally squeezed vacuum states. The magnitude of the gravitational squeezing of the vacuum can be estimated from the squeezing condition, which simply states that substantial gravitational squeezing of the vacuum occurs for those quantum electromagnetic field modes with *wavelength* (λ in meters) > *Schwarzschild radius* (r_S in meters) of the mass in question (whose gravitational field is squeezing the vacuum). The Schwarzschild radius is the critical radius, according to general relativity theory, at which a spherically symmetric massive body becomes a black hole; i.e., at which light is unable to escape from the body's surface. We can actually choose any radial distance from the mass in question to perform this analysis, but using the Schwarzschild radius makes equations simpler in form. The general result of the gravitational squeezing effect is that as the gravitational field strength increases the negative energy zone (surrounding the mass) also increases in strength. Table 3 shows when gravitational squeezing becomes important for example masses. The table shows that in the case of the Earth, Jupiter and the Sun, this squeeze effect is extremely feeble because only ZPF mode wavelengths above 0.2 m – 78 km are affected. For a solar mass black hole (radius of 2.95 km), the effect is still feeble because only ZPF mode wavelengths above 78 km are affected. But note from the table that quantum black holes with Planck mass will have enormously strong negative energy surrounding them because all ZPF mode wavelengths above 8.50×10^{-34} meter will be squeezed; in other words, all wavelengths of interest for vacuum fluctuations. Black holes with proton mass will have the strongest negative energy zone in comparison because the squeezing effect includes all ZPF mode wavelengths above 6.50×10^{-53} meter. Furthermore, a black hole smaller than a nuclear diameter ($\approx 10^{-16}$ m) and containing the mass of a mountain ($\approx 10^{11}$ kg) would possess a fairly strong negative energy zone because all ZPF mode wavelengths above 10^{-15} meter will be squeezed.

Table 3. Substantial Gravitational Squeezing Occurs When $\lambda \geq 8\pi r_S$ (For Electromagnetic ZPF; adapted from Davis, 1999a**)**

Mass of body	Schwarzschild radius of body, r_S	ZPF mode wavelength, λ
Sun = 2.0×10^{30} kg	2.95 km	≥ 78 km
Jupiter = 1.9×10^{27} kg	2.82 m	≥ 74 m
Earth = 5.976×10^{24} kg	8.87×10^{-3} m	≥ 0.23 m
Typical mountain ≈ 10^{11} kg	≈ 10^{-16} m	≥ 10^{-15} m
Planck mass = 2.18×10^{-8} kg	3.23×10^{-35} m	≥ 8.50×10^{-34} m
Proton = 1.673×10^{-27} kg	2.48×10^{-54} m	≥ 6.50×10^{-53} m

•**Recommendations**:

- Theoretical Program 1: A one to two year theoretical study (cost ≈ \$80,000) should be initiated to explore the recently discovered $K < 1$ (FTL) solutions to equation (2.33) in order to define, characterize and model the negative energy density source(s) that induce the FTL vacuum modification. The study should also identify potential lab experiments designed to test theoretical predictions.

- Theoretical Program 2: A one to two year study (cost ≈ \$80,000) should be initiated to conduct a detailed review of the negative energy generation schemes summarized above to define their characteristics, performances and requirements. The study should develop technical parameters for each of the schemes in order to identify potential lab experiments.

- Experimental Program 1: An experimental study should be conducted to test Forward's (1998) Casimir energy extraction proposal. An experiment definition study will be required to estimate the experimental method, procedure, equipment needs and costs.

- Experimental Program 2: An experimental study using ultrahigh-intensity lasers should be conducted to test the Optically Squeezed Laser Light proposal. An experiment definition study will be required to estimate the experimental method, procedure, equipment needs and costs.

- Experimental Program 3: An experimental study using ultrahigh-intensity lasers should be conducted to probe QED vacuum physics and vacuum modification as well as test elements of the PV-GR model. A starting point for this program would be to use such lasers to perform the Ding and Kaplan (1989, 1992, 2000; see also, Forward, 1996) experiment. This is an important fundamental physics experiment to do, because it can distinguish between the rival quantum vacuum electromagnetic ZPE fluctuation and fluctuating charged particle source field theory models, which would settle the acrimonious debate over whether the vacuum really fluctuates or not. R. L. Forward (1999) told the author that a Nobel Prize rides on performing this experiment and settling the issue once and for all. The Ding and Kaplan proposal is already designed to probe QED vacuum physics and vacuum modification. [The essence of the Ding and Kaplan proposal is to demonstrate that a form of photon-photon scattering predicted by QED gives rise to 2nd-harmonic generation of intense laser radiation in a DC magnetic field due to the broken symmetry of interaction (in the Feynman "box" diagram approximation). This effect is possible only when the field system (optical wave + DC field) is inhomogeneous, in particular when a Gaussian laser beam propagates in either a homogeneous or inhomogeneous DC magnetic field. In other words, a vacuum region is filled with a DC magnetic field that polarizes the virtual particle pairs (a.k.a. virtual photons) in the vacuum. This polarized vacuum then scatters incident ultrahigh-intensity laser photons of frequency ν (energy E), thereby generating outgoing photons of frequency 2ν (energy $2E$).] An experiment definition study will be required to estimate the experimental method, procedure, equipment needs and costs.

- Experimental Program 4: An experimental study using ultrahigh-intensity lasers should be conducted to establish the extreme physical conditions necessary to test the strong-field limit of general relativity with an emphasis on generating spacetime curvature and negative energy in order to induce a putative micro-wormhole. (Experimental Programs 3 and 4 could be done together to determine whether Puthoff's PV-GR theory or Einstein's general relativity theory is the correct model for nature.) A Nobel Prize is in the offing if this question were to be addressed and settled. An experiment definition study will be required to estimate the experimental method, procedure, equipment needs and costs.

3.0 q-TELEPORTATION

3.1 Teleportation Scenario

Future space explorers and their equipment will need to easily and quickly travel from an orbiting spacecraft to the surface of some remote planet in order to get their work done, or military personnel in the United States need to easily and quickly travel from their military base to another remote location on Earth in order to participate in a military operation, or space colonists will need quick transport to get from Earth to their new home planet. Instead of using conventional transportation to expedite travel the space explorer, military personnel or space colonist and/or their equipment go into the "Teleporter" (a.k.a. "Transporter" in Star Trek lingo) and are "beamed down" or "beamed over" to their destinations at light speed. The mechanism for this teleportation process is hypothetically envisioned to be the following:

1. Animate/inanimate objects placed inside the teleporter are scanned by a computer-generated and -controlled beam.

2. The scan beam encodes the entire quantum information contained within the animate/inanimate object(s) into organized bits of information, thus forming a digital pattern of the object(s).

3. The scan beam then dematerializes the object(s) and stores its pattern in a pattern buffer, thus transforming the atomic constituents of the dematerialized object(s) into a matter stream. Alternative 1: The dematerialization process converts the atoms into a beam of pure energy. Alternative 2: The scan beam does not dematerialize the object(s).

4. The teleporter then transmits the matter/pure energy stream and quantum information signal in the form of an annular confinement beam to its destination. Alternative: Only the quantum information signal is transmitted.

5. At the receiving teleporter the matter/pure energy stream is sent into a pattern buffer whereby it is recombined with its quantum information, and the object(s) is rematerialized back into its original form. Alternative 1: The receiving teleporter recombines the transmitted quantum information with atoms stored inside a reservoir to form a copy of the original. Alternative 2: The quantum information is reorganized in such a way as to display the object on some three-dimensional (holographic) visual display system.

Problem: This generic scenario is modeled after teleportation schemes found in SciFi. There are a lot of important little details that were left out of the teleportation process because we simply do not know what they are. This technology does not yet exist. And we are left with the question of which one of the alternative processes identified in items 3 – 5 one wants to choose from. The above scenario is only an outline, and it is by no means complete since it merely serves to show what speculation exists on the subject. The above scenario describes a speculative form of what we call *q-Teleportation*.

There are questions to be addressed in the above scenario. Does the teleporter transmit the atoms and the quantum bit information signal that comprises the animate/inanimate object or just the quantum bit information signal? There are $\approx 10^{28}$ atoms of matter combined together in a complex pattern to form a human being. How does one transmit this much information and how do we disassemble that many atoms? Computer information gurus would insist that it is not the atoms that matter but only the bits of information representing them when considering the transmission of large "bodies" of information. But are humans simply the sum of all the atoms (and the related excited atom quantum states) that comprise them? We could possibly learn to reconstitute a beam of atoms into a chemically accurate human being. However, would this also include the reconstruction of a person's consciousness (personality, memories, hopes, dreams, etc.) and soul or spirit? This question is beyond the scope of this study to address, but it is nevertheless one of the most important concepts awaiting a complete scientific understanding.

For the teleporter to process and transmit the quantum bit information signal that encodes the animate/inanimate object's pattern will require stupendous digital computer power. For each atom comprising the object we must encode its location in space (three position coordinates), its linear and angular momentum (three vector components for each quantity), and its internal quantum state (electron orbital-energy levels and their excitation/de-excitation and ionization states, binding to other atoms to form molecules, molecular vibrational/rotational states, bound nuclei states, spin states for electrons and nuclei, etc.), etc. If we assume that we can digitally encode all of this information for a single atom with a minimum of one kilobyte (1 byte = 8 bits, 1 bit ≡ 0 or 1) of data, then we will require a minimum of 10^{28} kilobytes to encode and store an entire human being (in three-dimensions). To digitally store and access this much information at present (and for the foreseeable future) is nontrivial. It will take more than 2,400 times the present age of the universe (≈ 13 billion years) to access this amount of data using commercially available computers (operating at ≈ 10 gigabyte/sec). Top-of-the-line supercomputers will not reduce this time significantly. The computer technology needed to handle such a large data storage requirement simply does not exist. The largest commercially available computers can store ≈ 40 gigabytes on a single hard drive. We will need ≈ 10^{20} of these hard drives to store the encoded information of just one human being. Also, wire and coaxial/fiber optic cables do not have the physical capacity to transmit this amount of data between devices. These numbers will not be significantly different for macroscopic inanimate objects. The information processing and transfer technology required for the teleportation system may become possible in 200 – 300 years if improvements in computer storage and speed maintains a factor of 10 – 100 increase for every decade. There is speculation that emergent molecular, bio-molecular (DNA-based systems) and quantum computer technology may achieve the performances required for a teleportation system. In the former case molecular dynamics mimics computer logic processes and the ≈ 10^{25} particles in a macroscopic sample will all act simultaneously, making for far greater digital information processing and transfer speeds. Researchers have given no formal performance estimates for this emergent technology. In the latter case quantum computing would take advantage of entangled quantum states of subatomic matter or photons, whereby digital logic processes would occur at light speed. This technology is in its infancy, and there has been no clear direction on what performance levels will be possible in the future. This topic will be discussed further in Section 3.2.3.

In the above teleportation scenario we might consider dematerializing animate/inanimate objects into a matter stream consisting of only the object's constituent atoms or atomic subcomponents (protons, neutrons and electrons) and transmitting them at the speed of light (or close to it). To push atoms or subatomic particles to near the speed of light will require imparting to them an energy comparable to their rest-mass energy, which will be at a minimum of one order of magnitude larger than the amount of energy required to break protons up into free quarks. The energy required to completely dematerialize (or dissolve) matter into its basic quantum constituents or into pure energy is alone stupendous. At first one will have to impart to every molecule within the object an energy that is equivalent to the binding energy between atoms (atomic binding energy ~ chemical energy ~ *several* eV) in order to break apart the molecules comprising the object's macro-structure. After this an energy equivalent to nuclear binding energies (≈ *several* × 10^6 times atomic binding energy, or ≈ *several* MeV) must be imparted to every free atomic nucleus inside the object in order to break apart the protons and neutrons residing within each nucleus. And last, an energy equivalent to the binding energy that holds together the three quarks residing within each proton and neutron must be imparted to each of the free protons and neutrons within the object. According to the Standard Model and experimental data, the quark binding energy is practically infinite. But all is not lost, because the Standard Model also predicts that if we could heat up the nuclei to ≈ 10^{13} °C (≈ 10^6 times hotter than the core temperature of the Sun, or ≈ 10^3 MeV), then the quarks inside would suddenly lose their binding energies and become massless (along with other elementary matter). This heat is also equivalent to the rest-mass energy of protons and neutrons. Therefore, to heat up and dematerialize one human being would require the annihilation of the rest mass-energy of all 10^{28} protons-neutrons or the energy equivalent of 330 1-megaton thermonuclear bombs. Compare this stupendous explosive energy with the explosive yield of the largest thermonuclear bomb ever detonated on Earth, which was a 50-megaton bomb that was built by Andrei Sakharov in the USSR and detonated on October 30, 1961; it was called "Tzar Bomba." Its first incarnation (ca. early October 1961) comprised a uranium fusion tamper, which gave an estimated explosive yield of ≈ 100 megatons. But the weapon was too heavy (27 metric tons) for a bomber to carry, so the tamper was replaced by one made of lead, which reduced both the weight and the yield. In the end we see that it is not a trivial problem to simply heat up and dematerialize any human or inanimate objects. The technology to do so does not exist unless we invoke new physics to get around the energy requirement.

Finally, we must consider the resolution and aperture of the optics required to scan and transmit the animate/inanimate object's matter (or energy) stream. The Heisenberg quantum uncertainty principle fundamentally constrains the measurement resolution of conjugate observable quantities, such as position and momentum or energy and time. The measurement of any combination of (conjugate) observables with arbitrarily high precision is not possible, because a high precision measurement of one observable leads to imprecise knowledge of the value of the conjugate observable. The quantum uncertainty principle makes it impossible to measure the exact, total quantum state of any object with certainty. The scan resolution of a teleportation system is defined by the wavelength of light used to illuminate the object's atomic/subatomic constituents and record their configurations. To resolve matter at atomic/subatomic distance scales requires that the energy of the scanner light (photons) be extremely large (according to the uncertainty principle); and during the scan this large light energy will be conveyed to the constituents, causing them to drastically change their speed and direction of motion. This means that it is physically impossible to resolve an object's atomic/subatomic particle components and their configurations with the precision necessary to accurately encode and later recreate the object being teleported. To resolve atomic/subatomic particles requires wavelengths smaller than the size of these constituents, which will typically be 1 Å – 1 fm. Such wavelengths are in the gamma ray part of the spectrum, and this becomes a major technical problem for us because at present there is no gamma ray electro-optics with which to work with. Now consider the example of teleporting an object from the surface of a planet back to its spacecraft in orbit some *several* × 10^2 – 10^3 km away. The optical aperture required to illuminate and scan an object with ≈ 1 Å – 1 fm resolution from orbit will be >> *several* × 10^2 – 10^3 km. If we are to consider teleporting an object from planet to planet or from star to star then the aperture required will be >> *several* × 10^8 – 10^{13} km. These technical problems are truly insurmountable unless totally new physics becomes available.

3.2 Quantum Teleportation

It turns out that there does in fact exist a form of teleportation that occurs in nature despite the numerous technical roadblocks described in the previous section. It is called *quantum teleportation*, which is based on the well-known concept of quantum entanglement. Erwin Schrödinger coined the word "entanglement" in 1935 in a three-part paper (Schrödinger, 1935a, b, c, 1980). These papers were prompted by the Einstein, Podolsky and Rosen (1935; denoted hereafter as EPR) paper that raised fundamental questions about quantum mechanics, whereby Einstein had loudly complained that quantum mechanics allowed physical processes resembling "spooky action at a distance" to occur. EPR recognized that quantum theory allows certain correlations to exist between two physically distant parts of a quantum system. Such correlations make it possible to predict the result of a measurement on one part of a system by looking at the distant part. On this basis, EPR argued that the distant predicted quantity should have a definite value even before being measured, if quantum theory is complete and respects locality (a.k.a. causality). EPR concluded that, from a classical perspective, quantum theory must be incomplete because it disallows such definite values prior to measurement. Schrödinger's perspective on this argument gives the modern view of quantum mechanics, which is to say that the wavefunction (a.k.a. quantum state vector) provides all the information there is about a quantum system. In regards to the nature of entangled quantum states, Schrödinger (1935a, b, c, 1980) stated that, "The whole is in a definite state, the parts taken individually are not." This statement defines the essence of pure-state entanglement. Schrödinger went on to give a description of quantum entanglement by introducing his famous cat experiment.

To better understand the concept of quantum entanglement/teleportation we will focus on the quantum wavefunction (a.k.a. quantum state function). Any quantum system such as a particle that possesses a position in space, energy, angular and linear momentum, and spin is completely described by a wavefunction. This is usually symbolized in a variety of ways, and we choose to represent a generic wavefunction using the traditional "bra-ket" notation of quantum mechanics: $|\varphi\rangle$. Anything that we want to know about the particle is mathematically encoded within $|\varphi\rangle$. As we discussed in the previous section the wavefunction can never be completely known because there is no measurement that can determine it completely. The only exception to this is in the special case that the wavefunction has been prepared in some particular state or some member of a known basis group of states in advance. By measuring one of the properties of a quantum system, we can get a glimpse of the overall quantum state that is encoded within $|\varphi\rangle$. According to the quantum uncertainty principle the act of doing such a measurement will destroy any ability to subsequently determine the other properties of the quantum system. So the act of measuring a particle actually destroys some of the information about its pristine state. This makes it impossible to copy particles and reproduce them elsewhere via quantum teleportation. However, it turns out that one can recreate an unmeasured quantum state in another particle as long as one is prepared to sacrifice the original particle. The trick is to exploit the EPR process to circumvent the quantum uncertainty principle.

As discussed previously, EPR discovered that a pair of spatially separated quantum sub-systems that are parts of an overall quantum system can be "entangled" in a non-local (i.e., non-causal) way. When two particles come into contact with one another, they can become "entangled." In an entangled state, both particles remain part of the same quantum system so that whatever you do to one of them affects the other one in a predictable fashion. More precisely, a measurement on one of the entangled sub-systems puts it into a particular quantum state, while instantaneously putting the sub-system with which it is entangled into a corresponding quantum state, while the two sub-systems are separated by arbitrarily large distances in spacetime (even backwards in time!). A simple example of this phenomenon is to prepare a pair of photons in the same quantum state such that they are entangled, and then allow them to fly apart to remote locations without any form of communication occurring between them along their journey. Measuring the polarization of one of the pair of entangled photons induces the other photon, which may be light-years away, into the same state of polarization as that which was measured for its entangled twin. The basic operation of quantum teleportation can be described as determining the total quantum state of some large quantum system, transmitting this state information from one place to another, and making a perfect reconstruction of the system at the new location. In principle, entangled particles can serve as "transporters" of sorts. By introducing a third "message" particle to one of the entangled particles, one could transfer its properties to the other one, without ever measuring those properties.

Historically, quantum entanglement was never reconciled with the quantum uncertainty principle and the requirement of locality (or causality) in observed physical phenomena, thus it became a paradox in quantum theory. A three-decade debate began following the appearance of the EPR paper over whether quantum entanglement (a.k.a. "spooky action at a distance") was a real quantum phenomenon or not, and this debate came to be called the "EPR dilemma." Einstein's only solution to the dilemma was to suggest that quantum mechanics was incomplete and needed a reformulation to incorporate local hidden-variables that can account for observed physical phenomena without violating causality. Bell (1964) later solved the EPR dilemma by deriving correlation inequalities that can be violated in quantum mechanics but have to be satisfied within every model that is local and complete. Such models are called “local hidden-variable models.” Bell showed that a pair of entangled particles, which were once in contact but later moved too far apart to interact directly (i.e., causally), can exhibit individually random behavior that is too strongly correlated to be explained by classical statistics. Bell’s inequalities make it possible to test whether local hidden-variable models can account for observed physical phenomena in lab experiments. Groundbreaking experimental work by Aspect et al. (1982a, b) along with further theoretical and experimental work done by others (Freedman and Clauser, 1972; Aspect, 1983; Aspect and Grangier, 1985; Hong and Mandel, 1985; Bennett and Wiesner, 1992; Tittel et al., 1998a, b; Tittel and Weihs, 2001) demonstrated violations of the Bell inequalities, which therefore invalidated the local hidden-variable models. The key result of recent theoretical and experimental work is that an observed violation of a Bell inequality demonstrates the presence of entanglement in a quantum system.

3.2.1 Description of the q-Teleportation Process

The experimental work of Bennett et al. (1993) followed by the theoretical and experimental work of others (Vaidman, 1994; Kwiat et al., 1995; Braunstein, 1996; Braunstein and Kimble, 1998; Pan et al., 1998; Stenholm and Bardroff, 1998; Zubairy, 1998; Vaidman and Yoran, 1999; Kwiat et al., 1999) made the breakthrough that was necessary to demonstrate the principle of quantum teleportation in practice. It was a remarkable technical breakthrough that settled, once and for all, the nagging question of whether quantum entanglement could be used to implement a teleportation process to transfer information between remotely distant quantum systems non-causally (i.e., at FTL speed). It is easy to describe how quantum teleportation works in greater detail. Figure 6 compares conventional facsimile transmission with the quantum teleportation process seen in Figure 7. In a conventional facsimile transmission the original document is scanned, extracting partial information about it, but it remains more or less intact after the scanning process. The scanned information is then sent to the receiving station, where it is imprinted on new paper to produce an approximate copy of the original. In quantum teleportation (Figure 7) one scans out part of the information from object A (the original), which one wants to teleport, while causing the remaining, unscanned, part of the information in A to pass, via EPR entanglement, into another object C which has never been in contact with A. Two objects B and C are prepared and brought into contact (i.e., entangled), and then separated. Object B is taken to the sending station, while object C is taken to the receiving station. At the sending station object B is scanned together with the original object A, yielding some information and totally disrupting the states of A and B. This scanned information is sent to the receiving station, where it is used to select one of several treatments to be applied to object C, thereby putting C into an exact replica of the former state of A. Object A itself is no longer in its original initial state, having been completely disrupted by the scanning process. The process just described is teleportation and not replication, and one should not confuse the two. There is a subtle, unscannable kind of information that, unlike ordinary information or material, can be delivered via EPR correlations/entanglement, such that it cannot by itself deliver a meaningful and controllable message. But quantum teleportation delivers exactly that part of the information in an object that is too delicate to be scanned out and delivered by conventional methods.

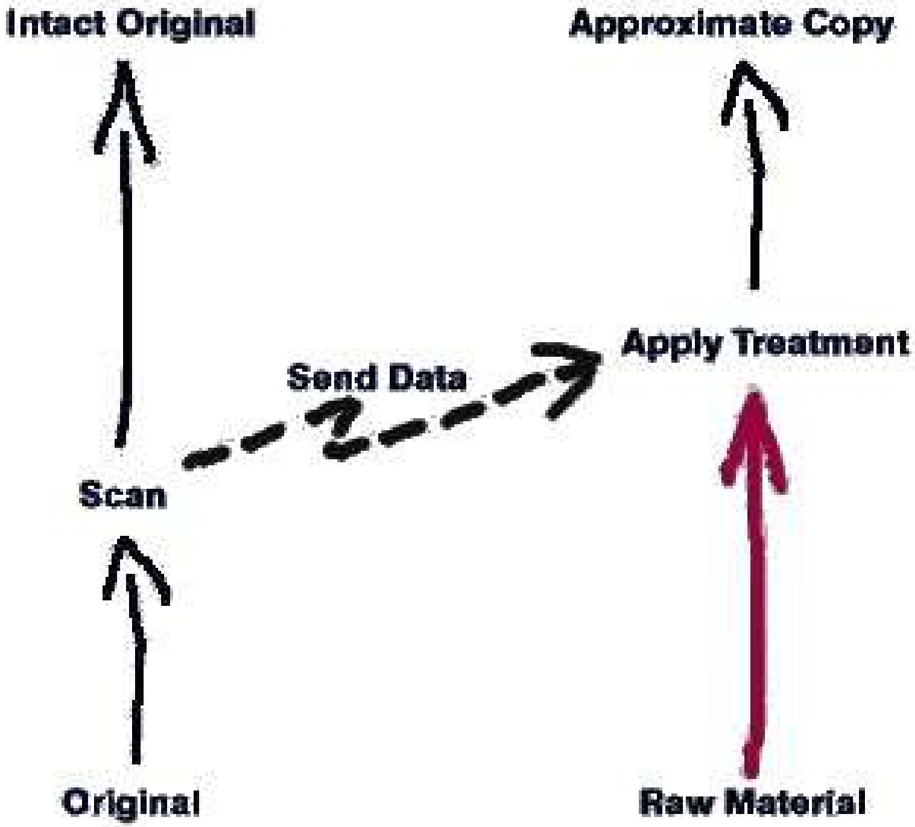
Intact Original
Approximate Copy
Apply Treatment
Send Data
Scan
Original
Raw Material

Figure 6. Classical Facsimile Transmission (Modified IBM Press Image)

Approved for public release; distribution unlimited.

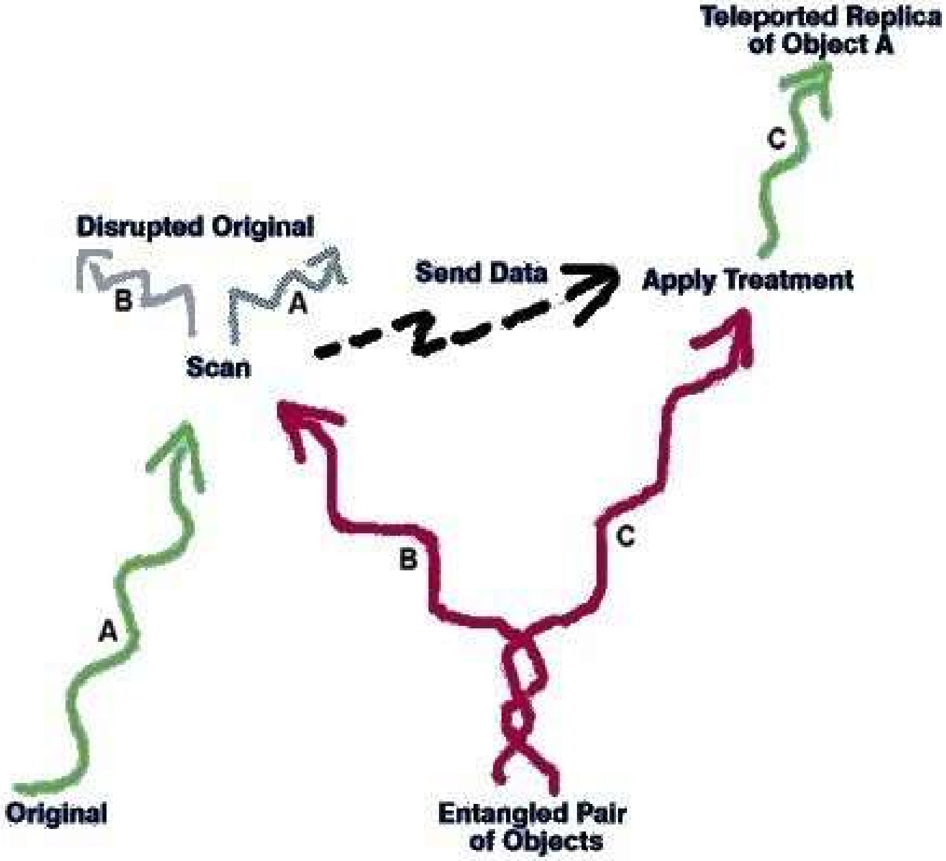
Teleported Replica of Object A
C
Disrupted Original
B
A
Send Data
Apply Treatment
Z
Scan
B
C
A
Original
Entangled Pair of Objects

Figure 7. Quantum Teleportation (Modified IBM Press Image)

Approved for public release; distribution unlimited.

We now go one more final step to give a simplified outline of the actual teleportation process according to Bennett et al. (1993). They propose a multistep procedure by which any quantum state $|\chi\rangle$ of a particle or a photon (that correspond to an N-state system) is to be teleported from one location to another. For example, $|\chi\rangle$ might be a two-level system that could refer to the polarization of a single photon, the nuclear magnetic spin of a hydrogen atom, or the electronic excitation of an effective two-level atom. The following scenario outlines the q-Teleportation process in a very simplified way:

1. Prepare a pair of quantum subsystems $|\varphi\rangle$ and $|\psi\rangle$ in an EPR entangled state so that they are linked together. $|\varphi\rangle$ and $|\psi\rangle$ are maximally entangled and together constitute a definite pure state superposition even though each of them is maximally undetermined or mixed when considered separately.

2. Transport $|\varphi\rangle$ to the location of the teleportation transmitter and transport $|\psi\rangle$ to the location of the teleportation receiver. (In the technical literature the transmitter is called "Alice" and the receiver is called "Bob.") The transmitter and receiver can be many light years apart in space. Note that the two subsystems are non-causally correlated via entanglement, but they contain no information about $|\chi\rangle$ at this point. The two subsystems represent an open quantum channel that is ready to transmit information.

3. Now Alice brings the teleported state $|\chi\rangle$ into contact with the entangled state $|\varphi\rangle$ and performs a quantum measurement on the combined system $|\chi\rangle|\varphi\rangle$. Bob and Alice have previously agreed upon the details of the quantum measurement.

4. Using a conventional classical communication channel, Alice transmits to Bob a complete description of the outcome of the quantum measurement she performed on $|\chi\rangle|\varphi\rangle$.

5. Bob then subjects $|\psi\rangle$ to a set of linear transformations (i.e., suitable unitary rotations) that are dictated by the outcome of Alice's quantum measurement. The quantum subsystem Bob originally first received is no longer in state $|\psi\rangle$ after the linear transformations because it is now in a state identical to the original state $|\chi\rangle$. Therefore, $|\chi\rangle$ has in effect been teleported from Alice to Bob.

Bennett et al. (1993) showed in their experimental work that this scheme requires both a conventional communication channel and a non-causal EPR channel to send the state $|\chi\rangle$ from one location to another. In addition to this, a considerable pre-arrangement of entangled states and quantum measurement procedures is required to make the process work. Bennett et al. (1993) analyzed the information flow implicit in the process and showed that Alice's measurement does not provide any information about the quantum state $|\chi\rangle$. All of the quantum state information is passed by the EPR link between the entangled particle states $|\varphi\rangle$ and $|\psi\rangle$. We can think of the measurement results as providing the "code key" that permits the EPR information to be decoded properly at Bob's end. And because the measurement information must travel on a conventional communications channel, the decoding cannot take place until the code key arrives, insuring that no FTL teleportation is possible.

The q-Teleportation scheme teleports the state of a quantum system without having to completely measure its initial state. The outcome of the process is that the initial quantum state $|\chi\rangle$ is destroyed at Alice's location and recreated at Bob's location. It is very important for the reader to understand that it is the *quantum states* of the particles/photons that are destroyed and recreated in the teleportation process, and not the particles/photons themselves. The quantum state or wavefunction contains the information on the state of a particle, but is not a directly observable physical quantity like mass-energy. The quantum information contained within a state is available in the form of probabilities or expectation values.

Therefore, q-Teleportation *cannot* teleport animate or inanimate matter (or energy) in its physical entirety. However, some experts argue that because a particle's or a photon's quantum state is its defining characteristic, teleporting its quantum state is completely *equivalent* to teleporting the particle/photon even though the original particle's/photon's quantum state (and defining characteristic) was completely destroyed in the process (more on this in Section 3.3). Therefore, no quantum cloning is possible and we are left with a (near-perfect) copy of the now-destroyed original after teleportation (Wootters and Zurek, 1982; Barnum et al., 1996). And finally, classical information itself *cannot* be teleported faster than the speed of light via the non-causal EPR channel; however, quantum information can (more on this in Section 3.2.3).

3.2.2 Decoherence Fundamentally Limits q-Teleportation

Finally, the reader must understand that the q-Teleportation scenario described in the previous section was simplified because we unrealistically assumed that Alice and Bob shared an EPR entangled pair that was free of noise or decoherence. Decoherence is the process, whereby an object's quantum states degrade when information leaks to or from the environment (i.e., environmental noise) through stray interactions with the object. In reality, Alice and Bob have quantum systems that interact directly or through another mediating quantum system like two ions in an ion trap that interact through phonon modes of the trap, or Rydberg atoms in a laser cavity that interact via photons (Sackett, 2001; Raimond et al., 2001). Decoherence degrades the fidelity of the quantum link (i.e., the set of pure EPR entangled pairs) between two quantum systems, thus introducing a certain level of error in the exchange of quantum information between the systems.

In a real-world example of an application of q-Teleportation to quantum computation (discussed in the next section), we can devise an array of interconnected ion traps with each trap holding a small number of ions that are coupled by ions that are moved between the traps or by traveling photons (Wineland et al., 2002). The quantum link (or EPR interaction) between a pair of systems is subject to noise or decoherence through photon loss or heating of the phonons. At present, decoherence imposes a fundamental limit on our ability to perform quantum information processing. Research is continuing on whether decoherence can be reduced, circumvented, or otherwise be (partially or totally) eliminated. Dür and Briegel (2003) have taken the first step towards this goal at rudimentary level by showing that fault-tolerant quantum computation can be achieved in the presence of very high noise levels occurring in the interaction link between small quantum systems, if one assumes that local quantum processing on each end is nearly error free. They showed that the interaction link can have an error rate of two-thirds.

3.2.3 Recent Developments in Entanglement and q-Teleportation Physics

Quantum teleportation physics is still in its infancy. Both theoretical and experimental developments are advancing in many different directions, but are far from maturity at this point in time because the field is still evolving at present. Technical applications of entanglement and q-Teleportation are just becoming conceptualized for the first time, while a small number of basic physics breakthroughs and their related applications are in experimental progress at present. The research community is still in the process of discovering the full nature of entanglement and q-Teleportation, its rules, and what roadblocks nature has in store for its applications and further progression. The literature cited in this study is by no means complete, and only represents a subset of the entire field, because the research is still evolving.

An important application of quantum entanglement and q-Teleportation was the discovery made by Shor (1994, 1997) that computation with quantum states instead of classical bits can result in large savings in computation time. For example, the best algorithms take exponentially more resources to factor ever-larger numbers on a classical computer. A 500-digit number needs 10^8 times as many computational steps to factor as a 250-digit number. The latter classically requires $\approx 5\times10^{24}$ computational steps, or about 150,000 years computing time at terahertz speed, to factor. Shor found a polynomial-time quantum algorithm that solves the problem of finding prime factors of a large integer.

He showed that his algorithm rises only polynomially so that a 500-digit number takes only eight times as many computational steps to factor as a 250-digit number. And by using the quantum factoring algorithm, a 250-digit number requires only $\approx 5\times10^{10}$ steps or < 1 second to factor at terahertz speed, so that a 500-digit number will take ≤ 1 second to factor. No classical polynomial-time algorithm for this problem exists at present. This breakthrough generated a cottage industry of research into quantum computing and quantum information theory.

IBM (2001) constructed a prototype quantum computer that uses the nuclear spins of seven atoms that are part of a large molecule with the iron-based chemical composition $H_5C_{11}O_2F_5Fe$. The computer uses entangled nuclear spins for storage and has a capacity of seven qubits (qubits are defined in the bulleted list in the next two paragraphs below). All of the Fluorine atoms in the large molecule are Fluorine isotope 19 and two of the Carbon atoms are Carbon isotope 13. All the other non-hydrogen atoms have even isotope numbers and no nuclear spins. The objective of the prototype quantum computer was to factor the number 15 into its two prime factors 3 and 5 by using Shor's quantum factoring algorithm. The quantum computation required that a sample of $\approx 10^{18}$ of the large molecules be placed in a magnetic field and manipulated by nuclear magnetic resonance (NMR) techniques. This mechanism allows the spins to function as qubits, whereby Schor's algorithm can be performed via manipulation of the NMR fields. NMR was used to implement quantum computing in this prototype, because the nuclear spins are well isolated from decoherence as a result of the very long decoherence time (the time after which quantum coherence is lost due to environmental noise) in the system.

To factor larger numbers will require a system that uses more than seven qubits. It is estimated that a quantum computer using ≈ 36 qubits could very quickly perform computations that would require a conventional computer ≈ 13 billion years to perform. And such a computer could solve one of the technical problems of human teleportation discussed in Section 3.1. However, a scale-up in the number of qubits is difficult because the IBM prototype has reached the technology limit of NMR quantum computing. The prototype's operation requires that all of the qubits must be in the same molecule. And molecules with more than seven spins that can be used as qubits are not feasible at present. However, there are alternative technologies for quantum computing that show promise for scaling up the number of qubits. The technologies of nuclear spin orientation of single atom impurities in semiconductors, electron spin orientation in quantum dots, and the manipulation of magnetic flux quanta in superconductors all show promise of providing a basis for scalable quantum computers. Finally, the primary technical problem in quantum computing at the present time is decoherence, and this must be eliminated or otherwise mitigated before new quantum technology can become competitive with conventional computer technology.

A byproduct of the recent quantum computing and information research is that a modern theory of entanglement has emerged. Researchers now treat entanglement as a quantifiable physical resource that enables quantum information processing and computation. Entanglement is no longer treated as a paradox of quantum theory. It has been recently discovered that (Nielsen and Chuang, 2000; Nielsen, 2003; Terhal et al., 2003):

- various kinds of pure and mixed entangled states may be prepared in addition to the simple pure-state superpositions that was described in the previous section

- the members of an entangled group of objects do not have their own individual quantum states, only the group as a whole has a well-defined state (i.e., "the whole is greater than the sum of its parts")

- entangled objects behave as if they were physically connected together no matter how far apart they actually are, distance does not attenuate entanglement in the slightest – it has been demonstrated that information can be teleported over 40 km using existing technology (H. Everitt, Army Research Office, 2000)

- if something is entangled with other objects, then a measurement of it simultaneously provides information about its partners

- some quantum systems can have a little entanglement while others will have a lot

- the more entanglement available, the better suited a system is to quantum information processing

- decoherence degrades the fidelity of the quantum link (i.e., the set of pure EPR entangled pairs) between two quantum systems, thus introducing a certain level of error in the exchange of quantum information between the systems; thus limiting our ability to perform quantum information processing (see more on this issue in the next paragraph below)

- mixed entangled states may be measured, distilled, concentrated, diluted, and manipulated

- the basic resource of classical information is the bit (i.e., the two values 0 and 1), while quantum information comes in quantum bits (i.e., qubits) that are described by their quantum state; qubits can exist in superpositions that simultaneously involve 0 and 1, thus giving them an infinite range of values; groups of qubits can be entangled; qubits must be insulated against decoherence, so that the coherent state of the quantum system in a quantum computer is preserved for a time that is long enough to set up a calculation, perform it, and read out the results

- quantum computers processing qubits or entangled qubits can outperform classical computers; functional requirements of quantum computers:

    - ❖ they must have the ability to initialize any qubit in a specified state, and to measure the state of a specific qubit
    - ❖ they must have universal quantum gates, which are logical elements capable of arranging any desired logical relationship between the states of qubits
    - ❖ they must also have a processor capable of interlinking quantum gates to establish rules and boundary conditions for their inter-relationships – in a quantum computation, the arrangement of quantum gates connects the qubits in a logical pattern, according to a program or algorithm, and after an interval the qubits assigned to the result are read out

- quantum error correction codes exist, whereby qubits are passed through a circuit (the quantum analogue of logic gates) that will successfully fix an error in any one of the qubits without actually reading what all the individual qubit states are; no qubit cloning is required

- a completely secure quantum key can be generated and distributed (for communication and decoding of encrypted messages) using entangled photons has been demonstrated (Tittel et al., 2000; Jennewein et al., 2000; Naik et al., 2000); any eavesdropper's attempt to intercept the quantum key will alter the contents in a detectable way, enabling users to discard the compromised parts of the data

- in an experiment which verified that EPR entanglement obeys Special Relativity (Seife, 2000; Scarani et al., 2000; Gisin et al., 2000; Zbinden et al., 2000a, b), and involving a photon detector moving at relativistic speeds (for example, Bob moves away from Alice at close to the speed of light), investigators determined that quantum information via EPR photon pair entanglement must travel > 10^7 times light speed (the photon detectors were 10.6 km apart)

- investigators are still developing quantitative laws of entanglement to provide a set of principles for understanding the behavior of entanglement and how it is used to do information processing

- investigators are working to develop an understanding of the general principles that govern complex quantum systems such as quantum computers

Other developments are equally as interesting or compelling. For example, the quantum state of the object we wish to teleport does not have to describe single microscopic systems like photons, ions, atoms or electrons. Quantum states can describe large collections of atoms like chemical compounds, humans, planets, stars, and galaxies. Hartle and Hawking (1983) even derived the quantum wavefunction of the Universe in closed form, although, it was extremely simplified and excluded the presence of quantum matter-energy. So it has become possible to consider teleporting large quantum systems. We summarize the more recent spectacular developments in the following:

- Generation of entanglement and teleportation by Parametric Down-Conversion (Bouwmeester et al., 1997; Zeilinger, 2003): EPR entangled photon pairs are created when a laser beam passes through a nonlinear β-barium borate or BBO crystal. Inside the crystal (BBO, for example) an ultraviolet photon (λ = 490 nm) may spontaneously split into two lower energy infrared photons (λ = 780 nm), which is called parametric down-conversion. The two "down-conversion" photons emerge as independent beams with orthogonal polarizations (horizontal or vertical). (The orthogonal polarization states represent a classic example of the discrete quantum state variables that can be teleported. Other examples of discrete quantum variables that have been teleported using other schemes include the nuclear magnetic spin of a hydrogen atom, electronic excitations of an effective two-level atom, elementary particle spins, etc.) In the two beams along the intersections of their emission cones, we observe a polarization-entangled two-photon state. For the experimental realization of quantum teleportation, it is necessary to use pulsed down-conversion. Only if the pulse width of the UV light, and thus the time of generating photon pairs is shorter than the coherence time of the down-converted photons, then interferometric Bell-state analysis can be performed. In this type of experiment, the pulses from a mode-locked Ti:Saphire laser have been frequency doubled to give pulses of ≈ 200 fs duration (1 fs = 10^{-15} second). The interfering light is observed after passage through IR filters of 4 nm bandwidth giving a coherence time of ≈ 520 fs. After retroflection during its second passage through the crystal, the UV pulse creates another pair of photons. One of these will be the teleported photon, which can be prepared to have any polarization. Beam splitters and photon detectors are used to perform the Bell-state analysis during the standard teleportation process that ensues. See Figure 8 for a schematic showing the layout of a standard parametric down-conversion entanglement-teleportation experiment.

- Teleportation of squeezed states of light and continuous quantum state variables (Furusawa et al., 1998; Sørensen, 1998; Braunstein and Kimble, 1998; Opatrný et al., 2000; Braunstein et al., 2001; Zhang et al., 2002; Bowen et al., 2002; Bowen et al., 2003; Zeilinger, 2003): Squeezed light (see Section A.2 in Appendix A) is used to generate the EPR entangled beams, which are sent to Alice and Bob. A third beam, the input, is a coherent state of unknown complex amplitude. This state is teleported to Bob with a high fidelity only achievable via the use of quantum entanglement. Entangled EPR beams are generated by combining two beams of squeezed light at a 50/50 beam splitter. EPR beam 1 propagates to Alice's sending station, where it is combined at a 50/50 beam splitter with the unknown input state, in this case a coherent state of unknown complex amplitude. Alice uses two sets of balanced homodyne detectors to make a Bell-state measurement on the amplitudes of the combined state. Because of the entanglement between the EPR beams, Alice's detection collapses Bob's field (EPR beam 2) into a state conditioned on Alice’s measurement outcome. After receiving the classical result from Alice, Bob is able to construct the teleported state via a simple phase-space displacement of the EPR field 2. Quantum teleportation in this scheme is theoretically perfect, yielding an output state which equals the input with a fidelity $F = 1$. In practice, fidelities less than one are realized due to imperfections in the EPR pair, Alice’s Bell measurement, and Bob’s unitary transformation. By contrast, a sender and receiver who share only a classical communication channel cannot hope to transfer an arbitrary quantum state with a fidelity of one. For coherent states, the classical teleportation limit is $F = 0.5$, while for light polarization states it is $F = 0.67$. The quantum nature of the teleportation achieved in this case is demonstrated by the experimentally determined fidelity of $F = 0.58$, greater than the classical limit of 0.5 for coherent states. The fidelity is an average over all input states and so measures the ability to transfer an arbitrary, unknown superposition from Alice to Bob. This technique achieves the teleportation of continuous quantum state variables, as opposed to the discrete quantum state variables used in the Bennett et al. (1993) teleportation protocol and its variants. The teleportation of a squeezed state of light from one beam of light to another demonstrates the teleportation of a continuous feature (of light) that comes from the superpositions of an infinite number of basic states of the electromagnetic field, such as those found in squeezed states. This line of research also involves the experimental demonstration of the mapping of quantum states from photonic to atomic media via entanglement and teleportation. Hald et al. (1999) reported on the experimental observation of a spin-squeezed macroscopic ensemble of 10^7 cold atoms, whereby the ensemble is generated via quantum state entanglement/teleportation from non-classical light to atoms.

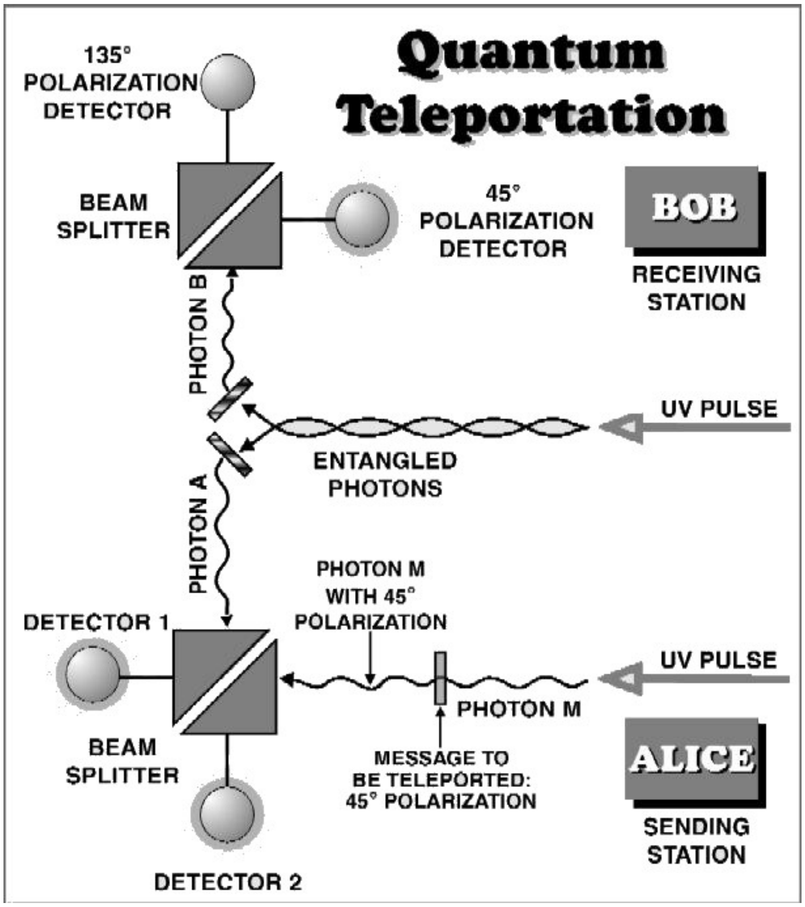
Quantum Teleportation
135° POLARIZATION DETECTOR
BEAM SPLITTER
45° POLARIZATION DETECTOR
BOB
RECEIVING STATION
PHOTON B
UV PULSE
ENTANGLED PHOTONS
PHOTON A
PHOTON M WITH 45° POLARIZATION
DETECTOR 1
UV PULSE
PHOTON M
BEAM SPLITTER
MESSAGE TO BE TELEPORTED: 45° POLARIZATION
ALICE
SENDING STATION
DETECTOR 2

Figure 8. Quantum Teleportation (From www.aip.org)

At the sending station of the quantum teleporter, Alice encodes a "messenger" photon (M) with a specific state: 45 degrees polarization. This travels towards a beam splitter. Meanwhile, two additional entangled photons (A and B) are created. The polarization of each photon is in a fuzzy, undetermined state, yet the two photons have a precisely defined interrelationship. Specifically, they must have complementary polarizations. For example, if photon A is later measured to have horizontal (0 degrees) polarization, then the other photon must collapse into the complementary state of vertical (90 degrees) polarization. Entangled photon A arrives at the beam splitter at the same time as the message photon M. The beam splitter causes each photon to either continue toward detector 1 or change course and travel to detector 2. In 25% of all cases, in which the two photons go off into different detectors, Alice does not know which photon went to which detector. This inability for Alice to distinguish between the two photons causes quantum weirdness to kick in. Just by the very fact that the two photons are now indistinguishable, the M photon loses its original identity and becomes entangled with A. The polarization value for each photon is now indeterminate, but since they travel toward different detectors Alice knows that the two photons must have complementary polarizations. Since message photon M must have complementary polarization to photon A, then the other entangled photon (B) must now attain the same polarization value as M. Therefore, teleportation is successful. Indeed, Bob sees that the polarization value of photon B is 45 degrees: the initial value of the message photon.

- Entanglement of Atoms (Hagley et al., 1997; Sackett et al., 2000): EPR entanglement at the level of atoms has been experimentally demonstrated using rubidium atoms prepared in circular Rydberg states (i.e., the outer electrons of the atom have been excited to very high energy states and are far from the nucleus in circular orbits). The experimental apparatus produces two entangled atoms, one atom in a ground state and the other atom in an excited state, physically separated so that the entanglement is non-local. And when a measurement is made on one atom, let us say the atom in a ground state, then the other atom instantaneously presents itself in the excited state – the result of the second atom wave function collapse, thus determined by the result of the first atom wave function collapse. This work is now evolving towards the demonstration of entanglement for molecules and larger entities followed by teleportation of their states. Bose and Home (2002) have improved on this concept by proposing a single, simple generic method by which any atoms, ions and macroscopic objects can be entangled and teleported.

- Teleportation of an Atomic State via Cavity Decay (Bose et al., 1999; Sackett et al., 2000): It has been shown how the state of an atom trapped in a cavity can be teleported to a second atom trapped in a distant cavity simply by detecting photon decays from the cavities.

- Biological Quantum Teleportation (Mavromatos et al., 2002): There are several obstacles to teleporting large complicated objects, especially biological entities. Decoherence is the primary obstacle. That is because observable quantum effects in biological matter is thought to be strongly suppressed due to the macroscopic nature of most biological entities and the fact that such systems live at near room temperature, and there is always contact between biological entities and the environment (the source of decoherence). These conditions result in very fast collapse of pertinent quantum wavefunctions to one of the allowed classical states of the biological entity. Mavromatos et al. (2002) propose a daring model that predicts dissipationless energy transfer along shielded macromolecules at near room temperatures as well as quantum teleportation of states across microtubules and perhaps neurons. It is proposed that under certain circumstances it is in principle possible to obtain the necessary isolation against environmental decoherence, so that meso/macroscopic quantum coherence, and entanglement extending over scales that are larger than the atomic scale, may be achieved and maintained for times comparable to the characteristic times for biological and cellular processes. Microtubules are comprised of tubulin that is a common polar protein found in the cytoskeleton of eukariotic cells, which is especially enriched in brain tissue. The model treats microtubules as quantum mechanically isolated high-Q QED cavities, exhibiting properties analogous to those of electromagnetic cavities routinely used in quantum optics. The model builds a microtubule network that achieves quantum teleportation of coherent quantum states, leading to decoherence-resistant bulk quantum information processing and computing within the biological matter. It is speculated that the model can explain how consciousness works, and how the brain processes and computes information.

- Teleportation of a laser beam with embedded radio signal (Bowen et al., 2003): The teleportation of a laser beam from one part of a lab to another has been demonstrated. Investigators embedded a radio signal into a laser beam, then disintegrated the beam and reassembled it a meter away, virtually instantaneously. The laser beam was destroyed in the teleportation process, but the radio signal survived. The laser light at one end of an optical communications system was disassembled and its replica was recreated elsewhere in the lab. Even though the laser beam did not survive teleportation, its encoded message did. This system could be used to transport secure data, such that it could become possible to construct a perfect cryptography system. When two parties want to communicate with one another, one can enable the secrecy of the communication to be absolutely perfect.

- Entanglement and Teleportation of a Macroscopic Ensemble of Atoms (Julsgaard et al., 2001): Expanding upon the earlier work of Hald et al. (1999) and Sackett et al. (2000), investigators experimentally demonstrated the entanglement of two macroscopic objects, each consisting of a cesium gas sample containing ≈ 10^{12} atoms. Entanglement is generated via interaction of the samples with a pulse of light, which performs a non-local Bell measurement on the collective spins of the samples. The entangled spin-state can be maintained for 0.5 milliseconds. The teleportation of macro-ensemble atom quantum states is expected to follow this experiment. This work is evolving towards the experimental demonstration of the Bose and Home (2002) proposal, which proved that there is a single generic process that can entangle and teleport any atoms, ions and macroscopic objects.

- Entanglement/teleportation of internal state and external motion information of atoms (Opatrný and Kurizki, 2001): Investigators propose an experiment for transmitting an atom's full information, including its "external" states, such as its energy of motion. This procedure replicates the quantum features of the external motion of a particle. For example, if particle-to-be-teleported C yielded a diffraction pattern after passing through two slits, then the same pattern would be produced by particle B, which receives the teleported information. The researchers propose the following idea: Dissociate a very cold molecule with a laser pulse into two atoms (called A and B). Then manipulate the two atoms so that they become entangled: each one is in a fuzzy state individually, but has a precisely defined relationship with its partner. Then let one of the entangled particles (such as A) collide with particle C, whose unknown state should be teleported. After their collision, the momentum values of the collision partners A and C are measured. With that information, the researchers know how to "kick" and deflect atom B, so that the motion of B precisely emulates that of particle C. The investigators say that state-of-the-art equipment for studying atomic collisions and quantum effects makes this experiment difficult, but feasible, to do. If this proposal proves to be correct, then the implication is that it will become possible to experimentally expand this concept to the teleportation of a large ensemble of atoms, such that the entire physical motion and quantum states of the ensemble can be teleported. This could lead to the future development of a teleportation process similar to what was discussed in Section 3.1.

- Laser-like Amplification of Entangled Particles and Entangled-Photon Lasers (Lamas-Linares et al., 2001): Entangled particles are notoriously difficult to create in bulk. To create entangled photons, for example, researchers use the parametric down-conversion technique to send laser light through a barium borate crystal. Passing through the crystal, a photon sometimes splits into two entangled photons (each with half the energy of the initial photon). However, this only occurs for one in every ten billion incoming photons. To increase the yield, researchers added a step: they put mirrors beyond the crystal so that the laser pulse and entangled pair could reflect, and have the chance to interact. The entangled pair and reflected laser pulse interfere constructively to generate fourfold more two-photon pairs or interfere destructively to create zero pairs. Following these steps, the researchers increased production of two-photon entangled pairs, and also of more rare states such as four-photon entangled quartets. This achievement could represent a step towards an entangled-photon laser, which would repeatedly amplify entangled particles to create greater yields than previously possible, and also towards the creation of new and more complex kinds of entangled states.

This list is by no means complete as new developments in this field continue to arise.

3.3 Conclusion and Recommendations

Given the incredible advancements that have been made in the entanglement and teleportation of macroscopic objects the size of 10^{12} atoms, we are still very far away from being able to entangle and teleport human beings (and even simpler biological entities such as cells, etc.) and bulk inanimate objects (tools, technical equipment, pencils and pens, weapons platforms, communications devices, personal hygiene supplies, etc.). There still remain four essential problems:

- One needs an entangled pair of such bulk objects.
- The bulk objects to be entangled and teleported must be in a pure quantum state (as in a Bose-Einstein condensate, for example). And pure quantum states are very fragile.
- The bulk objects to be entangled and teleported must be extremely isolated from the environment to prevent the onset of decoherence.
- The Bell-state measurement of animate or inanimate objects during entanglement/teleportation will require extracting an amount of information (in bits) that equals or exceeds the number of atoms contained within the object. This infers that the computer storage and processing requirements to entangle and teleport a complete bulk object will be astronomically huge (recall the discussion in Section 3.1).

It is difficult to imagine how we can achieve an extreme level of environmental isolation for an object, let alone a living being that breathes air and radiates heat. Experiments with atoms and larger objects must be done in a high vacuum to avoid collisions with molecules. Thermal radiation from the walls of a teleportation apparatus would easily disturb a tiny amount of matter. At present, decoherence imposes a fundamental limit on quantum entanglement and teleportation. Decoherence is the primary reason why we do not routinely see any quantum effects in our everyday world. Research is continuing on whether decoherence can be reduced, circumvented, or otherwise be eliminated. And some minor progress has been made in that direction.

In q-Teleportation it is the *quantum states* of the objects that are destroyed and recreated, and not the objects themselves. Therefore, q-Teleportation *cannot* teleport animate or inanimate matter (or energy) in its physical entirety. However, some experts argue that because an object's quantum state is its defining characteristic, teleporting its quantum state is completely *equivalent* to teleporting the object, even though the original object's quantum state (and defining characteristic) was completely destroyed in the process. This goes to the heart of what is meant by identity. When an object has all the right properties and features, it will be the same object that one observes whether it was observed now or 24 hours ago. Quantum physics reinforces the point that objects of the same type in the same quantum state are indistinguishable from each other. One should, according to this quantum principle, be able to swap all the atoms in a particular object with the same atoms from a mound of raw materials, and reproduce the original object's quantum states exactly with the end result that the new object is identical to the original. Last, we do not know how to put a human being into a pure quantum state or what doing so would mean for biological functioning (including brain function), but we do know how to put $\leq 10^{12}$ gas atoms/ions and a beam of photons into a pure state in practice. Further research will be required to ascertain whether microbiological and higher-level biological systems, in addition to bulk inanimate matter, can be put into pure quantum states and entangled/teleported.

To perform a Bell-state measurement on (bulk) animate or inanimate objects, during the entanglement/teleportation process, to extract and encode its information will require extracting an amount of information (in bits) that equals or exceeds the number of atoms contained within the object. An object containing a few grams of matter will require the extraction of $> 10^{28}$ bits of data. A simple virus of $\approx 10^7$ atoms would require the extraction of $\geq 10^8$ bits of information during the entanglement/teleportation process, whereas the extraction of a minimum of 10^{28} kilobytes will be required to encode and store an entire human being. This is beyond the capability of present digital electronic computer technology to store and process. It is difficult to see how far computer technology will advance towards meeting this requirement.

It is difficult to fathom what will be in store for the teleportation of human beings given some possible future technology. What about the effects of the q-Teleportation process on the human consciousness, memories and dreams, and the spirit or soul? We know from quantum physics that "the whole is greater than the sum of its parts." So what happens to the fundamental characteristics of a human being when he/she steps into the teleporter-transmitter, where their quantum states (i.e., their complete identity) are destroyed during the quantum entanglement/teleportation process, and then their copy is created at the teleporter-receiver an instant later? What will things be like during the entanglement process? Will a teleported individual's consciousness, memories and dreams, and spirit/soul be successfully and accurately teleported or not? This is a major ethical and technical question that will have to be addressed by future research.

•**Recommendations**:

- Broad-spectrum Quantum Computing Technology Development Program: At present, the Quantum Information Science Program (QISP) is coordinated by the U.S. Army Research Office with funding and support from the Army, the National Security Agency, DARPA, and the Office of the Deputy Director of Defense for Research and Engineering. The Naval Research Lab and the CIA are both involved in their own programs. The CIA vets new commercial development of computer technology and computer information processing via its In-Q-Tel company (reference 44). This includes R&D on quantum entanglement and teleportation for computer, information processing and secure communications. QISP was funded for $19 million in 1999. The program involves 34 projects by researchers at 21 universities, three government laboratories and two corporate laboratories. QISP goals include building a quantum computer, developing quantum information processing, and further advances in quantum teleportation. The AFRL should join QISP and provide partnership funding on the order of $1 million per year. An alternative to this would be for AFRL to collaborate with In-Q-Tel and participate in its technology R&D venture capital programs. This R&D investment would allow the Air Force to acquire very advanced quantum physics and related technological applications that can support its mission. The R&D investment benefits would include the development and implementation of quantum computing/information processing and secured quantum communications technology, which can significantly enhance the performance and security of Air Force computing and communication systems infrastructure, and aerospace weapons systems.

- Quantum Cryptography: A dedicated research program should be implemented to develop a mature quantum cryptography technology. Theoretical and experimental work is in progress among a small number of select groups (QISP, In-Q-Tel, universities, etc.), but this field is not advancing fast enough for practical applications to become available to meet increasing adversarial threats against secured military and intelligence communications. The goal of proposed quantum cryptography research is to bring the theoretical and experimental foundation of quantum cryptography and secure quantum information processing to maturity, and to fully develop and implement quantum entanglement/teleportation-based cryptography technology. Recent experimental work has demonstrated that a completely secure quantum key can be generated and distributed for the communication and decoding of encrypted messages using entangled photons. Any eavesdropper's attempt to intercept the quantum key will alter the contents in a detectable way, enabling users to discard the compromised parts of the data. There is much more work that needs to be done in this area. I recommend that the AFRL implement a

$1 million/year program for five years in order to advance the state-of-art in quantum cryptography technology.

- Quantum Decoherence: Decoherence is the primary reason why we do not routinely see any quantum effects in our everyday world. And it imposes a fundamental limit on quantum entanglement and teleportation via the interaction between entangled/teleported quantum systems and their local environment. In order to advance quantum entanglement/teleportation physics and develop applied technologies, it is necessary that a research program be implemented by the AFRL to explore whether decoherence can be significantly reduced, circumvented, or otherwise be eliminated. An insufficient number of small university groups have slowly made minor progress in this direction. I recommend that a $500,000 - 750,000 per year R&D program be conducted for five years to overcome this technical challenge.

- Pure Quantum States: In order to entangle and teleport quantum particles and bulk objects, they both must be prepared in a pure quantum state. And pure quantum states are very fragile to decoherence. A technical challenge for entanglement/teleportation physics is whether the requirement for pure quantum states can be relaxed and how much decoherence will play a role in this situation, what technical challenges will arise when increasing the size of entangled/teleported matter to larger macroscopic scale ($>> 10^{12}$ atoms), and whether matter of mixed composition (such as a gas or Bose-Einstein condensate of mixed atomic elements) can be entangled/teleported in both pure and mixed quantum states. I recommend that a $250,000 – 500,000 per year research program be conducted for five years to study this problem.

- Entangling Bulk Matter and Bell-State Measurement to Extract Information: Recent experiments demonstrated the entanglement of two macroscopic objects, each consisting of a cesium gas sample containing $\approx 10^{12}$ atoms. Entanglement was generated via interaction of the samples with a pulse of light, which performs a non-local Bell measurement on the collective spins of the samples. In order to push the envelope on this development and take it to higher practical levels, it will be necessary to ascertain the limit on the size and composition of bulk matter entanglement (given the decoherence and pure-state constraints); and to determine what other quantum states can be used for entanglement, what other Bell-state measurement techniques can be used, and whether multiple quantum states can be entangled. The chief technical challenge is the computer technology that will be required to facilitate the huge amount of data that must be extracted, processed and stored from bulk matter quantum states during the Bell-state measurement process. I recommend that a $500,000 – 1 million per year research program be implemented for five years in order to explore these questions and ascertain what solutions may be technically available, and to develop such solutions.

- Biological Quantum Teleportation: The Mavromatos et al. (2002) theoretical model for biological entanglement and teleportation is a remarkable concept that could result in the development of a workable physics theory of consciousness. The model has potential applications to advanced quantum computing/information processing physics and the physics of psi phenomena (see Chapter 5). A research program should be implemented to continue the Mavromatos et al. (2002) work and bring their model to theoretical maturity. It is recommended that this program be funded at $500,000 – 800,000 per year for five years. A parallel or follow-up program should be implemented to experimentally test this model and ascertain any useful technological applications. One application that should be explored in the proposed research program is advanced, ultra-fast, ultra-high-capacity quantum computing and information processing using natural and/or artificial biological systems. The parallel or follow-up experimental research program should be funded at $800,000 – 1.5 million per year for five years.

- FTL Communication: Experiments verifying that EPR entanglement obeys Special Relativity (Seife, 2000; Scarani et al., 2000; Gisin et al., 2000; Zbinden et al., 2000a, b) determined that quantum information via EPR photon pair entanglement must travel > 10^7 times light speed. Can this mechanism be exploited to achieve FTL communication? If so, then the potential military and commercial applications will be revolutionary, and the science and industry of communications will be forever transformed. A comprehensive theoretical and experimental research program should be implemented to answer this question. It is recommended that this program be funded at $700,000 – 1 million per year for five years. A modest experiment definition study should be funded at $80,000 for one year to delineate the most promising experimental approaches to be used for the larger research program. [There is much controversy and debate over FTL (a.k.a. superluminal) signals/communication, and the reader should see the selected superluminal references in the Teleportation References section of this study.]

- New Entanglement/Teleportation Breakthroughs: The most exciting developments in quantum teleportation physics has included the teleportation of a laser beam with an embedded radio signal, the teleportation of squeezed states of light (and hence, continuous quantum state variables), the teleportation of photon states to atoms/ions (from light to matter!), the entanglement of two similar/dissimilar quantum particles that are created by two (independently) different particle sources, the laser-like amplification of entangled particle/photon pairs, parametric down-conversion entanglement and teleportation (of discrete quantum state variables), quantum cryptography with unbreakable keys, the teleportation of quantum information at speeds > 10^7 times light speed, the entanglement and teleportation of macroscopic (10^{12} atoms) matter quantum states, etc. There is also the yet-untested proposal to entangle/teleport the external physical motion and internal quantum state information of atoms. This shows that quantum physics sets no apparent limit on what it is that can be teleported/entangled and how it is to be teleported/entangled, or where it is to be teleported/entangled. At present teleportation technology requires fiber optic and coaxial cables to teleport quantum state information from one location to another. Can we avoid the use of cables and teleport through free space? [Note: Before this report went to press, Aspelmeyer et al. (2003) reported their outdoor experiment that demonstrated the distribution of quantum entanglement (of laser photons) via optical free-space links to independent receivers separated by 600 m across the Danube River (during inclement nighttime weather), with no line of sight between them. This experiment is revolutionary and begins the step toward conducting satellite-based distributed quantum entanglement.] We have not discovered all the possibilities that nature has in store for us. The present breakthrough discoveries will likely introduce novel military and intelligence technology applications in the near and far future. But further R&D must be conducted in order to discover new applications for these recent breakthroughs, to make additional breakthroughs and discoveries, and to advance the state-of-art in quantum teleportation physics to meet future challenges to the Air Force mission. I recommend that a two-track R&D program be implemented over five years. The first track should be funded at $250,000 – 750,000 per year for the purpose of developing new entanglement/teleportation breakthroughs in quantum teleportation physics. The second track should be funded at $750,000 – 1.5 million per year for the purpose of developing applications for any new breakthroughs with the proviso that such applications benefit the Air Force mission and have commercial dual-use capability to leverage advance technology in the private sector.

4.0 e-TELEPORTATION

4.1 Extra Space Dimensions and Parallel Universes/Spaces

A literature search for proposed e-Teleportation concepts based on the conveyance of objects through extra space dimensions and/or parallel universes/spaces has yielded only one result (see Section 4.2). The present state-of-art in research on parallel universes/spaces and extra space dimensions has been strictly limited to the work on developing a grand unified quantum field theory and a quantum theory of gravity, whereby the former necessarily includes the latter. Quantum gravity/unified field theory research has been evolving since the 1920s when Kaluza and Klein published the first papers to describe a model for the unification of gravity with electrodynamics. Many of the more prominent theories today invoke extra spatial dimensions, the existence of parallel universes/spaces, or both in order to quantise gravity and/or to unify gravity with the other forces of nature. It is beyond the scope of this study to provide an in-depth review of all of the research that has been done in this area, so we list below a select few of the historically prominent models that have largely gained a secure foothold in present-day research:

- Kaluza-Klein Electromagnetic-Gravity Unification Theory/Modern Kaluza-Klein Gravity Theories (Kaluza, 1921; Klein, 1926; de Sabbata and Schmutzer, 1983; Lee, 1984; Appelquist et al., 1987; Kaku, 1993, 1994; Overduin and Wesson, 1998): It was originally suggested that Maxwellian electrodynamics and Einstein gravitation could be unified in a theory of five-dimensional Riemannian geometry, where the gravitational and electromagnetic potentials together would determine the structure of spacetime. The fifth space dimension is curled up into a ball of space with a radius slightly larger than 10^{-35} m, and it was originally regarded as having no physical significance because it was simply a mathematical tool used to catalyze unification. At present, the generic name of Kaluza-Klein stands for a wide variety of approaches to quantising and unifying gravitation with other quantum fields using any number of dimensions greater than four.

- Superstring Theories (Green, 1985; Kaku, 1988, 1993, 1994): These theories come in a wide variety of interrelated concepts, and they are a highly evolved form of Kaluza-Klein theories. They are based on the dynamics of string-like fundamental quanta, whereby the observed fundamental particles are manifested by the vibrational ground or excitation states of a quantum string (open or closed loop). The superstrings are $\approx 10^{-35}$ m (i.e., the Planck length) in size. There are different versions of these theories that require ten, eleven or twenty-six extra space dimensions to unify and quantise gravity, whereby the extra dimensions are curled up (i.e., compactified) into balls of space with a radius $< 10^{-35}$ m. These theories later evolved into versions that are now called F- and M-theory. The mathematics behind this class of theories is very ugly, and it is difficult for even the best superstring theorists to make simple or sophisticated calculations and predictions. And so far, this class of quantum gravity theories has escaped experimental verification.

- D-Brane and 3-Brane Theories/Parallel Spaces (Rubakov and Shaposhnikov, 1983a, b; Polchinski, 1995; Antoniadis et al., 1998; Randall and Sundrum, 1999a, b; Weiss, 2000; Pease, 2001; Arkani-Hamed et al., 1998, 2000, 2002): D-brane theory is a recent incarnation of the original superstring theories in which open strings, corresponding to the fundamental particles of the standard model (quarks, leptons, gauge bosons), have their free ends stuck on a (hypersurface) membrane called a D-brane (D = Dirichlet boundary conditions). But the graviton, which corresponds to a closed loop of string, can propagate in all the dimensions. It provides both unification and quantization of gravity by assuming that there are n new spatial dimensions in addition to the three infinite spatial dimensions we know about. And the extra space dimensions are $\approx 10^{-35}$ m in extent. A very recent alternative version of this model is called "3-brane" theory. In this theory, each of the n extra space dimensions is of finite extent $R \approx 2\times10^{(32/n)-17}$ centimeters. The space spanned by the new dimensions is called "the bulk." In this theory, the particles of the standard model live within our familiar realm of three spatial dimensions, which forms a three-dimensional (hypersurface) membrane or "3-brane" within the bulk. The propagation of electroweak and strong nuclear forces is then confined to our 3-brane. However, at distances (r) less than R, gravity (via gravitons) propagates in the full $(3 + n)$-dimensional space, whereby its strength falls as $r^{-(2+n)}$ with increasing separation r. When $r > R$, the gravitational force reverts to its normal Newtonian r^{-2} falloff because there is no longer any extra-dimensional space for it to spread into. If $n = 1$, then the size of the extra-dimension would have to be $R \approx 2\times10^{15}$ cm (or 2×10^{10} km = 133.3 AU; 1 AU = 1.5×10^{8} km is the mean Earth-Sun distance) in order to account for the weakness of gravity, but an extra space dimension this large would have already made itself obvious in the observed dynamics of the solar system. For this reason, investigators have discounted the possibility that $n = 1$. If $n = 2$, then the size of both extra space dimensions would have to be $R \approx 0.2$ cm (or 2 mm). In any case, inconspicuous neighboring 3-branes may be separated from the 3-brane we live on by only a fraction of a millimeter, or even much smaller distances, across the higher-dimensional bulk. Such neighboring 3-branes may be distant folds of our own 3-brane, with the same physics, but able to influence us across shortcuts through the bulk. Or they may be completely separate 3-branes possessing their own fundamental laws and parameters of nature that are completely different from our own. Several tabletop Cavendish-type experiments are now looking for sub-millimeter deviations from Newtonian gravitation as a first step towards verifying 3-brane theory, and other experiments are now being planned or are already underway (Pease, 2001). At present the preliminary experimental results have been negative for the existence of extra space dimensions, and the experimental data suggests that two extra space dimensions are now constrained to length scales << 0.2 – 0.3 millimeters while seven extra space dimensions can be no larger than 2 femtometers (Pease, 2001).

- Parallel Universes/Parallel Spaces (Everett, 1957; Wheeler, 1957, 1962; DeWitt, 1970; DeWitt and Graham, 1973; Jammer, 1974; Davies, 1980; Wolf, 1988; Kaku, 1994; Visser, 1995 and Section 2.1): There are only two other research tracts that are concerned with parallel universes besides 3-brane theory. The first tract is the traversable wormhole research that was discussed in Section 2.1. Traversable wormholes can connect many different universes in the "multiverse" (i.e., a conglomeration of many universes), and these are called inter-universe wormholes. However, traversable wormhole physics (a.k.a. Einstein's General Relativity Theory) does not provide a physical prescription for the existence and nature (i.e., fundamental parameters and physical laws) of other putative universes. The difference between inter-universe and intra-universe (i.e., two distant regions of one universe are connected with each other) wormholes arises only at the level of global geometry and global topology. Local physics near the throat of a traversable wormhole is insensitive to issues of intra-universal or inter-universal travel. An observer in the vicinity of the throat, while making local measurements, would not be able to tell whether he was traveling to another universe or to a remote part of our own universe. And one cannot rely on the topological (as opposed to geometrical) information to determine which is the case, because topological information is not enough to uniquely characterize an inter-universe connection. And General Relativity Theory does not fix the topology of spacetime, so we cannot ascertain the existence of other universes. [Note: Traversable wormholes are also geometrically possible for higher dimensional spaces.] The second tract is the "Many Worlds" interpretation of quantum theory. This version of quantum theory requires the simultaneous existence of an infinite number of equally real worlds, all of which are more-or-less causally disjoint, in order to interpret consistently the relationship between observed phenomena and observers. The theory was proposed in an attempt to overcome a number of deep paradoxes inherent in the interpretation of the theory of measurement and quantum theory. The Many Worlds theory argues that quantum theory requires the existence of a "superspace" of worlds spanning the range of all possible quantum observations (or quantum measurements). Through our acts of measurement we are imagined to trace a path through the mesh of possible outcomes. All the "worlds" are causally disjoint, and the uncertainty of quantum observation can be interpreted as an artifact of our access to such a limited portion of the superspace of possible worlds. The evolution in the superspace as a whole is entirely deterministic.

At present, none of the theoretical concepts outlined above have been brought to a level of technical maturity, where it becomes meaningful to ascertain whether any form of e-Teleportation is theoretically possible between extra space dimensions and different or parallel universes/spaces. However, as mentioned in the item on parallel universes/parallel spaces, there is the exception that traversable wormholes (three- and higher-dimensional) provide a solid physics principle for the implementation of teleportation between parallel universes/spaces. And traversable wormholes can be devised to connect 3-branes together. See Section 2.1 for the discussion on teleportation via traversable wormholes. Also, Kaluza-Klein theories, superstring theories and D-brane theory all have the common feature that their extra space dimensions are $\leq 10^{-35}$ m in extent, which makes it impossible for any useful form of macroscopic-level teleportation to occur between space dimensions. Last, it is not yet possible to do theoretical calculations or even experimentally verify most of these theories. Three-brane theory is the best parallel space theory there is, with the possibility that macroscopic-level teleportation is possible between space dimensions (only if the extra space dimension(s) has length scale(s) >> millimeters). But this theory is still in the stage of maturing theoretically and achieving experimental verification (or falsification). Therefore, we can go no further in this section.

4.2 Vacuum Hole Teleportation

An unusual teleportation concept has been proposed by Leshan (1999, 2002), which describes the teleportation of objects throughout our universe by using the geometrical properties of spacetime. The proposal posits that there is a "zero-space" that exists outside the boundary of our universe, whereby this zero-space is a "point form" space, where the distance between any two points is always equal to zero. Leshan also calls this space a "hole." Further requirements and assumptions of the model are:

- time does not exist as a property in zero-space
- the cosmological principle (i.e., there are no privileged frames relative to another place or point in the universe) requires that the boundary or border of the universe must pass through every point of space
- virtual holes (or zero-space) in spacetime must exist at every point of the universe, which are also called "vacuum holes"
- vacuum holes exist as virtual particles

The last item is interesting because it implicitly says that vacuum holes (a.k.a. zero-space) must also be virtual particles, and in Section 2.2 we showed that virtual particles are a representation of the vacuum ZPF. Therefore, this infers that vacuum holes can be considered to be vacuum zero-point fluctuations in

Leshan's model. Thus, a teleportation mechanism can arise in this model because distances between zero-space and any other point in the universe are zero, so that the vacuum holes can potentially exist at every point in the universe simultaneously. Therefore, if an object is sent "out of the universe" and into a vacuum hole (a.k.a. zero-space), then the object can appear at random at any spacetime point in the universe.

The mechanism for teleportation in this model is:

- to send an object outside of the universe by creating a closed surface (i.e., "hole sphere"), which consists of vacuum holes, around the object;
- while inside the hole sphere, the object then ceases to exist because objects cannot really exist outside of the universe;
- however, the object simultaneously exists at any other remote location in the universe (via the cosmological principle) at the instant it became enclosed by the hole sphere;
- therefore, it has been teleported to some remote location in the universe

Leshan points out that the teleportation device must curve spacetime so that the starting and destination points in the universe coincide, and the curved geometry must be similar to that of a black hole for an instant, so that a channel between the two points can be formed. (This sounds suspiciously like creating a traversable wormhole via an Einstein-Rosen bridge, which can be made traversable by perturbing the Schwarzschild spacetime metric an infinitesimal amount.) There is no space to traverse, so therefore there will be no passage of time during teleportation. The only expenditure of energy in this teleportation scheme is the energy that will be needed to curve spacetime.

This teleportation concept is very convoluted. Leshan does not offer any further explanations that are useful nor does he offer any precise technical description for the vacuum holes, and how they are to be produced and manipulated. There is also no mathematical physics derivation published by Leshan to support this concept. I am totally unable to evaluate this concept in the absence of a rigorous theoretical framework. This concept is too sketchy and full of technical "holes" to seriously consider it any further for this study. The reader should note that it has already been demonstrated that traversable wormholes are the best physical principle available to implement teleportation between universes and extra space dimensions.

4.3 Conclusion and Recommendations

At present, none of the theoretical concepts explored in this chapter have been brought to a level of technical maturity, where it becomes meaningful to ascertain whether any form of e-Teleportation is theoretically possible between extra space dimensions and different or parallel universes/spaces. However, there is the exception that traversable wormholes (three- and higher-dimensional) provide a solid physics principle for the implementation of teleportation between parallel universes/spaces. And traversable wormholes can be devised to connect 3-branes together. Kaluza-Klein, superstring and D-brane theories do not allow for any useful form of macroscopic-level teleportation to occur between space dimensions, because these theories require that the extra space dimensions be $\leq 10^{-35}$ m in extent. Last, it is not yet possible to do theoretical calculations to make predictions or even to experimentally verify most of these theories. Three-brane theory is the best parallel space theory there is with the possibility that macroscopic-level teleportation is possible between space dimensions. But this theory is still in the stage of maturing theoretically and getting experimental verification.

•**Recommendations**:

- The recommendations outlined in Section 2.3 are relevant to the investigation of the possibility for e-Teleportation.

Approved for public release; distribution unlimited.

5.0 p-TELEPORTATION

5.1 PK Phenomenon

P-Teleportation is a form of psychokinesis (or PK) similar to telekinesis but generally used to designate the movement of objects (called apports) through other physical objects or over great distances. Telekinesis is a form of PK, which describes the movement of stationary objects without the use of any known physical force. And PK is essentially the direct influence of mind on matter without any known intermediate physical energy or instrumentation. Rigorously controlled modern scientific laboratory PK, and related psychic (a.k.a. “psi”, “paranormal” or parapsychology), research has been performed and/or documented by Rhine (1970), Schmidt (1974), Mitchell (1974a, b, see also the references cited therein), Swann (1974), Puthoff and Targ (1974, 1975), Hasted et al. (1975), Targ and Puthoff (1977), Nash (1978, see also the references cited therein), Shigemi et al. (1978), Hasted (1979), Houck (1984a), Wolman et al. (1986, see also the references cited therein), Schmidt (1987), Alexander et al. (1990), Giroldini (1991), Gissurarson (1992), Radin (1997, see also the references cited therein), Tart et al. (2002), Shoup (2002), and Alexander (2003).

A well-known theoretical/experimental/operational program directed by H. E. Puthoff, R. Targ, E. May and I. Swann was conducted at SRI International and the NSA, and sponsored at various times by the Central Intelligence Agency (CIA), the Defense Intelligence Agency (DIA), and the Army Intelligence and Security Command (INSCOM) over more than two decades; and the program was later carried on by E. May at SAIC (Alexander, 1980; Puthoff, 1996; Targ, 1996; Schnabel, 1997; Tart et al., 2002). This was called the Remote Viewing program, and it was a compartmentalized special access program possessing a variety of codenames during its 22 years of operation. Remote viewing involves precognition and clairvoyance, and it allows a practitioner to acquire information irrespective of intervening distance or time. The Remote Viewing program ended in 1994 and President W. J. Clinton officially declassified it in 1995. The reader should note that the very first U. S. military-intelligence R&D programs on psi, PK and mind control were conducted by H. K. (Andrija) Puharich, M.D., L.L.D during his military service at the Army Chemical and Biological Warfare Center at Fort Detrick, Maryland in the 1940s-50s. Puharich had an interest in clairvoyance and PK, and dabbled in theories for electronically and pharmaceutically enhancing and synthesizing psychic abilities. While in the Army, Puharich took part in a variety of parapsychology experiments, and he lectured Army, Air Force and Navy groups on possibilities for mind warfare. He was a recognized expert in hypnotism and microelectronics.

PK phenomenon was also explored in the Remote Viewing program. Col. J. B. Alexander (USA ret.) credits professional aerospace engineer Jack Houck for “capturing PK phenomenon and transitioning it into an observable form” (Houck, 1982, 1984a, b; Alexander et al., 1990; Alexander, 2003). During the past three decades, Houck (along with Alexander) held a number of PK sessions, whereby attendees are taught the PK induction process, and initiate their own PK events using various metal specimens (forks, spoons, etc.). Individuals were able to completely bend or contort their metal specimens with no physical force being applied whatsoever. Numerous government science advisors and senior military officials took part in and/or witnessed these events, which took place at the Pentagon, at officers’ or scientists’ homes, and at one quarterly INSCOM retreat attended by the commanding general and a group of colonels and generals commanding INSCOM units around the globe. Spontaneous deformation of the metal specimens was observed at the PK session conducted during the INSCOM retreat, causing a great deal of excitement among those present. Other notable trained observers were also present at this session, and they critically reviewed the events. Psychic Uri Geller (1975) is the original model for demonstrating

PK metal bending. During a talk that he gave at the U.S. Capitol building, Uri caused a spoon to curve upward with no force applied, and then the spoon continued to bend after he put it back down and continued with his talk (Alexander, 1996). Jack Houck continues doing extensive experimental work and data collection on micro- and macro-PK phenomena. Scientifically controlled PK experiments at the Princeton University Engineering Anomalies Research Laboratory were conducted by Robert Jahn (Dean Emeritus of the School of Engineering), who reported that repeatedly consistent results in mentally affecting material substances has been demonstrated in the lab (Jahn and Dunne, 1987). In the 1980s, Jahn attended a meeting on the PK topic at the Naval Research Laboratory, and warned that foreign adversaries could exploit micro- or macro-PK to induce U.S. military fighter pilots to lose control of their aircraft and crash.

Very early investigations of, and experiments on, p-Teleportation occurred during the 19th and early 20th centuries. Many cases that were studied, and the experiments that were performed, were undoubtedly due to fraud, and few experiments have occurred under controlled conditions during that period. However, most of the credible, scientific reports of p-Teleportation phenomenon and related (controlled) experiments occurred in the late 20th century (see for example, Alexander et al., 1990; Radin, 1997). Some of that scientific work involved the investigation of Uri Geller and a variety of other recurrent spontaneous PK phenomena (Hasted et al., 1975; Puthoff and Targ, 1975; Targ and Puthoff, 1977; Nash, 1978; Wolman et al., 1986). Psychics Uri Geller (1975) and Ray Stanford (1974) claimed to have been teleported on several occasions. Most claimed instances of human teleportation of the body from one place to another have been unwitnessed. There are also a small number of credible reports of individuals who reported being teleported to/from UFOs during a UFO close encounter, which were scientifically investigated (Vallee, 1988, 1990, 1997). But there are a larger number of such reports that are anecdotal, whereby the witness data tends to be unreliable. However, we will confine our discussion to the controlled laboratory experiments that have been performed and reported.

One of the more interesting examples of controlled experiments with Uri Geller was one in which he was able to cause a part of a vanadium carbide crystal to vanish (Hasted et al., 1975). The crystal was encapsulated so it could not be touched, and it was placed in such a way that it could not be switched with another crystal by sleight of hand. A more spectacular series of rigorously controlled (and repeatable!) laboratory experiments occurred in the Peoples Republic of China (PRC). In September 1981, an extraordinary paper was published in the PRC in the journal Ziran Zazhi (transl.: Nature Journal), and this paper was entitled, “Some Experiments on the Transfer of Objects Performed by Unusual Abilities of the Human Body” (Shuhuang et al., 1981). The paper reported that gifted children were able to cause the apparent teleportation of small objects (radio micro-transmitters, photosensitive paper, mechanical watches, horseflies, other insects, etc.) from one location to another (that was meters away) without them ever touching the objects beforehand. The experiments were operated under exceptionally well-controlled conditions (both blind and double-blind). The researchers involved included not only observers from various PRC colleges and medical research institutes, but also representatives from the PRC National Defense Science Commission. Because of the involvement of the latter, it was deemed necessary that an unclassified Intelligence Information Report be prepared by the DIA (see Shuhuang et al., 1981), which included a detailed English translation of the article.

Additional research carried out by the Aerospace Medicine Engineering Institute in Beijing was reported in the July 1990 issue of the Chinese Journal of Somatic Science (Kongzhi et al., 1990; Jinggen et al., 1990; Banghui, 1990), which was also translated into English by the DIA. Reported in several articles are experiments involving the videotaping and high-speed photography of the transfer of test specimens (nuts, bundles of matches, pills, nails, thread, photosensitive paper, chemically treated paper, sponges dipped in $FeCl_3$, etc.) through the walls of sealed paper envelopes, double layered KCNS type paper bags, sealed glass bottles and tubes with sealed caps, and sealed plastic film canisters without the walls of any of these containers being breached. All of the Chinese experiments reported using gifted children and young adults, who possessed well-known extraordinary PK ability, to cause the teleportation of the various test specimens. In all the experimental cases that were reported, the test specimens that were teleported were completely unaltered or unchanged from their initial state, even the insects were unaffected by being teleported. The experiments were well controlled, scientifically recorded, and the experimental results were always repeatable.

The Chinese papers are all extremely interesting and very well written, and they show photographs and schematic diagrams of the various experimental setups. The experimental protocols were explained in lengthy detail, and thorough data and statistical analysis were presented in the results. The combined results from the several Chinese experiments showed that:

- different research groups designed different experimental protocols, used different gifted psychics, used different sealed containers, and used different test specimens (live insects, bulk inanimate objects, and even radio micro-transmitters were used to track the location of the specimens) that were to be teleported;

- the time required for the teleportation of test specimens through various barriers was anywhere from a fraction of a second to several minutes, and this was not dependent on the test specimen that was used, the sealed container that was used (or its barrier thickness), which experimental protocol was used, or which psychic was being used

- the high-speed photography/videotaping recorded in one series of experiments that test specimens would physically "meld" or blend with the walls of sealed containers; and recorded in a different series of experiments that test specimens would simply disappear from inside the container only to reappear at another location (after seconds to several minutes of time transpired), such that the test specimen did not actually undergo total material disintegration/reintegration during teleportation – this data is important, because without the aid of electronic monitoring instruments, the average person's sensory organs and usual methods of detection are temporarily unable to perceive the test specimen's (ambiguous) existence during the teleportation process;

- the radio micro-transmitter used as a test specimen in one series of experiments (Shuhuang et al., 1981) transmitted a radio signal to several stationary electronic instruments/receivers, so that the specimen could be tracked and monitored (via signal amplitude and frequency measurements) during the teleportation process; the experimenters discovered that there was large fluctuations in the intensity (in both amplitude and frequency) of the monitored signal to the effect that it would either completely disappear or become extremely weak (to the extent that the monitoring instruments could scarcely detect it) – it was discovered that there was a definite correlation between the change in strength (i.e., radical frequency shifts were observed) of the monitored radio signal and the teleportation of the test specimen, such that the weak or absent signal indicated that the specimen was "nonexistent" (or in an altered physical state) during teleportation (note: the monitored signal amplitude and frequency of the micro-transmitter specimen were stable before and after teleportation);

- before and after "passing through the container wall/barrier", the test specimen and the container's wall/barrier are both complete solid objects;

- the gifted psychics were never allowed to see (they were blindfolded in many experiments) or touch each of the test specimens or the sealed containers before and after experiments were conducted, and only the experimenters touched the specimens and containers (using both blind and double-blind protocols);

- the experimental results were all repeatable

- the conditions for fraud and sleight of hand were totally eliminated, and multiple independent outside witnesses (technical and military-intelligence experts) were present at all times to ensure total fidelity of the experiments

The experimental radio micro-transmitter and high-speed photography/videotaping data offer an important clue on what the teleportation mechanism is, and this will be discussed further in Section 5.1.1. The Chinese were unable to offer any significant physics hypothesis that could explain their results. Some researchers stated that it is necessary to invoke a new physics, which somehow unifies the human consciousness (i.e., physics of consciousness) with quantum and spacetime physics, in order to understand p-Teleportation and related PK phenomena. The researchers were amazed by their repeated results, and were barely able to fathom the altered "state of being" that test specimens underwent during teleportation.

It is also important to point out that during the Cold War the DIA produced three (now declassified) reports on the parapsychology research of the Soviet Union and its Warsaw Pack allies (LaMothe, 1972; Maire and LaMothe, 1975; DIA Report, 1978; other related studies were reported by Groller, 1986, 1987). The purpose of the reports was to collate and summarize collected intelligence, describe in great detail, and assess the Soviet Union and Warsaw Pact R&D on parapsychology and paraphysics. The reports outlined the history of pre-revolutionary (Czarist) Russian, and WWII and post-WWII era Soviet R&D on psychotronics, human mind/behavior control, and the entire spectrum of parapsychology. The Soviet information also mentions the psychotronic/parapsychology R&D materials that Soviet military forces took from various Nazi research centers in and around Germany at the end of WWII. The entire spectrum of parapsychology phenomena was explored by the Soviets, which resulted in the generation of a wealth of experimental data and related scientific research literature. One DIA report noted that there was an East versus West science debate in the Soviet literature over whether paranormal phenomenon and related experimental data was real or even scientifically sound in comparison to western scientific practice and philosophy. Another DIA report lists the names and affiliations of all the researchers, as well as the names of the various Soviet and Warsaw Pact research centers, that were involved. Also, Pratt (1986) reviews and summarizes the history of Soviet psychotronics research.

The U.S. military-intelligence establishment was concerned with the possibility that the Soviets and their Warsaw Pact allies were conducting psychotronics and mind control R&D in order to discover how to exploit and control powerful phenomena that could be used against the U.S. and its allies. LaMothe (1972) chronicled how the Soviets had been researching methods of influencing human behavior for over sixty years. The Soviets and their allies extensively explored an influence technology that they called "controlled offensive behavior", which is defined as "research on human vulnerability as it applies to methods of influencing or altering human behavior" (LaMothe, 1972). Also, LaMothe (1972) describes the revolutionary techniques the Soviets studied to influence human behavior, which included: sound, light, color, odors, sensory deprivation, sleep, electromagnetic fields, biochemicals, autosuggestion, hypnosis, and parapsychology phenomena (such as psychokinesis, telekinesis, extrasensory perception-ESP, astral projection, clairvoyance, precognition, and dream state, etc.). The LaMothe (1972) report became an aid in the development of countermeasures for the protection of U.S. and/or allied personnel.

Psychotronics is the general term that was used in the former Soviet Union/Warsaw Pact countries to categorize many psychic phenomena undergoing scientific research. The conclusions that were reached in the DIA reports are that within the category of psychotronics, the Soviets identified two discrete skills (LaMothe, 1972):

- bioenergetics: those phenomena associated with the production of objectively detectable effects such as psychokinesis, telekinesis, levitation effects, transformations of energy, i.e. the altering or affecting of matter

- bioinformation: those phenomena associated with the obtaining of information through means other than the normal sensory channels (i.e., ESP), such as telepathy, precognition, and clairvoyance, i.e., using the mind to tap into the thoughts of others or to acquire present or future information about objective events in the world

These phenomena involve using the mind and/or some “field” of the body to affect other minds and inanimate objects irrespective of intervening distance or elapsed time, and without engaging any conventional tools. Bioenergetics and bioinformation are two classifications that form a single branch of science the Soviets preferred to call biocommunications. Soviet biocommunications research is primarily concerned with exploring the existence of a definite group of natural phenomena controlled by laws that are not based on any known (energetic) influence. The types of biocommunication (a.k.a. psychotronics) phenomena includes special sensory biophysical activities, brain and mind control, telepathic communications or bioinformation transceiving, bioluminescent and bioenergetic emissions, and the effects of altered states of consciousness on the human psyche. Psychotronics and remote viewing provide capabilities that have obvious intelligence applications. The Soviets and their Warsaw Pact allies invested millions of dollars in psychotronics R&D because they understood this, and saw the potential payoff for military and intelligence applications.

The U.S. response to Soviet psychotronics R&D programs was the Remote Viewing program. In addition, the U.S. Army began the JEDI Project in 1983, which sought to increase human potential using teachable models of behavioral/physical excellent by unconventional means (Alexander et al., 1990). The JEDI Project was essentially a human-performance modeling experiment based on neuro-linguistic programming (NLP) skills, whereby advanced influence technologies to model excellence in human performance was used. The program ran under the auspices of the Army INSCOM and the Organizational Effectiveness School, and was sponsored by a U.S. government interagency task force. Finally, it should be pointed out that the program had successfully trained several hundred people, including members of Congress (such as Al Gore, Jr. and Tom Downey), before being terminated.

There is a wealth of factual scientific research data from around the world attesting to the physical reality of p-Teleportation and related anomalous psi phenomena (Mitchell, 1974b; Targ and Puthoff, 1977; Nash, 1978; Radin, 1997; Tart et al., 2002). The skeptical reader should not be so quick to dismiss the subject matter in this chapter, because one must remain open-minded about this subject and consider p-Teleportation as worthy of further scientific exploration. The psychotronics topic is controversial within the western scientific community. The debate among scientists and scientific philosophers is highly charged at times, and becomes acrimonious to the point where reputable skeptical scientists cease being impartial by refusing to examine the experimental data or theories, and they prefer to bypass rational discourse by engaging in ad hominem attacks and irrational “armchair” arguments.

P-Teleportation and related phenomena are truly anomalous, and they challenge accepted modern scientific paradigm. Lightman and Gingerich (1991) wrote, “Scientists are reluctant to change paradigms for the purely psychological reasons that the familiar is often more comfortable than the unfamiliar and that inconsistencies in belief are uncomfortable.” And theories change over time when anomalies enter the picture. Anomalies are particularly helpful for they point to the inadequacies of an old model and point the way to a new one. Anomalous scientific facts are unexpected and difficult to explain within an existing conceptual framework. Kuhn (1970) describes scientific discovery as a complex process, in which an anomalous fact of nature is recognized, and then followed by a change in conceptual framework (i.e., paradigm) that makes the new fact no longer an anomaly. Kuhn stated that, “Discovery commences with the awareness of anomaly, that is, with the recognition that nature has somehow violated the pre-induced expectations that govern normal science.” This statement neatly describes exactly what transpired during the historical revolution that took place in physics between the classical mechanics/electrodynamics age in the 19th century and the quantum/atomic/nuclear/relativistic age in the 20th century. And this isn’t the only time in human history that scientific paradigms have dramatically changed. The discovery of p-Teleportation already commenced in the 20th century, so let us continue the discovery and create a new physics paradigm for the 21st century.

5.1.1 Hypothesis Based on Mathematical Geometry

The Chinese researchers reported in their teleportation experiments that high-speed photography/videotaping recorded test specimens physically “melding” or blending with the walls of sealed containers, and in a different series of experiments the test specimens would simply disappear from inside the container only to reappear at another location (after seconds to several minutes of time transpired). They also reported in the series of radio micro-transmitter experiments that there were large fluctuations in the intensity (in both amplitude and frequency) of the monitored signal to the effect that it would either completely disappear or become extremely weak (to the extent that the monitoring instruments could scarcely detect it); and they discovered that there was a definite correlation between the change in strength (i.e., radical frequency shifts were observed) of the monitored radio signal and the teleportation of the radio micro-transmitter, such that the weak or absent signal indicated that the specimen was “nonexistent” (or in an altered physical state) during teleportation. This data is important because without the aid of electronic monitoring instruments, the average person’s sensory organs and usual methods of detection are temporarily unable to perceive the test specimen’s (ambiguous) existence during the teleportation process. This data offers an important clue on what the teleportation mechanism is.

It is beyond the scope of this study to propose a complete self-consistent physics theory of consciousness/mind, which explains how the mind can activate p-Teleportation and related psychotronics phenomena. This topic has been under study in recent decades by a legion of medical science, bio- and neuro-physiology, psychology, mathematics, philosophy, and physics experts. Many different theories with varying degree of theoretical maturity and self-consistency have been proposed over the years, and most of them have not yet been experimentally tested for various reasons. However, some first-order experimental work has been done (Mitchell, 1974b; Targ and Puthoff, 1977; Wolman et al., 1986; Radin, 1997; Tart et al., 2002). Ironically, quantum mechanics theory, and the related physics of quantum entanglement and teleportation, has become the primary focus of all of the physics theories of consciousness/psychotronics that have been recently proposed (see for example, Shan, 2003). Wolman et al. (1986) and Radin (1997) provide a review and discussion on recent theories and experiments that are based on quantum physics theory (see also, Walker, 1974; Targ and Puthoff, 1977; Mitchell, 1999, and the references cited therein; Tart et al., 2002). It appears that the physics of q-Teleportation (Chapter 3) has tremendous relevance to the physics of p-Teleportation and psychotronics.

In the following I propose a parsimonious first-order hypothesis that can explain the gross features of both the Chinese p-Teleportation data and the other reported p-Teleportation phenomena. But I will refrain from including any role that might be played by quantum phenomena since the scientific community has not yet settled that particular issue. (However, it is apparent that quantum theory and quantum phenomena will likely play a key role in a formal physics theory of PK and psychotronics.)

First-Order Hypothesis:

- ☐ Fact 1: The mature discipline of mathematical geometry developed the properties of higher dimensional spaces (Reichenbach, 1957; Manning, 1977; Rucker, 1977). An example of one such property that is of relevance to the hypothesis: One can visualize a four-dimensional world by using color as the 4th dimension. We can think of a three-dimensional world, whereby objects pass through one another if their colors (i.e., four-dimensional locations) are different (Reichenbach, 1957). For example, color can be used as a 4th dimension to see how a knot in three-dimensions can be untied in a 4th spatial dimension without moving the ends of the cord. That is because a cord cannot stay knotted in four-dimensional space, because the extra degree of freedom will cause any knot to slip through itself. Two other interesting and relevant examples are that the links of a chain may be separated unbroken in the 4th dimension, and a flexible sphere may be turned inside out without tearing in the 4th dimension (Manning, 1977; Rucker, 1977).

- ❑ Proposition 1 and Fact 2: It has been proposed that our space actually possesses a slight four-dimensional hyperthickness, so that the ultimate components of our nervous system are actually higher dimensional, thus enabling the human mind/brain to imagine four-dimensional space (Hinton, 1888, 1904; Rucker, 1977). If this is the case, then the three-dimensional nets of neurons that code thoughts in our brain may form four-dimensional patterns to achieve four-dimensional thought. The "bulk" space in 3-brane theory (see Section 4.1), and experimental data from the Remote Viewing program (see Section 5.1), provide support for this concept. Can we see into the 4th dimension and have four-dimensional thoughts? Yes, we can. Proof (see, Rucker, 1977, 1984): If you look at a Necker cube for a while, it spontaneously turns into its mirror image and back again. If you watch it do this often enough, the twinkling sort of motion from one state to the other begins to seem like a continuous motion. But this motion can only be continuous if it is a rotation in four-dimensional space. The mathematician August F. Möbius discovered in 1827 that it is in fact possible to turn a three-dimensional solid object into its mirror image by an appropriate rotation through four-dimensional space (a.k.a. hyperspace rotation). Thus, it is actually possible for our minds to perform such a rotation. Therefore, we can actually produce four-dimensional phenomenon in our minds, so our consciousness is four-dimensional. Rucker (1984) shows another dramatic example of being able to see into the 4th dimension via a "Neck-A-Cube."

- ❑ Fact 3: Another property of higher dimensional geometry (Reichenbach, 1957; Rucker, 1977, 1984) is that one can move through solid three-dimensional obstacles without penetrating them by passing in the direction of the 4th (spatial) dimension. The 4th dimension is perpendicular to all of our normal three-dimensional space directions, and so our three-dimensional enclosures have no walls against this direction.

- ❑ Conclusion and Hypothesis: Therefore, the results of the Chinese p-Teleportation experiments can simply be explained as a human consciousness phenomenon that somehow acts to move or rotate test specimens through a 4th spatial dimension, so that the specimens are able to penetrate the solid walls/barriers of their containers without physically breaching them. No real dematerialization/rematerialization of the specimens takes place. The intensity fluctuations of the radio micro-transmitter specimen's electromagnetic signal, and the apparent blending of the other specimens with the walls of their containers, represent the passage of the specimens through a 4th spatial dimension. During teleportation the radio signals emitted by the micro-transmitter became weak/non-existent and fluctuated, because they were spreading out into the 4th dimension and became undetectable in our three-dimensional space. The weak signals that were ("barely") detected represent the leakage of a portion of the radio signal back into our three-dimensional space from the 4th dimension during teleportation. The observed blending of the other specimens with the walls of their containers is how the movement/rotation of the specimens through the 4th dimension was visually interpreted by the mind (along the lines of the Necker cube or Neck-A-Cube examples).

5.2 Conclusion and Recommendations

We will need a physics theory of consciousness and psychotronics, along with more experimental data, in order to test the hypothesis in Section 5.1.1 and discover the physical mechanisms that lay behind the psychotronic manipulation of matter. P-Teleportation, if verified, would represent a phenomenon that could offer potential high-payoff military, intelligence and commercial applications. This phenomenon could generate a dramatic revolution in technology, which would result from a dramatic paradigm shift in science. Anomalies are the key to all paradigm shifts!

•**Recommendations**:

- There are numerous supporters within the U.S. military establishment who comprehend the significance of remote viewing and PK phenomenon, and believe that they could have strategic implications. Bremseth (2001), a U.S. Navy SEAL, attended the Marine War College and studied the Remote Viewing program, and interviewed many of the former program participants. Bremseth then wrote his thesis on the topic, and concluded that the evidence supported continued research and applications of remote viewing. A research program improving on and expanding, or implementing novel variations of, the Chinese and Uri Geller-type experiments should be conducted in order to generate p-Teleportation phenomenon in the lab. The performances and characteristics of p-Teleportation need to be delineated in order to develop a refined hypothesis. Such a program should be designed so that an operational model for p-Teleportation can be developed and implemented as a prototype. An experimental program similar in fashion to the Remote Viewing program should be funded at $900,000 – 1,000,000 per year in parallel with a theoretical program funded at $500,000 per year for an initial five-year duration. The role of quantum physics theory and related quantum phenomena (i.e., entanglement and teleportation) in p-Teleportation and psychotronics should be explored in this program (see for example, the Biological Quantum Teleportation recommendation in Section 3.3). An experiment definition study should be conducted first to identify and propose the best experiments for this program, which should be funded at $80,000 for one year.

6.0 REFERENCES

1. Aczel, A. D. (2002), Entanglement: The Greatest Mystery in Physics, Four Walls Eight Windows Press, New York
2. Aharonov, Y., Reznik, B. and Stern, A. (1998), “Quantum limitations of superluminal propagation,” Phys. Rev. Lett., 81, 2190-2193
3. Aharonov, Y. and Albert, D. (1981), “Can we make sense of the measurement process in relativistic quantum mechanics?,” Phys. Rev. D, 24, 359-370
4. Alexander, Col. J. B. (2003), Winning The War: Advanced Weapons, Strategies, And Concepts For The Post-9/11 World, St. Martin’s Press, New York, pp. 238 – 244
5. Alexander, J. B. (1996), “Uri’s Impact on the U.S. Army,” posted on http://www.urigeller.com
6. Alexander, Lt. Col. J. B. (1980), “The New Mental Battlefield: ‘Beam Me Up, Spock,’” Military Review, vol. XL, no. 12
7. Alexander, Col. J. B., Groller, Maj. R. and Morris, J. (1990), The Warrior’s Edge, W. Morrow Co., New York
8. Ambjϕrn, J. and Wolfram, S. (1983), “Properties of the Vacuum. I. Mechanical and Thermodynamic,” Annals Phys., 147, 1-32
9. Antoniadis, I., Arkani-Hamed, N., Dimopoulos, S. and Dvali, G. (1998), “New dimensions at a millimeter to a fermi and superstrings at a TeV,” Phys. Lett. B, 436, 257-263
10. Appelquist, T., Chodos, A. and Freund, P. G. O. eds. (1987), Modern Kaluza-Klein Theories, Addison-Wesley, Menlo Park
11. Arkani-Hamed, N., Dimopoulos, S. and Dvali, G. (2002), “Large Extra Dimensions: A New Arena for Particle Physics,” Physics Today, 55, 35-40
12. Arkani-Hamed, N., Dimopoulos, S. and Dvali, G. (1998), “The hierarchy problem and new dimensions at a millimeter,” Phys. Lett. B, 429, 263-272
13. Arkani-Hamed, N., Dimopoulos, S., Kaloper, N. and Dvali, G. (2000), “Manyfold universe,” Journal of High Energy Physics (JHEP online physics papers), http://jhep.sissa.it/archive/papers/jhep122000010/jhep122000010.pdf0012
14. Aspect, A. (1983), Trois tests expérimentaux des inégalités de Bell par mesure de corrélation de polarisation de photons, Ph.D. thesis No. 2674, Université de Paris-Sud, Centre D’Orsay
15. Aspect, A., Dalibard, J. and Roger, G. (1982a), “Experimental Tests of Bell’s Inequalities Using Time-Varying Analyzers,” Phys. Rev. Letters, 49, 1804-1807
16. Aspect, A., Grangier, P. and Roger, G. (1982b), “Experimental Realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A New Violation of Bell’s Inequalities,” Phys. Rev. Letters, 49, 91
17. Aspect, A. and Grangier, P. (1985), Lettere al Nuovo Cimento, 43, 345
18. Aspelmeyer, M., et al. (2003), “Long-Distance Free-Space Distribution of Quantum Entanglement,” Science, 301, 621 - 623
19. Banghui, W. (1990), “Evidence of the Existence of Abnormal States of Matter,” Chinese J. Somatic Sci., First Issue, 36 [translated into English by the Defense Intelligence Agency]
20. Barcelo, C. and Visser, M. (2002), “Twilight for the energy conditions?,” Int. J. Mod. Phys. D, 11, 1553
21. Barnum, H., Caves, C., Fuchs, C., Jozsa, R. and Schumacher, B. (1996), “Noncommuting Mixed States Cannot Be Broadcast,” Phys. Rev. Lett., 76, 2818-2821
22. Bell, J. S. (1964), “On the Einstein Podolsky Rosen Paradox,” Physics, 1, 195
23. Bennett, C. H., et al. (1993), “Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels,” Phys. Rev. Lett., 70, 1895-1899
24. Bennett, C. H. and Wiesner, S. J. (1992), “Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states,” Phys. Rev. Lett., 69, 2881-2884

25. Bennett, G. L., Forward, R. L. and Frisbee, R. H. (1995), "Report on the NASA/JPL Workshop on Advanced Quantum/Relativity Theory Propulsion," AIAA-95-2599, 31st AIAA/ASME/ASE/ASEE Joint Propulsion Conference and Exhibit, San Diego, CA
26. Birrell, N. D. and Davies, P. C. W. (1982), Quantum fields in curved space, Cambridge University Press, Cambridge
27. Blaauboer, M., et al. (1998), "Superluminal pulse transmission through a phase conjugating mirror," Optics Communications, 148, 295-299
28. Boschi, D., et al. (1998), "Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolski-Rosen channels," Phys. Rev. Lett., 80, 1121-1125
29. Bose, S. and Home, D. (2002), "Generic Entanglement Generation, Quantum Statistics, and Complementarity," Phys. Rev. Lett., 88, 050401
30. Bose, S., Knight, P. L., Plenio, M. B. and Vedral, V. (1999), "Proposal for Teleportation of an Atomic State via Cavity Decay," Phys. Rev. Lett., 83, 5158-5161
31. Bouwmeester, D., et al. (1997), "Experimental quantum teleportation," Nature, 390, 575-579
32. Bowen, W. P., et al. (2003), "Experimental investigation of continuous variable quantum teleportation," Phys. Rev. A, 67, 032302
33. Bowen, W. P., Treps, N., Schnabel, R. and Lam, P. K. (2002), "Experimental demonstration of continuous variable polarization entanglement," Phys. Rev. Lett., 89, 253601
34. Brassard, G., Braunstein, S. and Cleve, R. (1998), "Teleportation as a quantum computation," Physica D, 120, 43-47
35. Braunstein, S. (1996), "Quantum teleportation without irreversible detection," Proc. Royal Acad., 53, 1900-1903
36. Braunstein, S., Fuchs, C., Kimble, H. and van Loock, P. (2001), "Quantum versus classical domains for teleportation with continuous variables," Phys. Rev. A, 64, 022321
37. Braunstein, S. and Kimble, J. (1998), "Teleportation of continuous quantum variables," Phys. Rev. Lett., 80, 869-872
38. Bremseth, Cmdr. L. R. (2001), Unconventional Human Intelligence Support: Transcendent and Asymmetric Warfare Implications of Remote Viewing, Thesis, Marine War College
39. Buttler, W., et al. (1998), "Practical free-space quantum key distribution over 1 km," Phys. Rev. Lett., 81, 3283-3286
40. Center for Quantum Computation Web Site: http://www.qubit.org
41. Chan, H. B., et al. (2001), "Quantum Mechanical Actuation of Microelectromechanical Systems by the Casimir Force," Science, 291, 1941-1944
42. Chiao, R. and Steinberg, A. (1998), "Quantum optical studies of tunneling and other superluminal phenomena," Physica Scripta, T76, 61-66
43. Chown, M. (1990), "Can photons travel 'faster than light'?," New Scientist, 126, 32
44. CIA In-Q-Tel Web Site information: http://www.cia.gov/cia/publications/inqtel/
45. Cole, D. C. and Puthoff, H. E. (1993), "Extracting Energy and Heat from the Vacuum," Phys. Rev. E, 48, 1562
46. Davies, P. C. W. (1980), Other Worlds, Dent, London
47. Davis, E. W. (1999a), Research Summary Report #1 to Dr. Hal Puthoff, IASA: Brief Summary of "Lorentzian Wormholes From The Gravitationally Squeezed Vacuum", NASA Research Center Online Library – Interstellar Studies (available from the author)
48. Davis, E. W. (1999b), Research Summary Report #2 to Dr. Hal Puthoff, IASA: Brief Summary of "Gravitational Vacuum Polarization. Parts I – IV", NASA Research Center Online Library – Interstellar Studies (available from the author)
49. de Felice, F. (1971), "On the gravitational field acting as an optical medium," Gen. Rel. Grav., 2, 347-357
50. de Oliveira, E. C. and Rodriguez, W. (1998), "Superluminal electromagnetic waves in free space," Annalen der Physik, 7(7-8), 654-659

51. de Sabbata, V. and Schmutzer, E. eds. (1983), Unified Field Theories of More Than 4 Dimensions, Proc. Int'l School of Cosmology and Gravitation (Erice), World Scientific, Singapore
52. Deutsch, D. (1998), The Fabric of Reality, Penguin Books
53. DeWitt, B. S. and Graham, N. eds. (1973), The Many Worlds Interpretation of Quantum Mechanics, Princeton University Press, Princeton
54. DeWitt, B. S. (1970), Physics Today, 23, 30
55. DIA Report (1978), Paraphysics R&D – Warsaw Pact, Defense Intelligence Agency, Report No. DST-1810S-202-78, DIA Task No. PT-1810-18-76, Washington DC (authors' names redacted)
56. Dicke, R. H. (1961), "Mach's principle and equivalence," in Proc. of the Int'l School of Physics "Enrico Fermi" Course XX, Evidence for Gravitational Theories, ed. C. Møller, Academic Press, New York, pp. 1-49
57. Dicke, R. H. (1957), "Gravitation without a principle of equivalence," Rev. Mod. Phys., 29, 363-376
58. Ding, Y. J. and Kaplan, A. E. (1992), "Nonlinear Magneto-Optical Effect in Vacuum: Inhomogeneity-Originated Second-Harmonic Generation in DC Magnetic Field," J. Nonl. Opt. Phys., 1, 51-72
59. Ding, Y. J. and Kaplan, A. E. (1989), "Nonlinear Magneto-Optics of Vacuum: Second-Harmonic Generation," Phys. Rev. Lett., 63, 2725-2728
60. Drummond, I. J. and Hathrell, S. J. (1980), "QED vacuum polarization in a background gravitational field and its effect on the velocity of photons," Phys. Rev. D, 22, 343-355
61. Dür, W. and Briegel, H.-J. (2003), "Entanglement purification for Quantum Computation," Phys. Rev. Lett., 90, 067901
62. Einstein, A., Podolsky, B. and Rosen, N. (1935), "Can quantum mechanical description of physical reality be considered complete?," Phys. Rev., 47, 777-780
63. Evans, J., Nandi, K. and Islam, A. (1996a), "The Optical-Mechanical Analogy in General Relativity: New Methods for the Paths of Light and of the Planets," Am. J. Phys., 64, 1401-1415
64. Evans, J., Nandi, K. and Islam, A. (1996b), "The Optical-Mechanical Analogy in General Relativity: Exact Newtonian Forms for the Equations of Motion of Particles and Photons," Gen. Rel. Grav., 28, 413-439
65. Everett, H. (1957), Rev. Mod. Phys., 29, 454
66. Forward, R. L. (2001), Personal Communication, Salt Lake City, UT
67. Forward, R. L. (1999), Personal Communication, Los Angeles, CA
68. Forward, R. L. (1998), "Apparent Method for Extraction of Propulsion Energy from the Vacuum," AIAA-98-3140, 34th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, Cleveland, OH
69. Forward, R. L. (1996), Mass Modification Experiment Definition Study, PL-TR-96-3004, Phillips Laboratory-Propulsion Directorate, Air Force Materiel Command, Edwards AFB, CA
70. Forward, R. L. (1984), "Extracting electrical energy from the vacuum by cohesion of charged foliated conductors," Phys. Rev. B, 30, 1770-1773
71. Freedman, S. J. and Clauser, J. F. (1972), "Experimental Test of Local Hidden-Variable Theories," Phys. Rev. Lett., 28, 938-941
72. Friedman, J. et al. (1990), "Cauchy problem in spacetimes with closed timelike curves," Phys. Rev. D, 42, 1915-1930
73. Furusawa, A., et al. (1998), "Unconditional quantum teleportation," Science, 282, 706-710
74. Furuya, K., et al. (1999), "Failure of a proposed superluminal scheme," Phys. Lett. A, 251, 294-296
75. Geller, U. (1975), Uri Geller: My Story, Praeger Publ., New York
76. Giroldini, W. (1991), "Eccles's Model of Mind-Brain Interaction and Psychokinesis: A Preliminary Study," J. Sci. Explor., 5, no. 2

77. Gisin, N. (1990), "Weinberg's non-linear quantum mechanics and superluminal communications," Phys. Lett. A, 143, 1-2
78. Gisin, N., Scarani, V., Tittel, W. and Zbinden, H. (2000), "Optical tests of quantum non-locality: from EPR-Bell tests towards experiments with moving observers," Ann. Phys. (Leipzig), 9, 831-841
79. Gissurarson, L. R. (1992), "The Psychokinesis Effect: Geomagnetic Influence, Age and Sex Difference," J. Sci. Explor., 6, no. 2
80. Green, M. B. (1985), "Unification of forces and particles in superstring theories," Nature, 314, 409
81. Greenberger, D. (1998), "If one could build a macroscopical Schrodinger cat state, one could communicate superluminally," Physica Scripta, T76, 57-60
82. Groller, Capt. R. (1987), "Soviet Psychotronics – a Closer Look," Military Intelligence, PB 34-87-1 (Test), pp. 43-44
83. Groller, Capt. R. (1986), "Soviet Psychotronics – a State of Mind," Military Intelligence, 12, no. 4, pp. 18-21, 58
84. Hagley, E., et al. (1997), "Generation of Einstein-Podolsky-Rosen Pairs of Atoms," Phys. Rev. Lett., 79, 1-5
85. Hald, J., Sørensen, J. L., Schori, C. and Polzik, E. S. (1999), "Spin Squeezed Atoms: A Macroscopic Entangled Ensemble Created by Light," Phys. Rev. Lett., 83, 1319-1322
86. Hartle, J. B. and Hawking, S. W. (1983), "Wave function of the Universe," Phys. Rev. D, 28, 2960-2975
87. Hasted, J. B. (1979), "Paranormal Metal Bending," in The Iceland Papers: Selected Papers on Experimental and Theoretical Research on the Physics of Consciousness, Puharich, A. ed., Research Associates Publ., Amherst, WI
88. Hasted, J. B., Bohm, D., Bastin, E. W. and O'Reagan, B. (1975), "Scientists confronting the paranormal," Nature, 254, 470-472
89. Haugan, M. P. and Will, C. M. (1977), "Principles of equivalence, Eötvös experiments, and gravitational red-shift experiments: The free fall of electromagnetic systems to post—post-Coulombian order," Phys. Rev. D, 15, 2711-2720
90. Hawking, S. W. and Ellis, G. F. R. (1973), The Large-Scale Structure of Space-Time, Cambridge Univ. Press, Cambridge, pp. 88-91 and 95-96
91. Hegerfeldt, G. (1998), "Instantaneous spreading and Einstein causality in quantum theory," Annalen der Physik, 7(7-8), 716-725
92. Heitler, W. (1954), The Quantum Theory of Radiation (3rd ed.), Oxford University Press, London, p. 113.
93. Herrmann, F. (1989), "Energy Density and Stress: A New Approach to Teaching Electromagnetism", Am. J. Phys., 57, 707-714
94. Hinton, C. H. (1904), The Fourth Dimension, Sonnenschein, London
95. Hinton, C. H. (1888), A New Era of Thought, Sonnenschein, London
96. Hochberg, D. and Kephart, T. W. (1991), "Lorentzian wormholes from the gravitationally squeezed vacuum," Phys. Lett. B, 268, 377-383
97. Hong, C. K. and Mandel, L. (1985), "Theory of parametric frequency down conversion of light," Phys. Rev. A, 31, 2409-2418
98. Houck, J. (1984a), "Surface Change During Warm-Forming," Archaeus, 2, no. 1
99. Houck, J. (1984b), "PK Party History," Psi Research, 3, no. 1
100. Houck, J. (1982), "PK Party Format," unpublished paper
101. IBM Press Release (2001), "IBM's Test-Tube Quantum Computer Makes History," http://www.research.ibm.com/resources/news/20011219_quantum.shtml
102. Jahn, R. G. and Dunne, B. J. (1987), Margins of Reality: The Role of Consciousness in the Physical World, Harcourt Brace Jovanovich, New York

103. Jammer, M. (1974), The Philosophy of Quantum Mechanics, Wiley-Interscience, New York, pp. 507-521
104. Jennewein, T., Simon, C., Weihs, G., Weinfurter, H. and Zeilinger, A. (2000), "Quantum Cryptography with Entangled Photons," Phys. Rev. Lett., 84, 4729-4732
105. Jinggen, H., Xinghai, Y. and Laijing, S. (1990), "Investigation into the 'Force' in Parapsychological Writing," Chinese J. Somatic Sci., First Issue, 32 [translated into English by the Defense Intelligence Agency]
106. Julsgaard, B., Kozhekin, A. and Polzik, E. S. (2001), "Experimental long-lived entanglement of two macroscopic objects," Nature, 413, 400-403
107. Kaku, M. (1994), Hyperspace: A Scientific Odyssey Through Parallel Universes, Time Warps, and the 10th Dimension, Anchor Books-Doubleday, New York
108. Kaku, M. (1993), Quantum Field Theory, Oxford University Press, New York
109. Kaku, M. (1988), Introduction to Superstrings, Springer-Verlag, New York
110. Kaluza, T. (1921), "Unitätsproblem der Physik," Sitz. Preuss. Akad. Wiss. Phys. Math., K1, 966
111. Kaplan, A. E. and Ding, Y. J. (2000), "Field-gradient-induced second-harmonic generation in magnetized vacuum," Phys. Rev. A, 62, 043805-(1-9)
112. Kim, Y.-H., Kulik, S. P. and Shih, Y. (2001), "Quantum Teleportation of a Polarization State with a Complete Bell State Measurement," Phys. Rev. Lett., 86, 1370-1373
113. Klein, O. (1926), "Quantentheorie und fünfdimensionale Relativitätstheorie," Zeits. Phys., 37, 895
114. Kongzhi, S., Xianggao, L. and Liangzhong, Z. (1990), "Research into Paranormal Ability to Break Through Spatial Barriers," Chinese J. Somatic Sci., First Issue, 22 [translated into English by the Defense Intelligence Agency]
115. Kuhn, T. S. (1970), The Structure of Scientific Revolutions, 2nd ed., Univ. of Chicago Press, Chicago
116. Kwiat, P. G., et al. (1995), "New high-intensity source of polarization-entangled photon pairs," Phys. Rev. Lett., 75, 4337-4341
117. Kwiat, P. G., et al. (1999), "Ultrabright source of polarization-entangled photons," Phys. Rev. A., 60, R773-R776
118. Lamas-Linares, A., Howell, J. C. and Bouwmeester, D. (2001), "Stimulated emission of polarization-entangled photons," Nature, 412, 887-890
119. Lamoreaux, S. K. (1997), "Measurement of the Casimir Force Between Conducting Plates," Phys. Rev. Letters, 78, 5-8
120. LaMothe, Capt. J. D. (1972), Controlled Offensive Behavior – USSR, Defense Intelligence Agency, Report No. ST-CS-01-169-72, DIA Task No. T72-01-14, Washington DC
121. Latorre, J. I., Pascual, P. and Tarrach, R. (1995), "Speed of light in non-trivial vacua," Nucl. Phys. B, 437, 60-82
122. Lee, H. C. ed. (1984), An Introduction to Kaluza-Klein Theories, Proc. Chalk River Workshop on Kaluza-Klein Theories, World Scientific, Singapore
123. Lee, T. D. (1988), Particle Physics and Introduction to Field Theory, Harwood Academic Press, London
124. Leggett, A. J. (1999), "Quantum Theory: Weird and Wonderful," Physics World, 12, 73-77
125. Leshan, C. (2002), "Proposal for Teleportation by Help of Vacuum Holes," in Gravitation and Cosmology: From the Hubble Radius to the Planck Scale, Proc. of a Symposium in Honour of the 80th Birthday of Jean-Pierre Vigier, Amoroso, R. L., Hunter, G., Kafatos, M. and Vigier, J.-P. eds., Kluwer Academic Publ., Boston, pp. 515-516
126. Leshan, C. (1999), "Thought Experiment to the Border of Universe," J. Theoretics, 1, no.4
127. Li, L.-X. and Gott, J. R. (1998), "Self-Consistent Vacuum for Misner Space and the Chronology Protection Conjecture," Phys. Rev. Lett., 80, 2980-2983
128. Lightman, A. P. and Gingerich, O. (1991), "When Do Anomalies Begin?," Science, 255, 690-695

129. Lightman, A. P. and Lee, D. L. (1973), "Restricted proof that the weak equivalence principle implies the Einstein equivalence principle," Phys. Rev. D, 8, 364
130. Maierle, C., Lidar, D. and Harris, R. (1998), "How to teleport superpositions of chiral amplitudes," Phys. Rev. Lett., 81, 869-872
131. Maire, L. F. and LaMothe, Capt. J. D. (1975), Soviet and Czechoslovakian Parapsychology Research, Defense Intelligence Agency, Report No. DST-1810S-387-75, DIA Task No. PT-1810-12-75, Washington DC
132. Mandel, L. and Wolf, E. (1995), Optical Coherence and Quantum Optics, Cambridge University Press
133. Manning, H. P. (1977), The Fourth Dimension Simply Explained, Peter Smith Publ., Gloucester, MA
134. Mavromatos, N. E., Mershin, A. and Nanopoulos, D. V. (2002), "QED-Cavity model of microtubules implies dissipationless energy transfer and biological quantum teleportation," http://arxiv.org/abs/quant-ph/0204021
135. McConnell, A. J. (1957), Applications of Tensor Analysis, Dover Publ., New York, pp. 163-217
136. Mead, F. B. and Nachamkin, J. (1996), "System for Converting Electromagnetic Radiation Energy to Electrical Energy," United States Patent No. 5,590,031
137. Milonni, P. W. (1994), The Quantum Vacuum: An Introduction to Quantum Electronics, Academic Press, NY
138. Mitchell, E. D. (1999), "Nature's Mind: the Quantum Hologram," National Institute for Discovery Science, Las Vegas, NV, http://www.nidsci.org/articles/naturesmind-qh.html
139. Mitchell, E. D. (1974a), "Appendix: Experiments with Uri Geller," in Psychic Exploration: A Challenge for Science, Mitchell, E. D., White, J. ed., G. P. Putnam's Sons, New York, pp. 683-686
140. Mitchell, E. D. (1974b), Psychic Exploration: A Challenge for Science, White, J. ed., G. P. Putnam's Sons, New York
141. Mittelstaedt, P. (2000), "What if there are superluminal signals?," Eur. Phys. J. B, 13, 353-355
142. Mittelstaedt, P. (1998), "Can EPR-correlations be used for the transmission of superluminal signals?," Annalen der Physik, 7(7-8), 710-715
143. Mittelstaedt, P. and Nimtz, G. eds. (1998), "Workshop on Superluminal Velocities," Annalen der Physik, 7(7-8), 591-592
144. Morris, M. S. and Thorne, K. S. (1988), "Wormholes in spacetime and their use for interstellar travel: A tool for teaching general relativity", Am. J. Phys., 56, 395-412
145. Mourou, G. A., Barty, C. P. J. and Perry, M. D. (1998), "Ultrahigh-Intensity Lasers: Physics Of The Extreme On A Tabletop," Physics Today, 51, 22-28
146. Naik, D. S., Peterson, C. G., White, A. G., Berglund, A. J. and Kwiat, P. G. (2000), "Entangled state quantum cryptography: Eavesdropping on the Ekert protocol," Phys. Rev. Lett., 84, 4733
147. Nash, C. B. (1978), Science of PSI: ESP and PK, C. C. Thomas Publ., Springfield, Ill.
148. Nielsen, M. A. (2003), "Simple Rules for a Complex Quantum World," Sci. Am., 13, 25-33
149. Nielsen, M. A. and Chuang, I. L. (2000), Quantum Computation and Quantum Information, Cambridge University Press
150. Nielsen, M., Knill, E. and Laflamme, R. (1998), "Complete quantum teleportation using nuclear magnetic resonance," Nature, 396, 52-55
151. Nimtz, G. (1998), "Superluminal signal velocities," Annalen der Physik, 7(7-8), 618-624
152. Opatrný, T., Clausen, J., Welsch, D.-G. and Kurizki, G. (2000), "Squeezed-Vacuum Assisted Quantum Teleportation," Paper No. 7thCEWQO/015, in Proc. 7th Central-European Workshop on Quantum Optics, Hungary
153. Opatrný, T. and Kurizki, G. (2001), "Matter-Wave Entanglement and Teleportation by Molecular Dissociation and Collisions," Phys. Rev. Lett., 86, 3180-3183
154. Overduin, J. M. and Wesson, P. S. (1998), "Kaluza-Klein Gravity," http://arxiv.org/abs/gr-qc/9805018

155. Pan, J.-W., et al. (1998), "Experimental entanglement swapping," Phys. Rev. Lett., 80, 3891-3894
156. Pease, R. (2001), "Brane new world," Nature, 411, 986-988
157. Peres, A. (2000), "Classical intervention in quantum systems. II. Relativistic invariance," Phys. Rev. A, 61, 022117(8)
158. Perry, M. D. (2000), "The Amazing Power of the Petawatt," Science & Technology Rev. (LLNL-DoE publication), March issue, 4-12
159. Perry, M. D. (1996), "Crossing the Petawatt Threshold," Science & Technology Rev. (LLNL-DoE publication), December issue, 4-11
160. Polchinski, J. (1995), "Dirichlet Branes and Ramond-Ramond Charges," Phys. Rev. Lett., 75, 4724-4727
161. Pratt, J. G. (1986), "Soviet Research in Parapsychology," in Handbook of Parapsychology, Wolman, B. B., Dale, L. A., Schmeidler, G. R. and Ullman, M. eds., McFarland and Co. Publ., Jefferson, NC, pp. 883-903
162. Preskill, J., Lecture Notes: http://www.theory.caltech.edu/people/preskill/ph229/
163. Puthoff, H. E. (2003), Personal Communication, Institute for Advanced Studies at Austin, Austin, TX
164. Puthoff, H. E. (2002a), "Polarizable-Vacuum (PV) Approach to General Relativity", Found. Phys., 32, 927-943
165. Puthoff, H. E. (2002b), "Polarizable-Vacuum Approach to General Relativity", in Gravitation and Cosmology: From the Hubble Radius to the Planck Scale, eds. R. L. Amoroso, G. Hunter, M. Kafatos, and J.-P. Vigier, Kluwer Academic Publ., Dordrecht, the Netherlands, pp. 431-446
166. Puthoff, H. E. (1999a), "Polarizable-vacuum (PV) representation of general relativity," http://arxiv.org/abs/gr-qc/9909037
167. Puthoff, H. E. (1999b), Personal Communication, Institute for Advanced Studies at Austin, Austin, TX
168. Puthoff, H. E. (1996), "CIA-Initiated Remote Viewing Program at Stanford Research Institute," J. Sci. Explor., 10, 63-76
169. Puthoff, H. E. (1993), "On the Feasibility of Converting Vacuum Electromagnetic Energy to Useful Form," Int'l Workshop on the Zeropoint Electromagnetic Field, Cuernavaca, Mexico
170. Puthoff, H. E. (1990), "The Energetic Vacuum: Implications for Energy Research," Spec. in Sci. & Technology, 13, 247
171. Puthoff, H. E., Little, S. R. and Ibison, M. (2002), "Engineering the Zero-Point Field and Polarizable Vacuum for Interstellar Flight," J. British Interplanetary Soc., 55, 137-144
172. Puthoff, H. E. and Targ, R. (1975), "Physics, Entropy and Psychokinesis," in Proc. Conf. Quantum Physics and Parapsychology (Geneva, Switz.), Parapsychology Foundation Publ., New York
173. Puthoff, H. E. and Targ, R. (1974), "PK experiments with Uri Geller and Ingo Swann," in Research in Parapsychology 1973, Roll, W. G., Morris, R. L. and Morris, J. D. eds., Scarecrow Press, Metuchen, New Jersey, pp. 125-128
174. Quantum Information: Special Issue (1998), Physics World, 11, no. 3
175. Radin, D. (1997), The Conscious Universe: The Scientific Truth of Psychic Phenomena, HarperEdge-HarperCollins Publ., New York
176. Raimond, J. M., Brune, M. and Haroche, S. (2001), "Colloquium: Manipulating quantum entanglement with atoms and photons in a cavity," Rev. Mod. Phys., 73, 565-582
177. Randall, L. and Sundrum, R. (1999a), "Large Mass Hierarchy from a Small Extra Dimension," Phys. Rev. Lett., 83, 3370-3373
178. Randall, L. and Sundrum, R. (1999b), "An Alternative to Compactification," Phys. Rev. Lett., 83, 4690-4693
179. Rarity, J. G. (2003), "Getting Entangled in Free Space," Science, 301, 604 - 605
180. Reichenbach, H. (1957), The Philosophy of Space and Time, Dover Publ., New York
181. Rhine, L. E. (1970), Mind over Matter: Psychokinesis, Macmillan, New York

182. Rubakov, V. A. and Shaposhnikov, M. E. (1983a), "Do we live inside a domain wall?," Phys. Lett. B, 125, 136-138
183. Rubakov, V. A. and Shaposhnikov, M. E. (1983b), "Extra space-time dimensions: Towards a solution to the cosmological constant problem," Phys. Lett. B, 125, 139-143
184. Rucker, R. (1984), The Fourth Dimension: A Guided Tour of the Higher Universes, Houghton Mifflin Co., Boston, pp. 45-49
185. Rucker, R. (1977), Geometry, Relativity and the Fourth Dimension, Dover Publ., New York
186. Sackett, C. A. (2001), Quant. Inf. Comput., 1, 57
187. Sackett, C. A., et al. (2000), "Experimental entanglement of four particles," Nature, 404, 256
188. Scarani, V., Tittel, W., Zbinden, H. and Gisin, N. (2000), "The speed of quantum information and the preferred frame: analysis of experimental data," Phys. Lett. A, 276, 1-7
189. Scharnhorst, K. (1990), "On Propagation of Light in the Vacuum Between Plates," Phys. Lett. B, 236, 354-359
190. Schein, F. and Aichelburg, P. C. (1996), "Traversable Wormholes in Geometries of Charged Shells," Phys. Rev. Letters, 77, 4130-4133
191. Schmidt, H. (1987), "The Strange Properties of Psychokinesis," J. Sci. Explor., 1, no. 2
192. Schmidt, H. (1974), "Psychokinesis," in Psychic Exploration: A Challenge for Science, Mitchell, E. D., White, J. ed., G. P. Putnam's Sons, New York, pp. 179-193
193. Schnabel, J. (1997), Remote Viewers: The Secret History of America's Psychic Spies, Dell Publ., New York
194. Schrödinger, E. (1980), Proc. Am. Philos. Soc., 124, 323
195. Schrödinger, E. (1935a), Die Naturwissenschaften, 48, 807
196. Schrödinger, E. (1935b), Die Naturwissenschaften, 49, 823
197. Schrödinger, E. (1935c), Die Naturwissenschaften, 49, 844
198. Seife, C. (2000), "'Spooky Action' Passes a Relativistic Test," Science, 287, 1909-1910
199. Shan, G. (2003), "A Primary Quantum Model of Telepathy," http://cogprints.ecs.soton.ac.uk/archive/00003065/
200. Shigemi, S., Yasuo, O. and Akihira, T. (1978), "Some Observations with Scanning Electron Microscope (SEM) of the Fracture Surface of Metals Fractured by Psychokinesis," Japan PS Soc. J., 2, no. 2
201. Shor, P. W. (1997), "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer," SIAM J. Sci. Statist. Comput., 26, 1484
202. Shor, P. W. (1994), "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer", in Proc. 35th Annual Symposium on Foundations of Computer Science, IEEE Computer Society Press, p. 124
203. Shoup, R. (2002), "Anomalies and Constraints: Can Clairvoyance, Precognition, and Psychokinesis Be Accommodated within Known Physics?," J. Sci. Explor., 16, no. 1
204. Shuhuang, L., et al. (1981), "Some Experiments on the Transfer of Objects Performed by Unusual Abilities of the Human Body," Nature Journal (Peoples Republic of China), 4, no. 9, 652 [Defense Intelligence Agency Requirements and Validation Branch, DIA Translation LN731-83, Intelligence Information Report No. 6010511683 (1983)]
205. Siegfried, T. (2000), The Bit and the Pendulum, John Wiley & Sons
206. Sørensen, J. L. (1998), Nonclassical light for atomic physics and quantum teleportation, Ph.D. thesis, Univ. of Aarhus
207. Srikanth, R. (July 1999), "Noncausal superluminal nonlocal signaling," http://arxiv.org/abs/quant-ph/9904075
208. Stanford, R. (1974), "Interview," Psychic, 7
209. Stenholm, S. and Bardroff, P. (1998), "Teleportation of N-dimensional states," Phys. Rev. A, 58, 4373-4376

210. Swann, I. (1974), "Scientological Techniques: A Modern Paradigm for the Exploration of Consciousness and Psychic Integration," in Proc. First Int'l Conf. on Psychotronic Research, United States Joint Publications Research Service, Document No. JPRS L/5022-1, Virginia
211. Targ, R. (1996), "Remote Viewing at Stanford Research Institute in the 1970s: A Memoir," J. Sci. Explor., 10, 77-88
212. Targ, R. and Puthoff, H. E. (1977), Mind-Reach: Scientists Look at Psychic Ability, Jonathan Cape Ltd.-Anchor Press, London
213. Tart, C. T., Puthoff, H. E. and Targ, R. eds. (2002), Mind at Large: Institute of Electrical and Electronics Engineers Symposia on the Nature of Extrasensory Perception, Hampton Roads Publ. Co., Charlottesville, VA
214. Terhal, B. M., Wolf, M. M. and Doherty, A. C. (2003), "Quantum Entanglement: A Modern Perspective," Physics Today, 56, 46-52
215. Thorne, K. S. (1993), "Closed Timelike Curves," GRP-340, CalTech, Pasadena, CA
216. Tittel, W. and Weihs, G. (2001), Quantum Inf. Comput., 1, 3
217. Tittel, W., Brendel, J., Zbinden, H. and Gisin, N. (2000), "Quantum Cryptography Using Entangled Photons in Energy-Time Bell States," Phys. Rev. Lett., 84, 4737-4740
218. Tittel, W., Brendel, J., Gisin, B., Herzog, T., Zbinden, H. and Gisin, N. (1998a), "Experimental demonstration of quantum correlations over more than 10 km," Phys. Rev. A, 57, 3229-3232
219. Tittel, W., Brendel, J., Zbinden, H. and Gisin, N. (1998b), "Violation of Bell Inequalities by Photons More Than 10 km Apart," Phys. Rev. Lett., 81, 3563-3566
220. Vaidman, L. (1994), "Teleportation of quantum states," Phys. Rev. A, 49, 1473-1476
221. Vaidman, L. and Yoran, N. (1999), "Methods for reliable teleportation," Phys. Rev. A, 59, 116-125
222. Vallee, J. (1997), Personal Communication, Science Advisory Board of the National Institute for Discovery Science, Las Vegas, NV
223. Vallee, J. (1990), Confrontations: A Scientist's Search for Alien Contact, Ballantine Books, New York
224. Vallee, J. (1988), Dimensions: A Casebook of Alien Contact, Ballantine Books, New York
225. van Enk, S. (March 1998), "No-cloning and superluminal signaling," http://arxiv.org/abs/quant-ph/9803030
226. Visser, M., Kar, S. and Dadhich, N. (2003), "Traversable Wormholes with Arbitrarily Small Energy Condition Violations," Phys. Rev. Lett., 90, 201102
227. Visser, M. (1997), Personal Communication, Washington University, St. Louis, MO
228. Visser, M. (1995), Lorentzian Wormholes: From Einstein to Hawking, AIP Press, New York
229. Visser, M. (1990), "Wormholes, baby universes, and causality", Phys. Rev. D, 41, 1116-1124
230. Visser, M. (1989), "Traversable wormholes: Some simple examples", Phys. Rev. D, 39, 3182-3184
231. Volkov, A. M., Izmest'ev, A. A. and Skrotskii, G. V. (1971), "The propagation of electromagnetic waves in a Riemannian space," Sov. Phys. JETP, 32, 686-689
232. Walker, E. H. (1974), "Consciousness and Quantum Theory," in Psychic Exploration: A Challenge for Science, Mitchell, E. D., White, J. ed., G. P. Putnam's Sons, New York, pp. 544-568
233. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. and Zeilinger, A. (1998), "Violation of Bell's Inequality under Strict Einstein Locality Conditions," Phys. Rev. Lett., 81, 5039-5043
234. Weinberg, S. (1992), Dreams of a Final Theory, Vintage Books, pp. 88-89
235. Weinberg, S. (1989), "Testing Quantum Mechanics," Ann. Phys., 194, 336-386
236. Weiss, P. (2000), "Hunting for Higher Dimensions: Experimenters scurry to test new theories suggesting that extra dimensions are detectable," Science News, 157, 122-124
237. Westmoreland, M. and Schumacher, B. (March 1998), "Quantum entanglement and the non existence of superluminal signals," http://arxiv.org/abs/quant-ph/9801014
238. Wheeler, J. A. (1962), Monist, 47, 40

239. Wheeler, J. A. (1957), Rev. Mod. Phys., 29, 463
240. Wheeler, J. A. and Zurek, W. H. eds. (1983), Quantum Theory and Measurement, Princeton University Press
241. Will, C. M. (1993), Theory and Experiment in Gravitational Physics (rev. ed.), Cambridge University Press, Cambridge, Section 2.6
242. Will, C. M. (1989), "The confrontation between gravitation theory and experiment," in General Relativity: An Einstein Centenary Survey, eds. S. W. Hawking and W. Israel, Cambridge University Press, Cambridge, Chapter 2
243. Will, C. M. (1974), "Gravitational red-shift measurements as tests of nonmetric theories of gravity," Phys. Rev. D, 10, 2330-2337
244. Wilson, H. A. (1921), "An electromagnetic theory of gravitation," Phys. Rev., 17, 54-59
245. Wineland, D. J., et al. (2002), "Quantum information processing with trapped ions," http://arxiv.org/abs/quant-ph/0212079
246. Wolf, F. A. (1988), Parallel Universes: The Search for Other Worlds, Simon and Schuster, New York
247. Wolman, B. B., Dale, L. A., Schmeidler, G. R. and Ullman, M. eds. (1986), Handbook of Parapsychology, McFarland and Co. Publ., Jefferson, NC
248. Wootters, W. K. and Zurek, W. H. (1982), "A single quantum cannot be cloned," Nature, 299, 802-803
249. Zbinden, H., Brendel, J., Gisin, N. and Tittel, W. (2000a), "Experimental Test of Non-Local Quantum Correlation in Relativistic Configurations," http://arxiv.org/abs/quant-ph/0007009
250. Zbinden, H., Brendel, J., Tittel, W. and Gisin, N. (2000b), "Experimental Test of Relativistic Quantum State Collapse with Moving Reference Frames," http://arxiv.org/abs/quant-ph/0002031
251. Zeilinger, A. (2003), "Quantum Teleportation," Sci. Am, 13, 34-43
252. Zhang, T. C., et al. (2002), "Quantum teleportation of light beams," http://arxiv.org/abs/quant-ph/0207076
253. Zubairy, S. (1998), "Quantum teleportation of a field state," Phys. Rev. A, 58, 4368-4372

APPENDIX A – A Few Words About Negative Energy

A.1 A General Relativistic Definition of Negative or Exotic Energy

We saw in equations (2.10a-c) that the surface energy and stress-tension densities of the material required to create and thread a traversable wormhole must be "negative." For surface stress-energy, and volume stress-energy in general, this is "negative" in the sense that the material we must deploy to generate and thread the traversable wormhole must have an energy density (ρc^2, ρ = mass density) that is less than the stress-energy density (τ), or we can write this condition as: *mass-energy* $\rho c^2 \leq$ *stress-energy* τ. On the basis of this condition, we call this material property "exotic." Therefore, the term "negative" is just a misnomer in this context. The condition for ordinary, non-exotic forms of matter that we are all familiar with is *mass-energy* $\rho c^2 >$ *stress-energy* τ. This condition represents one version of what is variously called the weak (WEC), null (NEC), average (AEC), dominant (DEC), strong (SEC) or "standard" energy conditions (that are mere hypotheses!), which allegedly forbid negative mass-energy density and gravitational repulsion (antigravity) between material objects to occur in nature. Hawking and Ellis (1973) formulated these energy conditions in order to establish a series of mathematical proofs in their study of the application of general relativity theory to cosmology and black hole physics.

However, there are general theorems of differential geometry that guarantee that there must be NEC violations (meaning exotic matter-energy is present) at a wormhole throat (Visser, 1997). In view of this, it is known that static radial electric or magnetic fields are borderline exotic when threading a wormhole, if their tension were infinitesimally larger, for a given energy density (Herrmann, 1989; Hawking and Ellis, 1973). Other exotic (energy condition violating) matter-energy fields are known to be squeezed quantum states of the electromagnetic field and other squeezed quantum fields (see Section A.2 for the discussion on squeezed quantum states), gravitationally squeezed vacuum electromagnetic zero-point energy (see Section 2.3 for the discussion on Gravitationally Squeezed Vacuum Energy), Casimir (electromagnetic zero-point) energy and other quantum fields/states/effects. These examples represent forms of matter-energy that possess negative energy density. Since the vacuum is defined to have vanishing energy density, anything possessing less energy density than the vacuum must have a negative energy density. With respect to creating wormholes, these have the unfortunate reputation of alarming physicists. This is unfounded since all the energy condition hypotheses have been experimentally tested in the laboratory and experimentally shown to be false - 25 years before their formulation (Visser, 1990 and references cited therein). Further investigation into this technical issue showed that violations of the energy conditions are widespread for all forms of both classical and quantum matter-energy such as planets, stars, black holes, neutron stars, people, space dust clouds, etc. (Davis, 1999b; Barcelo and Visser, 2002). In addition, Visser (1995) showed that all (generic) spacetime geometries violate all the energy conditions. Violating the energy conditions commits no offense against nature.

A.2 Squeezed Quantum States and Negative Energy

In quantum mechanics the energy (E) and frequency (ν) of a quantum oscillator system, such as electromagnetic radiation (or light), are interchangeable via the Planck relation $E = h\nu$ ($h = 2\pi\hbar$). And from the Heisenberg quantum uncertainty principle, we know that the conjugate variable to the frequency is the oscillator phase (φ), such that $\Delta\nu\Delta\varphi \geq \hbar$ is obeyed. Phase is difficult to measure and is ignored in characterizing complex quantum systems.

Recent theoretical and experimental work has shown that in many quantum systems the limits to measurement precision imposed by the quantum vacuum zero-point fluctuations (ZPF) can be breached by decreasing the frequency noise at the expense of increasing the phase noise (thus maintaining $\Delta\nu\Delta\varphi \geq \hbar$), while at the same time the variations in frequency, and therefore the energy, are reduced below the ZPF such that the energy becomes "negative." "Squeezing" is thus the control of quantum fluctuations and corresponding uncertainties, whereby one can squeeze the variance of one (physically important) observable quantity provided the variance in the (physically unimportant) conjugate variable is stretched/increased. The squeezed quantity possesses an unusually low variance, meaning less variance than would be expected on the basis of the equipartition theorem. We can exploit quantum squeezing to extract energy from one place in the ordinary vacuum at the expense of piling up excess energy elsewhere (Morris and Thorne, 1988).

Appendix B – *TH*εμ Methodology

In the formalism of the *TH*εμ methodology, the functions T and H are introduced by requiring that the Lagrangian for the motion of particles (with charge e_a and mass m_{0a} for the ath particle), under the joint action of gravity and the electromagnetic field A_α ($\alpha \equiv$ spacetime vector components), be expressed in the canonical form

$$L = \sum_a \int \left(-m_{0a}\sqrt{T - H{v_a}^2} + e_a A_\alpha {v_a}^\alpha \right) dt + \left(8\pi\right)^{-1} \int \left(\varepsilon \mathbf{E}^2 + \mu^{-1}\mathbf{B}^2 \right) d^3x\, dt \qquad \text{(B.1);}$$

where the arbitrary functions T, H, ε, and μ are functions of the metric (a.k.a. gravitation field), ${v_a}^\alpha$ is the ath particle four-vector velocity, and A_α is the electromagnetic field four-vector potential, ***E*** and ***B*** are the electric and magnetic field strengths, and (B.1) is in geometrodynamic natural units ($\hbar = c_0 = G = \varepsilon_0 = \mu_0 = 1$). The Lagrangian characterizes the motion of charged particles in an external gravitational field by the two functions T and H, and characterizes the response of the electromagnetic fields to the external gravitational field by the two functions ε and μ. For all standard (metric) theories of gravity, the four functions are related by

$$\varepsilon = \mu = \sqrt{\frac{H}{T}} \qquad \text{(B.2);}$$

and ***every*** metric theory of gravity satisfies this relation, such that the Einstein Equivalence Principle is satisfied.

AFRL-PR-ED-TR-2003-0034
Primary Distribution of this Report:

AFRL/PRSP (15 CD)
Dr. Frank Mead
10 E. Saturn Blvd
Edwards AFB CA 93524-7680

AFRL/PRSA (1 CD)
Dr. Jean-Luc Cambier
10 E. Saturn Blvd.
Edwards AFB CA 93524-7680

AFRL/PR (1 CD)
Dr. Alan Garscadden
1950 Fifth Street
Building 18
Wright-Patterson AFB, OH 45433-7251

AFRL/PR Technical Library (2 CD + 1 HC)
6 Draco Drive
Edwards AFB CA 93524-7130

Chemical Propulsion Information Agency (1 CD)
Attn: Tech Lib (Dottie Becker)
10630 Little Patuxent Parkway, Suite 202
Columbia MD 21044-3200

Defense Technical Information Center
(1 Electronic Submission via STINT)
Attn: DTIC-ACQS (Pat Mawby)
8725 John J. Kingman Road, Suite 94
Ft. Belvoir VA 22060-6218

Dr. Eric Davis (10 CD)
4849 San Rafael Ave.
Las Vegas, NV 89120

Dr. Dana Andrews (1 CD)
P.O. Box 36
Leavenworth, WA 98826

Dr. Jim Degnan (1 CD)
AFRL/DEHP
Kirtland AFB, NM 87117

Dr. Greg Benford (1 CD)
Physics Department
University of California
Irvine, CA 92717

Dr. Jim Benford (1 CD)
Microwave Sciences, Inc.
1041 Los Arabis Ln.
Lafayette, CA 94549

Dr. Gary L. Bennett (1 CD)
7517 West Devonwood Dr.
Boise, ID 83703

Dr. Mitat Birkan (1 CD)
AFOSR/NA
801 N. Randolf St.
Arlington, VA 22203

Dr. Jon Campbell (1 CD)
P.O. Box 295
Harvest, AL 35749

Mr. Michael Libeau
NSWCDD
17320 Dahlgren Rd.
Attn: Code G23 Libeau
Dalgren VA 22448

Dr. Phil Carpenter (1 CD)
US Dept. of Energy
Oak Ridge National Laboratory
P.O. Box 2008, MS: 6269
Oak Ridge, TN 37831

Dr. Brice N. Cassenti (1 CD)
Pratt & Whitney Aircraft
400 Main Street – MS: 163-07
East Hartford, CT 06108

Dr. Chan K. Choi (1 CD)
Purdue University
School of Nuclear Engineering
West Lafayette, IN 47907

Dr. Terry Kammash (1 CD)
University of Michigan
Nuclear Engineering Dept.
Ann Arbor, MI 48109

Dr. Ingrid Wysong (1 CD)
EOARD
223-321-Old Merylebone Rd.
London NW1 5th
United Kingdom

Dr. Robert Frisbee (1 CD)
JPL, MS 125-109
4800 Oak Grove Dr.
Pasadena, CA 91109

Dave Froning (1 CD)
Flight Unlimited
5450 Country Club
Flagstaff, AZ 86004

Geroge D. Hathaway (1 CD)
Hathaway Consulting Services
39 Kendal Ave.
Toronto, Canada, Ontario
Canada M5R 1L5

Clark W. Hawk, Director (1 CD)
Propulsion Research Center
University of Alabama in Huntsville
5000 Technology Drive, TH S-266
Huntsville, AL 35899

Alan C. Holt (1 CD)
NASA/Johnson Space Center
Code OD
Houston, TX 77058

Dr. Steven Howe (1 CD)
19 Karen Lane
Los Alamos, NM 87544

Mike Kaiserman (1 CD)
Raytheon Missile Systems Company
Bldg 805, M/S C3
Tucson, AZ 85734

Dr. George Miley (1 CD)
University of Illinois, Dept. of Nuclear Engr.
214 Nuclear Engineering Laboratory
103 South Goodwin Ave.
Urbana, IL 61801

Marc Millis (1 CD)
NASA Glenn Research Center
M.S. SPTD-2
21000 Brookpark Road, MS: 86-2
Cleveland, OH 44135

Dr. Jordon Kare (1 CD)
222 Canyon Lakes Pl.
San Ramon, CA 94583

Ron J. Kita (1 CD)
87 Shady Springs DR.
Doyelstown, PA 18901

Dr. Gerald L. Kulcinski (1 CD)
Nuclear Engineering Dept.
University of Wisconsin
1500 Johnson Dr.
Madison, WI 53706

Dr. Geoffrey A. Landis (1 CD)
Sverdrup Technology
21000 Brookpark Rd., MS 302-1
Cleveland, OH 44135

Dr. Michael LaPointe (1 CD)
NASA Lewis Research Center, MS: SPTD-1
21000 Brookpark Rd.
Cleveland, OH 44135

Dr. Sheldon Meth (1 CD)
DARPA
Tactical Technology Office
3701 N. Fairfax Dr.
Arlington, VA 22203

Dr. Michael M. Micci (1 CD)
Prof. of Aerospace Engineering
233 E. Hammond Bldg.
University Park, PA 16802

Dr. Hal Puthoff (1 CD)
Institute for Advanced Studies
4030 Braker Lane, West
Suite 300
Austin, TX 78759

Dr. Eric E. Rice (1 CD)
Orbital Technologies Corp.
402 Gammon Place, Suite 10
Madison, WI 53719

Dr. Aurthur Morrish (1 CD)
DARPA/ATO
3701 N. Fairfax Dr.
Arlington, VA 22203

Dr. Paul Murad (1 CD)
Sr. Analyst, Director for Intel Production
Missile & Space Intel Center
Defense Intelligence Agency
Washington, DC 20340-6054

Dr. Brian Palaszewski (1 CD)
NASA Glenn Research Center
21000 Brookpartk Road, MS: 5-10
Cleveland, OH 44135

Dr. Alan Pike (1 CD)
DSAS
1988 Crescent Park Drive
Reston, VA 20190

Dr. Dennis Pelaccio (1 CD)
SAIC
8100 Shaffer Parkway, Suite 100
Littleton, CO 80127

Ben Plenge (1 CD)
101 W. Eglin Blvd
Suite 342
Eglin AFB, FL 32542-6810

Dr. James Powell (1 CD)
Plus Ultra Technologies, Inc.
25 East Loop Rd.
Stony Brook, NY 11970-3350

Mr. Charles A. Yost (1 CD)
Electric Spacecraft Journal
73 Sunlight Drive
Leicester, NC 28748

Dr. George Schmidt (1 CD)
NASA HQ
300 E. Street SW
Washington, DC 20546

Steve Squires (1 CD)
Directorate of Applied Technology
Test and Simulation
STEWS-DATTS-OO
WSMR, NM 88002

Robert Talley (1 CD)
Topaz 2000, Inc
3380 Sheridan Dr.
Suite 172
Amherst, NY 14226

Dr. Kenneth D. Ware (1 CD)
Defense Nuclear Agency
Simulation Technology
6801 Telegraph Road
Alexandria, VA 22310

Dr. Feiedwardt Winterberg (1 CD)
University of Nevada
Desert Research Institute
Reno, NV 89507

Dr. Young Bae (1 CD)
1101 Bryan Ave.
Suite C
Tustin, CA 92780

Dr. Thomas M York (1 CD)
1215 Inverary Pl.
State College, PA 16801

Dr. Robert J. Barker (1CD)
AFOSR/NE
801 N. Randolf St.
Arlington, VA 22203

Time Travel: Time Dilation

Daniel Buzzo
University of the West of England
Department of Computer Science and Creative Technology,
Coldharbour Lane,
Bristol, UK
daniel.buzzo@uwe.ac.uk

In this paper I discuss the theory and actuality of the effects of Time Dilation, as predicted by Einstein's Special Relativity. Alongside this the research investigates the visible, experiential effects of most personal form of Time Travel from a personal, phenomenological perspective, illustrated through digital media.

The body of work, this paper and an accompanying book of modified digital images, recalls the Hafele-Keating experiment of October 1971 and parallels the Quantified Self experiments of diary photography and personal analytics in a holistic, interconnected way.

The work tracks 12 months of air travel and illustrates with practical data the physics and mathematics behind the theory and actuality of Time Dilation. The research combines the interdisciplinary worlds of Computer Arts with Data and Scientific Visualisation to create a tangible collision between the visualisation of the personal and illustration of theory

The observations are documented in ethnographic, photo essay illustrations as an investigation into the mathematics and physics of personal and interpersonal time. Using situated imaging presented as a practice-based body of work to communicate the abstract world of very small time periods and the direct relevance of Einstein's work on the personal and perceptual world we inhabit. Contrasting Newtonian mechanics and Einsteinian space-time the work seeks to illustrate the personal nature of how the reality of time travel influences every aspect of the interpersonal

Time dilation. Time travel. General relativity. Special relativity. Visual thinking. Photography. Visualisation. Practice-based research. Art.

1. INTRODUCTION

In 1971 two American researchers, J.C. Hafele and R. E. Keating, conducted an experiment.

They placed two caesium atomic clocks on passenger airplanes and flew them in opposite directions around the globe to test a hypothesis first posited in Einstein's theory of special relativity.

Their aim was to measure the effects of relativity on these highly accurate devices - the predictions indicated that the eastward flying clock would lose 40ns and the westward gain 275ns. A nanosecond is equal to one billionth of a second (10^{-9} or $^{1}/_{1,000,000,000}$ s). To compare, one nanosecond is to one second as one second is to 31.7 years.

> During October, 1971, four cesium atomic beam clocks were flown on regularly scheduled commercial jet flights around the world twice, once eastward and once westward, to test Einstein's theory of relativity with macroscopic clocks. From the actual flight paths of each trip, the theory predicted that the flying clocks, compared with reference clocks at the U.S. Naval Observatory, should have lost 40+/-23 nanoseconds during the eastward trip and should have gained 275+/-21 nanoseconds during the westward trip ... Relative to the atomic time scale of the U.S. Naval Observatory, the flying clocks lost 59+/-10 nanoseconds during the eastward trip and gained 273+/-7 nanosecond during the westward trip, where the errors are the corresponding standard deviations. These results provide an unambiguous empirical resolution of the famous clock 'paradox' with macroscopic clocks. (Hafele & Keating 1972)

The results showed that compared to stationary clocks in the laboratory the eastward clock lost 59ns and the westward gained 273ns. Whilst there has been discussion of the accuracy and therefore reliability of the measurements the experiment is commonly held up as one of the first to give practical evidence of the physical effects of time dilation.

2. SPECIAL RELATIVITY AND TIME DILATION

In special relativity Einstein States that the duration of time will pass at different rates when comparing an object at rest versus an object in motion. The greater the velocity of the object in motion, the closer to the speed of light, c, the greater the effect of what is known as 'time dilation'

The effect of this phenomenon is significant at speeds close to the constant of light speed but becomes exponentially less as the object in question approaches rest.

$$t' = \frac{t}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Figure 1: *The mathematical relationship of time, velocity and the speed of light describing Time Dilation*

Unless our velocity is a substantial fraction of the constant speed of light the dilation effect (γ) is approximately 1. This, combined with our extremely limited lifespan is why we don't immediately notice time dilation at ordinary speeds. The fastest speed most humans will achieve, an airplane travelling near the speed of sound, gives a time dilation where $\gamma = 1.0000000000005$.

In other words, 1 second for a passenger on the plane would pass as 1.0000000000005 seconds for a stationary observer.

Time effectively passes more slowly for the traveller relative to the observer.

3. THE REALITY OF THE EFFECT

Whilst there is discussion of the validity of the results of the 1971 Hafele & Keating experiment, particularly around the 'corrections' made to their data at publication, their initial observations are clearly aligned with the predicted results to be expected from the mathematics for time dilation.

As Reinhardt *et al* (2007) note:

> Time dilation is one of the most fascinating aspects of special relativity as it abolishes the notion of absolute time

In their 2007 paper they investigate the effects of time dilation with even more accurate atomic clocks, adding more experimental evidence consistent with special relativity. There are also numerous other experiments, such as those into muon decay, that provide additional practical evidence that the effect described is real and measurable.

Figure 2: *Over Dorset: Duration 55 minutes*

4. IMPLICATIONS FOR EVERYDAY LIFE

The practical implications for everyday life at first appear intangible when one considers the seemingly miniscule time periods being discussed. Even taking some of the most extreme examples one can imagine being feasible, such as that of a cosmonaut living in a fast low earth orbit, circumnavigating the earth every 90 minutes gives γ of approximately 1.0000000003.

If one considers our astronaut in orbit for two years, the Time Dilation due to special relativity would give an increase in lifespan of approximately 20 milliseconds. A millisecond is a thousandth (10^{-3} or $^{1}/_{1,000}$) of a second.

This gap of 20 milliseconds, or two hundredths of a second, are suddenly well within human scale perception and understanding. This is a common measured gap in motor racing and fast sports like downhill skiing. Even the record for the 100m sprint is now is 9.58 seconds, run by Usain Bolt at the World Athletics championships in Berlin in 2009.

Scales in milliseconds are spaces and durations that even humans with our modest lifespans can perceive and understand.

As Ed Lu, Science Officer on the International Space Station writes, from orbit, in his blog post on experiments in 'Relativity':

> After our 6 months in space, we will have actually aged slightly less than everyone else on the ground because of an effect called time dilation. It isn't by much (about 0.007 seconds), but it is one side benefit of flying in space!

Though the effects of time dilation at human scale speeds are often imperceptibly small they are nonetheless real. Seeing the implications of this reflected in everyday life reveals the truly personal nature of the time travel that we undertake in everything we do.

The realisation that time dilation is a tangible and observable phenomena that affects ourselves and everything around us was given heightened poignancy by spending large amounts of time flying at altitude and high speed. Observing the curvature of our planet and seeing the dark sky at the edge of space overhead, watching the mountains and deserts as they slid effortlessly under the wings of successive aeroplanes. Being able to observe the size of the individual against the backdrop of a macro-cosmic scale all the while reflecting on the fact that to a subatomic particle travelling near the speed of light, the universe is about four weeks old and it can be crossed in a matter of months.

> Nature can produce even larger particle energies. Some particles striking the Earth's upper atmosphere have energies that exceed 2*1020 eV. If such particles are protons (with mass of about 1 GeV), their speeds would be 0.999 999 999 999 999 999 999 995 c. For them, γ is 1011. Now the age of the universe is about 13 billion years for us, but for such particles, the age of the universe would be about (13 billion years/1011), ie about a month. Such a particle could cross the visible universe in a matter of months (their time).

This twist of perspective, the deliberate 'making strange' that Shlovsky (1917) describes occurs when translating the abstract of the macrocosmic to the intimacy of the personal. The act of comparing and contrasting macro-scale physics and mathematics with recordings of human scale observations evidence reveals useful and potentially significant insights into the nature of the world we actually inhabit.

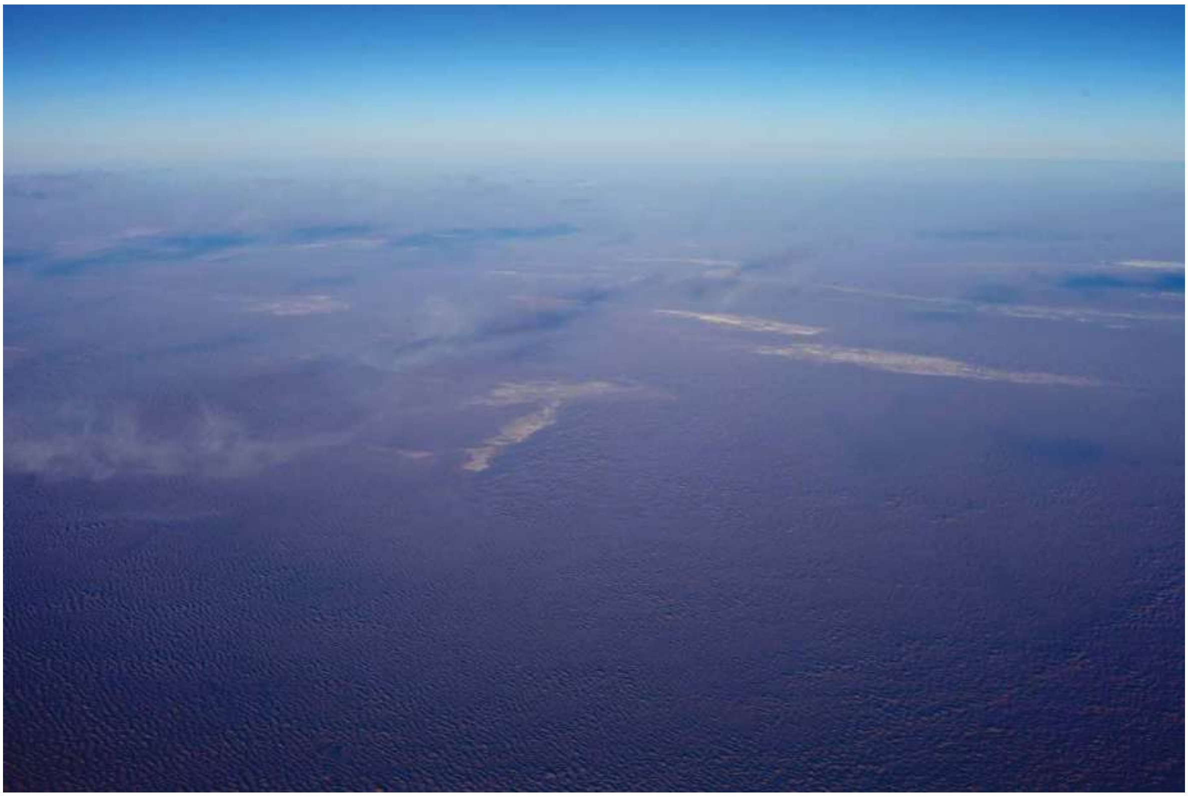

Figure 3: *Sunset over Northern China. Duration: 10 hours 50 minutes*

5. THE PICTORIAL AS PRACTICE-BASED RESEARCH

As Blevis (2011) so eloquently puts it, as an introduction to the extensive list of uses of digital imagery in design thinking.

> **Visual thinking** is the use of imagery and other visual forms to make sense of the world and to create meaningful content. Digital imagery is a special form of visual thinking, one that is particularly salient for HCI and interaction design

Alongside the descriptions of digital imagery as a *'form of information'* and a '*shared and externalised memory and cognition'* (Blevis 2012) I would argue that practice-based research, when used as an investigative tool, can assist in revealing and documenting subtle but important non-verbal evidence and also provide compelling and persuasive arguments within presentation of research findings and conclusions. The research approach taken in this area is explicitly practice-based, rooted within contemporary lens based digital media and video art. The investigative material presented as the counterpart to this research paper was collated from hundreds of minutes of digital video and several hundreds of digital still images taken while airborne on commercial passenger aircraft during January - December 2013. Taken over the course of 44 flights, totalling over 80 hours airborne the visual material is published as a 42 page visual essay, also containing text elements and descriptions taken from this academic paper.

***Figure 4**: Sunset over London. Duration: 55 minutes*

6. REALISATION OF EFFECT ON THE PERSONAL

Having recorded the number of flights, durations and distances the calculations based upon the mathematics of special relativity, as discussed earlier show that.

Given 44 flights totalling 85 hours and 20 minutes or approximately 5120 minutes flying time. gives 307,200 seconds of travel per year.

If $\gamma = 1.0000000000005$ when travelling at or near the speed of sound, gives a dilation of 0.0000001536 seconds.

To slip one second from base reference, such as the clock sat on my kitchen table I need to travel at my present annual rate for 6,510,416 Years.

Or to put it a different way, travel for a little over 260,000 years to slip one frame out of sync on a standard rate video camera.

7. CONSIDERING PHOTOGRAPHY IN EVIDENCE

When collating the increasing amounts of video and photographic material recorded whilst airborne, travelling at or near the speed of sound, the overwhelming sensation of velocity, of speed, of travel is ever-present. Viewed at a macro scale the visual material illustrates this distancing, sliding nature of objects moving in a larger framework. It supports the realisation that the terra firma so often taken as a baseline for observation is in itself only a small part of a larger perspective. The perspective that the proprioceptive nature of our observations of time, informed by the physicality of the form we have evolved and the life span perspective we consider as a norm and useful context are challenged by the linking of these measurable effects. The effects of the micro-scale variations in time seen in contrast to the macro-scale movement around the curve of the earth.

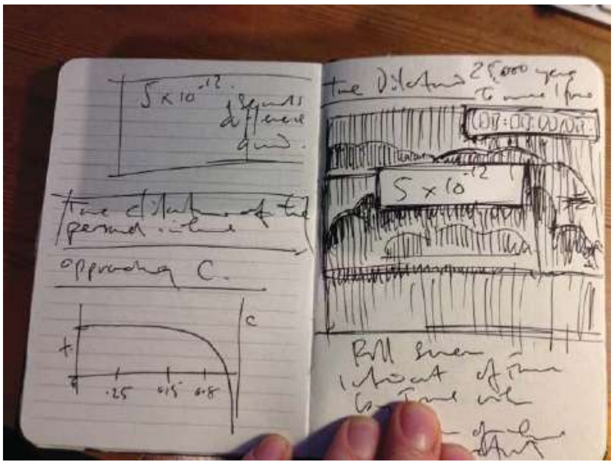
5 x 10^-12
time dilation of the period ... approaching C.
c
t
.25
.5
.8
5 x 10^-12

***Figure 5:** An airplane travelling near the speed of sound gives a time dilation where γ = 1.0000000000005*

The visual material, when presented as an adjunct to the evidence of the universal but intimately personal nature of the effects of time dilation help communicate the simple fact that we are all time travelling. Pursuing our own clearly delineated paths and durations relative to each other. Rather than there being one universal time that exists everywhere, for everyone, all at the same moment the personal implication that there is real evidence to prove that this is an illusion is both moving and compelling.

Sitting in an airline seat riding at the edge of space, looking down on the oceans, seas, cities, forests and deserts, seeing terra firma for the illusion that it is. Not a constant underlying grid of space and time - the universal yard stick and the universal clock, such as Newtonian physics promised as it replaced the Aristotelian events-make-time view of the world that had reigned before in popular consciousness.

Realising that the H&K caesium clock from 1971 could be sat beside me, ticking away in its own little timetrack, ticking to a completely separate rhythm than its counterpart clock sat on my kitchen table at home. Realising that the slipping and sliding of our lives against each other is our own intimate version of time travel, we all move forward as future becomes present becomes past but the durations we experience, and that actually occur, and when ***now*** happens is subtly different for everyone. The time we each experience actually does expand and contract as we move in space.

Figure 6: *Frost over Mongolia. Duration: 10 hours 55 minutes*

The illusion of the Newtonian ideal of a universal timescale, the idea of a 'universal now' is disappearing and being replaced by the ideas of general relativity. Of a universe full of individual 'nows' all unique and all moving independently. This change is akin to seeing the universe of unique, individual 'nows' at a human, personal scale and seeing a life travelling, sliding, back and forth in time against the 'nows' of those around one.

The important element this body of work seeks to describe is not physics or even a discussion of physics but an explanation of physics and the revelatory moment of seeing it in 'real' experienced life. This work seeks to relate that revelatory moment and articulate it with the artefacts, the book and the paper, telling the story of seeing how each and every one of us is travelling in time, independently, uniquely, every day.

> To see a World in a Grain of Sand
> And a Heaven in a Wild Flower,
> Hold Infinity in the palm of your hand
> And Eternity in an hour. (William Blake)

8. REFERENCES

Bergson, H. (1923) *The Creative Mind: An Introduction to Metaphysics*, Citadel Press [1992]

Blevis, E. (2011) Digital imagery as meaning and form in HCI and design: an introduction to the Visual Thinking Backpage Gallery. Interactions 18, 5 (September 2011), 60-65.

Blevis, E, Churchill. E, Odom. W, Pierce.J, Roedl.D,, and Wakkary.R. (2012) Visual thinking & digital imagery. In *Proc. CHI EA '12*. ACM, New York, NY, USA, 2715-2718.

Callender, C., and Edney, R. (2001). *Introducing Time.* Icon Books. ISBN 1-84046-592-1.

Hafele J.C., and Keating. R. E. (1972) Science 177, 166.

Lu. E, 'Expedition 7, Relativity' Blog Post. http://spaceflight.nasa.gov/station/crew/exp7/luletters/lu_letter13.html (retrieved 20 March 2014).

Muon Experiment in Relativity. http://hyperphysics.phy-astr.gsu.edu/hbase/relativ/muon.html (retrieved 5 March 2014).

Reinhardt, S. *et al.* (2007). Test of relativistic time dilation with fast optical atomic clocks at different velocities. Nature Physics 3 (12): 861–864.

Shklovsky, V. (1917) *Art as Technique*. In: Russian Formalist Criticism: Four Essays. Trans. Lee T. Lemon and Marion J. Reis. University of Nebraska Press, Lincoln, 1965.

The Quantified Self. http://quantifiedself.com/ (retrieved 4 December 2013).

Toothman, J. How Do Humans age in space? HowStuffWorks. http://science.howstuffworks.com/humans-age-in-space.htm (retrieved 24 April 2012).

US 20030230675A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2003/0230675 A1**

(43) **Pub. Date: Dec. 18, 2003**

(54) **ROTOR INDUCTANCE PROPULSION SYSTEM**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
John St. Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/170,847**

(22) Filed: **Jun. 12, 2002**

Publication Classification

(51) **Int. Cl.**^{7} **B64G 1/40**; B64G 1/42

(52) **U.S. Cl.** **244/172**

(57) **ABSTRACT**

The invention is a spacecraft having a circular, domed hull around which dual electrically-charged rotors one above the other are counter-rotating on the edge of the hull. Embedded in the hull are three solenoids which create a positive vector potential at the rotors. The surface charge density times the radius times the vector potential times the area of the rotors creates an angular momentum in the vertical direction. This angular momentum produces a positive spacetime curvature over the dome of the hull and a negative spiking spacetime compression over the rotors. By machining circumferential grooves of decreasing height along the radius of the rotor, a negative surface inductance is generated. This negative inductance gradient times the negative spacetime compression time the rotor current density squared times the rotor area squared is a positive lift force on the spacecraft.

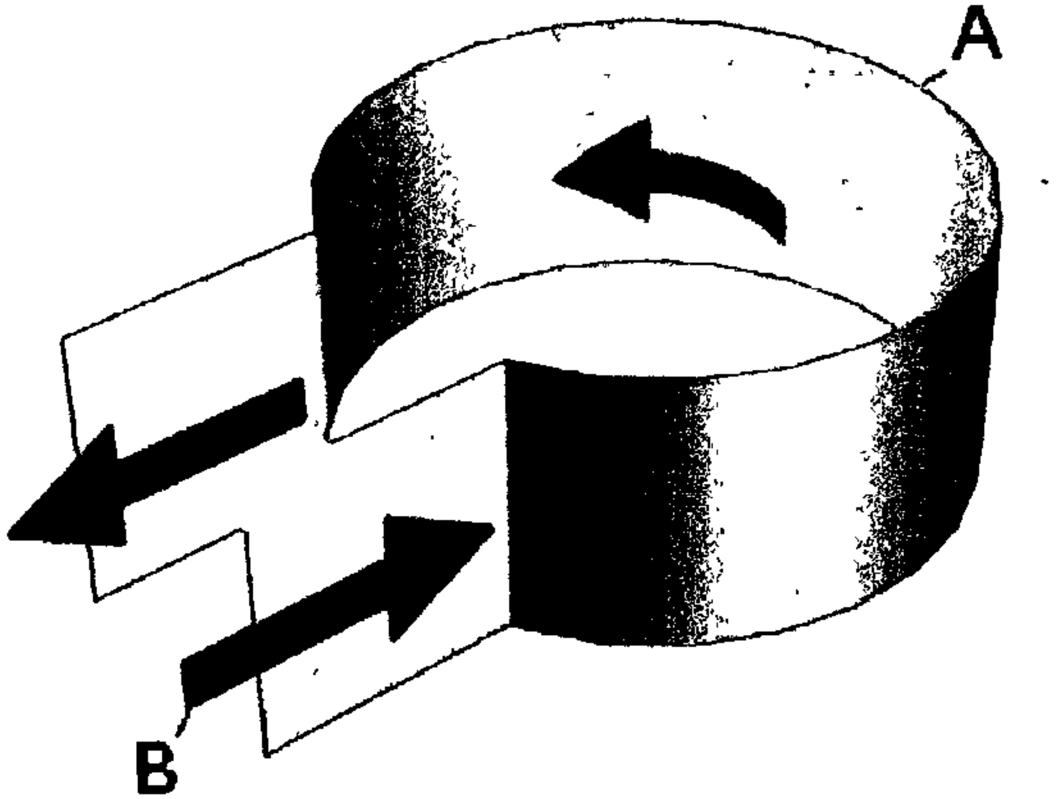

Figure 1

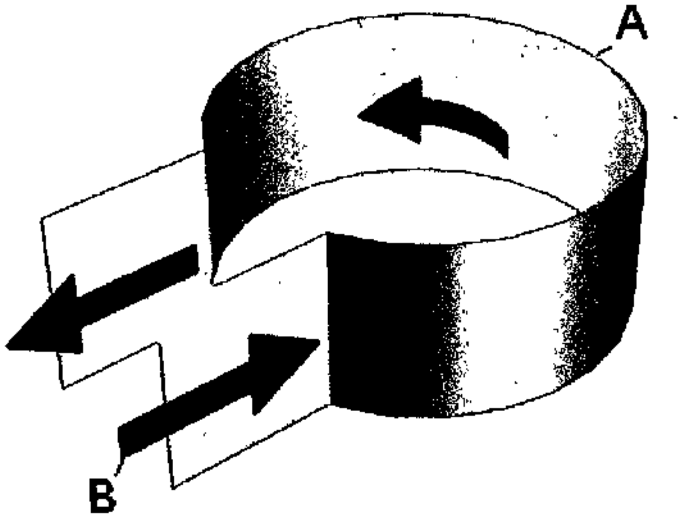

Figure 2

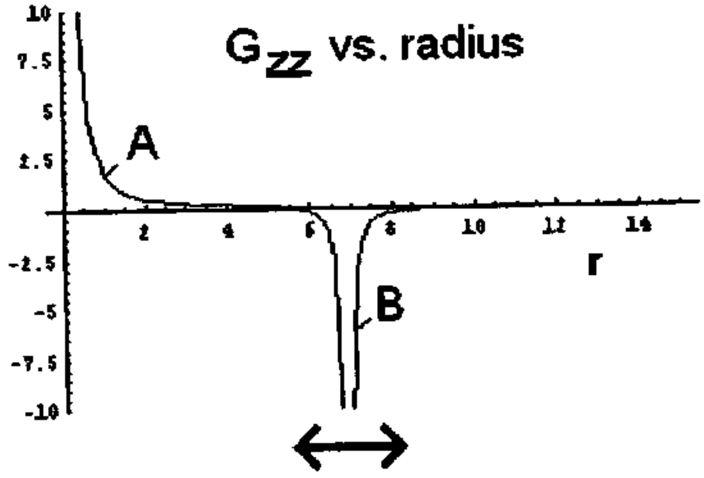

Figure 3

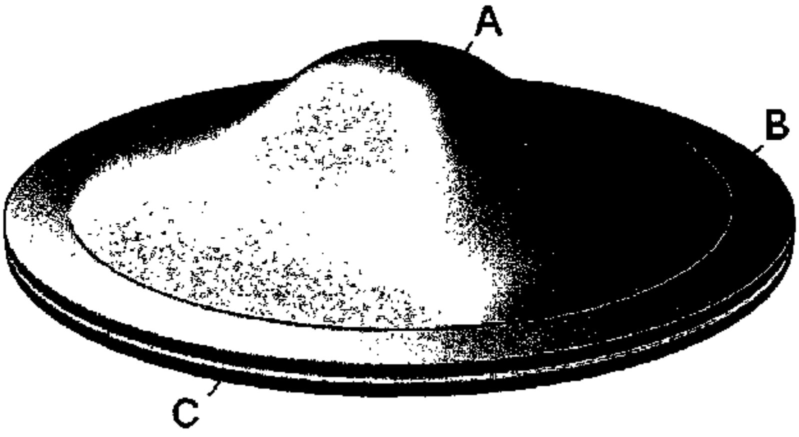

Figure 4

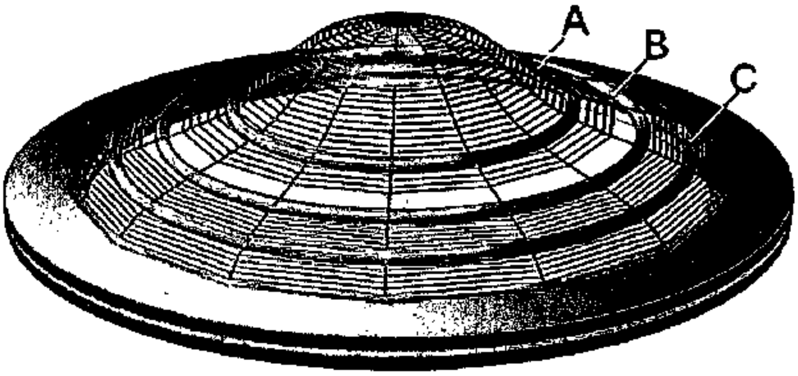

Figure 5

$$A = \frac{n I a^2}{2 \varepsilon_0 c^2} \frac{1}{r'}$$

Figure 6

$$A \equiv \frac{1}{m} \frac{\text{coul}}{\text{sec}} m^2 \frac{m^2 \text{kg}\, m}{\text{coul}^2 \text{sec}^2} \frac{\text{sec}^2}{m^2} \frac{1}{m} = \frac{\text{kg}\, m}{\text{sec}\, \text{coul}}$$

Figure 7

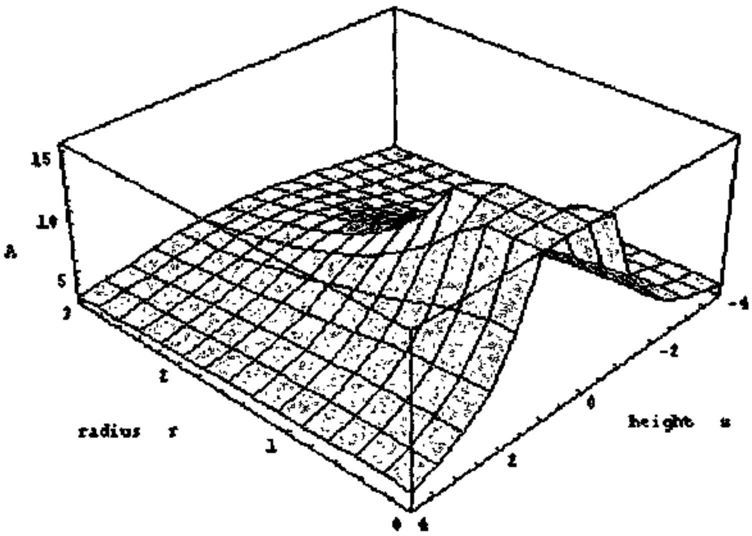
Vector Potential versus radius r and height z

Figure 8

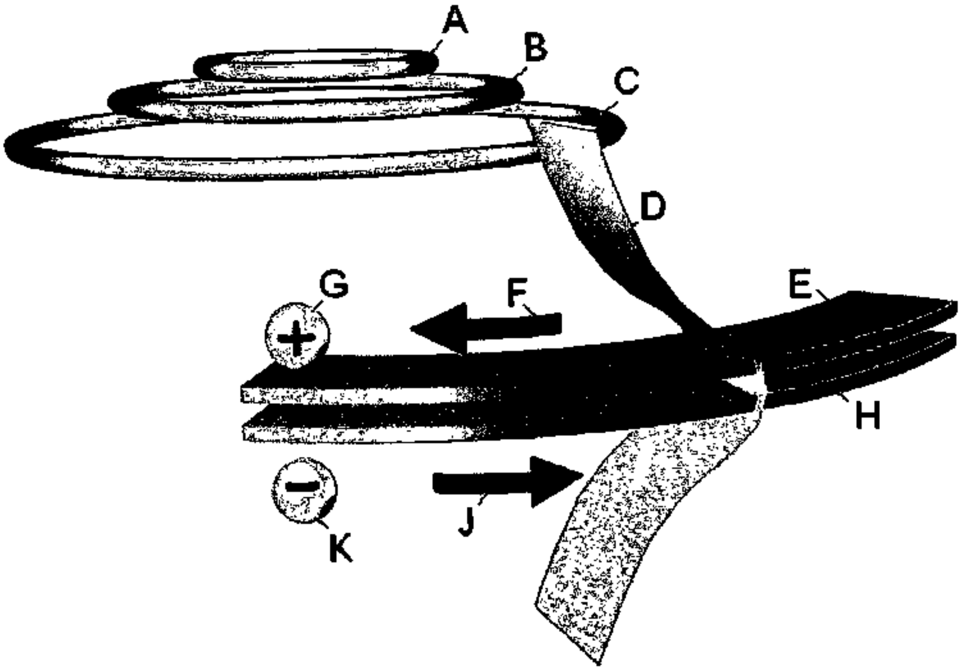
A
B
C
D
G
F
E
H
J
K

Figure 9

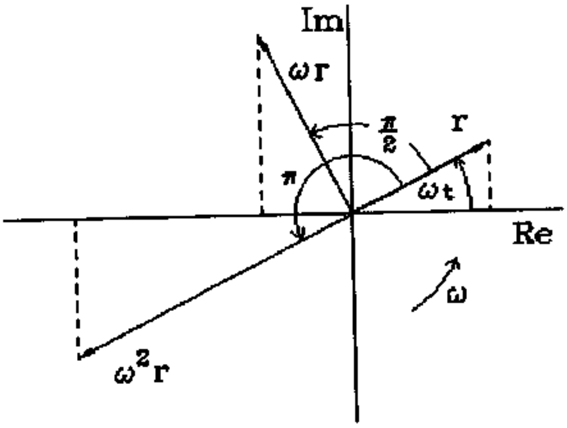
Im
ωr
$\frac{\pi}{2}$
r
π
ωt
Re
ω
$\omega^2 r$

Figure 10

$$
\begin{aligned}
S_k &= \left[-\sigma r e^{i\omega t} + \sigma r e^{-i\omega t}\right] \frac{n I a^2}{2\varepsilon_0 c^2} \frac{n_z \text{area}}{r'} \\
&= -\left[e^{i\omega t} - e^{-i\omega t}\right]\left[\sigma r \frac{n I a^2}{2\varepsilon_0 c^2} \frac{n_z \text{area}}{r'}\right] \\
&= -\left[e^{i\omega t} - e^{-i\omega t}\right] K
\end{aligned}
$$

Figure 11

$$S \equiv \frac{\text{coul}}{\text{m}^2}\,\text{m}\,\frac{1}{\text{m}}\,\frac{\text{coul}}{\text{sec}}\,\text{m}^2\,\frac{\text{m}^2}{\text{coul}^2}\,\frac{\text{kg}\,\text{m}}{\text{sec}^2}\,\frac{\text{sec}^2}{\text{m}^2}\,\frac{\text{m}^2}{\text{m}} = \frac{\text{kg}\,\text{m}^2}{\text{sec}}$$

Figure 12

$$e^{i\omega t} - e^{-i\omega t} = 2 i \,\text{Sin}[\omega t]$$

Figure 13

$$S_k = -2 i K \,\text{Sin}[\omega t]$$

Figure 14

$$ds^2 = \ldots - 4\frac{S}{r}dtd\theta = \ldots 8\frac{iK\,Sin[\omega t]}{r}dtd\theta$$

Figure 15

$$g_{\mu\nu} = \begin{array}{c|cccc|} & t & r & \theta & z \\ t & -1 & 0 & 4\frac{iK\,Sin[\omega t]}{r} & 0 \\ r & 0 & 1 & 0 & 0 \\ \theta & 4\frac{iK\,Sin[\omega t]}{r} & 0 & r^2 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

Figure 16

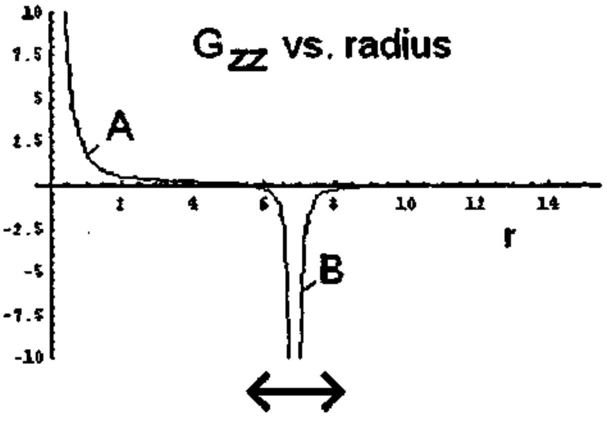

Figure 17

$$\text{Magnetic Energy} = -L\,K\,J^2\,A^2$$

Figure 18

$$\text{Magnetic Energy} \equiv -\frac{\text{kg}\,\text{m}^2}{\text{coul}^2}\frac{1}{\text{m}^2}\frac{\text{coul}^2}{\text{sec}^2\,\text{m}^2}\text{m}^4 = -\frac{\text{kg}\,\text{m}^2}{\text{sec}^2} = -\text{joule}$$

Figure 19

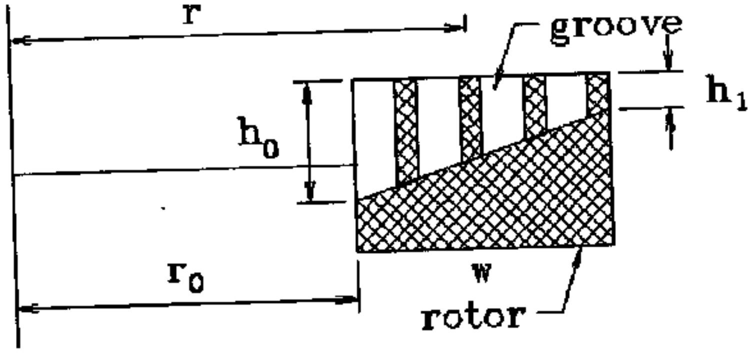

Figure 20

$$h = h_0 - (r - r_0)\left(\frac{h_0 - h_1}{W}\right)$$

Figure 21

$$\frac{dL}{dr} = \mu n^2 2\pi w \frac{\partial}{\partial r}\left(r\left(h_0 - (r - r_0)\left(\frac{h_0 - h_1}{w}\right)\right)\right)$$

$$\frac{dL}{dr} = \mu n^2 2\pi w\left(h_0 - \frac{(h_0 - h_1)(2r - r_0)}{w}\right)$$

Figure 22

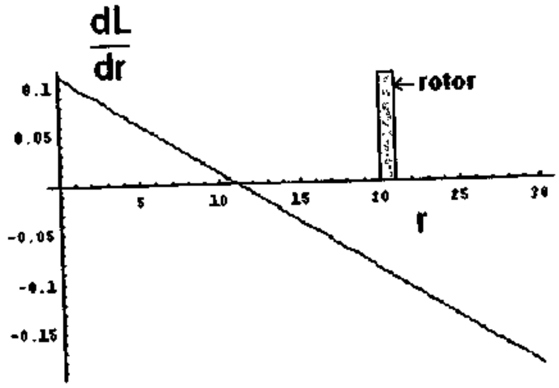

Figure 23

$$F = \frac{dL}{dr} K_c J^2 A^2 = \frac{kg\,m}{coul^2}\frac{1}{m^2}\frac{coul^2}{sec^2\,m^2} m^4 = \frac{kg\,m}{sec^2} = newton$$

ROTOR INDUCTANCE PROPULSION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] The invention, which is the object of my present application, is a spacecraft with a circular, domed hull around which are located dual electrically-charged counter-rotating rotors. The top surface of the upper rotor is etched with circular metallic grooves which give the rotor a surface inductance. The groove height decreases from the inside radius to the outside radius of the rotor giving it a radial inductance gradient. The surface charge density times the angular velocity produces a current density. The counter-rotating rotors produce a negative spacetime curvature over the rotors. The negative surface inductance gradient times the negative spiking spacetime curvature times the current density squared times the area squared is the positive lift force on the rotor.

REFERENCE WORKS

[0002] *Gravitation*, Wheeler

[0003] *Traveling-Wave Tubes*, The Bell System Technical Journal, Vol. XXIX, January 1950, No. 1, J. R. Pierce.

[0004] *The Magnetron as a Generator of Centimeter Waves*, The Bell System Technical Journal, Vol. XXV, April 1946, No. 2, J. B. Fisk, H. D. Hagstrum, P. L. Hartman.

[0005] *Lectures on Physics*, Richard Feynman

[0006] *Geometry of Electromagnetism*, Paul Hammond

BACKGROUND OF THE INVENTION

[0007] I was reading several articles about the development of the magnetron during World War II in the Bell System Technical Journal. I was trying to understand why the device resonates because it must contain a spring constant which would arise from an inductance and capacitance due to the geometry of the cavity. As given by Feynman, inductance of a solenoid is the permeability of space times the turns per length squared times the volume of the solenoid. Referring to **FIG. 1**, the magnetron cavity (A) has a circular region connected to a planar region. The electrical current flows on the sides of the cavity shown by the arrows (B). In this case, there is only one turn per height of the cavity times the volume of the cavity times the permeability which produces the inductance. The two planar regions produce a capacitance across the ends which creates the resonant frequency.

[0008] From my previous patent application Dual Rotor Propulsion System I know that the two rotors produce a current density in the angular direction along the rotor. If I spread out the magnetron cavity into a circular groove around a rotor, then the current would flow on the side walls enclosing the groove volume. The rotors also produce a spacetime curvature profile as shown in **FIG. 2**. Curve (A) is a positive spacetime curvature tension over the dome. Curve (B) is a negative spiking spacetime curvature over the rotors. The curvature is measured in inverse meter squared. So the surface inductance times the negative curvature times the current density squared times the rotor area squared is magnetic energy. The differential of the magnetic energy would be a force. So there has to be a gradient of the surface inductance. The volume element of a groove is equal to the circumferential length times the height times the width of the groove. The easiest is to vary the height with radius. Since the curvature is negative, the gradient has to be negative as well in order to get a positive lift force. Thus the height goes from large to small from the inside to the outside radius.

SUMMARY OF THE INVENTION

[0009] The invention relates to a spacecraft with a domed, circular hull of elliptical cross-section having dual electrically-charged counter-rotating rotors located one above the other on the edge of the hull. The upper rotor is positively charged and rotates clockwise with a negative angular velocity per the right-hand rule. The lower rotor is negatively charged and rotates with a positive angular velocity. The current density is the surface charge density times the velocity of the rotor. This particular combination of velocity and charge produces an angular momentum which creates a negative spiking spacetime curvature over the rotors.

[0010] The top surface of the rotor is etched or machined with circular grooves around the rotor. This creates a surface inductance which is equal to the permeability of space times the turns per length squared times the volume of the groove. In this case, there is only one turn per height of the groove. If the height of the groove decreases from one groove to the next, then there is a negative surface inductance gradient in the radial direction. So the lift force on the rotors would be the negative surface inductance gradient times the negative spacetime curvature times the current density squared times the rotor area squared.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0011] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0012] **FIG. 1**. Perspective of magnetron cavity.

[0013] **FIG. 2**. Spacetime curvature G_{zz} over hull and rotor.

[0014] **FIG. 3**. Perspective view of spacecraft with dual rotors.

[0015] **FIG. 4**. Wire frame view of three solenoids.

[0016] **FIG. 5**. Vector potential equation for solenoid.

[0017] **FIG. 6**. Units of vector potential.

[0018] **FIG. 7**. 3D graph of vector potential using three solenoids.

[0019] **FIG. 8**. Perspective of vector potential along rotors.

[0020] **FIG. 9**. Rotor mechanics diagram using exponential representation.

[0021] **FIG. 10**. Angular momentum equation due to vector potential and rotating charged rotors.

[0022] **FIG. 11**. Units of angular momentum.

[0023] **FIG. 12**. Exponential equation for twice imaginary sine of the angle.

[0024] **FIG. 13**. Angular momentum equation for g metric tensor.

[0025] **FIG. 14**. Angular momentum term for elemental line length ds^2.

[0026] **FIG. 15**. The g metric tensor containing the angular momentum.

[0027] **FIG. 16**. The spacetime curvature tension G_{zz} versus radius due to angular momentum.

[0028] **FIG. 17**. Magnetic energy equation.

[0029] **FIG. 18**. Units of magnetic energy.

[0030] **FIG. 19**. Cross-section of rotor showing groove height gradient.

[0031] **FIG. 20**. Equation for the height h of the rotor groove as a function of radius.

[0032] **FIG. 21**. Equation for the groove inductance gradient.

[0033] **FIG. 22**. Plot of the groove inductance gradient as a function of radius showing that it is negative at the rotor.

[0034] **FIG. 23**. Lift force on rotors due to inductance gradient, compression curvature, rotor current density and area.

DETAILED DESCRIPTION OF THE INVENTION

[0035] 1. Referring to **FIG. 3**, the spacecraft has a circular, domed hull of elliptical cross-section with dual electrically-charged counter-rotating rotors one above the other on the edge of the hull.

[0036] 2. Referring to **FIG. 4**, embedded within the hull are three solenoids or current loops carrying a constant electrical current in the positive sense per the right-hand rule.

[0037] 3. Referring to **FIG. 5**, Feynman has shown that the vector potential {A} of a solenoid is equal to the number of turns per length {n} times the current {I} times the radius of the solenoid {a} squared divided by half the permittivity of space $\{\epsilon_0\}$ times the speed of light {c} squared times the radius {r'} prime from the center of the coil to some location in space such as the rotors. It has been found by physicists Bohm and Aharanov that the vector potential field is not confined to an infinitely long solenoid as is the magnetic B field. The vector potential has units of kilogram-meter per second-coulomb as seen in **FIG. 6**.

[0038] 4. Referring to **FIG. 7**, the three solenoids of varying radius and area produce a positive vector potential at the centerline of the rotors as seen in the graph.

[0039] 5. This graph is then rotated ninety degrees so that it can be located in relation to the rotors as seen in **FIG. 8**. The vector potential (D), which is created by solenoids (A,B,C), passes through rotors (E) and (H). Rotor (E), which has a positive charge (G), is rotating clockwise (F), and rotor (H), which has a negative charge (K), is rotating counter-clockwise (J).

[0040] 6. Referring to **FIG. 9**, rotor mechanics uses the exponential function for the harmonic motion of the rotor. The radius is $\{r\ e^{i\omega t}\}$ which when differentiated with respect to time becomes a velocity $\{i\ \omega\ r\ e^{i\omega t}\}$ where the imaginary {i} is a 90° phase lead which makes the velocity tangential to the rotor.

[0041] 7. The rotor surface charge {σ sigma} is rotating around at some radius {r}. For the upper rotor the surface charge density is positive (+σ) but the rotor has a negative angular velocity {−ω}. For the lower rotor, the surface charge density is negative {−σ} but the rotor has a positive angular velocity {+ω}. So the combined surface charge rotation is $\{-\sigma\ r\ e^{i\omega t}+\sigma\ r\ e^{-i\omega t}\}$. This charge rotation times the positive vector potential due to the solenoids times the rotor surface area is equal to angular momentum S as shown in **FIG. 10**. The units are given in **FIG. 11**. The angular momentum is equal to the negative of the difference of the exponentials times a constant. I then recalled that this difference is equal to twice imaginary sine of the angle as shown in **FIG. 12**.

[0042] 8. Referring to **FIG. 13**, the angular momentum is equal to minus two times the imaginary number times a constant times the sine of the rotational angle. In most of my spacecraft designs, the electromagnetic fields determine the flow rate of angular momentum. When the flow rate is integrated with respect to time, the angular momentum becomes imaginary. In this invention, the angular momentum is imaginary due directly to the rotors. The importance of being imaginary is that the radius resonates with the angular momentum. In some spacetime curvature equations, the denominator has a term equal to the radius to the fourth power plus twice the square of the angular momentum. In spacetime units, angular momentum is viewed as meter squared. So the square of imaginary angular momentum is negative angular momentum equal to negative meters to the fourth power. So at some radius, these two terms are equal, the denominator goes to zero, and the spacetime curvature becomes infinitely-large, creating a huge spike.

[0043] 9. This next section calculates the spacetime curvature from the equation for the angular momentum.

[0044] 10. Referring to **FIG. 14**, the elemental length ds^2 is curved by the presence of the angular momentum which is one of the energies, such as mass, charge, and electromagnetic fields, that Einstein showed can curve spacetime. Because the angular momentum is pointing in the vertical z-direction, due to the direction of the unit normal vector to the surface of the rotor, it rotates around in the angular direction {dtdθ} as found in cylindrical coordinates {t,r,θ,z}.

[0045] 11. In gravitational physics there is a g metric tensor which is a measure of length in spacetime coordinates. It is a 4 by 4 matrix with rows and columns equal to the cylindrical coordinates. Referring to **FIG. 15**, the diagonal of the matrix is $\{-1,1,r^2,1\}$ where the minus one corresponds to time which is Einstein's convention. Half the angular momentum goes in the {tθ} slot of the g metric tensor, and the other half goes in the {θt} slot.

[0046] 12. From this g metric tensor, Einstein's G curvature tensor can be calculated in the various directions. In Einstein's General Theory of Relativity, his equation is $G=8\pi T$ where G is the spacetime curvature measured in inverse meter squared, and the T tensor is the stress-energy-momentum matrix containing all the electromagnetic pressures, mass and momentum components that curve spacetime. The spacetime curvature tension G_{zz} in the vertical direction, as a function of radius, is shown in **FIG. 16**. A positive curvature indicates that there is a spacetime tension over the hull which produces lift. The curvature has a large positive spike over the dome of the hull (A) which means that there is a large lift force over the center of the hull. The curvature then falls off and spikes with a negative spacetime curvature compression over the rotors (B). This curvature also oscillates back and forth a short distance due to the sinusoidal term. The problem is to convert this compression into a lift force which this invention solves.

[0047] 13. This next section shows how the spacetime compression over the rotors generates lift.

[0048] 14. As I mentioned, I have been reading some of the World War II magnetron scientific papers of the Bell System Technical Journal. It turns out that inductance of a solenoid is equal to the permeability of space times the number of wire turns per length squared times the volume of the solenoid. Imagine having a copper strip in the shape of the magnetron cavity in **FIG. 1**. The electrical current (B) flows around the height of the strip (A), so there is only one turn per the height. Then that is multiplied by the volume of the cavity and permeability to get the inductance. The inductor is storing magnetic energy equal to half the inductance times the current squared. In this invention, current density J in amp per meter is used rather than current. So taking a hint from this information, the magnetic energy would have to be related to the inductance times the square of the current density times the area squared of the rotor and times the negative spacetime curvature. The curvature is measured in inverse meter squared. That product produces a negative magnetic energy when in fact I want a positive force. This means that there has to be an inductance gradient in the equation instead of pure inductance and, furthermore, the gradient has to be negative in order to cancel out the negative sign of the compression curvature.

[0049] 15. The equation for the magnetic energy in terms of the inductance $\{\mathcal{L}\}$, curvature {K}, current density {J} and area {A} is shown in **FIG. 17** with the units in **FIG. 18**. The inductance in the equation is proportional to the volume of the circular groove in the top surface of the rotor. That volume is equal to $\{2\pi\, r\, h\, w\}$ where {h] is the height of groove (depth) and {w} is the width located at some radius {r}. In terms of machining, it would be more difficult to machine a variable width groove rather than a deeper groove of a constant width. So I am going to say the gradient is in the radial direction with the height of the groove decreasing going from the inside radius to the outside radius as depicted in **FIG. 19**.

[0050] 16. **FIG. 20** gives the equation for the height of the groove along the radius of the rotor.

[0051] 17. **FIG. 21** gives the inductance gradient by differentiating the inductance with respect to the radius. The initial groove height ho is not very large, and the inductance gradient becomes negative at the rotor as seen in the graph of **FIG. 22**.

[0052] 18. Referring to **FIG. 23**, the lift force on the rotor is now positive due to the combined negative inductance gradient times the negative spacetime compression. The square of the negative current density is positive also. This means that that the top surface of the lower rotor can also have a surface inductance gradient which would double the lift force.

What I claim as my invention is:

1. A spacecraft having a circular, domed hull with dual electrically-charged counter-rotating rotors one above the other located on the edge of hull.

2. Said hull having embedded within it three or more solenoids which generate a positive vector potential at the rotors.

3. Said upper rotor having a positive surface charge density, and rotating clockwise in the negative direction per the right-hand rule.

4. Said lower rotor having a negative surface charge density, and rotating counterclockwise in the positive direction per the right-hand rule.

5. Said rotor surface charge density and velocity creating a negative current density on both rotors.

6. Said vector potential and rotating surface charge density on rotors generating an angular momentum in the vertical direction.

7. Said angular momentum, generating a spacetime curvature tension over the dome of the hull, and a negative oscillating spacetime curvature compression over the rotors.

8. Said rotors having circumferential grooves of decreasing height machined into the top surface of the rotors in order to create a negative surface inductance gradient.

9. Said negative surface inductance gradient times the negative spacetime curvature compression times the rotor current density squared times the rotor area squared generating a positive lift force on the spacecraft.

* * * * *

US 20060014125A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0014125 A1**

St. Clair (43) **Pub. Date: Jan. 19, 2006**

(54) WALKING THROUGH WALLS TRAINING SYSTEM

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
4A
52 KINGS COURT
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/890,635**

(22) Filed: **Jul. 14, 2004**

Publication Classification

(51) **Int. Cl.**
G09B 19/00 (2006.01)

(52) **U.S. Cl.** **434/258**

(57) ABSTRACT

This invention is a training system which enables a human being to acquire sufficient hyperspace energy in order to pull the body out of dimension so that the person can walk through solid objects such as wooden doors.

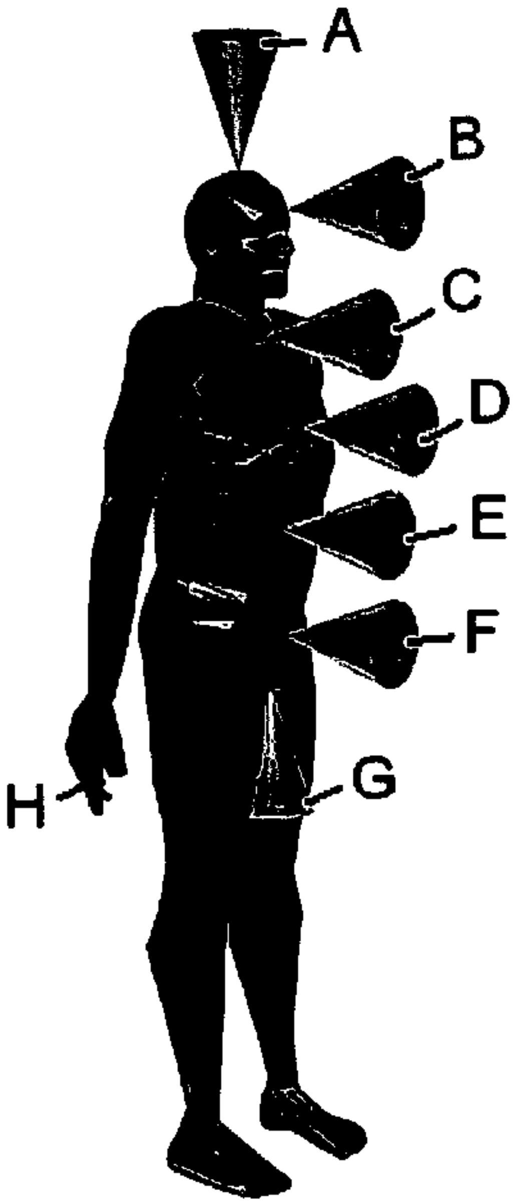

Figure 1

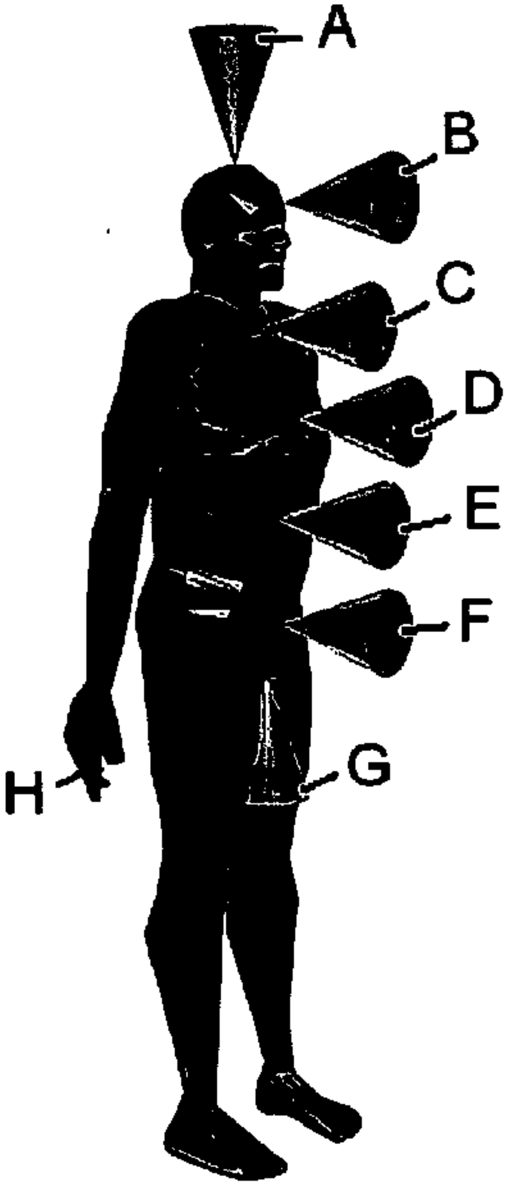

Figure 2

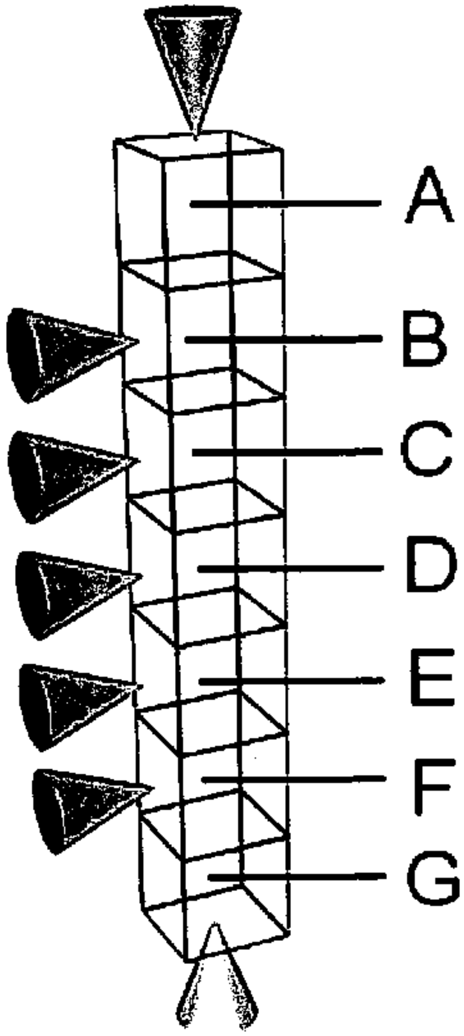

Figure 3

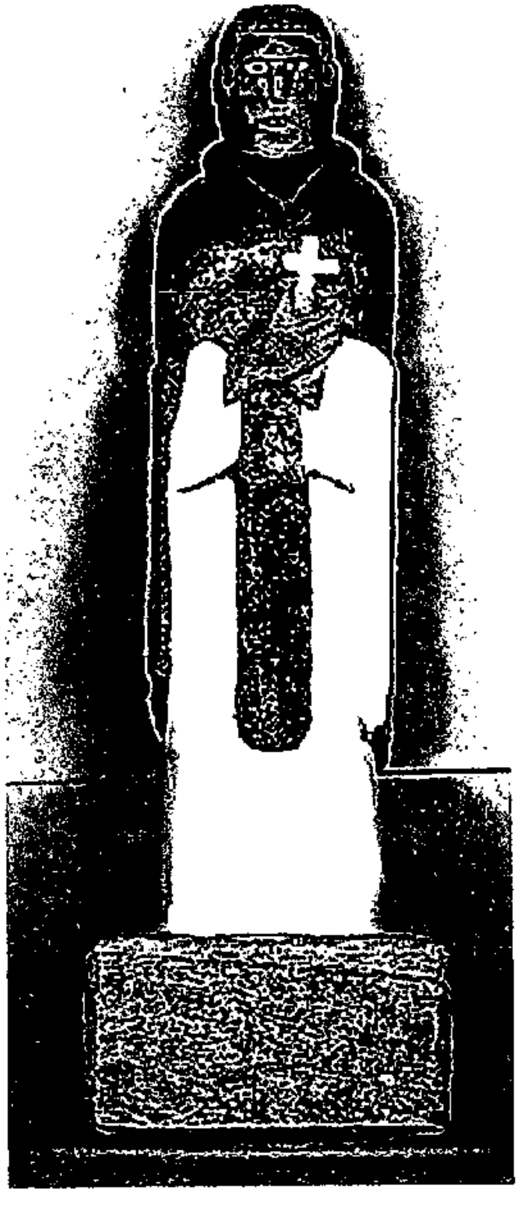

Figure 4

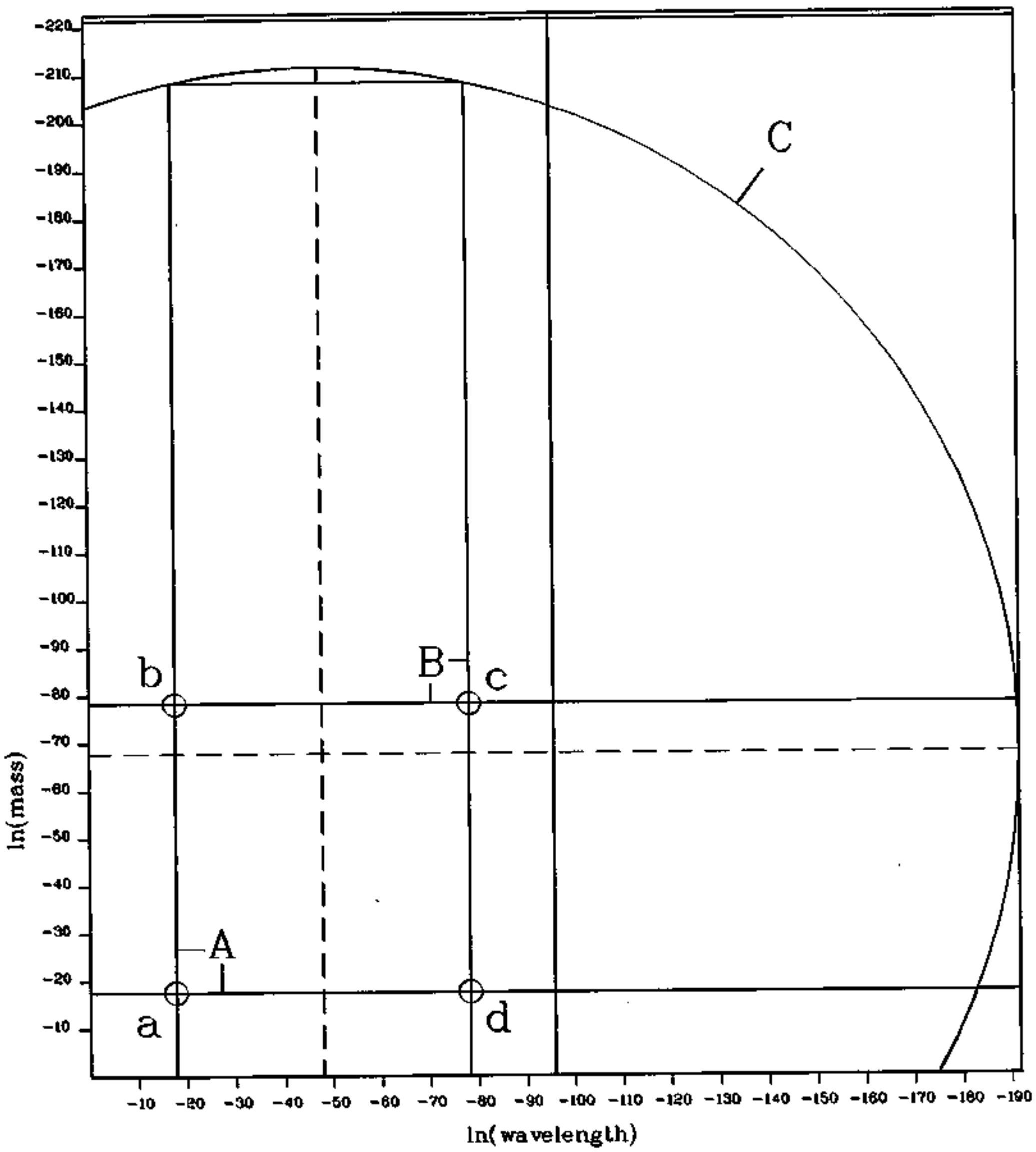
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
C
B
b
c
A
a
d

Figure 5

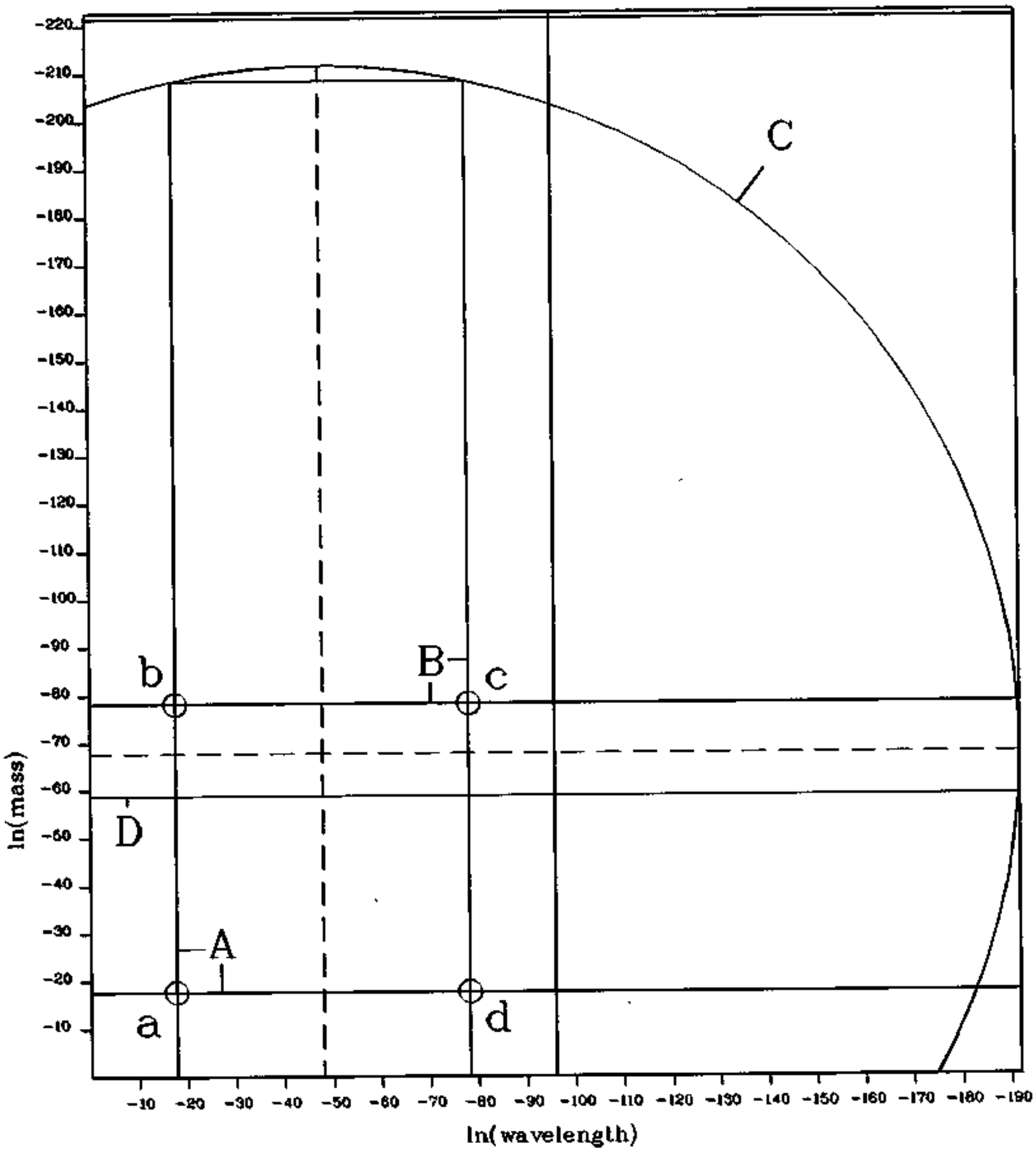
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
C
b
B
c
D
A
a
d
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 6

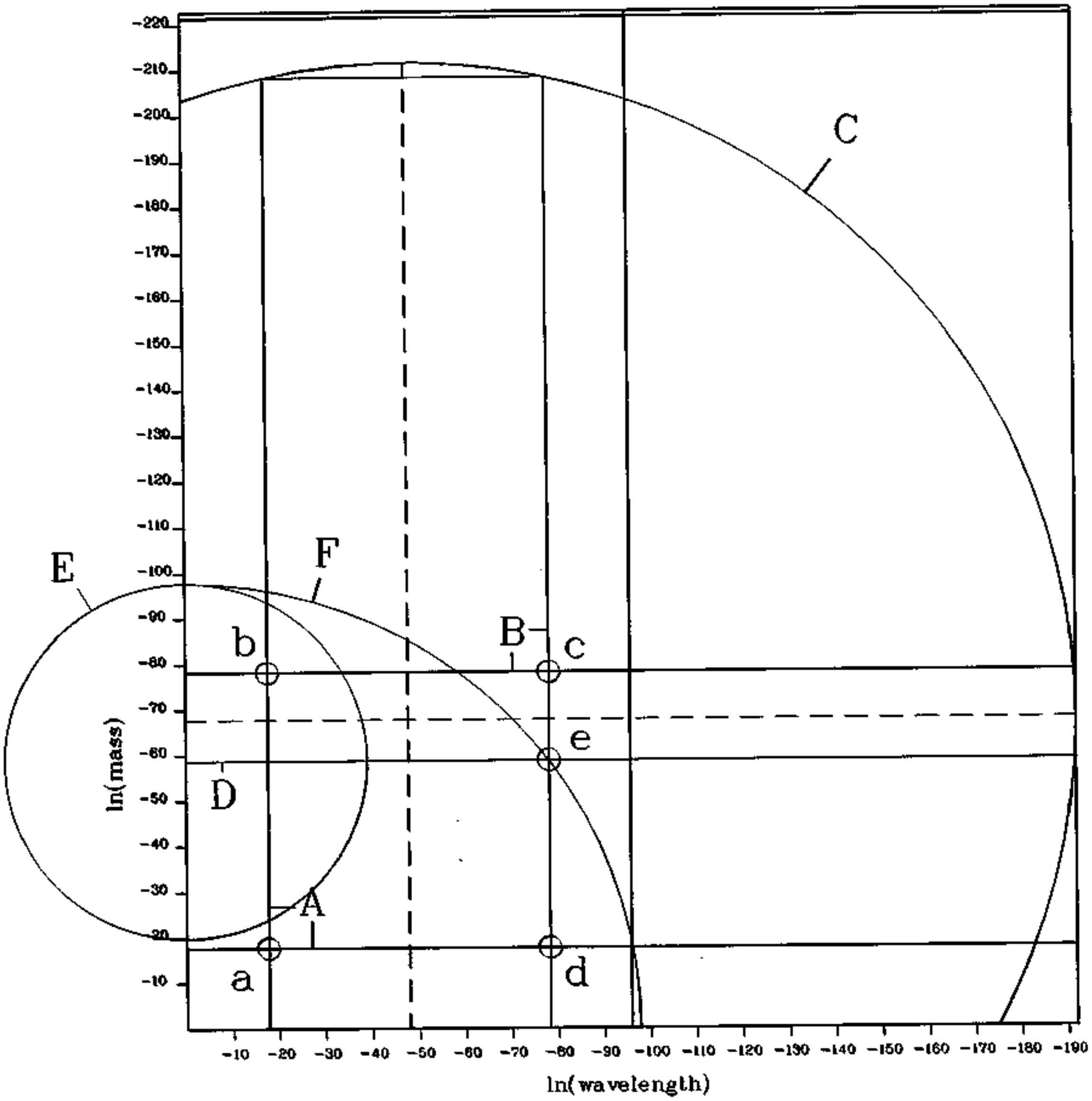
C
E
F
B
b
c
e
D
A
a
d
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190
ln(wavelength)

Figure 7

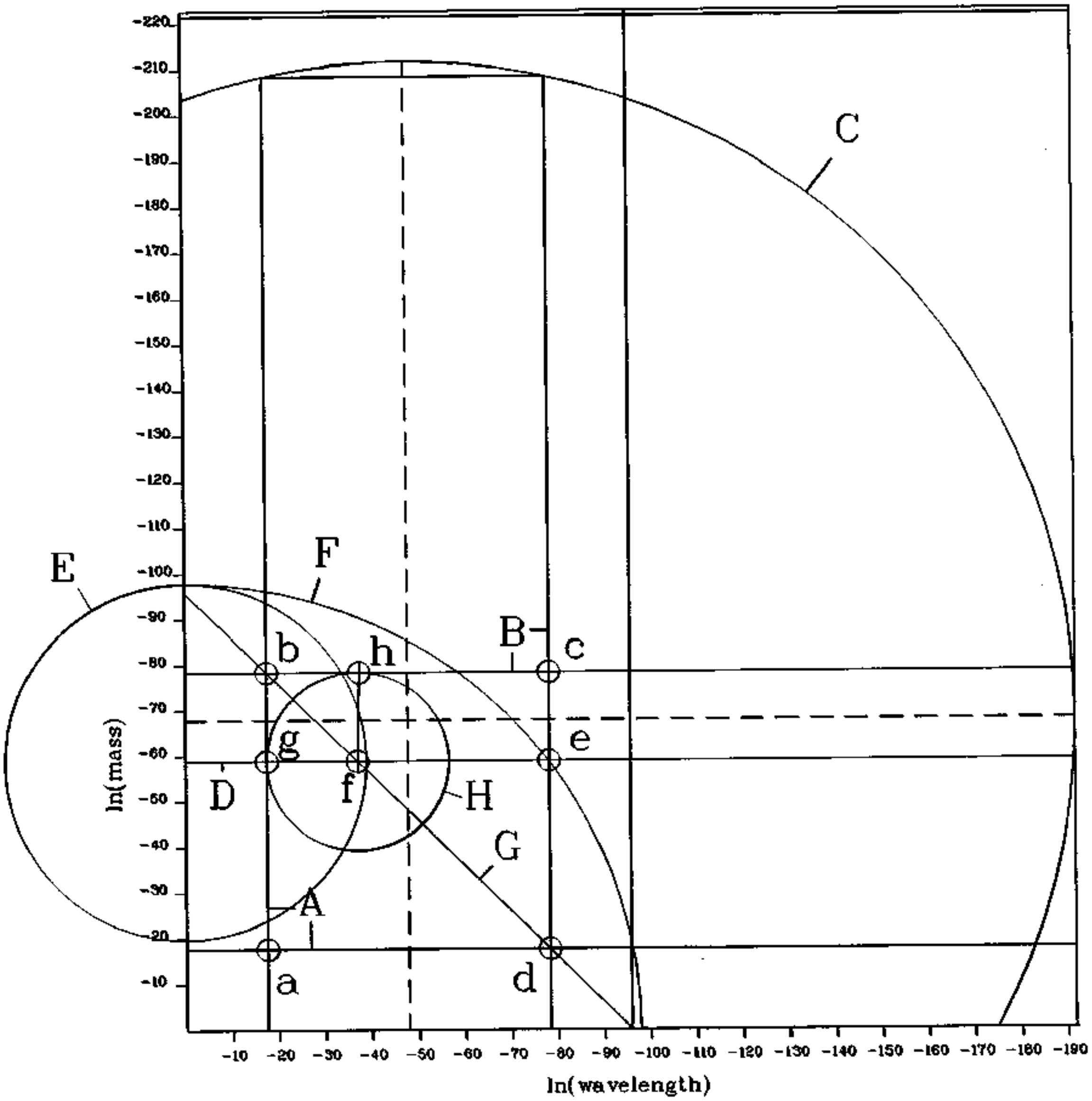
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
C
E
F
B
b
h
c
g
e
f
D
H
G
A
a
d

Figure 8

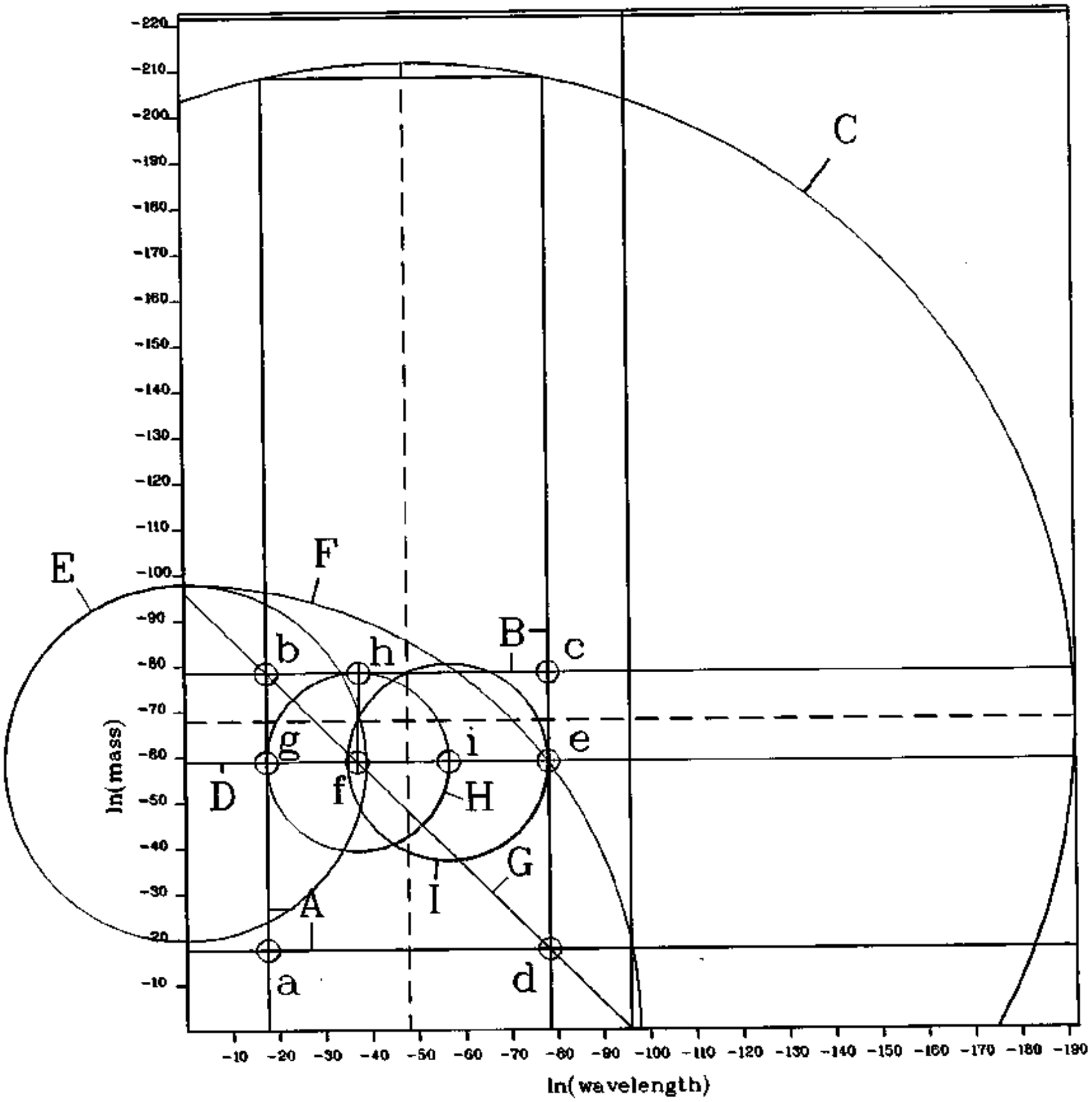
C
E
F
B
b
h
c
g
i
e
D
f
H
A
G
I
a
d
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 9

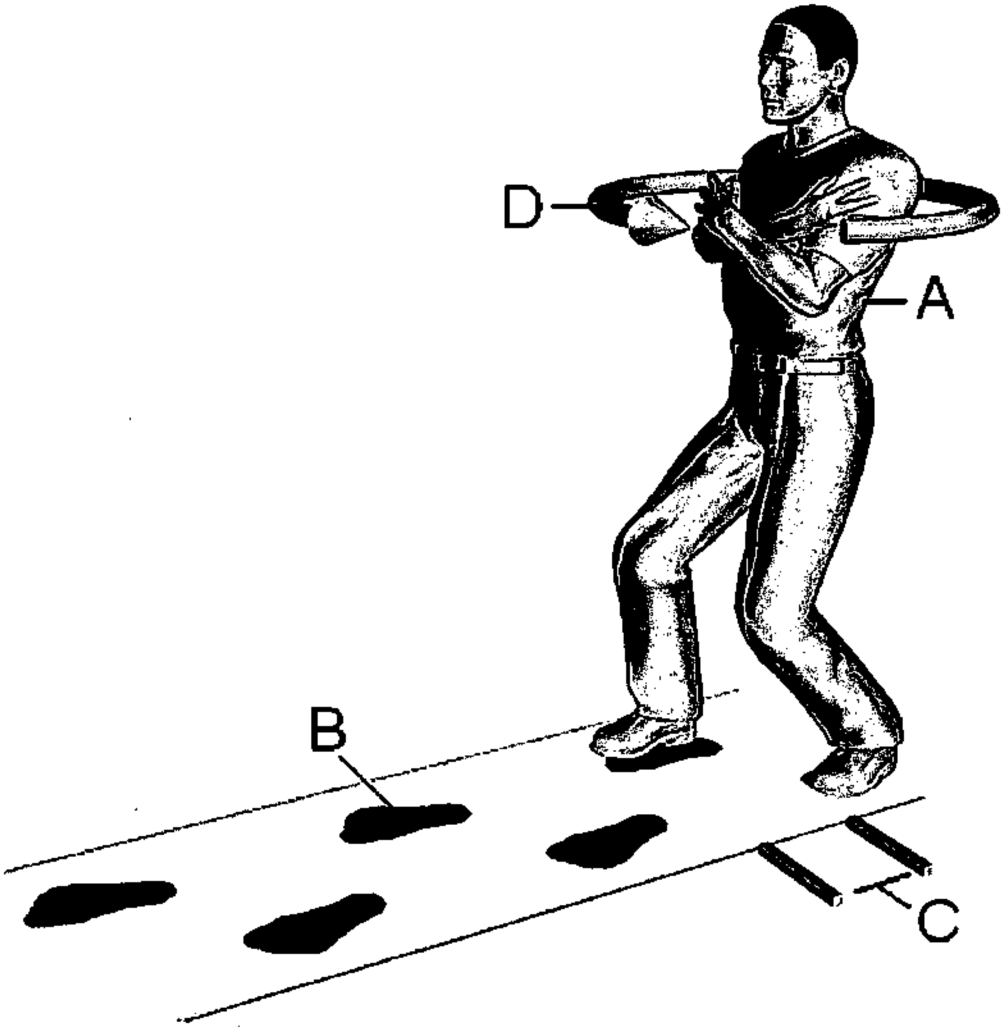

Figure 10

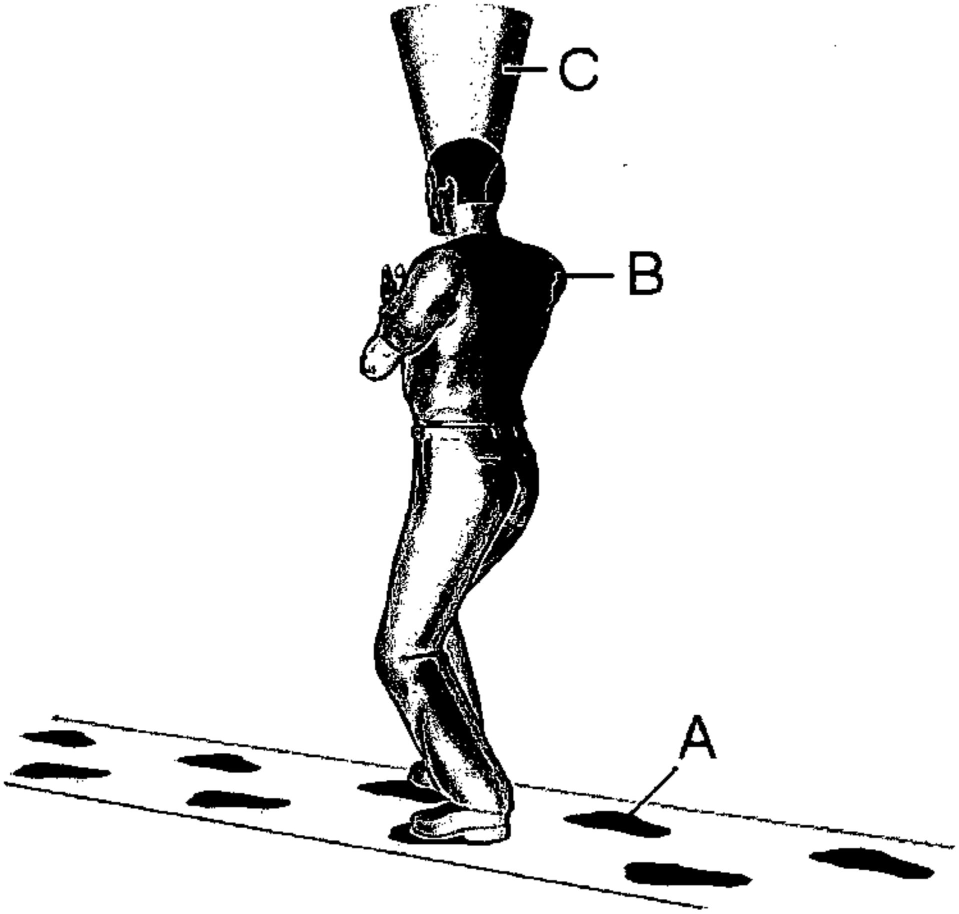

Figure 11

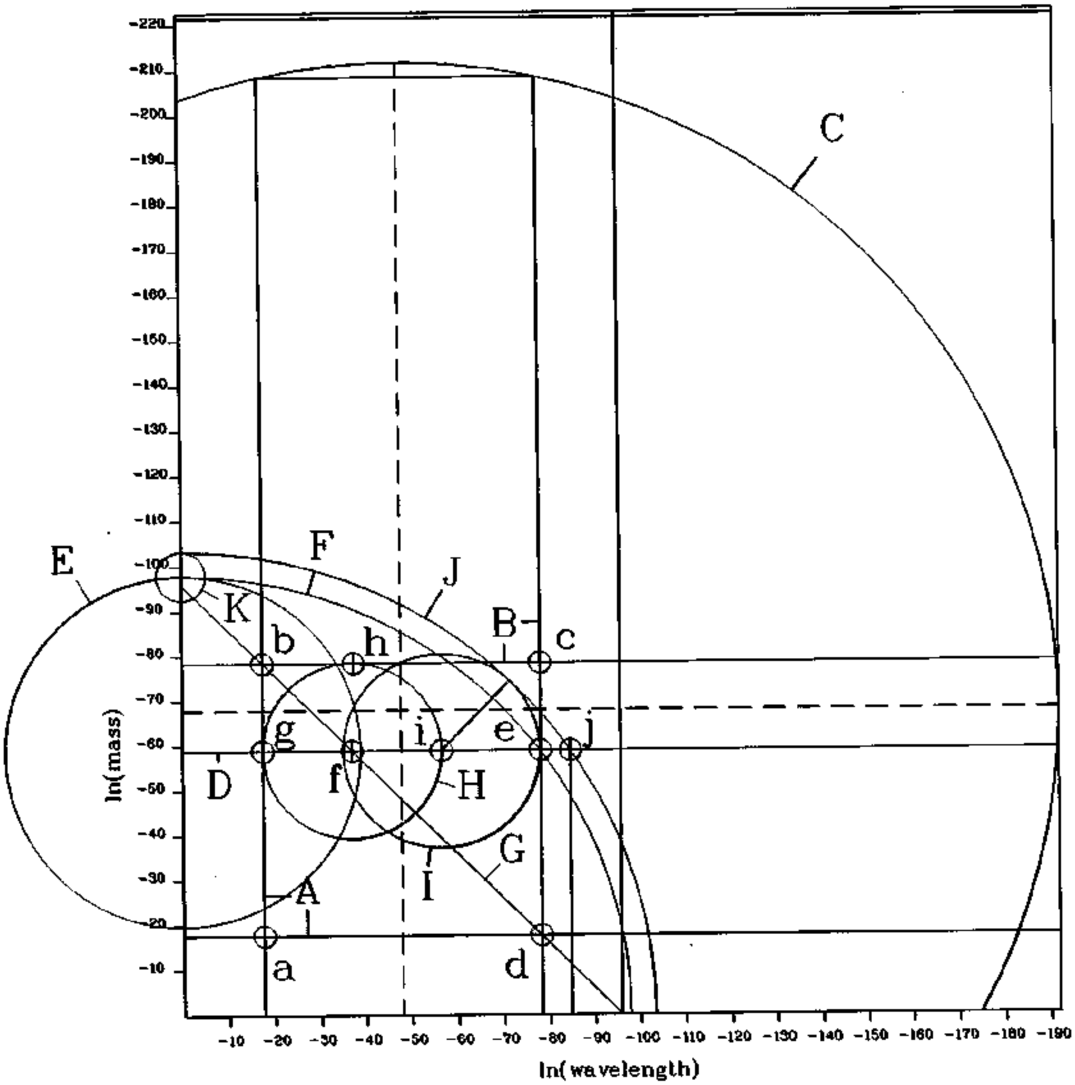
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
C
E
F
J
K
B
b
h
c
g
i
e
j
D
f
H
I
G
A
a
d

Figure 12

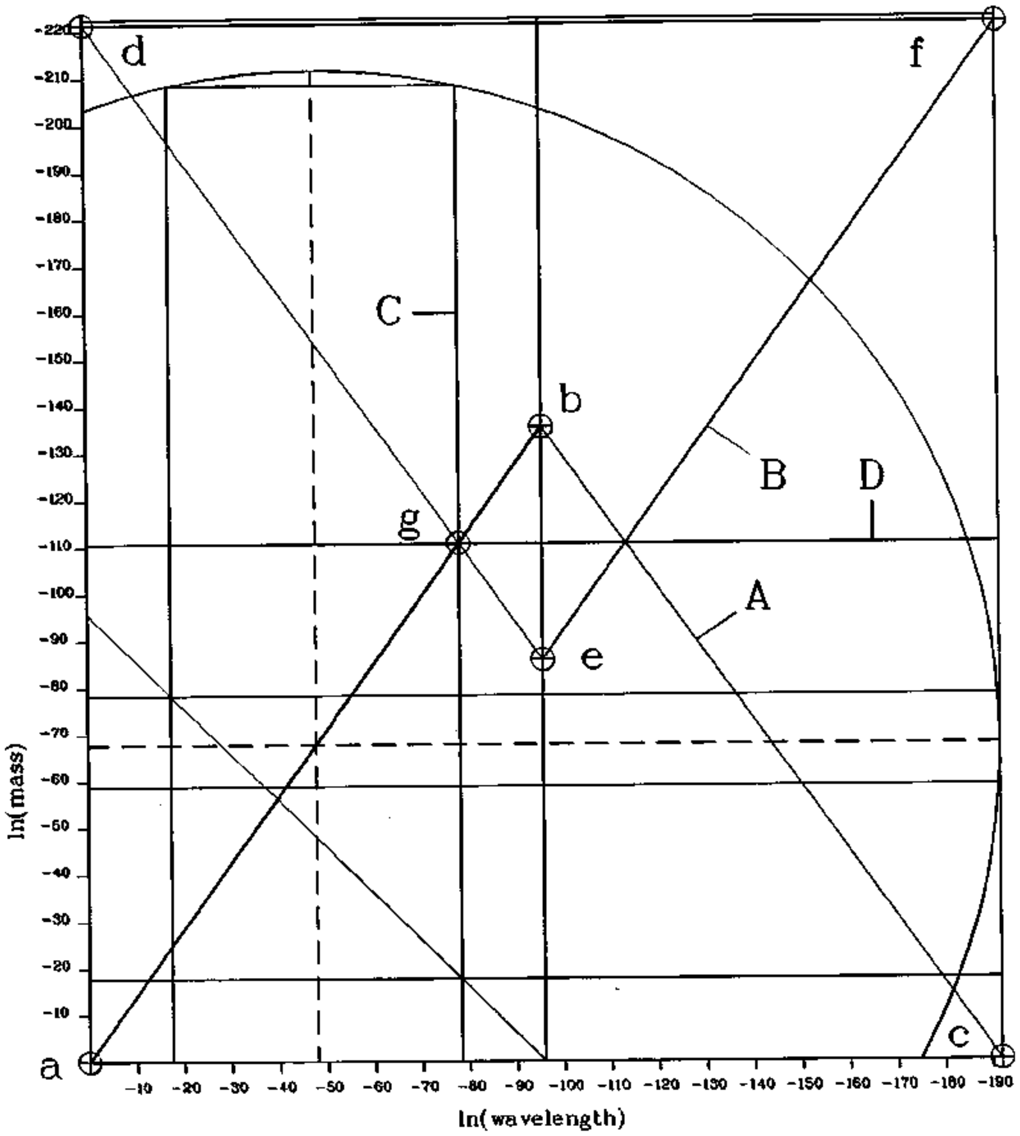
d
f
C
b
B
D
g
A
e
c
a
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 13

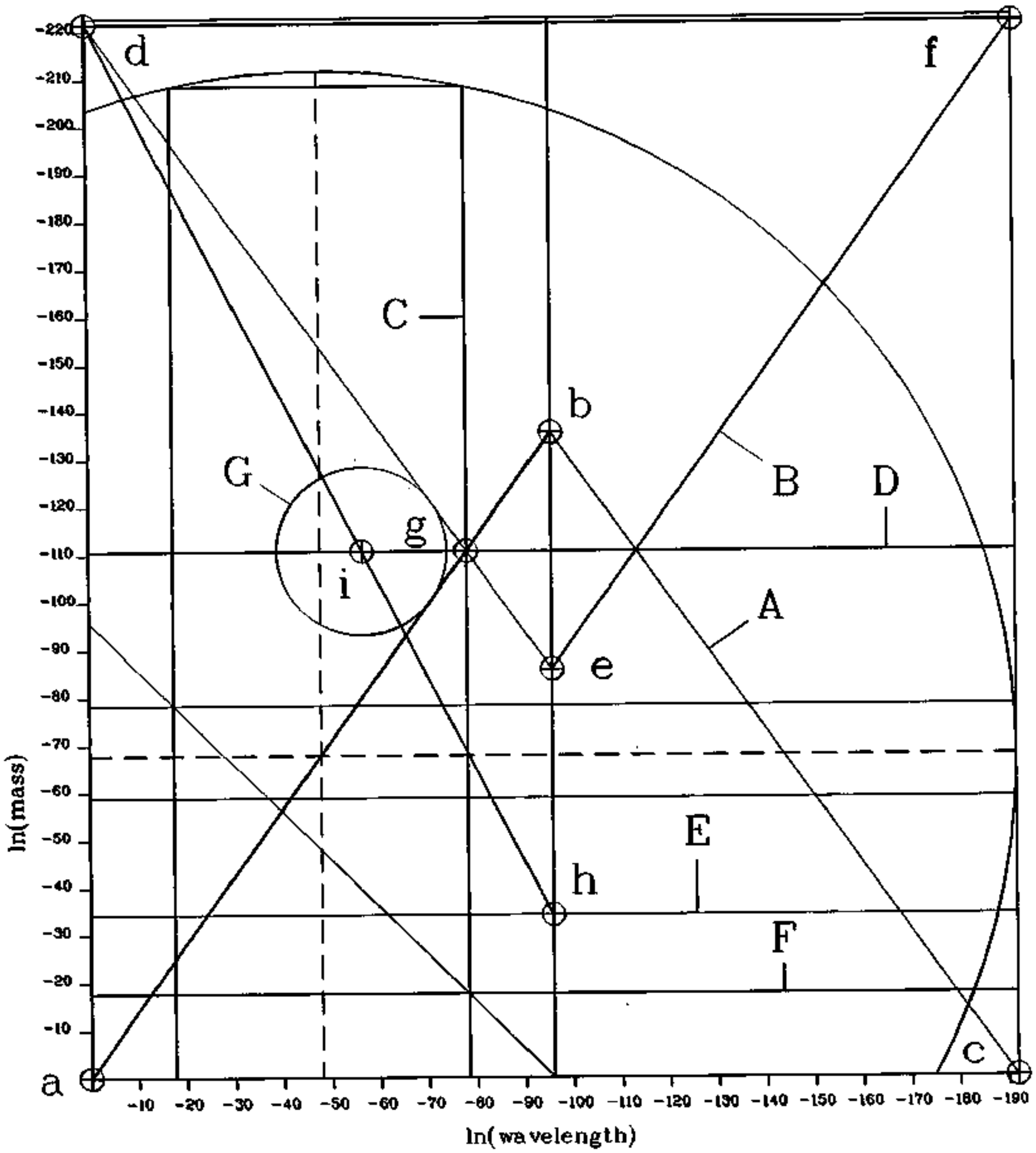
d
f
C
b
G
g
B
D
i
A
e
E
h
F
c
a
ln(mass)
ln(wavelength)
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190

Figure 14

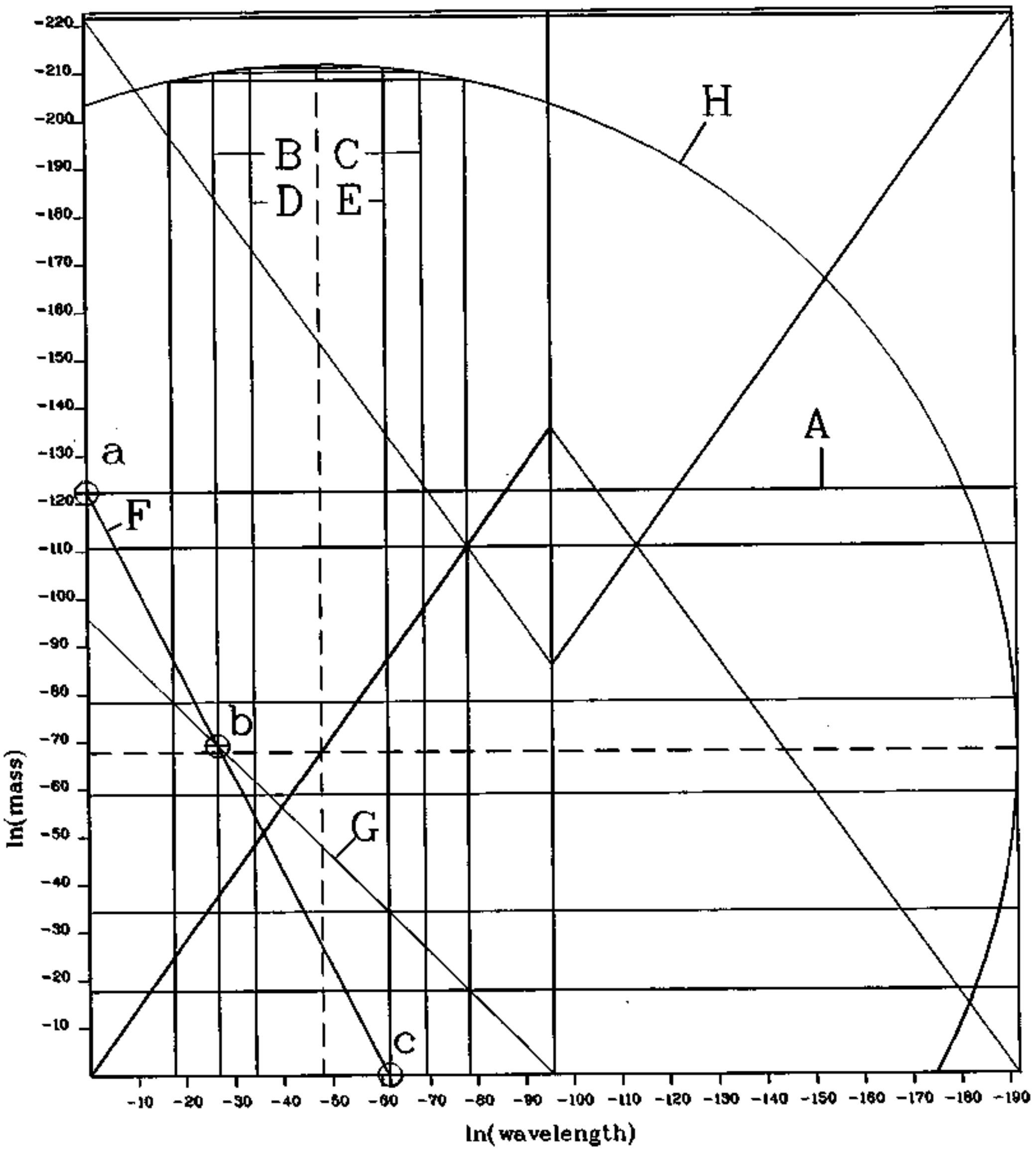
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
a
b
c
A
B
C
D
E
F
G
H

Figure 15

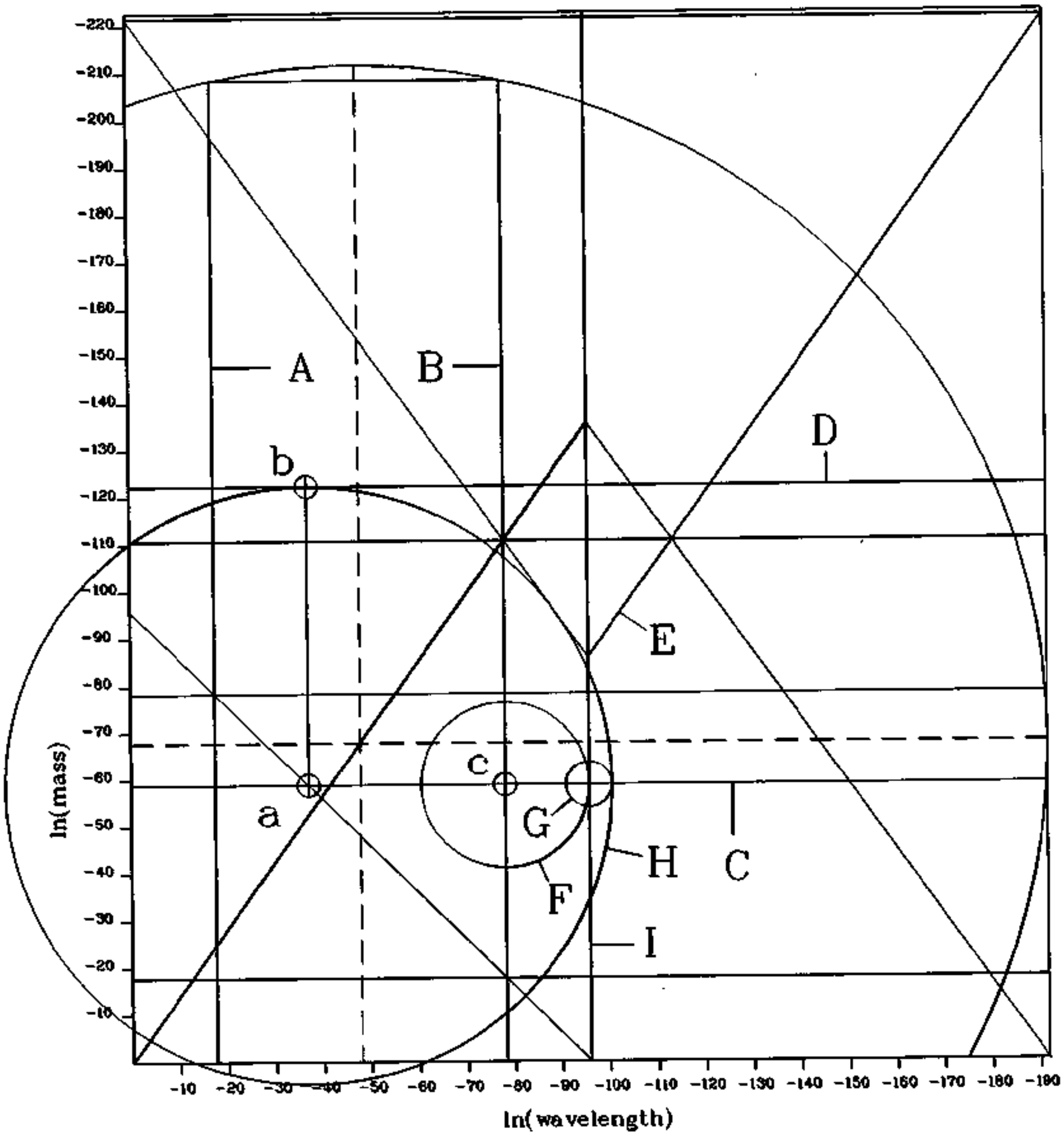
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
A
B
D
b
E
c
a
G
H
C
F
I

Figure 16

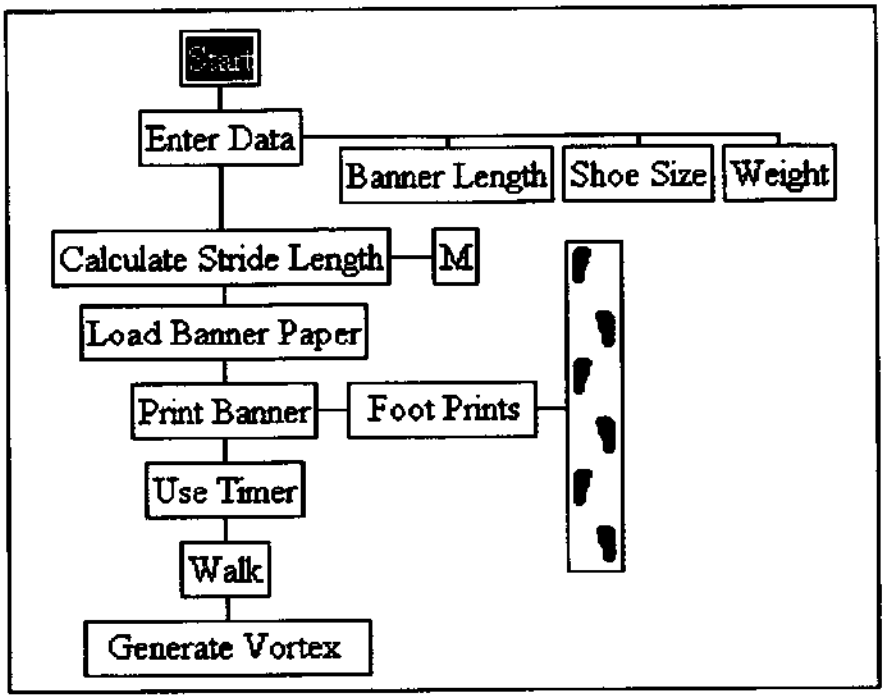
Start
Enter Data
Banner Length
Shoe Size
Weight
Calculate Stride Length
M
Load Banner Paper
Print Banner
Foot Prints
Use Timer
Walk
Generate Vortex

Figure 17

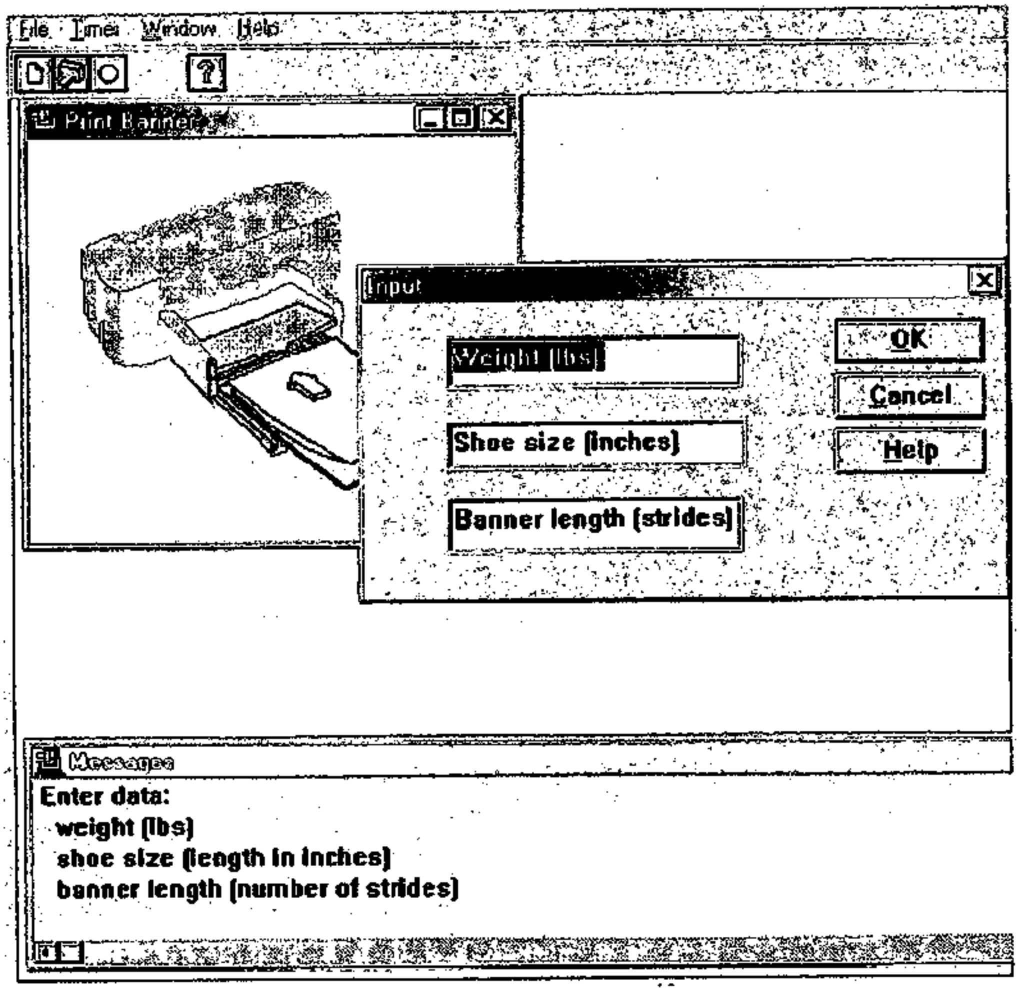
File
Timer
Window
Help
Print Banner
Input
Weight (lbs)
Shoe size (inches)
Banner length (strides)
OK
Cancel
Help
Messages

Enter data:
weight (lbs)
shoe size (length in inches)
banner length (number of strides)

Figure 18

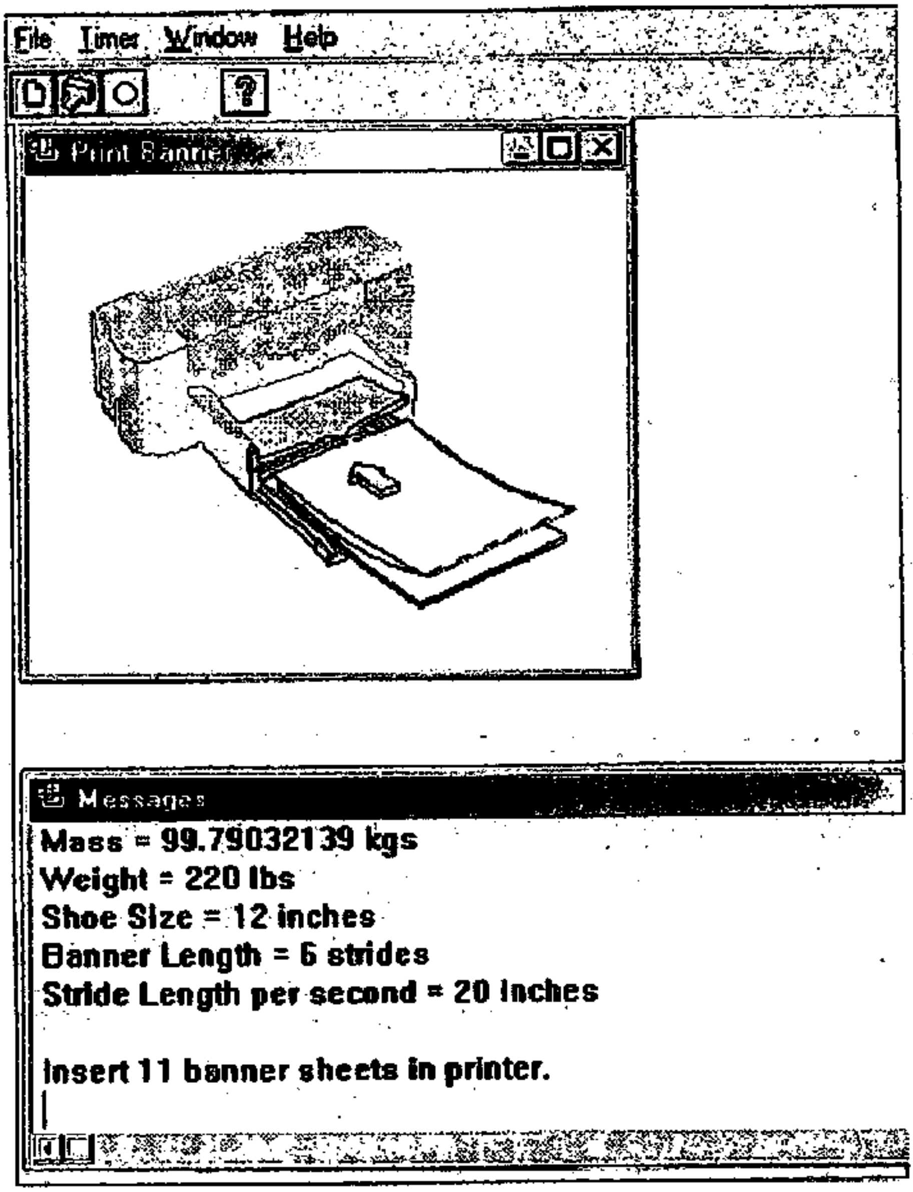
File Timer Window Help
Print Banner
Messages

Mass = 99.79032139 kgs
Weight = 220 lbs
Shoe Size = 12 inches
Banner Length = 6 strides
Stride Length per second = 20 Inches

Insert 11 banner sheets in printer.

Figure 19

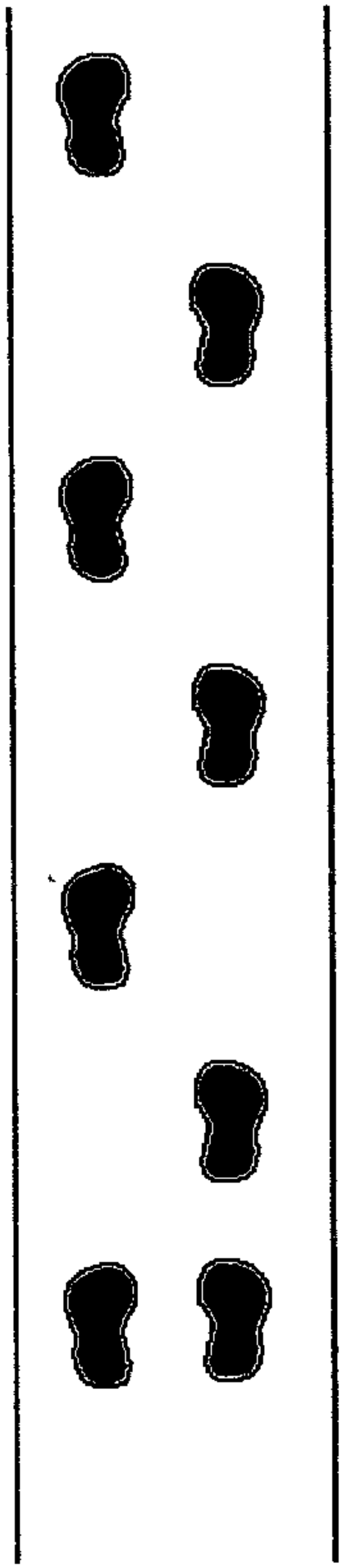

Figure 20

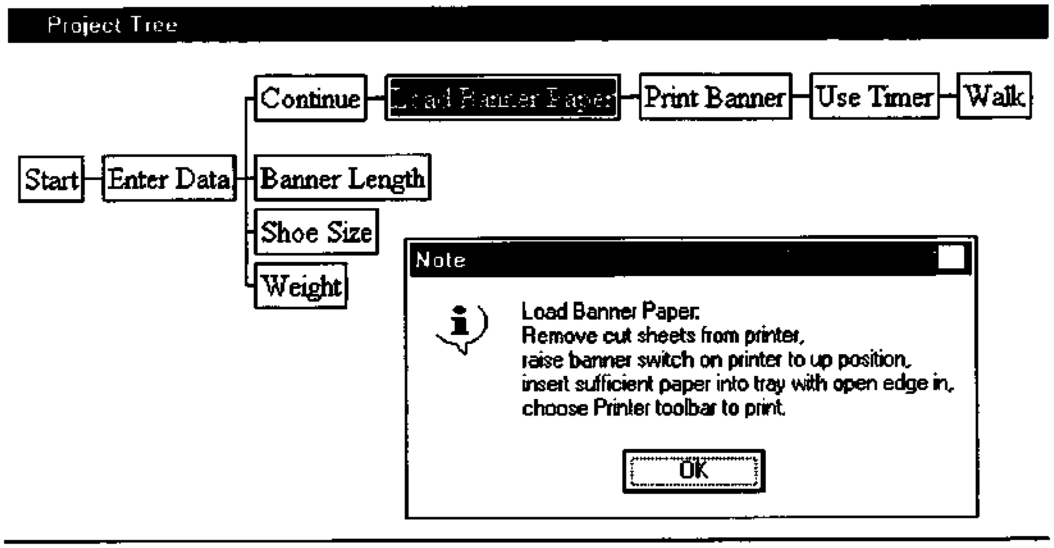
Project Tree
Start
Enter Data
Continue
Load Banner Paper
Print Banner
Use Timer
Walk
Banner Length
Shoe Size
Weight
Note
OK

Load Banner Paper.
Remove cut sheets from printer,
raise banner switch on printer to up position,
insert sufficient paper into tray with open edge in,
choose Printer toolbar to print.

WALKING THROUGH WALLS TRAINING SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a training system that enables a human being to acquire sufficient hyperspace energy in order to pull the body out of dimension so that the person can walk through solid objects such as wooden doors.

BACKGROUND OF THE INVENTION

[0002] A human being is a hyperspace energy being living in a physical container or body that is comprised of 67% water. This high percentage of water makes this invention possible. Referring to **FIG. 1**, the hyperspace energy being receives energy from our dimension through seven vortices that run the length of the body. Each vortex connects to a separate hyperspace dimension having its own particular frequency. This arrangement allows for the development of seven modular energy components corresponding to the mind, spiritual eye, voice, body, abdomen, plasma energy ball (battery), and ground connection.

[0003] Vortex (A), known as the top vortex, supplies energy to the mind and provides a channel of communication to other entities in the universe. This channel has been tested up to 100,000 light years which is the diameter of the galaxy.

[0004] Vortex (B), known as the spiritual eye, has a hollow cone-like appearance surrounded by white, misty, low-density hyperspace energy. Because the eye is modular, the mind can project it to vast distances, a process known as remote viewing.

[0005] Vortex (C), known as the voice and hearing module, can also be projected to enormous distances in order to communicate verbally with other entities. Using the proper remote viewing instruments, it is possible to project both the eye and verbal modules to see and talk at the same time.

[0006] Vortex (D), known as the heart vortex, provides protective energy to the upper internal organs, arms and hands.

[0007] Vortex (E) provides protective energy for the lower organs such as the intestines, liver and kidneys.

[0008] Vortex (F), known as the battery of the body, consists of an orange-colored plasma energy ball about one to two inches in diameter. The body becomes paralyzed if this ball is removed from the body. When the hyperspace energy being leaves the body, the vortices close and the battery takes over in order to keep the body functioning. This vortex also plays a role in the creation of the astral energy baby that attaches to the fertilized egg.

[0009] Vortex (G) is the ground vortex which rotates in the counter-clockwise direction in order to provide an energy ground for the electrical circuit. All the other vortices rotate in the clockwise direction as seen from the front such that energy flows into the vortex according to the right-hand rule of physics.

[0010] Vortex (H) is the hand vortex which rotates counterclockwise on the right hand and clockwise on the left hand as seen looking at the palm. Thus there is a rotating flow of hyperspace energy between the two hands when the palms are facing each other.

[0011] Referring to **FIG. 2**, each vortex feeds energy into its own hyperspace module shown by the lettered box. In terms of quantum mechanics, each box is actually a potential energy well in which each module develops separately. At the time of death of the body, these modules are joined together as a single energy being. The process is powered by the plasma energy ball battery which also contains the logical instructions for assembly.

[0012] Different people, as tested by the pendulum on the hand vortex, have different vortex sizes. Three people were tested. The first person had almost no rotational movement showing very little energy. The second person had a vortex radius of one inch. Another had a vortex radius of four inches which covered his entire hand. The latter also has the ability to lift another human being, lying prone on a table, off the table by flowing low-density hyperspace energy into the person through the hand vortices. He has also experienced walking through a solid wooden door with a dog at his side. What this means is that there is a way, as described in this invention, of creating large energy vortices which will enable a person to acquire sufficient energy to walk out of dimension through solid objects such as wooden doors.

[0013] Researching the historical records, referring to the statue in **FIG. 3**, a humble black Catholic Dominican friar of the Santo Rosario Convent, by the name of San Martin de Porres, living in Peru in the 1500's, also developed this ability to walk through doors. For his beatification, many witnesses came forward to recount his extraordinary abilities. For example, a witness, who worked in the Convent, went to the cell of San Martin to ask for something to eat. When he reached the cell, he saw San Martin leaving with some medications apparently to heal someone sick. The witness waited by the open door of the cell for his return. After having waited awhile without being distracted by anything else, he saw said venerable brother fray Martin come out from his cell from the inside and call him by name. The witness was terrified, not understanding how this was possible.

[0014] The woman who gave me the statue told me that he would walk miles and miles each day to visit the poor. The task of walking means that there is a velocity involved. Because the body has mass, then there is a mass times a velocity, or momentum, involved in this ability. Notice also that the statue shows him walking with his right arm crossed over his left arm in the form of an X.

[0015] What this walking momentum means in terms of physics is given in the following analysis. The human body consists of 67% water. A water molecule consists of two hydrogen atoms and one oxygen atom having the atomic formula H_2O. The atomic weight of one atom of hydrogen is 1.008 awu. The atomic weight of one atom of oxygen is 16.000 awu. The molecular weight of one atom of water is therefore:

Weight of two atoms of hydrogen 2 × 1.008 awu =	2.016 awu
Weight of one atom of oxygen 1 × 16.000 awu =	16.000 awu
	18.016 awu

The formula weight is just the atomic weight expressed in grams. Thus the formula weight of water would be 18.016

grams or 0.018016 kilograms. According to Avogadro's law, the formula weight contains $N=6.02\times10^{23}$ molecules. Thus the mass of one water molecule is the formula weight divided by the number of molecules:

$$\text{mass} = \frac{.018016\ \text{kg}}{N} = 2.99269103\cdot10^{-26}\ \text{kg}$$

$$\ln\left(\frac{\text{mass}}{1\ \text{kg}}\right) = -58.77103943$$

[0016] According to Einstein's Special Theory of Relativity, energy is equal to the mass times the speed of light squared. The energy of a photon is equal to Planck's constant h times the frequency f of the photon. Equating these two energies shows that

$$E = mc^2 = hf = h\frac{c}{\lambda} = \frac{h}{c}\frac{c^2}{\lambda}$$

$$\Rightarrow m\lambda = \frac{h}{c}$$

$$\Rightarrow \ln(m) + \ln(\lambda) = \ln\left(\frac{h}{c}\right) = \text{base} = -95.91546344$$

which says that the natural logarithm of the mass plus the natural logarithm of the wavelength is equal to the natural logarithm of Planck's constant divided by the speed of light c, known as the base constant in the tetrahedron diagram. This diagram plots the mass versus wavelength in natural logarithms. Notice that the left hand side of the equation is the sum of mass and wavelength, so the right hand side must also be the same.

$$\ln\left(\frac{h}{c}\right) = \ln\left(\frac{\Omega\Lambda\ 2\pi\Lambda\ c}{c}\right) = \ln(\Omega\Lambda) + \ln(2\pi\Lambda)$$

where ΩΛ, known as the Planck mass, is the linear mass Ω of the universe times the bottom dimensional limit of the universe Λ, and 2πΛ is the bottom dimensional wavelength, known as the Planck wavelength. That is, our dimension is bounded by the Planck box having sides Planck mass and Planck wavelength. These boundaries have values:

$\ln(\Omega\Lambda)=-17.64290101$

$\ln(2\pi\Lambda)=-78.27256243$

[0017] Referring to **FIG. 4**, these two lines are plotted on the tetrahedron diagram. The Planck mass line (A) reflects off the sphere (C) and returns as the Planck wavelength (B) which shows the dual nature of quantum physics. This creates the Planck box (a,b,c,d) which is the boundary of our dimension.

[0018] Referring to **FIG. 5**, the mass of the water molecule is plotted as horizontal line (D) on the diagram. The energy of the water molecule is the mass times the speed of light squared.

$E=mc^2$

$\ln(c)=19.51860099$

$\ln(mc^2)=\ln(m)+2\ \ln(c)$

[0019] Referring to **FIG. 6**, a circle (E), having a radius equal to the speed of light squared, centered on the mass of the water molecule at the vertical axis, generates a circle (F), centered at the origin, that intersects (e) the mass of the water molecule at the Planck wavelength. That is, this intersection point sits right on the Planck box boundary between space and hyperspace. The radius of circle (F) is actually the mass of the water molecule divided by the speed of light squared.

$$E = \frac{m}{c^2} = m$$

$$\Rightarrow \frac{1}{c^2} = 1$$

$$\Rightarrow c = 1\frac{\text{meter}}{\text{sec}}$$

$$\ln(c) = 0$$

Taking the positive square root, the speed of light is one meter per second at the Planck boundary. The experiments with brain hemisphere resonance show that the resonant frequency of the human energy field is between 1 Hz to 5 Hz which is well below the hearing threshold of 20 Hz. Because the traveling wave has a wavelength of 0.3048 m and the speed of light is unity at the boundary, the frequency should be

$$f = \frac{c}{\lambda} = \frac{1\frac{m}{\text{sec}}}{.3048\ m} = 3.28\ \text{Hz}$$

which is within the middle of the experimental male range. At this resonant frequency, the human energy being pops out of the body. This represents only a first stage in the development of the energy being. But what is really wanted is to have both the physical body and the energy field move out of dimension together as San Martin did.

[0020] Referring to **FIG. 7**, the 45° base line (G) is added to the diagram. Notice that the Planck mass intersects (b,d) the Planck wavelength on this line because they sum to the base constant. The mass of the water molecule crosses this line at point (f). A circle, centered (f), with a radius equal to the speed of light, is tangent to the Planck wavelength (h) and the Planck mass (g). Since mass times velocity is momentum, the diagram says that the momentum of the water molecule is tangent to the boundaries of the Planck box which separates space from hyperspace. In order to get to point (e) from the momentum of the water molecule, a second circle is added to the momentum.

[0021] Referring to **FIG. 8**, a circle (I), centered on the water molecule mass at the speed of light circle (i), is made tangent to the Planck wavelength at point (e). The momentum M of space is equal to the Planck mass times the Planck scale times the speed of light. At point (e), the speed of light is unity, so that the momentum is just the Planck mass in momentum units:

$$M = \Omega\Lambda c = 2.176634194 \cdot 10^{-8}\ \mathrm{kg}\frac{m}{s}$$

Circle (I) has a radius equal to

$$\text{ratio} = e^{21.572952} = 2338912700$$

Therefore the walking momentum in order to get to point (e) is the momentum of space M times this ratio

$$M_w = M \cdot \text{ratio} = 50.909573606\ \mathrm{kg}\frac{m}{\mathrm{sec}}$$

The stride length L per second that a person of mass W has to walk is the walking momentum divided by the mass W times a period T of one second

$$L = \frac{M_w}{W}T$$

For a person with a mass W of 99.79 kg (220 pounds), then the stride length L is 20.08 inches or one foot and eight inches. The person has to walk this length in one second on each foot.

[0022] Looking at the statue of San Martin, his arms are crossed over each other. The vortex of the right hand points backward, and the left hand vortex points forward due to the reversed rotation. Referring to **FIG. 9**, this creates a rotational energy channel (D) around his body (A). The stride length (C) is calculated according to the body mass, and then a banner printout is made showing where the footprints (B) are to be placed each second. The question is: "What happens when one walks the walk?".

[0023] On the very first experiment, referring to **FIG. 10**, what happens is that, after taking only six strides on the banner printout (A), a huge spinning vortex (C) develops over the top of the head and the vertex locks onto the heart vortex in the center of the chest (B). In everyday life, this vortex is not created because normal walking is much faster and the hands are held at the side of the body. The energy rush through the pineal gland is so intense that one feels immediately sleepy and starts yawning excessively due to the increased flow of melatonin.

[0024] After practicing with the banner printout, long walks were made through the park. In this case, a vertical white line rotated around a vertical axis located about six feet perpendicular to the path on the right side of the body. When the walking speed was correct, this white line would lock onto the centerline of the body. Speeding up or down caused the white line to lose synchronization and rotate away. This white line is related to the ability to levitate the body. San Martin had so much energy that, according to witness testimony, he could float horizontally in the air with his head resting against the bowed head of Christ on a carved wooden cross. Thus San Martin's energy sources were channeling energy from Christ, collective broom energy as described in a separate patent application, and the walking momentum vortex energy.

[0025] During the early part of the 20th century, a man's parents were lying in bed dying of tuberculosis. With their permission, he placed a weighing scale under each of their beds. When they passed away, he found that each scale registered a loss in weight of 2.5 ounces. This is equivalent to 0.071 kg, which is the mass of the human energy being.

[0026] After conducting a number of experiments with water vortices draining from a cylindrical tank, it can be stated from Bernoulli's theorem that the potential energy plus the kinetic energy is a constant

$$gz + \frac{1}{2}mv^2 = const$$

The shape of the inner surface of the water circulation has a velocity proportional to the inverse of the radius, so the shape of the surface is

$$(z - z_0) = \frac{k}{r^2}$$

which says that the height of the vortex is proportional to the inverse of the square of the radius.

[0027] The hand vortex area ratio between the second test subject and the third test subject is equal to the square of their radii:

$$\text{ratio} = \frac{(1\ \text{in})^2}{(4\ \text{in})^2} = \frac{1}{16}$$

Because the speed of light at the boundary was determined to be one meter per second, the energy of the third test subject is

$$E = \left(\frac{.071\ \mathbf{kg}}{16}\right)\left(\frac{1\ \mathbf{m}}{\mathrm{sec}}\right)^2 = 4.4375 \cdot 10^{-3}\ \text{joule}$$

$$\ln(E) = -5.417664124$$

[0028] Referring to **FIG. 11**, a circle (K), having this radius, is added to the energy of the water molecule (E), to produce augmented energy circle (J). This circle (J) intersects the mass of the water molecule outside the Planck box at point (j). This means that the increased hyperspace energy moves the water molecule, and hence the body, out of dimension. Furthermore, circle (J) is tangent to the walking momentum ratio circle (I) which keeps the geometry locked together.

SUMMARY OF THE INVENTION

[0029] It is the object of this invention to create a training system that allows a person to develop the ability to walk around out of dimension, passing through solid objects. This invention is based on one of the most remarkable relationships between the water molecule and the boundary between space and hyperspace. The mass of the water molecule is equal to the energy of the water molecule at this boundary. Because the body is composed of 67% water, the body sits on the boundary such that any additional increase in energy would move the body out of dimension into hyperspace. Because human beings are actually hyperspace energy beings living in physical bodies, the additional energy required to move the body out of dimension comes from increasing the energy of the hyperspace being. One source of this energy comes from walking cross-handed at the proper velocity in order to generate a large hyperspace energy vortex that flows energy into the potential wells of the hyperspace being. This increased hyperspace energy will then allow the person to walk around out of dimension through solid wooden doors. Because the door and the person are in two slightly different dimensions at the same moment, it appears that the person is walking through the door. After passing through the closed door, the person then returns to our dimension and emerges in the interior of the closed-door room.

[0030] This technique can be used in reverse to heal an infected hand instantaneously. A salve made from the St.Mary's herb is applied to the skin of the infected hand. The hyperspace energy then flows through the right-hand vortex such that the infected hand and the salve are taken slightly out of dimension. What happens is similar to when a short piece of straw is embedded in a hard wooden telephone pole as a tornado passes over the pole. The straw and pole are taken out of dimension such that they briefly merge together. As the tornado moves on, both objects come back to dimension merged together. Thus the salve (straw) is merged with the bacteria (pole) in hyperspace such that the bacteria is killed instantly. Removing the hand vortex brings the infected hand back into dimension cured.

[0031] Based on this information and the results of many experiments, this invention creates a large vortex by walking at a certain velocity with the hands crossed over the chest. The proper walking momentum is created by a computer program that inputs the person's weight, shoe length, and the number of strides to be taken. The program then prints out a banner showing the footprints where the person has to step each second. When a person obtains sufficient energy from these methods, the person is then tuned to the subspace geometry of the universe as will be shown using the tetrahedron physics diagram.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0032] **FIG. 1**. Energy vortices of the human body.

[0033] **FIG. 2**. Seven potential wells fed by the energy vortices of the body.

[0034] **FIG. 3**. Carved wooden statue of San Martin de Porres who could walk through solid wooden doors.

[0035] **FIG. 4**. Tetrahedron diagram showing boundaries of the Planck box of dimension.

[0036] **FIG. 5**. Tetrahedron diagram showing mass of water molecule.

[0037] **FIG. 6**. Tetrahedron diagram showing water molecule energy and mass are equal at the Planck box boundary.

[0038] **FIG. 7**. Tetrahedron diagram showing that water molecule momentum is tangent to the boundaries of the Planck box.

[0039] **FIG. 8**. Tetrahedron diagram showing the momentum ratio required to reach the Planck wavelength boundary from the water momentum.

[0040] **FIG. 9**. Perspective view of crossed-hand momentum walking using banner printout.

[0041] **FIG. 10**. Perspective view of vortex generated by momentum walking.

[0042] **FIG. 11**. Tetrahedron diagram showing how additional hyperspace energy supplied to the potential wells of the hyperspace energy being enables the human body to be pulled out of dimension.

[0043] **FIG. 12**. Tetrahedron diagram showing the inverted tetrahedrons whose crossing represents the merging of two worlds between space and hyperspace.

[0044] **FIG. 13**. Tetrahedron diagram showing that the proton wavelength is determined by the Planck mass tangent to the inverted tetrahedrons.

[0045] **FIG. 14**. Tetrahedron diagram showing that the mass of the universe determines the electron and proton elementary particles.

[0046] **FIG. 15**. Tetrahedron diagram showing that momentum walking together with the increased energy of the hyperspace energy being is tangent to the mass of the universe.

[0047] **FIG. 16**. Computer program block diagram for printing banner footprints.

[0048] **FIG. 17**. Computer program input dialog window.

[0049] **FIG. 18**. Computer program calculation of stride length per second.

[0050] **FIG. 19**. Six-stride screen banner printout for 220 lb. person.

[0051] **FIG. 20**. Project tree showing help information by double clicking on node.

DETAILED DESCRIPTION OF THE INVENTION

[0052] 1. Referring to **FIG. 12**, a tetrahedron (A, path abc) is added to the diagram. The tip of the tetrahedron (e) falls on the base constant which is equal to Planck's constant divided by the speed of light. A second tetrahedron (B, path def) is inverted around horizontal line (D) which has a geometrical relationship to the base constant. The line is located at

$$\text{centerline} = \frac{2}{\sqrt{3}}\ln\left(\frac{h}{c}\right) = -110.7536373$$

Notice that the intersection of the two inverted tetrahedrons (g) occurs at the Planck wavelength which is the boundary between space and hyperspace. Line (D) is referred to as the merging of two worlds or the connecting of two worlds, a phrase obtained by means of remote viewing. That is, it is the dividing line between space and hyperspace. This is the reason that the two boundaries intersect at this point.

[0053] Referring to **FIG. 13**, the proton wavelength (E) is added to the diagram. The proton wavelength has a value of the electron wavelength divided by 1836.1527

$$\ln(\lambda_p) = \ln\left(\frac{\lambda_e}{1836.1527}\right) = -34.26005901$$

A line (hd), from the base constant at the proton wavelength (h), to the corner of the inverted tetrahedron (d), intersects the merging of two worlds line at point (i). A circle, with a radius equal to the Planck mass (G), centered (i), is tangent to the inverted tetrahedrons. Thus the proton is defined by the base constant and the geometry of subspace. The reason that the proton is tangent to both tetrahedrons is because the electron and proton follow one single path between space and hyperspace. Thus there is only one single particle in nature. Because the particle enters our space at two different locations, we see the one particle as two distinct elementary particles. This relationship can be seen in Library of Congress tetrahedron diagram tet0565.

[0054] Referring to **FIG. 14**, the mass of the universe MU is equal to the linear mass Q of the universe times the radius R of the universe which is 10^{26} meters

$$\ln(MU) = \ln(\Omega R) = 122.3347509$$

as shown on the diagram as line (A). The electron wavelength (B) reflects off the circumscribing sphere (H) and returns as the electron mass (C). The distance between reflection points is the hyperspace charge which is equal to the charge of space less the electron charge. So the electron goes from wavelength to electric charge to mass. The proton wavelength (D) reflects off the sphere and returns as the proton mass (E) which intersects the horizontal axis at point (c). A line (abc) from the mass of the universe at the vertical axis (a) to the proton at the horizontal axis (c) intersects the electron (b) which determines the electron's mass and wavelength since this point is on the 45° base line. What this means is that cosmology determines the values of the elementary particles.

[0055] Referring to **FIG. 15**, the mass of the water molecule (C) intersects the 45° base line at point (a). A circle (F), with a radius equal to the Planck mass is centered on point (c) at the Planck wavelength boundary (B) where the mass of the water molecule numerically equals the energy of the water molecule and where the speed of light is unity. The Planck mass is tangent to the base constant which is the vertical centerline (I) of the diagram. The large circle (H), centered (a) on the water molecule, is tangent to the mass of the universe (b) and tangent to the inverted tetrahedron (E). The difference between this circle (H) and the base constant (I) is the energy the hyperspace being has to acquire in order to be tuned to the mass and geometry of subspace. This difference, shown as circle (G), has a mass

$$m = e^{-4.792671}\ \mathrm{kg} = 8.29\cdot 10^{-3}\ \mathrm{kg}$$

Thus the hand vortex radius ratio has to be the square root of the mass of the hyperspace energy being divided by this tangent mass, or:

$$r = \sqrt{\frac{.071\ \mathrm{kg}\cdot 1\ \mathrm{in}^2}{e^{-4.792671}\ \mathrm{kg}}} \approx 3\ \text{inches}$$

which is a vortex radius that is three times larger than that of a normal person, but one inch smaller than the third test subject who had a hand vortex radius of 4 inches. This is the reason that the third test subject was able to walk through walls and teleport to other locations because his energy was sufficiently large enough to cross over the inverted tetrahedron into a co-dimension of hyperspace. Notice also that dotted circle (J) with a radius equal to the mass of the water molecule, centered (c) on the boundary, is tangent to the tetrahedron (K). This makes the combined geometry tangent to the inverted tetrahedrons and the mass of the universe.

[0056] 2. Referring to **FIG. 16**, a computer program generates a banner printout with footprints spaced for walking according to the weight, shoe size and length of banner desired. Some banners could fit in a small room, or be placed on the floor of a long corridor. As shown in the block diagram the program inputs these three variables with error checking. Then the program calculates the stride length L per second from the equation

$$L = \frac{M}{W}T$$

[0057] 3. In the above equation, the value of the momentum M, as determined by the tetrahedron diagram, is preferably 50.9095736 kg m/s. The weight of the person is converted to mass W in kilograms. The stride period is preferably 1 second. Referring to **FIG. 17**, the data is entered in the dialog input window.

[0058] 4. Once the stride length has been calculated, the program displays the stride length and the required number of banner sheets in the message window, as seen in **FIG. 18**.

[0059] 5. The print banner menu is selected and the foot prints are printed on continuous banner paper. A six-stride scaled screen version of the banner is shown in **FIG. 19**. The banner paper is then placed on the floor, and a one-second beeping timer is activated from the toolbar or menu. The person then walks beside the printed foot prints, taking one stride per beep, which produces the correct walking momentum to generate the hyperspace vortex. The vortex, which forms in only six strides, brings additional hyperspace energy into the quantum potential wells of the hyperspace energy being.

[0060] 6. A help system consists of a project tree which explains the various steps in using the program. Double mouse clicking on a project tree node displays the help instructions in a dialog window as shown in **FIG. 20**.

[0061] 7. In summary, the purpose of the training system is to substantially increase the energy of a human being who will then have the capability of walking through walls, body levitation, instantaneous healing of infections, full-body teleportation to another location, remote viewing at vast distances in terms of light-years, and looking into hyperspace co-dimensions. The third test subject and I have been able to experience all the above phenomena. He did it through augmenting his energy, and I have done it through the application of electromagnetic fields, by spinning on my vortex accelerator machine and using this invention.

I claim:

1. A training method comprising the steps of:

generating a banner having a plurality of footprints spaced at regular intervals wherein the banner is placed on the ground;

generating a periodic audible signal, whereby the audible signal repeats at a regular interval of time equal to the period; and

walking on the banner by tracing the footprints spaced at regular intervals, wherein one step is made with each audible signal.

2. The method of claim 1, wherein the step of generating the banner further comprises:

providing a person's actual mass in kilograms; and

determining the stride length based upon the following equation:

$$L=(M/W)*T$$

where L is the stride length in meters, M is a constant of approximately 50.91, W is the mass of the person in kilograms, and T is the period of the audible signal in seconds, and

wherein the footprints on the banner are spaced at the stride length.

3. The method of claim 2, wherein the audible signal is a beep and the period of the audible signal is one second.

4. A training method for a person comprising the steps of:

generating an audible signal having a fixed period;

generating a banner having regularly spaced indicia for identifying preferred step locations, wherein the distance between adjacent indicia is determined by the following formula:

$$L=(M/W)*T$$

where the distance between adjacent indicia in meters is L, a constant of 50.9095736 is equal to M, the person's mass in kilograms is W, and the fixed period in seconds is T; and

walking on the marked path by stepping upon each of the regularly spaced indicia wherein one step is made with each period of the audible signal.

5. The training method of claim 4 wherein the fixed period is one second.

6. The training method of claim 5, wherein the regularly spaced indicia are footprints.

* * * * *

US 20030209636A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2003/0209636 A1**
St.Clair (43) **Pub. Date: Nov. 13, 2003**

(54) BOBBIN ELECTROMAGNETIC FIELD PROPULSION VEHICLE

(76) Inventor: **John Quincy St.Clair**, San Juan, PR (US)

Correspondence Address:
John St.Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/142,583**

(22) Filed: **May 9, 2002**

Publication Classification

(51) **Int. Cl.^{7}** **B64G 1/32**

(52) **U.S. Cl.** **244/166**

(57) ABSTRACT

This invention relates to a spacecraft which generates its own magnetic moment and magnetic field gradient in order to produce lift on the hull. The magnetic moment is generated by a large area solenoid located in the hull. A toroidal core wrapped with electrical bobbins at intervals along said core produces a traveling magnetic wave along its surface. This magnetic wave creates a spacetime curvature, similar to a tilted plate, which causes the formation of a magnetic field gradient. Power is not critical because the system uses a magnetic vortex wormhole generator to lower the speed of light in order to efficiently create highly relativistic fields due to Lorentz transformation.

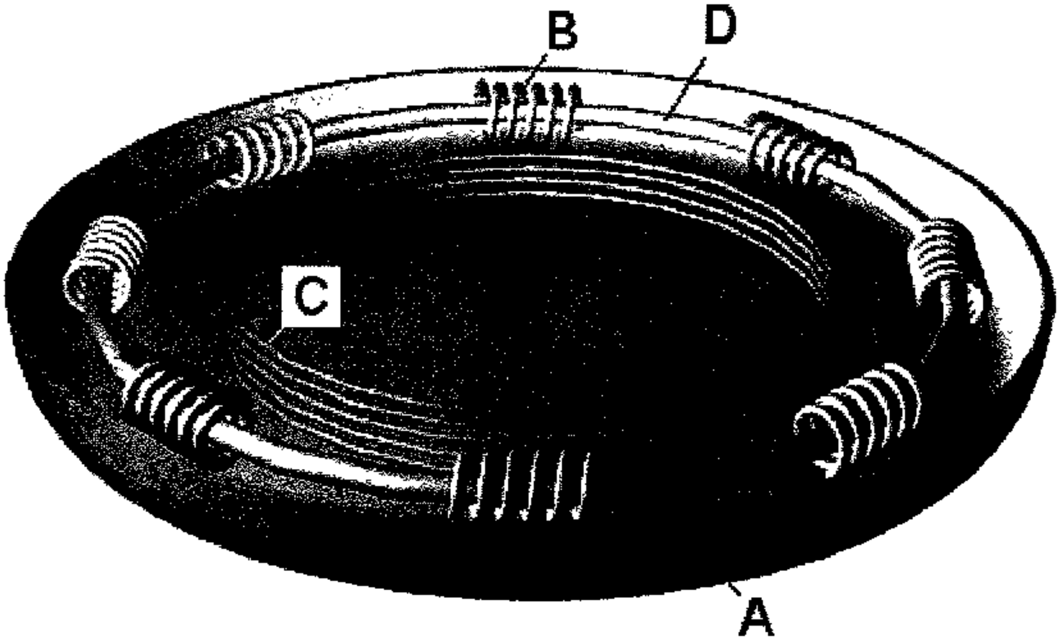

Figure 1

$$\text{Force} = \mu \cdot \frac{\partial B_z}{\partial z} = \text{amp} \cdot \text{m}^2 \frac{\text{kg}}{\text{sec} \cdot \text{coul} \cdot \text{m}} = \frac{\text{coul}}{\text{sec}} \cdot \text{m}^2 \cdot \frac{\text{kg}}{\text{sec} \cdot \text{coul} \cdot \text{m}} = \frac{\text{kg} \cdot \text{m}}{\text{sec}^2} = \text{newton}$$

Figure 2

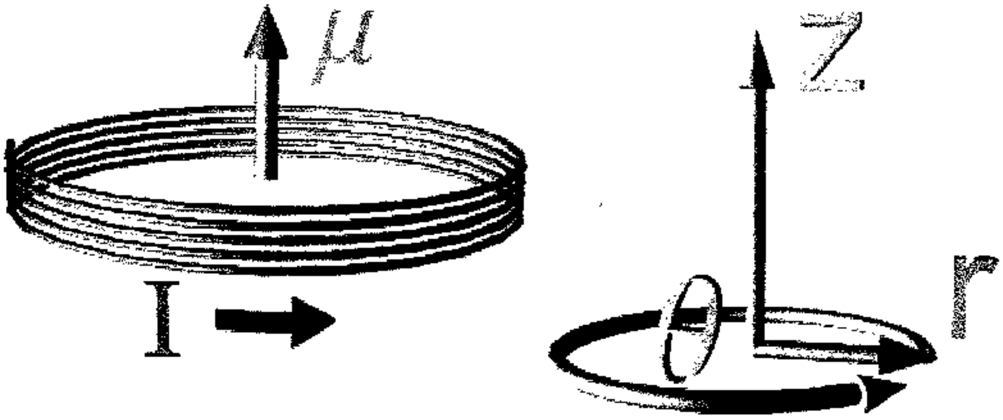

Figure 3

$$B_Z = \mu n I$$

Figure 4

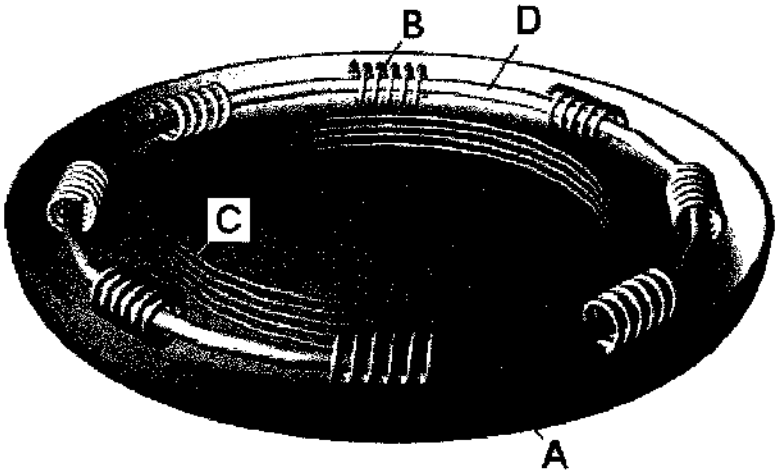

Figure 5

$$v = \sqrt{\frac{\omega}{\sigma\mu}}$$

Figure 6

$$g_{\alpha\phi} = \begin{array}{c|cccc|} & t & r & \theta & z \\ t & -1 & 0 & \mathrm{Sin}[\theta - t] & 0 \\ r & 00 & 1 & 0 & 0 \\ \theta & \mathrm{Sin}[\theta - t] & 0 & r^2 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

Figure 7

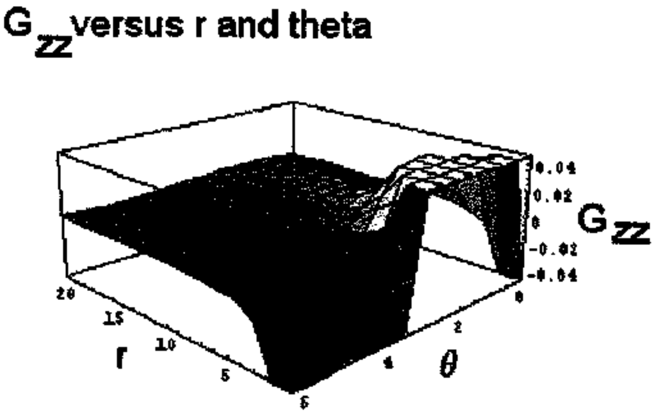

Figure 8

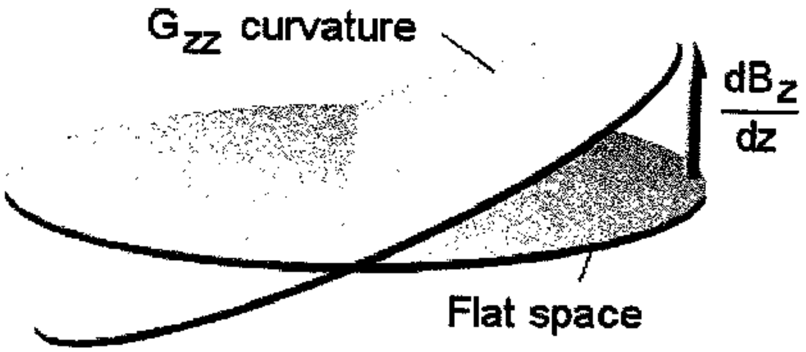
G_zz curvature
dB_z/dz
Flat space

BOBBIN ELECTROMAGNETIC FIELD PROPULSION VEHICLE

BRIEF SUMMARY OF THE INVENTION

[0001] This invention, which is the subject of my present application, is comprised of a toroidal core around which are wound a plurality of electromagnetic bobbins. The bobbins are pulsed electrically to create an amplified magnetic wave that travels around the core. At the same time, a circular electrical conductor carrying direct current creates a magnetic field around its area which results in the formation of a magnetic moment. This magnetic moment, together with the spacetime curvature distortion created by the traveling magnetic wave, produces a lift force on the vehicle.

REFERENCE PAPERS

[0002] Levitron, Hones, U.S. Pat. No. 5,404,062.

[0003] *Geometry of Electromagnetic Systems,* Paul Hammond, page 179.

BACKGROUND OF THE INVENTION

[0004] The idea for this invention comes from experiments I have done using thin transformer laminations wound in intervals with bobbins of wire connected to a frequency generator. Pulsing the bobbins electrically creates a slow traveling magnetic wave along the surface of the lamination. The velocity of the wave, as shown by Hammond in the reference paper, is the square root of the frequency divided by the conductivity and permeability of the material. From Einstein's General Theory of Relativity, this type of wave around the circumference creates a spacetime curvature distortion in the vertical direction that looks like a tilted plate. The magnetic field traveling around the circumference has to follow this curvature. This creates a magnetic gradient in the z-direction which together with the magnetic moment developed by a solenoid produces lift on the vehicle.

SUMMARY OF THE INVENTION

[0005] It is the object of this invention to create a magnetic field gradient and magnetic moment in order to produce a lift force on the hull of a vehicle. The magnetic field gradient is produced by a traveling magnetic wave which produces a tilting-plate spacetime curvature around the hull. The magnetic moment is created by a simple circular wire carrying direct current around its area.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0006] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0007] **FIG. 1**. The lift force on the vehicle is equal to a magnetic moment times the magnetic field gradient.

[0008] **FIG. 2**. Magnetic moment μ created by direct current I flowing counterclockwise around the wire coil. Cylindrical coordinates are shown to the right.

[0009] **FIG. 3**. The wire coil solenoid creates a magnetic field in the z-direction equal to the permeability of space times the number of turns per length of the solenoid times the current in the windings.

[0010] **FIG. 4**. Perspective view of bobbin spacecraft.

[0011] **FIG. 5**. Magnetic wave velocity along core.

[0012] **FIG. 6**. The g metric tensor in cylindrical coordinates with the sinusoidal wave function in the $\{t,\theta\}$ slots.

[0013] **FIG. 7**. The spacetime curvature G_{zz} in the vertical direction created by the magnetic wave traveling around the core.

[0014] **FIG. 8**. Tilted plate spacetime curvature showing magnetic gradient.

DETAILED DESCRIPTION OF THE INVENTION

[0015] 1. Referring to the equation in **FIG. 1**, the lift force on the bobbin spacecraft is equal to its magnetic moment times its magnetic field gradient. The magnetic moment has units of electrical current, measured in amps, times the area enclosed by the current. Thus the units of the magnetic moment are amp-m^2. The magnetic moment can be created by a large, circular coil of wire carrying direct current I as shown in **FIG. 2**. The spacetime cylindrical coordinates $\{t, r, \theta, z\}$ are shown to the right of the drawing where t is time, r the radius, the horizontal angle θ, and the vertical height z.

[0016] 2. The wire coil solenoid also produces a magnetic field in the vertical z-direction equal to the permeability of space μ times the number of turns per unit length of the solenoid n, times the current I in the winding. The equation is shown in **FIG. 3** which can be developed from Ampere's law that the magnetic field around a loop is equal to the current passing through the loop.

[0017] 3. Referring to **FIG. 4**, the bobbin spacecraft consists of a large, horizontal, highpermeability toroidal core (D) wrapped at intervals with electrical bobbins (B) which can be pulsed electrically to create a traveling magnetic wave around the core. Interior to this core is a large direct current solenoid (C) which produces said magnetic field in the z-direction. These devices are enclosed in a circular hull (A) containing the coils on the outer rim and a cabin area in the center.

[0018] 4. Referring to **FIG. 5**, the velocity of the magnetic wave on the surface of the core is equal to the square root of the wave frequency ω divided by the conductivity a times the permeability μ of the core material. When the first bobbin is pulsed, a wave starts to propagate along the surface of the core. As the wave passes the second bobbin in sequence, another electrical pulse is generated to amplify the wave. After many cycles, the wave traveling around the core becomes larger and larger in amplitude.

[0019] 5. A traveling wave has a wave function equal to a sinusoidal function with an argument of the angle θ around the periphery less the time t, or $\mathrm{Sin}[\theta-\omega t]$.

[0020] 6. In gravitational physics, there is a g metric tensor which is a measure of length in spacetime coordinates. When mass or electromagnetic fields are involved in a certain region of space, a curvature of space is created. The curvature of space can then be calculated directly from this metric tensor. The result is Einstein's G curvature tensor which shows the spacetime distortion. The g metric tensor is a 4 by 4 matrix having rows and columns equal to time t, radius r, angle theta, and vertical height z. The diagonal from top left to bottom right has a signature equal to $\{-1, 1, r\hat{}, 1\}$.

[0021] 7. Because the magnetic wave traveling around the core is varying with time t in the theta direction, the wave function has to go into the $\{t,\theta\}$ and $\{\theta,t\}$ slots of the metric tensor, as shown in **FIG. 6**.

[0022] 8. From this metric tensor, Einstein's G curvature tensor is calculated using a general relativity software program. The spacetime curvature in the vertical z-direction is contained in the G_{zz} component. A 3D plot of this curvature with respect to radius and theta is shown in **FIG. 7**. The axis on the right side is the angle around the periphery, and the axis on the left is the radius which goes from 0 to 20 meters. The vertical axis of the graph is the spacetime curvature in the vertical z-direction. The center of the vertical axis is zero. At a small radius, there is a sinusoidal curvature which is positive from 0 to π and which is negative from π to 2π. What this looks like is a tilted plate as depicted in **FIG. 8**.

[0023] 9. In flat spacetime with no electromagnetic fields or mass, the curvature would be the horizontal plate as seen **FIG. 8**. Due to the traveling magnetic wave, the spacetime curvature looks like the tilted plate. The flat-space magnetic field was pointing up in the vertical direction, having no divergence and no gradient. In curved spacetime, however, the magnetic field becomes tilted just like the plate and a magnetic gradient dB_z/dz is created. This spacetime curvature gradient of the magnetic field times the magnetic moment of the second coil produces a lift force on the hull which is anchored to these coils.

What I claim as my invention is:

1. A spacecraft that generates a magnetic moment and a magnetic field gradient in the vertical direction in order to create a lift force on the hull.

2. A toroidal core wrapped with electrical bobbins at intervals around the core whose purpose is to create and amplify a magnetic wave that travels along the surface of the core.

3. A circular, direct-current carrying solenoid, located radially inside the toroidal core, to create the magnetic moment.

4. Said magnetic surface wave creates a spacetime curvature, similar to a tilted plate, which produces a magnetic field gradient in the vertical direction.

5. A circular hull, with interior cabin, electrical power supply to drive the coils, and a computer to calculate and sequence the activation of the electrical bobbins.

6. Magnetic sensors, in the form of small coils, located on the toroidal core which can detect the velocity and position of the traveling magnetic wave as feedback to the computer control system.

7. A magnetic vortex generator, either rotating magnet or dual coil, to produce a wormhole through which low linear mass and low speed of light hyperspace energy can enter the hull in order to create highly relativistic electromagnetic fields.

8. A variable current generator to modulate the direct current in the coil so as to produce a variable magnetic moment that can control lift, hover and descent.

9. An electrical power supply using mechanical flywheel stored energy together with a large area capacitor, resupplied with energy by solar cells located on the outer surface of the hull.

* * * * *

US 20040164824A1

(19) **United States**
(12) **Patent Application Publication**
St. Clair
(10) **Pub. No.: US 2004/0164824 A1**
(43) **Pub. Date: Aug. 26, 2004**

(54) HYPERSPACE ENERGY GENERATOR

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
John St. Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/372,556**

(22) Filed: **Feb. 21, 2003**

Publication Classification

(51) **Int. Cl.^{7}** **H01P 3/06**
(52) **U.S. Cl.** **333/243**

(57) ABSTRACT

This invention is a braided gold wire coaxial cable of micron size which generates hyperspace energy by coupling to the tetrahedral geometry of subspace, dimension and the Planck mass.

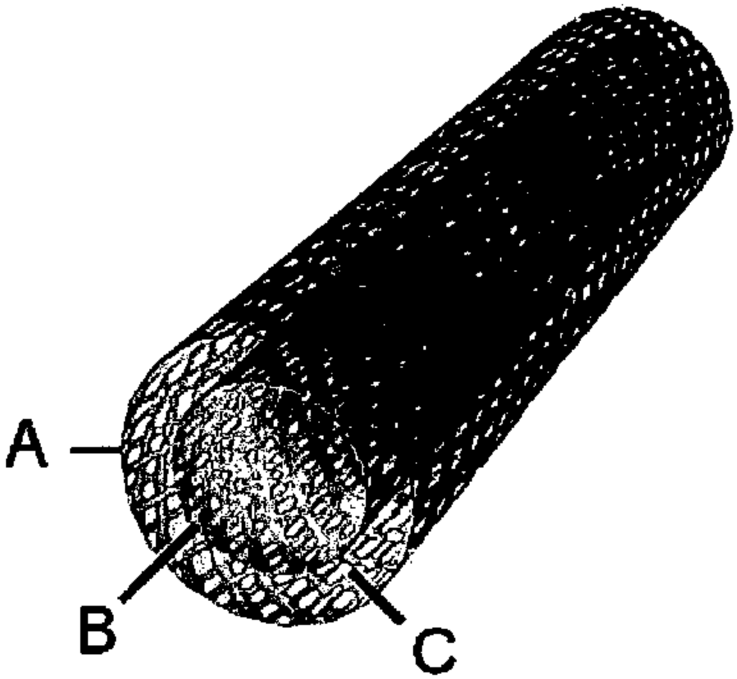

Figure 1

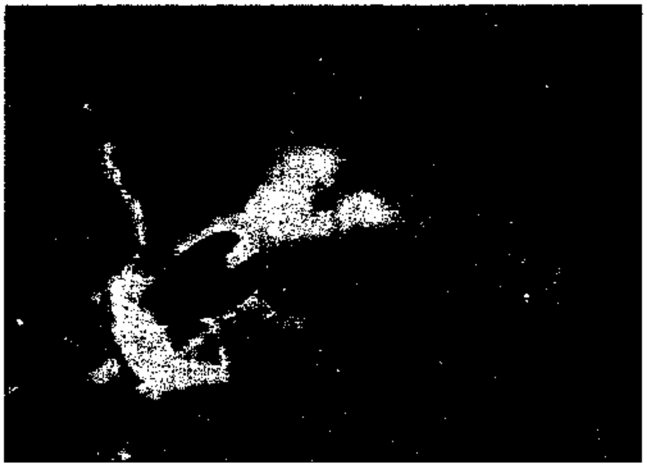

Figure 2

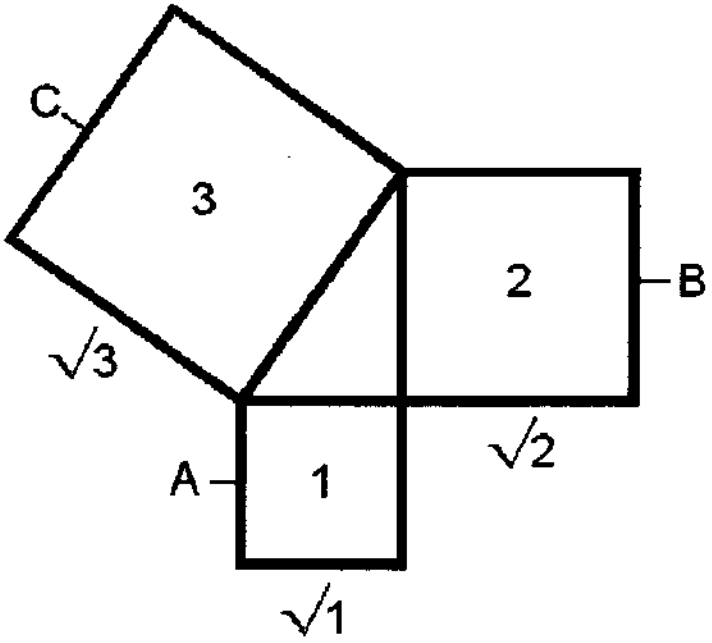

Figure 3

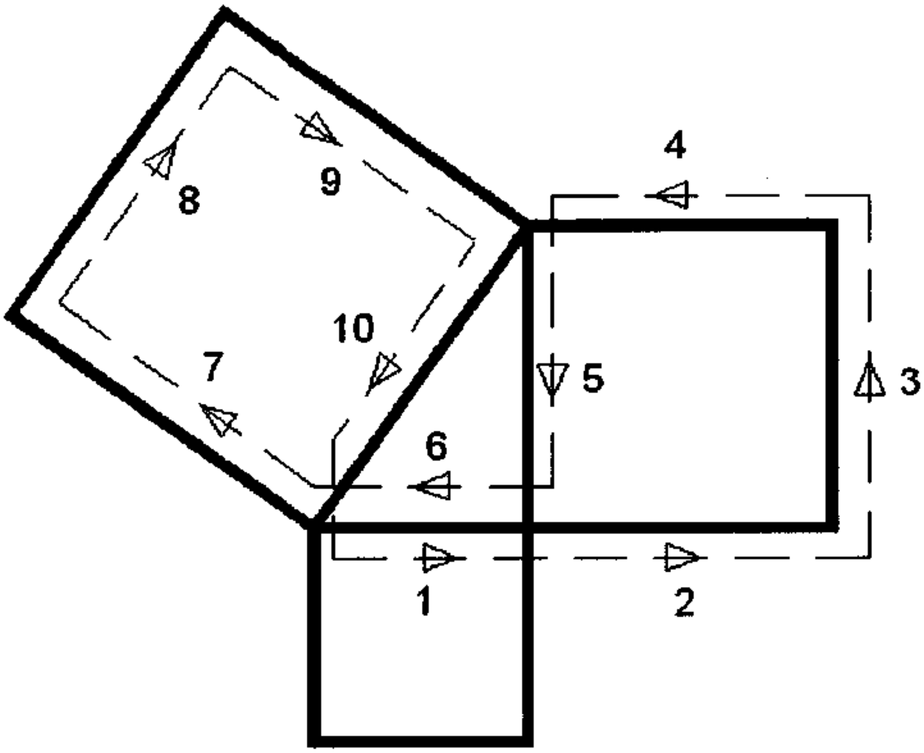
4
8
9
10
7
5
3
6
1
2

Figure 4

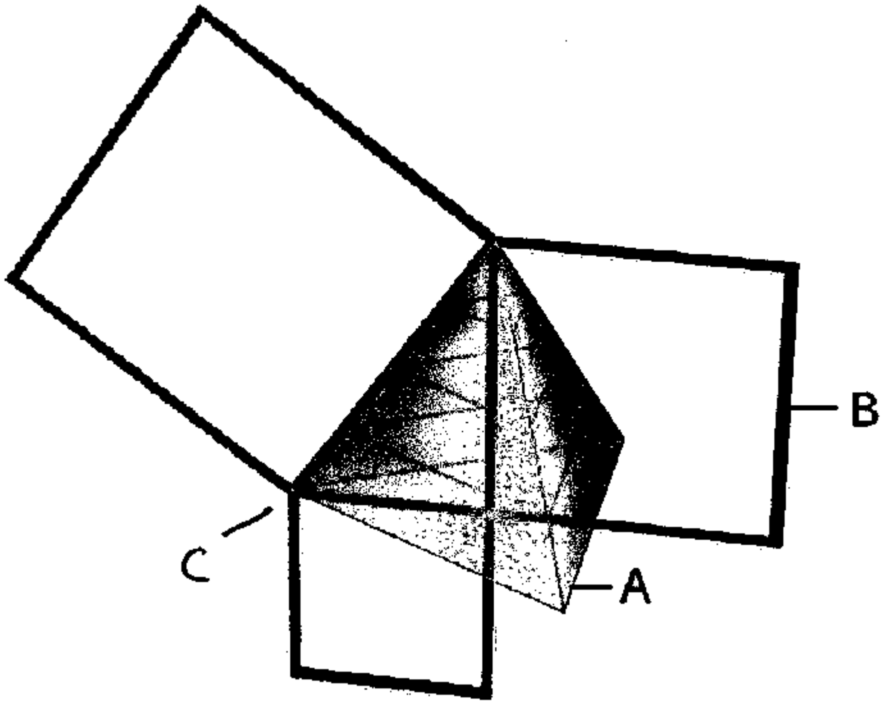

Figure 5

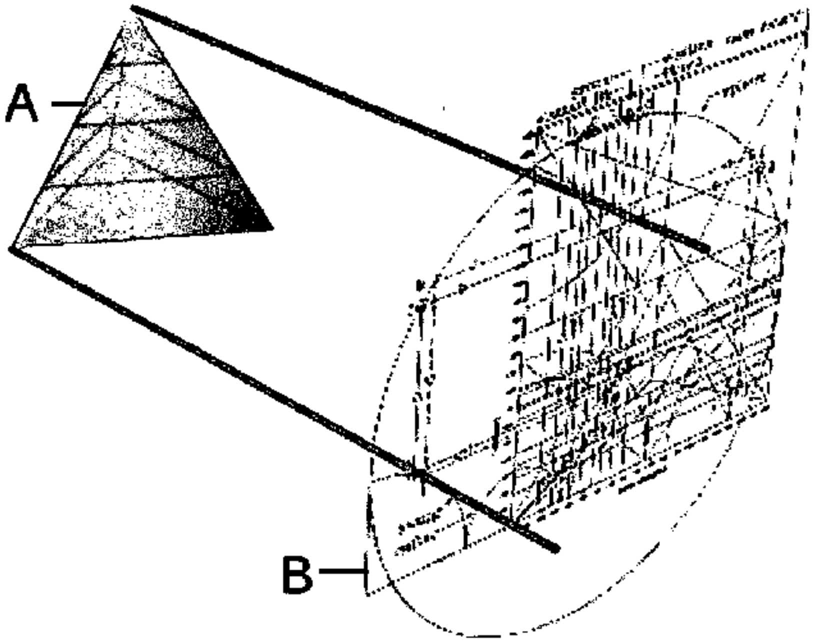

Figure 6

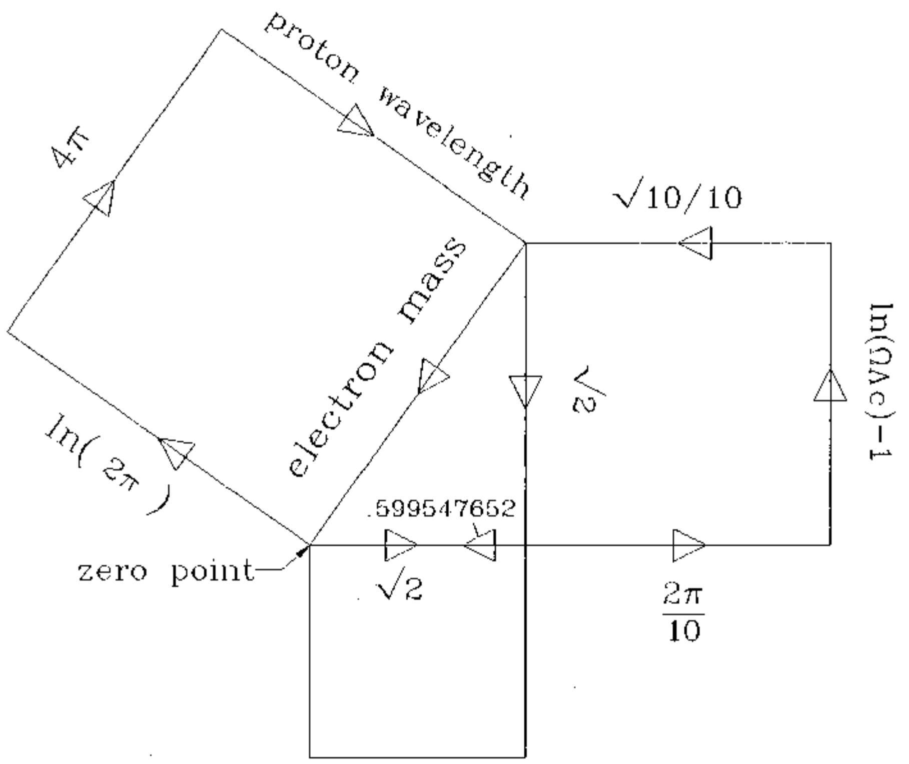
proton wavelength
4π
√10/10
electron mass
ln(ΩΛc)−1
√2
ln(2π)
.599547652
zero point
√2
2π/10

Figure 7

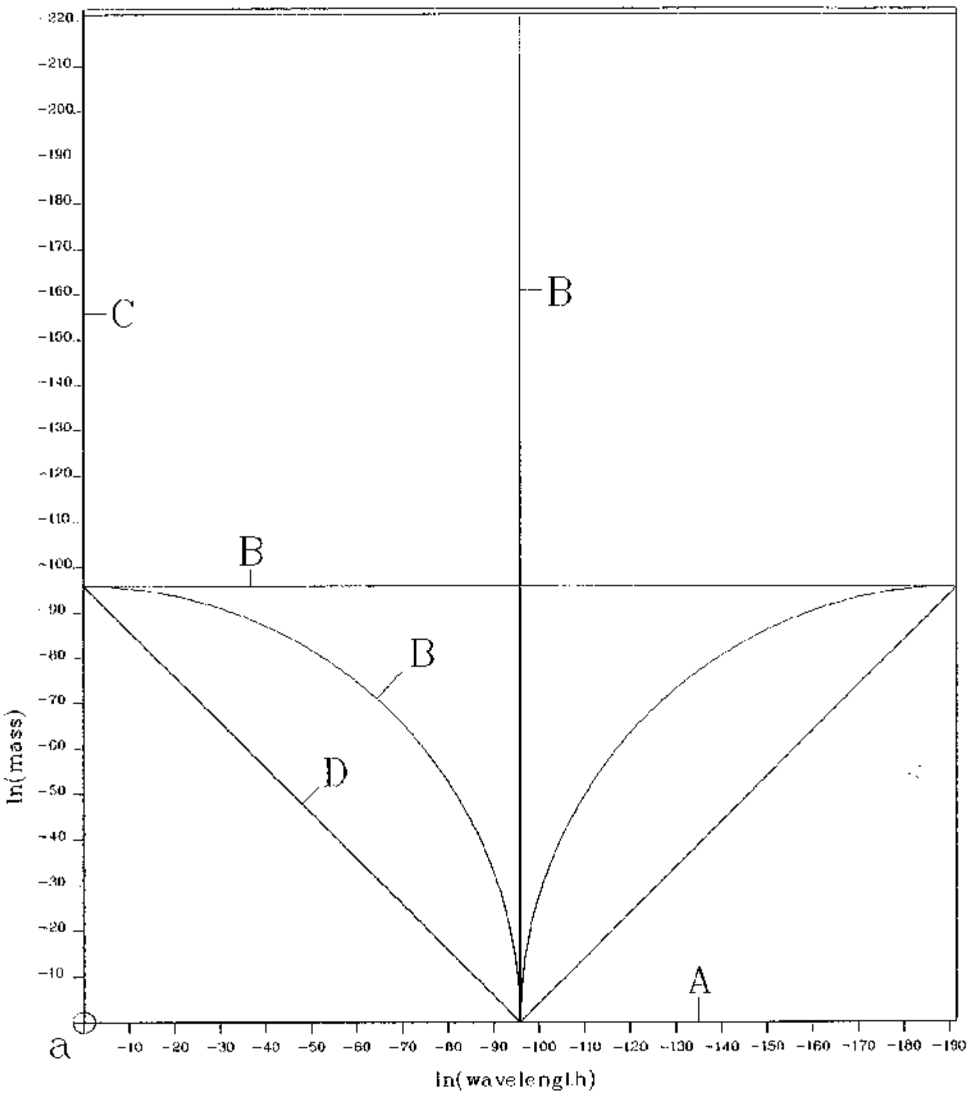
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
C
B
B
B
D
A
a
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 8

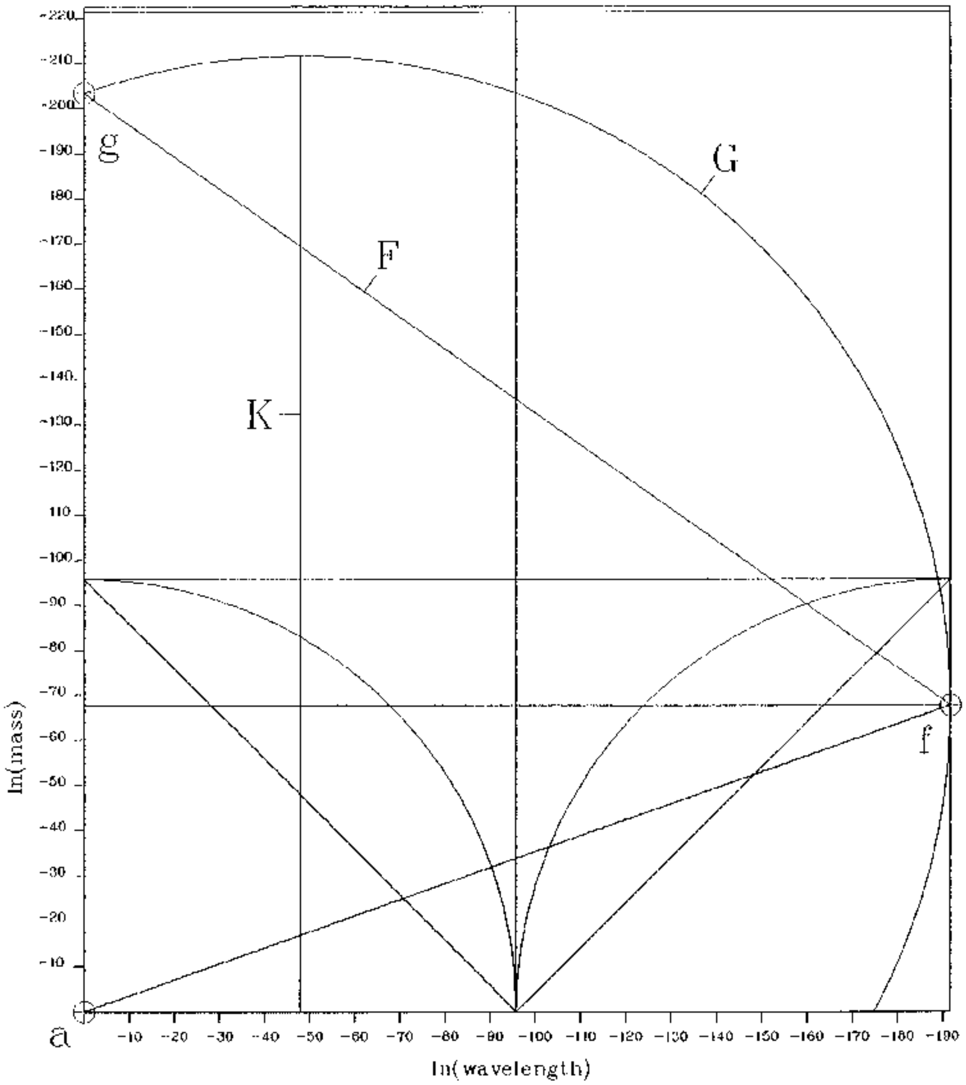
g
G
F
K
f
a
ln(mass)
ln(wavelength)
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190

Figure 9

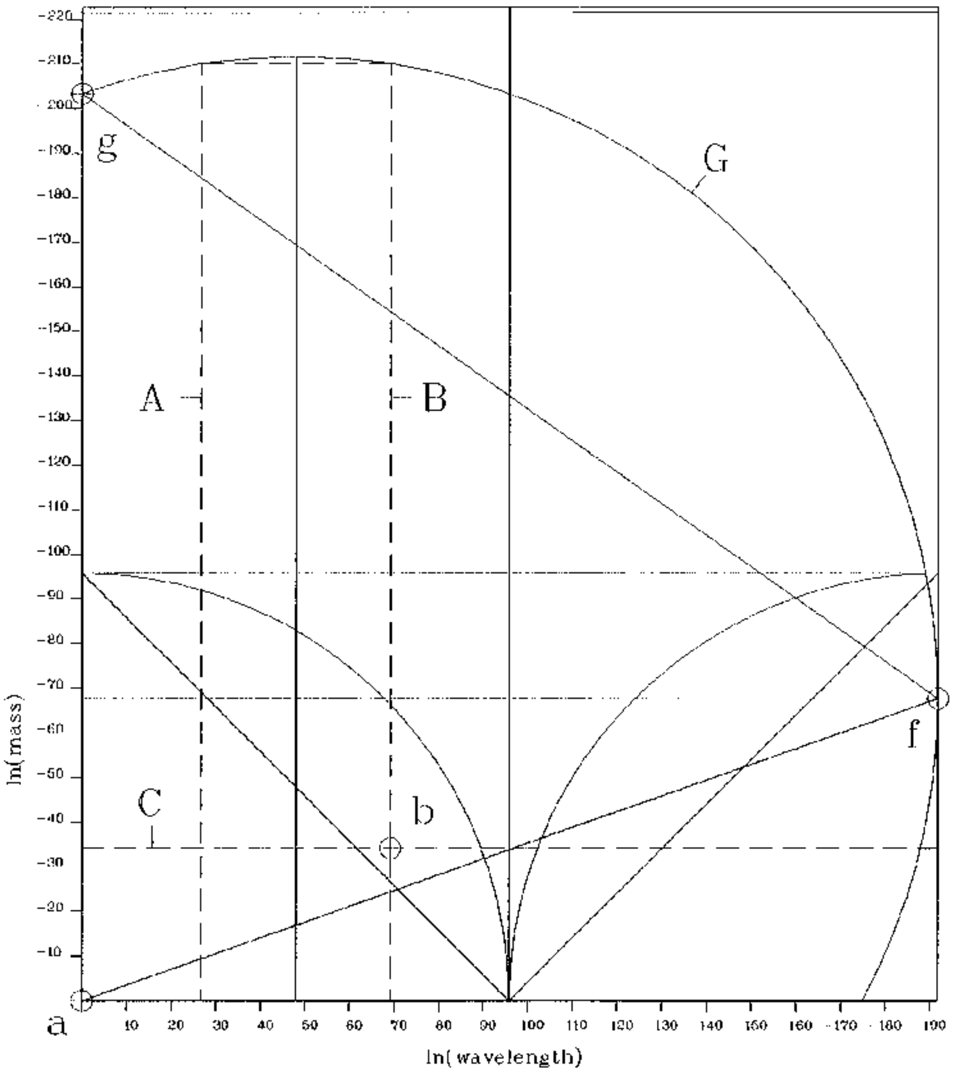
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
g
G
A
B
C
b
f
a
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
ln(wavelength)

Figure 10

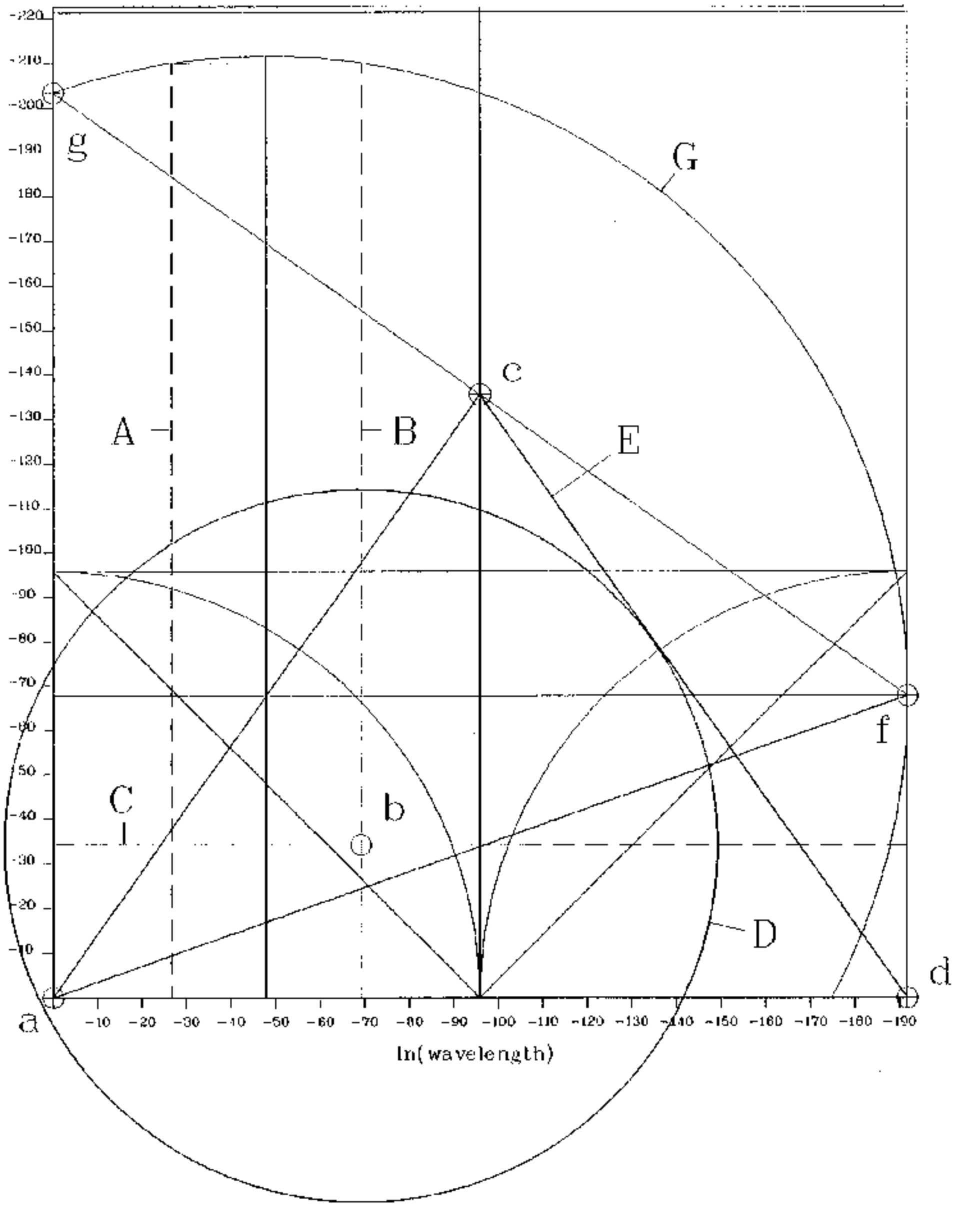
-220
-210
-200
-190
180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
g
G
c
A
B
E
C
b
D
f
d
a
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 11

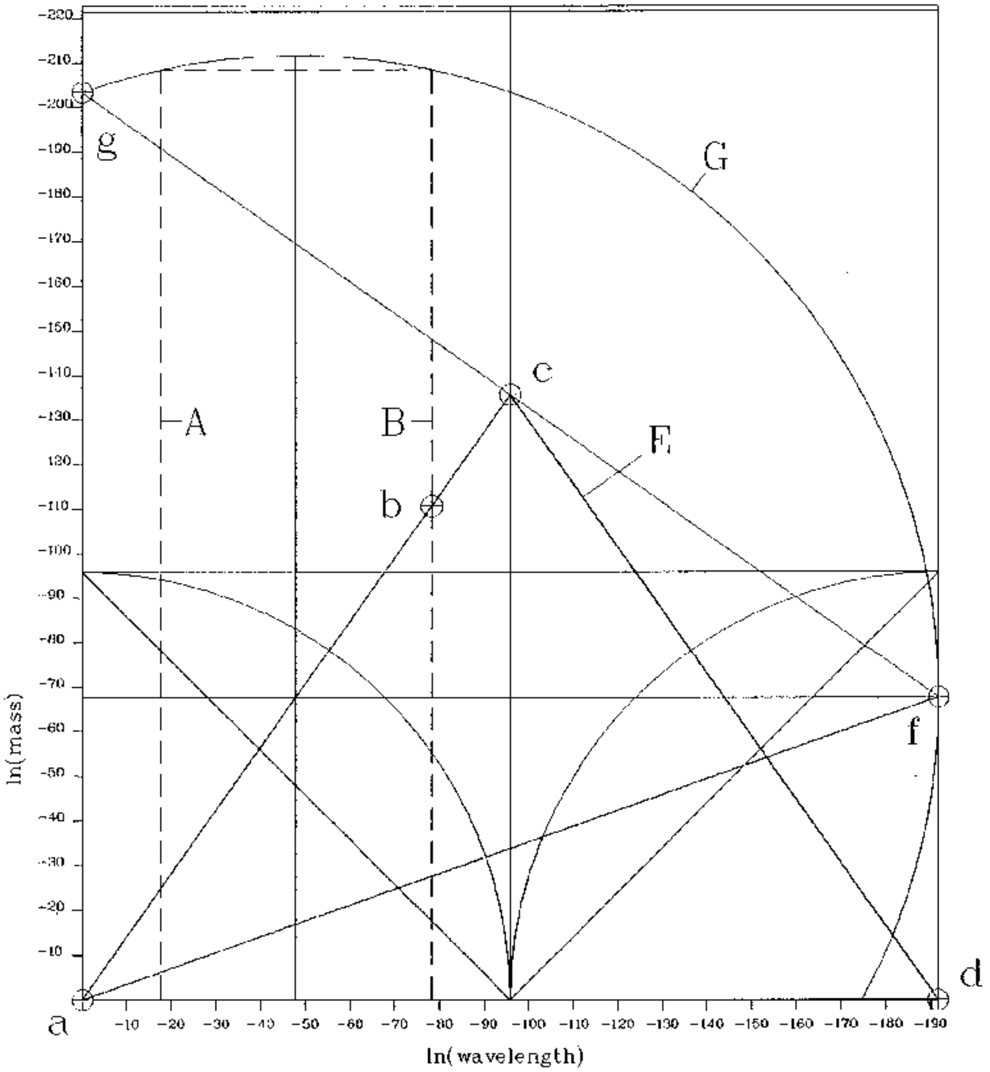
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
g
G
c
A
B
E
b
f
d
a

Figure 12

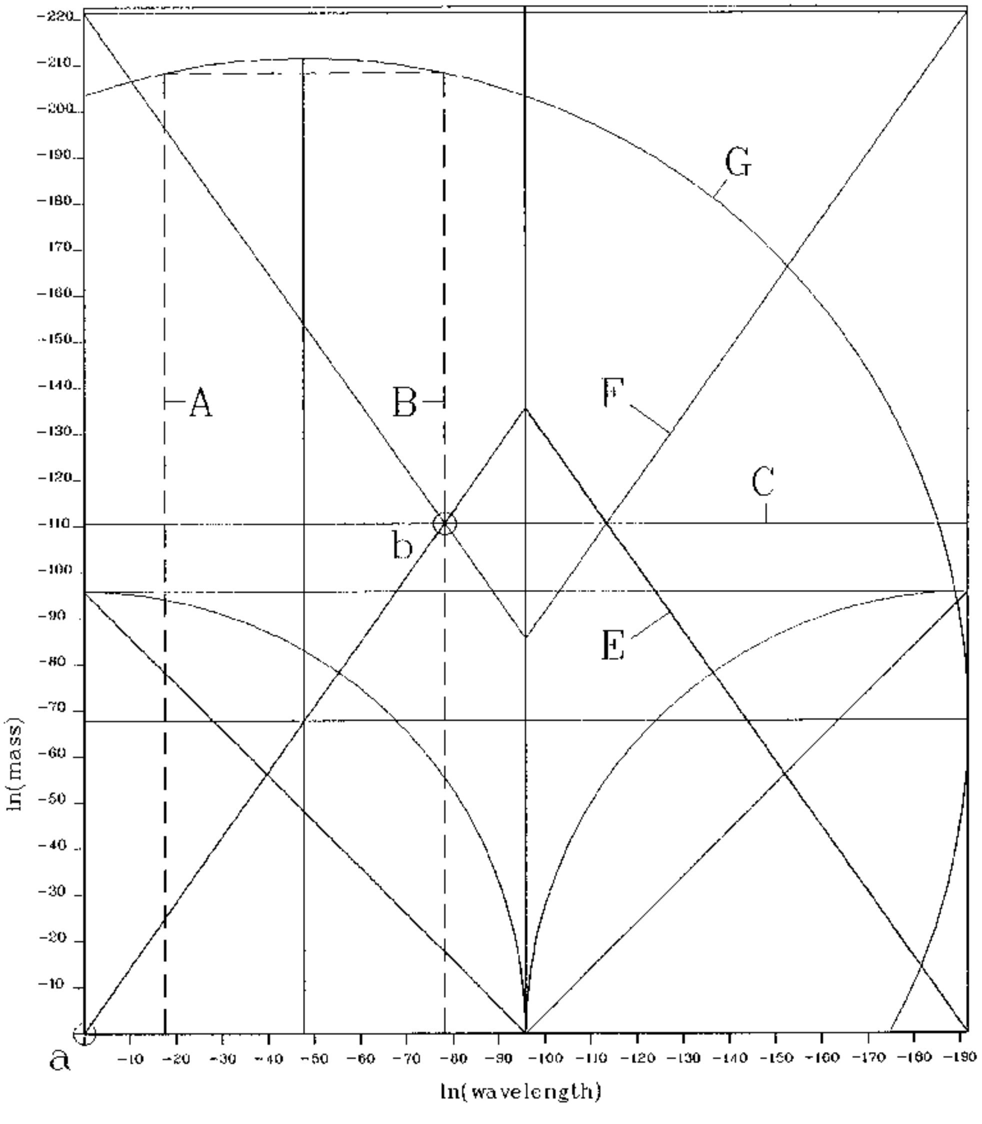
-220
-210
-200
-190
-180
170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
A
B
C
E
F
G
a
b

Figure 13

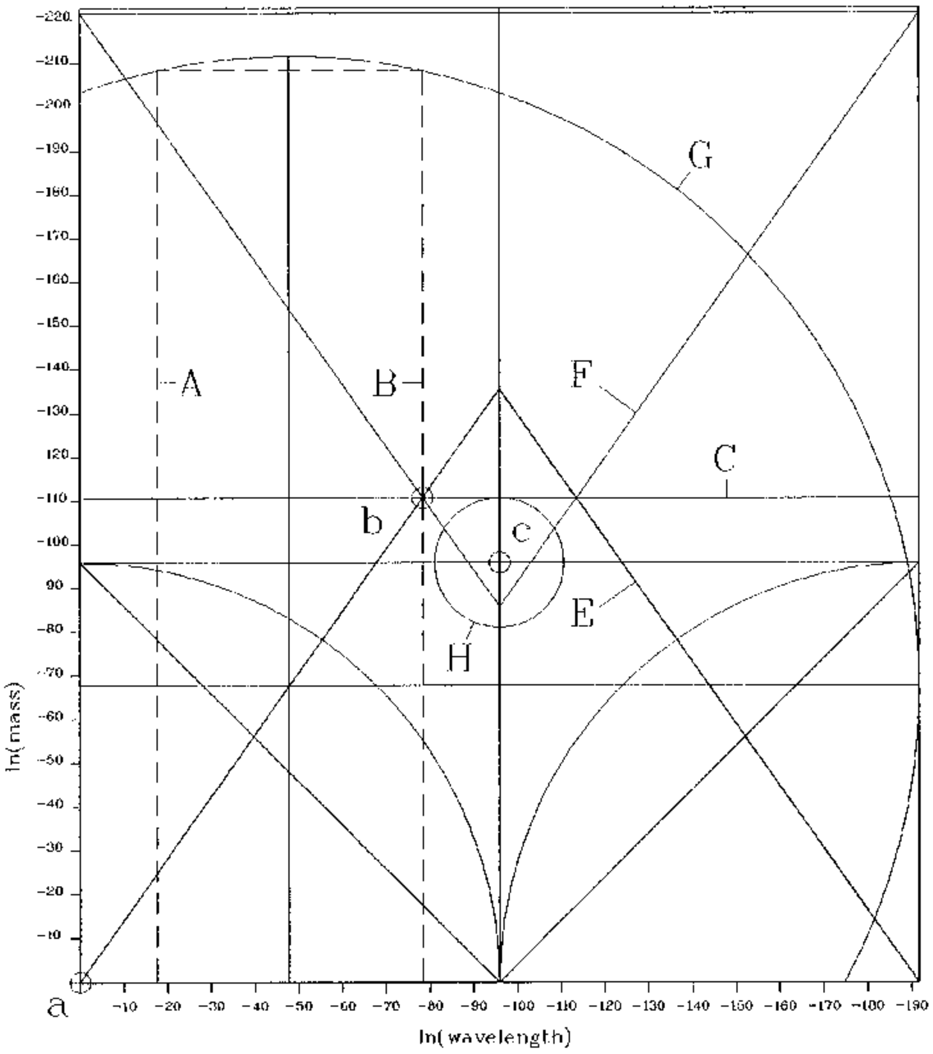
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
120
-110
-100
90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
A
B
C
E
F
G
H
a
b
c

Figure 14

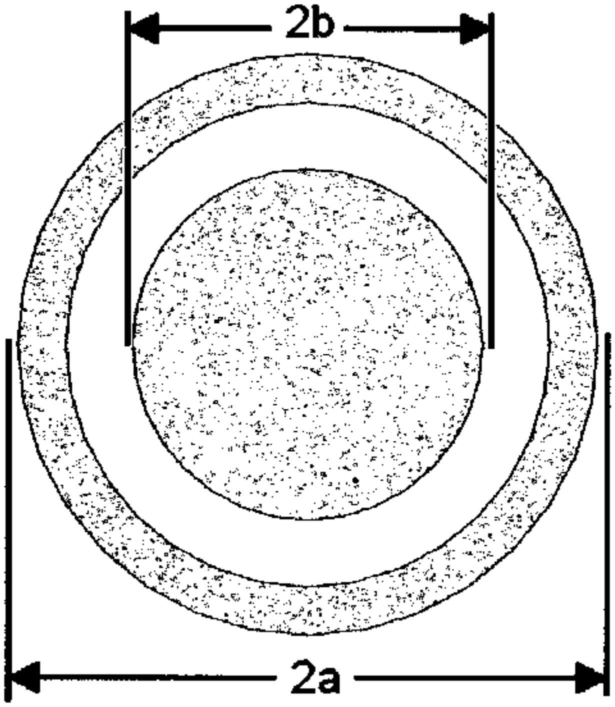

Figure 15

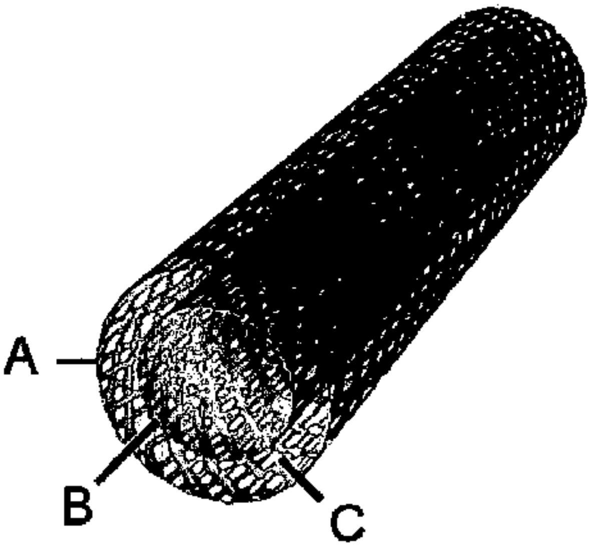

HYPERSPACE ENERGY GENERATOR

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a braided gold wire coaxial cable of micron size which generates hyperspace energy.

BACKGROUND OF THE INVENTION

[0002] Electrical experiments with micron-sized braided gold wire coaxial cable show that it is capable of generating substantial amounts of hyperspace energy. Referring to the electron microscope photograph shown in **FIG. 1**, the white mist emanating from the cable is low-density hyperspace energy that is flowing in from a co-dimension of our universe. The dimensions of the cable are of such a particular size as to couple the cable to the tetrahedral geometry of subspace, the dimension of space, the Planck mass and the linear inductance of the universe.

[0003] According to physicist Dr. Edward Witten of Princeton University, space has twenty-four dimensions, of which ten dimensions are non-redundant. Imagine taking a path around the Pythagorean triangle, as known as the planar tetrahedron, with sides equal to $\{\sqrt{1}, \sqrt{2}, \sqrt{3}\}$ as shown in **FIG. 2**. There are three squares denoted the one-square (A), the two-square (B) and the three-square (C). Each square has four sides. The edge of each square can be traversed in two directions. Thus the total number of dimensions is

$$dim_{total}=2\cdot(4+4+4)=24$$

[0004] Referring to **FIG. 3**, there is a path starting at the corner of the triangle, along the one-square **(1)**, around the two-square **(2,3,4,5)**, back along the one-square **(6)**, around the three-square **(7,8,9,10)** and back to the corner of the triangle. The numbering of the edges shows that there are ten edges. Because the path is traversed in only one direction, the number of reduced dimensions is

$$dim_{reduced}=1\cdot(10)=10$$

[0005] Referring to **FIG. 4**, the planar tetrahedron (B) forms one edge of the three-dimensional tetrahedron (A). Rotating the planar tetrahedron ±120° produces the other two edges. The tetrahedron has four faces which are equilateral triangles. The ten dimensional path starts and ends at (C), the corner of the tetrahedron known as the zero point.

[0006] Referring to **FIG. 5**, the projection of the 3D tetrahedron (A) onto a plane is called the tetrahedron diagram (B) which is the main diagram of the new geometrical physics known as Δphysics. All the constants of physics can be derived geometrically from the tetrahedron diagram and its associated planar tetrahedron. An example of this is shown in **FIG. 6** where the edges are given specific constants related to tetrahedral geometry, dimension, curvature, and the mass and wavelength of the elementary particles such as the electron and proton. The ten dimensional path includes the following constants

[0007] a. electron wavelength

[0008] b. proton wavelength

[0009] c. 47π The solid angle of the sphere. The tetrahedron is circumscribed by a sphere.

[0010] d. ln(2π) The natural log of the curvature. The subspace geometry is a logarithmic manifold. The tetrahedron diagram plots the logarithm of mass versus the logarithm of wavelength.

[0011] e. 0.599547652 A constant related to fractal dimension and the speed of light factor 2.99792458.

[0012] f. $\sqrt{2}$ The edge of the two-square.

[0013] g. $\sqrt{10}/10$ The square root of ten dimensions per 10 dimensions.

[0014] h. ln(ΩΛc)−1 The natural log of the momentum of space less one.

[0015] i.

$$\frac{2\pi}{10}$$

[0016] The curvature of space per 10 dimensions.

[0017] j. $\sqrt{2}$ The edge of the two-square.

[0018] The length of each edge is multiplied by the constant assigned to that edge. The ten edges have the order of $\{\sqrt{1}, \sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{1}, \sqrt{3}, \sqrt{3}, \sqrt{3}, \sqrt{3}\}$. What subspace geometry does is to multiply the edge length, such as $\sqrt{1}$, times the square root of two $\sqrt{2}$. Then it takes the square root of that number and multiplies it by the next edge, which is $\sqrt{2}$, times the curvature per 10 dimensions

$$\frac{2\pi}{10}.$$

[0019] It then takes the square root of that number and so on. In equation form, this looks like the following calculation

$$\sqrt{3a\sqrt{3b\sqrt{3c\sqrt{3d\sqrt{1e\sqrt{2f\sqrt{2g\sqrt{2h\sqrt{2i\sqrt{1j}}}}}}}}}}=80.1104395$$

[0020] where the letters correspond to those in the list of constants. The square root sum total is equal to the Planck scale Λwhich is the bottom dimensional limit of the universe. The sum of the ten constants per a speed of light circumference is equal to unity

$$\frac{\sum\limits_{n=10} s_n}{2\pi\ln(c)}=1.000000000$$

[0021] where the log of the speed of light is

$$\ln(c)=\ln(299792458)=19.51860099$$

[0022] and multiplying by 2π is the circumference of a circle with a radius equal to the speed of light.

[0023] Referring to **FIG. 7**, the tetrahedron diagram plots the natural logarithm of mass on the vertical axis (C) versus the natural logarithm of wavelength on the horizontal axis (A). The reason for this is that the mass of the electron times its wavelength is equal to the mass of the proton times its wavelength which in turn is equal to Planck's constant h divided by the speed of light, known as the base constant (B). If two numbers multiply, they sum in logarithms. In subspace geometry, the sum of the logarithm of the mass of the electron plus the logarithm of the wavelength is equal to the logarithm of the base constant which has a value of −95.91546344.

$$\ln(m_{electron}) + \ln(\lambda_{electron}) = \ln\left(\frac{h}{c}\right) = -95.91546344$$

[0024] What this means is that the mass and wavelength slide on a 45° base line (D) which has end points on the vertical and horizontal axes equal to the base constant.

[0025] Referring to **FIG. 8, a** line (af) drawn from the origin at the tetrahedral angle of 19.47122063, equal to the asin(⅓), creates a tetrahedron (F) along path (afg). This tetrahedron is circumscribed by a sphere (G) with sphere diameter (K).

[0026] The Planck scale path calculation showed that the electron mass and the proton wavelength were the last two edges. The electron mass has a value of

$$\ln(m_e) = -69.17083217$$

[0027] and the electron wavelength has a value of

$$\ln(\lambda_E) = -26.74463127$$

[0028] Referring to **FIG. 9**, the electron wavelength (A) is plotted as a vertical line on the tetrahedron diagram. The wavelength reflects off the circumscribing sphere (G), and returns as the electron mass (B). So the diagram incorporates the concepts of both classical physics (point mass particles) and quantum mechanics (wave particles).

[0029] The proton wavelength has a value of

$$\ln(\lambda_p) = -34.26005901$$

[0030] The proton wavelength (C) is plotted as a horizontal line in order to get the intersection (b) with the electron mass.

[0031] Referring to **FIG. 10, a** circle (D) with a radius equal to the Planck scale is drawn centered (b) on the intersection of the electron mass with the proton wavelength, which are the last two edges of the Planck scale calculation. A line (ac) from the origin to the intersection of the base constant with the rotated tetrahedron creates the vertical tetrahedron (acd). As can be seen, the Planck scale is tangent to the tetrahedron on side (cd). This tetrahedron is the projection of the 3D tetrahedron shown before in **FIG. 5**. What this means is that the tetrahedral geometry of subspace determines the bottom limit of our universe. And this bottom limit, called the Planck scale, contains within itself the mass and wavelength of the elementary particles, curvature, dimension and planar tetrahedral geometry. Tetrahedron diagram tet0565 shows that the electron and proton are one and the same particle because the electron path rotates counterclockwise around the curvature and then returns clockwise as the proton path. This path occurs moving through space and hyperspace. Because the single particle enters our universe from hyperspace at two different positions, we see it as two distinct particles. Thus the tetrahedron diagram shows that hyperspace exists.

[0032] The speed of light is equal to the inverse of the square root of the permeability μ of space times the permittivity ϵ of space

$$c = \sqrt{\frac{1}{\mu\varepsilon}}$$

[0033] The permeability is linear inductance or inductance per length which you would find in a solenoid for example. The permittivity is linear capacitance or capacitance per length which you find in a capacitor. In an electrical circuit, the inductance and capacitance form a resonant circuit. The resonance frequency can be changed by changing the inductor or capacitor. In a similar manner, the speed of light is not constant, but can be lowered by increasing the permittivity. Hyperspace energy has a high permittivity and therefore a low speed of light. This low speed of light gives hyperspace energy a luminescent quality which is seen as a white mist **(FIG. 1)**.

[0034] From Einstein's General Theory of Relativity, the stress pressure T on spacetime is proportion to the square of the ratio of the electric field E to the speed of light c.

$$T = \left(\frac{E}{c}\right)^2$$

[0035] Thus substantially lowering the speed of light creates an enormous spacetime pressure which can be used to generate the lift force on electromagnetic field propulsion vehicles. Furthermore, the electric field is subject to the Lorentz transformation

$$E = \frac{E_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

[0036] The electric field E_0 moving in a frame velocity of v, can quickly attain relativistic proportions because the speed of light could be 1 meter per second, rather than the enormous value in our universe of 299792458 meters per second. Thus one would like to permeate the hull of the electromagnetic field propulsion vehicle with this hyperspace energy in order to increase the electric field and hence the spacetime curvature around the hull which produces the enormous lift force on the vehicle. The method of bringing in this hyperspace energy is to use braided gold wire coaxial cable which is coupled to the geometry of subspace. The subspace geometry is contained in the Δphysics tetrahedron diagram.

[0037] Just as space has a linear inductance and linear capacitance, it also has a linear mass Ω or mass per meter.

Physicist Dr. John A. Wheeler of Princeton likes to invert this and call it "mom" for meter of mass. The Planck mass is equal to the Planck scale Λtimes the linear mass Ω

$$\ln(m_{\text{Plank}})=\ln(\Omega\Lambda)=-17.64290101$$

[0038] Planck's constant h is equal to 2π times the Planck scale squared times the linear mass Ω times the speed of light c.

$$h=2\pi\Lambda^2\,\Omega c=2\pi\Lambda(\Omega\Lambda)c$$

[0039] which shows that Planck's constant is actually the circumference of a circle of radius Planck scale times the Planck mass times the speed of light. The base constant is therefore

$$\text{base}=\frac{h}{c}=\frac{2\pi\Lambda(\Omega\Lambda)c}{c}=(2\pi\Lambda)(\Omega\Lambda)=-95.91546344$$

[0040] which is an area, known as the Planck box, bounded by the Planck wavelength (2πΛ) and the Planck mass. Everything outside the Planck box is hyperspace. Everything inside the Planck box is our universe. Thus the boundary between space and hyperspace is the Planck wavelength and the Planck mass. In logarithms, notice that the Planck mass and Planck wavelength, just like the electron and proton, sum to the base constant.

[0041] Referring to **FIG. 11**, the Planck mass (A) and the Planck wavelength (B) are plotted on the diagram and reflected off the sphere. The Planck wavelength intersects the tetrahedron at (b) which is the boundary between space and hyperspace known as the centerline of the diagram. The centerline has a value equal to the base times the square root of 4/3.

$$\textit{centerline}=\sqrt{\frac{4}{3}}\,\textit{base}=-110.7536373$$

[0042] Referring to **FIG. 12**, the centerline (C) is drawn on the diagram and the tetrahedron (E) is mirrored (F) across the centerline to indicate the co-dimensions of hyperspace.

[0043] Referring to **FIG. 13, a** circle (H), centered at the base at the base (c), tangent to the centerline (C), has a radius equal to base times the square root of 4/3 less one

$$R=\left(\sqrt{\frac{4}{3}}-1\right)\textit{base}=14.83817383$$

[0044] This is the length that has to traversed in order to cross over the centerline from the base constant of our universe to the co-dimension of hyperspace. Furthermore, the Planck mass, which is the other boundary, has to be crossed in order to get to either axis as seen by the length between the vertical axis and line (A). It can be looked at also as the length needed to go from the Planck wavelength (B) to the tangent point of circle (H) in order to reach the center of the mirror tetrahedrons.

SUMMARY OF THE INVENTION

[0045] This invention is a braided gold wire coaxial cable of micron size that is coupled to the subspace geometry of the universe for the purpose of bringing in low-density hyperspace energy into our universe from the co-dimensions of hyperspace. The dimensions of the coaxial cable are of such particular size as to enable it to couple to the ten dimensions of space, the 3:1 geometrical ratio of the tetrahedron, the coaxial wave function based on the logarithm of the ratio of the outer radius to the inner radius of the cable, the length between the base constant of our universe and the centerline between mirror tetrahedrons, and the geometrical relationship between the Planck mass and the linear inductance of the cable.

[0046] Referring to **FIG. 14**, the coaxial guide has an outer braided conductor **(A)** and an inner braided conductor **(B)** in which the outer and inner conductors are at radii a and b respectively. The linear inductance **L** of the cable is equal to the natural logarithm of the ratio of the radii times the permeability μ of space divided by the curvature 2π

$$L=\ln\left(\frac{a}{b}\right)\frac{\mu}{2\pi}$$

[0047] It can be shown that the ratio of the area-to-volume ratio of the tetrahedron to the area-to-volume ratio of its circumscribing sphere is 3:1. It is also the ratio of the area of the three-square to the area of the one-square on the planar tetrahedron. It is also the tetrahedral angle asin(⅓) equal to 19.47122063°. It is also a maximum work condition between the velocity ratio of a fluid stream and a moving vane such as in turbomachinery. So the first constraint on the radii is

$$\exp\left(\ln\left(\frac{a}{b}\right)+1\right)\equiv 3$$

[0048] The second constraint is that the radius c of the circle, equal to the difference between the base constant and the centerline, is related to ten dimensions. The value of the radius projected into our universe is

$$c=\left(\sqrt{\frac{4}{3}}-1\right)\text{base}=.3596358547\cdot\text{microns}$$

[0049] and the dimensional constraint is

$$\left[\ln(c)+\ln\left(\frac{1}{a}\right)\right]^2=10$$

[0050] The coaxial wave function constraint is

$$\frac{\ln(a)^2}{2\pi\ln\left(\frac{a}{b}\right)} \blacksquare \ln(c)^2$$

[0051] The linear inductance of the coaxial cable has to be such that it gets geometrically across the Planck mass which is the second boundary of our Planck box

$$\exp\left(\ln(\Omega\Lambda) - \ln\left(\frac{\ln\left(\frac{a}{b}\right)}{2\pi}\mu\right) + 1\right) \equiv 3$$

[0052] With these four subspace constraints, the outer radius a and inner radius b of the cable are

[0053] a=8.342461828 microns

[0054] b=7.559058141 microns

[0055] so the cable has an outside diameter of roughly 16 to 17 microns.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0056] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0057] **FIG. 1**. Electron microscope photograph of the coaxial cable and hyperspace mist.

[0058] **FIG. 2**. The planar tetrahedron.

[0059] **FIG. 3**. The ten dimensional path around the planar tetrahedron.

[0060] **FIG. 4**. The relationship between the planar tetrahedron and the 3D tetrahedron.

[0061] **FIG. 5**. Projection of the 3D tetrahedron onto a plane to create the tetrahedron diagram.

[0062] **FIG. 6**. Dimension, geometry, curvature and elementary particles that make up the ten dimensional Planck scale path.

[0063] **FIG. 7**. Tetrahedron diagram showing base constant.

[0064] **FIG. 8**. Tetrahedron diagram showing rotated tetrahedron and circumscribing sphere with electron mass and wavelength reflecting off sphere.

[0065] **FIG. 9**. Tetrahedron diagram showing intersection of electron mass with proton wavelength which are components of the Planck scale path.

[0066] **FIG. 10**. Tetrahedron diagram showing Planck scale tangent to tetrahedron.

[0067] **FIG. 11**. Tetrahedron diagram showing Planck mass and Planck wavelength intersecting tetrahedron.

[0068] **FIG. 12**. Tetrahedron diagram showing mirror tetrahedrons across centerline, the boundary between space and hyperspace.

[0069] **FIG. 13**. Tetrahedron diagram showing distance between the base constant and the centerline used to calculate the dimensions of the coaxial cable.

[0070] **FIG. 14**. End view of coaxial cable showing radii a and b used in the subspace geometry constraints in order to couple to hyperspace.

[0071] **FIG. 15**. Perspective view of braided gold wire coaxial cable.

DETAILED DESCRIPTION OF THE INVENTION

[0072] 1. Referring to **FIG. 15**, the coaxial cable has a braided gold wire outer conductor (A) and a braided gold wire inner conductor (B) separated by a dielectric (C). The open braiding promotes the conduction of the electromagnetic wave while allowing the hyperspace mist to seep out of the braid and permeate the surrounding material in which it is embedded.

[0073] 2. The radius of the outer conductor a and the radius of the inner conductor b have the following values in order to couple the cable to the tetrahedral geometry of subspace.

[0074] a=8.342461828·microns

[0075] b=7.559058141·microns

[0076] 3. Even though the wire size is very small, the cable can be made in limited lengths using the new nanotechnology and silicon micromotors.

I claim:

1. A coaxial cable which has:

a) a braided gold wire outer conductor of radius 8.34 microns;

b) a braided gold wire inner conductor of radius 7.56 microns;

c) a thin dielectric separator between the two conductors; and

d) an open weave to allow the hyperspace mist to seep out of the cable and permeate the surrounding material in which the cable is embedded;

2. A specific relationship between the physical dimensions of the coaxial cable, given in items (**1***a*) and (**1***b*), to the following tetrahedral subspace couplings:

a) the ratio of the area-to-volume ratio of the tetrahedron to the area-to-volume ratio of its circumscribing sphere, equal to 3:1, with a coupling to the natural logarithm of the ratio of the radii of the conductors;

b) the ratio of the area of the three-square of the planar tetrahedron to the area of the one-square, equal to 3:1, with a coupling to the Planck mass and the linear inductance of the cable;

b) the distance between the base constant and $\sqrt{4/3}$ times the base constant, equal to −14.83817383 in natural logarithms, with a coupling to the outer radius of the conductor, the ten dimensions of space, and the co-dimensions of hyperspace; and

c) the wave function of the coaxial guide given in terms of item (2*c*), the natural logarithm of the ratio of the radius of the outer conductor to the radius of the inner conductor, and the curvature 2π.

* * * * *

US 20060070371A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2006/0070371 A1**

(43) **Pub. Date: Apr. 6, 2006**

(54) **ELECTRIC DIPOLE MOMENT PROPULSION SYSTEM**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/958,436**

(22) Filed: **Oct. 5, 2004**

Publication Classification

(51) **Int. Cl.**
F03H 1/00 (2006.01)
(52) **U.S. Cl.** **60/203.1**

(57) **ABSTRACT**

A spacecraft propulsion system utilizing a rotating octagon of trapezoidal electrically charged flat panels to create an electric dipole moment that generates lift on the hull.

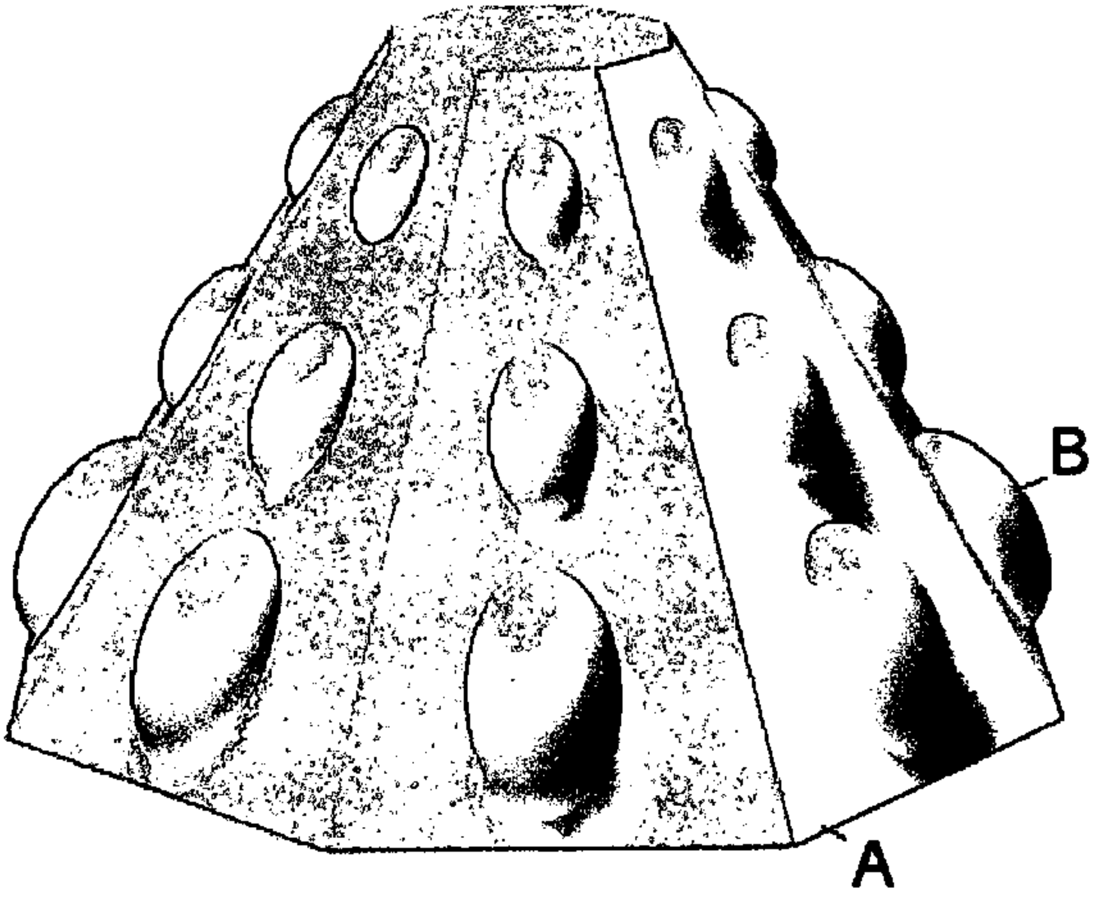

Figure 1

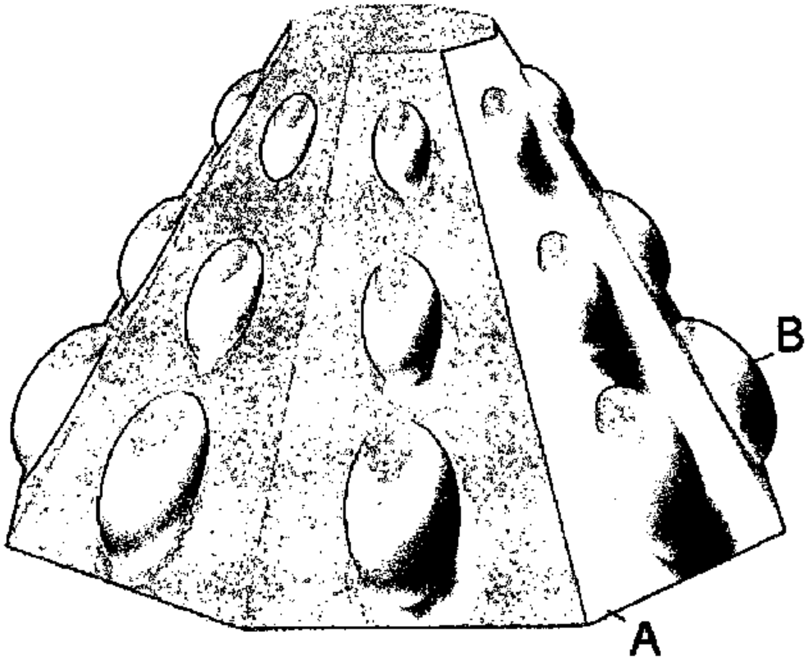

Figure 2

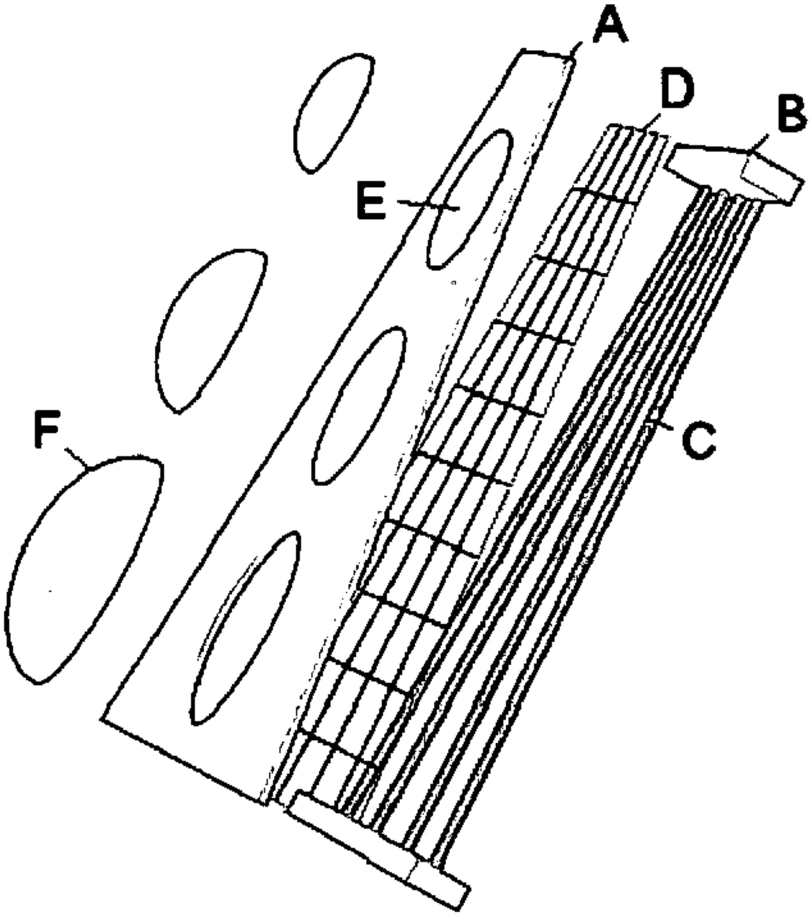

Figure 3

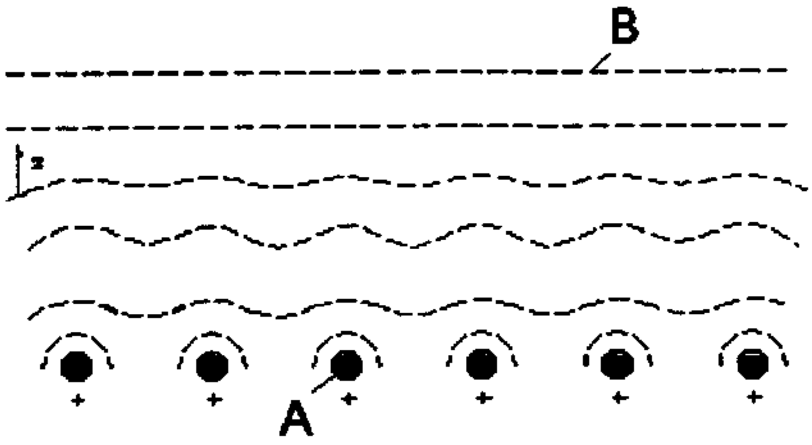

Figure 4

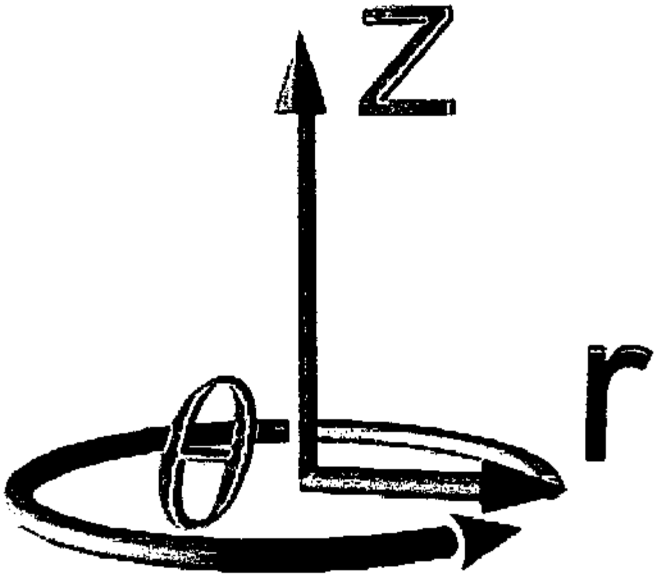

Figure 5

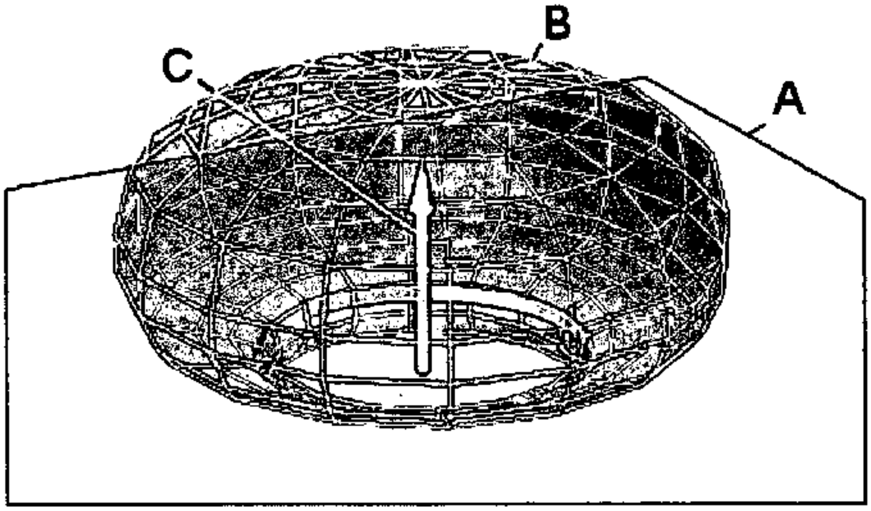

Figure 6

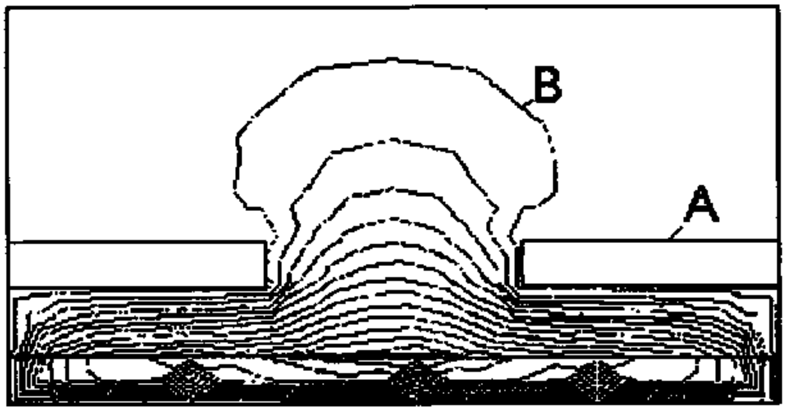

Figure 7

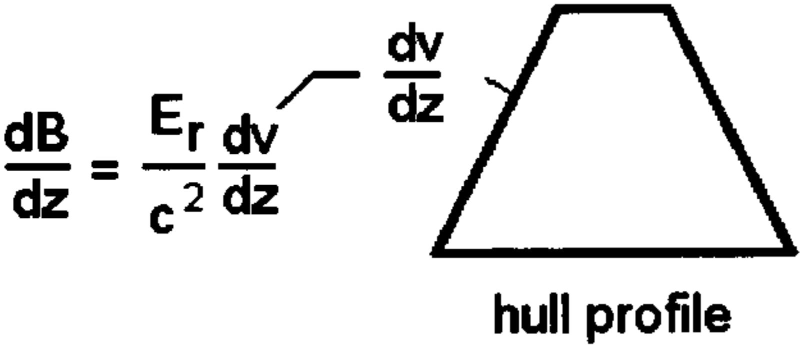
$\frac{dB}{dz} = \frac{E_r}{c^2}\frac{dv}{dz}$
$\frac{dv}{dz}$
hull profile

Figure 8

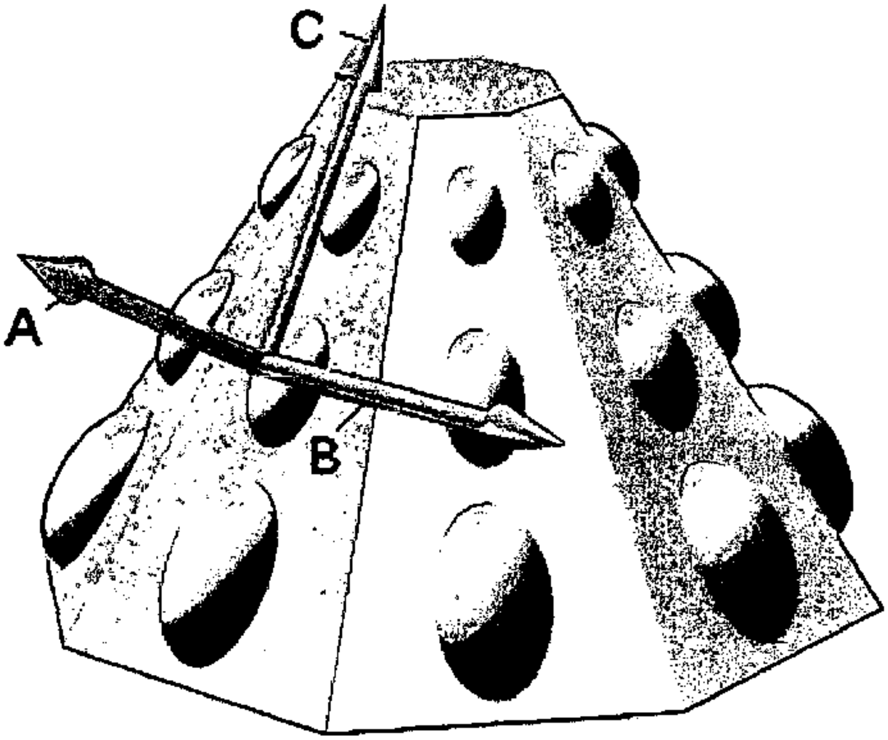

Figure 9

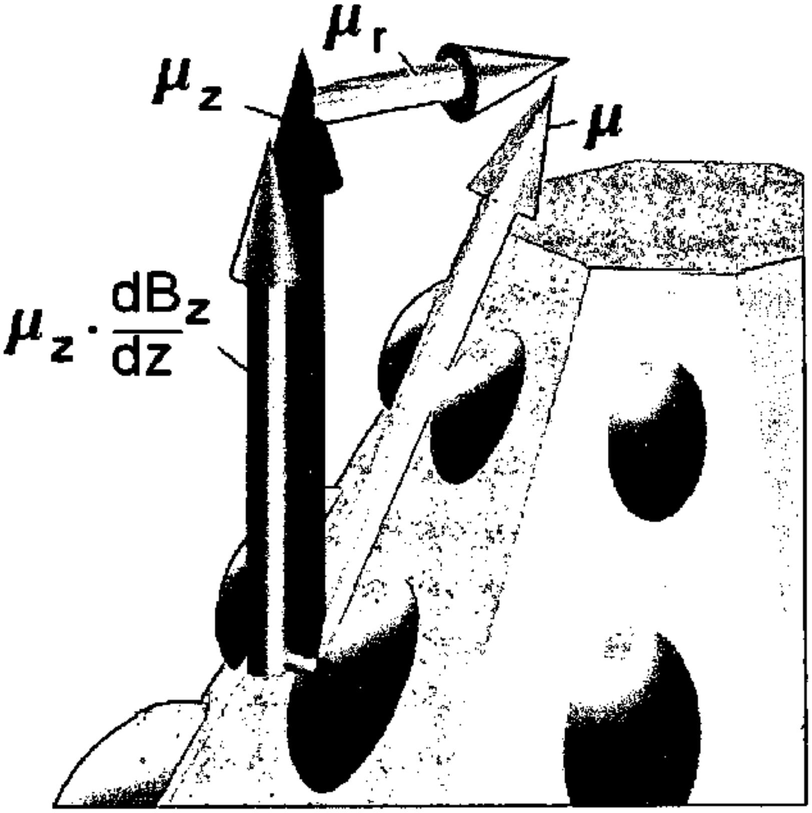
μ_r
μ_z
μ
$\mu_z \cdot \frac{dB_z}{dz}$

Figure 10

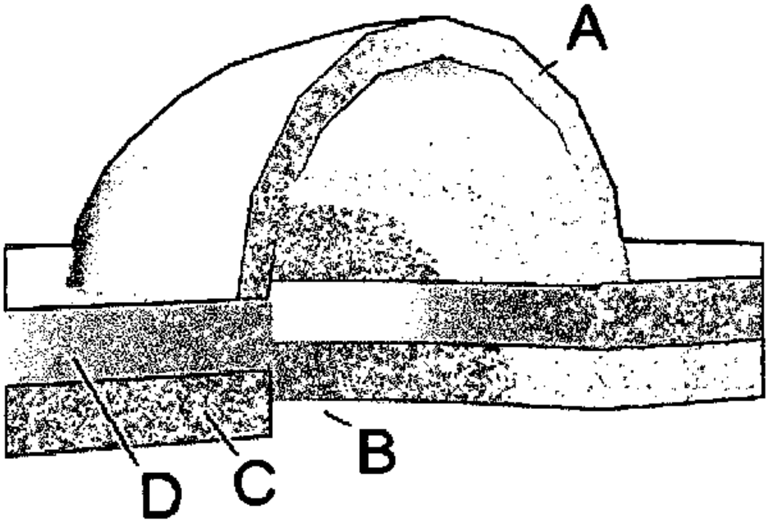

Figure 11

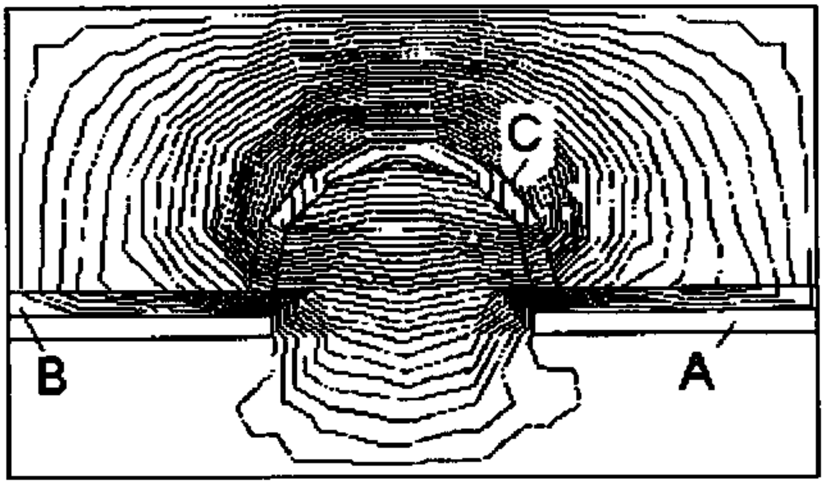

ELECTRIC DIPOLE MOMENT PROPULSION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] The invention is a spacecraft utilizing trapezoidal electrostatically charged flat plate panels which form a pyramidal hull. A panel contains three holes each of which produces a potential energy ellipsoidal bubble that creates an electric dipole moment. The rotation of the hull generates a magnetic moment and a magnetic field gradient in the vertical direction that produces a lift force on the spacecraft.

BACKGROUND OF THE INVENTION

[0002] It is known from electrodynamics that a hole in a conducting plane forms a potential energy bubble. This bubble creates an electric dipole moment from which it is possible to develop a magnetic moment. A rotating tilted hull produces a velocity gradient that generates a magnetic field gradient in the vertical direction. This combination produces a lift force on the spacecraft. A very large potential energy bubble is produced provided that the hole protrudes out of the plate in an ellipsoidal shape. Furthermore, a double cladding, in which each layer around the hole has a different permittivity, confines the field to the outside of the hull for even better results.

[0003] The planar potential energy is created by a grid of electrically charged wires or rods running the length of each panel. A circular potential energy from each rod very quickly sums to form a flat sheet of energy which emerges from the hole to form the potential energy bubble.

SUMMARY OF THE INVENTION

[0004] The invention relates to a spacecraft utilizing a rotating octagon of trapezoidal electrically charged flat plate panels to form a hull in the shape of a pyramid. Each panel has three protruding ellipsoidal bubbles that produce an electric dipole moment from a planar potential energy field created by a group of charged rods parallel to the panel. Because the panels are tilted and the hull is rotating, there is a tangential velocity gradient in the vertical direction. This creates the magnetic moment. Because the hull rotates, the radial electric field produces a magnetic field gradient in the vertical direction. This combination of magnetic moment and magnetic field gradient produces a lift force on the hull of the spacecraft.

[0005] On the underside of each panel is a group of high voltage electrically charged rods which run parallel to the panel. These wires or rods produce a planar electrical potential field underneath the holes in the panel. This potential energy field then bubbles out of the holes in the panel to create a large ellipsoidal potential energy field above the hull. The potential energy bubble carries an electric dipole moment which when rotated with the hull generates a magnetic moment in the vertical direction.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0006] **FIG. 1**. Perspective view of electric dipole moment spacecraft.

[0007] **FIG. 2**. Perspective exploded view of one panel with the ellipsoidal domes, flat hull panel with three holes, the charged rod grid and the planar potential energy field.

[0008] **FIG. 3**. Planar view of flat potential energy field produced by electrically charge wire rods.

[0009] **FIG. 4**. Perspective view of cylindrical coordinates $\{r, \theta, z\}$.

[0010] **FIG. 5**. Perspective view of ellipsoidal potential energy field emerging from hole in plate which produces an electric dipole moment.

[0011] **FIG. 6**. Planar view of field lines of potential energy bubble emerging from plate hole.

[0012] **FIG. 7**. Planar view of sloping hull profile needed to get a velocity gradient.

[0013] **FIG. 8**. Perspective view of hull showing electric dipole moment, the tangential velocity of the hull, and the magnetic moment.

[0014] **FIG. 9**. Perspective view of the magnetic moment components in the radial and vertical direction whereby the lift force is generated by the dot product of the vertical magnetic moment with the magnetic field gradient.

[0015] **FIG. 10**. Perspective view of cross section of dome showing two layer cladding with different permittivities to enhance potential energy field.

[0016] **FIG. 11**. Planar view of enhanced potential energy field with two layer cladding.

DETAILED DESCRIPTION OF THE INVENTION

[0017] 1. Referring to **FIG. 1**, the spacecraft is a rotating octagon of trapezoidal electrostatically charged flat panels which form a closed hull (A). Each panel has three ellipsoidal domes (B) of varying size centrally located along the major length of the panel. The purpose of the dome is to create a large ellipsoidal potential energy bubble over the hull which develops an electrical dipole moment. Because the hull is rotating, a magnetic moment is created in the vertical direction. A magnetic field gradient created by the rotating electric field on the hull in combination with the magnetic moment produces a lift force on the hull.

[0018] 2. Referring to **FIG. 2**, the trapezoidal hull panel (A) contains three ellipsoidal holes (E). A group of wires or rods (C) running parallel to and just underneath the panel are electrically charged to a high voltage at the end terminals (B). The rods produce a planar potential energy field (D) just under the holes in the panel. The field emerges from the holes in the shape of an ellipsoidal bubble and is amplified by an ellipsoidal dome (F) on the outside of the hull.

[0019] 3. Referring to **FIG. 3**, the group of parallel rods (A) are given a linear charge λ in units of charge per meter. The electric field E developed by the rod is the linear charge divided by the circumference of a circle of radius r around the wire times the permittivity ϵ of space. The analysis of this arrangement shows that within a few grid width spacings, the potential energy field ϕ due to the electric field E_0 has become planar (B) in the z-direction given by the equation

$$\phi = -E_0 z$$

[0020] 4. Referring to **FIG. 4**, the following analysis is done in cylindrical coordinates {r,θ,z}.

[0021] 5. Referring to **FIG. 5**, the ellipsoidal potential energy (B) emerges through the hole in the panel plate (A). In doing so it creates an electrical dipole moment (C) shown by the arrow normal to the hole area.

[0022] 6. Referring to **FIG. 6**, the bubble (B) emerges through plate (A).

[0023] 7. Because the bubble has the shape of an ellipse, the centroid y of the bubble would be four thirds the radius a divided by π as given by

$$\bar{y}_{ellipse} = \frac{4}{3}\frac{a}{\pi}$$

[0024] The electric dipole moment is then given as the charge q times the centroid y. The charge of the hole is equal to the permittivity E times the electric field E emerging from the hole times the area of the hole of radius a

$$q = \varepsilon_0 E\pi a^2 \frac{\text{coul}^2}{\text{m}^2\ \text{newton}}\frac{\text{newton}}{\text{coul}}\ \text{m}^2 = \text{coul} = \text{charge}$$

[0025] 8. The electric dipole moment p is the centroid y times the charge q

$$p = qy = \varepsilon_0 E\pi a^2 \frac{4}{3}\frac{a}{\pi} = \frac{4}{3}\varepsilon_0 E a^2\ \text{coul}\cdot\text{meter}$$

[0026] The electric dipole moment p times the hull velocity v is equal to a magnetic moment μ which is what creates the lift force on the hull

$$\mu = pv\ \text{amp}\cdot m^2$$

[0027] 9. The rotating hull creates the electric dipole moment velocity so that the entire hull develops a magnetic moment. In tensor notation, the magnetic moment μ is in the vertical z-direction because there is a radial component of the electric dipole moment times the velocity. The velocity is the radius r in the radial direction times the angular velocity ω in the z-direction

$$\mu^z = p_r x^r \omega^z$$

[0028] 10. The force F on the hull is the gradient of the dot product of the magnetic moment μ with the magnetic B field

$$F = \nabla(\mu \cdot B)$$

[0029] 11. By electrically charging the hull of the vehicle, a radial electric field is produced. By rotating the hull, the radial electric field changes with time. Thus Maxwell's equations will involve the curl of the magnetic field in the radial direction because the radial electric field is varying with time

$$(\nabla \times B)_r = \frac{1}{c^2}\frac{\partial E_r}{\partial t}$$

[0030] 12. The cross product involves the magnetic field in the theta direction which is zero

$$\frac{1}{r}\frac{\partial B_z}{\partial \theta} - \frac{\partial B_\theta}{\partial z} = \frac{1}{c^2}\frac{\partial E_r}{\partial t} = \frac{1}{r}\frac{\partial B_z}{\partial \theta}$$

[0031] Substituting the derivative of the electric field E

$$E_r = E_0 e^{i\omega t}$$

[0032] and integrating with respect to angle theta gives the vertical magnetic field B as the tangential velocity v times the radial electric field E divided by the speed of light c squared

$$B_z = \frac{v}{c^2} E_r$$

[0033] 13. The force on the hull is the gradient of the magnetic moment μ times the magnetic field B. In the equation for the magnetic field, the only available variable to work with in order to get a gradient of the magnetic field comes from the velocity.

$$\frac{dB_z}{dz} = \frac{E_r}{c^2}\frac{dv}{dz}$$

[0034] 14. Referring to **FIG. 7**, because the hull is in the shape of a pyramid, the velocity is a function of the height z of the hull. Using eight flat sides keeps the radial electric field pointing in the same direction in each panel. Each panel has three domes to produce the magnetic moment for a total of 24 magnetic moment generators.

[0035] 15. Referring to **FIG. 8**, the electric dipole moment (A) points in the radial direction, the rotating hull produces a tangential velocity (B), and the result is a magnetic moment (C) along the panel.

[0036] 16. Referring to **FIG. 9**, because the magnetic moment is parallel to the panel, there are vertical and radial components of the magnetic moment. The vertical magnetic moment creates the dot product with the magnetic field gradient, which is equal to the lift force.

[0037] 17. **FIG. 10** shows a cross-section of the dome (A) and the plate hole (B) with double cladding to enhance the field. The upper cladding (D) has a low relative permittivity in the range of 2 to 40, and the lower layer has a high relative permittivity in the range of 1200 to 4000.

[0038] 18. Comparing **FIG. 11** to **FIG. 7**, this dome and cladding configuration creates a much larger electric dipole moment compared to a hole in the plate. The wavy lines are the equi-potential energy lines from the dome (C) and the upper layer (B) and the lower level (A).

I claim:

1. A spacecraft propulsion system comprising:

a rotating octagon of trapezoidal electrostatically charged flat panels which form a closed sloping hull in the shape of a pyramid;

panels each having three holes covered by three ellipsoidal domes of varying size centrally located along the major length and axis of each panel; and

a grid of high voltage electrostatically charged rods located on the interior side of each panel such that a planar potential field is produced parallel to and under each panel hole.

2. The domes, holes and rotating charged hull of method 1 producing:

an ellipsoidal potential energy field emerging from the holes and generating an electric dipole moment on the outside of the hull;

a magnetic moment in the vertical direction due to the rotating electric dipole moment; and

a rotating electric field in the radial direction which generates a corresponding magnetic field gradient in the vertical direction proportional to the velocity gradient of the sloping panels of the hull.

3. A lift force on the spacecraft hull generated by:

the magnetic moment times the gradient of the magnetic field in the vertical direction; and

a dual surface layer hull cladding having different permittivities which enhance the electric dipole moment whereby the upper cladding has a low relative permittivity in the range of 2 to 40, and the lower layer has a high relative permittivity in the range of 1200 to 4000.

* * * * *

US 20060038081A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2006/0038081 A1**

(43) **Pub. Date: Feb. 23, 2006**

(54) ELECTRIC DIPOLE SPACECRAFT

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/912,621**

(22) Filed: **Aug. 4, 2004**

Publication Classification

(51) **Int. Cl.**
B64G 1/40 (2006.01)

(52) **U.S. Cl.** **244/171.5**

(57) ABSTRACT

This invention is a rotating spacecraft that produces an electric dipole on four rotating spherical conducting domes perturbing a uniform spherical electric field to create a magnetic moment interacting with the gradient of a magnetic field that generates a lift force on the hull.

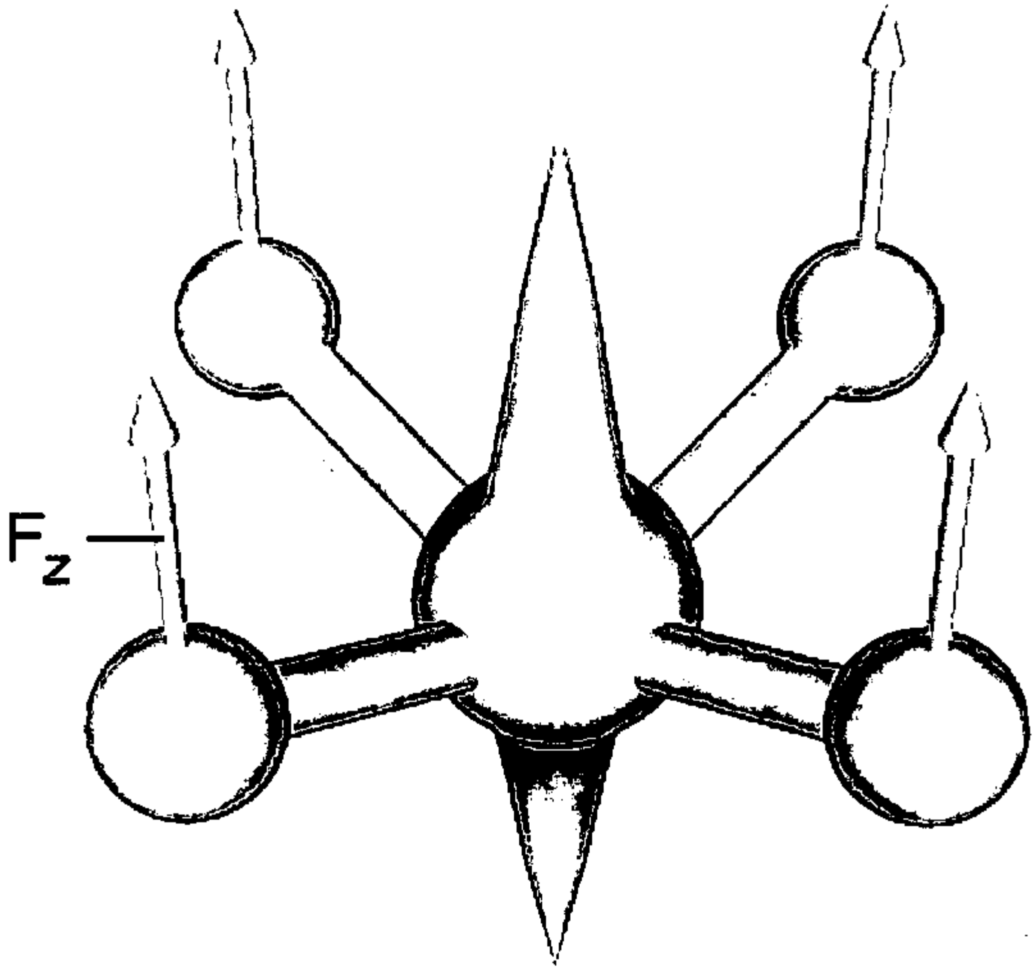

Figure 1

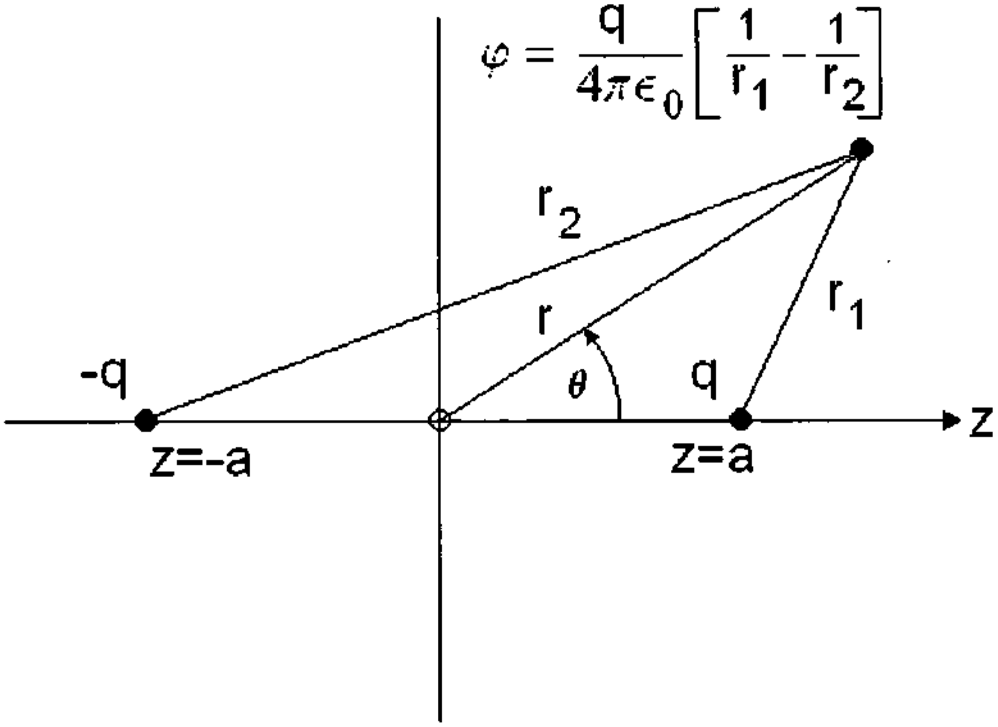
$\varphi = \frac{q}{4\pi\epsilon_0}\left[\frac{1}{r_1} - \frac{1}{r_2}\right]$
r_2
r_1
r
θ
-q
q
z=-a
z=a
z

Figure 2

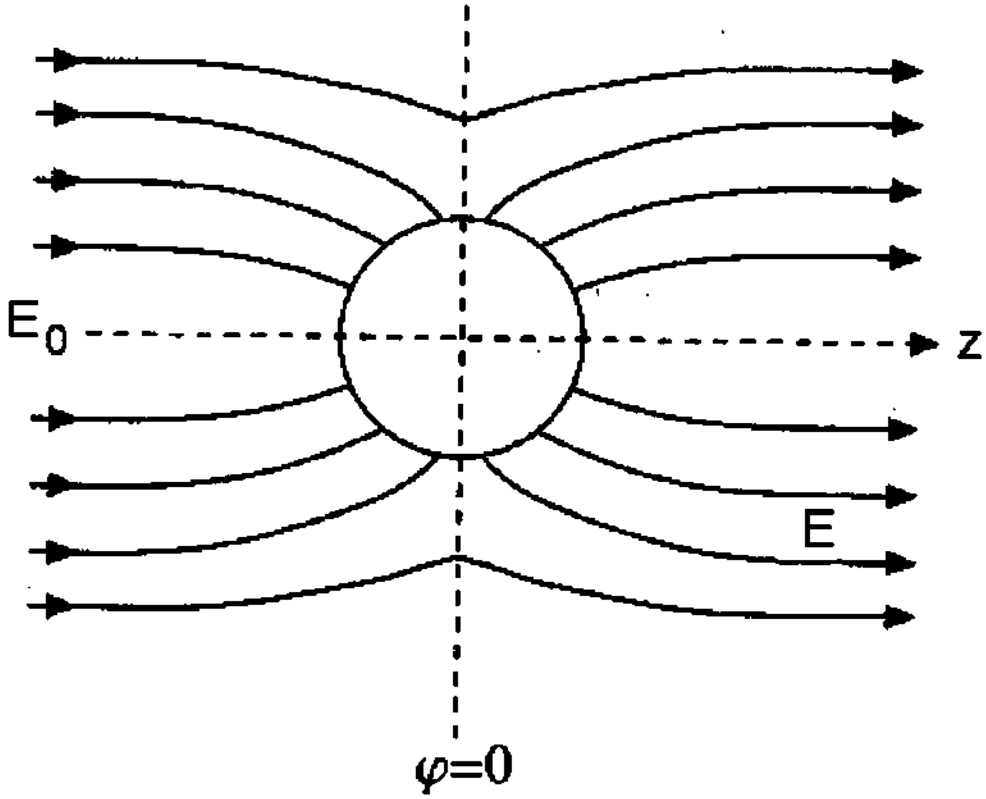
E_0
z
E
$\varphi=0$

Figure 3

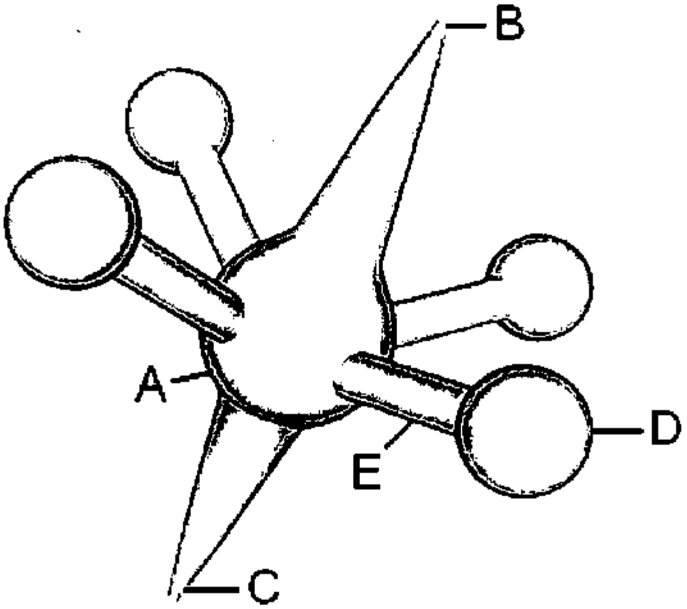

Figure 4

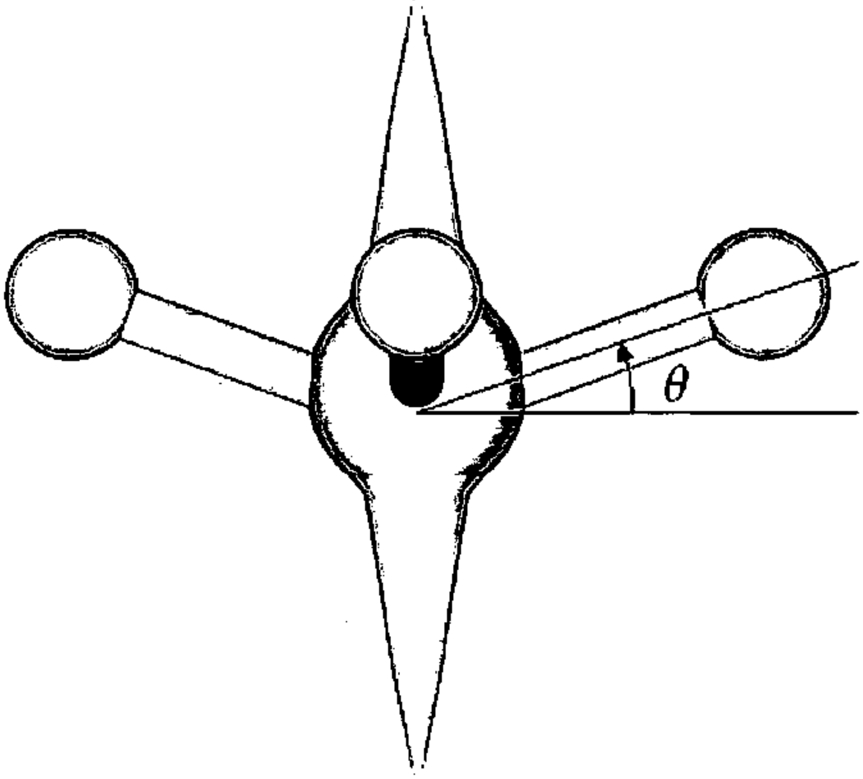

Figure 5

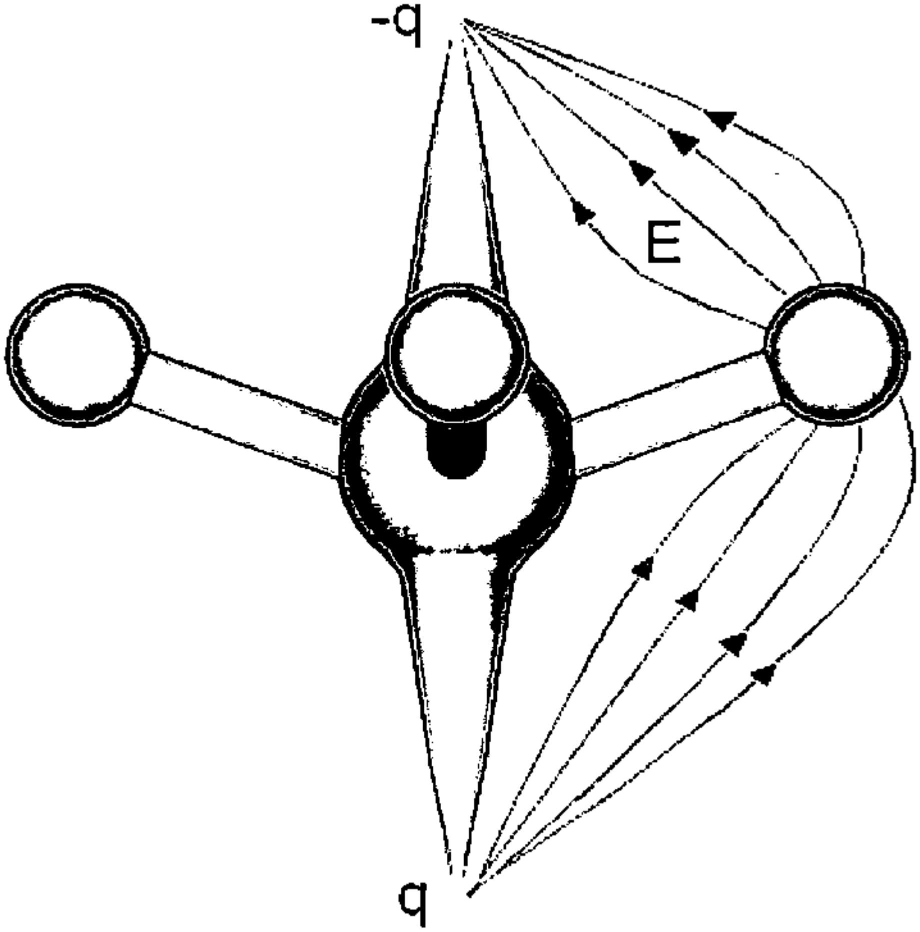

Figure 6

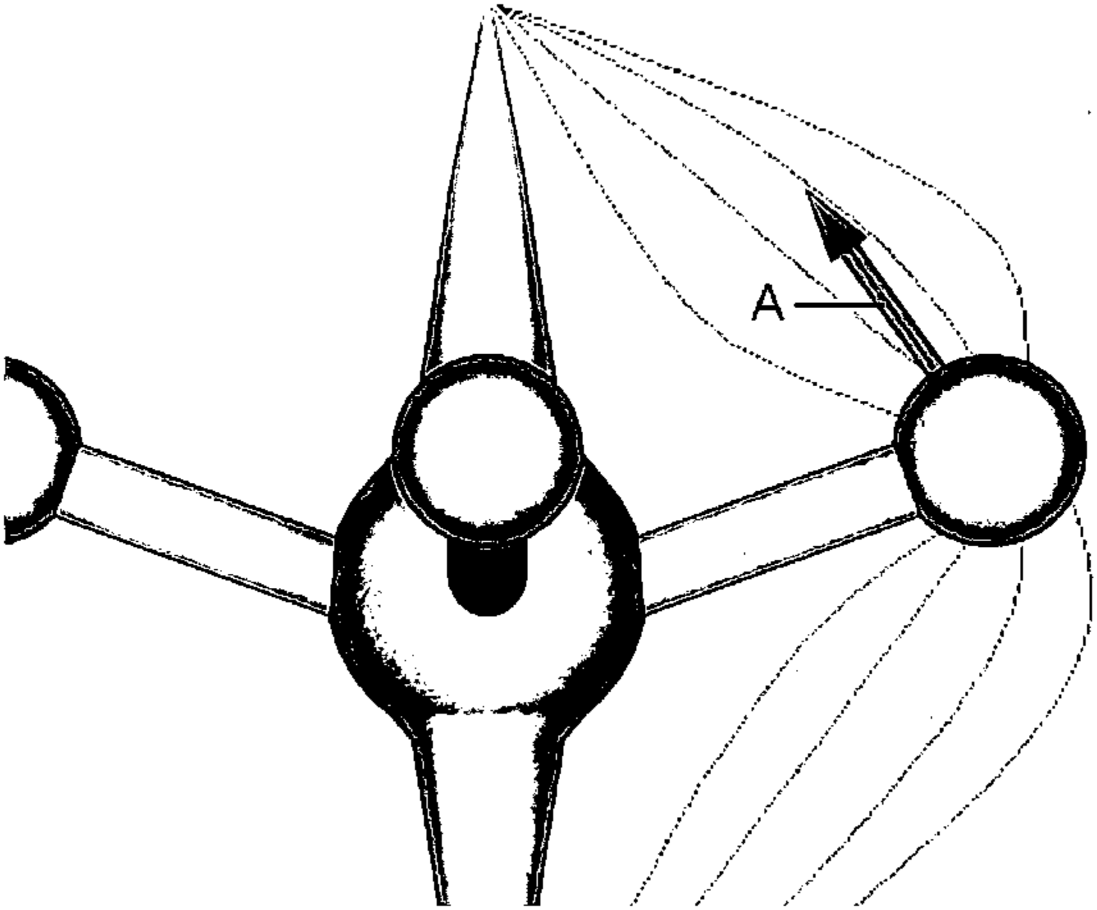

Figure 7

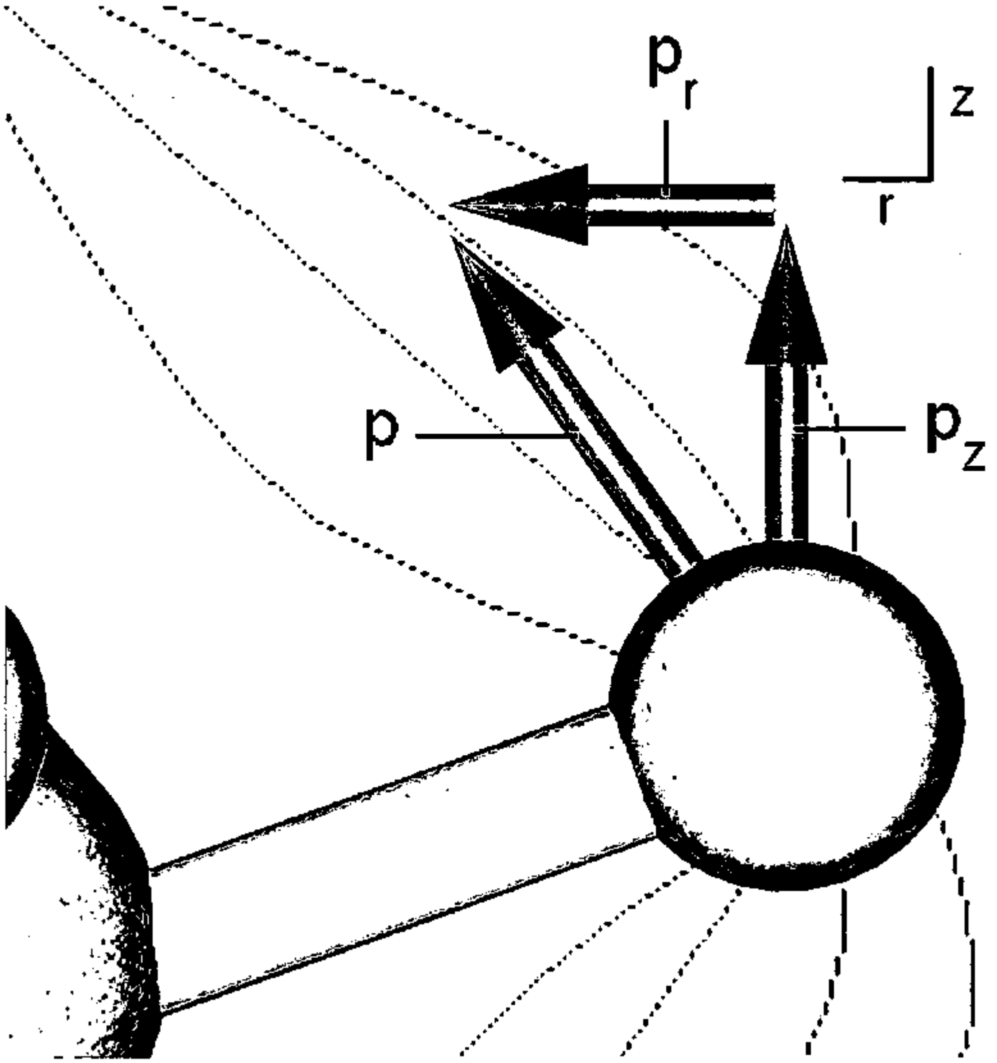

Figure 8

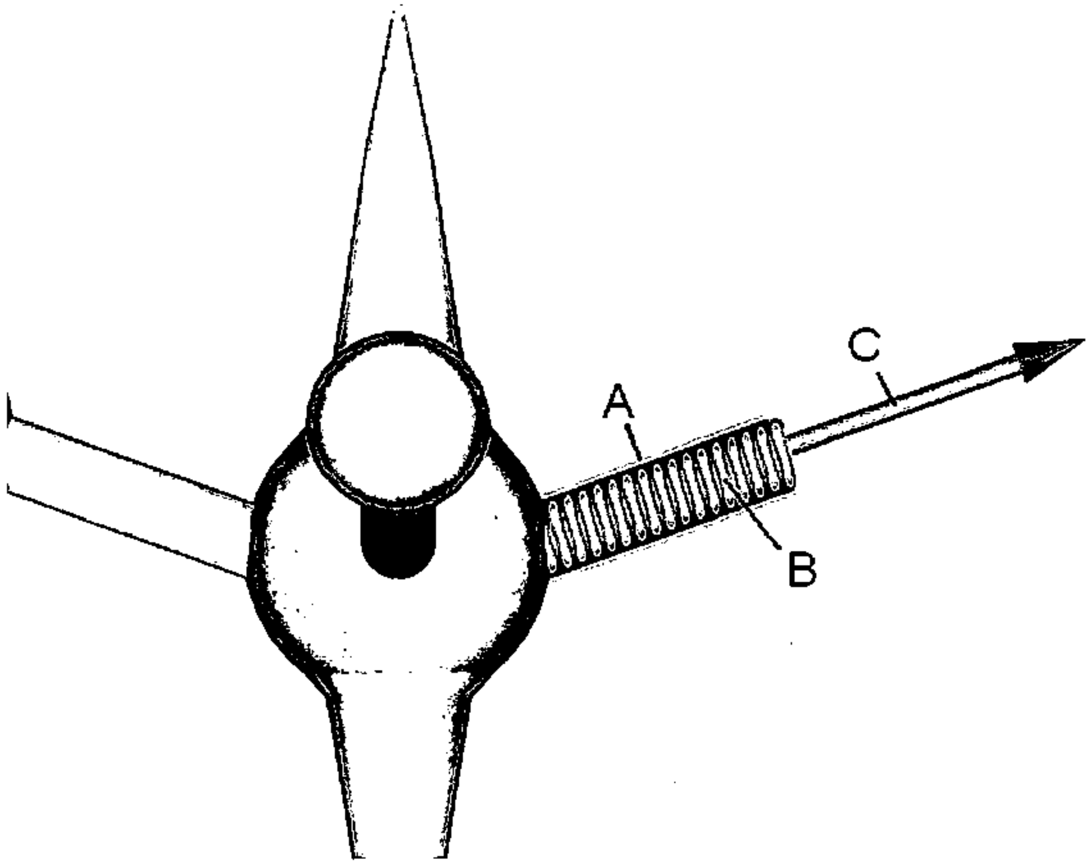

Figure 9

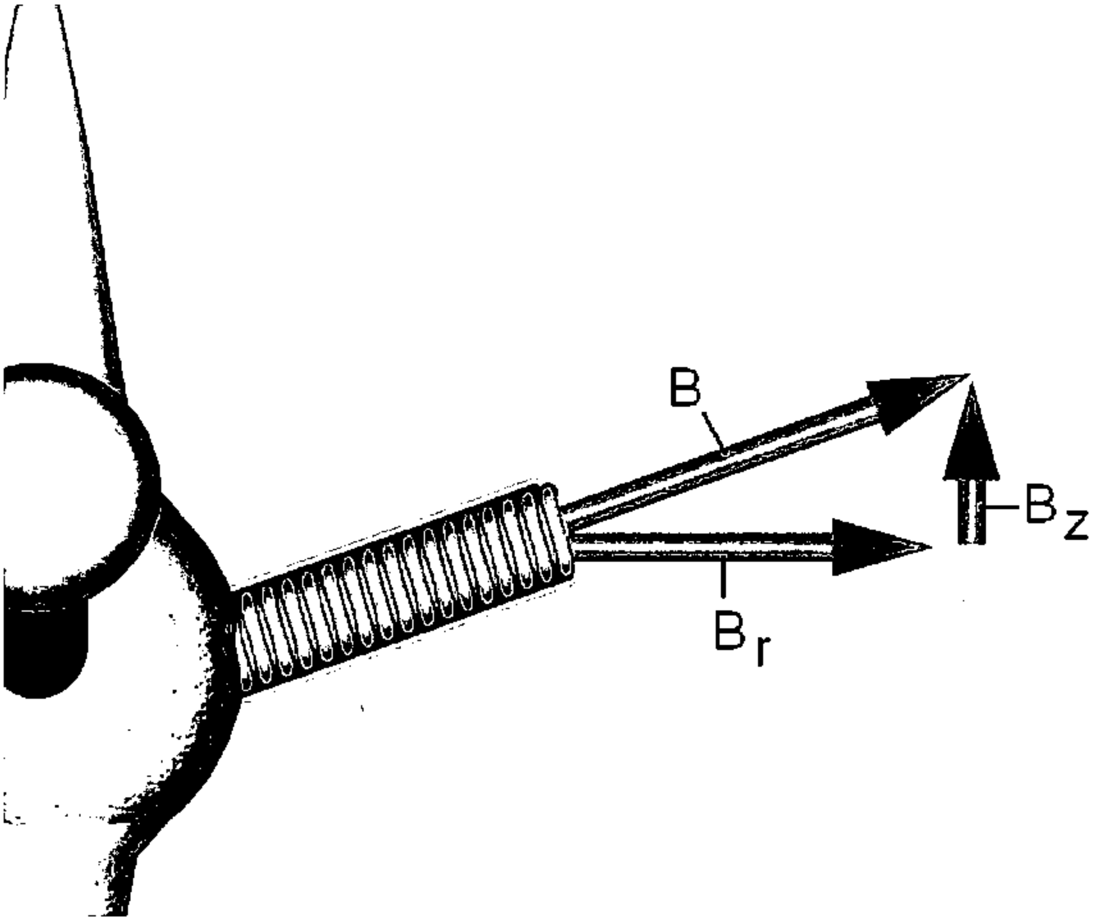

Figure 10

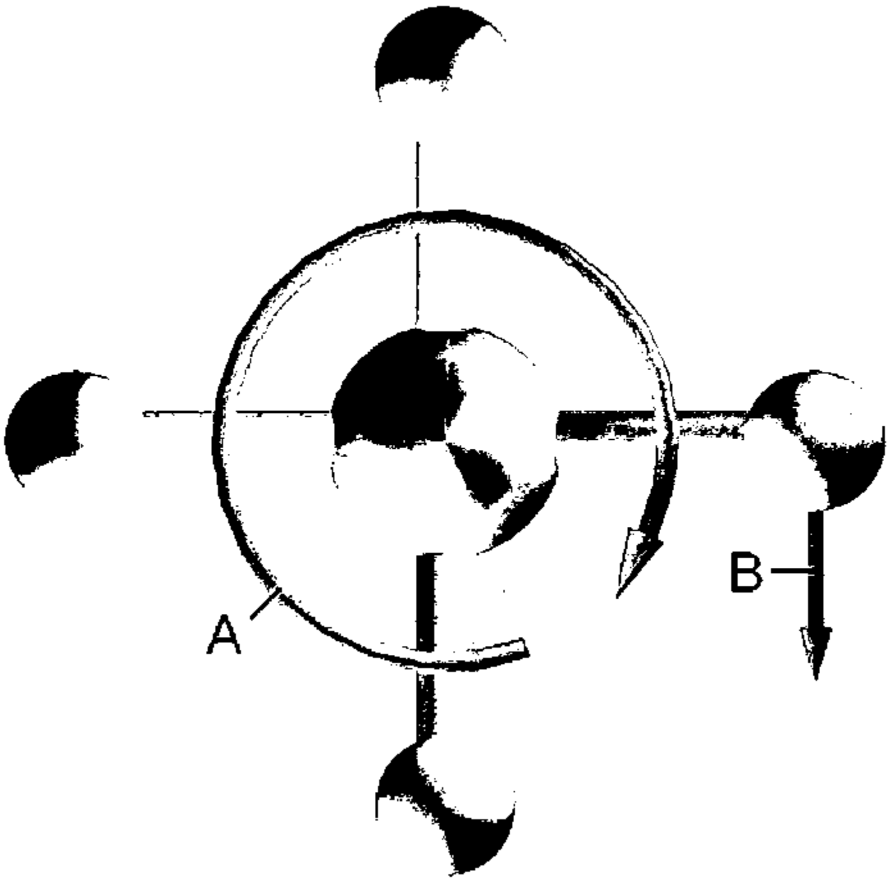

Figure 11

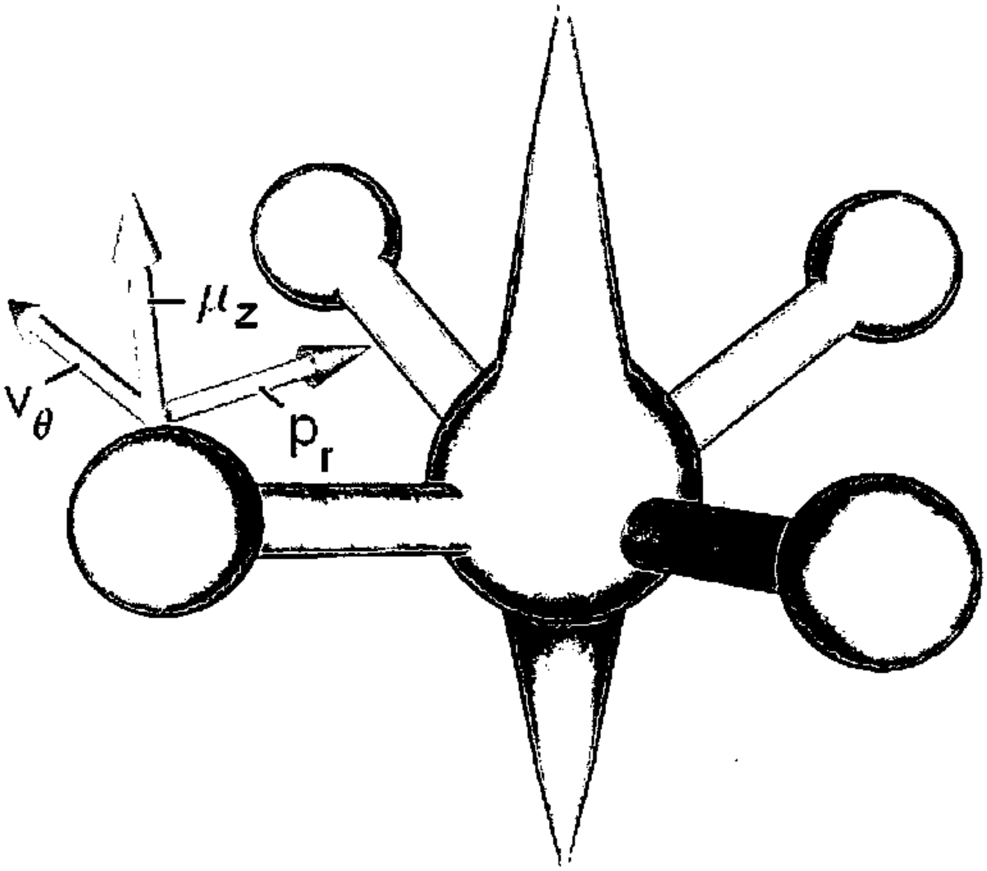
μ_z
v_θ
p_r

Figure 12

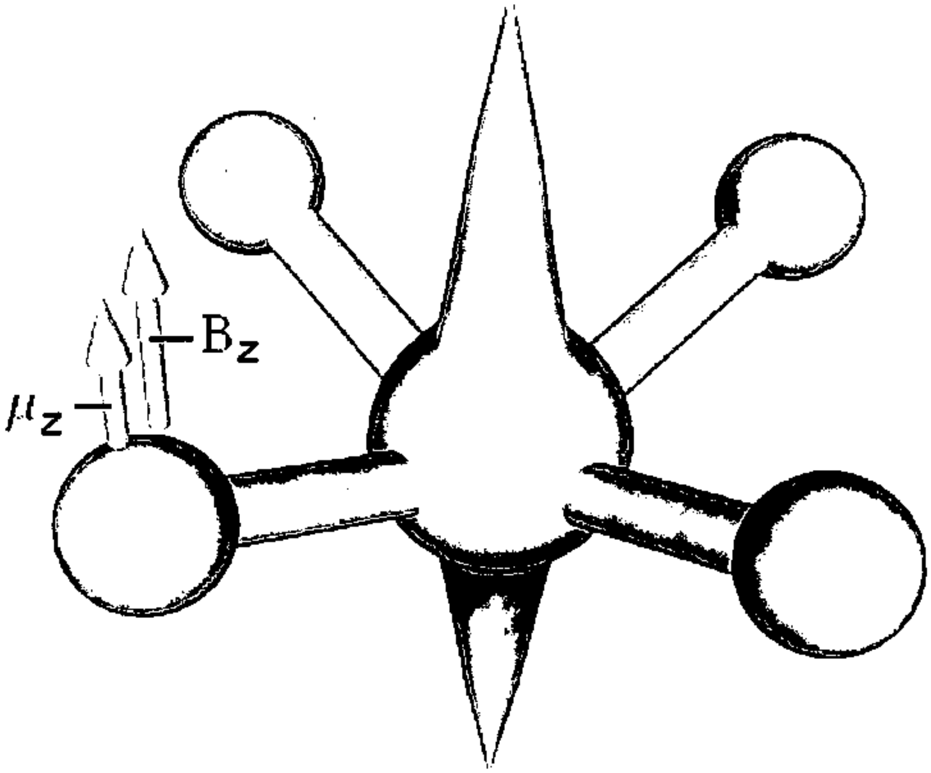

Figure 13

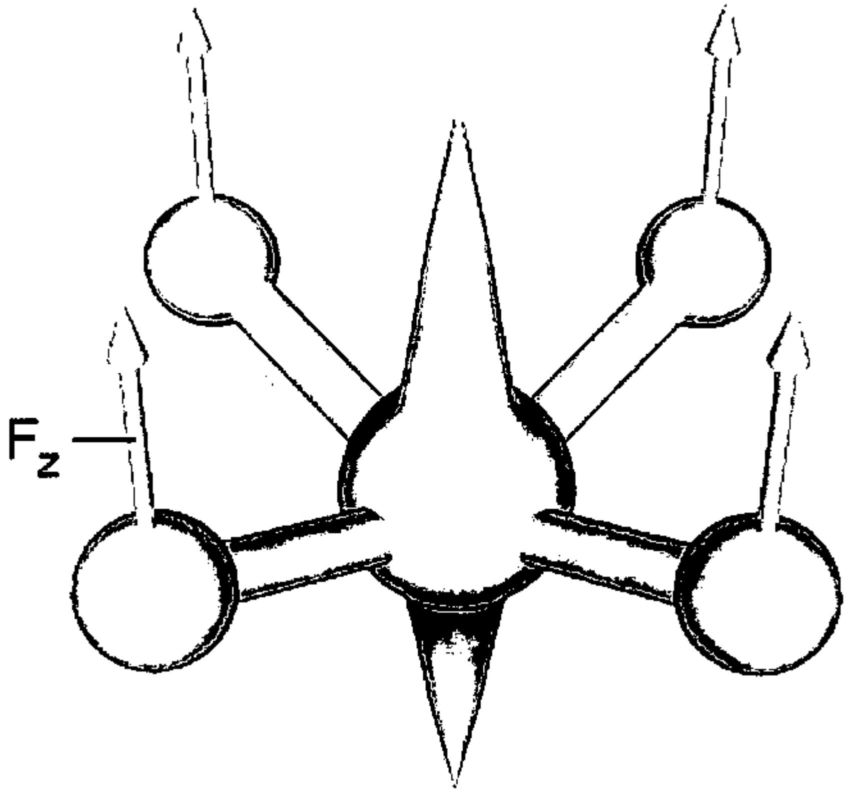

Figure 14

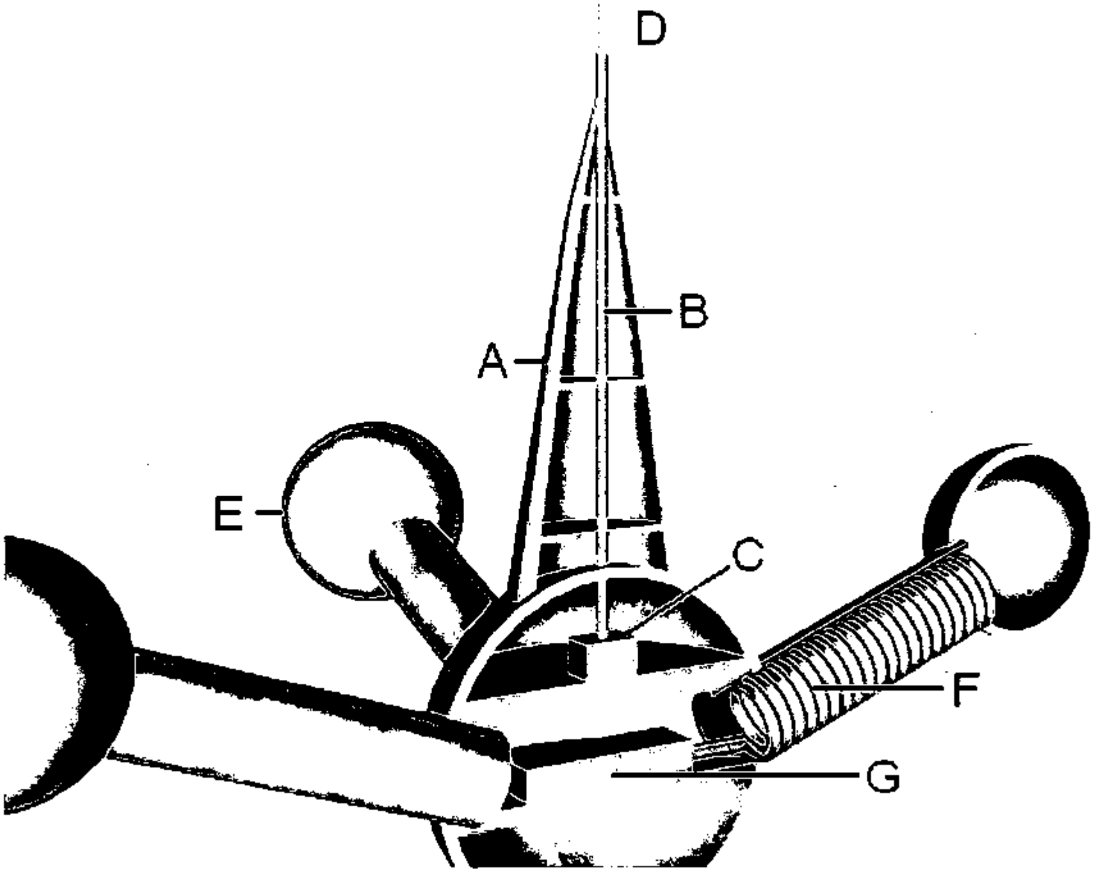

ELECTRIC DIPOLE SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a rotating spacecraft that utilizes four spherical conducting domes perturbing a uniform electric field in order to create a lift force by means of a magnetic moment times the gradient of a magnetic field.

BACKGROUND OF THE INVENTION

[0002] An electric dipole p is two electrical charges of opposite sign {q, −q} separated by a distance a.

$$p=qa=\text{coulomb}\cdot\text{meter}$$

[0003] If this dipole is moving with a velocity v, it produces a magnetic moment μ.

$$\mu = pv = qav = \text{coulomb meter}\frac{\text{meter}}{\text{sec}} = \frac{\text{coulomb}}{\text{sec}}\text{meter}^2 = IArea$$

[0004] The magnetic moment is equal to a current I circulating around an area. The magnetic field B has units of kilogram per second per charge coulomb.

$$B = \frac{\text{kg}}{\text{sec } coul}$$

[0005] The gradient of the magnetic field in the vertical direction z has units of

$$\frac{\text{dB}}{dz} = \frac{\text{kg}}{\text{sec } coul \text{ meter}}$$

[0006] This gradient interacting with a magnetic moment creates a force F measured in newtons.

$$F_z = \mu\frac{\text{dB}}{dz} = \frac{coul\ \text{m}^2}{\text{sec}}\frac{\text{kg}}{\text{sec } coul\ m} = \frac{\text{kg } m}{\text{sec}^2} = \text{newton}$$

[0007] In terms of vectors, the force is equal to

$$F=\nabla(\mu\cdot B)$$

which is the gradient ∇ of the dot product $(\cdot)$ of the magnetic moment with the magnetic field. This means that the magnetic moment has to be aligned with the field. The lift force on the spacecraft would then be the magnetic moment in the vertical z-direction μ_z times the magnetic field in the z-direction B_z. For constant magnetic moment, the gradient affects the magnetic field only, resulting in the same force equation

$$F_z=\nabla(\mu_z\cdot B_z)=\mu_z\nabla B_z$$

[0008] Referring to **FIG. 1**, the electric dipole has a positive charge q located on the z-axis at a distance a from the origin of the graph. A second negative charge −q is located at a distance −a from the origin. The positive charge produces an electrostatic potential ϕ_1 at a radius r_1 equal to the charge q divided by 4π times the permittivity of space ϵ_0

$$\varphi_1 = \frac{q}{4\pi\varepsilon_0}\frac{1}{r_1}$$

where the permittivity is linear capacitance, measured in farads per meter. The electrostatic potential has units of volts

$$\varphi = \frac{coul}{\left(\frac{\text{farad}}{\text{meter}}\right)}\frac{1}{\text{meter}} = \frac{coul}{\text{farad}} = \text{volt}$$

because the charge in coulombs held by a capacitor is equal to the capacitance, measured in farads, times the capacitor voltage. Because the second charge has the opposite sign, the potential ϕ_2 at a radius r_2 to the same point in space is

$$\varphi_2 = \frac{-q}{4\pi\varepsilon_0}\frac{1}{r_2}$$

[0009] The total potential ϕ at some point in space is equal to the sum of the two potentials, or

$$\varphi = \varphi_1 + \varphi_2 = \frac{q}{4\pi\varepsilon_0 r_1} - \frac{q}{4\pi\varepsilon_0 r_2} = \frac{q}{4\pi\varepsilon_0}\left[\frac{1}{r_1} - \frac{1}{r_2}\right]$$

[0010] As seen in the diagram, the point of space is a distance r from the origin. Using the law of cosines, radius r_1 can be written as

$$r_1 = (r^2 + a^2 - 2ar\cos(\theta))^{\frac{1}{2}} = r\left(1 + \left(\frac{a}{r}\right)^2 - 2\left(\frac{a}{r}\right)\cos(\theta)\right)^{\frac{1}{2}} = r(1 - 2xt + t^2)^{\frac{1}{2}}$$

where t is the ratio of the charge location over the radius, and x is cos(θ). The potential for positive charge q_1 can be written

$$\varphi_1 = \frac{q}{4\pi\varepsilon_0}\frac{1}{r_1} = \frac{q}{4\pi\varepsilon_0}\frac{1}{r}(1 - 2xt + t^2)^{-1/2}$$

[0011] Dropping the factor $q/4\pi\epsilon_0 r$, the square root can be expressed in terms of the Legendre polynomial $P_n\cos(\theta)$ of the nth power

$$g(t, x) = (1 - 2xt + t^2)^{-1/2} = \sum_{n=0}^{\infty} P_n(x)t^n$$

where the absolute value of t is less than one. The polynomial coefficients of t^n can be obtained by using the binomial theorem to expand the generating function g(t,x) as

$$(1-2xt+t^2)^{-1/2}=\sum_{n=0}^{\infty}\frac{(2n)!}{2^{2n}(n!)^2}(2xt-t^2)^n$$

which evaluates to

$$\frac{0!}{2^0(0!)^2}(2xt-t^2)^0+\frac{2!}{2^2(1!)^2}(2xt-t^2)^1+\frac{4!}{2^4(2!)^2}(2xt-t^2)^2$$
$$1t^0+xt^1+\left(\frac{3}{2}x^2-\frac{1}{2}\right)t^2+\text{order}(t^3)$$

[0012] The first three Legendre polynomials are therefore

$$P_0=1$$
$$P_1=x$$
$$P_2=\frac{1}{2}(3x^2-1)$$

[0013] The electrostatic potential for both charges of the electric dipole is

$$\varphi=\frac{q}{4\pi\varepsilon_0}\frac{1}{r}\left\{\left(1-2\left(\frac{a}{r}\right)\cos(\theta)+\left(\frac{a}{r}\right)^2\right)^{-1/2}-\left(1+2\left(\frac{a}{r}\right)\cos(\theta)+\left(\frac{a}{r}\right)^2\right)^{-1/2}\right\}$$

[0014] The potential can be evaluated in terms of the Legendre polynomials as

$$\begin{aligned}\varphi&=\frac{q}{4\pi\varepsilon_0 r}\left[\sum_{n=0}^{\infty}P_n(\cos(\theta))\left(\frac{a}{r}\right)^n-\sum_{n=0}^{\infty}P_n(\cos(\theta))(-1)^n\left(\frac{a}{r}\right)^n\right]\\&=\frac{2q}{4\pi\varepsilon_0 r}\left[P_1(\cos(\theta))\left(\frac{a}{r}\right)+P_3(\cos(\theta))\left(\frac{a}{r}\right)^3+\dots\right]\end{aligned}$$

[0015] The first and most dominant term when the radius is much greater than location a is equal to

$$\varphi=\frac{2aq}{4\pi\varepsilon_0}\frac{P_1(\cos(\theta)0}{r^2}$$

which is the electric dipole potential and 2aq is the dipole moment

$$p=2aq$$

[0016] Now imagine a constant electric field E_0 which is perturbed by a conducting sphere of radius a. The unperturbed electrostatic potential outside the sphere would the negative of the electric field times the radius times the Legendre polynomial, or

$$\phi_1=-E_0rP_1$$

[0017] The electrostatic potential perturbed by the charges is the voltage E_0a times the radius a times the a of the dipole moment times the Legendre polynomial divided by the radius squared

$$\varphi_2=E_0aaa\frac{P_1}{r^2}=E_0a^3\frac{P_1}{r^2}$$

[0018] The total potential outside the sphere is the sum of the two potentials equal to

$$\varphi=-E_0rP_1+E_0a^3\frac{P_1}{r^2}=-E_0P_1\left(r-\frac{a^3}{r^2}\right)=-E_0rP_1\left(1-\left(\frac{a}{r}\right)^3\right)$$

[0019] Referring to **FIG. 2**, the previously uniform electric field is shown perturbed by the neutral conducting sphere. The center of the sphere is taken as the origin and the z-axis is oriented parallel to the original uniform field.

[0020] The electric field induces a surface charge density σ on the sphere equal to the negative of the permittivity of space times the gradient of the electrostatic potential

$$\sigma=-\varepsilon_0\frac{\partial\varphi}{\partial r}\Big|_{r=a}=3\varepsilon_0E_0\cos(\theta)$$

[0021] The electric field also induces an electric dipole moment on the sphere equal to the

$$p=\frac{qa}{r}\cdot\frac{2a^2}{r}=4\pi\varepsilon_0E_0a^3$$

with units of coulomb-meter. If this sphere is rotating around a central axis at a velocity v, it will create a magnetic moment μ equal to the dipole moment times the velocity.

$$\mu=pv$$

with units of ampere-meter^2.

SUMMARY OF THE INVENTION

[0022] As shown in the preceding background section, a neutral conducting sphere placed in a uniform electric field will generate a magnetic moment when rotated around a central axis. The electric field can be created by two points charges of opposite sign separated by a distance between them.

[0023] Referring to **FIG. 3**, the spacecraft has a spherical cabin (A) to which are attached cone-shaped electrostatic towers (B,C) above and below the cabin along the direction of travel in the z-direction. Because the electric field goes from the positive charge to the negative charge, the tip of the lower tower has a positively charged electrode, and the upper tower has a negatively charged electrode. Four equally-spaced neutral conducting spheres (D) are connected to the cabin by non-conducting tubes (E). The tubes make an angle with the cabin such that the distance (CD) is greater than distance (DB). The angle θ of the tube with respect to the cabin can be seen in side view **FIG. 4**.

[0024] Referring to **FIG. 5**, the charges create a uniform spherical field between the towers. The conducting spheres perturb this field such that the electric field **(E)** points toward the upper tower in a manner similar to that shown previously in **FIG. 2**.

[0025] Referring to **FIG. 6**, the conducting sphere produces an electric dipole moment (A) pointing at an angle toward the upper tower.

[0026] Referring to **FIG. 7**, by the law of addition of vectors, the electric dipole $\bar{p}$ can be represented by two orthogonal vectors pointing in the vertical z-direction p_z and in the inward radial direction p_r.

[0027] Referring to **FIG. 8**, the hollow tube (A) connecting the cabin with the conducting sphere contains a spiral-wound electrical solenoid (B) which produces a magnetic field (C). This magnetic field B can be decomposed into two orthogonal vectors pointing in the vertical z-direction B_z and in the outward radial direction B_r as shown in **FIG. 9**.

[0028] Referring to top-view **FIG. 10**, the spacecraft has a clockwise angular velocity ω (A) which gives the conducting sphere a velocity v as shown by the vector (B). By the right-hand rule of physics, the angular velocity vector points in the negative z-direction. The angular velocity in the z-direction crossed with the radius r in the radial direction produces a velocity v in the clockwise θ-direction using cylindrical coordinates {r,θ,z}.

$$v_\theta = w'_z \times r_r = -\omega r$$

[0029] Referring to **FIG. 11**, the negative radial dipole moment p_r crossed with the negative velocity v_θ of the sphere produces a positive magnetic moment μ_z in the z-direction.

$$\mu_z = p_r \times v_\theta = (-p_r)(-v_\theta) = pv$$

[0030] Referring to **FIG. 12**, the magnetic field B_z in the vertical z-direction is dotted with the magnetic moment μ_z in the z-direction to produce a force F_z in the vertical z-direction on each conducting sphere (**FIG. 13**).

$$F_z = \nabla(\mu_z \cdot B_z) = \mu_z \nabla B_z$$

[0031] The magnetic field that is produced by the solenoid actually curves away and around. Thus there is a gradient of the field in the z-direction.

[0032] The force can also be expressed in tensor notation. The magnetic B field in the vertical direction is part of an electromagnetic 4×4 matrix Faraday tensor F

$$F^\alpha_\beta = \begin{array}{c|cccc|} & \beta = t & r & \theta & z \\ \alpha = t & 0 & 0 & 0 & 0 \\ \alpha = r & 0 & 0 & B_z & 0 \\ \alpha = \theta & 0 & -B_z & 0 & 0 \\ \alpha = z & 0 & 0 & 0 & 0 \end{array}$$

which shows that the magnetic field is located in slot $F^r{}_\theta$ of the Faraday tensor. In tensor notation the subscripts and superscripts have to match up on both sides of the equation. Matching subscripts and superscripts on the same side of the equation cancel. In this case, the electric dipole moment is in the radial direction p_r. The velocity can be represented as a time derivative of the θ-coordinate x^θ or

$$v^\theta = \frac{\partial x\theta}{\partial t}$$

[0033] Thus the force component in the z-direction becomes

$$F_z = p_r v^\theta F^r_{\theta,z} = (-p)(-v)\frac{\partial B_z}{\partial z} = pv\frac{\partial B_z}{\partial z}$$

where the angular and radial tensor components cancel and comma-z (, z) represents differentiation of the magnetic field in the z-direction.

[0034] The spacecraft design also has an inherent motion control system for moving in various directions. If the magnetic field of one solenoid arm is increased or decreased, the force on that sphere will be increased or decreased. Thus the spacecraft can turn in a particular direction.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0035] **FIG. 1**. Electric dipole.

[0036] **FIG. 2**. Uniform electric field perturbed by electric dipole.

[0037] **FIG. 3**. Perspective view of spacecraft.

[0038] **FIG. 4**. Angle of solenoid tube.

[0039] **FIG. 5**. Electric field perturbed by conducting sphere.

[0040] **FIG. 6**. Electric dipole generated by conducting sphere.

[0041] **FIG. 7**. Orthogonal vector components of electric dipole.

[0042] **FIG. 8**. Magnetic field produced by tube arm solenoid.

[0043] **FIG. 9**. Orthogonal vector components of magnetic field.

[0044] **FIG. 10**. Angular velocity of hull.

[0045] **FIG. 11**. Magnetic moment produced by radial electric dipole and sphere velocity.

[0046] **FIG. 12**. Dot product of the magnetic moment with the magnetic field.

[0047] **FIG. 13**. Vertical lift force on all four conducting spheres.

[0048] **FIG. 14**. Perpective view of spacecraft interior.

DETAILED DESCRIPTION OF THE INVENTION

[0049] 1. Referring to the cut-away view **FIG. 14**, the construction of the spacecraft is a thin-wall insulating thermoplastic having a dielectric constant in the range of 20 kilovolts per millimeter (A). An insulated electrode (B) runs from the cabin power supply and high-voltage transformer

(C) to the tip of each tower (D). The four spheres (E) are silver plated to make them conducting. The tube solenoids (F) are driven by a direct current power supply (G).

[0050] 2. The present model uses 3D computer design software and stereolithography fabrication techniques to create the thin-wall, low-weight, hollow structure of the hull. The computer model is sliced into many thin horizontal slices. A laser, mounted on an x-y table, draws out the slice on a table immersed in a bath of liquid polymer. Due to its sensitivity to the light, the liquid polymerizes. The table is then lowered a few thousandths of an inch more and the process is repeated. Thus making hollow spherical and conical shapes is extremely easy to do. Parts can be designed and stored in *.STL stereolithography files for transmission by Internet e-mail to the service bureau machine shop which sends the finished parts back the next day by express mail.

I claim:

1. A spacecraft comprising:

a. a spherical cabin;

b. an electrostatic conical tower mounted on top of item (**1***a*), supporting a vertically-mounted negatively-charged insulated electrode at the tip of the tower;

c. an electrostatic conical tower mounted on the bottom of item (**1***a*), supporting a vertically-mounted positively-charged insulated electrode at the tip of the tower;

d. a vertical electric dipole created by items (**1***b*) and (**1***c*);

e. a high-voltage transformer to drive item (**1***d*), mounted in item (**1***a*);

f. four tubular arms, mounted at 90° around and extending at an angle from item (**1***a*);

g. four solenoids, each of which is mounted axially inside item (**1***f*);

h. a direct current power supply to drive item (**1***g*);

i. four silver-plated conducting spheres, each of which is mounted on the end of item (**1***f*);

2. an electrostatic lift system that:

a. produces a uniform spherical electric field by means of item (**1***d*) which envelopes item (**1***i*);

b. produces a perturbed electric field due to the presence of item (**1***i*);

c. produces an electric dipole moment in the direction of item (**1***b*) due to items (**2***a*) and (**2***b*);

d. produces a vertical magnetic moment due to the clockwise angular velocity of item (**1***a*) combined with item (**2***c*);

e. produces a vertical lift force on item (**1***i*) due to item (**2***d*) combined with the magnetic field gradient in the vertical direction produced by item (**1***g*); and

f. creates a motion control system by varying the current to item (**1***g*) in order to increase or decrease the effect of item (**2***e*) on a particular item (**1***i*).

* * * * *

US 20060144035A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0144035 A1**
St. Clair (43) **Pub. Date: Jul. 6, 2006**

(54) PHOTON SPACECRAFT

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/027,969**

(22) Filed: **Jan. 3, 2005**

Publication Classification

(51) **Int. Cl.**
F03H 1/00 (2006.01)
(52) **U.S. Cl.** **60/203.1**; 60/204

(57) **ABSTRACT**

A spacecraft propulsion system utilizing photon particles to create negative energy over the hull in order to generate a lift force on the hull.

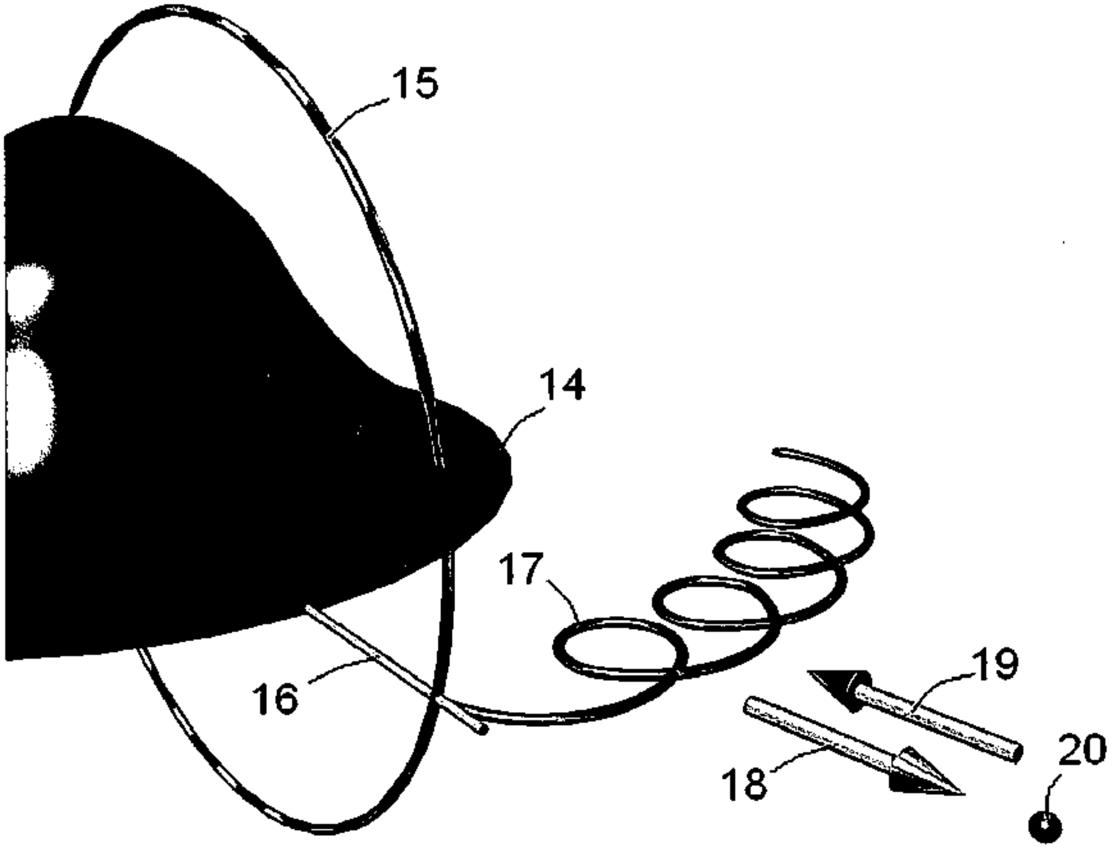

Figure 1

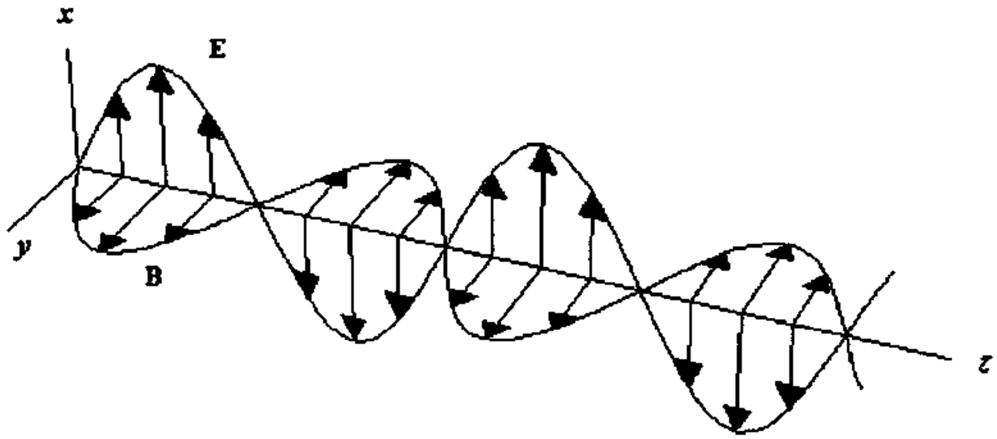

Figure 2

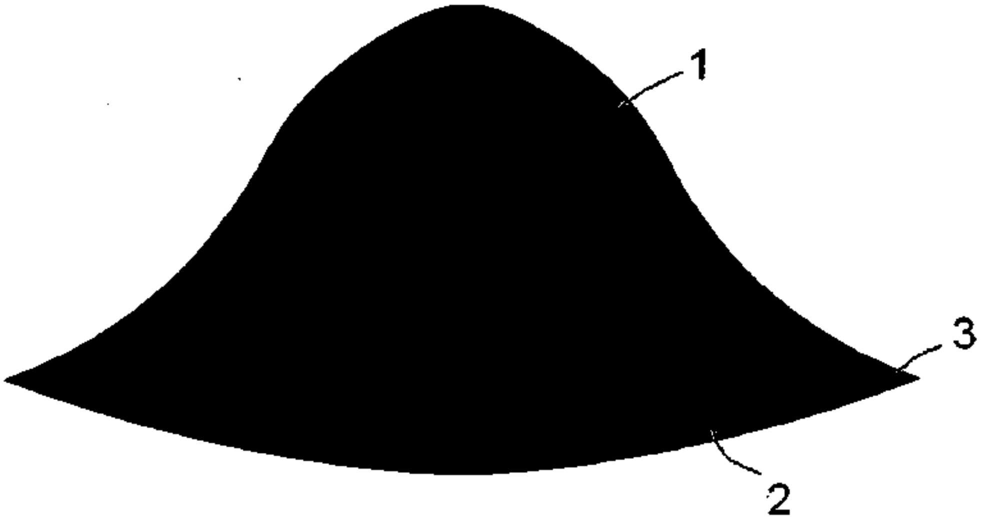

Figure 3

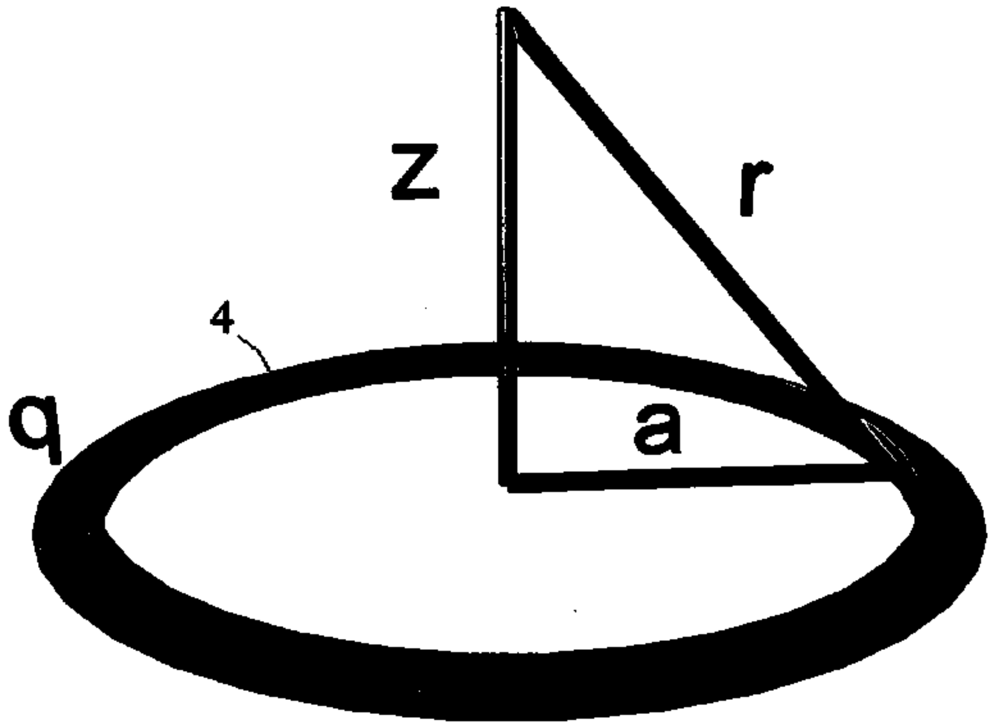

Figure 4

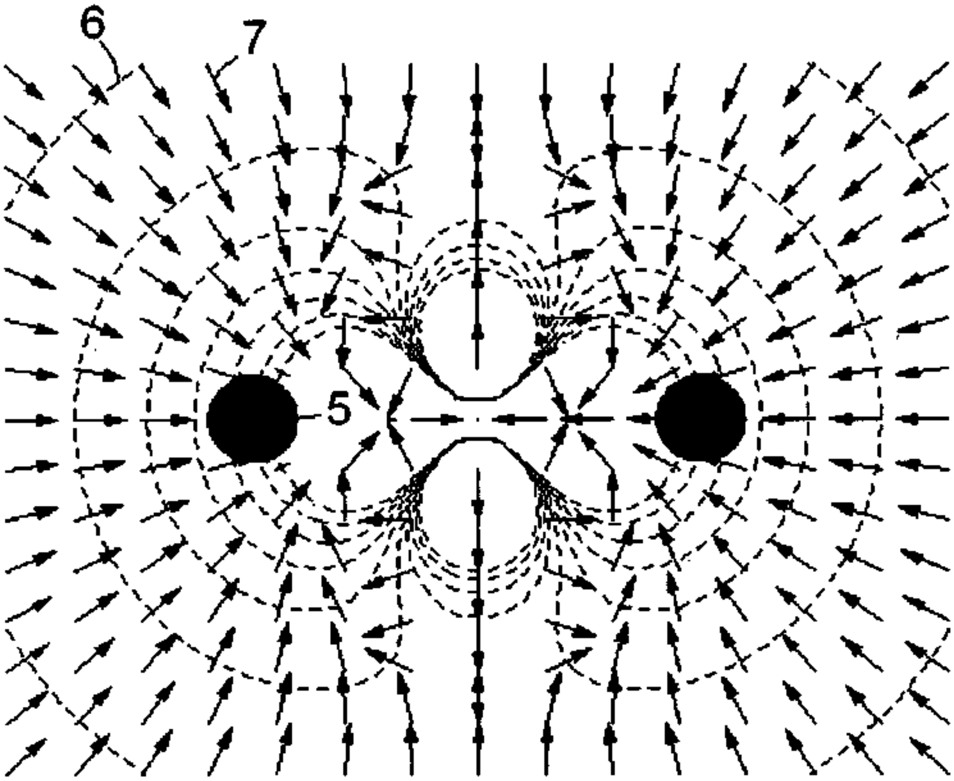

Figure 5

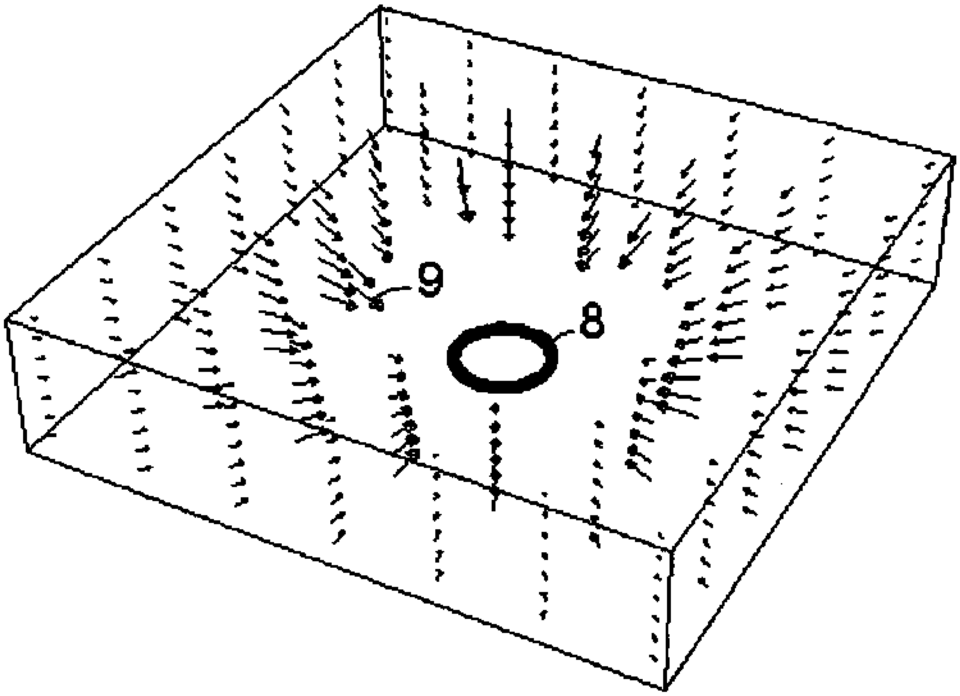

Figure 6

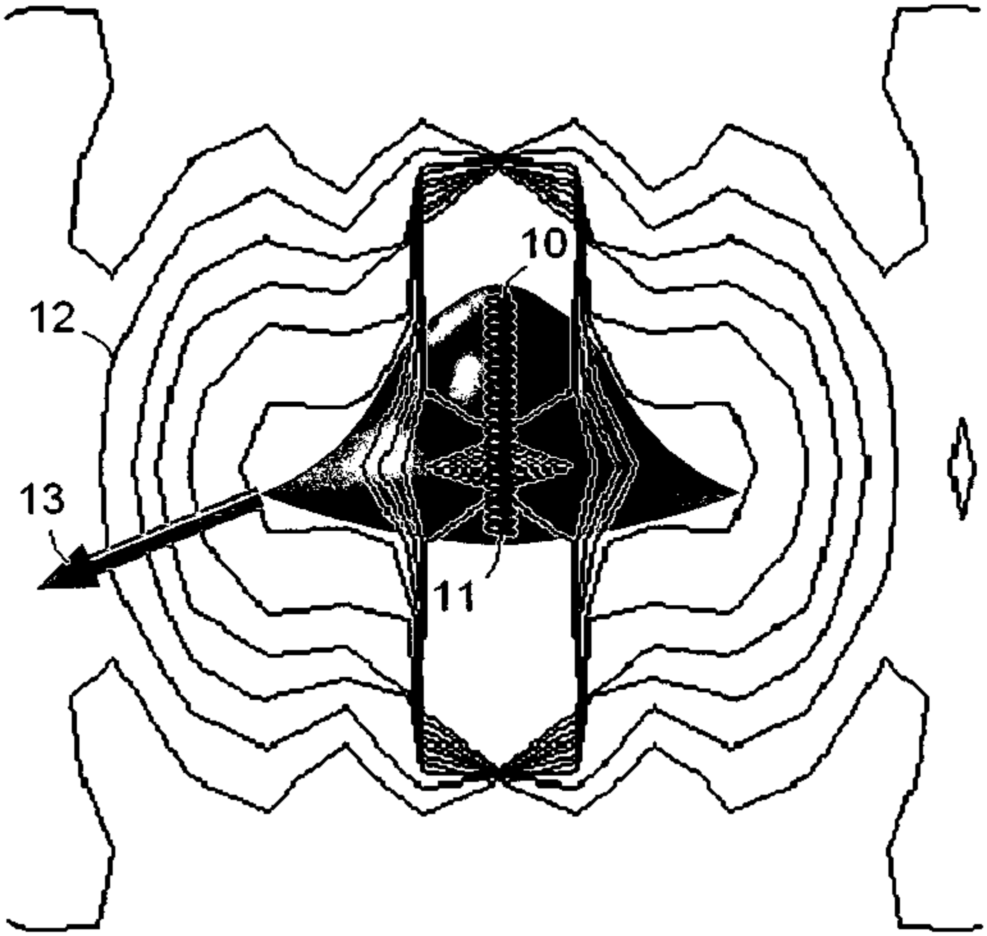

Figure 7

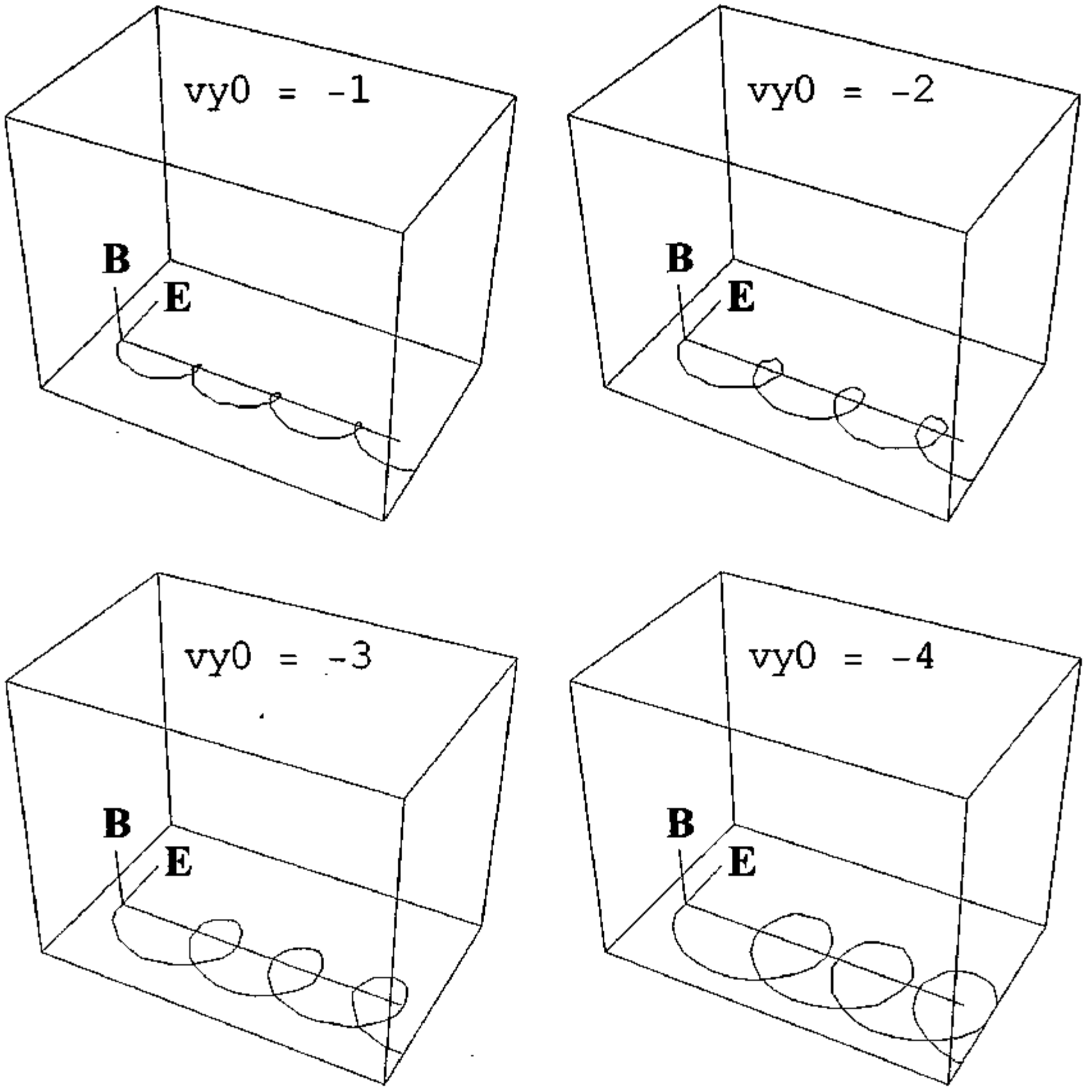

Figure 8

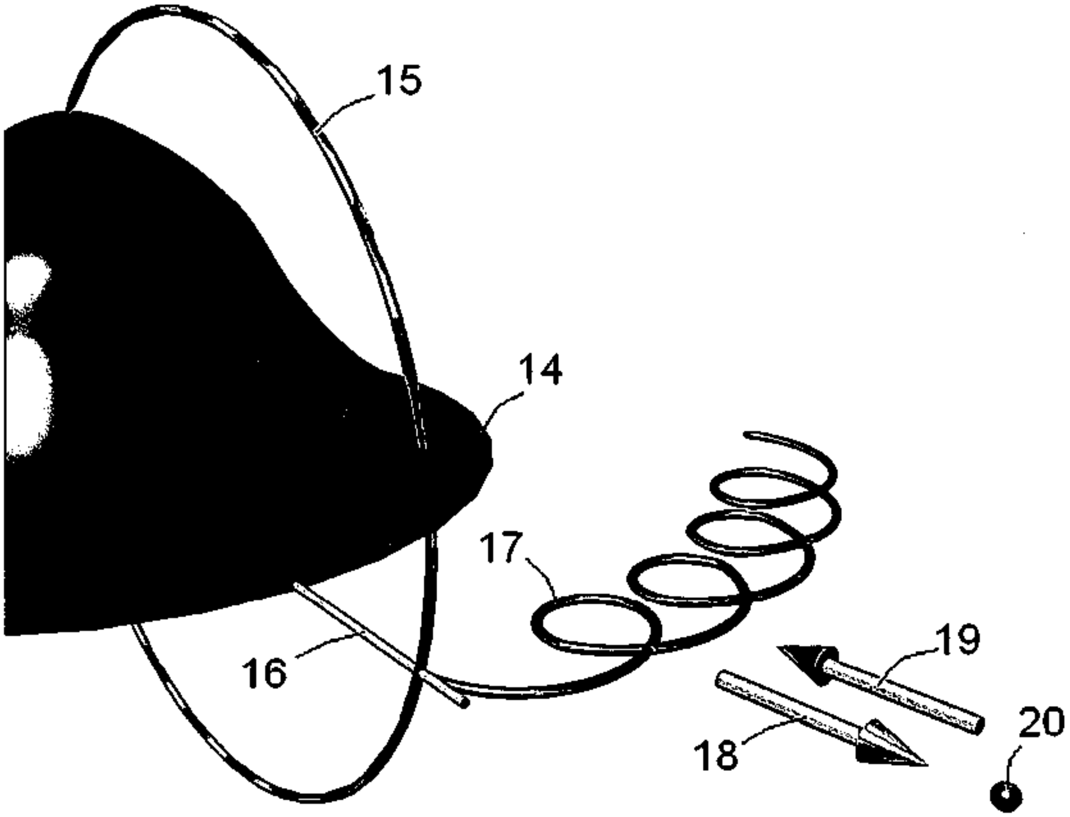
15
14
17
16
19
20
18

Figure 9

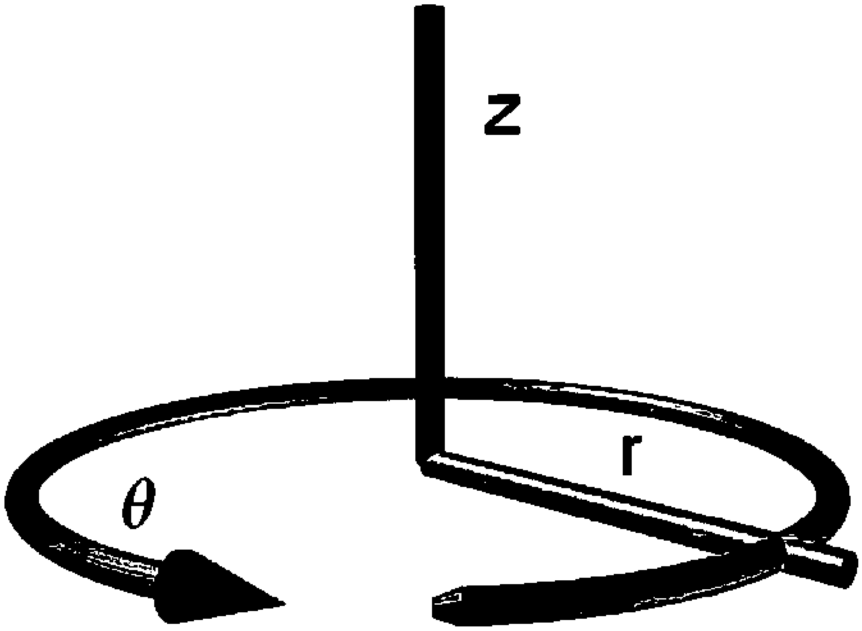

PHOTON SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a spacecraft propulsion system that employs photon particles to generate a field of negative energy in order to produce lift on the hull.

BACKGROUND OF THE INVENTION

[0002] Referring to **FIG. 1**, an electromagnetic wave traveling in the z-direction consists of an electric E field vibrating in the x-direction and a magnetic flux density B field vibrating at right angles in the horizontal y-direction. The energy-stress-momentum of this photon can be analyzed using Einstein's General Theory of Relativity and the Faraday F tensor. The Faraday tensor is a 4×4 matrix containing the electromagnetic wave components as shown here in general where c is the speed of light

$$F^{\alpha}_{\beta} = \begin{array}{c|cccc|} t & 0 & \frac{E_x}{c} & \frac{E_y}{c} & \frac{E_z}{c} \\ x & \frac{E_x}{c} & 0 & B_z & -B_y \\ y & \frac{E_y}{c} & -B_z & 0 & B_x \\ z & \frac{E_z}{c} & B_y & -B_x & 0 \end{array}$$

For this particular photon, this tensor is

$$F^{\alpha}_{\beta} = \begin{array}{c|cccc|} t & 0 & \frac{E_x}{c} & 0 & 0 \\ x & \frac{E_x}{c} & 0 & 0 & -B_y \\ y & 0 & 0 & 0 & 0 \\ z & 0 & B_y & 0 & 0 \end{array}$$

[0003] The elemental spacetime length ds squared is equal to sum of the squares of the Cartesian elemental lengths

$$(ds)^2 = -(dt)^2 + (dx)^2 + (dy)^2 + (dz)^2$$

The coefficients of this equation, {-1,1,1,1} are the diagonal components of the g metric tensor

$$g_{\alpha\beta} = \begin{array}{c|cccc|} t & -1 & 0 & 0 & 0 \\ x & 0 & 1 & 0 & 0 \\ y & 0 & 0 & 1 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

The stress-energy-momentum tensor T can then be calculated for the photon using the Faraday tensor and the g metric tensor in the following equation from gravitation physics

$$4\pi T^{\mu\nu} = F^{\mu\alpha} F^{\nu}_{\alpha} - \frac{1}{4} g^{\mu\nu} F_{\alpha\beta} F^{\alpha\beta}$$

The stress-energy-momentum tensor indicates the curvature of space due to the application of electromagnetic fields, mass, angular momentum and charge. The mass of the Earth, for example, generates a negative curvature of spacetime such that objects fall toward the mass. The T tensor, which is also a 4×4 matrix, contains the momentum or flux terms in the first row and first column. The normal pressure stress is located along the diagonal. The shearing stresses are located off the diagonal. The energy term is in the upper left corner as depicted here,

$$T^{\mu\nu} = \begin{array}{c|cccc|} t & \text{energy} & flux_x & flux_y & flux_z \\ x & -flux_x & pressure_x & shear_{xy} & shear_{xz} \\ y & -flux_y & shear_{yx} & pressure_y & shear_{yz} \\ z & -flux_z & shear_{zx} & shear_{zy} & pressure_z \end{array}$$

[0004] Since $B^2 = E^2/c^2$, the stress-energy-momentum tensor for the photon is therefore

$$T^{\mu}_{\nu} = \begin{array}{c|cccc|} t & \frac{-E^2}{c^2} & 0 & 0 & \frac{+E^2}{c^2} \\ x & 0 & 0 & 0 & 0 \\ y & 0 & 0 & 0 & 0 \\ z & \frac{-E^2}{c^2} & 0 & 0 & \frac{+E^2}{c^2} \end{array}$$

This remarkable result shows that the photon is actually a negative energy particle (top left corner) which is pushed along by a positive pressure wave (lower right corner). The particle has a positive flux (upper right corner) in the z-direction, as well as a balancing negative flux in the lower left corner so that the overall momentum of the universe remains the same. All four components cancel and we see the photon as a massless particle moving at the speed of light.

[0005] Thus the key idea behind this invention is that it is possible to cancel out the pressure term and leave a stationary vibrating electromagnetic field of negative energy over the hull of the spacecraft. The importance of negative energy is that it is a prerequisite to generating wormholes between space and hyperspace.

[0006] Hyperspace consists of the those co-dimensions which have different physics constants such as a low speed of light. The existence of hyperspace, which has a white misty look, is not a well-known scientific concept. Experiments with our magnetic vortex wormhole generators, hyperspace torque generator, full body levitation using Chi Kung breathing, arm levitation by spinning the co-gravitational K field, full body teleportation through hyperspace a distance of 100 meters using a pulsed gravitational wave, jumping into hyperspace, having a plate of toast enfold off the breakfast table and disappear into thin air, walking through walls and doors out-of-dimension, looking into other dimensions, remote viewing through subspace to distances of 100,000 light years, and other electromagnetic experiments carried out by co-researchers, have shown us the reality and existence of hyperspace.

[0007] Referring to **FIG. 2**, the spacecraft consists of an upper (**1**) and lower (**2**) hull attached by ceramic insulators to a circular ring (**3**). The ring provides support and is attached to an outer sharp-edged rim which is electrostatically charged to a potential −V. The purpose of the charged rim is to generate a radial electric E field around the vehicle.

[0008] Referring to **FIG. 3**, the radius of the ring (**4**) is equal to a. The distance from a point on the ring to the z-axis is r. The potential on the z-axis is therefore the charge divided by the distance,

$$potZ = \frac{q}{\sqrt{a^2 + r^2}}$$

This potential is expanded as a series in terms of inverse radius r

$$potZout = \frac{35qa^8}{128r^9} - \frac{5qa^6}{16r^7} + \frac{2qa^4}{8r^5} - \frac{qa^2}{2r^3} + \frac{q}{r}$$

The potential outside the ring can be written in terms of the Legendre polynomials P

$$Vout = \sum_{n=0}^{s} \left(\frac{a}{r}\right)^{n+1} A[n] LegendreP[n, \mathrm{Cos}(\theta)]$$

where s is the number of terms in the expansion. By equating the known particular solution potZout on the z-axis with the general Vout solution, the coefficients A[n] are found to be

$$A(0) = \frac{q}{a}$$

$$A(1) = 0$$

$$A(2) = \frac{-q}{2a}$$

$$A(3) = 0$$

which are substituted back into the Vout equation to get the potential outside the ring.

[0009] Referring to **FIG. 4**, the potential (dotted lines **6**) looking at a slice through the ring (**5**) is shown together with the electric E field. The negative gradient of the potential is the electric field (**7**) shown by the direction of the arrows. The importance of this diagram is that the electric field points in the radial direction toward the negatively charged ring. The force on an electron is the electron charge times the electric field

$$F = q_e E_r = -|q_e|(-|E_r|) = +F$$

Because the electron charge is negative and the radial field points in the negative direction toward the ring, the force on the electron is positive. Thus the electron moves away from the ring in the positive radial direction. A 3-dimensional plot of the ring (**8**) and the electric field (**9**) is shown in **FIG. 5**.

[0010] The stress-energy-momentum generated by a radial electric field is calculated using the Faraday F tensor

$$F_{\alpha}^{\beta} = \begin{vmatrix} 0 & E_r & 0 & 0 \\ E_r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{vmatrix}$$

The g metric tensor has to be given in spherical coordinates {r,θ,ϕ}

$$g_{\alpha\beta} = \begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin(\theta)^2 \end{vmatrix}$$

where θ is the angle from the vertical to the radius r. The stress tensor T^{rr} along the radial direction is

$$T^{rr} = \frac{E_r^2}{8\pi c^2}$$

which shows that the pressure is negative along the radial line equal to the square of the radial electric field divided by the square of the speed of light. Because the field is squared, it doesn't matter that the electric field points in the negative direction. The square makes it positive, but the overall curvature pressure is negative. Thus this negative pressure cancels out the positive pressure propelling the photon along. The second key idea of the invention is how to generate this photon moving in the radial direction.

[0011] It has been known for a long time in physics that an electron moving in a circular path will emit photons in a process known by the German word Bremsstrahlung which is translated as "breaking radiation." There are several types of radiation such as classical Bremsstrahlung involving a charged particle making a collision with another charged or uncharged particle in which photons are emitted. The quantum mechanical Bremsstrahlung involves the sudden appearance or disappearance of a charged particle which also emits radiation. In space, having a field of wormholes in which the electrons are spiraling down into hyperspace would result in the emission of photons by the quantum mechanical method. Also, in the atmosphere, having collisions with air molecules results in emission of photons in the classical way.

[0012] In order to get the electrons to spiral around and emit photons, a crossed electromagnetic field is used as shown by the following equation

$$F = q(E_r + v_r \times B_\theta)$$

where the velocity v is in the positive radial direction due to the force of the electric field. The velocity crossed with a magnetic flux density B field in the θ-direction makes the electron move sideways back and forth in a wiggling motion.

[0013] Referring to **FIG. 6**, a direct current solenoid (**10**), represented by multiple current loops, running vertically through the center of the hull, generates a magnetic field that curves around the outside of the hull, as shown by contour lines (**12**). The north pole (**11**) is at the bottom of the hull. A radial arrow (**13**) from the electrostatically-charged rim is perpendicular to the magnetic field lines. The cross product in the force equation becomes the electron radial velocity times the magnetic field v_r B_θ.

[0014] Referring to **FIG. 7**, the electric field is in the y-direction and the magnetic field is in the z-direction. The flat looping path in the x-direction is the motion of the electron. The electron, which has a negative charge, starts to move in the direction opposite to that of the electric field. In this particular diagram, the electron acquires a velocity in the negative y-direction. Then a sideways force in the x-direction is produced due to the cross product of the velocity with the magnetic field times the negative charge

$$-q(-v_y \times B_z) = +F_x$$

Depending on the magnitude of the velocity, various size loops can be produced.

[0015] In terms of the hull coordinates, because the flat loop is in the plane of the electric field which points in the radial direction, the electron emits light in the radial direction. This condition means that the negative radial pressure created by the electric field cancels the radial pressure of the photon. Thus the photon becomes a stationary vibrating quantum of negative energy. This has the appearance of a luminescent light source. The stress tensor for this condition is therefore

$$T_\mu^\nu = \begin{vmatrix} -\frac{E^2}{c^2} & 0 & 0 & \frac{E^2}{c^2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{E^2}{c^2} & 0 & 0 & 0 \end{vmatrix} = -\frac{E^2}{c^2} \text{residual negative energy}$$

[0016] residual negative energy

which leaves a residual negative energy per photon.

[0017] Referring to **FIG. 8**, the negatively charged rim (**14**) produces a radial electric field (**16**) that crosses the magnetic B field (**15**) of the solenoid. Electrons emitted by the charged rim then encounter this crossed field which makes them spiral (**17**) around the hull. Because of the tight loop, the electron emits Bremsstahlung radiation in the radial direction (**18**). The positive pressure field of the photon, which is directed in the radial direction, is canceled by the negative pressure field (**19**) created by the electric field. Because the photon energy is negative, a stationary vibrating electromagnetic quantum of negative energy (**20**) surrounds the hull.

[0018] This negative energy and the pressure stress created by the electromagnetic fields open up wormholes between space and hyperspace. The potential head is positive from hyperspace into space because the energy of hyperspace is more positive than the negative energy field. The low-density hyperspace energy fills the hull and its surrounding space with a white misty hyperspace energy which makes the spacecraft lighter in mass, and therefore lighter in weight within a gravitational field. The actual physics is more complicated still because the electrons find that the resistance of hyperspace is lower than the resistance of space. Thus they spiral down the wormholes which results in a sudden disappearance of charge. The quantum mechanical effect of this is to radiate even more photons which in turn produce even more negative energy.

[0019] The lift on the hull is generated by the radial electric field. In cylindrical coordinates, the g metric tensor is

$$g_{\alpha\beta} = \begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

Using this metric tensor, the pressure stress in the vertical direction T^{zz} is

$$T^{zz} = \frac{E^2}{8\pi c^2}$$

which is a positive curvature over the hull. The mass of Earth produces a negative curvature in which objects fall toward the mass. By counteracting this negative curvature with a more than positive curvature, lift is developed on the spacecraft. Because the negative energy lowers the effective mass of the vehicle, the acceleration is large with a modest electric field. Moreover, in our dimension, the speed of light is 299792458 meters per second. Hyperspace energy has a speed of light equal to one meter per second. Thus the stress is amplified by a factor of

$$A = \left(\frac{299792458\,m/s}{1\,m/s}\right)^2 \approx 9 \cdot 10^{16}$$

Because electromagnetic fields are relativistic, motion in a low-velocity-of-light energy field amplifies their strength.

SUMMARY OF THE INVENTION

[0020] It is the object of this invention to create a spacecraft propulsion system that produces wormholes between space and hyperspace using negative energy in order to generate lift on the hull. It was discovered in the Riemannian curvature calculations of gravitation physics that negative energy is required to keep open the throat of the wormhole. From experiments with the magnetic vortex wormhole generator, it is known that the proper combination of electromagnetic fields, together with this negative energy, can create a wormhole through which smoke can be blown into hyperspace.

[0021] Referring to **FIG. 9**, the directions of force, velocity, and electromagnetic fields are referred to in the cylindrical coordinate system {r,θ,z}. An electrostatically charged sharp-edged ring in the θ-direction around the hull of the spacecraft produces a radial electric field. A vertical solenoid in the z-direction through the center of the hull produces a magnetic field which is perpendicular at the rim to the electric field. With the current in the solenoid flowing in the clockwise (−θ) direction, using the right-hand rule, the magnetic field points in the upward z-direction outside the rim. Because the rim is charged to a negative voltage, the electric field points toward the hull in the negative radial (−r) direction. Electrons emitted by the rim travel outward (+v) because the charge on the electron is negative which, together with the negative electric field, produces a positive radial force. The radial force on the electron causes it to acquire a velocity which interacts with the magnetic field. The cross product of the velocity (+v) with the positive (+B) magnetic field produces a sideways force on the electron in the negative θ-direction. However, because the charge on the electron is negative, the force is

$$F=-q\{v_r,0,0\}\times\{0,0,B_z\}=\{0,qB_z v_r,0\}$$

which is positive in the θ-direction. It is this sideways force that produces a flat spiraling or looping motion whereby the electron emits photons, known in German as Bremsstahlung radiation, in the radial direction. The photon, which is actually a quantum of negative energy, has a positive radial pressure which propels it along. Because the radial electric field produces a negative pressure in the radial direction, the two opposite fields cancel in the radial direction to form a residual stationary vibrating negative energy. Thus the hull becomes surrounded by negative energy which, together with the pressure stresses created by the electric field, generates wormholes between space and hyperspace.

[0022] The gravitational potential between hyperspace and space is positive because the hyperspace energy is more positive than the negative energy around the hull. Thus the low-density, low-speed-of-light hyperspace energy flows through the wormhole and fills the hull. This has the effect of reducing the effective mass of the hull. Because the electric field generates a positive pressure over the hull in the vertical z-direction, there is an upward force on the vehicle due to the pressure times the hull area. Since the vehicle has a low mass, there is a modest upward acceleration on the spacecraft equal to the force divided by mass.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0023] **FIG. 1**. Perspective view of an electromagnetic wave.

[0024] **FIG. 2**. Perspective view of spacecraft.

[0025] **FIG. 3**. Perspective view of charged ring.

[0026] **FIG. 4**. Planar plot of the radial electric field produced by charged ring.

[0027] **FIG. 5**. Perspective view of radial electric field around ring.

[0028] **FIG. 6**. Planar view of magnetic flux density field contour lines.

[0029] **FIG. 7**. Perspective view of electron motion in crossed electric and magnetic fields.

[0030] **FIG. 8**. Perspective view of production of negative energy around hull.

[0031] **FIG. 9**. Perspective view of cylindrical coordinate system {r,θ,z}.

DETAILED DESCRIPTION OF THE INVENTION

[0032] 1. The hull is made from a single sheet of aluminum which has been stretched to its yield point by hydraulic cylinders. An upper and lower die is CNC machined to the profile of the hull. The soft sheet is then clamped in the die where it takes on the smooth shape of the hull without any wrinkles. The hull is extremely rigid after forming and does not require any structural reinforcements.

[0033] 2. A section of the aluminum ring is made in a 3D computer graphics program. The model is stored as a stereolithography file (*.stl). The computer model is then sent via Internet e-mail to the stl server who prints the part in an ultraviolet light-cured polymer. The part is returned the next day by Express Mail. Using a rubber blanket mold to create several ring sections, the entire ring is assembled together in another wooden mold box having thin circular laminate-coated particulate wall boards on either side of the ring. Then a liquid rubber mold is poured on top of the ring and allowed to harden overnight at room temperature. Since the rubber mold is flexible, the ring can be extracted fairly easily. This ring model is then sent to the foundry where it is cast in aluminum using the lost wax process in which a wax mold evaporates out of the sand casting. We are also experimenting with non-magnetic copper casting metals containing beryllium having good conductivity.

[0034] 3. A 11.5 cm plastic pipe is mounted on a rotating fixture driven slowly by a microcontoller, stepper motor, and power electronics board. Using a large diameter insulated wire, such as a 17 AWG with a wire diameter of 0.127 cm, the wire is wound slowly on the pipe and expoxied so that the windings don't come loose. The solenoid is then mounted vertically in the hull supported by the support ring and driven by a current generator located nearby on the test rig.

[0035] 4. The ring is driven by a high voltage electrostatic generator similar to the night vision scope high voltage power supplies. The ring charge is isolated from the hull by ceramic insulators.

I claim:

1. A spacecraft propulsion system comprising the components:

an aluminum horizontal circular structural support ring;

an aluminum hull in the shape of a high dome on top and shallow dome on the bottom attached to the circular support ring using ceramic insulators;

an electrostatically negatively-charged sharp-edged circular ring, preferably of non-magnetic aluminum or copper, attached with ceramic insulators to the outside of the support ring;

a solenoid mounted through the center of the hull in the vertical direction and attached to the center of the support ring;

an electrostatic high-voltage generator to drive the outer electrostatic ring; and

a direct high-current generator to drive the solenoid.

2. The method of claim 1, wherein a negative radial electric field is generated around the hull by placing a negative potential on the sharp-edged electrostatic ring using the electrostatic generator.

3. The method of claim 1, wherein the current-driven solenoid generates a vertical magnetic field around the hull with the north pole of the solenoid facing down through the bottom of the hull which causes the magnetic flux density field to point up outside the rim.

4. The method of claim 1, wherein electrons are emitted radially by the sharp edge of the charged ring.

5. The methods of claims **2**, **3** and **4**, wherein the crossed electromagnetic fields cause the electrons to spiral around in flat loops during which photons are emitted in the radial direction.

6. The methods of claims **2** and **5**, wherein the negative radial pressure created by the electric field cancels the positive radial pressure of the photon to leave a residual quantum of negative energy per photon around the hull.

7. The methods of claims **2** and **6**, wherein the pressure stress created by the electric field, and the negative energy combine to form wormholes between space and hyperspace.

8. The method of claim 7, wherein low-density hyperspace energy of a higher gravitational potential flows through the wormholes to fill the hull and surrounding space around the hull with the effect of reducing the effective mass of the spacecraft.

9. The method of claim 2, wherein the electric field generates a positive pressure in the vertical direction over the hull which together with the hull surface area, generates an upward lift force on the hull.

10. The method of claim 6, wherein the negative energy, having a low light speed, amplifies the strength of the electromagnetic fields and pressure stress fields.

11. The methods of claims **4** and **7**, wherein the electrons spiral down the low resistance wormholes into hyperspace such as to create a sudden disappearance of electrical charge which quantum mechanically causes a large emission of additional photons.

* * * * *

US 20040200925A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2004/0200925 A1**

St.Clair (43) **Pub. Date: Oct. 14, 2004**

(54) CAVITATING OIL HYPERSPACE ENERGY GENERATOR

(76) Inventor: **John Quincy St.Clair**, San Juan, PR (US)

Correspondence Address:
John St.Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/372,557**

(22) Filed: **Feb. 21, 2003**

Publication Classification

(51) **Int. Cl.^{7}** **B64C 27/22**
(52) **U.S. Cl.** **244/10**

(57) **ABSTRACT**

A hyperspace energy generator that uses cavitating oil bubbles within a magnetic field in order to create wormholes between space and hyperspace for the purpose of permeating the hull of a spacecraft with low-density hyperspace energy.

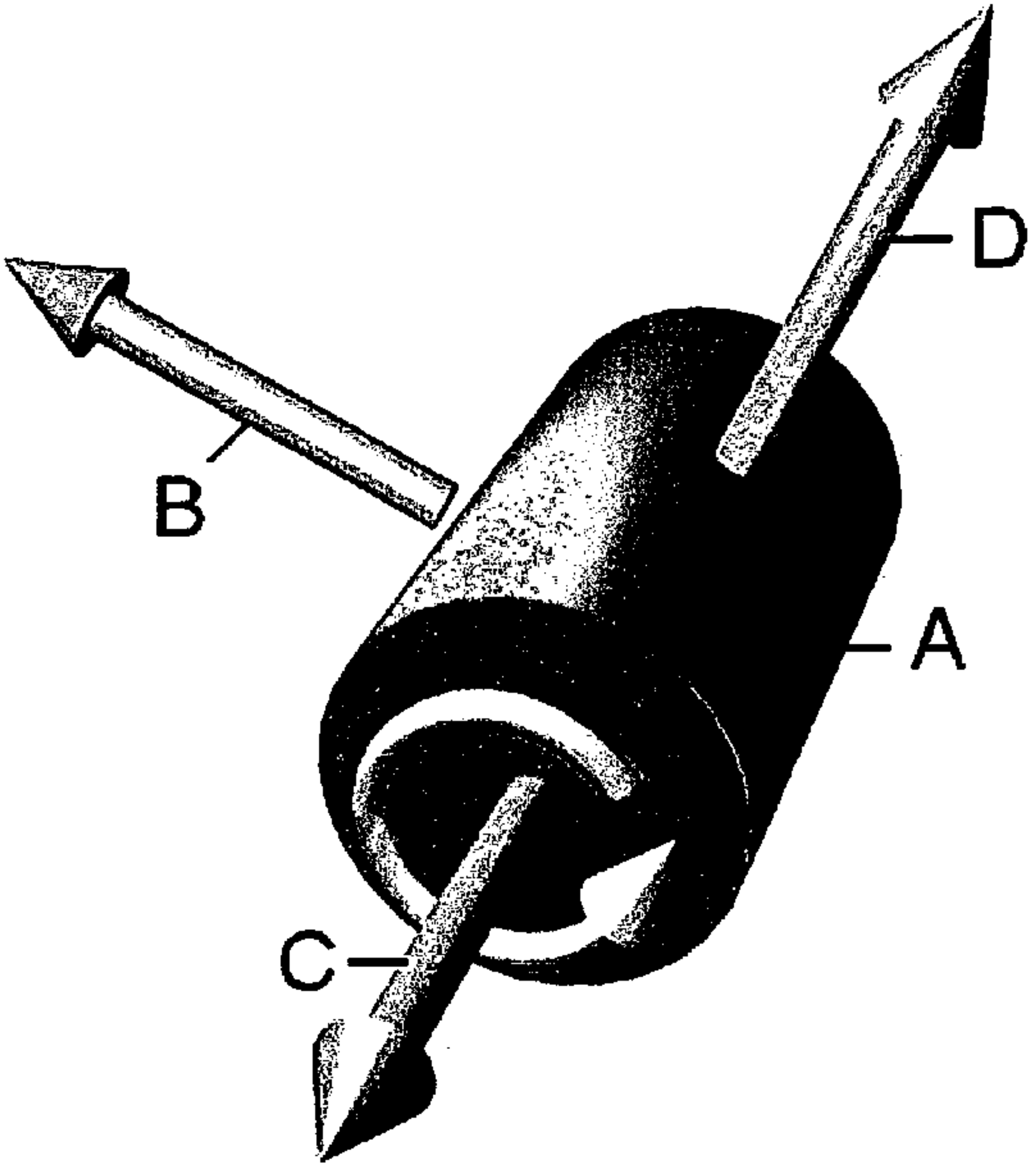

Figure 1

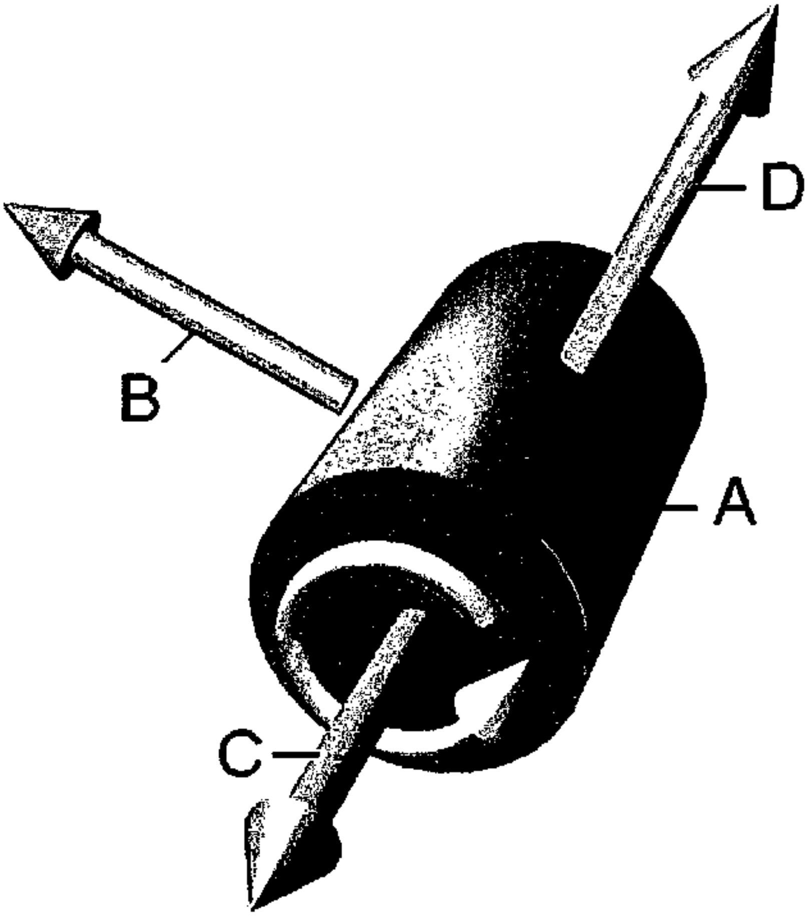

Figure 2

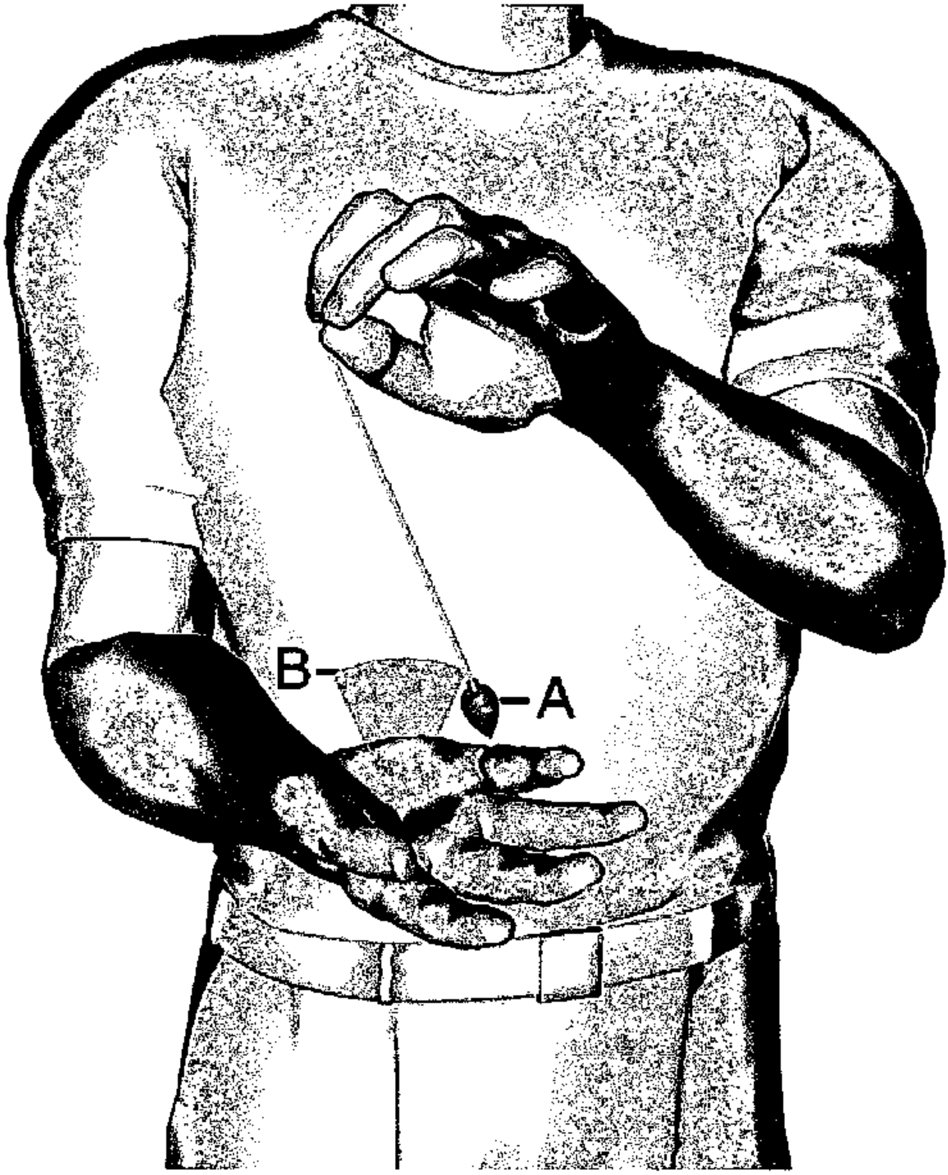

Figure 3

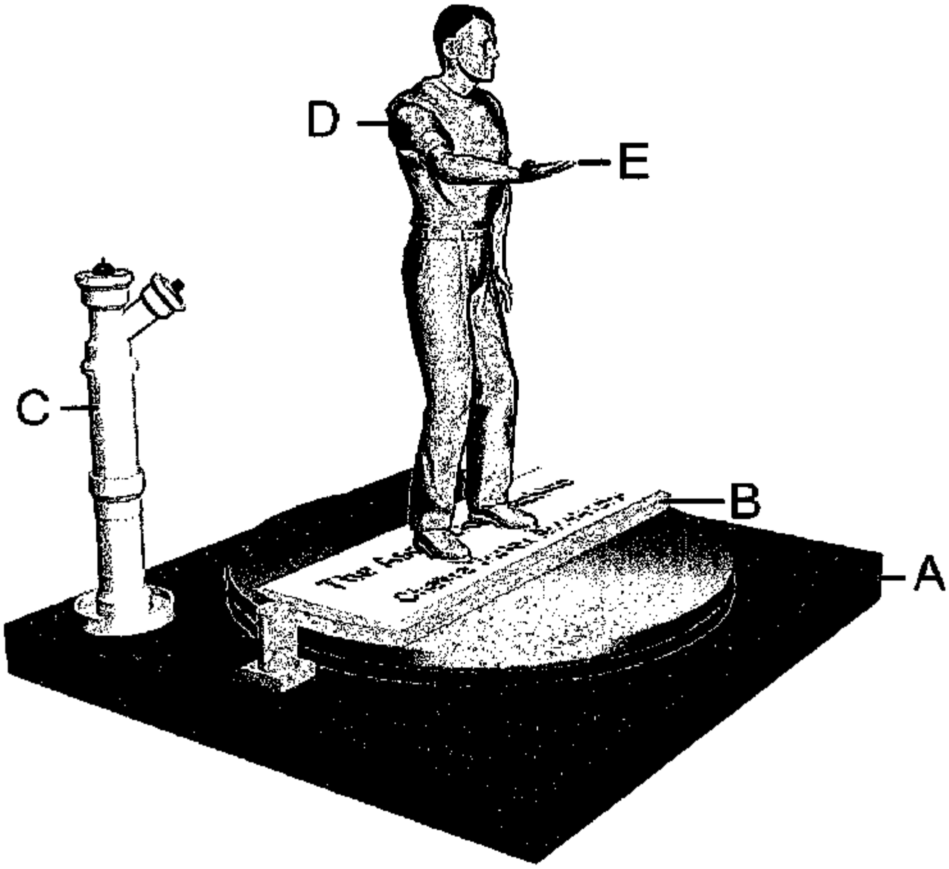

Figure 4

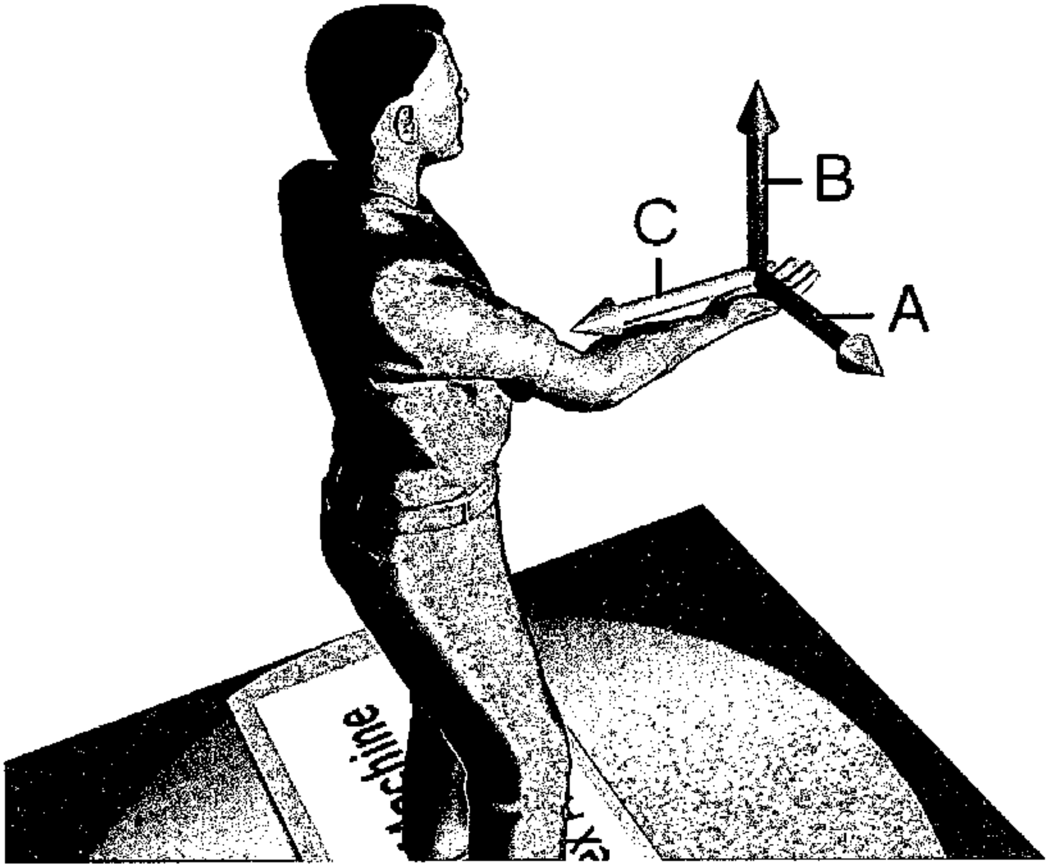
B
C
A
Machine

Figure 5

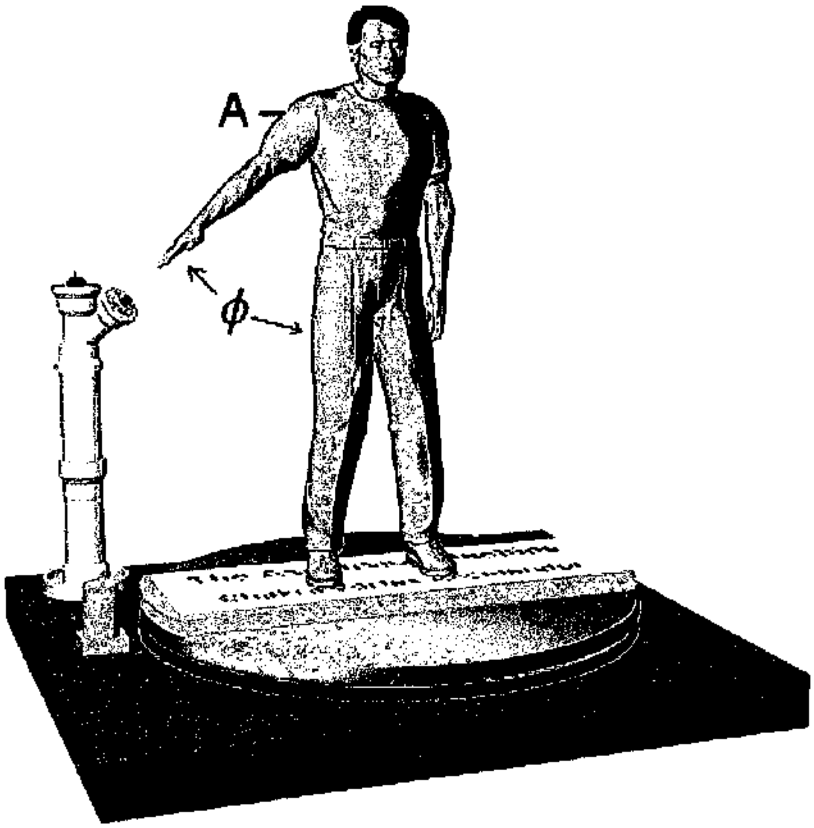

Figure 6

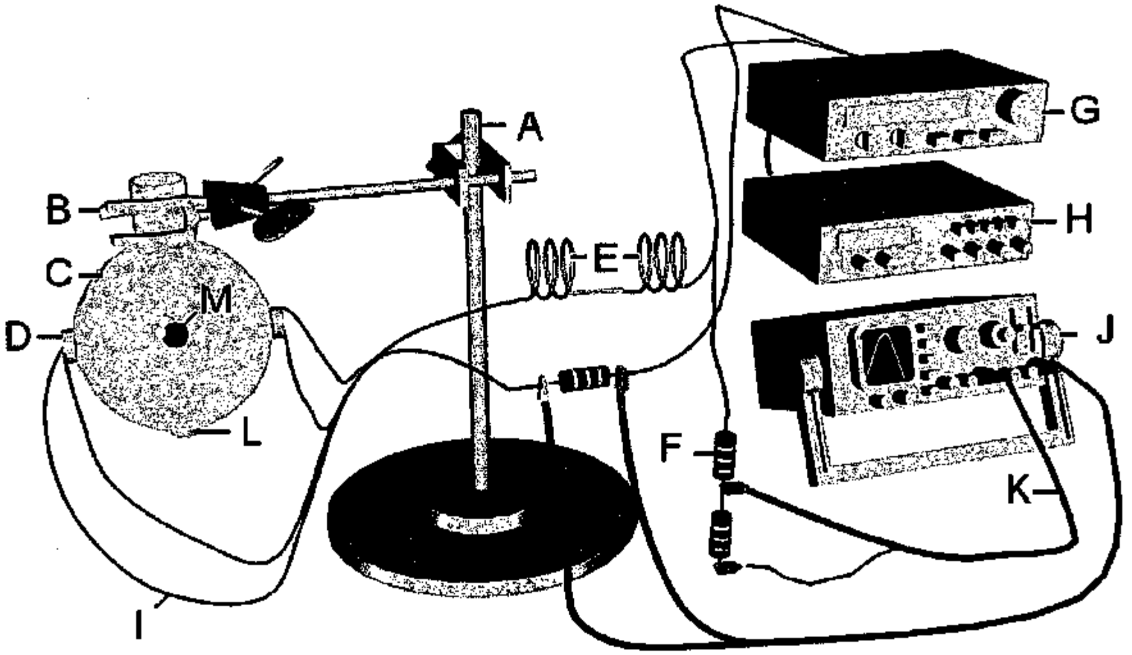
A
B
C
M
D
E
G
H
J
L
F
K
I

Figure 7

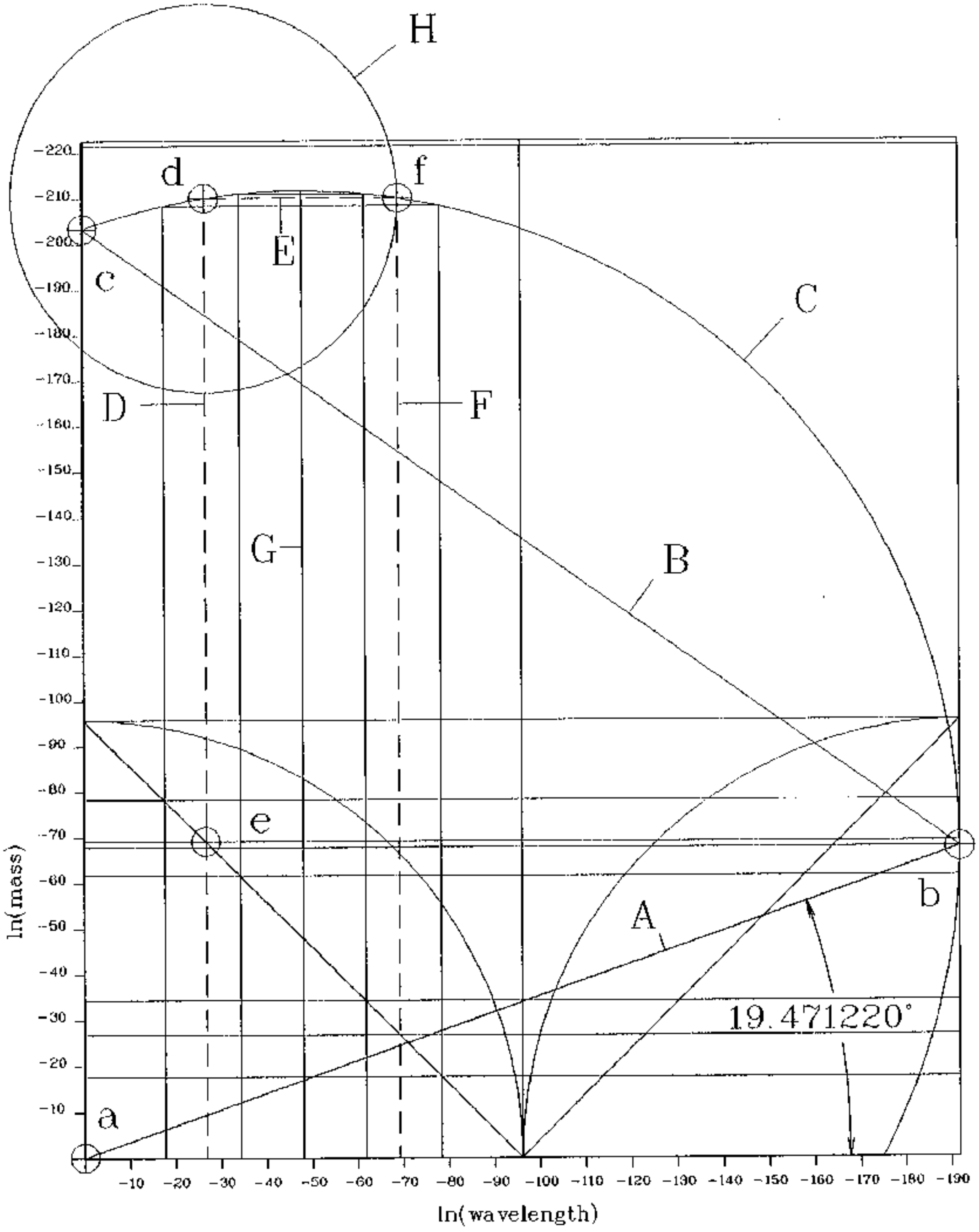
H
d
f
c
E
C
D
F
G
B
e
b
A
a
19.471220°
ln(mass)
ln(wavelength)
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190

Figure 8

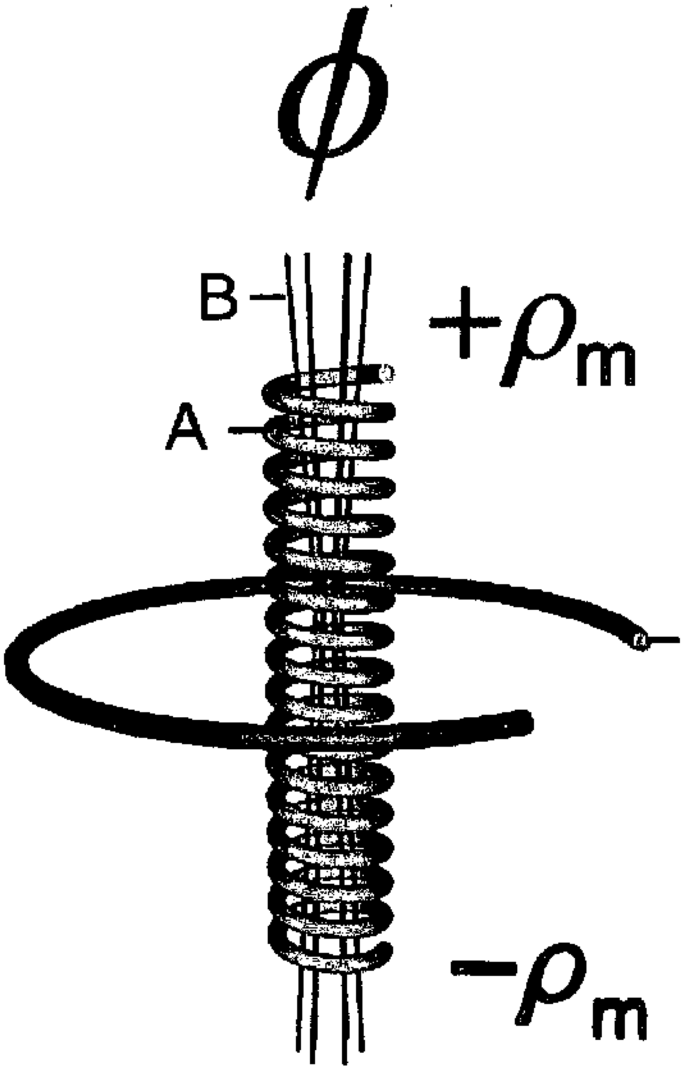
ϕ
B
$+\rho_m$
A
C
$-\rho_m$

Figure 9

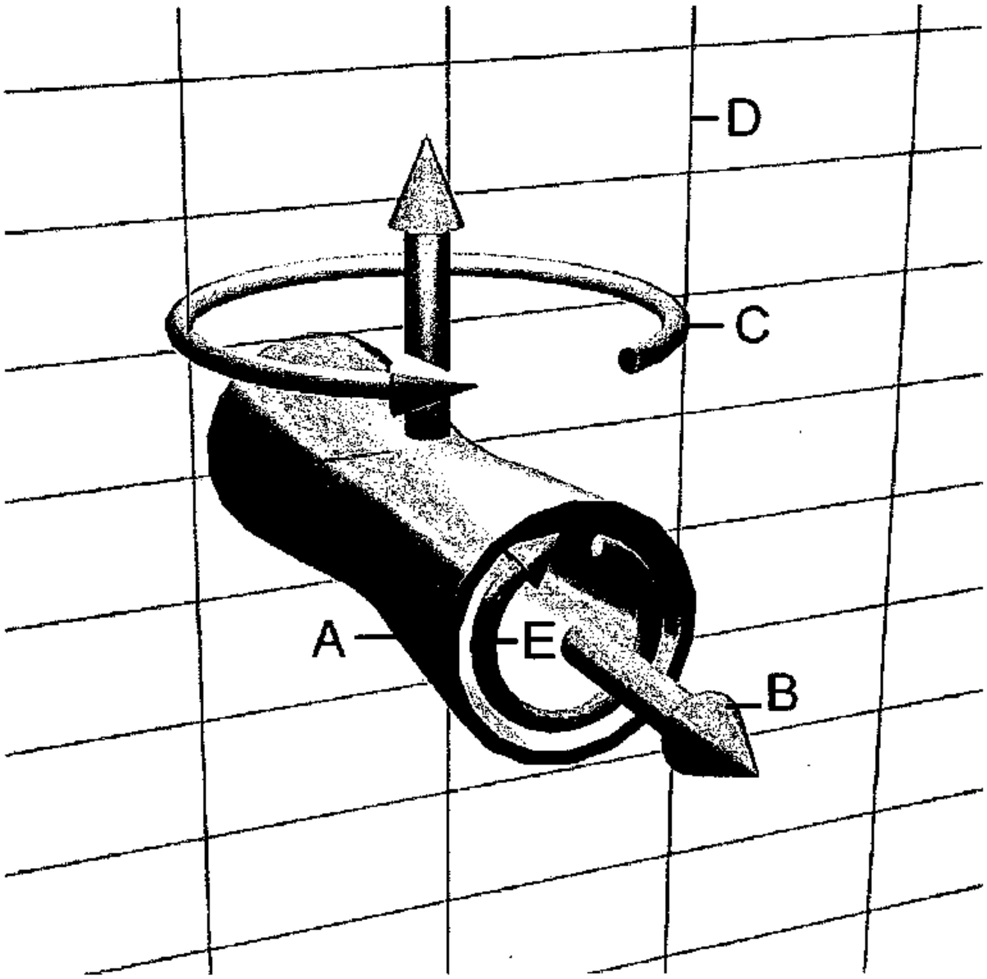

Figure 10

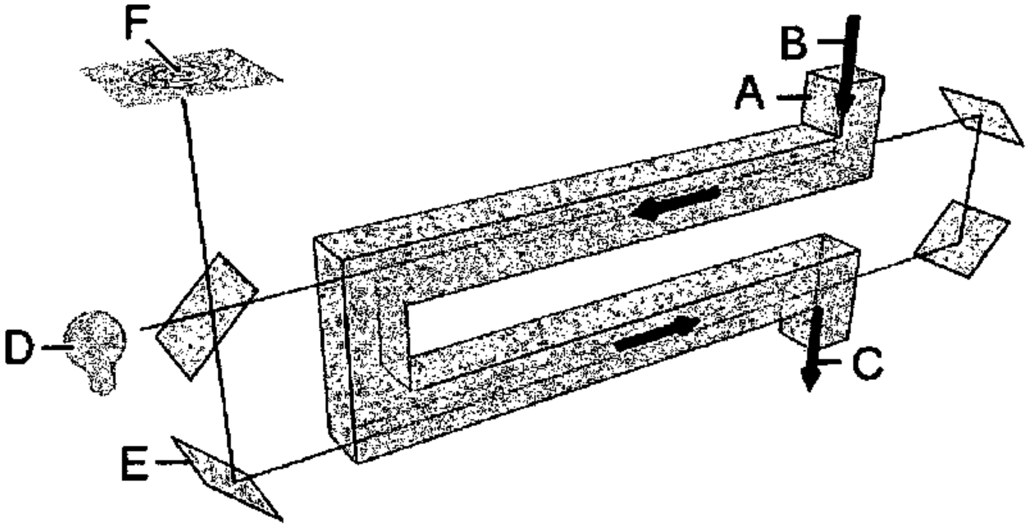

Figure 11

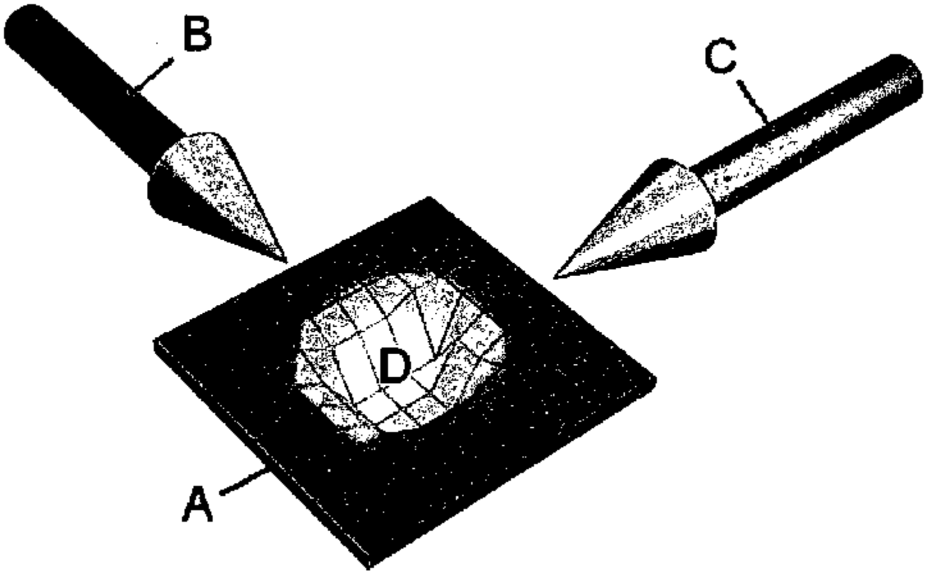

Figure 12

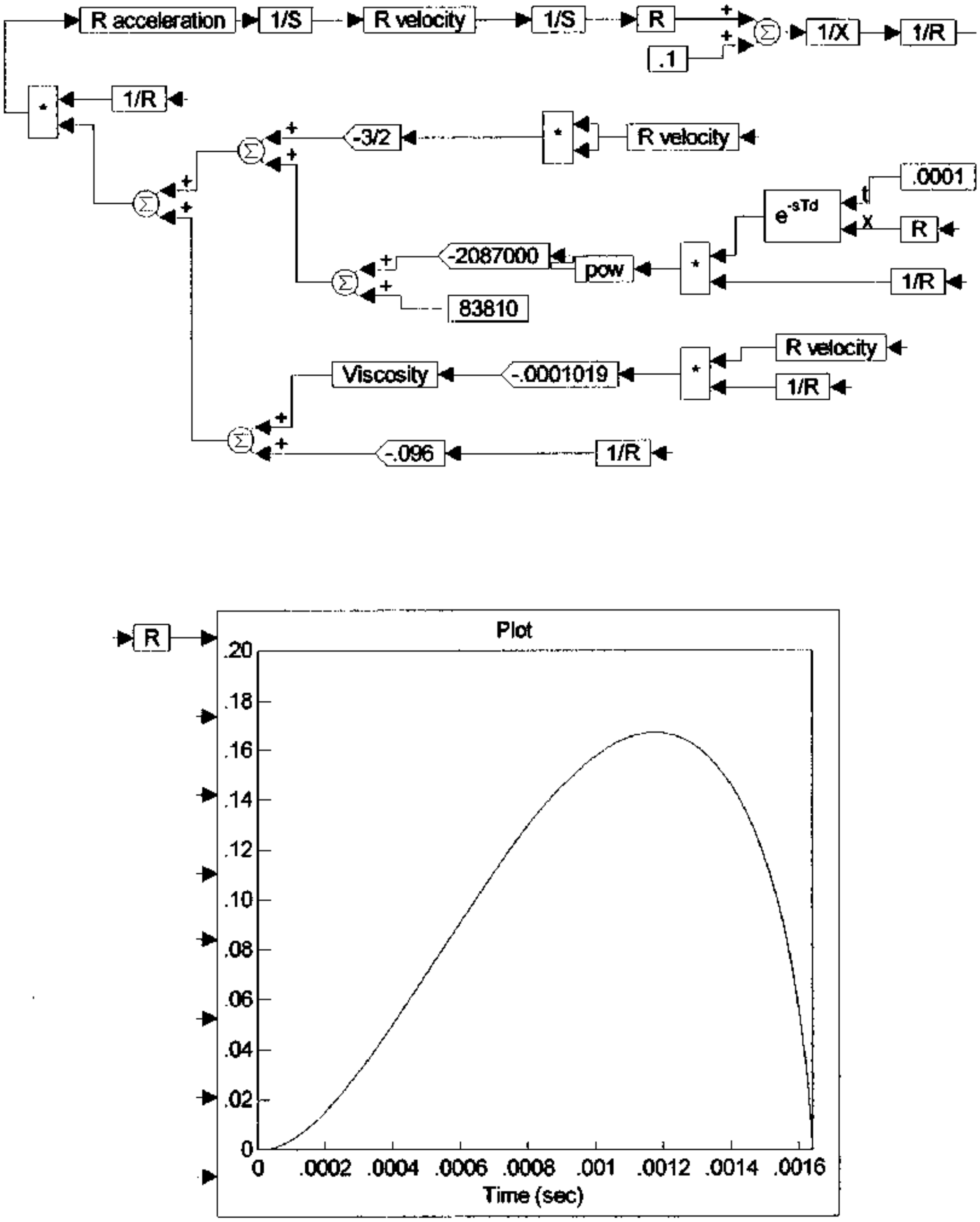
R acceleration
1/S
R velocity
1/S
R
.1
+
+
Σ
1/X
1/R
*
1/R
Σ
-3/2
*
R velocity
.0001
e^{-sTd}
t
x
R
-2087000
pow
*
1/R
Σ
83810
Σ
Viscosity
-.0001019
*
R velocity
1/R
Σ
-.096
1/R
R
Plot
.20
.18
.16
.14
.12
.10
.08
.06
.04
.02
0
0
.0002
.0004
.0006
.0008
.001
.0012
.0014
.0016
Time (sec)

Figure 13

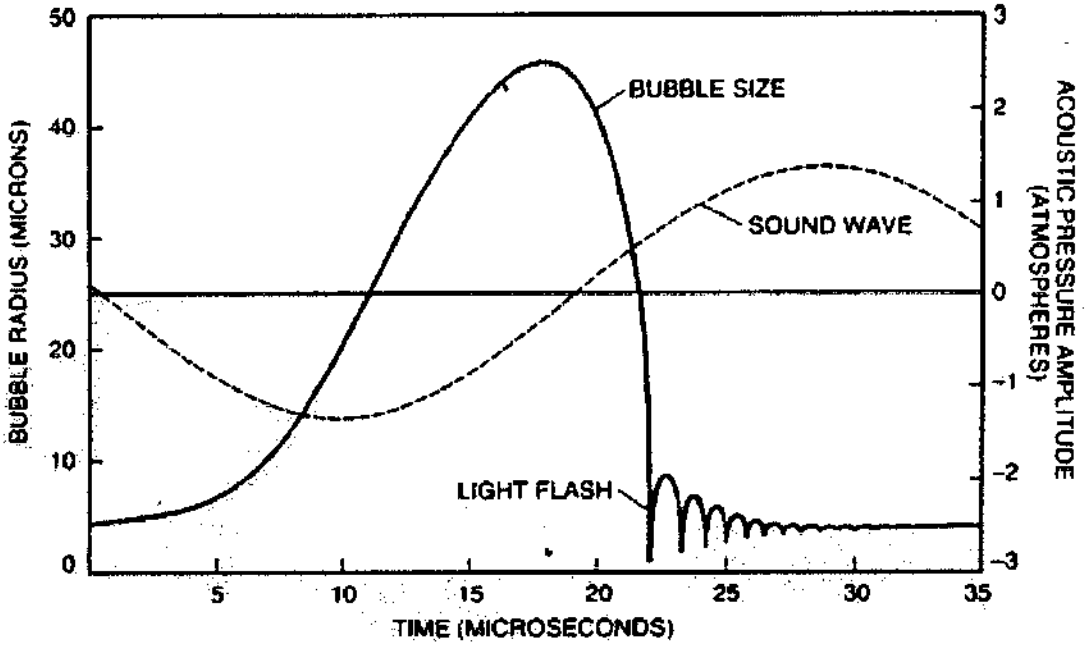
BUBBLE SIZE
SOUND WAVE
LIGHT FLASH
BUBBLE RADIUS (MICRONS)
50
40
30
20
10
0
ACOUSTIC PRESSURE AMPLITUDE (ATMOSPHERES)
3
2
1
0
-1
-2
-3
5
10
15
20
25
30
35
TIME (MICROSECONDS)

Figure 14

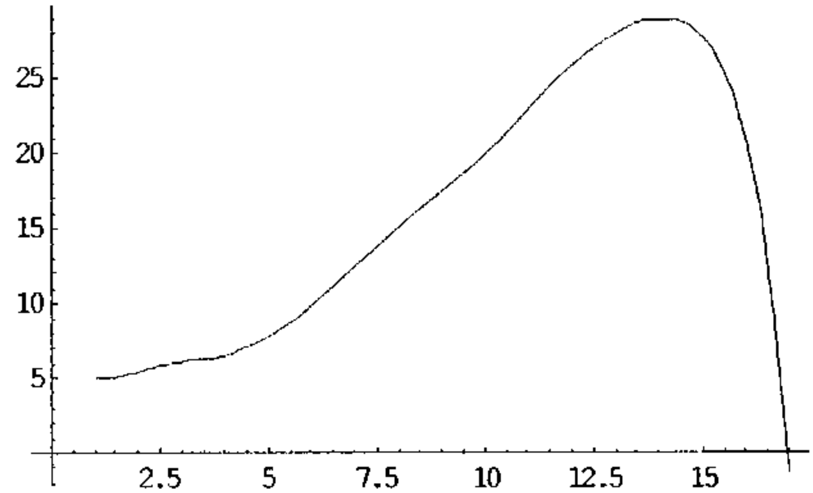

Figure 15

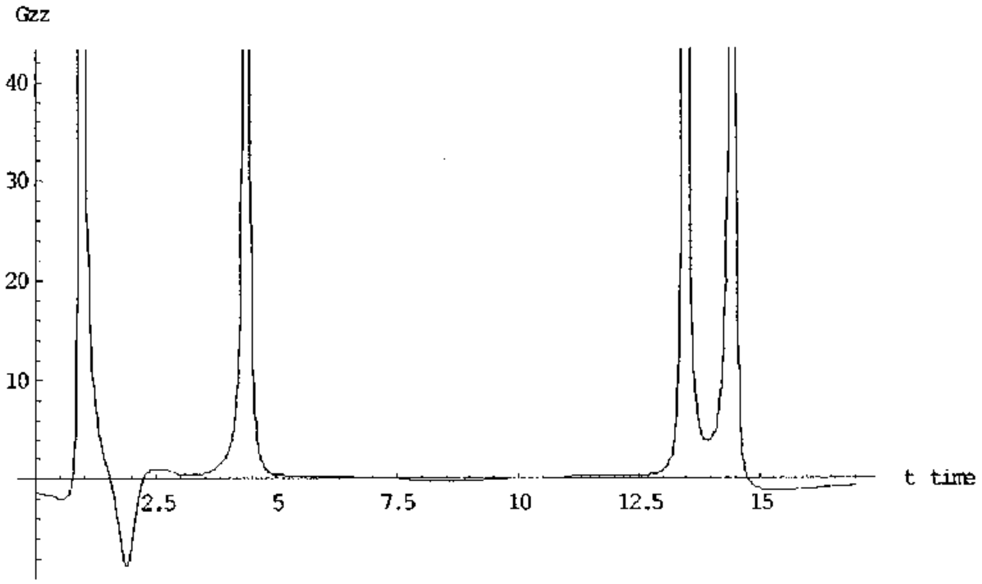

Figure 16

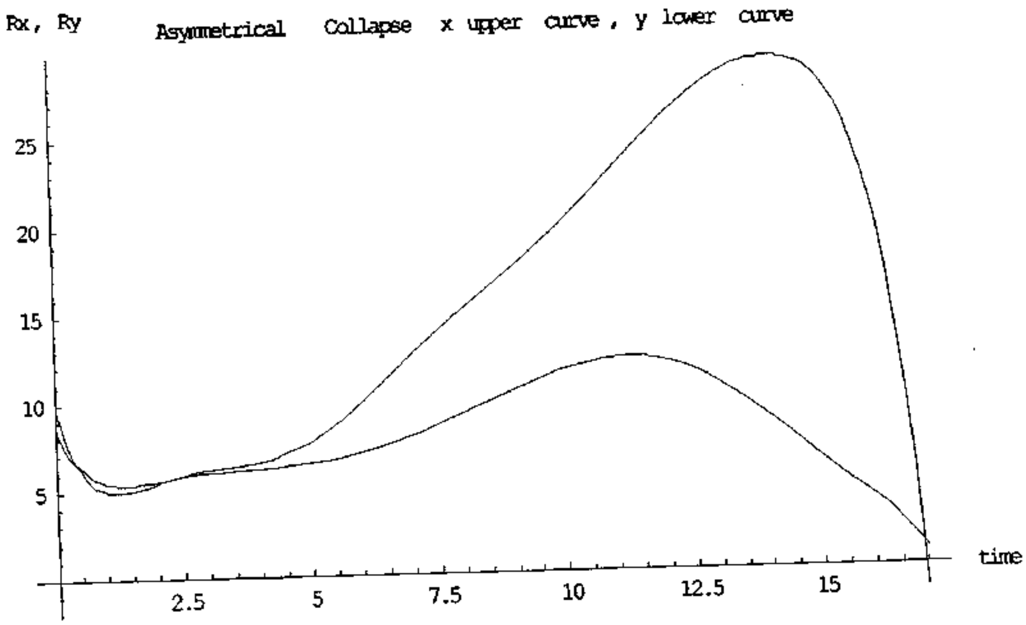
Rx, Ry
Asymmetrical Collapse x upper curve, y lower curve
25
20
15
10
5
2.5
5
7.5
10
12.5
15
time

Figure 17

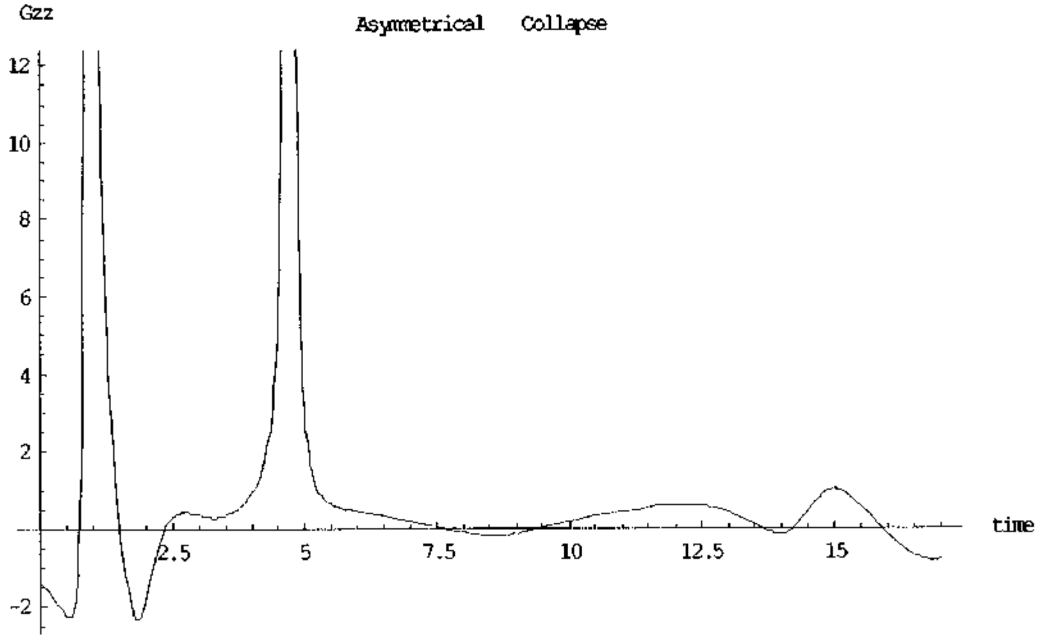

Figure 18

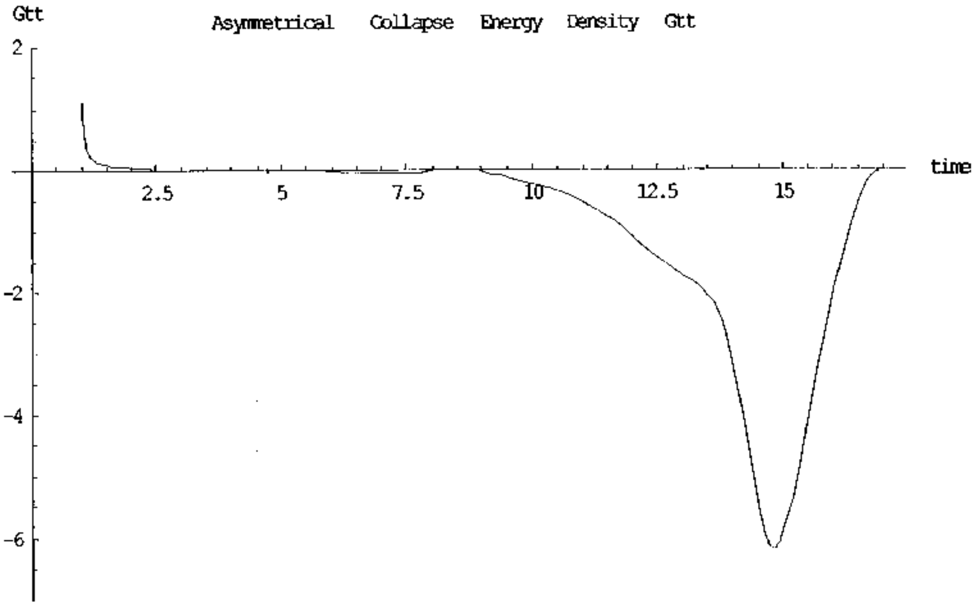
Gtt
Asymmetrical Collapse Energy Density Gtt
2
-2
-4
-6
2.5
5
7.5
10
12.5
15
time

Figure 19

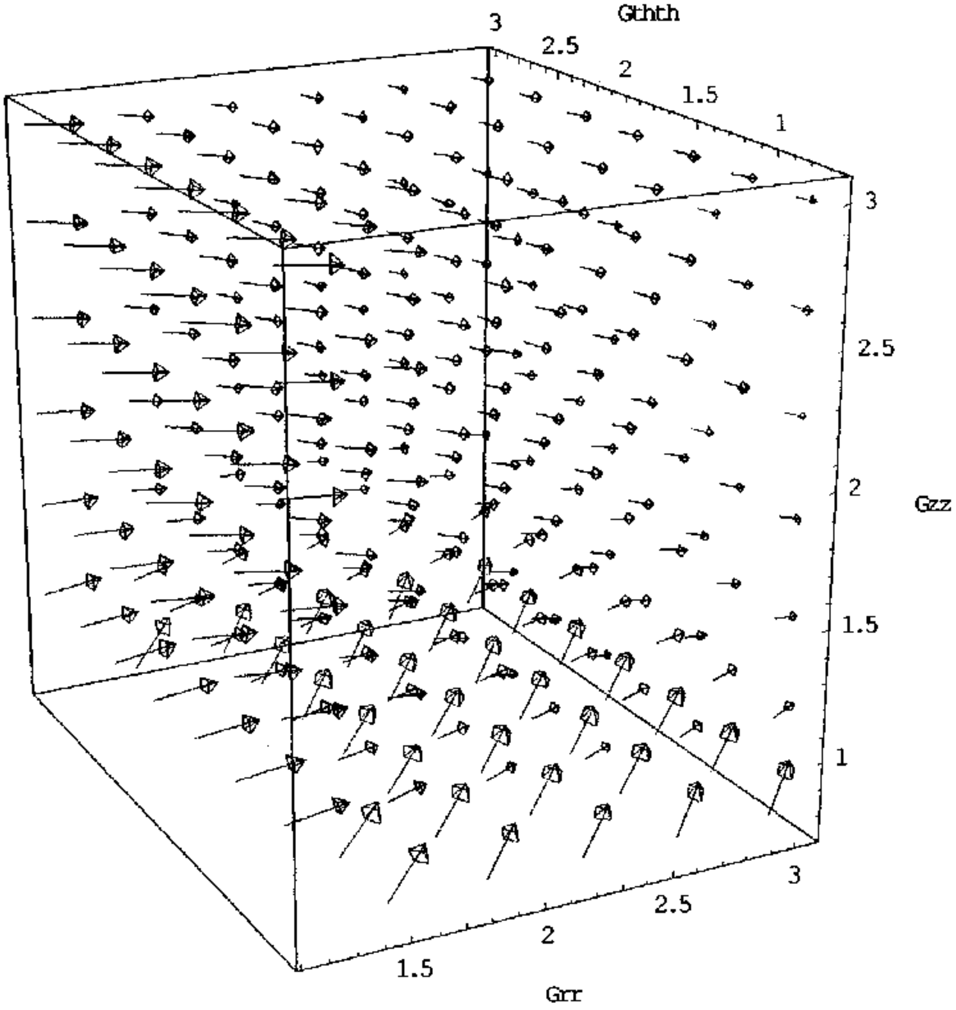

Figure 20

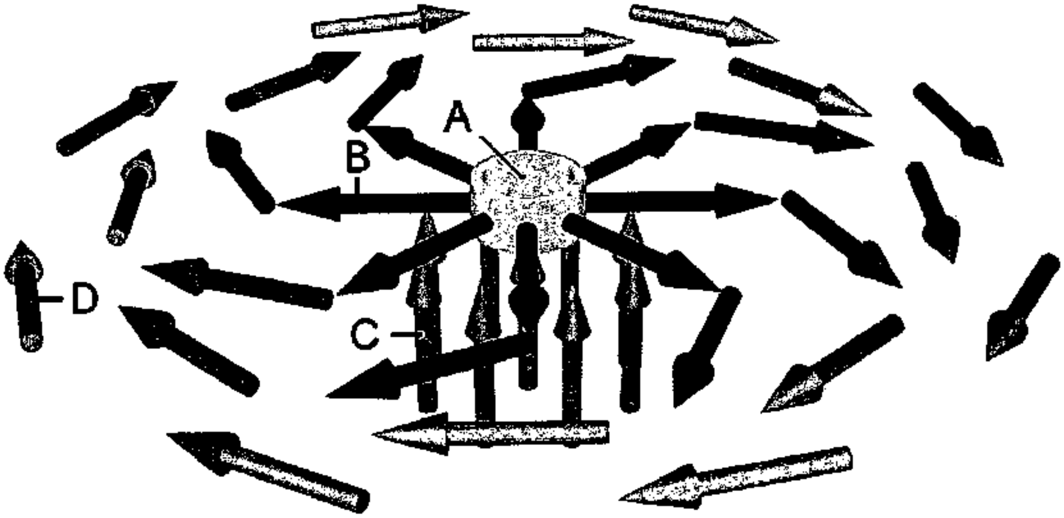

Figure 21

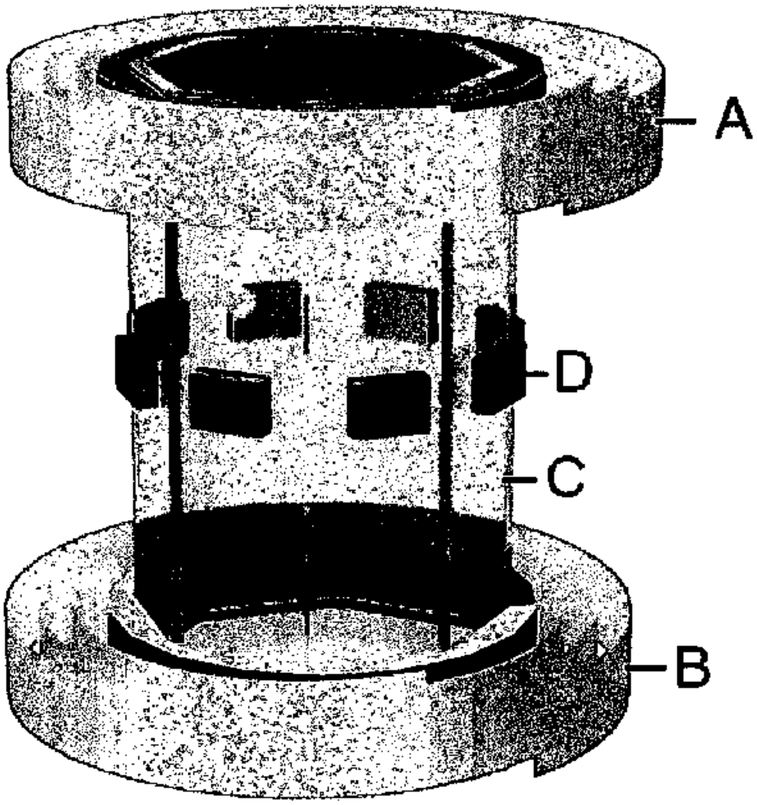

Figure 22

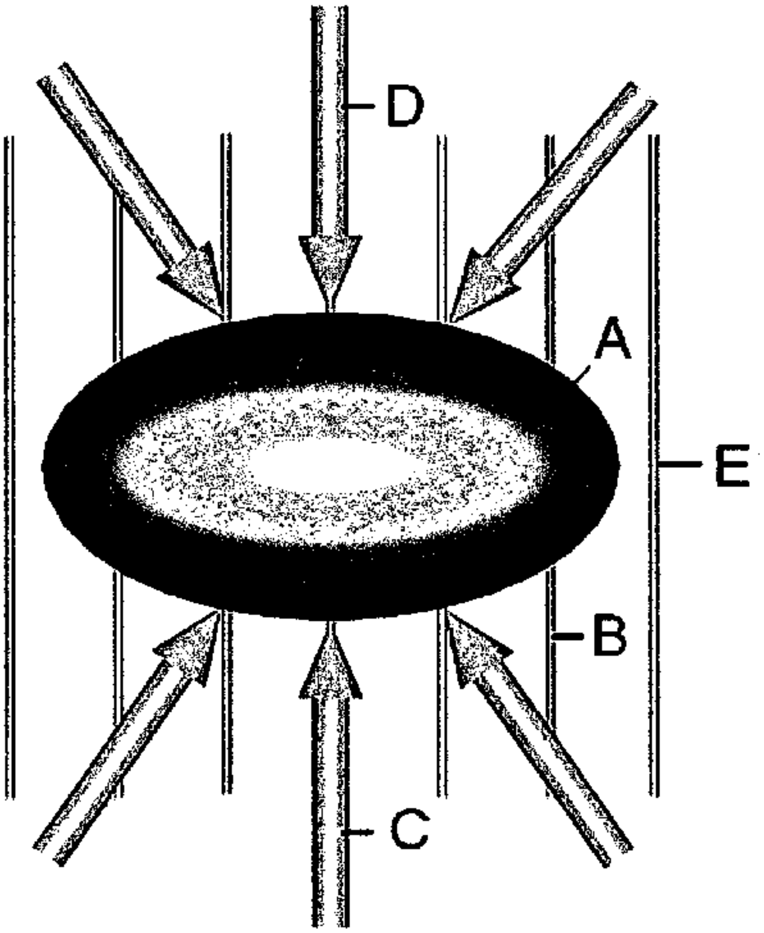

Figure 23

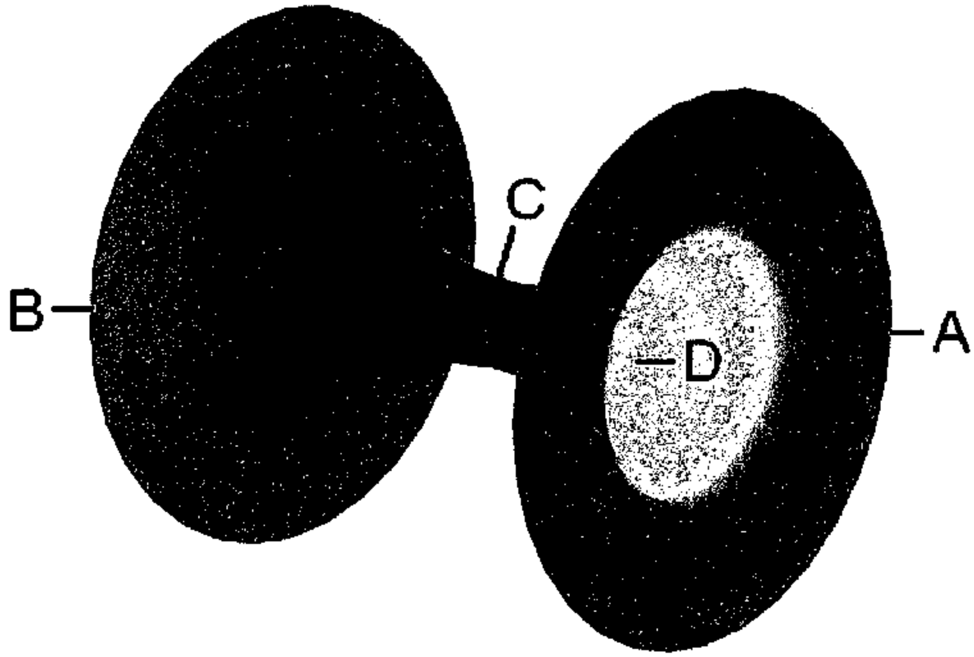

Figure 24

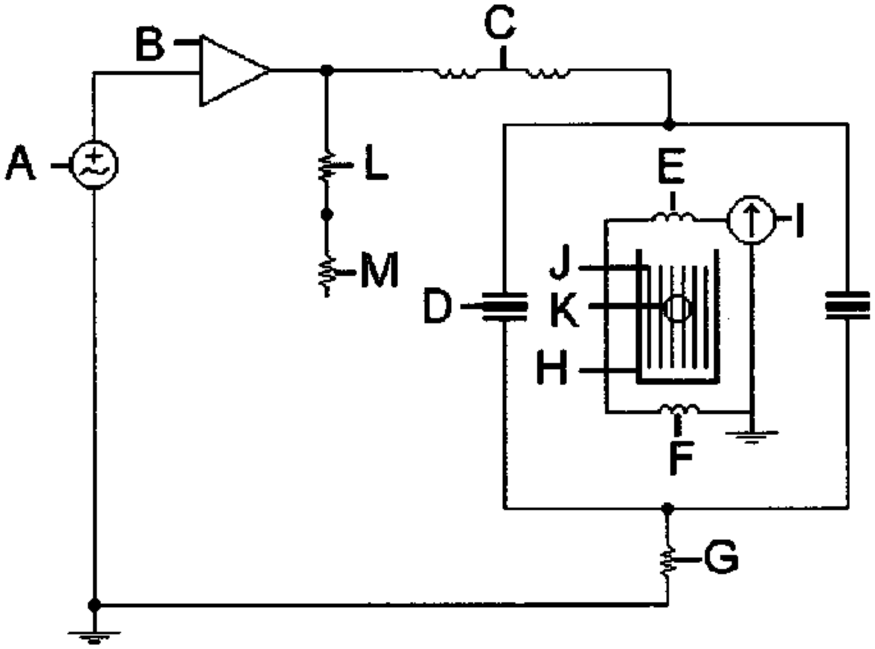
A
B
C
D
E
F
G
H
I
J
K
L
M

CAVITATING OIL HYPERSPACE ENERGY GENERATOR

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a hyperspace energy generator that uses cavitating oil bubbles within a magnetic field in order to create wormholes between space and hyperspace for the purpose of permeating the hull of a spacecraft with low-density, low-speed-of-light hyperspace energy.

BACKGROUND OF THE INVENTION

[0002] As shown by physicist Dr. John Archibald Wheeler of Princeton University, there is a second term in Newton's gravitational equation which includes an additional force depending on whether or not the body is spinning.

$$F=mg+m(v\times\Omega)$$

[0003] where F is the force, m the mass, g the acceleration of gravity, v the forward velocity of the body, and Ω the angular velocity of the body. Referring to **FIG. 1**, a rotating cylinder (A) is moving with a velocity (B) while spinning counterclockwise as shown by the angular velocity vector (C). The velocity crossed with the angular velocity vector produces a force in the upward direction (D).

[0004] Referring to **FIG. 2**, a pendulum (A) held in the left hand over the right hand will spin in the counterclockwise direction due to the hyperspace energy vortex (B) emanating from the palm of the right hand. This vortex arises because human beings are hyperspace energy beings that live in physical containers located in this universe. The vortex might be considered a wormhole between space and hyperspace. Notice that the angular momentum vector of the vortex points up out of the palm of the hand. If the hand were spinning from left to right, then there would be a force exerted, as shown by the above equation, that would bring hyperspace energy into the right arm.

[0005] Referring to **FIG. 3**, in order to test this hypothesis, a Chakra Vortex Accelerator was built for the purpose of spinning this hand vortex. The motor-driven machine has a base (A) on which is mounted a warehouse palletizing ring and platform (B). A control column (C) has a speed control and on/off switch. A person (D) stands on the platform with his right arm outstretched and palm facing upward (E). The platform spins from left to right.

[0006] Referring to **FIG. 4**, the velocity vector (A) points to the right. The angular momentum vector of the rotating chakra energy vortex points up. The cross product $A\times B=C$ produces a force that brings in high permeability hyperspace energy into the arm.

[0007] Referring to **FIG. 5**, after spinning 99 times per day for a few months, it was noticed that the right arm (A) would remain out at an angle ϕ from the body without making any exertion to do so. The left arm, which was not receiving the energy, remained at the side of the body. That is, the low-density hyperspace energy in the right arm was making the arm lighter and producing a subsequent anti-gravitational effect. The conclusion was that it might be possible to invent some device which would produce large amounts of hyperspace energy. This would make possible relativistic electromagnetic fields which can be used to produce the lift force for electromagnetic field propulsion vehicles.

[0008] Then in the February 1995 issue of *Scientific American*, an article appeared entitled *Producing Light from a Bubble of Air* which gave the details for producing sonoluminescence from a cavitating bubble using a kit for sale containing the piezoelectric transducers. Referring to **FIG. 6**, the experimental setup includes a ring stand (A) with a three-finger clamp (B) holding a 100 milliliter flask containing degassed water produced by a Walsh vacuum/pressure pump. Piezoelectric drive transducers (D) are mounted on the sides of the flask in order to create sound pressure waves which create the cavitation bubble (M) in the center of the flask. The transducers are connected by insulated wire (I) to a pair of inductors (E) having a mutual inductance between them. These inductors are driven by a frequency generator (H) connected to an audio amplifier (G). The oscillation signals are picked up by the oscilloscope (J) which is connected to the resistors (F) using the probes (K). The piezoelectric drive transducers act electrically as capacitors. The inductors are wired in series with the capacitance such that the combination resonates at the frequency at which acoustic resonance occurs. Bubbles introduced into the flask tend to coalesce in the center of the flask Above a certain sound intensity, the bubble will collapse and emit a dim light visible to the unaided eye in a darkened room, a process called sonoluminescence.

[0009] Further research showed that there were some earlier doctoral theses available. One of them by Dr. Steven DeWitt Horsburgh entitled *Radial Instabilities of a pulsating air bubble in water* mentioned a most interesting observed phenomenon of the surface oscillations called a "beaconing effect." When a bubble exhibited a stable or long-lived shape oscillation lasting more than two minutes, the bubble had a tendency to slowly precess. The precession rate was on the order of a few hertz which can be taken to mean around 3 Hz. Visually, the beacon looked like a light house beacon. It was a very obvious effect and when the beacon crossed the photodiode, the output voltage increased by at least an order of magnitude.

[0010] Upon reading this, it became apparent that sonoluminescence is involved with gravitational hyperspace physics. Einstein's theory says that inertia is a manifestation of the geometry of spacetime. It also says that geometry is affected by the presence of matter to an extent proportional to the factor G/c^2 where G is the gravitational constant and c is the speed of light. The inverse of this factor is the linear mass Ω of the universe. A spherical shell of mass m with a radius R will create a drag angular velocity on a pendulum equal to

$$\omega_{drag}=\frac{4}{3}\frac{G}{c^2}\frac{m_{shell}}{R_{shell}}\omega_{shell}=\frac{4}{3}\frac{\frac{m_{shell}}{\Omega}}{R_{shell}}\omega_{shell}$$

[0011] The mass of the shell divided by the linear mass (mass per length) converts the mass into a length which is then divided by the radius of the shell which is also a length. The numerical factor 4/3 is found only by a detailed calculation. The calculation, starting from a flat background spacetime manifold, showed the effect of the moving current of mass on the metric. Expressed in polar coordinates, the metric acquires a non-zero coefficient $g_{\phi t}$. Inserted into the equation for geodesic motion, this off-diagonal metric coefficient gives rise to a precession according to the above equation. The numerical factor corresponds to a pendulum located anywhere inside the rotating shell of mass.

[0012] In other doctoral theses, books and physics articles on cavitation, light is emitted after the bubble starts to collapse. A very sharp peak of light is produced which lasts in terms of picoseconds. Thus this type of sonoluminescence is very different than the beaconing effect where the bubble remains stable for long periods of time. The beaconing effect is actually a rotating wormhole between space and hyperspace.

[0013] Briefly what happens is that there is ionized gas in the bubble which releases electrons. As the electrons spiral down through the wormhole from space into hyperspace, they give off light. This situation puts the electrons in contact with the charge of hyperspace. Space has a charge which is equal to the sum of the electron charge q_e and the hyperspace charge q_h. The charge of space q is equal to the Planck mass times a conversion factor between mass and charge. The Planck mass is equal to the linear mass Ω of the universe times the bottom dimensional limit of the universe known as the Planck scale Λ. The conversion factor is the square root of the gravitational constant G times the permittivity of space ϵ. The permittivity is linear capacitance or capacitance per unit length. Thus the charge of space is

$$q=\Lambda\Omega\sqrt{G\epsilon}=q_e+q_h$$

[0014] Because the electron comes in contact with the hyperspace charge, the drag angular velocity in the above equation has to be modified. Notice that if the angular velocity of the shell is modified by the electron fine structure constant α, and the mass involved is the mass of the electron, moving through a radius into hyperspace of the Planck scale Λ, then the drag angular velocity would be

$$\omega_{drag}=\frac{4}{3}\frac{\frac{m_{electron}}{\Omega}}{\Lambda}\omega_{shell}\frac{1}{\alpha}$$

[0015] The angular frequency of the shell is 2π times the electron frequency f_e.

$$\omega_{shell}=\omega_{electron}=2\pi f_e$$

[0016] The inverse of the electron fine structure constant α is equal to Planck's reduced constant ℏ times the speed of light c times the factor 4π times the permittivity of space ϵ divided by the square of the charge of the electron q_e.

$$\frac{1}{\alpha}=\frac{\hbar c4\pi\varepsilon}{q_e^2}n=137.037916527$$

[0017] Planck's constant is equal to the Planck scale Λ squared times the linear mass Ω times the speed of light.

$$\hbar=\Lambda^2\Omega c$$

[0018] Notice that the fine structure constant has a square in it and Planck's constant also has a square. If the electron charge were traveling through the Planck scale into hyperspace, then there would be a linear charge equal to the electron charge divided by the Planck scale

$$\frac{1}{\alpha}=\frac{\hbar c4\pi\varepsilon}{q_e^2}n=\frac{\Lambda^2\Omega cc4\pi\varepsilon}{q_e^2}n=\frac{4\pi\Omega\varepsilon c^2}{\left(\frac{q_e}{\Lambda}\right)^2}n$$

[0019] The constant equal to the linear mass Ω times the permittivity times the square of the speed of light is actually the square of the linear charge χ of space. So the fine structure constant can be written

$$\frac{1}{\alpha}=4\pi n\frac{(\chi)^2}{\left(\frac{q_e}{\Lambda}\right)^2}=4\pi n\frac{\left(\frac{q}{\Lambda}\right)^2}{\left(\frac{q_e}{\Lambda}\right)^2}=4\pi n\left(\frac{q}{q_e}\right)^2=4\pi n\left(\frac{q_h+q_e}{q_e}\right)^2=$$

$$4\pi n\left(\frac{q_h}{q_e}+1\right)$$

[0020] So the electron fine structure constant is related to the ratio of the charge of hyperspace to the charge of the electron, which is the reason that the equation had to be modified. The electron moving across the Planck scale wormhole is modified by the hyperspace charge it is seeing in a hyperspace co-dimension. This confirms that the beaconing effect is a rotating wormhole into hyperspace.

[0021] The electron fine structure constant is actually the fastest velocity in the Bohr atom. The fine structure constant is the ratio of the electron orbital velocity to the speed of light.

$$\alpha=\frac{v_n}{c}$$

[0022] where the permitted quantized n orbital speeds are v_n/n or v/1, v/2, v/3 and so on. Using a value of v/3, the fine structure constant would be

$$\frac{1}{\alpha}=\frac{3c}{v}=4\pi n\left(\frac{q_h}{q_e}+1\right)=12\pi\left(\frac{q_h}{q_e}+1\right)$$

[0023] which when substituted into the drag angular frequency equation

$$\omega_{drag}=\frac{4}{3}\frac{\frac{m_e}{\Omega}}{\Lambda}2\pi f_e12\pi\left(\frac{q_h}{q_e}+1\right)=17.809\ \frac{\text{radians}}{\text{sec}}$$

[0024] The frequency of precession is

$$f_{drag}=\frac{\omega_{drag}}{2\pi}=2.83\text{ Hz}$$

[0025] which is close to the 3 Hz observed in the experiment. The ionized electrons are moving at a slower velocity in the bubble. The 1/3 ratio could also come from the subspace tetrahedron geometry where the tetrahedral angle is determined by this ratio

$$\theta = \operatorname{asin}(1/3) = 19.47122063°$$

[0026] This is an alternative possibility because this angle determines the charge of hyperspace according to the tetrahedron diagram of Δphysics. In geometrical physics, all the constants of physics are determined geometrically by the tetrahedron. The tetrahedron diagram plots the natural logarithm of mass versus the natural logarithm of wavelength. The tetrahedron is circumscribed by a sphere which reflects the electron wavelength into the electron mass. Thus the diagram combines the dual reality of classical physics, involving point mass particles, and quantum physics, involving wave-like particles. The scientific discovery that the electron and the proton are one and the same particle was made using this diagram. It also shows that hyperspace exists.

[0027] Referring to the tetrahedron diagram in **FIG. 7**, a line drawn at the tetrahedral angle produces the tetrahedral lines (A,B) along path (abc). This tetrahedron is circumscribed by sphere (C) with diameter (G). The electron wavelength (vertical dotted line D) reflects off the sphere (d,f) and returns as the electron mass (F). The distance (df) between reflection points is equal to the natural logarithm of the hyperspace charge. It can be shown that the electron charge is related to the geometry around the top of the sphere. Part of the geometry is outside the sphere which makes the charge immune to relativistic velocities.

[0028] Referring to **FIG. 8**, a current-carrying solenoid (A) produces a magnetic flux (B) through the center of the coil. The end of the solenoid behaves like a magnetic monopole. One end has a positive magnetic charge density $+\rho_m$ and the other has a negative magnetic charge density $-\rho_m$. One of Maxwell's electromagnetic equations states that the divergence of the magnetic field is equal to the magnetic charge density.

$$\nabla \cdot B = \rho_m = +\rho_m - \rho_m = 0$$

[0029] Because the magnetic field is solenoidal, the sum of the charges in our spacetime is equal to zero, which is the usual Maxwell equation. In the case of the wormhole, one pole of the magnetic flux is in our spacetime and the other pole is in a co-dimension of hyperspace. In our universe, there is a net magnetic charge density, and Maxwell's equation becomes

$$\nabla \cdot B = \rho_m$$

[0030] When working with bar magnets, bringing two north poles together shows that there is a spring constant involved. The electrons on the surface of the pole spin in the same direction which creates an electrical current around the surface. The magnetic pole strength g is therefore a spring constant K divided by the current per area J.

$$g = \frac{K}{J} = \frac{kg\,m}{s^2}\frac{1}{m}\frac{s\,m^2}{q} = \frac{kg\,m^2}{s\,q}$$

[0031] The magnetic charge density is then the pole strength per volume V or

$$\rho_m = \frac{g}{V} = \frac{kg\,m^2}{s\,q}\frac{1}{m^3} = \frac{kg}{m\,s\,q}$$

[0032] The divergence of the magnetic field has units of

$$\nabla \cdot B = \frac{1}{m}\frac{kg}{s\,q}$$

[0033] which are the units of the magnetic charge density. Notice that the units can also be formatted as a pressure per electrical current in the wire.

$$\rho_m = \frac{kg}{s\,q\,m} = \frac{kg\,m}{s^2\frac{q}{s}m^2} = \frac{\text{force}}{\text{area}}\frac{1}{I} = \frac{\text{pressure}}{\text{current}}$$

[0034] The current flowing through the coils of the solenoid, or around the pole face of the bar magnetic creates a pressure on spacetime. And it is this pressure that can punch through spacetime into hyperspace. Referring again to **FIG. 8**, an electrical charge winding n times around the end of the spacetime solenoid, as seen by path (C), is equal to the charge divided by Planck's constant times the pole flux

$$2\pi n = \frac{q}{h}\int_{Surface} B = \frac{q}{\hbar}\Phi$$

[0035] which shows that the flux Φ is quantized by the winding number n

$$\Phi = n\frac{2\pi\hbar}{q} = n\frac{h}{q}$$

[0036] where the flux is equal to the winding number n times Planck's constant h divided by the charge. Since the flux is the pole strength in webers, the pole strength g has a value of

$$g = n\frac{h}{q}$$

[0037] The magnetic charge density is the pole strength divided by the Planck volume or

$$\rho_m = \frac{g}{\Lambda^3} = n\frac{h}{q}\frac{1}{\Lambda^3} = n\frac{2\pi\Lambda^2\Omega c}{q\Lambda^3} = \frac{2\pi n}{\Lambda}\frac{\Omega c}{q} = 2\pi n\frac{B}{\Lambda}$$

[0038] The magnetic B field of space divided by the Planck length of the wormhole, acting as a linear magnetic field, times the winding number is equal to the magnetic charge density.

[0039] Referring to **FIG. 9**, the wormhole (A), connecting space and hyperspace (D), is precessing (C) in the counterclockwise direction with the angular velocity vector in the vertical z-direction. The magnetic flux (B) through the wormhole is in the radial direction. The mouth of the wormhole is an area with a normal vector in the radial direction also. This arrangement creates an electric field which circulates around the interior periphery of the wormhole. Maxwell's equation says that the curl of the electric field E is equal to the negative time rate of change of the magnetic B field.

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

[0040] This equation is difficult to manage in differential form. As both Maxwell and Feynman have said, it is better to use the integral form involving Stokes' theorem where the area da is converted into a line integral ds which goes well with Maxwell's idea of flux linkage.

$$\int_{surface} (\nabla \times E) \cdot n\, da = \oint E \cdot ds = -\frac{\partial}{\partial t} \int_{surface} B \cdot n\, da = -\frac{\partial}{\partial t} \Phi = -\Phi\, \omega_{drag}$$

[0041] The flux Φ, which is precessing around at the drag angular velocity calculated previously, is equal to the negative of the electric field around the interior perimeter (E) of the wormhole. This is the reason that the beaconing effect lasts for such a long length of time. The force on the electron is equal to the charge of the electron times the electric field. Because the charge is negative, the electron rotates counterclockwise in the opposite direction to that of the electric field. The electrons continuously circulate around the inside periphery of the wormhole. Quantum theory predicts that any radiated electromagnetic energy from an accelerated electron will give off one or more discrete quanta, or photons. The electron is accelerating because it is moving in a circular path. Because the water vapor is ionized in the cavitating bubble due to shock waves, surface oscillations and high temperatures, there are huge numbers of electrons available for producing an intense light which, as registered by the photodiode, increased by an order of magnitude when the beacon rotated by the photodiode.

[0042] From gravitational physics it is known that negative energy has to be produced in order to create a wormhole. Most experimenters are using collapsing spherical bubbles that do not produce negative energy and therefore do not produce the beaconing effect. This next background section shows that if the bubble collapses asymmetrically, then negative energy is produced.

[0043] Referring to an experiment carried out by Fizeau, **FIG. 10**, water flows (B,C) in opposite directions in two tubes of water (A). A light beam from a lamp (D) is projected through both tubes using half-silvered mirrors (E). Fizeau measured an interference (F) between light propagating with the flow and light swimming against the current, indicating that the flow of the medium does affect light propagation. A moving medium turns out to drag light to an extent quantified by a dragging coefficient

$$\alpha = 1 - \frac{1}{n^2}$$

[0044] where n is the index of refraction of the medium. Einstein's addition theorem of velocities states

$$v = \frac{v' + u}{1 + \frac{v'u}{c^2}} = \frac{\frac{c}{n} + u}{1 + \frac{cu}{nc^2}} = \frac{\frac{c}{n} + u}{1 + \frac{u}{nc}}$$

[0045] In the limit of slow flows compared with the speed of light in a vacuum, the effective velocity of light v in the medium flowing at a velocity u is

$$v \approx \left(\frac{c}{n} + u\right)\left(1 - \frac{u}{nc}\right) \approx \frac{c}{n} + \alpha u$$

[0046] The relationship between cavitating bubbles and light dragging is that when the bubble collapses, the surface has a tremendous acceleration and velocity which creates internal shockwaves that bounce back and forth within the medium. At the same time, the water vapor and air are ionized which creates short bursts of light. Thus there is light that is being dragged along by the collapsing surface of the bubble. This moving medium appears to light as a change in the metric of spacetime resembling, it turns out, a gravitational field. The velocity of the medium affects the elemental length ds^2 of spacetime as given by the metric for light dragging

$$ds^2 = c^2dt^2 - dx^2 + k(c^2dt^2 - u \cdot dx)^2$$

[0047] where k is equal to

$$k = \frac{\alpha}{c^2 - u^2}$$

[0048] and dx is the elemental length in the {dx,dy,dz} directions. The velocity u in the dot product is equal to the velocities {ux,uy,uz} in the x, y and z-directions. Expanding the equation produces the following metric

$$ds^2 = (1+k)dt^2 + (ku_x^2 - 1)dx^2 + (ku_y^2 - 1)dy^2 - dz^2 - 2ku_x dxdt - 2ku_y dydt + 2ku_x u_y dxdy$$

[0049] where velocity uz is zero and the speed of light is equal to one.

[0050] The dt^2 energy density term is modified by one plus the k factor. Thus the index of refraction of the medium, the speed of light and the overall light dragging velocity u affects the energy density. It can be shown that the speed of light in hyperspace is much lower. The difference of the squares of the velocities in the denominator creates the possibility for amplification of this term.

[0051] The factors of the elemental lengths are then inserted into the g metric tensor which is then used to calculate the spacetime distortion generated by the collapsing bubble. The g metric tensor is a 4×4 matrix having rows and columns that correspond to the elemental lengths {t, x, y, z}.

$$g_{\alpha\beta} = \begin{array}{c|cccc|} & \beta = t & \beta = x & \beta = y & \beta = z \\ \alpha = t & 1+k & -ku_x & -ku_y & 0 \\ \alpha = x & -ku_x & ku_x^2 - 1 & ku_x u_y & 0 \\ \alpha = y & -ku_y & ku_x u_y & ku_z^2 - 1 & 0 \\ \alpha = z & 0 & 0 & 0 & -1 \end{array}$$

[0052] In Einstein's General Theory of Relativity, the G curvature tensor, which is a 4×4 matrix having units of inverse meter squared, is equal to a stress-energy-momentum T tensor that is calculated from the electromagnetic fields. The stress tensor T combines energy density, electromagnetic flux, and pressure terms in one matrix.

$G = 8\pi T$

[0053] Most of the physics constants are linear constants such as linear mass which is mass per unit length. The permittivity of space ϵ is linear capacitance or capacitance per meter. The permeability of space μ is linear inductance or inductance per meter. The speed of light converts time into meters. What this means is that energy, pressure and momentum can all be converted into meters using these linear constants. The inverse squared of the length in meters is the curvature which is measured in units of m^{-2}.

[0054] Referring to **FIG. 11**, a square bubble (A) in the xy plane collapses with a velocity in the x-direction (C), and a velocity in the y-direction (B). This produces a curvature in the vertical z-direction (D). The curvature in the z-direction is the curvature tensor component G_{zz}. which can be computed directly from the g light dragging metric tensor. The result of the calculation is an equation in terms of the velocity, acceleration and the time rate of change of the acceleration, known as jerk.

[0055] Because the profile of the bubble radius expanding and collapsing with time has been published by those researching cavitation, it is possible to differentiate the profile to get the velocity, acceleration and jerk. These equations can then be inserted into the equation for the G curvature tensor.

[0056] The radius R of the bubble depends on the air density inside the bubble ρ, the viscosity of the water μ, the pressure of the P of the environment, the surface tension σ of the water, the ratio of specific heats γ, and the velocity $\dot{R}$ and acceleration $\ddot{R}$ of the bubble radius.

$$R\ddot{R} + \frac{3}{2}R + \frac{4\mu}{\rho}\frac{\dot{R}}{R} + \frac{2\sigma}{\rho R} + \frac{P_0}{\rho} - \left(P_0 + \frac{2\sigma}{R_0}\right)\left(\frac{R_0}{R}\right)^{3\gamma} = 0$$

[0057] Referring to **FIG. 12**, the above equation for the bubble radius is shown in the feedback simulation program. The diagram starts by solving the equation for the acceleration which is then integrated twice (1/S) to get the velocity and position. These variables are then fed back to the input through the various constants involving the viscosity, surface tension, pressure and air density. The graph at the bottom of the simulation shows the bubble radius with time.

[0058] Referring to **FIG. 13**, the bubble profile with time is reproduced from the *Scientific American* article mentioned previously. This profile was then fitted with a polynomial equation which is plotted in **FIG. 14**. By differentiating the equation, the velocity, acceleration and jerk are obtained which can then be inserted into Einstein's G curvature tensor.

[0059] Referring to **FIG. 15**, the curvature component in the vertical direction G_{zz} is plotted with respect to time. In the *Scientific American* plot, the light flash occurs at the end of the collapse marked "LIGHT FLASH" around 22 microseconds (**FIG. 13**). This corresponds to the two intense positive curvature G_{zz} spikes at the end of the collapse. This represents a symmetrical spherical bubble collapse where the velocities are the same in the x and y directions.

[0060] The energy density component G_{tt} for this symmetrical collapse is given by

$$G_{tt} = \frac{-2u_y^2 a_x^2 + 4u_x u_y a_x a_y - 2u_x^2 a_y^2}{4(u_x^2 + u_y^2 - 2)}$$

[0061] Looking at this carefully, notice that if the velocity in the x-direction u_x is equal to the velocity in the y-direction u_y, and the acceleration in the x-direction a_x is equal to the acceleration in the y-direction, then the equation reduces to zero.

$$G_{tt} = \frac{-2u^2a^2 + 4u^2a^2 - 2u^2a^2}{8(u^2 - 1)} = 0$$

[0062] which means that the symmetrical collapse with equal velocities does not produce any negative energy with which to create the wormhole. Only sharp pulses of light are emitted at the end of the bubble collapse where the intense positive curvature spikes are encountered. In order to get the rotating wormhole beaconing effect, the velocities have to be different.

[0063] Referring to **FIG. 16**, the bubble is given an asymmetrical profile as shown by the two curves. The upper curve is in the x-direction, and the lower curve is in the y-direction. The collapse in the y-direction is not as severe.

[0064] Referring to **FIG. 17**, the asymmetrical collapse has the curvature spikes at the beginning similar to the symmetrical case of **FIG. 15**. Notice that there are no spikes at the end of the collapse. However, the energy density is much different.

[0065] Referring to **FIG. 18**, the energy density G_{tt} is plotted as a function of time. Rather than being zero as in the symmetrical case, there is a large region of negative energy density as the bubble collapses. This negative energy is what creates the wormhole. Because there are no large curvature spikes at the end of the collapse, the wormhole bubble is stable and precesses for a few minutes.

[0066] Experiments with the pendulum show that some people have much larger auras than others. The pendulum swings over the entire hand. One particular person who has enormous vortices was working in a television repair shop. When a television set is brought in to the shop, the large capacitor has to be discharged first. The repairman with the large chakra vortices was standing behind a second repairman who was approaching the television set. As the second repairman got closer to the set, an enormous streaming blue spark sailed over his head and zapped the first technician. Hyperspace has a much higher permittivity than our spacetime. The resistance of space R is equal to the square root of the permeability μ over the permittivity ϵ.

$$R = \sqrt{\frac{\mu}{\varepsilon}}$$

[0067] Since the technician with the large chakra vortices was producing large quantities of hyperspace energy, his resistance was very much lower due to his larger permittivity. The spark grounded on his body. The free electrons, which are created by ionization of the hydrogen atoms of the water vapor, see a very low path of resistance and flow toward the wormhole. There they are trapped in accelerated motion due to the presence of the circular electric field and give off substantial numbers of photons which produces the beaconing lighthouse effect.

[0068] The elemental length ds^2 in cylindrical coordinates $\{t, r, \theta, z\}$ for a mass M is given by

$$ds^2 = -\frac{dt^2}{1-\frac{M}{r}} + \left(1-\frac{M}{r}\right)dr^2 + r^2 d\theta^2 + dz^2$$

[0069] The mass M distorts the radius r of the elemental length which is what gives the mass a curvature or gravitational field. Negative energy, which is produced at the end of the bubble collapse, is equal to a negative mass times the speed of light squared. Thus mass M in the equation goes to negative mass, or $M \rightarrow -M$. Furthermore, negative energy inverts the mass factor so that the metric becomes

$$ds^2 = -\left(1+\frac{M}{r}\right)dt^2 + \frac{dr^2}{1+\frac{M}{r}} + r^2 d\theta^2 + dz^2$$

[0070] The g metric tensor contains the factors of the elemental lengths along the matrix diagonal

$$g_{\alpha\beta} = \begin{Vmatrix} -\left(1+\frac{M}{r}\right) & 0 & 0 & 0 \\ 0 & \frac{1}{1+\frac{M}{r}} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & 1 \end{Vmatrix}$$

[0071] As before, Einstein's G curvature tensor in the radial G_{rr}, angular $G_{\theta\theta}$, and vertical z-direction G_{zz} is calculated from this negative energy g metric tensor.

[0072] Referring to the 3D G curvature plot in **FIG. 19**, the strength of the curvature is given by the length of the arrow, and the direction of the curvature is given by the arrow. The curvature in the radial direction G_{rr} is along the bottom axis from the box corner origin to the right. The curvature in the angular direction $G_{\theta\theta}$ is the bottom axis from the origin to the left. The height of the box is the curvature in the vertical direction G_{zz}. Notice in general that the curvature is perpendicular to the left face (θz-plane) and then starts to curve to the right and begins to point toward the right face (rz-plane). On the bottom face (rθ-plane) there is an upward vertical curvature.

[0073] In cylindrical coordinates, the θz-plane is actually a vertical tube of radius r. So at small r, the curvature is perpendicular to the tube. At larger radius, the curvature starts to point in the θ-direction, wrapping around the tube. At small radius, there is also a vertical curvature parallel to the sides of the tube.

[0074] Referring to **FIG. 20**, the curvature in cylindrical coordinates is seen to be the structure of the wormhole. The wormhole (A) is formed due to the radial pressure (B) which pulls open the wormhole. One end of the tube connects to hyperspace, and the other end connects to our spacetime. There is also a shearing pressure (D) along the sides of the tube which slices spacetime open and helps shape the tube. Moving outward along the radius, the curvature turns sideways and produces a shearing pressure (D) in the θ-direction which gives the wormhole a vortex appearance. These wormholes can also be very large, enabling one to peer into a co-dimension as attested to by the inventor.

SUMMARY OF THE INVENTION

[0075] As described in the previous section, the cavitating bubble has to collapse asymmetrically in order to generate the negative energy which creates the wormhole. In order to collapse asymmetrically, the forces on the bubble surface have to be unequal. The force on the bubble is given in terms of the a dragging coefficient, the bubble radius R, the surface charge density D_r, the magnetic B field in spherical coordinates $\{B_r, B_\theta, B_\phi\}$. The fields are integrated over the entire solid angle Ω of the bubble to determine the force F.

$$F_r = -\left(1-\frac{1}{n^2}\right)\frac{R^2}{2}\oint d\Omega\left[\left(1+\frac{1}{n^2}\right)(D_r^2 + B_r^2 - B_\theta^2 - B_\phi^2)\right]$$

[0076] The integration of the tangential components of the magnetic field such as $\{B_\theta, B_\phi\}$ over the solid angle of the bubble will sum to zero. The surface charge density D_r and the radial magnetic field B_r are the only fields which can be used to create an unbalanced force. If there is no surface charge density, then the force equation reduces to

$$F_r = -\alpha\frac{R^2}{2}\oint d\Omega\left[(2-\alpha)\left(\frac{\chi^2}{\Omega_m}B_r^2\right)\right]$$

[0077] where the conversion constant is the linear charge squared divided by the linear mass of the universe χ^2/Ω_m.

[0078] In this invention, referring to **FIG. 21**, the magnetic field is created by two cylindrical solenoids (A,B) which are wrapped around the top and bottom of the glass container (C). The container has flat sides in the shape of an octagon. Each solenoid consists of a continuous roll of very thin sheet copper which are separated by beaded mylar line. The mylar beads separate the sheets electrically and provide air cooling to the coil which can get rather hot. This makes it easy to assemble the coil because the copper roll and mylar line are wound together as the solenoid is formed. The lines of magnetic flux flow vertically through the container yet do not interfere with the transducers (D) which are mounted on the flat sides of the container.

[0079] Referring to **FIG. 22**, the cavitating bubble which forms in the liquid in the container experiences a vertical magnetic field (B) due to the two solenoids. These flux lines create a radial force on the top and bottom of the bubble (D,C). The flux lines on the outside (E) do not create any radial force because they are tangential to the surface. The spherical bubble flattens into an asymmetrical ellipsoidal shape (A) which enables the creation of the wormhole and lighthouse beaconing effect.

[0080] Referring to **FIG. 23**, the spacetime curvature of the wormhole (D) creates a tunnel (C), known as the throat of the wormhole, between flat space in our universe (A) and flat hyperspace (B) which exists in a co-dimension of our universe. In general, the co-dimensions have different physics constants. Going back to the chakra vortices of the human energy field, the vortices remove energy from space and deliver it to several hyperspace co-dimensions. This energy builds up human aura modules at different frequencies in astral space. Astral space has a very low speed of light and small linear mass. This is the reason that human beings can leave their bodies and pass through solid concrete walls, or project their spiritual eye to remote locations for viewing. On the other hand, the linear mass and speed of light of space is very much greater than hyperspace. Therefore, the pressure of the space is much greater than the pressure in hyperspace. In terms of the wormhole with a throat area A, the pressure would be the linear mass times the speed of light squared over the area

$$P_{space} = \frac{\Omega_s c_s^2}{A}$$

[0081] The pressure of hyperspace on the other side of the wormhole is

$$P_{hyperspace} = \frac{\Omega_h c_h^2}{A}$$

[0082] Because the linear mass and the speed of light of hyperspace are lower than that of space, there is a positive net pressure going into hyperspace.

$$P_{net} = P_{space} - P_{hyperspace}$$

[0083] This is the reason that the human vortices can obtain energy from this universe. Because the cavitating bubble produces negative energy, the linear mass is negative on the side of space. The pressure equation is then given by

$$P_{net} = -\frac{P_s}{A} - \frac{P_h}{A} = -\frac{P_s + P_h}{A}$$

[0084] which shows that there is a reverse pressure going from hyperspace into our space. The asymmetrical cavitating bubble generates hyperspace energy. The hyperspace energy flowing into our universe also gives the light generated by the wormhole a soft white misty look.

[0085] There are many reasons for generating this hyperspace energy. Electromagnetic fields are subject to the Lorentz transformation which involves comparing the frame velocity with that of the velocity of light.

$$E' = \frac{E}{\sqrt{1 - \frac{v^2}{c^2}}}$$

[0086] If the ratio of the velocity v to the velocity of light c is close to unity, then huge relativistic fields can be produced. These fields can curve spacetime to such an extent that a large lift force can be produced on the hull of an electromagnetic field propulsion vehicle.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0087] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0088] **FIG. 1**. Rotating cylinder demonstrating additional lift force.

[0089] **FIG. 2**. Use of the pendulum.

[0090] **FIG. 3**. The Ascension Machine, Chakra Vortex Accelerator.

[0091] **FIG. 4**. Hyperspace energy brought into body by rotation of hand chakra.

[0092] **FIG. 5**. Antigravitational effect using low density hyperspace energy.

[0093] **FIG. 6**. Sonoluminescence experiment.

[0094] **FIG. 7**. Tetrahedron diagram showing hyperspace charge.

[0095] **FIG. 8**. Magnetic monopole.

[0096] **FIG. 9**. Precessing wormhole.

[0097] **FIG. 10**. Fizeau light dragging experiment.

[0098] **FIG. 11**. Asymmetrical velocity collapse of bubble radius generates wormhole.

[0099] **FIG. 12**. VisSim feedback simulation of the collapsing bubble.

[0100] **FIG. 13**. Cavitating bubble radius as a function of time.

[0101] **FIG. 14**. Digitized bubble radius of **FIG. 13**.

[0102] **FIG. 15**. Spiking spacetime curvature of collapsing bubble.

[0103] **FIG. 16**. Dual velocity profile required to produce negative energy.

[0104] **FIG. 17**. Spacetime curvature for dual velocity collapse.

[0105] **FIG. 18**. Negative energy generated from asymmetrical collapse of bubble.

[0106] **FIG. 19**. 3D plot of spacetime curvature using negative energy.

[0107] **FIG. 20**. Wormhole vortex.

[0108] **FIG. 21**. Perspective of cavitation tank equipped with magnetic coils.

[0109] **FIG. 22**. Flattening of bubble by magnetic field.

[0110] **FIG. 23**. Hyperspace energy generated by reverse pressure gradient of wormhole.

[0111] **FIG. 24**. Solenoid and transducer driving circuits.

DETAILED DESCRIPTION OF THE INVENTION

[0112] 1. Referring to **FIG. 24**, the transducer driving circuit uses a frequency generator (A) driving a linear amplifier (B) into a pair of variable inductors (C) which resonate with the capacitance of the transducers (D) attached to the sides of the octagonally-shaped glass container (H). Resistors (L,M,G) provides voltage measuring points for the oscilloscope.

[0113] A DC current source (I) drives the two solenoids (E,F) which produce a vertical magnetic field (J) through the container. The magnetic field produces an asynmetrical force on the bubble surface (K) such that the collapsing bubble produces negative energy. The spacetime curvature produced by the light dragging of the collapsing bubble surface, together with the negative energy, create a wormhole between space and hyperspace. The negative spacetime energy produces a negative pressure which brings low-density hyperspace energy into our dimension.

I claim:

1. A hyperspace energy generator having:

a) a glass container in the shape of an octagon acting as the resonating chamber;

b) two piezoelectric transducers, mounted and operating acoustically as a pair on opposite sides of item (1a);

c) an opaque oil-like liquid with a high index of refraction filling item (1a) acting as the resonating liquid in which the cavitating bubbles are formed;

d) two thin sheet copper solenoids, located on top and bottom of item (1a), that produce a vertical magnetic field through items (1a) and (1c);

e) two variable inductance coils for creating a resonant frequency with item (1b);

f) a linear amplifier driving items (1b) and (1e);

g) a variable frequency generator driving item (1e);

h) a direct current electrical generator driving item (1d);

2. a cavitation system which:

a) produces one or more cavitating bubbles that collapse asymmetrically due to the differential force exerted on the bubble surface by the presence of a vertical magnetic field generated by item (1d) referred to above;

b) generates negative energy due to light dragging of the collapsing bubble surface;

c) creates wormholes between space and hyperspace due to the presence of item (1b);

d) generates a negative pressure between space and hyperspace that forces low-density hyperspace energy into our dimension; and

3. optional dual electrically-charged plates that:

a) are mounted on the sides of item (1a);

b) create a differential electric charge density across the bubble surface for the purpose of creating an additional method of asymmetrical bubble surface collapse.

* * * * *

US 20030209637A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2003/0209637 A1**

(43) **Pub. Date: Nov. 13, 2003**

(54) **ROTATING ELECTROSTATIC PROPULSION SYSTEM**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
John St. Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/142,582**

(22) Filed: **May 9, 2002**

Publication Classification

(51) **Int. Cl.^{7}** **B64G 1/40**; B64G 1/42

(52) **U.S. Cl.** **244/172**

(57) **ABSTRACT**

This invention relates to a spacecraft propulsion system utilizing thrusters comprised of a motor-driven electrostatically charged cylinder rotating within an electrostatically charged annular ring for the purpose of creating a spacetime curvature stress-energy tension in the horizontal direction. The thrusters are augmented by magnetic vortex generators, either embedded in the cylinders or located above each thruster, for the purpose of increasing the permittivity of space by permeating each thruster with low density hyperspace energy generated by a wormhole created between our space and hyperspace. A combination of three thrusters mounted on the underside of the hull of the spacecraft provide thrust and yaw motion control.

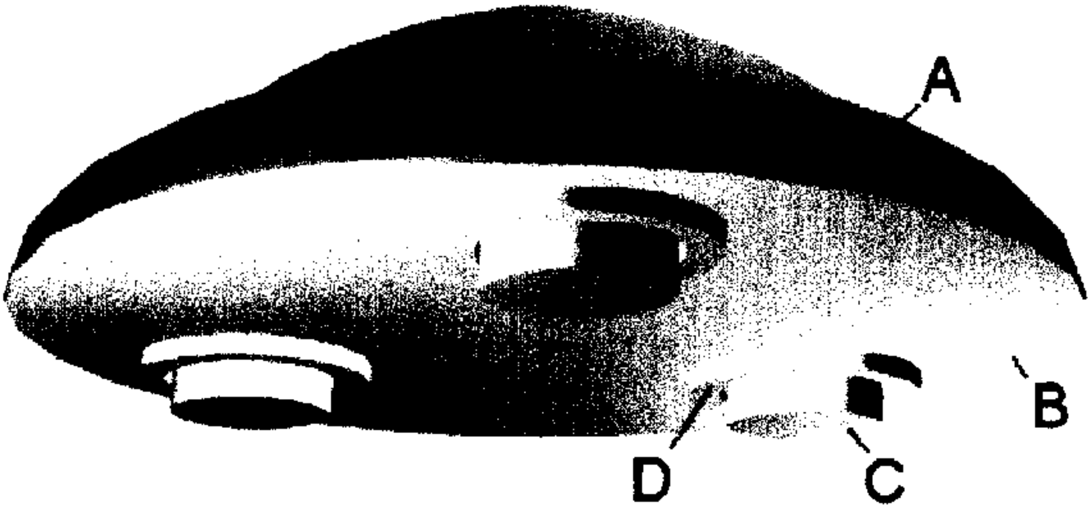

Figure 1

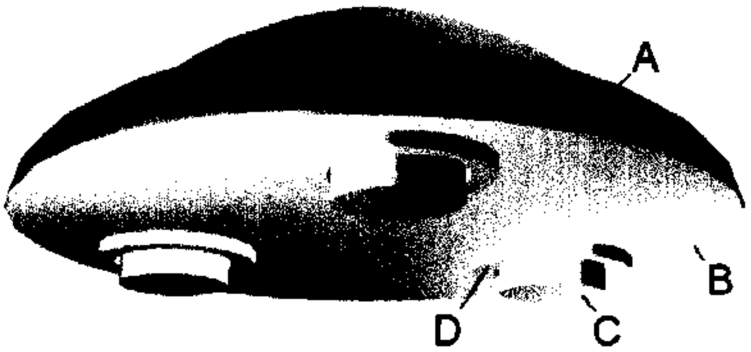

Figure 2

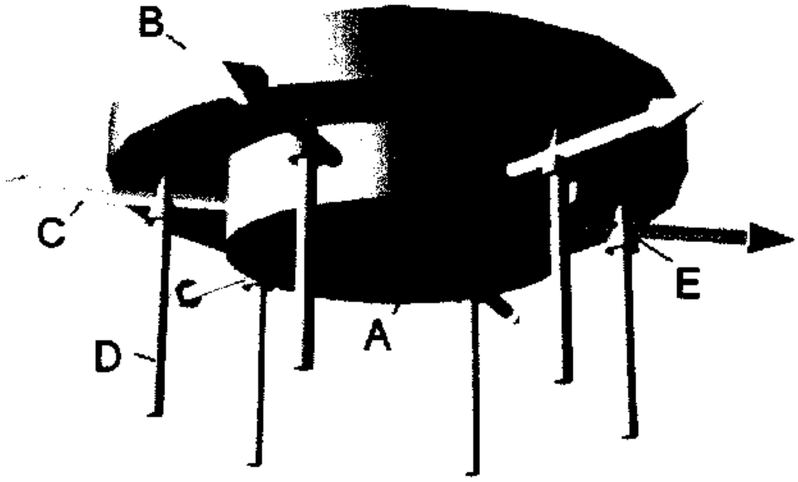

Figure 3

$$T^{zr} = -\frac{\varepsilon_0}{\Omega c^2}\frac{E_z E_r}{4\pi} = \frac{\mathrm{coul}^2}{\mathrm{m}^2\mathrm{n}}\frac{\mathrm{n}}{\mathrm{coul}}\frac{\mathrm{n}}{\mathrm{coul}}\frac{\mathrm{m}}{\mathrm{kg}}\frac{\mathrm{s}^2}{\mathrm{m}^2} = -\frac{\mathrm{kg\,m}}{\mathrm{s}^2}\frac{\mathrm{s}^2}{\mathrm{kg\,m}^3} = \frac{-1}{\mathrm{m}^2}$$

Figure 4

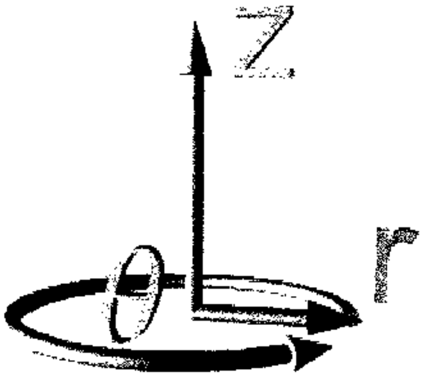

Figure 5

$$F^{\alpha}{}_{\beta} = \begin{array}{c|cccc|} t & 0 & E_r & 0 & E_z \\ r & E_r & 0 & 0 & 0 \\ \theta & 0 & 0 & 0 & 0 \\ z & E_z & 0 & 0 & 0 \end{array}$$

Figure 6

$$T^{zr} = -\frac{E_r E_z}{4\pi}$$

Figure 7

$$\frac{d}{dt} S_r = \varepsilon_0 \, \varepsilon_{r\,t\,z} x^t T^{zr} n_r \mathrm{Area} = \varepsilon_0 \, t \frac{E_r E_z}{4\pi} n_r \mathrm{Area}$$

Figure 8

$$\frac{\mathrm{coul}^2}{\mathrm{m}^2\,\mathrm{n}} \mathrm{sec} \frac{\mathrm{n}}{\mathrm{coul}} \frac{\mathrm{n}}{\mathrm{coul}} \mathrm{m}^2 = \mathrm{kg} \frac{\mathrm{m}}{\mathrm{sec}^2} \mathrm{sec} = \mathrm{kg} \frac{\mathrm{m}}{\mathrm{sec}}$$

Figure 9

$$\frac{d}{dt} S_{r,t} = \frac{\varepsilon_0}{4\pi} e^{i\omega t} E_r E_z \,\mathrm{area} + \frac{i\varepsilon_0}{4\pi} e^{i\omega t} E_r E_z \, t\, \omega \,\mathrm{area}$$

Figure 10

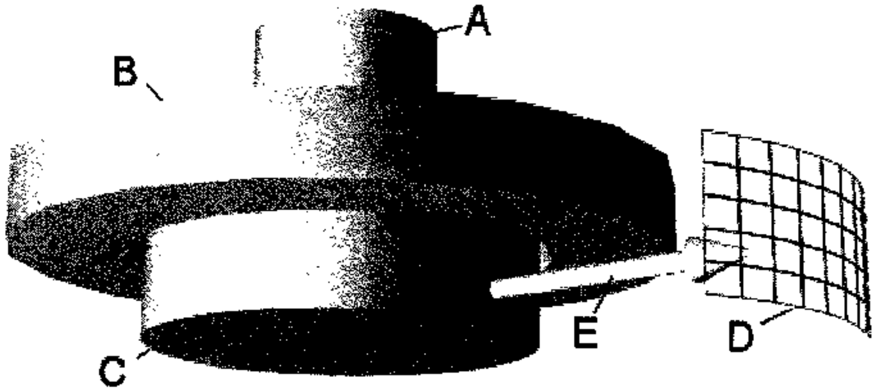

Figure 11

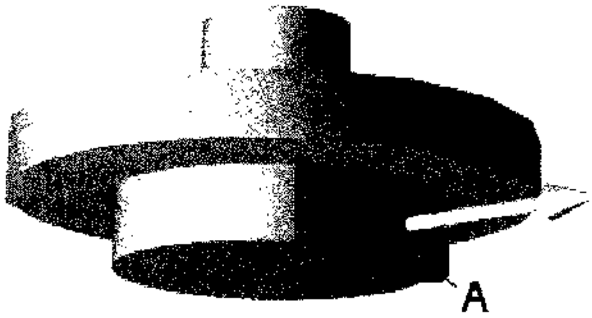

Figure 12

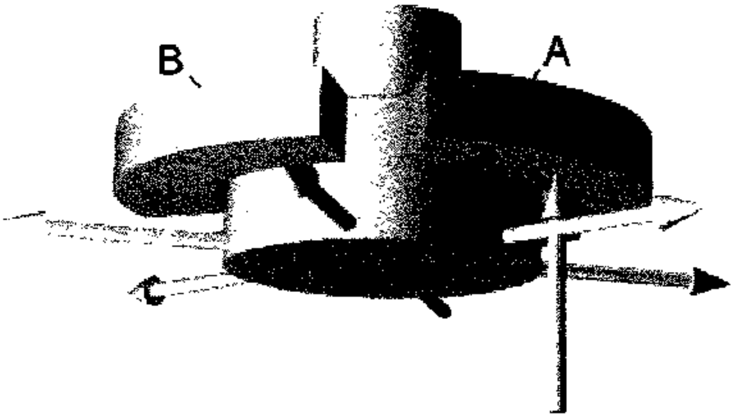

ROTATING ELECTROSTATIC PROPULSION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] The invention, which is the object of my present application, is a spacecraft propulsion system which develops a spacetime curvature tension utilizing a combination of a rotating radial electrostatic field and a fixed vertical electrostatic field. The two fields create a stress-energy T^{zr} gradient in the radially direction which is equal to force. The radial field is created on the side of a charged rotating cylinder on the underside of the hull. The vertical field is created by an annular charged ring concentric with the cylinder. Three rotating cylinders are located in a triangle on the bottom of the hull in order to produce a force in any direction in the horizontal plane.

REFERENCE PAPERS

[0002] *Gravitation*, Wheeler, page 80.

BACKGROUND OF THE INVENTION

[0003] When working with Maxwell's equations in tensor notation, it became apparent that a tensor can change identity depending on what permutation of variables is involved. For example, one single equation can involve both charge density and current density. And all of Maxwell's equations can be reduced to just two equations.

[0004] In the tensor equation for momentum, if the lever arm is length then the equation is equal to the flow rate of angular momentum. If the lever arm is time, then you get linear momentum. And if the field rotates with time, then the time rate of change of linear momentum is a force which is the basis for this invention.

[0005] Einstein said that mass curves space and space tells mass how to move. In this sense, generalized mass can be mass, electromagnetic fields, charge or angular momentum which create a spacetime curvature that produces a force on the spacecraft.

SUMMARY OF THE INVENTION

[0006] The invention relates to a spacecraft utilizing a rotating electrostatically charged cylinder and a concentric annular charged ring to create a stress-energy spacetime curvature in the horizontal plane on the spacecraft's underside hull. A motor drives the rotating cylinder which extends below the hull. A charged surface produces an electric field in the direction normal to the surface. The vertical and rotating electric fields combine to create a rate of change of linear momentum which creates a horizontal propulsive force on the hull.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0007] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0008] **FIG. 1**. Perspective view of spacecraft hull (A) with the three electrostatically charged rotating cylinders (C) surrounded by their annular charge rings (D) extending down on the underside of the hull (B).

[0009] **FIG. 2**. Perspective view of individual thruster showing rotating cylinder and ring.

[0010] **FIG. 3**. Stress-energy tension created by two electric fields and having units of curvature.

[0011] **FIG. 4**. Cylindrical spacetime coordinates $\{t, r, \theta, Z\}$.

[0012] **FIG. 5**. Faraday electromagnetic F tensor containing electric fields in the radial and vertical directions.

[0013] **FIG. 6**. Stress-energy tensor T showing it is equal to the product of the two electric fields divided by 4π.

[0014] **FIG. 7**. The tensor equation for flow rate of angular momentum S.

[0015] **FIG. 8**. The units are linear momentum due to the time lever arm.

[0016] **FIG. 9**. The rate of change of linear momentum is the horizontal force produced by the two electric fields.

[0017] **FIG. 10**. The angular momentum flows through an area whose normal vector is in the radial direction.

[0018] **FIG. 11**. Cylinder with only one electrostatically charged segment.

[0019] **FIG. 12**. Annular ring with three individual segments which can be charged separately to create a force in a particular direction.

DETAILED DESCRIPTION OF THE INVENTION

[0020] 1. Referring to **FIG. 1**, the spacecraft comprises an upper hull (A) with three rotating electrostatically charged cylinders (C) with their concentric annular electrostatically charged rings (D) located on the spacecraft's bottom hull (B).

[0021] 2. In a closer view of one of the cylinders seen in **FIG. 2**, motor-driven rotating cylinder (A) has an electrostatically charged surface which produces an electric field (C) normal to said surface. The fixed, electrostatically charged annular ring (B), which is concentric with the cylinder, produces a vertical electric field (D) normal to its surface. This crossed field (E) creates a negative spacetime curvature tension which is the product of the two fields divided by 4π as seen in the equation, **FIG. 3**.

[0022] 3. Notice that the equation involves the permittivity of space ϵ_0 divided by the linear mass of the universe Ω and the speed of light. This produces units of inverse meter squared which is the spacetime curvature. In Einstein's General Theory of Relativity, the spacetime curvature tensor is equal to the stress-energy tensor or $G=8\pi T$ where G is the curvature, and T is the product of the electromagnetic fields. The problem with this equation, which has been resolved with this invention, is that the linear mass Ω times the speed of light c is an enormous number. Even with the square of enormous electric fields, the curvature would be too small even to notice, and little force would be generated.

[0023] 4. In another patent application of mine entitled Magnetic Vortex Generator, it was shown that a rotating cylinder containing embedded and stacked bar magnets produces a negative mass and negative spacetime spring constant. It can be shown that this combination produces a small wormhole or interdimensional connection between our space and hyperspace along the centerline of the rotating cylinder. Co-dimensions of hyperspace have different physics constants. A low pressure region of hyperspace has a very low mass density and a very low speed of light. The wormhole allows this low density hyperspace energy to enter into our space and permeate the cylinder and annular ring. The permittivity is proportional to the inverse of the speed of light squared. The hyperspace speed of light, obtained from my tetrahedron physics diagram, is 8971 meters per second. The speed of light in our dimension is 299792458 meters per second. As shown by the enclosed reference calculation, the hyperspace permittivity is about a trillion times larger. Because the force is equal to this new permittivity times the electric fields squared times the area around cylinder, the force is greatly amplified by this increase in the permittivity of space.

[0024] 5. It is pertinent to this invention how the stress-energy is created due to the two electric fields in the vertical and radial direction. In gravitational physics, there is a Faraday F tensor which contains all the components of the electromagnetic fields. It is a 4 by 4 matrix whose rows and columns correspond to the coordinates of spacetime which in cylindrical coordinates are $\{t, r, \theta, z\}$ where t is time, r the radius, theta the horizontal angle and z the vertical height. These coordinates are shown in **FIG. 4**.

[0025] 6. The radial electric E_r field and the vertical electric E_z field can be inserted into the Faraday tensor seen in **FIG. 5**. The sign of the vertical field is positive because it points in the positive z-direction due to the fact that the annular ring has a negative charge. The cylinder has a positive charge. This produces the negative stress-energy tensor T^{zr} as drawn in **FIG. 6**.

[0026] 7. In order to calculate the force on the cylinder, it is necessary to calculate the flow rate of angular momentum. Momentum is mass times velocity or mass meter per second. If the mass is moving in a circle, then there is a lever arm times the momentum which makes it mass meter squared per second. If this is differentiated with respect to time, then a flow rate of angular momentum is produced with units of mass meter squared per second squared. As mentioned previously, tensors have this dual nature where depending on the permutation of the variables, it means one thing or another. In this case, the lever arm will be time, rather than length which converts the flow rate of angular momentum into just linear momentum. If you change linear momentum with respect to time, then you get a force.

[0027] 8. The flow rate of angular momentum S is shown in **FIG. 7**. The force has to be against the area whose normal vector is in the radial direction which is also the direction of the momentum. So S has an r subscript indicating that it flows in the radial direction. The permutation tensor ϵ has three subscripts which keeps track of the tensor notation. The first subscript is the same as the momentum subscript. Permutations of the coordinate variables which are in order have a plus one sign. Permutations which are in reverse order have a minus one sign. Permutations in which the variable are repeated are zero. For example, $\epsilon_{tr\theta z}=-\epsilon_{rt\theta z}$ because the r and t are in reverse order in cylindrical coordinates. Because the permutation tensor starts with r, then we can have permutations such as $\{r,t,z\}$ which is a reverse order negative permutation. This negative sign cancels the negative sign of the stress tensor. The reason this permutation is chosen is because the first subscript on the stress-energy tensor is now z. Because the normal to the area is in the radial direction, then the stress-tensor matches the electric fields that we have available, which are also in the zr-direction.

[0028] 9. In this case, the second subscript of the permutation tensor is time, rather than length. So the units become, as shown in **FIG. 8**, those of linear momentum. To me, this was very surprising. Then I realized that the radial electric field rotates with time which means the differential of the linear momentum produces a force on the hull. In exponential notation, the radial electric field rotates with $\text{Exp}[i\omega t]$. This is multiplied by the time lever arm, so the term that has to be differentiated is $t\ e^{i\omega t}$. The time lever arm saves the differentiation by making one term real so that the force is real. This is shown in **FIG. 9**. The units of the first term are real newtons. In the second term, the time t multiplied by the frequency cancels out in terms of units, so the units are still force, but imaginary

[0029] 10. There is an area involved in the force equation which is depicted in **FIG. 10**. The charged rotating cylinder (C) located inside the charged annular ring (B) is driven by motor (A). The radial electric field (E) is normal to the area (D) whose normal vector is also in the radial direction. The angular momentum flows through this area which surrounds the cylinder. In doing so, it curves spacetime which produces the force.

[0030] 11. If the radial electric field is continuous around the cylinder, then the net force is zero. Referring to **FIG. 11**, one section (A) of the rotating cylinder is charged, which means that there is a force once per cycle in a selected direction depending on when the annular ring is charged. The other option, referring to **FIG. 12**, is that the annular ring is divided up into two or three sections with the feature that the charge can be turned on (A) or off (B) on a particular segment. As the rotating electric field goes around, one of the annular ring sections will have an electric field in order create a force on that side.

[0031] 12. Since there are three force cylinders, this allows for yaw motion control so that the hull of the spacecraft can pivot to change direction. After the directional change, the two back cylinders can be synchronized to produce thrust in the forward direction.

[0032] 13. There could also be a polarity change for the vertical electric field such that a positive stress-energy is produced which would reverse the direction of the thrust.

What I claim as my invention is:

1. A spacecraft propulsion system utilizing three electrostatically charged motor-driven cylinders each one of which rotates within a charged annular ring to produce a spacetime stress curvature tension in the horizontal plane on the underside of the hull.

2. Said rotating cylinder comprised of bar magnets embedded in the cylinder and stacked in groups at intervals around the periphery of the cylinder with the purpose of increasing the permittivity of space by permeating the cylinder and ring with low linear mass, low speed of light hyperspace energy by means of a wormhole between our space and hyperspace. The larger permittivity is to increase the force. Said technique is contained in my patent applications Magnetic Vortex Wormhole Generator and Magnetic Vortex Generator.

3. Said rotating cylinder and annular ring having a suitable metal surface for forming and maintaining the electrostatic charge.

4. Said rotating cylinder having one or more segments which can be electrostatically charged individually or together to produce a force in a particular direction.

5. Said annular ring having one or more segments which can be electrostatically charged individually or together to produce a force in a particular direction.

6. A combination of three such thrusters providing a force in any particular direction or for yaw motion control.

7. Electric polarity switching of the fields in order to reverse the spacetime curvature and therefore change the direction of thrust.

8. The use of a magnetic vortex generator located above each thruster in order to permeate the cylinders and rings with low density hyperspace energy which would substitute for the embedded magnets in each cylinder.

* * * * *

US 20060072226A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0072226 A1**
St. Clair (43) **Pub. Date: Apr. 6, 2006**

(54) REMOTE VIEWING AMPLIFIER

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/957,391**

(22) Filed: **Oct. 2, 2004**

Publication Classification

(51) **Int. Cl.**
G02B 1/00 (2006.01)
(52) **U.S. Cl.** **359/896**; 359/894

(57) ABSTRACT

An apparatus which enhances the ability of a person to perform remote viewing by connecting the human spiritual eye to the tetrahedral geometry of subspace.

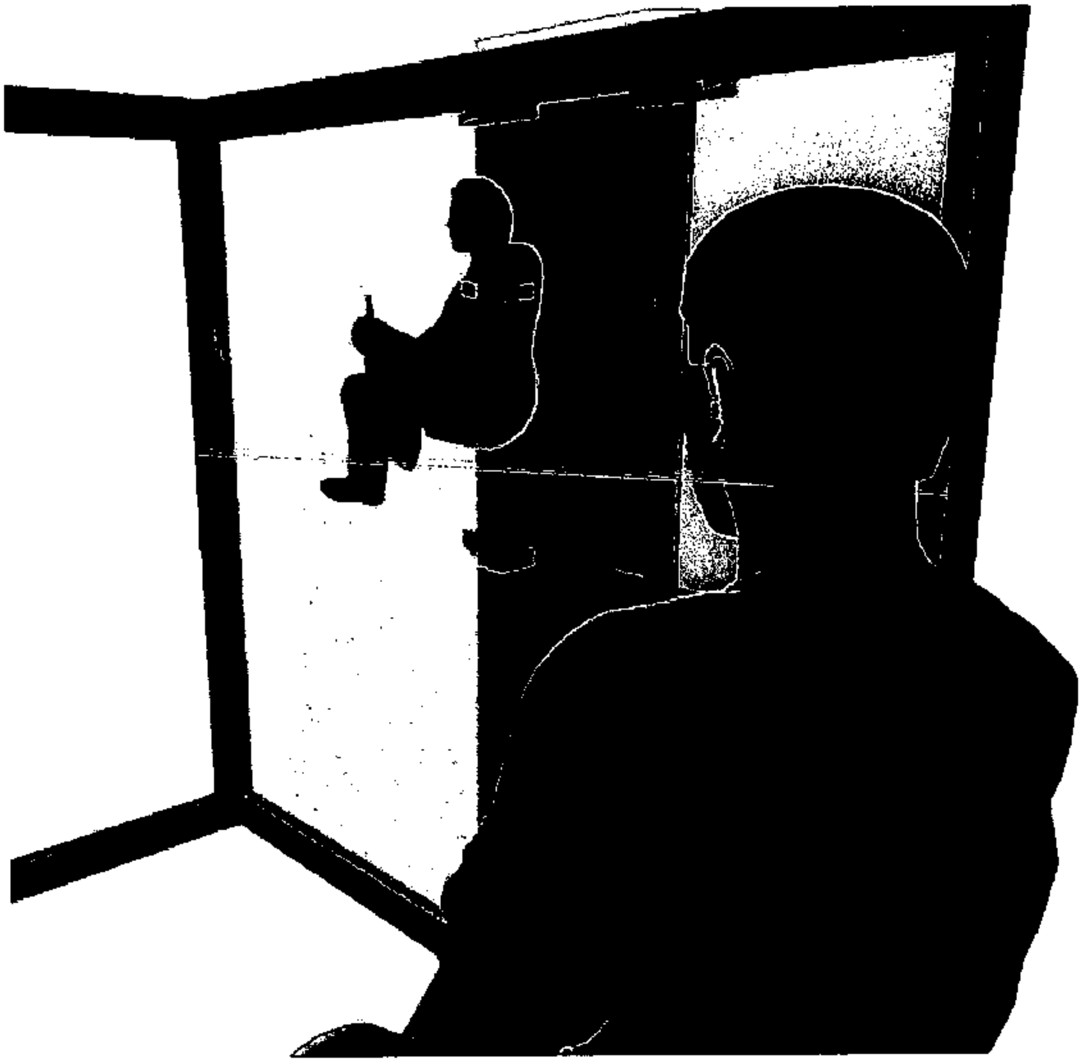

Figure 1

Figure 2

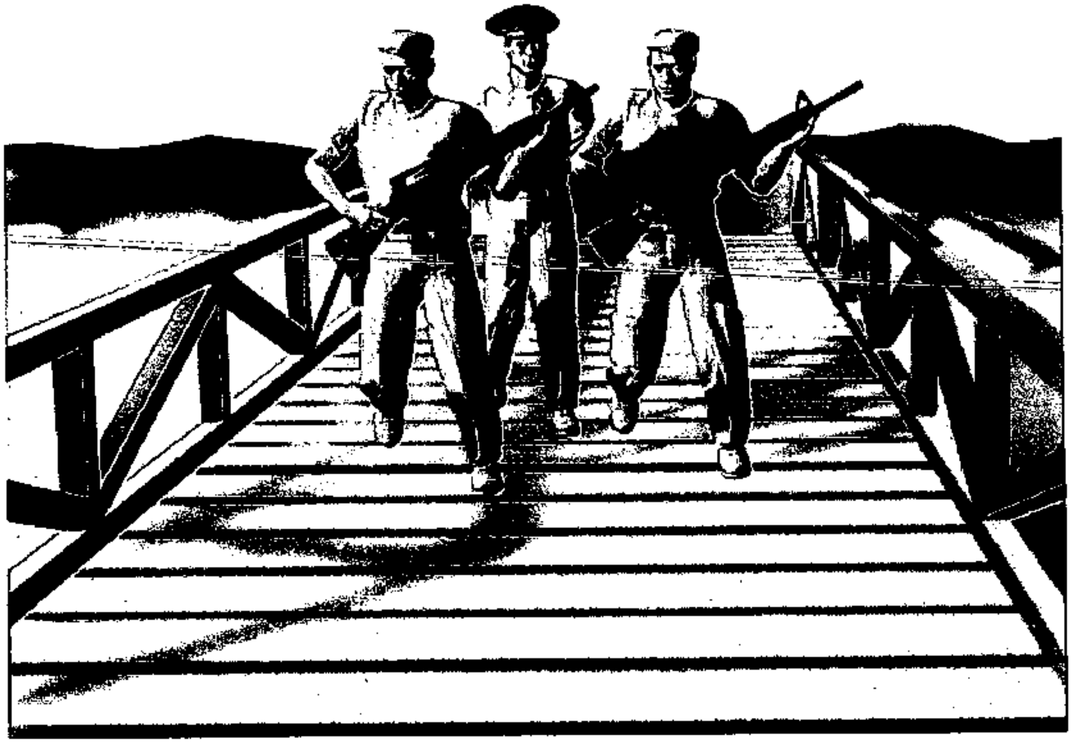

Figure 3

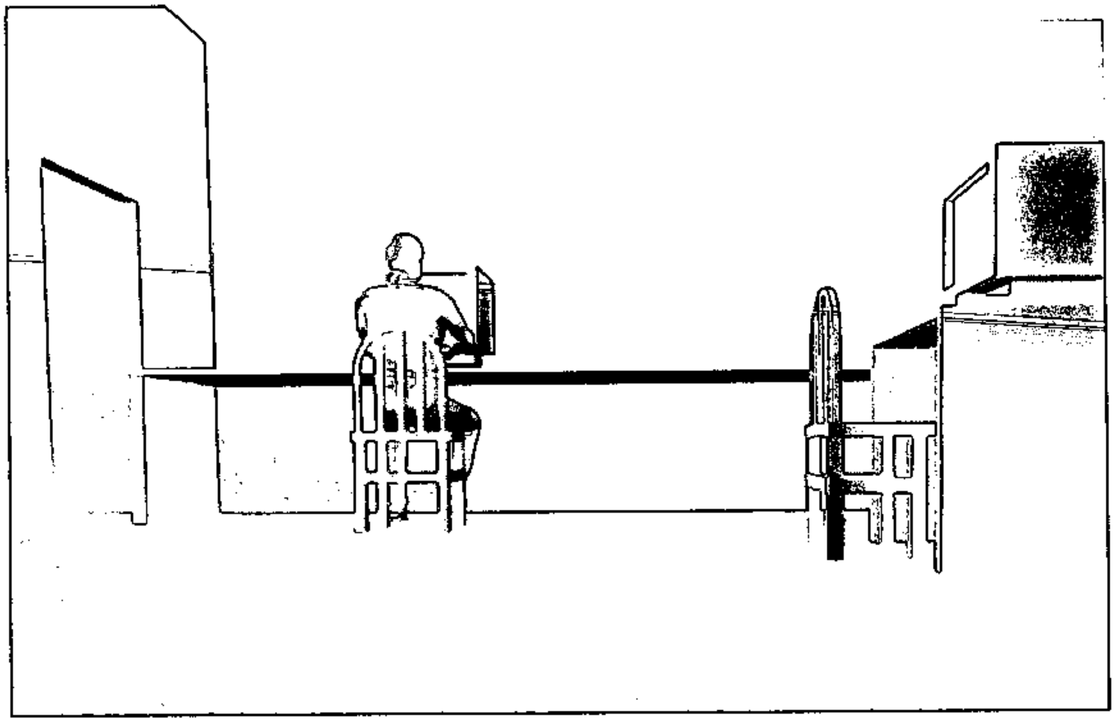

Figure 4

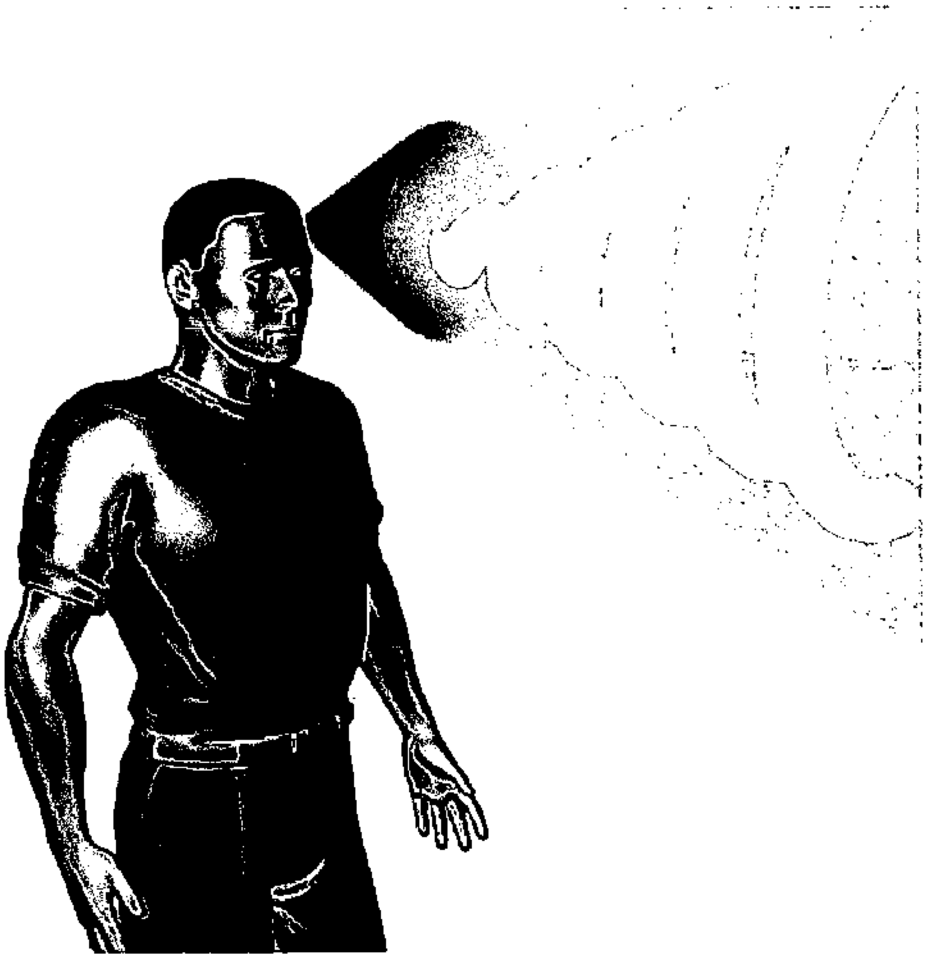

Figure 5

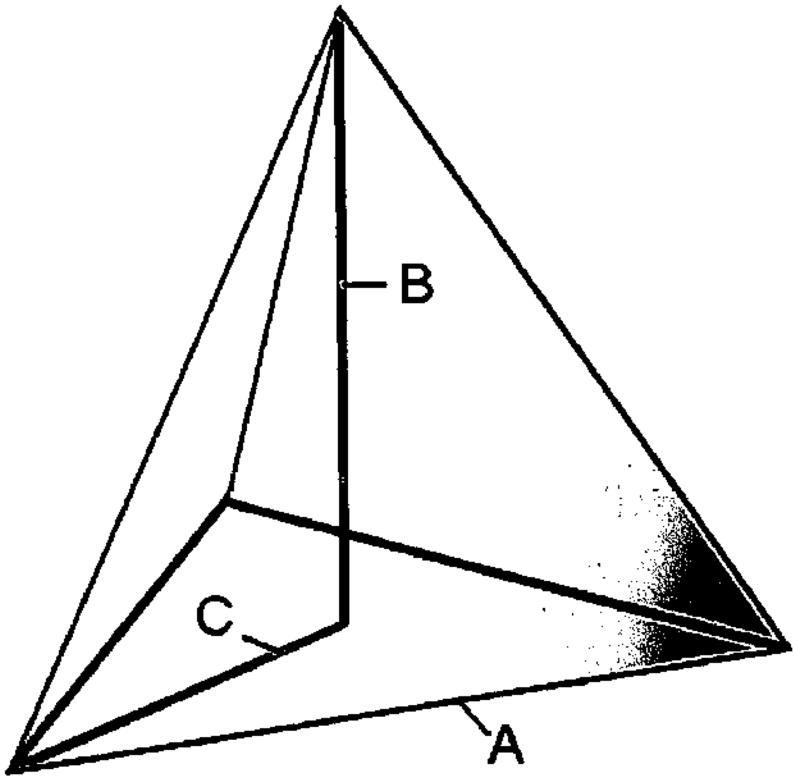

Figure 6

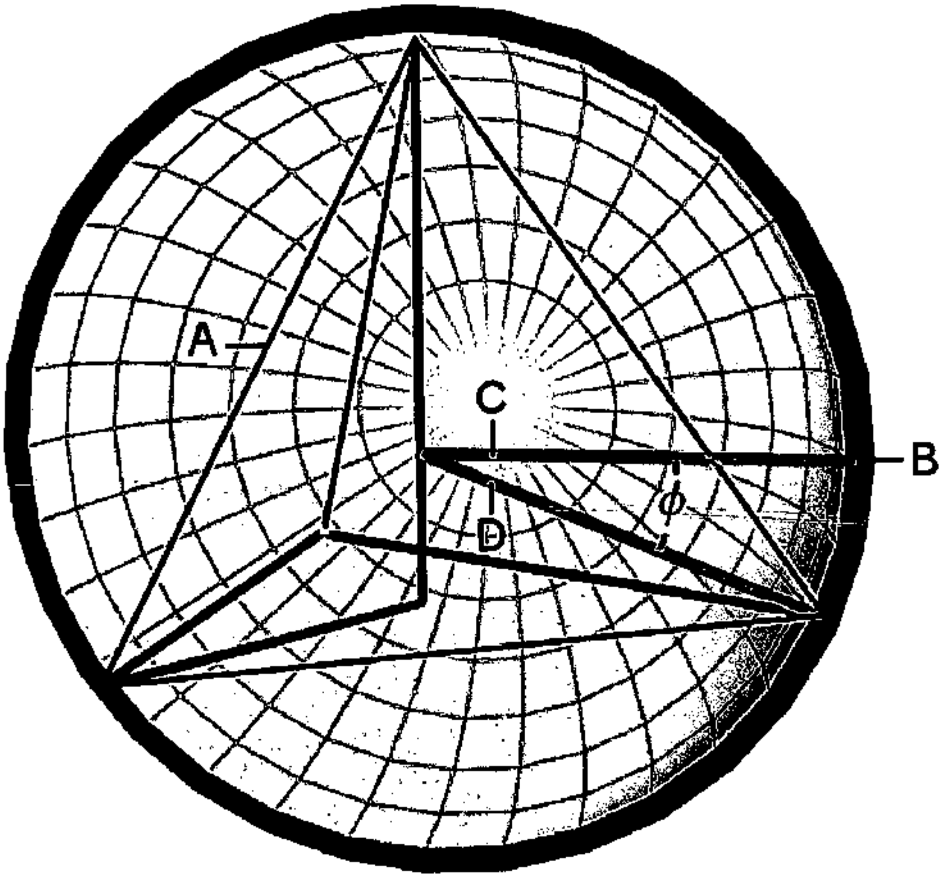

Figure 7

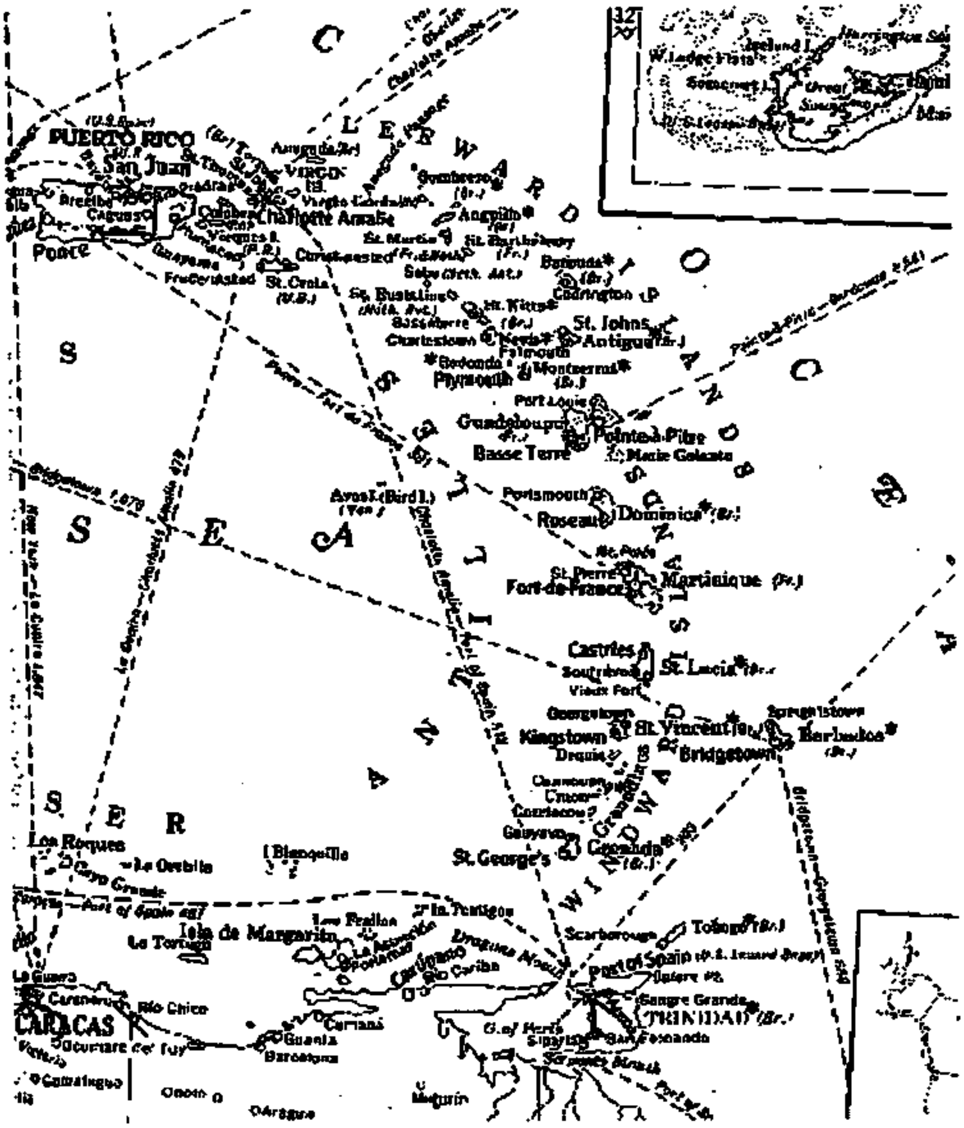
PUERTO RICO
San Juan
Ponce
Caguas
Arecibo
Charlotte Amalie
VIRGIN IS.
Anegada
Sombrero
Anguilla
St. Martin
St. Barthelemy
Christiansted
Frederiksted
St. Croix
Saba
St. Eustatius
Barbuda
Codrington
St. Kitts
Basseterre
Charlestown
Nevis
St. Johns
Antigua
Redonda
Montserrat
Plymouth
Falmouth
Port Louis
Guadeloupe
Pointe-à-Pitre
Basse Terre
Marie Galante
Aves I. (Bird I.)
Portsmouth
Roseau
Dominica
St. Pierre
Fort-de-France
Martinique
Castries
Soufrière
Vieux Fort
St. Lucia
Georgetown
Kingstown
St. Vincent
Speightstown
Barbados
Bridgetown
Bequia
Canouan
Carriacou
Gouyave
St. George's
Grenada
Los Roques
La Orchila
Cayo Grande
Blanquilla
Los Frailes
Los Testigos
La Tortuga
Isla de Margarita
Porlamar
La Asunción
Carúpano
Rio Caribe
Scarborough
Tobago
Port of Spain
Sangre Grande
TRINIDAD
San Fernando
Siparia
La Guaira
Caracas
Rio Chico
Cumaná
Guanta
Barcelona
Ocumare del Tuy
Maturín
Dragons Mouth
Serpents Mouth
C
L E E W A R D
I S L A N D S
W I N D W A R D
S
S E A
S E R
O
C
32
Ireland I.
W. Ledge Flats
Somerset I.
Great Sound
Harrington Sd.
Hamil

Figure 8

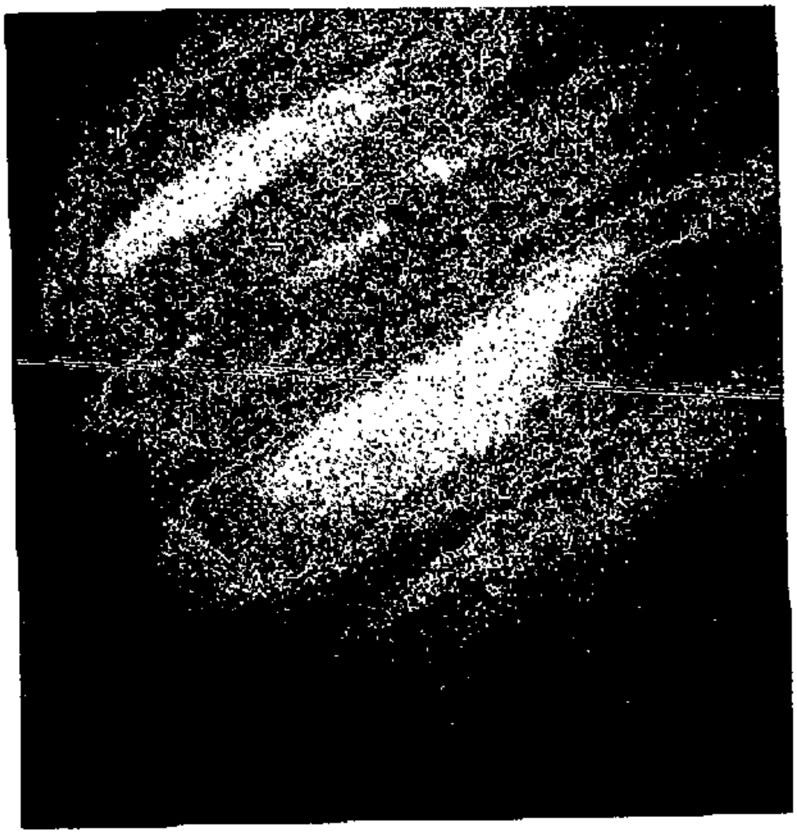

Figure 9

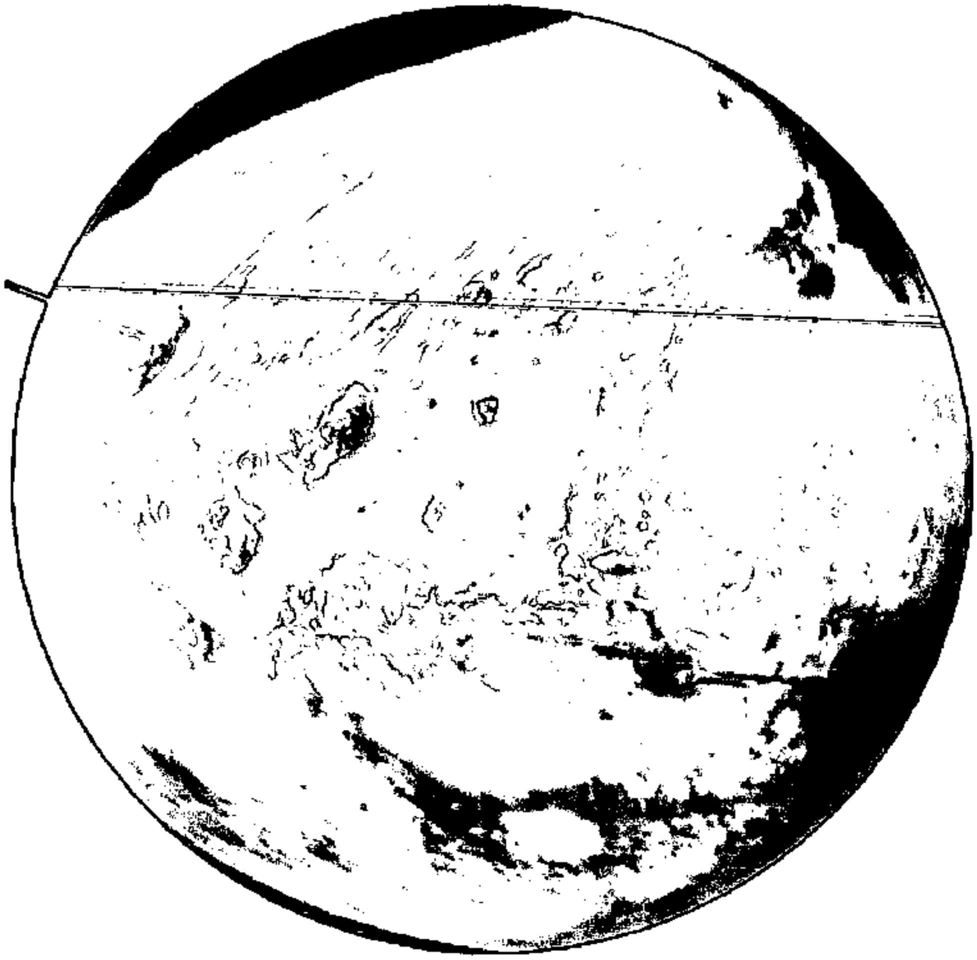

Figure 10

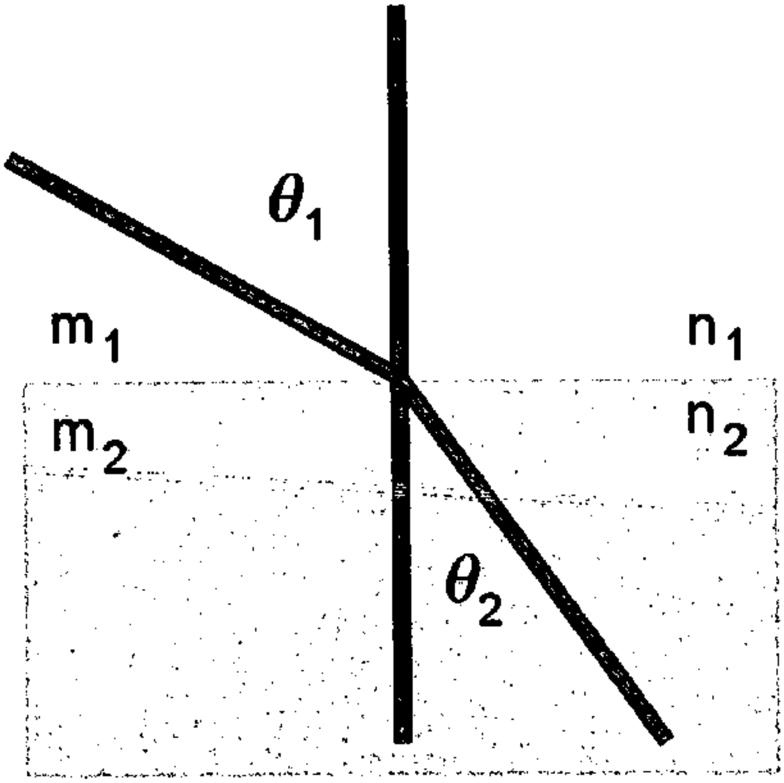
θ_1
m_1
n_1
m_2
n_2
θ_2

Figure 11

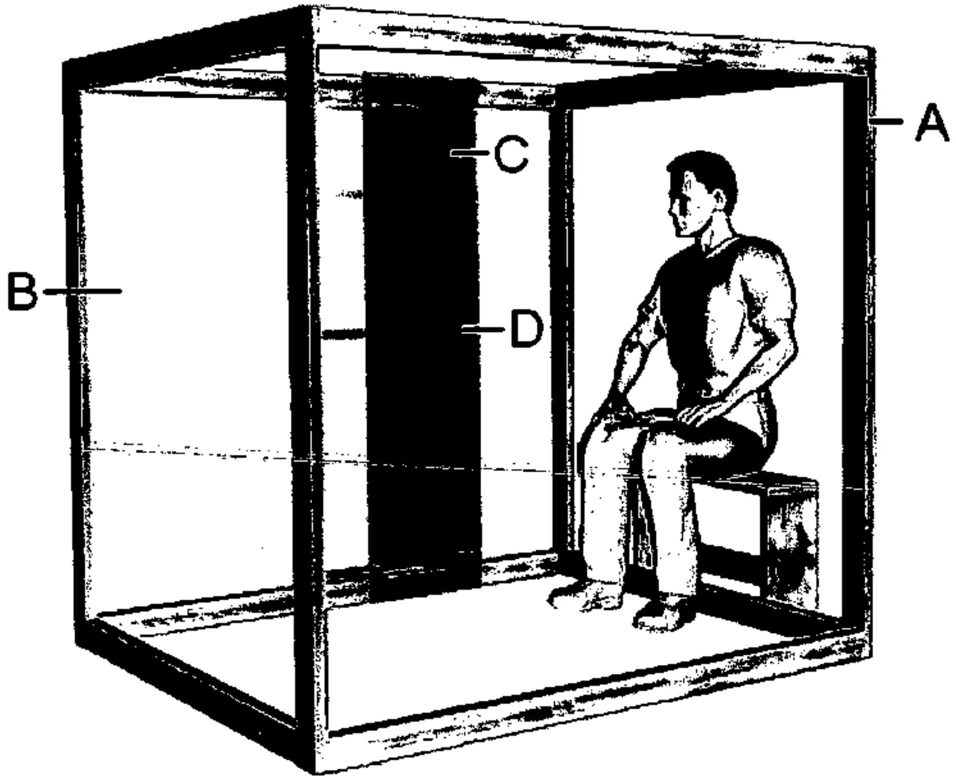

Figure 12

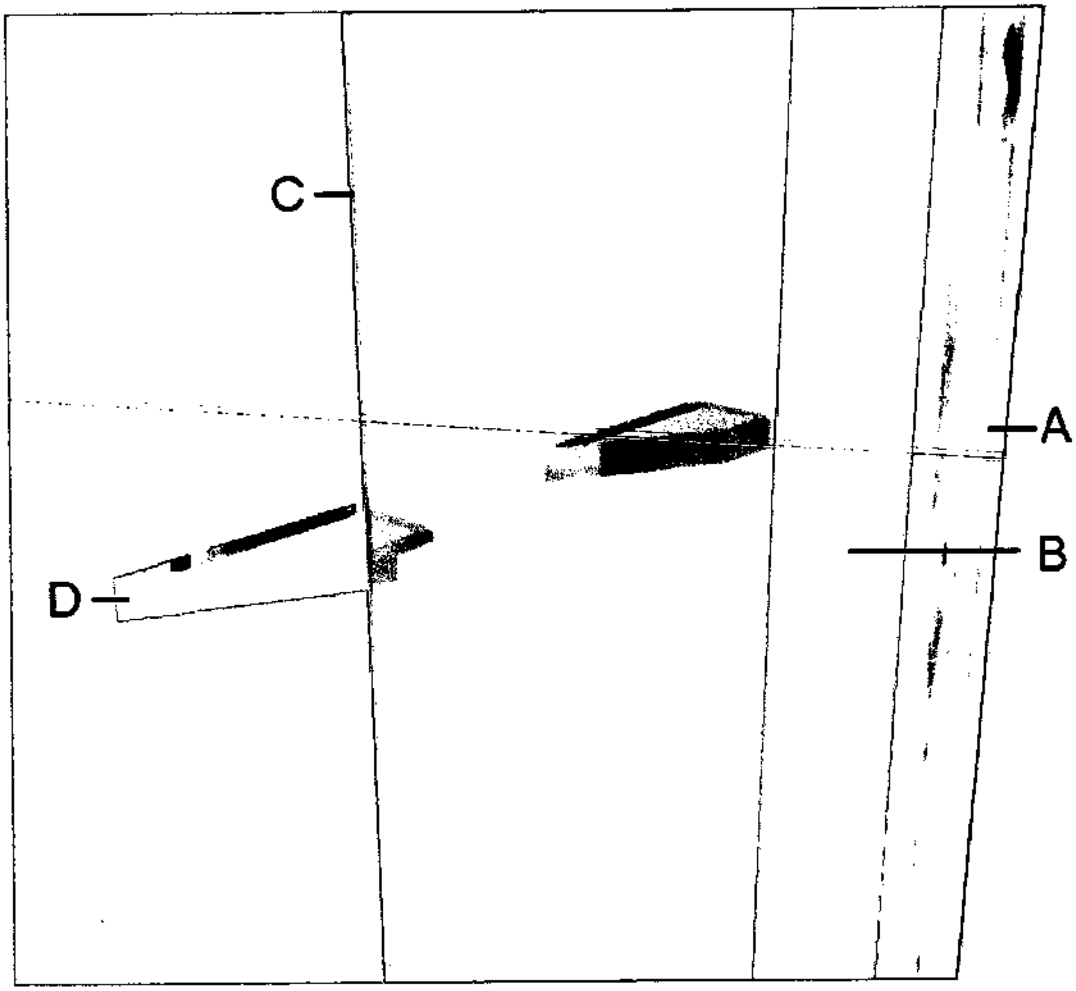

Figure 13

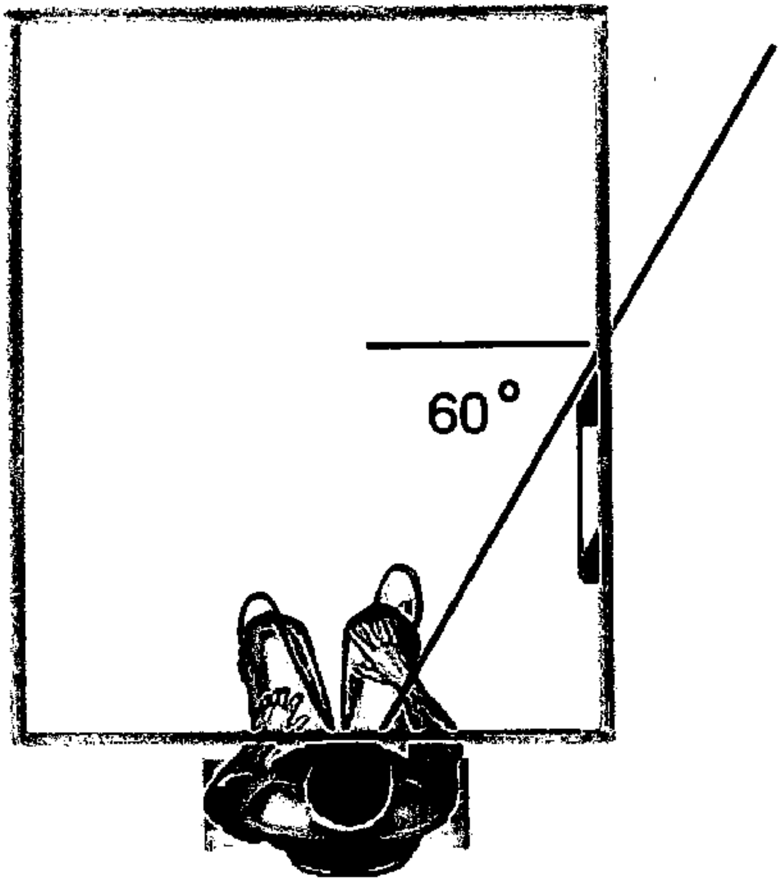

Figure 14

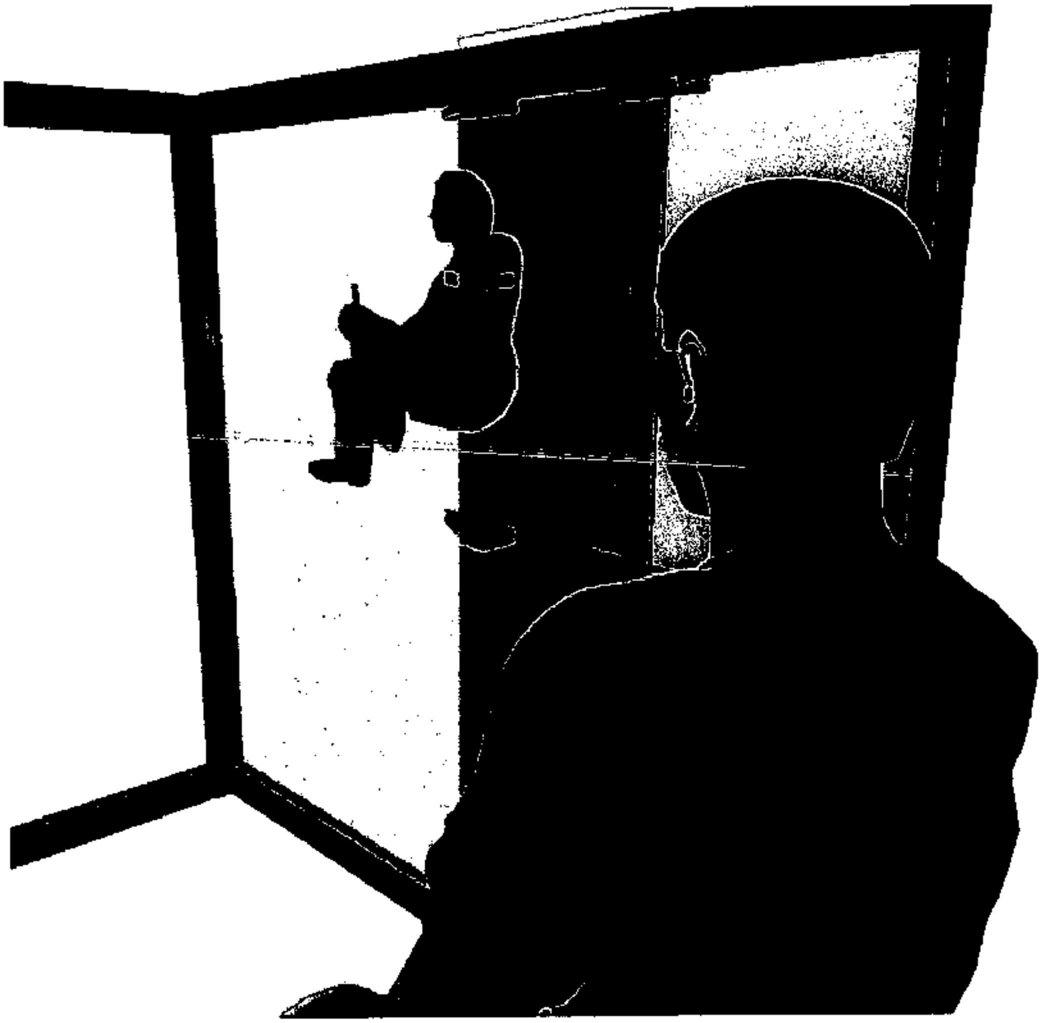

REMOTE VIEWING AMPLIFIER

BRIEF SUMMARY OF THE INVENTION

[0001] This invention enhances the ability of a person to perform remote viewing by connecting the spiritual eye to the tetrahedral geometry of subspace.

BACKGROUND OF THE INVENTION

[0002] Remote viewing is the projection of spiritual modules of the human energy field to distant locations in order to see, communicate and interact with other entities who live in subspace, space and hyperspace co-dimensions of the universe.

[0003] One of my first remote viewings was made at night to a distance of 10,000 miles on the sunlit side of the earth. My spiritual eye and body projected together while my mental facilities remained in my physical body. I found myself looking down on a palm tree from a height of about one hundred feet. The palm tree had several coconuts in it as seen in **FIG. 1**. I then gave the command to lower myself to the ground. At that moment I went sailing down past the coconuts, barely missing the tree! Finding myself on a pathway through the tropical forest, I then came to an extremely long wooden bridge which crossed over a river gorge. On the other side of the bridge I could see three soldiers running toward me as shown in **FIG. 2**. The two soldiers in front were carrying rifles and wearing light blue berets. The man running behind them was wearing an officer's cap with a red band. My first reaction was that I was going to be shot. I edged over on the right side of the wooden railing. They ran right past without seeing me. I then asked to see the building that these soldiers were guarding. Everything went dark, and then I found my spiritual eye peeking out of the floor of a computer room as seen in **FIG. 3**. There was one man using a computer on the opposite side of the room near an open door. He got up from his chair and came over to sit in front of a second computer located a few feet from where I was located. From the glare of the computer monitor, I could clearly see his face. Everything went dark as my spiritual eye and body projected back to my physical body.

[0004] Another time my spiritual eye, spiritual body and mind were standing outside the closed front door of my condominium. Upon patting my legs with my hands, I couldn't find the keys in my trousers. When I looked down, I realized I wasn't in my physical body. I then shot through two solid walls of concrete and returned to my awakening body.

[0005] What these two examples show is that the human spiritual energy system is modular. The reason it is modular is because there are seven hyperspace co-dimensions, each vibrating at a slightly different frequency, which receive energy from space through seven vortices located along the physical body. Over a lifetime, these vortices build up the human aura. Because all energy systems have to be grounded, the remaining six modules are the legs, body, arms/hands, voice, eye and mind. When the entire group is out-of-body as a single entity, then the soul energy powers the body in a manner similar to a battery. The soul looks like a two-inch diameter orange ball of plasma. If the soul is removed from the body, then the body becomes paralyzed except for a small movement of the eyelids. Upon death, all these separate modules are assembled into a single energy being. A light cord from the soul provides the battery power and information required to join these modules together. A second light cord from the pituitary gland transfers the modules into the energy being for assembly. This energy information transfer is the reason that people in a near-death situation say they saw their entire life flash before them.

[0006] Moving to a larger picture of things that are happening in the galaxy, I was able to make contact with the Pleiadian Federation which is located about 400 light years from earth. The Federation is a group of over one hundred intelligent beings that were brought to the Pleiades from around the galaxy. One member of the Federation calls itself the Intelligent Insect Beings. They are the ones who fly the black triangles over Belgium and France for the purpose of evacuating human beings back to the Pleiades for relocation on a planet called Earth II. The reason for this evacuation was that it was not known if it would be possible to win the battle of Revelations, which would take place about two years later here on earth.

[0007] As could be expected, the humans were angry and hysterical at being abducted. So the Intelligent Insect Beings asked me if I could calm them down. I was in telepathic communication with them, and they were in telepathic communication with their computer system, which meant that I could have my thoughts displayed to the humans on the computer monitor. It turned out that one woman was from Central America and only spoke Spanish. The Intelligent Insect Beings didn't speak Spanish so they thought there would be no way I could communicate with her. So I told them that I would spell the Spanish words letter-by-letter and she could then read my message. Since most Spanish people are Catholics, I thought a religious message would be of importance to her. I spelled out the phrase, "Que Dios te bendiga." which means, "May God bless you." I also asked the two beings to put their hands together in a form of prayer, and the woman followed suit. At that moment the Intelligent Insect Beings were reading her mind to understand her emotional state. They said, "She is crying tears . . . " After a slight but tense pause, they continued, "of joy!" They said the woman had a big smile on her face and was successfully transitioned into her new life.

[0008] One year later, the Blond Aliens of the hundred member council of the Pleiadian Federation remote viewed me so that I could celebrate with them the success of their mission to earth. The Blond Aliens, which is not their real name, fly the Beamship spacecraft. This hyperspace vehicle can teleport itself to any location in the galaxy. It dematerializes into the black void for an instance and then rematerialize anywhere in the galaxy. This ship can be seen in Billy Meier's video from Switzerland along with a picture of their envoy, Semjase.

[0009] Just recently I had a demonstration of the Beamship's ability. I heard an emergency distress call by a commander of a spacecraft who said that they had a fire onboard their spacecraft. I immediately remote viewed a Federation maintenance mothership, they got the frequency and location of the ship, and dispatched the Beamship in time to rescue the commander. The burned-out cables, which had caused the fire, were repaired by the Federation and the commander arrived home safely on his planet in his own spacecraft.

[0010] The Federation then received a message from the commander saying that he wanted his planet to join the Federation now that there was this type of communication available. So the Federation visited his planet for the signing ceremony, and I was invited to attend the proceedings by remote view. Word got around about this, and within three weeks another 20 planets joined the Federation for the same reason.

[0011] At one ceremony, which was attended by Admiral Third Class of the Pleiadian Defense Department, His Highness of the planet was signing the document of incorporation. At that moment, I could see him signing, so I exclaimed, "He is signing with his left hand." The Admiral was almost apoplectic at hearing this. After giving her the signed document, His Highness held up his hand and asked me how many fingers he was holding up. He had a fist so I said none. Then he put his index finger out and I said one. He then made a fist again and I said none. Then he held out all fingers and I said five. The Admiral said that he had a big smile on his face as he went to announce the agreement because he knew, even though his planet was located **90,000** light years away on the other side of the galaxy, he could instantly communicate any problems to the Federation. So this is the importance of developing inventions that can enhance our remote viewing ability because one day it will mean that we can become a vital part of the Pleiadian Federation.

SUMMARY OF THE INVENTION

[0012] Referring to **FIG. 4**, the spiritual eye of the human energy system is located at the pituitary gland in the forehead. It has the shape of a hollow cone which is composed of the misty white energy of hyperspace. Light coming into this vortex is then transferred by a light cord to a visual energy module which is located in a co-dimension of hyperspace. Because these modules are interconnected by light cords, the mind module is able to interpret the visual pattern the eye is seeing. More importantly, the mind can give logical instructions to this spiritual eye module for it to rotate around or move in a particular direction.

[0013] The reason that hyperspace has a white misty look to it is that the speed of light is very much less than the speed of light in our spacetime. The Lorentz transformation says that the distance L is shortened relativistically to a distance L' in a way related to the ratio of the velocity v of the object to the velocity of light c.

$$L' = L\sqrt{1 - \frac{v^2}{c^2}}$$

If the velocity of light is very low, then a small velocity creates an enormous contraction in length. By moving through hyperspace, therefore, enormous distances can be traversed. And this is the reason that it is possible to project the spiritual eye, voice and hearing to remote locations in the galaxy.

[0014] The universe is composed of subspace, space and hyperspace which are co-dimensions of each other. Subspace is defined by the geometry of the tetrahedron which is a four-sided solid whose faces are equilateral triangles having three 60° angles. Referring to **FIG. 5**, a tetrahedron whose sides are the square root of three (A), has a height of the square root of two (B), and base length equal to the square root of one (C). This forms the basic number set $\{\sqrt{1}, \sqrt{2}, \sqrt{3}\}$.

[0015] Referring to **FIG. 6**, the tetrahedron (A) is circumscribed by the sphere (B). Rod (C) is the sphere radius. A second rod (D), of equal length to rod (C), from the center of the sphere to the corner of the tetrahedron makes an angle ϕ of

$$\phi = \mathrm{ArcSin}\left(\frac{1}{3}\right) = 19.47122063'$$

So the four corners of the tetrahedron touch the sphere.

[0016] This tetrahedral geometry can be seen throughout the planets of the solar system Referring to **FIG. 7**, the islands of the Caribbean curve down from Puerto Rico to Venezuela forming an island vortex. The low density hyperspace energy releasing from the corner of the tetrahedron softens the rock mantle. The hot magma then rises through the rock with the least resistance. This creates a circular arc of volcanic islands along the edge of the vortex.

[0017] Referring to **FIG. 8**, the Giant Red Spot of Jupiter is located at a southern latitude of 19.5°. This vortex is so large that the entire earth can fit in it.

[0018] Referring to **FIG. 9**, the Olympus Mons volcano is located at a northern latitude of 19.5° as shown by the marker. This volcano is the size of France. Notice the fallen plume of volcanic debris toward the north east.

[0019] The double harmonic of the tetrahedral angle is twice 19.5° or 39° which is the location of the Silver Bridge in Point Pleasant, West Va. A large wormhole opened up around the bridge during Christmas rush hour when the bridge was full of cars. Due to the low density hyperspace energy, the rivets holding the cables down popped loose and all the cars were dumped into the river. A computer simulation using Schrodinger's quantum mechanics equation for a particle in a potential well shows that as the energy becomes less dense, the particle is no longer contained in the potential well. The electron jumps out. Thus the atomic bonds are broken which softens the rivets. This is the first time that there has been an understanding of the failure mechanism of this bridge.

[0020] After downloading from the Internet several pages of the index of refraction of a wide range of materials, I noticed that the index of refraction for Plexiglas was 1.50. Another source said it was 1.51. One of the Internet sites had a movable flashlight which showed the incident ray and the refracted ray. For Plexiglas, surprisingly enough, the incident ray was coming in at an angle of 60° to the normal, and the light was refracted at 35.26°, both of which are tetrahedral angles. The angle of the equilateral face of the tetrahedron is of course 60°. The angle at the top of tetrahedron is the arc-cosine of the ratio of the height over the edge length.

$$\theta = \mathrm{ArcCos}\left(\frac{\sqrt{2}}{\sqrt{3}}\right) = 35.26^{\circ}$$

According to Snell's law, the index of refraction n, times the sine of the angle $\sin(\theta_1)$ of the ray leaving material m_1, is equal to the index of refraction n_2 times the sine of the angle of refraction $\sin(\theta_2)$ of the ray entering material m_2. Referring to **FIG. 10**, the equation is

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$$

The index of air n_1 is equal to one. The index n_2 of Plexiglas is 1.50. If the incident ray is at $\theta_1 = 60°$ to the normal, then the output angle is

$$\theta_2 = \mathrm{ArcSin}\left[\frac{n_1}{n_2}\mathrm{Sin}(\theta_1)\right] = \mathrm{ArcSin}\left[\frac{2}{3}\frac{\sqrt{3}}{2}\right] = 35.26^{\circ}$$

which is equal to the angle of the tetrahedron.

[0021] Then I recalled several months earlier that I had gone to the Subway restaurant to get a sandwich. I was sitting by the Plexiglas window communicating with the Admiral whose mothership was in earth orbit. She mentioned that they were bringing two people aboard. At that moment I looked out through the window and I could see both of them clearly and easily through my spiritual eye. To my amazement, I saw both of the captives start to pull out guns from behind their backs. I then projected by spiritual hands which resulted in preventing the attack on the security guards. To say the least, the Admiral was rather surprised at these events. She then asked me to look at the design of the window because she thought it had something to do with my enhanced remote viewing capabilities. Looking at the Plexiglas, I noticed that on the edge of the large window pane there was a shorter piece of Plexiglas which was mounted parallel to the window pane. This smaller panel acted to protect the yellow neon fluorescent tube. I took the measurements of the design using a piece of paper that I found near the table.

[0022] I then went home and designed up a mounting bracket with my 3D computer software. I had already installed the stereolithography software that converts the design to the *.STL file format. How stereolithography works is that it slices the design into many thin horizontal sections. The machine has a platform which is mounted in a bath of liquid polymer. An ultraviolet laser, mounted on an xy-table, then traces out the slice. Because the liquid polymer is light sensitive, it polymerizes immediately into solid plastic. Then the platform is lowered a few thousandths of an inch and the second slice is added. This process eventually builds up the complete 3D part. Using the Internet, the *.STL file is sent by e-mail to the stereolithography service provider who returns the part overnight. So the next day I had the piece from which I made a plastic mold and several additional pieces for mounting the Plexiglas bracket on a full sheet of Plexiglas that I ordered locally. It never occurred to me to measure the angles, so after I got the index of refraction for Plexiglas, I measured the incident angle and it turned out to be 61°. So then I realized that the spiritual eye was being diffracted across these two Plexiglas plates which connected it to the tetrahedral geometry of subspace. Because subspace is the foundation of space, this created a much more efficient route for remote viewing. The result is that this invention has allowed me to make contact with 430 alien civilizations. Since then I have been awarded the Aphysics prize for my work in the invention and elaboration of the tetrahedron diagram of which there are now over 4000 graphs. The scientific discoveries contained in the diagram are (1) the electron and proton are one and same particle, (2) the existence of hyperspace, (3) how mass can be taken out of dimension, (4) cosmology determines the elementary particles, (4) all the physics constants and the tetrahedral geometry are contained in the two 360° circles of the infinity symbol and (5) all the physics constants are determined geometrically and projected from a subspace manifold into our dimension. For my work in Revelations, I was awarded four beautiful galloping riderless white horses of the Apocalypse.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0023] **FIG. 1.** Remote viewing the top of a palm tree containing several coconuts.

[0024] **FIG. 2.** Remote viewing three soldiers running across a wooden bridge.

[0025] **FIG. 3.** Remote viewing computer building that soldiers were guarding.

[0026] **FIG. 4.** Spiritual eye of human aura.

[0027] **FIG. 5.** Tetrahedron.

[0028] **FIG. 6.** Tetrahedron circumscribed by sphere.

[0029] **FIG. 7.** Caribbean volcanic island vortex.

[0030] **FIG. 8.** Giant Red Spot on Jupiter.

[0031] **FIG. 9.** Olympus Mons volcano on Mars.

[0032] **FIG. 10.** Snell's Law of Refraction.

[0033] **FIG. 11.** Perspective view of remote viewing station.

[0034] **FIG. 12.** Wedge-shaped spacers for diffraction panel.

[0035] **FIG. 13.** Remote viewing angle.

[0036] **FIG. 14.** A remote viewing session taking place on a planet located 90,000 light years from earth on the other side of the galaxy showing His Highness signing with his left hand the document that allows his planet to become a member of the Pleiadian Federation.

DETAILED DESCRIPTION OF THE INVENTION

[0037] 1. Referring to **FIG. 11**, the remote viewing station is a rectangular box wooden frame (A) on which is mounted on one side a large sheet of Plexiglas (B). The diffraction panel (C), made of a shorter length of Plexiglas, is mounted with acrylic glue on the Plexiglas sheet using clear polyoptic molded plastic spacers (D).

[0038] Referring to **FIG. 12**, the wedge-shaped spacers (D) hold the diffraction panel (C) to the sheet of Plexiglas (B). The angle of the wedge is 30° which makes the incident angle 60° to the normal.

[0039] Referring to the top view perspective **FIG. 13**, sitting on the wide bench, the remote viewer can adjust his sight along the wedge for proper alignment at an angle of 60°. Due to the 1.50 index of refraction of Plexiglas, the spiritual eye is diffracted across the edge of the first panel and then refracted across the second panel at the tetrahedral angle of 35.26°.

[0040] Referring to **FIG. 14**, the remote viewing image is seen superimposed on the large sheet of Plexiglas which acts as the viewing screen.

I claim:

1. A remote viewing station comprising:

(a) a rectangular box frame made of wood having a length of six feet, a width of four feet and a height of six feet;

(b) a large sheet of quarter inch Plexiglas, having an index of refraction of 1.50, mounted on the right side of item (1a);

(c) several wedge-shaped clear plastic spacers, three inches in width and eight and a half inches in length, having a wedge angle of 30° that are mounted on the interior right side of item (1b);

(d) a shorter sheet of quarter inch Plexiglas, having a length of one and a half feet, mounted on item (1c) parallel to item (1b);

(e) a wide bench on which the remote viewer sits so that the remote viewer can align his sight along the wedge angle of item (1c);

2. A remote viewing amplifier that:

(a) diffracts the spiritual eye of the remote viewer across the edge of item (1d) at an incident angle of 60° to the normal;

(b) refracts the spiritual eye at 35.26° to the normal due to the refractive index of item (1a) according to Snell's Law of Refraction; and

(c) aligns the spiritual eye with the tetrahedral geometry of subspace due to the diffraction/refraction combination of item (2a) and (2b).

* * * * *

US 20060145019A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0145019 A1**

St. Clair (43) **Pub. Date: Jul. 6, 2006**

---

(54) **TRIANGULAR SPACECRAFT**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/017,093**

(22) Filed: **Dec. 20, 2004**

Publication Classification

(51) **Int. Cl.**
B64G 1/40 (2006.01)

(52) **U.S. Cl.** **244/171.5**

(57) **ABSTRACT**

A spacecraft having a triangular hull with vertical electrostatic line charges on each corner that produce a horizontal electric field parallel to the sides of the hull. This field, interacting with a plane wave emitted by antennas on the side of the hull, generates a force per volume combining both lift and propulsion.

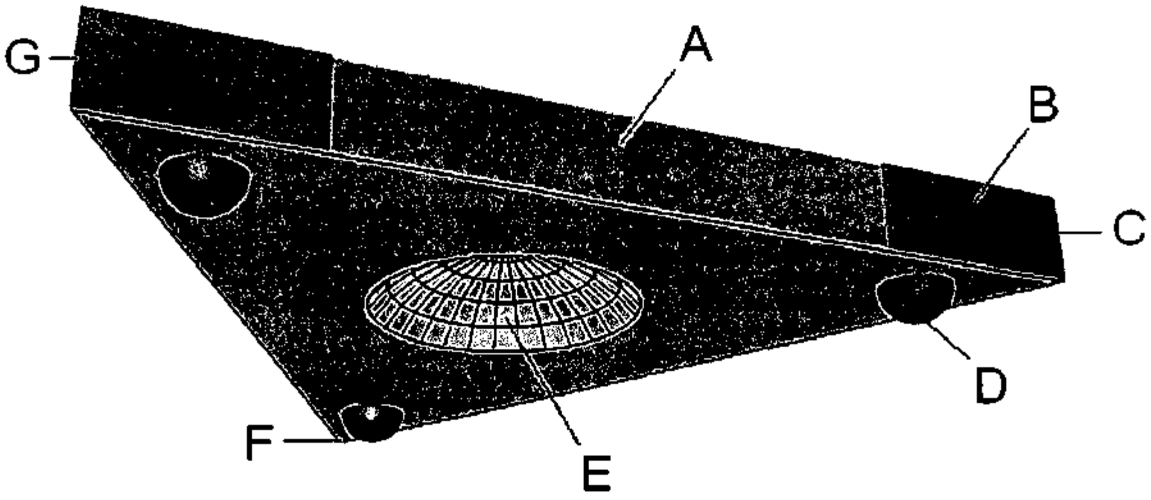

Figure 1

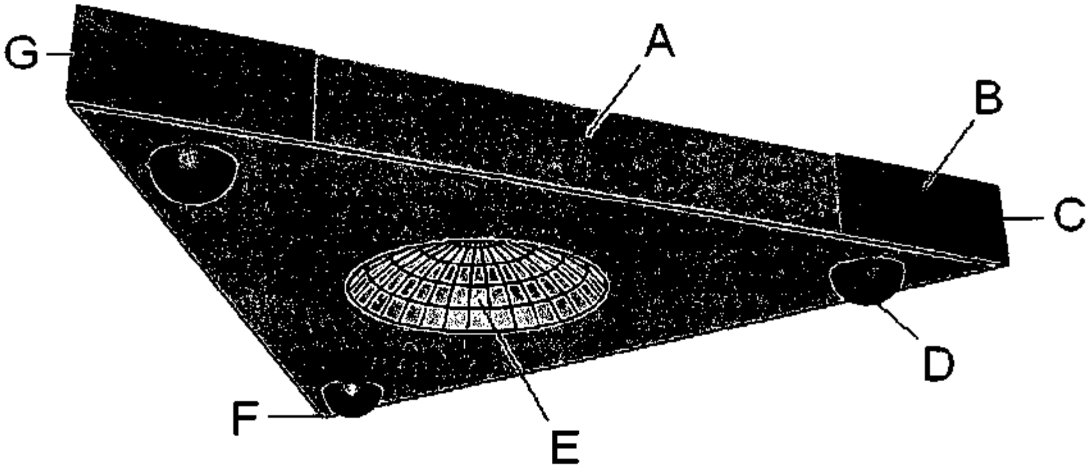

Figure 2

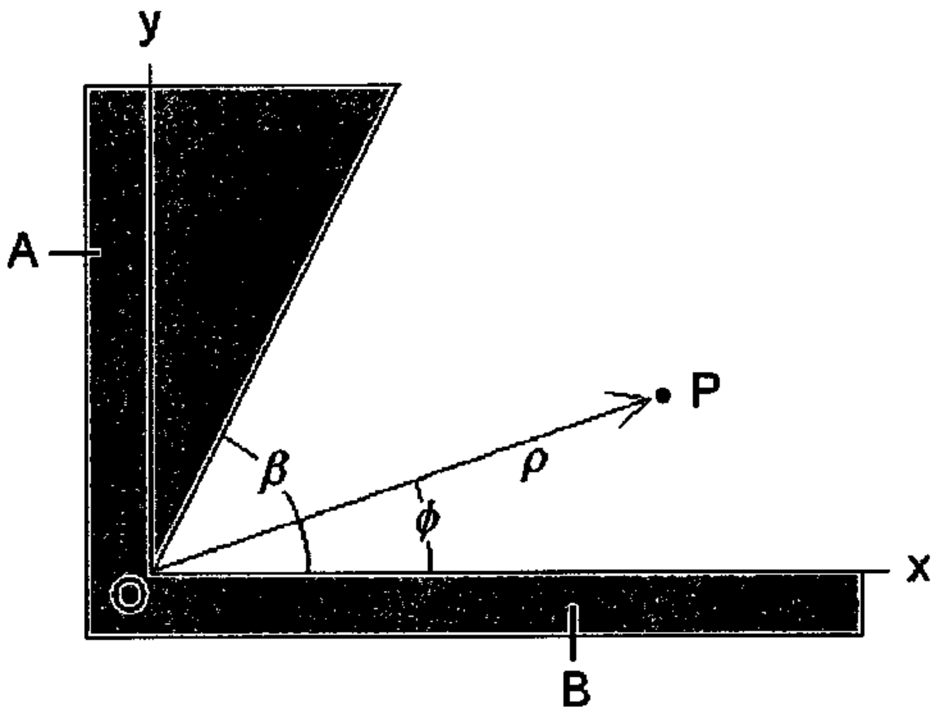

Figure 3

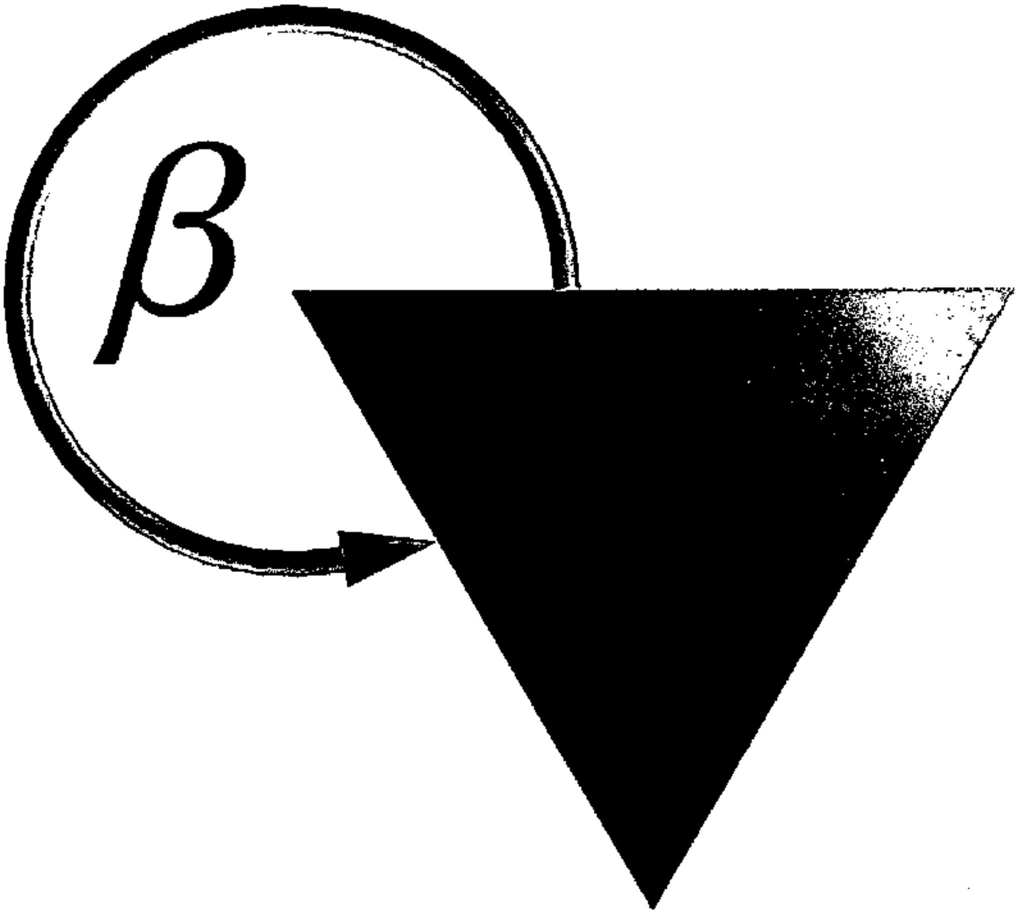

Figure 4

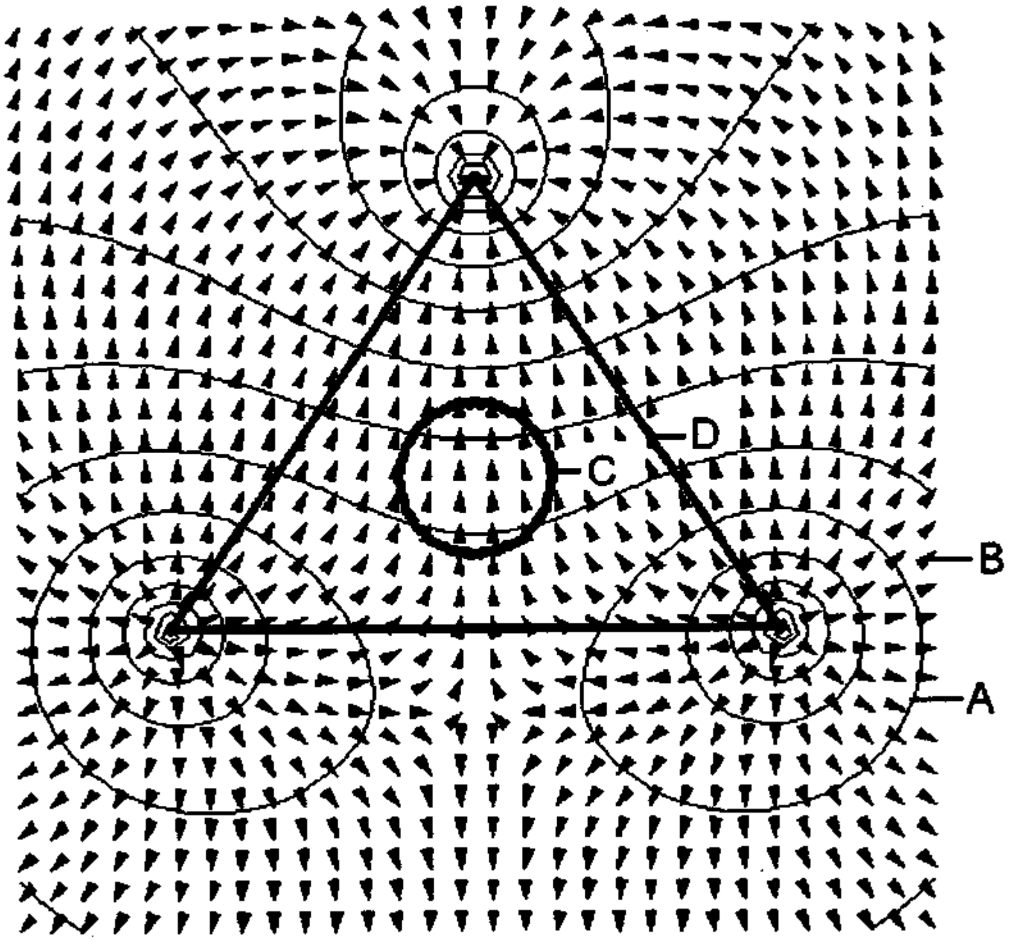

Figure 5

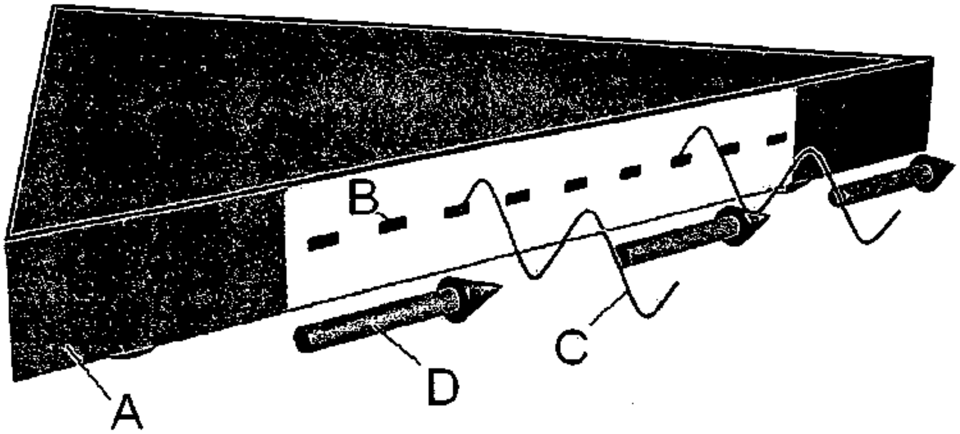

Figure 6

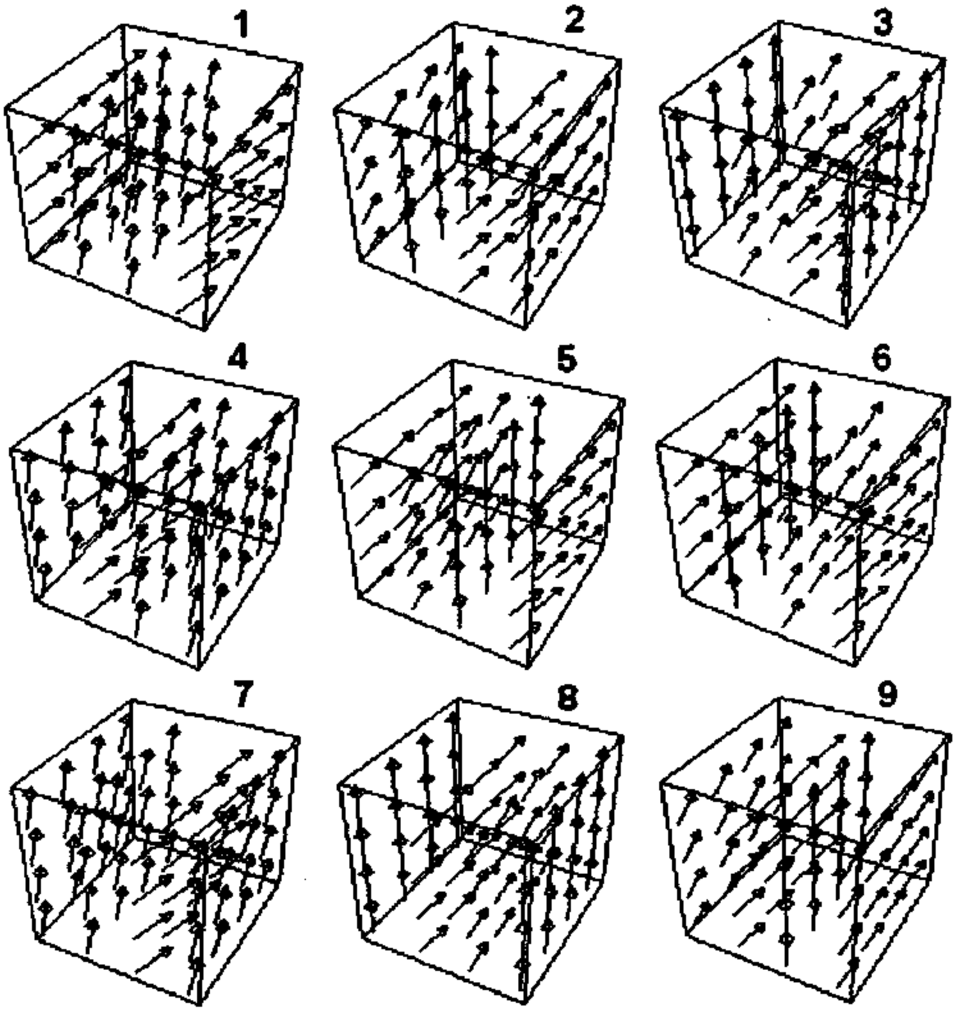

Figure 7

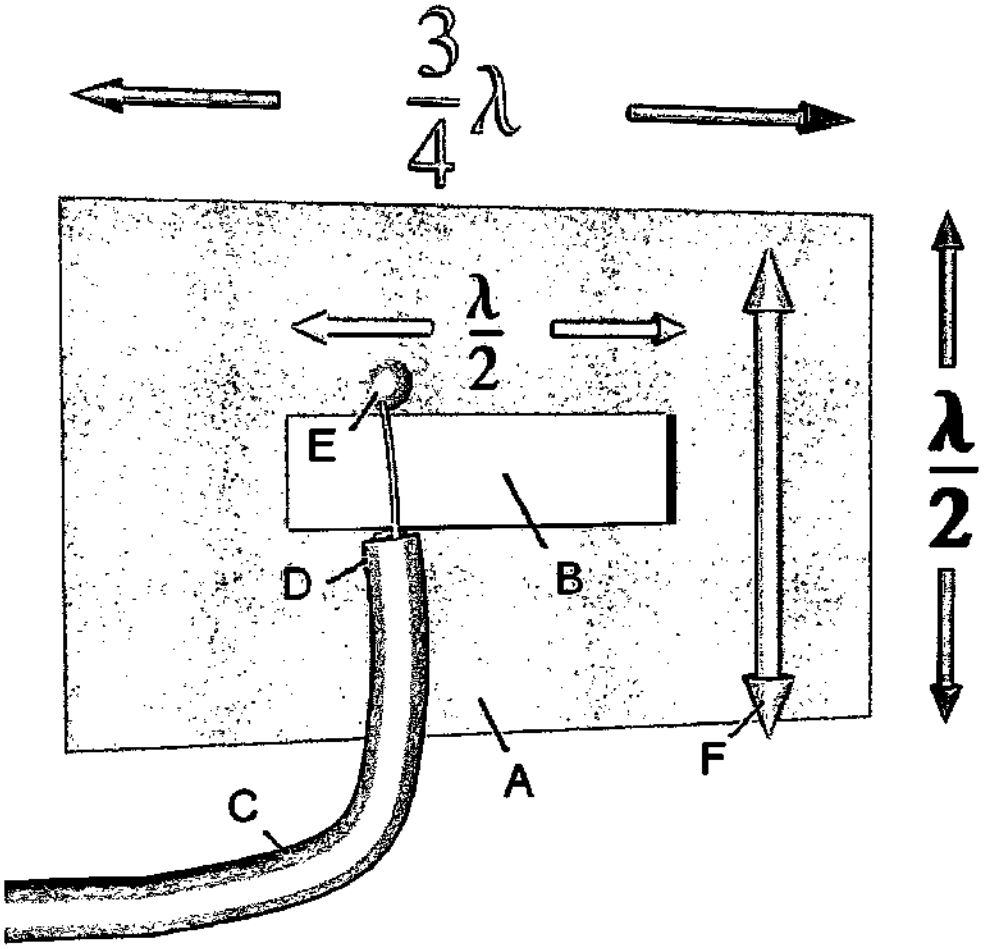
$\frac{3}{4}\lambda$
$\frac{\lambda}{2}$
$\frac{\lambda}{2}$
E
B
D
A
F
C

TRIANGULAR SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a spacecraft having a triangular hull with vertical electrostatic line charges on each corner. The line charges create a horizontal electric field that, together with a plane wave emitted by antennas on the side of the hull, generates a force per volume providing a unique combination of both lift and propulsion.

BACKGROUND OF THE INVENTION

[0002] Referring to **FIG. 1**, the spacecraft has a hull in the shape of an equilateral triangle. A parabolic antenna (E) is centrally located in the bottom of the hull. An array of horizontal slot antennas is located along the side of the hull (A). Each back corner (F,G) has a corner conducting plate which is charged to a positive voltage +V. The forward corner (C) has a conducting plate charged to a negative voltage −V. A motion control hemisphere (D) is located on the bottom surface in each of the three corners.

[0003] Referring to **FIG. 2**, two planes (A,B) intersect at the origin O at an opening angle β. Each plane (x,y) is charged to a voltage V. The potential at point P is determined in polar coordinates {ρφ}. The Laplace equation for the potential Φ in polar coordinates is given by:

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\Phi}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2\Phi}{\partial\phi^2}=0$$

Using a separation of variables solution, the potential is given as the product of two functions:

$$\Phi(\rho,\phi)=R(\rho)\Psi(\phi)$$

which when substituted into the Laplace equation becomes:

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right)+\frac{1}{\Psi}\frac{d^2\Psi}{d\phi^2}=0$$

Since the two terns are separately functions of ρ and φ respectively, each one has to be constant with the sum of the constants equal to zero:

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right)=\nu^2 \quad \frac{1}{\Psi}\frac{d^2\Psi}{d\phi^2}=-\nu^2$$

These two equations have solutions:

$$R(\rho)=a\rho^{\nu+b\rho-\nu}$$

$$\psi(\phi)=A\cos(\nu\phi)+B\sin(\nu\phi)$$

The azimuthal angle φ is restricted to a value in the range 0≦φ≦β. The boundary condition is that the potential Φ is equal to V for any radius ρ when φ=0 and φ=β. This means that v has to be an integer value of π so that the sine function is zero:

$$\sin(\nu\beta)=\sin\left(\frac{m\pi}{\beta}\beta\right)=\sin(m\pi)=0 \quad m=1,2\ldots$$

which in turn means that the coefficient A of the cosine term has to be zero in the solution above. Choosing b=0 makes the general solution for the potential equal to:

$$\Phi(\rho,\phi)=V+\sum_{m=1}^{\infty}a_m\rho^{m\pi/\beta}\sin(m\pi\phi/\beta)$$

which shows that when the angle is zero, the sine is zero and the potential is V. If the angle is β, then there is a multiple of π such that the sine is zero again.

[0004] Because the series involves positive powers of the radius, for small enough ρ, only the first term *m*=1 in the series is important. Thus around ρ=0, the potential is approximately

$$\phi(\rho,\phi)\approx V+a_1\rho^{\pi/\beta}\sin(\pi\phi/\beta)$$

[0005] The electric field component is the negative gradient of the potential:

$$E_\phi(\rho,\phi)=-\frac{1}{\rho}\frac{\partial\Phi}{\partial\phi}=-\frac{\pi a_1}{\beta}\rho^{(\pi/\beta)-1}\cos(\pi\phi/\beta)$$

The surface charge distribution σ at φ=0 and φ=β is equal to the electric field perpendicular to the surface times the permittivity of space ϵ_0:

$$\sigma(\rho)=\varepsilon_0E_\phi(\rho,0)=-\frac{\varepsilon_0\pi a_1}{\beta}\rho^{\frac{\pi}{\beta}-1}$$

Notice that if angle of intersection β is less than π, then the equation says that there is a very small radius to a positive power which means little charge density accumulation.

[0006] Referring to **FIG. 3**, the value of β, in the case of the triangular hull, is equal to 360° less 60° for a total of 300° or:

$$\beta=\frac{300}{180}\pi=\frac{5}{3}\pi$$

$$\rho^{\frac{\pi}{\frac{5}{3}\pi}-1}=\frac{1}{\rho^{\frac{2}{5}}}$$

which says that there is a charge density singularity to the two fifths power for small radius. Thus, the corner plates on the hull create a huge line charge density along the sharp vertical corner edge. The equation for the potential of a line charge density is given as:

$$\Phi(x,y)=-\frac{\lambda}{2\pi\varepsilon_0}\mathrm{Ln}((x-x_0)^2+(y-y_0)^2)$$

where λ is the charge per unit length in the vertical z-direction, and x_0 and y_0 are the location of the line charge in the xy-plane.

[0007] Referring to **FIG. 4**, the triangular hull (D) is plotted together with the potential contours (A) and the electric field arrows (B) created by the three corner line charges. The line charges are perpendicular to the paper. Notice that the electric field arrows are parallel crossing the center parabolic antenna (C). The electric field is also parallel to the sides (D) of the triangle.

[0008] Referring to **FIG. 5**, along the side of the triangle (A), an array (B) of horizontal slot antennas emit electromagnetic waves that have a vertically polarized electric E field (C). These traveling waves interact with the electric field (D) produced by the line charges on the corners of the triangle.

[0009] Using differential forms mathematics, this combination of fields is represented by the Hodge star of the differential of the wedge product of the two fields. The antenna electromagnetic field is a combination of a traveling magnetic field B_w, and electric field E_w. The stationary field E created by the line charges is perpendicular to the traveling wave.

$$*d(E\wedge(B_w+E_w\wedge dt))\frac{\varepsilon}{c}=\frac{\text{force}}{\text{volume}}$$

where ϵ is the linear capacitance of space and c is the speed of light. Thus there is a force per volume around the hull.

[0010] This combination of fields produces a spacetime curvature as determined by Einstein's General Theory of Relativity. The traveling electric field has an amplitude in the vertical z-direction and travels in the x-direction

$$E_w = E_z\cos(x-t)$$

The Faraday electromagnetic tensor contains all the electric and magnetic fields in all the {x,y,z} directions. The first row and first column contain the two electric fields

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} t & 0 & E_x & 0 & E_z\cos(x-t)\\ x & E_x & 0 & 0 & 0\\ y & 0 & 0 & 0 & 0\\ z & E_z\cos(x-t) & 0 & 0 & 0 \end{array}$$

The stress exerted on spacetime occurs in the xx, yy and zz-direction as calculated from the stress-energy tensor T of gravitational physics

$$4\pi T^{\mu\nu}=F^{\mu\alpha}F^{\mu}_{\alpha}-\frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

where g is the metric tensor for Cartesian space

$$g_{\alpha\beta}=\begin{array}{c|cccc|} t & -1 & 0 & 0 & 0\\ x & 0 & 1 & 0 & 0\\ y & 0 & 0 & 1 & 0\\ z & 0 & 0 & 0 & 1 \end{array}$$

where the diagonal components are the coefficients of the elementary spacetime length ds squared

$$(ds)^2=-(dt)^2+(dx)^2+(dy)^2+(dz)^2$$

The calculation produces three stresses T^{xx},T^{yy} and T^{zz} in their respective {x,y,z} directions.

[0011] Referring to **FIG. 6**, these three stresses are plotted together as a 3D vector field animated over time in nine frames. The graphs show that there is a lift force as depicted by the vertical arrows as well as a force of propulsion as shown by the interspersed horizontal arrows. With the passage of time, these vectors exchange places with each other so that the lift becomes the propulsion and vice versa, creating a wavy stress-energy field around the hull.

SUMMARY OF THE INVENTION

[0012] This invention is a spacecraft with a triangular hull having charged flat plates on the vertical corners of the three sides. The two rear corners are charged to a potential V. The forward corner is charged to a potential −V. The 60° angle on the corner creates a line charge density singularity that produces a huge horizontal electric field pointing from the back to the front of the craft which is also parallel to the sides of the triangle. An array of horizontal slot antennas located on the sides of the triangular hull produce an electromagnetic wave with the electric field polarized in the vertical direction. This combination of fields produces a spacetime force in both the vertical and horizontal directions such that the spacecraft receives a lift force and a force of propulsion.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0013] **FIG. 1**. Perspective view of triangular spacecraft.

[0014] **FIG. 2**. Drawing of the intersection of two charged plates in order to calculate the charge density in the corner.

[0015] **FIG. 3**. Perspective view of the corner angle β for the equilateral triangle.

[0016] **FIG. 4**. Planar 2D graph showing the electric field produced by three line charges on the corners of the triangular hull.

[0017] **FIG. 5**. Perspective view of electric field produced by the linear charge interacting with the traveling electromagnetic wave produced by the slot antenna.

[0018] **FIG. 6**. 3D vector animation of the lift and thrust force generated by the fields.

[0019] **FIG. 7**. Perspective view of slot antenna.

DETAILED DESCRIPTION OF THE INVENTION

[0020] Referring to **FIG. 7**, the antenna (A) is made out of sheet copper in which a rectangular horizontal slot (B) has been notched out using a die press and sheet metal fixture. A coaxial cable from the amplifier and frequency generator is attached across the slot by soldering the outer cable (D) to one side of the slot and the inner cable (E) to the other side of the slot. This creates the positive and negative charges across the gap which forms the vertical electric field (F) which radiates out perpendicularly to the copper sheet.

[0021] Although the invention has been described with reference to specific embodiments, such as a particular antenna system, those skilled in the art will appreciate that many modifications and variations are possible without departing from the teachings of the invention. All such modifications and variations are intended to be encompassed within the scope of the following claims.

1. A spacecraft comprised of the following components:

(a) a triangular hull in the form of an equilateral triangle;

(b) two copper plates attached on opposite vertical sides at each of the three corners of the hull (**1***a*) such that a sharp vertical edge is formed where they come together;

(c) an electrostatic generator used to charge the back two copper-cladded corners (**1***b*) to a high positive voltage, and the third forward copper-cladded corner to a high negative voltage;

(d) a horizontal slot antenna array mounted-on the sides of the hull; and

(e) a frequency generator, antenna and coaxial cables to drive the antenna array (**1***d*).

2. To create, by claims (**1***a*, **1***b*, **1***c*), an intense vertical line charge at the corners (**1***b*) and a horizontal electric field that that is parallel to the sides of the hull (**1***a*);

3. To create, by claims (**1***d*,**1***e*), an electromagnetic wave with a vertically polarized electric field traveling outward from the side of the hull (**1***a*); and

4. To create, by claims (**2**,**3**), an interaction of the electrostatic field (**2**) with the electromagnetic wave (**3**) such that a combined spacetime curvature pressure is generated on the hull in the upward and forward direction to produce lift and propulsion respectively.

* * * * *

US 20060112848A1

(19) **United States**
(12) **Patent Application Publication**
St. Clair
(10) **Pub. No.: US 2006/0112848 A1**
(43) **Pub. Date: Jun. 1, 2006**

(54) **PERMANENT MAGNET PROPULSION SYSTEM**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/001,217**

(22) Filed: **Dec. 1, 2004**

Publication Classification

(51) **Int. Cl.**
B61C 3/00 (2006.01)
(52) **U.S. Cl.** **105/49**

(57) **ABSTRACT**

This invention is a propulsion system for a train that uses permanent magnets mounted on a rotating iron cylindrical plate carrying a radial current in order to create a spacetime curvature distortion which pulls the locomotive along the track.

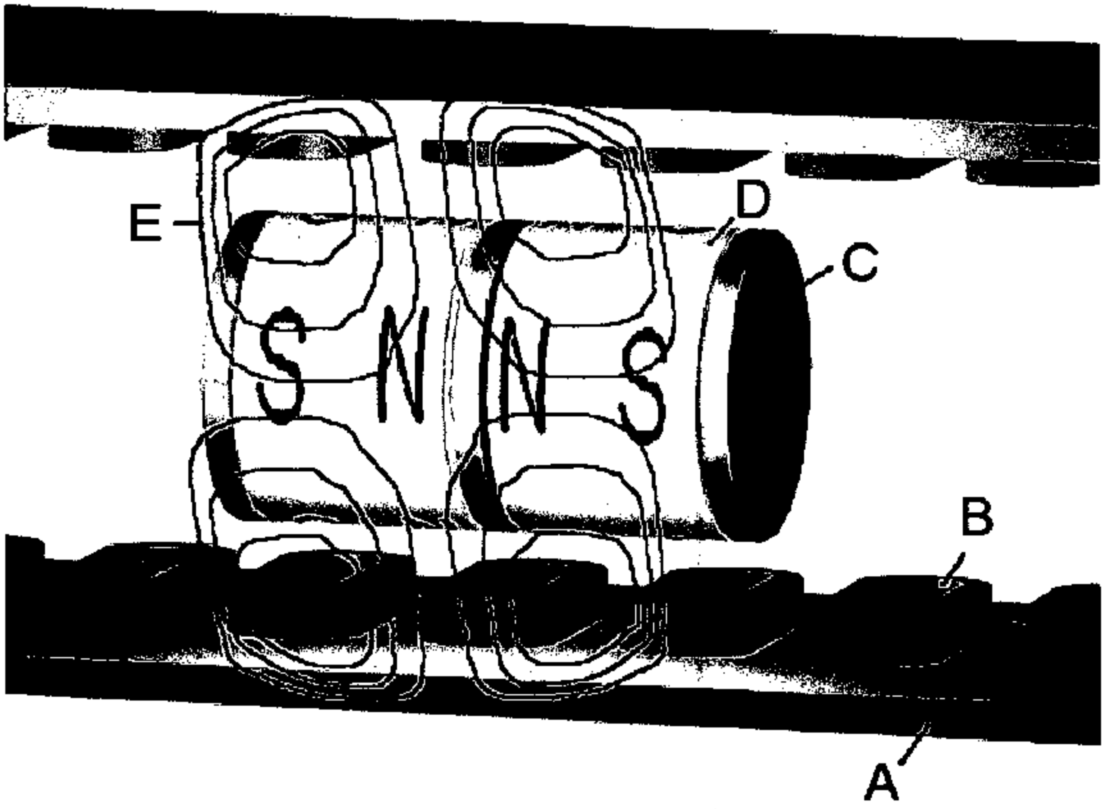

Figure 1

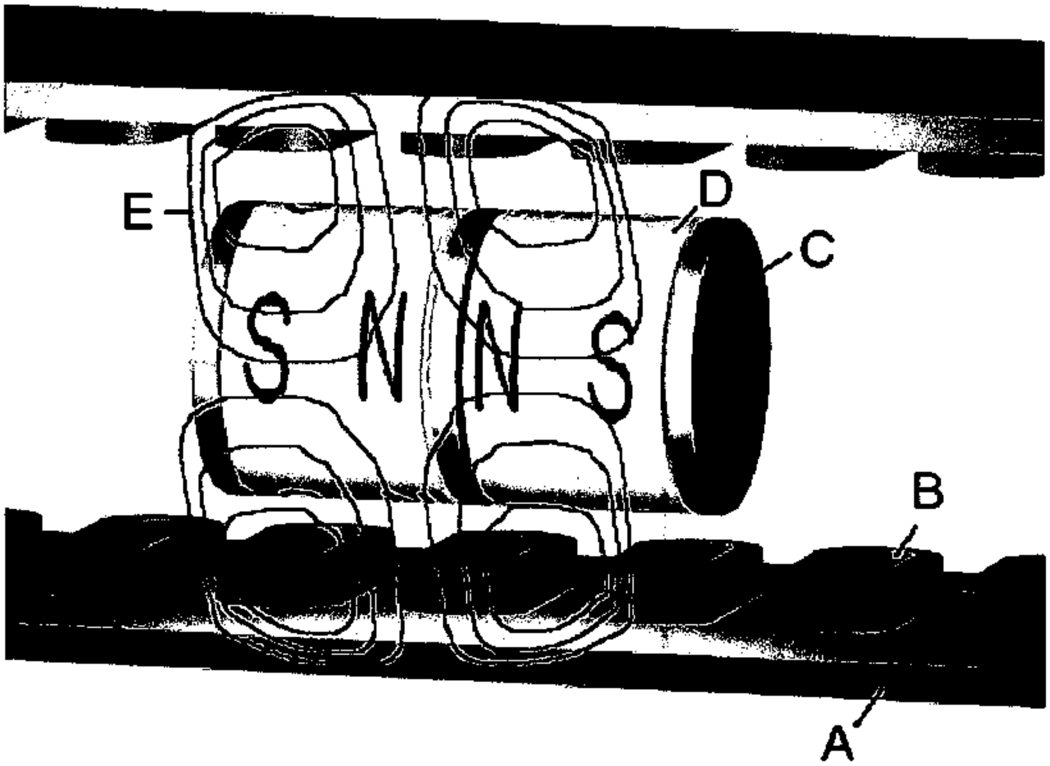

Figure 2

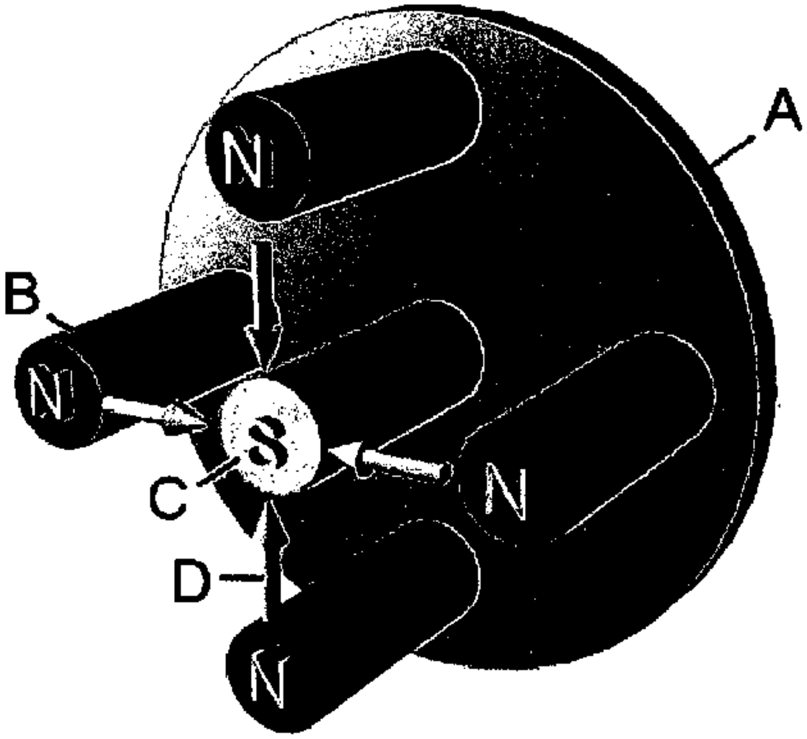

Figure 3

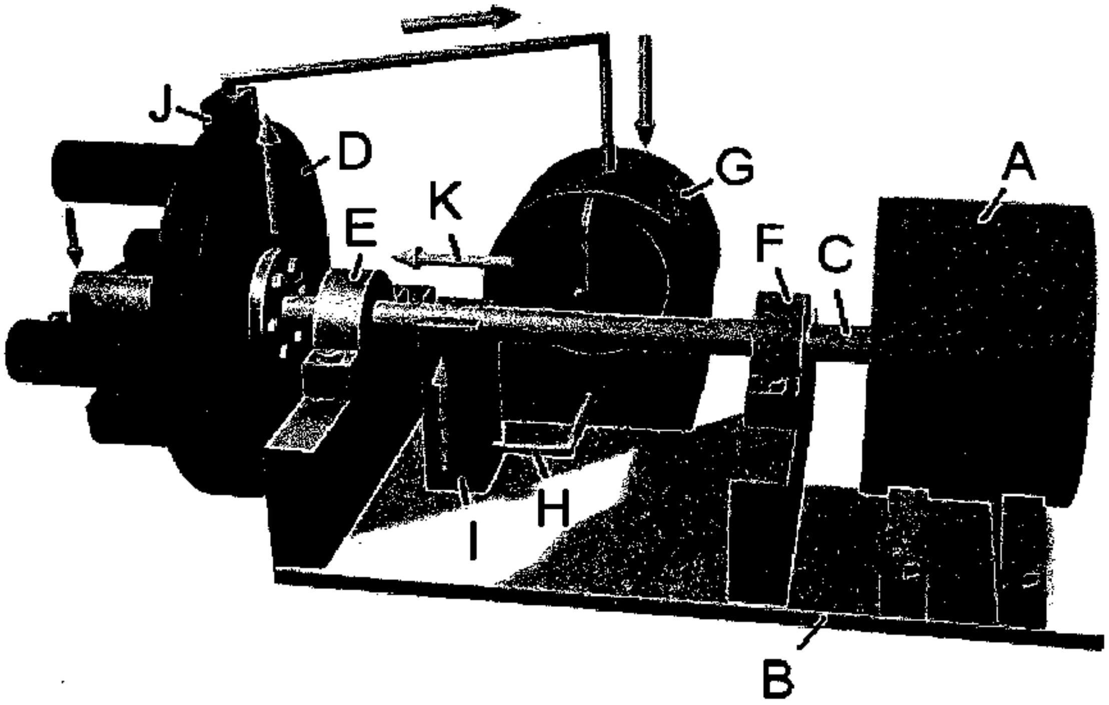
J
D
G
A
K
E
F
C
H
I
B

Figure 4

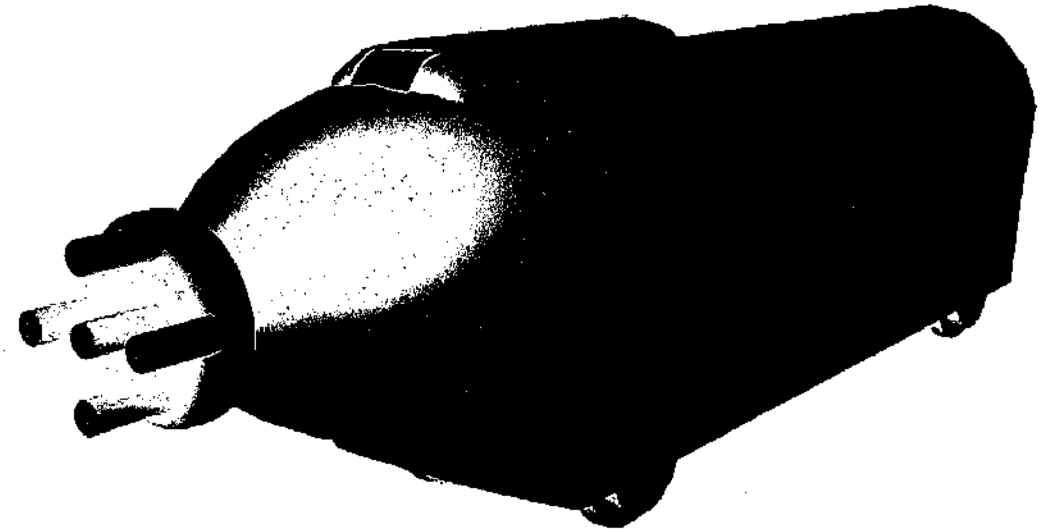

PERMANENT MAGNET PROPULSION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a propulsion system for a train that utilizes spinning cylindrical magnets in order to create a spacetime pressure distortion ahead of the vehicle that pulls the locomotive along the track.

BACKGROUND OF THE INVENTION

[0002] At the present time, referring to **FIG. 1**, proposed permanent magnet propulsion systems use a dual railway track (A) supporting a series of coil windings (B) located along the track. The vehicle is attached to two permanent magnets (D) between steel pole pieces (C). The north pole of each magnet faces the interior pole piece such that the magnetic flux path (E) follows the center pole piece up through the railway bed and then back to the south pole of the magnet. As the magnets move along the track, the coil windings are activated at the correct time by Hall sensors. With the coil energized as a north pole, the permanent magnet north pole is repelled which drives the vehicle along the track The problem with this design, and other similar designs, is that it is not practical to wind huge numbers of sensor-activated electrical coils along a steel track.

[0003] From Einstein's General Theory of Relativity, it is known that a spacetime curvature pressure develops perpendicular to direction of vibration of the electric and magnetic field. As an example, the photon has an electric field vibrating in the vertical y-direction and a magnetic field vibrating in the horizontal x-direction. The spacetime curvature pressure is therefore along the z-axis of radiation which pushes the negative mass of the photon along. Thus in order to create a spacetime curvature pressure in the z-direction along the track which would pull the train forward, a magnetic flux density field is required in the radial direction.

[0004] Referring to **FIG. 2**, four equally-spaced north permanent magnets (B) surrounding a centrally-located south permanent magnet (C) are mounted on an iron cylinder which acts as the radial flux return path. The magnetic flux density field (D) is in the radial direction from the north pole to the south pole. In order to provide strength, the magnets are molded onto a steel shaft and coated with epoxy so that they don't rust. During the molding process, a capacitor-discharge magnetizer is used to create the magnetic field of the magnet.

[0005] In Cartesian coordinates {-ct,x,y,z}, the elemental spacetime length ds squared is the sum of the squares of the incremental lengths {cdt,dx,dy,dz}

$$(ds)^2=-(dt)^2+(dx)^2+(dy)^2+(dz)^2$$

where the speed of light c is unity. The coefficients (-1,1,1,1) of this equation make up the g metric 4×4 tensor

$$g_{\alpha\beta}=\begin{array}{c|cccc|} & [t & x & y & z] \\ t & -1 & 0 & 0 & 0 \\ x & 0 & 1 & 0 & 0 \\ y & 0 & 0 & 1 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

[0006] The Faraday electromagnetic tensor contains the magnetic fields which determine how the spacetime length ds is curved. For a magnetic flux density field in the x-direction, Bx, and a magnetic flux density field in the y-direction, By, the Faraday tensor is

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} & t & x & y & z \\ t & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & -By \\ y & 0 & 0 & 0 & Bx \\ z & 0 & By & -Bx & 0 \end{array}$$

The stress-energy-momentum tensor T, which determines how space is curved, is calculated from the following equation

$$4\pi T^{\mu\nu}=F^{\mu\alpha}F^{\nu}_{\alpha}-\frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

The stress-energy in the z-direction ahead of the locomotive is

$$T^{zz}=\frac{B_x^2+B_y^2}{8\pi}=\frac{B_r^2}{8\pi}$$

where the sum of the squares of the fields in the x and y directions is the radial B field. In Einstein's General Relativity Theory, the curvature G tensor is equal to the stress-energy tensor divided by 8π. The G tensor is the curvature of space having units of inverse radius squared.

$$G=\frac{T}{8\pi}$$

Therefore the curvature G_{zz} generated along the z-direction ahead of the train is proportional to the square of the magnetic flux density field

$$G_{zz}=\frac{1}{r^2}=\frac{G\varepsilon}{c^2}\frac{B_r^2}{8\pi}=\frac{1}{meter^2}$$

where G is Newton's gravitational constant (not to be confused with the curvature tensor), ϵ is the linear capacitance of space, and c is the speed of light. The linear mass of space Ω is the speed of light c squared divided by the gravitational constant G, so that the equation can be written as

$$\frac{G\varepsilon}{c^2}\frac{B_r^2}{8\pi}=\frac{\varepsilon}{\Omega}\frac{B_r^2}{8\pi}=\frac{1}{\frac{\Omega}{\varepsilon}}\frac{B_r^2}{8\pi}$$

where the conversion factor is the square of the magnetic vector potential A

$$\sqrt{\frac{\Omega}{\varepsilon}} = \frac{\text{kgm}}{sec coul} = A$$

which is actually the momentum per charge. Therefore the curvature equation can be written as

$$\frac{1}{r^2} = \frac{1}{8\pi}\left(\frac{B_r}{A}\right)^2$$

This equation shows that it is necessary to create a magnetic vector potential together with the radial magnetic flux density field in order to create a curvature of space. Looking at the units of A shows that it is a mass momentum per charge

$$A = \frac{\text{kg}}{sec}\frac{m}{coul} = \frac{m\omega^2 r}{I}$$

or a mass m rotating with angular velocity c) per current along the radius. In terms of the invention, what this means is that the mass of the iron cylinder has to be rotating and there has to be a radial electrical current I in order to produce the linear charge along the radius. The differential mass dm depends on the circumference times the differential radius dr, the mass density p, and the length L of the cylinder

$$dm = \rho 2\pi r L dr$$

so that the magnetic vector potential becomes

$$A = \int_0^R \frac{\rho 2\pi r L\omega^2 r}{I}\,dr = \frac{2}{3}R^3\rho\pi L\frac{\omega^2}{I}$$

The value of A for the iron cylinder is

$$L = .2m$$

$$\rho = 7866\frac{kg}{m^3}$$

$$R = 1m$$

$$\omega = 2\pi f = 6.28\ \text{sec}^{-1}$$

$$I = 3000000\ \text{amp}$$

$$A = .04335\frac{kgm}{\text{sec}coul}$$

$$Br = 1.2tesla$$

$$\frac{1}{8\pi}\left(\frac{Br}{A}\right)^2 = 30.47m^2$$

$$r_{curvature} = \sqrt{8\pi}\left(\frac{A}{Br}\right) = .181m$$

What makes this possible is that the new N-machines can easily generate a minimum of 6 million amps which is twice the value of the electrical current above.

[0007] Referring to **FIG. 3**, the assembly consists of a large induction motor (A) mounted on the train's base plate (B) driving a motor shaft (C) attached to the iron cylinder (D). The shaft is held in place by two thrust bearings mounted in two pillow blocks (E,F). The current-generating N-machine (G) is electrically connected by a copper bus (H) to a copper-beryllium brush (I) on the motor shaft with a similar return brush (J) on the edge of the iron cylinder. The current (K) flows through the motor shaft to the center of the rotating cylinder and then radially outward to the edge. The magnetic flux density flows from the north poles of the outer permanent magnets to the central south pole, along the central magnet to the center of the rotating cylinder and then radially outward to the south poles of the outer magnets.

[0008] The thrust F developed is the radius of curvature of spacetime r_c calculated above times the magnet flux density field times the current I

$$F = \frac{r_c B_r I}{\sqrt{8\pi}} \approx 30000lbf$$

Using conservation of tensor coordinates, the radius of curvature is in the z-direction, the magnetic flux density field is in the radial direction and the current is in the radial direction

$$F^z = x^z B_1 I^r$$

where the radial indices cancel, leaving the z-index as the direction of the force.

SUMMARY OF THE INVENTION

[0009] It is the object of this invention to create a spacetime curvature in front of a train locomotive in order to pull the vehicle along the track It is known from gravitational physics that a spacetime curvature is generated perpendicular to the direction of vibration of the electric and magnetic field. A radial magnetic field, which can be produced by permanent magnets attached to the flat faces near the rim of a iron cylinder rotating about the z-axis, will create a curvature in the z-direction. Four cylindrical north-pole-oriented magnets produce a radial magnetic flux density with is channeled into a central cylindrical south-pole-oriented magnet. The flux lines then flow radially outward through the steel rotating cylinder and reconnect with the south poles of the four outer magnets. The rotating iron cylinder generates the equivalent of a magnetic vector potential when an electrical current flows from the center of the cylinder to the edge. This current is generated by an N-machine current generator. The square of the magnetic flux density divided by the magnetic vector potential is equal to the spacetime curvature. The square root of the inverse of the spacetime curvature is the radius of curvature. The thrust developed is this radius of curvature times the magnetic flux density field times the current.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0010] **FIG. 1**. Perspective view of proposed permanent magnetic propulsion system using coil windings on the steel track.

[0011] **FIG. 2**. Perspective view of permanent magnet rotor assembly.

[0012] **FIG. 3**. Perspective view of system showing motor drive, N-machine and permanent magnet rotor.

[0013] **FIG. 4**. Perspective view of locomotive and rotor/magnet assembly.

DETAILED DESCRIPTION OF THE INVENTION

[0014] 1. The permanent magnets are made of neodymium-iron-boron material which is heated to its melt temperature and injection molded around a steel shaft threaded at one end while at the same time a pulsed magnetic field is applied to the material using a charge-discharge magnetizer. Because of the iron in the material, a coat of epoxy is applied to the magnet in order to protect it from the environment. Holes are drilled into the iron plate 90° apart near the rim, threaded, and then the steel shaft with the magnet is then inserted. Another hole is drilled and tapped in the center of the circular plate for attaching the south pole magnet which is used as the return path for the magnetic flux.

[0015] 2. Another easier way to make the magnets is to purchase short lengths of tubular NdFeB magnets and then stack them on the steel shaft with a cylindrical iron pole piece on the end of the shaft. The pole piece then holds the magnets down in place when the shaft is threaded into the plate.

[0016] 3. Referring to **FIG. 4**, the propulsion system is mounted inside the train cabin such that the rotor/magnet assembly extends out in front of the locomotive where the spacetime curvature is generated.

I claim:

1. A train propulsion system consisting of the following components:

a. a rotating iron cylindrical plate rotor of high relative permeability driven by an induction motor and horizontal steel motor shaft mounted in pillow block thrust bearings;

b. four cylindrical magnets, each molded to a steel support shaft threaded into the iron plate at 90° intervals around the rim of the plate with their north poles facing away from the plate;

c. a fifth cylindrical magnet molded to a steel support shaft which is threaded into the center of the iron plate with the south pole facing away from the plate;

d. an N-machine current generator supplying a radial electrical current from the center of the rotating plate by means of a copper-beryllium brush on the motor shaft (**1***a*) and another similar brush on the outside edge of the rotor.

e. a locomotive train on which the components are mounted such that the rotor/magnet assembly extends out in front of the locomotive with the rotor's angular velocity vector pointing along the track.

2. a closed magnetic flux path along a radial path in air from the north poles of the four outer magnets (**1***b*) to the south pole of the central magnet (**1***c*), through the center magnet and then radially outward through the rotor (**1***a*), returning back through the four outer magnets, such that the flux and electrical current (**1***d*) flow in the same outward radial direction through the rotor.

3. the creation of a spacetime curvature due to claims (**1***a* through **2**) that produces a large force on the locomotive equal to the radius of the spacetime curvature times the flux times the current.

* * * * *

US 20060180473A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0180473 A1**

St. Clair (43) **Pub. Date: Aug. 17, 2006**

(54) **WATER ENERGY GENERATOR**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/060,037**

(22) Filed: **Feb. 17, 2005**

Publication Classification

(51) **Int. Cl.**
C25D 5/34 (2006.01)

(52) **U.S. Cl.** **205/210**

(57) **ABSTRACT**

A water energy generator that generates electricity using a magnetic vortex wormhole generator and a water injector/vacuum chamber to produce low density hyperspace energy that causes the hydrogen atoms of water molecules to decay into electron pairs.

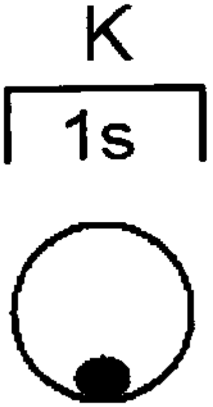

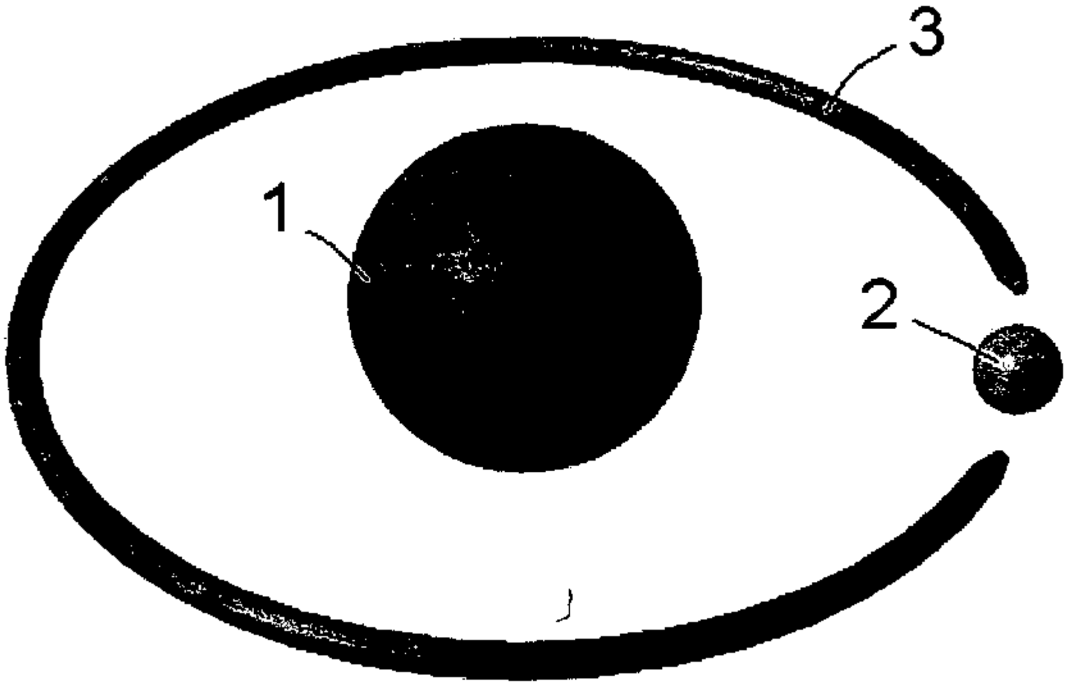

Figure 1

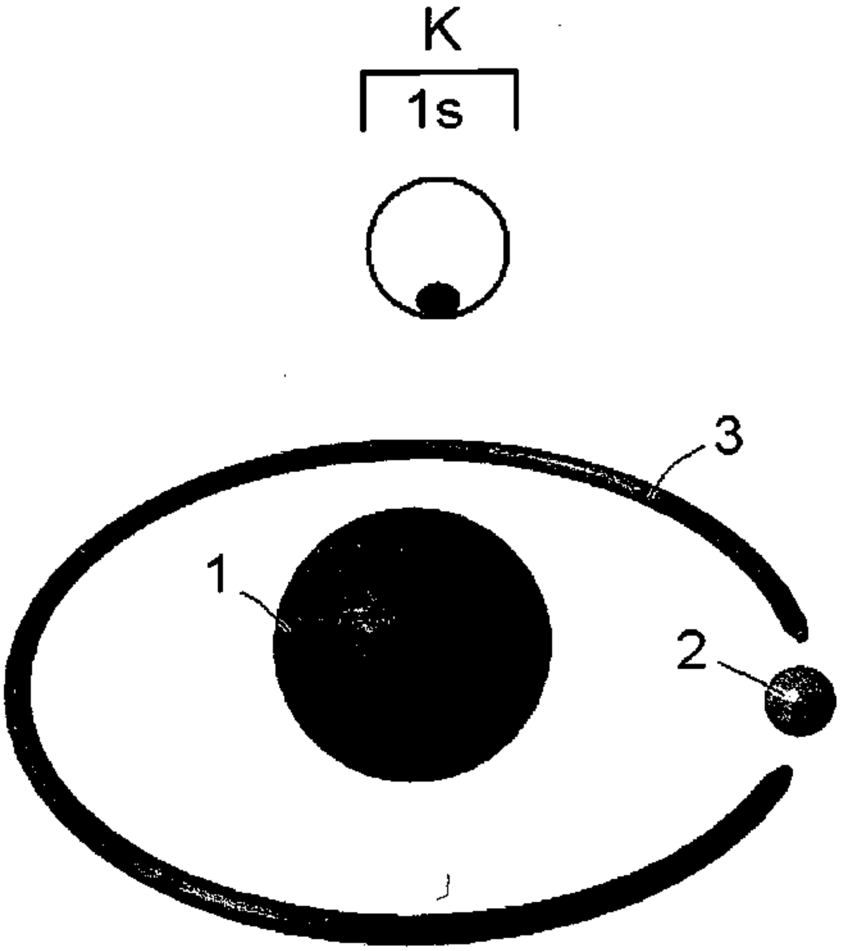

Figure 2

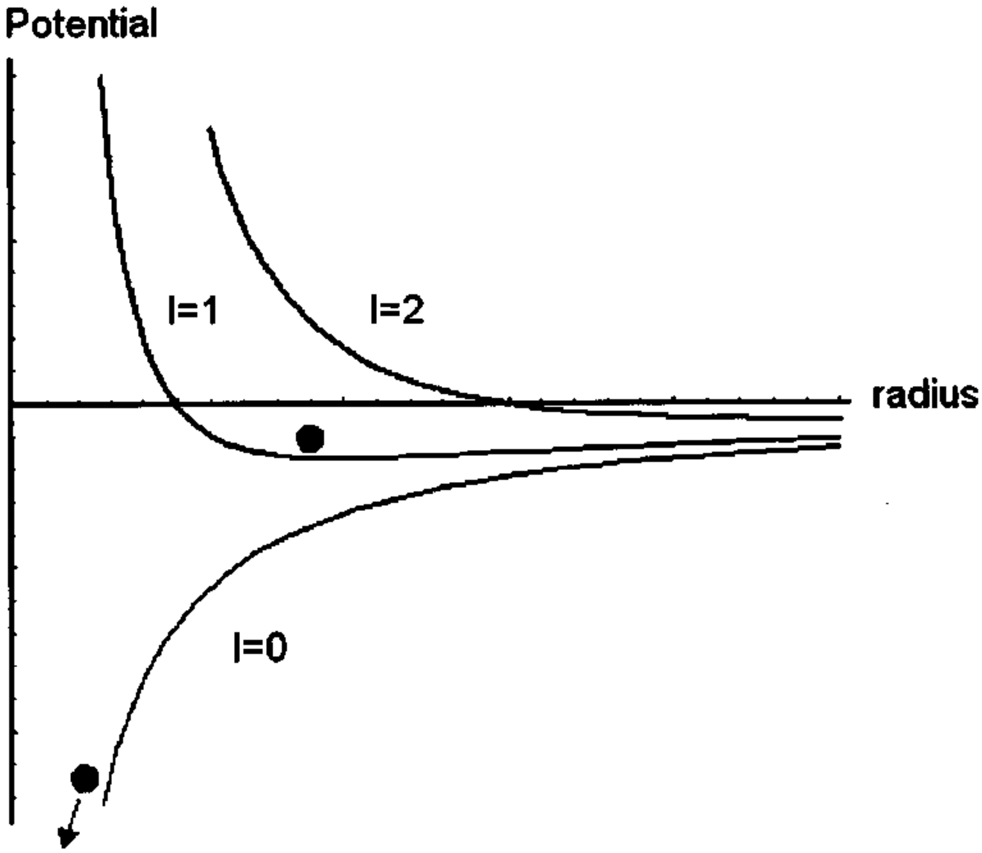

Figure 3

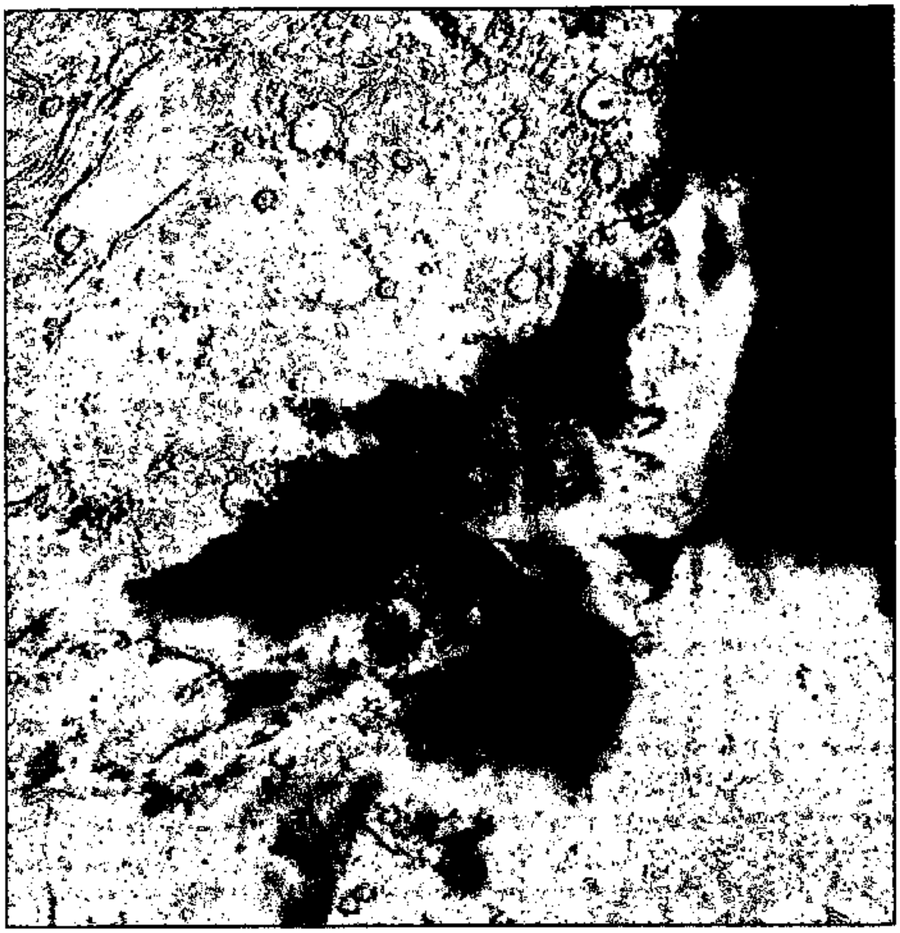

Figure 4

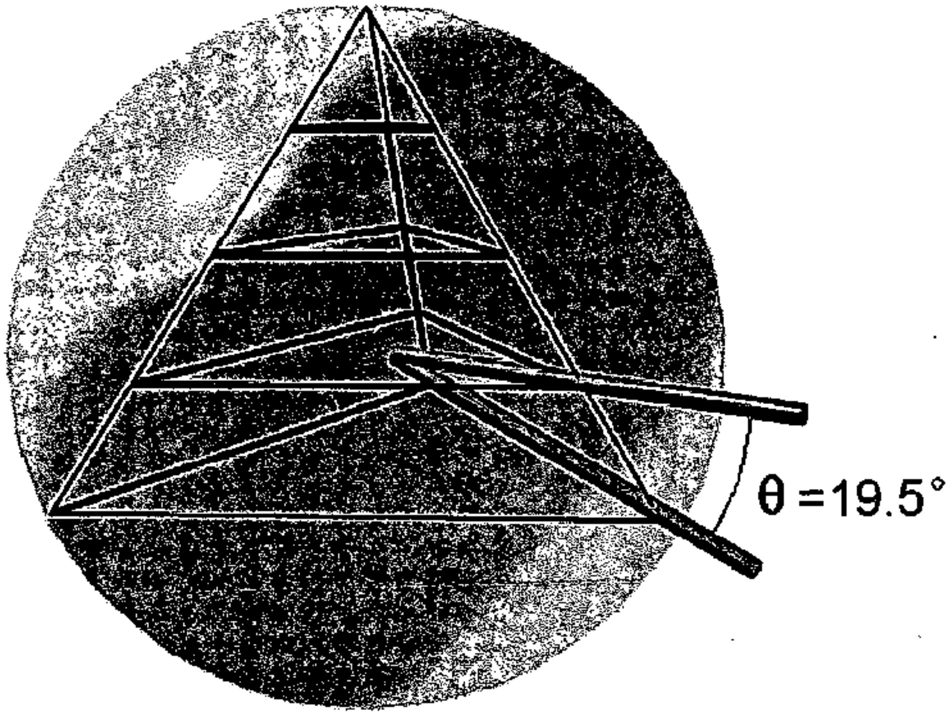

Figure 5

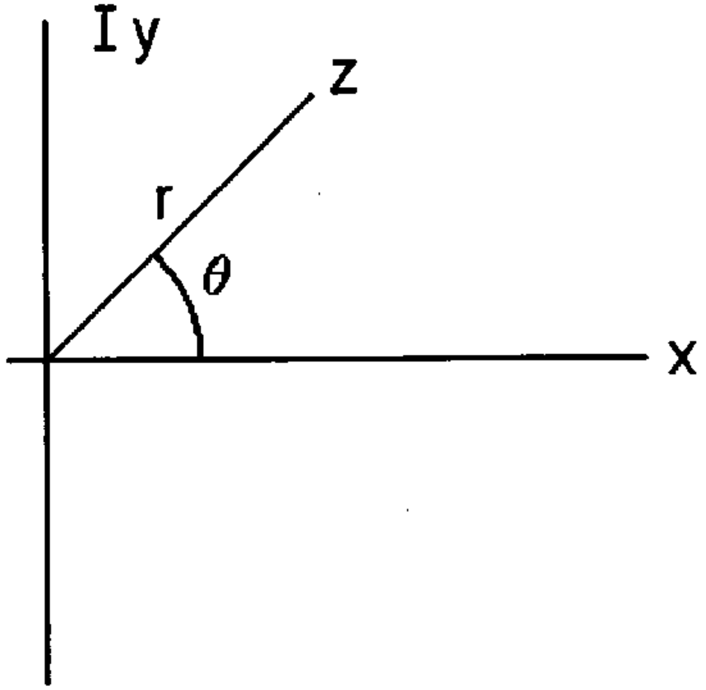
Iy
z
r
θ
x

Figure 6

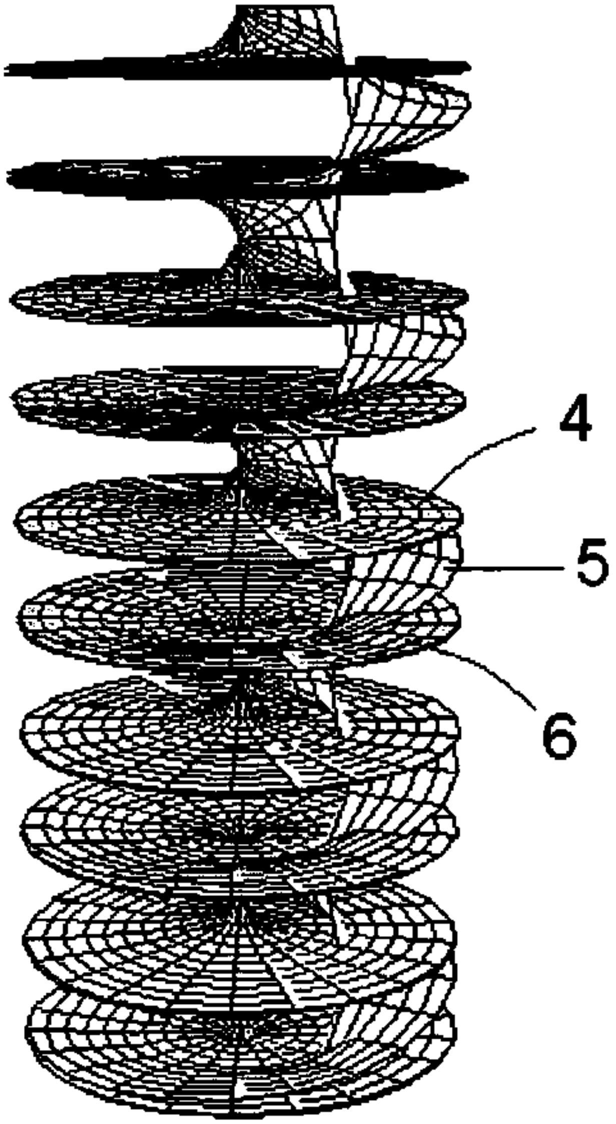

Figure 7

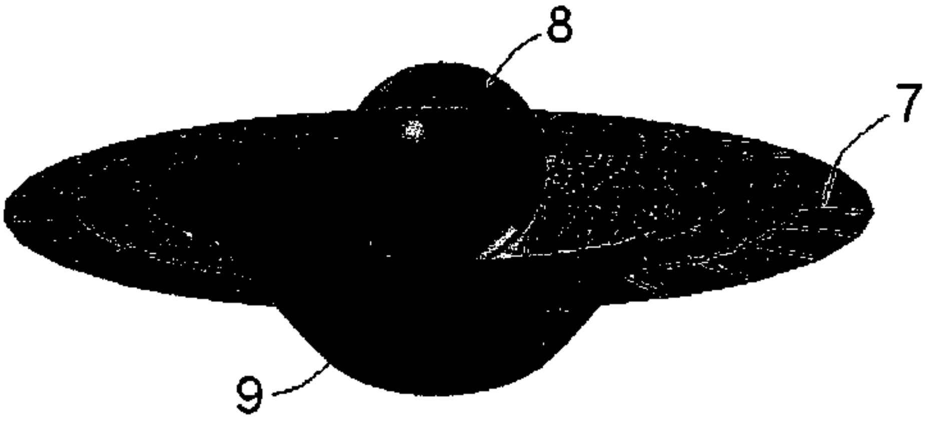

Figure 8

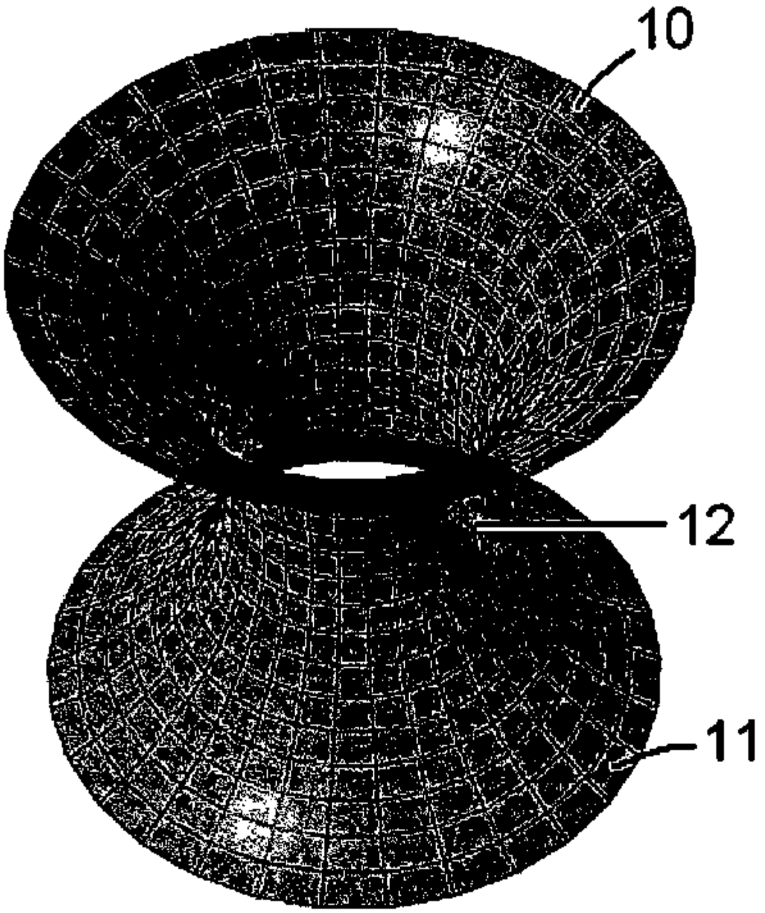

Figure 9

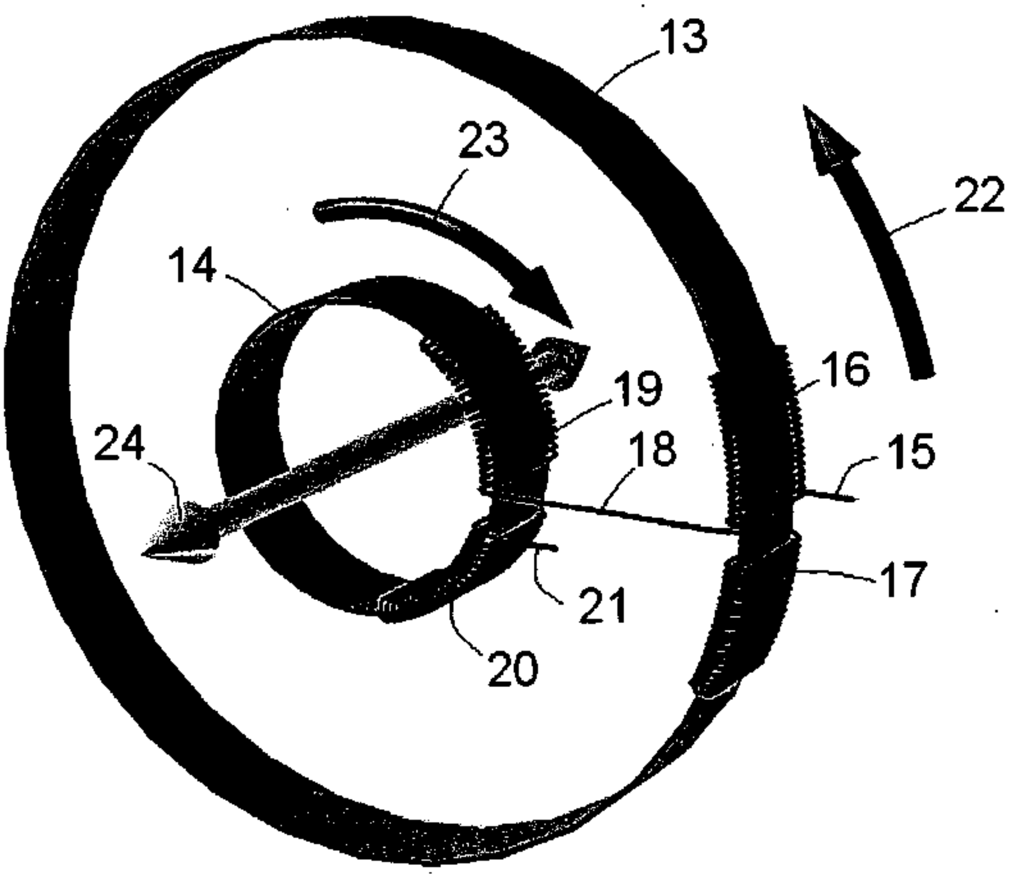
13
23
22
14
16
19
24
18
15
17
21
20

Figure 10

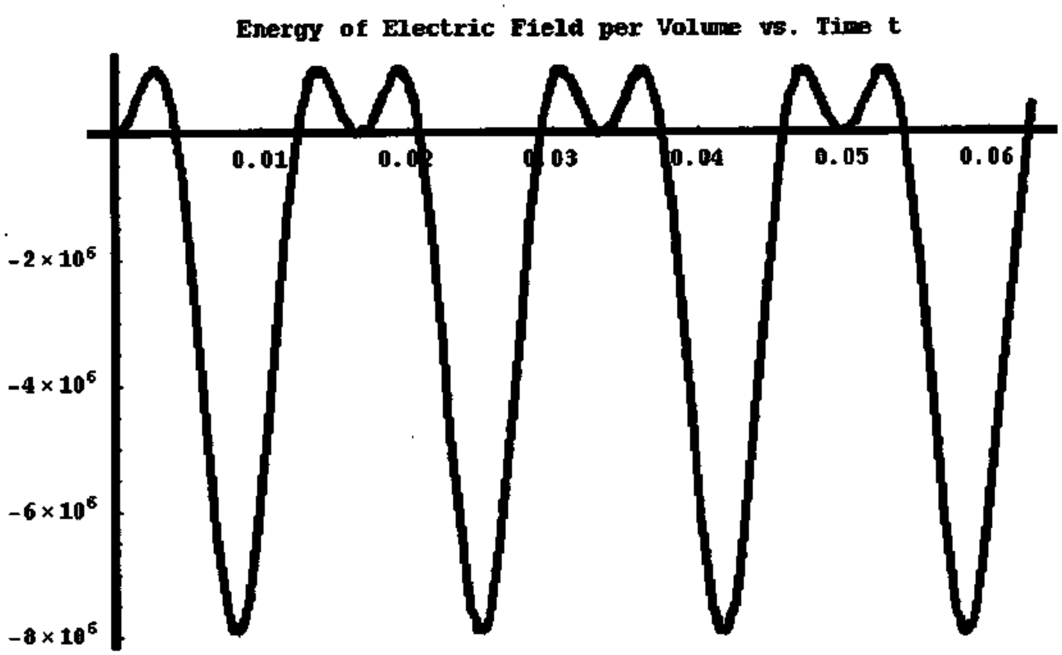
Energy of Electric Field per Volume vs. Time t
0.01
0.02
0.03
0.04
0.05
0.06
-2 × 10⁶
-4 × 10⁶
-6 × 10⁶
-8 × 10⁶

Figure 11

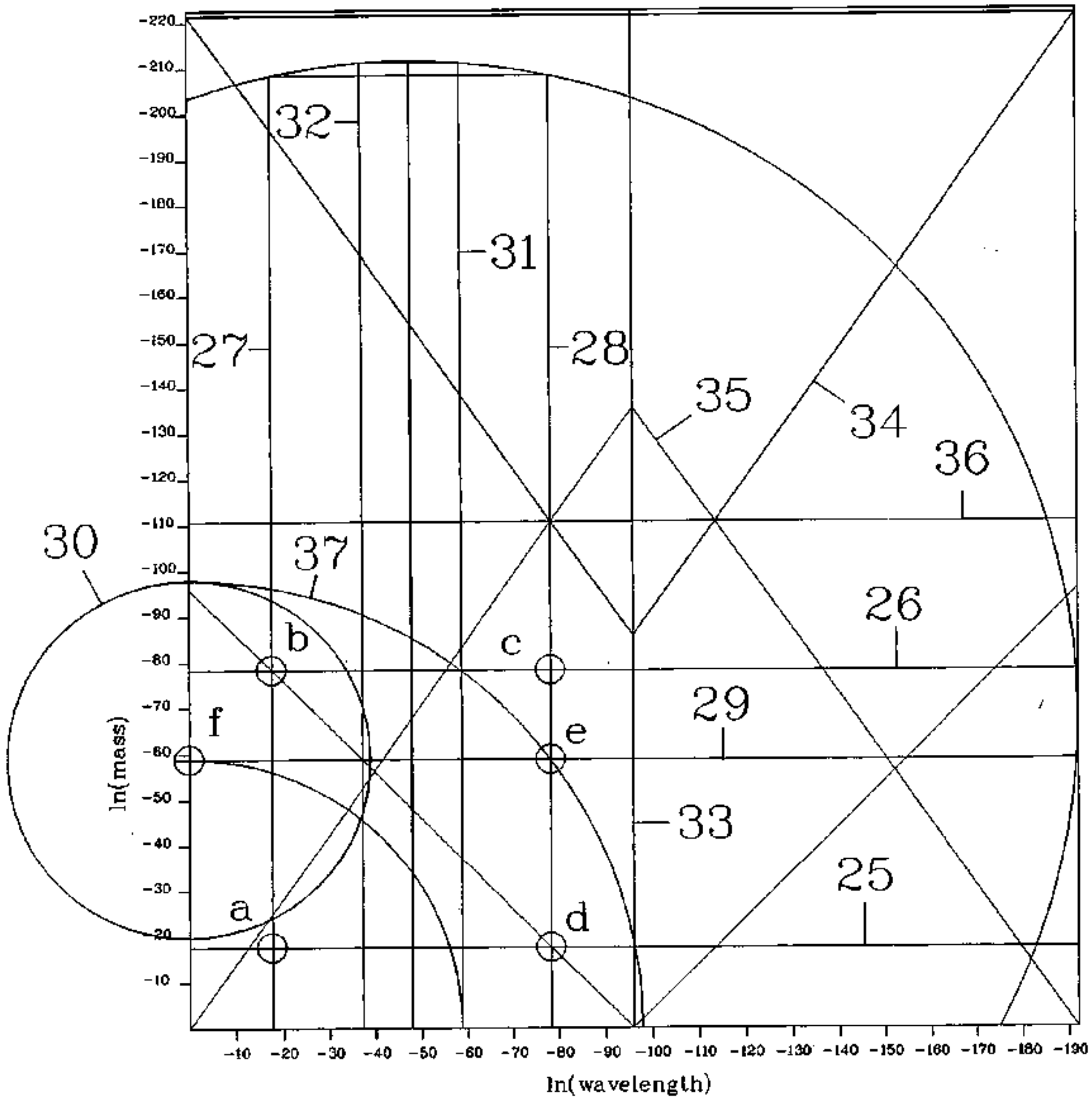
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
32
31
27
28
35
34
36
30
37
26
b
c
29
f
e
33
a
d
25

Figure 12

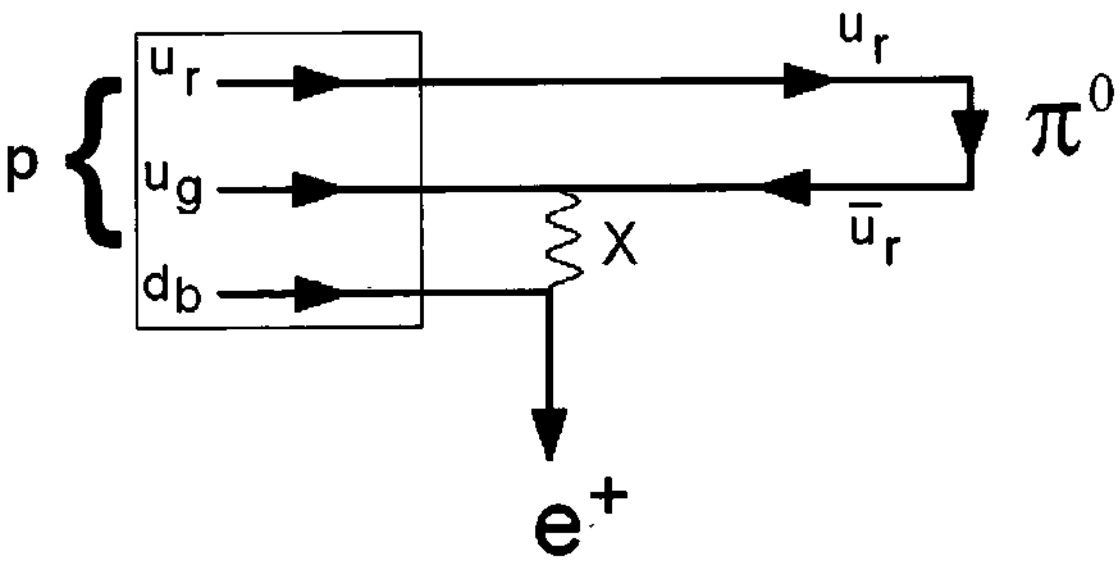
p
u_r
u_g
d_b
u_r
π^0
$\bar{u}_r$
X
e^+

Figure 13

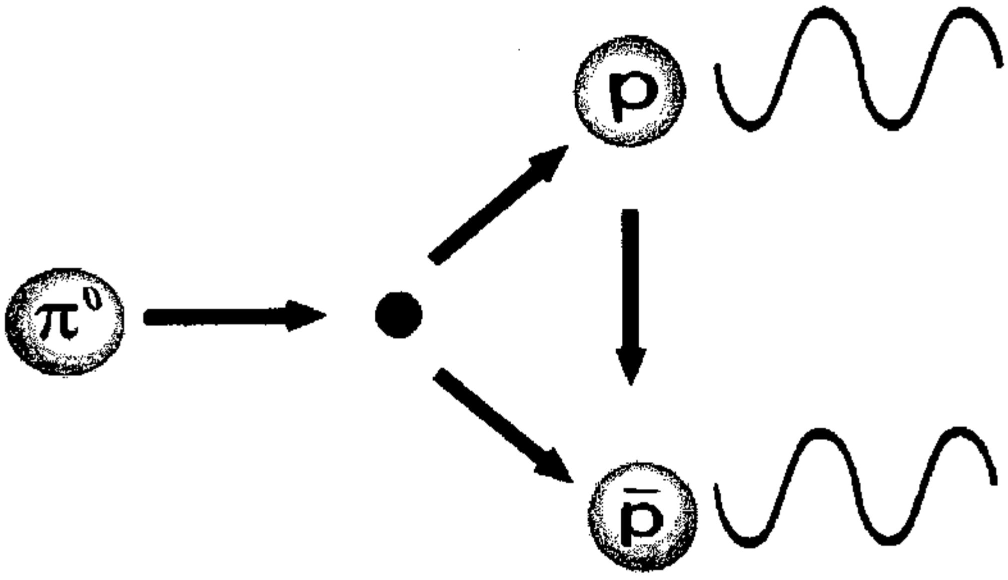

Figure 14

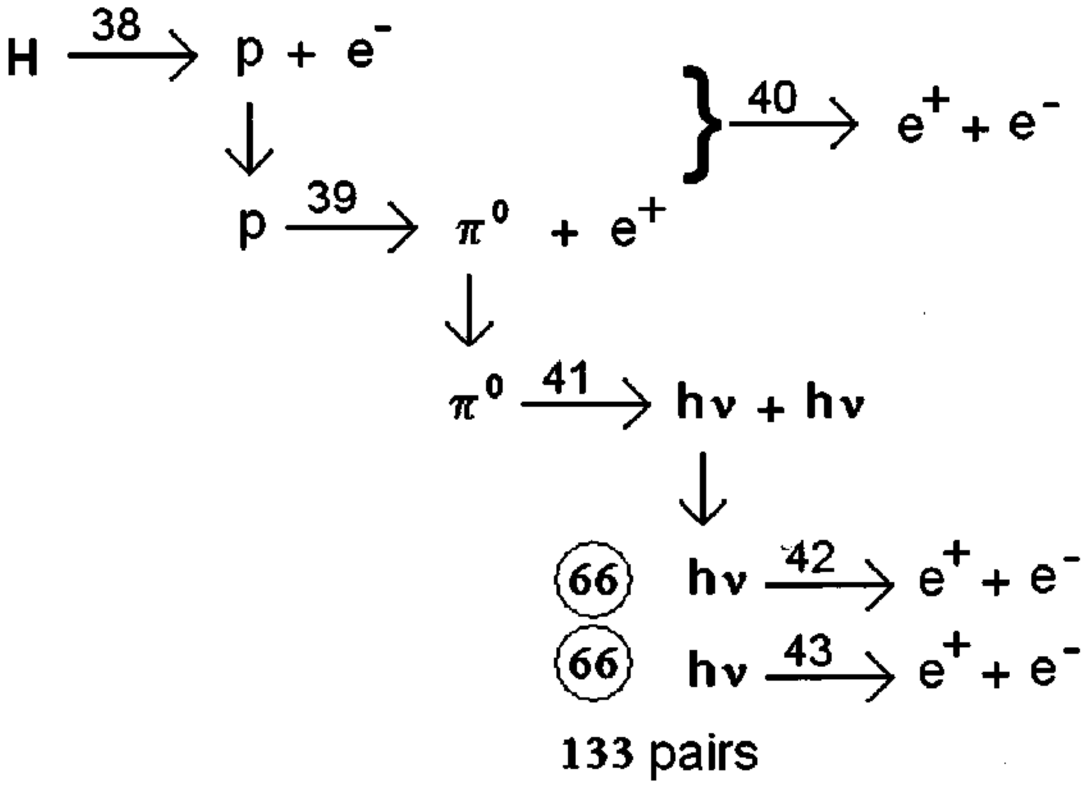
H
38
p + e-
40
e+ + e-
p
39
π0 + e+
π0
41
hν + hν
66
hν
42
e+ + e-
66
hν
43
e+ + e-
133 pairs

Figure 15

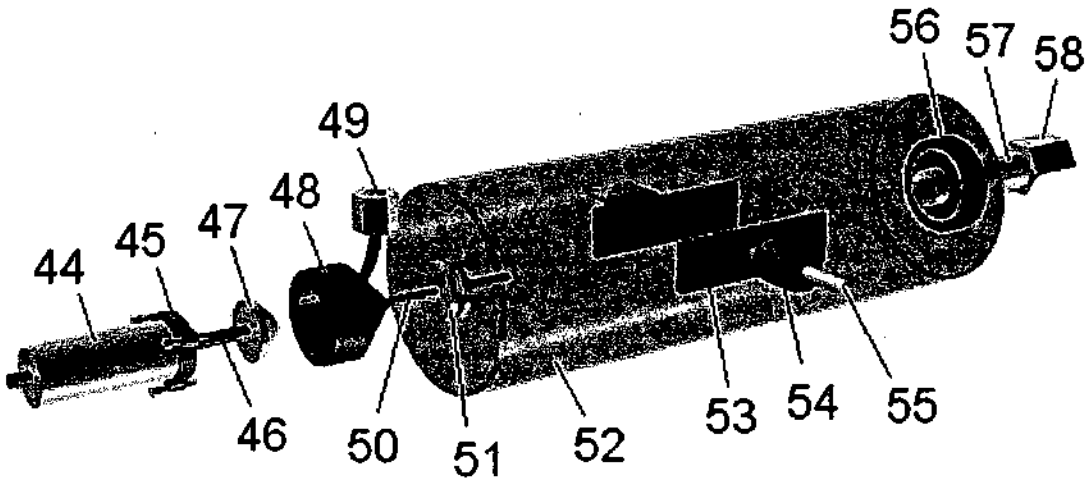
56
57
58
49
48
47
45
44
46
50
51
52
53
54
55

Figure 16

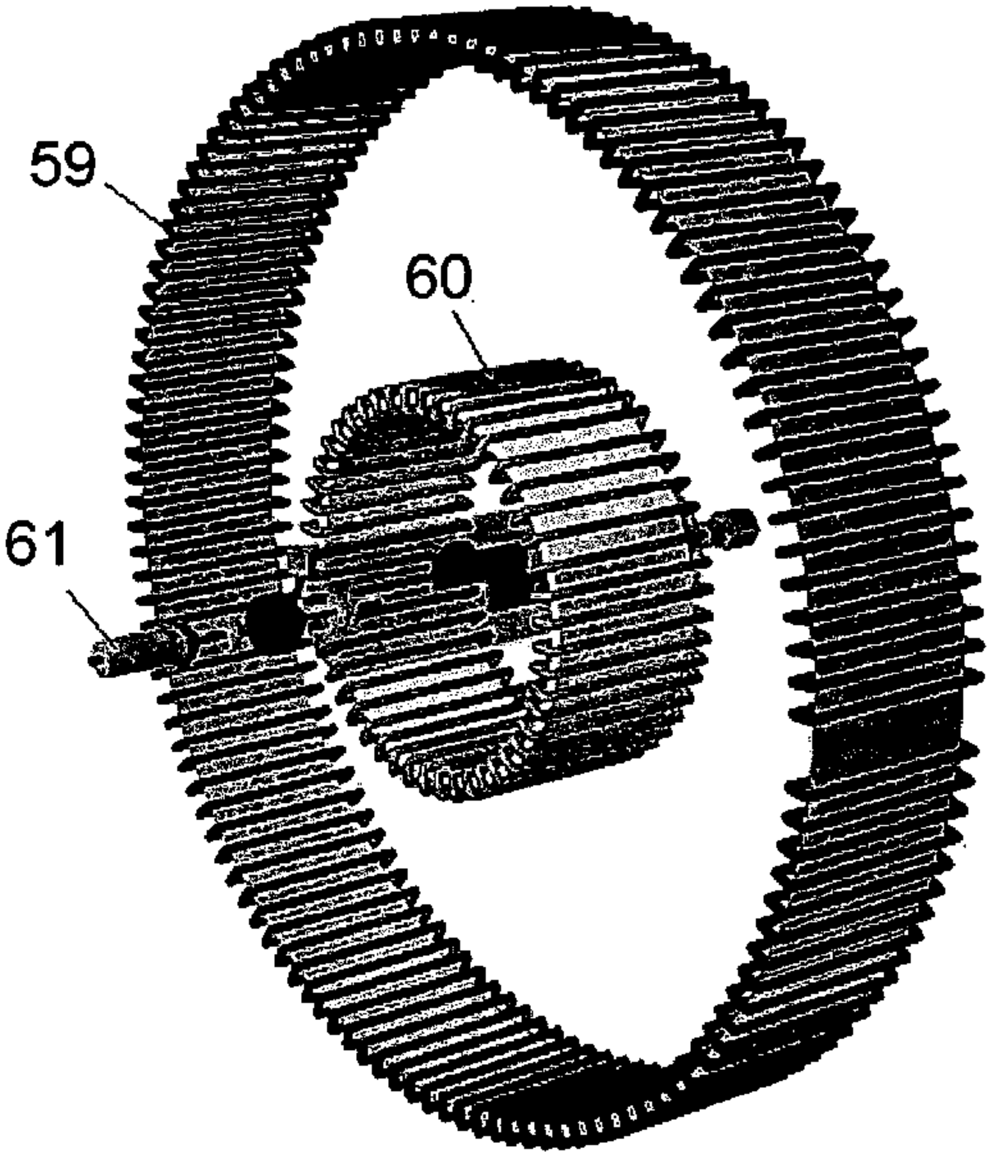

Figure 17

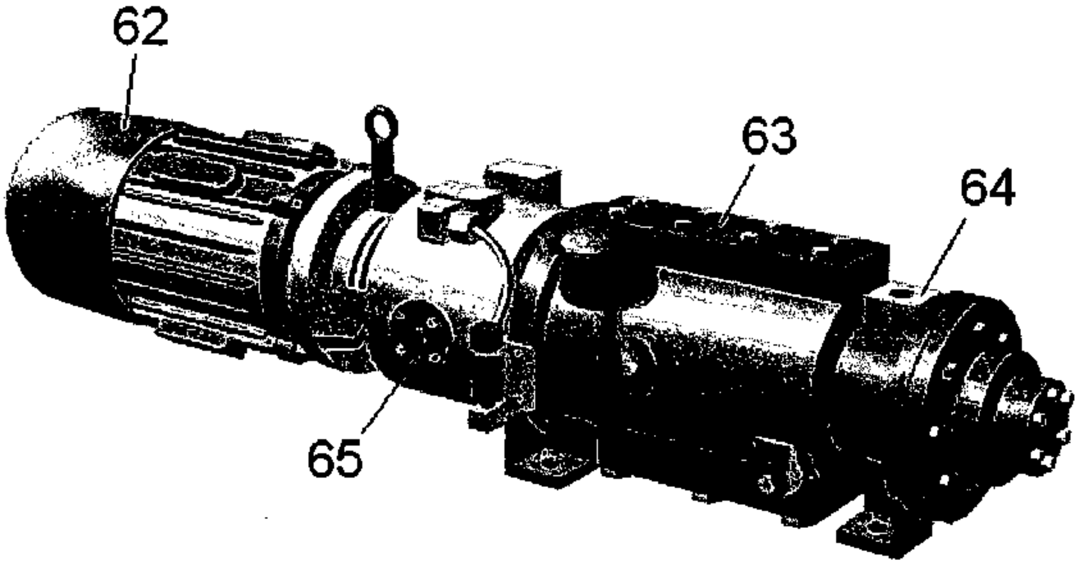

WATER ENERGY GENERATOR

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is an energy generator that uses the transition through wormholes of the hydrogen atoms of water molecules to break the bonds of the atoms and convert the protons into photons and electrons which can be collected for energy.

BACKGROUND OF THE INVENTION

[0002] A molecule of water consists of two hydrogen H atoms and one atom of oxygen O which has the chemical formula

H_2O

The hydrogen atoms can be separated from the oxygen atom by electrolysis. In this process, a direct current of electricity, such as from a battery, is passed through water decomposing it into hydrogen and oxygen. Pure water, however, is a poor conductor of electricity. It is therefore necessary to add some substance to form a solution that will conduct an electric current. Such a solution that will conduct electric current is called an electrolytic solution. A small amount of sulfuric acid or sodium hydroxide is added to the water to form an electrolytic solution. Water electrolyzed yields hydrogen plus oxygen

Water->hydrogen+oxygen

$2H_2O \rightarrow 2H_2 + O_2$

Because this invention is to be used on spacecraft, the oxygen can be used for breathing and the hydrogen can be used to produce energy that will be used to create the electromagnetic fields which provide lift and propulsion.

[0003] Referring to **FIG. 1**, a single hydrogen atom consists of one proton (1) in the nucleus and one electron (2) moving in an orbital (3) around the nucleus. In quantum physics notation, there is only one main K shell and one orbital containing a single electron in the 1 s subshell. An orbital is an energy level containing one or two electrons in a subshell of an atom. Only a total of two electrons may be placed in one orbital with the added constraint that the electrons spin in opposite directions. Looking at the 1 s orbital, it can be seen that only half of it is occupied. What this means is that it is possible to add another hydrogen atom in which its electron occupies the other position in the orbital to from the hydrogen molecule H_2.

[0004] In 1925, a physicist by the name of Edwin Schrodinger developed a wave equation, which bears his name, that models the hydrogen atom. Even though the classical picture of **FIG. 1** showing a distinct electron orbiting the proton is easy to visualize, in reality the electron is a wavefunction Ψ whose square indicates the probability of finding the electron at a particular point. This then was the start of quantum physics where it was found that the electron energy can only take on certain discrete values.

[0005] A traveling wave moving in the positive x-direction can be represented by the function $\Psi_1=\Psi_1(x,t)$ of the form

$$\Psi_1=A_1 \cos(2\pi(x/\lambda-v\ t))$$

where x is the distance along the x-axis, A the wave amplitude, λ the monochromatic wavelength, v the velocity of the wave and t time.

[0006] A similar monochromatic wave moving to the left can be represented by

$$\Psi_2=A_2 \cos(2\pi(x/\lambda+v\ t))$$

where the sign of the velocity is reversed. The superposition of these traveling waves results in a standing wave, or stationary waves, of the form

$$\Psi=\Psi_1+\Psi_2=A \cos(2\pi x/\lambda) \cos(2\pi v\ t)$$

which is now a product of a spatial-dependent term A cos(2πx/λ), and a time-dependent term cos(2πv t). Taking the partial derivative of **Ψ** twice with respect to x

$$\frac{\partial^2\Psi}{\partial x^2}=-\left(\frac{2\pi}{\lambda}\right)^2\Psi$$

The momentum p of a particle is equal to Planck's constant h divided by the mass m of the particle

$$p=h/\lambda=\frac{2\pi\hbar}{\lambda}$$

where h bar is the reduced Planck constant. Thus Schrodinger's equation can be written as

$$\frac{\partial^2\Psi}{\partial x^2}=-\frac{p^2}{\hbar^2}\Psi$$

For a particle whose potential energy is V(x), the total energy is the kinetic energy K plus the potential energy

$$E=K+V=(p^2/2m)+V$$

Therefore,

$$p^2=2\ m\ (E-V)$$

and Schrodinger's equation becomes

$$\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}+(E-V)\Psi=0$$

The potential V is just the Coulomb potential of the product of two charges e divided by the radius r between them

$$V=\frac{e^2}{4\pi\varepsilon r}$$

where ϵ is the linear capacitance of space. A general wavefunction can be separated into a radial R part and a spherical harmonics part Y

$$\Psi=R_{n,1}Y_{1,m}(\theta,\phi)$$

where the first term is called a radial wavefunction which describes the "in-out" motion of the electron. When Schrodinger's equation is separated, it is found that the radial part of the wavefunction, R, must be a solution of the quantized differential equation

$$\frac{\hbar^2}{2mr}\frac{d^2(rR)}{dr^2}+\left(E-\frac{l(l+1)\hbar^2}{2mr^2}+\frac{e^2}{4\pi\varepsilon r}\right)R=0$$

This can be expressed in a simpler form by defining the function

f=rR

which is then found to satisfy the equation

$$\frac{\hbar^2}{2m}\frac{d^2 f}{dr^2}+(E-V_{eff})f=0$$

where the effective potential is given by

$$V_{eff}=\frac{l(l+1)\hbar^2}{2mr^2}-\frac{e^2}{4\pi\varepsilon r}$$

where the 1's are the quantum orbital numbers. For s orbitals equal to 1=0, the effective potential is just the electrostatic potential of the nucleus. For 1 greater than zero, the first term is equal to the kinetic energy owing to the angular motion of the electron at a distance r and with angular momentum √1(1+1)h.

[0007] Referring to **FIG. 2**, the effective potential is graphed as a function of radius r and the orbital numbers. For orbital number 1=0, which is the lower of the three curves, the potential does not provide a stable position for the electron (black disk) and the negatively-charged electron just crashes into the positively-charged proton nucleus as shown by the arrow. For orbital 1=1, the first term is called the centrifugal repulsion which together with the electrostatic potential provides for a stable position for the electron as seen in the middle curve. The potential energy is negative which creates a slight valley in which the electron obtains a stable orbit. Higher orbital numbers 1=2 produce a similar valley potential further out on the radius. The 1=1 orbital does not allow the electron to come near the proton which provides for a stable hydrogen atom. So the key to this invention is how to destabilize this hydrogen atom energy system and produce usable energy which can propel the electromagnetic spacecraft and run other hyperspace inventions. Refer to my patent applications such as Dipole Moment Spacecraft, Dual Potential Hull Spacecraft, Photon Spacecraft, Electromagnetic Field Propulsion System, Full Body Teleportation, Magnetic Vortex Wormhole Generator, Electric Vortex Wormhole Generator, Sulfur S8 Wormhole Generator, Cavitating Oil Hyperspace Energy Generator, Rotor Inductance Propulsion System and Triangular Spacecraft.

[0008] Many of these patent applications involve wormholes and hyperspace which are not well-known concepts in the scientific community. Hyperspace consists of those dimensions which are co-dimensional with our spacetime. The reason I know about hyperspace is because (1) I have been in hyperspace on a number of occasions and have experienced Einstein's time dilation according to his General Theory of Relativity, (2) I have experienced more than one full-body hyperspace teleportation over a distance of 100 meters, (3) we have been able to create a wormhole between space and hyperspace with the magnetic vortex wormhole generator in which smoke was blown through one side of the coil into hyperspace, a first contact verified by the Grey Aliens, (4) I have seen the green mist associated with moving out of dimension and crossing over into hyperspace, (5) I have looked into another dimension and have seen another building, a car and a man who waving at me in the presence of an artificially-intelligent Cyborg with the "high-tech look" from the Pleiadian Defense Department, (6) I can remote view through hyperspace subgeometry to distances of 100,000 light years to the edge of the galaxy and have made first contact with around 500 extraterrestrial civilizations involving the use of patent applications such as Remote Viewing Amplifier, Quantum Dot Energy Cylinder and Walking Through Walls Training System, (7) I am the only person on Earth who has communicated with the designers of the crop circles found in England and explained their design to them in terms of subspace geometry, (8) I can walk through walls as a hyperspace energy being, and (9) I have discovered how anti-gravity is possible using low density hyperspace energy, not to mention all the other research work on these electromagnetic field propulsion vehicles. For my work in developing the geometry of the subspace manifold known as the tetrahedron diagram, I was given the Aphysics award by the Admiral, who is third in line to the Admiralty of the Pleiadian Defense Department. For my work in Revelations, she awarded me four beautiful galloping white riderless horses of the Apocalypse. The reason that the Pleiadian Defense Department was involved was that the Admiral had the task of creating the energy being that would protect the subspace manifold during the battle of Revelations which took place in the year 2001. So these are some of my personal experiences in the field of hyperspace physics.

[0009] As to the scientific aspect of hyperspace physics, observations of the solar system have noted that large vortices occur on the planets at a latitude of ±19.5°. On the planet Jupiter, for example, the Giant Red Spot vortex, which is the size of two planets like Earth, is located at −19.5° latitude.

[0010] Referring to **FIG. 3**, the Olympus Mons volcano, which is the size of France, occurs in the northern hemisphere at 19.5°. The plume of volcanic ash can be seen being blown to the upper right.

[0011] Here on Earth, just north of me at +19.5° in the Caribbean, there is a slow moving rock mantle vortex that curves the islands down toward Venezuela. Since the tetrahedron has three corners, another corner occurs at 120° to the west where the Marshall Islands are located. The reason that the volcanic islands form is that low density hyperspace energy softens the atoms and molecules to such an extent that the atomic bonds are broken. The hot magma from the core has an easier path to the surface through the soft rock compared to the hard rock of surrounding areas. As the vortex rotates, the rising hot magma creates volcanoes which form the chain of islands.

[0012] As another example of this, the Silver bridge, which is located at a double harmonic of 39° between Ohio and West Virginia, collapsed because a large wormhole opened up and flooded the bridge with low density hyperspace energy. The metal rivets softened and pulled out of the plates holding down the suspension cables. Thus the roadway tilted to one side and all the cars fell into the river.

[0013] If you look at the quarried granite polygonal megalithic stone blocks that were used to build Machu Picchu, the blocks are put together as though they were soft putty. They fit together perfectly. Machu Picchu, or Great Picchu, is the Quichua name for a sharp peak which rises ten thousand feet above the sea. How this was achieved is described in my patent application *Rotating Magnetic Vortex Generator* which shows that rotating permanent magnets can create a wormhole between space and hyperspace. The low density hyperspace energy then floods the block as to make it weightless. The block is then floated up from the quarry which is 2500 feet below Machu Picchu. The huge 1000-ton blocks found in the platform at Baalbek in the Cedar Mountains of Lebanon were also lifted into position in this manner.

[0014] Using a technique called Chi Kung breathing, not necessarily known to the Chinese, one of our researchers has been able to levitate himself to a height of six feet. He might have gone higher but he lost his balance and his arm touched a nearby metal pipe. He then floated back to the ground. Contracting the abdominal muscles causes hot air to rise in the lungs while at the same time cold air is inhaled through the nose. This combination of hot and cold air mixing together creates environmental oscillators having a temperature fluctuation. This changes Planck's constant such that the body goes out of dimension and low density hyperspace energy fills the body.

[0015] The final example comes from Biblical times. Jesus and his father Joseph were both carpenters. A chair that Joseph had made for a client was found to be too small. So Jesus placed his hands on either side of the chair and stretched the chair to fit. The phrase in Spanish is "mas silla" which means that Jesus made more chair. In the dictionary, the Spanish word for putty is "masilla." The reason that Jesus could do this is that he passed hyperspace energy through his hand vortices which softened the wood. See my patent application called *Hyperspace Torque Generator* which shows this in more detail. The hand vortex is actually a planar co-gravitational K field that can rotate a pendulum around in circles. This solenoidal field corresponds to the magnetic H field. The electric field corresponds to the linear gravitational g field. The equations are identical for both sets of fields. I learned that Jesus actually reincarnated here on Earth in order to identify the planet on which the battle of Revelations would take place.

[0016] In summary, these are some examples as to how low density hyperspace energy can soften and break the bonds of atoms and molecules.

[0017] As mentioned previously, the astronomical observations suggest that the structure of the universe is related to the tetrahedron. Referring to **FIG. 4**, the corners of a tetrahedron circumscribed by a sphere, touch the sphere at an angle of 19.47122063°. The ratio of the area-to-volume ratio of the sphere to the area-to-volume ratio of the tetrahedron is ⅓ which is the magic ratio in physics. The arcsin(⅓) is 19.47°.

[0018] In order for all the physics constants to be projected into our universe and the co-dimensional hyperspace dimensions, there has to be a tetrahedral subspace manifold. The only mathematical function which allows multiple rotations of the complex plane is the logarithmic function Log[z].

[0019] Referring to **FIG. 5**, the complex number z in the complex plane is equal to x+I y where x and y are real constants and $I^2=-1$. The complex number z can also be written in terms of a radius and an angle

$$z=r\ \mathrm{Exp}[I\ \theta]$$

Because the angle can rotate more than once in 2π m multiples, the complex number z is actually

$$z=r\ \mathrm{Exp}[I(\theta+2\pi\ m\ )]$$

Taking the logarithm of z

$$\mathrm{Log}[z]=\mathrm{Log}[r]+I(\theta+2\pi m)$$

[0020] Referring to **FIG. 6**, it can be seen that it is not possible to go around multiple times on the same surface due to a branch cut along the origin where Log[0] is undefined. What happens is that as you rotate around on surface (4) and get to the origin, you go down along the branch (5) to the next hyperspace plane (6). Thus the universe is composed of many hyperspace co-dimensions. From personal experience, I estimate that the distance separating the two dimensions is about 3 meters when I was looking into another dimension at the man waving at me. The physics is even more complicated because there is a Lorentz dimensional transformation in which higher dimensions appear smaller and lower dimensions appear larger. In one case I was looking at a huge mothership at a much higher dimension and it looked like a tiny toy model spacecraft. They fired a laser cannon at me, and I then curved space which made the beam change course. You can see why I was awarded the four horses.

[0021] The antilog of Log[z] is

$$e^{\mathrm{Log}[z]}=z=re^{I\ \omega t}$$

where the angle is the angular frequency ω times the time t. The subspace geometry remains stationary or fixed, but the projection of the antilog into our dimension generates a system which is frequency dependent. This is why we get oscillating fields.

[0022] From Einstein's General Theory of Relativity, it is known that various kinds of energy can curve spacetime such as mass, electromagnetic fields, angular momentum and electrical charge. The elemental spacetime length ds in cylindrical coordinates {t,r,θ,z}, known as the Schwarzschild metric, shows that spacetime can be curved using mass M and charge Q

$$(ds)^2=-(dt)^2\left(1-\frac{2M}{r}+\frac{Q^2}{r^2}\right)+\frac{dr^2}{\left(1-\frac{2M}{r}+\frac{Q^2}{r^2}\right)}+(rd\theta)^2+(dz)^2$$

where you can see that the elemental time dt is dilated by the mass M and the elementary radius dr is reduced. I have actually experienced this time dilation in which, as I was jumping into hyperspace, I was hit by a car which broke my shoulder. When I came back into dimension which appeared to be a few seconds later, I found that I had an 8-inch blood ring down the right side of my chest. My shoulder blade, which sticks up about two inches, is still broken to this day.

[0023] Referring to **FIG. 7**, flat space (**7**) can be curved (**9**) by a massive body (**8**) such as a planet. For example, the mass of the Earth gives space a negative curvature such that objects tend to fall toward the center of the mass (bowl). On the other hand, using electromagnetic fields, it is possible to produce a positive curvature such that a spacecraft rises by falling upward.

[0024] Referring to **FIG. 8**, if the fields are intense enough, then a wormhole (**12**) forms between space (**10**) and hyperspace (**11**). This depiction is called an embedding diagram because there is no open space going through the wormhole. An object traversing the wormhole moves along the surface from one dimension to another. In order to open the throat (**12**) of the wormhole, negative energy is required. Referring to my patent application Dual Potential Hull Spacecraft, negative energy is produced by the interaction of a microwave beam with an oscillating magnetic H field. In terms of differential forms mathematics, this is given as the Hodge Star * of the differential d of the wedge ^ product of the magnetic H field with the electromagnetic wave {B,E}

$$*d(H\wedge(B+E\wedge dt))=\frac{\partial(-\rho)}{\partial t}$$

which says that there is an increasing rate of change of negative energy (−ρ). Due the presence of negative energy together with the spacetime curvature pressure produced by the electromagnetic fields, wormholes open up between space and hyperspace. There is a positive gravitational potential between hyperspace and space because the low density hyperspace energy is more dense than the negative energy in this dimension. Thus the hyperspace energy flows into our dimension which reduces the mass of the spacecraft. The upward spacetime pressure stress over the hull due to the electromagnetic fields creates a lift force on the nearly massless vehicle. Because the hyperspace energy has a speed of light equal to one meter per second, the electromagnetic fields become relativistically strong since they obey the Lorentz transformation. Due to the low mass, high spacetime lift pressure and strong EM fields, the vehicle can attain very high rates of acceleration.

[0025] Another method to produce a wormhole is to use bucking magnetic fields which is described in my patent application *Magnetic Vortex Wormhole Generator*. In gravitation physics, the Faraday F tensor, which is a 4×4 spacetime metric {t,x,y,z}, contains all the components of the electromagnetic fields in the various spatial directions {x,y,z}

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} t & 0 & E_x & E_y & E_z \\ x & E_x & 0 & B_z & -B_y \\ y & E_y & -B_z & 0 & B_x \\ z & E_z & B_y & -B_x & 0 \end{array}$$

where contravariant index a refers to the rows, and the covariant index β refers to the columns. For example, the component $F^t_x = E_x$ is the electric field in the x-direction. If there were two magnetic bucking fields in the x-direction the Faraday tensor would be

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} t & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & 0 \\ y & 0 & 0 & 0 & B_x-B_x \\ z & 0 & 0 & -(B_x-B_x) & 0 \end{array}=\begin{vmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{vmatrix}$$

which says that the Faraday tensor is zero. Thus no spacetime curvature is generated from two bucking magnetic fields at the same position. On the other hand, if the two bucking magnetic fields are concentric at different radii, then the Faraday tensor becomes

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} t & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & 0 \\ y & 0 & 0 & 0 & B_x\delta(x_1)-B_x\delta(x_2) \\ z & 0 & 0 & -(B_x\delta(x_1)-B_x\delta(x_2)) & 0 \end{array}$$

which is not zero due to the presence of the Kronecker δ delta function which locates the fields at different positions. The spacetime stress-energy-momentum tensor T is then computed from the following equation

$$4\pi T^{\mu\nu}=F^{\mu\alpha}F^{\nu}_{\alpha}-\frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

where g is the metric tensor containing the coefficients of the elemental spacetime length ds. With the mass M and charge Q term equal to zero, there being just electromagnetic fields involved, the g metric tensor in cylindrical coordinates becomes

$$g_{\alpha\beta}=\begin{array}{c|cccc|} t & -\left(1-\frac{2M}{r}+\frac{Q^2}{r^2}\right) & 0 & 0 & 0 \\ r & 0 & \left(1-\frac{2M}{r}+\frac{Q^2}{r^2}\right)^{-1} & 0 & 0 \\ \theta & 0 & 0 & r^2 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

$$=\begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

[0026] Referring to **FIG. 9**, the magnetic vortex generator has two concentric, thin, flat cylindrical silicon-iron cores (**13**,**14**), each consisting of a stack of three 0.020 inch thick transformer laminations wrapped with insulating tape. Using insulated magnetic wire, a flat helical coil (**16**) is wrapped counter-clockwise around the outer laminations. The coil receives power through connection (**15**). The winding then continues around the core until coil (**17**) where the wire (**18**) is extended to the inner core and the wrapping of coil (**19**) proceeds in the clockwise direction. The inner core wrapping terminates on coil (**20**) with the extension of the second power connection (**21**). Because the coils are wound in opposite directions, the generator produces two bucking magnetic fields at different radii (**22**,**23**).

[0027] According to Maxwell's electromagnetic equations, the curl of the magnetic flux density B field times the square of the speed of light squared is equal to the rate of change of the electric E field

$$c^2\nabla xB = \frac{\partial E}{\partial t}$$

Multiplying both sides by the elemental area of the core and integrating over the area

$$c^2\int_0^{r_0}\nabla xBdA = \frac{\partial}{\partial t}\int_0^{r_0}E\cdot ndA = \frac{\partial}{\partial t}\int_0^{r_0}E\cdot n2\pi rdr$$

The curl of the field can be converted into a line integral around the core contour using Stokes' Theorem

$$c^2\int_0^{r_0}\nabla xBdA = c^2\oint_{r_0}B\cdot ds = c^2\int_0^{2\pi}Br_0d\theta$$

[0028] The magnetic flux density B field oscillates with angular frequency ω

$$B=B_0e^{I\omega t}$$

Substituting this into the equation and integrating over time for the inner core field

$$-c^2\int_0^t\int_0^{2\pi}B_0e^{I\omega t}r_0d\theta = \int_0^{r_0}E2\pi rdr = E\pi r_0^2$$

This can be solved for the electric E field that is produced by the oscillating magnetic flux density B field

$$E_0 = \frac{2IB_0c^2(-1+e^{I\omega t})}{r_0\omega}$$

The electric field for the outer core is the same equation except that the opposite sign of the magnetic flux density B field and the radius r1

$$E_1 = \frac{-2IB_0c^2(-1+e^{I\omega t})}{r_1\omega}$$

Adding these two fields together is the total electric field E

$$E = E_0 + E_1 = \frac{-2B_0c^2(-1+\mathrm{Exp}(I\omega t))(r_0-r_1)}{r_0r_1\omega}$$

The electrostatic energy of the field is equal to half the linear capacitance of space times the summation over the volume of the dot product of the electric field with itself

$$U = \frac{\varepsilon}{2}\int E\cdot EdV$$

Because the electric field points in the z-direction out of the coil, the dot product is actually the square of the electric field.

[0029] Referring to **FIG. 10**, the energy U per volume is plotted as a function of time with a radius ratio r_1/r_0 of 3/1. As the graph shows, the magnetic vortex wormhole generator produces mostly negative energy which is required in order to create the wormholes. Because the stress-energy-momentum T tensor is also the square of the electric field, this graph gives the spacetime curvature pressure. Thus the electric field produces both the pressure and negative energy required to open up wormholes between space and hyperspace.

[0030] Going back to the effective potential equation,

$$V_{eff} = \frac{l(l+1)\hbar^2}{2mr^2} - \frac{e^2}{4\pi\varepsilon r}$$

it can be seen that the first term is divided by the mass of the particle. In current gravitation physics, the mass of the particle is invariant with velocity. It does not obey the Lorentz transformation. The mass is related to the energy E of the particle and its momentum p by

$$m^2=E^2-p^2$$

In different inertial frames moving with a relative velocity v, the energy and the momentum obey the Lorentz transformation, but no matter what the relative motion, the mass of the particle is constant. In the first term of the effective potential, the mass is constant which leaves just Planck's constant.

[0031] Having worked for over ten years on the subspace manifold, known as the tetrahedron diagram, I found a most incredible intersection on the diagram when working with the water molecule. The water molecule has two hydrogen atoms and one oxygen atom as mentioned previously. The atomic weight of one atom of hydrogen is 1.008 atomic weight units (awu). The atomic weight of oxygen is 16.000 awu. Therefore the molecular weight of water is

$$\text{Weight of two atoms of hydrogen } 2\times 1.008\ awu = 2.016\ awu$$

$$\text{Weight of one atom of oxygen } 1\times 16.000\ awu = \frac{16.000\ awu}{18.016\ awu}$$

The gram molecular weight is the atomic weight expressed in grams, so there are 18.016 grams in Avogadro's number of molecules. So the mass per molecule in logs is

$$\mathrm{Log}\left[\frac{18.016\ \mathrm{gram-mol}\Big/\left(1000\ \frac{\mathrm{gm}}{\mathrm{kg}}\right)}{6.02\times 10^{23}\ \mathrm{mol}}\right]=-58.77103943$$

[0032] Referring to **FIG. 11**, the tetrahedron diagram plots the natural logarithm of mass versus the natural logarithm of wavelength. The reason for this is that mass times wavelength is equal to Planck's constant divided by the speed of light c, known on the diagram as the base constant. At the present time there are over 4000 diagrams which are copyrighted in the Library of Congress. In logarithms, the product of two numbers is the sum of the two numbers. This means that the sum of the mass and the wavelength are equal to the base constant which has a value of

$$\ln[m]+\ln[\lambda]=\ln[h/c]=-95.91546344=\text{base constant}$$

Our dimension has a lower limit on mass and length known respectively as the Planck mass and the Planck wavelength. The Planck mass is the linear mass Ω of the universe times the Planck scale Λ. The Planck wavelength is circumference of a circle of radius Planck scale. In terms of logs, the Planck mass and Planck wavelength are

$$\text{Planck mass}=\ln(\Omega\Lambda)=-17.64290101$$

$$\text{Planck wavelength}=\ln(2\pi\Lambda)=-78.27256243$$

[0033] When these values are plotted on the tetrahedron diagram shown in **FIG. 11**, the Planck box (abcd) is formed which are the boundaries of our dimension in subspace. The line numbering is as follows

Planck mass	25, 27
Planck wavelength	26, 28
mass of water molecule	29, 31
speed of light squared circle	30
base constant	33
inverted tetrahedrons	34, 35
centerline	36

The energy of the water molecule, circle (**37**), is equal to the sum of the water molecule mass (**29**) plus the speed of light squared circle (**30**). The energy circle (**37**) intersects the mass of the water molecule (**29**) at the Planck wavelength (**28**, point e), which is the boundary between space and hyperspace. What this means is that the mass is equal to the energy at the Planck box boundary. The only way that this is possible is if the speed of light c is equal to one meter per second

$$E=m\ c^2=m\ c=1\ \text{meter/second}$$

A water molecule traversing a wormhole into hyperspace undergoes a change in the speed of light from 299792458 m/s to 1 m/s.

[0034] Planck's constant h is equal to the Planck mass ΩΛ times the Planck scale Λ times the speed of light c.

$$h=\Omega\Lambda\Lambda c$$

By having the speed of light go to 1 m/s, the orbital term in the effective potential V_{eff} is reduced by a factor of the speed of light squared equal to 9×10^{16}. This unbalances the equation to such an extent that only the Coulomb potential term remains.

$$\frac{l(l+1)\hbar^2}{2mr^2}\ll\text{coulomb term}$$

The electron is attracted to the proton nucleus because the centrifugal term no longer provides a stable orbit for the electron. Thus the atomic binding is destroyed and the water molecule becomes soft as putty.

[0035] Referring to **FIG. 12**, the collision of the electron with the proton together with the enormous change in the proton's energy causes the proton p to become unstable and decay. According to the Standard Model of particle physics, the elementary particles are composed of smaller particles known as quarks. The six quarks have been named up u, down d, strange s, charm c, top t, and bottom b. The subscript on the quark indicates one of three colors {red r, blue b, green g). As shown in the diagram, the proton is composed of three quarks $\{u_r, u_g, d_b\}$, two of which are up quarks of which one is red and the other green, and a third blue down quark. The proton p decays into a positron e^+ which is an electron with a positive charge, and a neutrally-charged pion π^0 particle through the exchange of an X boson particle. The pion has a mass between the electron and the proton.

$$p\rightarrow\pi^0+e^+$$

[0036] Referring to **FIG. 13**, the pion π^0 then decays into a proton p and antiproton $\bar{p}$ which annihilate each other to produce two photons shown on the right by the traveling waves. So the overall energy exchange is

$$p\rightarrow\frac{2h\nu}{c^2}+e^+$$

where hv is the energy per gamma photon with frequency v. The electron of the hydrogen atom would then annihilate the positron for additional photon energy.

[0037] Referring to **FIG. 14**, the hydrogen H atom is composed (**38**) of the proton and electron as seen in the upper left corner. The proton decays (**39**) into the neutral pion and a positron. The electron from the hydrogen atom and this positron form one electron pair (**40**). The pion then decays (**41**) into two gamma photons which produce an electron pair production energy cascade into 132 pairs (**42,43**) for a total of 133 electron pairs. These electrons can then be captured electrostatically and used for the production of electricity.

[0038] Referring to **FIG. 15**, the electrons are captured with the water droplet injector . The plunger (**46**) of a spring-loaded cylindrical solenoid (**44**) is attached to a tapered piston (**47**). By means of ring collar and bolts (**45**), the solenoid is bolted to the injector (**48**). A supply of purified water is attached to the water inlet connection (**49**). When the solenoid is activated, it pulls back slightly so that water can enter the valve. When the solenoid is deactivated, the piston forces the water droplet out through the nozzle (**50**) into a cylindrical glass vacuum chamber (**52**). Two cylindrical glass disks (**51**) hold the nozzle in place. On the other end of the vacuum chamber is the sealed-tube connection (**56**,**57**,**58**) to the vacuum pump. In the middle of the vacuum chamber, two metal plates (**53**) are attached through sealed glass collars (**54**) to electrical pins (**55**). The plates are electrostatically-charged with opposite charges so as to form a capacitor. This creates an electrical field between the plates which attracts the electrons to the positively-charged plate.

[0039] Referring to **FIG. 16**, the vacuum tube and water injector (**61**) are mounted along the centerline of the inner (**60**) and outer (**59**) magnetic vortex wormhole generator coils. The low density hyperspace energy traversing the wormhole along the centerline of the coils causes the injected water molecules to soften and decay into a cascade of electrons. The oscillating electric field along the centerline causes the electrons to vibrate back and forth. The crossed electric field between the charged capacitor plates causes the electrons to flow toward the positively-charge plate in order to produce electricity.

[0040] Referring to **FIG. 17**, the vacuum tube is connected to the vacuum pump through a hose connection to the pump air inlet (**64**). A 5 Hp electric motor (**62**) drives dual rotating flights of screws which trap the air and move it toward the exhaust outlet (**65**) shown with no muffler. On a spacecraft operating in the vacuum of outer space, this component would not be needed.

SUMMARY OF THE INVENTION

[0041] It is the object of this invention to generate electricity by using low density hyperspace energy to soften water molecules such that the atomic binding is broken which causes the hydrogen nucleus to decay into a cascade of electron pairs. These electrons are then collected on a positively-charged plate in order to produce electricity. The water molecules are softened by flooding them with low density hyperspace energy that is produced by a magnetic vortex wormhole generator. The generator creates negative energy and a spacetime curvature along the centerline of two concentric coils. This combination opens up wormholes along the centerline. Because the gravitational potential of low density hyperspace energy is greater than the negative energy, the hyperspace energy flows through the wormhole from hyperspace into our dimension. The hyperspace energy has a speed of light equal to one meter/second. This causes a change in Planck's constant h such that the proton orbitals of the hydrogen atom are unable to produce a centrifugal repulsion which keeps the electron in orbit. The Coulomb potential term dominates and the electron is attracted to the proton. Due to the vast change in the speed of light, and the collision of the electron with the proton, the proton becomes unstable and decays into a neutral pion and a positron. The pion then decays into two gamma photons which produce a large cascade of electron pairs.

[0042] A water injector, consisting of a solenoid-activated valve and nozzle, injects water droplets into a vacuum chamber which is positioned along the centerline of the two concentric coils where the wormholes form. Due to the low density hyperspace energy passing through the wormholes into our dimension, the water molecules soften and decay into electrons which are collected on an electrostatically-charged capacitor plate having a positive charge located in the glass vacuum chamber.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0043] **FIG. 1**. Perspective view of hydrogen atom K shell.

[0044] **FIG. 2**. Graph showing potential binding energy of hydrogen atom.

[0045] **FIG. 3**. Perceptive view of Olympus Mons volcano at +19.5° Mars latitude.

[0046] **FIG. 4**. Perspective view of tetrahedron inscribed in sphere.

[0047] **FIG. 5**. Graph showing complex plane.

[0048] **FIG. 6**. Perspective view of hyperspace co-dimensions of logarithmic manifold.

[0049] **FIG. 7**. Perspective view of embedding diagram showing curvature of space caused by a mass.

[0050] **FIG. 8**. Perspective view of wormhole embedding diagram.

[0051] **FIG. 9**. Perspective view of coils of magnetic vortex wormhole generator.

[0052] **FIG. 10**. Graph showing that generator produces negative energy.

[0053] **FIG. 11**. Tetrahedron diagram showing that the speed of light at the Planck box boundary at the water molecule is one meter/second.

[0054] **FIG. 12**. Perspective view of proton decay into neutral pion and positron.

[0055] **FIG. 13**. Perspective view of pion decaying into two gamma photons.

[0056] **FIG. 14**. Diagram showing decay of the hydrogen atom into electrons.

[0057] **FIG. 15**. Perspective view of water injector and vacuum chamber.

[0058] **FIG. 16**. Perspective view of vacuum chamber mounted along centerline of magnetic vortex wormhole generator.

[0059] **FIG. 17**. Perspective view of vacuum pump used to evacuate vacuum chamber.

DETAILED DESCRIPTION OF THE INVENTION

[0060] 1. The coils of the magnetic vortex wormhole generator are made of three stacks of 0.020 inch silicon-iron transformer laminates. These are washed to remove the oil, and then wrapped with insulating tape in order to keep the laminations together. Using a very long bench made of wooden planks and 2×4 sawhorses, the outer coil is wrapped counter-clockwise right to left using a large spool of 14 AWG magnetic wire. A thin spacer is used between windings in order to reduce the winding capacitance. Once the outer coil is wound, the wire is continued to the second inner coil which is wrapped clockwise, leaving enough wire between coils such that when the coils are mounted in the wooden frame, the coil is one continuous winding having an input and output connection. Using an inductance meter, the inductance of the coil is measured. Using a standard frequency of 60 Hz, the capacitance of a sheet metal capacitor is calculated such that the generator is electromagnetically resonant at this frequency. The generator is connected to the line voltage by a 1:1 isolation transformer which is connected to a small primary coil wrapped on a toroidal core whose similar secondary coil is connected to the sheet metal capacitor and inductance coil. Resonance is achieved by adjusting the spacing and overlap of the sheet metal.

[0061] 2. The vacuum chamber is made of a glass tube with sufficient wall thickness to withstand the vacuum pressure. A number of glass blowing techniques are used to make the glass-electrode connection for the capacitor plates. Then circular pieces of glass plate are cut out and ground to the inside diameter of the tube, fitted with the nozzle and vacuum connection, and then heat sealed to the chamber. The chamber and water injector are then attached to a wooden bracket mounting which is doweled and glued to the wooden frame of the generator.

I claim:

1. A water energy generator system comprising the components:

a magnetic vortex wormhole generator and driving resonant electrical circuit;

a water droplet injector;

a vacuum chamber and vacuum pump; and

an electrostatic electron capture system.

2. By means of claim (**1**), a magnetic vortex wormhole generator comprising two concentric cylindrical coils of different radii wound in opposite directions, made of thin sheet silicon-iron transformer laminations wound with one continuous length of magnetic wire providing a single input and single output connection to the driving electrical circuit.

3. By means of claim (**2**), a coil winding method and oscillating driving circuit producing bucking electric fields along the centerline of the generator normal to the coils which create a spacetime curvature pressure and negative energy.

4. By means of claim (**3**), the generation of wormholes between space and hyperspace along the centerline of the generator such that low density, low speed of light hyperspace energy flows through a positive gravitational gradient from hyperspace to space.

5. By means of claim (**1**), a water injector comprising a solenoid-activated water valve, water supply connection, seal and nozzle for injecting water droplets into the vacuum chamber.

6. By means of claims (**5**) and (**4**), the softening and particle decay of the water molecules by the hyperspace energy into neutral pions, positrons, gamma photons and finally a cascade of electron pairs.

7. By means of claims (**6**) and (**1**), the capture of the electrons on electrostatically charged capacitor plates located in the vacuum chamber for the purpose of producing electrical energy.

8. By means of claims (**1**) and (**2**), a resonant electrical driving circuit comprising a line isolation transformer connected to a primary coil wound on a toroidal coil whose secondary output coil is connected to a capacitor and the inductance coil of the generator such that the capacitance of the capacitor and the inductance of the coil form a highly resonant electrical circuit.

* * * * *

US 20060168937A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2006/0168937 A1**

(43) **Pub. Date: Aug. 3, 2006**

(54) MAGNETIC MONOPOLE SPACECRAFT

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/048,026**

(22) Filed: **Jan. 31, 2005**

Publication Classification

(51) **Int. Cl.**
F03H 1/00 (2006.01)

(52) **U.S. Cl.** **60/203.1**

(57) ABSTRACT

A spacecraft propulsion system that utilizes a dual method of providing lift on the hull by means of magnetic monopoles and electromagnetic spacetime curvature pressure.

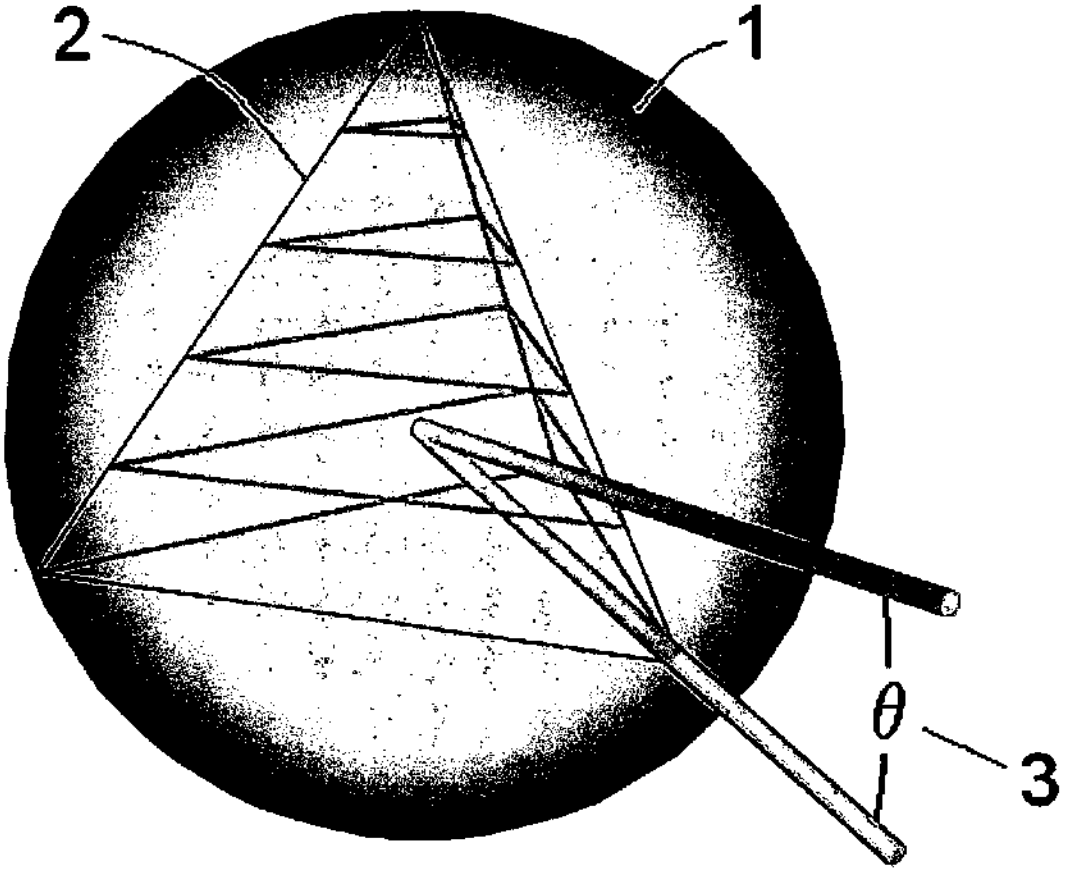

Figure 1

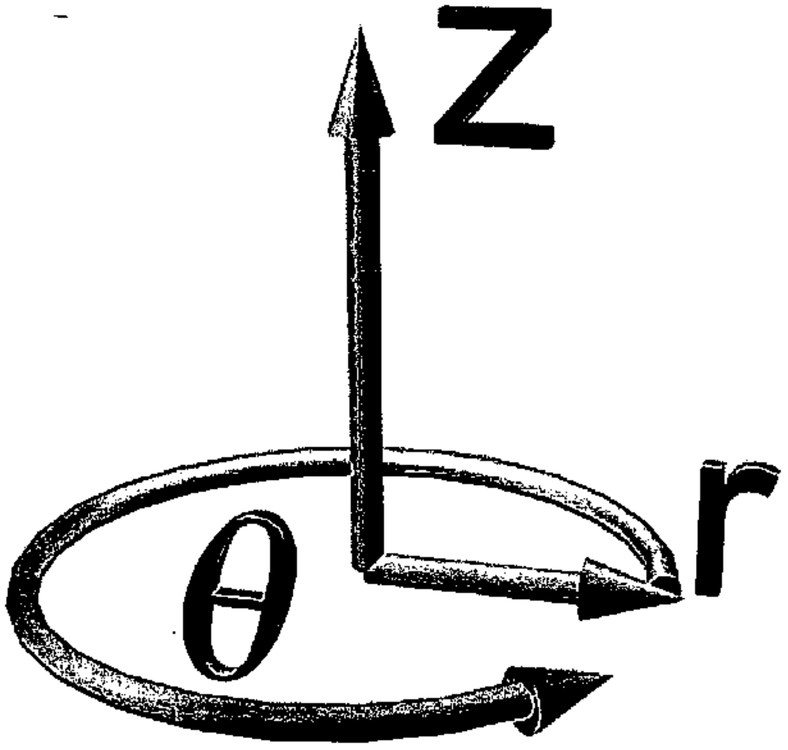

Figure 2

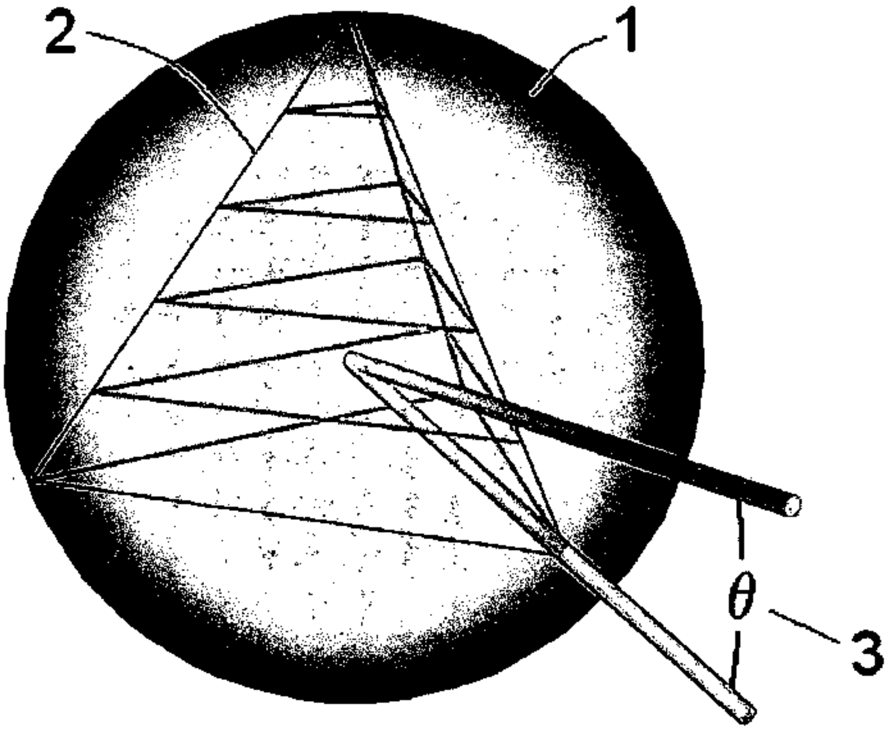

Figure 3

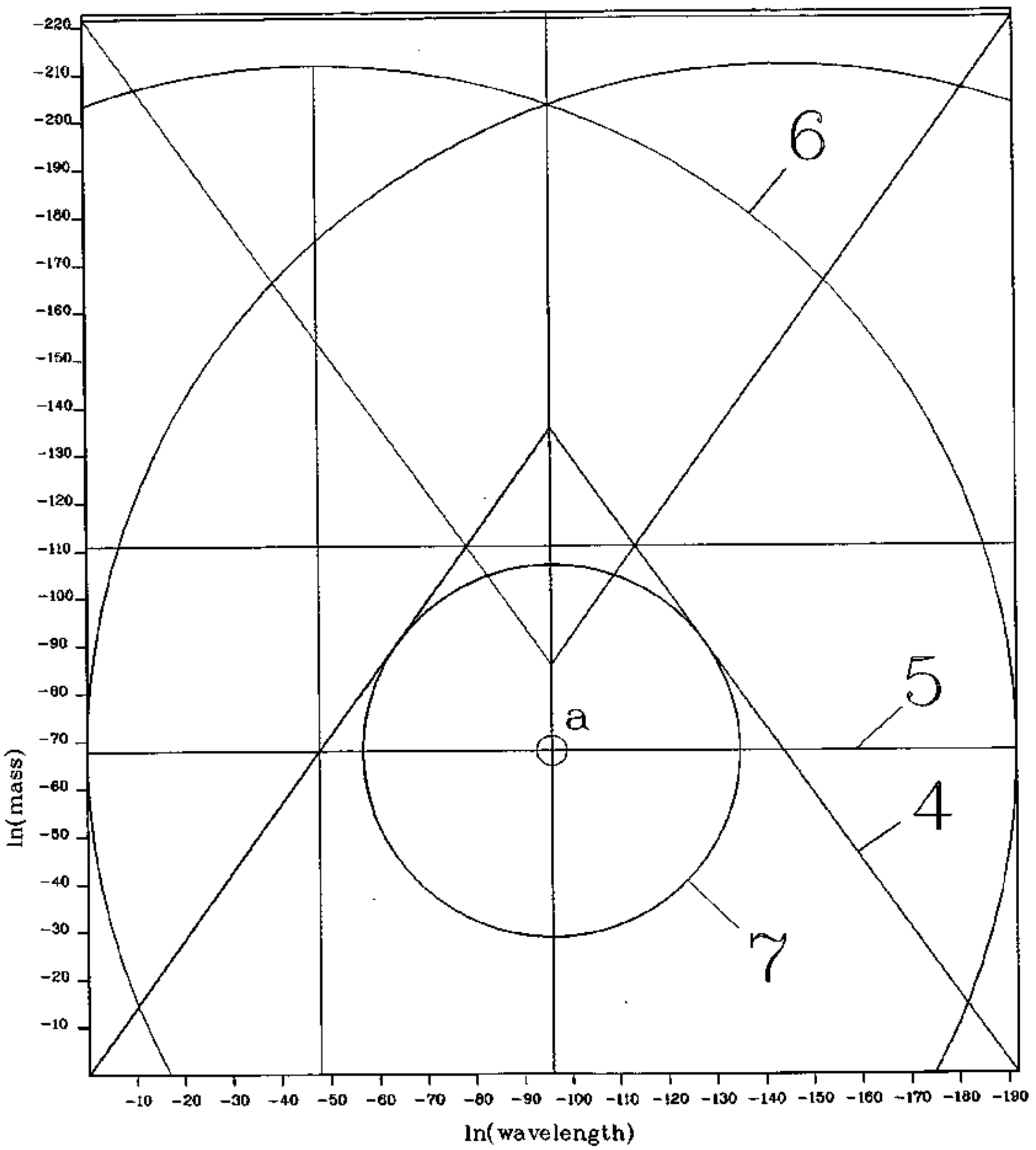
6
5
4
7
a
ln(mass)
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(wavelength)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190

Figure 4

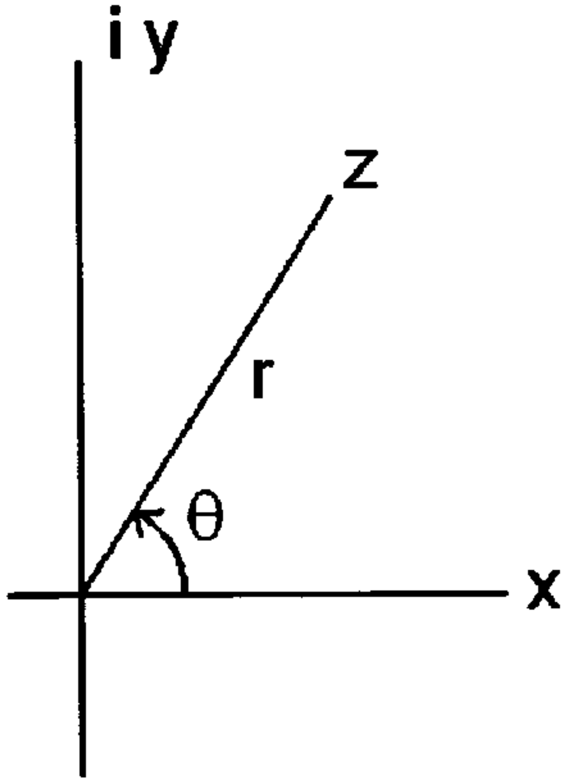

Figure 5

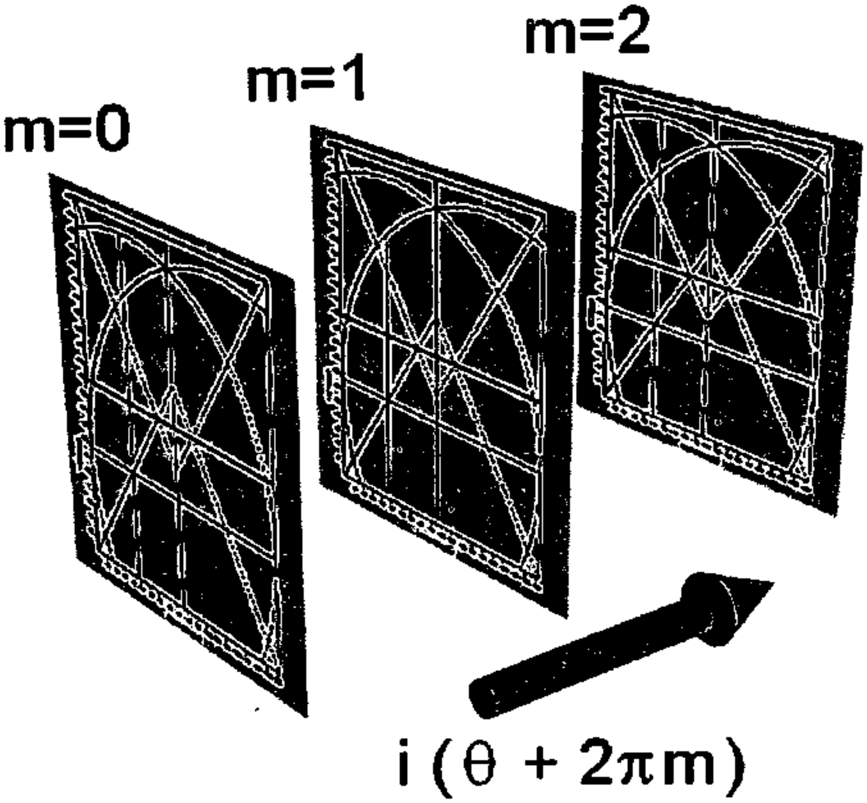
m=0
m=1
m=2
i (θ + 2πm)

Figure 6

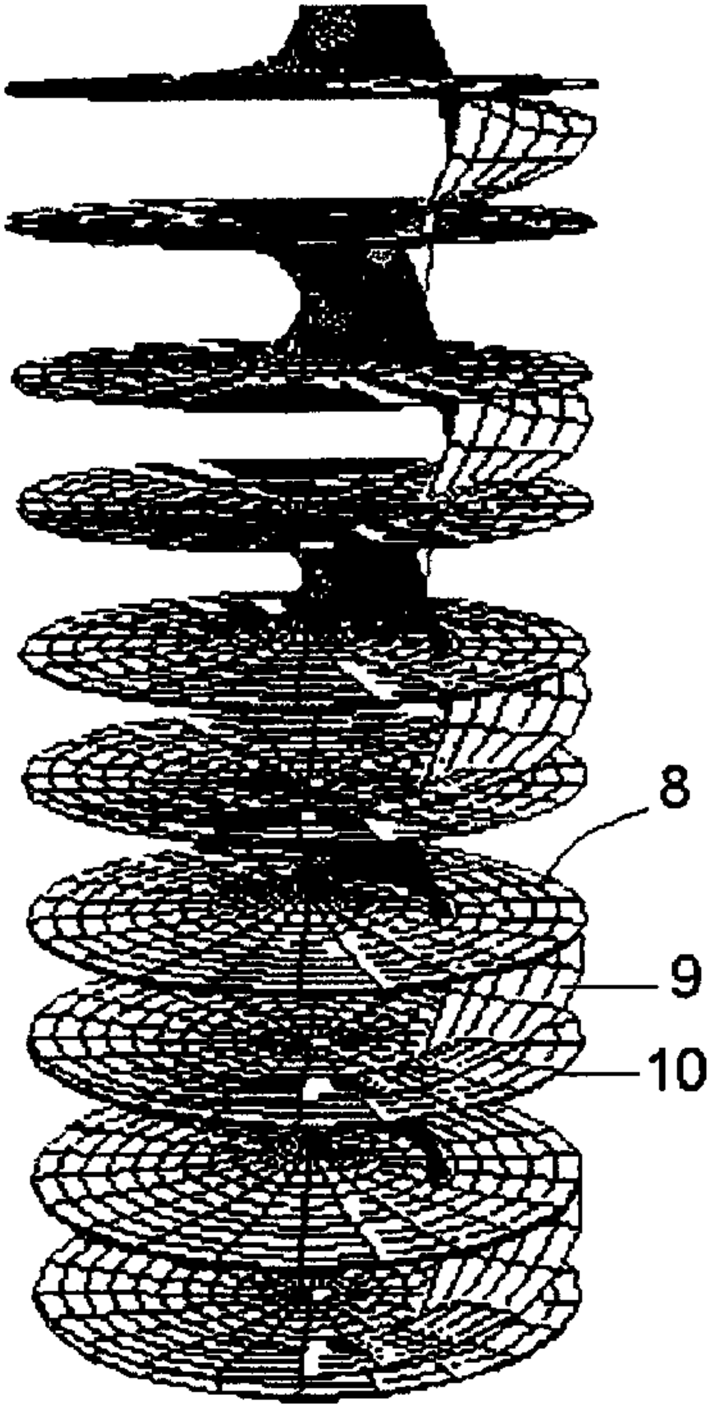

Figure 7

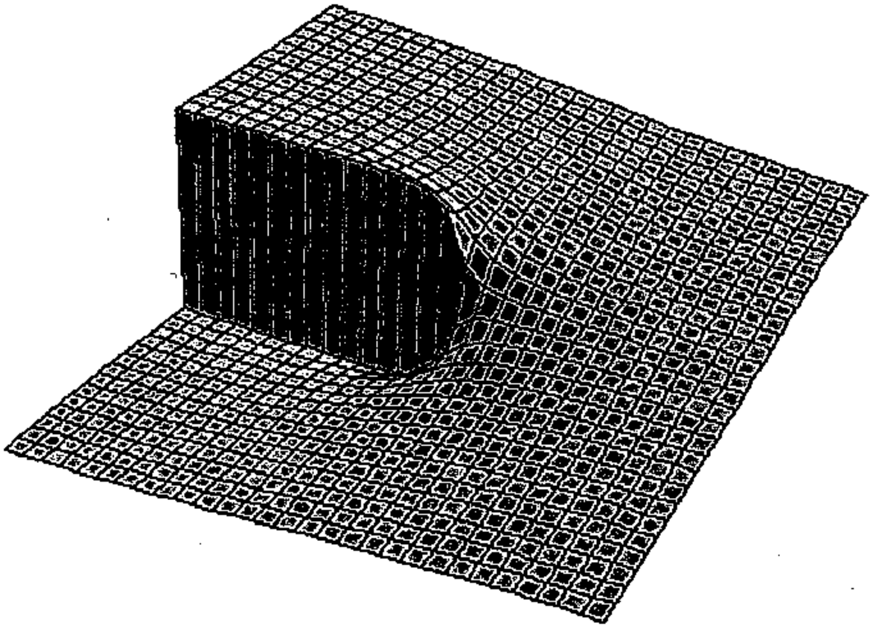

Figure 8

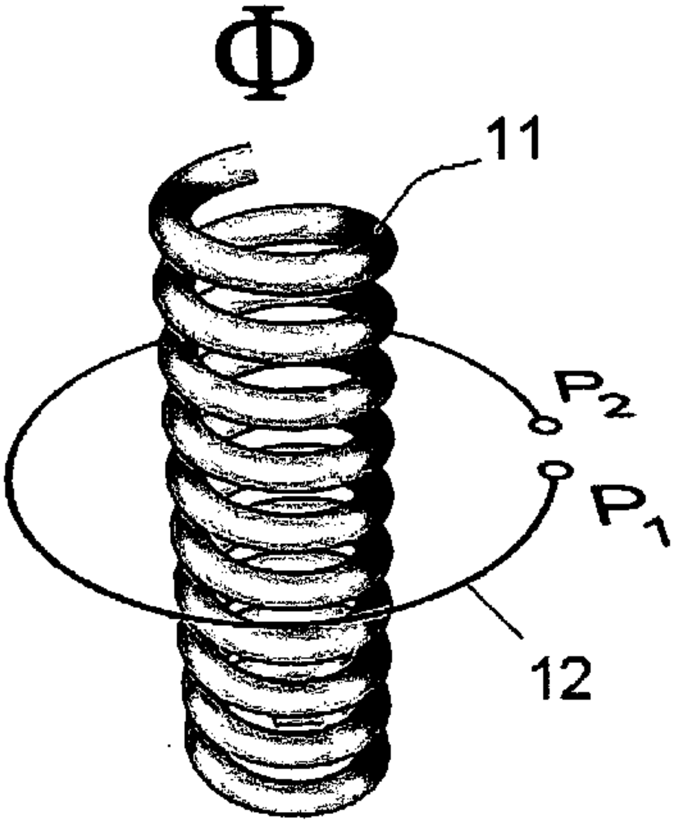

Figure 9

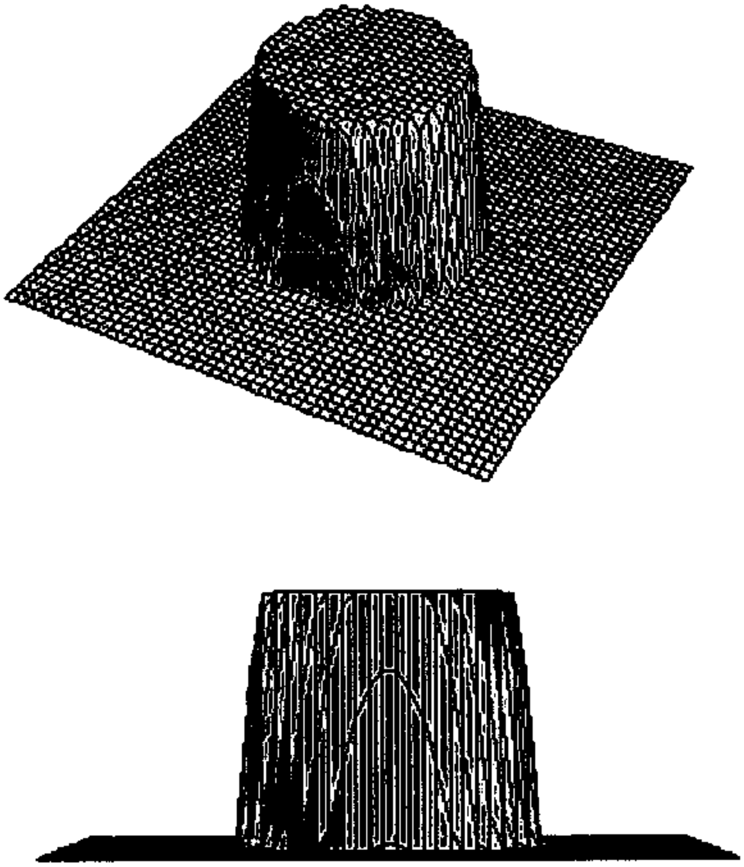

Figure 10

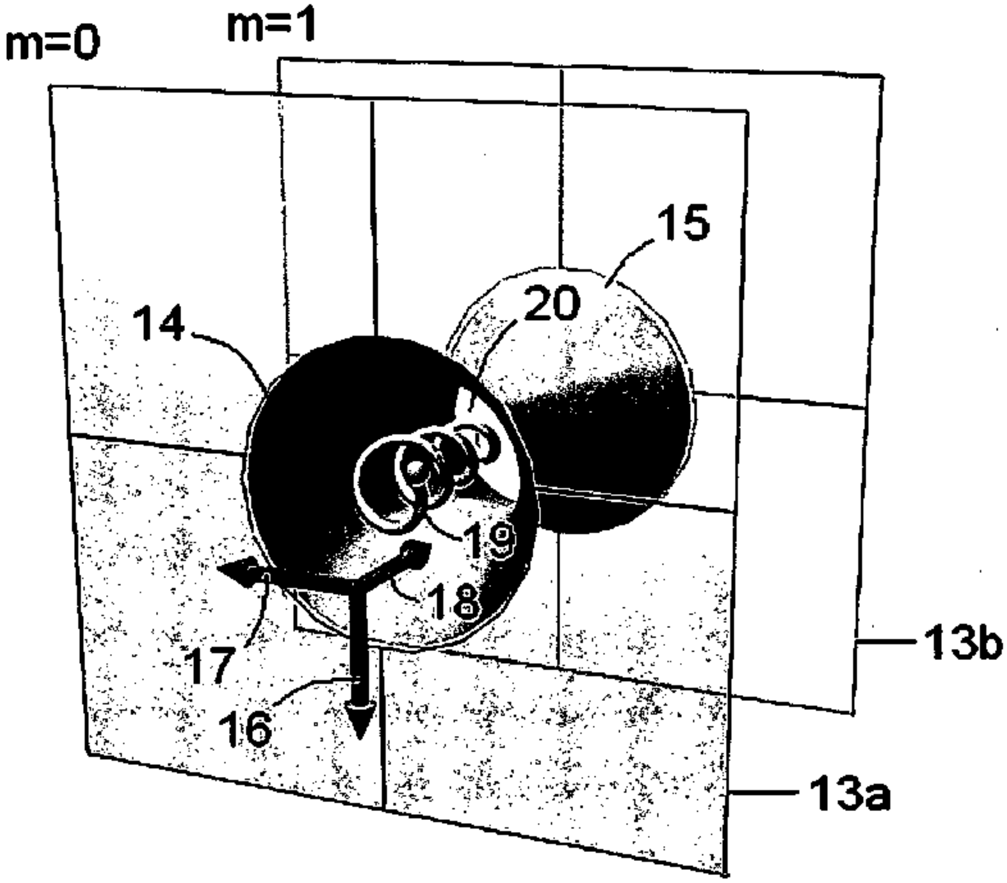
m=0
m=1
15
20
14
19
18
17
16
13b
13a

Figure 11

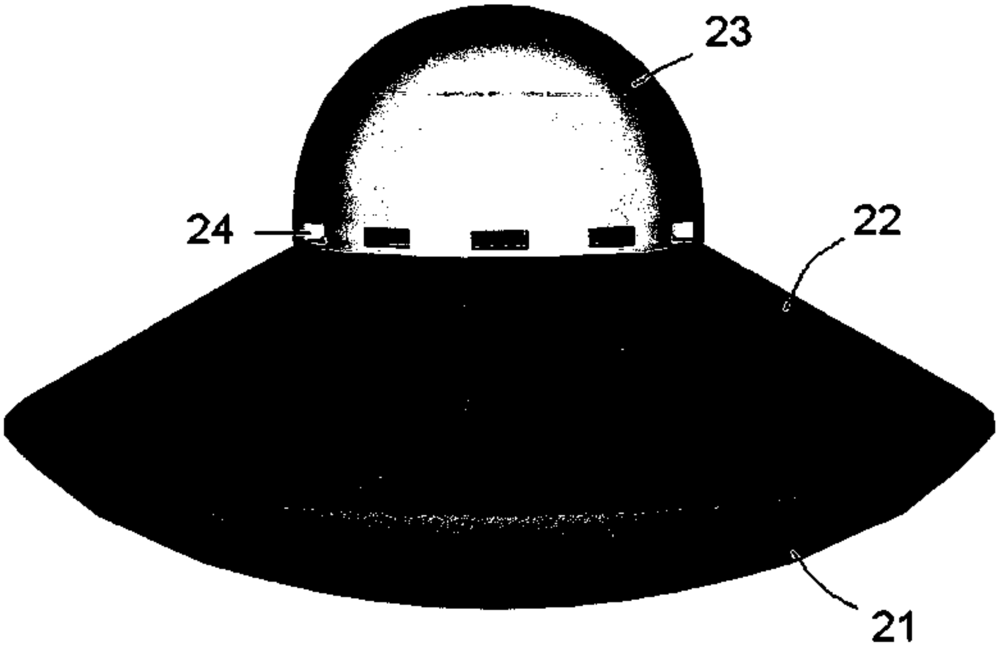

Figure 12

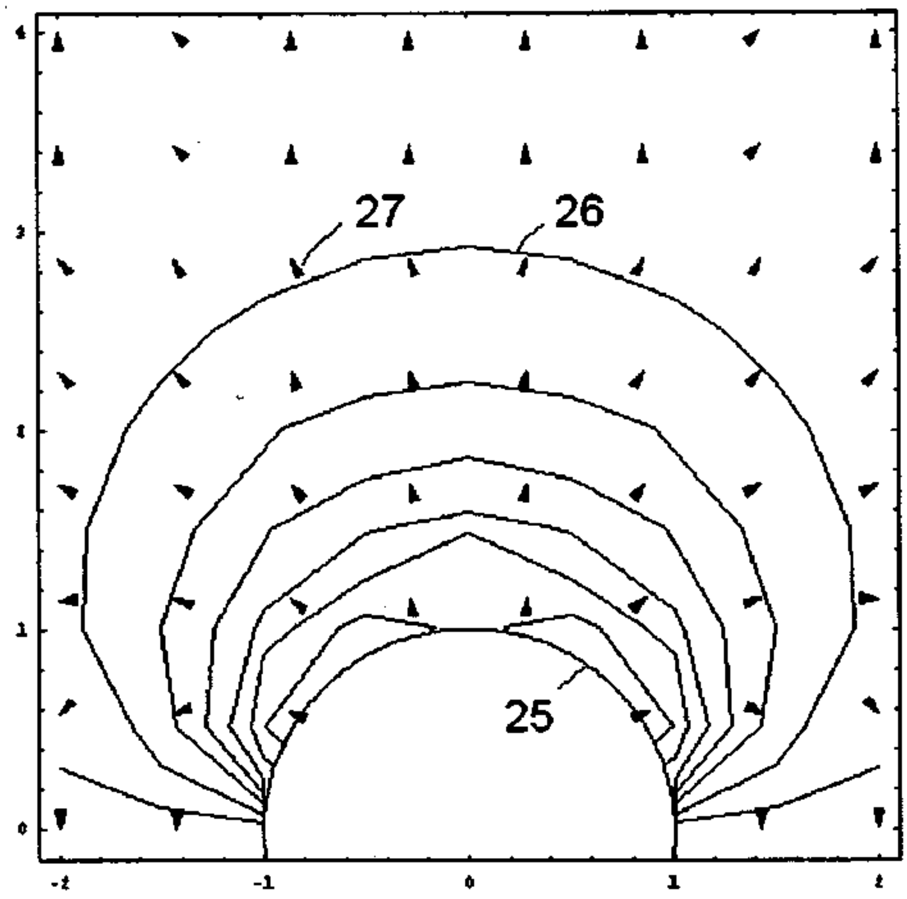

Figure 13

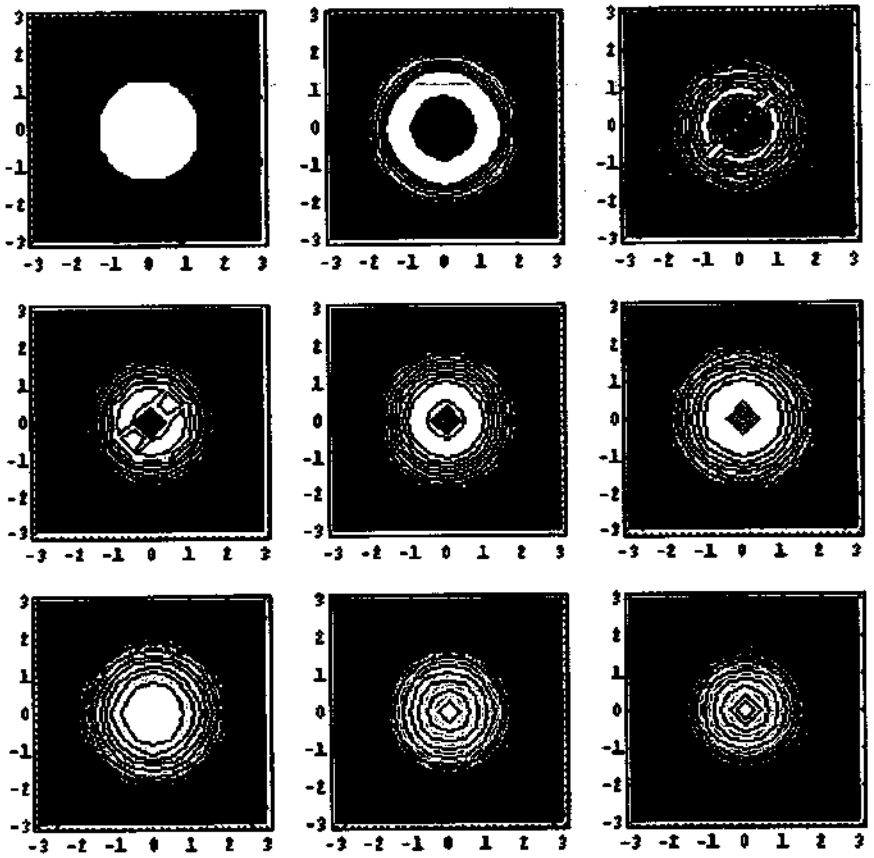

Figure 14

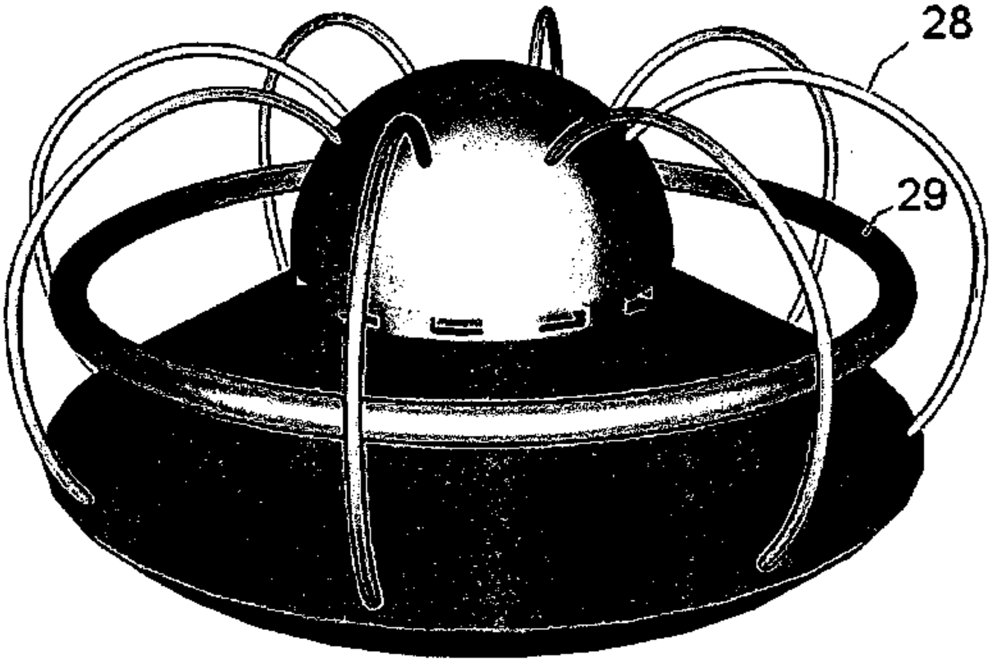

Figure 15

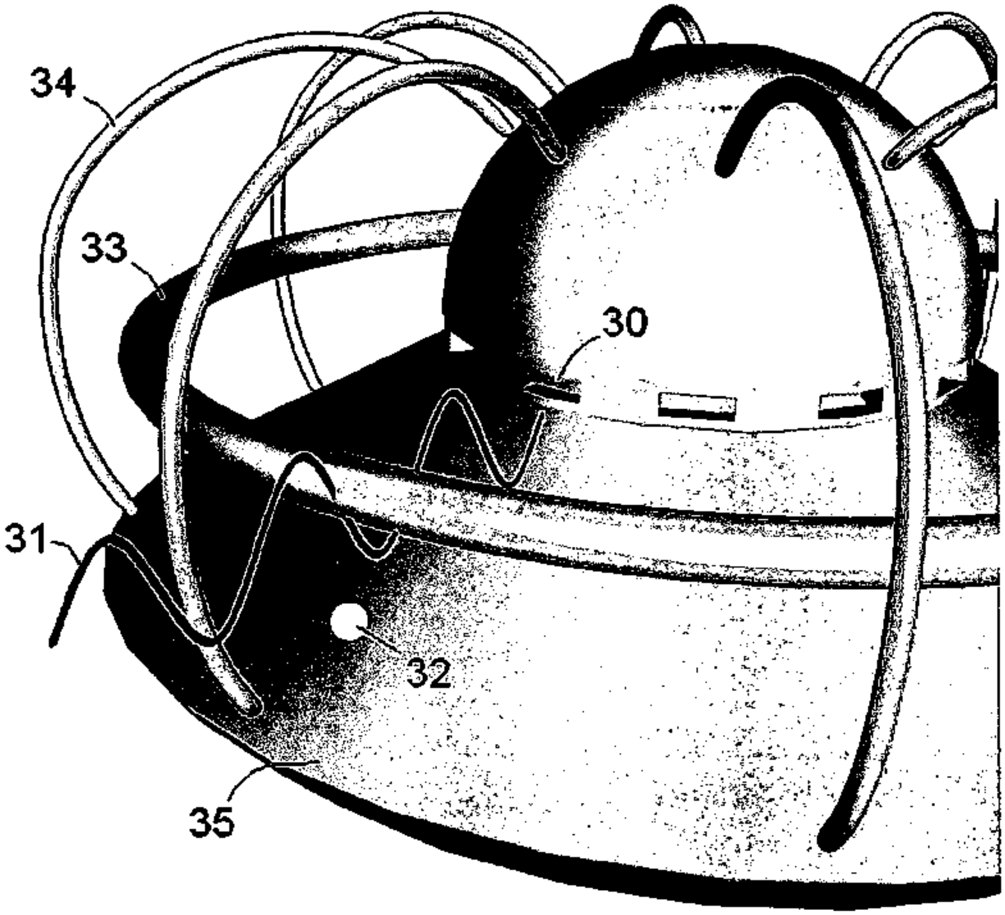
34
33
30
31
32
35

Figure 16

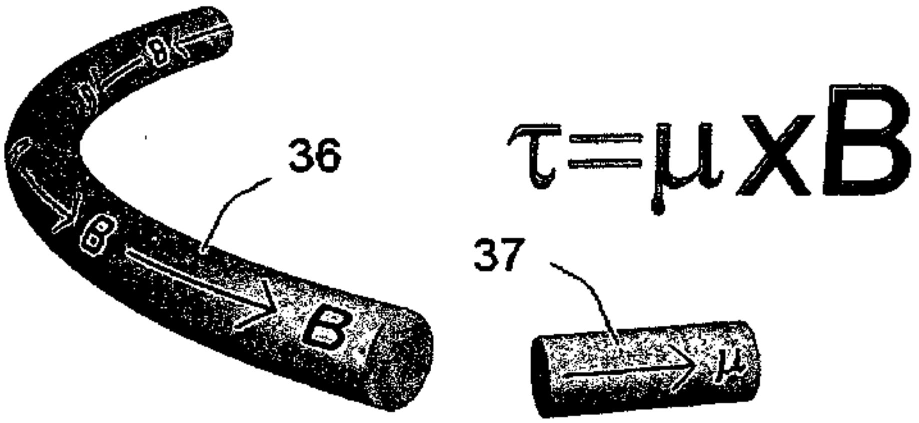
B
B
B
36
τ=μxB
37
μ

Figure 17

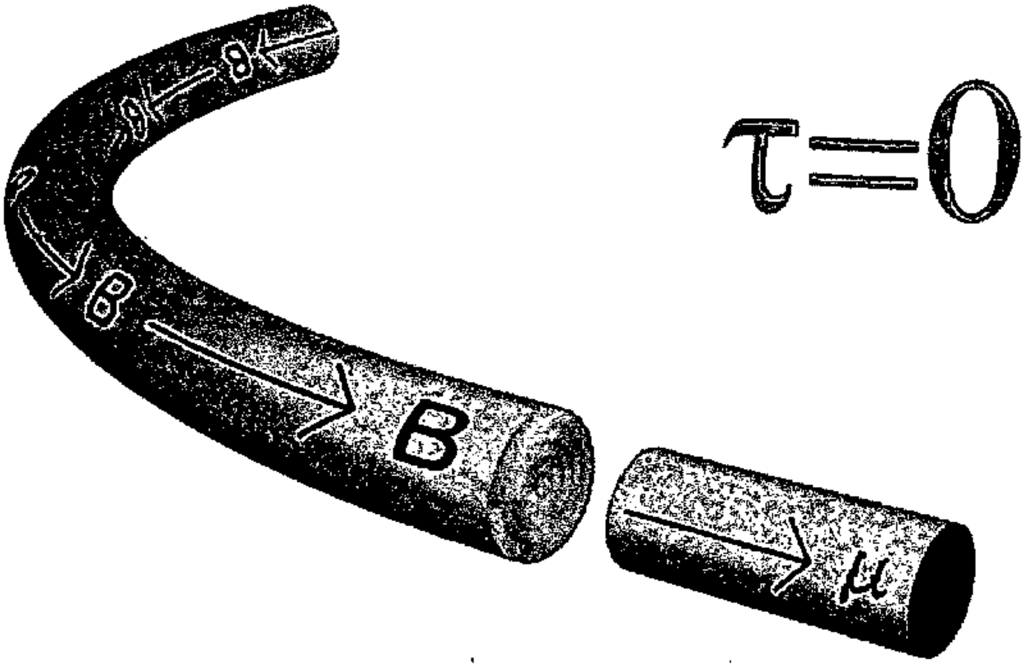
B
B
B
B
B
μ
$\tau=0$

Figure 18

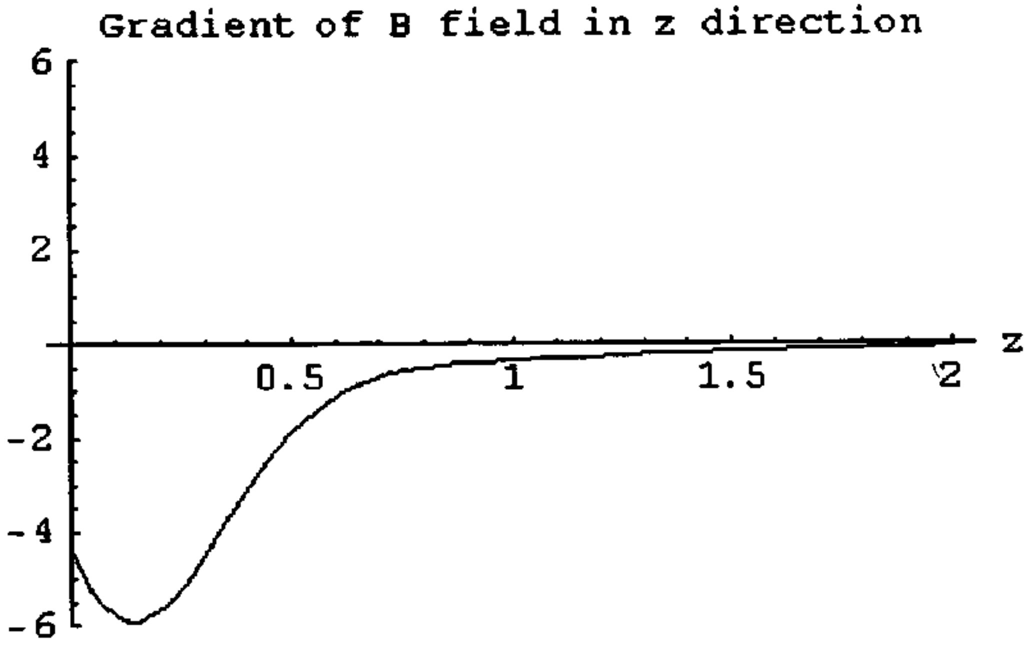

MAGNETIC MONOPOLE SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a spacecraft propulsion system that generates a field of wormholes which are threaded with a magnetic field. Acting as two attracting magnets, the spacecraft's north magnetic field is attracted to the constantly regenerating south magnetic monopoles of the wormholes which provides lift on the hull.

BACKGROUND OF THE INVENTION

[0002] According to one of Maxwell's electromagnetic equations, the curl of the electric E field is equal to the negative time rate of change of the magnetic flux density B field.

$$\overline{V} \times E = -\frac{\partial B}{\partial t}$$

The curl can be thought of as a circulation around a closed loop specified by the right-hand rule where the fingers curl in the direction of the electric field and the thumb represents the changing magnetic flux density field through the area of the loop. At no time is the electric field diverging around the loop. That is, the divergence of the curl is zero which is a well-known vector operation

$$div\ \mathrm{curl}E = \frac{\partial}{\partial t}(divB) = 0$$

The partial derivatives of divB are zero at all points in space. Performing the integration, therefore, the divergence of B is equal to a constant

div B=constant

[0003] Referring to **FIG. 1**, the following discussion is made in cylindrical coordinates {r,θ,z}. In cylindrical coordinates, the divergence of the radial B field, div B, is equal to a constant C

$$Br'[r] + \frac{Br[r]}{r} = C$$

where the prime (') represents differentiation with respect to the radius r. The solution to this equation resides in determining the constant C. In the vacuum of space without any wormholes, the constant C is zero. Because the spacecraft is surrounded by a field of wormholes, there is a magnetic flux density field threading each one. That is, each wormhole is actually a magnetic monopole, and therefore the entire field of monopoles constitutes a large magnet with one pole in this dimension and the other pole in the hyperspace dimension.

[0004] The concept of the wormhole involves a new type of scientific thinking involving the creation of a gateway between our spacetime and that of a hyperspace co-dimension. The gateway is created electromagnetically as shown by my patent applications Rotating Magnetic Vortex Generator, Magnetic Vortex Wormhole Generator, and Sulfur 8 Wormhole Generator. The gateway can also be created ultrasonically through bubble cavitation as shown in my patent application Cavitating Oil Hyperspace Energy Generator. In one experiment, smoke was blown through one side of the coil of the magnetic vortex wormhole generator and no smoke came out the other side. The smoke was blown through the wormhole into another dimension.

[0005] The existence of hyperspace is not generally known in the scientific community. The reason it exists can be thought of in the following manner. Referring to **FIG. 2**, the corners of a tetrahedron (**2**) circumscribed by a sphere (**1**) touch the sphere at an angle (**3**) of −19.47°. Looking at the planets of the solar system, the Giant Red Spot vortex of Jupiter, which can hold two planets the size of Earth, is located at this angle. On Mars, the Olympic Mons volcano, which is the size of France, is located at north 19.5° Here in the Caribbean there is a slow moving rock mantle vortex at north 19.5° that curves the islands down toward Venezuela. So the geometry of space is related to the tetrahedron. What this suggests is that there is a subspace manifold whose tetrahedral geometry projects all the constants of physics into our dimension.

[0006] Referring to **FIG. 3**, this subspace geometry is shown in the tetrahedron diagram which plots the natural logarithm of mass on the vertical axis versus the natural logarithm of wavelength on the horizontal axis. With this diagram, it is possible to visualize the entire universe on a single sheet of paper. Triangle (**4**) is the tetrahedron whose vertical centerline shown by the small π circle (a) is equal to Planck's constant h divided by the speed of light c, known as the base constant. The base constant in logs is equal to −95.91546344. Line (**5**) is the diameter of the sphere (**6**). A circle (**7**), centered on (a), with a radius equal to the speed of light squared, is tangent to the tetrahedron. It can be shown that all the constants of physics, such as the speed of light squared and the Planck mass and the Planck wavelength, are determined geometrically by this logarithmic subspace tetrahedral geometry.

[0007] Referring to **FIG. 4**, the complex plane consists of a real horizontal axis, and a complex vertical axis where the value of the complex number z is given by a radius r and an angle θ

$$z=re^{i(\theta+2\pi m)}$$

The logarithm of z is

$$\mathrm{Log}[z]=\mathrm{Log}[r]+i(\theta+2\pi m)$$

where m is an integer m=0,1,2 . . . corresponding to multiple rotations of 2π. What this means in terms of the tetrahedron diagram, referring to **FIG. 5**, is that there are multiple diagrams separated by 2π rotations. Each multiple is another hyperspace dimension. Only the log manifold has this characteristic. Referring to **FIG. 6**, the hyperspace dimension, shown as 2π circular surface (**8**), has a cut (**9**) on the undefined Log[0] origin line such that another hyperspace dimension is created below it (**10**).

[0008] Referring to **FIG. 7**, this branch cut does not bring one back to the original surface nearing the origin. It takes one down to another level of the universe into another hyperspace dimension. I can attest personally that I have been able to look into another hyperspace co-dimension as well as jump into another dimension.

[0009] Furthermore, Dr. Stephen Hawking of Cambridge University has shown that our dimension is connected to a wormhole through complex time. That is, the hyperspace dimension is rotated forward by 90° which makes it orthogonal to us. While this is the mathematical explanation as to why there are hyperspace co-dimensions, I can attest personally to the fact, as described in my patent application Full Body Teleportation, that I was teleported through hyperspace and returned to our dimension over a distance of 100 meters. Because hyperspace exists, it is then possible, using electromagnetic fields, to open wormholes between our dimension and other hyperspace dimensions.

[0010] Referring to **FIG. 8**, the constant C in the magnetic flux density B field equation is determined as follows. The end of a solenoid (**11**) is equivalent to a local magnetic source of flux Φ which represents the wormhole. A charge (**12**) encircling the solenoid has the same value at P_1 and P_2 but there is a phase difference of 2 π n where n is an integer equal to the number of times the charge encircles the solenoid. The change in phase is equal to the charge q divided by Planck's reduced constant times the flux for a solenoid of radius r=R

$$\Delta\theta = 2\pi n = \frac{q}{\hbar}\Phi = \frac{q}{\hbar}B\pi R^2$$

Solving for the magnetic flux density field threading the wormhole

$$B = \frac{2\pi n h}{q\pi R^2}$$

The differential equation becomes

$$Br'[r] + \frac{Br[r]}{r} = C[2]$$

Solving the equation for the radial field Br

$$Br[r] = \frac{C[2]r}{2} + \frac{C[1]}{r}$$

Equating this Br field with the B field and solving for the C[2] constant

$$C[2] = \frac{-2[-2\hbar n + qrC[1]]}{qr^2}$$

Substituting for C[2], the C[1] constant drops out and therefore the wormhole magnetic flux density field becomes a constant inside the throat radius R

$$Br = \frac{2\hbar n}{qR^2}$$

but the divergence moving from outside to inside the radius is a constant due to the discontinuity.

[0011] Referring to **FIG. 9**, the wormhole has a constant cylindrically-shaped magnetic field of radius R with a negative south pole due to the negative charge on the electron, Using a wormhole radius equal to one hundred times the electron radius with n equal to 10 turns, the magnetic flux density B field has a magnitude of 1.4×10^6 tesla.

[0012] Referring to **FIG. 10**, the wormhole surface (**14**) is a connection through a throat area (**20**) between space and hyperspace (**13**) to another surface in hyperspace (**15**). The electron path (**19**) is actually spiraling down the surface of the wormhole. The magnetic flux density field (**18**) points into the wormhole such that the cross product of the electron velocity (**16**) with the field (v×B) points toward the outside rim of the wormhole. Because the electron has a negative charge, the force (**17**) on the electron (−q v×B) is inward toward the centerline of the wormhole. Since one pole of the field is in space, and the other pole is in hyperspace, the wormhole appears to us as a magnetic monopole. The key to this invention is how to generate this magnetic wormhole field.

[0013] Referring to **FIG. 11**, the aluminum spacecraft is constructed of a flattened, shallow spherical lower hull (**21**), a circular upper hull with a flat sloping surface (**22**), a spherical dome cupola (**23**), and a cylindrical section (**24**) housing a circular array of radial microwave waveguides. Surfaces (**22**) and (**23**) are electrically charged, using high-voltage transformers, to an alternating electrostatic potential such that the potential on the dome is +V when the sloping hull has a −V potential and vice versa. This creates an electric field from the positively charged surface to the negatively charged surface.

[0014] Referring to **FIG. 12**, the dome (**25**) is charged to a positive potential. The spherical potential lines (**26**) emanate from the dome and curve around toward the sloping hull. The negative gradient of this potential field is the electric E field (**27**) which is perpendicular to the potential lines. The electric field lines from the dome then terminate on the sloping hull which can be seen in the lower right hand corner of the graph.

[0015] Looking down from above at the top of the dome, referring to **FIG. 13**, the oscillating electric field generates a circular oscillating magnetic field around the hull at various elevations. The last two graphs in the right-hand corner are near the top of the dome as seen by the smaller radius contour lines.

[0016] Referring to **FIG. 14**, the oscillating electric field lines (**28**) are between the dome and the sloping hull. The oscillating magnetic field (**29**) encircles the hull at the level of the microwave waveguides.

[0017] From gravitation physics, it is known that negative energy is required to open up the throat of the wormhole. In terms of differential forms mathematics, the negative energy is created by wedging the magnetic field with the radial microwave beams of the waveguides. This generates an increasing time rate of change of negative energy ρ as shown by

$$*d(B\wedge(B_{wave}+E_{wave}dt))=\frac{\partial(-\rho)}{\partial t}$$

where (*) is the Hodge Star operator, (d) the differential operator and (ˆ) the wedge operator which joins the circular magnetic flux density B field with the electromagnetic wave (B_{wave}, E_{wave}).

[0018] Referring to **FIG. 15**, the interaction of the circular magnetic field **(33)**, generated by the oscillating electric field **(34)**, interacts with the radial electromagnetic field of the microwave beams **(31)** of the waveguides **(30)** to generate negative energy **(32)** over the sloping hull **(35)**.

[0019] This combination of fields also creates the wormhole field over the hull. The spacetime curvature pressure T in the vertical z-direction is equal to the square of the circular magnetic flux density field

$$T^{zz}=\frac{B^2}{8\pi}$$

This stress-energy-momentum tensor can be thought of as a spacetime curvature proportional to the inverse of the radius squared, or as a pressure term which acts on the surface area of the hull. Thus there is the combination of a pressure stress and negative energy which creates the wormhole field over the sloping hull. Even though the magnetic flux density B field is oscillating, it is the square of the field which creates the stress. Thus the tension is still in the positive vertical z-direction.

[0020] Due to the low speed of light of hyperspace, the hyperspace energy is low density. Thus there is a positive gravitational potential between hyperspace and our dimension such that the hyperspace energy flows through the wormholes and onto the hull of the spacecraft. When this happens, a white mist forms over the hull. The effect of the hyperspace energy is to lessen the mass of the spacecraft, and relativistically increase the strength of the electromagnetic fields due to the smaller speed of light.

[0021] Referring to **FIG. 16**, the circular magnetic flux density field **(36)** is interacting with the magnetic monopole **(37)**. The monopole acts as one pole of a magnet and therefore has a magnetic moment equal to the area times the electrical current circulating through it as the electrons move from our dimension into hyperspace. The great physicist Maxwell **(1870)** pointed out that the energy associated with charges and poles is potential energy and that therefore these objects tend to move in a direction that will decrease the potential energy, similar to a brick sliding down an inclined plane. Now to reduce the potential energy is the same as to reduce the field which gives a measure of the potential energy. If two like charges are brought together they strengthen one another's field, while opposite charges reduce one another's field. Thus like charges repel and unlike ones attract.

[0022] However, the energy associated with electric currents is not potential but kinetic in origin, because it is associated with moving charges. Now in mechanics it is well known that bodies move in a direction to increase their kinetic energy, if there is a source of external energy. The magnetic field of currents is a measure of kinetic energy and currents will try to move in a direction that will increase the field. Thus similarly-directed currents attract and unlike currents repel. In **FIG. 16**, the monopole will act so as to increase the field of the circular magnetic B field generated by the spacecraft. The monopole will then rotate into alignment with the flux tube. This increases the field strength of the flux tube. There is a torque τ on the monopole equal to the cross product of the magnetic moment with the B field

$$\tau=\mu\times B$$

[0023] Referring to **FIG. 17**, when the monopole is aligned with the flux tube, the cross product is zero and there is zero torque on the monopole. Notice that even if the oscillating B field points in the opposite direction, there is still zero torque since the torque depends on the sine of the angle between them. The sin(0°) or sin(180°) is the same zero value. With the monopole aligned with the flux tube, the kinetic energy is now maximized. That is, both the flux tube and the monopole point in the θ-direction.

[0024] The magnetic B field can be represented as a vector having three components. The only component is in the angular direction

$$B=\{B_r,B_\theta,B_z\}=\{0,\ B_\theta(z),\ 0\}$$

where the field varies, as was seen in **FIG. 13**, in the vertical z-direction. The magnetic monopole field also points in the θ-direction

$$\mu=\{0,\mu_\theta(z),0\}$$

where there is a negative gradient of the monopole field in the z-direction due to the location of the microwave waveguides near the sloping hull.

[0025] The force F on the monopole is the gradient of the monopole's magnetic moment μ with the magnetic flux density B field

$$F=\nabla(\mu\cdot B)=\{0,\ 0,\ \mu_\theta B_\theta'(z)+B_\theta(z)\mu_\theta'(z)\}$$

which says that there is a force on the monopole in the z-direction equal to the magnetic moment times the gradient of the magnetic field in the z-direction plus the magnetic field times the gradient of the magnetic moment in the z-direction.

[0026] Referring to **FIG. 18**. differentiating the magnetic field in the z-direction shows that the gradient is negative outside the hull. This can also be seen visually in **FIG. 13** where the graphs decrease in intensity.

[0027] A negative gradient for both the magnetic moment and the field means that the force on the monopole is negative. The force on the tube connected to the hull is therefore the negative of a negative, yielding a positive lift force.

$$F_{monopole}=-\{0,\ 0,|\mu_\theta B_\theta'(z)+B_\theta(z)\mu_\theta'(z)|\}=-F_{tube}$$

$$F_{tube}=+\{0,\ 0,|\mu_\theta B_\theta'(z)+B_\theta(z)\mu_\theta'(z)|\}$$

which says there is an upward lift force on the hull due to the combination magnetic monopole and flux tube. This lift force is in addition to the vertical lift force generated by the spacetime curvature created by the electromagnetic fields themselves.

SUMMARY OF THE INVENTION

[0028] This invention is a spacecraft propulsion system that utilizes electromagnetic fields and microwaves to generate negative energy and a spacetime curvature over the hull. The hull consists of a hemispherical dome, a circular array of radial microwave waveguides, a sloping flat hull, and a shallow spherical hull on the bottom. Alternating current high voltage transformers connected to the dome and the sloping hull generate a curving oscillating electric field between the dome and the sloping hull. Due to this oscillation, a horizontal circular oscillating magnetic flux density B field is created around the dome.

[0029] By firing the microwaves at right angles to the B field, negative energy is created over the hull. The negative energy and spacetime curvature pressure generate wormholes between space and hyperspace. Because hyperspace has a low speed of light and positive gravitational potential, low density hyperspace energy flows through the wormholes and onto the hull. The effect of the hyperspace energy is to lessen the mass of the vehicle and to increase the strength of the electromagnetic fields. Because the resistance of hyperspace is less than the resistance of space, electrons spiral down the wormholes into hyperspace. This creates a magnetic field through the wormhole with one pole in our dimension and the other pole in hyperspace. Thus a field of magnetic monopoles is created over the hull.

[0030] The magnetic monopoles, which represent kinetic energy, align themselves with the magnetic flux tubes in order to maximize the total magnetic field. Because there is a gradient of the monopoles and field in the vertical direction, a negative force develops on the monopoles equal to the gradient of the dot product of the magnetic moment of the monopole with the B field. Thus the opposite reaction is a positive force on the flux tubes attached to the hull which is equivalent to bringing the north pole of a magnet together with the south pole of a second magnet. Because the hull constantly regenerates the wormhole field, the hull experiences a constant upward lift force. This is in addition to the lift generated by the spacetime curvature pressure which is proportional to the square of the magnetic flux density B field.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0031] **FIG. 1**. Perspective view of cylindrical coordinate system $\{r, \theta, z\}$.

[0032] **FIG. 2**. Perspective view of tetrahedron circumscribed by sphere.

[0033] **FIG. 3**. Tetrahedron diagram showing speed of light squared is determined by the tetrahedron.

[0034] **FIG. 4**. Complex number z representation in the complex plane.

[0035] **FIG. 5**. Perspective view showing multiple log manifold hyperspace dimensions.

[0036] **FIG. 6**. Perspective view of orthogonal hyperspace dimensions.

[0037] **FIG. 7**. Perspective view of Log[z] showing cut along origin.

[0038] **FIG. 8**. Perspective view of charge phase shift around a magnetic flux.

[0039] **FIG. 9**. Perspective view of wormhole magnetic flux density B field.

[0040] **FIG. 10**. Perspective view of magnetic monopole wormhole.

[0041] **FIG. 11**. Perspective view of spacecraft.

[0042] **FIG. 12**. Graph showing electrostatic potential and electric field over dome.

[0043] **FIG. 13**. Animation showing circular magnetic field around dome at increasing elevation.

[0044] **FIG. 14**. Perspective view of electric and magnetic fields around hull.

[0045] **FIG. 15**. Perspective view of generation of negative energy.

[0046] **FIG. 16**. Perspective view of monopole misaligned with flux tube with torque.

[0047] **FIG. 17**. Perspective view of monopole aligned with flux tube at zero torque.

[0048] **FIG. 18**. Graph showing negative gradient of flux tube in the z-direction.

DETAILED DESCRIPTION OF THE INVENTION

[0049] 1. The aluminum hull is made by a technique called stretch forming which uses hydraulic cylinders to stretch a large sheet of aluminum to its yield point. This makes the aluminum sheet soft and pliable. Using a die which has been CNC machined to the desired hull profile, the sheet is then die pressed into a very rigid, smooth and lightweight structure requiring no other support. There are actually three dies consisting of a spherical dome, sloping hull, and shallow spherical dome.

[0050] 2. The rest of the hull consists of a cylindrical hull with a radius equal to the upper dome. A segment of this hull is designed on a 3D computer graphics program and stored as a stereolithography *.stl file. The file is then transmitted over the Internet to a server who prints up the part on an xy-plotter with an ultraviolet laser and ultraviolet light sensitive polymer bath. The computer model is sliced by a special program into many thousands of slices which are printed one over the other until the part is completed. The server returns the part next day by Express Mail. Several parts are then molded using liquid plastic such as to form the complete ring. A sand mold is then constructed from all the molds to form a plastic cylindrical hull having the waveguide slots molded into it. The waveguide silver-coated aluminum boxes are then installed in the slots and connected to the frequency generators and amplifiers. The purpose of the plastic waveguide cylinder is to separate the electrostatic charges on the dome and the sloping hull In this particular case we used a dome from another spacecraft design which saved on the cost of the die.

I claim:

1. A spacecraft propulsion system consisting of the following components:

a lower hull made of aluminum sheet having a shallow spherical profile;

a circular flat sloping hull made of aluminum sheet attached to the top of the lower hull on the periphery;

an electrically-insulated plastic-molded tubular cylindrical hull containing slots for mounting an array of radial microwave waveguides, attached to top of the flat sloping hull;

a hemispherical cupola in the shape of a dome made of aluminum sheet mounted on top of the insulated cylindrical hull;

an array of rectangular microwave waveguides mounted in the waveguide slots of the cylindrical hull;

a high-voltage alternating current transformer with one side electrically attached to the upper dome and the other side electrically attached to the flat sloping hull; and

a frequency generator and amplifier to drive the microwave waveguides.

2. By means of claim (1), an oscillating electric field is created between the upper dome and the sloping hull using the high-voltage alternating current transformer.

3. By means of claim (2), an oscillating circular magnetic flux density field is generated around the sloping hull and upper dome.

4. By means of claims (1) and (3), negative energy is generated by the radial microwave beams of the waveguide array impinging on the circular magnetic flux density field around the hull.

5. By means of claim (3), a positive spacetime curvature pressure constituting a lift force is developed over the hull in the vertical direction proportional to the square of the field.

6. By means of claims (4) and (5), a field of wormholes between space and hyperspace are generated over the hull.

7. By means of claim (6), due to the positive gravitational potential between hyperspace and space, low-density hyperspace energy flows through the wormholes onto the hull to reduce the mass of the spacecraft and strengthen relativistically the electromagnetic fields.

8. By means of claim (6), electrons, emitted by the charged hulls, spiraling down the wormholes, generate a field of magnetic monopoles with one pole in space and the other in hyperspace.

9. By means of claims (6) and (3), the magnetic monopoles maximize their kinetic field energy by aligning with the magnetic flux tube.

10. By means of claim (9), the gradient in the vertical direction of the dot product of the magnetic moments of the monopoles with the magnetic flux density field is a negative force on the monopoles and an equal but opposite positive lift force on the magnetic flux tubes attached to hull.

11. By means of claims (10) and (5), a dual method of providing apositive lift force on the spacecraft is constituted.

* * * * *

US 20060073976A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0073976 A1**

Pohlman (43) **Pub. Date: Apr. 6, 2006**

(54) METHOD OF GRAVITY DISTORTION AND TIME DISPLACEMENT

(76) Inventor: **Marlin B. Pohlman**, Tulsa, OK (US)

Correspondence Address:
STITES & HARBISON PLLC
1199 NORTH FAIRFAX STREET
SUITE 900
ALEXANDRIA, VA 22314 (US)

(21) Appl. No.: **10/954,767**

(22) Filed: **Oct. 1, 2004**

Publication Classification

(51) **Int. Cl.**

F16C 39/06	(2006.01)
G01V 7/00	(2006.01)
G09B 23/06	(2006.01)
F41B 6/00	(2006.01)
H01S 3/09	(2006.01)

(52) **U.S. Cl.** .... **505/166**; 505/180; 434/300; 73/382 G; 505/164

(57) ABSTRACT

A method for employing sinusoidal oscillations of electrical bombardment on the surface of one Kerr type singularity in close proximity to a second Kerr type singularity in such a method to take advantage of the Lense-Thirring effect, to simulate the effect of two point masses on nearly radial orbits in a 2+1 dimensional anti-de Sitter space resulting in creation of circular timelike geodesics conforming to the van Stockum under the Van Den Broeck modification of the Alcubierre geometry (Van Den Broeck 1999) permitting topology change from one spacelike boundary to the other in accordance with Geroch's theorem (Geroch 1967) which results in a method for the formation of Godel-type geodesically complete spacetime envelopes complete with closed timelike curves.

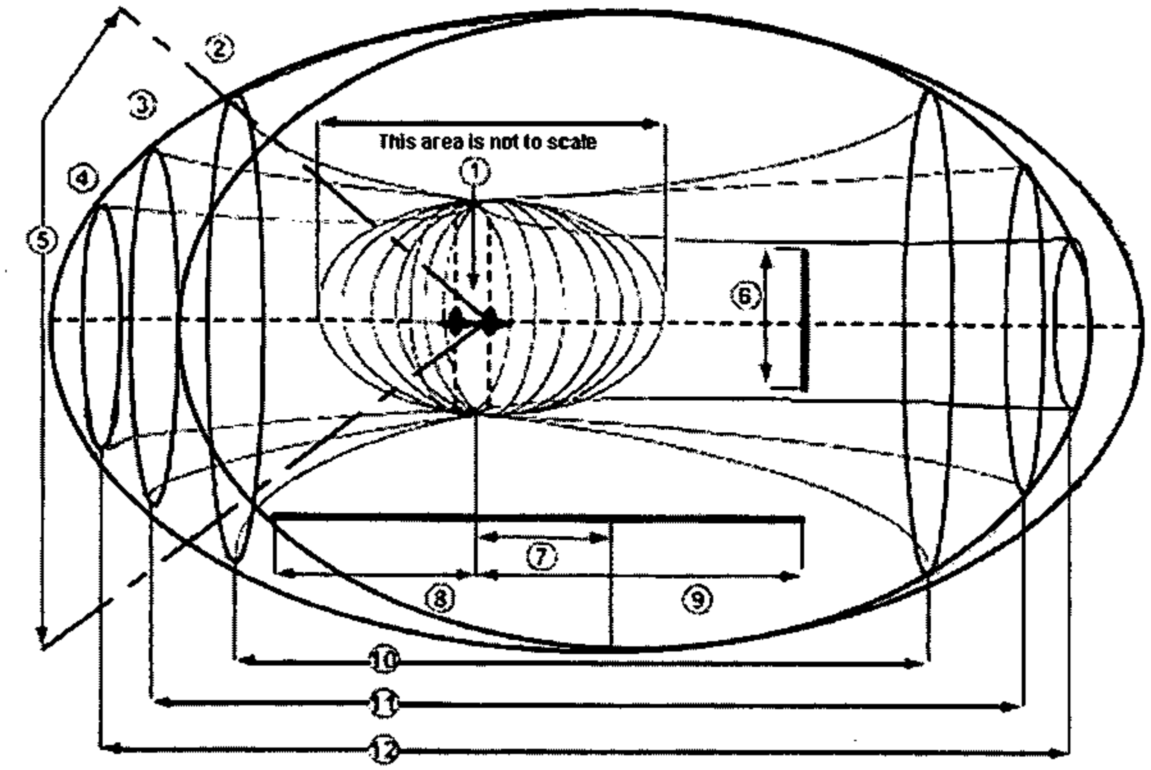

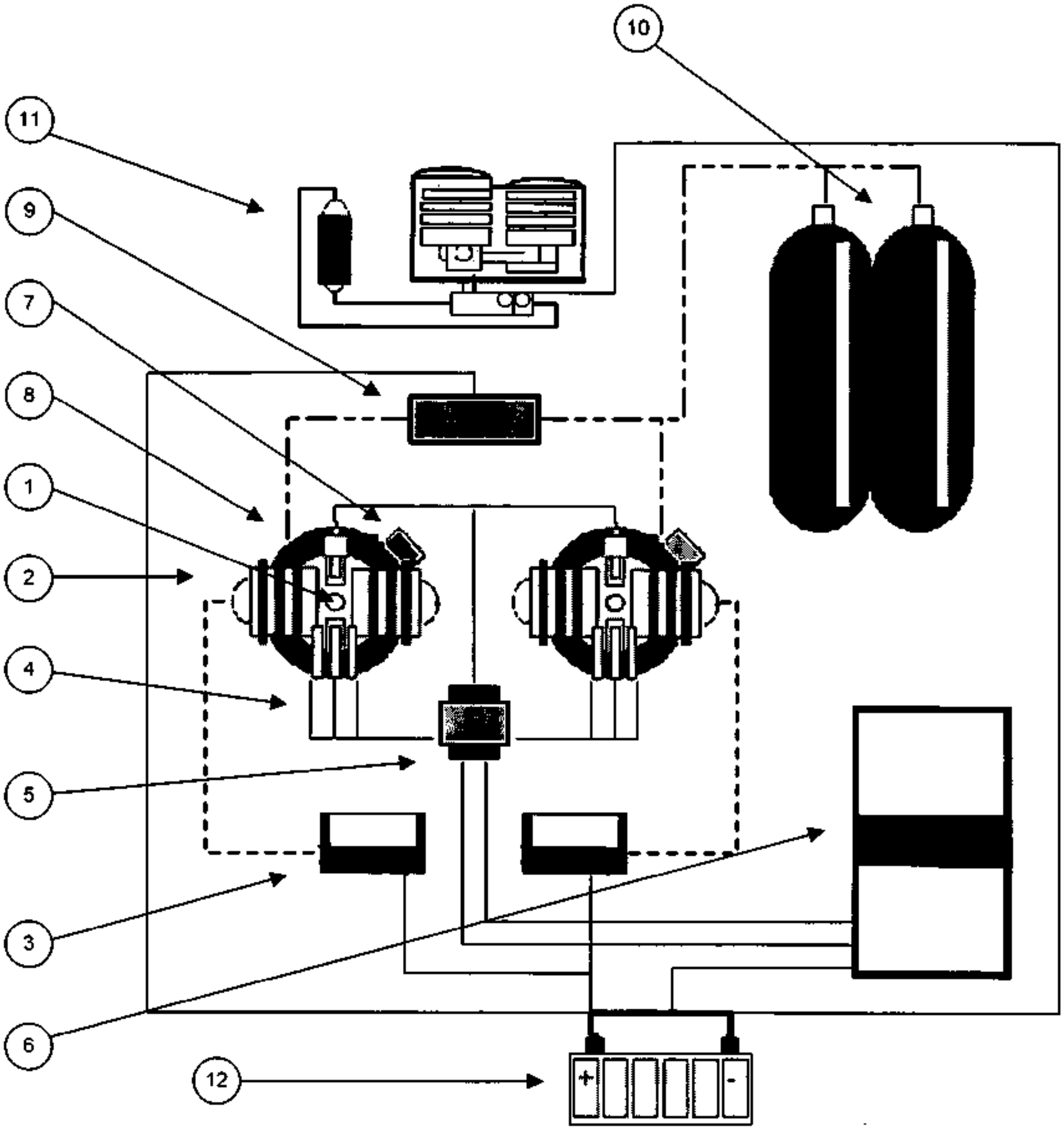
10
11
9
7
8
1
2
4
5
3
6
12
+
-

Figure 1

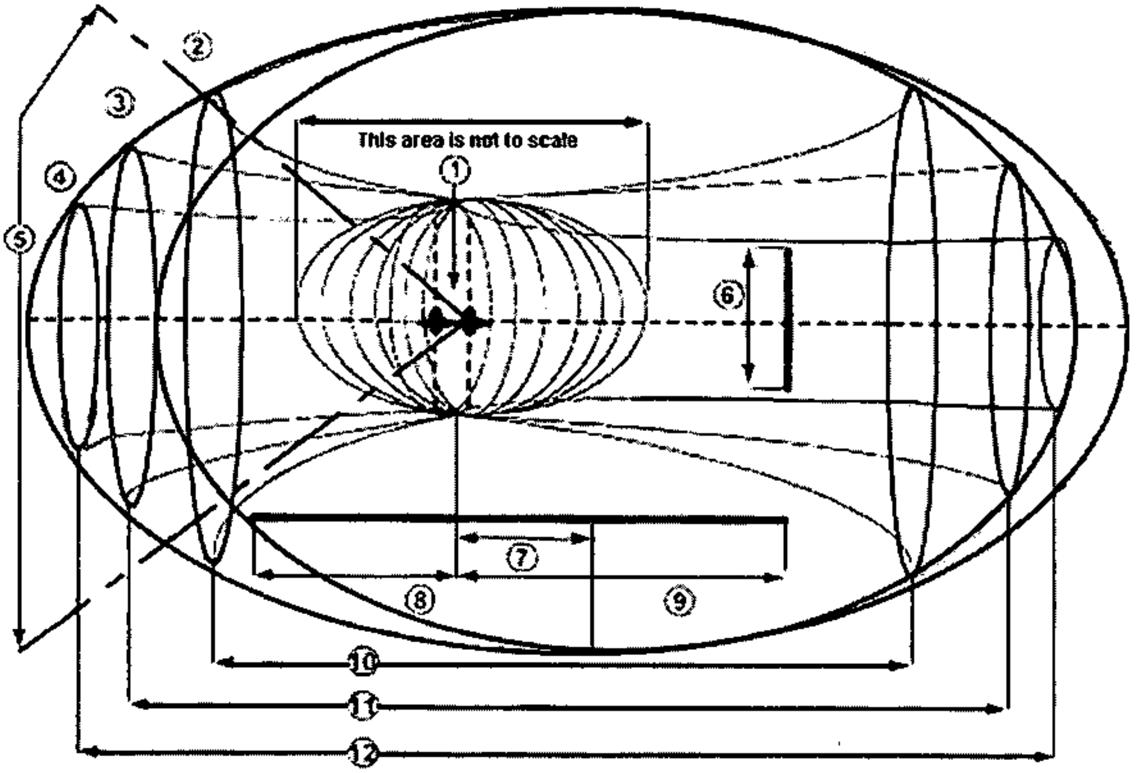
This area is not to scale
①
②
③
④
⑤
⑥
⑦
⑧
⑨
⑩
⑪
⑫

Figure 2

METHOD OF GRAVITY DISTORTION AND TIME DISPLACEMENT

FIELD OF THE INVENTION

[0001] The present invention relates to the use of technical time displacement devices, which operate by the modification of gravitational fields. These drive systems do not depend on the emission of matter to create thrust to take advantage of time dilation, but rather create a change in the curvature of space-time, in accordance with general relativity. This allows travel across topologies by warping space-time, to produce a topology change from one spacelike boundary to the other in accordance with Geroch's theorem (Geroch 1967)

THEORETICAL BACKGROUND OF THE INVENTION

[0002] The concept of gravity should be examined in the light of quantum gravity and in turn as a component of quantum physics itself. The fundamental minimal quantum of energy in quantum physics is Planck's constant; h. Thus in accordance with the energy equivalence formula $E=mc^2$, the fundamental minimum quantity of mass (m_q) can therefore be derived, from known constants by; $m_q=h/c^2$ (1). Taking this minimal mass, it is possible to show that the formation of all matter, the forces of nature and indeed space-time itself derive from this single quintessential quantity.

[0003] Thus if the number of quintessences in a system is; $n_q=m/m_q$; then the total Energy of the system is more logically given by, the energy of a single quintessence (h); directly multiplied by the number of quintessences (n_q) in that system, thus

$$E=hn_q=mc^2 \qquad (1\text{ a}).$$

[0004] Furthermore, this minimal mass, termed quintessence, can form the basis of the existence of a quantum gravitational field in the form of a space-time lattice, from which quantum gravity may be derived from first principles. Furthermore, the conglomeration of these quintessences also accounts for the formation of the elementary particles and the forces acting between them, as in superstring theory. This concept explains the formation of matter and the forces of nature on a quantum mechanical basis and directly explains the existence of wave particle duality. Thus as $n_q=m/m_q$; the frequency of light and matter (f) is determined, directly, from the number of constituent quintessences. This leads automatically to the fundamental equation, derived from (1), $f=n_q=E/h$, where n_q is the number of quintessences, which leads directly to the frequency of both light and matter. This in turn leads directly to a Universal wave equation for matter and light $\lambda=c/\beta n_q=hc/\beta E$ (2), where β is the relative directional velocity, v/c. As the momentum, $p=\beta\times E/c$, then this equation also gives the standard de-Broglie wave equation, $\lambda=h/p$ in agreement with current theory and experiments[1].

[0005] Using the Universal wave equation, the standard equation for special relativity, $m'=m_o/(1-\beta^2)^{1/2}$, derives from first principles. Also from these observations, a modified Dirac wave equation may be derived, $E\psi=(-j\beta\cdot\nabla+\beta m)\psi$ (2a), the results of which have been recently verified by a paper in which the orbitals of electrons were experimentally directly visualised[2]. Moreover, a fundamental equation for general relativity can be formulated, where G is the gravitational constant and r_q is the given radius of quintessence; $G=9(r_q)^2c^4/\lambda\beta E$ (3), such that the Universal wave equation is in direct agreement with general relativity[3]. Thus special and general relativity and quantum mechanics can be unified.

[0006] From here it is possible to proceed in a number of ways; the geometric structure of the electron and the forces of Nature may be derived from first principles and in turn the structure of the quarks, including the top and bottom, otherwise known as truth and beauty can be seen. Moreover, the presence of a space-time lattice results in an understanding of quantum EPR effects. By allowing a theoretical flow of energy through the space-time lattice it can be shown that:

[0007] Energy is not bound by space-time

[0008] Thus logically accounting for phenomena such as entanglement and quantum tunnelling. Quintessence can also be used to explain, logically, the inner physics of a black hole, the missing mass of the Galaxy, the continuing expansion of the Universe, Guth's inflationary theory and the Big Bang. Hence, it is now possible to understand the Universe, including space-time, matter and the forces of nature from the radius, mass and vibration of a single quantity, quintessence.

[0009] With this understanding of space-time, matter and the forces of Nature, and in particular gravity, it is possible to demonstrate that the modification of gravitational fields, and in turn the warping of space-time, can be technically readily achieved.

[0010] Using standard equations for special relativity, $m'=m_0/(1-\beta^2)^{1/2}$, it can be demonstrated that by differentially increasing the velocity of electrons, by applying a differential current, their mass can be increased in a specific way. In turn by increasing the mass of electrons, by general relativity, the number of gravitons emitted from these electrons can be modulated. By multiplying this effect using an ultracentrifugational device the differential graviton emission can be manifestly amplified. This in turn, in accordance with general relativity, will cause a change in the curvature of space-time.

[0011] This effective warping of space-time does not, of necessity, imply superluminal velocities, but does allow the creation of warp drive systems, which do not depend on the creation of thrust by the ejection of material as used in current space technologies.

Part 1—Fundamental Laws of Physics

[0012] Quintessential Mass

[0013] The quantum physical, minimum component of energy is Planck's constant; h. To define the minimal component of mass, using the standard energy equivalence formula; $E=mc^2$, such a minimal mass (m_q) would be required to have the value equivalent to; $m_q=h/c^2$ (1). The total mass of a system (m) would then be; $m=n_qm_q$, where (n_q) is the number of these minimal units. Thence, the total energy of a system can be derived from the minimal energy; h, multiplied by the number of these energy units (n_q). Thus as, $E=mc^2$, then also $E=m_qn_qc^2$ and substituting $m_q=h/c^2$, the energy equivalence formula has the more logical formulation; $E=hn_q$(1a). Thus the energy of a system is equivalent to the minimal energy unit; h, multiplied by the number of those minimal energy units (n_q).

[0014] Taking this minimal mass/energy, it is possible to show that all matter, the forces of nature and space time can be constructed from this single quintessential quantity. Moreover, using this quantity the laws of physics can be derived from first principles. Thus, a priori, all components of the physical universe, including space-time, can be constructed from this minimal mass component, termed quintessence.

[0015] Wave Particle Duality

[0016] If the presence of quintessence accounts for the structure of matter and if matter itself forms from the number of quintessences, then the frequency of matter and thus wave particle duality directly arises from first principles. Specifically the wavelength of matter derives from the vibration of quintessence from which it is constituted. Thus the frequency (f) and in turn the wavelength of light and matter is directly equivalent to the number of quintessences contained within it. We find that the actual frequency of light can be directly derived from first principles from the effective mass of the photon (m.) and thus by the number of quintessences (n_q) it contains.

[0017] Thus for light conventionally:

$$f=E/h$$

[0018] and if $E=mc^2$, and $h=m_qc^2$, then

$$f=m_\gamma c^2/m_qc^2$$

and

$$f=m_\gamma/m_q=n_q$$

Thus

$$f=n_q \tag{4}$$

[0019] Thus the formula for the frequency of light E=hf is now readily explained by the observation that the frequency is determined quite directly from the number of quintessences n_q within the photon.

[0020] The wavelength is thus also given by:

$$\lambda=c/f=m_{qc}/m_\gamma=h/p$$

[0021] We can now show that the frequency of matter also has the same derivation from quintessence, as has the frequency of light. The frequency of matter is again equivalent to the number of quintessences it contains. Thus the wave particle duality of matter itself can be explained by its composition from quintessence. The amount of quintessences contained within a electron sphere will depend on the number of quintessences constituting the electron and those passing through it as a result of its relative velocity β^2 (where $\beta=v/c$); effectively its relativistic momentum (p). The frequency will then be related to the total number of quintessences. Thus for matter,

$$f=\beta^2n_q \tag{4a}$$

[0022] Thus it is possible to derive the conventional de Broglie wave equation for matter from first principles. Thus, as $\lambda=v/f$, we have:

$$\lambda=v/\beta^2n_q \tag{5}$$

[0023] thus as $n_q=E/h$

$$\lambda=hc/\beta E \tag{2}$$

[0024] and as conventionally $\beta E/c=p$, then for matter:

$$\lambda=h/p$$

[0025] Provided that in the de Broglie equation, the momentum of the object is calculated using the relativistic mass, thus accounting for the total number of quintessences n_q in an object, this gives an accurate value for the wavelength of matter[1].

[0026] Thus the wavelength of matter follows directly from its constituents, quintessence. As matter is made of quintessence, similarly to light, its frequency depends on the number of quintessences n_q within it, traveling relative to the speed of light. Moreover, $\lambda=hc/\beta E$, underpins a fundamental relationship between wavelength and energy. Furthermore, this is mathematically the same as the term $\lambda=hv/\beta^2E$, giving a relativistic expression for the wavelength of matter, from which the relativistic equations may be directly derived

[0027] Wave Equations

[0028] The derivation of wave particle duality from first principles also now allows the derivation of a modified wave equation for matter.

[0029] To derive his wave equation Shrodinger commenced with the de Broglie equation using momentum (p). For lower energies the momentum of an electron is conventionally derived from the kinetic energy of the electron and the mass of the electron m_0. Thus conventionally:

$$E_k=\frac{1}{2}mv^{1/2} \quad \text{and} \quad p=m_0v$$

Thus

$$E_k=p^2/2m_0$$

then

$$p=\sqrt{(E_k\cdot 2m_0)}$$

[0030] and conventionally, the de Broglie equation can also be written as:

$$\lambda=h/p=h/\sqrt{(E_K\cdot 2m_0)}$$

[0031] In turn the Shrodinger wave equation directly derives from the square of the above classical non relativistic term for kinetic energy:

$$\lambda^2=h^2/E_k\cdot 2m_0$$

$$\text{thus } E_k=\frac{h^2}{2m}\cdot\frac{1}{\lambda^2}$$

$$\text{As } E=E_k+V$$

$$\text{then } E\psi=-\frac{h^2}{2m}\cdot\frac{d^2\psi}{dx^2}+V\psi=jh\cdot\frac{d\psi}{dt}$$

[0032] However, the Shrodinger equation, may be refined by taking into account relativity. Thus the true values for the energy are given by the relativistic momentum (p).

[0033] A fundamental relativistic wave equation for ψ, and its logical derivation may now be developed through the concept of quintessence as a fundamental constituent of matter.

[0034] The amount of quintessences in the electron is determined by the number of quintessences forming the electron at rest, plus the amount of quintessences passing through it due to its relativistic velocity, which will determine the relativistic momentum (p) of a particle.

[0035] The frequency of matter can now be readily calculated from first principles to give a more accurate result. Thus as matter is made of quintessence, similarly to light, its frequency is equal to the number of quintessences n_q within it. The wavelength will depend on its velocity travelling relative to the speed of light and thus multiplied by the relative velocity compared to $c(\beta=v/c)$;

[0036] Hence for matter as previously shown:

$$\beta=v/\beta^2 n_q=hc/\beta E \tag{2}$$

[0037] And conventionally

$$E=(p^2c^2+m_0{}^2c^4)^{1/2}$$

[0038] Using these equations, we can now, also, reformulate the Shrodinger wave equation, which has the advantage that relativity can be treated in a quantum mechanical way. Thus if the wave energy of matter is defined as:

$$E_\lambda = \frac{\beta\sqrt{p^2c^2+m_0^2c^4}}{c^2}$$

thus

$$E_\lambda = \sqrt{(\beta^2p^2/c^2)+\beta^2m_0^2}$$

[0039] which in complex space generalises to

$$E_\psi=(-j\beta\cdot\nabla+\beta m)_\psi \tag{2a}$$

[0040] As the term

$$\alpha = \frac{e^2}{hc}4\pi\varepsilon_0;$$

also represents the ground state ratio of the velocity of the electron to c. Thus $\alpha=\beta=v/c=1/137$.

[0041] Thus, also

$$E_\psi=(-j\alpha\cdot\nabla+\beta m)_\psi$$

[0042] This is thus the standard relativistic equation that Dirac was able to construct from the Shrodinger wave equation. This relativistic equation can be derived from the modified wave equation. This takes into account the relative mass energy which the quintessential wave equation contains.

[0043] Where importantly the term $\beta\cdot m$ is the mass m, multiplied by the ratio of the relative velocity to light $\beta=v/c$, and the term a is also essentially the relative velocity of the electron.

[0044] The Dirac equation was an empirical formula which worked mathematically, nevertheless even Dirac admitted it was not logically understood. The importance of these equations is that they show that the existence of quintessence allows the waveparticle duality of matter to be explained and mathematically derived from first principles, Thus the frequency of matter or even light is simply determined by the number of quintessences it contains.

[0045] Indeed, a recent publication in Nature has suggested that the direct visualisation of the orbitals of electrons shows that these are in very close agreement with theory. However, there is a significant departure from theory, in the interstitial molecular regions, suggesting that the higher velocities of the electrons obey the modified Dirac equation. Thus these orbitals were in keeping with the modified Dirac equation, which itself may be derived from the wave equation above, $\lambda=hc/\beta E$

[0046] The Shrodinger wave equation will approximate to the correct values until v approaches c. Indeed the Shrodinger equation will give similar answers as that derived from equation, under most experimental conditions.

[0047] However, equation 2 and its derivative may have advantages over standard Shrodinger theory with relativistic speeds. Furthermore, equation 2, conceptually shows that the wave particle duality of matter derives from the principle that the frequency of matter is directly equal to the number of quintessences it contains. Importantly it also mathematically allows relativity and quantum mechanics to be united.

[0048] With $v \approx c$, the modified Dirac equation will yield more accurate results, particularly compared with the Schrodinger equation. We also find that the equation $\lambda=hc/\beta E$ is equivalent to the de Broglie wave equation, $\lambda=h/p$, provided we use the relativistic mass in the de Broglie equation. Given this, these equations yield accurate experimental results.

[0049] Thus we find that the modified formulation of de Broglie wave equation $\lambda=hc/\beta E$ leads directly to a modified Dirac relativistic wave equation and is supported by recent experiments which measure the wavelength of matter and demonstrate the electron orbitals experimentally from these wave equations for matter.

[0050] Wave Particle Duality and Relativity

[0051] From here it is possible to proceed in several ways using the relativistic wave equation. It is apparent that the reintroduction of the term for relative velocity into the wave equations will enable the reintroduction of special relativity into quantum mechanics. In particular we should now be able to derive the term

$$\sqrt{\left(1-\frac{v^2}{c^2}\right)}$$

as a special case of quantum mechanics.

[0052] Thus if:

$$\lambda = hc/\beta E$$

As $E=\sqrt{p^2c^2+m_0^2c^4}$, squaring

$$\lambda^2 = \frac{h^2c^2}{\beta^2\cdot(p^2c^2+m_0^2c^4)}$$

-continued

Conventionally $p^2c^2 = \frac{E^2v^2}{c^2}$

then

$$\lambda^2 = \frac{h^2c^2}{\beta^2 \cdot ((E^2v^2/c^2) + m_0^2c^4)}$$

Thus as $\beta^2 = \frac{v^2}{c^2}$ and $m_0^2c^4 = E_0^2$, then:

$$\beta^4E^2 + \beta^2E_0^2 = \frac{h^2c^2}{\lambda^2}$$

hence $\beta^4E^2 = h^2c^2 \cdot \frac{1}{\lambda^2} - \beta^2m_0^2c^4$ thus

$$\beta^2 = \frac{h^2c^2}{\beta^2E^2} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4}{\beta^2E^2}$$

As $E^2 = m^2c^4$

$$\beta^2 = \frac{h^2c^2}{\beta^2m^2c^4} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4}{\beta^2E^2}$$

Substituting $h = m_qc^2$

$$\beta^2 = \frac{m_q^2c^6}{\beta^2m^2c^4} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4}{\beta^2E^2}$$

As $m_q/'m = 1/n_q$ (eq. 2)

$$\beta^2 = \frac{c^2}{\beta^2n_q^2} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4}{\beta^2E^4}$$

Thus if $f = \beta^2n_q$; (eq. 7a)

$$\beta^2 = \frac{v^2}{f^2} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4f^2}{\beta^2E^2}$$

As $1/\lambda^2 = f^2/v^2$

$$\beta^2 = \frac{v^2}{f^2} \cdot \frac{f^2}{v^2} - \frac{\beta^2m_0^2c^4f^2}{\beta^2E^2}$$

Thus

$$\beta^2 = 1 - \frac{\beta^2m_0^2c^4}{\beta^2E^2}$$

As $E^2 = m^2c^4$

$$\beta^2 = 1 - \frac{m_0^2}{'m^2}$$

Hence

$$m_0/'m = (1-\beta^2)^{1/2}$$

Thus

$$'m = m_0 \Big/ \left[1 - \frac{v^2}{c^2}\right]^{1/2}$$

[0053] Thus this derivation now allows relativity as a universal case of the quintessential wave nature of matter.

[0054] The original premises on which special relativity was based were: that the speed of light is a constant and that all observers are equal. As the speed of light has dimensions of length and time but not apparently of mass, the relativistic change in mass is not accounted for. Using quintessence logically and directly accounts for the relativistic mass changes.

[0055] Moreover, relativity can be derived from the de Broglie equation, and visa versa, directly, thus linking relativity and quantum mechanics by taking into account the existence of quintessence mass.

[0056] Hence, it is now possible to derive the relativistic equations for mass and in turn for space and time from the quintessential wave equation, thus deriving special relativity as a universal case of quantum mechanics and thus uniting special relativity and quantum mechanics. This now allows a further understanding of the nature of space-time.

[0057] The Space-Time Lattice

[0058] The understanding of the true nature of space-time and how it is formulated in three dimensions of real space is crucial. To simply assume that space-time exists, and thence not to question the nature of that existence, denies a deeper understanding of the universe.

[0059] In order to understand the nature of space-time itself, at the quantum level a further look at the nature light and the photon is necessary. Since Einstein's description of light as a particle (the photon) and the description of the photoelectric effect, the standard picture of light as simply a wave can, no longer be applied. If light was to exist as a photon, it could not exist in one dimension, as ordinary waves do, it would need to be three dimensional, with the addition of time. Let us suppose, in this case, that a photon is a three dimensional helical ringlet of light, travelling in the x vector, and spinning around the x-axis. Conventionally this ringlet has a radius; $r=\lambda/2\pi$. The ringlet itself would be vibrating in the y and z vectors. The vectors x, y and z would represent the photon, the substance of which, would be travelling in the x direction and oscillating in the y and z vectors, which would represent oscillatory energy. This in turn would allow it to act as a wave, and create oscillatory electromagnetic fields.

[0060] It is important to re-examine space-time itself in this light, this would have one directional vector with two vector dimensions of energy, one of capacitance and one of electrical permeability, thus accounting for the well known constants of free space; the permittivity of free space (ϵ_0) and the permeability of free space (μ_0) respectively. The vector dimension of direction x, would be the direction of travel and those "quintessences" travelling in an outwardly direction would account for none other than the expansion of the universe. Three of these quintessences would naturally constitute three dimensional visible space-time. These constituents of space-time would interact with the generations of the other vector dimensions reciprocally. Thus one quintessence would sweep out one vector of permeability and one vector of permittivity, through which the other two quintessences could travel, and vica versa, creating a three dimensional space-time lattice.

[0061] The permittivity of free space, (ϵ_0) which is equivalent to capacitance, would as with capacitance plates, be determined by the effective separation between quintessences. The permeability of free space (μ_0) is in fact a force, measured as $4\pi\times10^{-7}$ N/A^2, would result from the force produced by the vibration of quintessence and would be dependent on the density of quintessence. Hence these two parameters would be reciprocal and thus the product of these two would therefore be a constant, which is recognised as none other than the speed of light.

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$$

[0062] This space time lattice would in effect be created by quintessences travelling in all directions with a speed of c within the lattice. The quintessences of the space time lattice would in effect produce a non-static ether. A non-static ether is fully compatible with special and general relativity. Indeed such an ether explains how space time can be curved as in general relativity. Furthermore, the existence of a non-static ether, was espoused by Einstein in his University of Leyden lecture on general relativity of May 5, 1920. In Einstein's own words;

[0063] "According to the general theory of relativity space without ether is unthinkable."

[0064] Recent evidence from a number of sources now strongly support the presence of this non-static ether, in the form of quintessence. An editorial from a major journal states "combined with other observations such as those of distant Supernova, the QMAP results corroborate the prevailing theory of inflation with the twist that the Universe is only one third matter (both ordinary and dark) and two thirds quintessence, a form of energy possibly inherent in empty space".

[0065] If we take into account the existence of quintessence and as such a three dimensional space-time lattice, matter which is intrinsically made of constituents of charge would interact with this lattice to produce the effects of mass. Mass would be perceived as a result of matter (whose constituent particles appear to contain charge) interacting with this lattice directly due to the inhibition of motion by the lattice's electrical permeability and permittivity vectors, which would form the existence of complex space. These quintessences would in the direction in the y and z vectors produce small vibrations of the order of the Planck length (10^{-35} m), whilst passing through the vectors of permeability and permittivity, thus producing the effects of mass.

[0066] The vibration would endow quintessence itself a (non rest) mass m_q equivalent, to the minimal mass of:

$$m_q = h/c^2 = 7\times373\times10^{-51}\ \text{kg}\cdot\text{sec} \tag{1}$$

[0067] The presence and magnitude of Planck's constant (h) and especially the speed of light (c) is thus explained. Indeed, the speed of light

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$$

is not in itself a fundamental quantity.

[0068] As the energy equivalence formula is $E=mc^2$, the minimal mass of a single quintessence, would thus be the minimal mass, h/c^2, hence again:

$$m_q = h(\mu_0\epsilon_0) = h/c^2 = 7.373\times10^{-51}\ \text{kg}\cdot\text{sec} \tag{1b}$$

or

$$m_q c^2 = h \tag{1c}$$

[0069] It is postulated by general relativity that the shape of space time itself can be altered, indeed the presence of the space time lattice now allows this to be altered by altering the density of quintessence. It is further clear that if quintessences underly the structure of the space-time lattice, they may also underly the structure of matter itself.

[0070] With regards a single quintessence, this passing through an energy vector of the space-time lattice would appear as a vibrating string. In a similar way to string theory, the conglomeration of these quintessences would produce the constituents of ordinary matter. Thus the general equation for the number of quintessences (n_q) in an object of mass (m) would be

$$m/m_q = n_q$$

[0071] The mass of the electron (m_e) for example, would be directly determined by the number of quintessences in the electron, multiplied by the mass of quintessence.

[0072] Quintessence and Complex Space

[0073] Quintessence is postulated to constitute the fundamental nature of space-time. Three quintessences each travelling in their respective x vectors at 90° to each other would create three dimensional real space-time. These quintessences would in the direction in their respective y and z vectors produce small vibrations of the order of the Planck length (10^{-35} m), this would create the vector dimensions of permeability and permittivity. The result would give space-time 9 dimensions of space as in superstring theory. However, unlike superstring theory the six hidden dimensions would not be "curled up so as to be so small as to be invisible" these six dimensions would be present in complex space. Thus, only three of these dimensions would represent ordinary three dimensional particulate space time i.e. three dimensional objects. The other six dimensions produced by the vibrations of quintessence would form complex space.

[0074] The mathematics of complex space, using imaginary $\sqrt{-1}$ or (j) numbers, is assumed in the standard formulation of the Shrodinger wave equation. Thus the presence of complex space is an integral part of quantum mechanics.

$$-\frac{h^2}{2m}\cdot\frac{d^2\psi}{dx^2} + V\psi = jh\cdot\frac{d\psi}{dt}$$

[0075] The mathematics of complex space is also an essential and integral part of the principles and application of modern electronic and control engineering. Indeed it has been well recognised for some time that each direction vector in electronic engineering can, be associated with complex vectors.

[0076] As this complex space consists of the vectors of permittivity and permeability it would only be "felt" by charged particles as in the electron. Nevertheless, as all particles are fundamentally composed of charged particles the effects of complex space would be felt by endowing these particles with mass and in turn kinetic energy.

[0077] In conventional complex space, a 2 dimensional Cartesian Argand diagram is mathematically used. However, in order to formulate the equations for particles a three dimensional Argand diagram is essential. This will have three dimensional vectors, one real vector and two imaginary vectors. Three of these diagrams will be required to fully describe the nature of particles, each with a real vector in the x, y and z vectors, respectively. Nevertheless, in the instance below the real vector is the x vector and the two imaginary vectors are given by $({}_jy, {}_jz)$

[0078] The Three Dimensional Argand Diagram

[0079] The beauty of a three dimensional Argand diagram is that the complex conjugate (i.e. the mirror image which confers mathematical reality on the coordinates) is formed by the value of the minus coordinate in the other complex vector dimension. Thus the complex conjugate of $(C_x^{1/2} + {}_jc_y^{1/2} + {}_{-j}c_z^{1/2})$ is $(c_x^{1/2} + {}_jc_y^{1/2} + {}_jc_z^{1/2})$. These two sums when multiplied thus give a real number solution.

[0080] Furthermore it is clear that nine dimensions of space time are necessary in the general relativistic equations. By including complex space we thereby create the nine dimensional spacial metrictensor and the metric energy tensor of matter necessary for computations for general relativity From here we can begin to understand the true structure of matter.

[0081] Energy and the Space-Time Lattice

[0082] The presence of numerous experimental data for quantum tunnelling. and indeed the recent observations by Nicholas Gisin, on the entanglement of distant photons now returns us to EPR experiments.

[0083] Using the quintessential modification of the de Broglie wave equation, gives us an insight into these teleportation and EPR effects.

[0084] As

$$\lambda = hc/\beta E \tag{2}$$

and

$$E = hn_q \tag{1a}$$

then

$$\lambda = c/\beta n_q \tag{2b}$$

[0085] Importantly, as indicated by equation (2b), energy having no quintessence; would have a wavelength of infinity. Specifically pure energy containing no quintessences, would have a lambda of infinity. According to quantum mechanics an infinite wavelength would result in the probability of that energy being anywhere. As energy itself has no electrical charge it would not be impeded by the permittivity and permeability of the three dimensional space-time lattice. Moreover, energy would not be detectable in three dimensional space-time, unless it interacted with matter, as in the EPR experiments. Indeed, energy is not observed when not bound to any form of mass or particle. Thus equation 9d, takes us to our original assertion

[0086] Energy is Not Bound by the Space-Time Lattice

[0087] Thus, as the EPR experiments suggest the existence of energy separate from matter and thus separate from the three dimensional space-time lattice, it is interesting to find that experiment suggests the existence of free energy in a continuum separate from space time to produce the effects of quantum teleportation

[0088] This is not, however, teleportation across an additional dimension, this is a term to describe in partially familiar terms the dissociation of energy from the three dimensional space-time lattice. As time is inextricably linked to each dimension of space, the effects of energy would be inextricably linked to the events, such as the creation of virtual particles, we see interacting within space-time.

[0089] It is unlikely that observers have any direct day to day experience to explain quantum events. Nevertheless, quantum physics may have given us a window into the hitherto hidden workings of the Universe. Thereby, the mystery of the uniformity of the Universe, across distances which the speed of light could not apparently traverse, is readily explained by the fact that the free energy contained in the Universe is not bound by the space-time lattice.

[0090] In the case of light, due to the exceedingly small masses involved, there would be relatively easy exchange of matter with free energy within a photon. This would make the photon the ideal experimental tool to look for energy which is not bound by matter and in turn energy which is not bound in space-time. Indeed, very recently Furusawa et at. have reported to have observed the transference of energy as photons from A to B, without those photons traversing space-time. This finding which has been supported using other experimental techniques, is very important as it suggests the existence of such a quantum continuum.

[0091] We have already seen strong experimental data using photons, atomic spins and other data for quantum teleportation which have recently been published which support these findings. According to the above equations the teleportation would vary in a predictable fashion, as with photons, in line with the wavelength of the light used, relative to the size of vibration of quintessence. As regards matter, the results do confirm that the effect of quantum tunnelling is indeed dependant on the wavelength of matter and the size of that matter.

Part II—Particle Physics

[0092] Electron Structure

[0093] Understanding the electron is fundamental to the understanding of the elementary particles. The hidden nature of the electron may recently have been revealed through observations by Horst Stormer, Daniel Tsui and Robert Laughlin for which a Nobel prize has recently been awarded. They describe a quasi electron particle of charge ⅓e. This has been described on a quantum basis as a vortex of energy, bound as a quasi particle in one dimension x, but not bound in the other two dimensions y and z, allowing dispersion in space-time as a vortex. What is more intriguing are the experimental conditions in which this occurs. First of all a two dimensional electron gas is created and held between two capacitance plates. A magnetic force is then applied in the remaining dimension, virtually creating a one dimensional passage through which only a quasi electron appears to be able to pass.

[0094] Given the presence of charge of ⅓e, then three of these quasi electrons could form an entire electron in three dimensional visible space time. Nevertheless, each would have energy and hence a wave function which would be present in the other vectors. This electron could thus follow the probability functions as described by the Shrodinger wave equation for ψ (otherwise termed as "essence" by Shrodinger)

[0095] If the mass of the electron (m_e) is constituted from quintessence, using the formula:

$$m_e/m_q=n_q$$

[0096] Then an electron would be constituted from:

$$\frac{9.11\times10^{-3}\ \text{kg}}{7.373\times10^{-51}\ \text{kg}\cdot\text{sec}}=1.235\times10^{20}\ \text{quintessences/sec.}$$

[0097] Thus taking into account the mass-energy content of quintessence (m_q) it is independently possible to derive the magnitude of the charge of an electron (e) using the following equation.

$$e=\sqrt{\frac{m_q\varepsilon_0}{\left(\frac{4}{3}\right)\pi hc}}=1.61\times10^{-19}C$$

[0098] This is in close agreement with the experimentally observed charge on the electron of 1.602×10^{-19} C.

[0099] Interestingly substituting $m_q=h/3c^2$ in the above equation we have:

$$e=\sqrt{\frac{\varepsilon_0}{3\left(\frac{4}{3}\pi c^3\right)}} \tag{6}$$

[0100] This can also be written as

$$e=\sqrt{\frac{\varepsilon_0}{3\left(\frac{4}{3}\right)\pi c^3}} \tag{6a}$$

[0101] Equation (6) has a number of very special implications, if re-examined, firstly three of these quasi electron spheres appear to be required to constitute the charge of the electron. More intriguingly, it indicates that the charge is related to the volume of a sphere with an apparent radius of c. Thirdly it indicates that the square of the charge of an electron (e) is proportional to the permittivity of free space (so). The charge given from equation (3) is in close agreement with the measured charge of the electron. Furthermore a more exact value for the charge of the electron (to seven decimal places) can be deduced by taking into account the gravitational field of the Earth (see Gravity and the Charge of the Electron). Furthermore the charge of the electron (e) can now be derived from first principles. Thus, equation (3) corroborates the evidence that the electron is indeed composed of three quasi electrons in keeping with recent experimental findings.

[0102] The significance of the electron, composed of three spheres each with a radius of c, is not immediately clear, but can be understood if the frequency of rotation of the electron is taken into account. Thus if the diameter of the electron was approximately 10^{-19} m, then its spin would need to be

$$\frac{1}{c}\times10^{-15}$$

m approx. eq. 10^6 cycles/sec. Thus given a very high rotation rate an electron could have an effective radius of 1/c and still occupy subatomic sizes. Indeed these observations might be used to estimate the rate of rotation of the quasi electron and its size (see Appendix 1).

[0103] With regards a single quintessence, this passing through an energy vector of the space-time lattice would appear as a vibrating string. In a similar way to string theory, the conglomeration of these quintessences would produce the constituents of ordinary matter. The electron, for example, would be constituted from approximately 1.235×10^{20} quintessences.

[0104] The dimensions of the equation for the electron can be readily resolved by considering each of the three vector dimensions. The exact dimensions of the equation need to be considered in the light of the nature of space-time itself. These dimensional equations help explain the nature of matter. Indeed the equation for the electron may be necessary for the full understanding of gravity

[0105] Complex Space and Electron Structure

[0106] The presence of complex space also now further explains the conformation of the electron, and its formulation at the quantum level, and the presence of particles, anti-particles and their spin up and spin down characteristics.

[0107] Indeed the short form equation for the charge of the electron (−e) can now be rewritten as a metric tensor with three dimensions in real space and six in complex space.

[0108] Thus if three of the x, y and z vectors are in real space and six vectors in complex space, where c is the speed of light in the real space vector, $_jc$ is the speed of light in the complex vector and $_{-j}c$ is the complex conjugate of $_jc$, thus the electron can be mathematically represented by the equation:

$$-e\equiv\varepsilon_{qe}\Big/\left(\frac{4}{3}\pi\right)^{1/2}\cdot\begin{array}{ccc}(c_x)^{1/2}\cdot & (jc_y)^{1/2}\cdot & (-jc_z)^{1/2}\\ + & + & +\\ (-jc_x)^{1/2}\cdot & (c_y)^{1/2}\cdot & (jc_z)^{1/2}\\ + & + & +\\ (jc_x)^{1/2}\cdot & (jc_y)^{1/2}\cdot & (c_z)^{1/2}\end{array}$$

[0109] Which now elegantly gives the real number solution

$$e = \frac{\varepsilon_{qe}}{\sqrt[3]{\left(\frac{4}{3}\right)\pi c^3}}$$

[0110] Where $e = \epsilon_{qe}$ is given as the permittivity of free space for a single quasi electron Equation 4 represents a "complex" tensor

[0111] Whilst the two dimensional Argand diagram has four quadrants, the three dimensional Argand diagram has eight cubic sectors. Two of these cubic sectors are diametric opposites and can represent "real" particulate objects. These have the primary coordinates x, y, −z; as in the electron described above, and the −x, −y, z, with the real vector x now having a minus sign. These two "real" cubic sectors, therefore, mathematically represent particles and their anti-particles.

[0112] The mathematical presence of the two primary diagonal mirror images (x, y, −z and −x, −y, z) now allow the introduction of the concept of antiparticles. This extension of the maths into a three dimensional Argand diagram thus results in the automatic formulation of the maths of antiparticles. Thus the charge of the positron ($^{+}$e) is formulated by the shortened form equation, where the real vectors now each have the minus sign, and therefore exist in the −x, −y, z sector of the three dimensional Argand diagram.

$$e = \varepsilon_{qe} \Big/ \left(\frac{4}{3}\pi\right)^{1/2} \cdot \begin{matrix} -c_x^{1/2} \cdot {}_{-j}c_y^{1/2} \cdot {}_{j}c_z^{1/2} \\ {}_{j}c_x^{1/2} \cdot -c^{1/2y} \cdot {}_{-j}c_z^{1/2} \\ {}_{-j}c_x^{1/2} \cdot jc_y^{1/2} \cdot -c_z^{1/2} \end{matrix}$$

[0113] The three dimensional Argand diagram also accounts for chirality and indeed the up and down spin of the electron. There are two other "real" primary coordinates in the Argand diagram, these are themselves the partial mirror images of the above coordinates (i.e. x, −y, z and −x, y, −z). In particular the y axis is of the opposite sign, thus in particles the y axis is in the downward direction, to form down spin particles and in anti-particles in the up direction, to form the antiparticle The up spin electron is given by eq. 8 and hence the down spin electron (−e|↓≦) is given by the equation

$$-e|\downarrow\rangle = \varepsilon_{qe} \Big/ \left(\frac{4}{3}\pi\right)^{1/2} \cdot \begin{matrix} c_x^{1/2} \cdot {}_{-j}c_y^{1/2} \cdot {}_{j}c_z^{1/2} \\ {}_{j}c_x^{1/2} \cdot c^{1/2y} \cdot {}_{-j}c_z^{1/2} \\ {}_{-j}c_x^{1/2} \cdot jc_y^{1/2} \cdot -c_z^{1/2} \end{matrix} \tag{9}$$

[0114] Thus the three dimensional Argand diagram accounts directly for the presence of antiparticles and the spin up and spin down particles seen in nature. It also accounts for the necessity of the electron to form a square root spherical object, as complex space depends on √−1, otherwise known as j.

[0115] Electron Pairing and Superconductivity

[0116] As the quintessences making up the electron are in a square root conformation, each of these quasi electrons would have a tendancy to pair to form an entire sphere.

[0117] The square root sphere structure of electrons with up and down spins can now superimpose to produce a complete sphere of varying extents. This produces electron pairing as seen at the atomic and molecular levels. It also accounts for the Pauli exclusion principle. This pairing thus accounts for the reactivity of the valence electrons and the electron probability densities, which in turn accounts for the existence of chemistry.

[0118] Furthermore, it is possible to account directly for superconductivity from first principles. For if both the complex and real vectors of the electron combine completely, the product of an up and down spin electron form a perfect superimposed sphere with radius c, with a charge of 2.59×10^{-38} C, denoted by the formula:

$$e^2 = \frac{\varepsilon o}{3\left(\frac{4}{3}\pi c^3\right)} = 2.59\times10^{-38}C \tag{6b}$$

[0119] As with standard superconducting theory, superconductivity can be explained by the formation of "Cooper" electron pairs, where the electrons are forced to pair by the presence of positive crystal charge in particular formation, at supercooled temperatures. In addition the electron pair now forms a stable entity whose angular momentum cancels.

[0120] It additionally becomes clear that the charge of two separate electrons (2e) is 3.2×10^{-19} C, but the charge of the combined electrons (e^2) is 2.59×10^{-38} C. This electron pair thus appears to have 19 orders of magnitude less charge than the electron and in turn 19 orders of magnitude less resistance. It is this effective reduction in charge and in turn resistance, which may account for superconductivity. When observed directly any electrical interaction with the Cooper electron pair will, however, result in the release of the full charge of both electrons, so that the full electrical charge put in will be equal to that coming out of the apparatus.

[0121] The Fine Structure Constant

[0122] Intruiginty from our knowledge of the electron we can further define the term .alpha., the fine structure constant; from the structure of the electron. Thus as the standard term

$$\alpha = \frac{e^2}{hc\times 47\pi\varepsilon_0};$$

substituting the term

$$e^2 = \frac{\varepsilon_0}{3(4/3)\pi c^3} \tag{eq. 6}$$

and

$$h = m_q c^2 \tag{eq. 1}$$

we find:

$$\frac{2\pi}{\alpha} = m_q[3(4/3\pi c^3)]^2$$

or

$$\frac{2\pi}{\alpha} = \frac{m_q e^4}{\varepsilon_0^2}$$

[0123] For brevity we may represent the quasi electron structure as $(4/3)\pi c^3=\theta$; to signify its threefold symmetry, thus

$$\frac{2\pi}{\alpha} = m_q(3\theta)^2 \tag{10}$$

[0124] Indicating that the fine structure constant of the electron (a) is indeed related to its dimensional structure. Again taking into account the effects of gravity the fine structure constant can be derived from first principles to nine decimal places (see Gravity and the Charge of the Electron).

[0125] Fundamental Forces and Particle Structure

[0126] In order to understand the fundamental forces and the nature of fundamental particles, an overview is required. Thus, there are three major forces; strong, electro-weak and gravity, each mediated by three force particles the gluon, photon and graviton respectively. These in turn, influence three types of particle, the quark, lepton, and by general relativity space-time itself. Each of these are composed of particles with multiples of charge of ⅓, which are themselves in three generations, and are present in three dimensions of real space. It is important that a comprehensive view of nature explains this threefold symmetry.

[0127] Using the Standard Model of particles, it is well accepted there exist quark particle charges of −⅓, −⅔ and +⅓ and +⅔ in quarks and anti-quarks. Given that each particle is made up of three quarks the presence of these fractional charges support the association of the fractional charges in this way to form three dimensional charged particles. In stable particles each of the three quarks would have a vector in one dimension, giving the three quarks together an existence in three dimensional visible space time. The particles that bind the quarks (gluons) are themselves required, in stable particles, to have three different color charges, one color in each dimension, for the particle to exist in three dimensional space-time. Furthermore, there are three generations of quarks (and indeed leptons).

[0128] The Standard Model (or a modification of this) and in particular the observation of quarks and indeed quasi electrons with fractional charge of ⅓ and ⅔ in both cases, indicates that particles are constituted from the equivalent of three of these quasi particles to form an electron and quarks to form baryons. In the normal three dimensions the energy would be carried by the particle, However, because each particle is constituted of three quasi particles and in each quasi particle or quark one visible dimension would be the direction vector, in the other two hidden dimensions of each vector the waves would carry energy. Thus each particle would be associated with vibration, which would account for wave particle duality and Heisenberg's uncertainty principle in three dimensional visible space-time.

[0129] These observations lead us directly to the previous postulate that the structure of the electron is composed of none other than three (root) spheres, and that this equation for the electron allows the determination of the charge of the electron from first principles, thus:

$$e = \frac{\varepsilon 0^2}{[3(4/3\pi c^3)]} \tag{6}$$

[0130] In addition the mass of the proton (m_p) can be directly calculated from the ratio of the mass (m_e) of the electron, given by the equation:

$$\frac{m_e}{m_p} = 5.45\times10^{-4} = 3\frac{\pi}{\sqrt{c}} \tag{11}$$

[0131] Strictly we should write,

$$\frac{m_e}{m_p + m_e} = 3\pi\sqrt{c};$$

which is much more elegant.

[0132] Which now gives

$$\frac{m_e}{m_p} = \frac{1}{\sqrt{c}\,/3\pi - 1} = 5.4462\times10^{-3}$$

[0133] This is in very close agreement with the experimentally derived ratio of the proton to electron masses which is also; 5.4462×10^{-3}

[0134] Thus the correlation factor between theory and experiment has a maximum error <0.00001.

[0135] If we combine equation 3:

$$e = \sqrt{\frac{\varepsilon_0}{3(4/3)\pi c^3}}$$

and equation 13:

$$\frac{m_e}{m_p} = 3\frac{\pi}{\sqrt{c}}$$

the positive charge of the proton (e_p) is given by:

$$e_p = \sqrt{\frac{\varepsilon_0}{3(4/3)\pi c^3}} \times m_e \times 3\frac{\pi/\sqrt{c}}{m_p} = e \tag{12}$$

[0136] The stable nuclear proton conformation can thus be represented by the short form equation:

$$p=3^{+}\sqrt{3(4/3)\pi c^3}\times 3(\pi/\sqrt{c}) \tag{13}$$

[0137] This forms a stable 3×3 conformation as with the stable electron structure.

[0138] Importantly the term ($\pi/\sqrt{c}$) is the 90° solution to the Shrodinger wave equation for an electron confined in a space with radius c!.

[0139] Thus the standard equation for an electron confined in a one dimensional box is given by:

$$E\psi(x) = -\frac{h^2}{2m}\frac{d^2\psi(x)}{dx^2}$$

[0140] If the one dimensional box has a length 2L the quantum amplitude (A) can only be non zero between x=0 and X=2L and the standard solution for the amplitude is none other than:

$$A=(1/L^{1/2})$$

[0141] Thus in one dimension the standard solution to the Shrodinger wave equation is:

$$\psi(x)=(\pi/L^{1/2})\mathrm{Sin}\ x/L$$

[0142] Thus not only is the electron charge derived from the equation for three spheres each with a radius of c (eq. 3); but the proton mass and charge can also be derived from the standard solution to the Shrodinger wave equation for a an electron confined in a space of radius c!.

[0143] The term ($\pi/c^{1/2}$) itself would thus most logically represent the gluon which is present in the proton. These gluons would bind the quasi electrons together to form the fundamental particles

[0144] The masses of all the known particles, including the up and down quarks, the W boson, the muon, charm, strange, the tauon, truth and beauty can thus also be derived from first principles in this fashion, and have the quasi electron as their basic constituent particle (see Appendix 1).

[0145] Thus the structure of the muon (μ) can also be derived from the ratio of the mass of the electron (m_e) and the mass of the muon (m_μ):

$$m_e/m_\mu=4.7\times10^{-3}=c^{1/3}$$

Thus

$$\mu=\epsilon_0{}^{1/2}\times m_e/m_\mu\times 3^{-}(4/3=\pi c^3)^{1/2}\times(\pi/c^{1/3})$$

[0146] Where the charge of the muon is in this equation equivalent to that of the electron e. In this case ($\pi/c^{1/3}$) can be considered to represent a specific high energy photon. Thus the structure of the muon, written in short form is:

$$\mu=3^{-}(4/3\pi c^3)^{1/2}\times(\pi/c^{1/3}). \tag{14}$$

[0147] Moreover the structure of the tauon can be calculated from the ratio of the mass of the electron and that of the Tauon (1.79 Mev);

[0148] Thus

$$0.511\ \mathrm{Mev}/1.79\ \mathrm{Gev}=2.85\times10^{-4}$$

$$m_e/m_\tau=(\pi/c)^{1/3}\times(\pi/c)^{1/9}=2.85\times10^{-4}$$

[0149] As the charge of the tauon is equivalent to the charge of the electron, hence the structure of the Tauon is given by the above equation

$$e_\tau=\epsilon_{qe}\times(m_e/m_\tau)\times 3^{-}(4/3\pi c^3)^{1/2}\times(\pi/c)^{1/3}\times(\pi/c)^{1/9}=e$$

[0150] This equation accurately predicts the charge −1; and mass of the Tauon (.~1.78 Gev). Thus the structure of the Tauon can in short form be given by the equation

$$\tau=-3(4/3\pi c^3)^{1/2}\times(\pi/c)^{1/3}\times(\pi/c)^{1/9} \tag{15}$$

[0151] Furthermore a more exact value for the mass of the muon and tauon can be deduced by taking into account the gravitational field of the Earth in a similar way to identifying the exact charge of the electron. In addition it may be necessary to take into account a possible mass value of the neutrino to arrive at a precisely accurate mass value of the muon and tauon. Nevertheless, the mathematical proof of these short form equations lies in the fact that they can very closely identify the charge and the masses of these particles, from first principles, as in equations.

[0152] Overall the mathematical geometrical structure of all the particles can be derived from the quasi electron, which is in turn derived from quintessence. Thus, the short form particle structures can now be derived from first principles. This includes the quasi electron (qe) and electron (e), from which the quarks (u,d) and in turn the stable proton (p) and stable neutron (n) and alpha particle (a) respectively are derived. The general structure of the force carrying bosons the photon. (g) and the gluon (γ) and the intermediate vector boson (W) can be given. It will also intriguingly be possible to derive, according to their generation, the structure of the strange (s) charm (c), beauty (b, or bottom) and truth (t or top) quarks directly from the structure of the muon (μ) and Tauon (τ) respectively.

[0153] Using the term $\Theta=(4/3\pi c^3)$, where, −/+ represents the charge of the quasi electron, we find:

[0154] 1st Generation:

$$(q_e)={}^{-}\Theta^{1/2} \tag{6c}$$

$$e=3^{-1/2} \tag{6}$$

$$d={}^{-}\Theta^{1/2}\cdot 3(\pi/c^{1/2}) \tag{16}$$

$$u=2^{+}\Theta^{1/2}\cdot 3(\pi/c^{1/2}) \tag{17}$$

$$s={}^{-}\Theta^{1/2}\cdot 3(\pi/c^{1/2}) \tag{18}$$

[0155] 2nd Generation

$$\mu=3^{-}\Theta^{1/2}\cdot(\pi/c^{1/3}) \tag{14}$$

$$c=2^{+}\Theta^{1/2}\cdot(\pi/c^{1/3})\cdot(\pi/c^{1/4}) \tag{19}$$

$$b={}^{-}\Theta^{1/2}\cdot(\pi/c^{1/9})\cdot(\pi/c)^{1/4} \tag{20}$$

[0156] 3rd Generation

$$\tau=3^{-}\Theta^{1/2}\cdot(\pi/c)^{1/3}\cdot(\pi/c)^{1/9} \tag{15}$$

$$t=2^{+}\Theta^{1/2}\cdot(\pi/c)^{1/3}\cdot(\pi/c)^{1/9}\cdot(\pi/c)^{1/4} \tag{21}$$

[0157] Particle Gluons (g):

$$g_1=(\pi/c^{1/2}) \tag{22}$$

$$g_2=(\pi/c)^{1/2} \tag{22a}$$

$$g_3=(\pi/c^{1/4}) \tag{22b}$$

$$g_4=(\pi/c)^{1/4} \tag{22c}$$

[0158] Particle Photons (.gamma.):

$$\gamma_1=(\pi/c^{1/3}) \tag{23}$$

$$\gamma_2=(\pi/c)^{1/3} \tag{23a}$$

$$\gamma_3=(\pi/c^{1/9}) \tag{23b}$$

$$\gamma_4=(\pi/c)^{1/9} \tag{23c}$$

[0159] Intermediate Vector Boson ($W^{+/-}$):

$$W^+=3^+\Theta^{1/2}\cdot 2(\pi/c^{1/6})^6 \tag{24}$$

$$W^-=3^-\Theta^{1/2}\cdot 2(\pi/c^{1/6})^6 \tag{25}$$

[0160] Stable* Proton:

$$p=3^+\Theta^{1/2}\cdot 3(\pi/c^{1/2}) \tag{13}$$

[0161] Stable* Neutron:

Stable * Neutron:

$$n = \begin{matrix} {}^+\Theta\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot{}^-\Theta\uparrow^{1/2} \\ {}^+\Theta\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot{}^-\Theta\uparrow^{1/2} \\ {}^+\Theta\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot{}^-\Theta\uparrow^{1/2} \end{matrix}$$

* Stable nucleonic neutron and proton conformations differ slightly from the Standard Model, this is due to the sharing of quasi electron and quasi positron particles within the nucleus, which allows stabalisation of these particles by the formation of stable 3×3 structures. The Standard conformations which describe non-nucleonic neutrons and protons are additionally given in Appendix 1.

[0162] Alpha particle (α):

$$\text{Alpha particle }(\alpha) := \alpha = \begin{matrix} [3^+\Theta]\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot[3^-\Theta]\uparrow^{1/2} \\ [3^+\Theta]\uparrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot[3^-\Theta]\downarrow^{1/2} \\ [3^+\Theta]\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot[3^-\Theta]\uparrow^{1/2} \end{matrix} \tag{27}$$

[0163] The mathematical proof for these structures and their decay mechanisms is lengthy and is thus fully contained in Appendix 1. All the particle structures are accurately mathematically defined by the masses of these particles.

[0164] The structure of these particles all contain the quasi electron and thus the metric tensor structure necessary in the formulation of the gravitational equations is sustained. The respective forces created by the gluon and the photon are important as they tell us the behaviour of matter and also lead to the likely structure of the graviton

[0165] Particle Spin and Size

[0166] The significance of the electron, composed of three spheres each with a radius of 1/c, is not immediately clear, but can be understood if the frequency of rotation of the electron is also taken into account. Knowing the structure of the electron has led us to deduce its charge and thus may lead us estimate its size and spin. Thus these observations might be used to calculate the radius and rate of rotation of the electron.

[0167] Let us suppose, that nature is truly beautiful, and that the radius of the fundamental quasi electron is indeed 1/c, and in turn the radius was balanced by the velocity of rotation 2π/c. This can be directly confirmed mathematically by taking into account the known spin of the electron, h/4π. Thus the actual spin of the electron may be calculated form the known energy of the spin.

[0168] The radius of the electron is not up till now known, but the radius of a quark has been estimated, and this is the radius derived from deep inelastic collisions of the proton. These estimates reveal a radius of approx. $r_p=1.18\times10^{-15}$ m This value may be used to assist in confirming the spin of the proton in revolutions per sec. (revs) and in turn the spin and size of the electron. Firstly we may proceed to estimate the spin of the proton. Thus as h=E·t (Joules×sec) and h=E·t=F·d·t (Joules×sec), then the spin;

$$h/4\pi=F\cdot d\cdot t \tag{28}$$

[0169] As F=ma, where $\alpha=(\text{revs}\cdot 2\pi)^2 r_p$ and m=the mass of the proton, then

$$h/4\pi=m(revs\cdot 2\pi)^2 r_p\cdot d\cdot t$$

[0170] The actual distance (d) traveled in a circle of half integer spin in 1 second is: $\text{revs}\cdot\pi r_p$ thus:

$$h/4=m(revs\cdot 2\pi)^3 r_p^{\ 2}/2$$

[0171] Hence:

$$revs=[h/m(2\pi)^4 r_p^{\ 2}]^{1/3}$$

[0172] Taking the effective mass the proton as 1.6726×10^{-27} kg, then the rate of spin of the proton in revolutions/sec is:

$$revs=5.65\times10^6 \text{ cycles/sec}$$

[0173] From the frequency of the specific rotation of the proton, given the half integer spin associated with the proton, we can thus mathematically confirm the relationship between the radius of a particle and its spin:

$$r_p\times revs/2=1/c \tag{29}$$

[0174] Furthermore, the fundamental radius of 1/c seen in geometric structure the quasi electron, is also reflected in the rotation rate and radius for the proton, thus as above 1/c+½ revs=1.85×10^{-15} m. Moreover, this means the actual half integer velocity of rotation is none other than 2π/c in metres/sec. So that the particle is in harmonic balance.

[0175] Using the fundamental formula h/4π=F·d·t, it is possible to obtain accurate estimates of the radius and spin rates of the electron, or indeed any particle, using the same principle of harmonic balance. Using the formula:

$$revs=[h/m(2\pi)^4 r_e^{\ 2}]^{1/3}$$

[0176] It appears there are two unknowns, the radius if the electron and its revolution rate, however, in accordance with the equation, r_p=2/c.revs, which gives the revolution rate of the proton, the same principle may also be used for the electron, by substituting r_e=2/c.revs, such that:

$$revs=hc^2/4m_e(2\pi)^4 \tag{30}$$

[0177] Taking the mass of the electron 9.109382×10^{-31} kg, the rate of revolution of the electron is:

$$revs=1.048\times10^{10} \text{ cycles/sec}$$

[0178] Which gives a predicted radius of the electron as

$$r_e=6.336\times10^{-19}\text{ cycles/sec}$$

[0179] So the half integer rotation velocity (revs.πr) is 2π/c !, for the electron in keeping with the harmonic balance of the electron.

[0180] The same principle may be used to obtain an accurate estimate of the spin and radius of the muon, or any other particle. Using the above formula

$$revs=hc^2/4m_e(2\pi)^4 \tag{31}$$

[0181] Then as the mass of the muon is 1.8823×10^{-28} kg then the revs of the muon $f_\mu=5.070\times10^{-7}$ cycles/sec and the radius r_μ is thus 1.316×10^{-16} m.

[0182] It is now possible to begin to explain how the muon and the other subatomic particles are formed. If a quasi electron is complexed with another structure the total geometric structure needs to maintain harmonic balance. So the frequency of rotation would need to match geometric structure with which the quasi electron was complexed

[0183] Intriguingly we find asymptotic convergence for the formulas for frequency and mass occurs, when the geometric structure complexed with the quasi electron has the structure represented by $(\pi/c^{1/3})$ [giving the frequency divided by two, because the single integer spin of the force carrying particles compares to a half integer spin for the muon]. So that

$$(f_1 3\pi/2)^{1/3}=f_\mu$$

[0184] When the ratio of the masses of the electron (m_e) and muon (m_μ) are related, such that:

$$m_e(3\pi/c^{1/3})=m_\mu$$

[0185] Indeed we find that (allowing for the neutrino) this ratio is very close to the actual ratio of the mass of the electron to the mass of the muon, determined experimentally.

[0186] Furthermore, we have seen that these geometric structures, representing harmonics of the speed of light, which either match the frequency or the amplitude of vibration of the quasi electron, mathematically define the masses of the particles and the fundamental forces of Nature.

Part III—Quantum Gravity

[0187] Quantum General Relativity

[0188] Given the overall energy "complex" energy tensor structure of the electron and the metric tensor, assumed in general relativity, the quantum nature of gravity itself can now be explored. The spherical complex tensor for the electron and the positron give the mathematical quantum structure and energy tensor for all the other particles. Together with the time dimension these nine space dimensions account for the 10 parameters present in the metric tensor necessary to formulate the equations for gravity using Riemann geometry and thus forms the basis of quantum gravity. Intriguingly the metric tensor at each point in space time is required to consist of a collection of ten numbers, Consequently, ten dimensional space-time hypotheses, such as this or superstring theory, do automatically yield general relativity.

[0189] Furthermore, the mathematical representation of the graviton and the gravitational constant may be directly estimated from the knowledge of the mass and radius of quintessence. Thence, the force of the vibrations of quintessence lead directly to quantum gravity.

[0190] The radius of quintessence should be approximately in keeping with the Planck length estimate (r), which is conventionally derived from the standard dimensional equation:

$$r_q^2=Gh/c^3 \tag{32}$$

[0191] Given the nine spacial parameters present in the metric tensor, used in general relativity we find that the actual formula for r_q^2 is mathematically in agreement with theory when:

$$9r_q^2=Gh/c^3 \tag{33}$$

[0192] This again supports the 9 dimensional view of space and the size of the vibrations of quintessence can thus be estimated.

$$r_q=1.35\times10^{-35}\text{ m} \tag{33a}$$

[0193] This value is in agreement with the Planck length. Indeed if the above equation is correct then we find that we can derive the standard equation for the general relativistic increase in radius, r', (eq. 34) directly from first principles and arrive at a more fundamental equation for quantum gravity. As

$$r'=G\cdot M/3c^3 \tag{34}$$

[0194] By substituting eq. 33) into equation 34, a fundamental relationship between r' and M is obtained.

$$r'/3r_q^2=GMc^3/Ghc^3=Mc/h$$

[0195] And substituting the quintessential equation, $h=m^{q}c^2$ (eq. 1) then:

$$r'/3r_q^2=M/m_qc=n_q/c \tag{35}$$

[0196] Hence the ratio of the change in radius to that of the radius of quintessence squared, is proportional, by a factor of c, to the ratio of the mass M of an object to that of the mass of quintessence, effectively the number of quintessences. Thus the change in radius, r' due to gravitation, is related to none other than the ratio of the mass and radius of an object to the mass and the square of the radius of quintessence. Thus again the gravitational change in radius is directly related to the number of quintessences.

[0197] Naturally, this would be exactly what would be logically expected if quintessence, like the equation for the charge of the electron (eq. 6) forms from a root sphere. Thus the change in spacial radius of a normal sphere is dependant on the square of the quintessential radius.

[0198] This increase in apparent radius represents none other than the (gravitational) binding energy for quintessence.

[0199] The meaning of the above dimensional equation (33) might itself be further understood by substituting the mass of quintessence (where $m_q=h/c^2$) into the equation. Thus in nine dimensions the gravitational constant (G) may be more logically given as,

$$9(\pi r_q^2/m_q)=G\pi/c \tag{36}$$

[0200] Where πr_q^2 is the cross sectional area of quintessence and m_q is the effective mass of quintessence, and thus

($\pi r_q^2/m_q$) represents the effective mass per unit area which quintessence exerts. This equation reduces to:

$$9r_q^2/m_q=G/c \tag{37}$$

[0201] From this we may derive the standard general relativistic relationship for the apparent change in radius (r') around a mass (M), from an understanding of the mass m_q and number (n_q) of quintessences. As $m_q=M/n_q$, then:

$$3r_q^2=G\cdot M/3c\cdot n_q \tag{38}$$

[0202] Then if

$$n_q=r'c/3r_q^2 \tag{39}$$

[0203] thus directly substituting for n_q in eq. 38:

$$r'=G\cdot M/3c^2 \tag{34}$$

[0204] The importance of this is that the gravitational change in radius now logically derives from equation 36, which describes the gravitational force as resulting directly from the mass of quintessence exerted/per unit area of quintessence.

$$9(\pi r_q^2/m_q)=G\cdot\pi/c \tag{36}$$

[0205] Thus equation 34 is the conventional equation for the general relativistic increase in radius (r') in a gravitational field, which is here derived from the underlying nature of quintessence. Thus the gravitational constant is derived from the mass and radius of vibration squared of quintessence from first principles.

[0206] Indeed it is apparent that a more fundamental equation for gravitation now exists, for equation (39) is mathematically accurate and numerically agrees with eq. 34:

$$r'/3r_q^2=n_q/c \tag{39}$$

[0207] These equations may be readily mathematically verified. If in accordance with standard general relativity, the apparent increase in radius r' is:

$$r'=GM/3c^2 \tag{34}$$

[0208] Then given that the mass of the Earth is 5.9745×10^{24} kg;

$$r'=1.478\times10^{-3}\text{ m}$$

[0209] Accordingly if $r'=3r_q^2\cdot n_q/c$; (eq. 39). Given the number of quintessences n_q constituting the Earth is M_E/m_q, then

$$n_q=5.9745\times10^{24}/7.3725\times10^{-5}=8.104\times10^{74}$$

[0210] As $r_q^2=1.823\times10^{-70}$ (eq. 33a) then:

$$r'=1.478\times10^{-3}\text{ m}$$

[0211] Thus equation 39 gives the same answer as the standard equation and may be understood on a logical basis. Indeed the meaning of c in the equation may be understood as it has been previously shown as being the basis for the radius of matter (eq. 6). Hence the general relativistic change in radius, r', is none other than the effective binding energy for quintessence.

[0212] Quantum Gravity and Wave Particle Duality

[0213] Quantum gravity can now be readily linked with quantum mechanics, indeed any observations which are self consistent must be able to do so easily.

[0214] The frequency of light has been previously derived

$$f=E/h=n_q$$

[0215] Thus the formula for the frequency of light (E=hf) has previously been explained theoretically by the simple observation that the frequency is determined quite directly from the number of quintessences (n_q) within the photon. The same principle has also been shown to apply to matter.

[0216] Let us now follow these equations for matter by calculating the wavelength of a photon from the Gravitational constant as an example; and also as a test of these observations and to demonstrate that the gravitational equations can also apply to the quantum world.

[0217] If

$$n_q=r'c/3r_q^2 \tag{39a}$$

[0218] where r' is the general relativistic increase in radius, and r_q is the radius of quintessence (eq. 33). Where $f=E/h=n_q$, substituting for n_q, then the frequency of the photon f_γ(where β=1) is given by:

$$f_\gamma=r'c/3r_q^2$$

[0219] Using the standard equation, $r'=GM/3c^2$ (eq. 34); we may substitute for r', thus we have:

$$f_\gamma=GM/9r_q^2c\text{ 15 Thus}$$

$$f_\gamma=\frac{G}{9r_q^2c^3}\cdot m_\gamma c^2\text{ and as }E=m_\gamma c^2;$$

$$f_\gamma=\frac{GE}{9r_q^2c^3} \tag{40}$$

Indeed as $9r_q^2=Gh/c^3$, then $f_\gamma=E/h=n_q$

[0220] It is possible to also demonstrate that the same relationship holds for the wave equation for matter. If we take the relativistic wave energy of matter, which has been previously derived,

$$f=\beta^2n_q$$

[0221] This includes the term for the number of quintessences flowing through the electron, in the complex vectors of space-time, to give the relativistic electron momentum (p) and a term for the rest mass, thus substituting into (40)

$$f_\gamma=\frac{GE}{9r_q^2c^3} \tag{40}$$

[0222] As $f=\beta^2n_q$ for matter then the equation expands to:

$$f_m=\frac{G}{9r_q^2c^3}\cdot\beta^2E\text{ As }\lambda=v/f\text{, then}$$

$$\lambda=\frac{9r_q^2c^3v}{G\beta^2E} \tag{41}$$

[0223] Then the equation again reduces to:

$$G = \frac{9r_q^2c^4}{\lambda\beta E} \tag{3}$$

[0224] Equations 3, 40 and 41 are important as they show that the quantum wavelength of any particle of rest mass m can be derived from the gravitational constant G. Thus linking quantum mechanics to quantum gravity.

[0225] It is therefore important to confirm the numerical accuracy of the above equation (40). We can do this by comparing the result to the standard computation of the de Broglie equation, in a range where de Broglie itself is likely to be most accurate; which according to these observations is in the low energy range (see section on Wave Particle Duality).

[0226] If we take an electron with an energy of 0.1 KeV the wavelength is conventionally given (where the kinetic energy of the electron E_k is given by the product of the charge of the electron (C) and the potential applied eV=0.1 Kev), by the standard equation:

$$\lambda = h/p = h/(E_k \cdot 2m_0)^{1/2} \text{ thus}$$

$$\lambda = 6.63\times10^{-34}/[1.602\times10^{-19}\times1\times10^2\times18.22\times10^{-31}]^{1/2}$$

$$\text{hence}$$

$$\lambda \cong 1.23\times10^{-10}m \text{ Using}$$

$$\lambda = \frac{9r_q^2c^4}{G\beta E} \text{ Where } E = \gamma m_0c^2 \tag{3a}$$

[0227] At 0.1 Kev, electron velocity is 6×10^6 r/sec, thus $\beta=2\times10^{-2}$ and $\gamma=1/(1-v^2/c^2)^{1/2}=1.0002$. Thus:

$$\gamma = \frac{9\times1.82\times10^{-70}\times80.78\times10^{32}}{6.76\times10^{-11}\times\beta\times1.0002\times9.11\times10^{-31}\times8.998\times10^{16}}$$

$$\lambda = 1.21\times10^{-10}m$$

[0228] Divergence between the de Broglie equation and the above equation (2) occurs at intermediate and high energies where it is generally accepted that the standard de Broglie equation may be less accurate. The values for eq. 2 and de Broglie are compared to recent experiments, which demonstrate a relativistic curvilinear plot for wavelengths of matter in keeping with eq. 40.

[0229] The de Broglie equation in the non-relativistic format yields a simple log/linear scale, which is not in keeping with relativity; whereas eq. 3 is dependent on relativity and mathematically accounts for both relativity in calculating the wavelength. Indeed recent experiment on quantum tunnelling through a wire mesh strongly suggests that the relationship between energy and wavelength is relativistically curvilinear$^{(ref\ 1)}$. Furthermore equation 3a suggests a fundamental relationship between energy (E), relative velocity (v/c=β), gravity (G) and the quantum wavelength (λ)

$$\lambda = \frac{9r_q^2c^4}{G\beta E} \text{ Indeed as } {}^9r_q^2 = Gh/c^3 \text{, then} \tag{3a}$$

$$\lambda = hc/\beta E \tag{2}$$

[0230] Equation 2 is the very same as the Universal wave equation derived form first principles for the wavelength of tight and matter, which allowed a relativistic solution to the equations for wave particle duality (see Wave Particle Duality). This now indicates that these quintessential equations are compatible with relativity, quantum mechanics and quantum gravity.

[0231] Graviton Structure

[0232] From these observations, if the value for the gravitational constant is substituted into the equation (35) we may now estimate the probable geometric structure of the gravitation, which is the force particle mediating gravity by acting on quintessence. Thus the Gravitational constant has been previously derived from the vibration of quintessence by the equation:

$$G\cdot(\pi/c)=9(\pi r_q^2/m_q) \tag{36}$$

[0233] This is in accurate agreement with the value for G(6.67×10^{-11} N m^2 kg^{-2}). This suggests that the most probable mathematical representation of the graviton (ϕ), the third force carrying particle is

$$\phi=(\pi/c) \tag{42}$$

[0234] Thus the gravitational constant (G) can be given by the mass and radius of quintessence and the structure of the graviton

$$G=9\pi r_q^2/\phi m_q \tag{43}$$

[0235] This shows the gravitational force to be related to the fundamental radius of quintessence space time, and the graviton.

[0236] Quantised General Relativity

[0237] The classical general relativistic formula, as given by Einstein is:

$$R_{\mu\nu}-1/2g_{\mu\nu}R=-\kappa T_{\mu\nu}$$

[0238] Where R is effectively the curvature of space-time, $R_{\mu\nu}$ denotes the contracted Riemann tensor of curvature and $T_{\mu\nu}$ is the "energy tensor" of matter.)

[0239] If we substitute the energy tensor matrix of the electron (eq. 9)×time, for the energy tensor of matter $T_{\mu\nu}$; and the metric tensor of the space-time lattice×time for the contracted Riemann tensor we can arrive at the same solutions for general relativity.

[0240] Furthermore, in his published paper on General Relativity, Einstein, defined the constant κ as:

$$\kappa=8\pi G/c^2$$

[0241] Therefore Einstein's equation should be written as

$$R_{\mu\nu} - 1/2g_{\mu\nu}R = \frac{8\pi G}{c^2}\cdot T_{\mu\nu} \tag{43}$$

[0242] Einstein himself was apparently not happy about the right hand component of the equation. However, we find that this part of the equation can now be explained and quantised by substituting the gravitational constant, $G=9\pi r_q^2/\phi m_q$, (eq. 42a),

[0243] Giving:

$$R_{\mu\nu}-1/2g_{\mu\nu}R=-\frac{8\pi^2 9r_q^2}{\varphi m_q c^2}\cdot T_{\mu\nu} \tag{44}$$

[0244] By substituting $m_q \cdot c^2=h$, and further substituting h=h/2π, we arrive at a quantised solution to Einstein's equations. Where A_q is the surface area of quintessence ($A_q=4\pi r_q^2$); φ is the graviton [φ=(π/c)] and h is Plancks constant. thus:

$$R_{\mu\nu}-1/2g_{\mu\nu}R=-\frac{9A_q}{\varphi\hbar}\cdot T_{\mu\nu} \tag{45}$$

[0245] The gravitational equation can now be further understood on a logical basis. The term $A_q=(4\pi r_q^2)$, where represents standard term for the surface area of a sphere of quintessence for the 9 space dimensions of the space time lattice, h is the energy content of quintessence×time and φ is the graviton, thus the right hand term now represents a true "metric energy tensor" of matter.

[0246] This leads directly to the standard solution to the field equations, for the general relativistic increase in radius r' of an object, where A is the surface area of a sphere of a given mass M, such that

$$r=\sqrt{(A/4\pi)}-r=GM/3c^2 \tag{34}$$

[0247] Furthermore, although equation 45, gives the same solutions as Einstein's equation, which is essentially correct, the difference is that the equation is now dependant upon Planck's constant (h), and moreover the radius of quintessence, which now defines a quantised solution to the equations.

[0248] Graviton Force Characteristics

[0249] Similar to the photon, the previously derived equation (42) for the graviton [φ=(π/c)] appears to also mathematically represent a helical ringlet of quintessence, but with a spin of 2. For the photon, taking the direction of motion as the x vector and its axis of spin also as the x vector, would account for the electromagnetic force and its attraction and repulsion characteristics. In the case of the gluon component ($\pi/c^{1/2}$), if the direction vector is x, then the axis of spin would be in the y vector, the same as quasi electrons, accounting for the particle binding characteristics of the gluon force. In the case of the graviton, if the direction of motion was in the x vector, the graviton spin axis would be in the z vector thus, as will be demonstrated, accounting for the gravitational force.

[0250] The spin axis of the graviton can also be derived using the known characteristics of the electron. If an electron is travelling in the x direction, then its spin axis is determined by the by the sign of the $_j$y vector (up or down). This view is in agreement with conventional theory, which indicates that the electron spin is similar to a rotating planet orbiting the sun, (the electron even appears to have orbital precession). As the electron passes through the space-time lattice, this spin would generate the formation of gravitons. This would occur as a result of the ejection of the excess quintessence passing through the electron. As the electron spins, the ejection of these gravitons would occur at a tangent to the electron's direction of motion. The ejection of the gravitons would occur, similarly to the ejection of energy of a pulsar or quasar, through the equivalent of the north and south poles of the electron. Thus, propelling the graviton in the direction of the electrons y vector. The ejection of the graviton would re-orientate and impart a specific angular momentum to the gravitons which would thus end up spinning on its own z axis. If for instance the graviton is released from an up spin electron the graviton will be rotating clockwise and its leading edge will displace quintessence downwards. In turn this will provide an upwards force.

[0251] This picture accounts for Fleming's left hand rule, is logical and provides an explanation for the magnetic force around a wire. According, to the left hand rule if the direction of the current is in the x vector, the magnetic field is in the z vector, and the force is upwards, in the y vector, in accordance with the above model. Therefore, this particular spin axis and the structure of the graviton results in its force characteristics. As the graviton is very small compared to the electron and both have different rather rapid spin axis it is difficult for these to bind and interact. Nevertheless, because the graviton has a spin of 2, and as it spin axis is perpendicular to its direction of motion, in the z vector, it readily displaces space-time quintessence to produce gravity. Thus because the graviton is able to displace space-time, it is capable of escaping a black hole. How else could the effects of gravity be felt beyond a black hole?

[0252] Quantum Gravity and Electromagnetism

[0253] With the above electron model of graviton production the nature of magnetism can be understood from first principles, Furthermore, the presence of a space-time lattice links relativity, and the forces of gravity with the electromagnetic and other forces of Nature. Indeed, evidence for these links may first date back to the 1820's, when Andre Ampere first defined the Amp. The force of attraction between two parallel wires 1 metre apart each carrying 1 Amp in a vacuum was defined as none other than the permeability of free space (2×10^{-7} N per metre of conductor). Thus conventionally the magnetic field strength around a long straight wire is given as:

$$B=\mu_0 I/2\pi r$$

[0254] Where I is the current and μ_0 is the permeability of free space ($4\pi\times10^{-7}$ N A^{-2})

[0255] The attraction between two wires both carrying negative charge is, however, counterintuitive as negative charges should repel. A conventional explanation overcomes this by invoking the presence of a magnetic field which is created by the current by the production of virtual photons. Thus we appear to have an explanation for the effects of magnetism which involves virtual photons, however, these photons are not observed. More accurately, according to conventional special relativity the magnetic field is none other than the electric field viewed relativistically.

[0256] A more satisfactory explanation, therefore, lies in the interaction between the electrons and the space time lattice. The moving electrons in the two wires interact with the lattice to produce gravitons; which are in phase when both streams of electrons are traveling in the same direction. The gravitonic waves interact constructively to disperse the space time lattice between the wires and induce an attractive force between the two wires, which produces in effect the permeability of free space. Thus this force results from the vibration of quintessence itself.

[0257] Conversely in two wires with current going in opposite directions the graviton waves are in anti-phase and would interact destructively between the wires. The gravitonic waves traveling radially outward from the wires would, however, disperse the lattice outside the two wires and produce apparent repulsion between the wires, which is exactly what is observed. These effects of electricity suggest that gravitons act as waves and that phase is important.

[0258] This effect is also seen with the north and south poles of ferromagnets. Nevertheless, with matter other than iron, cobalt or nickel, the graviton emission cannot be phased as the atoms are unable to align and magnets do not appear to exist with other materials.

[0259] In ordinary magnetic system the release of gravitons from the north pole would be exactly balanced by those released from the south pole of the magnet and hence there would be no net force on the magnet until an external magnet or electrical current were applied.

[0260] Overall the magnitude of the forces in electrical systems where electrical conduction occurs are well defined by the permeability and permittivity of free space μ_0, and ϵ_0. Where v is the constant velocity of the charge and ϵ is the electric field produced by the charge.

$$B=[\mu_0\epsilon_0]v\cdot\epsilon$$

[0261] These observations suggest that the forces of electricity which produce magnetism are indeed related to the permittivity and permeability of free space and that these quantities are exerted by an apparent vacuum. Thus the effects of magnetism could be explained by none other than the phased effects of gravitational waves on the space time lattice.

[0262] Electromagnetism is of further interest to quantum gravity, particularly if we combine the standard equations, $B=\mu_0 I/2\pi r$ and $B=[\mu_0\epsilon_0]v\cdot\epsilon$, substituting for B we have:

$$2\pi r=I/\epsilon_0 v\epsilon \tag{46}$$

[0263] Thus $2\pi r$ is proportional to the inverse of ϵ_0. Thus as space time is dispersed by gravitons the permittivity field will increase in the same way capacitance increases with separation of plates. Because of the inverse relationship between ϵ_0 and $2\pi r$, as ϵ_0 increases the circumference of a circle and the apparent ratio of π is to r, will appear to diminish in accordance with general relativity. This not an actual diminution in the circumference of a circle but the effective reduction of the resistance to motion in a circular path in this field.

[0264] Incidentally, the above observations, also lead us directly to Schrödinger's formula for the average equilibrium distance (r) between an electron with charge (e) in orbit around a proton, which is conventionally given by:

$$r=n^2 4\pi h^2\epsilon_0/me^2$$

[0265] Where ϵ_0 is again the permittivity of free space, m is the mass of the electron and n is an orbital integer, h is Planck's constant and e is the charge of the electron. Furthermore if $e=[\epsilon_0/3(4/3\pi c^3)]^{1/2}$ (eq. 3); then the equation at n=1, for the electron orbital radius elegantly simplifies to:

$$r=4h^2c^3/m$$

[0266] Hence the orbital radius of the electron is related to spin of the electron (h) and its mass (m).

[0267] Quantum Gravity and the Charge of the Electron

[0268] The equation for the charge of the electron (eq. 1) contains the term so (permittivity of free space) which according to these observations should vary in a gravitational field.

$$e=[\epsilon_0/3(4/3\pi c^3)]^{1/2} \tag{6a}$$

[0269] If we combine the standard equations, $B=\mu_0 I/2\pi r$ and $B=[\mu_0\epsilon_0]v\cdot\epsilon$, substituting for B we have:

$$2\pi r=I/\epsilon_0 v\epsilon \tag{46}$$

[0270] Thus $2\pi r$ is proportional to the inverse of ϵ_0. Thus as space time is dispersed by gravitons the permittivity field will increase in the same way capacitance increases with separation of plates. Because of the inverse relationship between ϵ_0 and $2\pi r$, as ϵ_0 increases the circumference of a circle and the apparent ratio of π to r, will appear to diminish in accordance with general relativity.

[0271] Thus ϵ_0 rises when space-time is dispersed by the gravitons that produce the gravitational field, This occurs in a similar way to the process by which capacitance increases with separation of plates in a capacitor.

[0272] Nevertheless, as c is a constant and as $c=[\mu_0\epsilon_0]^{-1/2}$, then if ϵ_0 rises then μ_0 falls. This is entirely consistent as μ_0, which represents the force that quintessence exerts, would be reduced if the quintessence space time lattice is dispersed.

[0273] Furthermore, as $\mu_0=4\pi\times10^{-7}$ N A^{-2}; then as μ_0 falls, then the apparent ration π to r, also falls in a gravitational field. This is largely the same as stating, as does general relativity, that the apparent radius r', rises in a gravitational field. So this view is consistent with general relativity.

[0274] Nevertheless, to derive an exact value for the charge of the electron we must account for gravity in the above equation. We will take the specific example of the Earth's gravitational field in order to obtain the exact value for the electron. If in accordance with standard general relativity, the apparent increase in radius r' is:

$$r'=GM/3c^2 \tag{34}$$

[0275] Then given that the mass of the Earth is 5.9745×10^{24} kg; then

$$r'=1.47864\times10^{-3}\ \text{m}$$

thus

$$2\pi r'=9.29057\times10^{-3}$$

[0276] Which is the incremental factor by which ϵ_0 must increase in Earth's gravitational field. So to correct ϵ_0 to account for gravity, ϵ_0 must be divided by the incremental factor, $2\pi r'$. Similarly as effectively π decreases in a gravitational field, to correct π to account for gravity it must be multiplied by this incremental factor. So the equation for an electron in a zero gravitational field is:

$$e=[\epsilon_0/3(4/3\pi c^3)]^{1/2}\div(1+2\pi r')=1.6022\times10^{-19}\ \mathrm{C} \quad (6b)$$

[0277] This now gives the charge of the electron as measured in a zero gravitational field as 1.6022×10^{-19} C, which is the same as that measured on Earth. Notably these observations appear to suggest that the charge of the electron is the same irrespective of the gravitational field.

[0278] Virtually unlimited degrees of accuracy for the charge of the electron and for the fine structure constant (α), may be achieved by taking into account 2nd and nth order gravitometric effects. Thus if we take into account the effect of gravity upon the radius of the Earth it is also important to take into account an effect upon the instruments with which we measure quantities, this would be a second order gravitometric effect. Thus taking into account 2nd order effects (r''), we have a very small, but nevertheless relevant change, such that: $r''=r'(1+2r')$. Thus $2r''=9.3180486\times10^{-3}$, and thus:

$$e=[\epsilon_0/3(4/3\pi c^3)]^{1/2}\div(1+2\pi r'')=1.6021765\times10^{-19}\ \mathrm{C} \quad (6c)$$

[0279] This agrees exactly to the nearest 7 decimal places with the maximum accuracy of the experimental value for the charge of the electron. Furthermore by taking into account the nth order gravitometric effect, it is theoretically possible to predict accuracy for the charge of the electron to 3n decimal places. This mathematically accuracy confirms the structure of the electron from first principles and indeed the theoretical effects of gravity on the permittivity of free space (ϵ_0).

[0280] This returns us directly to the fine structure constant for the electron which is conventionally given by: $\alpha=e^2/hc\cdot4\pi\epsilon_0$. If $\alpha=e^2=\epsilon_0/3(4/3\pi c^3)$, accordingly the quintessential equation for α is structurally given by: $2\pi/\alpha=m^q[3\Theta]^2$ (where $\Theta=4/3\ \pi c^3$; see The Structure of the Electron and Matter), we must now take into account the effects of gravity, as above, thus:

$$2\pi\alpha/m_q[3\Theta]^2\div(1\_2\pi r'')^2=0.007297353$$

[0281] Where the gravitational term for the increase in radius r'' allows the mathematical derivation of $\alpha=0.007297353$, and the above equation is in agreement with the conventional experimental value for $\alpha=0.007297353$ to the nearest 9 decimal places.

[0282] Hence the term $(1\_2\pi r'')^2$ is in accordance with these observations for the effect of gravity on electromagnetic forces. To a maximum accuracy governed by current knowledge of the mass of the Earth and the Gravitational constant and thus the term for the gravitational increase in radius r'. These observations can also be used to accurately predict the magnetic moment of the electron

[0283] Thus the presence of the fine structure constant can now be further understood, by deriving the constant from first principles; specifically from the actual dimensional conformation for the charge of the electron: $e=[\epsilon^0/3(4/3\pi c^3)]^{1/2}$ (eq. 6).

[0284] Overall the fine structure constant α (allowing for the term r' which is the general relativistic increase in the radius of the Earth due to gravitation) is given by none other than the formula for the mass of quintessence and from the structure of the electron, which can now be derived from first principles to seven decimal places or more.

[0285] Quantum Gravity and the Electron Magnetic Moment

[0286] The theoretical origin and nature of magnetism remains obscure in current electromagnetic theory. An explanation suggests these magnetic effects are produced by photons, although no photons have ever been observed. To get round this difficulty it is postulated by physics that magnetism results from "virtual" photons. However, Maxwell's equation for electromagnetism states that the photon has no net magnetic effect.

$$\delta B_x/\delta B_y/\delta y+\delta B_z/\delta z=0$$

[0287] Thus magnetism could not, by the above standard equation, be derived from a photon real or virtual.

[0288] In addition observational data suggests that black holes have powerful magnetic fields and as in theory photons are unable to escape from black holes (except for small quantities in the form of Hawking radiation), it would be difficult to explain these magnetic fields on the basis of photon emission.

[0289] Einstein postulated that magnetism was merely due to special relativity .sup.(ref 17). The postulate for the nature of magnetism in these current observations, states that the magnetic force results from relativity due to none other than the phased emission of gravitons (why postulate two invisible forces, magnetism and gravity, when one, the graviton, will do). This view as previously discussed (Quantum Gravity and Electromagnetism) is entirely compatible with standard relativity(ref 19). Thus with the graviton origin of magnetism, the equation for the magnetic moment of the electron should have an expression in terms of quintessence and in turn the gravitational force and in particular the graviton.

[0290] The standard term for the magnetic moment of the Bohr Magneton (SIB) is:

$$\mu B=eh/4\pi m_e$$

[0291] In standard quantum mechanics the Bohr Magneton, μB, however, needs to be corrected to agree with experiment. The "correction factor" is termed "ϵ"; where $\epsilon=(\alpha/2\pi)-0.328\alpha^2/\pi^2=0.001159641$. Thus theory reveals μ_e, the magnetic moment of the electron where:

$$\mu_e=(eh/4\pi m_e)[1+(\alpha/2\pi)-0.328\alpha^2/\pi^2]$$

[0292] The conventional derivation of the term ϵ above, is given from the fine structure constant, $(\alpha/2\pi)$ which is theoretically consistent. However, a rather arbitrary mathematical correction term; $0.328\alpha^2/\pi^2$ needs to be used in this standard equation. This appears ad hoc and needless to say, more accurate measurements show, the electron magnetic moment to the Bohr magneton ratio, $1+\epsilon=1.001159652$, which suggests the correction factor is indeed incorrect. Nevertheless, this correction factor is essential for "renormalisation" and thus for quantum mechanics to work.

[0293] Quantum gravity readily explains the discrepancy between the theoretical Bohr Magneton (μB) and the actual measured magnetic moment of the electron (μ_e). In accordance with the above chapter (Quantum Gravity and the Charge of the Electron)

[0294] Thus the significant mathematical discrepancies can be removed by accounting for the effects of quantum gravity.

[0295] Thus taking the charge of the electron (e), using the equation for the Bohr magneton and the effects of quantum gravity such that gravitational change in radius is r". The magnetic moment of the electron is given by:

$$\mu_e=(eh/4\pi m_e)(1+[\alpha/2\pi+(1+r'')])$$

[0296] This gives an electron magnetic moment to Bohr magneton ratio of 1.00115968. Thus the mathematical term for the magnetic moment of the electron is given, avoiding the arbitrary and dubious term $0.328\alpha^2/\pi^2$ used in the standard equation, simply by accounting for quantum gravity.

[0297] It is now possible to unite the equations for gravity and magnetism by substituting the fundamental key equations of quantum gravity. Thus if: $h=3m_qc^2$ (eq. 1b) and $m_e=m_qn_q$ (eq. 2). Then we can express the magnetic moment of any particle with the charge of the electron, including the proton, in terms of the number of quintessences (n_q) in that particle.

$$\mu B=ec^2/(4/3\pi n_q); \tag{47}$$

[0298] Given that the postulated structure of the graviton is: ($\phi=\pi/c$) (eq. 42), then substituting we have

$$\mu B=3ec/4\phi n_q \tag{48}$$

[0299] Showing that the equations for the magnetic moment are compatible with the gravitational equations given earlier. Principally, the quintessential equations now allow the determination of the magnetic moment of any charged object from the equation for the graviton and directly from the number of quintessences it contains. In conventional physics the magnetic moment of the electron requires a correction factor, $(1+(\alpha/2\pi c)-0.328\alpha^2/\pi^2)$, to derive the correct experimental value. These observations herein, indicate that the correction factor is more logically (1+r"), where r" is the general relativistic increase in radius around a gravitational body. This suggests that magnetism is not only affected by gravity, but can, as shown as above, be derived using the quantum gravitational equations.

[0300] Quantum Gravity and Special Relativity

[0301] Ordinary matter passing through the lattice would produce gravitons which would interact with space-time as described by general relativity. The quantity of gravitons would be determined by the apparent mass and in turn these would apparently curve space time. The geometry of this "curvature" is elegantly described by general relativity using Riemann geometry, specifically using metric tensors. Intriguingly the metric tensor is not a single number, but at each point in space time it is required to consist of a collection of ten numbers, Consequently, ten dimensional space-time hypotheses, such as this or superstring theory, may automatically yield general relativity

[0302] General relativity is indeed very elegant, nevertheless there was a logical step yet to answer. That is, how do gravitons shape space time? This can now be readily answered by considering the interaction of a three dimensional space time lattice with gravitons themselves to produce the effects of gravity. The effects of gravity are as such to compel a body in motion towards the gravitational object and to a much smaller extent visa versa. This effect can only be produced If gravitons repel quintessence (the constituents of the 3D lattice). Indeed, it has been stated that in order to explain cosmic inflation and the "flatness" of the Universe that quintessence must shun (or be shunned by) matter.

[0303] In descriptive terms a body close to a large mass will have a tendency to move toward it because the three dimensional lattice would be less dense as it approached the surface of the large mass. Overall there would be less resistance to motion in the direction of the large mass, and the motion in this direction would be facilitated by the vibration of quintessence.

[0304] In general relativity the principle governing motion is the geodesic of least distance, this can be re-expressed using similar equations using least action. Furthermore, the concept of motion due to the vibrations of quintessence is more logically and experimentally compelling.

[0305] These observations can now be used to link general and special relativity. Thus as we approach the speed of light, the mass of an object travelling through the space-time lattice would approach infinity, directly because the number of quintessences passing through a body would increase with increasing velocity, hence the equation:

$$m'=m_0/(1-v^2/c^2)^{1/2} \tag{50}$$

or

$$m'=m_0/(1-v^{2\,[\mu}{}_0\epsilon_0])^{1/2} \tag{50a}$$

[0306] In turn this would generate increasing gravitons and accordingly this would explain the observed effects of special relativity. Time itself is due to passage through the space-time lattice, and where the space-time lattice is dispersed by gravitons, time and length are reduced with increasing velocity and hence increasing space-time lattice dispersion, similar to the way in which gravity alters space-time

[0307] As a result:

$$t'=t(1-v^2/c^2)^{1/2},\ l'=l(1-v^2/c^2)^{1/2}$$

[0308] Thus resulting in the effects of special relativity.

[0309] Quintessence and Black Holes

[0310] To address the relationship of the space-time lattice to gravity directly, it is important to discuss the concept of quintessence with regard to general relativistic equations. The standard general relativistic equation for the apparent increase in radius (r) due to the curvature of space time around a gravitational object, which has also been previously derived from first principles (eq. 36), is:

$$r'=GM/3c^2 \tag{34}$$

[0311] This can also thus be written as:

$$3r'=GM[\mu_0\epsilon_0] \tag{51}$$

[0312] This standard equation, is in keeping with the above observations. Specifically, as the mass increases, ϵ_0 increases, in turn the radius will appear to increase (relative to π).

[0313] The above observations now allow us to examine the effects with regard to the interior of black holes themselves. The event horizon would represent a critical density for quintessence, in which light could not escape. The Schwarzschild radius would now be given by:

$$R_s=2GM[\mu_0\epsilon_0]$$

[0314] The event horizon will occur at the point at which there is less resistance to circular motion than motion in a straight or partially curved line. Given that π is proportional to $1/\epsilon_0$. The event horizon should occur when the permittivity has increased by a factor of π.

[0315] Effectively because the permittivity of free space rises, π decreases. This is entirely in keeping with general relativity which predicts the effective change in the ratio of the radius to the circumference as given by the conventional equation, where r', is again the apparent change in radius.

$$r=GM/3c^2$$

[0316] Hence π will effectively decrease as we approach the event horizon of a black hole, When π decreases to 1, the circular circumference is equal to the diameter and moreover, inside this limit it is shorter for light to travel in a circle. Thus light cannot escape the event horizon.

[0317] This can give us great insights into the workings of space-time, for flat Euclidean space the standard equation is:

$$e^{i\pi}=-1$$

[0318] In accordance with general relativity, the ratio of the radius to the circumference changes in a gravitational field, and effectively π=1, at the event horizon, thus the boundary condition for the shape of space-time at the event horizon now has the direct equation:

$$e^{f}$$

[0319] Within a black hole as the permittivity of space increases by a factor of 2π an object within it will complete two rotations rather than travel in a straight line. In effect exceeding the speed of light by 2π. Hence, the condition for space-time is represented by the equation:

$$e^{i/2}$$

[0320] Thus an increase in the permittivity of free space by a minimum factor of π, to produce a black hole is estimated to result from an increase in mass by a factor of approx. 10^6 (the ratio of the mass of the earth and that of a putative black hole).

[0321] Continuing with the subject of a black hole, according to the model inside the black hole, the gravitons produced by the matter present would be in equilibrium with the density of the space-time lattice. Increasing the rate of rotation of the matter in the black hole for instance would thus increase the production of gravitons and its effective mass and increase the radius of the event horizon. A density gradient of the space-time lattice would continue to exist within the black hole. Progressively closer to the center of a black hole matter itself would be increasingly compressed and the spherical structure of the quasi electron would be predicted to collapse. This collapse would result in the formation of an exotic form of matter in the form of pure quintessence in a black hole.

[0322] This pure quintessence would produce the singularity at the centre of the black hole. The larger the black hole in terms of mass the more pure quintessence would exist at its core.

[0323] Quintessence and the Big Bang

[0324] Quintessence theory not only predicts the occurrence of the Big Bang, but allows a prediction for the value of the entire mass of the Universe, from first principles.

[0325] In accordance with quintessence theory the big bang resulted from the explosion of an immense black hole singularity, which was constituted from pure quintessence.

[0326] On the basis of quintessence, there will be a critical mass for Big Bang event; thus if entire space-time between quintessence is compacted so that no further quintessence can be accommodated, the addition of further quintessence would destabilize the immense black hole, resulting in the Big Bang.

[0327] It is possible to predict this critical mass, using the radius of quintessence as a benchmark. Given the nine spatial parameters present in the metric tensor, used in general relativity we find that the actual formula for the radius of quintessence; r_q^2, is mathematically in agreement with general relativistic theory when:

$$9r_q^2=Gh/C^3 \tag{33}$$

[0328] This again supports the 9 dimensional view of space (so crucial in superstring theory). Moreover, the size of the vibrations of quintessence can thus be calculated as:

$$r_q=1.35\times10^{-35}\ \text{m} \tag{33a}$$

[0329] The volume of each quintessence is thus:

$$4/3\pi r_q^3=1.0306\times10^{-104}\ \text{m}^3 \tag{33b}$$

[0330] So to be accommodated within unit volume of space time, with no intervening apparent space time, (given that each of 9 overlapping quintessences are required) would require approx.

$$9\times10^{104}\ \text{quintessences}$$

[0331] As the mass of quintessence is $m^q=h/c^2=7.373\times10^{-51}$ kg sec (eq. 1). Then the mass of the Universe, to two decimal places, is:

$$1.18\times10^{53}\ \text{kg} \tag{33c}$$

[0332] This is in close agreement with a recent estimate of the mass of the Universe from COBE and other satellite data, which estimates the mass to be 100 trillion trillion trillion trillion tonnes (10^{53} kg)

[0333] Moreover, the early formation of the galaxies can be readily explained, it is likely that in such a big bang some very small black holes might have prevailed and that these formed the seeds of the galaxies we see today.

[0334] The event horizon, calculated from the Schwartzschild radius, of such an immense black hole is about 10^{26} m, which would have allowed Guth's inflationary component to the early expansion of the Universe.

[0335] In addition, inflation may result directly from the observation that once electrons have formed from the primordial soup of quintessence, they emit gravitons which in turn repel space time, which might also result in another cosmic inflationary cycle.

[0336] Most importantly quintessence theory explains the Big Bang from first principles and is capable of accurately predicting the mass of the Universe.

[0337] The Nature of Energy

[0338] These observations allow a fundamental understanding of energy. The quantum physical, minimum component of energy is Planck's constant; h. To define the minimal component of mass, using the standard energy equivalence formula; $E=mc^2$, such a minimal mass (m_q)

would be required to have the value equivalent to; $m_q=h/c^2$ (1). The total mass of a system (m) would then be; $m=m_q n_q$, where (n_q) is the number of these minimal units. Thence, the total energy of a system can be derived from the minimal energy; h, multiplied by the number of these energy units (n_q). Thus as, $E=mc^2$, then also $E=m_q n_q c^2$ and substituting $m_q=h/c^2$, the energy equivalence formula has the more logical formulation;

$$E=hn_q \tag{1a}$$

[0339] Thus the energy of a system is equivalent to the minimal energy unit; h, multiplied by the number of those minimal energy units (n_q)

[0340] This leads directly to a deeper understanding of wave particle duality and the wave nature of matter.

[0341] This is encapsulated by the quintessential energy formulae

[0342] As conventionally $\beta \cdot E/c=p$, then

$$\lambda=h/p=hc/\beta E \tag{2}$$

and

$$E=hn_q \tag{1a}$$

then

$$\lambda=c/\beta n_q \tag{2b}$$

[0343] Importantly, as indicated by equation (2b), energy having no quintessence; would have a wavelength of infinity. Specifically pure energy containing no quintessences, would have a lambda of infinity. According to quantum mechanics an infinite wavelength would result in the probability of that energy being anywhere. As energy itself has no electrical charge it would not be impeded by the permittivity and permeability of the three dimensional space-time lattice. Moreover, energy would not be detectable in three dimensional space-time, unless it interacted with matter, as in the EPR experiments. Indeed, energy is not observed when not bound to any form of mass or particle.

[0344] Thus equation 2b, takes us to our original assertion regarding the existence of pure energy.

[0345] Energy is not Bound by the Space-Time Lattice

[0346] Thus, as the EPR experiments suggest the existence of energy separate from matter and thus separate from the three dimensional space-time lattice, it is interesting to find that experiment suggests the existence of free energy in a continuum separate from space time and matter to produce the effects of quantum teleportation.

[0347] This is not, however, teleportation across an additional dimension, this is a term to describe in partially familiar terms the dissociation of energy from the three dimensional space-time lattice. As time is inextricably linked to each dimension of space, the effects of energy would be inextricably linked to the events, such as the creation of virtual particles, we see interacting within space-time. It is unlikely that observers have any direct day to day experience to explain quantum events. Nevertheless, quintessence theory may have given us a window into the hitherto hidden workings of the Universe. Thereby, the mystery of the uniformity of the Universe, across distances which the speed of light could not apparently traverse, is readily explained by the fact that the free energy contained in the Universe is not bound by the space-time lattice.

[0348] In the case of light, due to the exceedingly small masses involved, there would be relatively easy exchange of matter with free energy within a photon. This would make the photon the ideal experimental tool to look for energy which is not bound by matter and in turn energy which is not bound in space-time. Indeed, very recently Furusawa et al. have reported to have observed the transference of energy as photons from A to B, without those photons traversing space-time. This finding which has been supported using other experimental techniques, is very important as it suggests the existence of such free energy.

[0349] Overall, quintessence theory gives an a priori explanation for the concept of mass, the elementary particles, the forces of nature and quantum effects. It can equally be used logically to explain the inner physics of a black hole, the missing mass in the Galaxy, the expansion of the Universe, Guth's inflationary theory and predicts the Big Bang, from first principles.

[0350] Part IV: Applied Theory with the Intent to Create Closed Timelike Curves

[0351] Electron Bombardment of the Photosphere to Induce Gravitational Shift

[0352] Leveraging the above relationship between mass and quintessence we derive that if an electromagnetic radiation with velocity v strikes the event horizon singularity of rest inertial mass mi, and U is the electromagnetic energy absorbed by the singularity, then, according to Maxwell's prediction, a momentum $q=U/v$ is transferred to it. Mass shift $d \cdot m_b$, dependent on the external electromagnetic energy, equals the inertial mass shift dependent on the increment of energy in the particle. Since in this case the inertial mass shift does not depend on velocity V, i.e., it is related only to the momentum q absorbed, it can be obtained by making p=0 in variation $\Delta H=H'-H=c[q^{2}+(m_i c)^2]^{1/2}-(m_i c)^2$ from the particles inertial Hamiltonian. Consequently, the expression of d_{m_p}, is written as:

$$d_{m_p}=\Delta H/c^2=m_i\sqrt{1+U/m_i c^2\left[\varepsilon_r u_{r/2}\left\{\sqrt{1+d/w\varepsilon^2+1}\right\}^2=1\right.}$$

[0353] Comparing now the expression of m_i and m_g we have $m_g=m_i-2dm_b$. By replacing m_b in this equation, given by equation above, we obtain the expression of the correlation between gravitational mass and inertial mass, i.e.,

$$m_g=m_i-2\sqrt{1+U/m_i c^2\left[\sqrt{\varepsilon_r u_{r/2}\{[d/we]^2+1\}^2-1}\right.}\cdot m_i$$

[0354] We see that only in the absence of electromagnetic radiation on the event horizon (U=0) is the gravitational mass equivalent to the inertial mass. Note that the electromagnetic characteristics, e, m and s do not refer to the singularity itself, but to the outside medium around the singularity (photosphere) in which the incident radiation is propagating.

[0355] Stable CTC Solution From Modified M-Theory

[0356] Our innovation is a method of creating an event where a dual membrane or dual boundary condition exists. We do this using generalization from a Misner space which has been modified if one analytically continues the maximal extended Misner Metric so that $ds^2=-du^2+dw^2+(dx^2)^2+(dx^3)^2$ to the Euclidean section so that u=iota zeta we obtain a Misner instanton on the section where w and zeta are both real. The Euclidean time, t, and the closed spacelike coordinate are both periodic, the later having a period of $2\Pi t^2$. Going back to the Lorentzian sector we find that the period of the closed coordinate becomes linearly dependent on the physical time. Using automorphic fields in the Hadamard function one can obtain a quantized condition for time. This gives us a figure on the order of the Plank time. This confines such a stable wormhole condition to the general area of the Plank scale in its modified form which is that area defined by the Membrane itself. As such, these Plank scale wormholes are the true source of the true virtual aspects of the vacuum and quintessence. The effects of the dual singularity system can be viewed as an overlap zone of two distinct space-times which have boundary conditions on both sides. As an object accelerates towards C it is this same boundary or horizon that object encounters when the Time defined horizon solution for the universe is imposed. At this point drawing upon Van Den Broeck's alterations for this space-time geometry to create a single closed Friedman-Robertson-Walker spacetime the space-time geometry can be represented by this equation $ds^2=dct^2-B^2[(dx-\beta fdct)^2]+dy^2+dz^2$ B can be any function that is large near the displacement device. We then consider this transformation as extended to four dimensional space-time with arbitrarily time dependent acceleration. We also present the device frame energy density T00 from a four dimensional calculation and note that the 4d classical calculation is everywhere finite.

[0357] Consider an Alcubierre interval given according to a remote frame's cylindrical coordinates by:

$$ds^2=(1-\beta^2f^2)dct^2+2bfdctdz-dz^2-dr^2-r^2df^2$$

where f is a function that is 1 at the location of the device and zero far from it.

[0358] Starting out with the first transformation z'=z−òct beta dct. Where b is first expressed here as a function of time ct. With some algebra for simplification this results in

$$ds^2=[1-\beta^2(1-f)^2]dct^2-2b(1-f)dctdz'-dz^2-dr^2-r^2df^2$$

Let g=1−f and this becomes

$$ds^2=[1-\beta^2g^2]dct^2-2bgdctdz'-dz'^2-dr^2-r^2df^2$$

[0359] Notice that this returned the original intervals form with a reversal on the sign of b and a reversal of the boundary conditions for g. Now we notice that at r=0, this interval becomes the interval for special relativity transformed to cylindrical coordinates. Thus, we have found a transformation to a frame based local to the device. One can also verify that in these coordinates the relevant affine connections vanish at r=0. a further proposed modification to this field we will reintroduce a time dilation term into the devices frame's interval. Only we will use different boundary conditions for it. We will keep A=1 both at the location of the device, and far from it, but allow it to become large in the warped region. This is achieved by the simplest means possible in the proposed field generation method. Since rotating kerr singularity under bombardment would produce an equal negative energy region. With the effect focused outward away from the device slightly by the relative alignment of the kerr singularity the actual inward going portion of such fields would overlap and cancel their effect out in the region of the device. This yields a space normal time region, which is again restored at the fringes of the outward going field. From the standpoint of a geometric picture of space-time around the device and extending outwards one has actually created a dual event horizon situation as far as time goes. The inner one is the shell of the canceled out field inside of which normal time flow is restored. The outer one is formed at the boundary where normal time resumes and within which we have a negative energy field.

SUMMARY OF THE INVENTION

[0360] The present invention is A method for the generation of a pseudo 2+1 dimensional anti-de Sitter space (DeDeo & Gott 2002) using two Kerr type positively charged rotating dilation singularities where one singularity is maintained as a axis of rotation or "reference" singularity, and the other "target" singularity is subjected to a differential electron flow so as to simultaneously pass above the photosphere of said singularity in its direction of rotation—prograde orbit—and contrary to its direction of rotation—retrograde orbit—to release a directed flow of gravitons in a sinusoidal oscillation simulating a rotational effect of the "target" singularity around the axis of rotation provided by the "reference" singularity, resulting in the creation of timelike curves in a compact time-oriented manifold permitting topology change from one spacelike boundary to the other in accordance with Geroch's theorem (Geroch 1967) which results in a method for the formation of G odel-type geodesically complete spacetime envelopes complete with closed timelike curves.

BRIEF DESCRIPTION OF THE DRAWINGS

[0361] **FIG. 1** is a schematic representation of the mechanism employed to house the components necessary to generate a 2+1 dimensional anti-de Sitter space, resulting in the creation of timelike curves in a compact time-oriented manifold

[0362] **FIG. 2** is a schematic representation of the G odel-type geodesically complete spacetime envelope created by the mechanism complete with closed timelike curves

DETAILED DESCRIPTION OF THE INVENTION

[0363] Principles of Gravity Distortion Time Displacement Systems

[0364] The theoretical understanding of quantum gravity allows the design of time displacement systems from first principles. It is unlikely that gravitons or Kerr singularities can be controlled in a precise way using current technology. Nevertheless, an understanding of three dimensional space-time and matter, does allow the design of elementary displacement systems. That is, systems whose displacement rely on direct warping space-time as opposed to the ejection of material to provide thrust resulting in time dilatational effects.

[0365] The background for these systems are already partially understood and quintessence theory allows their further development. For this invention, the formation of black holes in the laboratory represents a crucial step in understanding the mechanisms that underlay gravitational physics and in turn the warping of space-time. The existence of black holes permits the localized application of the Axial torsion Spin-Rotation Coupling Effect (Zhang & Beesham 2002) resulting in a Rotating Frame with Relativistic Factor (Zhang 2003) which can be used in the creation of a Alcubierre space time bubble under the Van Den Broeck modification of the Alcubierre geometry (Van Den Broeck 1999) resulting in a method for the generation of a pseudo 2+1 dimensional anti-de Sitter space (DeDeo & Gott 2002) using two Kerr type positively charged rotating dilation singularities where one singularity is maintained as a axis of rotation or "reference" singularity, and the other "target" singularity is subjected to a differential electron flow so as to simultaneously pass above the photosphere of said singularity in its direction of rotation and contrary to its direction of rotation to release a directed flow of gravitons in a sinusoidal oscillation simulating a rotational effect of the "target" singularity around the axis of rotation provided by the "reference" singularity. In this instance the space-time lattice would be repelled by gravitons in such a way as to disperse space-time quintessence in a circular fashion around each of the singularities, producing multiple event horizons around the simulated axis of rotation provided by the reference singularity.

[0366] The release of gravitons from the target singularity is controlled by differentially governing the electron flow across the photosphere with the use of powerful electric currents. In turn the differential direction of flow across the photosphere of gravitons would determine the direction of motion through the space time lattice and the resulting time displacement from one spacelike boundary to the other in accordance with Geroch's theorem (Geroch 1967) when implementing this approach it is important to remember that the black hole is not excited by the smashing of the clump as it "hits" the horizon. The hole is rather excited when the metric perturbation associated with the clump is "felt" by the background metric. The excitation event therefore constitutes a smooth process whereby in-fall of a clump from .rmb and through r+ serves as a source in the Teukolsky (1973) equation for small perturbations to the Kerr geometry (with appropriate boundary conditions at r+ and r∞). This is an important distinction with a great deal of relevance to the practical engineering of the displacement unit since we need to gauge the "driving" of QNR modes in terms of an e_ective coupling from clump in-fall.

[0367] What results is a method for resonant driving of the quasi-normal ringing (QNR) wave modes of the Kerr geometry of the target singularity. The micro black hole hyper-accreting at rates ÿ M. 1 M□sec. 1 from a neutrino cooled disk is pushed through to oscillate near resonance of its (l,m=2, 2) quadrupole QNR frequency due to the in-fall of compact mass over-densities from the cusp in e_ective potential on a dynamical time scale. This mode is induced via induced magneto-rotationally induced fluid dynamics in the ultra-relativistic region of the flow bounded from below by the marginally bound orbit radius: rmb If the QNR modes are fed resonantly for a few seconds of hyper-accretion, the enhanced amplitude of the oscillations yields a very high rate of energy deposition into gravitational waves. Indeed, the integrated energy deposition is large enough to "evaporate" the equivalent of a factor of a few times the total rest mass-energy of a single clump into gravitational waves, which in turn interact with the reference singularity

[0368] Application of the method described in the previous section results in translation outside of the cosmological horizon, where F(R)<0. The topology of the resulting geodesic, for large constant R, is mathematically equivalent to a Euclidean cylinder of the condition R×Sn where T is the coordinate along the cylinder. I± are located outside the future/past cosmological horizons, where R is timelike and T is spacelike. In the case of a rotating Kerr black hole, there are two circular photon orbits that can exist in the equatorial plane and be exploited to create a tipler sinusoid. One is a prograde orbit moving in the same direction as the black hole's rotation, while the other is a retrograde orbit moving against the black hole's rotation. Their radii are respectively given by

$$r1=2M(1+\mathrm{Cos}(2/3\ \arccos(-|\alpha|/M)))$$

$$r2=2M(1+\mathrm{Cos}(2/3\ \arccos(|\alpha|/M)))$$

where a is the angular momentum per unit mass of the black hole. The orbits fall in the range M greater or equal to r1 greater or equal to 3M greater or equal to r2 greater or equal to 4M The fact that a prograde photon or in or case cooper pair with Bose characteristics orbits the black hole at a smaller radius than a retrograde one can be attributed to the well-known Lense-Thirring effect, i.e., the dragging of inertial frames due to the black hole's rotation which we control and influence via electromagnetic induced load coupling with the open magnetic field lines threading the BH horizon. This dragging would cause charged cooper pairs to revolve around the black hole relative to a static observer at infinity. Thus, to such an observer, a prograde cooper pair would have to orbit at a smaller radius to compensate for the 'extra' angular momentum acquired, while a retrograde one would have to orbit at a larger radius to compensate for the 'lost' angular momentum. Indeed, in the limit of zero rotation, these two orbits coincide at r D 3M, giving the single circular orbit of the Schwarzschild black hole. Now, recall that orbits around the Schwarzschild black hole are necessarily confined to a plane passing through its center, because of the spherical symmetry of the space-time. However, the Kerr black hole space-time has only an axial symmetry (in addition to being stationary), and this raises the possibility of non-planar orbits. One could, for example, contemplate the existence of spherical Boseon orbits—orbits with constant coordinate radii that are not necessarily confined to the equatorial plane—around the Kerr black hole. Such orbits would be a nontrivial generalization of the two circular photon orbits that lie in the equatorial plane. At first it may seem a little surprising that such spherical orbits could even exist, but there is an interesting reason as to why they are possible. Note that an object in a spherical orbit would, in addition to moving around the black hole in the azimuthal direction, be undergoing some periodic motion in the latitudinal direction. This is only possible if there is a conserved quantity associated with motion in this direction, just as angular momentum is necessarily conserved by its rotational motion in the azimuthal direction. (This result can be seen, for example, using action-angle variable) Now, because the Kerr space-time has only axial symmetry, geodesics in it should have only two constants of motion, namely energy and angular momentum. However, Carter discovered the remarkable fact that geodesics in the Kerr space-time possess a third constant of motion. It turns out that Carter's new constant governs the motion of geodesics in the latitudinal direction, although it is not related to any obvious space-time symmetry. Thus, spherical timelike orbits, which assume eigenlike properties around the Kerr black hole, are possible and when coupled with charged Bosons may be exploited to manipulate the relative event horizons of the "target" singularity. The location and the temperature of the modified event horizon depend on the time, charge and angle of incidence of the cooper pairs. The Fermionic spectrum of Dirac particles displays a spin-rotation coupling effect due to the interaction between the particles with spin-1=2 and the black holes with rotation. The effects arise from the interaction between the spin of Dirac particles and the rotation of the evaporating black holes. The feature of this spin-rotation coupling effect is its dependence on different helicity states of coupled particles with spin-½ and its irrelevance to the mass of particles.

[0369] In order to design a mechanism for time displacement exploiting the Carter asymettry cited above we utilize two positively charged top spin rotating kerr type black holes aligned on demand to create a simulated rotational effect around a central axis provided by the "reference" singularity. In the case of the Kerr blackhole (singularity), this is accomplished by an inverse Blandford-Znajek (BZ) process utilizing the magnetic flux of open field lines connecting the horizon and an induced remote load. (Ding-Xiong Wang, Kan Xiao & Wei-Hua Lei, 2001) permitting the microscopic blackhole (singularity) to be rotated along its horizontal axis in at relativistic centrifugational speeds. A differential current is then applied in such a way as to pass through the entire photosphere in the desired direction. As a result the gravity field can be manipulated by three factors that affect it in distinct ways. Adding electric charge to the singularities increases the diameter of the inner event horizons. Adding mass to the singularities increases the area of gravitational influence around the singularities. Rotating and positioning the polar axis of the singularities affects and alters the resulting gravity sinusoid

[0370] In order to create a sinusoid capable of inducing a topology change from one spacelike boundary to the other in accordance with Geroch's theorem The electric charge in the upper half of the photosphere would be maximised. The electrons will have a vector in the left to right direction as the singularity spins clockwise. If a maximised current is applied to the singularity in the same direction this will result in a increase in the velocity of the electrons relative to the centre of gravity of the singularity, due to the flow of current. In turn, according special relativity and to the space-time lattice model, this wilt result in an increase in the relativistic mass if the electrons and in turn by general relativity an increase in the release of gravitons.

[0371] Conversely in the lower half of the singularity the electrons will have a vector of motion in the right to left direction due to the spin of the singularity. This will be relativistically slowed by the differential current applied in the same direction as the current above, and hence in the opposite direction to the direction of rotation. The charge can be separately applied and adjusted to ensure that the electrons are relativistically stationary relative to the centre of gravity. In turn this will minimize the relativistic mass and result in a decrease in the release of gravitons for the lower half of the singularity.

[0372] The overall result will be a greater release of gravitons in one direction and a lesser release of gravitons in the converse direction. The effect will be enhanced by the use of a multi-phasic current simultaneously applied. This will result in the release of multi-phasic gravitons which will disperse space time in the singularity with increase in density in the converse this effect can also be produced and supplemented with the use radio frequency pulses, The radio frequency pulses must be designed to produce a change in the spin of the particle to enhance the release of gravitons in the desired direction

[0373] With the use of large currents the drift velocity of the electrons across the photosphere could be greatly increased. Within this region the electrical resistance is virtually eliminated. Thereby allowing large currents to be induced with minimum total power output.

[0374] The result is the production of cooper paired electrons of high speed and hence high relativistic mass in the desired half of the singularity, whilst producing low speed and thus low mass paired electrons at the converse of the singularity, in accordance with special relativity. The imbalance in the rotating singularity will be continuously present creating a dynamic warping of space-time. In effect, the differential current flow, will produce differential graviton production and in turn, by general relativity, the warping of space-time

[0375] As the cooper pairs cross the photosphere, accretion of the Bose particles results. Near-hole accretion across the target singularity is then motivated by magneto-rotationally induced, ultra-relativistic disk dynamics in the region of the flow bounded from below by the marginally bound geodesic radius rmb. As the particles impelled have high spin values, a largely coherent magnetic field in this region has the dynamical implication of compact mass segregation at the displacement nodes of the non-axisymmetric, MRI modes. This results in prolific gravitational wave emission coincident with the gamma-ray stage. The gravitational wave emissions are then manipulated to influence the reference singularity to produce a variable gravitational sinusoid which is then used to mathematically approximate the gravametric distortion, inducing a topology change from one spacelike boundary to the other in accordance with Geroch's theorem.

BIBLIOGRAPHY

[0376] Mesoscopic behaviour of the neutral Fermi gas ^{3}He confined in quantum wires. S. L. Phillipson, A. M. Genault, S. N Fisher, G. R. Pickett, P. J. Y Thibault. Nature, 395: 578-580.

[0377] Direct Observation of d-orbital Holes and Cu—Cu Bonding in Cu$_2$. J. M. Zuo et al. Nature 401: 49-52.

[0378] The Foundation of the General Theory of Relativity, A Einstein. Translated from "Die Grundlage der Allgemeinen Relativittstheorie". Annalen der Physik, 49, 1916. Published in, Einstein, The Principle of Relativity, Dover publications, 1952.

[0379] Sidelights on Relativity. Ether and the Theory of Relativity Albert Einstein. Dover Publications 1983.

[0380] Inflation is dead long live inflation. George Musser. Scientific American, July 1998, p 9.

[0381] The flip side of the Universe. George Musser. Scientific American, September 1998, p 13.

[0382] The Fifth Element, Marcus Chown. New Scientist, Apr. 3, 1999, pp 29-32

[0383] The Quantum Universe. Tony Hey and Patrick Walters. Cambridge University Press 1987. (Solving the Shrodinger equation pp. 165-167).

[0384] Complex Numbers, Mathematics for Engineers. W. Bolton; Longman Scientific & Technical; Chap. Introducing Complex Numbers, p 1-11.

[0385] Unconditional Quantum Teleportation. A. Furusawa, I. Sorensen, S. L. Braunstein, C. A. Fuchs, H. J Kimble, E. S. Polzik. Science, Oct. 23, 1998. Vol. 282: 706-709.

[0386] Complete quantum teleportation using nuclear magnetic resonance. M. A. Neilsen, E. Knill, R. Laflamme. Nature, Nov. 5, 1998. Vol. 396: 52-55.

[0387] Experimental quantum teleportation. D. Bouwmeester et al. Nature, 390: 575-579 (1997).

[0388] Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolski-Rosen channels. D. Boschi, S. Branca, F. De Martini, L. Hardy, S. Popescu. Phys. Rev. Lett. 80: 1121-1125 (1998).

[0389] W. Tittel, J. Brendel, H. Zhuniden, N. Gisin, Phys. Rev. Lett. 81: 3563-3566. (1998) into the Vortex. Charles Seife. New Scientist, Oct. 24, 1998. p 7.

[0390] Direct observation of a fractional charge. R de Picciotto, M Reznikov, M. Heibium, V Umansky, G. Bunin, D. Mahalu. Nature, 389: 162-164.

[0391] Tables of Physical and Chemical Constants. Kay & Laby. Longman Scientific & Technical.

[0392] Physics in Focus. Michael Brimicombe. Section, Wave Mechanics; p 521. Nelson Special Relativity. A. P. French. Nelson. Chapter 8, Special Relativity and Electricity, pp 256.

[0393] Universe In Balance. Jeff Peterson. New Scientist, 168: p 27

[0394] The Accelerating Universe. Mario Livio. John Wiley and Sons, Inc. (2000)

[0395] Escape from Earth, Charles Seife, New Scientist, Feb. 6, 1999, p 6.

[0396] The Discovery of the Top Quark. Tony Liss and Paul Tipton. Scientific American; September 1997, pp. 26-41.

[0397] The Mystery of the Nucleon Spin. Klaus Rith and Andreas Schaifer. Scientific American: July 1999, pp 42-47.

[0398] And you're glue. Frank Wilczek. Nature; 400, pp 21-23.

[0399] The warp drive: hyper-fast travel within general relativity Miguel Alcubierre —Classical and Quantum Gravity, vol. 11, L73-L77, (1994)

[0400] Photon propagation in a stationary warp drive space-time: Claes R. Cramer (1995)

[0401] Some thoughts on the Implications of Faster-Than-Light Interstellar Space Travel: I. A. Crawford, Quarterly Journal of the Royal Astronomical Society, vol. 36, 205-218, (1995)

[0402] Physical and Cosmological Implications of a Possible Class of Particles Able to Travel Faster than Light: Luis Gonzalez-Mestres: Contribution to the 28th International Conference on High Energy Physics, Warsaw (Poland), (1996)

[0403] Warp drive and causality: Allen E. Everett, Physical Review D, vol. 53, 7365-7368, (1996)

[0404] A Superluminal Subway: The Krasnikov Tube, Allen E. Everett & Thomas A. Roman, Physical Review D, vol. 56, 2100-2108, (1997)

[0405] Quantum effects in the Alcubierre warp drive spacetime, William A. Hiscock, Classical and Quantum Gravity, vol. 14, L183-L188 (1997)

[0406] The unphysical nature of "Warp Drive", Michael J. Pfenning & L. H. Ford; Classical and Quantum Gravity, vol. 14, 1743-1751, (1997)

[0407] On the Possibility of a Propulsion Drive Creation Through a Local Manipulation of Spacetime Geometry; Vesselin Petkov; Presented at the 34th

[0408] AIAA/ASME/SAE/ASEE Joint Propulsion Conference, (1998)

[0409] 'Operational' energy conditions, Adam D. Helfer; Classical and Quantum Gravity, vol. 15, 1169-1183, (1998)

[0410] No warp drive; D. H. Coule; Classical and Quantum Gravity, vol. 15, 2523-2527, (1998)

[0411] Quantum Inequality Restrictions on Negative Energy Densities in Curved Spacetimes, Michael John Pfenning; Doctoral Dissertation, (1998)

[0412] Hyperfast Interstellar Travel in General Relativity; S. V. Krasnikov; Physical Review D, vol. 57, 4760, (1998)

[0413] Superluminal travel requires negative energies; Ken D. Olum; Physical Review Letters, vol. 81, 3567-3570, (1998)

[0414] Hyper-fast travel without negative energy, Eric Baird; (1999)

[0415] Warp drives, wavefronts and superluminality, Eric Baird; (1999)

[0416] A traversable wormhole, S. Krasnikov; (1999)

[0417] Speed Limits in General Relativity; Robert J. Low; Classical and Quantum Gravity, vol. 16, 543-549, (1999)

[0418] Null geodesics in the Alcubierre warp drive spacetime: the view from the bridge; Chad Clark, William A. Hiscock & Shane L. Larson; Classical and Quantum Gravity, vol. 16, 3965-3972, (1999)

[0419] A 'warp drive' with more reasonable total energy; Chris Van Den Broeck; Classical and Quantum Gravity, vol. 16, 3973-3979, (1999)

[0420] On the warp drive space-time, Pedro F. Gonzalez-Diaz; Physical Review D, vol. 62, 44005-44012, (2000)

[0421] On the (im)possibility of warp bubbles; Chris Van Den Broeck; Summary of talk delivered at STAIF-2000, (2000)

[0422] Reduced Total Energy Requirements for a Modified Alcubierre Warp Drive Spacetime; F. Loup, D. Waite & E. Halerewicz Jr. (2001)

[0423] Warp Drive With Zero Expansion, Jose Natario; Classical and Quantum Gravity, vol. 19, 1157-1166, (2002)

[0424] A Causally Connected Superluminal Warp Drive Spacetime; F. Loup, R. Held, D. Waite, E. Halerewicz, Jr., M. Stabno, M. Kuntzman & R. Sims, (2002)

[0425] Weak Energy Condition Violation and Superluminal Travel; Francisco Lobo & Paulo Crawford (2002)

[0426] On the Problems of Hazardous Matter and Radiation at Faster than Light Speeds in the Warp Drive Space-Time; C. B. Hart, R. Held, P. K. Hoiland, S. Jenks, F. Loup, D. Martins, J. Nyman, J. P. Pertierra, P. A. Santos, M. A. Shore, R. Sims, M. Stabno & T. O. M. Teage, (2002)

[0427] 2+1 gravity, chaos and time machines; Bengtsson I and Brannlund J, J. Math. Phys. 42 (2001) 3565

[0428] Time Travel in Einstein's Universe: the Physical Possibilities of Travel through Time, Gott J R; Houghton Mifflin Company, Boston, 2001

[0429] Gravitational memory of natural wormholes, Anchordoqui L A Trobo M L Vucetich H Zyserman F Mod Phys Lett A 15(6): 429-438, 2000

[0430] A traversable singularity-free wormhole. The gravitational memory phenomenon Gott time machines, BTZ black hole formation, and Choptuik scaling, Birmingham D, Sen S; Phys Rev Left 84 (2000) 1074-1077

[It is shown that the Gott time machine provides a mechanism for black hole formation]

[0431] Negative energy, wormholes and warp drive, Ford L H, Roman, T A; Sci Am: 46-53, 2000 Jan (reply: 2000 May p. 12)

[0432] Warp drive space-time; Gonzalez-Diaz P F; Phys Rev D62: 044005 (2000)

[0433] Misner-brane cosmology, Gonzalez-Diaz P F; Phys Lett B486: 158-164 (2000)

[0434] Kinks, energy conditions and closed timelike curves; Gonzalez-Diaz P F; Int J Mod Phys D 9(5): 531-541, 2000. (gr-qc/0007065)

[0435] Will we travel back (or forward) in time? Gott J R; Time, Apr. 10, 2000 pp 68-70

[0436] Charged universes of the Godel type with closed timelike curves; Klepac P and Horsky J, Class Quant Grav 17(13): 2547-2555, 2000

[0437] The Ori-Soen time machine, Olum K D; Phys Rev D61 (2000) 124022

[0438] How much energy do closed timelike curves in 2+1 spacetimes need?, Tiglio M H, Phys Rev D61 (2000) 08 1503

[0439] Energy conditions and their cosmological consequences, Visser M and Barcelo C gr-qc/0001099

[0440] Cylindrically symmetric spinning Brans-Dicke spacetimes with closed timelike curves Anchordoqui L A, Perez Bergliaffa S E, Trobo M L, Birman G S; Mod Phys Lett A14 (1999): 1105-1112

[0441] From spinning to non-spinning cosmic string spacetime; De Lorenci V A De Paola R DM Svaiter N F, Class Quant Grav 16(10): 3047-3055, 1999

[0442] Bounds on negative energy densities in static space-times, Fewster C J, Teo E; Phys Rev D 5910(10): 4016, 1999

[0443] Decohering a charged scalar field in a time-machine wormhole background, Flagga M SN; Int J Theor Phys 38(8): 2121-2161, 1999

[0444] Rotating black holes in unified theories; Gibbons G W. Progress of Theoretical Physics Supplement. (136): 18-28, 1999

[0445] Supersymmetric rotating black holes and causality violation; Gibbons G W and Herdeiro C AR, Class Quant Grav 16 (11) 3619-3652, 1999

[0446] Perdurance of multiply connected de Sitter space; Gonzalez-Diaz P F, Phys Rev D 59, 1999, 123513

[0447] Nonorientable spacetime tunneling; Gonzalez-Diaz P F, Garay L J, Phys Rev D 59 (1999) 064026

[0448] Can the universe create itself?; Gott J R, and Li X L, Phys Rev D 58, 1998, 023501. Also: astro-ph/9712344

[0449] Supersymmetric rotating black holes and causality violation; Holst S, Matschull H-J Class Quant Grav 16, 1999, 3095

[0450] Godel-type universes in string-inspired charged gravity; Kanti P, Vayonakis C E, Phys Rev D 60, 1999, 103519

[0451] Quantum field theory and time machines; Krasnikov S., Phys Rev D 59: 4010, 1999 Jan 14 Dec. 1998

[0452] Time machines constructed from anti-de Sitter space; Li L X, Phys Rev D 59 (1999) 084016

[0453] Wormholes in string theory, Vollick D N., Class Quant Grav 16(5):1599-1604, 1999

[0454] Time machines and the breakdown of unitarity; Antonsen F and Bormann K, Int J Theor Phys 37(9): 2383-2393, 1998

[0455] An exterior for the Godel spacetime; Bonnor W B, Santos N O, MacCallum M A H, Class. Quant. Grav. 15 (1998) 357-366

[0456] All the worlds a time machine, Chown M, New Scientist 157(2124): 38-41, 1998 Mar. 7.

[short exposition of M. Hadleys idea that an elementary particle is a local ST warp that it permits closed timelike curves. Hence general relativity could reproduce (explain) the effects of QM]

[0457] Quantum time machine; Gonzalez-Diaz P F Phys Rev D 58 (12): 4011, 1998

[0458] A design for a quantum time machine, Gray A quant-ph/9804008

{misleading title. On QM measurement problem; quantum "seeing into future."}

[0459] Time machines with non-compactly generated cauchy horizons and "handy singularities"; Krasnikov S V, gr-qc/9711040 (To appear in the proceedings of the 8th Marcel Grossman Conference)

[0460] A singularity-free WEC-respecting time machine, Krasnikov S V, Class Quant Grav 15 (1998) 997-1003

[0461] Self-consistent vacuum for Misner space and the chronology protection conjecture, Li L X and Gott J R, Phys Rev Lett 80, 2980, 1998

{Misner space stable to quantum flutuations; against CPC}

[0462] Decoherence caused by topology in a time-machine spacetime; Mensky M B and Novikov I D, Int. J. Mod. Phys. D5 (1996) 1-27

[0463] Godel metric as a squashed anti-de Sitter geometry, Rooman M, Spindel Ph, Class Quant Grav 15 (1998) 3241-3249

[0464] Testing causality violation on spacetimes with closed timelike curves; Rosenberg S, Phys Rev D 57(6):3365-3377, 1998

{politzer TM used to investigate the scattering experiment; interesting results on agreement with the no CTC case if there is enough TM's}

[0465] A new non-Hausdorff spacetime model for resolution of the time travel paradoxes; Sharlow M F Annals of Physics 263(2): 179-197, 1998

[0466] Particle creation near the chronology horizon; Sushkov S V, Phys Rev D 58, 1998, 044006

[0467] Divergences in the effective action of acausal spacetimes, Cassidy M J, Class Quant Grav 14 p 523, 1997 {shows that a self-consistent quantum vacuum for Misner space exist, allowing time travel}

[0468] The 1-loop instability of Misner space; Cassidy M J, Class Quant Grav 15 (8),

[0469] Superluminal subway—the Krasnikov tube; Everett A E and Roman T A, Phys Rev D 56 (4): 2100-2108, 1997

[0470] Time machines and the principle of self-consistency as a consequence of the principle of stationary action (ii)—the Cauchy problem for a self-interacting relativistic particle, Carlini A and Novikov I D, Int J Mod Phys D 5(5): 445-479, 1996 Oct.

[0471] Warp drive and causality, Everett A E, Phys Rev D 53 (12) pp 7365-7368, 1996 Ringholes and closed timelike curves, Gonzalez-Diaz P F Phys Rev D 54(10): 6122-6131, 1996

[0472] Time machine as four-dimensional wormhole; Guts A K gr-qc/9612064

[0473] Gott time machines in the Anti-de Sitter space; Holst S, Gen Rel Grav 28 (1996) 387-403

[0474] The quantum stability of the time machine, Krasnikov S V, Phys Rev D54 (1996) 7322-7327

[0475] Must time machine be unstable against vacuum fluctuations?, Li L X, Class Quant Grav 13 (1996) 2563-2568

[0476] Three-dimensional billiards with time machine, Mensky M B and Novikov I D, Int J Mod Phys D 5(2): 179-192, 1996

[0477] Decoherence caused by topology in a time-machine spacetime, Mensky M B and Novikov I D, Int J Mod Phys D 5(1): 1-27, 1996.

[0478] Improved time-machine model; Soen Y and Ori A, Phys Rev D 54(8): 4858-4861,

[0479] Unitarity restoration in the presence of closed timelike curves; Anderson A, Phys Rev D51 (1995) 5707-5715

[0480] Problems in time-machine construction due to wormhole evolution; Antonsen F and Bormann K, Int J Theor Phys 34 1995, p 2061

[0481] Time machines: the principle of self-consistency as a consequence of the principle of minimal action; Carlini A, Frolov V P, Mensky M B, Novikov I D, and Soleng H H, Int J Mod Phys D4 (1995) 557-580; Erratum-ibid. D5 (1996) 99

[0482] Nonlinearity in quantum theory and closed timelike curves; Cassidy M J, Phys Rev D52 (1995) 5676-5680

[0483] Unitarity of Quantum Theory and Closed Time-Like Curves; Fewster C J, Wells C G, Phys Rev D52 (1995) 5773-5782

[0484] Time machines without closed causal geodesics; Low R J, Class Quant Grav 12(5): L 37-L 38, 1995 May.

[0485] A remark on kinks and time machines; Yurtsever U, Gen Rel Grav 27(6): 691-693, 1995

[0486] The warp drive: hyper-fast travel within general relativity; Alcubierre M, Class Quantum Grav 11, L73 (1994)

[0487] Energy-momentum restrictions on the creation of Gott time machines; Carroll S M, Farhi E, Guth A H and Olum K D, Phys Rev D 50 pp 6190-6206, 1994

[0488] Kinks and time machines; Chamblin A, Gibbons G W and Steif A R, Phys Rev D 50(4) pp R2353-R2355, 1994

[0489] The breakdown of quantum mechanics in the presence of time machine, Goldwirth D S, Perry M J and Piran T, Gen Rel Grav 25 pp 7-13 1994

[0490] Quantum propagator for a nonrelativistic particle in the vicinity of a time machine; Goldwirth D S, Perry M J, Piran T and Thorne K S, Phys Rev D 49(8): 3951-3957, 1994

[0491] Time machine, self-consistency and the foundations of quantum mechanics; Gonella F, Found Phys Lett 7(2): 161-166, 1994

[0492] Time walk through the quantum cosmic string; Gonzalez-Diaz P F, gr-qc/9412034

[0493] Unitarity and causality in generalized quantum mechanics for nonchronal spacetimes; Hartle J B, Phys Rev D 49 (12) pp 6543-6555, 1994

[0494] (2+1)-dimensional spacetimes containing closed timelike curves; Headrick M P, Gott J R; Phys Rev D 50, 1994, 7244-7259

[0495] Time travel, Jackiw R.; Ukrainskii Fizicheskii Zhurnal. 39(1): 6-14, 1994.

[0496] On the classical stability of a time machine; Krasnikov S V, Class Quant Grav 11 (11): 2755-2759, 1994 Nov.

[0497] Isometries between Gott's two-string spacetime and Grant's generalization of Misner space, Laurence D, Phys Rev D 50 pp 4957-4965

[0498] Negative Energy Densities in Extended Sources Generating Closed Timelike Curves in General Relativity with and without Torsion; Soleng H H, Phys Rev D49 (1994) 1124-1125

[0499] Multiparticle solutions in 2+1 gravity and time machines; Steif A R, Int J Mod Phys D 3(1): 277-280, 1994

[0500] VanVleck determinants—traversable wormhole spacetimes; Visser M

[0501] Algebraic approach to quantum field theory on non-globally-hyperbolic spacetimes; Yurtsever U, Class Quant Grav 11 (1994) 999

[0502] Cosmic strings and chronology protection; Grant J DE, Phys Rev D 47, 2388 (1993)

[0503] Complex geometry, quantum tunneling, and time machines; Li L X, Xu J M and Liu L, Phys Rev D 48(10): 4735-4737, 1993

[0504] Quantum field theory in spaces with closed timelike curves; Boulware D G, Phys Rev D46 (1992) 4421-4441

[0505] Global structure of Gott's two-string spacetime; Cutler C, Phys Rev D 45 p. 487-494, 1992 {Cauchy horizons for Gott's machine (two-string space-time)}

[0506] "Cosmic flashing" in four dimensions; Ford L H and Roman T A, Phys Rev D 46, 1328-1339 (1992)

[0507] Unitarity of interacting fields in curved spacetime; Friedman J L Papastamatiou, Phys Rev D 46 4442, 1992

[0508] Feilure of unitarity for interacting fields on spacetimes with closed timelike curves; Friedman J L Papastamatiou, and Simon J Z, Phys Rev D 46 4456-4469, 1992

[0509] Particle creation for time travel through a wormhole; Kim S W, Phys Rev D 46 2428-2434, 1992

[0510] Time machine and self-consistent evolution in problems with self-interaction; Novikov I D, Phys Rev D 45 p. 1989, 1992

[0511] Structure of the singularity inside a realistic rotating black hole, Phys Rev Lett} 68, 2117-2120; Ori A (1992)

[0512] Simple quantum systems in spacetimes with closed timelike curves; Politzer H D, Phys Rev D 46 4470-4476, 1992

[0513] Quantum mechanics near closed timelike lines; Deutsch D, Phys Rev D 44 (10) p 3197-3217 1991

[0514] Billiard balls in wormhole spacetime with closed timelike curves: Classical theory; Echeverria F, Klinkhammer G and Thorne K S, Phys Rev D 44 (4) p. 1077-1099, 1991

[0515] Closed timelike curves produced by pairs of moving cosmic strings: Exact solutions; Gott J R, Phys Rev Left 66, 1126, (1991)

[0516] Inner structure of a charged black hole: An exact mass-inflation solution; Ori A, Rev Left 67: 789, 1991

[0517] On the outcome of Kerr-like collapse; Charlton N and Clarke C J S, Classical and Quantum Gravity 7 No. 5 (May 1990)

[0518] Cauchy problem in spacetimes with closed timelike curves; Friedman J, Morris M S, Novikov I D, Echeverria F, Klinkhammer G, Thorne K S and Yurtsever U, Phys Rev D 42 (6) 1915-1930, 1990 {QM fixes the TM classical billiard paradox (Polchinski's prdx. "CTC's may be not as nasty as people have assumed". The principle of self-consistency in time travel reviewed.}

[0519] Traversable wormholes from surgically modified Schwarzschild spacetimes, Visser M, Nucl Phys B 328 pp 203-212

[0520] Traversable wormholes: some simple examples; Visser M; Phys Rev D 39: 3182-3184 1989

[0521] Minimal acceleration requirements for time travel in Godel spacetime; Malament D, Journal of Mathematical Physics 26, 1985, p 774-777

[0522] Timelike nongeodesic trajectories which violate causality: A rigorous derivation; de Felice F, Nuovo Cimento 65 B pp 224-232, 1981

[0523] Rotating cylinders and the possibility of global causality violation; Tipler F J, Phys Rev D 9 pp 2203-2206, 1974

[0524] Global structure of the Kerr family of gravitational fields, Carter B, Phys Rev 174, p 1559, 1968

[0525] Complete analytic extension of the symmetry axis of Kerr's solution of Einstein's equations, Carter B, Phys Rev 141, p 1242, 1966

[0526] Acta Physica Polonica, Staruszkiewicz A 24 1963, p 734

[0527] An example of a new type of cosmological solutions of Einstein's field equations of gravitation, Goedel K, Rev Mod Phys 21 p. 447, 1949

[0528] Gravitational field of a distribution of particles rotating about an axis of symmetry; Van Stockum W J, Proc R Soc Edin 57 pp 135-154, 1937

[0529] Appendix A: The Synthetic creation of micro singularities. Throughout this patent application the creation of micro singularities has been maintained. While at first glance this may seem to be a deficiency in the patent application it is in fact an engineering problem that is in the process of being solved. In the International Journal of Theoretical Physics 41 (11): 2073-2090, November 2002 L. J. Garay published "Black Holes in Bose-Einstein Condensates". In this article Garay showed that there exist both dynamically stable and unstable dilute-gas Bose-Einstein condensates that, in the hydrodynamic limit, exhibit a behavior completely analogous to that of gravitational black holes. This discovery coupled with creation of color glass condensate (CGC) an extreme form of nuclear matter in which a nucleus travels at near-light (relativistic) speed which flattens like a pancake in its direction of motion and spawns a large number of gluons hold the key to high energy singularity formation.

I claim:

1. A method for the generation of a pseudo 2+1 dimensional anti-de Sitter space comprising the steps of:

creating two Kerr type positively charged rotating dilation singularities, including the steps of

maintaining one of the singularities as a axis of rotation reference singularity,

maintaining the other of the singularities as a target singularity, and

subjecting the target singularity to a differential electron flow so as to simultaneously pass the differential electron flow above a photosphere of said target singularity in a direction of rotation thereof and contrary to the direction of rotation thereof, in order to release a directed flow of gravitons in a sinusoidal oscillation simulating a rotational effect of the target singularity around the axis of rotation provided by the reference singularity.

2. A method of generating a force around a body, comprising the steps of:

employing sinusoidal oscillations of electrical bombardment on the surface of one Kerr type reference singularity in close proximity to a second Kerr type target singularity to take advantage of the Lense-Thirring effect, wherein the electrical currents employed in the bombardment are passed simultaneously across the photosphere of said reference singularity in its direction of rotation and contrary to its direction of rotation to release a directed flow of gravitons in a sinusoidal oscillation simulating a rotational effect of the target singularity around the axis of rotation provided by the reference singularity;

creating timelike curves in a compact time-oriented manifold of Godel-type geodesically complete spacetime envelope under the Van Den Broeck modification of the Alcubierre geometry, resulting in the creation of timelike curves in a compact time-oriented manifold permitting topology change from one spacelike boundary to the other in accordance with Geroch's theorem.

* * * * *

US 20070285325A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2007/0285325 A1**

St. Clair (43) **Pub. Date: Dec. 13, 2007**

(54) **CHI ENERGY AMPLIFIER**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911

(21) Appl. No.: **11/448,456**

(22) Filed: **Jun. 7, 2006**

Publication Classification

(51) **Int. Cl.**
H01Q 1/00 (2006.01)
H01Q 13/10 (2006.01)

(52) **U.S. Cl.** **343/720**; 343/767

(57) **ABSTRACT**

A Chi energy amplifier that utilizes a co-gravitational K field to generate a flow of hyperspace energy into or out of a ceramic dome containing a slot antenna.

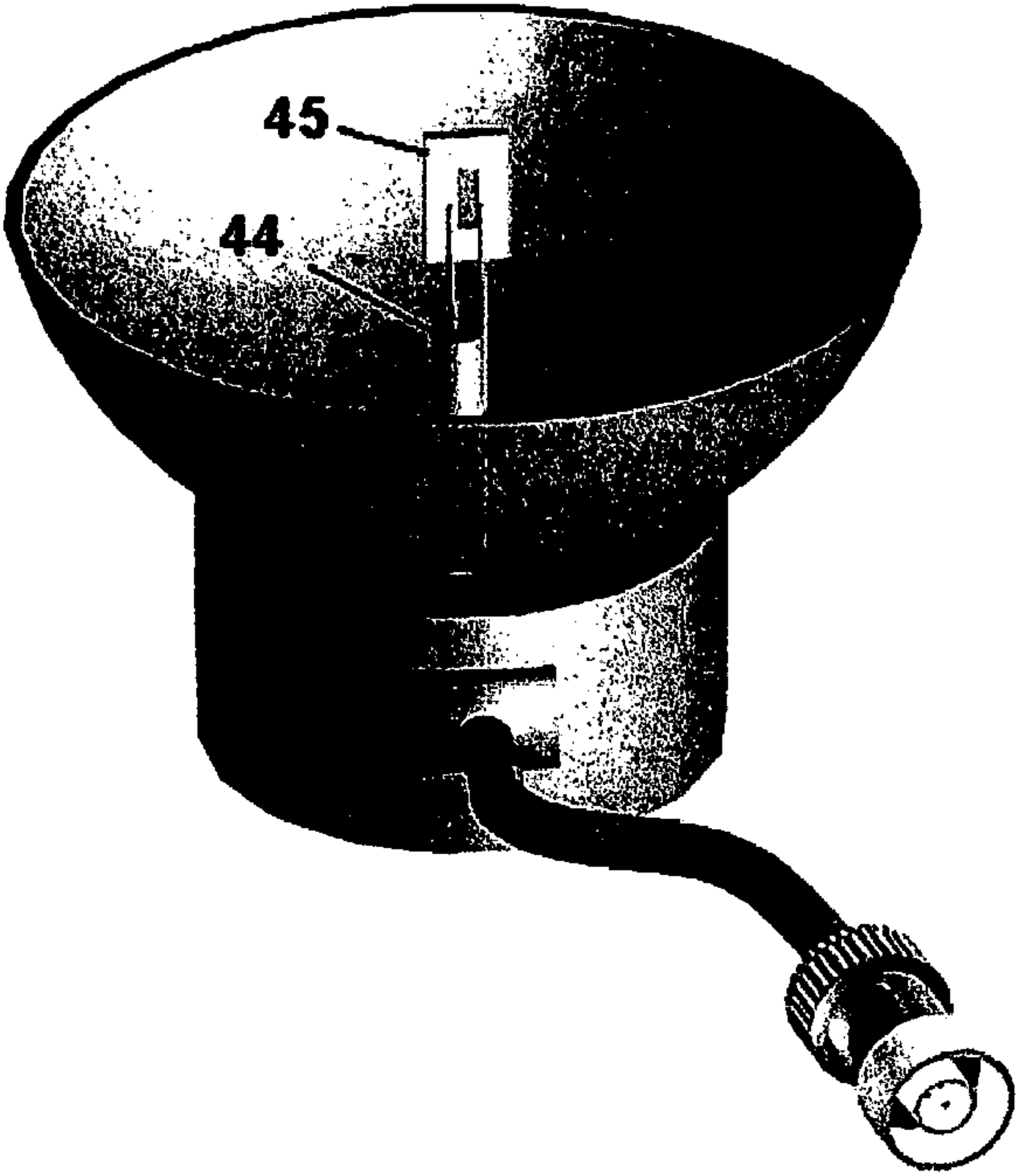

Figure 1

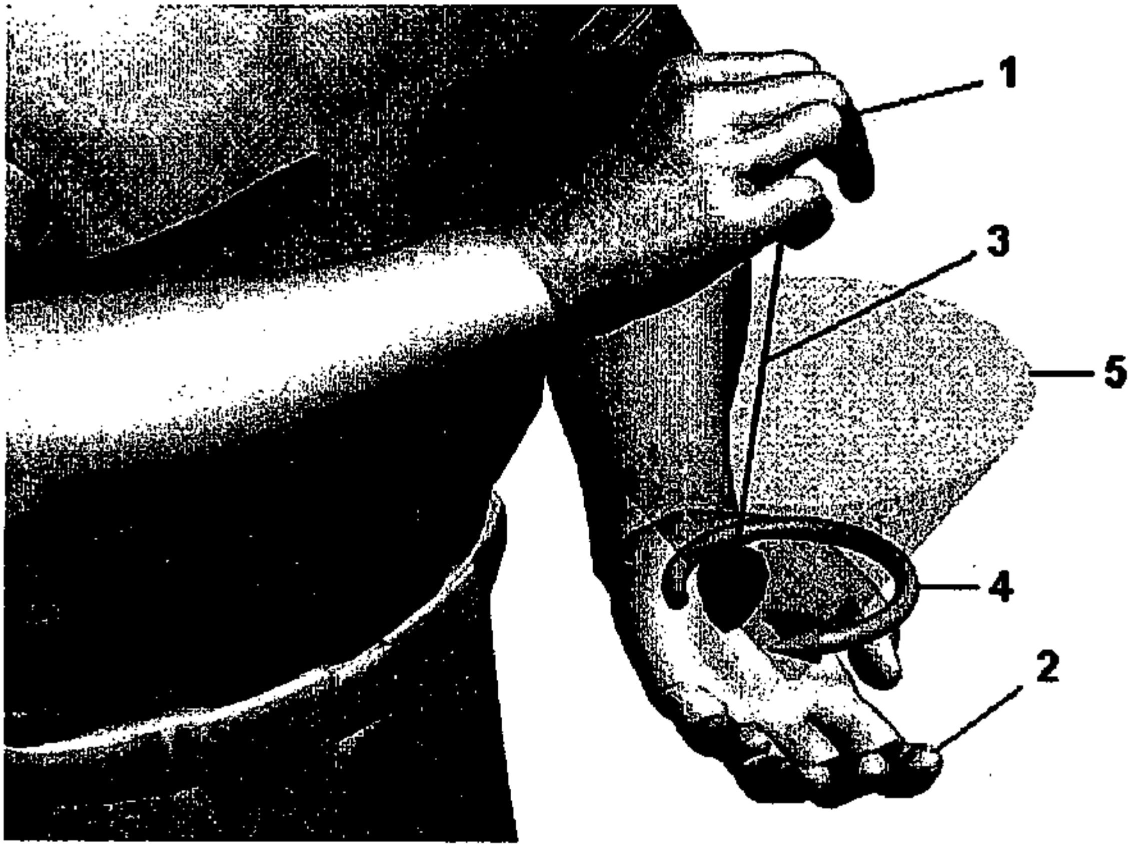

Figure 2

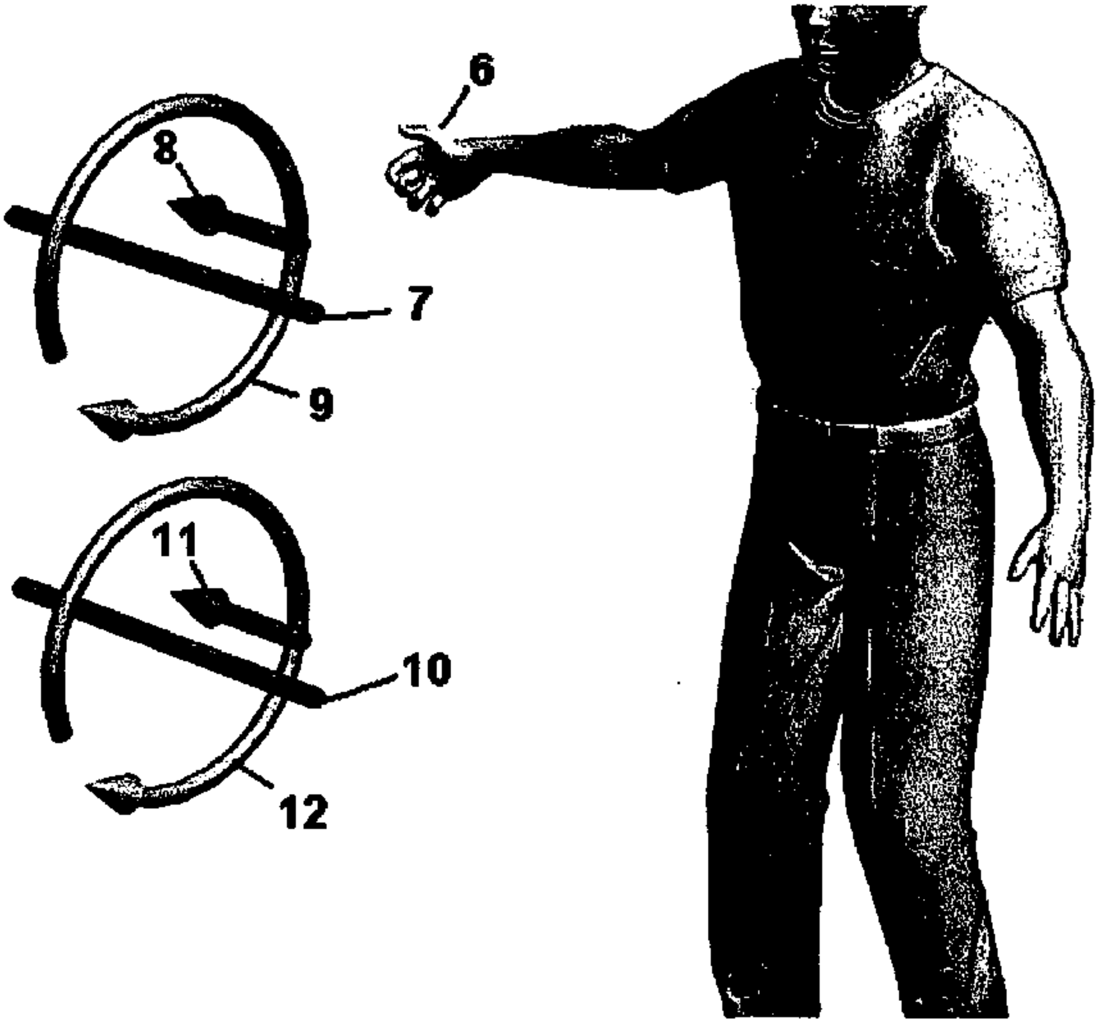
6
8
7
9
11
10
12

Figure 3

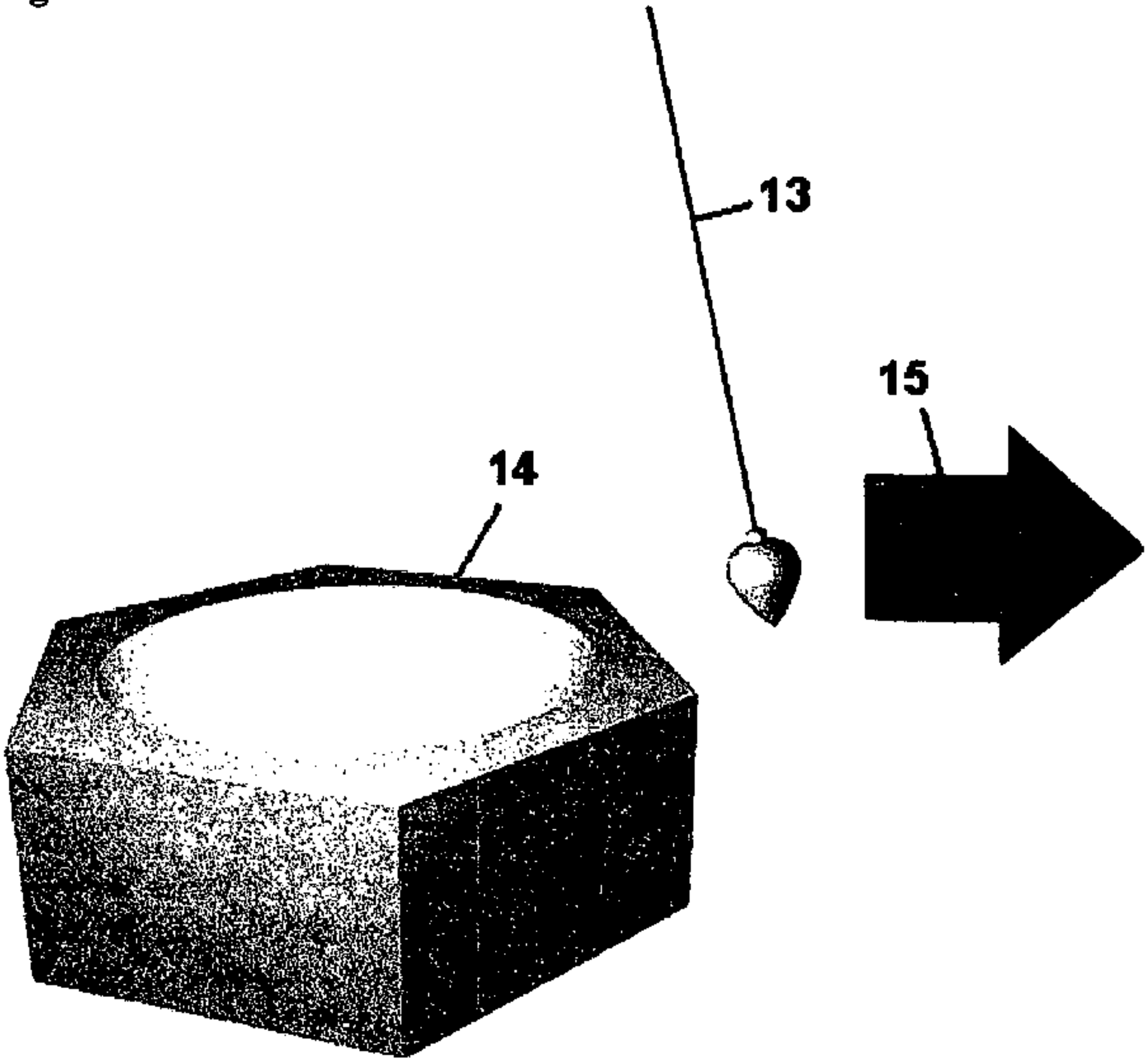

Figure 4

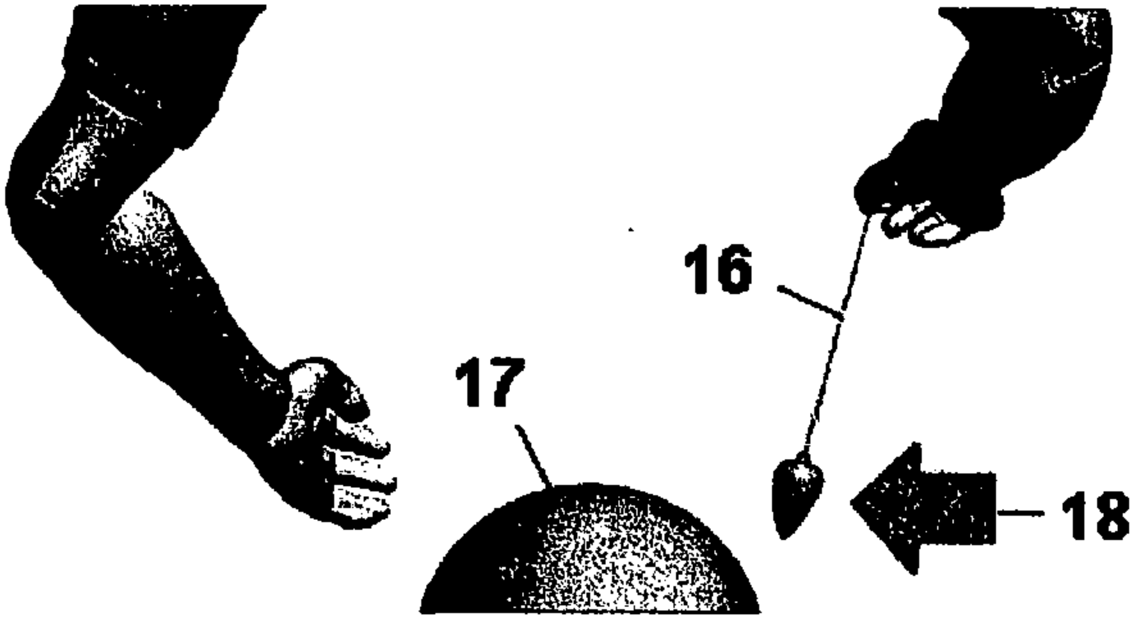

Figure 5

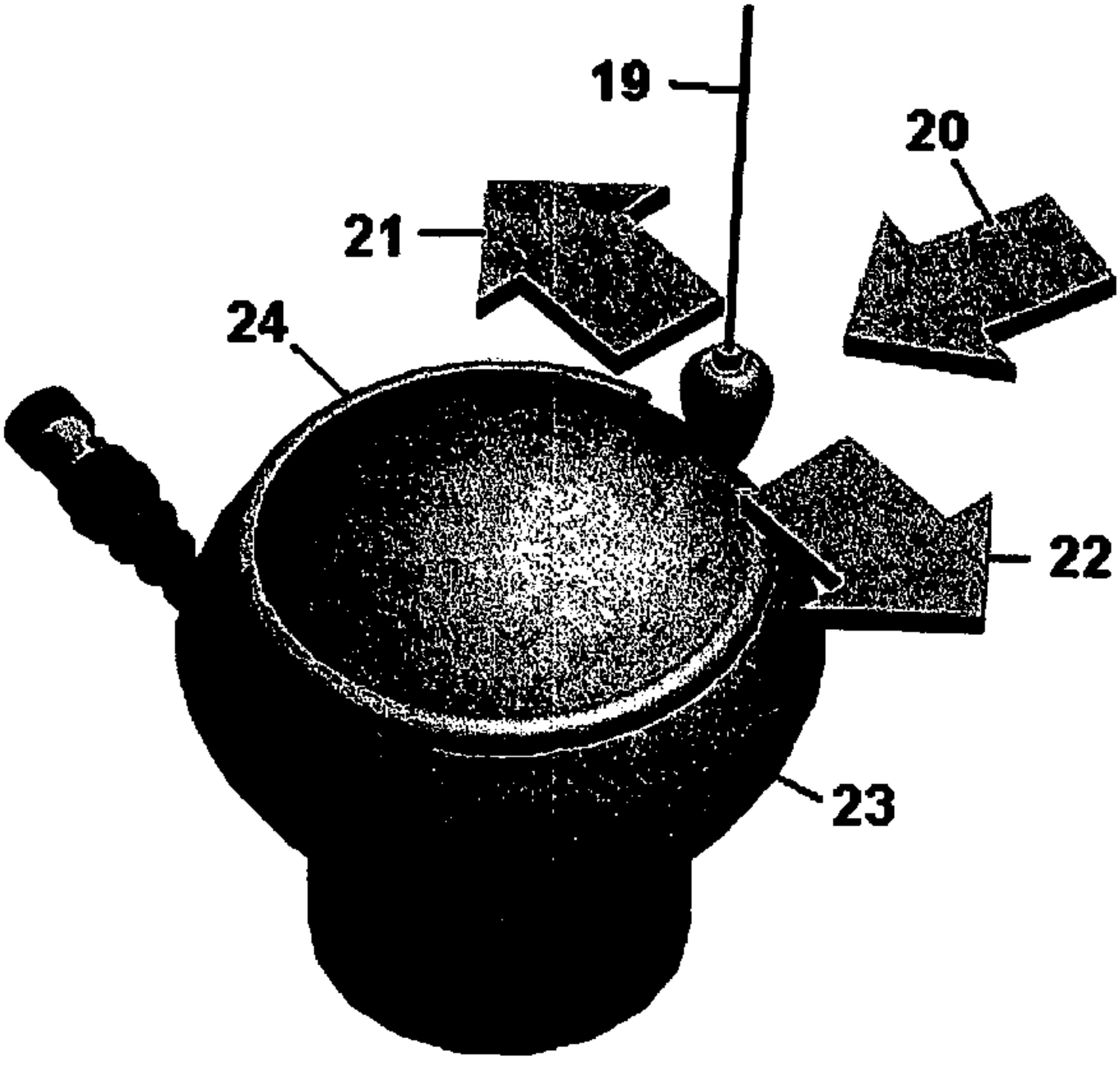
19
20
21
24
22
23

Figure 6

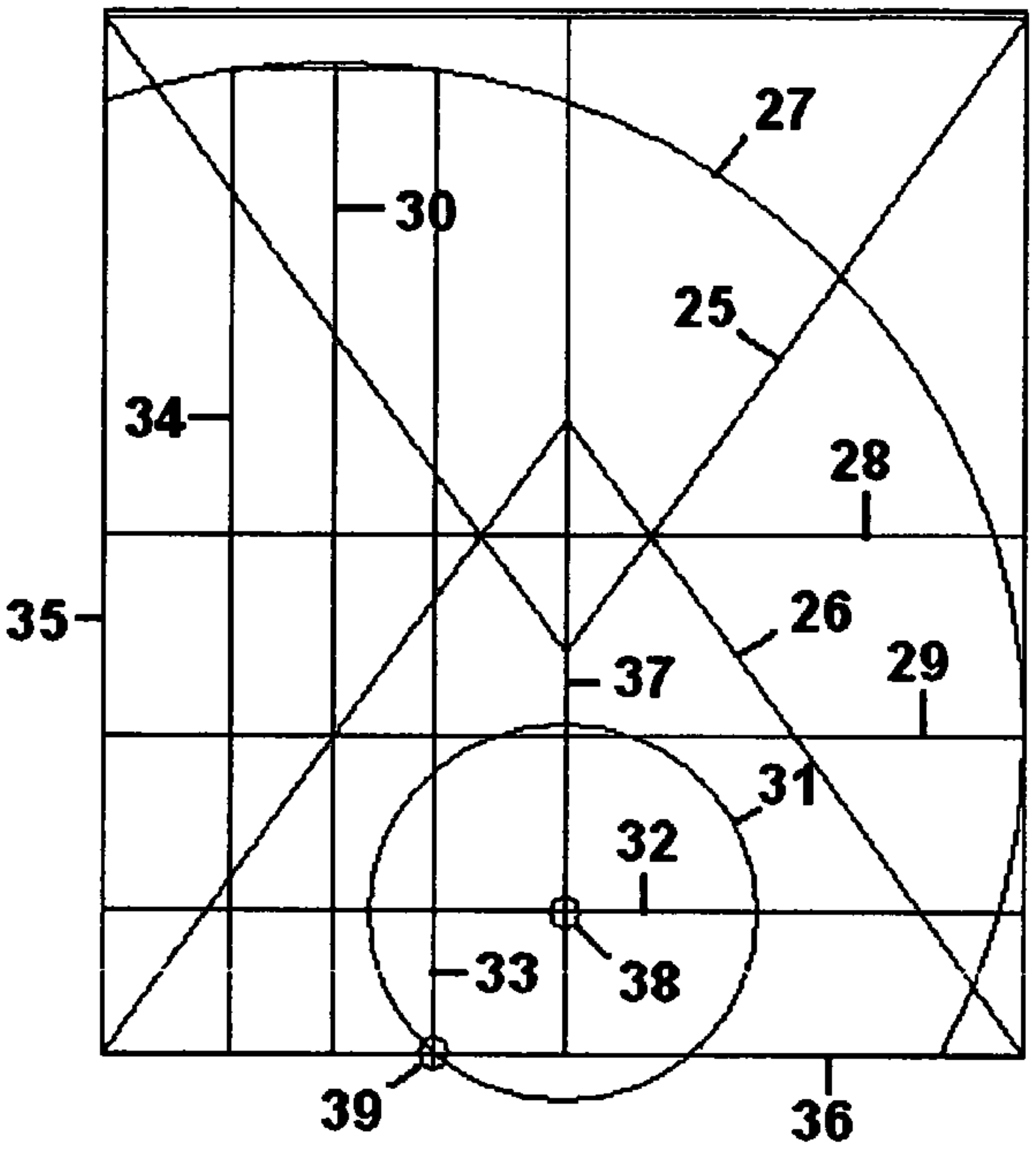
27
30
25
34
28
35
26
29
37
31
32
33
38
39
36

Figure 7

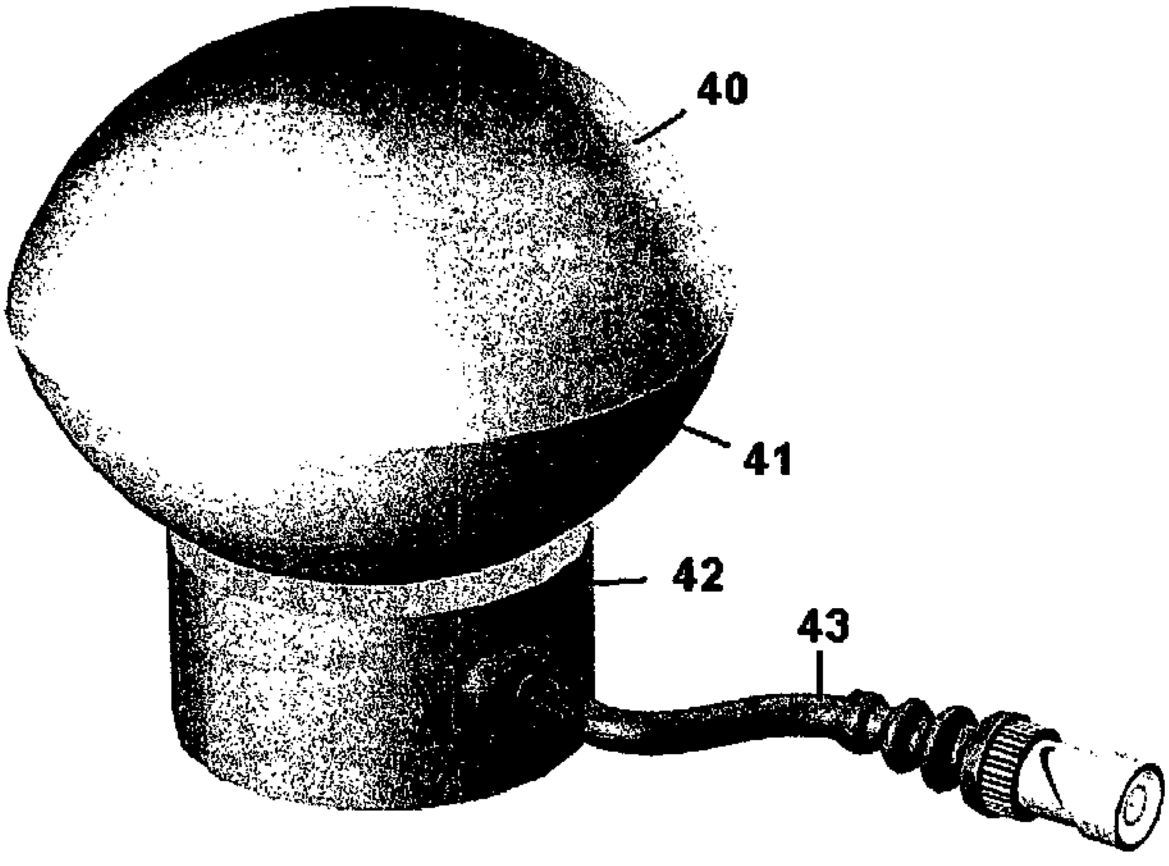

Figure 8

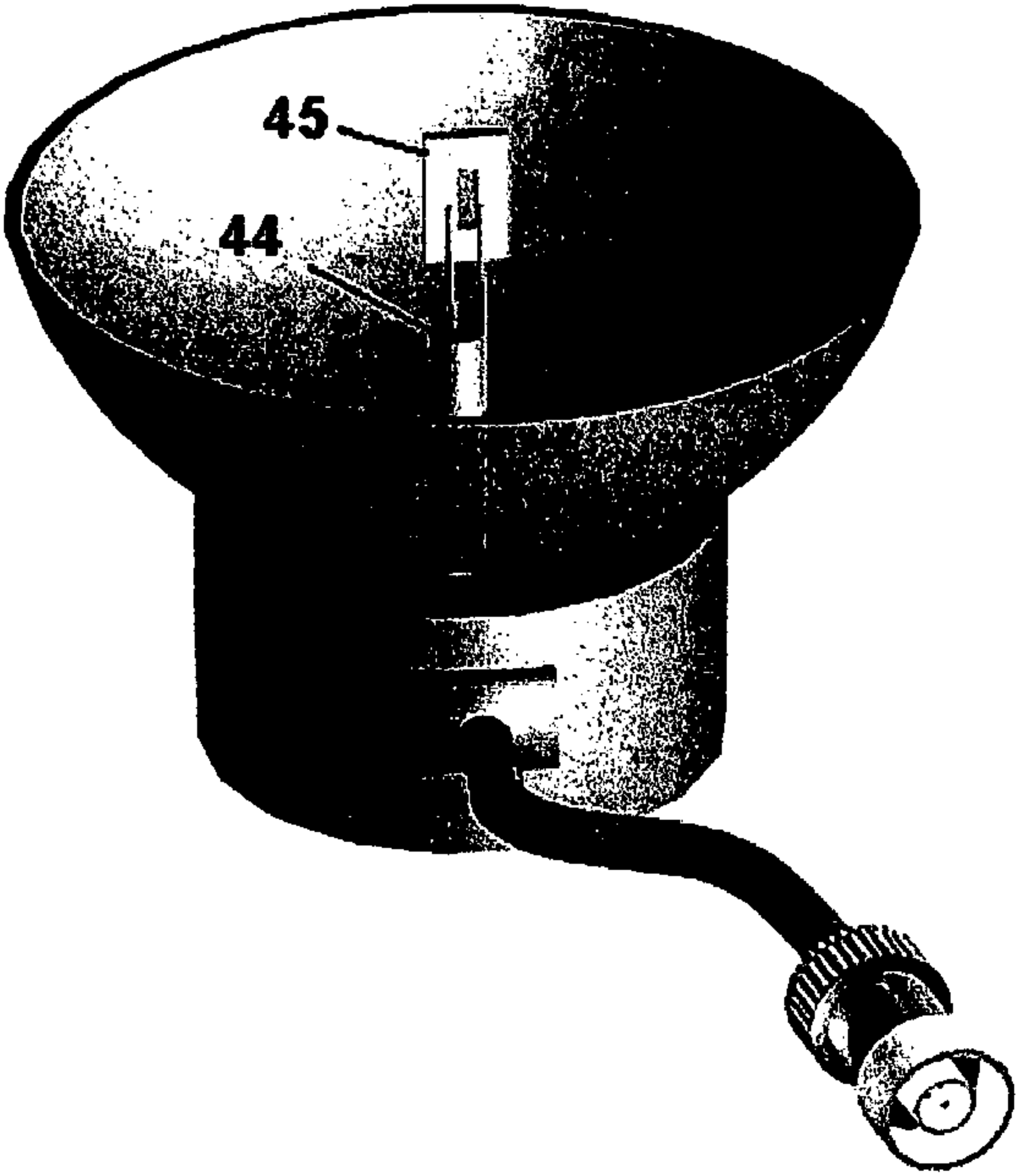

Figure 9

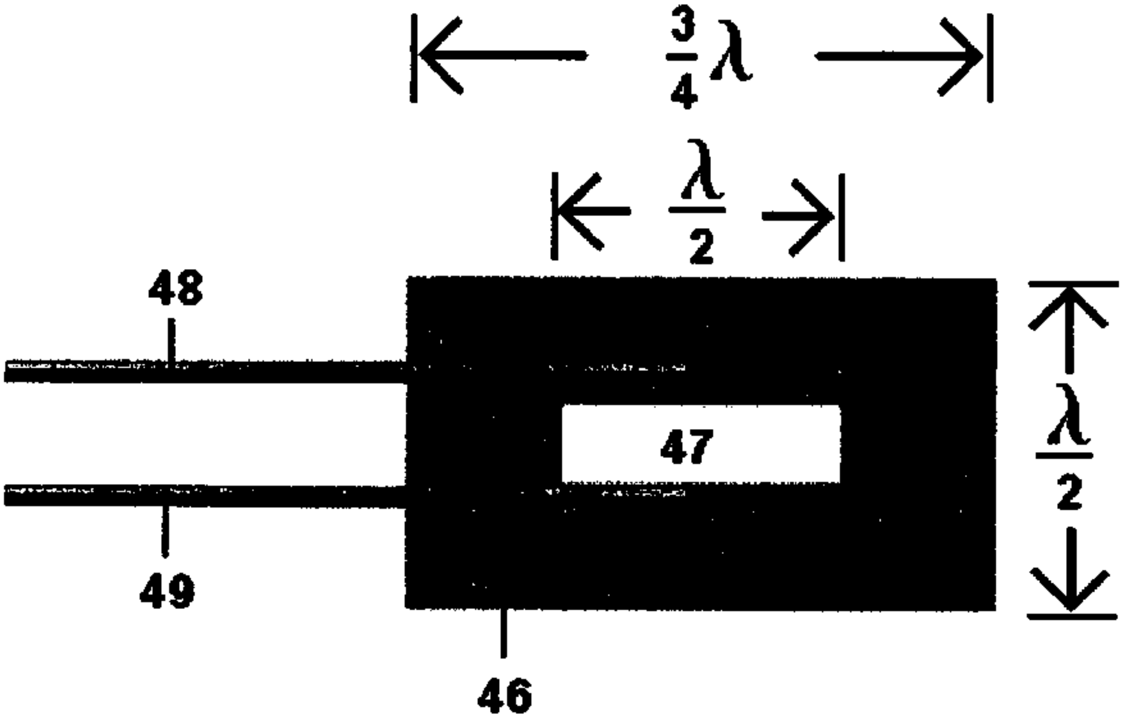
$\frac{3}{4}\lambda$
$\frac{\lambda}{2}$
48
47
$\frac{\lambda}{2}$
49
46

Figure 10

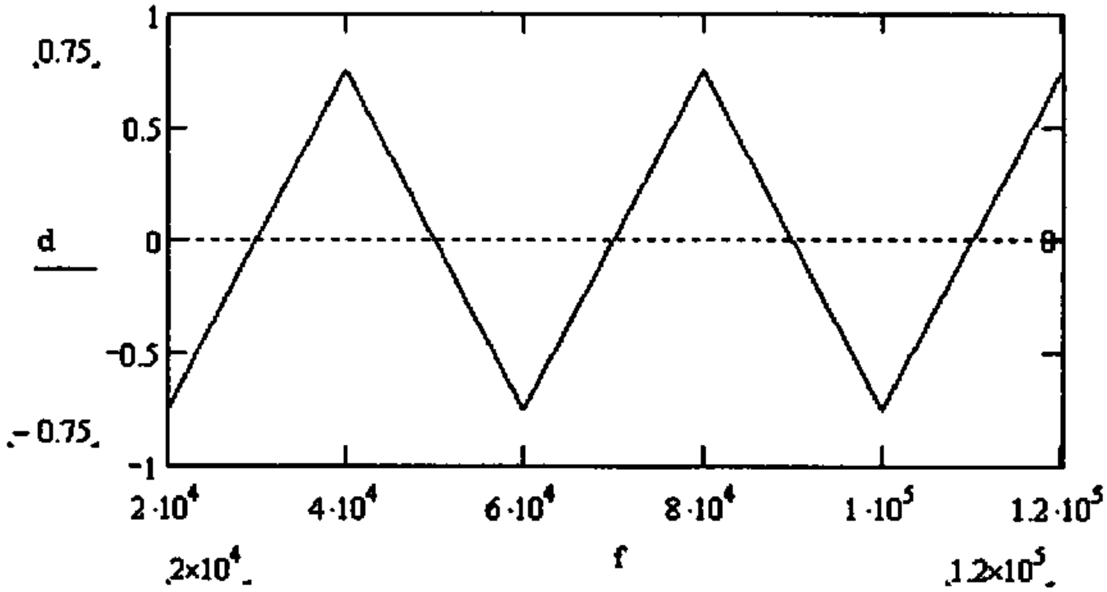
1
0.75
0.5
d
0
-0.5
-0.75
-1
2·10⁴
4·10⁴
6·10⁴
8·10⁴
1·10⁵
1.2·10⁵
2×10⁴
f
1.2×10⁵
0

Figure 11

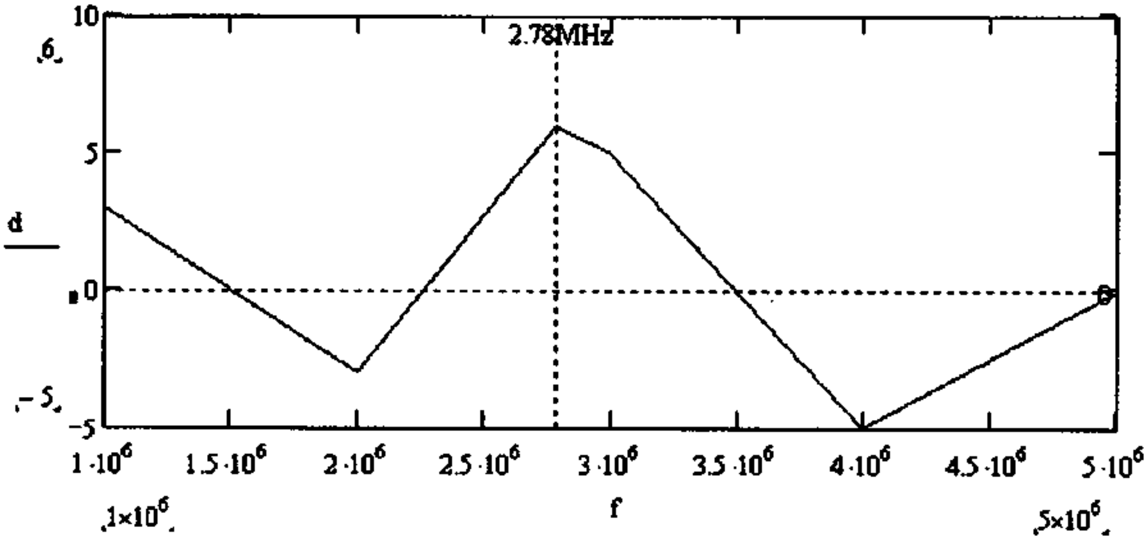
10
6
5
d
0
−5
−5
2.78MHz
1·10^6
1.5·10^6
2·10^6
2.5·10^6
3·10^6
3.5·10^6
4·10^6
4.5·10^6
5·10^6
1×10^6
f
5×10^6

Figure 12

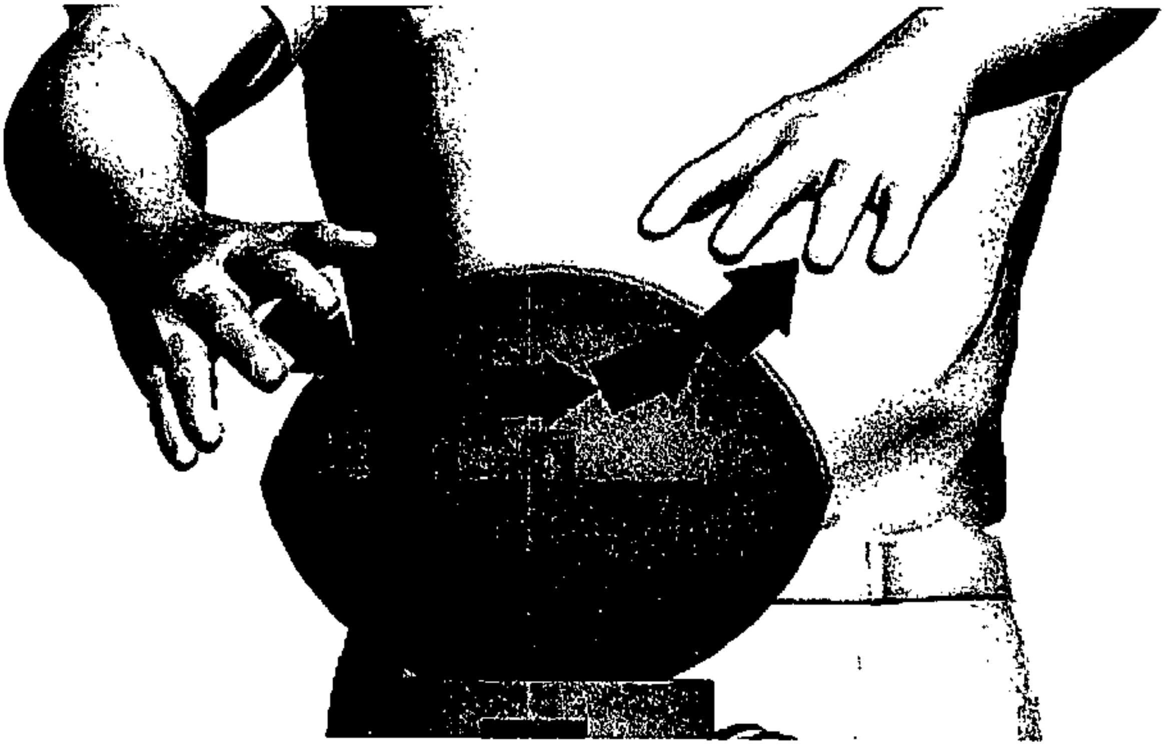

CHI ENERGY AMPLIFIER

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is an energy amplifier that controls the direction of the second co-gravitational K field by means of a slot antenna located in a resonating sphere. The hyperspace mass flow rate entering into this dimension is determined by the frequency of the antenna. This energy flow is used to increase the energy of the human energy field known as Chi.

BACKGROUND OF THE INVENTION

[0002] Referring to FIG. **1**, a pendulum suspended by a string (**3**) is held by the string with the right hand (**1**) over the upright palm of the left hand (**2**). The pendulum swings in circles shown by the clockwise direction of the arrow (**4**). The pendulum swings at a constant frequency at about 1 to 2 Hz. What this means is that there is a second gravitational K field (**5**) that can cause a mass to rotate in circles.

[0003] Everyone is familiar with the linear gravitational g field which is defined as Newton's gravitational constant G times the mass of the earth divided by the square of the radius of the earth. This is the field that causes objects to accelerate radially toward the earth's surface. Newton's gravitational constant G is equal to the speed of light c squared divided by the linear mass Ω of the universe. The speed of light is

$$c = 299792458 \frac{\text{meters}}{\text{second}}$$

$$\Omega = 1.346812891 \cdot 10^{27} \ \frac{\text{kilograms}}{\text{meter}}$$

$$G = \frac{c^2}{\Omega} = 6.673200002 \cdot 10^{-11} \frac{meter^3}{second^2 \ \text{kilogram}}$$

The mass and radius of the earth are

[0004]

$$ME = 5.977 \cdot 10^{24} \ \text{kilograms}$$

$$RE = 6371.03\ 10^3 \ \text{meters}$$

$$g = G\frac{ME}{RE^2} = 9.82 \ \frac{\text{meters}}{second^2}$$

[0005] From electromagnetism, the electric E field is linear and the B magnetic field is circular or forms closed loops. The electric field starts and ends on electric charges. Since there are two electromagnetic fields, it makes sense that there are also two gravitational fields. A flow of electric current through a straight wire causes a circular magnetic B field to form around the wire. In a similar manner, a flow of mass through a channel causes a circular gravitational K field to form around the channel.

[0006] Referring to FIG. **2**, the right hand rule (**6**) shows that if the flow is along the direction of the thumb of the right hand, then the field curls around in the direction of the fingers. If an electric current is flowing through the wire (**7**) in the direction of arrow (**8**), then a counter-clockwise magnetic field circles the wire shown by arrow (**9**). In the bottom view, a mass flowing through a channel (**10**) in direction (**11**) generates a counter-clockwise gravitational K field shown by arrow (**12**). This field is the reason that the pendulum swings in circles. The gravitational K field only exists if there is a mass flow. In the wire there is also an electric field driving the electrons, so it must be that there is a g gravitational field driving the mass along the channel.

[0007] The correspondence between electromagnetic and gravitational constants must be the following:

Electric	Gravitational
Charge q	Mass m
Electric Field E	Gravitational Field g
Magnetic Field B	Co-gravitational Field K
Linear Electric Charge λ	Linear Mass Ω
Permittivity of Space ϵ_0	$-1/4\pi G$
Permeability of Space μ_0	$-4\pi G/c^2$
Convection Current Density J	Mass Current Density J

[0008] Maxwell's equation for the curl or circulation of B involves the current density J and an electric field E changing with time t

$$\nabla \times B = \mu_0 J + \frac{1}{c^2}\frac{\partial E}{\partial t}$$

The corresponding gravitational equation is therefore

[0009]

$$\nabla \times K = -\frac{4\pi G}{c^2} J + \frac{1}{c^2}\frac{\partial g}{\partial t}$$

[0010] If the g field is constant with time, then the circulation of the K field only depends on the mass density flow through the channel. Notice the minus sign in front of the first term which says that if energy is entering our dimension through a channel, then the K field is in the counterclockwise direction as indicated by the pendulum. As seen in FIG. **1**, the energy is leaving through the left hand because the rotation is clockwise. On the right hand, the pendulum rotates in the counterclockwise direction indicating that energy is coming from the hand. Thus there is a flow of energy between the hands known as the Chi energy. The purpose of this invention is to amplify this energy.

[0011] The g gravitational field is due to a negative space-time curvature created by the mass of the earth. In many physics books this is depicted as a bowl. So the idea was to acquire a ceramic kiln which could produce ceramic bowls using earthenware clay and a bowl mold. The curvature created by the bowl, if any, could be detected by the pendulum.

[0012] Referring to FIG. **3**, a pendulum (**13**) held in the right hand with the left hand near the side of the concave mold (**14**) starts to swing toward the right, away from the plaster mold as shown by arrow (**15**).

[0013] Referring to FIG. **4**, a pendulum (**16**) held in the right hand over a positive curvature dome (**17**) swings toward the top of the dome as shown by arrow (**18**).

Supposedly the K gravitational field rotates in circles, but these experiments showed that the pendulum movement was radial.

[0014] Referring to FIG. **5**, the pendulum (**19**) swings radially (**20**) toward the top of the dome (**23**) because the forces (**24**) around the dome counteract each other as shown by arrows (**21, 22**). Thus there is no sideways pressure on the pendulum.

[0015] The next question was what is the value of the co-gravitational K constant? In electromagnetism the electric field is the speed of light times the magnetic field. So the g gravitational field should be the speed of light times the co-gravitational K field.

$$\begin{aligned} K &= \frac{g}{c} \\ &= \frac{1}{c}G\frac{MU}{RU^2} \\ &= \frac{1}{c}\frac{c^2}{\Omega}\frac{MU}{\frac{MU}{\Omega}RU} \\ &= \frac{c}{RU} \end{aligned}$$

$$\ln(K) = -40.34861143$$

[0016] where the radius RU of the universe is 10^{26} meters and c is the speed of light. The speed of light divided by the radius is angular frequency measured in Hz. Recall that the pendulum over the hand has a frequency f of 1 Hz which makes the angular frequency o equal to 2π radians per second since ω=2πf So the question is whether or not this value of K means anything in terms of hyperspace physics?

[0017] The universe has a geometrical sub-manifold based on the tetrahedron. All the physics constants of the universe are determined by this tetrahedral geometry. As shown in diagram tet0565 stored in the Library of Congress, the electron and proton are one and the same particle. There is a continuous clockwise path from the electron wavelength, through electric charge to electron mass and returning out-of-dimension back along a counterclockwise path as the proton. Because the paths are in opposite directions, the proton has a positive charge and the electron has the opposite negative charge. Because this path crosses into our dimension from hyperspace, we see two different particles. Thus Nature only has one particle. The tetrahedron diagram also shows that our dimension is offset from the origin of the diagram by the Cabibbo angle which is found in particle physics. Our dimension is determined by the Planck mass and the Planck wavelength which are the bottom limits of our dimension, known as the Planck box. If the proton mass is centered on the Planck mass, a tangent line to the circle can only be made by using the down quark and up quark which comprise the proton (duu). And the tangent line is drawn at the Cabibbo angle. So the diagram verifies some important experimental physics data.

[0018] Referring to FIG. **6**, the tetrahedron diagram is drawn with a vertical axis (**35**) corresponding to the natural logarithm of mass. The horizontal axis (**36**) is the natural logarithm of wavelength. The inverted tetrahedrons (**25, 26**) cross at the centerline (**28**) which is known through remote viewing as "the merging of two worlds." The centerline is the separation point between space and hyperspace. The circumscribing sphere (**27**) has a horizontal diameter (**29**) and a vertical diameter (**30**). The electron wavelength (**34**) reflects off the sphere and returns as the electron mass (**33**).

[0019] Because the K co-gravitational field is related to circular or spinning motion, as well as vorticity per area, it might be imagined that it has something to do with the spinning electron. The energy E of the electron is equal to the mass m_e of the electron times the speed of light c squared.

$$\begin{aligned} &m_e = 9.1093897\cdot 10^{-31}\ \text{kilogram} \\ &c = 299792458\cdot\frac{\text{meter}}{\text{second}} \\ &E = m_e c^2 \\ &\ln(E) = -30.13363019 \end{aligned}$$

[0020] The tips of the inverted tetrahedrons pass through the base constant, vertical line (**37**), which is equal to Planck's constant h divided by the speed of light.

$$h = 2.210260685\cdot 10^{-42}\ \text{kilogram meter}$$

$$\text{base} = \ln\left(\frac{h}{c}\right) = -95.91546344$$

[0021] On the tetrahedron diagram, the co-gravitational circle K (**31**) is centered on the base (**37**) at the electron energy (**32**) shown by the small π circle at the intersection (**38**). As can be seen, the K circle determines the electron mass (**33**) at the horizontal axis (**36**) as shown by intersection (**39**). Because the electron is the proton, it also sets the constants of the proton such as wavelength, charge and mass. Thus the K co-gravitational field determines our elementary particles. Because these particle paths go in and out of dimension, it means that hyperspace exists. And because the K field is circular, depending on the direction of the energy flow, another purpose of this invention is to control the direction of the field.

SUMMARY OF THE INVENTION

[0022] Referring to FIG. **7**, the Chi energy amplifier consists of two ceramic domes (**40, 41**) resting one on the other such as to form a hollow internal clam-like structure. The domes are mounted on a cylindrical ceramic base (**42**) containing a reinforced passageway for the coaxial cable and BNC connector (**43**). The BNC connector plugs into the frequency generator (not shown) which has a frequency range of 0.4 Hz to 5 MHz.

[0023] Referring to FIG. **8** with the upper dome removed, the coaxial cable (**44**) extends vertically through the base into the lower dome. The cable is soldered to a slot antenna (**45**). The center conductor of the cable is soldered to the left side of the slot and the ground shielding is soldered to the right side of the slot. Thus there is a voltage difference on the two sides which produces an oscillating electric field across the slot which radiates electromagnetic energy into the hollow dome.

[0024] A close-up of the slot antenna is shown in FIG. **9**. The antenna is a 0.050" thick copper sheet (**46**) with a width of ¾ wavelength by a height of ½ wavelength. A slot (**47**) of ½ wavelength is electric-discharge machined into the copper sheet (**47**). The center conductor (**48**) from the coaxial cable is soldered to the top of the slot. The ground wire (**49**) is soldered to the bottom of the slot. This configuration produces a voltage difference between the inner top and bottom surfaces of the slot. The frequency generator produces an oscillating voltage across the slot which generates an electric field that radiates from the antenna into the ceramic cavity.

[0025] The wavelength of the antenna and the dimensions of the ceramic cavity have to be tuned to the geometry of our dimension. Newton's gravitational constant G is equal to

$$\ln(G) = -23.4303342$$

$$f = e^{23.4303432} * \mathrm{Hz} = 15\ Ghz$$

$$\lambda = \frac{c}{f} = .020 \text{ meter} = .787 \text{ inch}$$

$$\text{slotwidth} = \frac{\lambda}{2} = .393 \text{ inch}$$

$$\text{antennawidth} = \frac{3}{4}\lambda = .590 \text{ inch}$$

The frequency of the antenna is the inverse of the gravitational constant. The reason for this is that when a circle of radius natural logarithm G is centered on the proton on the tetrahedron diagram, the circle intersects the corner of the Planck box which bounds our dimension between space and hyperspace. It is the low-density hyperspace energy that the amplifier will bring into this dimension.

[0026] The size of the ceramic cavity based on the angular frequency ω=2πf is

$$\frac{\frac{\omega^2}{c^2} + \frac{1}{ab}}{\frac{\omega^2}{c^2} + ab\left(\frac{\omega^2}{c^2} - \frac{1}{a^2}\right)\left(\frac{\omega^2}{c^2} - \frac{1}{b^2}\right)} = \frac{\tan\left(\frac{\omega}{c}(b-a)\right)}{\omega}\frac{c}{b-a}$$

where b is inside radius of the cavity, a is the radius of an object at the center of the cavity and c is the speed of light. Using the object radius as a=0.020 meter to account for the antenna, the inside radius of the cavity is b=4.25 inches which is the size of the 8.5 inch diameter ceramic dome using ¼ inch thick clay. The frequency is then equal to

$$f = \ln\left(\frac{\omega}{2\pi}\right) = 23.45$$

which is close to the inverse of the gravitational constant G of −23.43. Thus the amplifier is tuned to the gravitational constant of the universe both in size of the cavity and the frequency of the slot antenna. Using a SMD surface mount capacitor and inductor in the picofarad and nanohenry range, it is possible to get up to this high frequency, but it was found that it was not necessary because the amplifier works at lower frequencies that are pulsed.

[0027] Referring to FIG. **10**, the swinging movement of the pendulum measured in inches at the side of the dome is plotted against a range of frequencies from 20 kHz to 120 kHz. At 20 kHz, the pendulum swings ¾ inch away from the dome. Then at 40 kHz, the pendulum swings ¾ inch toward the dome. This pattern is repeated until 120 kHz. What this means is that the amplifier can control the direction of the co-gravitational field using frequency. The tetrahedron diagram shows that this phenomenon is related to the two axes of the circumscribing sphere of the tetrahedron.

[0028] Referring to FIG. **11**, the swinging of the pendulum is plotted for a range of frequencies from 1 MHz to 5 MHz. At a frequency of 2.78 MHz the pendulum was swinging strongly from the vertical to the horizontal, a distance of six inches. On the tetrahedron diagram, the distance from our base constant to the centerline between inverted tetrahedrons is equal to 14.838168. This corresponds to a frequency of

$$f=e^{14.838168}*\mathrm{Hz}=2.780574169\ \mathrm{MHz}$$

which is close to the experimental value.

[0029] Referring to FIG. **12**, the Chi energy flowing between the hands, shown by the arrows, is amplified by placing the hands across the dome. This energy from the right hand mixes with the hyperspace energy entering the dome from hyperspace. The combined energy is then absorbed in the left hand vortex. The effects of this amplified energy are simply amazing and have to be experienced to appreciate what it means.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0030] FIG. **1**. Perspective view of pendulum movement over left hand vortex.

[0031] FIG. **2**. Perspective view of right hand rule of physics showing similarity between electromagnetism and gravity.

[0032] FIG. **3**. Perspective view of plaster mold cavity showing negative curvature.

[0033] FIG. **4**. Perspective view of dome showing positive curvature.

[0034] FIG. **5**. Perspective view of dome showing radial pendulum movement.

[0035] FIG. **6**. Graph of co-gravitational K field on tetrahedron diagram.

[0036] FIG. **7**. Perspective view of Chi Energy Amplifier with BNC coaxial cable.

[0037] FIG. **8**. Perspective view of slot antenna in dome.

[0038] FIG. **9**. Perspective view of slot antenna showing dimensions.

[0039] FIG. **10**. Graph of pendulum movement versus frequency of antenna, 20 KHz.

[0040] FIG. **11**. Graph of pendulum movement versus frequency from 1 MHz to 5 MHz.

[0041] FIG. **12**. Perspective view of amplifying Chi energy.

DETAILED DESCRIPTION OF THE INVENTION

[0042] The dome of the amplifier is made of red earthenware clay from Minnesota having a cone 06 kiln firing temperature of 1828° F. over a period of 7.5 hours. The clay is placed between two hardwood ¼ inch thick slats. It is then rolled flat with a rolling pin. The sheet of clay is cut in half and one half is placed in a 9-inch diameter bowl mold using a soft sponge to push it into position. The second half is then added to the first with the seam between worked flat with a metal kidney and elephant ear sponge. The top of the mold is scrapped flat to create the bowl shape. The moist clay is left to dry for one day after which it falls out of the mold due to shrinkage. After completely drying, the dome is then placed in a kiln which runs the computer-controlled cone 06 temperature firing profile. A preheat at 180° F. for one hour is required to make sure the dome is completely bone dry. It takes 7 to 8 hours to fire the dome with another 12 hours to cool down naturally.

[0043] The base is also made of ¼ inch clay. A circular shape is cut out and fitted with a ribbon of extruded clay from a clay gun to make the cylindrical wall. Using a hole cutter made of hollow tubing, a hole is made in the side of the base for the coaxial cable. The second lower dome has a hole cut in the bottom when the clay is leather hard.

[0044] A 0.050" thick copper sheet is EDM machined to cut a half wavelength slot in the copper. The inner conductor of the coaxial cable is soldered using silver epoxy to the one side of the slot. The ground shielding of the cable is silver epoxied to the other side of the slot. The other end of the coaxial cable has a BNC connector which plugs into the frequency generator.

[0045] The pendulum is made out of soft brass rod and machined on a lathe into a plumb bob shape. A hole is drilled in the stem to hold the 6" long string.

I claim:

1. A Chi energy amplifier comprising:

a. a hollow ceramic dome in the shape of a clam shell with a hole in the lower half that provides access to a coaxial cable;

b. a cylindrical ceramic base on which the dome is mounted with holes that provide access for a coaxial cable into item (a);

c. a slot antenna soldered to the coaxial cable, located midway inside item (a), such that the inner conductor of the cable is soldered to one side of the slot and the ground shielding is soldered to the other side of the slot;

d. a frequency generator connected to the other end of the coaxial cable by means of a BNC connector for the purpose of providing an oscillating voltage to item (c);

e. an electromagnetic wave generated by items (c, d) that resonates inside item (a);

f. a co-gravitational K field generated by items (a, b, c, d, e) whose direction can be controlled by the frequency of item (d); and

g. a flow of hyperspace energy into or out of the dome due to item (f).

* * * * *

455-619 AU 233 EX
FIP8106 OR 685,958

55

No. 685,958. **Patented Nov. 5, 1901.**

N. TESLA.
METHOD OF UTILIZING RADIANT ENERGY.

(Application filed Mar. 21, 1901.)

(No Model.)

Electric steping motor energized by corpuscular energy from sun

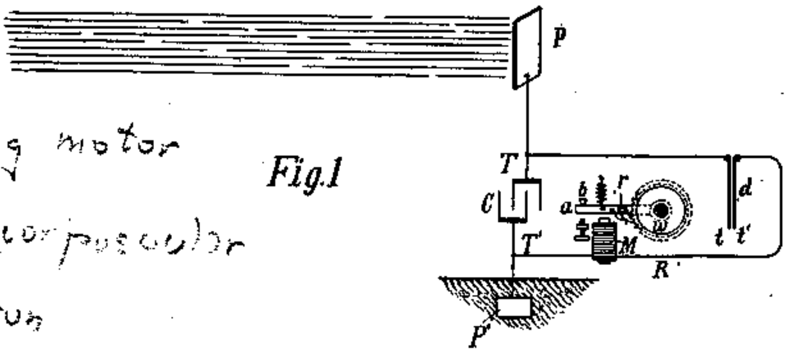

Fig.1

No refs.

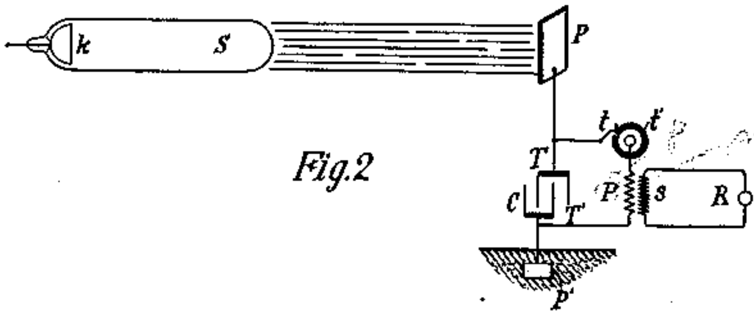

Fig.2

Witnesses:
Raphaël Netter
M. Lawson Dyer

Nikola Tesla, Inventor
by Kerr, Page & Cooper Attys

THE NORRIS PETERS CO., PHOTO-LITHO., WASHINGTON, D. C.

UNITED STATES PATENT OFFICE.

NIKOLA TESLA, OF NEW YORK, N. Y.

METHOD OF UTILIZING RADIANT ENERGY.

SPECIFICATION forming part of Letters Patent No. 685,958, dated November 5, 1901.

Application filed March 21, 1901. Serial No. 52,154. (No model.)

To all whom it may concern:

Be it known that I, NIKOLA TESLA, a citizen of the United States, residing at the borough of Manhattan, in the city, county, and State of New York, have invented certain new and useful Improvements in Methods of Utilizing Radiant Energy, of which the following is a specification, reference being had to the drawings accompanying and forming a part of the same.

It is well known that certain radiations—such as those of ultra-violet light, cathodic, Roentgen rays, or the like—possess the property of charging and discharging conductors of electricity, the discharge being particularly noticeable when the conductor upon which the rays impinge is negatively electrified. These radiations are generally considered to be ether vibrations of extremely small wave lengths, and in explanation of the phenomena noted it has been assumed by some authorities that they ionize or render conducting the atmosphere through which they are propagated. My own experiments and observations, however, lead me to conclusions more in accord with the theory heretofore advanced by me that sources of such radiant energy throw off with great velocity minute particles of matter which are strongly electrified, and therefore capable of charging an electrical conductor, or even if not so may at any rate discharge an electrified conductor either by carrying off bodily its charge or otherwise.

My present application is based upon a discovery which I have made that when rays or radiations of the above kind are permitted to fall upon an insulated conducting body connected to one of the terminals of a condenser, while the other terminal of the same is made by independent means to receive or to carry away electricity, a current flows into the condenser so long as the insulated body is exposed to the rays, and under the conditions hereinafter specified an indefinite accumulation of electrical energy in the condenser takes place. This energy after a suitable time interval, during which the rays are allowed to act, may manifest itself in a powerful discharge, which may be utilized for the operation or control of mechanical or electrical devices or rendered useful in many other ways.

In applying my discovery I provide a condenser, preferably of considerable electrostatic capacity, and connect one of its terminals to an insulated metal plate or other conducting body exposed to the rays or streams of radiant matter. It is very important, particularly in view of the fact that electrical energy is generally supplied at a very slow rate to the condenser, to construct the same with the greatest care. I use by preference the best quality of mica as dielectric, taking every possible precaution in insulating the armatures, so that the instrument may withstand great electrical pressures without leaking and may leave no perceptible electrification when discharging instantaneously. In practice I have found that the best results are obtained with condensers treated in the manner described in a patent granted to me February 23, 1897, No. 577,671. Obviously the above precautions should be the more rigorously observed the slower the rate of charging and the smaller the time interval during which the energy is allowed to accumulate in the condenser. The insulated plate or conducting body should present as large a surface as practicable to the rays or streams of matter, I having ascertained that the amount of energy conveyed to it per unit of time is under otherwise identical conditions proportionate to the area exposed, or nearly so. Furthermore, the surface should be clean and preferably highly polished or amalgamated. The second terminal or armature of the condenser may be connected to one of the poles of a battery or other source of electricity or to any conducting body or object whatever of such properties or so conditioned that by its means electricity of the required sign will be supplied to the terminal. A simple way of supplying positive or negative electricity to the terminal is to connect the same either to an insulated conductor, supported at some height in the atmosphere, or to a grounded conductor, the former, as is well known, furnishing positive and the latter negative electricity. As the rays or supposed streams of matter generally convey a positive charge to the first condenser-terminal, which is connect-

RADIANT ENERGY.

EXAMINER

ed to the plate or conductor above mentioned, I usually connect the second terminal of the condenser to the ground, this being the most convenient way of obtaining negative electricity, dispensing with the necessity of providing an artificial source. In order to utilize for any useful purpose the energy accumulated in the condenser, I furthermore connect to the terminals of the same a circuit including an instrument or apparatus which it is desired to operate and another instrument or device for alternately closing and opening the circuit. This latter may be any form of circuit-controller, with fixed or movable parts or electrodes, which may be actuated either by the stored energy or by independent means.

The rays or radiations which are to be utilized for the operation of the apparatus above described in general terms may be derived from a natural source, as the sun, or may be artificially produced by such means, for example, as an arc-lamp, a Roentgen tube, and the like, and they may be employed for a great variety of useful purposes.

My discovery will be more fully understood from the following detailed description and annexed drawings, to which reference is now made, and in which—

Figure 1 is a diagram showing typical forms of the devices or elements as arranged and connected in applying the method for the operation of a mechanical contrivance or instrument solely by the energy stored; and Fig. 2 is a diagrammatical representation of a modified arrangement suitable for special purposes, with a circuit-controller actuated by independent means.

Referring to Fig. 1, C is the condenser, P the insulated plate or conducting body, which is exposed to the rays, and P′ another plate or conductor, all being joined in series, as shown. The terminals T T′ of the condenser are also connected to a circuit including a receiver R, which is to be operated, and a circuit-controlling device *d*, which in this case is composed of two very thin conducting-plates *t t′*, placed in close proximity and very mobile, either by reason of extreme flexibility or owing to the charater of their support. To improve their action, they should be inclosed in a receptacle from which the air may be exhausted. The receiver R is shown as consisting of an electromagnet M, a movable armature *a*, a retractile spring *b*, and a ratchet-wheel *w*, provided with a spring-pawl *r*, which is pivoted to armature *a*, as illustrated. The apparatus being arranged as shown, it will be found that when the radiations of the sun or of any other source capable of producing the effects before described fall upon the plate P an accumulation of electrical energy in the condenser C will result. This phenomenon, I believe, is best explained as follows: The sun as well as other sources of radiant energy throw off minute particles of matter positively electrified, which, impinging upon the plate P, communicate an electrical charge to the same. The opposite terminal of the condenser being connected to the ground, which may be considered as a vast reservoir of negative electricity, a feeble current flows continuously into the condenser, and inasmuch as these supposed particles are of an inconceivably small radius or curvature, and consequently charged to a relatively very high potential, this charging of the condenser may continue, as I have found in practice, almost indefinitely, even to the point of rupturing the dielectric. Obviously whatever circuit-controller be employed it should operate to close the circuit in which it is included when the potential in the condenser has reached the desired magnitude. Thus in Fig. ② when the electrical pressure at the terminals T T′ rises to a certain predetermined value the plates *t t′*, attracting each other, close the circuit connected to the terminals. This permits a flow of current which energizes the magnet M, causing it to draw down the armature *a* and impart a partial rotation to the ratchet-wheel *w*. As the current ceases the armature is retracted by the spring *b* without, however, moving the wheel *w*. With the stoppage of the current the plates *t t′* cease to be attracted and separate, thus restoring the circuit to its original condition.

Many useful applications of this method of utilizing the radiations emanating from the sun or other source and many ways of carrying out the same will at once suggest themselves from the above description. By way of illustration a modified arrangement is shown in Fig. 2, in which the source S of radiant energy is a special form of Roentgen tube devised by me having but one terminal *k*, generally of aluminium, in the form of half a sphere with a plain polished surface on the front side, from which the streams are thrown off. It may be excited by attaching it to one of the terminals of any generator of sufficiently-high electromotive force; but whatever apparatus be used it is important that the tube be exhausted to a high degree, as otherwise it might prove entirely ineffective. The working or discharge circuit connected to the terminals T T′ of the condenser includes in this case the primary *p* of a transformer and a circuit-controller comprising a fixed terminal or brush *t* and a movable terminal *t′* in the shape of a wheel with conducting and insulating segments which may be rotated at an arbitrary speed by any suitable means. In inductive relation to the primary wire or coil *p* is a secondary *s*, usually of a much greater number of turns, to the ends of which is connected a receiver R. The terminals of the condenser being connected as indicated, one to an insulated plate P and the other to a grounded plate P′, when the tube S is excited rays or streams of matter are emitted from the same, which convey a positive charge to the plate P and condenser-terminal T, while terminal T′ is continuously receiving negative electricity from the plate

P'. This, as before explained, results in an accumulation of electrical energy in the condenser, which goes on as long as the circuit including the primary p is interrupted. Whenever the circuit is closed, owing to the rotation of the terminal t', the stored energy is discharged through the primary p, this giving rise in the secondary s to induced currents which operate the receiver R.

It is clear from what has been stated above that if the terminal T' is connected to a plate supplying positive instead of negative electricity the rays should convey negative electricity to plate P. The source S may be any form of Roentgen or Lenard tube; but it is obvious from the theory of action that in order to be very effective the electrical impulses exciting it should be wholly or at least preponderatingly of one sign. If ordinary symmetrical alternating currents are employed, provision should be made for allowing the rays to fall upon the plate P only during those periods when they are productive of the desired result. Evidently if the radiations of the source be stopped or intercepted or their intensity varied in any manner, as by periodically interrupting or rythmically varying the current exciting the source, there will be corresponding changes in the action upon the receiver R, and thus signals may be transmitted and many other useful effects produced. Furthermore, it will be understood that any form of circuit-closer which will respond to or be set in operation when a predetermined amount of energy is stored in the condenser may be used in lieu of the device specifically described with reference to Fig. 1, and also that the special details of construction and arrangement of the several parts of the apparatus may be very greatly varied without departure from the invention.

Having described my invention, what I claim is—

1. The method of utilizing radiant energy, which consists in charging one of the armatures of a condenser by rays or radiations, and the other armature by independent means, and discharging the condenser through a suitable receiver, as set forth.

2. The method of utilizing radiant energy, which consists in simultaneously charging a condenser by means of rays or radiations and an independent source of electrical energy, and discharging the condenser through a suitable receiver, as set forth.

3. The method of utilizing radiant energy, which consists in charging one of the armatures of a condenser by rays or radiations, and the other by independent means, controlling the action or effect of said rays or radiations and discharging the condenser through a suitable receiver, as set forth.

4. The method of utilizing radiant energy, which consists in charging one of the armatures of a condenser by rays or radiations and the other by independent means, varying the intensity of the said rays or radiations and periodically discharging the condenser through a suitable receiver, as set forth.

5. The method of utilizing radiant energy, which consists in directing upon an elevated conductor, connected to one of the armatures of a condenser, rays or radiations capable of positively electrifying the same, carrying off electricity from the other armature by connecting the same with the ground, and discharging the accumulated energy through a suitable receiver, as set forth.

6. The method of utilizing radiant energy, which consists in charging one of the armatures of a condenser by rays or radiations, and the other by independent means, and effecting by the automatic discharge of the accumulated energy the operation or control of a suitable receiver, as set forth.

NIKOLA TESLA.

Witnesses:
M. LAWSON DYER,
RICHARD DONOVAN.

New Energy Technologies

Published by Faraday Laboratories Ltd. Issue #3 (6) 2002

Research on advanced propulsion systems and new energy sources

Time Control experiments

Also in this issue:

* New experiments: magnet monopole beam cures cancer
* Electrogravity experiments
* High efficiency ceramic heater
* New aether experiments
* Instantaneous telecommunication

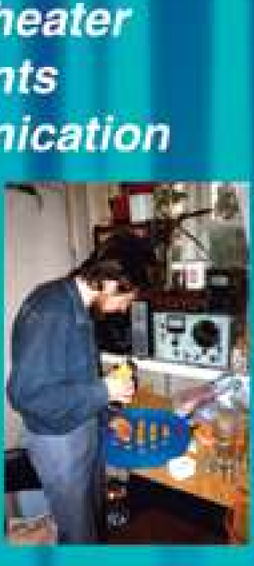

Editorial

Dr. Chernobrov and NET editor Olga Leontyeva.

The head of the Moscow research group "Kosmopoisk" Dr. Vadim A. Chernobrov (our readers certainly remember his articles on experiments with time) has recently visited the editorial office of New Energy Technologies in Saint-Petersburg, Russia. His visit was caused by discussion on the Time Machine project (investigation of active properties of time), started under agreement with Faraday Labs Ltd. Our magazine will inform readers on the topic in later issues.

NEW ENERGY TECHNOLOGIES
New Book

Dear Reader!
All articles of the previous 6 NET issues will soon be published as a new book in English (400 pages, more than 100 photos, figures and drawings, format A4).
You can order by e-mail: book2001@faraday.ru
or fax 7-812-3806564
Price with airmail delivery is 83 USD.

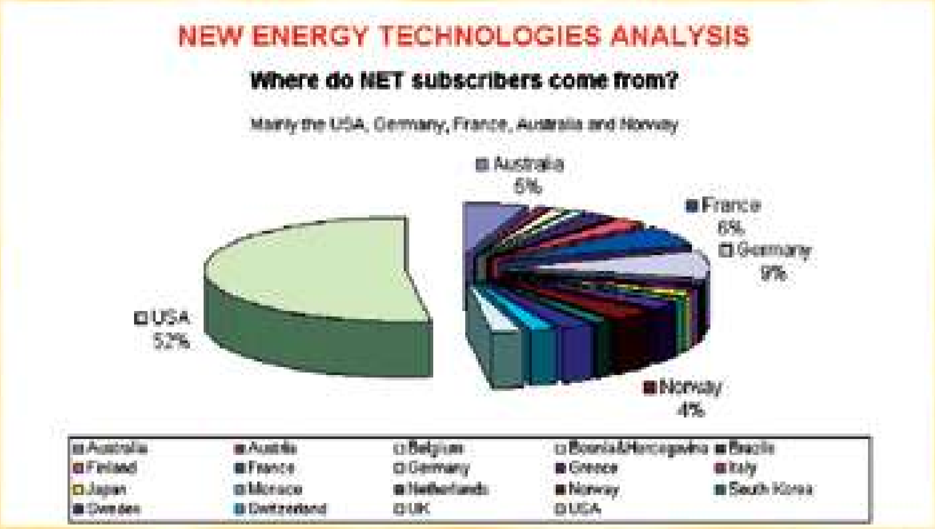

Publisher Faraday Lab Ltd

Editor-in-Chief Alexander V. Frolov
Editor Olga O. Leontyeva
Designer Ekaterina F. Yakovleva
Translator Antonina V. Nikolaeva

Scientific Advisory Board
Kyrill P. Butusov
Ivan M. Shakhparonov

New Energy Technologies Magazine
is published by Faraday Lab Ltd
six times a year

Correspondence Address:
POBox 37, St.Petersburg,
193024, Russia
Phone/fax: 7-812-380-6564
net@faraday.ru
www.faraday.ru

Subscription: $90 any foreign
Air mail delivery included

PRINTED IN RUSSIA
Copyright © 2002 by Faraday Lab Ltd.
All rights reserved.

LUTEC 1000

News from Australia

Magnet motor-generator

www.lutec.com.au

"There are a couple of other major factors that we won't go into here, suffice it to say that our current prototype demonstrates 1500% more " out" than "in"... "

"The Lutec 1000 will be sold throughout Australia within three years. The majority of buyers are expected to be the home owners, although there is strong support for our product throughout the entire strata of society and business. Market leadership should be maintained for the next five years. The product is sold through commission only sales agents and our staff at the plant, mainly to anyone who needs or uses electricity."

John Christie (left) is the co-founder, CEO and chairman of Lutec (Aust.) Pty Ltd.

Ludwig (Lou) Brits (right, above) is the co-founder, managing director and head of development of Lutec (Aust.) Pty Ltd. He is Chairman and director of TTI Pty Ltd. TTI Pty Ltd is a state government recognised R.A.P. (remote area power) system provider.

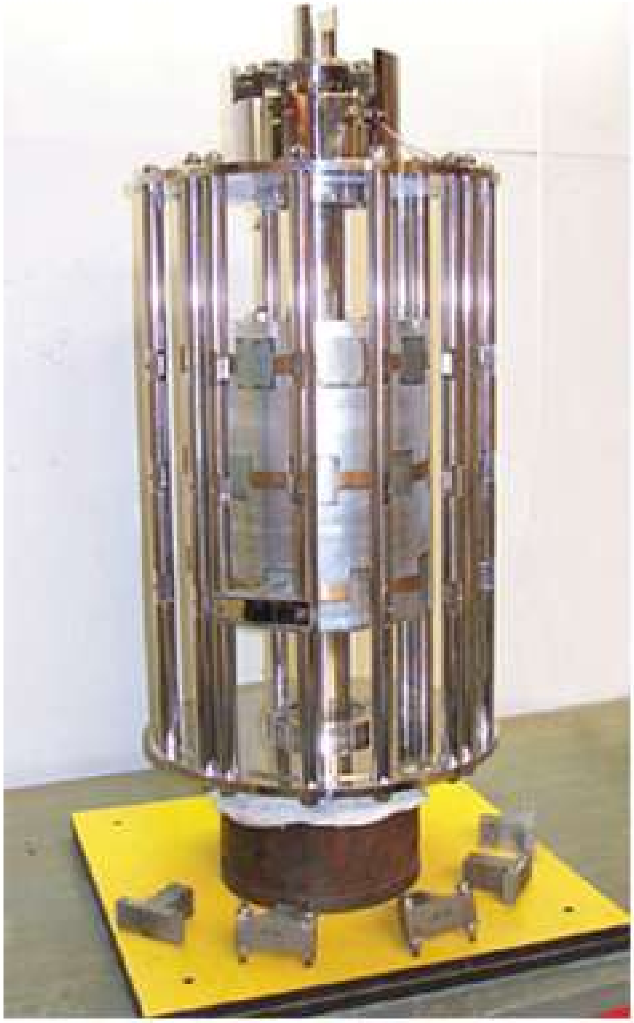

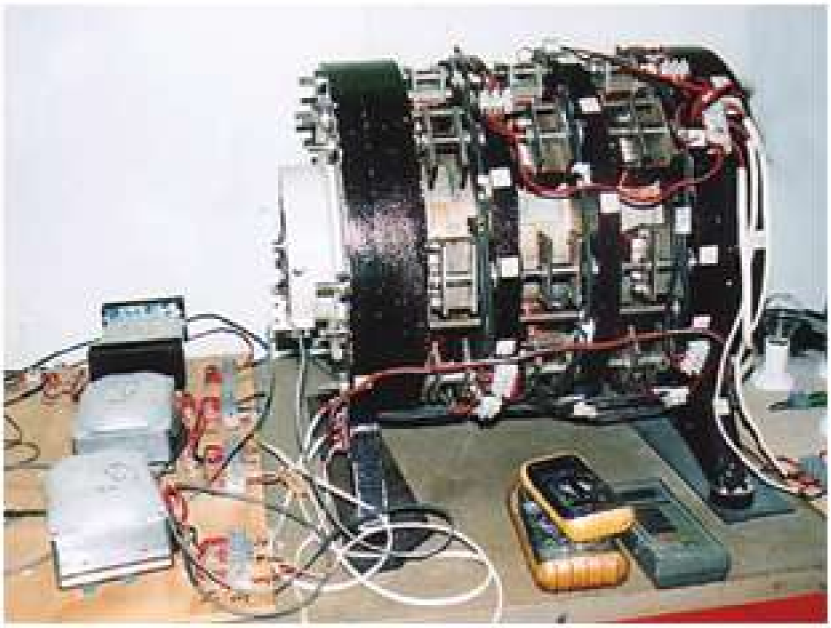

Electrogravity

by Tim Ventura
tventura@attbi.com

In this issue a report by Tim Ventura, USA on experimenting with asymmetrical capacitors.
www.americanantigravity.com

NEW ENERGY TECHNOLOGIES #6

1. Large-Scale Shakharov condition, **David Noever and Christopher Bremner**
2. Matter as a resonance longitudinal wave process, **Alexander V. Frolov**
3. Physical Principles of The Time Machine, **Alexander V. Frolov**
4. Time Machine Project by **Alexander V. Frolov**
5. Kozyrev-Dirak radiation, **Ivan M. Shakhparonov**
6. The Electrical Vortex Non-Solenoidal Fields**, S. Alemanov**
7. Physical Mechanism of Nuclear Reactions at Low Energies, **V.Oleinik, Yu. Arepjev**
8. The Evolution of Lifter Technology, **T. Ventura**
9. Reality and consciousness in education and activity, **A.Smirnov**
10. Old new energy, **Y. Andreev, A. Smirnov**
11. On the influence of time on matter, **A. Belyaeva**
12. Life without diseases and ageing-preventive electrical bio-heater features, **A. Belyaeva**
13. Technical report, on Belyaeva's high efficient ceramic heater, **Sh. Mavlyandekov**
14. Fundamental properties of aether, **A. Mishin**
15. Effect of Magnetic Blow Wave Field on Wine Systems, **I. Shakhparanov and others**
16. Nikola Tesla and Instantaneous Electric Communication, **V. Korobeynikov**
17. The Unitied Gravitation theory, **I. Kuldoshin**
18. New Sources of Energy from the Point of View of Unitary Quantum Theory, **L.G. Sapogin, Yu.A. Ryabov, V.V. Graboshnikov**
19. Antigravitation Force and antigravitation of matter. Methods of its creation, **A. K. Gaponov**
20. The capacitor, which has energy of atomic bomb (Review of A. Gaponov's research)

consequence of many commonly accepted concepts and dogmas of the modern "scientific perspective of natural phenomena". This crisis situation in modern physics is a direct consequence of many conservative scientific viewpoints, unfortunately supported and protected by modern official academic science. The evolution of our consciousness has been influenced by many undoubtedly well known experts and has been evolving for a long time in the environment of specific scientific vacuum and requires immediate revival. Even methods used for dissemination of new knowledge should be improved, if one actually wishes to accelerate the progress of Humankind.

The perspective for practical applications of new previously unknown scientific phenomena and effects looks very attractive, and they may be achieved by cooperative efforts of the human intellect. New breakthrough technologies of the 21st Century will require serious changes of many commonly accepted concepts and dogmas in fundamental physics. This process of progressive development cannot be stopped.

REFERENCES

1. Richard P. Feynman, Robert B. Leighton, Matthew Sands. The Feynman Lectures on Physics, Addison-Wesley, 1964, Vol. 2, Ch. 1. Paragraph 6 "Electromagnetism in Science and Technology" (the very end of paragraph)
2. J. Maxwell, Selected Works on the Electromagnetic Field Theory, Gostekhizdat, Moscow (1954).
3. G. V. Nikolaev, Non-contradictory Electrodynamics. Theories, Experiments, and Paradoxes, Publishing House of the Tomsk State University, Tomsk (1997).
4. A. S. Kompaniets, in: Theoretical Physics, State Technical and Theoretical Press, Moscow (1957), pp. 126-128.
5. R. T. Sigalov, T. I. Shapovalova, Kh. Kh. Karimov, and N. I. Samsonov, New Research of Forces of the Magnetic Field, FAN Press of the Uzbekskaia SSR, Tashkent (1975).
6. Ya. I. Frenkel, Electrodynamics. Vol. 1, United Scientific and Technical Presses, Leningrad/Moscow (1934).
7. G. V. Nikolaev and B. V. Okulov, Inertial Properties of Electrons, deposited at VINITI, No. 4399-77, Moscow (1978).
8. Observations of the Aharanov-Bohm Effect, Nature, No. 7, 106 (1983).
9. G. V. Nikolaev, Scientific Vacuum. Crisis in Basic Physics. Is There Any Way Out?! Publishing House Kursiv, Tomsk (1999).

Large-Scale Sakharov Condition

David Noever and Christopher Bremner

NASA Marshall Space Flight Center,
Space Sciences Laboratory
Mail Code: ES76, Huntsville AL 35812

Editor's note: This article was presented by the autors for publication in New Energy Technologies. For the first time it was published in 1999 by the American Institute of Aeronautics and Astronautics, Inc. All copyrights belong to the authors.

Abstract

Recent far reaching theoretical results have used the quantum vacuum noise as a fundamental electromagnetic radiation field to derive a frequency (ω) dependent version of Newton's gravitational coupling term, $G(\omega)$. This paper reconciles the cut-off frequency with the observed cosmological constant, and then briefly puts forward a realizable laboratory test case in the 10 - 100 MHz frequency range. One analogy is drawn between the classical vacuum energy experiments with attraction between two closely spaced plates (Casimir cavity) and the arbitrarily dense material boundaries possible in Bose condensates, such as irradiation at MHz frequencies of superfluid helium or superconductors.

Theoretical Background

Zel'dovich [1] first suggested that gravitational interactions could lead to a small disturbance in the (non zero) quantum fluctuations of the vacuum and thus give rise to a finite value of Einstein's cosmological constant in agreement with astrophysical data. Using dimensional analysis and the suggestion by Zel'dovich, Sakharov [2] derived a value for Newton's gravitational constant, G, in only one free parameter, frequency, ω:

$$G \sim c^5 \Big/ h \int \omega d\omega \sim 1 \Big/ \int \omega d\omega$$

where c is the speed of light and h is the Planck constant. The free parameter in frequency when integrated over all values from zero to high frequencies must contain the usual integration cutoff value (Planck frequency on observable electromagnetic phenomenon).

Puthoff [3] and others [4 5] have extended Sakharov's condition in a relativistically consistent model to determine constants of proportionality. His model derives an acceleration term in first order expansion (in flat space time), then equates inertial and gravitational mass (by the equivalence principle) to make contact with the gravitational constant, G, directly as:

$$G = \left(\pi c^5 / h\omega_c^2\right) \sim 1/\int \omega d\omega$$

which is the Sakharov condition [2,3]. This paper revisits the meaning of the cutoff frequency, ω_c, for radiation interactions, of which the quantum vacuum [6-10] and

Planck frequency are only the leading terms, and for which linear combinations of forces can introduce other plausible frequencies. One purpose of this reexamination is whether the resulting gravitational coupling constant, G, can be reconciled with the anticipated energy density of the universe [11] without resorting to extreme space time curvature and thus yield enough critical density to contain the expansion of the universe. Finally we particularize the case to the high-density fluctuations possible in Bose condensates [12], a potential experimental test case for how the effects of vacuum noise might manifest observably.

One far-reaching consequence of the vacuum energy model is the attractive force of gravity becomes reducible to the radiative interaction between oscillating charges, e.g. the zero point field (ZPF) applied to subatomic charges. Mass and inertia arise from the fundamentally electromagnetic ZPF oscillations.

This random background gives the usual quantum mechanical energy spectrum from particle field effects:

$$\rho(\omega)d\omega \sim \omega^3 d\omega$$

a very important dimensional relationship, since the third power in frequency avoids anomalous Doppler shifts from velocity boosts, or stated alternatively is the correct spectra for a Lorentzian (non accelerated) invariant radiation field [13].
More specifically, the energy spectrum [3] can be written as:

$$\rho(\omega)d\omega = [\omega^2 / \pi^2 c^3][h\omega / 2]d\omega =$$
$$= h\omega^3 / 2\pi^2 c^3 d\omega \sim \omega^3 d\omega$$

which is an expression in the first parenthesis of the density of the normal modes and in the second parenthesis of the average energy per mode. When this energy density is integrated over all frequencies, the ω^3 divergence produces well known infinities in the integration limit of high frequencies, thus an assumed cutoff frequency (appropriate to experimental observation limits at the Planck frequency), is usually introduced:

$$\omega_\rho = \left(c^5 / hG\right)^{1/2}$$

For mass, m, moving in an accelerated reference frame $g = -a = Gm/r^2$, the resulting energy spectrum includes a gravitational spectral shift [3],

$$\Delta\rho'(\omega)d\omega = h\omega / 2\pi^2 c^5 [Gm / r^2]^2 d\omega \sim 1 / r^4 d\omega$$

a kind of short range ($1/r^4$) gravitational energy shift, but electromagnetic in origin when zero point fluctuations are included. (N.B. To account for equal gravitational mass effects in neutrons and protons, the ZPF oscillations must involve subatomic charges, or 'parton' effects. The assumption derives from high frequency interactions of ZPF wherein these subatomic particles are asymptotically free to oscillate as independent or free particles as quantum noise).

A further far reaching consequence [3] is mass itself becomes interpretable as a dependent quantity derived from a damped (with decay constant Γ) oscillation driven by random ZPF:

$$m = \Gamma c^3 / G = 2h\Gamma / \pi^2 c^3 \int \omega d\omega$$

with the only two free parameters, the damping factor Γ, and again the frequency, ω. The internal kinetic energy of the system contributes to the effective mass.

This leads to an overall average spectral density, written in terms of mass as:

$$\Delta\rho'(\omega) = m^2 c^5 \omega / 2h\omega_c^4 r^4$$

for the electromagnetic field distribution near ($1/r^4$) to the mass, m, which in detail is half electric and half magnetic.

One additionally attractive feature is the correspondence between this derivation and the view of gravity as a dynamical scaleinvariance breaking model (e.g. symmetry breaking near the Planck mass energy [14]). A final result includes the force calculation between two ZPF radiation oscillators of the correct form yielding Newton's average force law

$$< F > = -Gm^2 / r^2$$

Thus, for a Newtonian force to first order in a flat space time, Sakharov [2] could be credited for proposing gravity as not a fundamentally separate force and Puthoff [3] and co workers [4-5] applied the vacuum electromagnetic field to equate gravity to a long-range radiation force (e.g. van der Waals like force). Higher order oscillatory gravity modes vary as $\left(\sin[\omega_o / \omega_c]\right)^2$.

To first order, a weak G coupling constant, $G = \left(\pi c^5 / h\omega_c^2\right)$, appears for high frequency cutoff at the Planck scale. A corollary in analogy to electromagnetic shielding by ordinary matter can be rationalized as the problem of frequency mismatch at high Planck frequencies, e.g. ZPF cannot be fundamentally shielded. In other words, frequency mismatch precludes gravity shielding by matter.

The purpose here is to revisit the only free parameter, the frequency cutoff, more in the spirit of a mass resonant frequency. The motivation for this approach can be summarized as: 1) the generality of other complementary radiation effects without relying on ZPF alone (e.g. other isotropic, homogeneous radiation sources); 2) the weak coupling constant, G, yields a vastly smaller than observed size of the universe (e.g. too small cosmological constant) when the Planck frequency is used as a cutoff value; and (3) the particle mass, $m = \Gamma c^3/G$, can be viewed as a renormalized or 'dressed' mass with a resonant interaction potential that is frequency dependent in its coupling constant, G, and with 'bare' mass that is large, $m_o \sim \left(m_\rho^2 / m\right)$, where the experimentally unobservable, $m_\rho = \left(hc/G\right)^{1/2}$ is the Planck mass.

In particular, why this large 'bare' mass does not generate a large gravitational field is not a unique anomaly in the Sakharov derivation, since similarly large vacuum point energies are common to field theories. The important point is that the derivation $G(\omega)$ is general however to any isotropic radiation field with the Lorentz invariant energy spectra $\left[\rho(\omega) \sim \omega^3\right]$, thus the candidates for the cutoff frequency of the particular radiation source can be interpreted as a Planck scale only if the rest mass, m_o, is not composed of many terms, rather than just the ZPF leading term. Since the ZPF is akin to a van der Waals force [3 5], polarizability (in charge and mass) must be considered, but without also excluding any number of linear combinations that might have alternative cutoff frequencies, ω_c, or damping terms, Γ, 'ala particle physics interpretations for resonant masses during renormalization. In other words, once a gravitational energy spectrum, $\rho(\omega)$ is postulated that is Lorentzian invariant, many fundamental sizes (or corresponding frequency values) are smeared (or dressed) by any number of characteristic frequencies between zero and the high frequency electromagnetic (Planck) cutoff ω_ρ. Quite simply, is the expression, $\omega_c = \omega_\rho$, a requirement for all radiation sources?

Many types of particle oscillations may satisfy the general requirements of a Sakharov condition, each having a characteristic mass (and energy) as in calculating the mass of any fundamental particle at its resonant frequency (including underlying partial charges or dense bosons). This brings the calculation to a consideration of the high density fluctuations characteristic of a Bose condensate [15 19]. While the high density variation may intrinsically be of interest, the exploration has more to do with reconciling the ZPF interpretation of the Sakharov condition with the observed cosmological constant [14].

A "top down" view of calculating the cutoff frequency imposes the self consistency test for the cosmological constant, Λ, from the outset. To calculate, the total frequency integrated energy density of the universe must be included:

$$\rho(E) = \int \rho(E) dE = h\omega_c^4 / 8\pi^2 c^3$$

which must have a mass equivalent, contribute to the universe's curvature, and thus have a fundamental relation to the critical density to contain the expansion of the universe [14 15]. The mass - equivalent ZPF to reach the universe's critical density [15], $\rho \sim 10^{-29}$ g cm^{-3} would necessarily limit the cutoff frequency for gravity to the value, $\omega_c < 7 \cdot 10^7 s^{-1}$, or between 10 -100 MHz.

A higher frequency greatly overshoots the cosmological constant, Λ, and induces extreme curvature in the universe. This problem has been cited frequently and stated most bluntly, as either ZPF or the cosmological constant requires revision. The relevance here arises from similarly large positive coupling terms in quantum gravity [15], which also generate a local gravitational Instability for typical upper limits on the cosmological constant, $\Lambda/8\pi G < 10^{12}$ cm^{-4}.

Rather than to dwell on the inconsistencies that plague attempts to reconcile quantum gravity, we particularize the problem to a case where the restriction to Planck scale becomes less clew, namely the high density fluctuations and universal scaling introduced in a Bose condensate. A Bose condensate, such as superfluid helium or superconductors [15 19], becomes of potential interest, mainly because of its arbitrarily dense boundaries and the classic Casimir experiment [20 22] which allows such dense material boundaries (two closely spaced conducting plates), if available, to modulate the background quantum fluctuation of ZPF. In other words, the matter-ZPF interaction becomes measurable by the observed attraction between two material boundaries. What dense boundaries might generate in Bose condensates remains a subject of great interest.

The significant case to investigate is whether Casimir-like interactions [20 22] will not only couple to ZPF radiation at a scale comparable to the quantum noise (or other radiation field), but also alter the value imposed by the Sakharov condition for G. It remains an open question whether this potential coupling interaction shares, as in ordinary critical phenomenon, the density correlation function, Φ, that is both independent of the coupling strength (or universal in renormalization) and consistent with the observed average energy density of the visible universe.

Thus the purpose here has bow to restate the Sakharov condition in the gravitational coupling constant, G, based on its only free parameter, a frequency cutoff, ω_c. Any potential relevance arises from similarly large values for the positive coupling term in quantum gravity, which generate conditions for a local gravitational instability for typical upper limits on the constant, $\Lambda/8\pi G < 10^{12}$ cm^{-4}.

To restate the Sakharov condition, matter in the vacuum provides boundaries for reduced 'Casimir like' modes available for otherwise isotropic radiation from quantum fluctuations (broad spectral noise). That this view reproduces Einstein gravity has been examined, including the full relativistic derivation [4-5]. The details of the appropriate mass, however, remain buried in the kinetic energy of general internal particle ('parton') motion [3]. Any appeal to a specific parton representation is limited only by essentially free particles with high frequency interactions, including underlying partial charges or dense bosons. The basis of considering arbitrarily high-density fluctuations in Bose condensate in analogy to the ZPF-Casimir experiment remains both an empirical and theoretical case to examine. There exist laboratory scale cases [15-19] where resonant radiation in the required 10-100 MHz range appear to produce anomalous effect for such Bose condensates as superconductors, but further work to confirm these results would be needed. In other contexts, these effects have been discussed as the Schiff-Barnhill effect for superconductors interacting with a gravitational field [23], but for the static rest moss rather than an effective mass in a conduction band.

Experimental Propositions

J. Weber [24,25] proposed the use of a superconducting Bose condensate for gravity wave detection, principally because of its potentially higher signal to noise ratio in carrying electrical signals upon length dilations in a relativistic framework for gravity waves travelling near the speed of light. W. Weber and Hickman [26] derived an experimentally testable relation based on torquing of a charged capacitor parallel to a gravity field, with

$$\tau = 2E_g / \pi[\alpha / (1-\alpha)^{1/2}]$$

where the capacitor will rotate relative to the gravity vector, for $\alpha = 2GM / rc^2$, is Schwarzschild radial coordinate $[dR = dr(1-\alpha)^{1/2}]$, E_g is dependent on the capacitor charge and geometry of the plates, $E_g = [Q^2 d/2\varepsilon WL(1-\alpha)^{1/2}]$, for a plate separation, and radial dimensions, W and L, charge Q, and ε the permittivity of free space. For plate separations of 2 mm on Earth, the maximum torque is approximately $\tau = 10^{-12}$ Nm, when charged to 2/3 dielectric breakdown. While not entirely promising for detection of such low torques, the large separation (2 mm) distance between capacitative plates naturally prompts generalization to the classic Casimir force [21] experiments only recently confirmed experimentally [20]. In particular, we rewrite the torque values to include the frequency terms derived with the Sakharov condition

$$[G = (\pi c^5 / h\omega_c^2)]:$$

$$\alpha = 2M\pi c^3 / h\omega_c^2 r$$

The appeal of this formulation is that a frequency dependent torque is derived, which further makes contact with proposals to modulate the Casimir capacitative plates for continuous extraction of energy [27]. This result requires further investigation experimentally, particularly to compare with previous reports for anomalies in AC- tuned electrical capacitors [28].

References

1. Ze1'dovich Ya. B. JETP Letters, 6, 345, 1967.
2. Sakharov A. Vacuum quantum fluctuations in Curved Space and the Theory of Gravitation, Sov. Phys. Reports, 12,1968,1040 1041.
3. Puthoff H. E. (1989) Gravity as a zero-point fluctuation force, Physical Review A, 39(5): 2333 2342, March 1, 1989.
4. Haisch B., Rueda A., Puthoff H.E., (1994) Inertia as a Zero Point Field Lorentz Force, Physical Review A, 49:678 694.
5. Haisch B., Rueda A., Puthoff H.E.,"Inertia as a Zero Point Field Force" Physical Review A 49, N 2, 678 (1994).
6. Ambjorn J. and Wolfram S. (1983) Properties of the Vacuum, 1. Mechanical and Thermodynamic, and Properties of the Vacuum, 2. Electrodynamics, Annals of Physics
7. Ambjorn J. and Wolfram S. (1983) Properties of the Vacuum. 1. Mechanical and Thermodynamic, Annals of Physics, 147:1 32.
8. Fulcher et al., "The Decay of the Vacuum," Sci. Am., vol. 241, p. 150, Dec. 1979
9. Puthoff, H.E. "Source of Vacuum Electromagnetic Zero Point Energy" Physical Review A 4 0, 4857 Nov 1 (1989); Errata and Comments, Physical Review A 4 1, March 1(1990); Physical Review A4 4, 3382, 3385 (1991)
10. Senitzky I.P., "Radiation Reaction and Vacuum Field Effects in Heisenberg Picture Quantum Electrodynamics", Phys. Rev. Lett. 31(15), 955 (1973). As pointed out by Puthoff [3] the relativistic results for the Sakharov condition have so far been encouraging, while the consequences for nuclear interactions in all coordinate frames have not been fully explored.
11. da Costa L. N., Freudling W., Wegner G., Giovanelli R., Haynes M. P., and Salzer J. J. (1996) The Mass Distribution in the Nearby Universe, Astrophysical Journal Letters, 468: LS L8 and Plate L 1
12. Modanese G. (1996) Theoretical analysis of a reported weak gravitational shielding effect, Europhy. Lett., 35(6):413 418.
13. Shupe M.A. The Lorentz invariant vacuum media, Am. J. Phys. 53, 122 (1985). A cautionary note is that lower frequency cutoffs can violate Lorentzian invariance, thus allowing a moving detector to reveal absolute motion by recording Doppler sifted frequencies. Standard methods might treat such effects like the cancellation of terms that remove anomalous ZPF infinities from field theories, but these topics remain to be explored.
14. Zee, A. Phys. Rev, Lett, 42,417 (1979); Phys. Rev. D. 23, 858, (1981).
15. Torr D. G. and Li. N. (1993) Gravitoelctric Electric Coupling vVia Superconductivity, Foundations of Physics Letters, 6(4): 371 383.
16. Unnikrishan C. S. (1996) Does a superconductor shield gravity? Physics C, 266:133 137.
17. Podkletnov E. and Nieminen R (1992) A Possibility of Gravitational Force Shielding by Bulk $YBa_2 Cu_3 O_{7-x}$ Superconductor, Physics pp.203: 441 444.
18. Li N, and Torr D. G. (1992) Gravitational effects on the magnetic attenuation of superconductors, Physical Review B, 46(9): 5489 5494. A simple consequence of the Sakharov condition $G = (\pi c^5 / h\omega_c^2) \sim 1/\int \omega d\omega$, can be written for the gravitomagnefic permeability as: $\mu_g = 4\pi G / c^2 = 4\pi^2 c^3 / \left(h \int_\circ^{\omega c} \omega d\omega\right) \sim 1/\int \omega d\omega$ which suggest that the same frequency resonance implied by the ZPF derivation will share similar consequences for vector gravity

effects. See also, DeWitt, B. S. Superconductors and Gravitational Drag, Phys. Rev. Lett. 16, 102(1966).

19. Li N., Noever D., Robertson T., Koczor R., and Brantley, W. (1997) Static Test for a Gravitational Force Coupled to Type II YBCO Superconductors, Physics p., 55, 287.

20. Lamoreaux S. K. (1997) Demonstration of the Casimir Force in the 0.6 to 6 mm Range, Phys. Rev. Letters, 78:5 8.

21. Milonni P.W. et al., “Radiation pressure from the vacuum: Physical interpretation of the Casimir force”, Phys. Rev. A, Vol. 38, No. 3, 1621 August 1988.

22. Milonni P. W. (1994) The Quantum Vacuum, Academic Press, San Diego, CA.

23. Schiff L.I. and Barnhill M.V. Bull. Am. Phys. Soc. 11, 96, (1966) and refn. 18.

24. Weber J. (1960), Detection and Generation of Gravitational Waves, Physical Review, 117(1)306 313.

25. Weber J. (1966) Gravitational Shielding and Absorption, The Physical Review (The American Physical Society), 146(4): 935 937.

26. Weber W. and Hickman H. (1997) A possible interaction between gravity and the electric field, Spec. Science Tech. 20, 133 136

27. Forward R.L. “Extracting electrical energy from the vacuum by cohesion of charged foliated conductors” Phys. Rev. B, Vol. 30, No. 4,1700 August 1994

28. Woodward, J. F. (1992) A Stationary Apparent Weight Shift From a Transient Machian Mass Fluctuation, Foundations of Physics Letters, 5:425 442.

The Problem of Electron and Physical Properties of Time:

To the Electron Technologies of the 21st Century

V.P. Oleinik

*Department of General and Theoretical Physics,
National Technical University of Ukraine
“Kiev Polytechnic Institute”,
Prospect Pobedy 37, Kiev, 03056, Ukraine **Institute of
Semiconductor Physics, National Academy of Sciences,
Prospect Nauky 45, Kiev, 03028, Ukraine;
e-mail: yuri@arepjev.relc.com

> ***“...it is necessary to periodically subject to the deepest revision the principles, which were recognized as final and were no longer discussed”.***
> ***Louis de Broglie***

Abstract

The results of an approach based on the synthesis of standard quantum electrodynamics and of the ideas of self-organization in physical systems are briefly outlined. The quantum model of electron as an open self-organizing system is constructed, with the physical mechanism of self-organization consisting in the back influence of the own field created by electron on the same electron. The own field is considered as a physical property of electron, intrinsically inherent in electrically charged matter, which is included in the definition of the particle from the very beginning. The own field of electron endows the particle with wave properties and represents a bearer of superluminal signals, which can be used for the creation of qualitatively new communication systems. Because of the inseparable link between space and time, the force in relativistic mechanics is the cause of change not only of the velocity of particle, but also of the course of time along the particle’s trajectory. For this reason the flow of time in some area of space depends on the character of physical processes, occurring in it, and, therefore, time can be controlled by slowing down or accelerating its course with the help of material processes. The conclusions of the paper are not in conflict with the special theory of relativity (STR); they are a direct consequence of relativistic equations of motion and represent an essential development of the generally accepted notions about space and time. At present all the necessary prerequisites are available, both theoretical and technical, for the practical mastering of the own fields of particles and of the physical properties of time.

1. Introduction. The Problem of Electron and Future Outlook

Electrodynamics, what is this? What is its value for man? Electrodynamics is the theory of electromagnetic interaction, one of four interactions existing in nature. Its role in the life of society is seen from the fact that the most part of natural phenomena, which we encounter at every step, is of electromagnetic origin: it is due to the interaction of electromagnetic field with electrically charged particles entering into atoms and molecules. It is fair to say that electromagnetism plays a crucial role in the life of mankind as it determines the ways of technical advance of society [1].

The key problem of quantum electrodynamics is the problem of electron, which can be formulated as follows: to construct from the first principles a non-contradictory model of electron, which takes into account experimental facts, i.e. to find the dynamical equation capable of describing the unique physical properties of electron, its internal structure, its behaviour when it interacts with electromagnetic field.

Electron was discovered a little more than 100 years ago, in 1897. With discovering the electron the revolution in physics began, which has resulted in unprecedented technical advance of society. The summit of development was reached in the middle of the 1950s and then the long period of evolutional development followed, when new physical principles were used to describe various physical processes and phenomena. The violent development of physics became slower in the 1970s and was replaced by stagnation in the subsequent years. The stagnation in electrodynamics continuing already over a period of several decades is gradually giving place now to a new ascent. The new scientific revolution is starting, which is associated with electron again, much as it happened hundred years ago. The reason is that electron is the most unique particle storing in itself the deepest mysteries of nature and the degree, to which they are disclosed, determines the wants to stay anonymous, until his patent application is done and university verification tests will be done). The claims are: 1200 Watts coil out with about 1076.4 Watts in into the driving motor at 3450 RPM. 8 amps 117volts at no load 9.2 amps 117 volts at full load. The output of about 1200 Watts is already a total overunity operation! As they just increase the input power by about 140 Watts only between idle and load state and they get 1200-Watts output it seems indeed a case, where Lenz law is violated! This generator also has NO motor effect! If you supply current to the coil, the permanent magnet in the center will not rotate; cause the flux just stays inside the toroid core! There you can see, that the back drag does not influence the mechanical rotation of the magnet!" Stefan used very good criterion to prove high efficiency of the design: **There is no back-torque effect! It is most important aspect of Gramm's generator.** You can contact directly Stefan Hartmann: Keplerstr. 11 B, 10589 Berlin, Germany. Tel: +49 30 345 00 497, FAX: +49 30 345 00 498 email: harti@harti.com info@ccard.net (Please, note: Dr. Harman referred to my old web site www.time-machine.spb.ru which is closed now).

So, basic principles of MEG and Φ-machines are the same. It was patented more than 100 years ago. Primary magnetic flux is topologically separated in two (or more) fluxes, which are mutually compensated in the ring core. Advantages of MEG are absence of moving parts since special input coils produce changes of primary flux. Also level of saturation in ferromagnetic material obviously should be corresponding to intensity of primary magnetic field, which is created by the permanent magnet, Fig.4.1.

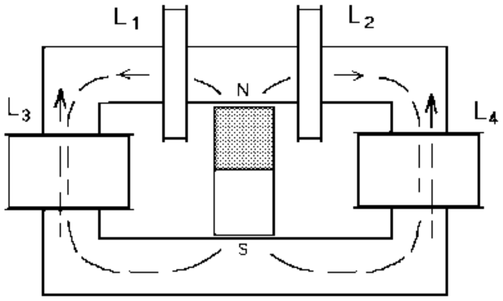

Fig.4.1

Besides MEG the same principle can be (and already was!) realized in many other systems. So, there is no any news in the USA patent #6,362,718 granted for "The Motionless Magnetic Generator". What did they claim? You can find it in the patent: "The first input coil and the first output coil **extend around portions of the first magnetic path**, while the second input coil and the second output coil **extend around portions of the second magnetic path."** Yes, it is the same bi-directional principle we discussed above: two parts of the magnetic flux and each coil produce effect to reduce flux due to this superposition.

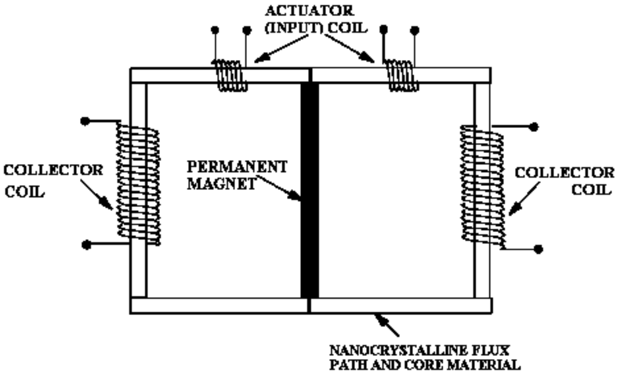

Fig.4.2
Diagram of prototype by Bearden.

In conclusion I'd like to confirm our sincere interest to develop joint work with all new energy research teams if they are not trying to obscure the issue of the technology by means of complex theoretical constructions and common words about zero point energy.

Matter as a Resonance Longitudinal Wave Process

Alexander V. Frolov

Abstracts

There is experimental data on gravitation anomalies for cases of resonance irradiation of the Bose condensates (superfluid helium or superconductor) at 10-100 MHz frequencies. It is developed by the author in frames of his aether theory that can be used for practical applications in aerospace and new energetics.

ZPF or aether fluctuations

The fundamental electromagnetic radiation field (Zero Point Field) ZPF or the quantum vacuum noise is a recently accepted physical notion. In the article [1] David Noever and Christopher Bremner used it to derive a frequency – dependent version of Newton's gravitational coupling term G. On the other hand we can consider the quantum vacuum noise as aether fluctuations. Dr. Alexander Mishin [2] described experiments on registration of these processes by means of special equipment. Both approaches (ZPF and aether fluctuation) allow to conclude that mass and inertia arise from these oscillations. However if we are considering the oscillation as some aether process then we can assume and describe some physical mechanism of this process.

One of consequence of the vacuum energy model, which is described in [1] is that "the attractive force of gravity becomes reducible to the radiative interaction between oscillating charges..." Let's clarify which kind of radiation can be created by oscillating electric charges.

There are many different sources to find the answer on this question and one of them is the article by Prof. Kirill P.Butusov [3] on symmetrization of Maxwell's equations and practical methods of generation of longitudinal waves in vacuum. So, ZPF model has a direct relation with the aether model since indirectly it leads to the question of longitudinal waves in vacuum. Physically they are waves of density of energy and in the aether model the waves are areas of more dense and more rarefied aether. Let's note that there are standing waves besides moving waves.

To consider the interaction of some mass particles and the fundamental field the notion of subatomic charges "partons" was introduced [1]. So, the mass itself "becomes interpretable as a dependent quantity derived from a damped oscillation driven by random ZPF" [1]. The authors wrote about "internal kinetic energy" of the mass particle and it can be considered as a function of ZPF oscillation frequency. In the aether theory of mass there is a similar notion of "aether vortex", which represents some cyclical process of some frequency and it is possible to calculate its kinetic energy. This aether vortex model of matter elements allows to assume real methods to change parameters of vortex and to get changes in parameters of existence of the matter. On the other hand we can discuss the possibility to change some physical parameters of aether in areas of the vortex to get the same result. This possibility follows from the well-known N.Kozyrev's experiments, which were named "investigation of active properties of time". N.Kozyrev used chronal (temporal) approach in his theory. We have to change his notion "the density of time" to "the density of aether" to get a direct link between his experiments and the aether theory of mass.

N.Kozyrev and others have [4,5,6] experimentally demonstrated that irreversible processes in matter produced changes of aether density in the area of the experiment. Detectors of different type can register this change. It is obviously that any matter element (i.e. the aether vortex) in this area of changed aether density should get more inner (kinetic) energy or slow the inner motion. From the chronal point of view these are changes of inner time of this matter element.

Gravity shield

One more interesting point that is discussed in the article by Noever and Bremner [1] is a problem of gravity shield. The authors show that resonance interaction with ZPF produces "the particle mass" and it can be viewed as "a renormalized or "dressed" mass with a resonant interaction potential. Similar resonance approach is used in the conception of de Broglie's matter waves. Also the authors [1] mentioned the existence of an experimentally unobservable mass. In this case ZPF cannot be fundamentally shielded by matter since "frequency mismatch precludes gravity shielding by matter" [1]. The only way to get screening of ZPF fluctuations seems to be very complex: it is necessary to provide frequency matching for whole wavelength band of the oscillations.

I think some specialization is necessary here to explain experimental gravity anomalies with Bose condensates experiments (superfluid helium or superconductors): **special process in matter can be used as the gravity screen and this approach does not involve the frequency-matching problem.**

We have concluded above that any matter element is a resonance process and its energy is derived from ZPF. It is useful to note that these are longitudinal wave oscillations of energy density in aether. In this case, the gravity shield problem can be solved in frames of the aether vortex conception of matter.

The longitudinal wave is a moving (or standing) areas of rarified and thickened aether. Let's consider the moving wave, which is responsible for gravitation attraction effect. How can we stop, re-direct or reflect longitudinal wave in aether by means of aether vortexes (matter elements)? **We can produce interaction with this wave only by means of other longitudinal waves.**

In macro-level this idea can be realized as longitudinal wave generator. Electromagnetic processes, which can be used as sources of directed longitudinal waves, are known and some of them are described in [3]. In other way the gravity shield can be produced as longitudinal waves generated by natural aether vortexes (i.e. by matter elements) if the matter exist in a special exited state, for example for cases of resonance irradiation of superfluid helium or superconductor at 10-100 MHz frequencies.

Matter element as resonance process

In [1] the authors wrote that it is possible to calculate "the mass of any fundamental particle at its resonant frequency." There is the question: what is the general basis of whole spectrum of stable elements masses?

In 1996 the author published the article "The concept of mass process" [7]. At first in this work physical sense and notion of 3-dimensional curvature was introduced. By analogy with known mathematical notion of linear curvature $\rho_1 = \frac{1}{r}$ (where r is radius) and uniform surface curvature $\rho_2 = \frac{2}{r}$ it was proposed to calculate curvature of a 3-dimensional space as

$$\rho_3 = \frac{3}{r} \qquad (1)$$

The radius r in this case means that in a 3-space there is some periodical process. In other words, 3-dimensional matter is a resonance process.

Further, de Broglie used formulations E=hf and E=pc (where p is momentum, h is Planck constant, f is frequency and c is velocity of light) to derive the following:

$$hf=pc \qquad (2)$$

that allows us to get the well-known formulation

$$\lambda = \frac{h}{p} \tag{3}$$

There is another logical branch of this idea that leads to the understanding of the mass properties of matter as a resonance process. Instead of E=pc in [7] it was proposed to use E=mc^2. In strength of the wave-particle duality we can write the equation

$$mc^2 = hf \tag{4}$$

and from this equation the mass can be presented as resonance electromagnetic oscillations

$$m = \frac{h}{c^2} f \tag{5}$$

Let's note that f=1/T, where T is some period of oscillation. So, we can write the following

$$m = \frac{h}{c^2} \frac{1}{T} \tag{6}$$

where $\frac{h}{c^2}$ is new constant between mass and period of time.

There is an important conclusion: any mass is a process and there is some period of time, which corresponds to this mass. In other words, there is no physical sense of time separate from some process of existence of mass. Product mass and period is a constant value, which was named as a chronal constant

$$mT = \frac{h}{c^2} = const \tag{7}$$

The chronal constant is a parameter of some real space and it is equal to $0.73725 \cdot 10^{-50} [Js^2 / m^2]$

Also in this work [7] there was a demonstration of several examples of newly discovered physical law: spatial curvature of some natural objects (proton, planet, DNA molecule) is a whole number. There is some analogy with the nuclear physics notion of wave number. From this fact we can assume that main natural matter elements exist in main resonance states. For example, if Bohr radius is 0.52917 Angstrom, then we can find the wave-length $l = \pi d$ and the linear curvature is $\rho = 1/l = 3.0075 \cdot 10^9\ (m)$ and 3-dimensional curvature of this object is $\rho = 3/l = 1.0025 \cdot 10^9\ (m)$ that is unit of mater, corresponding to simplest atom, i.e. unit matter engine. Let's note that it is near the unit and some distortion of 0.0025 means non-ideal resonance state of the system.

Calculations for planet Earth in [7] were based on the known period of orbital rotation T=31557600 sec that corresponds to frequency of electromagnetic oscillations

$$f = 1/T = 3.168861 \cdot 10^{-8} (1/s) \tag{8}$$

and wave-length

$$\lambda = c/f = 9.46... \cdot 10^{16} (m) \tag{9}$$

The curvature (if this wave-length is considered as radius of the resonator) is following wave number

$$\rho = 1057.00 \cdot 10^{-20} (1/m) \tag{10}$$

Also we can use other known data about the planet. Daily rotation period of our planet is known T=86400 sec and we can calculate its wavelength $\lambda = 3469{,}82(m)$ and corresponding curvature (wave-number). Sure, it is also a whole number with a good accuracy:

$$\rho = 2882 \cdot 10^{-7} (1/m) \tag{11}$$

The laws of physics in macro cosmos and micro cosmos are similar. From these calculations it was assumed that whole formation of mass spectrum of stable chemical elements of matter is determined by similar physical mechanism.

Creation of mass

In short we can summarize that technology of longitudinal waves in aether is a real basis for creation of matter with mass and inertia properties. N. Tesla used this method to produce different objects: from ball lightning up to electrons. Velimir Abramovic says in his article [8]: "The principle of resonance and harmonic oscillation of aether seems to be so clear that all problems of modern physics, especially a problem of energy conversion, will be solved with its development. By means of his vacuum tube Tesla got protons, electrons and neutrons directly from aether and reproduced them at any distance. Instead of giving a possibility to the bundle of protons to move through space to some place, he created conditions for momentary appearance of arbitrary quantity of particles in the given place."

Any objects can be classified as aether vortex and parameters of this vortex determine its mass, electric charge and other properties of matter.

The "parton" as element of matter in [1] is a useful tool for description of physical properties of aether.

Longitudinal waves in Woodward's experiment

In [1] the authors state that resonant radiation in the required 10-100 MHz range appears to produce anomalous effects for such Bose condensates as superconductors. In my opinion it is a particular case of discussed above technology of longitudinal waves in aether due to possibility of transformation of transverse electromagnetic waves in longitudinal waves in the superconductors. This transformation in plasma is a well known physical mechanism.

More facts to prove this idea: by Woodward [9] there is a special requirement, i.e. the frequency of mechanical vibrations should be twice the frequency of electrical oscillations in the capacitor, which demonstrates the weight anomalies. But from the other hand it is a common rule for creation of longitudinal weaves in plasma! Also it is a necessary condition for generation of parametrical oscillations! So, we can assume that basis of the effects in [1] and [9] is a generation of longitudinal wave in aether.

Conclusion

Any element of matter can be considered as resonance process of aether oscillations, which are longitudinal waves. There is an analogy with description of these longitudinal waves and well-known matter waves by de Broighl. Experimenting on the longitudinal waves generation and especially experiments on standing waves to get gradient of aether pressure allows to develop gravity control technology.

References

1. D. Noever and C. Bremner, NASA Marshall Space Flight Center, Large Scale Sakharov Condition, 35th AIAA/ASME/SAE/ASEE joint Propulsion Conference and Exhibit, 1999, Los Angeles, Ca, USA.

2. Alexander M. Mishin, The physical system of artificial biofield (experimental research of ether), New Energy Technologies, #1, 2001, p.45.

3. Kirill P.Butusov, Symmetrization of Maxwell-Lorenz equations, New energy Technologies, #2(5), 2002, p.14.

4. Kozyrev N.A. Selected Works, 1991, Leningrad, publ. by Leningrad State University.

5. L. Shikhobalov, N. Kozyrev's ideas today, New Energy Technologies #2 (5), 2002, p. 20.

6. Alexander V. Frolov, Kozyrev on possibility of decrease of mass and weight of the body, New Energy Technologies #2 (5), 2002, p. 35.

7. The Concept of Mass Process, Frolov A.V., Proceedings of the congress "New Ideas in Natural Sciences", 1996, St.Petersburg, published by PiK Co., p.123-134.

8. Tesla, Velimir Abramovic, New Energy technologies, #1(4) 2002, p.17.

9. Woodward, J.F. A Stationary Apparent Weight Shift From a Transient Machian Mass Fluctuation, Foundations of Physics Letters, 5, p.425-442, 1992.

Gerlovin's Theory of Activation

Alexander V. Frolov

It is a review of the famous book by Ilia L. Gerlovin "Basis of unified theory of all interactions in matter"published in 1990, St.Petersburg, Russia. We hope this article let you discover some new aspects of physical vacuum structure to develop more new experimental methods. Comments made by Alexander V. Frolov, Editor.

In [1] the author wrote about different methods to activate water solutions: mechanical, thermal, acoustic, magnetic and electrical. One of the known methods is an activation by means of electrohydraulic method. There is also some information about activation of other mediums, mainly liquids, but also some gases and solid bodies.

There are no theoretical explanations of these facts to explain all aspects of these phenomena. Furthermore, complexity of interpretation of these phenomena in frames of common physical notions induced some scientists to announce these phenomena as non-existing and "illegal".

Ilia L. Gerlovin formulated the physical principles of theory of activation of mediums on the basis of new physical theory, the Theory of Fundamental Field (TFF). These principles are based on his two important conclusions from the TFF:

a) "Space around us is not empty, physical vacuum consists of material physical objects, i.e. elementary particles of vacuum (EPV). These particles are responsible for main activation processes;
b) Force interactions between atoms in molecule, between molecules in crystals has not spherical symmetry in the crystals of solid bodies, but an axial symmetry and the interactions are changing in time with very high frequency of about 10^{18} Hz. This feature of force interactions also makes its own contribution to the activation of mediums." [1, p. 314]

So, it was assumed that the phenomenon of activation of mediums can be defined as anisotropy of force interactions, which leads to **"meta-stable state, which can be called *structurally activated state of the given structure*".**

Here is some difference in principle between chemical term "activation", which characterizes a transformation of molecule or atom in some active state with an increased energy, which is sufficient to provide a chemical reaction. It is ***energy activation*.** **Gerlovin described new notion, a *structural activation:* "This phenomenon can be classified as some change of structure of activation object. With this, energy of molecule can have no changes, and *active properties***

Coupled with aetherodynamics time conception, which was suggested by Alexander V. Frolov, the works on the control of space-time parameters gain the possibility for development and commercial application. As a theoretical basis there are those N. Kozyrev works where his conception of "time density" are replaced by that of "aether density" according to Frolov.

In September 2002, Faraday Labs Ltd Company plans to complete testing of the first experimental system, and to start the patenting and research of applied aspects, first of all in medicine.

Physical Principles of the Time Machine

Alexander V. Frolov

Experimental success of research team headed by Dr. Vadim A. Chernobrov, Moscow was reported in [1]. The time course can be controlled as rate of any process in local space-time (inner space of the Time Machine). It can be decelerated or accelerated by means of special "converging electromagnetic waves". Ordinary waves move from the source whereas special "converging waves" move to some central point, i.e. into the focus of the system. In Chernobrov's design of the Time Machine this process is organized by means of several spherical envelops, which consist of several electromagnets. Electronic control unit controls the processes in this design. Dr. Chernobrov reported about 3% change of the time course in 4th version of the system, which was tested with a human inside. The goal of Dr. Chernobrov's work is to research the medical aspects and experimental investigation of the principles. Several important conclusions were obtained from the project: the time course can be controlled and character of the changes is different for acceleration and deceleration mode.

Other known publication and research projects on the same topic seem to be very far from any commercial and practically useful application. Obviously the topic is very new and fantastic for most of scientific community and at first we have to clarify the physical principles of the time control project, which is started by Faraday Labs Ltd.

In this project we believe that notion of time is one of possible description of real physical properties of our Universe. So, it is not mathematical abstraction but some aspect of physical reality and we can discover some physical properties of time. Russian astrophysicist N. A. Kozyrev [2] developed a theory of active properties of time and according to his point of view there are two properties: time course and time density. Prof. Kozyrev demonstrated experimentally that time density in area of some process (changes of matter) is dependent on entropy parameters of the processes. In [3] it was demonstrated that Kozyrev's experiments could be interpreted in aether theory and it has led to simple physical conclusions and clear experimental perspectives: **time course and its density can be explained and controlled as parameters of aether.** Directions of aether flow and density of aether are subjects of experimenting with non-reversible changes in matter, for example, in crystallization or melting processes. Also it is possible to use special electromagnetic processes, for example, Chernobrov's "converging waves" or other longitudinal waves as methods of aether compression or rarefaction. If we assume that process of existence of elements of matter physically can be explained as aether vortex processes then its rate is a parameter of **aether income/outcome balance** (aether inflow in element of matter and aether outflow from the element of matter). It was also described in Time Rate Control (TRC) theory [3]. To control this balance it is necessary to develop technology of longitudinal waves generation, its focusing and resonance effects. The previous research and experimenting on the topic has been made by N. Tesla.

Let's assume that we have some technology to change parameters of time course. How should we organize this local space-time (what is spatial topology of the design)? There is a very interesting experiment to get the answer: rotation of a heavy cone (for example, lead cone) entrains surrounding aether, so a vortex appears, which is a toroidal formation of aether (rings). The rings can exist in space for a long period. The further question is: Why does the beam of light (laser beam) directed to the cone by tangent create a luminous ring? We can assume that due to natural properties of photons (light propagates along the geodesic line in space) some autonomous closed toroidal space should be created in such experiment. The next thought is: since space and any matter exist in time then we can speak about some autonomous time. The general conclusion is to be the following: autonomous 4-dimensional space-time can be created as toroidal aether vortex.

Here is point to note some aspects of research project by Prof. Robert Mallett, Connecticut University, USA. In fact, sometime next year, he hopes to produce the first piece of technology that eventually will allow him to build a time machine. By Mallett it will be a device that employs lasers "to twist space". Why is he going to close the beam of light? His theoretical background is knowledge about black holes, i.e. understanding of the connection between gravity and curvature of space-time. In Einstein's theory both matter and energy can bend space and time. So Prof. Mallett assumes that curvature of space-time can be changed not only by mass (like a black hole) but it can be affected by energy of photons. This has led Prof. Mallett to consider the possibility of using a circulating beam of light to twist space and to create closed loops in time. It is predicted that a spinning neutral particle, when placed in the ring, is dragged around by the resulting gravitational field [4]. From the first view it is the same approach we have considered above (experiment with aether toroidal rings). But proposals by Prof. Mallett differ in principle from the aether conception.

The main aspect of this technology is a creation of autonomous (self-closed) toroidal space-time. Autonomous geodesic world line of this space-time is self-closed. Any photon should be circulating in this system due to its properties: photon is always moving along the straight line of the space.

More deep understanding of this technology follows from the explanation of photon as oscillation of aether. Any photon can be considered as result of relative motion of the matter (observer) in absolute space (immovable aether). Usually a photon is considered as moving object in space. But we can assume that observer is in the motion and the photon is oscillations of the absolute space (immovable aether). Which approach is more real one? Sure, it is more easy to consider a photon as moving object but let's remember fact of our real motion in the Universe and fact of the Universe expansion.

So, ideas by Prof. Mallett are very far from the **aether nature of the time phenomenon.** He follows the black holes theory and general understanding of space-time distortion due to mass or energy presence. Also he knows that a light beam should be closed in a ring. However Prof. Mallett is very far from physical basis of the effects. **The key of time rate control is technology of artificial aether flow, creation of aether vortex systems (AVS), management on density and direction of aether flow.** There are several technical methods to produce it. Any light beam should be curved in self-closed "light ring" if it is placed in a toroidal aether vortex and we can say that this system has own space-time.

What does "some changes of time course" mean? We can measure it as some changes of standard rate of oscillation process, for example, some stable wavelength of laser beam or quartz oscillations. There is a well-known experiment with two atomic clocks (one of the clocks is placed on the roof of some building and another one is placed on surface of planet). Due to vertical component of gravity the time course should be different and it can be measured. How can we organize difference in these measurements if both atomic clocks are placed in the same altitude?

It is necessary to consider gravity nature in frames of the aether conception. Two atomic clocks demonstrate difference in measurements due to difference in aether flow density. Hence, by means of aetherodynamics methods it is possible to control the rate of oscillation processes in the atomic clocks and in any matter (i.e. time course itself).

The aetherodynamics methods have a clear analogy with electrodynamics: **motion of charge produce field and there is the induction law**. Really, classical electrodynamics can be considered as particular case of the aetherodynamics. So, physical sense of any field is stress or deformation (it is some static field) or oscillations of aether.

Let's introduce the notion of chronal (temporal) charge to consider some technical aspects. In electrodynamics we assume an electric charge as element of matter with positive or negative electric properties and we have to compare it with some reference (zero charge or test charge). Let's note that in any case we have to consider "charge of some particle" but not an "abstract charge". So, we can postulate that any element of matter has **zero chronal charge** if it is moving from Past in Future with standard (usual for measurements of surface of our planet) time course. If the time course (i.e. existence of some element of matter) is decelerated then it can be measured as decrease of standard oscillation frequency of the matter. Time course acceleration means some increase of standard oscillation frequency of the matter. Let's determine that in the first case it is negative chronal charge and in the second case it is some positive chronal charge. Atomic clock is one of possible methods to measure zero chronal charge or to find some relative positive or negative difference.

It is predicted here that motion of chronal charge should produce a chronal field. Some provisional data was received by Frolov from simple experiments on the rotation of a heat source. Accelerated motion of chronal charge (changes of density of chronal current) should produce aethero-induction effect that is an analogy (or more general case) of Faraday's induction effect. This effect can be detected as secondary (induced) deceleration of time course in nearest area of accelerated time matter. Another case is a secondary (induced) acceleration of time course in the nearest area of decelerated time matter.

Technical realization of aethero-induction method seems to be very close to idea, which is described in classical epic "Back to the Future". At first, it is necessary to create or to collect some chronal charge in a "flux condenser" and then to accelerate it in space up to some velocity. According to the aether conception, this creation of the chronal charge is a real technical process.

It is assumed that estimated chronal effects are demonstrated as some threshold field, i.e. space-time has some stable discrete energy levels and changes of its curvature should have discrete threshold mode. All new aspects disclosed in this paper are the subject of a patent process. Faraday Labs Ltd organizes experimental program on the topic. Practical application of this technology is new energy systems and propulsion methods.

References

1. Chernobrov V.A. Experiments with a man in the Time Machine, New Energy Technologies, #3 November-December 2001, p.6.
2. Kozyrev N.A. Selected works, Publ. by Leningrad State University, 1991.
3. Frolov A.V. Practical application of the Time Rate Control theory, New Energy Technologies, #3 November-December 2001, p. 15.
4. Mallett, R.L., Weak gravitational field of the electromagnetic radiation in a ring laser, Phys. Lett. A, 2000.

investigated completely yet. It was found also that harmful effect on biological systems is not related to the process of movement in Time itself but is a result of the difference of the Time rate value in various parts of a body (a biological system).

Inside of the laboratory setup it was also discovered that Time could be changed with some inertia. Areas of space having different Time rates have vague borders. With sufficient difference in Time rate the human can see an area with a different Time rate as some white mist. Higher the difference – the mist is denser, that can be used as an alarm signal for biological systems. It is possible to consider Time-travel as possible and (after experiments with mice) there are reasons to suppose it will be safe for travelers if they follow certain rules. It is especially necessary to emphasize: the trips through Time (due to new discovered properties of Time) can't affect the Past and they can't change our past history. All the so-called paradoxes for the traveler in Time (for example when "he meets himself in the Past" or "he kills his grandfather in his childhood" have clear solutions in 3-dimensional Time.

It is possible to consider as a proven fact that Time has more than one dimension, i.e. O. Bartini's theoretical calculations are confirmed by these experiments: Time has 3 dimensions. Hence our Earth world can be considered as a 6-dimensional object: length, width, height, age or date of Time, variant of a History or erosion of Time, density or rate of Time. The concept of "the Arrow of Time" as fourth dimension (moment of Time) is a particular case of the concept of sixth dimension (rate of Time) that leads to the physical concepts of gravitation and energy and they are simultaneously connected. Concepts of the " Einstein-Rosen bridges" known since 1916 or "worm-holes" introduced into science by John Willer in the 50's, are travels in 5th and 6th dimensions, i.e. the "classical" Time travels, which were described by H. Wells.

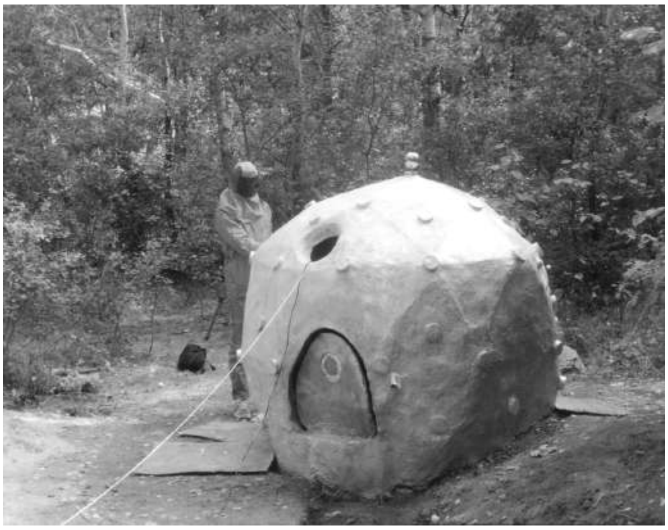

Editor's: As the reader could note, the author does not disclosure the secrets of the TM design. From the photo you can see the electromagnets, which form the regular stereometrical construction as well as the cables from the TM to the control unit. Dr. Chernobrov mentioned the converging electromagnetic waves only. So, to understand how it works, it is necessary to get a clear notion of the converging electromagnetic waves. Let's imagine the ripple effect created by a stone in the water. The waves move from a central point to periphery. The converging waves are just an opposite process: the waves move from periphery to the central point. Is it possible in Nature? Yes, sure. Dr. Chernobrov wrote: "Let's throw a hoop on the water and inside of the hoop we'll see converging waves." The Time Machine technology by Dr. Chernobrov is based on the similar principle.

Time Machine Project

Alexander V. Frolov

Scientific Expert of the Russian Physical Society,
General Director, Faraday Lab Ltd
Tel/fax: 7-812-380-6564 Tel: 7-921-993-2501
Email: director@faraday.ru alex@frolov.spb.ru

May 29, 2002

Faraday Labs Ltd and Dr. Vadim Chernobrov have signed the agreement on scientific-research work on investigation of active properties of time.

In the course of the previous experimental works, carried out by Dr. Chernobrov's research team during the period from 1984-2002, four versions of Time Machine had been made and tested. At these devices (the biggest system is about 1 meter in diameter) the effects of deceleration and acceleration of time course were created and measured. The principles of control of time course velocity were based on the interconnection of electromagnetic processes and physical properties of space-time. Special electromagnets, operating in pulse mode, are placed at the spherical frame. They create the so-called "converging wave", which by Alexander Frolov is a longitudinal wave in nature.

Alexander V. Frolov, General Director Faraday Labs Ltd and Ph. Dr. Vadim A. Chernobrov have just signed the Contract

Coupled with aetherodynamics time conception, which was suggested by Alexander V. Frolov, the works on the control of space-time parameters gain the possibility for development and commercial application. As a theoretical basis there are those N. Kozyrev works where his conception of "time density" are replaced by that of "aether density" according to Frolov.

In September 2002, Faraday Labs Ltd Company plans to complete testing of the first experimental system, and to start the patenting and research of applied aspects, first of all in medicine.

Physical Principles of the Time Machine

Alexander V. Frolov

Experimental success of research team headed by Dr. Vadim A. Chernobrov, Moscow was reported in [1]. The time course can be controlled as rate of any process in local space-time (inner space of the Time Machine). It can be decelerated or accelerated by means of special "converging electromagnetic waves". Ordinary waves move from the source whereas special "converging waves" move to some central point, i.e. into the focus of the system. In Chernobrov's design of the Time Machine this process is organized by means of several spherical envelops, which consist of several electromagnets. Electronic control unit controls the processes in this design. Dr. Chernobrov reported about 3% change of the time course in 4th version of the system, which was tested with a human inside. The goal of Dr. Chernobrov's work is to research the medical aspects and experimental investigation of the principles. Several important conclusions were obtained from the project: the time course can be controlled and character of the changes is different for acceleration and deceleration mode.

Other known publication and research projects on the same topic seem to be very far from any commercial and practically useful application. Obviously the topic is very new and fantastic for most of scientific community and at first we have to clarify the physical principles of the time control project, which is started by Faraday Labs Ltd.

In this project we believe that notion of time is one of possible description of real physical properties of our Universe. So, it is not mathematical abstraction but some aspect of physical reality and we can discover some physical properties of time. Russian astrophysicist N. A. Kozyrev [2] developed a theory of active properties of time and according to his point of view there are two properties: time course and time density. Prof. Kozyrev demonstrated experimentally that time density in area of some process (changes of matter) is dependent on entropy parameters of the processes. In [3] it was demonstrated that Kozyrev's experiments could be interpreted in aether theory and it has led to simple physical conclusions and clear experimental perspectives: **time course and its density can be explained and controlled as parameters of aether.** Directions of aether flow and density of aether are subjects of experimenting with non-reversible changes in matter, for example, in crystallization or melting processes. Also it is possible to use special electromagnetic processes, for example, Chernobrov's "converging waves" or other longitudinal waves as methods of aether compression or rarefaction. If we assume that process of existence of elements of matter physically can be explained as aether vortex processes then its rate is a parameter of **aether income/outcome balance** (aether inflow in element of matter and aether outflow from the element of matter). It was also described in Time Rate Control (TRC) theory [3]. To control this balance it is necessary to develop technology of longitudinal waves generation, its focusing and resonance effects. The previous research and experimenting on the topic has been made by N. Tesla.

Let's assume that we have some technology to change parameters of time course. How should we organize this local space-time (what is spatial topology of the design)? There is a very interesting experiment to get the answer: rotation of a heavy cone (for example, lead cone) entrains surrounding aether, so a vortex appears, which is a toroidal formation of aether (rings). The rings can exist in space for a long period. The further question is: Why does the beam of light (laser beam) directed to the cone by tangent create a luminous ring? We can assume that due to natural properties of photons (light propagates along the geodesic line in space) some autonomous closed toroidal space should be created in such experiment. The next thought is: since space and any matter exist in time then we can speak about some autonomous time. The general conclusion is to be the following: autonomous 4-dimensional space-time can be created as toroidal aether vortex.

Here is point to note some aspects of research project by Prof. Robert Mallett, Connecticut University, USA. In fact, sometime next year, he hopes to produce the first piece of technology that eventually will allow him to build a time machine. By Mallett it will be a device that employs lasers "to twist space". Why is he going to close the beam of light? His theoretical background is knowledge about black holes, i.e. understanding of the connection between gravity and curvature of space-time. In Einstein's theory both matter and energy can bend space and time. So Prof. Mallett assumes that curvature of space-time can be changed not only by mass (like a black hole) but it can be affected by energy of photons. This has led Prof. Mallett to consider the possibility of using a circulating beam of light to twist space and to create closed loops in time. It is predicted that a spinning neutral particle, when placed in the ring, is dragged around by the resulting gravitational field [4]. From the

Kozyrev-Dirak Radiation
Its influence on animals

Dr. Ivan M. Shakhparonov

International Academy of Energy-Informational Sciences

In the experiment with animals, that were made in the Center of Oncology Researches (COR) at the Russian Academy of Medical Sciences (RAMS), on the applying of Kozyrev-Dirak's Focused Radiation (KDFR), it has been found that KDFR decreases the quantity of glucose in the blood, reduces its tenacity, promotes the strengthening of immunity and the rise of the quantity of marrow cells.

Introduction

This time researchers in Russia and abroad experiment on ball lightnings by means of nonoriented circuits, which are similar to the electric analogues of Mobius band, also by means of Klein bottle and their combinations. Non-oriented fields are investigated very intensively now. Accordingly, organisms of the researchers, who observe the interactions of such fields with a matter, are also changed, thus they should take it into account on making such experiments. The aim of the article is to show in which way the fields of nonoriented circuits influence on animal and human organism. Besides the article has for its object the prevention of negative consequences, which can appear for experimenters through the research process.

Experiments with animals that were carried out in 1992-1993 in Russian Academy of Medical Sciences (RAMS) had not been published in proper time because there were no quantitative methods of radiation detection. Later, in 1996 they were developed [1] and KDFR parameters were measured in that geometry, which were applied in RAMS. In 1998 powerful and super-powerful KDRF sources were obtained. These sources were applied (and are applied now) in the researches at the controlled radioactive decay [2]. Kozyrev and Nasonov [3] and later also Lavrentyev with the collaborators [4,5] have proved experimentally that the Sun and some stars generate the radiation, which has early unknown properties. We suggest that the radiation, discovered by Kozyrev [3], and the radiation, which is researched by us and by other experimenters with nonoriented circuits, are of the same phenomenon. At first, it should be noted that on interaction between Kozyrev-Dirak radiation (KDR) and a matter made it colder. As it was demonstrated above [1], cooling effect can be explained by matter re-magnetization under the influence of KDFR beam (adiabatic demagnetization). According to the still unpublished data, KDFR bunch destroys matter lattice by the way of it's moving. However, after a couple of week matter reconstructs it to the almost tabulated points, without defects, blockness and other damages, which are peculiar to other natural crystal structures. **In the definite sense nature demonstrates the way to rejuvenate compound structures**. As it is well known, vital functions of biological systems on the Earth depend on the structure and composition of water. Therefore, we have a right to expect considerable changes in the vital functions of biological organisms under the influence of KDFR.

Experimental Devices

In experiments with animals there were applied the devices, which concentrated KDR (KDCR) and had 50 Wtt aggregate electrical power. The description is presented in [2].

KDFR indication was obtained by calorimetric method [1], along the way of movement of the main bunch (with 10 cm across diameter) and at angle of 45° from the geometrical axis of a device.

Researches of Bleeding Duration

Let us consider KDFR influence on the blood composition of animals. At the experiment 24-28 gram weighting, pelletized fed male mice were used. In the process of the experiment it was discovered that 3 and 4 hour processing of mice with KDFR at the distance 2.5 m and at the presence of animals in the sphere of maximum radiating power, caused some changes of fibrillation system. The bleeding duration was determined according to Duke method. Two groups of animals were used at the experiments: a group with 4-hour duration of KDRF processing and a group with 6-hour duration. Time of bleeding was considered in dynamics at 1, 2, 7, 14, 21, 28 and 35 day (Fig. 1). The bleeding duration of the intact animals was determined by the value 128±11 sec. After the applying of KDFR there was noticed some increase of bleeding duration to 261±15 sec and 223±21 sec on the first day after the stopping of the influence. In the subsequent periods bleeding duration gradually decreases up to the level of physiological norm. The whole normalization of the index is observed at the animals, which were processed by KDFR during 4 and 6 hours, on 28-35 day up to 115±12 and 133±18 sec correspondingly. In the process of observations at the animals, the correlation between time of fibrillation and periods of KDRF processing of the animals has not been revealed (Fig. 1).

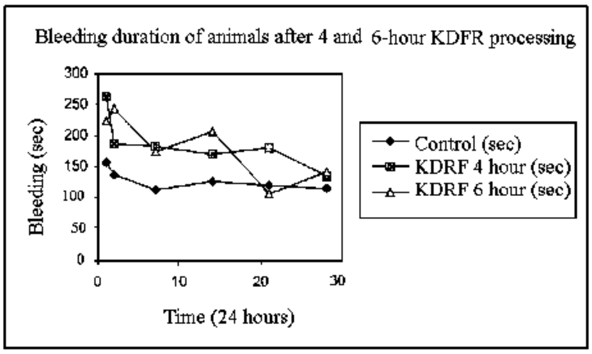

Fig.1

In the course of the experiment the strongly marked chronometric hypocoagulation was discovered due to the extension of the parameter "K" or, probably, because of the change of aggregation properties of platelets (Table 1).

For the determination of biochemical indexes serum was obtained from 5-8 ml of rats' venous blood. Beforehand (24 hours before the slaughter), forage was taken away from the animals. The determination of biochemical indexes was provided by means of biochemical analyzer HITACHI. As a result of the experiment it was determined that at the first day after influence of KDRF there was a tendency of decrease of the glucose content (Table 2). Other indexes varied in the limit of physiological norm.

Table 1 Parameters of thromboflexogramm after KDFR, 4 hours

Animal #	Parameters of thromboflexogramm			Fibrio gene	Fibrinal activity
	R (sec)	K (sec)	Ma (mm)		
1	72	∞	10	-	-
2	102	∞	18	-	-
1	90	∞	10	275	75
2	180	150	52	315	90
1	180	∞	5	-	100

Table 2 KDFR influence on the glucose content in blood of the rats

Time (days) after the experiment	KDFR 4 hours (mmole/l)	Test (mmole/l)	KDFR 6 hours (mmole/l)	Test (mmole/l)
1	3.14	6.12	6.39	7.27
10	7.59	9.35	8.90	6.69
30	6.05	6.69		

Research of haemopoiesis system

Several criteria were considered: the dependence of biological effect on the distance, on the power flux density, on the duration of processing. Besides, KDRF influence on mice survival was considered.

At the experiment 24-28 gram weighting, pelletized fed male mice were used. The marrow was examined in the dynamic at 1, 3 and 7 day after KDFR influence. Six animals were taken on each point. After the decapitation of the mouse their thighbones were taken out and after that the absolute number of myelokaryocytes was calculated by the standard method in Goryaev chamber.

Dependence of the biological effect on the distance

In all experiments the maximum flux density along the geometrical axis of KDCR device was a constant. There were used four temporal modes of the influence (1, 2, 3, 4 hours) and three points of long distance between KDSR and the biological object (0.5; 1.5; and 2.5 m). At 0.5 m distance there were no differences in the number of marrow cells in comparison with the control cells. With the increase of distance between KDCR and the object from 0.5 to 1.5 m some tendency to the increase of the number of marrow cells up to the 7th day was observed. Four-hour KDRF processing caused the increase of the number of karyocytes up to $29.99 \pm 1.25 \times 10^6$ ($P<0,001$). Thus, any dependence of the biological effect on the exposition has not been revealed. For instance, at the 7th day after one hour of the exposition the number of karyocytes was equal to $28.45 \pm 1.87 \times 10^6$ at the same time after 3-hour processing it came to $27.65 \pm 0.74 \times 10^6$. Alongside with the change of the distance to the biological object from 1.5 to 2.5 m the tendency towards the increase of the number of marrow cells has kept within the same limits $28.27 \pm 1.32 \times 10^6$ and $29.57 \pm 0.88 \times 10^6$.

Dependence of the biological effect on the power flux density

The comparative investigation of KDFR influence on the biological object in the coverage of KDCR (along its geometric axis) and outside the coverage has demonstrated that alongside with the increase of the radiation intensity there was a tendency towards the decrease of stimulative influence of KDFR on haemopoiesis.

Dependence of the biological effect on the duration of processing

On processing the animals at distances up to 2.5 m from KDCR and on increase of exposition to 3-4 hours it is possible to obtain reliably significant difference in the number of marrow cells from the physiological norm to the 7th day.

KDRF influence on the survival of mice

The experiments, determining the survival reaction of animals, were made by means of gamma radiation. 30-day survival is the criterion of determination. Conditions for the experiments are the following: in the coverage of KDCR and aside the coverage, (the distance between the KDCR and the object is 2.5 m in the coverage of KDCR and 0.5m outside the coverage). Time of influence is 4 hours. Animals of both sex were used. Two groups of animals were used. The test group was put to the gamma radiation in the diapason of doses, which caused marrowy syndrome, i.e. from 7.5 to 8.5 Gr. The second group of animals after the irradiation in the same diapason of doses was repeatedly processed with KDFR. Time of the influence is 4 hours at 7.5 Gr gamma radiation on 15 mice in one bath and 8.5 Gr on 15 mice in another bath. Total gamma radiation of the animals was made by means of the source ^{137}Cs with the dosage rate of 5.2 Gr/min. Gamma radiation in this dose diapason causes death of the animals during the development of the marrowy syndrome, i.e. from the 6th to the 20th day along with the aplasia of haematogenic tissue. Combination of gamma radiation and KDFR sometimes leads to the slight increase of the number of survived animals. If the animals are irradiated by gamma rays at first and then by KDFR, the death control at 7.5 Gr radiation is equal to 5.5% from the total number of the animals and at the following KDFR processing 16% of the irradiated animals die. However, 67% of the animals in the tested group have died after KDFR processing and after the coming next gamma radiation with 8.5 Gr total dose. And in the group, which was processed with KDFR, only 46% of the animals died.

Immunity strengthening

For the investigation of KDFR influence the following tests were chosen: activity of natural killers and T-killers, which had been obtained by the immunization in vitro in the unidirectional mixed culture of lymphocytes and also in the reaction of blast transformation on the specific mitogen [6, 7]. All tests were made on the 7th day after a single KDFR influence. Unfortunately, data have been obtained with the applying of radioisotope preparations. Though the experiments of this kind were successful and though they have demonstrated the increase of some immune reactions' level, there is a certain doubt in the relevancy of radioisotopes application [2]. Thus the series of experiments was made. These experiments were aimed at the investigation of KDFR influence on the development of the swelling process. The aim of the experiment is the investigation of KDFR influence on the development of Ehrlich cancer and sarcoma-37, which were repeatedly inoculated to mice. At the first stage of the experiment there was a single KDFR influence on the mice repeatedly inoculated with sarcoma-37, on the 2nd day after the repeated inoculation of the swelling cells to the animals. The repeated inoculation was made intramuscularly in a right thigh, in a dose 10^6 of cells per a mouse. Time of KDFR influence for the first group of animals was equal to 1 hour; for the second one it came to 2 hours; and for the third one it came to 4 hours. Each group consisted of 6 animals. The test group consisted of six mice with repeatedly inoculated sarcoma-37 and which had not been processed with KDFR. As a result, the average lifetime of tested animals was equal to 9 days. The average lifetime of animals of KDFR groups came to 48 days (for 1-hour KDFR influenced group); to 12 days (for 2-hour group); and to 31 days (for 4-hour group). Thus, the average lifetime of the experimental group came to 29 days. Besides, in the group, which has been processing with KDFR during 1 hour, **the half of mice had survived (three of six mice).**

At the second stage of the experiment the repeated (5 times during 2 hours) KDFR influence on the mice was applied. These mice have been inoculated with sarcoma-37 at seven days before the beginning of the influence. As a result, the average lifetime of the animals was equal to 27 days, and for the mice, which were processed with fivefold KDFR influence, the average lifetime was equal to 76 days. The obtained results are the evidence of inhibition of swelling development for the animals, which were processed with KDFR influence. This leads to the increase of lifetime of such animals in comparison with the test. **Thereby, at a great extent the results of the previous experiments on the strengthening of immune status after KDFR influence were confirmed.**

Results and discussion

Let us make a conclusion. At the KDFR influence on animals' organism the following effects were observed: decrease of blood viscosity; strongly pronounced hypocoagulation; decrease of contents of glucose. Increase in the number of karyocytes and the extended lifetime of the animals, infected with Ehrlich cancer and sarcoma-37 were also observed.

As for human being, the researches in this area have not been carried out yet and they are still confined to the single observations. It is possible to give an example from the author's practice. In 1975 nonoriented circuit of 3kWtt power was examined. Field strength was measured. The author of the article had been working in the field for about 8 hours. And after five hours after the experiment I had felt bad. That time it was nothing known about the influence of the new radiation on human organism. The arrived ambulance has quickly diagnosed that I was close to hypoglycemic coma. On several hours after the intravenous glucose injection, my state has become normal. **Now we know that before the experiments with powerful KDFR bunches it is necessary to eat sugar.** Thus we believe that the data, which were obtained after the experiments with animals, can be applied to a human being. We can suggest that the manifestation of the symptoms of the KDFR influence on human organism depends on the power of the applied source, on the total mass of the organism and on the time of it in the coverage of the irradiation. From aforesaid it is clear that the experiment with powerful KDFR sources is far from being harmless and it is better to make it distantly after exclusion of man presence near experimental stands and devices. At the same time it is quite obvious that on applying of small capacity and fixed time of irradiation it is possible to develop methods for curing of human diseases, which are considered now as incurable (for instance of diabetes, some diseases of haematogenic system, of cancer and possibly of AIDS.

References

1. Proceeding of the International Scientific Conference "New Ideas in Natural Sciences" Problems of Modern Physics, p. 176-187
2. Journal of new energy, Vol. 3, #4,1999, I.M. Shakhparonov "Interaction between Kozyrev – Dirak radiation and radionuclides", p. 85-89
3. Kozyrev N.A. Selected works, Leningrad State University, 1991, part 3
4. Lavrentyev M.M. Eganova I.A. Luzet M.L., Fominykh S.F./ / Reports by AS USSR. – 1990, Vol. 314, #2, p. 352-355
5. Lavrentyev M.M., Gusev V.A., Eganova I.A., Luzet M.L., Fominykh S.F.// Reports by AS USSR. – 1990, Vol. 315, #2, p. 368-370
6. Methodological materials on experimental pharmacological and clinical trials of immune modulating effect of pharmacological remedies. - Ministry of Health USSR, M., 1984
7. Talmadge J.E. and Chiragos M.A. Comparison of immunomodulatory and immunotherapeutic properties of biologic response modifiers. Springer Seminar Immunopathol., 1985, 8, 429-443

Effect of Magnetic Blow Wave Field on Wine Systems

I.M.Shakhparonov (Corresponding author), S.A.Grin,
S.R.Tsimbalaev, L.N.Kreindel, V.N.Kocheshkova,
A.I.Podlesny, S.Yu.Gelfand

AGD Firm, Peschanyi pereulok, House No.20P korpus No.1, Lfl. 33
125252, Moscow, A- 252, Russian Federation

Russian Institute of Canning Industry, Shkolnaya Street. 78.
142703 Vidnoe 3. Moscow Region, Russian Federation

Introduction

Authors communicate the data on influence of Magnetic Blow Wave (MBW) field on several wineproducts. It was found, that MBW did not lead to significant changes in the major components of the wineproduct (sugar, organic acids, minerals). At the same time the taste and aroma of treated wine become more pleasant; content of heavy alcohols and wine stone in the treated samples was less than in non treated ones. A mechanism of transformations was also discussed.

Keywords: *Magnetic Blow Wave (MBW), Wineproduct, GLC of aroma compounds and ethanol, HPLC of sugars, Atomic Absorption Spectrometry (AAS) of minerals, Heavy alcohols and aldehydes, Wine stone, Turbidity tendency, Organoleptic evaluation*

Magnetic Blow Wave (MBW) was obtained for the first time during the investigations on ball lighting generation under the laboratory conditions (Shakhparonov 1994). MBW as a physical object is interesting because of some facts, which suggest that MBW is a magnetic monopole. The MBW can also interact with the matter and transforms it in a definite way. Typical example is an elementary carbon in the form of graphite, which is transformed by such magnetic treatment into ferromagnetic substance (ibid). The graphite, which is initially diamagnetic, transforms to paramagnetic one with general radiation doze of about $7\cdot10^{19}$ neutrons/cm^2. Other types of radiations could not affect this way (Svoistva 1975). So one unit of MBW can be considered as $1\cdot10^5$ of neutron masses. This fact may be regarded as an indirect evidence for assuming that MBW and magnetic monopole are the same things. In the absence of excited radioactivity a slow MBW [$v/c < 1\cdot10^4$] occurs, which does not ionize atoms (Devons, 1963). Therefore, their interaction with the matter can be observed only indirectly. No data exist on the interaction of MBW with organic substances. The experiments and results reported in the present communication may be a starting point for development of technology and to formulate the methods for vintage wine and best quality spirit production.

Materials and Methods

Assuming that MBW and magnetic monopole are the same things, a number of conditions were selected for all experiments. The MBW source and the samples were placed in the same axis and the axis was oriented according to magnetic meridian direction. Such magnetic orientation is appropriate, as the energy of magnetic monopole theoretically increases in a magnetic field (Devons, 1963). All of samples were placed at 250 cm distance from MBW source, in hermetically closed glasses. It should be noticed that MBW could penetrate through many other barriers, for example into cast iron reservoir with wall thickness of 5 cm (Amaldi, 1970).

The quality investigations were made by using of standard equipment. HPLC, equipped with refractometric detector was used for sugars estimation. Separation of organic acids in forms of their ethyl esters and acid esters was carried out chromatographically using a column packed with polyethylenglycol succinate and the following temperature option: initial temperature is 120°C, final temperature is 220°C, temperature growth rate: 8°/min. GLC was also employed for determination of ethanol. Minerals content conclusion: in order to obtain complete information about any system, it should be destroyed. However, destruction of tissues of the man in order to get information about their state is a too high price to pay for the information about his health.

However, the above Van Hoven's criterion can be satisfied with the minimum influence, when the cells are not destroyed and the atoms of these cells, being primary sources of torsion spectrums to be registered, are bring into the non-equilibrium state by means of outer disturbing influence.

In order to choose the frequency of the disturbing torsion influence correctly, it is necessary to take into account the role of water in physical and biochemical organization of tissues of the human organism.

At the same time, it is necessary to take into account the resonance torsion frequencies of various human organs. Finally, it turns out that the signal of torsion disturbance should be rather sophisticated considering both these factors. The TORDI system is a ready-to-use production device. Nevertheless, it is important to understand that the model is not the limit of scientific and technical potential incorporated in it and that enhanced variants of the system will appear with the course of time.

Summing up, I would like to draw your attention once more to the fact that work on torsion technologies is not limited by the directions that were discussed here. Actually, as it was pointed out in the beginning, on-going development includes all branches of economy, industry, agriculture and medicine, as well as all problems of everyday life. Technologies that we mentioned are the forerunner of the fact that the mankind is on the threshold of the age of torsion technologies, which, we believe, will change our life in the 21st century more than all the scientific and technical revolutions of the 20th century.

Reference

1. Cartan E. Compates Rendus. Akad.Sci., Paris, 1922, v.174.
2. V.Melnikov, P.Pronin. Problem of gravitation constant stability and additional interactions. Itogi Nauki I Tekniki, ser. Astronomy, v.41, Gravitation and Astronomy, Moscow, VINTI, 1991.
3. G.I.Shipov Theory of the Physical Vacuum. Nauka, Moscow 1997.
4. I.Ternov, V.Bordovitsin. On the modern interpretation of classical spin theory of Ya.Frenkel. UFN, 1980, v.132, No.2.
5. V.Bagrov, V.Bordovitsin. Classical spin theory. Izvestiya VUZ, Phys.Series, 1980, No.2.
6. F.I.Belinfante. On the spin augular momentum of mesons. Physica VI, 1939, v-6, No.9.
7. M.Markov. The very early universe. Proc of the Nuttfield Workshop, Eds. G.V.Gibbson, S.W.Hawking, S.T.Siklov, Cambridge, 1988.
8. Hideo Uchida.A method apparatus for detecting a fluid. Patent England, No 511662, May 24, 1978.
9. Anatoly Akimov. Heuristic discussion of the problem of finding long-range interactions. EGS-concepts. CISE VENT, preprint N7A, Moscow, 1992.
10. IITAP RANS, TORTECH USA, Horizonts of the Science and Technology XXI age Proc, Editor A.E.Akimov, Folium, Moscow, 2000, vol.1 (in Russian).
11. N.Kozyrev, Astronomical observations by means of the physical properties of time. In "Flarestarse" International Symposium in Byurakan, 1976,Armenian Academy of Sciences Publ., Yerevan, 1977 (in Russian)
12. Bouwmecster D. et.al. Nature, v.390, 11 dec, 1997.
13. G.I.Shipov. Theoretical and Experimental Research of Inertial Mass of Four-Dimensional Gyroscope. ITTAP RANS, preprint N10, Moscow, 2001, (in Russia).
14. The way of correction of metal alloy microstructure. Patent Russian, RU 2107105, 1998.
15. A.Dolgov, Yu.Zel'dovich, M.Sazshin. Cosmology of the Early Universe. MGU Publ., Moscow, 1988.
16. I.A.Wheeler. Einstein's vision. Springer Verlag, 1968.
17. Convegno Internazionale:Quale Fisica per 2000, Proc. Bologna, 1991. see: The Manual of Energy Devices and Systems.
Complied D.A.Kelly, D.A.K. WLFUB, Burband, California, 1986, Publ. N1269/F-289.
18. Daytlov V.L. Polarization Model of the Heterogeneous Physic Vacuum. Institute Mathematical, Sibirians Academic Science, 1998 (in Russian).

The Electrical Vortex Non-Solenoidal Fields

Sergey B. Alemanov
E-mail: alemanov@inbox.ru Phone 7 (095) 323-6848

A mistake was found in the electrodynamics: it is detected that all electrodynamics' postulates corresponds to the experimental facts, but vortex electric fields has unclosed inductive lines.

When the magnet is moving, then the current of magnetic induction is moving together with it. From known velocity of motion **v** and the value of magnetic induction **B**, it is possible to calculate the intensity **E** of appearing vortex field according to electrodynamics formula of transformation of fields **E=vB**.

If to change the E=vB on induction $D=\varepsilon_0 E$ in formula of fields' transformation, that will get $D=\varepsilon_0 Bv$, where D is electric induction, B is magnetic induction, v is velocity of motion, ε_0 is electric constant.

Herewith the appearing electric induction is always transverse to the direction of motion. It is possible to formulate the rule of origin for electric induction under the condition of rectilinear motion: if to dispose the right hand palm so four fingers shows the motion direction of the magnetic flow (the field), connected with moving magnet, and the vector B fells into palm, then the moved aside big finger will indicate the direction of vector D. The given rule is like the rule for Lorenz' force, but on the contrary (the difference is in frame). In the first case the charge moves, but the magnet rests. Here the magnet moves, but the charge, which points the direction for lines of force of electric induction, is immovable. So, there it is the rule for left hand, but here, on the contrary, it is the rule for right hand. Thereby, if the charge moves, but the magnet is immovable, then the rule of left hand uses for determination of the force. But if the magnet moves, but the charge rests, then the rule of right hand uses for determination of the force.

The origin of electric force is connected with that, the vortex electric field $D=\varepsilon_0 Bv$ appears around moving magnet (the magnetic field does not act on immovable charges).

In common literature on electrodynamics there is no any difference between electric vortex field and solenoidal field, but these are different notions. The sign of solenoidal field is the closed lines of electric induction (the flow of vector D through the closed surface is a zero), but for the vortex field the sign is following: the work of forces can be different from zero under the condition of motion along a closed line. That is to say, the vortex fields can agitate the rotational currents.

From the electrodynamics textbook: "The work of forces of vortex electric field can be different from zero, when the electric charge is moving along a closed line."

For instance, when the magnet moves, the vortex electric field appears and this field can be solenoidal or not, depending on magnet's orientation. Let's take such example: the magnet moves evenly, rectilinearly, and it's poles are oriented transversely to direction of motion. According to the rule of origin for electric induction ($D=\varepsilon_0 Bv$ that is the rule of right hand), the appearing vortex electric flow is not a solenoidal, since the lines of electric induction are not closed. Its begins in one conditional area of disturbance (+), accompanies the moving magnet, and it finish in another area of disturbance (-). For presentation it is enough to consider only two areas (+) and (-), represented on Fig.1. These dissimilar areas of disturbance appears because that flow of magnetic induction inside the magnet has the inverse direction, that outside the magnet.

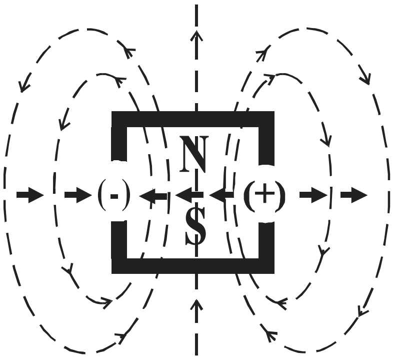

Fig.1

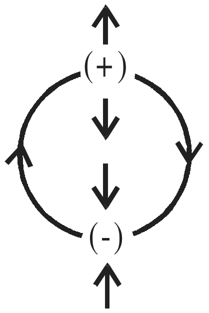

Fig.2

That moving disturbance of electric and magnetic fields presents itself as transverse electromagnetic disturbance. Also, it is necessary to notice, that under such magnet's motion, the appearing vortex electric field is not closed, but the current of electric displacement, connected with it, is closed (a currents are always closed). In given example, for clarity, it is possible to present a intensity of electric field through the Lorenz' force, if to take the frame, in which the magnet rests, and the test charge moves.

On the Fig.1 the moving magnet is conditionally represented (motion is toward to the text, magnet is moving away). N and S are poles of magnet. The direction of lines of electric induction, appearing when the magnet is moving, specified by arrows → and ←. Part of the lines begins in positive area (+) and finishes in negative area (-), the areas are placed on the ends of magnet. **The flow of electric induction through closed surface is not a zero; that is to say, these areas of disturbance are moving electric charges.**

From the electrodynamics textbook again: "The flow of vector D through any closed surface is equal to algebraic amount of external charges, covered by this surface. In the electrodynamics these postulates has the same role, as Newton' laws in classical mechanics."

Thereby, according to postulate, it is necessary to consider the appearing dissimilar areas of disturbance (+) and (-) to electric charges, or it is necessary to change the postulate.

It is interesting, that a part of lines of electric induction, which placed frontal and behind magnet, starts and finish at infinity, since the distribution of magnetic induction around magnet has not determined borders.

For clarity, it is possible to make following calculation. For instance, the coil (loop or turn) with current, as a magnet, moves evenly and rectilinearly, but its magnetic poles are oriented transversely on motion direction. Under such motion the lines of electric induction are not closed, and the dissimilar areas of electric field's disturbance appears in space on the edges of this coil.

On Fig.2 the moving coil with current is conditionally represented. It moves from left to right side of the page. The arrows on the coil indicate the direction of current. The appearing dissimilar areas of disturbance of electric field are marked by signs (+) and (-). Knowing, that in medium of the coil $B=\mu_0 I/2r$ and according to $D=\varepsilon_0 Bv$, it is possible to find the electric induction, appearing in the center, between two dissimilar areas $D=\varepsilon_0\mu_0 Iv/2r$, where I is current in the coil, r is radius of the coil, v is velocity of motion, ε_0 is electric constant, μ_0 is magnetic constant. The electromagnetic disturbances in transverse electromagnetic waves has the similar field construction, there also dissimilar areas of disturbance of electric field exists, that is to say the lines of electric inductance are not closed. Only the currents of electric displacement and magnetic induction are closed.

Let's consider another example: magnet moves rectilinearly, but its poles are oriented longitudinally to direction of motion. According to the rule for origin of electric induction ($D=\varepsilon_0 Bv$ is the rule of right hand), the appearing rotational electric flow is solenoidal, since in this case the inductive lines become closed lines. **Usually in books on the electrodynamics such moving magnet is considered, and the wrong conclusion is thereof done, that vortex electric field is always**

solenoidal, herewith it is forgotten, that poles of the magnet can be oriented not only along the direction of motion, but across also.

From the electrodynamics textbook: "The vortex electric field differs from electrostatic field that it is not related with any electric charges and its lines of intensity are closed lines."

From theory and from experiments it follows, that under transverse motion of magnet the lines of disturbance of vortex electric field can be unclosed and, accordingly, the flow of induction through the closed surface is not a zero. Then there is a direct discrepancy to facts in modern electrodynamics. It is strange, but for the whole history of researches in magnetism the transverse magnet's motion was not is considered. It leads to revising of electrodynamics' postulates, which plays such role in electrodynamics, as the Newton's laws plays in classical mechanics. The postulates, giving invalid belief about field processes, accordingly, do not allow to make some correct calculations. Fallaciousness of these postulates was one of the reasons, on which the electrodynamics could not to consider and to calculate the discrete electromagnetic waves (photons), where the magnetic field also is the transverse field (the field construction and calculation of photons are represented on the page http://www.comail.ru/~alemanov). **That is to say, not only particles has the charges, but areas of disturbance of field (without particles) are the charges also,** where the flow of electric induction through the closed surface is not a zero. Thereby, the vortex electric fields can be not only as closed flows of induction, but as well as inducted electric charges, accordingly, the laws for electric charges are valid for induced electric charges also. For instance, in the law of conservation of charge: if somewhere the area of disturbance with positive sign appears, that negative area appears also.

From the electrodynamics textbook: "The vortex electric field is generated by the variable magnetic field. Its force lines are always closed, like force lines of magnetic field."

But before this fundamental postulate, confirming, that force lines of vortex electric field are always closed, it was necessary to consider all variants of change for the magnetic field, including the variant of the transverse motion of the magnet. That is to say, the consideration of physical processes could not be unilateral. Faraday considered the longitudal motion of magnet and discovered the electromagnetic induction, but the transverse motion of magnet that have the principle importance for understanding of field processes in electrodynamics was not considered. Thereby, the longitudal motion of magnet brings to arising a vortex electric field with closed force lines, but transverse motion of magnet brings to arising a vortex electric field, where the lines of forces are not closed. In this case it lead to induced electric charges. It is necessary to notice, that this is first mistake, detected in electrodynamics postulates for all time of existence of electrodynamics.

From the electrodynamics textbooks: "…Gauss' theorem is valid not only for electrostatics, but also for electrodynamics, which using a variable in time electromagnetic fields. We are not sure if this hypothesis is valid or it is not valid… Only the experiment can give the answer on this question. The whole collection of experimental facts speaks in favor of this hypothesis." But, unfortunately, the experiment with transverse motion of magnet was not considered seriously in this textbook.

(Editor's note: Well-known Searl's experiments and Godin & Roshchin's experiments are based on such transverse motion of magnets (rollers). In Alemanov's article it was demonstrated that in this case the experiment should lead to induced electric charges. Really it was detected in experiments. Hence this missed aspect of electrodynamics is very important for development of the new energy technologies.)

Gravito-Inert Mass

J.A. Asanbaeva

720000, Kyrgyzstan, Bishkek, Kadyrov's Scientific Center
+996 (312) 47-25-40, +996 (312) 65-02-83
bondarenko@nazaraliev-centre.com

Nature of mass is one of the important problems of modern physics. It is accepted to consider that the mass of elementary particle is determined by fields, which are connected with it (electromagnetic, nuclear and others). However, we didn't create any quantitative theory of mass. There is no theory to explain why masses of elementary particles form a discrete spectrum of values and to allow determining this spectrum.

Mass (m) is a physical value, one of characteristics of matter, which defines its inert and gravitational properties. Accordingly, we distinguish inert mass (m_i) and gravitational mass (m_g).

Inert mass (m_i) characterizes dynamical properties of a body, its property to accelerate under the action of the force $\left(\vec{F}_i\right)$ and according to the second Newton's law is considered to be constant coefficient of proportionality for the given body between $\vec{F}$ and acceleration $\vec{a}$.

$$\vec{F}_i = m_i\vec{a} \qquad (1)$$

Gravitational mass (m_g) is a source of gravity field. Every body creates its gravity field, which is

18. Kozyrev N.A. *Selected Transactions* (Leningrad University Press, Leningrad, 1991) (in Russ.).

19. Lavrent'ev M.M., Eganova I.A., Medvedev V.G., Olejnik V.K., and Fominykh S.F. *On Scanning of Celestial Sphere by Kozyrev's Sensor.* Doklady AN SSSR, 323 (4), p.649-652 (1992) (in Russ.).

20. Akimov A.E., Kovaltchuk G.U., Medvedev B.P., Oleinik V.K., Pugatch A.F. *Preliminary Results of Astronomic Observations of the Sky by the Kozyrev Procedure.* Academy of Sciences of Ukraine, Central Astronomic Observatory. Preprint CAO-92-5P, 1992, p.16 (in Russ.).

21. *Oleinik V.P. Superluminal Transfer of Information in Electrodynamics.* SPIE Material Science and Material Properties for Infrared Optoelectronics. 3890 p. 321-328 (1998) (http://www.spie.org /).

22. Oleinik V.P. *Faster-than-Light Transfer of a Signal in Electrodynamics.* Instantaneous Action-at-a-Distance in Modern Physics (Nova Science Publishers, Inc., New York, 1999), p.261-281.

23. Eganova I.A. *The World of Events Reality: Instantaneous Action as a Connection of Events through Time.* Instantaneous Action-at-a-Distance in Modern Physics (Nova Science Publishers, Inc., New York, 1999).

24. Lavrent'ev M.M. and Eganova I.A. *Physical Phenomena Predicted and Revealed by N.A.Kozyrev, in the Light of Adequacy of Space-Time to Physical Reality.* Philosophy of Science, 1(3), p.34-43 (1997) (in Russ.).

25. Logunov A.A. *Lectures in the Theory of Relativity and Gravity.* A Present-Day Analysis of the Problem. (Nauka, Moscow, 1987) (in Russ.).

26. Oleinik V.P. *The Newest Development of Quantum Electrodynamics: Self-Organizing Electron, Superluminal Signals, Dynamical Inhomogeneity of Time.* Physical Vacuum and Nature. 4 p.3-17 (2000) (in Russ.).

27. Jefimenko O.D. *Electromagnetic Retardation and Theory of Relativity* (Electret Scientific Company, Star City, 1997).

28. Oleinik V.P., Borimsky Ju.C., Arepjev Ju.D. *Time, What is this? Dynamical Properties of Time.* Physical Vacuum and Nature. 5 p.65-82 (2001); *New Ideas in Electrodynamics: Physical Properties of Time.* Semiconductor Physics, Quantum Electronics and Optoelectronics. 3 N4 p.558-565 (2000). E-print: quant-ph/0010027.

29. Borimsky Ju.C., Oleinik V.P. *The Course of Time in Classical and Quantum Systems and Dynamical Principle.* Physical Vacuum and Nature. 6 (2001) (in print).

The Principle of Self-Organization, which can be formulated as follows: any material object represents an open self-organizing system whose internal structures are formed with the participation of the whole universe. Apparently, the Principle of Self-Organization, incorporated in nature as one of the integral properties of matter, is nothing more nor less than a spirit (or absolute idea, or creator) which operates the world and creates all its variety.

Physical Mechanism of Nuclear Reactions at Low Energies

V.P. Oleinik* and Yu.D. Arepjev

> *Tell me what the electron is,*
> *and I shall explain to you everything else.*
> *W. Thomson*

Abstract

The physical mechanism of nuclear reactions at low energies caused by spatial extension of electron is considered. Nuclear reactions of this type represent intra-electronic processes, more precisely, the processes occurring inside the area of basic localization of electron. Distinctive characteristics of these processes are defined by interaction of the own field produced by electrically charged matter of electron with free nuclei. Heavy nucleus, appearing inside the area of basic localization of electron, is inevitably deformed because of interaction of protons with the adjoining layers of electronic cloud, which may cause nuclear fission. If two or more light nuclei occur "inside" electron, an attractive force will appear between the nuclei that may result in the fusion of nuclei. The intra-electronic mechanism of nuclear reactions is of a universal character. For its realization it is necessary to have merely a sufficiently intensive stream of free electrons, i.e. heavy electric current, and as long as sufficiently a great number of free nuclei. This mechanism may operate only at small energies of translational motion of the centers of mass of nuclei and electron. Because of the existence of simple mechanism of nuclear reactions at low energies, nuclear reactor turns out to be an atomic delayed-action bomb, which may blow up by virtue of casual reasons, as it has taken place, apparently, in Chernobyl. **The use of cold nuclear reactions for production of energy will provide mankind with cheap, practically inexhaustible, and non-polluting energy sources.**

Introduction

Nuclear reactions at low energies, occurring in physical and biological systems, and, in particular, the cold fusion (CF) of nuclei, attract ever increasing attention (see review articles [1,2]). This is explained by the fact that research on CF (in what follows, by cold fusion we shall understand any nuclear reactions at low energies) opens up the way to the solution of the problem which was set more than 50 years ago in the field of controlled thermonuclear reactions (CTR) and which has not been solved that is the problem to provide mankind with cheap fuel. An important point is that CF allows to create not only cheap, but also non-polluting energy sources, as nuclear reactions at low energies are not accompanied by radiations dangerous to health (γ-radiations, streams of fast neutrons and other particles). Note that the energetic problem facing mankind is presently of special interest in connection with the fact that, according to expert evaluations, the oil-and-gas resources in the world will suffice only for some decades. For this reason the study of CF is among the most important problems of physics.

It is necessary to note that, relying on the standard theory of nuclear reactions describing nuclear processes in vacuum, experts in the field of nuclear physics, engaged in CTR, reject the very possibility of existence of nuclear fusion at low energies. Two basic objections are raised against CF:

1. at low energies the penetrability of Coulomb barrier around nuclei is so small that the probability of nuclear fusion is practically equal to zero;
2. distinction between the atomic and nuclear energy scales is so great that the energy, which might be evolved as a result of nuclear fusion, could not be transferred directly to atomic lattice; therefore the energy above should be emitted in the form of streams of γ-quanta, fast neutrons and other particles. However, such streams of sufficient intensity have not been registered.

The answer to the first objection against existence of CF is that at the heart of CF are nuclear processes occurring in environment, and the basic role is played here, apparently, by collective effects caused by interaction of nuclei with particles of environment in which the nuclear reaction takes place. The laws governing the behavior of interacting nuclei in vacuum are inapplicable to the description of CF of nuclei [3]. Nuclear reactions occurring at low energies submit to completely different laws, which can be established only provided that collective effects mentioned above are taken into account. For this reason the standard theory of nuclear reactions in vacuum can by no means refute the existence of CF.

As to the impossibility of transferring the energy between levels of various scales, we can give an example of the phenomenon of sonoluminescence (luminescence of a liquid when a sound wave causing cavitation passes through it) [4], in which the energy transfer from an acoustic wave to electromagnetic field occurs with appreciable probability in spite of the fact that the distinction between energies of acoustic phonons and quanta of light reaches 11 orders.

As early as 10 years ago J. Schwinger, the Nobel winner and the known expert in the field of the theory of elementary particles and quantum electrodynamics, asserted that it is impossible to deny the reality of CF phenomenon [3,4]. Since then the CF phenomenon for nuclei was repeated hundreds times in laboratories all over the world, tens of patents on the ways of energy generation on the basis of CF were registered and enormous number of experimental works were published, which not only confirmed the existence of effect, but also contained its detailed analysis.

The most convincing evidence for the existence of nuclear reactions at low energies seems to give the mass-spectrometric research of reaction products [5] as well as research on biological systems [6]. Detailed study of electric explosion of foil made of especially pure materials in water, described in [5], suggests that at electric discharges transformation of chemical elements occurs. Study of optical spectrum of plasma arising at discharge and of the mass-spectrometric analysis of sediments, which remained after the discharge, shows that in plasma there is an appearance of a significant number of chemical elements which were not presented in the initial material of explosive foil and electrodes and also that the isotope structure of the foil material changes appreciably. The change of experimental conditions, for example, of energy contribution in foil, its mass and dimensions results only in redistribution of intensity of plasma spectral lines, i.e. in the change of statistical weight of chemical elements in plasma, but the composition of chemical elements remains unchanged and it essentially depends on the material of foil. As it is seen from the received results, nuclear reactions, which take place at electric discharge, are not accompanied by the occurrence of a neutrons stream and γ-radiation and proceed at low energies of atomic nuclei.

The research mentioned above as well as many others, carried out by different researchers in different laboratories, **allow to draw a conclusion that existence of nuclear reactions at low energies is reliably established.**

The development of research on CF is hampered by the absence of theory of the phenomenon. As noted by Schwinger [3,4], the situation in CF is closely parallel to that one in high-temperature superconductivity: reality of the last, as a result of careful experimental research, is completely established, though theory of the phenomenon is absent till now.

In [5], to account for the transformation of chemical elements, the hypothesis is put forward that at the electric explosion of foil in the plasma channel magnetic monopoles are formed which may overcome the Coulomb barrier even at insignificant kinetic energy due to the great magnitude of their magnetic charge. The monopole, appearing not far from a nucleus, causes its polarization: those nucleons of the nucleus, which are situated more close to the monopole, experience stronger influence of the last, than the nucleons situated on the opposite side of the nucleus. As a result, a deformation of the nucleus arises (the nucleus is lengthened), which may result in nuclear fission.

Obvious drawback of this mechanism of nuclear reactions is that magnetic monopoles have yet to be found out in nature.

Numerous attempts to construct a consistent theory of CF (see reviews [1,2]) have not been crowned with success. As it was noted above, for the CF to be described, the account of the collective effects may be important caused by interaction of nuclei with environment, in which nuclear reaction takes place. But does it suffice to take into account these effects in order that the theory of the phenomenon is constructed? The analysis of the experiments on transformation of chemical elements at low energies and on the CF of nuclei suggests that the discussed phenomenon does not fall within the domains of exotic ones: it seems to occur in nature constantly, at every step, in both physical and biological systems. Therefore, it is natural to expect that nuclear reactions at low energies should have a simple physical explanation.

However such explanation, which is not beyond the scope of existing representations, is yet to be found. Does not it mean that we are facing here the situation similar to that which has arisen in physics at the end of the 19th century and which has been figuratively described in the words: on the light sky of physics there are only two small dark clouds – the radiation of absolutely black body and the Michelson experiments? Let us remind that in order for these clouds to be removed, it has taken the revision of physical notions about electromagnetic field as well as about space and time.

As is noted in [8], there is a simple physical mechanism of nuclear transformations at low energies which existence follows from the quantum theory of electron as an open self-organizing system [9]. If two or the greater number of light nuclei appear inside free electron, more precisely, inside the area of basic localization of the particle, because of interaction of nuclei with electrically charged matter of electronic cloud, a force of attraction appears between the nuclei which may result in fusion of nucleus. This means that **cold nuclear reaction represents an intra-electronic process** which character is defined by physical properties of the own field produced by electrically charged matter of electron. The purpose of this paper is more detailed consideration of the mechanism above stemming from the spatial extension of electron.

In section 2 physical ideas are formulated and basic results are schematically presented of quantum theory of electron as an open self-organizing system. The theory outlined is necessary to elucidate the origin of the mechanism resulting in the occurrence of nuclear reactions of fusion and fission at low energies. The essence of the developed approach consists in that **the own field created by electron is treated as a congenital, integral physical property of electron, intrinsically inherent in the particle by the very nature of things and for this reason the own field and self-action are included in the definition of the particle** at the initial stage of formulating the theory. As is seen from the received results, electron represents a quantum (elementary excitation) of the field of electrically charged matter. It is a solition, which physical and geometrical properties are described by the non-linear and non-local dynamical equation similar to the known Dirac equation.

In section 3 the application of quantum model of self-organizing electron to nuclear reactions at low energies is considered. It is noted that because of the presence of simple physical mechanism of nuclear reactions at low energies, which is of a universal character, nuclear reactors represent, in effect, nuclear delayed-action bombs, which from time to time may blow up by virtue of the casual reasons. Hence, though nuclear stations may provide mankind with energy, however atomic engineering is a very dangerous way of energy production. The only acceptable way of solving the energetic problem consists in the use of nuclear reactions at low energies.

Quantum model of electron as an open self-organizing system

The basis for the standard formulation of quantum electrodynamics (QED) is the hypothesis that electron is a structureless point particle which does not experience self-action. This assumption results in serious difficulties – the divergences of mass and charge of electron and the impossibility to explain stability of the particle (see, for example, [10-12]).

The difficulties mentioned above are very serious. According to Dirac, **the difficulties of QED "in view of their fundamental character can be eliminated only by radical change of the foundations of the theory, probably, radical to the same extent as transition from the Bohr orbits theory to modern quantum mechanics"** ([13], p. 403). "Correct conclusion", Dirac emphasizes, "is that the basic equations are incorrect. They should be changed in such a way that divergences do not appear at all".

The main reason of occurrence of difficulties is the assumption that electron is a point-like particle. Therefore, abandonment of this hypothesis is inevitable. As an analysis of the problem shows, the key to constructing a consistent quantum theory of electromagnetism lies in taking account of the Coulomb self-action of electron, i.e. the back action of the own field created by charged particle in environmental space upon the same particle. In the special case that the particle is at rest in an inertial reference frame, own field of the particle turns into static Coulomb field.

E.Schrödinger who suggested the historically first physical interpretation of quantum mechanics put one of the boldest ideas concerning the problem of electron forward. According to Schrödinger's hypothesis, the quantity $e|\Psi(\boldsymbol{r})|^2$ (e and $\Psi(\boldsymbol{r})$ are charge and wave function of electron, respectively) is the density of spatial distribution of electron's charge and, consequently, **the linear sizes of electron are the same as those of atom** [14,15]. However, they did not succeed in substantiating the interpretation and, for this reason, it was rejected by the majority of physicists [16].

An important step to the correct understanding of the physical nature of electron was made by A. Barut and by his collaborators [16-18] who formulated and developed quantum theory of electromagnetic processes on the basis of self-energy picture (the Self-Field QED). Using expression for the total own energy of electron, they managed to calculate the Lamb shift and other radiative corrections and to show that radiative phenomena may be described in terms of the action function, without using the second quantization method. As is pointed out by Barut [17], "the correct quantum equation of motion for radiating electron is not the Dirac or the Schrödinger equation for bare electron, but an equation containing an additional non-linear self-energy term".

New lines of approach to the problem of electron are offered in [9, 19-24]. The formulation of electrodynamics is considered which represents a synthesis of standard quantum electrodynamics and ideas of the theory of self-organization [25]. The physical mechanism of self-organization of electron consists in self-action. Taking into account the self-action means that electron is treated as a feedback system.

Let us outline schematically the results of the formulation of quantum electrodynamics in which electron is an open self-organizing system.

Editor's note: The authors develop mathematics by using Lagrangian functions, 7 equations. You can contact the authors for more information about.

Thus, the negative result is received: we have tried to take into account self-action of electron in a natural way by supplementing the Lagrangian function with the self-energy term, but we came to an equation that has no reasonable physical solutions at all. This result seems to mean that the standard theoretical scheme reaches here the limits of its applicability and so, remaining in its framework, it is impossible to solve the problem of electron and elucidate the physical nature of electromagnetic interaction.

Essentially new point, which is introduced in [9] into quantum mechanics consists **in the replacement of the model of isolated system described by harmonic oscillator with the model of open system**. Let us advance the arguments indicating the inevitability of using the model of open system as a basis of the description of interaction between microparticles [26].

Note, first of all, that quantum particle theory based on the use of the models of isolated system is, strictly speaking, physically meaningless. Really, any observation conducted on a system represents a process of interaction of the system with the means of observation. But in case of microparticles (quantum particles) this interaction is not weak and consequently it is inadmissible to neglect it, i.e. microparticles should be necessarily considered as essentially non-isolated systems.

A starting point of the standard formulation of quantum mechanics is the physical idea that interaction between physical fields can be reduced to collision of the particles corresponding to these fields, the particles before and after collision being considered as free ones. According to these representations, quantum mechanics is based on the notions of "bare", non-interacting particles, with the interaction between them being considered as an additional factor which can only insignificantly alter the physical properties of non-interacting particles. However, such an approach to interaction between physical fields is obviously of an idealized character because particles constantly interact "with vacuum as with some kind of physical medium in which the particles move" [27]. Interaction of particles with vacuum fluctuations is not small and it cannot be removed.

It is well also to bear in mind that the necessary intermediary at studying micro-objects are the means of observations (the devices) with the classical field corresponding to them which should be taken into account in consistent quantum theory [28]. Inclusion in theoretical scheme of arbitrarily weak classical external field results in occurrence of non-zero width Γ of energy levels of "dressed" particles. The basic impossibility to isolate a real particle from vacuum fluctuations of the field and from the classical sources connected to the means of observation is indicative, thus, of necessity to take into account the non-zero width of energy levels of real particles [26].

The use of the harmonic oscillator model, when describing the interaction of electromagnetic radiation with substance, seems to be the main source of serious difficulties of the standard formulation of quantum theory, as such an approach means apparent neglect of those physical processes which, proceeding constantly, are responsible for inseparable coupling of real physical system to surrounding medium. Introducing artificial notion about switching on and switching out of interaction of oscillator with radiation field, we are able to calculate within the framework of existing theory the width of energy levels of oscillator, but we cannot assert with certainty that such an approach results in correct description of interaction.

From the reasoning given above it is seen that they are the models with energy levels of non-zero width that should form the basis for the description of interaction of radiation with substance. It is necessary to formulate such a quantum theory, which would take into account the energy levels of non-zero width Γ. The case in point is that one should introduce an infinitesimal damping Γ into the initial set of equations describing interaction of charged particles with electromagnetic field. Such an approach means the violation in infinitesimal of homogeneity of physical system relative to translations in time. Necessity of violating the homogeneity of time follows from that fact that in the usual approach (with $\Gamma = 0$) the states of the system of interacting fields have degeneracy of infinitely large multiplicity in relation to time translations. According to the fundamental Bogoliubov's concept of quasi-averages [29], when describing the behavior of degenerate systems, one should include into Hamiltonian an infinitesimal term removing degeneracy. In the theory presented here degeneracy of states of quantized fields relative translations in time is removed by introducing the infinitesimal damping Γ into Lagrangian. Thereby the degeneracy under study is removed already in the initial, zero-order approximation, which is of fundamental importance for the approach based on perturbation theory.

Formulation of the physical idea that quantum friction arises at the very elementary level - at the level of one particle is given in monograph [26]. Impossibility to isolate real particle from the surrounding world is that property which should be taken into account already in the one-particle theory (for each kind of particles), even before switching on the interaction with other particles. Model of the particle as an open system ($\Gamma \neq 0$) is attractive owing to the fact that from the very beginning the degeneracy of states relative to time translations is absent in it, the degeneracy, which is removed in standard approach by taking into account the interaction of particle with vacuum field fluctuations and classical fields. The basis for the developed formulation is the fundamental concept of quasi-averages supplemented with the requirement that the equations of motion of the particle with $\Gamma \neq 0$ follow from the action principle. It should be emphasized that the non-zero damping Γ is introduced into electrodynamics with the aim to establish the structure of the Lagrangian function, which takes into account the property of openness of physical system. After establishing the structure, the limiting transition $\Gamma \to 0$ is fulfilled.

In our opinion, the development of quantum theory will be inevitably connected with the use of models of open system; as such models reflect more completely the physical essence of interrelations in the real world. It is necessary, thus, to define more exactly the concept of openness of physical system, which, on the one hand, would describe real system accurately enough and, on the other, would be simple enough to describe the particular physical processes.

As open system has the richer physical contents in comparison with isolated system, some essentially new mathematical ideas are needed for its description. First of all, it is necessary to increase the number of independent dynamical variables describing the particle as open system. In papers [9,19-24], as a basis for the description of self-acting electron, the simplest model of open system is used which can be described by the Morse-Feshbach-Bateman Lagrangian function [30,31] and which was successfully used for the description of dispersive medium (the review of articles, in which applications of the model of open system to electrodynamics of dispersive medium are considered, is given in monograph [26]). In this model the number of dynamical variables is doubled as compared with the isolated system, namely, to each dynamical variable of "bare" particle, Ψ , there correspond two dynamical variables, which are denoted by Ψ and $\tilde{\Psi}$. These quantities are considered as components of the wave function describing the quantum state of self-acting particle. One of them, say, Ψ , corresponds in a sense to the particle alone (to the "bare" particle) and the other, $\tilde{\Psi}$, to the surrounding medium, in which the particle moves.

Editor's note: You can contact the authors directly for more information (8-16 equations).

Equation (16) coincides in its appearance with the usual Dirac equation for charged particle in an external field described by 4-potential . However, in reality, it differs essentially from Dirac's equation. The distinction consists in that **equation** (16) **is non-linear and non-local**, with the non-locality being of both spatial and time character. Potential ($A_{||}$) and vortex ($A_{\perp}$) components of the 4-potential, entering equation (16), differ from each other by their physical nature: the former describes the Coulomb field and is expressed quadratically in terms of the wave function components of electron, and the latter describes transverse electromagnetic waves and is expressed in terms of vortex electromagnetic field. As a detailed analysis shows, solutions to the basic dynamical equation describe the clots of self-acting electrically charged matter, localized in space, i.e. the particle is a soliton.

The internal energy spectrum of electron is discrete with an indefinitely large number of levels, and to each value of internal energy E_k (k is the set of quantum numbers) there correspond certain linear dimensions and geometrical form of the region of localization of electron's charge. Dimensions and the number of extreme of wave function increase with increasing the value of energy E_k. The distribution of electric charge of atomic electron in the ground state consists of the range of basic localization with the linear dimensions of the order of Bohr radius a_0 ($a_0 \sim 10^{-10} m$) and of the tail stretching up to infinity. It is essential that because of non-linearity of the dynamical equation of electron, wave function does not obey the superposition principle. By virtue of this, electron acquires the properties of absolutely rigid body: the perturbation acting on electron at an instant of time in the range of basic localization becomes known at the next instant $t + 0$ at any distance from the particle.

In Fig. 1 the results of calculation are represented schematically, carried out on the basis of equation (13), of the distribution of electric charge in atomic and free electrons in the ground (a) and first excited (b) states.

According to [9,19], the atom represents a system of nuclear and electronic solitons interacting with each other, the internal energy spectrum of the hydrogen atom, due to electromagnetic interaction, being of a zoned character. The occurrence of zoned structure of energy spectrum of hydrogen atom is explained as follows. Free nucleus, because of existence of Coulomb self-action, has a discrete internal energy spectrum. As the interaction of nucleus with electron is small in comparison with the energy of Coulomb self-action of the nucleus, it can be taken into account by perturbation theory. From here it follows at once that each energy

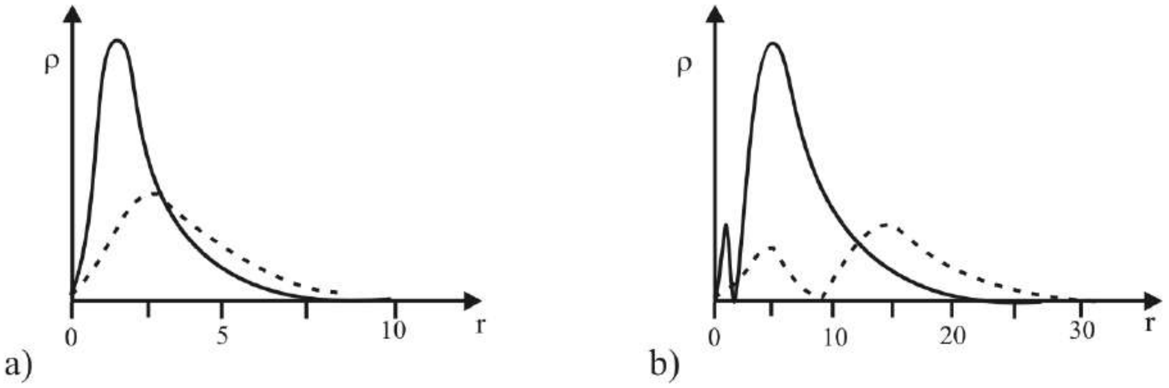

Fig. 1. Density of electric charge (ρ) of electron in the ground state (a) and in the first excited state (b): the continuous lines correspond to electron in the hydrogen atom, and the dotted ones to free electron, r is the distance from the center of mass of electron measured in Bohr radii.

level of free nucleus is split in a zone. There are indefinitely many zones (Balmer's replicas) and in each of them there are indefinitely many energy levels. The lowest zone coincides with the usual Balmer spectrum.

Physical mechanism of nuclear reactions at low energies

The quantum theory presented above schematically of electron as an open self-organizing system is indicative of the existence of the following mechanism of nuclear reactions at low energies [8].

If there occur in the region of basic localization of free electron, which linear sizes in the ground state of the particle are several times as large as those for hydrogen atom (see Fig. 1), two or the greater number of nuclei, each of them attracts on itself the adjoining areas of electronic cloud, resulting in compression of the electronic cloud as a whole. As a result, there appears automatically an attraction of the nuclei, which proved to be "inside" electron, on each other (see Fig. 2).

Calculation shows that the Coulomb barrier around nuclei is deformed, its height decreases and the probability of penetration through the barrier accordingly increases due to tunnel transition. Under certain conditions this process may result in fusion of nuclei. Obviously, the process in question can occur only at small energies of translational motion of the centers of mass of electron and nuclei: nuclei should be "inside" electron long enough for them to have time to come nearer to each other as a result of electron-nuclear interaction. This mechanism of nuclear fusion is of a universal character. In order for it to be realized, it is necessary to have only a stream of free electrons intensive enough, i.e. heavy electric current, and as long as sufficiently great number of free nuclei.

If heavy nuclei appear "inside" free electron, owing to their interaction with the electronic cloud there occurs polarization of nuclei. Because the own field of electron interacts with protons more strongly than with neutrons, nuclei are deformed (become extended), and this process may result in the decomposition of nuclei to fragments (in nuclear fission).

As is noted in [7], the official version of the reasons for Chernobyl accident contains serious contradictions, a number of facts concerning the accident has no convincing explanations, and this circumstance forces to search for the true reasons for the happening, since

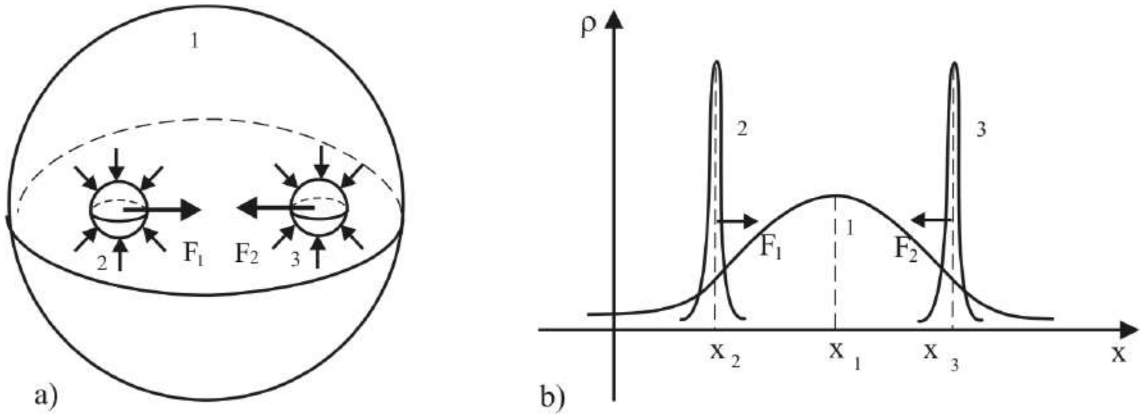

Fig. 2.
The schematic image of interaction of nuclei with electronic cloud: (a) 1 is the region of basic localization of electron, 2 and 3 are nuclei, F_1 and F_2 are the attractive forces between nuclei, which appear at the expense of electronic cloud compression induced by Coulomb forces; (b) ρ is the charge density, 1 is electronic soliton, 2 and 3 are nuclear solitons, X_n (n=1, 2,3) are coordinates of the centers of mass of particles.

"not having understood the mechanism of the one tragedy, we sooner or later shall become witnesses of the other". The authors hypothesize that the reason of the accident was penetration into the nuclear reactor of magnetic monopoles, which have caused the decay of nuclei ^{238}U, and this has resulted in production of delayed neutrons, growth of power output of the reactor and explosion. As an argument in favor of the assumption, the fact is presented that nucleus ^{238}U are disintegrated under the action of "strange" radiation appearing at explosion of foil.

In the opinion of the authors of [5,7], "strange" radiation is created by those magnetic monopoles, which form bound states with nuclei of atoms. These compound particles give the abnormally wide tracks similar to those of a creeping caterpillar, and also the tracks of complicated shape reminiscent of spirals and gratings. Character of tracks changes when imposing magnetic field, which, as the authors believe, is an argument in favor of the assumption above. There are also some special tracks very similar to scratches and ink spots. "Strange" radiation is of spherical form, it resembles a ball lightning, and its duration is more than ten times as great as that of the current pulse arising at electric discharge. With the course of time the luminous sphere (the ball-like plasma formation) is dividing into many small "balls".

It is our opinion that "strange" radiation is caused by free electrons in excited state arising in the area of electric discharge. According to [9, 19], linear sizes of the region of basic localization of such electrons can make many tens of sizes of atom. The heavy nucleus, for example, the nucleus ^{238}U, appearing inside the electronic cloud, is inevitably deformed because of interaction of protons with adjoining layers in the distribution of electric charge of electron, and this deformation can cause nuclear fission. If two or the greater number of light nuclei appears "inside" electron, then attractive forces arise between nuclei, which may result in fusion reaction. When electric discharge is strong enough, the areas of basic localization of some electrons can overlap, and if a nucleus lands in the area of overlap, because of Coulomb attraction of nucleus on the adjoining layers of electronic clouds, a bound state may be formed, of two electrons and the nucleus, characterized by the relative stability and significant spatial extension.

Obviously, if the concentration of free electrons is great enough, there may be formed some relatively stable bunch of plasma consisting of great number of free electrons and nuclei, which in virtue of chaotic movement of nuclei and because of the absence of preferred directions should have approximately spherical form. Let us note that atomic electrons, belonging to additional energy zones of atom (Balmer's replicas associated with nuclear self-action, see Section 2) can contribute to "strange" radiation.

As is seen from above, to account for the reasons for Chernobyl accident, there is no need to involve magnetic monopoles. The scenario of development of events during the accident, described in [7], seems to be quite plausible if only to understand by initiators of nuclear fission not hypothetical monopoles but free electrons, which powerful pulse might arise as a result of electric discharge in the region of turbo-generators.

The existence of simple physical mechanism of nuclear reactions at low energies, indicated in this paper, implies that nuclear reactors are, in effect, nuclear delayed-action bombs, which will blow up from time to time. Explosion of nuclear reactor may take place because of casual short circuit at an electric subcircuit, owing to which there appears an intensive stream of free electrons. This stream, having got for any reasons in nuclear reactor, may initiate explosion of the reactor. It follows from here that though nuclear stations may provide mankind with cheep energy, atomic energetics represents a very dangerous way of producing energy (as well as the energetics using controlled thermonuclear fusion). The only acceptable way of resolving the energetic problem consists in the use of nuclear reactions at low energies.

According to the results obtained, nuclear reactions at low temperatures occur "inside" electron under the action of own field of particle. Hence, to elucidate physical mechanism of CF, it is necessary to study in detail intra-electronic processes and physical properties of own fields of particles. Note that the own field, by its physical properties, essentially differs from the field of electromagnetic waves: this is the field of standing waves of matter, it is of purely classical character and may not be reduced to the set of photons. The own field of charged particle plays in nature a special role, consisting in that it transforms environmental space into the physical environment (physical vacuum) with the properties of absolutely rigid body [32].

As it was repeatedly noted in the literature [1,2], experiments on CF are badly reproduced, and this fact gives rise to doubt the very existence of the phenomenon. Bad reproducibility of results seems to be explained by the fact that CF depends upon great number of parameters: upon electric current density, concentration of free nucleus, concentration of impurities and dislocations in samples, sizes of samples etc. In order to obtain reproducibility of results, it is necessary that all these parameters, describing the environment in which nuclear reactions occur, be the same in various experiments, but to achieve this as a difficult task.

In conclusion we shall dwell upon the problem of linear dimensions of electron, which is of special interest in connection with the mechanism of nuclear reactions indicated here. The inference that the dimensions of electron in the ground state of atom are of the order of Bohr radius, i.e. of the order of atomic dimensions, following from dimension considerations [9,19] and confirmed by quantum model of electron, seems completely unexpected. At first sight, it is in conflict with both the theory of quarks and experimental data on scattering of electrons. According to quark models, the radius of electron corresponding to its quark structure makes up the quantity of the order of 10^{-22} m [33]. It is necessary to emphasize, however, that the above-mentioned magnitude of linear dimensions of electron refers to the internal structure induced by Coulomb field. The last is long-distance and consequently the linear dimensions of internal structures produced by it (i.e. spatial inhomogeneities in the distribution of electric charge in various quantum states) should considerably exceed the dimensions of quark structures connected with electron. There seems to exist a hierarchy of internal structures of particle produced by Coulomb forces, nuclear forces, inter-quark interactions etc. characterized by the smaller and smaller linear sizes.

As to the experiments on scattering of high energy electrons, according to which the internal structure of electron is not manifested up to distances of the order of $10^{-16} \div 10^{-17}$ m, two arguments, at least, can be adduced in favor of that there is no contradiction here with the experiment. Firstly, in experiments on scattering, investigators were trying to register the details of internal structure of electron within intervals much smaller than Bohr radius, which is why it is not surprising that results of experiments proved to be negative: at high energies electrons behave like point particles, their internal structure has no time to be manifested. Secondly, the results of experiments were analyzed from the point of view of standard representations about electron, which refer to a point particle, but are obviously inapplicable to real, self-acting electron. According to the predictions of quantum theory of electron as an open self-organizing system, real electron is a special object - soliton, i.e. such a cloud of electrically charged substance which, when interacting with other particles, tends to keep its sizes and geometrical form.

At present there is as yet no scattering theory of this kind of particles and for this reason it is impossible to predict with certainty how can the internal structure of electron be manifested in experiments on scattering.

References

1. Storms E. A *Critical Review of the "Cold Fusion" Effect.* J. Sci. Explor., **10**, #2, p.185 (1996). See also:
http://www.jse.com/storms/1.html
2. Storms E. *Cold Fusion Revisited.* Infinite Energy, 4, #21, p.16(1998).
3. Schwinger J. *Cold Fusion: A Hypothesis.* Z. Nat. Forsch. A 45, 756 (1990); *Cold Fusion: A Brief History of Mine.* Infinite Energy, 1, #1, p.10 (1995).
4. Schwinger J. *Nuclear Energy in an Atomic Lattice I.* Z. Phys. D15, 221 (1990); Prog. Theor. Phys. 85, 711 (1991); *Energy Transfer in Cold Fusion and Sonoluminescence.*
http://www.geocities.com/CapeCanaveral/Lab/4520/theory.htm
5. Urutskoev L. I., Liksonov V. I., Tsinoev V. G. *Experimental Detection of "Strange" Radiation and Transformation of Chemical Elements.* Applied Physics, M., 2000, p.83-100. Urutskoev L.I., Liksonov V.I., Tsinoev V.G. *Observation of Transformation of Chemical Elements During Electric Discharge*
http://arxiv.org/ftp/physics/papers/0101/0101089.pdf
6. Vysotskii, V. I., Kornilova A. A., and Samoylenko I. I. *Experimental Discovery of Phenomenon of Low-Energy Nuclear Transformation of Isotopes (Mn55=Fe57) in Growing Biological Cultures.* The Sixth International Conference on Cold Fusion, Progress in New Hydrogen Energy (Ed. M. Okamoto) Oct. 13-18, 1996, Hokkaido, Japan, Vol. 2, p. 687; Infinite Energy, 2, #10, p.63 (1996).
7. Urutskoev L.I., Geras'ko V.N. *On the Possible Mechanism of Chernobyl Accident*.
http://www.recom.hotmail.ru/CHER_INC_1.doc
8. Oleinik V.P. *To Electronic Technologies of the 21st Century: on the Threshold of Revolution in Communication Systems. Collection of Reports*, Millenium 2002, International Conference "To Innovations in the 21st Century", Odessa, April 13, 2002, p.268-273.
9. Oleinik V.P. *The Problem of Electron and Superluminal Signals*. (Contemporary Fundamental Physics) (Nova Science Publishers, Inc., Huntington, New York, 2001).
10. Berestetsky V.B., Lifshits E.M., Pitaevsky L.P. *Relativistic Quantum Theory*, part 1. (Moscow, Nauka, 1968).
11. Medvedev B.V. *Foundations of Theoretical Physics*. (Moscow, Nauka, 1977).
12. Dirac P.A.M. *Relativistic Wave Equation of Electron*. Progress in Phys. Sciences, 129, #4, p.681–691 (1979).
13. Dirac P.A.M. *The Principles of Quantum Mechanics.* (Moscow, Nauka, 1979).
14. Schrödinger E. *Quantisierung als Eigenwertproblem*. Vierte Mitteilung. Ann. der Physik, Bd. 81, S.109–139 (1926).
15. Schrodinger E. *Selected Works on Quantum Mechanics*, Edited by Polak L.S. (Moscow, Nauka, 1976).
16. Barut A.O. *Schrodinger's Interpretation of as a Continuous Charge Distribution*. Ann. der Physik, Bd. 45, S.31-36 (1988).
17. Barut A.O., van Huele J.F. *Quantum Electrodynamics Based on Self-Energy: Lamb Shift and Spontaneous Emission without Field Quantization*. Phys.Rev., A32, #6, p.3187–3195 (1985).
18. Barut A.O., Dowling J.P. *Quantum Electrodynamics Based on Self-Energy*: Spontaneous Emission in Cavities. Phys.Rev., A36, #2, p.649–654 (1987).
19. Arepjev Yu.D., Buts A.Yu., Oleinik V.P. *To the Problem of Internal Structure of Electrically Charged Particles. Spectra of Internal Energy and Charge Distribution for the Free Electron and Hydrogen Atom.* Preprint of the Inst. of Semiconductors of Ukraine, N8-91 (Kiev, 1991) 36 p.(in Russ.).
20. Oleinik V.P. *Quantum Electrodynamics Describing the Internal Structure of Electron*. Quantum Electronics. #44, p.51-59 (1993) (in Russ.).
21. Oleinik V.P. To *the Theory of the Internal Structure of Electron. Second Quantization and Energy Relations.* Quantum Electronics. #45, p.57-79 (1993) (in Russ.).
22. Oleinik V.P. *Quantum Theory of Self-Organizing Electrically Charged Particles. Soliton Model of Electron*. Proceedings of the NATO-ASI "Electron Theory and Quantum Electrodynamics. 100 Years Later." (Plenum Press, N.-Y., London, Washington, D.C., Boston, 1997), p.261-278.
23. Oleinik V.P. Nonlinear *Quantum Dynamical Equation for the Self-Acting Electron*. J. Nonlinear Math. Phys. 4, #1-2, p.180-189 (1997).
24. Oleinik V.P. *Quantum Equation for the Self-Organizing Electron.* Photon and Poincare Group (Nova Science Publishers, New York, Inc., 1999), p.188-200.
25. Nicolis G., Prigogine I. *Self-Organization in Non-Equilibrium Systems* (Wiley-Interscience, 1977).
26.Oleinik V.P., Belousov I.V. *The Problems of Quantum Electrodynamics of Vacuum, Dispersive Media and Intense Fields.* (Kishinev, Shtiintsa, 1983).
27. Bogoliubov N.N., Shirkov D.V. *Introduction to the Theory of Quantized Fields.* (Moscow, Nauka, 1976).
28. Bohr N. *Selected Scientific Works*, V.2. (Moscow, Nauka, 1971).
29. Bogoliubov N.N. *Quasi-Averages in the Problems of Statistical Mechanics.* In the Book: Statistical Physics and Quantum Field Theory. (Moscow, Nauka, 1973).
30. Morse P.M., Feshbach H. *Methods of Theoretical Physics*, V.1. (Moscow, Foreign Literature, 1958).
31. Dakker H. *Classical and Quantum Mechanics of the Damped Harmonic Oscillator.* Phys. Reports, 80, # 1, p.1-112 (1981).
32. Oleinik V.P. *Superluminal Signals, the Physical Properties of Time, and the Principle of Self-Organization.* Physics of Consciences and Life, Cosmology and Astrophysics, #1, p.68-76 (2001).
33. Dehmelt H. *Experiments with an Isolated Subatomic Particle at Rest*. Progress in Phys. Sciences, 160, #12, p.129–139 (1990).

The Evolution of Lifter Technology

Tim Ventura

tventura6@attbi.com

INTRODUCTION

Readers of the electric-spacecraft journal might know a little about the Lifter technology popularized recently be Jean-Louis Naudin, but they probably don't know the whole story. In the short amount of time that has transpired since the publication of that article, this technology has both literally and figuratively taken off – going from a "proof-of-concept" prototype by Naudin to an international group of researchers investigating how to give the lifter higher-performance and greater efficiency. With the first commercial products now on the horizon, if you haven't taken the time to read up on lifter technology, this is the perfect time to do so. . .
To give you a complete up-to-date overview of where this technology is, where it is going, and what I think it is capable of, let me start with the basics – an overview of how I became involved with Electrogravity research and what eventually led me to become involved with lifter technology.

MY BACKGROUND

I started college at 16 years old, back in 1992 – at the same time, I purchased a kit containing "hoverboard plans" from Hovertech, Inc. The moment that I received that $20 white-manilla envelope in October 1992 was the moment that I became involved with what has now been nearly 10 years of electrogravity research.

I worked with Bill Butler – the president and chief-scientist of Hovertech – on a variety of different antigravity, Electrogravity, and levitation ideas from approximately 1992 through 1996. While putting in my college time, I was also taking distinct advantage of the enormous college library at Western Washington University to read up on everything that might possibly relate to Electrogravity. I read books on standard electronics and physics theory alongside with books by the masters of this science, such as TT Brown and Nikola Tesla.

Bill and I played with several different ideas – many of them only peripherally related to Electrogravity. For instance, I published a manuscript initially in 1996 describing Tesla's theory on how to reliably produce Ball-Lightning using a standard Tesla coil – the information courtesy of WWU's excellent library. Bill also assisted me with obtaining video footage of a Searl-effect conference that he attended in Denver in the early 90's – this footage was an excellent overview of Searl's design and construction concepts for what he believes is the next major technological step in aviation and space travel.

Bill and I eventually found different paths, and in some ways drifted apart. Bill moved into Geomagnetic levitation research and started intense investigation on the patents of How Wachspress and the magnetic dipole levitator. I went to more traditional technologies – eventually becoming a UNIX system administrator for AT&T Wireless.

I hadn't heard from Bill Butler in about 6 months when he sent me a short email containing the words "hey, check this out" – and a link to Jean-Louis Naudin's "Lifter Experiments" home-page. I visited the site, watched all of the video clips, and then watched them again. This was the technology that I had been waiting for!

LIFTER TECHNOLOGY

I can say without a doubt that the lifter technology is completely revolutionary, but you might not realize how profoundly revolutionary it is until you've stopped to think about it for a bit. What is it about the lifter that makes it so unique, especially when so many inventions claim to produce more and better electromagnetic thrust? The answer is simple – the lifter works repeatedly.

Jean-Louis Naudin started a figurative bonfire when he decided to replicate a "proof-of-concept" experiment by a small Huntsville, AL aerospace contracting firm. The lifter initially came into being in the mind of Jeff Cameron – the chief scientist of Transdimensional Technologies – in the 1970's from experiments conducted with high-power military and research-grade lasers. A device in the lasers called a "pre-ionizer" was used to apply a high-voltage to the lasing-medium to facilitate better performance. Repeated operation of the pre-ionizer had a common side-effect of horribly twisting the wire and foil combination out of shape, which required a decent amount of work to repair.

Jeff Cameron realized that the torsional effect on the pre-ionizer was a side-effect of some unknown force acting on the pre-ionizer apparatus, and he began a long-term investigation into what was causing the apparatus to deform. His eventual results indicated that a force in the foil collector in the pre-ionizer was causing a net-thrust in the entire pre-ionizer apparatus that was making it twist and move on its mounts within the laser – the lifter came to him later as a three-dimensional device to demonstrate this force.

Naudin's genius became readily apparent not through a giant breakthrough in technology, but rather in a more subtle fashion – he replicated the lifter experiments of Transdimensional Technologies and published videos, articles, and complete construction plans on his website to allow others to do the same. In a manner similar to the open-source software movement, Naudin had taken an incredible scientific find that might have otherwise been overlooked and done and incredibly charitable and intelligent thing – he gave it away for others to play with. By following Naudin's instructions, inventors all over the globe began to slowly replicate the

Transdimensional Technologies experiments and thereby validate the proof of concept that Jeff Cameron had created to show that his "mystery force" was real after all. Naudin of course took advantage of these replications of the experiment by showcasing them on his own website – which in turn lends additional credibility to his research.

As far as technology goes, the lifter demonstrates that science and engineering have more than their share of humorous irony. For the years that I researched Electrogravity and antigravity claims, all of the devices that I had seen required something "magic" to make them work. For instance, Bob Lazar's UFO-claims could have been reverse-engineered except that they require 'element 115' to make them work – an element chemically related to Bismuth that is theorized to potentially have electrogravitic properties. I will come back to the possible electro-gravitational properties of Bismuth in a bit, as it turns out that this element may in fact provide some use for future lifter technology.

The Searl-effect disc is an even better example of the "magic" usually involved with building a working Electrogravity device. Searl's ideas seem valid enough, but although he supposedly demonstrated several working prototypes in the 1950's, he is currently pursuing millions of dollars in research funding in to replicate those experiments in a modern-day setting.

The irony involving lifter technology is that while inventors all over the world have been searching for the perfect electro-gravitational device for decades, the possible working proof of concept for many of these theories has been sitting in front of us the whole time – the lifter costs less than $10 in parts to build, and none of them are magic – in fact, for my experiments, all of them were at stores within 2 blocks of my house — balsa wood from the craft store, aluminum foil from the supermarket, 30-gauge magnet wire from the local Radio Shack, and an old computer monitor for the high-voltage power-supply.

LIFTER PHYSICS

Whether or not Jeff Cameron knew it at the time he constructed his lifter prototype, what he was actually building was a 3 dimensional representation of a drawing on a patent application by TT Brown in the 1950's. In the patent application, the drawing shows a positively charged wire suspended over a grounded foil body which was meant to demonstrate the most basic Biefeld-Brown effect generator. While Brown's drawing is a little different than Jeff's design, the resemblance is uncanny enough to indicate that both of these men had the same basic force in mind.

TT Brown's patent indicates that this Biefeld-Brown effect generator works due to a gradient electrostatic-field between the wire and the foil – in essence, these two elements compose a low-efficiency, high-voltage air-gap capacitor in which the difference in geometries between the two capacitive elements generates a net-directional force from the larger element towards the smaller element. Jeff Cameron seems to have a practical axiom that goes along with this scientific philosophy, which is that there must be both a leakage current and a capacitance between the wire and the foil in order for the lifter to function.

Conventional physics says that two capacitor elements of different sizes will not generate a net-directional force, so what gives? This is actually the thinking that convinced me to abandon my research into Biefeld-Brown effect technology in 1996 – physics says it doesn't work. What the books say will happen is that since the wire can only maintain a lower-capacitance than the foil, the overall capacitance between the two elements will be reduced to be equivalent to that on the smallest element (or plate) in the capacitor. This, of course, assumes a 2-element series-wired capacitor, such as the lifter.

I can give you the conventional physics answer to this small riddle by simply saying that the lifter uses a manifestation of ion-wind. This would state that the electrons crossing the air-gap cause a breeze that causes thrust – since the breeze would be traveling down from the wire to the foil, the thrust would be up, as demonstrated in testing. In the ion-wind explanation, the electrons are emitted from small-diameter of the positively charged wire in such great abundance that they move a significant airflow down to the foil where they are absorbed and transported electrically back to the HV power-supply's electrical ground.

Conventional physics would seem to have the theoretical answer to why the lifter causes lift, but in the experimental setting, which is what we now have an abundance of thanks to Jean-Louis Naudin, the conventional physics explanation doesn't suffice. Experimentally, there are several deviations from the ion-wind explanation that seem to invalidate it. For instance, if you completely contain the lifter in a plastic-enclosure, it will still generate lift – this would not be the case if a breeze was responsible for lifting the device. How could it be, if the breeze is limited to the inside of an enclosure which itself is levitating?

A more compelling proof that Biefeld-Brown is something other than ion-wind comes from Purdue University, where the lifter experiment was replicated inside a vacuum-enclosure with positive results. While ion-propulsion can work in space, it usually assumes that there is argon, krypton, or other noble gas to be used as the propellant – the vacuum enclosure showed that with no gas available for transport the lifter showed a moderate improvement in performance.

The vacuum enclosure tests are definitely compelling evidence that something else is going on other than ion-wind – at least compelling enough for NASA to file patent number 6,317,310 – "Apparatus and Method for Generating Thrust using a Two Dimensional, Asymmetrical Capacitor Module". The NASA patent description – which can be accessed from Naudin's lifter website – is as vague is it is compelling in that NASA is basically requesting a patent on any technology that generates force using two geometrically dissimilar capacitive plates. Disregarding the fact that this patent was issued nearly 50 years after TT Brown's patent using nearly identical descriptions and pictures, and also disregarding the fact that NASA also doesn't understand why the lifter generates thrust, it seems apparent the this phenomena is gaining credibility in engineering circles while physicists seemingly continue to deny that anything is going on.

THE EVOLUTION OF LIFTER TECHNOLOGY

Every good movie always has a sequel, and in technology, if at first a major government agency 'liberates' your idea, it may seem that a sequel is in order. In the case of the lifter, it would appear that the NASA patent would cover this technology to at least some degree – at least until someone overturns this patent under the prior-art rule – which means that the next generation has to be considerably more advanced to escape having the research and development be forfeit to the government.

The pursuit of more advanced versions of the lifter technology is currently underway by several independent inventors, as well as Transdimensional Technologies themselves. Most of the private research by inventors has delved into improving the current lifter design to produce a greater force output and utilize less power to do so. Because the lifter is so simplistic in design, many of these enhancements have been of a very basic nature.

Jean-Louis Naudin was the first independent inventor to do serious work with improving the technology behind the lifter – and even so, the majority of his work has utilized similar materials in more complex arrangements. Naudin has demonstrated dramatically increased lifting forces by building a "lifter inside a lifter" for demonstration purposes. Naudin has also done a great deal of work in taking breaking up the concept of the single triangular lifter into a parallel series of lifting cells – which means that these cells, working in parallel, can contributed to greater stability and higher force output than any single lifting element.

Saviour – an independent inventor working with Jean-Louis Naudin – has done some of the most interesting improvements on lifter design since those by Naudin himself. Saviour's concerns have not focused around the "bigger is better" philosophy that many inventors have stuck by – he has done several experiments to determine the radiation output, remote-controlled applications development, and materials analysis and improvement on the lifter that others have not had the time or expertise to conduct.

A recent experiment by Saviour demonstrates just how this gentleman's foresight is helping other experimenters – Saviour substituted nichrome heating wire for the common lightweight wire used for the emitter, and demonstrated that the lifting force greatly increased when a higher potential 12-volt charge was used to heat the emitter wire in conjunction with the standard high-voltage charge coming off it.

Transdimensional Technologies – the developers of the initial lifter design – are taking the approach to optimizing lifter performance to another level. They are currently not-so-secretly working on a 2nd generation lifter, which will consist of a 1-piece layered material to replace the current wire and foil design.

The layered material approach to the lifter is an idea that Jeff Cameron may or may not have had after some lengthy discussions with Travis Taylor – the man responsible for testing some anomalous materials known as "Art's Parts".

Art's Parts were some pieces of material sent by an unknown person to the Art Bell radio talk-show with a note stating that the they were pieces of UFO wreckage taken from the often-cited "Roswell crash" in 1947. Whether or not the pieces of material actually came from that crash is unknown, but Art Bell did the honorable thing by sending them to an acquaintance in US Army research named Travis Taylor for a professional scientific investigation.

Taylor, who apparently tested the materials after-hours in a world-class research lab to avoid potential classification by his superiors, used an electron-microscope to determine that the layered materials were actually pieces of metal – containing several hundred microscopically thin layers of magnesium and bismuth. Taylor also tested the layered-metal with a high-voltage apparatus, which seemed to indicate that when a voltage was applied to the material, the layered metal would move – and in some cases levitate.

Taylor reported his findings to Art Bell and sent video clips of his high-voltage experiments, which eventually made it back to a permanent home on the Art Bell radio show website. In addition, Taylor conveyed his belief that the only manner in which the pieces of metal could properly be produced was through an advanced form of electron-deposition technology, due (apparently) to an absence of oxygen-molecules between the different layers of metals. Additionally, the layers of metal were too thin to have been mechanically produced.

Jeff Cameron indicated that Transdimensional Technologies maintained some contact at one point in time with Travis Taylor, apparently as professional colleagues in the defense community in Huntsville, AL. I am not an expert on this relationship, other than to say that to the best of my knowledge these two individuals knew and contacted each other, and that this is how Jeff Cameron might have come up with the 2nd generation lifter idea.

ADVANCED LIFTER TECHNOLOGY

As an inventor, I couldn't care less whether or not the idea for the technology came from a crashed UFO. To be perfectly honest, I'm not what you would call a "believer" anyways, although I have often wondered about it. My point is not to attempt to lend any credibility to "Art's Parts", but rather to tie in the properties of the anomalous material's high-voltage movement with the underlying theory of lifter operation.

Even mentioning a UFO in a respected publication or article is the kiss of death in today's world – and I wouldn't do it if it wasn't an intricate part of the story. The other interesting thought is that the layered material is once again partially composed of Bismuth – which is thought to possibly have some of the same electro-gravitational properties as Bob Lazar's Area 51 "element 115". Is there a similarity, or merely a coincidence between a claim that hasn't gained credibility and a technology currently under development?

The lifter in its own right is essentially a layered material. One of those layers is the emitter wire, which is highly charged with about 30kV worth of electrons, another layer is the air-gap, which is approximately 3 cm in height, and the final layer is an electrically-grounded "skirt" of aluminum foil that surrounds the lifter. It is also reasonable to expect that there are only two possible forces at work in the lifter – one of which being a possible ion-wind effect moving down from the emitter to the foil, and the other being a possible Biefeld-Brown effect, moving up through the foil to the emitter.

There are a few shortcomings in the lifter as a design that might be overcome if we could transition the layered material from one containing an air-gap to one that does not. For instance, the lifter is currently a rather delicate object, in that having a wire under tension as the emitter makes construction difficult for future automated assembly. Additionally, because the air-gap requires struts to support the emitter wire, a trade off involving the weight versus the strength of the struts is additionally involved in any current implementation of lifter technology.

Some of the other changes that would be helpful to implement when transitioning lifter technology from one type of air-gap to another are changes in the materials used to increase the dielectric capacity. High-K dielectric materials may be used to increase the displacement of electrons in the material to enhance charge transport. And since increasing the dielectric potential of the layered materials also increases the breakdown resistance, it means that thinner materials can be used.

Designing a lifter without an air gap would accommodate lower voltage requirements between the foil and the emitter. The voltage would not have to create the large e-field gradient to create a leakage current across such a large void. Therefore the overall voltage across the device could be greatly reduced, without much cost in thrust. A lower operating voltage in turn means that a lower-output power-supply can be used for a given amount of current, which increases the overall efficiency.

Transdimensional Technologies recent research is utilizing the layered materials approach to eliminate the air-gap and substitute for it high-k dielectric materials that may allow higher overall performance. Although they have not yet released details about the exact composition or thickness of the materials that they are working with, they claim to currently have a 10% reduction in weight using a low-voltage current across the thickness of their newest device.

FUTURE LIFTER TECHNOLOGY

Thanks to the tremendous amount of research being done on lifter technology by Transdimensional Technologies and a loosely affiliated group of inventors around the world, the future of lifter technology seems very bright at this point.

Transdimensional hopes to release some breakthrough research to allow replication of their newest 2nd generation experiments in the very near future, and along with that stands the massive body of research and advancements being done by inventors and researchers such as Jean-Louis Naudin, Saviour, the Lifters-group, and myself.

My personal goals are to attempt to assist Transdimensional Technologies in popularizing this technology to increase awareness of it and help "spread the word" about what it is and how it can potentially help the world.

Imagine if instead of getting in your car and driving through the usual maze of thoroughfares and side streets you were able to simply type in your destination and have a flying vehicle take you there automatically. The lifter technology offers to potential to transform the current transportation market by offering point-to-point aerial transport without the need for roads or freeways.

Additionally, unlike the magnetic-levitation ("Maglev") technologies that are currently being promoted as the future of transportation, the lifter does not require a specially constructed and exorbitantly expensive track to operate – the greatly reduces the per-unit cost on the technology and opens the door for wider adoption by the general public for transportation solutions.

Other individuals are currently working to see if lifter technology may offer cost-effective methods of transport into space, which would reduce the cost greatly and allow a one-piece, reusable method of moving things into orbit.

LIFTER RESOURCES

All of the research involved with the lifter technology is available to the public on the internet. The list of resources below are some of the better and more common resources to obtain detailed lifter information.

American Antigravity
→ http://tventura.hypermart.net

→ The author's website that includes video clips, complete instructions, and other related lifter information.

Jean-Louis Naudin's "Lifter Experiments Website"
→http://jnaudin.free.fr
→ A very in-depth website containing video clips, complete instructions,

World-Wide Lifter Replications
→ http://jnaudin.free.fr/html/lftwrld.htm
→ An overview with photos and video from many of the independent inventors who have replicated the lifter experiments.

Transdimensional Technologies, Inc
→http://www.tdimension.com
→ The home page for Transdimensional Technologies, the developers of the lifter design.

Blaze Labs (Saviour's Research Website)
→http://bel.150m.com
→An excellent site on research into lifter enhancements, radiation testing, sealed devices, power supplies, and other topics relating to lifter technology.

Lifter Builders Group
→ http://groups.yahoo.com/group/Lifters
→ An email group for the exchange of research findings for those interested in building lifters or staying current on the state of the technology.

NASA Patent #6,317,310
→ The NASA patent regarding obtaining thrust from an asymmetrical two-dimensional capacitor, grant Nov 13, 2001.

Research on the Capacitance Converter of Environmental Heat to Electric Power

N.E. Zaev

143970, Moscow region, village Saltykovka,
Granitchnaya Str., 8 529-9664

Nickolay E. Zaev works on creation of the prototypes of converter energy, which do not require any fuel. The direct conversion of environmental heat to electric power is possible in the processes of "charge-discharge" in non-linear condensers or by means of "magnetization-demagnetization" of ferrites. Such converters of energy create cold and electric power without any fuel.

Theory of the converter, results of early experiments on the generation of microwatt power, methods and features of research are given in this article. The methods of generation of a few watts power are described in details. The possibilities and difficulties of creation of powerful capacitance converters are discussed in this article.

I. Grounds of research.

1.1. From positions of orthodox physics there is no subject of research. It is evident that the energy of charging (C) A_c condenser C_x is always equal or more than the energy of discharging (D) A_d, i. e. always $A_c \geq A_d$. Only the advanced analysis shows that it is not always true. Exactly, in C_x, where $\frac{\partial C}{\partial V} < 0$ an inequality $A_d > A_c$ is possible, and in C_x, where $\frac{\partial C}{\partial V} < 1$, then the work $A_c > A_d$. Therefore we should discuss the nonlinear capacitors (NC). In the end of 1969 I noticed a systematic inequality $A_d > A_c$ during the measurement of A_c and A_d of many capacitors with different dielectrics. Theoretical grounds and results of measurements of this phenomenon are given in the publications in 1984 [1], [2, page 73]. On the industrial standards NC (varicond), ceramic condensers VK2-ZSH, $4 \cdot 6{,}8 \cdot 10^{-9}\mu F$ with an optimal voltage about 95 V it was stated that $\frac{A_d}{A_c} \sim 1{,}21$ with the power to about $98 \cdot 10^{-6}$ Wt and "generated" extra power is equal to $21 \cdot 10^{-6}$ Wt.

1.2. In [1] and [2] the strict theoretical proofs of realization of $A_d > A_c$ (there are four of them) are given. On 1m^3 of dielectric $|A_d| - |A_c| = -\frac{1}{2} a \cdot \varepsilon_0 \cdot E_c^{\ 3}$ (E_c is an intensity of the field, V/m; ε_0 is a dielectric constant of vacuum, a is a coefficient of nonlinearity of the capacitor). Below we state one more proof more connected with the parameters of circuit.

It is well known that with the charge of a linear capacity from the source of constant voltage V_0=const through the resistor R=const it gets an energy $A_c = \frac{C \cdot V_0^{\ 2}}{2}$ exactly equal to the output energy in the time of charging t_c. The output energy irradiated from the load R is a Joule heat $\Theta = R \cdot \int_0^{t_c} i^2 \cdot dt$ [3, page 546]. If NC (nonlinear condenser) is charged, then there are no proofs of such equation. The NC are the variconds or other capacitors, which have $\frac{\partial C}{\partial V} > 0$ in the interval $V=0 \div V_k$. For the variconds V_k is some voltage, which corresponds to the maximum C_v. If $V > V_k$, then $\frac{\partial C}{\partial V} < 0$. For some other capacitors V_k is a voltage breakdown.

For further consideration let's believe that in the operating area of the given sample of varicond a function

Reality and Consciousness in Education and Activity

A.P. Smirnov

Vice president of International Club of Scientists
190031, Saint Petersburg, Kazanskaya str., 36
Tel: +7 (812) 312-0508
E-mail: science@shaping.org

Relation of thought to existence is the main question of philosophy as science on general laws of Nature was formulated but it still did not interpreted and solved in the frames of generally accepted logic standards. The ways to solve it lead to futile discussions of materialists and idealists, to senseless disputes of determinists with eclectics and apologists of the "chance". This discussion lost its sense without a determination of terms under discussion and condemned debaters to have subjective "gustatory" senses, which were changing while aging and depended on the extent of received and conceived knowledge. Such is the situation in this link of World studying, which does not allow creating a logic chain of reasoning in the understanding of cognizable things.

A paradoxicality of all things that happen is connected with incorrect translations and interpretation of wisdom of ancient philosophers and scornful attitude both to the knowledge of distant past and classical heritage, which highlighted the elements of natural-science approach to Weltanschauung.

According to Plato, **an ideal thing** is a visual thing, which can be felt by our organs of sense. Therefore, the understanding of objective reality is mediated by the crowd of our feelings in such a way that perception of reality by means of these feelings gives us a notion of the World. Hence, **our notions about reality are the subject of research in science, but not the World itself, i.e. the World outside of our consciousness.** So, what should be studied in our notions about the World? Let us refer to the wisdom of ancient scientists again: " The World is given in motion and its laws are the laws of motion". Then, we should speak about laws, order, i.e. about relation and interrelation in the phenomena of motion. This is the distinctness in notions and actions (determinism) to predetermine further development of reality cognition logic, i.e. what has an influence on us and determines specific character of our perception. Further we can speak about formation of ideas about reality, which require some premises, principles to organize these ideas. These principles are given in classical heritage, in "Dialogues" by G. Galilee [1] and "Mathematical principles of natural philosophy" by I. Newton [2]. A notion of force as a measure for momentum was introduced, which manifests in action and disappears from the body after the action is over, and the body keeps its new state due to the inborn "inertia force". **But the force itself cannot do anything without its application with a certain speed.** Then we introduce a notion of action as a product of acting force F_A and the speed of action V_A.

We offer a law of interaction, which determines the interaction between action of the cause and the effect appeared during this action as a reaction, i.e. the product of the force of reaction F_R and the speed of reaction V_R. Thus, this interaction between the cause and the effect is determined by the transfer of action from one object to another in equal quantity, but with appearance of new quality, which is determined by specificity of interacting objects according to fundamental law of interaction:

$$F_AV_A=-F_RV_R.$$

Unfortunately, an incorrect interpretation of interaction manifestation as an opposite counteraction became strong in our mind. This manifestation is perceived as a compensation of cause by action of the effect. Moreover, the incorrect way of writing of the mathematical form of Newton's third law manifestation established in textbooks and scientific literature due to the incorrect translation as $F_A=-F_C$. This very tragic situation for the science suppressed the development of logic in description of processes. Chance and statistic approach to the description of phenomena has taken place in our perception. This approach is based on the model of non-interacting elements, in which there is no order stipulated by the interrelation of elements. The science has developed this model and its properties, and this fact predetermined the evolution of notions about real World.

This ideology penetrated in mathematics, which for sake of physics began to study properties of objects, but not operations with them. Moreover, a possibility to reflect specific character of real physical processes in the interconnection of cause-effect relations by mathematical operations is not realized. It is essential, that fundamental law of interaction establishes manifestation and description of elementary act of cause and effect interrelation, **the law of manifestation of a Fact.** It means that order in the World is conceived through manifestation of concrete facts. The action of law of interaction lies in the basis of these facts.

So, there is a conclusion: the World is perceived through the discrete manifestation of motion forms evolution. Hence, the discrete mathematics of finite discrete aggregate can be applied to describe the World, but not the continual mathematics, which lies in the basis of traditional orthodox physics. All these circumstances lead to numerous problems and difficulties in description of our notions of reality, to the plenty of used principles, which are in contradiction to each other, as R. Feinmann noticed once [3].

And what we can get from determinism, which is based on fundamental law of interaction, law of cause and effect interrelation? **The change of force value in a reaction takes place, i.e. the change of value of the potential gradient, i.e. the change of energy**

concentration. This circumstance is visually demonstrated by the operation of Archimedean lever as well as in all phenomena of the real World. This is Archimedean lever, where the loss of speed takes place, but there is a gain in force. And the load raised on a lower height than the way, which was made by the applied force, will give a huge power during its free fall. This power is higher in so many times, in how many times the time of the load fall is less than the time of action spent on its raising! And this is the fact, which determines specific character of **creation in the real World**. We should attribute both quantitative and qualitative characteristics to energy. This is the side of energy manifestation, which is reflected in Plank's formula: energy is proportional to frequency.

Manifestation of fundamental law of interaction also lies in the basis of general universal regularity of evolution of real many-particle systems with the change in external conditions. This process develops in multistage way, and on the each stage the logarithm of the ratio between the event happened and the event to happen always is equal to the work of external forces. In other words, the relation of the event happened to the resource is in exponential dependence on the initial conditions and extent of external influence. Exponential character of development of processes is the evidence that Nature develops according to the law, which conserves itself during evolution. This regularity, which manifests everywhere, can be naturally called **the Principle of Order.**

Fundamental law of interactions and Principle of Order appeared to be enough to describe and understand phenomena in the observed World. And it is natural to expect that this principle of Nature manifest in finer World also. This World includes lower and higher frequencies, which are not available for us yet to watch this wide-range frequency-wave emanating Universe.

From all aforesaid we should make a conclusion that the logic, which exists in the traditional physical tool, appeals to the model and principles of the World of non-interacting elements using the range of regularities, which also reflect some features of the real World, but they do not include fundamental law of interaction and Principle of Order, which are necessary and sufficient to describe reality. Descriptions existing in traditional physics are phenomenological ones and concern only those aspects of the phenomena under investigation, which do not include possible qualitative changes during development of processes, because the main property of real processes of interactions (creation of new energy property) was excluded.

The current situation in physics had a strong influence on formation and development of other sciences, other fields of knowledge, since the logic of reflection of cause-effect links was initially excluded. These are the links to determine existence, i.e. existence of constant creation of the World. All these circumstances give grounds to fundamentally revise educational programs, first of all, in physics, philosophy, mathematics, chemistry and biology. A change to the offered logic of cognition, which is based on the Principle of Order and fundamental law of interaction, will fundamentally change our notions about the World as well as will open big opportunities for new technique and technology. A Man has got huge opportunities in cognition and existence, but due to his immorality and features of incorrect aims in the logic of cognition he cannot use these gifts of Nature. We present wider and deeper view on the World and a Man in it, which allow analyzing, watching and operating with those fields of reality, which manifest in finer World, World of higher-frequency energies and other structures of fields. Logic of cognition had not touched these structures yet.

References

1. Galilee Galileo. Selected works in 2 volumes. M.: Nauka, 1964.
2. Newton Isaac. Mathematical principles of natural philosophy. News of Nickolaevskaya Marine Academy, issue IV, V, Petrograd, 1915-1916. Volumes I, II and III, 620 p.
3. Feinmann Richard. Character of physical laws. M.: Nauka, 1967, 160 p.

Old New Energy

Y. I. Andreev, A.P. Smirnov

St.Petersburg, Russia
E-mail: science@shaping.org
Internet: http://www.physical-congress.spb.ru

Two kinds of energy, accumulated energy [1] and free energy [2], are considered as an inexhaustible source of natural energy created by Nature itself. It is ecologically clean and possible to be renewed in natural conditions.

The energy accumulated in substance is released as a result of partial decay of substance in elementary particles. At that, the acquired defect of mass is so small that it does not change chemical properties of substance and is compensated in natural conditions. Physical mechanism of energy-release lies in the fact that an electron in plasma layerwise takes sufficiently smaller elementary particles (electrino) from positively charged atoms or fragments of substance (ions). Electrino give their kinetic energy to plasma, heat it up and move beyond the bounds of reaction zone in the form of thermal and optical radiation. There is no substance, which could not take part in such process of energy-release, i.e. phase transfer of higher form (PTHF). The most appropriate, available and low-cost substances are air and water, which play the role of nuclear fuel in PTHF. It is turned out that usual combustion is also a process of PTHF, in which oxygen is a nuclear fuel and organic fuel is a donor of electrons. In the process of combustion oxygen atoms get the defect of mass equal

concentration. This circumstance is visually demonstrated by the operation of Archimedean lever as well as in all phenomena of the real World. This is Archimedean lever, where the loss of speed takes place, but there is a gain in force. And the load raised on a lower height than the way, which was made by the applied force, will give a huge power during its free fall. This power is higher in so many times, in how many times the time of the load fall is less than the time of action spent on its raising! And this is the fact, which determines specific character of **creation in the real World**. We should attribute both quantitative and qualitative characteristics to energy. This is the side of energy manifestation, which is reflected in Plank's formula: energy is proportional to frequency.

Manifestation of fundamental law of interaction also lies in the basis of general universal regularity of evolution of real many-particle systems with the change in external conditions. This process develops in multistage way, and on the each stage the logarithm of the ratio between the event happened and the event to happen always is equal to the work of external forces. In other words, the relation of the event happened to the resource is in exponential dependence on the initial conditions and extent of external influence. Exponential character of development of processes is the evidence that Nature develops according to the law, which conserves itself during evolution. This regularity, which manifests everywhere, can be naturally called **the Principle of Order.**

Fundamental law of interactions and Principle of Order appeared to be enough to describe and understand phenomena in the observed World. And it is natural to expect that this principle of Nature manifest in finer World also. This World includes lower and higher frequencies, which are not available for us yet to watch this wide-range frequency-wave emanating Universe.

From all aforesaid we should make a conclusion that the logic, which exists in the traditional physical tool, appeals to the model and principles of the World of non-interacting elements using the range of regularities, which also reflect some features of the real World, but they do not include fundamental law of interaction and Principle of Order, which are necessary and sufficient to describe reality. Descriptions existing in traditional physics are phenomenological ones and concern only those aspects of the phenomena under investigation, which do not include possible qualitative changes during development of processes, because the main property of real processes of interactions (creation of new energy property) was excluded.

The current situation in physics had a strong influence on formation and development of other sciences, other fields of knowledge, since the logic of reflection of cause-effect links was initially excluded. These are the links to determine existence, i.e. existence of constant creation of the World. All these circumstances give grounds to fundamentally revise educational programs, first of all, in physics, philosophy, mathematics, chemistry and biology. A change to the offered logic of cognition, which is based on the Principle of Order and fundamental law of interaction, will fundamentally change our notions about the World as well as will open big opportunities for new technique and technology. A Man has got huge opportunities in cognition and existence, but due to his immorality and features of incorrect aims in the logic of cognition he cannot use these gifts of Nature. We present wider and deeper view on the World and a Man in it, which allow analyzing, watching and operating with those fields of reality, which manifest in finer World, World of higher-frequency energies and other structures of fields. Logic of cognition had not touched these structures yet.

References

1. Galilee Galileo. Selected works in 2 volumes. M.: Nauka, 1964.
2. Newton Isaac. Mathematical principles of natural philosophy. News of Nickolaevskaya Marine Academy, issue IV, V, Petrograd, 1915-1916. Volumes I, II and III, 620 p.
3. Feinmann Richard. Character of physical laws. M.: Nauka, 1967, 160 p.

Old New Energy

Y. I. Andreev, A.P. Smirnov

St.Petersburg, Russia
E-mail: science@shaping.org
Internet: http://www.physical-congress.spb.ru

Two kinds of energy, accumulated energy [1] and free energy [2], are considered as an inexhaustible source of natural energy created by Nature itself. It is ecologically clean and possible to be renewed in natural conditions.

The energy accumulated in substance is released as a result of partial decay of substance in elementary particles. At that, the acquired defect of mass is so small that it does not change chemical properties of substance and is compensated in natural conditions. Physical mechanism of energy-release lies in the fact that an electron in plasma layerwise takes sufficiently smaller elementary particles (electrino) from positively charged atoms or fragments of substance (ions). Electrino give their kinetic energy to plasma, heat it up and move beyond the bounds of reaction zone in the form of thermal and optical radiation. There is no substance, which could not take part in such process of energy-release, i.e. phase transfer of higher form (PTHF). The most appropriate, available and low-cost substances are air and water, which play the role of nuclear fuel in PTHF. It is turned out that usual combustion is also a process of PTHF, in which oxygen is a nuclear fuel and organic fuel is a donor of electrons. In the process of combustion oxygen atoms get the defect of mass equal to 10^{-6} %, which constitutes the so small value that it cannot change chemical properties of oxygen and does not call killing radioactive emanation.

There is a possibility to use energy properties both of oxygen and nitrogen of free air in the process of PTHF. To do this it is necessary to destroy nitrogen molecule at least in atoms or smaller fragments by some initiating influence. It is achieved by electrical discharge, magnetic flow, explosion and other means. These means consume much less energy than produced in PTHF. In particular, such processes were achieved in combustion engines. Such nitrogen mode of operation and combustion is accompanied by oxidation to H_2O, but not to CO_2, which is more effective in energy and ecological aspects. Accordingly, the power of engine increases and organic fuel is saved. Exhausts from this process mainly contain water vapor [3].

PTHF processes with excessive power release (more than consumed power) were also obtained in heat-generators operating with water.

Free energy diffused in the surrounding space could be transformed into mechanical, electrical or another kind of energy by means of vibration-resonance, electromagnetic and other energy systems. Classification of these systems as well as physical mechanism of energy transformation is given in [2]. The known Searl's engines can serve as an example of energy systems working with free energy.

The developed physical mechanisms of energy-release processes will allow to create industrial, stably operating, ecologically clean energy systems, which do not consume organic and nuclear kinds of fuel, harmful for humankind.

References

1. Andreev Ye.I., Smirnov A.P., Davydenko R.A., Klucherev O.A. Natural enegetics. – SPB, Nestor, 2000, 126 p.
2. Andreev Ye.I., Andreev S.Ye., Glazyrin Ye.S. Natural energetics –2. – SPb, Nevskaya Zhemchuzhina, 2002, 104 p.
3. Patent 2179649, Russia, 2000 / Andreev Ye.I., Smirnov A.P., Davydenko R.A.

On General Nature of Forces

Dr. Evgueni D. Sorokodoum

Entrepreneur and General Director, Vortex Oscillation
Technology Ltd, Volochaevskaya Street, 40-B, Flat 38,
109033, Moscow, Russia
Telephone: 7-095-362-8084
e40dum@dol.ru http://vortex.viptop.ru

Introduction

In techniques and in our life we got used to certain physical notions concerning force. We usually use these notions in creation of automobiles, airplanes, rockets and other techniques, but we don't think about **the origin of forces** in general. Usually appearance of force in continuum is connected with presence of momentum gradient.

A number of works, which describe various versions about origin of a force appeared [1, 2, 5, 8, 17, 21, 22, 23, 25, 30, 35, 36, 38, 39]. Different mechanisms of appearance of force are considered in these articles. Usually they consider origin of a force in one of the fields, which are: electromagnetic, gravitational and others. **There was a theoretical attempt to connect the force initiation with energy gradient [33]. Experimental proof of force initiation due to energy gradient was obtained in the works [7, 38].**

Below we made an attempt to show the general regularity of force initiation, which is connected with non-uniform distribution of energy in space. With this process, physical nature of any kind of energy and specific mechanism of force initiation does not play any role. These are only particular cases of general nature of force initiation.

General nature of forces

We are surrounded by space, which is full of energy. Here we mean the energy of any nature: mechanical, thermal, electromagnetic and others. Energy is related with material world and its value is connected with the volume.

Any particle (volume) of continuum has energy:

$$A = A(x, y, z, t) \qquad (1)$$

where *x, y, z* are Eighler's coordinates of the center of particle, *t* is time.

Transmission of energy from one point of space to another one can take place by various methods, both in connection with energy transmission by material particle itself (which is a "carrier" of energy in this case) and without such transmission (for example, with wave motion). For the volume degenerated in ideal point the energy will be zero. That's why it is more comfortable to operate with the energy density concluded in the increasing constantly. We are the first who analytically got the law of gravity of the masses from the known equation of thermal conductivity. Appeared that on the relatively small distances (in the bounds of the Sun System) the law of gravity by Newton remains valid, but on the larger distances the sudden decrease goes on (Gauss integral), which naturally solves the famous Zelinger's paradox of gravity.

As a conclusion we should note that in the bounds of a stable galaxy of a spiral kind there is the circulation of ether. Ether moves from the periphery of the galaxy to its center (nucleus) by two spiral branches. This becomes apparent as a weak magnetic field (8-10 micro Gauss). In the nucleus of the galaxy there is the impact of two strings as well as there is formation of the spiral toroidal circles (protons). Then the protons form the adjoined vortexes around themselves (electron shells) and from the proton- hydrogen gas the stars are forming, which are moving to the periphery by the same branches. There they dissolve in ether at the periphery since the protons will loose their energy and stability due to the viscosity. Ether which have got the freedom will return to the nucleus of the galaxy and this process is going on in our galaxy for hundreds milliard years and it will keep going until the new center of vortex formation will begin to concentrate ether. Then the new galaxy will appear and our galaxy will disappear. But it will not happen soon and we have enough time to understand that we should return to the concept of ether in modern science.

Experimental Demonstration
of Cosmic Influence on the Earth Life in N.A. Kozyrev's Researches
("On the Influence of Time on Matter")

Alexandra L. Belyaeva

Bishkek, 720075, Russia 8th Location, 46, apt.80
Tel.: 7-996-31-41-25-79
E-mail: artem_belyaev@mail.ru
Internet: www.newphysics.h1.ru

Editor's note: this article represents a part of the big scientific conception "World models in the new scientific progress". On applying of this conception a great number of practical technical devices have been created (as an example of such device we offer the description of universal electrical bio-heater, which was created by the group of researchers from Bishkek, Kyrgyz Science Technical Center "Energy" during the *work on ceramic electroconvector).*

We have to note that the position of our editorial board concerning "time" and Kozyrev's work is not in a good correlation with the authors' one.

Nicolay Alexandrovich Kozyrev scientifically and experimentally discovered the action of relations' interconnection, which was falsely named as time. Time cannot cause action because it is absolute and does not have any physical sense (Samat Kadyrov. Monograph "Theory of unified field").

Author's note: relations' interconnection is an interaction of structurally similar objects. It is a nuclear resonant gain-frequency process: in a stationary electric field, which is modeled by systematic organization, there is a development of similar to structural one, in-focus rays of powerful regular coherent radiations. These coherent radiations are determined by properties of chemical components of interrelated substances.

According to N.A. Kozyrev, it is ought to expect not identical density of relations' interconnection in space. Some processes decrease density; others on the contrary increase density of relations' interconnection. Action of the increased density is weakened according to the law of reversed squared distances; it is shielded by a solid matter, at thickness about 5cm, and is reflected by a mirror, according to the familiar optics law. The action of the decreased density on a detector is shielded, but does not reflected by a mirror. Properties of a matter can be changed under the influence of relations' interconnection. In this sense there is a big advantage in changes of electric current conductivity of resistor, which is brought into Witson bridge and is located near some process. For instance, in order to increase density it is useful to realize the process of evaporation of a volatile liquid; and for density decrease the process of cooling of a warmed-up agent can be realized. Due to these processes, change of conductor resistance is actually realized with opposite signs. Increase of density of the conductor with positive temperature coefficient leads to decrease of its resistance. At negative temperature coefficient there is an effect of the opposite sign, in the direction of changes, caused by temperature changes. Such correspondence to fall in temperature should be observed at changes of other properties of a matter, **because disorder in a matter structure is reduced along with fall in temperature**. The researches have shown the following results at the resistor, which was situated near processes of acetone evaporation on cotton wool and of solution of sugar in water. The relative resistance change of resistor was observed at the 6th or 5th digit after comma (or even at the 4th digit if resistors had especially high temperature coefficient).

There is now a possibility to study the Universal World not only by means of the investigated spectrum of electromagnetic oscillations, but also through physical properties of relations' interconnection.

At many researches the influence of relations' interconnection on resistor electroconductivity was investigated. Acetone evaporation (at 10-15 cm distance from the resistor) was applied there as the process, which controls sensitivity of a system. However, the process of evaporation can influence on the resistor not only with density increase, but also due to temperature increase that occurs at evaporation. In order to take into consideration this cooling effect, (in the area of evaporating acetone) temperature was measured by Beckman mercurial thermometer with 0.01°C multiplying factor. The first experiments (without thermal protection) have shown the fall in temperature by several hundredth of degree. This fall was enough to cause the changes of resistor electroconductivity. However, the thermometer had been keeping on the demonstration of practically the same fall in temperature at thermal insulation of the resistor. The thermometer reacted on the radiation of relations' interconnection at acetone evaporation.

The part of the thermometer with a placed in a pasteboard tube mercury tank was laid round with cotton wool and put into a glass retort. The experimental process was fulfilled near the retort, and the reading of mercury altitude in capillary was determined by the scale of the thermometer through the closed window in the next room. The mercury altitude was decreased at dissolution of sugar in water (with steady temperature) and it was increased at the release of the squeezed spring, which was placed near the thermometer.

The radiation of the relations' interconnection was observed from many stars. It is caused by the inner processes, which take place on these heavenly bodies. The Sun (with its turbulent processes) radiates the relations' interconnection besides the searched electromagnetic radiation. Actually, if sunlight is recovered with a thin screen, the significant influence on the resistor will be discovered. The influences of the Sun to the Earth through the relations' interconnection become doubtless. These influences of the Sun should have a particular significance in vital functions of organisms, because it brings the beginning for life support. The totality of the researches demonstrates the dependence of matter state from the changes of the general background of the relations' interconnection. The drift of the devices (that show daily changes) usually stops about at midnight and then changes its direction. As for the seasonal course, there is a density decrease of the relations' interconnection in spring and summer; and there is an increase of it in autumn and winter. It is connected with the absorption of the relations' interconnection by the vital functions of plants and with the return of it at their fading. There are indications at the seasonal changes of chemical processes. For instance, reaction of polymerization has more difficulties in its realization in springtime. V. Zhvirilis observations of minimum and maximum light admission by means of the crossed Nickolya prisms can be explained by the crystalline reconstruction of these prisms.

By Kozyrev, as being invisible, vital source is disseminated everywhere in Nature, thus possibility of its accumulation is the only necessary thing. Such a possibility is realized in vital organisms because all vital functions counteract to the usual course of systems' destruction. The ability of organisms to keep and accumulate this counteraction is the reason, which determines the great role of biosphere for the Earth life. But even if we assume, that spreading of life in Space is one of its peculiar properties, biosphere will not have a decisive significance.

Cosmic bodies (and first of all stars) can serve as the reservoir, which gathers vital source. Enormous stocks of energy flow out of stars in a very weak degree through the radiation of comparatively cold external layers. Inner stars energy is preserved so well, that even at the lack of supplement, matter of the Sun would become cold only at one third degree per year. For the Universe the creative source carries the relations' interconnection. Thus cosmic bodies are necessary for support of life.

Author's note: We apprehend relation's interconnection as natural radioactive background. In fact, it is a nuclear resonance gain-frequency interaction of inertial masses that depends on living systems, especially on its rituals and that regulates its survival. Cosmic bodies regulate this process. Humanity is able to control nature only obeying to natural laws. In-focus beams of powerful laser streams are formed in the electric field of living system organisms. The creation of proton-antiproton pair in the living cells, alongside with the process of the absolute release of energy serves as a creative vital force. The process of radiation, support, absorption of energy by the organization (assembly of particles) is realized through the relation's interconnection and regulates its total mass.

Humanity is able to control nature only obeying to natural laws.

Life without Diseases and Old Aging Preventive Electrical Heater with Programmed Features

Alexandra L. Belyaeva

Bishkek, 720075, Russia 8th Location, 46, apt.80
Tel.: 7-996-31-41-25-79
E-mail: artem_belyaev@mail.ru
Internet: www.newphysics.h1.ru

Universal electrical bio-heater is intended for heating of rooms and preventive clearing of an air atmosphere from disease-producing organisms at continuous exposition (continuous work). The principle of its work fundamentally differs from those of the existing analogues. Carbon crystals are in the basis of bio-heater, which makes it environmentally appropriate.

Bio-heater represents a range of ceramic cylinders, jointed with metal plates on top and underneath. These plates play the role of load-carrying structure. It is used in production areas and living rooms for heating alongside with destruction of pathogen microorganisms. One bio-heater with 0,2 kWtt power is oriented for heating of the area with volume 35-45m^3 (in the future production of modernized models powered from solar cells is planned).

As distinct from the usual oil heater, preventive electrical bio-heater destructs agents of infectious diseases, whereas, according to the researches, oil heater stimulates their reproduction.

Absolute ecological cleanness is obtained by release of the quarters from the effect of increased atmospheric dampness with the temperature, appropriate to sanitary code. Any type of mold or fungus disappears in the quarter and in the future these forms do not renew their existence (even after removal of bio-heater).

The absence of injurious radiations is attained by the following: features of raw material, which is used during the process of electrical bio-heater production; radiation is normal during bio-heater working. Pollution-free temperature influence is attained by favorable infrared radiation.

Among the other properties of electrical bio-heater there are following: fire-safety; explosion proof; chemical inertness; enormous effectiveness from the point of view of electric energy demand. Structural simplicity facilitates its durability; there is nothing in bio-heater to be broken.

Technical aspects (applied Know How): In the process of technologic production of ceramic cylinders, from which bio-heater is consisted, **diamond-like cellular ceramic structure** with superimposed combination of atoms of lattice elements is created. Rhythmic work of cells, which form ceramic mixture, leads to resonance and creates a kind of blow wave (at micro level). This blow wave physically destroys microorganisms that have no calcium framework. It is related only to those microorganisms that are agents of infectious diseases, such as: staphylococcus, enterococcus, enterobacterium, etc. Thus parameters of the evoked blow wave coincide with vibration frequency of the definite types of bacterium and elementals. These blow waves cause the similar effect in room near of bio-heater, e.g. colonies of microorganisms are noticeably decreased there (even at the absence of bio-heater in the nearest room).

Due to its self-organization, bio-heater works in the range of living systems, it is approached to them. There is a realization of active connection with living coaly forms of biological systems. Actually the work of bio-heater is adjusted to them. Bio-heater properties can be programmed at the process of its production.

Bio-heater is a patented product. Patent KR #464 MKI C 04 V 33/24 “Ceramic mixture, possessing heat-radiating properties”. Application #20010075.1 at Patent KR #464 MKI C 04 V 33/24 “The way of creation of energy, renewable, programmed hard-phase ceramic-carbon mass structure”. Application at Patent KR #464 MKI C 04 V 33/24 “Technology of producing of electrical heaters with anti-resonant air prophylactic effect”.

Finale product (FP) purchase is not more expensive than those of existent models of electrical heaters. Cost value is noticeably brought down on organization of the scaled production. It is ought to take into consideration that from all existent types of heating, from the customer’s point of view, this one is the most energy-efficient. Manufacturing of such bio-heaters can be organized on the base of acting industrial production of ceramic fabrics. It will require some expenses. Moreover production service is rather cheap because there is no need in maintenance staff.

Electrical bio-heater can be applied everywhere, where there is a need in: a) economical heating; b) decrease of air moisture; c) disinfections of rooms.

As for the life cycle of bio-heater it does not become obsolete morally and technically. It is produced from the materials, which are not liable to wear.

The invention has a certificate of KR Gosstandart. From the end of 1998 the first unimproved modification of bio-heater (with power 0,6 KWtt) were put into serial production in Bishkek (with small test production runs). This time bio-heaters are readily used as medical equipment in hospitals and maternity hospitals in Bishkek. **Inventor: Alexandra L. Belyaeva.**

Technical Report

The comparison of quantity of heat energy, required for heat of rooms, and of heat quantity, which is produced by Belyaeva's electroconvector.

Mavlyanbekov Sh.Yu.

Deputy Director KSTC "Energy"
Internet: www.newphysics.h1.ru

Editor's note: this calculatious demonstrates the advantages of the device, which at 340 Wtt energy consumption produces about 700 Wtt of heat power.

The calculation of heat output, coming from the ceramic electroconvector to a room, was based on the basis of research statement of EVNA-0.2/220 electroconvector's influence on air micro flora of industrial rooms at 23.10.01. - 06.11.01. period.

The researchers were carried out in the arbitrary room in a four-storied large-panel building. This room was on the 3rd floor, with facing east windows. The room was of 52.5 m^3 air-space, 3.5 m height and 15 m^2 area. The calculation of heat, was made on the basis of "Methods for calculation of the requirement in heat and electric energy of buildings". These methods were registered by Department of Justice of Kyrgyz Republic on 08.09.2000, #154. According to the normative data, temperature of inner air (t_{in}) in the room must be equal to +20° C. In Bishkek planned specified temperature of external air (t_{ex}) for heating is minus 23°C. The average temperature of heating period is t_{av}= -0.9°C, specific heat characteristic of the building is: q=0.4 Kcal/m^3 h °C.

Medium quantity of heat energy, which is required for heating, is determined by the formula:

$$Q_{av}heating = q \cdot V \cdot (t_{in} - t_{ex}) \cdot 1.12 \cdot \\ \cdot [(t_{in} - t_{av}) \div (t_{in} - t_{av})] \text{Kcal/h}$$

$$Q_{av}heating = 0.4 \cdot 52.5 \cdot (20+23) \cdot 1.12 \cdot \\ \cdot [(20+0.9) \div (20+23)] = 492 \text{Kcal/h}$$

Thus at the average annual temperature of the heating period, which is: t_{av}= - 0.9°C, the quantity of heat energy required for this room, comes to 492 Kcal/h.

According to the research statement, the trials of the electroconvector with 200Wtt power were carried out at the following external air temperature: +10.2°C; +8.5°C; +10°C; +6.6°C. The calculation data and results of its examination are brought together in a table. The parameters of electroconvector with 340Wtt power are demonstrated in the same table.

Table 1
Table of determination of heat entry and heat consumption's correspondence in the experimental room

#	External air temperature, °C	Inner air temperature °C	Normative heat consumption KWtt (Gcal/h)	Experimentalheat consumption KWtt (Gcal/h)	Percentage depending on normative heat consumption %	Economy of heat energy kWtt (Gcal/h) comparing with power consumption of the device 0.2 kWtt (0.000172 Gcal/h)	Economy of heat energy kWtt (Gcal/h) comparing with power consumption of the device 0.34 kWtt (0.000292 Gcal/h)
1	- 0.9	+20	0.572 (0.000492)		100	-	
2	+10.2	+20	0.267 (0.00023)		100	-	
	+10.2	+16		0.158 (0.000136)	59	0.2-0.158=0.042	0.34-0.158= =0.182
3	+8.5	+20	0.314 (0.00027)		100		
	+8.5	+17		0.233 (0.000200)	74	0.2-0.233=-0.033	0.34-0.233= =0.107

4	+10	+20	0.273 (0.000235)		100		
	+10	+23		0.355 (0.000305)	130	0.2-0.355=-0.155	0.34-0.355= =-0.015
5	+6.6	+20	0.366 (0.000315)		100		
	+6.6	+19		0.338 (0.000291)	92	0.2-0.338=-0.138	0.34-0.338= =0.002

Calculation data demonstrate a considerable economy of heat energy at daily unevenness of external air temperature.

Heat productivity of the new structure of electric convector with 340Wtt power was calculated on the assumption on the suggestion that heating of the room is carried out by the irradiation at the process of heat exchange.

$$E = \varepsilon \cdot C_o \cdot T^4 \cdot 10^{-8} \qquad \text{Wtt/m}^2$$

where: C_0 =5.67 Wtt/m^2 K^4 is a radiant emittance of blackbody, ε=0.93 is an emissitivity factor of the surface of earthenware duct tube; T=70°C=343 K is the temperature of the surface of earthenware duct tube.

On substitution of the known values into the formula we get:

$$E = 0.93 \cdot 5.67 \cdot 343^4 \cdot 10^{-8} = 727 \ \text{Wtt/m}^2$$

As the area of irradiation surface is equal to S=0.96 m^2, then quantity of heat, which is evolved by the convector, comes to:

$$E_k = S \cdot E = 0.96 \cdot 727 = 698 \ \text{Wtt (or 600 Kcal/h)}$$

The quantity of heat, which is required for the heating of the room, is 492 Kcal/h (at the external air temperature equal to minus 0.90 and temperature in the room equal to plus 200).

Thus, electric convector with 340 Wtt power is able to heat totally the room with 60m3 area.

Editors note: 340 input and 700 output!!!

Longitudinal Waves in Vacuum: Creation and Research

Ph. Dr. Kirill P. Butusov

190121, Saint Petersburg, Angliysky prospect, 5-18
Tel: (812) 113-8511

The author presents a new elegant system, which is the symmetrized Maxwell's equations. In practice it gives a possibility to create the longitudinal waves in vacuum. This system is of great importance in telecommunications and aerospace technigue.

There is a stable paradigm in electrodynamics that the existence of the longitudinal waves in vacuum is impossible. This paradigm played its negative role preventing scientific minds from solving this problem. However, Maxwell was not as categorical in his opinion on this question as his following were.

Particularly he wrote: "Science of electromagnetism as well as optics is not able to confirm or deny the existence of longitudinal oscillations."

Maxwell's dynamic equations are usually considered as partial derivatives in time. However, the total derivative in time includes the so called substantial derivative, which was shown in the equations for the moving coordinate system. In particular, one of these equations was written by Maxwell himself to explain the phenomenon of electromagnetic induction discovered by Faraday. This induction takes place in the conductor moving across the field lines of electromagnetic field:

$$\vec{E} = \vec{V} \times \vec{B}\,; \qquad \text{(I)}$$

Other equations were obtained later by other scientists. In the table I below Maxwell's equations are given in a split form. Their static and dynamic parts are given separately as well as the equations for moving and fixed coordinate systems. Such matrix concept of Maxwell's equations allowed finding their incompleteness. Really, the analysis of the matrix shows its high symmetry. However, full symmetry of the system of equations is broken by the absence of the equation (X). It seems to be strange and calls a desire to remove this defect in such an elegant system of equations.

A new equation is introduced in the Table 1 for the full symmetry of the matrix:

$$\nabla \cdot \rho = -\frac{1}{c^2} \cdot \frac{\partial \vec{j}}{\partial t}\,; \qquad \text{(X)}$$

Fundamental Properties of Aether

Alexander M. Mishin

Author's note: In the article the principles determining major properties of aether are formulated on the basis of an empirical material.

Real aether [1-6], the primary and superfine essence of which is still a secret, has turned out to be absolutely non-standard superfluid three-dimensional material medium, which simultaneously is at solid, liquid and gas phases. The first master phase of aether is a specifically solid absolute space or an energetical "bottom" of the Universe ("celestial stronghold"). At that the solid phase is considered as mesomorphic vortical-wave structure, which has particular holographic properties. Classical matter represents to be one of the stable and energetic space-time levels of the Universe. Aether vortexes exceed all conceivable space scales, have quasi-material properties and create a great number of stereo-dynamic subspaces (parallel worlds).

The first basic principle, to which aether entirely follows, is the principle of the least disturbance (the least action). Many well-known and unknown physics laws are the subsequent of this principle. In particular, any motion in macroscopic aether happens in such a way to minimize the interaction with the matter of our world, with zero moment of the disturbance momentum. In the classical physics this principle has been reflected as Le Shatelye principle, as variation principle, laws of thermodynamics etc.

The second principle is the principle of fractality, which confirms the similarity of forms and properties of quantum aether vortex structures regardless of their space scale. This principle also determines the Universe as stereodynamically multivariate system in the form of hierarchy of vortical-wave structures of the unified aether (fractal matreshka). On the researching of the macroscopic objects of the Universe it is possible to make a conclusion about microcosm structure if taking into account the changes of frequencies and velocities of action transmission.

In the third place there is a principle of physical autonomy, which confirms that any solitary mass (for example a planet) creates aether system. The particular principle of relativity, which reflects one of the fractal properties of the Universe, can be applied to this system. Such autonomous mass becomes similar to the miniuniverse with its aether subspaces, which repeat the basic phases of the Universe spectrum in more narrow (which depends on the size of mass) frequency band of space-time frequencies. Thus, in the local system of the Earth solid aether reproduces the structure of gravitational field with energy "bottom" in the mass center. As the result such spherical body occurs to be an energy drain and warms up from within.

The fourth is the principle of interaction between matter and vortex-wave forms, which do not depend to the spectral part of the Universe, that is quasimatter. This is the principle of new interaction in nature. The value of energy interaction in each experiment diminishes in time according to exponential law that is explained by the forming of energy informational or adaptation barrier, which separates parallel worlds and reflects the properties of vortex tenacity of aether as superfluid medium. At that, time of interaction is proportional to the size of quasimatter and the barrier for the earthly conditions is lowered at the indefinite period, on the assumption of only thrice-repeated observation of forces (triad law).

According to this principle, aether dynamic experiments in the earth laboratory do not have classical repeatability that, from the one hand, gives occasion to doubts in the objectivity and scientific character of the non-traditional experiments and from another hand it is the most reliable test feature of macroscopic aether motions. Biosystems have special relations with this principle.

The fifth is the principle of many-dimensional autobalance of forces. All vortex and linear motions of macroscopic aether organize themselves in the way that in the band of space-time spectrum of the local system (usually with the aid of fluid and gas aether) occurs to be self-balanced, that is they have zero resulting impulse and the moment of impulse due to the existence of the proportionate antivortexes and antistreams of another spectral structure at the same space volume. The self-balanced vortex structures and streams are practically closed for the outer watch from the direction of our material world, at least with respect to the methods of classical physics. The principle of autobalance of forces reflects aether properties as unified synergetic system and has a significant applied meaning.

Let call the principle of viability of aether dynamic systems as the sixth principle. Only a stereodynamic multivariate system is a viable one, that is a system, which during a definite period of time has the opportunity, called as life cycle, to realize interconcerted self-oscillating processes of vortex-wave character simultaneously at different phase states (subspaces, layers) of aether. The most important features of such a system are its space-time quasimaterial (vortex-wave) broadbandness and finite time of existence, which is determined by the conditions of creation of the energy-informational barrier. Self-oscillation regime demands the presence of an energy source, oscillatory circuit (a pendulum) of any character, intensive process (of negative tenacity) and a channel of positive feedback (negative entropy).

In the sense, referred above, any material system is viable and occurs to be a big system in the form of coordinated community of multivariate subsystems. In its turn each big system as a part of the hierarchy is a constituent of bigger system, until everything is embraced by the Biggest System, that is the Universe.

The seventh principle of the universal energy interchange is the physical realization of the law of unity and struggle of oppositions. This principle determines spontaneous creation of thermodynamic and antigravitation potentials. Any local matter mass (a body), situated in the open space, creates an exchange process with the surrounding aether volume in the way that more fine-structure fluid aether is absorbed by the body, and the less power-consuming gas aether is radiated. As the result the body as a heat engine gets energy due to the cooling of aether exteriors. At that, antigravitation forces acts between bodies and aether exteriors, which have different temperature.

This principle, which establishes the existence of antipodes of the second law of thermodynamics and Newtonian attraction, is realized mainly in cosmic scales and explains in which way the energy is created in the bowels of planets and stars and why the Universe is stable as regards to gravitation. Obviously, the most unexpected for the modern Physics is the discovery of non-traditional nuclear processes where conditional reactions of decay and fusion occur at the usage of quasimatter.

More deep research of new experimental results and of the stated above scientific principles lets to determine the priority-driven strategic tendencies in Physics, to open more entirely the laws of mechanics and thermodynamics of many-dimensional aether, including the theory of non-traditional waves and new types of electromagnetism. At that, the supreme aim is the research of differences in aetherodynamics laws on the Earth (in a laboratory) and in outer space, the unknowing of these differences has caused logical insularity, false all-sufficiency of classical physics, which had refused as "not wanted" the aether conception and fundamental Universal laws.

References

1. Mishin A.M. On the new properties of physical vacuum, gravitational field and mass. DD USSR, 1988, p. 44
2. Mishin A.M. Experimental results on the registration of aether wind // New ideas in Natural Sciences. Series: Problems of research of the Universe, issue 18. – SPb: RAS, 1995, p. 24-33
3. Mishin A.M. The Aether Model as Result of New Empirical Conception. New Ideas in Natural Sciences. (On materials of the International Conference). Part I – SPb: RAS, 1996, p.95-104
4. Mishin A.M. The physical system of artificial biofield // "New Energy Technologies" – SPb: Faraday Labs Ltd, 2001, issue #1, p. 45-50
5. Mishin A.M. Antigravitation and new energy processes // "New Energy Technologies" – SPb: Faraday Labs Ltd, 2001, issue #2, p. 37-41
6. Longitudinal thermomagnetic effect // New Energy Technologies – SPb: Faraday Labs Ltd, 2002, issue #2(5), p. 38-41

Irving Langmuir and Atomic Hydrogen

Nicholas Moller

PO Box 201
34008 Eretria
Greece

Editorial

In this paper Dr. Nicholas Moller describes the history of development of Atomic Hydrogen technologies in details. It is remarkable that this technology can be applied not only for welding processes but also as a clean free energy source. It is important to note that in this case the hydrogen process does not involve a consumption of hydrogen, which is not combusted in the process. Atomic hydrogen is not really a fuel but rather a medium, gateway or a super-conductor of ZPE form the vacuum of space, converting ZPE radiation and ultra-high frequency electrical energy into infrared (heat) radiation.

Irving Langmuir.

This is the story of Irving Langmuir who was the first to develop a theory on Atomic Hydrogen on the basis of empirical research and experimentation. His work in this field lasted from 1909 to 1927. During this period he was employed by the Research Laboratory of General Electric Company. Patents and discoveries developed by Langmuir during his time with General Electric were to a considerable extent instrumental in laying the foundations for what is today one of the largest corporations in the world.

The question that gave birth to this article, is why his work and discoveries on Atomic Hydrogen were the only work that received hardly any attention at all and why his revolutionary breakthrough was deprived of world attention for almost 100 years? This question becomes even more relevant when taking into consideration the high standing he enjoyed with his contemporaries (including being awarded the Nobel Prize in Chemistry) and it is better to make it distantly after exclusion of man presence near experimental stands and devices. At the same time it is quite obvious that on applying of small capacity and fixed time of irradiation it is possible to develop methods for curing of human diseases, which are considered now as incurable (for instance of diabetes, some diseases of haematogenic system, of cancer and possibly of AIDS.

References

1. Proceeding of the International Scientific Conference “New Ideas in Natural Sciences” Problems of Modern Physics, p. 176-187
2. Journal of new energy, Vol. 3, #4,1999, I.M. Shakhparonov “Interaction between Kozyrev – Dirak radiation and radionuclides”, p. 85-89
3. Kozyrev N.A. Selected works, Leningrad State University, 1991, part 3
4. Lavrentyev M.M. Eganova I.A. Luzet M.L., Fominykh S.F./ / Reports by AS USSR. – 1990, Vol. 314, #2, p. 352-355
5. Lavrentyev M.M., Gusev V.A., Eganova I.A., Luzet M.L., Fominykh S.F.// Reports by AS USSR. – 1990, Vol. 315, #2, p. 368-370
6. Methodological materials on experimental pharmacological and clinical trials of immune modulating effect of pharmacological remedies. - Ministry of Health USSR, M., 1984
7. Talmadge J.E. and Chiragos M.A. Comparison of immunomodulatory and immunotherapeutic properties of biologic response modifiers. Springer Seminar Immunopathol., 1985, 8, 429-443

Effect of Magnetic Blow Wave Field on Wine Systems

I.M.Shakhparonov (Corresponding author), S.A.Grin,
S.R.Tsimbalaev, L.N.Kreindel, V.N.Kocheshkova,
A.I.Podlesny, S.Yu.Gelfand

AGD Firm, Peschanyi pereulok, House No.20P korpus No.1, Lfl. 33
125252, Moscow, A- 252, Russian Federation

Russian Institute of Canning Industry, Shkolnaya Street. 78.
142703 Vidnoe 3. Moscow Region, Russian Federation

Introduction

Authors communicate the data on influence of Magnetic Blow Wave (MBW) field on several wineproducts. It was found, that MBW did not lead to significant changes in the major components of the wineproduct (sugar, organic acids, minerals). At the same time the taste and aroma of treated wine become more pleasant; content of heavy alcohols and wine stone in the treated samples was less than in non treated ones. A mechanism of transformations was also discussed.

Keywords: *Magnetic Blow Wave (MBW), Wineproduct, GLC of aroma compounds and ethanol, HPLC of sugars, Atomic Absorption Spectrometry (AAS) of minerals, Heavy alcohols and aldehydes, Wine stone, Turbidity tendency, Organoleptic evaluation*

Magnetic Blow Wave (MBW) was obtained for the first time during the investigations on ball lighting generation under the laboratory conditions (Shakhparonov 1994). MBW as a physical object is interesting because of some facts, which suggest that MBW is a magnetic monopole. The MBW can also interact with the matter and transforms it in a definite way. Typical example is an elementary carbon in the form of graphite, which is transformed by such magnetic treatment into ferromagnetic substance (ibid). The graphite, which is initially diamagnetic, transforms to paramagnetic one with general radiation doze of about $7 \cdot 10^{19}$ neutrons/cm^2. Other types of radiations could not affect this way (Svoistva 1975). So one unit of MBW can be considered as $1 \cdot 10^5$ of neutron masses. This fact may be regarded as an indirect evidence for assuming that MBW and magnetic monopole are the same things. In the absence of excited radioactivity a slow MBW [$v/c < 1 \cdot 10^4$] occurs, which does not ionize atoms (Devons, 1963). Therefore, their interaction with the matter can be observed only indirectly. No data exist on the interaction of MBW with organic substances. The experiments and results reported in the present communication may be a starting point for development of technology and to formulate the methods for vintage wine and best quality spirit production.

Materials and Methods

Assuming that MBW and magnetic monopole are the same things, a number of conditions were selected for all experiments. The MBW source and the samples were placed in the same axis and the axis was oriented according to magnetic meridian direction. Such magnetic orientation is appropriate, as the energy of magnetic monopole theoretically increases in a magnetic field (Devons, 1963). All of samples were placed at 250 cm distance from MBW source, in hermetically closed glasses. It should be noticed that MBW could penetrate through many other barriers, for example into cast iron reservoir with wall thickness of 5 cm (Amaldi, 1970).

The quality investigations were made by using of standard equipment. HPLC, equipped with refractometric detector was used for sugars estimation. Separation of organic acids in forms of their ethyl esters and acid esters was carried out chromatographically using a column packed with polyethylenglycol succinate and the following temperature option: initial temperature is 120°C, final temperature is 220°C, temperature growth rate: 8°/min. GLC was also employed for determination of ethanol. Minerals content was examined with Atomic Absorption Spectrometry (AAS). Electronic spectra of samples were obtained with double beams UV Vis spectrophotometer equipped with permanent wavelength scanning. Redox potential was measured with EV-74 potentiometer.

The aroma alterations in the wine samples were investigated by GLC method after preliminary concentration of aromas by solid phase adsorption. The concentration was carried out by barbotation of inert gas (nitrogen) through liquid and consecutive catching of volatiles with tube trap, filled by Polysorb 1 sorbent (Lur'e 1972). The well-known analogue of Polysorb 1 is Porapak Q. The tube may be regarded as a short chromatographic column, and volatiles go through it according to their retention times. The choice of sorbent was motivated by the fact, that retention times of water and ethanol was rather small (ibid). Thus, a concentration process can be ended at the moment, when water and ethanol have passed through the column, as the other volatiles remained bonded. The aroma desorption was made with ethyl ester. The analysis of the concentrates obtained was carried out with gas chromatograph equipped with flame ionisation detector (FID), column 3 m x 3 mm, packed by Carbovax M on the Supelcoport. Temperature for the analysis was programmed from 100 to 190°C with increase of 1°/min. Isothermal conditions in the borders had durations of 2 and 40 min respectively. The "mild" conditions of separation were also employed (initial oven temperature was 80°C with isothermal condition duration 5 min, temperature growth rate 1°/min, final temperature 150°C and isothermal condition duration 40 min).

Optical activity was tested with Spectropol at D line of Na (580 nm). The samples were evaluated organoleptically by a group (12 persons) of workers from Russian Institute of Canning Industry. Turbidity tests were made under the methods of Valuiko et al (1987). In some cases, qualitative tests were completed by MPL turbiditymetric measurements. Before testing samples were filtered. Determinations of heavy alcohols and aldehydes contents were carried out in accordance to National Standard (GOST, 5363-67) as follows below. Determination of the constituents of "heavy spirits"(i pentanol, i butanol) was based on reaction of the sample with salicylic aldehyde in a presence of H_2S0_4. Rose colour develops if sample contains the heavy alcohols. The density was measured with Vis-photometer and the quantative determination was carried out using standard graph made with mixture solution of i pentanol and i butanol. A method for determination of aldehydes content is based on a reaction of fuchsine sulphite. The developed colour was measured with Vis-photometer. Calibrating plot constructed basing on typed solutions was used for quantification.

Results and Discussion

Investigations of wine quality changes after MBW treatment were performed using two samples of portwine ("Zemfira") type wine. Sample 1 was a reference (non treated) and sample 2 was treated with MBW.

In both samples, the fructose and glucose levels were practically the same and amounted to 43.8 ± 3,32.22.5 g/*l* respectively. Sucrose and maltose were absent. Total sugar content was 76.0 g/1 though the level marked on the label was 80 g/*l*. It is thus apparent that the treatment of wine with MBW does not lead to noticeable changes of sugars content. Results of organic acids determinations are given in the Table 1.

Table 1
Main organic acids content, g/1

Acid	Treated wine	Initial wine
Lactic	0.0265	0.00187
Oxalic	0.010	0.0088
Succinic	0.209	0.18
Malic	4.56	4.22
Tartric	0.0805	0.0895
Citric	0.401	0.483

Standard deviation for the determination method was estimated as 7 %. This fact shows that differences in organic acids content are not significant. It should be noted that a tendency of slight increase in light acids (up to malic) in the treated wine was observed in contrast to noticeable change in more heavy acids. The ethanol content of both of samples was 181 and 184 g/*l* for non treated and treated samples, respectively, though the label on the bottle indicated 190 g/*l* concentration. Standard deviation was 5 %. Thus, MBW treatment does not lead to significant changes in alcohol content.

Atomic Absorption Spectrometry (AAS) data indicated that the samples were practically identical in terms of K, Na, Ca, Mg, Fe, Cu and Zn contents (data are not shown).

Similarly, spectra of treated and non treated wines, diluted 150 times before photometring, were practically identical, thereby pointed out that polyphenols are unchanged.

When wine is industrially treated with IR or microwave heating, ultrasonic, ultraviolet and g radiation, different reactions occur and there include redox reaction, esterification, condensation, hydrolysis, Maillard reactions, etc (Kishkovsky 1988). Most of reactions are accompanied by redox potential changing. Increase in Redox potential points out the increase in concentration of oxidants, i.e. oxygen, peroxides, and other compounds, which are electron acceptors. Redox decrease is a result of oxidation processes (ibid). Redox potential was practically constant ($\Delta E =$ 145 mV and 150 mV in samples # 1 and 2 respectively). Evidently, oxidation processes, like they occurred during heat treatment, were absent during the MBW treatment.

One of the important reactions to be considered is the Maillard reaction. Essentially it appears in form of browning, decrease in reducing sugars and amino acids, and new aromas formation. While our result evidences on absence of irrelevant aromas, alterations of wine colour, and sugar content, thereby indicating insignificant contribution of Maillard reaction on wine quality changes due to MBW treatment. Technological treatment often leads to esters accumulation that improves wine aroma. It's well known that the most important in this context are the esters of C_6-C_{14} fatty acids (Kishkovsky 1988). During heat treatment, storage and other physical influences, different kinds of acid esters accumulate. These have weaker aroma than esters of fatty acids. But their appearance proves the existence of esterification processes. A comparison of aromas chromatograms of the samples 1 and 2 proves occurrence of changes in concentrations of the individual substances (increasing of peaks length with retention times of 13.10, 100.9; decreasing of peaks length of 54.85 min). An order of peaks exit of different volatiles is given in the Table 2.

Table 2

Exit order of different volatiles

Exit order, from published data for Carbovax 20 M	Retention time in our experiments, min	Exit order, from published data for Carbovax 20 M	Retention time in our experiments, min
Acet aldehyde	3.8	i-Pentanol	23.56
Ethyl acetate	4.71	i-Amyl butyrate	
Diacetyl	4.82	Acetone	
Methanol	4.91	n-Pentanol	27.92
Ethanol	6.01	i-Amyl valerate	
n-Propanol		Ethyl lactate	
i-Butanol	13.58	Ethyl caprilate	52.6
Butyl acetate	13.59	Acetic acid	
i-Butyl acetate		Diethyl succinate	
Ethyl valerate		Ethyl laurate	
n-Butanol	17.73	Phenyl ethanol	
Amyl acetate	20.70	Diethyl malate	

A comparison of retention times of components with peaks of standard substances of wine aroma indicates that butyl acetate and i-butanol are very close to peak 2. Data on chromatographic separation with mild condition showed that i butanol and butyl acetate peaks exited simultaneously. Organoleptic evaluation recorded a nice smell in the treated wine, thereby due to the formation of butyl acetate. Data indicated the presence of ethyl malate, ethyl tartrate and ethyl citrate in the samples, in addition to two peaks corresponding to ethyllactate and ethyl oxalate. The large experiment error does not allow any inference on changes of their height after magnetic treatment. Identification of peaks with retention time factor especially in such complex system is not unquestionable. However, the best way is to use the chromato mass spectrometer, which allows inference according to their individual mass-spectrum.

Organoleptic evaluation can depend on aliphatic alcohols content. Determination with GLC shows (Table 3), that their quantity in the both samples is rather small with respect to average values taken from literature for this type of wines. Thus such changes can not be recognized with such evaluation. For both of samples pH was equal to 4.0.

Table 3
Aliphatic alcohols content, mg/1

Alcohol	Sample #1	Sample # 2	Average literature values (Kishkovsky 1988)
Methanol			80-350
i-Propanol			0,3-3
n-Propanol	less than 20		5-50
i-butanol	less than 20	less than 20	20-100
n-Butanol	less than 10	less than 10	2-10
i-Pentanol	less than 20	less than 20	100-250

The results of optical activity measurements indicated, that both samples are not optical active. Filtration, clarification and dilution could not change the optical activity. Perhaps, there is a compensation of different forms of D- and L- compounds in the samples, thus total activity was very close to zero, and magnetic influences could not change equilibrium between the forms.

Organoleptic evaluation of more delicate taste and aroma of the treated sample with respect to non treated one, MBW treated sample as more complete, harmonic, noble, and natural in contrast untreated sample was recorded as excessively bitter and sour in spite of practically the same pH of samples.

It is interesting to test the tendency of wine to make a different kind of turbidity after the MBW treatment. Data showed that both samples were not positive for protein turbidity. In term of reversible colloid turbidity formation, after storage at 7.5°C for 1 day, the MBW treated sample was homogeneous, in contrast to the formation of different phases with different refractometric numbers in untreated sample. Both the phases in untreated sample were liquid, with a density very close to each other, but the borders of phases were like broken lines when crystallization begins in crystallization process. This alteration in untreated sample may be due to micelle state changes or of structurization of product.

The tendency test for polysaccharide turbidity based on the reaction with phenol in presence of H_2S0_4 and determination of the derivative formed by photometrically, indicated, that difference in concentrations of polysaccharide in the both samples are very small, the levels being 119 and 106 mg/*l* for untreated and treated samples respectively. These values are close to range of polysaccharide stability (150-200 mg/*l*), and thus do not allow any conclusion on changes of relative stability of the samples. A tendency for polyphenols turbidity, due to polyphenols associates precipitation upon addition of salt did not show differences. Turbidity, as determined in MPL apparatus, was 15 FEM as against value of 0.2 FEM before testing in untreated sample. These numbers were respectively 14 and 0.3 FEM for treated sample. Thus it indicates that both the samples are very stable with respect to polyphenols turbidity and that the magnetic treatment does not lead to alteration in the polyphenols stability.

The data on the colloid stability indicate, that both the samples showed rather high resistance against protein, polysaccharide and polyphenols turbidities. Besides, treated sample showed higher stability with respect to reversible colloid turbidities.

It is interesting to investigate as to how heavy alcohols and aldehydes, which are often produced, when low-grade technology is used, are affected by magnetic treatment. For these studies, a system of simple mixture, consisting only of spirit and water, was used. Commercial vodka bottled in standard 0.5 l bottles and artificial solutions, containing 40 % of food derived rectified spirit were used. Data showed, that MBW treatment significantly influenced the heavy alcohols content, as the reduction in heavy alcohol was more than two times. In addition, it reduced aldehydes by more than 3 times in vodka, and more than 30 % in rectified spirit. Data indicate that efficiency of aldehydes removal is higher when the sample contained higher level of aldehydes. Thus, the MBW treated vodka and rectified spirit will be better than untreated one. It is however stressed that untreated samples were also recorded as good by sensory panel. So, limits for aldehydes are usually present in high quality vodka established by National Standard (GOST 5363- 67) are 6 - 15 mg/*l*. Thus the organoleptic evaluation of samples does not allow to find difference in aldehyde levels in these samples.

Table 4

The main results of heavy alcohols and aldehydes determination (mg/1) in vodka and solution, contained 40 % of rectified spirit

Substance	Non-treated vodka	Treated vodka	Non-treated spirit	Treated spirit
Alcohols	8.7	2.55	3.38	1.5
Aldehydes	1.5	0.4	0.6	0.4

Sediments formation and its character were also evaluated. The sediment in treated grape juice was dense and more dark, the formless, non crystalline sort, and gel like form. The volume of the sediment occupied up to 30% of total volume. The sediment did not sink or float, nor it stick to the walls of glass. It was found that 100 ml of juice gave about 155 mg of dry sediment. Microscopic investigations showed an absence of any kind of bacteria or fungi in the sediments.

The effects of high energy of magnetic influences on sediment were also investigated. The experiments were carried out with "Portwine Erevanski, vol. 0.5 1, white, spirit content 19 vol %, sugar 10 %, prepared according to GOST (National Standard) 7208-84". Crystalline sediment appeared on the walls and especially on the bottom of the bottle after the BMW treatment. An amorphous precipitate was also presented, and it can be separated by decantation. Crystalline sediment, after washing with ethanol and drying to constant weight, weighted 69.2 mg, and was of bright brown colour. A tartrate content as a tartrate acid, of the sediment was 59% mass. If it is considered as a tartar (a wine stone) of potassium sodium tartrate, then tartar content in sediment works out to be 86%. If it is considered as a tartar of dipotassium tartrate, tartar content in sediment will be 93%.

Generalization of data shows the positive effect of magnetic treatment on the wine samples, leading to harmonic taste of treated wine and absence of non pleasant tastes. Most of the changes were found to bein the flavour and taste components, which were minor substances in the product. For example esters concentration changes during the treatment. At the same time the content of major components, such as sugars, organic acids, particularly, heavy organic acids, and especially ethanol remain constant. It seems logical from kinetic point of view, when simple processes, like esterification, are preferable with respect to many stages reactions, and reactions with high activation energies, which can go at hard conditions. Also, it seems logical that magnetic treatment may influence on electrical state of colloid species. Thus magnetic treatment can be considered as mild, selective in the comparison with many other physical methods. Nevertheless, the changes lead to acceptable energetical and nutritious value of the product. Data show that difficult problems, such tartar removal, can be solved by MBW treatment.

References

1. Amaldi E. et al in: Preprint CERN, Report 63 13. Search of Dirak Magnetic Pole, 1970.
2. Devons S. Search for Magnetic Monopole, Sci., Progr. (No.204), 601 (1963).
3. GOST (National Standard) 5363 67. Vodka. Metody ispytaniy (Testing Methods).
4. Kishkovsky Z.N., Skurikhin I.M., Khimiyavina (Wine Chemistry), Moscow: VO "Agropromizdat", 1988, 253 p. (in Russian)
5. Lurle A.A. Sorbenty i khromatograficheskie nositely (Handbook on Sorbents and Chromatographic Supporters), Moscow, Khimiya, 1972, 320 p.(in Russian)
6. Shakhparonov I.M. in: Sharovaya molniya v laboratorii (Ball Lighting in the Laboratory) / Collection of Articles. Moscow: Khimiya, 1994, 400 p.
7. Svoistva konstrukzionnykh materialov na osnove uglegoda (The Properties of Constructional Materials Based on Carbon) / Handbook/ Nagorny V.G., Kotosonov A.S., Ostrovsky V.S., Dymov B.K., Lutkov A.I., Anufriev Yu.P., Barabanov V.N., Belgorodsky V.D., Kuteinikov A.F., Virgelev Yu.S., Sokker G.A. Moscow, Metallurgy, 1975, P. 73 77 (in Russian)
8. Valuiko G.G., Zinchenko V.I., Mekhuzla N.A. Stabilizatsiyavina (Wine Stabilization), Moscow: "Agropromizdat", 1987, 160 p. (in Russian).

The Fundamentals of the New Principle of Motion

By The Group Studying Inertialess Natural Processes (GSINP)
123430, Moscow, Mitinskaya Str., 40-1-244 Email: gibip@mail.ru

P. Sherbak

The concepts of active and passive interaction between the moving object and the space form the basis of the new principle of motion.

So as to be more understandable, let's consider what is the old principle of motion. For this we will use the concept of a moving object and the space in which the object is moving. Naturally, material objects and the space can't interact between each other directly, because the space is the philosophical category. In this case we can understand physical essence of natural phenomena easily. In our view, the material objects interact with some fundamental energy of space (FAM), which fills all space with a different density. Thus the energy (FAM) is inalienably connected with the space. One of the first names of this energy is "ether" in the early scientific works. So, for the simplicity we will accept that the object and the space interact between each other.

Thus, all existing methods of motion which have been invented by mankind till the present time are based on activity of the material object that means the one expends some energy to produce the motion, and at the same time space is passive, it means that space does not need to spent any energy to move the object. And so in common case space tries to keep the object in the former state interfering with accelerated movements of the object (in accordance with **the 1st , the 2nd and the 3rd Newton's laws**).

It should be noticed that such method of motion (for the speed, which is much less than the speed of light) takes place both in animate and inanimate natures. In this case the level of energy of motion and reaction of space (or an environment) are not very high. Incidentally, the energy of object can be of different types: electrical, chemical, biochemical, mechanical etc. The common consequence of this type of motion is existence of the ***inertia***. The classical physics can't answer the question: "what is inertia?» The same situation is applied to the concept of ***mass,*** which is closely connected with inertia. The classical physics says that ***the mass is a measure of inertia.***

There is the new principle of motion of material object: the object is passive and space is active. In this case it's more favorably for space in the energy aspect to move the passive object and to spent some power then to keep the object in the present place in the former state of immobility (in accordance with **the 1st, the 2nd and the 3rd Newton's laws**). And so we should introduce **the 4th law of** Newton's **mechanics**. It says that ***there are the systems of coordinates in which the body is moving not rectilinearly with acceleration when this body is in the state of immobility***.

The basic and the main differences of the offered principle of motion from the existing methods at the end of the XX century are the following:

1) The absence of inertia of motion;

2) There are no limits for the speed of motion;

3) The absence of "fuel reserves" "on board" of the moving material object.

Nikola Tesla and Instantaneous Electric Communication

Vladimir I. Korobeynikov

Russia
E-mail: elen@mail.infos.ru

Nikola Tesla (1856-1943), an outstanding inventor, was and still remains one of the most mysterious persons in the history of electrophysics. Whereas the most scientists were moving together in direction of microparticles investigations, as the basis of matter structure and of nature itself, he was going in opposite direction. He had a keen interest in the investigation of electric charge of the Earth as a whole. He was looking for the ways to influence on it, to control its state and methods of its regulation.

Therefore, exactly, the most of his searches, experiments, the purpose of constructions and buildings, created according to his conceptions, cause perplexity and misunderstanding of scientists even in nowadays.

The most mysterious of his main experiments were made in USA after 1904. After Nikola Tesla death in 1943, all his diaries and records over a period from 1904 year had mysteriously disappeared. Probably they were stolen (it was known, what to take). Lost records could “cast light” on one of the most “strange” of his buildings in the form of the enough tall tower, on the top of which a specially created toroidal transformer was placed. This transformer could create there a huge electric potential up to the billion volts.

Nikola Tesla switched on this tower-device, what caused the fright and even panic in mind of people from nearby settlements. Of course! Because of very high electric potential there began air ionization, which spread very high to the atmosphere accompanying by the effect of color play. Such luminous, color-playing sky caused even a horror of people, who knew nothing about the experiment made and its goals. They did not guess that Tesla by means of the electric charge, created of the tower, was influencing on the electric charge of the Earth as a whole (about 600000 Coulomb). There was a global scale in Nikola Tesla’s investigations.

There is no point in detailed analysis of the fact that the potential of the tower top influenced on the Earth charge. Interaction of charges-balls with the distortion of field lines, distortion-distribution of charge on their surfaces, induced charge, is beautifully described even in school physics textbooks. In Nikola Tesla investigations the Earth had the role of one of the charged balls. It was possible by changing of charge on the tower to deform electric charge distribution on the whole Earth surface at once. This deformation (electric currents) could be fixed at once in every point of the Earth surface. It is alluring to use this effect for data transfer telecommunication, both on the Earth, and in space.

After such introduction the question “How does the system of instantaneous electric communication for any distance look like and work?” is still opened. First of all, the readers need to know, that such instantaneous communication is possible in principle. The proving it theoretical calculations, are rather difficult for popular interpretation. Some part of readers can take it on trust, and those who are most interested in can apply to works of Oleinik V.P. (quantum physics) the professor from Kiev Polytechnic University. At the minimum there are two necessary works: Oleinik V.P. “Faster-than-light transfer of a signal in electrodynamics. Instantaneous action-at-a-distance in modern physics” (Nova Science Publishers. Inc. New York. 1999) and Oleinik V.P. “Latest development of quantum electrodynamics: self-organizing electron, faster-than-light signals, dynamical heterogeneity of time.” (Physical vacuum and nature. 4. 3-17. 2000).

“PC” magazine has devoted a rather significant article entitled “Computers and teleportation” to V.P. Oleinik works, concerning instantaneous electric communication (“PC” #6, 2000). Note, that the author of the given article has also found the possibility of instantaneous electric communication, but by means of materialistic methods, absolutely different from Oleinik’s ones, what is most important – two different solutions point to the possibility of this communication. “PC” #6, 2000 in the article “Circles on fields” cited mathematical formulae of the structure of electron electro-magnetic field as an illustration (it refers to the Earth too) that the author of this article has got.

The most attentive readers of that article could notice, that one vector Hz absolutely “ignores” Special Theory of Relativity, since its mathematical expression does not include the velocity of light, whereas it presents in other vectors as a product of electric and magnetic conductivity. Magnetic line of this Hz vector goes to infinity and returns back from infinity. It surrounds the whole Universe. It is alluring to use exactly this (Hz) line for the **instantaneous** communication for any distance.

It is not so difficult to do it. In the Fig. 1 the easiest and most available for understanding line of the **instantaneous** electric communication is shown. A rotating charged dielectric ball (an “electron”, isn’t it?) is used as transmitter. The ball can be electrically charged up to the limit of charge flow-out into the ambient space. Around the charged rotating ball there appears electro-magnetic field, entirely analogous to the electro-magnetic field of the Earth (and of the electron too). The central magnetic line Hz goes to the infinity and returns

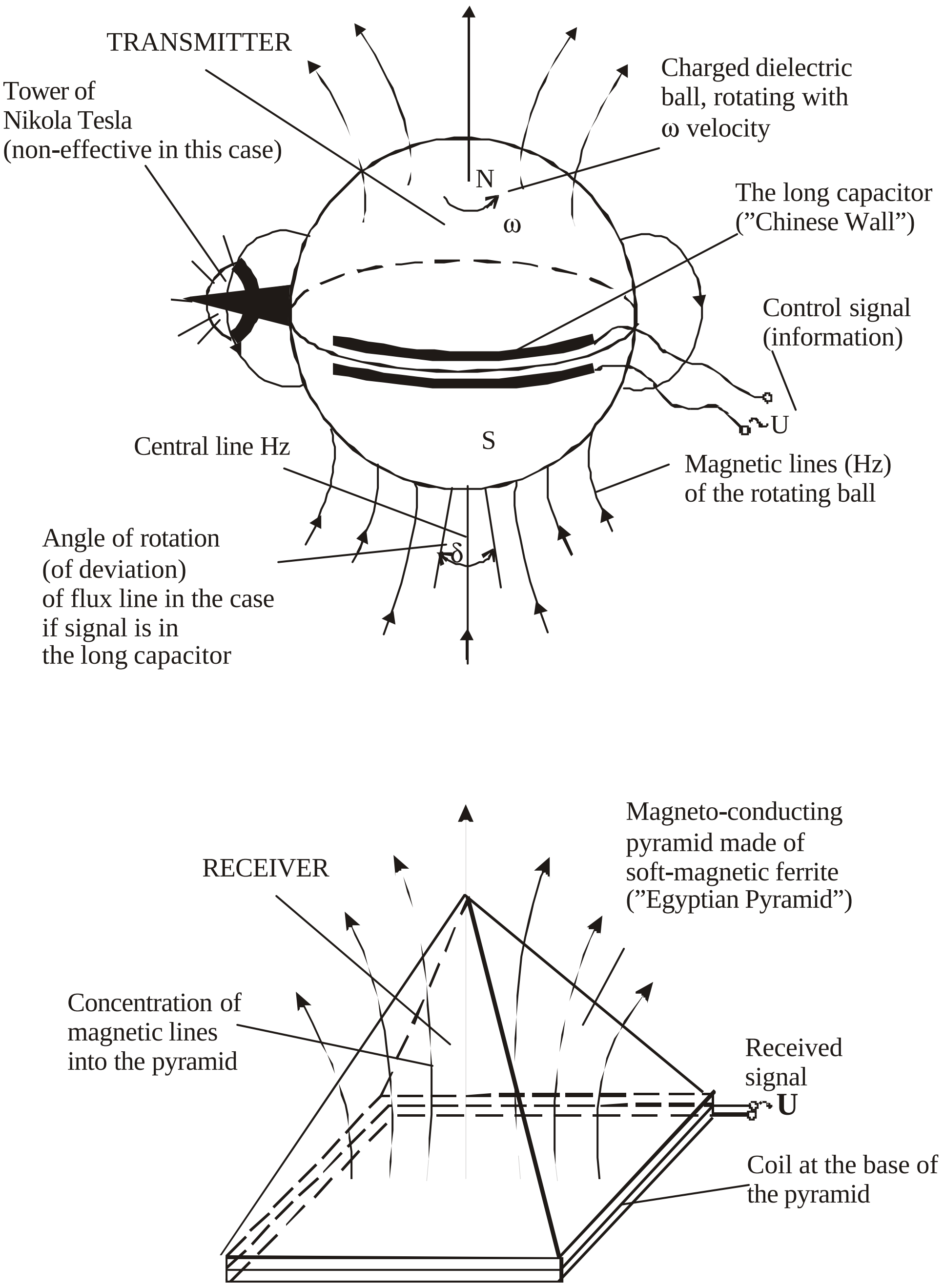
TRANSMITTER
Tower of Nikola Tesla (non-effective in this case)
Charged dielectric ball, rotating with ω velocity
N
ω
The long capacitor ("Chinese Wall")
Control signal (information)
U
S
Central line Hz
Magnetic lines (Hz) of the rotating ball
Angle of rotation (of deviation) of flux line in the case if signal is in the long capacitor
δ
RECEIVER
Magneto-conducting pyramid made of soft-magnetic ferrite ("Egyptian Pyramid")
Concentration of magnetic lines into the pyramid
Received signal
U
Coil at the base of the pyramid

Fig.1

Line of the instantaneous electric communication on the basis of rotating charged ball and pyramid

back from it to the opposite side of the ball. In the same way the central magnetic line of the Earth (Hz) goes from one pole to the infinity through the whole Universe and returns from it to the center of the opposite pole.

If by the information to force the rotating ball (electron) to "wag by tail" (by Hz vector) which stretches through the whole Universe, then this "wagging" can be controlled instantly in every point of the Universe. While the rotating ball has a steady distribution of the surface charge, the line Hz does not change its dynamic position in the Universe.

If such distribution of the surface charge is broken, space position of the line Hz also will change. On mounting the Tower of Nikola Tesla on the surface of rotating ball and measuring the potential on this tower in time with an information it is possible to change the charge distribution on the ball, and, respectively, the space position of central magnetic line (Hz) in the whole Universe at once. Big disadvantage of the Tesla tower is that maximum influence on charge is executed in the point under the tower, and farther it began decreasing roughly (exponentially), according to physics laws.

Hence it is advisable to influence on the whole surface of charged ball, but not on some point of it. It is possible to influence at once on the very big part of surface by the long capacitor, placed on the perimeter (equator) of the charged rotating ball. Because of optimality reasons, this capacitor length should not exceed a quarter of the ball perimeter (equator) length. Charging and discharging this long capacitor on the ball equator by the data signal, only the position (angle 5) will be changed, not a value of the infinitely long magnetic line (Hz) in the Universe. It is a data transfer.

The natural question appears: "How to make on the Earth the most powerful transmitter for the **instantaneous** electric communication?" The answer suggests itself: "It is necessary to use the Earth itself as a rotating charged ball." It is not effective to use Nikola Tesla's Tower to deform the Earth electric charge. To place on the Earth very long (about thousands km) capacitor is quite easier. However, it must be placed not on the equator exactly, but moved a little bit because of the initial heterogeneity of the Earth surface charge distribution, caused by the presence of continents and oceans. It will be necessary to find the line of electric equator, where the amount of charge north and south of it is similar. This line will not be ideally straight and will be situated near the 30th parallel.

As a matter of fact, this grand capacitor is already built, but is half-broken. This capacitor is very well known – it is a Great Chinese Wall. The ancient, powerful Chinese Tzcin' Shi Huandi empire adapted and used it (capacitor) for the protection from nomads incursions. How unexpectedly and originally it is! In this case the electric iron would be the best tool for spiking. It is clear enough that the charged ball (as well as the Earth) will "wag by tail", which stretches through the whole Universe and does not change its energy, but only changes its position in space in time with information. Now we can go on to the question, concerning the way to control the Earth "wagging by tail" in the Universe, and thus to read information **instantaneously** in any point of the Universe.

In the Fig. 1 it is shown the input device of the electric communication receiver, made of the magnet sensitive material (it can be soft-magnetic ferrite) in the form of pyramid, with the proportions of well-known Egypt pyramids. Magnetic field lines of the far space pass through the pyramid from the top to the base and are concentrated by pyramid. If there is no signal (the "wagging by tail" of the far planet-transmitter is absent), then the magnetic flow, coming through the pyramid, does not change, and induced voltage in the coil, placed in the base of the pyramid, is absent (no information). If "wagging by tail" begins, then the magnetic flow, coming through the pyramid, will change, and it will cause the appearance of voltage on the coil in the base of the pyramid in time with the information.

Thus, the signal is received instantly. Here it is necessary to remind once again the difference between the instantaneous and usual radio transmission. The usual radio transmitter for the transmission of the information uses the **energy** distortion of space by the information. This energy change in space happens with the velocity of light and hence there is the loss of time for information passing. In the considered case there is no **energy** change in space, there is only a change of magnetic lines position (Hz).

This is exactly the vivid and fundamental difference between the usual electric communication and the instantaneous one. **In other words, in usual transmitter during the fixed time interval there is the change of signal energy (instantaneous value), whereas in instantaneous transmitter there is no this change (only information). This is exactly the fundamental difference.**

Evidently, to receive instantly the signal from the opposite part of our Galaxy, we need rather big pyramid, in order to concentrate a big amount of field lines into the oscillatory circuit under the pyramid. The question can appear: why the pyramid, why not a cone? The point is that lines of the Earth magnetic field (the very lines that compass needle reacts on) in the any place of the pyramid horizontal section have the same density of distribution and are directed strictly parallel to the pyramid base. The cone in its horizontal section cannot provide such uniformity of distribution that is why it is not advisable to use it. From the space magnetic field lines pass through and concentrate in the pyramid strictly at right angle to the pyramid base.

This is the riddle of pyramids wonderwork. Any person coming into a pyramid, at the same moment feels the change of mental and physical condition of organism; whish is very different from that it was before the entering into a pyramid. Of course! Visitors come inside, into concentrated magnetic field lines of the powerful and functioning magnetic core of the receiving electric circuit, what is absent outside the pyramid.

It is strange, but most of tourists are afraid of the ill effect, which can be produced by electrical systems on their health, but there they stand in a queue to feel this effect in pyramids. Concentration and division of magnetic field lines are the easy and effective way to reject a noise, created by the Earth magnetic field.

It is clear, pyramids should be oriented very thoroughly, so that lines of the Earth magnetic field would be strictly parallel to the base and to the opposite (East-West) sides of pyramid. To get such exactness of orientation in modern conditions is very problematically.

The most convenient place to build a pyramid (pyramids) is on the electric equator, in the place of its intersection with the electric meridian. Such place is located in Egypt, near its capital Cairo. And again we meet a paradox: such pyramids are already built on the Earth, but they are half-broken. And Egypt was not less powerful than the ancient Chinese empire.

The Egyptian dynasty of Pharaohs has “completed” and adapted pyramids to burial-vaults, where mummies of dead Pharaohs were buried. Perhaps, it is even more incredible than in China. The impression is given that ancient powerful civilizations on the Earth had a competition between themselves, who will use radio-engineering constructions for instantaneous galactic communication in the most incredible way. Let’s give to a reader an opportunity to select a “winner”.

It must be noted that «PC» already published information that the Chinese Wall and Egyptian Pyramids are radio engineering constructions, intended for the instantaneous galactic communication (PC #114, 1997, etc).

There appears an interest in the possibility to produce very simple and manufacturable systems of instantaneous electric communication right now. Radio-electronic industry can produce them, but still does not guess about it.

Furthermore, such systems of instantaneous electric communication can be created at home, and even senior pupils are capable to use them. In the Fig. 2 there is shown the construction of instantaneous electric communication line in comparison with the usual one. It can be produced even at home conditions. Two permanent magnets, connected between each other by analogous poles, are used as transmitting circuits.

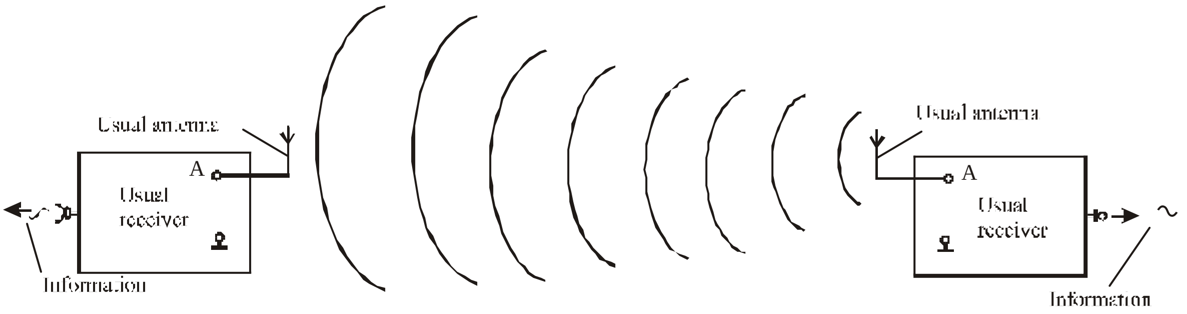
Usual antenna
A
Usual receiver
Information
Usual antenna
A
Usual receiver
Information

Line of the usual (Einsteinian) radio communication

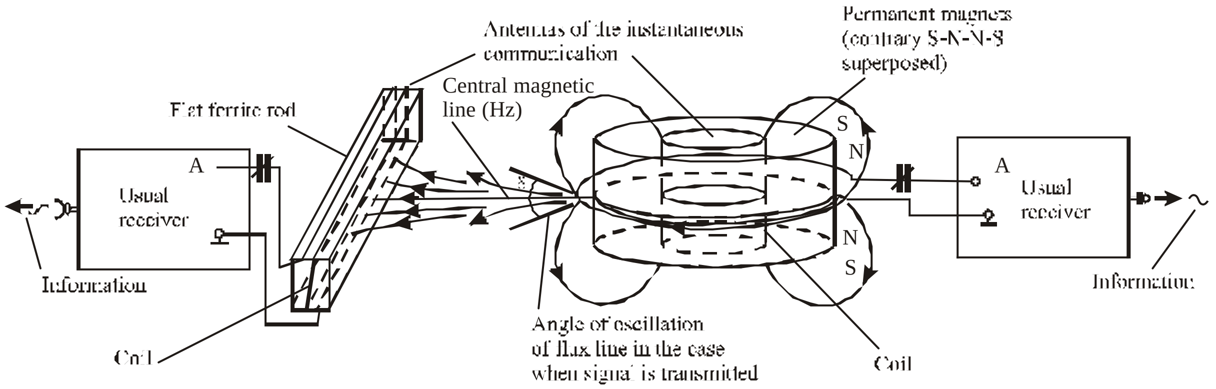
Antennas of the instantaneous communication
Permanent magnets (contrary S-N-N-S superposed)
Central magnetic line (Hz)
Flat ferrite rod
A
Usual receiver
Information
Coil
S
N
N
S
A
Usual receiver
Information
Angle of oscillation of flux line in the case when signal is transmitted
Coil

Line of the instantaneous (Maxwellian) radio communication

Fig. 2

Transformation of the usual line of communication of usual receiver-transmitter into the instantaneous one

Permanent magnets can be replaced by electromagnets. In the magnet connection point there is a coil, which while the signal passing through it will change its position (angle 5) in the space of the central magnetic line (Hz), coming out from the place of two magnets connection. Receiving circuit is available to be made of the flat ferrite, but coil must be winded along, through butt-ends of core, so that the whole internal part of the coil would be maximal (in area extent) filled by ferrite.

The obtained coil can be completely “winded” (screened) by flat ferrite of big size. For more clearness of the experiment the central magnet line of transmitting part must be directed strictly along the axis of the receiving coil.

Now, if we give the alternating voltage (information) from transmitter to the transmitting coil, fixed on the permanent magnet, then the receiver, connected to the circuit of the transmitting coil, placed on the flat ferrite, will detect an alternating voltage (information). Maximum effect is achieved at the resonance (coincidence of transmitter and receiver frequencies).

It is checked. It works. The dullest experts in radio-electronic (after the reading above) can rejoin, without making an experiment, that it is an absolute nonsense that any communication is out of the question. Coils with absolutely perpendicular axes, besides one of them is screened, do not interact with each other.

And here the most interesting thing starts. In the Fig.2 as it was mentioned above, the usual communication line and the instantaneous one were compared. The usual transmitter cannot generate the vector Hz that is why systems of usual and instantaneous communication cannot see each other in strict sense. What does it mean? It means, that in the same city it is possible to transmit on the one-carrier frequency ("what a nightmare!") two absolutely different television channels without any noises of one to another.

Usually by frequency match of a transmitter to the working frequency of another one, the radio communication is broken, but here it does not happen. Here some additional explanation should be given. As the vector Hz, which "ignores" the theory of Einstein is received from Maxwell equations, it follows that the usual (Einsteinian) system and the instantaneous (Maxwellian) one work on mutually perpendicular electromagnetic field lines (vectors).

In the Fig. 2 such difference is shown clearly. These are just "jokes" of complex numbers, when one value is absolutely perpendicular to another and nevertheless together they form a single whole. In other words it means, that two greatest persons in science Einstein and Maxwell as a matter of fact are something like "Siamese twins", completely grown together at the angle of 90 degrees, even by heads. On the one hand every one is on his own, but nevertheless they are the common (complex) organisms.

Hence there are a lot of misunderstandings on happened phenomena. How many scientists tried to find some mistakes of Einstein? They produced very convincing proofs concerning instantaneous interactions in nature. These scientists did not suspect that time and still do not guess now that they already for a long time are "walking" in the complex physics, which still does not exist. Einstein and Maxwell ("Siamese twins"), each occupies his own part of the complex number (complex physics) and they cannot be already taken off from there.

The only third, free "vacancy" is left to throw on the both of them at once the common "collar" and "reins", i.e. to fasten them ("twins") together by module and argument as any complex number. In this case no matter how the one part of complex number "ignore" the other one, only its argument will change, and module always will be equal to the conditional unit. Only this single fact in principle changes the conception about "appearance" and "disappearance" of elementary particles! Even considered instantaneous and usual communications on the same receivers and transmitters in a complex conception (complex physics) eliminate appeared "misunderstandings" of all kinds.

Advantages of the instantaneous (Maxwellian) electric communication are especially evident during the connection with long-distance spacecraft. At present in the interval between sending of control signal to the station in region of Solar System peripheral planets and getting the reply it is possible to have a small break for dinner (it is very convenient).

In the case of the instantaneous electric communication use, duty operators will have "no dinner". Moreover, the system of instantaneous electric communication can realize two-way communication underwater and from underwater to overland. It is clear that input and output circuits of such system must be covered by slushing composite for the protection from aggressive effect of the salt sea-water. Such systems of instantaneous communication are very required to submarines.

Now, when readers know and understand the principle of operation of instantaneous (Maxwellian) electric communication systems and their advantages over usual ones (Einsteinian), we can only wait, when radio-electronic industry will start to produce these very required systems.

ELECTRIFYING TIMES

an online and published magazine about Electric, Hybrid, Fuel Cell Vehicles, advanced batteries, ultra capacitors, fuel cells, microturbines, free energy systems, events and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908
fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

Institute for Planetary Synthesis

P.O. Box 128, CH-1211 Geneva 20,
Switzerland
Tel. 41-022-733.88.76, Fax 41-022-733.66.49
E-mail: ipsbox@ipsgeneva.com
website: http://www.ipsgeneva.com

The Unified Gravitation Theory

(The unified super-principle, which controls the Universe)

I. P. Kuldoshin

Orenburg, Neftyanikov str., h. 2, apt. 9,
460019, Russia

(Editor's comments by Alexander V. Frolov)

A forum of the leading USA physicians took place in the White House in March 1998 in presence of President Clinton. There was only one question: "When will the nature of Gravitation be opened?"

The well-known USA physician-astrologer S. Hoking declared that it possibly would occur in twenty years and it would be **the Unified Theory of All.** So, the scientific world by default called it the greatest discovery of the future.

Some time later a new hypothesis pretending to this discovery has got its birth in Orenburg. Despite this fact this hypothesis would gain recognition and status of the **Greatest Discovery of Mankind** only by 2018 that was predicted by S. Hoking.

To present day there have been written a lot of hypotheses on this problem but they haven't been recognized. Many scientists consider our Universe as living and functioning according to the unified and rigid laws in Macro and Microworld, which provides automatic regulations of all its processes due to **circulation of radiant energy of the Universe life in cosmic space**. This energy is inexhaustible and environmentally clean, and Mankind may learn using it in the nearest time for the welfare and for prevention of **contradirectional irreversible ecological** catastrophe.

There is no alternative for humankind to escape and it will not appear in the future. Only cosmic energy will save us. From the book "Secret Doctrine" by E.P. Blavatskaya we can get complete information about the fact that a highly developed civilization of Atlases on the Earth had a "**General Theory of All**" yet 10-12 thousand years before our civilization. They had no automobiles, but instead they had flying objects (aircrafts) "Vimana" of various types as well as ships and submarines, on which they also used **Cosmic energy.**

While reading an abstract in General Soviet Encyclopedia, I got acquainted with the theory of "Aether wind", which was abolished in the beginning of the XX century, and then I understood that this theory contains a deposit to discover the nature of gravitation.

The nature of gravitation is the only one and there are no alternatives in theoretical as well as in physical sense. When scientific world of entire planet abolished the theory of "Aether wind", it lost the possibility to discover the nature of gravitation. The XX century was marked by a revolutionary development of scientific and technical progress, but there was an almost 100-year stagnation in cognition of the Universe elements.

The theory of "Aether wind" supposed that all Cosmos is filled with aether particles flying with the speed of light (these particles are "neutrino" according to modern understanding). **The role of gravitation, carrier of light and retarding medium in Cosmos** was attributed to this motion of particles.

But this theory allowed chaotic motion of particles, which is impossible in mechanism of the Universe, which is adjusted up to automatic mode. Besides, motion of these particles is not possible without an absolute buffer unit, which prevents their head-on collision at the speed of 600000 km/sec (it is thermonuclear explosion and death of matter, i.e. the Universe). E.P. Blavatskaya wrote that Cosmos is filled with radiant energy of the Universe life, luminophore, electromagnetic aether. Thereby she predicted a ready solution to make correction in the uncompleted theory of "aether wind". On the basis of above stated and due to the **un-assumed dawning up,** the theory of "aether wind" was completed. It was the ground to develop a hypothesis of radiant "aether wind". Particles of this wind (neutrino) are electromagnetic particles and move with the speed of light in all directions as **contradirectional paired single-stream flows** (like electrical current in twin-wire cable). Due to this, an absolutely stable concentration of these beams in cosmic space is provided according to the principle "what has come in, the same has gone out".

The hypothesis formulates new views on the problem of structure of elements of the Universe material world.

Some separate conclusions do not match the views of modern scientific thought on the problems of physical principles of material world structure and functioning of the Solar system.

List of topics of the hypothesis

1. The hypothesis disclosures the operating environment of a super-mechanism, which controls the Universe (it is a radiant "aether wind").

2. It disclosures the nature of retarding mechanism of flying objects in Cosmos (its name is Lorenz-Fitzgerald compression).

3. It proves the absence of Universal gravity and beams of light as we usually conceive it. (The beams of aether wind collide and compress matter. An alternative to the notion about beams of light is a temperature wave impulse on the beam of aether wind. It explains why the speed of light doesn't depend on the speed of the source of light. Light is a "passenger" on the beam of "aether wind").

4. This hypothesis disclosures the mechanism of stablization of rotary and orbital movement of the Universe matter in macro- and micro world due to retarding medium in Cosmos.

5. It disclosures the mechanism of reverse rotation of Venus due to the forces of autorotation.

6. It disclosures the mechanism of reverse orbital movement of planets and satellites of planets. (Such a planet had not been opened yet, but there are 6 satellites in the Solar system, which move counter to the others, and it is not an occasion, but **a particular case of the effect of aether wind beams**).

7. It disclosures a real nature of Tungusska catastrophe. (There were about 100 hypotheses, but **neither of them** was recognized to be true).

8. It disclosures **the nature of gravitation** and gives an explanation that gravitation can be:
   - usual (vertical);
   - horizontal;
   - circular

It is important to note: not the entire matter takes part in gravitation, but 1/3, i.e. 33,3% of matter.

9. It disclosures the **nature of Levitation** and proves that 1 liter of water on the surface of the Earth can have the weight from 0 up to 3 kg.

10. It disclosures the role of gyroscope effect in life support of the Universe. The gyroscope effect allows transformation of translation energy of radiant "aether wind" to the rotational energy for practical needs of humankind.

11. This hypothesis gives scientific and technical recommendation for creation of cosmic energy converters.

12. It gives scientific and technical recommendation for producing of levitation effects for any technical systems.

13. It disclosures the possibility of cosmic flights with the super-light speed.
14. It explains experiments on metering of horizontal gravitation (The first experiment was made on February 27, 1999).

15. It disclosures the particle (neutrino) of original matter of the Universe and gives its characteristic. (Ancient thinkers called modern "neutrino" as "Aether", and it was not occasionally, because its diameter is in 10^{25} times smaller than atom's diameter.

All matter of the Universe consists of the same indivisible particles "neutrino" presented by three groups:
- "energy" group, which is in the beams of "aether wind";
- building group, which forms the part of any micro particle;
- free group (neutral-reserve) as a building material for new matter and operating environment of all electromagnetic processes.

All neutrino of three groups rotate with the speed of 3×10^{43} rps (equatorial speed of neutrino is equal to the speed of light).

Fields are formed in every particle as a result of rotation:
- strong field of a small volume doesn't allow particles to close up;
- weak field of a big volume is a general mechanism of gravitation.

As scientists write at the present time, the World is subdivided on a dense world (which we can see) and fine world (invisible). At that the density of such world is in 10^{15} times less than density of water.

It is known in science that all matter of the Universe both great and small rotates and is a gyroscope. Particles of matter get rotation with their birth, thus the fields are born in them simultaneously. **Matter cannot exist without rotation, which generates fields.**

All mechanism of interaction between three groups of particles is based on the mutual repulsion. This is the only mechanism, which always and automatically is able to create the necessary stable interval between the particles and only this mechanism provides the function of gravitation.

Many scientists of the late XX came close to the discovery of the nature of gravitation, but they didn't accept a thought to conceive **the motion of aether particles as a pair-counter flow**. And there are three necessary conditions to realize gravitation:

1. The particles should have the fields of repulsion.
2. The contradirectional flows should envelop the particle of matter from two sides.
3. While one beam is passing a matter mass then the force of fields should decrease and gravitation effect should appear.

Mechanism of gravitation

Gravitation appears due to the intersection *(Editor's note: interference)* of fields, produced by beams particles and fields of the visual matter. As it was mentioned above, the beams are paired and contradirectional. Usually the beams in cosmos are mutually balanced and they do not call gravitation effects.

But on the surface of the Earth the contradirectional beams are not similar in their power. The powerful beams come from above, i.e. they only penetrate the atmosphere, and the weakened beams come from below, i.e. they penetrated all the Earth. Thus, gravitation appears.

Gravitation is a unique property of "aether wind" beams to loose part of their power during penetrating of matter mass. **Gravitation** is the difference of forces of contradirectional beams. *(Editor's note: Really other authors reported this idea also. I cannot find who was the first in discussion about gradient of aether as nature of gravitation.)*

Horizontal gravitation

As a particular case, there is **horizontal** gravitation on the surface of the Earth. It appears on the boundary between lowland (of the sea) and plateau. In this case one beam goes above the surface of the Earth (water), and the counter beam penetrates mountain range and weakens. The first measurements of horizontal gravitation effect were made on February 27, 1999 on the route Orenburg – Samara at 49 km before Syrtinskiy slope.

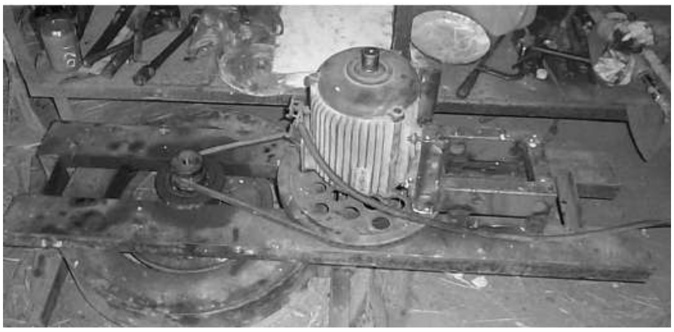

A leaden load (0,5 kg) on the float (a piece of foam plastic) moved on the water surface (not in the sea but in basin) towards the mountain.

Horizontal gravitation is much more weaker than usual gravitation, but it can reach the value that makes water to flow at some angle upwards.

Circular gravitation

Only fast-rotating bodies can create circular gravitation.

All bodies rotate by their orbits around the Sun in the open space of Solar system **due to circular gravitation guided by rotating the Sun.** Furthermore; circular gravitation always is direct (co-directional to the Sun rotation) and reversed gravitation on the periphery of Solar system. A planet with reverse orbital movement had not been discovered until now, but 6 satellites of planets in the Solar system have reversed orbital movement.

Here is the proof of the fact that circular gravitation appears only around the fast-rotating bodies and **slow-rotating bodies, for example, the planets Venus and Mercury cannot form circular gravitation**, that's why they have no satellites.

Our Sun is a prototype of mechanism to transform translation energy of aether wind beams into rotary energy.

(Editor's note: According to Kozyrev, any star is a transformer of time (chronal type of energy) into heat energy. Really, the aether wind can be considered as the chronal type of energy in our understanding and for our usual three-dimensional measurement equipment. To my mind it is a clear link to notion of 4-dimensinal objects, i.e. the time. ***Time can be described by parameters of the aether wind, i.e. its velocity, direction and density. So, we can say that quantitatively time can be described by formulations for kinetic energy of the aether movement. From the other hand it is equivalent of heat energy, which can be measured by usual methods after transformation of the longitudinal waves of the aether in transverse electromagnetic waves).***

Therefore, any mechanical disk rotating very fast will create a circular gravitational field, which is able to rotate all bodies in the direction of the disk (for example, a rim mounted on its bearing co-axially with rotating gyroscope). I designed and tested a similar device in January 2000. A gyroscope (of 200 mm diameter and 3 mm thick) was over-speeded up to 18 thousand rpm. Rotation of gyroscope called slow (but with a good momentum) rotation of the rim of 15 kg weight.

The gyroscopes with the mass of 0,5 kg, 15 kg and 90 kg were tested during summer of 2001. All them called rotation of the rims.

(Editor's note: There are other experimental facts. Fast rotation of mass should produce rotation of some part of nearby aether. Self-closed aether forms vortex and if photon is trapped by this vortex, then

experimenters can see "ring of light" near rotating mass. The rings or self-closed photos can exist in the same place after the mass was stopped or removed away.)

Nowadays gyroscopes in military devices are over-speeded up to hundreds of thousands of rpm. The more rates the gyroscope has, the more energy the rim will produce if it is connected to some generator. But these research works led to single-valued conclusions that gyroscopes themselves cannot produce big quantity of additional cosmic energy not jointly with permanent electromagnets. The Sun as well as planets has natural electromagnetism and their circular gravitation increases in many times due to the presence of electromagnetic fields.

(Editor's note: I think it is obviously that a rotating magnet can involve into the rotation much more quantity of aether than any simple rotating mass. In some theories any magnetic field is considered as circulation of aether particles.)

Electromagnetic fields are the unique boosters of circular gravitation. So, the gyroscopes themselves cannot produce necessary quantity of cosmic energy per unit mass of gyroscope without using of electromagnetism.

In October 2001, I got a copy of 24 patents description. There were patents on "perpetual motion machines". But since such "perpetual motion machines" cannot exist in reality, then we can explain them as gyroscopical transformers of cosmic energy. Efficiency of these transformers varies from 150% up to 10^6 % and practically all of them work using gyroscope. But nowadays only the transformer (Bauman's machine) works in Switzerland, in Maethernitha theological community, Linden city. Some systems have been working from 1980 and producing total power of 750 kWtt, the gyroscopes of 2m diameters are provided with constant magnets.

Besides, there are ready transformers of cosmic energy in Russia. The Professor of Moscow State University, Academician of Russian Academy of Natural Science Leonid Leskov spoke about them in the first half of 2001. He actually said that Mr. Chubais does not allow innovation of energy transformers, which are ready for commercialization (see newspaper "Raduga", Samara, July 2001).

I assume that any kinds of such transformers work on the energy produced by "Aether wind" beams. Perhaps our earth ancestry (Atlases) used this energy to fly as well as extraterrestrials. I remember information about flying platforms, which were designed in Germany in 1943-1945. Nowadays there are publications that there are not less than 10 captured extraterrestrial's spacecrafts on the Earth, and some samples were tested in Russia and the USA.

On the basis of all above-mentioned it becomes extremely clear that the main secret of Nature was discovered, and let's representatives of conservative science don't pull the wool over people's eyes to prove that "it is impossible". It is possible! Physics is an experimental science in its main part, and there is no completed theory until now.

As a result, I'd like to make some conclusions: The secret of Gravitation nature was discovered not in connection with new scientific investigation, but due to dawning up and understanding of the fact that gravitation since earliest times was produced by "Aether winds", which fill all cosmic space. Instead of improvement of "Aether wind" theory, academician science abolished it and forgot it such as some scientists of nowadays don't have an idea of it. While abolishing of "Aether wind" theory, scientific world spent 100 years in vain to find an alternative to it. A real Cosmic scientific and technical progress was slowed down during this term. Without this progress all humankind will kill environment of the Earth in 30-40 years!

Rush hours for humankind to turn to cosmic energy came, we have not even an hour to wait, and otherwise we will loose a chance to survive. Today the scientific and technical level is such that taking into consideration the buildup made by inventors – enthusiasts, who created more than 50 types of Cosmic energy transformers, it is possible to begin repetition work in one year. Now there is the only barrier to do it, i.e. market relations in energetics developed during last 100 years.

Let's look into near Future. The process of energy resources (coal, oil, gas) formation in bowels of the Earth took hundreds million years. There was period of clean ecology in the World Ocean, on land and in the atmosphere. **And all it catastrophically had been diminishing during 2-nd half of the XX century**. **There are about 40 years for our civilization to reach the boundary of having no chance to support normal life on the Earth.** An irreversible process of struggle for survival using underground environment and protection from mortal ecology will begin. Our close posterity will not forgive us this betrayal.

Is there any solution? Yes, there is.

It is necessary to publish the descriptions of all "perpetual motion machines" models as well as unprofitable publication of short technical documentation in the Internet and magazines, which will give a chance to many companies, research groups and individuals to re-produce them. But at first we should choose the models, which are the most reasonable in technology and prime cost. Such a way of replication of the models will give people confidence, interest and reliable information on existence of inexhaustible salutary cosmic energy. And the victory will be the reward for courageous, enterprising and advanced people.

which generates electrical voltage $E_{\updownarrow} = \ell \frac{dB_{rot}}{dt}$ in the structure of vacuum. This voltage generates Gravitational Impulse itself $G = 4\pi E_{\sigma} S \cdot (\Delta r_g)^2$, where $\Delta r_g = e_0 \frac{E_{\updownarrow}}{b}$.

The supposed "Gravitational Impulse" in the experiment by Podkletnov is modeled by a quarter of cosine curve.

Duration of this curve is determined by the decrease of magnetic field "trapped" into the superconductor due to the partial heating of semi-conductor emitter after plasma passed the discharge of 2MV with the current strength of 10000 A. The formula of the model is the following:

$$X'' = Ae^{-2\pi f_0 D_0 t} \cos(2\pi f_0 \sqrt{1-D_0^2}\, t) \qquad (1)$$

the calculation is made for the frequencies of 30, 3, 0,3 and 0,03 Hz and acceleration of 12 m/sec^2, which appears for the mass of the pendulum 30 g with the force horizontally to gravitation 0,03·12=0,36 N.

It can be supposed that it is necessary to make more careful solution of the problem to find the effect on the pendulum by its reaction, which is known from experiment. We should apply the more correct use of spectral method of solution of differential equation for the pendulum with setting of impulse effect. Further, having the recording of temporal function of magnetic field by Hall-effect devices and using Maxwell formulas, we should find electrical field acting in physical vacuum. This field will give us the force of gravitational impulse.

The experiment by V. Roshchin and S. Godin is simpler for physical modeling (Editor's note: the author assumes it is simpler than Podkletnov's effect). All input and output parameters are known to the authors, i.e. force of the magnets, frequency of variable magnetic field in the local place of space vacuum, change of gravity. Furthermore, there are known cylindrical formations of magnetic "loops" around the device and their approximate arrangement with the intervals divisible by the half of rotor radius. Effects of temperature decrease at 8° C in cylindrical atmospheric formations can be simply explained by adiabatic decrease of air pressure due to the decrease of gravitation between molecules of air. Formulas for estimation of decrease of gravitational and inertial forces are the same that for Podkletnov's experiment:

$$E_{\updownarrow} = \ell \frac{dB_{rot}}{dt}, \qquad (2)$$

$$G = 4\pi E_{\sigma} S \cdot (\Delta r_g)^2, \text{ where } \Delta r_g = e_0 \frac{E_{\updownarrow}}{b}. \qquad (3)$$

References

1. Podkletnov E., Giovanni M. Impulse Gravity Generator Based on Charged Superconductor with Composite Crystal Structure.
2. Roshchin V., Godin S. Experimental research on physical effects in dynamic magnetic system // Letters to the Journal of Theoretical Physics, 2000, vol. 26, issue 24.
3. Rykov A.V. Principles of full-scale physics. // Institute of Earth Physics RAS, M.: 2001, 58 p.

New Sources of Energy from the Point of View of Unitary Quantum Theory

L.G. Sapogin,

Department of Physics, Technical University (MADI)
Leningradsky pr. 64, A-319, 125829, Moscow, Russia
E-mail:sapogin@cnf.madi.ru, sapogin@cnf.madi.msk.su

Yu.A. Ryabov,

Department of Mathematics, Technical University (MADI)
Leningradsky pr. 64, A-319, 125829, Moscow, Russia
E-mail:ryabov@vmat.madi.ru

V.V. Graboshnikov

Representative in Moscow of Sceptre Electronics Ltd.
Millennium House Business Center
12, Troubnaya Street, Moscow 103045, Russian Federation.
E-mail: sceptre@orbita.ru

Abstract

The Unitary Quantum Theory (UQT) is a new version of the field quantum theory, which has been developed by the principal author (Prof. L.Sapogin) of this paper for over 25 years. The theory is directly related to the problem of new energy sources, and this paper can be of interest for Journal of New Energy for it is the UQT (and not the classical Newton mechanics or the modern standard quantum mechanics) that provides a theoretical basis for the development of new sources of energy and for the explanation of the operation principles of the existing and functioning over unity devices.

The fundamental provisions of the UQT and a number of results received on the basis of it were published in many scientific journals and reported at international conferences (see [1-6], etc.). Generally, the UQT as expressed by the language of formulae and equations represents a new mathematical model of interaction and movement of elementary particles in the form of a complicated system of non-linear integral-differential equations, an important property of this model principally defines the trajectories and velocities of the particle movement in space (unlike the standard quantum theory, which directly defines only the probabilities of the presence of the particles at a certain point in space). Another, and the most essential (for the problem of new energy sources) property of the UQT is the absence of the energy conservation laws and the impulse for single particles in it. That is why the UQT makes theoretically possible processes of energy generation as if from nothing, if they are regarded from the classical mechanics point of view or the standard quantum theory (while the UQT is able to explain the phenomenon), as well as creation of a device with efficiency above 1. In other words, the UQT provides for a theoretic possibility of making a perpetual mobile!

In the 1970's, when the UQT started to be developed, there was nearly no data of the observed phenomena, or any experimental results confirming this unusual theory. Today, such data are abundant. For example, such processes can be named as generation of excessive heat energy during cavitation of very small water bubbles; generation of excessive electric energy in an anomalous gas discharge; excess generation of electric energy when electric current passes through proton-conducting ceramics, etc. Besides, and still more important, operating devices that have been created much more energy than it was necessary for these devices functioning: electric current generators "Testatica" (Switzerland); thermal cell CETI (J.Patterson, USA); heat generators (Yu.S. Potapov, Moldavia, J.Griggs, USA); electric current generators (P.Correa, A.Correa, Canada); electric engines on magnetic ceramics (Japan), and others. The said phenomena and operation principles of the above-mentioned devices can be explained with the help of the UQT.

In this paper we will also touch upon such an important problem as cold nuclear fusion. The feasibility of this nuclear process, which is categorically denied by the standard quantum theory and nuclear physics specialists, was predicted by the author of the UQT as far back as in 1983. This phenomenon was discovered in 1989 (electrochemical experiments, M.Fleischmann, S.Pons). Many subsequently received experimental data confirmed the existence of nuclear reactions under very small energies, of nuclear transmutations in plants and biological objects, very slightly connected with generation of energy [7-8]. From the point of view of the UQT, which provides an explanation of the cold nuclear fusion mechanism, this process can be applied in practice (after the relevant devices are designed) for generation of energy, for production of isotopes, and for nuclear waste liquidation.

Introduction

Inventors, as well as swindlers of all kinds, had long ago been trying to construct, or at least design, a *perpetual mobile*, i.e. an imaginary machine that produced work without any outside energy. Peter the Great even founded the Imperial Russian Academy of Sciences for such research, but the modern Russian Academy of Sciences does not like to recollect this circumstance. On the other hand, the French Immortals in 1755 decided not to consider any *perpetual mobile* projects at all, and, as we would see, were quite right as regards the Newton mechanics. The brilliant success of the classic mechanics has strengthened still more the sacred belief of the mankind in the Divine Infallibility of the Conservation Laws, and today it is nearly indecent to express any doubts about these laws.

Let us first of all find out the origin of the conservation laws in ordinary mechanics. Practically any textbook will tell you that the Energy Conservation Law (ECL) follows from the homogeneity of time, the Impulse Conservation Law from the homogeneity of space, and the Angular Momentum Conservation Law from the isotropy of space. That is why many people have an impression that the conservation laws themselves follow only from the quality of time and space, which is today an undoubtedly relativistic notion. But, for example, the angular momentum is not a relativistic notion. So, such a narrow approach is not altogether correct, and it is necessary to turn to the second Newton law, or the equation of relativistic dynamics and the system insularity. However, *the qualities of the time and space ensue exactly from the analysis of the Newton mechanics*, though they are often construed incorrectly. Let us remind you the correct interpretation.

Homogeneity of time suggests that *if at any two moments of time two similar experiments are made in similar closed loop systems, the results thereof will not differ.*

Homogeneity and isotropy of space mean that *if a closed loop system is moved from one part of space to another, or is oriented differently, nothing will change.*

The making of the fundamental energy and impulse conservation laws from the Newton equation is very simple. Let us put down the main equation of dynamics as

$$\mathbf{F} = \frac{d\mathbf{P}}{dt}$$

For closed loop system $\mathbf{F}=0$ (no external forces operating) and the equation integral will be $\mathbf{P} = Const$ the impulse conservation law.
Now let us take the main equation of dynamics as:

$$\mathbf{F} = m\mathbf{a} = m\frac{d\mathbf{v}}{dt}$$

and multiply it scalarwise by $\mathbf{v}$

$$\mathbf{F}\cdot\mathbf{v} = m\frac{d\mathbf{v}}{dt}\mathbf{v} = \sum_{i=1}^{3} m\frac{dv_i}{dt}v_i = \sum_{i=1}^{3} m\frac{d}{dt}\left(\frac{v_i}{2}\right) = \frac{d}{dt}\left(\frac{mv^2}{2}\right)$$

where v is the module of the velocity vector $\mathbf{v}$. For closed loop system $\mathbf{F}=0$, and the equation integral will be

$$\frac{mv^2}{2} = Const$$

one of the forms of the energy conservation law. From the definition of the angular momentum for a particle,

$$\mathbf{L} = [\mathbf{r} \times \mathbf{P}]$$

Differentiating both parts by *t,* we get

$$\frac{d\mathbf{L}}{dt} = \left[\frac{d\mathbf{r}}{dt} \times \mathbf{P}\right] + \left[\mathbf{r} \times \frac{d\mathbf{P}}{dt}\right]$$

Since the impulse vector is parallel to the velocity vector, the first bracket will equal to zero. On the basis of the resultant equation and the definition of central force as not creating any momentum, we get

$$[\mathbf{r} \times \mathbf{F}] = 0 \quad \text{or} \quad \mathbf{L} = Const.$$

In case of the central force in an unclosed system, the angular momentum is preserved by value and direction. The angular momentum conservation law for a closed loop system results in the same way as the impulse conservation law from the equation of the rotary motion dynamics:

$$\mathbf{M} = \frac{d\mathbf{L}}{dt}$$

For a closed loop system, the momentum of external forces $\mathbf{M}=0$ and the integral of the equation will be the angular momentum conservation law

$$\mathbf{L} = Const$$

In relativistic dynamics the emergence of the energy and impulse conservation laws separately can be easily received from the relativistic ratio for energy and impulse

$$E^2 = P^2c^2 + m^2c^4$$

Term m^2c^4 is an invariant, i.e. the same in all reference systems. In other words, it is a certain constant. This equation can be represented in a slightly different form

$$E^2 - P^2c^2 = Const$$

For the equation to be valid, it is required that

$$E = Const \text{ and } P = Const$$

and this none the other but conservation laws for energy and impulse.

Strictly, relativistic mechanics has a conservation law for 4-impulse vector P^μ, but we will not dwell on these details, because small energies are what we are interested in.

In the classical theory, the energy conservation law states that the energy of a closed loop system remains unchanged, so, if the energy of such a system is designated at a moment $t=0$ as E_0, and at the moment t as E_t, $E_0 = E_t$.

Conservation laws in ordinary quantum mechanics

The standard quantum theory formulates the energy conservation law in the same way. In quantum mechanics we have the same movement integrals as in classic mechanics. A certain value *L* will be a movement integral, if

$$\frac{d\hat{L}}{dt} = \frac{\partial \hat{L}}{\partial t} + \left[\hat{H}, \hat{L}\right] = 0 \qquad (1)$$

Since $\left[\hat{H}, \hat{L}\right]$ is defined by commutator of operator $\hat{L}$ and of Hamilton's operator, any value *L*, *not depending explicitly on time*, will be the movement integral, if its operator commutes with the Hamilton's operator. When value *L* does not explicitly depend on time, the first item in (1) turns to zero. There remains

$$\frac{d\hat{L}}{dt} = \left[\hat{H}, \hat{L}\right] = 0 \qquad (2)$$

and for the movement integrals not explicitly depending on time the Poisson quantum bracket equals zero. From (1) and (2) it follows that the average value of the movement integrals does not depend on time:

$$\frac{d}{dt}(L) = 0$$

All good papers on the quantum theory prove that probability $w(L_n, t)$ to find at any moment *t* any value of the movement integral, i.e. L_n, does not depend on time. Further, is constructed as the movement integrals not explicitly dependent on time. Since operators $\hat{\mathrm{L}}$ and $\hat{\mathrm{H}}$ commute, they have common proper functions, which are functions of stationary states. Let us note that the latter follows from solutions of the equation without time, which was received from the full equation with imposition of requirement

$$\Psi(\mathbf{r}, t) = \Psi_0(\mathbf{r}) \exp\left(i \frac{E}{t}\right)$$

equivalent to search of only periodic solutions. Further, quite naturally, *there appeared an equation without time with actually imposed conservation laws, because now nothing depends on time*. Expansion by such proper functions looks as follows:

$$\hat{L}\Psi_n = L_n\Psi_n \qquad \hat{H}\Psi_n = E_n\Psi_n$$

where

$$\Psi(x,t) = \sum_n c_n \Psi_n(x) \exp\left(-i \frac{E_n}{\hbar} t\right) = \sum_n c_n(t) \Psi_n(x) \qquad (3)$$

$$c_n(t) = c_n \exp\left(-i\frac{E_n}{\hbar}t\right) = c_n(0)\exp\left(-i\frac{E_n}{\hbar}t\right)$$

Since (3) is expansion into proper functions of operator $L_{n'}$ probability does not depend on time:

$$w(L_n, t) = |c_n(t)|^2 = |c_n(0)|^2 = Const$$

Since energy is a movement integral and probability *w (E,t)* to find at a moment *t* an energy value equaling E, does not depend on time, then:

$$\frac{dw(E,t)}{dt} = 0$$

Let us note once again that it is *the probability to find a certain value that does not depend on time, but not the value itself, which for any separate event is accidental and can assume a wide range of values.*

The quantum energy conservation law in the above form suggests a possibility of defining energy at a given moment without subjecting it to uncontrolled change, which raised no doubts in classic mechanics. But in the quantum theory the energy, without changing its value, can only be measured to

$$\Delta E \geq \frac{\hbar}{\tau}$$

where r - is measurement duration. Formally, it does not present any difficulties for the energy conservation law, since energy is a movement integral, and we have much time to make long measurement. For example, let us make measurements during time r, then leave the system to itself for time *T*, and then define energy again. The classic quantum energy conservation law states that the result of the second measurement will coincide with the result of the first measurement to $\Delta E \approx \frac{\hbar}{\tau}$. But even in the ordinary quantum theory all this is not consistent enough. For the real vacuum fluctuations can interfere, that always influences the results of a single process, but their influence disappears after the passage to an ensemble of events. Here we have a violation of the conservation law due to vacuum fluctuations, though existence of movement integrals, unlike in the Unitary Quantum Theory (UQT).

The generally accepted quantum theory carefully avoids the question of conservation laws for individual events in the case of small energies. This question is either not discussed at all, or it is said that *the quantum theory does not describe individual events. Yes, it does describe individual events, but it can only predict a probability of this or that result.* It is clear that in this case there are no conservation laws for individual events (it is wrong to speak about it in case of an accidental result of an individual event), and they appear only after the averaging by large ensembles of events. Essentially, it can easily be proved that classic mechanics follows from quantum mechanics after summing up by a large number of particles, because for a sufficiently big mass, the length of the de Broglie wave becomes much less than the body dimensions, and no quantum-wave qualities can be talked about.

Conservation laws in Unitary Quantum Theory

In the UQT [1-14] any quantum particle is not a point, but a source of field like in the ordinary quantum mechanics, but it represents a bunched field (wave packet) of a certain unified field. The dispersion equation of such a nonlinear field turned out to be such that the wave packet (particle) during its movement periodically appears and disappears, and the envelope of this process coincides with the de Broglie wave. Numerous particles during their periodic disappearance (spreading in the Universe) are repeated appearance from vacuum fluctuations. A theory of quantum measurements has been built, and the probability interpretation follows from the mathematical formalism of the quantum theory [10,11], and it is not postulated as in conventional quantum mechanics. Unfortunately, the main UQT equation turned out very complicated, for it is a system of 32 nonlinear integral-differential equations, which could require for their solution some new mathematical methods. But from this the relativistically invariant Hamilton-Jacoby equation, and the Dirac equation system strictly follow.

Papers [13,14] give a solution of the simplified scalar integral-differential UQT equation, which gave a localized solution for the form of a wave packet representing a particle. It turned out that the integral from a bilinear combination of such a solution for the whole volume gives with the precision of 0.3% the value of a non-dimensional elementary electric charge [13,14], which was essentially its first theoretical calculation. Then, this solution in the form of a periodically appearing and disappearing wave packet (which square describes the density of a spatial charge) can be replaced by an oscillating charged particle [15-18], the movement whereof will be described by the conventional Newton equations:

$$m\frac{d^2\mathbf{r}}{dt^2} = -2Q\mathbf{GRAD}U(\mathbf{r})\cos^2\left(\frac{mt}{2\hbar}\left(\frac{d\mathbf{r}}{dt}\right)^2 - \frac{m\mathbf{r}}{\hbar}\frac{d\mathbf{r}}{dt} + \varphi_0\right) \quad (4)$$

$$m\frac{d^2\mathbf{r}}{dt^2} = -2Q\mathbf{GRAD}U(\mathbf{r})\cos^2\left(-\frac{m\mathbf{r}}{\hbar}\frac{d\mathbf{r}}{dt} + \varphi_0\right) \quad (5)$$

where *m, Q,* $\mathbf{r}$ - mass, charge, and radius-vector of the particle, $U(\mathbf{r})$ – external potential, φ_0 - initial phase.

Since $E = -\mathrm{GRAD}U$, and a magnetic field also exists, the Lorenz force should also be calculated for $\mathbf{F} = \frac{Q}{c}[\mathbf{v}\times\mathbf{H}]$, but in the electromagnetic wave **E** and **H** are equal, and for small energies value $\frac{v}{c} \to 0$, and force **F** can be ignored. Both these equations produce qualitatively similar results for different problems, but the first non-autonomous equation evidently does not have any movement integrals at all, and any hope for analytical solutions is very unreal. But for the second autonomous equation such hope still exists. Let us note that these equations describe more accurately the experimental results of scattering on the coulomb potential than the classic Rutherford formula! Application of these equations for the tunnel effect and scattering on short potential also produces correct results, but in this case passage through a high barrier (tunnel effect) will be defined by the initial phase. Of greatest interest, however, is the harmonic oscillator problem.

It is possible that a change in the properties of a material point in the process of its movement is just another step in the material point movement theory. In conventional mechanics this idea is not altogether new. There are Meshchersky's equations for bodies with a changing mass, and Tsiolkovsky's equation for a rocket. But so far, in the conventional quantum theory, the particle has a permanent and stable in space and time set of properties, and in the UQT all the parameters of the particle are changed and oscillate during movement.

It should be noted that *Newton did not introduce the notion of a material point at all,* and it would be ridiculous to think that he was not able to have this natural and rather trivial idea. Most probably, and it is not by chance, for today many troubles of the field quantum theory are rooted in the approach to the particle, as to the point, the most vivid example being a large bouquet of divergences. Nevertheless, this approach is very convenient and should only be used correctly. Let us also remember, that in accordance with the Newton corpuscular theory, beams of light were to be regarded as a flow of certain particles. They are emitted by a shining body in all directions and move in an empty space or a homogeneous medium evenly and straight, i.e. in the same way as the ordinary material particles do in the absence of any external or interaction forces. Newton explained reflection and refraction of light beams on the surface of border between two homogeneous mediums by the effect of certain forces on this border, in the direction perpendicular to the surface. These forces changed the normal velocity component, but did not touch upon the tangential one, which allowed to derive the reflection and refraction laws. However, the inability of such a theory to account for the light partial reflection and passage phenomena, as well as the Newton rings (which he himself discovered), led him to bouts (or fits) theory, which is quite modern, although nearly forgotten. Newton believed that for full explanation of all the processes it was necessary to suggest that some light particles could experience reflection bouts, and others – passage bouts. Let us imagine light falling to a flat surface, which is partially, reflects and partially passes. With quantum description of this phenomenon, a particle connected to the falling wave at the time of hitting the surface has a certain probability of passing or being reflected, and Newton simply used the word "bouts" in place of the word "probability".

It is absolutely clear that all descriptions of processes by the equation with an oscillating charge will be an approximation, because it is evident that no movement equations for a material point can describe even the simplest interference processes on a semi-transparent mirror, during which a material particle should be divided in two parts which will later eliminate each other by destructive addition. It is surprising, but the numerical solution of the problem of scattering on a short potential (the Ramsauer effect) for equations (4) and (5) gives the correct diffraction picture! But if we want to describe an individual particle correctly in the conventional quantum mechanics, the picture becomes inexact and purely probabilistic. At every given moment of time a particle can exist in only one of the mutually incoherent states, because one particle cannot move in different directions simultaneously (it cannot have many impulses at the same time). Nevertheless, there seems to exist a whole class of processes, where description with the help of equations (4) and (5) have certain sense. It is well known that in all experiments the local energy and impulse conservation law in individual quantum processes are true only under high-energy values. But under small energy values it is not so, at least because of the ratio of uncertainties and the probabilistic character of all the quantum theory predictions, and the idea of a global, not local ECL, is invisibly present in the quantum mechanics, and is certainly far from new.

In the strict UQT and the quantum measurement theory, a great role belongs to unavoidable vacuum fluctuations. It is clear that these fluctuations are totally unpredictable and non-invariant in relation to space and time translations. The same can be said otherwise: there are no habitual properties of time and space in this theory. Space-time is now not homogeneous and not isotropic. For example, if the system is transferred to a new point in space, or a certain experiment is repeated at another time, at the point where particle parameters are studied, and it interacts with the macro-device, a new value of vacuum fluctuations (different from the previous one) can appear and produce a different result. Of course, all this is only true for small energies and individual events.

Still more destructive is the UQT for the notion of a closed system. For individual events under small energy values this notion is simply unacceptable for the following reason: vacuum fluctuation at the location of the particle (e.g. in a potential pit) can be sharply changed at any moment. It can be caused by different factors – the nature of vacuum fluctuations itself, or the tunnel effect of another random particle.

Sometimes it is stated that conservation laws follow from the Nether theorem, though these results are present in the works by D. Gilbert and F. Klein. For any physical system, the movement equations from which can be received from the variation principle, each one-parametric continuous transformation that leaves the variation functional invariant, corresponds to one differential conservation law, and there exists a clearly conserved value. It is easy to see, however, that vacuum fluctuations imposed on the varied function (integral of Lagrangian) do not in sum remain unchanged during parametric transformations (at least today it seems so), and this consideration does not work without preliminary of ensemble.

And now we are in for a little philosophy. The local Energy Conservation Law (ECL) in individual processes follows from the Newton equations for closed systems. It would be naive to think that its local formulation will be preserved forever, and would be a bad mistake to transfer the ECL from the Newton mechanics to the quantum processes without any changes, because the latter are more fundamental.

References to the first principle of thermodynamics are, strictly speaking, groundless, because this principle is a postulate. For example, well-known Russian mathematician N. Luzin, in a letter to an inventor wrote that the first principle of thermodynamics is the result of unsuccessful attempts of the mankind at building a *perpetual mobile*, and strictly follows from nothing. *Today it may be said with a great degree of certainty that no sophisticated machine in the framework of the Newton mechanics can be a perpetual mobile,* and the decree of the French Academy of 1755 not to consider any *perpetual mobile* projects is still valid. We will only add that now it is valid only for those projects that are based exclusively on the Newton mechanics.

There is the tendency in modern physics to reduce ECL, especially in theory, to the rank of a secondary derivation from the movement equations (movement integrals). Some physicists restrict the ECL to the framework of the first principle of thermodynamics, others, like D. Blokhintsev [37], think it quite probable that with the development of a new theory the form of the ECL will undergo certain changes. F. Engels wrote in his "Dialectics of Nature": "…none of the physicists actually regard the ECL as an eternal and absolute law of nature, a law of spontaneous transformation of the movement forms of the matter and the quantitative constancy of this movement in all its transformations". But many people do not share this opinion. M. Bronstein in his book "Structure of Matter" wrote: "The ECL is one of the main laws of the Newton mechanics. Nevertheless, Newton did not ascribe to this law the general character that this law actually possesses. The reason for this *erroneous* (italicized by authors) opinion of Newton of the ECL is very interesting…". It is now clear that in view of the above, such an opinion was not at all erroneous. Let us remind you that Newton predicted many things, even the UQT, in his "bouts theory".

On the other hand, the authors of quantum mechanics realized that there was no conservation law for single quantum processes under small energies at all. The idea that the construction of ECL, together with the second law of thermodynamics, was a statistical law, true only on the average and inapplicable to individual processes with small energy, first occurred to Schroedinger, and later to Bohr, Kramers, Slater, and Gamov. In 1923 Bohr, Kramers and Slater made a desperate attempt to develop the theory, where the energy and impulse conservation laws in case of scattering would be true only statistically, on an average for long periods of time, but would be inapplicable to elementary events. Lev Landau even called it "Bohr's wonderful idea".

Later, however, the authors gave up this approach and, besides, this idea at that time did not follow from the quantum theory equations, and the authors, *to come out of the predicament, simply declared that quantum mechanics did not describe individual events at all*. Thus, the most vivid paradox of the quantum science was removed by a simple ban on thinking about it! But the ingenious idea that conservation laws do not apply to individual quantum processes and emerge only after the averaging by the ensemble of particles remains alive. This idea might have been a little premature, and, possibly, should be a little different.

The Unitary Quantum Theory (UQT), on the contrary, individual particles, and the difference in their behavior is accounted for by the initial phase of the wave function. In this case, local conservation laws do not exist for a single particle, and measuring the initial phase or some other parameters for an individual particle is quite a different matter. It is not true that the UQT has given up probabilistic description. Probabilistic interpretation remains, but the probability now is strongly dependent on the initial phase. Although the equations with an oscillating charge can determinately predict a particle's behavior, the measurements can be made only with the help of a macro device, which will give only a probabilistic result. Impossibility to determinate measurements does not change anything, for the UQT provides for a possibility of influencing the probability value, which was earlier unavailable. The existing Von Neumann theorem about hidden parameters does not effect our result, but the relevant discussion is too cumbersome, and we will leave it out.

*In other words, **all the requirements**, wherefrom the classical conservation laws follow, are now absent. We can hardly expect the conservation laws for individual particles to be preserved under small energies in such a situation.* Today we are convinced that the classical energy, impulse and angular momentum conservation laws for individual quantum objects are not valid under small energy values because of periodic appearances and disappearances of the particle. All direct experimental tests of the conservation laws were made for large energy values, and for small energies of an individual particle only probabilistic results can be received, and, in this case, it would be indecent even to recall the conservation law.

Energy generation and perpetual mobile

Let us make the following imaginary experiment. For simplification purposes we will use in our reasoning a certain quantum ball-particle. When a classical ball approaches a wall (perpendicularly for simplification), the speed of the reflected ball is always equal to the initial speed (we ignore friction and regard the ball and the walls as absolutely elastic). In the case of a quantum ball, the speed of the reflected ball will acquire in different experiments with absolutely equal initial conditions a whole range of values: some balls will be reflected at a speed greater than the initial speed, others – at a speed equal or lower than the initial speed, and all this is described by quantum mechanics.

Let us ask the following question: what if a second wall is found, parallel to the first one, in order for the ball to increase its speed after each reflection from the wall? Then we will have increased ball energy without any special efforts on our part. Such phenomena appear in the problem of particle oscillations in a potential pit (not necessarily parabolic) on the basis of equations (8) and (9), when four types of solutions are possible, three of which are most important for us: stationary, "maternity home", and "crematorium". In the two latter solutions traditional conservation laws do not work. These solutions are presented in Fig. 1. Such oscillator behavior explains many experimental facts. From the physical point of view, it means that in stationary solutions with fixed discrete energies (conventional quantum mechanics) the speed of the particle reflected from the wall will be equal to the speed of the falling particle. If the speed of the particle is decreased after each reflection, it will mean the "crematorium" solution, and if it increases, the "maternity home". Scenarios for situations will depend on the initial phase of the wave function and the particle energy. In ordinary situations the "crematorium" and "maternity home" solutions always compensate each other, and we find conservation laws.

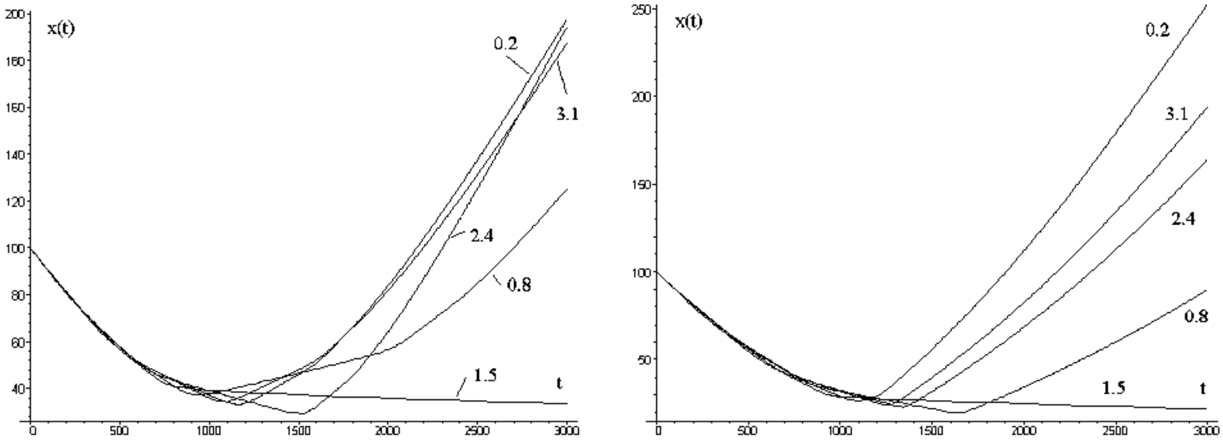

Fig. 1. Dependence of the distance between the moving charge and the nucleus on time for autonomous and non-autonomous equations.

The task of the future developers of new energy systems of the 21st century will consist in creating such initial conditions for a great number of particles making up a body that only the "maternity home" solution would be realized, and the "crematorium" solution would, if possible, be suppressed.

It follows from the above that if the unitary quantum theory ideas are applied correctly, there is no fundamental taboo for a perpetual mobile. Such a taboo, as it was shown, does not formally exist even in conventional quantum mechanics (no conservation laws for individual processes with small energies), and, in order to generate energy, they should be somehow accumulated (all random processes with excess energy should be grouped together). But conventional quantum mechanics refuses to describe individual events and is unable to offer any ways for such grouping. The unitary quantum theory seems to offer such an opportunity.

However, the great idea of free energy generation was distorted by effort of some research associations interested to keep their stability degree, that everyone who started speaking about it was considered to be a crazy man.

Modern experimental physics has verified the correctness of conservation laws either for very large energies in individual quantum events, or for big macroobjects, when automatic averaging by ensemble is made, but the area of very small energies for individual events today is a *terra incognita*.

In order to see how the conservation laws for reflection (repulsion) of an individual particle from the Coulomb heavy nucleus with different values of the initial phase are violated, we have solved numerically one-dimensional equations (8) and (9) under the different initial conditions:

$$\hbar = 1, m = 1,\ 2Zze^2 = 1,\ x_0 = 100,\ V_{x0} = -0.1$$

In the Fig. 2 the distances between the moving particle and repulsive nucleus are shown as a time function, for different initial phases in cases of non-autonomous and autonomous equations.

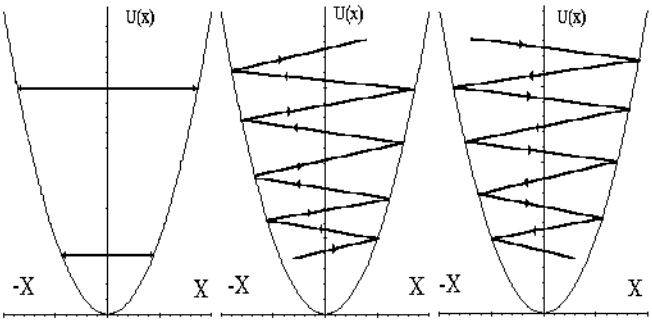

Conventional solution Mathemity home solutions Crematorium solutions

Fig.2 Three types of solutions for oscillator

It is evident from the calculations that the speed of a reflected particle can be equal, lower or higher than the speed of a falling particle. This situation seems to be true for all potentials.

Calculations were also made for other potentials: harmonic oscillator, Yukawa, Gauss, dipole, hyperbolic secant, and Wood-Saxon, and the quality results were nearly the same. If we sum up the impulse of all the particles falling with different phases and compare it with the summarized impulse of all the reflected particles, the summarized reflected impulse, for example, for the Coulomb potential, will be several percent higher than the summarized impulse of the falling particles. For other potentials such a small deviation can even be in the opposite direction. On the whole, this problem is very complicated and requires additional research, because all this is also dependent in quite a complex way on the initial conditions (initial speed, phase and distance).

Philosophically, any categorical taboos, like the impossibility of creating a *perpetual mobile*, are absolutely unacceptable. If everyone is convinced of it forever, the conservation laws and *perpetual mobile* taboos will remain unshakable as long as the human civilization exists. Of course, the funeral of the Conservation Laws can be very prolonged. Anyway, we are not going to do it, and our article might be just a cleanup for the future tomb, and the splendid funeral with all the necessary honors will be organized by future generations. *On the other hand, these laws will never die out completely and will surely be applied, but such spheres of science and technology will appear, though small at first, where these laws are not valid.*

The truth should be accepted irrespective of where it comes from. Words of F. Engels from the “Dialectics of Nature” will be quite appropriate here: “When the solar system ends its life circle and shares the fate of all the finite things, when it falls victim to death, what will happen next? Thus, we come to a conclusion that the heat emitted into the universe should have an opportunity, in a way yet to be established by the natural sciences, to turn into another form of movement, where it can be accumulated again and start functioning. And in this casethe main problem that prevented transformation of dead suns , back to red-hot nebulas will disappear”.

The question of whether the conservation law exists in global form (we have already proved its not being local) remains open, because nothing leads to it except the inertia of the human mind. This inertia was based on the Newton laws, which were replaced by quantum laws. This mental inertia leads to a situation, when in case excessive energy is generated during solution of movement equations, a question arises how it can happen, and where it comes from. Of course, if a particle (e.g. a photon) falls on a semitransparent mirror, the packet is divided into two halves, which, due to imposition of vacuum fluctuations, will be recorded by photomultipliers as full-fledged photons [1-5]. The result is that energy is taken as if from vacuum: two photons appear in place of one. Another photon can be divided on the mirror into two halves, but they will not be recorded by the meters, and the energy will allegedly pass into vacuum. So, at one time we borrowed energy from vacuum, and then gave the same amount of energy back to vacuum at another place. You can think like that, and this process might take place. But if we consider the equation with an oscillating charge, the energy and impulse conservation laws are not valid there for solution of the movement problem, and vacuum fluctuations have nothing to do with it at all. As for the question of where the energy comes from, it is the result of our mental inertia, and is, essentially, an atavism imposed by the Newton mechanics. But the latter appears as a result of an extreme passage from quantum mechanics, which is more fundamental.

It is interesting to note that there is a bomb in the logical definition itself of the energy conservation law. If energy is something that cannot appear or disappear and always simply passes from one form to another, the only value satisfying this condition is zero. We are far from assuming that energy does not exist. But the problem of existence is solved differently in different philosophical systems, and the mathematical approach seems to be the most correct one: *an object exists if it is free from contradictions*. Energy has bad luck in this case, for under such an approach it should be zero.

Some cosmologists (for example, British prof. Fred Hoyle) are very willing to have a process, in accordance with which the Universe has certain places where energy appears from other certain places, in which it is eliminated. Besides, any philosopher at least a little bit familiar with astronomy, looking at the bright night sky, will see the birth of matter and its expansion into a still greater space. But for this purpose the Global Energy Conservation Law is superfluous and only denies what is observed. The head reels…

Cold Nuclear Fusion and Nuclear Transmutation.

Let us approach the epoch-making experiments made by Fleischmann and Pons in March 1989 [30] from the positions of the equation with an oscillating charge. One of the authors predicted in 1983 [9] the possibility of such nuclear reactions under very small energies. Without going into well-known details, we will sum it up very briefly: cold nuclear fusion exists, and there are no people or theories capable of giving a clear explanation. The chain of various mechanisms meant to explain this intriguing phenomenon is growing, but few really believe in them. The reason is as follows.

When a charged particle interacts with the nucleus, the potential energy is like in Fig. 3, where the right top part of the curve is conditioned by mutual Coulomb repulsion between the nucleus and the charged particle.

The repulsion potential will be

$$U(r)=\frac{Zze^2}{r}$$

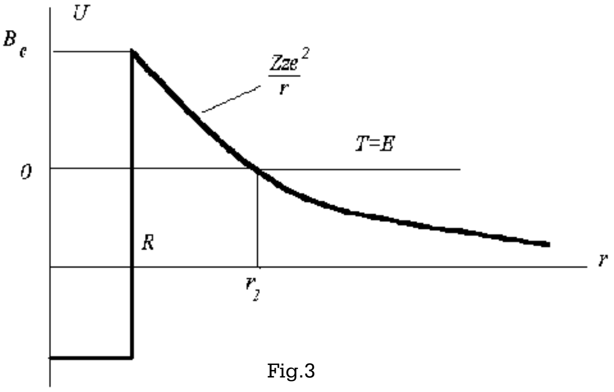

Fig.3

where Z – charge of the nucleus, and z – charge of the approaching particle, e – charge of the electron, r – distance between the particle and nucleus. When $r=R$ (critical distance), then the potential energy curve goes sharply down, which is due to the emergence of intense nuclear gravitation, the potential whereof today appears more complex than could be imagined mathematically. If the charged particle overcomes the Coulomb barrier with a height of

$$B_c=\frac{Zze^2}{R}\approx\frac{Zz}{\sqrt[3]{A}}MeV$$

it will further get into the nuclear gravitation area and a nuclear reaction will take place.

Let us look at the nuclear interaction of a charged particle with kinetic energy $T<B_c$. From the point of view of classical mechanics, there will be no nuclear reaction in this case, because the particle will approach the nucleus and at a certain distance $r<R$ from the top of the Coulomb barrier will turn back and be reflected from it. However, from the point of view of quantum mechanics, there exists a tunnel effect, and the probability of such a tunnel passage, or transparency of potential barrier D is described by a well-known formula:

$$D\approx\exp\left(-\frac{2}{\hbar}\int_{r_1}^{r_2}\sqrt{2\mu(U-T)}dr\right) \qquad (6)$$

where $\mu=\frac{Mm}{M+m}$ is reduced mass. The bottom integration limit r_1 coincides with the nucleus radius R, and the top limit r_2 can be found from the condition $T=\frac{Zze^2}{r_2}$. After integration we will get

$$D=\exp(-2g\gamma)$$

where $g=\frac{R}{\lambda_{B_c}}$; $\gamma=\sqrt{\frac{B_c}{T}}\arccos\left(\sqrt{\frac{T}{B_c}}\right)-\sqrt{1-\frac{T}{B_c}}$,

and value $\lambda_{B_c}=\frac{h}{\sqrt{2mB_c}}$, the de Broglie wavelength corresponding to the kinetic energy of the particle equal to the barrier height $T=B_c$. If $T<<B_c$, expression (6) is easily transformed to look as

$$D=\exp\left(-\frac{2\pi RB_c}{\hbar v}\right)=\exp\left(-\frac{2\pi Zze^2}{\hbar v}\right) \qquad (7)$$

where v is velocity.

Let us now see what the shocking cold nuclear fusion will look like on the basis of the above considerations. The deuteron energy in an ordinary electrolytic Fleischmann-Pons cell will be about 0.025 *eV*, and the height of the Coulomb barrier for this case is $B_c=\frac{Zze^2}{\sqrt[3]{A}}=0.8MeV$. In classical mechanics it would be just naive to talk about overcoming such a barrier with a height dozens times greater than the kinetic energy. Let us now see how the tunnel effect will improve the situation. Let us assess the value of g and γ for the case of collision between two deuterons with such energy:

$$g=\frac{R\sqrt{2mB_c}}{\hbar}=1.9;$$

$\gamma=\sqrt{\frac{B_c}{T}}\arccos\left(\sqrt{\frac{T}{B_c}}\right)-\sqrt{1-\frac{T}{B_c}}\approx 8883$ and the probability of such a process will be $\exp(-2\cdot 1.9\cdot 8883)\approx 10^{-7328}$, i.e. practically pure zero. The fusion cross-section will be defined by the product of nuclear cross-section and the tunneling probability:

$$\sigma=\sigma_{nucl}D$$

and, in the case under review, is also a very small value. If the clash parameter of deuterons is not zero, the emergence of centrifugal potential

$$U = \frac{\hbar^2 l(l+1)}{2mr^2}$$

will still further lower the probability of such interaction.

It is these very circumstances that make the nuclear physics scientist think that there is no cold nuclear fusion as such. For example, such a serious and responsible edition as Encyclopedia Britannica 2001 found no place for the notion of cold nuclear fusion at all. Such an official position can be understandable only from the point of view that quantum mechanics is absolutely true and unshakable. Despite this, for the 12 years since the Fleischmann-Pons experimental discovery, nearly 30 international conferences have been devoted to this subject, there are lots of books and magazines on this subject, and the number of articles on the problem is nearing ten thousand. Today the situation is gradually developing in the positive direction, and the research in the field of hot nuclear fusion, which has already wasted over $90 billion for 45 years, is slowly coming to naught.

But today there exist well known experimental data on cold nuclear fusion. They are numerous and various. We will dwell only upon the most important and sufficiently reliable results. Thus, the classical view of electrolysis of a palladium cathode saturated with heavy hydrogen in heavy water identifies an anomalous quantity of heat energy up to 3 kWt/cm^3, or up to 200 Mj per small sample. Products of nuclear reaction have also been found: tritium (10^7 - 10^9t/s), neutrons with energy of 2.5 MeV (10-100n/s), and helium. Absence of He^3 among the reaction products shows that heat is not generated by reaction $d+p$. Besides, emission of charged particles ***(p, d, t, γ)*** is observed. Similar processes are observed in case of a gas discharge on a palladium cathode, of phase passage in different crystals saturated with heavy hydrogen, irradiation of deuterium mixture with a powerful sound or ultrasound flow, in cavitating microbubbles in heavy water, in a tube with palladium powder saturated with heavy hydrogen under a pressure of 10-15 atm., etc. In certain reactions (e.g. $d + t \rightarrow \alpha + p$) neutrons of 14 MeV are absent, and such a strange situation occurs in other cases too. Activity of Li^6 , Li^7 in reactions with heavy hydrogen and protons failed to be discovered, whereas reaction

$$K^{39} + p \rightarrow Ca^{40}$$

was well recorded even in biological objects. But the most intriguing fact of all these processes is the shortage of nuclear reaction products for explanation of the emerging heat effects. Thus, in certain cases the number of nuclear reaction products (tritium, helium, neutrons, quanta) should be millions of times greater in order to account in some way for the quantity of the generated heat. Generation of such a big amount of energy cannot be accounted for by either chemical or nuclear reactions, or by phase passages. The well-known interaction $d+d$ goes along three channels:

D + D —> T (1.01) + p (3.03) *(Channel 1)*
D + D —> He (0.82)+ n(2.45) *(Channel 2)*
D + D —> He + γ(5.5). *(Channel 3)*

All these reactions are exothermal. The third channel has a very small probability. It was experimentally discovered that they could occur under very small energies. In a molecule D_2 the equilibrium position between atoms is 0.74A and in accordance with the conventional quantum theory, these two deuterons could accidentally enter a nuclear fusion reaction. But the interaction value is very small $\lambda_{D_2} = 10^{-64} c^{-1}$. There is a known estimate that in the water of all seas and oceans there are 10^{43} deuterons, and in 10^{14} years there will be only one fusion.

It follows from the aforesaid that the main problem impeding the occurrence of the $d+d$ reaction lies in the existence of a very high Coulomb barrier. Our approach allows for this problem to be solved, and there is such an opportunity in the UQT. The UQT equation solutions show that the distance to which deuterons can approach each other is strongly dependent on the phase of the wave function (by the way, it is absolutely clear intuitively).

Let us consider the one-dimensional problem [15-18,31]. There is a stationary nucleus with charge *Ze* at the point of origin, and another nucleus is approaching it along axis x (charge *ze*, mass *m*) at a certain initial velocity. The non-autonomous and autonomous equations of such a problem will look as follows:

$$m\frac{d^2x}{dt^2} = -\frac{2Zze^2}{x^2}\cos^2\left(\frac{m}{2\hbar}\left(\frac{dx}{dt}\right)^2 t - \frac{m}{\hbar}\frac{dx}{dt}x + \varphi_0\right) \quad (8)$$

$$m\frac{d^2x}{dt^2} = -\frac{2Zze^2}{x^2}\cos^2\left(-\frac{m}{\hbar}\frac{dx}{dt}x + \varphi_0\right) \quad (9)$$

Since an analytical solution was not found for all the areas of initial phases , numerical methods were applied with the following initial values: $Z=z=1$, $e=1$, $m=1$, $x_0=-10$, $\hbar=1$ for different initial velocities and initial phase values. As had been expected, braking or acceleration of the particle happens only when the charge is large. But at the last stage, under certain initial phases close to $\frac{\pi}{2}$, a wonderful process occurs: velocity, charge and repulsing force are very small. Due to phase ratios, the small charge is not changed for a long time, which means that the particle (or rather what is left of it) is not influenced by any forces, and it is crawling at a permanent small speed for a very long time ("the snail effect") inside the field of another particle, and can come very close to the center. Such movement with a very small charge and a small speed can last for several hours, and disconnection of the external field will not effect this movement. This process reminds of quiet and invisible scout penetration into the enemy territory. This phenomenon occurs only in certain phase areas, and can be conveniently called a phase hole, which is illustrated in Fig. 4 resulting from the solution of equation (8).

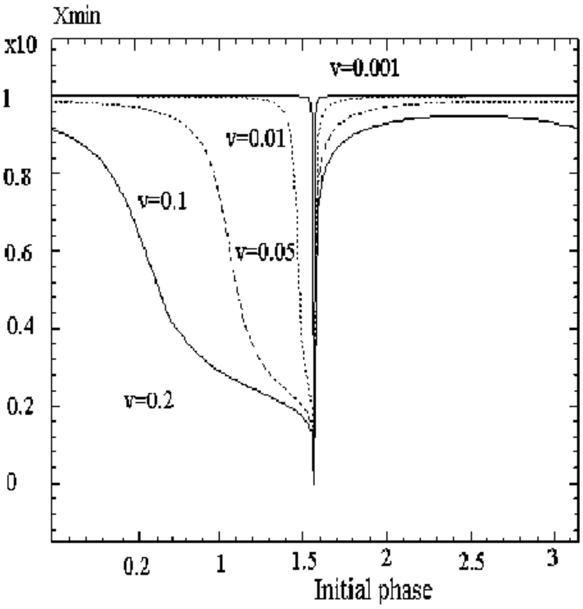

Fig.4. Distance to the turning point of the moving charge depending on the initial phase value for different initial velocities.

Let us note in passing that now we can account for one of the nuclear physics anomalies, which has a tendency to be totally ignored. Under a nucleon energy of 1 MeV, its velocity is 10^9 cm/s, nuclear radius is 10^{-12} cm, and the passage time of the nucleus is 10^{-21}s, but the time period in which the nucleon passes is usually anomalously long - 10^{-14} and even more, and it is absolutely unclear what the nucleon is doing in the nucleus so long. In our model it is easily explained by the “snail effect”.

For the same equation, the minimum distance was calculated between the charges dependent on velocity (Fig. 5) for different initial phase values. For comparison, Fig. 5 also shows the result of the classical calculation based on the Coulomb law. It is obvious from Fig. 4 and 5 that the minimum distance to which charges can approach each other is nearly independent of kinetic energy, but with reduction of speed the initial phase area width is reduced as well. In other words, reduced energy brings also reduced probability of a nuclear reaction.

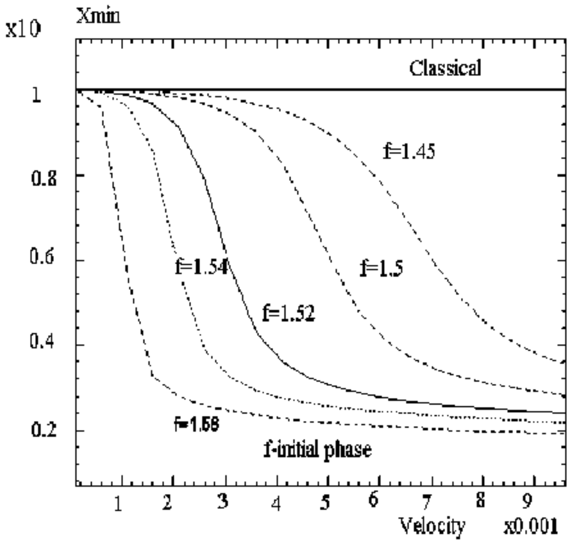

Fig.5 Minimum distance between charges depending on the initial velocity for different initial phase values.

The same results are true for the autonomous equation (9). Under the conventional quantum theory, the ratio of the reaction speeds in the tritium and neutron channels should be close to unity: $\frac{t}{n}=1$. But in many experiments on cold nuclear fusion this value is very different from unity and equals $\frac{t}{n}=10^9$. In different experimental situations it is reproduced by different experimental groups with a very high accuracy. This very intriguing problem has so far received no simple explanation. Let us dwell on a possible cause for such aphenomenon. With a small velocity in the phase hole, neutrons are affected by nuclear gravitation forces, and protons are affected by electrostatic repulsion forces. Under the effect of this momentum, the deuteron will have enough time to turn in such a way that the neutron parts of the deuteron would faceeach other. After the neutron gravitation the nuclear forces will be saturated, which will weaken the proton connection, and one of the protons will leave the system. This reaction can be conditionally presented in the following form:

$$d+d \to p+\left(n+d\right)\to p+t$$

This reminds of the Oppenheimer-Phillips effect.

It is well known, however, that under big energies, the probabilities of the first and second reaction channels are the same, and this phenomenon should somehow be accounted for. Increased probability of the neutron channel with growing energy can be connected with the appearance of secondary neutrons in the reaction *T + D = He + n (14.1 MeV).* In a deuterium-rich environment, a big part of the resultant tritons will pass to neutrons in the process of this reaction, which has a cross-section of 5 barns under an energy value of 70 KeV. According to assessments in [32], the number of such secondary neutrons per one triton is $7.9\cdot10^{-12},1.7\cdot10^{-9},2.7\cdot10^{-6}$ for energy tritons 10, 20 and 100 KeV respectively. Thus, the prevalence of $\frac{t}{n}>10^6$ can be expected only in those reactions, where tritium is born with energies over 40 KeV.

It should not be assumed, however, that the phase hole phenomenon in its whole area leads to a nuclear reaction. It can be assumed that reduction of the Coulomb repulsion is followed by reduction of strong interaction. But how? Today nobody knows the exact equation of the strong interaction potential. Besides, the particle approaches the turning point $X_{\min}$ is rather "thin". Will it be able to take part in a full-fledged nuclear reaction, or will it fly through, like it happens with the electron in the s – states of the atom? There are very narrow phase areas, when soon after the particle stops the charge grows quickly and is sharply accelerated. The charge can even be maximum in the nuclear force effect area. May be, it is this narrow phase area that is responsible for cold nuclear fusion, and in case of strong interactions the phase hole mechanism must be operating as well.

It was discovered long ago that nuclear transmutations have a mass character (especially in plants and biological objects), but they have little to do with energy generation. Examples of such reactions:

$$Mn^{55} + p \rightarrow Fe^{56} \ ; \ Al^{27} + p \rightarrow Si^{28} ;$$

$$P^{31} + p \rightarrow S^{32} ; K^{39} + p \rightarrow Ca^{40}$$

In reactions of this type, a very slow proton (with practically zero kinetic energy) penetrates the nucleus in the above-mentioned way and remains there. *No intranuclear energy is generated, because both before and after the reaction the nucleus remains a stable object.* In classical nuclear physics, the nucleus usually became unstable after it was penetrated by a charged nucleon with a large kinetic energy and always broke into parts, and the nuclear debris had an even greater kinetic energy. Reactions of the above type were considered impossible under small energies and for this reasons were not studied by classical nuclear physics. It seems to be *a completely new type of nuclear transmutations, not recognized by modern nuclear science, but experimentally discovered rather long ago.* Today there is a great deal of experimental material confirming mass nuclear transmutation phenomena. Moreover, there are many projects of neutralizing nuclear excess with the help of this technology. Journals *Infinite Energy, New Energy, Cold Fusion, Fusion Facts*, etc. and Internet are full of such projects.

Of course, a change in the nuclear charge will result in restructuring of electronic atom shells, but the energy related to this process will be about several electron-volts and is nothing in comparison with the energies of nuclear reactions from several to hundreds million electron-volts. By the way, nuclear engineers are accustomed to such energy ranges in nuclear reactions. It was this circumstance that made them deny *a priori* all nuclear processes in biology, because under such energy values of the debris dozens and hundreds thousand of complex biological molecules will be destroyed.

Quite a long time ago, Lois C. Kervran [33] wrote a book about nuclear transmutations in biology, and now, nearly 20 years later, its second edition was published! It gives, evidently for the first time, numerous *experimental* data of the above-mentioned phenomena. The reaction of the official science is very interesting. For example, well-known physicist Karl Sagan, after reading the book about such *experimental* data, advised Kervran to read elementary textbooks in nuclear physics!

Some time later a research was made by Panos T. Pappas [58], who studied one of the well-observed nuclear reactions in biological cells:

$$Na_{23}^{11} + O_{16}^{8} = K_{39}^{19}$$

Classical biology has long known about the existence of equilibrium, when the ratio between the number of *K* and *Na* ions is maintained with greatest accuracy despite the shortage or even absence of K ions in food. Later, in work [59] this nuclear reaction was even called the life equation, and the existence of such nuclear reactions in biological objects was proved by M.Sue Benford with the help of direct physical methods.

All thermonuclear fusion programs are based on blunt heating and compression of the reacting material. Despite the progress achieved, the head of the works in England, Dr. Alan Gibson [34], established several years ago that the model reactor design would be created not earlier than in 50 years. Today, this point of view is generally accepted. Even if the reactor is once made (although the authors have grave doubts about it), it will be very complicated, expensive, and harmful for the environment.

Classical approaches have so far not yielded any positive results, despite multi-billion investments and a great number of physicists, engineers, service personnel, and managers involved. It is only natural that this army of researchers is a potential impediment for all alternative projects of new power engineering. It has been noted that *"viability" of any idea is proportionate to the number of people involved and investments made.* For these reasons, the Fleischmann-Pons works were given a hostile reception in the USA and other countries.

All the controlled thermonuclear fusion programs are accompanied with the adjective "controlled", although there is no control whatever. It is simply that the initial quantity of the reacting substance is *prudently* made very small. For example, a ball of lithium deuteride during laser reduction has a diameter of several mm. So far no one has been seriously considering the question of utilizing the energy of an explosion of such a ball, which is approximately equal to the energy of an explosion of a box of antitank grenades.

The straightforward approach to fusion used by the modern science is very natural, because quantum mechanics has no methods of influencing this process. The future of really controlled nuclear fusion systems may be not on the way of primitive and blunt method of heating and compression of the material, but following UQT on the way of using collisions of nuclei having small energy and corresponding fine adjustment of the wave function phase.

It is essentially possible in case of imposition of the external controlling electromagnetic field on the reacting system, which contains quasi-fixed ordered deuterium atoms and free deuterons. The same properties can be demonstrated by special atomic grid geometry. Diffraction scattering of a deuteron flow on such grids will lead to automatic deuteron selection by energies and phases.

It seems that in the Fleischmann-Pons electrochemical experiments such an ordered system existed in the *Pd-D* grid, and some phasing occurred, which accounts for the results of these experiments [30].

Today it appear to us that the cold nuclear fusion processes will be effectively used for nuclear waste liquidation and production of isotopes.

Many researchers [35,36] discovered that the quantity of heat generated in the process of electrolysis of ordinary water on nickel electrodes (there is no hope for nuclear reactions in such systems) is the same as in the electrolytic cell with heavy water. It confirms other measurements, which showed that the quantity of nuclear reaction products is millions times less than is required for such an amount of generated heat, and its origin remains a mystery.

Further we will give certain concrete data demonstrating the phase values of a deuteron with an oscillating charge, under which the deuteron can approach the nucleus to a critical distance of 10^{-12} cm or less, i.e. giving the data to estimate the value of the above-mentioned phase "hole" in the interval $(0,\pi)$ of the phase change.

Assume that the stationary nucleus with the charge q is placed to the coordinate origin x=0 and the deuteron with the same charge q is placed at the initial moment t=0 to the point $x_0 < 0$on the x-axis, and the deuteron velocity equals $\dot{x}_0 = v_0 > 0$. The units of mass, length and time are chosen in such a way that $m = 1, \hbar = 1, c = 1$ (*m* - deuteron mass, *c* - light velocity). Charge *q* equals 0.085137266. Our units are connected (to 4 significant figures) with the system (kg, m, s) as follows:

1 mass unit = 3.345×10^{-27}kg,
1 length unit = $1.049 \cdot 10^{-16}$ m,
1 time unit = $3.502 \cdot 10^{-25}$ s.

The electron velocity corresponding to its energy of 1eV equals $5.931 \cdot 10^{7}$ cm/s. The deuteron velocity corresponding to such energy will be assumed to be 3680 times less, and in our units it will be $5.372 \cdot 10^{-7}$ (if $c = = 3 \cdot 10^{10}$ cm/s). Then the deuteron movement towards the nucleus is described by the equation

$$\ddot{x} = -\frac{2q^2}{x^2}\cos^2\left(\frac{1}{2}(t+t_*)\dot{x}^2 + x\dot{x} + \varphi_0\right), \qquad (10)$$

where the parameter t_* is defined under the condition that the argument of cosine equals φ_0 for $t = 0, x = x_0, \dot{x} = \dot{x}_0$ (thus $t_* = -(2x_0)/\dot{x}_0$), and this parameter may be considered as the initial moment of so called local time.

In the interest for us are namely solutions of eq. (10) under very small deviation ε from phase φ_0 and so we put $\varphi_0 = \frac{\pi}{2} + \varepsilon$ and rewrite eq.(10) in the following form:

$$\ddot{x} = -\frac{a}{x^2}\sin^2\left(\frac{1}{2}(t+t_*)\dot{x}^2 + x\dot{x} + \varepsilon\right), \qquad (11)$$

where *a* = 0.0144967 Let the initial x_0 to be equal - 500000 of our length units (i.e. approximately $5 \bullet 10^{-9}$ cm) and the initial deuteron velocity v_0 to be equal to the velocity v_{00} corresponding to the deuteron energy of 1 eV or less. But it turned out that the precision of numerical integration of this equation under such initial conditions and under values $|\varepsilon| = 10^{-15}$and less is small and besides the interval of the integration must be very large. That is why this equation also had to be transformed by passing to "slow" time $\tau = |\varepsilon| t$ to the equation relative to the variable $w = \left(\frac{dx}{d\tau}\right)^2$ as a function of *x:*

$$\frac{dw}{dx} = -\frac{2a}{x^2}\left\{\frac{1}{\varepsilon^2}\sin^2\left[|\varepsilon|\left(\frac{1}{2}(\tau+\tau_*)w + x\sqrt{w} \pm 1\right)\right]\right\}, \qquad (12)$$

where $\tau_* = -(2x_0)/\sqrt{w(x_0)}$ and +1 if ε >0, and -1 if ε < 0. It must be added also the equation for τ as a function of *x* :

$$\frac{d\tau}{dx} = \frac{1}{\sqrt{w}}. \qquad (13)$$

The system of equations (12, 13) is, so to say, a "model" system describing fairly accurately the deuteron movement under all values of |ε| from 10^{-24} to $10^{-6.}$ Numerical integration of this system was fulfilled under different values of e and under following initial conditions:

$$w(x_0) = 2.103, \tau(x_0) = 0, x_0 = -500000, \tau_{\bullet} = 689573.18 \qquad (14)$$

It may be noted that the initial deuteron velocity v_0 equals 1.450172 (following the relation $\dot{x}_0 = |\varepsilon|\sqrt{w(x_0)}$) for given initial $w(x_0)$ and for |ε| = 10^{-7} , i.e. such velocity is approximately 3.7 times less than velocity v_{00} corresponding the deuteron energy of 1eV. If $|\varepsilon| = 10^{-6}$ then the velocity v_0 is approximately 2.7 times greater than velocity v_{00} .

It turned out that the numerical tables for values of w, τ obtained under different values of $\varepsilon < 0$ in the interval (-10^{-24}, -10^{-6}) don't differ essentially from each other. The following table is true to three-four significant figures for τ and $\dot{x}/|\varepsilon| = \sqrt{w}$:

x	τ	$\dot{x}/\lvert\varepsilon\rvert$
-500 000	0	1.450
-50 000	$1.426 \cdot 10^6$	0.0493
-500	$1.002 \cdot 10^7$	0.000489
-200	$1.067 \cdot 10^7$	0.000440
-100	$1.090 \cdot 10^7$	0.000425
-80	$1.100 \cdot 10^7$	0.000423

If reducing the table values of x to centimeters, we obtain the following corresponding approximate values:

$$5 \cdot 10^{-9}, 5 \cdot 10^{-10}, 5 \cdot 10^{-12}, 10^{-12}, 0.8 \cdot 10^{-12}$$

The time interval ΔT , in which the deuteron reaches the critical distance 10^{-12} *cm* from the center, is $67350/|\varepsilon|$ of our time units or $(1.090 \cdot 10^7/|\varepsilon|) \cdot 3.502 \cdot 10^{-25}$ seconds. If nuclear forces are not taking into account then the deuteron may approach the distance less 10^{-12} *cm* .

We present here the table, where are given the initial deuteron velocities v_0 in velocities shares v_{00} and the corresponding time intervals ΔT (in seconds) for different values of ε.

ε	$\dfrac{v_0}{v_{00}}$	ΔT (s)
-10^{-6}	2.7	$3.82 \cdot 10^{-12}$
-10^{-7}	0.27	$3.82 \cdot 10^{-11}$
-10^{-22}	$0.27 \cdot 10^{-15}$	$3.82 \cdot 10^{4}$ (≈ 10.6 hours)
-10^{-23}	$0.27 \cdot 10^{-16}$	$3.82 \cdot 10^{5}$ (≈ 10.6 hours)

Let us note that the given data change essentially under positive values of ε (10^{-6}, 10^{-7}, etc.). There is some asymmetry of solutions behavior under negative and positive values of ε. The calculations show the minimum distance $|x|_{\min}$ more than 500 of our lengths units even for relative big initial $w(x_0)$ = 10000. Thus, if we limit ourselves to the condition that the deuteron energy is not over $(0.27)^2$ eV at a distance of $5 \cdot 10^{-9}$ cm from the central nucleus, and the whole process of deuteron movement towards the nucleus does not exceeds approximately 10.5 hours, then the interval $(\frac{\pi}{2} - 10^{-7}, \frac{\pi}{2} - 10^{-23})$ is approximately the sought phase "hole" in the whole interval ($(0, \pi)$) of phase change φ_0 in eq. (10).

If many deuterons with the energy not more than $(0.27)^2$eV at the distance $5 \cdot 10^{-9}$ *cm* from the nucleus are equally distributed along their phases φ_0 , the ratio of the length of this "hole" to π, equaling approximately $0.3 \cdot 10^{-7}$, is equal to a share (or a relevant percentage of $0.3 \cdot 10^{-5}$) of deuterons overcoming the Coulomb barrier.

The above figures express at least the order of probability of the cold nuclear fusion occurrence, and this order is absolutely incompatible with the figures in the classical quantum mechanics mentioned above. Let us note once again that a one-dimensional problem was solved, and in case of an accurate analysis (not zero sighting distance will be taking into account) this probability will be lower. Let us also pay attention to the large time intervals ΔT calculated if $|\varepsilon|$ is very small. It explains well the effect (observed by many researchers) of continuation of cold nuclear fusion reactions during even many hours after the disconnection of the voltage in the electrolytic cell. This effect was named even "life after death".

As for the analysis of the deuteron movement with the help of the autonomous equation, the calculations lead to initial velocities v_0 , exceeding the above mentioned numbers, although the general motion picture is the same. But the autonomous equation is interesting, because in the area of those values $x, \dot{x}$, under which the product $x\dot{x}$ is modulo small, it is possible to replace $\sin(x\dot{x})$ with $x\dot{x}$, and the eq.(11) under ε=0 to replace with simplified equation (describing the deuteron motion from initial point $x_0 > 0$ to center)

$$\ddot{x} = a\frac{(x\dot{x})^2}{x^2} = a\dot{x}^2$$

This equation has a very simple analytical solution. Without giving very simple calculations, we will present the final formulas.

Let us take the following initial conditions:

$$x(0) = x_0 > 0, \quad \dot{x}(0) = -v_0 < 0$$

Then

$$\dot{x}(t) = -\frac{v_0}{1 + av_0 t}, \; x(t) = x_0 - \frac{1}{a}\ln(1 + av_0 t)$$

It follows from these formulas that the velocity of a particle moving in accordance with the initial equation never turns to zero, and under

$$t = t_* = \frac{\exp(ax_0) - 1}{av_0}$$

$x(t_*) = 0$, i.e. the particle reaches the center of the nucleus, its velocity at this moment being

$$\dot{x}(t_*) = \frac{-v_0}{1+av_0t_*} = -v_0\exp(-ax_0)\ ,$$

so that it *passes through the nucleus and moves further.*.

For example, let a=0.0144967, x_0 = 1000 ($\approx 10^{-11}$ cm), $\dot{x}(0) = 5.37\cdot 10^{-10}$ (≈ 16 cm/s).

Under such initial data, the product $x\dot{x} = -0.0000537$, so it is quite possible to replace $\sin(x\dot{x})$ with $x\dot{x}$.
In this case,

$$t_* \approx 2.3\cdot 10^{7} \qquad (\approx 8\cdot 10^{-18}\text{ s}),$$

$$\dot{x}(t_*) \approx -29.9\cdot 10^{-17} \qquad (\approx 9\cdot 10^{-6}\text{ cm/s})$$

These figures fit well into the reasonable framework, so the autonomous model can also be of use for the movement analysis in the problem under review. The phenomenon of particle passage through the Coulomb potential accounts very well for the existence of pendulum orbits in the Bohr-Sommerfeld model, when in states *1s,2s,3s* etc. the electron passes through the nucleus. Such states in the strict theory and experiment have no impulse, so in the Bohr-Sommerfeld model they were discarded as absurd. Now they have a right to existence. Further, the experimental data for angular distribution of non-elastic scattering by nuclear reactions (including reactions with heavy ions) reveal the big amplitude of the scattering forward. It is impossible to explain such effect by the formation of intermediate nuclei but it is may be explained from the viewpoint of our UQT.

General Principles of Creating New Energy Sources

In the ancient classical *perpetual mobile* idea it is supposed that energy is just created and not taken from outside (impossibility of *a perpetual mobile* is the first law of thermodynamics). There have appeared lately many articles and even books dwelling on the idea of energy generation from vacuum. We are not in complete agreement with many of these works, and we will dwell only on some of them, which, in our view, can be of interest. One of the main ideologists of this completely new sphere in science are Daniel C. Cole and Harold E. Puthoff, and their first serious work entitled «Extracting energy and heat from the vacuum» was published in Physical Review E, vol. 48, #2, (1993). In this work authors use the Casimir forces [60] making them produce useful work. The appearance of such forces in vacuum is understandable intuitively: if in a stormy sea we put vertically into the water two big parallel plates, on the outside part of these plates the waves will hit them at random, and between the plates there will be no waves. Then, the hitting of the waves outside the plates will produce a gravitation force between them (the Casimir force discovered experimentally long ago), which the authors of this interesting work are going to exploit. It is easily seen that in this idea energy is generated from vacuum fluctuations.

Our approach is altogether different. When the equation with an oscillating charge was solved for the quantum oscillator, 4 types of solutions were discovered. For us only two of them matter – "crematorium" and "maternity home". In one solution ("crematorium") the particle slowly falls to the bottom of the pit and finally turns into a "specter" (under the strict unitary quantum theory it disappears, spreads about the Universe and contributes to vacuum fluctuations everywhere). In the other solution ("maternity home") the particle can even be born of a very small fluctuation, or accumulate a sufficiently big energy. Let us underline once again that both these processes are not at all logically connected. In other words, there are such systems where energy will disappear completely (electrolytic baths), or increase unlimitedly (it might be our Universe).

It is the energy conservation law that presents the strongest impediment in all cosmological approaches. However, universes with birth of matter have long existed in scientific cosmology independently of us. There is known the theory of British astronomer Fred Hoyle based on the idea of continued creation of matter from nothing. The question of whether such an approach is realized in nature and whether the energy emitted by quasars is the result of work produced by a certain gigantic pit, is the most intriguing question of the future.

It is yet unclear whether the values of appearing and disappearing energy in these solutions are equal. But neither in the strict UQT nor in the equation with an oscillating charge vacuum (as a big set of random oscillations) is needed for energy generation. Of course, UQT admits of such an energy exchange with vacuum. For example, during split of a photon on a semitransparent mirror, at one time both halves of the photon will not be registered and will give their energy to the vacuum and disappear for the observers for good, at another time there will appear two photons out of one, and the lacking energy will be taken from vacuum. But *the movement equations (4) and (5) themselves know nothing about vacuum and can generate energy due to their nature* (they are noninvariant relative to the coordinate translations) and the conservation laws we are so accustomed to do not exist for them.

Let us remind you once again that the latter follow from the Newton equations, and the Newton equations result from averaging by a big number of events, while for individual events of small energies no conservation laws in quantum physics exist.

In other words, it can be said philosophically that a motion of a small wave packet, once started, will give birth to other movements (energy) and, consequently, to matter. Since most various and breath-taking speculations are possible, up to the creation of a universe, we will stop here.

Thus, the generated or disappearing energy in our approach can be manifest not only in the changes of the particle velocity during movement in a certain potential, but in the appearance or disappearance of the particles themselves as well. A change of the particle velocity in movement is most easily discoverable, and it is the velocity increase that can be used for generation of heat or electric current. There can be energy systems, which exploit the fact itself of charge oscillation and the consequence of it. It is very probable that these phenomena, contradicting the most fundamental laws of modern science, have been long discovered and even applied. But these are the very phenomena that are the easiest to be exploited at the first stage of development of such new energy technologies.

When an energy generation mechanism is used, crematorium-type solutions should be suppressed. But all the quantum processes are built on the basis of elementary acts, and each of them is impossible to be controlled separately. But if the probabilities of such processes are controlled, they, being multiplied by the great number of participants in the process, automatically become macroscopic variables of quantum kinetics, and the process itself becomes possible. It can easily be achieved, if process participants with correlated initial phases are selected.

Let us remind you that the Newton and relativistic classical mechanics follow from the strict UQT, while the Newton movement equations with the resultant energy and impulse conservation laws follow from the oscillating charge equation with averaging by the particle ensemble composing a classical body (material point). But these conservation laws are nonexistent for individual microparticles in our theory, and they appear only in case of averaging by the ensemble of particles. Thus, if the energy-generating processes are accumulated, and the processes where energy disappears are suppressed, a classical *perpetual mobile* can be created.

But the UQT and the oscillating charge equation have other differences not only from the equations of classical mechanics, but also from some equations of electrostatics and electrodynamics.

There is a fundamental theorem of circulation for the electric field. Let us dwell on it in more detail. Let us have a vector field **E,** which can be an electrostatic or a gravitation field.

$$\mathbf{E} = \mathrm{P}(\mathrm{x},\mathrm{y},\mathrm{z})\mathbf{i} + \mathrm{Q}(\mathrm{x},\mathrm{y},\mathrm{z})\mathbf{j} + \mathrm{R}(\mathrm{x},\mathrm{y},\mathrm{z})\mathbf{k}$$

Line integral

$$\Gamma = \oint_l (\mathrm{P}\,dx + \mathrm{Q}\,dy + \mathrm{R}\,dz) = \oint_l \mathbf{Edl} \qquad (15)$$

is called circulation of vector field **E** by contour *l*. Of course, circulation depends not only on **E**, but also on the passage direction accepted in contour *l*; by changing the passage direction we will change the circulation sign. Form (15) is convenient for mathematicians, but for our purposes we will present equation (15) in a different way. If both parts of equation (15) are multiplied by electric charge *q*, on the right we will get the integral of force $q\mathbf{E}$ by way **dl,** i.e. work for moving the charge along a closed loop contour. It is well known that this value is zero.

$$\int_l q\mathbf{Edl} = \mathbf{0} \qquad (16)$$

If this value were not zero, an energy source could be created. For this purpose a charge should be moved in electric field **E** from point *a*, located in the high voltage area of the field, to point *b,* located in low voltage area of the field, and then back, but along another route. The values of work from $a \to b$ and from $b \to a$ would be different, and we could extract work from the field without making any changes in the system. When the charge is constant, it is certainly true, so for a macroscopic constant charge this theorem is an analog of the energy conservation law. The authors have not come across such an interpretation of the energy conservation law in other works. If the charge is microscopic, then in the UQT it changes, depends on time, coordinate and velocity, so work from $a \to b$ and from $b \to a$ will be different, in this case work can in principle be extracted from the field without any changes made in the system.

Discussion of Experimental Results

Let us now get down to explaining some very unusual experimental results, which the authors have nothing to do with, and which they sometimes regard rather skeptically. The point is that the sphere of new energy sources is the headache of all the human civilization, and in this sphere, like nowhere else, the dividends can be exorbitantly high, and for this reason there are in this sphere a lot of swindlers (even among the theoreticians) and simply erring people. The official science of the world does not so far believe in such research, but the most suspicious fact is the great multiplicity of such works. The authors are not inclined to regard all these people as swindlers or erring, because the UQT can offer a beautiful and simple interpretation of certain phenomena.

There are strange plants with the efficiency over 100%. They are even manufactured in small quantities and are rated among energy-saving devices already termed over unities. Japanese researchers take these problems very seriously, and the leading role in studying this problem belongs not to the USA, but to Japan, which even finances many US institutes in this framework. The total Japanese expenditures for this research exceed $200.000.000 a year. It can be forecast that with the Japanese mentality and the state policy of exporting not natural resources, but superhigh technologies and intellect, Japan will find itself among the leading countries early in the 21st century. We think that our readers will not be surprised to hear that Russia has not allocated a single cent for this program, and all research was made on pure enthusiasm.

In the USA such works do not get official governmental support either (like, for example, the dying out hot nuclear fusion problem), but a great number of private firms and individual businessmen are conducting large-scale research. The following US journals are devoted to the subject: *Journal of New Energy, Infinite Energy, Cold Fusion, New Energy News, Fusion Facts*, and *NET-Journal* (Switzerland).

Switzerland, Italy, Germany, and France are also among the countries where the new energy problems are seriously researched from the cold nuclear fusion point of view.

A very young sphere of power engineering has emerged and is quickly developing, which researches many new energy sources. In future those new energy-saving sources will first be used, which will considerably differ from the existing ordinary energy transformers in that they will generate additional energy that can be used in the interests of the mankind. The development of civilization will then be limited not by long-expected reduction of natural fuel resources, but by heat pollution of the environment.

Let us enumerate just a few of the new energy directions:
1. The Patterson fuel cell (CETI).
2. Supermagnet-superengines of Takahashi, Aspden and Adams
3. Swiss plant Testatika.
4. Engines operating on water.
5. Hypersound Griggs pump, the Potapov and Schaffer heat generators.
6. Schoulder and Fox cluster systems.
7. N – machines of Farade, Bruce de Palma, Newman, Searl, Tewari, etc.
8. PAGD reactor of Canadian researchers P. Correa and A. Correa.

This list can be complemented with the surprising experimental results received by physicists A. Samgin and A. Baraboshkin (Russia, Institute of High-Temperature Electrochemistry under the Russian Academy of Sciences, Ekaterinburg) [24,25] and T.Mizuno [26] (Japan). They appear to have used, totally independently of each other, special proton-conducting ceramics, which, when electric current runs through them, generate a thousand times more heat energy than the electric energy consumed. In some experiments by T.Mizuno this value even exceeded 70000(!). T. Mizuno in a personal talk with one of the authors of this report said that he feared very much the radiation sickness. But no α, β, γ radiation or nuclear debris was found, and the nuclear processes are not responsible for such energy generation. Such proton-conducting (or, to be more exact, deuteron-conducting) ceramics was made using the power metallurgy methods by agglomeration under high temperatures. In other words, all the chemical processes in it had long been over. The origin of such an amount of excessive energy is absolutely incomprehensible in the framework of conventional science, for they cannot be accounted for either by nuclear or chemical reactions, or by phase passages. At first the authors of this experiment supposed nuclear fusion reactions of the D+D type. At our request, A. Samgin replaced heavy hydrogen (deuterium) during ceramics production with ordinary hydrogen. If the effect of such huge energy generation was connected with the nuclear D-D reactions, all the anomalous heat effects would have disappeared, but they persisted. After such a large quantity of energy was generated, the tablet disintegrated into powder.

These effects can easily be explained by UQT from the harmonic oscillator theory point of view. When the tablet is agglomerated, there remain in it some caverns of a size of hundreds Angstrom units. When direct or alternating current flows through it, the protons and deuterons in their movement (there are few electrons in such ceramics) get into these caverns, and a process can start which is described by the “maternity home” solution. A particle accumulating energy, oscillates in such a pit, and finally the energy will be sufficient both for heating and for destruction of the pit walls (tablet turning into powder). The same processes seem to be taking place in a palladium electrolytic cell with heavy water, and in a nickel electrolytic cell with ordinary water, which accounts for anomalously large heat generation, not related to nuclear processes.

It would be good to verify experimentally the dependence of the tunnel effect on the initial phase. But it seems us that it is more important for our opponents, since both cold nuclear fusion (CNF) and discovery of nuclear transmutations *(which, from the point of view of modern science, are even more absurd than the existence of CNF*) evidently cannot be accounted for in any other way. Besides, such a direct experiment is of a fundamental value. There are today a lot of people and groups in the world, who pin great hope on exploiting the nuclear transmutation phenomenon for the purposes of processing and recycling of nuclear wastes, and the question of industrial generation of tritium for military purposes using CNF methods was under consideration in Los-Alamos. Internet magazines are full of such information. We are not giving Internet addresses here, because everything is constantly changing in this live system.

Let us analyze some of the above-mentioned devices. The first, the oldest and the most mysterious information was information about internal combustion engines operating on water.

Let us give just one example. When we were students, one of our teachers, the late Professor G.V. Dudko (1959) told us that in 1951 he had participated in the testing of an internal combustion engine [39,55-57]. The device represented a hybrid of a diesel and an ordinary carburetor engine, where a gas of petrol was needed to start it and then ignition was switched off, and an ordinary fuel pump sprayed into the cylinder warmed up and strongly compressed water with special additives (which the inventor himself put into the tank in small quantities, and which, as we now understand, represented the principal secret). The engine was installed on a boat. The researchers were riding for two days in the Azov Sea, and only water vapor was the engine exhaust. Professor Dudko himself drew the water fuel overboard and poured it into the tank. They needed much water, several buckets a day, but there was no shortage of it… The question of why, if everything was so great, these engines are still not in use, can occur only to a person who has never lived in Russia.

From the point of view of the solutions of the harmonic oscillator problem, the following theoretical possibility exists [40,44,47,55-57]: if water with the necessary additives (which, evidently, represent the secret of many invented engines operating on water) is compressed and sprayed into the cylinder, each drop of water, when it gets into the cylinder after being compressed, will start dilating and will pass by inertia the equilibrium position. As a result, caverns (empty volumes) can be formed in it, with a size of several dozen of Angstrom units. If a free proton (or some other microparticle) gets into such a cavern in the required phase (it is supposed that the task of the additive is exactly this), the "maternity home" solution will be realized and some of the drops will explode… Later we heard and read many times about various Russian inventors, who had successfully created and tested engines operating on ordinary water with some mysterious additives.

Of course, the possibility of catalytic water decomposition with small energy consumption before spraying into the cylinder is not at all excluded. There are films and information in Internet about testing of cars operating on water, which is catalytically (with small energy consumption) decomposed into oxygen and hydrogen. Such power engineering would be ecologically absolutely clean, and the only restriction would lie in heat pollution of the environment.

An ideal solution for the motor transport could also lie in use of some new types of electric energy generators. The UQT even admits of the possibility, which was long observed in the experiments of Nicolas Tesla and in those made by Canadian physicists the Correas, who even received a patent for a system generating energy from vacuum fluctuations (as they believe) [45]. The readers could have got acquainted with our detailed theory of these processes in [46]. But the ideal system for the automobile would certainly be Testatika.

Any imagination will be amazed at the thermal cell CETI created by James Patterson, USA [27], in which takes place the electrolysis of specially made nickel balls in ordinary water. The US paper «Fortean Times» [1] 85, 1995, wrote about it: "December 4, 1995 will go down into history. On this day a group of independent experts from 5 US universities was testing a new source of energy with a stable output heat power of 1.3 kWt. The consumed electrical energy was 960 times smaller". All experts note that the generated heat is of mysterious origin and cannot be explained by chemical or nuclear reactions, as well as by phase passages. The US ABC TV showed on February 7 and 8 1996 in the «Nightline» and «Good Morning America» cycles of programs about the development by Patterson of a new energy source generating hundreds of times more energy than it consumes. The mysterious nature of the generated heat was again underlined. It is interesting [34] that Motorola tried to buy the CETI patent from its authors for $20.000.000, but met with a refusal. We are sure that Motorola had invested a certain amount of money into the study of this problem before making such a serious offer. All that happens within the Patterson element has nothing to do with nuclear reactions (although Patterson told one of the authors that he was of a different view), and, in our opinion, can be accounted for by exactly the same processes as were described above for proton-conducting ceramics.

The sonoluminescence phenomenon, when certain liquids start shining if weak ultrasound is run through them, also looks very mysterious. No satisfactory explanation has so far been found for this experimentally proved phenomenon, discovered by Moscow University Professor S.N. Rzhevkin in 1933. As Nobel Prize winner Professor Yulian Schvinger said, "it has no right to exist, but it does exist" [38]. This phenomenon can also be explained from the above-mentioned positions.

There are also heat generators (Yu. Potapov [21-23], Moldavia, James J.Griggs [28], and Huffman [29], Schaffer - USA). In them many cavitating bubbles are formed during circulation of ordinary water, in which excessive energy is generated, with the output to input energy ratio approaching 1.7. In these experiments and plants no chemical or nuclear reactions can take place, and thousands of Potapov's heat generators have been manufactured for heating homes. In such devices (they are very different in appearance) a great number of cavitating bubbles are created in a flow of water. This is achieved either with the help of interrupting the water flow with a special rotor (J.Griggs, Huffman, Schaffer), or the water flow is twirled by a special helix and then enters the zone of sharp dilation, where cavitating bubbles are formed (Yu. Potapov). In general, it should be said that cavitation remains a great puzzle for theoretical hydrodynamics and science. For example, forged multi-ton screw propellers of big nuclear submarines under certain operation modes and geometry of the surrounding forms can be destroyed by cavitation within only a few hours. It happens because of huge energy generated in cavitating bubbles.

Under certain values of phase and energy, a particle in the pit, each time reflecting from the walls, will have a greater velocity than that of a falling particle (this is within the uncertainty relation), and after many reflections will accumulate a fairly big energy which will be generated in the form of heat or bremsstrahlung when the pit is destroyed, and, finally, the energy of the oscillations of such a particle accumulated in the pit will always be transformed into heat in an ordinary solid body or a liquid. This physical idea immediately accounts for both sonoluminescence (although for sonoluminescence in general this mechanism is less primitive), and energy generation in proton-conducting ceramics, nickel during electrolysis in ordinary water (CETI element), and water bubbles of commercial heat generators. The theory predicts that the samples should be fissure due to increased pressure on the walls of the potential pit with the growth of energy, which fact also takes place, since both ceramic samples and nickel balls in the CETI element finally disintegrate. It is evidently for these reasons that any metal containing much hydrogen in its grid becomes fragile and is quickly destroyed, which fact is well known to engineers.

The small number of experiments does not so far allow for making concrete conclusions as to what particles generate energy in pits (microbubbles). Besides, for at least an electron to disappear a pit of about 0.5 MeV is required, while in a solid body the pits are about several eV deep, and what seems to happen is only loss of kinetic energy, and not disappearance of particles. The fact that this process requires very deep potential pits, which do not exist in a solid body, does not change the essence of the matter.

Of course, under ordinary conditions, both competing solutions usually take place at once: "maternity home" and "crematorium", which compensate for each other and the energy is preserved. For energy generation, the "maternity home" solution should prevail. Both these processes take place simultaneously and compete with each other, but, formally, they are not connected in space and time. The complexity of the energy generation problem lies in suppressing the "crematorium" solution by a careful selection of different parameters and promoting the "maternity home" solutions. So far we cannot say for sure what the optimum dimensions of such cavitating bubbles are, or which object oscillates in them, because for this purpose special experiments are needed, which so far have not been staged.

Of course, the inexorable Robber in the form of the Carnot principle stands in the way of transformation of the heat generated in a heat generator or ceramics into electrical or mechanic energy. In accordance with this principle, all mechanic or electrical energy can be transformed into heat, but the reverse process is always connected with big losses.

If there are experiments and plants in which energy generation contradicting the conventional conservation laws is discovered, there should also exist opposite ones, where energy disappears completely, i.e. the "crematorium" solution prevails. It proved to be true. There are such modes during electrolysis in electrolytic baths, under which the temperature of the solution in the bath is strongly reduced for unaccountable reasons, and this fact has no explanation at all. This phenomenon long ago was noted by attentive industrial engineers, and it is called the "bath-freezing" mode [49,50]. Chinese physicist Swe-Kai Chen from Taiwan in his experiments [48] stably observed the same phenomena. It is quite easily explained: a particle with a velocity exceeding the most probable velocity in this distribution gets into caverns on electrodes and after some oscillations reduces its velocity, which becomes smaller than the most probable one, and then the particle leaves the cavern at a small speed, and the same process can happen to another energetic particle. This leads to the cooling of the cell in the case of such mass processes.

The problem of ferromagnet magnetization (the Easing model) can also be reduced to the orientation of a magnetic doublet by the external magnetic field, and then it is essentially the harmonic oscillator equation with a slightly different return force ($F = \frac{\mu}{r^3}$) and all the conclusions made earlier remain in effect. That is why magnetization should also produce energy generation effects. This proved to be true. For the general public everything began on May 17, 1996, when Frode Olsen from the research group "Free Energy" showed on the Norwegian TV (TV2) a surprising film about a "dynamic sculpture" made by artist and sculptor Reidar Finsrud from Skaarer, Norway. The author of this "dynamic sculpture" *had no idea about physics* and had been making it for 12 years. Einstein's idea of how discoveries are made conveniently comes to mind at this point: everyone knows that a certain thing cannot be done, but there is a man who does not know it, and it is he who makes the discovery.

This "dynamic sculpture" accompanied by an "explaining" poster «*perpetual mobile*» represents an iron well-polished ball with a diameter of 2.7 inches weighing about 2 pounds. The ball is rolling along a circle on close guides resembling two parallel skids with a diameter of 25 inches past the poles of three permanent magnets, where it is magnetized. In the area of three permanent magnets three more mobile magnets are installed on special mobile 5-inch long levers, and these magnets, when the ball passes them, are slightly inclined (due to the ball gravitation) and, after the ball passes them, are raised by the holding springs (sway like yokes). The ball makes a complete turn in 3 seconds. All this magic (***they say the ball had been rolling along the close contour for more than a year***) does not have any sources of energy and is installed for everyone to see in a Norwegian picture gallery on a special stand covered with a glass cover. The authors only saw a good TV film about this installation and were mostly surprised at the fact that the ball had not stopped during uninterrupted shooting (about 20 minutes).

We are well acquainted with circus tricks, but it is absolutely incomprehensible how such a trick could be staged using some secret methods. It is clearly seen that the ball in its movement always partially transfers its energy to the three long swaying pendulums, but there is no way to use them for pushing the ball and making up for friction, this being the only trick that could, in our view, be applied here. All the rest is clearly visible and contains nothing suspicious.

Let us estimate the generated energy. At an initial speed of about 1m/s the ball stops after 30 seconds, if all the magnets are removed. It means that the energy consumed in 30 seconds is about 0.5 joules, or 1/60 Watt. The total energy generated in a month is 43,200 joules, and this is huge energy, much greater than that of a good shell!

It is clear (if the word is relevant here at all) that when the ball is approaching the permanent magnet and the process of magnetization is going on, it is accelerated, but when it mechanically gets past the equilibrium position and, moving away from the magnets, becomes demagnetized, the gravitation (which now starts slowing the ball down) will be slightly less than it was at the moment of the ball's acceleration. This small difference in forces provides for small positive work to overcome friction. Energy generation and similar things during magnetization had been *predicted* by one of the authors in magazines *Infinite Energy* vol.1, No.2, p.38, (1995); *Proceedings of the ICCF*5, p.361, April 9-13, (1995), Monte-Carlo; *Cold Fusion*, No 11, p.10, (1995); *Chinese Journal of Nuclear Physics* (vol. 19, ^{1}2, 1997). The quantum-mechanic processes are very complicated, but some of them can be understood.

All keen physicists were quick to understand it, and J. Naudin in France made a similar, but much simpler experiment. A ball of a soft magnetic material is swaying along parallel U-formed skids in a system of four magnets. Near the bottom of the U-form there is a small smooth step. It may have been made to make the magnetization and demagnetization processes different in time, which is very important. If there are no magnets, nothing interesting happens and oscillations are quickly (in a few seconds) damped. If the magnets are present, oscillations go on up to 3 hours 27 minutes. It appears that in this case the author failed to find good material and parameters of the plant, so friction was not compensated completely. In all these experiments demagnetization of permanent magnets does not happen, because the experiment is repeated many times with the same results.

And now a few vague words about demagnetization processes. During magnetization of the ball, the magnetic moments of its atoms are oriented (like the hands of a compass) along the field lines. When the ball leaves the magnetic field area, the atom magnetic moments are disoriented under the influence of the heat motion, and it becomes demagnetized. In the unitary quantum theory the share of the oriented magnetic moments in the external field can be bigger than in the conventional quantum mechanics (the "maternity home" solution), and the ball gravitation can be stronger due to it. Disorientation of these moments happens similarly in both theories. It seems to be for this reason and due to the difference in magnetization and demagnetization time that a difference in magnetic forces occurs when the ball approaches the magnet or moves off from it.

The scientists of the older generation will remember that a similar toy was shown in the 30's to David Gilbert, who said it was the most interesting thing he had ever seen. A question arises as to why it has not yet been realized. We do not know a physical-mathematical answer to this question, and it is not our task to analyze the social reasons of this phenomenon. Japan has a different mentality, and there is a governmental program for generating energy from permanent magnets. Takahashi [51] even seems to have made an electric engine with an efficiency of up to 318%!

Still more mysterious is the long-known problem of energy shortage in many biochemical reactions with ferments (enzymes). For example, in the well-studied reaction of disintegration of polysaccharides in the presence of lysozyme the following happens: a polysaccharide molecule gets into a special cavern in a big lysozyme molecule, and some time later its debris are thrown out of it (Fig. 6). The broken binding energy of the polysaccharide is about 3 eV, while the energy of the heat movement is only 0.024 eV. From the standard science point of view, it is absolutely unclear where lysozyme takes the energy to break the polysaccharide. No satisfactory mechanism for explanation of such reactions (and they are very numerous) was found, and all this was "swept under the carpet", as physicists say. The UQT provides for a completely new look at the catalytic processes, which has an incomprehensible source of energy reducing the molecule activation energy. From our point of view, this process is a variant of the "maternity home" solution for oscillator.

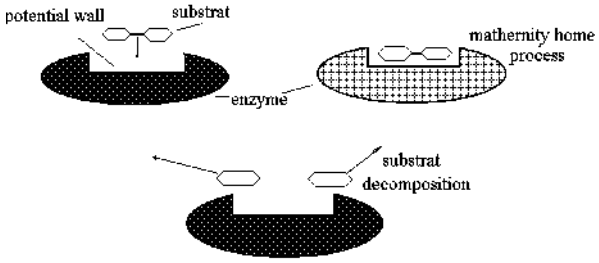

Fig.6. Break of polysaccharide molecule by lysozyme.

The most surprising thing is that in all the cases generation of excessive energy cannot be accounted for by chemical reactions or phase passages. If nuclear reactions do sometimes happen (which should not be according to modern science), they can account for only a hundredth or a thousandth share of the generated heat energy. There is no doubt that all these are effects of new physics, for in the framework of the old physics all this is simply unexplainable.

But the existence of a plant that produces *out of nothing* about 10 kilowatt of direct current electric energy with a voltage of 300 V seems nearly impossible. The story was described by one of the authors in three different magazines, and we will just give a brief resume [52-54].

In summer 1999, at the invitation of Swiss physicists (Director of the Institute of New Energy Sources in Egerkinhen Adolph Schneider), one of the authors visited several research organizations. It is interesting that there is such an institute in small Switzerland, and there is none in big Russia. The purpose of the invitation was very simple: to explain the operation of a plant generating energy out of nothing, i.e. a *perpetual mobile*. In Switzerland such plants are called Testatik Machine M/L Converter from religious group «Methernitha» (Address: Methernitha, CH-3517 Linden, Switzerland, phone: ++41 31 97 11 24).

Such machines exist today in the Swiss town of Linden near Bern. Part of the town belongs to the Religious Christian Community, which is fenced and heavily guarded. There are about 250 members of the Community, many of them are physicists, graduates of the universities of Geneva, Lozanne, Bern. It is not only a research laboratory, they have their own TV center, a film studio, a small furniture plant, shops, garages, residential blocks, and support services. You will probably have guessed that this community does not consume any energy, and this is the most accurate fact in the whole story, for the inquisitive journalists have found out that no money from them comes to the accounts of the local power station, which provides power for all the town. In a cellar of one of the houses they have a power station that produces energy… out of nothing. The author of this inexhaustible source of free direct current energy is Swiss physicist Paul Baumann. Let us briefly describe these fantastic plants: they are of four types (sizes) with capacities of 0.1, 0.3, 3 and 10 kWt. Externally, the plant resembles very much the standard electrostatic machine with Leyden jars often used in physical demonstrations. There are two acryl disks with 36 pasted narrow sectors of thin aluminum, which rotate in different directions. In the first samples ordinary gramophone records were used for disks. The machine is started by pushing the disks in different directions by fingers. The rotation speed is 50-70 turns per minute. After the start disks rotate independently and can be easily stopped by hand, the direct current voltage is about 300-350 V, and the current is up to 30A. The mechanical energy used for rotation (only 100 mWt, according to measurements made by Austrian Professor S. Marinov) is hundreds of thousand times smaller than the generated electrical energy. The biggest plant for 10 kWt has plastic disks with a diameter of over 2 m, the smallest one – 20 cm, the weight of the plants is small enough, the 3-kWt machine weighing about 20 kg.

The charge separation process (which consumes energy!) practically does not slow down the disks. Connection of a load in the form of a 200-Wt bulb does not change the rotation speed either. No cooling or heating of the air or machine parts during long operation takes place, only a slight smell of ozone is felt. The system is noiseless, compact, environment-friendly, and can be installed anywhere.

The Community management thinks, and quite rightly, that wide spread of such systems in the world will lead to a heat explosion, because all the energy generated by the mankind finally finds itself in an energy dump (is transformed into heat), and all this can finally lead to overheating of the environment. They absolutely do not believe (and not without grounds) in the capability of the mankind as a whole to negotiate reasonable use of this invention, and they think that the harm caused by it will be greater than from nuclear, bacteriological, or conventional weapons. Their main idea for the mankind is to live in balance with the environment and to make full use of the energy of the wind, the sun, the water, etc. For this reason the Community is heavily guarded, and they are not going to donate their main discovery to the mankind.

Professor Stephan Marinov visited the Community twice (in July 1988 and in February-March 1989). He was even given such a plant with a capacity of 100 Wt (300 V, 0.3A), which he studied in his laboratory. As far as we now know, even the inventor of this machine does not fully understand its operation principle, so he contacted Marinov out of sheer curiosity of a scientist.

In 1989 Professor Marinov published a book “Thorny Path to Truth – Documents of Violation of Conservation Laws” in International Publishers East-West. The book contains a lot of photos, a measurement report, and a description of the plant. He also organized a research group called “Free Energy” within the Community (Methernitha Group Stephan Marinov Free Energy).

There are very interesting words in this book: “I can state without any doubts that this machine is a classical *perpetual mobile* in its pure form. After the initial push, it goes on rotatingby itself for an indefinitely long time, constantly producing electrical energy in the amount of 100 Watt… It is still unclear, however, how it all can happen…”. As far as we know, nobody has managed to build a similar plant elsewhere.

We have an approximate idea of how the plant operates. The idea is as simple and ingenious as that of the wheel, which is absent from the surrounding nature, so the inventor could not borrow the idea. We will just show that the existence of such a plant is in full conformity with the UQT. It is natural that the plant operates on the basis of the charge separation principle. Let us have two metal spherical surfaces with a hole, isolated from the earth and from each other. If, with the help of an insulated stick, we transfer the first electron from Ball A to the internal surface of Ball B through the hole, a difference of potentials will occur, and if we transfer the second and the subsequent electrons, Ball A will attract the transferred charge, while Ball B will repulse it, and energy will have to be spent during the transfer of charges (Fig. 7).

Let us remind you that under the existing circulation theorem (16), the charge transfer work will consume the same amount of energy as will later be generated during the passage of electric current resulting from charge separation. But in the UQT the circulation theorem (16) for *an individual elementary charge* is not valid. Thus, we can select the time and route, along which the charge will be transferred in such a way, that

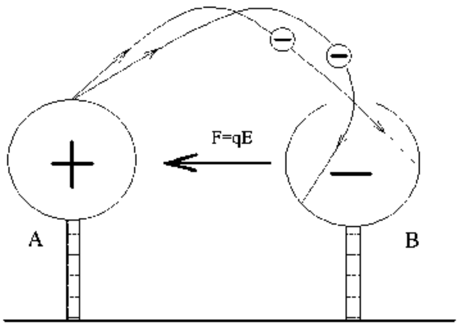

Fig.7. Work for moving the charge depends on method of movement and route.

the charge value during the transfer will be close to zero, and, consequently, the electrostatic force and the charge transfer and separation work will be close to zero too. For example, instead of selecting the route you can wait for the charge to be reduced to zero and then transfer it quickly, and when the charge increases, immediately stop the transfer and fix the charge. Or you can duly select the route and velocity. There are many options. This was evidently realized by Paul Baumann, who is *so far* practically unknown to the official science, and who can find consolation in the idea that the inventor of the wheel will never be known at all. The problem of simple arrangement of all this is just a matter of technique.

You cannot help, thinking that all these might be just tricks. The history of *perpetual mobile* abounds in evidence of downright swindling and frauds, and not a single positive result before, and who can guarantee that the information given above will not prove to be another swindle?

First of all, if all the people always piously believe in the unquestionable stability of the energy conservation law, there will never be any progress in this sphere, and it is then unexplainable how man got down from the palm at all. Secondly, to justify the proposed rebellious position, the following idea comes in mind: if 30 years ago somebody had told the authors (who were then already professors) that at the beginning of the next millennium they would deal in such research, it would have seemed not only a silly joke, but an absolutely impossible thing as well. But, as Voltaire said, "He is silly who does not change".

In conclusion we wish to express with certainty that the time of theoretical recognition and of practical universal using of overunity devices will come soon and become the epoch of new energetics. The people of our planet will regret that so much oil, coal and gas was burned causing terrible ecological losses.

The authors thank astronaut V.A. Dzhanibekov and Professor A.P. Buslaev.

References

1. Sapogin, L.G. Unitary Field and Quantum Mechanics. In: Investigation of Systems. Academy of Sciences of the USSR, Vladivostok, No 2, p. 54-84, 1973 (in Russian).
2. Sapogin L.G. On Unitary Quantum Mechanics. Nuovo Cimento, vol.53A, No 2, p.251, 1979.
3. Sapogin, L.G. A Unitary Quantum Field Theory. Annales de la Fondation Louis de Broglie, vol.5, No 4, p.285-300, 1980
4. Sapogin, L.G. A Statistical Theory of Measurements in Unitary Quantum Mechanics. Nuovo Cimento, vol.70B, No 1, p.80, 1982.
5. Sapogin, L.G. Statistical Theory of Detector in Unitary Quantum Mechanics. Nuovo Cimento, vol.71B, No 3, p.246, 1982.
6. Boichenko, V.A. and L.G. Sapogin On Equation of Unitary Quantum Theory. Annales de la Fondation Louis de Broglie, vol.9, No3, p.221, 1984.
7. Sapogin, L.G. and V.A. Boichenko On Solution of One Non-linear Equation. Nuovo Cimento, vol.102B, No 4, p.433, 1988.
8. Sapogin, L.G. and V.A. Boichenko On Charge and Mass of Particles in Unitary Quantum Theory. Nuovo Cimento, vol.104A, No 10, p.1483.
9. Sapogin, L.G. Clearcut Picture of Microworlds. "Journal Technique for the Young" (Tekhnika Molodezhi). Moscow, No 1, p.41, 1983 (in Russian).
10. Sapogin L.G., "Statistical Theory of Measurements in Unitary Quantum Mechanics," Nuovo Cimento, vol. 70B, No.1, p.80, 1982.
11. Sapogin L.G., "Statistical Theory of Detector in Unitary Quantum Mechanics," Nuovo Cimento, vol. 71B, No. 3, p. 24M, 1982.
12. Boichenko V.A., Sapogin L.G., "On Equation of Unitary Quantum Theory," Annales de la Fondation Louis de Broglie, vol. 9, No. 3, p.221, 1984.
13. Sapogin L.G., Boichenko V.A., "On Solution of One Non-linear Equation", Nuovo Cimento, vol. 102B, No.4, p.433, 1988.
14. Sapogin L.G., Boichenko V.A., "On Charge and Mass of Particles in Unitary Quantum Theory," Nuovo Cimento, vol. 104A, No.10, p.1483, 1991.
15. Sapogin L.G. "Deuteron Interaction in Unitary Quantum Theory", and "On Mechanisms of Cold Nuclear Fusion". In: Proceedings of the Forth International Conference on Cold Fusion, vol.4, Theory and Special Topics Papers TR-104188-V4, July 1994, p.171-178, Hawaii. 1994.
16. Sapogin L.G. "Deuterium Interaction in Unitary Quantum Theory", and "On Mechanisms of Cold Nuclear Fusion". In: Fusion Source Book. International Symposium on Cold Nuclear Fusion and Advanced Energy Sources, Belarussian State University, Minsk, May 24-26, p.91-98. 1994.
17.Sapogin, L.G. "Cold Nuclear Fusion and Energy Generation Processes in Terms of theEquation". Chinese Journal of Nuclear Physics vol.19,#2, p.115-120, 1996 .
18. Sapogin, L.G. «Cold Nuclear Fusion and Energy Generation Processes in Terms of the Schrödinger Equation». Infinite Energy [E.Mallove, editor], vol.1, No 5,6, p.75-76, 1996.
19. Sapogin, L.G. «Energy Generation Processes and Cold Nuclear Fusion in Terms of the Schrödinger Equation». In: Proceedings of the Sixth International Conference on Cold Fusion, Progress in New Hydrogen Energy, October 13-18, 1996, Japan, vol.2, p.595-600.
20. Sapogin, L.G. "Energy Generation Processes in Terms of the Schrödinger Equation."Proccedings îf the 2nd Russian Conference CNFNT (in Russian) p.18-24, Sochi, September 19-23,1994.
21. Potapov J.S. Patent of the Russian Federation No 2045715 Heat Generator and Device for Heating of Liquids. Registered on the 10-th of October 1995; priority from the 26-th of April 1993 (in Russian).
22. Potapov J.S. Water as a Source of Life and Energy. Enerjia-takarekossadi, Revu. p.25- 29, September 1998, Budapest.
23. Potapov J.S., MD, Patent No 649 Instalatie pentru obtiinerea enerjiei electrice si-temice. Buletin Oficial de Proprietate Industriala, No12, p.18-19, Chisinau (in Moldavian).
24. Samgin A., Baraboshkin A. et al. Influence of Conductivity on Neutron Generation Process in Proton-Conducting Solid Electrolytes. In: Proceedings of the 4th International Conference on Cold Fusion. Palo Alto, USA, v.3, p.51-57, 1994.
25. Samgin A. Cold Fusion and Anomalous Effects in Deuteron Conductors during Stationary High-Temperature Electrolysis. In: Proceedings of the 5th International Conference on Cold Fusion. April 9-13,1995, Monte-Carlo, p.201.
26. Mizuno T., Enio M., Akimoto T. and K. Azumi Anomalous Heat Evolution from $SrCeO_3$-type proton conductors during absorption/desorbtion of deuterium in alternate Electric Field. Proceedings of the 4th International Conference on Cold Fusion, December 6-9,1993, Hawaii, vol.2, p.14., EPRI, Palo Alto, USA, 1994.
27. Patterson J.A System for Electrolysis, U.S. patent No 5,494,559,27 Feb.1996; Miley G.H. and J.A. Patterson in: Proceedings of the 6th International Conference on Cold Fusion, Progress in New Hydrogen Energy, October 13-18,1996, Japan, vol.2, p.629-644.

28. Griggs J. Calorimetric Study of Excess Heat Production within Hydrosonic Pump System Using Light Water. Fusion Source Book. International Symposium on Cold Fusion and Advanced Energy Sources, Belarussian State University, Minsk, Belarus, May 24-26, p.248-253, 1994.
29. Huffman M.T. From a Sea of Water to a Sea of Energy, Infinite Energy, vol.1, No 1, p. 38-45, 1995.
30. Fleischmann M., Pons S. Electroanal. Chem., v.261, p.301, 1989.
31. Sapogin, L.G. and I.V.Kulikov "Cold Nuclear Fusion in Unitary Quantum Theory". Chinese Journal of Nuclear Physics, vol.17, No 4, p.360-370, 1995.
32. Cryz W.: Rivista Nuovo Cimento, 1, Special No, 42, 1969.
33.Kervran Lois C. Biological Transmutations. Swan House Pub. Co, NY, 11223, 1972.
34.Private Communication.
35. Notoya R., Noya Y., Ohnisi T. Fusion Technology. vol. 26, p. 179-183, 1993.
36.Swartz M. Journal of New Energy vol.1, #3, 1996.
37. Blokhintsev D.I. On Energy Conservation Law. In: Works on Methodological Problems of Physics, p.51, 1993, Print of Moscow State University. (In Russian).
38. Schwinger J. Casimir "Energy for Dielectric". In: Proceedings of the National Academy of Sciences, vol.87, p.8370-8372, 1990, "Cold Fusion: Does it Have a Future?" Journal "Cold Fusion", vol.1, #1, page 14-17,1994.
39. Sapogin L.G. «What Can Our Power Engineering Be Like in the Next Millennium», Journal Business-Match, 14, 1998 (In Russian).
40. Sapogin, L.G. On One of Energy Generation Mechanisms in Unitary Quantum Theory. Infinite Energy [E.Mallove, editor], vol.1, No 2, p.38-39, 1995.
41. Sapogin, L.G. On One of the Energy Generation Mechanisms in Unitary Quantum Theory. Proceedings of the ICCF5, p.361, April 9-13,1995, Monte Carlo.
42. Sapogin, L.G. Energy Generation Processes and Cold Nuclear Fusion in Terms of the Schrödinger Equation. In: Proceedings of the Sixth International Conference on Cold Fusion, Progress in New Hydrogen Energy, October 13-18, 1996, Japan, vol.2, p.595-600.
43. L.G. Sapogin, "On One of Energy Generation Mechanisms in Unitary Quantum Theory". Proceedings of the 2nd Russian Conference CNFNT (in Russian) p.18-24, Sochi, September 19-23, 1994; Cold Fusion, No 11, p.10, 1995.
44. L.G. Sapogin, "On One of Energy Generation Mechanisms in Unitary Quantum Theory". Cold Fusion, No 11, p.10, 1995.
45. Correa Paulo and Correa Alexandra XS NRG in Technology, Infinite Energy, vol.2, #7 p.18-38, Nr 8 p.10-15, #9 p.33-37, 1996. US Patents, numbers: 5.416.391, 5.502.354, 5.449.989.
46. Sapogin L.G. Theory of Excess Energy in PAGD Reactor (Correa reactor). In: Proceedings of ICCF-7, Vancouver, April 1998; Infinite Energy, No 20, 1998, p.49.
47. Sapogin L.G. "New Source of Energy?" Journal "Acknowledgement and Physical Reality", Moscow, vol. 2, #1, page 34-40,1997, (in Russian).
48. Swe-Kai Chen, Chu-Yung Liang «Observation of Cell Temperature Drops». In: Proceedings of ICCF-7, Vancouver, April 1998, p.68-72.
49. Jakimenko L.M. Electrolysis of Water. Chimia Press, Moscow, p.p. 33, 86, 90-114, (1970) (in Russian).
50. Pfleiderer N. Electrolysis of Water. pp. 12, 17-18, 1935, Leningrad (in Russian).
51. Rothwell J. Yasunori Takahashi´s Supermagnets, Infinite Energy, vol.1, No 5,6, p.33, 1996.
52. Sapogin L.G. "Is This Really True?", Infinity Energy, N 28, 2000.
53. Sapogin L.G. «Perpetual Mobiles Operating in Switzerland»,magazine «Chudesa I Prikluchenia» # 2, 2000.
54. Sapogin L.G. «They Say There are No Perpetual Mobiles. Then what is it?", magazine "Samolet", 4, 2000.
55. Sapogin L.G. XXI Century - New Sources of Energy? In: Chudesa i Prikliuchenia, Moscow, No 11, p.32-35 (In Russian) 1996, and No 3, 1998.
56. Sapogin L.G.,Kulikov I.V. "Neue Quantenfeldtheorie und prozesse zur electromagnetischer und thermisher energie mit overunity effekt", DVR-Mitglieder-Journal 2/2000.
57. Sapogin L.G. "The 21st Century: Will it Bring a New Quantum Picture of the Universe and New Energy Sources?", Journal of New Energy,vol.2, #3/4,1999.
58. Panos T. "Electrically Induced Nuclear Fusion in the Living Cell ", Journal of New Energy vol.3, #1, 1998.
59. M.Sue Benford, R.N. M.A. "Biological Nuclear Reactions: Empirical Data Describe Unexplained SHC Phenomenon" Journal of New Energy vol.3, #4, 1999.
60. Schwinger J. "Casimir Energy for Dielectric". In: Proceedings of the National Academy of Sciences, vol.87, p.8370-8372, 1990, "Cold Fusion: Does it Have a Future?" Journal "Cold Fusion", vol.1, #1, page 14-17,1994.

KOZYREV-DIRAC EMANATION.

INTERACTION WITH MATTER and METHODS OF DETECTING

Dr. Ivan M. Shakhparonov

125252, Russia, Moscow,Pestchanny Pas. 20-1-33
phone/fax 8·095-198-2012

In this paper the authors show the possibility of creation of a new kind of emanation. The magnetic monopole beam can be made in space as a result of focusing of some natural substance. Special devices based on the Moebius band elements make the given focusing. This emanation is able to magnetize graphite and organics, decrease the radioactivity, and influence the oncology diseases .The time reverse technology is realized in such devices.

Experimental data, which allow making a conclusion about existence of previously unknown emanation, are presented in this report. Here are descriptions of experiments and methods of measurement. The effects of interaction between new type of emanation and matter have been obtained.

Till the present moment theoretical physics didn't pay attention to the nonoriented configurations and spaces. The reason of this situation is the fact, that from the philosophic point of view it is not possible to determine and locate the area of the nonoriented topological structures in our world. We (eight scientific teams) joined our forces and we needed more than 30 years to solve this problem by an experimental approach.

The fundamental tenet of the casual mechanics developed by Kozyrev can be formulated as follows. There are two types of energy in the Universe. The positive or «right» energy acts as a factor of the entropy increase. The negative, or «left» energy tends to decrease the entropy, i.e. it acts as a factor, which regulates the entropy increase. The «right» energy is transformed to the «left» one and this fact may be interpreted as a course of time from the past to the future. When the energy is transformed from the «left» to the «right» form, time is reversed. Kozyrev supposed [1] that through revolving of a body together with a

6. Activation by nuclear magnetic resonance;
7. Activation by electronic paramagnetic resonance;
8. Activation by electrochemical force.

All these methods can be used as possible way to high efficient energy systems. Gerlovin wrote: “Usually 1, 6 and 7-th methods of structural activation are realized in catalysis simultaneously. Besides, catalysis differs from macroscopic methods because it has the most minimal distances from the sources of activator fields to the activated molecules. And finally, an active participation of force fields created by nuclei of atoms and significantly more active participation of disturbed EPV is possible in catalysis. That’s why catalysis is the most effective method of structural activation. The detailed account of this method exceeds the limits of this article and we can only annotate it.” [1, p.333]

Information stated above is only a small part of the questions appeared under consideration in Gerlovin’s theory of fundamental field (TFF). Other important questions should be considered with a new experimental data.

References

1. Ilia L. Gerlovin, “Foundations of unified theory of all interactions in matter”, published in 1990, St.Petersburg, Russia.

2. To the question of multipolarity, Alexander V. Frolov, New Energy Technologies, #1 (4) 2002, St.Petersburg, Russia.

3. N. A. Kozyrev, Selected works, St.Petersburg, LGU, 1991.

4. Practical application of time rate control (TRC) theory, Frolov A.V., New Energy Technologies, #3, 2001, p.15, St.Petersburg, Russia.

Antigravitation Force and Antigravitation of Matter. Methods of its Creation

Anatoly K. Gaponov

Sadovaya Str. 195, Novosibirsk, 630009, Russia

Part I

For a long time there is an opinion in physics about antimatter as a possible source of antigravitation, but the researches on this subject came into a dead end. The existent presentations and formulas forbade the conclusion about antigravitation, but our conducted investigations brought us to the possibility to get antigravitation of substance and to the paradoxical conclusions concerning the next:

1. Two types of space exist:
    a) The Absolute space
    b) The Relative space

2.a The Gravitation Field is the relative space, which has accelerated motion, directed to the center of a planet.

2.b The Antigravitation Field is the relative space, which has accelerated motion, directed from the center of a planet.

3. Gravity force does not depend on mass of a body! The mass can be presented in three versions:

a) m_k –mass as amount of atoms.

b) $W_{m_A^e}$ - electronic-atomic energy in mass.

c) $W_{m_G^m}$ – mechano-gravitational energy in mass.

On the basis of the stated notions we offer to revise the essence of force not only in Coulomb’s formula, but in Newton’s formula too.

$$F = K\frac{q_1 \cdot q_2}{R^2} \qquad F = P\frac{m_1 \cdot m_2}{R^2}$$

It is well known, that mechanical energy can be bringing in electrostatic charge, where mechanical energy runs (turns) into energy of electric field, where

(mechanical energy) $F \cdot R \rightarrow \frac{E_2 \cdot V_{m^3}}{2}$ (electrical energy)

Similarly it is also possible to insert mechanical energy into mass of a body. As the result, the mechanical energy will turn into energy of gravitational field, where

(mechanical energy) $V \cdot F \cdot t \rightarrow \frac{g^2 \cdot V_{m^3}}{2}$ (gravitational energy)

Since the volume of the Earth is constant, the acceleration of gravitational field will be increased.

It should be logical to expect, that when removing the mechanical energy from mass the inverse process will occur, that is to say acceleration reduction of gravitation field will occur.

In his works I. Newton affirmed about existence of two spaces:

The Absolute space - is an immovable non-rotatable space, which represents a limited cube, with our planet in the center.

The Relative space – is a movable space. It can move with acceleration in the absolute space.

Editor’s note: In aether conception this means two parts of aether: some part is involved into the motion with the mass, but another part of aether is immovable.

The main mistake in search of aether consisted in the following: Maikelson’s experiments were aimed on search of relative velocity between bodies and space. However, it was the **relative acceleration** between bodies and space that was necessary to search for.

To quote the conclusions of I. Newton: "Body can keep the quiescent mode or mode of rectilinear uniform motion ..." By this, he postulates, that the relative linear velocity between solids and space does not exist. But we know that for rotation it exists (the famous experiments with revolving pail of water).

The gravitational field is the accelerated "falling" relative space, which represents a spherical form. If relative space moves, thus the question appears: where does it move? There is only answer: it moves in the absolute cubic space.

In Einstein's theory there is notion of unified and curved space in gravitational field, but the contradictions appear here, and on concerning that N. Tesla writes: "Only by presence of force field it is possible to explain the observed motion of celestial bodies, but thus the hypothesis of curvature of space is not necessary. The whole scientific literature on this subject is futile and doomed on oblivion". [1]

The fact that gravitation is the accelerated moving relative space can be proved by observation of accelerated moving rocket, where the acceleration in rocket is equivalent to the acceleration in gravitational field. Accelerated movement of rocket is relatively, that allows speaking about either acceleration of rocket motion in immovable space, or accelerated motion of space in immovable rocket!

The anti-gravitational field is the relative space, which has accelerated motion from the center of a body (for example: rotating cylinder, Earth satellite and etc.) But it is possible to create the model of anti-gravitation without rotations. On the basis of analogy between mechanical and electric energy comes to conclusion that gravity between bodies does not depend on mass of the body, but on mechanical-gravitation energy, contained in this mass, which is possible to contribute or to extract from. **Therefore, this is the internal gravitation energy.**

Part II

The "Mass" can be considered as a measure of three different conditions of matter:

m_k - as a measure of amount of atoms, representing a "framework" or "container", in which two types of independent energies are concentrated.

$W_{m_A^e}$ - as a measure of electric energy, which can be either accumulated or extracted, and it have a "compressed" form.

An example of accumulation of electric energy in mass is a big cylinder, rotating with linear velocity, close to velocity of light, in this cylinder the mass of electric and magnetic fields of atoms increases. There are another possible ways to contribute and to extract the said energy from mass.

And finally W_mM_G is the mass, which can be a measure of mechanical energy, or it can be either inserted or extracted from the matter (it can be identified as the gravitational mass). This gravitational mass is what we put our attention on, because it affects upon gravitation and it is able to create antigravitation.

"In his time N. Tesla worked on more general problem, which is the problem of matter and energy. And he has found, as he believed, the new physical principle, on the ground of which he brought forth his gravitational theory that was named dynamic gravitation. But he did not tell about it until almost the end of his life". [2] **Really, dynamic gravitation is the energy of motion.**

Let's take the following indications:

V – mechanical velocity
F – force
t – time.
In this case the product $W = V \cdot F \cdot t$ has the dimension of energy. Hereinafter, let's take

I – strength of electric current
U – difference of potentials
t – time.

Then $W' = I \cdot U \cdot T$ has the dimensional of energy. Thereby, $W \sim W'$ that is to say, the following products are accepted as equivalent:

1. $V \cdot F \cdot t \sim I \cdot U \cdot T$

2. In previous materials it was reported about untraditional way for accumulation of energy, under the condition, in which at constant current I the product $q = U \cdot t$ will depend on amount of inserted energy in unchangeable circuit L = const, in which the energy can be accumulated by untraditional way not only in electric capacity, but also in inductance.

Similarly the energy can be accumulated by untraditional way in a moving body, under the condition V=const and m_k=const (the product $qr = F \cdot t$ will depend on inserted energy and have unlimited value). **Exactly this charge will create the powerful gravitational fields.**

3. Let's take: F is mechanical force, R is distance. Then the product $F \cdot R$ has the dimensionality of energy. For the uniform electric field the product $E_a \cdot \frac{E^2 \cdot V}{2}$ also has the dimensionality of energy. In this case E_a is constant, E is intensity of electric field, V is volume. Thereby, $F \cdot R \sim E^2 \cdot V$. Similarly, $V \cdot R \cdot t \sim g^2 \cdot V$

We have received the correlations of resemblance for heterogeneous physical values, on the ground of which the following physical experiments can be offered:

On the grounds of the above-mentioned analogies it can be assumed that the accumulation of "compressed" energy is possible in mechanics, as well as in electricity. Since the velocity is relative, that the mass can have zero velocity relatively a observer, who moves with this mass, but the force field will remain unchangeable, since it depends on already invected mechanical energy.

Let's note that:
When the "compressed" electric energy is accumulated, the power field does not change.

When the "compressed" mechanic energy is accumulated, the power field increases.

Now we have come to the amazing conclusion that the gravitational force does not depend on mass of matter, but it depends on mechanic energy, which is included in this mass. This energy is unstable and at contact with land it is disappearing, and at zero gravity it can be saved for a long time.

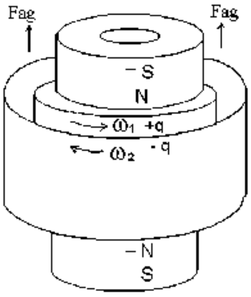

Fig. 1 The first way to obtain the antigravitational force.

1. The magnets are not revolved.
2. The cylinders of charged capacitor are revolving in different directions.

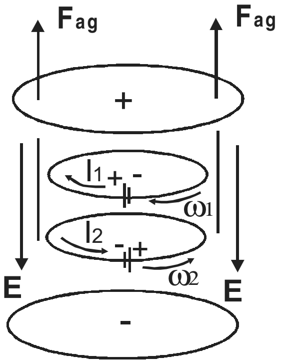

Fig. 2 The second way to obtain the antigravitational force.

1. The capacitor plates are charged and not revolving.
2. The current circuits are revolving in different directions.

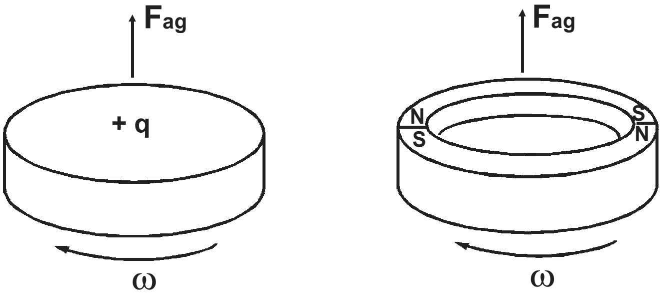

Fig. 3. The third way to obtain the antigravitational force

a) electric
b) magnetic

1. The disk and the ring are made from electrical current conductive material.
2. When these disks rotate, the currents, which emit the mechanic-antigravitational energy in the manner of heat, are formed there.

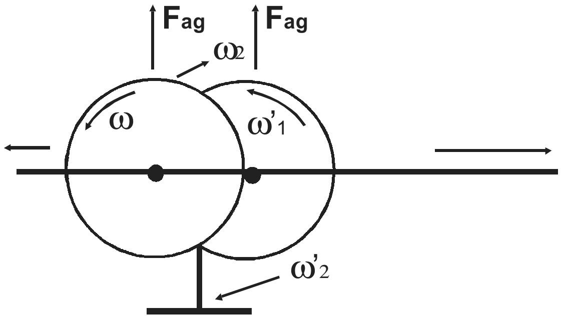

Fig. 4. The fourth way to obtain the antigravitational force.
Mechanical method.

1. This is an extraction of energy from matter. It was reported in details on the 10 international symposium in Volgo-Donsk, Russia.
2. The difference with electric circuits is that it is possible not only to extract the mechanical energy, but also to insert additional energy in the system.

References

1. Magazine Inventor and Rationalizator, Russia, #9, 1979.
2. Magazine Inventor and Rationalizator, Russia, #9, 1979, p.28

The Capacitor, which has the Energy of an Atomic Bomb

(Review of Anatoly K. Gaponov's research by Eugenie and Marina Golomolzins)

Is it possible to place a pail of water into a one-liter jar? At the first look the answer is obvious: certainly not! However, the inventor from Novosibirsk, Anatoly Gaponov thinks differently. He does not "press" water, but electrical energy, placing an energy equivalent of atomic bomb into ordinary electrical capacitor.

Everybody using ordinary batteries knows its defect: they need frequent recharging. Gaponov's capacitor is slightly smaller than a matchbox. Just come home by electrical automobile, take out the capacitor from engine, and then put it into the pocket. For home needs you can just insert the capacitor into plughole to power the light, boiler, and TV system. In general, each electronic device can have its own capacitor, then an electrical wiring is not necessary. After one

On the grounds of the above-mentioned analogies it can be assumed that the accumulation of "compressed" energy is possible in mechanics, as well as in electricity. Since the velocity is relative, that the mass can have zero velocity relatively a observer, who moves with this mass, but the force field will remain unchangeable, since it depends on already invected mechanical energy.

Let's note that:
When the "compressed" electric energy is accumulated, the power field does not change.

When the "compressed" mechanic energy is accumulated, the power field increases.

Now we have come to the amazing conclusion that the gravitational force does not depend on mass of matter, but it depends on mechanic energy, which is included in this mass. This energy is unstable and at contact with land it is disappearing, and at zero gravity it can be saved for a long time.

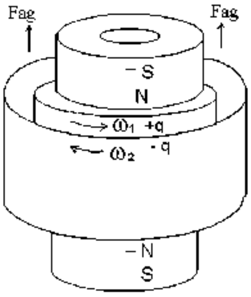

Fig. 1 The first way to obtain the antigravitational force.

1. The magnets are not revolved.
2. The cylinders of charged capacitor are revolving in different directions.

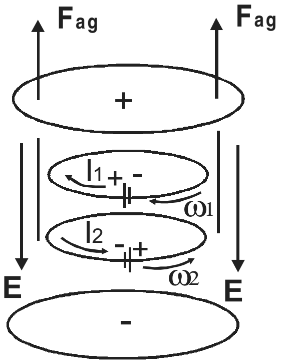

Fig. 2 The second way to obtain the antigravitational force.

1. The capacitor plates are charged and not revolving.
2. The current circuits are revolving in different directions.

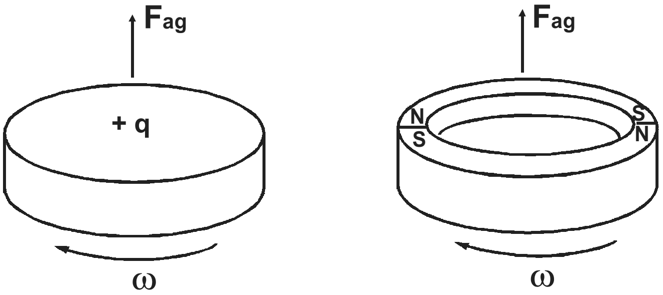

Fig. 3. The third way to obtain the antigravitational force

a) electric
b) magnetic

1. The disk and the ring are made from electrical current conductive material.
2. When these disks rotate, the currents, which emit the mechanic-antigravitational energy in the manner of heat, are formed there.

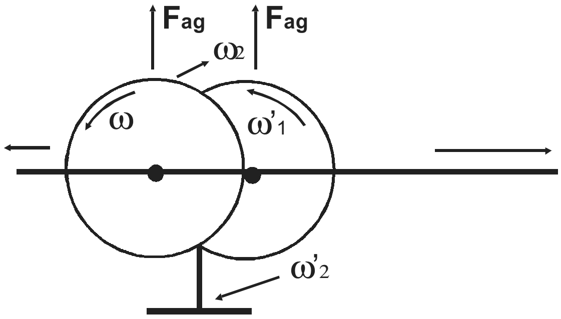

Fig. 4. The fourth way to obtain the antigravitational force. Mechanical method.

1. This is an extraction of energy from matter. It was reported in details on the 10 international symposium in Volgo-Donsk, Russia.
2. The difference with electric circuits is that it is possible not only to extract the mechanical energy, but also to insert additional energy in the system.

References

1. Magazine Inventor and Rationalizator, Russia, #9, 1979.
2. Magazine Inventor and Rationalizator, Russia, #9, 1979, p.28

The Capacitor, which has the Energy of an Atomic Bomb

(Review of Anatoly K. Gaponov's research by Eugenie and Marina Golomolzins)

Is it possible to place a pail of water into a one-liter jar? At the first look the answer is obvious: certainly not! However, the inventor from Novosibirsk, Anatoly Gaponov thinks differently. He does not "press" water, but electrical energy, placing an energy equivalent of atomic bomb into ordinary electrical capacitor.

Everybody using ordinary batteries knows its defect: they need frequent recharging. Gaponov's capacitor is slightly smaller than a matchbox. Just come home by electrical automobile, take out the capacitor from engine, and then put it into the pocket. For home needs you can just insert the capacitor into plughole to power the light, boiler, and TV system. In general, each electronic device can have its own capacitor, then an electrical wiring is not necessary. After one

or two years you will just have to come into electric service station and charge your magic capacitor like a gas balloon. Meantime, this research work began from hypnosis.

Anatoly Konstantinovich Gaponov (by birth from Kaluzhskaya region) is Ziolkovsky' countryman. In his youth Gaponov was brought by fate into Sakhalin, where he showed hypnotic abilities. As an inquisitive person Gaponov had organized a research group, started experiments and soon he understood that human brain had incredible possibilities.

A mental prick was made distantly to the hypnotized man, and he uttered a cry of pain. The ability to see people through, to define and to avoid organism's faults, was revealed in a hypnotic trance. It was possible to inspire pleasant emotions, to force "watching" a film on the given subject, as if on the screen. An uneducated person became an erudite, as if being connected to a certain global information database. Thus an idea to make an amazing experiment was appeared.

In one of the experiments Gaponov hypnotized the person with four classes education, and asked, if it was possible to transmit the electric current without wires? The hypnotized person gave the answer, that it was possible. For that it is required to convert the electric energy in x-ray radiation. And what afterwards? Afterwards it is required to focus that rays. By what? By the lens made from quartz glass, gold coated. It was a miracle! The person told about things that in usual condition he had no idea of! The information was received from somewhere outside.

Further quite an amazing thing has occurred. Gaponov asked the hypnotized person, if it was possible to intensify the abilities of hypnotist's brain?

He answered, that he could. "He turned me round and stared at the back of my head, - recalls Anatoly. - And suddenly the smile began to tear my mouth. I could not do anything with myself. When my mouth was sprawled literally from ear to ear, the hypnotized person in some inhuman voice declared that experience could not be continue since the cerebral hemorrhage would occur. I was hardly able to give the order to stop the experiment".

Thereby, the experiments with hypnosis gave the beginning to the thirty-years period of inventions in the field of accumulation and transmission of energy. After the return to native Kaluzhskaya region, Gaponov was occupied with physics, development of logical thinking and became the town champion in chess.

The necessary books fell into his hands by themselves: some time a certain acquaintance gave it to read; another time he found the last copy in a bookstore. As a majority of self-taught inventors, Anatoly preferred practical experimentation. In quest of laboratory for realization of his own ideas, he moved to Novosibirsk. **As a result, in 1980 Gaponov has made experimental system for compression of energy.** From the school Physics we have known the notion of "electric arc" – it is a small blue lightning between two electrodes. Gaponov has tamed this lightning in such a way, that having drawn apart two wires, which executed the role of electrodes, by hands and got the arc by length up to half meter. Anatoly confirms that in principle, it is possible to create an arc of any desired length under any amperage.

One of the experiments found out one more enigmatic characteristic of electric discharge. During electric photography of arc a person happened to be between the camera and the system. On typing pictures, the researchers have found with surprise, that the electric arc was perfectly seen through the person. That is to say, it created the invisible field, for which material object was not an screening obstacle, and which was fixed on the film.

The further experiments with electric arc have allowed to get a new source of energy, as well as to open the possibility of setting light and sound on fire! Just imagine, you ring up a bell, it's sound waves spread at once in all directions, and then flash up with bright blaze.

(Editor's note: this experimental facts are rare modern evidences of possibility to create longitudinal electric waves. It is clear analogy here with sound waves in air since they are longitudinal waves also. Alexander V. Frolov)

When the problem of energy source was solved, Gaponov turned to the problem of energy accumulation. According to Gaponov, he has provedexperimentally the possibility of charging of an ordinary capacitor with any amount of energy. This statement sounds paradoxically: how it is possible to place the unlimited amount of contents in limited volume? However, this is not a simple way.

Gaponov believes that energy "placing" occurs not in space, but in time by means of his system! In what way? Imagine, that you fill one-liter jar with water. But already after an instant the water-filled jar is in past, and that present one is once again ready to be filled. And so ad infinitum. Water as if it fills a certain "time reservoir", and a jar is just a neck of this "time reservoir".

(Editor's note: This method is described in other articles also but usually it is pure mathematical discussion about Minkovsky space-time and theoretical proposals. Gaponov's experiments are realization of fantastical idea to take power from the time flow, i.e. from Past or from Future to get over-unity in Present space. Alexander V. Frolov)

"It is possible to demonstrate one more example, - Anatoly Gaponov adds. – Let's charge the capacitor with the expectation, that it will supply the light bulb

for one second. Thereby, on the Earth this light bulb will be on only for an instant. But if the same capacitor with light bulb is placed in rocket and dispersed around the Earth at the velocity, closed to velocity of light, time on the board of rocket will be so slowed that the light bulb on rocket will be glowing infinitely long for an observer from the Earth. It means, that in any case it is the same energy quantity, but in one case it's action is sprawling for a second, and in another one it is sprawling for eternity! It is possible to say, that in my system I have created the condition corresponding to this hypothetic rocket".

The system for accumulation of electric energy could be charged by ordinary wall plug 220 VAC. Time period of charging is different and depends on the certain scheme of the system. By the way, sea electric slopes are the certain natural analogues of such capacitor. Some elements of internal device of these sea creations reminds the "pump" elements for placing of electric energy into "temporal jar".

Finally, the third Gaponov's invention is the system for transmitting of energy without wires. As well as in two previous cases, there is an experimental device. Anatoly Gaponov speaks that he has succeeded in getting the essence of experiments for transmitting of energy, which were conducted by Tesla.

It is clear, that the main advantage of this method is an absence of wires and losses of electric energy. The electricity could be transmitted directly into any point, where receiving equipment placed, let say from Kaluga to Sahara. However, this is not so interesting for anybody, since for the present day Anatoly Gaponov's inventions don't have demand.

" The first system was created twenty years ago", - says Mr. Gaponov. – "Now I am fifty five, but things have not budged an inch". He adds dreamily: "Eh, if only I had a laboratory and some money...".

Gritskevitch's Hydro-Magnetic Dynamo

Oleg V. Gritskevitch

RUSSIA, 690002, VLADIVOSTOK,
Okeansky prospect, 99 - ap.112
phone/fax: (7-4232) 424-674 Email: ogri@mail.primorye.ru
Russian Academy of Energy and Information,
Russian Academy of Natural Sciences

Editorial: The article presents construction and operation of Oleg V. Gritskevitch's hydro-magnetic dynamo, which is an example of very powerful new energy system. The prototype in Armenia has been produced over 1500 KWtts power during several years.

The author was born on 14 August 1936 and grew up in Vladivostok, Russia. He is married and has a son Boris. Gritskevitch is a physicist by education. He worked in the Far - East branch of the USSR Academy of Sciences. Since 1985 he has been working independently as an inventor. He has more than 70 patents on inventions ranging from household engineering up to high technologies, which he has been trying to apply in our country and met big difficulties. After numerous attempts to receive the patents the author was convinced that outflow of the information occurred. Therefore he has received the state certificates as on know-how (on a French way of patenting), for all his inventions.

Introduction

During the Institute for New Energy 1999 Symposium, I lectured on my hydro-magnetic dynamo. This paper is my attempt to explain the construction and operation of my dynamo.

To fool investigators of my secrets, I have an occasion provided misleading information. For example, the drawing accompanying the Russian patent referenced below shows a cylinder across the toroid to fool readers. The real dynamo only has the toroid without the cylinder. Even its name "hydro-magnetic dynamo" is somewhat deliberately misleading.

I have some familiarity with the new energy field. Nearly all purported new energy devices are fairly small electrical generators. The dynamo may be the only new electrical generator which most nearly meets all the requirements of an ideal large-scaled electrical generator. My dynamo really is the single most valuable invention the world has ever known.

Alexander V. Frolov of St. Petersburg recommended me to contact with Dr. Patrick Bailey, Institute for New Energy since Pat has lots of contacts who could possibly help me with patenting my invention of a new source of energy in USA.

I conducted the work on the theory and creation of the electrostatic generator-converter «Hydro-magnetic dynamo» about 20 years. (See dynamo history below.) The first primitive equipment was created when I worked in Academy of Sciences. During that time various changes were introduced in the generator and in the theory of its work. It is now possible to manufacture, install, and apply it in industry.

For the first time I made the public report on this work in 1991 on a symposium in Volgodonsk city. The report received the positive replies and reviews of the experts of a nuclear industry in USSR. The same year I was accepted in International Nuclear Society. In these years I offered development of this technology to different state bodies and private enterprises. But there was the only answer: "It is very interesting and perspective project, but there is no money for it".

NEW ENERGY TECHNOLOGIES #7

1. Gravitonics is Electronics of the XXI Century, Spartak M. Poliakov, Oleg S. Poliakov
2. Experimental Research on Gravitational Propulsion System. Review
3. Beamship Technology: A Re-working of Early 20th century Discoveries, Russell Anderson
4. Commercial Antigravity, Tim Ventura
5. Action without Reaction. New Gravidynamic Paradox, Yuri N. Ivanov
6. New Possibilities of Vortex Electric Power Devices, Stanislav A. Lisnyak
7. Investigation of Electric Energy Transmission Processes in non-Metallic Conducting Channels, D.S. Strebkov, A.I. Nekrasov, S.V. Avraamenko
8. Microcosm – Universe - Life, Alexey G. Shlienov, Ernest L. Petrov
9. Hypothesis of a Theory of Everything, Jack P. Gibson
10. The Charge and Mass of a Photon, Dzabrail Kh. Baziev
11. The Homopolar Motor: A True Relativistic Engine, Jorge Guala-Valverde, Pedro Mazzoni, Ricardo Achilles
12. Weight Reductions Generated by Bucking-Field Permanent Magnets, William C. Simpson
13. Advanced Nuclear Waste Decontamination Technologies, Mark Porringa
14. Experimental Investigations of the Radioactive Isotope Half-Value Period Changing in the Local Volume ofCause-Effect Relations, Igor A. Melnik
15. Tilley Electric Vehicle
16. The Problem of Time: Force as the Cause of Change in the Course of Time, Valentin P. Oleinik
17. Time and its Physical Relationships, Andrew Michrowski
18. Time as Phenomenon of the Expanding Universe, Michael H. Shulman
19. The Experiments on Thermo-Gravitation. Review
20. Letters to the Editor from Hector D. Perez Torrez

New Energy Technologies

Published by Faraday

Issue #4 (7) 2002

Research on advanced space propulsion methods and new energy systems

History of invention: Poliakov's vortex drive

Patented by Research Institute of Space Systems named by Krunichev, Russia

Also in this issue:

- * New gravidynamic paradox
- * Experiments on weight reduction
- * New possibilities of vortex energetics
- * Experiments on homopolar motor
- * Physical relationships of time

Gravitonics is Electronics of the XXI Century

Hypothesizes, Conclusions, Speculations

Spartak M. Poliakov,
Oleg S. Poliakov

60-let SSSR str., 1-167, Friazino,
Moscow area, 141120, Russia
Tel: (096) 564-65-67

(This article was published in "ELECTRONICS: Science, Technology, Business" magazine 5/2000 p. 8-13)

Today we can easily insist that ways to solve main gravitonics problems are already defined, at that the practical realization of any of them will mean the break through in engineering. So, what successes has the Russian science already achieved in the area of gravitonics, and what priorities can we lose here in the nearest future?

Introduction to Gravitonics

The electronics of the "past century" uses electron as a ball, which has weight, radius, electric charge and magnetic moment. These very parameters define electron behavior in the electrostatic, magnetostatic and electromagnetic fields. But electron abilities are not limited by it; electron spin and internal microstructure features remain unaccounted and unclaimed. Evenly speaking, General Theory of Relativity (GTR) of Einstein was a power impulse for mathematical physics development and gave birth to many productive ideas. But the main problem, that is the secret of gravitation, remains undisclosed… In works of K.P. Stanjukovich [1] and A.Z. Petrov [2], who carefully followed GTR, it was shown that this theory described neither energy, nor impulse of gravitational radiation, i.e. it can not explain gravitation. About 20 years ago V.B. Braginsky, today's RAS Corresponding Member, came up with an idea: "if the propagation speed of gravitational signal is higher than the velocity of light, there will be already another theory, not GTR!" Maybe, the reason is the postulation of equality of gravitational and electromagnetic radiation velocities?…

Today the approximate theory of gravitational radiation sources can be built on the basis of the following simple considerations: if during annihilation of "electron-positron" pair there creates the pair of gamma-quantums with energy about 0.511 MeV, then the pair of back gamma-quantums with the energy about 0.511 MeV, could create "electron-positron" pair. Is it possible to assume, that electron, positron and gamma-quantum with the energy about 0.511 MeV are just three stages of one and the same object?! If it is possible, then for the rational description of the given object we will have to suppose the existence of subparticles, named by us uniquantums [3], or named by other authors microleptons [4].

On the basis of Heisenberg uncertainty relation conformably to the energy and duration of quantum of electromagnetic radiation, measured by laboratory means, it is possible to calculate the minimal "electrical length" of photon (i.e. quantum geometrical extension in free space in wave-length units), which is equal to 137λ, and in the uniquantum theory it is equal to 137 uniquantum-antiuniquantum pairs. On the basis of these conceptions it is possible to construct the spatial microstructure model of electron. So, what kind is it?

We think, that electron can be represented as thin-walled spheroid, walls of which are two light (C) barriers, separating the "internal" part of electron from the "external" one. From the traditional physics point of view "over-barrier" space is an "imaginary" one. This very space can contain the gravitational mass of electron. The radius of the gravitational spheroid is equal to the half of the classic electron radius, and its imaginary weight is 137 times more than the rest mass of electron. Being "cut" off by the double light C-barrier, uniquantums of the spheroid internal part are as if non-existent for the outer world, and the rest mass of electron is formed by magnetic energy of three uniquantums on the external orbit with the classic electron radius. This very spheroid, rotating with the tangential velocity C, let us get the precise value of the electron spin.

The study of presented model shows, that:

- The "electromagnetic" rest mass of electron is "magnetostatic";
- The gravitational mass of electron is an imaginary value and it is 137 times more than the rest mass of electron;
- The gravitational radius of electron is two times less than the "classical" one;
- The "internal" gravitational radius of electron is 45.7 times more than the external one, i.e. the internal space is compressed per 45.7 times (!);
- The spin is equal to the classical one, but this value is imaginary one (!);
- The value of the "effective" electron charge is three times more than the classical tabulated value;
- The native magnetic field of electron is equal to $8{,}9 \cdot 10^{13}$ Oersted;
- The gravitational constant is equal to 10^{33} $cm^3/g.s^2$, i.e. it is about 10^{40} more than the "world" gravitational constant of the Earth;
- The gravitational energy of electron is equal to $137 \cdot 0.511$ MeV, i.e. 137 times more than the equivalent energy of the rest mass of electron.

The model is paradoxical. But it can be tested experimentally! Comparing "electromagnetic" rest mass of electron with the relation of electromagnetic energy to gravitational one, it is possible to determine the connection between magnetostatic and gravitational energy of electron, and, therefore, with energy of the magnetized ferromagnetic.

Gravitational constants of the Earth and of the electron differ in about 10^{40} and can be described by the same simple equation:

$$\gamma_{loc} = 2k\gamma_0 \cdot \omega^{3/2},$$

where k - is a parameter of the gyroscope shape, γ_0 - is absolute universal constant, equal to 1/137 and ω - is native rotation frequency of the gyroscope.

Let us assume, that gravitational constants of all objects should be described by this equation. By substitution of the new gravitational constant into the known equation of the gravitational energy $W = \gamma m^2 / r$ we will get the equation of the gravitational energy of rotating gyroscope with any size (from electron up to the Galaxy!). Thus, the main point of the "non-Einsteinian" theory of gravitational energy sources comes to the thing that any rotating object and any magnetized ferromagnetic have their own gravitational energy, and the sources of gravitational radiation can be only nonlinearly moved objects, or objects which are in the state of change of phase (for example, permanent magnet during its demagnetization). It is the gravitational theory and explanation of "strong" and "weak" interactions!

Laboratory test of the equations

Magnetostriction

J.P. Joule found the effect of change of ferromagnetic linear sizes and volume during magnetization as early as 1842. Magnetostriction is widely used in modern technique, but in the physical encyclopedia of 1963 there is the following honest acknowledgement: "For the most ferrites both longitudinal and transverse magnetostriction is negative; the reason of it is still unclear."

In the scientific literature magnetostriction is usually defined as $\lambda=\Delta L/L$. However, during the change of external field to some arbitrary and enough small value ΔH, it is advisable to define magnetostriction as $\lambda=1/L\cdot\Delta L/\Delta H$, since in magnetostriction experiments the value $\Delta L/\Delta H$ (or $\partial L / \partial H$) is changed. By means of the suggested equation $W=137(BHV)=BHV/\alpha$, which connects magnetic energy with the gravitational one, it is possible to get enough simple equation for the magnetostriction:

$$\lambda = 1/L \cdot \partial L / \partial H = \alpha \cdot k / (B \cdot H)_s \cdot H^2 \cdot \partial \mu / \partial H$$

where $(B\cdot H)_s/\alpha$ - is the density of gravitational energy in the point of magnetic saturation, k – is the parameter of share of gravitational field in the magnetostriction effect, H - is magnetic bias, $\partial \mu / \partial H$ - is differential magnetic conductivity.

The new equation qualitatively corresponds to four known features of magnetostriction [5], namely:

- The magnetostriction sign is defined by the sign of $\partial \mu / \partial H$, i.e. by the course of the magnetization curve, measured in the direction of calculated component of the linear magnetostriction;
- Graphical sum of three linear components of magnetostriction, calculated by three main axes of the anisotropy form of the model, is always negative and numerically close to the value of the volume magnetostriction;
- Magnetostriction is an even effect, since the equation includes squared value of the external magnetic field;
- Dependence of magnetic conductivity μ from the filed H and hence dependence $\partial \mu / \partial H$ has a hysteresis nature. Therefore, the magnetostriction is a hysteresis phenomenon too.

So we have the right to "close" the question of physical encyclopedia on the cause of magnetostriction. **Magnetostriction is the secondary gravitational effect of ferromagnetic "self-constriction" in its own gravitational field.**

Gravitational-optic effects of GTR

Distortion of the light beam, passing near the Sun and the photon frequency bias in the field of terrestrial gravity (the Nobel experiment of Paunda and Rebki) are the main arguments in favor of GTR canonization. It is very attractive to repeat these experiments in laboratory conditions, basing on our conception of the origin of gravitational field.

The acceleration of gravity, used in experiments with ferromagnetic, reached the value $4.72\cdot10^{15}$cm/s^2, i.e. about $4.8\cdot10^{12}$g. At such values of acceleration there is no necessity to introduce a definition "space masses". In these experiments there was used the optically transparent ferromagnetic, which was the saturated solution of manganese chloride in water at room temperature. The experiment on the beam distortion was made in 1975 [7]. It was shown, that this effect is the result of two simultaneous processes. The first is an intense drift of magnetic ions, which forms the gradient of index coefficient that causes the light beam distortion. Another process is a relatively weak gravitational beam distortion, for which, nevertheless, the relation of deviation angle to the track length (the length of the dish is about 100 mm) is turned out to be about 10^{10} more than in "Einsteinian" gravitational-optic experiments.

The experiment on bias of the optic radiation frequency [8] was made in 1978-1980 and was repeated in 1983. With use of heterodyne and interferometrical methods of measurement we were succeeded to observe effects of "red" and "blue" frequency biases in the non-uniformly magnetized ferromagnetic by means of simple displacement of the working dish (with the length about 40 mm) from one side of the magnet gap to another. The maximum displacement is about 10^{-5}, that is about 10^{10} more than in the experiment of Paunda and Rebki.

Problem of the propagation speed of gravitational radiation

There are still only few publications about such fundamental parameter as the propagation speed of gravitational radiation; it is able to speak only about pages, or even lines! Let us refer to major sources:

I. Newton: *"The propagation speed of gravitational interaction is equal to infinity."* It is an argument, because otherwise we would have to bring the "delay" parameter Δt into the Law of Gravity, what is not noticed in real conditions of star observations [9].

P.S. Laplace in 1787, taking into account observation errors of that time, showed, that gravitational interaction speed was about $50 \cdot 10^6$ times more than the light propagation speed, i.e. it was about $1{,}5 \cdot 10^{18}$ cm/s [10].

A. Einstein: *"The propagation speed (of gravitational interaction) is equal to the light velocity".* This statement is postulated.

Even during the change of propagation speed of gravitational radiation between the Earth and the Moon it is impossible to define the signal delay about 10^{-11} s, i.e. we cannot measure directly the propagation speed of gravitational radiation (supposing that we have both generators of the gravitational radiation and receivers of it). But this speed can be estimated by the reflection impulse, what exactly was made in 1987 [3]. And its value is about $9 \cdot 10^{20}$ cm/s!

On the basis of conservation law of impulse of unidirectional radiator with arbitrary energy type we can get a simple equation:

$$F/(dW/dt) \cong 10^4 \cdot V/C^2 \qquad \text{[g/Wt]}$$

where V – is the speed of radiation propagation, F – is tractive force in grams, dW/dt – is power of radiation in Watts and C – is velocity of light.

For making the experiment there were constructed, produced and adjusted: sensible scales with one degree of freedom (sensitivity of balance is about 1g at the oscillator mass together with the moving element of scales which is about 50kg); the indication system of small mass changes (phase-meter receiver); gyroscopic system, changing the mass in the dynamic mode (there are 16 possible operating modes – from the rotation with steady and variable angular speed up to the forced precession with the variable angle of precession, with the "right" and "left" rotation of all load-bearings elements at option); power sources and commutation automated system. The period from idea up to its realization took about two years (1985-1987) [3]. Taking into account the real parameters of the system, the program of calculation was drawn and propulsive burns were calculated. The results of machine computation can be compared with real impulses, demonstrated on the screen of the oscilloscope.

If strange speed value 177 C^2 is discarded, then the middle speed value is close to C^2, i.e. to $9 \cdot 10^{20}$ cm/s! Of course, we would like to think that this is the second fundamental matter speed of our world, which we has approached experimentally …

Gravitational receiver

During the creation of gravitational antennas and receivers there appear almost insuperable difficulties from the modern fundamental science point of view. That is why it is advisable to look at this problem from another side. At first, it is necessary to consider gravitational radiation interaction not with the mass, which it goes through without losses, but with the gravitational field of independently gravitating mass, when the interaction must be the most effective because of the principle of physical processes reversibility. At second, it is necessary to choose some critical parameter of auto-gravitating receiver as a value, which is directly measurable by gravitational detector. For example, angular velocity of free rotation of thin disk with big diameter, the frequency of magnetization precession during NMR (nuclear-magnetic resonance) or NFMR (non-linear ferromagnetic resonance) etc. can be chosen as such a value.

In 1987 there was the first successful attempt to receive the gravitational impulse. The source of external signal was gyroscopic precessing system with the variable angle of precession (the propagation speed of gravitational radiation was measured by it). Double gyroscope, setting in motion by one electric motor, but with the opposite directions of rotation, was used as a detector. Between disks there was placed the source of light, impulses of which, passing through disks openings, were registered by photodiodes. Their signal came into differential circuit of data processing. The memory oscilloscope reproduced impulses of gravitational radiation. At that radiating system and memory oscilloscope was started up simultaneously. During the work process there appeared a problem of exciting of slow auto-oscillations of gyroscope-detector. This problem together with the low frequency of auto-oscillations of mechanical system led to a conclusion that this research direction is not very promising. However, the fact of detection was proved!

Gravitational engine of continuous action

Only about nine years passed since the appearance of the idea about engine up to its realization! In 1997 the engine was produced and tested. The engine with weight about 28 kg was made "weightless" on the magnetic hanger, and longitudinal draft, appearing in accordance with the impulse conservation law, was measured by micrometer detector of longitudinal shifts (sensitivity is about 50g/point). Such engine could be built still in the beginning of the last century… However, it has a secret **that is a gyroscope with the variable radius, working in the continuous mode.**

The engine power is defined by the formula

$$\frac{dW}{dt} = 5k\gamma_0 \omega^{3/2} \left(\frac{m^2}{r^2} \right) \cdot \frac{dr}{dt}$$

In June of 2000 there were made experiments with the model of gravitational engine, which represents a gyroscope with the variable radius (see photo on the 1st cover page). The mercury was used as rotating fluid. Tests were made in the Research Institute of Space Systems named by Krunichev. In three experiments, at a certain speed of rotating fluid there was fixed a decreasing of the engine weight (38,5 kg) up to 1.0-1.5 kg (2-3%). The specific impulse of the engine was equal to 2.5-3.0 kg per kilowatt of electric power. Analysis shows, that the increasing of propulsion force is possible at optimization of design and operating modes.

Some preliminary resume

In the magazine "Foreign Literature" #1, 1967 the article "For hundred years forward…" by Jack Marabini was published. There were made some conclusions about prognostic work of firm Rand Corp., including the area of gravitational technique. Namely:

- Development of communication facilities on gravitational waves in 2000;
- Creation of spaceships with antigravity engines in 2050;
- Transformation of gravitational energy into electric one in 2100.

In the article it was noted, that the most "fantastic" predictions of this firm, as a rule, come true passing ahead.

According to our crude estimations, the propagation speed of gravitational radiation is "C" times as much than the velocity of light, but we know neither laws of attenuation and propagation of gravitational waves, nor laws of their reflection and refraction, nor laws of their interaction with the substance… The large routine work is expected: making of measurements and investigations, tabulating of obtained data, publishing and society familiarization of the results, their "popularization". It is necessary to learn to use gravitational radiation and to protect oneself from its accidental influences, to design standards and dosimeters, etc., i.e. to repeat the way of radio engineering and nuclear physics comprehension.

For that we need generators and receivers of gravitational radiation. It means that the financial support is necessary. And engineers are sure to be ready to pay the highest price for the chance to give to the Mankind spaceships, systems of instantaneous communication with them and real perspectives for the very long history.

Conclusion

We have already passed the long way, if not in space, then in time. We have made:

- Gravitational engines of continuous action with the specific impulse about 2.5 kg/kWt [11];
- Transformers of gravitational energy into thermal and electrical ones [12];
- Communication system based on gravitational waves [13];
- Receivers of gravitational (microlepton) radiation of biological and mineral objects [5];
- Devices for control of "laboratory time" flow (time machine) [14].

The main goals of the authors were to attract readers' attention to the problems, which demand an urgent solution. Some questions were decided, and even seemed to be clear. It also seems to be clear what to do further. And what do you think about it?

References

1. Stanjukovich K.P. Dr. Phys.-Math. Sc. CTR and gravitation // Seminar of the Moscow nature investigators society. // M. – 1966.
2. Petrov A.Z. Some peculiarities of CTR // Seminar of gravitational chair of Kazansky SU. (Preprint / Ukraine Academy of Science) // Kiev – 1971.
3. Poliakov S.M., Poliakov O.S. Introduction into experimental gravitonics. // M.: Prometey – 1991.
4. Patent #2113000 RF. Method to search minerals by their own radiation, device for its realization and microlepton detector / / Ohatrin A.F., Ohatrin A.A., Sizov V.S. Priority from 21.07.1997.
5. Kirensky L.V. Magnetism. // M. Nauka – 1967, p.141.
6. Vonsovsky S.V. Magnetism // M. Nauka – 1971, p.404.
7. Poliakov S.M., Martynov V.F. Method of deflection and focusing of optic radiation. A.n. #2187534, 10.11.1975.
8. Poliakov S. M., Olihov I. M. Dorofeev V. A. and others. Displacement of optic radiation frequency in non-uniformly magnetized ferromagnetic // Proceedings of the Moldova Academy of Science. Series of Physicotechnical and mathematical sciences. // Kishinev – 1983, #2, p. 57-59.
9. Isaac Newton. Mathematical beginning of Natural Philosophy // Transl. Moscow-Leningrad – 1936.
10. Laplace P. S. Statement of the world system. // Leningrad, Nauka, vol. 4, p. 197-198.
11. Poliakov S.M., Poliakov O.S. The beginnings of experimental gravitonics // In: Proc. of Int. Conference "New Ideas in Natural Science" // S.P. – 1996, p. 529-536.
12. Potapov Ju.S., Fominsky L.P. Eddy energy // Cherkassy, Kishinev OKO-Plus – 2000, p. 387.
13. Akimov A.E, Shipov G.I. Torsion fields and their experimental manifestations // Proc. of Int. Conference "New Ideas in Natural Science" // S.P. – 1996, p.22-225.
14. Chernobrov V.A. Secrets of time // M., AST – 1999.

• •

About the authors:

Spartak M. Poliakov. Graduated from Kishinev State University. Profession is physicist-experimenter. He works in electronic industry for about 47 years. S. M. Poliakov is the author of more than 50 scientific works. One of his latest books is "Introduction into experimental gravitonics". Interests: microwave engineering, gravitational electronics, faster-than-light communication transformation of gravitational energy into electric one.

Oleg S. Poliakov. Graduated from Moscow Institute of Electronics. Profession is "Semi-conductor electronics". Interests: computer engineering, industrial gravitonics. He is a co-author of "Introduction into experimental gravitonics" and "Self-tutorial of computer work".

BEAMSHIP TECHNOLOGY

by Russell Anderson
russanderson3@hotmail.com

Read in this issue a report by Russell Anderson, USA, on beamship Model Flying Craft

www.americanantigravity.com

Publisher Faraday Lab Ltd
Editor-in-Chief Alexander V. Frolov, Editor Olga O. Leontyeva,
Designer Ekaterina F. Yakovleva
Scientific Advisory Board Kyrill P. Butusov

New Energy Technologies Magazine, is published by Faraday Lab Ltd, six times a year. Correspondence Address:
POBox 37, St.Petersburg, 193024, Russia
Phone: 7-812-380-6564, net@faraday.ru, www.faraday.ru

Subscription: $90 any foreign. Air mail delivery included
PRINTED IN RUSSIA Copyright © 2002 by Faraday Lab Ltd.
All rights reserved.

Experimental Research on Gravitational Propulsion System

Editor's: It is a review of the article by ***V. A. Menchikov****, the Director of Research Institute of Space Systems, named by Krunichev, Russia. The article was published in"Polyet" magazine #10,2001, p.38-39, Russia. It scrutinizes the matters on development of propulsion systems based on the unconventional approach to the problem of gravity, i.e. gravitational engines. It also cites the results of the gravitational engine model research made by means of the experimental facility, created in the Khrunichev Research Institute of Space Systems.*

The device, transforming rotary motion into unidirectional motion, looks like S.M. Poliakov's one. It also operates with rotation of liquid, which causes the propulsive force. Truly speaking, Poliakov had an agreement with Research Institute of Space Systems named by Khrunichev in 2001. Some funds were assigned to develop the device, however the project, into which Poliakov had put a lot of work, still remains unrealized. Besides, the scientist's name is not even mentioned in the patent.

***

Scientific and technological advance opens to mankind more and more wide abilities to use space for the solution of global problems. In many respects the complete realization of these abilities will be defined by the developement of means used for delivery of payloads into the space. In the XXI century the dominating use of reactive chemical and electrical propulsion systems in rocket-space technique as well as low application level of engines with other physical principles can be the factor of an "inhibitory" influence for the development of such techniques. It is caused by the fact that created rockets practically do not correspond to noticeably increased standards of safety, operating costs, costs for transport operations execution and ecological influence on the environment.

Thus, there becomes to be urgent the problem of development of alternative approaches towards the creation of propulsion systems, made for the rocket-space technique on the base of unconventional ideas and engineering solutions. A rather old-established idea of creation of gravitational engine should be concerned as one of such ideas. It is based on the unconventional approach to the problem of gravity. Nowadays many countries take part in solution of the gravitational problem, namely Russia, USA, Japan, etc., and if till recently only some scientists and inventors showed the interest to this problem, then now it arouses interest of research-and-production majors. Unfortunately, now it is not possible to speak about sufficient theoretical or practical development of this idea. However, the interest is so considerable, that practically separate experiments on this subject were made earlier and they are still made nowadays. After all, stakes are very high and are defined by applied nature of the problem (the ability to create qualitatively new engines for the rocket-space technique), as well as by its scientific significance.

One of the directions to solve the problem of the creation of gravitational propulsion systems is the realization of associated theoretical and experimental methods of the search of physical processes, leading to the antigravity effects appearance, which cannot be adequately described by existed theoretical conceptions. V. Shauberger's patent, based on the postulate of gravitational energy radiation by "disturbed rotating mass" can be considered as an example of such practical realization. Taking into account a number of known experimental results, a model of gravitational engine and experimental system for estimation of this model parameters were made in Research Institute of Space Systems named by Khrunichev to provide the practical realization of Shauberger idea (Fig.1). It is the metal construction, which provides the model displacement in upward direction with the ability of its rotation around vertical axis.

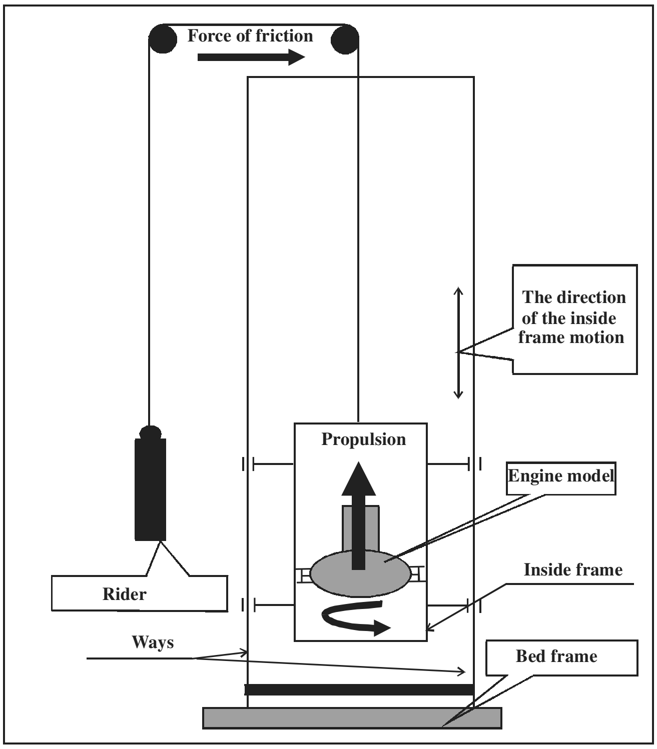

Fig.1
Schematic circuit of the experimental system

Later, to increase the system sensitivity it was improved and the block system of suspension was replaced by the lever frame.

Turning angle of the stand frame, where the model of gravitational engine is suspended, depends on the following: weight and geometry characteristics of the frame; weight characteristics of the engine (of counterweight); engine propulsion and frictional forces in bearings. Laser indicator of frame turning angle and vertical ruler let increase the gauge of lifting height of the gravitational engine model proportionally to the arm of light beam.

On using this system there was made a wide cycle of tests, which qualitatively confirmed the presence of propulsion force. Tests were recorded by video. The analysis of experimental results shows, that acting time of propulsion was about 12 s at each switching. During repeated switching of the gravitational engine model in different conditions it is able to create the propulsion, the value of which at 40…50 s of operation can be about 3 standard units of propulsion force (1 standard unit of thrust is about 10gs), and while using the powerful electromotor it can be about 80 standard units of propulsion at the intervals up to 4 s.

Research Institute of Space Systems named by Khrunichev, works on automation of experimental researches and on development of laboratory resources for factor analysis of appearance of the propulsion vector with the usage of the described model of gravitational engine.

Patent

The patent was published in the Bulletin of Patent Information in 2001.
(11) 20946
(51) 7 F 16 H 27/00
(21) 2001121237/20
(24) 01.08.2001

Editorial: V.A. Menchikov together with A.F. Akimov, A.A. Kachegan and V.A. Svetlichnyi have got this patent. Dr. Spartak M. Poliakov, being the author of the principle, is not mentioned at all in the patent.

NEWS REVIEW

Boeing Tries to Defy Gravity

According to Jane's Defence Weekly (UK), http://www.janes.com, Boeing, the world's largest aircraft manufacturer, has admitted it is working on experimental antigravity projects. These projects are able to overturn a century of conventional aerospace propulsion technology and alter the entire aerospace business. Boeing uses researchers by Yevgeny Podkletnov, who claims to have developed a device, which can shield objects from the Earth's gravity. Many conventional scientists, who have not been able to reproduce Dr Podkletnov's results, view his project, named «GRASP» (Gravity Research for Advanced Space Propulsion) with suspicion.

Dr Podkletnov claims to have countered the effects of gravity in an experiment at the Tampere University of Technology in Finland in 1992. The scientist says he found that objects above a superconducting ceramic disc rotating over powerful electromagnets lost weight. The researches have shown that the reduction in gravity was small, about 2%, but the implications - for example, in terms of cutting the energy needed for a plane to fly - were immense.

His devise, named "impulse gravity generator" is capable to produce a beam of "gravity-like" energy that can exert an instantaneous force of 1,000g on any object – enough, in principle, to vaporize it, especially if the object is moving at high speed. Laboratory installation has already demonstrated the 4in (10cm) wide beam's ability to repel objects a kilometer away and that it exhibits negligible power loss at distances of up to 200km.

Applications of the device can include space launch systems, artificial gravity on spacecraft, aircraft propulsion and "fuel-less" electricity generation ("free energy"). However, observers say that Podkletnov's device could be engineered into a radical new weapon, for example, adapted for use as an anti-satellite weapon or a ballistic missile shield.

Documents, obtained by reliable sources, show that Boeing is taking Dr Podkletnov's research seriously. It is also possible, Boeing admits, that "classified activities in gravity modification may exist". The paper points out that Podkletnov is strongly antimilitary and will only provide assistance if the research is carried out in the "white world" of open development.

Boeing is the latest in a series of high-profile institutions trying to replicate Dr Podkletnov's experiment. The military wing of the UK hi-tech group BAE Systems is working on an anti-gravity programme, dubbed Project Greenglow. The US space agency, Nasa, is also attempting to reproduce Dr Podkletnov's findings, but a preliminary report indicates the effect does not exist.

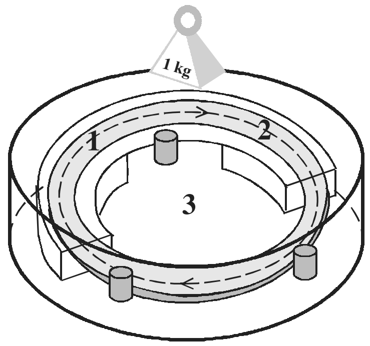

1. Solenoids create magnetic field
2. Spinning, super-conducting ceramic ring
3. Liquid Nitrogen acts as coolant

BEAMSHIP TECHNOLOGY

by Russell Anderson
russanderson3@hotmail.com

Read in this issue a report by Russell Anderson, USA, on beamship Model Flying Craft

www.americanantigravity.com

Publisher Faraday Lab Ltd
Editor-in-Chief Alexander V. Frolov, Editor Olga O. Leontyeva,
Designer Ekaterina F. Yakovleva
Scientific Advisory Board Kyrill P. Butusov

New Energy Technologies Magazine, is published by Faraday Lab Ltd, six times a year. Correspondence Address:
POBox 37, St.Petersburg, 193024, Russia
Phone: 7-812-380-6564, net@faraday.ru, www.faraday.ru

Subscription: $90 any foreign. Air mail delivery included
PRINTED IN RUSSIA **Copyright © 2002 by Faraday Lab Ltd.**
All rights reserved.

Beamship Technology: a Re-working of Early 20th century Discoveries

Russell Anderson

Applied Electrogravitics 377 Avon Rd., Ste. D-119
Devon, PA 19333 russanderson3@hotmail.com
484-255-1072

Some Basic Background

The concept of an electric aero-spacecraft with no moving parts was initiated by the Yugoslavian electrical wizard **Nikola Tesla**, who lit the entire world 100 years ago, at the turn of another century, with his revolutionary AC electric current. In 1916-17, Dr. Francis Niepher performed meticulous mass-deflection experiments under rigorous scientific conditions with lead spheres suspended by wires with shielded and unshielded containers. An accounting of this important series of experiments is in TRANSACTIONS OF THE ACADEMY OF SCIENCE OF ST. LOUIS VOL.23, 1916 and 1917. Related article is in THE ELECTRICAL EXPERIMENTER, March 1918.

Before 1905, **George S. Piggot** was routinely suspending small silver balls to water globules, corks, wood, using the electrostatic field from a specially designed Wimshurst machine in a glass container under several atmospheres of pressure to raise the current level. Output voltage was typically 500KV. The field was propagated by a charged sphere. A small curved conducting plate on the floor acted as a ground. He observed unusual patterns of blue dots with filaments over the suspended objects, sometimes with an anomalous 1/2 cm "dark band" on the suspended objects. Piggot states, "It is my firm conviction that that somewhere on the outer confines of our planet there exists a similar contracting belt thru which naught but the gravitational vibrations of the sun penetrate, and these vibrations absolutely annihilate or absorb all other less powerful ones". If the force was Coulombic in nature, objects would be first attracted, and then strongly repelled by the charged metal sphere. After the objects were suspended, Piggot found he could remove the conducting ground plate, and the objects still floated, suspended. The phenomenon of levitation was accompanied by "luminous halos".

In 1925-27, **Albert Einstein** released his scientific "gem", his "zur Einheitlichten Feldtherie", or the Unified Field Theory for Gravitation and Electricity, to the press and the scientific community. It combines electricity, magnetism, and gravitation into a single mathematical expression, showing how High-Voltage/Low Current electricity (Electrogravity) -and conversely Low-Voltage/High-current (magnetogravity) "acceleration-fields" (G-field) could be produced using then-available relatively LOW-technology. Indeed, a very simple technology. The unifying field is the electrical field (because it can produce gravitation and repulsion fields, as well as magnetism). His Crowning work was released with much press write-ups and fanfare, then it was quickly forgotten as if the scientific community and the world had suffered some kind of collective amnesia!

All of the readers of this magazine need no introduction to the pioneering work of American Scientist **Thomas Townsend Brown**, who was playing around with an X-ray tube around the same year as Einstein's Unified Field Theory was released. He filed his first patent for this newly and accidentally discovered "electrogravitational-effect" which causes motion in a high-voltage condensor or capacitor configuration. He was only 17 at that time. The discovery that high-voltage/low amperage electrostatic potentials applied to an object causes motion in the direction of the positive pole, and electrical charges naturally move to the OUTER surface of an enclosed charge-conductor, held strong prospects for what Brown would later name the "space-car", and wrote an article "HOW I CONTROL GRAVITATION". His pioneering work, and demonstration of devices in Hawaii during World War II, drew attention from the department of Naval Intelligence. He was invited to work on "Project-Rainbow" (the Philadelphia Experiment for Electromagnetic Stealth) because of his pioneering work on what was starting to be understood very covertly as a true WARP DRIVE. Experiments with certain new and classified arc-welding apparatus at the Philadelphia Navy Yards to weld armor-plate for battleships was (by use of banks of primitive but powerful avalanche-discharge capacitors) producing anomalous and unexplained effects, such as disappearing tools and other apparatus in the heavily shielded welding chamber. These strange effects were accompanied by a strange "blackout -zone" which, like Piggot's early work, was not optical in nature. TT Brown's devices in his AH Bahnson Labs home movies lift more than their own weight and move inside vacuum chambers in these films. TT Brown later founded NICAP in 1956, which became the most respected UFO data gathering and hard scientific organization in the world, besides the US department of Naval Intelligence itself, and the Foreign Technology Division at Wright-Patterson Air Force base in Ohio.

My background and work

I primarily have a background and degree in computer programming, electronics, most fields of science, Flying Saucer Technology research (almost 30 years worth), Radio/Control fixed and rotary-wing aircraft since 1972. I have been experimenting and working with high-energy and electrogravitic devices and systems since 1987. I built my first small High-Voltage generators starting around this time. I built kits from Information unlimited and elsewhere.

In late January 1990 I built my first working 2-foot flying discs, which were a direct replication of Thomas Townsend Brown's most important representation of his electrogravity-propelled scale-model vehicular concept, from US Patent #2,949,550. In January 1992, I built a 120KV high-voltage/low-current electrostatic generator from an Information Unlimited kit, primarily for force-field propulsion research.

In June 1999, I built a tower and rotor apparatus to complete the experiment, and I powered it with the output from a 100KV generator I built from an Information Unlimited Kit. The results were spectacular, and taught me a great deal about what was involved in producing and maximizing the Biefeld-Brown Electrogravitational effect. **I suspected from my research, and my experiments, that the basic effect was not due to current-flow and resulting ion-wind. When there was current-flow, the effect is attenuated, power consumption goes up, and thrust goes DOWN.** In late June of 2000, I presented this working TT Brown Electrokinetic Apparatus with larger 1-meter discs at the 2nd Antigravity Conference in Reno, NV, hosted by Jim Cox. A VHS videotape of this working and spectacular presentation at the first part of the conference is available from www.soundphotosynthesis.com

Now that I had mastered producing horizontal thrust, vertical thrust, or antigravity, was the next goal. About this same time, there was buzz all over the Internet about claims of two or three individuals who sounded credible at the time who had successfully replicated the many multi-layered "gravity-warp capacitor" or "electric rocket". Oddly, these claims could not be verified, and the individuals making the claims disappeared back into the woodwork. Such actions are bizarre and hinder the progress of true science, which is undergoing a shift in paradigms right now, if not a change in dogma.

I spent months die-cutting hundreds to thousands of tinfoil and aluminum-foil circular-notched conductor plates and wax paper and mylar rings. I did some initial testing with a Tin and wax-paper 400-layer gravity warp-capacitor heap, according to plans I had acquired from H & A Industries in 1992, and what was on Bill Beatty's amateur science site. No one else has come forward with positive results on this tedious and time-consuming device. So much work for so little effect! And, if you short out the stack with too much power, you must tediously and laboriously search thru hundreds of layers to find the dielectric layers with the telltale carbonized holes. The Electric rocket has been recently successfully replicated and tested in hard vacuum and patented recently by Hector Serrano. The Serrano effect is identical to the so-called Biefeld-Brown Electrogravitational effect. They are one in the same thing. I may dust off my completed 400-layer grav-cap, but I hardly find it worth the time and effort, because of my recent work starting in early October 2001.

The Lifter and The Evolution to Beamship Model Flying Craft

Although I had attempted a few small "Hagen" patent-type antigravity (VTOL) models in the early 90s, I found their performance poor at best and their power consumption high. In late summer, 2001, someone, I forget who, on the JLN's lab list of researchers and anomalous science-experiment and technology enthusiasts ran across a website owned by Transdimensional Technologies, of Huntsville, Alabama (famous for NASA research facilities, the late Dr. Rolf Schaffranke, author of the important ETHER TECHNOLOGY, under the pseudonym "Rho Sigma", and Dr. Tom Bearden) had produced a hovering device. From my previous work, I recognized it immediately as TT Brown's Electrokinetic Apparatus that I had successfully replicated and demonstrated before a live audience years earlier. I noticed the capacitors were made from Aluminum FOIL, not the thin-but-heavy Aluminum sheet stock from Home Depot that I had been using for years, (I had assumed that to make my 3-foot discs hover and ascend vertically, I would have to use voltages in the hundreds of kilovolt range, and generate high x-ray, UV, and possibly gamma-ray emissions as a by-product, in other words, a typical flying saucer with all the associated radiological effects that have been documented for over half a century) so they could lift their own weight. The result matches almost exactly the simplest graphical representations of TT Brown's patent from 1960, and De Seversky's Ionocraft patent from 1964, which was a thin foil cathode plate with a thin anode wire separated from the cathode by stand-off insulator posts. I was eager to reproduce these devices (I don't know how I overlooked this simple solution, it was all sitting in those old 1960s Brown and De Seversky patents I have studied for 15 years previously) and many people around the world, especially the webmaster of the JLN Lab's site French researcher Jean-Louis Naudin, who began replicating many different types of larger and more sophisticated devices, some of which resembled model spacecraft, and began amassing tables of very useful data, that researchers could use as basic guidelines to follow. I replicated the first hovering device, the "Lifter" (so-called by Transdimensional Technologies) as a 1-foot triangle, with 2-inch foil cathode and #42 enameled copper magnet wire. To energize it I used a commercial power supply from Gamma High-Voltage Research that I had acquired from Ebay some years ago. It was perfect for antigravity research, having full metering, and variable voltage from 0 to 40KV, and current limiting from 0 to 1.5 milli Amperes of current. The heavy 1-meter discs of the Electrokinetic Apparatus were too heavy and the rotor-friction to great for this low-powered device (60 Watts, maximum), but for the lifter, it proved ideal.

My first "lifter" antigravity device worked, but its performance was less than ideal. It had to be stripped of its lower balsa-wood frame and some of its foil before it would degravitate (counterbary), and it "maxed-out" the current-limited power supply at 33KV I 1.5mA, for a stable hover (actually this is an upward flight configuration, because the device is tethered to the lab table with 3 sewing threads). That is 49 Watts. The concept of a hovering TT Brown Electrokinetic Apparatus had been proven to my satisfaction, however, and I initiated more research into past works and patents to raise efficiency to workable levels. The performance was slightly better than my early 1990s "wire-grid" type devices. I found this slightly encouraging.

After a couple months reading and research (why reinvent the wheel, its all been done before), I started to replicate larger models in February and March 2002, but kept coming up against a size-barrier with the Multicellular (grid) approach that many researchers had assumed would raise thrust, and efficiency. This approach obviously did neither, as no one seemed to be able to produce hovering devices above a certain size, the current consumed (adding to total wattage consumed) was prohibitive with the low-powered (still high-voltage, low-current) devices that most of the mostly amateur researchers were using. Researchers around the world started to replicate different versions of the basic lifter 1 (an 6 to 12-inch equilateral triangle). The lifters are always tethered to the testing surface with 3 strings to keep them from going dangerously unstable and possibly short-circuit when they reach the limit of the umbilical supplying power to the device.

From my previous Biefeld-Brown effect replications years earlier, and from carefully reading Brown's EK Apparatus patent, I knew that increasing the diameter of the wire would reduce leakage current created by coronal discharge, mostly coming from the forward electrode, which in the 2 and 3-foot saucers consisted of an arc of copper tubing in the front quadrant of the saucer, or disc. Corona robs power (amperage) from the disc that otherwise would be used to "propel" the disc. Increasing the diameter of the copper tubing, as per Brown's patent if the effect was due primarily to ion-wind, more current and current flow between the electrodes would be desired to effect more air movement. But this is not what I saw in the saucers. **There was apparently another, far more powerful but subtle force effecting silent propulsion of the saucers that had nothing to do with charge-transfer and ion-momentum.**

In February of this year, I undertook an effort to replicate and improve performance and reduce power consumption of the lifter device, based on data from my electrogravitic work of years past. I started by using thicker diameter enameled copper magnet wire, #35 to #30 diameters. I first built a 1-foot equilateral triangular basic "Lifter-1", weighing only 3.5 grams. On March 16th, I built a lifter with the thicker #35 enameled copper wire.

I made the three sides 1-foot long and exactly 2-inches high. After experimentation, I found the optimum spark gap for my High-Voltage power supply (Gamma High Voltage Research 40KV with current limiting to 1.5mA). The small silver-colored device leapt off the test table and pulled violently against its anchor strings to a distance of about a foot. This seemed like a great deal of force for such low power. The large discs of my TT Brown EK apparatus required a good deal higher voltage to initiate motion in the direction of the anode. The device consumed 26KV I 0.56mA DC, which calculates out to 14.56 Watts. I was getting more excited, because this was the best efficiency seen of any result yet posted.

On March 19th, I tested 2 lifters glued together in a "diamond" shaped configuration.

This 2-foot device weighed 6.0 grams, with the same #35 wire and a 2 and 5/8" air gap. It took 25KV to nullify the weight of the device, and it achieved a stable hover at 35KV I 0.8mA. That is 28 Watts. This is about what I had initially expected, double the power for double the Watts. Still, this was far less overall power going into the device to achieve a stable hover than my first primitive and radically shorn and trimmed device. After lift off to the extent of the anchors, I found I could reduce power slightly and maintain a stable hover. On march 22nd at 3:49 pm I got the diamond lifter to achieve a stable hover with a 2 and 5/8" air gap at 29.5 KV I 0.32mA. This was only 9 Watts! This was unheard-of efficiency. I was further encouraged to build and test larger hovering devices to see how large I could get them with my low-powered commercial power supply.

I then built a "lifter-2", which consists of three 1-foot triangular capacitor cells taped together. It weighs 11.4 grams. March 30th at 3:22 pm, the device achieved a stable hover at 38KV I 0.57mA for 19.76 Watts total power. The larger device was more energy efficient than a device 1/3 the size. I wanted to see how far this could go, so I added three more lifter cells to make a 6-cell device, 3 feet on each of its three sides. I was eager to check the performance of this fairly large device. This was the diameter of my horizontally propelled TT Brown Discs.

This device weighed 21.6 grams. I kept the spark gap the same distance on this device. However this device failed to achieve counterbary (lift). It just sat on the test table, filling the air with the smell of ozone and making a sizzling sound (corona noise). I noticed that the current maxed-out on the power supply at a fairly low voltage and would not go any higher.

I concluded in my disappointment that all that wire from all the inter-connections to the cells was causing corona leakage and robbing current, which otherwise would be used by the device for propulsion.

The idea then hit me that perhaps I could make a device with the same outer diameter as the 3-foot device, but have greater efficiency because of a much shorter length of wire. I built basically a 3-foot (1-meter) version of the first 1-foot device. This device weighed 16 grams. It lifted off the table with amazing force and hovered stably with I 52mA 30KV which is 15.6 Watts. Not only had I achieved a larger-size device, but far better power efficiency for a much larger and heavier device. I was overjoyed! I saw that I had a great deal of lifting force to spare. I had not even come near the limit of my power supply. I added extra bracing at the corners and extra balsa and a triangular paper "payload -tray" in the center of the device, supported by three 1/16"x1/16-inch balsa stock. The extra bracing and payload area added approximately 2 more grams. With a 5 gram payload, the device consumed 39.9 KV I 0.99mA, for total power consumption of 39.5 Watts. I was really encouraged at that point, because I knew that these results were unheard-of, in terms of energy efficiency. I had solved the problem of decreasing efficiency by dispensing with a "grid-based" device. Increasing the area of the capacitor plate was one of the factors that increased performance and efficiency, lessening the input power requirements with increasing size. Now the Biefeld-Brown effect could be properly studied, now that most of the ion-flux had been eliminated, resulting conservation of energy by the device, and resulting in greatly increased propulsive force.

Since I now knew the limit of payload for the device at the power level I was using, I added a balsa framework that approximated a central cabin area, and three small styro-foam spheres on the center of the straight sides on short lengths of balsa. The device no longer looked like a test device, but now looked like a scale model spacecraft. I remembered the Edouard "Billy" Meier UFO contact case, and knew that all his original photographs and movie footage of extraterrestrial spaceships the extraterrestrials themselves called "beamships" (there are several styles and variations, all with different specific functions and capabilities, some manned, some remote-controlled "telemeter discs" that had a tri-hemispherical undercarriage that I knew from past research were propulsion condensors) and that the original un-tampered photos all passed rigorous analysis using the latest and most sophisticated computer and other equipment, case detractors not withstanding. Also the spiritual messages of these genetic brothers of Man and their accounting of humankind's history and origins from far across space rang true and struck a chord with me. I decided to name this new 1-meter model spacecraft Beamship Variation I. The three sides of the device performed the same function on this device that the three spherical or hemispherical capacitors often seen on the underside of full-size "beamships" (Daylight-disc-type UFOs), which illustrates a similar if not identical propulsion methodology to full-size 3 to 7 meter and larger "off-Earth-built" aero-spacecraft. Clearly the propulsion methodologies were exactly the same in the model as in the full-size flight device.

I immediately built a 4-foot diameter model with a full cabin framework and internal payload area and achieved even greater performance and efficiency. I was ecstatic. This 4-foot device I dubbed Beamship Variation II. I received a suggestion from Mr. Tim Ventura of American Antigravity that performance could be increased by using small diameter stainless-steel wire. It seemed unlikely to me that smaller diameter wire would increase performance, it contradicted Brown's patent, and my own past research with large electrogravitic discs. But Stainless steel has a high number of free electrons in the outer valence atomic shells (electron orbits).

So I obtained some #40 stainless locally and the results confirmed Mr. Ventura's suggestion. Corona noise was heard at a much higher power level, and was greatly attenuated in volume. Leakage current was less, and the two Beamships now had more thrust with less power input. They even carried more payload at less power input. Variation II weighs 21 grams and loft a payload of 6 grams at 40KV I 1mA for 40 Watts total power. Again, this was unheard-of efficiency. The anode wires sang a strange harmony as the Beamships floated in the air, stably at any altitude, from floor to ceiling, without any fuel or visible means of support. This was very Beamship-like.

I thought that now since corona discharge on the anode wire was less, I could decrease the spark-gap distance without creating a spark (which kills lift). Thrust seemed initially to increase, but efficiency went down because there was current-flow now, and current consumption went way up. The supply would now max-out at 37KV I 1.5mA and would not increase because of the current limiting. The Beamships now were noisy, as the foils chattered loudly because of all the ion-wind that was now rushing downward along and past the foils. I used a concert fog machine to observe the ion-flux vector, and filmed it digitally with my Logitech web-cam, and with VHS analog video.

Analysis of the fog -tests showed a circular vortex of air surrounding the anode wire that flared out into a downwash of air below the Beamship. I was disappointed, because I thought then that the thrust action of the device was due to simple ion-transfer. A useful-enough effect, but of questionable use in the vacuum of space without an ionizing medium.

Beamship Variation III

I reasoned I had just about enough power in the supply to build and fly a 6-foot (2-meters) Beamship. Since the balsa came in 3-foot lengths, this was simple. As with all the lifter devices and the more evolved and efficient Beamship-series model aircraft, construction techniques are extremely simple and require little skill to assemble. Weight of the Beamship Variation III is 42 grams, with 6-feet of length on its 3 sides. Height of the foils was still 2". Full frame and cabin, with Searl "IGV"-type landing legs, to support the weight of this heavy and very large device. I set the spark gap at 2 and 1/2 inches. At 12:15pm EDT, May 12th, 2002 the Beamship was weightless at 32KV with current maxed-out at 1.5mA.

The device barely lifted off, and "hopped", across the floor once or twice at full power. It had the same loud rattling of the foils due to the terrific downwash of electrified air. I needed to raise the power level. I increased the distance of the spark gap to 2 and 3/4". Now the Beamship took off straight up with power to spare, as if it was one of the smaller craft. Beamship Variation III is weightless at 30KV I 0.85mA (25.5 Watts), and airborne into a stable hover at 35KV I 1.35mA. That is only 47.25 Watts. It can carry a payload of 5 grams, or 5 grams worth of additional framework and structure, to the limit of the power supply, which is 60 Watts (40KV I 1.5mA).

The 42-gram, six-foot model aero-spacecraft only consumes 47.25 Watts at hover, but my first small and trimmed device ate up 49 Watts! Clearly, using the single-cell Beamship methodology had a huge advantage over the "multi-cellular" design that other researchers had built and tested, seemingly reaching an impasse in terms of size and efficiency, which my large single-cell Beamship technique had seemingly solved. In early April my 1-meter Beamship, weighing 22 grams including 5-gram payload, consumed 39.6 Watts. So the 42-gram, 6-foot Beamship used only 7.65 Watts more total input power at stable hover that the 1-meter Beamship. Power-to-weight ratio for the 2-meter Beamship V. III works out to 1.125 Watts to lift 1 gram stably of scale model electric spacecraft. With little to no ion wind, the Beamship had plenty of upward force and achieved stable counterbary at greater efficiency than I had ever heard or read about. Also, I was not aware of any devices in scientific history that has achieved this type of counterbary for this little input power and this colossal size. I performed additional fog-tests with two red semiconductor-emitted laser beams in the plane of one side of the device, one above the wire, one below the foil. The Beamship without all the ion-wind was nearly silent again, except for "singing" and softly "thrumming" anode wires. These laser-beam tests further confirmed the marked absence of ion-wind with a larger spark gap.

At 1.125 Watts-per gram at 57KV I 1.4mA (78.8 Watts) would lift 89.775 grams worth of electric spacecraft. So not only is the larger size in a single cell far more efficient that the "grid" design, in thrust and power consumption, but the reduction in current from increasing the spark gap raised power level to the device, while dropping power consumption of the device. My initial suspicion of ion-wind producing most of the thrust in the Biefeld-Brown effect had been disproven also because of the great weight of the device. I never would have discovered this important fact if I had stayed with smaller devices, trying to raise their efficiency. And I never would have discovered the efficacy of the Biefeld-Brown effect if I had stayed with the multi-cellular "lifter" methodology. One of the factors that raises the level of propulsive force (if "propulsion" is the right word) is increasing the area of the plate, according to TT Brown's patents. So the larger size single-cell capacitor's ability to reduce power consumption and effect greater propulsive force and upward acceleration, was easily explained by the Biefeld-Brown Effect. Brown had been vindicated. My gut feeling had seemingly been confirmed: this was our first warp-drive (reaction-less drive).

My experiments had yielded greater efficiency, and greater size and weight of VTOL hovering models than any that I had ever heard or read about. I still am having difficulty taking in these facts; and it is very awe-inspiring to see such a large device de-gravitate and hover stably at any altitude, from floor to ceiling.

Next for me is a higher-powered supply (60KV), moving up to a three-meter Beamship Variation IV, improvements to the cathode such as a thin, symmetrical airfoil shape, as Brown suggests in his patents, tungsten wire, and then carbon-wire for the anode, and full heat-shrink coverings on the frames, running lights, onboard lasers, onboard digital/proportional Radio/Control, and now that we know the power requirements, and have a good handle on efficiency, onboard power generation. I have already begun designing with my associates our own custom-made outboard and inboard battery-powered power supplies, and finally will cut the power umbilical to the model electric spacecraft permanently, and Beamship technology advances further. If the on/off duty cycle is pulsed at a low frequency, power input can be reduced by two-thirds, at least. Experiments conducted by Jean-Louis Naudin last fall (2001) confirm this phenomenon, suggested by Brown and De Seversky in their patents. Clearly, onboard power can easily be effected, using modern miniaturization and circuitry. Pitch and roll vector control can be achieved by electrically isolating the anode wires on the three sides of the ship, and varying pulses to these three wires. Yaw control can be achieved by simply installing a horizontal-double sided capacitor inside the ship near one corner. Simple full 4-channel flight control is thus achieved.

The Beamship series aircraft are fascinating research and entertainment devices (see cover page), and are the vanguard of a whole new generation of radio/controlled antigravity model aircraft with no moving parts and dead silent propulsion. But they are more than that. The Beamships, if allowed, could probably rise up at any speed thru the atmosphere, right up to near-Earth-orbit, and probably keep on going out into limitless space. No need to achieve ballistic escape velocities of miles-per-second. This is non-ballistic flight. They even have a certain amount of wind resistance outdoors and indoors because the electrical field causes air to flow AROUND the model flying craft, not into it. This is such a safer, environmentally cleaner, vibration and nearly silent and more pleasant method of aero- space travel than carrying tons of explosive reaction mass, which can and does explode. No more use of heat energy to effect transportation.

The 21st century has begun in earnest!

Beamship series aircraft are available for sale for research and hobby/entertainment use right now through the American Antigravity website: www.americanantigravity.com. Look for the Applied Electrogravitics antigravity technology website late spring, 2002. You can contact me, Russell Anderson for details on pricing, and new and improved variations, and power supplies for outboard and onboard drive, which are currently in design stages.

Editor's note: More ideas on development of T.T. Brown's patents are on our web site: http://www.faraday.ru. Read about T-capacitor!

Data Table 1

Antigravity device	Weight of device	Wire type	Voltage/current	Total Watts	Payload	Payload/power
1-foot "Flyer-1"	3.2 grams	#42 enameled copper	33KV I 1.5mA	49	None	————
1-foot "lifter-1"	3.5 grams	#35 enameled copper	26KV I 0.56mA	14.56	None	————
2-foot "diamond-lifter"	6.0 grams	#35 enameled copper	29.5KV I 0.32mA	9	None	————
2-foot, 3-cell "lifter-2"	11.4 grams	#35 enameled copper	38KV I 0.57mA	19.76	None	————
3-foot, single-cell "Beamship Variation I"	16 grams	#35 enameled copper	30KV I 0.52mA	15.6	None	————
Beamship Variation I-fully-rigged	18.5 grams	#35 enameled copper	————	39.5	5 grams	39.9KV I 0.99mA
4-foot "Beamship Variation II"	21 grams	#40 stainless-steel	————	40	6 grams	40KV I 1mA
6-foot "Beamship Variation III"	42 grams	#40 stainless-steel	35KV I 1.35mA	47.25	None	————

References

1. Ether Technology, by Rho Sigma,
2. Lost Science, by Gerry Vassilatos,
3. The Philadelphia Experiment, by William Moore,
4. High Energy Electrostatic Research Vol. 1, Antigravity and UFOs, by Raymond A. Nelli, Homemade Lightning, by R.A. Ford,
5. Antigravity – the Dream Made Reality, by John Searl, The Law of the Squares Book 4, by John Searl,
6. The Antigravity Handbook, compiled by D.H. Childress,
7. The Coming Energy Revolution, by Jeanne Manning,
8. Agnew H. Bahnson Labs home movies (1957-60), JLN labs website, Bill Beatty's "weird-science" website, American Antigravity website (www.amercanantigravity.com).

Patent references

1. A.H. Bahnson, Jr. Electrical Thrust Producing Device, USP # 2,958.790, filed May 12,1958, granted Nov. 1, 1960.
2. Marcel Jean-Joseph Pages (France), # 1,253,902 Petit-Viton, filed Jan 5, 1960, granted January 9th, 1961
3. G.E. Hagen, Flying Apparatus, #3,120,363, filed Sept. 11th, 1958, granted Feb. 4th, 1964
4. T.T. Brown, Electrokinetic Apparatus, #2,949,550, filed July 3rd, 1957, granted Aug. 16th, 1960
5. T.T. Brown, Electrokinetic Generator, #3,022,430, filed July 3rd, 1957, granted Feb. 20th, 1962
6. A.P. De Seversky, Ionocraft, # 3,130,945, filed Aug. 31st, 1959, granted April 28th, 1964
7. V. Gradecek, Electric Aerospace Propulsion System, # 3,177,654 filed Sept. 26th 1961, granted April 13th, 1965

SmartPAK Technology™

invention by William Alek

the world's first commercial all solid-state system that provides a "standard" platform for experimenters, researchers, and developers to do energy-related practical applications, experiments, and perform exploration of highly efficient alternate energy system

The theory of operation is based on the amount of energy that is required to magnetize and de-magnetize ferromagnetic materials utilizing a core/coil/magnet assembly. It has been discovered that it takes MORE energy to magnetize a suitable core assembly than to de-magnetize it. The SmartPAK system is designed to measure and collect the difference, and store the excess energy for later use.

The SmartPAK device is controlled by a 68HC908GP32 micrcontroller programmed to measure input/output voltages and currents, calculate COP, and contains software algorithms for a complete "turn-key" power management system. The device features a "standard" user interface, which allows the user to design their own custom core/coil/magnet "head assemblies", and immediately test and display in real-time its' performance.

William Alek

President and CEO of INTALEK, INC

3506-43rd. Place
Highland, Indiana 46322-3129
e-mail: wsalek@intalek.com

Correspondence Address: POBox 37,
St.Petersburg, 193024, Russia
Phone: 7-812-380-6564,
net@faraday.ru, www.faraday.ru

Commercial Antigravity

Tim Ventura

tventura6@attbi.com

Introduction

Let me begin by posing a simple question that I would like the reader to keep in mind throughout this article: How far away is commercial antigravity? I'm not talking about a laboratory experiment where a giant magnet is used to levitate a frog, or secret UFO experiments that the government isn't sharing with the business world, but a real, viable antigravity solution to what I consider to be the most pressing issue facing the world today – transportation.

Who can answer a question like that? How far away is commercial antigravity? The author has read numerous scientific texts on the subject, and is familiar enough with contemporary theories of gravity, antigravity, and electromagnetism to suggest that most scientists believe that commercial antigravity is at least 100 years away from existence. But the author believes that most scientists are wrong.

Defining terms

Any article about antigravity would not be complete without properly defining the terms to be used. In this article, antigravity is not used in the strict sense of the word. The author's intent is to discuss a method of propulsion, which for all intents and purposes can be considered antigravity, and may include antigravity – but also may include several other forms of similar propulsion. The reasoning behind this is that experience has shown that the majority of people in the world don't care how something works – they care what it can do for them. This article is about the effect of Antigravity – not the cause.

Real Antigravity would consist of an apparatus used to either reduce the apparent mass of an object or reduce the effects of gravitational attraction between the Earth and an object. An example of an apparatus that may in fact do this is the Podkletnov superconductor apparatus currently being tested by NASA.

This Podkletnov device essentially consists of a spinning superconductor that self-levitates above a pool of liquid nitrogen and supposedly creates a "beam" or "shaft" of antigravity (or reduced gravity) directly above it as it operates. The levitation of the superconductor itself is not antigravity – it is a well-known side effect of ceramic-superconductors called the "Meissner Effect". The Meissner effect is simply a side effect of the superconductor's interaction with the Earth's magnetic field, and is easily explained by physics.

Podkletnov claims that when he worked with a team of researchers investigating superconductors in Russia around 1991, the smoke from the tobacco pipe of a fellow researcher began to climb steeply in a column directly above the superconductor. The researchers began to think that they were on to something, and Podkletnov subsequently performed several follow-up experiments that led him to the conclusion that the levitating superconductor produced a shielding effect between the Earth and anything positioned directly above the superconductor. He reasoned that gravitational shielding would provide a "column" of reduced gravity above the superconductor that should extend up and away from the Earth indefinitely. Podkletnov calculated that with a rapidly spinning levitating superconductor he had achieved a 2% loss in weight for anything directly over the superconductor.

Podkletnov's research is interesting and compelling, and it would fall into the category of "real" antigravity – but I am writing about Podklentov's type of research as well as enormous amount of research and theory available on electromagnetic propulsion systems. These can be considered "effective" antigravity.

The Harrier jet fighter can swivel its engine exhaust-nossels to create vertical lift, which resembles antigravity in that it is VTOL takeoff. However, the Harrier does not use effective antigravity because it has all of the functionality and side effects of an aircraft. A helium blimp would be a closer example to "effective" antigravity, but it too is not – because it works on basic aerodynamic principles.

Aerodynamics is not effectively antigravity – aerodynamics is instead expensive, difficult to manufacture, prone to explosive failure, and highly unreliable. This is not to suggest that a jet aircraft is unreliable, because it has a variety of backup systems, but that the technology itself is unreliable in that a jet is adversely effected by the medium that it uses to propel itself. Air pressure, humidity, temperature, and strong winds all cause a degree of unreliability. In addition, turbine engines stretch the limit of what mechanical engineering can achieve – which is why they are prone to break if even something as small as a bird gets sucked in during flight.

Antigravity is not about moving the air around — it is about a medium-agnostic means of air transportation that produces vertical and/or directional lift without relying on air-pressure like a wing or blimp. Antigravity is an electromagnetic or electrogravitic system for reducing the weight of an apparatus to allow it to lift more easily. Antigravity is pushing a button and having your vehicle take off without runways, noisy engines, minimum flight-speeds, propellers, or any of the other drawbacks that limit conventional aircraft from achieving popularity similar to what an automobile might have.

The author's definition of Antigravity for the purposes of this article is confined to electromagnetic or electrogravitic devices that reduce the weight of an object to enable it to take off without conventional thrust-producing apparatus. This definition might also fit many of the classical characteristics known at one time as "the electric spacecraft".

1. *Business Analysis versus Scientific Analysis*

The author disagrees with the majority of scientists as to when commercial antigravity will become possible for some very basic and obvious reasons. To begin with, the majority of credentialed physics-related scientists come from a theoretical school of thought, which tends to limit their world-view to only contain those things that are currently or potentially explained by theory. The author, however, comes from an experimental school of thought that seeks to capitalize on existing observational data without the rigorous need to explain every last detail of its functionality. The author is an engineer, not a scientist – and engineers don't need to totally understand how something works in order to make it better.

This difference between the engineering point of view and the physicists is also different in the manner in which they seek out observational data. A physicist looks towards naturally observable data, and in the event that none exists they look towards current theories to explain potential future observations. The engineer is more open to ideas that are less rigorously tested from the perspective of scientific method, but are currently observed as potential solutions to real-world problems.

2. *Potential Technologies Overivew*

Phycists currently tend to dismiss the entire concept of Electrogravity, and the reasoning behind their logic is very sound. To begin with, Electrogravity is not observable in nature. In addition, many of the claims by those persons who submit Electrogravity and antigravity devices for public review are faked, exaggerated, or just plain wrong.

Physicists are responsible for maintaining a working body of theoretical knowledge, and if they were to admit results such as Schnurer's without skeptical scrutiny it would undermine the very fabric of technology itself. If the Podekletnov results were to be accepted as fact at face value without rigorous prrof, imagine the amount of money that would be wasted in attempts to build enormous Antigravity vehicles based on this theory.

In the middle of the spectrum lies the concept of Maglev, which is mentioned here only for the purpose of specifying that Maglev is not commercial antigravity. It has been mistakenly thought of as antigravity by many because it utilizes a magnetically-levitated train to improve the velocity of the train and reduce transit time between stops. In reality, Maglev is not really a vehicle at all.

The definition of a vehicle would be a device that transports itself as well as its passengers and cargo between two points. This is why an automobile is considered a vehicle but an escalator or elevator is not – the automobile transports its entire propulsive apparatus to another location, but an escalator or elevator does not move – it merely repositions its cargo between points. The Maglev train is not really a vehicle at all – it is actually a very long electromagnetic armature that transports people and cargo between its ends at high speed. While it may serve a commercial need, it is not to be confused with Antigravity.

On the opposite end of the spectrum is the author of this article – who has built and successfully tested over 30 electromagnetic "Lifters" at the time of this writing. The Lifter is a device based on research by Transdimensional Technologies and related to research by Thomas Townsend Brown that demonstrates an antigravity effect when a High-voltage DC current is applied to it.

Currently, the exact method of propulsion for the Lifter is being debated. It is thought to be one of two things – either an effective form of "ion-wind" propulsion, or else a form of field-effect propulsion based on an as-yet unknown force. While the debate about the exact nature of this propulsion is important with regard to future research, in reality it does not change the observational data that demonstrate that this technology works perfectly, consistently, and reliably.

The Lifter design was demonstrated by the author in a continuous mode of operation for over 7 hours straight on Sunday, April 21st, 2002, at the Seattle Center "EarthDay and Renewable Energy Exhibition". During this seven hour period of time, the author's Lifter hovered at a tethered height of 12 inches from the surface of the table, powered by a 30 watt load from a simple computer monitor.

This article is not meant to get into the details of methods of antigravity, only to suggest that it already exists in the form of electromagnetic propulsion systems if nothing else. The author is confident that in time physicists will find a theoretical reason for why the Lifter operates as it does, but for the time being the fact of its operation overshadows the method of its operation.

3. *Market Needs*

Commercial Antigravity doesn't require a 2% loss in weight to operate – it will require something akin to a 200% loss in weight. A commercial antigravity device will have to demonstrate exceptional performance to gain market acceptance, but not for the reasons that might immediately come to mind.

One might believe that skepticism from the scientific community would prevent antigravity technology from gaining the scientific acceptance needed to become a commercially accepted engineering discipline. The long term view, however, shows that this is not the case – engineering and market forces drive innovation, and formal science plays a supporting role in explaining and quantifying the innovations that engineers have already commercially qualified as valid.

The real roadblock to success for commercial antigravity is market acceptance. The author's demonstration of the Lifter technology at the Renewable Energy Exhibition helped him to realize that the vast majority of consumers have no idea what antigravity technology could be used for, much less what they themselves could use it for. The same thinking was apparent at the dawn of the age of personal computing, when the idea of having a computer in the home was a completely foreign concept.

So in brief, a market does not exist for antigravity technology, which is why inventors working with this technology have been unable to find appreciable support for their work. Many inventors look at this technology and ask, "how could the public not understand how valuable a technology like this is?" – but that isn't the problem. The problem is that most innovators with an interest in antigravity are so closely tied to the science behind the technology that they fail to review and address the business needs that drive the market acceptance of a new technology. In other words, people don't buy antigravity – they buy solutions. People don't buy cars to simply have a car – they buy cars because people need transportation needs that they have to fulfill. People don't buy computers because they want to have a computer – they buy computers because they want to share and process information and communications.

Marketing Requirements

How will antigravity technology gain the market acceptance to become a commercially viable technology? There are a variety of ways in that antigravity technologies will become commercially viable, but only after antigravity is no longer sold as antigravity – it needs to be sold as a personal or business solution.

The solutions that antigravity technologies are best prepared to provide at the moment are in the realm of transportation technology. This includes moving people and cargo to destinations in a similar manner to conventional transportation technologies such as aircraft or automobiles.

With regard to providing transportation solutions, antigravity has the ability to incorporate the best features of both contemporary automotive and aerospace technologies into a single technology that will serve point-to-point transportation needs better than either of the two aforementioned technologies could by itself.

For a moment, assume that a person wants to travel from Los Angeles to New York in a short period of time. Currently, the most convenient method of transportation to accomplish this would require the person to take an automobile to the airport, and from there take an aircraft from the Los Angeles airport to the New York airport. After departing at the New York airport, the passenger must then take another vehicle to their intended destination.

Commercial antigravity technology could serve a dual-purpose short and long-range transportation role, taking on the aspects of both ground transport as well as air transport.

Product Delivery Requirements

In order to deliver commercial antigravity as a viable solution to business needs, a variety of work will need to be completed on the various component systems of this technology to turn it from what is currently a "proof of concept" into a commercial reality.

Let us assume for a moment that we have developed a working device based on Antigravity or some method of Field-Effect Propulsion. While this is the critical stepping stone to success, this is by no means the end al be all of the development cycle.

To begin with, the technology must be perfected to the point of being both economical and reliable. As it stands now, the market already has technologies in place that fulfill some or all of the requirements for the technology that Antigravity is being developed to replace. In order to serve as an effective replacement for these technologies, antigravity technology must then demonstrate that it both costs less in terms of operation and manufacture, as well as being more reliable than conventional air-transportation solutions.

I mention reliability in light of the recent negative media attention surrounding several recent commercial airline crashes. From a marketing perspective, air-travel disasters provide a great deal of negative publicity for the airline industry. Since the airline industry has a mostly successful track record of delivering passengers and cargo, people are for the most part willing to forgive the occasional air-disaster. However, with a new technology such as antigravity-based air-transport, there is not a long enough track record to permit public acceptance of air-disasters. One substantial disaster in the early days of antigravity could serve to forever damage the credibility of this new technology.

With regard to being economical, any type of antigravity system that intends to surpass existing methods of air-transport must be able to do so at a less-expensive rate to own and operate, and must have a vehicular lifespan at the very least similar to conventional air-transport devices. This would allow the total cost of ownership (TCO) to be less for an antigravity vehicle than it would otherwise be for a conventional craft.

There is one caveat to acceptance of antigravity technology as compared to conventional aircraft, which is simply that if antigravity vehicles are able to operate in an environment or manner that precludes conventional aircraft, then they should be able to gain a market niche without immediately having to surpass conventional aircraft in the area of TCO.

Assuming that we can develop a propulsion system that is both more reliable and less expensive to operate for the transportation of passengers and cargo, we then have to build up the skeleton of a vehicle compatible with this form of propulsion around the actual propulsion system.

For instance, a conventional aircraft has pitch, yaw, and several other flight controls, but for an antigravity vehicle there is a high likelihood that some or all of these controls will not be required, thereby changing the dynamic of flight associated with the craft. This will require new methods of pilot certification and flight-qualification, as well as requiring a control-philosophy to be created surrounding how the craft will operate.

I use the phrase "control-philosophy" instead of simply "control layout" because one of my assumptions about antigravity propulsion systems is that they will allow more flexibility in the design process for engineers to determine how the craft "should" fly, as opposed to an aircraft or helicopter, in which the components determining speed and handling are based primarily on an interaction between the design of the craft and the atmosphere.

In brief, an antigravity cargo-transport may have very different needs for flight than perhaps a lightweight passenger vehicle would, although there would also need to be a consistency between the control-systems of these devices to reduce the need for extra pilot training and competency testing.

Therefore, it should be apparent from the last few paragraphs that not only are there several propulsion-system related challenges involved with developing a commercial antigravity device, but there are also several challenges in the design, training, support aspects of this technology that also factor into the requirements to be complete before a complete product can be delivered.

Conclusion

At the beginning of this treatise, I posed the simple question of **"how far away is commercial antigravity"**. The reader, I expect, probably interpreted that question in terms of time, which is the usual measurement of questioning when new technologies will become part of our lives.

However, as I have attempted to demonstrate throughout this article, **the time component is much less important to the development process than is the distance component – that is, how far away from commercial antigravity we are.** When I use the word distance, I mean specifically what tasks must be completed in order for antigravity to go from being a proof-of-concept approach to a new form of propulsion-system to being a completed vehicle ready for manufacture.

I have attempted not to address the legal implications of antigravity technology with regard to certification for general or specific use – my thought on this is that the discussion of legal ramifications of antigravity is best left for another time. This is due primarily to the size and scope of that discussion, which is beyond what I am attempting to analyze in this article.

So, in finale, how far away from commercial antigravity we are depends not so much on time as on the rate at which we can perform the work required to provide the underpinnings on top of which the technology can be built. This seems important to me, as it underscores how close we appear to be to a working method of antigravity propulsion, and how we might consider focusing resources and goals to achieve the realization of this common dream that we share.

EDITORIAL: PERPETUAL MOBILE OF 1902

There is one more interesting example of perpetual mobile, which was described in the collected articles [1]. The motor shown in Fig.1 was invented in 1902. The vessels *b, c, d* and *e* are mounted on a shaft *a*, and have one side *f* tangential to the shaft, and the other side radial. Compressed air is forced into each vessel through the valves *p*. It is stated that under "the action of the internal pressure of the vessels, and after a slight impulse has been given to same, in the direction of the arrow, the whole apparatus will begin to move and continue to do so without ever stopping, the velocity corresponding to the pressure established within the vessels".

Really simple... Let's try to examine it.

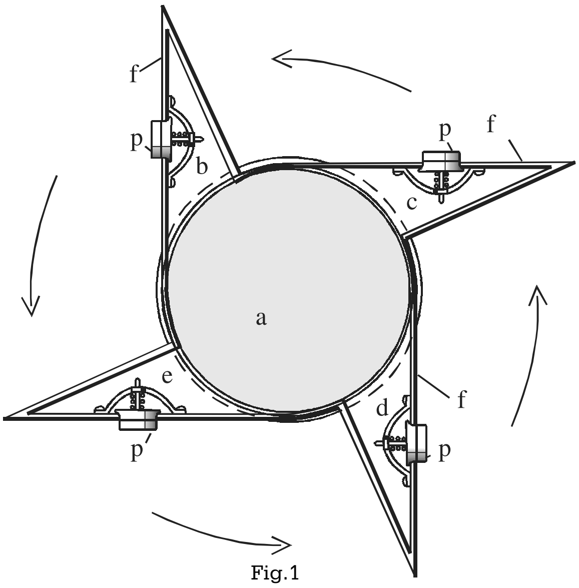

Fig.1

Reference

R.A. Ford. The Perpetual Motion Mystery. A Continuing Quest, 1987
Lindsay Publications, Inc., Bradley, IL 60915

Action without Reaction

New Gravidynamic Paradox

Yu. N. Ivanov
Academician of Russian Academy
of Natural Sciences

Saikina, 11/2, app.4, Moscow, Russia
Tel: (095) 277-33-76/276-50-39
mirit@narod.ru
http://www.mirit.org

Editorial: Yu.N. Ivanov discovered a new unusual physical phenomenon that is a gravidynamic paradox. The sense of the paradox is the existence of the situation, which is not forbidden by physical laws and which allows any man (even a child) easily hold suspended or carry things with 100kg or more weight by means of some simple device. This phenomenon is just a part of the significant applied topic and a demonstration of the possibility to cancel weight characteristics of any material object. Per se it is a work on the obtaining of the methods to eliminate aircrafts weight.

We can use only resistant things as a support!

Galilee-Newton's Laws are in the basis of classical mechanics. Thought concerned fundamental, these laws do not explain the main thing, i.e. the internal processual essence of the phenomena, described by them. There is another scientific line – the quantum mechanics, which tries to discover the essence of deep processes. There is no connection between these lines, therefore they as if exist independently. There are also long-time conversations about necessity to combine the quantum and the classical approaches. We suppose, that practically the combining has already happened, since thanks to rhythmodynamics, there appeared phase, frequency and velocity of light in formulae of classical mechanics (without these attributes the quantum mechanics is impossible). It was revealed, that phase-frequency method of late mechanics formulae presentation gives a real physical sense both to formulae themselves, and to concepts which seemed vague before *(namely: cause of motion, force, velocity, inertiality, gravitation)*. In this sense, the upcoming rhythmodynamics fills up the gap between the main physical lines and draws them together. But let's change the vector, since **the subject of the given article is to concern cases in which the third Newton law is not directly valid.**

Newton formulated the third law in the following way:

"Action always has equal and opposite interaction, in other words, actions of two bodies on each other are equal and directed in opposite sides".

This law represents the fact that one-way action of one body to another cannot exist in nature, but there is only an interaction between them, i.e. there is no action without reaction [1].

We must note, that the third Newton law is valid only for systems with 100% feedback, appeared in the interaction period. For example, during the magnet influence on the iron object it is revealed, that this object also starts to influence upon the magnet, i.e. to attract it with the same force. It happens because in the period of action of magnetic field the iron object itself becomes the source of magnetic field (Fig.1). Here the magnetic field, as an especial environmental condition, acts as a mediator, i.e. exists by itself in the interval between object and magnet .

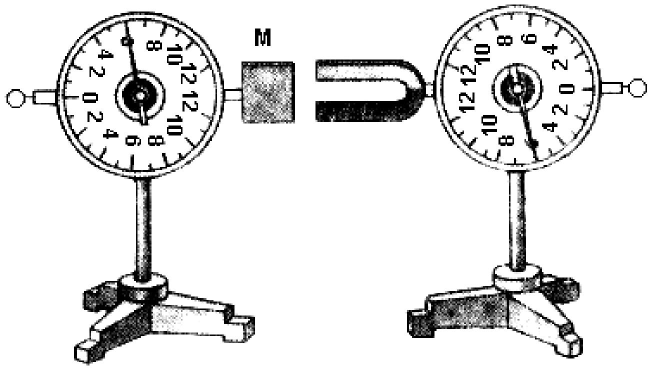

Fig. 1

Interaction of magnet and metal body M. Here the action is equal to reaction, i.e. the third Newton law is valid.

It is considered to be that magnetic field, created by electric current, spreads with the velocity of light. If current impulse is short, then the magnetic field spreads independently from the conductor and no matter if there is current in the conductor at present, or not. If the distance up to the object is big, then the magnetic field remains between source and object for some time, and has no influence upon the object. In this sense the portion of magnetic field, placed on the path between source and body, can be considered as moving independent "entity", i.e. the space by means of itself carries its changed state (magnetic properties) from one place to another. In this period magnetic properties in the form of quantum can influence neither on the source, from which the magnetic quantum is already detached, nor on the object, which the magnetic quantum still does not reach.

But having reached the object, magnetic field changes the state of this object. If as the result of influence the object becomes a source of magnetic field, then the part of the field is reradiated in the direction of source, i.e. the object itself becomes the source for some time, and by this it is able to influence on the first primary source. In this situation the principle of action and reaction works, because the feedback takes place.

If the feedback does not appear during the period of influence of one body to another, then the action and reaction law is not valid in the system. Let's demonstrate it by the example of the mechanical experiments, in which there is a feedback between objects.

Let there is a device (Fig. 2), which throws off two water (air) streams in the opposite directions in such a way, that reactive forces completely compensate each other. At that the thrown down stream compensates the gravity too. In this case the device will fly without falling, i.e. it will have zero weight.

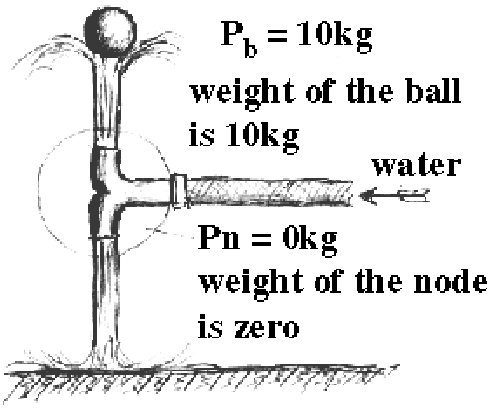

Fig. 2

The ball, suspended in the water stream, does not put pressure on the source. It happens because there is no feedback between the ball and the source. If you have such a device, then, lifting the ball by means of it, you will not feel the weight of the ball. The reason: the ball hangs due to the kinetic energy, which the stream passes to it and and this ball can not influence on the source through the water stream.

The appearance of the ball with the weight P_b does not influence on the weight P_n of the device. And even in the case if we press the ball down, i.e. essentially displace its location in the stream, the device will feel nothing.

The absence of device weight is evident for us; therefore we can easily move the device if its original (vertical) orientation remains unchanged.

Let's change the situation and place the ball (body with the shape, which is steady for hanging) in the upper stream. This ball with the weight about 10 kg is placed in a way to be kept by this stream at some distance, for example at 0.5 m. Will the device react on the changed situation, i.e. on the ball appearance? Will the weight of the ball, hanging in the stream, be added to zero weight of the device (0 + 10 kg)?

The calculation shows that it will not: $\Sigma P=0+10=0$ [kg]. There is only an illusion that the ball is supported by the stream and that it is a part of the system. As a matter of fact, the ball is detached from the system and hangs due to the water kinetic energy. It is easy to check it by making an experiment in a bathroom: "Remove the douche sprayer, then direct the stream up and holding it by the one hand with the aim of weighting, try to influence on the source by the other hand through the stream. You will be surprised to feel that the hand, which is supported by the stream or any object, which is hanged in the stream, does not influence on the source".

At first sight the situation seems to be a paradox, however it very illustrates the possibility of the force action without reaction. So, the action can be one-way, i.e. calling no reply reaction in the form of pressure on the action source.

In such an unusual way it is possible to keep the ball of bigger weight (100 kg and more), at that to move it easily by means of the device and at the same time not to feel the presence of additional weight in the stream. **It is possible only in the case when there is no feedback between body and source,** i.e. the body, hanged in the stream, has no ability to act on the force source. Thus, we can both hold and move the heavy ball without additional efforts, and also lift it up to any height (for example, 100 m). As a matter of fact, we deal with the **new gravidynamic paradox**.

It is interesting then, how to solve the task of the following type: "Let the body with 100 kg weight is hanged in the water (air) stream in such a way, that it does not change the velocity and direction of liquid outflow from the source nozzle. How much energy the operator must spend to lift this body to the height about 10 m?" (Here it means that the operator must take the device, which supports the ball, and, moving upstairs, lift the body, which hangs in the stream, by means of this device to the height about 10 m). If to solve this problem correctly (the condition is that the operator lifts himself together with the device and the ball), we will find out, that operator's energy is spent only for lifting of the device, which creates water streams. The operator, lifting together with the device, will not even notice that in the stream there is a body with about 100 kg weight **(this is the sense of the paradox).**

The situation only seems to be absurd, and even paradoxical for theorists, but it is not a hopeless one.

We can also observe the effect of action without reaction in ultrasonic field of the source. If the source is fixed on scales (Fig.3) and body is hanged, as it was made in the stream, the scales will show only the precise weight of the source and will not react on the weight of the hanged body, no matter how heavy it is.

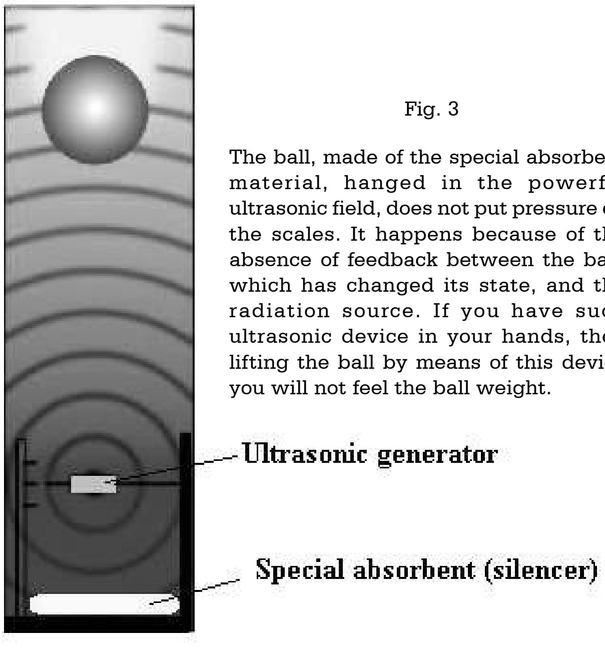

Fig. 3

The ball, made of the special absorbent material, hanged in the powerful ultrasonic field, does not put pressure on the scales. It happens because of the absence of feedback between the ball, which has changed its state, and the radiation source. If you have such ultrasonic device in your hands, then lifting the ball by means of this device you will not feel the ball weight.

The experiment of such type was made in one of the secret institutes. Water was used as an acoustic environment. The heavy ball easily floated up from the vessel bottom i.e. it behaved as if it had no weight.

The similar phenomena but such of the other organizational level, sometimes become apparent in nature. The phenomenon has its very name because the event happens unexpectedly and is not analyzable through laboratory research [3].

"A spherical ball lightning about the size of a football ball flew over a village Galtsovka of Altaisky Krai at a height of twenty to thirty meters. The first shed in its way, with ferroconcrete poles, was crushed and collapsed. Flying over a slated house, the ball lightning tore away the roofing slate together with nails, raised it into air and pulled it along, scattering its parts all over the village. Flying over a tractor station, the ball lightning crushed a frame, welded of metal angles and covered with tarpaulin. On the approaching of the ball lightning to another frame, it was at first dragged along the ground, and after the ball lightning had passed it, the frame was lifted and carried at a distance of 300 meters. The weight of the frame was no less than 100 kg".

Further the author analyzes the situation from point of view of the known physical laws: "The frame of hundreds kilograms weight was carried by fields of the flying ball lightning. However, for some reason the ball lightning kept on flying straight and did not note that some metal frame was caught to it. If the ball lightning, as it is usually considered, has a density of air and it is a weightless formation, then why the frame with the weight not less than 100 kg could not change its flight trajectory even in some extent?"

It draws attention that in some cases the ball lightning field pushed objects away, and in other cases it attracted them. It can be explained only by fact that in each case the ball lightning field specifically influenced on internal properties of objects, and then objects themselves somehow reacted on their new state (they changed their motion) . The feedback absence (in other way it happens between magnet and iron object) allows the ball lightning not react upon the things, which take place in its field. If these objects themselves became sources of similar field, i.e. reradiated it, they would influence on the ball lightning trajectory. **Most likely, in the given case, there was realized the situation when action caused no reaction.**

Conclusions

In the context of known physical laws the particular problem of direct action without feedback was formulated and solved. The solution of this problem provides deep understanding of how to control weight characteristics of material objects in open systems. As it was shown by the example with water (which is only an illustration of more fundamental processes) we can "deceive" the nature, but only through the deep understanding of the processual character of the concerned events.

By the example of ultrasound we show, that in principle there can exist such field flows, which are able to influence on bodies without the feedback effect. There were defined some conditions and criteria of bodies and methods of influence, at which the third Newton law is not directly valid (it is not published in this article).

The described type of influence can be called as pressure. But during consideration of processes at atomic or deeper level, i.e. from the position of rhythmodynamics, we find out something of another kind, that is phase-frequency one. It prompts to us, what kind of technologies will exist and how our aircrafts will look in nearest future. But not everyone is able to understand it at once.

Moscow 15.05.2002

For sponsors: mirit@yandex.ru
Our website: http://www.mirit.org

References

1. Physics textbook for the 9th class of secondary school // M., Prosveschenie, 1982

2. Ju.N. Ivanov, Rhythmodynamics // M., Noviy Tcentr, 1997

3. G.V. Nikolaev, Scientific vacuum // Tomsk, Kursiv, 1999

Comment

In the gravitation field we have a balanced open system, in which appearance of the new body at first sight must cause this system reaction such as weight increase. However, this is not the case.

The given problem is solved in two independent steps:

1) examination of interaction between water source and water (reactive effect, which is compensated by counter flow);

2) examination of interaction between water, detached from the source, and the body (unit time impulse per unit area).

It is impossible to examine the interaction between the source and the body because water has no rigidity. There are no means to influence on the source through the intermediate stream, therefore the principle of action and reaction is valid for each step individually, but not for both! But in this case the third Newton law must be developed as following: "If there is a 100% feedback between two bodies, their interactions are equal and inversely directed. And if there is no feedback, the action of one body to another causes no reaction, i.e. the action is not equal to reaction". And this is already another law!

New Possibilities of Vortex Electric Power Devices

Stanislav A. Lisnyak

Prospect Stoletiya Vladivostoka, 112-37
690069, Vladivostok, Russia
Tel/Fax: (7-4232)310-554
Email: Stasvladcomru@mail.ru

The earlier published article [1] was devoted to the explanation of the excess energy output at the rotation of flow medium in vortex tube (VT). The researches on VT demonstrated by the real examples that the excess energy output is a real fact, which is determined by natural physical processes. These processes take place in the fluid at its vortex (rotational) motion in the closed volume by means of phase transformations (skips) of the fluid state. However, the practice shows that for obtaining of patents and benefits for the real vortex devices, which are made for heat generation, it is necessary to write a little about their real technical characteristics, though technical and constructive possibilities of vortex devices can be much better. Temperature of water of these devices could not be higher than 120°C, and value of fluid pressure is confined within 5…6 kg/cm^2. Circulation pumps of pressure over 6 kg/cm^2 are used to increase thermal power output of vortex heat generators (VHG). It causes the discredit of VHG since at such a method their generative possibilities reduces to 100% value and less! We set a problem to discover the ways and engineering solutions for the essential increase of the coefficient of VHG energy conversion to not less than 200% value.

Our experience in development and exploitation of deep-sea physical devices has given some engineering solutions in this way. Actually, we have the opportunity to use the generally applicable circulation pumps, which have pressure not more than 4.0 kg/sm^2. At the same time we can raise the temperature of the heat carrier to 300°C and higher! and the pressure can be increased to 1000kg/sm^2! Such fluid parameters allow to use this fluid as a heat carrier in powerful hot-water generators of steam engines, turbines etc. By the way, this method allows to make the value of temperature of water enough to supply its PYROLESIS! The main point of the method is the following: all the closed system of water-filled VHG works at pressure with the value under 1000kg/sm^2. It is provided by means of the special device, the so-called automatic pneumohydraulic block (APHB). This lets to raise the temperature of working fluid to 1000°C with no change in its aggregative state. The practical functional scheme of such a device is shown in Fig.1.

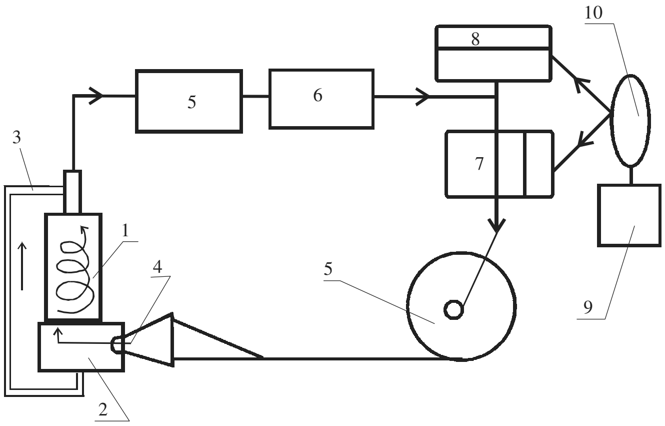

Fig. 1

1 – vortex tube; 2 – entry vortex chamber; 3 – by-pass;
4 – upstream end; 5 – circulation pump; 6 – heat-exchangers;
7 – APHB; 8 – volume of fluid leakage compensation;
9 - compressed-air flask; 10 - gas pressure regulator

The device works in the following way. The whole device is filled with working fluid without entrapped gas. At the operation of the circulation pump (5), fluid through the upstream end (4), placing in the entry vortex chamber (2), swirls, accelerates and gets into the vortex tube (1), where its "energy saturation" is realized. Then heated fluid gets into the heat-exchangers (6) for heating or for other purposes. After passing through the heat-exchangers, fluid gets into the pneumohydraulic device (7), where there is kept up the proper level of fluid pressure in the whole system. To prevent the system breakdown at the accidental fluid leakage, there is a device (8), which compensates such leakages. Compressed-air flask (9), with the volume under 1000kg/cm^2, and gas pressure regulator (10) keep up the selected level of the pressure in the system. The use of blocks (7,8,9,10) excludes evaporation in the system and prevents the breakdown of the circulation pump (5).

This VHG scheme allows to double the effectiveness of YUSMAR devices [2] only at the expense of rise in working fluid temperature in 2…3 times.

Observation of any fluid swirls and film documents of windspouts arrive to conclusion that all vortex structures are rotation bodies, created by the lines of the second order: $Y=aX^2$. In other words, as a result of the rotation in the swirl, air or fluid mass gets the acceleration of the second order. Taking into account the aforesaid, it is evident that in order to form the classical swirl in the VT, the very VT should be a tubular body of rotation. This body is created by the curve (see below Fig.2).

In YUSMAR and similar devices the vortex fluid motion take place in straight cylinder and 1/3 of this area is used for fluid deceleration that causes vortex flew disruption. This results in the impossibility to increase additional heat in the straight tube in more than 1.54 times. It is caused by the fact that the main vortex formation takes place only in the vortex chamber and the flows separation does in the tube itself. Then, this vortex formation is right away disrupted by different plate brakes! It suggests itself that output part of VT should be made in the spiral form, expanding at the flow passage.

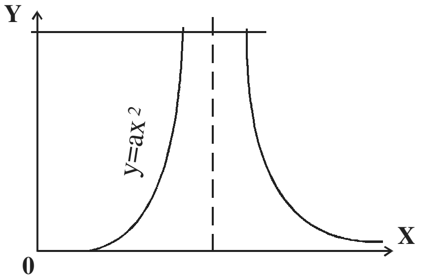

Fig. 2
Generatrix of the wall of vortex tube

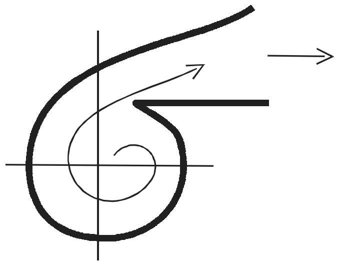

Fig. 3
Upper part of the tube

Such section, made on the outlet of the vortex tube, helps to slow down the flow avoiding its disruption, since axial-cold flow, placed along the tube axis, should stay intact. Two of such engineering solutions allow raising the level of the additional VHG energy up to 180…200%.

YSMAR devices are designed for one tube with the assigned heat output. However, if to take into consideration and to use the aforesaid conclusions, then we can raise energy conversion coefficient (ECC) above 200% at the series connection of two and more VT. The scheme of such a device is shown in Fig. 4.

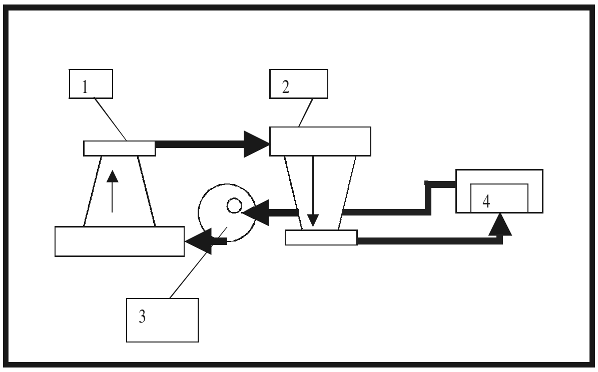

Fig. 4

Where: 1 – the first VT; 2 – the second VT; 3 – circulation pump; 4 – heat exchanger

According to Fig.4, the total energy output of the presented device is:

$$Q=K^2$$

Where: Q is the total output of the device; K is energy conversion coefficient of one VT; 2 – quantity of serially connected VT; P – the assigned power of the circulation pump.

In our earlier article [1], concerning VT application, we supposed that there is a straight transformation of vortex motion of ionized fluid into electric current. Under studying of many articles, devoted to the methods of water ionization, we discovered a significant one. It demonstrates that at the determined temperature, pure water increases its ionization capacity up to 3 orders without changing in the aggregative state [3]. The diagram in Fig.5 shows such dependence.

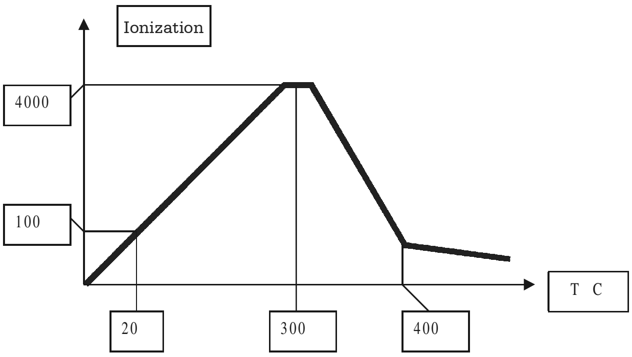

Fig.5
Diagram of the dependence of pure water ionization from temperature

The diagram demonstrates that water at 300°C temperature increase the quantity of ions in 4000 times as compared to 0°C and in 40 times as compared to 20°C. If to take into consideration that first experiments on VHG modernization help us to find the ways of fluid heating up to the practically any temperature without changing of its phase state, then on using the dependences (see Fig.5) it seems possible to create vortex fluid electric generators of the forward transformation. The aforesaid pneumohydraulic block, which is used for VHG functioning, allows to keep up water temperature at 300°C and pressure at about 90kg/cm^2 without threat of water evaporation! Presence of any rotation of fluid medium always causes appearance of two vortex flows. **These flows always move and rotate in opposite directions and if we do not put obstacles for their motion then they transform into each other and can exist without energy supply for an indefinite period of time.**

The presented VT form (see Fig.6) provides the producing of two fluid flows in the tube. These flows do not influence to each other and there is only their reciprocal overflow without disruption of the flows in the central part of VT. The tube presents a tubular body of rotation, which is created by hyperbola. There is positive angular acceleration of fluid in the lower part of the tube; maximum speed of rotation of the axial and peripheral flows is in the middle part and negative angular acceleration, i.e. deceleration and transfer of kinetic energy into heat energy, is in the upper part of VT.

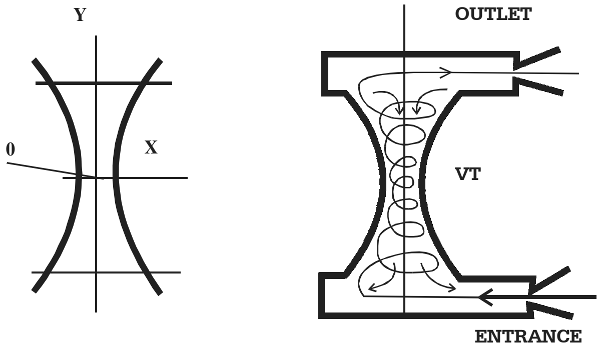

Fig.6

The maximum heating of fluid and its polarization is in the point Y=0. The polarization potential in the point can reach 10.000 V at 5m/sec flow speed for 10 cm diameter in the point. The cold flow, which comes along VT axis, is opposite to the peripheral hot flow. These flows close into one system in lower and upper chambers of VT. Thus, the presented VT classically modeled "rolling" and "unrolling" of water area.

There are no single vortexes in nature. Two vortexes, rotating in opposite directions, always appear while fluid mediums rotation! Science is not still able to describe energy overflows from one vortex center to the center of another one by means of mathematics. However we believe that this moment is close. There are maximum energy transformations of rotating water in the point Y=0.

The potential of the inner flow is equal to the outer flow by its value but is opposite by its sign. Potential difference is maximum in the point Y=0, in which removal of electric charges is the most effective.

It is appropriate mention here the name of Romanian engineer and researcher Henry Koanda, who in 20th of the last century discovered the so-called "Koanda effect" (attachment of fluid jet to the surface of usual kettle at pouring of it out the cups). Basing on this effect he suggested to make new type of aircrafts, which could have advanced bearing capacity and maneuverability. Unfortunately, money and conservatism of aircraft-industrialists did not allow realization of these engineering solutions.

At the same time, Koanda charged his disciple Patrick Flanagan with the job to research all **water properties**. The result was amazing! It was turned out that water actually has infinitely many phase states and when it is moving, then it can trap energy from the environment by some way. The publications on the point appeared in our press in the early eighties, late nineties of the last century. Most likely, they had become the basis of the invention, made by Potapov, i.e. vortex tubes for water!

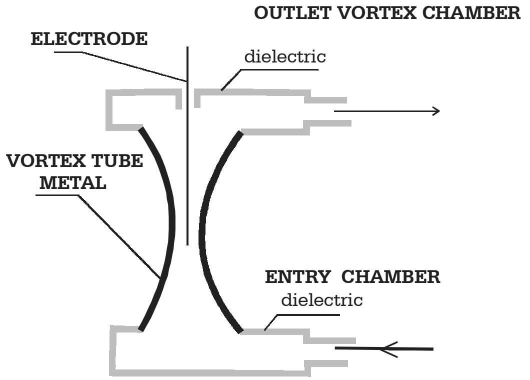

Fig.7
Scheme of vortex electric generator

In the presented scheme of electric generator the entry and outlet vortex chambers are made of dielectric material in the spiral form. This material certainly should withstand not less then 300°C of fluid temperature and about 100kg/sm^2 of pressure. Fluoroplastic or ceramic can be used as such a material.

From aforesaid it seems to be possible to submit the new scheme of electric generator for readers' consideration (see Fig.8).

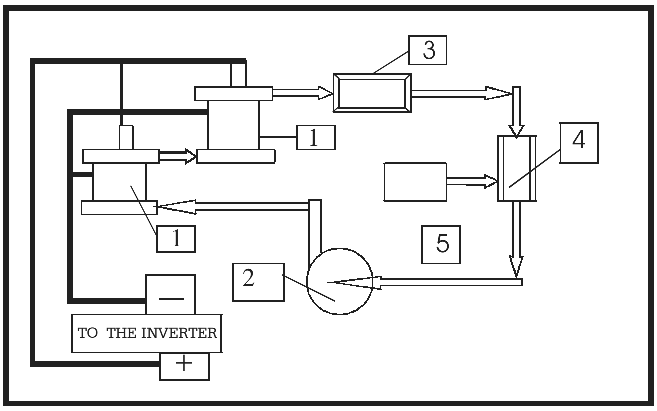

Fig.8
Generalized scheme of vortex electric generator
1 – vortex tubes; 2 – circulation pump; 3 – heat exchanger; 4 – automatic pneumohydraulic block; 5 - compressed-air flask

Actually we could finish the article if it were no disputes about excess energy, which is released at vortex motion of fluid in tubes. Having used a simple experiment, which any inquisitive person can repeat, we found the positive solution (see the scheme of the experiment in Fig. 9).

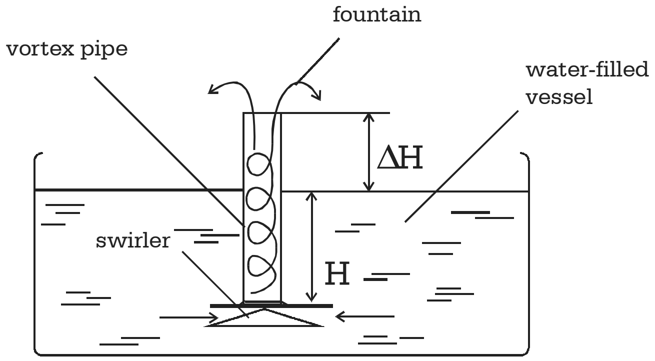

Fig.9
Demonstration of energy increase of fluid, rotating in vortex tube

A pipe with not less than 60 cm length and about 1-inch diameter is placed in a reservoir with not less than 50 cm depth. The lower part of the pipe is formed as an arm circular generator with about 15...17cm diameter. The gap width is about 2 cm. If we place the pipe with closed top end, which then will be open, into the reservoir, then water is splashed out above 20 cm of the reservoir level. Thus at the least 20% energy increase of energy mgH is observed.

By the way, we suggest orthodoxies of physics and other who have doubts to make one more simple experiment. Try to boil water in a can. When water starts to boil, swirl it with a spoon to the right or to the left. Then you will suddenly discover that all water surface is calm and only in the center of the can there is a water-steam mixture, which is about 20% higher than the level of water surface. You will also see that water is absolutely calm at walls of the can. This experiment can be reproduced anytime and anywhere, even in space, since it is a demonstration of **vortex** fluid motion and its actual influence on everything, which is around!

All presented engineering solutions are practically reproducible in any laboratory. The authors of the article would welcome the opportunity to co-operate with science and industry representatives for realization of the ideas in real serial production devices.

References

1. O.V. Gritskevitch, S.A. Lisnyak, On possibilities of vortex energetics //New Energy Technologies. 2002, #1(4), p.2-3
2. Patent RF #2045715, 1993
3. Chemical and Engineering News. 2000 #1, p.26

Investigation of Electric Energy Transmission Processes in non-Metallic Conducting Channels

D.S. Strebkov, A.I. Nekrasov, S.V. Avraamenko

All-Russian scientific Research Institute of Electrification
of Agricultural Economy
2 Vishnyakovsky passage, 1 Moscow, 109456, Russia,
energy@viesh.msk.su

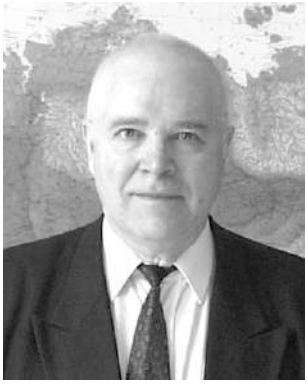

D.S. Strebkov

(Editor's notes by Alexander V. Frolov)

It has been found that at low frequency (1-25 kHz and higher) electric power can be transmitted with low losses from generator to receiver along single channel made of non-metallic conductive media such as water in plastic tube carbon thread, layer of damp soil, ITO films on glass substrate, laser and electronic beams. Transmitted power, as well as for traditional three phase lines, is limited by natural power of the transmission line and line capacity may reach at high voltage and pulse and operation modes the value 10^9 Wtt.

Introduction

The well-known methods of electric power transmission are based on transmission of active energy by means of conductivity currents in closed circuit. Electromagnetic energy spreads along power transmission lines (PTL) as progressing waves of electromagnetic field or field of charge [1]. Line wires made of aluminum or copper are conductive (guide) channels. Electromagnetic energy stream moves along these channels from generator to energy receiver and backwards to the generator. Maximum transmission possibility of 3-phase PTLs is limited by losses on the line resistance, by peak voltage (which is determined by electric strength of the insulation) and by electromagnetic stability of the line.

The modern approach to provide the electromagnetic stability consists in rigid regulation of line parameters by means of high-speed shunt reactor and consequence capacitive compensation for the purpose to except changes of electromagnetic power flows and to suppress resonant properties of a line [2].

In Tesla works [3] and in the researches of Russian scientists [4] a method of active power transfer was offered. This method supposes to transfer active power by means of electromagnetic capacitive current assisting with resonant properties of a single-wire line (SWL), made of a metal conductor. The purpose of the present work is a research of an opportunity to use non-metal conducting mediums for transmission of electric energy.

A pipe with not less than 60 cm length and about 1-inch diameter is placed in a reservoir with not less than 50 cm depth. The lower part of the pipe is formed as an arm circular generator with about 15...17cm diameter. The gap width is about 2 cm. If we place the pipe with closed top end, which then will be open, into the reservoir, then water is splashed out above 20 cm of the reservoir level. Thus at the least 20% energy increase of energy mgH is observed.

By the way, we suggest orthodoxies of physics and other who have doubts to make one more simple experiment. Try to boil water in a can. When water starts to boil, swirl it with a spoon to the right or to the left. Then you will suddenly discover that all water surface is calm and only in the center of the can there is a water-steam mixture, which is about 20% higher than the level of water surface. You will also see that water is absolutely calm at walls of the can. This experiment can be reproduced anytime and anywhere, even in space, since it is a demonstration of **vortex** fluid motion and its actual influence on everything, which is around!

All presented engineering solutions are practically reproducible in any laboratory. The authors of the article would welcome the opportunity to co-operate with science and industry representatives for realization of the ideas in real serial production devices.

References

1. O.V. Gritskevitch, S.A. Lisnyak, On possibilities of vortex energetics //New Energy Technologies. 2002, #1(4), p.2-3
2. Patent RF #2045715, 1993
3. Chemical and Engineering News. 2000 #1, p.26

Investigation of Electric Energy Transmission Processes in non-Metallic Conducting Channels

D.S. Strebkov, A.I. Nekrasov, S.V. Avraamenko

All-Russian scientific Research Institute of Electrification
of Agricultural Economy
2 Vishnyakovsky passage, 1 Moscow, 109456, Russia,
energy@viesh.msk.su

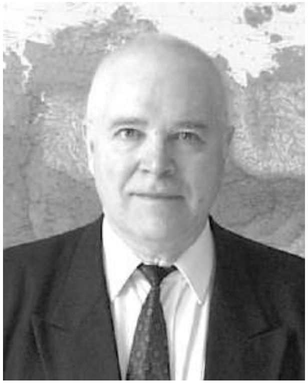

D.S. Strebkov

(Editor's notes by Alexander V. Frolov)

It has been found that at low frequency (1-25 kHz and higher) electric power can be transmitted with low losses from generator to receiver along single channel made of non-metallic conductive media such as water in plastic tube carbon thread, layer of damp soil, ITO films on glass substrate, laser and electronic beams. Transmitted power, as well as for traditional three phase lines, is limited by natural power of the transmission line and line capacity may reach at high voltage and pulse and operation modes the value 10^9 Wtt.

Introduction

The well-known methods of electric power transmission are based on transmission of active energy by means of conductivity currents in closed circuit. Electromagnetic energy spreads along power transmission lines (PTL) as progressing waves of electromagnetic field or field of charge [1]. Line wires made of aluminum or copper are conductive (guide) channels. Electromagnetic energy stream moves along these channels from generator to energy receiver and backwards to the generator. Maximum transmission possibility of 3-phase PTLs is limited by losses on the line resistance, by peak voltage (which is determined by electric strength of the insulation) and by electromagnetic stability of the line.

The modern approach to provide the electromagnetic stability consists in rigid regulation of line parameters by means of high-speed shunt reactor and consequence capacitive compensation for the purpose to except changes of electromagnetic power flows and to suppress resonant properties of a line [2].

In Tesla works [3] and in the researches of Russian scientists [4] a method of active power transfer was offered. This method supposes to transfer active power by means of electromagnetic capacitive current assisting with resonant properties of a single-wire line (SWL), made of a metal conductor. The purpose of the present work is a research of an opportunity to use non-metal conducting mediums for transmission of electric energy.

Methods

For realization of the experiments a single-wire energy system (SWES) was used. Its electric circuit is shown in Fig. 1 a,b. SWES consists of the following parts: high-frequency generator (1) of 28V voltage AC and 1 kWtt power; transmitting (2) and receiving (3) Tesla transformer with conductive channel (4) between them; rectifier (5) and electric load (7) as an incandescent lamp or electric motor (220V, 1 kWtt power). High-voltage winding of Tesla transformer is made in the form of cylindrical winding on the ferrite core with 50-100mm in diameter, 4000-6000 turns. The internal end of a high-voltage winding is connected to the conducting channel, and the external end is not connected at all (free end). A low-voltage winding that consists of 40-60 turns provides power supply of Tesla transformer. At the submission of electric power from the high-frequency generator to the winding of power supply zero potential appears on the free end of the high-voltage winding, and voltage with corresponding to the generator 1-25 kHz frequency is formed on the conducting channel. Besides, Tesla transformer as the spiral antenna generates electromagnetic waves of 5-10 cm length , which corresponds to the turn length of a high-voltage winding. Current resonance mode is formed in a supply circuit of Tesla transformer. At the same time, voltage resonance mode (of frequency equal to the generator frequency) is formed both in the reception and transmitting circuit of high-voltage windings and in the conducting channel (1).

The following materials were used as the conducting channels: filled with water or sea water polythene tubes with 10 mm diameter and 1,5 m long; plastic tray with soil layer of 150×10×400mm; film of tin dioxide – indium oxide (ITO) on 0,3 micrometer thick and 300 Ohm resistant glass substrate; graphite thread of 0,1 mm diameter, 500 mm long and with 100 Ohm resistance. For comparison steel and copper wire of 0,1 mm diameter and 5 m long was used also.

The voltage on the conducting channel changed in limits 1-10 kV, generator frequency changed from 1 up to 25 kHz. Voltage, current and capacity were measured on SWES output and on the load by standard electric measurement equipment.

Results and discussion

The replacement of the metal conductive channel to the channel made of non-metallic conductive materials does not effect to any reduction of possibilities of transmitting SWES or heating of a material of conducting channels. The consecutive connection of the channels did not resulted in reduction of transmitted power. Circuits break in the conducting water channel by the creation of an air gap led to the occurrence of arc discharge of reactive capacitor current. However this discharge did not cause increase of water temperature at transmitted power 300 Wtt and voltage 4,5 kV within 1 hour, that confirms the absence of energy losses in the conducting channel. The increase of water temperature did not result in decrease of transmitted power. Water PH reduction from neutral value up to 4 was detected.

The increase of sea salt concentration in water up to the level of 5-7 g/l did not increase transmitted power in comparison with tap water. However the replacement of tap water to dionized water with 16 MOhm resistances resulted in 100 % decrease of transmitted power.

Thus, it is experimentally shown, that conducting channels, produced from non-metal materials, have quasisuper-conductive properties in SWES at resonant mode. Possible explanation of this effect is the absence of active conduction current in the channel and the dominant role in the process of energy transfer belongs to displacement current, for which Joule Lentz law is not valid [11]. In the supply circuit of Tesla transformer current is practically reactive, and in resonance conditions active values of inductive and capacitive currents are equal. Vectors of these currents are opposite in their phases. Current of the high-frequency converter is spent for losses (component less than 2 %) in supply circuit wires and in the core of Tesla transformer and also for creation of reactive current in the conducting channel. In the mode of voltage resonance we have measured voltage active values of high-voltage inductance windings and conducting channel, interturn capacitance of windings and the capacities of the conducting channel. **We have discovered that they were equal and their phases were opposite to each other.** Losses from transmission of capacitive current through active resistance of the conducting channel are insignificantly small. Corona discharged losses and leakage current could be decreased by isolation of conducting channel. In this case active current and magnetic field of the line are equal to zero. Electric field of the line has maximum value. As well as in usual PTLs, maximum transferred power is limited by charge power of the line. Angle between vectors of voltage is equal to zero in the beginning and at the end of a line. Quality factor of SWES at frequency 5 kHz is in 100 times above than usual PTLs at frequency 50 kHz. In the conditions of resonance it leads to the significant increase of voltage along the conducting channel and it also leads to transmission power.

In usual PTLs voltage changes along the line are insignificant. The angle between vectors of voltage in the beginning and in the end of PTL constitutes the value, which is proportional to the wavelength of line.

On the basis of the researches the methods and devices are offered for transmission of electrical energy through plastic water guide, electro-insulated from ground, through irrigational channels and through isolated pipelines, which are used for gas, oil, hot and cold water transportation. Also the energy can be transferred through fiber-optic cable with conducting film on the surface, through all-carbon composite cable and through electro-insulated part of a ground and water surface, including highway parts. There are also methods and devices, designed for the following transmissions. These methods can be applied for transfer of power to stationary and mobile units. There are also generated requirements to electrical safety and to restriction of use of drinking and hot water from pipelines, which are under electrical voltage. These requirements and restrictions are generally the grounding of pipeline parts, which are located on certain distance from the generator. This distance is equal to the whole number of half-waves and for it SWES voltage is equal to zero. In the case of a side pipeline it is necessary to ground the parts of the pipeline that are placed on the distance of odd number of quarter-wavelength from the main pipeline. For 5kHz frequency the quarter-wavelength is equal to 15000 m.

N. Tesla grounded one end of high-voltage windings of his transformer on the receiving and on the transmitting end of SWES. Tesla considered this condition as the necessary one for the transfer of power along the Earth. The results of our researchers demonstrate that it is not necessary to use metal self-closed conductor (and current lines in the Earth) for transmission of electric energy on low frequency (1-25 kHz).

For this frequency energy could be transferred from the generator to the receiver if we have single-wire guide system created as non-metal conducting channel. By the similar way electromagnetic energy is transferred by laser beam or microwave-beam. But in our case we can obtain high degree of efficiency that is caused by slight losses on energy absorption and energy emanation. Thus one of the ends of high-voltage winding at the energy generator will have zero potential and remain free. The symmetric end of a high-voltage winding on the reception end should be connected to some natural capacity 6 (Fig. a), which can represent the case of a balloon or frame of a tractor. In our experiments we used metallic safe-box as such natural capacity.

Editor's note: In 1887, October 11, the famous Russian scientist Pavel Yablotchkov got the France Patent #120684, which described the method to increase efficiency of electrical circuit by means of "atmosphere electricity". It was confirmed in many experiments that output power can be twice more than power provided for the circuit from primary electric generator. For that it is necessary to use a single-plate capacitor. The special feature by Yablotchkov, that provides maximum efficiency, is the high degree of air ionization. For this Yablotchkov proposed to use the special capacitor, which consists of a big number of metallic needles. By its view this construction reminds of a hedgehog. Thus we can assumed; that above described natural capacity (balloon or frame of a tractor) serves as a collector of free electrons. By Yablotchkov the efficiency of such systems can be increased by means of maximization of ionization process. It will not lead to the increase of losses if the second end of the high-voltage winding is not grounded.

In the other method of energy transfer, a condenser-diode block 8 was connected to the conducting channel on the receiver end. This block is one of the known circuit of voltage doubling, Fig.1 (b). On the condenser 8 electric energy is transferred through electronic switch 9 to load 7. In this case the entire length of conducting channel 4 and Tesla transformer winding 2 at generator must be equal to odd number of quarter wave-lengths.

Non-metal conducting channel (for instance, fiber-optic or coal-plastic cable) can be used for transfer of electric energy not only along but also as perpendicular to the Earth (for example, to relay aerostat or sounding balloon).

SWES conducting channel can be also created by ionization of air ions with laser beam [13]. Neodymium laser with double frequency and with energy 1 Joule in impulse is able to create 10^{15} cm^{-3} ion concentration in air. This concentration is sufficient for streamer initiation and for transfer of electric energy through the conducting channel. Ionization potential, time of ion existence and of excited molecule state, coefficient of multiquantum absorption, all these determine the limiting length of conducting channel in atmosphere that is equal to 300 km and its wave resistance at 200-400 Ohm. Voltage which is necessary for SWES comes to the quantity 0,5 MV – 15 MV, that depends on the length of a channel.

We suggest to use relativistic electron bunches of high energy as the conducting channel out of the atmosphere. As distinct from laser bunches they do not have divergency. In this connection the Moon or artificial conducting body, where the energy receiver is placed, can be used as natural capacity 6. Whereas energy generator can be installed on the Earth or on its satellite. Transmission range of electric energy is determined by the length of the generated conducting channel. The entire length of the conducting channel in the beginning and in the end must be equal to the whole number of half-waves. Here the length of high-voltage windings of two Tesla transformers must be taken in considerations. Electric energy, transmitting through the conducting channel, can exceed the energy of electron and laser beams generators in 10-100 thousand times. These generators play the role of a directing system (of usual SWEG wires), along which the transfer of electric energy proceeds.

It was offered to use colliding and crossing electron and laser beams with conducting transitional bodies as conducting channels for transfer of energy from the Space to the Earth and back. On the heights up to 30 km compositional coaly and fiber-optic cables can be used. To create the global energy system of the Earth it was also suggested to apply single-wire energy system and conducting layers of the Earth ionosphere as the conducting spherical channel [14].

Thus for electric energy transfer at the frequency 1-25 kHz and higher in the resonance mode a single-wire channel from the following non-metal conducting mediums can be used: water, damp soil, coal-plastic, oxide film, ionized air channels, that are created by laser beams in the atmosphere, ionosphere conducting layers, and also beams of relativistic electrons out of the atmosphere. These non-metal conducting channels in the resonance mode have negligibly small resistance losses especially if to be compared with metal conductors, which are used in the known non-resonance methods of energy transfer by means of active conduction currents in the closed circuit. Electric energy in the resonance mode can be transferred with small losses from the generator to the receiver along the single-wire channel, made of non-metal conducting materials. The transfer can be realized at the frequency 1-25 kHz and higher, to any distance and to any direction relatively to the Earth. The transmission capacity is limited by charge power of a line as well as in the usual PTL. At high voltage the transmission capacity can reach the quantity from 10 Wtt to 10^9 Wtt in the pulse and streaming modes.

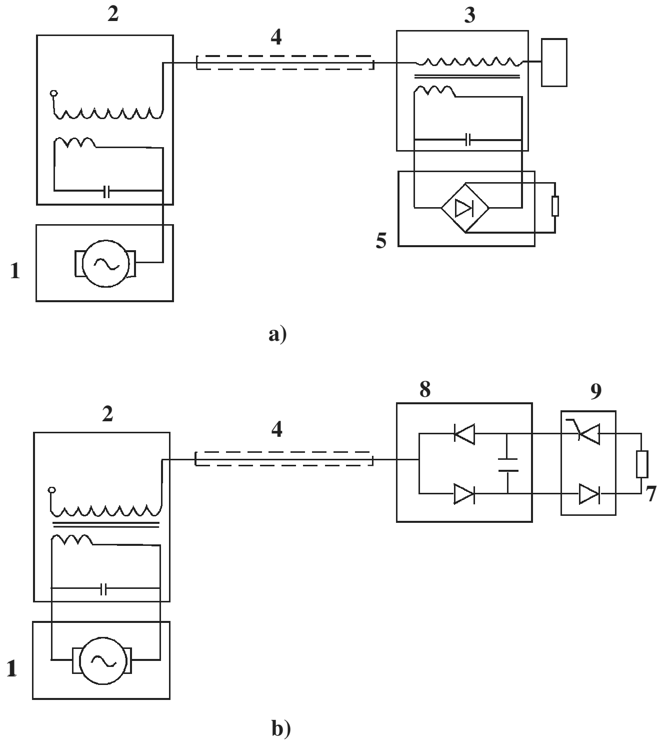

Fig. 1
Electric circuit of single-wire energetic system
with non-metal conducting channel
a) SWES with symmetric array of Tesla generators

b) SWES with diode-capacitor block in the end
of conducting channel

References

1. A.M.Kalashnikov, Ya.V Stepuk. Bases of radio engineering and radiolocation. Oscillating systems. M: Department of USSR Defense 1965, - p.127-158.
2. G.N.Alexandrov, S.V. Smolovic, Flexible lines for electric energy transmission over long distances. V Symposium "Electrical Engineering-2010", October 12-22, 1999, Moscow Region P., 35-42.
3. N.Tesla Apparatus for transmission of electrical energy. US Pat #349621, 15.05.1900.
4. S.V. Avraamenko Mode of electric devices supply and device for its realization. Pat. RF #210649 11.04.1995. Publ. 10.04.1998. Bulletin #10.
5. D.S.Strebkov, S.V.Avraamenko, A.I. Nekrasov Device for power supply of mobile electrical aggregation. Pat. RF #2158206, 15.04.1999. Publ. 27.10.2000. Bulletin #30.
6. D.S.Strebkov, S.V.Avraamenko, A.I. Nekrasov Mode of supply of electric transport devices and device for its realization. Pat. RF #2136515, 26.08.1998. Publ. 10.09.1999. Bulletin #25.
7. D.S.Strebkov, S.V.Avraamenko, A.I. Nekrasov One-trolley system of power supply of mobile electrical aggregations. Materials of the international scientific and technical conference on automation of agricultural manufacture. Minsk 6-8 July 2000, - p. 65-66.
8. D.S.Strebkov, S.V.Avraamenko, A.I. Nekrasov Research on single-wire system of electric energy transmission. International conference Ecology and agriculture. Technique. – Saint-Petersburg, 2000, - p.50-55.
9. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. Single-wire electric system for renewable-based electric grid. Third European Solar Congress "Eurosun 2000", June 19-22, 2000, Copenhagen, Denmark p., 10.
10. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. SWEPS – a novel electric power transmission technology. Planetary Association for Clean Energy Newsletter, 2001. – vol. 11, - 3, - p. 7-11.
11. I.E. Tamm. Bases of electricity theory. – M.Science, 1976. – p. 133, 397-400.
12. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. Mode and device for electric power transmission. Pat. RF #2172546, 24.01.2000. Publ. 20.08.2001. Bulletin #23.
13. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. Mode and device for electric power transmission. Pat. RF #2143775, 25.03.1999. Publ. 27.12.1999. Bulletin #36.
14. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. Mode and device for electric power transmission. Pat. RF #2161850, 14.07.1999. Publ. 10.01.2001. Bulletin #1.

Microcosm, Universe, Life

Alexey G. Shlienov
Budapeshtskaya, 66-1-77,Saint Petersburg, 192286 Russia

Ernest L. Petrov
Marshala Zaharova, 50-809, Saint Petersburg, 198330 Russia,
Fax.: (812) 587-9349

The authors of this article have prepared the 3rd edition of the book "Microcosm, Universe, Life" and now they are looking for a publisher to edit Russian and English version. In particular, the mechanism of energy transformation in the Universe (Fig.1) is described in the book. Matter radiates energy in the form of photons (as well as neutrinos). In general this energy is observed in the form of Cosmic Microwave Background Radiation (CMBR). Photons and neutrinos transfer the most of energy to de Broglie longitudinal photons. Matter absorbs energy mainly in the form of longitudinal photons. Thus, a cycling process of energy transformation (as well as matter transformation) occurs in the Stable Universe.

The 3rd edition is devoted to the description of 6 World Systems (Table 1), at that the 6th System is based on Unified Field Theory, which is developed by the authors, and proceeds from the following:

(Editor's: The understanding of the energy transformation mechanism allows to develop new fuel-less energy sources based of mutual transformations: longitudinal photons⟵⟶transverse photons).

Table 1

System	Principal Ideas
The 1st – Geocentric (Ptolemaic system)	Geocentrism and Anthropocentrism
The 2nd – Classic (Copernicus – Newton's system)	Description of the Universe on basis of Newton's Law of Gravity
The 3rd – Compromise (Tycho Brahe's system)	The compromises between two first systems
The 4th – Fractal (Charlie's system)	Non-heterogeneity of Large-Scale distribution of matter in the Universe
The 5th – Relativistic (Einsteinian system)	Description of the Universe on basis of gravitational field equations. Anthrop principle
The 6th – based on Unified Field Theory	Electromagnetic nature of all physical interactions. Irreversibility of all elementary micro-processes in combination with circular character of transformations in Large Scale of space ($c/H=R$ order) and time ($t_H=I/H$ order).

1. For theoretical results, which agree with the facts, it is necessary to use assigned inertial system (as Lorentz did), which, as we have known, is connected with CMBR. Obtained by this way decisions can be applied to the other systems, in particular, with use of Lorentz transformation for mass, energy, momentum, time and length.

2. Theory must agree with the whole known collection of facts, starting with such established empirical generalizations as basic principles of Natural science (Giordano Bruno, Lyell) and conservation laws (Mayer, Joule, Helmholtz, Faraday, Newton, Huygens, Vernadsky).

3. All physical objects could be presented as the systems, consisting of quanta of positive electric charge (protons), quanta of negative electric charge (electrons in fermions or boson states, which are part of neutrons

and atomic nuclei) and quanta of energy, impulse, moment of momentum (apeirons).

4. The leading role in the Universe energetic belongs to de Broglie longitudinal photons (of the first class, i.e. with the spin I=0), energy of which is twice as much as apeiron energy and is equal to *h H*, where *h* – is Plank's constant, *H* – is Hubble's constant.

5. Longitudinal photons, polarizable and depolarizable by matter particles, are the carriers of gravitational, magnetostatic and strong interaction. Electromagnetic interaction is carried by photons with spin I=±1 and weak interaction is carried by neutrinos and apeirons with spin I=±0.5.

6. All elementary microprocesses are irreversible, that correlates with circular character of transformations in Large Scale of space and time.

7. In particular, there is an irreversibility of the process of photons and neutrinos motion. At each segment, equal to the length of de Broglie wave, photons and neutrinos lose energy *h H*, which is equal to longitudinal photon energy. In such a way, Hubble's law can be presented as the following:

$$\lambda = \lambda_1 \exp\left(\frac{r}{R}\right) = \lambda_1 \exp(Ht),$$

$$Z = \frac{\lambda}{\lambda_1} - 1 = \exp\left(\frac{r}{R}\right) - 1 = \exp(Ht) - 1,$$

where λ, λ_1 – are observable and laboratory wave lengths, r – is distance, R – is radius of gravitation interaction, which is equal to the radius of Metagalaxy, t – is time, Z – is cosmological red shift.

8. As resulted upon this process, the excess of longitudinal photons is absorbed by matter. At that, mass is considered as measure of capacity, which is absorbed by matter in form of longitudinal photons. Thus, energy of the matter should be increased according to the law:

$$E(t) = m_{\circ}c^2 \exp\left(\frac{\varepsilon'}{c^2}t\right) \approx m_{\circ}c^2 + m_{\circ}\varepsilon' t,$$

at that,

$$\frac{\varepsilon'}{c^2} = \frac{H}{137^2},$$

where ε'- is the capacity, which is absorbed by unit mass in form of longitudinal photons, c – is velocity of light. In particular, electron of the 1st Bohr orbit in hydrogen atom (i.e. in this case at de Broglie wave length) absorbs energy *h H* at 1 period.

9. Being in ionization state, intergalactic matter is the general portion of average density of matter in the Universe ρ_{av}, besides, the average value of absolute velocity of baryon component is close to c/137. Let's consider this correlation to be exactly executed.

CMBR is the most powerful cosmic radiation. Its specter is close to that one of black body at temperature T_F=2.726K, i.e. its spectral density is maximal at frequency ν=160GHz. Cosmological red shift causes the increase of spectral density in radio-region (ν<160GHz) and the decrease of it in microwave region (ν>160GHz). Each of these processes is compensated by the inverse Compton effect, i.e. by the dispersion of radio-photons at matter corpuscle, first of all at protons. Calculations demonstrate, that average energy of absorbent radio-photons is equal to $0.45\cdot10^{-15}$erg (ν=68GHz) and average energy of radiated microwave photons is equal to $2.17\cdot10^{-15}$erg (ν=330GHz). There is one re-radiated microwave photon per one absorbed radio-photon. At that, the concentration of photons and CMBR spectrum remain unchangeable. Syunyaev and Zeldovich concerned the close inverse Compton-effect of CMBR dispersion on electrons in clusters of galaxies. Actually such an effect was discovered in 2 clusters of galaxies.

Thus, on transferring of energy to photons, matter corpuscles must fill the deficiency of energy by receiving it from longitudinal photons. In fact, there are observed demonstrations of longitudinal photons, they are "static fields". In this case it is a cosmic magnetic field, which accelerates charged particles of matter (Alfven). More detailed consideration let us to find a virial correlation between 4 main components of energy density:

$$\frac{\rho_{av}c^2}{137^2} = E_F = 2E_M = 2E_{kin},$$

where $r_{av}c^2$- is an energy equivalent of mass density of matter; E_F- is energy density of CMBR; E_M - is average energy density of magnetic field; E_{kin} - is average density of kinetic energy. Thus, some kinds of energy circularly transfer to another, which are interrelated (Fig. 1).

$$\frac{\varepsilon_{av}\rho_{av}V}{1.26} = E_F HV = 2E_M HV = 2E_{kin}HV = \\ = \frac{(\varepsilon_{av} - \varepsilon')\rho_{av}V}{0.26} = \varepsilon'\rho_{av}V,$$

where ε_{av} - is average capacity, radiated by unit mass in form of photons; V – is volume of 10^{81} cm^3 order, according to which the averaging is made.

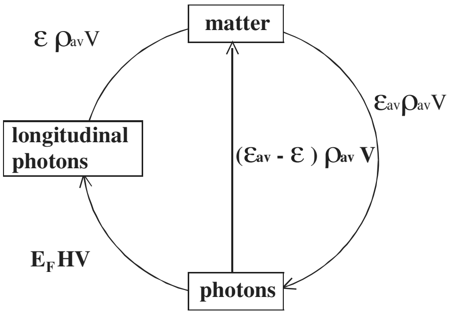

Fig.1

Energy transformation in Large Scale ε_{av}=0.0942 erg gr^{-1}sec^{-1} – is average energy, which matter unit radiates in unit time; ε'=0.07476 erg gr^{-1}sec^{-1} – is energy, which matter unit absorbs in form of de Broglie longitudinal photons in unit time; ρ_{av}=$0.8730\cdot10^{-29}$ gr cm^{-3} – is average matter density in the Universe; E_F=$4.18\cdot10^{-13}$ erg cm^{-3} – is energy density of cosmic microwave background radiation (CMBR); $V\approx10^{81}$cm^{-3} –is volume at which the averaging is made; H=$1.562\cdot10^{-18}$ Hertz.

10. These correlations include pressures, densities and temperature T_F, i.e. we have the equation of the state of intergalactic matter, and thus of Metagalaxy and the Universe. Let us present this equation in simple form:

$$E_F = \frac{\varepsilon' \rho_{av} R}{c} \approx \frac{\varepsilon_{av} \rho_{av} R}{c}$$

and let us compare it with correlation for density of radiant energy on surface of a star * or of star system

$$E_* = \frac{\varepsilon_* \rho_* R_*}{3c} \approx \frac{\varepsilon_* \rho_* R_*}{c}.$$

It is an especially amazing analogy between Metagalaxy (and the Universe) and large-scale cosmic system *g*, for which

$$E_g \approx \frac{\varepsilon_g \rho_g R_g}{c},$$

at that $E_g \approx E_F$, $\varepsilon_g \approx \varepsilon_{av} \approx \varepsilon'$.

This equation agrees with the whole observed data of globular clusters, galaxies, groups and clusters of galaxies, in particular, with empirical correlations, which are magnitude - angular diameter.

11. The stated approach lets to determine the values of many fundamental constants by different ways. As a case in point, the results of definite values ε_{av} and ε', erg g^{-1} c^{-1} are presented in the Table 2.

Table 2

#	Data	ε_{av}	ε'
1.	The boarder between two parts of the Main Sequence of stars	of 0.1 order	of 0.1 order
2.	The results of recalculation of observed star characteristics as respects to their centers	of 0.1 order	of 0.1 order
3.	The local minimum of star luminosity function near the Sun (according to G.A. Starikova's data)	of 0.1 order	of 0.1 order
4.	The correlation: mass - luminosity of white dwarf stars	of 0.1 order	of 0.1 order
5.	The correlation: mass - luminosity of neutron stars	of 0.1 order	of 0.1 order
6.	The correlation: mass - luminosity of globular clusters	less than 0.5	less than 0.5
7.	The correlation: mass - lumonosity of elliptic galaxies	more than 0.06	more than 0.06
8.	The correlation: mass - luminosity of spiral galaxies and irregular galaxies *Ir I*	less than 0.5	less than 0.5
9.	The correlation: mass - luminosity of galaxies as a whole	of 0.1 order	of 0.1 order
10.	The correlation: mass – luminosity of clusters of galaxies	of 0.1 order	of 0.1 order
11.	Empiric values of H, ρ_{av}, E_F constants	of 0.1 order	of 0.1 order
12.	The most exact values of h, e, G, E_F constants	0.0942	0.07476

12. Uncontradictory description of Microcosm and the Universe promotes the better understanding of Life, which is inseparably unified with them. The statistical data manipulation of more than 100 catalogues of microphysics and cosmic objects let us to get more than 1000 empirical correlations and diagrams, and to determine, that they agree with theoretical correlations, which were received according to the ideas on circular character of energy transformations (Fig. 1) and of matter in the Universe.

The Russian edition of the book is mailed out: the 1st edition (1995) at the cost of $3; the 2nd edition (1998) at the cost of $6.

The authors are thankful for valuable discussions to Yaroslav G. Klyushin and to Alexander V. Frolov.

HYPOTHESIS OF A THEORY OF EVERYTHING

Jack P. Gibson

www.crikcritter@infoave.net

This work grew out of an interest in curious occurrences, physical, mathematical and metaphysical. Many things just do not fit normal knowledge or are of unknown origin. Items such as gravity, time, mathematical equations of higher order, ghosts, e.s.p., etc. After a number of years, the problem began to appear to be not of this physical dimension but of a grander scale. A multi-dimensional space-time format seems to offer solutions to understanding most of the problems. In the following hypothesis all known physics and mathematics are considered to be valid. Metaphysical aspects, while not accounted for by physics and mathematics, are considered to be present in nature, either real or as a defect in physical processes, simply because they are experienced by so many people. The hypothesis will be kept simple because the details require the work of many specialists beyond my abilities and there is an endless list of speculative problems to which the hypothesis could be applied. To save time and distractions to the reader the following is stated as fact, knowing full well it may not be so, in order to present the most concise statement of the hypothesis.

The Universe, from the smallest to largest unit, is a quantum probability based fractal illusion. All matter exists in three-dimensional space, both internal and external. Any grouping can be considered a fractal space. It is both real and imaginary at the same time acting according to Eulers equation:

$$(e^{\wedge}(i^{*}Pi))+1=0$$

more useful in this instance is the general form:

$$e^{\wedge}ix = -1=\cos x + i \sin x = -\tan 45 \deg$$

which allows use of a variable where Pi , a constant, is used in the original. The key to understanding of multi-dimensional space time is the realization that Pi is a variable! As we know Pi is the lock on three-dimensional space. Because of its mathematical nature it can neither grow nor diminish, thereby locking in the three dimensional form of the fractal unit under consideration. Such a fractal unit is considered to be at rest. But Einstein's equation $E=mc^2$ includes speed or movement. When objects are moving relative to a fixed point, the moving object is foreshortened in the direction of movement. A circle becomes an ellipse and Pi becomes indeterminate letting the circumference fluctuate, according to its relative speed, between two and Pi in a circle of unit diameter. If one goes beyond the value Pi the Universe grows out of bounds without limit becoming unstable. If the value goes below two then the Universe blinks out. I.e., one has exceeded the value of the constant c. Since all matter moves relative to a fixed fractal point and there are infinitely many points in the Universe of constantly changing velocities, the Universe is constantly adjusting its reality to the relative speeds of its components. Thus you get at times some weird things. Since not only time changes but the structure and observable mass also, one finds such things a matter appearing and disappearing as the small units of chaos move through the Universe. Changes in the local probability structure cause matter or reality to change in the image structure of the Universe.

To delve further into the aspects of what is happening, let us consider the following. Einstein's familiar formula gives the relationships for our three-dimensional fractal space. The inverse of the formula gives the speed with which any fractal space can change, including the Universe. There is both a positive and negative component to the change. If we consider our fractal space to be positive then we may consider any other point to be negative. There are an infinite number of such points and while the effect on our space may be small from any one point; the effect adds up. Any moving or changing of mass must effect the entire Universe. Mass and inertia are tied together in such a system. Measured mass is the attempt of the mass to return to zero relative speed in its associated fractal space according to the laws of physics, i.e. return to its lowest energy state. Inertia is the same effect of the changing state of the mass from one-dimensional state to another. Both are the result of efforts to change the dimensional state of mass. The positive aspect of the square root of the ratio of energy/mass is the view we behold from our position in space while the negative is the view from the opposite position.

In addition to the view of the Universe as being made up as fractal space, all moving at different relative speeds to any other fractal space, one must consider the make up of space having at the same time dimensional space, according to Euler's formula given above. There are at any moment an infinite number of spatial dimensions, defined by the relative speed of the fractal units involved, extending throughout the Universe. Each fractal unit is constantly changing its position in dimensional space. There are thus constantly changing energy states throughout the Universe both in relative position in real space and in dimensional space. While on one hand the distances across the Universe are immense in relative space, many points across the Universe are present in the same dimension at the same moment, possibly allowing for instance travel across space without the time penalty. This results in what might be considered a tuned circuit for the Universe or its components.

If we consider the structure of the relative Universe to be stored energy analogous to magnetic fields and the dimensional fields to be capacitive, then we can plot the results on a conventional two-dimensional chart where x is the horizontal axis and the vertical is the imaginary axis. A well-known construction in electrical physics and which is known to very often follow relativity closely. That throws the negative side of i into the second quadrant and the positive into the fourth quadrant, if we consider the Universe to be reflective and x to be negative when i is positive. Euler's equation does not reflect a totally positive or negative result. One wonders at this point as to other equations where unity is set to one or higher dimensional equations where the three dimensional components are factored out leaving another component. The fourth is considered to be time but how do the remaining fit into a Universe as this hypothesis describes? The line of thought can be followed further into electrical analogy, which I leave to the reader. The important result is that there should be a resultant Q, figure of merit, of the Universe resulting in nodes, or peaks, of probability. The resultant tuned frequencies of the Universe can be calculated from known factors of the Universe. Such an effect should show nodes of reality where matter occurs.

Properly applied the above hypothesis gives speculative answers to most of the problems facing science today. Questions such as:

Is the Universe open or closed?

Both. At each fractal component of the Universe there is a separate universe each reaching limits beyond which it cannot exist. At the same time there is somewhere a top Universe that must ever remain open.

What is the structure of time?

Time results from the ever changing fractal Universe and goes forward because all justification of the Universe causes a corresponding change in reality. Time moves slower at higher relative speeds and as all time is perceived to be slower in other fractal units, changes in our time are perceived to always go forward.

What are such phenomena as ghosts, spirits, apparitions, etc?

Events such as these are fragments of probability left behind due to sudden changes in the probability of the illusion we perceive to be reality. Having little energy they are perceived but are not able to interact with the more forceful real reality. They float free in probability space interacting at random intervals with the more forceful realities of the real world.

Where is the missing matter of the Universe?

There may be clouds of matter floating about in deep space, however, most of the approximately 90% of matter missing will be found in the layers of dimensions making up space and which, though interacting, do not appear in our reality. Calculations show only about ten dimensions are active in any one reality. The rest curl up in a ball or knot. While an essential part of space or reality they are not perceivable from the fractal Universe of which they are a part.

Parapsychology?

Events of this type can be attributed to interconnections on the dimensional, or imaginary, level where there can be interactions between fractal units though not directly associated with each other in relative space. The manifestation varies or is temporary due to the constantly changing and adjusting due to interactions of the probabilities of the Universe. Due to the innate probabilities of each individual, one may be more susceptible or sensitive than other individuals.

Bible, prophesies, angels, demons, aliens, etc.

If one takes the Bible at face value, with some leeway for it's age and many translations, it pretty well describes what it purports to describe. Historic predictions and conditions of today. If we assume this hypothesis has any value then the passage in which God says "Let us make man in our image." takes on new meaning. If we assume the probability basis of the Universe and the constant changing due to justification of reality and time then the miracles listed do not seem so in violation of the physical world. The miracles only require some ability to control reality. Certainly well within the providence of a creator. It speaks of heavenly beings capable of traveling through the Universe in real time and some who interfered with mankind on the Earth and that they are still doing so today. If this hypothesis is correct then it is readily seen how such events could occur. Creation events closely compare with the way todays computers are made and organized. If man can do it, why could not the original creator? If one ignores a creator, then one is hard pressed for an answer as to how the Universe was created.

UFOs

If one accepts any part of the above then the functioning of UFOs begins to be understandable. They move through space by making spatial jumps through fractal nodes. Their appearing around magnetic and electrical sources maybe due to some, on our part unrealized, easy entry and exit to other space at those points. They are able to make right angle turns at high speeds because the speed is only from our viewpoint. From the viewpoint of the craft it is simply changing locations in space. The turning on edge may be only a different orientation of space at that point.

Particles "Out of the Vacuum"

As we shoot atomic matter at higher and higher speeds into nuclei we are able to fracture the construction of the building blocks of matter. The high speed fractal construction of probabilities of the unit used smashes into the target with enough speed to mix the probabilities resulting in new nodes from the total probabilities present. Various probabilities are ejected resulting in short lived particles in unstable nodes. At times fragments of probabilities appear out of nowhere as they form up to combine into larger nodes. Travel distance and speed observed may give a clue to the basic frequency of the Universe.

Black Holes

The midpoint of the tuned circuit of the relative structure and dimensional structure of the Universe. As matter spirals in, from the relative structure, to be torn apart into its basic probabilities it passes through to the dimensional structure of space which holds 90% of the matter. There it is available for recollection and use by the relative space.

Probability. Basic structure of the energy of space.

Everything operates at random but within preferred patterns set by the frequencies involved. If you look long enough anything can be found. As a result we find odd bones of prehistoric creatures that may have never existed. Weird things can happen. Odd pieces of structure can occur such as metaphysical events. The Q of space is sometimes sharp and sometimes rounded, spreading out to include things not really meant to be in the overall scheme of the Universe. Christ said “If you had the faith of a mustard seed, you could say to the mountain move and the mountain would move”. (For those not of the Christian faith, this may be considered not as any proof but as a representative statement of the meaning of this paragraph.) That is in keeping with the quantum statements that nothing exists until it is seen and that we have some control over the probabilities. If the structure of space is based on probability then the next important question is ‘What is a probability’. We look to probability as a simple mathematical work. But, what makes probability work? The entire Universe is based on mathematics but we have no idea what is underneath the mathematics. Are dimensional planes perhaps better represented by other number bases? Are some of the unsolvable mathematical problems solvable in other number bases?

Gravity

All forces in the Universe are the same. Gravity happens to be the one that works on our fractal space. Others, molecular, atomic, and nuclear are the same but work in different fractal space. All are the result of matter trying to reach its lowest potential.

The above are only random questions chosen without any order. Most other problems can in some way be answered in line with the hypothesis, if not in detail. Too many questions can be fitted to the hypothesis not to take a serious look at it even though it may seem a little far out. The number of answers from one simple statement of the structure of the Universe defends the idea better than details.

One possible proof, and a relatively simple one, occurs to mind. Einstein’s theory of relativity was proved when NASA flew a clock in space and then determined that time did slow down. To test the above hypothesis I would like to see four clocks used in the following manner.

Four atomic clocks in sets of two each. Two to be left on the Earth and two to be placed in space for an appropriate time. At the end of that time one of the clocks on the Earth to be taken into space and compared to the two already in space. Bring one of the clocks in space down to the Earth and compare to the clock left on the Earth.

The clock taken to space should read slower compared to the clocks in space because the ones in space were at rest compared to the one on the Earth.

The clock brought back from space should read slower, the same as the first one flown by NASA, because the clock on the Earth was at rest compared with the one in space.

The two clocks left in space should be returned after an appropriate time and compared to the ones on the Earth. The original in space should now be slower than the original on the Earth and its mate from space. The fourth clock is unpredictable.

The author apologizes for any errors in theory, mathematics, etc. There was no one to consult on the hypothesis which is an original work of my own and drawn on many references of others from the past If it has any value I am indebted to the work of all the others whose work I relied on. Any errors are entirely my own and not attributable to others. This work is freely published in the public domain to be used by any and all who wish to do so. It is not to be copyrighted or patented in any manner so as to restrict others rights to the hypothesis or it’s use.

• •

JACK P. GIBSON

About the Author:

EDUCATION: University of North Carolina at Raleigh (Nuclear & Electrical Engineering), Air University, Gunter Air Force Base (Radio Fundamentals, Motorized Equipment) Western Electric Co Engineering Center

The Charge and Mass of a Photon

Dzabrail Kh. Baziev

Sirenevy blv., 65 building 4, apt.55
105484, Moscow, Russia
(095) 461-9398, (866-22)-5-10-11

Abstract

The discovery of true elementary particle named **electrino**, which has a constant mass and constant positive charge and which is a material carrier of magnetic field, electric current and all kinds of emanation, was firstly practically applied at the modernization of wireless telephone. In the review V. Anpilogov writes: "The question on the influence of low intensity microwave radiation on human health still remains open for discussion for more than 50 years" [3]. Discussion on this question has been already inappropriate still it was definitely proved in the patent application "Device, which is made to put away the charged particles flux from the head of the user of mobile wireless phone". During the talk on mobile wirelessphone $2{,}6\cdot10^{22}$ particle/sec pass through user's brain, whereas their energy is $7{,}65\cdot10^{-19}$ Joule, and the energy of binding of molecules in protein polymers ranges at $4{,}3\text{-}7{,}6\cdot10^{-19}$ Joule according to L. Poling. From that we can conclude that microwave radiation of the radiotelephone with top antenna leads to the destruction of user's brain tissue. The author of the patent application D. Baziev (#2001105456, 28.02.01. in Russia and #PCT/RU02/00054 of the international application) offers to fix the antenna on the low part of the radiotelephone and to produce radiotelephones with minimal length within 10 cm. In this case the diagram of the directional radiation pattern is on the level of the user's chin and the brain is out of the direct exposure area. This harmless of the microwave radiation could not be proved before the discovery of "electrino".

Introduction

A systems analysis of all experimental and observation material gathered in physics, astronomy, and astrophysics from the times of Galileo let us reveal the following:

1. Experimental materials do not agree with the existing physical theory.
2. There is a certain fundamental disadvantage of the experimental material, which prevents to build a consistent theory.
3. This fundamental disadvantage consists in the absence of a charge antipode of electron, which is in the form of a true elementary particle with a positive charge and finite mass.
4. The proton and positron are not true elementary particles and neither of them can be a charge antipode of the electron because they are subjected to splitting.
5. Discovery of the second true elementary particle with a positive charge could restore the charge symmetry in physics, thus leading to a radical revision of the existing theoretical physics and resolving its current crisis state.

Searching for this particle required to ascertain physical nature of Planck's constant. This became possible only after the structure of a light beam had been understood. Namely, it was the photon sector velocity, known as Millikan constant μ, rather than the speed of light *c*, that proved to be a constant, viz.:

$$\mu = \lambda_i^2 \nu_i = 119.916984\ \text{m}^2/\text{s} = \text{constant}, \qquad (1)$$

where λ_i and ν_i are the wavelength and frequency of the *ion* monochromatic beam in the light beam.

This new quantity elucidated the physical nature of Planck's constant:

$$h = m_\varepsilon \mu \frac{\sqrt[3]{4\pi/3}}{2} = \\ = 6.6262681\cdot10^{-34}\ \text{kg m}^2/\text{s} = \text{constant}, \qquad (2)$$

where m_ε is the mass of the second (after electron) true elementary particle to be called "electrino". From this expression we have

$$m_\varepsilon = \frac{2h}{\mu\sqrt[3]{4\pi/3}} = \\ = 6.85575729963\cdot10^{-36}\ \text{kg} = \text{constant}. \qquad (3)$$

The electrino has a positive charge ε determined by

$$\varepsilon = \frac{m_\varepsilon n_\varepsilon e}{m_u - n_e m_e} = \frac{-3.229526609098\cdot10^{-54}}{1.6578584539\cdot10^{-27}} = \\ = 1.98764431671\cdot10^{-27}\ \text{C}, \qquad (4)$$

where $m_u = 1.66057\times10^{-27}$ kg is the mass of an elementary atom accepted as a mass equivalent of one atomic unit; $n_e = 3$ is the number of electrons in one elementary atom; $e = 1.6021892\times10^{-19}$ C is the charge of an electron; $m_e = 9.038487\times10^{-31}$ kg is an improved value of electron mass; $n_\varepsilon = 2.418198867\times10^{8}$ is the number of electrinos in an elementary atom.

Thus, it is obvious that Planck's constant is the angular momentum of the electrino. Moreover, it was Planck's constant that concealed the second true elementary particle, which is the charge antipode of the electron discovered by J.J. Thompson as far back as in 1897.

The solution of Planck's constant has become a basis for the synthesis of the new theory of physics [1]. This theory in particular shows that the electrino is the carrier of the magnetic field and electrical current. It is a photon of radiation of all ranges, and serves as a universal carrier of energy and information. The electrino plays the role of a neutrino in moving along the first order trajectories.

The first experiment

An extraordinary importance and novelty of the new theory required **an experimental proof of the electrino**. For that several experiments were made in the Institute of General and Inorganic Chemistry, Moscow. The experiment was based on the following effects predicted by the theory.

1. If assume that electrino exists and that light beam is a flux of particles having positive charge and finite mass [1] we can conclude the following. At the discharge of the dc source through an incandescent lamp in which the current is converted to light and irreversibly emitted, the source weight in charged state must differ from its weight in discharged state. If we prove this difference experimentally we may say that light does consist of material particles of finite mass and a dc charge carried away by light is positive because an incandescent lamp (W=15 Wtt) does not emit electrons, which are the carriers of negative charge.
2. The second effect to prove was that the weight of a discharging dc source is *increasing* whereas its weight when charged is *decreasing*.

To prove the validity of these predictions, several sealed containers with different dc sources inside were fabricated. The electrodes were brought out through glass insulators. The batteries were discharged through an electric lamp radiating in the visual and infrared ranges. The weight of containers was measured before and after discharge process with accuracy $\Delta W = \pm 0{,}02$ mg; balance error was equal to $\Delta = \pm\ 0{,}05$ mg the standard deviation of the measurements was within $\sigma = \pm 0{,}03$ mg; the buoyancy was calculated for each measurement of weight. In this paper, we present test results of only one container with four generally marketed GP rechargeable cells connected in series. The total battery voltage reached 5400 mV at 6000 mA/h charge capacity. The discharge was interrupted when the voltage dropped to 4000 mV, the duration of the discharge was measured accurate to one second. Two series of experiments were run: one in air, the other, under argon. Each series had ten charge-discharge cycles (Table 1 and Fig. 1). The total amount of the experiments and detailed discussion of results have been summarized in a recently published brochure [2].

The results of the above tests allow us to make the following conclusions:

1. Both galvanic and rechargeable cells during a discharge through an electric lamp show sufficient changes in their weight and charge thus proving that photons have a finite mass and a positive electric charge.
2. A new elementary particle, named electrino, derived from Planck's constant in August 1982, and published in May 1994, thus gets a complete and absolute experimental confirmation.

The second experiment

One of the concepts of the new theory is that the speed of light in vacuum is a function of photon frequency along the beam axis, according to the proportions:

$$c_i = \sqrt{\mu\nu_i} \quad [\text{m/s}], \tag{5}$$

$$\nu_i = \mu / \lambda_i^2 \quad [\text{s}^{-1}]. \tag{6}$$

According to the new theory, for the velocity of monochromatic light (solar light or mercury-discharge lamp, but not a laser) with a wavelength of λ_r=6.8×10^{-7} m (mid-point of the red spectral line), we have

$$\nu_\text{r} = \mu / \lambda_\text{r}^2 = 2.59336038 \cdot 10^{14}\ \text{s}^{-1}, \tag{7}$$

$$c_\text{r} = \sqrt{\mu\nu_\text{r}} = 1.76348505882 \cdot 10^{8}\ \text{m/s}, \tag{8}$$

which is 58.823% of the speed c_v = 2.9979246×10^8 m/s of a violet beam with a wavelength of 4 × 10^{-7} m.

We have to account that, according to this theory, the laser beam is not a true light beam though it is created of electrinos. The speed of laser beam is equal to the speed of beam plus the speed of current in the conductor, viz.,

$$\nu_0 = 2.8992629 \cdot 10^8\,\text{m/s} = \text{const} \tag{9}$$

If we select a monochromatic beam of ultra-violet light with a wavelength of λ_1 =4 · 10 $^{-8}$ m then its velocity will be c_1=10 C:

$$c_1 = \mu / \lambda_1 = \frac{119{,}916984 m^2/\sec}{4 \cdot 10^{-8} m} =$$
$$= 2{,}9979246 \cdot 10^9\, m/\sec$$

$$\nu_1 = \mu / \lambda_1^2 = \mu / 16 \cdot 10^{-16} m^2 = 7{,}4948115 \cdot 10^{16} \sec^{-1}$$

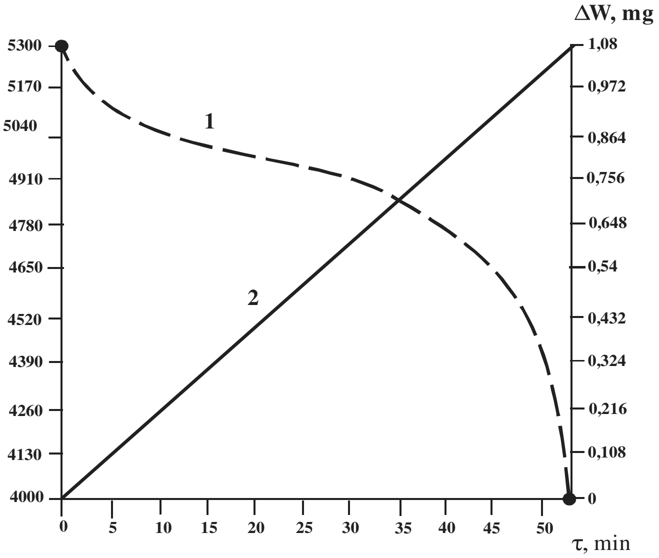

Fig. 1
Voltage drop [mV] of a battery and (2) weight increment [mg] of container #6 in an argon atmosphere during the second discharge cycle [minutes]. X-direction – is spark duration in minutes; Y-direction – is voltage of a battery; auxiliary Y-direction – is weight incensement (mg).

$$c_1 = \sqrt{\mu \cdot v_1} = \sqrt{8{,}987551907 \cdot 10^{18} m^2 / c^2} = 2{,}9979246 \cdot 10^9 m / \sec$$

This experiment has been run yet, and it is offered for experimental verification with further publication of experimental results.

Table 1

Weight of container #6 (under argon) in charge-discharge experiments

		Charged battery		Discharged battery			Charge
Run	Measured value $W \pm \sigma$, mg	Buoyancy Gmg	Real weight $W_0 = W + G$	Measured value $W \pm \sigma$, mg	Buoyancy Gmg	Real weight $W_1 = W + G$	weight $\Delta W_z = W_1 - W_0$, mg
1	126825.13±0,02	85.031	126910.166 ±.0.02	126825.901 ±0.01	85.002	12691.903 ±0.01	0.737
2	126825.107±0.02	86.572	126911.679 ±0.02	126826.221 ±0.01	86.538	126912.759 ±0.01	1.080
3	126825.21±0.01	86.782	126911.992 ±0.01	126826.279 ±0.01	86.560	126912.839 ±0.01	0.847
4	126825.187±0.01	86.563	126911.749 ±0.01	126826.493 ±0.02	86.385	126912.878 ±0.02	1.128
5	126825. 65±0,04	86.290	126911.941 ±0.04	126826.65 ±0.01	85.836	126912.941 ±0.04	0.770
6	126827.28±0.00	85.187	126912.467 ±0.00	126827.990 ±0.01	85.204	126913.194 ±0.01	0.727
7	---	---	---	---	---	---	---
8	126826.98±0.00	86.182	126913.162 ±0.00	12682 7.897 ±0.02	86.308	126914.205 ±0.02	1.042
9	126826.95±0.00	86.307	126913.257 ±0.00	126827.757 ±0.01	86.402	126914.159 ±0.01	0.902
10	126827.25±0.00	86.294	126913.544 ±0.00	126828.35 ±0.00	85.729	126914.079 ±0.00	0.535

References

1. D.K. Basiev, Osnovy obyedinyonnoy teorii fiziki (Foundations of Unified Theory of Physics), Pedagogika, Moscow, 1994
2. D.K. Basiev, Zaryad i massa fotona (The charge and mass of a photon), Pedagogika, Moscow, 2001
3. V.P. Anpilogov, "The century of quality", #3, 2001, p.60

About the Author

When a child Dzabrail Kh. Baziev became a political exile along with all Balkarian people; in 1956 he was rehabilitated and in 1957 graduated from the secondary school with steady purpose to become an outstanding chemist. His first higher education is biological and chemical (two-profile department of Kabardino-Balkarian State University). In 1965 Baziev also completed his post-graduate study in Biological Department of Moscow State University.

At 1979 Dz.Kh. Baziev had published 25 scientific articles on ornithology and ecology, however, this period became a sudden turn in his scientific work. In order to elaborate the basis of theoretical biology, he came to the conclusion that biological process could not be described without understanding of its physical essence. On the other hand, it had become evident that there was no any solution for physical phenomenon, which is the basis of biological process.

In 1983 the scientist gave descriptions of his first fundamental discoveries. He found the solution of the physical essence of Plank's constant and pioneered the use of it in gases analysis. It was a revolutionary step in science since the synthesis of science was realized at the fundamental level for the first time.

As the result of this considerable work there was a new interdisciplinary theory of physics. The author created new thermodynamics of real gases and new electrodynamics, which differs from maxwellian one and it considers *electrino* as the carrier of magnetic field and electric current.

Dzabrail Kh. Baziev has succeeded in the systematic analysis of the vast experimental material and discovered that ***electrino*****, as the true elementary particle with positive charge, is able to provide the connection of physics with biology, chemistry and other parts of the fundamental science. Thus it leads us to the radical reconsideration of all conceptions, existing in natural science.**

The Homopolar Motor: A True Relativistic Engine

Jorge Guala-Valverde

Norpatagonica-R&D Dept. S.Fe 449 Neuquen, Argentina
Q8300BG1, Confluencia Tech University Neuquen, Argentina
gual@ieee.org

Pedro Mazzoni

Fundación Julio Palacios Neuquen, Argentina
fundacionjuliopalacios@usa.net

Ricardo Achilles

Confluencia Tech University Neuquen,Argentina,
RA Biosystems Neuquen, Argentina
achilles@ieee.org

This article discusses experiments, which enable the identification of the seat of mechanical forces in homopolar-machines. Authors provide a suitable variation on a recent work "The Unipolar Dynamotor: A Genuine Relational Engine" [3], where "relational" implies "absolutely relativistic". The authors' view agrees with both Weber's recognition in the 19th century of the importance of relative motion in electromagnetic phenomena [4] and Einstein's 1905 statement concerning electromagnetism [5].

The Faraday disk: a reversible engine

The essential components of the homopolar machine, first conceived by Faraday in 1832, are shown in Figure 1. A conducting disk, free to rotate in the neighborhood of a permanent magnet, is attached to the end of a shaft. A closing wire provides a conducting path between two arbitrary points of the disk. Such a device exhibits *reversible* behavior.

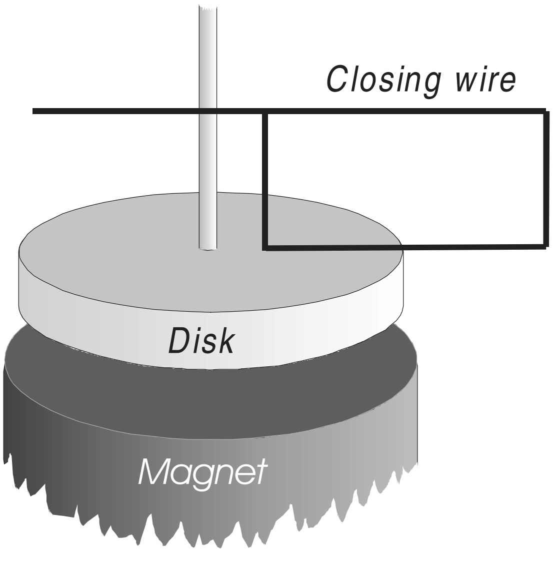

Fig.1
Faraday's setup magnet, disk and closing wire

A radial current path of length L takes place in a region of the disk when direct current (*dc*) from an external source is injected into the closing wire. The interaction of the current with the magnetic field produces a Laplace force [6]

$\boldsymbol{F} = \int_a^{a+L} I(d\boldsymbol{r}x\boldsymbol{B})$ causing the rotation of the disk. This set-up is the *motor configuration.*

When the disk is spun by an external source of mechanical energy, an *emf* appears in it. The displacement of free charges is produced in this case by the Lorentz force $\boldsymbol{f} = q(\boldsymbol{v}x\boldsymbol{B})$, converting the conducting disk into an *emf* source able to drive *dc* through the whole disk plus closing-wire circuit. This set-up is the *generator configuration*.

A seemingly curious fact occurs in the motor configuration, when *dc* is injected into the circuit with the disk attached to the magnet. Both disk and magnet turn together.

Two rival theories, a relativistic and an absolutistic one, have been applied to understand the observed facts:

In relativistic view, generator configuration makes sense only when there is relative motion of the magnet with respect to either the disk or the closing wire. Also, a motor configuration will only take place if the *possibility of relative motion* between magnet and either disk or closing wire is enabled.

Thus, in the relativistic framework, with the magnet attached to the disk, the closing wire becomes the "active" part for the production of mechanical forces or *emf*. In this case the disk itself behaves as a "passive" element providing a closing-circuit current path.

Conversely, in the eyes of an absolutist, a generator configuration is enabled only because of the disk or closing-wire *absolute* motion. Here, absolute means "relative to a frame where the preponderance of the mass of the universe is at rest" [7,8]. In our case, the lab frame acts as an acceptable absolute-motion reference. Thus, from an absolutistic view, the magnet's rotation with $\partial B/t = 0$ in each point of the surrounding space is unable to produce an *emf* on nearby conductors. When in a motor configuration, *dc* is injected in the circuit, and the absolutist assigns the observed rotation to the magnet "dragging" by the conductor. Here, the closing wire acts as a "passive" circuit element.

New experimental work, complementary to that currently known on the subject, introduces arguments in favour of the relativistic viewpoint. The related experiments, whose underlying physics rests upon a modified version of the original Faraday setup, are described in the following sections.

The asymmetrical rotor

Figure 2 shows the disk-shaped ceramic permanent magnet creating the axial magnetic field ***B***. The removal of a 12° sector introduces a field-reversion region. Outgoing and ingoing ***B*** field lines are represented by the and symbols, respectively.

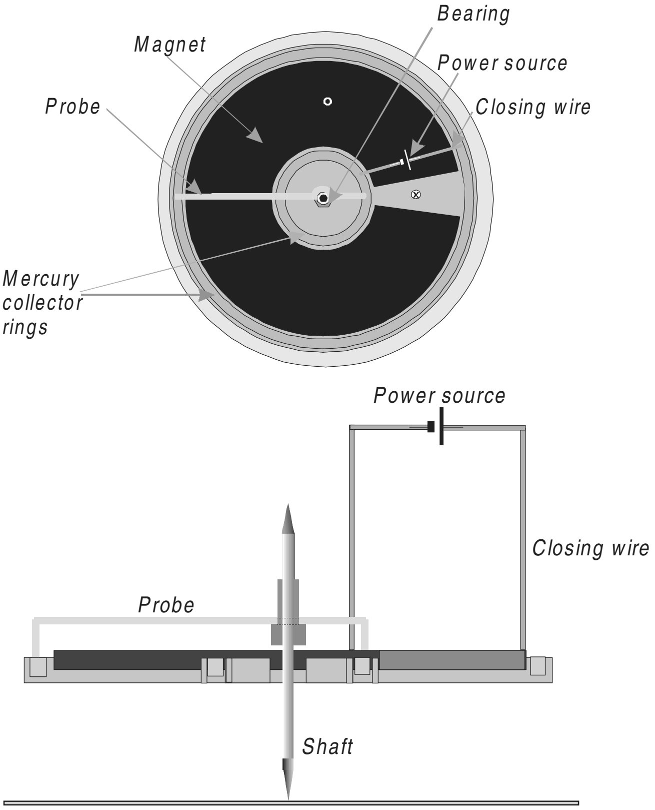

Fig.2
Layout of the Asymmetrical Rotor applied to the experiments

Two mercury collector rings are embedded in a wood cylinder. One is located close to the hollow-disk magnet inner rim and the other in the proximity of the outer rim. The magnet's inner and outer radii are 25 and 75 mm, respectively, and its height 25 mm. Its average flux density 2 mm above the magnet has been estimated to be 0.05 T based on a generator experiment with a rotating copper disk. The magnet-and-wood-cylinder body (the asymmetrical rotor from here on) is firmly anchored to a vertical shaft terminated in sharp points at both ends. While the lower one lays on a hard-polished surface, a conical bearing, enabling its almost frictionless rotation, centers the upper one.

Unlike the series-connected conductors diametrically anchored to the shaft in the Guala-Valverde case [3], only one radial conductor wire, a probe located 2 mm above the magnet's face, was considered. By mounting it on a bearing, its free rotation is permitted with its ends remaining in contact with both collector rings. A 12V lead-acid battery applied to the closing wire feeds the probe through the collector rings. In the first four experimental cases presented the closing wire remains firmly anchored to the lab. In two complementary experiments, rotation of the closing wire mounted on two shaft-centered bearings is allowed. Its behavior as a probe occurs by the injection of *dc* from an additional closing-circuit wire anchored to the lab.

Experimental

Six experiments performed are described below:

1. Rotor anchored to the lab, probe free to rotate above the magnet's upward magnetic-field region: A radially-ingoing injected *dc* in the 0.2 A range was enough to overcome conductor-bearing and mercury-wire contact friction. A net *counterclockwise rotation* of the probe took place.

2. Probe anchored to the rotor above the magnet's upward magnetic-field region, with both free to rotate: A radially-ingoing injected *dc* in the 5 A range was enough to overcome conductor-plus-rotor inertia and friction. A net *counterclockwise rotation* of the probe took place.

3. Rotor anchored to the lab, probe free to rotate above the magnet's downward magnetic-field region: A radially-ingoing injected *dc* in the 0.2 A range was enough to overcome conductor-bearing and mercury-wire contacts friction. A net *clockwise rotation* of the probe took place.

4. Probe anchored to the rotor above the magnet's downward magnetic-field region, both free to rotate: A radially-ingoing injected *dc* in the 5 A range was enough to overcome conductor-plus-rotor inertia and friction. A net *counterclockwise rotation* of the probe took place.

5. Rotor anchored to the lab, closing wire free to rotate above the magnet's upward magnetic-field region: A 0.4 A *dc* injected in the inner collector ring was enough to overcome conductor-bearing and mercury-wire contacts friction. A net *clockwise rotation* of the closing-wire took place.

6. Rotor anchored to the lab, closing wire free to rotate above the magnet's downward magnetic-field region: A 0.4 A *dc* injected in the inner collector ring was enough to overcome conductor-bearing and mercury-wire contacts friction. A net *clockwise rotation* of the closing-wire took place.

Discussion of results

Experiments (1) and (3) can be explained using either absolutist or relativistic viewpoints because of the coincidence of the probe motion relative to the lab with the probe motion relative to the magnet.

Experiment (2) can be explained by a trivial absolutist argument founded on a hypothetic probe "dragging effect" on the magnet. A relativistic viewpoint recognizes the "active" rotational torque on the closing wire rather than on the probe where, hinging on Newton's third law, the whole action may be split in two:

Magnet-probe. The magnet produces a counterclockwise torque on the probe, and the probe exerts an equal but opposite torque on the magnet.

Magnet-closing wire. The magnet exerts a clockwise torque on the closing wire, and the wire an equal but opposite torque on the magnet.

With the probe attached to the magnet, there is no chance for relative motion between them. Consequently, due to the action-reaction cancellation, rotation is forbidden. Conversely, with the closing wire mechanically decoupled from the magnet, relative motion of the latter is permitted. The torque exerted by the closing wire on the magnet is responsible for the observed rotation.

Experiment (4): Due to its similarity with (2) a trivial relativistic explanation is applicable to the counterclockwise torque exerted by the closing wire on the magnet. There is no known plausible absolutistic explanation for it. As quoted above, the hypothetical dragging effect would produce a clockwise rotation in this case. The consideration of the experiments (2) and (4) suffices to reject the dragging hypothesis.

Complementary experiments (5) and (6) confirm the short-range extension of the field-reversion region founded on the closing-wire *clockwise* rotation (6). Briefly speaking, the closing wire is not sensitive to the field reversion and the magnet's *counterclockwise* reaction explains at once the outcome of (4). Clearly, experiments (5) and (6) show that the torque on the closing wire is independent of its location on the magnet.

Figure 3 depicts the two rotational torques involved in (2) and (4).

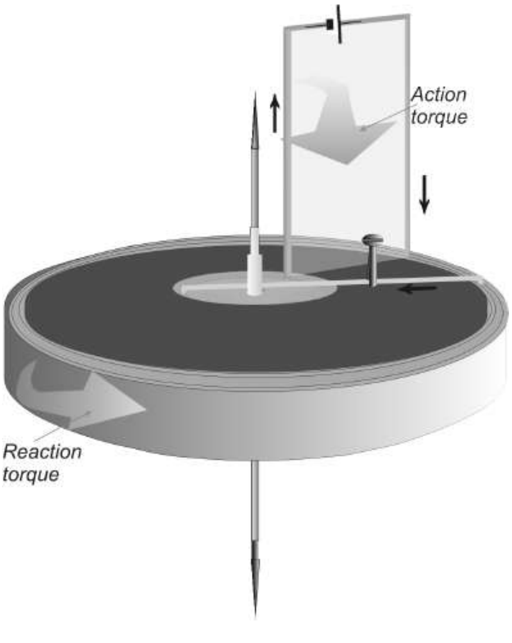

Fig.3
Rotational torques acting on the magnet and on the closing wire

Topological and miscellaneous considerations

One of the keys to the success of the above described experiments lies in the dynamotor's magnet design (see Fig.4). The short-range field reversion region allows the inversion of the Laplace force on the probe, making the force on the closing wire insensitive to that ***B**-field* reversion.

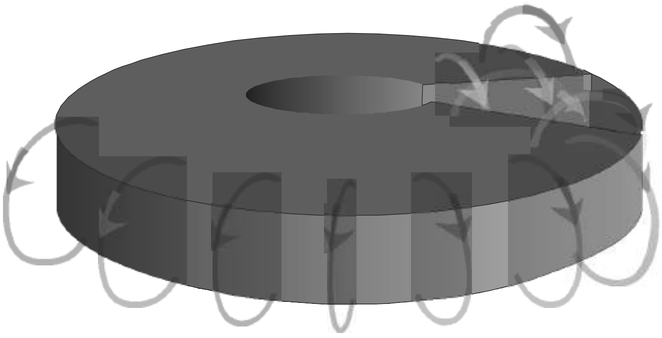

Fig. 4
The magnet's field-revision region

In all the above cases the electromagnetic forces between probe and closing wire were neglected because of its small magnitude compared to the predominant magnet-wire interaction forces.

The observed torques became, in all the experiments, independent of the location of the contact points between closing wire and collector rings. Also, the closing wire shape exhibited no noticeable influence on torques. These observations can be easily explained using the $div\mathbf{B} = 0$ fundamental law, Laplace force, and some elementary topological considerations.

Kennard [1], Bartlett [1]· Panosky [7,8], Muller [9], Wesley [10] and some of this article's authors took absolutistic viewpoints when dealing with homopolar phenomena [11,12]. On the contrary, Weber [4], Assis [13], and Kelly [14] adopted a relativistic framework on the issue from the beginning.

By attaching the magnet to the disk in the original Faraday setup, the relative rotation between disk and closing wire remains unchanged. Therefore, in a generator configuration, the disk plus magnet rotation at with the closing wire at rest in the lab is entirely equivalent to the closing-wire rotation at – with the disk plus magnet at rest. This fact introduced a correct but physically "colorless" weak relativism to the homopolar generator description: the "unipolar generator really has three components, the magnet, the cylinder and the meter (including the contacts). A relative motion of the last two, not the first two, is required" [1].

A growing interest in basic electromagnetism [15,27] can not be ignored, and from time to time some authors, attempting to catch "free energy" from the space, have claimed the design of homopolar engines with efficiency greater than unity, as can be checked by searching for *homopolar motor* on the Internet. The strict application of Newton's third law precludes the above non-physical possibility.

It is worthwhile to stress that the homopolar machine is a famous example where Faraday's flux rule fails. This fact worried Faraday himself and is clearly discussed by Feynman [28] who emphasized that the correct physics is always given by the Lorentz force law and the Maxwell fundamental equation $curl\ E = -B/t$. Homopolar induction is fully understood using only the Lorentz force. Our experiments enhance the relativistic structure of the Lorentz force because the only relevant velocity is the velocity of the conductor relative to the magnet.

Acknowledgments: To Profs. C.N. Gagliardo and A. Ipohorski-Lenkiewicz for the conceptual comments on this development.

References

1. J. Guala-Valverde and P. Mazzoni, "The principle of relativity as applied to motional electromagnetic induction". Am. J. Phys 63, 228-229 (1995).

2. J. Guala - Valverde, P. Mazzoni and Ro Blas, The quasilinear generator and = 0". Am. J. Phys. 65, 147-148 (1997).

3. J. Guala-Valverde and P. Mazzoni, "The unipolar dynamotor: a genuine relational engine". Apeiron 8, 41-52 (2001).

4. A.K.T. Assis, Weber Electrodynamics (Kluwer, Dordrecht, 1994).

5. A. Einstein, "Zur elektrodynamik bewegter körper". Annalen der Physik 17, 891-921 (1905). See also The Principle of Relativity. A.Einstein, H.A. Lorentz, H. Weil and H. Minkowsky, pp. 35-65 (Dover, New York, 1952).

6. G. Cavalleri et al. "Experimental proof of standard electrodynamics by measuring the self-force on a part of a current loop". Phys. Rev. E 58, 2505-2517 (1998).

7. W. K. H. Panofsky. Private communication (letter to J.Guala-Valverde, dated Jan 30, 1995).

8. W. K. H. Panofsky and M. Phillips, Classical Electricity and Magnetism. (Addisson Wesley, NY, 1955).

9. F. Müller, "Unipolar induction experiments and relativistic electrodynamics". Galilean Electrodynamics 1, N°3, 27-31 (1990).

10. J.P. Wesley. Private communication (letter to J.Guala Valverde dated Nov 01, 2001). See also "Progress in Space-Time Physics", Benjamin Wesley, Blumberg (1987).

11. J.Guala-Valverde and P. Mazzoni, "The motional electromagnetic induction revisited". Russian Academy of Sciences, 3rd Conference, pp.186-188. St. Petersburg (1995).

12. J. Guala-Valverde and P. Mazzoni, "The unipolar Faraday generator revisited". Galilean Electrodynamics, 4, 107-108 (1993).

13. A.K.T. Assis and D.S. Thober, "Unipolar induction and Weber's electrodynamics". Frontiers of Fundamental Physics, Plenum Press, NY, pp. 409-414 (1994).

14. A.G. Kelly, "Experiments on unipolar induction". Physics Essays 12, 372-382 (1999).

15. A.E. Robson and J.D. Sethian, "Railgun recoil, Ampere tension, and the laws of electrodynamics". Am. J. Phys. 60, 1111-1117 (1992).

16. A.K.T. Assis, Relational Mechanics (Apeiron, Montreal, 1999).

17. P. Graneau, "Electromagnetic jet-propulsion in the direction of current flow". Nature 295, 311-312 (1992).

18. T. E. Phipps and T.E. Phipps Jr., "Observation of Ampere forces in mercury". Phys. Lett. A 146, 6-14 (1990).

19. P.Graneau and N. Graneau, Newtonian Electrodynamics (World Scientific, Singapore, 1996).

20. V.F. Mikhailov, "The action of an electrostatic potential on the electron mass". Annales de la Fondation Louis de Broglie 24, 161-169 (1999); 26, 33-38 (2001).

21. A.K.T. Assis, "Comment on experimental proof of standard electrodynamics by measuring the self-force on a part of a current loop". Phys. Rev. E 62, 7544 (2000).

22. P. Graneau and N. Graneau, "Electrodynamic force law controversy". Phys. Rev.E 63, 058601-2 (2001).

23. T.E.Phipps and J. Guala-Valverde, "Simulation of amperian current elements by magnetic toroids". 21st Century Science & Technology 11, 55-65 (1998).

24. M. Bueno and A.K.T. Assis, Inductance and Force Calculations in Electrical Circuits (Nova Science Publishers, Huntington, N.Y. 2001).

25. H. Aspden, "Anomalous electrodynamic explosions in liquids". IEEE Trans. on Plasma. Sc. 14, 282-285 (1986).

26. G.M. Graham & D.G. Lahoz, "Observation of static electromagnetic angular momentum in vacuo". Nature 285, 154-155.

27. V. Dvoeglazov, "Essay on the nonmaxwellian theories of electromagne-Tism". Hadronic Journal Suppl. 12, 241-245 (1997).

28. R. Feynman, The Feynman Lectures on Physics, II. (Addisson Wesley,NY, 1964).

Weight Reductions Generated by Bucking-Field Permanent Magnets

LAB REPORT ON SmCo RING MAGNET EXPERIMENTS

Experiments conducted by:
William C. Simpson

New Horizons Research 600 Meridian Street Extension,
#302 Groton, CT 06340
Tel. (860) 405-1157

The following experiments were conducted at the *Coastal Environmental Laboratory* (CEL) at Avery Point, Groton, Connecticut. (41° 19‘ 0.17" N. latitude x 72° 3‘ 50.27" W. longitude x 35 feet elevation above mean sea level) I wish to thank the personnelat the CEL for their generosity for providing the use of their *Mettler Toledo*® Model AG104 electronic scale for the measurements taken in the proceeding experiments. The AG104 electronic scale is an enclosed pan unit with a maximum mass range of 101-grams with 0.0001-gram readability.

The purpose of these experiments was to see if there is any detectable weight change when permanent magnets are forced together with their like-poles facing each other. The magnets were weighed individually, in both directions, with their field poles oriented vertically. The sums of the two individual magnet weights (magnet #1 and magnet #2) in each vertical orientation were compared to the weight measurements taken when they were assembled using the nylon bolt and wing nut depicted in **DIAGRAM 1**. The specifications for the two Samarium Cobalt magnets used in the following experiments are shown in **DIAGRAM 1**.

The first set of experiments with the SmCo Ring magnets were conducted January 14, 2002. An inverted paper cup was used to raise the test sample magnets 2.75" above the AG104 electronic scale pan in order to minimize possible magnetic interaction with the scale-sensing element, as depicted in **DIAGRAM 2**. The tare adjustment was used to set the scale readout to 0.0000-gram with the cup in place. The magnets were weighed individually. Magnet #1 weighed 9.9450-gram with the N pole facing up and 9.9397-gram with the S pole facing up. Magnet #2 weighed 9.9520-gram with the N pole facing up and 9.9443-gram with the S pole facing up.

The second set of experiments with the SmCo Ring magnets were conducted February 4, 2002. These experiments were shielded with Mu 80 magnetic

Editorial: This article is one more link between magnetizm and theory of aether, to my mind. It is possible to assume that in his experiments the author creates small but detectable changes in density of aether, that demonstrates itself as the weight changes.
Alexander V. Frolov.

DIAGRAM 1

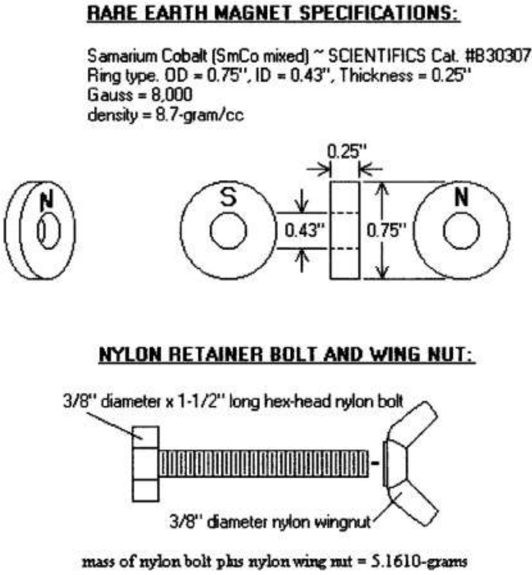
RARE EARTH MAGNET SPECIFICATIONS:
Samarium Cobalt (SmCo mixed) ~ SCIENTIFICS Cat. #B30307-30
Ring type. OD = 0.75", ID = 0.43", Thickness = 0.25"
Gauss = 8,000
density = 8.7-gram/cc
0.25"
N
S
0.43"
0.75"
N
NYLON RETAINER BOLT AND WING NUT:
3/8" diameter x 1-1/2" long hex-head nylon bolt
3/8" diameter nylon wingnut
mass of nylon bolt plus nylon wing nut = 5.1610-grams

shielding material as depicted in **DIAGRAM 3**. The scale was tare adjusted to 0.0000-gram with the entire set of Mu 80 shield pieces in place. Then the magnets were weighed individually. Magnet #1 weighed 9.9483-gram with the N pole facing up and 9.9486-gram with the S pole facing up. Magnet #2 weighed 9.9527-gram with the N pole facing up and 9.9542-gram with the S pole facing up. The Nylon bolt and wing nut were placed in the Mu 80 shield can (without the magnets) and the scale was tare adjusted to 0.0000-gram. Therefore, the readouts would only be reading the weight of the bucking magnets.

The first column in **TABLE 1**, the vertical measurements, is the distance of separation d, or air gap, of the magnets. The second column shows the January 14, 2002 weight measurements of the two magnets, as shown in **DIAGRAM A**. The third column shows the January 14, 2002 weight measurements of the two magnets, as shown in **DIAGRAM B**. The fourth column shows the February 4, 2002 weight measurements of the two magnets, as shown in **DIAGRAM C**. The fifth column shows the February 4, 2002 weight measurements of the two magnets, as shown in **DIAGRAM D**.

The horizontal measurement, as depicted in **TABLE 2**, **TABLE 3**, **TABLE 4** and **TABLE 5**, are through the four basic magnetic compass headings; North, East, South, and West respectively. They are referenced to **DIAGRAM E** and **DIAGRAM F** for the January 14, 2002 experiments and **DIAGRAM G** and **DIAGRAM H** for the February 4, 2002 experiments. The corresponding graphs of the force change plots, **GRAPH 1**, **GRAPH 2**, **GRAPH 3**, **GRAPH 4**, and **GRAPH 5** accompany each table. The forces were converted from the mass readings, which are a scalar measurement, to dynes. The convention used for the force vector was chosen as plus (+) for up, or a weight reduction, and minus (-) for a weight increase.

DIAGRAM 2

01/14/2002 EXPERIMENNTS

CUP HEIGHT: 2.75" above scale pan. 1.75" upper diameter, 2.5" lower diameter

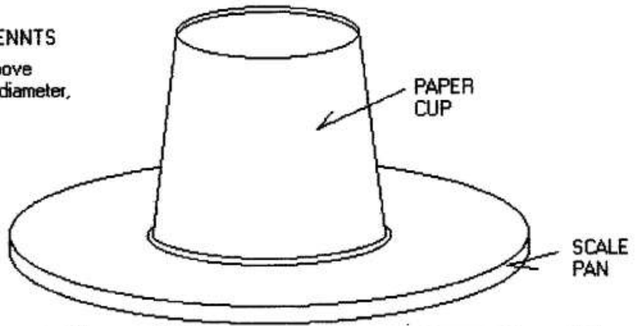

DIAGRAM 3

02/04/2002 EXPERIMENTS

MU METAL SPECIFICATIONS:

Source: NATIONAL ELECTRICAL ALLOYS
Oakland, NJ
SPECIFICATION: MIL N 14411C COMP 1
GRADE: HY MU 80 SHIELDING ALLOY
COLD ROLLED BRIGHT
DESCRIPTION: COIL
HEAT # 982011207
CHEMICAL ANALYSIS:
Ni: 80.16%
Mo: 4.6%
S: 0.0007%
C: 0.010%
P: 0.004%
Mn: 0.486%
Si: < 0.010%
Fe: BALANCE
MECHANICAL PROPERTIES:
HARDNESS: HV1 = 159
GRAIN SIZE: 8.5
COERCIVE STRENGTH:
HC = 0.0080 Oe
PERMEABILITY: MUMAX = 355000
SATURATION: 7900 Gauss
THICKNESS: 0.010"
DENSITY: 8.747-gram/cc

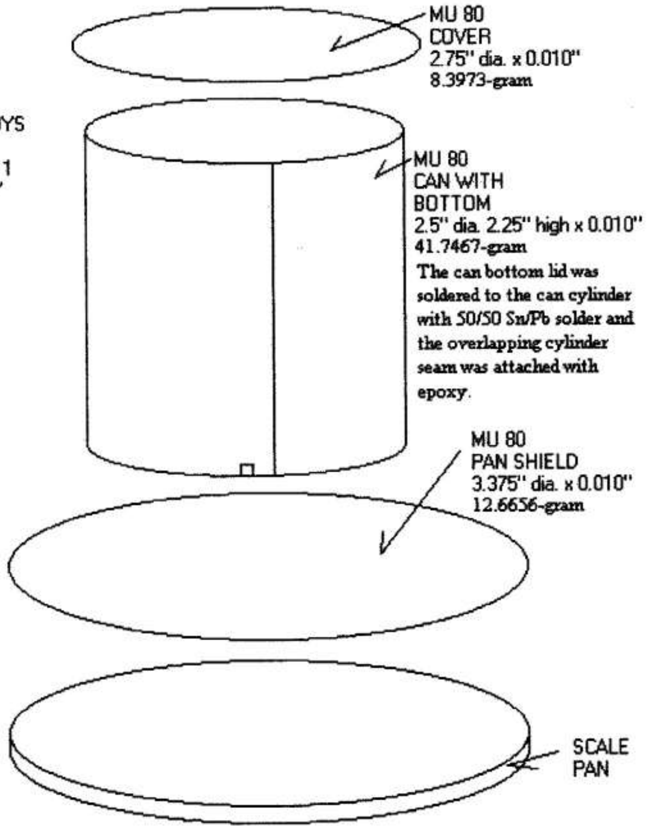

The can bottom lid was soldered to the can cylinder with 50/50 Sn/Pb solder and the overlapping cylinder seam was attached with epoxy.

DIAGRAM E

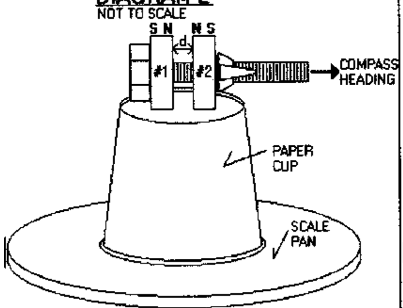
NOT TO SCALE
S N
N S
d
#1
#2
COMPASS HEADING
PAPER CUP
SCALE PAN

DIAGRAM F

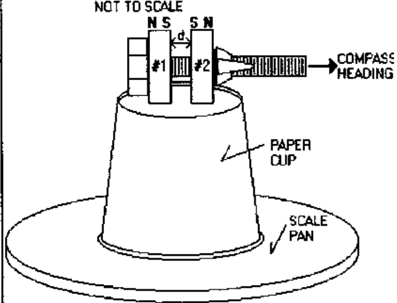
NOT TO SCALE
N S
S N
d
#1
#2
COMPASS HEADING
PAPER CUP
SCALE PAN

DIAGRAM G

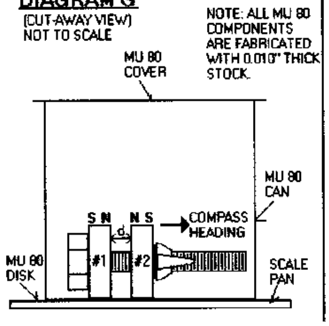
(CUT-AWAY VIEW)
NOT TO SCALE
NOTE: ALL MU 80 COMPONENTS ARE FABRICATED WITH 0.010" THICK STOCK
MU 80 COVER
MU 80 CAN
S N
N S
d
COMPASS HEADING
#1
#2
MU 80 DISK
SCALE PAN

DIAGRAM H

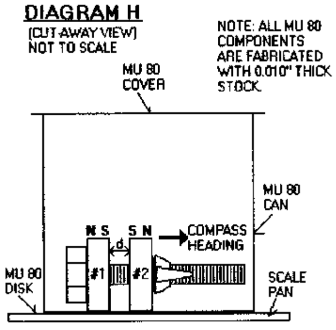
(CUT-AWAY VIEW)
NOT TO SCALE
NOTE: ALL MU 80 COMPONENTS ARE FABRICATED WITH 0.010" THICK STOCK
MU 80 COVER
MU 80 CAN
N S
S N
d
COMPASS HEADING
#1
#2
MU 80 DISK
SCALE PAN

ELEVATION VIEW OF THE VERTICAL EXPERIMENTS

NOTE:

January 14, 2002 Experiments: Start time: 13:27 hours EST, End time: 15:40 hours EST.

February 4, 2002 Experiments: Start time: 13:18 hours EST, End time: 14:50 hours EST.

TABLE 1:

	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap d	**Diagram A** N to N vertical 19.8893-gram (Sum of 1 & 2)*	**Diagram B** S to S vertical 19.8917-gram (Sum of 1 & 2)*	**Diagram C** N to N vertical 19.9025-gram (Sum of 1 & 2)*	**Diagram D** S to S vertical 19.9013-gram (Sum of 1 & 2)*
0.0 Inch	19.8759-gram 0.0134-gram weight reduction (0.06737%)	19.8757-gram 0.0160-gram weight reduction(0.08045%)	19.8760-gram 0.0265-gram weight reduction(0.13324%)	19.8758-gram 0.0255-gram weight reduction(0.12821%)
1/8 Inch	19.8761-gram 0.0132-gram weight reduction(0.06637%)	19.8751-gram 0.0166-gram weight reduction(0.08346%)	19.8760-gram 0.0265-gram weight reduction(0.13324%)	19.8763-gram 0.0250-gram weight reduction(0.12570%)
1/4 Inch	19.8763-gram 0.0130-gram weight reduction(0.06536%)	19.8748-gram 0.0169-gram weight reduction(0.08497%)	19.8760-gram 0.0265-gram weight reduction(0.13324%)	19.8768-gram 0.0245-gram weight reduction(0.12318%)
3/8 Inch	19.8765-gram 0.0128-gram weight reduction(0.06436%)	19.8753-gram 0.0164-gram weight reduction(0.08246%)	19.8760-gram 0.0265-gram weight reduction(0.13324%)	19.8777-gram 0.0236-gram weight reduction(0.11866%)
1/2 Inch	19.8774-gram 0.0119-gram weight reduction(0.05983%)	19.8753-gram 0.0164-gram weight reduction(0.08246%)	19.8757-gram 0.0268-gram weight reduction(0.13475%)	19.8809-gram 0.0204-gram weight reduction(0.10257%)
5/8 Inch	19.8776-gram 0.0117-gram weight reduction(0.05883%)	19.8754-gram 0.0163-gram weight reduction(0.08195%)	19.8779-gram 0.0246-gram weight reduction(0.12368%)	19.8806-gram 0.0207-gram weight reduction(0.10408%)
11/16 Inch	19.8777-gram 0.0116-gram weight reduction(0.05832%)	19.8748-gram 0.0169-gram weight reduction(0.08497%)	19.8773-gram 0.0252-gram weight reduction(0.12670%)	19.8832-gram 0.0181-gram weight reduction(0.09100%)

* Magnets #1 and #2 were individually weighed in the orientation used in each experiment and their separate weights were added together.

The following equation was used to calculate the weight changes, in dynes, in the proceeding graphs.

$$\overrightarrow{\Delta F} = k \cdot \Delta\text{mass} \cdot g$$

where

$$k = 980.665 \cdot \text{dyne} \cdot \text{gm}^{-1}$$

and

$$g = 9.80665\ \text{m} \cdot \text{sec}^{-2}$$

which is the local rate of gravitational acceleration.

The product of the measured changes in mass, Δmass, and g is denoted as follows on the graphs:

$$\text{NNopen}_{n,1} = \Delta\text{mass} \cdot g$$

for N-to-N pole facings in the open (or unshielded).

$$\text{SSopen}_{n,1} = \Delta\text{mass} \cdot g$$

for S-to-S pole facings in the open (or unshielded).

$$\text{NNshield}_{p,1} = \Delta\text{mass} \cdot g$$

for N-to-N pole facings shielded with Mu 80 shielding.

$$\text{SSshield}_{p,1} = \Delta\text{mass} \cdot g$$

for S-to-S pole facings shielded with Mu 80 shielding. In the subscripts, ${}_{n}$ and ${}_{p}$ refers to the respective number of data points per plot. The subscript ${}_{1}$ refers to the vertical change in force (weight change) axis and ${}_{0}$ refers to the horizontal distance d axis. In **GRAPH 1**, ${}_{n} = {}_{p}$.

GRAPH 1:

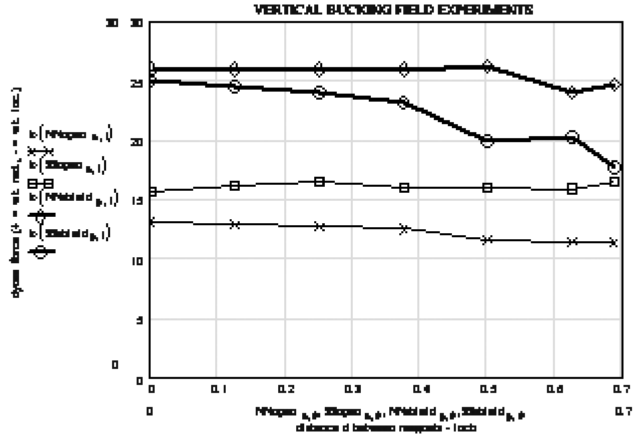

ISOMETRIC AND ELEVATION VIEW OF THE HORIZONTAL EXPERIMENTS

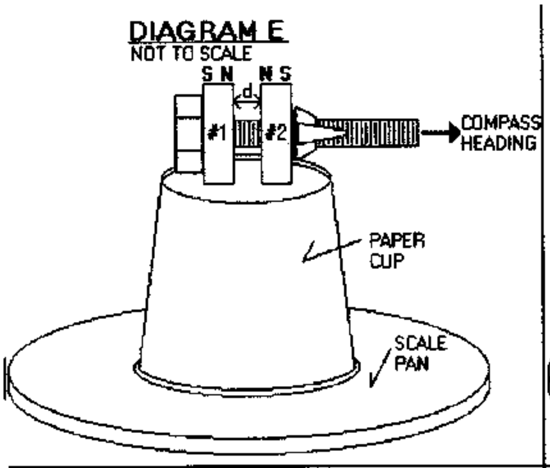
DIAGRAM E
NOT TO SCALE
S N
N S
d
#1
#2
COMPASS HEADING
PAPER CUP
SCALE PAN

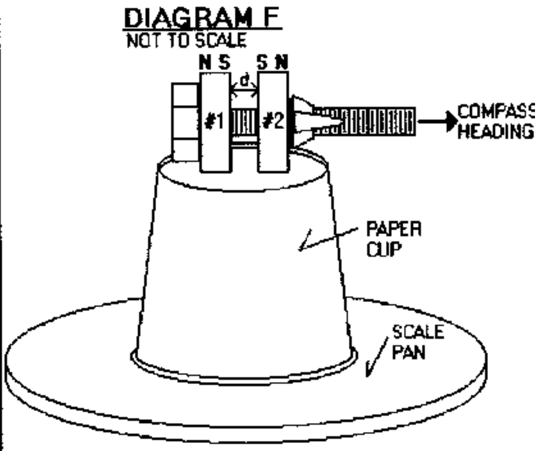
DIAGRAM F
NOT TO SCALE
N S
S N
d
#1
#2
COMPASS HEADING
PAPER CUP
SCALE PAN

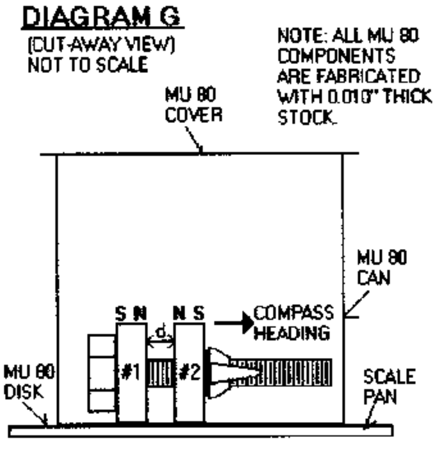
DIAGRAM G
(CUT-AWAY VIEW)
NOT TO SCALE
NOTE: ALL MU 80 COMPONENTS ARE FABRICATED WITH 0.010" THICK STOCK.
MU 80 COVER
MU 80 CAN
S N
N S
d
#1
#2
COMPASS HEADING
MU 80 DISK
SCALE PAN

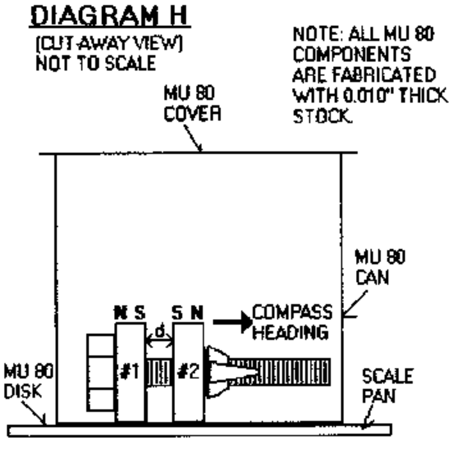
DIAGRAM H
(CUT-AWAY VIEW)
NOT TO SCALE
NOTE: ALL MU 80 COMPONENTS ARE FABRICATED WITH 0.010" THICK STOCK.
MU 80 COVER
MU 80 CAN
N S
S N
d
#1
#2
COMPASS HEADING
MU 80 DISK
SCALE PAN

MAGNETIC DECLINATION FROM THE LOCAL TOPOLOGICAL MAP:

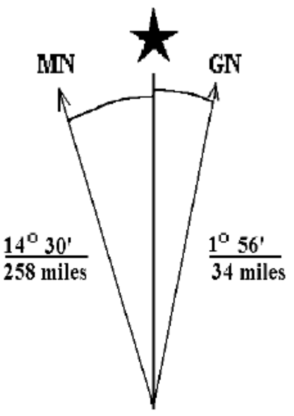

1984 Magnetic Declination ~ U. S. Geological Survey
New London, Connecticut Quadrangle Topological Map

TABLE 2

North # **Heading**	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap d	**Diagram E** N to N horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram F** S to S horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram G** N to N horizontal 19.9019-gram (Sum of 1 & 2)*	**Diagram H** S to S horizontal 19.9019-gram (Sum of 1 & 2)*
0.0 Inch	19.8735-gram 0.0170-gram weight reduction (0.08547%)	19.8778-gram 0.0127-gram weight reduction(0.06385%)	19.8890-gram 0.0129-gram weight reduction(0.06479%)	19.8810-gram 0.0209-gram weight reduction(0.10497%)
1/8 Inch	19.8728-gram 0.0177-gram weight reduction(0.08899%)	19.8790-gram 0.0115-gram weight reduction(0.05782%)	N/A **	N/A **
1/4 Inch	19.8721-gram 0.0184-gram weight reduction(0.09251%)	19.8798-gram 0.0107-gram weight reduction(0.05379%)	N/A **	N/A **
3/8 Inch	19.8705-gram 0.0200-gram weight reduction(0.10055%)	19.8800-gram 0.0105-gram weight reduction(0.05279%)	N/A **	N/A **
1/2 Inch	19.8698-gram 0.0207-gram weight reduction(0.10407%)	19.8811-gram 0.0094-gram weight reduction(0.04726%)	N/A **	N/A **
5/8 Inch	19.8693-gram 0.0212-gram weight reduction(0.10658%)	19.8818-gram 0.0087-gram weigh t reduction(0.04374%)	N/A **	N/A **
11/16 Inch	19.8696-gram 0.0209-gram weight reduction(0.10508%)	19.8827-gram 0.0078-gram weight reduction(0.03921%)	19.9190-gram 0.0171-gram weight increase(0.08588%)	19.9029-gram 0.0010-gram weight increase(0.00502%)

Compass Heading is approximate
* Magnets #1 and #2 were individually weighed with pole faces oriented vertically, with N up then with S up, and the results were averaged and added.
** Due to time constraints, these measurements were not taken.

GRAPH 2:

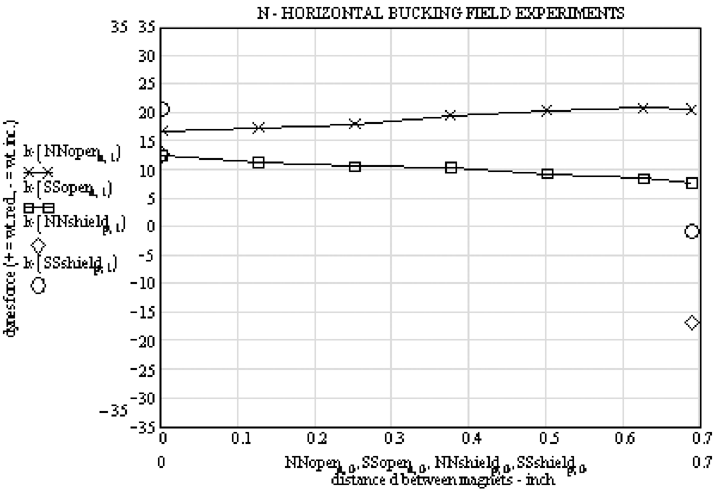

NOTE:
The calculated mass of each magnet was determined by the following formula based upon the manufacturer's dimensions and density value.

$$\text{volume} := \pi\cdot\left[\left(\frac{0.75\cdot\text{in}}{2}\right)^2-\left(\frac{0.43\cdot\text{in}}{2}\right)^2\right]\cdot\left(\frac{1}{4}\right)\cdot\text{in}$$

$$\text{volume} = 1.214963\ \text{cm}^3$$

$$\text{density} = 8.7\cdot\text{gm}\cdot\text{cm}^{-3}$$

$$\text{mass} = \text{density}\cdot\text{volume}$$

$$\text{mass} = 10.570177\ \text{gm.}$$

This is higher than the magnetized mass of each magnet.

However, some tables give a lower density for the SmCo magnet, 0.300lb/in^3, which equals:

$$\text{density} = 8.303971\cdot\text{gm}\cdot\text{cm}^{-3}$$

$$\text{mass} = 10.089017\ \text{gm. .}$$

This is still higher than the magnetized mass of each magnet. Does the SmCo material become slightly lighter in weight when it is magnetized?

TABLE 3

East # **Heading**	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap D	**Diagram E** N to N horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram F** S to S horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram G** N to N horizontal 19.9019-gram (Sum of 1 & 2)*	**Diagram H** S to S horizontal 19.9019-gram (Sum of 1 & 2)*
0.0 Inch	19.8763-gram 0.0142-gram weight reduction (0.07139%)	19.8755-gram 0.0150-gram weight reduction(0.07541%)	19.8928-gram 0.0091-gram weight reduction(0.04570%)	19.8731-gram 0.0288-gram weight reduction(0.14464%)
1/8 Inch	19.8777-gram 0.0128-gram weight reduction(0.06435%)	19.8750-gram 0.0155-gram weight reduction(0.07793%)	N/A **	N/A **
1/4 Inch	19.8782-gram 0.0123-gram weight reduction(0.06184%)	19.8743-gram 0.0162-gram weight reduction(0.08145%)	N/A **	N/A **
3/8 Inch	19.8779-gram 0.0126-gram weight reduction(0.06335%)	19.8722-gram 0.0183-gram weight reduction(0.09200%)	N/A **	N/A **
1/2 Inch	19.8792-gram 0.0113-gram weight reduction(0.05681%)	19.8715-gram 0.0190-gram weight reduction(0.09552%)	N/A **	N/A **
5/8 Inch	19.8814-gram 0.0091-gram weight reduction(0.04575%)	19.8712-gram 0.0193-gram weight reduction(0.09703%)	N/A **	N/A **
11/16 Inch	19.8815-gram 0.0090-gram weight reduction(0.04525%)	19.8720-gram 0.0185-gram weight reduction(0.09301%)	19.9338-gram 0.0319-gram weight increase(0.16021%)	19.9030-gram 0.0011-gram weight increase(0.00552%)

Compass Heading is approximate
* Magnets #1 and #2 were individually weighed with pole faces oriented vertically, with N up then with S up, and the results were averaged and added.
** Due to time constraints, these measurements were not taken.

GRAPH 3:

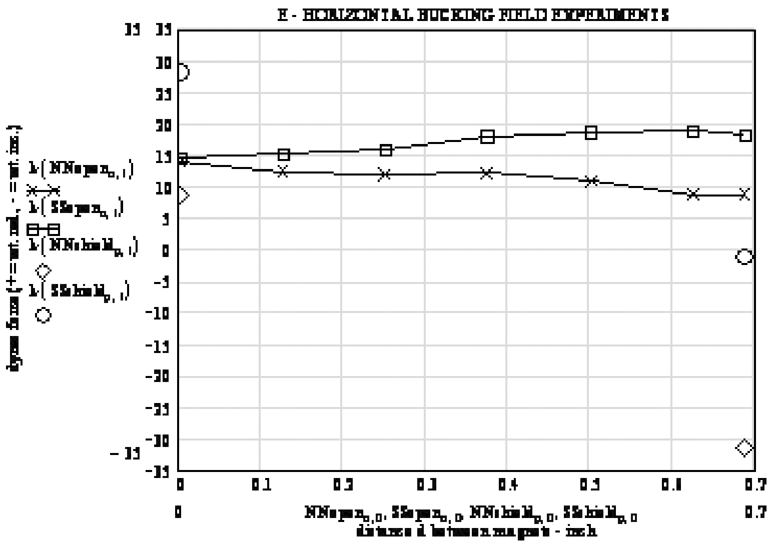

TABLE 4

East # **Heading**	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap D	**Diagram E** N to N horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram F** S to S horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram G** N to N horizontal 19.9019-gram (Sum of 1 & 2)*	**Diagram H** S to S horizontal 19.9019-gram (Sum of 1 & 2)*
0.0 Inch	19.8746-gram 0.0159-gram weight reduction (0.07994%)	19.8769-gram 0.0136-gram weight reduction(0.06837%)	19.8896-gram 0.0123-gram weight reduction(0.06178%)	19.8660-gram 0.0359-gram weight reduction(0.18030%)
1/8 Inch	19.8720-gram 0.0185-gram weight reduction(0.09301%)	19.8788-gram 0.0117-gram weight reduction(0.05882%)	N/A **	N/A **
1/4 Inch	19.8709-gram 0.0196-gram weight reduction(0.09854%)	19.8797-gram 0.0108-gram weight reduction(0.05430%)	N/A **	N/A **
3/8 Inch	19.8704-gram 0.0201-gram weight reduction(0.10105%)	19.8803-gram 0.0102-gram weight reduction(0.05128%)	N/A **	N/A **
1/2 Inch	19.8711-gram 0.0194-gram weight reduction(0.09753%)	19.8814-gram 0.0091-gram weight reduction(0.04575%)	N/A **	N/A **
5/8 Inch	19.8698-gram 0.0207-gram weight reduction(0.10407%)	19.8825-gram 0.0080-gram weight reduction(0.04022%)	N/A **	N/A **
11/16 Inch	19.8703-gram 0.0202-gram weight reduction(0.10156%)	19.8817-gram 0.0088-gram weight reduction(0.04424%)	19.9423-gram 0.0404-gram weight increase(0.20290%)	19.8810-gram 0.0209-gram weight reduction(0.10497%)

Compass Heading is approximate

* Magnets #1 and #2 were individually weighed with pole faces oriented vertically, with N up then with S up, and the results were averaged and added.

** Due to time constraints, these measurements were not taken.

GRAPH 4:

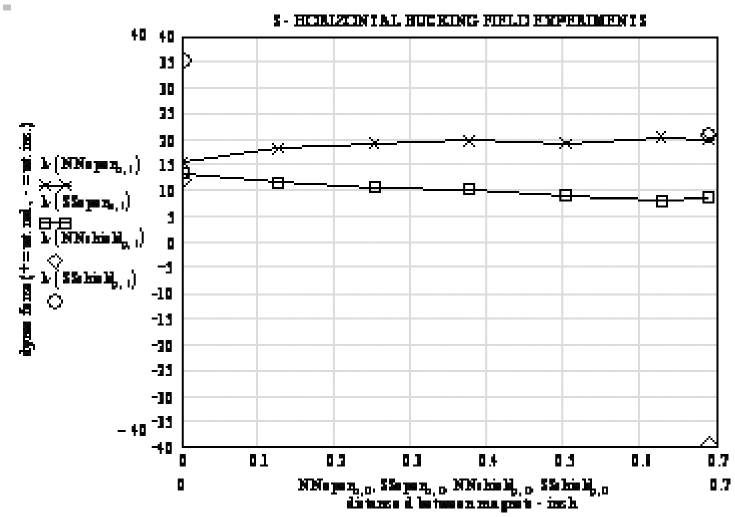

TABLE 5

West # **Heading**	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap D	**Diagram E** N to N horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram F** S to S horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram G** N to N horizontal 19.9019-gram (Sum of 1 & 2)*	**Diagram H** S to S horizontal 19.9019-gram (Sum of 1 & 2)*
0.0 Inch	19.8777-gram 0.0128-gram weight reduction (0.06787%)	19.8748-gram 0.0157-gram weight reduction(0.07893%)	19.8933-gram 0.0086-gram weight reduction(0.04319%)	19.8724-gram 0.0295-gram weight reduction(0.14816%)
1/8 Inch	19.8767-gram 0.0138-gram weight reduction(0.06938%)	19.8737-gram 0.0168-gram weight reduction(0.08446%)	N/A **	N/A **
1/4 Inch	19.8777-gram 0.0128-gram weight reduction(0.06435%)	19.8733-gram 0.0172-gram weight reduction(0.08647%)	N/A **	N/A **
3/8 Inch	19.8783-gram 0.0122-gram weight reduction(0.06134%)	19.8727-gram 0.0178-gram weight reduction(0.08949%)	N/A **	N/A **
1/2 Inch	19.8806-gram 0.0099-gram weight reduction(0.04977%)	19.8722-gram 0.0183-gram weight reduction(0.09200%)	N/A **	N/A **
5/8 Inch	19.8811-gram 0.0094-gram weight reduction(0.04726%)	19.8715-gram 0.0190-gram weight reduction(0.09552%)	N/A **	N/A **
11/16 Inch	19.8803-gram 0.0102-gram weight reduction(0.05128%)	19.8723-gram 0.0182-gram weight reduction(0.09150%)	19.9427-gram 0.0408-gram weight increase(0.20491%)	19.8976-gram 0.0043-gram weight reduction(0.02160%)

Compass Heading is approximate

* Magnets #1 and #2 were individually weighed with pole faces oriented vertically, with N up then with S up, and the results were averaged and added.

** Due to time constraints, these measurements were not taken.

GRAPH 5:

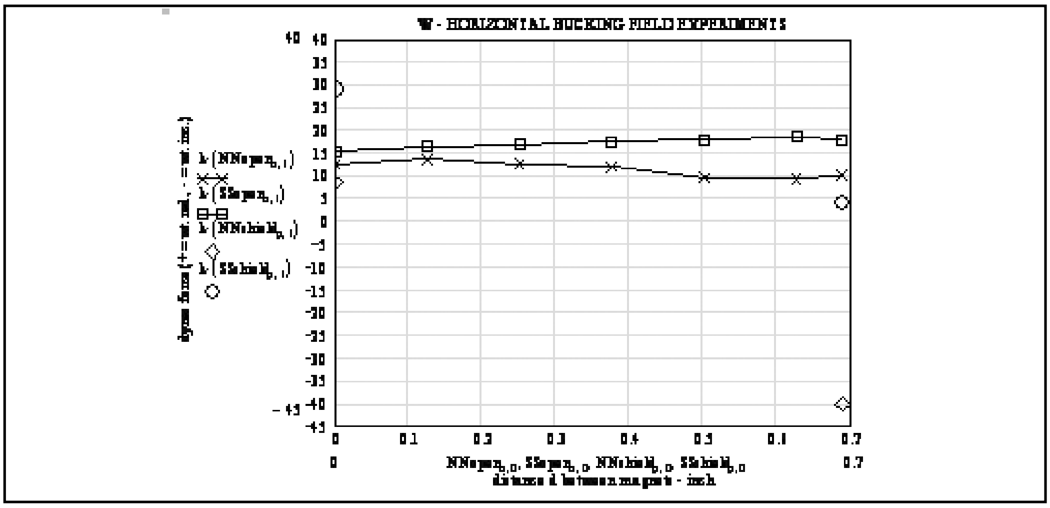

RING MAGNET #1, N UP: 9.9483-grams
RING MAGNET #2, N UP: +9.9527-grams
TOTAL WEIGHT: 19.9010-grams
WEIGHT INCREASE WITH OPPOSITE POLES ATTRACTING, AS DEPICTED IN **DIAGRAM I**, IN MU 80 SHIELD where 19.9861-grams was the measured value:
19.9861-grams – 19.9010-grams = 0.0851-gram weight increase.

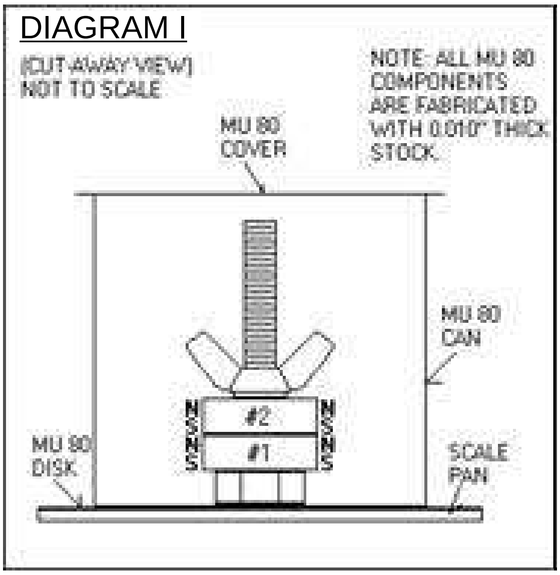

THREE RIVERS COMMUNITY COLLEGE, THAMES VALLEY CAMPUS LAB REPORT ON SmCo RING MAGNET EXPERIMENTS

The following experiments were conducted at the *Thames Valley Campus* (TVC) of the *Three Rivers Community College*, Room #207 Chemistry Laboratory, in Norwich, Connecticut on March 8, 2002. (41° 30' 34.62" N. latitude x 72° 6' 13.63" W. longitude x 115 feet elevation above mean sea level) I wish to thank the instructors at *Three Rivers* for their generosity for providing the use of their *Sartorius*® Model # 2442 analytical balance for the measurements taken in the proceeding experiments. The *Sartorius* Model # 2442 analytical balance is an enclosed pan unit with a maximum mass range of 200-grams with 0.0001-gram micrometer readability and a precision of 0.05-mg standard deviation.

The purpose of these experiments was to compare the Avery Point vertical measurements conducted on the *Mettler Toledo*® Model AG104 electronic scale, in **TABLE 1**, with the *Sartorius* Model # 2442 analytical balance measurements recorded in **TABLE 6**. The same two Samarium Cobalt magnets (magnet #1 and magnet #2) weighed individually in each vertical orientation were compared to the weight measurements taken when they were assembled using the nylon bolt and wing nut depicted in **DIAGRAM 1**. The specifications for the two Samarium Cobalt magnets used in the following experiments are shown in **DIAGRAM 1**.

The first column in **TABLE 6**, the vertical measurements, is the distance of separation d, or air gap, of the magnets. The second column shows the weight measurements of the two magnets, as shown in **DIAGRAM A**. The third column shows the weight measurements of the two magnets, as shown in **DIAGRAM B**. An inverted paper cup was used to raise the test sample magnets 2.75" above the *Sartorius* Model # 2442 balance scale pan in order to minimize possible magnetic interaction with the balance, as depicted in **DIAGRAM 2**. The fourth column shows the weight measurements of the two magnets, as shown in **DIAGRAM C**. The fifth column shows the weight measurements of the two magnets, as shown in **DIAGRAM D**. These experiments were shielded with

Mu 80 magnetic shielding material as depicted in **DIAGRAM 3**. The resultant data of **TABLE 6** is plotted on **GRAPH 6**. The results of the previous experiments at *Avery Point*, from **GRAPH 1**, and the recent *Thames Valley* experiments, from **GRAPH 6**, are plotted on **GRAPH 7** for comparison. The Mu 80 magnetically shielded experiments on the *Sartorius* Model # 2442 analytical balance at *Thames Valley* are in close agreement with the data collected with the AG104 electronic scale at Avery Point. However, the *Thames Valley* data collected for the unshielded experiments is somewhat smaller in weight reduction. I attribute this to external interference. The *Thames Valley* setup included a nonferrous tabletop, as did the Avery Point setup. However, the *Thames Valley* balance table consisted of a steel frame and legs, which may have altered the readings. The Mu 80 shielding provided a more intrinsic method for accurate data collection.

The horizontal measurements were not taken in this set of experiments due to time constraints. A final experiment was attempted to replicate the relative weight increase with the opposite poles of the ring magnets "stuck" together, as depicted in **DIAGRAM I**; however, the magnets shattered during assembly!

TABLE 6

	03/08/2002 TRCC EXPERIMENTS, @ TVC: OPEN, WITH PAPER CUP		03/08/2002 TRCC EXPERIMENTS, @ TVC: WITH MU 80 SHIELDING	
Air gap d	**Diagram A** N to N vertical 19.8734-gram (Sum of 1 & 2)*	**Diagram B** S to S vertical 19.8726-gram (Sum of 1 & 2)*	**Diagram C** N to N vertical 19.8912-gram (Sum of 1 & 2)*	**Diagram D** S to S vertical 19.8929-gram (Sum of 1 & 2)*
0.0 Inch	19.8699-gram 0.0035-gram weight reduction (0.01761%)	19.8696-gram 0.0030-gram weight reduction(0.01510%)	19.8691-gram 0.0221-gram weight reduction(0.11110%)	19.8706-gram 0.0223-gram weight reduction(0.11210%)
1/8 Inch	19.8699-gram 0.0035-gram weight reduction(0.01761%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8688-gram 0.0224-gram weight reduction(0.11261%)	19.8707-gram 0.0222-gram weight reduction(0.11160%)
1/4 Inch	19.8695-gram 0.0039-gram weight reduction(0.01962%)	19.8695-gram 0.0031-gram weight reduction(0.01560%)	19.8687-gram 0.0225-gram weight reduction(0.11312%)	19.8712-gram 0.0217-gram weight reduction(0.10908%)
3/8 Inch	19.8695-gram 0.0039-gram weight reduction(0.01962%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8692-gram 0.0220-gram weight reduction(0.11060%)	19.8730-gram 0.0199-gram weight reduction(0.10004%)
1/2 Inch	19.8689-gram 0.0045-gram weight reduction(0.02264%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8695-gram 0.0217-gram weight reduction(0.10909%)	19.8738-gram 0.0191-gram weight reduction(0.09601%)
5/8 Inch	19.8686-gram 0.0048-gram weight reduction(0.02415%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8702-gram 0.0210-gram weight reduction(0.10557%)	19.8746-gram 0.0183-gram weight reduction(0.09199%)
11/16 Inch	19.8680-gram 0.0054-gram weight reduction(0.02717%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8707-gram 0.0205-gram weight reduction(0.10306%)	19.8756-gram 0.0173-gram weight reduction(0.08697%)

* Magnets #1 and #2 were individually weighed in the orientation used in each experiment and their separate weights were added together.

NOTE:
March 8, 2002 Experiments: Start time: 12:12 hours EST, End time: 13:47 hours EST.

GRAPH 6:

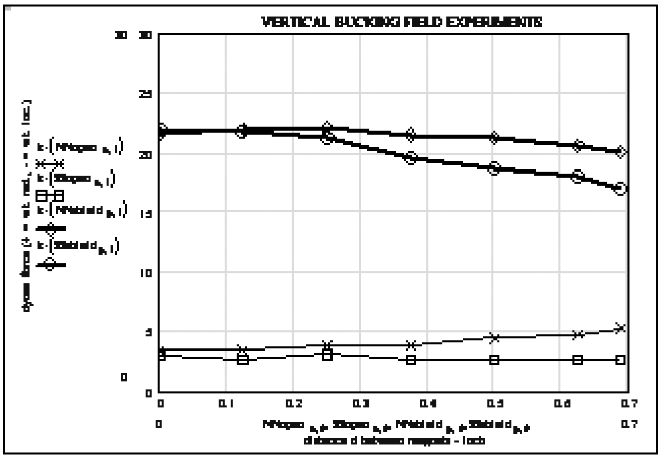

MAGNETIC DECLINATION FROM THE LOCAL TOPOLOGICAL MAP

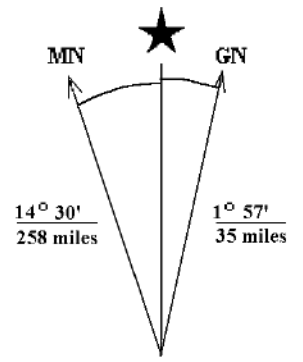

1983 Magnetic Declination~ U. S. Geological Survey
Norwich, Connecticut Quadrangle Topological Map

GRAPH 7:

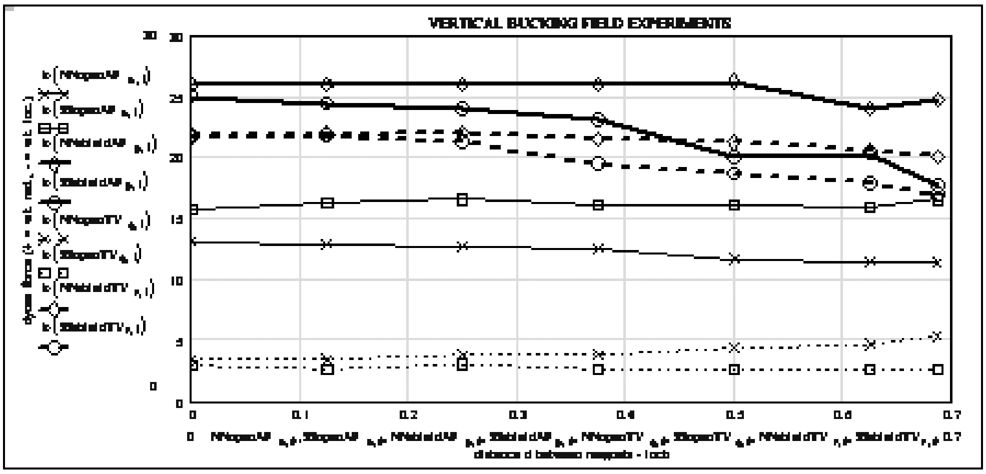

KEY TO **GRAPH 7**:

For N-to-N pole facings in the open (or unshielded) at Avery Point (AP):
NNopenAP
For S-to-S pole facings in the open (or unshielded) at Avery Point (AP):
SSopenAP
For N-to-N pole facings shielded with Mu 80 shielding at Avery Point (AP):
NNshieldAP
For S-to-S pole facings shielded with Mu 80 shielding at Avery Point (AP):
SSshieldAP
For N-to-N pole facings in the open (or unshielded) at Thames Valley (TV):
NNopenTV
For S-to-S pole facings in the open (or unshielded) at Thames Valley (TV):
SSopenTV
For N-to-N pole facings shielded with Mu 80 shielding at Thames Valley (TV):
NNshieldTV
For S-to-S pole facings shielded with Mu 80 shielding at Thames Valley (TV):
SSshieldTV

The subscripts for the data points $_{n} = {}_{p} = {}_{q} = {}_{r} = 7$ are all the same value in **GRAPH 7**.

Advanced Nuclear Waste Decontamination Technologies

Mark Porringa

Zeropoint Techtonix Inc, 430 Bass Lake Road, R R # 1,
Deep River, Ontario K0J 1P0, Canada
(613) 584-2960 fax: (613) 584-4616 porringam@aecl.ca

It is a review of nine alternative, peer-reviewed techniques as candidates for the global clean-up of nuclear waste.

The following is a reasonably comprehensive list of potentially effective nuclear waste treatment methods that might be employed to treat the entire range of radioactive wastes that have proven to be such a daunting and horrendously expensive problem for the nuclear industry (in all its forms) with major, long term implications for the environment.

A wide variety of methods will probably be required to accommodate the many different radioactive waste sources including high and low level, solids, liquids and gases. Process names used here are in some cases just convenient labels used to categorize and set them apart from each other.

Theories on several of these processes are still quite speculative and solid evidence that would pass conventional peer review is still lacking. This is after all a very new field of science.

Some of these technologies are already well protected by international or national patents, with additional US and international patents pending, and further patents may be obtained on new developments as they are made.

The Brown's Gas-Metal Matrix Process:

The BG-MMX process utilizes a patented electrolysis cell of the Australian Prof. **Yull Brown**'s design that is said to produce a stoichiometric mixture of monatomic hydrogen and oxygen or possibly a quasi-stable water molecule raised to a high-energy state. This gas has some very peculiar properties including the ability to sublimate tungsten (6000°C) with an implosive flame that burns cool in air with a temperature of only 130°C.

The gas is used to heat a proprietary mixture of metals and/or metal oxides including the radwaste to be neutralized. A highly exothermic radiant reaction appears to result in the immediate reduction of radioactivity approaching 95% of the original levels judging from preliminary tests, within seconds of treatment. The process is conjectured to be effective with high level solid wastes and possibly gasses, but probably not liquids. The high temperatures involved may also preclude the processing of more volatile wastes.

Since 1991, this technology has been successfully demonstrated, on a small scale, at least 50 times to US, Chinese, Japanese and United Kingdom officials on a variety of nuclear waste products including Americium, Cobalt, Uranium, and Plutonium. The technique can be applied for the immediate decontamination of stockpiles of nuclear waste materials being held near nuclear power plants. The process is very simple, safe, and inexpensive to develop further into robotics application for on-site treatment with no foreseen environmental effects.

Photoremediation

The *Photoremediation* process of the American Dr. **Paul Brown** is essentially conventional physics, *albeit* applied in a new and novel way. The process involves the use of a high-energy electron beam impinged on a target, which in turn produces a monochromatic gamma radiation that is tuned to induce *Photofission* and *Photoneutron* reactions in the target material causing rapid neutralization of radioactive isotopes. The efficiency claimed exceeds 500% due to the high cross-section reactions in the Giant Dipole Resonance region. The 10 MeV electron beam produces typical fission reactions in the 200MeV range effectively turning high level solid wastes such as spent fuel into an energy source. The process is apparently intended for on-site treatment with some waste-partitioning required, an aspect which may not be desirable in certain countries.

While this idea is similar in topology to a system being developed by **Los Alamos National Labs**, Dr. Paul Brown's approach offers several advantages: no need for extensive chemical pre-processing and the energy required to effect transmutation is greatly reduced. No new technology needs to be developed, yet the engineering of such a photon reactor must be completed and it could itself become a practical method for generating power.

ZIPP Fusion

The ZIPP fusion process, identified by **Mark Porringa**, induces a wide variety of fusion reactions, resulting from the radial compression of individual diatomic and other simple molecules dissolved or suspended in a light water, carbon arc electrolysis cell. A variety of other cell configurations are envisioned.

The process appears to produce only stable isotopes, which should therefore make it capable of stabilizing a wide variety of radioactive waste materials. The theory on the process draws from Condensed Charge phenomena, Brown's Gas implosion, cavitation bubble collapse and sonoluminesence - all variations of the Casimir effect - which is believed to cohere the Zero-point energy of Quantum Vacuum Fluctuations. Transmutations using variations of this basic process may be applicable to a wide variety of nuclear wastes and appears capable of operating with an efficiency exceeding 100%.

A major implication of this process is that the Strong force of the nucleus is understood as an ultra close range Casimir effect. **Oakridge** Nuclear Laboratories in the US in conjunction with several international collaborators have just (this month, in fact) announced a deuterium cold fusion process based on the essential elements of the ZIPP Fusion process first reported in 1998. The process is very simple and inexpensive to develop.

RIPPLE Fission

The RIPPLE Fission process is an adaptation of existing potential technology utilizing a supersonic ionized gas to aerosol a counter flow heat exchanger that envelopes the radioactive waste aerosol in a vacuum induced plasma vortex which appears to disrupt the matter stabilizing influence of the Quantum Vacuum fluctuations resulting in "gentle" low recoil fission reactions which produce only stable fission products, with excess neutrons being prompt converted to protons via quenched Beta emissions. The process is apparently proven with conventional non-radioactive wastes and is believed applicable to the entire spectrum of radwaste without the need for waste partitioning. This process is also conjectured to operate with over-unity efficiency.

The LENTEC Processes

The Low Energy Nuclear Transmutation Electrolytic Cells of the Cincinnati group produce a variety of transmutation reactions using a variety of exotic electrolysis cell designs that generally produce condensed charge clusters composed primarily of up to 10^{11} electrons each. These electron charge clusters produced with the use of special electrodes can penetrate the nuclei of larger atoms in solution and transmute these atoms into stable elements.

The range of design and operating protocols and potential applications are essentially limitless provided for the waste that is dispersed in the electrolyte. The reported transmutation of thorium to stable titanium and copper by the Cincinnati Group and by the Salt Lake City group is one of the most dramatic examples of this type of treatment process. Application to other high-level liquid transuranic fissionable wastes such as surplus Plutonium seems likely. The glaring absence of normal fission yield energies is perplexing but probably explicable as another form of low recoil fission reaction, similar to RIPPLE fission.

The Plasma Induced/Injected Transmutation - PIT Processes (also known as HDCC)

Plasma Induced/Injected Transmutation processes run include a gamut from recent achievements dating back to the **Oshawa-Kushi** cold plasma transmutations reported in 1964. The patented high-density charge cluster process (HDCC) was first discovered by **Kenneth Shoulders** and added on to by **Harold E. Puthoff**. Later, the late **Stan Gleeson** discovered HDCC in properly processed solutions. Still later, **Alexander Ilyanok** of Belarus discovered HDCC, followed by **Vasiliy Baraboskin** in Russia.

The production of Condensed Charge Clusters and various plasma glow discharge phenomena in a variety of gaseous atmospheres is again implicated as the underlying cause with what should be by now an obvious connection with the coherence of Zero-point energy from the Quantum or Stochastic vacuum.

Desk-top high energy particle accelerators have also been envisioned, based on the "piggy back" principle, in which the clusters permit acceleration of "piggy-backed" heaver +ions to extremely high energies capable of causing fusion and transmutations in target materials including those in solution and the materials of which the electrodes are composed. Brown's Gas implosion and cavitation bubble collapse reactions are also believed to be prevalent in these types of cells due to the prevalence of electrolysis.

A high-density charge cluster technology was discovered and used by Stan Gleeson to stabilize radioactive liquid wastes and has been developed further in the last 4 years by a group led by **S-X Jin** and **Hal Fox**. Best results for radioactive liquids have been demonstrated in the processing of thorium for a 30-minute period and achieving a reduction of radioactivity of about 90% from a liquid sample.

Kervran Reactions

The very compelling evidence compiled by French Nobel Candidate Dr. **Louis Kervran** has identified a wide range of nuclear transmutations in biological systems that have not been adequately explained. Coherence of Zero-point energy via Casimir effects within the *Somatid* particles identified by the Canadian **Gaston Naessens** is implicated as a possible cause. A wide variety of *in vitro* and *in vivo* reactions are believed to be possible as proven in nature and numerous experiments typically involving a reaction medium composed of a dielectric fluid such as water. Highly radiation resistant microorganisms have been found thriving in the core of nuclear reactors indicating the possibility of microorganisms being capable of transmuting some bioactive nuclear wastes in the course of the normal metabolism of such organisms.

The Monti Process

The Italian **Roberto A. Monti's** process involves confined explosions involving proprietary mixtures of materials that include radioactive waste. Ignition of such mixtures causes nuclear transmutations resulting in reduced radioactivity (to near-background levels) following combustion, gradually over 1 to 4 days.

This technique has been confirmed by the Italian **ENEA** and is supported by the French CEA scientists as a serious candidate for treatment of waste stockpiles. The system, as currently designed, required waste to be inserted into a chamber.

Higher group symmetry electrodynamics

Extremely weak, non-classical, higher group symmetry electromagnetic fields were found during a 1991 experiment made by **Glen Rein** to alter significantly the level of radioactivity in materials, even those in the environment. The experiments suggest that higher group symmetry electrodynamics modulate the quantitative and /or qualitative properties of radioactive species. If the non-classical fields directly affect the radioactive species, it is likely that the appropriate field parameters will be discovered to neutralize radioactive emissions. In 1999, a theoretical basis for the phenomenon was developed by the Welsh physicist, **M. W. Evans**, with the participation of Lt. Col. (retired) **Thomas E. Bearden**.

The technology is extremely simple and could be applied with minimum logistics for treating massive structures, *in-to* outdoors, such as the Chernobyl disaster site.

Psitronics Group Systems, International

Robert "Paul" LeBreton, 2901 Hwy. 6, HC 77 Box 42, Laguna, NM 87026 USA

E-mail: wizzard9@earthlink.net Phone:1-(505)-836-7534

Psi/Group's Magnetic Motor - Funding for New Prototype Sought:

As the engineer who designed the self-sustaining magnetic motor being advertised by Psitronics Group Systems, International. It perhaps falls upon me to explain (as well as I am able) the methods used by me in this motor's design:

The actual "picture" description is totally Intellectual Property; Protected as a Trade Secret.

However: This magnetic motor is a "perpetually imbalanced configuration" of permanent magnets; an unequal number of magnets on the rotor & stator; that revolves in a "self-sustaining" manner as the magnets seek balance... There are no electrical components in the design unless one wished to insert alternator windings in the stator to provide an electrical output...

We have discovered: through a prior unsuccessful prototype that using a para-magnetic material (we used aluminum) for the rotor & stator was self-defeating; as with the powerful rare earth magnets used in the design aluminum destroyed the magnetic fields... On further analysis of the failed prototype it was deduced that a "latch up" condition would occur between rotor & stator magnets; unless the rotor magnets were canted or skewed at an angle from the stator magnets... The plastic prototype contemplated corrects both of these problems....

We are seeking an investor willing to put up $4000 (DUS) for the prototype and who can offer $40,000 return on investment "if it proves successful" ...We have a Global Distribution 'ready to go" as well as a pledge of $ Millions towards Manufacturing & Marketing from EarthTech, International on submission of a running motor...

Very Respectfully Yours,
Paul LeBreton

http://home.earthlink.net/~wizzard9
(Psi/Group Website)

Inquiries should be emailed to:
wizzard9@earthlink.net

Experimental Investigations of the Radioactive Isotope Half-Value Period Changing in the Local Volume of Cause-Effect Relations

Igor A. Melnik

Pr.Frunze 232, Tomsk, 634021, Russia, Ph. (3822) 244555
breg@mail.ru

Nuclear-design methods of the elemental analysis are based on the property of the radioactive isotope decay rate constant. Half period as a constant quantity is defined by the time feature (the time flow is uniform in the every space point). The given postulate is confirmed by the long-term results and raised no doubts. But in last years some researches, investigating enough fine effects by change of the registered radiation intensivity, came to the conclusion about influence of space cycles on the time flow [1]. In his turn, N. A. Kozyrev during the investigation of active physical properties of time came to the conclusion about violation of its uniform flow in the local volume of the cause-effect relations, created by the cyclic motion (rotation, oscillation) of bodies [2,3,4].

Thus, the author get an idea to use a radioactive isotope cezium-137 as a “sensor”, measuring the changes of time flow uniformity in the certain local volume of cause-effect relations. In basis there are following arguments: owing to a conception of time uniformity and considerable cezium-137 half-value period, the source activity must be permanent during the experiment time. By the time flow (period) change, i.e. changes of uniformity in the local volume of cause-effect relations, the half-value period – T is changing in the direct proportion. Source activity, respectively, is changing in the inverse proportion according to the law exp (1/T) in relation to the external space volume. Hence, registered gamma-quantum intension (amplitude impulse distribution) in the absence of the cause-effect relations is proportional to the function $N_0 \sim \exp(\ln 2/T_0)$. Then the proportion $N_0/\Delta N$, where $\Delta N = N_0 - N$ is made, and by means of it a half-value period difference is defined by the formula

$$1/\Delta T = \left| 1/T_0 + \ln(\Delta N/N_0)/\ln 2 \right| \qquad (1)$$

N_0 – selective average amplitudes of impulses at the case of static liquid (in the absence of the cause-effect relations);
N - selective average amplitudes of impulses at the case of rotation of liquid.

In this case, there were investigations of the determination of the gamma-quantum intension change (i.e. change of a half-value period) dependence on the angular velocity of the activator rotation, and also on the coordinates location and amount of cause-effect relations (Fig. 1).

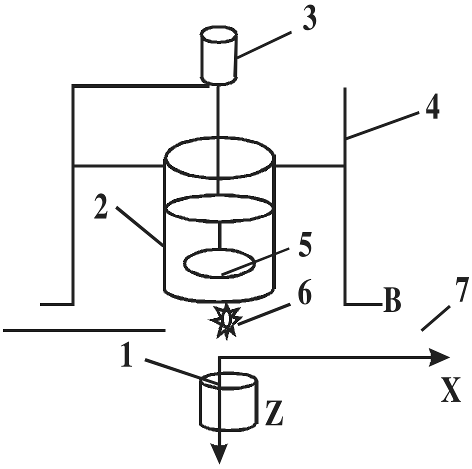

Fig.1

The following devices were used as measurement equipment: semi-conducting detector (SCD-63V) (1), preamplifier (PAG-2K), amplifier (BAI-3K) and analyzer (AMA-02F1). Energy gamma-line of 662 KeV was measured with the resolution 4 KeV. The gamma-quantum source was glued to the detector housing at the distance of 100 mm from its surface; so, any spatial change through coordinate axes was executed together with the detector to avoid even a tiny change of the source-detector geometry.

The vessel with a liquid (2) was placed above the source (6), vortex fluid motion was created by an activator (5), placed on the rod of electric motor (3). The glass with a liquid, connected with the motor and frame (4), was displaced regarding to the center of the revolved activator, what was the reason of the glass vibration in horizontal line, and, hence, one of frame sides vibration in vertical line (v). In its base, the frame was a square with the side of 200 mm. To assign the direction of the space cause-effect relations, one of frame sides was vibrating, and the opposite one (on the X axis) was fixed on the table (7). An experiment was made in such a way, that vibrating parts did not adjoin with the detector. The radius of glass is 50 mm, the distance from central axis up to the vibrating frame part is 100 mm. N-selection of the every value on the diagram (Fig. 2) corresponded to twenty measurements, roots from the average variance of numbers distribution $D \approx 160$ impulses.

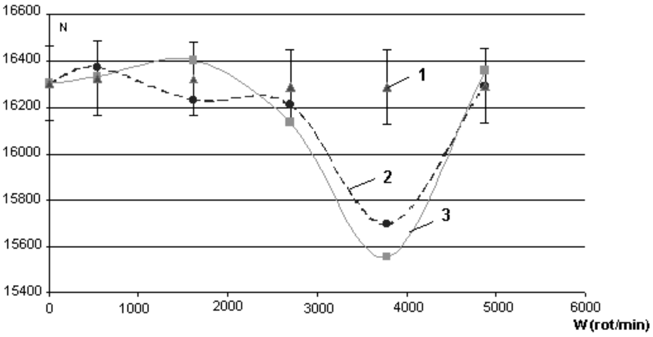

Fig. 2
Dependence of photopeak-N area from the angular velocity of rotation – W, where 1 – is photopeak area at static liquid, 2 – is photopeak area at clockwise rotation, 3 – is photopeak area at anticlockwise rotation

During two months eighteen experiments were made (at the one geometry) of the given effect fluctuation study, and there were no considerable deviation. An effect of the time deceleration always appeared at the angular speed of the activator rotation W=3780 r/min (during the vibration of the whole perimeter of the frame base). If there was a vibration on the only one side of the frame base (on the X axis), then this effect appeared at the higher speed of rotation. Unfortunately, the maximum motor speed of rotation W_{max}=4880 r/min did not let to define the precise position for the next rotation point. During the experimental results extrapolation $W_2 \approx 5930$ r/min ($W_2 \approx \pi W/2$) was obtained.

In the table there are selective middle amplitudes of impulses -$N=\sum N_n/n$ roots from the numbers distribution average variance D, dispersion of the simple average G at the different directions of the liquid rotation (for the rotation point W=3780 r/min). The half period is estimated by the equation (1) and formula $T=T_0+\Delta T$ where T_0=30.2 year (the cezium-137 half-value period).

Liquid motion	N, imp.	D	G	T, year
Static liquid	16304.0	160.7	35.9	30.200
Clockwise	15696.0	162.0	36.2	30.412
Anticlockwise	15553.0	165.0	36.9	30.423

During the investigation of the activator angular speed of rotation dependence on the registered gamma-quantum intension, **there was discovered the time deceleration effect in the fixed point of ambient space of the glass with the liquid.** The effect was versatile, during the further study of the reasons of the effect disappearance, the following regularities were found:

1. Intension change appeared only at the case of asymmetrically fixed glass, when there appeared horizontal oscillations, transmitted to the vertical oscillations of the frame;

2. An effect disappeared, if the gamma-quantum source was placed in the glass center (~ 2 ÷ 3 mm down from the glass bottom);

3. At the same speed of rotation but without liquid, with an eccentric activator, the time deceleration effect disappeared.

Will consider the system with liquid, its internal chain consisted of three cause-effect relations:

1. Activator-liquid;
2. Liquid-glass (frame);
3. Frame-table.

Glass oscillations were transmitted to the frame through the hard cohesion. There was only one cause-effect relation – an activator (a frame) – a table. Thus, amplitude of the time flow changing was, most likely, influenced on by the amount of cause-effect relation, and also, there was observed a quantum effect, dependent on the activator and liquid rotation frequency.

The time deceleration maximum amplitude appeared during the source displacement from the central axis of glass along the X coordinate to the distance ~ 5 mm. There appeared the necessity to check the given effect on space points along the all coordinate axes in the radius R ≈ 100 mm from the central point of the glass bottom (an extreme vibration point along the X axis was placed on this distance).

On this purpose there were made some experiments, results of which are shown on the (Fig. 3), (Fig. 4).

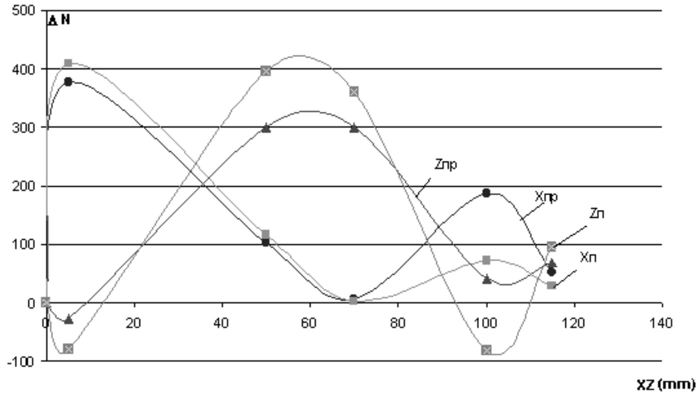

Fig. 3
Dependence of the increase of photopeak -ΔN area from the spatial interval in Cause-Effect Relations by X and Z coordinates, where X_cZ_c – is clockwise rotation, $X_{ac}Z_{ac}$ – is anticlockwise rotation

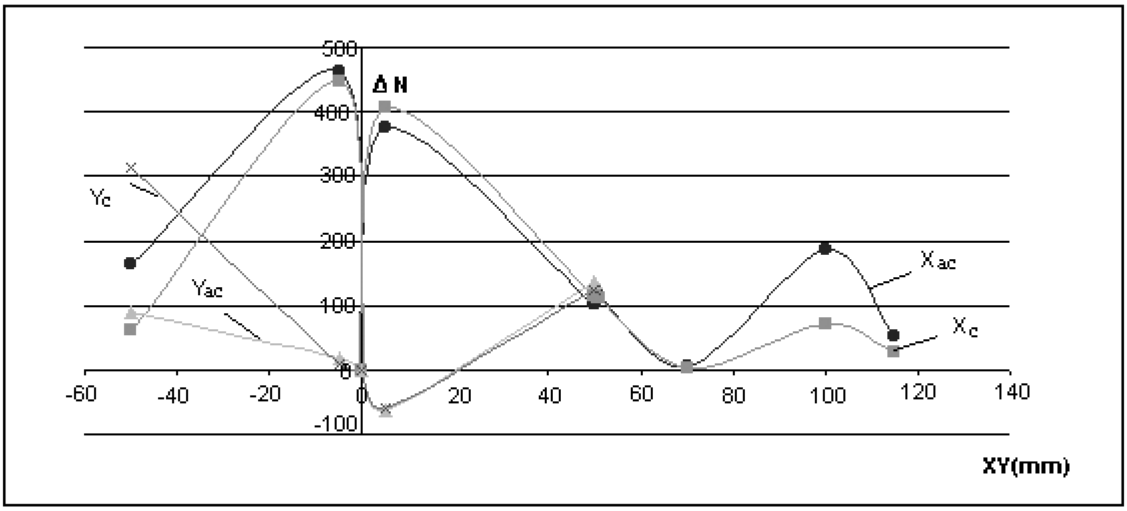

Fig. 4
Dependence of the increase of photopeak -ΔN area from the spatial interval in Cause-Effect Relations by X and Y coordinates, where X_cY_c – is clockwise rotation, $X_{ac}Y_{ac}$ – is anticlockwise rotation

Let us suppose that positive axis Z is directed from the glass bottom center to the ground and axis X – from the central axis of glass to the vibrated part of the frame (North-West), in this case Y is directed to the North-East. Along the all coordinates, in every adjusted space points on the distance R_i from the glass bottom center (R_i = 5; 50; 70; 100; 115 mm), there were forty measurements. At that in every point after ten measurements connected with the rotation there were measurements at the static liquid. It was made to avoid any systematic errors. The amplitude of the sample average impulses at the static liquid - $N_0=10550$ imp. In this case the root from the dispersion $\Delta=120$, dispersion of the simple average $G=19$. Amplitude increment was calculated by the formula $\Delta N=|N_0-N|$. Following regularities were found:

1. If considering the difference of impulse amplitude (in the positive coordinate space) in the first cause-effect relation evidence (activator-liquid), i.e., to put the difference between the initial (5 mm) and final (50 mm) points of amplitude registration (their middle evidence between "clockwise" and "anticlockwise") $\Delta N_5^{50}=|N_5-N_{50}|$ into the formula (1), then relations $\Delta T_Z/\Delta T_X=\Delta T_X/\Delta T_Y\approx1.11$ ($\pi/2\sqrt{2}\approx1.11$)

2. Change of impulse amplitude difference along X is the antiphase of changes along Z and Y;

3. At the rotation clockwise and anticlockwise there takes place a periodical inversion of amplitudes difference relation;

4. At the liquid rotation clockwise and anticlockwise, there appears an obvious distinction in evidences of the increment ΔN in negative region of coordinate axes;

5. Along Y and Z-axes in points (5; 100 mm) there is observed insignificant time acceleration.

The whole cause-effect relation system was defined as an internal (activator-table), as an external one (ground - system center of gravity). **In the internal space volume time at the certain conditions breaks its own uniformity, at that the time period change is nonlinear and is defined by its quantum nature.** Therefore, standard clocks in causal relations must be considered regarding to the center of gravity and location of the measurement point in the internal system space.

References

1. Shnol S.E., Zenchenko T.A., Zenchenko K.I., PozharskyE.V., Kolombet V.A., Konradov A.A. Regular change of the statistical distributions thin structure as a result of cosmophysical causes. Kushnirenko E.A., Pogozhev I.B. Comment to the article of Shnol S.E. and Co. Progress of physical sciences. Vol. 170, #2, p.213-218, 2000.

2. Kozyrev N.A. Causal or asymmetric mechanics in the linear approximation. Pulkovo, 1958, p. 90.

3. Kozyrev N. A. Astronomic observations by means of time physical properties. Symposium transactions, Erevan, 1977, p. 209-227.

4. Kozyrev N.A. About possibility of bodies mass and weight under the influence of time active properties. Eganova I.A. Analytical survey of modern time-keeping ideas and experiments. Novosibirsk, 1984, p. 92-98.

5. Melnik I.A. Experimental researches of the time flow local deceleration. VINITI 5.02.92, N1032-mg92, #4, p.70.

UP-TO-DATE INVENTION

TILLEY ELECTRIC VEHICLE

Tilley Foundation, Inc.
131 Hiwassee Road
Lebanon, TN 37087, USA
615-443-5315
E-mail: cktilley@bellsouth.net

Nowadays there stays urgent the development of technologies, which could possibly be combined into a reliable, commercially successful advanced self powered vehicle with additional desirable but currently commercially unavailable features.

The advantages of electric vehicles seem to be evident – they are clean, quiet, powerful, require much less maintenance than gasoline or diesel-fueled vehicles, and are inherently much simpler and easier to manufacture. Their drawbacks have been a short range, long battery recharging time, and a heavy, bulky battery pack.

After several years of personal accomplishments in the alternative energy industry, **Carl B. Tilley, the inventor from USA,** was convinced that it was possible to build an electric car that could be powered without the help of external power to keep the battery charged.

The concept to produce a useful electric performance car that would last more than a few hours and would be economical to run, safe to drive around town or across the United States and never use a drop of fuel challenges the future of transportation as we know it today.

With the establishment of the **Tilley Foundation, Inc.**, in the year 2001, Carl Tilley set out to prove it could be done. It was an ambitious project and it broke ground on the facility in Tennessee that would build the first self-generating electric car.

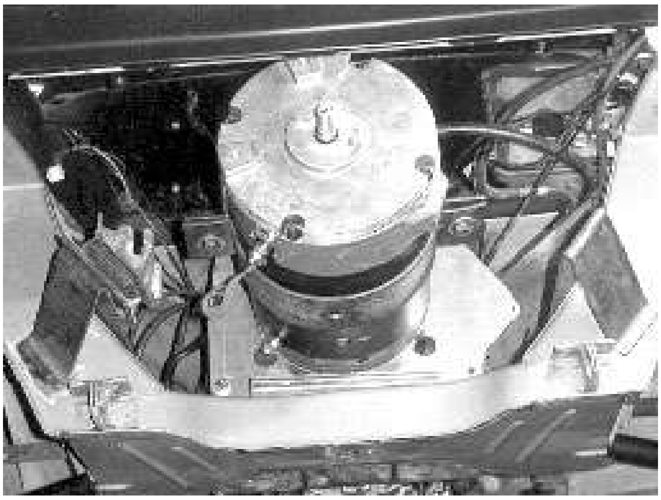

Construction of a 1,800 square foot building, that was powered with another recently developed electric device, began in the year 2002. Electricity for this car is provided from a different energy invention, which was void of any outside power supply. It is ironic that one alternative energy device actually built the invention to power and build the electric car.

The **Tilley Electric Vehicle** (**TEV**), converted from a 1981 DeLorean, energizes the imagination and defies what has been accepted as a standard in the area of transportation for years.

From the selection of the proper car to be converted, to the advanced technology, which is on board, the **TEV** performs comparably to gasoline-powered vehicles.The difference is you have no need for fuel and you do not have to stop the vehicle to charge it after driving. There is no pollution and you can cruise the highways at the same speed as any other vehicle.

There is a new car on the road today. A car built with technology that defies the concept of fossil fuel powered cars, and can run coast to coast without ever relying on the battery being charged from an outside source.

The long awaited transportation revolution and the end of our reliance on fossil fuel has now begun **....the Tilley Electric Vehicle.**

The demonstration of a DeLorean powered by an electric motor and 12-volt standard car batteries is supposed to be on September 7, 2002, at the Nashville Superspeedway, USA. The battery bank is kept in a charged condition by the "on board" charger which is the device invented by Carl B. Tilley. Racing legend Bobby Allison is one of the guest drivers for this demonstration.

Details and results of the demonstration read in our next issue

The Problem of Time: Force as the Cause of Change of the Course of Time

Valentin P. Oleinik

Department of General and Theoretical Physics,
National Technical University of Ukraine "Kiev Polytechnic Institute", Prospect Pobedy 37, Kiev, 03056, Ukraine;
e-mail: yuri@arepjev.relc.com
Internet: http://temporology.bio.msu.ru/OLEINIK/oleinik.htm.

Abstract

Material processes occurring in a physical system under the action of a force field necessarily influence the course of time along the trajectory of motion of particle. A general relationship is obtained which relates the course of time on one path section of a particle when moving in a force field to that on the other path section in the same inertial reference frame. According to the results obtained, the force in relativistic mechanics is not only the cause of acceleration of particle relative to an inertial frame of reference, but also the cause of change in the course of time along the particle's trajectory. ***Therein lies the physical content of the dynamical principle underlying the special theory of relativity (relativistic mechanics).*** *The applications of the theory developed to homogeneous fields - to the field of gravity and electromagnetic field, and to the gravitational field produced by a point mass particle are considered. Physical properties of the state of imponderability of particle in an external force field are investigated. It is noted that the change in the course of time in a force field is in no way connected with the change in space-time metric and is a direct consequence of the causality principle of relativistic mechanics.* ***The existence of dependence of the course of time on the state of motion of particle in a force field points to the feasibility of controlling the course of time using force fields.***

Time is among the most common concepts, which are used constantly both in everyday life and in science. This is because all the events and material processes in the world happen in space and develop in time and, hence, the laws that govern space-time connections are the most general and hold for all the forms of matter. Nevertheless, time remains one of the most mysterious concepts of physics; its physical essence is not adequately revealed up till now [1-4]. The concept of time with difficulty yields to logical analysis.

From the point of view of common sense the essence of time is that time characterizes the duration of events and processes, indicates their natural sequence, at which the present, going away to the past, gives place to the future.

Isaac Newton gave a clear-cut characteristic of the concept of time, to which the majority of physicists adheres: "The absolute, true, and mathematical time in itself and by virtue of its nature flows uniformly and regardless to any other object". Though, according to Newton, time flows equally and uniformly and does not depend on the processes, occurring in the world, the daily experience speaks in favour of the fact that the course of time is not uniform. Depending on circumstances in our history, it seems to us that time either flies swiftly or hangs heavy on our hands; sometimes it even changes suddenly, by leaps. In connection with these speculations the question arises of whether the subjective sensations of non-uniformity in the course of time familiar to everyone have an objective basis.

In Newtonian mechanics time is of an absolute character, it does not change as one passes from one inertial reference frame to another and represents merely a parameter, whose change at the will of explorer results in the change of state of a mechanical system in accordance with the equation of motion.

In relativistic mechanics time remains a parameter describing the development of system. But now time and space are intimately linked with each other to form a single system, i.e. the 4-dimensional space-time. In going from one inertial frame of reference to another time gets entangled with spatial coordinates, so that time in one reference frame represents a "mixture" of time and coordinates in the other. Time ceases to be universal, the same in all inertial reference frames; it takes on a relative character.

The indissoluble association of time and space takes on special importance in the light of the concept of physical field, which was called by Einstein the most important discovery in physics after Newton. According to this concept, the occurrence in space of a force field means that space turns into a physical environment, which is capable to interact directly with other bodies and gains, thus, physical properties, becoming an active participant of physical processes. In view of the fact that **space and time are indissolubly related to each other, the presence of a force field in some area of space must necessarily result in the appearance of physical properties of time caused by the motion of body in this area.**

Thus, from the synthesis of the notion of space-time and of the idea of physical field it follows with necessity that **the course of time in a given region of space should depend on physical processes in this region, i.e. time, as well as space, should have physical properties [5-8].**

It should be emphasized that in special theory of relativity (STR) time and spatial coordinates are independent and formally equal in rights quantities, which determine the position of elementary events in space-time. On the other hand, time stands out in relation to spatial coordinates. The special role of time is due, from the viewpoint of geometry, to the pseudoeuclidity of geometry of the 4-dimensional space. From the physical point of view, it is associated with the dynamical principle (causality principle), according to which the state of motion of a physical system at an instant of time t uniquely defines its behaviour at the next instant of time $t+0$. The significance of dynamical principle lies in the fact that it relates the temporal evolution of system to the physical processes caused by force fields and in doing so it allows one to determine the course of time in the system, its possible dependence upon the character of physical processes, and not just the sequence of events and their duration.

The idea about the existence of the physical properties of time belongs to N. Kozyrev [9]. By introducing into mechanics an additional parameter taking into account the directivity of the course of time, Kozyrev has formulated causal (asymmetrical) mechanics from which it follows that time has physical properties. According to the results of theoretical and experimental investigations conducted by Kozyrev and his followers [9-13], events can proceed not only in time, but also with the help of time, **information being transmitted not through force fields, but via a temporal channel, and the transfer of information happens instantaneously. The appearance of additional forces, associated with the physical properties of time and capable to fulfill work, testifies that time can serve as a power source.**

In the papers by I. Eganova [12] and M. Lavrent'ev and I. Eganova [13] it is stated the problem of direct experimental research of the physical properties of time to ascertain the relations of a new type between phenomena and to discover new methods to change the state of substance. In [14] O. Jefimenko investigated the dynamical effect of the slowing-down of time.

According to [6-8], the conclusion that physical properties of time exist follows strictly from relativistic mechanics, without introducing any additional hypotheses. The physical properties of time are of purely dynamical nature: their existence results from dynamical principle. **The availability of physical properties of time is manifested in that time has a local inhomogeneity: its course along the trajectory of motion of a point particle in a force field is continuously changed, and this change in the course of time is a result of the action upon the particle of a force field in the inertial reference frame, in which the motion is considered.**

Editor's note: The author gives a detailed consideration of the physical content of the local dynamical inhomogeneity of time. Considering the motion of a classical point particle under the action of a force field in the inertial reference frames, that moves relatively to each other, Valentin P. Oleinik derives 73 equations, that help him to obtain the following conclusion (mathematical details and physical reasoning may be found in http://temporology.bio.msu.ru/OLEINIK/oleinik.htm.*)*

The elucidation of the physical nature of time is one of the most important problems of theoretical physics. The purpose of research on the problem of time is to study the physical properties of time, i.e. to ascertain the possible interrelation between time and material processes. In particular, it is of interest to find out

- whether the flow of time depends upon physical processes and whether the back influence exists (i.e influence of the change of the time course on physical processes);
- what mechanisms of the change of the course of time are available;
- what factors are capable to speed up or to slow down the flow of time.

In papers [5-8] on the basis of Lorentz transformations relating to coordinates of points, lying on the trajectory of motion of particle in a force field, the phenomenon of local dynamical inhomogeneity of time is predicted. The main result consists in the proof that material processes occurring in a physical system under the action of a force field necessarily influence the course of time along the trajectory of motion of particle. The case in point is the change of the course of time along particle's trajectory in one inertial reference frame as compared with that in the other.

In this paper the next step is made: the relationship is obtained which relates the course of time on one path section of a particle when moving in a force field to that on the other path section in the same inertial reference frame. The main idea underlying the approach developed results from the analysis of Lorentz transformations and consists in that **the course of time of a particle moving by inertia, i.e. of a particle being not exposed to force, should be uniform.**

As is well known [17,18], the existence of dependence of the course of time upon the gravitational field potential is predicted with the general theory of relativity (GTR). According to GTR ([17], p.303), time flows differently at the different points of space in one and the same reference frame. Since "gravitational field is nothing more nor less than a change of the space-time metric" ([17], p.313), one can assert, apparently, that the change in the course of time is due, in the view of GTR, to the change of the 4-space metric. It should be emphasized that in the present paper gravitational field is considered as an ordinary force field, and the particle motion is supposed to occur in pseudo-Euclidian space-time. The main formulas of the article, (22) and (25), describe the change in the course of time in an arbitrary force field at different spatial points in one and the same inertial reference frame. As is seen from the results received, the change in the course of time in a force field is in no way connected with the change of space-time metric. It is conditioned by the force field action on particle in inertial reference frame and is a direct consequence of the dynamical principle underlying relativistic mechanics.

It should be emphasized that the existence of dependence of the course of time on the state of motion of particle in a force field points to the feasibility of controlling the course of time using force fields.

Note an important peculiarity of the non-inertial reference frame, in which the imponderability state of a particle is attained: **there is such a space-time region in which the reference frame at hand can be approximately considered as inertial.** In connection with the fact that such reference frames (it is natural to call them **quasiinertial** in contradistinction to the **true inertial** reference frames) are, generally speaking, not equivalent to each other (see previous section), the derivation of a rigorous criterion for inertial reference frame acquires especial significance. The dynamical criteria for defining the inertial and non-inertial states are considered in the papers by B. I. Peschevitsky [19]. The heliocentric reference frame seems to be among the quasiinertial reference frames, being inertial with adequate accuracy only in a restricted region of space (for example, within the limits of our Galaxy) [16].

The author is grateful to Yu.D. Arepjev for his interest to the paper and stimulating discussions.

References

1. Blokhintsev D.I. Space and Time in the Microworld (Nauka, M., 1970) (in Russian).
2. Prigogine I. From Being to Becoming: Time and Complexity in the Physical Sciences (W. H. Freeman & Co., San Francisco, 1980).
3. Prigogine I. and Stengers I. Order Out of Chaos (Bantam Books, New York, 1983).
4. Logunov A.A. Lectures on the Theory of Relativity and Gravitation: Modern Analysis of the Problem (Nauka, M., 1987) (in Russian).
5. Oleinik V.P. The Newest Development of Quantum Electrodynamics: Self-Organizing Electron, Superluminal Signals, Dynamical Inhomogeneity of Time, Physical Vacuum and Nature, 4, 3-17 (2000).
6. Oleinik V.P., Borimsky Ju.C., Arepjev Ju.D. New Ideas in Electrodynamics: Physical Properties of Time. Semiconductor Physics, Quantum Electronics & Optoelectronics, 3, #4, 558-565 (2000). E-print: quant-ph/0010027.
7. Oleinik V.P. Superluminal Signals, Physical Properties of Time, and Principle of Self-Organization, Physics of Consciousness and Life, Cosmology and Astrophysics, 1, 68-76, (2001); Borimsky Yu.C., Oleinik V.P. The Course of Time in Classical and Quantum Systems and Dynamical Principle, Physical Vacuum and Nature, 6, (2001) (in print).
8. Oleinik V.P. The Problem of Electron and Superluminal Signals. (Contemporary Fundamental Physics) (Nova Science Publishers, Inc., Huntington, New York, 2001).
9. Kozyrev N.A. Selected Transactions (Leningrad University Press, Leningrad, 1991) (in Russian).
10. Lavrent'ev M.M., Eganova I.A., Medvedev V.G., Olejnik V.K., and Fominykh S.F. On Scanning of Celestial Sphere by Kozyrev's Sensor, Doklady AN SSSR, 323(4), 649-652 (1992) (in Russian).
11. Akimov A.E., Kovaltchuk G.U., Medvedev B.P., Oleinik V.K., Pugatch A.F. Preliminary Results of As-tronomic Observations of Sky by the Kozyrev Procedure. Academy of Sciences of Ukraine, Central Astronomic Observatory. Preprint CAO-92-5P, 1992, p.16 (in Russian.).
12. Eganova I.A. The World of Events Reality: Instantaneous Action as a Connection of Events through Time, Instantaneous Action-at-a-Distance in Modern Physics (Nova Science Publishers, Inc., New York, 1999).
13. Lavrent'ev M.M. and Eganova I.A. Physical Phenomena Predicted and Revealed by N.A.Kozyrev, in the Light of Adequacy of Space-Time to Physical Reality, Phylosophy of Science, 1(3), 34-43 (1997) (in Russian).
14. Jefimenko O.D. Electromagnetic Retardation and Theory of Relativity (Electret Scientific Company, Star City, 1997).
15. Khaikin S.E. Forces of Inertia and Imponderability (Nauka, M., 1967).
16. Sivukhin D.V. The General Course of Physics. Mechanics (Nauka, M., 1979).
17. Landau L.D. and Lifshits E.M. The Classical Theory of Field (Pergamon Press, Oxford. 1985).
18. Pauli W. Theory of Relativity (Nauka, M., 1983).
19. Peschevitsky B.I. Dynamical Criteria for Inertial and Noninertial States, p.1 and 2. Phylosophy of Science, 1(7), 79-84 (2000); 1(9), 99-109 (2001) (in Russian).

EAGLE-RESEARCH

Energy Solutions
since 1984

A Research Organization
that Develops & Distributes
Practical Energy-Saving
Methods & Devices

4 Energy Way, P.O. Box 118
Porthill, Idaho 83853
FAX: 250/492-7480

Technical questions to:
wiseman@eagle-research.com

Raum&Zeit

EhlersVerlag GmbH, Geltinger Str.
14e, 82515 Wolfratshausen
Telefon: 08171/41 84-76, 08171/41 84-66
Internet: http://www.raum-und-zeit.com

Institut für Raum-Energie-Forschung (IREF)
i.m. Leonard Euler

Global Scaling Theory
Gravitation theory
Mathematic Biology
Gravitation energetics
Hydrogen-Helium Energetic
Information technologies
Telecommunication technologies
Transport technologies

Time and its Physical Relationships

Andrew Michrowski

President, The Planetary Association for Clean Energy, Inc., 100
Bronson Avenue, # 1001, Ottawa, Ontario K1R 6G8 Canada.
(613) 236-6265; fax: (613) 235-5876
pacenet@canada.com

> *Time cannot be absolutely defined, and there is no inseparable relation between time and signal velocity.*
> *Albert Einstein [1]*

Time does not exist by itself

The phenomenon of time emerges in relationships – as an expression of properties of physical bodies and changes that occur to them.

Time is a factor of energy. Time has to do with the *increase* and *decrease* of energy. For example, as energy is brought down to a "zero level", time is "eliminated", so in a sense, time cannot be "compressed" - only eliminated. In the zero-energy level, electrons occupying this level in unlimited numbers are available through state transitions for the building of matter and the vacuum [2]. So it is the extent and the nature of energy flow that determines the characteristics of time.

How do we know all this?

In his seminal foundations of physics work of the early 1900s, **Sir Edmund T. Whittaker** demonstrated by canonical quantization that there exist physical, time-like and longitudinal photons *in vacuo* [3,4,5]. The 2 scalar potential functions, F and G which completely characterize an electrodynamic field due to electrons in the ether are:

$$F(x, y, z, t) = \sum \frac{e}{4\pi} \sinh^{-1} \frac{\bar{z}'-z}{\left\{(\bar{x}'-x)^2 + (\bar{y}'-y)^2\right\}^{1/2}}$$

$$G(x, y, z, t) = \sum \frac{e}{4\pi} \tan^{-1} \frac{\bar{y}'-y}{\bar{x}'-x}$$

In these equations, for the fundamental case in which a field is due to any number of electrons moving in any way, we observe that time emerges only through the displacement of energy. Another way of putting it would be that time emerges through change in energy.

These photons have an independent physical existence. Whittaker himself observed, after computation that the "total disturbance at any point, due to this system of waves, is independent of the time, and is everywhere proportional to the gravitational potential due to the particle at the point" [6]. **A. D. Sakarov** admitted that the gravitational field is a "conglomerate of loose things and not a fundamental field of nature at all [7].

Everything electromagnetic, and probably gravitational, starts from these potentials, not fields, and under certain circumstances, there may exist photons without fields being present at all. In the *vacuo*, the longitudinal light photon travels in the direction of the beam, like an energy capsule, as a scalar four-potential-function energy standing-wave field, with many different frequencies, with an internal symmetry based on circular polarization [8,9], an energy field or nexus that "has an end, but no beginning". The time-like and space-like parts of the four potential are photons with spin –1, 0 and +1 that are longitudinally directed, and which are observed in the ***Compton*** and the photoelectric effect [10].

Movement of light affects time

Philip S. Callahan designed an elegant experiment that shows how variations in the movement of light affect time [11]. Changes in exposure settings of photographs of same objects, including coherent light laser spots results in shift of position of images. The more coherent the light, the less apparent is a shift in time. He suggested that time is neither absolute nor independent of photon activity of space.

We can also state, as a corollary, that the *movement* of light generates time. After all, the electrodynamics is associated with photons. This is particularly significant in time engineering, as will be seen later.

Time and life

Now, it is well known that the ability of cells to sense the presence of light is a primary function of life itself. If a cell can sense light, it is alive; if it cannot, it is dead [12]. Callahan was able to observe that time increments were detected on light-detecting surfaces of living organisms such as the cuticles of leaves.

The Russian time researcher, **Nikolay A. Kozyrev** considered that living systems "consume" time for their life-energy [13]. **Velimir Abramovich** suggests that living organisms are naturally-driven "time machines". They each have an inbuilt time that serves as a "code" to structure their own physical totality and to regulate their own functioning. Thus their "local time" acts as a "time operator" ("time condition") frequency [14]. The nature and level of electromagnetic fields in living systems could therefore be considered as indicators of their "energy level" and how it affects and adjusts the inherent rate of time-flow.

Time and consciousness

We append to these statements the notion that, yes, time is "consumed" by living organisms – but only if they are conscious. Essentially, time can be perceived by measurement, which always requires a conscious observer.

When we measure, we observe the differentials of energy flow. It is the measurement of these differentials - as we note from the Whittaker equations of scalar fluxes - that allows us to the measure the equivalence of time. When we measure differentials, we are effectively creating our notion of time.

It is difficult for humans to conceive that time does not exist when humans think about it. **Saint Augustine** speaks for all humans with his observation:

> *While I do not think about time, I know that time exists.*
> *When I begin to think about it, I stop understanding*
> *what it is.*

With such conditioning, humans cannot come to grips with the notion that the Universe has no beginning and no end. It is like the paradigm of absolute nothingness – people cannot deal with the concept of no time because, effectively, consciousness, like life, consumes time. One could advance the notion that thought itself is time in motion. When we think, we are creating time. Life is movement – and the brain registers energy differentials and their associated fluxes of scalar potentials.

Time and causal mechanics

Since time is the result of movement of energy (and accordingly, light), then time is associated with the cause-effect processes linked to a first position of energy to the next. In other words, time is innate to causal mechanics.

Variations in energy flow, logically, lead to variations in the characteristics of time. They also permit variations in causality.

Conversely, changes in time lead to changes in physical structure [15]. This phenomenon was proven by Kozyrev and has been labelled as the ***Kozyrev effect*** [16]. He demonstrated that changes in the course of time affect the performance of electronic components. Figuratively speaking, a transistor is like a corral in which light behaves in a certain way. By changing the way that light behaves, a transistor becomes valueless.

To create an extra-temporal causality (with linkages to another level/dimension) one must change the light movement. Changing the light patterns means changing the behaviour of a wave, and its concordant photonic reality. By altering the photons, we make them go to another dimension.

Changes in time can effect change in living systems [17]. The engineering of "time-polarized" waves has the promise to target and amplify natural healing procession in living organisms. It is these natural activities that would restore diseased cells to their original and healthy condition [18]. Such time engineering may also include the reversal of the effects of AIDS, smallpox, anthrax, and most bio warfare agents, with treatment times of only a few minutes per week, with no more than 3 treatments required in all. Eventually it may also be possible to reverse the effects of genetic disorders, effect limb regeneration, and cure spinal cord injuries [19].

Engineering causality

Variation in the density of energy determines variation in the "the course of time". By the "speed of course of time" is meant the rate of causal transformation and the input of additional forces into systems (including mechanical tensions). According to Russian scientists [20, 21,22] there are interesting possibilities in deploying engineering for causal mechanics.

In preventive time engineering, one could delay the approach of a known cause and to artificially close the consequence loop, and thereby annulling it from ever achieving an effect. In other words, one can make the effect "happen" before its normal time, disrupting space structure with its related "speed of the course of time". One can also make it "happen" after its expected deadline. This technology could have interesting implication in strategic situations (preventing a extra-planetary body from attaining a collision course).

According to the Russian experience, when the spatial structure is disrupted by time-engineered causal mechanics, the affected region undergoes relative greater entropy (or, less order). The volume of space is forced out to somewhere else, generating torsion fields, much like a balloon will drift away from denser air. A similar phenomenon occurs when the velocity of a mass increases, the force emerges against it, called "inertia" [23]. Now, to increase velocity, the energy of the mass increases. In the case of time, as the movement of energy increases, interaction or reaction with another dimension ensues to compensate for that primary change [24].

The boundary layer between the two states of space can then act as a mirror and the approaching agent may be reflected back to its source. For example, a light beam may be reflected back to its emitter, in full, or in part, depending on the engineering.

With space-time engineering, we could develop teleportation systems [25]. The course of time goes from the past to the future, in the direction of greater disorder. Going back to the past represents deceleration. The etheric continuum is perceived by Kozyrev is containing variations of "density" of structural elements. The "denser" the etheric region, the slower is the course of time. A zone of accelerated time course would be forced out into "rarefied" ether. A zone of slowed time would be forced into denser ether, hence the basis for a teleportation technique.

In vivo experiments with a Time Machine

For over 15 years, a Russian association of scientists has conducted experiments with acceleration and deceleration of time with 4 prototypes of time machines

[26]. Light-heartedly, the units are called "muskrat traps" since the experiments conducted in remote forest were disguised as a high-tech electromagnetic technology for trapping muskrats. The time machine units are spheres ranging in diameters of 30cm, 1 meter and 2.1-meter. The shells are encased with coils designed to produce convergent waves.

Team leader **Vadim A. Chernobrov** describes converging electromagnetic waves as moving from a periphery to a central point. They are observed when a hoop is thrown into the water and inside the hoop the waves converge. If a potential is applied to do work and to initiate the energy differential process, the other reverse direction scalar (the reverse-time energy flow) must react. Thus, compensation of time - in the form of the deceleration or acceleration of the rate of time - can take place [27].

The first trials involved mice, in which most (25 out of 31) died. Eventually, there were successful 2-hours runs of time travel. An experiment with a dog that was clearly frightened also showed no ill effects. This led to experiment with humans, the first being **Ivan Konov** who, on August 26, 2001 decelerated into the past by 3% of planetary time during a half-hour trial. Dozens of others have experienced the phenomenon and report such sensations as: quicker pulse, giddiness, itching skin, body twisting, numbness at extremities and a case of an out-of-body experience. Harmful effects on living systems do not appear to be linked to the change of the rate of time, but rather to the variations of the time rate value among regions of a living organism.

Some individuals reported visual experiences such as "starry sky", "luminous vortices" and colour spots. Individuals observing outside of the time travel machine noted headaches. The most interesting phenomena occurred just before the start-up: significant presence of ozone for several hundreds of meters around the machine, which was located in a forest. Also noted were strange lighting effects in the sky above the apparatus, accompanied by sounds that inexplicably appeared to generated from inside.

Factors affecting the rate of time

The experiments yielded interesting observations: the phenomenon of the rate of change of time varies according to the hour of the day and according to the lunar phase. The rate can be influenced by a variety of external inputs, including mechanical vibration.

The transition into the future differs from than into the past. It is like movement from any point in a tree – where downwards represents past time. There are many paths possible towards the future – upwards, along the branches, but only one towards the past – downwards to the trunk. The return from the past time is possible only if the time traveler does not interfere with occurring events – or the possibility of returning to another branch of the tree. However, a return move from any variant of future time is possible regardless of the behaviour of the traveller.

The Russian time-travel experiments point to a relationship between inertia and time. In changes of rates of time, the region adjacent to the spheres develops a boundary layer effect, appearing as an aura of "white mist". The greater the time differential, the denser is the mist.

A similar phenomenon has been observed, and captured on film, with some experiments conducted by **John Hutchison** involving remote-controlled lifting and disruption of objects [28,29]. It may be that the ***Hutchison effect*** involves causal mechanics.

Time and frequency

Time may be viewed as a process – or "change-of-space" in any direction that does not exist in our dimension [30,31,32]. In physics, new properties are commonly acquired as the result of change in some property: charge, current, induced magnetic field, etc. Here, the new property becomes a new dimension. For the frequency of oscillations, the formula is $f = 3/\lambda$ [l/s], where λ is wavelength in [m]. Here, the velocity of light is equal to 3 (the 10^8 mathematical degree is omitted since it is a question of scale of measurement only).

Therefore, the analogy between our dimension and frequency gives us a new notion, which is curvature ρ: $\rho=3/r$ [l/m] and $f = 3/\lambda$ [l/s]. The 3-dimensional radius is represented by $R=\lambda/\rho= r/3$ [m] and time as a period of oscillation has the relationship: $T = \lambda/f$ [s]. **Time can be considered to be equivalent to the radius of curvature ρ if the linear radius r and the wavelength λ are the same. This is a condition for the spatial resonance effect.** Note that in this analysis, *"m"* and *"s"* are unlike when length is measured in meters in our dimension. However, for a new dimension it is possible to use equal units for (m) of 4th dimension and conventional "second". Furthermore, light (photons) is possible in our dimension only as a process in such a spatial resonator.

Experiencing time dilation

Based on the above discussion, it is proposed that we experience changes in rate of time in our daily life. During sleep, as our energy level decreases, so it can be argued that time decreases and we "go" into another dimension. Aspects of what is observed during the dream-state do not obey the rules of our familiarity of physical existence, and conventional causality - because the rate of time is different. As we start our dreaming, and as we emerge from dreaming, the recalled experiences resemble more our regular experiences. Before and after sleep, our brain frequency tends to resemble the daytime's. In effect, understanding our dream state's sense of time may be a reference key to comprehending the physics and causality mechanics of time engineering.

Time and planetary gravity

The Earth's mass appears to be in continual growth. 250 to 350 million years ago, our globe may have been half size - with all of the continents as one landmass [33]. In ancient sediments, the natural angles of slope in sand beaches greatly exceeded those of today, indicating that gravity has on our planetary surface has increased 8-fold, several times, during the last 1.5 billion years. Yet the planet's average density may not have changed – only the acceleration of free fall [34,35].

There could also be another explanation possible for the change of slope of beaches. The pull of the moon may have been different in the past.

Also, the planetary magnetic fields could have increased over the millions of years, through interactions with the solar flux of hydrogen atoms. The sun is in constant explosion – production of energy differential. Earth and other planets could be responding with harmonics to compensate for the solar activity, leading to an overall increase in magnetic fields. Such a phenomenon may give the illusion that the physical body is growing larger. In other words, with time, the force of gravity would alter planetary mass and energy.

Assuming that the Earth's density has not increased, it is possible account for, mathematically, the relative increase – growth processes - of the nuclei of terrestrial atoms, including the doubling of mass of nucleons and of electrons [36]. Such calculation also accounts for the emission of 2 different photons by hydrogen atoms (also known as the "red shift" described by **E. Hubble**).

The phenomenon can be explained with the time-like and longitudinal photons described by the bi-directional 2 scalar potential functions [37,38]. Could it be suggested that matter is gravity minus time? In other words, gravity is related to accumulation of energy differential (time) in mass.

The value of gravity varies throughout the planet, in part because of the poles and in part due to local density of matter. **Kirill P. Butusov** has noted a correlation between places of civilization and regions of greater gravity. A faster rate of evolution may be associated with such gravity zones. In these zones there would be a greater influx of scalar potentials. Time would be more "authenticated" by these energy flows into conscious beings. Butusov reminds us that time has a "positive energy" and flowing into nuclei of atoms or a "negative energy" flowing out of the nuclei of atoms [39].

The outflow would be representative of gravitational energy. Longitudinal waves are known to be able to enter and to leave nuclei. Such flows of time must come from other dimensions. After all, the surface of any elementary particle separates our dimension from another. This leads to an interesting possibility in which all time-associated processes between dimensions are synchronized.

Time reconsidered

Arguments have been laid that suggest that time is equal to energy differential (including movement of light). Living systems "consume" time as part of their consciousness and measuring processes.

It is possible to engineer causality. An understanding of higher-order electrodynamics is required. Techniques exist for the generation of scalar potentials. Causality technology has many applications. They include: therapeutics, energy generation, consciousness technology, inter-dimensionality, defence systems, teleportation, and of course, "time travel".

Appreciation is expressed for guidance in this discussion by Thomas E. Bearden and Bernard de Montréal.

References

1. Einstein, Albert. How I created the theory of relativity. [Translation by Yoshimara A. Ono.] /Physics Today. Volume 35 (8), August, 1982. p. 45-7. [Lecture given by Einstein in Japan, on learning of his Nobel Prize.]
2. Phipps, Jr. Thomas E. Heretical Verities: mathematical themes in physical description. /Classic Non-Fiction Library, Urbana, 1986. p. 397-405
3. Whittaker, Edmund T. On the partial differential equations of mathematical physics. /Mathematische Annalen. Volume 57, 1903. p. 333-355.
4. Whittaker, Edmund T. On an expression of the electromagnetic field due to electrons by means of two scalar potential functions./ Proceedings of the London Mathematical Society. Volume 1, 1904. p. 362-372.
5. Whittaker, Edmund T. A history of the theories of aether and electricity. Volume 1 and 2. London: Thomas Nelson & Sons, 1953.
6. Whittaker, Edmund T. On the partial differential equations of mathematical physics. /Mathematische Annalen. Volume 57, 1903. p. 354.
7. Sakarov, A. D., Vacuum fluctuations in curved space and the theory of gravitation. /Soviet Physics Doklady. Vol. 12, 1968. p. 1040-1. [English translation.] Original in: Doklady Akad. Nauk. USSR, Vol. 177, 1967. p. 70-1.
8. Evans, Myron W., Physica B. Volume 182, 1992. p. 227
9. Evans, Myron W., and S. Kielich (editors). Modern nonlinear optics. [Special topical issue of Ilya Prigogine and Stuart A. Rice: Advances in Chemical Physics. Wiley, New York 1992, 1993, 1997, and 2001] Volume 85 (2).
10. Institute for Advanced Study (Budapest). On Whittaker's F and G fluxes, Part III: the existence of physical longitudinal and time-like photons. // Higher symmetry electrodynamics: a collection of AIAS papers. [Special Issue. Journal of New Energy.] 2000. p. 7-1 – 7-5.
11. Callahan, Philip S. and Kevin. Sinewaves of spacetime. / Newsletter of the Planetary Association for Clean Energy. Volume 9(2&3), April 1997. p. 8-9.
12. Jibu, et al. Evanesecent (tunnelling) photon and cellular "vision". /Biosystems. Vol. 43, 1997. p. 65-73.
13. Kozyrev, Nikolay. A. Selected works. University of St. Petersburg. 1991. [In Russian]. Also: On the way to understanding the time phenomenon: Construction of time. A. P. Levich, editor. World Scientific Press. 1995.
14. Abramovich, Velimir. Experimenting with Time. /New Energy Technologies. Volume 3, November-December, 2001. p. 13-5
15. Frolov, Alexander V. Practical application of time rate control (TRC) theory. /New Energy Technologies. Volume 3, November-December, 2001. p. 15-18.
16. Kozyrev, Nikolay A. The possibility of experimental study of the properties of time. /National Technical Service, US

Department of Commerce, Springfield, Virginia. (JPRS 45238). [Undated.]

17. Bearden, Thomas E. Renaissance healing, a concept whose time has come. /Proceedings of the United States Psychotronics Association. Volume 9, 2000. p. 1-2.
18. Bearden, Thomas E. Vacuum engines and Prioré's methodology: the true science of energy-medicine. I and II, /Explore! For the Professional. Volumes 6 (1), and (2) 1995.
19. Bearden, Thomas E. Private communication. February 2002.
20. Frolov, Alexander V. Practical application of time rate control (TRC) theory. /New Energy Technologies. Volume 3, November-December, 2001 p 15-8.
21. Kozyrev, Nikolay. A. Selected works. University of St. Petersburg. 1991. [In Russian]. Also: On the way to understanding the time phenomenon: construction of time. A. P. Levich, editor. World Scientific Press. 1995.
22. Kozyrev, Nikolay A. and V. V. Nasonov. On some properties of time found by means of astronomical observations. / Collection: Problems of research of the universe. Volume IX, M-L, 1980.
23. Puthoff, Harold E. The energetic vacuum: implications for energy research. /Speculation in science and technology. Volume 13 (3). p. 247.
24. Frolov, Alexander V. Trilateral spacetime effect. / Newsletter of the Planetary Association for Clean Energy. Volume 9(2&3), April 1997. p. 10-13.
25. Frolov, Alexander V. Practical application of time rate control (TRC) theory. /New Energy Technologies. Volume 3, November-December, 2001 p 15-8.
26. Chernobrov, Vadim. Experiments with a man in the Time Machine. /New Energy Technologies. Volume 3, November-December, 2001 p. 6-8.
27. Frolov, Alexander V. Trilateral spacetime effect. / Newsletter of the Planetary Association for Clean Energy. Volume 9(2&3), April 1997. p. 10-13.
28. Calante, Pelayo and Andrew Michrowski (compilers). The Hutchison File. The Planetary Association for Clean Energy, 1996. p. 16-7.
29. Hathaway, George D. The Hutchison effect – a lift and disruption system. /New Energy Technology. The Planetary Association for Clean Energy, 1990. p. 77-103.
30. Frolov, Alexander V. Trilateral spacetime effect. / Newsletter of the Planetary Association for Clean Energy. Volume 9(2&3), April 1997. p. 10-13.
31. Ouspensky, Pavel D. A new model of the universe. New York, 1971.
32. Frolov, Alexander V. The application of potential energy for the creation of power./ New Energy News. Volume 2(1), May, 1994.
33. Kirilov, I. V. Increase in volume of rocks is one of the reasons of tectonic deformations. /Academy of Science of the USSR Geologist. Volume 1, 1963.
34. Smirnov, L. S. and Ju. N. Lubina. On the possibility of studying the change in gravity together with geological time. /Reports of the USSR Academy of Science. Volume 187, Number 4. 1964.
35. Maxlow, James. Earth expansion – the definitive proof. / Nexus. Volume 8 (3), May 2001. p. 47-51.
36. Butusov, Kirill P. Time is a physical substance. /New Energy Technologies. Volume 3, November-December, 2001. p. 8-13.
37. Bjerknes, K. A. Vorlesungen ueber hydrodinamische fernkraefte. Leipzig, 1900-2.
38. Yarkovsky, I. O. Universal gravitation as a result of formation of ponderable matter inside heavenly bodies. Moscow, 1889.
39. Butusov, Kirill P. Time is a physical substance. /New Energy Technologies. Volume 3, November-December, 2001. p. 8-13.

New Energy Technologies

PRINT EDITION ADVERTIZING RATES

Full Page Ad

B/W - $400

Color - $800

Half Page Ad

B/W - $200

Color - $400

Quarter Page Ad

B/W -$100

Color- $200

Ask about discounts for multiple ads

New Energy Technologies is distributed to scientific laboratories in 20 countries around the world.

To advertise contact us:
Faraday Labs Ltd, Phone/fax 7-812-3806564,
http://www.faraday.ru, office@faraday.ru

Time as Phenomenon of the Expanding Universe

Michael H. Shulman

7 (095) 357-80-33, 7 (095) 421-11-31
E-mail: shulman@dol.ru

Preface

In 1992 the author of this words, having been stimulated by Kozyrev's book [1], started to reflect on Time meaning and Time and Space asymmetric roles during the Universe Expanding. The development of these ideas has led to the work [2], it is available now (the Russian version only) on the website of the Institute of time nature explorations (grant #00-07-90211 of the Russian fund of basic research).

I would like to thank the Chairman of the Russian Interdisciplinary Temporology Seminar Dr. A. Levich from Moscow State University for his constant support and friendly interest. Also I would like to thank A. Moskowsky for the 20-year discussion of the physics history and philosophy.

The said work [2] pretends to revise radically a number of basic physical concepts of the Space - Time, Motion and Energy nature. It includes a detail analysis and mathematical calculations. Only a brief account of the main part of this work is presented below.

Introduction

The nature of time is not yet enough clear for natural science. In Newton mechanics time was presented as some universal formal parameter. Its value rises steadily at every point of the Universe by unknown for us reason. Each physical process occurs in space in correspondence with the time course.

In the Special Relativity (SR) time and space are integrated to the common 4D-continuum. However, in this theory the time component having imaginary factor seems also to be "exotic". In this concept the increase of time is also implied in each reference frame.

The General Relativity (GR) allowed linking the time properties with gravitational fields and the space geometry. The time currency started to be associated with a spatial expansion of the Universe.

The theoretical physics traditional approach to the process description is based upon the considering of time course as primary (original) one. There are also in the modern physics [3] several attempts to deduce the time concept as secondary one from different fundamental (microscopic) concepts. However, the third way (inverse to the first one) is possible and forms the basis of this paper. A starting point of this way is the following question: "Does any universal process exist which could *generate* physical time?"

(Editor's note: The same question was formulated and the answer was proposed by Alexander V. Frolov in 1996, report "Matter as process", Scientific congress "New Ideas in Natural Sciences". It was assumed that similar process can be produced by special technical methods also.)

Such fundamental cosmic process really exists and it is well known in the modern science. **It presents the Universe expansion** and was opened at the first third of the 20th Century by the American astrophysicist E. Hubble and others [4]. It means the general increase of distances between all 3D-bodies. The same scattering of two-dimensional-figures happens on the surface of some spherical balloon during air incoming. The centre of this sphere does not belong to the surface; all points of the sphere (the Universe) are equivalent.

Some time earlier the theoretical physics had come to the same results. As it is well known, the Einstein's GR was published in the 1916. After that Friedmann (1922) proposed the concept of the expanding Universe. For example, in the book [5] a description of the basic cosmic model is given. Hereinafter this model is called "Einstein-Friedman model", or "EF-model". In this model the Universe is presented as 3D- hyper surface of a 4D- sphere with increasing radius. Of course, the curvature of the 3D-hyper surface increases with time too.

Basic hypothesis relative to time nature

Some simple and pictorial views consist a basis of the new concept. Hereinafter it is called briefly as "The Spherical Expanding Universe Theory (SEUT)".

In the SEUT, as well as in the EF-model, in every time the Universe represents the 3D-hyper surface of a 4D-sphere. However, there is one very important difference. In the Einstein's theory the spatial components of the metric tensor are opposite in sign to the time's one. For example, we may consider time as imaginary quantity, then spatial coordinate as a real one. On the contrary, in the SEUT the 4D-continuum is considered as purely

Euclidean, all the four coordinates are real quantities. The usual spherical geometry can be used on a surface of the 4D-sphere.

As it is well known, in the GR the Age of the Universe can be *calculated* using the EF-model or a similar one. Usually, the radius-age dependence is not a direct proportionality in such GR models. However this result may be deduced if to neglect the global pressure of matter that fills the Universe. In [2] it is demonstrated, that an account of the static pressure of matter follows to the Universe radius linear dependence on its age (see below).

On the contrary, in the SEUT the time universal course is *manifested.* The Universe age is *identified* with a current Universe radius divided by the velocity of light. Numerous important consequences may be deduced from this statement. On the other hand, it allows avoiding many other arbitrary postulates in the model.

Mechanical motion and maximum velocity in the SEUT

The SEUT states, there is no unlimited set of independent mechanical motions. Only world lines of moving bodies exist. Each of them has some inclination relative to the time line, which presents a normal to the hypersurface of the 4D-sphere. It is *an inclination angle that defines the spatial motion velocity*. At the increase of sphere radius the intersection point of word line and current hypersurface "moves" with exact correspondance with a modern physics prediction.

In particular, immovable objects (stars) have zero inclination, their world line are normal to the hypersurface. Hence they "scatter" according to the Hubble law, their mutual velocity is proportional to mutual distance. If a body world line has some inclination relative to normal, the angle is more than zero. But it can't exceed 90°, therefore *maximum* mechanical motion *velocity* appears *naturally*, it is equal to the velocity of light.

Let us consider three variants of motion (see Fig.1).

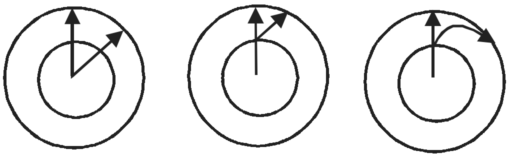

Fig. 1
Phenomenon of a "motion" of point on a sphere surface for immovable (at the left), uniformly moving (at the centre) and non-uniformly moving (at the right)

The left picture illustrates the Hubble effect. The right picture presents a general case of motion with acceleration. The central picture corresponds with an inertial motion; its world line is direct. In this case the moving body displacement increases proportionally to the Universe radius increment. So, *inertial* motion is not postulated in the SEUT, it appears as natural model consequence.

At a large 4-sphere radius values all the relationships of SR and usual mechanics laws are applicable approximately in the SEUT. A Special Relativity light cone transforms to all the hyper surface of the 4D-sphere. But the analogy is not complete, because an absolute remote SR area degenerates to this 3D- hyper surface in the SEUT (see Fig. 2).

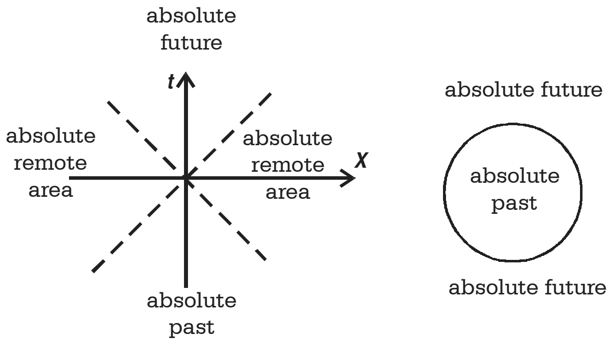

Fig. 2.
4D- continuum areas in the SR (at the left) and in the SEUT (at the right)

SEUT and Minkowski geometry

Let us consider small increments of time and space coordinates along body world lines during the Universe expanding. It is enough to consider a small area of the Universe, so we can neglect its curvature. Then it is acceptably to replace approximately concentric hyper surfaces ("isochrones") by parallel hyper planes. The space-hold corresponds with a representative point "drift" perpendicularly to isochrones, an inertial motion corresponds with displacement along inclined direct lines between isochrones.

At each time the Universe is represented by a certain isochrone that contains all the real spatial points. Let us accept that 4D-sphere radius increment divided by velocity of light presents invariant measure of (absolute) time increment. We will also state that this quantity has the same value in each *inertial* reference frame, i.e. at a motion along each *direct* world line.

Let the angles of world line inclinations from normal direction are enough *small*. Then metric relationships like Minkowski geometry ones appears in our purely Euclidean 4D-continuum. In particular, well known relationship

$$c^2 ds^2 = c^2 dt^2 - dr^2$$

can be deduced from the Pythagorean theorem. It connects a spatial component *dr* with a time component *dt* (at moving reference frame) through velocity of light *c*. Here *ds* is an absolute time interval (between two 4D-events at a immovable reference frame). **Hence, if velocities aren't very high, the Lorentz transform is correct in different inertial reference frames.**

On the Einstein's relativity principle

If the world lines inclinations from exact normal direction cannot be accepted as small, then Minkowski geometry relationships are correct approximately only. It means that Einstein's relativity principle is correct (in the SEUT) only for reference frames that move with velocities enough small relative to *selected* reference frame. Such reference frame is linked hardly with a body at (absolute) rest, i.e. drifting along radial world line.

The selected reference frame existing reminds of old ether theories that contradict to the Special Relativitiy views. It seems, these theories became a thing of the past irretrievably. In fact, the velocity of light in vacuum is constant everywhere and everywhen. However, the reference frame existing can be detected in principle as a light signal frequency bias, i.e. with the help of Doppler effect. Well, this phenomenon is really detected by the modern astrophysics!

The temperature diagram of the Cosmic Microwave Background Radiation (CMBR) coming to the Solar system from all the sides of the Universe is presented on Fig.3. The data was registrated during 4 years by the Cosmic Background Explorer (COBE) satellite (NASA Goddard Flight Centre, COBE Science Working Group).

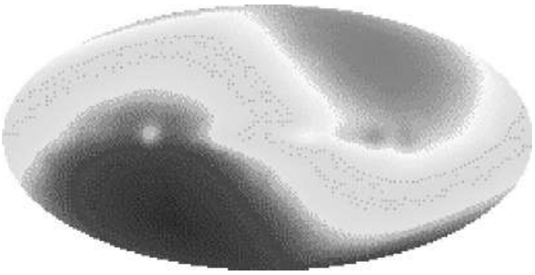

Fig. 3
The anisotropy of the Cosmic Microwave Background Radiation (CMBR)

The well-known Russian scientist J.B. Zeldovitch in the Editorial Addition to [6] in connection with earlier experiments notes that careful measurements **allowed to find out some anisotropy of CMBR.** An antenna oriented to the Lion constellation detects that the radiation temperature is 0.013% more, than mean temperature. The radiation temperature in the opposite direction is 0.013% less, than mean. **Generally, a temperature varies continuously between these two values.** The isotropy presents only for some imaginary observer. ***The Solar system, Earth move to the Lion constellation relative to this observer having velocity 390 ± 60 km/s.*** Hence, as a result of the Doppler effect, a incoming radiation seems to be more hot, and an overtaking radiation seems to be more cold. This example shows that for observer of any point of the Universe this CMBR is isotropic. We may consider this observer and the connected reference frame as selected one. The selected reference frame existence at the Universe every point looks like the physicists commonly held view preceding to Relativity. They thought that the light presented ether oscillations occupying whole the Universe. They thought also that a reference frame connected with ether was preferable, or selected. They tried to detect the Earth motion relative to ether. We know that these experiments gave the negative result: any ether doesn't exist. But the Universe evolution follows that when CMBR is observed (and only in this case!), the selected reference frame (called sometime "new ether") appears. ***The new ether or CMBR just realises the motion according to Hubble's law.***

The modern NASA's data allows to put the relation of the Solar system velocity to the velocity of light equal to 0,15%. It is enough small value justified Special Relativity and Minkowski geometry relationships application. But can we believe this phenomenon to be an exhaustive proof of the SEUT accuracy?

To test it we propose a not complicated observational experiment. If the CMBR anisotropy is due to the real selected frame existence, then it may be detected for any electromagnetic radiation. In particular, an anisotropy of solar radiation has to exist at the different year periods. It has to be detectable in August, when the both solar radiation and CMBR come to the Earth from the Lion costellation side (see Fig. 4). In February these sourses are opposite in disposition relative to the Earth, therefore the solar radiation anisotropy direction has to be opposite. The expected effect value (with account of the Lion constellation straight ascendancy and obliquity of the ecliptic) is approximately equal to 300 km/s, i.e. nearly 0,1% of the velocity of light. In November and May the anysotropy has to be practically absent.

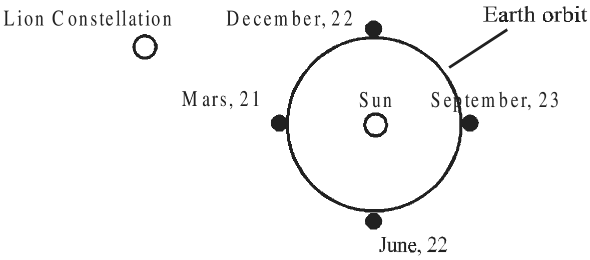

Fig. 4
The Sun and Lion constellation disposition relative to the Earth

Probably, an analogous SEUT test in a laboratry is realizable with help of artificial radiation sources.

Particle mass, energy and impulse

So, we consider the Universe as expanding 3D-hypersurface of a 4D-sphere. Mass localisation places in the Universe present the points of the hypersurface intersection by world lines. So, **these world lines have a real physical meaning, not abstract illustrative** this one. We may expect this physical meaning to be more essential than simple word expression.

Particularly, while the Universe global analysis is making, we may suppose that such fundamental particle feature as its mass at rest presents some *relative* value. Such relation (some kind of a quantum number) may include, for example, 4D-sphere (the Universe) diameter and some characteristic size like de Broglie wave period that is inversely proportional to the mass. This hypothesis may make clear *inertia* nature as two characteristic times *relation*. It may also explain the *rest-energy* notion.

Meanwhile, the Universe radius increases with time. Well, what happens to mass? If a de Broglie wave period increased proportionally to time, we couldn't generally detect the Universe expansion, including famous "red shift". But if particle wave periods are constant, then *matter mass has to rise proportionally to the Universe age and size*.

In the Relativity (like Minkowski geometry) we use vectors having imaginary projection to time axis and real projections to space ones. Particularly, it is true for velocity, acceleration, and energy-impulse 4-vectors. As against, vectors having all the real components are only used in the SEUT. At that, a 4-interval value (length in pseudo-Euclidean space) of some relativistic vector answers the absolute time axis projection of a corresponding SEUT-vector, and imaginary component of a relativistic vector (time of motion) answers the corresponding vector length in purely Euclidean continuum of the SEUT. For example, the energy-impulse vector module presents such quantity. Its projection to absolute time axis is energy at rest divided by velocity of light, and its projections to spatial axis are impulse components. This quantity is *constant* while the particle movement is *inertial* one.

A jump to *non-inertial* motion in the SEUT is connected with a corresponding state vector *changement* law. So, if particle motion velocity changes, its *energy at rest* doesn't change, therefore full acceleration at a time interval can be calculated using the difference between new and old impulse values. Thus, the non-uniform motion equation in the SEUT can be found like SR as time derivative of an impulse expression.

The force-acceleration relation depends on a mutual orientation of force and velocity vectors in the both SR and SEUT. But in Relativity a reference frame velocity can be choised arbitrary, for example it can be zero, then the relation will be equal to one.

On the contrary, in the SEUT an *absolute* velocity is presented, it is defined by the world line inclination relative to the normal. Let the Earth move with any velocity relative to the immovable (selected) reference frame. Then we will be able to detect the absolute velocity using two measurements, the first one along the world line, and the second one in a perpendicular direction.

If this absolute velocity is really defined by direction and value following from CMBR anisotropy effect, then we can expect a relative difference near $2{,}25 \cdot 10^{-6}$ between longitudinal acceleration and transversal one.

Local gravitational fields of particles

What does the SEUT talk about body gravitational fields? Let us imagine all the bodies as immovable and drifting exactly along the *radial* world lines. If there is a mutual gravity attraction effect between two bodies in such Universe, an observer will detect some curvature of their world lines. They will seem to be bending one to another instead of a radial divergence. In essence, in this case we may replace a world line by a gravitational field line. Then the analogy allows us to identify an Universe isochronous intersection with an equal potential surface that these field lines have to be *normal* to this surface. So, we arrive to a presentation that a Universe isochronous intersection is not strictly concentric hypersurface. It is perturbed by some kind of craters (see Fig.5), that centres correspond with gravitating bodies.

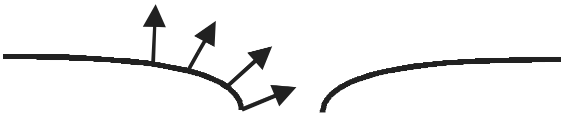

Fig. 5
Local body gravitation field

The inclination angle of a crater profile relative to non-perturbed sphere hypersurface is equal exactly to the inclination angle of a normal relative to strict radial direction. Hence, a local gravitation field intencity measure agrees practically in each point with body velocity measure that we used earlier. It authorizes energy concept using for both mechanical motion and gravitation phenomenon.

SEUT and General Relativity

Let us discuss some GR's aspects. Is it acceptable to neglect pressure of matter? When Einstein searched for his early cosmological static model solution, he had to introduce a cosmic constant in his equation. This constant answered a ***negative matter pressure***, that Einstein could not determine a meaning. In a non-stationary model a solution exists independently on cosmic constant presence, therefore it may be put often as zero. As rule, bodies' velocities may be put as zero too; therefore (*dynamic*) pressure is usually neglected.

However, we insist on necessity to account a *static* pressure of gravitating matter. Really, it can be ignored in the case when Einstein's relativity principle is applicable. Accordingly with it a gravitation field can be *always* replaced by reference accelerated frame. In this case a purely kinematical side is only accounted. However, *not every* field may be considered (even locally) as uniform one (see Fig. 6). Let the radius of a field source (or a *probe particle*) have the same order that the mutual distance. Then the Einstein's equation connecting space geometry with matter physical features seems to be incomplete. More precisely, it is incorrect to put exactly equal to zero a static pressure in the matter density tensor, it is necessary to introduce its (unknown, calculable) value accounting **the *continuum deformation energy*.**

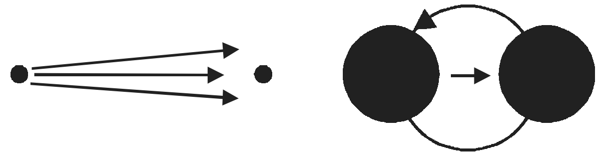

Fig. 6
At the left a locally uniform gravitation field is presented, the right field can't be considered as uniform one even locally

Are the proposed amendments important? Specialists know, that the Metagalaxy gravitation radius is *in fact* comparable with its real size. **In [2] we have showed that the Universe radius was *less* than its gravitation radius.**

It is shown also in the same work [2], that a resulting gravitational pressure in the Universe is *negative* and it is responsible for the Einstein's cosmic constant. This fact seems to be evident because a gravitation force aims to *gripe* any matter accumulation. **But it is possible to conclude it by a non-trivial way.** Let us consider a matter pressure as a reactive force operating within a uniform sphere of incompressible liquid. A pressure dependence on internal density is presented in [7]. We can see [2], **that if the sphere gravitation radius exceeds its geometric radius not more than approximately 1% (or it is still less), then a sudden negative change of the pressure appears at the internal abroad. This phenomenon may be explained as a volume "expansion" due to a metrics perturbation.**

The consideration of static pressure allows not only to find out a *new* (linear in time) cosmologic solution, but also to calculate a dependence of the Universe gravitational pressure on the radius of the Universe. This *negative* value has the representation like that one for a non-relativistic sphere (a star or a planet, for example). Also we would like to note, the formulation of density-Universe radius in the SEUT is exactly the same that the formulation of so-called *critical* density in the EF-model.

Two very important circumstances are clarified for all that. First, the **Universe mass was turned out as *linearly increasing* function of the 4D-sphere radius, and it is not a constant. A famous Einstein's programme is realized unexpectedly in the SEUT: a matter features (density) are reduced to a space features (curvature).** In other words, a necessity to introduce a mass distribution in the equations externally ("by hands") is eliminated in the SEUT. This operation is need in the GR to find out a spatial metrics changement law.

Secondly, the seeming paradoxicality of University mass (and energy!) non-conservation make us to reflect on conditions that the *accomplishment* allows to the energy conservation law correctness (see Fig.7).

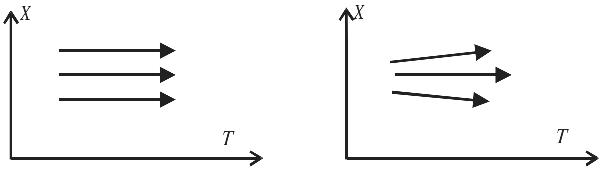

Fig. 7
If space features *do not depend* on time, then the closed system energy is *constant* (at the left). But if space features *evolve* in time, then the isolated system energy *have not to be constant*.

It is evident, we believe, that the energy can be exactly constant only in such physical system (or in whole the Universe) for which space features (in particular, a curvature) are strictly constant in time also. However, the both modern physics and SEUT issue from the opposite concept. Hence, this state can only be accomplished approximately, moderately of a bit of the modern rate of a relative space curvature evolution. This rate has order 10^{-10} per year for the modern Universe.

N.A.Kozyrev [1] basing on astrophysical observations stated the common star radiation origin that has to be due to the time-energy transformation. Accordinly with the SEUT, the relative increment of star mass (and its rest energy) is equal to the Universe age relative increment. It is interesting that the Sun mass relative lost due to the radiation consists in 10^{-15} per year, i.e. five orders less than mentioned above energy increment.

Universe origin and closed geometry

The cosmological EF-model could not say anything on the Universe origin. On the contrary, the work [2] approach allows an obvious way to study the problem.

As it is noted in [7], the metrics of any sphere area having a non-zero density is perturbed relative to Euclidean one, its geometry agrees with 4D-sphere hypersurface geometry. For non-collapsing sphere its gravitation potential relief is like a very small *"pit"* that gravitation radius is much less than its geometric size. However, when the density rises, the metrics perturbs more and more, and the pit transforms to some kind of *"crater"*. The crater is connected with the external surface by a narrow neck. Only this neck or its part is visible for an external observer, and the gravitation insurmountable barrier transmutes an object central area into "a lost world".

From the point of view of the external world, the central area presents a "black hole" absorbing irreversibly all the matter and radiation. On the other hand, for our Universe inhabitant the "navel-string" connecting with the external world has to seem a spherical "white hole", to which a matter and radiation are coming continuously (and, may be, carry out an information on the external world features). There is an old Russian fiction science book called "Sannikov Land", where an internal gigantic trench concerning central small star is described. Our model seems to be like this picture.

Is it possible that we live in such a black hole? The present hypothesis answers affirmatively this question. The negative sign of the matter pressure due to continuously increasing of our world size allows to such conclusion. And the University self-closing can be physically explained by the way.

On the other hand, as it is shown in [2], in spite of matter average density negligibly small, the Universe gravitation radius is more than its geometrical radius; hence, it presents a black hole. This condition accomplishment agrees the Universe self-closing, the boundary absence in spite of its finite volume. Also, this fact confirms our assumption that the Universe rest energy localisation area does not exceed the gravity operation area.

A specific model of star collapse is created in the modern General Relativity. In general, it can study in *three* different reference frames. As a rule, the "point mass" models are used. *The first* model is linked with an external observer; the *second* one accompanies a matter falling to the black hole. *The third* model presents an internal reference frame, i.e. an observer within collapsing object.

From an external observer's point of view the matter falling time to the black hole is infinitely large. However, in an accompanying reference frame it is finite. Since in the accompanying reference frame time and space coordinates are expressed through the both types of external reference frame coordinates. What is more, in the internal reference frame the time and space coordinates quite trade places, the metrics tensor components are depending on time. Further, any matter point history in this accompanying reference frame starts at the zero moment and finishes after same universal time period in a special (singular) point, after which nothing exists ("time barrier").

As we believe, another lacing between internal and external collapse pictures will be possible, if we consider a non-point collapsing object. Nobody wonders now at a situation, when a time period can be finite in one reference frame and infinite in another one. **Therefore, we can believe, unlimited black hole *collapse* in the external Super-Universe may seem to present unlimited *expansion* of our Universe observing inside.** This expansion seems to *start* from a singular point, and the same point is the history end of all the matter of the external Super-Universe that fall to the black hole. I would like to note especially, it does not mean that internal time pass in opposite to external one. **Rather, it is possible to state, time within a black hole passes ortogonally to external one.**

References

1. Kozyrev N.A. Selected works. - Leningrad University, 1991.
2. Shulman N.H. Site of the Institute of time nature exploration (Russian-language version): www.chronos.msu.ru , Project of Chair-laboratory of time as phenomenon of expansion, 2001.
3. Vladimirov Yu. S. Relational theory of space-time and interactions. Part 2. Theory of physical ineractions. Moscow university, 1998.
4. Sharov A.S., Novikov I.D. A man, who discovered the Univerce explosion. Edvin Hubbl's life and work. Moscow, Science, 1989.
5. Einstein A. The essence of Relativity. Moscow, IL, 1955.
6. Wineberg S. First three minutes. The modern approach to the Univerce origin. Moscow, Energoizdat, 1981.
7. Tholmen R. Relativity, thermodynamics and cosmology. Moscow, Science, 1974.

INFINITE ENERGY
Cold Fusion and New Energy Technology
Including Conventional and Renewable Energy

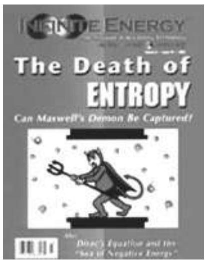

Infinite Energy Magazine
❋ Cold Fusion ❋ New Energy
❋ New Science
❋ New Technology

Subscriptions, 6 Issues Per Year
$29.95 North America
$49.95 Foreign

Single Sample Copy
$5.95 North America, $10.00 Foreign

Infinite Energy Magazine
P.O. Box 2816-FV
Concord, NH 03302-2816
Phone: 603-228-4516 Fax: 603-224-5975
http://www.infinite-energy.com

ELECTRIFYING TIMES

an online and published magazine about Electric, Hybrid, Fuel Cell Vehicles, advanced batteries, ultra capacitors, fuel cells, microturbines, free energy systems, events and exhibitions worldwide

63600 Deschutes Mkt Rd, Bend Oregon, 97701
541-388-1908, fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

Institute for Planetary Synthesis

P.O. Box 128, CH-1211 Geneva 20,
Switzerland
Tel. 41-022-733.88.76, Fax 41-022-733.66.49
E-mail: ipsbox@ipsgeneva.com

website: http://www.ipsgeneva.com

The Experiments on Thermo-Gravitation

This review of the article by Alexander P. Schegolev, the scientist from Saint Petersburg, was received by NET editors from our reader vladrim@mail.ru and presents his own view about Schegolev's experiment [1]. The opinion of our editorial board is not fully in agreement with the author's conclusion. Also we have no information who is vladrim@mail.ru.

It is known from the science history that the very experiment gives a push to revising of old knowledge; it also checks the new designs and conclusions. The knowledge of physical theories is based on the experiment; the experiment confirms hypothesizes or refuses them. Making experiments, we ask questions to the nature. And it always answers us on the clear pointed question. However…

Michael Faraday was trying in vain to find out the relationship between gravitation and electricity. These experiments were repeated on the modern level. Further, there was made an attempt to screen gravity force, the influence of environment on gravity was also trying to be researched. Many efforts were spent on finding out of the gravitational radiation falling to the Earth from depths of the Universe, as well as artificial generation of gravitational waves.

The experimental studies of weight change must be under attention. In particular, the attempt to find the differences in weight after sharp turning of body from the quiescent state into the state of rotation around its own axis was undertaken. The experiments of the influence of strong magnetic field and temperature on the weight of a body were made.

But alas! No changes of weight were discovered in previous experiment *(Editorial: Perhaps, the author has no information about sucessful results. We reported about such results in our magazine)*. Analyzing these experiments and theoretical premises, the author of the given article has made an unusual experiment imitating the heat motion in the Earth from core to peripheries. For that experiment there was used the steel ball with 100 mm diameter. The cone hole was made in the ball up to its center. The ball was installed by the hole upwards on laboratory scales with 50 mg scale factor and laser beam was directed inside (into the hole). The directed heat flow outgoing from the center of the ball was created by this way. While the surface temperature increasing, the arrow of scales became to be rejected aside weight decrease. The temperature was measured by the contact thermocouple. After an hour and half, approximately, when the temperature was 300 degrees Celsius the laser was switched off. The difference (the decrease) in weight against initial value (in the cold state) was four grams per 4200 grams of the test body.

Further, the ball was getting cold slowly on the scales and its arrow was creeping to the initial position. For acceleration of this process, the ball was periodically blown on by pressure air. It was the check, if the air (heated by the ball) influences on scales or not. The stability of its data independently on actions of pressure air stream excludes such suggestion. When the temperature of the ball became the same with room temperature, its weight return to the value nearly initial. However it is necessary to add, that this "nearly" is about 200 mg. This confirms that the change of the weight was not occurred because of the evaporations of metal from action of laser beam, and we were the witnesses of absolutely unusual phenomenon. *(Editorial: Perhaps, it is necessary to take into account the deformation).*

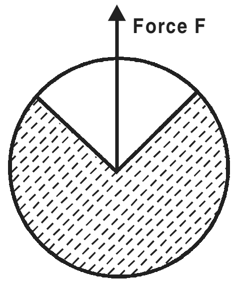

Fig. 1
The ball with cone hollow

That is all concerning the question, which was given to nature by the author of the given article. However, one more doubt remains: did the heat radiation (coming from the ball) affect on mechanism of scales? To check this, the control experiment was made, under which the same ball was heated by usual way in the electrical furnace. In spite of its temperature was about 600 degrees Celsius and the ball was placed on scales for a long time (until full cooling down), the arrow did not move from the initial position. Thus, we have really turned out face to face with a phenomenon requiring an explanation.

If the weight change in this experiment has happened in consequence of gravitational interaction, then, therefore we must revise some fundamental concepts. Today it is difficult even to assume the consequences. The only thing, which is possible to say certainly, is **that divergent on radius uni-directional (single-vector) heat flow** is the object for observation, absolutely unknown to us or unnoticed before, which holds the ensemble of surprises. Up to now, we have dealt only with **chaotic** heat motion, which was researched by the thermodynamics and heat transmission.

For analogy, it is possible to give the example **of the electric current arising, which is possible to get only under directed electrons moving.** In general, more than ten experiments were made, and all of them have given the same result: the weight of the body was decreased. Who will solve this enigma?

Editor's note: The question is presented above. The answer is the aetherodynamics theory. Any directed (unidirectional) heat flow is also a flow of aether. The mass changes are the demonstration of natural mechanism of existence of this mass as aether vortex (Alexander V. Frolov).

References

1. Collected reports "Untraditional scientific ideas about nature and its phenomena", #1, club FENID, Gomel, 1991.

UP-TO-DATE INVENTION

The RotoVerter from arkresearch

Editor's: The description of this invention was received from Hector D. Perez Torrez, designer of the "RV" Rotoverter and TV transverter OU "transformation" devices.

Here is the first prototype, used to test a 3PH generator. Light bulb was a 100W 120V, AC generator required 380W for excitation to 120VAC plus the light bulb load for a total of 480W. Input to rotoverter was 240W for a total of 200% efficiency from the prime mover. Friction loss and prime mover loss was not calculated, PF unity, 120VAC unto a purely resistive load excitation pure DC120VDC3.1666667A (380W) + 100W = 480W. -240W=240 excess.

It is a common 3PH dual winding 230/460V motor wired to 480V run at 120v as a ROTOPHASE converter. That's the beauty of it, OFF the shelve components preferably totally enclosed motors (so fan can be removed) low friction bearings, from 3HP to 7.5HP. Re-rating hp can be from 1/7 to 1/3 the original HP (some motors can run up to 1/2 HP capacity but will require extreme care and cooling.... and exact tuning to load).

System requires starting and running capacitance; the perfect starting capacitance is the one that gives same plate current value as 480V but running at 120V. The best running capacitance is the one that gives the lowest Amperage reading at programmed load. (ALL capacitors are AC 370V OIL "no electrolytic").

Some motors are better than others; also for generator use you must choose the over-unity generator and motor combination to obtain OU. It looks easy to gain it, but requires quite a good choice of items and common sense. In this Research we have found even a few turns or a run weld in stator can change drastically the results. Up to date best performance has being obtained from US (motors 7.5 HP, BALDOR 3,5, HP, GE 3HP totally enclosed, dual winding 230/460V). Higher HP will require professional expertise in industrial electricity and power eng. (not recommended for amateur experimenter). Anyone engaged in this experiments must be familiar with safety procedures and basic electric and mechanic knowledge. Voltages and rotating machinery must be handled with great care. SEE basic diagram and connection.

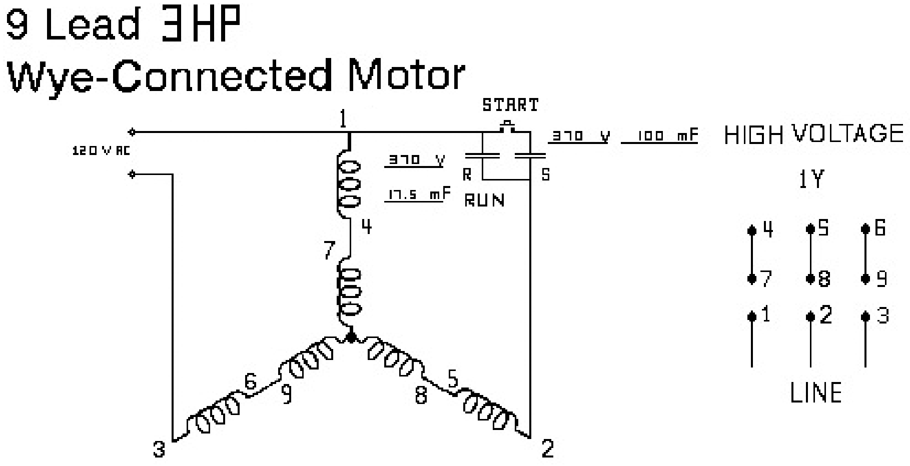

Fig. 1

Latest UPDATE BELOW 7-01-02

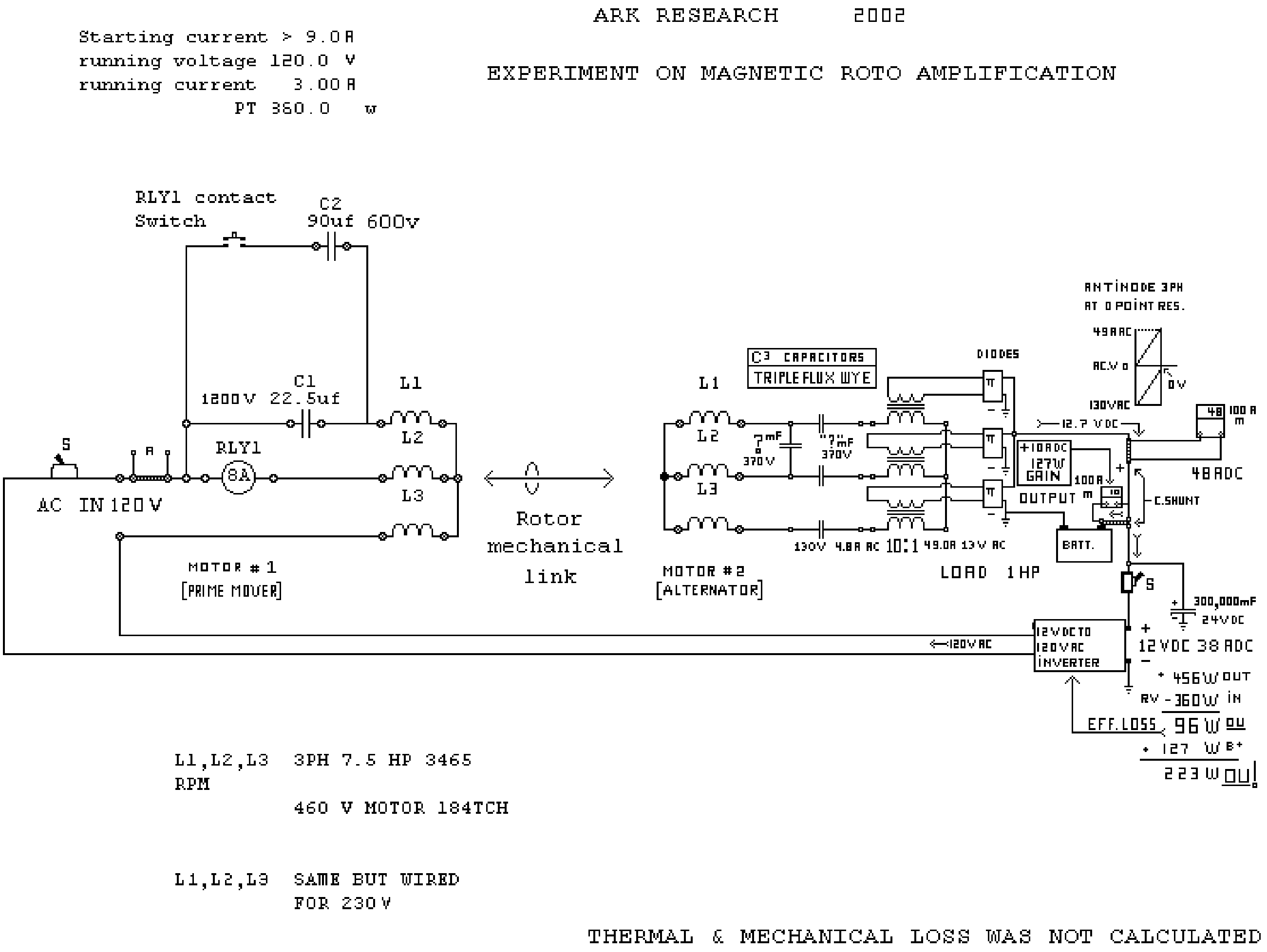
ARK RESEARCH 2002
EXPERIMENT ON MAGNETIC ROTO AMPLIFICATION
Starting current > 9.0 A
running voltage 120.0 V
running current 3.00 A
PT 360.0 w
RLY1 contact Switch
C2 90uf 600v
C1 1200V 22.5uf
L1
L2
L3
S
A
RLY1
8A
AC IN 120V
MOTOR # 1 [PRIME MOVER]
Rotor mechanical link
C3 CAPACITORS
TRIPLE FLUX WYE
L1
L2
L3
?mF 370V
"?"mF 370V
130V 4.8A AC 10:1 49.0A 13V AC
MOTOR # 2 [ALTERNATOR]
DIODES
ANTINODE 3PH AT 0 POINT RES.
49A AC
AC.V 0
0V
130VAC
12.7 VDC
48
100 A m
48 ADC
+10ADC 127W GAIN
OUTPUT
100A m
10
C.SHUNT
BATT.
LOAD 1HP
S
300,000mF 24VDC
12VDC TO 120VAC INVERTER
120VAC
12VDC 38 ADC
+ 456W OUT
RV -360W IN
EFF.LOSS 96 W OU
+ 127 W B+
223 W OU!
L1,L2,L3 3PH 7.5 HP 3465 RPM
460 V MOTOR 184TCH
L1,L2,L3 SAME BUT WIRED FOR 230 V
THERMAL & MECHANICAL LOSS WAS NOT CALCULATED

Fig.2

Subject: RV Basic level 1 Full MODEL Disclosure (Public domain)

This is the RV (Level 1) plan of a full looped RV system:

Battery provides primary power for 12VDC to 120VAC. As to run prime mover (Roto-converter), second motor acts as ansquirrel cage self-exited generator, a triple flux-capacitor LC tank tuned to best Standing wave condition as to create standing wave current node internally in battery at 0 voltage to battery "negative resistor". At 0 volts "voltage" a negative current is created as to maintain a reverse flow (charge) to battery exceeding the forward drain of the inverter demand, detuning system with a forward charge at 10 amperes with a voltage rise of approximately 0.83333 V over the battery. Voltage charge produces OU transform from the 0 point standing wave component. System gains energy from stochastic resonance within the LC tank components draining energy from "thermal" signature of the Aether and K thermodynamic-thermoelectric ambient heat (Electron spin). This is a full disclosure of an operational and tested device. System is made of standard off shelve items. Tuning is made by changing capacitor values and the proper selection of standard items for its construction, 3PH motors, 10:1 12V or 5:1 24V transformers with the proper core and winding values (standard) off shelve, diode bridges capacitor (all standard).

Warning: System opens (NON standard) space-time anomaly as time is reversed due to aether energy transform....
Extreme warning: Do not exceed 10KW Nor use of trans-uranic elements near unit as they may reach critical mass by regressive "inversed" decay. PB 206, U235, U238 must be maintained as far as possible from unit.
Extreme Warning: Dangerous Toy, Eldridge, Event horizon and hell raiser effects may be created by protoplasmic fields at extreme potentials PK.
Use wisely. You Can create Paradise or open the gates to Armageddon.

The choice is yours now....

Hector D Perez Torrez

Lifter Project
by Jean-Louis Naudin
http://www.jlnlabs.org

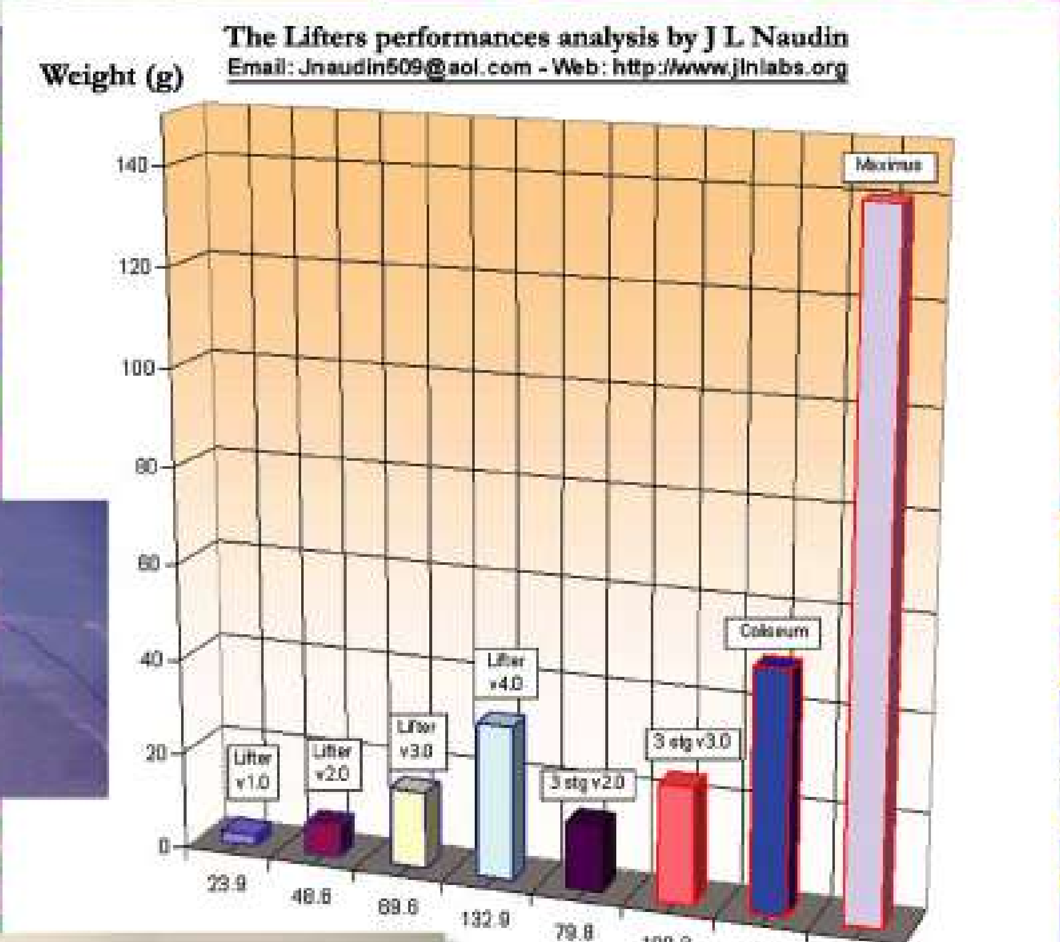

Mouse aboard!

The 1st Worldwide
Electronaut Lifter Project

ISSN 1684-7288

New Energy Technologies

Printed version ISSN 1684-7288
Online version ISSN 1684-9698

Published by Faraday Laboratories Ltd. *Issue #1 (10), 2003*

Research on advanced aerospace systems and new energy sources

Single-Wire Electric Power Transmission 20 kWt!

News from France

Electrostatic propulsion: 60g payload

Also in this issue:

- Cold Fusion by Plasma Electrolysis of Water
- News of Hydrogen Energetics
- Unusual Permanent Magnet Motors
- New Energy and Propulsion Systems

New Energy Technologies magazine
Back issues 2001-2002

Julia V. Sukhova
Translator

Kate M. Pomerantseva
Designer

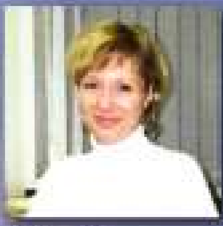

Olga O. Leontyeva
Editor

New Energy Technologies
History of invention: Poliakov's vortex drive
Also in this issue:
7 KW output
35% weight loss
Time Machine
Time Control experiments
Also in this issue:
Research on advanced propulsion systems and new energy sources
September-October 2001
Acad. Oleg Gritskevich's Hydro Magnetic Dynamo: 1500 KWt power without any fuel!
Also in this issue:
MEG
500%

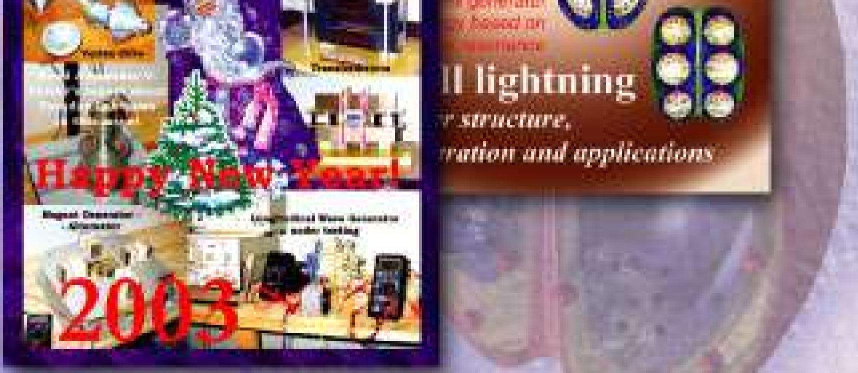
Happy New Year!
2003
ll lightning
r structure,
ration and applications

Kirill P. Butusov
Scientific Advisor

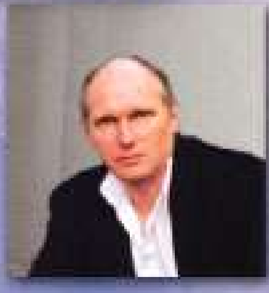

Alexander V. Frolov
General Director
"Faraday Laboratory" Ltd

Searl Effect Technology

http://www.sisrc.com

Experimental craft construction (self-rotating permanent magnet system)

15 kWt Searl Effect Generator

LENR Experiments
(Low Energy Nuclear Reactions) "Cold Fusion"

http://www.lenr-canr.org

Read in this issue!

NEW ENERGY TECHNOLOGIES #10

1. Testatika. Review of well-known electrostatic system.
2. Cold Fusion by Plasma Electrolysis of Water. Ph.M. Kanarev, T. Mizuno
3. Cold Fusion: What is it and what does it mean to science and society? E. Storms
4. Technical Introduction to LENR-CANR (Low Energy Nuclear Reactions). E. Storms
5. It started in 1989… (History of Cold Fusion Development). P. Hagelstein
6. A Science Tutorial (Cold fusion). T. Chubb
7. LENR (Low Energy Nuclear Reactions) Experiments.
8. On Electrogravitation. A.V. Lemeshko
9. Avalanchedrive Propulsion Device. M.L. Filho
10. Hydrogen-Powered Vehicles. P. Behr, G. Schneider
11. Unusual Permanent Magnet Motors. T. Valone
12. Hydrogen Energetics. A. Pashova
13. On the Longitudinal Electromagnetic Waves. A.V. Frolov
14. Space and Terrestrial Transportation and Energy Technologies. T.C. Loder
15. Commercialising the “Searl Effect”.
16. Interesting Publications of the Last Year.
17. “Lifter” Project
18. New Electric Fire Technology. V.D. Dudyshev
19. New Effect of “Cold” Evaporation. V.D. Dudyshev
20. Conceptual Hurdles to New Millennium Physics. D. Reed
21. Resonant Phenomena Occurring in Alternating Current Circuit. A.V. Frolov
22. Books Review

TESTATIKA

Review of well-known electrostatic system

Editor comments by Alexander V. Frolov
General Director, Faraday Lab Ltd

A running "free energy" machine coming from Switzerland, Europe. It was developed over a 20 years Research period by a religious group called Methernitha. This group lives in Linden, Switzerland. The head engineer of this superb machine, Mr. Paul Baumann discovered its principles while observing nature. He claims its running principle was found by studying the lightning effects from nature. The documents being offered here shed light on this energy marvel.

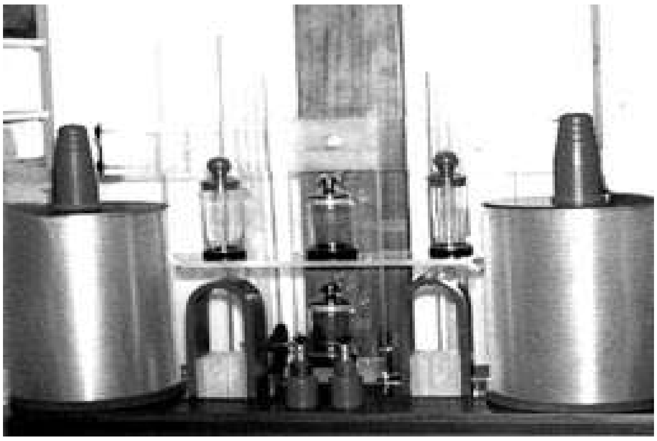

These documents are the result of work of people who have witnessed this wonder machine over the years... And here are the facts:

- The Testatika is a rotary-type machine that runs on it's own, once started by revolving it's two discs by hand!
- The machine not only runs on it's own energy, but produces also a huge amount of excess power, at least 3 KWtt of power! This is almost enough to supply a small house with one machine!
- The machine is about 70cm wide, by 40cm deep and is about 60cm in height!
- It delivers from 250 volts to about 320 volts direct current, depending on the humidity in the air! At this voltage it can supply at least 10 Amperes of pulsating direct current!
- Testatika is not a perpetual motion device, but an energy machine that collects it's huge amount of energy from the ions contained in air. However, there are some technological secrets implemented to overcome the normal drag-resistance of conventional generators. This is but one of it's secrets held by the Methernitha group.

There are many different sized Testatika machines, some of the smaller units deliver only about 200 to 300 watts, none of which are mass-produced. These devices are still laboratory prototype units.

Methernitha is a spiritual community (Christian Alliance) http://www.methernitha.com. They wrote: "We are fortunate in gaining the experience that paradoxically the most beautiful and useful results can be achieved by just using the most simple means. Never did we use any borrowed capital because we want to stay free Swiss citizens and do not want to be hindered or even bound in any way in the pursuance of our aims."

The two contrary-rotating discs generate an electrostatic charge. One disc represents the earth, the other the cloud. Using grid electrodes the charges are bound. After that they are collected by non-contacting so-called antenna keys and then sorted.

After being initially turned on by hand, the discs rotate by themselves according to the electrostatic laws about attraction and repulsion. A rectifying diode keeps the cycles in steady state. Otherwise the impulses of attraction and repulsion would accumulate and cause the discs to run faster and faster. The correct speed is of great importance and for optimal power generation the discs have to run quite steady and slow.

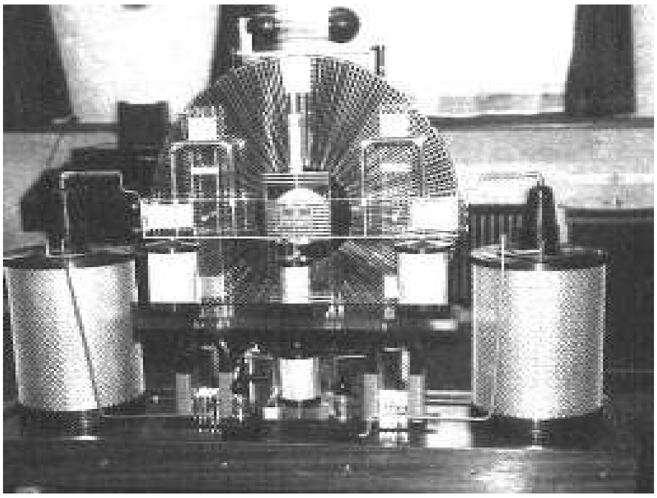

By means of grid condensers the energy is stored and then uniformly discharged, at the same time reducing the high voltage and building up power with additional devices. Finally the machine supplies a uniform direct current, which varies according to the size of the model. The machine furnishes about 3-4 kWt permanent output, depending on humidity, whereby the electric potential ranges from 270 to 320 Volt. High humidity of the atmosphere prevents the build-up of electric potential. The drier the air is, the better.

(Editor: It is necessary to note that the analogues method was used by Russian inventor Pavel N. Yablochkov who in 1877 take out a patent #120684 "The system of distribution and amplification of electrical currents by means of atmosphere electricity..." As it was shown by Yablochkov devices, power doubled in lamps, though current of consumption did not increse.)

No doubt, through the so far achieved results one main objective has been reached, namely to prove that it is possible to use Free Energy. Nevertheless the research work is not yet completed.

To the educated physicist many things of this machine may seem impossible, maybe even crazy. Maybe he is also offended by the conceptions used to explain the whole. Only partly we could use the concepts of conventional physical terminology to explain and define only approximately the functions and properties of the various parts of the machine.

After all it will be necessary to create some more new concepts like the one we have already used before, when we termed the non-contacting collectors of electric charges as antenna keys.

This machine puts experts, which are just trained in conventional physics to a very hard test, because its mode of action is not explainable with the state of the art of officially accepted physical knowledge, or at the most only partially explainable. However also a trained specialist should remain free and independent in his thinking, and should avoid to be limited by the temporal framework of publicly admitted knowledge in any science.

It has to be noted that the established science was already many times forced to change or give up some of its very fundamental concepts. Think about Galilee, to name only one example. Our human society almost condemned this man as a sorcerer and magician, just because he investigated and discovered a truth that seemed unacceptable by the established science of the days.

The book knowledge of any times is not wrong, but it is incomplete, and therefore allows to draw wrong conclusions.

There was used information from the official web site: http://www.methernitha.com

Official adress of Methernitha: Methernitha Genossenschaft Administration, Moosbuehlweg 2, 3517, Linden, SWITZERLAND

Official E-mail of Methernitha: info@methernitha.com

New Energy Technologies

PRINT EDITION ADVERTIZING RATES

Full Page Ad

B/W - $400

Color - $800

Half Page Ad

B/W - $200
Color - $400

Quarter Page Ad

B/W -$100
Color- $200

Ask about discounts for multiple ads

New Energy Technologies is distributed in the leading Scientific Laboratories in Russia and has subscribers in more than 20 countries all over the world

To advertise contact us:
Faraday Labs Ltd, Phone/fax: 7-812-380-3844
http://www.faraday.ru, office@faraday.ru

COLD FUSION BY PLASMA ELECTROLYSIS OF WATER

Ph.M. Kanarev
The Kuban State Agrarian University,
Department of Theoretical Mechanics
13, Kalinin Street, 350044 Krasnodar, Russia
E-mail: kanphil@mail.kuban.ru

Tadahiko Mizuno
Faculty of Engineering, Hokkaido University,
Kita-ku, North 13, West-8 Sapporo 060-8628, Japan

Abstract: It has been disclosed that transmutation of the atomic nuclei of alkaline metals and the atomic nuclei of the cathode material takes place during plasma electrolysis of water.

Key words: atom, nucleus, proton, neutron, electron, cathode, low-current.

INTRODUCTION

Cold nuclear fusion is the first hypothesis of a source of additional energy in heavy water electrolysis. Fleischmann and Pons, the American electrochemists, are the authors of this hypothesis [1]. They reported about it in 1989. Since that time a large number of experiments has been carried out in order to obtain additional energy from water [2], [3], [4], [5], [7], [8], [9], [10], [11], [12]. We continue to discuss this problem.

THE FIRST EXPERIMENTAL PART

In order to check this hypothesis, the following experiments were performed. Two cathodes were made of iron with mass of 18.10 g and 18.15 g. The first cathode operated during 10 hours in KOH solution; the second cathode operated during the same period in NaOH solution. Mass of the first cathode remained unchanged; mass of the second one was reduced by 0.02 g. The voltage by plasmaelectrolysis process was 220 V and the current (0.5-1.0) A (Fig.1). The indices of the consumption of the solution and the gases being generated were as follows (Table 1).

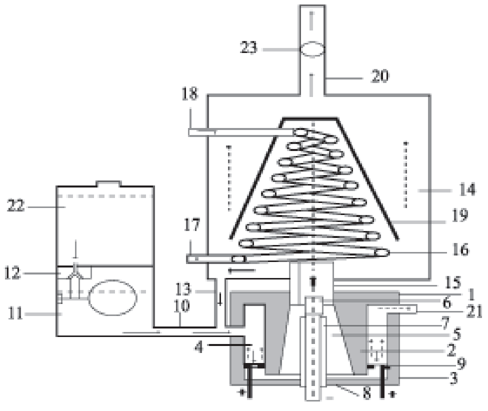

Fig. 1.
Diagram of gas generator. Patent # 2175027:
1 - lid of the reactor; 3 - body of the reactor; 6 - the cathode; 9 - the anode; 11 - solution dosing unit; 16 - cooler; 20 - pipe for gas release; 23 – anemometer

Table 1
Experimental results

Indices	Water consumption, kg	Volume of gases,m^3	Energy expenses, kWh/m^3
KOH	0.272	8.75	0.28
NaOH	0.445	12.66	0.21

In order to increase safety of experimental results, the volume of the gases introduced with the help of anemometer is reduced twofold.

It is known that from one litre of water it is possible to produce 1220 litres of hydrogen and 622 litres of oxygen. Quantity of the gases generated by the plasma electrolytic process is much greater than it is possible to get from consumed water (Table 1) [6]. It gives the reason to think that not only water molecules, but also the nuclei of alkaline metals and the atomic nuclei of the cathode material serve as a source of these gases. The analysing experiment has been performed in order to check this fact.

Tadahiko Mizuno, the famous Japanese scientist (the co-author of this article), who works at the Division of Quantum Energy Engineering Research group of Nuclear System Engineering, laboratory of Nuclear Material System, Faculty of Engineering, Hokkaido University, Japan, kindly agreed to perform chemical analysis of the cathode samples with the help of the nuclear spectroscopy method (EDX). Here are the results of his analysis. The content of chemical elements on the surface of non-operating cathode is as follows (Table 2).

Table 2
Chemical composition of the cathode surface prior its operation in the solution

Element	Fe
%	99.90

The new chemical elements have appeared on the working surface of the cathode, which works in KOH solution (Table 3).

Table 3
Chemical composition of the surface of the cathode, which operates in KOH solution

Element	Si	K	Cr	Fe	Cu
%	0.94	4.50	1.90	93.00	0.45

The chemical composition of the surface of the cathode, which operates in NaOH has proved to be different (Table 4).

Table 4
Chemical composition of the surface of the cathode, which operates in NaOH solution

Element	Al	Si	Cl	K	Ca	Cr	Fe	Cu
%	1.10	0.55	0.20	0.60	0.40	1.60	94.00	0.65

Thus, the hypothesis concerning the participation of the nuclei of alkaline metals and the atomic nuclei of the cathode material in the formation of gases during plasma electrolysis of water has experimental confirmation. Let us carry out the preliminary analysis of the data being obtained (Tables 2, 3, 4).

THE FIRST THEORETICAL PART

In any of these cases, the atoms and the molecules of hydrogen are formed. The part of it is burned and the other goes out with the steam. We have already shown that the processes of fusion of the atoms and the molecules of hydrogen and its isotopes result in occurrence of additional thermal energy [6]. Numerous experiments show that up to 50% of additional thermal energy are generated during the plasma electrolysis of water, it is less than the results of the calculations originating from the existing cold fusion theories [6]. That's why it is necessary to analyse energetics of the particle creation process during the atomic nucleus transmutation.

Having considered the model of the electron we have found out that it can exist in a free state only when it has a definite electromagnetic mass [6]. Being combined with the atomic nucleus it emits a part of energy in the form of the photons, and its electromagnetic mass is reduced. But stability of its condition does not become worse, because the energy carried away by the photons is compensated by binding energy of the electron in the atomic nucleus [6].

If the ambient temperature is increased, the electron begins to absorb the thermal photons and to pass to higher energy levels of the atom reducing binding with it. When the electron becomes free, it interacts with the atom only if the ambient temperature is reduced. As this temperature is reduced, it will emit the photons and sink to lower energy levels [6].

If the electron is in a free state due to an accidental external influence on the atom and the environment has no photons, which are necessary for it to restore its mass, **it begins to absorb the ether from the environment and to restore its constants in such a way**: mass, charge, magnetic moment, spin and radius of rotation. The electron acquires the stable free state only after it has restored its all constants [6].

Thus, if an interchange of the free state and binding state with the atom takes place due to the accidental influences on the atom, **the electron restores its electromagnetic mass every time due to absorbing the ether. It means that actually it plays the role of a converter of the ether energy into the thermal photon energy.**

The Japanese investigators Ohmori and Mizuno [4] registered neutron radiation during plasma electrolysis of water and reported that not only the nuclear process, but also the process of the electron capture by the free protons can be the source of this radiation.

As hydrogen plasma is generated during the plasma electrolytic process of water electrolysis, there exists a tendency of the capture of the free electrons by them.

It is known that rest mass of the electron is $m_e = 9.109534 \cdot 10^{-31}$ *kg*, rest mass of the proton is $m_p = 1.6726485 \cdot 10^{-27}$ *kg*, and rest mass of the neutron is $m_n = 1.6749543 \cdot 10^{-27}$ *kg*. The difference between the mass of the neutron and the mass of the proton is equal to $\Delta m_{np} = 23.058 \cdot 10^{-31}$ *kg*. It is $23.058 \cdot 10^{-31} / 9.109 \cdot 10^{-31} = 2.531$ of the mass of the electron. Thus, the proton should capture 2.531 electrons in order to become the neutron. The question arises at once: what will happen to the remained of electron mass $(3.0-2.531)m_e = 0.469m_e$? The disturbed balance of masses in this process is explained by modern physics in a simple way: a neutrino is created [6].

As the neutrino has no charge, it is very difficult to register it. If the neutrino takes the excess mass away or replenish the lacking one, can the elementary particles execute this process by themselves?

As the photons are emitted and absorbed only by the electrons, the proton, which absorbs the electrons, cannot convert the remainder of mass of the third electron into the photon. If the electron is absorbed by the third one and gives more than a half of its mass to the proton in order to convert it into the neutron, the remaining part of mass ($0.469m_e$) of the electron, which has no possibility to become the photon, is converted into a portion of the ether, which "is dissolved" and mixed with the ether in the space. The fact that plasma has no photons with the mass corresponding to the part of mass of the third electron, which has not been absorbed by the proton during its conversion into the neutron, can serve as a proof of such affirmation. Let us calculate energy of such photon [6].

The difference the mass of the neutron and the proton is equal to $\Delta m_{np} = 23.058 \cdot 10^{-31}$ *kg*. If we subtract this value from the mass of three electrons, we'll get mass m_F, from which the photon should be formed [6]

$$m_F = 3m_e - \Delta m_{np} = 3 \cdot 9.109534 \cdot 10^{-31} - 23.05810^{-31} = 4.270602 \cdot 10^{-31} \text{ kg} \quad (1)$$

If the photon is formed from this remainder of mass m_F, its energy will be [6]:

$$E_{ph} = m_F \cdot C^2 = \frac{4.270602 \cdot 10^{-31} \cdot (2.997924 \cdot 10^{8})^2}{1.602189 \cdot 10^{-19}} = 23.956126 \cdot 10^{4} eV \quad (2)$$

This value of energy corresponds to roentgen spectrum, that's why the creation of each free neutron should be accompanied by the creation of one roentgen photon. If it does not take place, we have two opportunities: the first one – we should think that in the case when the neutron is created, the neutrino was formed from mass m_F=4.270602 × 10_{-31} *kg* and flew away in the unknown direction; the second one – there were no conditions for the formation of the photons in the process being considered, and mass, which failed to be formed as a particle, "was dissolved" in the ether. Which variant is closer to the truth [6]? There is no exact answer, but it is known that the Japanese scientists registered only neutron radiation with intensity of 50,000 neutrons per second, and they failed to register roentgen radiation [4].

If in this process the roentgen photons were created, they would not exceed heat efficacy of the plasma electrolytic process, because they would not be the thermal photons. The thermal photons are radiated and absorbed when the electrons make the energy transitions to the energy levels, which are the most remote from the atomic nuclei, where the infrared photons and neighbouring ones from the optical range of the spectrum with energies of ≈(0.001-3.3) eV are generated (Table 5) [6].

Table 5
Electromagnetic spectrum bands

Bands	Wave-length, m	Energy, eV
1. Low- frequency band	$\lambda \approx (10^{7}...10^{4})$	$E \approx 10^{-15}...10^{-11}$
2. Broadcast band	$\lambda \approx (10^{4}...10^{-1})$	$E \approx 10^{-11}...10^{-6}$
3. Microwave band	$\lambda \approx (10^{-1}...10^{-4})$	$E \approx 10^{-6}...10^{-3}$
4. Relic band (maximum)	$\lambda \approx 1 \times 10^{-3}$	$E \approx 1.2 \times 10^{-3}$
5. Infrared band	$\lambda \approx (10^{-4}...7.7 \times 10^{-7})$	$E \approx 10^{-3}...1.6 \times 10^{-2}$
6. Light band	$\lambda \approx (7.7 \times 10^{-7}...3.8 \times 10^{-7})$	$E \approx 1.6 \times 10^{-2}...3.27$
7. Ultraviolet band	$\lambda \approx (3.8 \times 10^{-7}...10^{-9})$	$E \approx 3.27...1 \times 10^{2}$
8. Roentgen band	$\lambda \approx (10^{-9}...10^{-12})$	$E \approx 10^{2}...10^{5}$
9. Gamma band	$\lambda \approx (10^{-12}...10^{-18})$	$E \approx 10^{5}...10^{9}$

Thus, the neutron fusion processes in plasma electrolysis of water will not generate additional thermal energy. But the appearance of the neutrons in plasma will promote the formation of the nuclei of deuterium and, possibly, of tritium. As the balance of masses remains almost unchanged, we have no reason to expect that **additional energy will take place** when deuterium and tritium are formed. But it is sure to appear during **fusion of the atoms of deuterium and tritium, i.e. the hydrogen atoms [6].**

In order to become a proton, the neutron should radiate something, which mass is Δm_{np}=23.058 × 10^{-31} *kg*. Let us convert this mass into energy [6].

$$E_{ph} = \Delta m_{np} \times C^2 = \frac{23.058 \times 10^{-31} \times (2.998 \times 10^{8})^2}{1.602 \times 10^{-19}} = 1.294 \times 10^{6}\, eV \qquad (3)$$

This energy corresponds to the gamma range photons, i.e. not to the thermal photons, and this process does not give additional energy. Thus, if the process of the formation of the helium atoms takes place during plasma electrolysis of water, it should be accompanied by gamma radiation. If there is no such radiation, but the helium atoms are formed, the neutrino takes away the above-mentioned portion of mass Δm_{np} or this mass, which has no opportunity to be formed as the photon, "is dissolved" in the environment, i.e. it is transferred into the state of the ether [6]. As the roentgen photons and the gamma photons are not the thermal ones, this process gives no excessive thermal energy [6].

Another variant is possible. When the atoms of alkali metal bombard the cathode atoms, they are destroyed completely and destroy the atoms of the cathode materials. Under the notion "completely" we'll understand such state when both the atom and the nucleus are destroyed. In this case, the protons of the destroyed nuclei begin to form the hydrogen atoms. The process of fusion of the atoms and the molecules of hydrogen generate additional thermal energy [6]. But one should bear in mind that if plasma disintegrates water molecule into hydrogen and oxygen and if these gases contact plasma, hydrogen is combined with oxygen, and water is formed. Noise generated by plasma is hydrogen microexplosions. Taking into consideration the above-mentioned fact the larger the volume of hydrogen burnt in plasma, the smaller its volume in the gas-vapour mixture. It means that such reactor operation modes are required when quantity of burnt hydrogen is minimal one. Our theory allows us to have such results.

As iron is the cathode material, the nuclei of its atoms are the targets of the atomic nuclei of potassium, alkaline metal. During the transmutation of the iron nuclei (Fig. 2 b), the atomic nuclei of chromium (Fig. 2 a) and the atomic nuclei of copper (Fig. 2 c) are formed [6].

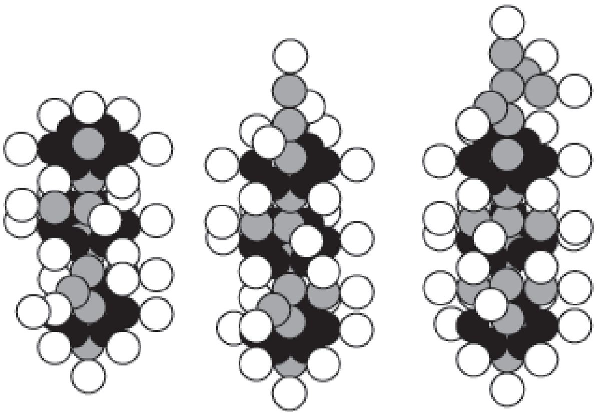

a) Cr (24,28) b) Fe (26,28) c) Cu (29,34)

Fig. 2.
Diagrams of the atomic nuclei of: a) chromium, b) iron, c) copper

When the atomic nucleus of iron (Fig. 2 b) pass into the atomic nucleus of chromium (Fig. 2 a), two protons and two neutrons are released; two atoms of deuterium or one atom of helium can be formed from them. If the neutrons pass into the protons, four atoms of hydrogen are formed.

It is easy to see (Fig. 2) that the atomic nucleus of iron (Fig. 2 b) should lose two upper protons and two neutrons in order to pass into the atomic nucleus of chromium (Fig. 2 a).

Three additional protons and six neutrons (total 9 nucleons) are required for the formation of the atomic nucleus of copper (Fig. 2 c) from the atomic nucleus of iron. As on the cathode surface (Table 3) the number of chromium atoms, which probably are formed from the atomic nuclei of iron, four times more than the number of atoms of copper, then the solution is sure to have superfluous protons and neutrons of the destroyed atomic nuclei of iron, and we can determined their approximate relative quantity.

Let us suppose that four nuclei of the iron atoms pass into the nuclei of the chromium atom. The total quantity of free protons and neutrons (nucleons) is equal to 16. As one atom of copper falls on each four atoms of chromium, 9 nucleons are spent for the formation of one nucleus of the copper atom, and 7 nucleons remain free.

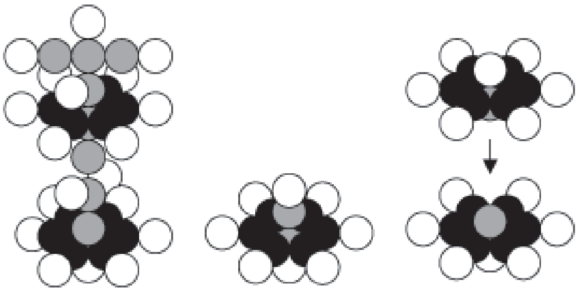

a) K (19,20) b) O (8,8) c) Si (14,14)

Fig. 3.
Diagrams of the atomic nuclei of:
a) potassium, b) oxygen, c) silicon

Let us see what is formed when the nucleus of the potassium atom is destroyed. Potassium is situated in the first group of the fourth period of the periodic law. Its nucleus contains 19 protons and 20 neutrons (Fig. 3 a) [6].

In Fig. 3 a, we can see a weak link of the nucleus of the potassium atom [6]. It is situated in the middle of its axis neutrons. When the transmutation of the nuclei of the potassium atoms takes place, the nuclei of the oxygen atoms can be formed (Fig. 3 b) as well as its isotopes and the nuclei of the silicon atoms (Fig. 3 c).

The analysis of the structure of the nuclei of the potassium atom (Fig. 3 a) shows that it is the most probable source of the nucleus of the silicon atom (Fig. 3 c), which atoms appear on the cathode (Table 3).

It is easy to count that during the destruction of one nucleus of the potassium atom and the creation of one nucleus of the silicon atom 5 free protons and 6 free neutrons, i.e. 11 nucleons, are formed.

Thus, the transmutation of the nuclei of the iron atoms and the potassium atoms results in the formation of free protons and neutrons. As the protons cannot exist in free state, the hydrogen atoms are created from them. If the protons are connected with the neutrons after the destruction of the nuclei of the iron atoms and the potassium atoms, the formation of deuterium, tritium and helium is possible.

Let us pay attention to the main fact – absence of the sodium atoms in the cathode material. It is natural that the potassium atoms have appeared on the cathode, which operated in KOH solution (Table 3). Why are no sodium atoms on the cathode, which operated in NaOH solution? The answer is as follows: the nuclei of the sodium (Fig. 4,a) atoms are completely destroyed during the plasma electrolytic process. The presence of potassium on the surface of the cathode, which operated in NaOH solution (Table 4), can be explained by insufficient ablution of the reactor after the operation with KOH solution.

As free protons and neutrons appear during the destruction of the nucleus of the sodium atom (Fig. 4,a), some nuclei of this element begin to form the atomic nuclei of aluminium (Fig. 4, b), chlorine (Fig. 4, c) and calcium (Fig. 5).

But not all free protons and neutrons are spent for the construction of the atomic nuclei of aluminium, chlorine and calcium. A part of them is spent for the hydrogen atom formation.

If we knew the total quantity of transmutating atomic nuclei of iron, potassium and sodium as well as the exact composition of the gases generated during the plasma electrolytic process, it would be possible to determine the atomic nuclei being formed from additional nucleons. Now we can only suppose that the majority of new nuclei are the protons, i.e. the nuclei of the hydrogen atoms. The increased volume of the gases generated during the plasma electrolytic process is explained by it [6].

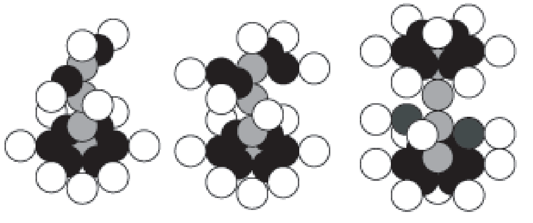
a) Na (11,12) b) Al (13,14) c) Cl (17,18)

Fig. 4.
Diagrams of the atomic nuclei of:
a) sodium, b) aluminium, c) chlorine

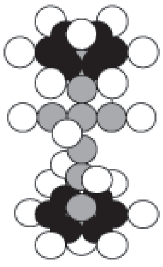
Ca (20,20)

Fig. 5.
Diagram of the nucleus of the calcium atom

The analysis of these Tables shows that transmutation of the nuclei of iron, of which the cathodes are made, results in the formation of chromium and copper in both cases. Apparently, aluminium, chlorine and calcium are formed from the destroyed sodium nuclei. In any case, free protons and neutrons are formed.

But not all free protons and neutrons are spent for the formation of the atomic nuclei of copper, aluminium, chlorine and calcium. A part of them is spent for the formation of the hydrogen atoms. In any case, the atoms and the molecules of hydrogen are formed. The analysis has shown that plasma electrolytic process extracts not more than 0.005 kg of alkaline metal from one litre of the solution. It appears from this that if all neutrons of the atomic nuclei of the molecules of water and alkali metals are transferred into the protons and the atoms and the molecules of hydrogen are formed, the formed volume of gas will be considerably less than the one registered during the experiment (Table 1). A question arises: where do additional gases come from? In order to get the answer on this question we made the next experiment.

THE SECOND EXPERIMENTAL PART

First of all we take into account, that high temperature of plasma forms the conditions when a set of various processes takes place at the cathode. First of all, water is boiled and evaporated. At the same time, one part of water molecules is disintegrated with a release of the atomic hydrogen; another part of the molecules forms the orthohydrogen molecules. A part of water molecules is disintegrated completely and is released at the cathode together with hydrogen and oxygen. A part of hydrogen is combined with oxygen again generating microexplosions (noise) and forming water. During plasma electrolysis of water, water vapor, hydrogen and oxygen are released simultaneously. If vapor is condensed, gas mixture is released. In order to measure gas flow rate the electronic anemometer have been used. Diameter of the electronic anemometer was equal to internal diameter of the gas make tube (23, Fig. 1). Its readings were registered and processed by the computer. The experiment was performed dozen times, and each time its readings were reproduced with small deviations [11]. But we had no hydrogen analyzer, that's why the results being obtained cannot be considered as final ones. We admonished it in all editions of the book Water is a New Source of Energy with such a phrase: "We abstain from lending an official status to these results with the hope to get necessary financing and to repeat them with a complete set of the necessary devices" [12, page 176].

In the middle of the year of 2002 we received small financing, which allowed us to make a new reactor and to buy some measuring instruments, in particular the scales with the measurement limit up to 600 g and accuracy of 0.02 g. Careful preparation allowed us to increase duration of continuous operation of the reactor (to 10 and more hours) and to register solution consumption for gas production.

The main difficulty of operation with the hydrogen is in the fact that its mixture with air (4-74)% or oxygen (4-94)% is combustible, and the fact was emphasized more than once during the experiments that made the researches be very careful. The second difficulty during hydrogen quantity measurements generated by the plasma electrolytic reactor is in the fact that its molecule has the smallest dimensions, that's why it penetrates easily to the places where the molecules of other substances do not penetrate. Molecular hydrogen diffuses easily even into metals. For example, one volume of palladium absorbs up to 800 volumes of hydrogen.

Gas flow speed was measured with the help of various anemometers, its readings being registered with the help of the computer. Numerous measurements and numerous analysis of gas flow speed measurement accuracy with the help of the anemometers showed that error of a conventional anemometer can be 100%.

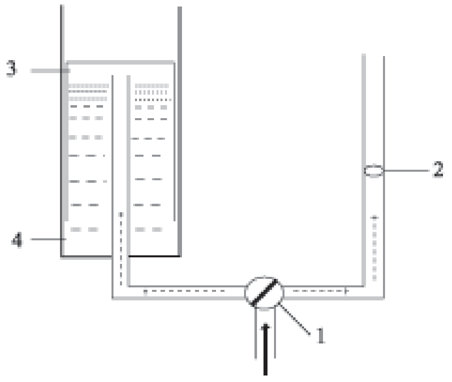

Fig. 6.
Diagram of measurement of flow rate of the gas and its volume:
1 - tap for gas flow movement direction switching,
2 – anemometer, 3 – graduated tank, 4 – water tank

It is known that it is possible to produce 1220 litres of hydrogen and 622 litres of oxygen from one litre of water. Quantity of the gases generated by the plasma electrolytic process is much greater than it is possible to get from consumed water (Table 1). It was a strong reason for a search of the measurement error. For this purpose, the diagram of measurement of flow rate of the gases and their quantity was used (Fig. 6).

The results of the measurements were as follows. The anemometer showed that 200 litres of gas mixture penetrated through it during 10 minutes. Nearly one litre of gases was in the graduated tank during this period.

Thus, the measurement of gas flow with the help of the anemometers distorted the result 200 fold. It should be mentioned that the reactor operated in the production mode of hydrogen and oxygen in the cathode zone. As a result, their mixture burst. The pulses of these explosions increased the readings of the anemometer.

It has become necessary to return to the reactor operation modes when no oxygen is released in the cathode zone. Our theory allows us to do this easy.

PROTOCOL
of tests of the first model of low-current Electrolyzers

It is known that it is possible to produce 1.22 l of H_2+ 0.622 O_2= 1.843 (H_2+O_2) from 1 ml of H_2O

Table 6
Experimental results

Indices	1	2	3	Average
1-duration of experiment, hour	1	1	1	1
2-voltage, V	70	70	70	70
3-current, A	0.038	0.080	0.098	0.072
4 – power, W	2.7	5.60	6.44	4.91
4-volume of consumed solution, ml	1.67	3.98	4.32	3.32
5-density of the solution, kg/l	1.04	1.04	1.04	1.04
6-volume of consumed water, ml	1.60	3.83	4.15	3.19
7-volume of the gas mixture being produced, l	2.95	7.06	7.85	5.95
6-volume of hydrogen being produced, l	1.95	4.67	5.07	3.80
7-energy consumption per 1 l of hydrogen, W·h/l	1.38	1.20	1.27	1.28
8-energy consumption per $1m^3$ of hydrogen, kWh/m^3	1.38	1.20	1.27	1.28
9-existing energy consumption for production of 1 m^3 of hydrogen from water, kWh/m^3	4.00	4.00	4.00	4.00

CONCLUSION

Transmutation of the atomic nuclei of alkaline metals and the atomic nuclei of the cathode material during plasma electrolysis of water existed. Plasma electrolytic process opens new prospects in study of matter on the nuclear, atomic and molecular levels. The low-current electrolysis allows us to get the inexpensive hydrogen from water.

REFERENCES

1. M. Fleischmann, S. Pons and M. Hawkins. Electrochemically Induced Nuclear Fusion of Deuterium. J. Electroanal. Chem. 261, 301 (1989),
2. ICCF-7 ACCEPTED ABSTRACTS. Infinite Energy. V. 4, Issue 20, p. 59…69.
3. Harold L. Fox. Cold nuclear fusion: essence, problems, influence on the world. View from USA. Production group SVITAX. M.: 1993, 180 pages.
4. T. Ohmori, T. Mizuno. Strong Excess Energy Evolution, New Element Production, and Electromagnetic Wave and/or Neutron Emission in Light Water Electrolysis with a Tungsten Cathode. Infinite Energy. Issue 20, 1998. Pages 14-17.
5. T. Mizuno. Nuclear Transmutation: The Reality of Cold Fusion. Infinite Energy Press. 1998. 151 pages.
6. Ph. Kanarev. Foundations of Physchemistry of Microworld. Krasnodar, 2002. 330 pages.
7. Kenneth R. Shoulders, “Method of and Apparatus for Production and Manipulations of High Density Charge”, U.S. Patent 5,054,046, issued Oct 1, 1991.
8. Ken Shoulders & Steve Shoulders, “Observations on the Role of Charge Clusters in Nuclear Cluster Reactions”, J.of New Energy, vol. 1, no 3, pp 111-121, Fall 1996, 7 refs, 22 figs.
9. Hal Fox, Robert W. Bass, & Shang-Xian Jin, “Plasma-Injected Transmutation”, J. of New Energy, vol. 1, no 3, Fall 1996, pp 222-230, 23 refs, 4 figs.
10. Shang-Xian Jin & Hal Fox, “High Density Charge Cluster Collective Ion Accelerator”, J. of New Energy, vol. 4, no 2, Fall 1999, pp 96-104, 47 refs, 4 figs., 3 tables.
11. Ph.M. Kanarev. Water is the Main Power Carrier of Future Power Engineering. Journal of New Energy. An International Journal of New Energy Systems. Vol. 6, No.2. Pag. 101-121.
12. Ph.M. Kanarev. Water is New Source of Energy. The Third Edition. Krasnodar 2002. 194p. (In English).

LENR
(Low Energy Nuclear Reactions)

http://www.lenr-canr.org

Editorial: There is a collection of some papers on LENR (Low Energy Nuclear Reactions), also known as Cold Fusion. CANR, Chemically Assisted Nuclear Reactions, is another term for this phenomenon. These original scientific papers are reprinted with permission from the authors and publishers.

Website http://www.lenr-canr.org features a growing library of scientific papers about LENR and an extensive bibliography of journal papers, news articles and books about LENR.

COLD FUSION: What is it and what does it mean to science and society?

Edmund Storms

Cold fusion is important because it promises to be a new source of pollution-free, inexhaustible energy. In addition, it is important because it reveals the existence of a new way nuclei can interact that conventional scientific theory predicts is impossible. What then is this phenomenon that suffers such promise and rejection?

Energy can be obtained from the nucleus in two different ways. On the one hand, a large nucleus can be broken into smaller pieces, such as is experienced by uranium in a conventional nuclear reactor and by the material in an atom bomb. This is called fission. On the other hand, two very small nuclei can be joined together, such as occurs during fusion of deuterium and tritium in a Hot Fusion reactor and in a hydrogen bomb. This process, called fusion, also takes place in stars to produce much of the light we see.

The fission reaction is caused to happen by adding neutrons to the nucleus of uranium or plutonium to make it unstable. The unstable nucleus splits into two nearly equal pieces, thereby releasing more neutrons, which continue the process. As every one now knows, this process produces considerable waste that is highly radioactive. The uranium used as fuel also occurs in limited amounts in the earth's crust. As a result, this source of energy is not ideal, although widely used at the present time.

The normal hot fusion reaction requires two deuterium or tritium nuclei to be smashed together with great energy. This is accomplished by raising their temperature. However, this temperature is so high that the reactants cannot be held in a solid container, but must be retained by a magnetic field. This process has proven to be very difficult to accomplish for a time sufficient to generate useable energy. In spite of this difficulty, attempts have been under way for the last 40 years and with the expenditure of many billions of dollars. Success continues to be elusive while the effort continues.

Cold fusion, on the other hand, attempts to cause the same process, but by using solid materials as the container held at normal temperatures. The container consists of various metals, including palladium, with which the deuterium is reacted to form a chemical compound. While in this environment, the barrier between the deuterium nuclei is reduced so that two nuclei can fuse without having to be forced together. Because the process causing this to happen is not well understood, the possibility is rejected by many conventional scientists. Difficulty in producing the process on command has intensified the rejection.While this difficulty is real, it has not, as many skeptics have claimed, prevented the process from being reproduced hundreds of times in laboratories all over the world for the past 13 years. As you will see by reading the reviews and papers in our Library (see http://www.lenr-canr.org/LibFrame1.html), the process continues to be reproduced with increasing ease using a variety of methods and materials.

What is the nature of this process and why has it been so hard to understand? To answer this question, a person needs to understand the nature of the barrier that exists between all nuclei. Because all nuclei have a positive charge in proportion to their atomic number, all nuclei repeal each other. It is only the surrounding electrons that hold normal matter together, with the nuclei being at considerable distance from each other, at least on the scale of an atom. When attempts are made to push the nuclei closer, the required energy increases as the nuclei approach one another. However, when deuterium dissolves in a metal, it experiences several unique conditions. The surrounding metal atoms produce a regular array that is able to support waves of various kinds. These waves can be based on vibration of the atoms (phonons), vibration of the electrons, standing waves of electromagnetic energy, or a wave resulting from conversion of the deuterium nuclei to a wave. In addition, the high density of electrons can neutralize some of the positive charge on the deuterium nuclei allowing a process called tunneling, i.e.. allowing passage through the barrier rather than over it. The mechanism of this neutralization process is proposed to involve a novel coherent wave structure that can occur between electrons under certain conditions. All of these wave processes have been observed in the past under various conventional conditions, but applying them to the cold fusion phenomenon has been a subject of debate and general rejection.

While the debate based on wave action has been underway, people have proposed other mechanisms. These include the presence of neutrons within the lattice. Normally, neutrons are unstable outside of the nucleus, decomposing into a proton, an electron, and a neutrino. Presumably, this reaction can be reversed so that neutrons might be created in a lattice containing many free electrons and protons. Having no charge, the neutron could then interact with various atoms in the lattice to produce energy. These neutrons might also be hidden in the lattice by being attached to other nuclei in a stabilized form, to be released when conditions were right. Several particles normally not detected in nature also have been proposed to trigger fusion and other nuclear reactions.

While search for a suitable mechanism has been underway, an understanding of the environment that triggers the mechanism has been sought, the so-called nuclear-active-environment. Initially, this environment was thought to exist in the bulk of the palladium cathode used in the Pons-Fleischmann method to produce cold fusion. It is now agreed that the nuclear reactions only occur in the surface region. Recent arguments suggest that this surface layer does not even require palladium for it to be nuclear-active. Nuclear reactions have now been produced in a variety of materials using many methods. The only common feature found in all of these methods is the presence of nano-sized particles of material on the active surface. If this observation is correct, four conditions seem required to produce the nuclear reactions. First, the particle must have a critical small size; second, it must contain a critical concentration of deuterium or hydrogen; third, it must be constructed of certain atoms; and fourth, it must be exposed to a source of energy. This energy can take the form of a sufficiently high temperature, a significant high flux of hydrogen through the particle, application of energetic electrons or charged particles, or application of laser light of the proper frequency. Until, the importance of these factors is understood, the effect will continue to be difficult to replicate.

Technical Introduction to LENR-CANR

Edmund Storms

At low energies, the Coulomb barrier prevents nuclei from coming together and fusing to form a single nucleus. To initiate a nuclear reaction, several methods are used. Nuclear reactions are normally initiated by pushing two atoms together with enough force to overcome the Coulomb barrier by brute force, or by using neutrons which penetrate the nuclei without seeing a barrier. (Neutrons have no electrical charge, so the Coulomb barrier does not stop them.) These forces are normally provided by high-temperature plasma or by accelerating ions to high energies. In contrast, LENR describes the mechanism and conditions that cause a variety of nuclear reactions to take place with relatively low activation energy. These unique conditions reduce the need for excessive energy. The normal method forces the nuclei together, while the new method encourages them to come together. The challenge has been to understand the unique characteristics of the necessary solid structure such that this structure could be generated at will.

Because the proposed method is unique, at odds with current nuclear theory, and is still difficult to reproduce, support for studies in many countries, but not all, has been very limited. Nevertheless, considerable information has accumulated over the last 13 years since Profs. Stanley Pons and Martin Fleischmann showed the world the possibilities inherent in this phenomenon. Much understanding is buried in conference proceedings and reports that are not available to a serious student. This information will, as time permits, be made available on this site. Students of the subject are also encouraged to use this site to interact with other people in the field and provide objective critiques of the work published here.

PHENOMENA DISCUSSED IN SOME OF THE PAPERS

At least 10 ways have been demonstrated to produce anomalous heat and/or anomalous elemental synthesis. A few of these methods will be described here. For course, not all of the claims are worthy of belief nor are they accepted by many people. Nevertheless, the claims will be described without qualifications in order to provide the reader with the latest understanding.

conditions. The surrounding metal atoms produce a regular array that is able to support waves of various kinds. These waves can be based on vibration of the atoms (phonons), vibration of the electrons, standing waves of electromagnetic energy, or a wave resulting from conversion of the deuterium nuclei to a wave. In addition, the high density of electrons can neutralize some of the positive charge on the deuterium nuclei allowing a process called tunneling, i.e.. allowing passage through the barrier rather than over it. The mechanism of this neutralization process is proposed to involve a novel coherent wave structure that can occur between electrons under certain conditions. All of these wave processes have been observed in the past under various conventional conditions, but applying them to the cold fusion phenomenon has been a subject of debate and general rejection.

While the debate based on wave action has been underway, people have proposed other mechanisms. These include the presence of neutrons within the lattice. Normally, neutrons are unstable outside of the nucleus, decomposing into a proton, an electron, and a neutrino. Presumably, this reaction can be reversed so that neutrons might be created in a lattice containing many free electrons and protons. Having no charge, the neutron could then interact with various atoms in the lattice to produce energy. These neutrons might also be hidden in the lattice by being attached to other nuclei in a stabilized form, to be released when conditions were right. Several particles normally not detected in nature also have been proposed to trigger fusion and other nuclear reactions.

While search for a suitable mechanism has been underway, an understanding of the environment that triggers the mechanism has been sought, the so-called nuclear-active-environment. Initially, this environment was thought to exist in the bulk of the palladium cathode used in the Pons-Fleischmann method to produce cold fusion. It is now agreed that the nuclear reactions only occur in the surface region. Recent arguments suggest that this surface layer does not even require palladium for it to be nuclear-active. Nuclear reactions have now been produced in a variety of materials using many methods. The only common feature found in all of these methods is the presence of nano-sized particles of material on the active surface. If this observation is correct, four conditions seem required to produce the nuclear reactions. First, the particle must have a critical small size; second, it must contain a critical concentration of deuterium or hydrogen; third, it must be constructed of certain atoms; and fourth, it must be exposed to a source of energy. This energy can take the form of a sufficiently high temperature, a significant high flux of hydrogen through the particle, application of energetic electrons or charged particles, or application of laser light of the proper frequency. Until, the importance of these factors is understood, the effect will continue to be difficult to replicate.

Technical Introduction to LENR-CANR

Edmund Storms

At low energies, the Coulomb barrier prevents nuclei from coming together and fusing to form a single nucleus. To initiate a nuclear reaction, several methods are used. Nuclear reactions are normally initiated by pushing two atoms together with enough force to overcome the Coulomb barrier by brute force, or by using neutrons which penetrate the nuclei without seeing a barrier. (Neutrons have no electrical charge, so the Coulomb barrier does not stop them.) These forces are normally provided by high-temperature plasma or by accelerating ions to high energies. In contrast, LENR describes the mechanism and conditions that cause a variety of nuclear reactions to take place with relatively low activation energy. These unique conditions reduce the need for excessive energy. The normal method forces the nuclei together, while the new method encourages them to come together. The challenge has been to understand the unique characteristics of the necessary solid structure such that this structure could be generated at will.

Because the proposed method is unique, at odds with current nuclear theory, and is still difficult to reproduce, support for studies in many countries, but not all, has been very limited. Nevertheless, considerable information has accumulated over the last 13 years since Profs. Stanley Pons and Martin Fleischmann showed the world the possibilities inherent in this phenomenon. Much understanding is buried in conference proceedings and reports that are not available to a serious student. This information will, as time permits, be made available on this site. Students of the subject are also encouraged to use this site to interact with other people in the field and provide objective critiques of the work published here.

PHENOMENA DISCUSSED IN SOME OF THE PAPERS

At least 10 ways have been demonstrated to produce anomalous heat and/or anomalous elemental synthesis. A few of these methods will be described here. For course, not all of the claims are worthy of belief nor are they accepted by many people. Nevertheless, the claims will be described without qualifications in order to provide the reader with the latest understanding.

The most studied method involves the use of an electrolytic cell containing a LiOD electrolyte and a palladium cathode. Current passing through such a cell generates D+ ions at the cathode, with a very high effective pressure. These ions enter the palladium and, if all conditions are correct, join in a fusion reaction that produces He-4. Initially palladium wire and plate were used, but these were found to form microcracks, which allowed the required high concentration of deuterium to escape. Later work shows that the actual nuclear reaction occurs on the surface within a very thin layer of deposited impurities. Therefore, control of this impurity layer is very important, but rather difficult. The use of palladium is also not important because gold and platinum appear to be better metals on which to deposit the impurity layer. This method is found, on rare occasions, to generate tritium within the electrolyte and transmutation products on the cathode surface. Different nuclear reactions are seen when light water (H_2O) is used instead of D_2O, although the amount of anomalous energy is less when H_2O is used. These observations have been duplicated hundreds of times in dozens of laboratories.

Application of deuterium gas to finely divided palladium, and perhaps other metals, has been found to generate anomalous energy along with helium-4. Both palladium-black as well as palladium deposited as nanocrystals on carbon have shown similar anomalous behavior. In both cases the material must be suitably purified. Palladium deposited on carbon can and must be heated to above 200/260°C for the effect to be seen. When deuterium is caused to diffuse through a palladium membrane on which is deposited a thin layer of various compounds, isotopes that were not previously present are generated with isotopic ratios unlike those occurring naturally.

A plasma discharge under H_2O or D_2O between various materials generates many elements that were not previously present. When the electrodes are carbon and the plasma is formed in H_2O, the main anomalous element is iron. This experiment is relatively easy to duplicate.

Several complex oxides, including several superconductors, can dissolve D_2 when heated. When a potential is applied across a sheet of such material, the D+ ions are caused to move and anomalous heat is generated.

If deuterium ions, having a modest energy, are caused to bombard various metals, tritium as well as other elements not previously present are generated. These ions can be generated in a pulsed plasma or as a beam.

When water, either light or heavy, is subjected to intense acoustic waves, collapse of the generated bubbles on the surrounding solid walls can generate nuclear reactions. This process is different from the fusion reaction claimed to occur within a bubble just before it disappears within the liquid because neutrons are not produced in the former case, but are produced in the latter case. This method has been applied to various metals in heavy water using an acoustic transducer and in light water using a rotating vane which generates similar acoustic waves...

HOW TO EXPLAIN THE CLAIMS

A major problem in deciding which model might be correct is the absence of any direct information about the nature of the nuclear-active-environment. At this time, two important features seem to be important, the size of the nanodomain in which the reactions occur and the presence of a deuterium flux through this domain. The domain can apparently be made of any material in which hydrogen or deuterium can dissolve. Until the nature of the nuclear-active-state (NAS) is known, no theory will properly explain the effect and replication of the claims will remain difficult.

When fusion is initiated using conventional methods, significant tritium and neutrons are produced. In addition, when other elements are generated, they tend to be radioactive. This is in direct contrast to the experience using low energy methods. These products are almost completely absent and, instead, helium-4 is produced. When radiation is detected, it has a very low energy. This contrasting behavior, as well as the amount of anomalous energy, has made the claims hard to explain using conventional models. This difficulty has been amplified by a failure of many skeptics to recognize the contrasting effect of the environment, a plasma being used in the older studies and a solid lattice of periodic atoms being present as the new environment.

Over 500 models and their variations have been proposed, some of which are very novel and some are variations on conventional ideas. Most models attempt to explain the nuclear reaction once the required environment has been created, without addressing what that unique environment might be like. These models involve conversion of a proton (deuteron) to a neutron (dineutron), creation of an electron structure that is able to neutralize the barrier, conversion of deuterium to a wave which interacts without charge, and the presence of otherwise overlooked neutrons and/or novel particles. Many of the models will have to be abandoned or seriously modified once the nature of the nuclear active environment is understood.

It Started in 1989 . . .

Peter Hagelstein

Many of us recall the controversy surrounding the announcement of claims of observations of fusion reactions in a test tube that were made in 1989. At the time, these claims were greeted with considerable skepticism on the part of the physics community and the scientific community in general.

The principal claim of Pons and Fleischmann

The principal claim of Pons and Fleischmann in 1989 was that power was produced in palladium cathodes that were loaded electrochemically in a heavy water electrolyte. The evidence in support of this was a measured increase in the temperature in the electrochemical cell. There was no obvious evidence for nuclear reaction products commensurate with the claimed heat production. Fleischmann speculated that perhaps two deuterons were somehow fusing to He-4 through some kind of new mechanism.

Rejection by the physics community

This claim was not accepted by the physics community on theoretical grounds for several reasons:

First, there was no mechanism known by which two deuterons might approach one another close enough to fuse, since the Coulomb barrier prevents them from approaching at room temperature.

Second, if they did approach close enough to fuse, one would expect the conventional dd-fusion reaction products to be observed, since these happen very fast. Essentially, once two deuterons get close enough to touch, reactions occur with near unity probability, and the reaction products (p+t and n+He-3) leave immediately at high relative velocity consistent with the reaction energy released. To account for Fleischmann's claim, the proposed new reaction would seemingly somehow have to make He-4 quietly and cleanly, without any of the conventional reaction products showing up, and would somehow have to arrange for this to happen a billion times faster than the conventional reaction pathway. Most physicists bet against the existence of such a magical new effect.

Third, the normal pathway by which two deuterons fuse to make He-4 normally occurs with the emission of a gamma ray near 24 MeV. There was no evidence for the presence of any such high energy gamma emission from the sample, hence no reason to believe that any helium had been made.

Finally, if one rejects the possibility that any new mechanisms might be operative, then the claim that power was being produced by fusion must be supported by the detection of a commensurate amount of fusion reaction products. Pons and Fleischmann found no significant reaction products, which, given the rejection of new mechanisms, implied an absence of fusion reactions.

An alternate explanation is proposed

The physicists decided in 1989 that the most likely reason that Pons and Fleischmann observed a temperature increase was that they had made an error of some sort in their measurements. When many groups tried to observe the effect and failed, this led most of the physics community to conclude that there was nothing to it whatsoever other than some bad experiments.

The claim of Jones

A second very different claim was made at the same time in 1989 by Steve Jones. This work also involved electrochemistry in heavy water and the observation of reaction products corresponding to the conventional dd-fusion reactions. The initial publication showed a spectrum of neutron emission that Jones had detected from a titanium deuteride cathode loaded electrochemically. The response of the physics community was skeptical, as the signal to noise ratio was not particularly impressive. Given the polarization of the physics community in opposition to the claims of Pons and Fleischmann (which were announced essentially simultaneously), the physicists were not of a mood to accept much of any claims that fusion could happen in an electrochemical experiment at all. Jones went to great lengths to assure fellow scientists that his effect was completely unrelated to the claims of Pons and Fleischmann, and was much more reasonable.

Also rejected

Physicists had reason to be skeptical. Theoretical considerations indicated that the screening effects that Jones was relying on were not expected to be as strong as needed to account for the fusion rates claimed. As this experiment could not seem to be replicated by others at the time, it was easy for the physics community to reject this claim as well.

Cold fusion, weighed and rejected with prejudice

Cold fusion, as the two different claims were termed, was dismissed with prejudice in 1989. The initial claims were made near the end of March in Utah, and the public refutation of the claims was made at the beginning of May. It only took about 40 days for the physics community to consider the new claims, test them experimentally, and then announce loudly to the world that they had been carefully weighed and rejected.

Following this rejection, physicists have treated cold fusion rather badly. For example, Professor John Huizenga of Rochester University was selected to be co-chair of the DOE ERAB committee that met to review cold fusion and issue a report. Shortly afterward, he wrote a book entitled Cold Fusion, The Scientific Fiasco of the Century, in which he discusses the claims, the experiments, and the extreme skepticism with which the new claims were greeted. Robert Park discusses the subject in his book entitled Voodoo Science. You can find many places where physicists and other scientists happily place the cold fusion claims together with claims of psychic phenomena.

A Science Tutorial

Talbot Chubb

First it is important to recognize that there are four distinct types of energy production:
1) chemical energy, that powers our cars and most of our civilization;
2) nuclear fission energy, as used to generate about 15% or our electricity;
3) hot fusion nuclear energy, which powers the sun and most stars;
4) cold fusion nuclear energy, which appears as unexplained heat in a few experimenter's laboratory studies and which most scientists believe is impossible.

The three types of nuclear energy produce 10 million times as much heat per pound of fuel than occurs with chemical energy. How do these types of energy differ? To understand this question you need to know some chemistry and physics.

Lesson 1

Nature has provided us with two types of stable charged particles, the proton and the electron. The proton is heavy, normally tiny, and has a positive charge. The electron is light, normally large and fuzzy, and has a negative charge. The positive charge and the negative charge attract each other, just like the north pole of a magnet attracts the south pole of a magnet. When you bring two magnets together with the north pole of one facing the south pole of the other, they pull together, bang! When they bang into each other they release a little bit of energy in the form of heat, but it is too small an amount to easily measure. To pull the magnets apart you have to do work, which is another way of saying you have to use up energy. It's almost like pulling a rock back up a hill. Rolling the rock down a hill actually creates a little heat, and pulling the rock back up the hill takes energy. In the same way the positive charge of the proton pulls on the negative charge of the electron and they stick together releasing energy in the process. The result is a hydrogen atom, designated H. A hydrogen atom is nothing but a fuzzy electron hugging a compact proton. The proton is the nucleus of the hydrogen atom. If you knock the electron off the hydrogen atom you get a positive ion H+, which is nothing more than the original proton. An ion is the name applied to an atom or molecule that has lost or gained one or more electrons, hence is no longer electrically neutral.

Lesson 2

As you know, nature has provided us with more than one type of atom. We have oxygen atoms, nitrogen atoms, iron atoms, helium atoms, etc.. How do these atoms differ? The answer is that they all have different types of nuclei (plural of nucleus, from the Latin). And these different nuclei all have different numbers of protons inside them, which means they all have different plus charges. The nucleus of the helium atom has 2 protons inside it, hence has plus 2 charge, and requires 2 electrons to neutralize its charge. When 2 electrons stick to it, it becomes a helium atom. The oxygen nucleus has 8 protons and has charge 8. When 8 electrons stick to it, it becomes an oxygen atom. The nitrogen atom has 7 electrons, and the iron atoms something like 26. But all the atoms are built more or less the same way, with a compact positively charged nucleus embedded in a cloud of fuzzy electrons. The difference in size between the compact nucleus and the fuzzy electrons is enormous. The sun has a diameter only about 100 times that of the earth. The electron cloud on an atom has a diameter which is about 100,000 times that of the nucleus. Cube these numbers to get the difference in volumes.

Lesson 3

We now are in a position to understand what chemical energy is. The atoms, all electrically neutral, can actually join with each other and release more energy. This is another way of saying that they can join into more stable configurations. The electrons in an atom try to configure themselves so as to get as close as possible to their nucleus, but their fuzzy nature requires that they take up a certain volume of space. However, if they join together with the electrons of another atom they can usually find a tighter configuration that leaves them closer to their beloved beginning of May. It only took about 40 days for the physics community to consider the new claims, test them experimentally, and then announce loudly to the world that they had been carefully weighed and rejected.

Following this rejection, physicists have treated cold fusion rather badly. For example, Professor John Huizenga of Rochester University was selected to be co-chair of the DOE ERAB committee that met to review cold fusion and issue a report. Shortly afterward, he wrote a book entitled Cold Fusion, The Scientific Fiasco of the Century, in which he discusses the claims, the experiments, and the extreme skepticism with which the new claims were greeted. Robert Park discusses the subject in his book entitled Voodoo Science. You can find many places where physicists and other scientists happily place the cold fusion claims together with claims of psychic phenomena.

A Science Tutorial

Talbot Chubb

First it is important to recognize that there are four distinct types of energy production:
1) chemical energy, that powers our cars and most of our civilization;
2) nuclear fission energy, as used to generate about 15% or our electricity;
3) hot fusion nuclear energy, which powers the sun and most stars;
4) cold fusion nuclear energy, which appears as unexplained heat in a few experimenter's laboratory studies and which most scientists believe is impossible.

The three types of nuclear energy produce 10 million times as much heat per pound of fuel than occurs with chemical energy. How do these types of energy differ? To understand this question you need to know some chemistry and physics.

Lesson 1

Nature has provided us with two types of stable charged particles, the proton and the electron. The proton is heavy, normally tiny, and has a positive charge. The electron is light, normally large and fuzzy, and has a negative charge. The positive charge and the negative charge attract each other, just like the north pole of a magnet attracts the south pole of a magnet. When you bring two magnets together with the north pole of one facing the south pole of the other, they pull together, bang! When they bang into each other they release a little bit of energy in the form of heat, but it is too small an amount to easily measure. To pull the magnets apart you have to do work, which is another way of saying you have to use up energy. It's almost like pulling a rock back up a hill. Rolling the rock down a hill actually creates a little heat, and pulling the rock back up the hill takes energy. In the same way the positive charge of the proton pulls on the negative charge of the electron and they stick together releasing energy in the process. The result is a hydrogen atom, designated H. A hydrogen atom is nothing but a fuzzy electron hugging a compact proton. The proton is the nucleus of the hydrogen atom. If you knock the electron off the hydrogen atom you get a positive ion H+, which is nothing more than the original proton. An ion is the name applied to an atom or molecule that has lost or gained one or more electrons, hence is no longer electrically neutral.

Lesson 2

As you know, nature has provided us with more than one type of atom. We have oxygen atoms, nitrogen atoms, iron atoms, helium atoms, etc.. How do these atoms differ? The answer is that they all have different types of nuclei (plural of nucleus, from the Latin). And these different nuclei all have different numbers of protons inside them, which means they all have different plus charges. The nucleus of the helium atom has 2 protons inside it, hence has plus 2 charge, and requires 2 electrons to neutralize its charge. When 2 electrons stick to it, it becomes a helium atom. The oxygen nucleus has 8 protons and has charge 8. When 8 electrons stick to it, it becomes an oxygen atom. The nitrogen atom has 7 electrons, and the iron atoms something like 26. But all the atoms are built more or less the same way, with a compact positively charged nucleus embedded in a cloud of fuzzy electrons. The difference in size between the compact nucleus and the fuzzy electrons is enormous. The sun has a diameter only about 100 times that of the earth. The electron cloud on an atom has a diameter which is about 100,000 times that of the nucleus. Cube these numbers to get the difference in volumes.

Lesson 3

We now are in a position to understand what chemical energy is. The atoms, all electrically neutral, can actually join with each other and release more energy. This is another way of saying that they can join into more stable configurations. The electrons in an atom try to configure themselves so as to get as close as possible to their nucleus, but their fuzzy nature requires that they take up a certain volume of space. However, if they join together with the electrons of another atom they can usually find a tighter configuration that leaves them closer to their beloved nuclei. For example, 2 hydrogen atoms can join together into a more compact configuration if each hydrogen atom contributes its electron to a 2-electron cloud, which the separate protons share. In this manner they form a grouping of the 2 electrons in a single cloud, together with the 2 isolated protons spaced apart from each other but still within the electron cloud. The result is a heat-producing chemical reaction $H + H => H_2$. (The => means "goes to" or "becomes".) The H_2 configuration is the hydrogen molecule, and when you buy a tank of hydrogen gas, H_2 molecules is what you get. Furthermore, the 2 electrons of the H_2 molecule and the 8 electrons of the O atom can find a still more compact configuration by combining their electrons to create the water molecule H_2O, plus heat. The water molecule is really a single cloud of electrons in which are embedded the three point-like nuclei to form a minimum energy configuration. So when we burn oil or coal we reconfigure the electrons to produce more stable configurations of point-like nuclei embedded in electron clouds, liberating heat. So much for chemical energy.

Lesson 4

We have slid over one point in the above discussion. How does Nature make a nucleus containing two or more protons in the first place. After all, each of the protons has a positive charge, and the positive charges repel each other very strongly when they are separated by a tiny distance, equal to the distance across a nucleus. The repulsion of like charges is just like the repulsion between the north poles of two magnets when they are pushed together the wrong way. Something must overcome this repulsion, or else the only kind of atoms we would have would be those of hydrogen. Fortunately, this is not what we observe. The answer is that there is a second kind of force which acts on protons. This is the nuclear force. The nuclear force is very strong but requires particles to almost sit on each other to have any effect. Also, there is a second kind of heavy particle, which is just like a proton, except that it has no positive or negative charge. It is not pushed away by the proton's plus charge. This other kind of particle is called the neutron, since it is electrically neutral. A peculiar fact of life is that it exists in stable form only inside a nucleus. When not in the nucleus it changes into a proton, an electron and a very light anti-neutrino in about 10 minutes. But it lasts forever inside a nucleus. Anyway, the neutron and the proton very strongly attract each other once they get close enough together, and then they combine to form a highly stable pair called a deuteron, which we designate D+. The single deuteron, when it combines with a single electron, forms the heavy hydrogen atom called deuterium, designated D. A second nuclear reaction occurs when two deuterons make contact. When they can be forced together so as to make contact, the 2 deuterons fuse, making a doubly charged particle. The grouping of 2 protons and 2 neutrons is even tighter than the proton-neutron grouping in the deuteron. The new particle, when neutralized by 2 electrons, is the nucleus of the helium atom, designated He. Larger groupings of neutrons and protons exist in nature and serve as the nuclei of carbon, nitrogen, oxygen, and iron, etc. atoms. All of these groupings are made possible by the very strong nuclear force, which is felt between particles only when they are in contact or share the same nucleus-size volume of space.

Lesson 5

We can now understand normal nuclear energy, which is really nuclear fission energy. During the early history of the universe massive stars were formed. In the explosion of these massive stars, lots of different types of nuclei were formed and exploded back into space. Second and later generation stars and planets were formed from this mix, including the sun. In the explosion process probably every possible stable configuration of protons and neutrons was produced, plus some almost-stable groupings, such as the nucleus of the uranium atom. There are actually 3 different types of uranium atom nuclei, called uranium-234, uranium-235, and uranium-238. These "isotopes" differ in their number of neutrons, but they all have 92 protons. The nuclei of all uranium atoms can go to a lower energy configuration by ejecting a helium nucleus, but this process occurs so rarely that the Earth's uranium has already lasted over 4 billion years. But the uranium nuclei are unstable in another way. In general, groupings of protons and neutrons are happiest if they have about 60 protons-plus-neutrons. The uranium nuclei contain more than three times this number. So they would like to split in two, which would release a lot of heat. But nature doesn't provide a way for them to split apart. They have to first go to a higher energy configuration before splitting in two. However, one of the three forms of uranium nucleus found in nature called uranium-235 and designated ^{235}U, gains the needed energy if it captures a neutron. The energized nucleus that results from neutron capture then splits apart with the release of an enormous amount of energy, and incidentally with release of additional neutrons. The additional neutrons can then split more uranium-235 nuclei, keeping the reaction going. This is what happens in nuclear power plants, where the heat, which is the end product of the nuclear splitting process, is used to boil water, generate steam, and turn electrical generators. (One also gets lots of radioactive products, which are a nuisance to dispose of safely.)

Lesson 6

We are now also in a position to understand hot fusion nuclear energy. As mentioned in lesson 5, the groupings of protons plus neutrons is most stable when the numbers of neutrons and protons approximate those found in the nucleus of an iron atom. Just as uranium has too many neutrons plus protons to be comfortable, so the light elements like hydrogen, helium, carbon, nitrogen and oxygen have too few. If the nuclei can be made to make contact under proper conditions, they can combine to create more stable groupings, plus heat. This is the process of fusion. Nature has found a way of doing this in stars like the sun. All Nature has to do is heat compressed hydrogen hot enough and wait long enough and hot fusion will occur. If Nature were to start with deuterium, which already has a paired proton and neutron, the task would be relatively easy in a star. Temperature is a measure of how much speed an atom of a given type has as it bangs around inside a cloud of such atoms. The higher the temperature, the higher the speed and the closer the atoms get to each other momentarily during a collision. In a star the temperatures are high enough that all the electrons quickly get knocked off the atoms, so one is really dealing with a mixed cloud of electrons and nuclei. At very high temperature the nuclei occasionally get close enough during collisions for the pulling-together short range nuclear force to turn on. Then the nuclei can stick together and go to a lower energy grouping of protons plus neutrons, releasing heat. Hot fusion nuclear energy is an attempt to carry out this process in the lab, using deuterium and mass-3 hydrogen (whose nucleus is a compact grouping of 1 proton and 2 neutrons) as the gas. Hot fusion requires that the gas be contained at temperatures of hundreds of millions of degrees, which can be done with the help of magnetic fields, but only for 1 or 2 seconds. The hope is to contain the gas for longer times. During the period of high temperature containment nuclear reactions occur during collisions. The main form of energy release is ejection of high energy neutrons and protons. The proton energy quickly converts to heat. The neutron energy can also be converted to heat but makes the equipment highly radioactive. It then becomes difficult to repair the equipment, which could make hot fusion a poor candidate for commercial power production. In any case hot fusion power is a dream that is still probably at least 50 years away. But most scientists view hot fusion as the only way to achieve fusion power. Hot fusion produces less radioactivity than fission power, is environmentally benign, and has a virtually limitless fuel supply on earth (many millions of years at present energy usage rates).

Lesson 7

So now we come to cold fusion. Cold fusion may provide an easier and non-radioactive way of releasing nuclear fusion energy. Cold fusion relies on a different way of letting the protons and neutrons in one nucleus make contact with those in another nucleus, so that the nuclear force can bring them into a more stable configuration. The requirement for any nuclear reaction to occur is that the reacting nuclei occupy the same volume of space. This condition is called particle overlap. In hot fusion particle overlap is brought about briefly by banging the nuclei together so as to overcome momentarily the repulsion of the two positive charges which try to keep the particles apart. In cold fusion particle overlap conditions are achieved by making deuterium nuclei act as fuzzy objects like electrons in atoms, instead of like tiny points. When either light or heavy hydrogen is added to a heavy metal, each hydrogen "atom" occupies a position inside the metal where it is surrounded by heavy metal atoms. This form of hydrogen is called interstitial hydrogen. With interstitial hydrogen the electrons of the hydrogen atom become part of the pool of electrons of the metal. Each hydrogen nucleus oscillates back and forth through a negatively charged electron cloud provided by the electrons of the metal. They can be thought of as moving back and forth like the pendulum in a grandfather clock. This vibration exists even at very low temperature, due to a peculiarity of a branch of physics called quantum mechanics. The vibration is called zero point motion. The nucleus then becomes a fuzzy object, like the electrons in an atom. But this amount of fuzziness is not enough to permit a hydrogen nucleus to make contact with another hydrogen nucleus. To get two or more hydrogen nuclei to share the same volume one must go one step further. In a metal electrical current is carried by electrons that act more like vibrating matter waves than like point particles. If electrons did not become wave-like inside solids, there would be no transistors and no present day computers. This wave-like kind of electron is called a Bloch function electron. The secret of cold fusion is that one needs Bloch function deuterons. One needs wave-like deuterons inside or on the surface of a solid in order that two or more deuterons share the same volume of space. But once the Bloch function deuterons are created, the nuclear force comes into play and the protons and neutrons making up the deuterons can rearrange themselves into the more nuclearly stable Bloch function helium configuration, with release of heat. To study cold fusion the experimenter has to force deuterons to assume the wave-like form and keep them in the wave-like state. Cold fusion experiments demonstrating release of excess heat show that this can be done. But at present no one knows how to do it reliably. Since cold fusion promises millions of years of energy without the problems of global warming or radioactivity, a real effort should be made to learn how.

For more tutorials go to www.hometown.aol.com/cffuture1 and www.hometown.aol.com/cffuture2

LENR (Low Energy Nuclear Reactions) Experiments

Review from
http://www.lenr-canr.org

Have you ever wondered what a physics laboratory looks like? They are seldom spacious or organized the way they are shown in movies. Most LENR researchers work at universities or home laboratories, with tight budgets in a crowded space. They keep old, broken equipment on shelves to scavenge parts for new experiments. In this section we present some photographs of equipment provided by researchers, and close up pictures of equipment. The actual cells, cathodes and other equipment used in electrolysis experiments often have an ad-hoc, home-made appearance, because they are made by hand. They have to be; they are unique, one-of-a kind prototypes. Nothing quite like them has ever been made before.

A visitor seeing a LENR experiment the first time may feel disappointed. It looks like any other electrochemical experiment. The heat or neutron flux produced by the experiment are so small they can only be detected with sensitive instruments. A null cathode that produces no effect looks exactly like an active cathode. The difference between one cathode and another is in the microscopic structure, or the traces of elements mixed in with the palladium. Only one kind of cold fusion looks dramatic i. e. the glow discharge reaction.

Here are a few photographs of cold fusion cells and devices (also see the cover page).

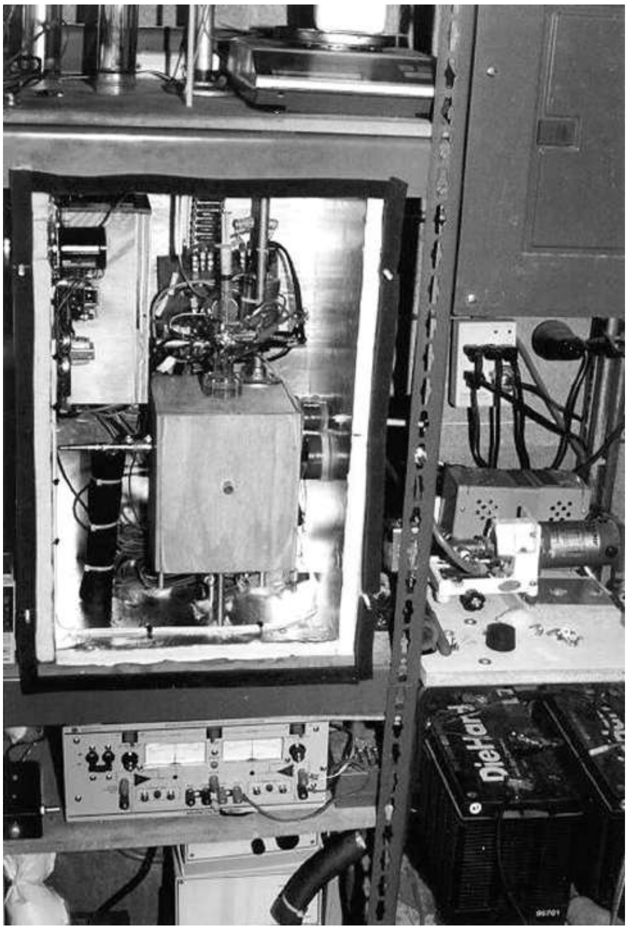

Fig.1
Box Calormeter

On Fig. 1 there is a calorimeter constructed by Edmund Storms. Note the DieHard® battery, lower right, that serves as an unninterruptable power supply. A power failure can ruin an experiment. Whenever possible, inexpensive, ordinary materials and instruments are used. However, experiments are never cheap, and they cannot be done on a shoestring. The equipment, arranged for another experiment (see Fig. 2), costs about $40,000.

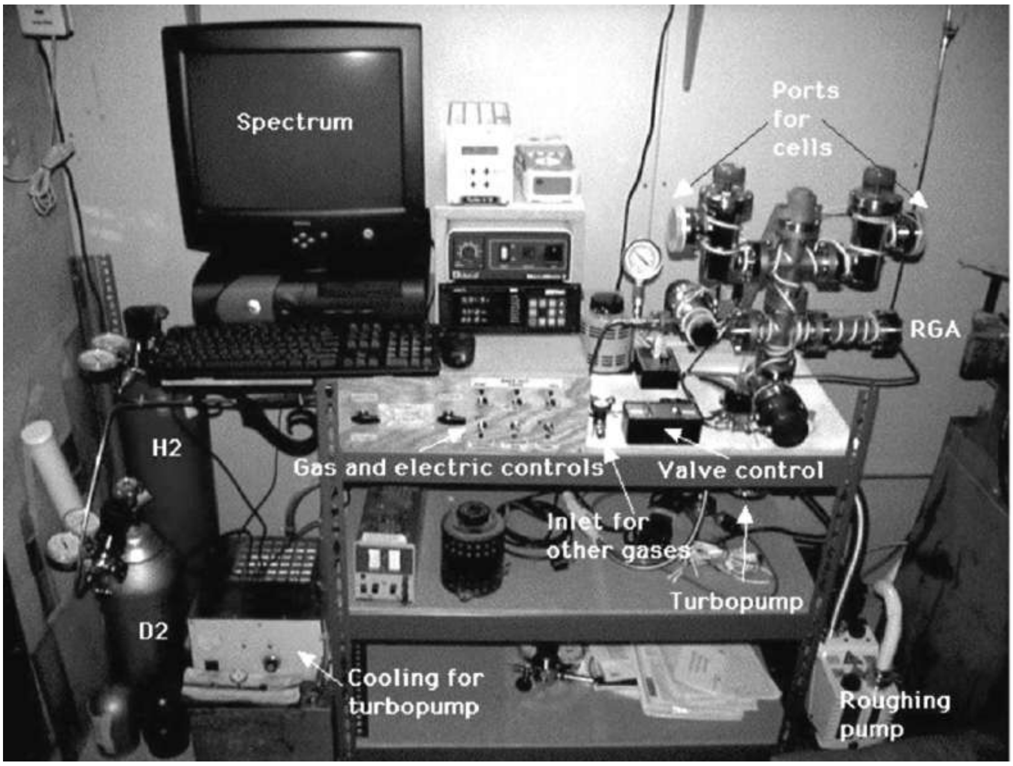

Fig.2 (On the left)
Vacuum system to prepare particles for gas loaded cold fusion cells, courtesy E. Storms.

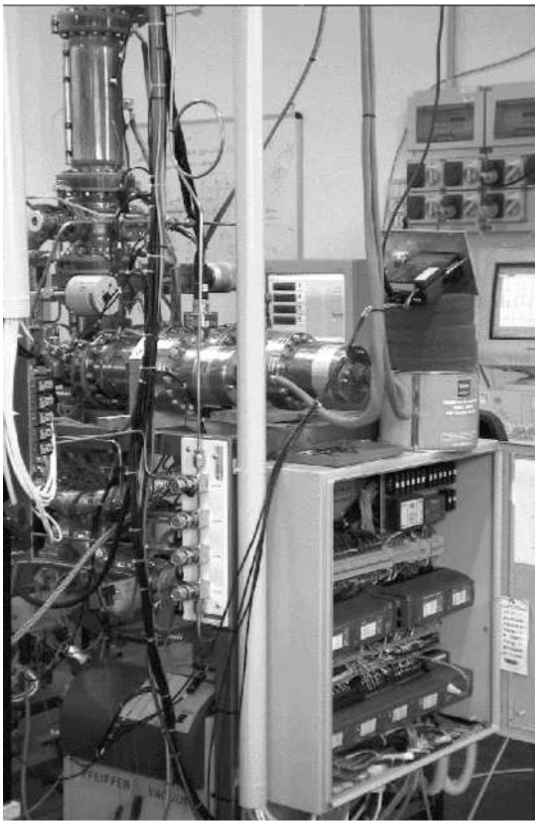

Fig.3

A high resolution mass spectrometer used for on-line helium detection during a cold fusion experiment at C. R. ENEA Frascati. (http://www.frascati.enea.it/nhe/)

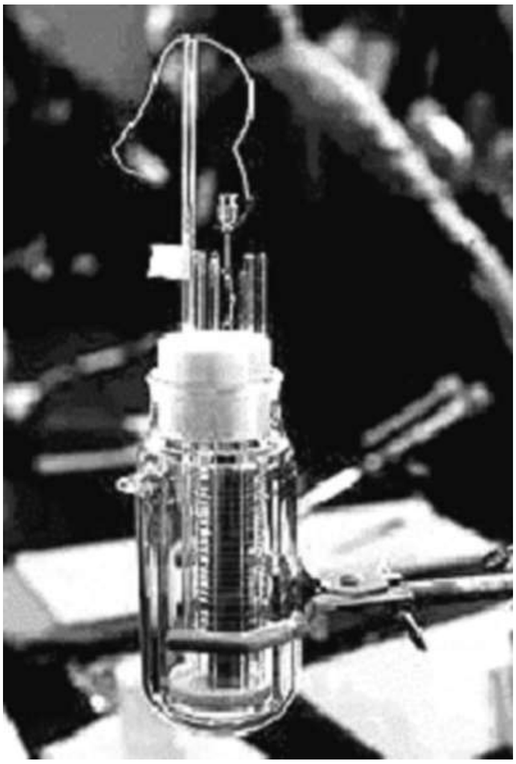

Fig.4

A cell at ENEA Frascati

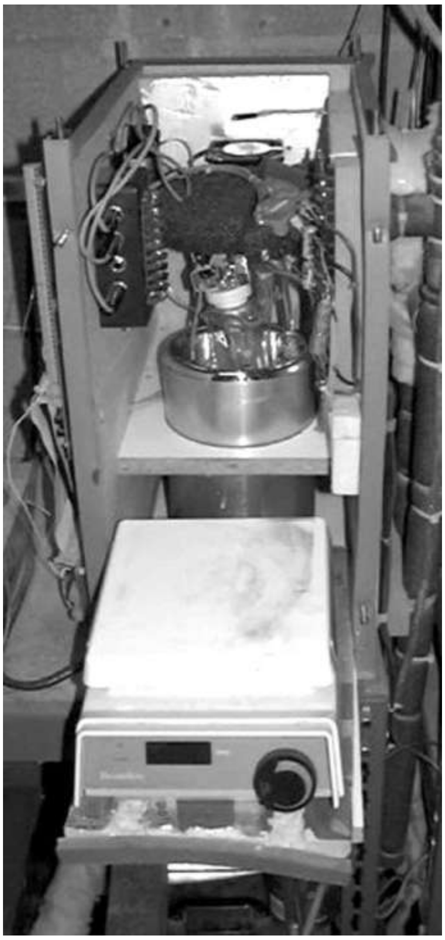

Fig.5

A flow-type cell, courtesy E. Storms

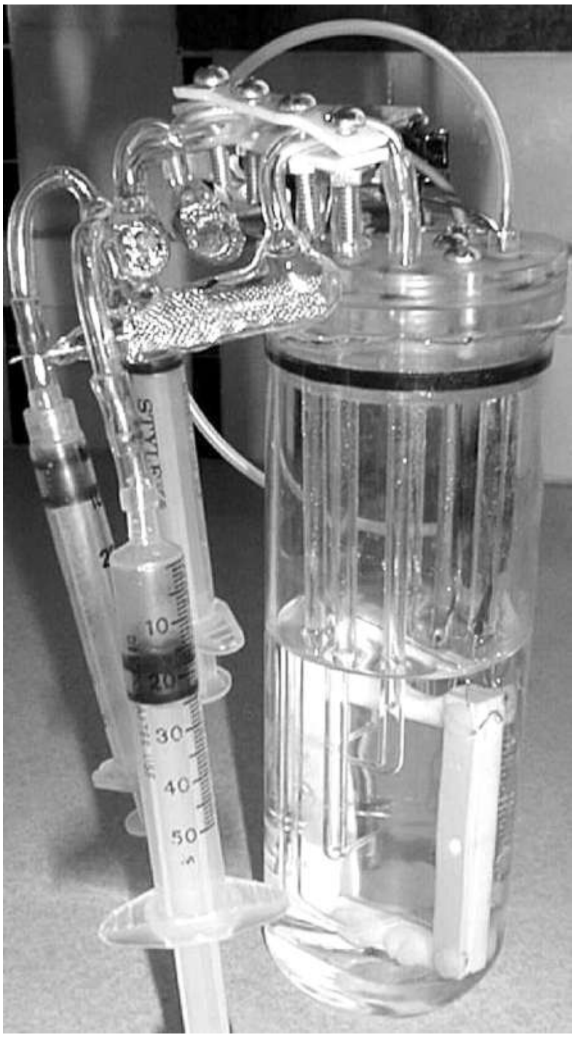

Fig.6

Close up of a Miley-style cell, courtesy E. Storms

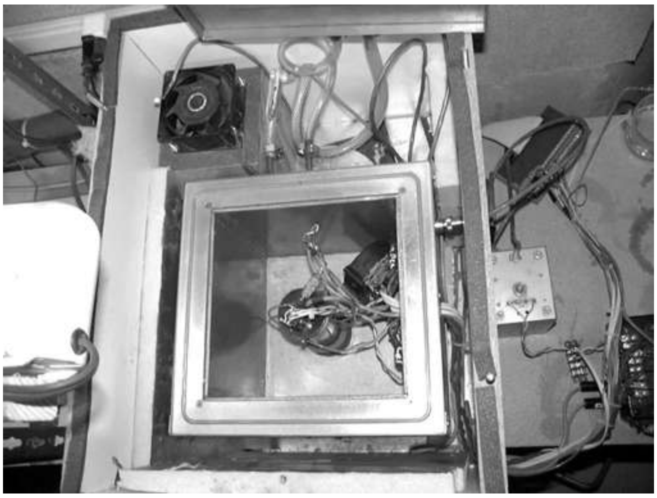

Fig.7
A cell installed inside a Thermonetics Seebeck calorimeter with the lid removed, courtesy E. Storms

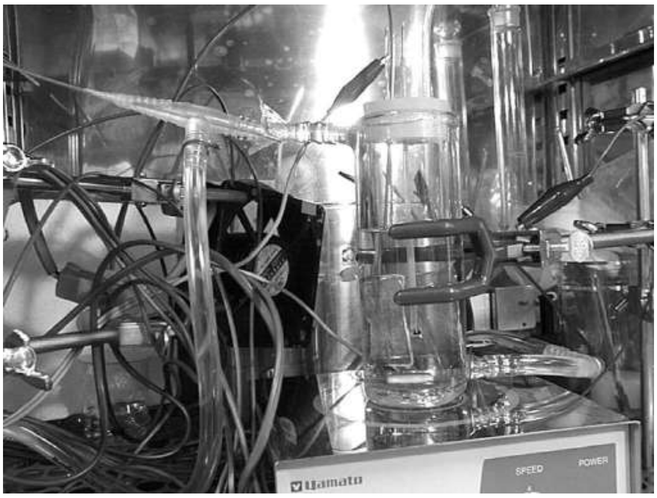
Yamato
SPEED
POWER

Fig.8

On the Fig. 8 you can see a glow discharge electrochemical cell at Hokkaido University, courtesy T. Mizuno. The cell is installed inside a crowded constant temperature air-cooled chamber. It placed on a magnetic mixer. Cooling water is pumped through the plastic tubes attached to the top and bottom. The muffin fan at the back circulates the air in the chamber

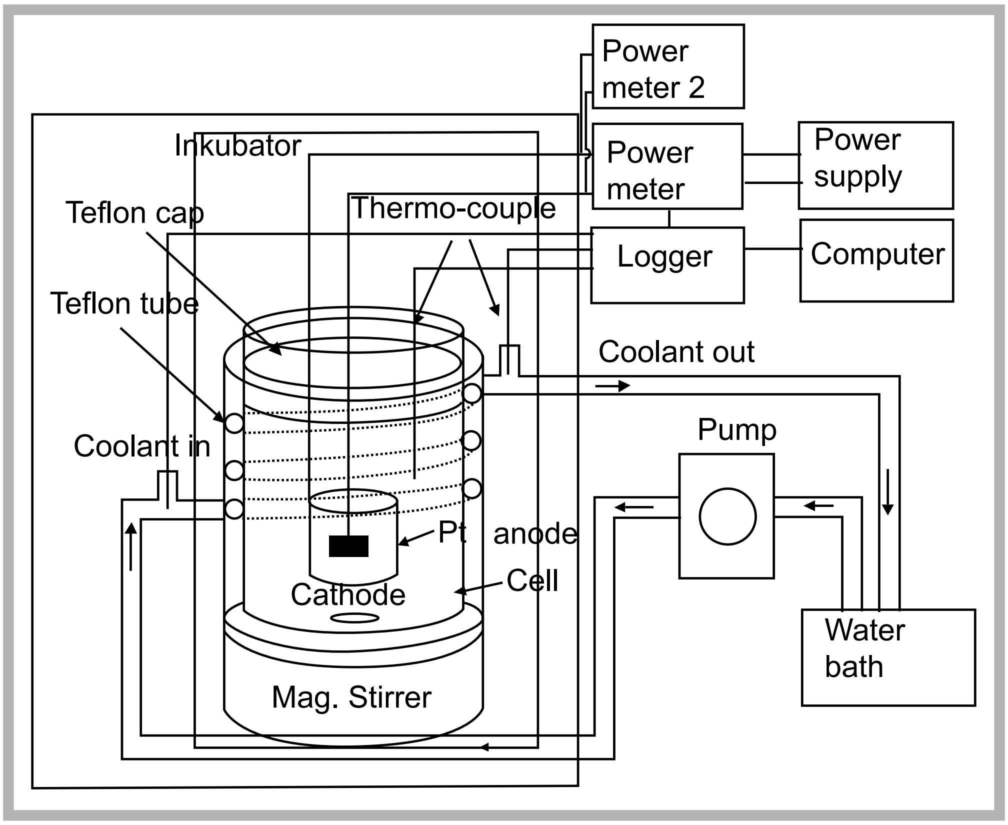
Power meter 2
Inkubator
Power meter
Power supply
Teflon cap
Thermo-couple
Logger
Computer
Teflon tube
Coolant out
Pump
Coolant in
Pt
anode
Cathode
Cell
Water bath
Mag. Stirrer

Fig.9

A schematic of the calorimeter shown above.

Hydrogen power engineering

Faraday Laboratories Ltd, Moscow, and Spectrum Investments Ltd, London, started joint R&D project on hydrogen power engineering. Photo: Alexander V. Frolov, Faraday Laboratories Ltd and Nicholas Moller, Spectrum Investments Ltd. The project includes designing and building of prototype to use hydrogen recombination process for heat generation.

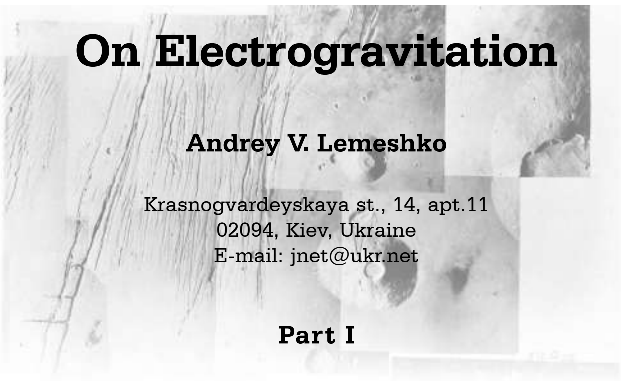

On Electrogravitation

Andrey V. Lemeshko

Krasnogvardeyskaya st., 14, apt.11
02094, Kiev, Ukraine
E-mail: jnet@ukr.net

Part I

"let's go to the Mars!"

Nowadays it becomes evident that exploration of deep space by means of rocket equipment is impossible. Aircrafts, which run on jet thrust, are too expensive and unreliable. Besides, they are too slow-speed if were considered on a scale of solar system. For example, the flight to the remote areas of our planetary system can drag on tens years. But the main disadvantage of rockets is their low load-carrying capacity. Thus, the most powerful transport system "Progress" can deliver less than 100 ton of load to the orbit of the Earth, and only 1 ton to the orbit of the Mars. Rather natural question arises as to whether any alternative ways of space exploration exist? Are there any transport systems, which can compete with rocket equipment on the interplanetary routs of future? Actually, at present there is only one technology, which is alternative to rocket ones – it is the so-called light sailers, using light pressure for acceleration. Laser and maser can be also used for their acceleration from satellites. However, there are also some disadvantages here.

There is also one more transport system, which is undeservedly consigned to oblivion. It uses electricity in acceleration aggregate. This transport system or the electric propulsion system is a result of researches and investigations of two little-known, but extremely talented American physicians and inventors, Thomas Townsend Brown and Paul Alfred Biefeld (the last one was also a professor of Astronomy).

Thomas T. Brown was the first who discovered electric principle of this propulsion system. Thus, in twenties of the XX century Thomas T. Brown, being the student of secondary school, became interested in X-rays. He wanted to reveal if the rays, outgoing the X-ray unit, could take positive effect. For his researches the inventor had got Coolidge Tube, the device, which radiates X-rays and consists of a glass tube. Deep vacuum is created in this glass tube. Anode and cathode are placed in the tube. Cathode radiates electrons and anode brakes them. X-rays are created at striking of electrons against anode. Brown used Coolidge Tube in the way nobody had done before. For ease of handling the researcher suspended the device on wires, which run to anode and cathode parallel to the Earth. During the operation of the device he noticed that every time when current fed to the tube, the tube was moved aside and tended to make slight progressive motion. Subsequent experiments demonstrated that the deviation value depended only on the value of voltage, which is between anode and cathode. The higher voltage is, the more deviation. X-rays were neither here not there. Much later at participation of Paul A. Biefeld, the same tendency to move was revealed at plane disc capacitors. The reason, which causes this motion, is the same, i.e. voltage, which exists between plates and not between anode and cathode as in Coolidge Tube.

Having run series of experiments, the researchers proved that electric energy could be directly transformed to mechanical. They also devised the physical principle, which were included in the scientific annals as Biefeld-Brown effect.

Editor's: Let us note that earlier Ampere and Faraday have been working on Electrokinetic effects which are easily detected at strong current. Modern researches on this point are known as "Sigalov's effect" (force appears in electric conductor which is bent at angle and at current impulse).

The essence of this effect lied in the fact that charged disc electric capacitors tend to move in the direction of their positive pole. At formulating of the essence of this effect the scientists consciously did not mention Coolidge Tube in order to demonstrate that according to their experimental observations, electron streams were not involved in generation of electromotive force. This effect of the direct transfer of electric energy to mechanic one (Biefeld-Brown effect) is observed, when current is fed to the Coolidge Tube or to the plane disc capacitor.

However, all this was realized much later, when in 1923 Brown became a student of Dennis University in Greenville, Ohio, where he met his teacher and later on a coauthor, Paul A. Biefeld. In school time the inventor came to the conclusion that the reopened by him phenomenon has nothing in common with X-rays and that high voltage, used for rays generation, is the basis of this phenomenon. Besides, Brown designed a device named "Gravitor". It looked like a usual bakelite box and was made as a simple construction, which consisted of several flat aluminum plates. The plates were placed as coins in a rouleau and separated with dielectric. It had only to place the box on the scale and connect it to some energy source with 100 kV voltage as the device either lost or put on about one percent of its weight that depended on polarity. An outside observer, who was not familiar with the essence of this effect, could think that mass was really changed. *(Editor's: Weight but not the mass is changed here. Weight is the force. There is a possibility of partial compensation of the weight by another force.)* But this was not the case. For better understanding of the processes, which take place at gravitor operating, let us make a mental experiment. Let us take a weight, put it on the scale and then raise it a little. The scale will certainly show that the weight has become lighter. After that let us press the weight then the scale will show that it has become heavier. However, in fact mass does not change. In the same way motive force, caused by high voltage, slightly raises the gravitor and at changing of polarity presses it. At that it seems that there is some change in weight. Thus, as if by chance, the scientist has discovered effect of direct transition of electric energy into mechanical one. *(Editor's: Let us specify the author's terminology. In T.T. Brown's Patent of 1927 it was mentioned that there* *was generated a force which partially compensated weight. The matter was always about the generation of x-force while gravitational field remained unchangeable).*

Having graduated from the University, from 1926 up to 1952 Brown already by himself was being occupied with development of his electrical propulsion system. At the same time whether in chase of fame or in hope to obtain money for his researches, T.T. Brown began asserting that he had discovered something new in physics, viz "electrogravitation", i.e. some connecting effect between gravitation and electricity. *(Editor's: In my opinion, T.T. Brown has never posed a problem in such a way. The matter always concerned the fact that X-force appears while gravitational field does not change.)* As for the development of the invention, it came to the decrease of weight of the device and increase of DC voltage, which fed the plates of the "gravitor". As a result the devices could fly and lift weight which greatly exceeded their own. Thus in 1953 in his laboratory Brown demonstrated the flight of aircraft which had a disk form. The device made 51 m/c (180 km/h), flying at circular course of 6 m diameter. This device was nothing but very lightweight capacitor in which by means of the special construction electromotive force was redirected in two planes upward and sideward. It made the device rise in air and fly around the mast. Current of 50kWtt voltage was supplied to the central mast by the wire. Then speed of "flying disks" became higher, the payload capacity was increased but that was again resulted by decrease of weight and increase of the supplied voltage. T.T. Brown so succeeded in decreasing of the thickness of capacitor plates and in making his devices superlight, that the witnesses of these demonstrations called these devices as "air films".

Later on all works in this area were placed on secret list and then closed because of lack of prospects. Thus, accounts and experiments demonstrated that it was impossible to overcome terrestrial attraction entirely and go to outer space by means of the gravitor because there are no such natural materials which could take huge electric and heat loads. Besides, sufficiently massive electric power station would be needed to produce required currents. *(Editor's: Thereupon the following question appears: why the author looks for "currents" while the essence of the technology is in applying of high voltage but not of the currents?!)* Moreover some alternative was made, i.e. jet engines. Gravitor could not compete with them by above-mentioned and quite objective reasons. In time, since the mechanism of transition of electric energy into mechanical one was not disclosed and Brown's populist idea has received some publicity, this effect became surrounded with idle talks and guess-works. However the scientists who researched this very effect, has developed only three theory wh ich could explain its existence.

The first theory was suggested by T.T. Brown, the discoverer of this effect. To his dying day Brown asserted that he had discovered the effect which could connect gravitation and electricity, i.e. "electrogravitation". But this theory can be easily refuted by practice. It is enough to put Brown's bakelite box to the scale in such a way as to place the plates of the "gravitor" at right angle to the surface of the scales. Then poles of the "gravitor" will be placed at the same level and parallel to ground and as a consequence electromotive force will influence in no way on the scale, since this force is directed sideward but not upward or downward. The scale will accurately show that there are no changes in weight and hence there is no **electrogravitation**. *(Editor's: The author will obtain propulsive force and it is directed not upward and downward but sideward. However the effect exists in any case. The changing of weight appears only when propulsive force generated by the device is directed against the weight force, i.e. upright.)*

The second theory was suggested by T. Brown's opponents. They asserted that electromotive force appears as a result of the so called "electron wind", i.e. electron streams which exist between the plates of the "gravitor". In their argumentations Brown's opponents rest upon the following facts which are well-known to every physicist. Fast electrons, as well as photons, have dual properties, i.e. interacting with matter and environment, they behave both as a wave and as a material particle. Since light or photons carry energy, then they heat bodies which are in their way and put pressure on the illuminated surface etc. Similarly electron streams carry heat or kinetic energy and hence they can put pressure on the anode of Coolidge Tube and get it moving forward. This theory seems to be interesting, however it explains appearance of electromotive force only in Coolidge Tube, but there is no explanation why discoid capacitors tend to move. Point is that there are no free electron streams in the capacitors. Usually between the plates of capacitors there is a dielectric which brings electron streams to nothing. If there is no electron wind, then what makes capacitors move?

Editor's: There is always a surface charge and reactive outflows of ions. But it is just a partial explanation of Biefeld-Brown's effect.

The third theory is based on the following fact. It is well known that electrons can have heat or Brownian chaotic motion. Free electrons, placed in a conductor, are in the similar chaotic heat motion before voltage is applied to this conductor. The higher temperature of the environment is the higher speed of this heat motion. According to school course of physics, heat energy or energy of chaotic motion of molecules (electrons) can be entirely transferred into mechanical one. If all molecules (electrons) start to move simultaneously and in the same direction, and if they push some piston toward this direction then heat energy of molecules (electrons) will be entirely transferred into mechanical work. In other ways, this piston will move until molecules (electrons) calm down. It is considered that this process is unrealizable in practice.

However, it seems that this is not quite the case, i.e. it is possible to regulate chaotic heat motion of electrons. Obviously this process takes place in the capacitor and in Coolidge Tube. It is no secret that when negative electrons try to move to the neighbor positive charged electrode, then dielectric layer which is placed between the plates of the capacitor prevents it. Electrons rest upon this layer and start to heat. Their internal energy is also increasing. The higher voltage is the higher heating of electrons, i.e. their capacity to have chaotic heat motion increases. However, chaotic motion of electrons is not observed on the capacitor plates, since the directed force (voltage) influences electrons. Electrons tend to move upward, downward, to the right and to the left, but voltage prevents it. Thus they just are forced to move in space in one direction and with the same speed. Capacitor or gravitor has to move alongside with electrons because plates of the capacitor (gravitor) become a piston which is pushed by heated electrons. This piston is in progressive motion. It starts to move under the influence of electrons, or rather together with electrons. The higher voltage is, the higher temperature of electrons and hence the higher their speed. Therefore speed of capacitor (gravitor) increases at voltage increase. There is a usual chain, i.e. electric energy transfers into heat energy and heat energy transfers into mechanical one. Approximately the similar process takes place in Coolidge Tube, but there the heat of electrons is caused by thermoelectric emission on the cathode. Electrons evaporate from the cathode surface. Voltage, existing between anode and cathode, makes electrons move in one and the same direction. They move towards anode, which serves as piston and takes heat push upon itself, i.e. it transfers heat energy of electrons into mechanical work. It is clear that the higher voltage is the more intensively electrons are heated. Therefore electrons more intensively influence anode and the suspended tube is deviated to a greater extent. (*Editor's: Nowadays this very explanation is the most admitted one.)*

Resting upon Biefeld-Brown's theory and experimental results and at sufficient financing, already within the next few years it would be possible to create a unique electric engine which can operate on hot or quick electrons. Besides there will appear a possibility to orbit the first spaceship equipped with a propulsion system similar to gravitor one's. In near future scheduled cargo-and-passenger service could be organized within planetary system. First of all such transportations can be organized between the Mars, Moon and Venus orbits that will let to start active colonization of these planets.

Thus, more than 200 flights to the Moon and back and about 70 flights to the Mars are possible by means of the gravitor which is made of modern fusions and dielectric materials and which is placed on the Earth orbit. At that there is a minimum risk for its breakdown during the flight, because gravitor has no movable parts and no explosives are used. Negative consequences for the crew in an improbable emergency are insignificant. Even if high voltage causes disruption of dielectric layer between the electrodes and voltaic arc appears then the spaceship, equipped with such a propulsion system, will mechanically arrive its destination. The source of energy will be its only trouble spot and it will happen only if nuclear reactor instead of solar batteries is used. There is also an idea to attach dwelling capsules and cargo bays, equipped with such electric engine, one after another, like carriages are attached to a locomotive. It will let to equip this interplanetary vehicle in several stages by compactly packed modules which are delivered from the Earth. Unfortunately, this project exists only on paper and support of government or financial institutions is not expected in the near future. Faint hope is rested upon enthusiasts but there are too few of such people, besides almost all of them are rather hardly related to space flights and science. Most of scientists even hear nothing about Biefeld-Brown Effect, because it has never become widely known, since all works of this field, except the very first experiments from the USA, were run and then closed in secrecy.

Part II

"...it's time to go to the stars"

Due to the theory of relativity there is an ingrained belief of several generations that it is impossible to move with more than 300 000 m/c speed, i.e. to move faster than light, since the postulates of this theory read that mass is directly depended on speed. Calculations demonstrate that while speed of some accelerated body verges towards light speed its mass will increase and in the point of equilibrium (V=c) it will unrestrictedly increase. Hence acceleration will tend to zero and speed will not virtually increase, no matter how long acceleration speed acts on the object. In other words, the accelerating mass dissipates speed. In principle it is correct but another conclusion can be also drawn from the dependence of mass and speed, for example, the following one: to accelerate an object to the speed which exceeds light speed, mass of the accelerated body should decrease and its speed should increase. Any physicist will say that this statement is correct. At first sight it is practically impossible to realize this idea therefore it was wrongly give to writers of science fiction. They added this idea to their armory and the hypothetical device, which realizes it, was named as gravitational engine. Let us realize this idea from the practical point of view and basing only on facts and logic.

Thus let us apply to the facts and give the description of one experiment. Professor of Pulkovo Observatory N.A. Kozyrev is the author of this experiment, he was also the first to discover the phenomenon of lunar volcanism. Kozyrev used a gyroscope, i.e. the device consisting of two rings of different diameter. These rings are placed athwart one in another and movably jointed. Top is attached to the inner ring through the cardan. The scientist put usual thermos filled with hot water near the beam scale. The top of the gyroscope was preliminarily spun up counterclockwise. At that balance indicator showed that gyroscope top at 90 gram weight became 4 milligram lighter in weight. Then Kozyrev started to pour water of room temperature inside the device through the vinyl chloride tube fixed into the hole, which was made in the plug of the thermos. Seemingly it is impossible to influence upon the gyroscope operating and its weight by a balloon filled with hot water, which is being cooled. Besides the thermos consists of a double-walled container which practically entirely excludes heat exchange with the environment. However balance indicator moved at one or two points, i.e. such connection existed.

This experiment can be divided into two steps. At the first stage the thermos filled with hot water is placed near the gyroscope and we can register the decrease of the top weight. At the second stage water of room temperature is poured into the thermos and the scale again shows that the weight has been changed. Somehow Kozyrev explained the second stage of this experiment. He assumed that when we pour cold water in the thermos, then the equilibrium of the system is broken because some irreversible processes take place there. Cold water can not cause temperature increase of water in the thermos. Until the system come to equilibrium at a new level, i.e. until the same temperature is set in the whole volume of the thermos, this process will cause the condensation of time which "additionally" influences the top. However the scientist missed that the weight of the top had decreased long before the moment when cold water was poured into the thermos, i.e. before irreversible processes appeared there. *(Editor's: Evidently, the author of the article is not well familiar with Kozyrev's researches and with works of other experimenters who investigated this problem.)* In other words, at the first stage of the experiment there is nothing to condense time and that insignificant weight condensation, which was equal to 4 milligram, absolutely does not fallen into his theory. Some other process is presented there.

It is quite possible that at the first stage of the experiment we deal with nullification of the top weight, i.e. mass of the top tends to zero. Some conclusion arises from this assumption, i.e. hot water can influence the mass of the top in some still incomprehensible way. *(Editor's: The point is not in water temperature but in the process of changing of the entropy.)* Thus there is a simplest and primitive model of the gravitational engine, viz there are hot water along with gyroscope and the required decrease of mass at the "outlet". Certainly, it is very bold assumption but it can be easily checked. Near the gyroscope there should be placed thermos filled with not boiling water (of 100°C temperature) but with water of 50°C temperature. If the weight of the top decreased by one half (for example, by 2 milligram instead of 4 milligram), then we can surely ascertain that the hotter water in the thermos is the more decrease of mass. Besides, each 100°C of water heating will decrease mass of the top by 4 milligram. It is not very difficult to calculate the temperature at which mass of the top will verge toward zero and the top will start accumulating negative mass.

Let us now assume that our experiment was successful then water as well as any other matter cannot be heated to such high temperature. However it is possible to heat electrons. They are very compact and have insignificant mass and can be heated to very high temperature. It is necessary to take several plates then to place a dielectric between them and to apply voltage. Electrons rest on the dielectric and start heating. The top can be rotated anti-clockwise by means of electric engines, i.e. capacitor and gyroscope should be combined in one system with common center of gravity. It is not very difficult, as the saying goes, it is just an engineering problem. The higher temperature of electrons is the more decrease of the top mass. Finally there is a moment when mass of the capacitor-gyroscope system will reach zero and this electric machine will rise to the surface of the gravitational field of the Earth or any other planet. It will be Biefeld-Brown's force that will make our apparatus taking off the gravitational field of the Earth. This very force will start to accelerate our apparatus in the space vacuum and top of the gyroscope will lose its mass which appears at acceleration. Thus it is quite possible to reach speed, which is close to the light one, or even to overcome the threshold of light speed.

Part III

"...what is new?"

What is implied by gravitational waves? This question has been exciting the scientists from the half of the XX century. But still nowadays, in spite of all attempts of the scientists to dissolve the problem, these waves were not discovered. There is a standard way to discover them, which is based on general relativity. Basing on this theory, scientists assume that gravitational waves should change in some way the weight of material objects. According to this assumption the scientists suspend very heavy balls to very precision scale and then try to trace changes of their weight. There were only negative results. It seems that mass changes so quickly and to such short period of time that scale fails to react upon these changes. However there is another way which seems to be very promising.

Dependence of time and gravitation is a well-known fact. Thus speed of time course depends on force of the gravitational field of an object. Stars and planets can be considered as an example. The more intense attraction near them the more slowly time passes there. In other words, the greater mass of the planet the more slowly time passes near it. It is possible that oscillations of masses, which generate gravitational waves, in some way change speed of the time course along the whole way of the propagation of gravitational waves. In other words, gravitational waves are always attended with slight time fluctuations. Time starts to pass a little faster or a little more slowly as compared with its usual course. These changes can be traced by means of usual quartz plates. Let me remind of the fact that quartz plates are used at some models of clocks to keep time. Thus, oscillations of gravitational masses produce in space both gravitational waves and time fluctuations (chronowaves), which can be easily detected. Possibly these chronowaves are the part of gravitational waves. This assumption is already a scientific fact which is unfortunately ignored by more than one generation of researchers.

Professor N.A. Kozyrev was the first who surmised about the existence of time ways (chronowaves). He called them as "time flows". The scientist worked out and run a simple experiment to confirm his hypothesis. Telescope and quartz plates were used. Quartz plates were placed in focus of the telescope, which was pointed to some bright star. Its objective lens was covered with black paper or tin plate in order to exclude the influence of light beams. At that quartz plates reacted upon the presence of chronowaves. Thus there was changed the frequency of oscillations of plates which were placed in focus of the telescope.

Besides the scientist discovered that in focus of the telescope there were changed electroconductivity and volume of some matters. It is explainable since some parameters of space as well as speed of passing of some physical processes and phenomena depend on the speed of time course. Outwardly it appears in change of electroconductivity and volume. In spite of the fact that during the experiments tin plate (thick metal cover) was used some skeptics insisting that the point is in infrared radiation which is produced by the heated cover. However they was silenced by Kozyrev's researches. Basing on the fact that we usually see stars not where they really are, but at the place they were at the moment of emanation of light signal, the scientist calculated the location of Procyon star, which is the nearest to the Earth. Kozyrev pointed the telescope to this "clear" part of the sky zone that allowed to avoid the influence of infrared radiation and to fix the location of the star. Thus it is nothing but chronowaves generated by the star that can change oscillation frequency of quartz plates.

After Kozyrev's death this field of research is not considered by official science and the very existence of chronowaves is denied. Only some theorists warily impute the above-listed effects to the action of some torsion fields and waves the existence of which is rather questionable. However chronowaves exist and they can be used in practice, for example, to get information from any point of our galaxy in several seconds. **It is quite possible since chronowaves have no mass and thus they are propagated more quickly than light waves.** However it is still a question what will transmit this information to us because the transmitter of these waves yet has not been invented. *(Editor's: It is interesting what mass of light waves the author means. It is a pity that the author of the article did not read 9 previous issues of NET magazine! There was a detailed demonstration of the connection of Kozyrev's theory and experiments with the theory of longitudinal waves in aether. If to replace Kozyrev's notion of "waves of time density" by the notion of "waves of aether density" then we will get a real experimental approach which can help to solve the problem of antigravitation and time control.)*

Of course theoretically chronowaves can be also generated without participation of great gravitational waves (it seems that heat processes can also generate them). However no particular success was achieved in this field. Thus by means of chronowave transmitter it is possible just to "listen" to remote stars the light of which will reach the Earth in many years. This receiver is very simple and it can be made by anyone who knows a little about radio engineering. *Under home conditions it is better to use a miniature incandescent bulb (lamp with filament tungsten) as a receiver and usual ohmmeter as measuring instrument.*

If in usual receivers radio waves are excited by weak electric currents existing in antenna then chronowave antenna is constantly under tension. Chronowaves do not excite current there but change some characteristics of it, influencing on electroconductivity of the matter which was used for the antenna (let us remember Kozyrev's experiments). Later on these insignificant changes are transferred in audio signal or fixed by any other means. That is the principle of operation of the device.

Editor's: These experiments on the registration of flicker effect in more detail are described in A.M. Mishin works which were published in previous issues of our magazine. As the reader could notice, the editors' opinion in many respects does not coincide with the author's point of view. (Alexander V. Frolov)

References

1. *G.Ya. Myakishev, B.B. Buhovtsev,* Physics, *M.: Prosveschenie, 1991*
2. *A.S. Kuzovkin, N.N. Nepomnyaschy,* What happened to torpedo-boat "Eldridge"? *M.: Znanie, 1991, Series "Znak voprosa" #3, p.37*
3. *S.N. Zigunenko,* How the time machine is made? *M.: Znanie, 1991, Series "Znak voprosa" #5, p.35*
4. A. Einstein and gravics, *Collection of articles, M.: Mir, 1979*
5. *A.D. Chernin* Physics of time, *M.: Nauka, 1978*
6. *A.I. Veinik* Thermodinamics of natural processes

Time Machine

Alexander V. Frolov gave this interview to the Spain magazine "Mas Alla de la Ciencia" ("Science"). The interview concerns some questions about scientific-research work on investigation of active properties of time. Let us remind that in 2002 Faraday Lab Ltd and Vadim A. Chernobrov (Moscow) started the join Time Machine Project.

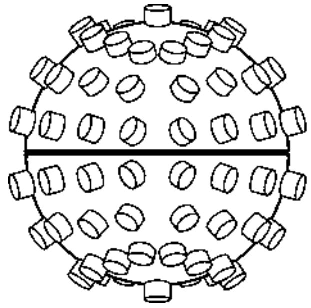

Miguel Segui: How does it value the results obtained by the prototype of time machine?
Alexander V. Frolov: Usually it is about 3% changes, i.e. 3 sec per 100 sec of experiments.

M.S.: How other devices are preparing inside the project Time Machine?
A.F.: Now we have project with special electromagnets, which are emitters of the longitudinal waves. So, other devices are electronics parts to provide impulse current. But it is not the only one version of the system. If the emitters of the longitudinal waves are based on other principles so all device has other design.

M.S.: Do you believe that some day the time trips will be possible?
A.F.: Yes, sure. We have the clear theory for it and it is based on understanding of physical sense of the aether, its relation with notion of time and with concept about existence of elements of matter in space-time. According to aether-vortex theory any element of matter is created from aether and its temporal parameters depend on density of aether.

M.S.: What practical applications can have the prototypes that are developing?
A.F.: Any change in the time course is the method of influence onto biological systems and the medical aspect is the main application. Also any physicist know relation between G (the gravity constant) and time course (rate of time). For example, time is slower on the surface of Earth than in space. So, local time changes (which are changes in density of aether or direction of aether wind) can be used to get propulsion force without reactive mass

flow. Also we can hope that some new materials can be produced in special area of slowed or accelerated time course.

M.S.: Which are the main difficulties of developing of these devices at the moment?
A.F.: For us now it is a part of the work that is related with designing of the powerful current electronics circuits. Also there are no known analogies for patent work, but we hope to complete our patent claim in March of 2003.

M.S.: Is it foreseen that they carry out experiments with human beings and the machine of modification of the time, just as V. Chernobrov carried out?
A.F.: We, i.e. Faraday Laboratories Ltd, are developing small unit for testing of the principles. In future we'll work with other design and with participation of the pilot (human passengers). But it will be the system of other kind. Let me explain: now we are working to get small local change of the aether density, so called "chronal charge". It is analogy with electric charge, which produces electric field in space. If the electric charge is moving then it will produce magnetic field. According to our plans after confirmation of the principles we are going to create design with moving "chronal charge" to generate "chronal field". There are suppositions that this field can be used practically for antigravitation propulsion transport.

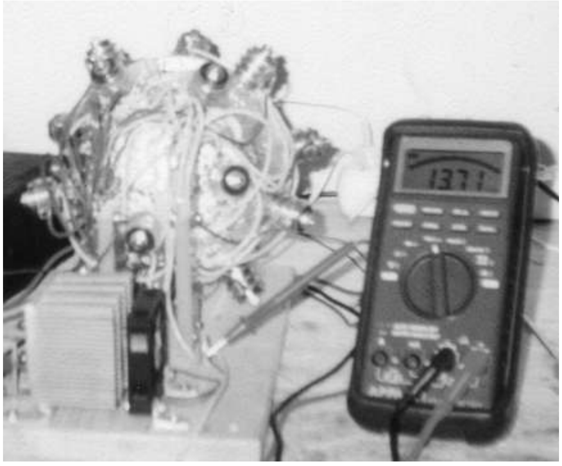

"FRCM" – Avalanchedrive
Fractioning and Replacement of a Constant Mass

Murilo Luciano Filho, Brazil

murilo@intrumenti.com.br, avalanchedrive@hotmail.com

For the while, this project has only these two names, as above. There is still a lot to do upon this development that involves free energy and / or "over-unity". It is designated to be of free use and application by everybody and everywhere, without any royalties just like it was a very old idea, which is now opened for entire community use. The inventor has enough strong reasons for this will.

This is a kind of donation, a personal donation, for the very best future of the Earth. Pioneers are needed (late pioneers?) Any person may be involved in this development to make money since there are no barriers for that. This means that this is an open idea and in this way it must stay. The principles are new and original and the energy involved is quite huge. In the inventor's mind there is realized the possibility to create an expert hands-on group to finish the project. In this article there is a rough draw which seems to be enigmatic but in fact it can clear the understanding of the principle and show the situation at which body may offer controlled spatial variation of length for the same weight.

The short text and a schematic draw (see Fig.1) are the way to show the principles and a few possible construction solutions. The photo shows a part of the device which can clear the so called weight localized rarefaction. Possibly the understanding of capacity for an observer will be a little forced. If necessary author can send the full patent requirement text, which contains 54 pages and 32 figures, not in English but in Portuguese.

FRCM – "AVALANCHEDRIVE"

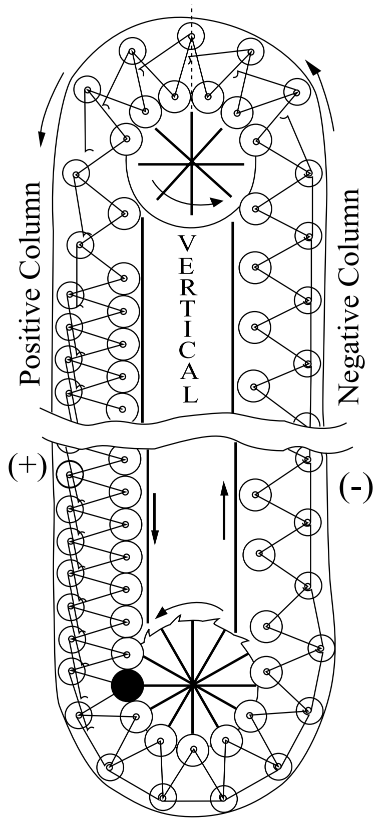

Fig. 1

Understanding of **FRCM / avalanchedrive** is a challenge in itself. When you start to learn about FRCM you will find complete original problems to deal. **Free your mind basing on some mechanical and physics knowledge and then avalanchedrive will come to make a part of your life. Feel free to work.**

PATENT REQUIREMENT RELEASE

This patent requirement is basically the application of a new physical principle which could be defined and found at least 400 years ago, so simple it is. Surely it is one of such things which was let behind by human mind and concern. FRCM is odd and amazing. The key for this project is a special body that can behave like a solid and like a fluid, at the same time, with combined characteristics just as designed and required.

The mass potential of a specially designed segmented body is hold and managed, and throughout proper means is sequentially fractured while suspension balance is kept. This causes, after start, a constant and cyclic movement, since arrangement goes, and the conducting and forcing form of the mass is not interrupted.

Also called "avalanchedrive", this principle work is based on just three main parts or components: 1st is the special segmented body – a chain; 2nd is the stationary and external assembling to hold and conduce the chain; 3rd is the straight vertical, in order to achieve the maximal weight storage position in the arrangement.

The chain is an endless repetitive zigzag construction with weights and defined angles that looks like an entire and elongated ellipse. Its design makes easy the change of its profile and form (mutant profile) that will be contracted or elongated, according to the way it is organized, supported and driven. When contracted or expanded, the chain assumes its maximal or minimal weight, in comparatively the same heights and level, but at different cross sections (tunneling?).

For distribution of the chain track the external assembling forms a kind of duct or rail, which is also elliptical. It allows to the chain to be contracted and expanded, to change direction, move, and pass away, as desired. The chain follows the way that the rail allows, in four different zones: mass-united zone, or positive column, or weight capacitor; the lower reversion input zone; the liberation zone, or negative column, with same height of the positive one; and the upper reversion zone, and its "zero point", or non-resistance top.

Both reversion zones are composed mainly of wheels. However, while the lower zone has to hang, open and bolt the chain, in order to change its profile and natural falling-down trajectory, the upper has a very passive function, that is just to let the chain cross over, and avoid any turn back possibility.

Fact is that in all circuit the chain follows with passivity the "voids" of trails, but assumes a very strong positive and pressure action in mass-united zone, especially in the button, where to the wheel there is applied all the active potential weight of the device, formed by the entire stack of the full contracted state, forcing movement.

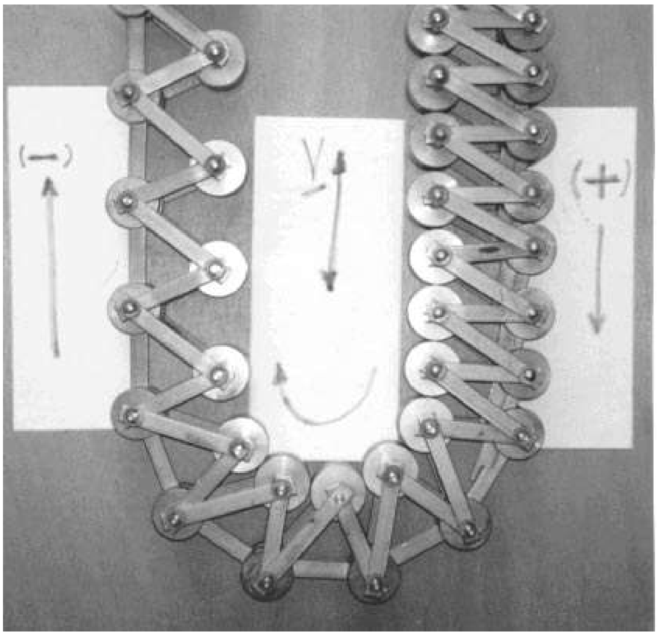

Photo of FRCM (Avalanchedrive)

It is easy to understand how and why FRCM works. Weight of the positive column is directly connected – as a compressed spring – against the relatively lighter negative column, through lower reversion wheel, where happens a kind of flux (the "solid flux") of the chain, that never can find a state of rest. Cycles of "breaking" of the chain can beconsidered as the instant of touch of each corner of the chain to the lower step wheel. To complete the circuitry, the chain excess, which is much lighter, passes over the upper wheel and turns back to the positive column top. It is impossible to turn the full subtracted potential to its source but mechanically the easier job to be done in FRCM.

FRCM manages and converts mass in energy...

The chain flows from the positive column, naturally finds changed profile and reverses to opposite way, and then it is pushed up to the negative column and to top. Finally it is collected, friendly, to restore the potential, at non-stop run! The only alternative to the chain is to move and move, and apply the positive surplus of weight to wrench under axle of the wheel.

This FRCM principle allows develop the equipment which is going to be able to take energy, work and force from gravity potential, so as it happens from wind and water at electrical generator machines, today. The potential is kept in such critical situation, that the only way for it is to escape and move. Besides in consequence of the suspending mass state the linear constant repetition of event will suffer. Water and other fluids do not offer "mutant profile" with fixed value of pressure, just as solid materials. There are many arguments for this view.

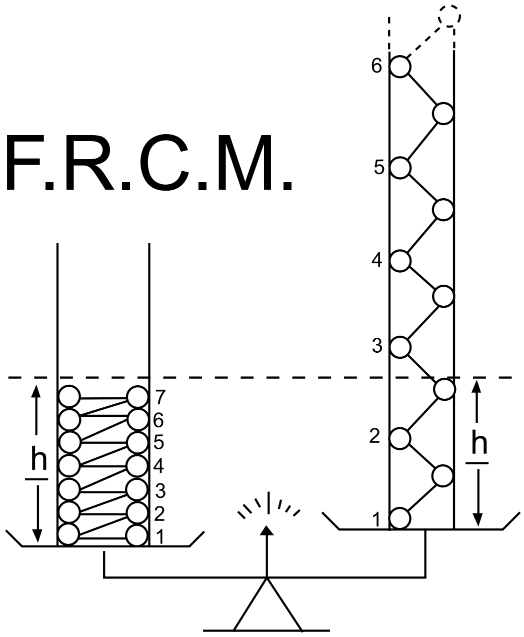

Fig. 2
Segment of a chain

Considering all mechanical losses, a part of the mass will be very active and applied in eccentric, or radial position to the wheel, even with the risk to cause too high speed. FRCM manages and converts mass in energy.

See a non-scale schematic draw (Fig. 2) and photo of a chain segment to compare densities between columns at right and at left.

A theoretical case can be calculated: at total chain weight - 100, general losses - 20, total negative resistance - 30, active or positive surplus is 50 – 30 = 20. This surplus that may be of 2kg, 200kg or 2,000kg, depending on scale, will force and free edge of the wheel, while the total weight keeps confined in the device. In the above case the difference of balance is 1.66. This means that the flowing velocity of the negative pile is 1.66 times higher than the falling of opposite positive pile. Just like a "solid venture", this change and relation is fixed, and in other cases may be varied or calculated to 3, 4 or even more! In all cases the speed must be hold that is not very easy job while one faces gravity acceleration.

The author of FRCM conceptions feels like these ideas have been around for many centuries, and that they are now discovered just like old draws. And this can be the bigger and the best moment to release this. Many thousands of different designs and solutions are coming in the future at looking for the better performance, costs and durability. Author thinks it to be strange that previous guesses could not reach this macro-physical concept. To build FRCM circuits, even re-applied or used old pieces, for example from ships and trains, will be useful.

This is only a start to a "new-old" development and clue to the sources of the modern mechanical engineering. Certainly some oppositions should be overcome. It is expected that other names are also going to be created and proposed to FRCM, but "avalanchedrive" seems to be a good one.

Author expects to get the invention and concept credits but not royalties and would also be glad to keep working on this concern. So, this message is a kind of invitation...

Editor: Unfortunately we have no photo or video documents of the operative device. We wait for readers' comments on this invention.

Complete text of the patent is avaible in Pourtuguese only.

ELECTRIFY
an online and published
magazine about Electric,
Hybrid, Fuel Cell Vehicles,
advanced batteries, ultra capacitors,
fuel cells, microturbines, free energy systems,
EVents and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908 fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

Institute for Planetary Synthesis

P.O. Box 128, CH-1211 Geneva 20,
Switzerland
Tel. 41-022-733.88.76, Fax 41-022-733.66.49
E-mail: ipsbox@ipsgeneva.com
website: http://www.ipsgeneva.com

Hydrogen-Powered Vehicles at Least a Decade Away

Peter Behr
Greg Schneider

Washington Post
http://www.washingtonpost.com/wp-dyn/articles A635122003Jan29.html
Thursday, January 30, 2003; Page A09

This information is forwarded as a courtesy from:
http://www.integrityresearchinstitute.org

President Bush's vision of a hydrogen-powered, non-polluting "Freedom" car for the next generation of American motorists pulled out silently from a Newport Beach, Calif., garage with Gregg Kelly at the wheel, bound for his office 10 miles away.

Kelly, president of a California robotics company, happens to drive a Toyota prototype of a hydrogen-fueled car, one of a handful in the United States today.

It will take at least a decade before a host of technological, economic and political barriers are overcome, permitting a fleet of these revolutionary vehicles to appear on U.S. highways, industry experts said. But by including the goal of hydrogen fuels in his State of the Union message Tuesday, Bush has opened the door to a fledgling movement that has already attracted a surprising coalition of supporters.

Environmentalists, automakers, oil companies and coal producers, engineering labs and strategists have seized on hydrogen as an almost too-good-to-be-true power source. It is abundant in water and air, it burns cleanly and it could free the nation from its dependence on Middle Eastern oil.

"For President Bush to frame the goal as he did is significant," said Jeremy Rifkin, consultant and author of a book advocating a transition from oil and gas to hydrogen. "How much is made of this, time will tell."

The president said he hoped that Americans born today would learn to drive in hydrogen-powered cars, a schedule that auto experts said could technically be met. But whether these vehicles will be commercially available depends on a huge array of variables.

First, the technology is still incomplete and unaffordably expensive. The specially equipped Toyota Highlander that Kelly drives has no price tag. The Japanese manufacturer, after investing millions of dollars in research, lent the vehicle to a University of California research project that Kelly's company supports. "My checkbook isn't fat enough," Kelly said.

The car is powered by electricity generated in a fuel cell by chemically combining hydrogen and oxygen. The engine spits out water drops instead of the carbon dioxide and other pollutants generated by burning gasoline.

With foreign manufacturers committed to press ahead, Detroit's carmakers have had to accelerate their research programs. In June, Ford Motor Co. will unveil a prototype car that uses hydrogen to power an internal combustion engine – part of a "bridging strategy" to help ease hydrogen into the marketplace until fuel cells are fully developed.

General Motors Corp. has developed a fuel cell-powered, car-sized "skateboard" – four wheels attached to a platform less than a foot thick, to which any kind of car body could be buckled.

Rather than use fuel cells, BMW has refitted 10 of its $70,000 Model 745 sedans with hybrid engines that burn either gasoline or liquid hydrogen directly. It could be mass producing them by the end of the decade at a "reasonable" cost for its customers if there were enough hydrogen fueling stations to power them, said spokesman Gordon Keil. "We're trying to get [fuel suppliers] interested in hydrogen. We've not met with a lot of enthusiasm."

As daunting as the engineering challenge is the need for a national hydrogen fuel infrastructure – factories to produce the fuel, pipelines and trucks to distribute it and stations to store and sell it. Environmentalists dream of a totally "green" strategy in which solar or wind power is used to separate hydrogen from water – an approach whose costs now would be prohibitive. A nearer prospect is producing hydrogen from natural gas or coal, however in either case, the carbon dioxide byproduct would have to be injected underground to avoid a huge increase in greenhouse-gas emissions, experts say.

Rifkin argues that an eventual scarcity of oil and gas, decades ahead, will push prices of these fuels up to a point where hydrogen becomes cost-competitive. "It isn't a problem that will yield to technology alone," agreed David M. Nemtzow, president of the Alliance to Save Energy.

While all the major automakers are developing fuel-cell technology, most are cautious about hyping it. "We don't want to get too exuberant about it in that sense, overselling it," said Greg Dana, vice president for environmental affairs at the Alliance of Automobile Manufacturers.

Fresh in some minds is the experience of the Clinton administration, which launched a high-profile, $1.5 billion research venture with the

Detroit automakers a decade ago to produce an 80 miles-per-gallon family car. No cars emerged, and the Bush administration halted the venture in favor of its hydrogen strategy.

Some of the president's political opponents contend the hydrogen option is a way of deflecting criticism over administration policies favoring energy production over conservation. "The president seems content with the auto industry's approach: 'Don't make us do anything today'," said the Sierra Club's Daniel Becker. Others say it does not go nearly far enough. Sen. Byron L. Dorgan (D-N.D), chairman of the Democratic Policy Committee, said recently, "It's moving in the right direction. But his proposal is rather timid. I think we need a bolder plan."

Bush's spending plan for the hydrogen project, $1.5 billion over five years, represents a $500 million increase over his current budget. The administration proposes to earmark $273 million for the 2004 fiscal year, but did not offer many specifics yesterday. The funding would support research on fuel cells, vehicle technology and distribution issues.

The magnitude of the goal demands an effort on the scale of the Apollo Moon project, Dorgan said yesterday. "You have to set benchmarks for five, 10 years out."

UNUSUAL PERMANENT MAGNET MOTORS

Thomas Valone

Integrity Research Institute,
www.integrityresearchinstitute.org
1220 L St. NW, Suite 100-232, Washington, DC 20005
Email: iri@erols.com

Abstract

Permanent magnet motors that try to achieve unusual overunity efficiencies with changes in wiring geometry, electronic switching schemes and magnetic configurations often are not successful. There are some designs that should be regarded as conventional and others as promising. Hopefully this article will help the reader to tell the difference before investing or accepting investment. Note: patents can be viewed for free at www.uspto.gov and also http://gb.espacenet.com/espacenet.

Introduction

An article about permanent magnet (PM) motors would not be complete without first reviewing the basic configurations that are present on the market today. Commercial PM motors are necessarily DC motors since their magnets are permanently polarized before assembly. Many PM motors which use brushes are switching to brushless motors that promise less friction and wear. Brushless motors include electronic commutation or step motors. A step motor, often used in the automotive industry, offers more continuous duty torque per unit of volume than any other electric motor but it is often a lower speed motor. The electronic commutation design is applicable to the switched reluctance (SR) motor. The SR motor substitutes soft iron in the place of higher cost permanent magnets for the outer stator and instead has an inner PM rotor. Brushless motors in general produce torque from current in the armature by the application of Faraday's Law. The ideal PM motor has a linear torque vs. speed curve. There are both outer rotor and inner rotor designs that are standard in PM motors.

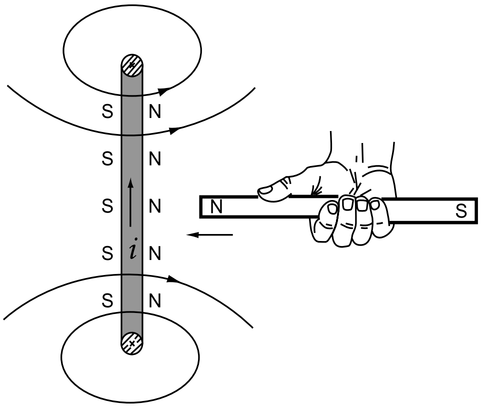

Fig.1

Lenz's Law
Induced B-field opposes motion.

To point out the focus of many of the problems with analyzing motors, the *Motion Control Handbook* (Designfax, May, 1989, p. 33) says that there is a "very important relationship between torque and back emf that is sometimes not understood." This relates to the electromotive force (emf) that is produced by the application of a changing magnetic field (dB/dt). In engineering terms, the "torque constant" (N-m/amp) equals the "back emf constant" (V/radian/sec). In physics, the motor terminal voltage

Detroit automakers a decade ago to produce an 80 miles-per-gallon family car. No cars emerged, and the Bush administration halted the venture in favor of its hydrogen strategy.

Some of the president's political opponents contend the hydrogen option is a way of deflecting criticism over administration policies favoring energy production over conservation. "The president seems content with the auto industry's approach: 'Don't make us do anything today'," said the Sierra Club's Daniel Becker. Others say it does not go nearly far enough. Sen. Byron L. Dorgan (D-N.D), chairman of the Democratic Policy Committee, said recently, "It's moving in the right direction. But his proposal is rather timid. I think we need a bolder plan."

Bush's spending plan for the hydrogen project, $1.5 billion over five years, represents a $500 million increase over his current budget. The administration proposes to earmark $273 million for the 2004 fiscal year, but did not offer many specifics yesterday. The funding would support research on fuel cells, vehicle technology and distribution issues.

The magnitude of the goal demands an effort on the scale of the Apollo Moon project, Dorgan said yesterday. "You have to set benchmarks for five, 10 years out."

UNUSUAL PERMANENT MAGNET MOTORS

Thomas Valone

Integrity Research Institute,
www.integrityresearchinstitute.org
1220 L St. NW, Suite 100-232, Washington, DC 20005
Email: iri@erols.com

Abstract

Permanent magnet motors that try to achieve unusual overunity efficiencies with changes in wiring geometry, electronic switching schemes and magnetic configurations often are not successful. There are some designs that should be regarded as conventional and others as promising. Hopefully this article will help the reader to tell the difference before investing or accepting investment. Note: patents can be viewed for free at www.uspto.gov and also http://gb.espacenet.com/espacenet.

Introduction

An article about permanent magnet (PM) motors would not be complete without first reviewing the basic configurations that are present on the market today. Commercial PM motors are necessarily DC motors since their magnets are permanently polarized before assembly. Many PM motors which use brushes are switching to brushless motors that promise less friction and wear. Brushless motors include electronic commutation or step motors. A step motor, often used in the automotive industry, offers more continuous duty torque per unit of volume than any other electric motor but it is often a lower speed motor. The electronic commutation design is applicable to the switched reluctance (SR) motor. The SR motor substitutes soft iron in the place of higher cost permanent magnets for the outer stator and instead has an inner PM rotor.

Brushless motors in general produce torque from current in the armature by the application of Faraday's Law. The ideal PM motor has a linear torque vs. speed curve. There are both outer rotor and inner rotor designs that are standard in PM motors.

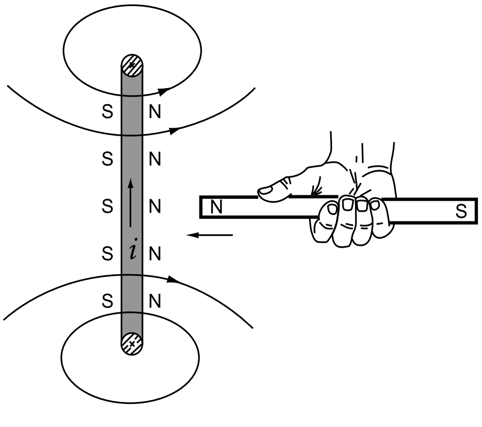

Fig.1

Lenz's Law
Induced B-field opposes motion.

To point out the focus of many of the problems with analyzing motors, the *Motion Control Handbook* (Designfax, May, 1989, p. 33) says that there is a "very important relationship between torque and back emf that is sometimes not understood." This relates to the electromotive force (emf) that is produced by the application of a changing magnetic field (dB/dt). In engineering terms, the "torque constant" (N-m/amp) equals the "back emf constant" (V/radian/sec). In physics, the motor terminal voltage is equal to the back emf minus the IR drop due to internal resistance. (Example: V = 8.3 v, back emf = 7.5 v, IR drop = 0.8 v.) This physics principle, also referred to as **Lenz's Law**, was discovered by Friedrich Lenz in 1834, three years after Faraday invented the homopolar generator. The oppositional nature of Lenz's Law, and its back emf, is built into a physical law called **Faraday's Law**, which is at the root of motor drive. The back emf is the reaction of the changing current in the coil. In other words, the changing magnetic field naturally creates a back emf because they are equivalent.

Therefore, it is recommended that Faraday's Law be carefully reviewed first before proceeding. An article such as "Faraday's Law—Quantitative Experiments" (*Amer. Jour. Phys.*, V. 54, N. 5, May, 1986, p.422) will help convince the valiant new energy experimenter that the change in flux which causes a back electromotive force (emf) is INHERENTLY equal to the back emf. It cannot be avoided or circumvented for excess energy benefit, unless the amount of magnetic flux change per time is also altered. They are two sides of the same coin. The energy into an inductive coil style of motor will naturally equal the energy out. Also referred to as "electrical induction," the changing flux "induces" a back emf.

Switched Reluctance & Field Switching Motors

To explore an alternative method for inducing motion, the "Permanent Magnet Motion Conversion Device" by Ecklin, patent #3,879,622, uses rotatable shutters for alternately shielding the poles of a horseshoe magnet. Repeated again in the Ecklin #4,567,407 "Biased Unitized Motor Alternator with Stationary Armature and Field," the idea of switching the magnetic field with a "flux switch" is common to these types of motors. To illustrate the underlying principle, Ecklin states, "The rotors of most of today's generators are repelled as they approach a stator and are attracted back by the stator as soon as the rotor passes the stator in accordance with Lenz's law. Thus, most rotors face constant nonconservative work forces and therefore, present generators require constant input torque." However, "the steel rotor of the unitized flux switch alternator actually aids the input torque for half of each rotation as the rotor is always attracted and never repelled. This construction makes it possible for some of the current or power fed to the motor windings to magnetically feed through a solid magnetic path to the AC output windings ..." Unfortunately, Ecklin still to this day has not achieved a self-running machine.

Also related is the Richardson patent #4,077,001 which discloses a low-reluctance keeper physically moving in and out of engagement with the ends of a magnet (p.8, line 35). Lastly, the Monroe patent #3,670,189 uses a related principle but accomplishes gating with the passing of rotor poles between permanent magnet stator poles. Monroe's claim 1, seems by its length and detail, to have almost guaranteed its patentability but of course its utility remains questionable.

It seems unlikely that as a closed system the Field Switching Motor can become self-running. In many examples, a small electromagnet will be necessary to help push the keeper into a synchronized rhythm. The Magnetic Wankel from Popular Science (June, 1979) can be compared in a basic manner to this type of invention. Also, the Jaffe patent #3,567,979 can also be compared (see abstract). The Minato patent #5,594,289 is also of a similar type as the Magnetic Wankel and quite intriguing to many people.

It has been found with inventions such as the Newman motor (U.S. Patent Application Serial No. 06/179,474), a nonlinear effect such as an impulse voltage is advantageous for overcoming the Lorentz force conservation effect of Lenz's Law. Also similar is the mechanical analog of the Thornson inertial propulsion device which uses nonlinear impact to transfer momentum along an axis perpendicular to the plane of rotation. A magnetic field contains angular momentum which only becomes apparent under certain circumstances such as Feynman's Disk Paradox, where it is still conserved. The impulse technique may possibly be used to advantage in this Field Switching Motor if the field switching can be done fast enough, with a rapid rise time, but more research is needed.

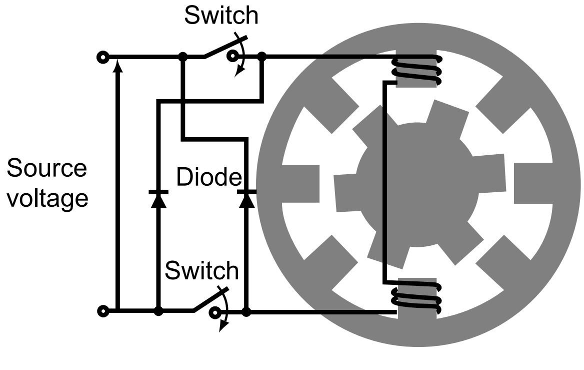

Fig. 2
Switched Reluctance Motor (*IEEE Spectrum* 1/97)

The best Switched Reluctance Motor that also has full accommutation is the Dr. Harold Aspden patent #4,975,608 which optimizes the performance of the coil input and operating above the knee of the B-H curve. Switched reluctance motors are also explained and praised in *IEEE Spectrum* (1/97).

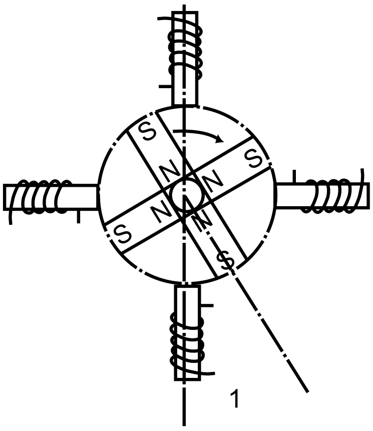

Fig. 3

Adams Motor

The Adams motor has attracted many followers including an endorsement from *Nexus* magazine as the best free energy motor they have seen. The performance of the machine, however, can be fully explained by Faraday's Law. The pulsing of adjacent coils which moves a magnetized rotor is actually following the same configuration as a standard switched reluctance motor. The delay that Adams speaks of in an Internet posting of his motor can be understood from the exponential voltage (L di/dt) of the back emf. The latest addition to this category, which gives credit to the Adams motor, comes from down under with PCT WO 00/28656 awarded to Brits and Christie in May, 2000. The simplicity of this motor is immediately obvious with the switchable coils and permanent magnet on the rotor. The patent also makes it clear that the "input DC current as supplied to the stator coil produces the magnetic repulsion force and is the only outside input to the overall system for total movement...." It is a well-known fact that all motors work on this principle. The key to their design is on p.21 of their patent where the inventors want to "maximize the influence of back EMF which tends to maintain rotation of the rotor/armature in a single direction." All of the motors in this field-switching category try to achieve this effect. Figure 4A of Brits and Christie disclose the voltage sources "VA, VB, and VC." Then, on page 10 it is stated, "At this time current is applied from the power source VA and continues to be applied until the brush 18 is no longer in contact with one of the contacts 14 to 17." There is nothing unusual about this design compared with the more sophisticated attempts listed previously in this section. All of these motors require an electrical power source and none of them are self-running.

When pulsing a coil with the passing of a permanent magnet, a suggestion that would help prove the claim for free energy is not to use battery power for the coil current. Instead, the amazing Weigand wires are recommended (*Pop. Sci.*, May, 1979) that exhibit a huge Barkhausen jump of magnetic domain alignment and a very well-defined pulse shape. Having a coil wrapped around a Weigand wire produces a substantial pulse of several volts with a changing external magnetic field passing a certain threshold. No electrical input power is required for this pulse generator.

Toroidal Motor

As compared to motors on the market today, the unusual design of the toroidal motor is similar to the Langley patent #4,547,713 with a two-pole armature in the center of the toroid. If a single-pole design is chosen, with for example North poles at each end of the armature, this would resemble the radial magnetic field for the armature which the VanGeel patent #5,600,189 uses. The Brown patent #4,438,362 assigned to the Rotron company, utilizes varying magnetization segments for a rotor in a toroidal air gap. The best example of a carousel toroidal motor is the Ewing patent #5,625,241, which also resembles the Langley patent mentioned above. Based upon magnetic repulsion, the Ewing invention uses a microprocessor-controlled carousel, basically to try and take advantage of Lenz's law and get a jump ahead of the back emf. The Ewing invention may be seen in operation, with co-inventor David Porter, in the commercial video, "Free Energy: The Race to Zero Point." Whether it may be more highly efficient than other motors on the market remains an open question. As the patent states, "it is also possible to operate the device as a motor using a pulsed direct-current power source" (col. 7, par. 30). It also contains a programmable logic controller and power control circuit which the inventors thought would send it over the top of 100% efficiency.

Unless a prototype proves to be successful in achieving a torque or force conversion linkage, the internally propelled magnet may be left without a practical application. Commercialization of these types of motors may not be favorable, since many competing designs are currently available on the market, with high flux linkage.

Linear Motors

The area of linear induction motors is well known in the literature. Schaum's Outline Series, *Theory and Problems of Electric Machines and Electromechanics* (McGraw Hill, 1981), explains that these are the same as cutting the rotor and stator of a standard induction motor and laying them out flat. The late Dr. Laithwaite, author of

Motion Without Wheels, was famous for monorail designs for trains in England based on linear induction motors.

The Hartman patent #4,215,330 is an example of one that achieves a linear motor transportation of a steel ball up a magnetized incline of approximately 10 degrees. Another invention in this category is the Johnson patent #5,402,021, which uses permanent arc magnets on a four-wheel cart, exposed to a parallel track of alternating permanent magnets which are in a fixed position. An even more amazing permanent magnet patent is the Johnson #4,877,983 which an eye witness has seen operating at the Johnson home in a closed loop for hours. It is reasonable to assume that a pickup coil could be positioned nearby so that each trip would result in a pulse of electricity to charge a battery. The Hartman patent could also be arranged in such a circular track so that perpetual motion of the first kind can finally be demonstrated.

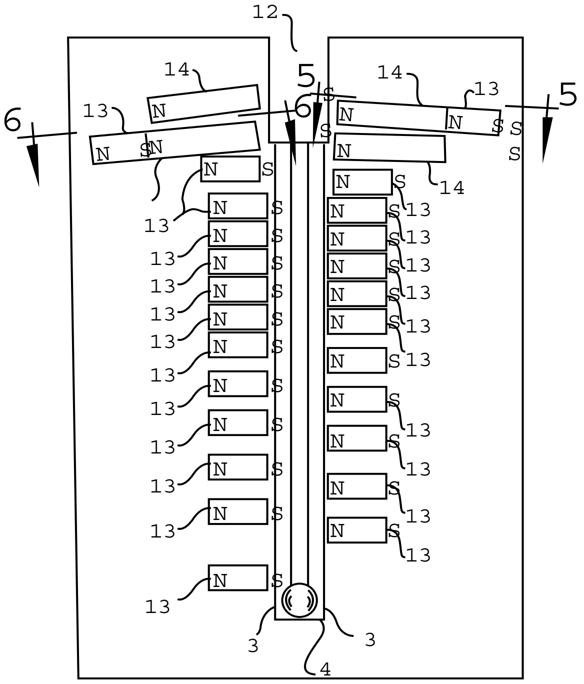

Fig. 4

Hartman patent #4,215,330

The Hartman patent is based upon the same principle as the famous electron spin experiment in physics called the Stern-Gerlach experiment. With an inhomogeneous magnetic field (one that is non-uniform) the force on an object with a magnetic moment is the gradient of the potential energy. Every physics textbook points out that this type of field, that is strong at one end and weak at the other end, will result in a unidirectional force on the magnetic object equal to dB/dx. That is exactly what the Hartman patent possesses (note spacing of magnets). Therefore, the resulting force propelling the ball up a ten degree incline, in the x direction, is in keeping with the laws of physics.

With state-of-the-art magnets, including ambient temperature superconducting magnets which are now finishing the development stage, a demonstration of impressive cargo weight will be shown to be transportable without maintenance electricity costs. Superconducting magnets have the unusual property of retaining the initial magnetized field for years, without the need for periodic energization to restore the initial field strength. Examples of the state of development of the superconducting magnet market can be found in the Ohnishi patent #5,350,958 (lack of cryogenics and lighting system output) as well as the reprinted article from *IEEE Spectrum*, July, 1997 on magnetic levitation.

Static Electromagnetic Angular Momentum

In a provocative experiment with a cylindrical capacitor, Graham and Lahoz (*Nature*, V.285, No.15, May, 1980) have expanded upon the proof published by Einstein and Laub in 1908 that the Lorentz force needs an additional term to preserve action and reaction. The article they cite has been translated and published in my book, *The Homopolar Handbook* (described below). Graham and Lahoz emphasize that there is a “real angular momentum density to r x (E x H)/c^2” and suggest how to see this energy effect in permanent magnets and electrets.

This is encouraging work, with an impressive source of Einstein and also Minkowski for its information. It is possible that it may have a direct application for the homopolar generator as well as the magnetic energy converter mentioned below since both have an axial magnetic field and a radial electric field like the cylindrical capacitor experiment of Graham and Lahoz.

Homopolar Motor

My book, *The Homopolar Handbook* (HH), covers experimental tests and history of the Faraday discovery, including Tesla’s contribution to it. Recently however, there have been new developments into a multi-rotor design of a homopolar generator, similar to the invention of John R. R. Searl.

Recurring interest in the Searl device, as pictured on the cover of *Antigravity*, the biography of Searl by John Thomas, should also center on the homopolar generator (HG). Preliminary analysis reveals that there are actually two separate HG phenomena occurring simultaneously, one which can be called the "revolution" effect (#1) and the second that could be called the "rolling" effect (#2). The first effect can be visualized as magnetized segments of an imaginary solid ring revolving around a common center. As suggested by drawings in HH, p.141-2, there are precedent designs that allow for segmenting an HG rotor.

With this model in mind, the #1 effect can be calculated, for 1 Tesla strength magnets, magnetized axially, adjacent to a single ring 1 meter in diameter, to produce more than 2 volts emf across each roller, (E-field directed radially from outer diameter of rollers to outer diameter of the adjacent ring) with say, 500 RPM. Note that this #1 effect is independent of any rolling of the magnet. The magnetic field in an HPG is tied to space and not to the magnet so rolling will not affect this large scale homopolar generator's Lorentz force effect (HH, p.10).

The #2 effect, located within each roller magnet, is the one noted in *Electric Spacecraft Journal*, Issue 12, 1994, (HH, p.160) where each roller, is a small homopolar generator. This effect is found to be somewhat weaker as it generates electricity from the center of each roller to its periphery. This design is like Tesla's HG (HH, p.81) where a rolling belt is contacting the outer edge of a circular magnet. With rollers in the vicinity of a tenth of a meter in diameter rolling, without slipping, around a 1 meter ring, approximately a half of a volt will be generated. The Searl design of ring magnetic material will normally strengthen the roller's B field.

It is important to realize at this point that the principle of superposition applies to these two effects. The #1 effect is a uniform E field across the diameter of the roller. The #2 effect is a radial effect as stated above (see HH, p.6-8). However, only the emf in the section of a roller between the two contacts, say at the center of the roller and its edge which contacts the ring, will actually cause current flow in any external circuit. This realization means that the effective voltage from the #1 effect will be half of the available emf, or a little more than 1 volt, which is still about double of the #2 effect. Upon applying superposition in the limited region indicated, we also find that the two effects oppose each other and the two emfs must be subtracted. The result of this analysis is that approximately one half of a volt of regulated emf will be present to generate electricity from a single set of rollers and one ring about 1 meter in diameter. As current is drawn, a Ball Bearing Motor effect will also take place (HH, p.54) that actually pushes the rollers along, assuming the roller magnets have a reasonable conductivity *(Thanks to Dr. Paul La Violette for this reminder)*.

In a related work, (*Tech. Phys. Lett.*, V. 26, #12, 2000, p.1105-07), Roshchin and Godin have published experimental results of their one-ring device, called a "Magnetic Energy Converter," with rolling magnets on bearings. It was designed as an improvement to the Searl device. Though my above analysis does not depend upon the ring being made of magnetic material, Roshchin and Godin did so. Their findings are encouraging and detailed enough for researchers to find renewed interest in this type of magnetic motor.

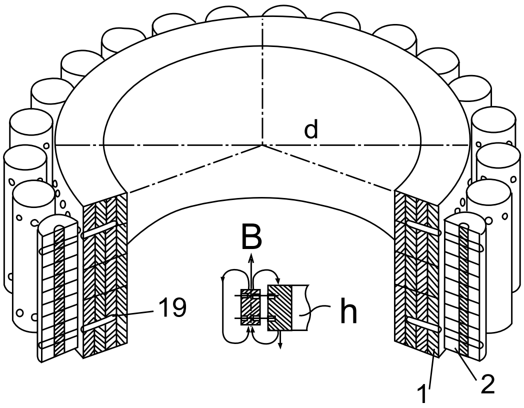

Fig.5

Magnetic Energy Converter in the experiment
by Roshchin and Godin (Russia)

Conclusion

So far, a couple of permanent magnet motors may have achieved perpetual motion, which exceeds 100% efficiency. Of course, conservation of energy concepts have to be considered and the source of the alleged extra energy examined. If permanent magnet field gradients do offer a unidirectional force, as the textbooks predict, then it is about time for its conversion toward useful work. The roller magnet geometry, now called a "magnetic energy converter" is also a unique style of magnetic motor. Exemplified in the Russian patent #2155435 by Roshchin and Godin, it is a magnetic motor-generator that shows potential for excess energy output. Since it relies upon the circulating cylindrical magnet rolling around a ring, the design is actually a generator rather than a motor. However, as they utilize the torque produced by the self-sustained motion of the magnets to run a separate electrical generator, it is working as a motor.

Reprint from Proceedings of Institute for New Energy Conference, 2001

Hydrogen Energetics

Review by Alla Pashova

A Russian writer Ludmila Ulitskaya, who is rather far from energy problems, says: "I'm quite sure that cheap electric energy sources have been already worked out and are hidden in oil king's safes. However such projects won't be taken out of the safes until the last oil drop is combusted. They don't want redistribution of money, world, power, and soon". Experts are sure to agree with the writer.

Now only ideas which can coexist together with the traditional oil power engineering and with a system of world oil distribution can escape the safes. Some expensive scientific toys promising future universal welfare are created to reassure ecologists, viz automobiles equipped with hydrogen engine. According to estimation of Japan representatives of motor car industry, installation of the hydrogen engine into such a car as a classical Sedan now comes to about $700 000. Approximate cost of a hydrogen bus named Kitaro achieves the astronomic amount of $1.25 million.

The hydrogen engine consists of several fuel elements, which are known as fuel cells. Hydrogen under pressure of 1.5-2.7 atmospheres comes to an anode. A cavernous catalyst splits the molecules into electrons and protons. The electrons set the electric engine in motion, and the protons come through a membrane acting to a cathode as an electrolyte. Here the catalyst combines them with the working electrons and with atmosphere oxygen into water molecules. It should be noted that the Europeans and the Americans are extremely charmed by this water streamlet which is exhausted instead of poisoning gases. Moreover, output of the fuel cell is 3-5 times more than output of the petrol engine.

The first hydrogen power station inaugurated in Las Vegas, Nevada, in 2002, (see Fig.1, 2) represents a system of the fuel cells with the proton penetrable membrane as an electrolyte (they are designed and produced by experts of Plug Power Inc.). The station produces hydrogen transforming and cleaning natural gas and generates electric power combining hydrogen stored in reservoirs with atmosphere oxygen. The USA Energy Department and Air Products Inc. together paid $ 10.8 million for this project.

It is necessary to note that the priority branch of the hydrogen engineering is providing new means of transport with fuel, and electric power production is just the second aim. The mentioned station produces electric power in addition to its main function. Nevertheless, if fuel market begins to decrease then the electric power payments will provide the "hydrogen magnates" with stable profit.

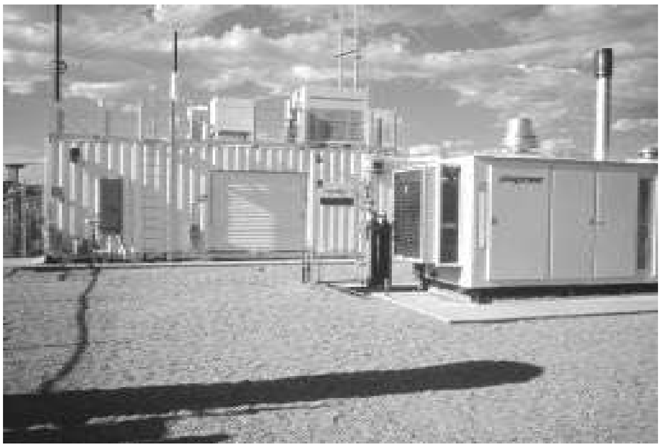

Fig. 1
Hydrogen power station

Nowadays the hydrogen automobiles allow oil kings keep the system of centralized fuel distribution and the infrastructure of service stations which will provide the means of transport with methanol, natural gas with hydrogen blend, and liquefied or gaseous hydrogen. Thus interested companies control the hydrogen engineering and scientific thought.

Moreover, oil can be applied as it is useful for board reforming (conversion), i.e. it can be used as a hydrogen source. The systems based on conversion have been known for a long time but there was need to work at them.

Russian experts improved the conversion system by replacement of the catalyst by plasma. Thus the conversion occurs in gas discharge of HVF appearing in the blend which needs to be converted. According to experts' opinion the Russian system is portable and has high output. In 2001 a group of experts of Russian Centre of Automobile Engine represented experimental prototype basing on Chevrolet Pickup S 10.

The ideal of modern hydrogen engineering is the following: "There should be minimal changes in the system of fuel distribution, non-polluted air, and thin streamlet of distil exhausted". However, evidently, the exhausted water is an ideal source for hydrogen fuel reproduction by means of high effective electrolyzing. If the closed system is worked out we would obtain an energy source powered by continuous chain of electrolyzing and recombination whose efficiency would be more than 100 %.

What are the known ways of hydrogen production? The greatest amount of the gas is produced on the base of catalytic conversion of hydrocarbon with water vapor. Temperature of the process depends on the catalyst composition. It is known that temperature of propane reaction can be decreased to 370° using bauxite as a catalyst. Approximately 95% of by-product carbon monoxide is further used in the reaction with water vapor.

A significant part of the total hydrogen production is given by a method of water gas. The essence of the method lies in reaction of water vapor with coke that is followed by the producing of carbon monoxide with hydrogen mix. The reaction is endothermal, it occurs at 1000°C. The heated coke is processed by the vapor; extracted purified mix of gases contains some hydrogen, some carbon dioxide, and great share of carbon monoxide. Further vapor processing of the carbon monoxide at 370°C increases the hydrogen extraction. The amount of carbon dioxide increases as well however it is easy to be removed by passing the gases mix through a scrubber sprayed by a water contraflow.

AIR PRODUCTS

Fig. 2

There is a well-known iron-vapor method according to which vapor is passed above iron at 500-1000°C. Hydrogen obtained by this method is usually applied to hydrogenise fats and oils. Composition of the iron oxide depends on the process temperature. Iron-3 oxide (Fe_3O_4) prevails at the temperature below 560° C. If the vapor is passed above the iron at the temperature above 560° C then iron-2 oxide (*FeO*) share increases. A slight admixture of carbon monoxide is removed when the heated mix passes above a catalyst. At this time the carbon monoxide is transformed into methane.

In the processes of oxidation and thermal cracking of hydrocarbons at soot production hydrogen is produced as a by-product.

The next effective method is methanol-vapor one. It is an endothermal reaction which is proceeds in common iron reactors at 260° C temperature and at 20 atmospheres pressure.

There should be also mentioned a catalytic reaction of ammonia decomposition at which nitrogen and hydrogen is produced. The reaction can be profitable if it is necessary to produce great amount of hydrogen. The method is useful if the obtained hydrogen is directly applied.

Electrolysis is the oldest method to produce hydrogen. Direct current comes to electrodes, the cathode collects hydrogen, and the anode accumulates oxygen. The technology makes hydrogen to be too expensive energy carrier. There is often used a technology of hot processing of water vapor at 700-900° C accompanied with light petrol and heavy liquid fuel which bleeds oxygen. This method is quite expensive as well.

As it is known, **wasteful power consumption of classical electrolysis** is caused by the fact that it is used to overcome forces of hydrate bound of ions with water molecules and to compensate endothermal effect of water decomposition. Therefore, ions reduction at corresponding electrodes requires more current intensity than in the case of absence of this physical phenomena. Hence production of 1 cubic meter of hydrogen requires 18-21.6 MJ, and total power consumption exceeds 50 MJ (taking into account the electric power production). Therefore hydrogen becomes too expensive.

In 1888 D.A. Lachinov patented a method of non-contact electrolysis. Improvement of this method promises huge amount of cheap hydrogen and appearance of high effective energy source operating as a closed cycle. The main problem of Lachinov's method is a gas layer covering the electrode in some time and impeding the reaction. It was proposed by I. A. Goryachev to prevent the gas layer using pulsating electric field. A.V. Frolov has proposed to increase electrolysis efficiency by means of rotation. Centrifugal force causes more swift detach of gas blebs from electrodes surfaces. Power consumption to overcome frictional force in the construction and to generate potential electric field is insignificant as compared with output. Hence efficiency of the system may exceed 100%.

There is also described the experimental device for production of hydrogen and oxygen from water where modulated voltage (not dc voltage) is applied to the electrodes. Modulation frequency is in connection with proper oscillation frequency of water molecules as well as with spatial structure. Possible efficiency of water electrolysis in spark discharge which removes oxygen admixture should be investigated . Water electrolysis at the radiation by electromagnetic waves of light range needs to be investigated as well.

Ikar Research Centre informs that plasma electrolysis can be the most probable source of cheap hydrogen. In 1987 a group of Russian scientists patented the first plasma-electrolytic reactor. Since the reactor had been worked out at a defense establishment then the patent was restricted and not covered by press. The patent contains a structural scheme of the reactor and some results of cleaning and disinfection of water by plasma. The patent does not represent any information of additional energy generated by plasma or hydrogen production.

In April of 1989 American scientists Pons and Fleishman published their experimental results on additional energy production at common water electrolysis. They claimed cold fusion to be source of this energy. However there is no reliable proof for this phenomenon.

In 1996 Yu. Beklyamishev, one of the co-authors of the first plasma-electrolytic reactor, published his experimental results. They demonstrated that there was additional energy in the plasma-electrolytic process. However he did not explain the energy source.

In 1998 there were published new experimental issues of appearance of additional energy in the plasma-electrolytic process. A group of Russian scientists tested one of the plasma-electrolytic reactors and officially fixed the additional energy. A report of the testing session was published in issue #22 of “Infinite Energy” magazine. In May of 1998 the third edition of a book named “Crisis of Theoretical Physics” was published by Prof. F. M. Kanarev. It contains data of additional energy production at plasma electrolysis of water defining the energy source. Soon Ohmori and Mizuno (Japan) published their results in works of Vancouver conference on cold fusion and in issue #20 of “Infinite Energy”. Ohmori and Mizuno observed neutron radiation occurring at a plasma process as well as iron, chrome, nickel, and carbon appearance on a wolfram cathode. That seemed to be a strong evidence of cold nuclear fusion at plasma electrolysis of water. Ohmori and Mizuno have explained neutron radiation as a result of electrons captured by protons. Nevertheless their conclusions raise doubts as cold fusion would have produced much more additional energy than it was fixed.

In 1996 there appeared the first publications where hydrogen atoms fusion instead of nuclear fusion was supposed as the additional energy source at usual electrolysis as well as at plasma source. In Russia first experimental results demonstrating energy consumption decrease for hydrogen production at plasma electrolysis of water were published in 1999 by Prof. F.M. Kanarev. (See F.M. Kanarev’s article in this issue of NET).

Fig. 3

Honda FCX-V3 uses hydrogen as a fuel. Hydrogen is fed under high pressure in the engine.

Nevertheless we have to note that high efficiency electrolysis is possible! O. Hvolson in his “Physics” (Berlin, 1923) explains a method of designing of a machine which can produce heat due to positive difference between energy of hydrogen combustion and energy consumed to hydrogen production by means of water electrolysis. In this case we do not consume energy to split the atoms. It occurs due to intermolecular forces at water dissociation by sulphuric acid ions. Energy is consumed only to neutralize charges of existent hydrogen ions and acid residue ions. Hydrogen combustion produces the energy which would have been consumed to split the atoms in the air. **Hence 67.54 Kcal of energy can be produced at 5 Kcal consumed. (F. Lepehin)**

High effective electrolysis application is clear future of energetics. Nowadays Stuart Energy Inc applies water electrolysis as a hydrogen source in a frame of hydrogen-fuel project. Experts of Stuart Energy Inc have designed a refueling unit which produces hydrogen and stores it in a 104-pound reservoir. Thus several auto are provided with fuel and it takes a couple of minutes to refuel.

It is quite simple to predict social and economical consequences of innovation of high effective energy source operating as a closed cycle energy system. Economic activity and private life of people will be independent of cities, of their industry and power engineering. People will move out of the cities and they will use compact and powerful energy sources there. Hence, decentralization of world economics will happen.

Regions of any country become more free from the central authority, the power of which nowadays is based on centralized fuel energetics.

On the Longitudinal Electromagnetic Waves

Alexander V. Frolov

General Director, Faraday Lab Ltd
Phone/fax: 7-812-380-3844
Email: office@faraday.ru

Concerning acoustic waves, we deal with longitudinal oscillations of the medium, since degree of air (medium) compression and rarefication changes along the direction of the wave spreading. Notion of the medium of electromagnetic waves propagation, i.e. notion of "aether", as well as the very possibility of the existence of longitudinal electromagnetic waves is disputable for some scientists. Let's clarify the situation and demonstrate that usual photon is longitudinal wave of energy density, which can be described as Umov-Poynting vector.

The transverse character of electromagnetic waves means that vector of electric field and vector of magnetic field are directed across the wave line. However, these vectors are just our subjective way to describe the process. Notions of electric and magnetic fields can be replaced by the notion of **unified helical field**, since to describe the motion of the point along the helical line there should be concerned the linear transfer (it is detected as electric field) and the process of rotation (it is described with the magnetic field). Thus notions of electrodynamics are just one of the ways to describe reality. Energy density of the given point of space is the real (objective) parameter of the wave. This energy density is described by the known Umov-Poynting vector $\overline{S} = \overline{E} \times \overline{H}$.

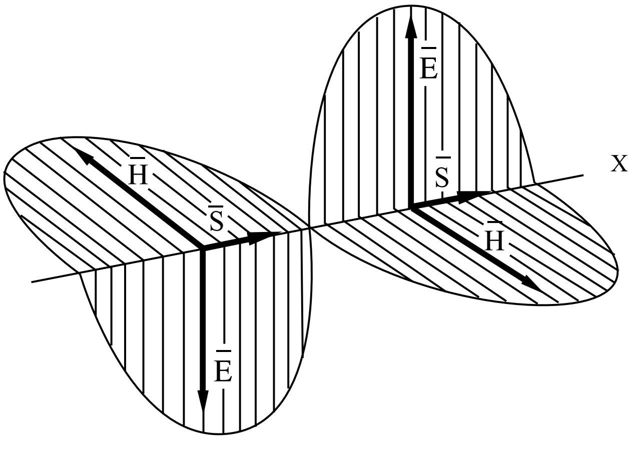

Fig.1

Let us consider electromagnetic wave in classical representation (Fig.1). Then we discover that direction of Umov- Poynting vector coincides with the wave line. The vector is unidirectional and "pulsating", because its quantity changes from zero to some maximal value and then it is reduced to zero, besides it takes place at half of the period of transverse wave. So, it is the explanation of the fact that frequency of longitudinal oscillations is twice as much than frequency of transverse oscillations (Fig. 2). **Previously it was known from the mechanism of energy interchange between longitudinal and transverse plasma oscillations and from the phenomenon of parametrical resonance. However the physical meaning of this phenomenon was not disclosed.**

From the presented here conception it follows that from the physical point of view, electromagnetic waves are **longitudinal oscillations of energy density**. In usual case these oscillations are unidirectional pulsations that determines the ability of the photon to be moving unidirectionally. Practically it is possible to create other types of photons, i.e. types of oscillations of energy density with qualitative new physical properties. It can be achieved by setting of definite functions and superposition of $\overline{E}$ and $\overline{H}$ vectors.

For example, in 1996 a conference "New Ideas in Natural Science" was organized by us in Saint-Petersburg, with the participation of 30 foreign guests and more than 100 Russian scientists. The report made by Academician Ignatyev, Krasnoyarsk, aroused great interest. During the experiments with rotating of crossed vectors $\overline{E}$ and $\overline{H}$ (see Fig. 3) there was created Poiting's vector, which corresponds to propulsion force of 60 N (about 6 kg).

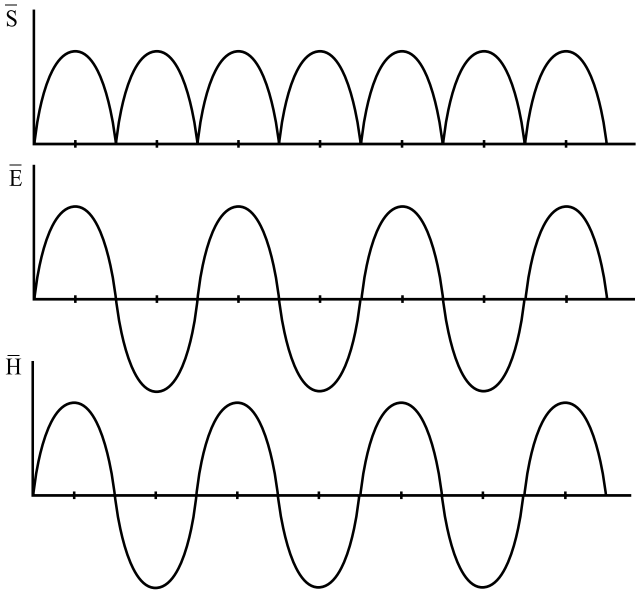

Fig. 2

The experiment is demonstrated on the photo (see Fig. 4), where the diameter of the device is equal to 4 meters and ends of the coils are to toroidal capacitors. This experiment was made in Krasnoyarsk.

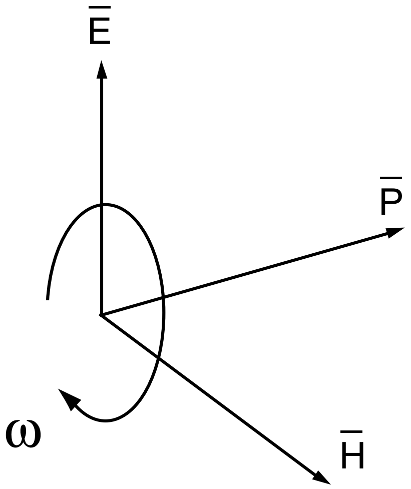

Fig. 3

Since it is impossible to consider vacuum, where some form of energy can exist, as “emptiness” then we can speak about vacuum as about some medium. Faraday and Maxwell, classics of electromagnetic theory, wrote exactly about deformations, stresses and expansion of aether. From this point of view electromagnetic waves are analogues of the waves of longitudinal deformation of elastic medium. More than 60 years ago Nicola Tesla wrote: “I showed that the universal medium is a gaseous body in which only longitudinal pulses can be propagated, involving alternating compressions and expansions similar to those produced by sound waves in the air. Thus, a wireless transmitter does not emit Hertz waves which are a myth, but sound waves in the ether, behaving in every respect like those in the air, except that, owing to the great elastic force and extremely small density of the medium, their speed is that of light.” [1]

Nowadays, the development of engineering and the new views upon the phenomena of electromagnetism, impel us to consider physical **vacuum as material medium of the special type**, which has the well-known properties, in particular electrical and magnetic properties. Moreover, this medium has energy, and energy density can be changed in case of propagation of any photon. Thus, as Tesla stated in his article “The True Wireless”: “The Hertz wave theory of wireless transmission may be kept up for a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history”.

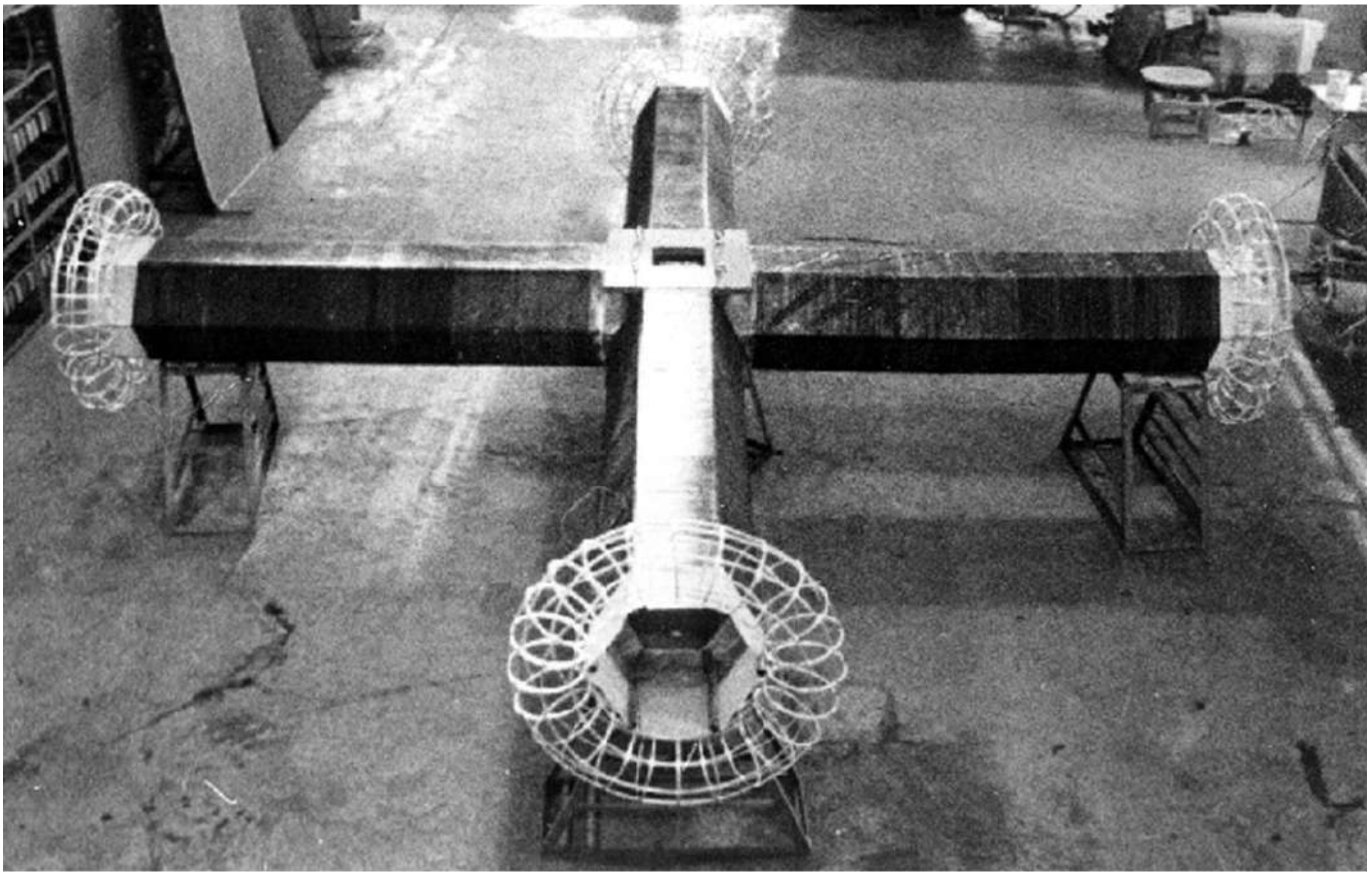

Fig. 4

References

1. “Pioneer Radio Engineer Gives Views on Power”, New York Herald Tribune, Sept. 11, 1932.

"OUTSIDE THE BOX" SPACE AND TERRESTRIAL TRANSPORTATION AND ENERGY TECHNOLOGIES FOR THE 21ST CENTURY

Theodore C. Loder

tedloder@aol.com

Institute for the Study of Earth, Oceans and Space
University of New Hampshire
Durham, NH 03824

Abstract

This paper reviews the development of antigravity research in the US and notes how research activity seemed to disappear by the mid 1950s. It then addresses recently reported scientific findings and witness testimonies - that show us that this research and technology is alive and well and very advanced. The revelations of findings in this area will alter dramatically our 20th century view of physics and technology and must be considered in planning for both energy and transportation needs in the 21st century.

Historical Background

Townsend Brown's Technology of Electrogravitics [1]

In the mid 1920's Townsend Brown [2] discovered that electric charge and gravitational mass are coupled. He found that when a capacitor is charged to a high voltage, it has a tendency to move toward the positive pole. His findings, which became known as the Biefeld-Brown effect, were opposed by conventional minded physicists of his time.

The Pearl Harbor Demonstration. Around 1953, Brown conducted a demonstration for military top brass. He flew a pair of 3-foot diameter discs around a 50-foot course tethered to a central pole. Energized with 150,000 volts and emitting ions from their leading edge, they attained speeds of several hundred miles per hour. The subject was thereafter classified.

Project Winterhaven. Brown submitted a proposal to the Pentagon for the development of a Mach 3 disc shaped electrogravitic fighter craft. Drawings of its basic design are shown in one of his patents. They are essentially large-scale versions of his tethered test discs.

Review of Issues from the 1950s

In 1956, a British research company, Aviation Studies (International) Ltd. published a classified report on Electrogravitics Systems examining various aspects of gravity control. They summarized the pioneering work of Townsend Brown and then described the use of electrogravitic thrust as follows:

"The essence of electrogravitics thrust is the use of a very strong positive charge on one side of the vehicle and a negative on the other. The core of the motor is a condenser and the ability of the condenser to hold its charge (the K-number) is the yardstick of performance " [3].

In one of their conclusions, based on Brown's work, they suggested that: "Electrostatic energy sufficient to produce a Mach 3 fighter is possible with megavolt energies " [4].

In spite of Brown's solid research, they later stated that, "One of the difficulties in 1954 and 1955 was to get aviation to take electrogravitics seriously. The name alone was enough to put people off" [5]. It seems that is as true today as it was in the 1950s.

A report by another British company, Gravity Rand, Ltd. in 1956, agrees with this assessment and states: "To assert electrogravitics is nonsense is as unreal as to say it is practically extant. Management should be careful of men in their employ with a closed mind or even partially closed mind on the subject." [6]

However, a trade press magazine, The Aviation Report, made numerous references to antigravity projects and listed many of the companies pursuing research in this area. Quotes from The Aviation Report listed in the Aviation Studies (International) Ltd. Report [7] are suggestive of what was going on behind the scenes.

In 1954 they predicted that: "... progress has been slow. But indications are now that the Pentagon is ready to sponsor a range of devices to help further knowledge."... "Tentative targets now being set anticipate that the first disk should be complete before

1960 and it would take the whole of the 'sixties to develop it properly, even though some combat things might be available ten years from now." (Aviation Report, 12 October 1954) [8].

During this time period many of the major defense and technology companies were cited as either having research projects or activities in this new field. For example: "Companies studying the implications of gravitics are said, in a new statement, to include Glenn Martin, Convair, Sperry-Rand, and Sikorsky, Bell, Lear Inc. and Clark Electronics. Other companies who have previously evinced interest include Lockheed, Douglas and Hiller." (Aviation Report, 9 December 1955) [9].

Others of these reports mention: AT&T, General Electric, as well as Curtiss-Wright, Boeing and North American as having groups studying electrogravitics.

During the same time period, the Gravity Rand report notes that: "Already companies are specializing in evolution of particular components of an electogravitics disk." [10]

However, in the area of predictions, the Aviation Report stated the following based on an extrapolation of technology development: "Thus this century will be divided into two parts – almost to the day. The first half belonged to the Wright Brothers who foresaw nearly all the basic issues in which gravity was the bitter foe. In part of the second half, gravity will be the great provider. Electrical energy, rather irrelevant for propulsion in the first half becomes a kind of catalyst to motion in the second half of the century." (Aviation Report, 7 September 1954) [11].

Looking back it is easy to say that they missed the mark. Did they really miss it by a half a century? Reading through these reports it is quite obvious that there was much interest in antigravity among a number of very high profile companies, as well as in the Department of Defense. What happened to this interest and why was it all downplayed during the following four plus decades? After all, T. Brown had shown that there is a demonstrable connection between high voltage fields and gravity. Why has it taken until the 1990s for more than just a few scientists to look at these results and publish on them in the open literature? A review of recent statements by former military personnel and civilians connected to covert projects begins to shed light on research activity in these areas over the last half century. **And it appears that there had been significant breakthroughs during this time period, well shielded from both the scientific and public eye.**

Recent Scientific Developments

In this section we consider developments in the antigravity field since the late 1980s and why the confluence of scientific findings and the testimony of witnesses associated with the military and covert groups indicates that a gravity solution with technological implications has been found.

Although general relativity has not been able to explain Brown's electrogravitic observations, or any other antigravity phenomenon, the recent physics methodology of quantum electrodynamics (QED), appears to offer the theoretical framework to explain electrogravitic coupling. Recent papers by members of the Institute for Advanced Study Alpha Foundation are putting a solid theoretical foundation onto the antigravity effects within the theory of electrodynamics and include papers by Evans [12] and Anastasozki et al [13].

Earlier in a 1994 breakthrough paper, Alcubierre showed that superluminal space travel is, in principle, physically possible and will not violate the tenants of the theory of relativity [14]. Puthoff [15] later analyzed these findings in light of the present SETI (Search for Extraterrestrial Intelligence) paradigms that insist that we could not be visited by extraterrestrial civilizations because of the speed-of-light limitations dictated by the general relativity theory. He suggests that superluminal travel is indeed possible. This leads to reduced-time interstellar travel and the possibility of extraterrestrial visitation, which our limited understanding of physics and scientific arrogance has "forbidden" in some sectors for most of the 20th century.

The second aspect of these physics findings deals with the zero point or vacuum state energy shown by the Casimir effect [16], which predicts that two metal plates close together attract each other due to imbalance in the quantum fluctuations. The implications of this zero point or vacuum state energy are tremendous and are described in several papers by Puthoff [17] starting during the late 1980s. Bearden [18] and colleagues have also written extensively on the theoretical physics of zero point energy and additionally have described various technological means of extracting this energy (for example see the recent paper by Anastasozki et al [19].). A theoretical book on zero point energy (and antigravity) was published by Bearden in 2002 [20]. There is significant evidence that scientists since Tesla have known about this energy, but that its existence and potential use has been discouraged and indeed suppressed over the past half century or more [21].

The coupling of the electrogravitic phenomena observations and the zero point energy findings are leading to a new understanding of both the nature of matter and of gravity. This is just now being discussed in scientific journals (though some evidence suggests that it has been understood for decades within the black project covert community). The question that is being addressed is: what keeps the universe running? Or more specifically, where do electrons get their energy to keep spinning around atoms? As electrons change state they absorb or release energy, and where does it come from? The simplistic answer is that it is coming from the vacuum state. Puthoff [22] describes the process as follows: "I discovered that you can consider the electron as continually radiating away its energy as predicted by classical theory, but **simultaneously absorbing a compensating amount** of energy from the ever-present sea of zero-point energy in which the atom is immersed. An equilibrium between these two processes leads to the correct values for the parameters that define the lowest energy, or ground-state orbit (see "Why atoms don't collapse," NEW SCIENTIST, July 1987). Thus there is a DYNAMIC EQUILIBRIUM in which the zero-point energy stabilizes the electron in a set ground-state orbit. It seems that the very stability of matter itself appears to depend on an underlying sea of electromagnetic zero-point energy."

Furthermore, it appears that it is the spinning of electrons that provides inertia and mass to atoms. These theories, linking electron spin, zero point energy, mass, and inertia have been presented in a number of recent papers, such as those by Haisch [23] and colleagues and provide us with a possible explanation of the Biefield-Brown effect. It appears that an intense voltage field creates an electromagnetic barrier that blocks the atomic structure of an atom from interacting with the zero point field. This slows down the electrons, reducing their gyroscopic effect, and thus reducing atomic mass and inertia, making them easier to move around.

Evidence of Extensive Antigravity Technology

The B-2 Advanced Technology Bomber

In 1993, LaViolette wrote a paper [24] discussing the B-2 bomber and speculating on its probable antigravity propulsion system, based on a solid understanding of electrogravitics, [25] the aircraft's design and the materials used in its manufacture. It appears that the craft is using a sophisticated form of the antigravity principles first described by T. Brown. Support for this thesis came from the Aviation Week and Space Technology (March 9, 1992), which reported that the B-2 bomber electrostatically charges its leading edge and its exhaust stream. Their information had come from a small group of former black project research scientists and engineers suggesting the B-2 utilizes antigravity technology. This information was supported by Bob Oechsler, an ex-NASA mission specialist who had publicly made a similar claim in 1990. These findings support the contention that there have been major developments in the area of antigravity propulsion which are presently being applied in advanced aircraft.

LaViolette later states the obvious that "the commercial airline industry could dramatically benefit with this technology which would not only substantially increase the miles per gallon fuel efficiency of jet airliners, but would also permit high-speed flight that would dramatically cut flight time." [26]

The Hunt for Zero Point [27]

This recent book contains some of the strongest evidence yet for major efforts and success in the field of antigravity technology. The author, Nick Cook, who for the past 15 years has been the Aviation Editor and Aerospace Consultant for Jane's Defense Weekly, spent the last 10 years collecting information for the book. This included archival research on Nazi Germany's antigravity technology and interviews with top officials at NASA, the Pentagon and secret defense installations. He shows that America has cracked the gravity code and classified the information at the highest security levels. **Because antigravity and its allied zero point energy technologies potentially offer the world a future of unlimited, non-polluting energy it has been suppressed because of the "huge economic threat".** His findings support those reported by many of the Disclosure Project witnesses cited above.

Antigravity Technology Demonstrations

Although T. Brown reported many of his findings nearly a half century ago, other experimenters have just recently begun to reproduce his work and report on it in the open literature and on the WWWeb. For example, Davenport [28] published the results of his work in 1995 supporting the findings of T. Brown, while Bahder and Fazi [29] in 2002 described their assessment of the forces associated with an asymmetric capacitor. Transdimensional Technologies [30] in the USA and J. Naudin [31] labs in France have posted on the WWWeb: diagrams, web videos, and data on their versions of antigravity "Lifters" based on an extension of Brown's work. It is a sad commentary on this whole area of research to see that public science is requiring us to demonstrate principles that were demonstrated nearly fifty years ago.

There have also been a number of other demonstrations of "antigravity" phenomena by researchers throughout the world. This includes the work of Brazilian physics professor, Fran De Aquino, and such devices as: the Searl Electrogravity Disc, the Podkletnov Gravity Shield and Project Greenglow, the Zinsser Kineto-baric Field Propulsion and the Woodward Field Thrust Experiments on Piezoelectrics. All of these are described in more detail by Greer and Loder. [32]

Implications of This Research

Antigravity and zero point energy research and their applications are finally being addressed by some of the open scientific community. This means there will have to be a rewriting of textbooks in this area so our new generation of students can apply this "new knowledge." Its application will lead to major breakthroughs in transportation technologies both earthside and in outer space. The implications are that we have the potential for human exploration of our solar system and beyond, if we have the will, within our lifetimes. It also means that the majority of 20th century space technology will be obsolete and in fact may already be so.

The zero point or vacuum state energy source is seen as a totally non-polluting energy source, which has the potential to replace all the fossil fuels on this planet. It also will provide the energy needed for long range space flights. This means that fuel cells and solar cells in common use today for space flight energy applications will only be needed until we transition to these new energy technologies.

Based on an analysis of trends in antigravity research over the last half-century and the information provided by numerous witnesses, it appears that there is both good and bad news. The good news is that it appears that we (at least covert projects) have already developed the theories of antigravity, and additionally have developed working spacecraft based on these principles. The bad news is that these technologies have been developed for at least several decades, at the public's expense and that human kind has been deprived of these technologies, while continuing to waste energy using less efficient and pollution enhancing technologies.

Supporting this contention is the following quote from Ben Rich, former head of the Lockheed Skunkworks. Just prior to his death, he stated to a small group after a lecture [33] that: "We already have the means to travel among the stars, but these technologies are locked up in black projects and it would take an act of God to ever get them out to benefit humanity…" He further went on to say that, 'anything you can imagine we already know how to do.' Strong words from a knowledgeable deep insider and words that support what a number of the witnesses stated as well.

As the reality of this knowledge begins to be understood, there will be an outcry among space scientists not on the inside for release of these technologies to allow all of us to explore space. There will be major changes in the way that NASA does its business, though predicting these changes is difficult. Not only has space exploration in the public sector suffered, but our planet's environment has suffered as well. Thus as this knowledge begins to sink in there will be an outcry among all concerned citizens on this planet for release of these technologies to allow all of us to reduce and ultimately eliminate global warming and environmental pollution that so threatens our way of life. These technologies will not only affect space travel technologies, but will also have a profound effect on transportation and energy production on the earth's surface.

In conclusion, we might consider the observation made by Halton Arp [34]: "We are certainly not at the end of science. Most probably we are just at the beginning!"

Acknowledgements

I thank the following people for many discussions while preparing and writing this paper including: S. Greer, A. Craddock, T. Bearden, P. LaViolette, M. McCandlish, D. Hamilton, T. Valone, E. Mallove, T. Loder, C. Loder, S. Patel and many of the courageous Disclosure Project witnesses.

References Cited

1. Quoted from: LaViolette, P. A. 2000. Moving Beyond the First Law and Advanced Field Propulsion Technologies. In T. Loder (ed.), "Outside-the-Box" Technologies, Their Critical Role Concerning Environ-mental Trends, and the Unnecessary Energy Crisis. Report prepared for The U.S. Senate Environment and Public Works Comm. www.senate.gov/~epw/loder.htm.
2. Brown, T. T. 1929. How I control gravity. Science and Information Magazine, Aug. 1929. Reprinted in Psychic Observer 37 (1): 66-67.
3. Aviation Studies (International) Ltd. 1956. Electro-gravitics Systems: An examination of electrostatic motion, dynamic counterbary and barycentric control. p. 14. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
4. Ibid. p. 27.
5. Ibid. p. 19.
6. Gravity Rand Ltd. 1956. The Gravitics Situation. p. 54. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
7. Aviation Studies (International) Ltd. 1956. Electrogravitics Systems: An examination of electrostatic motion, dynamic counterbary and barycentric control. p. 11. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
8. Ibid. p. 34.
9. Ibid p. 41.
10. Gravity Rand Ltd. 1956. The Gravitics Situation. p. 47. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
11. Aviation Studies (International) Ltd. 1956. Electrogravitics Systems: An examination of electrostatic motion, dynamic counterbary and barycentric control. p. 32. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
12. Evans, M. W. 2002. The link between the Sachs and O(3) theories of electrodynamics. In Evans, M.W. (ed.), Modern Nonlinear Physics, Pun 2. 2nd ed., Advances in Chemical

Physics 19:469-494.
13. Anastasovski, P.K., T.E. Bearden, C. Ciubotariu, W.T. Coffey, L.B. Crowell, G.J. Evans, M.W. Evans, R. Flower, A. Labounsky, B. Lehnert, M. Mészáros, P.R. Molnár, S. Roy, and J.-P. Vigier. 2001. Anti gravity effects in the Sachs theory of electrodynamics Foundations of Physics Letters. 14(6):601-605
14. Alcubierre, M. 1994. The Warp Drive: Hyper-fast travel within general relativity. Classical and Quantum Gravity, 11, L73.
15. Puthoff, H. E. 1996. SETI, The Velocity-of-Light Limitation, and the Alcubierre Warp Drive: An Integrating Overview, Physics Essays 9:156.
16. Lamoreaux, S.K. 1997. Demonstration of the Casmir force in the 0.6 to 6 μm range. Physics Review Letters. 78:5.
17. Puthoff, H. 1989. Gravity as a Zero-Point Fluctuation Force." Phys. Rev A., 39(5):2333-2342. Puthoff, H. 1989. Source of Electromagnetic Zero-Point Energy." Phys. Rev A, 40(9):4597-4862.
18. See the Tom Bearden web site for an extensive listing and copies of his papers at: www.cheniere.org.
19. Anastasovski, P.K., T.E. Bearden, C. Ciubotariu, W.T. Coffey, L.B. Crowell, G.J. Evans, M.W. Evans, R. Flower, A. Labounsky, B. Lehnert, M. Mészáros, P.R. Molnár, J.K. Moscicki, S. Roy, and J.P. Vigier. 2001. Explanation of the motionless electromagnetic generator with 0(3) Electrodynamics. Foundations of Physics Letters, 14(1):87-93.
20. Bearden, T. 2002. Energy from the Vacuum: Concepts and Principles. Cheniere Press, Santa Barbara, CA.
21. Valone, T. 2000. The Right Time to Develop Future Energy Technologies. in T. Loder (ed.). "Outside-the-Box" Technologies, Their Critical Role Concerning Environmental Trends, and the Unnecessary Energy Crisis. Report prepared for The U.S. Senate Environment and Public Works Comm. www.senate.gov/~epw/loder.htm.
22. Puthoff, H. 1990. Everything for Nothing. New Scientist, 28 July 1990, pp. 52-55.
23. Haisch, B., Rueda, A. and Puthoff, H. 1994. Beyond E =mc2; A First Glimpse of a Postmodern Physics, in which Mass, Inertia and Gravity Arise from Underlying Electromagnetic Processes. The Sciences, 34:26. Haisch, B., Rueda, A., and Puthoff, H. 1997. Physics of the Zero-Point Field: Implications for Inertia, Gravitation and Mass. Speculations in Science and Technology, 20:9. Haisch, B. and Rueda, A. 1998. An Electromagnetic Basis for Inertia and Gravitation: What Are the Implications for 21st Century Physics and Technology? in El-Genk, M.S. (ed.), Space Technology and Applications International Forum-1998, DOE CNF-980103, CP420, p. 1443. Haisch, B. and Rueda, A. 1999. The Zero-Point Field and the NASA Challenge to Create the Space Drive. Proc. NASA Breakthrough Propulsion Physics Workshop, NASA/CP-1999-208694, p. 55.
24. LaViolette, P. 1993. The U.S. Antigravity Squadron. p. 82-101. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.p.82-101.
25. LaViolette, P. A. 1992. Electrogravitics: Back to the future. Electric Spacecraft, Issue 4, pp. 23-28. LaViolette, P. A. 1993. A theory of electrogravitics. Electric Spacecraft, Issue 8, pp. 33-36
26. LaViolette, P. A. 2000. Moving Beyond the First Law and Advanced Field Propulsion Technologies. in T. Loder (ed.). "Outside-the-Box" Technologies, Their Critical Role Concerning Environmental Trends, and the Unnecessary Energy Crisis. Report prepared for The U.S. Senate Environment and Public Works Comm. www.senate.gov/~epw/loder.htm.
27. Cook, N. 2001. The Hunt for Zero Point: Inside the World of Antigravity. Broadway Books, NY. 256pp.
28. Deavenport, L. 1995. "T.T. Brown Experiment replicated. Electric Spacecraft Journal. Issue 16. Oct. 1995. (Reprinted in: Valone, T. (ed.), 1994. Electro-gravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005)
29. Bahder, T.B. and Fazi, C., "Force on an Asymmetric Capacitor," Army Research Laboratory, v51 ARL-TR.nb. Avail. at: http://xxx.lanl.gov/abs/physics/02110. 2002.
30. Transdimensional Technologies, 906-E Bob Wallace Ave., Huntsville, AL 35801.
31. http://jnaudin.free.fr
32. Greer, S.M. and T.C. Loder III. 2001. Disclosure Project Briefing Document, 492 pp. Available on CD from: The Disclosure Project, P.O. Box 2365, Charlottesville, VA 22902. Also available from: www.disclosureproject.org.
33. Lecture given at UCLA Engineering Department on March 23, 1993
34. Arp, H. 1998. Seeing Red: Redshifts, Cosmology and Academic Science. Montreal, Aperion. (p. 249).

Congress "The Time Machine"

Faraday Labs Ltd invites to participate in scientific congress devoted to experiments on changing of space-time properties. It is planned April 12, 2003, Moscow. Main topics of the congress are time and gravitation in the context of eatherodynamics, experiments and applied aspects of these technologies. The main report is "Method to Control Temporal Parameters of Physical Processes" by Alexander V. Frolov.

Organizing committee: Vadim A. Chernobrov (KOSMOPOISK research center) and Alexander V. Frolov (Faraday Labs Ltd). The registration fees are not required.

Please, contact us http://www.faraday.ru or **email** congress@faraday.ru **Phone/fax 7-812-380-3844**

Please send this pre-registration form by post: P.O. Box 37, St. Petersburg, Russia 193024 **or e-mail:** congress@faraday.ru

PRE-REGISTRATION FORM **for participation in the congress "The Time Machine"**

Last name:
First name:
Title/ Function:
Company/ Organization:
Address:
Postcode:
City:
Country:
Tel.:
Fax:
e-mail:

☐ I would like to submit an oral presentation (report). (Please, enclose one-page abstract).

☐ I intend to participate as a guest

☐ I'm interested in further information.

Date Signature

COMMERCIALISING the “SEARL EFFECT” for ENERGY and PROPULSION SYSTEMS

Review from web site http://www.sisrc.com/
Email address: admin@sisrc.com

The “Searl Effect”, discovered by J.J.R. Searl, is a method of extracting energy. Some current expressions for the source of this energy being The Space Fabric, Quantum Space Field, and Zero Point Energy.

SISRC Ltd has been formed to develop and license Searl Effect Technology (SET) on a global basis.

THE COMPANY

SISRC Ltd is the company formed to administer the implementation of the Searl Effect Technology, (SET) invented by J. R. R. Searl. SISRC Ltd designs, develops and implements Searl Effect Technology as it is initiated in various applications and territories. SISRC Ltd is the administrative hub of the group and will continue to be based in the UK. SISRC Ltd will grant production and marketing licenses for Searl Effect Technology applications to different companies in specific territories. Currently proposed related companies include:

* SISRC -Germany, SISRC -Iberia, SISRC - Sweden, SISRC - Australia, SISRC New Zealand.
* SISRC -AV (Audio Visual) provides computer graphics presentations of the SET.

BACKGROUND

The Searl Effect Generator (SEG) technology, as applied to the commercial market, had been previously developed to the point where a few prototype SEG generators were made, and used for electricity generation and motion. Commercial interest at that time focused on the SEG’s transport potential and, under commercial pressure to deliver a fully functioning system, the original generators were used and lost in a series of vehicular propulsion experiments and demonstrations. Funding was insufficient to continue with the manufacture of the required larger-scale pressurized cabin machines - resulting in the termination of the project at that time.

Although all the operating principles, precise proportions and weights of the required materials are known for three of the four required operating materials, the precise data of the original magnetic layer is uncertain. The objective of the current R&D programme is to re-establish the original magnetic layer using modern and more efficient materials.

Originally, the layered materials were constructed and magnetized by the now-defunct Midlands Electricity Board under the direction of John Searl (see colored photo of the experimental craft construction on the cover page).

Modern magnetic materials have advanced considerably, and old ones discontinued, so a series of tests need to be conducted to establish the optimum materials and processes. These tests need to comply with the working criteria required and must lead to a cost-effective manufacturing process.

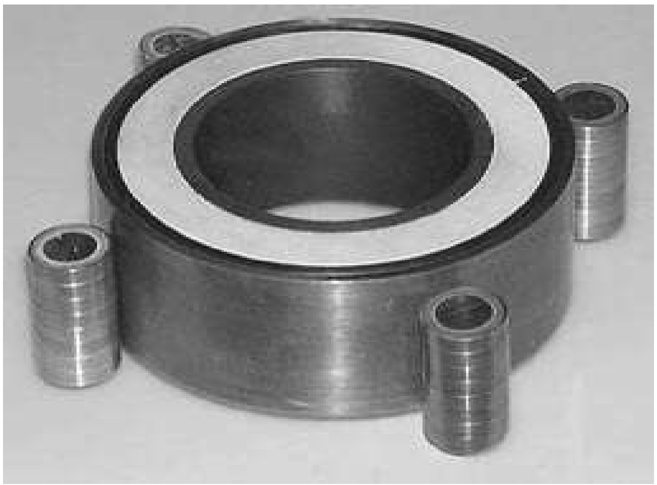

Fig.1
Experimental ring and rollers

In recent times SISRC has been re-establishing the original research. Due to the very limited funding that has been available, only a partially functioning demonstration prototype of the SEG principles has been possible. This prototype consists of the innermost of the three composite rings required and several rollers.

We would like to hear from anyone who worked on or was involved with Searl Technology prior to 1983. Also anyone who has any old technical data, photographs or films relating to the technology. Such information could greatly assist this technology to reach the marketplace for the benefit of all and would be dealt with in the strictest of confidence.

TECHNICAL DESCRIPTION

Physically the Searl Effect Generator (SEG) consists of three concentric rings each made of a composite of four different materials which are also concentrically attached to each other. The three rings are fixed to a base. Surrounding each of the rings, and free to rotate around them, are rollers - typically 10 on the first ring, 25 on the next and 35 on the outer ring. Surrounding the rollers on the outer ring are coils which are connected in various configurations to supply either AC or DC current at a variety of voltages.

Multiple magnetic poles are imprinted on the rings and rollers to form frictionless magnetic bearings. These also arrange the static charge into opposing charge clusters which cause the rollers to rotate around the circumference of the ring (as shown in Fig. 2).

Some expressions currently in use to describe the source of the energy for the SEG are The Space Fabric, Quantum Energy field of Space or Zero Point Energy. This is an unlimited and constant source of energy which can be made to flow when the correctly proportioned masses concerned are stimulated by the correct frequencies creating an 'open system'.

The idea of utilising this source of energy is currently the subject of various devices and experiments such as the 'Lamb shift', 'Casimir Effect' and the work of the Russian Nobel Prize winner Ilya Prigogine. However these devices and experiments tend to only prove the existence of the energy and not a method to create a coherent, ordered flow to produce useful power.

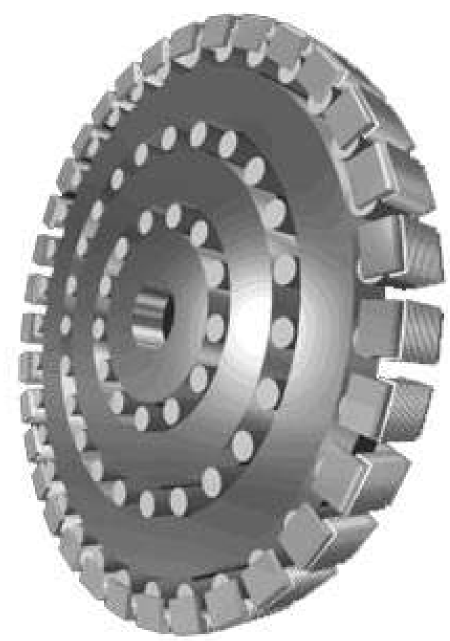

Fig. 2

15 KWt Searl Effect Generator (SEG)

In contrast, John Searl has discovered that, in order to create a steady and stable flow, all the masses of the device (and the stimulating frequencies) must conform to precise values determined mathematically by the 'Law of the Squares'. A machine constructed to these principles produces a stable and useful power output.

DEFINITION OF A SURFACE OPERATING S.E.G. (J. Searl) 15.09.00

A Linear motor operating on a magnetic bearing with the characteristics of an auto-transformer. The S.E.G. is defined as a device, which is constructed from 2,124 component parts, which make up 3 plates and 66 roller sets (see figures on the cover page). 12 of the 2,124 components create the 3 plates, which act as reaction components. 2,112 components are employed to create 66 roller sets, which act as active components. Each roller set consists of 8 segments constructed from 32 components. Each segment cannot be less than 34 grams in weight, as the law of the squares defines that value as the lowest value suitable for the reasonable generation of electricity. Therefore the smallest roller set must weigh 272 grams, making a grand total of 2,244 grams per roller set for the smallest surface bound SEG.

Tolerance should be about 0.05 gram per roller set. The larger the error the greater is the loss of power within the unit. Therefore the total tolerance for the smallest SEG roller set must not be greater than 3.30 grams - above this factor the SEG will not function. **For the best results, the tolerance across the total 66 roller sets should not exceed 0.05 grams!**

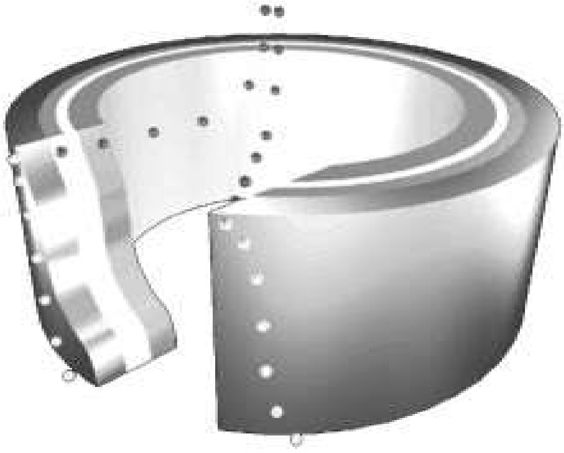

Fig.3

The S.E.G. is a step up rotary transformer. In fact it's a 'three rotary transformers' operational system, within a unit. The inner transformer output is fed in to the input of the second plate which increases the output of the second unit, which output feeds as an input to the final unit, thereby increasing its finaloutput to a very high voltage. Therefore the S.E.G. is just a prime mover for the production of clean electrical energy. The output windings must be designed to step down this output to 240 volts at 15KWt's.

During 1968 it being upon this issue, as development progress both here and in the USA, new findings are being created, and updates will be added to this page, as they are certified. When the SEG rollers are brought into close proximity to the SEG Ring, the Searl Effect resonant magnetic field causes negative ions and electrons to be drawn into and accelerated through the machine. This process is assisted by the highly electron-attracting rare earth metal Neodymium.

The unique mechanical and material arrangement of the SEG pulses the neodymium to continually release and replace the surplus electrons to provide electrical or mechanical power, or both.

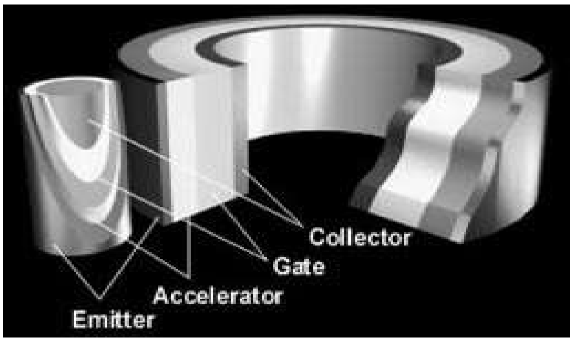

Fig.4

DEVELOPMENT

Prior to the production of complete 15KWt Searl Effect Generators (SEGs), a minimum of two demonstration Ring & Roller prototypes are to be produced.

The "Ring" is typical of the innermost of the three rings required to make a complete SEG. The 69-roller sets in a complete three ring SEG are identical. Six roller sets will be made for the prototype inner rings so that the Searl Effect can be demonstrated.

The production of a self-running Ring & Roller prototype will establish the precise sequence and parameters of electrical, magnetising and material processes required for the SEG and will demonstrate electrical power generation. An operating single inner ring with 1-8 rollers will also determine if a particular material will function correctly and at what level of power output.

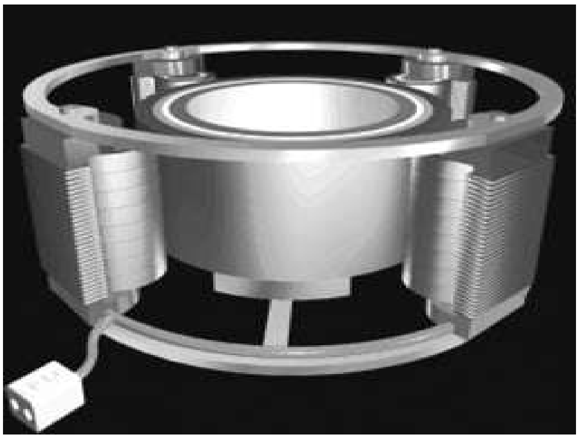

Fig. 5
Demonstration ring and rollers

The five phases of the development programme that commenced in October 1999 have been completed and the rollers successfully magnetised with 'Searl Effect' magnetic fields. A sixth phase is now required to apply the technology developed in Phase five to the "Ring". This sixth Phase resulting in a demonstration Ring & Roller prototype has duration of 6 months.

NEWS REVIEW

Scientific Breakthrough Liberates Energy Users from Fossil Fuel Dependence

Information from
http://www.genesisworldenergy.org/genesis_world_energy.htm

Story originally published by Diana Echeverria, USA

Director of Public Relations
Guy Rome & Associates, Inc.
208-345-4143 208-602-0325 (cell)

Technology breakthrough harnesses energy from the molecular structure of water

BOISE, ID - Dec. 5, 2002 - Genesis World Energy, a privately funded consortium created by a group of military and space program research and development specialists, today unveiled a scientific breakthrough that allows consumers to easily access the energy contained within the hydrogen and oxygen molecular structure of ordinary water. This scientific breakthrough provides a limitless, low cost and environmentally clean source of energy that can be implemented with minimal cost and effort. The viability of using water as an energy source, previously a theoretical concept, is now a reality. "Water has always been the source of life on this planet, now it will also transform the way we create energy" said Charles Shaw, corporate counsel and spokesperson for Genesis World Energy. "The implications for worldwide energy generation and consumption are nothing less than staggering."

The Edison Device

The first application of this technology is represented in the "Edison Device", a self-contained, self-sustaining energy generation unit. Roughly the size of an outdoor air conditioning system, the power source can be quickly and easily installed in any home or business to provide virtually unlimited energy from any available water source. The Edison Device utilizes the existing electrical wiring and natural gas plumbing in a home or business to replace the energy provided by utility companies. The home version of the Edison Device produces approximately 30 kWtt of combined gas and electrical energy per day. By comparison, the typical home uses between five to six kilowatts per day. The commercial model is capable of producing 100 kilowatts of energy per day. The energy generation portion of the devices has no moving parts. In fact, the only "mechanical" aspects of the equipment are small circulation pumps and micro-valves, making the Edison Device both silent and virtually maintenance-free. A minimum amount of water is used over an estimated 20+ years of service life.

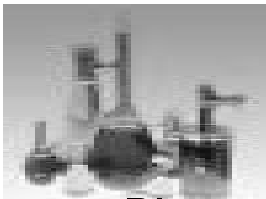

Interesting Publications of the Last Year

Discovery by a Group of Scientists Headed by Valerian Sobolev

Some time ago mass media announced that a sensational discovery was made by a group of Russian scientists headed by Valerian Sobolev. This discovery aroused ambiguous attitude of physic community. We have written about this fundamental and, to a certain extent, revolution discovery. Let us revise the fact that Sobolev claimed 7 scientific discoveries which were made, i.e. the process of depletion (a special electrochemical process), magnetic discharge, a new energy source, a method of low temperature plasma generation and superconductor. A laboratory system of the experimentally disclosed process as well as its industrial prototype is easy to be realized in practice for creation of new energy sources and wide-ranging production of ultrastrong materials. In many cases these materials can replace existing constructional materials.

The materials produced as a result of the process are multielement chemical compounds which are new states of matter. This new state of matter has **a time-changing magnetic field that can be EMF source in a coil of a generator.** Ordered structures of matters which are in the new modified state are nothing but a magnet charge. Materials containing the magnet charge and representing a continuous matters are **new energy sources.** Due to the new state matters become able to produce electric power as well as to generate low temperature plasma. It is easy to be realized in the process of common technological procedures. That promises designing of propulsors for “unsupported” transportation systems in near future.

Basing on realization of the disclosed process and using new materials the group of Russian Scientists headed by Sobolev together with American businessmen has created superfine and flexible glass for packing. The glass was called as “strong glass”. It seems to demonstrate higher pressure stability than steel. A method of production of these materials was patented in 1999. In the patent the scientists of Valerian Sobolev’s group are represented as authors of this new method, and the owner of the patent is Dynelec Corp (Columbus, Ohio, USA). You can read about the patent at the Web-site of http://patft.uspto.gov/netahtml/srchnum.htm (patent #5,964,913, October 12, 1999).

Sobolev’s group has appeal to the Russian government for sponsorship; however only foreign investors have provided funds for new energy sources developing. According to Russian Information Agency “News”, Sobolev’s group has signed a $168 million contract with one of Canadian companies for developing industrial production of energy sources. Manufacturing of such energy sources **can change energy supply system all over the world**. According to the contract the Canadian company is going to finance building of at least two plants. One of the plants is going to be built in Russia, and the other one is planning to be established in Canada. Each plant is supposed to produce 70 thousand of the self-running energy sources a year. Power output of the sources will come to 3-10 kWt. In other words, in the nearest future every person will be able to buy such an energy source and to apply it in household.

Single-Wire Electric Power System

(see photo on the cover page)

Experimental Results of Russian Scientists

Scientists of the All-Russia scientific research institute of electrification of agriculture (VIESH) academician D.S. Strebkov, engineer S.V. Avramenko, dr., A.I. Nekrasov, post-graduate student O.A. Roshchin developed a new method and the equipment for transmission of electric energy on a single-wire line using resonant idle operation mode and reactive capacitive currents for transmission of active electric power. Tesla transformers and frequency converter were used at the experiments.

First there was created the experimental sample of single-wire electric power system (SWEPS) with **10 kW** electric capacity and 3000 V voltage. As a result, obtained SWEPS electric parameters a hundred times exceeded parameters of a usual two-wire or three-wire alternative and direct current line.

In the end of 2002 there was made an attempt to increase transmitted capacity by using of more powerful condensers, which were installed in resonant circuit. There was developed the electric technique of **20kW** single-wire electric power system (SWEPS) with two Tesla transformers. As a result parameters of a low-voltage winding of the step-down transformer were changed. They **exceeded parameters of a usual two-phase or three-phase ac line or dc line in two hundred times.** At room temperature effective specific transmitted electric power was 4 MW/mm^2 and specific current density was 600 A/mm^2. These parameters could be achieved for existing methods of electric energy only at use of special materials in a mode of low temperature superconductivity. The measurements which were made at wire with diameter of 1mm, 100 microns and 80 microns, demonstrated that parameters did not depend on diameter of a wire. There was also experimentally confirmed the property of a single-wire line to transfer active power without essential losses on line resistance. Irradiation losses at low frequency 3,4 kHz were small. In single-wire power system the 80 microns wire plays a role of directing system for an electromagnetic energy flow, which runs from the generator to the load.

“Lifter” Project

We present update information on the experiments with Lifters or asymmetrical capacitors, which use High Voltage to produce a thrust. ***By Jean-Louis Naudin it was demonstrated with “Maximus” experiment, that a Lifter can be scaled up and also that such a device is able to lift up to 60 g of payload (total weight: 194 g). According to Naudin, it is now possible to build a craft which will use the Biefeld-Brown effect to fly silently and without moving parts only powered by electrical energy.*** *It was also declared that on January 8th, 2003, there was successfully done two historical flights with a mouse as a test pilot of the Lifter “Maximus”* ***(propellantless electrokinetic craft)****. Below there is an analysis of electrogravitation experiments made by Jean-Louis Naudin and Tim Ventura. (See color photos on the cover page).*

Review of Electrogravitation Experiments made by Jean-Louis Naudin and Tim Ventura

Tim Ventura

Tim Ventura Website: http://www.americanantigravity.com
Email: TVentura@seattle.telecomsys.com, tventura6@atbi.com

There is a spreadsheet containing a jpeg-snapshot of data that I have assembled based upon Jean-Louis Naudin’s published results (JLN Labs Website: http://jnaudin.free.fr. and http://www.jlnlabs.org). In the graphs that I’ve created, it seems to indicate that current plays more of a role in propulsion than voltage does — i.e.: for a given amount of power in watts, raising the current and decreasing the voltage seems to create a higher level of thrust.

In light of this, I have been increasing the thrust of my own Lifters by creating an electrical-bypass of the “load-resistor” on my power-supply’s high-voltage output. This has an interesting result:

I use a current-driven power-supply, which means that if no load is connected to it the voltage will build up on the high-voltage output until arcing occurs. In my case the voltage will build up to around 65kV, at which point the power-supply will automatically shut down. Conversely, if a very light load (such as a short circuit) occurs, the opposite effect happens and the power-supply will deliver much higher current at a much lower voltage until the supply is overdrawn and once again shuts down.

Normally the output load-resistor constrains the current, and in doing so it maintains the voltage on the high-voltage output at the nominal 50kV output level. However, bypassing the output load-resistor allows the power-supply to deliver voltage and current that are based almost entirely on the type of load connected to it.

When I connect a normal Lifter to the bypassed power-supply, the voltage will build across the air-gap until ions begin to flow in a conduction-current across the air-gap. Although this begins to happen at approximately 15kV, it seems to have some type of “peak efficiency’ at approximately 22.5 kV. The amount of current climbs to approximately 11mA at this voltage.

This method of experimental setup seems to allow the Lifter to “find its own sweet-spot” for operation, instead of using the output-resistor to “force” a specific voltage on it for operation. This appears to maximize the thrust output during operation.

Lifter Efficiency Spreadsheet

Description	Lift capacity (grams)	Length (cm)	Efficiency (g/m)	Lifter Weight (grams)	Voltage (kilovolts)	Corona Air-Gap (centimeters)	Current (mA)	Power (watts)
Lifter 1	3.3	60	5.5	2.3	41.9	3	0.57	23.9
Lifter 2	9.6	180	5.3	6.6	43.35	3	1.12	48.5
Lifter 3	20	360	5.6	16	27.5	3	2.53	69.5
Lifter 4	36	720	5	32	44	3	2.01	132.9
3-Stage Lifter 3	54	1080	5	24	30	4.5	8.06	240
Coliseum Lifter	90	2160	4.16	50	31.769	4.5	8.06	254.15

New Electric Fire Technology

Valery D. Dudyshev

Doctor of Technical Science, Professor,
Corresponding Member of Russian Ecological Academy
ecolog@samaramail.ru
http://www.intrasoft.ru/nizot

Abstract

The article proves importance and prospects of development of combustion technologies. In particular there is considered application of a new electric fire technology for environmentally appropriate combustion of any matters and gases. The technology uses electric fields as combustion catalyst. There is observed application of technology in heat-and-power engineering, transport heat engines, disposal units etc. [1, 2]. It is demonstrated that the new technology allows intensify combustion processes and increase their controllability (i.e. the control of temperature, gases pressure, gradient of heat conductivity, and so on). There is made a conclusion about availability of the technology to solve critical energetic and ecological problems of the civilization.

We hope that this article will excite interest of a wide circle of readers, i.e. of professionals in the area of combustion and heat technologies, physicists, ecologists, and those people who are interested in new scientific ideas.

Global Ecological Problems are the Problems of Ineffective Combustion of Hydrocarbon Fuel

The ecological problems become more and more serious and threaten to grow into world ecological catastrophe. The main cause of the atmosphere pollution in megapolises is imperfection of combustion technologies (of heat-and-power engineering, heat machines, transport engines, waste combustion…). It has been proved that their share of pollution of the planet atmosphere comes to 70-80%. The combustion technologies mean any technologies of combustion of fuel, matters and gases. The combustion technologies are the most popular technologies in the world. The modern civilization without the combustion technologies is impossible. There are many industries, which apply these technologies, such as heat-and-power engineering, transport, metallurgy, food industry, oil-and-gas refining industry, chemical industry, waste combustion neutralization.

Therefore the global ecological problems will not be solved until the mankind develops the combustion technologies. The article describes and discloses the essence of the new technology of ecologically appropriate and effective combustion of fuel and wastes of any type.

Why is it difficult for ecologists to conserve nature?

Modern methods and technology used for solving of the ecological problems consist in **analysis** of extent and sources of the environment pollution (ecological monitoring), and, moreover, in **refinement of atmosphere, water and soil, which are used in different technologies or have been polluted due to the imperfection of the technologies** (it is refinement from such toxic components as drinking and run-off water, exhaust and waste gases etc.).

Unfortunately, methods of measuring of the huge spectrum of toxic matters, which are thrown into the atmosphere at matters combustion, as well as the methods of their utilization are impartially difficult, expensive and imperfect. Even using of an ideal technics for measuring of the atmosphere pollution is not effective since it is a struggle with **consequences** of the combustion and other technologies imperfection but not with **causes** of the atmosphere pollution.

Ecological and energetic effectiveness of the known combustion technologies

It is known that real effectiveness of transformation of chemical energy of fuel in the combustion process is low. For example, in heat engines it comes to 25%, effectiveness of transformation of heat energy into electric energy in thermoelectric power stations does not exceed 40%. If energy consumptions of mining, processing and delivery of the fuel to consumers are taken into account then the summary efficiency of the existent combustion technologies (of heat machines and devices) comes to no more than 10 – 15%! It means that more than the half of chemical energy of the fuel is transformed into heat and different toxic matters and waste gases, which pollute the planet atmosphere. They cause acid precipitation, a global “greenhouse” effect, which threaten to cause the climate warming, the world flood, and final poisoning of the living nature. Thus the share of imperfect power engineering in the nature pollution comes to 70 – 80%! How is it possible to cleanse the planet atmosphere, especially the environment of world megapolises, in the situation when the amount of toxic matters produced by transport and industries is comparable with the amount of rest pure air?

Until the mankind learns burn matters and gases by effective and ecologically appropriate way the planet atmosphere will remain polluted and in near future it can become uninhabitable. Thus we will not solve the ecological problems until we gain the understanding of the combustion processes and processes of combustion transformation of matter chemical energy into heat energy, and then into other useful types of energy (i.e. electric energy, mechanical energy, light energy).

Physical Essence and Problem of Classical Combustion of Matters

Combustion is one of the most difficult phenomena, which are known by the humankind. From the scientific point this phenomenon is a chain reaction of sequential fragmentation of fuel particles into smaller charged radicals; it is physical chemical processes of transformation of chemical energy of intermolecular connections as well as combustion also includes physical processes of transformation of energy into heat and light on molecular and atomic levels. Many other processes, which proceed simultaneously, are involved.

From school years we know that combustion is a process of interaction of fuel with an oxidant that is accompanied by heat and light energy generation. In higher school the words of "as well as by cryptic energy of chemical connection of waste gases" are added to the school definition. The combustion processes are studied and improved by scientists and experts of different areas (chemists, physicists, heat-and-power engineering specialists, thermalphysicists etc.). There are known fundamental investigations of combustion chain reactions made by such Russian scientists as N.N. Semenov, Ya.B. Zeldovich and their followers.

Until now intensiveness of fuel combustion is increased by air blowing into the combustion zone that increases the amount of waste toxic gases thrown into the atmosphere. **Let us arouse several questions, which seem naive at first sight. Why is an oxidant (air or oxygen) needed for matter combustion? Is it possible to do it without any oxidant?** How does the combustion process begin and proceed? There are a lot of vague questions in physics of combustion. For example, how can the temperature and the intensiveness of the combustion be regulated? Can the heat conductivity of the flame be controlled? How can heat motion of particles be regulated in the flame and in the waste gases, and what can it cause? There is another problem of combustion. It is the very hydrocarbon fuels, which are applied at modern heat processes. The great Russian scientist, D.I. Mendeleev stated that to use oil is the same thing as to stoke a stove with banknotes.

Since the hydrocarbon fuels are complicated chemical matters and the combustion processes are imperfect then in the process of their combustion a great amount of different by-product matters and toxic gases are produced. They waste unused self-energy of fuel into the atmosphere and pollute our planet.

Physical essence of the new electric combustion technology

How is it possible to burn the hydrocarbon fuel by environmentally appropriate way? How can this environmentally appropriate technology be realized in practice? A brief answer is following**: it is necessary to make such conditions of combustion of organic fuel, and to introduce such a combustion catalyst that the energy of intermolecular and intramolecular connections of the organic fuel becomes absolutely free. At that the chemical energy can be transformed into energy of electromagnetic radiation, for example, into light energy and heat energy without generation of by-product polluting matters and gases. Then a working body (for example, water) is effectively heated by this directed concentrated electromagnetic radiation. In this process effects of electromagnetic waves reflection and concentration are used, or the electromagnetic radiation is directly transformed into electric energy.** In this case there is a new opportunity to regulate flame temperature and to extremely decrease quantity of the oxidant, i.e. to create organics combustion with minimal quantity of waste gases.

Is it principally possible to burn organic matters and gases without the atmosphere pollution using electric field as a catalyst? It is possible if parameters of this field are correctly chosen. More precisely it is possible if huge Coulomb forces of the electric field are correctly regulated by interaction of electrically charged particles of the fuel and the oxidant with force lines of the electric field.

One variant of an experimental device is represented in Fig. 1.

Devices for approbation of the new electric combustion technology

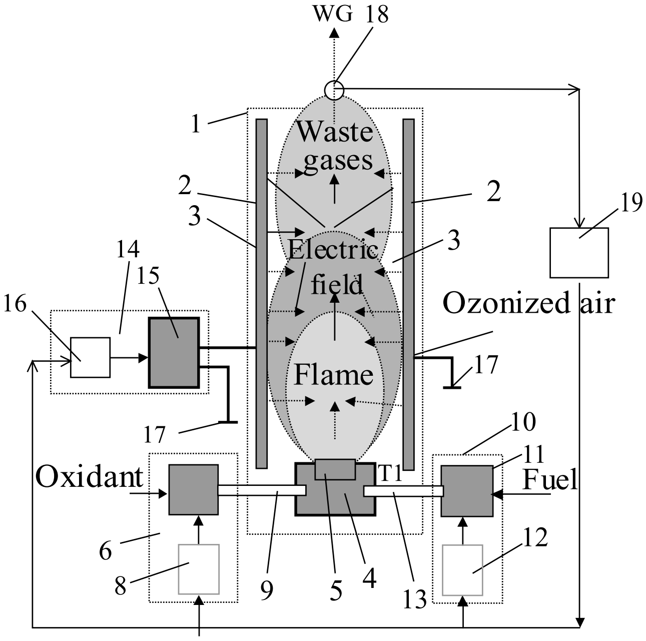

Fig.1 a
Plan of the experimental device

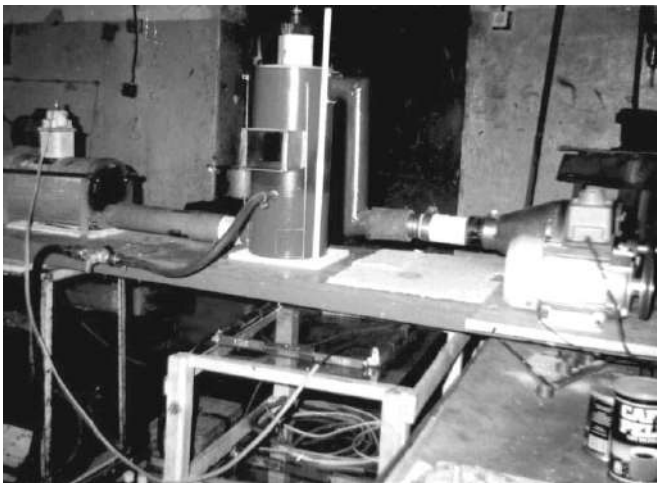

Fig. 1 b
Photo of the experimental system

A structure chart of the experimental device, which was designed to investigate the new electric combustion technology, is demonstrated in Fig.1a. The device contains a body (1) equipped with flat electrodes, which are insulated from the body (2), (they can be placed in either horizontal flat or vertical flat). The electrodes (2) are located on opposite inner walls of a combustion chamber (3). The device is equipped with a mixer (4) with a device for supplying of blended fuel to the zone of combustion. At the same time the device (5) is used for electrical combustion of the blend. The described device is equipped with an air track (6), which contains an oxidant activator (7) joined with a regulator (8) by a control circuit and with the mixer (4) by an air pipe (9). The regulator (8) serves for regulation of the oxidant activation extent. At the end of the air pipe it is possible to place a vortex device (it is not shown in the figure). Also the device is equipped with a fuel track (10) containing a fuel activator (11) and a regulator (12) of the activator and consumption of fuel. The block (11) is joined with the mixer (4) by a fuel pipe. The device is equipped with an electric combustion activator (14) containing a high-voltage transformer (15) of intensity and frequency joined to its regulator (16) by a control circuit. An electric outlet of the block (15) is joined to one of the electrodes (2), and the second outlet is safely electrically grounded by a grounding mat (17). The high-voltage wire of the outlet of the block (15) is connected with the electrode (2) through a bushing insulator (it is not demonstrated in Fig.1a). The device is equipped with a gas toxicity calculator (18) joined to the inlet of a mode optimizer (19), which is a control system for interacted regulating of all the parameters of combustion. For that the outlet of the mode optimizer (19) is joined to the inlets of control of the regulators (8), (12), (16).

The device operates in the following way. At first an activated oxidant of O1 is applied through the air track (6), activator (7) and air pipe (9) to the mixer (4). Then an activated fuel of F1 is applied through the fuel track (10) and the activator (11) to the same mixer (4). A blended fuel is prepared in the mixer (4) and then is atomized and combusted by a sparkle of the block (5). In the combustion chamber (3) flame and waste gases are processed by strong alternating electric field, which is generated by the block (14) in a gap between the electrodes (2). In the process of combustion waste gases toxicity is measured by the special toxicity sensor (18). Depending on current toxicity the parameters of flame combustion are regulated by the mode optimizer (19). More precisely the oxidant consumption and the extent of its activation are changed by the regulator (6), the fuel consumption and the extent of its activation are changed by the regulator (12), the intensity and frequency of the alternating electric field generated by the block (14) in the combustion chamber (3) is changed by the regulator (16). Let us note that the electric field in the combustion chamber (3), viz in the gap between the electrodes (2), catalytically influences on both the flame and the waste gases. The essence of the process lies in the fact that the processes of fragmentation and oxidation of fuel radicals and of toxic oxide molecules are accelerated. Intensiveness of combustion and of toxic gases refinement increases as well as the intensity of this field and its frequency do. As a result of the interconnected regulation of all the listed parameters fuel of any kind can be fully, intensively, “cleanly” combusted.

In Fig.1b there is a photo of the operating experimental device designed to investigate the processes of influence of electric field on the process of matters combustion and of cleaning of waste gases. The photo represents following things.

1. In the left side there is a furnace with a high-voltage electrode in a higher part of its body.
2. In the centre there is a vertical column of the electric combustion cleaning (after-burning) of toxic gases in the electric field; on the top of the column a high-voltage electrode can be seen. Gas pipes designed for toxic gases supply–withdrawal are connected with the column and located to the right and to the left from it.
3. A regulable blower is represented in the upper right side of the picture. It is designed for toxic gases withdrawal and connected with the gas pipe by the column.
4. At the foot of the picture there is a regulable high-voltage tension source (the electric field source), which is connected with the furnace and the column of toxic gases cleaning by high-voltage wires.

It is experimentally proved that combination of two stages of the combustion activation (i.e. in the furnace and the column) ensures ideal cleaning of toxic gases at combustion of any toxic matters.

Some Results of Experiments

Our experiments and investigations of many-sided influence of electric and high-voltage electromagnetic fields on the combustion process have proved that such practically ideal condition of fuel and matters combustion is possible to be realized in practice [1].

Low-powered static and alternating electric fields (i.e. the fields of constant sign and of variable sign) of more than 1kV/cm intensity were used as combustion catalysts as well as high-frequency electromagnetic fields of low power with some frequency of oscillations of molecules in the flame (flame power is 0.1 – 1% of heat power of the flame of the combusted organic fuel).

Peculiarities of Oxidants Application in the New Technology

It is known from the thermodynamics and combustion theory that optimal ratio of the oxidant mass, for example air, and the fuel is approximately 1:16 in an average combustion process. In our experiments with the electro-field catalyst the ecologically appropriate combustion of the hydrocarbon fuel (mazut, straw oil) was achieved at oxidant deficiency (for example, at the ratio of oxidant mass and the fuel of 1:1). For experts it means that there is a real possibility for 10 – 15 times decrease of amount of waste gases of any heat machines and to while their former power is the same. As the experiments demonstrate carbon and hydrocarbon are absolutely removed from the waste gases. In the experiments oxides of hydrogen and nitrogen were 4 – 8 times decreased and flame existence was 5 – 10 times increased.

The issues of the experiments prove the hypothesis about possibility of effective environmentally appropriate combustion as a process of direct transformation of chemical energy of organic fuel into electromagnetic radiation energy of the flame (including heat and optic diapasons). The transformation is accompanied by removal of polluting toxic components from the waste gases that occurs due to many times intensification of the combustion process under action of electromagnetic catalysts.

Operations of regulable activation of fuel oxidants (simultaneous or separate), which are introduced into the combustion technology, ensure additional improving of the combustion process. Especially it occurs at heavy oil fuel combustion and water-fuel emulsion combustion that was experimentally tested by us. The process becomes especially effective if the flame of the combusted activated blended fuel is additionally processed by alternating electric field. Due to introduction of the operation of all the combustion parameters regulation (of consumption of fuel, oxidant, of their activation extent, and of their combustion intensiveness extent), according to the information of the waste gases toxicity, it is possible to achieve effective combustion of fuel and waste products of practically all kinds.

Our experiments demonstrate that combination of the very flame procession with procession of the waste gases and air (i.e. an oxidant) by electric field is very effective for the waste gases cleaning. The essence of this additional cleaning of the waste gases lies in fragmentation of carbon particles and exhaust opacity by electric forces of an alternating field as well as in after-oxidation of some toxic oxides in the medium of an ozonized oxidant. Energy consumption for activation of the flame combustion by strong electric fields is small and does not exceed 1–3% of heat energy of the flame. Advantage of this invention is universality of the application for combustion of any inflammables. It is possible due to widening of the diapason of the electric field parameters regulation (of intensity and frequency), especially in the mode of their interconnected regulation.

The essence of catalytic action of the alternating electric field on the flame combustion process consists in effective breaking of dipole radicals of fuel by an activated (dipole) oxidant. Moreover, the essence lies in better mixing of layers of combusting flame with the oxidant that occurs due to removal of a doubled electric layer from the limit line of the flame. Thus this technical solution allows achieve new positive effects due to its significant peculiarities, i.e. extending of application area of the known electric combustion method over the combustion process of any inflammables as well as it allows significantly increase controllability of the flame combustion process.

Let us note that in the experiments on the combustion of organic fuel in strong electric fields the regulation of the flame temperature and of its existence was achieved at unchanged consumption of fuel and of oxidant. That was achieved by changing of the parameters of the combustion electric-field catalyst (intensity and frequency) of a longitudinal electric (electromagnetic) field. Rotation and stabilization of the flame was achieved by a rotating transversal electric field. Changing of the flame height was realized by the longitudinal electric field.

The investigation, which has been performed by us, proves that directly acting on the flame the very electric field which directly influences on the flame and emission of electrons flow (ideal type of oxidation) into the flame can most effectively intensify the combustion process and make it harmless for the humankind and for the environment!

It has already been experimentally proved that energy consumption required for generation and regulation of this electric field and of the electrons flow for intensification and environmental appropriateness of the combustion is quite small relatively the combustion energy and comes to fractions of a percent of the flame energy. Thus our experiments prove that the best **"oxidants" and combustion catalysts are** not superfluous air but **the electron and the electric field!**

Regulation of Heat Conductance of the Flame and the Heat Flow

As our experiments have demonstrated, the electric field can act as an effective combustion catalyst as well as a regulator of its intensiveness; moreover, it can control even a vector of heat conductance. It is proved by the experiments that the parameters of this field can be regulated as well as temperature of the flame and gradient of the flame heat conductance. Interesting experimental results were obtained measuring full heat of combustion of the same quantity of fuel at the usual method of fuel combustion and at the method, which uses combustion electric catalysts, even in the case of oxidant deficiency. In the last case the energy of fuel combustion almost 1.5 times increases that can be explained by fuller transformation of fuel chemical energy into electromagnetic radiation. At the usual combustion methods the chemical energy of the organic fuel was not completely used and remained as a cryptic summary chemical energy of intermolecular connections of many toxic waste gases, which were exhausted into the atmosphere by heat devices. Basing on the performed experiments it may be supposed that, evidently, specific heats of matters are 20-50% higher at this method than at their usual combustion method. The essence of the new electric combustion technologies consists in this new physics of combustion.

The author has already got patents for invention of Russian Federation [3 – 12] for the methods of regulation and intensification of matters combustion processes.

Some peculiarities of atomization, inflammation and combustion of organic fuel in electric fields

We have not discussed all the potential capabilities and advantages of the new electric combustion technology for different areas of technics. Let us demonstrate them in more details.

One of characters of the new electric combustion technology is **effects of oxidant ozonization, of electrostatic atomization and electrostatic injection of electrically charged particles of fuel and the oxidant that is followed by generation of the finest aerofuel opacity on a molecular level in a combustion chamber**. Naturally, such fine atomization of the fuel causes its easier evaporation, inflammation and combustion especially in the medium of the ozonized oxidant. **The mechanism of fuel electrostatic fragmentation is physically explained by Coulomb interaction of repulsion from each other of electrically likely charged fuel drops (particles) that is accompanied by their progressing fragmentation and corresponding decrease of their mass and electric charge**. As the experiments have shown the extent of the fuel drops (particles) fragmentation depends on initial electric potential of fuel charging and initial size of injected fuel drops (particles), which for their part depend on configuration and size of a fuel-injection nozzle as well as on pressure occurring in a fuel line of the heat machine.

Generated by dipolar high-voltage potentials catalyzing electric field can be introduced into the atomization zone or (and) the combustion zone of the combustion chamber of an engine. Its Coulomb forces accelerate motion of electrically charged particles of fuel and oxidant to an opposite electric potential, which is connected with, for example, a piston of an engine. Thus electric field acts as an electrostatic pump that allows decrease pressure in a fuel line as well as allows practically ideally atomize the fuel and mix it with the oxidant. Moreover, it lets simplify and improve the systems of injection and inflammation of the blended fuel in heat machines and devices, for example, in internal combustion engines or in boiler plants. Their application can cause additional improving of energetic and ecological characters of the combustion technologies. According to our information, such high-performance fuel-injection nozzles have yet been unknown in technics therefore they have not been applied in heat machines (injector internal combustion engine).

Is it possible to combust water? Water as a fuel!

More amazing experimental results were obtained at 40-80% dilution (emulsification) of liquid organic **fuel**, for example diesel oil, with **habitual water**. Summary energy as heat and light, which is generated in the process of such blend combustion, was not practically changed that can be explained by liberation of energy of chemical connections of water as well as of fuel. In the process of the experiments on combustion of the emulsion at first **it was transformed into the finest water vapor on the molecular level by means of capillary electroosmosis and of electrostatic atomization. Then dipolar water molecules, which had been electrified by field, were fragmented into hydrogen and oxygen, and then the hydrogen was effectively combusted in the medium of the ozonized oxygen.** Let us note that at this mechanism of water "evaporation" and of following splitting of water molecules into hydrogen and oxygen the electric field expends energy, and heat of organic fuel combustion just accelerates (catalyze) this process.

Editor: It is necessary to note that electric field can not expend energy. If there are no conduction currents then the field source does not decrease difference of potentials. Hence it can be concluded that effectiveness of such power systems is possible to be UNRESTRICTEDLY HIGH. The author writes about it below as about an "incomprehensible phenomenon".

The most amazing and incomprehensible fact is that the electric field transforms water into "vapor" and breaks the water molecule into hydrogen and oxygen practically **without any expenditure at minimum of energy consumption**. Electric power of the intensity high-voltage transformer, which is necessary for strong field generation, lies in the interval from several watts to tens of watts. As the experiments have demonstrated electric consumption of the high-voltage source practically is not changed at correct shunting of the flame by the electric field (i.e. through an air gap). Current consumption practically did not increase depending on mode of fuel atomization and its combustion (of flame height, of combustion intensiveness…).

There are only two conditions of effective work of static electric field as an electrostatic pump-atomizer of fuel and as a catalyst of blended fuel combustion. The first one is intensity of the field in the zone of fuel atomization and in the zone of flame combustion. The second one is enough emission of electrons into the blended fuel flow. In this mode Coulomb forces make work of electrostatic pressure generation and of electrostatic fragmentation of particles and molecules of fuel and water. The forces repulse likely charged drops of water. Heat motion of fuel radicals, which are combusted and fragmentized in the flame, impedes their chemical recompounding into water molecules in the zone of combustion. Moreover, it causes proceeding of a very physical chain reaction of hydrogen combustion in the flame.

Therefore additional light and heat energy, which is generated in the flame by combustion of hydrogen generated from water molecules that occurs in the ozonized oxygen, does not decrease summary energy of combustion of blend of fuel with water but increases it.

Thus using of electric and electromagnetic fields as the strongest catalysts of combustion of organic fuels and any matters clears great **prospects** of combustion technologies improving. Moreover it allows create ecologically appropriate heat machines and devices and 20-50% increase their effectiveness due to fuller transformation of chemical energy of inflammables into heat or light energy.

Ways of radical improvement of heat machines or about a new mechanism of transformation of heat energy into mechanical energy and into kinetic one

For clear realization of causes of extremely low output of modern heat machines that has generally lead the civilization to the ecological catastrophe it is necessary to understand their common principle of operation as well as a cause of heat energy non-effective using in them.

A mechanism of transformation of heat energy, which is generated by fuel combustion, is the same for all known heat machines. It consists in generation and transformation of superfluous pressure of a heated working body (i.e. gas, vapor etc.) into kinetic energy of a working element of the heat machine (i.e. a piston of an internal combustion engine, a turbine of an aviation engine, a reaction jet flowing out a rocket nozzle etc.) in special chambers of the machines.

It is also known that the higher the temperature of the heated working body, for example of a gas, the higher initial pressure in the working chamber of a heat machine. **Why output of heat machines is so low?** Any expert of thermal physics, thermodynamics, and heat machines can answer that efficiency (output) of heat energy transformation into mechanical one is defined by the second law of thermodynamics and is clearly illustrated by Carno heat cycle. According to these postulates, **output** of an ideal heat engine does not depend on a working matter and on the engine construction but is defined by temperatures of the working body in the starting point and in the final point of the cycle, i.e. by the temperatures of the heater and of the refrigerator of the **heat machine**. Real output of the heat machines is limited by heat stability of materials and by imperfection of engine constructions.

Nevertheless known thermodynamic processes and lows of heat motion of particles and molecules are significantly changed in strong electric fields and demonstrate new prospects for improving of heat machines.

A solution of the problem can be briefly formulated as: to increase output of heat machines it is necessary to use a new mechanism of redistribution of heat energy into energy of directed pressure of a compressed working body (i.e. gas, vapor, etc.) on the working element of the heat machine. At the minimal pressure it should be redistributed on sidewalls of the working chamber.

Can it be realized in practice? On the face of it the technical solution is unrealizable nevertheless it exists. It is possible to redistribute and to regulate heat motion (pressure) of heated gas inside a closed reservoir and, hence, its temperature in a certain direction. To achieve this aim **it is necessary to introduce static electric field into the combustion chamber of a heat machine and to orient heat motion of electrically charged and dipolar molecules of expanding working bodies (i.e. gas, vapor) along force lines of the electric field by electric Coulomb forces**. Something similar occurs, for example, in liquid crystal cells of electronic timer, in Kerr cells (electric optic effect) at periodical application of electric field to them that causes dipolar molecules turn along a field vector as well as change of light penetrability of these matters.

There is a difference between this known in electrooptics phenomenon and our case. The difference lies in the fact that polarized gas molecules remain movable along force lines of the electric field in contrast to liquid crystals, which do not. This fact causes redistribution of parameters of heat energy (i.e. pressure, temperature, and heat penetrability) of dipolar molecules of heated gas just along force lines of the field. The more the intensity of the field at the initial temperature of the working body is the more the difference of pressures on the sidewalls of the chamber and on the working element, for example, on an engine piston.

Heat motion of polarized particles of heated gas can be decelerated by combination of three electric fields of constant sign in accordance with all three coordinate axes; hence it is possible to significantly and quickly decrease temperature and pressure of the gas. In this case the heat energy of the heated gas makes jump transformation into electromagnetic radiation.

Editor: Historically this method, i.e. plasma retention by electric field, was proposed by Oleg Lavrentiev in 1948. He had 7-grade education and served as a sergeant in Armed Forces in Sakhalin. After he sent Stalin a letter containing a phrase, as "I know a secret of hydrogen bomb creation" he was invited to Moscow to set his ideas out to academicians. His ideas were not lost nevertheless the scheme of plasma electrostatic retention was not applied ("Expert" magazine #23, 18 of July of 2001). In 1950 Sakharov and Tamm proposed a scheme of a toroidal magnetic thermonuclear reactor which has been developed until now. Why is it so? It is difficult, expansive and practically unreal…

Thus introduction of strong electric fields into heat machines allows significantly increase effectiveness of transformation of heat energy of a working body into mechanical and kinetic energy of working elements of the machines by means of directed regulation of temperature and pressure of heated gas, for example, in the starting point and in the final point of a working stroke of a piston. In other words it allows increase output of the machines.

APPLICATION AREA OF NEW ELECTRIC COMBUSTION TECHNOLOGY

Practical application of the new combustion and thermal technologies

The new technology is universal and applicable in practically all spheres of technics. Therefore we believe that further development and application of this new technology of fuel and waste products combustion is very important for radical improving of all the heat machines and of all thermal technologies. As the civilization uses hydrocarbon fuel and matters in its life circle then just this new technology can solve critical ecological and energetic problems of the civilization.

It is impossible to view in details all proposed perspective technical solutions based on this technology [1-25] within one article. Therefore we give only one striking example.

Environmentally appropriate engine for motor transport

Since the transport, which uses thermal internal combustion engine (ICE), is most harmful for the environment then let us discuss ways of their ecological and energy improving. Several years ago the author patented "method of intensification of ICE operation" [7]. The essence of the invention consists in introduction of controllable electric field into combustion chambers during the whole operation period of an engine by special monoelectrode spark-plugs (Fig.2).

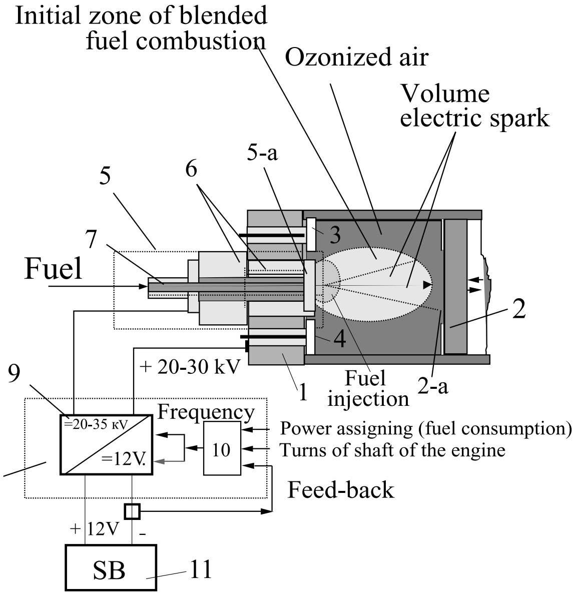

Fig.2a

Plan of an environmentally appropriate ICE (for transport)

1. – combustion chamber;
2. – piston, 2a –heat-resistant surface of the piston (zone of electric-spark discharge)
3. – admission valve (its canal is not indicated)
4. – exhaust valve (its canal is not indicated)
5. – combined mono-electrode spark-plug; 5a – fuel nozzle with hollow central electrode and end disk electrode
6. – electric insulator of the spark-plug – fuel nozzle
7. – fuel pump (for example, an electrostatic one)
8. – high-voltage regulable transformer (electric field source)
9. – force block of tension high-voltage transformer (=12V/25kV)
10. – system of regulation of electric field source (9)
11. – board current network =12V (SB is a storage battery of an automobile).

This structure chart briefly demonstrates basic components of a new ICE, where there is directed pressure of gases on a piston. In particular, many inventions of the author are realized in practice here, i.e. semi-digital spark-plug, electrostatic fuel nozzle (5), electrostatic fuel pump (7). A new mode of volume inflammation of blended fuel has been realized in practice, i.e. "a spark coming from a disk electrode (5-a) to a piston (2-a)". Since it is possible to regulate the electric field intensity generated by a high-voltage source of tension then an advance angle of blended fuel inflammation can be regulated, gases pressure on the piston at explosion stroke of ICE operation can be regulated as well.

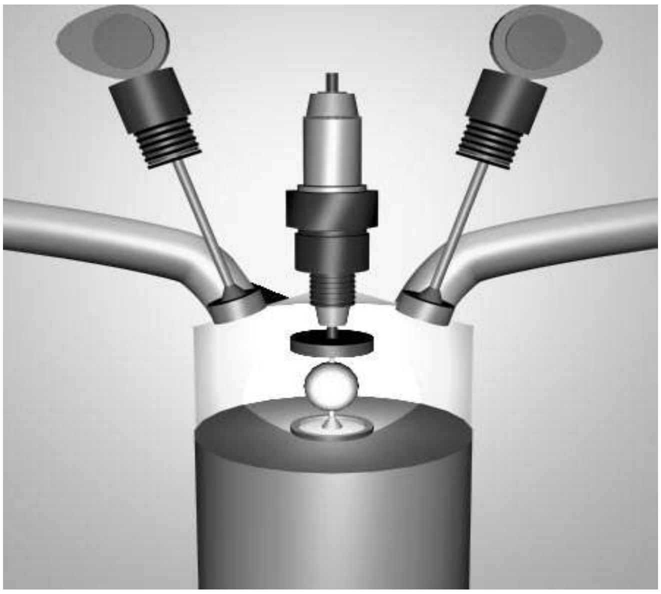

Fig.2b
Photo of a simplified construction of the new ICE

Fig.2b demonstrates simplified three-dimensional construction of the proposed energetically and ecologically perfect ICE. There are observable elements such as a combustion chamber, a piston, valves, a monoelectrode spark-plug, and an electrostatic fuel nozzle (in the centre on top), which is connected with the spark-plug. In the centre of the combustion chamber it can be seen volume inflammation of the blended fuel at spark appearance from end electrode to the piston.

As a result a summary positive effect of ecological, construction and energetic improving of ICE is achieved, i.e. effective cleaning of exhausted gases directly in the combustion chambers of the engine, significant simplification and perfection of a fuel injection system, improvement of system of distribution and electric inflammation of the blended fuel. In this engine a distributor (as well as its analogues) is removed at all since sparking and intensive inflammation of the blended fuel automatically appears between central electrode of the spark-plug and the piston which comes to the upper "dead" point. The central electrode of the spark-plug is constantly under high tension. The advance angle of ignition is regulated by change of the electric field intensity. Powerful multipoint ignition causes simultaneous intensive inflammation of the blended fuel that occurs throughout the whole chamber. The electric field as a powerful combustion catalyst intensifies the process of the blend combustion at the explosion stroke of the engine and at after-burning of waste gases directly in the combustion chambers at the following output stroke of the engine operation. Directed along the axis of the piston in the combustion chambers this electric field serves as a transformer of gases heat energy into mechanical energy of the thermal engine pistons. That is caused by the fact that the field orients heat motion of the expanding gases along the axis of the pistons in combustion chambers of the engine at explosion stroke of ICE just. It causes redistribution of heat energy and increase of the gases pressure on the very pistons that significantly improves effectiveness of transformation of heat energy of the fuel combustion into mechanical energy of the piston motion (theoretically the improvement is by two-three times), i.e. it two-three times increases output of a classical thermal engine making it come to 70-80%.

The electric field, which is introduced into the combustion chambers of an internal combustion engine, ensures significant economy of fuel (up to 30-40%) at saving of its working characteristics. It occurs due to fog electrostatic spray of fuel and to the fuel electrization as well as due to oxide ozonization. Moreover it occurs as a result of deep after-burning of hydrocarbon components of fuel, combusting blend and waste exhaust gases. This method allows effectively regulate temperature of the blend combustion in the chambers while required compression in cylinders is the same, for example, to minimize nitric oxide generation in the exhaust gases. As a result there is no necessity in external devices for cleaning of exhaust gases of internal combustion engines; ecological appropriateness improvement can be simultaneously achieved. Additional aerosol hydro-alkaline processing of these gases can allow totally clean **exhaust gases** of the transport, which operates on thermal engines.

Workability of such an environmentally appropriate engine in two-stroke and four-stroke variants is experimentally demonstrated. There are also experimentally shown decrease of all toxic components in exhaust gases, 20-30% increase of useful power of the engine, and simultaneous decrease of fuel consumption.

Thus practical application of the new electric combustion technology can exert revolutionary influence upon propulsion engineering development and upon increasing of ecological compatibility and effectiveness of the transport, which operates on thermal engines.

Diagram of operations of technology of effective using of refinery wastes in heat-and-power engineering

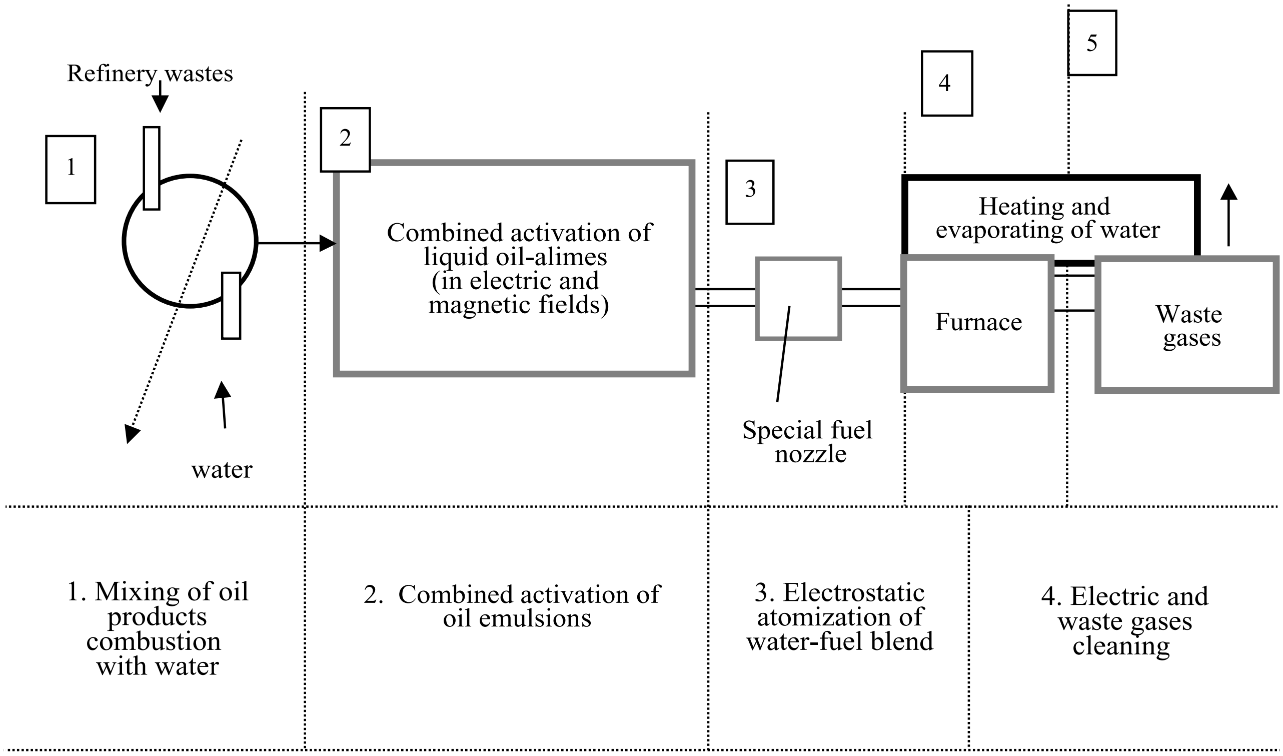

Refinery wastes
1
2
Combined activation of liquid oil-alimes (in electric and magnetic fields)
3
4
5
Heating and evaporating of water
Furnace
Waste gases
Special fuel nozzle
water
1. Mixing of oil products combustion with water
2. Combined activation of oil emulsions
3. Electrostatic atomization of water-fuel blend
4. Electric and waste gases cleaning

Fig.3

Technology of environmentally appropriate combustion of refinery wastes as fuel for boiler plants

The author has patented other technical solutions of improvement of different devices, which are based on these technologies. These solutions concerning boiler plants, gas turbine engines and jet engines and many other heat machines and devices allow increase their effectiveness and simultaneously significantly improve their ecological factors. As an example of application of this technology in heat-and-power engineering there is a diagram of operations of the technology of the environmentally appropriate combustion of refinery waste as a fuel for boiler plants (Fig.3). Unfortunately the limits of one article do not allow discuss these inventions in details nevertheless the author will realize it in practice in future works.

Conclusions:

1. The existent methods of solving of the ecological problems are ineffective as well as work of organizations, which deal with ecological and nature conservation activity. It is caused by the fact that they are oriented to search and remove consequences of ecological pollution of the nature instead of their causes.
2. The article demonstrates real causes of global ecological problems, which consist in imperfection of technologies of generation and transformation of energy.
3. Using of strong electric and electromagnetic fields of low power as combustion catalysts allow significantly improve ecological factors of combustion technologies and of devices, which realize the technologies in practice. In other words it allows realize in practice "clean combustion of fuels and wastes and possible existence of "clean" heat technics".
4. Application of electric and electromagnetic fields in the combustion technologies allows significantly increase effectiveness of transformation of chemical and heat energies of fuel into mechanical and electric energies.
5. The electric combustion technology allows realize in practice principally new ways to control combustion process as well as to control many thermal and kinetic processes (i.e. processes of pressure, heat conductance, temperature, etc.), i.e. allows increase output of heat machines.

6. The new electric combustion technology is a universal method of solving of ecological and energetic problems of the civilization and a universal way of radical improvement of technics (i.e. of the transport, boiler plants, oil-and-gas processing plants, combustion wastes recycling plants, etc.).

References

1. V.D. Dudyshev Electric combustion technology is an effective way to solve energetic and ecological problems //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #3 1997;
2. V.D. Dudyshev How is it possible to save the civilization and the nature from global ecology pollution? //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #11 1997;
3. V.D. Dudyshev Method of combustion control. Patent of Russian Federation #2071219 of 19.04.1994;
4. V.D. Dudyshev Method of intensification and regulation of combustion. Patent of Russian Federation #2125682 of 06.06.1995;
5. V.D. Dudyshev Method of electric combustion of refuse and wastes. Patent of Russian Federation #2117870 of 06.06.1995;
6. V.D. Dudyshev Method of electric combustion cleaning of waste gases. Patent of Russian Federation #2125168 of 15.11.1995;
7. V.D. Dudyshev Method of combustion intensification in a fire-chamber of a boiler plant. Patent of Russian Federation #2079766 of 14.06.1995;
8. V.D. Dudyshev Method of flame extinguishing, Copyright of USSR #1621234 of 12.03.1988;
9. V.D. Dudyshev Method of decrease of toxicity of exhaust gases of motor transport. Patent of Russian Federation #2117179 of 20.02.1996;
10. V.D. Dudyshev New method of environmentally appropriate matter combustion as a radical way of solving of global ecological problems. // Collected reports of II international congress of "Environment for us and future generations: ecology, business and human rights in new conditions", Russia, Samara, 1997;
11. V.D. Dudyshev New method of refuse electric combustion // Collected reports of III international congress of "Environment for us and future generations: ecology, business and human rights in new conditions", Russia, Samara, 1998;
12. V.D. Dudyshev Problems of ecological education and upbringing of the youth // Collected reports of II international practical conference of "Pedagogical activity as a cultural activity", Russia, Samara, 1999;
13. V.D. Dudyshev Environmental safety of motor transport //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #5 1997;
14. V.D. Dudyshev Perspective technical works out and inventions of ecological improvement of motor transport. //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #12 1998;
15. V.D. Dudyshev Method of ICE exhaust gases cleaning, i.e. electric combustion cleaning and electric filter at ICE output. Patent of Russian Federation #1404664;
16. V.D. Dudyshev Problems and ways of ecological improving of native motor transport. //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #11 1998;
17. V.D. Dudyshev Method of internal cleaning of exhaust gases of ICE. Patent of Russian Federation #2165031;
18. V.D. Dudyshev Method of intensification of ICE operation. Patent of Russian Federation #2135814;
19. V.D. Dudyshev Method of decreasing of toxicity of ICE exhaust gases and a device of its application. Patent of Russian Federation #2135814;
20. V.D. Dudyshev Method of intensification and regulation of combustion. Patent of Russian Federation #2125682;
21. V.D. Dudyshev Method of heat machine control. Patent of Russian Federation #2134354;
22. V.D. Dudyshev Environmentally appropriate ICE. // "New technologies" ("Novye tekhnologii") magazine #2 2001 Samara;
23. V.D. Dudyshev New electric combustion technologies. // "New technologies" ("Novye tekhnologii") magazine #3 2001 Samara;
24. V.D. Dudyshev New electric combustion technologies. // "New technologies" ("Novye tekhnologii") magazine #4 2001 Samara;
25. V.D. Dudyshev New electric technology of fire non-contact extinguishing and ignition prevention. // "New technologies" ("Novye tekhnologii") magazine #9 2002 Samara;

Infinite Energy Magazine

* Cold Fusion * New Energy
* New Science * New Technology

Subscriptions, 6 Issues Per Year
$29.95 North America
$49.95 Foreign
Single Sample Copy
$5.95 North America, $10.00 Foreign

Infinite Energy Magazine
P.O. Box 2816-FV
Concord, NH 03302-2816
Phone: 603-228-4516 Fax: 603-224-5975
http://www.infinite-energy.com

EAGLE-RESEARCH
Energy Solutions
since 1984

A Research Organization that Develops & Distributes Practical Energy-Saving Methods & Devices

4 Energy Way, P.O. Box 118
Porthill, Idaho 83853
FAX: 250/492-7480

Technical questions to:
wiseman@eagle-research.com

New Effect of “Cold” Evaporation and Dissociation of Liquids Based on Capillar Electroosmosis

Valery D. Dudyshev
Doctor of Technical Science, Professor,
Corresponding Member of Russian Ecological Academy

ecolog@samaramail.ru
http://www.intrasoft.ru/nizot

There is discovered new electric physical effect of intensive “cold” evaporation and dissociation of liquids and aqueous solutions into fuel gases **without any energy consumption** due to high-voltage capillar electroosmosis [1].

Prospects and problems of hydrogen engineering

Effective obtaining of hydrogen from water is a long-standing tempting dream of the civilization. That is connected with the fact that there is a huge amount of water on the planet, and hydrogen engineering promises the possibility to obtain unlimited quantity of “free” energy obtained from water. All the more, the very process of hydrogen combustion in oxygen medium, which is obtained from water as well, ensures ideal high-calorie and environmentally appropriate combustion.

Ecological and energetic problems are very actual nevertheless they have not been effectively solved. All known methods and devices of production of hydrogen and other fuel gases are ineffective since there is no a real high-performance technology of evaporation and splitting of liquid molecules. The main cause of ineffectiveness of the analogous consists in their difficulty and in energy consumption for breaking of intermolecular connections at dissociation of water liquid fractions.

Physical-chemical structure of even habitual tap water is quite complicated since there are numerous intermolecular connections, chains and other molecular structures in water. In particular, in habitual tap water there are different chains of oriented water molecules, which are peculiarly connected with admixture ions (cluster formations), its various colloidal compounds and isotopes, mineral matters as well as various dissolved gases and admixtures.

It is a paradox but in the living nature there is a long-standing effective way of electric capillar delivery and “cold” evaporation of liquid, which allows transform it into gaseous state without heat energy and electric energy supply line. This natural effect is realized in practice by plants, which deliver aqueous solution and make its “cold” evaporation by capillar electroosmosis. It is quite comprehensible that this natural energetically perfect technology is applicable in methods of liquids transformation into fuel gases. The author of this article has designed such experimental devices of cold electric capillar evaporation of liquids according to electric pumps of trees (Fig.1-3).

The simplest operating device, which experimentally realizes in practice the effect of high-voltage capillar electroosmosis of the “cold” evaporation and dissociation of water molecules, is demonstrated in Fig.1.

New Electroosmotic Capillar Method of Obtaining of Fuel Gas from Water

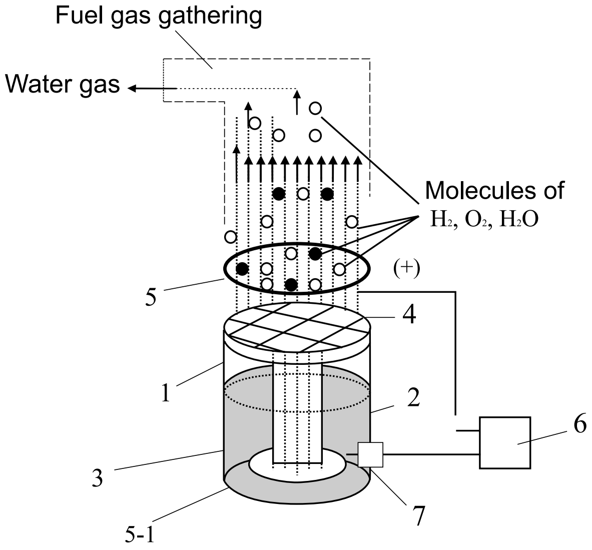

Fig.1
Simplest device of capillar electroosmosis of liquids

First experiments on the electric capillar dissociation of liquids were made with using of habitual water as well as its solutions and water-fuel emulsions of various concentrations as liquids. In all these cases fuel gases were successfully obtained in spite of the fact that these gases greatly varied due to their composition and heat capacity. The experiments on the electroosmotic evaporation and dissociation of liquids are realized in practice by the following way. First a wick (3) and a porous evaporator (4) are moisten with a water-fuel blend (emulsion) (2) then the blend (2) is poured into a reservoir (1). Then a high-voltage source of tension (6) is switched on and high-voltage difference of potentials (about 20 kV) is supplied to the liquid at some distance from the capillaries (i.e. from the wick (3) and the evaporator (4)). The source of electric field is joined to the device by electrodes (5-1) and (5). A plate perforated electrode (5) is placed above the evaporator (4) surface at a distance, which is enough to prevent an electric breakdown between the electrodes (5) and (5-1). Electrostatic forces of longitudinal electric field acts on the liquid. As a result dipolar polarized molecules of the liquid move along capillaries of the wick (3) and evaporator (4) from the reservoir to an opposite electric potential of the electrode (5) (i.e. electroosmosis occurred). Liquid molecules are detached from the evaporator (4) surface by these forces and transformed into a visual fog, i.e. the liquid is transformed into another aggregative state at minimal energy consumption of the electric field source (6). After that they provide electroosmotic elevation of this liquid. In the process of detaching and collision of evaporated liquid molecules with molecules of air and ozone as well as with electrons there is occurring partial dissociation between the evaporator (4) and the upper electrode (5) in an ionization zone. At the process a fuel gas is produced, which can come thorough a gas collector (7), for example, into combustion chambers of motor transport engine.

It has been experimentally shown that change of intensiveness of process of evaporation and dissociation of vapor molecules depends on change of distance from the electrode (5) to the evaporator (4). Moreover, this dependence is conditioned by the following factors, viz by changing of the evaporator area, kind of the liquid, quality of capillar material of the wick (3) and the evaporator (4), parameters of the electric field as well as the source of intensity (of power) (6).

The author's first experiments on this simplest device, which were organized in 1986, demonstrated that "cold" water fog (i.e. gas) appears in capillaries from liquid (i.e. water) at high-voltage electroosmosis without any observable energy consumption but just using potential energy of the electric field. This conclusion is evident since in the process of experimenting consumed electric current of the field source was the same and equal to the consumed current of the source idling. This current was not changed depending on the appearance of the liquid evaporation.

The experiments demonstrated that capillar electroosmosis evaporated quite significant quantity of water (1 liter) without any energy consumption for 10 minutes at a 10-centimeter-diameter capillar cylinder. That is to say that the consumed electric power (10 Watts) of the electric current source, i.e. of the tension high-voltage transformer (20 kV), was unchangeable and did not depend on mode of its operation. It has been experimentally stated that the whole consumed energy supplied by the current network is insignificantly small comparably with the energy of liquid evaporation. It can be explained by the fact that power was consumed only to generate the electric field and did not increase at liquid capillar evaporation that occurred due to work of an ionic pump and of a polarized pump. Hence the effect of cold electric-capillar evaporation of liquid is very economical in respect to the process energy consumption.

In spite of the fact that energetic essence of this process has not been disclosed it is evident that both "cold evaporation" and water dissociation are realized in practice by potential energy of the electric field. More precisely, the visual process of evaporation and water splitting into H_2 and O_2 at the capillar electroosmosis is realized by powerful Coulomb forces of this strong electric field.

Editor: It is not the most surprising fact. The most astonishment is aroused by school-day stereotypes that work of a field at a closed cycle is equal to zero. These stereotypes have been kept in mind of people for a long period of time. Everybody understands that a field can do work but if a body falling from some height is accelerated in a potential field and its kinetic energy increases then it requires energy consumption to relevate the body up to this height. Nevertheless, the analogy of a gravitation field with an electric one is not one-valued since the electric field may be generated only at a part of trajectory of the accelerated body motion. An electric field can be pulsating, it can be screened or it is possible to change its direction at the reverse part of the trajectory in such a way that the field constantly accelerates the body. Hence a principal conclusion can be made: summary work of a potential field may not be equal to zero. This conclusion has earlier been proposed by A.V. Frolov in his article published in the USA (Newsletter of the Institute for New Energy. May 1994. p. 1-4).

In principle this uncommon electroosmotic pump-evaporator-splitter is an example of the perpetual motion machine of the second type. Thus the high-voltage capillar electroosmosis of aqueous liquid provides really intensive and energetically free evaporation and splitting of water molecules into the fuel gas (H_2, O_2, H_2O) by means of using of potential energy of the electric field.

To produce more complete dissociation of water molecules into the fuel gas it is necessary to make the whole water molecules collide each other and be split into molecules of H_2 and O_2 in an additional transversal alternating field (Fig.2).

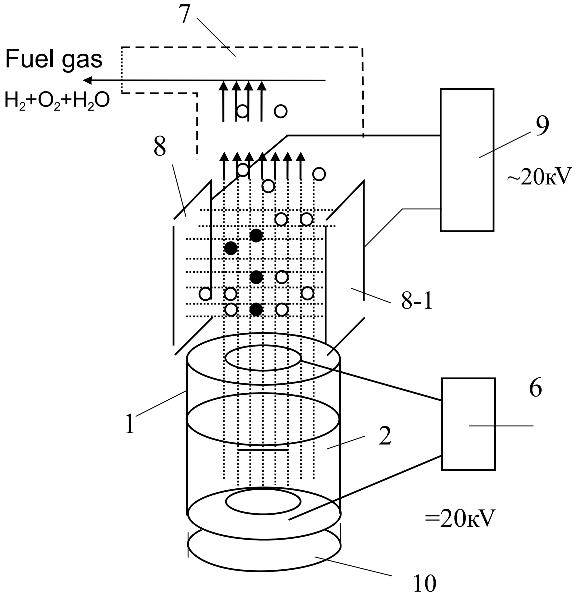

Fig.2

Device for production of fuel gas by electroosmosis. The device is equipped with an additional high - voltage splitting center of liquid molecules by electric alternating field.

At the second stage of water dissociation the energy of the second electric field is used, more precisely, powerful electrostatic forces are used to intensify oscillation resonant process of "collision-repulsion" of electrified water molecules represented as water gas. The result of this process is complete breaking of liquid molecules and generation of fuel gas molecules.

Conditions of optimal dissociation vary due to a kind of the liquid, to capillaries properties, and to the field parameters. These conditions are caused by required productivity of the process of dissociation of concrete liquid. Fig.2 demonstrates in details functional structure and composition of the device equipped with two sources of the electric field.

In the case of preliminary division of initially chemically neutral water into chemically active fractions (i.e. acid fraction and alkaline fraction) realization of the technology of production of fuel gas from water becomes possible at temperature below zero (up to –30°C). In winter it is quite important and useful for motor transport. This "fractional" electrically activated water does not freeze at degree of frost; hence the device designed for hydrogen production from such activated water can operate at environment temperature below zero and at degree of frost.

This principle of additional chemical activation of water (or liquid) is realized in practice in the device

(Fig.3). In contrast to the devices mentioned above (see Fig.1, 2) this device is supplied with an electrochemical activator of liquid (3) with evaporators (4).

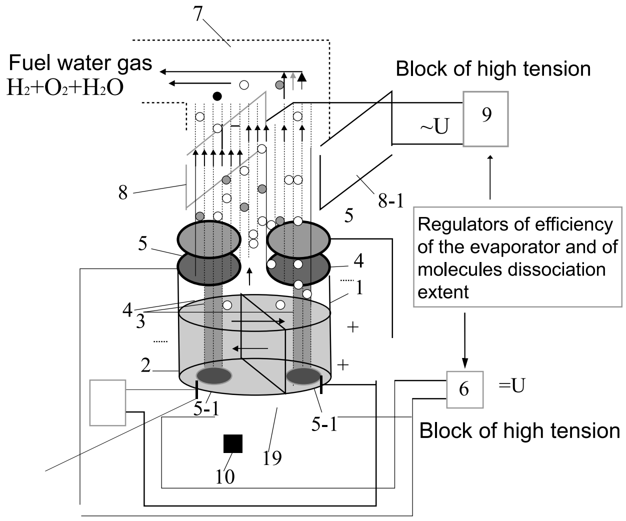

Fig.3

Hybrid device equipped with two sources of electric field and liquid electric activator

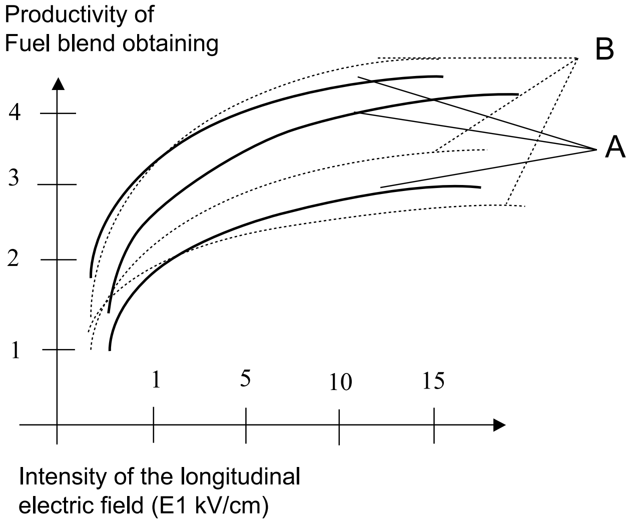

Fig.4

Productivity of the method increases as well as oscillation frequency of the second transversal electric field (A) and evaporating area (B) increase.

There is a known and low-expenditure progressive electric technology by Stanly Mayer, which has been worked out to obtain fuel gas from water (Mayer cells). Our technology is more progressive and ensures higher productivity than Stanly Mayer's technology does [3]. It can be explained by the fact that this electroosmotic effect of evaporation and liquid dissociation combined with a mechanism of the electrostatic pump and the ionic pump ensures intensive evaporation and dissociation of liquid as well as effective detachment of gas molecules from the dissociation zone. The process of gas molecules detachment is accompanied by acceleration from the upper end of the capillaries. Therefore in our case the effect of screening of the working zone of molecules electric dissociation is not produced. All these processes proceed at minimal energy consumption, which is similar to analogous energy consumption. Moreover, the process of fuel gas production does not slow down as the process in Mayer's technology does therefore gas productivity of our method is significantly higher than that of this progressive analogue at similar minimal energy consumption.

Some technical and economical aspects of the new technology realization

In the near future production run of these high-effective electroosmotic generators of fuel gas from practically any liquids (including tap water), which are based on the proposed new technology, can be established. At the first familiarization level it is especially easy and economically appropriate to realize in practice a variant of the device of transformation of water-fuel emulsion into fuel gas. A prime cost of the production-run device for generation of fuel gas from water of 1000 m^3/hour productivity comes to approximately 1 thousand of US dollars. Consumed power of such an electric generator should come to no more than 50-100 Watts. Therefore such compact and effective fuel electrolyzers can operate practically in any motor car. As a result heat engines can work on any hydrocarbon liquid or even on habitual water. Mass application of these devices for the motor transport can cause immediate energetic and ecological perfection of the motor transport as well as designing of an environmentally appropriate and economical heat engine. Approximate financial expenditures for working out and designing of the device for fuel gas obtaining from water, and for bringing the investigation of the first testing device of 100 m^3/sec productivity to an experimental-industrial model come to about 450-500 thousands of US dollars. That contains the expenditures for projecting and investigation, for designing of the very experimental device and of a testing bed, which is necessary for approbation and engineering development of the device. The author is interested in business and creative cooperation with those companies, which can provide this project with investments to bring the device to the experimental-industrial model and introduce the perspective technology into practice.

Conclusion

Electroosmotic "cold" evaporation and dissociation of water and aqueous solutions through capillaries is a perspective way of highly productive fuel gas production at minimum of energy consumption.

References

1. V.D. Dudyshev New electric combustion technology is an effective way of solving of energetic and ecological problems // "Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine #3 1997;
2. "New Energy News" magazine, May 1994. P. 1 – 5;
3. Stanly Mayer Method of production of fuel gas. Patent of the USA #4.936,961;

Conceptual Hurdles to New Millennium Physics

Donald Reed

E-mail: torsionpower@yahoo.com

Abstract

Time, space, energy and mass form the four-fold conceptual basis for gauging physical reality. The following paper, as a follow-up to a previous exposition [1] which focused on the necessity to change the current paradigm for time, continues in this same vein, but will also consider in greater depth the interrelations of this phenomenon with the other three yardsticks cited above. Again, information garnered from a wide variety of sources will be considered. We hope to provide, through just such a unique eclectic format, the beginnings of a possible fresh understanding of the workings of nature and perhaps ultimately furnish a conceptual basis for extending the structure of current physical theory to compatibly encompass the elements of a unified framework of physics and metaphysics.

Introduction

In the former article cited above, by investigating the following varied sources and research [2-17], we came to the inescapable conclusion that the phenomena of time and space are considerably more intimately related than is currently suspected by modern science, and will require a drastic re-working to fit the conclusions of the various evidence cited. Towards this end, the adoption of a more expansive paradigm for these yardsticks was suggested, which incorporates a fluid-field nature for time and space where both are derivatives of the fundamental ground-form of energy in flux or oscillation. These new approaches involve the inclusion of a Kozyrev-type of "substantial"(active) time-flow as opposed to the conventional "relational"(passive) concept of time, where it is used as a static parameter signifying duration. The substantial aspect of time presupposes that it is an essence which can and does affect physical processes, and that those same physical systems can cause a reverse action on time [8]. Such active fluctuating (deformable) types of time or space, implies the establishment of a revolutionary notion: positing completely *non-scalable* metrics for both time and space. This is in direct contrast to all contemporary orthodox models for physical reality, either in Einsteinean relativity (Special or General theory), quantum field theory, or even superstring/supersymmetric theory, all of which continue to consider both of these yardsticks from their relational (length or durational) standpoints. Consequently, all these formalisms require scalable metrics of some sort for their proper description (for refreshing viewpoints on this matter, see [7]). For instance, in the standard equations of quantum theory, time is regarded merely as an unchangeable static parameter. When we later consider applying these novel substantial aspects of time to the edifice of quantum theory, we will see that time must then become a "hidden variable". It will be seen that by considering time in this manner, a clearer and more tractable explanation of the inevitable probabilistic aspects of quantum theory, evidenced in the signature Uncertainty Principle, complementarity and non-locality will arise, without having to invoke the counter-productive and ill-conceived Copenhagen Interpretation.

Oscillatory Models for Time and Space

Upon accepting non-scalable metrics for time and space, we observe that a similar field nature for **energy** and **mass** must in turn also be postulated. Moreover, from the Killick description of how (sub-atomic) tachion-pairs operate [6], shuttling their energy back-and-forth in a free-wheeling but purposeful manner and creating oscillatory features of time and space by their action, we should consider the possibility of another unprecedented concept: a "value-motivated" energy might be at the foundation of the structural integrity of physical matter. This is conceivable when considering Killick's description of the tachion-pair dynamic/evolutionary cycle as necessarily including the 3-step process of observation – reflection – action (**trinitivity** of motion). Along the same lines, from the Smith book **The New Science** [5], we learn that a so-called "tempic field" energy exists, which could be described as the parent structure out of which our standard conception of "clock-time" (entropy changes) results. The tempic field is essentially a scalar (but not static) function which has vectorial nature only in terms of its distribution or *gradient* in space. Because of the nature of the tempic field, its derivatives – the electric and magnetic fields can operate on each other in a specific geometric/topological manner, to produce a local change in the *time-frame* of matter, to use a term coined by the entity Kryon [4]. This theoretical description of the alteration of inertial/gravitational mass and time-frame has possibly seen recent actual demonstration in the dramatic experiments of John Hutchison [10] and Rudolf Zinsser's "kinetobaric" effect [11], as well as unpublished research of both Wilbert Smith and Ken Killick with electric caduceus-wound coils [5,6].

New Models for Relativistic-Fluidic Vacuum-Structure and Possibility of its Manipulation

All of this evidence allows us to conclude that measurable changes in relativistic parameters of time, space, mass and energy might not only be a feature accompanying rapid uniform movement of physical objects (Special Relativity), or representative of large gravitating astrophysical objects (General Relativity). Indeed, by relying on current limited paradigms, contemporary physics may have missed ascertaining the possibility of being able to alter these same relativistic parameters by the artificial technological manipulation of the tempic (vacuum) fields of sub-atomic particles in stationary matter, by use of specific electromagnetic fields. Tom Bearden, for one, has articulated on these various "futuristic"-type technologies in his many writings over the years [15]. We have recently seen that some of the novel theoretical conceptions of the more visionary physicists such as David Hestenes on the zitterbewegung (vacuum "jitter") phenomenon exhibited by the electron [12-14], has finally caught up with some of these advanced ideas by presenting new mathematical demonstrations (using geometric Clifford-algebraic) manipulations of the Dirac equation. In summary, the Hestenes' study concludes that the Dirac wave function and its properties, including the Dirac equation and relations to physical observables such as energy-momentum, spin and position probability current, all possess heretofore obscured important geometric relations. These results imply that probabilistic features of the quantum theory of the electron/positron arise principally from the electromagnetic interaction of the accompanying zitterbewegung-spin field of these particles with the ambient dynamic vacuum. This is, of course, in direct contrast to the conventional received view that ascribes wave-particle duality as a property of matter that is completely independent of the nature of its interactions. This revolutionary geometric interpretation of electron dynamics incorporates in its model an electron spin which arises from a helical or spiral world-line in space-time. The essential unprecedented feature of the Hestenes' zitterbewegung idea is the association of the spin with the local circulatory-helical motion characterized by the *phase* of the electron wave function. Thus, we reach the conclusion that the complex phase factor of the electron wave function can be directly associated with an objective helical motion of the electron which is, in turn, a derivative of the zitterbewegung. One intriguing feature of this structure is a frequency of oscillation that is inversely proportional to the scalar radius of curvature of the particles' helical world-line. Moreover, the Clifford-algebraic analysis reveals that this oscillation frequency is identical to electron/positron *mass*, revealing a possible key variable particle mass-energy (*frequency measure*), which is in inverse relationship to particle size. It is clear that this new model has important classical implications not yet considered by established physics which nevertheless directly correspond to some of the key features of sub-atomic behavior enumerated above and previously. Here, we refer to the tachion-pair dynamics [1,6], and also to the microscopic dynamics of elementary particles revealed by the selected esoteric sources previously considered [1-5].

Another researcher who echoes many of these conclusions is C. Sano [21]. Sano built on the work of A.P. Smirnov [22] who postulated that Newton's Third Law of action/reaction actually modeled "screw" structures, implying that elementary particles (specifically electrons/positrons) possess chiral-spiral field configurations. Accordingly, Sano posited that all actions/reactions are transmitted between actors and reactors by parallel or perpendicular clutching of the rotating chains of electromagnetic spirals of the hidden electrons and positron-pairs of the vacuum. Also, similar to the Hestenes' development cited above, Sano postulated that the radius of the outer electromagnetic spiral surrounding the electron/positron pair can change, oscillating in size around either particle, and thus producing the particles' electric or magnetic character. Supporting his theory, Sano cites the key research of I.M. Shakhporanov [23] whose experiments claimed that magnetic monopoles were generated, by employing an electrical circuit based upon Moebius band topology. Some of the unusual phenomena demonstrated by this apparatus, tending to support the magnetic monopole hypothesis, was transformation of diamagnetic substances into paramagnetic, the ferromagnetization of normally non-magnetizable materials (graphite, etc.), acceleration of chemical reactions, acceleration or deceleration of the decay of radioactive materials, etc. Sano also claimed that extraction of energy from the vacuum was possible using rotational action/reaction using magnetic monopoles.

Referencing this research to the current exposition, we recall that the key element of Killick's hypothetical tachion-pair operation was the similar non-linear/non-orientable Moebius-type dynamic cycle which results in the corresponding deformation (compression or rarefaction) of time/space/energy parameters [1,6]. In this regard, the remarkable similarity of this model to the research of Sano and Shakhporanov may be much more than mere coincidence. In fact, later we shall examine new research outlining an apparatus (yet untested), which claims to overcome the Coulomb barrier in low energy nuclear reactions (LENR), and achieves nuclear fusion by producing local time-dilation of soft photons in proximity of the deuteron reactants.

Additional Key Russian Research

To motivate the reader's sensibilities towards consideration of some of these unorthodox ideas, we defer to the previous article [1] in which many of these notions and mentioned research are explored in greater depth, as well as consider the following corroborative research.

One intriguing example of such experiments is the work of V. Chernobrov, from which he has claimed to demonstrate both acceleration and deceleration of local time-rate, within a small spherical enclosed volume conducted since 1988 [18]. The experimental system to produce these effects, was a set of electromagnets, connected in series and parallel and installed inside the globe-shaped surface in several layers. In various arrangements including up to 3 to 5 of such layers, these electromagnetic working surfaces (EWS) of various diameters were installed inside each other (similar to the Russian toy doll "matrioshka") with the maximum EWS diameter was about 1 meter, and the minimum (internal) diameter of 115 mm. With this configuration, Chernobrov claimed to measure small but detectable deceleration (-30 sec/hr.) and acceleration (+30 sec/hr.) of time within the sphere to time monitored outside its field of influence. One interesting difference was noted between the slowing down vs. the speeding up of time. The deceleration occurred considerably more smoothly and steadily whereas with acceleration, sharp discontinuous jumps were observed. These instabilities accompanying time-acceleration was observed in connection with cycles of the moon, diurnal fluctuations, and also operator presence. Chernobrov also noticed another phenomenon that also was reported in connection with the alleged legendary Philadelphia experiment, where matter in different time-frames apparently exhibited strange relative optical phenomena [see 1,4]. Specifically, the human eye in the time-frame exterior to the experiment perceives matter within the time-frame created by the apparatus as either transparent or surrounded by a vague white mist. We note with interest that transparent "shimmering" effects of substances in the target area were also occasionally a feature of Hutchison effect [10].

Academician A. Chernetsky produced what he termed a self-generating discharge (SGD) in a plasma that exhibited longitudinal energy density waves from a structured vacuum. Some unusual effects noted were over-unity energy generation (C.O.P. > 1), as well as change of the electrical conductivity of matter (reduction of the resistance of resistor) placed between the capacitor plates of the SGD. A local structured vacuum was definitely produced since the resistance value remained unchanged even when the generator was turned off. Here reference must be made to the Kozyrev experiments [1,8], in which a type of similar *memory* process was activated ostensibly within the vacuum. Specifically, it turned out that in experiments with a vibrating torsion balance (or pendulum), at points of support the emerging additional forces did not disappear when the irreversible process (vibration) was stopped, but remained in the system for an appreciable time. The SGD plasma device may have also produced a detectable change in local time-flow rate as evidenced by decrease in frequency of a quartz oscillator placed in the discharge of the generator. The Chernetsky generator thus possibly caused anomalies in time-rate at a local space position. There were also basically anecdotal phenomena where psychic ability of personnel in proximity to the SGD was apparently enhanced [19]. The latter effects are not inconceivable once the intimate connection of consciousness to space and time is ascertained, from consideration of not only hypothetical "value-motivated" tachion-pair dynamics, but from the more prosaic studies as we shall see next.

Pathological Studies in Time-Perception

We can motivate further understanding for the claimed intimate connection of time to space by considering the results of a recent study of patients possessing a disjointed sense of time by Metod Saniga of the Slovak Academy of Sciences [20]. Saniga discovered the brain is hard-wired to perceive space and time as interconnected. Specifically, time pathology is apparently always accompanied by space pathology, in a sense that space either loses dimensions or acquires other dimensions. To quote Saniga "When time seems to stop, people often feel as if space becomes 2-dimensional. On the other hand, when the subject feels they perceive the past, present and future all at once, they simultaneously have the impression that space has infinite dimensions". This phenomenon is apparently not culturally endemic since Saniga illustrates that both pathologies cross cultural lines, evidencing similar studies cited from Italian, German and English psychological journals. In his report, Saniga combined mathematical models ("pencil-conics") and pathology reports of schizophrenic, drug-induced and other abnormal perceptions of time. His current work also encompasses studies of near-death experiences. He found that most of those who claim to have allegedly crossed over to the "other side" and back, tell similar tales. For that brief moment of near-death, the universal subjective experience of all individuals in this state of consciousness is that time loses its meaning. Although such evidence can at best be considered as anecdotal, since subjective conscious experience transcends the possibility of scientific proof, nevertheless perhaps these results from pathological and near-death studies also give us a hint towards new conceptions of time and space that necessitate both phenomena to be inextricably linked.

Investigation of the Possible Internal Properties of Time

One inescapable conclusion that results from all the above varied studies, either scientifically or psychologically based, and which cries out for future experimental verification, is the possible existence of an internal structure for time. Such a structure also implies the necessity for constructing a new edifice for physical reality that incorporates non-scalable metrics for the four fundamental yardsticks. We shall first investigate supporting evidence for internal time structure and subsequently the various implications of space, time, energy or mass that is non-scalable.

First, accepting an internal time structure, we are led to conclude that in quantum theory, time must then attain "hidden variable" status. One recent study that brilliantly articulates this notion, is the thought-provoking paper by X. Chen [24]. By positing 3 time variables as quantum hidden variables, Chen derives the Dirac equation classically. Moreover, he demonstrates that the non-intuitive property of "spin" of an electron or positron , arises naturally as a topological property of 3-dimensional time + 3-dimensional space. By extrapolating from this model, Chen then concludes that the inherent probabilistic aspects of quantum theory itself, as evidenced in wave-particle duality, Uncertainty Principle, quantum non-locality ("entangled" particles), etc., can be interpreted as the behavior of a single particle in 3 + 3 time-space. Chen postulates that the three dimensions of time geometrically form a "time sphere", with a generic point on the sphere possessing the following 3 coordinates: time radius, and two "time angles". The mathematical formalism arising from the process of stereographic projection from the north pole of this sphere, having radius of one-half, to any point on the spherical surface, incorporates both the positive-energy (north hemisphere) and negative-energy (south hemisphere) solutions to the Dirac equation (see Fig. 1 for clarification; note Z is a four-component spinor wave function). Also, by considering the evolution of a single particle, from the standpoint of this model of internal time-space, we can derive a picture of different paths on the time sphere of different weights, and on each path causality is satisfied. Through this process, Chen derives a purely classical explanation of the originally quantum-mechanically canonized Feynman Path Integral concept. This result emerges since each path from the time-sphere center to the surface corresponds to each Feynman path, and the surface of the sphere corresponds to the "surface" of the wave function. Chen also classically explains the processes inherent to Bose-Einstein-condensation (BEC) and superconductivity with this model. In such phenomena, two particles with the same spatial coordinates, cannot possess the same time angle, and thus will not have any interaction with each other and occupy the same quantum state as evidenced in BEC and superconductivity. Finally, Chen's solutions of the Dirac equation for a free particle correspond to Hopf bundles in monopole theory, and each Hopf fiber corresponds to each plane wave with different momentum states. This picture corresponds to wave packet diffusion in quantum theory. When a particle is in a fixed momentum state, each space point can contain only one Hopf bundle (one time angle), and various different time angles will be distributed in the whole space but with the same Hopf bundle. Consequently, that particle can be found everywhere. This picture appears to describe the Uncertainty Principle classically using 3 + 3 dimensional time-space. Along similar lines R. Kiehn [25] has also recently underscored the important but overlooked classical connection between spinors, minimal surfaces and the Hopf map.

Recently, Chen's work has been further developed by H. Kitada [26] who, by introducing both 3-dimensional time and energy operators, sheds new classical light on the uncertainty relation that holds between these two parameters as well. Moreover, Chen is not the only reseracher to derive the Dirac equation classically. R.A. Close has published a paper which presents a classical mathematical description of circularly-polarized waves in 3-dimensions [27]. It turns out that these chiral space waves are satisfied by a four-component wave function which satisfies a Dirac equation. Furthermore, much like the Hestenes' zitterbewegung interpretation of quantum theory, the term normally associated with electron mass in the Close equation, introduces a rotation or oscillation frequency of the propagation direction.

Classical models for nuclear processes have also recently been proposed which incidentally also imply internal time structure. Notable among these, C. Cagle claims to have developed a fusion energy device employing LENR based upon such a classical theory [28]. Normally, Deuterium nuclei are diffuse in momentum space (having high relative velocities). The process Cagle outlines produces a compactification of their momentum space (low relative velocities), so that the deuterons develop a common de Broglie wavelength that is greater than or equal to the inter-particle distance. This is claimed to be accomplished by passing soft x-ray photons near the region of two deuterons in a lattice structure of Lithium Deuteride. **The soft x-ray, when absorbed , produces a region of total *time-dilation*, causing two effects: first, ionization of atoms by strongly repelling any associated electrons in the area, and secondly and most importantly, the physical extent of the time-dilation causes two adjacent normally repelling Deuterium nuclei to overlap in a common momentum space; that is, their relative velocities achieve very low values.** Consequently, due to the time dilation, for a very short time their common de Broglie wavelength exceeds their inter-particle distance. Thus, they become strongly attractive and undergo nuclear fusion.

Cagle also claims to clarify misconceptions about the fusion process in a thermonuclear bomb. He maintains that it is not due to extreme kinetic energies (high temperatures) of particles as supposed in current paradigms, but due to nuclear processes caused by absorption of a time-dilated soft photon flux, produced by Compton scattering of hard x-rays from a fission igniter, passing through a foil of depleted Uranium. Again, just like the controlled nuclear fusion process, this causes overlapping of deuterons in momentum space (low kinetic energies) for a few pico seconds, and subsequent fusion with normal exothermal processes ensuing.

Furthermore, as a fundamental component of his LENR fusion device, which ostensibly produces time dilation and controls energy production from the fusion process, Cagle incorporates a topological field structure termed an electromagnetotoroid. The toroidal field current oscillates between toroidal and poloidal modes in a dynamic cycle which is amazingly virtually identical to Killick's description of the oscillation dynamics of a single unit toroidal tachion [1,6](also see Cagle's website for animated simulation).

Regarding dynamics of astrophysical structures, Cagle further claims that the stellar jet core star of HH30 is also a gigantic electromagnetotoroid. During the poloidal current mode, vast quantities of matter are produced and ejected along the poloidal axis when the mode changes from poloidal to toroidal. This leads one to consider the possibility that the source of superluminal gamma ray bursters, which have heretofore mystified astrophysicists and defied explanation in terms of current paradigms, might be the result of large- scale abrupt changes in the time-frame of celestial vacuum regions, subsequently causing a boost in light velocity.

Implications of Non-scalable Metrics

When postulating fluid-field natures for the four yardsticks of physical reality: time, space, energy and mass, we must assume the existence of a *non-scalable* metric for our vacuum structure. Such a non-scalable vacuum, absent of a defined reference frame, must of necessity possess a *dual* nature. That is, the primary feature of a dual vacuum structure implies that infinitely small quantities must be treated on an equal footing with those that are infinitely large. In other words, information encoded into infinitesimal elements, is also instantly present in the unbounded infinite extents as well. This is tantamount to the existence of a holographic information encoding and transmitting vacuum field, which would structure events in space and time as a non-Markovian chain. In a non-Markovian chain of events, the prediction about the next link in the chain, requires a knowledge of **all** links, not just the one preceding it.

Metod Saniga's research into mystical perceptions of superconscious reality appears to support such a primordial non-Markovian holographic mapping of time and space. To quote from a transcript of such a vision: "I wake up in a whole different world…a different space…This space was distinct from the one we all know. It had different dimensions, everything contained everything else. One was situated in a state of being in which the 'will be' (future) and the 'vanishing' (past) were already included, and this being was my consciousness. It contained it all. The 'being contained' was present very vividly in a geometric way in the form of circles of different sizes which again were all part of a unity since all of the circles formed exactly one circle. The biggest circle was a part of the smallest and vice versa."[20].

Such a structure of time and space can best be topologically encoded in a non-orientable structure such as the Moebius band or Klein bottle. Similar to the data presented in the above mystical revelations and in the previously treated Killick tachion-pair dynamics [1,5], in such configurations "inside" and "outside" lose their meanings and meld into one another; ordinary dichotomic relations distinguishing thesis from antithesis are sublated and supplanted by a higher unity defining the evolutionary process of the system. In the mathematical edifice abstracted from such a vacuum structure, duality rather than exclusivity of field structures holds sway. This in turn, presupposes a static-dynamic triality of fundamental field structures, characterized by the three magnitudes: zero, infinity and the mediating factor of unity, the latter being an indication that each of the other two field extremes are in perfect balance. Accordingly, we will assume that the stability (coherency) of non-scalable fields results whenever more than half their reality is in common.

Further development pertaining to a specific algebraic-geometric structure that would inherently incorporate this field model, will be carried out in a future paper. Nevertheless, the interested reader can skirmish on this frontier by consulting the relevant recent references to a Clifford-algebraic structure which encodes the projective duality of "space" and "counterspace" into a model for mechanics on the quantum level [29]. This promising model accounts classically for the phenomenon of quantum non-locality. It also heralds a new version of quantum field theory which, by treating electron-positron pairs as topologically non-orientable, is free of the detrimental divergences in self energy and charge, thus obviating the necessity for the ad hoc prescription of renormalization so endemic to current theory [30].

However, due to the limited scope of the current treatment, here we shall only make general comments on the significance of the above-mentioned concept of field triality. Although this model may appear foreign to current physical theory, we underscore that exactly such a system was delineated by W. Smith in the book, The New Science [5]. Observing, as we have [1], that this book in some parts of its exposition can cause confusion, we carefully choose the following selected passages that are most instructive in clarifying the notion of non-scalable metrics and in pointing the way to new paradigms (my comments in parentheses). General comment on non-scalable metrics: "Unity is half way between zero and infinity and always remains the fulcrum about which all other values regardless of scale pivot". On the electric field: "Between the two limits of zero and infinity and through unity there is divergence which we recognize as the electric field, but with the exception that there is no point charge at zero. Our awareness establishes the charge at radius unity with exactly half of it being 'inside' and the other half being 'outside'". On tempic field dynamics: "The tempic field being purely scalar in nature merely contributes to the manner in which changes can occur in the system. If, however, through some characteristic of configuration, the tempic field within a composite particle displays some irregularity, this will manifest as a 'vibration', and if the magnitude of the vibration is sufficient as to cause the interpenetrating fields to exceed the half-and-half point , the whole system will come apart: radioactivity. When a system does come apart, those field which were coherent and interpenetrating in the correct proportion will remain so and take off down the tempic field gradient and become radiated energy". On velocity of light/Planck's constant: "Within the universe we perceive, we are reasonably satisfied that the maximum value of all the coherent fields involved in our particles and radiated energy have the same value. Planck's constant is the numerical expression of this value, and the velocity of light is the expression of the numerical value of the (local) tempic field intensity". On nuclear fusion: "If a region is selected in which there are two fields of the same kind, same magnitude, same direction, such that very nearly half the reality of each is within the region, then the two fields are just on the verge of becoming coherent. If the fields are not coherent, the total energy in the region is the sum of the energies of the two fields; i.e., twice the square of the field intensity of each integrated over the region, or twice the energy of the two fields incoherent. This represents the 'packing energy' of bits and pieces of atomic nuclei, and also points the way to the precipitation of energy out of the cosmic background (vacuum engineering)". Notice how this description, written 45 years ago, of how a tempic field gradient (time-stress) operates between atomic nuclei to produce coherency (half-in half-out condition), perfectly parallels the recent Cagle view, which claims the fusion process takes place only as a function of time-dilation [28]. Finally, on gravitation: "With the exception of skew electric fields (electric field possessing tempic field gradient) all the other fields of an aggregate mass may be considered as static because they have no component of the tempic field at right angles to them. The skew fields (spiral-helical?) on the other hand, are dynamic because they do have a quadrature tempic field component. Furthermore, skew fields are largely incoherent (gravitation cannot normally be shielded) simply because the usual almost random orientation precludes their meeting the half-in requirement of form to become coherent. Since the gravitational field is due to the skew electric field, or stated differently, to the induction from the 'motional magnetic field' (longitudinally moved magnetic field), this is the logical region to explore for the mechanism by which fields may be produced to combine with the gravitational field (of Earth) to produce a resultant more to our liking (alteration of local gravitational potential)".

This last quote describes the gravitational field of a mass as a residual field phenomenon, similarly to the Sakharov and Puthoff conclusions which attribute gravitational and inertial mass to the interaction of accelerated or gravitating matter with the zero-point-vacuum-fluctuations (ZPF) causing a vacuum reaction force [31]. Moreover, similar to the Hestenes' claims, in their recent papers [32], Haisch/Rueda/Dobyns propose that, via this new interpretation of inertial mass as an acceleration-dependent electromagnetic (Lorentz) force, that a former postulate of quantum mechanics appears to be derivable classically via the interpretation of rest mass as the energy of the ZPF driven zitterbewegung; that is, the de Broglie wavelength of a moving particle, may be derived from Doppler shifts of the Compton frequency oscillations associated with zitterbewegung that occurs when a particle is placed in motion.

For further studies of physical theories with non-scalable metrics, the reader is directed to the excellent work of A.A. Nassikas [16]. Like few theories hitherto postulated, Nassikas posits a fundamental probability density function for vacuum energy, out of which oscillatory-deformable sub-atomic level physical aspects of time and space then emerge. This is the reverse to most contemporary paradigms, which view energy as a derivative of matter, time and space, even at the quantum level. Finally, A. Frolov has used Nassikas' theory to explain over-unity energy generation that has been demonstrated in some LENR electrolytic cells of the Pons-Fleischmann variety [17].

Editor: Actually Prof. Nassikas wrote in this patent that the idea which is described in the patent was proposed by Alexander Frolov. In 1996 these questions were discussed by us in Saint-Petersburg. Specifically, a Palladium cathode over-saturated with protons produces an imbalance of the vacuum engine in this area, with an accompanying change in the local time-rate and subsequent breakdown in the Coulomb barrier of the nuclei as per schemes articulated by C. Cagle [28] and T. Bearden [15].

Conclusions

In this journey through the latest research and speculations involving new interpretations of the four physical yardsticks: time, space, energy and mass, we hope the reader and technical specialists/theorists have been inspired to continue research along the lines suggested in the above dissertation. In this regard, certain open-ended questions present themselves that focus squarely on the viability of technological future developments in the new energy field. First, could the missing element in our eventual understanding of all devices featuring documented over-unity energy generation, change in mass (gravitational potential), LENR-based transmutation of elements and the amelioration of radioactivity, etc., be due the alteration of local time-flow in sub-atomic vacuum fields within the apparatus? Secondly, could the frustrating feature of the capriciousness of over-unity power and the like to yield to testable replication in many cases, be due to currently unperceived uncontrollable aspects of time and the vacuum which must be addressed in order to solve these problems? Thirdly, could incorporation of non-orientable topological structures such as the Moebius band, Klein bottle, etc. in new models of electrodynamic field structure, be instrumental in finding the answers to both of the previous questions? With sufficient probing for possible answers to such questions, the present author has the firm belief that new paradigms are certain to soon result that will crystallize these thoughts into viable scientific hypotheses subject to experimental testable verification. As an offshoot and definite bonus of this process, possibly a new more expansive understanding of the role of the vacuum in both nature and consciousness will be in the offing.

References

1. D. Reed, "A New Paradigm for Time: Evidence from Empirical and Esoteric Sources", Explore!-for the Professional, vol. 4/5 July/Sept., 2002.
2. J. Roberts, unpublished remarks as transcribed by W.H. Kautz.
3. J. Roberts, The Seth Material, Prentice-Hall, Inc., 1970, pp. 292-298.

J. Roberts, The "Unknown" Reality, Prentice-Hall, Inc., 1986.

4. L. Carroll, Alchemy of the Human Spirit (Kryon Book III), The Kryon Writings, 1995.

L. Carroll, Letters From Home (Kryon Book VI), The Kryon Writings, 1998.

L. Carroll, Passing the Marker (Kryon Book VIII), The Kryon Writings, 2000.

5. W. Smith, The New Science, Fern-Graphic Pub., 1964.
6. D. Reed & K. Killick, "Tachion Energy Theory", Energy Unlimited, 1978-1983.
7. A. Correa & P. Correa, "Consequences of the Null Result of the Michelson-Morley Experiment: The Demise of the Stationary Aether, the Rise of Special Relativity and the Heuristic Concept of the Photon", Infinite Energy, no. 38, July-Aug. 2001, pp. 47-64. A. Correa & P. Correa, "The Sagnac and Michelson-Gale Experiments", Infinite Energy, no. 39, Sept.-Oct. 2001, pp. 32-49.
8. N. Kozyrev, "On the Possibility for the Experimental Investigation of the Properties of Time", Time in Science and Philosophy, Prague, 1971, pp. 111-132.

N. Kozyrev, "An Unexplored World", Soviet Life, 1965.

A.P. Levich, On the Way to Understanding the Time Phenomenon (Part II: The Active Properties of Time According to N. A. Kozyrev), World Scientific, Singapore, 1996.

9. D. Reed, "Torsion Field Research and Implications for New Physics and Energy Technologies", Jour. New Energy, vol. 4, no. 2, Fall, 1999, pp. 151-164.
10. J. Hutchison, "The Hutchison Effect Apparatus", Proc. of the 1st Symp. On New Energy, Denver, May, 1994, p. 199.

J. Hutchison, "Inside the Hutchison Effect", Extraordinary Science, vol. III, issue 4, Oct-Dec. 1991, pp. 23-26.

11. T. Valone (ed), Mechanical Energy from Gravitational Anisotropy, Integrity Research Institute, Washington, DC, 1996.

W. Peschka, "Kinetobaric Effect as Basis For a New Propulsion Principle", Raumfahrt-Forschung, Feb, 1974 (in German); trans. By D. Reed, Infinite Energy, vol. 4, issue no. 22, 1998, pp. 52-56.

D. Reed, "Translator's Analysis and Comments on the Zinsser Effect Device, Infinite Energy, vol. 4, issue no. 22, 1998, pp. 57-59.

12. D. Hestenes, "Quantum Mechanics from Self-Interaction", Found. Phys., vol. 15, no. 1, 1985, pp. 63-87.
13. D. Hestenes, "The Zitterbewegung Interpretation of Quantum Mechanics", Found. Phys., vol. 20, no. 10, 1990, pp. 1213-1232.
14. D. Hestenes, "Zitterbewegung Modeling", Found. Phys., vol. 23, no. 3, 1992, pp. 365-387.
15. T. Bearden, Extracting and Using Electromagnetic Energy from the Active Vacuum, Association of Distinguished American Scientists, Huntsville, Alabama, 2000.

T. Bearden, "Giant Negentropy in the Common Dipole", Proc IC-2000, St. Petersburg, Russia, 2000(in Press)

16. A.A. Nassikas, " The Hypothesis of the Unified Field and the Principle of Its Dual Interpretation", Proc. Of III Int. Conf. On Problems of Space Time and Gravitation, Russian Academy of Sciences, St. Petersburg, Russia, 1996.
17. A. Frolov, "The Work Created by Means of a Potential Field", Proc of III Int. Conf. On Problems of Space Time and Gravitation, Russian Academy of Sciences, St. Petersburg, Russia, 1996.
18. V. Chernobrov, "Experiments on the Change of the Direction and Rate of Time Motion", Proc. Of III Int. Conf. On Problems of Space Time and Gravitation, Russian Academy of Sciences, St. Petersburg, Russia, 1996.
19. A.V. Chernetsky, "Processes in Plasma Systems with Electric Charge Division", Deutsche Vereinigung fur Schwerkraft-Feld Energie e. V. no. 25, Mar. 1991, no. 27, Feb. 1992.
20. M. Saniga,"Unveiling the Nature of Time-Altered States of Consciousness and Pencil-Generated Space-Times", Int. Jour. Of Transdisciplinary Studies, vol. 2, no. 2, pp. 8-17.
21. C. Sano, "Twisting and Untwisting of Spirals of Aether and Fractal Vortices Connecting Dynamic Aethers", Jour. New Energy, vol. 6, no. 2, Fall 2001, pp. 178-184.
22. A.P. Smirnov, "Hyper-Physics - Physics of Reality", Kneeling Before the Truth, pp. 62-100, Materials of the 2nd Int. Scientific Clarification Conf.-2000, St. Petersburg, Russia,(2000).
23. I.M. Shakhparonov, "Interaction of Kozyrev-Dirac Emanation Methods of Detection and Interaction with Matter", Proc. Of III Int. Conf. On Problems of Space Time and Gravitation, Russian Academy of Sciences, St. Petersburg, Russia, 1996, pp. 175-187.
24. X. Chen, A New Interpretation of Quantum Theory – Time as a Hidden Variable, Apr. 2001, xxx.lanl.gov/abs/quant-ph/9902037.
25. R.Kiehn, Spinors, Minimal Surfaces, Torsion, Helicity, Chirality, Spin, Twistors, Orientation, Continuity, fractals, Point Particles, Polarization, the Light Cone and the Hopf Map, www.uh.edu/~rkiehn.
26. H. Kitada, Three Dimensional Time and Energy Operators and an Uncertainty Relation, July 200, xxx.lanl.gov/abs/quant-ph/0007028.
27. R.A. Close, "Torsion Waves in Three Dimensions: Quantum Mechanics with a Twist", Found. Phys. Lett., vol. 15, no. 1, Feb. 2002, pp. 71-83.
28. C. Cagle, (2001), www.singtech.com/.
29. O. Conradt," Mechanics in Space and Counterspace", Jour. Math Phys., vol. 41, no. 10, 2000, pp. 6995-7028.
30. B. Fauser "Projective Relativity- Present Status and Outlook", Gen. Rel. Grav, vol. 33, 2001, pp. 875-887.
31. H.E. Puthoff, Phys. Rev. A, vol. 39,(1989), p. 2333. B. Haisch , A. Rueda & H.E. Puthoff, Phys. Rev. A, vol 49,(1994), p.678.
32. B. Haisch, A. Rueda & Y. Dobyns, "Inertial Mass and Quantum Vacuum Fields", Annalen der Physik, (Sept. 2000). B. Haisch & A. Rueda, "On the Relation Between a Zero-Point-Field-Induced Inertial Effect and the Einstein-de Broglie Formula", Phys. Lett. A, vol. 268, (Apr. 2000), pp. 224-227.

History Pages

Resonant Phenomena Occurring in Alternating Current Circuit

Alexander V. Frolov

Tel/fax: 7-(812)-380-38-44
E-mail: office@faraday.ru

There is a classical (according to manuals) notion that resonant phenomena can not influence on efficiency of a transformer or a motor as in a capacitive circuit or in an inductive circuit power produced at minimal power at the input of an oscillation circuit is reactive (let us remind that active power is measured at active resistance).

At the beginning of the last century this aspect was viewed differently. Advantages of a resonant case in alternating current circuits were used in practice. Let us appeal to a rare book by J. Claude-V. Ostwald named "Electricity and its applications by popular language" (I.N. Kushnerev Press, Moscow, 1914, p. 463).

"The phenomenon proceeds in a corresponding electric circuit as well as it occurs in hydraulic model: if self-induction and capacity parallel connected with each other are under influence of an alternating electric propulsion force then the total current coming through the system is equal to the difference of currents which pass two certain paths.

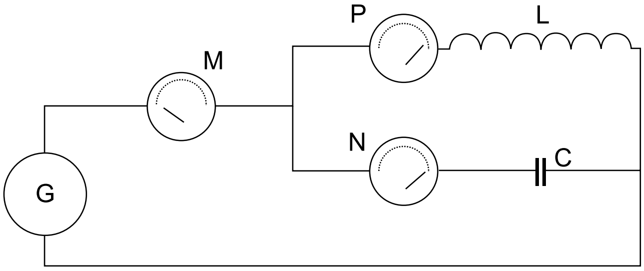

Fig.1

… join ammeters to the general circuit (M) and to either path (P and N). If P demonstrates 100 Amperes and N demonstrates 80 Amperes then the total current will be equal to 20 Amperes instead of 180 Amperes. Thus alternating current has its own "summary" so we should consider its condition. …capacity introduction compensates self-induction action in some way… let us begin to change the self-induction by moving a core into. What will happen if the current coming through the coil achieves 80 Amperes, i.e. is equal to the current observed in the path with a capacitor?

As you guess, since the total current is equal to difference between currents passing the paths then now it will be equal zero. It is an incredible case:the device produces zero current dividing into two paths and in each of these paths current is equal to 80 Amperes. It is a good example for the first acquaintance with alternating current, isn't it?"

Andrey A. Melnichenko (Moscow) is one of modern investigators who research this phenomenon. Any motor of alternating current can be considered as inductance. A circuit consists of motor coils and some capacitor connected in series with the motor winding. Thus if we take the circuit, consisting of motor coils and a capacitor which is connected in series to motor winding and adjust it in resonance, then mechanical power occurring on the motor shaft is produced at zero (minimal) power consumption of an alternating current source.

Melnichenko applied a simple method of voltage increase by meanse of resonance: he succeeded in obtaining of normal voltage for operation of standard motors of 50 Hz 220 V from a source of 50 Hz 110 V and 70 V. The circuit consumption for overcoming of its active resistance (of coils) can be considered as insignificant. Low frequency currents require a large capacitor. Nevertheless even at higher frequency, for example at 400 Hz, the system can be compact and effective. This method is worth to be applied in a scheme consisting of an alternating current motor in resonance mode and an electric generator which has a stable load. Change of load causes change of rotation speed therefore the system requires to be readjusted for resonance.

Experiments with powerful alternating current motors (about 100 Amperes as J.C. Ostwald wrote) working in resonance mode should demonstrate all advantages of the resonance mode application.

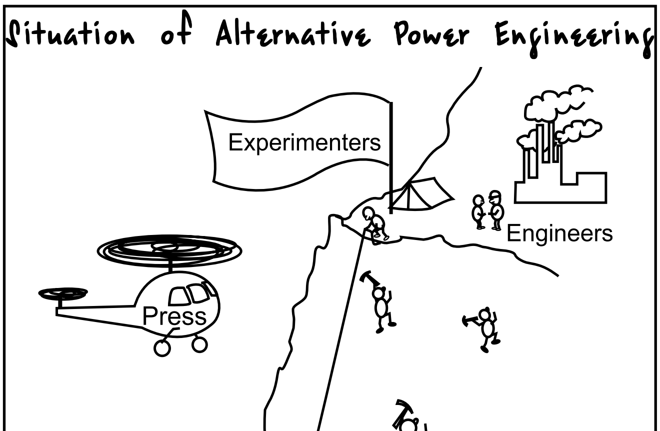
Situation of Alternative Power Engineering
Experimenters
Engineers
Press
Theorists

BOOKS REVIEW

Space Power Generator and Paramahamsa Tewari Update

Submitted by Toby Grotz, USA

http://www.tewari.org

A new generation of Space Power Generators is being built by Paramahamsa Tewari. The new design is expected to achieve substantial gains over previous models which has been measured over unity by experts in the utility industry as well as by independent consultants. Details are presented on the web site http://www.tewari.org.

From the Electron to a Perpetual System of Motion

Paramahamsa Tewari, B.Sc. Engg

As is well known, an electron, despite high-speed interactions with electric and magnetic fields and other particles of matter, remains unaffected structurally—maintaining its mass, charge, inertia, and locality after the interactions. This single fact itself is sufficient to postulate on the existence of some unique universal entity constituting the electron, such that the structure explains the known properties and behavior of the electron – as to how it sustains the collisions with the other particles and, yet, remains permanently indestructible without any reduction in the quantities of its basic properties. A theory that explains this enigmatic fact is framed and described in detail elsewhere. It reveals that the electron is a perpetually rotating system, a space-vortex (vortex of absolute vacuum) that through the process of motion in electric current, and interaction with external magnetic field can also lead to the development of a system partially violating Lenz's law and, thereby, capable of achieving perpetual motion in a machine. Such a conclusion has a new theoretical basis, as well as experimental confirmation – briefly outlined in this paper.

The basic understanding of the interaction of electric current with magnetic fields requires deeper knowledge of the fundamental nature of the electric current as well the magnetic field; this, in turn, leads to an inquiry on the nature of "electric charge" and its distribution in the electron structure. The origin of electron mass and charge, electrostatic and electrodynamic forces are described below with the vortex structure of electron, which further facilitates explanation of atomic structure and, thereafter, an over-unity efficiency generator. Due to brevity, predominantly qualitative-descriptions of the physical processes involved have been presented.

The full theory and diagrams are presented in From the Electron to a Perpetual System of Motion at http://www.tewari.org.

Universal Principles of Space and Matter (A Call for Conceptual Reorientation)

A new book, *Universal Principles of Space and Matter* has just been published by Paramahamsa Tewari. Arangements are being made with the publisher to provide these books outside of India. The price is expected to be about $30 depending on shipping costs from India.

A new theory of matter, that confronts with the existing concepts of space in the contemporary physics.

The basic phenomena dealt with are:

* Structural interrelationship between space and matter
* Origin of mass, inertia, and electric charge
* Creation of the electron from space

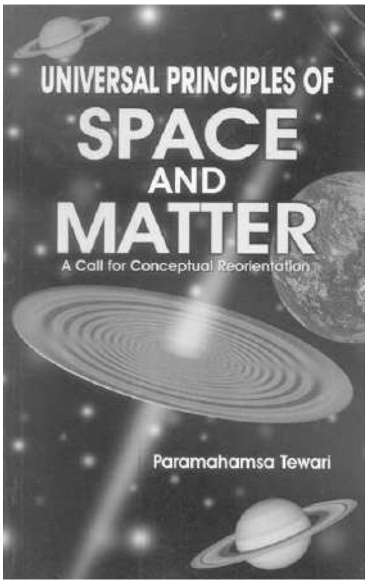

* Discovery of new fundamental equations on mass and charge
* Derivation of the presently known universal constants (gravitational, Coulomb's constant, dielectric, magnetic permeability, electron charge, electron mass, Planck's constant etc.) with the postulate of a single universal costant – the speed of light in the absolute vacuum
* Discovery of an inward force in nuclear structure against the Coulomb repulsive forces, hitherto unknown
* Relationship between light and gravity
* Theoretical derivation of the surface gravity of the Earth, Sun and the planets
* Discovery of electrical repulsive forces between the Sun and the planets, unknown in celestial mechanics
* Derivation of the planetary orbits
* Prediction of the continuous creation of universal matter at glactic centers and existence of electrical forces of interaction btween the stars and galaxies, so far least suspected
* Identifying the fundamental particle of matter
* Revelation of conceptual errors in the modern understanding of the basic nature of light
* Vindication of Einstein's conclusion of the speed of light as the limiting speed
* Vindication of Descartes' principle of property-less ether.

New Reviews of P. Tewari's work

New reviews of P. Tewari's work have been received from Dr. John A. Wheeler and Prof. Umberto Bartocci. A copy of Universal Principles of Space and Matter was forwarded this summer to Dr. John A. Wheeler at his summer residence in Maine.

Dr. Wheeler saw the development of modern physics while working with Neils Bhor at age 27 and was a colleague of Albert Einstein's at Princeton. Credited with being the "father of the Black Hole Theory", Dr. Wheeler is well schooled in quantum mechanics and relativity.

After receiving his copy of Universal Principle of Space and Matter Dr. Wheeler called the book "a unique gift". Further reviews of P. Tewari's work by Dr. Wheeler can be found at: http://www.tewari.org/Critical_Reviews/critical_reviews.html.

Prof. Umberto Bartocci

Universita' di Perugia
Dipartimento di Matematica e Informatica Via Vanvitelli
06100 PERUGIA - ITALY
Tel.: (39)-075-5002494
E-mail: bartocci@dipmat.unipg.it

This is a very singular book, in front of the current paradigm of contemporary Physics, a "cartesian-inspired" work which calls for reorientation in the foundations, by wisely warning that:

The conclusion of the modern physics that absolute space, time, simultaneity, and space filling media are discredited ideas is certainly premature (p. 178).

The books deals with arguments such as: Discovery of Charge and Mass Equations; Fundamental States of Cosmic Energy, Fields and Forces; Gravitation; Universal Constants; Motion of Electron; Atomic Structure; Light; Creation of Cosmic Matter (about Tewari's physical conceptions see also the second section of this Episteme's special issue http://www.dipmat.unipg.it/~bartocci).

JNET

New Energy Technologies collection of articles Japanese Version

Collection of articles from New Energy Technologies magazine, 2001-2002 is published in Japan.

On purchasing of the book, please, contact

Eiichi Yamamoto,
President Yama Trans Co. Ltd.

Email: admin@yamatrans.co.jp or
eyama@yamatrans.co.jp

Raum&Zeit

EhlersVerlag GmbH,
Geltinger Str.14e,
82515 Wolfratshausen
Telephon: 08171/41 84-76,
Telefax: 08171/41 84-66
Internet:
http://www.raum-und-zeit.com

Institut fur Raum-Energie-Forschung (IREF) i.m. Leonard Euler

- Global Scaling Theory - Hydrogen-Helium Energetic
- Gravitation theory - Information technologies
- Mathematic Biology - Telecommunication technologies
- Gravitation energetics- Transport technologies

Rotating device powered by permanent magnets

Minato Wheel

http://fdp.hemsida.net

Experiments made in Sweden

by Eric Vogels

Read about launch vehicle "Lightcraft" in the review "Commercial exploration of space"

http://www.lightcrafttechnologies.com

"Lightcraft" is made
for space flights
by means of megawatt
laser beam

Lightcraft Technologies, Inc.

Ball Lightning control

Experiments by Russian scientists

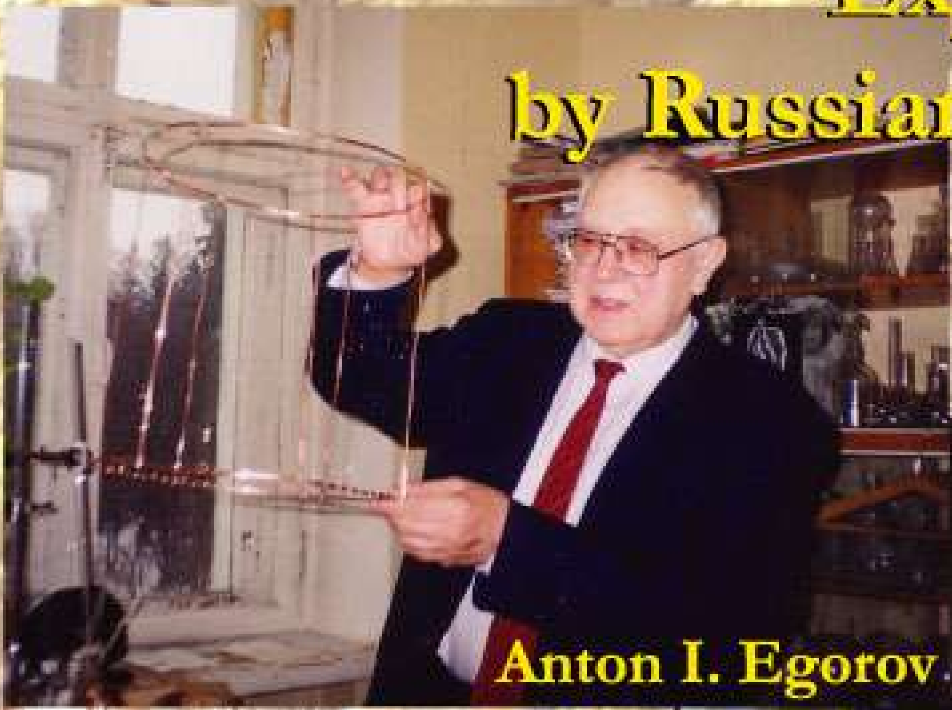

Anton I. Egorov

The standard ball formation

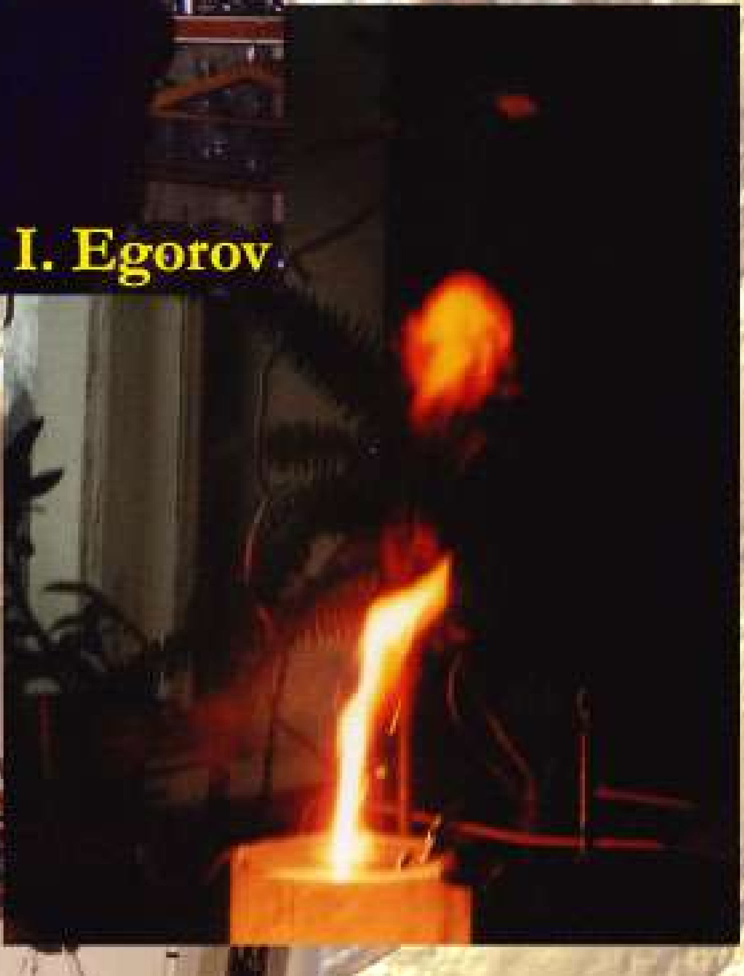

Read in this issue!

Tilley Foundation, Inc.

http://www.tilleyfoundation.com

K. Tilley
Electric Vehicle (TEV) - automobile of the future generation

ISSN 1684-7288

New Energy Technologies

Printed version ISSN 1684-7288
Online version ISSN 1684-9698

Published by Faraday Laboratories Ltd. *Issue #2 (11), 2003*

Research on advanced aerospace systems and new energy sources

Commercial exploration of space

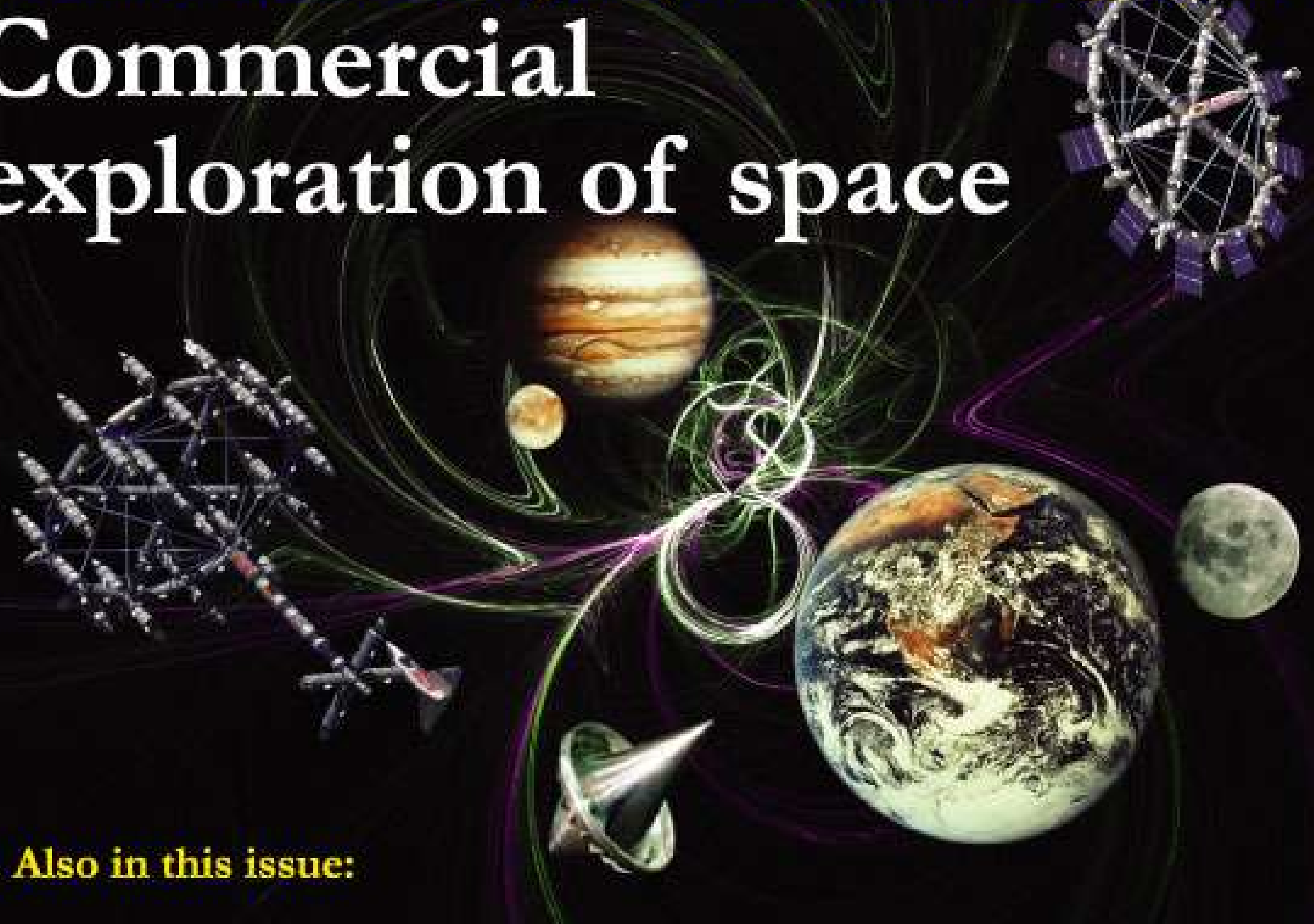

Also in this issue:

- New experiments on magnetic motors
- Ball lightning in laboratory environment
- Alternative aircrafts and space propulsion systems

NEW ENERGY TECHNOLOGIES #11

1. Minato Wheel. E. Vogels (Sweden)
2. Mechanism of Drive-Free motion. S.A. Gerasimov (Russia)
3. Production of Ball Lightning in Laboratory Environment. Report by A. Pashova (Russia)
4. Electric Discharge Produced into Air Half-Space. G.D. Shabanov, O.M. Zherebtsov (Russia)
5. Free Energy. A.V. Frolov (Russia)
6. Some General Properties of Matter and Energy Sources. E.D. Sorokodum (Russia)
7. Commercial Exploration of Space. Review by O. Leontyeva (Russia).
8. Prospects of Hydrogen Energetics. Ph.M. Kanarev (Russia)
9. Heater Producing Healthy Effect. A.L. Belyaeva and others (Kirghizia)
10. Inertial Propulsion Device. V.E. Senkevich (Russia)
11. Unlimited Accelerated Nonreactive Motion. Review.
12. Revolutionary Current. S.K. Baliyan (India)
13. Tilley Electric Vehicle. Review.
14. Trends in the Application of Motor-Generators. A. Akau (USA)
15. Electric Vehicle in Russia. Review by A. Pashova (Russia)
16. The Richard Clem Motor and the Conical Pump. R. Koontz (USA)
17. Spontaneous Flow Heat Pump. J.-L. Brochet (France)
18. Registration of Gravitational Waves. V.N. Petrov (Russia)
19. High-Energy Hyper-Low-Frequency Electric Field. Yu.V. Ivanko (Ukraine)
20. To the Question of Work Made by Electrostatic Field. A.V. Frolov (Russia)
21. Comments to Electro-Reactive LIFTER News.
22. Letters from Readers.
23. Advertisements (books)

Perfect Energy Source

Minato wheel. Experiments in Sweden

Eric Vogels, Sweden

E-mail: fdp@hemsida.net
http://fdp.hemsida.net

The self-running bicycle wheel was shown to the public at the First International New Science Symposium in 1997, held in Korea. The inventor of the wheel, Mr. Kohei Minato, was a speaker during the symposium. On www.keelynet.com you can find the quote of an eye-witness:

"The motor is actuated by moving the N pole of a large permanent magnet (the drive magnet) toward the wheel. As this magnet is moved toward the wheel, the wheel starts to spin. As the magnet is moved closer to the wheel it spins faster.

The acceleration of the wheel is rapid. So rapid in fact, as to be startling. To put it another way I was very impressed. The motor works. And it works very well. In the film clip (see http://fdp.hemsida.net) a slight pumping action of Minato's hand holding the magnet is apparent. When I braced my hand so that there was no pumping action, the motor still ran. In fact it seemed to run better.

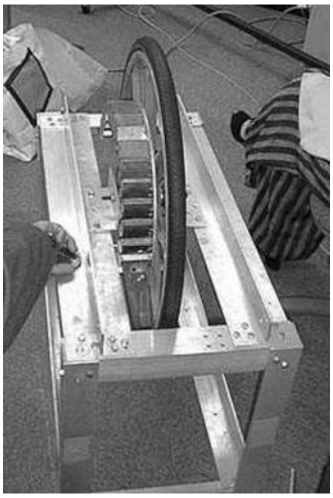

Fig. 1

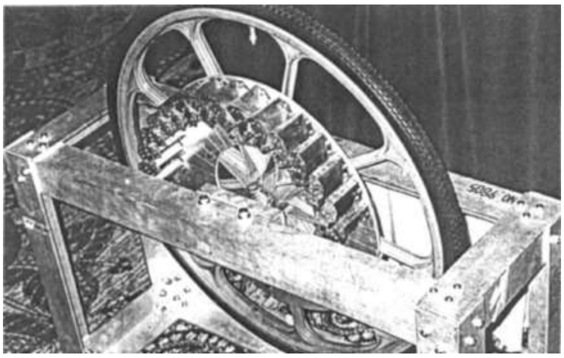

Fig. 2

Pumping action by the hand held magnet **is not** the power that drives the motor. When the drive magnet is moved away from the wheel it coasts rather quickly to a stop and comes to rest in a manner typical of any spinning bicycle wheel. Again when the wheel is at rest and a large magnet is moved up to the wheel it starts to spin. At no time it is necessary to touch the wheel to get it rotating."

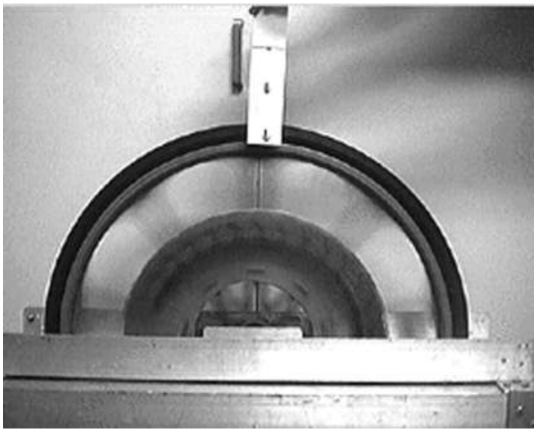

Fig. 3

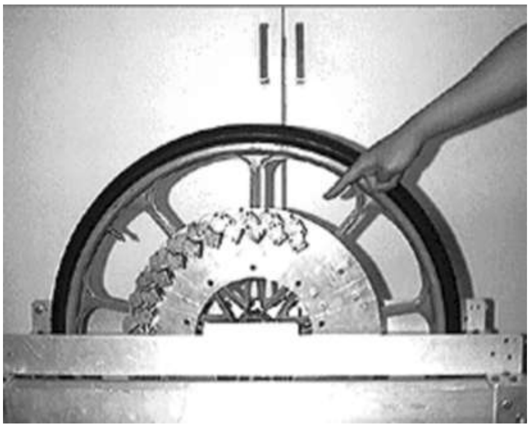

Fig. 4

Since the symposium in 1997 and the first publications in the Internet several people have tried to replicate the wheel, as it seemed to be a **perfect energy source for the future**. Some claims are made that replication was successful but no proof in the form of pictures or video clips are given.

Because I have a great interest in magnets and free energy and the fact that I did a lot of experiments to find a device that could be rotated only by means of permanent magnets, I decided to try to replicate the wheel by myself.

This journey started by searching on the Internet and collecting all the material that was available. A good way to sort all the found material was building my own website at http://fdp.hemsida.net where all my findings are presented.

One of the devices I build was made from a 'how to build a Minato wheel'-manual that I found on the Internet.

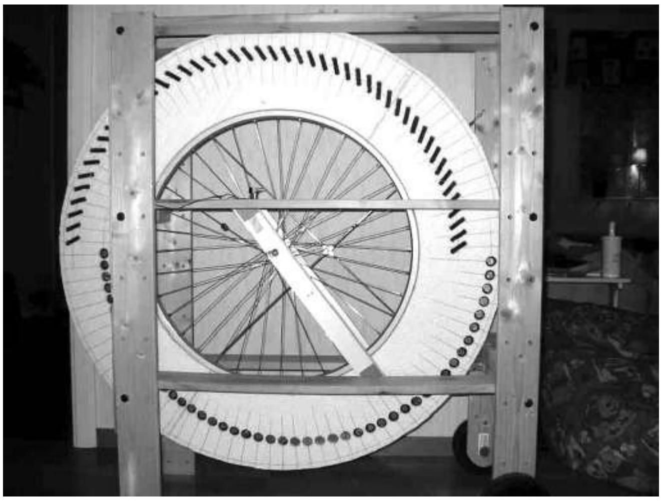

Fig. 5

Unfortunately this wheel did not show any movement at all. I mounted the magnets at several angles and distances from the rim without achieving any positive results. Since the size of the wheel (1 meter diameter) was not very practical to handle and since the results were negative I searched for an easier way for testing. I found that the use of Styrofoam, an old turntable and Neodymium disc magnets kept the whole project within the edges of my kitchen table. I developed a computer program to help me design wheels, since I want to be able to replicate my own wheel, in case I find a model that works. I do not believe in an engine that works after several weeks of tuning and that fails to function as soon as one magnet is replaced. Individual tuning is okay to optimise a working design but if a design does not show the potential to rotate by itself it's a flawed design.

After optimising the way to build the small wheels, I could design and make a wheel in 30 minutes. I made a test scenario and the results for every test were posted on the web site.

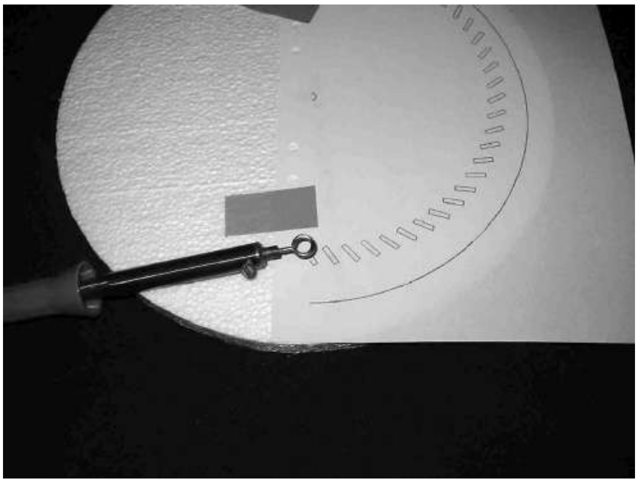

Fig. 6

I searched for the 'perfect' angle, number of magnets and position. The set-up with the strongest thrust was used as the base for the other experiment. By this time I realized that I was not really trying to rebuild the Minato wheel but that I wanted **a rotating device, powered by permanent magnets.**

When for the first time I decided to split a track of magnets that covered 180 degrees of the wheel in a number of smaller tracks, **the results became much better**.

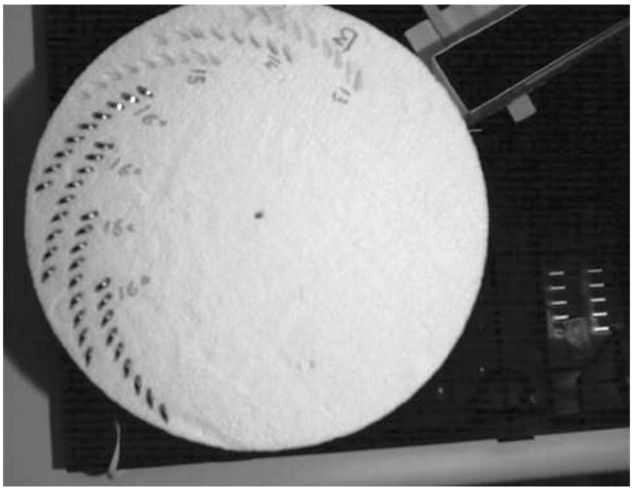

Fig. 7

By using this way of placing the magnets, every track differs a little bit from the next track, depending on its place on the wheel. The angle of the magnets influences on the thrust that is given by the stator magnet, depending on the speed that the track is entering the magnetic field of the stator. In other words: the track at the beginning of the wheel, with magnets covered side of the wheel, gives less thrust than the track at the end. This is important since a track with a lot of thrust has a big 'sticky spot' that should pass the stator. This 'sticky spot' stops the wheel during its rotation just before the first track enters the magnetic field of the stator.

During the last experiments the stator magnet changed place and pushed against the magnetic fields of the tracks from above. Simply because more thrust was developed by this way. The next wheel will have the tracks mounted on a different level. It means that the first track will have a bigger distance to the stator magnet than the last track. This should make the sticky spot weaker and the final 'kick off' stronger. The final wheel will be mounted vertically, like the Minato Wheel.

Thanks to the invention of the Neodymium magnet in 1983, magnets became much more powerful while the size and length decreased. Before 1983 a lot of the experiments could not be done since the size of the magnets made the positioning of several magnets in tracks like this was impossible. It is my personal believes that the strength of these magnets will lead to results that are impossible according to the physical law books, because these magnets did not exist when those books were written.

Until the moment I write this I have not found the perfect energy source yet but by exploitation all the material and findings on my web site I hope that more people get interested and will help searching for a great energy device that will not pollute the air which our children and grand children have to breath.

Mechanism of Drive-Free Motion

Sergey A. Gerasimov, Russia

Physics Department, Rostov State University, Rostov-on-Don, 344090
Email: GSIM1953@mail.ru

*The article presents results of the experimental study of the motion created by vibrations of an internal mass of the system of bodies accompanied by impacts of the unbalanced load with the external body. This type of propulsion drive is known by a number of names. Among these are the **vibrational propulsion device** and **reactionless machine**. Sometimes it is called the **inertioid** [1, 2]. The vibrational transposition is proved to go on by means of internal forces of a system of bodies [3-5]. The Tolchin's inertioid [6] is considered to be the first device that used the forces of inertia to create the reactionless infinite motion in space. We are not in a position to be a judge of this. There exist a number of projects of such machines but experimental results concerning such kind of motion are very limited. Below there are the experimental results on average velocities of such a motion and the description of a propulsion device a main particularity of which is absence of wheel-drive.*

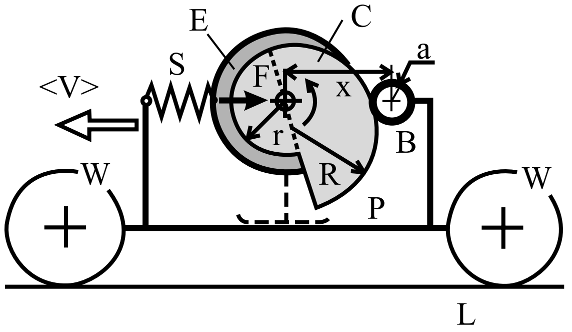

Fig. 1

Scematic representation of a vibratory-impact self-transposition

The unbalanced load in this device is an electric motor *E* of mass *m* that executes undumped vibrations relative to a platform *P* of mass *M* with four wheels *W* which can roll on a horizontal surface *L* with rolling friction the coefficient of which is *k*. The frictional force is determined to be a force which adjusts to keep the cart from motion across a surface. A disk cam *C* at one end of the axle of the motor ensures a prescribed character of the vibrations and elastic impacts between the load *m* and the follower *B* of radius *a*. In present experimental device, the cam *C* consists of two semicircles of different radii *r* and *R* as it is shown in Fig.1. In this experiment $r=0.03$m, $R=0.05$m and $a=0.01$m. The electric motor and the platform are coupled by a connecting spring *S*. The force *F* by means of which the electric motor is pressed to the follower varies linearly from $F=4$N at $x=0.04$m to $F=4.8$N at $x=0.08$m. The compressed spring *S* is necessary not only for creating close contact between the cam and the follower. The restoring force of the spring produces the transposition of the cart in a direction opposite to **F** when the follower moves without contact from $x=2R-r+a$ up to an impact at $x=r+a$. The impact suppresses the transposition. Another mode of transposition is also possible when the impact of the cam on the follower is a reason of the transposition of the cart in the direction of the force **F**. In this case the frictional force extinguishes the transposition of the cart.

Friction between the load of mass *m* and the cart is negligible since the corresponding effective coefficient is less than 0.001. The same is for the friction between the cam and the follower. The mass of the spring is $m_S=0.007$ kg , and the mass of the cam is $m_C=0.019$ kg. The total mass of the wheels is $m_W=0.082$ kg. This value can be useful for a theoretical

During the last experiments the stator magnet changed place and pushed against the magnetic fields of the tracks from above. Simply because more thrust was developed by this way. The next wheel will have the tracks mounted on a different level. It means that the first track will have a bigger distance to the stator magnet than the last track. This should make the sticky spot weaker and the final 'kick off' stronger. The final wheel will be mounted vertically, like the Minato Wheel.

Thanks to the invention of the Neodymium magnet in 1983, magnets became much more powerful while the size and length decreased. Before 1983 a lot of the experiments could not be done since the size of the magnets made the positioning of several magnets in tracks like this was impossible. It is my personal believes that the strength of these magnets will lead to results that are impossible according to the physical law books, because these magnets did not exist when those books were written.

Until the moment I write this I have not found the perfect energy source yet but by exploitation all the material and findings on my web site I hope that more people get interested and will help searching for a great energy device that will not pollute the air which our children and grand children have to breath.

Mechanism of Drive-Free Motion

Sergey A. Gerasimov, Russia

Physics Department, Rostov State University, Rostov-on-Don, 344090
Email: GSIM1953@mail.ru

The article presents results of the experimental study of the motion created by vibrations of an internal mass of the system of bodies accompanied by impacts of the unbalanced load with the external body. This type of propulsion drive is known by a number of names. Among these are the ***vibrational propulsion device*** *and* ***reactionless machine****. Sometimes it is called the* ***inertioid*** *[1, 2]. The vibrational transposition is proved to go on by means of internal forces of a system of bodies [3-5]. The Tolchin's inertioid [6] is considered to be the first device that used the forces of inertia to create the reactionless infinite motion in space. We are not in a position to be a judge of this. There exist a number of projects of such machines but experimental results concerning such kind of motion are very limited. Below there are the experimental results on average velocities of such a motion and the description of a propulsion device a main particularity of which is absence of wheel-drive.*

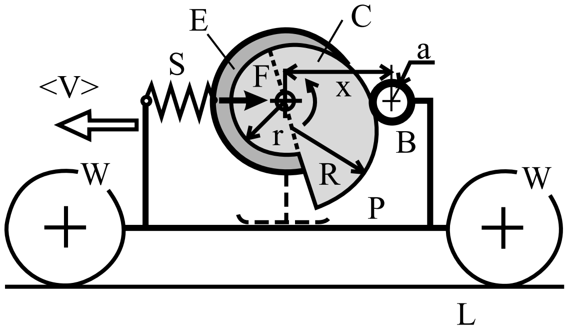

Fig. 1

Scematic representation of a vibratory-impact self-transposition

The unbalanced load in this device is an electric motor E of mass m that executes undumped vibrations relative to a platform P of mass M with four wheels W which can roll on a horizontal surface L with rolling friction the coefficient of which is k. The frictional force is determined to be a force which adjusts to keep the cart from motion across a surface. A disk cam C at one end of the axle of the motor ensures a prescribed character of the vibrations and elastic impacts between the load m and the follower B of radius a. In present experimental device, the cam C consists of two semicircles of different radii r and R as it is shown in Fig.1. In this experiment r=0.03m, R=0.05m and a=0.01m. The electric motor and the platform are coupled by a connecting spring S. The force F by means of which the electric motor is pressed to the follower varies linearly from F=4N at x=0.04m to F=4.8N at x=0.08m. The compressed spring S is necessary not only for creating close contact between the cam and the follower. The restoring force of the spring produces the transposition of the cart in a direction opposite to **F** when the follower moves without contact from x=2R-r+a up to an impact at $x=r+a$. The impact suppresses the transposition. Another mode of transposition is also possible when the impact of the cam on the follower is a reason of the transposition of the cart in the direction of the force **F**. In this case the frictional force extinguishes the transposition of the cart.

Friction between the load of mass m and the cart is negligible since the corresponding effective coefficient is less than 0.001. The same is for the friction between the cam and the follower. The mass of the spring is m_S=0.007 kg , and the mass of the cam is m_C=0.019 kg. The total mass of the wheels is m_W=0.082 kg. This value can be useful for a theoretical analysis. Thus, the total mass of the unbalanced body is chosen to be much larger than the mass of the cam or the spring: $m = m_E + m_C + m_S \approx m_E$.

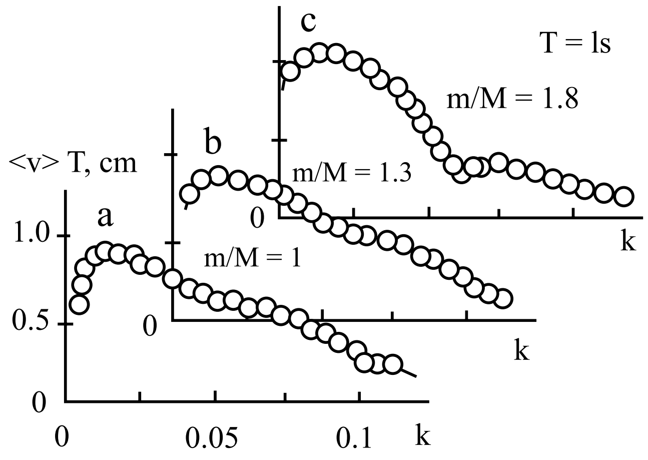

Fig. 2
Frictional coefficient dependence of average velocity of the platform at the period of vibrations T=1s for various mass ratios m/M:
(*a*) M=0.9 kg , (*b*) m=1.2 kg , (*c*) m=1.6 kg

At least in the case of irreversible mode, one could expect that the value of average velocity of transposition $<v>$ must be proportional to the number of impacts per unit of time. In the other words the path traveled by the cart per one impact must not depend on frequency of vibrations. This is a reason why the dependence of the product $<v>T$ versus the value of the friction coefficient k and ratio of masses d = m/M is investigated in this work. Measurements were carried out for two magnitudes of rotation period T= 1s and T=6s of the cam. The obtained results are presented in Fig. 2 and Fig. 3. First of all, the assumption mentioned above is not confirmed.

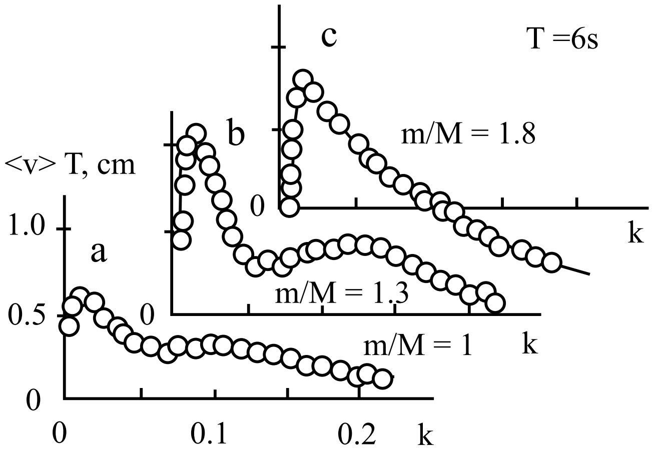

Fig. 3
Average velocity of the platform $<v>$ as a function of frictional coefficient k at T=6s for various mass ratios m/M: (*a*) M=0.9 kg , (*b*) m=1.2 kg , (*c*) m=1.6 kg

The value $<v>T$ for the period of the rotation T=1s sufficiently differs from that for T=6s measured at the same conditions. Besides, when the mass ratio is large the cart can change its direction of transposition. The reason of these results is not clear but this device provides a challenge to explain and investigate this kind of motion that enables us to discover the reactionless motion if it is probable. One should pay attention to a fact that the infinite transposition of the platform takes place even if the friction coefficient k is small.

References

1. Gulia N.V. *Inertia.* // Moscow: Nauka. 1982.
2. Aspden H. Anti-gravity Electronics. // *Electron and Wireless World.* 1989. V. 95. No 1635. P. 29-31.
3. Nappo F. Sulla Influenza Indiretta Delle Forze Inertne sul Moto del Baricentro. // *La Ricerca Scientifica (Rendiconti).* 1965. V. 8. No 1. P. 3-14.
4. Gerasimov S.A. Anisotropy of Frictional Force and Vibrational Transposition. // *Problems of Applied Physics.* 2001. V 7. P. 85-88.
5. Gerasimov S.A. Self-Similarity of Vibrational Motion in a Resistant Medium. // *Journal of Applied Mechanics and Technical Physics.* 2002. V. 43. No 1. P. 90-92.
6. Tolchin V.N. *Inertioid.* // Perm: Perm Book Publisher. 1977.

About the Author

Sergey A. Gerasimov graduated from Faculty of Experimental and Theoretical Physics of Moscow Physical Engineering Institute in 1976. Received Degree of PhD in Physics and Mathematics in 1987. Author of about 70 articles on Astrophysics, Atomic and Radiation Physics, Classical Electrodynamics and Mechanics. In present - Associate Professor of Department of General Physics of Rostov-on-Don State University. The fields of scientific interests are questionable problems in physics: self-interaction, self-transposition, unipolar induction.

ELECTRIFYING TIMES
an online and published
magazine about Electric,
Hybrid, Fuel Cell Vehicles,
advanced batteries, ultra capacitors,
fuel cells, microturbines, free energy systems,
events and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908 fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

Production of Ball Lightning in Laboratory Environment

Report by our correspondent Alla Pashova

Three years ago a working device producing a ball lightning in laboratory environment was designed in Saint-Petersburg Institute for Nuclear Physics (SPINP). The produced ball lightning is accessible for detail investigation and quite stable. The lifetime of the ball lightning comes to about one second that is rather significant for alike artificial formations. The experiments on the device made by scientists of SPINP, i.e. A. E. Egorov, G. D. Shabanov, S. Stepanov, are not supported or financed. Let us note that every scientist of the group searches proofs for his own hypothesis of nature and structure of the ball lightning at all.

A leading expert of SPINP, Anton I. Egorov, pays attention to dethronement of scientific myths:

— There is a myth of ball lightning that is created by mass media. The mythical ball lightning is a concentrate of mysterious energy which is extremely dangerous for a human. It destroys houses, kills animals, pursues people. After meeting it a human can lose his hair or teeth and different misfortunes begin to happen. Supposing a simple story of a farmer-eyewitness: "It thundered, and a fist-sized fire ball rolled down along a drainpipe. It fell into a barrel of water, the water gurgled. I came up to it and put my hand into the water. The water seemed to become warmer…". After republishing of the story by several newspapers a dramatic story about a ball lightning which has evaporated a barrel of water appears. No wonder that such familiar attitude to facts causes hundreds of hypotheses of ball lightning nature.

— **What is your hypothesis about ball lightning structure?**

— At the beginning of 90th I. D. Stakhanov, a member of Institute of Magnetism (IZMIRAN), developed a special method to interview eye-witnesses that resulted a right notion on ball lightning phenomenon. According to Stakhanov, ball lightning is a clot of hydrated plasma which is generated in wet air at electrical discharge.

Water as a chemical compound is remarkable for its anomalous properties: combining of two lightest elements does not generate gas but produces a high-boiling liquid. This is caused by extremely irregular distribution of electrons in a water molecule. Due to this property it acquires properties of an electric dipole. Water molecules interact with charged ions, aerosol particles, and with each other in a special way.

If a positive ion and a negative ion are introduced simultaneously into a puff of warm wet air then water dipoles immediately produce hydrate shells around the ions. When the hydrated positive ion approaches the hydrated negative ion additional water molecules are drown into gaps between them. As a result, there is a stable cluster in which the charged ions are conserved. The cluster consists of two ions of opposite charges and the hydrate shell. Water molecules prevent the ions' approaching and recombination, hence, lifetime of the ions in the cluster increases up to tens of minutes, i.e. in 12-13 orders. Interaction of clusters causes open-chain structures followed by space structures. That is to say, there is appeared a clot of cold hydrated plasma that accumulates great energy (up to 1 kylojoule per liter). The clot of plasma loses this energy at recombination of ions.

— **Could you, please, tell about the design of the device? What processes proceed when the device is operating?**

— Our task is to introduce an abundant population of ions into a puff of warm air saturated with water vapors. A base of the device for laboratory reproduction of ball lightning is a capacitor bank which is able to be charged up to 5.5 kV. The positive pole of the capacitor bank is connected to a ring electrode by means of a copper bar. The ring electrode is placed on a bottom of a polyethylene reservoir filled with water. The negative pole of the capacitor bank is connected to a carbonic electrode which is placed in the centre of the reservoir near water surface. A quartz pipe encloses the electrode in such a way that it is possible to drop water or to put some natural matter on it.

To generate a ball lightning 2-3 drops of water are put on the electrode. When impulse discharge occurs a bright plasma spout escapes the centre of the electrode that is accompanied by a quiet plop. A glowing plasmoid which is an artificial ball lightning parts from the plasma spout. It comes up slowly in the air and then disappears falling to pieces in 0.2-0.3 seconds.

We have made thousands of experiments for investigation of ball lightning properties, i.e. defining size, lifetime, colour, average temperature, excessive charge, content of a dust component.

It was ascertained that the artificial ball lightning is generated in a narrow interval of breakdown tensions. The average size of such a ball lightning is 12-20 cm, and its lifetime comes to 1 second. Temperature of the ball lightning is not very high, i.e. 50° C. This can be defined if to take into account speed of rising of the ball lightning. If the plasmoid is accepted as a puff of warm wet air of 14-centimeter diameter which in atmosphere comes up at 293 K at speed of 1-1.2 m/sec then, consequently, its temperature should not exceed 330 K.

Colour of the lightning varies greatly and depends on presence of aerosol of matter trapped in the moment of discharge. Usually the lilac central part of the plasmoid is surrounded by a diffusive yellowish layer. Some admixture of natrium salt and calcium makes the core of the plasmoid yellow or orange.

If the central carbonic electrode is replaced by an iron or copper or aluminum one then the fundamental character of the phenomenon does not change. However, colour of the plasmoid depends on a radiation spectrum of excited atoms of the electrode, i.e. iron plasmoids are whitish, copper plasmoids are greenish, aluminum ones are white with reddish shimmer.

— **A generated ball lightning exists for about 1 second. How can it be made more stable?**

— Lifetime of an artificial ball lightning depends on many conditions, i.e. size and geometrical form of the central electrode, voltage between the electrodes, value and duration of a current impulse, temperature and electroconductivity of water which is put on the central electrode. Besides, lifetime of the plasmoid can be changed by introducing an additional dispersive phase into it. We have tested tens of matters and begun to investigate suspensions of colloidal graphite and fine-dyspersated ferric oxide.

A suspension of 3 g of colloidal graphite, 8-10 ml of acetone (which played role of penetrating agent), and 90 ml of water is put on the central carbonic electrode. When an electrical discharge occurs a layer of the suspension forms a flying spherical plasmoid. It comes up slowly and disappears in 0.3-0.8 seconds. The core of the plasmoid has a colour of flame, i.e. colour of burning carbon.

To prolong the existence of the generated ball lightning without application of aerosols it will be possible to use the so-called "Faraday's cylinder" whose production has been already begun. For the same purpose G.D. Shabanov proposes to put a stopping potential on a probe of detention.

— **There is an opinion that physical nature of ball lightning is similar to the process of controlled thermonuclear fusion. In this case, if your work on generation of a stable ball lightning is successful then you will be a competitor of the expensive project of controlled thermonuclear fusion**.

— I think that it is totally incorrect. Hydrated plasma is the first enemy of the thermo-nuclear fusion as water molecules do not allow neutrons approach each other. Effective cold fusion should be realized in organic liquids, for example, in heavy acetone or in a water-free medium. Somehow, it should be an absolutely "dry" process. No experiments on real "dry" cold fusion have been realized in practice. It has not been also examined the surfaces on which the combination of heavy hydrogen atoms produces maximal heating.

Scientists should pay attention to two most effective cold fusion processes. The first one expects association of two atoms of deuterium occurring on a totally dry deuterated surface which consists of, for example, zirconium deuteride. In the moment of fusion of a deuterium molecule local heating appears, and neutrons depart. The other perspective method of realization of the cold fusion process requires an absolutely "dry" organic liquid, i.e. liquid acetone in which hydrogen atoms are replaced by deuterium atoms (C_3D_6O) or by atoms of cyclic compound of $C_6(D_2)_6$. A tellurium or zirconium tip of an ultra-sonic dispersant is placed into a reservoir of this liquid. Cavitation blebs are formed on the surface of the dispersant. Neutron output comes to 10^4 particles. Maximal neutron output, which was achieved by Lipson, an American experimenter, consists of 10^8 neutrons at a desired result of 10^{13}. Certain quantity of neutrons can be obtained during ultrasonic cavitation which is accompanied by a phenomenon of sonoluminescence. Due to sound resonance the only cavitation bleb is generated in acetone. When the bleb collapses weak glowing is observed. The cause of this phenomenon consists in gas heating occurring in the bleb that is the result of high pressure produced by its collapse. The burst can last from 1/20 up to 1/1000 sec. Light intensiveness depends on quantity of gas in the bleb. If gas is absent in the bleb then the glowing does not occur. Light emission of the bleb is very weak, it becomes visible if it is strengthen or in absolute darkness.

— Is it rightful that cold fusion is the future of world power engineering?

To my mind another direction seems to be more perspective, i.e. extraction of uranium from sea water and then its burning in heavy hydrogen reactors like one which exists in Canada. Photo voltaic accumulators can also become a successful approach of the alternative traditional fuel engineering. By the way, a working model of such a device to utilize free solar energy has recently been created in our Institute (official web site: http://www.pnpi.spb.ru).

Editor: Read the publications on this theme in following issues of our magazine. Below there is a description of other attempts to generate a ball lightning in laboratory environment or at home. Besides, we publish an article dedicated to the problem of laser control of ball lightning. In the article there are photos (also see the cover page) and a scheme of the working device designed by this scientific group.

Ball Lightning Experiments

Information from
http://www-personal.umich.edu/~reginald/ball_l.html

Ball Lightning in the Microwave

This is a great experiment to try at home. It requires a microwave oven, a candle, a toothpick and a lighter. First you put the candle (~1-2 inches in height) in the microwave (remove glass plate from bottom so candle sits on metal). Put the toothpick in the candle sticking straight up. Light the toothpick on fire so that flames are leaping off the tip. Shut the door quickly and turn on oven full blast. There will be loud popping noises and then balls of fire will leap from the toothpick and fly around inside the microwave while making a buzzing sound like a bee. If it doesn't work at first try to move the position of the candle in order to find the "hot spot" in the oven for it to work.

Real Ball Lightning Generated by Pulsed Power Inductor

This experiment is very DANGEROUS. The author did this one a few years ago. The Idea was got from an article about a guy generating ball lightning using a high current transformer (TBA). The author had an idea of using an inductor to store large amounts of electrical energy. The conductor on the end of the stick touches the metal ring. This completes the circuit, and start the inductor. Then there was used the air compressor nozzle to blow out the arc as the wire was pulled away. The copper wire explodes and creates rapidly rotating and burning molten balls of copper. These 'Balls of Fire' exhibit many of the properties of real ball lightning.

Electric Discharge Produced into Air Half-Space

Gennady D. Shabanov, Russia

Tel. +7(271)46477
E-mail: discharge@gtn.ru

Oleg M. Zherebtsov, Russia

Tel. +7(271)46272
E-mail: omzher@mail.ru

In the article a possibility of streak lightning control by means of low intensive laser emission is observed.

During investigation of an electric discharge into air half-space [1] it was noticed that this discharge can be controlled by a low intensive laser beam. Now there is a problem of laser control of lightning discharge [2]. However, numerous works on this approach demonstrate that "the hope to get a quick solution of the problem of lightning control by laser emission has not been confirmed" [3].

Careful observation of this problem in [2] has shown ways out the situation. The authors of [2] consider plasma channel produced (by means of laser) in free atmosphere at a possible greatest height to be of doubtless interest of the science of lightning. Finally, creation of the plasma channel should be of benefit for lightning protection. The authors of [2] give notice that "there are many difficulties of fundamental and practical importance on this path".

Weak theoretical understanding of lightning generation causes pessimistic estimations of solution of the problem. The article [2] notes particularly that "there are neither adequate theory, nor numerical calculations and qualitative understanding of the phenomena defining the speed of a leader… The situation of a theory of the leader channel is little better (from quantitative point of view)…".

Realistically the following statement can refer to the lightning discharge: "The electric discharge appeared to be very "unhandy" for theoretical description but the most interesting phenomenon in the experimental aspect" [4].

Experimental Part

In this work a capacitor bank with 0.6 mF capacity, which can be fed up to ~5 kV, was used to produce impulse discharge into air half-space. A scheme of the device is presented in Fig.1. At connection/disconnection of a discharger 5 a "spout" is let out of an electrode 3. The spout carries the potential of the cathode (virtual cathode) at a significant height into air half-space. A probe placed at the height of ~15 cm fixes a potential which is similar to the potential occurring at the cathode. The researches have demonstrated that the produced formation continues to glow for several hundreds of milliseconds (the glow is fixed from a zone located at 15-45 cm above the cathode). Typical time of the discharge comes to 100±20 msec and depends on the cathode material. Electric field generated in the spout comes to less than 8 V cm^{-1}.

Ball Lightning Experiments

Information from
http://www-personal.umich.edu/~reginald/ball_l.html

Ball Lightning in the Microwave

This is a great experiment to try at home. It requires a microwave oven, a candle, a toothpick and a lighter. First you put the candle (~1-2 inches in height) in the microwave (remove glass plate from bottom so candle sits on metal). Put the toothpick in the candle sticking straight up. Light the toothpick on fire so that flames are leaping off the tip. Shut the door quickly and turn on oven full blast. There will be loud popping noises and then balls of fire will leap from the toothpick and fly around inside the microwave while making a buzzing sound like a bee. If it doesn't work at first try to move the position of the candle in order to find the "hot spot" in the oven for it to work.

Real Ball Lightning Generated by Pulsed Power Inductor

This experiment is very DANGEROUS. The author did this one a few years ago. The Idea was got from an article about a guy generating ball lightning using a high current transformer (TBA). The author had an idea of using an inductor to store large amounts of electrical energy. The conductor on the end of the stick touches the metal ring. This completes the circuit, and start the inductor. Then there was used the air compressor nozzle to blow out the arc as the wire was pulled away. The copper wire explodes and creates rapidly rotating and burning molten balls of copper. These 'Balls of Fire' exhibit many of the properties of real ball lightning.

Electric Discharge Produced into Air Half-Space

Gennady D. Shabanov, Russia

Tel. +7(271)46477
E-mail: discharge@gtn.ru

Oleg M. Zherebtsov, Russia

Tel. +7(271)46272
E-mail: omzher@mail.ru

In the article a possibility of streak lightning control by means of low intensive laser emission is observed.

During investigation of an electric discharge into air half-space [1] it was noticed that this discharge can be controlled by a low intensive laser beam. Now there is a problem of laser control of lightning discharge [2]. However, numerous works on this approach demonstrate that "the hope to get a quick solution of the problem of lightning control by laser emission has not been confirmed" [3].

Careful observation of this problem in [2] has shown ways out the situation. The authors of [2] consider plasma channel produced (by means of laser) in free atmosphere at a possible greatest height to be of doubtless interest of the science of lightning. Finally, creation of the plasma channel should be of benefit for lightning protection. The authors of [2] give notice that "there are many difficulties of fundamental and practical importance on this path".

Weak theoretical understanding of lightning generation causes pessimistic estimations of solution of the problem. The article [2] notes particularly that "there are neither adequate theory, nor numerical calculations and qualitative understanding of the phenomena defining the speed of a leader… The situation of a theory of the leader channel is little better (from quantitative point of view)…".

Realistically the following statement can refer to the lightning discharge: "The electric discharge appeared to be very "unhandy" for theoretical description but the most interesting phenomenon in the experimental aspect" [4].

Experimental Part

In this work a capacitor bank with 0.6 mF capacity, which can be fed up to ~5 kV, was used to produce impulse discharge into air half-space. A scheme of the device is presented in Fig.1. At connection/disconnection of a discharger 5 a "spout" is let out of an electrode 3. The spout carries the potential of the cathode (virtual cathode) at a significant height into air half-space. A probe placed at the height of ~15 cm fixes a potential which is similar to the potential occurring at the cathode. The researches have demonstrated that the produced formation continues to glow for several hundreds of milliseconds (the glow is fixed from a zone located at 15-45 cm above the cathode). Typical time of the discharge comes to 100±20 msec and depends on the cathode material. Electric field generated in the spout comes to less than 8 V cm^{-1}.

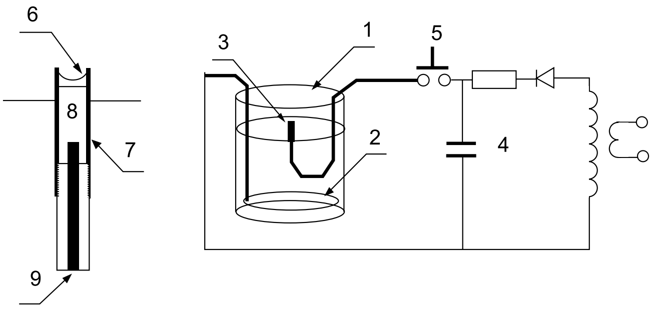

Fig 1
A scheme of a device made for producing of long-living plasmoids

1 – a polyethylene reservoir, 2 – a ring electrode, 3 – a central electrode, 4 – a capacitor bank of 0.6 mF capacity, 5 – a discharger, 6 –water or aqueous suspension drop, 7 – a quartz pipe, 8 – a carbonic or metal electrode, 9 – a copper bar.

The spout was acted by weak laser emission of less than 1 mW at 22-centimeter height. The laser beam was directed perpendicular to the discharge axis, the spout achieving the laser beam moved along the beam towards the laser.

In Fig.2 the spout has moved ~8.5 cm along the laser beam, and a usual sphere formation was generated. Due to the horizontal motion it was distorted if to be compared with usual sphere formation. The usual sphere formation is represented in Fig.3 (also see the cover).

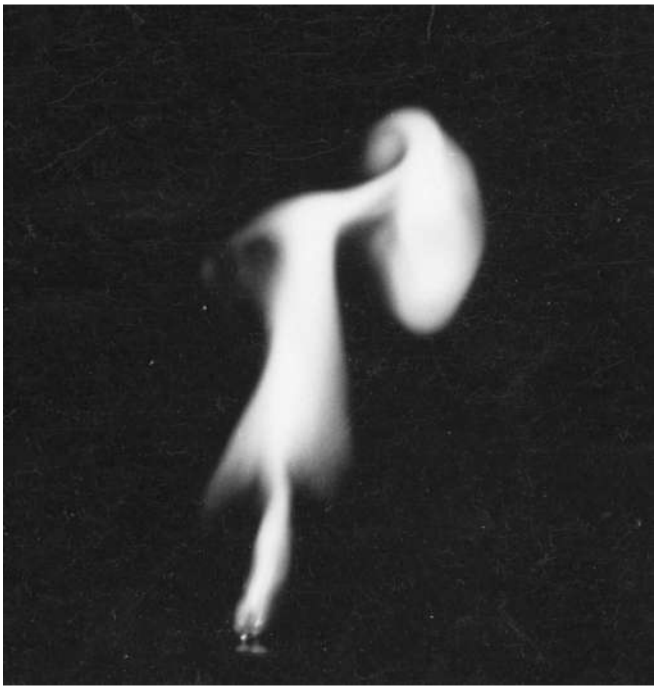

Fig. 2
Motion of the leader channel (spout) along the laser accompanied with the sphere formation

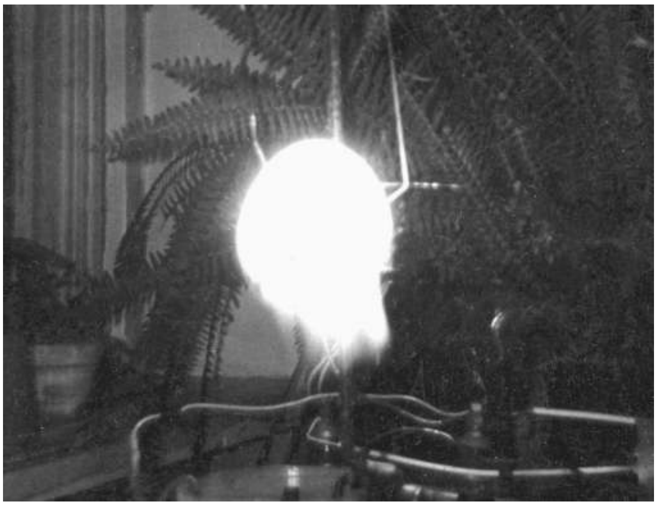

Fig.3
The standard ball formation

Discussion

As it is evenly mentioned in [5], results of model experiments are difficult to be applied directly to laser control of lightning because in short intervals characteristics of the discharge differ from characteristics of lightning discharges. Actually, at comparing considered current and the field existing in a lightning (i~100 A, ~3 V cm^{-1}) with these phenomena occurring in a laboratory spark (~1 A, ~300 V cm^{-1}) incorrectness of modeling of this process becomes obvious. From the other hand the authors of [2] note that “the leader channel is like a channel of electric arc… At current strength of ~100 A plasma of the channel of the arc is supported by fields the densities of which come to several volts by a centimeter. The lightning has such leader currents”.

Even an air arc of atmosphere pressure has a field of about 100 V cm^{-1} at currents of about 1 A. The generated discharges (spout) accompanied by the field of less than 8 V cm^{-1} (maximal current appearing in the discharge gap is 50-60 A) are more appropriate for modeling the leader channel of the streak lightning than typical discharges used for investigation of these processes, for example, like [6].

At first approximation the leader channel is considered as an ideal conductor in the article [2]. We make research to define the field in the discharge (spout) more exactly. The article [2] makes the following consideration about a cause of occasional generation of new leader heads: “the surface of *equipotential* plasma conductor (channel) *has* a property of *instability*. There is an occasionally generated sharp jut. An intensified field appears at the jut along the edge. Under the influence of the field the jut becomes to grow in any direction including at the significant angle to the weak outer field”. The aforesaid seems to explain why our leader channel changes its direction and runs at right angle to its initial motion. (Fig.2). Weak harmonic action of the laser beam to the leader channel causes generation of a new head which continues motion “on the significant angle”.

The effect of plasma motion towards a light beam is a demonstration of a general tendency of propagation of the discharges to an incident electromagnetic field [7]. "Area occupied by plasma usually increases towards the laser emission" [8]. Leading character of motion of this formation is confirmed in [9] as well as the interaction with the laser beam is confirmed in [10]. It should be taken into account that in the works [9-10] experiments were performed at devices which generated such a discharge. However, that discharge was by 2 orders weaker than the discharge generated by the device represented in this work. Naturally, the results were less defined. Comparison of the discharges is presented in [11].

According to our data, the leader channel (spout) has a very abrupt bound (less than 1 mm). In this layer the field can come to ~30 kV cm^{-1} (at height of 15 cm).

Conclusions

Due to the assumption [2] about *instability* of the surface of the *equipotential* leader channel there has been successfully performed the "control" action on an electric discharge which models a streak lightning. It has been achieved by means of weak harmonic oscillations made by the laser. This mechanism is supposed to be applied for streak lightning control.

References

1. G. D. Shabanov. Optical properties of glowing formations // Letters in the Magazine of Technical Physics. 2002, V.28, I.4, P. 81-86;
2. E. M. Bazelyan, Yu. P. Rizer. Mechanism of attraction of lightning and problem of lightning laser control (Nowadays Physics) // the UPS. 2000. V. 170, #7, P.753-769;
3. A.D. Starikov, Yu. A. Rezunkov. Lightning discharge control by laser emission. // the Optic Magazine. 1999, V. 66, #3, P. 15-16;
4. B. E. Meyerovich on the path to realization of electromagnetic collapse // the UPS. 1986. V. 149, #2, P.221-257;
5. Miki Megumu, Uada Acushi, Shindo Takatoshi. Characteristics of laser controlled discharges occurring in long air gaps // the Optic Magazine, 1999, V. 66, #3, P. 25-29;
6. P. Rembo, J. Biggert, V. Kubechek, J. Schwartz, A. Berstein, J. Dilet, R. Bernstein, K. Shtalkopff. Laboratory investigations of lightning discharge initiated by laser // the Optic Magazine, 1999, V. 66, #3, P. 30-35;
7. Yu. P. Rizer Laser spark and distribution of discharges. Moscow: Science (Nauka), 1974. P.308
8. A. A Mak, N. A. Solovyev. Introduction into physics of high temperature laser plasma. Leningrad: Publishing House of Leningrad University. 1991. P. 152;
9. S. E. Yemelin, A. L. Priozersky, V.S. Semenov, G. E. Skvortsov Characteristics of dynamic state propagation in stream of capillary discharge. // Letters in the Magazine of Technical Physics. 1997, V.23, I.19, P. 54-59;
10. R. F. Avramenko, A. Yu. Gridin, A. I. Klimov, V. I. Nikolaeva. Experimental investigation of power-consuming compact plasma formations. // Collected articles Ball lightning in laboratory. Moscow: Chemistry. 1994;
11. G. D. Shabanov, O. M. Zherebtsov, Experimental modeling of ball lightning analogue. // Theses of 10th Russian conference on ball lightning. Moscow (Dagomys). 2002. P. 17-20.

New Energy Technologies

PRINT EDITION ADVERTIZING RATES

Full Page Ad
B/W - $400
Color - $800

Half Page Ad
B/W - $200
Color - $400

Quarter Page Ad
B/W -$100
Color- $200

New Energy Technologies is distributed in the leading Scientific Laboratories in Russia and has subscribers in more than 20 countries all over the world

To advertise contact us:
Faraday Labs Ltd, Phone/fax: 7-812-380-3844
http://www.faraday.ru, office@faraday.ru

Free Energy

Alexander V. Frolov, Russia
Faraday Laboratories Ltd
http://www.faraday.ru, tel/fax 7-812-3803844

In general the conception was formulated in 1995 and today we can say that the concept of physical vacuum, which is a new source of energy, finds more and more supporters. The fundamental works about nature of "zero point energy" are published, for example, Andrew D. Sakharov [1], Hal T. Puthoff [2] and many other interesting works. It is clear now that space or "physical vacuum" has its internal structure, therefore it can be used as a new source of energy if we organize the process of changing of this structure. Moreover, according to some theories, the existence of reality (i.e. the World of some certain parameters of space and time) is defined by the function of probability of energy density. For example, Dr. Nassikas, Greece [3], proves that it is not possible to consider the space without its energy, and there is no space without energy. The internal structure of reality is coexistence of two forms: gravitational energy and electromagnetic energy. Any local increase of the first one should produce decrease of the second one, so the sum amount of the change is zero.

In some other articles about energy transformation processes (gravitational form of energy into electromagnetic heat radiation of mass, for example) we can find that both increase of entropy and the inverse processes (decrease of entropy) are possible and in this case the electromagnetic energy can be converted into the gravitational form, that is shown in details in the works of Nobel laureate I. Prigozhin, Order and Haos. Man's new dialog with Nature, London, 1984. Since **the direction of time** (the time course) **and direction of the entropy change** (increase or decrease) are related notions then free energy technologies are considered by Kozyrev as methods of practical application of natural time course [4] that is presented in aether-dynamics by Frolov as aether flow of some density and this aether density determines the time rate as hardness of cause-effect connections for any process, and also for the process of existence of matter in space-time of this aether density.

According to this theory, in any point of space it is possible to get power by means of energy transformations without any consumption of mass-fuel. We can say also that in this case some change of the energy density of space should be detected.

Let's clarify some determinations:

1. The Potential (lat. Potentia that means "force"). In physics this is scalar parameter, it's gradient expresses the intensity of field of a certain force. In common sense, the potential is possibilities, which exist for execution of some task, for completion of some work.
2. The Work is a quantitative parameter of energy transformations. The transformation means here change of form.
3. The Energy (Greek "Energie" means "action, activity") is a quantitative characteristic for different forms of motion.
4. The Power is amount of work per unit of time.

According to the given determinations by The Soviet Encyclopedic Dictionary, edition of 1988, Moscow, **the fact of presence of potential (scalar) field, for example, electrical or gravitational filed, is real possibility to produce some work if we can organize change of energy forms. Let's note that power source is not required to keep the potential field in force. It is free.**

...the potential (scalar) field can produce real work!

The example of this work, which is produced by the field: body falls in gravitational field and when it strikes on the ground then some part of its potential energy is transformed to heat, that is the work as transformation of energy forms. So, there is a conclusion: the potential (scalar) field can produce real work! But we have considered only a half of cycle and in classical case in the second half of the cycle it will be necessary to produce the same work against the field to raise the body to the initial point.

Let's formulate the task to produce the work periodically and to get the power in load from this process. Usual mistake is to accept the particular case (the same body returns back in the same field) as a single possible case. But in special case changes of the system are possible, for example, the field intensity is not a constant but some variable value (alternating or pulsing), or the body changes its own parameters. In this case in each of half-cycle of the process the field can produce real positive work to accelerate the body.

The main technological solutions are obvious: **it is necessary to create gradient of field in space** (full or partial screening of trajectory of the body, which is moving in the field) **or gradient of field in time** (pulsing mode of field). This is quite easy for electric and magnetic fields, but for system, which uses gravitational field to produce the work, we can assume changes of parameters of the body only.

It is possible to understand why professional physicists dislike the question about possibility to use scalar (potential) fields to produce useful work in a load since they think about The Law of Energy Conservation. To find mutual understanding it is necessary "to upgrade" this Law for 4-dimentsional case of real physical system. It is necessary to introduce **the notion of the structure of space-time** where the considered process is created. In other words, for real practical engineering tasks it is necessary to consider **space-time of the real process** but not an abstract space-time. Usually the energy density of technical processes is small in comparison with natural energy density of space (aether density), which is a result of real astrophysical processes, i.e. motion of planet, star, galaxy. If we discover structure of this real rhythm of the space-time of our planet then we'll be able to design it in our technical devices to use aether-dynamics as theoretical basis.

So, in orthodox physics there is the unchallengeable formulation: the work of potential field on closed trajectory of motion is equal to zero. Yes, it is right for only case: if one part of work is positive (the acceleration) and another part is negative (the deceleration). But one part of the trajectory or a part of periodic process with the negative work can be excluded by different methods: by means of spatial superposition, pulsing mode and change of interaction polarity or by the screening of electromagnetic interaction.

For example, minor changes in vacuum tube design (the grid is located under cathode) allows increasing the kinetic energy of electrons and to increase emission current by means of the potential on the grid only. Also vector potential of magnetic field or gravitational potential can be used by similar way.

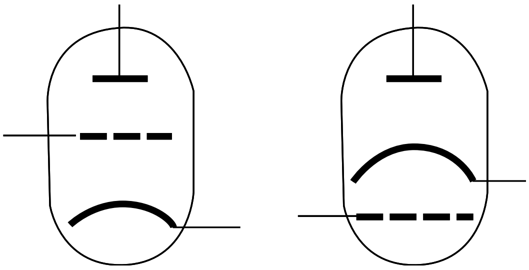

Fig. 1

Besides this way it is possible to use pulsed mode and switch-off the primary energy source before the emission electrons will reach the anode, Fig.2. In this case there is not the conductivity current between anode and cathode and the primary source is not discharging during its work.

Therefore, it is not a news that potential field can produce real work. In any textbook there are examples of positive and negative half-cycles that result to zero net work. But for the case of consequent execution of two processes the parameters of one of the processes can be changed, that **the work, which is produced on the closed trajectory can be formed as two or more parts of positive work.** Important aspect is following: it is necessary to determine the point (moment) of change the sign of the interaction and at that point (moment) to change the parameters of the process accordingly.

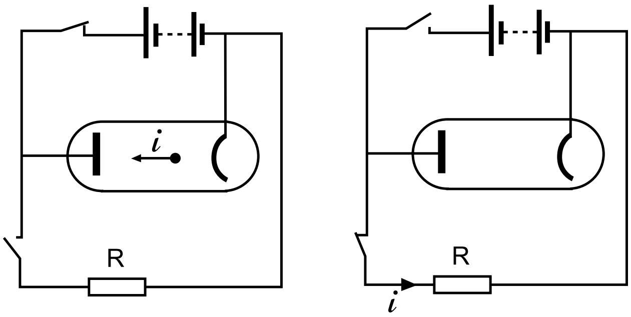

Fig. 2

The good sense requires the answer: what is the source of the power to produce this work if this proposed concept is right? Let's notice that before the considered examples, the existence **of power interactions of potential fields** must make the same question. For example, how the ordinary permanent magnet (i.e. its vector potential) holds a piece of metal making work against force of gravity without any fuel? How the potential field can move and accelerate ions, i.e. kinetic energy of ions can be increased without any fuel? What provides the forces of elasticity? Let's try to find answers from consideration of inner structure of electric and magnetic field, especially let's develop **our understanding of the notion of gradient, which describe properties of space-time design in area of this field.**

The concept of potential as bi-directional flow of photons and anti-photons was proposed by English mathematician E. Whittaker, and then it was developed by Dr. Thomas E. Bearden [5]. In this concept, the generalization of the Third Newton's law looks as the requirement of complementary pair to the process of the electromagnetic radiation. Since "process" means a change of information in time, this paired anti-process is reversed in time. Of course, it is development of process to its own future, but from our point of view, it goes from future to past. According to generalized Third Newton's law the radiation of photon is paired with anti-photon. In this case, the internal space-time structure of electric potential field is formed by two contrary flows of energy: photons spread from charge source and the anti-photons "inflow" into the charged mass. Let's notice that this concept defines the relationship of charge and mass. The charge without mass does not have any sense.

The flow of energy outgoing from charged mass is responsible for all phenomena of radiation. Incoming flow of energy is responsible for the gravitational interaction. Therefore, the notions "radiation" and "gravitation" can be considered as complementary pair; that is considered in details in concept of Josef Hasslberger [6].

...it is impossible to get "something" from "nothing"

There is a quite correct question: since "something" can not be created from "nothing" then what and where will be changing if we create the process of free power dissipations in local area of our space-time by means of asymmetrical potential interaction?

Let's try to present a "process" as some change of "information" in time and from this point we'll get the conclusion about results of this energy disbalance of the natural internal structure of potential field: in this case the entropy process in direct time (heat dissipation in load) should be balanced with equal anti-entropy process in reversed time. In other words, the disbalance is change of the space-time curvature and changes in the time rate.

So, it is impossible to get "something" from "nothing" but if we have understanding of the balance (process in time and anti-process in reversed time) then in the Universe the 4-dimentsional balance is not violated. It is assumed here that it is impossible to create one process but it is possible to create two opposite balanced processes, and each of them produce real work in the load. In astrophysical consideration it was stated in 1964 by Academician Gustav Naan, Estonia, Tartu. He wrote [7]:

"In our real world we could extract any amount of energy from vacuum if some technology provides at the same time the extracting of the same amount of energy for anti-world. Total sum value of energy is equal to zero."

Let's note also that similar conception of "dynamical zero" was applied in ancient India mathematics, i.e. "zero" is not "nothing" but it is hidden possibilities (balance of forces).

One more interesting question appears: what is about reality of the particle of matter after we have claimed that its mass-parameters are parameters of certain process? With this idea the material world as a whole is certain process and "stability" is not a property of some object but a parameter of its existence, which is a process in space-time of some certain energy density and some certain structure. By analogy the vortex in liquid is a process but it is not a liquid.

Thereafter, as electron was presented by Shredinger as wave packet and by Whittaker as function of two scalar potentials, the old idea of instability or **dynamic structure of matter** has the possibility of experimental testing and development as a teleportation technology. This idea is technology of space-time engineering, i.e. creation of the space-time with some certain parameters where curvature (or rate of time) determined rate of existence of the matter in space. It is related with some energy density, which can be increased or decreased. Since by this way it is possible to consider any material particle of substance as certain process, which is balanced by corresponding time-reversed process, so there is no theoretical difficulties to develop technologies of chemical transmutation, materialization, dematerialization, teleportation etc.

So, the using of potential energy to create a process of dissipation of power in load does not violates the Law of Conservation, and this Law can be generalized: the total energy of four dimensional system is amount of energy of processes in time and energy of processes in reversed time, it is constant and it is equal to zero.

Now let's try to review some free energy projects (the systems to produce work without consumption of fuel). Russian Peter the Great had intention to visit Germany in 1725 to test the Orferius' device. There is a very old description of "perpetually rotating wheel", which was made in India by inventor Bhaskar in 1150! From that times the mind of inventors is developed to other modern systems but due to the efforts of scientific groups, which are interested to keep in force the ideas of primitive materialism, the great idea of free energy transformation is distorted to such extent that any person, who began to speak about free energy, had a chance to get the name of "mad". Why? The reason is common understanding of the "power", which is "some work per unit of time" and generally it can be presented only as result of some transformation of matter structure, i.e. disintegration, chemical reaction, nuclear decay, nuclear synthesis or any change of structure of a matter. In any case, a material (the firewood, oil products or nuclear fuel) is considered as the fuel, regardless of its transformation method.

Some concepts did not consider the fields (electromagnetic, gravitational and others) as a kind of matter. So hypotheses, and even successful experiments on transformation of "non-material" type of energy in energy of material object (into the work) were not taken into consideration. The physics is a study about measurable and tangible quantities. New measurement methods let us work with a new physical phenomenon. So, we can see that **real situation in alternative energy is changing due to experimenters efforts** but not from the great theoretical team.

Some time ago the electric energy was not considered as a material object, but gradually people have been able to refuse the gas pipes, which were real material source of power for the gas light lamp, in favor of electric wires for electrical illumination. In a short time, I think, it will be possible to refuse the wires and we'll consider the aether (physical vacuum) as unlimited energy source if we organize **the process of transformation of space-time parameters.** It will be necessary to drop the old conception of "primary source of power", which should be connected with the consumer by some method and to develop conception of free energy source in any place "on-demand".

Let's consider, what the term "free energy" means today. The energy in general sense means the "ability of body to produce some work". The energy of closed system is constant. Certain device can look like "perpetual mobile", but nobody will be shocked because the way of energy "inflow" is known. For example, solar panel battery is obtaining its power from external source of light. But in general case, 3-dimensional observer can see nothing similar to the inflow of energy into the system if multi-dimensional energy balance is not analyzed. In other words, "perpetual mobile" is a right name since for its description it is necessary to use notions of "Time", "Eternity", "Causality" and other categories, which are more related with philosophy and religion but not with modern physics.

The "perpetual mobile idea" is really value to be the purpose of serious scientific work. In the book "Great experiments in physics" published by "World", 1973, Professor G. Lipson wrote: "Joule was the person of a very practical kind of mind and he was enthusiast of the perpetual mobile idea". I hope that it is not necessary to explain here who was Joule. However, "practicality" of free energy generators is obvious only for energy customers, but not for energy producers, who created the centralized system of energy distribution. This is the main reason of absence of alternative fuel-less energy systems in the modern market.

Let's consider the existing classification of "perpetual mobiles":

1. "Perpetual mobile" of the first kind is a design, which can "create energy". It is disputable aspect and all patent offices refuse to consider the patent claim of such type. They answer that "energy can not be created or destroyed" but energy can be transformed from one type into another type.

2. "Perpetual mobile" of the second kind is the collector of environmental heat. It is not a "thermo-pair" which uses temperature difference to produce electricity, but it is the "heat pump". This type of devices works with negative entropy, or more exactly, with sintropy. The terminology still is not defined but we can say that in entropy systems the produced work is equivalent to the dissipated heat and the work, which is produced by sintropy systems is equivalent to some absorbed environment heat.

3. The "perpetual mobile" of the third kind is a demonstration of perpetual motion without friction. The analogue is electric current in superconductors. Development of this idea is creation of systems with negative friction. In electromagnetic systems it corresponds to the known cases of the "negative conductivity", i.e. some electrical circuits in this mode can generate output power.

Additionally, let's show that "creation of energy" is possible in theory, for example: two processes of equal power compensate each other $A + B = 0$. Let's assume that some technology creates another (inverse) process as: $0 = A + B$ i.e. two processes of some power in sum create zero result. In general principle, also there is possibility to use many processes balanced situation as $A + B + C + \ldots = 0$ and this interesting conception is known as theory of multipolarity.

Let's consider some quite real (to my mind) examples from the history of development of free energy technologies. Nikola Tesla's investigations are not well known to modern scientists and engineers. In his works on development of wire-less telecommunication Tesla used flat spiral coils as a secondary winding of transformer. The magnetic field of such coil is radial and it is placed in the plane of the coil. In 1995 the author of this article experimented with similar flat spiral coils. I have to claim that when such coil is operating as secondary winding and the solenoid is the primary winding of the transformer then we can measure asymmetrical mutual induction, i.e. connection of active load (lamp) to output circuit of the transformer is free from input power in the primary winding. It is the simplest example of technical realization of the asymmetrical cause-effect connection, described by Kozyrev in his theory of active properties of time.

... "creation of energy" is possible in theory...

Another Tesla's invention is his resonance transformer. Modern electrical engineering describes transformer with forced electrical oscillations and radio engineering considers operation of resonance systems mainly. Tesla put the question on transformation of power in resonance transformer and by this way the efficiency can be more than 100%. With high frequency currents and high power level Tesla used single-wire terminals as loads, i.e. the lamps and other single-wire receivers of power (motors) were powered from high frequency changing electric field. Such single-wire terminal does not consume any power from primary source because it uses the change of potential in point of connection to conductor (let's note also that for maximum efficiency this point of connection should be one of the maximums of standing wave).

The notion about "free vibrations" belongs to Tesla and this term describes sinusoidal oscillations in electric circuit that is created after short non-sinusoidal impulse due to real vibrations of the free electrons. Resonance mode of the free vibrations can be the real way to excess power output.

Original Tesla's approach to electrodynamics allowed him to build in 1934 fuel-less car with electromotor, which was powered from some 12 vacuum tubes generator of unknown design.

The standing waves of electric field that were observed by Tesla during thunderstorm brought him to conclusion about possibility of the system to power remote energy consumers from energy generator without any transmission lines and without radiation methods. He assumed that it is necessary to create special standing wave of electric potential (or variable in time electric potential field) around the generator, then the unlimited number of loads (lamps, motors) in area of this potential field can be powered if they are tuned in resonance with oscillations of the generator. Let's notice that in each receiver the power can't be more than power of this generator but it is possible to install many independent "receivers" without mutual interference.

The modern investigations on these problems sound as sensation, because Tesla's works are unknown for the modern generation of scientists. Of course, modern electronics components and the tools facilities allow to create real "miracles" in comparison with the past age experiments. For example, engineer Avramenko described his work on single-wire power transmission in Journal of Russian Physical Ideas, 1991, No.2, and in journal "Inventor and rationalizator" 1992, No. 5, 6. The light bulb (or ventilator) was used as load of the single-wire power transmission line. This line can be made of high resistance material, for example, tungsten, but the power can be transmitted without heating of wire! It is possible to say that in this experiment a wire does not transmit power from the generator to the load, but the wire is the conductor of information signal, which is created by polarizational current opened and described by M. Faraday.

It is not difficult to repeat the experiments with single-wire line: it is necessary to place two diodes on the end of a line, which is connected to secondary windings of high voltage transformer (I used television set high voltage unit) by such a way that different poles of the diodes are connected to the line, Fig.3.

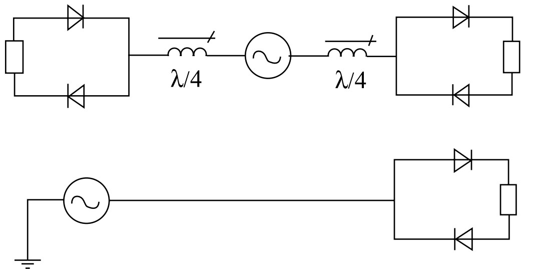

Fig.3

Two other poles of diodes create the source of potential difference (voltage), from what it is possible to charge the capacitor or to power the load (luminescent bulb). This diodes scheme is named as "Avramenko's diodes plug". In my experiments it was determined that high efficiency mode is resonance mode in real line and in this case the maximum change of potential is created in the point of connection of the "Avramenko's diodes plug". There is well known formulas to calculate this resonance as quarter-wave antenna vibrator system. Of course, high frequency and high potential allow creating high power in the load. In 2001-2003 New Energy Technologies magazine have published new experimental results on single-wire power transmission reported by research team from Moscow (Prof. Strebkov, Avramenko, Nekrasov and others). For example, they built and tested lines from 20 to 100 Kwtt power level!

... the electric potential field can create non-compensated force in the system and to produce a work without any power input.

One more well known researcher on the subject of free energy was Thomas Townsend Brown. He considered creation of reactionless propulsion force by means of electric forces only. Ionization is not considered here! According to his works, the electric potential field can create non-compensated force in the system and to produce a work without any power input. English patent by T. T. Brown #300,311 of August 15, 1927 describes the method to create propulsion force and power from electric energy source only. In this first patent it was claimed that in ordinary flat electric capacitor (two flat plates and dielectric between them), which is charged up to 50 kilovolts and more, Brown discovered propulsion force. This force moves the capacitor to positively charged plate direction. It is only one of his ideas and in his other patents of 1930 - 1965 Brown has described many new methods to create propulsion force and free power in load by means of electric field only (scalar potential field as a source!).

... the efficiency of "electric system can be million to one"

We have to exclude ideas on electrokinetic apparatus since it is just a reactive method and propulsion force is result of ionization flow. Especial case is Brown's **idea to create the asymmetry of electrostatic forces** by means of some special form of surface, USA patent #3187206 of June 1, 1965, application of May 9, 1956 (Fig 4).

The schemes and descriptions by the Brown's patents are undoubtedly the work of great practical value. As was mentioned by Brown, the efficiency of "electric system can be million to one" because the potential field can produce real work, for example, rotate some electro generator, but it does not change the primary source of field.

In 1927 T. Brown demonstrated the devices in Ohio, later he worked in France. His works in France were stopped and he returned in the USA. Detailed information on his works can be received from descriptions of patents http://www.soteria.com/brown and from the book “Electrograviticis Systems” by Thomas Valone, Integrity Research Institute, USA.

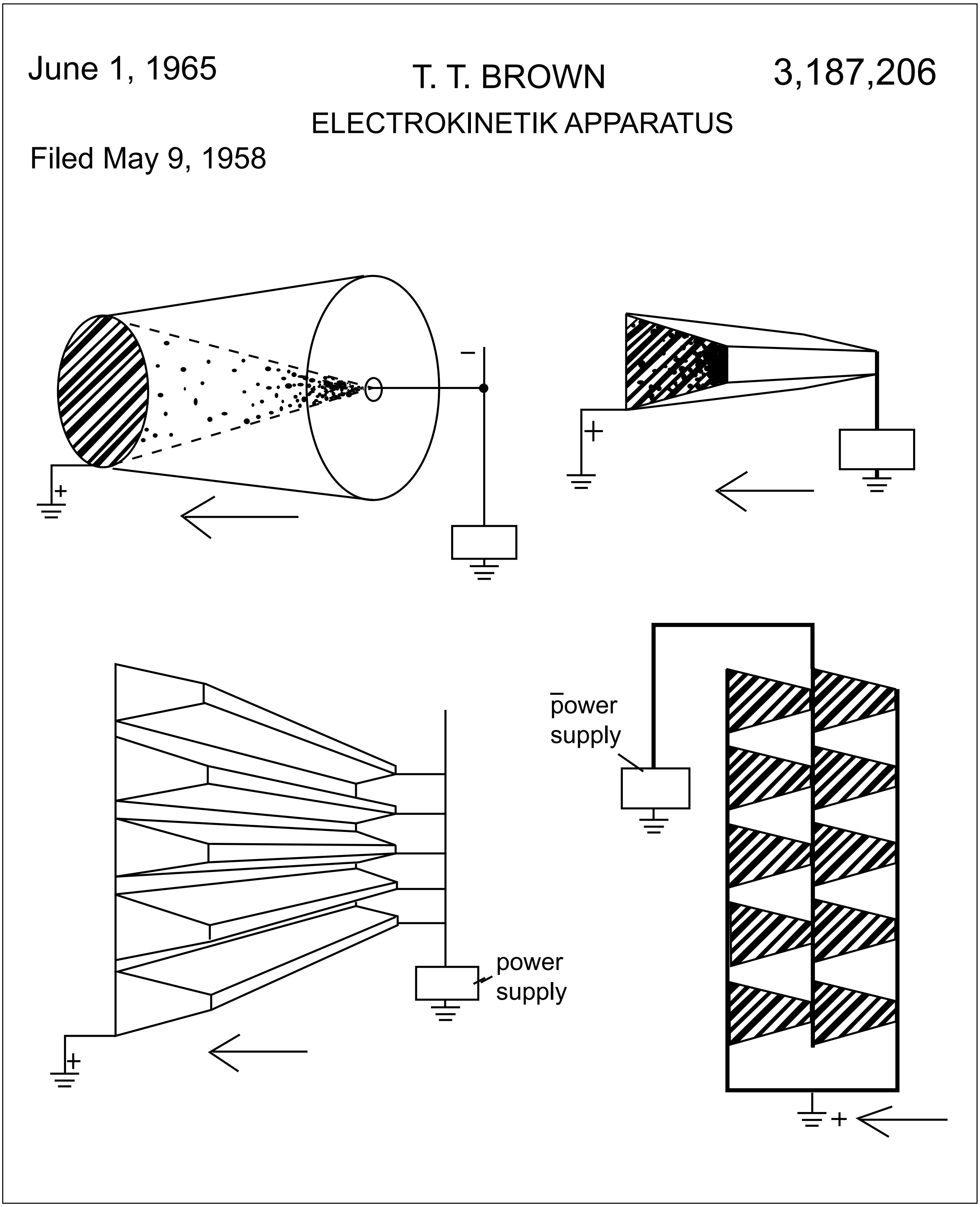
June 1, 1965
T. T. BROWN
3,187,206
ELECTROKINETIK APPARATUS
Filed May 9, 1958
−
+
+
power supply
power supply
+
+

Fig. 4

Electrokinetic apparatus by T.T. Brown

Besides, there is the simplest example of creation of real work by means of potential field that is also resulting from the Brown's works. Usually two plates of electric capacitor are equal to each other. However, if one of the plates is small and another one is large then the electric field between them is not isotropy field and in this case there is some **gradient of intensity of the field**. In such field any dielectrical object, for example a ball, will be polarized unevenly and due to this fact some force should make it to be moving aside, where the field has the greater intensity. I have to note for mathematicians that since the intensity is "gradient of potential" then gradient of intensity is the **gradient of the gradient**, i.e. it is the second derivative of potential. This idea is an example of the well-known rule: the change gives the new quality.

The example of creation of the propulsion force by means of potential field is also connected with "perpetuum mobile" task, since conductivity currents in the capacitor are very small and it almost does not require power consumption (after being charged once time) and the system can produce mechanical work permanently, in this case it is the work against the gravity force. In general case, if the design allows asymmetrical energy transformations, then the surplus output power and reactionless propulsion force can be created in this system.

Most likely future aerospace systems, which are based on the electrogravity, are the most perspective direction of free energy technologies. Why not for the power engineering? It is clear that now there are some existing fuel heating systems, heat stations and power industry to provide by some traditional way all current needs of the society and due to this fact the innovation of any free energy technology is related with hard competition. However, for commercial programs for cosmos the reactive rockets principles are not acceptable more. Each satellite during its operation should produce profit to compensate large expenses for the rocket-carrier. Only fuel-less propulsion systems can allow developing space commercial programs that are new infinite market. Let's note that the gravitational (reactionless propulsion) technologies are not related with cosmodrome and their cost can be quite acceptable to realize the project with private funds.

... future aerospace systems, which are based on the electrogravity, are the most perspective direction of free energy technologies.

There are some known examples of free energy systems.

In 1921 The Seattle Times, as well as Denver Post of August 8, 1921 published the articles about inventions of Alfred Earl Hubbard. His device included the central core with coil and eight remote coils placed around the central core. After primary impulse, the impulses in all coils were powered and rotating magnetic field was created in central coil. The power produced in the central coil was quite sufficient for self-excitation of the system and for producing of useful work in the load (motor). The boat and the car with electromotor, which was powered from the Hubbard's generator, were demonstrated.

In 1928, Lester Hendershot invented the electric generator of 300-Watts power. This device was designed of details used in radio-receiver to get oscillator (500 kilohertz) and non-inductive coil. Later, in 1970, William Cooper experimented with non-inductive bifilar coils. He used induction phenomenon in the case of zero magnetic component (two-wire winding or flat spire coil).

The Cooper's USA patent 3610971 of 1971 describes **the principle and the device to create power in secondary circuit without reaction on primary circuit**, as well as method to get reactionless propulsion force for aerospace application. Cooper also has found that specially designed coils can produce the field, which can not be screened and this field has some common parameters with the gravitational field.

The gravitation is considered by Cooper as a polarization of atoms in gravity field of planet. So he declares in description of his patent: "the electronic generator... of super high frequency creating the pulsing electric field of single polarity... acting in opposite direction to the Earth gravitational field... that to depolarized the atoms and to release them from the gravitation".

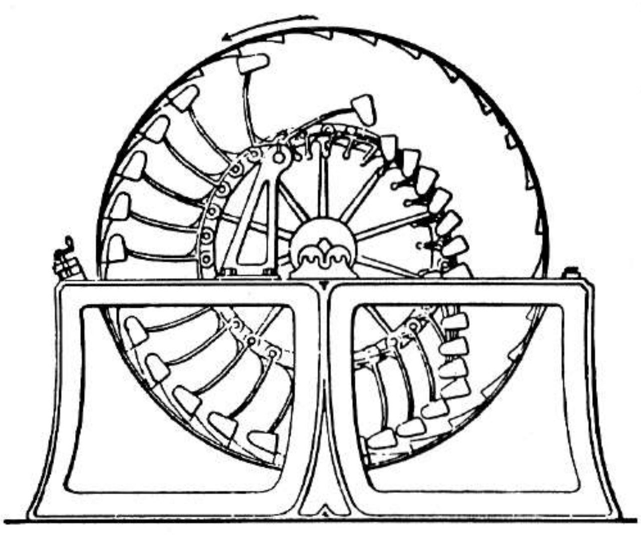

Fig.5

By the way, the gravitational field itself can be used to get a power. "Unbalanced wheel" is a well known design. The weights on the one side of the wheel, which is rotating in vertical plane, can be organized to be moving to axis, but on the another side the same weights can be organized to be shifted from the axis to the periphery of the wheel. By this way there is some constant shifted center of gravity in the system and it can be permanently rotating system. One of the inventors who built such wheel is Hugo R. Fraga, Havana, Cuba (The Perpetual Motion Mystery. R.A. Ford, Lindsay Publications Inc., Bradley, IL 60915, USA, 1987, see Fig. 5).

The earliest information about such systems is dated of 1150 and the name of inventor is Bhaskar, India. He used the tangential disposed pipes, which were half-filled with water. In France (1235-1240) William de Onnecure demonstrated self-rotation wheel with seven weights. In Italy (1438) Mariano de Jacopo has built the system made of eight bars disposed in the plane of rotation, and the bar can be fold up in the middle like elbow joint to provide rotation. One of the well-known and documented events of real demonstration of perpetual rotation wheel was mentioned in 1620. Edward Somerset (Second Marquis of Worcester), author of the book "Century of Inventions", 1963, has built and tested the wheel of about 4 meters in diameter, 14 weights of 25 kilograms each. The test of this machine was organized in London, in witness of King Karl, Grand Duke Hamilton and Grand Duke Richmond, and there are files in royal archives about this test. The descriptions of these and other systems are published in the book "Perpetual mobiles: past and present time" by Brodiansky, Moscow, Energoatomisdat, 1989.

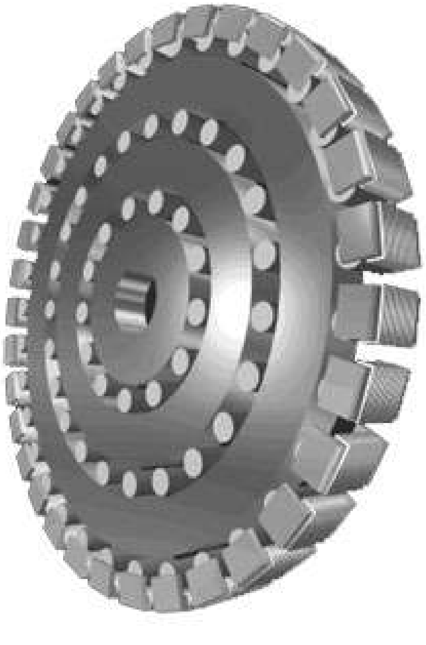

Fig.6

In different idea, which is known from Leonardo Da Vinchi drawings, lifting of water is produced by the helical "Archimedes' screw" of small diameter with large centrifugal acceleration, which reduces weight, but lowering of water was organized with the screw of other (large) diameter, so the force of weight is working to rotate this screw by the weight of the falling water. The paradox of these systems will be removed if to consider them as systems of variable topology, as it was done at the beginning of the present article. Practically, the **cycle is separated here into two processes** (the lifting and falling of mass) in the system with different parameters (the topology) for the first stage and the second stage of the cycle, but herewith **it is necessary to consider two different physical systems, not one and the same.** In that case, the classical theory can explain the work created by potential gravitational field of our planet as energy exchange between two different systems.

Other well known topic is research projects by John Searle, i.e. the "Searle's disks" (Fig. 6). It is necessary to note that the inventor mentioned in his articles anti-gravitational effect and free energy output also. The address is: John Searle, 13 Blackburn Lower Strand, Graham Park Estate, London MW9 5 NG, United Kingdom.

In several words we can say that rotor makes free electrons to be displaced to the peripheries of the system. With sufficient velocity it was mentioned that there is phosphorescence and ionization around the disk. It is possible to assume that main effect is based on well known Lorenz forces and understanding of the Poynting vector, which is circulating in this system. The ionization currents are closed through the space from periphery to the center, and self-rotation of the disk is provided by the classical Lorenz force, since the current interacts with magnetic field of the rollers.

In Russian experiments of 1992 Roshin and Godin built similar system of 7KWtt power output and they claimed that 100 kg axial force and areas of decreased temperature in environmental also were detected. Fig.7 demonstrates main parts of the system by Godin and Roshin.

It is necessary to note that similar anti-gravitational effects appearing for the case of over-unity operation were observed by different inventors independently. For example, in 1990 Floyd Sweet demonstrated his invention named as "vacuum triode amplifier" VTA. The barium magnets were pre-conditioned by special method to be used in special "trigger mode". This "bi-stable condition of magnet" provides possibility of transition from one direction of field to another direction due to the weak control signal, which was provided from external generator. It is known that if the material was pre-conditioned by the magnetic switching of 60 Hz frequency then its control signal must have the same 60 Hz frequency. A part of output power was closed to provide feedback and additionally some power can be used in output coil for the load. Tom Bearden studied the scheme of vacuum triode amplifier and confirmed that it demonstrates the work with negative energy. It means that the work in the load is connected with use of negative time. In this negative time, according to Bearden, the gravity is repulsing force. The experiments on VTA demonstrated that VTA decreased its weight according to level of the power, which is extracted from vacuum. Additionally we can

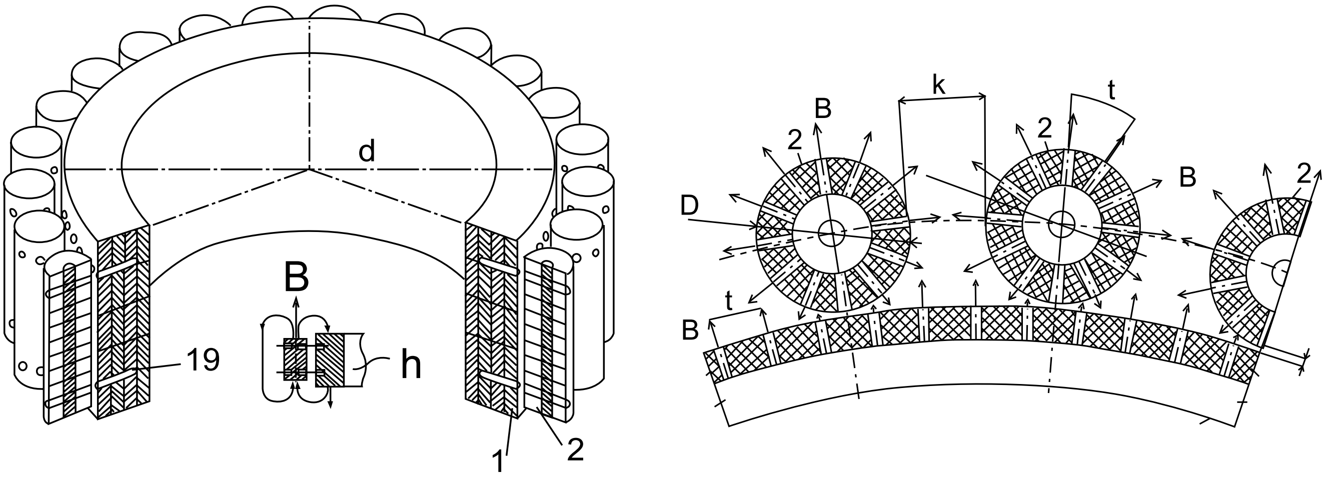

Fig.7

say that permanent magnets and coils of VTA were self-cooling during operation and the temperature difference was about 20 degrees in contrast with the environmental temperature.

One of VTA schemes includes two sets of magnets 4 x 6 x 1 inch, placed on two walls of frame. The attraction is created between them. Output and controlling coils are located between them. The axes of output coils are parallel to the magnetic field lines, but the axis of control coils are placed under 90 degrees angle. The secret of system is the "conditioning process", which "brings the magnets to the special condition". It is necessary to create great number of micro-cracks in the magnet due to re-orientations of magnetic domains. In such "half-magnet" the domains get the ability to be orientated in the same direction in a weak "control" magnetic field. In fact, they are not domains but large parts of magnet's material, separated by micro-cracks, that is to say acoustic domains. Many researchers repeat Floyd's works. Let's note that arc discharge of alternating current through magnet barium ceramics directly can provide the best results of the "conditioning" process. The coil of "conditioning" is not necessary in this case. The frequency of alternating current must be corresponded to the frequency of the control signal. Thereby, Sweet was creating **the bi-stable solid-state condition of magnetic substance.** The acoustic resonance is the reason of oscillations with the frequency according to the control weak magnetic field. Floyd Sweet died on July 5, 1995 at age of 83 years old. It is known, that his widow sent the archives to some great Automobile Corporation.

... the extraction of space energy is result of transformation of vacuum energy but from the other hand the energy density in this case determines so called "time rate"

It is interesting to note again that the magnets of VTA demonstrated self-cooling during operation, up to 20 degrees difference in comparison with the environmental temperature. It is one more example of relation between notion "density of time" and "energy". We can say that the extraction of space energy is result of transformation of vacuum energy but from the other hand the energy density in this case determines so called "time rate".

Kozyrev's articles [4] on causal mechanics theory and experiments describe possibility to use "time course for producing of useful work". Kozyrev introduced notion "density of time" and he demonstrated experimentally several methods how to change the density of time, which depends on irreversible processes intensity. Powerful "generator" of such processes is biosphere of our plane and it creates **season and daily changes of the density of time**. It is known that VTA power output also was variable in different time of day and night. It can be explained by Kozyrev. Next step in logical development of this idea is **to change notions from the "density of time" to "density of aether"** [10].

Relation between magnet phenomena and aether circulations was known from the beginning of the electrodynamics and now we can assume that VTA was real example of asymmetrical cause-effect connection, where the hardness of this connection depends on the density of aether. In this case the season and daily variation of the aether density are reason of VTA output power variations. So, we can make a conclusion: conception of **asymmetrical cause-effect connections** should be used as theoretical basis of all over-unity systems.

Generation of extra power in nonlinear materials (ferrites and dielectrics) was considered by Nikolay E. Zayev, Journal of Russian Physical Ideas, #1, 1991. Discovery was claimed as "Cooling of some dielectrics by changing electric field with generation of energy", Russia discovery #32-OT-10159, November 14, 1979; the inventions were also claimed as "Method of transformation of heat energy of dielectrics into electric energy", Russian patent claim #3601725/07 (084905), of June 4, 1983 and "Method of transformation of heat energy of ferrites into electric energy", Russian patent claim #3601726/25 (084904), of April 3, 1983. This theory is not about some transformation of space-time structure but in any case the practical application of conversion of environmental heat is very important because this work is real basis of free energy systems.

In one of Zaev's articles, which were published in Journal of Russian Physical Ideas, he quoted from K. Zialkovsky: "If the heat can be transmitted from cold body to hot body, then it must have enormous importance, and not only philosophical or scientific, but also practical importance. The Clausius postulate is not confirmed in this case. The gravitational force, as well as other reasons (number of the reasons is unknown) break this postulate... heat can be transmitted from cold body to hot body but only as the result of some exclusive conditions", published in Russian, "The Second beginning of thermodynamics", Kaluga, Russia, 1914. So, the inventor of free energy system should provide this "exclusive conditions" to organize collection and transformation of environmental heat in his free energy system.

Another well-known free energy system is Swiss electrostatic machine. In spiritual commune Methernitha, Linden in Switzerland, since 1980 several free energy devices have been generating total power of 750 kilowatts. From the technical point of view, these devices are modernized electrophore generator. Permanent magnets also included into design of these devices. The machine of 20 cm diameter produces about 200 watts, and a big machine has the disk of 3 meters and it produces about 30 kilowatts, Fig. 8.

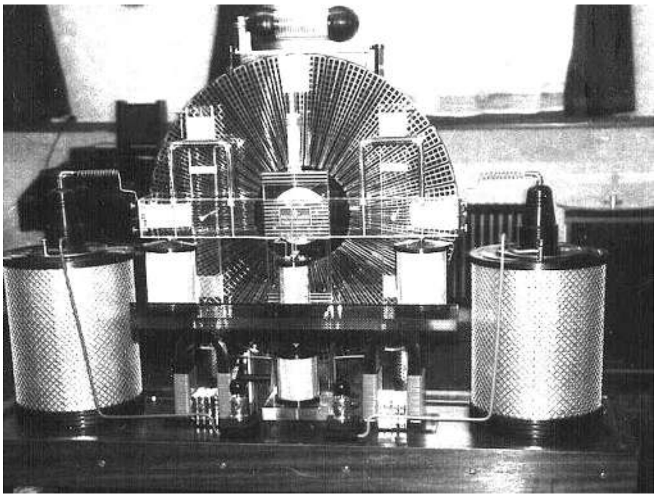

Fig.8

One of the modern technical decisions, which are very close to this Swiss machine is patent USA No. 4897592 by William Hide, January 30, 1990. This device is "system to generate power from electric field". It is one more example of work, which can be produced by potential field, in particularly by means of electric field that can be used as free source of power. The energy (the potential energy) is the possibility to produce work and the power is a work per unit of time, i.e. it is a process. The properly organized process, for example acceleration of rotor in Hide's machine, uses the potential field on the part of positive work (acceleration) and the author of the patent provides meta screen on the part of the trajectory, where the field decelerates the rotor.

... it is "perpetuum mobile" and it can not be patented even if it really works!

One more example is Reed's motor, which uses energy of permanent magnets. By the description of 1991 it is made of four disks (two immovable discs and two rotating discs), and eight magnets are placed on them. Howard Johnson used similar method, USA patent #4151431, Fig.9

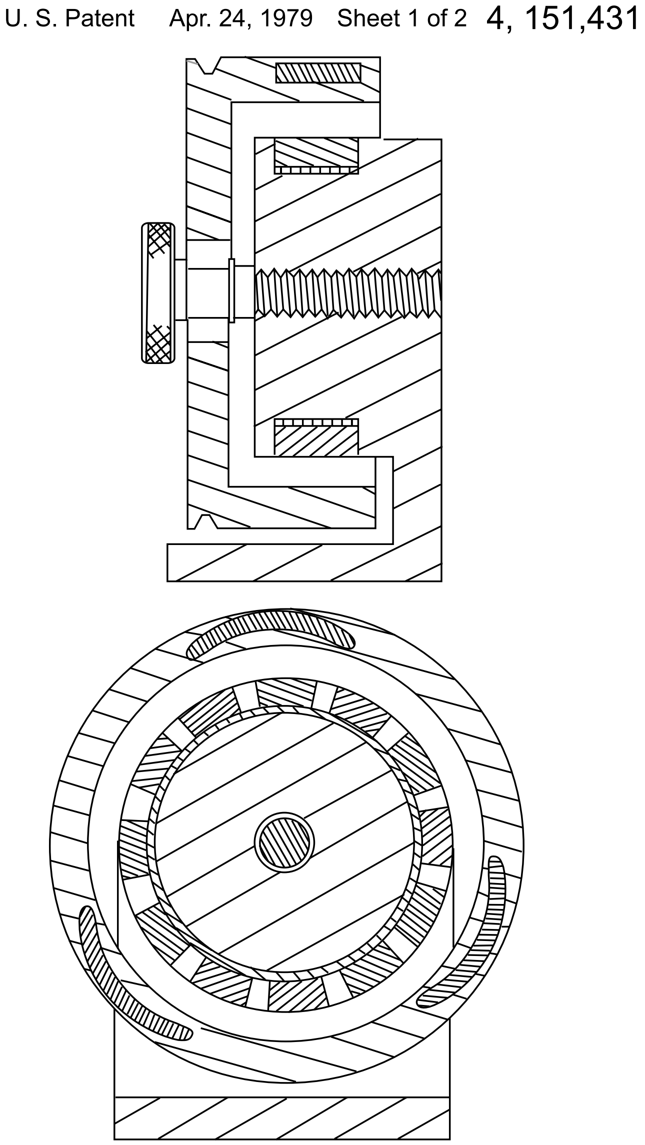

Fig.9

Journal "Science & Mechanics" of 1980 describes this invention. Hovard filled the application in 1973 but he has got the confirmation only in 1979! The reason of delay is very clear: it is "perpetuum mobile" and it can not be patented even if it really works! Johnson found clear and simple description of generation of power in his device and explained it as "extraction of power from inner energy of electron spin in ferromagnetic material".

From text of his description it follows that USA patent #4,151,431 of April 24, 1979, "Permanent magnet motor", author Howard R. Johnson, application #422,306 December 6, 1973 was developed from the prototype that is USA patent #4,074,153 of 1978, class of international classification H02K 41/00E:

"The invention is directed to the method of utilizing the unpaired electron spins in ferromagnetic and other materials as a source of magnetic fields for producing power without any electron flow as occurs in normal conductors, and to permanent magnet motors for utilizing this method to produce a power source. In the practice of the invention the unpaired electron spins occurring within permanent magnets are utilized to produce a motive power source solely through the superconducting characteristics of a permanent magnet and the magnetic flux created by the magnets are controlled and concentrated to orient the magnetic forces generated in such a manner to useful continuous work, such as the displacement of a rotor with respect to a stator. The timing and orientation of magnetic forces at the rotor and stator components produced by the permanent magnets to produce a motor is accomplished with the proper geometrical relationship of these components."

...conception of asymmetrical cause-effect connections should be used as theoretical basis of all over-unity systems.

It is reported that functioning Johnson's model produced about 5 kilowatts power free of any primary source of energy. Let's note that Johnson writes in his patent about permanent magnet as about system with "super conductive parameters". The currents of electrons in permanent magnet are manifestation of real superconductivity and for this case it is not necessary to provide the cooling for zero ohmic resistance. Moreover, the "resistance" must be negative since the magnet could save and renew its magnetized condition. Thereby, any permanent magnet is the example of perpetuum mobile of the third kind on the micro level. Also we can say the same about each atom.

Also, let's note that in general case motion (rotation) can be created due to the discussed above gradient of field, which in Johnson's device is result of asymmetry in "rotor-stator" system. By the similar way the gradient of velocity of airflow above and below wing creates the gradient of pressure and this fact produces great lifting power in airplanes.

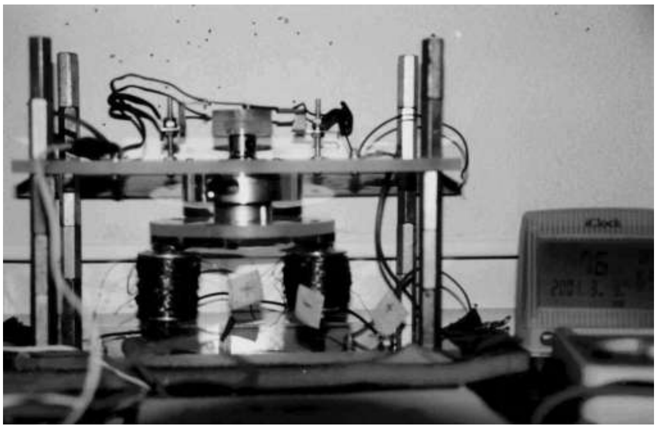

Fig.10

Well-known variant of permanent motor-generator is Adams motor, Fig.10. The rotor with radial orientated (the same pole outward) permanent magnets is rotating and creating inducted currents in stator coils, which are placed around rotor in the plane of rotation. From the point of traditional electrical engineering, any motor-generator without closed magnet flux is not high efficient device. However, exactly open magnetic flux of the Adams motor allows to take off power without deceleration of the rotor. We can assume that in this case phenomenon of electromagnetic induction is not important but in this design there is magnetic induction only, i.e. magnetization and demagnetization of cores in the field of the moving magnet. It is a perfect analogy with phenomena of electric induction that is "electrization by influence". Similar "magnetization by influence" differs from electromagnetic induction and secondary magnetic field in winding of generator is not related with deceleration of the rotor. Robert Adams works with Harold Aspden under patenting of their system. Adams is more than 70 years old but from our correspondence with him we can say that he is going to build demonstration version of 10 Kwtt generator.

...any permanent magnet is the example of perpetuum mobile of the third kind on the micro level.

There is also special name "alternators" for this class of devices, which use interruption of magnetic flux, for example it is the device by USA patent of John Echlin #4567407.

The experiments to investigate the alternator principles were organized also by the author of this article and it was demonstrated that ferrite core of the generator coil is self-cooling. The simplest experiment is based on electro motor, which rotates iron plate and it periodically appears in the gap between magnet and coil. But it is necessary to note that change of the flux in the coil area should be organized by such a way to decrease the flux of the field in the rapprochement half-cycle and to increase the flux for the moving off half-cycle. In this case the rotor is accelerated by the secondary field (back-EMF). From 1994 to 2003 several experiments were produced and main principles were claimed in the patent description, Fig.11.

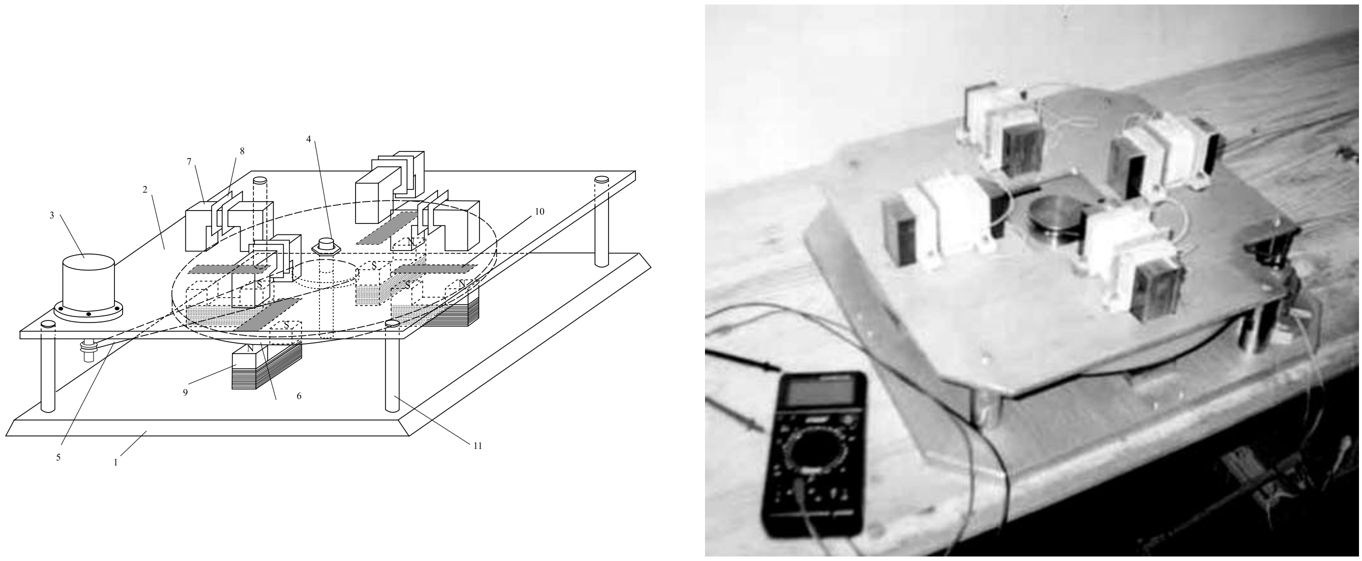

Fig.11

One more topic is "extraction of power from air" and Josef Swenson has conducted the series of simple experiments to develop it. The frequency of natural pulsations of electric field of planet is about 7.5 Hz and it is well known from Tesla's age. Swenson works with frequency 375 kilohertz and antenna of 10 meters. Please, contact for more details: Josef Swenson 423 North 15th Street, Moorhead, Minnesota 56560, USA. However, everybody remembers from the school story about simplest electrical experiments by Lomonosov and Rihman, who investigated arc discharge in gap between iron wire from a roof ("antenna") and ground wire. Let's include the resonance circuit "inductance-capacity" and diode rectifier to get some useful work in the load "from atmospheric electricity", Fig.12.

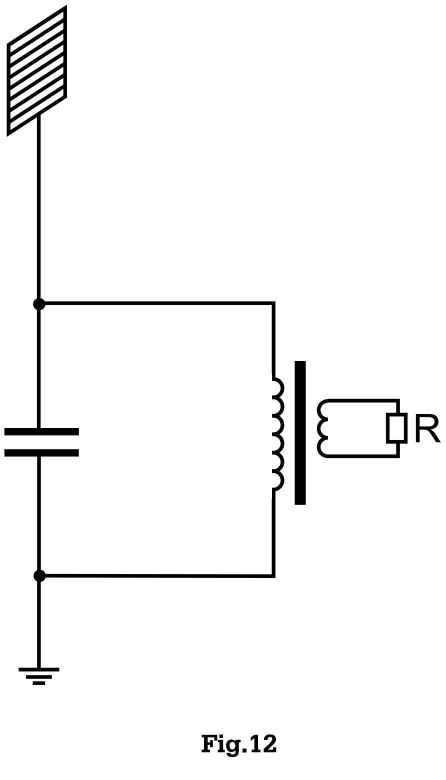

Fig.12

In 1900-1930 a lot of articles were published in technical press about Henry Moray. His demonstration systems produced more than 50 kilowatts free power output. It is known that Mr. Yakovlev (from USSR Foreign Department headed by Mr. Molotov) visited Morey in November of 1929 in New York to test his devices. The devices consist of capacitors, coils and special electronic-vacuum lamps.

In 1990 journal "Magnets", 2 (3) published article, which describe analogy between Moray's devices and Hubbard's coils, which can extract power by means of inner energy of nucleuses of materials if special resonance is created. In 1978 Cospray Research Institute has published the well-known book "The Sea of Energy" by T. N. Moray, in which theory of Moray is presented most completely.

RQM Raum-Quanten-Motoren Corporation (Schmiedgasse 48, CH-8640 Rapperswil, Switzerland, fax 41-55-2125209) offers for free energy devices of different power level: RQM 25 kilowatt and RQM 200 kilowatt. The principle of work is based on an invention by Oliver Crane and his theory. Web site http://www.rqm.ch. But I have to note that in present time they are developing capitalization of the company mainly (actives are above 650 million dollars) than innovation of the new technology.

One more historical example: in 1925-1945, Hans Koler demonstrated his free energy devices. The system was built in Germany and it produced 60 kilowatts of free power. The description includes six permanent magnets and coils, which were superposed in plane of hexagon form. Each magnet is core of the coil.

Another interesting topic is unipolar induction effect, which is well-known from Faraday. This effect creates electro motive force (EMF) in rotating disk if axial magnetic field is applied to the disk. One of the well-known practical developments in this area is device by Bruce de Palma. In 1991 he has published the results of his tests, from which it follows that deceleration of the rotor due to back EMF in the case of unipolar induction is less than in traditional electro generators. So output power of system can exceed the power, which is necessary to rotate the rotor.

Really, motion of electrons in magnetic field, which is perpendicular to plane of rotation, is the reason of the Lorenz force and this force acts in radial direction that can not be the reason of deceleration. Indian research on this topic is developing by P. Tewari. In New Zealand there is research group headed by Dr. Ashley Gray. In 1994 the leader of Japanese market MITI published report about progress in 40 KWtt unipolar generator, which uses superconductors for its electromagnets. The interest of Japan to alternative energy projects can be explained by the position of Japan on fuel market.

There is a well known rule: Demand is related with Proposals. It is easy to imagine the prospects of local introducing of free energy systems in one or several countries, if some producers of product will be able to exclude expenses on electricity and fuel from cost sales. Other countries of their own rich natural resources (for example, oil) will be in problematic position on the new international market, mainly due to the fact that their industry and transport are oriented to conversion and consumption of oil fuel that increase the cost of all products.

... increase of the spark gap produces more surplus power in the load of the circuit.

One more modern free device was invented by Wingate Lambertson, USA. In his device free electrons get the additional energy passing through the number of thin metal-ceramic composed layers. The units were designed by the authors and each unit can generate 1600 watts, and it is possible to connect them in parallel. The address of author: Dr. Wingate Lambertson, 216 83rd Street, Holmes Beach, Florida 34217, USA.

Especially the researches on free energy with plasma processes should be noted here. In 1980-1990 Alexander Chernetsky, Yury Galkin and others have published the results of experiments on creation of "self generated discharge" SGD. The electric arc was placed consecutively into secondary circuit of electromagnetic transformer and it produces real increase of power in load and reduction of consumption power in primary circuit of transformer. The author of the present article produced simplest experiments to investigate the arc (electric discharge) in electric circuits and possibility to create the mode of "negative resistance" in this circuit was confirmed. One of the effects was demonstrated in 1996 during the conferences "New Ideas in Natural Science", St.-Petersburg.

Adjusting parameters of arc (distance between two electrodes) it is possible to see that consumption current is decreasing until zero and then it can change its direction i.e. this system begins to generate the power. During similar experiment of 1971 by Dr. Chernetsky substation transformer in Moscow Aviation Institute was destroyed in result of strong "reversed current" impulse, which exceeded consumed power in 10 times more. According to Chernetsky's concept, the reason of this mode is well known phenomenon of plasma instability and **pinch-effect for great currents.** However, the author of this article tested device, which demonstrated similar effect (switching on the load in secondary circuit of transformer and in the presence of arc in this circuit, consumption power does not increase, but reduces) for small currents about 300 mA.

Since for pinch-effect it is necessary hundreds Amperes then it was offered another explanation: the surplus power in this circuit appears due to the acceleration of electrons in the gap between electrodes, i.e. particles of plasma are **accelerated by means of electric potential field between two electrodes**. It is noted during the experiments that increase of the spark gap produces more **surplus power in the load of the circuit.** To avoid mistakes the measurements of the consumed power were organized in DC (direct current) battery circuit and therefore there is no any reason to speak of phase shifts mistakes to try to explain skeptically this effect.

... waves of density of time are used by organisms for their vital activity.

Today theory and experiments on self-generating discharge are quite well developed to build free energy systems of any power scale. The reason of delay in its practical development is a complex problem: this work leaves the frames of classical physics. In his book "About physical nature of bio-energy and its simulation", Moscow, Publ. VZPI, 1989, Dr. Chernetsky considered the structure of biological fields and bio-energy processes in living organisms from the point of **longitudinal waves conception.** Self-generating discharge in the mode of negative resistance produces such longitudinal waves and they are self-sustaining (self-powered energetically) and it is considered as field of living object.

Really experimentators of the Chernetsky's group, who worked with SGD device detected influence of biologically active radiations and fields, which can be not screened by usual methods. It was reported by Chernetsky that parameters of this radiation can be selected to accelerate the development of plants and biomass or to suppress it. So, we should speak about artificial living biological system or biological form of energy for this class of free energy devices. Perhaps by the same plasma oscillation way all living organisms provide its vital activity, since long time ago it was known that metabolism and food can not provide enough energy for vital processes.

Nikolay A. Kozyrev also wrote about "reason of life" and he confirmed that waves of density of time are used by organisms for their vital activity. Between "waves of density of time" and "waves with longitudal component" there is a clear analogy. Kozyrev and Chernetsky experimentally demonstrated methods of creation of such waves.

...any free energy system should change the causality in surrounding space-time.

New Energy News magazine, March 1996, wrote about attempts to innovate for American aerospace stations special power source, which uses similar plasma technology: anomalous electric discharge. The USA patents #5416391 and #5449989 belong to Dr. Paulo Correa and Dr. Alexandra Correa, Canada. In their report on free energy Denver conference it was claimed that efficiency is about 483% .

Let's note one more free energy topic: high efficient electrolysis. Classical electrolysis as decomposition of electrolyte in electric field is a wonderful example of work, which is produced by field and it can be organized by such a way to be free from the primary energy source (battery). The traditional scheme uses closed circuit of current through electrolyte and the battery, so the battery is discharged during the operation with electrolytic cell. However any physics textbook confirms that **ions in electrolyte are moving due to electric field only, i.e. work to organize the displacement of ions and heat power, which is connected with this work, are produced by the potential field and expenses of the primary power are not required.**

The current through the battery, which is created in usual closed electric circuit destroys the primary difference of potentials in the battery but it is not some necessary condition. For correct organization of the experiment the products of electrolysis (gases) can be created almost free and their utilization (burning) can provide more heat power than input electric power.

Prof. Latchinov, the real member of Russian Physic-Chemical Society, who patented his method of electrolysis in 1888, mentioned that in some cases the electrolytic cell is freezing during its operation. This is an effect of the law of Conservation, which forces to return the surplus power by means of heat environmental energy. In the experiment by Latchinov the electrolytic cell can create gases of high pressure and consumed power is the same as for the case of low pressure gases. However, it is clear that high pressure gases can produce more work than low pressure gases do. This question was sensational problem in scientific societies of 1888 and it is not solved yet completely.

One of other variants of high efficient electrolysis was investigated by Igor Goriatchev, Russia. Instead of usual 3 Volts level of electrolysis, Goriatchev uses 0.2 Volt pulsing mode. He claimed ratio output/input as 1500% efficiency and he hopes to increase this value twice in more perfect design.

Professor Kanarev from Krasnodar wrote about his experiments on plasma electrolysis and has proved that in electrolysis of water output power can be more than input power.

Real example of "perpetuum mobile of the second kind" is invention by Russian engineer Albert Serogodsky (Moscow) and German engineer Bernard Sheffer (Berlin). They have patented new system for direct transformation of environmental heat into electricity, Germany patent #4244016. Retro-condensation of mixture of benzine and water is organized in closed system under temperature of 154 ^{0}C degrees. You can try to get more info from: Werkstatt fur Dezentrale Energleforschung, Pasewaldtstrasse 7, 14169 Berlin, Germany.

The fundamental theoretical researches on direct transformation of environmental heat to useful work were provided by Real Member of The Russian Physical Society Mr. Gennady N. Buynov, St.-Petersburg. The description of his project on "The Mono-thermal device" was published in journal "Russian Ideas" #2, 1992.

In 1995 the scientific journal of Russian Physical Society #1-6 published article "Perpetuum mobile of the second kind (paired gas-chemical cycle)". The author Gennady N. Buynov proposed to consider that **entropy function can be abortive**, i.e. it can be undetermined in some place if reversible chemical reactions are presented in the system.

Herewith, the circular integral of entropy is not a zero and in this case the heat function but not entropy function (according to the Gess law) becomes the function of condition.

Buynov offers to use four–oxide of nitrogen as working matter for this cycle. His works are excellent example of scientific enthusiasm, which (in combination with financial interest of the customers) could produce real results for Russia many years ago. Let's note that priority of this discovery belongs to Russia in spite of attempts of other scientists to claim and develop this idea now as their own work.

We can make a very interesting conclusion if we analyze history of so called “cold fusion” discovery. According to de-classified materials of 1960, priorities of Russia in this topic are obvious. In 1989 Pons and Fleshman have reported about results of their experiment. In 1995 Russian Journal “Inventor and Rationalizator”, #1 has published article about invention made by Ivan S. Filimonenko that was named in 1957 as “warm nuclear syntheses”. In 1957 he detected **extra power output in process of heavy water electrolysis** and he mentioned several applied aspects, for example, propulsion force and possibility to reduce radioactivity by means of this process. In 1960 Kurchatov, Korolev and Zhukov have supported the ideas of the author. Russian Government adopted secret resolution on this work:

1. Investigation of energy generation.
2. Development of propulsion without reactive mass flow.
3. Research on protection from radioactivity.

But in next several years this work was suppressed by people of nuclear power plant team. The only system of such type named as TOPAZ was innovated for Russian aerospace systems. World-wide innovation of this technology is a real way to introduce “warm reactors of syntheses” and it is not necessary to wait for results of high-cost “Tokomak” project and other thermonuclear researches. Let’s note that secondary effects (propulsion force and influence on radioactivity) are possible due to using of “free energy” aspect if output power is result of change of space-time parameters.

In 1994 Journal “Russian Ideas”, #1-6, it was published an interesting document “Conclusion of The Moscow City Council Commission on the question about Ivan Filimonenko’s discovery and its development”. In this document it was recognized that it is vitally necessary to renew this works. But in 2003 we still have nothing new on this topic. Why? It can be assumed that the Problem of innovation for this technology is possible military application of the methods since influence on radioactivity (for example, remote reduction of radioactivity of some object) is area of interests of the Defense Department. The fact that energy generators by Filimonenko can be used for quick restoration of the present ecological balance is not so important in this case. The same conclusion is about propulsion method, which was proposed by Filimonenko. Mr. Korolev knew about this method; however present space programs are still based on rockets and reactive principles, and anti-gravity flying machines we can see only in fantastic movies.

At the same time, development of commercial cold fusion projects was started in some countries, for example: Patterson Power Cell is introduced in Texas, USA (Clean Energy Technologies Inc., Dallas, Texas, fax 214-458-7690). More than thirty patents were owned by ENECO Corporation, which is collecting the main technological solutions in this area. The production of electrolytic thermal cells was started by Nova Resources Group., Inc., Colorado.

In August of 1995 Atomic Energy of Canada, Ltd. Company, which is member of The Planetary Association for Clean Energy, has published the review on modern methods of conversion of nucleus wastes and deactivation. Two new technologies were offered for introduction: contact processing by “Brown’s gas” and remote processing by scalar (torsion) fields. Let’s note that the technology proposed by Canadians and Filimonenko’s technology demonstrate the effect of influence on the rates of radioactive decay.

These examples are only small part of real situation. Main references on publications are foreign and it can lead to wrong conclusion that Russia is delayed in this direction of new technologies development. In fact, Russia has more talented inventors and researchers than any other country. However, condition for work, patenting and publications of ideas are not the same, and usually Russian technologies cannot reach the level of international market. This problem depends only on real and official state policy with respect to inventors and scientists. In real life financial support of scientific institutes is mainly subsidy for management of the institutes, but not for science. The inventions and discoveries always were made by a certain real person, but not by Institute or some scientific team.

In Russia of last age and in the other world on the whole the institutes and laboratories have been creating for a new scientific problems, discoveries or new directions in science. It was necessary to claim about discovery and provide priority for own country to get official support. Main schools of thought appear by this natural way. By the same natural way the necessary in existing of some scientific institute can be removed when the idea grows up to serial production stage. If there are no any fresh ideas in this school then the institute should be transformed in design office of the production plant on this topic. It is almost impossible to create a new research institute in modern Russia, so really new ideas (if they are not ranged in frames of some existing scientific directions) can not be developed and they can not create a new school of thought in Russia.

... it is possible to make a conclusion about grandiose misinformation of society...

People (carriers of the new ideas) have to leave Russia to realize their sensational ideas on free energy and antigravitation topics. Why they do not work with Russian Academy of Sciences? It is the rhetorical question. Usually after many years of correspondence with patent office or with official science bodies the authors can find that their ideas are published in "serious scientific magazines" but without any respect to their names...

Attachment 1 presents some information about patented technologies. When we study old and modern patent documents, it is possible to make a conclusion about **grandiose misinformation of society**, which leads to appearance of two different Worlds, separated by the level of knowledge: evident and hidden knowledge. The achievements of the second hidden world could change our planet, give a chance to remove all energy and ecological problems. Besides, we have discussed that some free energy systems (for example, self generated plasma discharge) have also the medical-biological aspects. This "influence" from free energy system is related with negative entropy processes in area of the operation of the system so some components of biological systems can be changed to less entropy. The design of free energy system defines the type of this influence. It was noted before that operation of any free energy system should be considered in high topology space-time as multidimensional system to see what the reason of the effect is.

So, since the rate of time course is determined by Nikolay A. Kozyrev as rate of cause-effect transformations, then any free energy system should change the causality in surrounding space-time. In this case we can assume that it creates quantum physics effects on the macro level. It is interesting to assume Heisenberg's uncertainty, tunnel effect and manifestation of wave characteristics (up to diffraction of objects) for macro level.

Some experimental data by Kozyrev is related with quantum effects, which were detected in his experiments with rotating and vibrating gyroscopes. This technology is the basis to design real teleportation systems, which will change parameters of space-time to allow changes of position of some material object from one point of space to another point of space (without transference on the distance between these points) by means of combining these points in the same place of space for some time.

The process of study of new technologies in alternative power industry and gravitation is developing actively in the world. Besides secret programs and institutes, it is possible to make a conclusion that activity of this work in countries of limited fuel resources is more efficient. Now Russia has rich **natural materials and oil resources but in the short time this advantage will not play any role in development of economy.**

Industrial and defense power of any country will depend on free energy technologies, knowledge on the biologically active energy and reactionless propulsion methods.

The new technology creates not only a new technical systems but also **new economical systems**. It is not physics but policy... Development of great Russian territories with industry of free energy technologies for home and commercial application is the way to change world economical balance in favor of Russia. It is difficult to say now what financial-economical group of modern Russia is most interested in development of this direction of science. The purpose of any financial-economical group is power on the market of energy resources, but when the free energy technologies are developing, then people and industrial or agricultural producers will be more independent from centralized system of energy and oil distribution, therefore, they will be more independent from influences of central authorities.

From the point of view of serious business, there is only one real argument in favor of developments of any really new technology: it should lead to increase of profit and expansion of the market. In this sense, it is possible to compare the new energy technologies with beginning of the steam machine age or with appearance of electric machines and illumination. This means the super profits and serious fight with competitors. To develop this way any efforts of scientists are not sufficient work. It is necessary to join them with the efforts of large business structures, which are interested to create new market of energy and power engineering in parallel to present monopoly fuel-energy market or to develop commercial exploration of space by means of new propulsion principles. In particular, telecommunication space satellites and projects on colonization of space could be the nearest commercial directions of the new market.

References

1. Sakharov A.D. "Quantum Fluctuations of Vacuum in Curved Space and Theory of Gravitation", Reports of Academy of Sciences USSR, vol. 12, 1968, page 1040.
2. Puthoff H.T., Ph. Review E, vol.48, num.2, p.1562-1565, "Extracting energy and heart from the vacuum"; Physical Review A, vol. 39, num. 5, "Gravity as a zero-point-fluctuation force".
3. Nassikas A.A., The Hypothesis of the Unified Field and the Principle of its Dual Interpretation, III International Conference "Problems of Space, Time, Gravitation", St.-Petersburg, 1994.
4. N.A. Kozyrev, Selected works, published by Leningrad State University, 1991, p.202, in Russian.
5. Gravitobiology, by T.E. Bearden, 1991, Tesla Book Company, P.O. Box 121873 Chula Vista, CA 91912, USA. Library of Congress Catalog Card Number 86-50553.
6. Hasselberg Josef, New Beginning of Thermodynamics, http://hasslberger.com, Hasslberger, Josef "The inventor and society" in Raum&Zeit #4, October 1989, Hasslberger, Josef "Vortex, the natural movement" in EXPLORE, No.5, Vol.3, 1992.

7. Gustav Naan, Symmetrcial Universe, Tarty Astronomical Observatory, Vol. XXXIV, 1964.
8. New Results of Development and Testing of Single-Wire Electric Power System, Prof. Dmitry S. Strebkov, Eng. Stanislav V. Avramenko, Dr. Aleksey I. Nekrasov, Eng. Oleg A. Roschin, New Energy Technologies, #5(8), 2002.
9. An Experimental investigation of the Physical Effects In a Dynamic Magnetic System. V.V. Roschin and S.M. Godin, New Energy Technologies, #1, 2001.
10. Practical application of time rate control theory. Alexander V. Frolov, New Energy Technologies, #3, 2001.

Attachment 1

Methods and devices to create power without external sources and reactionless motion systems

3913004 October 14, 1975, Method and equipment to increase electric power, Robert Aleksander.
4975608 December 4, 1990, Motor with switchable magnetic resistance, Harold Aspden.
5288336 Converters of heat into electricity, Harold Aspden, see also patents 5,065,085 and 5,101,632
4622510 November 11, 1986, Parametric electric machine, Ferdinand Kap.
2912244 1959, Gravitational System, Ottis Karr.
4006401 February 1, 1977, Electromagnetic generator, V. Rivas
3811058, 3879622 Motors with permanent magnets.
1835721 December 8, 1931, Motor with permanent magnets, A. Povel.
1963213 June 19, 1934, Magnetic motor, G. Pose.
1859643 May 24, 1932, Motor with permanent magnets, G.L. Worsington.
1859764 May 24, 1932, Magnetic device, G. Baugon.
2982261 Mac Klintok's Air motor.
4595843 June 17, 1986, Transformer of rotation magnetic flux, Robert Del Vechio.
4567407 January 28, 1986, Motor - alternator, John Eklin.
3368141 January 6, 1968, Transformer in combination with permanent magnets, K. Garon.
3890548 June 17, 1975, Motor with pulsing capacitor discharge, Edwin Gray.
4595852 June 17, 1986, Electrostatic generator, Robert Gandlach.
4831299 May 16, 1989, Homopolar generator of alternating current, Enakishy Khasaka.
4249096 February 3, 1981, Electric dynamo, Barbara Nikoks.
3610971 October 5, 1971, Electromotive generator of electric field, Williams Couper.
4897592 January 30, 1990, System producing power from energy of electrostatic field, Williams Hyde.
4151431 April 24, 1979, Motor with permanent magnets, Hovard Johnson.
4806834 February 21, 1989, Electric circuit of inductive conductors, transformers and motors, Erl Kening.
3374376 March 19, 1968, Electric generator, Raymond Kromry.
3977191 August 31, 1976, Power source, Robert Brett.
3670494, Method of converting of atomic energy in kinetic energy.
4709323 November 24, 1987, Converter of parallel resonance, Charles Lien.
5146395 September 8, 1992, Power source using two accumulating circuits, Richard Mac Kee.
4210859 June 1, 1980, Inductive device with two orthogonal windings, Paul Meretsky.
4500827 February 19, 1985, Linear electric generator, Thomas Merit.
4904926 February 27, 1990, Electric generator of magnetic motion, Mario Patsishinsky.
4945273 July 1990, High effective electric machine, Josef Pinkertone.
4883977 November 28, 1989, Converter of magnetic power, Dennis Regan.
4077001 Electromagnetic converter, Frank Richardson.
5018180 May 21, 1991, Conversion of energy, Kennet Shoulders.
4652771 March 24, 1987, Transformer, Theodore Speach.
477 2816 September 20, 1988, Conversion of energy, Jefry Spens.
4748311 May 31, 1988, Inverter, Fridrikh-Verner Thomas.
International patent H02K 31/00, 39/00 dated June 24 1982, Closed part of unipolar machine, Adam Trombly.
4687947 August 18, 1987, Electric circuit for conserving of power, Melvin Kobb.
4772775 September 20, 1988, Generation of plasma flux in electric arc, Sam Lich.
The USA patents 5416391 and 5449989, Paulo Correa.
4432098 and 4429280, Transmission of information by means of magnetic vector potential, Raynolds Gelinas.
Great Britain, #547668, January 30 (September 7) 1942, Motor with permanent magnets, Stenly Hichkok.
Great Britain, application # 2282708A, Motor with permanent magnets, Robert Adams, Harold Aspden.
4394230 USA patent, July 19, 1983, Henry K. Pukharich.
2251775 Great Britain patent, April 20, 1994, Thermoelectric conversion, Harold Aspden.
5288336 USA patent, Thermoelectric conversion, Harold Aspden.

The USA patents on electrogravitation

1363037 Goddard, December 21, 1920;
2004352 Simon, June 11, 1935;

2210918 Karlovitz, August 13, 1940;
2588427 Stringfield, March 11, 1952;
2231877 Bennet, 18 February, 1941;
2279586 Bennet, 14 April, 1942;
2305500 Slayter, December 15, 1942.
English patent 300,311 August 15, 1927, Townsend Braun.
French patent 1003484 November 1951, Electrogravitation;
3187206 June 1, 1965, Electrokinetic equipment, Townsend Brown.
3022430 February 20, 1962, Electrokinetic generator, Townsend Brown.
3018394 January 23, 1962, Electrokinetic converter, Townsend Brown.
2949550 August 16, 1960, Electrokinetic equipment, Townsend Brown.
1974483 September 25, 1934, Electrostatic motor, Townsend Brown.

Attachment 2

Organizations

- Faraday Lab Ltd, Lev Tolstoy Str. 7, St. Petersburg, 197376, Russia.
- The Russian Physics Society, 141002, Moscow reg., Mytishy, B. Sharapovskaya str. 3. Fax 095-2926511
- Academy for Future Sciences, P.O. Box FE, Los Gatos, CA 95031, USA.
- AERI, Advanced Energy Research Institute, 14 Devonshire Mews West, London W1N 1Fp, Great Britain.
- ADAS, Association of Distinguished American Scientists, P.O. Box 1472, Huntsville, AL 35807, USA.
- Borderland Sciences Research Foundation, P.O. Box 429, Garberville, CA 95440-0429, USA.
- Center for Action, P.O. Box 472, HCR 31, Sandy Valley, NT 89019, USA.
- Electrodynamics Gravity, Inc., 35 W. Tallmadge Ave., Akron, Ohio 44310, USA.
- Fusion Information Center, P.O. Box 58639, Salt Lake City, Utah 84158-0369, USA
- Gravity Power Research Association, 36 Mountain Road, Burlington, MA 01803, USA.
- GRI, Group Research Institute, P.O. Box 438, Nelson, New Zealand. Dr. Ashley Gray.
- High Energy Enterprises, P.O. Box 5636, Security, CO 80931, USA. Fax 719-4750582.
- Institute for Advanced Studies at Austin, 4030 Braker Lane W., Suite 300, Austin, TX 78759, USA.
- INE, Institute for New Energy, 1304 South College Avenue, Fort Collins, CO 80524, USA.
- Integrity Institute, 1377 K Street, NW, Suite 16, Washington DC, USA. Fax 202-543-3069.
- Orgone Biophysical Research Laboratory, Inc., P.O. Box 1395, E1 Cerrito, CA 94530, USA.
- Quantum Biology Research Laboratory, P.O. Box 60653, Palo Alto, CA 94306, USA.
- SEA, Space Energy Association, P.O. Box 11422, Clearwater, FL 34616, USA.
- Tesla Book Company, P.O. Box 121873, Chula Vista, CA 91912, USA.
- Tesla Incorporated, 760 Prairie Av., Craig, CO 81625, USA. Fax 303-824-7864.
- Journal Explore, P.O. Box 1508, Mount Vernon, Washington 98273, USA.
- Electric Spacecraft Journal, P.O. Box 18387, Asheville, NC 28814, USA. Fax 704-683-3511.
- Nexus New Times Magazine, P.O. Box 30, Maplepton Old, 4560, Australia. Fax 074-429381.
- Journal Cold Fusion Times, P.O. Box 81135, Wellesley Hills MA 02181, USA.
- Journal Infinite Energy, P.O. Box 2816, Concord, NH 03302-2816, USA.
- Journal 21st Century Science & Technology, P.O. Box 16285, Washington, DC, 20041, USA.
- Journal Cold Fusion, 70b Route 202N, Petersborough, NH 03458, USA.
- Brown's Gas International, 5063 Densmore Av., ENCINO, California 91436 USA.
- ENECO Inc., 391-B Chipeta Way, Salt Lake City, Utah 84108, USA. Fax 801-5836245.
- "Robert Adams and Company" 46 Landing Road, Whakatane, Bay of Plenty, New Zealand. Robert Adams,
- Methernitha, 3517 Linden, Switzeland. Manager Francis Bosshard.
- Swiss Association for Free Energy, P.O. Box 10, 5704, Egliswilli, Switzeland.
- Space Research Institute, Box 33, Uwajima, Ehime 79, Japan. Dr. Shinichi Seike. Fax 895-24-7325.
- Nuclear Power Corporation, 581 400 Karnataka, India. Project Director Dr. Paramahamsa Tewari.
- Cosmic Energy Foundation, Neptunuslaan 11, 3318 E1 Dordrecht, Netherlands. Dr. Martin Holwerda
- Sabberton Research, P.O. Box 35, Southampton S097BU, England, Dr. Harold Aspden.
- Labofex, Experimental and Applied Plasma Physics, Ontario, Canada, Dr. Paulo N. Correa.

Raum&Zeit

EhlersVerlag GmbH,
Geltinger Str.14e,
82515 Wolfratshausen
Telephon: 08171/41 84-76,
Telefax: 08171/41 84-66
Internet:
http://www.raum-und-zeit.com

raum & zeit
Laser harmonisiert Blut und Wasser
El Niño

Institut fur Raum-Energie-Forschung (IREF) i.m. Leonard Euler

- Global Scaling Theory
- Gravitation theory
- Mathematic Biology
- Gravitation energetics
- Hydrogen-Helium Energetic
- Information technologies
- Telecommunication technologies
- Transport technologies

Some General Properties of Matter and Energy Sources

Devoted to memory of Academician A.A. Harkevich

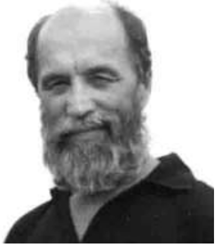

Dr. Eugene D. Sorokodum

Volochaevskaya Street, 40-b, Flat 38,
111033, Moscow, Russia
Telephone: (7) - (095) -362 80 84
Email: e40dum@dol.ru http://vortex.viptop.ru

Introduction

Nowadays there are appear more and more attempts to disclose new methods of obtaining energy from the ambient space, i.e. from heat, electromagnetic, gravitational fields, physical vacuum, aether, etc. (See, for example, publications in New Energy Technologies [1-3]). The publications demonstrate that researchers meet various incompletely investigated physical processes. These scientific areas are difficult to be investigated. All this makes obstacles for development of works on creation of new energy sources. **In this work there is made an attempt to demonstrate general properties of energy sources of any physical nature. This attempt is based on a general analysis of properties of matter.** A theory of transformers is used to investigate general properties of matter.

Action of all physical, technical and biological systems is expressed in energy transformation. Numerous theories based on particular (specific for some system) methods are applied for description of these systems. However, if the number of degrees of freedom and of elements inside a system increases then many theories are not able to describe operation of the systems. First these problems appeared in electrical engineering, radio engineering, automation and acoustics. The theory of transformers is applied to these fields of science. The theory represents a complicated system as a "black box" having several inputs and outputs. Operation of the numerous elements occurring inside the box is represented as some equivalent functions reduced to the inputs and outputs.

In the last decades methods of solution of mechanical dynamic tasks by the method of complex resistances are developed as well as representation of elements as linear transformers [11] and finite elements in liquid [14]. This tendency can be applied to the mechanics of liquids and gases. Now there are successfully developed those concepts which assume observation of models having very few degrees of freedom to be enough for analyzing processes in hydrodynamic systems [10]. However, a mathematical apparatus for description of transformers operation which is well-developed in these areas of science and engineering has a special view and is applicable only for these scientific areas.

Academician A.A. Harkevich developed a theory of a linear transformer up to the level of the general theory of transformers which is applicable for transformation of any types of energy [24]. The general theory of transformers proposed by A. A. Harkevich is applicable for description of various energy sources, flying and swimming objects, functioning of different animals' organs, and technological processes. In this work some general properties of matter and energy sources are investigated, according to the general theory of transformers.

General theory of energy transformer

The whole ambient space, from the microworld to the macroworld, is filled with energy. According to different theories, space is represented as a compact medium (i.e. having distributed parameters) or a medium consisting of a limited number of discrete elements (i.e. having concentrated parameters). On the analogy of hydrodynamics [10, 14] the compact medium can be represented as an equivalent system with a limited number of degrees of freedom. Hence, the whole space can be represented as some system consisting of elements and communications between the elements with a limited number of degrees of freedom. Energy exchange occurs due to the degrees of freedom. Energy transmission occurs if energy gradient is presented in the ambient space. Due to the energy gradient a force tends to realize transmission in the space [4]. The elements can be systems as well.

Hence, the more we observe dividing the elements on the systems and the systems on the elements the more we will penetrate into the microworld (i.e. atoms, elementary particles, physical vacuum, aether, etc.). The more we combine the elements in the systems, and the systems in the new larger systems the more we observe the macroworld (the Solar System, galaxies etc.). All the systems and the elements are interconnected. The systems and their elements are transformers of energy. Energy motion occurs in the smallest part of space. Consequently, the whole space can be represented as a system consisting of the energy transformers. In the general case due to every type of energy limits of a transformer consist of outer limits (i.e. communications with the macroworld) and inner limits (i.e. communications with the microworld). Dividing into the macroworld and the microworld occurs in relation to the size of a transformer for every type of energy. Types of energy coming through communications of a transformer and inside it may differ, i.e. mechanical, heating, electromagnetic, chemical and other known and unknown types of energy.

The processes in which a great number of interacting elements and different types of energy participate are very complicated and do not allow describe the processes accurately by modern mathematical methods. Hence, there is a problem to find such methods of solution of the tasks which without disclosing all the communications inside the element can give the understanding of the way an element moves in the system. The following premises can be made for the method of solution of the task of a system and its elements' motion:

1. Motion of all the elements in the space is characterized by energy exchange occurring among them;
2. We are interested in a certain limited area for every individual type of energy. The chosen limited area will be called as **a transformer**;
3. The transformer has degrees of freedom both on its limits (sides) and inside it (inner degrees of freedom);
4. Further this limited area (the transformer) which has or is able to have a limited number of degrees of freedom (sides) at its limit will be observed;
5. Interaction between this transformer and the ambient space occurs only through these degrees of freedom (sides) by means of energy exchange;
6. Motion of energy between the elements of the transformer occurs according to its degrees of freedom inside the transformer. There can be a limited or unlimited number of degrees of freedom.
7. All the space is full of the transformers. All the transformers adjoin each other without gaps. Energy exchange between the transformers occurs through their sides which do not have a size but reflect general kinematical and dynamic characteristics of the energy transferred through these sides.

Editor: The full variant of the article includes mathematical description of operation of energy transformers which is followed by these conclusions:

- There can be any number of different types of elements having a corresponding number of degrees of freedom and any type of energy inside the transformer. However, on every side of the transformer a generalized force has the same value and depends only on change of energy according to this degree of freedom. Therefore, two transformers are considered to be equal if the generalized forces (energy changes) on all the sides are equal. In this case it is not necessary for the inner structure of the transformer, number of inner degrees of freedom, and energy types of these two transformers to be equal. This is **a principle of equivalency of transformers at the equivalency of the forces acting at the sides of transformers**.

- In the general case the generalized force at any side depends not only on energy exchange occurring at this side but on energy exchange occurring among the sides or at other sides of the transformer. The generalized forces are produced by energy distribution in the space.
- A transformer is characterized by the fact that there are different types of energy at its different input sides (or the energy can be of the same type but having other characteristics of motion).

Resistances of the interaction provide information about physical properties of the transformer, in particular, about physical interaction between the sides. At that a number of inner degrees of freedom in this transformer as well as the reactions at the other sides do not play any role. Proper resistances of the sides and resistances of the interaction provide information about inner physical characteristics of the transformer reduced to equivalent values at the sides. Hence, it is possible to have two transformers having equal equivalent resistances of interaction for all the sides. However, according to their geometrical and constructional characteristics and types of energy, the transformers will be different. The resistances may be a function of kinematical characteristics **(a nonlinear transformer)** or of time **(a parametrical transformer)**. Moreover, it can have constant values **(a linear transformer)**.

It is appropriate to consider physical properties of matter included in the volume of a transformer only due to resistances at the sides of the transformer.

Structure of the transformer

Every inner degree of freedom can be represented as a series circuit. By analogy with electric circuits [6, 7, 11, 16-18] the transformer can consist of various circuits. An unlimited number of degrees of freedom allows represent a transformer consisting of an unlimited number of chains. The circuits have series, parallel, or mixed junctions. A part of a circuit whose elements have the same generalized displacement is called as a branch. The branch can consist of one or several elements. A place where three or more branches are joined is called as a multiple junction. A circuit is considered to be a closed path including several branches and multiple junctions. All the elements of a chain connected in series have equal generalized displacement (as well as equal speed and acceleration). The generalized force acting on the whole circuit is equal to the sum of forces acting on the elements of the series circuit.

The elements connected in series can be replaced by one element in such a way that the generalized displacement and the summary generalized force would not change. In the case of a parallel connection all the branches of the circuit are joined to the same pare of multiple junctions and are under influence of the same generalized force. Parallel circuits can be replaced by series equivalent circuit and vice versa [6, 7, 11, 17, 18]. Hence, a number of inner degrees of freedom and a quantity of chains may be decreased as well as increased. In this case resistances occurring on all the sides will be constant.

Structure of the linear transformer

The simplest transformer is a one-side transformer having one inner degree of freedom...

If to the input of the transformer we deliver generalized displacement (or force) which depends on time as a pure sinusoidal signal then the linear transformer will have sinusoidal reaction of the same frequency. For the sinusoidal signal with fixed frequency the linear transformer (including a transformer having an unlimited number of inner degrees of freedom) can be represented as an equivalent transformer reduced to this input as one chain. An equivalent transformer placing at this input at other fixed frequency of sinusoidal disturbance will be represented as a chain having one degree of freedom but different values of chain elements.

If the signal at the input of the transformer is of complicated time-periodical form then it can be decomposed to Fourier series. Every harmonics of the Fourier series will produce its own chain. The general complicated signal will produce the complex chain consisting of compound sum of elementary series chains. This complicated chain consists of elementary chains interconnected in series or in parallel and combined in multiple junctions and circuits. The construction principles of this complicated chain can be based on methods of identification and synthesis which are well developed in electrical engineering, automation for linear and non-linear transformers [9, 12, 13]. These principles are formally useful for transformers which transform energy of any physical nature. Identification produces so much elementary chains as the number of inner degrees of freedom. Every elementary chain reflects motion by one inner degree of freedom. Physical analogy occurring among motions of different nature is actual here [11, 16, 18]. Values of the generalized masses, elasticities, dissipative and active elements can depend on kinematical variables (non-linear chains), on time (parametrical chains), or they can be constant (linear chains).

If a series circuit includes mass and elasticity then it will have resonant frequency. At this frequency reaction of mass and elasticity will be absent in the summary reaction. It is equal to the fact that we will know nothing about the value of mass (and elasticity) and, moreover, about its presence at all. The mass can have huge value, but it will be absent for us. Hence, if we assume a transformer as an atom then the atom may include elementary chains having huge masses (much more than the mass of the Earth), but we will not know about it until we apply such frequency to the input which explicates this mass.

Any energy transformer including a transformer with distributed parameters can be represented as a system of elementary series and parallel chains connected in a certain way. A concrete transformer can have various types of these connections. All of them can be included into another combination of connections that is accompanied by change of values of masses, elasticities, dissipative and active elements. However, in this case equivalency of reaction should be fulfilled at a certain diapason of frequencies on all the sides of the transformer.

There is a special spectrum of resonant frequencies for every combination of connections of the transformer. There will be an unlimited number of degrees of freedom and, hence, an unlimited number of resonances for the transformer with distributed parameters (for example, an elastic nail, a compressible fluid having a limited size). Our influence on the transformer usually has a certain diapason of frequencies. As the result of a limited diapason of excitation frequency the transformer will represent a system having a limited number of inner freedom, even if the transformer has distributed parameters.

Values of equivalent elements depend on oscillation frequency. Moreover, according to known laws, several parallel circuits can be transformed into a series one and vice versa. Taking it into account the following conclusions can be made:

1. The generalized mass, elasticity and dissipative elements do not have constant values in nature. Their value depends on frequency of the process, i.e. they depend on the character of time-change of the disturbance acting on the transformer.
2. The generalized mass, elasticity and dissipative elements are products of time-space change of energy.

Types of energy transformers

Transformers can be nonlinear, parametrical and linear. The transformer can be **active** (having an internal source) and **passive** (having an external source). The passive transformer can never be an energy source (by definition). A transformer can accept or transform one type of energy (monoenergetic transformers) or several types of energy (polyenergetic transformers). The polyenergetic transformer includes a mechanism and corresponding elements and chains which allow transform one type of energy into another one. Fundamentally, all the transformers are polyenergetic transformers of energy, however many of them can be represented as monoenergetic ones since other types of energy participate weakly in them. Designing a transformer it is more convenient to put the energy source out the transformer. Let us further classify the passive transformers.

The passive transformers can belong to the following types:

1. Simple transformers

In this case energy is produced at the output due to transformation of energy coming to the inputs into energy of another type or having other characteristics. The following transformation can serve as the example.

An electric transformer: resistance of one value comes to the input, and the resistance of another value comes from the output (a type of energy is the same).

A furnace for combustion of fuel: energy of chemical connections comes to the input and heating energy comes from the output (energy of radiation may be neglected).

A Wind Generator, Hydroelectric Power Station: kinetic energy of moving air or water comes to the input, and electric energy comes from the output.

2. Energy intensifiers

In this case a transformer intensifies energy coming to one of the inputs due to energy coming to the other inputs (or to the other input). Below there is an example for a *monoenergetic transformer*. Energy (an electric signal having certain parameters) comes to the input of a transformer, and the output energy has the same characteristics but it is more intensive. It can occur due to electric energy coming to the other input of the transformer.

A heat pump represents an example for a *polyenergetic transformer*. Low potential heat energy of the ambient space (air, water, ground) comes to the input of the heat pump, and electric energy from a power network comes to the other input. The output heat energy has value which is equal to the sum of the coming heat energy of the ambient space and electric energy coming from the power network through the electric engine of the compressor. The output heat energy excesses energy coming from the power network by several times. The output energy is always lower than the sum of the low potential heat energy and the electric energy coming to the input. Hence, the output of a heat pump is always less than unit.

Efficiency of the heat pump consists in the fact that a consumer pays for electric power produced by power network only (use of the low potential heat of the ambient space is free). However, the output heat energy is several times more than the energy which can be produced by the electric power coming to the input.

3. Active oscillator

An active oscillator is a device of a certain type. Energy of time-constant characteristics comes to its input, and at output there is energy of time-periodical characteristics. Energy with time-periodical characteristics can come to the input as well since these characteristics are not connected with the time-periodical characteristics of the output energy. The transformer can operate in a mode of an active oscillator if it has the following features:

1. Indirect connection as a series circuit: one of the outputs is connected to an additional transformer whose output is connected to one of the inputs of the transformer. The additional transformer provides certain changes of the input signal (time delay).
2. Energy comes to the input (it can be time-constant or time-periodical).
3. The transformer must have a certain type of non-linearity.

Examples of the active oscillator

Internal combustion engine: the indirect connection is the system of combustion, the non-linear element is the system of admission and exhaust valves, and the energy source is energy of chemical connections of fuel.

Various electric generators: all of them have the indirect connection, the non-linear element, and the energy source.

The active oscillators differ from the electrical intensifiers qualitatively. The difference consists in the fact that the active oscillator can operate and receive energy from the ambient space without additional energy coming to it (from a power network, for example).

Theorists and experimenters try to create a self-supporting self-organizing system in their works on creation of alternative sources. This fundamentally correct tendency requires using theory of active oscillators. There is the developed theory in electrical engineering, automation, and other scientific fields [25]. To simplify application of these theories it is necessary to use analogy between equal processes which take place in different physical fields.

4. Rectifiers

Time-periodical energy comes to the input of the transformer. The output energy has time-constant or practically time-constant characteristics. If the transformer has a non-linear element or elements of a certain type (a diode or a one-sided valve) then it can operate in the mode of a rectifier. The examples are electrical rectifiers and pumps of a valve type.

Properties of energy transformers

General properties:

1. A transformer can transform energy of one type as well as transform one type of energy into another type.
2. The generalized kinetic, potential and dissipative energies are specific peculiarities of energy change in the space and time.
3. The generalized masses, elasticities and dissipative elements are products of energy change in space and time.
4. Values of the generalized masses, elasticities and dissipative elements (existing inside the transformer as well as reduced to the equivalent values on the sides of it) depend on a temporal character of changes of kinematical or dynamic disturbances. Values of the generalized masses and elasticities at the sides of the transformer are lower than the corresponding values inside the transformer.
5. Energy interchange between macrostructures and microstructures can occur through a transformer.
6. One area of space can "know" about another area of space through energy exchange only, i.e. through generalized dynamic and kinematical interactions. In the light of the fact that chains of different transformers can have similar physical properties (for example, an equal resonant structure) interference of these transformers seems to be possible including interference occurring at a great distance.

Properties of the passive monoenergetic transformer:

1. Transformation of characteristics of the same energy type.
2. The output energy can be both equal or lower than the sum of energies coming to all the inputs. The sum of the output energies decreases due to dissipative losses occurring inside the transformer.

Properties of the passive polyenergetic transformer:

1. One energy type coming to an input can be transformed into another energy type coming from the output.
2. The sum of energies of all the outputs can be equal to the sum of energies of all the inputs or less than it. The sum of the output energies decreases due to dissipative losses occurring inside the transformer. The output will always be lower than unit. If the energy coming from certain of the outputs (within the limits of one output) is considered as a positive effect then the output will be less than the mentioned value.
3. Energy of one and the same type comes to the input and from the output while other energy types can come from other outputs. This property can allow increase one type of energy coming from an output at the expense of the other types of energy coming to the inputs.

Analysis

This article is aimed to provide general properties of the energy transformers and to analyze some problems of creation of renewed energy sources, according to these properties. There are made various attempts to design a renewed energy source which would produce energy without fuel consumption. In the light of the above mentioned results, the researches on creation of the renewed energy source may be represented by the following way. Energy can be produced from the macrostructures and the microstructures.

In the case of the macrostructure, energy can be produced by a simple transformer, an electric intensifier or an active oscillator. Methods of production of energy from the ambient space by means of simple transformers (thermoelectric and hydroelectric power stations, wind generators, etc.) are developed sufficiently by the humankind. Works on producing energy from the ambient space by means of energy intensifiers are on its initial stage. Heat pumps, which utilize low potential heat energy by means of Freon-compressor heat pumps, are successfully introduced. One of the perspective approaches is replacing such a heat pump by a vortex oscillating heat pump which does not contain Freon and a compressor [3, 5, 19, 22]. In future it will be possible to replace the energetic pump by a new energy transformer of the type of an active oscillator. In this case the energy source will always obtain energy from the ambient space without using additional energy from a power network. Energy sources of the type of the active oscillator are internal and external combustion engines, stream engines, etc.

In the general case a transformer can contain different energy types and has communications with **outer microstructures**. It should be noted that the outer microstructures can be included in the geometrical size of the transformer. Energy of this microstructure can be huge; hence, this energy can be obtained from them for a very long period of time. Atomic reactors can serve as an example for this fact. The types of energy in these microstructures can be represented by known types (electromagnetic energy, nuclear energy) and unknown types (whose numerous quantity can be considered). These energy types exist in the ambient space. However, almost all the transformers of energy produced by a human (except the nuclear reactors) do not interact with the energies, and we do not obtain energy from these microstructures.

The task of creation of a new energy source consists in designing such an energy transformer which could obtain energy from the microstructures and transform it into the type of energy which is needed to a consumer. In this case all the designed transformers will not have over unit output. Efficiency of these transformers will be defined by the cost of the energy produced by them. This energy depends on a ratio of energy coming to one of the inputs (electric power) (for example, in heat pumps) to the output energy. If the output energy is much more than the artificial energy which comes to one of the inputs then this transformer can be very useful for a consumer.

New high effective energy sources are often claimed to be created. According to the observed facts, existence of such sources is practically possible. Nevertheless, realization of this source will depend on our knowledge of physics of the microstructures and methods of transformation of energy of the microstructures into such type of energy which can be used by a consumer. Physics of microstructures is investigated badly, and many types of energy are totally unknown. Therefore, modern scientists try to create the new energy sources at random. In this case the following results can be achieved:

1. The high effectiveness of an energy source (over unit output) does not correspond to reality.
2. The output energy of the energy source is actually higher than the input energy (the source of the energy intensifier type).
3. The energy source actually produces energy without artificial input energy (the source of the active oscillator type). It can occur if an inventor failed to recognize the source of energy coming from the microstructures. It can be connected with particular or total lack of exploration of physics of the microstructures. Moreover, the fact is caused by transcendent difficulty of disclosing energy sources from the macrostructures.

Conclusions

It is principally possible to create an energy source whose output energy would be higher than its input energy (an energy intensifier) or an energy source which would produce energy without additional input energy (an active oscillator).

1. Various energy sources or transformers of energy of the macrostructures and microstructures can be designed.
2. Practical value of the energy source is defined according to cost of its output energy instead of the output value.
3. Designing new energy sources it is useful to consider them from the point of view of transformers. It is worth to define all the communications of the transformer and to state experimentally energy at all the accessible communications of the transformer.
4. Developing theories of sources it is expedient to use physical analogy and to take into account ready results of works on theories of transformers, chains, etc. of electrical engineering, radio engineering, automation and other scientific fields.
5. In the light of the fact that it is planned to obtain energy from the microstructures it is necessary to pay attention to new probable types of radiation and to their influence on a human.

References

1. Frolov A.V. The work Advanced Energy and Propulsion Systems based on Chronal Reaction Method // New Energy Technologies. - 2001, 2. p.51-55.
2. Frolov A.V. Reactionless Propulsion and Active Force // New Energy Technologies. - 2001, 2. p.56-58.
3. Sorokodum E. Recycling of energy of the atmosphere by means of vortex // Russian technologies for industry “Renewable Energy” IWRFRI’2001,may 28-30, 2001, Saint-Petersburg. Abstracts. - 2001, p. 92-93.
4. Sorokodum E. On general nature of forces. New Energy Technologies. 1(4), 2002, p.30-36.
5. Sorokodum E. D. Vortex heat-generators. New Energy Technologies Issue. Vol.2(5), 2002, p.17-76.
6. Andreev V.S. Theory of nonlinear electric circuits. M.: Radio and Communication. 1982, p.280
7. Atabekov G.I. Theory of linear electric circuits. M.: Soviet Radio. 1960, p. 712
8. Berezkin E.N. Lectures on theoretic mechanics. Part 11. Dynamics of the system, dynamics of solid body, analytical dynamics. M.: MSU - 1968, p.314
9. Blakyer O. Analysis of nonlinear systems. Transl. from English. Edited by Hohlov R.V. M.: Mir. – 1969, p. 400
10. Gledzer E.B., Dolzansky F.V., Obuhov A.M. Systems of hydrodynamic type and their application. M.: Nauka – 1981, p. 366
11. Druzinsky I.A. Mechanical circuits //L.: Mashinostroenie. – 1977, p. 240
12. Kononenko V.O., Plahtienko N.P. Methods of identification of nonlinear mechanical oscillating systems
13. Zakrzhevsky M.V. Oscillations of essentially-nonlinear mechanical systems. Riga, “Zinatne”. – 1980, p.190
14. Ladikov Yu.P. Stabilization of the processes in continuums. M.: Nauka. – 1978, p. 432
15. Merkin D.R. To the question on forces’ structure // Applied mathematics and mechanics. - 1975, 39, 35. p. 929-932
16. Oltson G. Dynamic analogies. M.: State Edition. Foreign Literature. - 1947, p.223
17. Popov E.P. Theory of linear systems of automatic regulation and control. M.: Nauka. – 1989, p.302
18. Puhov G.E., Vasilyev V.V., Stepanov A.E., Tokareva O.N. Electrical modeling of the tasks of structural mechanics. Edition of Academy of Science of Ukrainian SSR. - 1963
19. Rodionov B.N., Sorokodum E.D. Vortex energetic. Constructional materials, equipment, technologies of the XX century. 3 (26). 2001, p. 28, 29
20. Sorokodum E.D. Theory of transformers for nonlinear biomechanical oscillating systems. Taganrog institute of radio engineering. – Dep. in VINITI, 7.05.1980, #1948-80 Dep. – 1980, p. 38
21. Sorokodum E.D. On the application of theory of nonlinear transformers for the calculation of complex mechanical oscillating systems. Reports at the scientific conference “Problems of nonlinear oscillations of mechanical systems”. Kiev. - 1978
22. Sorokodum E.D. Vortex source of energy. International scientific and social conference. Prospects of preservation and developing of the united civilization of the planet. Culture. Ecology. Space. 26-31 of May, 2002
23. Strelkov S.P. Introduction to the theory of oscillations. M.: Nauka. – 1964, p. 440
24. Harkevich A.A. Theory of transformers. State Energetic edition, M.-L.: 1948, p. 191
25. Harkevich A.A. Auto-oscillations. / Harkevich A.A. Selected works in 3 volumes, V. 2, M.: Nauka – 1948, pp. 253-379.

Commercial Exploration of Space

Review
by Olga Leontyeva, Editor
http://www.faraday.ru

Of late years the mankind has greatly advanced in space exploration. New spaceships are created, automatic apparatus are launched to the planets of solar system, space stations are orbited. A man has passed to outer space and begun exploring the planets of solar system. More and more specialists and people of different professions are involved in the development of space industry. However to present day only a few of them has succeeded flying into space.

At the present state of affairs space flight is very expensive and a serious reason should exist to forward a man into space if he or she does not relate to the aims of professional cosmonautics. However in last years there has appeared a certain breakthrough in the solving of the problem. Nowadays space technologies gradually turn from the sphere of experimental and scientific researches to the area of practical application. The time has come for a man to realize the real space flight without being the professional cosmonaut.

What are the ways to solve the problem? No doubt that special attention should be paid to the development of new space technologies, search of new types of fuel and attraction of investments to the sphere of space tourism.

April 28, 2001 can be considered as official date of birth of space tourism, when there was launched the space ship "Soyuz TM-32" with the first space tourist aboard. Almost in a year, on April 25, 2002 the space tourist #2 South African Republic person Mark Shattlword started on his space journey.

Today many companies give the opportunity to make a real space flight for all comers who have enough money and health. It is promised the providing of the most modern space technologies and the most perfect space equipment. It can be created a furor by the "Minimum program" of the Russian company "Atlas Aerospace" which is made by the members of Yu.A. Gagarin Center of training of cosmonauts.

In recent years at the international market there appear more and more companies which deal with search and encouragement of inventors who work on creation of alternative propulsion systems. Activity of "X Prize Foundation" company can be considered as an example.

The X PRIZE Foundation

722-A Spirit of St. Louis Blvd
St. Louis, Mo. 63005
Tel: 636-519-9449, Fax: 314-533-6502
http://www.xprize.org
E-mail: press@xprize.org

The X PRIZE is a $10,000,000 prize to jumpstart the space tourism industry through competition between the most talented entrepreneurs and rocket experts in the world. The $10 Million cash prize will be awarded to the first team that:

- Privately finances, builds & launches a spaceship, able to carry three people to 100 kilometers (62.5 miles)
- Returns safely to Earth
- Repeats the launch with the same ship within 2 weeks

The X PRIZE competition follows in the footsteps of more than 100 aviation incentive prizes offered between 1905 and 1935 which created today's multibillion dollar air transport industry.

For more than 30 years, the general public has waited for an opportunity to enjoy the space frontier on a first-hand basis. The X PRIZE Foundation is working to make space travel possible for all. The spaceships that compete for the X PRIZE are designed to carry passengers.

Since its inception in May 1996, the X PRIZE Foundation has registered more than 20 teams from seven countries to compete for the prize. The X PRIZE is fully funded through January 1, 2005, through private donations and backed by an insurance policy to guarantee that the $10 million is in place on the day that the prize is won. Additional funds are still being raised by the X PRIZE Foundation to implement the competition (judging, media, event management, etc.) and continue the Foundation's education mission.

The X PRIZE was inspired by the early aviation prizes of the 20th Century, primarily the spectacular trans-Atlantic flight of Charles Lindbergh in The Spirit of St. Louis which captured the US $25,000 (US$) Orteig prize in 1927. Through a smaller, faster, better approach to aviation, Lindbergh and his financial supporters, The Spirit of St. Louis Organization, demonstrated that a small professional team could outperform a large, government-style effort.

The Societal Benefits of the X PRIZE include:

- Creation of a new generation of heroes
- Inspiring and educating students
- Focusing public attention and investment capital on this new business frontier
- Challenging explorers and rocket scientists around the world; and,
- Vehicles built for the X PRIZE will eventually serve four different industries:
  - Space Tourism
  - Low-cost satellite launching
  - Same-day package delivery
  - Rapid point-to-point passenger travel.

In Fig. 1 it is demonstrated the typical X Prize trajectory.

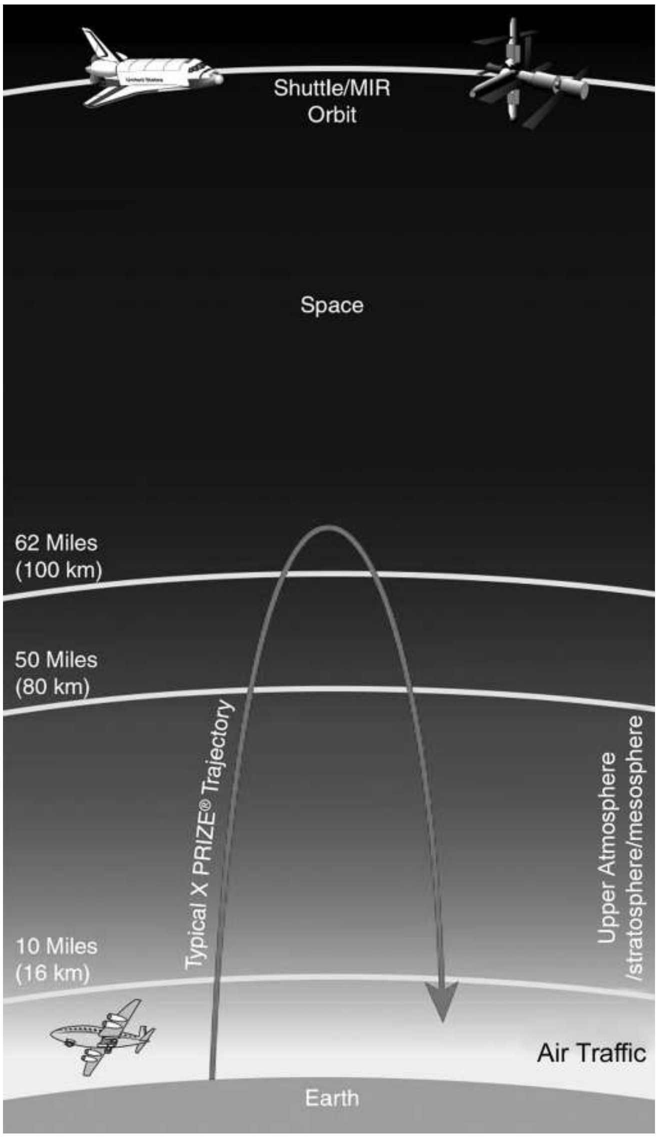

Fig. 1

Editor: Below we publish photos and brief comments about some official X PRIZE registrant teams. More detailed information you can find at http://www.xprize.org/imagefacts/photo1.html.

The Da Vinci Progect

http://www.davinciproject.com

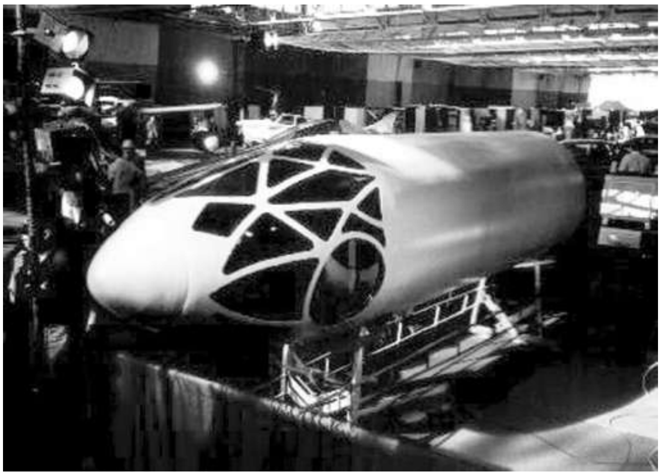

Fig. 2
"Wild fire"

The **da Vinci Project** will launch its spacecraft ("Wild Fire") from the world's largest helium balloon. The 3,270 kg (7,200 pound) rocket will be tethered 720 meters (2,400 feet) below the balloon and lifted over the course of an hour to an altitude of 80,000 feet. The 10,000 pound thrust, liquid oxygen, kerosene engines will fire the first stage and the rocket will fly an initial angular trajectory to clear the balloon. The spacecraft then will transition to vertical flight to its apogee of 120 km in space. The rocket will reach a maximum speed on both its ascent and re-entry of Mach 4, or 4,250 kph (2,650 mph).

An innovative ballute will protect and stabilize the rocket on re-entry. A flyable parachute will be deployed at 25,000 feet and the rocket will descend under control, guided by GPS, to a predetermined landing zone. The da Vinci Project has already successfully conducted full-scale rocket motor test and has built a full-scale mockup of their vehicle.

"Discraft" corporation

http://www.xprize.org/teams/teams.html

John Bloomer, the team leader, is an aerospace engineer and he has worked on many aerospace progects, including Apollo and holds more than 60 patents on a disc plattorm aircraft. Bloomer's ship utilizes "Blastwave" Pulsejets.

Flight Sequence

Fixed, 7850-ft^2-area, laminar-flow wing take-off at about 60 mph within about 150 ft, featuring climb with gradual air-breathing acceleration (according to a fixed program) at fixed angles, to exit the atmosphere at Mach 10 on an unpowered ballistic arc to reach 75-mi. altitude: return on down-leg of same unpowered arc to gradual power-on flare-out re-entry of the atmosphere in simple reverse sequence of the take-off velocity profile. Range above 100,000 ft is about 480 mi. which is covered in about 5 minutes.

"Kelly Space & Technology"

http://www.kellyspace.com

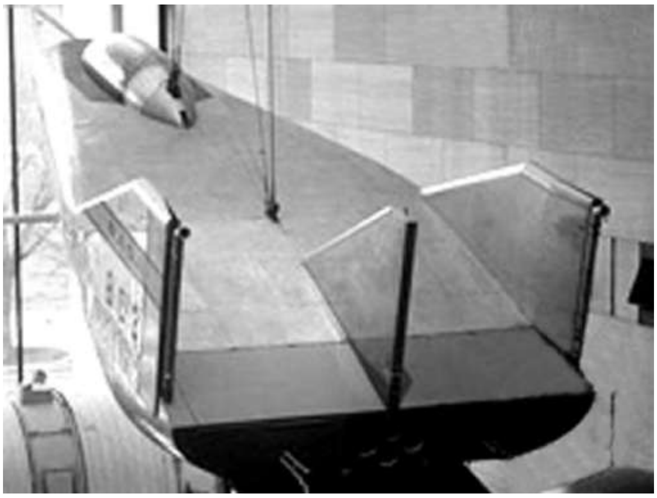

Fig. 3
"LB-X"

The vehicle is a rocket-powered delta wing glider with a liquid oxygen and kerosene liquid rocket engine. The spacecraft is prepared and fueled at the takeoff airport. The spacecraft is towed to release altitude behind a conventional jet powered aircraft such as Boeing 747 aircraft. Upon release from the tow aircraft, the main engine is throttled up for boost phase. The vehicle nominally coasts to an apogee altitude of 100 kilometers. The vehicle then glides to a landing at the takeoff airport for checkout and refueling prior to the next flight.

Flight Sequence

The lifting body will be towed to launch altitude behind another aircraft, and the rockets will be ignited. The craft will return to the landing site and make an unpowered, horizontal landing.

"Lone Star Space Access"

http://www.dynamicar.com

The Cosmos Mariner employs air-breathing jet propulsion for take-off and landing from conventional airports and rocket propulsion for ascent from cruise flight in the stratosphere to 30 or 40 nautical miles altitude. From there, the vehicle coasts to a target altitude of around 65 nautical miles. The air frame is designed to interface with two jet engines (turbofan or turbojet) each with 20,000-lbs static thrust or less. For rocket propulsion, the Cosmos Mariner will use the Aerojet AJ26-NK31A, a staged-combustion kerosene engine. The vehicle is designed to take off and land from conventional runways.

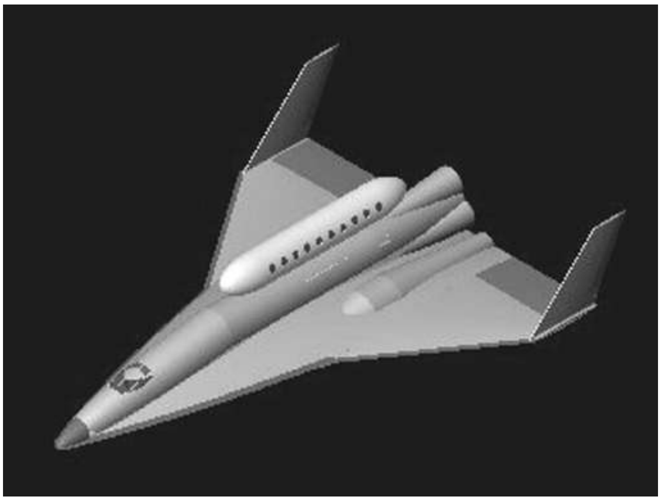

Fig. 4
"Cosmos Mariner"

Along with X Prize Foundation it should be mentioned ***ALLTRA*** *Company (Germany). ALLTRA consists of a small group of space experts. The main objective of ALLTRA is to 'sell' the space idea to a broad public and to identify future commercial opportunities in the space sector.*

On the official ALLTRA website http://www.alltra.de you can find collection of artist's views of the projects which are aimed at decrease in space ships value by means of using of new types of engines. Besides the Company deals with the development of space hotel projects.

JNET

New Energy Technologies
collection of articles
Japanese Version

Collection of articles from New Energy Technologies magazine, 2001-2002 is published in Japan.

On purchasing of the book, please, contact

Eiichi Yamamoto,
President Yama Trans Co. Ltd.

Email: admin@yamatrans.co.jp or
eyama@yamatrans.co.jp

Alternative Aircrafts and Space Propulsion Systems

Editor: Nowadays all over the world there are made numerous attempts to create new types of aircrafts and alternative space propulsion systems. Below we publish the review of some interesting devices which are presented in the modern market.

"Arbortech Pty.Ltd"

http://www.airboard.com.au

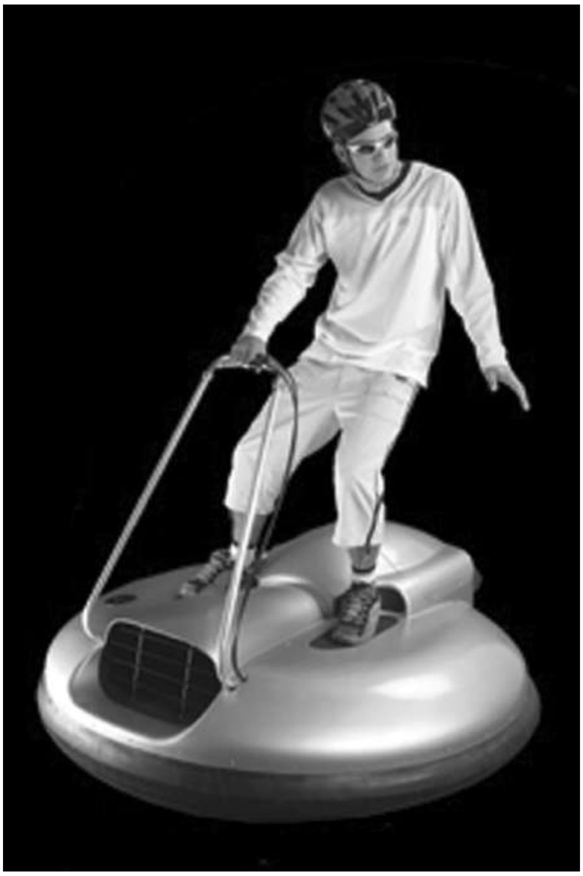

Fig. 1
Airboard 2000

General Technical Specifications

Total payload, including rider — 100 kg (220 lb)
Operating time — 1 hour on full tank of fuel
Construction — Fiber glass/High-impact plastic shell, Aluminum frame, Rubber skirt
Starting — Electric key-start, Battery included
Engine — Briggs & Stratton 4-stroke
Fuel tank capacity — 5 litre (1.3 US gal)
Fuel type — 85 Octane unleaded

How does it work?

Airboard uses Hovercraft air cushion principles to glide just above the ground. The air cushion is generated from a purpose designed engine and fan which are suspended below the Airboard shell. In addition to providing an air cushion for the Airboard to ride on the fan also provides a fast stream of air exiting from the rear of the craft to provide forward thrust – just like a hovercraft. To help provide better acceleration the Airboard also includes a unique friction drive wheel at the rear. When the rider wishes to accelerate forward, weight transfer is used to bring the friction drive into contact with the ground. By engaging the friction drive clutch the Airboard can be accelerated forward without losing the feel of hovering above the ground. The control of the Airboard is provided by weight transfer of the rider, similar to surfboards, skate boards and snow boards. In operation the Airboard can be started, stopped and steered in a controlled manner and this means that riders can perform stunts, trick maneuvers and race around tight tracks against each other. In this unique way the Airboard is the first ever vehicle to deliver the advantages of hovercraft vehicle without the disadvantages of proof acceleration and handling.

"Trek Aerospace, Inc."

http://www.solotrek.com

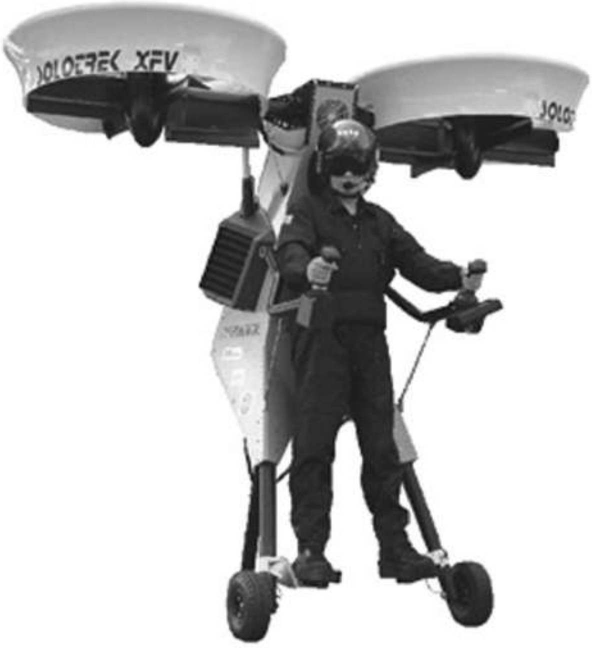

Fig. 2

Trek Aerospace

Preliminary Specifications and Predicted Performance

Normal Gross Take Off Weight	800 Lbs.
Fuel (15 U.S. Gallons)	98 Lbs.
Mission Payload, net of fuel	352 Lbs.
Takeoff/Landing Distance	0 (VTOL)
Maximum Speed	70 Mph
Range	120+ Miles
Hover/Loiter Endurance	2+ Hours
Engine Type	Advanced Int. Combustion
Fuel Requirements	Heavy-Fuel or Gasoline

Special Features: 100% fly-by-wire control system with electronic stability augmentation; Intuitive, easy to fly safely; Minimal field service requirements; Rugged, efficient power train.

DM AeroSafe

http://dmaerosafe.freeservers.com

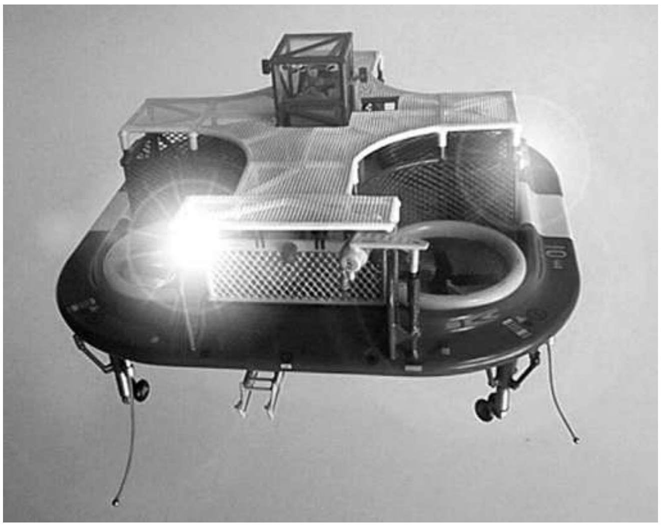

Fig. 3

EAGLE vertical take-off and landing aerial rescue platform

DM AeroSafe is a small research and development team, which has developed a totally new high-rise rescue technology to retrieve trapped people from areas which cannot be reached by conventional aerial ladder, conventional helicopter or a helicopter equipped with a Heli-Basket.

This technology could give rise to a new class of air transportation means, used for safe close-in maneuvering around tall structures, even inside the highly populated areas.

Eagle Aerial Rescue Platform's Performance Summary

Dimensions:

Length - 42 feet (12.8 m)
Width - 42 feet (12.8 m)
Height - 16 feet (4.8 m)

Weights:

Max. Gross weight - 5000 pounds (2268 kg)
Payload (Crew of two plus 10 rescued people) - about 2000 pounds (about 900 kg).

Propulsion System:

Four (4) variable collective pitch Ducted Propellers
Propeller Diameter - 7.8 feet (2.4 m)
Engines - four (4) four-cylinder radial piston aero engines with forced air cooling system 250 hp each (1000 hp total)

Performance:

Max. Duration - about 5 hours without refueling
Max. Airspeed - 40 knots (75 km/h)
Max. operating altitude - about 7000 feet (about 2200 m).

"LTAS/CAMBOT, Inc."

http://www.lvcm.com

Passenger Craft "Ltas 30-Xb"

The ships have rigid monocoque hulls, hybrid powered vectored thrust control and full active buoyancy control. (LTAS US Patent Pending).

This small 2-3 person craft at 70 to 80 feet in diameter will demonstrate ALL production systems and is designed for the LTAS FAA Type Certification program.

NASA's Langley Research Center

http://science.nasa.gov

NASA researchers are studying insects and birds, and using "smart" materials with uncanny properties to develop new and mind-boggling aircraft designs.

The "personal aircraft" that replaces the beloved automobile in people's garages may still lie in the realm of science fiction or Saturday-morning cartoons, but researchers at NASA's Langley Research Center (LaRC) are developing exotic technologies that could bring a personal "air-car" closer to reality.

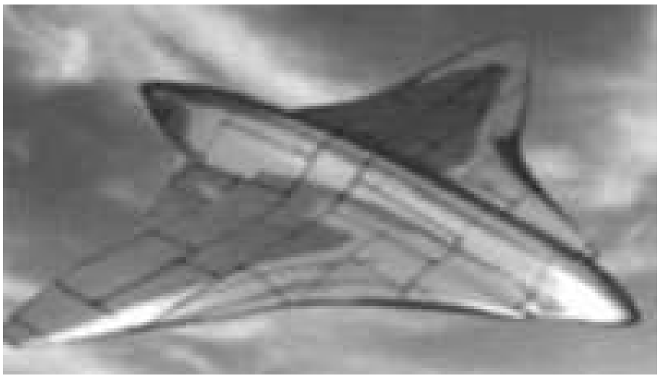

Fig. 4

And air-cars are just the beginning! Self-healing wings that flex and react like living organisms, versatile bombers that double as agile jet fighters, and swarms of tiny unmanned aircraft are just a few of the science-fiction-like possibilities that these next-generation technologies could make feasible in the decades ahead.

Laser Beam Flight

Lightcraft Technologies, Inc. (LTI)

http://www.lightcrafttechnologies.com

Lightcraft Technologies, Inc. is a new company committed to providing low-cost access to space through the use of beamed energy propulsion. (See color photos on the cover page).

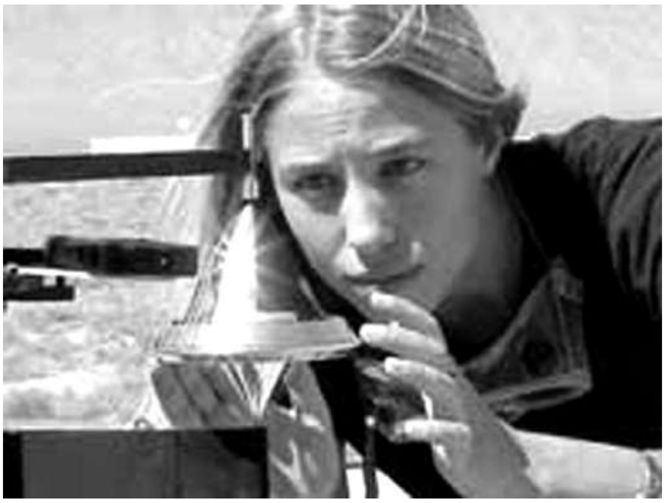

Fig. 1

In 2000 at the High Energy Laser Systems Test Facility (HELSTF), Lightcraft Technologies, Inc. (LTI) set a new world's altitude record of 233 feet (71 meters) for its 4.8 inch (12.2 cm) diameter laser boosted rocket - in a flight lasting 12.7 seconds. Although much of the flight was spent hovering at 230+ feet, the Lightcraft sustained no damage and will fly again. Besides setting the new altitude record, the craft demonstrated the longest ever laser-powered free flight and the greatest "air time" (i.e., launch-to-landing/recovery). LTI launched a total of seven vertical flights between 8:30 am and 11:30 am with three Lightcraft weighing less than 1.8 ounces (51 grams). Two of the flights by Lightcraft #3 reached 159 and 184 feet with the same propellant load!

...set a new world's altitude record of 233 feet!

The record flights were powered by the 10 kW pulsed carbon dioxide laser named "PLVTS" by the organization that owns it: the Directorate for Applied Technology, Test and Simulation (DATTS). Even though PLVTS was suffering from an arcing or grounding problem that caused it to run erratically, the laser power was still adequate to propel the craft to record altitudes.

What is a Lightcraft?

A Lightcraft is a 1kg launch vehicle, made from high temperature ceramic materials, that flies into space on a megawatt laser beam.

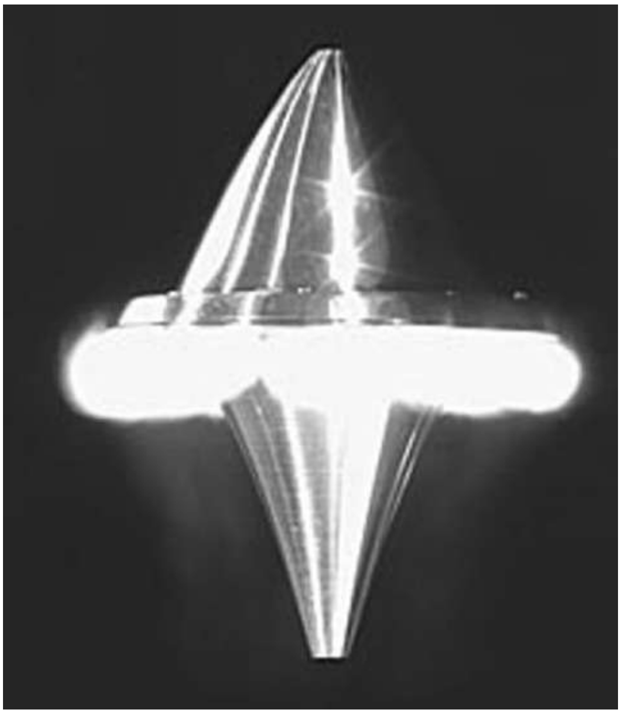

Fig. 2

The Lightcraft is both a single-stage-to-orbit launch vehicle and a satellite. If you have any further questions or comments, write or call LTI headquarters in Bennington,VT. The company representatives would be delighted to send you more information, or answer any inquiries over the phone.

How does it work?

A ground based laser is the power source that propels the Lightcraft into orbit. Lightcraft can deliver payloads into space for a fraction of the cost of traditional rockets because most of the engine stays on the ground, thereby unburdening the craft from having to lift the energy source for its propulsion system.

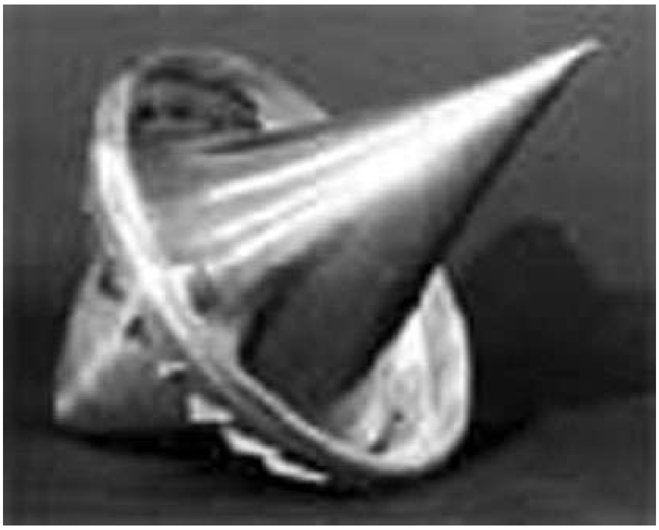

Fig. 3

The back side of the craft is a large, highly polished parabolic mirror that is designed to capture the laser beam projected at it from the ground. The mirror focuses the beam, rapidly heating the air to 5 TIMES the temperature of the sun, creating a blast wave out the back that pushes the vehicle upward. As the beam is rapidly pulsed, the vehicle is continuously propelled forward, on its way to orbit.

History of Lightcraft

1987 — Prof. Leik Myrabo invents Lightcraft for SDIO.

1997 — First successful wire-guided tests at WSMR. Solved flight stability difficulties, much like the Wright brothers did with the airplane.

1997 — Lightcraft broke Goddard's 41 ft., 1926 first successful rocket flight - but this time with no on-board fuel.

1998 — Record flight of 99 ft. with air breathing Lightcraft engine.

1999 — Record flight of 128 ft. with first rocket Lightcraft engine.

2000 — LTI sets new world record for highest flight (233 ft), longest flight time, and heaviest vehicle.

How can LTI reduce launch costs?

Conventional Launch — $175,000,000
The Lightcraft — $46,000

Chemical Rockets:

- carry massive propulsion source on board
- are expendable
- extremely costly
- prone to explosion due to fuel on board

Laser Propulsion:

- propulsion energy source remains on the ground!
- Lightcraft are inexpensive to manufacture and extremely light weight
- highly reusable power source is never subjected to the risks of flight

Lightcraft Technologies, Inc.

1914 Walloomsac Rd.
Bennington, VT 05201
Office: 802-447-6275
FAX: 802-447-8216
E-mail: tmyrabo@lightcrafttechnologies.com

Alternative space propulsion systems Star Drive

Mark R. Tomion, USA
http://www.stardrivedevice.com

Re: U.S. Patent **6,404,089** for the **Electrodynamic Field Generator**, **(EDF),**
issued June 11, 2002 to Mark R. Tomion.

The 'official' name of the "StarDrive device", per the U.S. Patent and international PCT Applications, is **Electrodynamic Field Generator.** The EDF Generator uses banks of permanent magnets and rotating Field Coils to produce a very-high DC rotor voltage, and plane-parallel ring electrode arrays to electrostatically expand and control that voltage *as applied to the hull*, so that huge quantities of external Field electrons may be accelerated to energy levels that are usually reached only with a particle accelerator! It's somewhat like a glorified arc welder whose output is deliberately shorted to its own housing, and the DC voltage and current across the emitter and collector housing sections can be thermionically increased to values that are generally observed only in lightning: but the Field's *current density* is limited to a value which falls short of damaging the hull!

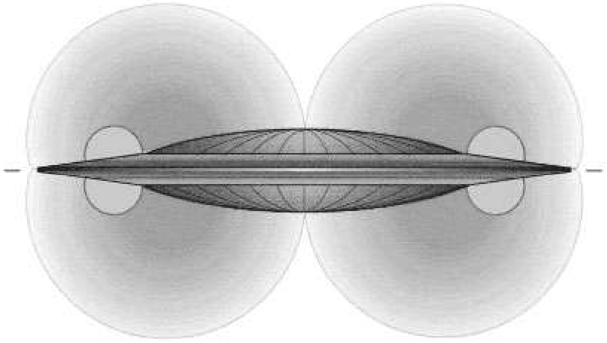

Fig. 1
StarDrive device hull & electrodynamic
Field configuration

As depicted above (Fig. 1), electrons circulating in the external Field envelope can achieve an impact velocity at the central collector sections which is very nearly that of light, and applied magnetic fields allow broad modulation of the Drive Field current's properties. The *Primary Arrays* shown in the generalized schematic diagram below (Fig. 2) have control grids which allow an arc resistance imbalance to be imparted to the otherwise symmetrical Field current, so that they *render the two relativistic current impulses variably non-isometric:* thereby yielding thrust that is essentially **reactionless**! And the simple DC *Primary Power System*, like the early Faraday disk dynamo, *is wholly rotor-based* **. . .**

Note: It can be seen that the propulsive thrust developed by a StarDrive vessel is essentially brute-force in nature – it's produced simply by means of a controlled variable imbalance in the continuous *physical impact* of the two external hemitoroidal electron current streams with the collectors! If these two Field currents were of equal magnitude, no net force would be developed. However, if the "lower" current stream is stronger than the "upper", the vessel will be propelled away from the stronger current – in the "upward" direction. Since there's no 'backward' exhaust produced in the process, this type of thrust is truly and demonstrably reactionless in nature.

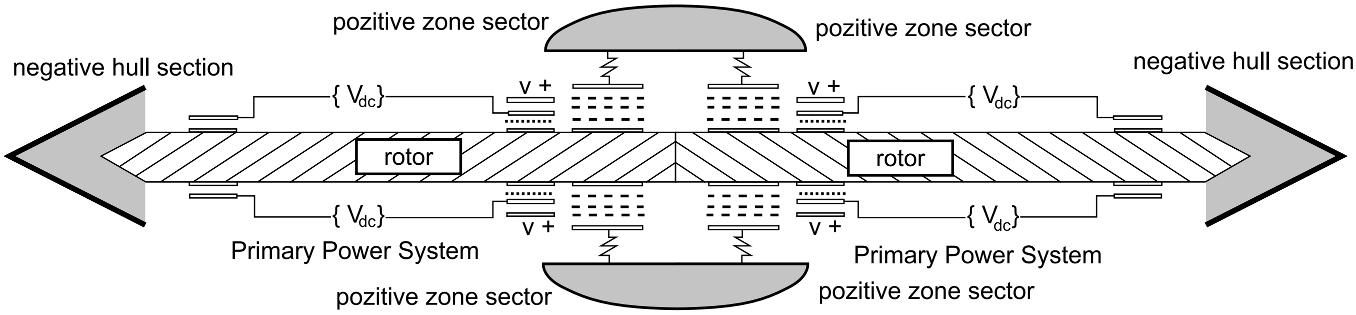

Fig. 2
Generalized schematic of rotor & dual induction ring assemblies

Needless to say, tremendous quantities of heat are produced in the StarDrive device's electron "targets" or collector housing sections, and liquid sodium must be pumped through each Primary Array's ceramic resistor network as a coolant. However, this excess heat in ground-based units may be used in the *commercial generation of electric power and desalinization of seawater*.

In fact, because an intense arc discharge field has the unique capacity to absorb vast quantities of quantum background energy, the EDF Generator is so efficient that the latter task may become truly cost-effective for the first time! And not only will large ***over-unity*** StarDrive Dynamo units be able to produce electric power at 60 to 720 MW output levels, they'll be able to do so for many years before the permanent magnet banks must be remagnetized!! *The only truly external input energy required* in the interim *is that necessary to initially bring the rotor up to speed* ***. . .***

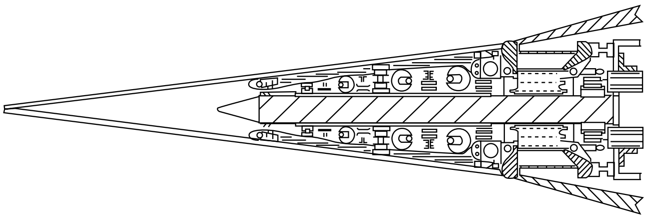

Fig. 3
A denumbered version of Fig. 1
from the EDF Generator Patent

Those of you who find this advanced technological prospect for the 21st century as exciting and fascinating as we do should consider making the inventor's book *StarDrive Engineering* a valuable addition to your personal or reference library.

A Layman's Description of the StarDrive Device

Nearly anyone who endeavors to gain an initial understanding of the *Electrodynamic Field Generator*, even a scientist or engineer, is likely to ask for a concise layman's explanation of "how it works". This isn't easily accomplished, for this device represents an unorthodox and multidisciplinary technology. However, the overview provided below will discuss the operative characteristics of the machine's simplified *electric power generation variant* in what is hoped to be the most direct manner possible. In this case, when people ask the question above, what they generally mean is: "By what means is *over-unity* operation achieved in a 'StarDrive' Dynamo?". It is assumed here that the reader has not only already become acquainted with certain basic aspects of the design, but also understands that over-unity operation is an absolute prerequisite for any viable system of light-speed interstellar propulsion.

By way of further background, the formative mechanical design basis of the EDF Generator is of course the original Faraday disk dynamo. For whatever reason, no one seems to have bothered trying to develop this simple machine into a more sophisticated *and patentable* form before now. The principal limitation of Faraday's disk dynamo in its original form was that, when one or more permanent magnets were used to pass flux directly through the plane of a solid conductive rotor, a rather high-loss heavy current at very low voltage was produced. The first design improvement was therefore to use rotor-mounted toroid field coils in order to generate a much higher-voltage primary current, and to segment the rotor to reduce "eddy current" losses. Also, the toroid coil configuration absolutely minimizes Lenz losses, or the magnetic "drag" that is experienced by any conventional generator's rotor.

To eliminate the use of brushes, it was necessary to apply traditional vacuum tube design and operating principles. Fortunately, most of the original patent work in this field has passed into the public domain, so incorporating certain aspects of that work into the EDF Generator presented no impediment to its patentability. By using plane-parallel electrode arrays instead of brushes to charge the rotor, it is possible to limit that portion of the induced rotor current which passes through the field coils to a very low level – effectively isolating them from the actual output circuit. More importantly, however, it then becomes possible to thermoelectrically charge the Generator's housing itself in such a way that *it* carries the device's full output circuit current instead!

The reason for this unorthodox design parameter is that it was desired to actually incorporate a standing electric arc field into the output circuit, to take advantage of the electron's inherent ability to absorb quantum background energy [including *zero point energy*, if and as necessary]. As a result of a thorough study of lightning, the inventor of the StarDrive device reasoned that the electrons comprising any naturally-occurring bolt of lightning had to *recover* (or absorb) an amount of ambient photonic energy equal to that which they expended in transit – in order to satisfy the conservation of energy principle. In such a case, it is not necessary that "we" do the *work* of moving charge *against* a potential gradient; the work may be done by the charge itself in being attracted *along* the potential gradient (or voltage level).

And since it is known from the field of welding that it takes less energy to sustain an arc than it does to initiate it, it therefore becomes possible to create an electrical circuit that outputs more energy than it requires as input. This is exactly what the EDF Generator does – by incorporating a standing arc field in its output circuit. Using the StarDrive device as our mechanism, "we" only provide the work-energy required to establish and maintain the external field's potential gradient, by initiating rotor rotation and bringing the thermoelectric elements up to temperature. The electrons in the electrodynamic field do all the rest . . .

Further Notes from the Inventor

In conjunction with one of our *StarDrive Generator* prototype project funding proposals, we are presently working on an interim proof-of-concept experiment for our proposed 24 kW air-cooled EDF Generator prototype. This full-scale mock-up of the 30"-dia. Generator's rotor and dual induction ring assemblies is intended to demonstrate the fundamental design principal discussed at the close of the Technical Overview (linked to our website's Method of Operation Summary page), whereby the voltage electrostatically induced on the rotor anode rings which power each Primary Array should be roughly one-third (1/3) of the Field Coil voltage (because of the capacitive dual induction ring geometry of the Primary Power System). For safety reasons, the Field voltage in all *air-cooled* StarDrive *Generators* will be limited by design to 850 VDC, and to 1,400 VDC in the larger *liquid-cooled* StarDrive *Dynamos*.

The experiment will also assist greatly in the derivation of remaining production model specifications, and will in fact incorporate production-quality rotor segments and electrode rings. Should this proof-of-concept experiment be successful, not only will the ability of our over-unity 24 kW Generator prototype to deliver large-scale DC output that's compatible with standard AC inverters (for utility grid distribution or off-grid conventional use) be virtually assured, but a major milestone incentive in our existing funding proposals will have been fulfilled as well. Further updates like those below will be forthcoming on the *News* page of http://www.stardrivedevice.com.

01/28/03 - International Patent Filings Secured!: We are very pleased to report that we were successful in our efforts to secure numerous international (PCT) Patent Application filings for the *Electrodynamic Field Generator* by the final deadline of January 21, 2003! These important filings were effected in Australia, Canada, the People's Republic of China, the European Union (including France, Germany, Italy, Spain, and the United Kingdom), India, Japan, Mexico, the Russian Federation, and South Africa.

The securing of these key Patent Application filings adds immeasurable value to our StarDrive Engineering Project overall, as it will greatly protect and enhance our investors' upside global market potential (assuming, of course, that our *24 kW StarDrive Generator prototype* is successful)! Be sure to check back from time to time for further updates as we proceed into the development phase of the Project.

11/07/02 - Academic Reference: Those of you kind visitors to our website who would like to have the benefit of an informed and unbiased academic opinion regarding the Electrodynamic Field Generator are welcome to contact John J. Tulip, Ph.D., Exec. Vice-Pres. of *American International University*. This consideration also applies of course to those parties who may be interested in securing a direct participation in our forthcoming EDF Generator Prototype Project. Dr. Tulip has not only expressed much-appreciated support of our efforts to introduce this important new technology, but has also had the opportunity to review our technical manual *StarDrive Engineering*. You may contact Dr. Tulip via e-mail at TulipJJ@aiuniversity.edu although we ask that you expect him to field serious and respectful inquiries only. Should you desire to speak with him by telephone, please be assured that he will endeavor to return your call whenever circumstances permit if you provide him with the proper phone number.

10/07/02 - Joint Venture Agreement Announced!: Mark Tomion, founder and president of Archer Enterprises and inventor of the recently-patented *Electrodynamic Field Generator*, is very pleased to announce that he has signed a Joint Venture Agreement with Affirm Technology Partners of Carlsbad, California to build a working prototype of his over-unity 'StarDrive' device's electric power output variant. The co-developers are planning to commence construction of a small air-cooled **StarDrive Generator** unit with a projected output rating of 24 kW and a housing diameter of only 30 inches, at a total weight of under 50 lbs., before the end of November 2002.

Should this exciting project be successful, it would represent an historic milestone in the development of over-unity electric power generation technology.

Not only is this small prototype EDF Generator expected to demonstrate *a minimum Coefficient of Performance in excess of 20:1*, but the design employed is completely linearly-scalable in a very broad range of sizes that would include liquid-cooled ***StarDrive Dynamo*** units with output ratings of up to 1 gigawatt! The most remarkable feature of these large Dynamo units is that an amount of recoverable thermal energy comparable to their respective electrical outputs will be made available for desalinating seawater, or for use in centralized municipal and industrial hydronic heating systems. And this capability would make the large-scale desalinization or distillation of water truly cost-effective for perhaps the first time ever.

Interested parties are welcome to contact Mr. Tomion at office@stardrivedevice.com (585-526-6817) for further information.

INFINITE ENERGY
Cold Fusion and New Energy Technology
Including Conventional and Renewable Energy

INFINITE ENERGY
The "Lifter" Phenomenon

Infinite Energy Magazine

* Cold Fusion * New Energy
* New Science
* New Technology
Subscriptions, 6 Issues Per Year
$29.95 North America
$49.95 Foreign
Single Sample Copy
$5.95 North America, $10.00 Foreign

Infinite Energy Magazine
P.O. Box 2816-FV
Concord, NH 03302-2816
Phone: 603-228-4516 Fax: 603-224-5975
http://www.infinite-energy.com

Institute for Planetary Synthesis

P.O. Box 128, CH-1211 Geneva 20, Switzerland

Tel. 41-022-733.88.76,
Fax 41-022-733.66.49
E-mail: ipsbox@ipsgeneva.com

website: http://www.ipsgeneva.com

Prospects of Hydrogen Energetics

Ph.M. Kanarev, Russia

E-mail: kanphil@mail.kuban.ru

Hydrogen is the only pollution free and inexhaustible energy carrier. But the implementation of such attractive properties of hydrogen is restrained by large expenses of energy for its production from water. Modern level of knowledge gives the opportunity to reduce these expenses [1], [2], [3].

It is known that a gram-atom is equal numerically to atomic mass of the substance, and a gram-molecule is equal numerically to molecular mass of the substance. For example, the hydrogen gram-molecule in the water molecule is equal to 2 grams, and the oxygen gram-atom is equal to 16 grams. The gram-molecule of water is equal to 18 grams. As the mass of hydrogen in the water molecule is 2x100/18=11.11% and the mass of oxygen atom is 16x100/18=88.89%, the ratio between quantity of hydrogen and oxygen is preserved in one litre of water as well. It means that 1000 grams of one litre of water contain 111.11 grams of hydrogen and 888.89 grams of oxygen.

One litre of hydrogen has mass of 0.09 grams, one litre of molecular oxygen has mass of 1.47 grams. It means that from one litre of water it possible to produce 111.11/0.09=1234.44 litres of hydrogen and 888.89/1.47=604.69 litres of oxygen. Thus, one gram of water contains 1.23 litres of hydrogen [1].

Now energy consumption for production of 1000 litres of hydrogen from water is 4 kWh and of one litre – 4 Wh. As it is possible to produce 1.234 litres of hydrogen, then 1.234x4=4.94 Wh are spent for production of one gram of water now.

Instruments and Equipment Used for the Experiment

A special experimental low current electrolyzer, a voltmeter of the highest accuracy (accuracy class 0.2, GOST 8711-78), an ammeter of the highest class of accuracy (accuracy class 0.2, GOST 871160), a balance with value of a division of 0.10 grams and 0.010 grams, and a stopwatch with value of a division of 0.1s.

Experimental Results

Indices	Sum
1 – duration of electrolyzer operation connected to the supply line, in 6 cycles t, min	6x5=30.0
2 – readings of voltmeter V, volts	13.6
3 – ammeter readings I, amperes	0.02
4 – power consumption (P=VxIxτ/60), Wh	0.136
5 – duration of electrolyzer operation disconnected from the supply line, in 6 cycles t, min	6x55=330.0
6 – solution mass change m, grams	0.44
7 – mass of evaporated water m', grams	0.02x6=0.12
8 – mass of water converted into gases m''=m-m', gram	0.320
9 – power consumption per gram of water converted into gases P'=P/m'', Wh/grams of water	0.425
10 – existing power consumption per gram of water converted into gases P'', Wh/grams of water	4.94
11 – reduction of power consumption for hydrogen production from water, K=P''/P', times	11.62
12– quantity of released hydrogen, ΔM =0.320x1.23x0.09=0.035, grams	0.035
13 – power content of hydrogen being produced (E=0.035x142/3,6)=1.397, Wh	1.397
14 – energy efficacy of water electrolysis process (Ex100/P), %	**1027**

Note: Gas output is clearly observed during many hours after the electrolyzer is disconnected from the electricity supply.

Conclusion

Low ampere water electrolysis is a way for production of inexpensive hydrogen from water and hydrogen energetics.

References

1. Ph.M. Kanarev. The Foundation of Physchemistry of the Micro World. Krasnodar, 2002. 320 pages
2. http://book.Kanarev.innoplaza.net
3. hppt://www.n-t.org/tp/ns/if.htm.

Heater Producing Healthy Effect

The First Device of Healthy Life Support in Fuel-Energy Industry

A. Belyaeva, L. Savelyeva, O. Bondarenko, Kirghizia

http://www.leteco.h1.ru
Email: leteco@mail.ru

Nowadays problems of rational use of resources including energy resources are very important. Lack of energy-efficient technologies causes consumption of great amount of electric power, coal, and mineral oil. From another hand, in the XXI century one of the main approaches of human society development is turned to healthy life-style and development of ecologically appropriate technologies. Hence, basic direction of heating systems development should correspond to, at the least, two requirements, i.e. effectiveness (energy-saving) and ecological appropriation (ecological functioning). Moreover, it is necessary to take into account additional requirements for modern works:

- Relatively low prime cost of the systems;
- Minimal operating costs;
- Availability of a system of temperature mode control;
- Use of domestically produced materials for the systems.

The aim of the work is creation of a universal electric device of natural ecologically appropriate materials. The device should maintain relative air humidity of natural atmosphere in a quarter, make its ecology healthier, meet all the listed requirements, and represent new generation energy-saving system designed for heating of industrial and domestic areas according to features of technical solution.

A.L. Belyaeva is the author of this work. The invention of this heating device was acknowledged as the best invention of Kirghiz Republic of the last two years. A.L. Belyaeva was called as a laureate of a competition of "The Best Inventing in Kirghiz Republic in 2001-2002".

Solving this problem the author based on work experience existing in the area of semi-conductors production. Actually, the model was worked out in the common area of electric engineering and industrial crystals growing. Using of knowledge and skills of the both areas made creation of this device possible.

It should be noted that initially the invention of the heating device, which is discussed here, was connected with necessity to improve microclimate and air atmosphere in a semi-conductor shop since the industrial process required optimal conditions. Hence, the device was observed as an element of the industrial process. However, soon other positive characteristics of the device were disclosed, i.e. when it was installed in a quarter all visible mould disappeared. Laboratory investigations stated that the operating device annihilated all pathogenic microflora. This rediscovered feature allowed using the device both in domestic and industrial area. After series of following laboratory investigations and bench tests construction of the device was changed and improved that allowed prepare it for line production.

Ceramic electroconvector: general characteristics

The working name of the final version of the model is *Ceramic Electroconvector TY 2971-006-22997241-2002.* Ceramic Electroconvector is an industrial and domestic electric heater of direct stationary action. It has high effective heat emission and satisfies ecological, sanitary, medical, and fire-prevention requirements.

At the same time it should be mentioned that by several characteristics this electroconvector differs from other known models of electroconvectors as well as from other existent heaters. It may be observed as a representative of an independent, specially created group of heating devices.

One of the main characteristics of the electroconvector is presence of constructive heating carbonic elements made of ecologically appropriate natural *non-metal* materials. Essentially, the electroconvector construction does not contain metal (the only metal part of the device is its supporting construction).

Production of heat of physiologically comfort zone requires 0.3 kW/h energy consumption which is 3-10 times less than energy consumption of known models of heating engineering. The Ceramic Electroconvector influences positively on ecology of a room. Room heating occurs better and more softly as conducting carbonic elements can be heated maximally up to 100° C. As a result oxygen is not burnt, and air is not overdried in a room. Achieved minimization of metal content in the constructive elements of the device increases the level of ecological compatibility both of electroconvector constructions and operation.

The device accumulates no static electricity, neutralizes harmful magnetic field generated by alternating current in the conducting element (it is typical for all the other electrical household appliances). Therefore, **the additional positive effect** is produced and higher ecological characteristics of functioning of this electroconvector are confirmed.

Insulating strength of ceramics prevents electrical shock accidents. Ecological compatibility of the electroconvector is provided by materials of the construction. The base of the device is a studied natural silicate fiber which has quantitative and qualitative content of useful chemical elements which are the closest ones to the group of medical adsorbents listed in a medical encyclopedia. The ceramics is adjusted to emit electromagnetic waves only in infra-red spectrum.

The electroconvector produces heat waves in the average IR spectrum (8.4-8.6 mkm) which is maximally approximated to the diapason of heat waves generated by a human (9.37 mkm). It annihilates humidity of buildings independently of outer space humidity. At the same time it neither burns oxygen nor overdries air.

Healthy effect is produced by all the constructive elements made of ecologically appropriate natural materials. Electric power is transformed into heat emission by conducting elements. This process causes a mode of generation of a continuous heat spectrum of radiation. The heat radiation is similar to heat spectrum of radiation generated by a human. At the same time, this feature together with resonant oscillations of the crystal lattice of the ceramic cylinders produces a destroying effect on pathogenic and conditionally pathogenic microorganisms.

From the point of view of room ecology the proved healthy effect of the operating electroconvector becomes very significant. The effect considers continuous presence of a human in the room, i.e. risk of pathogen infection through respiration objectively decreases.

Application of the ceramic electroconvector is especially actual in patient care institutions and children's institutions, in special precision industries and space technologies. It can also be used in saunas with dry vapor.

The ceramic electroconvector is designed for unsupervised continuous work.

Structure of the electroconvector producing the healthy effect

Shortcomings of known electric heaters are: great electric power consumption (0.75-3.0 kW/h), big number of metal details, complex technical performance, and use of a necessary additional blower as a ventilator. Big number of metal details decreases ecological compatibility and productivity of the heaters. The listed devices have to use high temperatures on heat-release surfaces for warming up rooms up to the level of physiological comfort. This causes increase of energy consumption. Moreover, using of metal heating elements influences on air and relative humidity in the room. According to available data, no existent electroconvectors have healthy influence on ecology of a room.

The represented ceramic electroconvector contains a carrying frame with horizontal boards which have convective gaps. The carrying frame has heat-release monolithic hollow ceramic cylinders whose walls contain longitudinal through holes. Heating carbonic conducting elements and de-energized carbonic rods are built in the through holes. The heating conducting elements are connected in parallel-series circuit at the output of the cylinders with it's ends placed into insulating supports of the carrying frame.

Constructive heating elements are the main differences of the device. All the heat generating constructive elements, i.e. conducting and de-energized heating elements, as well as heat-release surfaces of ceramic cylinders have contiguous spectra of infra-red radiation.

Energy-efficient effect

The electric scheme of connection of the conducting elements provides different modes of work of the device. According to these modes, energy consumption is in the interval of 0.05-0.3 kW/h. The mode of 0.05 kW/h is calculated on a supporting level of warming-up of a room. Maximal energy consumption (0.3 kW/h) corresponds to the superior limit of the temperature mode of heating of working heat-release surfaces of the ceramic cylinders. In this case temperature of the heat generating elements, i.e. de-energized carbonic rods and conducting carbonic elements) lies in the limit of max 100°C. This produces a significant potential resource of electric strength and of durability of the used elements. The de-energized carbonic rods function as heat accumulators at switching on and switching off the device. Heating the conductors the de-energized rods accumulate heat through the ceramic walls of the cylinder till their temperature becomes equal to the temperature of the conducting elements. Appearing electromagnetic resonance between the conductors and the de-energized carbonic rods intensifies infra-red radiation of the ceramic wall. Accumulative heat potential of the de-energized carbonic rods allows maintain uniform radial heating of the heat-release ceramic surface of the cylinder without decreasing energy consumption. In this case additional conducting elements of infra-red radiation become unnecessary.

The heating efficiency of the ceramic electroconvector was estimated by independent experts. Surface density of the radiation flow was calculated by formula of Stefan-Boizmann distribution law. Taking into account heating of the cylinders' surfaces up to 70° C this value came to 727 W/m^2. **Total heat generated by the electroconvector per hour comes to 600 kcal or 698 W** at the total area of the radiation surfaces of 0.96 m^2 (on the basis of 1 kcal=1.163 W/h according to [1]). The electroconvector consumes 300 W/h and produces 698 W/h. That is to say that it effectively transforms electric energy into heat energy. Operational modes of the electroconvector are based on analysis of the heating effect produced by different devices, i.e. a tube metal heater (TMH) having a conducting metal element of Nichrom, a ceramic cylinder having a Nichrom conducting element, and a ceramic cylinder having a conducting element of carbonic ribbon. The carbonic ribbon produces high heating efficiency. Temperature on the surface of the ceramic frame is up to 80° C and temperature on the conducting element is 100° C. Hence, the device equipped with the conducting element of the carbonic ribbon does not produce excessive heat radiation.

The electroconvector consumes 300 W/h and produces 698 W/h.

After one-year operation of the preproduction models in a private school it was noticed that number of respiratory illnesses and influenza among the pupils slumped, condition of skin became better, attacks of bronchial asthma among the teachers stopped, and allergic itch left the patients who suffered from allergy. A side effect was that flowers began to grow better, and those which had not blossomed began to blossom at least. Obviously, a combination takes place there: air cleaning of microorganisms, humidity normalization, warming comfort, and influence of pyramidal ceramic structures on water vapors which are transported by warm blasts.

Applied Know How

Ambient air in rooms is a complex substance including various chemical compounds, ions, dust parts, water vapors, infectious and potentially infectious microflora, etc. The electroconvector is represented by a ceramic hardphase crystal structure. Cold air blast moves close to the ceramic surface. It meets combined oscillation of crystal microlevel structures of ceramics. *Filtering electromagnetic screen* appears that breaks shell of pathogens and potential pathogens. It is a performance of **sanitation properties**. Atmospheric water contacts with structures of ceramics, then it is cleaned and structured. After that water cleans air which is breathed in by a human. Hence, the organism becomes healthier. It is a performance of the **healthy effect**. A clean room warmed uniformly improves human's health.

Main Performance Attributes of the Ceramic Electroconvector Producing Healthy Effect

Nominal required power	0.05-0.3 kWt/h
Voltage of the feeding network	220 ± 22 V
Frequency of alternating current	50 Hz
Temperature of the conducting element in operation mode	50…110° C
Temperature of the heat-release surface of the cylinder	39…90° C
Surface density of radiation flow	727 Wt/m^2
Quantity of radiated heat	600 kcal/h
Class of protection	1
General area of heat-release surfaces of the cylinders	9600 cm^2
Sizes	410 x 400 x 75 mm
Weight	18-21 kg
Durability	30 years

The ceramic electroconvector is designed as a floor construction for repetition work.

References

1. Heat Engineering Reference-book, Moscow, Energy (Energia), 1975, V. 1, p. 12.

About the author

After graduating from Polytechnic University of Frunze city Alexandra L. Belyaeva (1953) participated in launching of semiconductor production. She was given a rank of Perfect Inventor and Rationalyzer for invention and rationalization action. In 1987 her invention named Method of Production of Monosilicon Seed Crystals was introduced into manufacture, according to close copyright of USSR. Since that year this introduction has allowed grow monosilicon having better semiconductor properties. She was given a rank of Inventor of USSR for the invention.

Alexandra L. Belyaeva has acknowledged and given a rank of Inventor of Kirghiz Republic. Since 1994 in collaboration with Russian specialists of ceramics she has been dealing with organization of ceramic productions in Kirghizstan. Here are the inventions by Belyaeva: Method of Production of Seed Crystals and Its Variants (Patent #42 KG), Ceramic Mass for Insulators' Production (Patent #43 KG), Ceramic Mass Having Heat-Radiating Properties (Patent #464 KG).

In 1992 Belyaeva was rewarded with a Silver medal of VDNH (Exhibition of Achievements of National Economy, Moscow) for development of technologies for production of constructive nitride-cadmium items of semiconductor wastes.

Her Method of Production of Nonexpendable Seed Crystals is at introduction stage. The seed monocrystals produced by the new method will allow grow ribbon monocrystals having certain properties. It will be possible to produce items made of the monocrystals which will meet requirements of energy-saving and ecological appropriation. These devices represent non-aging powerful energy sources, various semiconductive devices with no inner microdeffects and with unlimited durability, structural water filters.

Since 2002 she has been dealing with organization of production of ecologically appropriate industrial-domestic appliances causing healthy effect. The invention of Ceramic Electroconvector Producing Healthy Effect is now introduced. Energy-saving of the new ecologically appropriate electroconvector exceeds energy-saving of all the existent heating systems.

Belyaeva is a laureate of competition of The Best Inventing in Kirghiz Republic in 2001-2002.

Inertial Propulsion Device

Vitaly E. Senkevich, Russia

E-mail: hhhhha@mail.ru

A body is at rest or moves linearly and uniformly until it is not influenced by an external force.
(School course of physics)

Forget everything that you were taught at school.
(Arkady Raykin)

This propulsion device consists of an engine and a body. The engine (see Fig.1) is remarkable for its movable stator (S) which can make free rotation as well as a rotor (R). When the engine starts the stator and the rotor begin to rotate in opposite directions. Thus the engine has two ends and one of them is connected to a flywheel (F). This flywheel begins accelerated rotation.

A cross-beam (CB) is connected to the second end of the engine where a rotating momentum appears. Under the influence of this momentum the cross-beam presses down one of the supports placed on the body (for example, Sup.1). As a result a force which is compensated by acceleration of the flywheel (F) appears on the support. When the flywheel is accelerated up to certain speed a control system (CS) switches the windings of the engine to change the direction into the opposite one (reverse). At that the cross-beam also tends to turn to the other side and presses down the second support (Sup.2). Thus the cycle repeats. It should be noted that forces acting at the supports are **codirected** and they move the whole device.

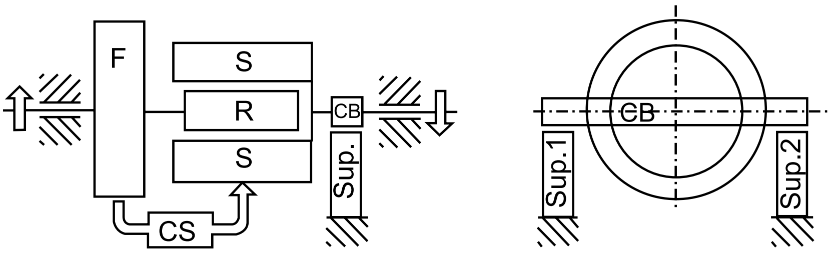

Fig.1

About the author

After graduating from Polytechnic University of Frunze city Alexandra L. Belyaeva (1953) participated in launching of semiconductor production. She was given a rank of Perfect Inventor and Rationalyzer for invention and rationalization action. In 1987 her invention named Method of Production of Monosilicon Seed Crystals was introduced into manufacture, according to close copyright of USSR. Since that year this introduction has allowed grow monosilicon having better semiconductor properties. She was given a rank of Inventor of USSR for the invention.

Alexandra L. Belyaeva has acknowledged and given a rank of Inventor of Kirghiz Republic. Since 1994 in collaboration with Russian specialists of ceramics she has been dealing with organization of ceramic productions in Kirghizstan. Here are the inventions by Belyaeva: Method of Production of Seed Crystals and Its Variants (Patent #42 KG), Ceramic Mass for Insulators' Production (Patent #43 KG), Ceramic Mass Having Heat-Radiating Properties (Patent #464 KG).

In 1992 Belyaeva was rewarded with a Silver medal of VDNH (Exhibition of Achievements of National Economy, Moscow) for development of technologies for production of constructive nitride-cadmium items of semiconductor wastes.

Her Method of Production of Nonexpendable Seed Crystals is at introduction stage. The seed monocrystals produced by the new method will allow grow ribbon monocrystals having certain properties. It will be possible to produce items made of the monocrystals which will meet requirements of energy-saving and ecological appropriation. These devices represent non-aging powerful energy sources, various semiconductive devices with no inner microdeffects and with unlimited durability, structural water filters.

Since 2002 she has been dealing with organization of production of ecologically appropriate industrial-domestic appliances causing healthy effect. The invention of Ceramic Electroconvector Producing Healthy Effect is now introduced. Energy-saving of the new ecologically appropriate electroconvector exceeds energy-saving of all the existent heating systems.

Belyaeva is a laureate of competition of The Best Inventing in Kirghiz Republic in 2001-2002.

Inertial Propulsion Device

Vitaly E. Senkevich, Russia

E-mail: hhhhha@mail.ru

A body is at rest or moves linearly and uniformly until it is not influenced by an external force.
(School course of physics)

Forget everything that you were taught at school.
(Arkady Raykin)

This propulsion device consists of an engine and a body. The engine (see Fig.1) is remarkable for its movable stator (S) which can make free rotation as well as a rotor (R). When the engine starts the stator and the rotor begin to rotate in opposite directions. Thus the engine has two ends and one of them is connected to a flywheel (F). This flywheel begins accelerated rotation.

A cross-beam (CB) is connected to the second end of the engine where a rotating momentum appears. Under the influence of this momentum the cross-beam presses down one of the supports placed on the body (for example, Sup.1). As a result a force which is compensated by acceleration of the flywheel (F) appears on the support. When the flywheel is accelerated up to certain speed a control system (CS) switches the windings of the engine to change the direction into the opposite one (reverse). At that the cross-beam also tends to turn to the other side and presses down the second support (Sup.2). Thus the cycle repeats. It should be noted that forces acting at the supports are **codirected** and they move the whole device.

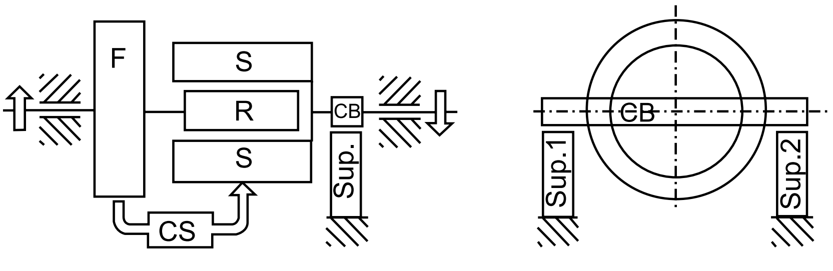

Fig.1

Unlimited Accelerated Nonreactive Motion

by G.P. Ivanov

Information from http://tts.lt/~nara/ruspopul.htm

Bias of centre of inertia of the closed system is a phenomenon exceeding the bounds of modern scientific conceptions. At the same time it is a way to the new marvelous world of earlier unknown laws and phenomena of nature. Researches made by G.P. Ivanov, Russia, have allowed him come to the following conclusion: it is a quite realizable task to create technical devices which can move under the action of nonreactive forces. However a purposeful scientific-research spade-work is required for the reliable registration of them by modern experimental facilities. According to the author during the whole XX century the known idea of latent impulse were misleading the scientists from the serious research of impulse-energy processes existing in systems at the presence of quasi-stationary electric and magnetic fields. According to G.P. Ivanov, it is related with the fact that the notion of "latent impulse" has nothing in common with the real momentum, since the very existence of "latent impulse" and "latent energy" which attends it, would make it impossible, for example, to adjust radio equipment since all effective capacitances change their value a hundred and thousand times as much (it depends on their orientation with respect to the magnetic field of the Earth).

Fig. 1, 2 demonstrate the patented by G.P. Ivanov method of realization of nonreactive motion (G.P. Ivanov, Yu.G. Ivanov. Method for production of propulsion. Patent #2172865, M., 2001). Fig. 1 demonstrates a device which consists of magnetized core with the attached metal electrodes.

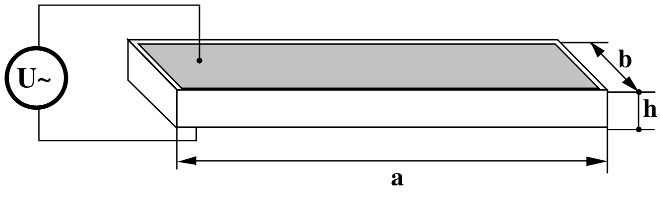

Fig. 1
The simplest "nonreactive" element

When alternating voltage is applied the device together with the center of inertia of the whole system (including power source and lead) will oscillate under the action of nonreactive force. It will move along the direction which is perpendicular to the vectors of electric and magnetic fields inside the core.

On Fig. 2 there is an analogous device supplied with a cylinder core. Magnetization of the core is defined by current of the coil which is wound around it (see Fig.2).

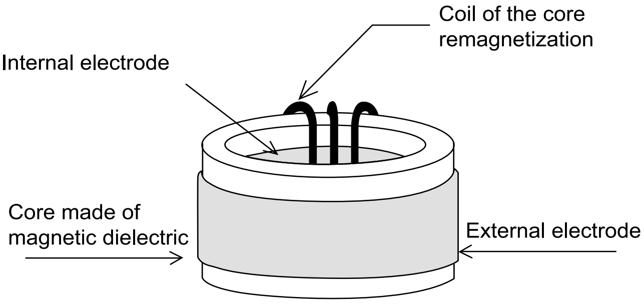
Coil of the core remagnetization
Internal electrode
Core made of magnetic dielectric
External electrode

Fig. 2
Diagram of the device which can be in unlimited accelerated nonreactive motion

Let us supply the electrodes of the device with such voltage which is enough to change field density in the core (for example, according to the law $E = E_0 \sin\omega t$), and the coil with voltage enough to change the magnetization (according to the law $M = M_0 \cos\omega t$). Then nonreactive force which is constant in its direction will act on the device along the axis of the cylinder. The average of the force is $F_{cp} = \pi f E_0 M_0 V/c^2$, where E_0, M_0 are the amplitudes of density values of the electric field and amplitudes of the core magnetization, **f** is frequency, **V** is core capacity. As a result the device can either move with acceleration or act against external forces. The researches allow the author to prove the validity of the following theorem: **At motion of the open-loop system (device) the work made by nonreactive force could not be realized by means of decrease of energy of the proper (appurtenant to the system) power source.** Where this energy comes from if there are no artificial power sources outside the device? However everywhere there is such form of matter as electrovacuum. It allows us come to a conclusion that nonreactive forces make work by means of decrease of electrovacuum energy.

Existence of electrovacuum does not contradict to the modern physical picture of the world. On the contrary providing the realization of laws of momentum and energy conservation, this idea originates from and organically supplements it. The theory of electrovacuum opens quite realizable prospects in different fields of human activity which seem to be fantastic and impossible. It also proposes unusual ways to solve many problems which are considered as insoluble. For example, it becomes possible to create nonreactive cosmonautics and private aircrafts. Energetics will be changed beyond recognition since power sources which require no material fuel will occupy the place of big and small modern electric power stations. The analogous list could be continued. There are presented quite realizable prospects related to inexhaustible reserves which are hidden in the ambient space.

Revolutionary Current

Suresh Kumar Baliyan, India

E-mail: suresh_baliyan@rediffmail.com
suresh_baliyan@yahoo.com

*Electricity is one of the most wonderful inventions of the 18th century. It can be produced by various methods in which different types of energy are used such as mechanical or chemical energy. In this article we discuss a new type of methods in which **electrostatic energy** is used.*

In this method we use **electrets** as an energy source. Electrets are permanent polarized dielectric material which is made by cooling dielectric material in a high intensity electric field. When we place an electron in the electric field of a point charge, it is influenced by a force. If the direction of the force is such that the electron moves in a closed path then it represents a current in the opposite direction along the closed path. Here closed path means a metal wire loop in which the electron movement is responsible for the current.

The charge of the electrets remains constant for many years (100) and there is no loss of energy because we use only the property of electrets that they apply force when a charge is placed in its electric field. Since here we create energy in the form of electric current then it violates the Law of conservation of energy.

Let us discuss the method in three steps:

1) Fundamental
2) Equivalent circuit
3) How we can use it as a potential source of electricity.

1. When we put a metal rod in the electric field of a point negative charge then the electron is influenced by repulsion force which moves in arrow direction and the electric field is cut by putting an earthed metal plate from any direction. When we place four charges on the corner of the metal frame and shield the electric field in particular direction as shown in Fig.1 then the force on the electron which is placed in metal frame will be in the direction of arrow and this makes a loop. The force influencing on the electron will be continuous, so the current which moves in direction opposite to the electron movement will induced in the metal frame.

2. The equivalent circuit of Fig.1 is shown in Fig2. It demonstrates that a 'V' volt battery and a resistance 'R' can be replaced by an arm of the metal frame. The polarity of the battery is shown in Fig. 1. A current of magnitude (V/R) will be induced in the circuit by this method.

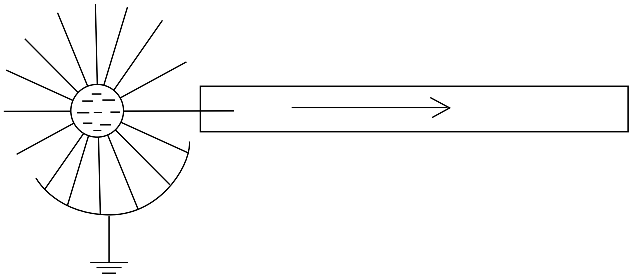

Fig. 1a

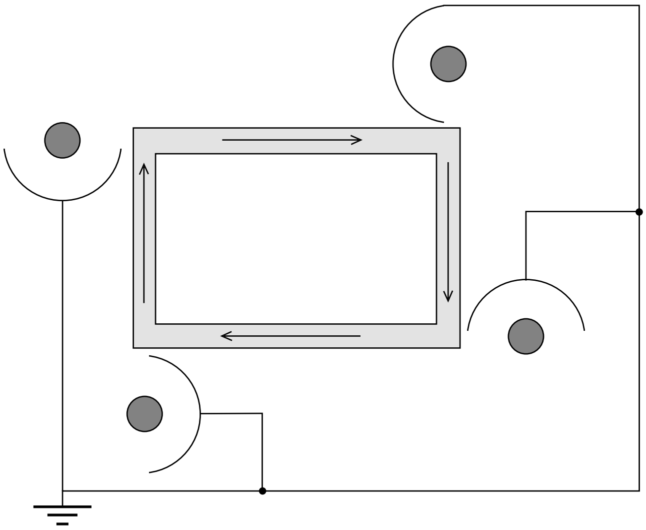

Fig. 1b

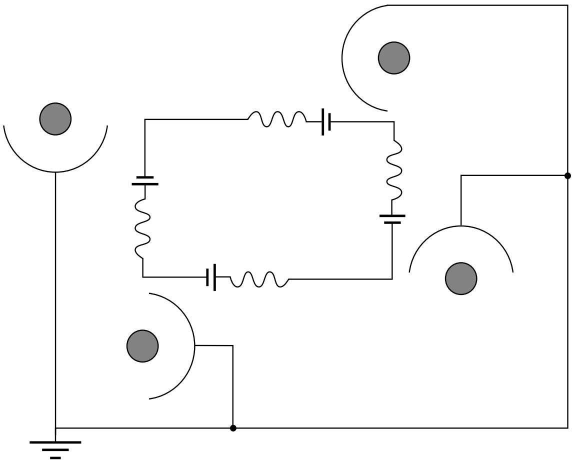

Fig. 2

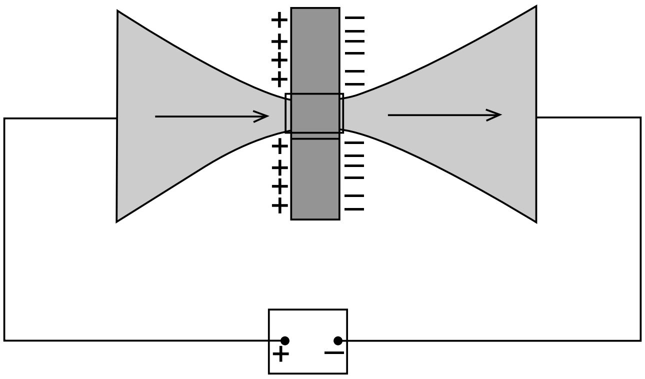

Fig. 3

3. We can use another device as a potential source (see Fig3.). Let us take an electret and make a hole at the centre. A hollow thin walled metal cylinder is fitted in the hole and a metal frame is taken as shown in the Fig. 3. We can use this metal cylinder as an electric field shield which may affect the direction of the force influencing on the electron. We may use this device as a potential source of electricity which will give electricity until the charges on the electrets vanish. The charge on the electrets remains constant for many years. Thus we can create energy and the problem of energy shortage can be solved by this method forever. We are trying to make a prototype of this but due to lack of equipment facility we are not able to demonstrate a working model. However we are trying to modify it.

NEWS REVIEW

Russian scientists are preparing new application of energy of geothermal waters

http://www.sciteclibrary.ru

A way for economical use and utilization of energy of geothermal waters was found in the Institute of Geothermal Problems (Russian Academy of Sciences). The new technology can be applied for heating in different purposes, and can compete with the power engineering, which uses the conventional energy carriers.

Geothermal springs are ideal for generating electricity by means of using of their permanent heat. This is especially important for Kamchatka, which is rich in these springs; however they are almost never used in mass power engineering, but only serve as an attraction for tourists, who admire beauty of Kamchatka geysers. At the same time the region itself freezes every year because of shortage of fuel for electric power stations.

What prevents us from using the geothermal springs?

The problem is that the storages of most of geothermal water fields have low and medium temperatures. This does not allow providing their competitiveness as regards conventional energy carriers. During tens of years this fact was an obstacle to the development of geothermal energy.

A solution was found in the Institute of Geothermal Problems. It was brought out that at many exploited geothermal fields wellhead overpressure exceeds 5-10 MPa and more. These waters contain fair quantity of dissolved organic gases. The scientists have found out that methane content of these waters exceeds 90 %. And until now at exploitation of the fields these types of energy have not been properly utilized.

The scientists of the Institute of Geothermal Problems have developed a technology of the optimum utilization of energy of geothermal waters and have increased the thermodynamic effect of the process.

This aim was fulfilled by means of transmission of thermal energy of geothermal water through the intermediate heat exchangers to the secondary heat carrier. Chemical energy of dissolved gases was used as an additional source of energy. The transmission was held by the use of primary and secondary separators. The distinctive feature of their technology is that associated potential energy of geothermal water is used as an additional source of energy. Expander and compressor placed on the same shaft are used as potential energy converters. Gas-holder and gas control point are used for utilization of energy of dissolved gases.

Thus on the bases of this technology geothermal energy can be effectively developed that will be competitive with fuel hydrocarbon energy.

Tilley Electric Vehicle

Tilley Foundation, Inc.

131 Hiwassee Road
Lebanon, TN 37087
http://www.tilleyfoundation.com
E-mail: cktilley@bellsouth.net

Editor: The advantages of electric vehicles were proved more than once. No fuel is required to set them in motion. Along with evident advantages of electric vehicles, such as their powerful characteristics and harmlessness for environment there is a side benefit, i.e. much less maintenance as compared with gasoline or diesel-fueled vehicles is required. Moreover it is easier to manufacture such vehicles.

In the previous issues of New Energy Technologies magazine we have already acquainted our readers with Tilley Electric Vehicle designed by Carl B. Tilley (USA). Below there is some up-to-date information and photos from the inventor. Besides we also publish an article devoted to the similar types of fuel-less motors.

After several years of personal accomplishments in the alternative energy industry, Carl B. Tilley was convinced that it was possible to build an electric car that could be powered without the help of external power to keep the battery charged.

The concept to produce a useful electric performance car that would last more than a few hours and would be economical to run, safe to drive around town or across the United States and never use a drop of fuel challenges the future of transportation as we know it today.

With the establishment of the Tilley Foundation, Inc., in the year 2001, Carl Tilley set out to prove it could be done. It was an ambitious project and it broke ground on the facility in Tennessee that would build the first self generating electric car.

...you have no need for fuel and you do not have to stop the vehicle to charge it after driving.

Construction of a 1,800 square foot building, that was powered with another recently developed electric device, began in the year 2002. Electricity from the building built the car from a different energy invention, that was void of any outside power supply. It is ironic that one alternative energy device actually built the invention to power and build the electric car.

From the selection of the proper car to be converted, to the advanced technology which is on board , the **Tilley Electric Vehicle TEV** performs comparably to gasoline powered vehicles. The difference is you have no need for fuel and you do not have to stop the vehicle to charge it after driving. There is no pollution and you can cruise the highways at the same speed as any other vehicle.

The rear mounted electric motor provides over 130 SAE net horsepower at 5,500 rpm. It offers a 3 speed automatic transmission that is smooth shifting and totally silent running. All this is combined with rack and pinion steering and a 35/65 rear weight bias that enables fast, sensitive handling and needs no power assistance.

...battery system will be fully charged at all times while in use.

It has 4 wheel disc braking for fast progressive, fade-free stopping. Counterbalanced gull-wing doors need only 14 inches of clearance. The rear sporty louvers are aerodynamic designed so that it almost eliminates any drag effect. All this with a stainless steel body makes for a great car.

Control center for the battery bank only allows what is needed to keep the batteries charged while in operation no matter what the speed or discharge from the battery bank. Your battery system will be fully charged at all times while in use. Simply get in, start the car and drive like any other vehicle.

A 1981 DeLorean was converted as the Tilley Electric Vehicle.Conversion of the car began in late June of 2002. State of the art metal fabrication to construct support for the electric motor, battery bank, control center and the TEV device was completed in July of 2002.

Several tests were made to validate the TEV technology. One of the last tests was made on September 7th, 2002. It has been demonstrated that after 17.3 miles driven on the Superspeedway at speeds ranging from 80MPH to 96MPH independent engineer certified batteries were full.

Assembly Pictures (see also the cover page)

Photos by Robert Gaither

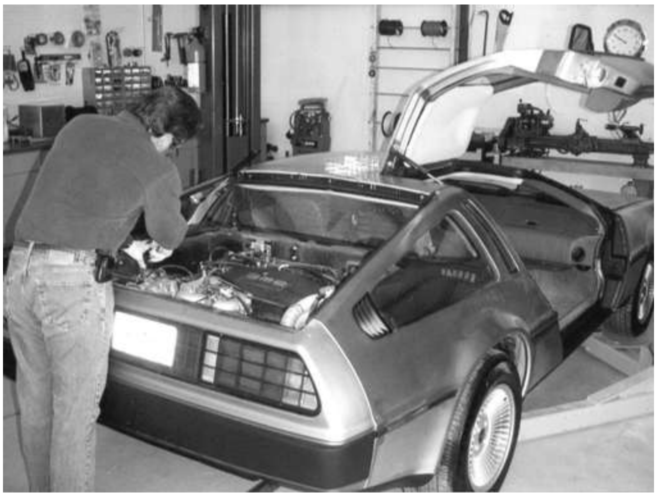

Fig. 1
Removing Gas Motor

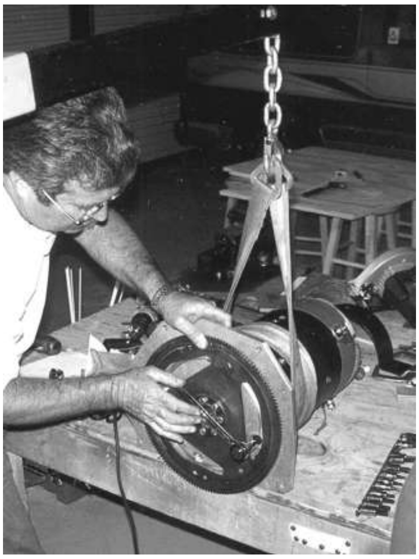

Fig. 4
Fitting Parts

Fig. 2
Fabrication

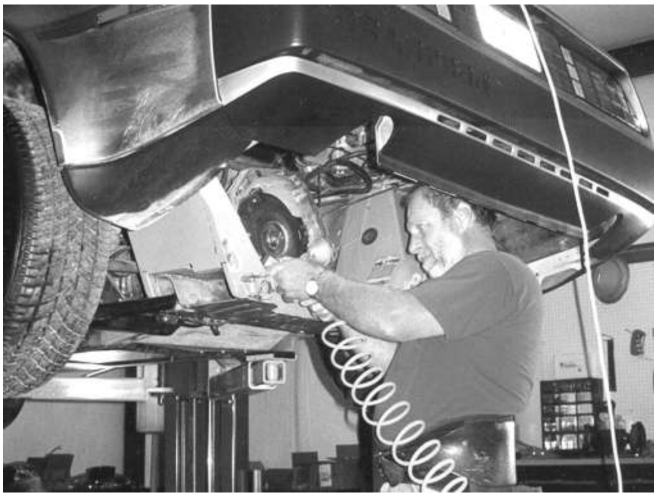

Fig. 5
Transmission Work

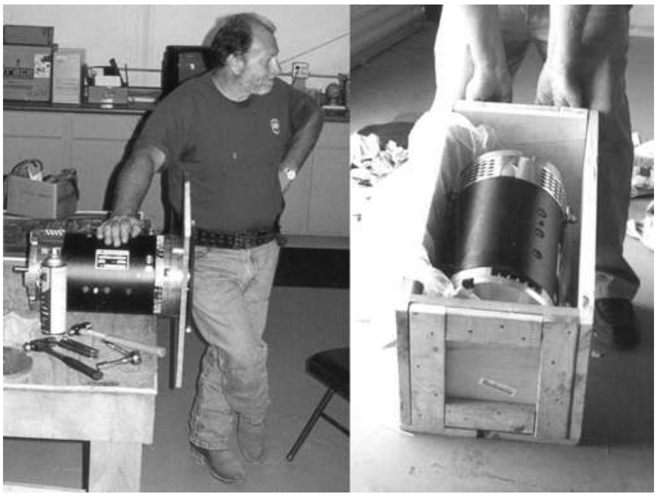

Fig. 3
Drive Motor

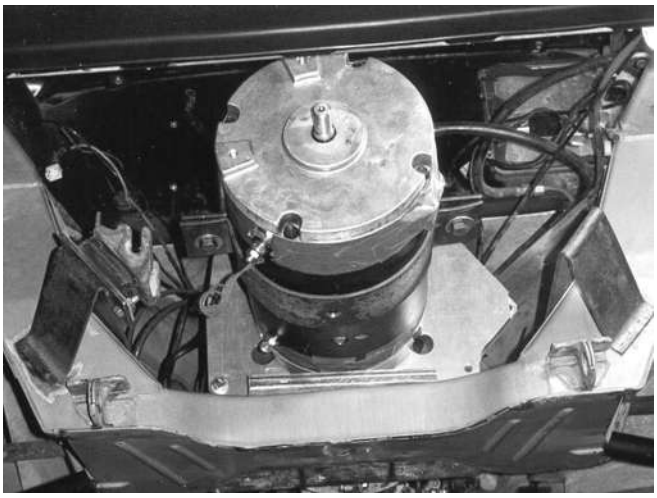

Fig. 6
Motor Mounted

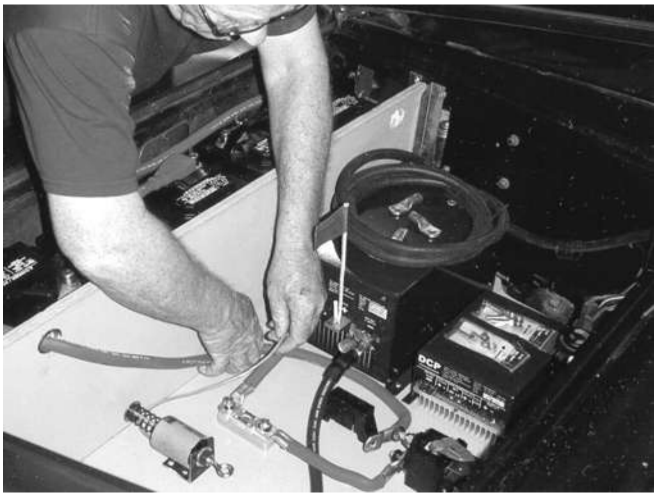

Fig. 7
Making Connections

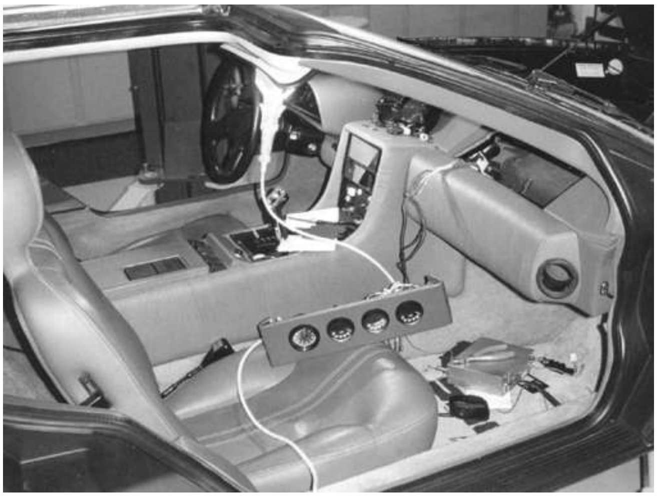

Fig. 9
Special Wiring &Guages

Fig. 8
About Ready

Fig. 10
Car Unveiling

Trends in the Application of Motor-Generators

Adrian Akau, USA

E-mail: adrianakau@aol.com

If the Tilley generator-motor should enter into the world market, some curious consequences may result, the most important being the transition of the Honda and Toyota hybrid cars into full fledged electrics.

First of all, a comparison should be drawn between the Tilley and the two hybrid cars presently being sold, the Honda Civic and the Toyota Pirus should be made.

The Honda Civic and the Toyota Pirus both use charging systems with a gasoline motor for better milage. The Tilley motor-generator is a stand-alone electric vehicle. The Honda motor-generator is just 60mm thick and provides (10kW or 13hp). The Toyota Pirus has a 44hp unit (American Version) which feeds power into electric motors at the wheels. Both Honda and Toyota gasoline motors are off at 0 mph. Only the electric system is used until the power demand reaches 10 kW; then the gasoline motor automatically kicks in. The Pirus is able to get higher milage in city driving than in country driving from the fact that the ratio of the power from the electric part to power to the gasoline part of the motor is greater at lower than at higher speeds, that is, less power is needed to run the car in city driving than in country driving because of the lower speeds and air resistance.

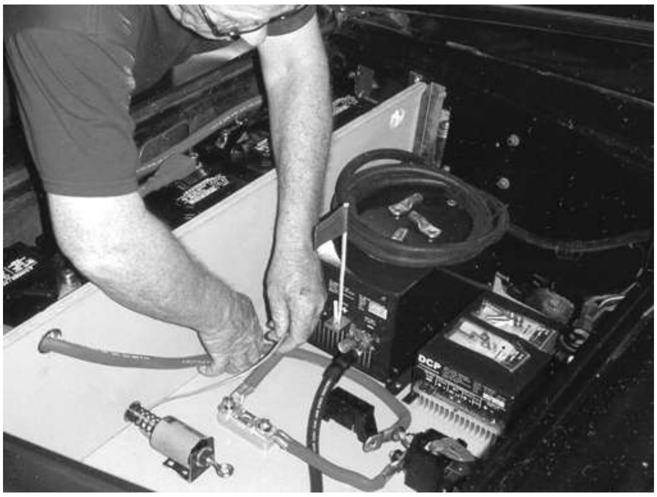

Fig. 7
Making Connections

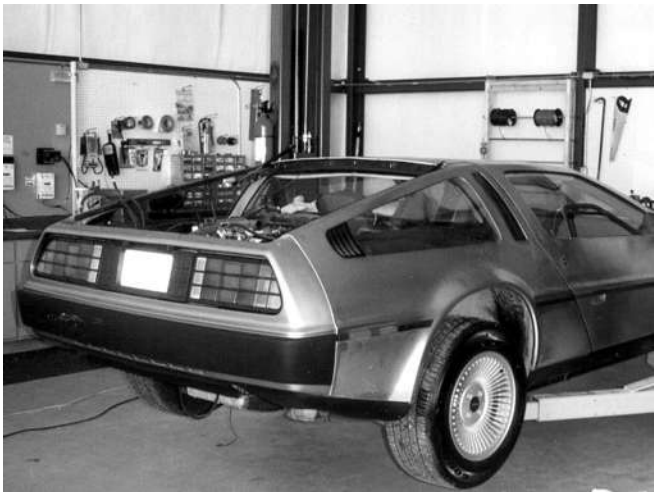

Fig. 8
About Ready

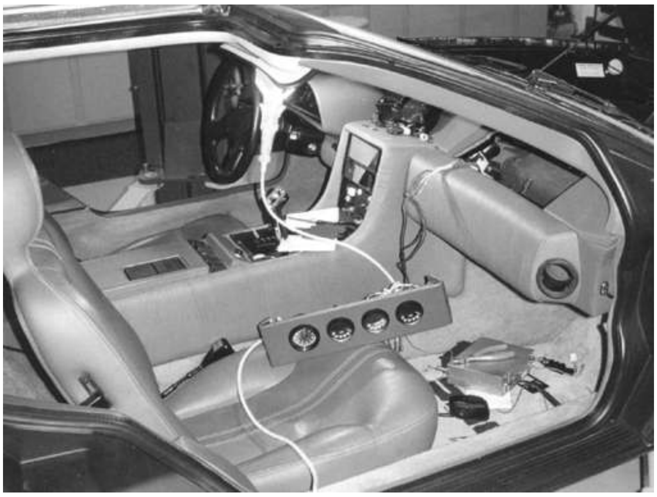

Fig. 9
Special Wiring &Guages

Fig. 10
Car Unveiling

Trends in the Application of Motor-Generators

Adrian Akau, USA

E-mail: adrianakau@aol.com

If the Tilley generator-motor should enter into the world market, some curious consequences may result, the most important being the transition of the Honda and Toyota hybrid cars into full fledged electrics.

First of all, a comparison should be drawn between the Tilley and the two hybrid cars presently being sold, the Honda Civic and the Toyota Pirus should be made.

The Honda Civic and the Toyota Pirus both use charging systems with a gasoline motor for better milage. The Tilley motor-generator is a stand-alone electric vehicle. The Honda motor-generator is just 60mm thick and provides (10kW or 13hp). The Toyota Pirus has a 44hp unit (American Version) which feeds power into electric motors at the wheels. Both Honda and Toyota gasoline motors are off at 0 mph. Only the electric system is used until the power demand reaches 10 kW; then the gasoline motor automatically kicks in. The Pirus is able to get higher milage in city driving than in country driving from the fact that the ratio of the power from the electric part to power to the gasoline part of the motor is greater at lower than at higher speeds, that is, less power is needed to run the car in city driving than in country driving because of the lower speeds and air resistance.

American car companies are beginning to develop their own starter-generators to "save an extra 10% on the gas milage". Are they doing this because they don't want to be left behind in the milage competition or because they really understand the nature of the starter-generator motors used by the companies producing the hybrids?

With the advent of the Tilley motor-generator, a new factor enters into the situation. If the system proves valid, does this mean that Honda and Toyota are out of the running? Hardly not! It is likely that these companies have allowed or even planned for this contingency with larger capacity motor generators which would do a similar job. With their fine cars and superb engineering, the transition from a hybrid to a pure electric might be costly but not too difficult to engineer. Each of the two hybrid companies may already have been built, tested vehicles of this nature. However, since the primary purpose of the car companies is to make money it is doubtful if this full-fledged fossil fuel-less type will be placed on the market unless competition so requires. A car using the Tilley motor-generator might certainly push them in this direction.

Should the Tilley remains undeveloped, one might expect the second level hybrid car to arrive in a few years from Honda and Toyota. These cars should have greater electric generator-motor capacities and smaller gasoline engines. These systems could give small cars much higher gas milage and could also be placed in larger, more elegant vehicles such as the Honda Accord and the Toyota Camrey.

The transition to fuel-less motors should hopefully be slow and painless for both the pocket book as well as the human psych. If scientists cannot be convinced by studying conventional scientific laws on how these motors work, perhaps they will better see the need for the promulgation of new laws in this area.

We know that at the present time, both Honda and Toyota motor generators (M.G.) have a powers assist function while the Tilley stands alone. Let us look at the present situation (we will use the Toyota U.S. M.G. because it is the more powerful version):

Honda and Toyota classify their M.G.'s as permanent magnet types; Carl Tilley says his is and advanced D.C. motor.

Power output:

Honda is given at 10 kW at 3,000 rpm or 13 hp
Toyota is 33 kW at 1,040-5,600 rpm or 44 hp
Tilley is up to 135 hp depending on rpm's

Voltage/Battery (Ni-MH=Nickle Metal Hydride, AH=amp hours)

Honda: 144V, 120 Ni-MH cells@1.2V ea., 6.5 AH
Toyota: 274V, 228 Ni-MH cells@1.2V ea.
Tilley: 144V,12-12V lead acid (Wal-mart), 1200AH

Physical Characteristics of each M.G.

Honda: 60 mm wide
Toyota: unknown
Tilley: 9" dia., 146 pounds

The main effort should be to implement the transition away from the use of fossil fuels. We know this can be done but it takes proper attitude. Proper attitude takes knowledge and knowledge takes willingness to learn. Willingness to learn takes desire to learn and desire to learn takes humility. Unfortunately, humility is a characteristic that many people, especially learned ones lack. *Editor's: Carl Tilley's comments are also given below.*

Very Good...you seem to know what we already were aware of...before we could mass produce our Tilley car I am sure some auto company would come out with basically the same thing...I do believe they already have it and they are just waiting for the right time. The good news is that at least the Tilley car lighted their fire to get them started.

Carl Tilley

Read in the next issue!

In April 12 of 2003 scientific conference "The Time Machine" was organized by Faraday Laboratories Ltd in Moscow. There were presented the reports by V.A. Chernobrov, A.V. Frolov, A.V. Rykov, E.D. Sorokodum, V.J. Kosyev, A.N. Solonyi, V.A. Atsukovsky. There was discussed design of the device made by Faraday Laboratories Ltd. as well as main principles of control on temporal characteristics of physical processes by changing of density of space energy, i.e. aether density according to patent claim.
Details are in the next issue!

Electric Vehicle in Russia

Review by Correspondent Alla Pashova, Russia

The director of *EV* (*Electric Vehicles*) company located in Tolyatti (Russia), Alexander Mukhanov states: “If during 2 centuries intellectual efforts of a scientific thought were concentrated on an electric engine instead of the internal combustion engine then now we would have driven electric vehicles, and the greenhouse effect would not have exist”.

The electric vehicle has driving-wheels which are set in motion by an electric engine, and it is powered by an electric battery. The electric vehicle appeared in Great Britain and France in 1880 and it is significantly older than the automobile equipped with the internal combustion engine.

At the beginning the speed and the fuel distance of the electric and gasoline vehicles were approximately the same. The main disadvantage of electric vehicles was a complicated system of recharge since there were no usual transformers of alternating current into direct one. Hence, the electric vehicle was recharged in a quite difficult way. An electric engine operated on alternating current was used to recharge it. The electric motor rotated a shaft of the generator which was connected to batteries of the electric vehicle. However in 1906 a rectifier which was simple to operate was invented. Despite its principle of operation was quite complicated (mercury vapor was used for transformation of alternating current into direct one) it has given a great stimulus for development of automobile industry.

There were no essential proves of advantage of gasoline vehicles as compared with electric vehicles. Nevertheless, even at that time there was an opinion that electric vehicles run more slowly and to shorter distances. However “gasoline” records were also insignificant and comparable with results produced by the electric vehicles. At the same time producers of electric vehicles did their best to demonstrate the advantages of these vehicles. If you remember, Walter Baker, one of the most famous American constructors and producer of the electric vehicles, achieved the speed of 130 km/h driving his car. An electric vehicle produced by Borland Electric company run from Chicago to Milwaukee (the distance of 167 km/h along non-asphalted road) without recharge. Next day after recharging the electric vehicle returned to Chicago by its own power without meeting any incidents. During the journey its speed came to 55 km/h.

In 1899-1900 a hereditary Russian nobleman Ippolit Romanov made an experiment on designing of the first Russian electric vehicle. Since these vehicles had been designed to run in Saint Petersburg then the city council required to produce them in the same place. In 1899 Romanov’s first electric vehicle was made. Its general design was borrowed from English cabs in which a cabman sat on a high dickey placed behind passengers. Romanov’s electric cab had two passenger seats situated in front of the dickey in a half-cabin having side and back windows. The cabin was equipped with a hood. The seat of a driver was placed behind and above the cabin; a box containing an accumulator was located under the driver’s seat. The vehicle was four-wheeled. Diameter of the front wheels was smaller than those of the rear wheels. The front wheels were fixed by elliptical springs and were driving ones. They were connected to two independent electric motors by a roller chain drive. The regulated rear wheels had a smaller diameter and were fixed by spiral springs.

This first electric vehicle was equipped with lead accumulator which had 36 banks. It required recharging every 60 versts (1 verst – 3500 feet). The total power of both the electric motors was equal to 4 hp. Design of the vehicular part of Romanov’s electric vehicle was borrowed from the models produced by an American company named *Morris-Salom*. This company had produced electric vehicles since 1898; however those models had wheels of greater diameter as they were equipped with pneumatic tires, while Romanov’s vehicle had wheels with light rubber rims. Both the electric vehicles had two electric motors, but the distinction of Romanov’s electric vehicle consisted in the presence of 6-row ball bearings in the wheels.

Romanov’s second cab was built in 1900. This model had entirely closed and glazed cabin for passengers. The chain drive was replaced by a gear; however the basic sizes remained the same. This model was equipped with an accumulator designed by Romanov. The general weight of the electric vehicle was 45 poods (750 kg), and the weight of the accumulator came to 22 poods (362 kg). It should be noted that weight of American and French electric vehicles exceeded these figures. The speed of both Romanov’s models of the electric vehicles came to 15 km/h.

Romanov’s first electric bus was built in the same 1900. The electric bus weighted 100 poods (1600 kg) could make speed up to 10 km/h. For reasons beyond Romanov’s control, traffic of the electric vehicles in Saint Petersburg was not organized.

A vehicular company named *“Frezer and Co.”* participated directly in production of Romanov’s electric vehicles, i.e. a running gear of these vehicles was created by this company. In January 1900 *The Cyclist* magazine reported that this company had built an electric vehicle which had already been tested. A picture of this first electric vehicle by *"Frezer and Co."* can be found in advertising publications of those years. Various firms proposed the innovation for sale and described its technical characteristics.

Frezer electric vehicle was four-seated, had two engines whose total power came to 7 hp. The advertisement booklets reported as following: "The accepted system of accumulator is remarkable for its solidity and life-time. Capacity of the elements is about 15 Amperes/ hour per a kilogram of electrodes. The size of the battery is defined according to calculation of 120 W/h per one ton and per one-kilometer of the distance; its weight is equal to 30-40% of the total weight of the vehicle including a payload". Weight of Frezer electric vehicle came to 70 poods (1120 kg). It made the speed of 15-18 km/h and required recharging every 35-50 versts of the path. The second variant of Frezer electric vehicle differed from the first one for less weight of the accumulator which was located above the rear axle.

In addition to the four-seated vehicles, *Frezer* Company produced two-seated electric vehicles equipped with engines of 3.5-hp power. The two-seated electric vehicle was remarkable for its more perfect steering tube. There is known one specimen of this model purchased by somebody from Riga (Latvia). Moreover, on the territory of *"Frezer and Co."* there was a station for recharging accumulators of the electric vehicles.

In the period of the Soviet Union pilot batches of electric vehicles began to be produced only after the II World War. In general there were small lorries applied for transportation inside great buildings, for example, in a main post office or an airport.

Now there is a great probability that Moscow will be the first Russian city where the electric vehicles can find mass application. Russian capital suffers from gas pollution which is especially heavy in the center of the city. At the same time the budget of the city allows acquire own ecologically appropriate transport. Now financing of production of the electric vehicles is realized through the Fund of Ecologization of Moscow Transport. The means are assigned due to income taxes coming to the budget of the city.

To tell the truth, citizens of big cities know electric vehicles very well, i.e. usual trolleybuses and trams are electric vehicles despite they are not autonomous. An autonomous public electric transport has one advantage which is especially valuable in the conditions of the cities overcrowded by cars. This advantage is their mobility. Trolleybuses are inseparable with their wires that causes decrease of their maneuverability. For example, they cannot pass several cars which have been parked wrongly. In this case a maneuverable and autonomous electric vehicle will run to the center of the road and pass the obstacle with ease.

A Moscow company named *Eltran* which deals with exploitation of electric transport has developed a model of EV which runs the distance of 60 kilometers without recharging. Total recharging of this electric vehicle takes about 5 hours, however a half of required energy they can obtain for 1 hour. Hence, if the accumulators are recharged before their full discharge then this vehicle can operate constantly. However, a developed infrastructure of charging stations is necessary for this aim.

From the economical point of view, EV does not yield to trolleybuses and trams which house more passengers but run twice more slowly than the autonomous and compact EV. Therefore EV does not experience difficulties in jams in the conditions of the traffic of the city. It maneuvers easily in thick traffic current.

Unfortunately, now EV cannot replace automobiles equipped with internal combustion engines. Imperfection of its accumulator batteries does not allow it to become a public vehicle. The battery imperfection causes a small run on one recharging, long cycle of recharging, and high price of the electric vehicle.

To tell the truth, EV has important advantages. Maintenance charges of EV are lower than the charges of the standard automobile requiring expenditures for support of cooling systems, powering systems and exhausting systems. The lifetime of the electric engine comes to 10 thousands hours. Quantity of operations for maintenance of the engine is minimal. For example, it is necessary to change brushes in the dc motor from time to time. As for a modern three-phase electric motor and synchronous ac motor they are practically maintenance-free.

EV is easy to drive. To start the car you should only insert the key, turn it and press the accelerator pedal. You do not need to make any manipulations with adherence or gearshift.

Use of EV is justified in so called zones of higher environmental requirements that are cities, parks, vacation resorts. Electric buses, developed by *Eltran* run in the public green space of All-Russia Exhibition Center. There are about ten compact commercial EV operating in Moscow.

Taking care of environmental sanitation the management of AZLK (big automobile plant) has developed EV on the basis of two production automobile models "Moskvich-2141" and "Moskvich-pikup-2335". "Moskvich-Elektro" (Fig. 1-2) at fully loaded mass of 2060 kg accelerates up to 60 km/h at 15 seconds. Maximum speed of the load-free vehicle is 110 km/hour. Cruising range at one battery charging is 100 km, and it comes to 80 km with 400 kg of load. At present Electrical Vehicle "Moskvich" is filled with imported component parts, which are cheaper but functional. In time it will be supplied with Russian equipment that will make it cheaper.

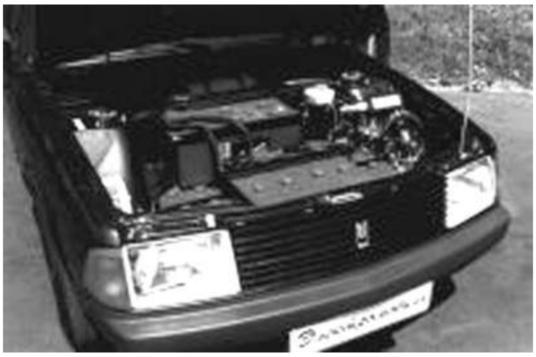

Fig. 1

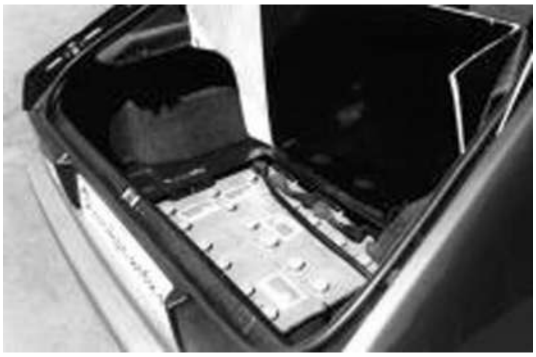

Fig. 2

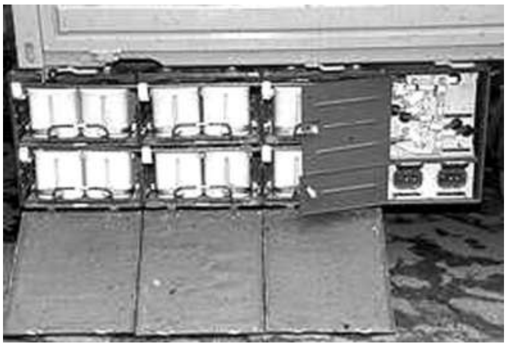

Fig. 3

The model of truck equipped with electric motor "ZIL-Electro", the authors of which being called "AVEKS" and "Optimum-electro", also enjoys the support of the Fund of Ecologization of Moscow Transport. Various privileges and free passage to any location in public green zones are promised to the future owners of "ZIL-Electro".

Forty-eight lead-acid traction batteries of roll type (Optima battery) are hanged in pairs along the ramps of the truck. They are united into 4 sections (Fig.3). Capacity of these batteries is enough for 70 km running. Probable charging station for these machines is the loading/unloading point. Charging time varies from 30 to 60 minutes. Substantial disadvantage of the model is that batteries take away a great percentage of carrying capacity. Among the advantages of the model there are small internal resistance, fast recovery (to 400 A), discharging to zero level without serious consequences.

A traction induction unit "ATAD-Optimum 50/120" serves as the engine for "ZIL-Electro". Its weight is only 100 kg and it doesn't require any diesel generator or transmission that compensates a little the loss of carrying capacity.

The engine is very simple. It is a non-contact rotor of a "squirrel cage" type rotating on the bearings. No brushes are used. Thus expensive maintenance is postponed for uncertain date. On the other hand there is asynchronism. It means that some equipment is needed for converting into three-phase current and further frequency and amplitude adjustment. However it is expensive and complicated.

In spite of all existing disadvantages, Electric Truck is considered to have a bright future ahead of it. There several reasons for that:

1 The vehicle is manufactured from cheap production component parts.
2 The system of optimum vector control of the drive is introduced. (The torque and shaft speed of rotation is regulated precisely, energy of the batteries is used rather efficiently.)
3 Microprocessor system with feedback coupling and great number of sensors (current sensor, heat sensor, velocity sensor, voltage sensor) minimizes losses, prevents the motor and batteries from overheating, protects mains from short circuit in case of an accident.
4 The vehicle is reasonably priced (for such an exotic object) – about $26 000 USD subject to serial production.

Volzhsky automobile plant has already been engaged in research of EV for a quarter of a century. Over 10 original models have been designed and produced for these years. They have been appreciated abroad as well as in Russia. These are electric vehicles "Poni", "Oka", "Elf", "Gnom", "Niva", "Rapan" "Lada-golf", etc.

The models manufactured on the basis of VAZ-1111 "Oka" became the most popular among them. These EV received many different rewards, and became world-renowned. For example, VAZ-111E "Oka", manufactured in 1993, kept all merits of a production midget car. Use of electric motor makes it a non-toxic and noiseless vehicle. Efficient fast-acting short circuit protection eliminates current rush, providing absolute car operating safety.

In EV of VAZ manufacture two direct-current motors are commonly used as a power-generating set. The first is a 25kW capacity motor with torque of 110 Nm.

Another one has 40 kW capacity and torque of 190 Nm. Motors of the first type are usually mounted on light vehicles, such as “Golf”, “Oka-Electro”, “Elf”, while high-power motors are usually used for VAZ-2108, VAZ-2109, “Niva”.

The first models of electric vehicles used to be equipped with nickel-zinc batteries. But as life time of these batteries is not long, it was decided to start using nickel-cadmium batteries, manufactured at Saint-Petersburg plants “Rigel” and “Istochnik”. Nickel-cadmium batteries are power-consuming, they withstand temperature of –40°C. The only drawback of these batteries is their cost, and as a consequence a high cost of the vehicle itself. Because of that the final choice has been made in favor of lead-acid storage batteries, which were mounted at “Gamma Golf” in particular, thus reducing its price by several times. The batteries are located under the driver and passenger seats. They allow discharge of 80-85% and provide cruising range without recharge of up to 100 km within city area and several times more on a highway. Power current is transmitted from the electric motor to the front wheels through single-reduction gear unit, which substitutes the transmission.

For the present time electric vehicles are not developed to take the place of automobiles with internal-combustion engine in all industries and spheres of application, but only as a specific vehicle for highly specialized utilization. They are utilized when use of internal-combustion engine is objectionable or impossible.

It is considered that so-called hybrid models of EV would be able to replace automobiles with internal-combustion engines. These are vehicles with two motors, an electric motor and an internal-combustion engine usually running on diesel oil. In this case the combustion engine works constantly, but under the operation modes of little air pollution. Besides, fuel consumption is also much reduced.

In this direction real commercial success is possible. For example, “Toyota-Pirus” has already been chosen as a personal vehicle by tens of thousands of customers, this being a good result for such an original construction.

Nowadays three circuits of hybrid propulsion system are generally known. The simplest one is a series circuit, in which combustion engine operates together with a generator, and battery or generator powers electric motor providing propulsion. Actually this is the same circuit of electric transmission that is used at quarry trucks. Its main disadvantage is great losses of energy at transmitting the torque to the wheels. In parallel circuit the output shafts of the motor-generator and combustion engine are rigidly bound, and it doesn’t allow working in steady-state mode (i.e. with permanent rotations and load).

In the third circuit, which is also the most commonly used there are an electric motor, combustion engine and a generator. The output shafts of all these units are bound by planetary differential. This allows almost lossless distribution of power between the units, and provides sustained performance of the combustion engine. The disadvantage of this circuit is the extreme complexity of coordination of units operating.

A hybrid EV was produced in Russia by automobile plant “Izhmash” though it has not been put in mass production yet. At the developing of the propulsion system instead of unique components there were used units of Russian mass automobile production that is one of the fundamental design solutions. “Orbit” Izh-21261 was used as a base for the device. It was equipped with gasoline engine (of 30 hp power and 650 cm^3 capacity), drive dc motor ПТ-125-12 (of 120V voltage and 49 H/m torque) and eight accumulators 6CE-55. Twin-engined propulsion system works according to the parallel circuit of engine joint. Electromotor serves as starter and at the acceleration it operates simultaneously with internal-combustion engine. On reaching some definite speed the electromotor changes its operation mode to the mode of generator and charges the accumulators. When the loads at gasoline engine increase then the electromotor begins to operate together with it. Operation of the hybrid propulsion system is controlled by starting regulation electronic equipment. Control unit, speed transducer of crankshaft of the internal combustion engine, detector of the accelerator pedal etc. are responsible for coordinated work of the engines.

During the tests fuel consumption of the hybrid vehicle “Orbit” was less by 20% than that of the usual model. For Russian prototype it is a good qualitative characteristic, however foreign designers reach better results, for example, “Crysler ESX3”, which was presented in 1998 at Geneva auto show, consumes 3.3 liter of diesel oil per 100 km of path.

Representatives of specialized Moscow companies producing EV assert unanimously that these are hybrid electric vehicles which should change public transport with internal combustion engine. To encourage the automakers it is useful to create the proper laws drawing on foreign experience. For example, in California, USA, automakers whishing to present their production in the state market should put out 2% of automobiles with null blast. Legal system also supports an automobile user who is provided for free parking and free or cheap recharge. The user is granted with maximum of discounts to encourage him buying EV which is still twice as expensive. However it is said that in one of the villages situated near Moscow where the so called new Russians live, everybody drive electric vehicles leaving their cars with gasoline engine outside the gates. It would be quite good if EV turns to the real vehicle from the extravagancy of nouveau riches and beloved creation of ecologists.

The Richard Clem Motor and the Conical Pump

An Investigation of the Clem Motor

Robert Koontz, USA
rdkoontz@attbi.com

In December of 1992 Jerry Decker posted an article on the KeelyNet BBS, about a self-running motor that developed excess useable power. The information, gathered from newspaper and individual sources, gave an anecdotal account of the motor invented in 1972 by Richard Clem of Flower Mound, Texas. New information has since been added and can be found on http://www.keelynet.com/energy/clem1.htm.

Richard Clem worked with heavy machinery for the city of Dallas. He used asphalt-spraying equipment, which pumped liquid asphalt. He noticed the asphalt pump would continue to run for up to 30 minutes after the power was turned off. It was this discovery that led to the development of the motor. Modifications he made eventually resulted in a substantial 350 horsepower output from a 200-pound motor. Clem is said to have often driven a car, powered by this motor, up and down Central Expressway in Dallas. He claimed it didn't use any fuel, and only needed a change of oil every 150,000 miles.

The motor had only one moving part, a cone shaped rotor mounted vertically on a hollow shaft. Spiral channels cut into the cone wound around its length and feed into peripheral nozzles at its large end. When fluid flowed through the spiral channels it was ejected out the nozzles and caused the cone to spin. At a certain velocity, the rotating cone became independent of the starter pump and began to operate by itself. At an operating speed of 1800 to 2300 RPM the fluid heated up to 300 F, requiring a heat exchanger. Vegetable oil was used because at 300 F water boils and conventional engine oil breaks down. A 12-volt battery was the only other power source. Clem never applied for a patent because his motor design was derived from the asphalt pump that was already patented. Fifteen companies turned him down before a large coal company offered to back him and signed contracts to sell the motor. Soon after the deal was signed, Richard Clem died of a heart attack

The above account contains only what I considered to be relevant for analysis of the Clem motor. Visit http://www.keelynet.com/ for the original material.

The gear pumps, typically used for asphalt spraying, do not match the description of the pump used by the city of Dallas back in 1972. There should be public records showing what equipment manufacture the asphalt sprayer was purchased from. Since the asphalt pump was patented, I searched for a pump patent that met the following criteria:

1) Patent issued on or before 1972.
2) Delivered pressure equivalent to a positive displacement gear pump.
3) Cone shaped rotor with spiral channels.
4) Self-propelling action.
5) Capable of pumping a viscous fluid like asphalt.
6) Large heat transfer to pumped fluids.

The following illustration (Fig.1) is from US Patent 3,697,190 (Truncated Conical Drag Pump). The patent was issued October 10, 1972 (criteria 1) and appears to match the description of the asphalt pump that Clem converted into his motor.

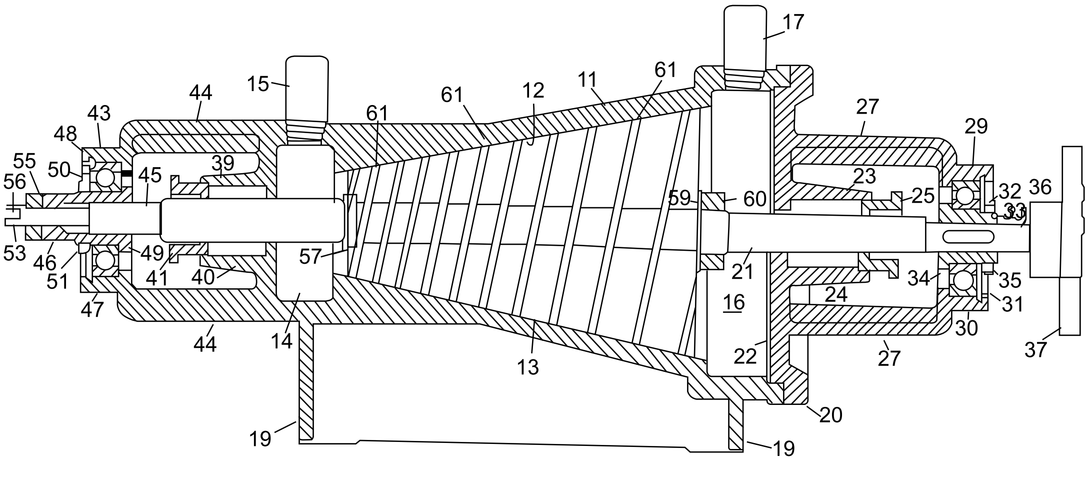

Fig.1

Housing 11, Conical interior wall 12, Conical rotor 13, Inlet chamber 14, Inlet pipe 15, Outlet chamber 16, Outlet pipe 17, Support feet 19, Detachable end cap 20, Rotor shaft 21, End cap wall 22, Boss 23, Packing 24, Adjustable gland nut 25, Bracket arms 27, Bearing boss 29, Bearing 30, Snap ring 31, Inner race 32, Sleeve 33, Shoulder 34, Retainer nut 35, Reduced diameter outer end 36, Coupling 37, Packing 39, Retainer 40, Gland nut 41, Bearing boss 43, Integrally formed bracket 44, Shaft reduced diameter 45, Bearing sleeve 46, Bearing 47, Snap ring 48, Inner flanged 49, Inner race 50, Nut 51,Shaft reduced diameter 53, Lock nut 55, Flat faces 56, Snap ring 57, Washer 59, Nut 60, Helical channel 61, Channel base 63, Channel sidewalls 64.

This is a high-pressure, low volume drag pump that can be used in place of conventional positive displacement pumps (criteria 2). It has a conical rotor that has a close fit clearance with the stationary housing wall. Delivered pressure is limited by back flow across the radial clearance and is inversely proportional to the square of the clearance. As a result, even a small increase in radial clearance would rapidly reduce pressure. The rotor is cone shaped so that the clearance can be controlled by axial adjustment of the rotor relative to the housing wall.

The conical rotor has two helical channels (criteria 3), in the form of square threads, spaced 180 apart for balance. The channel depth decreases as the rotor diameter increases. Fluid enters the channels at the small end of the rotor. The fluid is induced to rotate with the channel by boundary layer drag. The boundary layer is the thin layer of fluid adhering to the channel surface. Molecular cohesion tends to drag the adjacent fluid with the boundary layer. The fluid is also in contact with the housing wall. The boundary layer drag against this stationary wall slows the rotation of the fluid in the channels. Because the fluid rotates slower than the rotor, it is forced through the channels towards the large end of the rotor. In addition the fluid is forced towards the large end by centrifugal force.

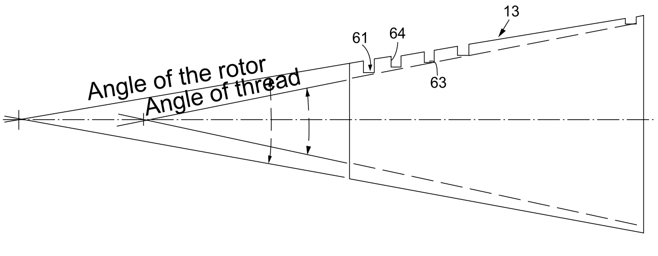

Fig.2

The above drawing illustrates the proportional decrease in channel depth as the rotor diameter increases. Why was this done? Note that as the diameter doubles so does the circumference. This means the fluid has to travel twice as far in the same time to maintain a constant slip velocity. By reducing the channel depth in half (cross-section area = depth x width) the fluid velocity is doubled thereby keeping the slip constant.

The spiral channels could be thought of as very long convergent nozzles. The increase in fluid velocity is in the opposite direction of the rotor spin. We should expect a reaction force from the acceleration of the fluid. This thrust would be directed tangent to the circumference and would increase the spin torque on the rotor. Even without the peripheral nozzles, that Clem later added, the pump rotor experiences a thrust force in a direction that would self-propel it (criteria 4).

Because fluid drag is the primary pumping force, it is well suited for viscous fluids like asphalt (criteria 5). The long channels also represent a large sliding surface area with frictional losses that would transfer heat to the pumped fluid (criteria 6).

All six of the patent search criteria have now been met. Of course this doesn't prove that it is the asphalt pump Richard Clem worked with.

A peculiar condition indicated by the patent is that as the velocity increases in the channels the pressure also increases. Bernoulli's Law requires the pressure to drop proportionally as the velocity increases. Assuming an ideal fluid without losses, when the channel depth is reduced in half, the cross section area is also half and this doubles the fluid velocity and the fluid pressure should drop in half. So what is going on here? There is a centrifugal component that would add to the fluid pressure.

My guess is it's too small to overcome the predicted pressure drop. Here is what I think may be going on. As the diameter and velocity increases the drag force propelling the fluid through the channel is proportionally greater. Energy is being added all along the length of the channel. Whatever the reason, if this high-velocity, high-pressure fluid is feed into tangent peripheral nozzles at the rotor large end, the energy will be converted to shaft horsepower.

The Clem motor is producing 350 shaft-horsepower and a large heat energy component. Where is this huge amount of energy coming from? Recent quantum mechanics zero-point field (ZPF) theories may point to the answer. From an article available at "BEYOND $E=mc^2$" Bernhard Haisch, Alfonso Rueda & H.E. Puthoff [1]:

"Our work suggests inertia is a property arising out of the vast, all-pervasive electromagnetic field we mentioned earlier, which is called the zero-point field (ZPF). The name comes from the fact that the field is held to exist in a vacuum-what is commonly thought of as "empty" space-even at the temperature of absolute zero, at which all thermal radiation is absent."

ZPF researchers theorize that mass, inertia and gravity are not intrinsic properties of matter but the interaction of matter with the zero-point field. By all pervasive it is meant that the ZPF exists not only in empty space but it is passing through your body right now and everywhere else. When you throw a stone you are interacting with this field since the ZPF resists change in motion. In essence the ZPF is the modern day aether.

The amount of energy making up the ZPF is thought to be enormous. Is the fluid acceleration in the Clem motor interacting with the ZPF in such a way as to rectify it and draw energy from it? Is it a hydraulic aether-diode? The fluid, in the Conical Drag Pump, flows through long convergent channels. Disregarding the boundary layer, is this accelerated flow laminar? Would such a long orderly flow entrain the aether energy?

From the perspective of the rotating channels the fluid appears as the discharge from a long nozzle. To exaggerate, if the fluid was held fast to the housing wall, the rotating channel would travel through the stationary fluid. This would be equivalent to achieving 100% efficiency. In reality the fluid is slipping against the stationary housing wall so that the rotating channel (nozzle) is moving faster than the fluid discharge velocity. Assuming the reaction thrust as the only propelling force, this would give efficiency greater than 100%. So, as the slip increases the reaction thrust decreases, but the efficiency increases.

Assuming the Conical Drag Pump is the pump Clem used, can it answer the following?

1) Why was a hollow shaft used?
2) Why was the cone mounted vertically?
3) Why was a starter pump needed?
4) How were the peripheral nozzles added?
5) How was the motor RPM regulated?
6) How did a large coal company get involved?
7) Was this kind of pump ever used in asphalt sprayers?

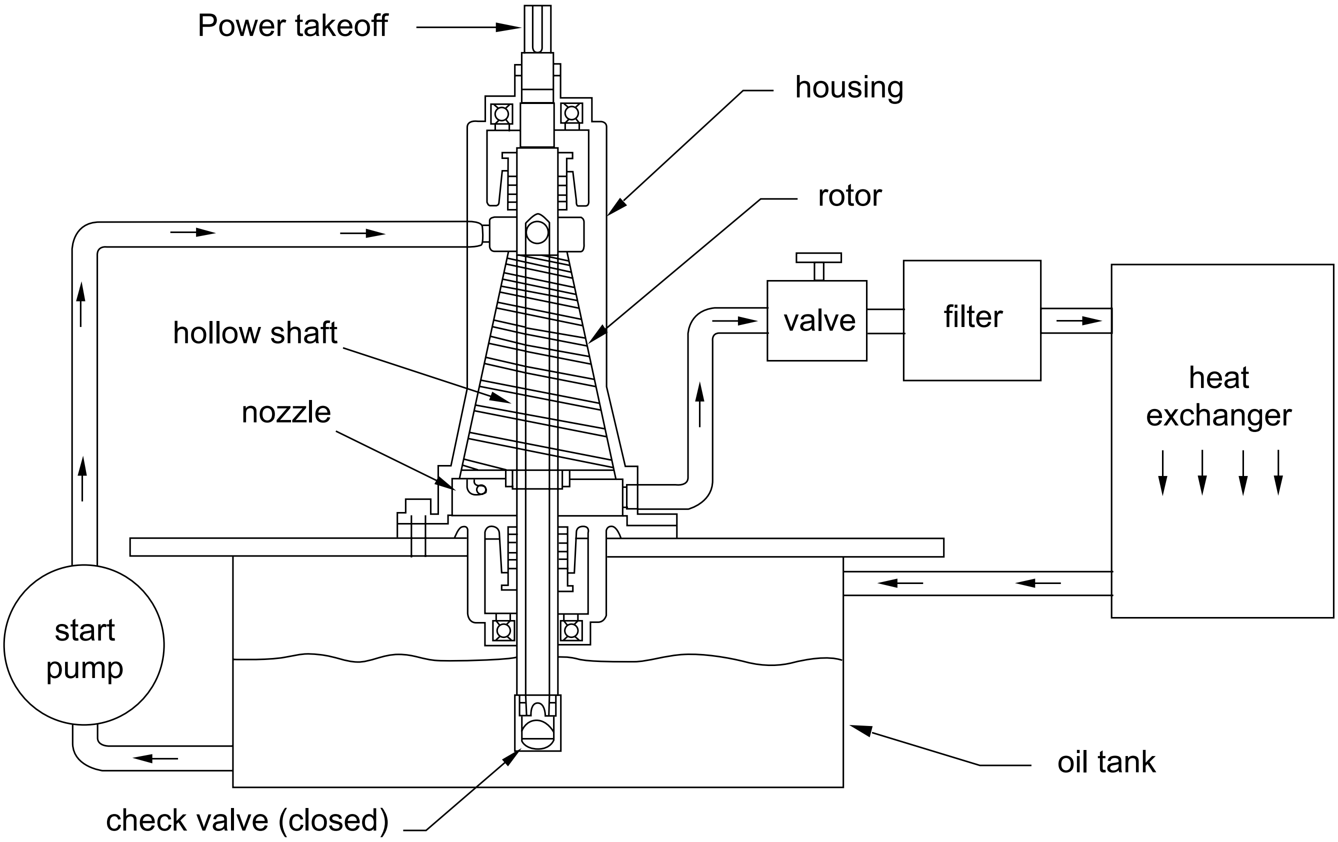

Fig.3

Clem motor shown in start mode
The arrows show oil flow

The above drawing shows a hypothetical Clem motor based on the Conical Drag Pump. The motor is mounted vertically so that the check valve on the hollow shaft is submerged down in the oil tank. The hollow shaft extends from the oil tank through the rotor into the inlet chamber. The start pump draws oil from the tank and forces it up the external feed line connected to the inlet chamber at the small end of the rotor. This fills the hollow shaft and forces the check valve closed. The oil flows into the spiral channels and out the peripheral nozzles. The reaction thrust of the nozzles spins the rotor. The oil flows through the return line, through the valve, filter, and heat exchanger and back into the tank. The start pump is most likely a standard hydraulic gear pump. It continues to pump until the rotor spins up to its operating speed. The combinations of a start pump and check valve would be a simple way to both prime the motor and spin up the rotor.

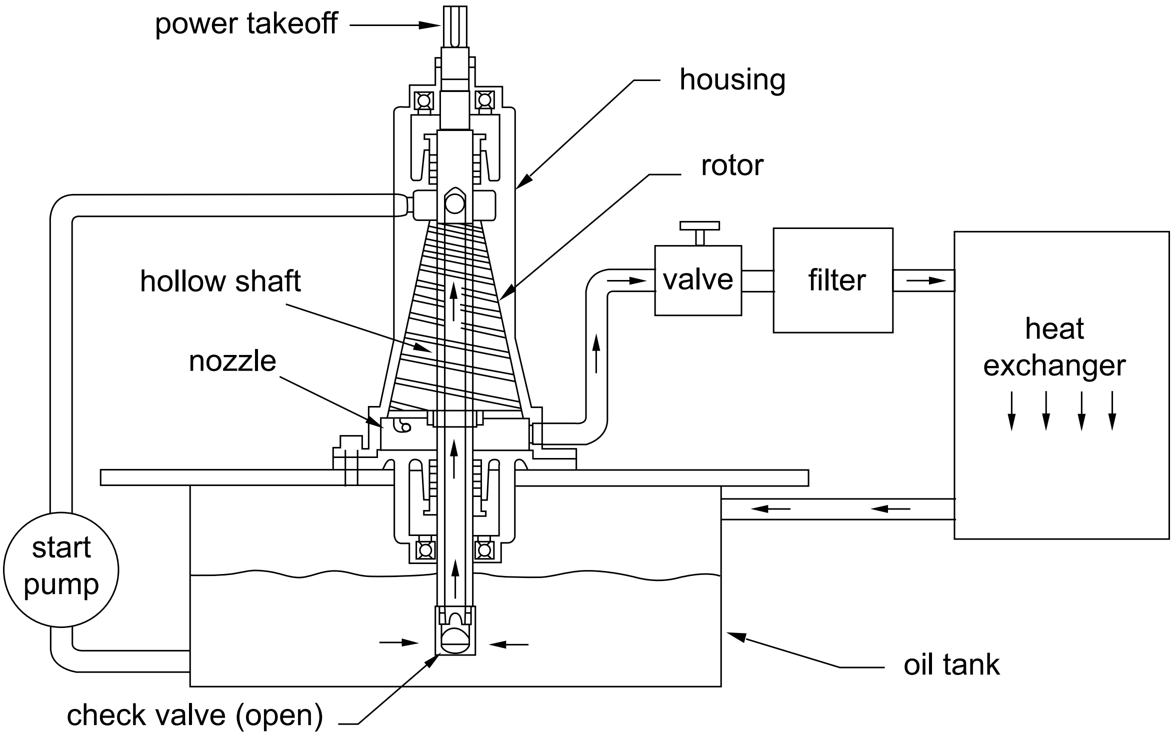

Fig.4

Clem motor shown in operation mode
The arrows show oil flow

Once the start pump is shut off the check valve is free to open. Oil is drawn up into the hollow shaft to the inlet chamber at the small end of the rotor. The spiral channels pump the oil down towards the large end of the rotor. A plate is attached to the large end of the rotor and fits with a close clearance with the housing wall. Nozzles attached to the outer edge receive high-pressure oil from the spiral channels. The jet reaction thrust from the nozzles delivers shaft horsepower to the power takeoff at the shaft top. Adjusting the valve to create hydraulic backpressure regulates the motor RPM. Closing the valve stops the motor.

When I first read about the Clem Motor I found it odd that a deal had been made with a coal company. Was there a connection with the pump? After finding the Conical Drag Pump patent, I wanted to contact the inventor Walter D. Haentjens of Barrett, Haentjens & Co., Hazleton, Pennsylvania. Otto Haentjens founded Barrett Haentjens & Co., in 1916. The business began in the coalmines of Pennsylvania with Otto Haentjens original patent on the balanced opposed impeller multi-stage volute pump. The company still supplies pumps to the coal industry. They have expanded to other markets and their pumps are installed in many industries worldwide. It is now known as Hazleton Pumps Inc., after its acquisition by The Weir Group.

I contacted Peter Haentjens, the VP/General Manager of Hazleton Pumps, by e-mail to find out if this pump had ever been put into production. He replied that they had not done anything with the patent.

An unusual pump design would have a tough time competing in the market with an industry standard like gear pumps. The Dallas asphalt sprayer may have been a one of a kind field test of the pump design. Or the pump manufacturer offered it for testing to an asphalt equipment company in the hopes of generating interest in it.

References

Bernhard Haisch, Alfonso Rueda & H.E. Puthoff, “BEYOND E=mc^2” //The Sciences, Vol. 34, No. 6, November / December 1994, pp. 26-31 copyright 1994, New York Academy of Sciences.

Spontaneous Flow Heat Pump

PACES : The Alternative Energy

Jean-Luc Brochet, France

http://www.new-energy-paces.com
Email: brochet.jean-luc@wanadoo.fr

"Paces" is a new high performance electronic process. It is not a traditional heat pump fitted with some electronic gadgets to improve its output, but a truly revolutionary process totally unused up to now. There are no moving parts, not even a moving fluid in it. "Paces" is an electronic process because for its development the pump's fundamental element uses micro-electronic techniques.

...an average of 9000MW could be produced, that is to say the equivalent of several nuclear power stations.

The principle consists in attracting the gas neutral molecules onto a so-called "electrostatic" plate, to accelerate them and heat the plate. An intense electric field is applied in a capacitor with micro-pores covering the plate. The neutral molecules are polarized and accelerated when they penetrate these micro-pores where the intense electric field exists. Each molecule accelerates and absorbs electrical energy when it enters into a pore. But in turn it decelerates and gives off an equal amount of energy upon its exit from the pore.

The molecules, by going back and forth between the plates, continually give off and take energy, thus resulting in a total consumption equal to zero.

The gas fluid does not undergo any change of thermodynamic state, contrary to all other heat pump processes. A thermodynamic cycle remains at the level of each gas molecule, but not at the level of the entire set of molecules.

There is more to this new process, it is not just heat pump process added to an already long list: **these new characteristics also enable some new applications.** Besides the traditional heat pump applications, due to its excellent coefficient of performance this process makes it possible **to produce electricity by cooling the environment.**

For instance, if the Rhine was equipped with heat exchangers into which diverted water could flow, thus cooling its temperature by just one degree, an average of 9000MW could be produced, that is to say the equivalent of several nuclear power stations. Please note that this process would also enable **both electricity and fresh water to be produced from seawater,** an interesting application for areas of the world where water is a scarce resource.

Thus we have at our disposal an ecological process for producing electricity, it would become conceivable to produce hydrogen in great quantities; vehicles would be hydrogen powered. Ozone, CO2 or any other type of pollution would become a thing of the past.

It will most probably be the energy source of future generations, ecological and excellent, inexhaustible, powerful yet completely harmless, stable and steady (contrary to the sun and wind); an invention capable of re-launching the world economy, in particular giving impetus to the currently depressed new technology sector.

However, you have probably noticed that this invention requires advanced technology, and the few French laboratories capable to develop this process have proven inaccessibility to do it for independent researchers such as the author of the article. Besides, the invention's claimed qualities, the revolutionary aspect was very often interpreted by examiners as 'utopian', even before they read the document! Having tired of battling against concept fixed for the past two centuries, the author of the article **has now decided to place this project within the hands of the international scientific and technological community,** in order to give the best chance of success to the invention.

It will most probably be the energy source of future generations!

Initial development requires the construction of a prototype; three possibilities are described on the website (http://www.new-energy-paces.com). As there is no longer legal protection abroad, anyone is free to experiment and to market, as he or she wishes, in all other countries outside of France.

Please help me to spread the word about this project, for example through scientific forums on the Internet. This will allow you to receive feedback and gather other opinions. Your participation, even the most modest, will be of use. To those of you who help this project to see the light of day, all of humanity will be indebted.

Registration of Gravitational Waves

Vladimir N. Petrov, Russia

173001, apt. 76, 10 Bolshaya Saint-Peterburgskaya st., Veliky Novgorod
E-mail: petvn@km.ru

Registration of gravitational waves is one of fundamental problems of physics. At the beginning of the last century A. Einstein's works predicted existence of these waves. Nevertheless, despite numerous attempts are made in different countries of the world these waves have not been fixed up today. Josef Veber, an American physicist working in 60th of the XX century, is sure to be called as the first experimenter who disclosed gravitational waves. Nowadays lots of many-million projects on registration of gravitation waves are developed and realized (LIGO (USA), GEO-600 (Germany-England), VIRGO (Italia), and TAMA (Japan)). However no positive results were achieved.

Failures in registration of gravitational waves can be explained by the fact that researchers have erroneous notion of the essence of these waves. This erroneous notion is laid in an experimental project that causes negative results.

Many researchers believe that influence of gravitational waves on the bodies should be resulted in deformation of these bodies (i.e. in change of mutual place of particles of a material body). All the attempts to disclose the gravitational waves are based on this notion. Now two types of ground gravitational antennas are supposed as priority.

The first type supposes the registration of mechanical oscillations of a massive test body that is initiated by influence of gravitational waves on it.

The second can register changes of a distance between space-apart free masses when gravitational waves act upon them. Great attention is paid to the second type of antennas in the developed projects. Failure attempts of registration of gravitational waves can be explained by their low intensiveness and insufficient accuracy of existing measuring equipments. Therefore, now there are made various attempts to register gravitational waves produced by such great cosmic phenomena as confluence of black holes. More accurate instruments and measuring equipments, for example, a laser interferometer, are developed. Nevertheless, positive results are hardly to be obtained.

Here is one of the fallacious explanations of action of gravitational waves on a body: *"Passing of gravitational waves changes an interval between objects transferring them one from another like two boats sometimes are approached sometimes moved from each other by sea waves".*

Following this example to explain influence of the gravitational waves on bodies placing on the Earth, it can be said that all bodies are in one boat and raising or sinking of the boat on sea waves does not cause change of the distance between the bodies. However, this example seems to be very simplified for explanation of influence of the gravitational waves on bodies. In fact, the case is much more difficult.

According to A. Einstein's theory, bodies moving with variable acceleration will radiate gravitational waves. **Gravitational waves represent the essence of changes of gravitational field occurring at light speed.** Since the gravitational field is a deformed space-time then the gravitational waves cause change of deformation of the space-time. Change of deformation of the space-time is accompanied with change of deformation of geodesic lines (i.e. lines by which all the bodies move).

A body moving by a geodesic line whose deformation is changed changes its acceleration. Change of acceleration of the body will change gravity applied to the body in this area of space. Hence, influence of gravitational waves on the body occurs through gravity which influences on it.

It is known that a fundamental difference between gravity and other forces appearing at a direct contact with a body consists in the fact that gravity provides all the elements of the body (all its points) with equal acceleration (deformation can not occur), and other forces influence on certain parts of the surface of a body and so cause its deformation.

Hence it is senseless to try to measure deformation of a body or change of a distance between test bodies occurring under influence of gravitational waves. It is necessary to note that, in particular, when a body is placed on an unmovable support relatively to the Earth, value and the direction of gravity coincides with weight of the body. Therefore, in this case gravitational waves can be registered at measuring of changes of the body weight.

An experiment on registration of the gravitational field of the Earth by measuring change of weight of a test body was made. Before the beginning of the experiment there was set a problem to register gravitational waves produced by the Mercury and the Venus at approaching to the Earth. The gravitational field in which the Earth moves changes due to superposition of gravitational waves of other planets on it. These changes were necessary to be registered.

The experiment was as follows: a test body (a 100 grams weight) was weighted on an electronic balance in equal time intervals of 14 days. The experiment was lasting for 12 months. In the experiment it appeared that change of a distance between the Earth and the mentioned planets causes change of weight of the test body. The change of weight of the body occurred in proportion to the distance between the planets. The test body was weighted in the same conditions and in the same place. The performance of the experiment is so easy that every person can made it.

It is known that weight of a body changes (decreases) at moving along the Earth surface from its pole to the equator. Change of weight of a body locating in the same place has not been fixed until now.

The fixed change of weight of the test body proves change of gravity applied to it and, consequently, change of the gravitational field in which the test body moves as well as the Earth does. **These changes of the gravitational field were caused by motion (superposition) of the gravitational waves produced by the mentioned planets at their approaching towards the Earth at the distance of the existent gravitational field.**

Hence, the Earth and all the bodies locating on it move in the constantly changing gravitational field, therefore, constantly change their weight. Nevertheless, as it is known, change of the gravitational field is the influence of gravitational waves on it. Hence, measuring changing of weight of bodies we can speak about registration of gravitational waves.

In search of the 4th dimension

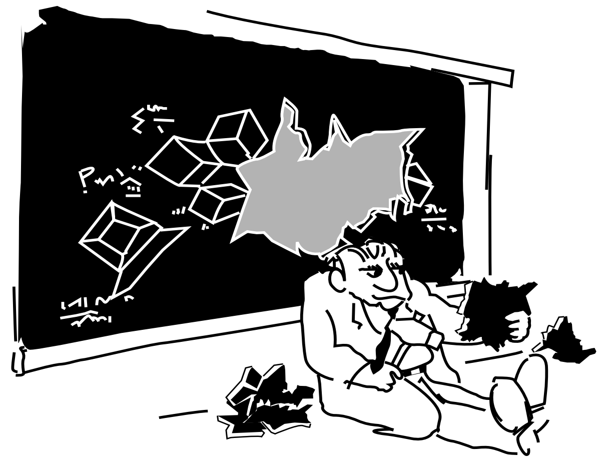

From http://home.ural.ru/~meop/index.html

High-Energy Hyper-Low-Frequency Electric Field

Yury V. Ivanko, Ukraine
E-mail: ux01w@hotmail.com

The article represents a practical and theoretical research on possibility of existence of a high-energy hyper-low-frequency electric field which is interpreted by modern measuring instruments and subjective perception as a static field. Besides there are observed a problem of the field registration, a hypothesis of relic origin of the registered field and a hypothesis of wave nature of the Universe.

A passenger on board a ship sailing in the ocean will never notice action of waves of high and low tide. At the same time, weathering waves of several numbers produce a dangerous storm. However, an ocean liner is designed for such a storm therefore the passenger would just experience some discomfort. It is another case if tide wave hides reefs… I wish our ship, i.e. the Earth, the Solar System, the Galaxy, and the Universe, not to be lead to such a place by any ignorant "navigator"!

One of the most mysterious persons in the history of electricity physics was an outstanding inventor Nicola Tesla (1856-1943). When the majority of scientists developed researches of microworld particles he followed the opposite scientific path. He was interested in the electrical potential of the whole Earth. He researched ways to influence on it, control its state and methods of its regulation. Therefore many of his patents, experiments as well as a purpose of constructions and devices built according to his ideas arouse perplexity and misunderstanding of modern scientists.

a test body was made. Before the beginning of the experiment there was set a problem to register gravitational waves produced by the Mercury and the Venus at approaching to the Earth. The gravitational field in which the Earth moves changes due to superposition of gravitational waves of other planets on it. These changes were necessary to be registered.

The experiment was as follows: a test body (a 100 grams weight) was weighted on an electronic balance in equal time intervals of 14 days. The experiment was lasting for 12 months. In the experiment it appeared that change of a distance between the Earth and the mentioned planets causes change of weight of the test body. The change of weight of the body occurred in proportion to the distance between the planets. The test body was weighted in the same conditions and in the same place. The performance of the experiment is so easy that every person can made it.

It is known that weight of a body changes (decreases) at moving along the Earth surface from its pole to the equator. Change of weight of a body locating in the same place has not been fixed until now.

The fixed change of weight of the test body proves change of gravity applied to it and, consequently, change of the gravitational field in which the test body moves as well as the Earth does. **These changes of the gravitational field were caused by motion (superposition) of the gravitational waves produced by the mentioned planets at their approaching towards the Earth at the distance of the existent gravitational field.**

Hence, the Earth and all the bodies locating on it move in the constantly changing gravitational field, therefore, constantly change their weight. Nevertheless, as it is known, change of the gravitational field is the influence of gravitational waves on it. Hence, measuring changing of weight of bodies we can speak about registration of gravitational waves.

In search of the 4th dimension

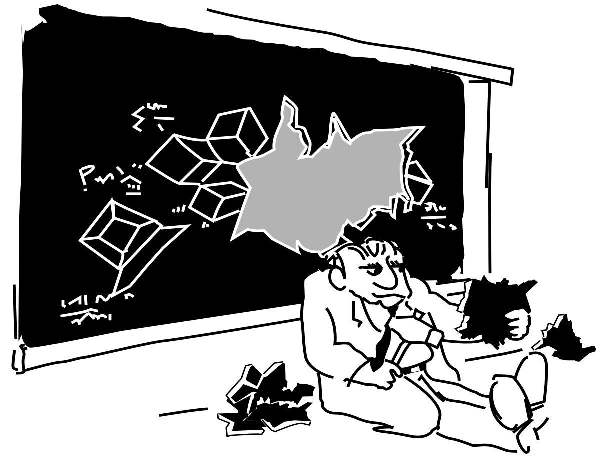

From http://home.ural.ru/~meop/index.html

High-Energy Hyper-Low-Frequency Electric Field

Yury V. Ivanko, Ukraine
E-mail: ux01w@hotmail.com

The article represents a practical and theoretical research on possibility of existence of a high-energy hyper-low-frequency electric field which is interpreted by modern measuring instruments and subjective perception as a static field. Besides there are observed a problem of the field registration, a hypothesis of relic origin of the registered field and a hypothesis of wave nature of the Universe.

A passenger on board a ship sailing in the ocean will never notice action of waves of high and low tide. At the same time, weathering waves of several numbers produce a dangerous storm. However, an ocean liner is designed for such a storm therefore the passenger would just experience some discomfort. It is another case if tide wave hides reefs… I wish our ship, i.e. the Earth, the Solar System, the Galaxy, and the Universe, not to be lead to such a place by any ignorant "navigator"!

One of the most mysterious persons in the history of electricity physics was an outstanding inventor Nicola Tesla (1856-1943). When the majority of scientists developed researches of microworld particles he followed the opposite scientific path. He was interested in the electrical potential of the whole Earth. He researched ways to influence on it, control its state and methods of its regulation. Therefore many of his patents, experiments as well as a purpose of constructions and devices built according to his ideas arouse perplexity and misunderstanding of modern scientists.

Having formulated my own hypotheses I often think why such a Great Genius as Nicola Tesla took his invention away with him.

Danger of this invention lies in simplicity of production of devices for generating (detecting) energy.

Practical Prerequisites

Prerequisites of this and other hypotheses were some undeclared effects occurring in practical wireless communication as well as investigation of devices for distant transmission of electric power.

Distant energy transmission was demonstrated by Nicola Tesla in 1892 in London and in 1893 in Philadelphia. However, even now the way he did it remains unknown.

In 1977-1985 I had to work a lot with radio transmitting equipment of different power of middle-wave (MW) and short-wave (SW) diapasons. According to fundamentals of physics, intensive electromagnetic field is produced around antenna curtain connected to a radio-frequency transmitter. There are brightly glowing gas-discharge electrovacuum devices at the distance up to several meters in this field. Attention was paid to the fact that more than ten daylight lamps of 80 W were successfully "glowed" at 40 W power of a transmitter. In this case a distant correspondent did not note significant decrease of a level of a received signal at the indicator of the receiver. It had to return to this fact in many years. The data of this experiment are represented in a particular document.

There is one more interesting effect which is disclosed when transmitting equipment (TX) and antennas (ANT) are adjusted. An incandescent lamp, which is connected to a certain place of a break of the antenna curtain or of a power feeder (L), glows brightly. The antenna circuit is not closed as galvanic element. (See Fig. 1).

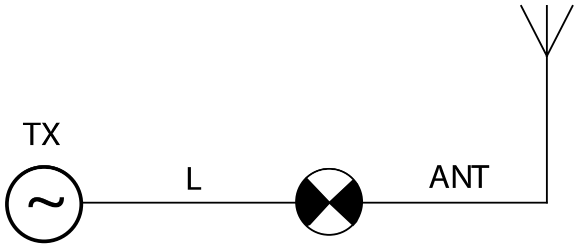

Fig. 1

Hence, the power line of the incandescent lamp is as if it is single-wire. It also had to return to investigation of this effect in many years. The data of this experiment are represented in a particular document. As in the case of the incandescent lamps, the correspondent did not note significant decrease of the level of the received signal at the indicator of the distant receiver.

S. V. Avramenko designed a plug of semiconductor diodes. Connecting this plug to a generator of alternating voltage of 10-10000 V by means of a single-wire line in the load of the plug he discovered current, which was continuous by the direction but pulsating by the value. Electric current measured in the single-wire line is very little, hence, the impression of superconductivity is produced, i.e. it is possible almost to avoid energy losses in the wires.

My experiments also represent a visual demonstration of the effect. For example, several links of light-emitting diodes connected inverse-parallel glow in the single-wire line. Garlands of light-emitting diodes and lamps having a single-wire power line can be produced. It is interesting that in this case power consumed by the generator do not increase practically (See Fig. 2, 3, 4). It was noted that if a standing wave is created around the generator "then an unlimited number of consumers would be able to apply a change of a value of this field to produce power in a load in their place" *(Editor: This phrase is in inverted commas since it is a part of the well-known article "Free Energy" by Alexander V. Frolov (Read in this issue, p. 15)).*

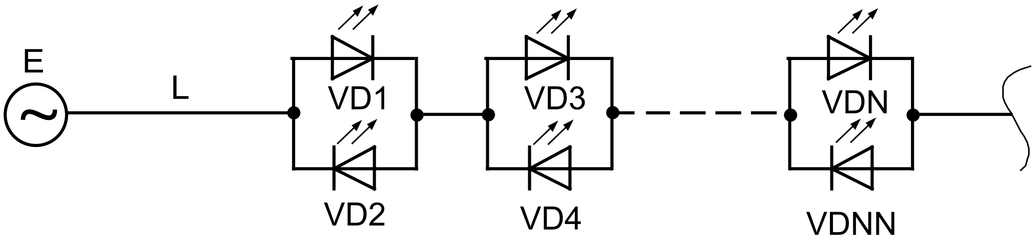

Fig. 2

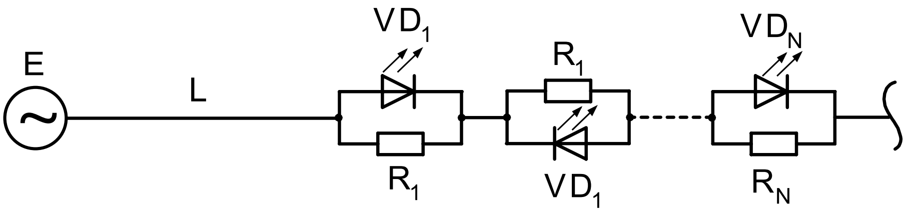

Fig. 3

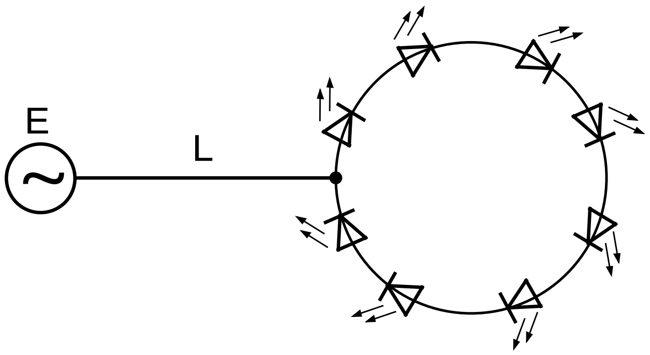

Fig. 4

Russian scientists investigated this phenomenon and proposed an explanation of producing active power in the load by means of reactive capacitive current. This was made applying resonant properties of the single-wire line made of a metallic conductor. Further, possibilities of application of nonmetallic conducting mediums for transmitting electric power were investigated. However, the researchers pay attention for the fact that this explanation is insufficient.

A task group named "Analysis" from Voronezh city (Russia) made a qualitatively new explanation of results of S. V. Avramenko's experiment. The works of this group from Voronezh are devoted to the comparison of properties and characteristics of so-called inertial and inertialess currents occurring in circuits at low frequencies. Their general and particular features have been disclosed. In the works there are demonstrated conditions at which conductivity caused by inertialess discharges predominates in conductors and conditions at which currents caused by conduction electrons predominate. There are demonstrated a description of S. V. Avramenko's experiments and the conclusions made by the researchers. The authors cite data of their test experiments which prove a part of conclusions made by S. V. Avramenko and his colleagues.

However this explanation does not have a right to exist as it does not have variants for practical application.

Results of my theoretical and experimental researches allow provide a **fundamentally new (!)** explanation of energy "transmission" having occurred in the experiments performed by S. V. Avramenko, investigations by D. S. Strebkov, V.A. Kuligin and other investigators. *(Editor: See publications of these works in previous issues of our magazine*). Moreover, they allow explain origin of many physical phenomena and processes including the origin of non-inertial charges and currents and their visualization.

First Hypothesis

Recently a number of scientists all over the world have declared a thought that surplus power is produced in the systems having overunit output due to inability to identify correctly all the existent energy sources. And a so-called phenomenology is observed due to the inability to deal with different types of energy (or fields).

In physics a negation of an overunit device follows a well known state which is considered to be true in other fields of knowledge, i.e. energy conservation law. It should be noted that it is formulated for isolated (closed) systems. I have devoted some time to the investigation of this state and to practical experiments and come to the following conclusion: there are **no isolated (closed)** systems in the nearest cosmos which can be observed by us!

Particular physical, chemical and other processes may actually be observed in closed circuits for making calculations with sufficient approximation.

Show me an isolated system, and I will prove that it is open.

Hence, according to fundamental physical laws, energy does not appear from nothing and not disappear to nowhere. However, the source of oscillations, i.e. our generator, does not provide a load with energy; nevertheless, power is produced in the load. Energy conservation law must be fulfilled! The classical formulation of the energy conservation law is as following: total energy of an isolated system is not changed in the course of time. According to this, the only fact remains valid, i.e. the single-wire system is not isolated from outer influence, hence, energy is consumed from some other source which is outer in relation to the system?! In this case our generator is one of sources of information about an amplitude, frequency, and phase.

Calculations and experiments have allowed made a following supposition. Avramenko's diode plug is a particular case of a device known in radio engineering for a long time, i.e. AM detector (a peak-detector, a limiter, a frequency mixer). I have examined the other known circuits of AM detectors and frequency mixers. The investigations are continued and by this time the supposition has been confirmed.

Let us accept an artificial generator (Tesla's transformer or any other one) as a main generator (or a main oscillator) and a hyper-low-frequency (HLF) oscillation of a huge amplitude as an envelope curve. Let us name it as RW ("relic wave") without placing emphasis on the sense of the name.

Let us suppose that the field can be of:
- local origin (artificial radio transmitters);
- atmospheric origin;
- geomagnetic Earth origin;
- the Moon – the Earth;
- the Sun – the Earth (the Sun as a nuclear reactor producing electromagnetic radiation, the Solar System);
- galactic origin;
- Universal origin – the Relic Wave; in contrast to the relic radiation occurring in the diapason of SHF (i.e. in

contrast to the convenient notion) it is the diapason of HLF.
- Etc.

The researches have demonstrated that the field of local and atmospheric origin may be excluded. The other variants require to be developed. Nevertheless, the variant of the Universal scale seemed to be the most probable. There is some reason for this supposition: the modern notion of the birth and the state of the Universe, i.e. the Big Bang and the expanding Universe. The known variants of scenarios of the Big Bang are based on a nuclear explosion of enormous power. The nuclear explosion is accompanied by an electromagnetic impulse (disturbance). It can be supposed that the Big Bang caused forming of a field having enormous amplitude values and long (great) period of oscillations. Probably, we deal with this field.

The second hypothesis about inconstancy, i.e. about wave nature of the Universe

A Russian proverb says: "Everything passes, everything changes". Constant electric voltage does not exist (!). That, what is considered as the constant voltage, really is a constant component (occurring at a certain time interval). This component is formed by combining of variable sinusoidal oscillations and/or is a quite low frequency fundamental. Hence, it is a non-sinusoidal "changing" having amplitude linear part at the time interval of the observer.

For example, in reality a usual storage battery of "direct current" can not support invariable voltage, i.e. when a discharge appears the voltage decreases (changes in time, oscillates with a non-sinusoidal oscillation and a quite low frequency fundamental), when a charge appears the voltage increases.

It is necessary to observe the constant voltage occurring after a rectifier and even a stabilizer as a constant component existing in the time interval from switching on to switching off. This will correspond to that is proposed by Fourier!

Show me a device producing constant voltage, and I will prove that the voltage is not constant.

Conclusions

If the supposition is totally confirmed then the statement that an electric field plays a fundamental role in the Universe will be valid.

We live in an electric field of a huge potential but we do not notice it. Since **our reference point is the amplitude value of RW field in our point of space** then the potential is equal to zero for our perception and for measuring instruments. Besides, **RW has an amplitude linear part at the time interval of the observer**. Hence, we consider RW field existing as unrevealed or as weak interactions in natural processes. Nevertheless, this influences directly on the whole surrounding material world.

Artificial production of heterogeneities in the RW field can cause its significant visualization.

According to the concept of the physical field, which was called by Einstein as the most important discovery in physics since Newton's time, every body generates a force field in the ambient space (for example, the electric field occurring if the body has an electric charge). This force field becomes apparent when a test body located in a certain point is influenced by a force.

Discovery of the fields of this scale allows define this concept more accurately. I would like to restate the information mentioned above in the following way: every non-homogeneity of physical vacuum (aether), which is registered by our perception or by measuring instruments as a physical object or a physical field, interferes with RW (according to other interpretations, it absorbs or rereflects RW, or becomes excited) producing new wave fields. This perceived physical field has frequency which is higher than RW frequencies. It is perceived as static in the time interval of observation. Hence, the field generates a charge and not the reverse. Any linear or non-linear transmission (motion) of a material or field object changes non-homogeneity which is generated by this object.

A question about measuring of parameters of the Universe remains open. Can an oscillation of the period of T=14 billions years and of the amplitude of A~T be seen, perceived, registered or measured?

The used literature and references are not cited in the article due to its great volume. The reader can familiarize with them at the web-site of http://www.efir.com.ua. The complete text of the article and other materials can be received from the author.

EAGLE-RESEARCH

Energy Solutions
since 1984

A Research Organization that Develops & Distributes Practical Energy-Saving Methods & Devices

4 Energy Way, P.O. Box 118
Porthill, Idaho 83853
FAX: 250/492-7480

Technical questions to:
wiseman@eagle-research.com

To the Question of Work Made by Electrostatic Field

Alexander V. Frolov, Russia

General Director, Faraday Lab Ltd
Tel./Fax.: 7-812-380-3844
Email: office@faraday.ru

We know some simple examples of work making due to the forces, which appear in electrostatic field. Since the field itself does not require energy consumption from the source (not taking into account the charge leakage), then free energy can be received with these forces. The first phenomenon that we are going to consider is a phenomenon of **electrostriction**. Mechanical forces appearing in dielectric under the influence of electrical field tend to deform it. Sometimes these forces are small, but they can be huge in special dielectrics. The character of deformation also depends on non-linearity of electrical field. Suffice it to remember that heating of a body takes place during deformation, and we can make a conclusion on possible using of the given effect in practice. Periodical deformation can be created due to the rotor or any other methods to change the field.

The second effect is not so popular. It was found by Hertz in 1881 that there is a rotation of dielectrical cylinder (or ball) in permanent electric field if it is created in some liquid or in gas (see Fig.1).

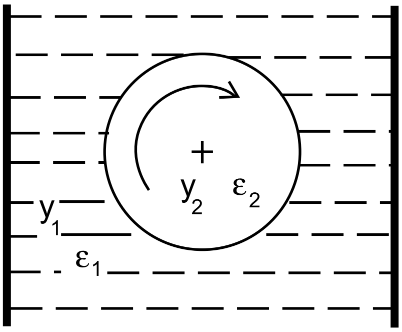

Fig.1

Here **$\varepsilon1$** is a dielectrical permittivity of the liquid, **$\varepsilon2$** is a dielectrical permittivity of the rotor, **$\gamma1$** is a conductivity of the liquid and **$\gamma2$** is a conductivity of the rotor. This effect was discovered by Hertz, then it was described by G. Quincke, Germany. Also Japan scientist I. Sumoto studied the effect in 1955. Modern research work was made by K.M. Polivanov, Moscow. There is the so called equation by Polivanov describing special conditions, which are necessary to begin rotation (**$\varepsilon1/\varepsilon2=\gamma2/\gamma1$** is the Polivanov's condition). Forces of electrostatic field produce this rotation and it is a real free energy system that can trap energy from inner structure of potential field. Let's try to describe the mechanism of this effect and then we'll be able to increase the power of the system up to the level of industrial application. So, why does it work? To my mind, there are some differences in conditions for polarization of the rotor and molecules of liquid or gas those are surrounding the rotor since there is a difference in permittivity and conductivity.

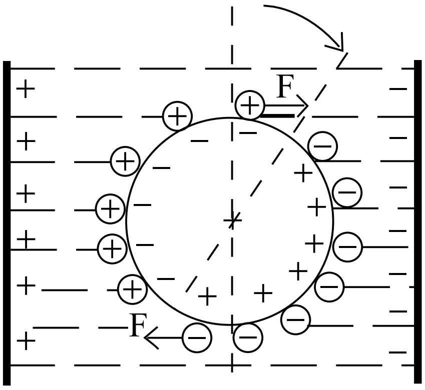

Fig.2

Due to this reason the molecules are polarized both by the field of electrodes and by the field of rotor. The rotor is polarized in the electrical field of electrodes, and molecules of liquid on its surface are polarized by the electric charge of the rotor, but not according to direction of the field. So, the rotor is surrounded by "screen" of molecules on its surface. The electric charge of this screen partially compensates the field of electrodes.

When some angle is created by the first initial turn, there is some part of the rotor surface, where molecules of liquid on the surface of the rotor are attracted to electrodes and it is the reason for future rotation. After some angle of the turn the polarization in this point of the rotor surface is changed but new molecules are incoming in the so called "sector of attraction" and the rotation is always accelerated. This well-known effect of 1881 is a very good example of possibility to produce useful work in load by means of electrostatic field only.

One more interesting effect is known as Faraday effect. In 1836 Faraday noted that raising of liquid upwards takes place in condenser with liquid dielectric (plates are installed vertically). During this process electrostatic field makes the work against gravity. A. Gyemant in 1926 established that value of this force is proportional to the square of voltage on condenser. Later in 1955 this phenomenon was studied in details by I. Sumoto, Japan. That's why in physical encyclopedias this effect is called "Faraday-Sumoto effect". In Tareev's book [1] it is pointed: "Under high voltage this phenomenon leads to flowing and turbulent boiling of liquid". I should remember for those, who see nothing unusual here, that it does not necessary to consume the power from initial source. Creating an initial field in electric capacitor, then we can use the received mechanical work. It is evident that the gradient of electrical field creates the conditions for movement of liquid dielectric. A task to create conditions for liquid circulation is more complex, but it has solution, if we take into account the presence of one more static field, i.e. gravitational field. Joint action of electrostatic and gravitational fields stipulates the circulation of liquid dielectric even in the simplest constructions. Besides, electrostatic filed can be partially screened.

The analyzed examples should draw attention of experimenters to these obvious ways of how to create useful work by means of electrostatic field.

References

1. Tareev B.M. Physics of dielectrical materials, Moscow, Energoisdat, 1982, p.199 – 200.

Editor: Below we publish the comments of our reader.

Letter On

Hertz-Quincke-Sumoto Effect

from Doug Marett

Email: doug.marett@sympatico.ca

…As far as I can tell, the original reference by Hertz is "On the Distribution of Electricity over the Surface of Moving Conductors", Wiedemann's Annalen, 13, pp.266-275,1881. Hertz's discussion of this phenomenon is cursory at best, having very little experimental work included. The real phenomenon appears to have been discovered by W. Weiler in 1893 (Zeitschrift fur den Physikalischen und Chemischen Unterricht, Vol.6, pp194-195). Weiler observed that a glass cylinder placed in a poorly conducting liquid between two spherical electrodes began to rotate when the electrodes were connected to an electrostatic generator. In 1896, George Quincke reported the same phenomenon and published a comprehensive report on it. This was in Annalen der Physik, Ser.3, Vol. 59, pp.417-486. Subsequent investigators have tended to attribute the discovery to Quincke, when in fact Weiler was the first to pioneer the work.

I do know that P.E. Secker et al. performed work on this field, references of which are available in the English language. These are:

P.E. Secker, et al., (1968) Journal of Applied Physics, Vol. 39, pp.2957-2961, and
P.E. Secker, et al., (1970) Journal of Physics D: Applied Physics Vol.3, pp 216-220.

I hope that this information might be of use to you.

BOOKS REVIEW

The Foundations of Physchemistry of Micro World

Ph. M. Kanarev

The Second Edition

The book was published in Russian and English; you can also read it at http://book.Kanarev.innoplaza.net

The new axiomatic of natural sciences is given in the book; on its basis, quantum physics and quantum chemistry have been returned to the classical way of development. The first steps are made on this way, which have led to discovery of the structure of the photon, the electron, and the principle of the formation of the atomic nuclei, the atoms and the molecules. The Planck's Law of radiation of perfect blackbody is given on the basis of classical concepts, and the connection of quantum phenomena with the laws of classical physics is proved.

The application of the new theoretical results to the solution of practical energy tasks on the basis of plasma electrolysis of water is shown. Due to this electrolysis, additional heat energy and the energy containing gases (hydrogen and oxygen) are generated. Cold Nuclear Transmutation of the atomic nuclei of alkaline metals and the atomic nuclei of the cathode material takes place during plasma electrolysis of water.

The book is intended for physicists, chemists and other specialists who are seeking the new directions for understanding the foundations of the micro world and the new energy sources.

Comments to Electro-Reactive LIFTER News

Alexander V. Frolov, Russia

General Director, Faraday Laboratories Ltd
E-mail: office@faraday.ru
http://www.faraday.ru

March 21, 2003 we got news from France about electro-reactive propulsion system.

Blaze Labs reached 100g PAYLOAD level.

You will find all the details of the Saviour's experiment at: http://blazelabs.com/exp14.htm.

This electro-reactive ionization-flying model requires 46 kVolt, 4 mA, i.e. about 200 watts to fly with 100 g useful load.

It is interesting to note that Jean-Louis Naudin's previous opinion of 1997 – 1999 about priority of the Frolov's asymmetrical capacitor now is changed. Sometimes people wish to re-write the history. It is possible and it is easy. Naudin's web page about http://jnaudin.free.fr/lifters/story.htm (see Fig. 1) is not started from the real beginning of the story. He forgot our discussion of 1997 and also photos and MPEG video I have send to him when he has started his way in electrograviticis.

If you remember the name of Naudin's first asymmetrical capacitor "Frolov's Hat" (see Fig.2) then you know why there is this name of this technology. But really it is not important to try for some serious reply from Jean since he has no personal post address, phone, or his photo on the web site...

Also it can be useful to visit this page http://www.faraday.ru/t-cap.html to know about something **more important than electric ionization flying models (Lifters),** since it is just a reactive ionization way, which is similar to rockets.

Ideas now developing by us are not aimed to increase the power level (from 1 gram to 100 gram propulsive force). It is **a qualitatively new idea**; it is really electrograviticis instead of "electro-reactive Lifter".

So**, what is about real "history of the question" instead of the French version?** The question here is not about a priority, it is nonsense after T.T. Brown's patents. As an example the NASA patents on asymmetrical capacitors can be mentioned since the ideas were opened before the NASA patent and there is nothing new in it. The problem is that people who develop reactive ionization Lifters develop only primitive reactive technology. **It is not essence of T.T. Brown research but it is the distortion of his ideas.**

So, the team in France is working to develop and support ideas of secondary and wrong method. Why? Perhaps they either do not see the real way or have payment for the wrong way **to mask a real work.**

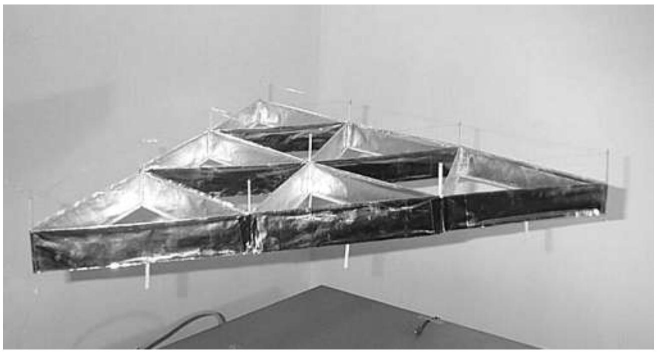

Fig. 1

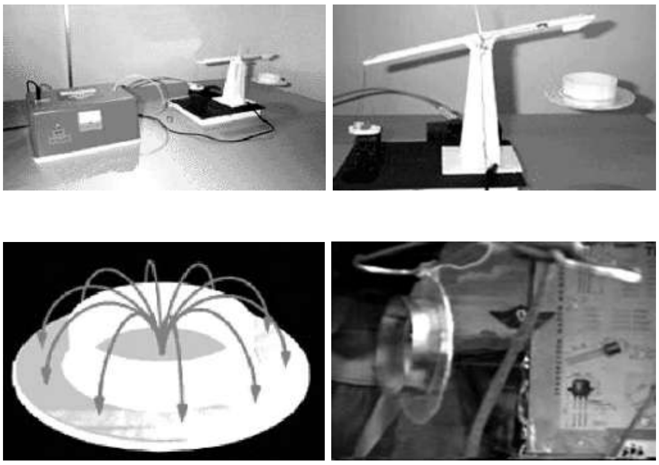

Fig. 2

Questions & Answeres

Josh Werrmann (Email: jsh111@yahoo.com): I wanted to ask you a few questions about an article titled “Highly Efficiency Electrolysis” by A.V. Frolov. In this article it is stated that Dmitry A. Latchinov patented a method for electrolysis, where one of the electrodes is insulated completely from the water. It is also stated that another scientist, Igor Goryachev, used a method of pulsed electric fields. I don’t know if you know, but inventor Stanley Meyer has a patent that covers both of these processes together, in a way. His patent is U.S. patent number: 4936961. Take a look at this. In the patent it states EXPLICITLY that there is a pulsed electric field, where a condition of NO CURRENT is preferred. This is very similar to the above to Russian scientists, except that the patent states a condition of resonance is necessary. I am wondering do the two scientists that are described require resonance in their devices? Or just plain pulsed electric fields, where one electrode is insulated from water.

Another patent I would like to refer you to is Archie Blue’s electrolyser, U.S. patent number: 4124463. This device is similar to A.V. Frolov’s idea, to get the oxygen and hydrogen bubbles off the electrodes, except he uses a blower to blow air in the cell, instead of rotating the cell like A.V. Frolov does. His invention also uses the conductivity current to work.

Alexander V. Frolov: As far as I know resonance is not created in this case. The main idea is the removal of gaseous film which appears on the surface of electrodes by rotation or some other methods.

NEWS REVIEW

Giant magnets lead the world to the new source of energy

http:// www.sciteclibrary.ru

The system of 150-tons magnets which have been developed by scientists of the USA, Japan and Russia opens a new era of safe and unlimited energy. In this system nuclear fusions serve as a source of energy. They accumulate heavy chemical elements not by means of nuclear reactions but by ultrahigh plasma pressure.

After supplementary research is held which aims to reduce manufacturing costs of the giant magnet weighing 925 tons, this magnet will be produced and demonstrated at the opening of the International Thermonuclear Experimental Reactor (ITER). This magnet in its turn will become a part of a bigger system, that is combined into a magnet weighing approximately 10 000 tons. The task of ITER is the demonstration of the process of nuclear fusion, as an energy source. During the process of nuclear fusion the light elements are combined by immense pressure, thus producing heavier elements. During this production process a lot of energy is emitted. The task of giant high-power magnets is to create magnetic fields, which must hold and control plasma, or to charge electrically the gas, in which the fusion is being held.

In Japan the system of 150-tons magnets serves as adjustment system of 925-tons magnets, which will be finally put into operation and heat up the ITER plasma. Two additional giant magnet systems will confine the plasma and will control its form. The model for one of them is presently being examined in Germany; the model of the other is in project.

Cylindrical 150-tons magnets have three main parts: external module, built by the Japanese team, internal module, built by the American team and a thin rod in the core which is equipped with process control equipment. Three different rods were separately checked, two of them were built in Japan, another in Russia.

Harnessing the Wheelwork of Nature

(340 pgs. Adventures Unlimited Publishers, 2002)

Thomas Valone
Email: iri@erols.com

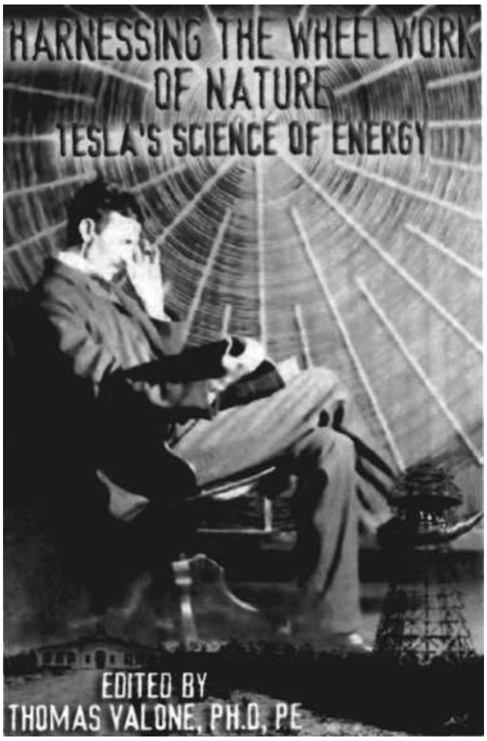

It is a new book by **Thomas Valone**, who edited this anthology in time for the **Wardenclyffe Tower Centennial (1903-2003)**. This book presents for the first time, the feasibility argument for Tesla's most ambitious dream, the wireless transmission of power. Pictured on the book's cover near his feet, the 187-foot Wardenclyffe Tower was Tesla's means to deliver natural 8 Hz electricity anywhere in the world, by longitudinal waves.

Unknown to most electrical engineers, Nikola Tesla's dream answers the energy crisis worldwide, saves electrical conversion losses, and provides a real alternative to transmission lines. In Dr. Corum's contributed papers, he explains Tesla's magnifying transmitter, which Tesla compared to a telescope. Corum points out that "the tuned circuit of his magnifying transmitter was the whole earth-ionosphere cavity resonator." This fact helps explain why Tesla stated, "When there is no receiver there is no energy consumption anywhere. When the receiver is put on, it draws power. That is the exact opposite of the Hertz-wave system...radiating all the time whether the energy is received or not." Thus, with Tesla's futuristic transmission of power, source dissipation will only be experienced when a load is engaged in a tuned receiver somewhere on the earth. This fact alone represents **a major leap forward** in electrical transmission efficiency, even one hundred years later.

Why wasn't the prototype of Wardenclyffe finished in 1903? Tesla offered this visionary conclusion: "The world was not prepared for it. It was too far ahead of time. But the same laws will prevail in the end and make it a triumphal success... Let the future tell the truth and evaluate each one according to their work and accomplishments. The present is theirs; the future, for which I really worked, is mine."

Up until now, there has been a general malaise regarding the lack of scientific comprehension of Tesla's greatest dream. For example, the Serb National Federation notes, "With the exception of the first biography of Tesla by John J. O'Neill, science editor of the *New York Herald Tribune*, and published in 1944, unfortunately no biographer since has had the necessary scientific/engineering academic credentials to discuss Tesla's work in the various fields." Contributors to ***Harnessing the Wheelwork of Nature*** are primarily physicists and engineers who are experts in Tesla technology. Their wealth of knowledge demonstrates their mastery of this extraordinarily progressive and technical subject. Finally, the best academic credentials have been brought to bear on the world's greatest electrical futurist.

This is a very readable and profusely illustrated reference volume on wireless transmission of power, besides being an excellent biographical gold mine of Tesla history. Nick Cook, editor of *Jane's Defence Weekly* and author of *The Hunt for Zero Point* says, "Tesla is one of the great overlooked geniuses of science and electricity. His full story deserves to be told. Tom Valone sheds important new light on his life and work."

Integrity Research Institute will hold a Tesla Energy Conference & Expo on November 8-9, 2003 to celebrate the Wardenclyffe Tower Centennial (1903-2003) in conjunction with the Tesla Wardenclyffe Project, Inc. (www.teslascience.org).

See www.IntegrityResearchInstitute.org for more details.

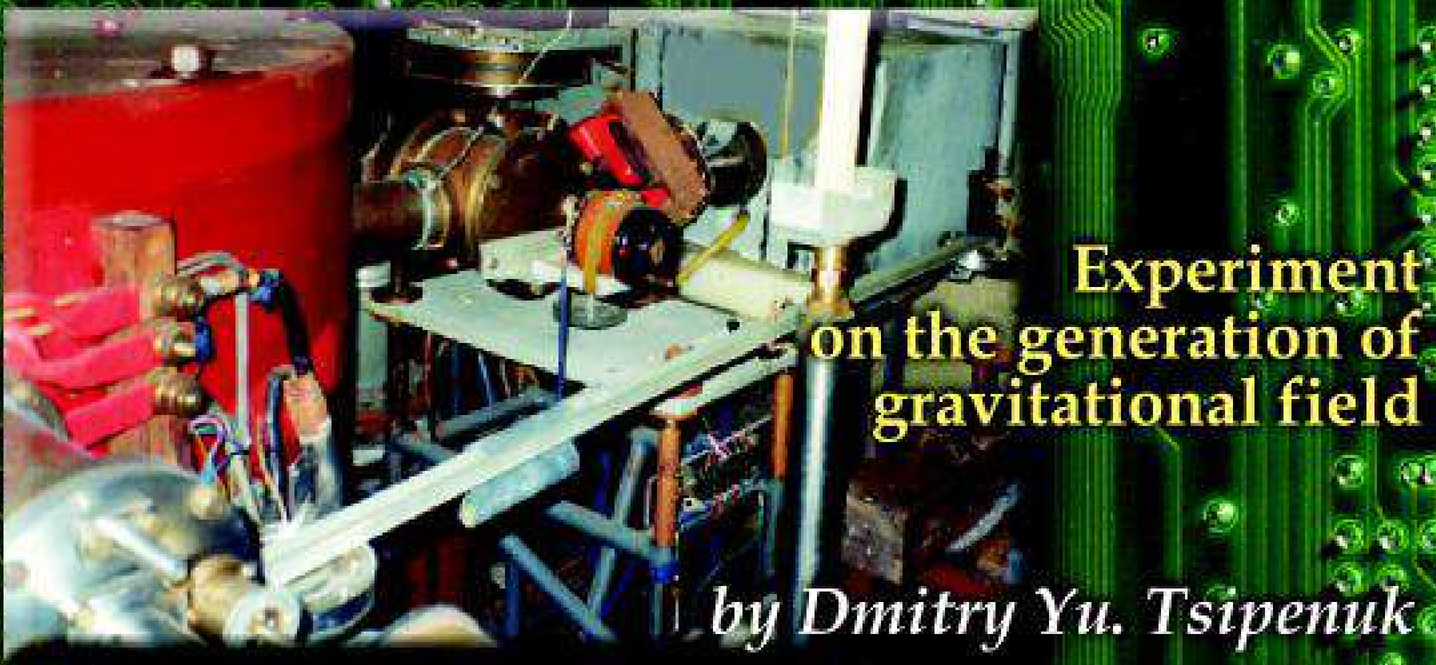

Experiment on the generation of gravitational field

by Dmitry Yu. Tsipenuk

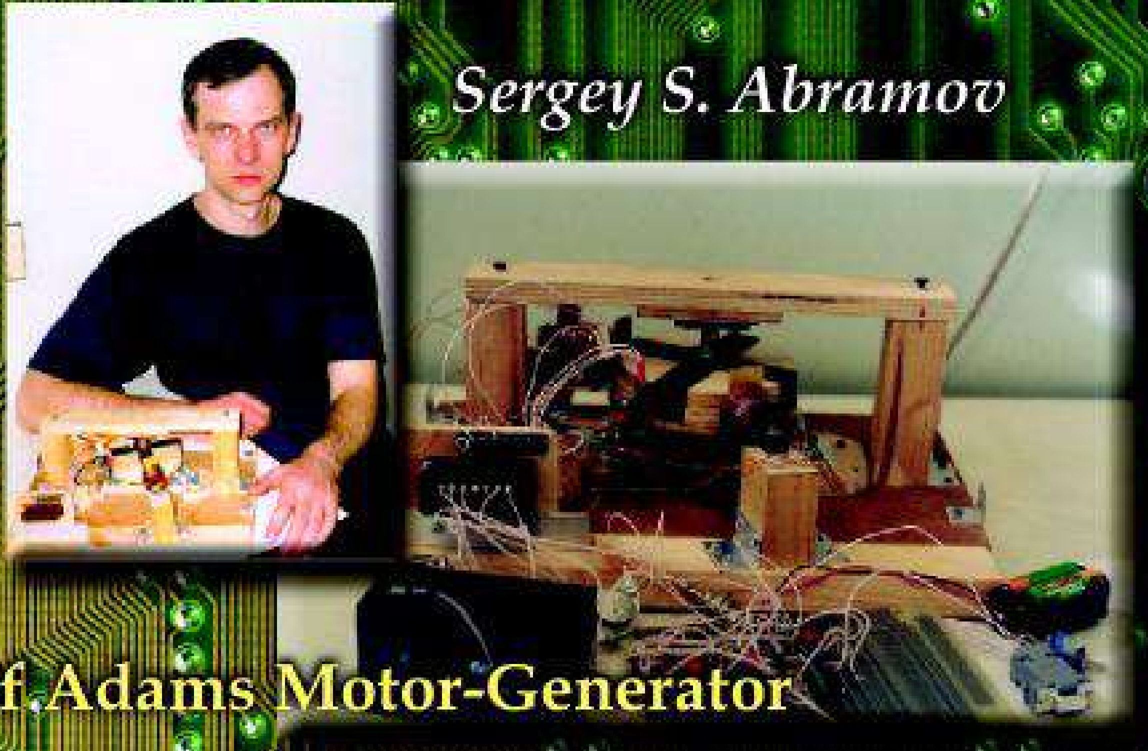

Sergey S. Abramov

Model of Adams Motor-Generator

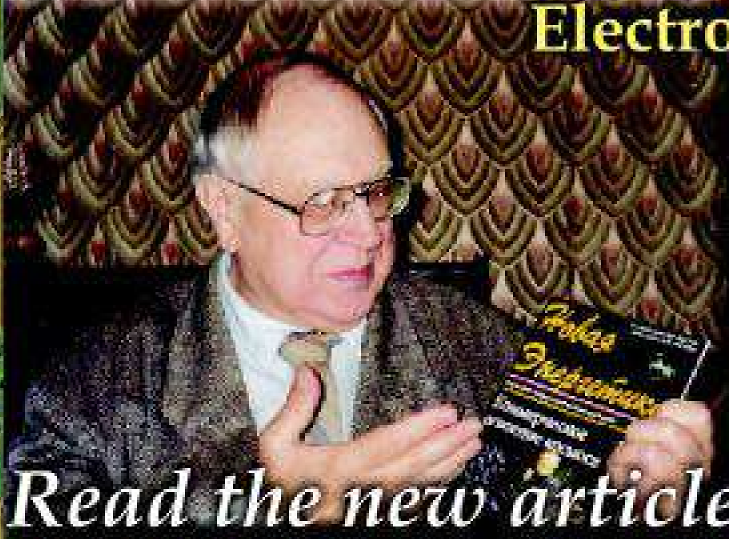

Electrolytic photoaccumulator for the XXI century energy systems

Read the new article by Anton I. Egorov

Flynn Generator

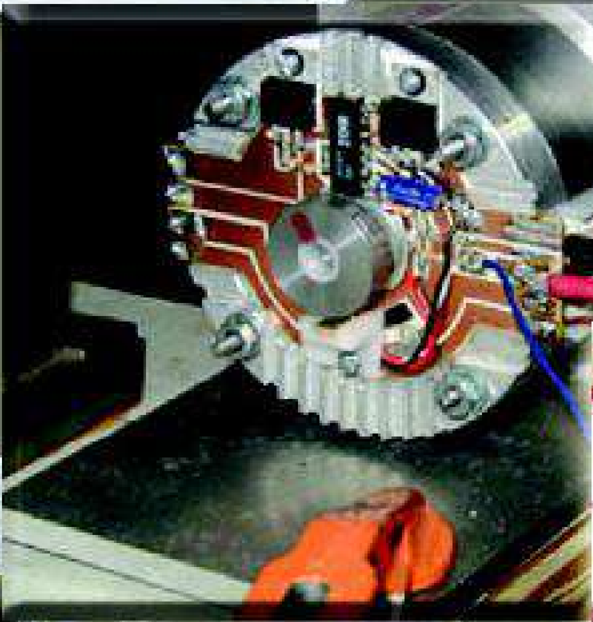

Original version of photomaterials is presented at http://www.flynnresearch.net

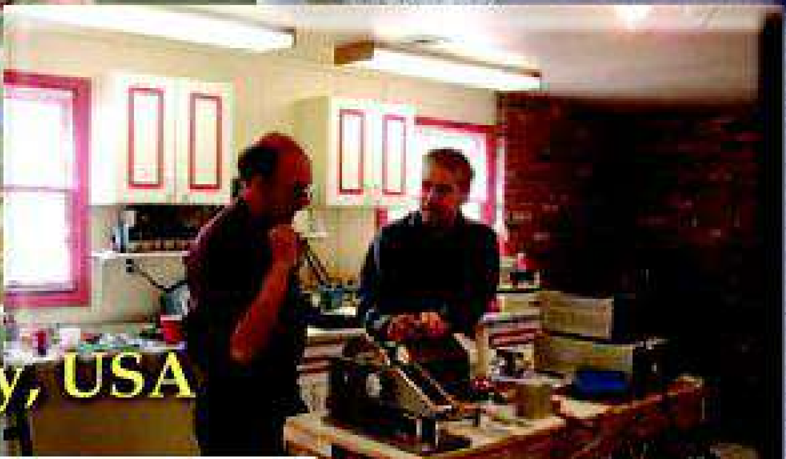

Joe Flynn Laboratory, USA

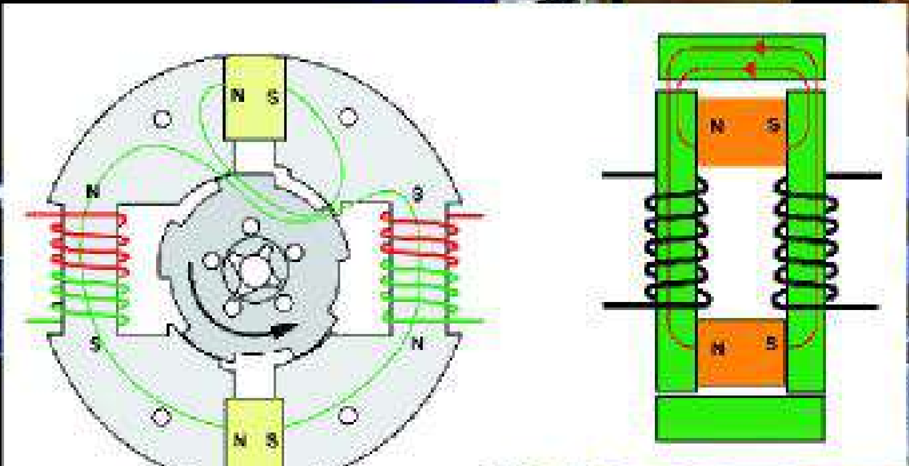

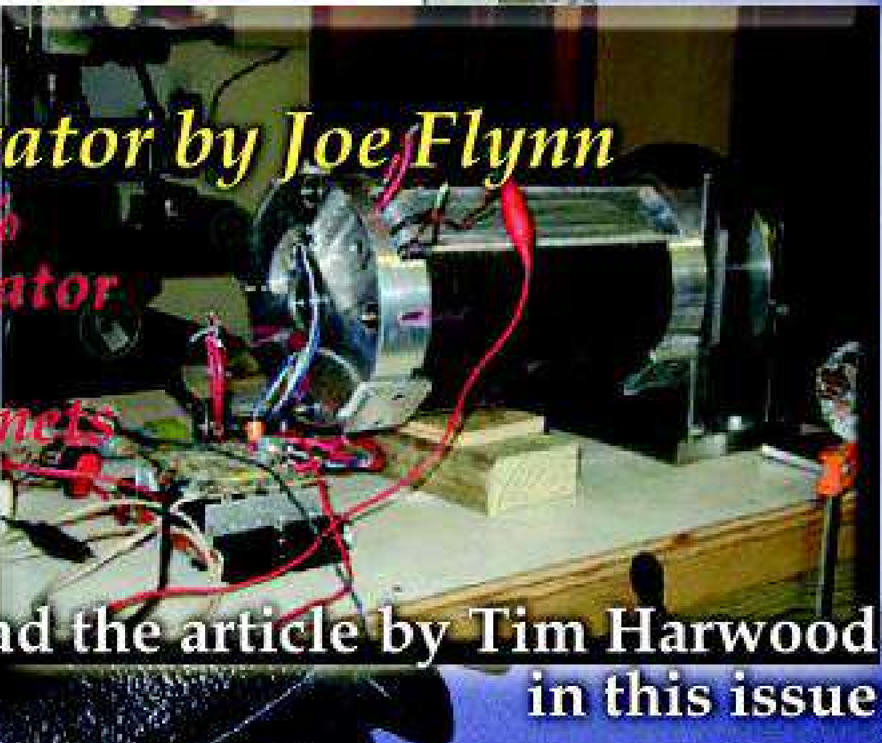

Electric generator by Joe Flynn

More than 350% efficiency generator operating on permanent magnets

ISSN 1684-7288

Read the article by Tim Harwood in this issue

Printed version ISSN 1684-7288
Online version ISSN 1684-9698

New Energy Technologies

Published by Faraday Laboratories Ltd. **Issue #3 (12), 2003**
Research on advanced aerospace systems and new energy sources

Time Machine

Conference in Moscow: discussion of the patent application on the method and device to control temporal parameters of matter

Time Machine

Experiments by research group "KOSMOPOISK" Summer, 2001

Device designed
by Faraday Lab Ltd, 2002

Alexander V. Frolov
Vadim A. Chernobrov
April 12, 2003, Conference hall
Hotel "Cosmos", Moscow

NEW ENERGY TECHNOLOGIES #12

1. Works on the Designing of Time Machines. Vadim A. Chernobrov (Russia)
2. Control of Temporal Parameters of Physical Processes. Alexander V. Frolov (Russia)
3. Etherodynamics as a New Field of Physics. Vladimir A. Atsukovsky (Russia)
4. Electromagnetic Gravitational Interaction Vadim Ya. Kosyev (Russia)
5. Medium For Existing of Matter in Nature. Anatoly V. Rykov (Russia)
6. Adams Motor. Sergey S. Abramov (Russia)
7. Field Transformation in the Model of Extended Space. Dmitry Yu. Tsipenyuk (Russia)
8. Joe Flynn's Parallel Path Magnetic Technology. Tim Harwood (USA)
9. Flynn Laboratory Photos
10. Flux-machine and Its Analogies. Review
11. Russian Patents on Alternative Energetics.
12. Gravitational Spaceships. Gregory R. Uspensky (Russia)
13. On Velocity Of Drive-Free Motion. S.A. Gerasimov , V.V. Stashenko (Russia)
14. Electromagnetic Self-Action. S.A. Gerasimov, A.V. Volos (Russia)
15. Uranium Photoaccumulator. Anton I. Yegorov (Russia)
16. Global Energy Prize (reportage)
17. Teleportation. Alla V. Pashova (Russia)
18. "Mass Defect" in Home Conditions. Prokofy V. Sherbak (Russia)
19. Aether as Unified Field. Alexander M. Mishin (Russia)
20. Global Energy. Philip M. Kanarev (Russia)
21. Fusion Processes Of Molecules Of Oxygen, Hydrogen And Water. Philip M. Kanarev (Russia)
22. Plasma Energy Power Generation. Bruce A. Perreault (USA)
23. News
24. The Marcus Device Controversy. Tim Ventura (USA)
25. Systems of conversion of thermal energy to mechanical one. Review
26. I. Prigozin
27. Letters

CONFERENCE "THE TIME MACHINE"

Editorial: On April 12 of 2003 a scientific conference "The Time Machine" was organized by Faraday Laboratories Ltd in Moscow, Russia. It was devoted to the experiments on control of space-time physical properties. At the conference there were discussed problems of time and gravitation in the context of etherodynamics, experiments and applied aspects of these technologies. Below we publish a review of the main reports presented at the conference.

The First Results of the Works on the Designing of Devices to Control Parameters of Physical Processes (of Time)

Further Plans on Designing of the Time Machine

Vadim A. Chernobrov, Russia

"KOSMOPOISK", Nagatinskaya 19A, 111533,
Moscow, Russia
E-mail: chernobrov@kosmopoisk.org

To conduct the experiments on the influence upon physical Time (density of space energy) some special devices were used as a general method of such influence. These devices can create converging waves which can cause the appearance of quasi-monopole in the confined space. *Quasi-monopole* is a part of space which has some parameters of hypothetical unitary monopole or bunch of such particles (in particular, it allows registering one magnetic pole by means of measuring equipment from the outside at some distance from this pole).

The pilot experiments have shown that it is very difficult (if not impossible) to create long-living quasi-monopole by means of permanent magnets or electromagnets operating on direct currents (in this case quasi-monopole represents a space with one outer and one inner magnetic pole). It can be explained by the fact that lines of force of the "inner pole" invariably find a weak spot in the heterogeneous surface of magnets and break out. As a result, at the device along with one "outer" magnetic pole there is a local output of magnetic lines of the "inner" pole.

During the designing of new devices there was a task to create a quasi-monopole situation in the confined space. This situation should be created not uniformly but transiently by pulsation method. Frequency of work of electromagnetic oscillators first of all was selected depending on linear dimensions of the devices. Selection of the frequency was made in such a way that one period of pulsation does not exceed the period of time which is necessary for electromagnetic waves to reach the center and opposite waves of the device.

At the designing of laboratory devices, which generate converging longitudinal waves, several different principles to obtain required parameters were considered in the multi-layer quasi-monopole, which works in high-frequency mode. Several types of the devices were realized at different degrees of development and with different success. It was shown that the simplest devices were that ones which used electromagnetic (solenoid) oscillators connected in series and in parallel. In different experiments between 3 and 5 such surfaces were used. These surfaces were called electromagnetic work surfaces (EWS). All layers of EWS of different diameters were mounted in series in each other (like matreshka). The outer layer was either mounted at force shell or simultaneously represented such a shell by itself.

The size of the maximal EWS was about 0.9 m, the diameter of the minimal (inner) EWS was equal to 115 mm that was enough to place laboratory animals inside the control detection devices. Laboratory animals were used to determine the consequences of the influence of converging spherical electromagnetic waves. Payload, (i.e. that which was directly used for transfer in Time-Space), included the afore-mentioned detection devices and (occasionaly) laboratory animals. The term "payload" was coined by analogy with the term which is used in cosmonautics. Volume of payload section was placed in the center of symmetry of the Time Machine (TM). In all the earliest Machines (except the 7th model) this volume still has not exceeded the volume of a football. The device with an outer diameter of 2.1 m and inner payload section of 1 m has the maximum size. It allows making human-aided experiments.

At different stages of the experiments the calculations were made by all available known modern methods to fix time. All types of electronic, quartz, mechanical and several specially made doubled quartz generators were used (there were compared frequency readings of measuring and etalon heat-insulated generators which were placed at a distance). Certain experiments used lightguiding diodes and some other methods. Before and after the experiment (more rarely during the experiment) readings of measurement clocks were periodically compared with those of an etalon clock and with signals of exact time which were transmitting by radio. Other physical factors causes side effects upon some types of measuring devices, for example, upon quartz-crystal clocks. However, doubling of measurement methods allowed essentially decreasing inaccuracy of measurements.

At some operating modes (which were not always predicted) the change of Time speed was attained (Prof. Nikolay A. Kozyrev called it density of Time t/t_0). This change came to about a fraction of a second per an etalon hour. Let us take usual "etalon" earth time as $t_o=+1$ then it becomes clear that it is speed range of $+0.99<t/t_o<+1.01$ which is researched in the experiments. Thus being placed inside the device, the inanimate objects and animals were transferred in the Future (with "above-zero" speed) more slowly or faster than those around them. It occurred at all operational modes of the device (deceleration or acceleration).

Difference of Time (gradient of Time speed or curvature of Space-Time field) was observed not only inside the TM, though there is no doubt that a maximal value of changed Time was registered inside the smallest "matreshka". As was expected, during the experiments Time change was also registered outside the device. However such change which had an opposite sign was smaller than the change inside the device approximately by an order of magnitude. It entirely corresponds to geometric laws, i.e. in proportion to the cube of the distance).

In other words, TM influences not only its inner part and pay load but also the environment. It bears a strong resemblance to jet propulsion but in Time and not in Space. It is a flight which is realized by rejection of Time instead of mass.

Thus it was determined that the processes of deceleration and acceleration of Time distinctly differ in their nature and consequences. In such a way the deceleration was considerably smoother and more stable. During acceleration there were observed sudden changes in readings and this operational mode was characterized by general instability and dependence on any (or many) external factors. Particularly, acceleration instability also consisted in the following: at fixed power the value speed of Time depended on the time of day and situation of the Moon. Possibly it also depended on some other factors including the presence of an operator or other people near the device. Even insignificant external influence, for example, mechanical shaking, caused the change of speed value which sometimes was rather significant.

Inside the laboratory devices there was also registered that **Time can change with some sluggishness**. After the changed Time speed influenced some physical object (for example, soil) then residual effects were observed at it for some period of time. These effects can be eliminated only by the influence of another speed of Time.

Preliminary general conclusions

The Present is the transfer or transformation of the polyvariant easily changeable Future into the univariant unchangeable Past. Thus flights in the Past (at "negative" density-speed t/t_0) and in the Future will happen in different ways. They can be compared with the motion of an ant along the tree: from any point of the tree (i.e. from the Present) there is only one way downwards open to the ant (i.e. in the Past) and many different ways upwards (i.e. in the Future). However among all the ways to the Future undoubtedly there are most probable, low-probable and almost improbable variants for development of the Future. The less probable this variant of the Future the more unstable and energy-intensive the motion to the Future will be. According to the "law of the crown of tree", return to the Present is possible only if when being in the Past the traveler does not interfere in the course of events and does not change the course of the past History. Otherwise the chronotraveler will return to the parallel Present from the Past by another branch of History. Penetration to the Future from the Present is hampered by the choice of the branch for transference. However return from any variant of the Future into the Present is possible at any scenario if there are no fusions of different variants of History…

In other words, we have a circumstantial evidence of the assumption that Time has more than one dimension. Thus there is an affirmation of the theoretical inferences of R. Bartini who believed that Time has 3 dimensions. Hence we can consider our terrestrial globe as 6-dimensional where the dimensions are: length; width; height; age or date of Time; variant of History or blur of Time; density or speed of Time. Thus the notion of "Arrows of Time" is completely absent in the fourth dimension (date of Time) but it is a special case of the notion of the sixth dimension i.e. speed of Time. At the same time the notion of speed of Time also relates to the physical notions of gravitation and energy. Thus the notion of "Einstein-Rosen bridge", which was introduced in 1916, or notion of "worm course", which was introduced by John Willer in the end of the fifties, are connected with transference in the 5th and 6th dimensions.

◆

Method and Device to Control Temporal Parameters of Physical Processes by Means of Changing of Energy Density of Space

Alexander V. Frolov, Russia

General Director, Faraday Lab Ltd
Tel./fax: 7-812-380-3844
Email: office@faraday.ru

This invention belongs to the methods and devices to provide control on rate of physical processes (that includes the process of the existence of matter in space-time) by means of increasing or decreasing of energy density of space (i.e. energy density of physical vacuum or density of the aether).

Let us consider the history of the invention:

Earlier there were proposed some methods and devices to influence the rate of physical and chemical reactions, biological processes or period of oscillation of the system. In the works by N.A. Kozyrev [1] there is a description of the experiments on the influence of some process (for example, process of evaporation or crystallization of matter) upon the period of another process, which serves as a detector and can be compared with reference oscillation process. In one case, the rate of oscillations of the detector decreases in the surrounding area near the process of matter evaporation. In another case, the rate of detector oscillations increases in the surrounding area near the process of matter crystallization. If we use a term of "entropy" then it is possible to say that the processes which are accompanied with entropy increasing (for example, conversion of matter from solid state into liquid one) influence on the matter (surrounding processes) in such a way that entropy of systems decreases. In another case, for example, near the process of crystallization, entropy of systems increases in the surrounding area near this process. Kozyrev used the term "wave of density of time" and he made a conclusion that in addition to "directivity" of time (time course) there are active properties of time, for example, "density of time".

To develop this approach for applied purposes it is necessary to use in-depth analysis of the physical sense the "law of the crown of tree", return to the Present is possible only if when being in the Past the traveler does not interfere in the course of events and does not change the course of the past History. Otherwise the chronotraveler will return to the parallel Present from the Past by another branch of History. Penetration to the Future from the Present is hampered by the choice of the branch for transference. However return from any variant of the Future into the Present is possible at any scenario if there are no fusions of different variants of History…

In other words, we have a circumstantial evidence of the assumption that Time has more than one dimension. Thus there is an affirmation of the theoretical inferences of R. Bartini who believed that Time has 3 dimensions. Hence we can consider our terrestrial globe as 6-dimensional where the dimensions are: length; width; height; age or date of Time; variant of History or blur of Time; density or speed of Time. Thus the notion of "Arrows of Time" is completely absent in the fourth dimension (date of Time) but it is a special case of the notion of the sixth dimension i.e. speed of Time. At the same time the notion of speed of Time also relates to the physical notions of gravitation and energy. Thus the notion of "Einstein-Rosen bridge", which was introduced in 1916, or notion of "worm course", which was introduced by John Willer in the end of the fifties, are connected with transference in the 5th and 6th dimensions.

◆

Method and Device to Control Temporal Parameters of Physical Processes by Means of Changing of Energy Density of Space

Alexander V. Frolov, Russia

General Director, Faraday Lab Ltd
Tel./fax: 7-812-380-3844
Email: office@faraday.ru

This invention belongs to the methods and devices to provide control on rate of physical processes (that includes the process of the existence of matter in space-time) by means of increasing or decreasing of energy density of space (i.e. energy density of physical vacuum or density of the aether).

Let us consider the history of the invention:

Earlier there were proposed some methods and devices to influence the rate of physical and chemical reactions, biological processes or period of oscillation of the system. In the works by N.A. Kozyrev [1] there is a description of the experiments on the influence of some process (for example, process of evaporation or crystallization of matter) upon the period of another process, which serves as a detector and can be compared with reference oscillation process. In one case, the rate of oscillations of the detector decreases in the surrounding area near the process of matter evaporation. In another case, the rate of detector oscillations increases in the surrounding area near the process of matter crystallization. If we use a term of "entropy" then it is possible to say that the processes which are accompanied with entropy increasing (for example, conversion of matter from solid state into liquid one) influence on the matter (surrounding processes) in such a way that entropy of systems decreases. In another case, for example, near the process of crystallization, entropy of systems increases in the surrounding area near this process. Kozyrev used the term "wave of density of time" and he made a conclusion that in addition to "directivity" of time (time course) there are active properties of time, for example, "density of time".

To develop this approach for applied purposes it is necessary to use in-depth analysis of the physical sense of the "time density" notion. Connection of notions about "time directivity" and "entropy of the system" was demonstrated in the work "Introduction into thermodynamic of irreversible processes", 1964 by Ilya Prigozhin [2]. In the work "Quantum fluctuations of vacuum in curved space and theory of gravitation" by A.D. Sakharov [3] the conception of vacuum structure was suggested. In the USA the work "Can the vacuum be engineered for space flight applications?" by H.E. Puthoff [4] is well-known. The author considered the applied aspects for study of vacuum structure and described the method and device to obtain propulsive force by means of changes in vacuum properties.

A substantive conception of time and methods for creation of waves of energy density were also considered by Prof. K.P. Butusov in the work "Time is a physical substance", 1991 [5]. In the book "What is The Time?" by Yu. G. Belostotsky [6] the connection between the notions of time and aether was demonstrated. This connection was considered from the point of view of astrophysics there.

We can also say that the modern conception of aether is successfully developed by V.A. Atsukovsky in his works [7].

In my articles, for example, "Physical principles of the Time Machine" [8], it was demonstrated that to develop experimental works on the topic it is useful to clarify the terminology and to consider **"waves of time density" as longitudinal waves of energy density in space.** In this case the notion of "time density" has a physical sense of energy density (aether density).

This approach can be realized practically by means of classical electro-technical and radio engineering methods and it is a development of aether-dynamical conception on the nature of electricity and magnetism by M. Faraday, "Experimental researches on electricity", volume 3, [9].

Let us consider a usual bipolar magnet from the point of view of aether-dynamics. Then it is possible to say that it is inflow and outflow of aether, i.e. it is a balanced energy system which does not change energy density in space. In this case it is evident that creation of magnetic monopole or modeling of quazi-monopole by means of electro-dynamical methods is a technical basis to create some local change of energy density in space.

Electric processes could also be used alongside with the magnetic phenomena. For example, in another book "Symmetrization of Maxwell-Lorenz equation" by Prof. Butusov [10] the creation of longitudinal wave was also considered. It was demonstrated that an electrically charged sphere can radiate longitudinal wave when the sphere radius is changing, i.e. when its surface is changing while the value of electric charge is constant.

One more well-known method is described in the book "Experimental gravitonics" by Polyakov [11]. There is a consideration of the generation of gravitational waves at the high-frequency magnetization and demagnetization of ferromagnetic material, i.e. at the powerful volume magnetostriction. Since at this phenomenon there are changes of matter density (i.e. changes of energy density in space, which is occupied by matter), then volume magnetostriction is a special case of changes of energy density.

Earlier Vadim A. Chernobrov had described a method and device to control temporal characteristics of physical and chemical processes by means of creation of the magnetic monopole (quasi-monopole). In this magnetic mono-pole there is a convergent wave, which is created by several sources situated in the spherical frame. According to this method in the multilayer spherical structure where every layer (the so called "electromagnetic work surface") is an assembly of electromagnets, by means of series connection of the layers the wave is created, which converges to the center of the device. The device has the same outside magnetic poles of the electromagnets (and the same inner poles) and thus a model of macroscopic magnetic monopole is created.

We assume that at in-phase operation of all sources of waves, interference of longitudinal waves provides some change of value of energy density of space in the focus of the system.

Experimental facts prove that detectors installed in the center of the device (for example, mechanical or electromagnetic oscillators) show change of period in their own oscillations. We have protected them by shielding them from heat radiation as well as from other kinds of electromagnetic influence. Thus it is possible to assert that the detectors decelerate or accelerate their oscillation period depending on the energy density, which is generated in the center of the device.

However for the experiments made by means of such a device, accurate adjustments of all wave sources are required to provide their in-phase operation. At the same time, operational stability of the system depends on the operational stability of each of the wave sources. Increase of the impulse frequency causes increase of the effect; however, it is limited by the parameters of electromagnets and generator of impulses. Besides, to increase the effect it is necessary to use more powerful energy sources since the current in the windings of electromagnets determines the value of the magnetic field of the created quasi-monopole.

Since the efficiency of such systems directly depends on the frequency and the value of changes of energy density in space, then in the next version of realization of this technology we suggest using of plasma shells instead of electromagnetic working surfaces. This will allow significant improvment in the specific parameters of the device.

Thus let us consider general engineering principles of operation and outline the ways to develop this method. Fig. 1 represents a three-layered electromagnetic emitter. This electromagnetic emitter is designed according to the invention in which the directed radiation of wave of energy density is created along the axis of the device.

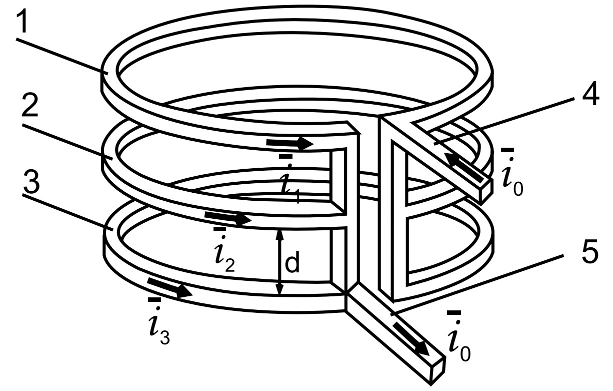

Fig. 1
Three-layered electromagnet

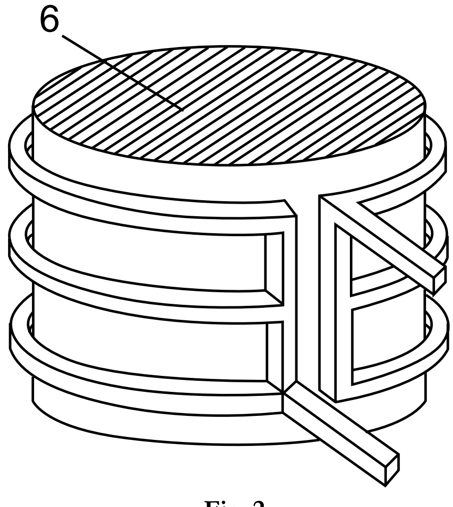

Fig. 2
Three-layered electromagnet with ferrite core

The device is designed according to the idea by Vadim A. Chernobrov for creation of the directed wave of energy density by means of phase shift in propagation of impulse front in three current branches, namely i_1, i_2, i_3. These branches are displaced along the electromagnet axis at some distance d.

The device works in the following way. When the pulsed generator is activated, front of current pulse i_0 appears at the output 4. Impulse front at branch 1 advances impulse front at branch 2 that is caused by spatial shift of current branches 1, 2, 3 relatively to each other along the electromagnet axis at the distance d. Impulse front at branch 2 in its turn advances impulse front at branch 3 for a certain time T. The second output of the electromagnet 5 is placed in such a way that impulse front at branch 1 will phase lag behind the impulse front at branch 2 (which in its turn will phase lag behind the impulse front at branch 3) for the same period of time T. Therefore at branch 5 the united impulse front is generated again.

Time T can be calculated in the following way:

$$T=d/c \text{ (seconds)} \qquad (1)$$

where c is a constant of propagation of impulse front. This constant is known as velocity of light.

At each impulse the T (i.e. the value of relative lag of impulse front) is a constant value. Thus high-frequency consequent excitation of layers of the electromagnet appears at each impulse. The frequency of the excitation is calculated in the following way:

$$f=1/T \qquad (2)$$

where T is relative lag of impulse front in seconds.

There is an example of frequency calculation: for the shift distance d=7 mm we can calculate a lag $T = (7/2.997924)x10^{-11}=2.335x10^{-11}$ (seconds) and frequency $f=1/T$ approximately comes to $4.28x10^{10}$ (Hertz).

Thus this design of three-layered electromagnetic emitter allows creating the waves of super-high-frequency band (for example of millimeter range) without the use of a semiconductor or other radio components.

It is very expedient to use the electromagnets designed with magnetostrictive material cores, that will significantly increase energy density of the longitudinal wave, which is generated by the multilayer electromagnet. In Fig. 2 there is an emitter with the core. In the case of high-frequency ferromagnetic magnetostrictive materials 6 the efficiency of emitter operation significantly increases.

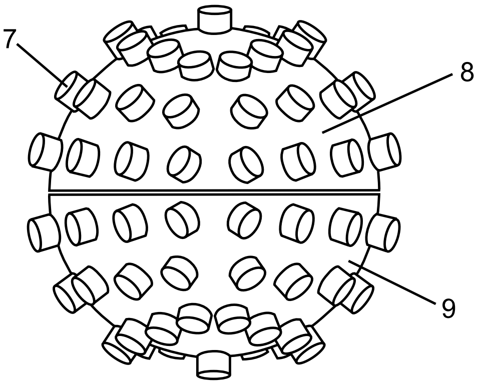

Fig. 3
Spherical design

Fig. 3 represents the spherical distribution of emitters 7 at the upper 8 and lower 9 hemisphere of the frame which could be opened in order to place detectors and different objects inside. It can also allow to establish in what way the changes of density of space energy influence the properties of different materials, velocity of physical and biological processes as well as chemical reactions. The installation of detectors 10 inside the device is shown on Fig. 4.

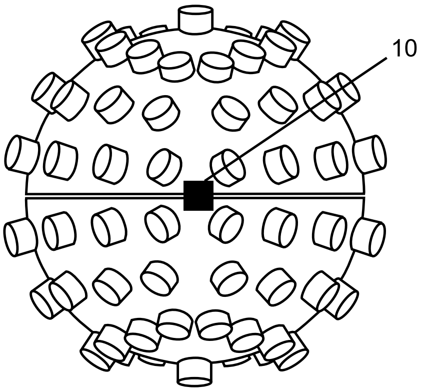

Fig. 4
Detectors inside the system

Another version of design is represented in Fig. 5, where the suggested method is realized by means of a spherical electric capacitor with three coats 11, 12, 13. Each capacitor coat is connected to the outlet of three-phase pulsed generator 14.

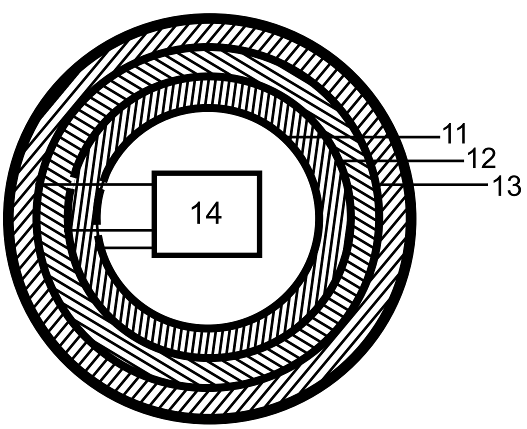

Fig. 5
Three-layered spherical capacitor

In this case the wave of energy density is created without electromagnetic emitters and this principle is not related to the modelling of magnetic quasi-monopole. The device operates due to the control unit, which provides the high-frequency changes of electric potential at each coat in such a way that the mode of high-frequency converging or diverging wave of energy density is created. In fact it is a standard three-phase generator but it produces not the rotation of the rotor of some electromotor but **"compression" or "decompression" of aether.** Aether is "pumped" in the center of the device or "pumped out" of the center.

In this case there is no need to tune separate sources of waves to make the device work in in-phased mode. It ensures reliability of the device operation if to be compared to the quasi-monopole. Besides, much less energy is necessary for the processes of charging and discharging of multi-layer spherical electric capacitor than for creation of magnetic field by means of conductivity currents.

Since the efficiency of such systems directly depends on the frequency and value of changes of energy density in space, we suggest the use of plasma shells instead of electromagnetic work surfaces for the next version of the device. It will allow significantly improvment in the specific parameters of the device. For that it is enough to place the electrodes of the multi-layers capacitor in a low-pressure gas area and these electrodes should be made as gauze electrodes. At that the wave is created in plasma, which is excited layerwise by several gauze electrodes placed in a spherical space between the inner and outer spherical bodies of the device. Therefore this version of design of the device can be considered as **the manipulation of the plasma method**.

Fig. 6 is a plan of one more design version. It is made as an inner frame 15 and outer frame 16, the space between them is filled with some gas 17. Three electrodes 18, 19 and 20 are connected to the three-phase pulsed generator 21. **Consequent excitation of plasma layers by electrodes 18, 19 and 20 creates the wave of energy density.** Propagation of this wave can be directed both to center of the device and from the center of the device.

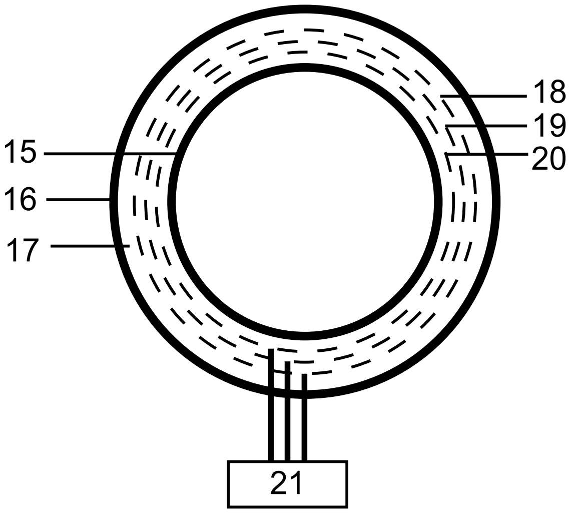

Fig. 6
Wave in plasma

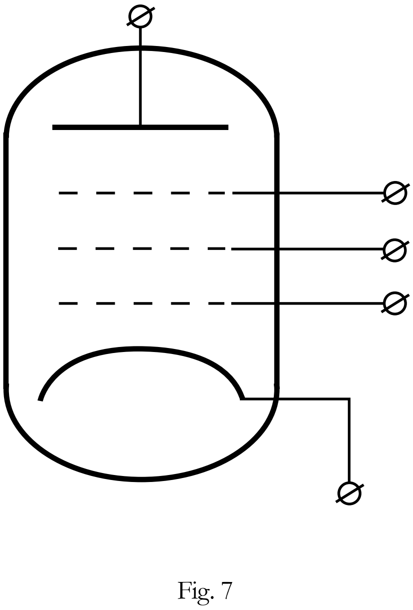

Fig. 7
Schematic electric circuit

In Fig. 7 there is a schematic electric diagram of the design.

To conclude. The work on time control has just started. We have clearly determined the main physical principles of operation of such devices, which can change the aether density in some given volume of space and thus influence temporal parameters of any physical process. The small experimental results today allow us to make real positive conclusion on availability of this method and on the possibility of its practical application in applied aspects. The first aspect is antigravitation propulsion technology and we are developing the methods to detect mass (weight) changes in the time control experiments to prove this applied possibility. Another area is medical applications of the changes in the aether density.

Russian Federation patent claim #2003110067 was filled April 9, 2003. At present time we are interested in marketing for this technology as well as in search of additional investment and partners.

References

1. N.A. Kozyrev, Selected works, Leningrad State University, 1991
2. I. Prigozhin, Introduction in thermodynamics of irreversible processes. M., 1964
3. A.D. Sakharov, Quantum fluctuations of vacuum in curved space and theory of gravitation, Reports of USSR Academy of Sciences, vol. 12, 1968, p. 1040
4. "Can the vacuum be engineered for space flight applications & Overview of theory and experiments", Dr. H.E. Puthoff, Infinity Energy, July-November, 1997
5. K.P. Butusov, Time as physical substance. Collection of articles "Problems of space and time in modern natural science", Issue 14, L. 1990, pp. 301-311
6. Yu.G. Belostotsky "What is The Time?" Saint-Petersburg, 1991
7. V.A. Atsukovsky "General etherodynamics", M., 1990
8. A.V. Frolov, Physical Principles of The Time Machine, New Energy Technologies #3(6), pp. 8-10, Saint-Petersburg, 2002
9. M. Faraday, Experimental researches on electricity, vol. 3, M. 1959
10. K.P. Butusov, Symmetrization of Maxwell-Lorenz equation. Problems of Space and Time in modern natural sciences, vol. 15, L. 1991
11. S.M. Polyakov, O.M. Polyakov "Introduction in experimental gravitonics", pp. 58-59, M. 1991
12. V.A. Chernobrov "Mysteries of Time", M. 1999

Aetherodynamics as a New Field of Physics

Theory and Experiments

Vladimir A. Atsukovsky, Russia
Email: atsuk@dart.ru

A crisis exists these days both in Natural science on the whole, and, in its basis, i.e. physics. This situation is caused by fundamental errors in the methodology of its theory. The postulate method and predominance of mathematics above physical sense can be also considered as the reasons for this. Crisis situation has also taken place in the past and the way out was always to use a deeper level of matter organization than the accepted one.

A large amount of material on elementary particles of matter is collected in physics. It was discovered that all these particles can be transformed in each other. Besides that, there was discovered the ability of vacuum to create such particles at strong electromagnetic fields. In this connection it becomes clear that all these particles and vacuum have a common building material, i.e. a physical medium which fills all the world space. This medium was renamed **aether** and the element of the medium was renamed **amer**.

The worked out methodology of aetherodynamics allowed determining that aether is a usual viscous coercible gas for which all laws of usual dynamics of gases are valid. It has allowed to determine its characteristics for near-Earth space and then to describe the structures of primary stable elementary particles (i.e. proton, neutron, electron, photon, atomic nuclei, atoms and some molecules) as well as physical essence of general fundamental interactions (i.e. strong and weak nuclear interactions, electromagnetic and gravitational interactions) and some physical phenomena.

Also developed was the model of aether circulation in the Universe within the limits of perpetually existent matter, usual Euclidean space, and evenly flowing time. Formation and decay of matter, formation and decay of celestial bodies and galaxies as well as functional classification of galaxies are also included there.

To confirm some theses the author and his research team conducted several experiments mainly on electromagnetism that gave positive results. The research was conducted on testing of the presence of aether wind. These research works have confirmed the results obtained by D. Miller in 1925.

Report by Vladimir A. Atsukovsky

Electromagnetic Gravitational Interaction

Vadim Y. Kosyev, Russia

http://uft.h1.ru; http://uic.nnov.ru/~kovy2
Email: uft@inbox.ru

In the article some mechanisms of electromagnetic gravitational interaction will be considered in the view of the Unified Theory of Field, Space and Time. See the complete description of The Unified Theory of Field, Space and Time at http://uic.nnov.ru/~kovy2; http://uft.h1.ru or in the book [1] which has the same title.

First of all, let us consider the structure of Space-Time where the radiation propagates. In our World gravitation is the unique substance which exists everywhere and gets over any obstacles. There are no known methods of insulation of gravitational field in modern science. It is impossible to imagine space and gravitation separately. Gravitation exists everywhere where there is some space. The gravitational field created by all masses of our metagalaxy is the aether in which cosmic objects moves and electromagnetic oscillations are propagated. The space surrounds us since the whole matter carries gravitational charge of only one sign. From astronomical research of cosmic space it follows that radius of metagalaxy is equal to the critical (gravitational radius) and hence gravitational potential is equal to $c^2 \sim 10^{17}$ $[m^2/c^2]$ in each point of space.

In our space the sum potential of the electric field is equal to zero. All bodies and space-time as a whole are electrically neutral. This assertion follows from two unique properties of gravitational space-time:

1. Values of electric charges of different signs of elementary particles are exactly equal.
2. There is an equal number of elementary particles which carry electric charges of different signs.

If electric potential is equal to zero in the given point of space then energy of alternating magnetic field passes entirely to the electric field and radiation gains the parameters of electromagnetic oscillations. However electric potential of positive and negative signs can prove its properties at the simultaneous charging of different local spatial regions. In the electric field the character of radiation differs from that one which is observed in ideal gravitational space-time (without electric potential).

In the presence of electric potential some part of magnetic energy is spent to create gravitational alternating field. Radiant energy is absorbed. Electrically charged spatial regions are perceived by us as filled with a strong absorbing substance. At the same time if the potential of the electric field can be compared to the potential of the gravitational field then spreading of radiation becomes impossible.

Change of both electric and gravitational field results in the creation of a magnetic field in the region of space-time which has a dual electrogravitational nature. Change in the magnetic field results in the creation of both electric and gravitational fields. The amplitude of electromagnetic and magnetogravitational constituents of the unified electromagnetic gravitational oscillations depends on field potential of opposite nature. The electromagnetic constituent is determined by gravitational potential and the magnetogravitational one is determined by electric potential. Transference of gravitational masses of matter in electrogravitational field-aether causes the creation of the proper magnetic field. Coming from the direction of magnetic field some force influences upon the moving electrically neutral masses. This force is similar to Laurence force.

Electromagnetic gravitational converter designed according to the Searl Effect (Fig. 1) visually demonstrates the mechanisms of electromagnetic gravitational interaction. A detailed description of the experiment is presented in the article [2] by Vladimir Roshschin and Sergey Godin (see http://www.n-t.org/tp/ts/dms.htm).

The device consists of a cylindrical stator of about 1 meter in diameter which is surrounded by 24 cylindrical rollers. Stator and rollers are made of magnetic material and they are magnetically linked (there is no contact between them). Vector of the magnetic field of the stator and rollers is vertically oriented along the axis of the cylinder but it has opposite direction. Rollers are mounted on the movable separator which circles round the stator. Each roller rotates on its axis in the same direction. The weight of the device is 350 kg. During rest, the device is electrically neutral and has proper gravitational charge, inertial mass, magnetostatic field of the stator and rollers.

On rotating of the magnetic system the following effects were observed:

- Depending on the frequency (up to 35%) the weight of the device was decreased.
- At a frequency of 550 rpm there is a spontaneous acceleration of the system and turns of the rotor abruptly increase (in quadratic dependence). To stabilize the mode it had to take off the excess energy to active load (up to 7 kWtt) by means of electromagnetic transducers.
- Magnetic field was distributed in a room as concentric surfaces of 5-8 cm thickness and 50-60 cm period. Depending on speed of rotation of the magnetic system there was a fixed temperature decrease from 0 to 8° within the regions of space which were curved by magnetic field.
- The glowing (corona discharge) appeared around the working converter.

Appearance of all mentioned effects is caused by the simultaneous presence of electric, magnetic and gravitational fields in the area of the converter.

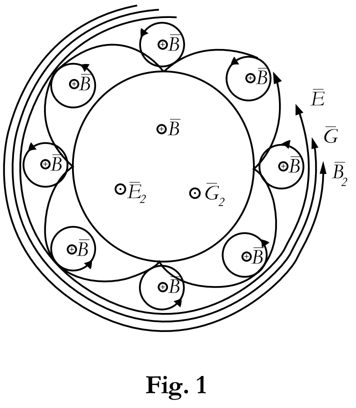

Fig. 1
Electromagnetic gravitational converter

Because of the Searl Effect a very complicated configuration of electric, magnetic and gravitational fields appears in the device. Rollers having their own field $\boldsymbol{B}$ rotate around the stator which has its field $\boldsymbol{B}$. Cycloidal motion of numerous magnets which are built in the rollers causes change in the magnetic flux. EMF $\boldsymbol{E}$ appears in the area of rotation of the rollers. EMF is directed along the perimeter around the device. The electric field creates stationary waves which are registered around the device as concentric electromagnetic gravitational walls. Magnetic rollers should be made non-conducting in such a way as to allow electric field penetrating inside. In the presence of the electric potential change of the magnetic flux causes appearance of gravitational field $\boldsymbol{G}$. This field is directed along the perimeter around the device. In the presence of electric potential the Lorenz force influences the gravitational masses of the rollers. These gravitational masses move in the external magnetic field of the stator. The Lorenz force is directed transversely to the velocity along the cycloid and in the direction of the field $\boldsymbol{B}$. The value of the Lorenz force depends on the electrical potential, magnetic intensity, mass of the rollers and their traverse speed. The electric potential in its turn depends on speed of rotation of rotor of the magnetic system. As a result the Lorenz force depends on speed according to square law. Rotating gravitational masses of the rollers generate magnetic field $\boldsymbol{B}_2$. Field $\boldsymbol{B}_2$ generates derived fields $\boldsymbol{E}_2$ and $\boldsymbol{G}_2$ etc.

Spontaneous acceleration of the system does not depend on the direction of rotation. Directions of fields and forces change at change of the direction of rotation that corresponds to acceleration in the opposite direction. When the direction of rotation is changed the weight of the device changes. When rotation is in one direction the field $\boldsymbol{G}_2$ is directed upwards (decrease of weight), when rotation is in another direction, it is directed downwards (increase of weight). In the presence of electrical potential many other interesting but still unstudied phenomena are observed (for example, curve of space, change of speed of time-flow, changes of ambient temperature…). Magnetogravitational converter transforms gravitational energy of space-time into energy of electric field which in its turn is transformed into mechanical energy of the rotating magnetic system and electric energy of the load. Energy resources of the future are not coal and gas reserves or nuclear power. They are in the inexhaustible energy of gravitational aether of our space-time.

Reference

1. V.Ya. Kosyev. Unified Theory of Field, Space and Time. Nizniy Novgorod, "Arabesk", 2000.
2. New Energy Technologies #1, 2003, p.35.

Medium For Existing of Matter in Nature

Anatoly V. Rykov, Russia
Email: rykov@uipe-ras.scgis.ru

By medium for existing of matter in Nature we understand physical vacuum which is a category determining all spheres of matter "life" (from particles of microworld to gravitational interactions in the Universe). Matter can not exist without this medium. Thus a question emerges, viz: how physical science can be developed without taking this fundamental aspect into consideration?

According to the Einstein postulate, in a void light always propagates with limited speed, which does not depend on the motion mode of a radiating body. It is a statement of deep physical sense which lies in the fact that only some physical medium can possess such a property. Indeed, if velocity of light, which was radiated by a body, further does not depend on speed of this body then it is possible only in a certain medium. For example, in air sound propagates with a certain speed which does not depend on the speed of the source and is determined only by density and volume elasticity of air. Thus, according to the Einstein postulate, there is a physical medium instead of void and velocity of light depends on the parameters of this medium (as is well known, velocity of light is equal to the square root from the product of inverse values of electrical and magnetic conductivity of vacuum).

The author of this article made an attempt to research this physical medium [1]. The author used a well-known experimental fact that at interaction of a photon with a charged real particle this photon transforms to electron-positron pair. Besides, it was taken into consideration that photon is an electromagnetic phenomenon. The research allows the following conclusion: the medium named in physics as vacuum (before the XX century it was called aether) is created by electric dipoles from (+) and (-) elementary charges. These dipoles are surrounded by magnetic (mass) continuum. There were determined dipole intervals, elasticity of decompression and ultimate strength of this interval. What does this model of medium mean?

1. This model is a physical validating for the "birth" of electron-positron pair at energy which is necessary to break dipole and create the "clot" of mass of these two particles. A certain model for the "birth" of mass must exist. It is not clear as yet but can be represented as a vortex of magnetic continuum inside the charge shell of electron and positron.

2. Let us suggest that values of (+) and (-) charges differ at 7.848981×10^{-41} Coulomb, then this difference is sufficient for the medium to be a source of gravitation and inertia. According to Newton's law of gravity, a weak electric charge of medium (all material bodies exist in this medium) causes all bodies to attract one another (under Coulomb's law). On the other hand, weak medium charge of a like sign creates repulsive forces which manifest themselves in the form of the expanding Universe. Thus it becomes clear the amazing uniformity of gravity and forces of negative pressure for the Universe. The mentioned difference of values of (+) and (-) charges of medium dipoles is not used as arbitrary parameter but it logically follows from its electrical structure [2].

3. Hypothesis of nature of gravitation is confirmed by calculations of the deflection angle of electric waves by the Sun. Calculated angle differs from the experimental value only in the fifth sign, which depends on the accuracy of several physical values used in calculation formula. There is practically an absolute coincidence with Einstein's theory. The difference is that it is a concept of space and time (i.e. geometry) which prevails in the general theory of relativity while in the nature of gravitation physical basing is used. Velocity of light is unstable and determined by medium state which depends on electric, magnetic and gravitational potentials.

4. The existence of "black holes" relates to medium structure and nature of gravitation. At the edge of "black holes" the ultimate acceleration from gravity is realized. It causes breakdown of connections in electric dipoles of medium, creation of matter and antimatter (the so called "evaporation" of black holes which was theoretically predicted by E. Hoking, England). However at the border of a black hole the velocity of light is equal to zero since its propagation

medium disappears. According to Einstein's theory it is such a deceleration of time that electromagnetic oscillations gain zero frequency.

5. Naturally the medium is a carrier of all types of electromagnetic radiation, beginning at stationary electric voltage and ending with super-high-frequency "photons", which can be determined not as electromagnetic but as magnetoelectric phenomenon. In the latter case magnetic continuum of medium has a crucial importance since it determines very small degrees of magnetoelectric disturbance. These degrees are thousands of times less than those of a hydrogen atom. It is reasonable that such small degrees creates illusion that a photon possesses properties of particles.

6. Medium structure directly leads to the notions of Quantum Mechanics, beginning at quantification of electronic "orbits" in atoms. This medium determines the "allowed" spots for electrons to be placed around a nucleus. Thus medium is a necessary place for the existence of all matter or matter of the Universe. The so called Compton length of electron wave is one of the validations of this statement. It is directly calculated to a high accuracy according to electric structure of vacuum.

7. Thus Plank's constant is not a mysterious "quantum of action", on the contrary it is entirely determined by medium parameters. Thereby the crisis (which conventional physics is accused of) can be logically overcame by the introduction of medium which is capable to take radiation in electromagnetic region only by quanta. Plank's constant is always presented at all quantum approaches. It is additional evidence in favor of the necessity to take into account the medium as natural place for existence of all the matter in Nature.

8. It is still a question what processes take place in the centers of galaxies. The observations show that the centers of galaxies create star matter. They often flow out the centers and settle themselves at approximately the same plane. It is an evidence of the fact that centers of galaxies quickly rotate and the favorable conditions for stars to reject matter are created. Similarly planetary systems are created around the rotating stars. It is thought that centers of galaxies are gigantic black holes. In the context of the concept of medium for matter existence it can be set up a hypothesis that being at some special state the medium creates stellar systems, i.e. galaxies.

9. Astrophysicists more and more trend to accept the existence of unknown "dark" matter, which occupies approximately 70% of all matter. It is said that due to antigravitation property this dark matter is responsible for the expansion of the Universe. Hypothesis on the nature of gravitation gives the affirmative replies to such suggestions as: expansion of the Universe is caused by weak electric charge of the medium and distributed mass of magnetic continuum is the very dark matter.

10. All elementary particles (electrons, positrons, mesons, protons, neutrons etc.) are in the medium and interact with it. In the case of electrons and positrons this interaction causes the great extension of dipoles of medium which directly adjoins to borders of particles. This extension helps photons to break the particles. In case of protons, there is such a breakdown of the medium at their borders that protons turn to be dressed in "coats" of virtual electrons and positrons. And medium is shown to be in entire state only when reaching the first Bohr orbit. This interaction of the medium and particles leads to the factors which determine the life time of some of them. Thus life time of a neutron is up to 30 minutes depending on the state which it has while leaving the nucleus. For mesons "the tearing" Coulomb forces, which exist between the medium and particles, are very powerful that makes the life time of mesons very short. However if the particles moves with a high speed relatively to the medium, then these forces noticeably decrease and if the speeds of motion is close to velocity of light then these forces become very small. The life time of such particles noticeably increases. In Einstein's theory it is said about "deceleration" of time. Finally the life time of particles is determined by the inner steadiness and outer influence of the medium. Protons have fantastic binding energy and they cause such great external destructive effect that their life time becomes immense.

11. When particles move in medium with high speeds, there is a concentration of magnetic continuum which is recognized as increase of mass of particles at increase of their speed. Evidently, we can continue the list of physical phenomena which relate to the presence of medium for existing of matter in Nature. However it seems to be enough to admit that medium (physical vacuum, aether) is of a great importance for the very existence of the Universe and for the processes which take place there. "Mechanism" of gravitation and inertia has a special importance since

it is the only thing which can make clear the real mechanism of Nature. It is possible that the methodology, which is accepted in physics, is necessary for more accurate description of phenomena but it is not enough for understanding of the nature of “space” and matter. Research of the medium for existing of matter will be able to meet the conditions sufficiency.

The understanding of vacuum structure, which is created by electric lattice from charge dipoles, surrounded with magnetic continuum, gives an opportunity to control the forces of gravitation and inertia. The medium of vacuum can be influenced by:

1. The radiation which has frequencies coming to the frequency from point #11 of the list of parameters.
2. Electric voltages which exist in vacuum (it is not very promising because of real matter breakdown).
3. Magnetic intensities (magnetic flux density). It is the most promising influence method (1-10 Tesla is enough to compensate terrestrial gravity).
4. Transformations of low accelerations to high accelerations of impact type.

Potential electric and magnetic energy is immense in vacuum medium. Reasoning from energy of one dipole, which is equal to 1.6x10 (-13) Joule, we will derive the value of electric energy equal to about 10 (+31) Joule per one cubic meter of the medium that is equivalent to mass annihilation of 10 (+15) kg!

References

1. Rykov A.V. Foundations of aether theory// RAS, M., 2000, p. 54
2. Rykov A.V. Hypothesis on the nature of gravitation// MSU, M., 2001, pp. 59-63

◆

Spartak M. Polyakov

(09 January, 1931 – 04 June, 2003)

Spartak M. Polyakov, a legendary physicist-experimentalist, known for his research in the field of gravitation, has died. Polyakov was the author of more than 50 research works, among them “Introduction into experimental gravitonics». Polyakov was the creator of the new model of photon. He also conducted numerous research in the field of electrodynamics, gravitational phenomena and superlight connection.

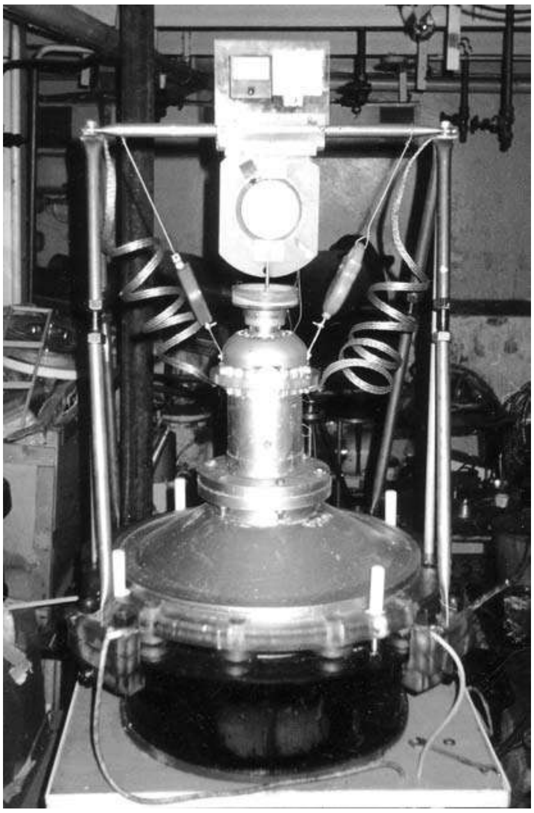

Spartak M. Polyakov graduated from Kishinev State University with degree in “General Physics”. For many years he worked at the top secret military engineering department “Istok” (Russia).

Plyakov’s alternative physical theory allows explaining such phenomena as magnetostriction and optical magnetization. The basic postulate of this theory states that the speed of propagation of gravitational waves is dozens of times faster than velocity of light regarded by the traditional science as the limit. In his experiments Polyakov demonstrated such methods of generation of gravitational waves as precession of gyroscope and remagnetization of ferromagnetic.

While working for “Istok”, Polyakov designed on his own a device which is possibly the first gravitational motor in history. Total specific thrust produced by rotation of gravitating mass comes to 2.5 kg/kWt. This index is close to that produced by the engines of modern helicopters, which is equal to 8 kg/kWt. Practical application and further development of Polyakov’s ideas can provide humanity with new kinds of communication, gravitational engines and free energy devices.

Adams Motor

Principle of Operation and Experimental Data

Sergey S. Abramov, Russia

Email: a_serge@cards.lanck.net

Adams' motor-generator represents the type of devices which use, as their creators claim, so-called "free energy". The term "Zero point energy" is also used in some sources [1, 2]. Due to quanta-mechanical fluctuations, this energy exists even at zero temperature. Adams' motor-generator belongs to the group of Switch reluctance motors [3]. Robert Adams, **former Chairman of the Institute of Electrical Electronics Engineers, USA (New Zealand section)**, designed this machine in the late sixties of 20th century. Similarly to all the devices utilizing free energy (referred to in come sources as "overunity devices"), Adams' device remains practically unknown to the general public. **The device is rather simple to assembled even at home, which I experimentally proved. However, it is necessary to be very careful while choosing the model parameters. As to the latter I have managed to collect quite a big number of instructions from existing sources; the summary of these instructions is presented below in this article. Based on these instructions, a low-power model can be assembled even without a mathematical analysis and modeling of electromagnetic field. Such model would certainly facilitate optimization of the device.**

Adams' motor is most frequently a DC machine; however, it can also use an AC source through a rectifier. In the latter case only the adjustment of the device and its control system can be provided.

Editorial: We disagree with some assertions of the author; however this article is of great interest.

My experiments on my own model do not yet allow me to make an unambiguously positive conclusion concerning the possibility to generate excessive energy. Experiments with my new control system designed on the base of AVR controller AT90s2313-10PI (it is produced by Atmel company (http://www.atmel.com) will allow to be more specific. Below there is a general analysis of the motor principle of operation and a number of recommendations concerning the construction and technology. I do not propose to take this as compulsory rules to follow; other technical solutions are possible.

Basing on the principle described in this article, R. Adams (the link to his articles is available at: http://www.aethmogen.com/wri/intro.shtml) created a few DC motor-generators which operate on permanent magnets. Some of them, according to information found on the Internet, have manifested 690% electrical efficiency and 620% mechanical efficiency. These devices operate at room temperature without overheating. My device has shown between 1 and 3 degrees overheating after an hour of functioning. However, it is easy to prove that such overheating is predictable for an average current of 0.15 A in coils of 35 mm long and 25 mm in cross sectional diameter. I have not been able to prove the data published on the Internet concerning the Adams motor capability to operate when the stator temperature is a few degrees lower than that of the environment. The temperature of the coil and of the power transistor is a good indicator of correctness of the circuit set-up and of functioning of the control circuit. There were cases when transistor and coil were noticeably heated after adjustment. Usually this was explained by a bad choice of points of transistor switching or by too extensive current impulses in the stator (which must amount to approximately 25% of period length). After the required adjustment the motor continued to operate almost without overheating.

Adams' motor was first mentioned in Australian Nexus Magazine in 1992. Later, Harold Aspden (Britain) proposed a slightly improved version of the motor and received Great Britain patent No. 282708 [4], which strongly reminds of the original version published by the above magazine. **Adams' device** represents an electrical motor and/or generator consisting of a rotor with radially directed permanent magnets and of a stator also constructed with a few radially directed and periodically magnetized cores with winding. (Fig. 1). In some models axial orientation of magnets and coils are also used.

Permanently magnetized poles of rotor can include any number of poles, even an odd number. Analogous poles of the magnets (all N poles or all S poles) are directed outside. A version with alternating poles is also possible; such model allows the torque to increase. In this case, after passing a rotor pole the stator is demagnetized by the current impulse and begins to be attracted by the magnet of different polarity. This circuit requires a more complicated control; on the other hand, it manifested rather good results in certain models.

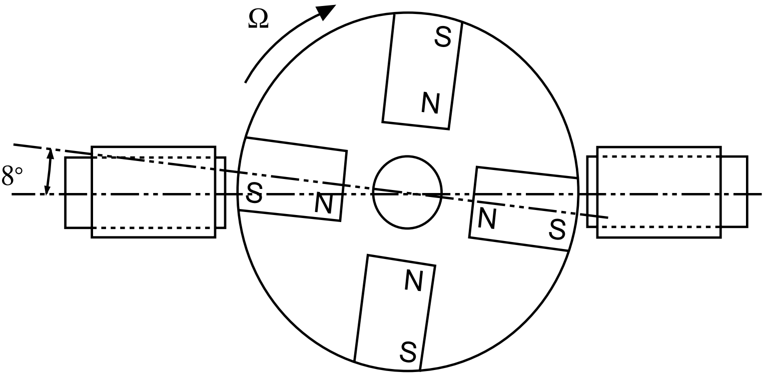

Fig. 1

Poles with winding placed on the rotor are radially oriented in order to obtain a supply of energy emerging as a result of the influence of counter emf from the rotor poles. Steel or iron cores are used for the poles of the stator with winding. It is also possible to use other materials, at that the core must have high magnetic inductivity and low level of magnetization reversal losses. The stator winding consists of a few hundred turns. The current inducted by the magnet in this winding will have the polarity which will cause repulsion of the magnet. Since the electromagnetic state of stator changes quite significantly and quite fast, then the stator core can be considered to be the most crucial element of the device. Ignoring this fact was one of the most frequent mistakes made by those who tried to reproduce Adams motor.

The current inducted in the stator is the function of:

- field size,
- number of winding turns,
- speed of flux changes.

Resultant parameters of this device cause each pole to be attracted or repulsed by stator poles when the rotor is in certain position in case when the rotor is unbalanced. To achieve this effect it is required to switch the input current in control coils after the signal from sensor of rotor position. R. Adams used a mechanical switch as a sensor. My device and a number of other devices use the signal from two Hall sensors. However, according to experimenters' information, better results are achieved if a position optical sensor is used.

Time of switching of impulses is determined by the size of the motor itself, i.e. the speed of motor rotation, location of rotor magnets towards the stator windings and the distance the rotor magnets pass while moving by the poles with stator winding.

It is necessary to take into account that any part of this motor can be modeled based on the existing electromagnetic theory and no part of the motor is in conflict with any laws of electromagnetism. There are so many ways to construct Adams' motor that any version may be considered to be correct.

One can say that the frequently pulsating electromagnetic process in the stator core is what allows Adams' device to function as a kind of diode which borrows energy from the field of permanent magnet but then does not return that energy in full.

5 stages can be defined in the periodical process which takes place in stator:

1. The magnet is attracted to the stator core. The permanent magnet is attracted to the iron core of the stator with winding. While doing so no consumption of electrical current takes place. It is as if kinetic energy is borrowed from an internal ferrite magnetic source and is supposed to be returned into the stator.
2. Stator core is magnetized. During the period when the magnet is positioned in front of stator core they both comprise a single magnetic conductor with an air gap and the stator core becomes an extension of the magnet side it faces. It is usually supposed that the energy "borrowed" on the first stage is getting back now.
3. Stator core is demagnetized. When the stator core becomes rotor magnet extension, the circuit closes and current impulse gets to stator windings. I have seen the instructions saying that the angle between

stator axles and rotor magnet must amount to 7-8 degrees as shown in Fig. 1. However, my model made it clear that at increase of speed of rotation it is necessary to start the coil a little earlier, when the magnet axle has not reached the stator axle. Probably this applies to the circuit with a Hall sensor only and if an optical switching is used the making angle will be different.

Magnetic field of this current acts to compensate magnetization of the stator, which is caused by the field of rotor magnet. Consequently, the summed current significantly compensates attractive force between rotor and stator and the rotor can freely rotate by using the inertia obtained at stage 1. This process is characterized by the fact that this current impulse is amplified by the current inducted in the stator winding by rotor magnet which, in accordance with Lentz law (1834), counteracts the power which induced it. Consequently, kinetic energy obtained as a result of attraction of the rotor to the stator at stage 1 is transformed into electrical demagnetizing impulse in stator winding during the period when the rotor and stator directions are congruent. This is the unique overunity characteristic of this model. However it is obvious that instead of returning this energy the motor transforms it into electromagnetic demagnetizing field.

4. Restoration: when the rotor is removed from the stator attraction zone the latter looses energy and returns to its initial demagnetized state. Decreasing electromagnetic field creates a current wave of reverse polarity which can be stored in the capacitor.
5. Reiteration of the process: This periodical process is renewed as described in stage 1 during the next magnetization of the stator, excluding the fact that emf preliminarily stored in the capacitor, on term of presence of suitable electrical circuit, can be used for facilitating stator demagnetization or even used to supply the load.

It may be briefly summarized that the frequently pulsating electromagnetic process in the stator core is what allows Adams device to function as a kind of diode which borrows energy from the field of the permanent magnet but does not return that energy in full. The important characteristic of such motors is that the stator windings are used for demagnetizing and not for magnetizing as it could seem from the first sight.

It is noteworthy that there is a small pause between attraction to the stator and repulsion from the stator. The effect of attraction to the core takes place a split second before the repulsion effect manifests clearly. This pause being the reason of electromagnetic asymmetry creates conditions necessary for achieving overunity effect. If the attraction to the core and repulsion by means of Lentz currents were taking place simultaneously and with the same power there would not be any overunity characteristics. That is why the rotor must be as lightweight as possible. From this point of view, T. Harwood's model is the most lightweight of the known models. In Harwood's device the magnets are mounted between two CD disks fixed on the shaft by means of plastic washers and glue. My model is heavier, which can be considered as one of its disadvantages.

The principle of operation of Adams' motor is based on the balance, which creates the electromagnetic asymmetry. To get the motor to operate the magnet must be attracted to the stator core which must have a smaller cross-section area in order to create attraction without any significant repulsion effect from the stator windings mentioned above. When stator and rotor axles are congruent the Lentz induced current must be sufficient for compensating the natural attraction of the magnet to the stator core. Consequently, stator windings must have enough turns for demagnetizing effect, but not to the extent that this effect fully manifests before the rotor reaches the stator axle when Lentz current has its maximal value.

During my experiments at 12 V voltage and on using two independently controlled stator coils the speed of rotation reached 3400 rpm. Please note that while reproducing such device it is necessary to take certain measures in order to ensure safety in case of possible breakdown. **The magnet disconnected from the rotor may be dangerous!**

Technological recommendations are as follows:

1. The device must be low power. It is better not to try to begin with a motor functioning in kilowatt range. This is feasible only on condition of having all necessary technological documentation which is not available at the moment.
2. The preferable voltage for the first model is 12 V. If the voltage is less the speed of rotation is too slow for indicating the expected characteristics of the device.
3. The best magnets are ferrite ones with dimensions 4x4x5 (where 5 is the magnet length). The practice has shown that at 12 V voltage neodymium-iron-boric (NdFeB) magnets cause a jerky rotation of the rotor.
4. The side of the stator core facing the rotor must

Control circuit with hall sensors

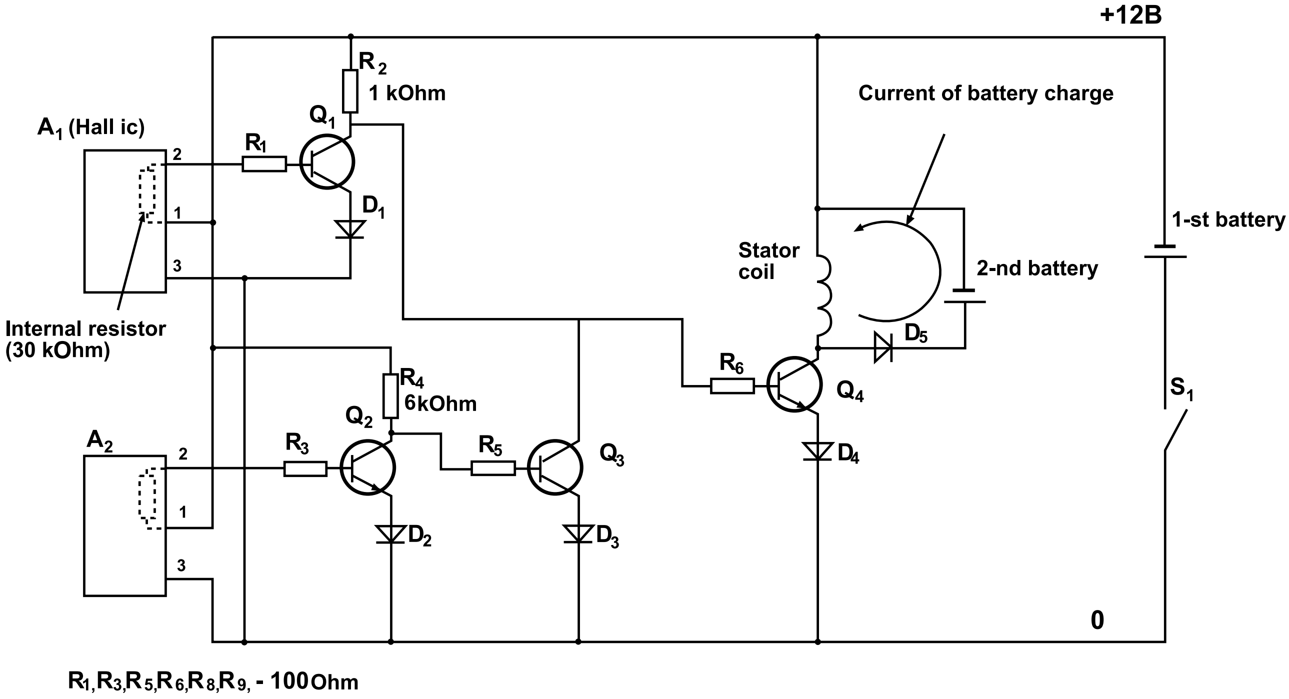

Fig. 2

have the size which is 4 times smaller than the corresponding side of the magnet. If cross-section area of the stator is larger, a bigger part of magnet field comes to the stator core when their axles coincide and thus there is nothing to induce Lentz' currents in the stator windings. It is a common point of view that it is necessary to try to minimize the volume of a device and to obtain the maximum efficiency at minimal material expense. In practice, it has always been the goal while designing electrical machines to decrease losses in stator I^2R. As for the Adams motor, it requires not only the use of disproportionate magnets but also stator windings with disproportionate number of turns specially designed to obtain maximum Lentz' currents that is achieved by hundred of turns of winding.

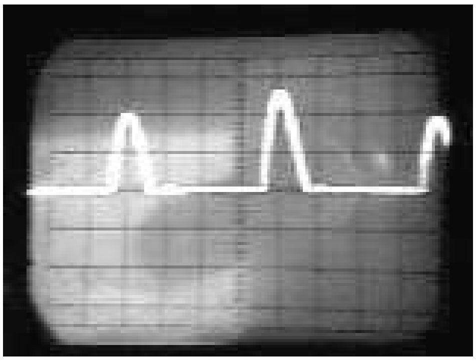

5. It is necessary to define approximately the effective zone of the magnet field. If practically applicable magnet field is equal to, for example, 8 cm, and the stator has a 10 cm long winding, then more than 20% of turns will not be efficiently crossed by magnetic force lines and will only create an excessive mass of the device. To define the degree of effective action of the magnet we can put a paperclip on the table and move it gradually in direction of the magnet until the paperclip is attracted to it. Actually if we take into account friction losses then the magnet influence zone will be a little larger. That is why the stator winding in the axle direction can be 10% larger than that in this experiment. The description of this test has been found on Tim Harwood's website (http://www.geocities.com/theadamsmotor/cdmotor.html).

6. The air gap clearance between the stator and the rotor must not exceed 1 1.5 mm.

7. Use as little metal in the device as possible. It is preferable that the metal is used in the core and stator windings only.

In order to increase the efficiency of this motor, it is necessary to build it in a manner, which allows removing the counter emf from the stator windings. To do that, this emf may be taken off and stored in the capacitor. The article published by Nexus magazine and Great Britain patent No. 282708 consider special generator windings; however, no sufficiently detailed data on operating rules are provided. Michael Smith (Australia, http://www.Fortunecity.com/greenfield/bp/16/content1.htm) has unequivocally informed me in one of his letters that

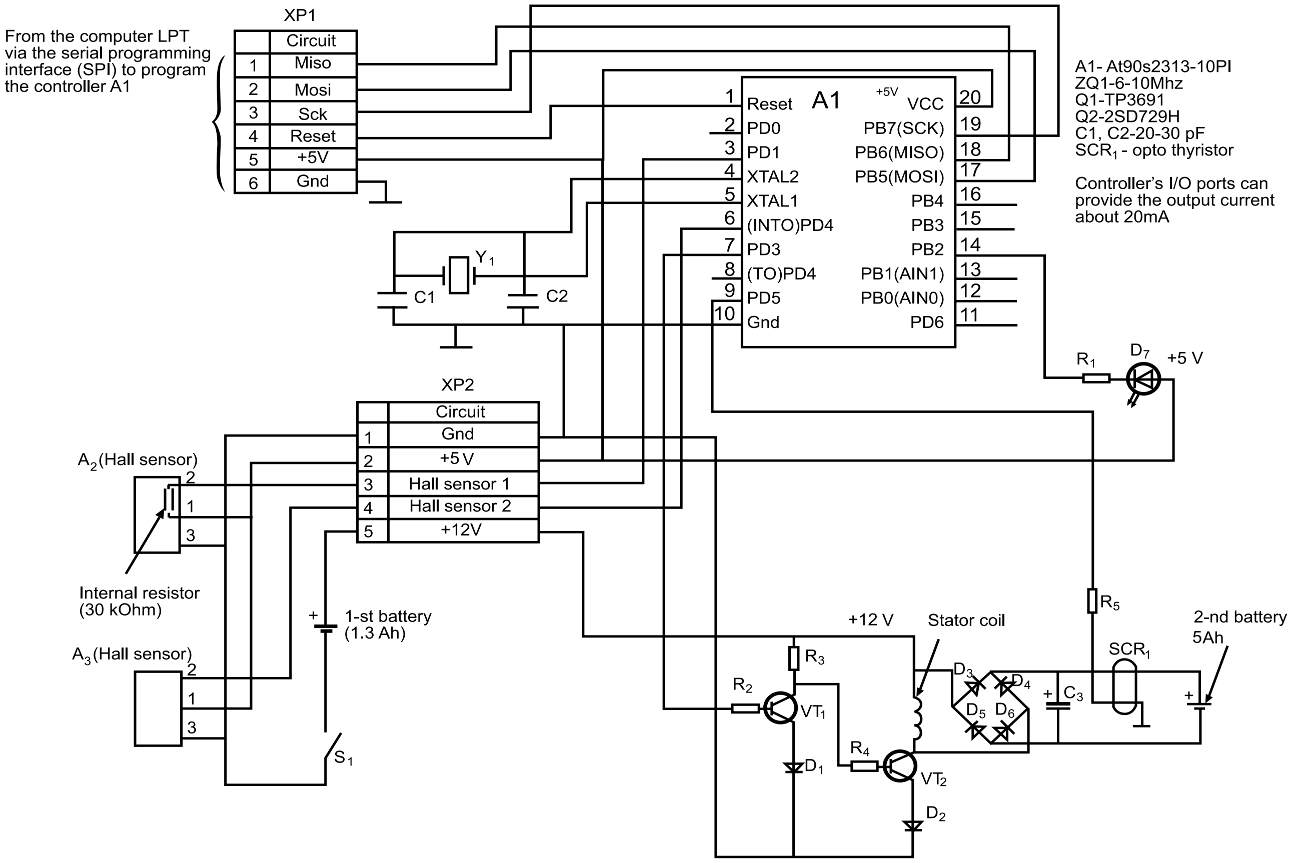
From the computer LPT via the serial programming interface (SPI) to program the controller A1
XP1
Circuit
1 Miso
2 Mosi
3 Sck
4 Reset
5 +5V
6 Gnd
A1
+5V
1 Reset
2 PD0
3 PD1
4 XTAL2
5 XTAL1
6 (INTO)PD4
7 PD3
8 (TO)PD4
9 PD5
10 Gnd
20 VCC
19 PB7(SCK)
18 PB6(MISO)
17 PB5(MOSI)
16 PB4
15 PB3
14 PB2
13 PB1(AIN1)
12 PB0(AIN0)
11 PD6
A1- At90s2313-10PI
ZQ1-6-10Mhz
Q1-TP3691
Q2-2SD729H
C1, C2-20-30 pF
SCR1 - opto thyristor
Controller's I/O ports can provide the output current about 20mA
Y1
C1
C2
XP2
Circuit
1 Gnd
2 +5 V
3 Hall sensor 1
4 Hall sensor 2
5 +12V
A2(Hall sensor)
2
1
3
Internal resistor (30 kOhm)
A3(Hall sensor)
2
1
3
1-st battery (1.3 Ah)
S1
R1
D7
+5 V
R5
+12 V
Stator coil
2-nd battery 5Ah
SCR1
R3
R2
VT1
D1
R4
VT2
D2
D3
D4
D5
D6
C3

Fig. 3

he has not achieved generation of excessive energy in such a system. During his experiments a two-battery circuit proved to be more effective. In this circuit, during a part of periodical process the energy is stored in the capacitor and then a command is transferred to a thyristor, which discharges it into the second battery. At that capacity, the second battery must be no less than 4 times more than the capacity of the first one. Otherwise the excessive energy does not have enough time to be stored during the discharge. I have also managed to provide the charge of the second battery, but the control circuit has not been assembled fully yet, as it is shown below in Fig. 4. Better results are expected when its final adjustments together with AT90s2313 controller program have been made.

In order to increase the efficiency of this motor, it is necessary to build it in a manner, which allows removing the counter emf from the stator windings.

I managed to start my first model of the motor in April, 2002. Afterwards, I spent approximately six months increasing its rotational speed from 750 to 3200-3400 rpm, decreasing vibrations and improving control circuit. In order to achieve a better rotor balance I had to re-construct it twice. It is very important to align the axles of the bearings with maximum precision; otherwise a considerable decelerating torque will manifest. The shaft penetrates the lower moving base. It is possible to rotate the base at a small angle for the precise alignment of axles and then to fix it with screws and nuts.

In order to decrease aerodynamic losses, two veneer parts are mounted between the magnets. Thus I managed to increase the speed, although the rotor weight has also been increased.

Stator cores are made of plates taken from a disassembled radio transformer. Tim Harwood used nails with winding, however, my own results with such core proved to be poor. Dimensions of the core are 10x11x50 mm.

The type of power source is also important. First, I connected a 9 Volt accumulator of "Krona" type in series with three metal-hydride 1.2 V accumulators. The speed of rotation did not exceed 1500 rpm. But when I used a lead/acid accumulator with 1.3 Ah capacity, the speed increased to 2600 rpm if there was one coil on the stator.

Four magnets mounted on the rotor have the dimensions of 20x20x35 mm and are fixed on a 105 mm glass fiber laminate disk. The current impulses proved to be too wide, by up to 40%. In order to make them shorter (down to 25-30%), I had to use the control circuit (Fig. 2) with two Hall sensors. At the signal from the first sensor the stator current is started, the second sensor switches it off. Many experimenters used a timer to control the impulse length while working with Adams' motor that is more practical since impulses are supposed to be wider during the start. I took this factor into account while designing the controller circuit. Current impulses are shown in Fig. 3. Their fronts are supposed to be shorter; it is probable that the coil has more inductivity than necessary. The impulses amplitudes are slightly different which is explained both by the difference in volume of induction of the magnets and by difficulty in achieving similar air gaps while working at home.

I have provided the charge mode of the 2-nd battery at my two-battery design. After 75 minutes of operation of the device the source lost 0.17 Volt whereas the second battery was charged at 0.36 Volt. The capacity of both batteries in this experiment was equal. Besides after such charging the second battery started to discharge quickly. The circuit where stator current charges the battery directly has to be considered ineffectual (see Fig. 1). To evaluate the charge, which is gained in non-hermetic accumulators, the density of electrolyte can be measured.

To obtain a more uniform torque I added the second stator which is controlled independently. This required installing two more Hall sensors and an additional power transistor. The angle between axles of the coils amounts to 135 (180-90.2=135) degrees. When the current is present in one part of the stator it is absent in another and vice versa. The speed has increased up to 3200-3400 rpm, and I deemed the further increasing of this characteristic unnecessary.

Increasing the number of circuit elements does not seem a good idea. Futhermore the adjustment process becomes more complicated. In order to improve the circuit of battery charge adding a timer circuit is required. Thus, I decided to use a controller circuit. A simple Basic program has been developed for AVR controllers. This program operates in mode similar to transistor circuit, but its capability can be significantly increased due to the built-in processor timers. Today program improvement is the most efficient way to solve the task of generation of excess energy in this circuit.

This article is meant to elucidate the principles of operation of one of the simplest devices which pretend to become an "overunity device", "free energy machine", "perpetual motion machine", whatever you call it. Probably someone will try to create such motor-generators on their own. I hope that my article will serve as a guide and will help to avoid the mistakes made by many experimenters (including myself) before they managed to build their own model.

References

1. Puthoff H.E. Source of vacuum electromagnetic zero-point energy. Phys. Rev. A 40.4857-4862; also 44. 3385-3386
2. Bernard Haich, Alfonso Rueda, Puthoff H.E. Physics of the zero-point field: implications for inertia, gravitation and mass. Speculation in Science and Technology, Vol.20, p.p. 99-114, 1997 (preprint version)
3. Miller, Switched Reluctancy Motors & Their Controls, ISBN: 10881855-02-03
4. Patent No. GB 2,282,708 (from NEN, Vol. 4, No. 8, December 1996, pp. 1-7).

INFINITE ENERGY
Cold Fusion and New Energy Technology
Including Conventional and Renewable Energy

Infinite Energy Magazine

* Cold Fusion * New Energy
* New Science
* New Technology
Subscriptions, 6 Issues Per Year
$29.95 North America
$49.95 Foreign
Single Sample Copy
$5.95 North America, $10.00 Foreign

Infinite Energy Magazine
P.O. Box 2816-FV
Concord, NH 03302-2816
Phone: 603-228-4516 Fax: 603-224-5975
http://www.infinite-energy.com

Prediction and Experiment on Field Transformation in the Model of Extended Space

Dmitry Yu. Tsipenyuk, Russia

119991, Vavilov str., 38, Moscow
E-mail: tsip@kapella.gpi.ru

A series of preliminary experiments on checking the possibility of generation of gravitational field at deceleration of charged massive particles in matter was carried out.

Introduction

In previous works [1-3] generalization of the special theory of relativity (STR) for the five-dimensional extended space with metric (+;-,-,-,-) was offered.

The model of extended space (ESM), combining electromagnetic and gravitational interactions, was made. For this, there was made (1+3)-dimensional space extension $M(T;\vec{X})$ of Minkovsky to (1+4)-dimensional space $G(T;\vec{X},S)$. Let us call it extended space. As the 5-th additional coordinate the value already existing in Minkovsky space, i.e, S interval

$$s^2 = (ct)^2 - x^2 - y^2 - z^2 \qquad (1),$$

is used.

Let us note that attempts to combine gravitation and electromagnetism have a substantial background.

Modern approaches to this problem trace back to the work of F. Klein [10] in which he proved that classical Hamiltonian mechanics can be represented as optics in the space of a great number of dimensions.

Then T. Kaluza tried to generalize Einstein's theory of gravitation to include electromagnetism in this theory as well [11]. He proposed to consider (1+4)-dimensional space with metric depending on potentials of the electromagnetic field. Kaluza's idea was evolved by O. Klein [12], G. Mandel [13] and V. Fock, and the model they had created got the name of the Kaluza-Klein theory. They proved that the trajectory of a charged particle has the form of a geodesic line with zero-length in 5-dimensional space.

In his works on 5-optics Y. Rummer [15] proposed to assign action dimensionality to the new dimension and to consider it periodical with the period equal to Plank's constant. Note that rest mass of particles in all these constructions unlike the model of extended space evolved in the works [1-7] was considered a fixed value. Subsequent development of multidimensional theories is given in the monograph [16].

A separate approach is represented by multidimensional constructions in the theory of strings and superstrings [17].

Approach to construction of (1+4) dimensional space evolved in [18] is close to the proposed model of extended space. Here it is proposed to use mass (matter) as the 5-th coordinate. However, in this model, as its originators admit, it is impossible, for example, to create the energy-pulse tensor. There is no such disadvantage in the extended space model [8].

Mechanics of a material point [1, 2, 7] and electrodynamics [1, 8] were made in the introduced extended space. Besides Lienar-Vihert potentials [6, 19] corresponding to such a model were also considered and properties of solutions of Maxwell augmented system of equations which are in conformity with these potentials were analyzed there.

Gravitational effects in extended space, such as the second orbital velocity, red shift and light deflection [4, 20, 21] were considered. It is proved that the formulas received in the general theory of relativity for calculation of values of these effects can be received by an absolutely different method within the framework of the extended space model as well.

It was proved [6, 19] that the fields in the model of extended space can change their signs. Such change of field-intensity sign and, consequently, change of sign of Lorenz force can be associated with radiation reaction of these fields which occurs when charged particles move with acceleration.

Thus, on the one hand, it was proved in the model of extended space that it is possible to get certain formulas describing gravitational effects of the general theory of relativity [4,20] using the technique of turns in extended space. On the other hand, it was proved that the electromagnetic field can be a source of gravitational field [6, 19].

Besides, a moving massively charged particle under deceleration can create a variable gravitational field around itself [9, 6, 19]. The following experiment was offered for experimental check of the latter assumption. In this experiment probable occurrence of gravitational field at deceleration of relativistic electrons was determined by change of oscillations of a massive torsion pendulum.

Experimental device

A narrow bunch of relativistic electrons from a microtron 1 (average power of the bunch is 450 Wt, electrons energy is 30 MeV) was directed to a brake target (position 2 or 3) made of wolfram exactly where deceleration of accelerated electrons took place.

A special torsion pendulum suspended on a vertical suspender 5 made of a springy metallic string with 1,8 mm diameter was placed near the brake target to register gravitational field which could probably appear at electrons deceleration. The length of the suspender made is 85 cm. The pendulum could rotate freely on the suspender only in horizontal plane.

The pendulum consisted of a light aluminum rod 4 (with a length of 120 cm) on the ends of which massive loads 6 and 7 made of non-magnetic material were fixed. The weight of each load was equal to 4kg. In the center a pendulum was fastened to a vertical suspender 5 by a special mounting preventing slippage during turns. To reduce the influence of magnetic inducings the pendulum was grounded and additionally screened by metallic grid from all sides. The period of free oscillations of the pendulum made were about 40 s.

Rigidity of the pendulum vertical suspender could be changed by means of limiting the length of effectively operating part of the suspender. As a result, the period of oscillations could be continuously changed within the limits of 40 to 27 s.

To reduce the influence of mechanical noise and to introduce additional attenuation in pendulum oscillations two liquid dampers 10 and 11 located near the pendulum massive loads were used.

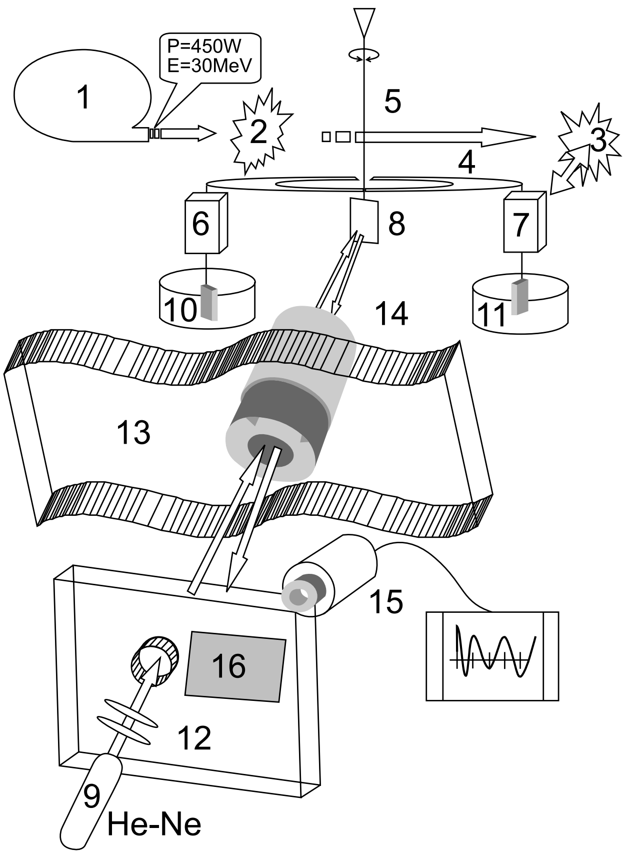

Fig. 1
Experimental device
1 – microtron, brake target made of wolfram – positions 2 or 3, 4 – rod, 5 – vertical suspender, 6 and 7 – massive loads, 8 – mirror, 9 – He-Ne laser, 10 and 11 – liquid dampers, 12 – optical system, 13 – concrete protection, 14 – observation channel, 15 – video system, 16 – screen.

Pendulum deflections were observed on a graduated screen by deflection of a laser beam reflected from a flat mirror 8. For this, the beam from a continuous He-Ne laser 9 through the optical system 12, which constricts the divergence angle of the laser beam, was directed to the mirror through a special narrow channel 14, located in concrete protection 13 around the microtron. By means of a video system 15 the beam reflected by the mirror was registered on the screen 16 located at a distance of 500 cm from the mirror. The video system allowed remote checking of vibrations of the laser spot and additionally enlarged the visual angle up to 12 times. The diameter of a focused laser beam on the screen was made 0.15 mm. The maximum turn angle of pendulum for the reflected beam to remain within the receiving channel was approximately 2 degrees. The accuracy of turn angle registration of the whole system was $5x10^{-4}$ degrees.

The pendulum was placed in such a way that one of the massive loads were close to the brake target at a distance of about 20 cm. There was also an opportunity to move the brake target from one end of the pendulum (position 2) to the other (position 3). This allowed changing the place of deceleration of electron bunch at constant parameters of all unaccounted mechanical noise and magnetic inducing. Thus, direction of pendulum torsion was changed under probable exposure of occurring gravitational radiation.

Below there is a photo of the experimental plant (see also the colored photo on the cover page).

Calibration Measurements

To make measurements it was necessary to select optimal parameters of the pendulum (masses of loads, suspender rigidity and the value of oscillations attenuation). On the one hand, while carrying out the measurement it is desirable that the amplitude of pendulum oscillations should be as maximal as possible. On the other hand, the beam reflected from the mirror should not go outside the limits of observation, restricted by the diameter of the narrow observation channel in radiation protection around the accelerator. Besides, the typical operating time of the loaded accelerator usually is 10-15 minutes. Necessity to accumulate the minimum of sufficient statistical data within this time limited oscillations period and the time of setting of the pendulum in a new equilibrium position at outside influence. All these requirements were as far as possible taken into account during selection of the final setting parameters.

An example of free oscillations of the pendulum in the presence of minor mechanical vibrations caused by the operating of vacuum pumps is given in Fig. 2, series I (the experiment took place on 31 May, 2001). The diagram shows the amplitude of laser beam oscillations on the screen 16 (upper and lower rows of values) depending on the number of oscillation. The laser beam is reflected from the mirror 8 which was fixed to the pendulum. The diagram also shows the current central equilibrium position (the central row) calculated by these amplitudes. Series I represents oscillations at influence of the background mechanical noise. Series II represents the response of the pendulum to minor permanent outside force. Accuracy in determination of position of the center of a light spot was 0.1 mm.

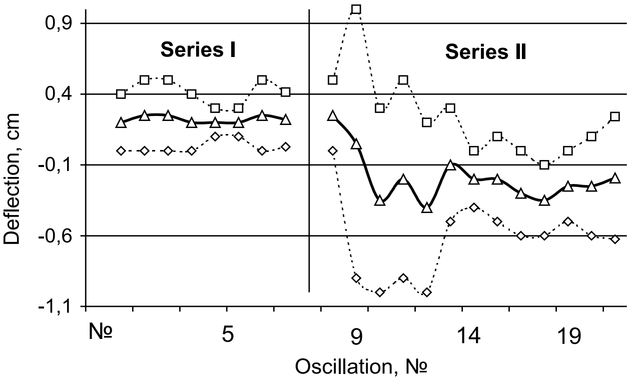

Fig. 2
Calibration Measurement 31 May, 2001

Measurement was carried out with one liquid damper in service and with increased rigidity of suspender (it was made by means of limiting the effectively operating suspender length). The period of free oscillations in these series was 29 sec. In this case there were set continuous oscillations of the pendulum around the average value of balance of 2.2 mm with the average amplitude of values fluctuation of about 0.2 mm.

Air cooling of one of the massive loads by a very light continuous air flow was carried out to study response of the pendulum to a minor constant external force. In this case (Fig. 2 Series II) noticeable change of pendulum oscillations already took place after 3-4 periods. Absolute setting of a new balance took place after 7-8 oscillations.

In the case of another series of calibration measurement (held on 7 June, 2001) both liquid dampers were used and rigidity of suspender was decreased. The period of free oscillations of pendulum was about 40 sec.

Addition of the second damper and decrease of suspender rigidity caused, on the one hand, increase amplitude of pendulum oscillations at influence of the external force and, on the other hand, in that case noticeable change of the equilibrium position of the pendulum took place after 1-2 oscillations.

Periodical checking of the invariability of initial central position of pendulum balance in time was also carried out. Thus, for example, in the series of measurement of 7 June 2001 measurement of equilibrium position was checked not only before the start of the main series of measurement, but also 2 hours after completion of the main works.

Experimental Results and Their Analysis

Within the period from 17 May, 2001 to 7 June, 2001 7 measurement series were carried out at various operation modes of the accelerator and at various pendulum parameters.

A record of check of equilibrium position of the pendulum was carried out before and after switching off the electron bunch (as well as during calibration measurement). At the same time, all electric inducing and mechanical noise remained stable within the whole measurement period. This was achieved by means of additional switching on all the devices which were used during measurement (water and vacuum pumps, magnetron, deflecting magnets, etc.) and their switching off only on completion of the measurements.

In Fig. 3 there are results on measuring the central position of the pendulum when brake target is in position 3 (see Fig. 1). Series I and III on the diagram correspond to check measurement directly before switching on and several minutes after switching off the electron bunch. Series II-A and II-B totally reflect pendulum oscillations at the time when the accelerator is operating (for about 10 minutes) and for some time after the bunch is off. Additionally, a trend line is drawn (averaging by 3 points).

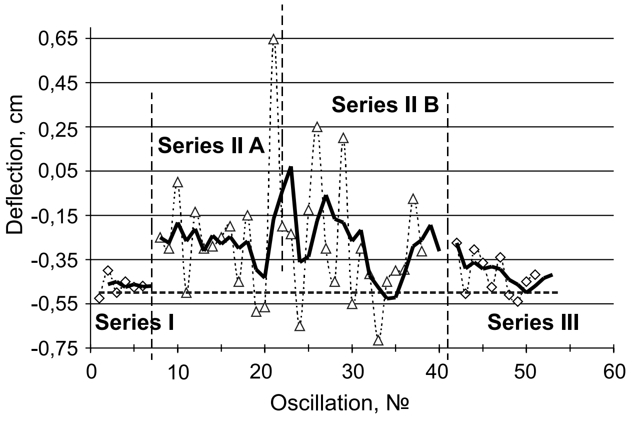

Fig. 3
Measurement of central position of pendulum equilibrium when brake target is in position 3.

Results of a similar experiment are given in Fig.4. The only difference in this experiment is that the brake target is in position 2. Also a trend line is added (averaging by 3 points). Series I and III are check measurements made directly before switching on and after switching off the electron bunch. Series II are pendulum oscillations at the time when accelerator is operating.

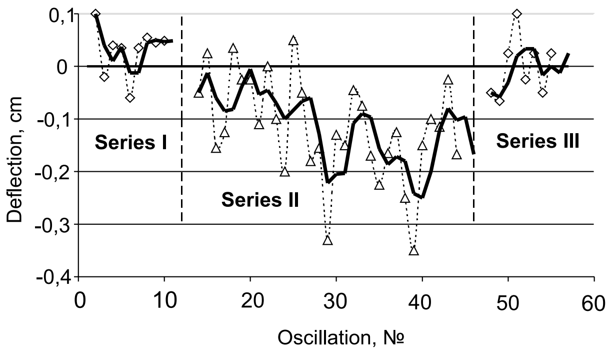

Fig. 4
Brake target in position 2

From qualitative comparison of trend lines (Fig. 3 and Fig. 4) it can be derived **that there is a correlation between switching on the electron bunch and the average deflection of the pendulum from equilibrium position if compared to checked series before and after switching on. At that, direction of deflection changes depending on what pendulum load the brake target is close to.**

Unfortunately, due to circumstances beyond the control of the author it seems to be difficult to improve experiments accuracy or to accumulate much of statistical data by now. Estimate of the value of the force which may cause such a shift of position of pendulum equilibrium was carried out. In the experiments this deflection did not exceed 1-2 mm (in the units of registering scale). Calibration of a rigid suspender (if it applies to a massive load at a pendulum end) gives the upper boundary of this force of not more than 10^{-6} N.

Conclusion

A series of experiments were made on checking the possibility to generate a field at deceleration of charged massive particles in matter.

Electrons accelerator was used as a source of charged particles. A narrow bunch of relativistic electrons (average beam power is 450Wt, electrons energy is about 30 MeV) was directed at a brake target made of wolfram where deceleration of accelerated electrons took place.

Measurement proved appearance of statistically reliable deflection of a torsion pendulum, one of massive loads of which was located close to a brake target by the time of deceleration of relativistic electrons.

Change of direction of pendulum torsion at shift of a brake target from one end of the pendulum to the other was also registered. The value of the force which causes pendulum deflection has the upper boundary of N.

Of course, these first experimental results on checking the predictions made on the basis of development of the model of extended space are of preliminary nature and need more thorough checking. That will be the basis of future experiments.

References

1. Tsipenyuk D. Yu. Andreev V. A, "Extended space and model of combined interaction", Short reports on physics FIAN, No.6, pp. 23-24, (2000); gr-qc/0106093
2. Tsipenyuk D. Yu., Andreyev V. A., "Explored in Russia" E-magazine,60, (1999); http://zhurnal.ape.relarn.ru/articles/1999/060.pdf
3. Tsipenyuk D. Yu., Andreyev V. A., "Structure of extended space", Preprint IOFAN, 5, 25p., Moscow, (1999)
4. Tsipenyuk D. Yu., "Field transformation in the model of extended space: prediction and experimental test", Gravitation&Cosmology, Vol.7, No.4(28), pp336-338, (2001); physics/0203017
5. Tsipenyuk D. Yu., "Field transformation in the extended space model: prediction and experimental test", Short reports on physics FIAN, No, 7 p.39-49, (2001); physics/0107007
6. Tsipenyuk D. Yu., Andreyev V. A., "Lienar-Vihert potentials for the extended space model", Short reports on physics FIAN, No 6 p.3-15, Moscow,(2002); physics/0302006
7. Tsipenyuk D. Yu., "Explored in Russia" e-magazine, 81, p.907-916, (2001); http://zhurnal.ape.relarn.ru/articles/2001/081.pdf
8. Tsipenyuk D. Yu., Andreyev V. A., "Electrodynamics in extended space", Preprint IOFAN, 9, 26p., Moscow, (1999)
9. Tsipenyuk D. Yu., Andreyev V. A., "Interaction in extended space", Preprint IOFAN, 2, 25p., Moscow,(2000)
10. Klein F. Zeits.f.Math.a.Phys., 375, (1901) (see translation in collection "Variation principles of mechanics", M.: Physmatgis, 1960)
11. Kaluza Th. Sitz.d.Preuss.Akad., 966, (1921) (see translation in collection "Albert Einstein and gravics", M.: Mir, 1979)
12. Klein O. Zeits.f.Phys., 37,895, (1926)
13. Mandel H. Zeits.f.Phys., 39,136, (1926)
14. Fock V. Zeits.f.Phys., 39,226,1926
15. Rumer Yu. B. Investigations on 5-optics, M., Gostehizdat, 1956
16. Vladimirov Yu. S. Dimensionality of physical space-time and integration of interactions. M., MGU, 1987.
17. Green M., Schwartz G., Vitten E. Theory of superstrings, v.1, 2, M.: Mir, 1990
18. Wesson P.S. Space, time, matter: Modern Kaluza-Klein theory. Singapoure, World Scientific publishing Co.Pte.Ltd, 1999
19. Tsipenyuk D. Yu., Andreev V.A., "Lienar-Vihert potentials and Lorenz force in extended space", Preprint IOFAN,1,20p., Moscow,(2001)
20. Tsipenyuk D. Yu., Andreev V. A., "Gravitational effects in extended space", Preprint IOFAN, 4, 15p., Moscow,(2001)
21. Tsipenyuk D. Yu., "Field transformation in the model of extended space: prediction and experimental test", Preprint IOFAN, 5, 24p., Moscow, (2001)
22. "Tsipenyuk D. Yu., report "Electromagnetic and gravitational interaction in extended space" at the Conference of Nuclear Physics Department of the Russian Academy of Sciences (RAN) "Physics of Fundamental Interactions", 27 November-1 December 2000, ITF, Moscow, http://www.itep.ru/
23. Tsipenyuk D.Yu. and Andreev V.A., "Interval as the fifth coordinate" proc. of 5-th Int. Conf. on Gravitation and Astrophysics of Asian-Pacific Countries, PFUR, Moscow, 2001, pp.29-30. http://rgs.da.ru/
24. Tsipenyuk D.Yu., "Direct transformation of the electromagnetic field to gravitational field in the model of extended space: predictions and possible way of its experimental test", proc. of 5-th Int. Conf. on Gravitation and Astrophysics of Asian-Pacific Countries, PFUR, Moscow, 2001, p.28. http://rgs.da.ru/
25. Tsipenyuk D.Yu., "Model of extended space: Prediction and test experiments on possible transformation of electromagnetic field into gravitational field", proc. of XVIth Workshop on High Energy Physics and Quantum Field Theory (QFTEP), Moscow, Russia, 2001, pp.398-405. http://theory.sinp.msu.ru/~qfthep/2001/Proceed2001.html
26. Tsipenyuk D.Yu., "Model of extended 5-d space and possible checking experiments", presented at 5-th Moscow international ITEP school of physics (30-th ITEP winter school of physics), 20-28 Febr., Moscow, 2002, http://www.itep.ru/
27. Tsipenyuk D.Yu., report "Second experiment on checking prediction on field transformation in extended space" at the scientific conference of nuclear physics section OFN RAN "Physics of Fundamental Interactions", ITF, 2-6 December 2002, Moscow, http://www.itep.ru/
28. Urbah V. Yu., Mathematical statistics, M., USSR Academy of Sciences Publishers, 1963.

Joe Flynn's Parallel Path Magnetic Technology

Tim Harwood, USA

Email: timharwood@usa.net

There is a widespread opinion common to the mainstream academic community and also to various alternative scientific forums, that some kind of exotic new physics will be required to design and implement over-unity technologies. That is to say electrical motors, electrical generators, or other apparatus, which produce an excess of magnetic force or energy above the value actually inputted.

However, this has recently been experimentally demonstrated not to be the case, as I have validated myself in simple $20 experiments undertaken at home with parts bought from the local hardware store. It is the purpose of this article to educate readers that with care, thought, and a little work, it can be demonstrated that existing textbook physical law, freely allows for the extraction of excess electrical energy from magnetic systems.

The Flynn Research Project

Joe Flynn has been engaged in magnetic flux research for over 25 years now. His work is long standing, comprehensive, and in later years, well funded. It is reported $7m has been spent to date, with over $1m alone developing a revolutionary high performance magnetic motor. His equipment is validated, and apparently already in mass production for select customers. Yet few researchers have heard of Joe Flynn. This surprising situation will hopefully shortly be changed, and Joe Flynn awarded the scientific accolades he deserves, for being the genius master mind behind one of scientific history's most outstanding research projects. Since many lines of research have been formulated and explored by Joe Flynn, the following article presents only a brief summary of some of his best art apparatus, but is nonetheless sufficient to convey the basic ideas.

The first illustration (Fig. 1) is taken from Joe Flynn's US patent 6,246,561, and explains a simple magnetic force multiplication experiment, which forms the basis for the Flynn magnetic art. If the windings on either side of the central magnet, which are normally connected in series, are properly pulsed, the field of the permanent magnet in the center will be diverted to the opposite side of the core flux path provided. Or in alternative language, the side of the core that is pulsed is demagnetized, relative to the field of the permanent magnet used in the apparatus. This is elementary textbook physics anyone can understand.

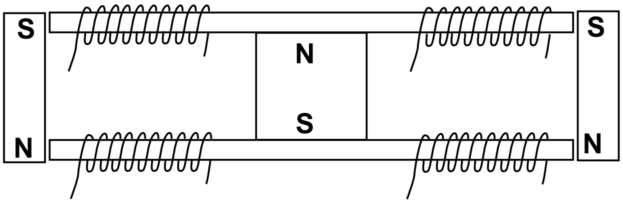

Fig. 1

1.75 times more force is delivered to the legs of the core than is provided by the electrical input to the control coils

So what is surprising about this apparently simple apparatus is that the armature on the side of the flux core will contain 1.75 times more units of magnetic force, that could be manifested by the electrical input to the apparatus alone. Since the ability to arbitrarily move the force from one point to another is the basis for motion or work, however simplistic, we therefore have a basis for a system that can be developed for practical technological purposes. Expressed in alternative language, we also have the capability to engineer a time varying magnetic field, without the need for moving parts, which will allow development of systems that output electrical energy. Both capabilities are highly desirable, and offer substantial opportunity for technical development.

Following on from this basic experiment, there is a second simple and logical improvement in layout illustrated in Fig. 2, which should be obvious, but has been shown not to be the case. In this instance, the pulse is centrally located, and a dual flux field layout employed, which both demagnetizes the core relative to one magnet, and magnetizes it relative to the other. Since the two actions are complementary, the input required to manifest the flux switching effect is halved, therefore doubling 'efficiency.'

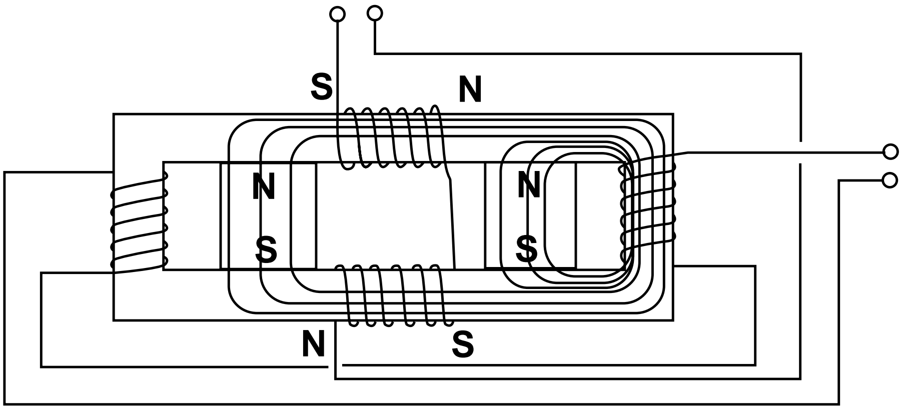

Fig. 2

3.47 times more force is delivered to the legs of the core than is provided by the electrical input to the control coils

It should be noted that while the efficiency is doubled, the absolute output may not be significantly improved. This is because the major weakness of this effect and technology is flux saturation of the core, with values depending upon the specific properties of the B-H curve of the core material employed, limiting the absolute output of both layouts the same.

The previous statements are not required to be taken on trust, and simple experiments have been proposed by Joe Flynn, such that anyone can validate this effect for themselves. Figure 3 is a simple experiment taken from the Flynn website (http://www.flynnresearch.net), that can be used to validate the principals put forth in this article.

Simple Magnetic Force Multiplication Experiments

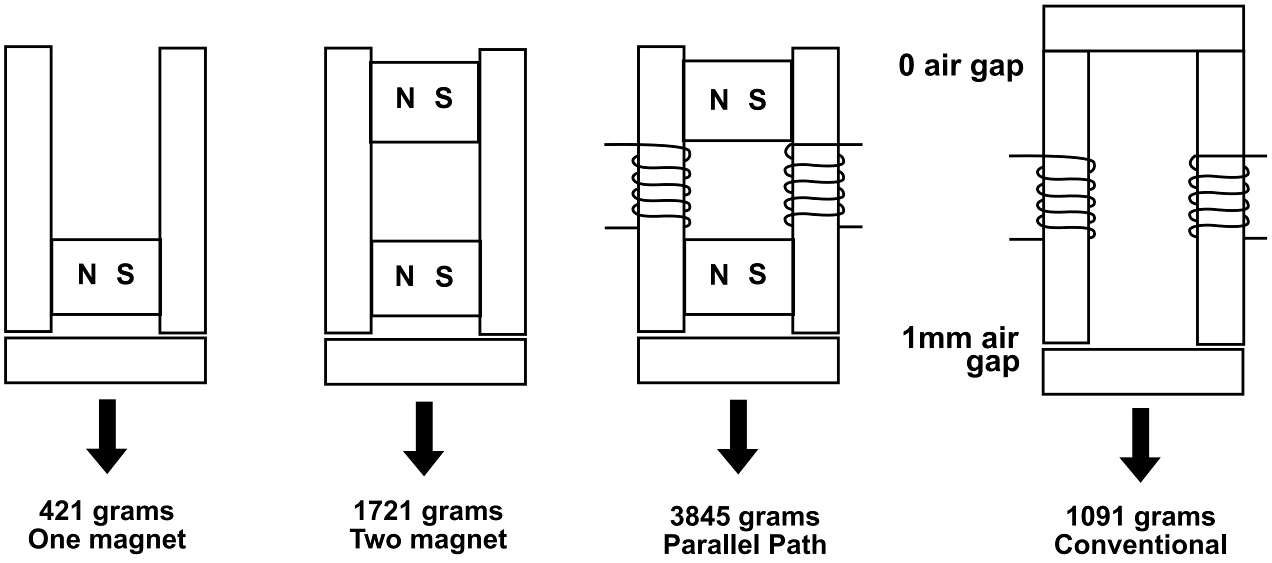

Fig. 3

Parallel Path can deliver 3.47 times more force to the legs of cores than any competing conventional technology

An even simpler non electrical flux experiment was proposed by GM in the Parallel Path Egroup. My apparatus is illustrated below in Figure 4. It is no more than magnets and steel staple strips, bought from a local hardware store for a total of under $10. The Parallel Path effect can be replicated with identical apparatus, at only a slight increase in cost and complexity, with the addition of a simple 12V polarity reversible power supply, such as those commonly sold to power computer speakers, among other applications.

Simple Magnetic Force Experiments

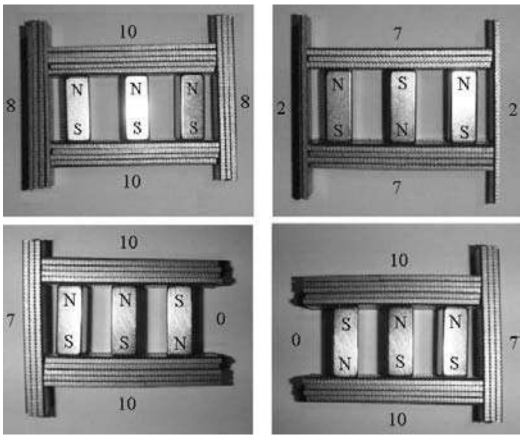

Fig. 4

Relatively small changes in layout produce large changes in force

Conservation of Energy / Field Potential

One of the aspects of the Flynn technology people find most difficult to understand, is how you can have a device that delivers 3.47 times more magnetic force than is electrically inputted, yet not violate accepted principals of text book physics, as stated in the introduction. The key point here is conservation of energy. Magnetic fields do not gain energy – they are conservative. You can only ever obtain less energy from a magnetic field minus losses, than is in fact present. I feel this apparent puzzle can not be better explained, than by reference to Joe Flynn's own words:

"Since the Parallel Path System produced 3.47 times more force than the conventional system, with the same electrical input, it appears to violate conservation, this is only true when observed from a traditional view point. The system contains three flux producing sources (2 magnets and an electromagnet) which together are capable of producing a far greater force than is actually produced. All of the flux sources together can produce a force of 13.11 units, therefore in the physical sense a loss of 1 - (9.01 / 13.11) = 31% is realized."

So the system is 350 % efficient, in terms of delivered magnetic force compared to net electrical input, yet still conforms to the accepted physical principals of energy conversation, by being only 69 % efficient, in terms of the fields present in the system. However surprising this result may appear, the analysis presented is in outline correct, with the difference between fields present in the system, and net electrical input, being the important concept presented.

Losses in the System and Optimisation

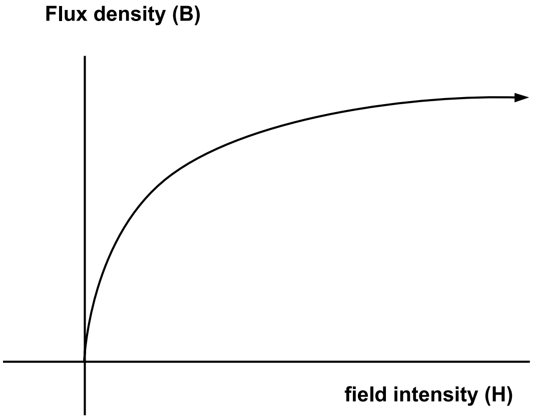

Fig. 5

In order to properly optimise flux cores, an appreciation of the physics that underlies the transfer of flux within a core is required. The normal magnetization curve, or B-H curve, is a mathematical relationship between applied field intensity H, and resultant flux density manifested in the core B. It varies according to core material, and the curve will shift, if there is a starting magnetism within the core, such as that provided by the field of a permanent magnet. If the start magnetism is excessive, the core is saturated, and will not properly respond to the applied force H. A simple B-H curves is illustrated in Fig. 5.

Hysteresis is a delay between applied magnetic force H, and resultant flux density B, that again varies according to material type. It also manifests as a delay between the termination of force H, and the manifestation of flux density B. So, the system will not turn on instantly, and will not turn off instantly, in simple terms. This is because the magnetic memory of the core, means a flux vector remains within it, even when the application of magnetic force H has been terminated. If we apply a reversed force H to the core, the basic B-H curve is now expanded as in Fig. 6, with the memory effect also illustrated.

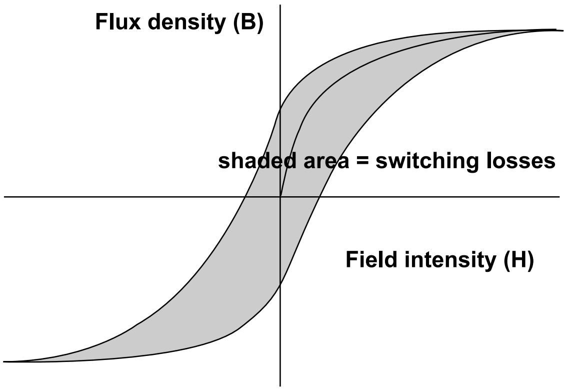

Fig. 6

Thus as can be seen, to return to the initial switched state, the remnance magnetism must now be overcome, hence input once in operation, will be greater than that required for the very first pulse. The area within the hysteresis curve gives a rough estimate for the amount of wasted energy, and along with other conventional sources of losses resultant in flux transfer within a core, is what reduces the efficiency of flux cores from maximum values of 2, or 4, down to values such as 1.75 or 3.47, typically.

Motor Apparatus

Simple Parallel Path Technology Demonstrator Motor

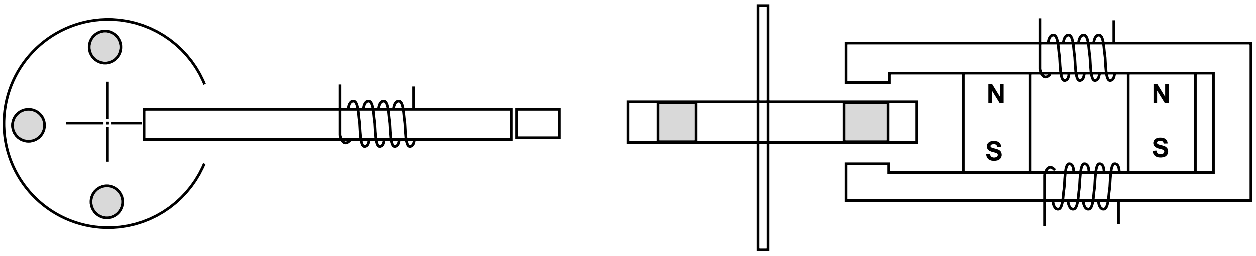

Fig. 7

Designed to demonstrate core principals, not provide over-unity

Although numerous practical applications abound for this effect, electric motor design remains the most outstanding opportunity. To this extent, again a few simple images, should be sufficient to explain how the basic flux switching apparatus, can be turned into a highly efficient electrical motor.

The first motor shown in Fig. 7 is one I have proposed to validate the flux switching effect at a most basic level. It illustrates the point made in the Flynn patent, that the armature of the core can be removed, and replaced with a motor flux path. This first motor is not claimed to be highly efficient, but it helps one to understand how the transition from simple flux core to motor takes place.

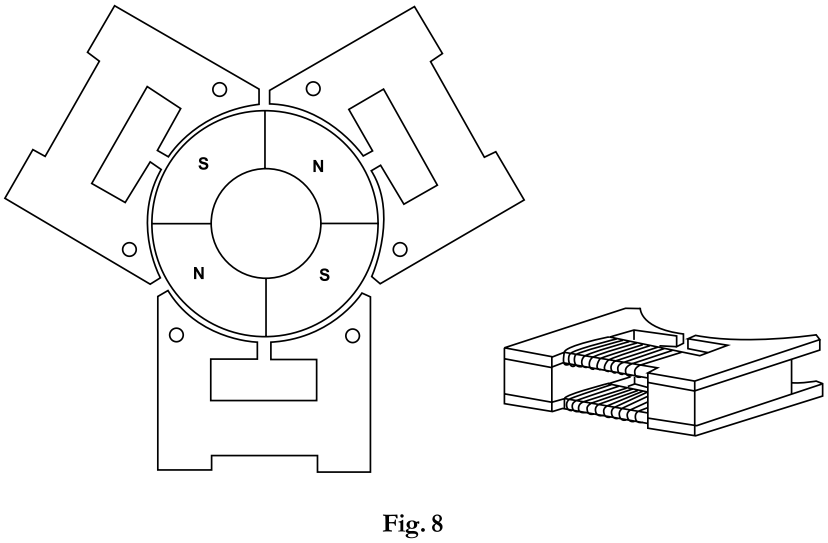

Fig. 8

Magnetic force is alternately switched from one leg to the other, imparting motion to the rotor sections

The next motor shown in Fig. 8 is again taken directly from the Flynn patent, and illustrates the next intermediate step to motor design. The fields of the permanent magnets are alternatively switched from one side of the surrounding flux cores to the other, alternately interacting with N and S poles on the rotor, imparting motion to the central rotor shaft.

With proper financial support, and the facilities to have metglas cores custom moulded, Joe Flynn was able to develop his final best art, shown in Fig. 9. No detailed performance numbers have been released for this motor, whose precise performance characteristics remain proprietary to Joe Flynn at this time. But the optimisation is so expert, it is stated to posses certain exotic properties, such as cool ambient operation, even during prolonged periods of continuous load. This 'cold running' is said to be of great interest to the American military, as it offers excellent stealth performance characteristics.

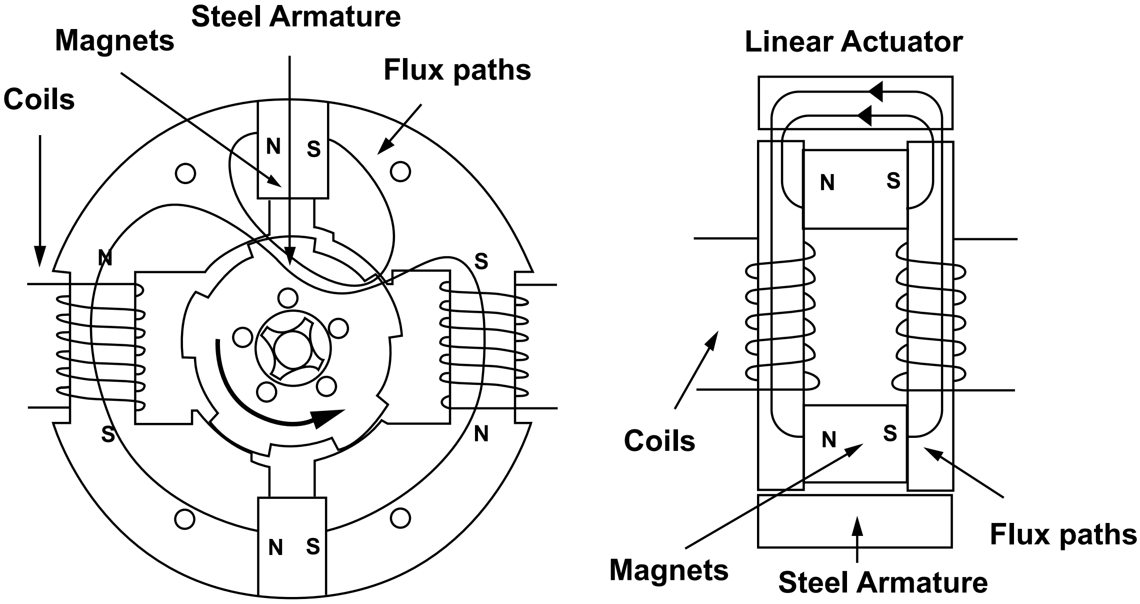

Fig. 9
Optimal Flux Core Motor Construction

Electrical Apparatus

Many readers will no doubt have noticed the similarity of the first illustration presented in this document, to the so called ‘Tom Bearden MEG.’ This is fair comment, and Joe Flynn has always highlighted this issue. However, it has been commonly stated Joe Flynn has simply developed mechanical apparatus, and the MEG with its electrical functionality, is distinct art, more advanced than the mechanical Flynn apparatus. However, this is shown not to be the case by a careful examination of the Flynn patent, in which the following is stated in the ‘Power Conversion’ section:

‘The construction (shown in Fig.10 A) utilizes four control coils and a single permanent magnet and the construction (shown in 10 B) uses two control coils and two permanent magnets. The flux that would normally be supplied by a primary winding is supplied by the static flux of the permanent magnet or magnets and the control coils convert this static flux into a time varying flux in a novel way. Both arrangements use two secondary coils, the secondary coils are placed in the region of the continuous flux path that would be occupied by an armature or rotor in the linear or rotary arrangements. The regions of the flux paths that perform work are the same in all cases.’

‘By alternating the polarity of the control coils during one cycle, one working region experiences an increasing flux and the opposite region experiences a decreasing flux and during the next cycle the opposite occurs. This results in the induction of a voltage in the secondary coils that is decided by the magnitude of the change in flux in the working region and the time in which this change occurs. The novelty of this discovery is that the primary flux inducing the voltage in the secondary coils is supplied by the permanent magnet or magnets and is far greater than the flux supplied by the control coils.’

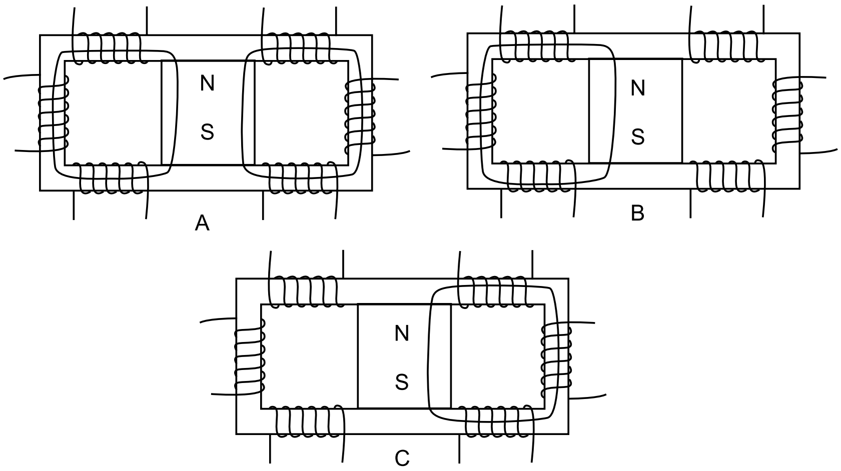

Fig. 10

Fig.10 taken directly from the Flynn patent makes the point even clearer. As can be seen, the device illustrated is in all functional respects absolutely identical to the so called 'Tom Bearden MEG.' This identical prior art, therefore calls into question the intellectual property rights of Magnetic Energy LTD, as the same invention can not be patented more than once. Joe Flynn has also stated that his intellectual property rights will be robustly defended, by legal action if necessary, and he regards himself and his company as being in possession of exclusive rights to the so called 'MEG' unit.

the input and output circuits must be closed in series. The disclosure of this technique amounts to putting the basic MEG methodology fully into the public domain.

As regards replication of electrical output orientated flux core devices, certain important details need to be stated. For example grade 8 ceramic magnets should be used, so as to avoid flux saturation of the core. A basic error, many early experimenters wasted time on. The requirement for strong magnets to obtain over-unity results, is as much of a myth, as the idea new physics is required.

But perhaps the greatest trade secret of the electrical devices, one which several lengthy non disclosure agreements are required to be signed before it can be disclosed, is that the input and output circuits must be closed in series. The disclosure of this technique amounts to putting the basic MEG methodology fully into the public domain.

The reason for this circuitry requirement is obvious enough, with only a little analysis. If the output circuit is closed when the input circuit is activated, then the input energy simply leaks into the output circuit, as in an ordinary transformer. So no flux switching effect is manifested, and the field of the permanent magnet is static in time. Thus you have an ordinary transformer, with reduced efficiency, because of the core flux saturation effect provided by the permanent magnet.

This is one of the most important point to make about the Flynn apparatus. If you approach it as if it is a normal piece of scientific equipment, then proper optimisation is not greatly problematic. For example more turns on the output coils, simply means more voltage and less current, exactly as standard textbook equations predict. Generally, problems only occur, if you imagine the effect is based upon exotic scalar type or vacuum energy physics, when in fact it is ordinary flux manipulation within a core.

Present Status of the Flynn Project

Initially Joe Flynn was remarkably open about his work and research. However, since performing a working demonstration of various advanced hardware samples for the American Department of Defence, little has been heard.

I want to clearly emphasize I do not speak for Flynn research, nor am I in any way connected with Flynn research, and by consequence, have no inside information whatsoever as to the present status of the project. But we all sincerely hope, that the project has not been swallowed whole by the American deep black military industrial research complex.

However, even if this is the case, it does not mean the technology is lost. Extensive and generous details have been provided by Joe Flynn of his research, both in his patent deposition, website, and other comments, such as to enable persons of scientific training and skill, to replicate the effects stated.

While replication of the electrical effect remains extremely demanding, the mechanical apparatus is very easy both to understand and replicate. There is no reason why scientists and home tinkers together, can not build Flynn type flux core motors, and explore over-unity flux manipulation for themselves. The future has arrived, and it is simpler and cheaper than anyone imagined to be possible.

Raum&Zeit

EhlersVerlag GmbH,
Geltinger Str.14e,
82515 Wolfratshausen
Telephon: 08171/41 84-76,
Telefax: 08171/41 84-66
Internet:
http://www.raum-und-zeit.com

Institut fur Raum-Energie-Forschung (IREF) i.m. Leonard Euler

- Global Scaling Theory
- Gravitation theory
- Mathematic Biology
- Gravitation energetics
- Hydrogen-Helium Energetic
- Information technologies
- Telecommunication technologies
- Transport technologies

Flynn Research Company

http://www.flynnresearch.net

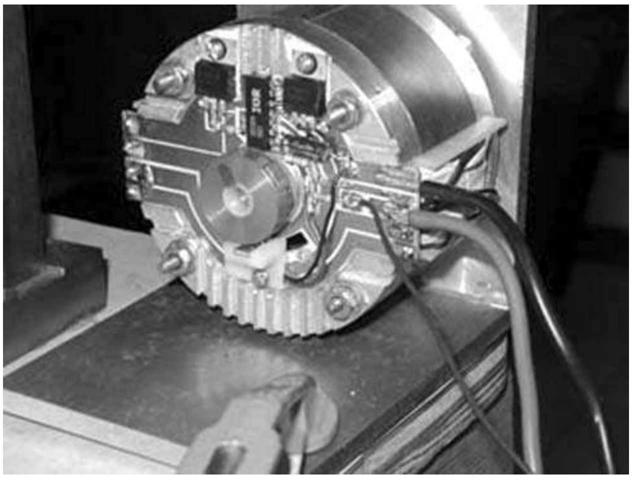

Photo 1

3.3" 1.2" stack height 1/2 hp weight 59 oz

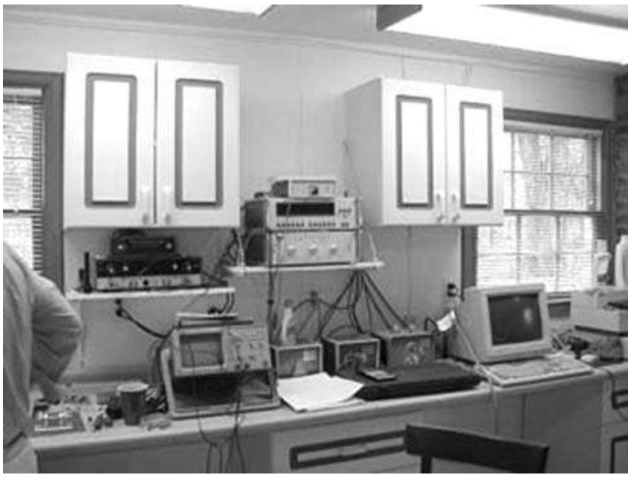

Photo 2

Dynamometers, Power Analyzer etc.

Editor: Below we publish information and photos, represented on Flynn Research Company website (http://www.flynnresearch.net), and a description of the patent on methods for controlling the path of magnetic flux.

Flynn Research is dedicated to research focused on the novel application of permanent magnet and electromagnetic technology. Magnetism is a field that has potential for tremendous growth from both a science and application perspective. The researchers of Flynn Research Company believe that magnetism is one of the least understood and most misinterpreted of the natural forces, with the exception of gravity.

Even though much work has occurred in the advancement of magnetic materials, little advancement has occurred in the basic understanding of magnetism. The field of magnetism is based on both theory and 'myth' and is generally described as a 'phenomenon'.

We still do not know what a line of flux is and how it travels through space. We know flux "appears" to originate within a permanent magnetic material, forms a loop from one end to the opposite end of the magnetic material, appears to occur without a time constant, produces a force (attractive and repulsive) between other permanent magnets and "magnetically permeable materials and has a relationship to current flow. Everything known about magnetism is based on "effect and relationships" without a clear understanding or provable definition of what is a line of magnetic flux.

Everything known about magnetism is based on "effect and relationships" without a clear understanding or provable definition of what is a line of magnetic flux.

The research is based on the "magnetic force of attraction effect" that is related to magnetic flux by the number of lines of flux occupying a given area or flux density squared. The simple fact that if the number of lines of flux occupying a given area "doubles" the force of attraction becomes four

Photo 3

8" dia x 8" height motor presentation

Photo 5

Technology Discussion

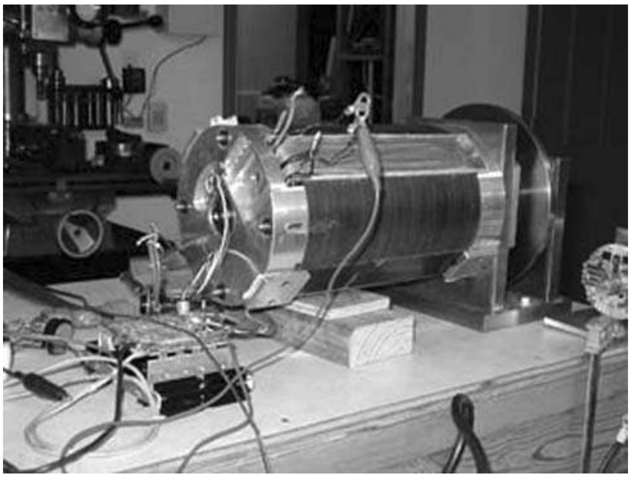

Photo 4

8" X 8" motor testing

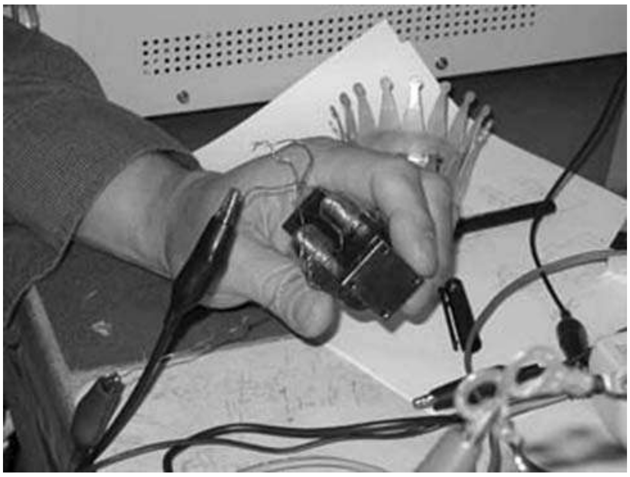

Photo 6

Parallel Path Test Apparatus

times greater. The basic model shows that with specific structures electromagnetism acting with permanent magnet materials can divide and multiply force in a manner that cannot be explained with conventional physics.

Rotary devices, linear devices, reciprocating devices and power conversion devices are now a patented product of this research. The goal of the Company is to commercialize their devices for use in an energy dependent world. Flynn Research is a research & development company and in cooperation with Magnetic Revolutions LLC licenses their technologies for use in the electric motor and power conversion industries.

The researchers do not construct their own test equipment nor write capture and data evaluation programs but use industry / scientific accepted products. Close to $1.5 million has been expended developing, protecting and evaluating the technology. Many fine minds have been involved in the development and evaluation of the technology. Many presentations have been given both to motor companies and academic groups. The prototypes have all been professionally built, which include a 1/2 hp 3.3" dia 1.2" stack height, the assembled motor weighs 59 oz and a motor 8" dia with a stack height of 8" currently undergoing testing.

United States Patent No. 6,246,561
June 12, 2001

Methods for controlling the path of magnetic flux from a permanent magnet and devices incorporating the same

Inventors: **Flynn; Charles J.** (Greenwood, MO)
Assignee: **Magnetic Revolutions Limited, L.L.C** (St. Louis, MO)
Appl. No.: **127056**
Filed: **July 31, 1998**

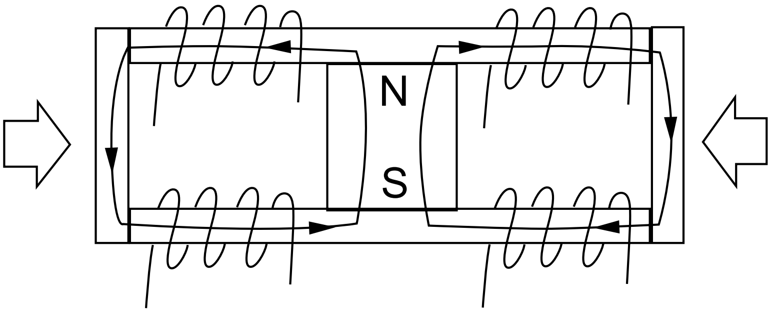

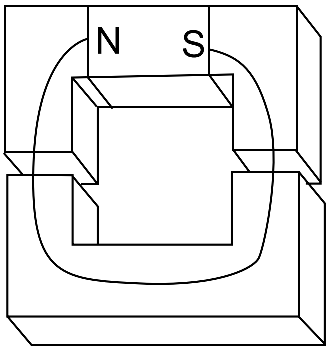

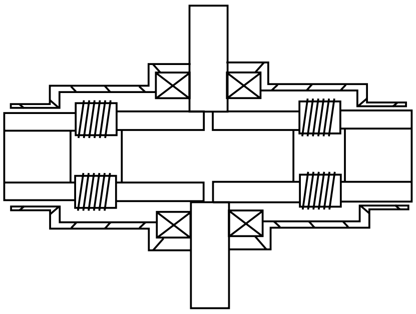

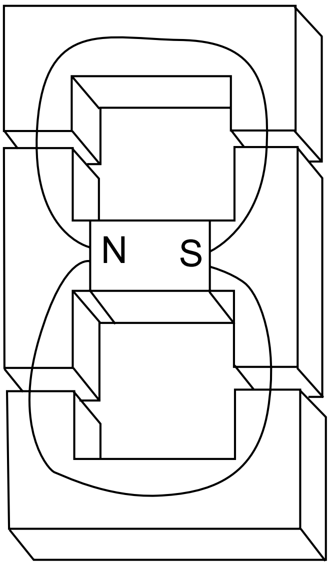

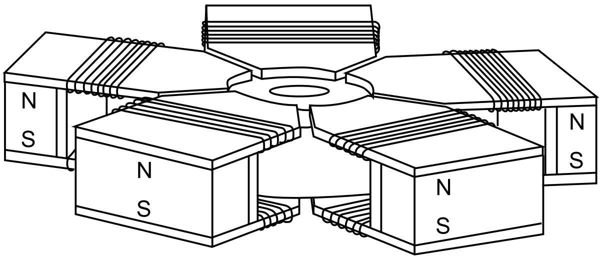

A permanent magnet device includes a permanent magnet having north and south pole faces with a first pole piece positioned adjacent one pole face thereof and a second pole piece positioned adjacent the other pole face thereof so as to create at least two potential magnetic flux paths. A first control coil is positioned along one flux path and a second control coil is positioned along the other flux path, each coil being connected to a control circuit for controlling the energization thereof. The control coils may be energized in a variety of ways to achieved desirable motive and static devices, including linear reciprocating devices, linear motion devices, rotary motion devices and power conversion.

Editorial: The principle of designing of the device (two counter parts of the flux are used) resembles Φ-machine or Gramm's generator. Read in this issue more details on these devices.

Flux-machine And Its Analogues

Review prepared by correspondent Alla Pashova, Russia

Flux-machine (or Ф-machine, since configuration of its field resembles the Russian letter "Ф") was designed by Alexander V. Frolov during the first half of 1990s. Ф-machine description was published more than once on the pages of scientific periodicals (1994, Institute of New Energy, Newsletter, June 1994, p.9.; 2002, New Energy Technologies, 2002, p.3), and it is well known to many researchers concerning themselves with such kinds of engineering.

Let us consider the arrangement of Ф-machine; see Fig. 1 (first published in 1994). A report was presented at the conference «New ideas in natural science», 1996 in Saint Petersburg (see photo in Fig. 1). The primary coil is mounted in the center of generator; two secondary ones are diametrically wound on a ring core. An air gap between ring and central magnetic circuits are of particular importance. **Two magnetic fluxes from two coils are balanced, and, thereby, there is no reaction in the primary circuit.**

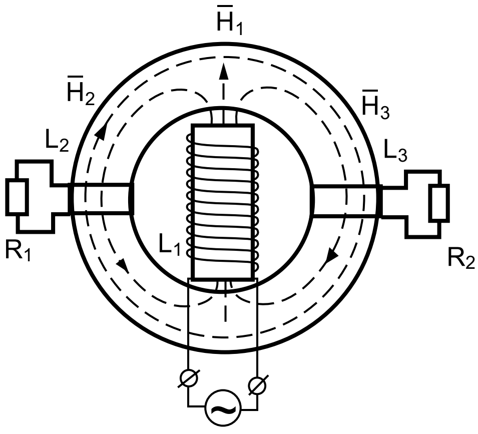

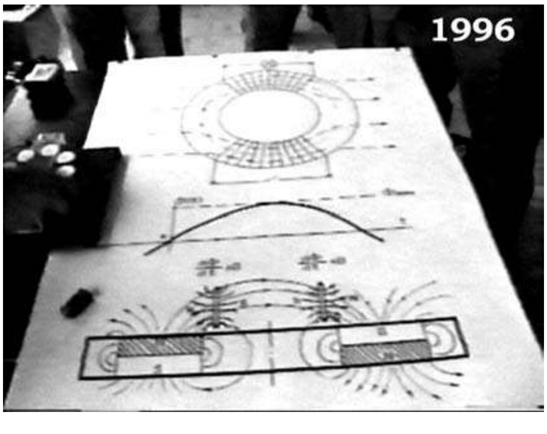

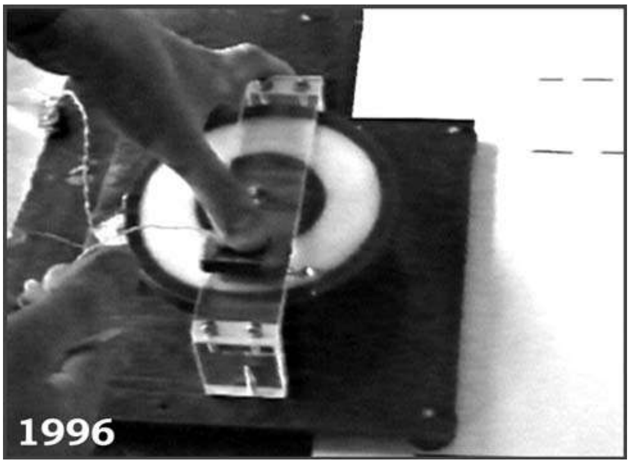

Fig.1
See also Fig. 7, p.29 – the analogy is evident

The device bears some similarity to the Gramm's generator (Zenob Theophyl Gramm, 1826-1901, Belgium-France, took out a patent for an electric ring-rotor generator in 1869). In Gramm's generator (Fig. 2) the ring rotor with a toroidal winding rotates. The toroidal winding touches two diametrically located conducting brushes. It turns out that in the ring of the rotor all winds of one rotor half create a field which is directed towards a field created by the other rotor half.

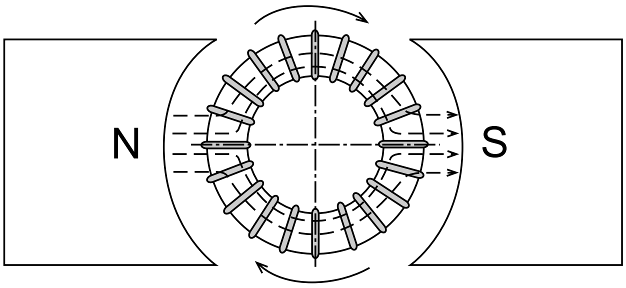

Fig. 2

An example of analogous device is represented in Fig. 3 (information is from the website: http://www.skif.biz/energy/arhiv1-6.shtml), and also in Fig. 4 (we have written about this invention in New Energy Technologies, Issue #5 (8), 2002, article by V.I. Boryak, Email: spin@i.com.ua).

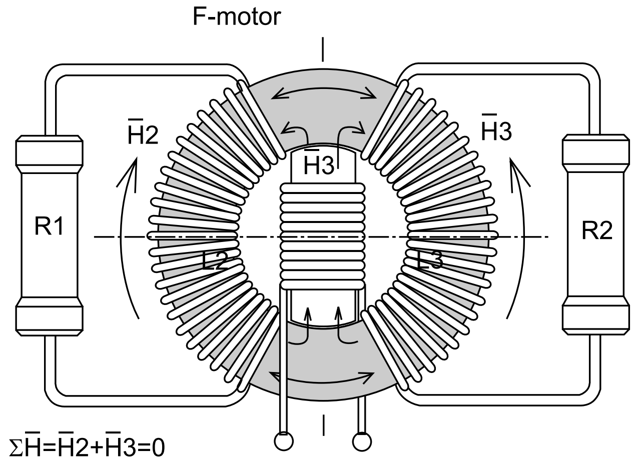

Fig. 3

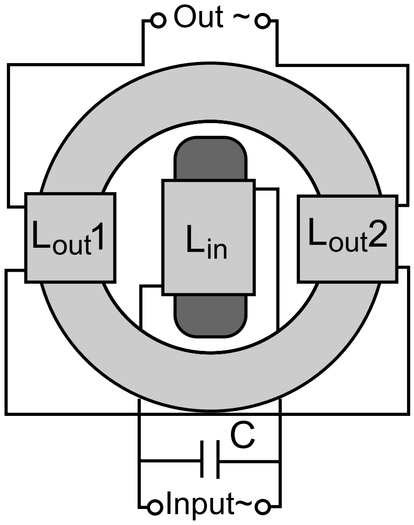

Fig. 4

In August 1999, according to Stephen Hartman (Hartman Multimedia Service, Email: hart@harti.com, info@ccard.net), a German research team designed a toroidal generator. As stated by the researchers, it obtained more than 1.200 W by load while efficiency exceeded 100%!

This subject has become very popular among different researchers. Fig. 5 illustrates a design with a magnet rotating in the center (designer - Olaf Berens, Email: olaf.berens@prognost.com).

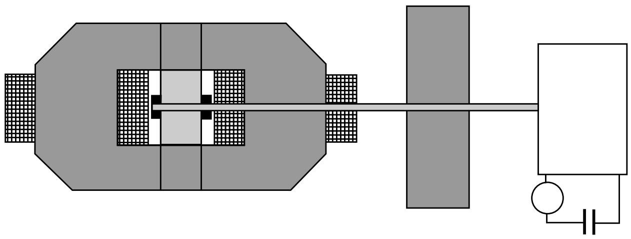

Fig. 5

One more example from the USA: in February, 2003 Donald Hofmann (USA) filed an application for patenting "Generators and transformers with toroidally wound stator winding». Below we publish the description of the patent.

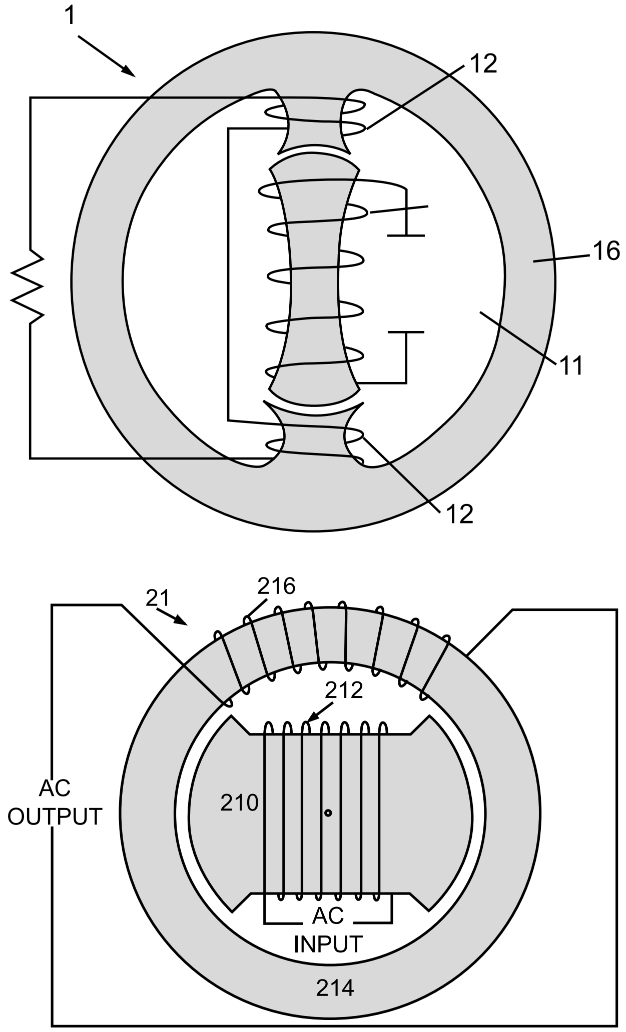

Fig. 6

Inventors: **Steven L. Sullivan, USA; David L. Goulet, USA; Donald Hofmann ,USA.**

Electrical generators are provided with toroidally wound stator windings electrically connected in series; a high permeability stator core, preferably an amorphous magnetic alloy material, glassy metal or HYPERCO™ 50 laminations; and a rotor. The toroidally wound stator coils and the stator core trap essentially all of the flux fields generated by the stator coils within the stator core. Since there is essentially no magnetic field leaving the stator, there is essentially no flux field interaction with the field generated by the rotor. The reduction of flux field interaction also reduces counter torque.

According to another embodiment of the invention, open positions of a rotor may be filled with magnetic material, preferably iron. Filling open portions of the rotor may assist in reducing undesirable flux leakage from the rotor. By trapping substantially all of the flux within the rotor, interaction between rotor and stator flux is reduced. In this manner, counter torque is reduced thereby increasing the overall efficiency of the device.

A prototype of this embodiment was constructed using a standard 10 kWtt gas-powered generator. The stator was removed from the generator and its housing. The stator windings were removed from the stator and rewound by hand using a No. 10 size copper wire with 180 winds wrapped in a toroidal fashion around approximately 180 degrees of the stator. The stator was then reinstalled in the modified housing and re-attached to the gas-powered generator. A standard 12 volt car battery was attached to the input of the rotor. A 12.92 volt input with 0.8 Amps was measured as the input to the rotor. The output was measured as 6.5 volts with a 40 Amp output. A series of input and output measurements were taken for various inputs.

Editor: Thus input power is 10 Wt, and output power is 260 Wt.

So, the old well-known ideas have found new ways of embodiment, though their essence remains unchanged. Let us note that Faraday Labs Ltd has been working on creating a prototype of ▯-machine with a toroid of 200 mm in diameter.

Please, read more about this in the next issue!

◆

Editorial: We publish the specification of several interesting patents, which were taken out by Russian inventors for recent years.

Russian patents on alternative energetics

- №93001754, 1996.06.27, Converter of gravitational forces into energy, V.V. Mironov and others
- №92008720, 1995.02.10, Slavic method to produce energy by conversion of gravitational forces, V.V. Mironov
- №93006696, 1995.04.30, Supporting converter of gravitational forces into energy, V.V. Mironov
- №2001121071, 2003.04.10, Method to create energy in liquid, heat-generator and heat-and-power device, A.Yu. Baurov and others
- №2132109, 1999.06.20, Generator of mechanical energy, A.Yu. Baurov and others
- №2001125794, 2003.05.10, Generator of static electricity, S.I. Danilov
- №2001120796, 2003.04.10, Electric power device, D.M. Beliy
- №98116077, 2000.05.10, Heat engine (monotherm, “perpetual mobile of second-type”), V.I. Likhachov and others
- №93048971, 1996.04.20, Float engine, I.P. Tchinarev
- №94023174, 1996.03.27, Gravitational-hydrostatic engine, I.P. Tchinarev
- №98112650, 2000.03.20, Water engine, V.A. Gylchuk
- №2001115953, 2003.03.20, Potential engine by Prokopenko, V.F. Prokopenko
- №2001115954, 2003.03.20, Potential amplifier by Prokopenko, V.F. Prokopenko
- №2001124364, 2003.04.20, Drum of gravitation (gravity), I.A. Strelnikov
- №95107181, 1997.02.20, Method to convert energy of physical vacuum into energy of physical space, O.K. Polev and others
- №92008695, 1995.04.20, Method to get overunity efficiency, Yu.V. Karasev and others
- №93006851, 1995.04.30, Principle, method and device for circulation of mass by force of gravitation and inertia, G.S. Kirichenko
- №93039885, 1996.05.20, Hydro-gravitational electric power station, L.S. Konkov
- №93036371, 1995.11.27, Method to convert energy created in result of interaction of earth’s gravity force and air bearing capacity into electric energy, V.I. Kovalenko
- №93025690, 1996.07.27, Method and device to convert heat energy into mechanical one, A.I. Lendyaev
- №2001122885, 2003.04.20, Rotary method to convert magnetic energy into mechanical one, L.N. Mezentsev
- №2001106783, 2003.02.10, Gravitational source of energy, A.D. Yurik and others
- №2001115968, 2003.03.20, Method to convert energy, V.S. Gorelyuh

According to another embodiment of the invention, open positions of a rotor may be filled with magnetic material, preferably iron. Filling open portions of the rotor may assist in reducing undesirable flux leakage from the rotor. By trapping substantially all of the flux within the rotor, interaction between rotor and stator flux is reduced. In this manner, counter torque is reduced thereby increasing the overall efficiency of the device.

A prototype of this embodiment was constructed using a standard 10 kWtt gas-powered generator. The stator was removed from the generator and its housing. The stator windings were removed from the stator and rewound by hand using a No. 10 size copper wire with 180 winds wrapped in a toroidal fashion around approximately 180 degrees of the stator. The stator was then reinstalled in the modified housing and re-attached to the gas-powered generator. A standard 12 volt car battery was attached to the input of the rotor. A 12.92 volt input with 0.8 Amps was measured as the input to the rotor. The output was measured as 6.5 volts with a 40 Amp output. A series of input and output measurements were taken for various inputs.

Editor: Thus input power is 10 Wt, and output power is 260 Wt.

So, the old well-known ideas have found new ways of embodiment, though their essence remains unchanged. Let us note that Faraday Labs Ltd has been working on creating a prototype of Φ-machine with a toroid of 200 mm in diameter.

Please, read more about this in the next issue!

◆

Editorial: We publish the specification of several interesting patents, which were taken out by Russian inventors for recent years.

Russian patents on alternative energetics

- №93001754, 1996.06.27, Converter of gravitational forces into energy, V.V. Mironov and others
- №92008720, 1995.02.10, Slavic method to produce energy by conversion of gravitational forces, V.V. Mironov
- №93006696, 1995.04.30, Supporting converter of gravitational forces into energy, V.V. Mironov
- №2001121071, 2003.04.10, Method to create energy in liquid, heat-generator and heat-and-power device, A.Yu. Baurov and others
- №2132109, 1999.06.20, Generator of mechanical energy, A.Yu. Baurov and others
- №2001125794, 2003.05.10, Generator of static electricity, S.I. Danilov
- №2001120796, 2003.04.10, Electric power device, D.M. Beliy
- №98116077, 2000.05.10, Heat engine (monotherm, “perpetual mobile of second-type”), V.I. Likhachov and others
- №93048971, 1996.04.20, Float engine, I.P. Tchinarev
- №94023174, 1996.03.27, Gravitational-hydrostatic engine, I.P. Tchinarev
- №98112650, 2000.03.20, Water engine, V.A. Gylchuk
- №2001115953, 2003.03.20, Potential engine by Prokopenko, V.F. Prokopenko
- №2001115954, 2003.03.20, Potential amplifier by Prokopenko, V.F. Prokopenko
- №2001124364, 2003.04.20, Drum of gravitation (gravity), I.A. Strelnikov
- №95107181, 1997.02.20, Method to convert energy of physical vacuum into energy of physical space, O.K. Polev and others
- №92008695, 1995.04.20, Method to get overunity efficiency, Yu.V. Karasev and others
- №93006851, 1995.04.30, Principle, method and device for circulation of mass by force of gravitation and inertia, G.S. Kirichenko
- №93039885, 1996.05.20, Hydro-gravitational electric power station, L.S. Konkov
- №93036371, 1995.11.27, Method to convert energy created in result of interaction of earth’s gravity force and air bearing capacity into electric energy, V.I. Kovalenko
- №93025690, 1996.07.27, Method and device to convert heat energy into mechanical one, A.I. Lendyaev
- №2001122885, 2003.04.20, Rotary method to convert magnetic energy into mechanical one, L.N. Mezentsev
- №2001106783, 2003.02.10, Gravitational source of energy, A.D. Yurik and others
- №2001115968, 2003.03.20, Method to convert energy, V.S. Gorelyuh

- №2002131190, 2003.04.20, Method to convert heat energy into useful work, G.R. Umarov and others
- №2001122676, 2003.05.27, Device to get energy by conversion of energy from one state to another, T.A. Trushina and others
- №2131636, 1999.06.10, Fuel-less engine, V.E. Alekseenko
- №5037775, 1996.10.10, Magnetic engine, V.E. Alekseenko
- №94003087, 1995.12.27, Gravitational engine, A.A. Voznyukov
- №93058095, 1996.06.27, Perpetual mobile, Yu.S. Parshutin
- №93056706, 1996.07.27, Perpetual mobile, A.V. Remezovich
- №93039002, 1996.01.20, Gas-dynamic perpetual mobile, A.I. Lendyaev
- №94011000, 1995.11.27, Perpetual mobile by Makagyan, V.S. Makagyan
- №97114829, 1999.06.27, Perpetual mobile by Makagyan, V.S. Makagyan
- №95114965, 1997.08.10, Electric perpetual mobile by Abrakitov, V.E. Abrakitov
- №9512209, 1998.02.20, Perpetual mobile, V.V. Sharov
- №99111572, 1999.09.27, Perpetual mobile, N.M. Rotar
- №98116343, 2000.05.10, Engine-electric saw (perpetual mobile), V.V. Smirnov
- №2000101256, 2001.11.20, Perpetual mobile, V.D. Rykov
- №2000120592, 2002.07.20, Perpetual mobile, K.I. Harazov
- №2001117783, 2003.03.20, Magnetic perpetual mobile, S.N. Verevkin
- №2001123502, 2003.04.20, Magnetic engine, A.E. Ryumin
- №5032711, 1995.02.20, Ferromagnetic engine, V.P. Levkin
- №94019782, 1996.01.10, Device for demonstration of conversion of permanent magnet energy, M.F. Ostrikov and others
- №95103846, 1996.11.27, Magnetic generator, M.F. Ostrikov and others
- №95112010, 1997.06.27, Method to convert permanent magnet energy into rotary motion, A.V. Starostin and others
- №2143170, 1999.12.20, Method to convert energy of mechanical loading into energy of repeated transference, V.I. Rahovsky and others
- №2000119415, 2002.06.20, Method to convert energy of fields of permanent magnets into mechanical work, Yu.S. Pilipkov
- №2001109811, 2003.02.27, Method to capture and convert energy of physical vacuum into electric energy, S.N. Kovalev
- №99121374, 2001.11.10, Perpetual mobile, E.M. Volodin
- №94026259, 1996.05.20, Device to convert magnetic field energy into kinetic energy, P. Imrish

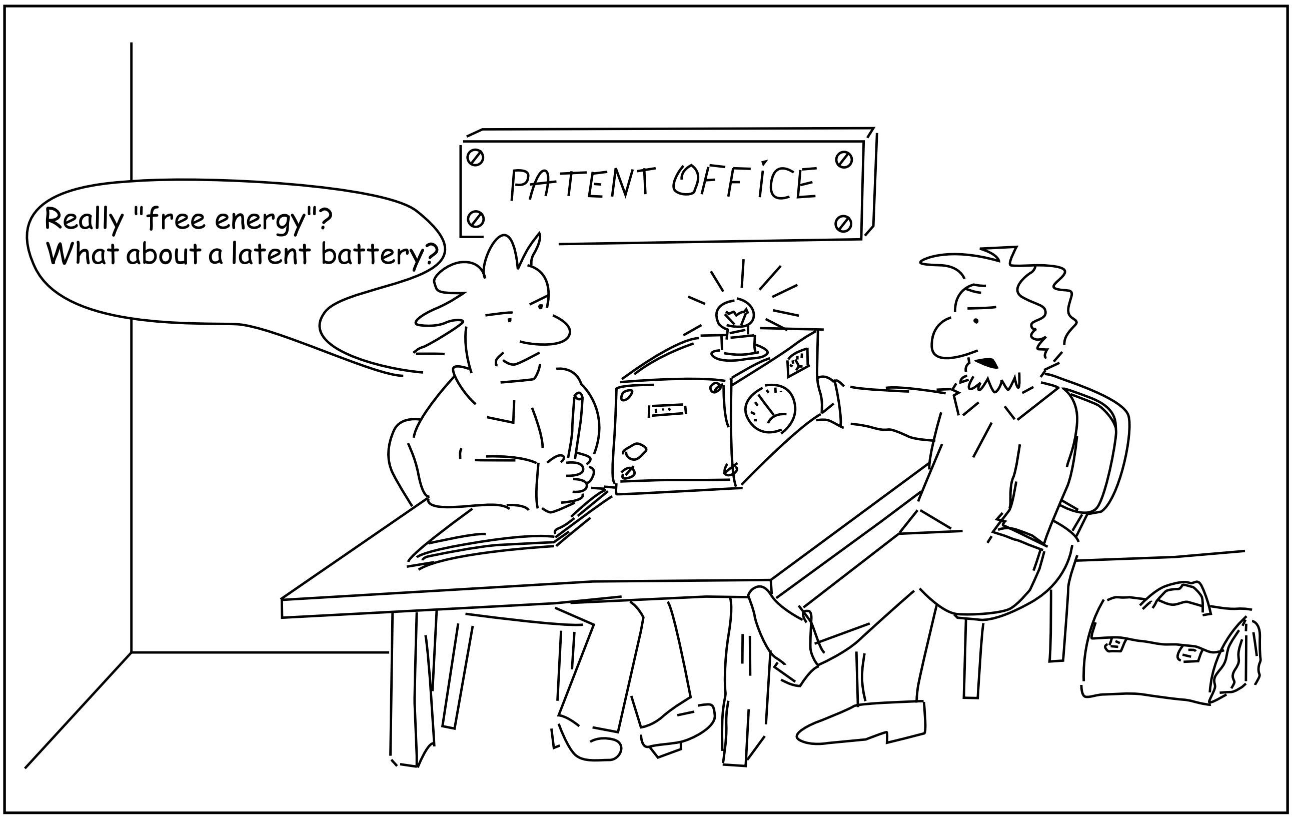
PATENT OFFICE
Really "free energy"?
What about a latent battery?

Device for conversion of magnetic field energy into kinetic energy

No. 94026259, 1996.05.20

Author: Pavel Imrish

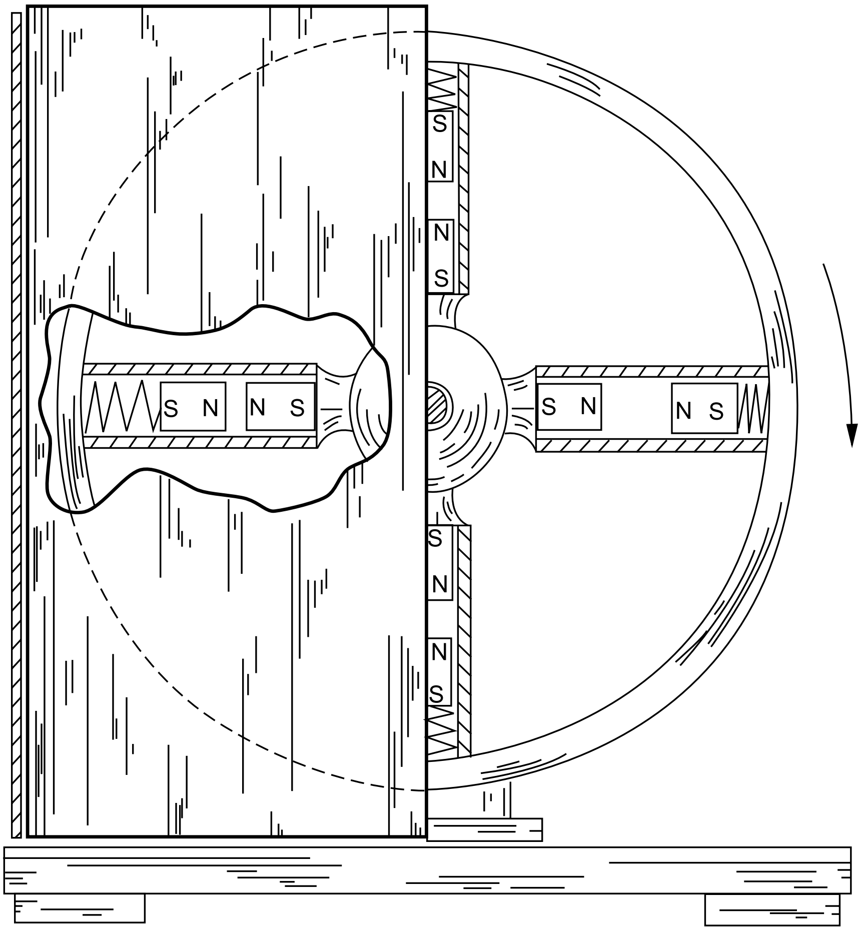

Fig. 1

The invention refers to a device for conversion of magnetic field energy into kinetic energy. The device has a rotating wheel with at least two radially opposing non-magnet holders used as guide ways for at least two permanent magnets, respectively. The permanent magnets are mounted in the holders so that there is a possibility to remove them. From the both sides of the wheel plates made of magnetic material are mounted axially; the plates surround the wheel partially. Depending on direction of the both poles, namely on location of the similar and opposite poles towards each other, compression element is situated on the holders, or between the wheel and the outer end of the permanent magnet located at the furthest radial distance from the axle of the wheel, or between two permanent magnets. (The device operation principle is obvious from Fig. 1).

Editor: We have found an example of a similar device whose description is presented in the book by R. Ford The Perpetual Motion mystery [1].

In this case, there is an analogous process, during which half of the cycle is screened and the rotor is supposed to rotate permanently (see Fig. 2).

1. R.A. Ford, The Perpetual Motion mystery. Lost Technology Series, p.23.

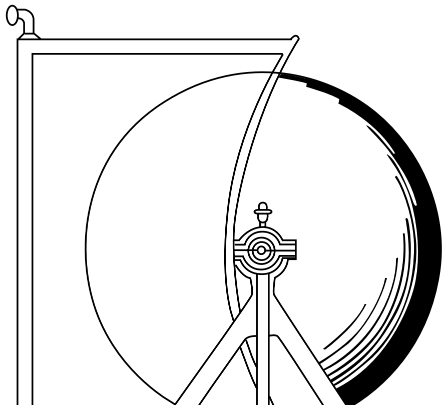

Fig. 2

Gravity spaceships

Review on Grigory R. Uspenskiy's works, Russia

http://space21.boom.ru/gravity.htm

The level of development of modern astronautics is primarily determined by the energy capabilities of space carriers. So, nowadays, the single-mission carrier rockets equipped with chemical engines allow space exploration within the solar system. These carriers have delimited the near-earth space in the form of a geostationary orbit, where the practical-purpose space complexes operate.

Newer and higher levels of quality will be reached by cosmonautics upon its mastering gravity power engineering. Flights towards the nearest stars will become possible and by the end of the next century this will be followed by travels throughout Galaxy.

Gravity engines are structurally simple (**two bodies of different density are rigidly connected**). To form some practicably significant value of thrust it is required to use absolutely new technology regarding creation and retaining of matter of high density which can be compared to that of atomic nuclei. For instance, at an engine mass of about 2 t, it is possible to attain 10 N thrust using bunches of aluminum and lead nuclei. The nuclei are approached to each other at a distance comparable with sizes of these bunches.

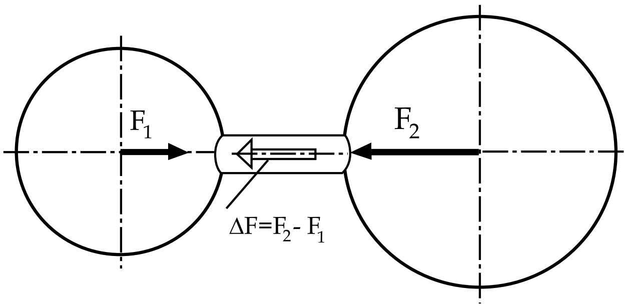

Fig. 1

Use of matter which is denser than nuclei bunch, increase of mass of the gravitating bodies, and reduction of the distance between the bodies causes increase of thrust. So, for a level of density of gravity holes (i.e. for extremely dense state of matter in which it is not attracted by other bodies) limit value of acceleration of the gravity engine reaches the order of 10^{100} ms^{-2}.

In engineering, we usually use units of distance as meters, centimeters, millimeters, and microns. Let us consider one of them – millimeters. With such a size of the gravitating bodies and distance between them, it is possible to obtain substantial thrust forces of the gravity engine and, accordingly, accelerations of its motion.

So, using an engine of 20 ton mass we will get the thrust of $2x10^4$ N, with mass of 200 ton we will get the thrust of $2x10^6$ N, and so on. Thus, by increasing the engine mass by one order of magnitude, we obtain increase of thrust by two orders, while sizes of the bodies remain unchanged. But acceleration, in this occasion, remains constant and equal to approximately 1 m/s^2.

Decrease in size of gravitating bodies of the engine leads to the corresponding increase of thrust by an order. Decrease of size is realized by increase of density of these bodies by 3 times up to 0.3 mm at the same masses. The acceleration they produce is increased by an order as well, i.e. up to 10 m/s^2. Decrease of sizes of the gravitating bodies and the distance between them down to 0.1 mm results in further increase of engine thrust by one more order and also increase of acceleration up to 100 m/s^2. If size of the bodies is 1 micron then acceleration increases up to 10^6 m/s^2.

It is apparent that even with up-to-date technological possibilities of miniaturization the gravity engine with masses of dozens and hundreds of tons is capable to create great thrust forces and form accelerations considerable enough to fly across the Galaxy. Therefore, the problem of creating a gravity engine is, mainly, the problem of producing and retaining the high-density matter. It will be possible after studying the mechanism of interaction between matter & gravitational substance followed by developing techniques for synthesizing of high-density, large-mass, and small-size substance from this matter.

It is possible to control the value of thrust of the gravity engine changing the distance between gravitating bodies. Thrust direction can be changed by turning the rigid connection of the gravitating bodies. The engine can be activated by bringing the gravitating masses together, and switching-off by moving them apart.

In a 24-hour period, such a ship can reach a speed of 10^6 m/s and cover a distance of about 10^{10} m. In a month, with a top speed of 3x 10^7 m/s the ship will cover the distance of 10^{13} m; in a year, with a speed of 4x 10^8 m/s the ship will cover a distance of 10^{15} m; in 10 years - 4x 10^9 m/s and 10^{17} m, respectively. Apparently, the ship with sufficient over-load can be used for flights within the solar system and for manned flights towards the nearest stars.

When transporting unbreakable and non-urgent cargoes, the acceleration can be increased up to a few dozens of unities. The flight duration might be also increased by several tens of years. This extends the area of utilizing the gravity ships with over-loads of several unities. Besides these ships can be used for transporting operations within near vicinities of the Galaxy.

The high-speed flight will demand protection against approach flow of matter. So, per second (with a ship speed of 10^{10} m/s) 10^{-12}kg of matter will approach the ship. At that density of interstellar material is 10^{-24} kg/m^3 and midship area is 10^2 m^2.

The great speed of approaching matter will create a substantial resisting force. At a speed of 10^{10} m/s, this force will come to the order of 10^{-2} N, whereas with a speed of 10^{18} m/s it will come to 10^{14} N. This is a great value, but as compared to thrust of such a ship which is to equal to 10^{17} N, the former value is a small one. That is why it is possible to overcome such resisting force.

Intercommunication with these ships might be, most likely, realizable by means of distortion of gravitational field. It is possible that the gravitational field distortion will be formed, on the contrary, by generating the matter from the gravitational field, and, most probably, by a method as yet unknown.

New books and DVD

Monograph **"General Etherodynamics. Modeling of matter structures and fields on the basis of conception of gas-like aether" by Vladimir A. Atsukovsky.** 2nd edition, M., Energoatomizdat, 2003; Brochure **"12 experiments on etherodynamics"** Zhukovsky, Publisher "Petit", 2003.

On questions about purchasing, please, refer to Email: atsuk@dart.ru

◆

Floyd Sweet Bench Test DVD

In order to demonstrate the reality of tapping free energy from the vacuum, we have now made available a very reasonably priced DVD of the bench tests of the late Floyd Sweet's Vacuum Triode Amplifier, with commentary by Tom Bearden.

During these two tests, the unit, which weighed about 6 lbs., can be seen to be putting out well over a million times more power than was put into it.

This DVD makes an excellent educational or instructional tool, and is priced at USD 9 including US domestic shipment. Running time is about 20 minutes.

Note: the full length Sweet DVD, which includes Sweet's proprietary conditioning process for the magnets, is also still available.

Order from http://www.cheniere.org/sales/order_by_credit_card.htm

On Velocity of Drive-Free Motion

S.A. Gerasimov , V.V. Stashenko, Russia

Physics Department, Rostov-on-Don State University,
Zorge St. 5, 344090, Rostov-on-Don, Russia

***Editorial:** We publish a summary of the article. The original text you can receive from the authors.*

A drive-free system made on the basis of a planar mechanism is described. The experimental results on average velocities of the drive-free propulsive system are presented in a scaled view.

During usual motion, a system is repealed from the surface in the case of its motion on a plane or from the medium when it moves in a resistance medium. This kind of motion is achieved by using a drive mechanism. In contrast to the usual motion, the drive-free propulsive system moves due to interaction of a body of a system with another body of the same system. Sometimes such machines are called inertioids [1] or vibration propulsive devices [2]. To produce such kind of motion it is sufficient to provide anisotropy of the resistance force [3-5] or asymmetry of the internal force [6]. One of the simplest ways to do it is to use a simple two-link planar mechanism. A device for such a motion is shown in Fig. 1.

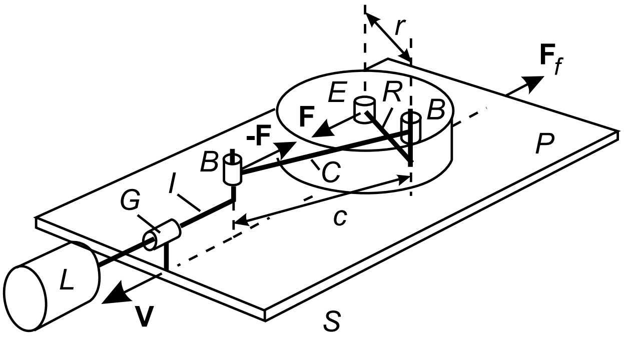

Fig. 1

Experimental device

This device with a total mass of M+m consists of a platform P on which an electric motor E rotating a drive rod R of length r is mounted. This is only one part of the planar mechanism. Another link is the connecting rod C of length c which connects the unbalanced load L and the drive rod R by means of two bell bearings B. The unbalanced load L of mass m on the rod I slides in guide G. The basic distinction of this mechanism is in that this variant of drive-free machine allows us to analyze the experimental data on average velocities of motion. Actually the only forces, which act on the platform and are collinear to the moving direction, are the force **F** caused by vibrations of the load L and the frictional force $\mathbf{F}_f$. The platform starts to move when the force acting on the platform becomes greater than the frictional force.

Let us discuss possibilities of the real propulsive system shown in Fig. 1. Measurements were made at r=0.01 m and at three mass ratios m/M. The values of frictional coefficients were measured for each parameter of the system and vary from k=0.25 to k=0.4. The system starts moving when theoretically the drift does not take place. When motion becomes reversible, the velocity of drive-free drift does not decrease.

Thus drive-free motion is possible even if the frictional coefficients are very small. When increasing the frequency of vibrations velocity does not decrease. We can not discuss the existence of some third force acting on the platform (we can suggest the existence of frictional force $\mathbf{F}_f$ and the internal force **F** caused by vibrations). Let somebody else take the liberty of making such a conclusion!

References

1. Tolchin V.N. Inertioid. // Perm: Perm Book Publisher. 1977. Blekhman I.I. Vibrational Motion. // Moscow: Nauka. 2000. Nappo F. Sulla Influenza Indiretta Delle Forze Inertne sul Moto del Baricentro. // La Ricerca Scientifica (Rendiconti). 1965. V. 8. No 1. P. 3-14.
2. Gerasimov S.A. Anisotropy of Frictional Force and Vibrational Motion. // Problems of Applied Physics. 2001. V 7. P. 85-88.
3. Gerasimov S.A. Self-Similarity of Vibrational Motion in a Resistant Medium. // Journal of Applied Mechanics and Technical Physics. 2002. V. 43. No 1. P. 90-92.
4. Gerasimov S.A. A Real Model of Vibratory-Impact Propulsive Device. // Problems of Applied Physics. 2000. V 6. P. 117-118.
5. Langhaar H. L. Dimensional Analysis and Theory of Models. // New York: Willey. 1951.

ELECTROMAGNETIC SELF-ACTION

S.A. Gerasimov, A.V. Volos, Russia

Physics Department, Rostov-on-Don State University,
Zorge St. 5, 344090, Rostov-on-Don, Russia

The torque of a self-action exerting on the mobile part of the reactionless engine is measured. It is shown that there exists a value of the height of the mobile electrode at which the torque of self-action is maximal.

There is an opinion that a body can not act on itself [1]. The existence of the so-called force of self-action is believed to contradict to the law of conservation of momentum. This is nothing but prejudice. Though the experimental confirmations of existence of self-action [2-4], conventional science is ready to refuse modern electrodynamics and make up a new theory of electromagnetism [1, 5] in order to save Newton's third law in use. It is comprehensible. First of all, these are the theoretical [5] and experimental [3] errors. The equivalence [5] of the Biot-Savart force is mathematically coarse since in this case the highest terms of expansion of force of self-action are cast out. This equivalence violates law of action and reaction and Ampere force law for which the principle of equality and collinearity of action and reaction forces is valid. Non-linearity of dependence of the torque of self-action on the current intensity [3] makes us doubt in the validity of these results. Moreover, typical values of force [2, 4] and torque [3, 6] are too small even if direct current in the circuit is significant. It is required to pass through the direct current of hundred amperes to produce the considerable displacement or turn of a body. As a result, small values of the effect caused ambiguity in the explanation of the phenomenon [1].

In fact, searching the reactionless propulsive devices we usually forget about the magnetic self-action which arises when a body consisting of a magnet and incomplete electric circuit can move violating law of action and reaction [7]. This violation is proved theoretically [8] but peculiarities of this motion are not well discussed. The principle of such a motion is shown in Fig. 1. This is the magnetic interaction between two current elements one of which is a part of a closed circular loop L and another is perpendicular to the first one. Since the force $d\mathbf{F}_{jm}$ acting on any current element of the loop L is perpendicular to the density of current $\mathbf{j}_m$ then the $\mathbf{Z}$-component of the torque $d\mathbf{N}_m=[\mathbf{r}_m \mathrm{x}\, d\mathbf{F}_{jm}]$ is equal to zero while torque which exerted on the second current element by the magnetic field of the loop, does not.

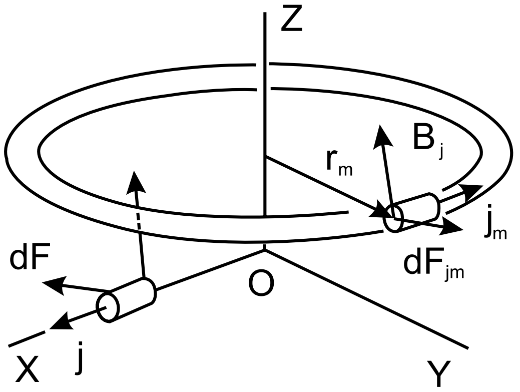

Fig. 1

Violation of law of action and reaction.
The force $d\mathbf{F}_{jm}$ with which the magnetic field $\mathbf{B}_j$ of another current element acts on the current of density $\mathbf{j}_m$, is not equal nor opposite to the force $d\mathbf{F}$ with which the magnetic field $\mathbf{B}$ of the first current element acts on the current of density $\mathbf{j}$.

An experimental device is shown in Fig. 2. A commercially available ring-shaped carbon-steel magnet M (inner radius – 20 mm; outer radius – 55 mm, height – 25 mm) of $2.2 \mathrm{x} 10^5$ A/m magnetization is located on the cover of a cylindrical electrode E suspended by thread T. The electrode E is furnished with a central electrode C of the same height h and 5 mm diameter. The radius of the electrode E is 130 mm. The open end of electrode E and the central electrode C are submersed in a conducting liquid L_+ and L_- placed in two volumes of a vessel V so that the depth of submergence of these parts is about 5 mm. In this experiment the conducting fluid is 10% solution of copper sulfate ($CuSO_4 5H_2O$). The vessel V consists of two thin coaxial plastic cylinders of 2 mm thickness. The radii of the plastic cylinders having the common bottom are equal to 10.5 cm and 18 cm. The height of the vessel is 12.5 cm. A disk and a ring-shaped electrodes S_+ and S_- of 1.5 mm thickness are placed on the bottom of the vessel to supply the direct electric current of intensity I. All conducting parts of the device are made of copper. The thickness of the electrode E is 1.5 mm.

When the direct current *I* of density **j** flows through the electric circuit, the magnet and the electrode *E* rotate in the direction of the magnetic force *d***F** which is proportional to the vector product [**j**x**B**] as shown in Fig. 2. This is a force of self-action by means of which the mobile part of this device consisting of the magnet *M* and the electrode *E* acts on itself. However there is no force of reaction which could cause such a rotation. This is really true since a ring-shaped magnet is equivalent to two cylindrical surfaces with the surface current of density $\mathbf{j}_m$. The force of self-action *d***F** acting on the cover of the electrode is compensated by the force of self-action $d\mathbf{F}_*$ acting on the cylindrical part of the electrode *E*. Magnetic induction field $\mathbf{B}_*$ "does its part" as shown in Fig. 2. Therefore, it would be appropriate to find out how the height of the electrode *E* influences the value of the torque **N**. Such an influence is demonstrated in Fig. 3.

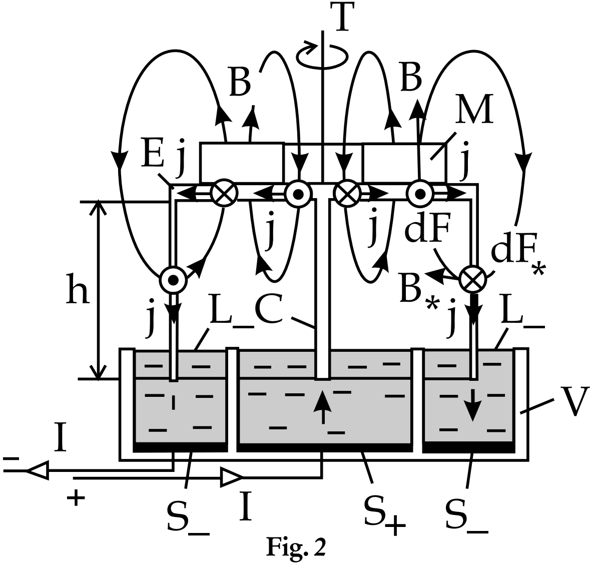

Fig. 2
Experimental device
(•) and (x) – directions of magnetic forces acting on various parts of the mobile electrode *E*

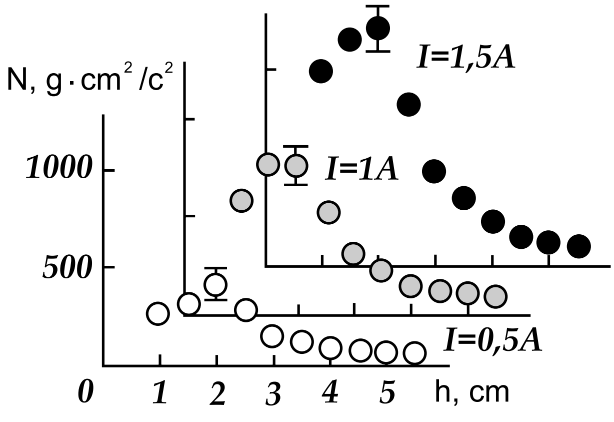

Fig. 3
Dependence of the torque *N* on the height of the electrode *h* at various values of the direct current *I* in the circuit

Thus there exists the value of the height *h* at which torque of self-action is maximal. This is the first result of the present work. At *I*= 1 A and *h*= 2 cm the electrode *E* turns at angle of 2.2 radian (126 °) with respect to the equilibrium position. Such a turn corresponds to $3.4\text{x}10^2$ $\text{gxcm}^2/\text{s}^2\text{rad}$, i.e to the constant of torsion of the thread. The second result is that the maximum value of the torque of self-action *N* at the current *I*= 1 A is even larger than the value of the torque produced by the traditional unipolar device [3] at the current 50 A.

In fact, this work represents an attempt to draw attention to a more effective and simple practical application of the self-action. It remains only to transfer electric current in an unclosed electrical conductor and such a possibility really exists.

References

1. Graneau N. The Finite Size of the Metallic Current Element. // Physics Letters A. 1990. V. 147. N. 2-3. P. 92-95.
2. Sigalov R.G., Shapovalova T.I., Karimov H.H., Samsonov N.I. Magnetic Fields and Their New Applications. // Moscow: Nauka. 1976.
3. Das Gupta A.K. Unipolar Machines. Association of the Magnetic Field with the Field-Producing Magnet. // American Journal of Physics. 1963. V. 31. N 6. P. 428-430.
4. Cavalleri G., Bettoni G., Tonni E., Spavieri G. Experimental Proof of Standard Electrodynamics by Measuring the Self-Force on a Part of a Current Loop. // Physical Review E. 1998. V. 58. N 2. P. 2505-2517.
5. Christodoulides C. Equivalence of the Ampere and Biot-Savart Force Law in Magnetostatics. // Journal of Physics A. 1987. V. 20. N 8. P. 2037-2042.
6. Serra-Valls A., Gago-Bousquet G. Conducting Spiral as an Acyclic or Unipolar Machine. // American Journal of Physics. 1970. V. 38. N 11. P. 1273-1276.
7. Gerasimov S.A., Volos A.V. On Motion of Magnet in Conducting Fluid. // Problems of Applied Physics. 2001. V. 7. P. 26-27.
8. Gerasimov S.A. Self-Interaction and Vector Potential in Magnetostatics. // Physica Scripta. 1997. V. 56. N 3-4. P. 462-464.

JNET

New Energy Technologies
collection of articles
Japanese Version

Collection of articles from New Energy Technologies magazine, 2001-2002 is published in Japan.

On purchasing of the book, please, contact

Eiichi Yamamoto,
President Yama Trans Co. Ltd.

Email: admin@yamatrans.co.jp or
eyama@yamatrans.co.jp

Uranium Photoaccumulator

Anton N. Yegorov, Russia

St.Petersburg Institute of Nuclear Physics, Russian Academy of Sciences, Gatchina, 188350, Leningrad region, Russia
http://www.pnpi.spb.ru, Email: bti@isppd.pnpi.nw.ru

Photoaccumulators are galvanic elements consisting of two communicating vessels, one of which is exposed to light and another is kept in darkness. Ionic composition of electrolyte, which fills these vessels, depends on intensity of illumination. If an inexpensive electrode is placed in each vessel, then a potential difference will appear. If the electrodes are closed at external working resistance the electron stream will begin to equalize ionic composition and the difference in illumination intensity will tend to disbalance it. As a result, a certain working mode will be established which will allow to utilize the photoaccumulator for production of industrial electrical energy. As distinct from the solar battery, which works while exposed to the light, the photoaccumulator stores some part of solar energy in its electrolyte which is supplied with electrical energy after sunset too.

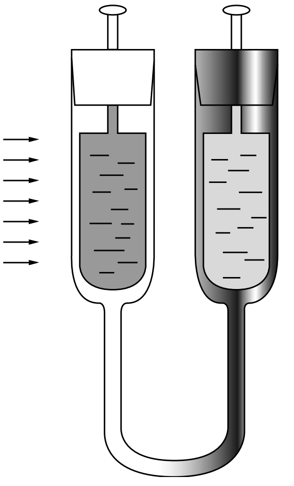

The idea of photoaccumulator is not a new one. It was established in the end of 19th century that it is feasible to create a photoaccumulator containing electrolyte made of the following mixture: $Fe^{2+} + Hg^{2+} = Fe^{3+} + Hg^{+}$.

This photoaccumulator was not adopted in practice because of high toxicity of the electrolyte and very low EMF (approximately 0.018). Development of technology is accompanied by a widening range of materials and matters which can be used in large scale production and power industry. Among such materials are, in particular, titanium, depleted uranium and rare-earth elements. Nuclear power industry uses U235 isotope only. Naturally occurring compound contains only 0.72% of U235. Hundreds of thousands of U238 from which U235 has been extracted are kept as useless stock.

In the end of sixties in one of radiochemical laboratories of LINP (Leningrad Institute of Nuclear Physics) P.N. Moskalyov repeated the same experiment over a long period of time. In the morning he placed a tightly closed retort filled with uranyl-chloride soluted in compound of water, spirit and hydrochloric acid on the window-sill. The solution in the retort changed its color from yellow to emerald-green. Before leaving the laboratory Moskalyov removed the retort with green solution from the windowsill to a closed chest. In the morning Moskalyov retrieved the retort with the solution (which was already yellow) from the chest and placed it on the sill. The solution became green again and the whole process repeated daily.

Chemical processes in uranium electrolyte are rather complicated: first, under the light influence the uranyl ion oxidizes ethyl alcohol, which is then transformed to aldehyde: $UO_2^{+2} + C_2H_5OH^{+} = U^{iv}(OH)_2 + C_2H_4O$.

In the dark ions of urany-4 interact with aldehyde and are transformed to UO_2 uranyl ion and spirit again. Thus, **this cyclic process can take place over a significant period of time being supported by the energy provided by light quanta of visible light which are transformed into the heat.**

The idea of utilizing this process in order to obtain electrical energy emerged in the late nineties. The mechanism of uranium photoaccumulator is very simple (Fig. 1): two glass vessels, one made of transparent glass and another of non-transparent glass, are connected with a black acid-proof rubber tube. The communicating vessels are filled with aqueous electrolyte containing 40% of spirit, 80-100 g/l UO_2Cl_2 and 0.2 m HCL. The electrolyte is previously activated by multiple cycles of exposing it to light and then placing it in darkness in order to obtain a significant volume of aldehydic fraction. The major difficulty of designing uranium photoaccumulator is connected with the choice of the material for electrodes. In galvanic elements carbonized platinum electrodes are usually used. For practical use platinized titanium gives the best fit. Titanium foil covered with superfine layer of platinum is currently widely used in electrochemical industry and is procurable.

In the transparent vessel the titanium petal-shaped electrodes are parallel the light flow, so the light quanta move freely through the solution.

In working condition each photoaccumulator produces 10 mA of current at 0.3 volt voltage, so they are to be connected in large batteries, both in series and in parallel.

It is also necessary to consider the aspect of safety. Until now, the depleted uranium has been used for one purpose only, namely, for manufacturing of cores for armor-piercing shells. When used in this manner, the uranium is pulverized, which leads to environmental damage. The battery of photoaccumulators is placed in hermetic metal chest; the front panel of the latter must be made of solid safety glass. Such a system will protect the photoaccumulator battery from malicious intent and the security staff from weak uranium radioactivity.

International Prize on Power Engineering “GLOBAL ENERGY”

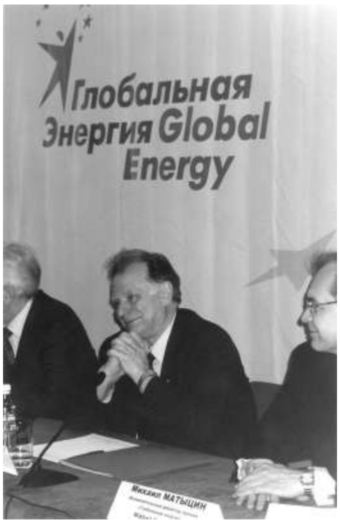

Zhores I. Alferov

International Prize “Global Energy” was first presented in St. Petersburg on June 15, 2003. The bonus fund of $900,000 was shared between three scientists “whose work has been and is still important both to the last century and future research”, said Zhores Alferov, Chairman of the Global Energy Prize Award International Committee.

Nick Holonyak, Professor at the University of Illinois (USA), was awarded for “fundamental contributions to the development of power silicon electronics and invention of the first semi-conducting light-emitting diodes in a visible part of the spectrum”. Thyristor invented by Holonyak allows the transformation of direct current into alternating and vice versa, to direct this process, to store and accumulate energy. This device is used for every electric locomotive, for frequency transformation. Such devices help to save significant amounts of energy, thus, about 30 percent of all energy produced in the world is processed through thyristors.

Глобальная Энергия Global Energy

Nick Holonyak’s other invention, i.e. semi-conducting light-emitting diodes in a visible part of the spectrum, has led to development of a new field of the modern power engineering. Supposedly, efficient and safe light-emitting diodes will replace daylight lamps in the next ten years.

“Although Nick Holonyak’s invention dates back to the end of the 50ies of the last century, it remains current” said the Chairman of Expert Commission, Academician Vladimir Fortov. “This is one of the key inventions leading to the search of new ways of sustainable energy saving”.

Diploma of Global Energy Prize laureate

Gennady Mesyats, Academician of Russian Academy of Sciences, and **Ian Douglas Smith**, Senior Scientist at The Titan Pulse Science Division, were awarded for giving birth to a new direction in the field of power engineering, i.e. power pulse energy. Due to their research, commutation of high-level currents at megavatt levels of voltage has become possible; a number of powerful current choppers have been created, futhermore, there have been developed transformer constructions which allow avoiding energy loss in power transmission lines. These research and innovations resulted in the creation of a number of unique pulse power machines and devices widely used in Russia and abroad (“Aurora”, “Helia” (USA), “Sinus”, “Gamma”, “Maus”, “Pik” (Russia) and others).

Teleportation

Review prepared by correspondent Alla Pashova, Russia

For conventional science, the term "teleportation" is not worthy of serious consideration unless ascertained by compulsory "quantum" teleportation. Thus, teleportation is unconditionally referred to microworld phenomena and, in fact, comes to distant information transfer. Spatial transference of a macroobject of definite mass has been excluded.

Teleportation of states

What does **quantum teleportation** mean? During active development of quantum theory, in 1935, the so-called EPR-paradox (Einstein-Podolskiy-Rozen paradox) was formulated in the well-known work "Can quantum-mechanical description of reality be full?" written by Albert Einstein, Boris Podolskiy, and Natan Rozen.

The gist of the paradox is as follows. There are two particles interacting for some time, thereby forming a common system. From the position of quantum mechanics, this coupled system can be described with a certain wave function. When interaction is over and particles scatter within arbitrarily large distances, they still will be described by the same function as before. At that, state of each separate particle cannot be known in principle that is apparent from uncertainty relation. Only when one of the particles enters a receiver, which records its parameters, the relevant characteristics of the other one emerge (exactly emerge, but not become known!). Thus instant unlimitedly distant "transmission" of quantum state of the particle is possible. Therewith, teleportation of the particle itself and transference of mass do not take place.

Einstein and his colleagues believed that existence of such particles predicted by quantum mechanics prove theory incompleteness. Thereof, the scientists inferred the necessity of other parameters (besides wave function) to describe quantum states. Otherwise, from the local viewpoint, correlations between elements of such a system could not be understood. It was far much later, when Bell showed that some of measurements could define these correlations and exclude any local hidden parameters. It was not until early 1980-s that famous experiments were performed finally eliminating a possibility of local hidden parameters.

In 1980, Alan Aspect experimentally proved that EPR-paradox in quantum world takes place indeed. Special measurements of state of EPR-particles indicated that EPR-pair not only has a common origin, but in addition, that one of the photons somehow "get to know" the way the second one was changed. In further experiments, existence of EPR-paradox was affirmed, even if particles of EPR-pair were removed from each other over a distance of 10 kilometers or so.

In 1993, Charles H. Bennett and his colleagues worked out a method to transfer the quantum state of some object of the microworld to another quantum object by means of EPR-pair and called this method "quantum teleportation". In 1997 a group of experimentalists under the direction of Anton Zeilinger for the first time implemented quantum teleportation of the photon state in the University of Innsbruck.

In such a way, researchers keep on improving the process of quantum teleportation. In 2001 Danish scientists managed to link gas particles spaced at a substantial interval from each other, by transmitting information about quantum state from one particle to another by means of laser. The quantum teleportation between two gaseous clouds were attained by Eugene Polzik and his colleagues in Orkhus University. They succeeded in coupling about million of cesium atoms, whereas the previous record was only four atoms.

Scientists of Australian National University destroyed a laser beam and nearly instantly recreated it in another point in space; in other words, they teleported photons of the laser beam. In contrast to previous similar experiments, the physicists managed to obtain the required result in 100 percents of cases. Ping Koy Lam, the head of the task group of Australian University, claimed that the first atom of solid substance was likely to be teleported within near three to five years. However, as most scientists admit, a task to teleport a human remains almost impracticable. Even teleportation of atoms, as compared to that of photons, is much more complicated process. It is even harder when dealing with molecules. It is basically possible (though practically very difficult) first to transfer a molecule to a minimum-energy state (ground state) causing it to radiate a certain sequence of photons. These photons will find themselves in a certain superposition containing all the "quantum" information, which was available in the molecule. Thereafter, it is possible to teleport photon states by means of EPR-pairs. Moreover, it is also required that the classically measured information about a molecule. If molecule of minimum-energy state is present in the receiver, then this molecule, by interacting with teleported photons in a required order, will transform to the quantum state identical to that of the initial one. Consequently, the quantum state of molecule of a certain material will be transmitted, actually with velocity of light. When that happens, the quantum state at the transmitting side will be destroyed.

The human organism comprises about 10^{27} atoms. To save and transmit information on properties of that number of particles seems to be practically unachievable. "Theoretically, nothing prevents us from doing that, but complexity of the problem is such that now no one seriously thinks about the solution" – states Ping Koy Lam.

Being a method of information transmission, quantum teleportation has found its application in quantum computers, whose information is stored in the form of a set of quantum states. Impossibility to wiretap and copy transmitted information is considered to be an advantage of such computers. Those researchers, who, nevertheless, wish to answer the question of "How to teleport matter, but not its state?" have to seek for more perspective theories and techniques.

Teleportation of material objects

Those people, who really want to realize instant spatial transportation of objects, i.e. teleportation, should refer to studying properties of Space and Time. Quantum teleportation has a certain finite velocity that cannot exceed that of light. The genuine teleportation assumes that an object should set off from a starting point to a finishing point (these points differ by a certain distance X), at that the transference time comes to zero. The object to be teleported is not changed or taken to atoms to be gathered later at a distant point of space according to information transferred to this point. (Fig.1).

The object disappears from one place and simultaneously appears in another place. How is that possible? A body will disappear from point A and appear in point B if to bend space in such a way as to let point A and point B coincide. Then the object will instantly appear in point B since there is no interval between points A and B. Teleportation could be realized by a device, which would make it possible to superpose points A and B.

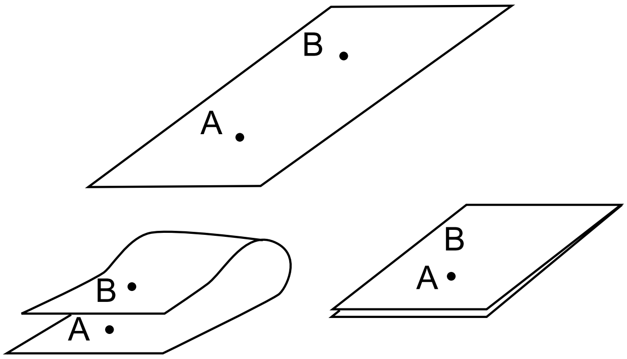

Fig. 1

Space is supposed to be unbendable without enormous energy consumption. However, Alexander V. Frolov points out that this issue is not so unambiguous: "Normally we consider space in connection with "natural" course of time existing in it. That is some degree of space curvature. Let us assume that it can be changed. The space curvature can be considered as acceleration or deceleration of time. Acceleration always demands, for instance in mechanics, some energy consumption. But if we "decelerate" time, energy is liberated and it can be accumulated for subsequent use".

K.Z. Leshan suggests surrounding a transferred object with a closed surface consisting of vacuum holes. Inside such a "hole sphere" there is geometry similar to that of a black hole. This place is absolutely isolated from the external Universe. No radiation is able to penetrate through the hole in space and time. For an observer, who is inside the hole sphere, the distance between sphere center and its border is infinitely large, since spatial metrics is ever-varying from the center to the border. Distances between the points is continuously shortened so that the distance between any couple of points comes to zero at the very hole surface. A matter transmitter can have inner or outer hole surface. In the first case a teleportation station has a spherical chamber to place the object. Equipment to produce holes should be located upon the outer side of the sphere. Such a station is capable to provide hundreds of launches a day by instantly expulsing spaceships to deep space over distances of millions of light years.

From microcosm towards macrocosm

Simeon Bocharov (member of Chemistry and Biochemistry Department, University of Delaver, Newark, USA) considers teleportation phenomenon using microobjects as an example and applying an interesting conception of protomatter.

In S. Bocharov's opinion, many paradoxes of the present-day science could be solved under condition of considering the whole existing matter to be a single continuum, i.e. protomatter. At this approach, the microcosm objects are not independent ones, being represented in the form of distortions of the very continuum and its manifestations here and now. Protomatter, whose distortion degree is beyond modern possibilities of detection, corresponds to vacuum. In the present-day paradigm redistribution of protomatter distortion corresponds to motion of particles. The important peculiarity is a refusal of such concepts as structure, dimensions, mass and other macroscopic characteristics with respect to microcosm objects, since here they are not considered as separate entities.

For experimental justification of his theory, S. Bocharov suggests to consider teleportation of microobjects incorporated in fullerens under low temperatures and pressure.

As a basis for reasoning he takes the fact that identical or similar conditions, whereon distortions are located, bring them to the state of identical or similar by lability/diffuseness. And vice versa, distortions, which are similar by lability/diffuseness, influence their near surroundings, creating, as a result, similar macroscopic states. In such a case protomatter distortion (in some spatial domains and under similar conditions) will cause appearance of similar distortions in another domain, whose degree of removal depends on presence of other distortions and their characteristics. In other words matter will be teleported from one spatial domain to another.

Possibility to realize teleportation of microobject causes the well-known phenomenon, which is nowadays described as embedding of particles into inner hollows of spheroid fulleren molecules without reacting with them. As object of teleportation there can be used distortions classified as elementary ones, such as hydrogen, helium or stable matters (noble gases). In both cases the influence of objects upon surroundings is minimized.

The researcher believes that in case of successful experiment, new pattern of matter structure will be confirmed, and valuable teleportation of microobjects (with prospects of such possibility for macrocosm objects) will be demonstrated.

One more approach to teleportation of physical macroobjects was considered in the work "Practical application of time rate control" (New Energy Technologies No. 3, 2001) by Alexander V. Frolov. It is assumed that density of space (aether) energy determines rate of passing of any processes including the very process of matter existence. Changes of aether density (increase or decrease) must result in the emergence of a force analogous to that of buoyancy, though acting towards the fourth dimension. This "chrono-motive force" (CMF) is also an analogue of electromotive force (EMF) and can be generated by analogy with electrodynamics. According to Alexander V. Frolov, laws of quantum mechanics as to discretization of levels of energy of material system, which exist in the domain of increased or decreased aether density, are valid in the macrocosm as well. Teleportation (as a transition from one state of the system to another one) can be studied with electron transference from one orbit to another as an example with the only difference that for the teleported object not only its location, but also the very spatial properties are changed. In aether of different density (after discrete transition) the same object will have different space around it, wherein time is decelerated or accelerated. Experiments of this field have already been in development stage.

Therefore, modern theoretical physics has handed the problem of teleportation to researchers-experimentalists, who possess sufficient breadth of mind. Perhaps, using teleportation technologies, they will manage to fundamentally alter the process of space exploration and raise our civilization up to radically new development level.

ELECTRIFYING TIMES

an online and published magazine about Electric, Hybrid, Fuel Cell Vehicles, advanced batteries, ultra capacitors,
fuel cells, microturbines, free energy systems, events and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908 fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

"Mass defect" in Home Conditions

Prokofy V. Sherbak, Russia

Mitinskaya Str., 40-1-244, Moscow, 123430, Russia
Tel.: (095)950-4135
Email: Prokofy@mail.ru

The idea to conduct experiments on "mass defect" demonstration appeared after a cycle of seminars, which took place in1993-1996 at the Moscow Aviation Institute under the direction of V. I. Patrushev, Doctor of Technical Sciences, Professor of "Designing of aerohydrodynamic systems" Department (the author of the article is also a graduate of the Department). At the seminars, an enterprising group studied and developed an applied theory of "Displacing fields" by Alexis V. Murlikin and an associate "Theory of energy exchange processes". Later the group was called **"Group studying non-inertial transference"** (non-inertial natural processes).

The experiment was aimed to test some theoretical suppositions, namely, that it is possible to influence on existent energy flows which input and output to the matter. At that the simplest way to influence indirectly is to change the medium the flows come through. In general, optic experiments on luminous flux refraction show the same. But light refraction and change of speed of light propagation at passing through different mediums is one thing; and change of mass of a material body is something new that has not ever been dealt with. Moreover you can consider light (i.e. electromagnetic radiation) to be the secondary manifestation of output energy flow. Besides mass defect is very important for nuclear physics because it "helps" to develop this branch and stimulates drawing "energy dividends" in splitting reactions, decay reactions and nuclear fusion reactions. But unfortunately we can not touch element nucleus and even million nuclei while we can easily touch metallic balls of 1 gram weight.

Let us note that a concept of mass is one of the most uncertain in physics. We can not measure mass directly. It is possible to do it only indirectly, by gravity, using scales (it does not matter what kind of scales are used), or by kinetic momentum (in this case mass is a measure of inertia) by means of dynamometers.

Thus it is evident that change of WEIGHT shown by scales is not MASS change as measure of matter quantity. Since one of the most important points of the theory of "Displacing fields" is creation of a vehicle of new generation, then one of the major tasks is to overcome gravity (to decrease weight). And this "home experiment" shows the way to solve it. However we can call it "home" only in part, because for weighing there were used electronic scales (mass-comparator with measurement accuracy of 0.1 microgram. (See Fig. 1).

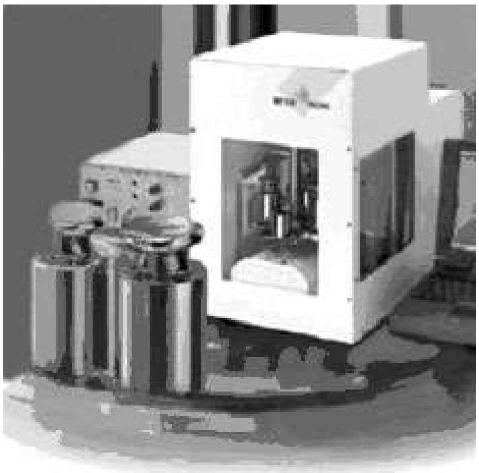

Fig. 1

Precision scales METTLER TOLEDO allows making:
High-accuracy weighing in the range
from 0.1 microgram to several tones
Measurement with resolution up to milliard points

Two sets were designed for weighing. The first one had an aluminum (Al) ball-kernel which weighed 1 gram and a lead (Pb) casing-shell consisted of two parts (half-spheres) which also has weight of 1 gram. The second set had on the contrary a lead kernel and an aluminum casing. The sets were disassembled and then hermetically assembled to make a single device with no gaps (See Fig. 2).

Since it had to be done with great accuracy and it was a piece article, not a Moscow product company undertook to produce it. And only one of the students, Nicholas Sorokin, a young boy with skilful fingers, had helped. He invented a technology which allowed making balls of the necessary size at home conditions. He was literally pickling every micron of metal with acid and finally got the necessary size and weight.

So, the weight of a non-assembled set was: 1 gram (weight of kernel) and 1 gramm (weight of shell). Weight of an assembled set, according to the "Theory of energy exchange processes" by Murlikin, is not equal to the simple sum of the non-assembled components, and mass defect would be different for various sets right up to the change of sign.

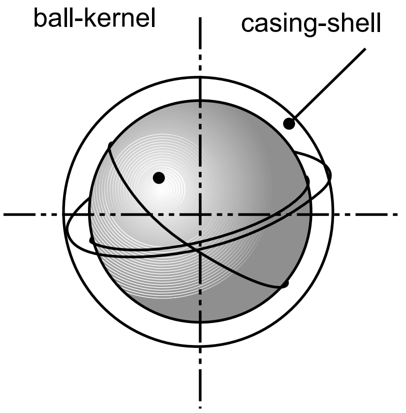

Fig. 2

Set of experimental balls

More than 50 experiments was carried out on weighing for every set. Assembled sets as well as non-assembled sets were weighed. **It was confirmed that there was an interaction between input and output flows for MATTER MAINTENANCE. It became apparent in the form of mass defect. Defect appeared in the sixth sign, i.e. it comes to the range from 0.0002% to 0.0007% of "kernel" mass (or "shell" mass). At that the lesser mass defect corresponds to aluminum shell, while the greater one corresponds to lead shell. Accuracy of weight measuring by the mass comparator exceeded the acquired result by 10 times. So we could not tell about inaccuracy of measurements.** So we had to apply to the "Russian center of testing and certification" which was very famous all over the world (Rostest-Moscow).

In general that is all, we have got an answer for ourselves, and the period of testing is behind. We have to advance; very much is still to be done. In addition let us mention one more surprise we got during the experiments. The matter concerns some time delay in receiving final results of every weighing. A qualitative picture of it is represented in Fig. 3. We did not expect that this effect would appear at such a rough level of measurement of energy exchange processes.

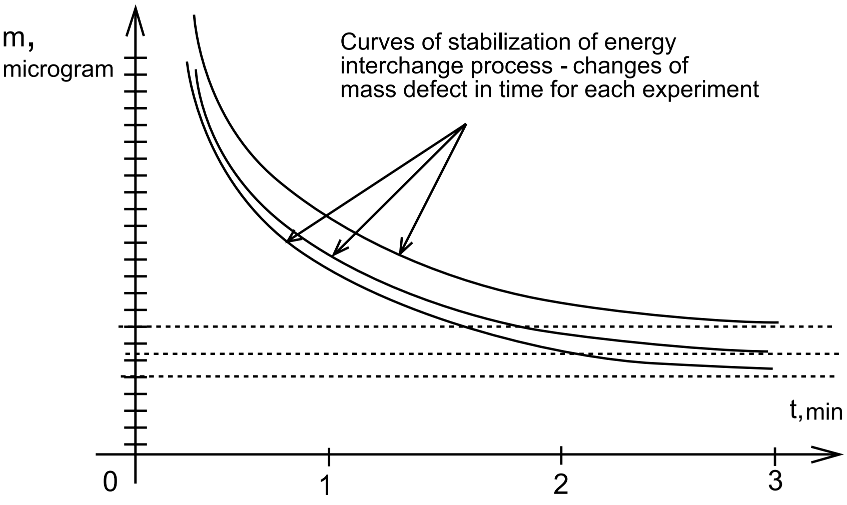

Fig. 3

Time stabilization of the process for receiving final weight

The subsequent report and detailed discussion of the experimental results by specialists (including physicists of various specializations) resulted in an unexpected non-recognition of facts and distrust. The main argument was a demand to increase the number of experiments up to 5-10 hundreds. And only then scientists would be ready to accept facts. It was quite a predictable reaction for the orthodox science in the middle of 90s. It remains to hope that if the discussion took place today it would have a paved way. However it is not a fact.

In conclusion let us note that this experiment on revealing mass defect ranks with famous experiments, in which rotating gyroscopes also demonstrate mass defect (decrease of weight) and even one of higher order than in our experiment. But in return we have no rotation, no chemical and nuclear reactions, while mass defect is presented.

Let us add that it is possible to complicate the task: to make a double casing and to choose other materials for "kernel" and "shell". (See Fig. 2). In our case low prices and availability made us use Al and Pb. The results will be more interesting, for example, with Li and Os (and with other supertransuranic elements).

Aether as Unified Field

Alexander M. Mishin, Russia

Planernaya Str., 79-208, St.Petersburg, 197373 Russia
E-mail: samish@mail.ru

Based on the original observations and experiments, the article gives more precise definitions of aether model, than which was previously elaborated. General laws of the Universe aetherodynamic are considered. There is described a new physical reality in the form of quasi-real aetherodynamic structures ("topological harmonics") generated by dynamics of the substantial world.

At the present-day stage of physics development words of Maugham, an English philosopher, take on particular topicality: "The great verities are too important to be new". Indeed, we are increasingly making certain that our remote ancestors were familiar with fundamental laws of the Universe; furthermore, they knew these better than we do. But which methodology of cognition of Nature and what kind of information sources did the ancient scientists have? Today, we can openly say: they had not ordinary experience only, but also esoteric, pretersensual one, without feeling any necessity to be divided into materialists and idealists. For more details on evolution of physical ideas, refer to work [1].

In discovering all over again, I also had to elaborate a "cellular taction" strategy, which enables the most important characteristics of aetherodynamic processes to be sensually perceived and evaluated. The main thing is that all my sensory observations are definitely reproduced at laboratory experiments while the simplest physical equipment is used.

We may get reliable knowledge about Nature from Nature itself only. Thus foundation of classical physics describing usual substantial world is also based on empirical facts studied by Galilei, Newton, Faraday and many other physicists-natural philosophers. I also would like to defend great Newton against attacks on his adherence to inductive methods which really brought imperishable values into science.

A new convolution of cognition is possible on condition of introduction into science of new concepts based on empirical study of physical entities previously unknown. Thereupon, I share the thought of Elena Ventzel, professional mathematician: "Mathematical methods are not useful, but harmful until phenomenon is not studied at a pre-mathematical, humanitarian level". It is ideological and logical language that may be used today to discuss the aether problem, whose solution calls for comprehension of natural-science facts, which are beyond traditional notions settled over centuries.

The long-term investigations have confirmed that exhaustive definition of a fundamental natural entity was given by Rene Descartes: "Universe does not contain anything but aether and its vortexes". This is not merely a postulate; this is an objective summary of empirical observations of Nature phenomena. In any academical model of microcosm there is no answer to the question: what kind of substance are electrons, quarks, strings made of? That is the very substance our forefathers called aether. Aether is unified field indeed; and it appears that for nearly his whole life Einstein had been searching for something that was under his foot.

It must be recognized that there is no undistorted, quiescent aether in Nature; as well as no zero-size vortexes exist. Therefore, in a customary sense, the aether concept is pushed behind transcendental horizon and turned into an inconceivable entity whose mystique is tantamount to religious ideas. However, all quantum vortex-wave distortions of the superfluid three-dimensional aether, which have an appearance of tori, threads, solitons and occupy spatio-temporal spectrum from zero to inter-universe scale, are material and contain energy under scientific interpretation. Any spaces such as turbulent fields, physical vacuum, gravitational fields and, of course, matter are material. All of them are etherodynamic structures of the real world. Spatial hierarchy of aether vortices is referred to as quantum stairs, fractal structure or "matreshka" [2-5].

To understand basic laws of aetherodynamics it is necessary to give a general definition for classical matter and other aether states. As a first approximation, let us give the name of matter to aether microscopic **left-screw** self-oscillating vortex-wave systems (elementary particles), whose existence and parameters are conditioned by higher hierarchical systems. This is our displayed world, with its electromagnetic fields and fundamental constants. Let us refer to vortex-wave structures of any vortex size, which in the considered space domain are **balanced** by a number of left- and right-screw vortices corresponding to matter and antimatter with prefix "quasi" to a free aether. Stable vortex lattices with alternating direction of helicity can be also considered in this case. Under natural conditions everything but classical matter or electromagnetic and gravitational fields can be related to the free aether. In the first place, these are the so-called physical vacuum and vortex fields, which attend gravitation. They represent anisotropic nonlinear subspaces and occupy definite places on spatio-temporal spectrum of the world aether. Particular free subspace creates an aggregate of topological harmonics [5] attending all the physical processes. Satellite of the gravitational field is distinguished because it belongs to specific class of stable vortex lattices.

Apparently, in general case, energy of free aether is virtual for our world due to its helical symmetry and quasi-matter, from the view of difference in sizes of typical vortices which correspond to protomatter. But this is not the entire point. There is also superfluidity of aether, whose vortex viscosity has marvellous properties. As a result, there are created phenomena of inertia and energy-information barrier separating material world from constant influence of quasi-real structures or turbulated space.

Classical inertia, as vortex viscosity, arises in the microscopic and more fine-structure free aether (in Newtonian space) and ever regularly reveals itself during acceleration of material bodies. However everything is different at stage of macrovortex aether, where inertia has properties of adaptation. Here it is affirmed that inertia phenomenon is "spread" throughout the spatial-temporal spectrum of vortex distortions of aether field. At that nature of inertia is changed. Within topoharmonic subspace in laboratory scales, the inertia law "diminishes" in time according to exponential dependence, whose time constant is determined by specific conditions and varied over wide limits (from seconds to many days). On the expiry of this time, inertial intercoupling transforms to the law of flicker effect. It may be said that flicker effect specify penetrability of energy and information barrier after expiry of time of paraphysical regular phenomenon [6]. Consequently, most of time the interaction with free macroaether is of random character and should be evaluated by mean observation.

To extend conceptual scope of our investigation it may be added that in classical hydrodynamics a liquid, along with ordinary inertia, has determined vortex macroinertia, which is created by vortices of the liquid itself. Therefore, it is necessary to keep on comparing processes at all levels of the unified field of aether. For instance, electric inductance is a kind of specific inertia, and topological harmonics of electromagnetic and other phenomena are subjected to the energy-information barrier (under laboratory conditions).

In the above arguments, the principle of relativity is latently and unconsciously presented. However, this complicated question needs to be separately considered. Physical nature of the energy-information barrier and reason for dependence of the inertia properties upon the size of vortices of free aether remain unclear, too. However, it can be supposed that classical elementary particles, as distinct from protomatter, have self-oscillating properties and are commensurable with vortices of the surrounding free world; and that the above-mentioned energy-information barrier does not exist for them on microscopic section of the spectrum. The gist of the barrier itself comes to "memorizing" topoharmonic processes by Earth's gravitational fields and laboratory equipment. Turbulent fields (physical vacuum) do not have such a "memory".

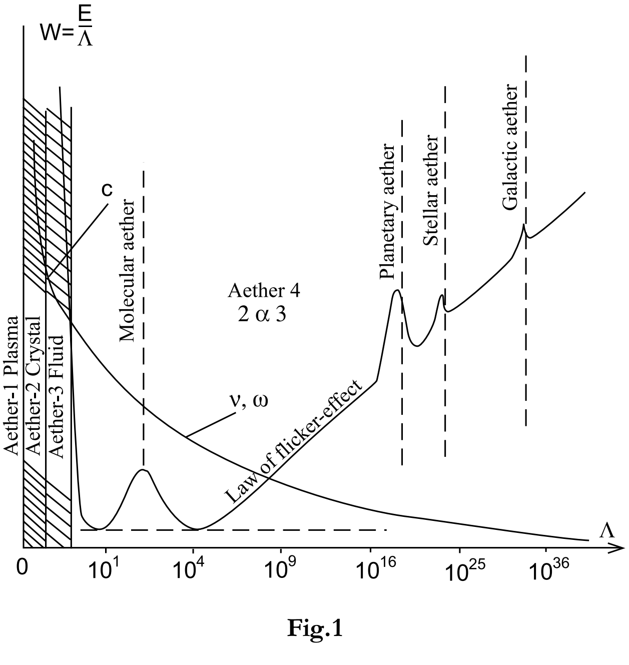

Fig.1

Fig.1 depicts diagrams of spectral density of vortices' energy; diagrams of rate of interaction transfer v and circular frequency ω depending on sizes of vortices of the world aether Λ (in picometers). As we can see, velocity of distortions propagation v is equal to velocity of light C in the physical vacuum, i.e. in electron sizes, exceeds C in nanocosm and tends to zero toward megacosms. Circular frequency of vortices rotation and course of physical time have the analogous qualitative dependence. However, vortices energy W has a feature and grows toward both ways from the lab scales; the aether numbers have been assigned conditionally, but all the specified phase states reflect some physical reality.

One of the most complicated questions is a definition of vortex matter in each particular case. The diagram demonstrates energy of free aether in the form of protomatter (simplest elementary particles – vortexes), and energy of classical elementary particles as helically definite self-oscillating emissions of aether field is localized in the most nonlinear subspace (physical or substantial vacuum). Unification of the elementary particles into atoms and macrobody does not denote direct transition to parallel world with different fundamental constants, whereas change of vortex size in the diagram exactly implies such a transition, though at a level of protomatter. This is a very intricate and crucial point. For example, a star is, unconditionally, an elementary particle of the relevant level of the Universe, even though comprising the ordinary matter.

Left wing of energy spectrum reflects property of the turbulent field of aether to gravitate, to sink down to the "bottom" transferring energy to the vortices of lesser scales. This phenomenon of distributed "gravitation" is available in every space point and explains a least-action principle. An illustrative analogue of this regularity is represented by the Earth's gravitational field and structure. Dense plasma assumed as iron core is situated in the center, and higher - there are magma, solid crust, water and air. These phase states exist in free aether within universe scales as well. Consequently, in femtocosm, and maybe at an even deeper energetic "bottom", there must be a plasmic subspace; evidently, it is assignable that aether translated from Greek means "fire".

Increase of energy to the right is connected with extending a dynamic band of the vortices spectrum, as the scale increases. In our world, this increase follows the flicker effect law, but this is only what we perceive across the energy-information barrier. Actually, at this section the vortex energy must increase well more strongly, because **our observation capabilities are limited** by accessible matter processes. Turning back to stars, one may affirm that modern science is unable to evaluate their real energy (evaluations are considerably underrated). The Galaxy vortex entrains such deep levels of multidimensional space-aether to its rotation that this causes violation of the Newton's law of inertia. Notion of "hidden mass" is unnecessary for explaining dynamics of galaxies. It is quite enough to take into account that differential rotation entrains the fine-structure aether (inertial subspace), and this process results in decreasing the inertia and external centrifugal forces. Unfortunately, Fig. 1 reflects only a characteristic size of the galactic vortex and does not visually demonstrate its active band of the spectrum. The finer is the space structure, the more inertial and closer to the absolute reference frame is the space. Obviously, the non-homogeneous, stereodynamically multidimensional and multiphase aether requires a perfectly new theory of random fields, which would also make allowance for the phenomenon of the material world.

Energy motion processes or spectral energy-cascades along abscissa axis are of decisive importance for existence of our world and the Universe. These processes have unusual properties. The "violet" cascade is attended with reduction of vortices size, absorption of matter energy and tends to increase energy concentration at the "bottom" of the world aether (near ordinate axis on Fig.1). This regularity experimentally confirmed in both spatial and temporal domains is a direct consequence of the above-mentioned phenomenon of distributed gravitation and clarifies an existence of the left wing of the spectrum. In classical physics "the violet" cascade effect is reflected in the least-action principle as well as in the thermodynamics laws.

The "red" energy cascade (right wing of the spectrum) corresponds to increase of vortices size and is accompanied with energy liberation. To start it up under earthy conditions special technologies are called for [7]. This has been confirmed with laboratory experiments and underlies the "perpetual motion" theory, where energy is pumped from nanocosm to the classical material world. Conception of artificial biofield [6], I have developed, considers methods of creating the aether vortex-wave macroscopic-scale distortions in physical systems. As a matter of fact, we deal with **ways of extending of natural spatio-temporal spectrum of aether (which settled into itself) by creation of a subspace of the topological subharmonics of the classical elementary particles, atoms and molecules**

(physical bodies), within the context of the etherodynamics fundamental law [5].

Even within action of fluctuating energy-information barrier, the spectrum right wing exhibits existence of natural processes in Nature that, supposedly, violate the known laws of thermodynamics [8]. But, as far as Nature cannot conflict with itself, we have to recognize that there is another physics, which "works" in geographic and greater scales, where role of the energy-information barrier demands to be reappraised. To all appearance, the universe vortex partialyl entrains plasmic subspace of aether into its rotation, and conditions for realizing "red" energy-cascade and generating matter are created on the sections of the maximum density gradient ("spectral" non-linearity). But the same way as a fish in ocean does not know reasons for storms and currents, we will never learn which force untwists the Universe handwheel, whose energy feeds our substantial world.

There are two components of free aether field that are the most essential for the present-day physics. The first one (left wing) is positioned on the spectrum beside the "bottom" in the form of dense liquid-crystalline "plasmatic" medium transforming to the Newtonian inertial subspace (physical vacuum). Since inertia is as regular as electric inductance is, the first subspace proves to be especially material and not such free, because it has properties which were imposed by the Universe rotation and distributed "gravitation". Let us conventionally limit the spatial spectrum of the first component with a nucleon size. The second component of aether space (right wing of the spectrum) stretches towards larger scales.

Except for known cosmic objects, this is an unknown ocean of the macrovortex and really free aether, or topoharmonics, which are born by dynamics of the substantial world. It is an experimentally proven fact that topoharmonics are generated by photons, electrons, and all the moving bodies [5, 9].

The second component has an ability to conversely affect the matter it was generated by as well as other material objects via "flickering" forces [10, 11]. This special subspace turns out to be doubly virtual, a genuine phantom. That is why, it has been winning the status of a real physical entity with such a difficulty, while Chinese philosophers were familiar with the same thousands years ago, separating entities "yang" and "yin".

The laboratory experiments indicate that topoharmonics have not only properties described in [5], but also an ability to interact with the space, creating a static domain of the excited aether in the course of propagation of the vortices-waves. At that physical bodies placing within this space volume are "charged" [2, 9]. The mentioned active domain (horizontal route) is "separated" by the Earth's gravitational field as follows: quasi-substantial constituent rapidly settles down creating vertical flow, but some phase structures are disintegrated very slowly (days, months, years).

The surprising thing is that biosystems equally belong to both aether components, not experiencing any constraint from the side of the energy-information barrier. The so-called human biofield is a topoharmonic part of physical body. It occupies its place on the spatio-temporal spectrum (Fig. 1). Quasi-substantial constituent (yang) and antiquasi-substantial one (yin) are strictly balanced, and acupuncture points and chakras are an analogue of Hartmann's geophysical grid, which is well-known to paraphysicists.

Conceptual generalization of physical ideas in the context of unified field of aether gives a constructive, cognitive impulse in solving of different problems, such as Nature's choice of exclusive laws for existence of "left-screw" material world as well as the problem of flicker-effect. At that, researching into the physical processes responsible for proton positive charge is of fundamental importance.

In conclusion, let us take note of conventional character of dividing the aether field into two components by the form of energetic spectrum. As a matter of fact, topoharmonics are generated by matter towards both larger spatial frequencies (subharmonics) and smaller ones (harmonics) and fill all levels of the Universe. The aforesaid is aimed at making perception of the most studied aspect of the **new physical entity** (vague due to diversity of hypothetical descriptions which exist in publications on this subject) more accessible.

Reference

1. A.P. Smirnov, I.V. Prokhortsev. Order principle. – Saint-Petersburg: ZAO «P and K», 2002. - 296 p.
2. A.M. Mishin. The aether model as a result of the new empirical conception. New ideas in natural sciences. (On materials of International Conference). Part 1 "Physical". - St.-Petersburg: RAS, 1996, p. 95-104.
3. A.M. Mishin. Fundamental properties of aether. Fundamental Problems of Natural Science. Series «Problems of the Universe research», fascicle 21 – Saint-Petersburg: RAS, 1999, p. 176-176.
4. A.M. Mishin. Physics of parallel worlds. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 23 - Saint-Petersburg: Publishing

House of the Saint-Petersburg State University, 2001, p. 258-269.
5. A.M. Mishin. Etherodynamics basic law. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 26, part III. - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2003, p. 206-218.
6. A.M. Mishin. Physical system with artificial biofield. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 23, - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2001, p. 258-269.
7. V.F. Zolotarev, V.V. Roschin, S.M. Godin. About space-time structure and some interactions. - Moscow: Publishers «PREST», 2000. - 309 p.
8. V. Starr. Physics of phenomena with negative viscosity. - Moscow: MIR, 1971.
9. V.V. Kvartalnov, N.T. Perevozchikov. Discovery of a "non-physical" component of OKG radiation. Journal «Parapsychology and Psychophysics». - Moscow: 1999, No. 2 (28), p. 64-67.
10. A.M. Mishin. Longitudinal thermomagnetic effect. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 26, part III. - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2003, p. 219-228.
11. A.A. Efimov, A.D. Shpitalnaya, Y.A. Zakoldaev. Eruptive prominences and earthquakes in the view of global space anisotropy. / «New Ideas in Natural Science». Series «Problems of the Universe research», fascicle 19, part II. – Saint-Petersburg: RAS, 1996, p. 403-413.

GLOBAL ENERGY

Philip M. Kanarev, Russia

The Kuban State Agrarian University, Department of Theoretical Mechanics
13, Kalinin Street, 350044 Krasnodar, Russia
E-mail: kanphil@mail.kuban.ru

Lately, the notion "Global energy" has become widely spread as a prize corresponding to this notion has been instituted in Russia. Since the Euclidean times, an unwritten law exists in science: to give definition to the notions, which are involved in scientific analysis. It is done for the purpose that everybody who uses this notion can understand the sense implied in it.

The Global energy problems have been discussed for a long time; they are well known. Depletion of natural energy carriers (oil, gas and coal) is the first problem. Environmental safety of energy carriers is the second problem. It is generally recognized that the solution of the issues connected with the aggregate of these two problems is of a global character. The results of scientific research, which should deal with global energy, originate from it. First and foremost, they are such scientific investigations, which result in the possibility to use an inexhaustible and environmental friendly energy carrier. It has been known for a long time. It is hydrogen produced from water.

Why hydrogen and why from water? At combustion of hydrogen, water is regenerated; thus, the energy carrier is inexhaustible. As far as inexhaustibility and environmental safety are concerned, hydrogen has no competitors. However implementation of these qualities is restricted by large energy expenses for hydrogen production from water. For hydrogen production, modern electrolyzers use from 10 to 20% more energy than it is produced at combustion of hydrogen.

It is easy to imagine what financial and intellectual resources of the world are included in the search of the ways to reduce energy expenses for hydrogen production from water. In Russia many scientific institutions of applied research and educational institutions deal with this problem. There is a Research Hydrogen Institute.

The associations of scientists on hydrogen energetics have been established in USA and Europe. This year, they carry out their regular scientific conferences, to which the author of this article has been invited.

Editorial: Let us note that it is not necessary to coin new scientific term "global energy", since quite for a long time the notion of "free energy" is used. This term represents the idea of the possibility to freely use energy by a consumer.

Energy problem is global not due to depletion of oil and gas, but due to their environmental unsafeness. However the confirmations exist that the world owners of energy resources do not bother about the problem of environmental unsafeness of modern energy carriers.

House of the Saint-Petersburg State University, 2001, p. 258-269.
5. A.M. Mishin. Etherodynamics basic law. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 26, part III. - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2003, p. 206-218.
6. A.M. Mishin. Physical system with artificial biofield. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 23, - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2001, p. 258-269.
7. V.F. Zolotarev, V.V. Roschin, S.M. Godin. About space-time structure and some interactions. - Moscow: Publishers «PREST», 2000. - 309 p.
8. V. Starr. Physics of phenomena with negative viscosity. - Moscow: MIR, 1971.
9. V.V. Kvartalnov, N.T. Perevozchikov. Discovery of a "non-physical" component of OKG radiation. Journal «Parapsychology and Psychophysics». - Moscow: 1999, No. 2 (28), p. 64-67.
10. A.M. Mishin. Longitudinal thermomagnetic effect. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 26, part III. - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2003, p. 219-228.
11. A.A. Efimov, A.D. Shpitalnaya, Y.A. Zakoldaev. Eruptive prominences and earthquakes in the view of global space anisotropy. / «New Ideas in Natural Science». Series «Problems of the Universe research», fascicle 19, part II. – Saint-Petersburg: RAS, 1996, p. 403-413.

GLOBAL ENERGY

Philip M. Kanarev, Russia

The Kuban State Agrarian University, Department of Theoretical Mechanics
13, Kalinin Street, 350044 Krasnodar, Russia
E-mail: kanphil@mail.kuban.ru

Lately, the notion "Global energy" has become widely spread as a prize corresponding to this notion has been instituted in Russia. Since the Euclidean times, an unwritten law exists in science: to give definition to the notions, which are involved in scientific analysis. It is done for the purpose that everybody who uses this notion can understand the sense implied in it.

The Global energy problems have been discussed for a long time; they are well known. Depletion of natural energy carriers (oil, gas and coal) is the first problem. Environmental safety of energy carriers is the second problem. It is generally recognized that the solution of the issues connected with the aggregate of these two problems is of a global character. The results of scientific research, which should deal with global energy, originate from it. First and foremost, they are such scientific investigations, which result in the possibility to use an inexhaustible and environmental friendly energy carrier. It has been known for a long time. It is hydrogen produced from water.

Why hydrogen and why from water? At combustion of hydrogen, water is regenerated; thus, the energy carrier is inexhaustible. As far as inexhaustibility and environmental safety are concerned, hydrogen has no competitors. However implementation of these qualities is restricted by large energy expenses for hydrogen production from water. For hydrogen production, modern electrolyzers use from 10 to 20% more energy than it is produced at combustion of hydrogen.

It is easy to imagine what financial and intellectual resources of the world are included in the search of the ways to reduce energy expenses for hydrogen production from water. In Russia many scientific institutions of applied research and educational institutions deal with this problem. There is a Research Hydrogen Institute.

The associations of scientists on hydrogen energetics have been established in USA and Europe. This year, they carry out their regular scientific conferences, to which the author of this article has been invited.

Editorial: Let us note that it is not necessary to coin new scientific term "global energy", since quite for a long time the notion of "free energy" is used. This term represents the idea of the possibility to freely use energy by a consumer.

Energy problem is global not due to depletion of oil and gas, but due to their environmental unsafeness. However the confirmations exist that the world owners of energy resources do not bother about the problem of environmental unsafeness of modern energy carriers.

In history of science, the facts of annihilation of the scientists who have won success in reduction of energy expenses for production of hydrogen from water have already been registered. The requesters of these actions think that they will lose their profits with the coming of hydrogen energetics. They do not understand that this process cannot be abrupt. It is impossible to replace the infrastructure of the existing energy carriers by infrastructure of hydrogen energetics within one or even ten years. Besides, infrastructure of hydrogen energetics will not be created on a blank place. It will be integrated into the existing infrastructure of energetics gradually; and its owners will automatically become the owners of hydrogen energetics. The first step has already been made in this direction. In the USA, a decision has been taken to equip all filling stations with pumps to fill the cars with hydrogen.

It is known that a priority of results of theoretical investigations is their publication in press. Usually, such priority is a personal one. Generally a patent is a priority of the results of experimental investigations. As a rule, this patent belongs to a group of authors. A published patent is a genie released from a bottle. No finesse of the authors to hamper a reproduction of experimental data given in a patent without the participation of the authors can stop the process of their implementation. Thus, the authors or a group of the authors who have filed an application for a patent are deprived automatically of the opportunity to influence the process of practical realization of their ideas.

In Russia, the state is a main patent holder of the global patents. Due to the well-known reasons, it has proved to be unable to control the course of scientific investigations and to forecast significance of their results.

It is known that if it becomes possible to reduce energy expenses for hydrogen production of water fivefold, it will be the cheapest energy carrier. Russia has already got technology, which reduces these expenses tenfold and more. However another way of looking is more perspective. Is there any use to decompose water into hydrogen and oxygen and to use hydrogen as fuel for heating, for example, of water in heating systems? Is it possible to make water generate heat? It turned out that, it is possible.

In Russia, three firms ("Yusmar", "Termovikhr" and "Noteka") sell cavitation heating equipment with energy performance index up to 150%. Official science looks awry at this activity, since such results conflict with one of the main laws of physics: law of conservation of energy. But market profit is stronger than this law.

Meanwhile, engineering practice has already proved that additional energy in the form of heat is generated in the ventilation systems and in the systems of water cavitation. Thorough scientific analysis of this problem shows that physical vacuum is the most probable source of additional energy in the systems of ventilation and water cavitation. Valence electrons of destroyed molecules of water take energy from physical vacuum and release it during repeated fusion of these molecules.

Why is additional energy generated in the air systems of ventilation and in systems of water cavitation? Because they are mechanical systems and mechanical destruction of chemical links requires half energy as compared with thermal destruction of these links. This is the main reason why one fails to increase energy performance index of cavitation processes over 200%.

Certainly, an increase of efficiency of any process by 30 or 50% is a good result; if it is obtained, it is possible to get even better one. But what if a water molecule is destroyed not mechanically, but electrodynamically? In this case, it becomes possible to find resonance frequencies of influence on the molecules and therefore to reduce considerably expenses of electrical energy for their destruction. Subsequent fusion of destroyed molecules will release determined quantity of energy unavoidably. It is a simple idea, and it has already been implemented. At electrodynamic influence on water molecules, electrical energy is converted into thermal energy with tenfold energy performance index. It means that if we spend 1 kWh of electric energy, we will get 10 kWh of thermal energy. To obtain such results, it is necessary to have knowledge of physchemistry of the microworld, which corresponds to knowledge of the 21st century, and it has already been published. Every month more than 1000 foreign scientists become familiarized with this knowledge at
http://Kanarev.innoplaza.net and
http://book.physchemistry.innoplaza.net.
The Russian speaking readers get this information from:
http://www.n-t.org/tp/ns/if.htm,
http://www.ikar.udm.ru/sb28-2.htm and
http://www.n-t.org/tp/ts/eb.htm.

I hope that now the readers will know the essence of the notion "global energy" and will understand those results of scientific researches which can really solve the related problems.

Energy Balance of Fusion Processes of Oxygen, Hydrogen and Water Molecules

Philip M. Kanarev, Russia

The Kuban State Agrarian University, Department of Theoretical Mechanics
13, Kalinin Street, 350044 Krasnodar, Russia
E-mail: kanphil@mail.kuban.ru

There is revealed the cause of appearance of additional energy during formation of covalent bonds in the fusion processes of oxygen, hydrogen and water molecules, besides the source of this energy is described.

Introduction

Engineering practice connected with servicing of ventilation systems allows revealing appearance of excessive thermal energy in circulated air. Similar phenomenon has been registered in systems of water circulation with the devices for its active cavitation. The results of our investigations explain not only a cause of these phenomena, but they give an opportunity to perform quantitative calculations for energy processes, which generate additional thermal energy [1], [2], [3], [4], [5].

Theoretical part

An oxygen atom is the eighth element of the periodic table. It is situated in the sixth group. The structure of its nucleus is given in Fig. 1 [1], [2], [3].

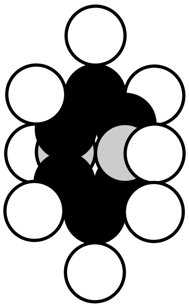

Fig. 1
Diagram of nucleus of oxygen atom: light – the protons, dark and grey – the neutrons

In Fig. 2 a diagram of the oxygen atom originating from the structure of its nucleus is given (Fig. 1). It has eight electrons. The electrons situated on the axis of symmetry are the most active ones (1, 2). Other six electrons situated in the plane, which is perpendicular to the axis line (a line of symmetry), by means of their total electric field remove electrons 1 and 2 from the nucleus at a large distance at that forming conditions for their large activity during the interaction with the electrons of the neighbouring atoms [1], [2], [3].

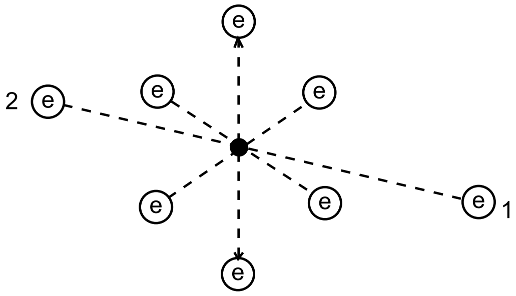

Fig. 2
Diagram of the oxygen atom

The least ionization energy of the electron of oxygen atom is equal to E_i =13.618 eV. Binding energy of this electron with the atomic nucleus corresponding to the first energy level is equal to E_1 =13.752 eV. Let us call this electron the first one. The calculation of energy indices of this electron, including its binding energies E_b with the atomic nucleus, according to the formulas (1) and (2), gives the following results (Table 1) [1], [2], [3].

$$E_{ph} = E_i - \frac{E_i}{n^2} = E_i - \frac{E_1}{n^2} \quad (1)$$

$$E_b = \frac{E_1}{n^2} \quad (2)$$

Table 1
Spectrum of the first electron of the oxygen atom

Values	n	2	3	4	5	6
E_{ph} (exp.)	eV	10.18	12.09	12.76	13.07	13.24
E_{ph} (theor.)	eV	10.16	12.09	12.76	13.07	13.24
E_b (theor.)	eV	3.44	1.53	0.86	0.55	0.38

The oxygen molecule structure is given in Fig. 3, a. It is formed by means of a connection of unlike magnetic poles of axis electrons of two oxygen atoms [1], [2], [3]. It is known that the fusion process of the oxygen molecules is accompanied with a release of 495 kJ/mole of energy, or in calculation for one molecule

$$E_b = \frac{495\cdot 1000}{6{,}02\cdot 10^{23}\cdot 1.602\cdot 10^{-19}} = 5.13eV. \qquad (3)$$

What principle does the Nature follow by distributing energy of 5.13 eV between the electrons of oxygen molecule (Fig. 3, a)? Energy of 5.13 eV is a thermal binding energy between the electrons 1 and 2' of two oxygen atoms (Fig. 3, a). When the oxygen molecule is formed, it is emitted in the form of the photons by the electrons, which enter into the bond. Hence it is equal to an amount of energies of two photons emitted by these electrons. Consequently, each contacting electron emits a photon with energies of 5.13/2=2.565 eV=E_b (Fig. 3). According to Table 1, in this case the valence electrons are situated between the second energy level and the third one [1].

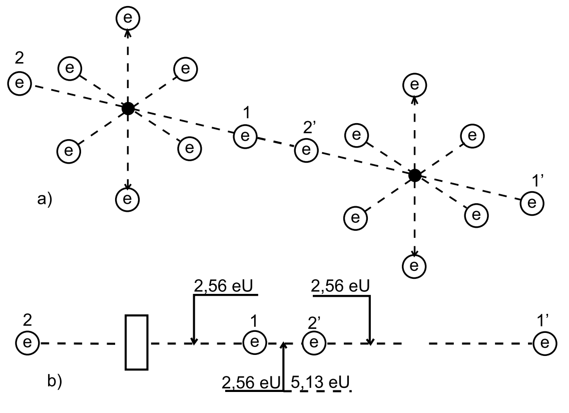

Fig. 3
Diagram of binding energy distribution between the electrons in the oxygen molecule

Two oxygen atoms are combined into a molecule in an excitation state. The excitation state is the state of an atom when its valence electrons are situated at such distances from the nuclei when the binding energy E_b between them is reduced to the thousandth of fractions of an electron-volt. In such state the atom can loose an electron and become an ion. Otherwise, without loosing electrons it is combined with an electron of the neighbouring atom by the valence electron, and a process of formation of oxygen molecule begins. It is an exothermic process when the axis valence electrons 1 and 2' emit photons, descend on lower energy levels and release 2.565x2=5.13 eV.

Let us pay attention to the fact that energy 5.13 eV is released by two electrons, which form a bond with energy of E_b=2.56 eV. In modern chemistry this bond is called a covalent bond. In order to break this bond it is necessary to use 2.56 eV of mechanical energy. For thermal cleavage of this bond, double quantity of energy is required, i.e. 5.13 eV. It is explained by the fact that the photon energy of 5.13 eV is absorbed by two electrons simultaneously. Only in this case, both electrons will be transferred to the highest energy levels with minimal binding energy E_b when they are disconnected, and each oxygen atom becomes a free one.

Thus, energy expenses for destruction of oxygen molecule depend on the method of influence upon the bond. During thermal action upon the bond it is destroyed when energy is 5.13 eV. During mechanical effect upon the bond, it is necessary to spend 2.56 eV of energy in order to destroy this bond. Therefore energetic of fusion process of the oxygen molecule depends on method of its destruction.

After thermal destruction of the oxygen molecule process of its formation begins from emission of the photons with energies of 2.56 eV by both valence electrons, and the previous electrodynamics binding energy (E_b=2.56 eV) is restored between the electrons of both atoms.

Thus during thermal destruction of the oxygen molecule the same amount of thermal energy is spent than that which is released during its further formation. No additional energy appears during thermal dissociation of oxygen molecule and at its further fusion.

If oxygen molecule is destroyed by a mechanical method, then it is necessary to spend 2.56 eV of mechanical energy for this purpose. Valence electrons of oxygen atoms are in a free state at lack of energy, which corresponds to such state, as there is no process of absorption of 2.56 eV of energy by each of them. The electrons cannot remain in such state; they should replenish immediately the energy, which they have failed to receive during a mechanical break of the bond between them. Where should they take it from? There is only one source: the environment, i.e. the physical vacuum filled with aether. They convert aether into energy of 2.56 eV immediately. The next stage is a connection of two oxygen atoms, whose valence electrons have replenished the reserves of their energy by means of aether. This process is accompanied by emission of the photons with energies of 2.56 eV by two electrons. Thus energy of absorbed aether is converted into thermal energy of the photons. If we spend 2.56 eV of mechanical energy for destruction of oxygen molecule, we will get double quantity of energy (2.56x2=5.13) eV during further fusion of this molecule. Additional energy is equal to 2.56 eV.

Much experimental data show that in ventilation systems thermal energy of circulated air exceeds electric energy spent for a fan drive. Now we know that this energy is generated at mechanical destruction of covalent bonds in the molecules of gases, which the air consists of.

Using the above-mentioned method, let us analyse energetic of water molecule, which sometimes generates additional thermal energy. A water molecule consists of one oxygen atom and two hydrogen atoms. Binding energies E_b of the hydrogen atoms with its nucleus are given in Table 2 [1], [2], [3].

Table 2

Spectrum of hydrogen atom

Values	n	2	3	4	5	6
E_{ph} (exp.)	eV	10.20	12.09	12.75	13.05	13.22
E_{ph} (theor)	eV	10.198	12.087	12.748	13.054	13.220
E_b (theor.)	eV	3.40	1.51	0.85	0.54	0.38

It is known that combination of hydrogen and oxygen is accompanied by an explosion, but its cause remains unknown. Let us try to find it.

Energy of fusion of hydrogen molecule is equal to 436kJ/mole, or 4.53 eV per a molecule. As the molecule consists of two atoms, then the above-mentioned energy is distributed between them. Thus energy of one bond E_b between the hydrogen atoms is equal to 2.26 eV (Fig. 4). At mechanical destruction of this bond 2.26 eV is enough. At thermal destruction of this bond double quantity is required (2.26x2=4.53 eV) [1].

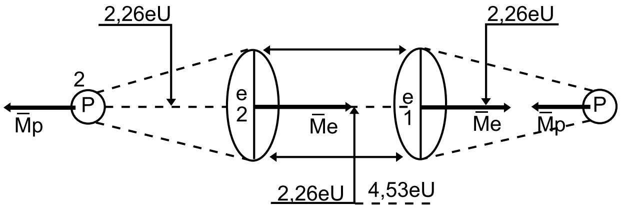

Fig. 4

Hydrogen molecule

In order to form two water molecules, it is necessary to break two hydrogen molecules and one oxygen molecule into atoms. At mechanical destruction of covalent bonds 2.26x2=4.53 eV is required to break two hydrogen molecules and 2.56 eV to break an oxygen molecule. Sum of these energies is equal to 7.13 eV. If the destruction processes of the above-mentioned molecules are carried out with a thermal method, then 4.53+4.53=9.06 eV is required for the destruction of two hydrogen molecules, and 5.13 eV is required for the destruction of one oxygen molecule. In total, it is equal to 14.19 eV. The difference between the energy spent for mechanical and thermal destruction of covalent bond of hydrogen and oxygen molecules is almost double.

It is known that during fusion of one mole of water 285.8 kJ or $285.8x1000/6.02x10^{23}x1.6x10^{-19}=2.96$ eV per a molecule are released. As a water molecule consists of one oxygen atom and two hydrogen atoms, 2.96/2=1.48 eV falls per the bond (Fig. 5). Hence the electrons of hydrogen and oxygen atoms in water molecule are between the forth energy level and the fifth one at the usual temperature (1.48/2=0.74 eV = E_b), Table 1,2 [1].

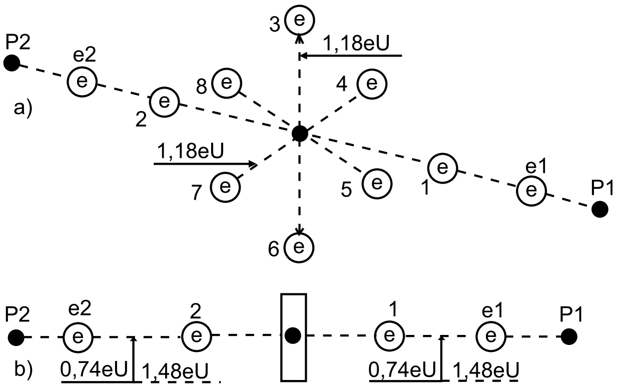

Fig. 5

Diagram of water molecule:
1, 2, 3, 4, 5, 6, 7, 8 are the numbers of the electrons of oxygen atom; P_1, P_2 are the nuclei of the hydrogen atoms (the protons); e_1 and e_2 are the numbers of the electrons of hydrogen atoms

Thus when two hydrogen molecules $2H_2$ and one oxygen molecule O_2 are destroyed by the thermal method, 14.19 eV are spent. As a result of fusion of two water molecules ($2H_2O$), 2.96x2=5.98 eV is released. There is some disbalance here since fusion process of water molecule is an exothermic one and 2.96 eV is released by one molecule. The given calculation shows that (14.19-5.98)/2=4.10 eV is absorbed during fusion of one water molecule. What is the cause of this contradiction?

The oxygen atom in the water molecule should reduce its volume when the transition from gaseous state into liquid state takes place. It will happen when the ring electrons of oxygen atom descend on lower energy levels (nearer to the nucleus). They will emit the photons and their total energy will be equal to energy spent to destruction of two hydrogen molecules and one oxygen molecule, i.e. 14.19 eV. Since two water molecules have 12 ring electrons, each of them will emit 14.19/12=1.18 eV=E_b (Fig. 5). It is more than energy (E_b=0.74 eV) of binding of axis electron with the nucleus, and it shows that the ring electrons are situated nearer to the nucleus than the axis ones.

In this case quantity of energy produced due to fusion of two water molecules (14.19+5.98) eV exceeds energy, which was spent for the destruction of two hydrogen molecules (9.06 eV) and one oxygen molecule (5.13 eV). Energy difference of 5.98 eV is divided between two water molecules. It means that 5.98/2=2.99 eV or 285.8 kJ/mole fall per a molecule. It corresponds to the existing experimental data completely [1].

The above-mentioned facts clarify a cause of the explosion, which takes place when hydrogen is combined with oxygen. Simultaneous transition of six ring electrons of each oxygen atom in the nascent water molecules to lower energy levels is accompanied by simultaneous emission of the photons, which generate explosion phenomenon.

Let us pay attention to the fact that two binding energies E_b between valence electrons e2 and 2 and between 1 and e1 are shown in Fig. 5, b. Energy of one electrodynamics bond is equal to E_b= 0.74 eV. If this bond is destroyed by the thermal method, 0.74x2=1.48 eV is required. This energy will be released during further fusion of the water molecule from hydrogen atom H and hydroxyl ion OH^-. In this case, no additional energy is generated.

Therefore the given bond is destroyed by the mechanical method spending 0.74 eV per a bond, each electron will have energy deficit equal to 0.74 eV after bond destruction. This energy will be immediately absorbed from the environment and will be emitted during the repeated fusion of the water molecule from the hydrogen atom H and the hydroxyl ion OH^-. At mechanical destruction of one bond of water molecule, the covalent chemical bond forms E_b= 0.74 eV of additional thermal energy, which is registered in systems of water cavitation constantly (as we have already noted) [1], [2], [3].

It is known that water molecules combine and form clusters. If the bonds between the molecules in the clusters are covalent ones, mechanical destruction of these bonds should be accompanied by a release of additional thermal energy as well [1], [2], [3].

Experimental Part

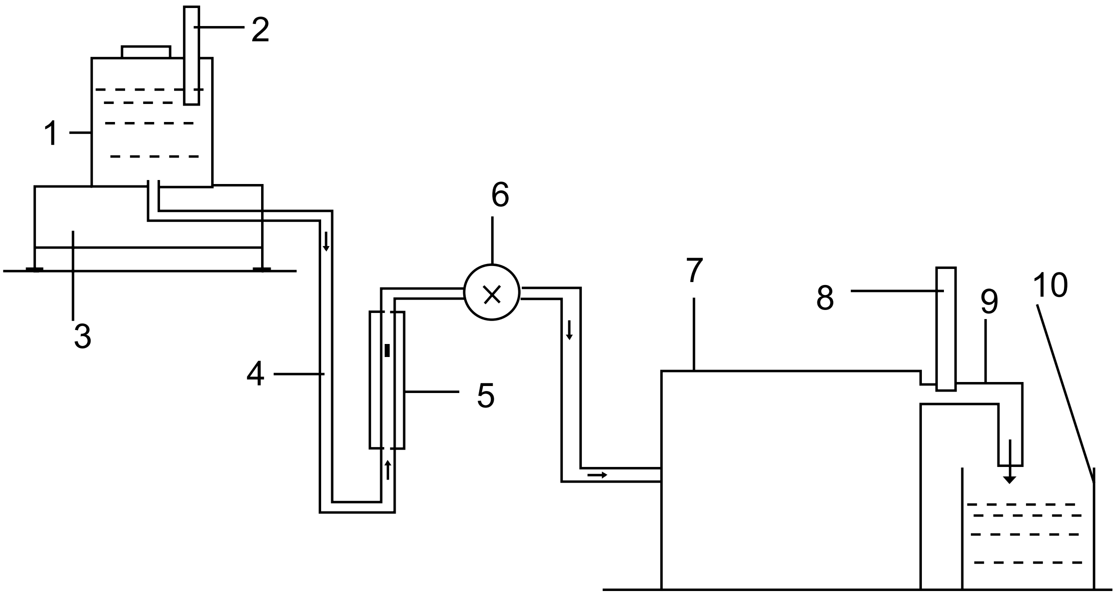

Fig. 6

Diagram of the experimental device: 1 - reservoir for solution; 2 - thermometer; 3 - electronic scales; 4 - solution supply duct; 5 - rotameter; 6 - feed solution regulator; 7 - a special thin plasma reactor is in the process of patenting; 8 - thermometer; 9 - discharge of heated solution; 10 - inlet reservoir

Thus, chemical bonds between the atoms in the molecules and the molecules in the clusters can be destroyed mechanically, by electrodynamic and thermal influence. We have already shown that the mechanical way of destruction of such bonds requires half energy as compared with thermal energy. It appears from this that energy expenses for electrodynamic destruction of these bonds should be less than thermal expenses as well. Electrodynamic impact on the bond gives the opportunity to form the resonance modes where energy expense for the destruction of these bonds is reduced to greater degree. In order to check this hypothesis a special experiment was carried out. It was connected with electrodynamic destruction of chemical bonds of water molecules with a changing frequency of impact. To test this hypothesis the check experiment was prepared and carried out by (besides the author of this article) A.I. Tlishev, G.P. Perekotiy, D.A. Bebko, D.V. Korneev. A diagram of the experimental device is given in Fig. 6. The results of this experiment are given in Table 3.

Table 3

Protocol of control test

Indices	1	2	3	Mean
1 – mass of the solution, which has passed through the reactor m, kg.	2.112	2.153	2.118	2.128
2 – temperature of solution at the input of the reactor t_1, degrees	24	24	24	24
3 – temperature of the solution at the output of the reactor t_2, degrees	33.5	33.5	33.5	33.5
4 – differential temperature of the solution $\Delta t = t_2 - t_1$, degrees	9.5	9.5	9.5	9.5
5 – durability of the experiment $\Delta\tau$, s	300	300	300	300
6 – reading of voltmeter V, B	25.0	25.0	25.0	25.0
7 – reading of ammeter I, A	1.40	1.40	1.40	1.40
8 – electric power consumption according to indices of voltmeter and ammeters, $E_2 = I x V x \Delta\tau$, kJ	10.50	10.50	10.50	10.50
9 – power spent for heating of the solution, $E_3 = 4.19 x m x \Delta t$, kJ	84.10	85.70	84.31	84.70
10 – reactor efficiency index according to the reading of voltmeter and ammeter $K = E_3 / E_2$	**8.00**	**8.16**	**8.03**	**8.06**

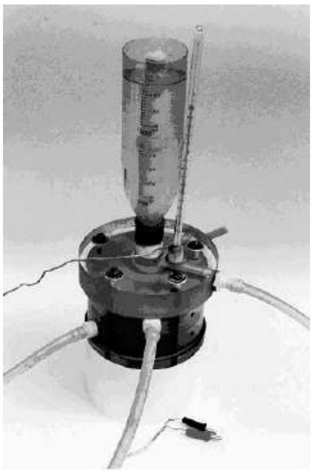

Fig. 7

Photo of heat reactor

In the Russian market three firms (Yusmar, Termovikhr and Noteka) sell cavitation water heating equipment with energy efficiency index of 150%. Soon, an air heating devices with the same efficiency will be produced. The processes of mechanical destruction of covalent bonds of the air gas molecules, molecules and clusters of water and their further fusion serve as a source of additional energy generated by these devices [1], [2], [3].

Conclusion

Analysis of energy balance of the molecules with covalent bonds shows the possibility of additional thermal energy formation during mechanical and electrodynamics destruction of these bonds.

References

1. Ph.M. Kanarev. The Foundation of Physchemistry of Micro World. Krasnodar. 2002. 320 pages. (In Russian).
2. Kanarev Ph.M. The Foundation of Physchemistry of Micro World. The second edition. (In Russian). http://www.ikar.udm.ru/sb28-2.htm
3. Kanarev Ph.M. The Foundation of Physchemistry of Micro World. The second edition. (In English). http://book.physchemistry.innoplaza.net
4. Kanarev Ph. M. Energy Balance of Fusion Processes of Molecules of Oxygen, Hydrogen and Water. (In English) http://Kanarev.innoplaza.net
5. Kanarev Ph. M. Energy Balance of Fusion Processes of Molecules of Oxygen, Hydrogen and Water. (In Russian) http://www.n-t.org/tp/ts/eb.htm.

The Fourth State Of Matter (Plasma Energy) Power Generation

Bruce A. Perreault, USA

http:// www.nuenergy.org
Email: nuenergy@cyberportal.net

Under the proper configurations and in the correct pre glow discharge environment more electrons can be released than what is required to trigger the release. This release of energy attains a high frequency oscillation that is indicative of the metal or metals involved in its release.

The release of high frequency energy from metals and semi metals (semi conductors) through the mechanism in the pre-glow discharge leads to the generation of radiant energy. This correlation has been identified through my findings. The release of high frequency energy from the atoms of electrical conductors is credited to my identification of the mechanism in the pre-glow discharge, or fourth state of matter. This has lead to extremely efficient alternative energy circuits.

The release of high frequency energy from the atoms of electrical conductors is credited to my identification of the mechanism in the pre-glow discharge, or fourth state of matter. This has lead to extremely efficient alternative energy circuits.

The pre-glow discharge condition occurs when a sufficient amount of charge difference is applied across the gap of two electrical conductors. It makes a hissing sound if the electrodes are surrounded by air. The electrical properties between this gap change: electrons become "stripped" from their respective atoms and are liberated. In this state, air is **ionized** and is transformed into a **plasma** and is no longer a gas. It is now in the fourth state of matter, the other three being solid, liquid, and vapor (gas). The fourth state of matter is a good conductor of electricity, its average resistance being much lower than that of the same substance in its gaseous state.

A good readily available spark gap combination is lead and iron. This combination generates plenty of radiant energy and causes noise to be heard on a radio in its vicinity. As a general rule, the better the thermoelectric difference between the dissimilar couple the more radiant energy that will be generated when it is exposed to a plasma field.

Early in the summer of 1981, the pre glow had first caught my attention. This is where I had observed it between two plastic insulated hook-up wires from radio shack. The plasma between the two wires was called into existence because of a high voltage power supply that I had obtained for my electrogravitic experiments. The two wires were twisted together and given a charge of around ninety kilovolts.

It was not until 1989 when I observed a high frequency signal on my oscilloscope coming from the pre-glow discharge phenomenon. These signals were coming from across the room and originated from a crude component that I had built. This crude component consisted of a guitar wire that was centered through a section of PVC tubing. Around this tube was wrapped a coil of insulated transformer wire. This was my first ion-valve. It was excited with the same power supply where I had observed the plasma effect several years previous.

I knew that my ion valve was generating high frequency oscillations using air ions but did not find any practical applications for it until six years later. It was in this year (1995) when I realized the importance of the ion valve and its associated high frequency ion field, known as radiant energy to Moray, it being the essential principle behind his energy receiver.

In one experimental set-up, it was confirmed that when a charged capacitor is discharged through a spark gap that the stored energy is transferred by way of high frequency electrical oscillations. It was found that these oscillations occur directly before the discharge of the capacitor, thus the term, pre-glow discharge. It was found that during the pre-glow discharge plateau, energy is transferred from a primary to a secondary electrical coil. Directly after pre-glow discharge, a surge of current is measured. At the point of current surge is where most of the stored energy in the capacitor is released as wasted heat. It was apparent that by maintaining the pre-glow discharge plateau a very efficient mode of energy transfer and conversion can be obtained. Logically what is required is a control mechanism. This is where my ion valve finds its niche. Upon further investigation, it is found that it keeps pre-glow discharge oscillations at their plateau, and suppresses current surge in a spark gap. It is a self-regulating quenching device. When it is placed in series with an air-core transformer and arc gap, its plasma hisses smoothly, the arc in the spark gap is almost silent and is very small. Continuous oscillations can be seen on an oscilloscope. A light bulb connected to the secondary will glow brightly. This result is not possible without the ion valve included in the circuit. The ion valve has the added feature of preventing back electromotive force, generated from inductive loads, from being discharged back through the circuit. Without the ion valve, energy is normally lost when it is discharged and is transformed into heat. Using the ion valve therefore has the effect of conserving energy.

Fundamental Radiant Energy Device

“Standard” scientific principles are not being violated. There is no fringe science involved with radiant energy power generation. This is a very old source of energy being extracted and harnessed in a unique way through my discoveries. What I have discovered is practically a limitless source of energy. Radiant energy has existed since the beginning of time. What I hold claim to is an improved method to generate and to convert radiant energy, kinetically active ions, cosmic energy, call it what you will, into useful electrical power. Think of the radiant energy power generator as a type of energy detonator that liberates great quantities of energy with only a small exciting spark. No laws of physics are being violated. No new laws are being implied. They are being expanded. The concept is not that much different from how a lighted matchstick is able to start a bonfire.

The power generated from my circuits comes from the transformation of matter into radiant electricity. It does not come from the splitting of atoms. To obtain high wattage from a spontaneously radioactive substance would require unsafe amounts of radioactive material.

Under the right conditions, ordinary matter can be made to generate intense surges of radiant energy that can be heard on a radio receiver as static noise. Build a device that can efficiently capture this energy and convert it into useful electrical currents and you will have yourself a powerful source of electrical power. This device will be powered by artificially disintegrating matter as described by Gustave Le Bon in his book “The Evolution of Matter” and in his book “The Evolution of Forces”.

Energy and matter are two distinct entities of the same manifestation. Matter represents a stable condensed form of energy. Heat, light, electricity, etc., are uncondensed vibrations of matter oscillating at differing rates. “Cosmic Energy” is the term that Le Bon used to define matter and energy as being one and the same manifestation. He theorized that when stable matter is disintegrated it is transformed into energy that we recognize as heat, light, electricity, radioactivity, etc.

Semiconductors will generate intense surges of electron oscillations that become powerful sources of radiant energy. A simple **ion valve**, or call it what you will, can be used to generate and convert this form of radiant energy into useful electromagnetic oscillations. A tuned transformer can be used to directly convert these oscillations into to a practical voltage and amperage. This unique valve is shown in Fig. 1. Before it can function it must be connected to a suitable voltage source. The circuit in Fig. 2 demonstrates the full working concept. Other embodiments are also possible. The proof of concept circuit reveals how radiant energy can be generated and converted into useful electrical currents. Many additional stages can be added for more power. Many other circuit and component configurations can also be used but the fundamental conversion principle remains the same. The circuit and components shown should give you a very clear idea of how the technology works. The actual mechanics and electronics of building and validating a radiant energy prototype are relatively simple. Contrary to what some people might think, dangerous levels of radioactivity are never used in my devices.

Ion-valve Converter Technology Explained

The ion-valve converter (ion-valve) shown in Fig.1 has an axial negatively charged tungsten cathode wire that extends the length its cylinder and is capable of emitting secondary electrons. The anode cylinder is positively charged and is made from a semiconductive material that will readily capture electrons.

Within a few milliseconds the accumulated negative ions are attracted to the positively charged onrushing atomic ions. When the negative and positive charges collide they neutralize each other generating high frequency electrical oscillations.

There appears to be a common thread shared between several alternative energy devices. It is the pre-glow discharge. The report on the **Hans Coler** device released by the British Government indicates that there is excess energy released when electrical contacts are opened and closed. The **Lester Hendershot** device utilized a buzzer circuit that opened and closed its electrical contacts. In the **Alfred Hubbard** coil pre-glow discharge flowed through electrical contacts, a distributor cap and radium soaked spark plug. The **Joseph Newman** motor used a sparking commutator. **Thomas Moray** invented a glowing, cold cathode discharge tube that was the heart of his radiant energy device. **Hermann Plauson** was granted U.S. Patent No. 1,540,998 that used spark gaps to convert atmospheric energy. **Frank Wyatt Prentice** was granted Canadian Patent No. 253,765 that detailed his invention, which lighted 50 sixty-watt carbon lamps with an input of only 500 watts. His invention utilized a spark gap driven high frequency tuned resonant system. **Chancy Britten** used ion-valves constructed with a central wire that was surrounded by a coil of wire which is described in his US Patent No. 1,826,727. Britten's valve was said to have lit up his home in the 1930's according to a local newspaper article of that time period. **Alexander Chernetski** experimented with what appears to have been a type of ion-valve that was filled with hydrogen gas. It is said that he got up to five times more energy out of his device than what he put into it. **Edwin Gray** was granted U.S. Patent No. 3,890,548 for his efficient spark gap driven capacitive-discharge motor. He improved on this patent by replacing the spark gaps with a pre-glow discharge switching tube. His U.S. Patents No. 4,595,975 and No. 4,661,747 describes this tube in detail. Gray's patents claim to conserve battery power by sending unused energy back to the supply batteries. On close examination, we also find that his switching tube is actually a quenching device.

Radiant energy was generated during the pre-glow discharge cycle that also contributed to recharging the batteries. **Paulo N. Correa** and **Alexander N. Correa** obtained patents to a pulsed pre-glow discharge system that recovers energy and recharges a battery.

I have found that radiant energy is generated when a plasma field is in contact with the atoms of an electrical conductor. Most importantly, the quantity of radiant energy is greatly increased when the plasma field occurs between the inter-electrode couple of two differing electrical conductors. That, the wattage produced will depend largely on the type of electrode materials that are used. I do not suspect that the above referenced inventors were aware of this amplifying mechanism.

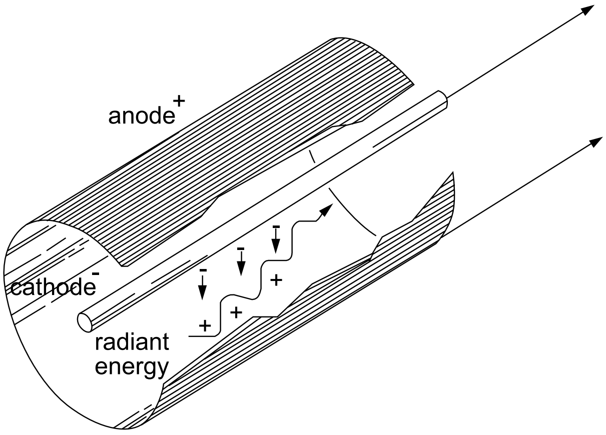

Fig. 1

Ion Valve

In the embodiment in Fig. 1 the negative charge on the wire in the ion-valve negatively ionizes any gas that contacts it. These ions rush towards the positively charged cylinder. When a metal ion that carries an excess electron hits head-on with an elementary differing positively charged onrushing metal ion a violent union of the two metals occurs. This causes the newly formed bimetallic alloy to violently oscillate, break up (dissociate) and to liberate a burst of electrons. The **"sea electron model"** helps to explain this effect. According to the sea electron model metals are bonded to each other through electron sharing. The model suggests that metal atoms be bathed in a sea of valence electrons. If this model is taken one step further it can be seen that when metallic atoms are dissociated from each other excess electrons are released in the form of electrical oscillations of high frequency (radiant energy). This occurs because the electrons no longer take part in the inter-atomic binding force that existed before the dissociation took place. It becomes clear that the freed electrons will add amperage to the output circuit to which it is connected. Henceforth, the equation **I x E = P** holds true in this system. Where, "I" represents the electrons (amperage), "E" electromotive force (ionic voltage), and "P" the power generated.

Obtaining Electrical Energy from the Transformation of Cosmic Energy

Matter is cosmic energy in a condensed state according to Le Bon and Moray. What this means is that matter can be excited its plasma state and caused to rapidly disintegrate transforming itself into electricity.

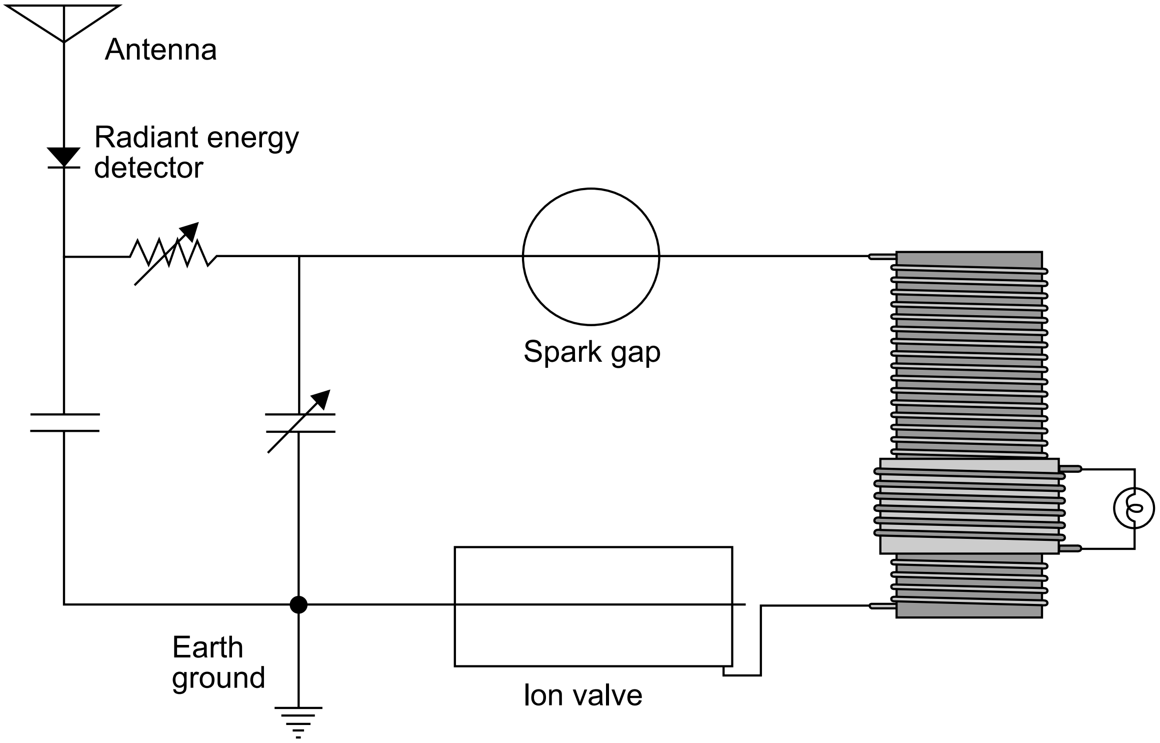

Fig. 2

Proof of Concept

The proof of concept shown will obtain energy from the dissociation of matter bringing the circuit to life. It does this by creating a plasma field between differing electrical conductors. Electrons are unleashed and feedback to the circuit where they are converted into useful power.

Plasma oscillating on the correct wavelength will act like a spark on a mass of explosive material but releasing electrical particles instead of heat. This is to say that the excited fourth state of matter (plasma) causes condensed energy (solid matter) to become uncondensed (radiant energy). The reaction will be far superior to the force that invoked it in the first place because the energy stored in matter is released. The energy that is condensed in the elements of matter is immense. The result is that an enormous amount of energy is released with only a slight loss of matter. Gustave Le Bon proved this when he demonstrated that the action of solar light and from electric arcs on bodies produced electric particles similar to those of uranium. He showed that it caused all bodies to disintegrate to different degrees.

> ***the excited fourth state of matter (plasma) causes condensed energy (solid matter) to become uncondensed (radiant energy). The reaction will be far superior to the force that invoked it in the first place because the energy stored in matter is released***

According to the law of conservation, when we give to a material body a determined quantity of energy, this energy might be transformed, but the body will never give back a quantity in excess of what it received. This principle is considered too self-evident to be disputed. It makes sense that matter can only give up energy that is given to it and is unable to create excess energy. Without violating this law, matter can be excited into giving up its stored inter-atomic energy. Condensed matter can become uncondensed if its oscillations become violent enough and henceforth transforming itself into radiant energy. No laws have been violated. They have been expanded. The first law of thermodynamics describes the principle of the conservation of energy. It states that "energy is not created or destroyed; it merely changes form". The fact is that the creation or destruction of energy is a result of matter being broken down or built up. They both go hand in hand.

Naturally occurring radioisotopes were created through the billions of years of **cosmic ray** bombardment of ordinary matter. Their matrixes became unbalanced. The correct trigger will cause these isotopes to seek their original balanced state. Enormous amounts of electrical energy can be obtained from the transformation of stored cosmic energy. These isotopes contain a very old energy reserve that can be released in a unique way with my discoveries.

Have you ever wondered about spontaneous radioactive decay? Is it really a super charged state of matter? If an atom can become ionized by either gaining or by losing electrons then why can there not be a nuclear ionization too? I hypothesize that atoms do become ionized on the nuclear level by gaining or losing neutrons. This may appear incidental but could be at the very heart of an ultra-chemistry. It is highly probable that this hypothesized phenomenon is behind natural spontaneous radioactive decay and explains many unanswered questions about nuclear science.

Liberated Energy

With a slight excitement from a pre-glow discharge in a spark gap, or even with no excitement at all, as we observe in spontaneously disintegrating radioactive bodies, such as U235, we can obtain large quantities of energy. Clearly, we did not create this liberated energy, since it already exists in matter, but we release it under the right conditions. This is being done without violating the law of energy conservation. The idea that matter could be transformed into energy was absurd before the acknowledgement of nuclear transformations.

A Nu Science is on the horizon. It involves the means of transforming matter into energy without splitting atoms. This science recognizes several isotopes of matter that spontaneously liberate energy as observed in naturally occurring radioisotopes. My research indicates that it is also possible to artificially speed up the natural decay process of condensed energy (matter) using a minute plasma field, as from a pre-glow discharge, etc… With a very small quantity of energy, we will be able to produce a very large quantity of energy without splitting the atoms.

Capture Capacitor

Nature offers us cosmic energy that manifests itself in many different forms. Electricity is only one of its manifestations. From this knowledge, we can obtain electrical power with no moving parts. The natural world contains many storehouses of this cosmic energy. Energy is all around us just waiting to be transformed. The capture capacitor is one such example. It is called a capture capacitor because it captures and transforms electrical charge into electrical current. It does this with the aid of naturally occurring radioactive matter (N.O.R.M.). Put a pair of dissimilar metal electrodes together with a porous ceramic dielectric material sandwiched between them, along with a weak electrolyte and you have built a capture capacitor. In a low power capture component there exists a minute impurity of N.O.R.M. It is always present in clays and is usually present in ceramic materials.

If you want a capture capacitor to generate more power all you have to do is to add additional amounts of radioactive material to the dielectric. Lead–210 is the optimum choice because it has a half-life of around 22.3 years and is a pure beta (electron) emitter. Its half-life is almost twice as long as tritium. This means that it is feasible to build a component that will put out power for several years with little maintenance. Lead-210 is a decay product of radon gas. The parent source of radon is uranium. Therefore, uranium ore can be powdered and mixed into the ceramic material.

The atomic ions emitted from naturally occurring or artificially induced radioactive transformations can be directly converted to electrical power. The circuit shown can be used to convert radiant energy into useful electrical currents. Additional stages can be added for more power. The circuit shown gives a general idea of how such an energy conversion device functions. By no means is this technology limited to this one circuit, configuration, or source of radiant energy.

EAGLE-RESEARCH
Energy Solutions
since 1984

A Research Organization that
Develops & Distributes
Practical Energy-Saving
Methods & Devices

4 Energy Way, P.O. Box 118
Porthill, Idaho 83853
FAX: 250/492-7480

Technical questions to:
wiseman@eagle-research.com

NEWS REVIEW

◆ A new type of high field superconducting magnet from Oxford Instruments has enabled physicists at the University of Nottingham to balance the force of gravity and levitate heavy and dense materials such as lead, gold and diamonds. The patented Minimum Condensed Volume™ (MCV™) Cryofree™ superconducting magnet, developed and manufactured by Oxford Instruments in collaboration with the University of Nottingham, is the first of its kind in the world. Unlike conventional superconducting magnets, the MCV magnet never needs to be refilled with liquid helium and yet can generate a field over 300,000 times stronger than the Earth's magnetic field. MCV has a very small footprint and, operating at 4.2 K with a room temperature bore, can generate highly stable magnetic fields up to 16 Tesla.

http://www.oxford-instruments.co.uk

◆ Polish inventor Zygmunt Orlowski called attention of the scientific society to his perpetual motion machine **METOZ**, which according to the author after further improvement will be able to extract energy from the earth's gravitation.

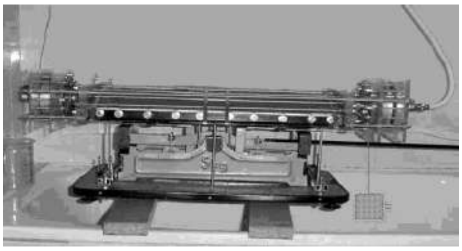

http://www.nets.pl/~metozor/perpetual_motion_machine.htm

◆ It is necessary to change behavior of people in order to decrease environmental damage. This was the message of **"Green Week 2003"** which took place since June 2 through June 5 in Brussels.

The conference's key issues were three aspects of United Europe environmental policy: sustainable production and consumption, renewable energy and water supply.

Green Week is the first environmental forum of such a scale. Three thousand participants from all over the world represented governments and funds, industry and mass media. The exhibition organized in the framework of Green Week included 62 displays from 14 European countries and were dedicated to the newest environmental technologies and projects. During the Green Week 26 conferences and seminars took place, as well as a great number of press conferences and actions for young people.

◆ **British researchers (**Gardner Watts company**) believe that they have made a groundbreaking scientific discovery after apparently managing to "create" energy from hydrogen atoms.**

The device represents "thermal energy cell" which appears to produce hundreds of times more energy than that put into it. If the findings are correct and can be reproduced on a commercial scale, the thermal energy cell could become a feature of every home, heating water for a fraction of the cost and cutting fuel bills by at least 90 per cent.

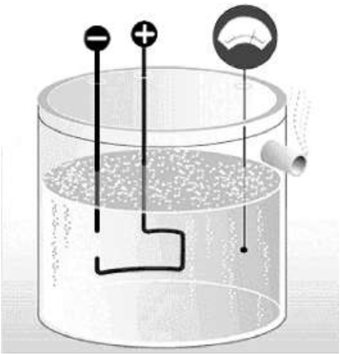

http://www.telegraph.co.uk

◆ **Gravity Conversion Rotary Device**

The below figure is from http://www.theverylastpageoftheinternet.com/newclaims/GravityMotor/gravity_motor.htm

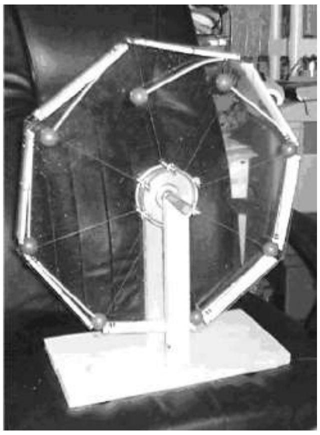

Does it really work?

THE MARCUS DEVICE CONTROVERSY

Tim Ventura, USA

Email: tventura6@attbi.com
http://www.americanantigravity.com
http://www.americanantigravity.com/marcus.html

Editorial: This overview is devoted to the fantastical device created by inventor Marcus Hollingshead that encountered more than its share of publicity and controversy between November 2002 and March 2003. You can participate in discussion on aggroup@yahoogroups.com.

The problem with Marcus has been that he was pressured by too many people to produce too fast results. I started talking to him in November, and he is always sounded like an honest, rational man with a good heart. The problem is that once people started to hear his private claims that he could lift something like 2 tons of weight, things kind of went south.

While people in general have been very well-meaning, they are desperate to solve many of our current energy-problems, and when Marcus appeared with a potential solution it got out of control. Pretty soon film crews were showing up at his house unannounced, and he was getting unwanted phone calls from people who had looked up his number online.

In my opinion, it is really a problem of a "social dynamic" at work here – too many people, with too much access, all at once. This also correlates to the type of person that different people are. For instance, I am a very open person, and I really do not value my privacy a heck of a lot — therefore, when TV crews drop by with 5 minutes notice (and they have), I tend to just let them in for a demo. However, Marcus, who is much more of a private person, seems to have become a bit overwhelmed at this proposition.
Also, keep in mind that I had initiated working with the media because I felt that it was time for the excellent work in the Lifters groups to see the light of day. There was some debate about this, but the Lifters group as a whole has greatly benefited from being about to tell people about this new technology.

Marcus had not actually "gone public" about his work. He had leaked information about his research to a few people, and from there it was leaked to several others, who in turn told their friends about it. Unlike the Lifter technology, which has hundreds of people that can talk about how it works and where it came from, Marcus technology was represented solely by himself. That means that everything surrounding his device came back ultimately to him alone. Without any real group ownership of the expertise behind the device

As I said before, Marcus was new to the publicity, and did not have time to prepare the resources that the rest of us have had to answer questions, provide demo materials, etc. Naudin and Saviour have a masterful manner with regard to teaching people about Lifter technology, but they have had years to prepare this information and perfect a method of presenting it to people.

Most of us in the AG groups tend to proselytize for the technologies that we believe in. I know that I do, and if you haunt the newsgroups much you will notice that everybody there believes in something and can go on for months non-stop telling the world about it (again, I do this). Marcus, however, does not really have this personality — he is more of a "take it or leave it" type of person.

For me, this was another indicator that he was the "real deal". You have to understand that I have never seen working prototype pictures of the Marcus device, and all that I had to go on was the feeling that I got from interacting with him on the phone and online. The fact that he did not try to "sell me" on the idea was a big plus. He told me about the device, and his story was always consistent. The other people that he talked to also had a consistent story – I have not talked to anybody and heard a different version than I heard from Marcus himself. That is a big plus from the credibility perspective.

Marcus initial goal (from when I started talking to him in November) was to gradually ease into working with the public to tell them about the technologies he was working on. Unfortunately, I probably aggravated the issue a great deal, because his early descriptions of the technology might have gone unnoticed if it had not been for some of the things he had mentioned about his research.

When I read Marcus' firsts posts on the device, my very first thought about his claims was that they were bogus. Judging from the reaction of the Antigravity newsgroup, I gathered that a few other people also had the same initial reaction. The reason might be a little selfish – after you hear claims but do not see results for a long enough period of time, you tend to protect yourself emotionally by becoming skeptical to the idea that somebody can be lifting hundreds of pounds of weight using a completely new and revolutionary technology.

Nonetheless, something that Marcus had said resonated with me. I remember that he was talking about his results being based on a study of the Earth's fields, which is something that I had read about myself during the college years. Telsa had done a lot of work involving the Earth's electrical and magnetic-fields, and I had always felt that this research might come in handy later as a potential basis for Antigravity research. However, I did not have any specific ideas, and began asking Marcus questions offline to find out more about his research.

In reality, I am nobody special, so I think that perhaps this would have happened in any event no matter what my reaction would have been. Certainly I am not the only person that took an early interest in his work, and from what I learned later he had already been a little bit too honest in the BBC online newsgroup before he ever became involved with the Antigravity newsgroups.

From January to mid-February Marcus became like somewhat of an addiction in the newsgroups. This was a strange phenomenon to watch, but if he did not answer questions in the newsgroups for even a day people started asking whether he had been kidnapped or the government had assassinated him.

You have to understand that most inventors tend to segregate their time into "public" and "private". I do, Naudin does, and everybody does this. When you talk to the public about what you have built, you really get into it — but when you are actually working on it, you tend to hole up in the garage and put the entire world on "ignore" until you have finished with whatever it is that you are doing.

Marcus really started to get publicity in late November 2002, and it built up over a few week period of time. By mid-December, he had a healthy following in the Antigravity newsgroup and people in the 5 or 10 other AG newsgroups were developing more than just a passing interest. Great stuff for Marcus – he had thought that he was the only person doing this research, and it turned out that lots of people shared this common interest with him. We had lots of really valuable and productive exchanges.

Marcus bided his time working on the projects – he had initially stated that he could lift around 200 pounds of weight with his November version (prototype #161, if I remember correctly). He did not have a digital camera, and he did not want to publish photos until he had a better prototype anyways. That was OK with me – I am used to work with a variety of inventors and came to realize a long time ago that everybody has their own style. I was content to wait until Marcus had additional data, and the only thing that I insisted upon was that sooner or later he ante-up and provide some photographic evidence.

Despite lacking photos or video evidence, Marcus did have documentation. In fact, he wrote more documentation about his device than I had ever hoped to receive. He cranked out about the equivalent of 30-pages of "real" documentation in a very detailed form within 1-month, and had the email equivalent of many more pages in postings about how the device worked online. Marcus had already been publishing details online for about 2.5 months.

Despite the incredible amount of published data that Marcus was able to complete, in some ways it made things worse for him than not having anything at all. The reason was the nature of his research. My work, currently focused on Lifters, was in a technology that is easy to explain – the device is built like this, it works like that, and it can be described in a manner similar to some other thing ....basic, easy to explain, and easy to compare.

After all, Lifters have been accused by science of being ion-wind devices. If nothing else, rebutting this claim gives us someplace to start – something to compare against to provide people with an idea about how the device works. Additionally, we had tons of video and photographic evidence to back up our claims. In Marcus case, the documentation that he provided made things worse because his technology is so utterly different. It is not ion-wind, nor the Biefeld-Brown effect, nor apparently the Searle effect. In fact, the more than you tried to pin down exactly what it was, the harder it became to describe it.

Also, its not like he could simply fall back onto describing the construction of the device **– reason being that it contained a series of interrelated coils with a unique configuration that were intrinsically tied to the operation of the device.** Every time he wrote a detailed summary of the construction of the device, we had come away with more questions than we had started out with. After a time I gave up trying to understand how it was constructed, because I only had bits and pieces.

Marcus was using an arrangement of 6-coils, and what are described as "bifilar windings". Essentially, this means that you wind two wires together around a coil and then tie one of the ends to another, making the two wires essentially one long wire wound "against itself" down the length of the armature its on. There are 6 of these armatures on the device, mounted around a central RP (or reference point).

Marcus created all of the terminology for these parts – this lends additional credibility to the idea that he had had 7 years of experience with it. These are ideas that you just do not "make up" for publicity. For instance, the bifilar windings on the armature-coils are self-cancelling fields, and even a basic-electronics education will tell you that. However, the interesting thing is that once you really start to get into the advanced "post-scientific" research on magnetic fields you start to see self-cancelling magnetic coils all over the place. The reason is apparently that while the "B-field" *(Editor: i.e. magnetic field)* of the coils cancels, the "A-field" *(Editor: i.e. vector potential)* that Tom Bearden keeps talking about does not, which means that you can isolate the really interesting effects from the coil without having to deal with the high field-intensity effects from normal electromagnetism.

Marcus had described these six armatures as being activated in a series, or order, to obtain results. Each of the coils was spun by an electric motor mounted on the armature, and when the coil was spun up to a certain speed and "stubs" mounted a periodic intervals on the side of the armature were fired in sequence the antigravity effect was created.

Marcus Antigravity effect was a directional force that was applied in a unique manner depending on which coils were activated. This means that by activating (for instance) coils 1 and 3, he could create an antigravity effect and move the device left – or something to that effect. Turning on and off each set of coils gave him movement on one axis of thrust, which corresponded to X, Y, or Z in a standard 3-D geometric graph. Since you can move in either direction down any given axis of thrust, this means that you have 6 axis of movement height (up and down), width (left or right), and depth (forward or backward). His device provided movement on any axis by either one coil or an arrangement of coils.

For a while Marcus did not want to talk about side effects, but I pressured him into it. The reason was that I have seen side-effects that were pretty darned strange in a number of different experiments, and I was willing to bet that if I pressured him a bit he would talk about what he had seen but did not want to reveal (for fear that people would not take him seriously). It turns out that I was right.....

Marcus had talked about the RP, or "Reference Point" upon several occassions. This is a multi-layered device that acts somewhat like a capactive element. It sits in the direct center of the Marcus device, and it is the reference-point for the entire device. The RP is manufactured from cast-iron, and Marcus says that he has a local company build them for a few bucks each by pouring iron into a mold based on one of his designs.

You can active the RP at the same time that you activate the coils. You have coils rotating around the RP, with "stubs" on the coils firing periodically at points corresponding to spots on the surface of the RP. Meanwhile, you also have an electrical charge on the RP itself, which means that there are a lot of elements interacting at once to create a very complex dynamic.

One of the interesting side effects was a darkening and "blurring" effect of the RP, as if light was being reflected off it. Another was a Star-Trek like "force shield" around the RP. These only occurred when all of the coils were activated at once. Also note that since the coils are opposing in nature (one for each direction on each axis), that when you are creating the force field effect the device cannot be levitating. However, the device is operational, and you have a force-field that you can apparently bounce a heavy hammer off without being able to penetrate it (Marcus description).

Another interesting note is that the force-field effect can be modified to create a vacuum. I am not sure how

Marcus noticed this, but he had said that it appeared to create a vacuum inside of the field's boundaries during operation. Interestingly, it also ran very cold – apparently down to a hundred degrees below zero, but only within the localized boundaries of the field.

I am not a professional physicist, but this is something that has been reported with the Searl effect and several other experiments, and the reason that I believe it occurs is because you are taking the kinetic energy of the device as well as some of the kinetic energy of its molecules (i.e.: heat energy), and injecting it into another dimension. After all, anytime we directly modify a gravitational field we are creating a dimensional effect, but in normal life we do not notice it because the boundaries between gravitational fields are very gradual.

Well, in the Marcus device the boundary between gravitational fields is not a several thousand mile-long gradient like the Earth's field is – instead of you have a gradient perhaps 10-times more intense focused within a 3 to 4 inch area. That means that you are creating essentially a rift in time and space (self-healing) that is kept open by the energy that you are pumping into the device.

the energy input is used only in putting the device into another frame of reference, but that any interaction between us and our "normal" frame would have to be modified for any interactions with the "modified" frame.

To the best of my knowledge, this is the most accurate description that I have seen yet for why his device does this. Marcus himself could not explain why these effects occurred, and to be honest I scared him more than a little when I told him my theory.

In the hypothesis that I just put forth on the origin and functionality of the Marcus device effect, the interesting this is that there really is not a direct correlation between input energy and the effect itself. That is because you are not using the input energy to actually cause an effect – you are essentially using the input energy to translate the device into another dimension (although it is still partially within ours).

Einstein called these "frames of reference". The idea is that the energy input is used only in putting the device into another frame of reference, but that any interaction between us and our "normal" frame would have to be modified for any interactions with the "modified" frame.

Again, this takes place in normal physics – but usually it involves motion, and normally it takes place over a vastly larger gradient of time/space/whatever. The difference between two frames is the difference between two locomotives moving at different speeds – the energy input goes into accelerating each locomotive up to that speed, but the actual work is performed if those two trains interact is solely a product of the frame of inertial reference (mass & velocity = energy).

Another excellent example is a helium-balloon. You put the energy into condensing the helium and pumping it into the balloon, but it is not the helium that creates lift – it is the surrounding atmosphere. In a very real sense a helium balloon gets its energy from the surrounding environment. The energy that you put into "maintaining flight" in a helium balloon is only the strain on the fabric used in holding the helium into the confines of the bag that encloses it. With the Marcus device, I would bet money that it is not the input energy creating these effects – it is instead the difference in frames of dimensional/inertial reference between "our" environment and the "device's" environment, whatever that difference may be.

You see, even though this sounds like a load of BS, magnetism is not real. This is not an idea that I am making up – in fact, it is an entire chapter in the physics 101 textbook kicking around downstairs in my house. Take two magnetic fields, and project them in the same direction at the same speed. Field A can not interact with Field B, because neither of the fields really exists – they are merely ripples on a pond, and that pond is the background of the time-space continuum. People that believe in relativistic gravitational effects take for granted that things like this happen – after all, gravity is not a force per se in relativity – only a modification of the dimensions of time-space. Well, most people do not realize that Einstein based his theory of gravitational force on his study of magnetism. Magnetism is the same thing – it is a modification of the fabric of time-space.

Take two magnetic fields that did not interact from a moment ago, and now change the direction that one of them is traveling in. Suddenly, the two fields that could not even see each other a moment ago now create an incredible "torque" – this is the same thing that we see in bar-magnets and electric motors. This is the "torsion field" research.

Bar magnets are different than pure fields because they are composed of the discrete fields of thousands of tiny magnetic "domains". That is why the fields from bar-magnets are never "invisible" to each other – it is because that the fields in bar magnets are too jumbled to ever really "line up" to the point of being invisible. Pure fields can do this, however, reinforcing the supremacy of Einstein's relativistic effects with relation to electromagnetism.

Anyhow, that is about it for now – I will try to come up with a more detailed picture for you on this if you like, but I expect that you will probably have more questions about Marcus theory and device in near future. It is a very complex idea underlying his research, but the basis of it is the idea of "rotating magnetic fields".

If you want to do additional research on your own into related fields of study, I would definitely look up "relativity and magnetism" – it provides an excellent example of how to visualize relativity theory in the study of magnetic field interactions. This is something that electronics engineers are not taught, which is why you have never heard of this in an engineering sense before. However, in physics is a well-known concept, although most physicists prefer to work with quantum mechanics rather than relativity these days.

In terms of rotating magnetic fields, you can look up "torsion fields" – again, the Russians have the best understanding of advanced torsion field physics, although their best material probably has not been translated yet. Additionally, rotating magnetic fields are the basis for the Searle Effect and the Hamel device.

Marcus was intrigued by Searle's claims, because they appear to have matched many of his own results. Interestingly, Marcus had developed his work within an "intellectual vacuum" and had never heard of Searle before I told him about the similarities. Nevertheless, force-fields, ice-cold operating temperatures, and antigravity seem to be a pervasive similarity between all manners of experiments that have been reported using rotating magnetic fields. I had assumed that Searle's claims were bunk, but after hearing about similar effects from Marcus research it led to reconsider some of the criticism surrounding Searle lately.

One a final note – one group that may lend additional credibility to this research is Godin and Roshchin in Russia. They obviously benefit from the years of knowledge that the Russians accrued in magnetic field systems theory, and they attempted a replication of the Searle effect device some years back (and claimed to get some results). I hear that they have been working on a revised version of their experimental setup recently and hope to have even better results in the very near future.

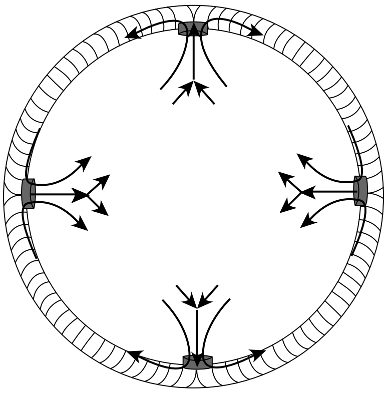

Fig. 1

Magnetic flow for a bifilar-wound toroid that is split at four points on the toroid. This configuration occurs when the coils in the nubs are charged to match the charge on the toroidal coil. The configuration is a quadra pole with a minimum magnetic field at the center

Fig. 2

Marcus recommends the Kikusui 6000L and 18000L multipurpose AC power-supplies

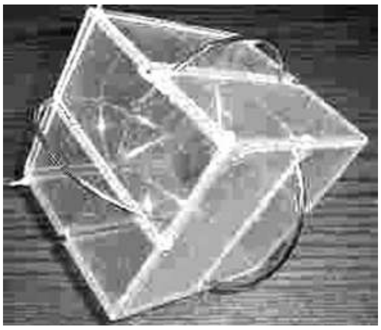

Fig.3

A scale-model mockup of the ring-arrangement within the Marcus device, showing the interlinking coils

History pages

Systems of conversion of thermal energy to mechanical one

Review prepared by correspondent Alla Pashova, **Russia**

There is a wide class of closed-cycle heat engines using differential temperature of water and ambient air. This kind of systems enables thermal energy to be directly converted to mechanical one so as to perform a useful work. We publish patent specifications of two analogous devices of this kind.

PATENT FOR INVENTION

Specification

of the engine, which uses differential temperature of water and ambient air.

To the patent by **P.I. Dobrokhotov** filled on October 22, 1928 (#34248).

Patent grant is published on March 31, 1930.
Patent is valid during 15 years from March 31, 1930.

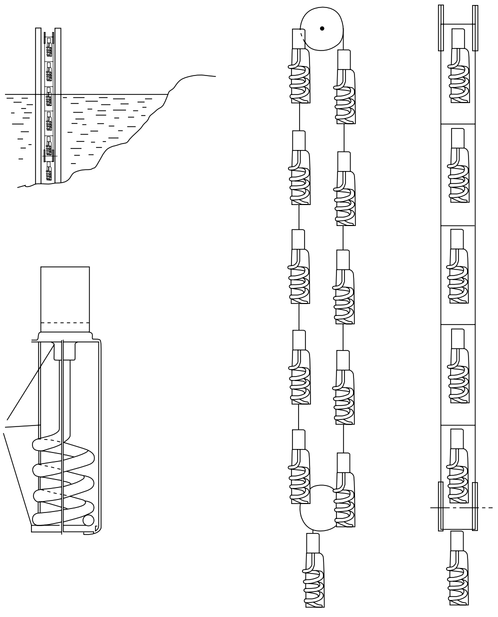

In 1928 **P.I. Dobrokhotov** proposed a design of such an engine. It represents an endless chain thrown over a pair of blocks and partly immersed into water. The chain consists of coiled pipes filled with a volatile liquid (such as sulfurous anhydride or ammonia). One end of each coil is fastened to a frame, while another end is fastened to a cylinder piston mounted on the same frame. When overwater elements are heated by means of ambient air, the liquid inside the coil evaporate. Vapor put pressure upon the pistons which being in motion compress air in the cylinder, and thereby reduce volume of each element. When submerging, the element gradually cools down, thereupon causing a decrease in pressure of vapor of the liquid, which is enclosed within coils. The pistons move in backward direction, and the volume of air available inside the cylinder increases. Increase of volume results in increase of water buoyancy force that sets the device in motion.

In 1980, **Y.S. Grigoriev** developed his forerunner's idea by improving the system described above. The inventor aspired to enhance the effectiveness of this method of conversion of heat energy to mechanical one. He suggested fixing the volume of chambers filled with a heat-sensitive actuating medium at its minimal magnitude before immersing of those chambers into warm water. In water the actuating medium heats and partially evaporates. At a maximum depth of submersion, the chambers are released from fixation to be expanded during emersion. Thus one may avoid expansion of the chambers during their submersion. Futhermore, there is a possibility to activate the device and get the useful work without preliminary spinup from an external drive.

INVENTION SU1404681 A1

23.06.88
Yu.S. Grigoryev
Prototype: USSR patent #13534
F 03 G 7/06, 1928

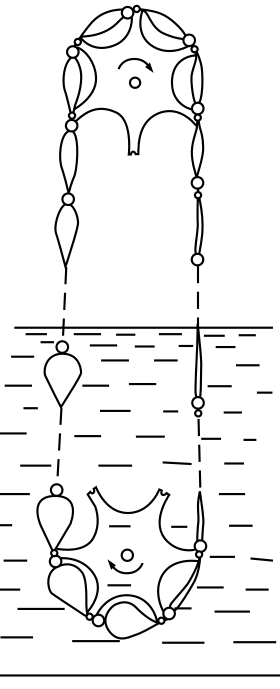

Iliya R. Prigozhin

(25 January, 1917 – 28 May, 2003)

Iliya R. Prigozhin, aged 86, Nobel Prize Winner, Belgian scientist of the Russian origin, died on 28 May in Brussels. Prigozhin made a number of prominent discoveries in the field of thermodynamics and statistical mechanics of nonequilibrium processes. In particular, he developed a concept of irreversibility as applied to thermodynamics. Prigozhin was studying time as a physical phenomenon. In 1977, he was awarded the Nobel Prize for chemistry "for works on thermodynamics of nonequilibrium processes".

The scientist formulated one of the major theorems of the theory of nonequilibrium processes. This theory was later named after him. According to Prigozhin's theorem, stationary state of the system corresponds to minimal entropy generation. The outstanding physicist was also an initiator of applying methods of theory of nonequilibrium processes in biology.

Thermodynamics principles were enunciated in the middle of the XIX century after the invention of a steam-engine, when interaction of heat, electrical, and mechanical work aroused a higher interest. In accordance with the first law of thermodynamics representing an energy conservation principle, energy neither disappears nor appears in any closed system, but transforms from one form to another.

The Prigozhin's theorem sounds as follows: stationary state of the system (under conditions impeding attainment of equilibrium state) corresponds to the minimal generation of entropies. If there are no such impediments, then production of entropy reaches its bare minimum, i.e. zero. The theorem was proved by Prigozhin in 1947.
In thermodynamics the nonequilibrium specifically open systems were of most interest for Prigozhin. In such systems either or both (matter and energy) interchange in reactions with an environment. At that, quantity of matter and/or energy increases or decreases in the course of time.

It is obvious that human society as well as the biological environment is an example of dissipative and nondissipative structures. In the 60s and 70s, Prigozhin developed his theory of dissipative structures and described the formation and development of embryos. Critical points of bifurcation in his mathematical model are correlated with a point, where a biological system becomes consecutive and stabilized.

For his scientific career time, the physicist was awarded over 40 scientific rewards and admitted as an honorary member into national academies of many countries all over the world.

◆

New Energy Technologies

PRINT EDITION ADVERTIZING RATES

Full Page Ad
B/W - $400
Color - $800

Half Page Ad
B/W - $200
Color - $400

Quarter Page Ad
B/W -$100
Color- $200

New Energy Technologies is distributed in the leading Scientific Laboratories in Russia and has subscribers in more than 20 countries all over the world

To advertise contact us:
Faraday Labs Ltd, Phone/fax: 7-812-380-3844
http://www.faraday.ru, office@faraday.ru

Russian Academy of Sciences
Russian Physical Society
Nuclear Society of Russia
Mendeleev Chemical Society of Russia
Moscow Lomonosov State University
Russian People's Friendship State University

The 11th Russian Conference on Cold Nuclear Transmutation of Chemical Elements and Ball-Lightning (RCCNT&BL-11) is to be held during September 28- October 5, 2003. The place of the Conference is in Dagomys near the city of Sochi.

The program of the Conference includes the following subjects:

- Experimental research in cold nuclear transmutation (fusion) and ball lightning;
- Theoretical models with respect to cold nuclear transmutation (fusion) and ball-lightning effects;
- Applied to these problems technologies and devices.

The Organizing Committee of the Conference is pleased to invite you to attend the Conference. The terms of your participation are as follows:

The full cast is $900, which will include the registration fee, hotel reservation and living, three daily buffet meals, conference proceedings, transportation from the Sochi airport and back, social dinner and special excursion or entertainment.

The languages of the Conference are Russian and English.

The total cost can be reduced to $800 if transferred before August 10 to the account of the Organizing Committee, which is to be announced on having received from you a confirmation about your decision to attend the Conference.

If you make a decision to take part in the Conference, please, let us know before July 1 about the title and abstract of your report.

Contact telephone: (7) (095) 196-9476 (ask Mr. Igor Goryachev), Fax: (7) (095) 196-6108
Email: gnedenko@kiae.ru; Bazhutov@erzion.madi.ru; bazhutov@izmiran.rssi.ru; bychvl@orc.ru
Address: I. Goryachev, Russia, 123182 Moscow, 1 Kurchatov Sq.,
Research Institute of High Technologies at RC "Kurchatov Institute"

Letters from our readers

Thank you for publication of my article. Your magazine becomes more and more popular.

My special thanks to Editor-in-Chief Alexander V. Frolov. Whatever they say or write in the press, he has done a great work and it is not only my own opinion. He has managed to do that everybody had been waiting for a long time. My colleagues (not only from Rostov-on-Don State University but also from other Rostov Institutes of higher education) ask me personally send the best regards to the editorial staff of your magazine.

Sergey A. Gerasimov

Free energy?

The scientific community says that it cannot exist. Where did we hear that before? Academician Valery Sobolev from the Academy of Sciences, the head of the research group working on the experiment says: "This is not some new physics, but just experiments that have been performed very seriously. Make surge of electrons from the environment, remove them, deplete the fused electrolyte with basic typical metals; what you derive from it is what cannot be produced by chemical methods. This is a new state of substance." As the researchers say, first they discovered a particular electrochemical process where high-temperature materials in a new state are the product. This in its turn allowed to discover a new state of substance, new class of materials, a new source of energy, a new method of cold plasma generation, a new superconductor. See whole article at: http://english.pravda.ru/science/19/94/379/9675_energy.html

Josef Hasslberger

Personal home page on physics, energy technology, social and economic issues: http://www.hasslberger.com

Interference disc generator with permanent magnets

Inventor Alan Francoeur, Canada

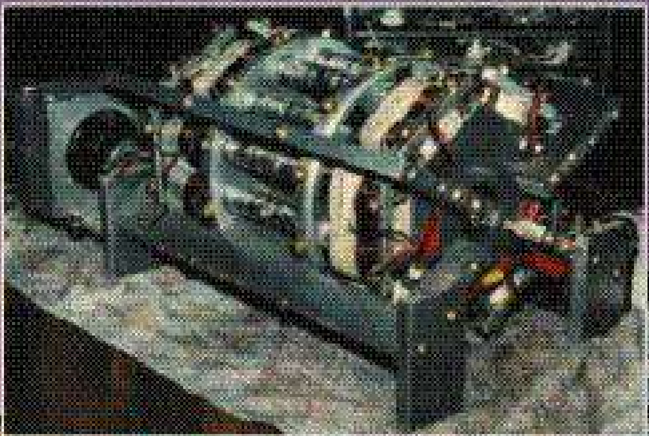

Carl Reich

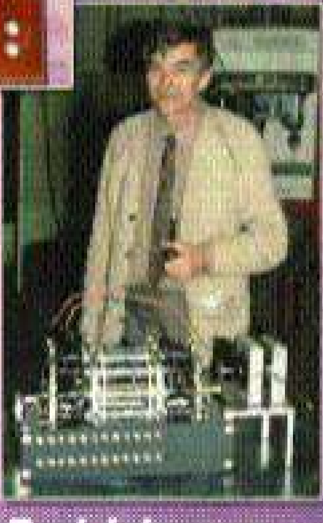

Bud Johnson

Vortex heat pump

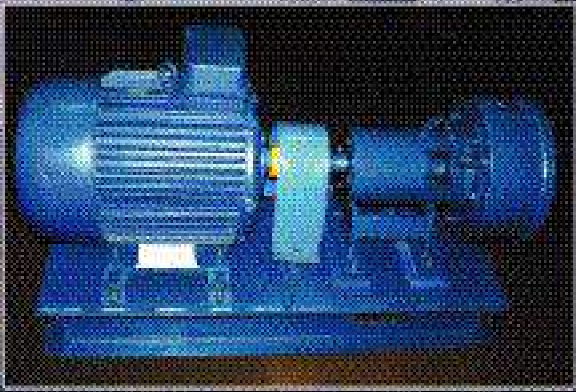

Efficiency 200% and more!

Inventor: Vladimir Susorov
(812) 342-2652
8-901-976-5151
susorow@rol.ru

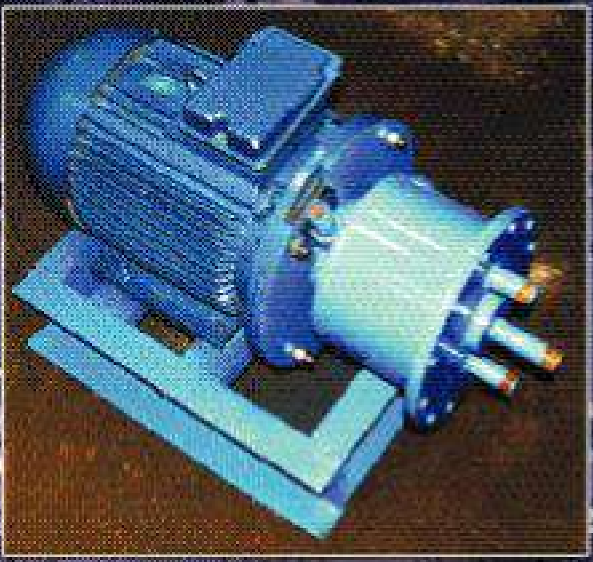

ISSN 5684-7288
9 771684 728803

Gravity-heat power system

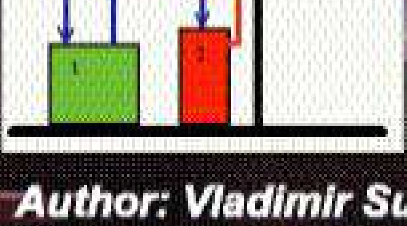

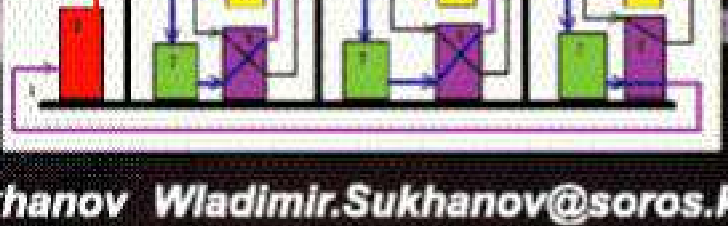

Author: Vladimir Sukhanov Wladimir.Sukhanov@soros.ksu.ru

New Energy Technologies

ISSN 1684 - 7288

Published by Faraday Lab. *Special Christmas Issue #5-6, 2003*

Research on advanced aerospace systems and new energy sources

In this issue:

- atomic hydrogen in power engineering
- magnetic generators
- capillary perpetuum mobile
- Hutchison effect

Photos from Faraday Laboratory: test bench for measuring the effectiveness of hydrogen dissociation and recombination processes

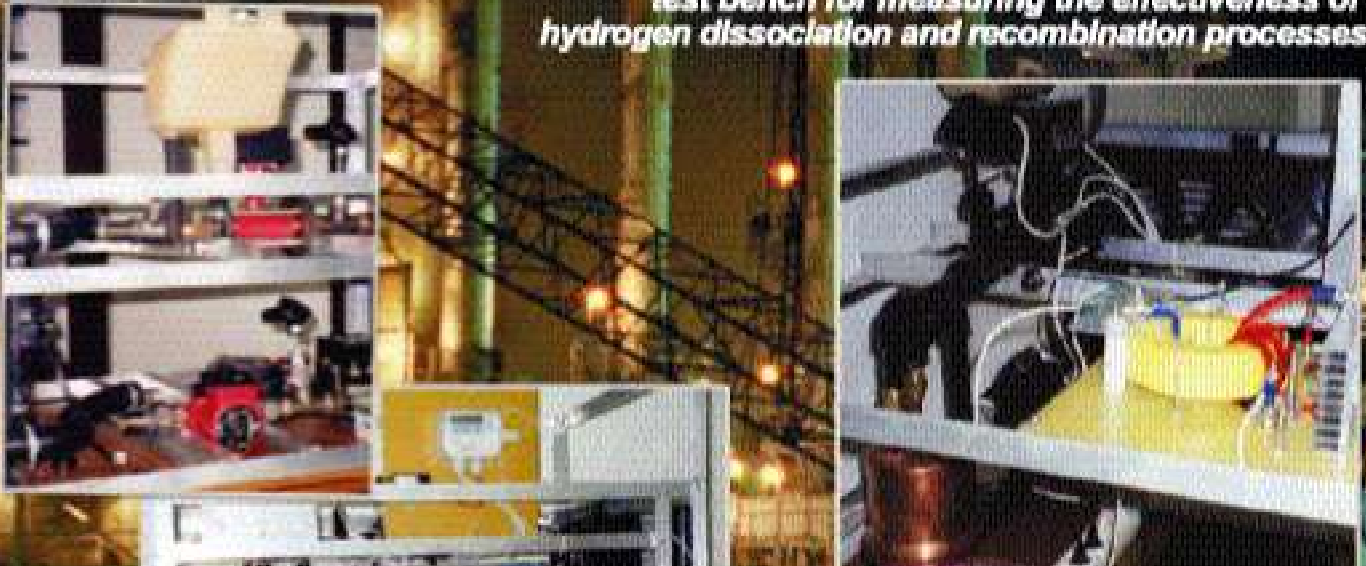

Bogomolov's Invention

Over-unity from elastic forces of spring, compressed by rotation.

Read in this issue!

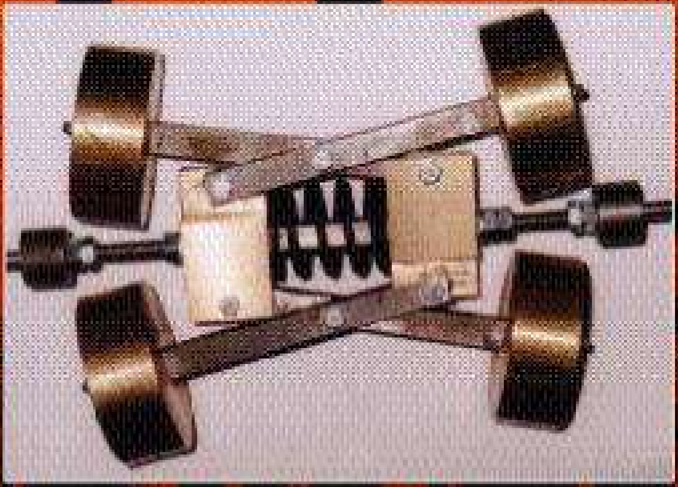

Spin Wave Laser of Ines Espinoza, USA

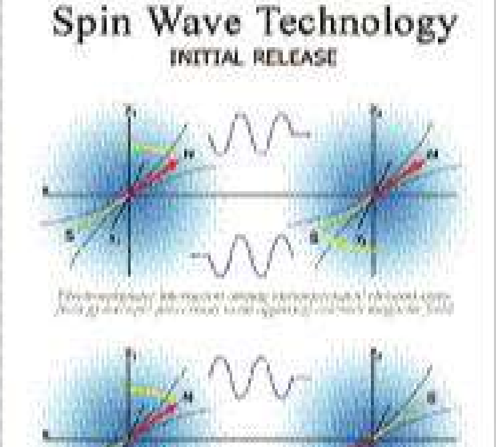

Spin Wave Technology

INITIAL RELEASE

George J Bugh

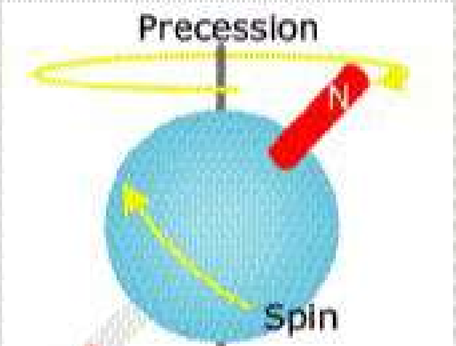
Precession
N
Spin
S
Electron Symbolic Diagram

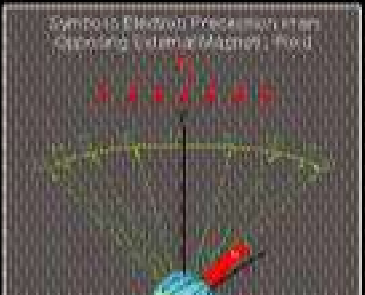

NEW ENERGY TECHNOLOGIES #14-15

1. News from our company
2. Novozhilov's motor, Y. Novozhilov, Russia
3. Self-action in the case of magnetic rotation, S. Gerasimov and others, Russia
4. Capillary motor, I. Elshansky, Russia
5. Gravity-heat energy system, V. Sukhanov, Russia
6. Modern perpetual mobile, Review
7. Free energy and antigravitation, (spin wave lasers), Ines Espinosa, USA
8. Mechanical analogies, Y. Ivanko, Ukraina
9. Cell of fine-plasma electrolyses, Ph. Kanarev, Russia
10. Smith Generator, Review.
11. Strategy of Over-unity Devices Marketing, R. Wood, USA
12. Hutchison effect, Mark Solis, USA
13. Heat energy conversion by means of non-linear dielectric, A Ries and others, Brasil
14. Gates motor, A. Akau, USA
15. "Cold electricity", A.Akau, USA
16. Gravity conversion rotary device
17. Torsion technologies, China
18. News
19. Antigravitation, B.Smith, USA
20. Electric levitation, MacLeod, Canada
21. Perpetual mobile, V. Likhatchev, Russia
22. Interference Generator, A.Francouer, Canada
23. New Paradigma of Time, D. Reed, USA
24. On the rotating permanent magnet system, A.Frolov, Russia
25. Inventor's week, USA
26. Patents on inertial drive topic
27. Bogomolov's experiment with spring drive, Russia
28. History of perpetual mobile
29. Tesla Energy Science Conference, USA
30. Perendev company, Review
31. Internet pages about free energy
32. Matveev's electrical generator, Russia
33. Perpetuum motor with magnets, France
34. The most interesting of 2003
35. Hydrogen power engineering, Review
36. News on hydrogen fuel cells
37. Electric discharge as part of the over-unity device, Review
38. Subscription page

Faraday Lab News

One of the most interesting research projects of our company in the year 2003 has been the research on hydrogen dissociation and recombination processes. It has taken us about half a year to build a special vacuum tube and in October, 2003 we obtained some experimental results on the heat output and efficiency of the system. The system we designed in our laboratory is shown in Fig.1 and also on the cover. The water-cooling system and the heat measurement system make it possible to measure the heat output within the range of 100 W to 5 kW with pinpoint accuracy. In the year 2004 we are planning to go on with our research work.

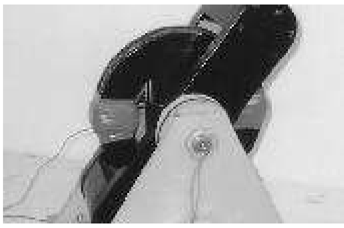

Fig.2

We have completed the second stage of designing and developing a high-efficiency magnetic generator with the neodymium permanent magnets of 1.25 Tesla. We confirmed the possibility of double output increase from the usage of two magnetic induction counter fluxes produced by the two coils wound on the generator core (Fig.2).

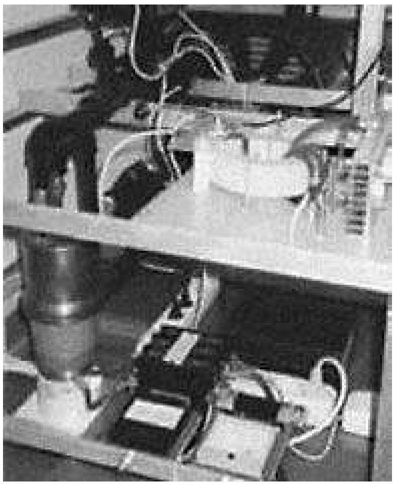

Fig.1

Our laboratory continues designing and perfecting a powerful "alternator"-type generator with permanent magnets and a drum-type rotor. We will give our readers more detailed information after we are aware of the patent examining result.

Our research on "The Control of Temporal Characteristics of Physical Processes by means of Ether Energy Density Change" is also awaiting the results of the patent examining group. This research on time and the construction of the "time machine" is not a pure theory but it has quite real prospects of being applied to the aerospace industry as a propulsion method.

It is worthy to note that now the alternative energy prospects in Russia do not entirely depend on financial factors. The cooperation is important between researchers, investors and organizations interested in the innovative development of Russian industry rather than in the development of Russia as the world primary natural resource is also important. Progressive political and public organizations should develop and introduce alternative energy programmes as well as render support to local researchers. The successful innovation of new technologies is impossible without this support and lobbying government. Let us hope that the year 2004 will bring along new opportunities and promising contacts.

Alexander V. Frolov
General Director Faraday Laboratory Ltd
http://www.faraday.ru

Novozhilov's Motor

Yury N. Novozhilov, Russia
Bratislavskaya Str., 25-6, 390035, Ryazan,
Tel.: (80912) 72-12-72

An interesting but a very simple motor has been designed. It operates on the basis of heating differences of its components due to temperature variations of contacting environments or due to variable radiation intensity.

Motor structure

The motor structure (see Fig.1) includes a metal or plastic wheel (1). In the center of the wheel there is an axis (2), on which spoke ends are fixed (3). Other spoke ends come through a hole in the wheel rim. On the spoke end there are weights, for example, in the form of small balls (4).

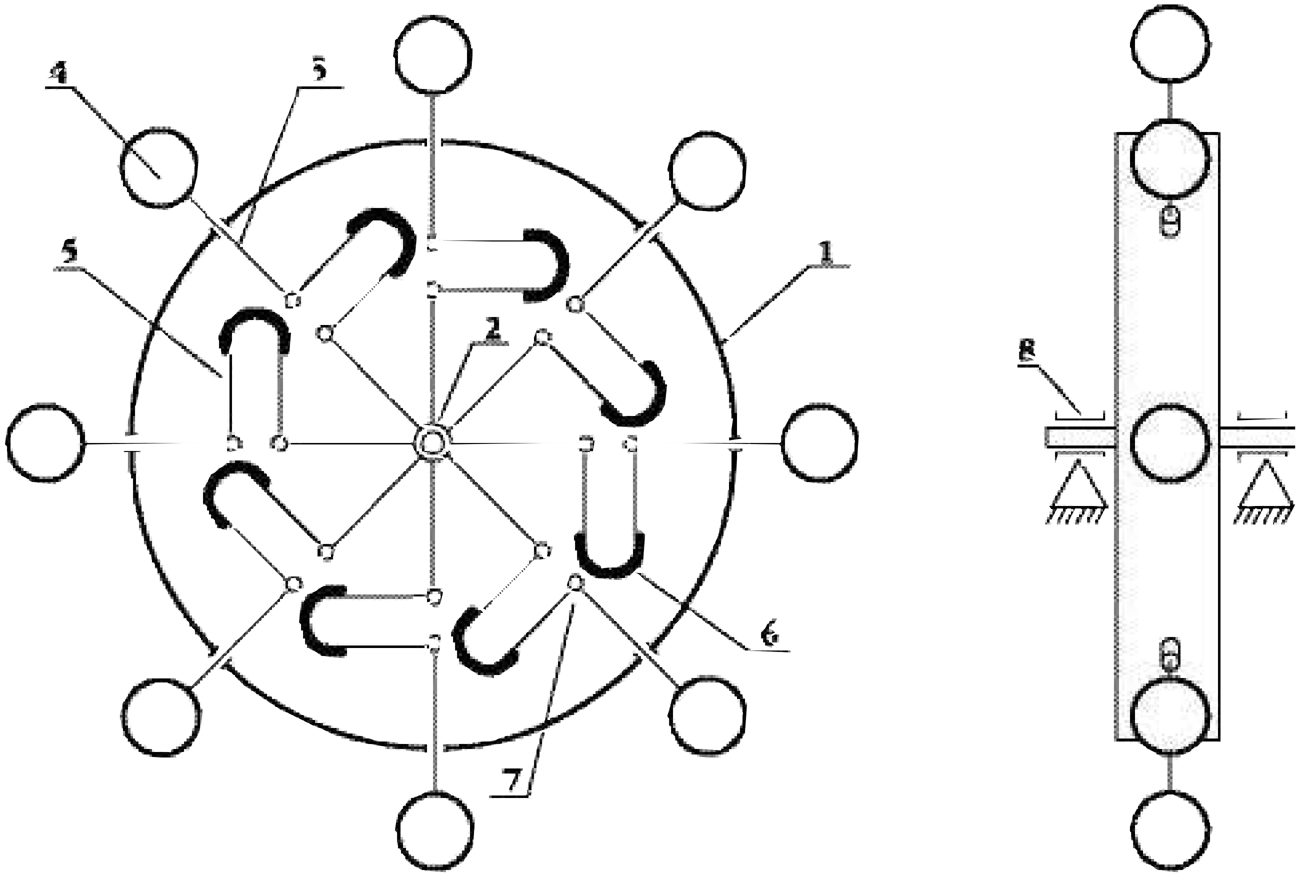

Fig.1
Motor

The structure peculiarity is preconditioned by a loop-shaped part placed on each spoke (5), which has a form of a metal strip or core. An arched bimetal component is built in the center of each loop-shaped part (6). Joints can be fixed at the joining point of loop-shaped parts and spokes to simplify the motor movement (7).

In the motor structure described arched bimetal components become straight when heated. The motor axis ends are placed in bearings (8).

To initiate the motor movement the loop-shaped parts need to be in mediums (zones) with different temperature. For instance, the lower part of the wheel may be placed into hot water (9) and the upper part will stay in the air (10).

Motor operation

When the motor is in a medium with identical temperature, for example, in the air, the temperature of all bimetal components will also be the same; therefore, they have identical form. In this case all spokes will similarly project over the wheel rim and all weights have similar distance from the wheel axis.

This state makes the whole system balanced and immovable. But if the motor wheel is partly placed into water with the temperature exceeding that of the ambient air **the bimetal components in water will become straight and shift the axis weights from the wheel axis.** This phenomenon initiates motor operation.

Let us consider this process in more detail.

On touching water, the bimetal component in the loop-shaped part becomes straight and shifts a weight from the wheel axis. The wheel becomes unbalanced, obtains torque and turns clockwise. This rotation makes the following bimetal component on the next loop-shaped part of the following spoke touch water. This bimetal component also becomes straight and shifts a weight from the wheel axis, generating an increased torque that turns the wheel clockwise.

At the same time a bimetal component comes from hot water on the left side of the wheel in zone “A”. The component cools and restores its initial form, i.e. it curves more and pulls the weight on the spoke to the rotation axis. This leads to a decrease of its counterclockwise torque that affects the wheel.

The loop-shaped parts on the left side (A) and the right side (B) of the motor wheel will cause asymmetry in the number of weights projected by the spokes. On the motor left side (A) the weights will be projected from the wheel axis when the spokes are located in corner “C”. On the right side (B) the weights on the spoke axes will be shifted from the axis when the spokes are located in corner “D”. Fig.2 shows that corner “D” is far larger than corner “C”. This means that more weights will be shifted on the right side of the wheel in zone “B” than on the left side in zone “A”.

Fig.2 shows that the wheel weights generating a clockwise torque are projected on the right side. On the left side in zone “A” only one weight is projected and it generates a counterclockwise torque. More weights shifted from the wheel axis on the right side will generate a greater clockwise torque than the weights shifted from the wheel axis on the left side. This is the cause of the wheel clockwise rotation.

This interesting effect is achieved due to the loop-shaped parts on the wheel spokes with bimetal components placed on them. When the next wheel spoke reaches corner “D”, the bimetal component connected with it touches hot water, becomes straight and shifts the corresponding weight further from the wheel rotation axis. The motor wheel continues rotation, making the next bimetal component touch water. It is also heated, becomes straight and shifts the corresponding weight from the wheel rotation axis. The wheel continues rotation. Bimetal components are touching water in succession. On the left side, in zone “A”, the bimetal components successively come out of hot water, cool in comparatively colder air, return to the initial sharply curved form and pull the weights on spokes to the wheel rotation axis. This is the reason for the wheel torque to decrease.

Mass movement from the axis in a rotating object commonly causes reduction of rotation speed. And, on the contrary, the rotation speed increases when the mass comes close to the axis. In motor zone “A” the weights come close to the rotation axis after the corresponding bimetal components come out of hot water, and the same number of weights move from the center in zone “B”, i.e. these effects on the rotation wheel compensate each other.

In essence this is a heat machine. The heat transfer from a warmer medium – water – to the cold medium – air – is achieved owing to the heat capacity of the band-form bimetal component. This is the reason for fast heating and cooling processes. Due to the fact that bimetal components are not large, not so much heat is needed for their heating.

The motor will operate when the air temperature exceeds that of water but in this case it rotates counterclockwise.

There is another model of this motor when nitanol components are used instead of bimetal ones. Nitanol is a memory metal alloy that changes shape when reaching certain temperature. In this instance, the nitanol component is normally curved but becomes straight when heated (for example, up to 50^{0} C), and the weight on the spoke end moves from the wheel rotation axis initiating rotation of the wheel. That means similar operation of a nitanol-based and bimetal-based motor.

The motor structure is very simple and its operation does not require power or fuel supply: it is enough to place its wheel into the water the temperature of which is different from that of the air.

The motor structure may be further developed: for example, a focused sun ray can be used for the heating purposes; the bimetal component can be heated with a common electric lamp on one side of the wheel, the required heat may come from a heating radiator, stove or burning gas-jet. Such motor may be heated when installed on a mantelpiece, etc.

This article describes the motor model that can be used as a toy. Being supplied with looking-glass units it may be applied to get various luminous effects or may be installed in the shop showcases, or else at exhibitions. This motor is also able to operate in a gravity force environment, even on the Moon. There direct sun rays may heat bimetal components to 200^{0} C, and the temperature of the components located in the shadow almost reaches absolute zero, i.e. the temperature difference in this instance is over 400^{0} C, which enables constant efficient motor operation.

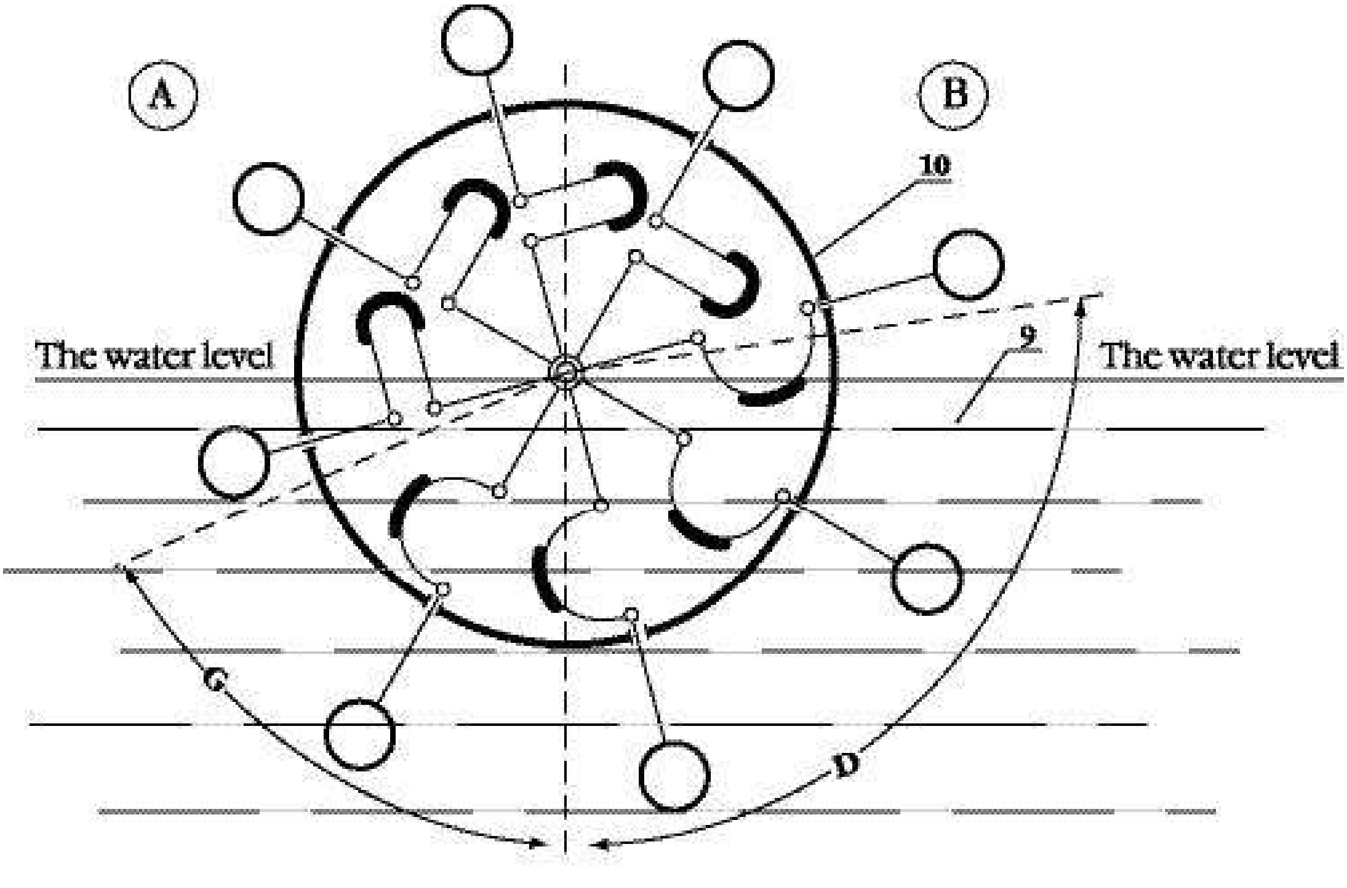

Fig.2
The motor in operation

Self-Interaction in Electromagnetic Rotation

S.A. Gerasimov , V.V. Stashenko, Russia
Physics Department, Rostov State University, Rostov-on-Don, 344090
Email:GSIM1953@mail.ru

Introduction

The electromagnetic rotation is the motion of a magnetized body ina direction perpendicular to both the vector of magnetization J_m and the direction of the electric current j in a liquid inside which the magnet is located (Fig. 1).

Editor: Readers can compare this vector composition and the device scheme of Godin and Roschin with the Searl experiments. The similarity is evident. *A. V. Frolov*

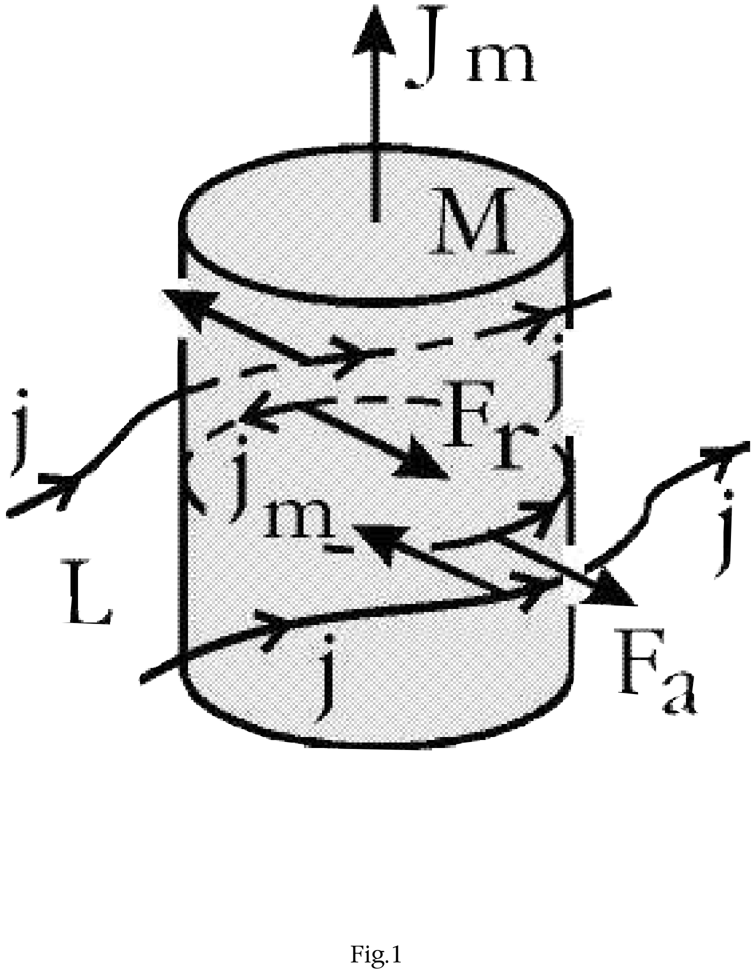

Fig.1
Electromagnetic rotation. A magnet M of magnetization J_m that is partly or entirely submerged in a conductive liquid L with a direct current of density j is moving in the direction of the force F_a+F_r

The force producing such a motion is sufficiently weak and, normally, the phenomenon is observed at rotation. Although, this is quite a well-explored phenomenon [1-6], nobody knows what makes the magnet rotate [4]. There is an opinion [2,3] that the motion of the magnet in the conductive liquid is caused by the forces of attraction $\mathbf{F}_a$ and repulsion $\mathbf{F}_r$ between currents of density $\mathbf{j}$ flowing in the liquid and surface magnetization currents $\mathbf{j}_m$ of the magnet, often called the Ampere currents as shown in Fig. 1.

The external magnetic field of a long cylindrical magnet originating far from either pole is zero. Therefore, no force is exerted on charges moving in the liquid far from the ends of the magnet. Nevertheless, the magnet still moves in the liquid with the direct current even if it is only a thin magnetized needle [5]. Thus, the deeper the magnet is immersed in the conductive liquid, the more liquid gathers around it. The deeper the immersion is, the higher the total force must be. That should be so.

In reality, it is quite different. During intermediate immersions when the magnet is partly submerged in the liquid, the torque N does not depend on the depth of immersion, in the other words, currents flowing round the cylindrical magnet do not influence the exerting force acting [6]. Moreover, during shallow immersions, the magnet moves together with the liquid. **This is none other than *the self-interaction* [7].** The conductive liquid under the magnet experiences the magnetic action. The direction of this force coincides with the direction of the magnet motion. **Thus, the part of the conductive liquid moving in this direction under the action of the magnetic field B, affected by the frictional forces in the liquid causes the motion of the magnet in the same direction.** The only way to find out the role of the self-interaction in this phenomenon is experimental investigation. It would be appropriate to exclude the currents flowing under the magnet. In this case the magnet must rotate in the opposite direction.

Experimental system

A commercially available cylindrical magnet M of magnetization J_m=1,95•10^5 A/m and a balanced load *P* connected by a rocker R are suspended by a thread *T* as shown in Fig. 2.

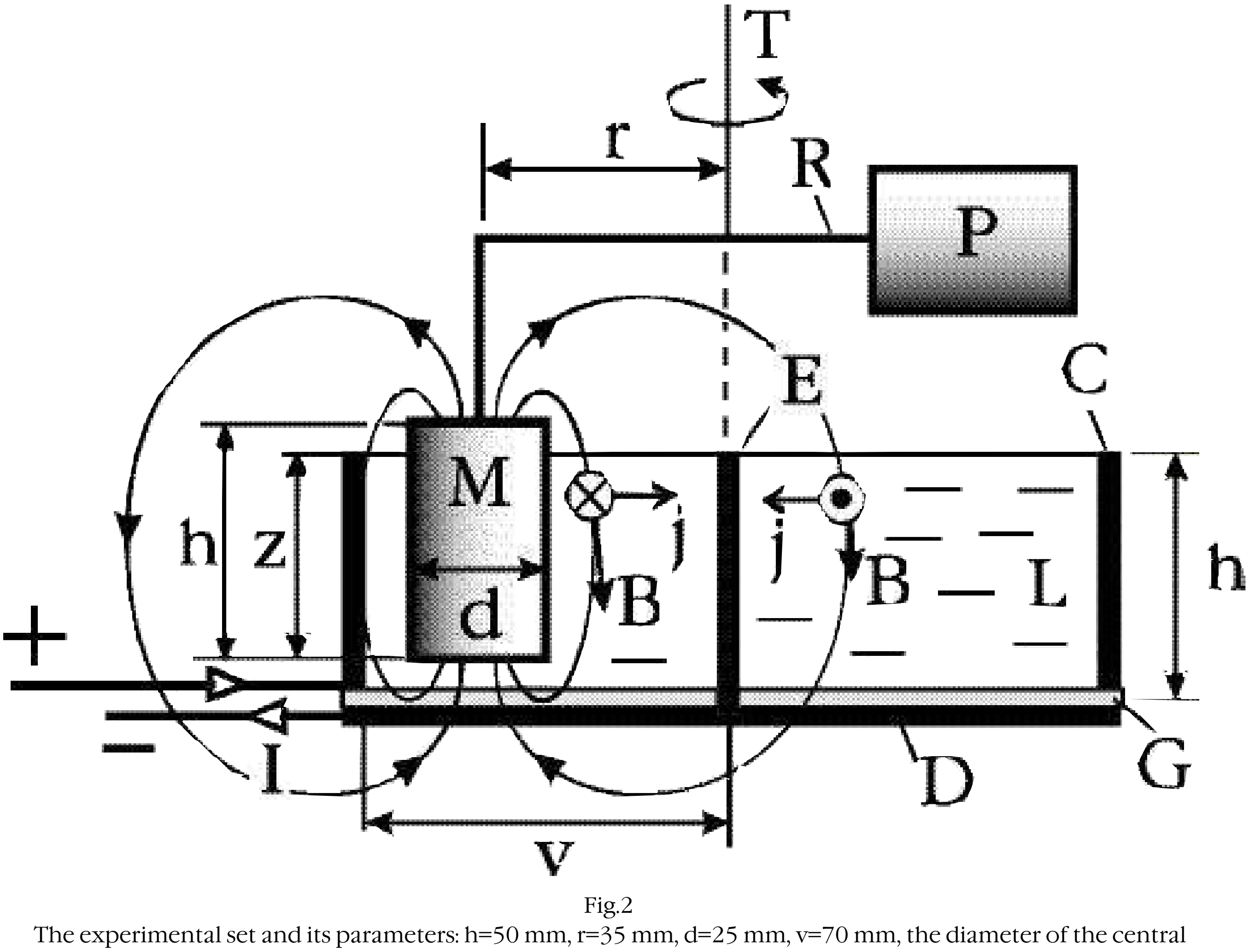

Fig.2
The experimental set and its parameters: h=50 mm, r=35 mm, d=25 mm, v=70 mm, the diameter of the central electrode E is 5 mm, the thickness of the bottom G is 2 mm. (×) and (·) are directions of the force acting on current elements of current density j in the magnetic field of inductance B

The magnet is immersed in the 5% copper sulfate solution ($CuSO_4$) so that the depth *z* of immersion can vary. A vessel containing the conductive liquid *L* is large enough for the magnet to be almost entirely immersed. The cylindrical surface *C* of the vessel and the central electrode *E* of the system are made of non-magnetic materials. The bottom of the vessel *G* is, of course, insulator. The magnet is coated with an insulating moisture-proof varnish. The length of the electrode *E* equals the height h of the vessel. A direct current of strength *I* passing through the cylindrical electrode *C*, the conductive liquid *L*, the central electrode *E* and the disk electrode *D* generate a torque of electromagnetic origin which results to the rotation of the magnet. The disk electrode *D* is intended to provide the system with the symmetrical supply of the current.

Experimental results

The size of this experimental device differs from that described previously [6]. **As a result we have obtained the torque N, which is about ten times more intensive than the former one.** But the main result is that the torque does not change its sign when the magnet is being immersed in the conductive liquid (Fig. 3).

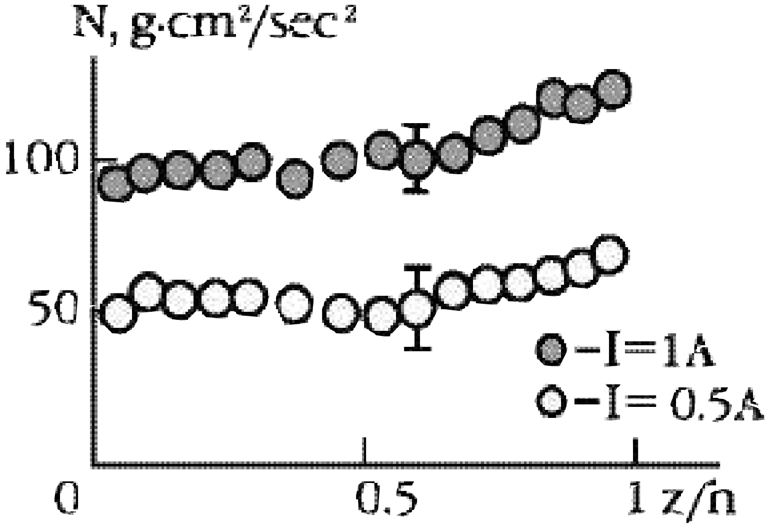

Fig.3
A typical experimental dependencies of the torque *N*
acting on the magnet depending on the depth of immersion *z*

When the magnet is almost entirely submerged in the liquid the depth of which equals the height of the magnet, no currents are flowing under and above it. In this case the rotation of the magnet is expected to be the result of drag action of the rotating liquid located in the intermediate field relative to the magnet. This part of the conductive liquid is rotating in the direction opposite to that of the motion of the magnetized body. It seems to be the right way. The magnet pushes the liquid away and, therefore, is moving in the opposite direction. No, this is wrong. If it was correct, this effect would also occur during the shallow immersion. At small *z* the magnet and the liquid rotate together. So far we still do not know what makes the magnet rotate.

S.A. Gerasimov

References

1. Faraday M. Experimental Researches on Electricity. // London: Ed. by R. and J.E. Taylor. 1839.
2. Ampere A.M. Theorie Mathematique des Phenomenes Electrodynamiques. // Paris: Blanchard. 1958.
3. Sigalov R.G., Shapovalova T.I., Karimov H.H., Samsonov N.I. // Magnetic Fields and Their New Applications. // Moscow: Nauka. 1976.
4. Gerasimov S.A., Ershov A.V. What Makes a Magnet Rotate? // Physics Education (Moscow). 2002. V. 22. P. 70-75.
5. Grabovsky M.A., Mlodseevsky A.B., Telesnin R.V., Shaskolskaya M.P., Yakovlev I.A. Lecture Demonstrations on Physics. // Moscow: Nauka. 1965.
6. Gerasimov S.A., Ershov A.V. Mechanism of Electromagnetic Rotation. // Problems of Applied Physics. 2001. V. 7. P. 34-35.
7. Gerasimov S.A. Self-Interaction and Vector Potential in Magnetostatics. // Physica Scripta. 1997. V. 56. N 3-4. P. 462-464.

Capillary Motor

Iosiph I. Elshansky, Russia
Koshtoyantsa St., 1a - 75, 119454 Moscow
Email: ewrik@bisinter.ru

For several decades the author has been organizing a contest for young inventors on the radio, television, in mass media and the Internet. There have been accumulated thousands of engineering solutions that belong to schoolchildren and youth. There are many original ones, including perpetual mobile. Though unfeasible in principle, they have interesting structure and competent engineering solutions.

For example, a child's dream to create an oblique roof over a city or field and direct the rain flow to a turbine to obtain mechanical work is of little use. But it's quite another matter if this child takes a pencil, a sheet of paper and tries to make a sketch or a drawing of such construction. Even if the construction is primitive and unfeasible, the child will get his first engineering experience and self-reliance.

I can still remember one peculiar case. Once the editors were instructed (it was in the times of the USSR) to assist an honored inventor of new machines. It was a Stalin Prize laureate, creator of a special type of lime that was widely applied in construction (I will not disclose his name). The inventor was given (that was not common practice) a large room in a shared apartment situated in a prestigious block of apartments in Moscow. It turned out that all room space, except for narrow passages at the walls, was occupied by a huge table. On the table up to the ceiling there were some pipelines, glass and metal vessels, retorts and other laboratory ware. The most impressive thing was retorts heated by gas coming through a hose from the shared kitchen. All that staff was bubbling, hissing and steaming. The inventor was creating a perpetual mobile!

I will give one more example of a situation common for the editors of the "Inventor and Rationalizer". Once we had a respectable-looking visitor who made an odd request to get the President of the French Academy of Sciences to respond to his letters about a perpetual mobile plan. The essence of Alexander Rodionov's (Maloyaroslavets, Russia) invention was that "in accordance with Newton's and Guerin's laws fluid flows up through capillaries and turns a wheel when flowing down (see Fig.1).

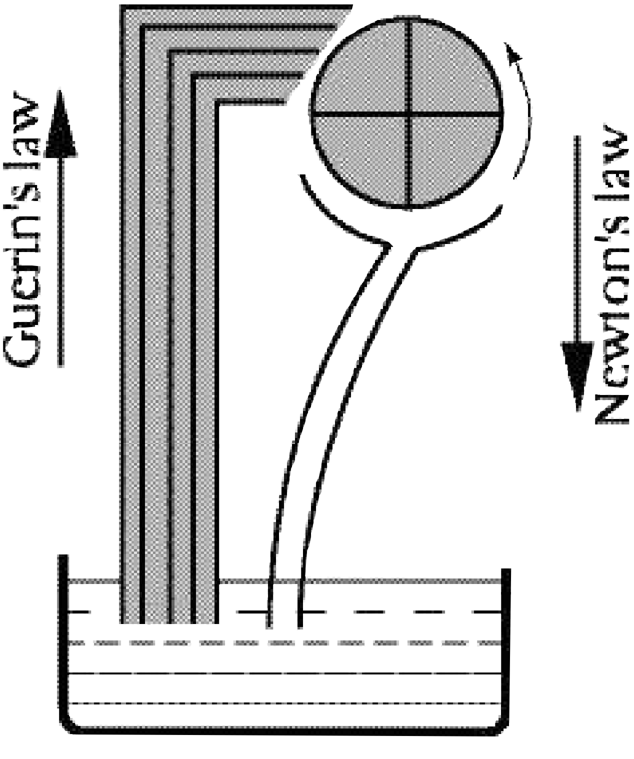

Fig. 1
Perpetuum-mobile

There are many similar projects in the history of engineering. Hence, this case may be considered as typical. We all know that the molecular adhesion forces (moistening) that push the fluid upwards will stick to it and the capillary motor will never work. But is the idea totally unfeasible? Since false modesty is inappropriate when it comes to perpetual mobile, I will present my ideas on this not so honorable subject for the reader's judgment.

It seems that all perpetual mobile inventors start not with a pencil and a sheet of paper but with experimenting on magnets, oblique plane, wicks, wheels, springs and other materials at hand. My first heat (perpetual) motor was also the result of an experiment.

The "Capillary" idea of Alexander Rodionov haunted me. I recalled that the great Lomonosov devoted lots of time to studying molecular adhesion and capillarity. They are present in everyday life, let alone the nature, especially in plants that would not be able to exist without them. How else would moisture flow up the trunks and stems? But, on the other hand, according to Mikhail V. Lomonosov's data, water flows up only for the maximum of several dozens of millimeters even through the thinnest capillary. But trees may have the height of several dozens of meters! If, as it is customarily considered, moisture flows from one capillary to another of its own accord, why not suppose that a capillary perpetual mobile is feasible? Explanations that moisture in plants flows up due to root pressure do not sound plausible. So where is the truth?

Years have passed since that visit of A. Rodionov. Now, just like I used to sympathize with, as it then seemed, a rather odd inventor, now my relatives sympathize with me. One can not blame them: for weeks and months there have been flowerpots all around the apartment - on the tables, window-sills and bookstands. With bottles, tubes and self-made water manometers to follow.

Once, when I was assembling a device, I could not find two similar glass tubes and had to put in the tube of thin polyethylene. But no matter how hard I worked the water in the communicating vessels was at different levels. In the glass tube the water level was always higher. It could not be otherwise, if you come to think of it, but isn't it the right time the words "made of equally wettable material" were inserted into the communicating vessels law?

Here is a primitive experiment: make a hole in the bottom of a plastic flowerpot containing soil and a plant, and insert there one end of an elastic tube. The other end is inserted into the bottom of a plastic bottle containing water (see Fig. 3). In accordance with the communicating vessels law, the water will flow from the bottle into the flowerpot. When we see that the soil has been made wet, the experiment shall get more complicated: elevate the flowerpot so that its bottom is higher than the liquid level in the bottle. In some days we shall notice that the soil has not dried up and the water level in the bottle has lowered. One should not be a Solomon to guess that the soil or the roots of the plant transfer the water from the lower to the higher level. Here is a typical natural perpetual mobile!

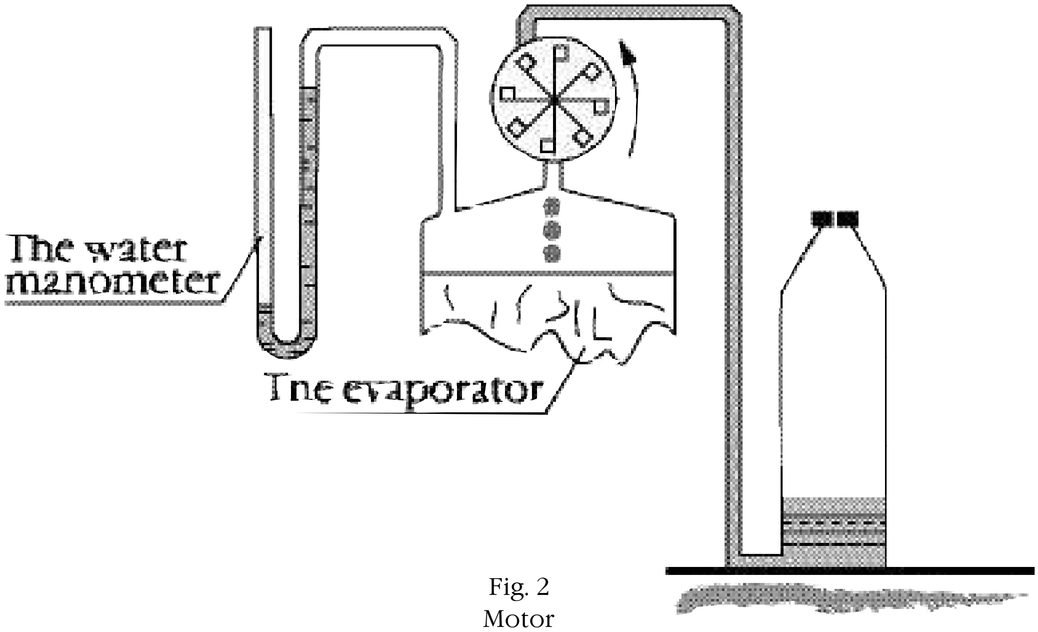

Fig. 2
Motor

It is no use referring to the molecular adhesion forces since they can not be applied at such height differences. What is it then? A molecular collision as in a boiling kettle? But a leaf has neither high temperature nor much space. Nevertheless, water evaporates both from the soil and leaf surfaces. At the same time there is rarefaction under the evaporating surface. Can it be the rarefaction that pulls up new portions of water?

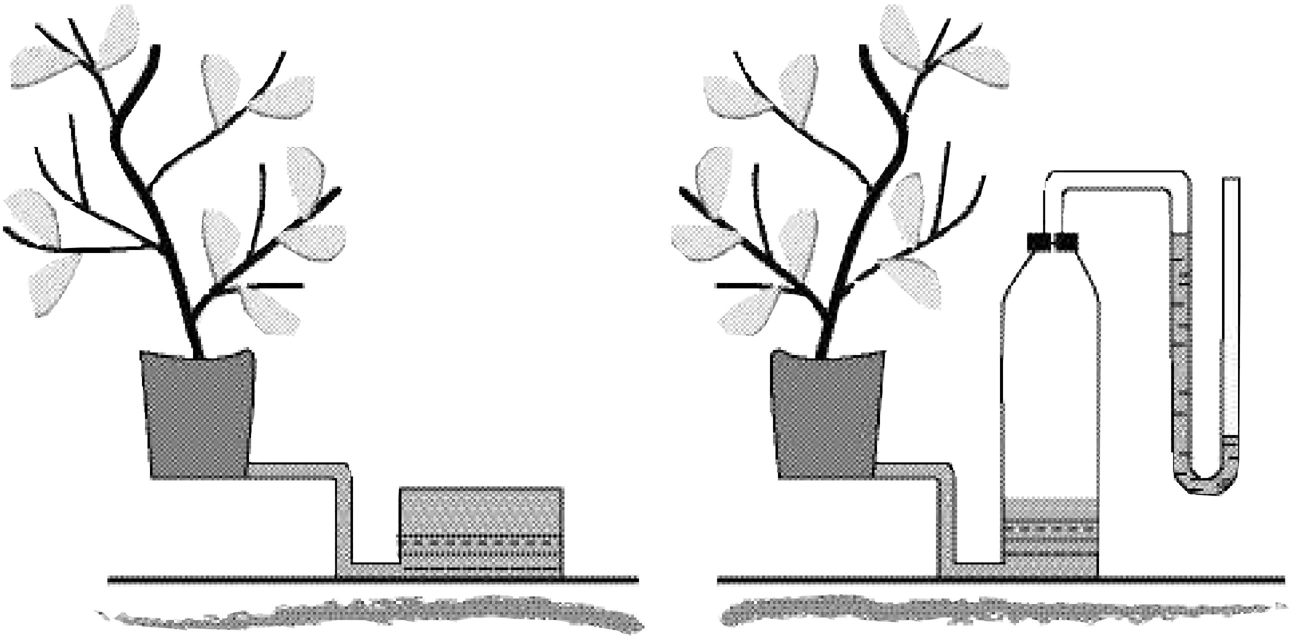

Fig. 3
Rarefaction

Let us make the experiment still more complicated. Put the leg pipe of a liquid manometer into the bottle cork. (Fig. 3) Almost immediately the water level in the closest leg to the bottle will start rising. It means that there is rarefaction in the bottle. Can it be the rarefaction that "rules" the plants?

Most likely, the essence of the mistake made by Rodionov and other authors of capillary motors consisted in attempts to get the water outflow from the capillary. **If it is not flowing out but evaporating, as it is common for the soil or plants, then, probably, vacuum will manifest itself and the capillary perpetual mobile will start working.**

I made my first capillary perpetual mobile many years ago (see Fig. 4). And it worked: though slowly, the wheel was steadily turning. No doubt, it was due to the heat inflow from the air. Therefore, such motor will work if there is a temperature difference and relative humidity of surrounding air is below 100%.

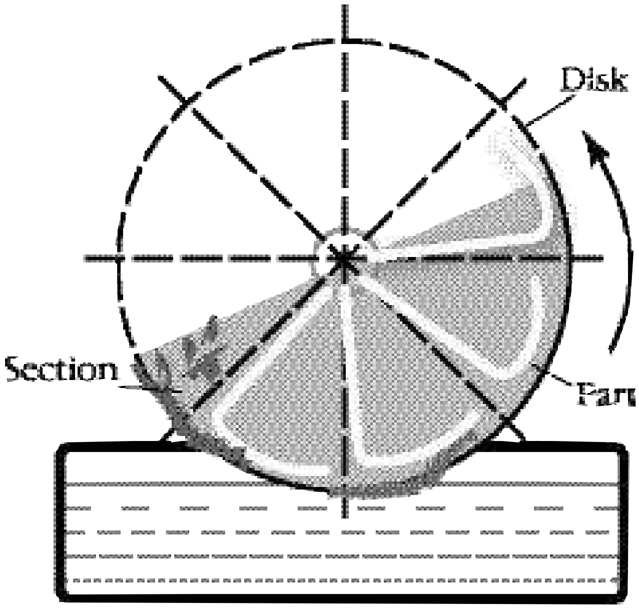

Fig. 4
Capillary Motor

The heat motor operating principle is clear from this figure. The disk (many disks on the common axis) is divided into isolated sections. Their surface is covered with a moisture-proof piece of material (filtered paper, cloth, etc.). The part of each section somewhat gets under the base of the next one. Hence, a small part of the base of one section and a part of the part following the section can get into water simultaneously. Due to the capillary moistening the section and the part will start to absorb water. Since the section is shifted about the center of gravity, the turning moment is bound to appear. The section that is out of water starts to dry up, the moment will increase and the wheel will start turning.

Having determined that, though slowly, the motor is turning, the author applied to a patent authority for a patent on the invention of the "Perpetual mobile", meaning it as a joke. I hoped that the experts (I knew some of them personally) will understand this joke. But, unexpectedly, the device was considered an invention, though under the name of the "Heat motor" (certificate of authorship USSR # 1455040).

Basically, the motor is similar to the project of A. Rodionov. I would not be surprised if he protested. The same capillaries and the spontaneous rise of liquid on the moisture-proof material. The only difference is that the liquid does not flow out but evaporates from the capillaries.

Almost without any reason - one can not predict his eccentricities - I have recently made one more capillary motor (see Fig. 3). Perhaps, it resulted from my experiments with plants. This motor directly and spontaneously lifts water to a higher level, which is more convincing. Moreover, water flows in the form of drops (or a stream) and turns the propeller! The motor is quite operable and has even been shown on TV. But do not try to verify it on your own - such work is useless without certain know-how. The author would not like to disclose the "secret" yet. I could still make a patent application.

NEWS

The Continuous Rotation Device

The essence of this invention is that in a closed circle the rotation of the object occurs without any outer influence or any power sources (electric energy, oil, diesel oil, etc.), which results in the rotation of the rotating part of the device. The torque can be easily enhanced by means of transmitting it to the reduction system. At that rate, the necessary rotation speed and rotation power can be acquired.

The device is small, easy to transport, mobile (both, in operation and switch-off), simple and inexpensive to make and assemble. It is economical, ecological and weather independent.

The device can be used in any industrial or household appliance field where the continuous rotation is required. It is especially important to note that this device can be used to generate energy on a large industrial scale as well as in small laboratories, in mountains and remote districts, in tunnels, at sea, etc. for it does not require any power lines.

Patent Claim Priority:
#20030059, April 19th 2003

For more information
contact
the author of the invention:
Eldar Sariyev,
Azerbaydzhan, Baku,
Tbilisskiy Prospekt, 75-26;
Tel: + (994 12) 92-47-73, 98-95-02;
Email: eldar_sariyev@yahoo.com

Gravity-heat Power System

Vladimir N. Sukhanov, Russia
Kosmonavtov Str., 35-39, 420061 Kazan
Email: Wladimir.Sukhanov@soros.ksu.ru

Natural hydrologic cycle is a widely recognized phenomenon. Water evaporates from the ground and water reservoir surfaces, generally, due to the Sun heat, then moves to the superstructure atmospheric layers where it is condensed, and precipitates onto the ground surface, but on higher levels compared with evaporation surfaces. In this process water is also purified and desalted. From the higher levels water flows to its main evaporation points, forming streams like brooks and rivers where water power plants could be established. Flowing to the lower surface layers with lower potential energy, water in the Earth's gravitational field performs work that can be utilized.

Natural hydrologic cycle principle is used in the gravity-heat power system (USA patent No. 3953971, International class. F03G7/04 of May 4, 1976).

In this system the temperature difference between ground surface layers (the mountain foot and its top) is used to produce electric power and obtain fresh water. The temperature difference does not change much with change of height, and the system effectiveness is quite moderate, which can be named as its drawback.

The system utilizes free environmental energy. Its efficiency (converted to the process maintaining energy) tends to infinity.

The author proposes a new gravity-heat power system that can substitute modern heat power plants. It contains all installation components as per USA patent No. 3953971 except for the heat exchanger on the ground surface whose functions are performed by a conventional steam boiler.

Gravity energy component **G** in the proposed system:

$$G = X_m \cdot T\, g\, H\,,$$

where

- X_m – fluid flow,
- **T** – time of operation,
- **g** – acceleration of gravity,
- **H** – tower height between the steam boiler and the freezer (condenser).

This formula does not allow for vapor density of operating fluid.

The vapor density is insignificant as compared with the operating fluid density.

Energy **P**, required to maintain operation:

$$P = X_m \cdot T\, q\,,$$

where **q** – specific fluid vaporization energy

Efficiency **K** – of the system should not be lower than 80%

To obtain efficiency that equals 120% the following is required:

$$H = 2{,}2q/g$$

Or **H** = 25,4 kilometers.

Construction of such towers is too complicated for modern technology. Therefore, a stratospheric balloon that is linked to the ground surface by a flexible twin-core hose rope may be used. One core is designed to send down fluid and the other to send up vapor. Flowing down to the ground surface through the hose, the fluid will generate fluid column pressure. The fluid will cross water-turbine generator cascades, after each of which the fluid pressure will decrease and the fluid will continue its way down.

There is an alternative. To make the tower several times lower the evaporation (at the tower foot) and condensation (at the tower/mountain top) should be phased (in the form of cascade) at various pressure values and with utilization of the same heat energy. For this purpose a heat carrier is used that will transfer heat from the tower (mountain) top to its foot. This operation will enable the tower height segmentation into **H/n** sections where **n** is the number of grades in a cascade.

Cascade structure is shown in Fig. 1

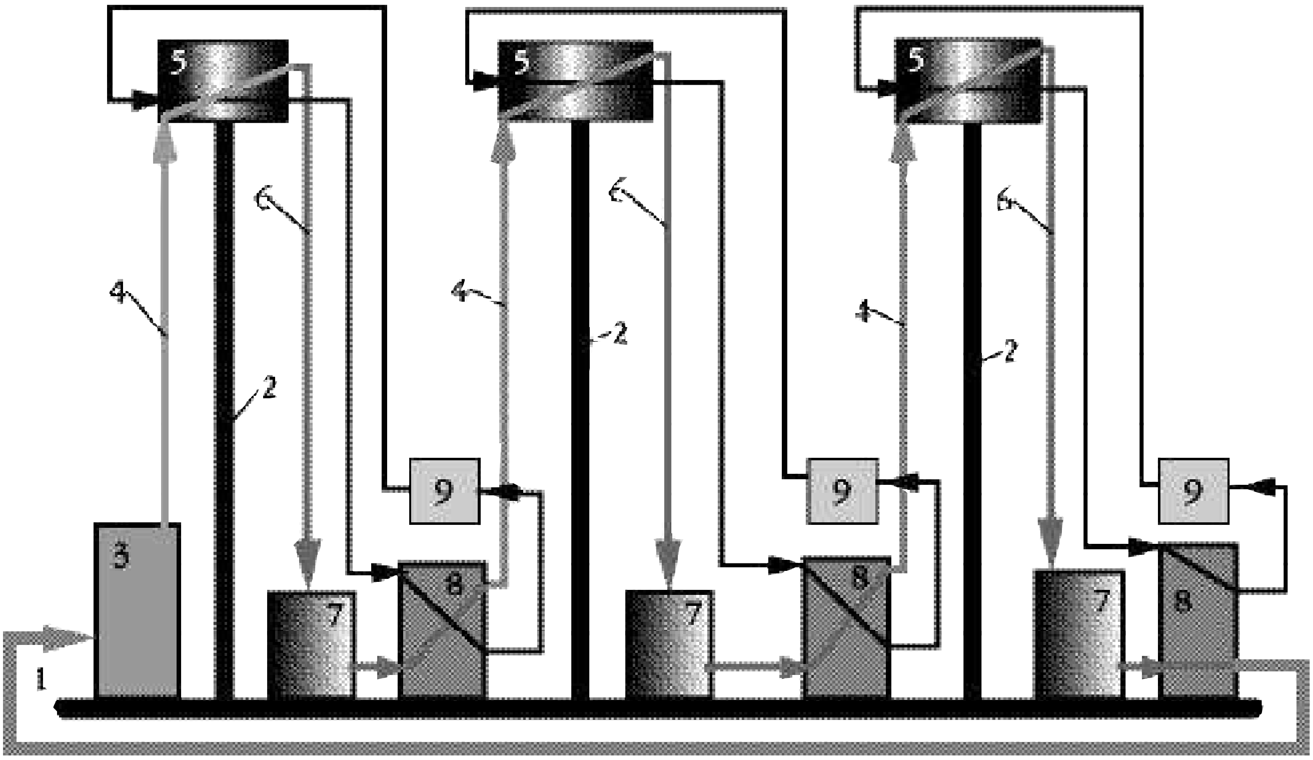

Fig. 1
Cascade

where

1 – ground surface,
2 – tower component or the smoke-stack of the steam boiler,
3 – steam boiler,
4 – steam pipe,
5 – heat exchanger and condenser,
6 – pressure pipeline,
7 – water-turbine generator,
8 – heat exchanger and evaporator,
9 – circulation pump,
10 – pipeline of the heat carrier circulation.

Cascade (operating fluid, water) with 500 meters of height can contain several dozens of grades (50 and more). **In this case the cascade efficiency may approach 120%.**

Decreasing the tower height to 500 meters (for water) will allow for utilization of conventional water-turbine generators and standard equipment.

In the proposed cascade the heat energy is required to maintain the cascade working capacity when Earth's gravitational field energy is utilized. In the author's opinion there are still other ways to use gravitation.

The system scale is its main drawback. Therefore, partial utilization of the proposed principle may seem more attractive. The proposed engineering solution was simple.

In a thermal power station the condenser is installed on the top of the tower (mountain). After the condenser the water lift in the water tower finished at the steam boiler pressure pipe (omitting the pump that was no longer necessary). Energy required for pressure pipe operation was saved. In other words, the water tower was utilizing the Earth's gravitation at the conventional juncture of vapor condensation in the steam plant.

Such proposal was made on June 17, 1982 in the patent application USSR N3453603/06, (101161) expert code 060701KH. However, the utilization of free and environmentally friendly gravity energy did not arouse interest even on the part of patent experts.

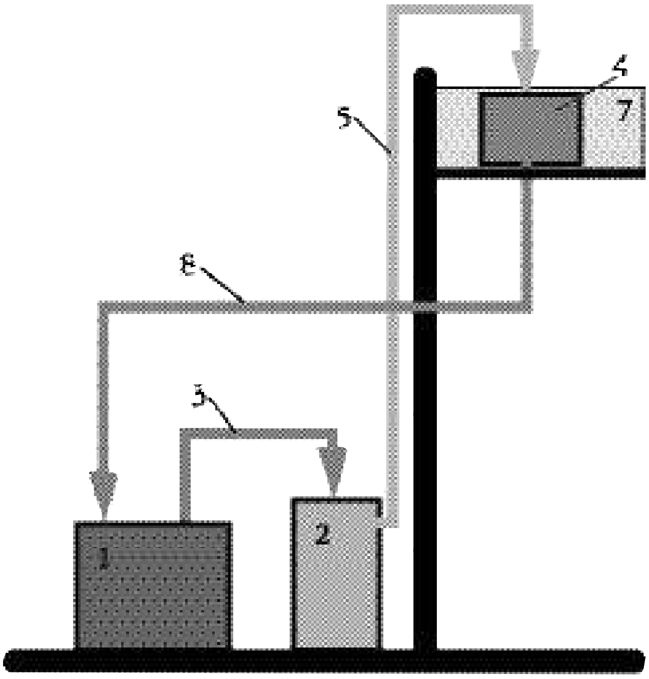

Fig. 2
System

The author is looking for serious partners.

The Alternative Energy Potential in Russia

According to the International Energy Agency (IEA) calculations, total investments into the energy industry constitute approximately 330 billion dollars annually. Almost half of this sum (150 billion annually) is allotted to the development of the electric power industry. This information only partially reflects the potential of the alternative energy development and does not include the market growth factor upon the introduction of autonomous power system technologies. Vast territories of the planet have not been sufficiently developed yet due to the absence of local energy resources and power lines. From this point of view, Russia is the country with an enormous consumer demand potential provided that the mass production of autonomous fuel-less energy systems guarantees low production prices.

Alexander V. Frolov
Editor-in-Chief

Faraday
Lab Ltd

Fuel-less energy R&D
R&D in alternative astronautics

Read our
New Energy Technonlogies
magazine

Lev Tolstoy Str. 7-601
St.Petersburg
197376, Russia
Tel: 7-812-3803844
www.faraday.ru

Modern Perpetual Mobile

Review prepared by correspondent Alla Pashova, Russia.

Based on the Internet pages and other information sources

It is common knowledge that any experimental information, which is against a conventional scientific paradigm, is considered to be false, extracted from the society and finally is turned into something **classified and elite.**

Politics-oriented science as a whole and natural science in particular often result in misrepresentation and falsehood of **popular knowledge** that turns it into the fighting weapon of socio-political organizations, states and their groups.

World-outlook and political contradictions of any society often determine science development strategies. In 1885 Nicola Tesla presented his transformer in action by means of which he lit carbon lamps in a 25 miles circle without any cords, conductors or switches using the 5000-HP turbine of the Niagara hydroelectric power station. Afterward, he obtained financial support for one of his energy projects.

On a special testing ground Nicola Tesla built vacuum-energy systems. However, in 1885 the systems, as well as the testing ground, were destroyed as it became evident that if they went on the mankind would never need organic fuel again. But the destruction of the systems and testing ground does not at all mean the destruction of papers and documents...Ever since that moment people have been trying to find "free energy".

The experiment of distant lighting of carbon lamps without using any cords was repeated only once by the physicist Filippov. He managed to light the lamps in Tsarskoe Selo employing the electric generating systems in St. Petersburg. Filippov died in 1914 under mysterious circumstances.

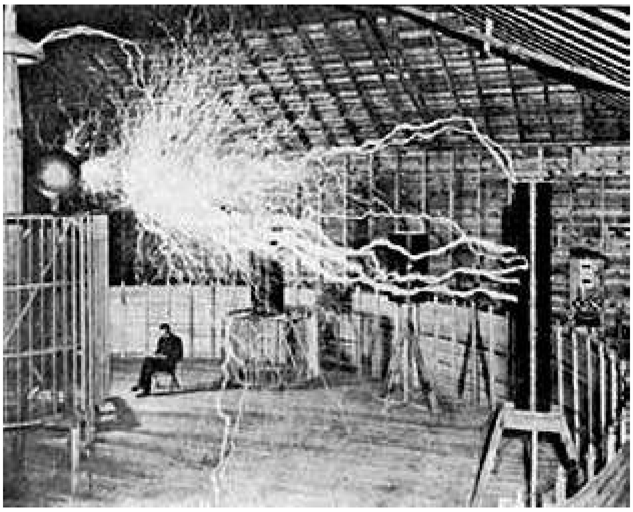

In 1917 the Portuguese American Andres invented non-convetional fuel for an internal combustion engine. The essence of his invention was in some inexpensive chemicals, which he dissolved in water (some drops of chemicals per pail of water). A special state committee tested the new fuel during the New York - Washington - New York automobile races. After the races one of the largest oil monopolies in the U.S. offered Andres 2 million dollars in cash for the documents and all rights for his invention and later locked all the information in their safes. Two days later after getting the money Andres disappeared. (The information was disclosed by V. Vasilevsky, the former chief of the scientific and technical intelligence department of the KGB of the USSR, who had been in charge of the department since 1930).

The above-mentioned examples show that there were a lot of attempts to develop revolutionary energy technologies. But what then are the grounds for the 'energy crisis' complaints? The answer is simple: the desire of a group of financial and energy multinational corporations for an economic, political and, finally, for a total control of the world. That is why there is no room for "free-energy" hunters in our science, not even as laboratory assistants, not to mention some higher positions.

Practice often negates scientific dogmas, especially when it comes to the efficiency of different devices. In general, the traditional understanding of efficiency is quite limited and should be forgotten. Instead of the traditional notion "efficiency" we should consider the notion "energy-conversion efficiency" (ECE), which has no limits. In 1974 in the USA a six-stroke internal combustion engine was designed which had a double ECE. The fifth stroke of the engine operation included water injection; the sixth was characterized by the water vapor introduction. Firstly, the engine had a higher ECE than that of the Carnot cycle. Secondly, taking the 55 % ECE of the contemporary internal combustion engine, the ECE of a six-stroke engine exceeds 1.

Practice often negates scientific dogmas, especially when it comes to the efficiency of different devices.In general, the traditional understanding of efficiency is quite limited and should be forgotten. Instead of the traditional notion"efficiency" we should consider the notion "energy-conversion efficiency"...

Before the war there appeared "studebekkers" with the fuel consumption of 5.5 l per 100 km. But the fuel-economy record was set by the Japanese. In 1986 they built up an automobile, which used only 0.055 l of fuel per 100 km (about 44 g). Unfortunately, now we do not have plants, which could manufacture automobiles of the kind.

In 1832 Tomson came up with the idea of **a heat-pump unit**. Stating the unity and interconvertibility of substance movements, he proved that the mechanical energy consumption can restore stray heat. At that time cross-Atlantic ships had low-temperature steam engines, which employed exhaust steam and ether vapors. The efficiency of those engines was already higher than the Carnot cycle efficiency.

Now with the help of the facts acquired during the last century we will prove the scientific inconsistency of the orthodox model of the second stage of thermodynamics. In 1941 the English physicists O. Hougen and K. Watson published their work featuring the experimental fact that the ammonia-saturated vapor pressure in liquids was higher than the general pressure. For example, with the general pressure of 50 atm and the temperature of 0^0 C, such pressure was 17 % higher. This means that having placed a semitransparent partition and a turbine between two containers it could have been possible to build up a natural-heat vapor-converter. This method was used by **von Platen, the Nobel prizewinner from Sweden, who constructed a self-rotating centrifugal machine, which provided a 1000 atm pressure when ammonia interacted with water. The heat emitted was enough to compensate friction losses**. It is also known that the Carnot principle cannot be applied to closed cycles. W. Vielstich, 'Combustion Cells', 1968: **"If an entropy change of a reaction is negative, it can exceed 1**. It means that a certain amount of energy can be obtained from the natural heat. Such an effect is possible in electrochemical generators, which employ a direct oxidation process before oxidation and dioxidation".

A demonstrative example of scientific experiments withdrawn from science is discovery # 13 "The knock power-transfer regularity" made on December, 18th, 1962, which makes it possible to create a mechanical "perpetual mobile". The experiment shows that the conventional "knock theory" does not work in practice: **the bouncing energy of a body after it has been knocked can be higher than its energy before it has been knocked**.

Trying to get some recognition, **E. Aleksandrov, Doctor of Engineering Science, made his demonstrative experiment in front of many different commissions: a chilled-steel ball when falling from a 10 m height onto a firmly-fixed chilled-steel plate made a 14-15 m-high bounce**. That was the notorious 'perpetual mobile' as it is. It went on like this until somebody decided to explain the result of the experiment by a metal lattice internal energy. At once, everything was clear. Since that moment the above-mentioned principle has been successfully employed in percussion-action machines. This principle can also be used in simple electric-power systems.

Now we can see that electro-technicians had already constructed demonstrative 'perpetual mobile' machines 10 years before politicians and economists started to speak about the 'energy crisis'.

In 1921 the mass media wrote about A. Hubbard who invented an electromagnetic generator without any external energy supply. The Hubbard generator was used as a boat motor.

> *...now we can see that electrotechnicians had already constructed demonstrative 'perpetual mobile' machines 10 years before politicians and economists started to speak about the 'energy crisis'...*

In 1928 L. Nidershot invented a 300 W electric generator, which did not require any external energy supply. The device consisted of a radio-technical oscillation generator (500 kHz) and a coil. After 68 years there appeared publications about A. Melnichenko who repeated the experiment.

In 1927 T. Brown (England) obtained a patent on the ways of originating a moving power and energy using an electric field. Later, in 1955 while working in France, he presented a system with the speed-capability of several miles per hour using a 2000 eV field. After that he had to stop his experiments and was sent to the USA.

In 1943 N. Tesla presented an electric-motor automobile. The energy was generated by a previously-unknown generator.

In 1960 Stovbunenko (whose research work results were subject to a special verdict of the Military-Industrial Complex) presented his motor, which made it possible to move in his old automobile 'Moskvich' around the city for the whole day utilizing the power of a standard automobile battery.

In 1980 there was a qualitative break through in electro-technical "perpetual mobiles". Baumann's electrostatic machines of the total capacity of 750 kW started functioning in the religious community of Liden (Switzerland), serving all daily needs of the village. (See the photos below)

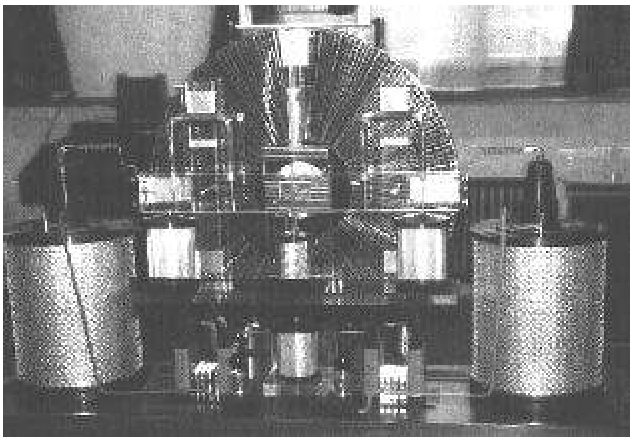

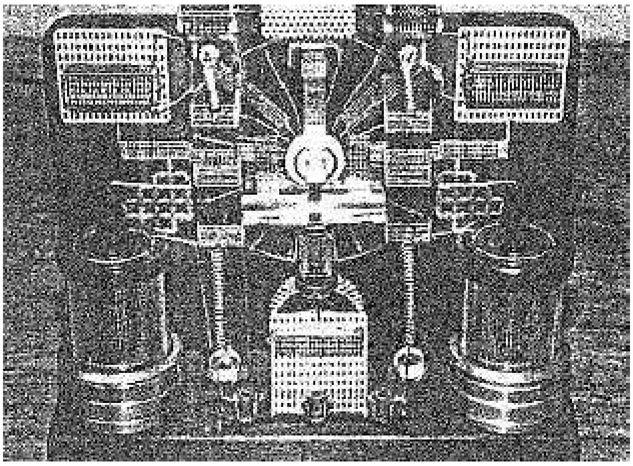

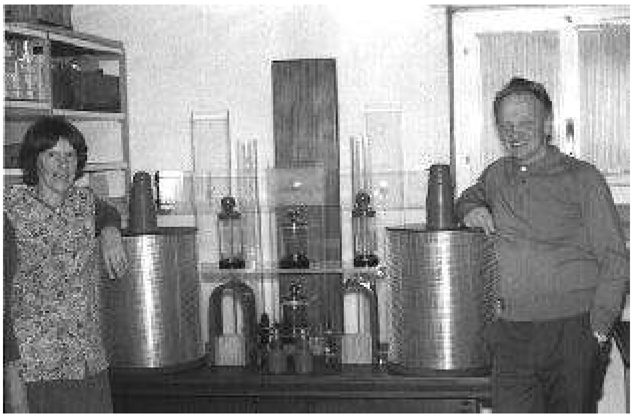

The photos are taken from www.free-energy.cc

Thus, in 1980 there appeared a community with no energy problems, no need for organic fuel and no fear for the 'energy crisis'.

In 1881 N. Sluginov discovered an energy imbalance that occurred in the water electrolysis process. In his experiments he got that the output energy was 30 % higher than the input energy. In 1980 American scientists recreated this energy imbalance and proved that while employing the steam turbine rejected heat, the water electrolysis efficiency reaches 120%.

This information introduces a story about an energy system invented in 1957 under the guidance of I. Filimonenko. The system did not simply generate energy in the form of a high-pressure vapor and output hydrogen and oxygen but also absorbed radiation. In 1960 the USSR Central Committee and Cabinet Council gave a special secret ruling, known as “The Three 'K's'” (Keldysh, Kurchatov, Korolyov) to go on with the research in this system. However, after Kurchatov's death the project was "oppressed" and later on after the death of Korolyov the project was completely cancelled. The special committee of the USSR Academy of Sciences stated that the system was functioning against "the Law of Nature". I. Filimonenko was dismissed and expelled from the party. Later, in 1980-1991 the experiments were partially renewed. Several testing systems were constructed in the Chelyabinsk region but the construction was never finished and the use of a mobile system for eliminating the damage caused by the Chernobyl accident was refused. This story is a demonstrative example of the traditional rejection of promising projects by science, which could be very beneficial for Russia.

Theoretical Background of 'Perpetual Mobile'

Originally, the term 'entropy' was used only in connection with the reasoning about limited or eternal nature of the Universe. Later it came to be used to describe the functioning principles of thermal machines. Now there exist more than 600 contradictory definitions for 'entropy', which are actually the results of a long and unfinished discussion between the supporters of Dekart and Leibnitz, who defined 'motion' in different ways. It is these permanent contradictions in the theoretical fundamentals of physics that make it impossible to explain the existence of a 'perpetual mobile'. However, standard physics has never denied the possibility of creating energy technologies allowing the ECE higher than 1.

We should remember that the basis of the so-called variational calculation is the mathematical apparatus technique of the conventional mechanics. It seems that if system behavior equations directly depend on time, this system cannot fall under any conservation laws. **This is the proof of a complete failure of the conventional energy conservation law!**

> *It is these permanent contradictions in the theoretical fundamentals of physics that make it impossible to explain the existence of a 'perpetual mobile'*

In time-independent thermodynamics the existence of 'perpetual mobile' can be explained by the system internal energy usage. There exist two kinds of thermodynamic systems. Some of them when being heated or affected in some other way start increasing their internal energy while other systems show the internal energy decrease. The latter are called negative absorption systems. The standard version of the first law of thermodynamics does not cover all the processes occurring in such systems. The output of a negative absorption system can be higher than its energy. The additional output is acquired by the gradual internal energy loss.

To cap it all, there are more than a hundred experimental facts that show the limited nature of the Maxwell theory. They were obtained by scientists from different research institutes and construction bureaus, were tested and registered. Thus, for example, in 1973 in the USSR the acoustic magnetoelectricity effect was discovered. The author of the discovery proved the interaction of electrons with ultrasonic waves with a 1000-times energy increase. This is against the Maxwell theory, which absolutely denies effects of the kind. But the limited nature of the theory was never accepted.

However, the examples of the machines in action prove that since 1834 when the first refrigerator was invented, when the cross-Atlantic ships were equipped with low-temperature ether-vapor steam-engines, the contemporary science had no right to introduce 'the second holy principle', especially in the orthodox form of 'the Carnot cycle efficiency'. The modern theoretical physics cannot provide grounds for the negation of creating over-unity devices. That is why we should rely on the physicists-experimentalists who develop the science in a practical rather than theoretical way.

Alekseenko's Fuel-less Motor

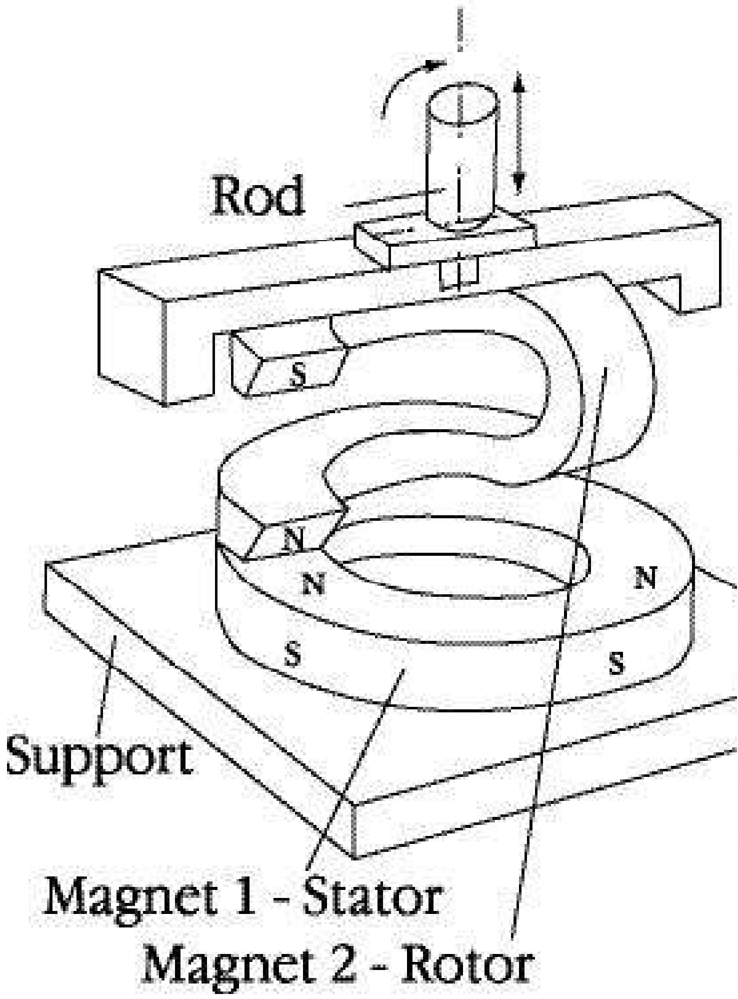

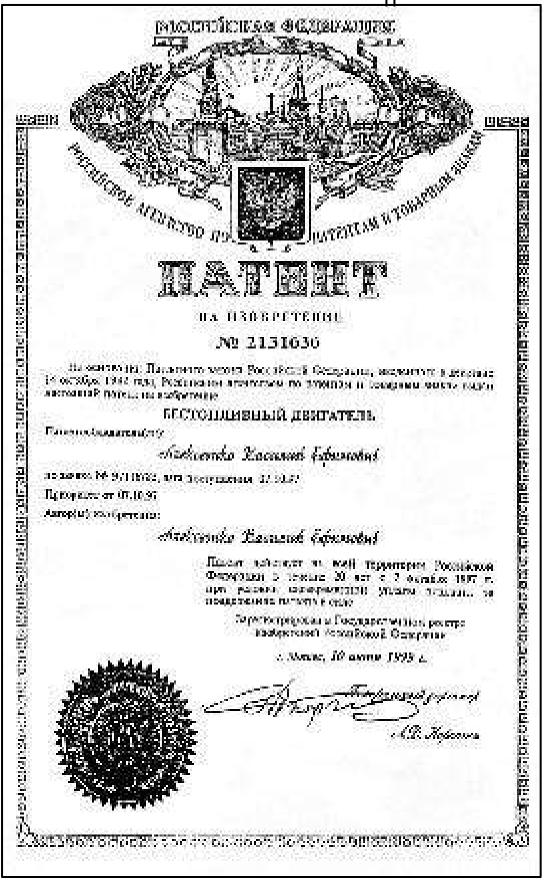
РОССИЙСКАЯ ФЕДЕРАЦИЯ

ПАТЕНТ

НА ИЗОБРЕТЕНИЕ

№ 2131636

БЕСТОПЛИВНЫЙ ДВИГАТЕЛЬ

Patent # 2131636
Vasiliy E. Alekseenko

Aerospace Defense Research Finds Free Energy and Antigravity Possible

Dr. Ines Espinoza, USA
email: dr.ines@vasantcorporation.com

Editor: it is worth mentioning that the first gravitational-wave physical mechanism and the mechanism of spin-effect "grazer" (gravitational lasers) designing were described by S. M. Polyakov and O.S. Polyakov in the book "An Introduction to the Experimental Gravitational Studies", Moscow, 1991.

After a 7-year research study, a senior staff aerospace defense engineer, George J. Bugh, concludes economically free energy and even antigravity are possible.

A 7-year study of "free energy" devices, sometimes called "over unity" devices, was done to determine if any of these devices generated power by transferring energy from unknown sources and if so to determine where the energy was coming from. Included in the study is research of related devices with claimed antigravity effects. The study attempted to determine validity of claims, commonalities of device characteristics and to determine how these devices could work. **The results conclude that some devices can generate economically free energy. This study also concludes there is a possible link between gravity and electromagnetism that can be exploited to generate antigravity or electrogravity effects.**

The majority of the study was to come up with a theory to explain how the devices could work. In his research, Mr. Bugh used mostly classical electrodynamics rather than quantum electrodynamics. In Quantum theory, the wave-like characteristics of matter are described using abstract probability waves. However, Bugh proposes that the wave characteristics of matter may also be described as coming from a very real sea of unseen electromagnetic standing waves among all matter. There is a slide show presentation at the website that explains the differences between the Quantum and Classical way of explaining particle interactions.

According to classical electrodynamics, all electrically charged particles, like quarks of protons and neutrons as well as orbital electrons for example, **should radiate away energy from precessional and precessional plus orbital motion**.

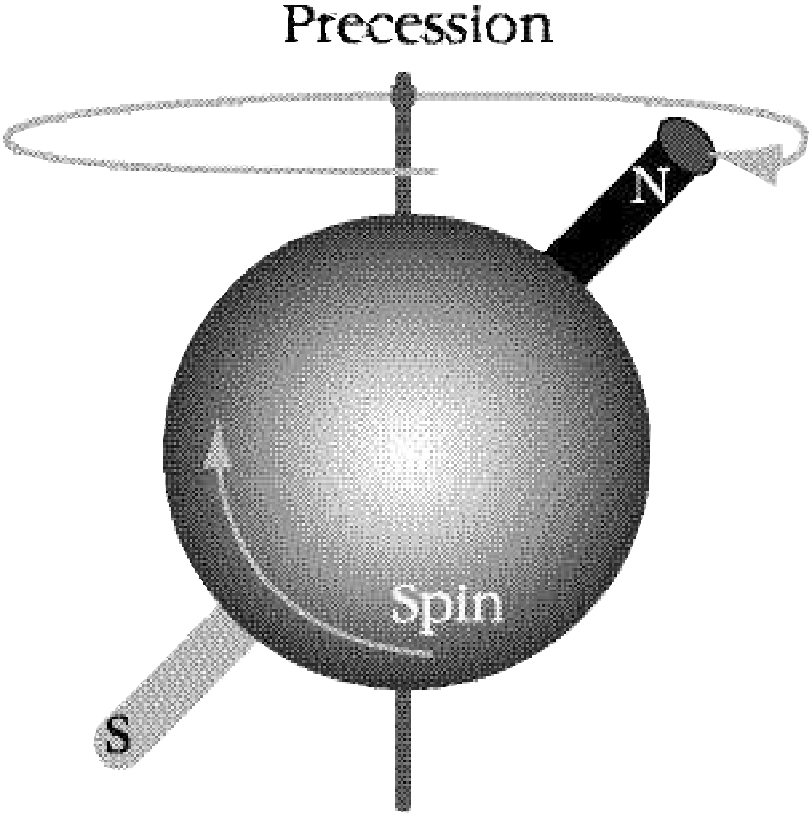

Fig.1
Electron motion

If in fact this really happens then all electrically charged particles can be radiating away energy all the time. However, all particles can also absorb just as much energy from all other radiating particles. The absorbed energy applies electromagnetic forces that naturally move all similar type particles into harmonious precessional motions with all other particles. This results in a vast sea of electromagnetic standing waves among all matter. Even free particles would move into precessional motions that are in sync with the established sea of standing waves.

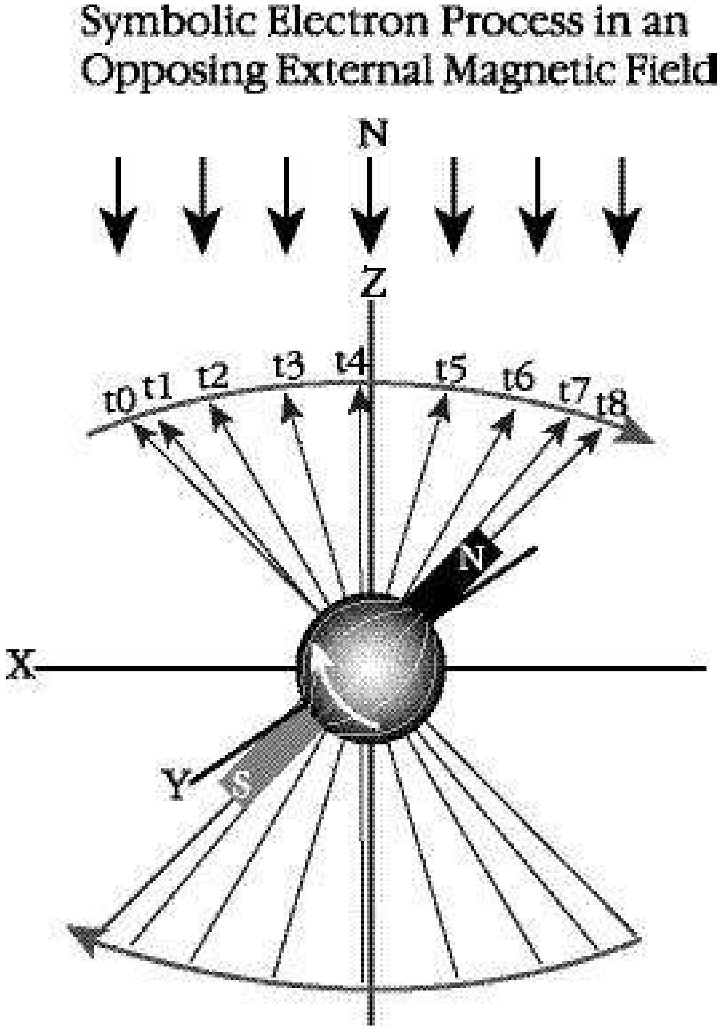
Symbolic Electron Process in an Opposing External Magnetic Field
N
Z
t0
t1
t2
t3
t4
t5
t6
t7
t8
N
X
Y
S

Fig. 2
Electron precession

There can be a hidden yet strong tendency towards harmony among all matter in the universe due to these unseen standing waves and spin interactions among all matter. This tendency can overcome to a great extent the tendency towards chaos and heat death of the universe. This tendency can also be exploited to perform work.

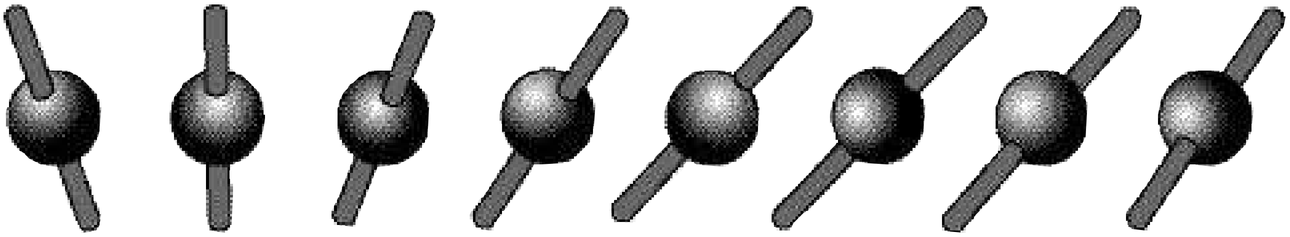

Fig. 3
Spin wave

This is an interaction among all matter that Ernst Mach alluded to as necessary to cause matter's characteristic of inertia. Einstein later called this Mach's Principle. Einstein studied Mach's ideas while developing his theory of General Relativity.

Bugh describes inertial resistance to acceleration as caused by electromagnetic forces. Changes in position of a mass will cause phase differences to develop between the precessional motions of the particles of that mass relative to the sea of standing waves. This in turn causes electromagnetic force that resists a mass from changing its position.

The research papers are published in a book and a CD titled "Spin Wave Technology Initial Release".

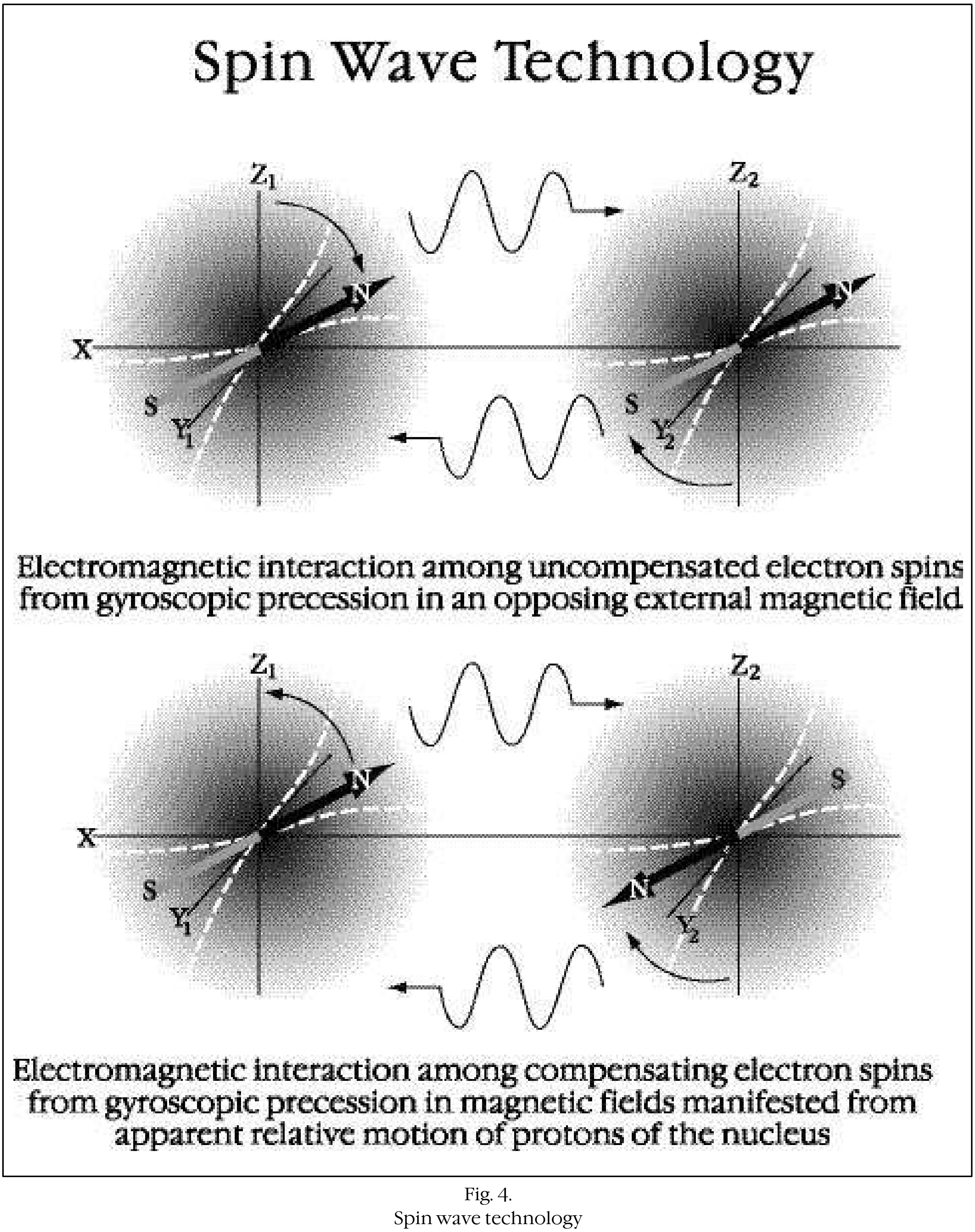
Spin Wave Technology
Z1
Z2
N
X
S
Y1
Y2
Electromagnetic interaction among uncompensated electron spins from gyroscopic precession in an opposing external magnetic field
Z1
Z2
N
X
S
Y1
N
S
Y2
Electromagnetic interaction among compensating electron spins from gyroscopic precession in magnetic fields manifested from apparent relative motion of protons of the nucleus

Fig. 4.
Spin wave technology

At the end of the book it suggests that it should be possible to create computer simulations of particle interactions using classical electrodynamics to demonstrate both inertial resistance and gravitational attraction and it should be possible to simulate devices that manipulate these particle interactions to demonstrate antigravity effects.

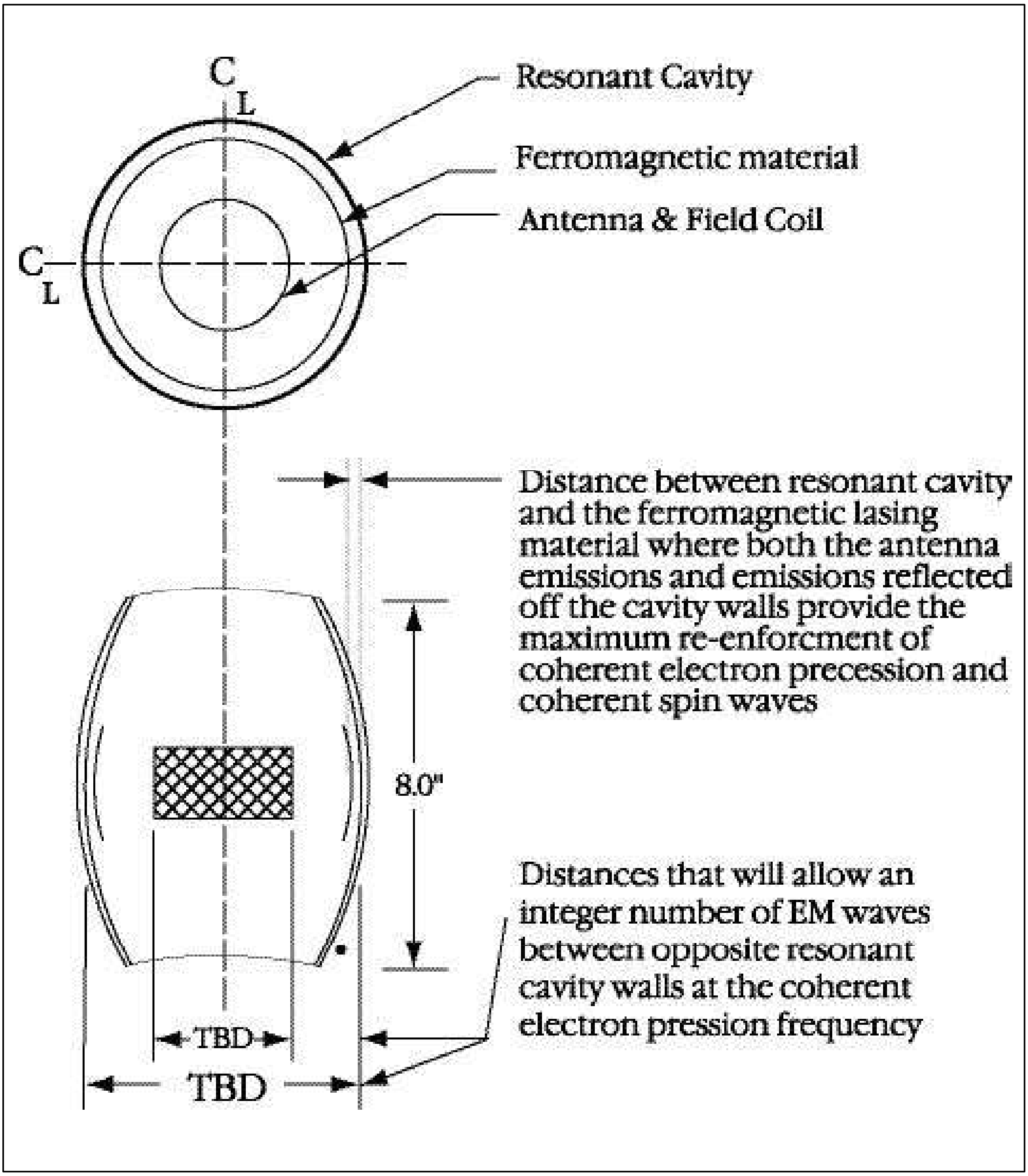
C
L
C
L
Resonant Cavity
Ferromagnetic material
Antenna & Field Coil
Distance between resonant cavity and the ferromagnetic lasing material where both the antenna emissions and emissions reflected off the cavity walls provide the maximum re-enforcment of coherent electron precession and coherent spin waves
8.0"
Distances that will allow an integer number of EM waves between opposite resonant cavity walls at the coherent electron pression frequency
TBD
TBD

Fig. 5
Spin wave laser

More information is available about the research results at the website: www.vasantcorporation.com.

On Physics of Electromagnetism
Mechanical Analogue or Pure Mechanics?

Yury V. Ivanko, Ukraine
E-mail: ux0lw@hotmail.com

After publication of the article "Study of possibility of a high-energy hyper-low-frequency electric field" in the "New Energy Technologies" magazine, Issue #2, 2003 I am often asked why I have mentioned the Big Bang in my hypothesis.

I would like to reassure atheists since I do not back the Big Bang hypothesis. But I had to base myself on something. At present the origin of the Universe as a result of the Big Bang has priority and is recognized by the academic science.

Both theorists and experts have long noticed a deep analogue between vortex motion of air currents and electromagnetism. My vision of the physics of electricity and electromagnetism is fully based on existence of a vortex capable physical environment. What is considered to be and is measured **as electric voltage, electric current, magnetic or gravitational interactions is nothing but similar-originating phenomena - the result of vortex flow environment affecting our detectors - dissimilar formations for these vortex flows**.

For instance, let us consider three vectors for the flow along the skin-layer of the line conductor - aerial emitter, together with alternating voltage attached to it (Fig.1):

1) longitudinal - showing as electric voltage, the gradient of potentials between the emitter feeding point and the emitter cold end, U;

2) tangential - showing as amperage - electric current, I;

3) radial - showing as the interacting force of two flows - magnetic field (attraction/repulsion), H (not to be mixed with the commonly used term of magnetic intensity).

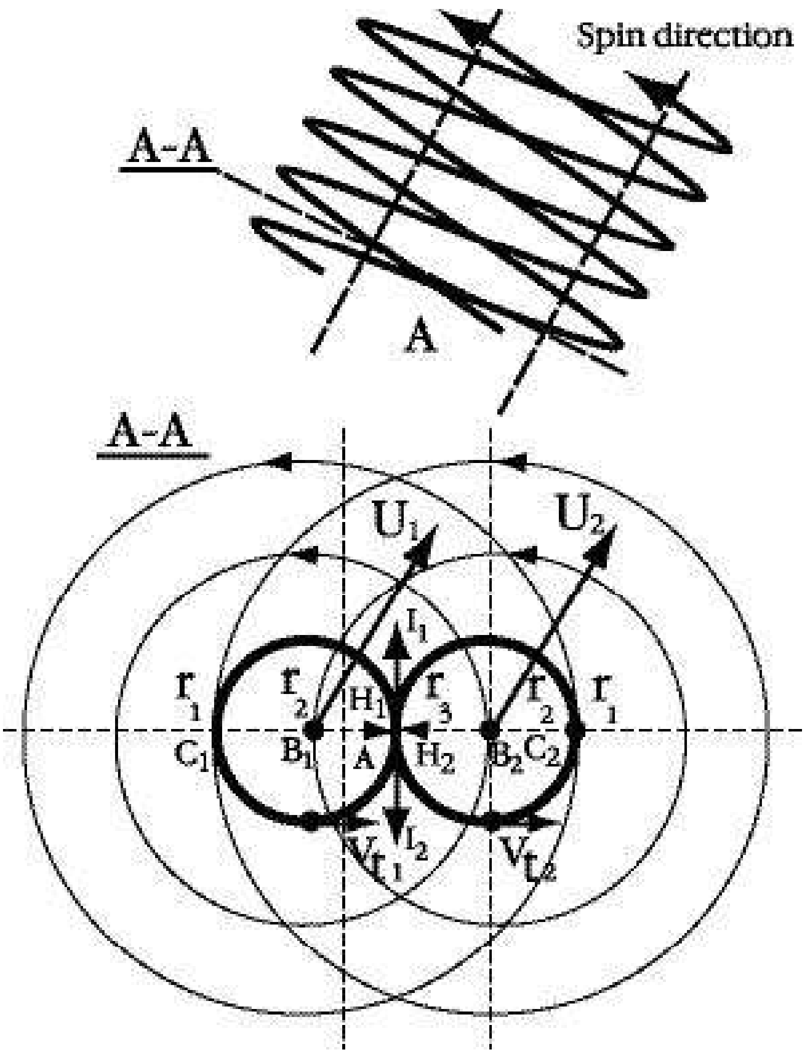

Fig.1
Interaction of two left-handed homogeneous parallel flows.

Let us consider it step by step:

Consider line conductor AB with length L in free space (Fig. 2).

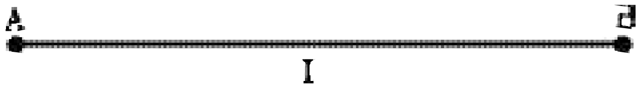

Fig.2
Line conductor in free space.

Let us assume that he is not affected by and fields. Then the gradient of potentials between A and B equals zero. Let us apply sinusoidal voltage from the generator output to point A (Fig. 3).

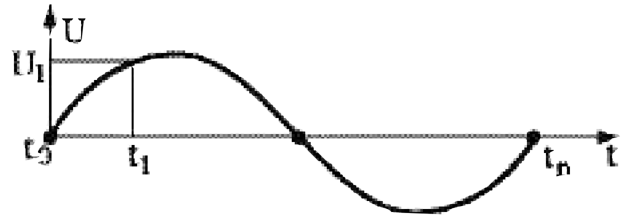

Fig.3
Sinusoidal voltage at the generator output.

Let us consider the initial time point t_0. The generator output voltage is U=0. At time point t_1 the generator output voltage has changed and took on value U_1.

It is necessary to note that (See Fig. 2) the propagation speed is finite: c=300000 km/sec. Point B will "learn" about the voltage alteration only in time $t_B=t_1+L/c$. Therefore, at time point t_1 between points A and B appears gradient of potentials $\Delta\varphi = U_1$. So what is the direction of vector U_1? Right, along conductor AB.

What practice says

One should not be a physicist or radio mechanic to notice small electro-shock of ungrounded metal objects close to aerials of powerful transmitters. Specialists term them as "electromagnetic inducers". Has an expert ever measured the actual characteristic of the field intensity distribution along a line conductor? Why should he? Everyone knows from school lessons how alternating voltage is distributed along the conductor, the length of which is comparable with the wave length (Fig. 4).

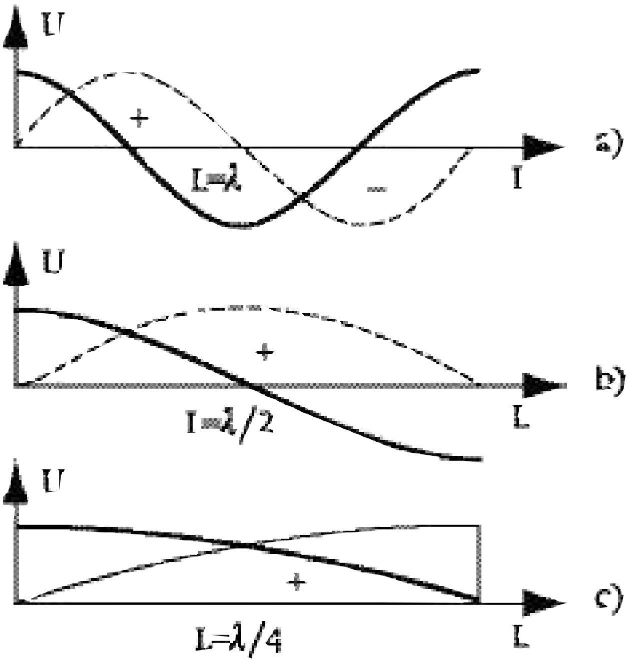

Fig.4
Classical representation of voltage distribution along a linear vibrator a) for λ, b) for $\lambda/2$, c) for $\lambda/4$.

Now let us consider what experimental measurements suggest.

An experiment of measuring the field intensity distribution along the line conductor is shown below. A classic field indicator (FI), assembled as per scheme in Fig.5, was used as an intensity measuring instrument.

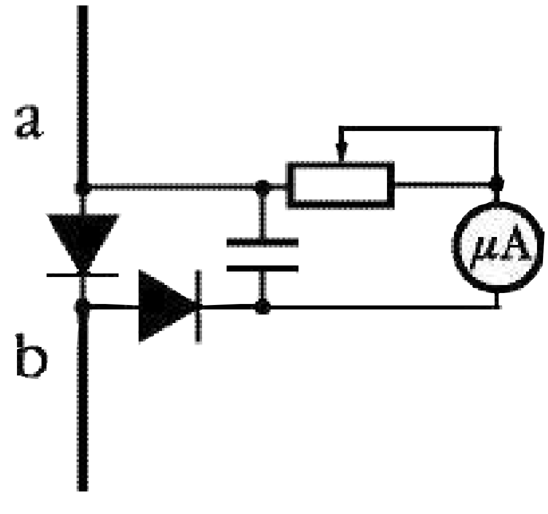

Fig.5
Field indicator principle diagram.

FI was located in immediate proximity to the vibrator but had no galvanic contact with it. Experimental frequencies were selected in accordance with the standing wave coefficient minimum (SWC) for λ, $\lambda/2$, $\lambda/4$. Sinusoidal signal generator was connected to one end of the vibrator (Fig. 6a).

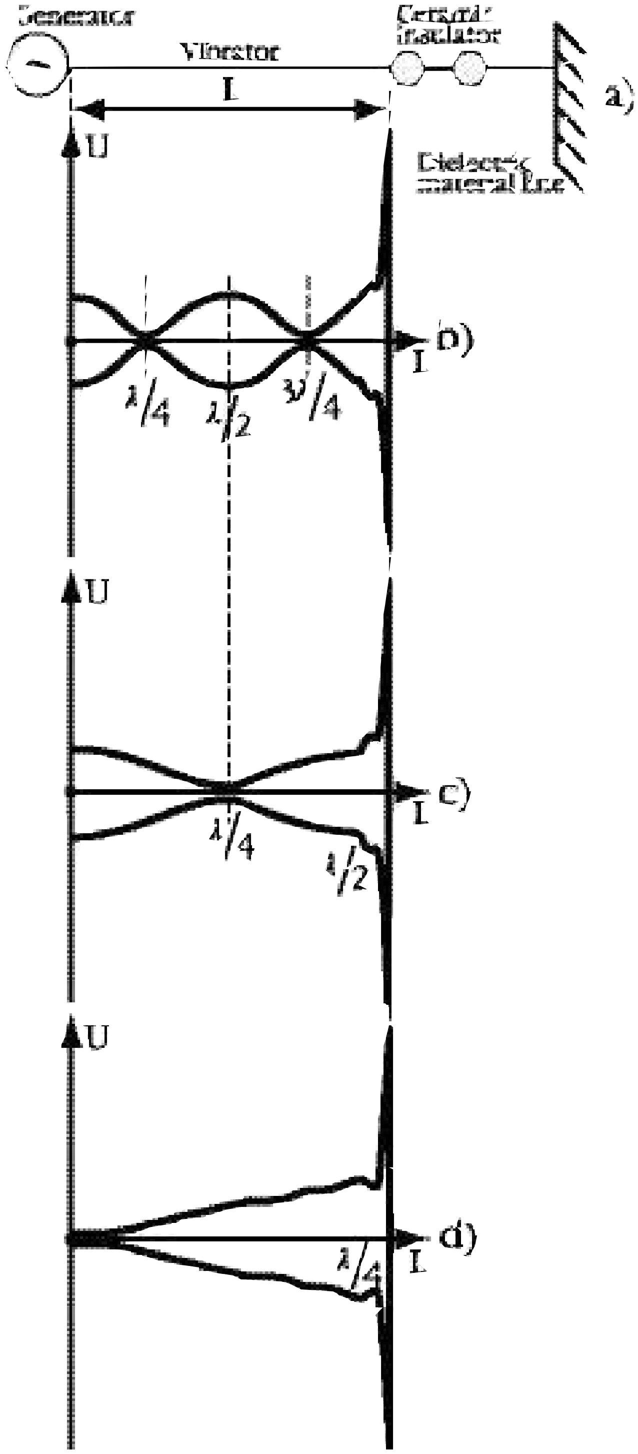

Fig.6
Measured field intensity
along a linear vibrator.

To better understand diagrams in Fig. 6 curves are drawn at the upper and lower ends of each axis L that corresponds to the vibrator location. The curves were run with FI horizontally located on both sides of the vibrator.

From Fig. 6d it is clear that the diagram is axisymmetric and looks like a conic funnel.

Meanwhile, the vibrator directional diagram measured by the monitor receiver far more distant than λ, shows the characteristic that coincides with the simulated one through the Matlab-6 system (See Fig.7). As you can see from this figure, the spatial characteristic of the directional diagram is in the form of a toroid.

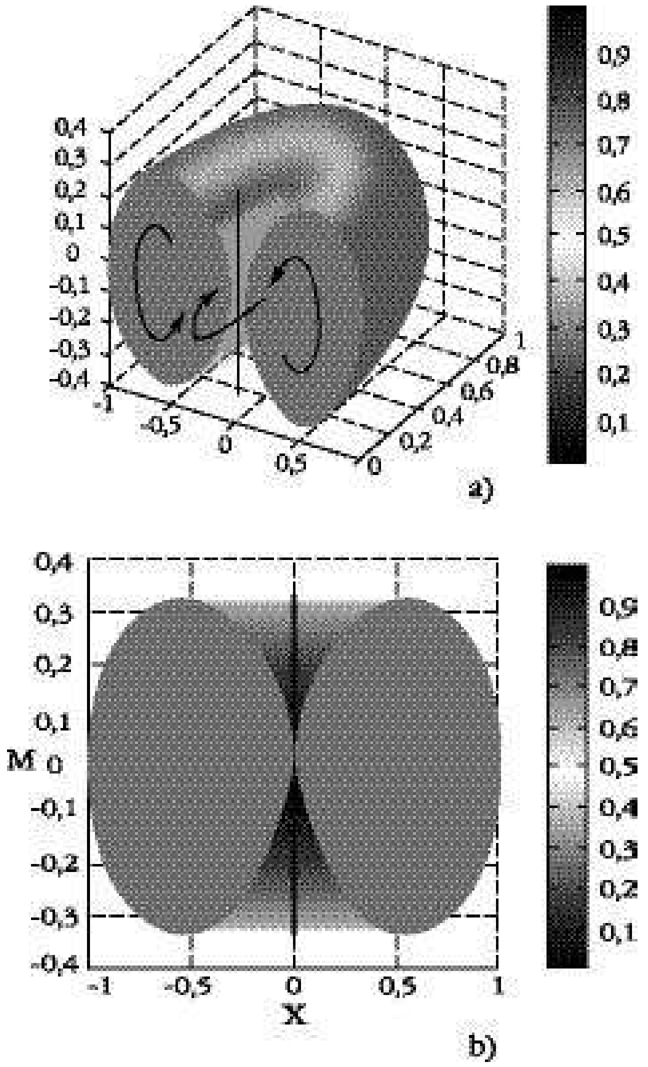

Fig.7
Matlab-6-simulated directional diagram of the semi-wave vibrator

A series of alterations was made to explore vortex ether dynamic processes with variable wave length and frequency in line conductors, solenoidal and flat coils, etc.

For certain purposes the solenoidal coil may be considered as a shortcut linear vibrator. On respective frequencies the field intensity distribution along the solenoid corresponds to the linear vibrator. This data may prove useful to many modern engineers of the Tesla transformers and generators. To increase voltage on the cold end a toroid capacitor is set in the Tesla transformers. Nuclear physicists use the hemisphere capacity for the particle accelerators. Thus, a voltage of millions of volts may be achieved.

Flat coils are best explored in terms of field intensity distribution (See Fig. 8)

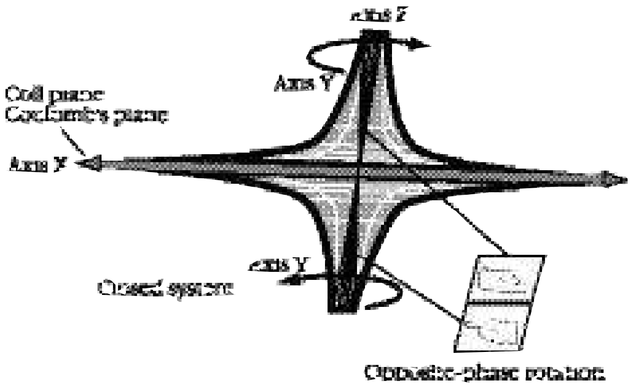

Fig.8
Field intensity distribution for a flat coil in a closed resonance circuit.

Field intensity distribution on a resonance frequency is correct.

The practical experiment for the case in fig. 6d has shown that the neon lamp that is connected with the vibrator end will be off. And this is with a 100 W generator! It is a paradox, one would say. The measurement diagram shows a "splash" of the field intensity at the vibrator end but we feel nothing. Indeed, when the FI is located as shown in Fig.9 for $\lambda/4$, the "electrodynamic vacuum" will generate. The device showed "0" intensity, which conforms neither with the diagram in Fig.6 nor with the diagram in Fig.7.

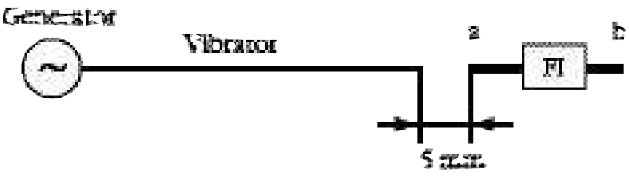

Fig.9
Field indicator shows "0" - "electrodynamic vacuum".

The same "vacuum" in the center of tornado was observed by people in the epicenter.

Remember what is measured with the field indicator (an ampere meter) along the linear vibrator. The Ampere force!

So, both people and their environment are in rotation process. The Earth rotates around its axis and around the Sun. The Sun rotates around the Galaxy center. The Galaxy rotates around the center of the Universe. The Universe rotates...

Every space point has an inceptive rotation impulse (torque). Let us remember the analogy. Water flowing out of the tub rotates counterclockwise in the northern hemisphere and clockwise in the southern hemisphere. If you set a right-handed torque in the northern hemisphere, the water will continue to flow counterclockwise.

The electromagnetism is the same. It has a natural left-handed rotation. Nevertheless, practical radiotechnology is familiar with right-handed fields artificially generated by transmitters.

The commonly used term of magnetic interactions is easier to understand if we consider the electromagnetic field as a vortex flow. Two left-handed homogeneous parallel flows are shown in Fig.1. At the point of interaction A of vector projection, the speeds of flow propagation on section plane **A-A** have opposite direction and compensate each other. The actual flow density is $\rho_1 < \rho_2 < \rho_3$. Reduced pressure occurs at point A. The flows are attracted. The resulting attraction vector for the whole flow is radially directed. The counter parallel flows will have composition, i.e. overpressure, and will repel.

One should note that graphic presentation in Fig.1 for vectors H and I for the etherdynamic systems is possible only in special cases. In dynamic vortex flows the current intensity vector is actually directed towards the flow movement and coincides with the vector direction of its propagation speed (V). The resulting vector H remains a radius-vector only for the considered point of interaction of two flows. The flow energy parameters are defined by its propagation speed V (Fig.1).

I am convinced of existence of ether vortex flows in the Universe. Modern astronomic and cosmological surveys confirm my hypotheses of aether Hyper-fields existence on a Universal scale.

The commonly used term of magnetic interactions is easier to understand if we consider the electromagnetic field as a vortex flow.

Even an amateur in aerodynamics and dynamic of gas vortex may make cosmological conclusions:

♦ vortex initiation implies the initial gradient flow of at least two counter flows;

♦ energy parameters of the emerging vortex flow can not exceed those of the parent flows

♦ there exist ether flows with energy exceeding that of our Universe

♦ our Universe has not originated from a point and is not finite. It is the spiral (vortex) structure of our Universe that testifies against the Big Bang hypothesis.

Conclusions:

Practical radiotechnology is familiar with the key difference between the field intensity distribution diagrams of aerial emitters in near and far zones. After combining diagrams 6d and 7b (Fig.10) it is possible to draw certain conclusions.

When transverse waves (radio waves) are formed around the aerial vibrator, the double transformation of the initial flow geometry starts:

♦ **the first flow is a homogeneous spiral flow on the conductor skin-layer;**

¨ **the second flow is a non-homogeneous vortex flow of the near zone;**

¨ **the third flow is a toroidal flow of the far zone.**

The toroid volume alteration in the far zone (as per the transmitter vibration frequency) forms volumetric planar transverse vibrations (spherical waves) that are called electromagnetic or radio waves. The gradient of potentials alterations (as per the transmitter vibration frequency) provoke changes of the Coulomb plane, which form longitudinal vibrations with a narrow directional diagram.

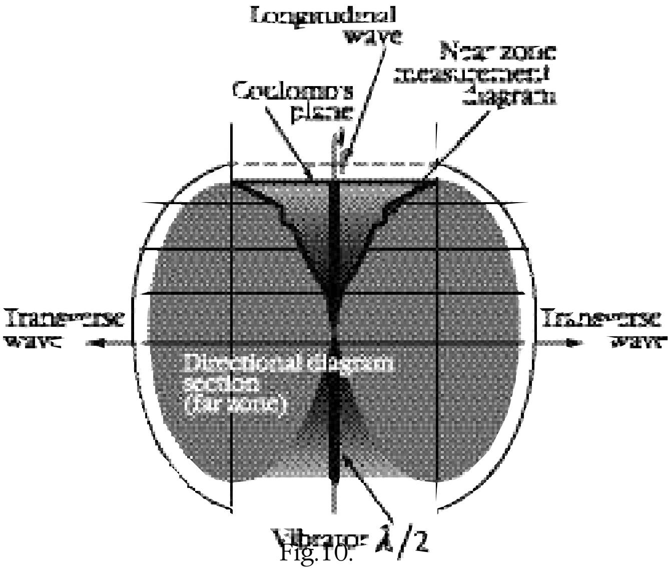

Fig.10.

The longitudinal flow is assumed to be formed as a chain of toroids (like smoke rings)

Ether-dynamic approach provides grounds for more substantial conclusions and assumptions. A new approach to energy and communication systems is made real. Further results of the experiments with transmitters and aerials, assembled on ether-dynamic principles, enabled to draw the following conclusions:

- all reasons, preventing transformation of the transmitter power into the power of transverse radio waves, are evident;
- all reasons for the transverse wave phase lagging with the increase of distance from the emitter, are also evident;
- besides the toroid that forms the transverse wave, the second flow generates a crucially new radiation that has a very narrow directional diagram along its longitudinal axis;
- the aerial ignores wave geometry for the second flow and its longitudinal derivative;
- the radiation along the second flow axis has a very high penetrating power;
- the radiation along the second flow axis has no signs of polarization.

The list of references is not shown in this article due to its size. It can be looked up at **www.efir.com.ua.**

NEWS

In October 2003 the second stage of the research on the "time control" was completed. In the photo below you can see Chernobrov V. A. and Frolov A. V. in the laboratory, October 31st 2003.

The second stage conclusions are quite interesting, however the usage of timers (chronometers) as detectors is proven to be inappropriate for they are subject to magnetization and their indication variations cannot be reliable in this case.

The next stage will feature the following method of detection of the time rate changes (the rate of the matter existence) that is the radiation wave-length measurement. If it is possible to achieve significant results then we will be able to detect the change of the laser ray colour in the area of the effect and its linear path deviations. The experiments will be held to register the weight changes of the detector at the expected influence of the produced effect on the ether density.

Our company looks for cooperation with corporations, which are interested in the application aspects of these technologies.

Frolov A. V.
General Director, Faraday Lab Ltd
7-812-3803844
http://www.faraday.ru

CELL OF THE THIN PLASMAELECTRIC HEAT GENERATOR

Ph. M. Kanarev, A.I. Tlishev, Russia
E-mail: kanphil@mail.kuban.ru

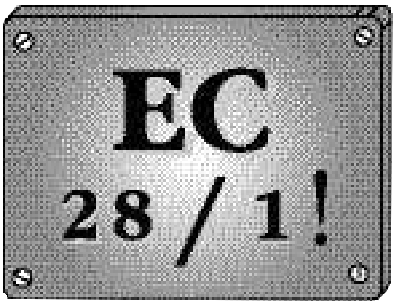

The main task of the experiment was to check the hypothesis: "Electrodynamic influence on the water molecules gives the possibility to significantly reduce energy expenses on destruction of their chemical bonds; further fu-sion of these molecules considerably increases the output of additional energy in the form of heat".

In order to solve this task, special experiments were carried out as regards the electrodynamic destruction of chemical bonds of water molecules by electric pulses of various frequencies.

Photo of the Thin Plasmaelectric Heat Generator

CHECK TEST RECORD FOR PLASMAELECTRIC HEAT GENERATOR

Table 1

	Indices	1	2	3	Avg.
1	mass of the solution, which has passed through the reactor m, kg.	0.470	0.432	0.448	0.450
2	temperature of solution at the input of the reactor t_1, degrees	22	22	22	22
3	temperature of the solution at the output of the reactor t_2, degrees	66	66	65	65.67
4	temperature difference of the solution $\Delta t = t_2 - t_1$, degrees 44	44	43	43.67	
5	durability of the experiment $\Delta\tau$, s	300	300	300	300
6	reading of voltmeter V, V	4.50	4.50	4.50	4.50
6'	**Reading of oscillograph , U', V**	**4.47**	**4.47**	**4.47**	**4.47**
7	reading of ammeter I, A	2.1	2.1	2.1	2.1
7'	**Reading of oscillograph , I',** A	**2.2**	**2.2**	**2.2**	**2.2**
8	electric power consumption according to indices of voltmeter and ammeters, $E_2 = I \cdot V \cdot \Delta\tau$, kJ	2.84	2.84	2.84	2.84
9	power the heated solution, $E_3 = 4.19 \cdot m \cdot \Delta\tau$, kJ	79.64	80.01	80.72	80.46
10	**reactor efficiency index K = E_3 / E_2**	**28.04**	**28.17**	**28.42**	**28.21**

Supply voltage and current were measured with the help of a voltmeter, an ammeter and an oscillograph (Fig. 1-4)

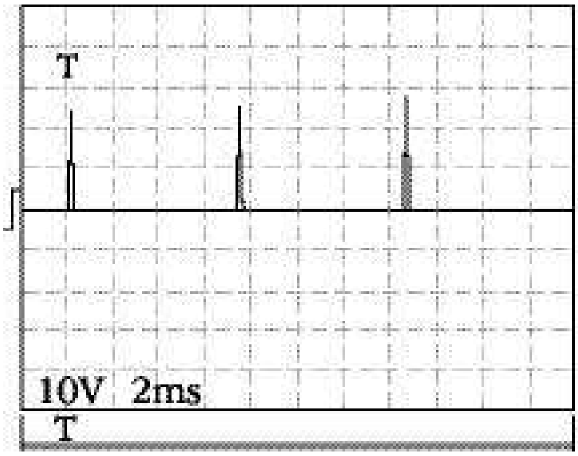

Fig. 1
Voltage

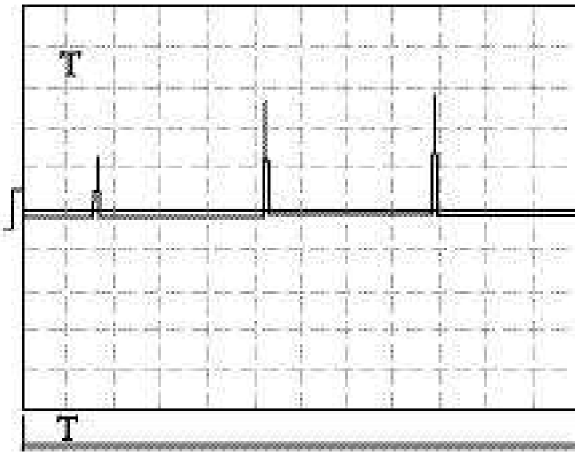

Fig. 2
Voltage

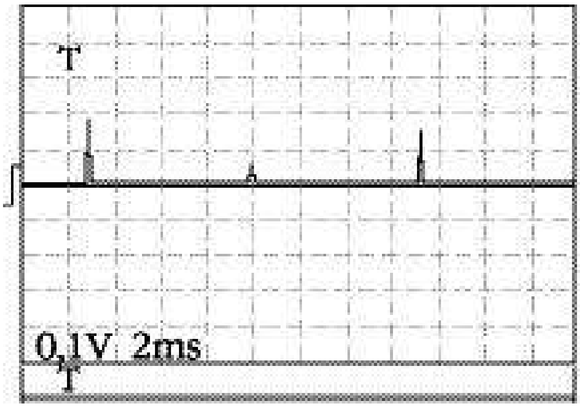

Fig. 3
Current

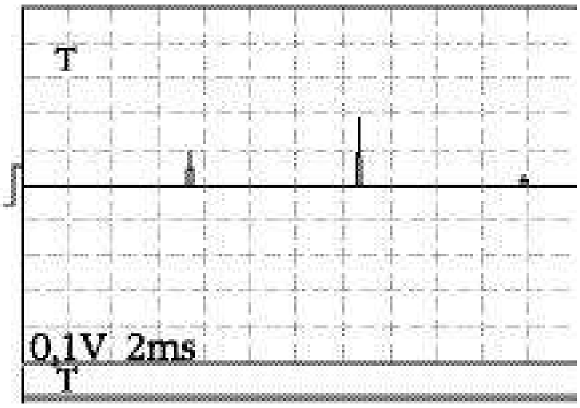

Fig. 4
Current

Process parameter calculation according to the oscillograms (Fig. 1-4) to the check test record (Table 1) gave the following results:

- Pulse scale 10.

- Average voltage amplitude according to Fig.1 and Fig. 2: $Ua_{av} = (23+25+28+10+26+29) \times 10/6 = 235$ V.

- Average current amplitude according to Fig. 3 and Fig. 4: $Ia_{av} = (20+6+17+7+10+19+3) \times 10/7 = 117$ A.

- Pulse repetition period T = 7.4 ms. Pulse duration $t_p = 0.28$ ms.

- Pulse frequency f = 1000/7.4 = 135.1 Hz. Relative pulse duration S = 7.4/0.28 = 26.32.

- Space factor Z = 0.5/26.32 = 0.019.

- Average pulse voltage $U_{av} = 0.019 \times 235 = 4.47$ V.

- Average current in pulses $I_{av} = 0.019 \times 117 = 2.22$ A.

Thus, it is possible to consider that the experimental test of energy efficiency of the water electric heat generator with the help of two methods gives practically identical results and confirms the above-mentioned hypothesis concerning the possibility to generate additional energy in the processes under consideration. It should be noted that as during measurements the pointer instruments of high class of accuracy of 0.2 have been used (relative conventional gauging error does not exceed 0.2%), and oscillographic measurement accuracy is much lower (usually, about 5%), the readings of the voltmeter and the ammeter should be considered more accurate.

Commercial efficiency of the water electric heat generator will depend on pulse generator economy. Since effi-ciency of powerful pulse generators can be near one unit, energy efficiency in industrial plants that use the considered heat generators should not differ greatly from the data obtained during laboratory research.

Simplicity and one hundred per cent reproducibility of the described experiments open a prospect for quick commercialization of the water electric heat generator.

REFERENCES

1. Kanarev Ph.M. The Foundation of Physchemistry of Microworld (the second edition). (In Russian) http://www.ikar.udm.ru/sb28-2.htm
2. Kanarev Ph.M. The Foundation of Physchemistry of Microworld. The second edition. (In English). http://book.physchemistry.innoplaza.net
3. Kanarev Ph.M. Water Electric Generator of Heat. http://Kanarev.innoplaza.net

Professor Kanarev looks for co-projects with prospective investors.

• Fantastic Projects •

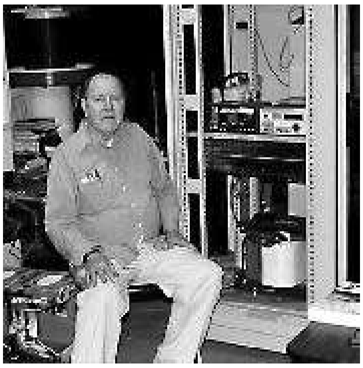

Smith's Generator

Review

TransWorld Energy,
8110 Bent Oak Lane Spring, Texas 77379, USA
Email: donsm1@earthlink.net
http://altenergy-pro.com

Don Smith is a University degreed professional in science & engineering. Served in three wars; WWII, Korea and Vietnam. He has enjoyed a long and successful profession as an executive in the petroleum industry. His life long outside interests include electronics and engineering. A popular walk on substitute for University Professors in physics, chemistry, biology and computer assisted drafting. Teaching style similar to Richard Feynman. He enjoys celebrity status in Japan and Saudi Arabia.

About 20 years ago, the book "Inventions, Researches and Writings of Nicola Tesla", resulted in his reconstructing, as a way of understanding the many theories and devices shown in the book. From this encounter, a very strong bonding resulted. The object then became putting his thinking inside that of Tesla, such that expanding upon and extrapolating areas not yet reached by Tesla. Technological advances in materials provide sourcing for devices not possible at earlier periods. For example, magnetic permeability which is the counter part of negative resistance is an open field for experimentation today.

Don does not sell his inventions, but trades licenses for shares in companies which then incorporate the new technology into their marketing systems. As such, he is on the Board of directors of several well established ventures located in Japan, Brazil, Mexico and Saudi Arabia.

Successful reproductions of Don's technology are present in Finland, Saint Petersburg - Russia, Yugoslavia, Romania, Japan, Hong Kong, and numerous other Countries. His book "Resonance Energy Methods" has a circulation of 40,000 copies in worldwide in many languages.

generators can be near one unit, energy efficiency in industrial plants that use the considered heat generators should not differ greatly from the data obtained during laboratory research.

Simplicity and one hundred per cent reproducibility of the described experiments open a prospect for quick commercialization of the water electric heat generator.

REFERENCES

1. Kanarev Ph.M. The Foundation of Physchemistry of Microworld (the second edition). (In Russian) http://www.ikar.udm.ru/sb28-2.htm
2. Kanarev Ph.M. The Foundation of Physchemistry of Microworld. The second edition. (In English). http://book.physchemistry.innoplaza.net
3. Kanarev Ph.M. Water Electric Generator of Heat. http://Kanarev.innoplaza.net

Professor Kanarev looks for co-projects with prospective investors.

• Fantastic Projects •

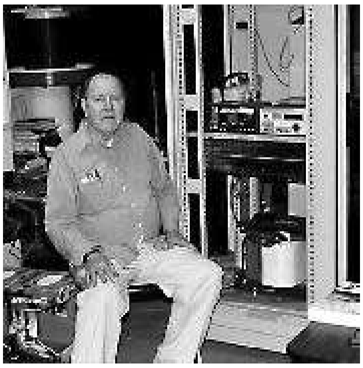

Smith's Generator

Review

TransWorld Energy,
8110 Bent Oak Lane Spring, Texas 77379, USA
Email: donsm1@earthlink.net
http://altenergy-pro.com

Don Smith is a University degreed professional in science & engineering. Served in three wars; WWII, Korea and Vietnam. He has enjoyed a long and successful profession as an executive in the petroleum industry. His life long outside interests include electronics and engineering. A popular walk on substitute for University Professors in physics, chemistry, biology and computer assisted drafting. Teaching style similar to Richard Feynman. He enjoys celebrity status in Japan and Saudi Arabia.

About 20 years ago, the book "Inventions, Researches and Writings of Nicola Tesla", resulted in his reconstructing, as a way of understanding the many theories and devices shown in the book. From this encounter, a very strong bonding resulted. The object then became putting his thinking inside that of Tesla, such that expanding upon and extrapolating areas not yet reached by Tesla. Technological advances in materials provide sourcing for devices not possible at earlier periods. For example, magnetic permeability which is the counter part of negative resistance is an open field for experimentation today.

Don does not sell his inventions, but trades licenses for shares in companies which then incorporate the new technology into their marketing systems. As such, he is on the Board of directors of several well established ventures located in Japan, Brazil, Mexico and Saudi Arabia.

Successful reproductions of Don's technology are present in Finland, Saint Petersburg - Russia, Yugoslavia, Romania, Japan, Hong Kong, and numerous other Countries. His book "Resonance Energy Methods" has a circulation of 40,000 copies in worldwide in many languages.

Abstract

Mr. Don Smith has discovered that Ohm's Law of Resistance does not apply to Magnetic Resonance that travels unrestricted for great distances. Therefore, multitudes of electrons are disturbed. The Magnetic moments are translated into usable electric energy. The Magnetic Dipole provides an unlimited source of electrical current. Mr. Smith's "Dipole Transformer" has been pending since last September.

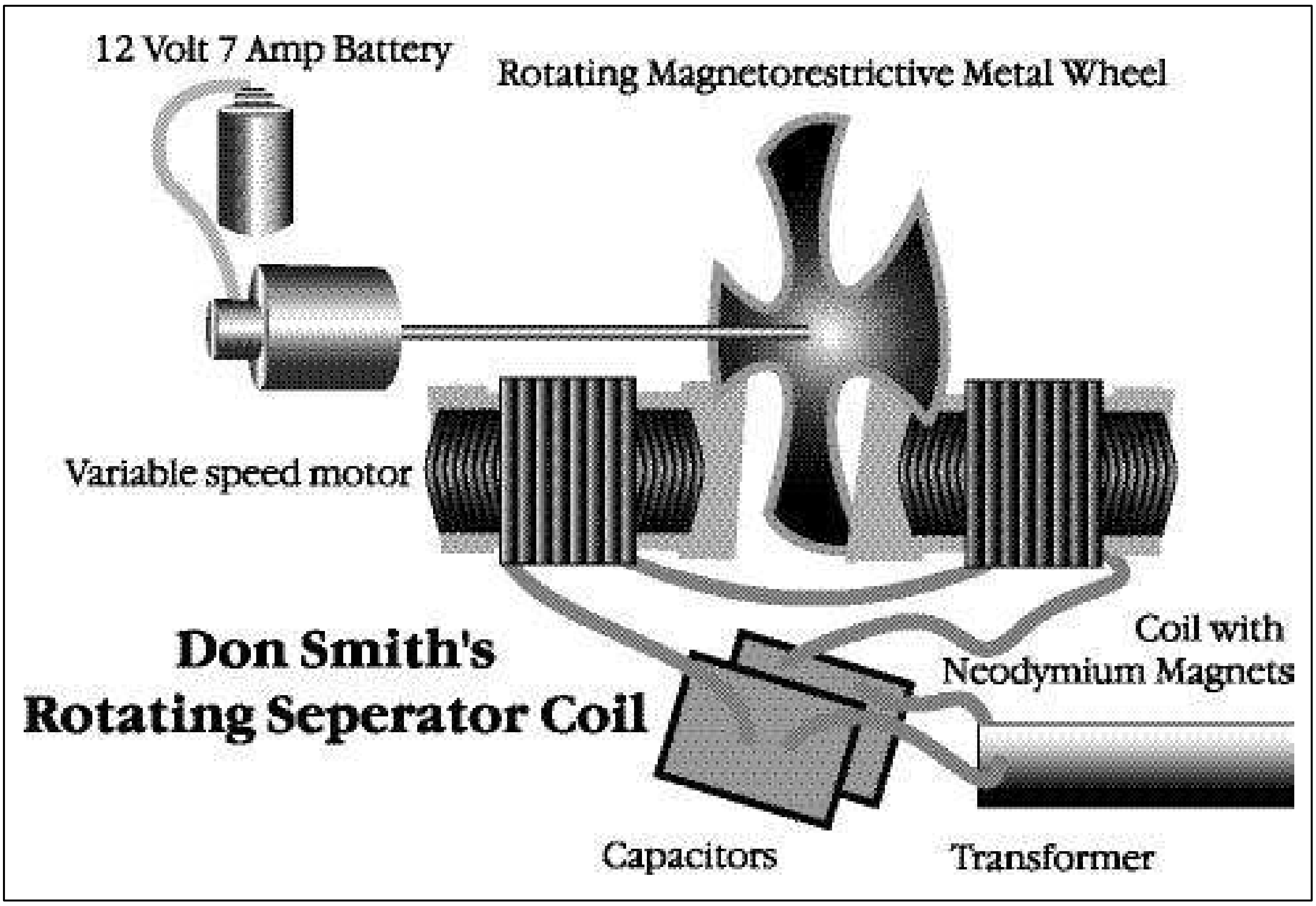

Fig.1
Smith's Energy Receiver

Useful energy occurs as the result of imbalances of ambient background energy and it is a transient phenomenon. In conventional circuits, the electrical field is in a closed system that is damped out with heat loss, which severely limits its utility. **The flip side of the electron generates magnetic waves, which is an open system, not subjected to heat loss. These being unrestricted are the universal source of energy. Universally present electrons are flipped to their electrical position, resulting in useful energy.** When properly constructed an electrical circuit can become self sustainable once started. An obvious corollary would be that the number of radio or television receivers in operation does not deplete the magnetic wave source providing the service. The key source of unlimited energy is Magnetic Resonance. To understand this requires putting a stake through outdated physics. An obvious example is the piano where the key impacts the one note giving one sound, which resonates with its two side keys providing a much higher level. Magnetic Resonant Energy clearly amplifies demonstrating more energy out than in. **Excess energy consequent to a reaction is called Free Energy and can be very large as is in the case of Magnetic Resonance.** The intentionally ignorant physicist makes a habit of ignoring this fact.

Useful electricity is obtained from disturbed electrons, which radiate magnetic fields and waves. The generator disturbs and collects this

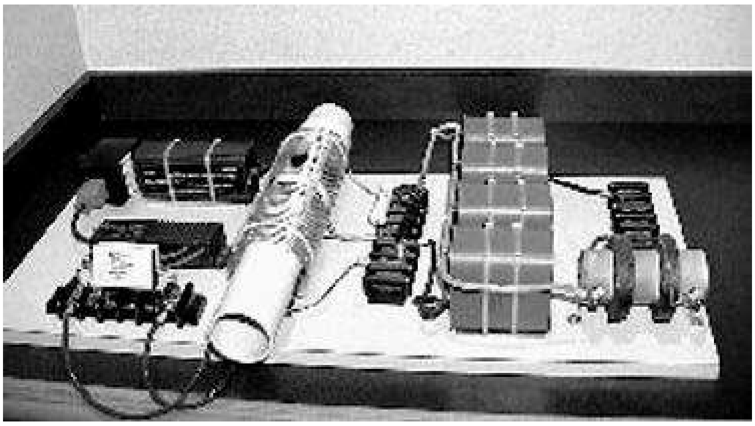

Fig.2
D. Smith's Energy Receiver

emitted electron energy. These electrons remain fully in tact and undiminished until the end of time. This Energy is free and the cost of accessing it depends on the ignorance of the collector. With an appropriate collector system an unending source of environmentally benign, inexpensive, available energy everywhere, is provided.

Editor: in September 2003 Donald Smith took part in the annual "Inventor's Week-end" Conference where he presented the report on his invention. In this issue you can find more details on the conference.

It is worth of a note that we have received ***a number of negative comments on this inventor's work****. However, we thought his articles to be quite interesting and decided to publish them. We would be glad to get readers' comments, who are familiar with the work and inventions of Donald Smith.*

NEWS

News from Farady Lab

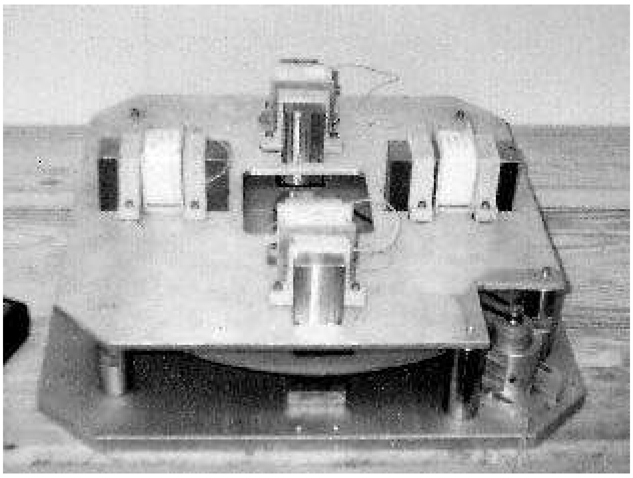

Disk permanent-magnet generator
Faraday Lab Ltd

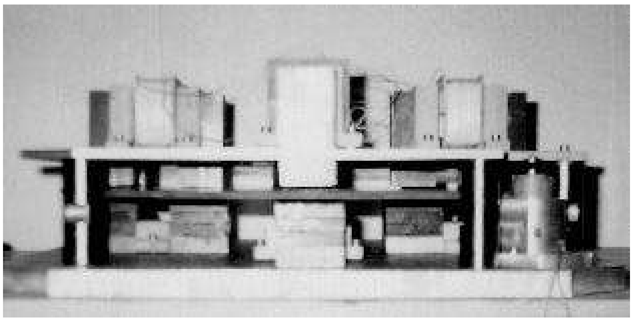

In our lab we are still working at the optimization of the alternator design (electromagnetic generator with the permanent magnets). We have introduced the drum rotor instead of the disk rotor into the new design of the alternator. The operation mechanism is the same, namely, the alternation of the magnetic flux near the generator coils occurs as a result of the rotating elements operation of the rotor, which "shunt" the magnetic flux.

Frolov A.V.

Strategies for Launching a New Over-Unity Device

Ryan S. Wood, USA
14004 Quali Ridge Drive, Broomfield, CO 80020

Problem description: With any singly successful over-unity energy device comes a host of business, technical and financial problems and opportunities. A public announcement permanently will lead to a "Pons & Fleishman" effect that at best delays public success and at worst derails the entire technological effort. Careful attention needs to be applied to the marketing launch, replication, patents, legal defense and finances.

Potential Solution: With a few working devices that generate100-500 horsepower it is relatively easy to hook these units to the electrical gird and generate monthly checks from the utilities. The law states that major utilities must purchase electricity from any supplier at "cost avoidance" rate now typically, between 6 and 10 cents per kilowatt-hour. This law, California Public Utilities Commission - Rule 21, for example, is designed to help small hydro-electric producers and solar or wind farms. There is a safety requirement of grid transfer switches, but they are not terribly complicated or expensive. The entire equipment set-up must be approved by the utility, however, if you start with a diesel electric generator or small hydro plant it can be approved, operating and stable before the clandestine switch or hybrid operation with an over-unity machine. The law is strongly supportive of the energy producer and disconnection for non-safety or maintenance reasons is exceeding rare.

A key benefit of this approach is that you have established repeatability by having multiple working sites, you have established that it can generate money and the technology does real work and you have created several independent locations where press and scientists can visit to validate the over-unity effectiveness in-situ.

Typical revenues would be as follows:

Generator Horse Power	100	200	300	400	500
Annual $, 90% uptime @ 6 c/kWh	$ 35,100	$ 70,199	$ 105,299	$ 140, 398	$ 175, 498
Monthly payment @ 6 c/kWh	$ 2,92	$ 5,850	$ 8,775	$ 11,700	$ 14, 625
Annual $, 90% uptime @ 10 c/kWh	$ 58,499	$ 116, 999	$ 175, 498	$ 233,997	$ 292, 496
Monthly payment @ 10 c/kWh	$ 4,875	$ 9,750	$14, 625	$ 19,500	$ 24, 375

It will take money without the strings of investors to champion and defend a new breakthrough technology into the energy marketplace and this strategy offers a quiet way to prepare for the ultimate launch and resulting scientific and media frenzy that will ensue. It will take years for the business to form, products to be built and marketing and acceptance and industry confidence to be established. All during those years this strategy can generate cash to support engineering and operations.

Please call if there are questions at 303-941-9663 (cell),
Home: 720-887-8071
Ryan S. Wood

• Fantastic Projects •

The Hutchison Effect – An Explanation

Mark A. Solis, USA
http://www.geocities.com/ResearchTriangle/
Email: your_neighbor_2000@yahoo.com

People often ask, "What exactly is the Hutchison Effect?" This brief essay is an attempt to answer that question to the satisfaction of the majority. First of all, the Hutchison Effect is a collection of phenomena which were discovered accidentally by John Hutchison during attempts to study the longitudinal waves of Tesla back in 1979. In other words, the Hutchison Effect is not simply a singular effect. It is many.

The Hutchison Effect occurs as the result of radio wave interferences in a zone of spatial volume encompassed by high voltage sources, usually a Van de Graff generator, and two or more Tesla coils.

Editor: We can call this effect "The H. Wells Effect" because he was first to describe it in his book "The Invisible man". The main character of the book used two sources of electromagnetic vibrations and Wells said that they were not Hertz waves but something else. "But the essential phase was to place the transparent object whose refractive index was to be lowered between two radiating centers of a sort of ethereal vibration..."("The Invisible Man", chapter 20).

A. V. Frolov

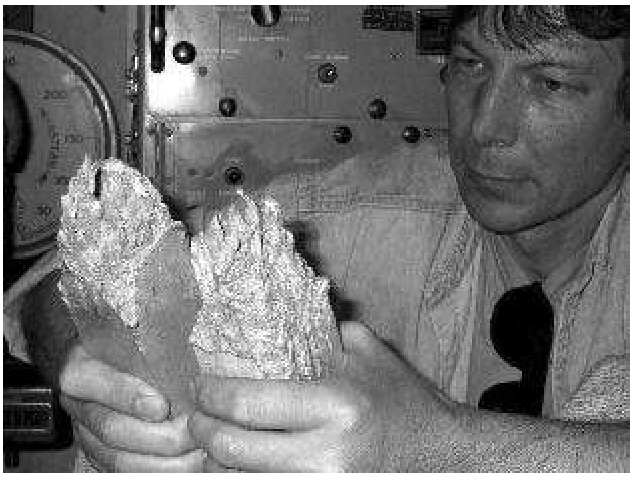

The effects produced include levitation of heavy objects, fusion of dissimilar materials such as metal and wood, the anomalous heating of metals without burning adjacent material, spontaneous fracturing of metals (which separate by sliding in a sideways fashion), and both temporary and permanent changes in the crystalline structure and physical properties of metals.

Fig.1 represents the proof of Hutchison Effect based on metal blocks. Deformation occurred at room temperature as a result of complex interaction of electromagnetic fields.

The levitation of heavy objects by the Hutchison Effect is not - repeat not - the result of simple electrostatic or electromagnetic levitation. Claims that these forces alone can explain the phenomenon are patently ridiculous, and easily disproved by merely trying to use such methods to duplicate what

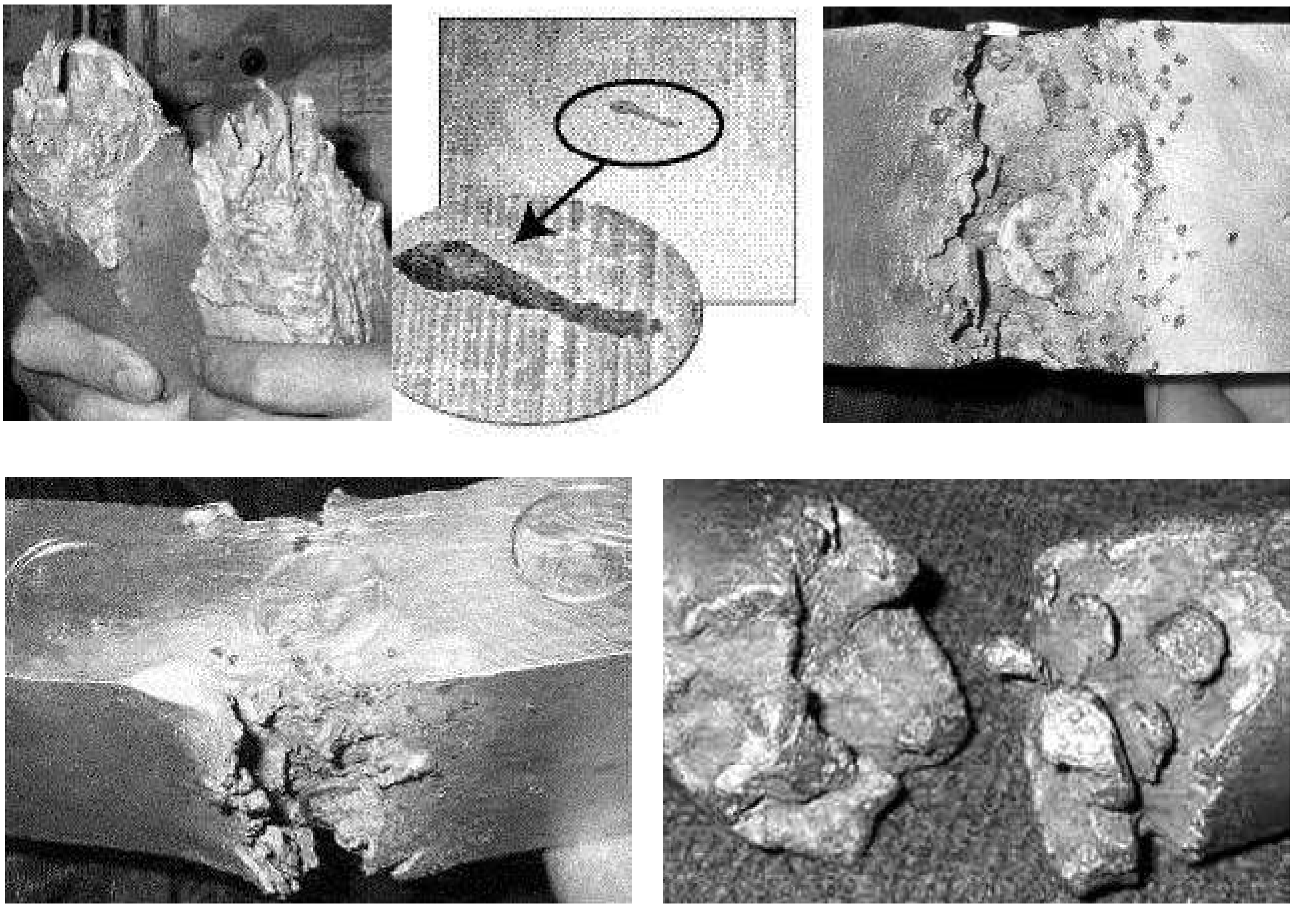

Fig.1

Cold melting of metals

The upper left: steel, The lower left: aluminium with coin prints and with a coin inserted in a partially open split; The lower right: fully torn aluminium bar; The middle upper and right: a section part of the aluminium block, into which a wood block is fused (brown matter)

http://www.rumormillnews.com/JOHN-HUTSHISON2.htm

the Hutchison Effect has achieved, which has been well documented both on film and videotape, and has been witnessed many times by numerous credentialed scientists and engineers. Challengers should note that their apparatus must be limited to the use of 75 Watts of power from a 120 Volt AC outlet, as that is all that is used by Hutchison's apparatus to levitate a 60-pound cannon ball.

The fusion of dissimilar materials, which is exceedingly remarkable, indicates clearly that the Hutchison Effect has a powerful influence on Van der Waals forces. In a striking and baffling contradiction, dissimilar substances can simply "come together," yet the individual substances do not dissociate. A block of wood can simply "sink into" a metal bar, yet neither the metal bar nor the block of wood come apart. Also, there is no evidence of displacement, such as would occur if, for example, one were to sink a stone into a bowl of water.

The anomalous heating of metal without any evidence of burning or scorching of the adjacent materials (usually wood) is a clear indication that possibly the nature of heat may not be completely understood. This has far-reaching implications for thermodynamics, which hinges entirely on the presumption of such knowledge. It should be noted that the entirety of thermodynamics is represented by the infrared portion of the electromagnetic spectrum, which is insignificant in a context of 0 Hz to infinite Hz. The anomalous heating exhibited by the Hutchison Effect shows plainly that we have much to learn, especially where thermodynamics and electromagnetism meet.

The spontaneous fracturing of metals, as occurs with the Hutchison Effect, is unique for two reasons: (1) there is no evidence of an "external force" causing the fracturing, and (2) the method by which the metal separates involves a sliding motion in a sideways direction, horizontally. The metal simply comes apart.

Doctor John Hutchison

Some temporary changes in the crystalline structure and physical properties of metals are somewhat reminiscent of the "spoon bending" of Uri Geller, except that there is no one near the metal samples when the changes take place. One video shows a spoon flapping up and down like a limp rag in a stiff breeze. In the case of permanent changes, a metal bar will be hard at one end, like steel, and soft at the other end, like powdered lead. Again, this is evidence of strong influence on Van der Waals forces.

The radio wave interferences involved in producing these effects are produced from as many as four and five different radio sources, all operating at low power. However, the zone in which the interferences take place is stressed by hundreds of kilovolts.

It is surmised by some researchers that what Hutchison has done is tap into the Zero Point Energy. This energy gets its name from the fact that it is evidenced by oscillations at zero degrees Kelvin, where supposedly all activity in an atom ceases. The energy is associated with the spontaneous emission and annihilation of electrons and positrons coming from what is called "the quantum vacuum." The density of the energy contained in the quantum vacuum is estimated by some at ten to the thirteenth Joules per cubic centimeter, which is reportedly sufficient to boil off the Earth's oceans in a matter of moments.

Given access to such energies, it is small wonder that the Hutchison Effect produces such bizarre phenomena. At the present time, the phenomena are difficult to reproduce with any regularity. The focus for the future, then, is first to increase the frequency of occurrence of the effects, then to achieve some degree of precision in their control.

The spontaneous fracturing of metals, as occurs with the Hutchison Effect, is unique for two reasons: there is no evidence of an "external force" causing the fracturing, and the method by which the metal separates involves a sliding motion in a sideways direction, horizontally. The metal simply comes apart.

Editor: On November 17th 2003 we got to know that the inventor John Hutchison can eventually lose his home laboratory. The property can be confiscated upon court order due to neighbours' complaints. Apparently, it is dangerous to conduct experiments of the kind at home and even within a town. We wish him success!

A. V. Frolov

A SIMPLE DEVICE FOR HEAT TO ELECTRICAL ENERGY CONVERSION BY MEANS OF FERROELECTRICS

A. Ries, A. Z. Simoes, M. A. Zaghete, J. A. Varela,
Institute of Chemistry, Universidade Estadual Paulista, Rua Prof. Francisco Degni s/n,
14801-970 Araraquara-SP, Brazil
Email: ries@posgrad.iq.unesp.br

E. Longo,
Department of Chemistry, Universidade Federal de Sao Carlos, Rodovia Washington Luis - km 235, 13565-905
Sao Carlos-SP, Brazil

Introduction

A few years ago, Zaev [1, 2] demonstrated experimentally the possibility of heat to electrical energy conversion by means of ferroelectrics. A capacitor filled with a ferroelectric material was charged and discharged periodically at a temperature just a little lower than the Curie point. According to Zaev, the capacitor must have a nonlinear capacitance dC/dV > 0, which means the capacitance increases with rising voltage. Besides these experimental results, it was proven by a theoretical calculation that the electrical energy obtained during discharging can reach up to 1.35 of the energy introduced to charge the capacitor. The present paper presents an inexpensive and simple electric circuit which can be used to prove heat to electricity conversion experimentally. Moreover, the physical mechanism of heat to electricity conversion is explained in the form of a thermodynamic cycle.

Thermodynamic cycle for energy conversion

The following thermodynamic cycle is composed of four reversible steps. It works only at a temperature just a little lower than the Curie point, where the dielectric constant is highly temperature dependent. Fig. 1 shows the relative dielectric constant of a representative ferroelectric ceramic as a function of temperature.

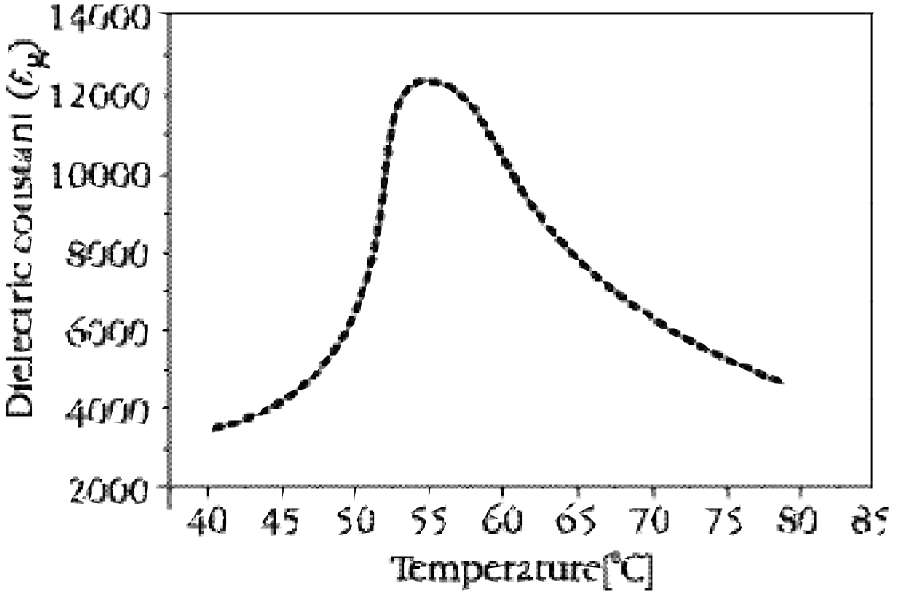

Fig. 1
Relative dielectric constant versus temperature for a barium strontium titanate ceramic (80% Ba, 20% Sr) prepared by the author. Curie point = 55⁰ C.

As it can be seen, just below the Curie point in the range from 50 to 55^{0}C, small changes of the temperature cause high changes of the dielectric constant.

Furthermore, to understand the conversion mechanism, it is necessary to understand the electrocaloric effect. When an electric field is applied to a dielectric medium, the latter is polarized. For ferroelectric materials, the parallel domain alignment is the most important part of the total polarization. During the successive alignment of the domains, some material parameters in particular entropy, heat capacity and temperature change. Since the material possesses less degrees of freedom in the polarized state, its entropy and heat capacity are reduced. As a consequence, provided that no heat exchange with the environment is possible (adiabatic conditions), the temperature increases. It is very important to understand that the energy for warming of the dielectric is not taken from the polarizing electric field. The increase in temperature is a result of the decreased heat capacity only.

From this, it can be derived easily what means "nonlinear capacitance".

When a capacitor is charged adiabatically which means its voltage increases faster than any heat exchange with the environment is possible, due to the electrocaloric effect, the temperature of the dielectric must also increase. As it can be seen from Fig.1, below the Curie point, an increasing temperature leads to a higher dielectric constant. Assuming that both temperatures, the one before charging and the other after charging are still below the Curie point, one can say that the dielectric constant increases with an increase in voltage.

Since the capacitance is proportional to the dielectric constant, the capacitor has a nonlinear characteristic $dC/dV > 0$. In the same way it can be concluded that adiabatic charging just a little above the Curie point reveals a nonlinear capacitor characteristic with $dC/dV < 0$.

Quantitative experimental measurements on the electrocaloric effect in some representative ferroelectric and antiferroelectric materials were published by Thacher [3].

STEP 1: Adiabatic charging of the capacitor

We consider a capacitor filled with a ferroelectric medium at an environmental temperature a little lower than the Curie point.

An adiabatic charging of a capacitor (no heat exchange) effects an increase in temperature and capacitance due to the electrocaloric effect. We assume that this increase in temperature is so small that the temperature remains below the Curie point. Finally, the capacitor is completely charged while the dielectric medium has a higher dielectric constant and therefore a higher capacitance than it posseses at the environmental temperature.

To go further to step 2, the capacitor is hold under conditions where no discharge is possible, e.g. it is disconnected from the power supply.

STEP 2: Thermal equilibration

After a short time, the charged capacitor has cooled down to the environmental temperature. While cooling down, the dielectric constant and capacitance decrease. But the charges of the capacitor plates remain constant. As a consequence, the voltage and the electrical energy increase. This can be readily recognized from the general capacitor equation $q = C V$ (q = charge, C = capacitance, V = voltage):

$$q = \text{constant} = C_1 V_1 = C_2 V_2$$

STEP 3: Adiabatic discharging of the capacitor

While discharging the capacitor the temperature and capacitance decrease (again heat capacity changes) due to the electrocaloric effect, leading to a further increase in the available electrical energy. The ferroelectric medium posseses now a final temperature below the temperature of the environment.

STEP 4: Thermal equilibration

In order to proceed with step 1, it is necessary to reach the environmental temperature. An inflow of heat from the environment to the dielectric material is required.

Electric circuit to observe energy conversion

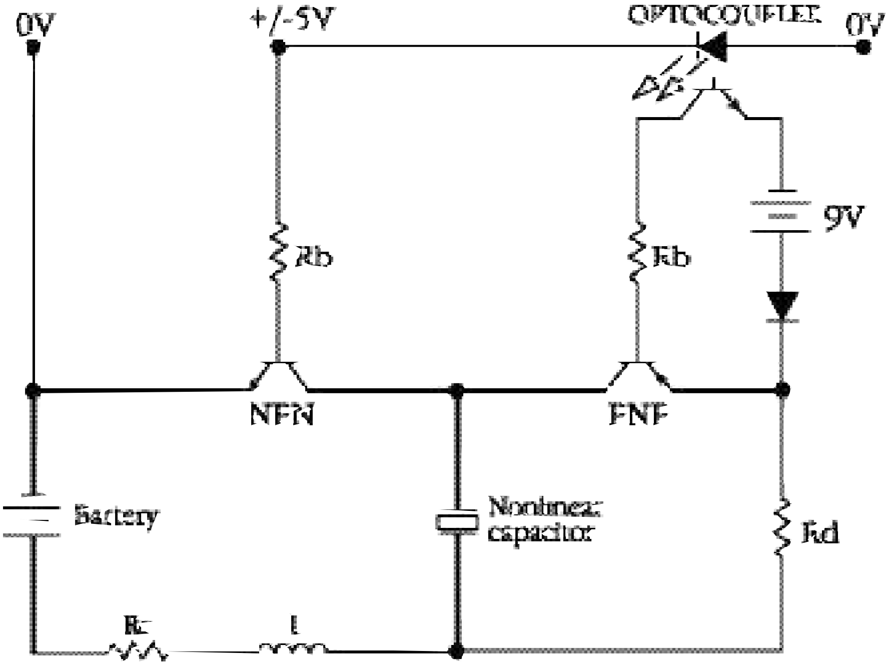

Fig. 2
Electric circuit for testing the energy conversion effect. The connections to the square wave generator are labeled with 0V and +/- 5V, the resistors Rb are to adjust the maximum base current of the transistors.

Assuming the experimenter has a nonlinear capacitor, the set-up shown schematically in Fig. 2 is proposed as a possible conversion device. The circuit contains two different types of transistors (NPN and PNP) which fulfill here the function of a switch.

The NPN junction transistor consists of two n-type semiconductors (called the emitter and collector) separated by a thin layer of p-type semiconductor (called the base). On the other hand the PNP junction transistor consists of a thin layer of n-type semiconductor lying between two p-type semiconductors. The base is the ON/OFF switch for the transistors. If a current flows to the base, there is a path from the collector to the emitter, where a current can flow (switch is ON). If there is no current flowing to the base, then no current can flow from the collector to the emitter (switch is OFF).

The charging-discharging process is controlled by a square wave generator which produces an output signal switching from +5V to -5V. The switching frequency is adjustable from 0.1 Hz up to at least 10 kHz. This output signal is applied to the bases of both transistors.

During the time the square wave generator output signal is +5V, a base current flows through the base-emitter junction of the NPN transistor which creates a low resistance path between the collector and the emitter. As a consequence, the capacitor is charged by the battery. The collector-emitter junction of the PNP transistor has a very high resistance, because here no base current flows. Therefore the capacitor is not discharged at the same time. When the output signal switches to -5V, a current flows to the base of the PNP transistor and not to the base of the

NPN transistor. Now the capacitor is discharged.

In order to understand the function of the inductance in Fig. 2, we analyze the thermodynamics of the charging process.

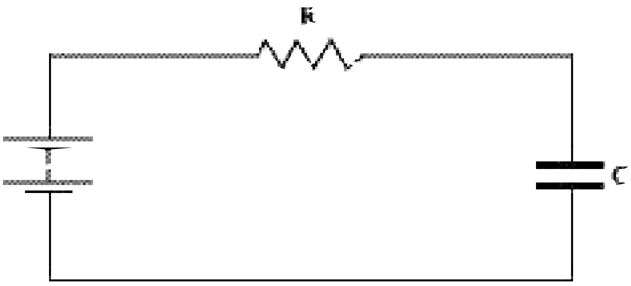

Fig. 3
Simple circuit for charging a capacitor

If a linear capacitor of capacitance C is charged to a voltage V through the load R according to Fig. 3, the energy $W = 0.5\ C V^2$ is stored in the capacitor. During the charging process, a time-dependent current I(t) flows:

$$I(t) = V/R \exp(-t/RC)$$

This current develops the heat Q in the load:

$$2Q = R \int I^2(t)dt = 0{,}5CV^2$$

Here the integration limits are from zero to infinity. This calculation shows, that the energy taken from the battery is the energy stored in the capacitor after charging plus the heat produced in the load, also 2•C V2. One can see that the heat Q is not dependent on the resistance value R.

The above used equations are only valid in case of a one-step charging process.

If the capacitor is charged stepwise to the final voltage V, e.g. using a ramp generator, and each voltage step effects a voltage increase , the total heat produced in the load is (N = number of voltage steps):

$$Q = N\ 0{,}5\ C\ (V/N)^2 = 0{,}5\ C/N\ V^2$$

Now the energy taken from the source is:

$$W = 0{,}5 \cdot (1 + 1/N) 1/2\ CV^2$$

If the number of steps N tends to infinity, the heat Q tends to zero.

This energy loss in the load dramatically influences the efficiency of any conversion device and must be taken into account, otherwise all experimental attempts to observe energy conversion would fail.

A more detailed discussion on charging a capacitor and the unavoidable energy losses was given by Heinrich [4] and Gupta et al. [5].

The inductance in the circuit (Fig. 2) effects a slower rising of the current during the charging process (due to self induction) and can reduce the heat generated in the load dramatically.

Acknowledgments

Financial support of this work was provided by the Brazilian agency FAPESP (www.fapesp.br) and the German Academic Exchange Service (DAAD), (www.daad.de).

References

1. N.E. Zaev, Yu.S. Spiridonov, Capacitance as a converter of heat to electrical energy, Russian Electrical Engineering, Vol. 69, (2000) pp. 69-72
2. N.E. Zaev, Research on the Capacitance Converter of Environmental Heat to Electric Power, New Energy Technologies, issue 2, (2001)
3. P.D. Thacher, Electrocaloric effects in some ferroelectric and antiferroelectric Pb(Zr, Ti)O3 compounds, Journal of Applied Physics, Vol. 39, (1963), pp. 1996-2002
4. F. Heinrich, Entropy change when charging a capacitor: A demonstration experiment, American Journal of Physics, Vol. 54, (1986), pp. 742-744
5. V.K. Gupta, G. Shanker, N.K. Sharma, Reversibility and step processes: An experiment for the undergraduate laboratory, American Journal of Physics, Vol. 52, (1984), pp. 945-947

• Fantastic Projects •

The Gates Motor

A. Akau, USA
Email: adrianakau@aol.com

The Gates motor company has produced a motor that does not run on fossil fuel or on electricity. It comes in a 28 HP and a 5,000 HP version. When linked to an electric generator, the larger motor will produce 2,200 kW.

My name is Adrian Akau. I am a recently retired school teacher, having taught science and math in the public school system, state of Hawaii for the past 36 years. One of the classes I taught was on energy.

Our world should be reaching the half-way mark in easily accessible oil by about the year 2016. Since the rate of oil usage is becoming higher each year, it is unlikely that the oil supply will last as long as the 140 years since the first well was drilled in Pennsylvania. We will become more dependent upon OPEC with all the political pressures and problems that accompany this type of dependency. As the easy oil is consumed, we must look for ways to extract less accessible oil such as from shale. The added expense will be passed on to the consumer.

The burning of coal is not the answer either because coal cannot supply convenient energy as does oil. An oil-based economy cannot transfer over readily to natural gas; there is not enough natural gas available and it is again, not a convenient and concentrated source as is oil. We also have serious pollution problems with the burning of fossil fuels (natural gas being the exception) but the burning of all fossil fuels have contributed to the serious problem of global warming.

It may take all of 50-80 years to transfer to a non-fossil fuel economy. The transfer must be made slowly so as not to disturb the infrastructure of our economy which is presently based upon fossil fuels. We must gradually create a new infrastructure that is non-fossil fuel dependent.

> *The motor incorporates a unique spring configuration to provide rotation*

The production model of the motor is 24 inches high by 48 inches in length. Connected to a generator it will provide sufficient power for about eight hundred US homes. It is designed to be tough and to last a long time. It uses a high grade airplane synthetic engine oil to prevent friction and the build up of heat. Only the oil filter needs to be changed. The motor has special needle bearings and portions of the motor are made of high quality steel.

The Gates motor is designed to run 24 hours/day, 365 days/year without incurring any fuel costs. The larger size motor would come in pairs; the spare motor being used as a back-up unit should the regular motor be turned off for maintenance such as to change the oil or in case there should be any problems with the first motor. The company is willing to enter into a contract for maintaining the motor after installation.

The Gates motor represents a major breakthrough in motor design technology and provides practical solutions to environmental and fuel conservation concerns. **Finally, a completely mechanical motor has been developed.** It can deliver the power needed in a wide variety of service applications without any waste products or emissions to adversely affect the environment. This revolutionary new motor utilizes spring power technology. Unlike conventional motors which must reach a maximum rpm level before the desired horsepower, the Gates motor provides maximum horsepower instantly by virtue of the torque stored in the springs.

The basic operating principle of the Gates motor involves a series of springs configured to provide the required motor rotation and power delivery. Consider, if you will, the operating principle of a grandfather clock. Once the clock is wound, it continues to operate until it runs down and stops. The springs must then be rewound so it will operate again. Suppose the clock is continually being rewound as it operates. Then the clock would run continuously. This condition is what forms the basis for the operating principle of the Gates motor. The motor incorporates a unique spring configuration to provide rotation. The springs inside the motor are wound and preset at the factory at the time the motor is assembled. The amount of torque wound into the springs at the factory determines the horsepower of the motor. Inside the motor are a multiple of springs positioned horizontally in a circular arrangement. As the motor operates, the springs are unwound a set number of degrees. A double ratchet system at one end of the motor rewinds the degrees used back into the springs. As the springs are unwinding, **the power generated from all the springs reset each spring (one at a time) during each revolution of the motor**. It is this reset action that produces and delivers power from the springs to the flywheel located at the opposite end of the motor. The flywheel serves to ensure the smooth operation of the motor and to convert the springs into useable horsepower. The smaller motor has 50 foot pounds of torque at a shaft speed of 3000 RPM and will produce 28 HP. The HP can be altered by factory adjustment. The speed of the engine is controlled by a hydraulic pump which provides pressure upon the power shaft.

Contacts: Gates Motor Corporation,
P.O. Box 715, Hauula, HI, USA 96717
Email:gagates@earthlink.net

Editor: We tried to get the video of this device operation from Gates Motor Co., but after a long consideration it was not provided since the authors are competitor-conscious.

Alexander V. Frolov

Cold Electricity

Adrian Akau, USA
Email: adrianakau@aol.com

Dr. Peter Lindemann in his book "The Free Energy Secrets of Cold Electricity" presents the story of this type of electricity which really is not electricity as electricity is normally meant to be understood. Cold electricity behaves differently than normal electricity. Regular electricity may be used in the process to generate this cold form but this transformation is not the type of change in voltage and current that occurs in a transformer. Rather, it is the extraction of a form of current from "normal" electricity by a process that uses high voltage.

In discussing the Edwin Gray motor, Dr. Lindemann relates a demonstration by Gray. Gray used a 6 volt car battery with lead wires running to a system he had devised using a voltage booster and a series of capacitors which raised the voltage to 3,000 volts. He then closed a switch that ran the high voltage into two electromagnets which caused a loud popping sound and cause the top electromagnet weighing a pound and a quarter to be propelled over two feet into the air. Gray claimed that only 1% of the system's energy was used with 99% going back to the battery. He said that he had 'split the positive' with his system.

Other evidence of using this fractionated electricity was Gray's use of a small motorcycle battery (15 amps which would normally produce W=V x A = 6 V x 15 A= 90 watts) to run concurrently six 15-watt electrical light bulbs, a portable 110 volt T.V. set and two radios. A glowing 40 watt light bulb running off the system was dropped into water without the glass breaking; the bulb gave off light but not the heat that accompanied a bulb run by regular electricity. This would mean that tungsten in the filament of the light bulb was not resisting the flow of this fractionated electricity and that this "cold electricity" was causing the luminous glow by some other means.

According to Gray's Patent #4,595,975, low voltage was chopped into pulsating DC by passing it through a multi-vibrator (buzzer like a door bell). Then the pulsed DC was sent through the low voltage winding (primary) of a transformer which changed it to pulsating high voltage DC current at the secondary side. The pulsating high voltage DC was rectified by a full wave bridge and changed into high voltage DC.

The high voltage DC was used to repeatedly charge a capacitor as the current was sent to briefly discharge across a spark gap (rated at 3000 volts). The discharge of the spark across the gap had to be in one direction only and its duration was controlled by the size of the capacitor and by the strength of a magnetic field encompassing the gap. This magnetic field had a quenching effect because it caused a back EMF each time the discharge occurred. The current from the discharge across the gap was then sent through a resistor and then to a vacuum tube (the conversion switching element tube).

Another of Edward Gray's patents "Efficient Electrical Conversion Switching Tube Suitable for Inductive Loads" (April 1987, patent 4,661,747) described the switching tube. It utilizes a low voltage anode (positive plate), a low voltage anode (positive plate) and one or more electro-static or charge receiving grids (located between the positive plates and the cathode or negative plate supplying the electrons). This tube's function was to 'split the positive'. A normal vacuum tube usually uses only a single anode at a particular voltage to control the voltage flow within the tube. The function of the charge-receiving grids was to collect the "cold electricity".

The most unusual part of this Power Supply circuit consisted of devices specifically designed to remove excess energy when the Switching Tube was functioning. A spark-gap protection device, served to "protect the inductive load and the rectifier elements from unduly large discharge currents". As stated, the rectifier elements are the two anodes and the grids in the Switching Tube. In a normal vacuum tube, no protection would be necessary because the maximum power of the tube would be based upon the electricity provided to heat the tube filament and to charge the anode (voltage x amps). Therefore, the excess energy comes from the process occurring as the spark gap affected electrical pulses pass through the Switching Tube. The process of extracting the "cold electricity" places danger of overload upon the system. (Please take note that the "electricity" from the Switching Tube is no longer the normal type of electricity we have from our batteries or from an electrical outlet but rather a completely new form with entirely different characteristics.)

The description continues, "Should the potentials (voltages) within the circuit exceed the predetermined values fixed by the mechanical size and spacing of the elements within the switching tube, the energy is dissipated (bypassed) by the protective device to the circuit common (electrical ground)" by two strategically placed diodes (devices that permit the passage of electricity in one direction only). The question now arises "What is the cause of this large amount of excess energy which must be dumped to ground through device acting like a lightning rod in order to prevent the circuit from burning out?"

To examine the cause of the excess energy source, we must go back over a hundred years to 1889. Heinrich Hertz had just announced in 1887 that he had discovered electromagnetic waves and Nicola Tesla was attempting to duplicate Hertz's experiments. Tesla used abrupt and powerful electric discharges produced with banks of capacitors charged to very high potentials and was able to explode thin wire (copper bus bars). He came to the conclusion that Hertz had mistakenly associated electrostatic inductions (electrified shockwaves in air) for true electromagnetic waves.

The explosion of the copper bars by means of the "disruptive discharges" from the capacitor bank produced sharp shockwaves which struck Tesla with great force across the entire front of his body. Tesla said they felt more like powerful gunshots rather than electrical sparks. They produced effects similar to lightning or to those produced by high voltage DC generators previously discussed; the simple closing of a high voltage DC generator caused a stinging shock. AC generators were not in use in this pre-AC era and it was shown, later on, that AC generators did not produce these effects.

This shock was first assumed to be the result of residual static charging. It stood straight out of highly electrified conductors, seeking ground paths which included workmen and switchboard operators. In long cables, Tesla estimated that this electrostatic concentration was several orders in magnitude greater than any voltage the DC generator could actually produce. It caused a hedge or crown of bluish needles or spicules to point at right angles to the cable or straight from the electrical cable line into the surrounding space. The bluish needles appeared the very instant the switch was closed and disappeared a few milliseconds later, after which the system functioned normally. However, anyone through whom the blue needles passed, especially in large regional power systems which used high exceptionally high voltages, usually did not survive. Generators rated at a few thousand volts produced hundreds of thousands or even millions of electrostatic volts during the start-up pulse. Highly insulated, heavily grounded relay switches had to be installed to protect workers from certain death.

Engineers at this time assumed **the effect to be caused by a "bunching" action which occurred when a powerful force was not able to move charge sufficiently quickly through a system.** (A similar "choking effect" also happened with large steam engines; if steam was introduced too rapidly, the steam engine could explode.) The metal of which the wire was composed somehow offered resistance to the charge carriers before they were able to move away from the generator terminals. It seemed that the wire acted as a blocking force to the electrons or to some unknown part of the electrons for the first few milliseconds, rather than as a conductor. The powerful, deadly bluish spikes sprang from the line until the current charges actually "caught up" with the applied electrical field. It was as if the spicules were a form of relieving pressure on the system by converting the voltage impulse into some other form.

Tesla began wondering why the electrostatic field moved faster than the actual charges. He believed that this effect could help him find electric waves better than his capacitors because the line resistance caused the electrostatic charges to bunch together into a density much greater that obtainable with his capacitors.

Tesla understood that ordinary capacitor discharges were oscillating currents or spark currents which "bounced" between each capacitor plate until their stored energy was dissipated. The high surge voltage of the DC generator exerted such a high one way pressure on the dense line charges, due to the millisecond resistance of the metal, that possible voltage "bounces" were also oscillations of current as was true with the capacitors. Tesla used every means possible to block the "back-rush" (bouncing) of the oscillating current in order to prevent the supercharge to prematurely decline; he wanted to maintain this powerful "bunching" effect as long as possible in order to study it and to make use of it.

His face and hands experienced a penetrating shockwave; a sharp pressure and an electrical irritation were brought on at the sudden closing of the switch. His face and hands were especially sensitive to these shockwaves which caused a "stinging" effect at close range. Tesla believed that he was being struck by material particles approaching the vapor state as they were thrust from bluish crowned wire.

Later on, he was to discover they were not gas particles. **He placed himself behind a glass shield but to his amazement, he still felt the shockwaves and stinging effects; the glass shield did not protect him.** These stinging rays could be felt at great distances from their sources and, to his amazement, would penetrate shields of both glass and copper. Tesla knew that normal electrostatic charges spread over the surface of a metal (copper, for example) shield so that the stinging effect could not have an electrostatic origin and therefore was not electrostatic in nature but was from some other yet unknown source.

Joseph Henry in 1842 had noticed the magnetization of steel needles by the discharge of a Leyden Jar (type of primitive condenser made with a glass bottle). The Leyden Jar was at the top floor of a building while the needles were in the basement. Dr. Henry, noting the passage of these magnetizing rays through brick walls, oak doors, stone and iron flooring and tin ceilings believed that the spark given off by the discharge released "light-like rays" that passed through the material obstacles with ease before magnetizing the needles.

Elihu Thomson, a physics instructor in 1872 had been attempting to make sparks from a Ruhmkorff Spark Coil more clearly visible for his physics students. He attached one pole of the coil to a cold water pipe and noted that the previously blue spark changed to white. He then attached the other pole to a large metal table top and produced a silver-white spark that would be clearly visible to all attending his lecture. He went to the door of the room to notify his colleague but received a strong shock from the door knob. Turning off the Ruhmkorff coil prevented the brass knob on the oak door to stop shocking. He returned with his friend, again turned on the coil and discovered that all metal objects in the entire building, no matter how distant from the coil or insulated from the floor, produced long and continuous white sparks upon the touch of a penknife or screwdriver.

The device Tesla perfected was far superior to the Ruhmkorff Spark Coil. The effect of his disrupter device was so strong that a single wire placed in an oil bath produced what he had originally but mistakenly thought to be gaseous streams so powerful that they visibly depressed the oil into a depth of about two inches. Tesla concluded that "besides the air, another medium is present".

> *...he placed himself behind a glass shield but to his amazement, he still felt the shockwaves and stinging effects; the glass shield did not protect him. These stinging rays could be felt at great distances from their sources and, to his amazement, would penetrate shields of both glass and copper...*

Tesla had been able to perfect his device by placing a capacitor between the switch and the D.C. generator thus increasing the power and protecting the generator windings in much the same manner as the diodes Edwin Gray's Switching Tube protected Gray's circuit. Tesla also raised the voltage and quickened the "make-break" rate of the switch to increase the power level by placing a powerful permanent magnet crosswise to the discharge path of the high voltage output of the DC generator wires. The magnetic field caused the discharge arc to automatically "blow out"; the charge passing through the wires in the presence of the magnetic field built up reverse emf (electromotive force or voltage).

Tesla believed he had discovered a new type of electricity with special characteristics. This form of electricity did not consist of alternating waves. **They were longitudinal waves** composed of successive shocking waves with effects that could be seen and felt at a distance. **Vector components of these shock waves were unidirectional. They were able to force charges in the direction of their propagation.** In his patent #787,412, "Art Of Transmitting Electrical Energy Through The Natural Mediums" (April 18, 1905), Tesla calculated the mean velocity of the waves propagated by his device to be 471,240 Km/sec. Knowing that the velocity of light is 300,000 Km/sec would indicate that the type of transmission to which Tesla is referring is different than standard electromagnetic radiation; these special longitudinal "Radiant Energy" waves moved faster than the speed of light.

After conducting hundreds of experiments, Tesla found that "Radiant Energy" longitudinal waves **could penetrate all materials and cause "electronic responses" in metals such as copper and silver.** Impulses exceeding 0.1 millisecond duration produced effects such as pain, mechanical pressures, explosion of thin wires and vibration of objects. Impulses of 1.0 microseconds caused the sensation of physiological heat; at still shorter lengths white light would fill the room. Impulses less than 100 microseconds were safe to work with and Tesla planned to use them for his power broadcast system since they could pass through all matter.

In 1890, Tesla discovered that placing a long single turn copper helix near his magnetic disrupter became covered with an envelope of white sparks. Effects were strongest when the helical coil was placed within the disrupter wire circle. In this "shockzone", the coil was surrounded by a blast of long, fluid like silvery streamers which clung to the surface of the coil, flowing over the coil at right angles to the windings. Tesla hypothesized the electrostatic-like effect was due to radiant transformation rules requiring measurements of discharge lengths and attributes of the helix used (number of turns, diameter, etc.). (Editor: resonance)

The new induction law he discovered showed that radiant shockwaves became much stronger when encountering segmented objects. The radiant shockwaves "flashed over" the outer skin of the helix from end to end without passing through the windings of the coil; shockwaves of 10,000 volts introduced to a 24 inch coil rise to 240,000 volts. The greater the resistance in the helix turns, the higher the maximum voltage. This was completely different from magneto-induction. Remember that he was not working with electricity but with radiant shockwaves.

> *Aether particles were incompressible and could easily pass through space and matter with a speed much greater than the speed of light. This was pure radiant matter but, at the same time, pure Radiant Energy. Cold electricity is one of the forms of this Radiant Energy*

This transformer he invented used radiant shockwaves to produce pure voltage without current. Each transformer had to be "tuned" by adjusting the disrupter to a specific impulse duration. At that point, voltage impulses could flow smoothly, flowing over the copper surface much like a stream of water in a pipe. No amperage could be detected but if the stream was aimed at a distant metal plates, "current" was produced which reached several hundred or even thousands of amperes. Tesla then began to wonder what made up this white, currentless stream.

Tesla determined that normal charge carriers (electrons) could not travel as quickly as the radiant pulse; no current moved the coil because the electrons were choked in the metal lattice of the coil. **The radiant pulse moving over the surface of the coil was not electronic in nature.** He placed the legs of a heavy U-shaped copper bus bar directly to the disrupter primary and then connected the short-circuited system to several incandescent lamps. These lamps glowed with a brilliant cold white light as with Gray's light bulb, proving that the power for the light was not electrical in nature. Tesla believed that the electrons were blocked from flowing through the wire while the radiant pulse was released over the coil surface as a "gaseous" pulse (splitting the positive, according to Gray).

Tesla came to believe that voltage could be viewed as streams of aether under various states of pressure and that his transformers affected the aether as to produce the luminous effects he observed. In his patents, he describes his "light-like rays" as tightly constricted aether streams propelled from his transformers along infinitesimal ray lines along which an incompressible movement occurs instantly through space on all points along its path. Aether streams were being drawn in through his transformer at higher natural pressure and then accelerated in the electrical discharge. The voltage in his transformer could control the brilliance of light in a room but this type of light was almost impossible to register on film. He could heat up a room or cause cool breezes by controlling the voltage in impulse duration in his transformer.

Tesla saw electrical current was really a complex combination of aether and electrons. Through the application of a disrupter, the electrons were removed from the gap by the magnetic field while the aether steams continued to flow through the circuit. He considered aether particles highly mobile with infinitesimally small mass and cross-section as compared to electrons. They were incompressible and could easily pass through space and matter with a speed much greater than the speed of light. This was pure radiant matter but, at the same time, pure Radiant Energy. Cold electricity is one of the forms of this Radiant Energy.

Edwin Gray, Dr. Nicola Tesla and Dr. Thomas Henry Moray all used radiant energy technologies. It is important to understand that the Laws of Thermodynamics and Maxwell's equations do not pertain to "Radiant Energy" technologies.

• Fantastic Projects •

Gravity Conversion Rotary Device

Review is prepared by editor Tatiana Ezhova

Here another fuelless perpetuum mobile is presented.

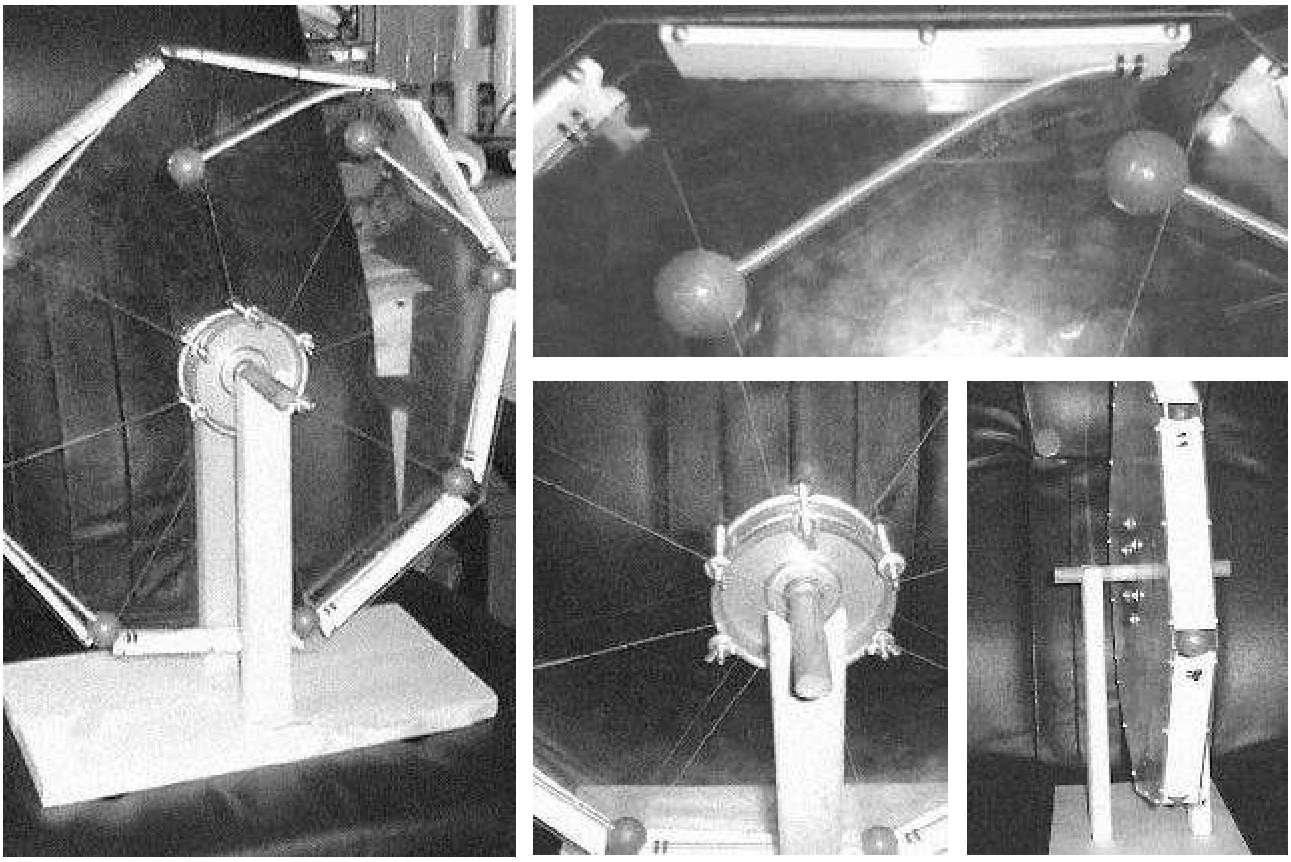

The system design

The design includes a wheel with spring spokes. Pegs may be fixed inside the wheel to balance movement of weights on springs. All parts should be homogenized as to dimension and weight. The author used 4 oz. lead balls and the 3/16" springs. The wheel measures 9" in radius as measured from hub center to center of each lead ball at outer periphery. This device is actuated and stopped by the hand. Its speed is self governing, it ramps up to speed quickly and simply stays there.

Note: you can find more detailed information about the device at http://www.greaterthings.com/News/FreeEnergy/Directory/GravityMotors/photos/. Also read in this issue article "Novozhilov's motor" that describes another wheel motor that does not require any fuel for its operation.

Torsion Technollogies

Xing-liu Jiang, Xiong-wei Wen, China
Science School,
Beijing University of Aeronautics and Astronautics, Beijing 100083
and
Li-jun Han, China
Department of Materials Science and Engineering,
Beijing University of Aeronautics and Astronautics, Beijing 100083

ABSTRACT

Highly localized nuclear activation in electrochemical systems and other electrical discharge processes have been observed by many laboratories in the world. This report is an attempt to explain such anomalous phenomena by using torsion field theory and axion model. Anisotopic behaviours of radiation products, burst character, "heat after death" of excess energy release in electrical discharge systems are considered to be interpreted by the torsion coherence of vortex dynamics with the zero-point energy induced by localized intense field emission of micro-protrusion of the cathode, and the dynamic Casimir effect of transient evolution of triple region of gas, liquid solution, and electrode protrusion. Axion model and Primakoff's effect are proposed for explanation of nuclear transmutation without noticeable gamma radiation.

Nuclear products with high concentration, unidentified tracks with highly collimated lines of low energy nuclear reactions in the electrochemical systems were recorded by CR-39 solid detectors and photo-films, and localized spots with chemical alterations were observed at our laboratory. It is suggested to carry out intensive study of vortex dynamic for explaining the anomalous phenomena in wide area of nature and laboratories. Analysis of vortex dynamics with wide range from pitting corrosion of electrochemical system, laboratory plasma, tornado, to quasar spiral model with extremely high energy cosmic rays in the center region, leads to a conclusion for that vortex dynamics creates torsion fields responding to the anomalous effects.

I. INTRODUCTION

Many laboratories in the world have observed nuclear reactions and **excess heat** in electrochemical systems. The mechanism of such anomalous phenomena is not being well understood according to normally accepted physics.

Nuclear products with high concentration and tracks with highly collimated lines of low energy nuclear reactions in the electrochemical systems were recorded by CR-39 solid detectors and films at our laboratory [1,2]. These facts suggest that quasar model with spiral structure and extremely high enrage cosmic rays in the center could be used to explain the mechanism. It is supposed to use the concept of torsion field to interpret the observed phenomena, typically, the properties of axion acceleration, memory effect, and the polarized nuclear reactions with torsion effect [3].

II. PHENOMENA IN NATURE AND AT LABORLATORIES

The vortex and spiral structures are the archetype that appears at all levels of nature and laboratories, for example, atom structure, vortex lattice in superconductors, dense plasma focus, lightning, quasar etc.

Recent report of dark matter annihilation at the galactic center describes that the cold dark matter near the galactic center is accreted by the central black hole into a dense spike [4]. Particle dark matter annihilation makes the spike to be a compact source of photons, electrons, positrons, protons, antiprotons and neutrinos. It reminds us for that there is a similarity among pitting corrosion with electrochemical noise, laboratory plasma pinching (dense plasma focus, for example), fast laser induced ion beams [5] and quasar spiral model with high-energy cosmic rays in spiral center in spite of large dimension difference. Comparing the experimental results of electrochemical cells with excess heat and nuclear transmutation to astrophysics phenomena, it is supposed that the investigation of vortex dynamics of torsion coherence with the zero-point energy is essential for tapping the zero-point energy.

As it is predicted by quantum mechanics that the vacuum is seething with active energy, even at temperature of zero point Kelvin. This zero-point energy (ZPE) can be thought as an infinite number of virtual photons that are popping out of the vacuum and going back in, but should be a measurable effect en masse. To exam the origin of ZPE background, the interaction of matter with the ZPE can be treated on the basis of charged point particles interaction with a background of electromagnetic zero-point radiation with spectral-energy density.

Based on observation in the experiments, the following features are considered in order to understand the experimental results with electrical discharge systems.

A. ELECTROCHEMICAL DOUBLE LAYER

In an electrolytic cell, the electrolysis with high conductivity and the electrochemical double layer with large layer-capacitance lead to a typical structure of the cathode potential distribution similar to the cathode drop of glow discharge in low gas pressure. For a compact layer the thickness of the double layer is equal to one ionic layer, across which there is a linear fall of potential. Thus, high electric field exists in some regions on the surface of the cathode. The local enhancement of the electric field on the cathode surface with the double layer is related to the protrusions and cracks similar to the tip discharge in air or in a vacuum. The current distribution depends strongly on the surface roughness and the work function of the electron emission. A high transient current density ($> 10^8 A/cm^2$) could be expected due to enhanced field.

B. ENERGY CONCENTRATION

On the cathode surface, the high persistent electrical fields ($>10^7 V/cm$) and large equivalent capacitance ($>25 \mu F/cm^2$) lead to a high energy concentration in the double layer [6]. The concentrated field on the tips of the protrusions or cracks after a long-loading period with deuterium on the palladium cathode surface creates a high transient electron flux because of the large distributed capacitance and the negligible inductance in a localized discharge mini-network. The experimental data show that the reactions take place only in some restricted areas that have specific properties. The idea of micro fusion due to the results of energy concentration and the high deuteron flux could be used to explain the nuclear transmutation.

C.TORSION FIELD AND THEIR EXPERIMENTAL MANIFESTATIONS

Elementary particles have the moment of quantity of motion, i.e. spin. If in any substance the spins of particle have a preferable direction, then it is interpreted as spin polarization of the substance. Every substance creates a torsion-field (or called spin-field or axial field) in the space surrounding it when polarized by spins [7]. The superposition of torsion field, generated by the atomic and nuclear spins of each molecule, determines the intensity of torsion field in the space surrounding each molecule. Torsion field has strong penetration ability and does not interact with the crystal lattice of substances. The torsion field created by rotation of some sort of matter is concentrated in two opposite beams propagating along the rotation axis. The intensity of torsion-field with some lower constant value can be retained for several weeks after the rotation is stopped. Such property of vortex mater has been recently observed in type II superconductors with magnetic flux line lattice [8]. The studies of vortex matter of type II superconductors have shown a number of puzzling phenomena associated with vortex motion, including: low-frequency noise and slow voltage oscillations; a history-dependent dynamic response, and memory of the direction, amplitude duration and frequency of the previously applied current.

Some behaviors of the torsion-field effect have been observed in the experiments of electrolysis regarding the so-called cold fusion, such as when the gas bubble chains come out from the protrusions of the cathode surface a long time after switching off the electrolysis potential. The more surprising thing is the heat after death, which has been recorded by many labs. This phenomenon could be explained by the persistence effect of the torsion-field produced by vortex dynamics of tip effect.

D. EXPERIMENTAL RESULTS IN THE ELECTROCHEMICAL CELL

CR-39 plastic films possessing with a high degree of optical clarity and isotope in track response, and are sensitive to neutron, proton, tritium, alpha, and other charged particles, were used to detect the products of nuclear reactions. The films of CR-39 were immersed in the NaOH electrolyte of heavy water and placed adjacent to the tips of the cathode [9]. After 110 hours, electrolysis experiment with the applied voltage 1 V and current 2 ma, the solid detector was etched by 6.25 N NaOH solution in 70^{0}C for 11 hours. The photomicrograph shows the cluster of tracks with a cycle crater of 100 μm in diameter and 25 μm in depth. According to the etching condition, the energies of the most of the particles, P, T, α are estimated roughly to be in the range of 1-4 MeV. On the backside of same region of the CR-39 film detector, one can clearly see a few of tracks in the circle area or nearby. Such tracks are believed to be created by recoil-protons of forward incident neutrons. High concentration of the cluster of nuclear tracks could be explained by the quasar spiral model and by the crystal channeling effect [Fig.1]. The experimental results of the generation of tritium with single crystal metal of Pd show that none of the generation of tritium has been revealed by using non-single crystal electrodes [10]. The importance of the crystal channeling effect for nuclear reactions in the electrochemical systems can be inferred.

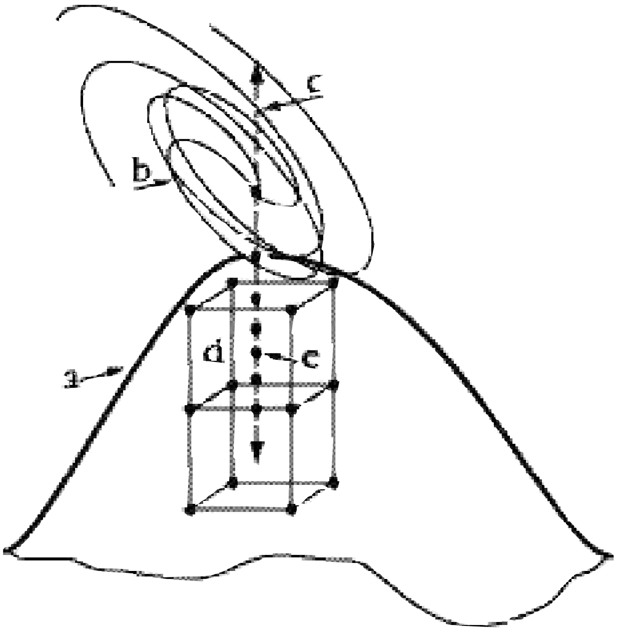

Fig.1
Schematic presentation of a micropinch spiral by tip effect in the electrolysis cell:
a) tip of the electrode; b) spiral structure of micropinch; c) electron beam; d) crystal channel; e) ion beam

To determine the spatial distribution of radiation active sites (RAS), Black-white 135 films of 27 DIN have been used to image the position of the RAS. After a 1,5 year of deposition in glass tube of finishing electrolysis experiments with light water electrolyte for more than 200 hours running, the patterns of RSA were clearly formed on the films after exposure of 100 hours [2]. The bright spots corresponding to the tips of palladium cathode edges can be seen due to the tip effect. The effect of magnetic field on the traces of the charged particles has been observed while the Pd samples exposing to the films, which were folding and wrapping up the sample, were inserted between couple magnets. The tracks stretching along the film surfaces confirm that the tracks were created by charged particles, electrons for example, with low energy about some keVs. Highly oriented tracks can be observed by autoradiography by using normal films locally[Fig.2].

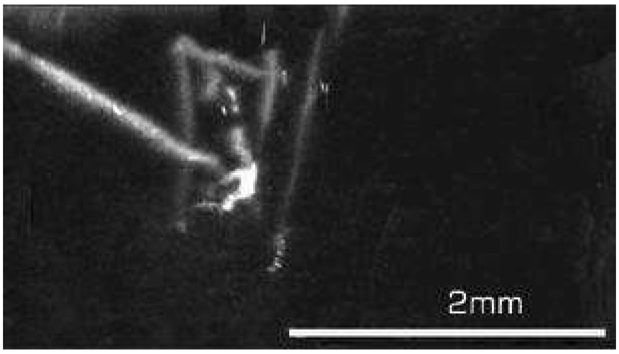

Fig.2
Autoradiography of charged particle tracks of beta delay isotopes on the surface palladium cathode. Some tracks of beta particles are paralleling the cathode surface.

E. SONOLUMINESCENCE AND “BUBBLE NUCLEAR FUSION”

Some scientists of Oak Ridge National Laboratory in America reported their articles of bubbles experiments in Science. Experimental results show that the radiation lights of sonoluminescence possess three characters: short duration with picoseconds; wide continual spectrum; highly oriented thin beams. **Based on those characters, vortex dynamics with axial acceleration of bubble collapse could be inferred.** Nuclear reaction with abnormal gamma radiation was observed [11]. Dr.Claudia Eberlein describes her conclusion on sonoluminescence that only the ZPE spectrum matches the light emission spectrum, which must be a ZPE phenomena [12]. The effect of torsion field on nuclear reactions along the axis of vortex should be taken into consideration for the low ratio of the nuclear products of n/T due to spin polarization of reaction particles.

Axion model and the Primakoff effect are proposed for explanation of nuclear transmutation without noticeable gamma radiation. Many laboratories in the world have been engaged in several experiments to search for axions, light neutral pseudoscalar particles yet to be discovered. The axion would be produced in the solar core through the Primakoff effect if its mass is a few electronvolts and could be detected in the laboratory.

III. CONCLUSION

Researchers of new energy study of "cold fusion" type should pay great attention to the general processes of electrolysis to find the key points, which could play a major role in the transit from the electrochemical processes to processes of torsion and the nuclear processes. From the authors' point of view, the evolution of double layers is of importance to understand the anomalous effect, typically on the protrusions of cathode. The change of space-time near the tips due to torsion field generated is expected to delivery the zero-point energy, and dynamic Casimir effect for the evolution of gas bubbles on the tips are expected to generate photons and excess heat [1]. The yields of transmutation products are related to the current distribution on the cathode surface. The cathodes of thin wire were benefit to the generation of nuclear reactions and excess heat for electrochemical systems. Careful examination of the evolution of electrochemical double layer will lead to a good understanding of pitting corrosion with electrochemical noise, and further to recognize the anomalous excess heat and nuclear reactions.

Heat after death was observed at many laboratories. It is believed that the persistence behaviors of torsion field can be used to explain such anomalous phenomena.

The contact between beads coated with a thin metallic layer or multilayered film and palladium black particles could be regarded as point-contact similar to the tip-effect [13].

We may say that it is expected to reveal the puzzles of the mechanisms of high-Tc superconductivity with pinning vortices by the torsion field theory [14,15].

REFERENCES

[1] Xing-liu Jiang, Jin-zhi Lei, Li-jun Han, Dynamic Casimir Effect in an Electrochemical System, J. New Energy, Vol.3, No. 4, 47(1999).
[2] Xing-liu Jiang, Chang-ye Chen, Li-jun Han, Tip Effect and Nuclear Active Sites, Proc. of the 7th Intern. Conf. on Cold Fusion, Vancouver, April, 1998, pp.175.
[3] Don Read, Excitation and Extraction of Vacuum Energy Via EM-Torsion Field Coupling Theoretical Model, J. New Energy, Vol. 3, No. 2/3, p.130(1998).
[4] P.Gondolo, J.Silk, Dark matter annihilation at galactic center, Phys. Rev. Lett. 83, 1719(1999).
[5] BPS, Laser light in , stream of protons out, Physics Today, Jan.2000,p.9.
[6] G.Korluem and J.OM. Bookris, Textbook of Electrochemistry Vol. II. Elsevier Publishing company Amstorsan , 1951. P.364
[7] A.E.Akimov, G.I.shipov, Torsion fields and their Experimental Manifestations, Journal of New Energy, 2(2), 67(1999).
[8] Y.Paltiel, E.Zeldov Y.N.Myasoedov el. al. Dynamic instabilities and memory effects in vortex matter, Nature 403, 398 (2000).
[9] X.L.Jiang, L.J.Han and W.Kang. Concentrated energy and Micro Nuclear Fusion, ICCF6, Oct 1996, Japan p.580.
[10] R.L.Matlock, F.E.Collins, G.R.Bancher, Anomalous tritium found in the recombined off gasses during electrolysis using crystal cathodes. Elem. Energy(cold fusion), 26, 28(1998).
[11] Taleyarkhan R P, West C D, Cho J S, Lahey Jr R T, Nigmatulin R I, Block R C. Evidence for Nuclear Emissions During Acoustic Cavitation. Science, 2002 295:1850~1862.
[12] Claudia Eberlein. Theory of quantum radiation observed as onoluminescence, Phy.Rev.Lett.53,2772(1996)
[13] G.H. Miley, G.Narne, M.J.Williams, J.A.Patterson, J.Nix, D.Cravens, and H. Hora, Quantitative Observation of Transmutation Products Occurring in Thin-Film Coated Microspheres During Electrolysis, Proceedings of ICCF-6, OCT. 1996, Japan, p. 629.
[14] T. Matsuda, K. Harada, H. Kasai, O. Kamimura and A. Tomomura, Observation of Dynamic Interaction of Vortices with Pinning by Lorentz Microscopy, Science, 271, 1393(1996).
[15] G.W. Crabtree and D.R. Nelson, Vortex Physics in High Temperature Superconductors, Physics Today, April 1997, p.32.

New Energy Technologies
PRINTED EDITION ADVERTIZING RATES

Full Page Ad
B/W - $400
Colour - $800

Half Page Ad
B/W - $200
Colour - $400

Quarter Page Ad
B/W - $100
Colour- $200

New Energy Technologies is distributed in the leading Scientific Laboratories in Russia and has subscribers in more than 20 countries all over the world

To advertise contact us:
Farady Labs Ltd, Phone/fax: 7-812-380-3844
http://www.faraday.ru, office@faraday.ru

Fantastic Projects

The Invention of Roman Solomyanny

Our previous issue featured an article about Roman Solomyanny who came to our office and told us amazing things. He said that during his army years (according to his words, he was in the military intelligence) he made a radio technology invention and later on built a free energy generator. The inventor claimed that his device had been working for a certain period of time without any fuel and had been producing power enough for heating and lighting his countryside house. We could not ignore such an intriguing piece of information. Moreover, Roman promised to build and present his device in 2003. We paid him for this interesting story to be published in the issue and prepared for the continuation... Alas, to no avail... His mobile phone was silent and we started to worry about the inventor's health. Fortunately, we managed to contact his parents on his home telephone (Roman lives in St. Petersburg) and they told us that the chap was OK. It seems he's lost his interest in us. But we still hope that Roman did invent something really exciting and not only polished up his talent for writing science fiction articles. Those people, who want to know the details about this invention or discuss it with the inventor himself, can get in touch with us for his contact information. Anyway, we will be very pleased to see him again full of new exciting stories. We have already introduced the 'Humour' section into the magazine and are planning to start the 'It Just Can't Be So' section.

Editor

Physicists are Joking

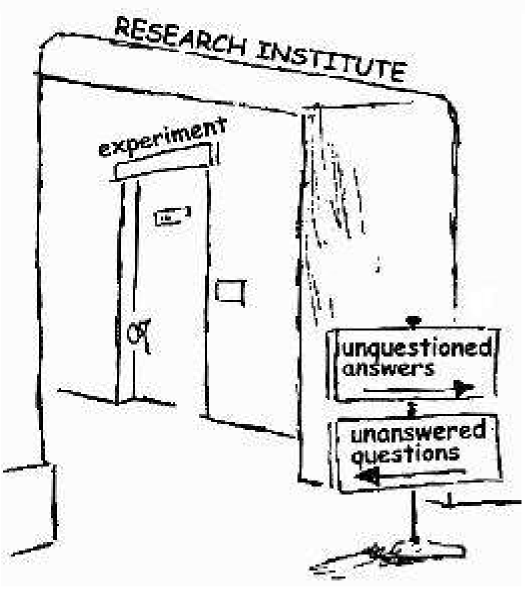

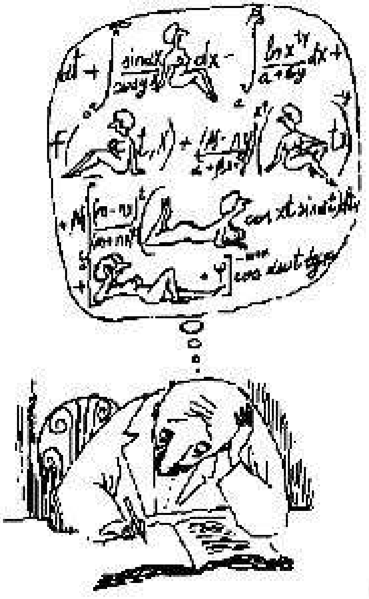

Anti-Gravity: The Holy Grail of the 21st Century

Bruce A. Smith, USA
PO Box 1676, Yelm, Washington 98597
Email: brucesmith@rainierconnect.com

A Primer on the Role of Electromagnetic, Electrostatic, and Torsion Fields in Anti-Gravity and Field-Effect Propulsion

Before visiting a University of Washington physics professor to talk about electrostatic propulsion, and hopefully anti-gravity, I realized: *Birds defy gravity. So do 747's for that matter. They apply the laws of physics and lift off the ground. That's anti-gravity, isn't it?* Yes, that's true, I suppose, in a metaphorical sense. But seagulls, jumbo jets and space crafts manifest anti-gravitation effects strictly speaking. What I want to address here is not the overcoming of gravity but the neutralizing of it.

Dr. Eugene Podkletnov, one of the foremost researchers in anti-gravity, and whose work is sought by NASA, Boeing and British Aerospace (now known as BAE Systems) describes the hunt for anti-gravity the greatest scientific quest of this century. He calls for an international effort, akin to the Manhattan Project that developed the atomic bomb, to conquer the secrets of anti-gravity, and usher in a new era of scientific understanding whose technological development will be at a scale so vast that the potential outcomes are merely hinted at by our previous achievements.

Just getting such a project off the ground will require unprecedented international cooperation, and public disclosure as well; the potentials are that vast, that scary, and that dangerous. Dr. Dan Marckus, noted British avionics expert, states in The Hunt for Zero Point, the seminal work-to-date on anti-gravity written by Jane's Defence Weekly aviation editor Nick Cook, that the secrets of anti-gravity in the wrong hands will make thermonuclear weapons look like firecrackers.

The secrecy surrounding anti-gravity research is phenomenal. Boeing refuses to publicly acknowledge any activity in anti-gravity development despite the fact that its competitor and sometime sub-contractor, British Aerospace (BAE Systems) is, the latter providing funds for four university research efforts as part of its Project Greenglow, one of which was a Podkletnov replication experiment headed by Dr. Clive Woods at the University of Sheffield,. Further, Nick Cook publicly, and privately to me in an email, states quite directly that George Muellner, former director of Boeing's ultra secret Phantom Works, claims Boeing sought the services of Dr. Podkletnov to unlock the secretes of his gravity-shielding research. Cook says that Muellner states Boeing was denied Podkletnov's services due to the objections of Russian officialdom, which the Russian-born Podkletnov must pay attention to, apparently, despite the fact that he works in Tempere, Finland. Dr. Podkletnov, wisely perhaps, chooses not to clarify these particulars despite our several emails.

Perhaps Boeing can deny any activity on anti-gravity because NASA is doing its own research, and as a prime contractor to NASA, such as by running the Space Shuttle Program, Boeing probably knows what NASA knows. NASA spent $600,000 recently in its Breakthrough Propulsion Physics program (BPP) to purchase Podkletnov replication equipment. Inexplicably, that equipment sits in boxes in NASA's Marshall Research Center in Huntsville, AL, awaiting more funding; according to an email I received from NASA propulsion researcher, Ron Koczor.

But enough with this business; what do we know about anti-gravity?

The search for that answer has taken me to some exciting and obscure places in this world, like the Aeronautics and Astrophysics lab at the Seattle campus of the University of Washington. I called those folks because Nick Cook in The Hunt for Zero Point, mentions that UW received a NASA contract to study theories of inertia as part of its BPP program. *That's a good place to start,* I thought, but it took backtracking to BPP Project Director Marc Millis at NASA's Glenn Research Center in Cleveland to find Dr. John Cramer at the UW Physics Department. His mission was to confirm with Dr. James Woodward the latter's 1996 preliminary research into the loss of gravitational mass in a targeted piece of metal from oscillating capacitors. Although Woodward's initial data appeared encouraging, NASA's Millis told me that their funding dried up before completing their research. Furthermore, the entire BPP became unfunded in 2002 and now in 2003 has become a hazy, privatized version of its former NASA subset self.

However, UW is continuing related research, such as magnetically confined fusion energy generators and that, for me, by using electromagnetism to contain an inner field makes it a close cousin of anti-gravity and field-effect propulsion. I spoke with Professor Uri Shumlak who told me that he and other UW staff from the Department of Aeronautics and Astronautics, along with a bevy of their grad students, are building a prototype of a fusion generator called HIT, which stands for Helicity Injected Torus.

This donut ring-shaped torus encloses a roundish chamber. Within that chamber a vacuum is first created, and then a volume of hydrogen gas is introduced and heated to a few million degrees Celsius, which separates the electrons and protons from their atoms turning the whole stew into a quasi-neutral foam of plasma. Then the torus envelopes the plasma with a magnetic field to keep it away from the sides of the chamber enabling the plasma mass to stay hot, and keep the rest of Seattle cool. (While I was standing next to his little eight foot long gizmo, Prof. Shumlak assured me there was no danger of a couple million degrees of heat escaping. The heat density of the plasma was "too low" for me to, well, break a sweat over. His quote was, "There's no more heat mass inside that chamber than what's contained in a cup of coffee." I sure hope you're right, Doc.)

Then, once the plasma field is contained, the magnetic field squeezes the plasma, fusing the nuclei of one hydrogen atom into another. As the hydrogen couples combine, a helium atom is created and a neutron is released, along with lots of energy in the form of heat. One day, such a generator will give us unlimited amounts of electricity, as the heat can produce electric voltage.

Lots of electrical power on the cheap the UW predicts; and the Department of Energy agrees, once the details of building reliable magnetic field generators are solved. What does magnetic fusion have to do with anti-gravity? Two things: first, magnets. Electromagnetism seems to be one of the major players in anti-gravity, particularly the use of electromagnetic fields to contain other fields, such as plasma fields in the HIT, or torsion fields, but more about that later. Secondly, the HIT works, or is about to. It's real and mainstream science embraces it, while anti-gravity is, well, a little more out there and reliable data is harder to obtain. So the technology of HIT lays a base that other research can build upon, such as not only containing other fields, but also building field effect propulsion systems, the most elementary of which is electrostatic propulsion, and aspects of that are already being applied by NASA.

Electrostatic propulsion uses electrical fields differently than electromagnetism does. In EM a current flows and creates a field, while in electrostatic systems the current is static and a charge builds up a field, such as in a capacitor.

These theories are utilized on NASA's Deep Space I, a probe bound for the outer reaches of our solar system. On Deep Space I, the propellant, a tankful of xenon gas, is excited electrostatically into positive ions. The engine has a negative charge at the exit end, so the charged xenon rushes out the tail pipe with a greater thrust than if it was just using conventional chemical propellants. In fact, the electrostatic propulsion system on Deep Space I allows it to fly at 60,000 mph, or 10,000 mph faster than it would with a conventional rocket. In addition, only 82 kg of xenon is needed for its entire mission, so with its smaller mass and weight Deep Space I will fly along side its intended target, a comet, and drag race on equal footing while filming and conducting studies. Again, not anti-gravity per se, but electrically charging Deep Space I's fuel-field sets the stage for a closer look at electrostatic propulsion.

Taking that closer look is Tim Ventura and his fellow researchers at American Antigravity, an organization based in Kirkland, WA. Ventura and his crew use electrostatic asymmetrical capacitors to create a field that levitates objects, such as their small, kite-like "lifters." These lifters are very light, weighing only a few ounces, and have balsa wood struts that support the capacitors. When two capacitors of different size receive their share of a 30,000 volt charge, the lifter lifts- no motors or wings.

How, no one really knows in my judgment; and the phenomenon is replete with controversy and mystery. But as one who has seen a lifter fly, let me tell you what one looks like and what I saw when Tim Ventura's took off.

Tim has been building lifters since he was a kid and has perfected a four-foot, by-four-foot, by four foot triangular lifter which has flown so many missions in his garage that the silver aluminum foil has turned white. The thin, chopstick-like balsa wood ribs that hold the aluminum foil in place are joined every few inches by a vertical strut (much like a telephone pole on a HO model railroad set) which sticks up and secures the copper or stainless steel wire of the upper capacitor. The ribs are intersected every ten inches or so by the strut of an interior triangle, since the whole lifter is composed of interconnected isosceles triangles which give the necessary strength to the balsa wood frame. All told there is about 30 linear feet of aluminum foil and a similar run of wire.

The lower and larger capacitor is a strip of aluminum foil stretched between the horizontal balsa wood struts. The second capacitor is a thin strip of 50 gauge wire mounted about one inch above the aluminum foil. As capacitors they store electrical charge but don't pass it on in a current.

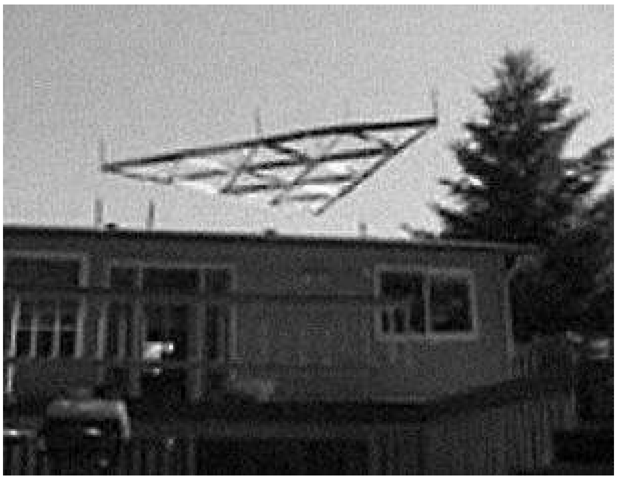

Fig.1
Lifter

The negative lead goes to the lower aluminum foil and the positive lead is attached to the upper wire. The three corners of the lifter are tethered to the work table so that the electrical leads from the power source are not broken off in flight.

The power source kicks out 15,000 volts at 250 watts. Tim uses a voltage generator made by Information Unlimited, Inc, but before the current reaches the capacitors, the voltage is stepped up to 30,000 volts by Tim's homemade voltage multiplier stack. At full throttle the lifter is straining at the tethers, bending the balsa wood frame near the point of fracture.

Throw the switch and at around 17,500 volts the lifter begins to quiver in take-off. At the full power of 30,000 volts the lifter is roaring and a noticeable downward breeze is observed. Many physicists call it "ion wind," and say that it is how the lifters fly. But what exactly is ion wind, and can it be the cause of flight?

"Ion wind is the movement of ionized air particles," according to Ventura, "which flow downward according to electrical charge." Here's his theory:

The positively charged wire on the top part of the lifter steals elections from the surrounding air, leaving the effected air molecules positively charged. These positively charged air molecules, or ions, then head downward toward the large source of negatively charged electrons generated by the aluminum foil. These air molecule ions are bigger and heavier than the electrons seeking them, so there is a net thrust downward pushing the whole lifter up. That's the theory, and frankly all I can do to verify the theory is to tell you what others tell me. Before I do that, though, let me tell you what I experienced standing next to a levitating lifter.

In flight the lifter emits a high whining, hissing buzz, and I could feel a good breeze coming up at me from the work table underneath the lifter. Also, standing next to the lifter but not touching it, the hair on the back of my head started to rise up in electrostatic-like fashion.

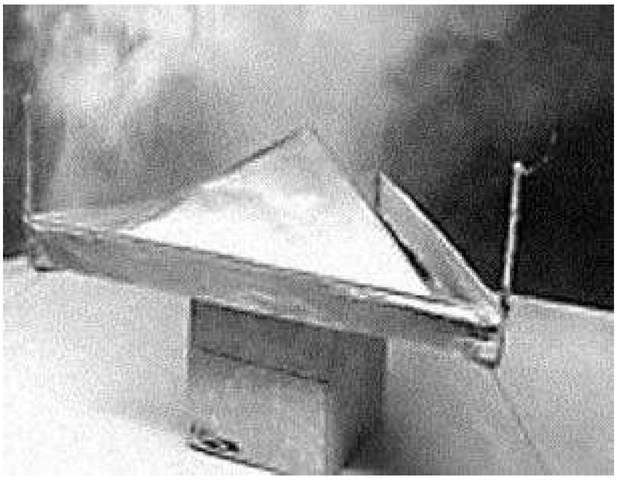

Fig.2
Lifter

To analyze the air currents Tim blew baby powder at the top of the lifter. The majority of the particulate cloud was drawn into the middle area of the lifter and then sucked downward. A kind of vortex was created at times, for intermittently I could see a cloud forming into an organized column beneath the lifter and then spreading out in 360 degrees once it hit the work table surface.

Electrostatic propulsion uses electrical fields differently than electromagnetism does. In EM a current flows and creates a field, while in electrostatic systems the current is static and a charge builds up a field, such as in a capacitor

Is that ion wind? Well, there certainly was a breeze, and it sure felt like air, but how would I know if it was ionized? Something definitely sucked the baby powder down, but was it more than just regular air blowing past me? Again, I don't know.

Is the movement of wind why lifters fly, regardless of whether it is ionic or not, or is the wind just a by-product and not the propulsion? Could the capacitors be creating a field that neutralizes gravity, allowing the craft to levitate? Or are they creating some kind of new field that is localized and the surrounding ambient field pushes this "field-bubble" up, much like a helium balloon is pushed up by the surrounding heavier air trying to fill the emptier "field" of the lighter helium?

Ventura thinks at least two phenomena are at work. Ion wind is definitely one he feels, for the breeze is self-evident. However, he thinks a second effect is at work, too, and many agree with him. Most speculation concerns what is called the Biefeld-Brown effect, the "Brown" being T. Townsend Brown, whose name is well known in early quantum research and whose work is prominently discussed in Nick Cook's The Hunt for Zero Point.

The Biefeld-Brown effect, according to Ventura, is the theory that high voltage, air-gapped capacitors with different or asymmetrical capacities generate a net directional force upwards from the larger element to the smaller element, which on the lifter is from the aluminum foil to the wire. This force then pushes against the ambient energy field of the surrounding area, perhaps pushing against a more rigid energy field of the zero point energy field.

Brown apparently made his case for these electromagnetic effects, receiving patents in the 1960's for his research. NASA's Dr. Jonathan Campbell at Marshall Research Center in Huntsville confirmed to me that he also, has received a patent recently for his research into the thrust effects of asymmetrical capacitors.

Editor: By the way, this patent is meaningless from the point of view of a prime claim. Before that there had been many publications on the subject.

A.V. Frolov.

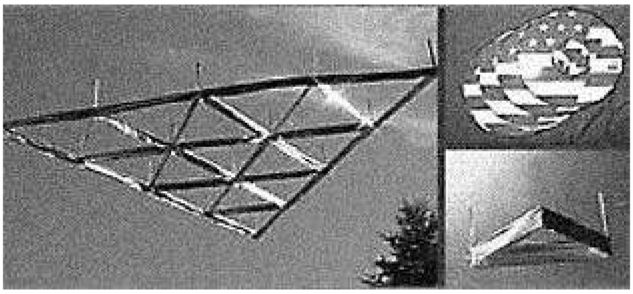

Fig.3
Lifter

However, prominent physicist Hal Puthoff, whose research is a broad swath across the fields of the 'new physics', featured in both The Field and The Hunt for Zero Point, and who was also the military's "Top Psychic" as the twelve-year director of the CIA's remote viewing squadron, has a different perspective: "I'm quite certain at this point that the so-called 'lifter' phenomena is just an electrostatic ion wind phenomena, not 'antigravity.'"

But Dr. John J. Rusek, Adjunct Professor of Aeronautics and Astronautics at both Purdue University and the United States Air Force Academy, says that "Initial findings of 'classroom' experiments with lifters show ionic wind to be way too small a factor, by three orders of magnitude." Dr. Rusek has formed a technological company, Swift Enterprises, to continue this research and bring it to the level that is "presentable to the mainstream physics community."

Along these lines, Jean-Louis Naudin shows on his extensive web site, not only how to build a lifter, but also several photos of research into the ion wind issue. Naudin's team has wrapped test lifters in plastic yet they still produce anti-gravitational effects.

Others may have a clue to the second or even a third force at work. Researcher Fran De Aquino, professor of physics at Maranhao State University in Sao Luis, Brazil, is described in the literature as showing that "bubbles of localized space-time" can exist in variance to the surrounding fields. Anecdotal experience suggests that the lifters may be undergoing such space-time anomalies.

Editor: At this point we need to interrupt the author as he has already completely confused the reader. I would recommend referring to Thomas T. Brown's patent: USA patent #3,187,206, 1965.

Alexander V. Frolov

The Hunt for Zero Point states that NASA sought the services of Dr. Eugene Podkletnov, and although their replication research languishes, The Hunt claims that researcher Ning Li, of Huntsville, AL is pursuing this line of research as a private contractor to NASA.

Another Huntsville operation, Transdimensional Technologies, is exploring these multi-faceted phenomena as well, and its extensive web site shows it to be a frequent contractor to NASA, including research into "asymmetrical capacitive propulsion," and capacitor-based devices to test "ion wind" forces. Jeff Cameron, of Transdimensional, is said by Ventura to be "the father of the lifter," having developed them while exploring anomalous torsional effects of high energy lasers. The lasers twisted and at the time it was considered a nuisance. But the unknown forces at work later led Cameron to found Transdimensional, develop lifter technology to a commercial level, and subsequently patent many pieces of related technology. Unfortunately, I have been unable to reach Jeff

Cameron or anyone at Transdimensional for any kind of confirmation.

Nevertheless, how would gravitons be blocked, or gravity shielded?

Dr. Hal Puthoff says there are two ways of looking at it. First, one can look at the issue from a quantum point of view, that there is a particle exchange between the gravitons and something else, and the net effect is anti-gravity. The how's and why's of that are speculative, so Puthoff turns to a classical approach for answers. He prefers looking at "engineering the vacuum." To do that one must first look at the vacuum.

As I understand it we are all in the vacuum, everything is. The "vacuum" is the matrix that contains all matter and all energy. It is the engineering perspective of the zero point energy field, or the *"Field"* as popularized by Lynn McTaggart in her masterpiece The Field. Puthoff shared with me statements from fellow researcher Dr. T. D. Lee that state: "The vacuum is the seat of energetic particle and field fluctuations, and ... is the seat of space-time structure ... that encodes the distribution of matter and energy.... The vacuum is energetic in its own right."

Thus energy can be drawn from the field; and spacecraft can have "vacuum propulsion systems, or propellant-less propulsion," in other words, field effect propulsion.

At any rate, more and more physicists are thinking that the vacuum can give them a whole lotta oomph, enough to propel spacecraft; and when they learn how to corral it, a whole bevy of new phenomena may be encountered, including anti-gravity. This new potpourri of research is being called by many the "new physics." And although his approach is classical, Dr. Hal Puthoff seems to be sensing what's out there waiting to be discovered.

Puthoff's current research as been to explore "the perturbation of atomic or molecular ground states, hypothesized to be equilibrium states involving dynamic radiation/absorption exchange with the vacuum fluctuations. In this model atoms or molecules ... are expected to undergo energy shifts that would alter the spectroscopic signatures of excitations involving the ground state." Puthoff says he's had no success so far with this approach, but his words remind me of De Aquino's speculation that objects lose mass as they absorb energy. Pull energy from the field around you and you lose weight. Bingo, lift-off. **But how does one pull energy from the field?**

Editor: Some solve this problem by means of the ether density changes, which are performed with the help of vortex longitudinal-wave technologies.

A.V. Frolov

Torsion fields might play a role here according to many, and the literature on anti-gravity is filled with the phrase, "torsional effects." But what exactly is a torsion field? "It has something to do with spin," Nick Cook told me on the phone. "You have a torsion field when you spin something. Add a little electromagnetism and you might have anti-gravity." That's the short-hand version of it and here's a deeper look.

Mike Wright, resident physics expert at BeyondTheOrdinary.Net web stream radio, told me this: "When forces create curvature (such as rotation) in more than two planes, a torsion field results. Not only does the object go around, but it goes around and 'down' or 'up', and the up/down movement is an additional acceleration in that dimension. EM and gravitational fields differ by having a magnitude of force and only one direction of movement.

"A tornado is a structure of air in air. A whirlpool is a structure of water in water. So, because more than two planes are involved, objects can be created from 'nothing'; that is to say that objects can be created from the medium of the environment, such as a tornado from two air masses of differing temperature."

So spin plus movement is the key. Again, Tim Ventura is on the hunt. He demonstrated to me **that spinning magnets will cancel out their magnetic fields sufficiently so that two magnets facing each other with like poles will not push each other away if one of the magnets is rotating**

perpendicular to the force of opposition. It's not anti-gravity, but it gets us closer to the heart of the matter.

Further, Russian physicists have been researching the torsional effects of both subatomic particle spin, and the loss of gravitational mass in planets from the angular momentum of their orbits. Spinning makes something happen, but what? Tornadoes and Mother Nature might have a few clues.

Tornadoes spin, in a sense, although no one is Oklahoma who has spent a night in a storm shelter during an F5 event would describe the tornadoes in the night sky as *spinning*. Nevertheless, tornadoes have anomalous effects that are legendary: blades of grass stuck into mirrors, a piece of straw embedded flawlessly into a tree trunk. How? It seems as if the laws of mass, gravity and inertia are melted as winds swirl at speeds up to 300 mph in an organized vortex pattern. Is this a clue to melting the pull of gravity?

Getting information to this question has not been easy. Many scientists claim not to have even heard of torsion fields, including particle physicists at major US universities. So, again I turn to Nick Cook and The Hunt for Zero Point

Dr. Dan Marckus says that if "you generate a torsion field of sufficient magnitude the theory says you can bend the four dimensions of space around the generator. The more torsion you generate the more space you perturb. When you bend space you also bend time."

Marckus continues, "If you dipped ...one of these whirlpools ... into the zero point energy field, the seething mass of latent energy that existed on an almost undetectable level all around us [in the field would] ... react in an almost magical way by directing that energy."

The torsion field, in effect, is "a pump, a 'coupling' device that could dip into, and then direct, energy out of the zero point energy field." "But," Marckus continues, "the vortex wasn't a three-dimensional phenomena or even a four-dimensional one. It couldn't be. For a torsion field to be able to interact with gravity and electromagnetism it had to be endowed with attributes that went beyond the three dimensions of left, right, up and down, and the fourth-dimensional time field they inhabited; something that the theorists for convenience sake labeled a fifth dimension-hyperspace."

Cook concluded that the torsion fields "bind with gravity ... to produce a levitation effect -- an antigravity effect," but "it wasn't doing so in the four dimensions of this world, but somewhere else." That somewhere else is hyperspace. **So how do we activate torsion fields and enter hyperspace?**

Dr. Eugene Podkletnov may have a clue.

Podkletnov, the Russian researcher working in Finland, has studied the gravity shielding effects of superconductors. Again, Nick Cook in The Hunt, relays vital information. Cook says Podkletnov claims, "If the superconductors are rotated considerably faster than 5,000 rpm ... the disc experiences so much weight loss that it actually takes off." Thus, torsion field creates levitation. I emailed Dr. Podkletnov to find out more about this issue. He replied: "[A] fast rotating object can, under certain conditions, cause the polarization of the volume that it occupies in space and around it. This polarization causes the gravitational effect as it modifies [the] local gravity field. The vortex of the polarized particles will create a vertical thrust with a certain force and spatial momentum. Some scientists call these polarized particles gravitons. The term graviton is an artificial one and at present we are not sure if it is a wave or a particle and what type of particle. Maybe it is a usual tachyon or a superluminal neutrino (a faster-than-light-particle). Polarization of the media means that the spins of electrons, protons, neutrons and of small subatomic particles that constitute the fabric of space or vacuum would be parallel. Then a kind of gravity well is formed and the objects tend to fall into this well. We observe this picture as an object rising to the sky. Polarization of the media (of space) causes some glow around the object as it acquires additional energy and because of it, the glow around some objects is observed."

What I understand from Dr. Podkletnov is that gravity is the effect of spin -- the spin of all subjected particles, from the sub-atomic level and up being parallel; thus they are all aligned to fall into the gravity well of earth. And spinning objects, such as his superconducting discs, when influenced additionally by an electromagnetic field, will experience a shift in the spin of the sub-atomic and atomic elements. They will be turned and not be aligned in parallel. Thus, they are able to levitate.

But how to polarize the media and get things spinning? Enter Dr. Marcus Hollingshed, an enigmatic figure allegedly from Cambridge University. Dr. Hollingshed claims to have built a six-ringed toroidal coil antigravity device, which achieved great effect using rotating magnetic fields. In January 2003 he announced on the Internet that he has developed a 160kg vehicle able to lift in excess of 2000kg and that it has both horizontal and vertical drive features. His device cannot only go up and down and sideways, but it can push things away and pull objects to it.

In addition, the field that the device purportedly generates is capable of being broadened and weakened, or narrowed and amplified in a lensing effect, with the field producing an absolute vacuum of 2.2m spherical diameter. Best of all, when it's cranked up the core of it goes invisible, although the term Dr. Hollingshed uses is that there is a "loss of reflected light." There are no reports of independent confirmation, and Nick Cook says he hasn't been invited to see it, so, he's skeptical.

Where does this leave us? Perhaps Dr. Podkletnov's words sum up our current situation. "Modern theoretical physics cannot give you the direct answer to your questions (levitation, torsion fields, etc.,) and a scientist who would agree to give you the answer cannot be regarded seriously, softly speaking. If you had asked Dr. Einstein if he were an expert on gravity, the answer would be NO. I can repeat his words: No, I am not a magician, yet; I am still learning."

ELECTRICAL LEVITATION

Wayne Macleod, USA
Email: dynamars2100@yahoo.ca, cwleod@shaw.ca

Gravity is the incremental slowing of time as we move closer to its source. If we imagine panes of glass stacked on top each other, one pane slightly denser than the one immediately above, a ray of light coming from above will constantly refract in a curved beam downward as it enters the stack. The same is true of a ray of light entering a gravity field. The analogy is not too far fetched because refraction is caused by a slowing of light in the denser medium. Of course, gravity affects material objects as well as light, but this is explained by every object traveling on a 'world line' in spacetime. We can use light as in the glass example because light is its own world line.

Why the world lines of objects take the paths they do can best be understood by compressing our normal 3 dimensional space into 2 dimensions, length and height only. When we throw a stone into the air it rises and falls in a parabolic arc in these two space dimensions. That path in space is a complete mystery until we consider another dimension, time, a dimension we can imagine measured at a right angle to the plane of the 2 space

energy and because of it, the glow around some objects is observed."

What I understand from Dr. Podkletnov is that gravity is the effect of spin -- the spin of all subjected particles, from the sub-atomic level and up being parallel; thus they are all aligned to fall into the gravity well of earth. And spinning objects, such as his superconducting discs, when influenced additionally by an electromagnetic field, will experience a shift in the spin of the sub-atomic and atomic elements. They will be turned and not be aligned in parallel. Thus, they are able to levitate.

But how to polarize the media and get things spinning? Enter Dr. Marcus Hollingshed, an enigmatic figure allegedly from Cambridge University. Dr. Hollingshed claims to have built a six-ringed toroidal coil antigravity device, which achieved great effect using rotating magnetic fields. In January 2003 he announced on the Internet that he has developed a 160kg vehicle able to lift in excess of 2000kg and that it has both horizontal and vertical drive features. His device cannot only go up and down and sideways, but it can push things away and pull objects to it.

In addition, the field that the device purportedly generates is capable of being broadened and weakened, or narrowed and amplified in a lensing effect, with the field producing an absolute vacuum of 2.2m spherical diameter. Best of all, when it's cranked up the core of it goes invisible, although the term Dr. Hollingshed uses is that there is a "loss of reflected light." There are no reports of independent confirmation, and Nick Cook says he hasn't been invited to see it, so, he's skeptical.

Where does this leave us? Perhaps Dr. Podkletnov's words sum up our current situation. "Modern theoretical physics cannot give you the direct answer to your questions (levitation, torsion fields, etc.,) and a scientist who would agree to give you the answer cannot be regarded seriously, softly speaking. If you had asked Dr. Einstein if he were an expert on gravity, the answer would be NO. I can repeat his words: No, I am not a magician, yet; I am still learning."

ELECTRICAL LEVITATION

Wayne Macleod, USA
Email: dynamars2100@yahoo.ca, cwleod@shaw.ca

Gravity is the incremental slowing of time as we move closer to its source. If we imagine panes of glass stacked on top each other, one pane slightly denser than the one immediately above, a ray of light coming from above will constantly refract in a curved beam downward as it enters the stack. The same is true of a ray of light entering a gravity field. The analogy is not too far fetched because refraction is caused by a slowing of light in the denser medium. Of course, gravity affects material objects as well as light, but this is explained by every object traveling on a 'world line' in spacetime. We can use light as in the glass example because light is its own world line.

Why the world lines of objects take the paths they do can best be understood by compressing our normal 3 dimensional space into 2 dimensions, length and height only. When we throw a stone into the air it rises and falls in a parabolic arc in these two space dimensions. That path in space is a complete mystery until we consider another dimension, time, a dimension we can imagine measured at a right angle to the plane of the 2 space dimensions. By multiplying the time of the object in flight by the speed of light, we have a three-dimensional coordinate system, not of space but of spacetime. We can then graph the world line of the object. So why does the object fall? It is because spacetime near massive bodies is curved. A remarkable fact is that the curved world lines of all objects in the same gravity field, whether of a thrown ball or fired bullet, have the same radius of curvature. Hence all objects fall with the same acceleration.

To neutralize gravity we must somehow 'straighten' the world lines of objects traveling in curved spacetime. It would be as if we had in our stack of glass panes a bubble, around which the density of glass became less on approach to the bubble. Then the ray of light, when close to the bubble, would refract opposite to the path it would normally take when traveling thought the rest of the glass. Similarly, if we could manufacture a 'time bubble' in a gravity field opposite to the time slowing known as gravity, we would have any object embedded in the time bubble isolated from that gravity field.

The Electrical Levitation theory can best be understood by first considering an analogy. Let us imagine a wheel spinning on an arm like a child's propeller toy. The arm also rotates, in a direction opposite to the spinning wheel. We consider the rate of spinning of the wheel from the point-of-view of two observers, one observer stationary on the ground, the other observer rotating with the arm. Obviously the two observers will not see the same rate of spinning on the wheel. Because the arm is rotating opposite the spin of the wheel, its rotation must be subtracted from the rate of wheel spin as seen by the stationary ground observer. This is not true of the observer rotating with the arm, who will see the spin of the wheel as if there were no arm rotation.

The concept is easier to envision with linear motion, such as of a man throwing a ball on a moving train. The velocity of the ball seen by the pitcher on the train will not be the same as its velocity seen by a stationary observer on the ground. As the stationary observer sees it, that velocity will have the train's velocity subtracted from it if thrown against the train's velocity. It is the same with rotary motion: the velocity of the wheel for the stationary observer is slower because the arm's rotation is subtracted from it.

But suppose the wheel rotation is *the same* for the two observers! Something must be different between them, and that is time. As explained by Special Relativity for linear motion, time is not a universal constant; it differs between observers depending on their relative motion. The property of nature that *is* constant, that produces this relative time difference, is the speed of light. So we must look for a similar constant in nature for a relative time difference between observers in our rotary example. We have one in electron angular momentum, known as electron "spin," $h/4\pi = 5.28 \times 10^{-35}$ kg-m^2/sec, where h is Planck's constant. Like the speed of light, this quantity is constant for all observers, whether the observer is on a rotating system or stationary on the ground. Here is the tool for producing our time 'bubble'.

Let us now imagine a series of concentric rings, all rotating in the same plane and in the same direction. Electric current is pushed through these rings in the same direction as the ring rotation. (Current here is considered the flow of electrons, not conventional positive current.) The rotating rings are sandwiched between two magnetic plates, the function of which is to maintain the angular momentum of the current electrons oriented properly with ring rotation.

Analogous to the above example, each electron takes the place of the spinning wheel and the rotating arm is replaced by the rings. Thus, because electron angular momentum is a universal constant, an observer of the electrons in this rotating system will not have the same time as a stationary observer outside it. If each current electron has its "spin" oriented opposite the rotation of the rings, time on the rotating system would run faster than for a stationary ground observer, the same as in empty space relative to the Earth. The rotating system would therefore have the world line of empty space, not that of the gravity field. Its spacetime would not be curved. Since gravity is a time phenomenon, such a system in a gravity field could not have the behavior of a normal object. This conclusion may seem erroneous considering that ring rotation could never reach relativistically significant velocities, but we have an analogy with magnetism that is caused by a Lorentz contraction of the distance between moving electrons. If we considered the diminutive drift velocity of electrons alone we would never conclude that magnetism was possible, but the effect is *accumulative* over trillions of electrons. The hypothesis here is that the same is possible with the trillions of free electrons oriented with a spinning ring, to produce an accumulative time effect.

We now come to the controversial part of this theory because the above treats electron angular momentum the same as angular momentum of a normal physical object, whereas the electron is a quantum particle and quantum particles have their own realities with bizarre consequences when analogized with our macro universe. For instance, the electron has quantum spin number 1/2, and such a particle would have 720 degrees in one rotation, not 360. Nevertheless, the electron *does possess* di-pole magnetism. It *does behave* as a spinning ball with negative charge. Electrical Levitation is therefore an empirical theory. The property of the electron that gives electron di-pole magnetism cannot be an actual physical rotation, but whatever that quantum property, if it produces di-pole magnetism there is reason to expect it to produce other macro physical phenomena as if it were.

The theoretical finding of General Relativity that time runs slower in a gravitational field was confirmed by the Pound-Rebka experiment in 1959. **It is not that gravity causes time slowing, gravity *is* time slowing.** Since time and energy are reciprocal, more time on our rotating ring system would mean less energy seen by an observer in that frame of reference than seen by a stationary observer on the ground, the opposite of the red shift of a gravity field. This energy difference must equal the energy of the mass in a gravity field that is to be levitated, its energy of weight, and lost. Experimenters should therefore be aware that their device might radiate. But this is not a free energy machine. The energy of levitation comes from its magnetic field, and the electron magnetic moment energy turns out to be $V_e = mc^2$ divided by the electrical current, *m* is the mass to be levitated and *c* is the speed of light. This is an enormous amount, but becomes practical if enough electrical current can be sent through the rings. That means the electrical resistance of the ring material must be very low. For copper it is not. No material currently exists with sufficiently low electrical resistance at room temperatures, but **superconducting materials** exist that at cryogenic temperatures experience a dramatic loss of electrical resistance.

An experimental device can therefore be envisioned using a superconducting disc substituting for the electric rings. The disc would serve as a conductor for an electric current and would therefore need to be sliced along one radius with an insulator placed in the notch. Both edges of the notch would be connected to a power source by brushes rotating with the disc. With a counterclockwise disc rotation as seen from the top, to have clockwise electron "spin" the magnetic field between the plates would have to be up. Only the moving free electrons of the current will be available for any time alteration effect. These will also **produce a magnetic field, which it is reasonable to assume would be of the same time alteration effect as the electrons, thus producing the required alternate time "bubble".**

The resources required for an experiment using superconducting material at cryogenic temperatures are beyond this writer's means and an experiment has not been attempted, but an experiment at the Tampere University, Finland, 1992, using a superconducting disc suggests that gravity shielding is possible. Owing to the immense advantages gravitational shielding would give to the present interest in space exploration, effort toward its development would be logical. This essay may offer clues on how to begin.

Editor: I find the article by Wane McLeod quite interesting as his approach is very similar to the idea I first presented in my report at the "Space, Time and Gravitation" Conference which took place in St. Petersburg in 1998. The report was included into The Proceedings of the Conference, Part I, 1999. Before that the brief version of the article was published in English in the 'ELECTRIC SPACECRAFT' magazine, Leicester, North Carolina 28748 USA, Issue 27, 1997 p.30-31.

I managed to demonstrate that the ideas of Thomas T. Brown, especially his USA patent # 3,187,206 of 1965, are something more than just the force asymmetry in the electric capacitor. According to the concept presented in my report, by creating a matter property gradient (in particular, the dielectric permittivity gradient) we actually change the curvature of the electric force line in space. Normally, the natural space curvature accounts for the electric field potential decrease with the distance increase from the surface of the charged object. By creating the dielectric permittivity gradient (described by T.T. Brown in his patent of 1965) we change the natural distance potential gradient law. We can both increase and decrease this change and even reverse it. With the dielectric permittivity change square function, the natural space curvature is completely compensated and with more extent it is reversed and can be turned to negative. At that rate, the potential is not decreased but increased with the increase of the distance from the surface of the charged object. This is the essence of my concept that has never been considered before by any author. By creating the gradient described by the quartic function we get the same distance potential gradient law as in the natural conditions only with a different sign!

T.T. Brown, who discovered a force in capacitors with a special dielectric, offered the practical application of this concept. However, he did not find the source of this force, which, according to my concept, is conditioned by two factors: the corresponding dielectric property gradient function and the elastic properties of the dielectric material, which is of no less significance. Normally, dielectric particles are attracted to the charged surface but when the reversion of the curvature of electric field force lines occurs, they are repulsed from it and we can use this propulsion force. Elastic deformations (including those occurring during the pulsating operation of the field) account for the reaction forces equivalent to the generated propulsion force. Thus, the momentum conservation law is in action, however, the device is still moving.

T.T. Brown also considered other ways of generating a force, such as the material magnetic property gradient. Logically, by analogy with the electric field we can consider the gravity field, which can be "designed" and "reversed" by means of creating a matter with the density gradient. At that rate, the gravity potential must change in accordance with a certain law when the distance from the gravity field source being increased.

We should also note that the local space volume levitation effect with the accelerated or decelerated time rate was first described in my articles "Physical Principles of the Time Machine", NET #3 (6), May-June 2002 and "Practical Application of the Time Rate Control (TRC) Theory", NET #3, November-December 2001.

Our company conducts experimental researches on creating new materials that possess the properties described. We are interested in serious business contacts with companies in the aerospace industry to further discuss eventual cooperation projects in this field, including *co-patenting*.

Alexander V. Frolov

Perpetuum Mobile Not in the Past but at the Present

Vladimir I. Likhachev, Russia
Krasnodarskaya str., 38/20-34, Moscow, 109386

In several issues of the "Nature and man. Light" magazine in 1995-97 my colleague E.Oparin and I wrote on the unscientific character of absolutizing the law of degradation of energy, on the important issue of creating the perpetuum mobile of the second kind utilizing infinitely available *environmental heat energy*. We even send a letter to Yury S. Osipov, the Russian Science academy President, but received no reply. Notwithstanding many official discussions, there are no well-reasoned objections to our arguments or proof on the part of "high" official science, because, in fact, we are right.

I have recently received one more confirmation: bright and feasible data on the negentropy cycle (the cycle with spontaneous reduction of entropy) with a chemically active working substance that will be described later.

After the publication of Victor M. Brodyansky's book "Perpetuum Mobile in the Past and at Present. From Utopia to Science and from Science to Utopia", Second Edition, one can not but return to these issues. More so because the foreword of V.A. Fabricant, member of the USSR Academy of Pedagogical Sciences, claims the law of degradation of energy to be the "law of nature" and supports shamefully the deformation of the monothermists, especially P.K.Oschepkov, by academicians P.Kapitsa, L.Artsimovich and I.Tamm ("Pravda" of November 22, 1959), followed by E. Velikhov, A.Prokhorov and V.Sagdeyev ("Pravda" of June 22, 1987).

After the publication of Brodyansky's book I met the author several times (notably at a special seminar in The Moscow Energy Institute) and expressed my opinion about his book: the book avoids serious discussion of the monothermists, stresses their failures which are quite normal in the initial development of perpetuum mobiles of the second kind.

For example, Brodyansky knew about Tsiolkovsky's discovery. Guay's book "On the Little-Known Hypothesis of Tsiolkovsky" is listed among the First Edition references. But Vladimir Brodyansky avoided analyzing this discovery both "in the past" and "at present". The Second Edition references even list the "Physical Thought in Russia" magazine, issue #1 of 1991, that contained for the fist and only time the modern edition of the "Law of Degrading of Energy" by Tsiolkovsky. "At present" V. Brodyansky knows my analysis of Boltsman's mistakes but he paid no scientific attention to my analysis either.

In Brodyansky's book the description level of chemical reaction cycles is even lower than that of Schpilrine. Schpilrine at least tried to present his arguments but here we see unsupported statements "justified" by the phrase "the detailed discussion would take too much space". Vladimir Brodyansky's analysis of anti-Stokes luminescence (pp.216-219) is also superficial. Dissipative and entropy processes prevail in it. But if there is evidence supporting higher frequency quanta than the frequency of radiating flux, **it is the evidence of negative entropy**. Vladimir Brodyansky not only is familiar with it and also shows it by the example of investigating the Sun and the Earth (p.247).

Vladimir Brodyansky's analysis of Rank's tube (pp.235-237) displays the same tendency. In the Rank's tube the dissipative processes also prevail. Its effectiveness as a refrigerator is lower than in the traditional schemes.

But one can not deny the existence of negentropy processes in it. It is these processes that "originate" the cold gas flow. And Brodyansky's "origination" on p.236 "The vortex tube in fact divides the coming gas into two fluxes: heated and refrigerated" - is totally false. That can be performed only by "Maxwell's demon" that does not exist in substance.

The refrigerated flux in the Rank's tube is formed by the "Tsiolkovsky's flux" by way of heat transfer "from cold to hot", from the tube center to its circumference by **centrifugal accelerations**. On its basis (to be more exact, on the basis of Finko's tube) the perpetuum mobile of the second kind can be created. And we have this engineering solution. E.Oparin and I made the corresponding patent application for an invention but the Federal Institute of Industrial Property again refused to consider such applications. Here one can see similarity with the causes of re-edition of Vladimir Brodyansky's book. And one need not be a big politician to admit the reality and understand the sources.

Environmental and, especially, anti-nuclear movement is invincibly expanding. Danger to human life and threat of ecological catastrophe are quite real. The Second Edition of Brodyansky's book protects behind scientific arguments those who use Russia as a source of easy money and then flee.

But we like our country - vast and austere. This austerity and space form REAL PEOPLE and support humanity and materialistic humanism of the energetic and talented, not just the rich.

In modern Russia the powerfully spread alcoholism, debauchery and drugs are the instruments of slavery. But the monothermy and development of alternative power engineering will become basis for Russia's revival and unprecedented prosperity.

Russians learn slowly, sometimes, very slowly, even disgracefully slowly. But we will learn. And then no one will stop us. The following tips are meant for those who are ready to risk their efforts and facilities in order to build alternative energy knowledge base: dissipative and negentropy processes are inseparable and very often their mechanism is the same. In our environment the dissipative processes are more effective and hide the negentropy ones. No dissipative process combination can generate the negentropy cycle. Look for examples in Brodyansky's book. To create the negentropy cycle and use the environmental energy at least one negentropy process is required that has effectiveness and negentropy exceeding entropy growth at all levels of the cycle.

News from “IntAlek” Company

William Alek [alekws@intalek.com] has updated the latest and greatest ZPOD system drawings:

hhttp://www.intalek.com/Index/Projects/SmartPAK/Projects/ZPOD/ZPOD_System05.pdf

This is release 5.0

Here is the latest electrical schematic: http://www.intalek.com/ZPOD/ZPOD_System05.gif

Based upon his "preliminary" tests thus far, he estimates the COP is around 2.0. This estimate is interpreted from the following scope traces:

http://www.intalek.com/ZPOD/in.GIF

Alan Francoeur's Generator
Canada

Alan L Francoeur, Canada
Email: al.f@shaw.ca

This is my writing and development about the Interference disc generator, and the permanent magnet dynamo machine, and other related inventions. I have always been intrigued with electricity and magnetism and aetheric energies starting at a young age. Back in 1980 I wanted to do something to help get my large vehicles better fuel economy. With this goal in mind, I experimented and designed my first heat exchanger vaporizer system to do my part to clean up our cars and trucks, fig. 1, 2, and 3 are two earlier vaporizer units under construction. For more information about this fuel system, please refer to my vaporizer fuel system report. The main body of this writing is referring the permanent magnet dynamo system that has been in the making since 1987, starting after the development of the interference disc generator.

Fig. 1

Fig. 2

Fig. 3

I invented the Interference Disc generator concept while I was employed at Giant Yellowknife Mines NWT from 1983 to early 1986. The idea of using metal blocking plates to shield magnetic fields stirred in my mind while working around heavy electrical mining equipment, and 1000hp and other electric motors with exposed coils that have heavy metal around the outer casing. I did a search at the time and found John Ecklin's work very interesting with his application of the shielding effect. During my off time in Yellowknife, I was experimenting with the interesting effect of producing induction with a stationary coil mounted beside a stationary magnet with a small air gap maintained between the magnet and the coil. During the construction of my first disc machine, I called it the Interference Shielding effect because of the interference the blocking disc does to cause induction, thus the Interference Disc Generator fits the description. The first transformer coils I used to test the interference shielding effect were used from old radios, and the magnets were alnico types. Back in 1986 in Calgary Alberta, a friend Bud Johnson and I constructed the first test disc generator model using alnico magnets, fig. 4, and later ceramic magnets were installed. Months later I designed and began construction of another larger Interference Disc machine with horseshoe shape neodymium 35 grade magnets mounted on the outside, and with the N and S magnetic poles facing toward the coils in the center portion of the machine. The coils and magnets in this larger machine are mounted in reverse of the first Interference Disc generator, fig. 5, 6.

Fig. 4

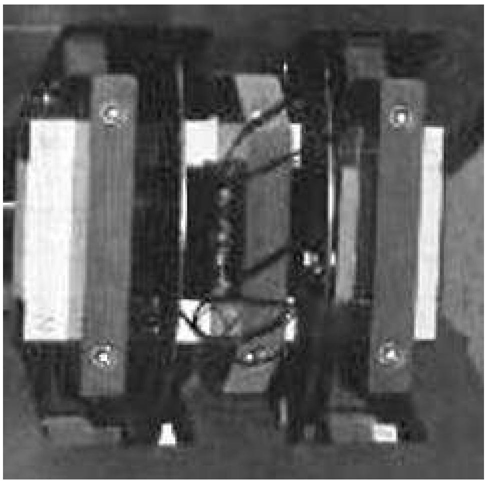

Fig. 5

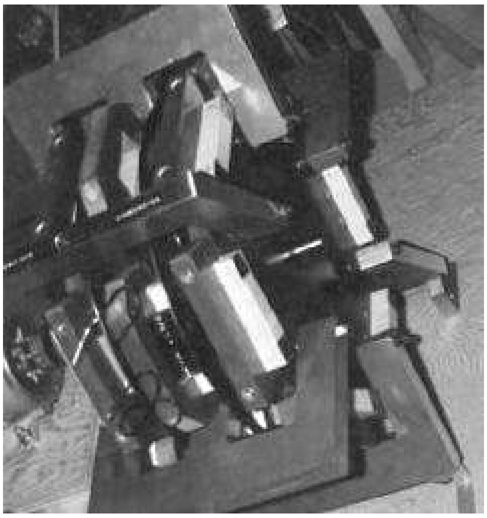

Fig. 6

General description of the interference disc generator

The Interference Disc electrical generator is made of a stator (114) consisting of an array of even number parallel mounted bar magnets (112) arranged and supported (114) in a circular fashion equidistant from each other, where the polarity of the magnets (112) are alternating when viewed at either end. Fig. 7

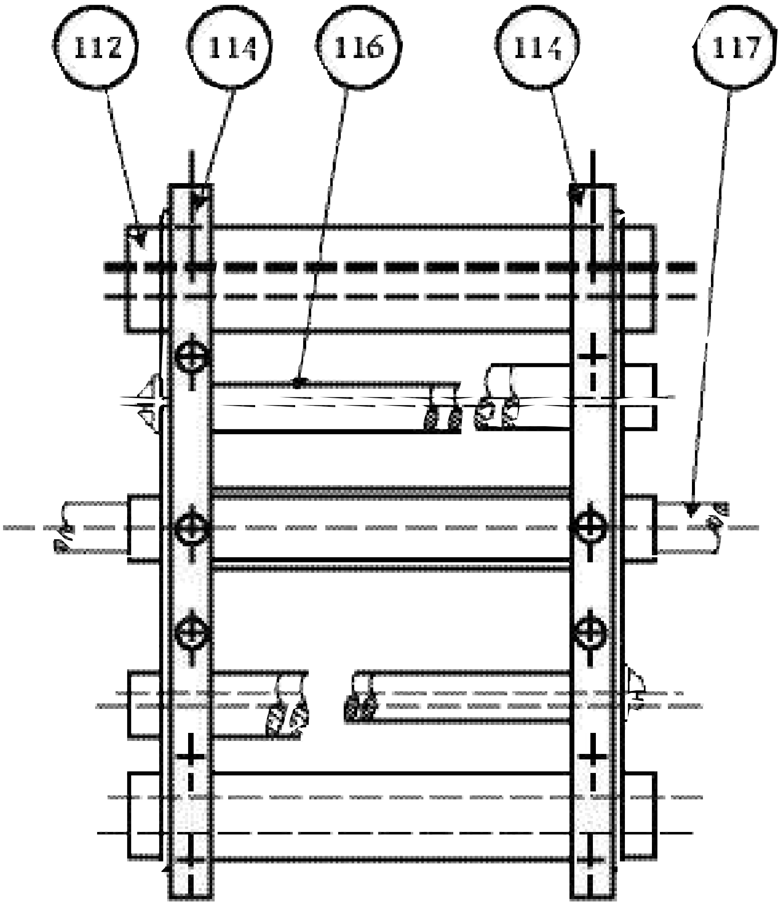

Fig. 7
Magnetic bars assembly (rotor)

Also in the stator fig. 8, (102) there are twice the number of coils (111) as bar magnets (112) wound in pairs on U-shaped cores (110). One half of the coil pairs mounted at one end of the bar magnets (112) in the same circular fashion, each corresponding to and aligned opposite to a pair of bar magnets (112) and separated from it by an air gap. On the other end of this pair of bar magnets (112) there are another set of coils (111) aligned likewise.

The rotor is made of two magnetically susceptible circular plates (109A or 109B) (such as iron or steel) Fig. 9, centrally mounted on an axle (107) fig. 7, which is in respect to the array of magnets and coils Fig. 9a.

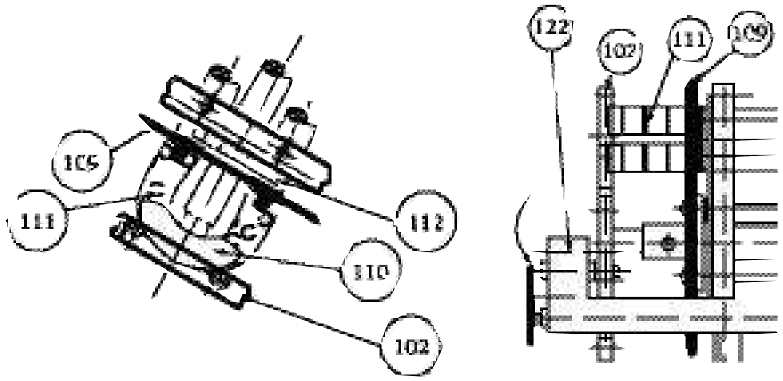

Fig. 8

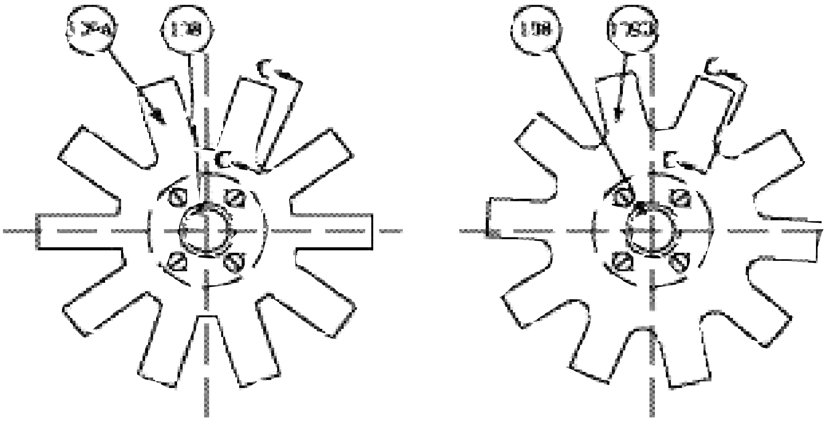

Fig. 9

The interference discs (109A or 109B) are separated from each other so that they can rotate in the air gaps between the magnets (112) and the coils (111). The interference discs (109A or 109B) have a number of **equidistant opening, either more or less in number than the number of the bar magnets (112)**.

The arrangement and size of the openings are such that when the discs (109A or 109B) are rotated by an outside motive force, they open the magnetic field between adjacent bar magnets (112), and the opposing coils (111) at both ends of the stator (114) simultaneously. Hence inducing an alternating electric current in the stationary coils (111) fig. 10.

By opening and closing the magnetic influx to the core (110) of the coils (111), an alternating current will be generated in the coils

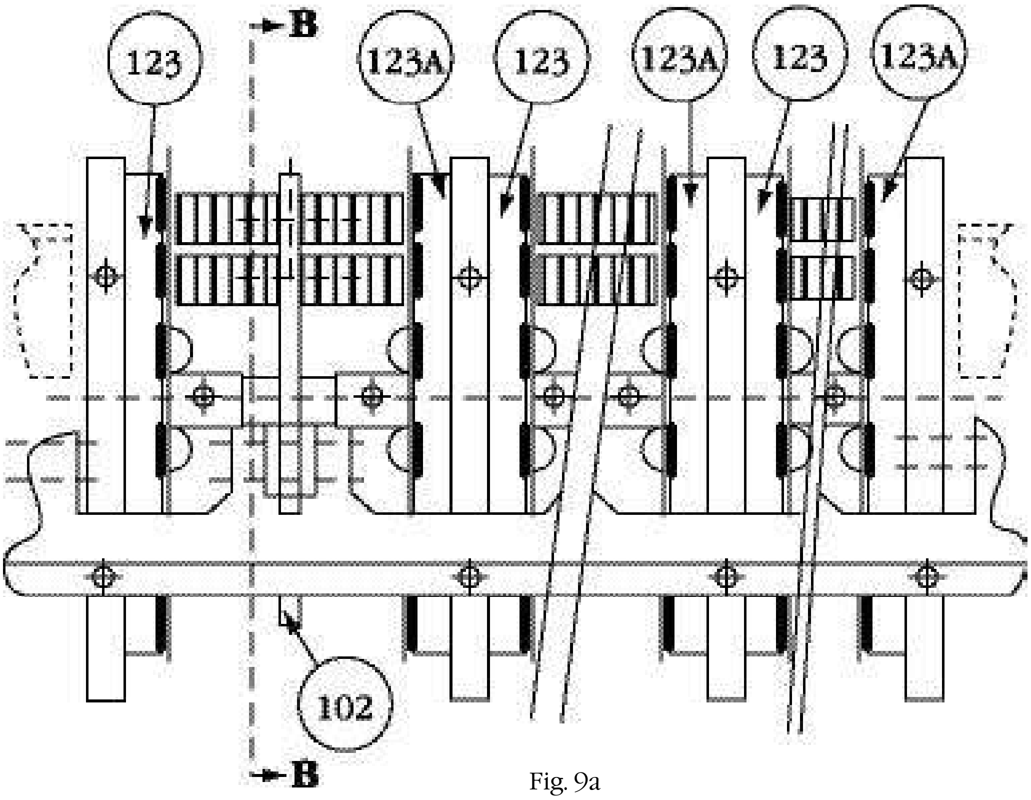

Fig. 9a
Magnetic bars assembly (stator)
Option II: multable magnetic coil-disk assembly

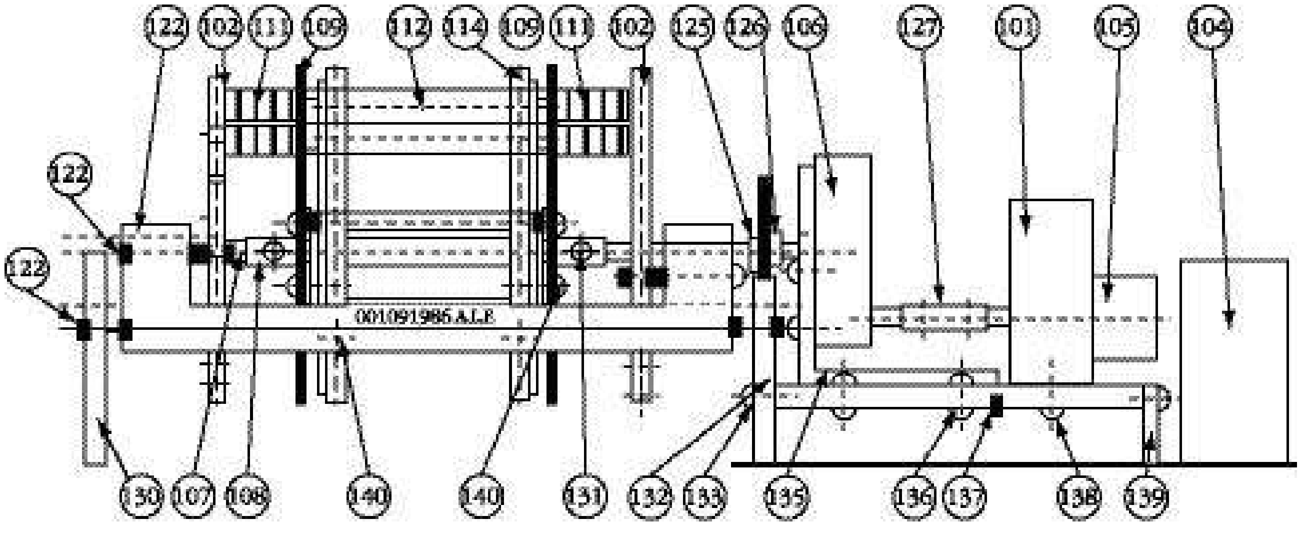

Fig. 10

(111) in a retro-order to the rotational direction of the interference discs (109A or 109B). Depending on the number of magnet-coil assemblies three phase AC current can be obtained. The regulation and attenuation of these currents can be achieved by known electric engineering methods (Fig. 10).

The efficiency of the unit would vary greatly depending on the speed of rotation, the width of the air gap, the strength of the magnets, and the materials used to construct the coils. I used nickel base amorphous metglass in the coils in both the smaller and larger machines, which as a result, shows greater efficiency as compared to conventional core laminates. Less energy is needed to spin the balanced disc rotor while causing a larger moving magnetic field to cause induction in the stationary coils. With this effect, it is easy to visualize the advantages this design has over conventional generators and motors. New testing of the coils will be conducted after the machine goes through all the upgrades to eliminate the problems with the warping interference discs.

The Interference Discs Generator described in this document **has the potential to produce more than unity effects**by easy turning of a magnetically balanced (non-warping) interference disc. Optimizing the device with closer tolerances and correctly selected materials and incorporating internal magnetic balancing, will add to the success and improved performance of this unique generator design.

Improvements can be made with this design while the operating principal remains the same. Looking at the photographs included you will see the 1st proof of concept generator frame is constructed mainly from aluminum, this metal was used only because it was inexpensive material and easy to work with fig. 11. Non magnetic and non-conductive materials should be used to construct the generator frame supports to eliminate the eddy current losses. All of the nuts and bolts used to hold the disc generators together worked well for making the test machines capable to have adjustable air gaps, however, they can all be eliminated with design changes that preset the gaps and tolerances to optimize the conditions of the interference disc effect.

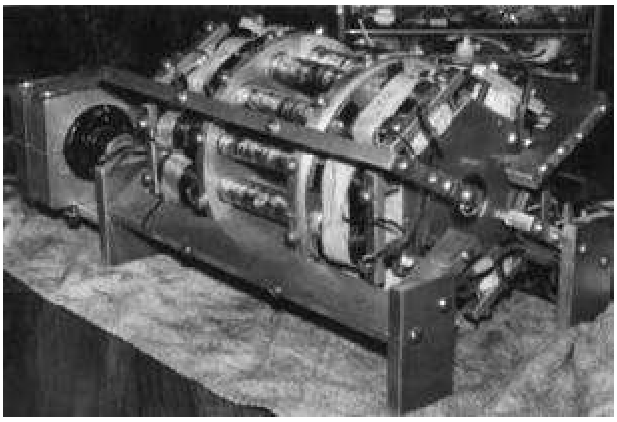

Fig. 11

Many improvements were adopted in the next larger #2 version of the Interference Disc generator fig. 12. Phanolic materials (non conductive and non magnetic) were used in the frame supporting the magnets and the coils, thus eliminating eddy current losses within the frame support. Large aluminum E bars were used to support the magnet frame support and the coil frame support with no noticeable eddy current losses. The aluminum used is positioned further away from the coils and magnets in the larger generator so as not to be affected by the magnets. Larger more powerful neodymium magnets and I shaped transformer coils were installed and tested in the larger machine which showed an increase in output power over the 1st machine fig. 13. Using more powerful neodymium magnets in this larger disc machine created some problems with the interference discs. Not only did they warp the discs they caused them to vibrate and chatter causing the fins to hit the magnets and the coils. I increased the air gap to over half an inch and moved the disc further away from the magnets to prevent damage. The effect was a reduction of magnetic flux reaching the coils that reduced the power output. The discs must be constructed much stronger and rigid if using neodymium magnet grade 35 or greater, then the air gap can be closed so more magnetic flux will reach the coils allowing for a greater electrical output.

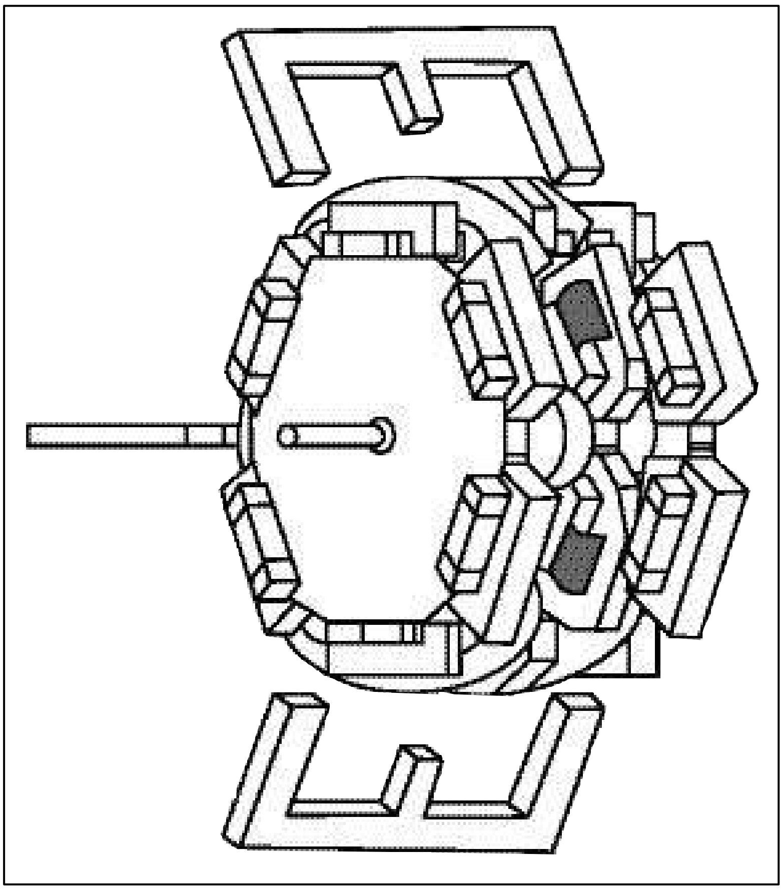

Fig. 12

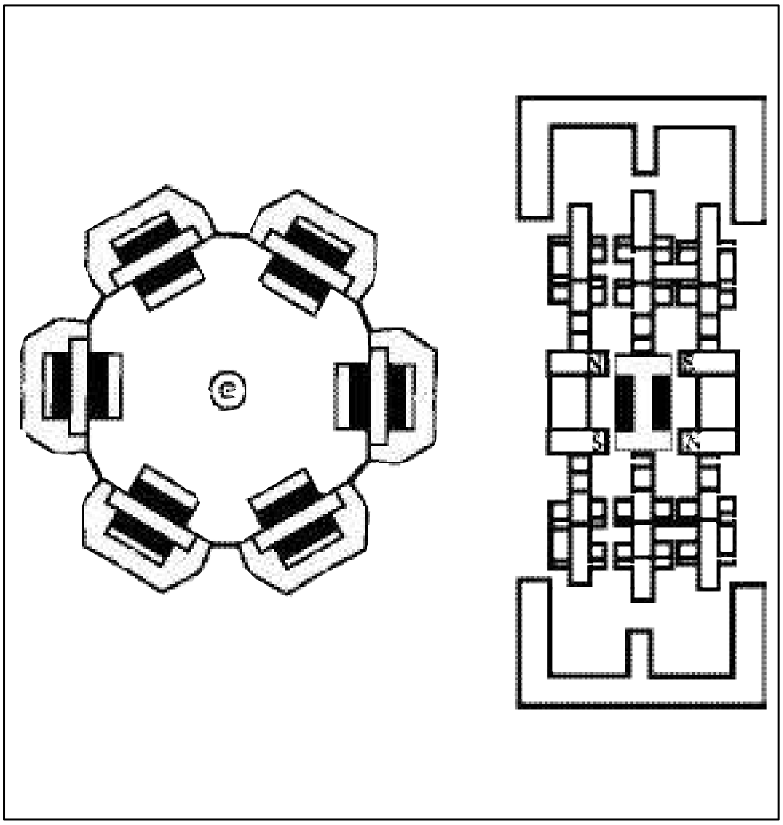

Fig. 13

The wiring of the coils circuit can be constructed by known engineering methods to achieve AC or DC power, and like any conventional transformer, wind the coils with heavy wire for more amps and more turns for more volts. The coils are stationary in this machine which makes it easy to harness AC power directly from the coils without brushes. All coils that are in phase can be wired in series or parallel, the number of the blocking plates on the interference disc determines the phasing of the coils.

The diagram in fig. 14, 14a shows the path of the magnetic flux movement in relation to the position of the blocking plates of the interference disc. The magnetic field in the core/coil is switching polarity as the interference disc fins alternately moves in and out of the air gaps on each side of the coil. As you can see, depending on the position of the disc, positive and negative induction takes place for AC output of the coils. John Ecklin has achieved this effect using rotating blocking plates. I have achieved the same effect using balanced rotating interference discs with multiple stacks of isolated coils and poles mounted in a circle equal distance from each other.

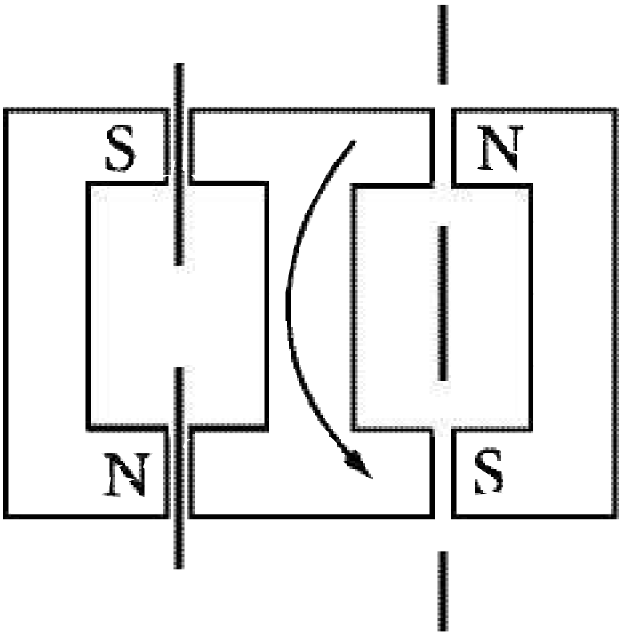

Fig. 14

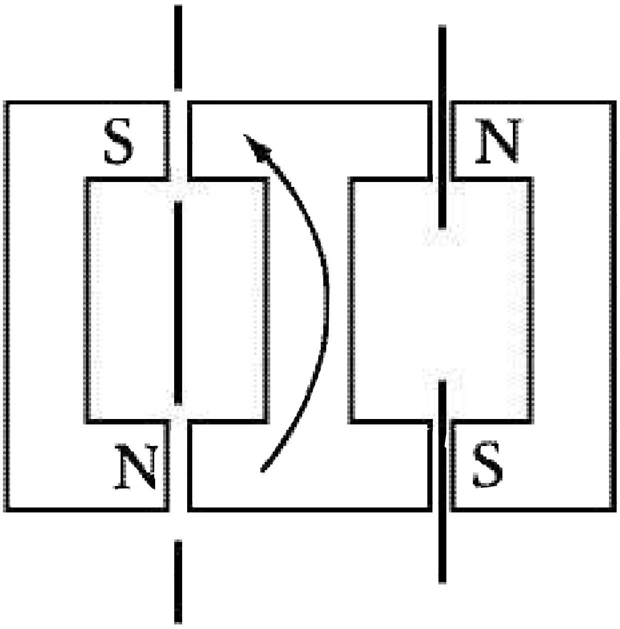

Fig.14a

The effect I reproduced with Gary Wesley's permanent magnet motor is related to the Interference Disc generator in the fact that blocking plates are used to shield the magnetic field, only we differ on methods to balance the blocking shields. Gary Wesley used springs to carefully balance the blocking plate, and I used a different number of interference fins working with the magnets to balance the blocking plates. The magnets in my Interference Disc machine become the springs.

More photos of both Interference Disc machines are shown in fig.15, 15a and 15b below.

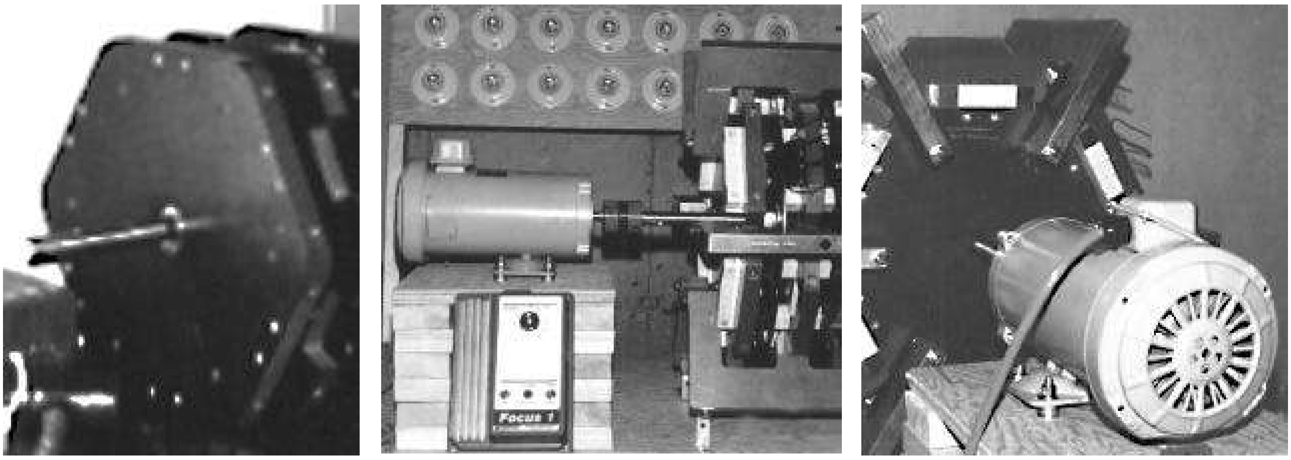

Fig. 15

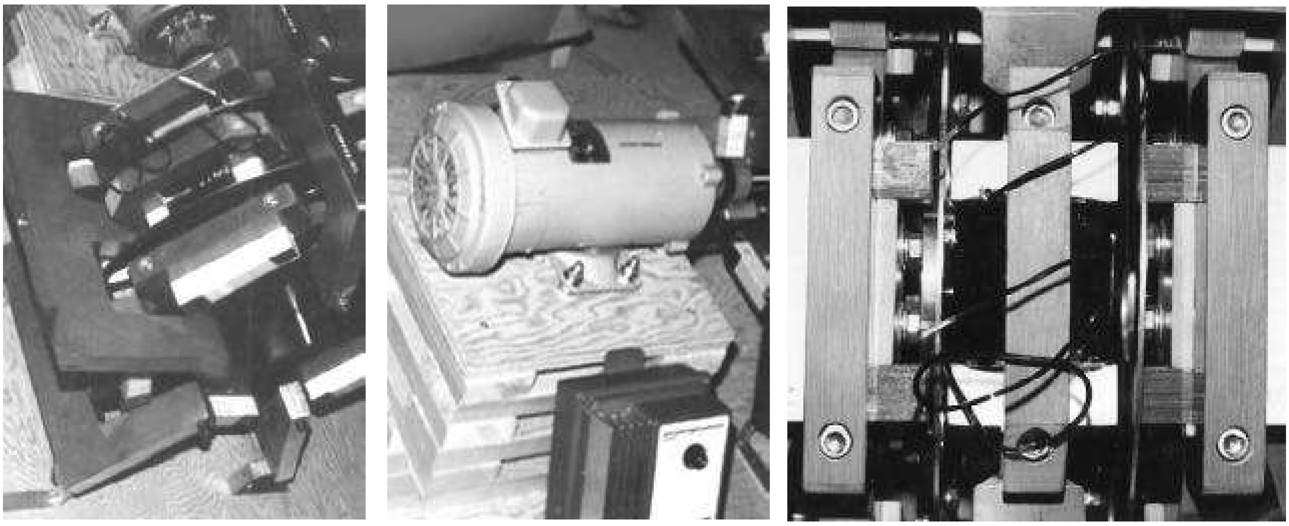

Fig. 15a

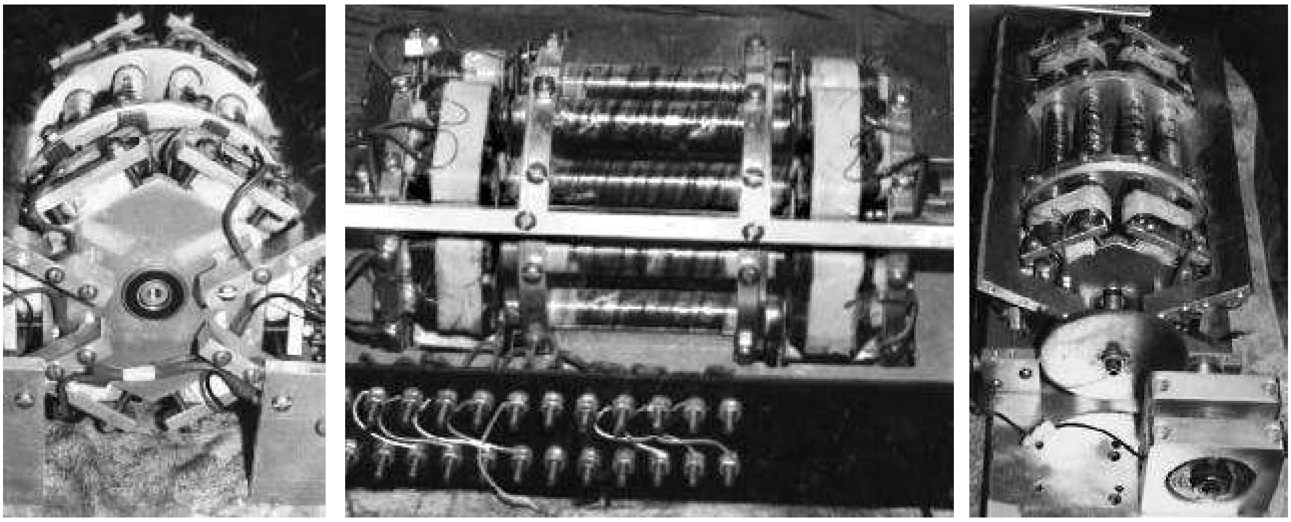

Fig. 15b

Closing comments on the disc generator

In closing this discussion of the Interference disc generator, I would like to go over and point out some of the concepts that I incorporated into the design of my permanent magnet dynamo.

As I made further progress and completed the later stages of the Interference disc generator, it became clear to me to use certain concepts of the machine and incorporate them into the design of my multiphase permanent magnet dynamo. Some of these concepts are explained here.

I have constructed several interference discs that have a different number of blocking shields, some with 10 blocking fins and some with 11 blocking fins and others with 12 blocking fins. They are all interchangeable on the machine to change the phase relations between all the isolated coils. The 12 finned disc will produce a single-phase dynamo with all the coils in phase, and the 10 finned disc on the machine will produce a three phase dynamo. One phase represents four groups of isolated coils, and the second phase represent the other four coils, and the four coils left are the third phase, for a total of twelve coils. All just by changing the number of fins on the interference disc and or adding more layers. Using 12 magnet poles and the 10 or 11 interference fins on the disc, the rotor discs becomes magnetically balanced and easy to rotate in spite of the load on the coils. The 12 finned disc has a larger holding force because of the same number of fins as magnetic poles. However, it can be reduced to almost nil by adding more layers of interference disc and balance them against each other. Many in phase coils can then be added in series or in parallel to build your power levels to any desired potential.

Both machines have stationary coils, and the disc generator has 12 magnetic poles (6N and 6S), and the permanent magnet dynamo also has 12 magnetic poles (6N and 6S). The disc generator uses 10 or more fins on the interference shielding plate per rotor, and the permanent magnet dynamo has 10 or more coils around each rotor. There is the same number of interference blocking plates as there are induction coils on each of the two difference machines and the number of blocking fins and or coils are interchangeable. The back emf effects are different on both machines, the Interference disc generator demonstrates easy turning of the discs while the stationary coils are loaded, and when producing alternating current the permanent magnet dynamo rotor slows down when a load is placed on the stationary coils. Both machines produce AC power out of the coils and the Interference Disc generator has a closed magnetic circuit through the coils, and the permanent magnet dynamo has open magnetic circuit through the coils.

The Permanent Magnet Dynamo

I would like to begin this topic at this point by describing my permanent magnet dynamo machine.

I designed this machine to be completely interchangeable into many different motor generator combinations. These combinations involve different number of magnetic poles and different number of interchangeable coil poles in the same machine, two versions are possible, one with isolated coils and another with the coils that share a common core fig. 16.

We can change the magnetically balancing effect (*holding force*) by changing the number of coil poles or (*interference disc fins*) in relation to the magnetic poles. This turns the machine into a multipurpose, multiphase, variable output, and variable frequency ac dc permanent magnet motor generator system. I first designed this unique multi purpose dynamo in 1987 after I built my third Interference Disc generator.

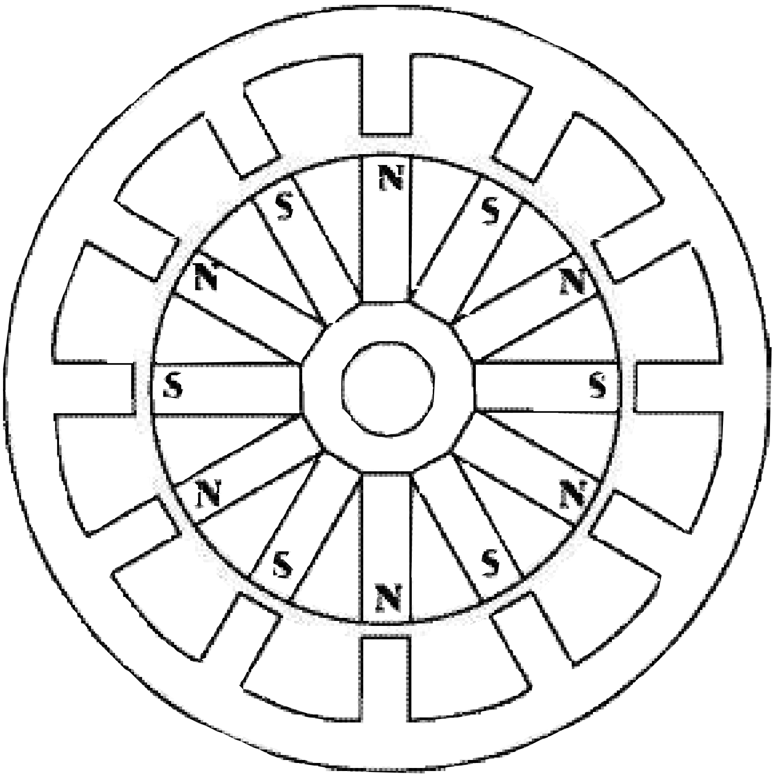

Fig. 16

Dynamo description:

Rotor:

It is well known that a bar magnet has its magnetic field concentrated at the N and S pole ends, and the half way line between the poles of a magnet is the neutral zone, this is the region where the magnetic field is canceled or neutralized. The *ferrous* keeper ring I designed for this rotor is mounted near the axle and it has machined flat decks so the magnets attach flush to the deck of the

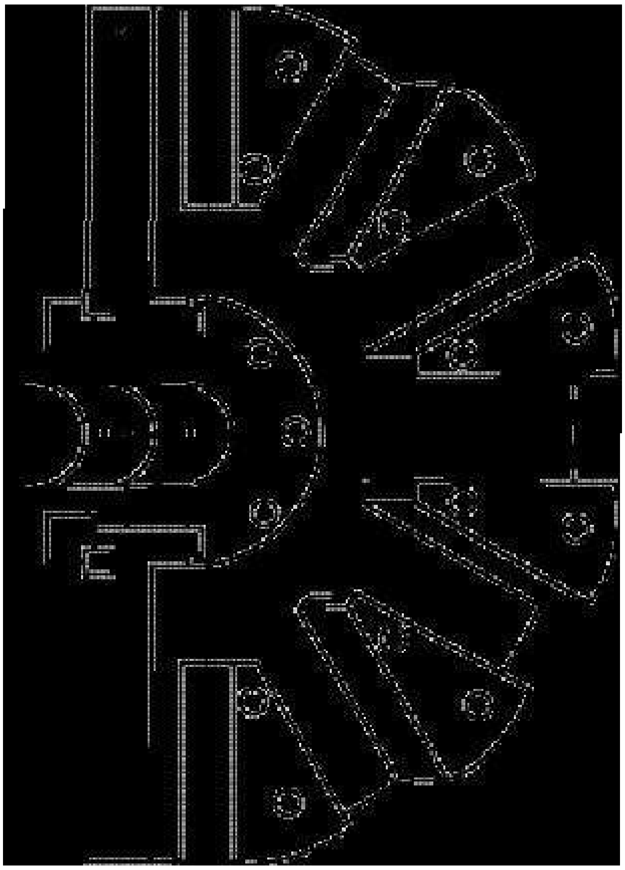

Fig. 17

keeper ring. Because of this feature, flux leakage is reduced at the keeper ring and the magnetic field concentrates at the rotor poles. With the poles of the magnet positioned near the axle and attached to the keeper ring in the rotor, the magnetic field that was there is now forced to add magnetic field density to the outer magnetic poles on the rotor rim. The density of the magnetic field at the poles is now greatly increased which makes a very powerful small size rotor. The *ferrous* inner keeper ring (*active iron*) that is installed inside this rotor design, is also the *neutral zone* of the magnetic field when viewing all the combined permanent magnets in the rotor. Shown is a cut away view diagram of the rotor, fig. 17.

I utilize rare earth rectangular shape neodymium grade 35 bar magnets with the dimensions of 3" long and 1"x1" on the end poles. There are 12 bar magnets mounted 30 degrees apart in NSNS configuration, and the bar magnets are attached to a precision machined inner keeper ring which makes the active iron the region of the middle of the magnetic field (the neutral zone). Another way to view this is to imagine two 3" long bar magnets attached together, you now have a 6" long magnet. Bend this magnet in half and install iron at the U and it now becomes a horseshoe magnet with two 3" long pole ends. It's a powerful rotor design in the fact that each of the 3" long bar magnetic poles has the field strength of a 6" long bar magnet. More poles on the rotor will allow for higher frequency ac generation at lower rpms, twelve magnetic poles on each rotor is shown in fig. 18, 19.

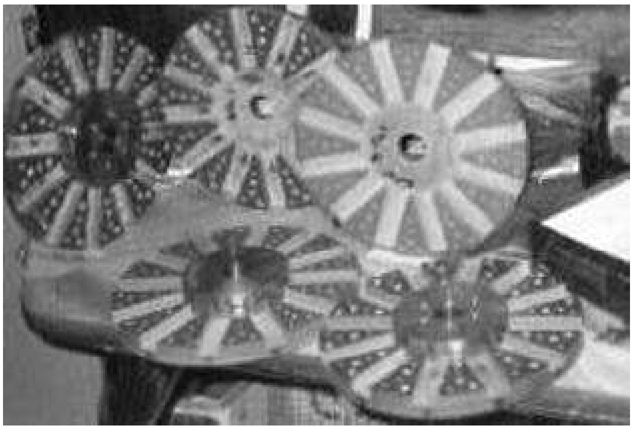

Fig. 18

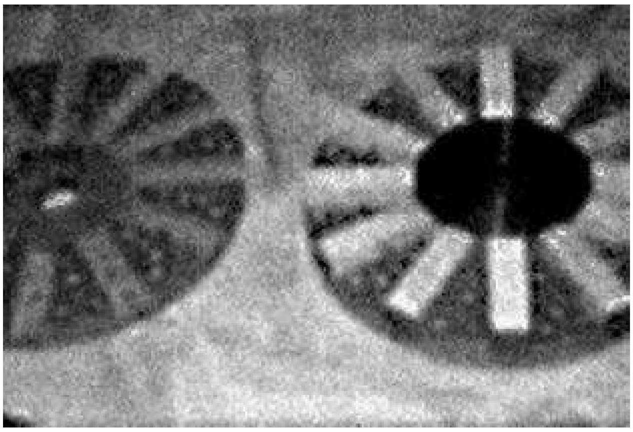

Fig. 19

The drawing in fig. 20 and picture in fig. 21, shows twelve magnet poles in each rotor in relation to the position of the coils. Ten *amorphous* cores are shown mounted around each of three rotors before they were wound with magnet wire fig. 21. The machine I have constructed houses a total of three rotors side by side that are mounted 10 degrees apart in relation to their *magnetic poles*, each rotor has twelve magnets installed for a total of 36 magnetic poles shown infig. 20. This dynamo concept can be constructed to any size with unlimited numbers of rotors and coil combinations to fit any power requirements.

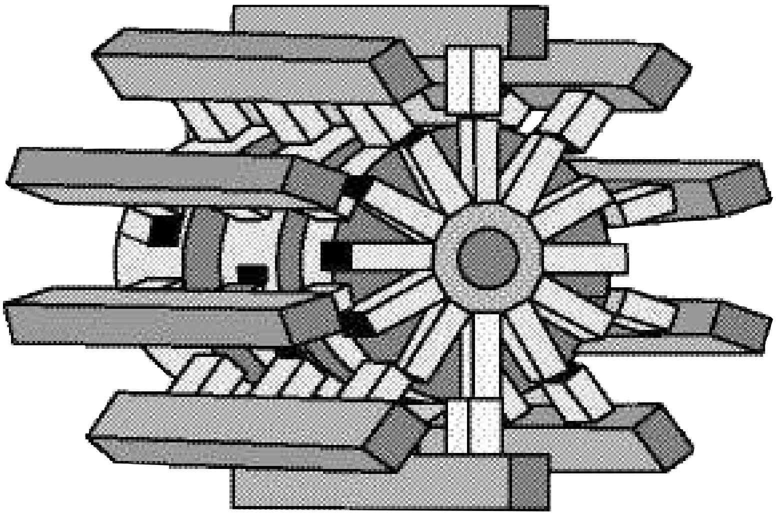

Fig. 20

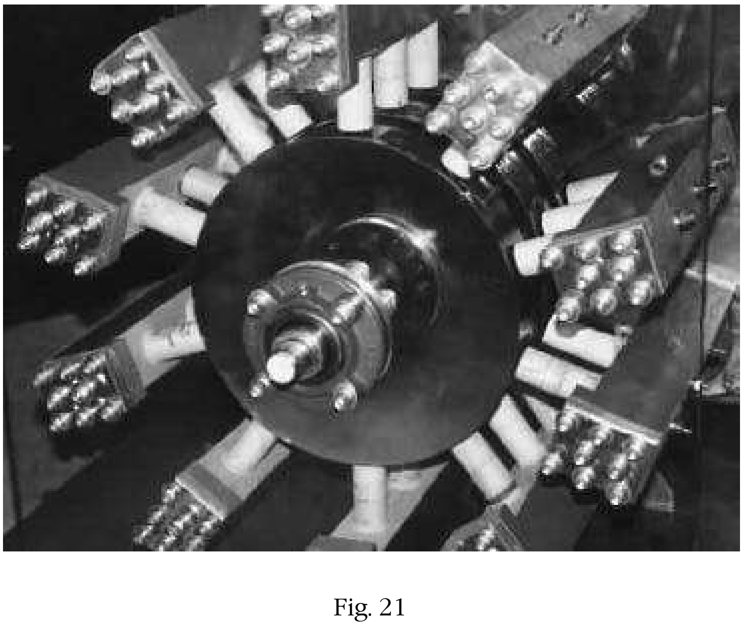

Fig. 21

It is well known that a magnet exerts a force of attraction to iron bringing them closer together, after they have come together, they are in their **static position**. A force must be exerted to either the magnet or the iron in order to pull the magnet and the iron apart, removing them from their static position. The term **static position** is used to describe the point when the magnet and the iron bar are at their nearest point.

As in the Interference Disc Generator, this Permanent Magnet Dynamo I am describing also uses an internal magnetic balancing effect which reduces startup torque on the rotor.

In this system, one set of magnets is in their static position in relation to the coils, such as the impulse coils. Another set of magnets on one side of the rotor is not in the static position, and there is another set of magnets that is not in its static position on the opposite side. One side is half in to the coils, and the other side is half out the coils, allowing the rotor to have minimal start up torque in turning the rotor. The magnets function in two jobs, the first with its velocity is to cause induction into the transformer coils producing ac power.

The second function, is the opposite pair magnetic fields on the rotor balancing each magnet out of their static position in relation to the coil positions at any given time, they cancel their holding force which puts the rotor at equilibrium. This reduces the amount of start up energy required to rotate the rotor resulting in higher efficiency. It allows the rotor and or multiple rotors to turn easily and efficiently despite the inherent powerful holding force of the rare earth neodymium magnets. With the three rotor machine configuration, I have connected 6 impulse coils in three phase to function as the motor impulse circuits with a hall effect brushless motor controller fig. 22.

Star wound three phase motor impulse coils.

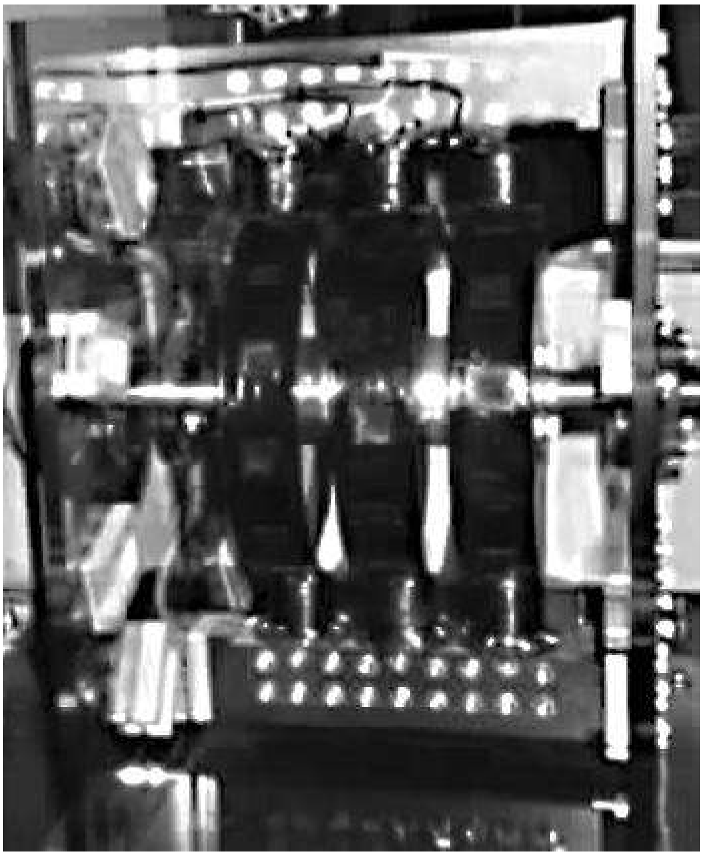

Fig. 22
Star wound
three phase motor impulse coils

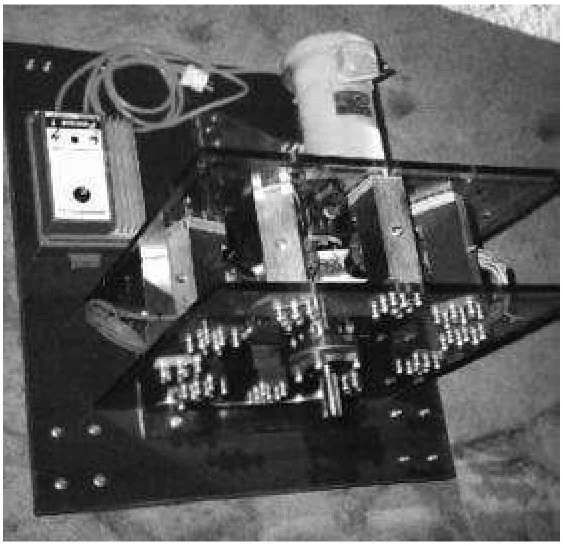

Fig. 23

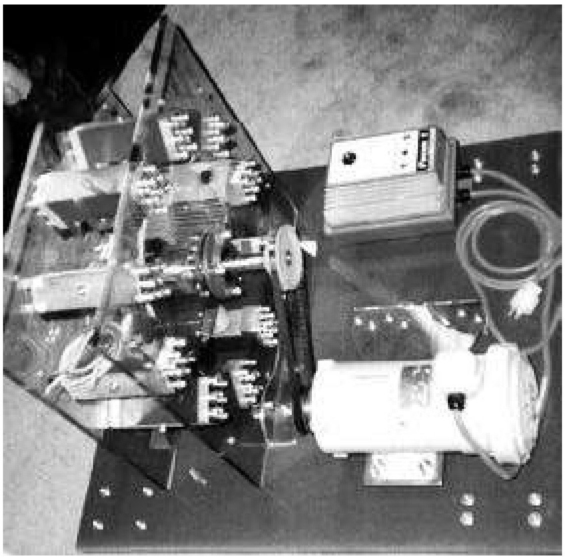

Fig. 24

Two more pictures of the one rotor dynamo option are shown with a 1hp dc motor turning the rotor with an 800 watt load on the coils, six 100 watt AC light bulbs, and two 100 watt 12 volt DC light bulbs, Fig. 25, 25a.

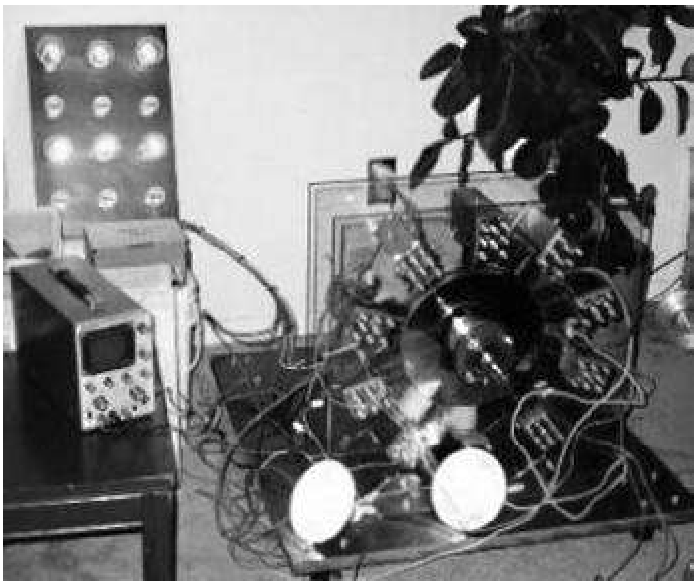

Fig. 25

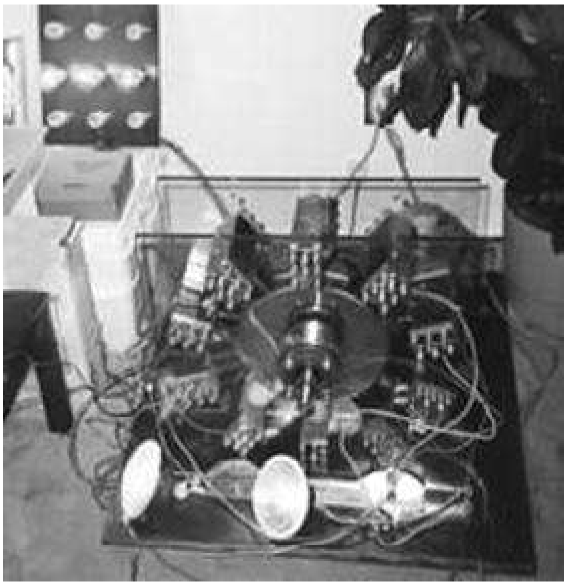

Fig. 25a

The efficiencies of this dynamo are so far inconclusive with regards to producing an endless self running machine. Our research is continuing to improve the efficiency by eliminating all conventional type ac and dc motors from this dynamo, and incorporating high voltage dc impulses of very short time duration's as Ed Gray did with his technology years ago. However, a description of my early self running attempts is included near the end of this dynamo report. These tests have manifested an abundance of heat in the 24 volt dc drive motor.

Another picture of the three-rotor dynamo configuration is shown with 12 regenerating ac to dc coils, and with the motor impulse coils removed fig. 26.

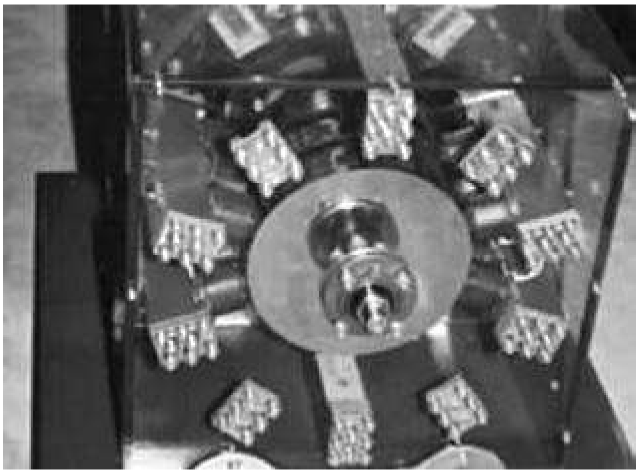

Fig. 26

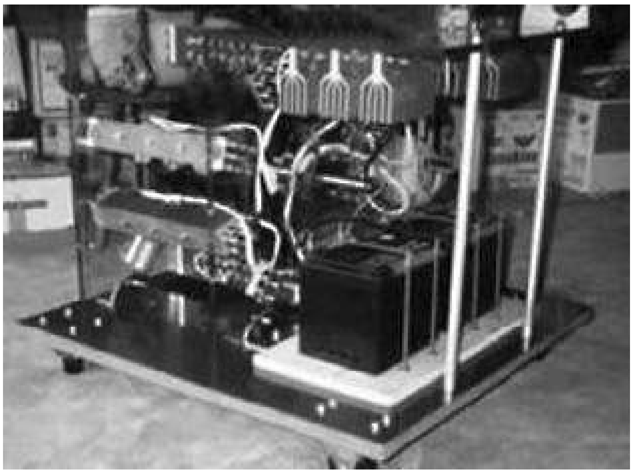

Fig. 26a

Another close up of the regenerating coils and three phase motor impulse coils is shown on the dynamo fig. 27.

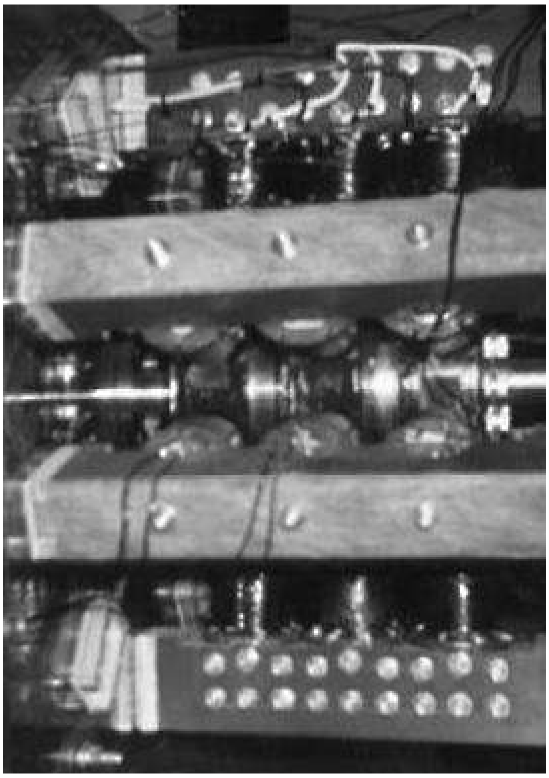

Fig. 27

Tests to produce an abundance of heat

I have testing the machine so as the output coils are connected to regenerate a 12 volt battery as the motor is turning the rotors. I used a bran new conventional permanent magnet 1/4 hp 24 volt dc motor that is running on 12 volts to turn the 110 pound 3 balanced rotors, and or the 35 pound 1 balanced rotor version.

The regenerating output coils produce AC to DC power with no noticeable ripple on the scope. The generated direct current from the coils is connected to go back into the battery, in which the motor uses to produce the torque needed to rotate the rotor. As this happens, the regenerated dc power from the coils is again going back into the battery first, and then to the motor. The result is that the battery supplies less power to the motor when the regenerating DC current is going into the battery. The motor takes on a load from two sources, one is from free wheeling the rotor with out loading or drawing current from the coils, and the other is the load induced on the motor when the coils are loaded to produce current.

The dc permanent magnet motor has shown that during one of many switch mode tests, it consumes 15 amps at 12 volts from the battery while a chosen number of regenerating coils supply ? the amount of current going back to the battery. The battery supplies 50% of the amps and the regenerating coils supply the other 50% amps. This effect can be changed depending on how many open isolated regenerating banks of coils are activated and loaded.

This test shows a 15 amp draw the motor while it is turning the rotors under load while regenerating current from the coils. It may be they balance each other and both battery and coils supply 7.5 amps each that make the total 15 amps draw on the motor. This accounts for the 15 amps total load that is shown on the dc drive motor. It produces an excessively large amount of heat in the dc motor that is useable output energy. On a larger scale the heat generated in a given motor can be enormous. More accurate measurements are needed to determine the true energy in verses energy out.

However, this electric heat pump effect can be altered to produce more or less heat depending on the size and type of the drive motor, and the amount of regenerating coils that are placed under load. Cooling tubes wound around the motor with water flowing through them is a way to pull the heat out for useful applications. Hot water from the drive motor and electric power from the coils at the same time.

After running the machine, the amorphous regenerating coils remain cool to the touch where as the conventional dc motor gets very hot with higher amps going through it. The machine also has a longer running time when the coils are regenerating the battery, as compared to running an outside equal load with the same amp draw, in that case the battery drains much faster.

This sheds light with the observation of high heat build up in the motor side (less efficient side) of the circuit while being cooler on the generator side (amorphous side) of the circuit. After hours of run time the battery begins to drain, when this happens, the battery begins to show tangible temperature changes from one side of the battery to the other. The whole negative side of the battery remains cool to the touch (temperature drop), while the whole positive side of the battery gets warmer to the touch (temperature rise). The effect is repeatable and only happens when the regenerating coils are connected to charge the battery as it is running. It may be approximately 20 degrees difference in temperature, depending on how many regenerating coils are activated sending its current to the battery. There must be exothermic reactions and endothermic reactions going on inside the battery motor generator combination that is causing this temperature difference. The load and heat generated can be changed the on the motor from the flick of a switch that activates and loads the output coils, which in turn dumps it's energy back into the battery pack or outside load. The efficiency ratio of the machine changes depending on how many and which banks of regenerating coils are activated, which again also effects the temperature difference manifesting on each side of the battery, motor, and amorphous coils.

Other tests

An Automotion dc brushless 3 phase motor controller with hall sensors was used to test the machine in fig. 26a. Tests have shown the 1hp dc motor and the three phase brushless dc motor controller will not be the answer to closing the loop on this motor generator machine. The brushless controller has a 50% duty cycle which will not manifest the electro radiant effect, and it also has a constant draw on the battery as a regular dc motor does. It is much less efficient than using very fast and short unidirectional dc impulses in the motor impulse coils to kick the rotor around. This brings the subject of Nikola Tesla's magnifying transmitter and Ed Gray's radiant energy impulse motors to light, and a possibly answer to closing the loop for a real auto-rotating machine.

Before I go further with this, I want to point out the many phase possibilities this dynamo has. If you look at the coil arrangements in the 10 coil machine, you can clearly see the coils that are in phase with each other. When a second and third rotor is installed with all the coils attached, one pair of coils on the first rotor is in phase with another pair of coils on the second rotor, and another pair of coils is in phase on the third rotor. With the ten coil to twelve magnet pole rotor assembly, the machine has five phases of isolated banks of coils. If twelve coil mounting bars are attached around the three rotors, and the three rotors are spiraled off set by 10 degrees, you will have a powerful three phase dynamo. If the rotors are not off set by 10 degrees and are lined up in phase with each other, all the coils on the machine will be in phase for a powerful single phase dynamo. The volts and or amperage can be tailored for an application by the winding of the transformer coils, and the machine produces 60hz at 600 rpm. This dynamo machine concept is also interchangeable into many other motor generator combinations, and sizes, and any volts or amps requirements, it all depends on the needs of the end user.

> *Cooling tubes wound around the motor with water flowing through them is a way to pull the heat out for useful applications. Hot water from the drive motor and electric power from the coils at the same time.*

It is possible to have super efficiencies with this system using unidirectional hv dc impulses of very short duration's at the make brake contact at the closure of the switch.

I have acquired two of Ed Gray's radiant energy impulse motors #4 and #5 to revive an almost lost technology. I am proposing to combine a special permanent magnet dynamo with less back EMF that is tailored to produce 5kv dc under 2000rpm. This dynamo will energize a 5kv dc capacitor bank and replace Gray's 12 volt to 5kv dc converter. The energized capacitor bank will discharge through the conversion tubes in the very same manner as in Gray's original circuit, only the capacitors will get energized from the high voltage dc dynamo instead of from a battery through an inverter. This version of a high voltage dc dynamo will be connected to and turned by Ed Gray's EMA motor, a starter motor will disengage after it begins running.

There are three banks of high voltage coils that are connected 120 degrees out of phase from each other, each bank represents 12 - 100 watt coils for a total of 1200 watts per phase. All isolated coils in each phase are wired in series to produce the high voltage dc potential to energize the capacitor bank. The isolated capacitor bank will discharge their potential through its conversion tubes in 120 degree increments each time the impulse coils are at their closes point toward each other. One isolated capacitor bank will discharge through the conversion tube with it disconnected from the dc dynamo circuit. The second isolated capacitor is at this point beginning to receive its charge from the dc dynamo, while the third capacitor phase is fully energized just before the switching circuit disconnects it from the dc dynamo for it discharge phase. This fully aligns all the isolated 3 coil phases in this permanent magnet dynamo system to the 120 degree three phase radiant energy impulse system of Ed Gray's EMA motor #5.

The plan is to eliminate all the batteries and converter and use a rotating high voltage dc dynamo in its place. Nikola Tesla used high voltage dc dynamos with capacitors as part of his magnifying transmitter to produce radiant electricity. So it seems logical to do the same thing, merging Ed Gray's radiant energy motor and our version of a high voltage dc permanent magnet dynamo. Could it regenerate a car, or power a home, and or a factory?

Shown below are pictures of Ed Gray's motors in fig 28, 29.

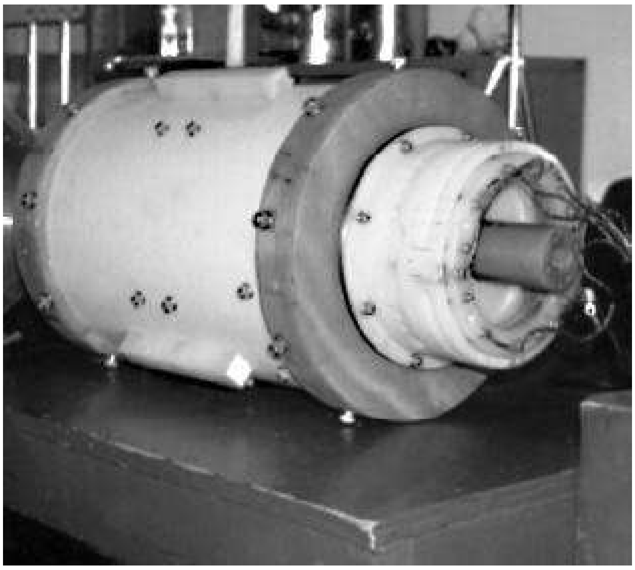

Fig. 28

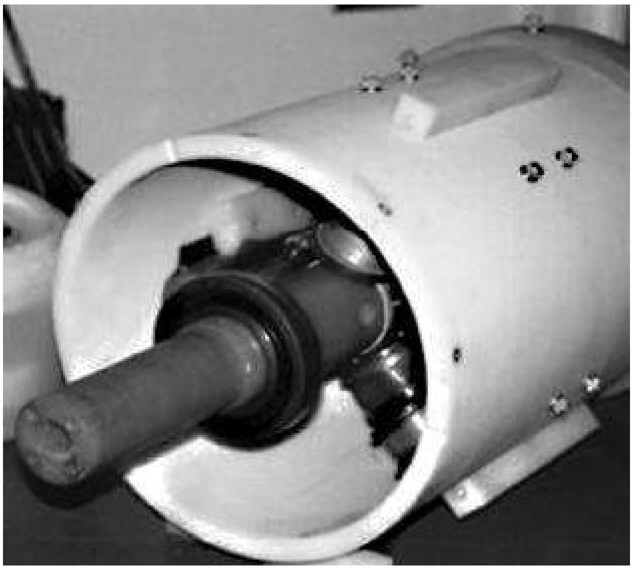

Fig. 29

I would like to acknowledge two friends of mine who have helped make the Interference Disc machine possible, Bud Johnson 1986 shown in fig. 30 and the late Dr. Carl Reich shown 1987 in fig. 31. I would also like to acknowledge Wilbert Griffith and Brian & Rose Langan for their contribution in helping us make these projects possible, and there are others who wish to remain anonymous. Bud Johnson also an inventor, was one of the first people who I met after relocating to Alberta from the NWT, he made arrangements for us to construct the first disc machine in a machine shop.

Fig. 30
Bud Johnson

Fig.31
Dr. Carl Reich

Dr. Carl Reich was a pioneer in the alternate medical field and he published a book along with Robert R. Barefoot called "The Calcium Factor", copyright 1992. Through Carl's dedication and hard work with alternate health and energy, Dr. Carl Reich contributions have also helped made the Interference Disc machine and our Permanent Magnet Dynamo possible.

I also would like to thank my wife Jan Francoeur for being part of making this research possible, she has spent countless hours to prepare this report and assist in the reconstruction of these machines.

The test results and hypothesis I have written are not complete, and the research and development of the described magnetic energy machines is on going, and as such, the potential and performance and true efficiencies of these machines are still yet to be determined. This author knows that the aether is real and free energy systems are possible, and I will continue to work until I have completed the integration of Ed Gray's radiant energy motors with our larger version dc dynamo machine. When this happens, a new round of testing will begin with our attempts to run the two machines together in closed loop regenerating mode.

I hope I explained these energy machines in a way that you all can understand, and more information will be released as progress is made. ***You may find this information useful with your quest to be energy independent. However, if you profit from any of my ideas I have described in this writing, then in all fairness, please remember to reward the inventor so research and development can continue.***

This is my gift to humanity.

Interesting Information

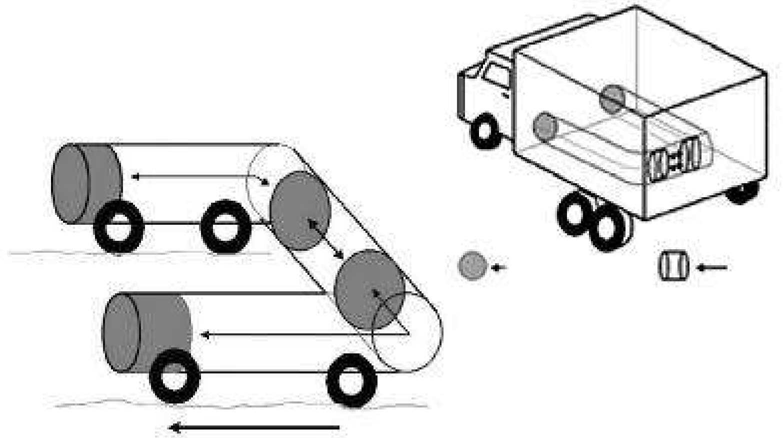

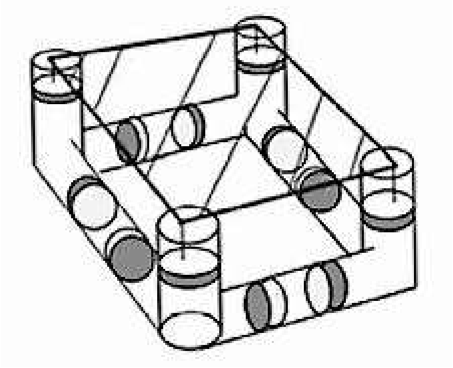

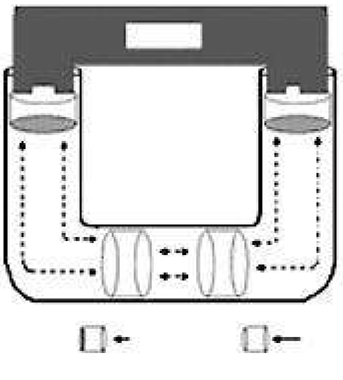

New Propulsion Letter
from Paul Toomer <jorjencar@leeds322.fsnet.co.uk>

Please take a look at this new web site on how to challenge Newton's third law of motion at

http://magnetic.propulsion.mysite.freeserve.com

A NEW PARADIGM FOR TIME

EVIDENCE FROM EMPIRICAL AND ESOTERIC SOURCES

Donald Reed, USA
1032 Borden Road, Depew, NY 14043
Email: torsionpower@yahoo.com

The author of this article
presents an experimental approach
to the teleportation problem
and time pace control
and
space regularity questions.

Part I of II

ABSTRACT

The following essay seeks to establish awareness, through a far-ranging careful examination of various empirically documented anomalous research results in the field of new energy, the long-suspected evidence provided by associated key legendary secret government project work in the USA and former USSR, maverick new theoretical models in foundational physics for elementary particles/fundamental electromagnetic wave-field structures, as well as remarkably corroborative related information from esoteric (psychically channeled) sources, of the necessity for the development of a new paradigm for mass, energy and especially time. After having been guided through this process, the reader will hopefully be motivated to acknowledge the importance of this imperative for a new understanding of the workings of nature, as well as gain hints for the associated future development of new viable sustainable energy sources and related technologies.

Introduction

When we come to examine the annals of physics over the past century, we find them replete with several competing core theories of the physical world, each attempting to demonstrate a unified conception of space, time, energy and matter. Notable among these, the most successful and formally canonized in academia, are special and general relativity, and quantum theory, the latter of which includes quantum electrodynamics and quantum chromodynamics as recent offshoots.

However, even these fundamental theories have fallen short in achieving this coveted goal, due to many reasons, not the least of which is their failure to account for the anomalous but substantial documented evidence continually presented over the years by new energy research and other related empirical evidence considered outside mainstream science. In this author's opinion, these weaknesses in current physics and its relative ignorance and/or selective omission of the findings of non-orthodox scientific research, stem primarily from an ill-conceived institutionalized conception of **time**, as an immutable linear flow against which everything involving **change** can be measured. This "relational" concept of time treats it as a specific passive property of physical systems and changes happening to them. It is a one-dimensional continuous and homogeneous entity geometrically describing the property of duration. This antiquated relational view of time is abstracted from our unique prejudiced viewpoint as sentient beings whose particular point of waking conscious focus is this camouflage physical reality, which is apprehended through the limitations circumscribed by the physical senses.

Accordingly, modern physics is built on the basis of this relational conception of time. However, the use of this conception has not so far resulted in resolving all the problems associated with time. Moreover, so far even an essential definition of time has not been formulated in physics, there are only operational definitions indicating different methods of measuring time intervals.

In contrast, the "substantial" conception of time, advanced in this paper, implies that time is an independent phenomenon of nature existing side by side with matter and physical fields, whose active essence can and does affect objects and processes occurring in the universe. Moreover, the following dissertation argues, by positing that active substantial **change** is the basic concept and relational **time** is its derivative, that those objects and processes can also have a reverse action on time.

Accordingly, we will attempt to show, through a wide-ranging examination of new energy research and other information, the necessity for the establishment of a fluidic, elastic, field nature for true active time which is malleable; one in which even the local pace of time, and mass and energy content can be influenced artificially by intelligent technological control, or naturally by conscious intent, or by a combination of these two methods. To help motivate the reader's awareness for this imperative, we will also examine several esoterically-based sources from key selected psychically channeled transcripts. These will be taken from the following sources: unpublished comments by Jane Roberts on an altered state of consciousness[1] (indicated in this manuscript by J.R.), or published transcripts from **The Seth Material** and The **"Unknown" Reality** (indicated in the manuscript by UR)[2], published chronicles from the entity Kryon, channeled by Lee Carroll [3], and concepts from Wilbert Smith's legendary discourse: **The New Science** (NS) [4]. For easier reference, all psychically channeled transcripts will appear in italics, with my additional explanatory comments in regular type within parentheses.

Granted, our unique investigative foray, ranging as it does from the concrete arena of knowledge represented by the cutting edge of visionary science to the intangible arcane realms bordering on the spiritual, will of necessity lack the exacting logic of scientific rigor.

Nevertheless, we hope to provide, through just such a unique eclectic format, the beginnings of a possible fresh understanding of the workings of nature and perhaps ultimately furnish a conceptual basis for extending the structure of current physical theory to compatibly encompass the elements of a unified framework of physics and metaphysics.

As "unscientific" as this proposed venture may appear, especially to the contemporary physicist or theoretician, we unabashedly press on, secure in the belief that the currently perceived "mutually exclusive" bodies of knowledge defined through psychic means and that of orthodox physics are more closely linked than is currently suspected. Indeed, it will be demonstrated that that some of the psychically defined data bears a striking resemblance to the tenets of present empirical knowledge.

New Research

Theories Indicate Necessity for Novel Time Concept

A body of work which postulates a fluid-field substantial nature for time, is the esoterically-based book written by Wilbert Smith in the early sixties, **The New Science**. Here, Smith outlined a unified theory of all physical interaction by positing that an active mass-free field energy he termed the *tempic field* exists, and is the parent field structure out of which our passive relational linear clock time emerges. To be sure, this book and its specific format of exposition has its weaknesses, not the least of which is a writing style which taxes one's comprehension to the limit. All researchers who have attempted to connect Smith's knowledge with the frontiers of current scientific knowledge, new energy research, etc., have been frustrated by his introduction of terms which may or not have the same meaning as the corresponding terms in mathematical science. This practice causes many passages to appear so vague as to cause total bewilderment on the part of the reader.

> *...the establishment of a fluidic, elastic, field nature for true active time which is malleable; one in which even the local pace of time, and mass and energy content can be influenced artificially by intelligent technological control, or naturally by conscious intent...*

In coming to investigate the source of Smith's theory 25 years ago, I came into contact with Kenneth Killick of Canada. Killick was the individual who served as the original mentor and for a time a colleague of Smith in the mid-1950's. Smith, whose background was in electrical engineering, found it very difficult to accept or understand the philosophical overtones of Killick's thought. Consequently, he resorted to the use of psychic mediums to try to verify the information provided to him by Killick. This is the origin of Smith's association with discarnate entities known to those familiar with his work as "the boys topside". Thus Smith's subsequent writings promote such confusion perhaps because he only obtained a partial understanding of these cosmic ideas, and he opted to writing in a kind of code basically to try to hide his own confusion. Through my own association with Ken, I discovered that The New Science is able to be put into reference with known facts once the spiritual-philosophical elements of Killick's teachings are duly integrated. When this is done, "Wib" Smith's book can teach much and can become a cornerstone to new energy science. Otherwise, it remains at best, a fragmentary enigma.

The work by Smith and Killick is pertinent to the theme of the present paper, since they both showed through experimentation with a special caduceus-wound coil, that so-called relativistic parameters of mass and gravity, energy, and time-flow, can be altered through intelligent artificial control. Adequate delving into some of the elements of these theories will enable us to glean new insight for transforming the current paradigm of time. Further details on Ken's so-called tachion energy theory can be found in this author's earlier expositions on this subject in issues of **Energy Unlimited** (1978-1982)[5].

Smith's main error is in postulating a fundamental "spin" dynamics based upon conventional continuous rotation. As Killick points out, such a concept of spin on the microscopic level, precludes any possibility of true evolutionary change, in the cosmic sense.

Also, as we shall see presently, the continuous spin concept obviates expression by anything in the universe, of its personal moral responsibility. However, with tachion energy theory intact and untrammeled, the impersonal dualistic concept of interaction between two polar entities, such as is expressed not only in ordinary classical Newtonian physics, but its 20th century successors, relativity and quantum theory, is supplanted with the omnipresent workings of a **trinity** at all levels and manifestations of reality.

Indeed, the fundamental tachion field (mass-free) energy has **three** components: two polar opposite entities (a positive "charge" and a negative "charge"), and the ability for these two to be in what is called "static-dynamic balance". The static-dynamic balance completes the trinity.

A simple physical macroscopic manifestation of static-dynamic balance can be seen in two people arm wrestling. As the energy of one participant is brought to bear against that of the other, we will see a vibration or oscillation once a balance of their energies is obtained. As more effort is expended, the vibrations will increase in frequency and decrease in amplitude. At this point we will see a static state relative to the two arms; neither causes loss of arm position of the other. But simultaneously we also have a dynamic situation in the rapid oscillation which maintains, and is in turn sustained by the muscular energy which produces the static state of the arms. Both states mutually support each other's existence. The two component entities (the arms) resonate together, but without losing their own identity.

The implications of this last statement, missing in current physical theory, is unique to the system of tachion energy. We will see in the following that the dynamic functioning of these sub-atomic tachions is such that by their inherent ability to manipulate space and the time-frame of matter through static-dynamic balance, they do not harm anything else in their environment, nor are they affected (forced to lose their own identity) by the ambient environment. Accordingly, inertial mass is not an inherent component of tachion energy, but is a derivative of certain tachion field modes.

We spoke earlier of tachion "charge". We keep the word in quotes to underscore that this is not to be equated to electrical charge, but merely signifies two polar energy states of the primordial ground-form, out of which all known physical forces come to be manifest. These tachion energy states can be defined in terms of "rotational" motion. Again "rotation" must also be placed in quotes as it does not signify the common idea of continuous spinning. Such a concept was suggested above in connection with evolutionary progress in its cosmic interpretation.

The tachion-pair could be described as executing a "back-and-forth" ratcheting movement. The tachion-pairs are analogous to the ends of a drum majorette's baton. But unlike the twirling motion of the baton, the tachion-pairs do not demonstrate the classic macroscopic continuous spin motion, but oscillate in a 3-component "clocking" action.

In Fig. 1 we have drawn a circle and have divided it into three 120 degree arcs. The first motion in the tachion cycle is an arc from the 0^0 -360^0 location, to the 240 degree position. This movement is stated philosophically as the question, "Can I go?" In the second phase it swings back from the 240 degree position to the 120 degree point, and then returns to the 240 degree position. This graphically represents an analysis of the first question, and can be phrased as the "Let me think" component. Finally, then it proceeds forward another 240 degrees taking it past the 360 degree point to 120 degrees. This last stage is the action motivated from the analysis in stage two, and is entitled, "Yes I can". In this activity it will be observed that all path lengths are 240 degrees.

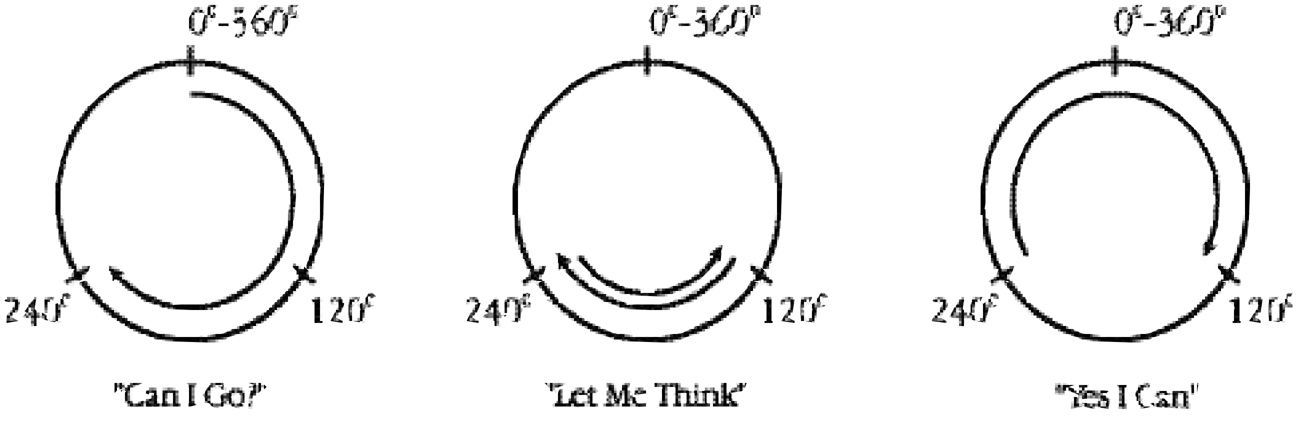

"Can I Go?"

"Let Me Think"

"Yes I Can"

Fig.1
The 3-Component Tachion Movement

However, as representative of true evolutionary progress, the tachion-pair does not remain confined to the plane in its 3-phase cycle. After each "Let me think" stage, the entire pair unit could be considered to advance "upward", perpendicular to its plane of "rotation". The combined motions can be viewed as an elevation up an inclined ramp, screw-fashion (Fig. 2).

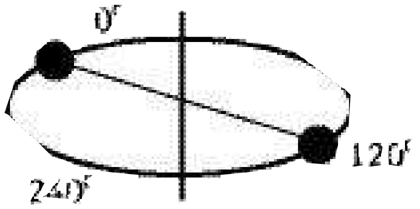

d) 240° Rotation, "Yes, I Can"

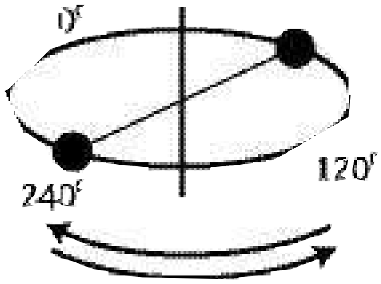

c) 120° & 120° Forward.
"Let Me Think"

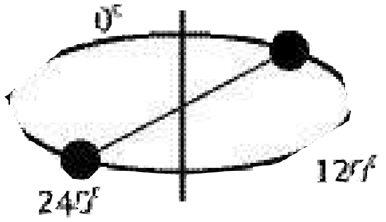

b) 240° Rotation, "Can I Go?"

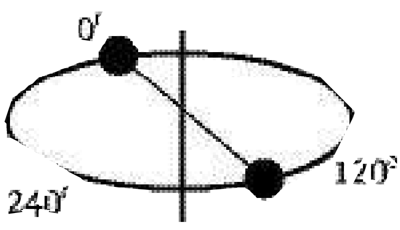

a) Initial Position

Fig.2
Tachion "Ratcheting"

Thus, while they clock themselves backwards in the plane, they never go back to where they were originally but only seem to do so from the point of view of an observer in the plane. This entire activity can be visualized as a diametrically opposed pair "rotating" in a plane, but there must still be considered an upward ratcheting movement. They execute the clocking motion simply because they adhere to the three principles of observation, analysis and motivation. A meaningful application of this idea to the structure of physics would eliminate the necessity for a totally impersonal type of interaction governed solely by force. Unlike all current theories of elementary particles, the tachion-pair, by its unique activity, thus expresses a personal moral responsibility. So this elemental energy of the universe never goes where it is not wanted, but only where it will not disturb its immediate environment.

The topological structure of each tachion in the pair is also significant in regards to this "ethical" modus operandi. First, tachions in their primordial state are massless, toroidal shaped fields which always occur in pairs. Like toroids, tachions will singly exhibit three motional degrees of freedom, to wit;

1. rotation around its major axis.
2. inner (P), or outer (N) rotation about its cross-sectional (poloidal) axis.
3. expansion and contraction of the field (each toroid pulsates radially about its cross-sectional axis).

As a pair-unit, the combined motions will give rise to the following possible orientations: two identical (N) rotations; (N) rotation - (P) rotation; two identical (P) rotations (Fig. 3).

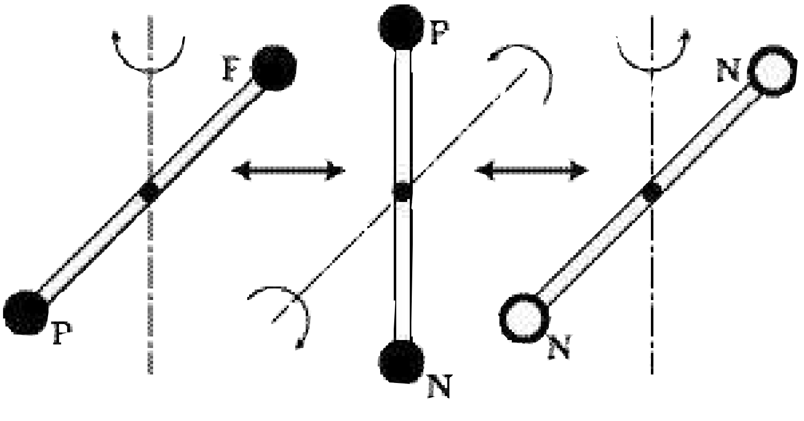

Fig.3
The Three Tachion Modes

The dual (P) or (N) rotations will always lie in the same plane, whereas in the "mixed" rotation mode, the orientations of the tachions will be mutually perpendicular. In the pair unit, the energy will shuttle back and forth between the tachions by virtue of their harmonious expansion and contraction. The contraction of one pair partner is immediately compensated for by an equal amount of expansion in the other. When one is fully expanded and the other completely contracted, they will reverse roles executing the second portion of the cycle. There will be two points in the clocking cycle where the tachions will share an equal amount of energy ("field equity") stage. In each cycle of a tachion-pair the field equity stage is passed twice, a fact which is significant when topology of the field dynamics is considered (see [5]). Now, looking at the field geometry for the tachion-pair unit at the two field equity stages of the cycle, we observe that the two geometries are not identical but are *mirror images* of one another via a non-orientable field structure. See Fig. 5, which shows the isomorphism between the dynamics on a Moebius band (a non-orientable topological structure) and the tachion pair clocking cycle.

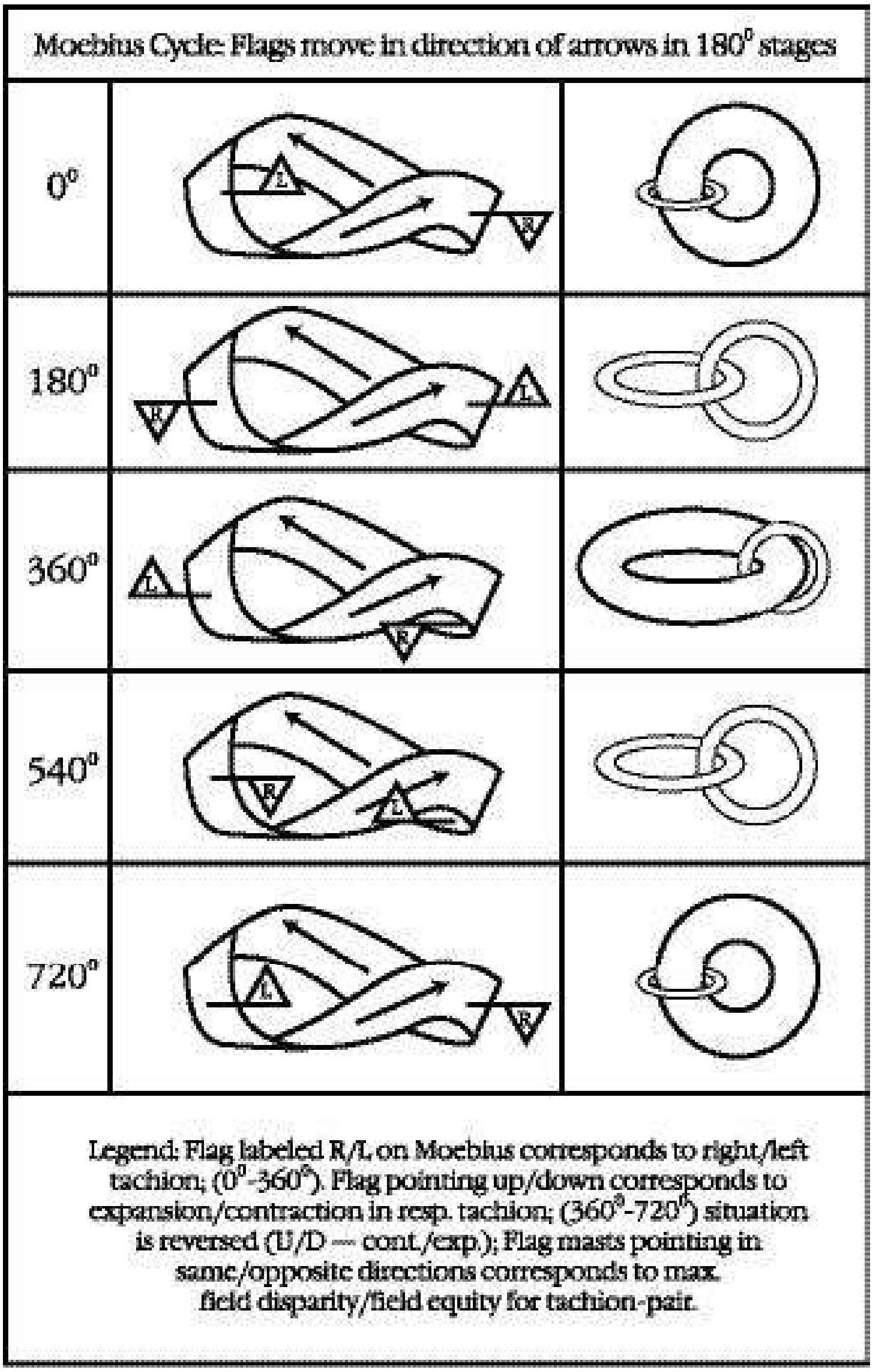

Moebius Cycle: Flags move in direction of arrows in 180° stages

Legend: Flag labeled R/L on Moebius corresponds to right/left tachion; (0^0-360^0). Flag pointing up/down corresponds to expansion/contraction in resp. tachion; (360^0-720^0) situation is reversed (U/D — cont./exp.); Flag masts pointing in same/opposite directions corresponds to max. field disparity/field equity for tachion-pair.

Fig.5

This suggests the remarkable idea that the space time points that the pairs generate are not, as commonly assumed, identical, static, and bereft of intrinsic character, but exhibit *enantiomorphic* topologies, which by virtue of their mirrored relationship exert upon one another a potential for dynamic interaction. For now we present a second diagram to illustrate the three pair energy modes: N-N, N-P, P-P, as well as one sequence of oscillation of the N-P pair (Fig. 4).

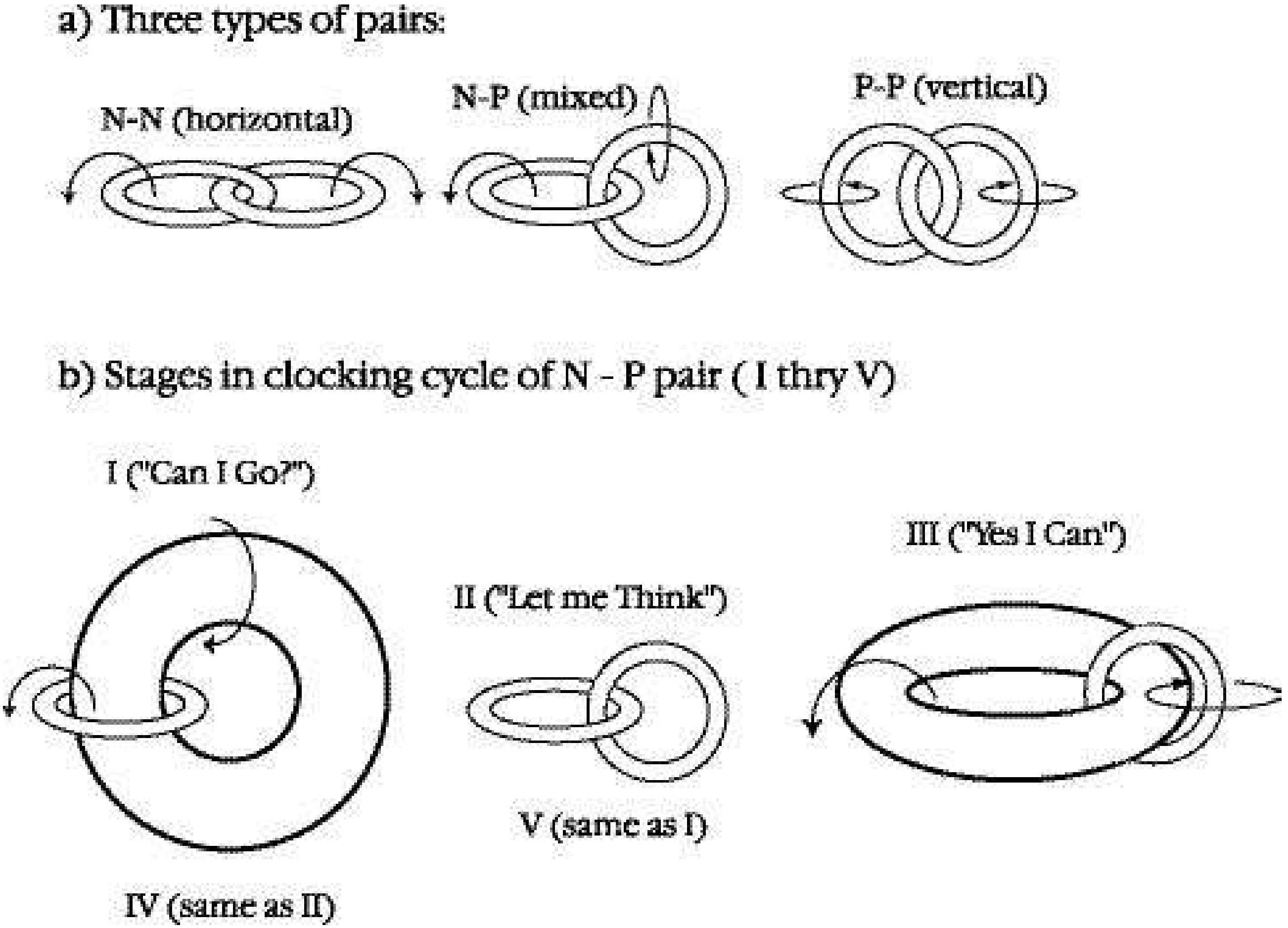

Fig.4

In this figure we note a unique trinity of field operation: two stages of maximum field disparity (I and III) connected by the intermediate condition of field equity (II). Upon postulating the additional requirement that the pair will create a new space-time location every time the field equity state is reached, we see that with each cycle the unit will not return to its original state with respect to the environment, but will advance to new evolutionary states with each pulsation. In accordance with the above comments, the tachion-pair unit "ascends" the evolutionary spiral via this unique action we have termed "clocking" motion. In acccordance with the description in Fig. 4A, in Fig. 4B, stage I represents "Can I go?", stage II represents "Let me think", and stage III represents "Yes I can".

Now, in this regard, the previous analogy of the tachion-pair as likened to the ends of a baton, where the baton rises and falls as it rotates, is quite apt. However, unlike the ends of the baton tachions are not rigidly connected but possess much freedom of movement outlined above. On the other hand, much like the baton ends, they are mutually supportive of each other's activity and requirements, never seeking to overpower, or to be eclipsed by the identity of their partner. In this manner they exhibit personal moral responsibility, as we have previously stated. By their mutual harmonious action tachion-pairs will progress in a true evolutionary fashion, while never harming anything else in the universe by their action.

> *...the dynamic functioning of these sub-atomic tachions is such that by their inherent ability to manipulate space and the time-frame of matter through static-dynamic balance, they do not harm anything else in their environment...*

Now, the above is an inherent characteristic of tachion-pairs as long as they go about their business individually- that is, in their pre-physical state. Now, when the three fundamental tachion-pair energy modes described above exhibit a slow clocking frequency, there is opportunity for them to join together whereupon we begin to see matter being formed. The element of mass, or inertia is provided by the binding activity of the (N-P) mode, which is a function of the slow oscillation of the pair unit. Besides the clocking frequency, the factor of inertial mass or density of physical substance is also dependent upon the amount of separation between the two pair components in each individual sub-atomic tachion-pair unit. In general, the denser or more massive an object is, the slower its microscopic tachions will clock, and the farther apart will be the individual toroid components. This is a classic push-pull situation, i.e., there is a key inverse static-dynamic balance relationship between tachion-pair pulsation frequency and their area of operation (density of field). The rest mass that any physical substance displays to macroscopic observation, is thus a direct function of the amount of energy supplied by the master field to sustain the oscillation of the pair-unit. Thus, a primary feature of tachion dynamics is the following: if we change the energy, we must expect to observe a corresponding change in the manifestation of the mass, and vice versa. Consequently, in tachion dynamics there is a built-in interchangeability of energy and mass, an idea which concurs with similar conclusions in Einsteinean relativity (notably, $E = mc^2$).

But here is where tachion theory and relativity part company. We have seen that our postulated "clocking" pulsation of the pair is the factor, which determines both space and time. Both **space** (area of pair operation), and **time** (frequency of pulsation) are mutually dependent on what **we shall call the master field of thought and intention, which determines the clocking frequency**. From the dynamic influence of the master field arises subsidiary conditions of time and space which are truly "fluid" qualities of tachion operation. This is indeed consonant with the assumptions of a proper unified field theory in which all the factors of our objective world as apprehended by the physical senses, are subsumed under the relative state of a guiding master field. Moreover, under the geometrical constructs of such a theory neither space nor time should be subservient to the other but should retain their unique qualities in a mutually supportive relationship. This is guaranteed in tachion theory under the fluid give-and-take of the pair operation. Relativity, on the other hand, despite its logical consistency, does not meet the above requirements. To his great credit Einstein recognized the need to integrate time with space in a synthesis that would retain the distinctive qualities of both. The result was a theory, which extended the classical Euclidean concept of 3-dimensional space as a kind of "cosmic container" of matter, to embrace a structure of space, which arises from the laws interrelating its material content. Thus, for instance, in his General Relativity theory gravitation became interpreted as a distortion of the spatial fabric by sources of mass-energy in a non-Euclidean framework. Unfortunately, as a foundation of his theory, Einstein mistakenly assumed a subservient role of time to space. He tried to "freeze" time down into spatial-like dimensions utilizing metric equations, which were variants of a basic form. Before a true understanding of gravitation is developed, the rationale for applying such metrics to this end should be seriously re-appraised. In short, instead of freezing time down into a spatial dimension in order to make it fit a world ruled by nothing but gravity, we must develop a conception of space sufficiently fluid to let true time have its proper place therein.

These and other similar ideas are articulated by Andrea and Paulo Correa in their recent insightful essay on the various serious shortcomings of relativity theory. They write: "Time is not treated by relativity as distinct in nature or in quality from the set of topological locations or lengths between points. Whether in the SR (Special Relativity) form of a flat Minkowski spacetime or in the GR (General Relativity) form of a curved Riemannian spacetime, the essence in relativity lies in its treatment of time as the fourth length of a 4-dimensional space. As the actual metric of a length of time is not fixed, only the intervals being invariant, one can no longer speak of the reality of timelines or of synchronicity" [6].

In their second installment, the Correas even imply the existence of a mass-free energy, which is an integral element of the Killick tachion energy theory: "From a strict physical viewpoint, only an energeticist position can make sense. While space and time may be considered to exist outside the function of matter, independently from it, *they cannot be conceived outside the function of energy*...the field remains conceptualizable (in Einsteinean relativity) only by the pseudo-Riemannian manifold, which, on its own, fails to analytically treat the difference in dimensionality between Space and Time, fails to differentiate between them as distinct manifolds, and fails to account for them as the intrinsic properties of energy in flux".

The tachion theory is such a model that delineates the functions of time and space in terms of a basic mass-free energy substratum. Now as we have seen, since tachion dynamics pre-supposes a static-dynamic balance feature connecting frequency of the pairs (fundamental time elements) with their area of operation ("fluid" space configurations), it represents such a foundation upon which to construct such a theory which will harmoniously treat time and space on an equal footing without neutralizing the unique structural features of each. Since the decrease/increase in the clocking frequency of the tachion-pairs will result in a corresponding increase/decrease of area of operation, the local *pace of time* will also decrease/increase at that space-time location. This is due to the push-pull relationship between energy and time. The intrinsic energy of a tachion-pair will be defined to be its energy of "motion" (energy in the "convergent", pre-physical state). Thus we see that an increase/decrease in the clocking frequency also implies that the intrinsic energy of the pair will also follow suit. Consequently, the intrinsic energy available for a manifestation will increase/decrease as time pace allotted for this purpose decreases/increases. Moreover, the product of intrinsic energy and time interval will always be constant due to the perfect trade-off between both factors in the tachion clocking action. This relationship is remarkably similar to the quantitative statement of the Uncertainty Principle in quantum mechanics in the form: $E \cdot t = h$. This formula states that the product of the energy exchange of a quantum mechanical particle and the time interval required for that exchange to take place, is equal to the so-called "quantum of action", or Planck's constant.

Thus, as a concomitant of tachion dynamics, which is argued could also be the basis of electron spin geometry among other things (see [5]), *time* is made malleable; no longer continuous, or rigidly uniform as in Newtonian conceptions, or subservient to space as in Einsteinean relativity, time "flow" can go to zero, and can even reverse under the application of certain electrodynamic field configurations such as provided by the caduceus coil and other devices [5]. With tachion dynamics even the laws of causality will no longer hold. This is due to the fact that tachions have a much broader field of operation than the purely physical realm. Thus, many unusual demonstrated features relative to quantum theory, such as the principle of Indeterminacy, quantum nonlocality (EPR experiment), "entangled particles" as well as hypothesized qualities of an *aether* of space, can find their source in this microscopic non-linear electrodynamic clocking oscillation attributed to the tachions. In fact, the ultimate expression of time as well as space to our sensory perception, could be a direct function of these substratum pulsations. From this consideration there emerges a new dynamic model for **fundamental time** as a "fluid-elastic" field effect arising from the unique ratcheting oscillatory motion of tachion-pair entities. Ultimately, it will be seen that tachions may even form the structural essence of consciousness itself, thus providing the long sought-for unifying link between physics and metaphysics.

N. Kozyrev

Other Recent Russian Research Implying New Dimensions for Time

Another key researcher who postulated an active substantial role for time was the Russian astrophysicist Nikolai Kozyrev [7]. He argued that the gap between the exact (physical) sciences and the natural sciences will not vanish until the principle of causality differentiating cause and effect is placed at the basis of the exact sciences. This implies the objective directionality of time or its flow. We encounter numerous manifestations of this principle in our life and in natural science. Yet this principle is new in both mechanics and physics as a whole. The natural scientist's time is not Minkowski's geometrical time, nor is it the time of mechanics, theoretical physics, etc.

> *Nikolai Kozyrev argued that the gap between the exact (physical) sciences and the natural sciences will not vanish until the principle of causality differentiating cause and effect is placed at the basis of the exact sciences*

As Kozyrev argued, mechanics has been using only the "geometrical" property of time, its duration size, i.e., the intervals between the events under study. Timed, these intervals acquire the same passive properties as the interval measured between two points in space. It is only this property of time that the objective sciences assumed to be existing objectively, postulating that all other time properties are subjective. However, if we want to differentiate causes and effects, in reality the *flow of time* must be a physical quantity, duly expressed mathematically and entering the equations of mechanics. The physical meaning and mathematical expression of the flow of time can be obtained from the space-time properties of causality. Thus, processes in the world occur not only in time, but also *with the aid* of time. Time flow is an active property owing to which time may act on a material system. Consequently, it is in the properties of time that the source maintaining the vital processes of the world must be sought.

From these arguments, we can also derive a fundamental property of time. Let us assume that we have changed by a certain technique the flow of time in a material system. In this process we have perhaps changed the stresses within the system and hence its energy.

However, it is impossible to change the total amount of motion of the system, i.e., to obtain a momentum equivalent to an external action. In other words, time may carry energy but no momentum. Time is a non-momentum material reality. From the above property of the flow of time, Kozyrev concluded that it follows directly that the rotation of a body may be used as a device for changing the flow of time in a system. Time flows into the system in the cause-effect direction, the rotation increases the flow, and hence the system obtains additional energy.

In attempt to verify his theories, in the 1950s, Kozyrev thus performed experiments with spinning gyroscopes and discovered, in agreement with his predictions, that the weight of the spinning system changes slightly depending upon the magnitude of the angular velocity and the direction of rotation. Although the effect detected was not large, the nature of the arising forces could not be explained by existing theories. By applying a vibrating framework to the gyroscope, Kozyrev also found that not only does time deliver an energy, but that near the cause-effect system *time density* changes. Time is rarefied near the generator and becomes denser near the receiver. In other words, time becomes stretched by cause and compressed by effect. In addition, the variation in time intensity is in inverse ratio to the first degree of the distance. This can be seen by the following consideration: Time is expressed by rotation and hence involves planes passing through a pole of the rotating system with any orientation in space. **In the case of lines of force issuing from a pole, their density decreases in inverse ratio to the square of the distance, while the density of planes can readily be shown to decrease with the first degree of the distance.**

Kozyrev discovered another property of time when investigating irreversible processes such as crystallization of a substance, condensation, or evaporation. This influence can be effected by any irreversible process and is registered by a measuring system without any preliminary excitation, i.e., when the cause and effect positions are not separated. In other words, apart from flow and density, time seems to have another property, causal orientation, resembling in a way the polarization of light [8]. This property proves to be so variable that even the sign of the observed influences may change and hence the reciprocal cause and effect positions become reversed.

Continuing his studies on the properties of time, Kozyrev (with V.V. Nasonov) conducted astronomical observations using a receiving system of a new type. When the telescope was directed at a certain star, the special detector positioned within the telescope registered the incoming signal even if the main mirror was shielded by metal screens. This fact implied that electromagnetic waves (light) had some component that could not be shielded by metal screens. When the telescope was directed, not at the visible but at the true position of the star, the detector registered a signal that was much stronger. The registration of the true position of different stars could only be interpreted as the detection of star radiation that had velocities billions of times the velocity of light. It was concluded that the radiation from stellar bodies hence had a component, which was non-electromagnetic. Kozyrev also found that the detector measured an incoming signal when the telescope was directed at a position symmetrical to the visible portion of the star relative to its true position. This fact was interpreted as a detection of the future positions of stars [9].

From the 1960's onward, additional Russian research groups continued and expanded upon the work began by Kozyrev on the unique non-electromagnetic radiation that was apparently a concomitant of all rotating bodies, whether accompanying macroscopic spinning objects of great mass (cosmological or terrestrial), or at the quantum level exhibited by elementary particles possessing recorded anomalous spin features which were not accountable by standard theories. This emanation which was exhibited by rotating bodies was termed torsion radiation [10]. Subsequently it was learned that this unique radiation was a part of all living or inert substances. This so-called "torsion field" research was carried on mainly in secret in the former Soviet Union, and only became known to the Western nations after the fall of the Iron Curtain. Many effects attributable to alterations in time and gravity have been reported through this research. For more information on this topic the reader is directed to this author's review in reference 10 and references therein.

Part II of II

Summary of Part I

It is the intention of this multi-part dissertation to motivate an imperative for considering the eventual establishment of a more expansive paradigm for the phenomenon of time, one which views time from the "substantial" as opposed to the orthodox "relational" standpoint. The substantial concept of time implies the existence of an **active** essence which can and does affect objects and processes occurring in the universe. Furthermore, approaching time from such a fundamental basis implies that by positing active substantial change as the ground form **prime mover** and relational time (**passive** "clock-time", or entropy changes) as its derivative, that those objects and processes can also have a **reverse** action on time.

Towards this end in Part I (hereinafter referred to as I), we considered the key research of the late astrophysicist N.A. Kozyrev who, in carefully controlled experiments which were subsequently successfully replicated, recorded phenomena which have hitherto been unexplainable by standard contemporary received knowledge- whether from the standpoints of Newtonian, Relativistic or Quantum physics. Kozyrev pointed out that only by considering time as an active essence, could these cited "anomalous" phenomena, which are virtual conundrums from the viewpoints of a relational (passive) concept of time, be duly formulated on a logical basis. Accordingly Kozyrev's associated "causal mechanics" incorporates an objective directionality of time flow, implying differentiation between "cause" and "effect". **Consequently, by applying a thermodynamically irreversible process to a physical system (example: vibration to a spinning gyroscope), the rotor in this instance would be observed to lose (or gain) a small but detectable percentage of its weight, depending upon the direction of rotation.** Moreover, in this process **linear momentum** (linear speed of rotation) is not changed, but energy is nevertheless delivered to the system by additional forces directed along the axis causing a change in **angular momentum**, and that near the cause-effect system of the irreversible process, **time density** is altered - rarefied near the generator and concentrated in proximity to the receiver. Kozyrev concluded that time is a non-momentum material reality.

Other experiments of an astrophysical nature showed that registration of the true position or future position (as opposed to visual-present-position) of different stars via the non-electromagnetic component of star radiation, demonstrated the existence of an energy essence, which was apparently instantaneous, or had a propagation velocity of at least a billion times the velocity of light. In this regard, Kozyrev opposed the conventional astrophysical paradigm, which posits energy generated/emitted by stars is a physical consequence of thermonuclear conversion. Instead he maintained that via rotation, stars converted time essence into non-physical (or pre-physical) energy. The latter was nevertheless detectable on Earth by transduction through reflecting telescopes shielded by metal screens, of stellar radiation by a Wheatstone bridge. Hence the "impact" of time was detected by resistor electric conductivity changes (see I and references therein).

Leaving no stone unturned in our investigation, we also took up a discussion of the related Killick-Smith tachion energy model for fundamental (pre-physical) reality. Here we coined a new term- a **trinitivity** of action, which was underscored as a key feature of the operation of tachion-pair entities. Accordingly, we posit the existence of two polar topologically orthogonal toroidal fields, with the unique ability, by the mutual shuttling back-and-forth of their energies, to exist in a third auto-generated state called "static-dynamic balance". Now, since tachions are hypothesized to transfer energy instantaneously, inertial/gravitational mass is not a primordial component of tachion operation (see I). Here we see a striking similarity to the Kozyrev conception of the operation of time in interaction with a physical system - where energy changes are manifested without linear momentum alterations delivered. Likewise, the Smith book, **The New Science** (see I) treats **the hypothetical tempic field as a scalar quantity, which only has direction in relation to its distribution (or "density" change). Only through establishing what Smith termed a *tempic field* gradient does the tempic field possess vectorial nature.** Similarly, Kozyrev viewed time like space as possessing both geometrical and physical properties. The geometric property of time is duration (its scalar component), while physically the course of time is similar to vectorial field propagation and the density of time to field intensity. Although Smith and Kozyrev apparently never were aware of each other's research, the following quote from Kozyrev reveals that they both made key profound discoveries about nature which indeed bear more than a superficial similarity: "Time density is a scalar quantity, just the one observed in the above experiments. Time density decreases with increasing separation from the process creating it. Therefore, a vector property corresponding to density gradient and tractable as time obtaining, should also be observed"[7]. However, in I we also enjoined the reader to observe caution and selectivity when attempting to separate the wheat from the chaff in Smith's book.

Accordingly, the previous ideas are substantially correct, whereas his concept of "continuous spin" of tempic field energy is basically flawed. Killick replaces this outmoded counter-productive notion with a key ratcheting motion, like the escapement of a watch, which is performed by the tachion entities composing a pair-unit. Through this unique "clocking-motion", a cycle of tachion-pair operation exhibits what could only be described as an **ethical modus operandi**, passing through the three stages of observation - reflection - action. By virtue of this 3-stage process, tachion-pairs demonstrate in their individual activity and in interaction with their environment - *personal moral responsibility*.

This is indeed an unprecedented notion which implies that a type of conscious thought process can be ascribed to tachion operation. This can be related to current notions from quantum physics, relative to quantum coherence of states - such as wave-function collapse, Indeterminacy, quantum non-locality (EPR experiment), "entangled" particles, etc. This may be an indication that in such cases, whenever the Principle of Causality is breached or at least compromised, we may be witnessing the unique effects of the inherent conscious microscopic non-linear clocking oscillation of the tachions.

Along these lines, although Kozyrev did not report observing any effects of time energy, which could be considered as "value-motivated", he did report on an equally mysterious "after-effect" in many experiments, which could only be construed as signifying a **memory** process. For instance, it turned out that in experiments with a vibrating torsion balance (or pendulum), at points of support the emerging additional forces did not disappear when the vibration was stopped, but remained in the system for an appreciable time. Interestingly, **their relaxation times were found to be independent of the mass of a body but were dependent on its density**. The largest effects with maximum preservation times were observed on porous materials like brick or volcano tuff. In enumerating on this unusual preservation action Kozyrev observed: "...a body placed for a certain time near an (irreversible) process and then brought to a torsion balance, produced the same effect on it as the process itself...aluminum showed no memory. The largest memorizing effect for processes of both signs has been shown by sugar"[7]. Also, his colleague Danchakov later observed in conjunction with biological experiments that water with decreased viscosity due to having been subject to the action of a process, exerted a distant action on water that had not been under such an action, causing its viscosity to decrease as well.

In this second part of the current dissertation, we will examine other important related information gleaned principally from esoteric sources, referenced in I. These have been obtained from books by Jane Roberts: **The Seth Material**, **The "Unknown" Reality** (UR) [2], unpublished transcripts (J.R.)[1], and transcripts from the entity Kryon channeled by Lee Carroll [3]. This information is remarkably corroborative with that from Kozyrev research and the tachion energy theory. Finally, we will cap off our exposition with a brief examination of a bold new maverick physical theory of subatomic electron structure, featuring variable time/mass attributes, which could be the catalyst for bringing the cited elements from esoteric knowledge and empirical physics research (Kozyrev, etc.) out from their hitherto relative obscurity, to the forefront in our ongoing mainstream investigations in probing the cutting edge frontiers of not only physics, but those of the natural sciences as well.

Other evidence

When one compares the transcripts from various esoteric sources it's remarkable how similar is their description of nature's fundamental operations at the pre-physical level. For instance, a careful reading of the Jane Roberts' channeled transcripts from the entity Seth, reveals a remarkable close connection with the particles termed EE (electromagnetic energy) units and the above Killick teachings on tachion pair dynamics. But first a word of caution to the reader. We should resist the tendency to put into the frame of reference of our sentient physical experience, structures whose essence is fundamentally non-physical. The customary physical characteristics of duality and dichotomy of fields, and the related yardsticks of mass, length and time, for the most part, may not apply in this situation.

It would be impossible, however, to formulate a comprehensive model for the EE unit dynamics were none of the factors for gauging physical reality to be able to be applied to higher dimensional realms. Fortunately, one characteristic does survive the transition to the non-physical framework- the element of **pulsation**. Seth reports in [2] how EE units, much like tachion-pairs are built up and dynamically pulsate in response to thought patterns. First, depending upon the relative intensity of the emotional energy, they will instantly be drawn together in clumps or may just as quickly disperse. Furthermore, being of a pulsating nature, these units can expand and contract. Theoretically, there is no limit to their size or frequency of pulsation. Seth describes their operation in the following manner: **Now, there are electromagnetic structures, so to speak, that are presently beyond your instruments, units that are the basic carriers of perception. Their size varies. Several units may combine, for example. To put this as simply as possible, it is not so much that they move through space, as that they use space to move through. There is a difference**.

On this point, if our conceptions remain "earthbound" we will not perceive any difference between these actions, but once we cast our mental gaze beyond physical experience, an important distinction becomes apparent. To understand this, we need only reflect on the similar non-classical logic that is the basis for the non-orientable Moebius topology that we have seen governs the dynamics of the tachion-pair unit [5]. Tachions, whose translational movement is governed by the relative pulsation frequency, apparently enter and leave our level of reality at two times (field equity stages) during each pulsation cycle. Before leaving our plane, the tachion-pair constitute what we call a space-time point.

Upon returning to our level, due to the Moebius topology of the clocking cycle, their essence creates a new space-time point with enantiomorphic (mirror image) characteristics to the first. The resulting progress of the units in physical reality is thus manifested by a "jerkily discontinuous" movement. Summing up, since the units **are** the point they create, it can now be understood that the units **use** space (or the nature of the space-time points) to move through, rather than the case of macroscopic matter which moves continuously (or appears to) through space.

Seth continues his description of the EE unit dynamics: *Laws of attraction and repulsion are used and, in a manner of speaking, thermal qualities are involved, and this is the only hint that your scientists have received of them so far. The units charge the air through which they pass, and* *draw to them other units. The units are not stationary in the way that, say, a cell is stationary. These units have no "home". They are built up in response to emotional intensity. They are one form that emotional energy takes. They follow their own rules of attraction and repulsion. As a magnet will attract with its filaments, so these units attract their own kind and form patterns which appear to you as perception. They are electromagnetic, in your terms, following their own patterns of positive and negative charge. In this instance, like definitely attracts like. The units are just beneath the range of matter. None are identical. However, there is a structure to them. The structure is beyond the range of electromagnetic qualities as your scientists think of them. Consciousness actually produces these emanations, and they are the basis for any kind of perception, both sensory and extrasensory. These emanations can also appear as sounds and you will be able to translate them into sounds long before your scientists discover their basic meaning. One of the reasons they have not been discovered is precisely because they are so cleverly camouflaged within* ***all*** *structures.*

Being just beyond the range of matter, having a structure, but a non-physical one, and being of a pulsating nature, they can expand and contract. They combine qualities of a unit and a field, in other words. Since they are beyond the range of matter, upon which matter is formed, they will not follow the laws of matter, although at times they may ***mimic*** *the laws of matter. It is almost impossible to detect an individual unit, for in its dance of activity it constantly becomes a part of other such units, expanding and contracting, pulsating and changing in intensity, in force, and changing* ***polarity****. It would be as if the positions of your north and south poles changed constantly while maintaining the same relative distance from one another, and by their change in polarity upsetting the stablility of the planet- except that because of the greater comparative strength at the poles of the units, a newer stability is almost immediately achieved after each shifting.* Even the altruistic motives of the tachion-pair as cited by Killick parallels similar behavior of these EE units.

Again from J.R.: *The behavior of these units changes in the following manner. When a unit is in the act of combining with another, it aligns its components in a characteristic way. When it is separating itself from other units, it will align its components in a different way. The polarities change in each case, within the units. The unit will alter its polarities within itself, adapting the polarity-design of the unit to which it is being attracted; and it will change its polarity away from that design on breaking contact*.

Ken Killick has described the mutual action of several tachion-pairs bonding, as the origin of standard sub-atomic particles possessing features known to modern science such as mass, charge, spin angular momentum, magnetic moment, etc. Thus, according to tachion energy theory, the electron attains its validity, and maintains its structural integrity, by virtue of relations between the component pair dynamics [see [5]). Since for the bulk of their independent existence, tachion pairs are in the pre-physical state, we can speculate that they may actually be superluminal entities. Remarkably, J.R. echoes this view with similar notions: *...the electron achieves its validity because of these* (possibly superluminal) *orbital units* (tachions?). *What we see as an electron is not made of the same things as the orbital units, however; it is an* ***effect****... when you change the structure of the electron, you are changing the relationships between the orbital units. This change will require an application of energy. It may be possible to construct a chamber that would give the electron more freedom.*

Whatever we do now in the handling of electrons is not giving them enough freedom and we are structuring the way in which they can appear. The next observation made by Seth could prove to be essential in clarifying the true nature of time: ... *If electrons were given more freedom they would affect* **our notion of time in perceiving them. The way the measurements (on electrons) are made forces the electron to appear**

in a particular way, but actually it's flitting around, seemingly in many places in rapid succession (pulsing in and out of physical reality). Turning to the structure of elementary particles in [2], Seth offers a similar observation in which he describes the electron as commonly perceived by scientific measurement apparatus, as merely a shadow or facade of the actual electron entity which constitutes a multi-dimensional structure.

In terms of a structural model for the electron, Seth says it would be more appropriate to view these units as composed of *interrelated fields* rather than as the current conception of a particle or a wave. Correspondingly, both of these criteria are key elements of the tachion theory, which posits the electron as composed of two-pair tachion fields. The electron achieves its stability-integrity from the interaction of these free-wheeling units.

Electron Spin and Structure of Fundamental Time

The electron "spin" feature is also claimed to be integral to the exhibit of "time's arrow" in UR [2]: *The electron spin determines time "sequences" from your viewpoint. In those terms, then, a reversed spin is a reversed time motion... electrons, however, spin in many directions at once, an effect impossible for you to perceive. You can only theorize about it. There are "electromagnetic momentums" thus achieved and maintained- certain stabilities that operate and maintain their own integrity, though these may not be "equal" at all portions of the spin. There are equalities set up "between" the inequalities... time, in your terms, then, is spinning newly backward as surely as it is spinning newly into the future. And it is spinning outward and inward* (pulsing inward and outward of physical reality) *into all probabilities simultaneously.*

It is vital that you understand this inward and outward "thrust" of time, however, and realize that from this flows the consecutive appearance of the moment. The thrusting gives dimension to time that so far you have not even begun to realize... This inward and outward thrusting allows for several important conditions that are necessary for the establishment of " relatively" separate, stable universe systems. Such a system may seem like a closed one from any viewpoint within itself. Yet this inward and outward thrusting condition effectively sets up the boundaries and uniqueness of each universal system, while allowing for a constant give-and-take of energy among them. By their unique operation , the EE-units, which form the elementary particles such as the electron, serve as the conduit for this energy flow once it is transduced by elements of thought that Seth terms "consciousness units"(CUs). Continuing in UR on this subject Seth **describes how time can be alternately created or "broken down"**: *The consciousness units serve as source points or "holes"* (Black holes or White holes?) *through which energy falls into your system or is attracted to it- and in so doing, forms it. The experience of forward time and the appearance of physical matter in space and time and all the phenomenal world, results. As CUs leave your system, time is broken down. Its effects are no longer experienced as consecutive, and matter becomes more and more plastic until its mental elements become apparent.* These observations are corroborated by Kryon in his dissertations to follow.

Specifically, in the Kryon channelings, further insight on this electron "freedom" is given in the context of change of frequency: *...There is something we call the electron haze which is around the nucleus of the atom. The space between the energy haze and the nucleus (which is great), varies more than you think in matter. As the space varies, the speed (*frequency*) of the haze must change. It is in the physics where the speed of the haze is strong and fast that you have a different time-frame than your own... you have not yet accepted that an object can seem to be stationary- yet traveling (*vibrating*) very fast. Not in a linear time from point A to point B, but in the electron haze of its vibrating parts.*

J.R. corroborates these comments in this dissertation: *Einstein considered relationships of this sort that hold in the far universe as one increases velocity towards the speed of light,* *but there are also "inner" situations (* altered **stationary** states*) that relate these quantities, and that are much easier to get into.* In regards to atomic structure to allow for frequency changes, Kryon adds: *What is it that makes the distance between the parts change? This is the puzzle for the scientists. For the area between the nucleus and the energy haze although vast, is not void. It is a patterned soup of energy (* the so-called "vacuum"*), and it is within the pattern of this null that changes the distance, and therefore the speed* (frequency*) of the haze.*

Mass, gravity and time are intimately related to this frequency change in the atom. Witness the comments from J.R.: ***Time exists as an invisible drag on matter and this is interpreted as gravity. Time exerts a definite force like gravity, and this is connected with the behavior of the electron.*** *However, since we always look at electrons in a certain way and gravity in a certain way, we miss the connection.* And this related comment as well: *There is another equation to be discovered that expresses the relation between time, mass and gravity.* Kryon reiterates this sentiment, along with relatively detailed scientific information: *Gravity is an absolute product of the characteristics of mass and time... one of which you may change.*

What has been missing so far in your thoughts is how gravity is related to time, and that the entire issue of gravity, mass and time is non-linear. You have yet to develop the formula of physics. It is not here yet, and we are going to give you three attributes of the formula, but without the formula. These attributes are understood but the way they relate are misunderstood. Your science is just now beginning to see the shadows of them, however. The first attribute of the formula is the ability to understand the ***density*** *of mass. Why is the electron haze so far from the nucleus? We have told you that you can change this distance, and in doing so, the density will change in the mass that is defined by the atoms being shifted. The electron haze will have to vibrate faster (this is possibly the "freedom" referred to earlier by J.R.), the closer it is to the nucleus; when the electron haze vibrates faster, its* ***time-frame*** *will shift. Therefore, the second attribute of the formula is a time-shift, and they go together. Also, be aware that you make an erroneous assumption that the actual speed (frequency?) of the electron haze must always be the same no matter what the distance. This is not understood yet either. There is a difference between speed (velocity/translatory rate) and vibratory rate, and it has to do with the actual physical definition of the* ***electron haze****.*

Mass discrimination is the ability to control the density of mass. With a density shift will come a time-shift, and with a time-shift will come a third attribute...we will call the **reality of location**, or where the matter is... in which dimensional reality it goes to when shifted. Here is an axiom: **shifts in dimensionality when matter is changed in this fashion create a reality where the matter has to be to exist in its new form**. It might be inches from where you changed it or miles. That depends on how much it got changed. The difficult thing to explain to you is a concept that has not been recognized... that matter has a **reality index**, and that its core attributes are linked to where it exists in time and space. Therefore, the three attributes that must work together are **density, time-frame and location**.

Electromagnetic Field Manipulation of Cosmic Lattice and Time-Space-Matter

In regards to tapping this (mass-free) energy-haze-soup, between the atomic nucleus and electron shells, Kryon speaks thusly: *There's something we are going to call APD (atomic phasic displacement). This is a term that is going to refer to the ability to tap the (Cosmic) Lattice for unlimited energy. The Cosmic Lattice is balanced, but it is not quiet. The Lattice has astounding power. It has a flow of energy that I cannot explain to you, for there is no paradigm model for it yet in your thinking. Therefore, you would not easily understand. The Lattice has vents that we can best describe as necessary for the flow of energy. It balances the slight inequity of the* *polarity. The vents also have to do with time...you will always find two vents together. One will be prominent, the other will be secondary. You might graphically and clearly see one, but you will have to look carefully to see its partner. There are always two. This is an axiom or physical rule of the Lattice energy, and of the universe:* This information remarkably parallels Killick's teachings on tachion energy pairs.

Witness this Kryon pronouncement:*Hiding within common atomic structure is a marvelous peek at something that will totally and completely mystify you, for it will seem to break all the laws of time and space. The" twins" are a pair of atomic parts that always relate to each other, and are always found in pairs. You will discover that when stimulated correctly, they will always move together as a pair. When you start separating them by distance to experiment, they will continue to move exactly together. If one's energy is converted, then the other will do the same. This will cause you to totally reexamine your ideas of time and space, for this condition will not follow the "ultimate speed" of transmission that you thought was correct... that of the speed of light. You will have discovered something that travels faster than you can ever measure*. Compare this statement with the torsion energy claimed by the Russians to travel at a billion times light speed [10].

In returning to the discussion on how to tap the Cosmic Lattice for energy, Kryon continues: *The vents are very necessary for the balance of your universal energy. The vents are also energy portals, (drains, to you), and are where the front of each Lattice cell touches the back...Like so many other physical processes, however, its going to take a tremendous amount of energy in order to unbalance even one cell's null attribute... Once you understand how to" prod" the null to unbalance itself, you will be rewarded with a steady flow of energy, far beyond what you put into it. This is accomplished since you create your own tiny "vent". An unbalanced cell creates a situation where the other cells around it will try to" feed" energy to the one that is unbalanced. This creates a tap that will pull upon the Lattice indefinitely, as long as your work matches the properties the Lattice expects to see.* Kryon further expands on the specific method to accomplish this...*Here's how it works. Two magnetic fields together, postured in the correct way- a way that is very three-dimensional in your thinking process- will create a "designed magnetic field" that is very specific. It's one you have never seen, and does not exist naturally. Start with trying several magnetic fields postured against one another- of unequal force and pattern, and at right angles. Don't make any assumptions. Think freely. Done in the right fashion, these two fields will create a third pattern which is unique and is the product of the original two. This third custom-created pattern is the one that you want to deal with, and is the one that has the potential of manipulating the Lattice. Once you have created it, you will know of its special qualities by how dramatically it changes the physics around it. And this important admonishment is given:* Here is a caution. *Keep this experiment away from your body. Keep the experiment in check with your scientific methods. Go slowly. Understand what you're seeing before you go on to the next step. Do not expose yourself to any magnetic fields. Remotely conduct all energy experiments. Remember that magnetics also plays an important part within your body (DNA components).*

...your math told you that there might be a possibility of invisibility and that was your goal. This was again because you did not understand that because the distance between the parts change, it doesn't necessarily mean that the overall size will dramatically change or that it might vanish...

Here is another. Understand that if you unbalance the Lattice too grandly and too greatly, you will have a ***time displacement***, *for the process also involves the property of* ***time****... One of the actual physical players in the creative process of unbalancing the Lattice is the manipulation of the time-frame of matter.* ***This is not time travel, but time displacement.*** *It is where you are actually addressing tiny parts of matter, and changing the time-frame they are in. When the inequity of time-frames meet one another (*matter mixed up with differing time attributes*), the result is a displacement of distance (*reality of location*). Although there is no horrendous danger for Earth within this time displacement, it can and will affect the local situation within the experiment. In other words, it can create a matter-distorting effect, completely stopping the experiment, and actually dislocating the parts. We are not going to say any more about this at the moment, but the more astute scientific minds reading this will go to the next obvious step... and the answer is "yes", the Lattice is also the key to quick travel of large physical objects... even of very short distances.*

Possible Documented Demonstrations of Time-Space Warping

Such fantastic transformations were possibly evident in the Hutchison experiments [11], in which metal samples were catastrophically fractured when exposed to radiation from the unique electromagnetic field effect. It was also noticed that many physical objects of different composition which were separate in space before the electromagnetic field was applied, became inexplicably embedded in each other once these objects were radiated by the field effect. For instance, **pieces of wood contained metal knives**, etc. Other metal samples were catastrophically fractured in their molecular-atomic crystal arrangement, producing a substance which was uniform and geometrically symmetrical at one end but appeared to have the resemblance of corrugated cardboard at the other.

Also, similar to Kryon's assertion of the inevitability of time-shift with mass altering phenomena, in Hutchison's work anomalous inertial impulses were recorded on film, sometimes resulting in the complete lift/levitation of various substances of different mass. We notice also anomalous inertial impulses were a feature of the Zinsser "kinetobaric" effect [12], in which due to excitation by **sawtooth microwave impulses with very fast rise time**, a sample attached to a torsion balance, deflected the balance for days and weeks after the original excitation was extinguished. Notice the similarity between these results and those of the Kozyrev experiments, some of which reportedly featured a memorizing preservation effect on a torsion balance, etc.

We also find additional confirmation from the basically anecdotal evidence from the alleged legendary Philadelphia Experiment in 1943, which apparently caused either large time- or space (teleportation) shifts in a Navy ship. In fact, in another piece Kryon himself possibly speaks directly about the Philadelphia Experiment, providing us with further insight into important possible new paradigms for time, mass and energy: *In all your questions regarding magnetics and the massless condition, you have never asked about what happened in your year 1943. You tried to create a massless condition with coarse equipment and little understanding of what you were doing. In the process you actually did create an unstable massless condition for a moment. Its instability created a situation where, instead of a true massless condition, you received one where the time-frame changed but the parts within the sphere of the time change did not have the fine-tuned synchronicity needed for a massless object. The result was an actual distance displacement of the object instead of a true massless condition. Indeed, humans were involved on the large object and their biology was damaged greatly. Your experiment was done in an atmosphere of desperation, and your goal was flawed. For your math told you that there might be a possibility of invisibility and that was your goal. This was again because you did not understand that because the distance between the parts change, it doesn't necessarily mean that the overall size will dramatically change (or that it mightvanish).*

Although this may seem like a paradox to you, the internal mechanics of small particle behavior supports this. The change is measurable but very small, much like what happens with heat and chill...Many of you have guessed correctly that ***magnetics and electricity play a critical role in the determination of the real attributes of mass****... and the magnetic variables that determine the mass product, are often working within very small particles to create the density of an object and also its time-frame. The time frame of a massless object is slightly different than your own, making you to appear slightly slower than the massless object. Its reaction to more traditional mass molecules around it is also predictable: because of the very slight time displacement, it tends to change the number of electrons within the atoms directly in contact with it. This is a clue on how to detect a massless object even if you can't see it.*

*The reason you felt the object (*ship*) would vanish was that you were able to simulate a "vanishing" in the laboratory with smaller objects. This observation was not consistent however, so you again were in desperation to try this experiment on a larger object. The "vanishing" was an illusion and was due to a distance displacement rather than an "in place" vanishing... does this give you a clue to long distance travel using magnetics and the vibratory rate of matter?... Only one human on the planet has ever captured a true massless condition, and even this was a coarse one and lasted for only for a few moments that were uncontrollable in scope. This massless condition was created in the primitive workshop of a great electrical scientist in your culture on the American continent not too long ago. If you were able to visit his workshop, you would note the holes in the ceiling and the patched glass-covered light port where his massless objects took off and wildly flew everywhere. If he had been born 50 years later he would have been able to control the attributes of the experiment. As it was, he did not have the precision tools you have now to direct and control such an experiment. It was his great passion to understand this phenomena, but because it was so uncontrollable and sporadic, he was never able to bring others to see it work, for he could not accurately create it consistently. This depressed him in his later years, for his was a great three-dimensional mind...* When asked the identity of this scientist Kryon replied*: The inventor of your multi-phase current, born in the land you now call Yugoslavia.*

From the above information, there emerges a remarkably consistent picture suggesting that atomic structure may not be as rigid or immutable as formerly thought. Changing the fundamental frequency of vibration of atomic structure can perhaps apparently not only change the density of matter by altering the distance of electron shells from the nucleus, but in so doing alter the time-frame of that object, its mass and also the reality of location of that substance. Certainly, the implications of the Kozyrev research, among others, can possibly form the basis for the eventual technological implementation of these futuristic principles. Perhaps those experimenting with the relatively detailed data provided by Kryon on APD can actually realize this coveted goal.

Maverick New Theories in Physics Suggest Fluid Nature for Time, Space, Energy

The Dirac theory is to be understood as a theory of the electron with electromagnetic interactions. It consists of the Dirac wave function and its properties, including the Dirac equation and relations to physical variables such as energy-momentum, spin and position probability current. In spite of its indisputable mathematical successes, the Dirac theory is still without a completely satisfactory physical interpretation. In particular, the so-called *zitterbewegung* ("jitter motion") is a distinctive feature of the Dirac theory which continues to be the subject of conflicting interpretations in the standard scientific literature. The concept of zitterbewegung was introduced by Schrodinger to interpret high frequency oscillations in free-particle wave packets of the Dirac theory. These oscillations with angular frequency $2mc^2/h$ arise from interference between positive and negative energy components of a wave packet. Schrodinger interpreted the oscillations as fluctuations in position of the electron about an average motion. However, this concept has not been interpreted in the bulk of the literature as having an objective significance, but has erroneously been viewed by standard theory as a mathematical artifact of the one-particle Dirac theory which does not appear in a correctly formulated quantum field theory.

However, as noted theoretical physicist David Hestenes suggests in the first of his recent article series in the journal, **Foundations of Physics** [13], it is only by interpreting the zitterbewegung literally and objectively as point particle motion, that a complete and coherent interpretation of the Dirac theory can be achieved. Moreover, this model has implications that have not yet been previously considered by established physics that relate directly to some of the key features of sub-atomic behavior enumerated in the tachion energy theory and also in the microscopic dynamics of elementary particles revealed by the selected esoteric channelings that we have previously considered in the present exposition.

In particular, it implies that the electron is the seat of a fluctuating electromagnetic field which oscillates with the DeBroglie frequency of the electron (1.6×10^{21}cycles/sec).

Thus, it tells us that a kind of electromagnetic wave-particle duality is implicit in the Dirac theory. Of course, this contradicts the conventional view that wave-particle duality is a property of matter which is completely independent of the nature of its interactions. If the zitterbewegung is an objective phenomenon, then it originates from electron self-interaction, since it persists in the absence of external fields. In this new theory which considers the zitterbewegung as an objective physical phenomenon, the electron mass and spin can be identified with the energy and angular momentum of electromagnetic interaction. It suggests that the self-interaction is such that there exist certain stable, non-radiating but accelerated states of motion; in particular, for a free particle, this implies motion in a circle with the radius of a Compton wavelength- the zitterbewegung. The zitterbewegung implies that some of the mass, at least, is kinetic self-energy associated with a spin, and it reconciles the mass with a zero mass of the bare electron. Thus, according to this bold new theory, the electron spin arises from a helical world-line in spacetime. To be sure, some researchers have previously postulated such a helical electron dynamics. Unfortunately, such models have failed to explain why a helical motion for spin should depend on interference between positive and negative energy states or why the zitterbewegung should depend on the way a wave packet is constructed, or even how the zitterbewegung can be the origin for electron spin despite the fact that it vanishes for plane wave states that certainly describe a particle with spin.

The Hestenes model does offer an explanation for such effects and accounts for the ubiquitous feature of spin angular momentum as a function of the zitterbewegung. The essential unprecedented feature of the Hestenes' zitterbewegung idea is the association of the spin with a local circulatory-helical motion characterized by the phase factor of the electron wave function.

Thus, we reach the conclusion that the complex phase factor of the electron wave function can be associated directly with an objective helical motion of the electron, which is, in turn, a derivative of the zitterbewegung. Although the idea of helical motion connected with the electron has been considered before, it has not previously been related to electron phase to produce a complete interpretation of the Dirac theory.

Moreover, in the next installment of his dissertation [14], Hestenes argues that the zitterbewegung is not only an objective dynamical phenomenon associated with the electron, but is a ubiquitous phenomenon, with manifestation in every area of quantum mechanics, even in the non-relativistic domain. For instance, by showing that spin angular momentum can be regarded as angular momentum of the zitterbewegung fluctuations, the zitterbewegung interpretation of the Dirac theory begun by Hestenes in reference 13 provides an explanation for the electron spin and magnetic moment in the physical circulation of momentum and charge. It also explains the mass as the energy of this circulation. Thus, the origin of the zitterbewegung is attributed to self-interaction of the electron with its own electromagnetic field. The relations derived by Hestenes in this article suggests that the interaction is of magnetic origin, since it has the form of a Larmor precession energy if spin angular momentum is proportional to a self-generated magnetic field. The so-called "rest mass" of the electron is thus a kinetic energy of self-interaction. It is this that gives the electron its inertial properties, and the flywheel-like nature of this inertia may be the ultimate origin of spin dependence in electron scattering. And the Heisenberg Uncertainty relations can now be attributed to the "zitterbewegung fact" that an electron cannot be confined to a region smaller than a Compton wavelength. Also, the stationary states of a bound electron exhibit a resonance of the orbital frequency with harmonics of the zitterbewegung frequency, which is imposed formally in the standard theory by requiring single-valuedness of the wave-function. Evidently, such resonances, so prominent in quantum mechanics, can be interpreted as zitterbewegung resonances. This leads to a new explanation of penetration of a potential barrier by sub-atomic particles as due to zitterbewegung fluctuations in momentum, and the Aharonov-Bohm effect as a shift in zitterbewegung phase. The zitterbewegung phase factor literally represents a physical rotation. The rotation rates of this phase in time and space directions are the source of the electron's energy, mass and momentum.

In his third paper [15], Hestenes draws a closer relation of the zitterbewegung dynamics to the Dirac theory by suggesting that the latter actually describes a statistical ensemble of possible electron motions, which are actually governed by the zitterbewegung substructure. Since the energy-momentum of the electron can now be interpreted in terms of a rotation rate in the spin-plane, Hestenes then derives a corresponding relation which defines a variable mass for the electron-another unprecedented idea which is in agreement with the information given by Kryon in his dissertations and Ken Killick's tachion theory. The mass m, the scalar radius of curvature r and the zitterbewegung frequency all covary with changes in the rotation rate in the electron spin-plane. Hestenes obtains a relation demonstrating that the electron mass is inversely proportional to the zitterbewegung frequency. This conforms to the relativistic concept of mass as a measure of energy content. But here, mass is concluded to be primarily a **frequency** measure. This also conforms to DeBroglie's original idea that the electron contains an internal clock with frequency determined by its mass, though for a free particle, the zitterbewegung frequency differs from the DeBroglie frequency by a factor of 2. Moreover, the new key relation derived by Hestenes: $m \cdot r = 1/2h$ (h, Planck's constant), says that this frequency measures the radius of curvature of the electron world-line, so it is a thoroughly geometrical quantity. Thus, as mass increases the radius of curvature decreases, in concert with the tachion dynamics as well. All this suggests that the electron mass relates our externally imposed time scale to a time scale intrinsic to the electron.

The reader can see that the information transmitted by J.R. is also remarkably similar in this connection. It should be evident that there is introduced a new concept of mass here, though, to a certain extent, it was already implicit in the Dirac theory. The formerly vague concept of mass as some kind of material stuff is completely gone. Also, no longer is vanishing mass a distinguishing feature of particles moving with the speed of light.

Corroborating Research on Oscillatory-Fluid Nature for Time

In his recent thought-provoking dissertation [16], Tom Bearden relates of the dynamic influence of time on a material system: "But we can also legitimately state that 'time is energy' and be rigorously accurate...Time is extremely compressed EM spatial energy...Without further elaboration, we speak of a 'mass' in which a small portion exists as 'masstime' rather than mass, as having been 'time-charged' or 'time-excited'.

The t-polarization wave in the time dimension is quite unique: The *spatial* energy of the wave is in equilibrium and not vibrating at all; instead, the photons comprising the wave are vibrating in their **time**-components... In short, mass m is changed into masstime mt by photon absorption... Rigorously, a mass does not really 'travel through time' continuously, per se, but proceeds with an overall serial change mechanism, driven by its total virtual and observable photon interactions, as m→ mt→ m→ mt→...We propose that this may account for the duality of particle and wave...The particle actually oscillates at a high rate between the m and mt states... Mass 'travels through time' by an extremely high oscillation between corpuscle-like state and wave-like state". Notice how Bearden's description of the macroscopic operation of time on a physical system closely parallels Hestenes' proposed model for of the zitterbewegung dynamics at the sub-atomic level which provides the electron (and possibly photons, etc.), with a time-substructure (its internal "clock").

We can see many other places in Bearden's treatise where the unique terminology chosen can now be put into direct reference with equivalent concepts and unique phraseology employed by many of the researchers examined in the current exposition. This cross-referencing certainly provides much needed clarification about key ideas for researchers currently involved with the development of this fledgling discipline that we shall term causal mechanics, after Kozyrev's designation.

All such efforts will help to place investigative endeavors currently scoffed at by establishment science on a firm rational foundation, as well as assist in synthesizing seemingly disparate but surprisingly related "anomalous" phenomena. For instance, to implement the science he has termed "vacuum engineering", Bearden often speaks of the necessity of establishing a "local time-stress of the **vacuum potential**". This process, which as we hope to have shown here is completely legitimate, is identical to what Smith calls creating a "tempic field gradient", Kozyrev's "alteration of time-density", or Alexander Frolov's "change in the local time-rate".

It can clearly be seen that such ideas can only be perceived as science fiction fantasy, if we observe strict adherence to all current received paradigms (even in superstring theory), which consider space and time purely from their relational standpoints. These theories must of necessity involve models employing scalable metrics for their proper description. However, once we embrace the novel view posited here that vacuum energy is primary and both local space and time are derivatives of this mass-free non-scalable dynamic non-local oscillating substratum, then we can conceive of space and time properly as complementary aspects of non-local conscious reality, neither fixed in a scalable manner or subservient to the other, but fluid and mutually interdependent features of energy in constant flux (see the related ideas voiced by A. Correa and P. Correa in I).

With this approach we will then come to acknowledge the technological feasibility of many of the "futuristic" claims made in the channeled transcripts. For instance, we can see that it will no longer be necessary to propel elementary particles to relativistic velocities, or to probe astrophysical systems possessing large gravitational fields, to recognize alterations in space, time and mass parameters, since these phenomena can equally as well be accomplished in the laboratory by modifying atomic structure of stationary matter through electromagnetic means. Once again, peruse the Kryon dissertations on this process, for it may be key to future technology.

For those who wish to further investigate these theoretical possibilities, the papers of **A. A. Nassikas** are recommended. Nassikas has postulated what very few researchers before him have theorized: the existence of a ground-form probability density function for energy, as a necessary precursor to the malleable (deformable) characteristics of time and space [18]. One of Nassikas' conclusions is significant in light of ideas articulated in the present exposition concerning the hypothetical tachion-pair model for reality (see I). Here we refer to the key push-pull oscillatory ("breathing") operation between space and time, which may be responsible for the structural integrity of matter as we know it (again re-read the Hestenes zitterbewegung model).

Researcher **Alexander Frolov** has taken Nassikas' theory even further to suggest an explanation for the **anomalous over-unity effect** that has continually been demonstrated in low-energy nuclear reactions (LENR) over the past decade [19]. In such so-called "cold fusion" cells the Palladium cathode is over-saturated with protons, producing a local imbalance in the vacuum engine, causing a local alteration in the time-frame of the experiment to compensate the local change in energy density. **Subsequently, as Frolov claims, this time-frame change converts non-local time-energy directly into heat energy, similar to Kozyrev's description of energy production by stars.** Perhaps this observation might provide the hitherto missing key in our understanding of fusion by electrolytic action.

Conclusions and Prospects

In the preceding paper it is sincerely hoped that the author has demonstrated ample evidence that will spur on researchers of a kindred spirit towards the development of new paradigms for space, energy, mass and especially time. All these concepts, as well as the foundations of current classical electromagnetic theory, are in need of a drastic overhaul before we can fully appreciate and understand the operations of nature as it interrelates the metaphysical-mental realms with the physical level of reality.

At any rate, the continual appearance of research results, such as the Kozyrev effect, which up to present apparently resists explication in terms of conventional scientific paradigms, certainly offers clear evidence of the incompleteness of our knowledge of nature, even in this enlightened era since the dawn of the new millennium.

Accordingly, hints as to how the Kozyrev effect arises in association with the postulated active properties of time, and how the latter interacts with known physical forces, can assist us towards the development of a more comprehensive paradigm which embraces a wider spectrum of human knowledge - one in which the current conundrums rampant in the foundations of theoretical physics, and biological science disciplines as well, which have hitherto prevented the establishment of a unified model of all physical interaction, can find satisfactory solution. Also, such empirical evidence which generally flies in the face of current scientific wisdom, presents the greatest challenge to our ongoing search for new sustainable energy sources which will be of absolute necessity in the future.

In this light, if formally integrated into scientific thought, the tenets of tachion energy theory, which is based on an altruistic modus operandi, also may hold profound implications for the status of our very social, political and religious structures as well. Indeed, unlike the impersonally motivated interactions believed to be the basis for force interactions in modern sub-atomic physics, which help sustain our illusory belief in the supremacy of the ego, the philosophical basis of tachion dynamics is an able reminder of the humility of purpose which should be the guiding force whatever our walk of life. Like tachion-pairs, in the ideal social structure individuals will work side-by-side, yet achieving a common accord. United in the circle of their activity they will work together joyfully in freedom, each with his own task, yet always conscious of the common bond.

With the full import of these simple truths appreciated, tachions may afford a rational justification of the formerly "unprovable" mystical doctrine of the omnipresent intelligent plan behind the working of the universe at all levels of conscious reality. We will come to realize the profound truth of Einstein's vision of ultimate reality couched in his heretofore cryptic statement, *"God is subtle, never malicious"*.

REFERENCES:

1 J. Roberts, unpublished remarks as transcribed by W.H. Kautz.
2 J. Roberts, **The Seth Material**, Prentice-Hall, Inc. 1970. Pp. 292-298.
The "Unknown" Reality, Prentice-Hall, Inc., 1986.
3 L. Carroll, **Alchemy of the Human Spirit**, (Kryon Book III), The Kryon Writings, 1995.
Letters from Home (Kryon Book VI), The Kryon Writings, 1998.
Passing the Marker (Kryon Book VIII), The Kryon Writings, 2000.
4 W. Smith, **The New Science**, Fern-Graphic Pub., 1964.
5 D. Reed & K. Killick, "Tachion Energy Theory"(Parts I-IV), **Energy Unlimited**, 1978-1983.
6 A. Correa & P. Correa, "Consequences of the Null Result of the Michelson-Morley Experiment: The Demise of the Stationary Aether , the Rise of Special Relativity and the Heuristic Concept of the Photon", **Infinite Energy**, no. 38, July-Aug 2001, pp. 47-64.
"The Sagnac and Michelson-Gale Experiments", **Infinite Energy**, no. 39, Sept.-Oct. 2001, pp. 32-49.
7 N. Kozyrev, "On the Possibility for the Experimental Investigation of the Properties of Time", **Time in Science and Philosophy**, Prague, 1971, pp. 111-132.
"An Unexplored World", **Soviet Life**, 1965.
A.P. Levich (ed), **On the Way to Understanding the Time Phenomenon (Part II: The Active Properties of Time According to N.A. Kozyrev)**, World Scientific, Singapore, 1996 (rare English language excellent critical review of Kozyrev research and theories).
8 M. Lavrentiev, "Detection of the Reaction of Matter to an External Irreversible Process", **Sov. Phys. Dokl.**, Mar 1991, vol. 36, no. 3, p. 243.
9 M. Lavrentiev et al., "Remote Effects of Stars on a Resistor", **Sov. Phys. Dokl.**, Sept 1990, vol. 35. No. 9, p.818.
"Detection of the True Position of the Sun", **Sov. Phys. Dokl.**, Nov 1990, vol. 35, no. 11, p. 957.
"Scanning the Celestial Sphere with a Kozyrev Detector", **Sov. Phys. Dokl.**, Apr. 1992, vol. 37, no. 4, p. 163.
10 D. Reed, "Torsion Field Research and Implications for New Physics and Energy Technologies" **Jour. New Energy**, vol. 4, no. 2, Fall 1999, pp. 151-164 (and references therein).
11 J. Hutchison, "The Hutchison Effect Apparatus", **Proc. Of the 1st Symp on New Energy**, Denver, May 1994, p. 199.
"Inside View of the Hutchison Effect", **Extraordinary Science**, vol. III, issue 4, Oct-Dec 1991, pp. 23-26.
12 T. Valone(ed), **Mechanical Energy from Gravitational Anisotropy**, Integrity Research Institute, Washington, DC, 1996.
W. Peshka, "Kinetobaric Effect as Basis for a New Propulsion Principle", **Raumfahrt-Forschung**, Feb 1974 (in German); trans. By D.Reed, **Infinite Energy**, vol. 4, issue 22, 1998, pp. 52-56.
D. Reed, "Translator's Analysis and Comments on the Zinsser-Effect Device", **Infinite Energy**, vol. 4, issue 22, 1998, pp. 57-59.
13 D. Hestenes, "Quantum Mechanics from Self-Interaction", **Found. Phys.**, vol. 15, no. 1, 1985, pp. 63-87.
14 D. Hestenes, "The Zitterbewegung Interpretation of Quantum Mechanics", **Found. Phys.**, vol. 20, no. 10, 1990, pp. 1213-1232.
15 D. Hestenes, "Zitterbewegung Modeling", **Found. Phys.**, vol. 23, no. 3, 1992, pp. 365-387.
16 T. Bearden, **Extracting and Using Electromagnetic Energy From the Active Vacuum**, Association of Distinguished American Scientists, Huntsville, Alabama, 2000.
"Giant Negentropy in the Common Dipole", **Proc IC-2000**, St. Petersburg, Russia, 2000 (in press).
17 A.A. Nassikas, **The Hypothesis and the Equations of the Unified Matter Field**, in New Ideas in Natural Sciences (Part I: Physics), St. Petersburg, Russia, 1996.
18 A. Frolov, **Work Created By Means of the Potential Field**, in New Ideas in Natural Sciences(Part I: Physics), St. Petersburg, Russia, 1996.

ELECTRIFYING TIMES

an online and published magazine about Electric, Hybrid, Fuel Cell Vehicles, advanced batteries, ultra capacitors, fuel cells, microturbines, free energy systems, events and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908 fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

On the Rotating Permanent Magnet System

Alexander V. Frolov, Russia
7 Lev Tolstoy Str., St. Petersburg, 197376, Faraday Laboratory Ltd.

In one of our issues we published an article about the results of the experiment conducted by S.M. Godin and V.V. Roschin. They started in 1990 at Moscow Aviation Institute (MAI) when a technical physics laboratory was created in Machine Industry Projects Institute at the "Aeropromservice" Association. Their research work was financed by private investors and was continued in close cooperation with Chair 310 of MAI, especially with Prof. PhD. L.K. Kovalev, Prof. PhD. V.V. Rybakov and PhD. K.V. Ilyushin. By 1991 "a device for inner energy of matter conversion" had been designed and tested (according to the Technical Design Assignment). The prototype was finished in the middle of 1993 (Fig.1).

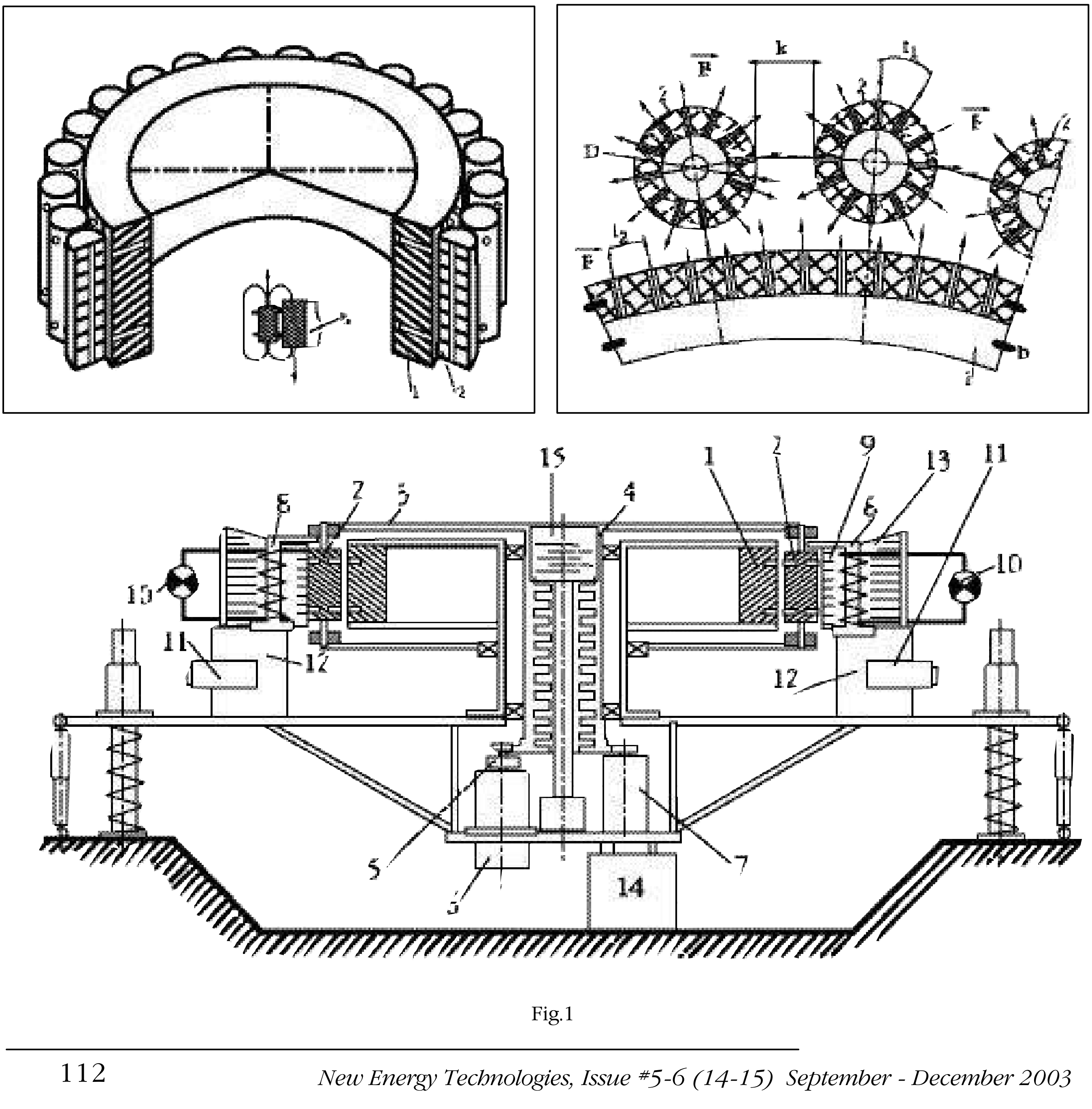

Fig.1

The experiments resulted in the possibility of creating a 7 kW output power in load after the rotor was sped up to 10 rps, but in the autumn of 1993 the laboratory was closed due to the investors' financial problems. Now their research work is continuing in Moscow and we hope the scientists will obtain new experimental results.

You can email S.M. Godin for more information at: smgodin@online.ru

Considering the theoretical basis of the functioning of these systems I should suggest a number of interesting details for prospective researchers.

S.M. Godin, V.V. Roschin et. al. (Physics of Negative Viscosity Phenomena. Prof. Victor P. Starr. Massachusetts Institute of Technology) spoke of vacuum (quantum medium) as **a physical medium possessing negative viscosity.** At that rate, it is worth making assumptions that can be useful to the developers of this topic. I think that standard (positive) viscosity of matter, from the physical point of view presupposes the presence of positive friction and a "work-to-heat" entropy transformation. Consequently, **negative viscosity** should correspond to the **negentropy cycle** "heat is converted to work". This is exactly the stray heat conversion once described by Tsyolkovsky in 1914. Apparently, **the medium of the kind supports self-sustained vortices** characterized by the surrounding medium heat absorption, which was mentioned yet not explained by S.M. Godin and V.V. Roschin.

It is possible that all the elements of matter are the self-sustainable vortex structures. To mathematically describe these physical mechanisms, the notions "heat-work-heat" transformation and "time" or "gravitation" should be interrelated. Russian philosopher Pavel D. Uspenskiy had been making related research since the year 1910.

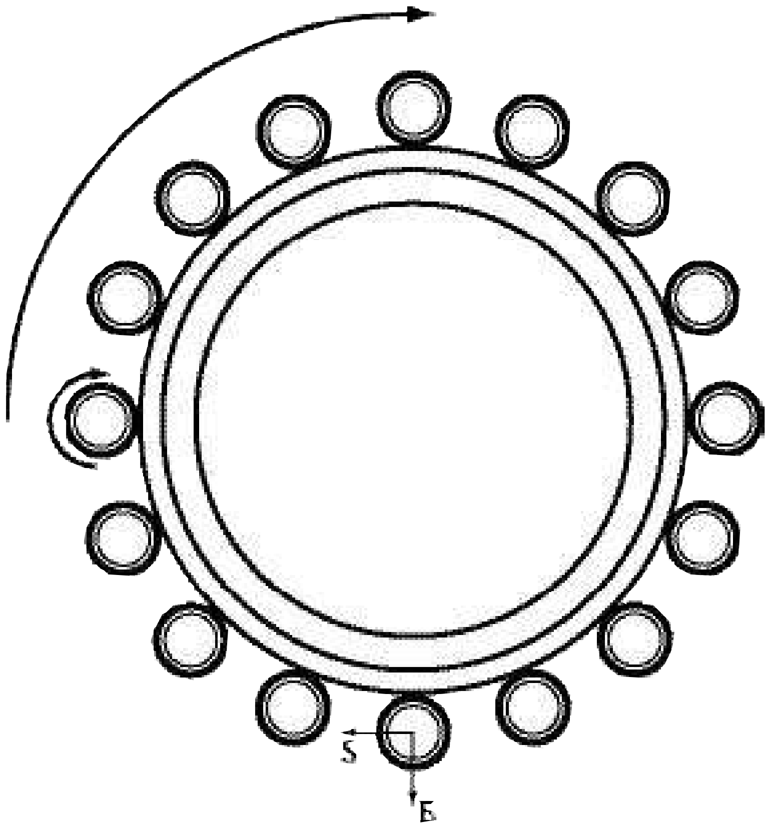

Fig.2

From the conventional point of view, this device (Fig.1) creates a circulating energy flux described by the Umov-Pointing vector (toroidal vortex) as the roller magnet field vector is vertical and the electric field is radial.

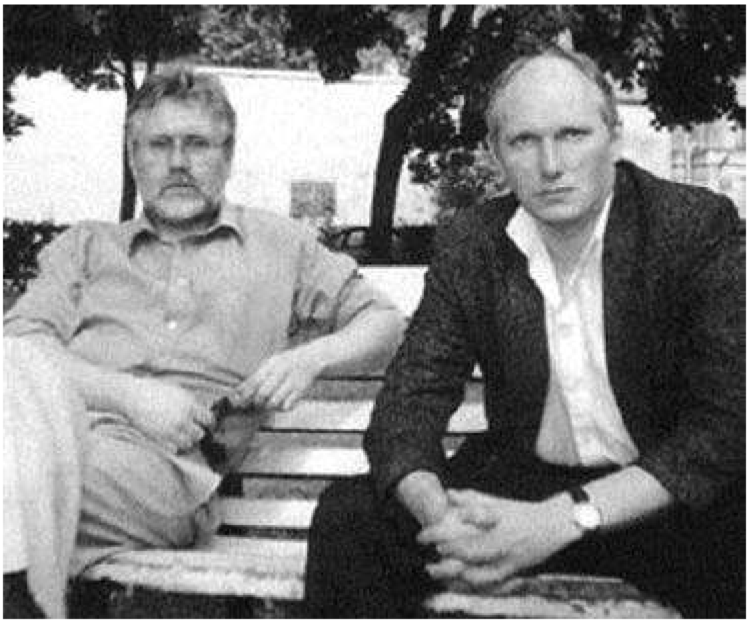

S.M.Godin (on the left) and A.V.Frolov (on the right), St. Petersburg, June 2001.

Inventor's Week Conference 2003, USA

Review is prepared by Editor Tatiana Ezhova

This year the annual Inventor's Week-end Conference was held on September 19-20 in Seattle, Washington. The conference featured several interesting inventions, including MEG.

Bill Alek demonstrated a device that he believed could let him detect Vortex-Energy. It was attached to a meter, and his belief was that if he walked into a naturally-occurring energy vortex carrying the device it would show him changes in gravitational and other energies. He cited the Vortices in Santa Cruz and Oregon as examples that might be good for experimentation.

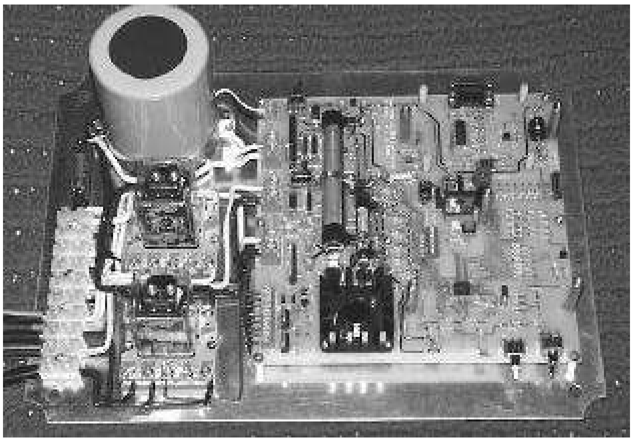

Bill Alek

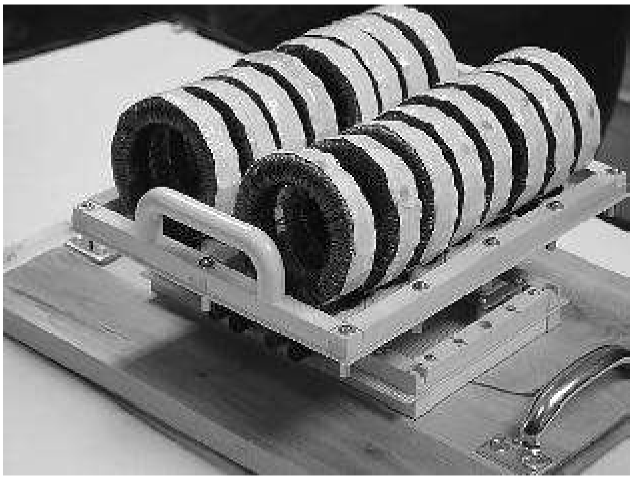

Z-Pod, Bill Alek

In his report "Introduction to Parametric Mass Fluctuation - a breakthrough in Energy and Inertialess Field Propulsion" Bill Alek presented the Z-Pod, developed to test a theory of solid-state parametric mass fluctuation. Here is the URL for a good overview of the device that he published online at www.intalek.com:

http://www.intalek.com/Index/Projects/SmartPAK/Projects/ZPOD/ZPOD_System03.pdf

Rio Von Sternberg

Rio Von Sternberg was presented as the original inventor of the MEG. He apparently has a patent on the device; however, from his presentation I understand that his configuration is somewhat different than the model that J.L.Naudin and T.E.Bearden have been experimenting with.

Alexander Peterson had a very interesting presentation on optical-mirror magnetic field systems. The crux of his presentation involved cutting precise notches into magnetic coils to separate them into identical optical-isomeric shapes. Apparently, by cutting a thin line directly down the center of the transformer-core is believed by Peterson to increase the efficiency of electric motors by up to 50%.

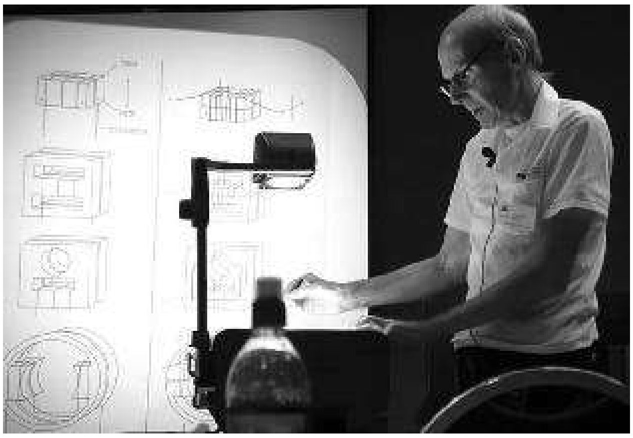

Alexander Peterson

Mark Plotkin, a well-known Washington DC attorney also took part in the conference. He is interested in new energy technologies and works with new technologies in collaboration with Mark Whitford at www.awetec.info. The photo of Mark shows him rotating a steel coil apparatus that was passed around the conference by Slim Spurling.

Mark Plotkin

Slim Spurling is an inventor from the Southwest, shown holding a helical-wound coil that he claims provides health benefits through an interface with zero-point energy. Spurling's belief is that this specific circumference of the coils provides a "Casimir-like Effect" that concentrates the energy.

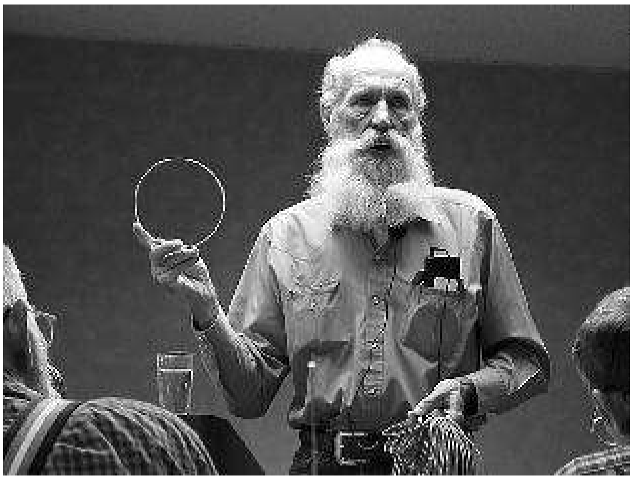

Slim Spurling

Donald Smith

Note: Donald Smith, a well-known inventor, also took part in this year's conference. His work is described in more detail in a separate article of this issue.

Full information on the conference and reports is presented at http://ionvalve.com/iw/2003/conference2003.htm.

USA "Inertia" Patents

Information from www.rexresearch.com/inertial/

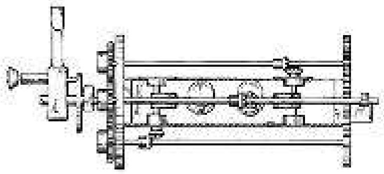

№ 6,290,622 (7/18/01)
Mechanical Force Generator
Murray, Lawrence D.

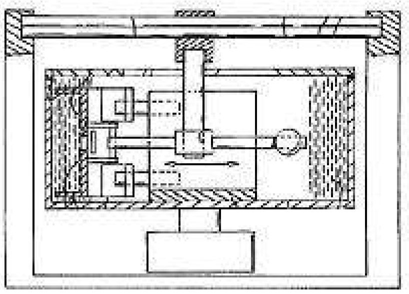

№ 6,259,177 (7/10/01)
Motion Imparting System
Deschamplain, Daviv

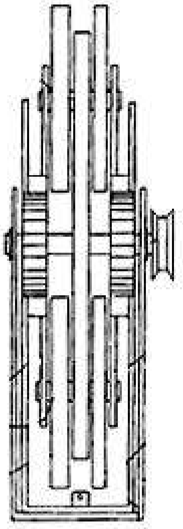

№ 5,890,400 (4/06/99)
Apparatus for Generating
a Propulsion Force
Oades, Ross A.

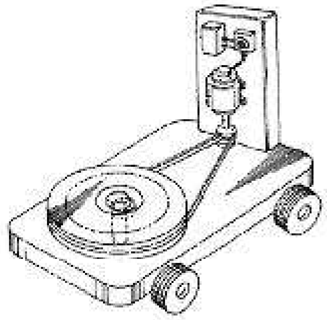

№ 5,937,698 (8/17/99)
Centrifugal Propulsi
on System
Kunz, William T.

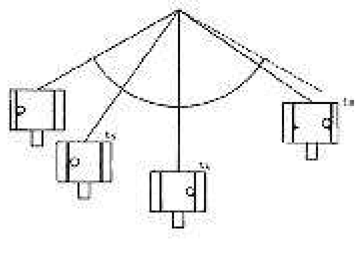

№ 6,089,511(7/18/00)
Method & Apparatus
for Propulsion
Rasmussen, James K.

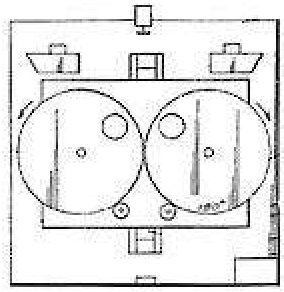

№ 5,966,986 (10/19/99)
Propulsion System
Laul, Virgil R.

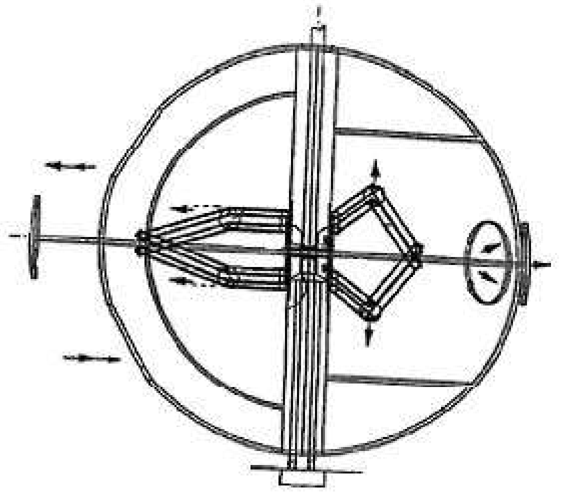

№ 5,673,872 (10/7/97)
Apparatus for Energy Transformation & Conservation
Shimshi, Ezra.

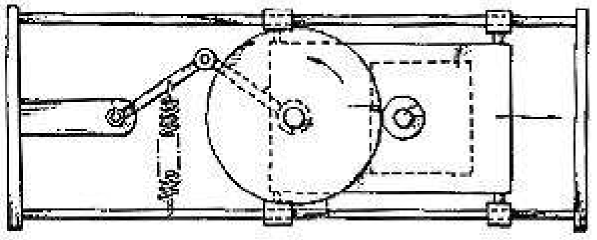

№ 3,492,881 (2/3/70)
Prime Mover
Auweele, Albert J. v.

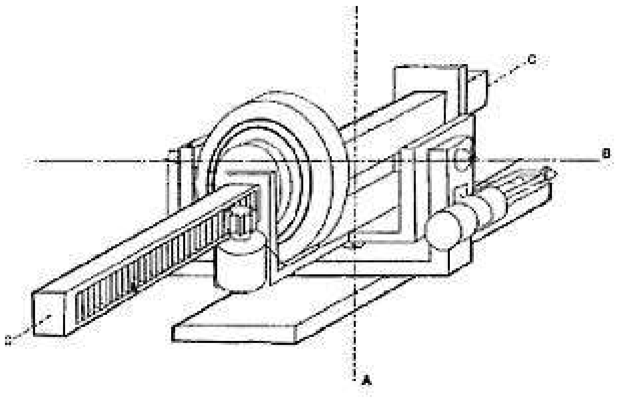

№ 5,860,317 (1/19/99)
Gyroscopic Propulsion System
Laithwaite, Eric R. & Dawson, Wm.

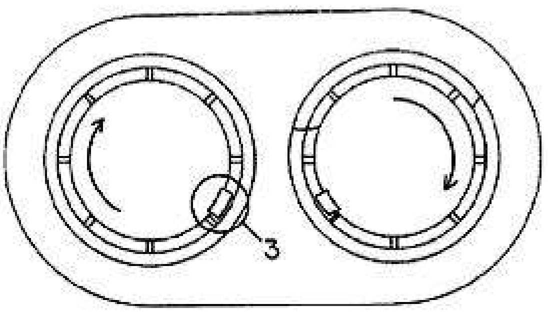

№ 5,831,354 (11/03/98)
Bootstrap Propulsion System
Stopplecamp, Timothy J.

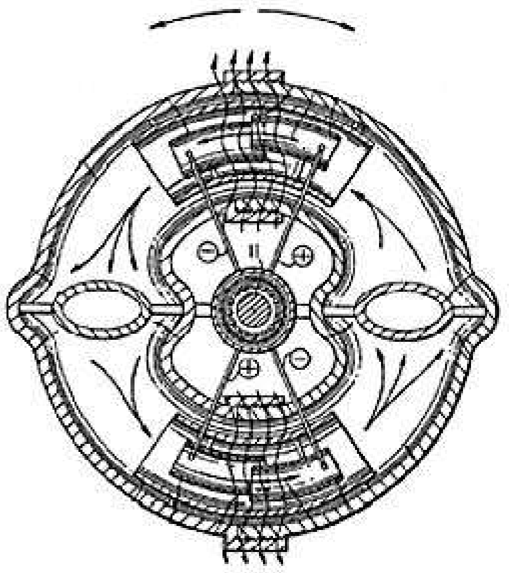

№ 5,410,198 (4/25/95)
Propulsion System
Butka, Kemal

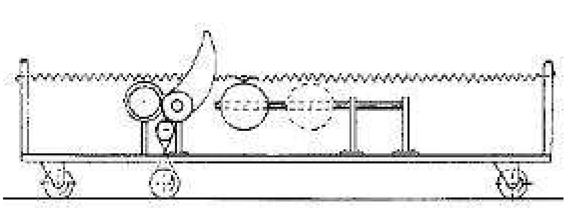

№ 5,685,196 (11/11/97)
Inertial Propulsion Plus Device & Engine
Foster, Richard E.

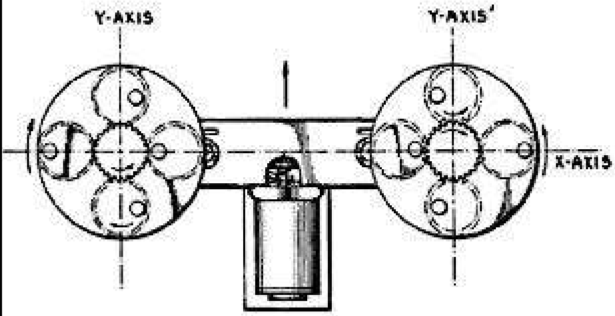

№ 5,167,163 (12/1/92)
Energy Transfer Device
McMahon, John C.

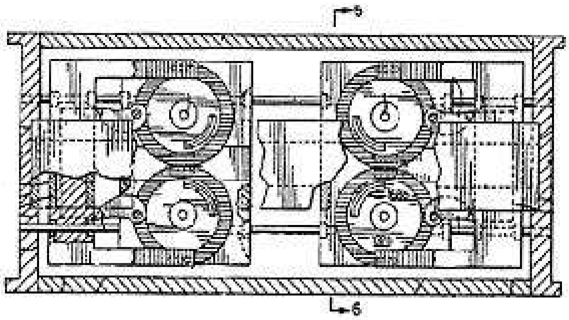

№ 5,182,958 (2/2/93)
Non-Linear Propulsion & Energy
Conversion System
Black, James W.

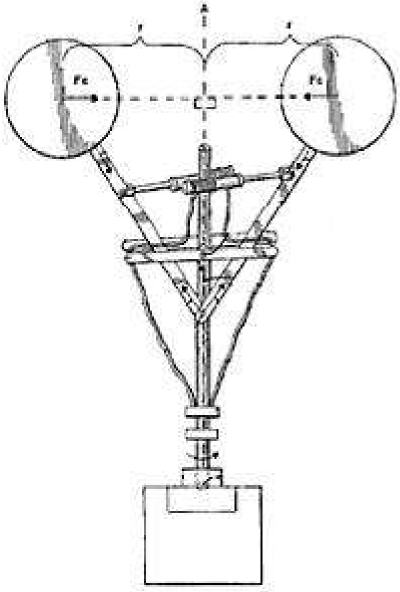

№ 5,557,988 (9/24/96)
Centripetally Impelled Vehicle
Claxton, John C.

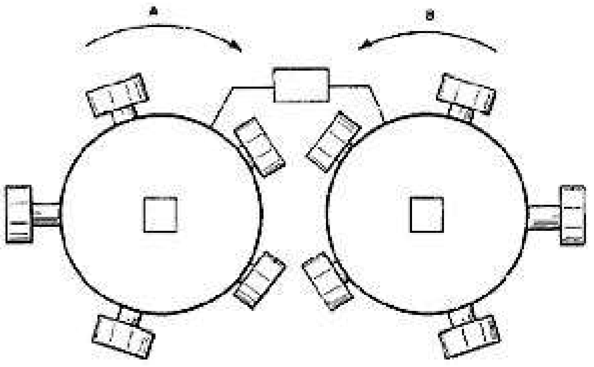

№ 5,782,134 (7/21/98)
Electromagnetically Actuated
Thrust Generator
Bouden, James D.

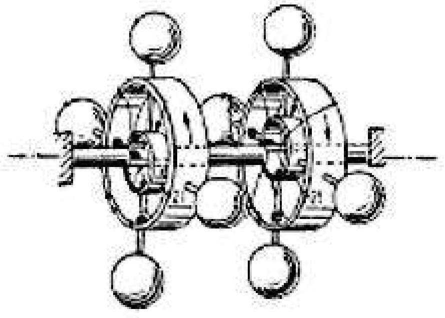

№ 5,156,058
Converting Rotary Motion
to
Lineal Motion
Bristow, Theodore R., Jr.

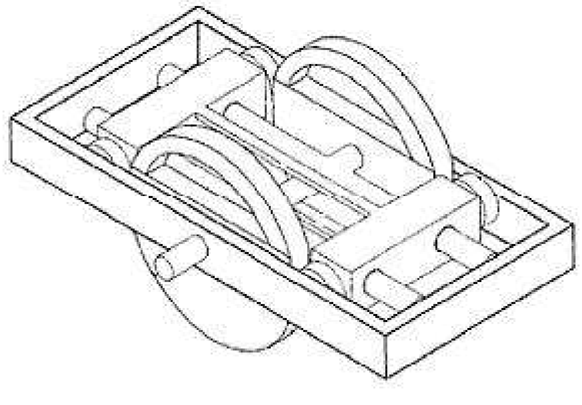

№ 5,488,877 (2/6/96)
Centrifugal
Inertia Drive
Lieurance, Richard L.

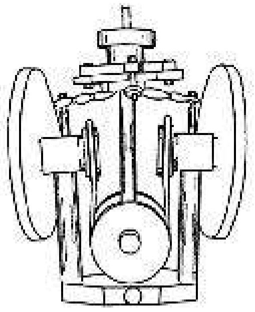

**№ 5,024,112(6/18/91)
Gyroscopic Apparatus
Kidd, Alexander D.**

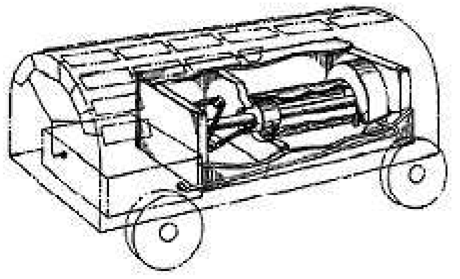

**№ 4,801,111(1/31/89)
System for Propulsion &
Positioning
Rogers, Charles E., et al.**

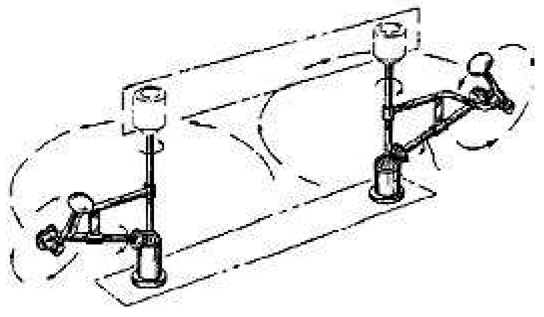

**№ 4,991,453(2/12/91)
Centripedal Device for
Concentrating
Centrifugal Force
Mason, Lyle M.**

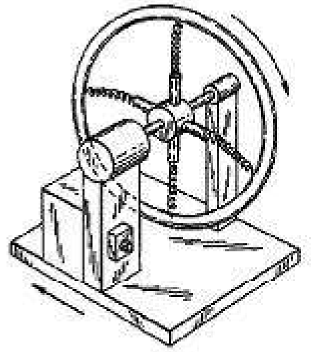

**№ 4,784,006(11/15/88)
Gyroscopic Propulsion
Device
Kethley, Lancelot I.**

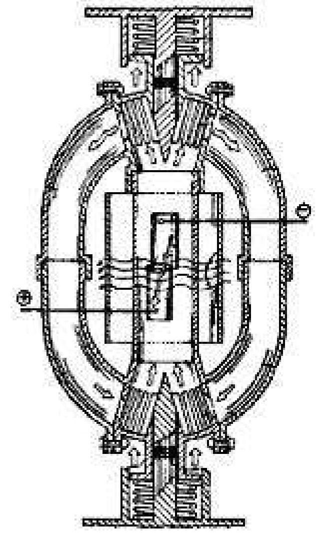

**№ 5,111,087(5/5/92)
Propulsion System
Butka, Kemal.**

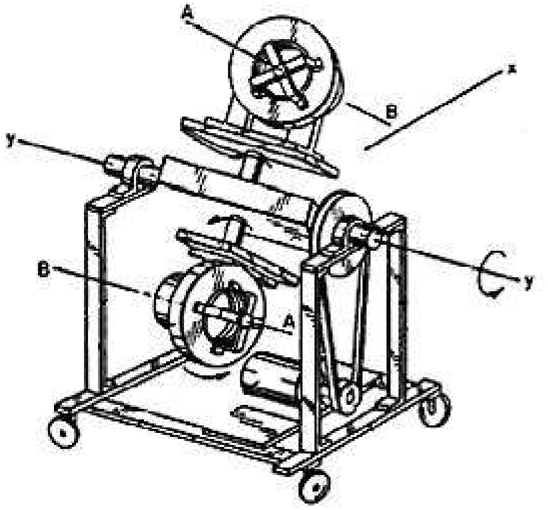

**№ 5,090,260(2/25/92)
Gyrostat Propulsion
System
Derloy, M.S.& Derloy M.**

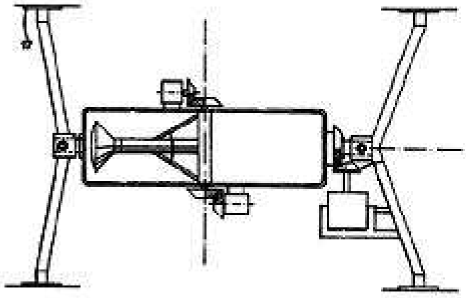

№4,884,456(12/5/99)
Centripedal Device for Concentrating Centrifugal Force
Zachystal, George J.

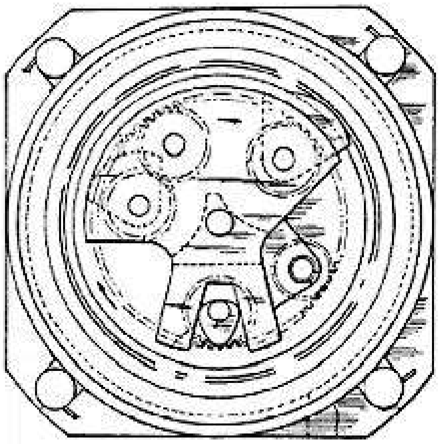

№5,042,313(8/27/91)
Conversion of Rotational Output to Linear Force
Montalbano, Paul J.

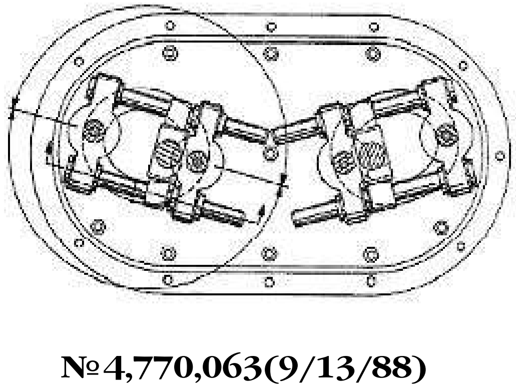

№4,770,063(9/13/88)
Universal Propulsion Powerplant & Impulse Drive Unit
Mundo, James D.

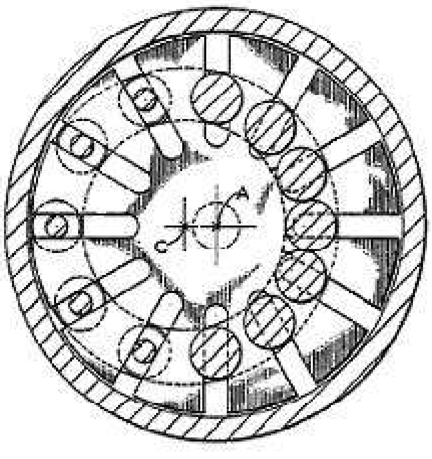

№4,788,882(12/6/88)
Flywheel
Fulop, Charles.

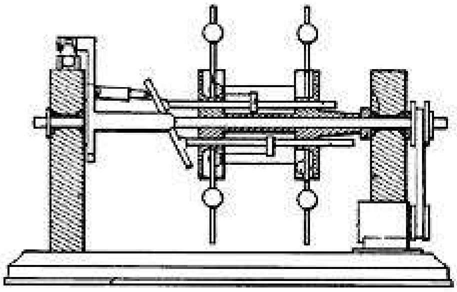

№5,054,331(10/8/91)
Controllable Gyroscopic Propulsion Apparatus
Rodgers, A.T.

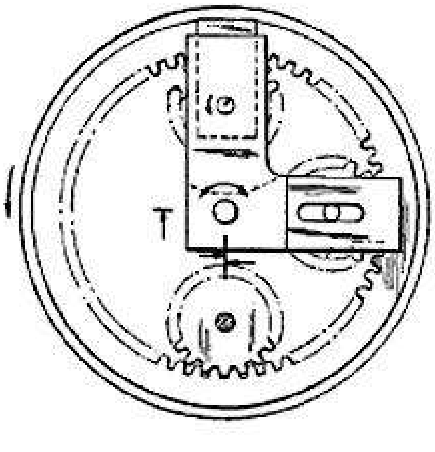

№4,856,358(8/15/89)
Conversion of Rotational Output to Linear Force
Montalbano, Paul J.

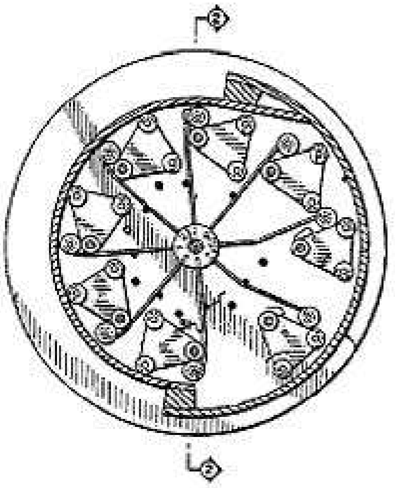

№4,744,259(7/17/88)
Apparatus for Producing a Directional Unit Force
Peterson, Oscar F.A.

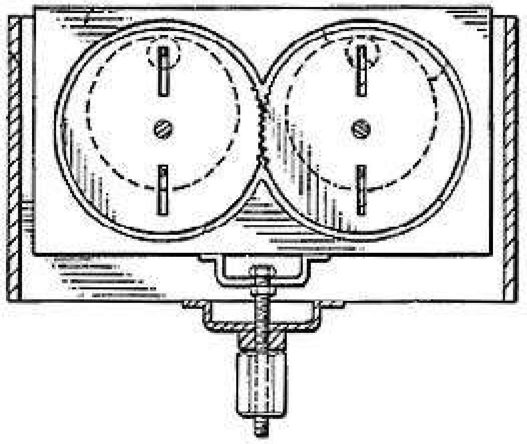

№4,712,439(12/15/87)
Apparatus for Producing a Force
North, Henry.

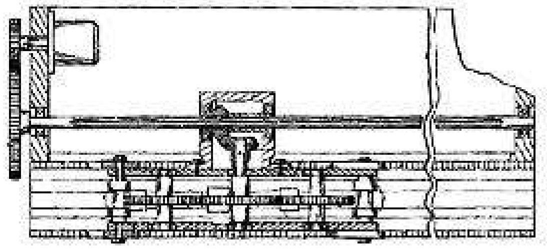

№4,242,918(1/6/81)
Mechanical Propulsion System
Srogi, Ladislaw G.

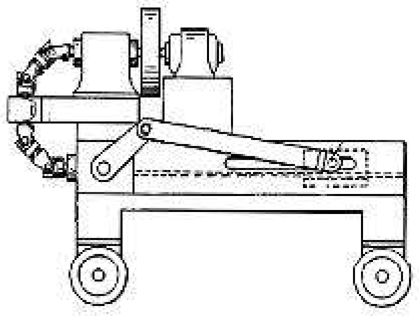

№4,674,583(6/23/87)
Impulse Drive
Peppiatt, Alvin C., et al

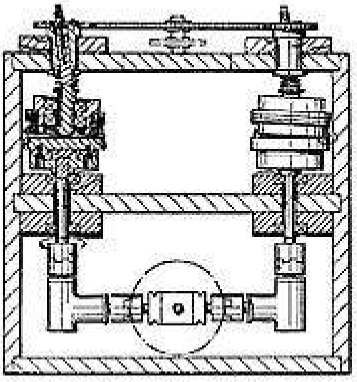

№4,409,856(10/18/83)
Propulsion System
de Weaver, Fred, III.

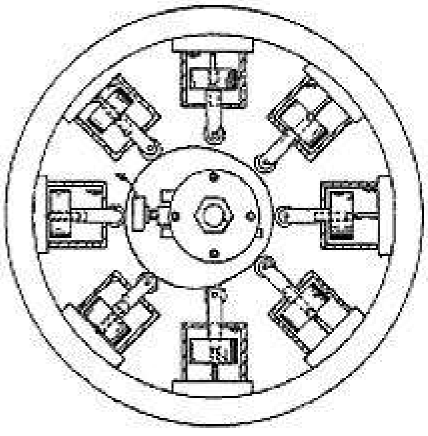

№4,579,011(3/1/86)
Propulsion Apparatus
Dobos, Elmer M.

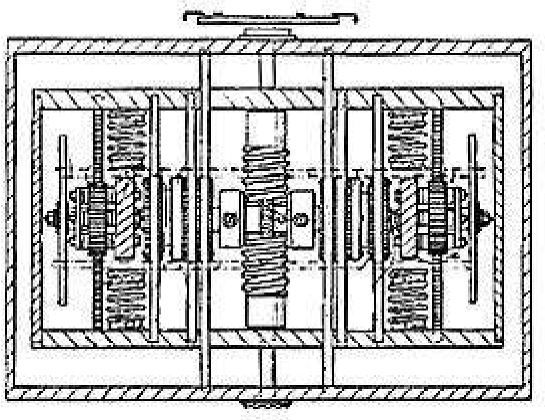

№4,577,520(3/25/86)
Mechanical Propulsion System
Colla, Joseph.

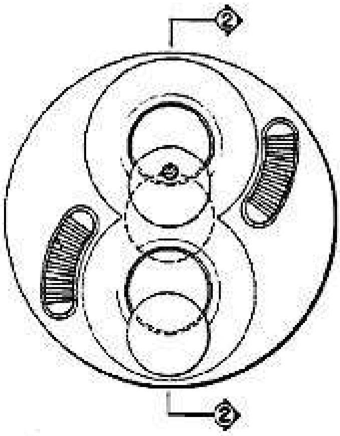

№4,631,971(12/30/86)
Apparatus for
Developing a Propulsive
Force
Thornson, Brandson R.

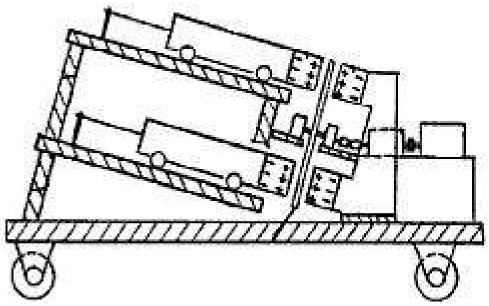

№3,889,453(6/17/75)
Propulsion System
Mast, Oscar.

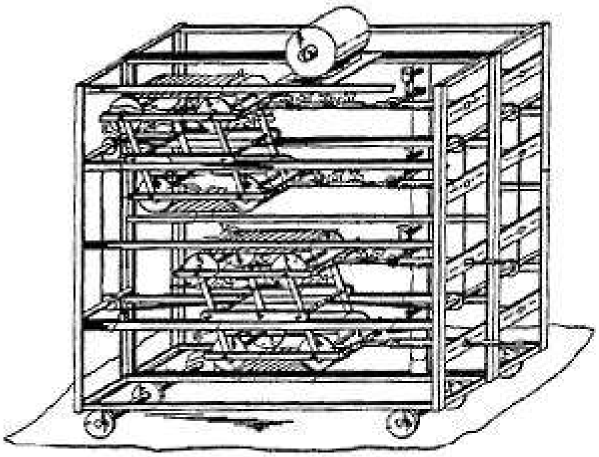

№4,726,241(2/23/88)
Continuous Force & Impact
Generator
Melnick, Harry S.

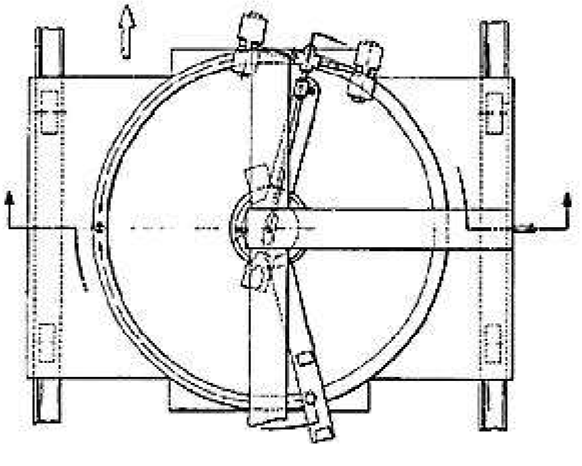

№4,238,968(12/16/80)
Conversion of Centrifugal Force
to Linear Force &
Motion
Cook, Robert L.

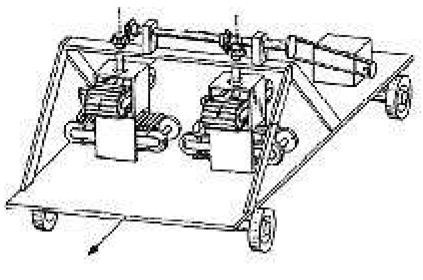

№4,261,212(4/14/81)
Unidirectional
Force Generator
Melnick, Harry S.

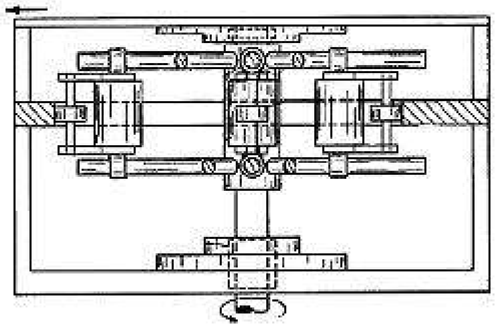

**№ 3,968,700(7/13/76)
Converting Rotary Motion
into a
Unidirectional Linear Motion
Cuff, Calvin I.**

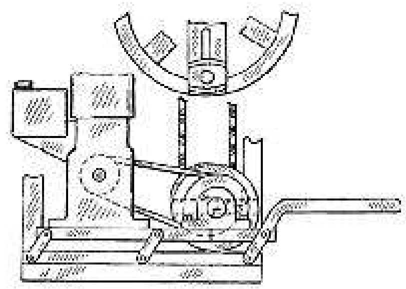

**№ 4,347,752(9/7/82)
Converting Rotary Motion
to
a Rectilinear Force
Dehen, Frederick L.**

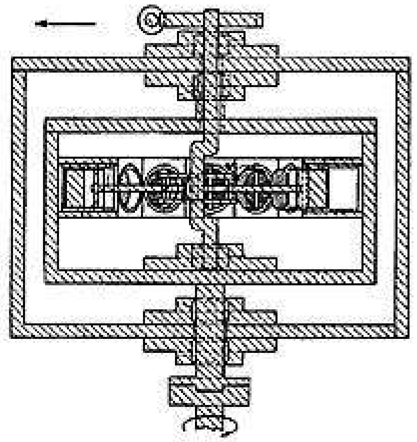

**№ 3,998,107(12/21/76)
Converting Rotary Motion
into
a Unidirectional Linear
Motion
Cuff, Calviv I.**

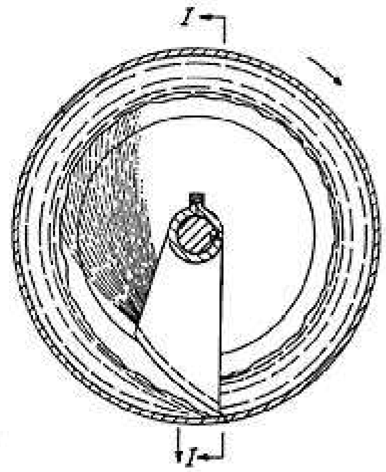

**№ 3,979,961(9/14/76)
Propelling an Object by an
Unbalanced Centrifugal
Force...
Schnur, Nicolas J.**

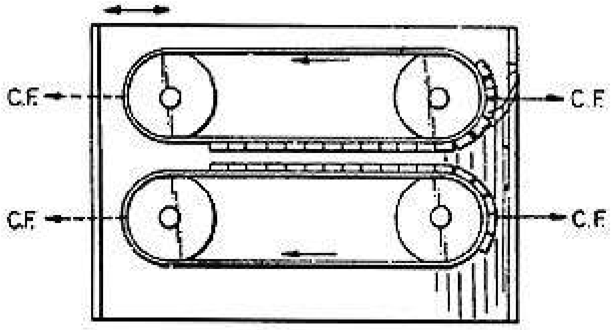

**№ 4,398,431(8/16/83)
Mechanical Power Transmitting
System
Melnick, Harry S.**

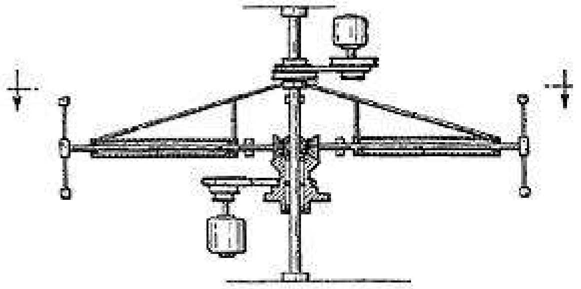

**№ 3,555,915(1/19/71)
Directional Force Generator
Young, Hersey W., Jr.**

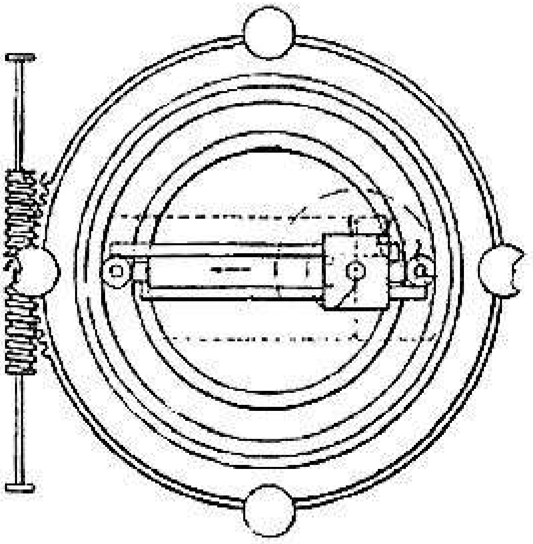

№ 3,750,484(8/7/73)
Centrifugal
Thrust Motor
Benjamin, Paul M.

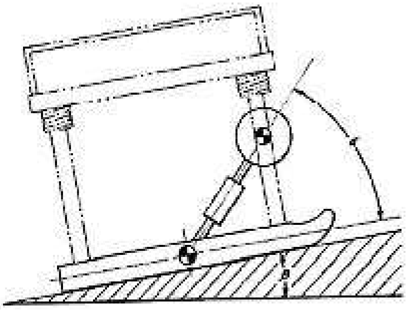

№ 3,916,704(11/4/75)
Vibratory Motion
Gaberson, Howard A.

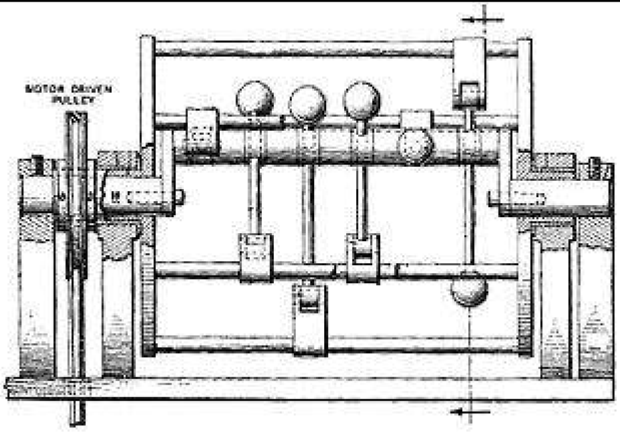

№ 3,810,394(5/14/74)
Centrifugal
Mechanical Device
Novak, Leo J.

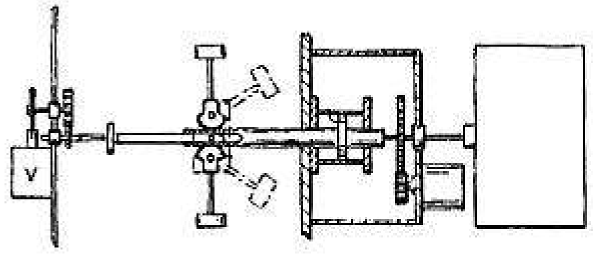

№ 3,863,510(2/4/75)
Inertia Engine
Benson, Everett H.

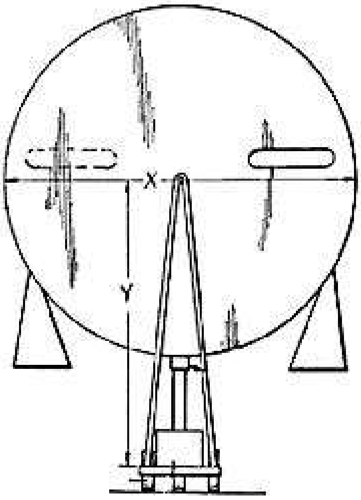

№ 3,756,086(9/4/73)
Propulsion System
McAlister, Roy E., et al.

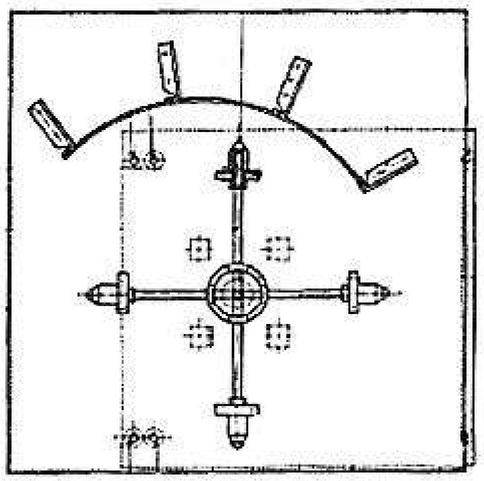

№ 3,807,244(4/30/74)
Device for Transforming Kinetic Energy
Estrade, Fernand.

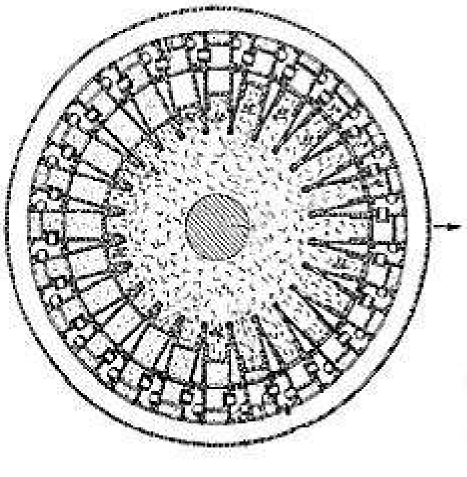

№ 3,584,515(6/15/71)
Propulsion Apparatus
Matyas, Laszlo B.

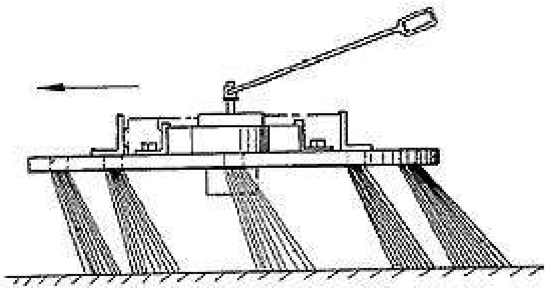

№ 3,196,580(7/27/65)
Toy Vehicle Having...Self-Contained
Drive Means
Rakestraw, Robert G.

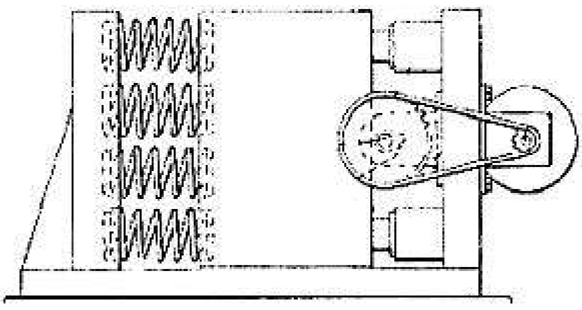

№ 3,266,233(8/16/66)
Inertia Propulsion
Device
Farrall, Arthur W.

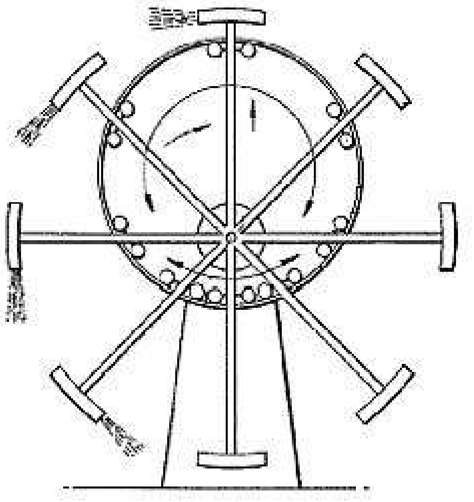

№ 3,177,660(4/13/65)
Propulsion
Apparatus
Haller, Paul.

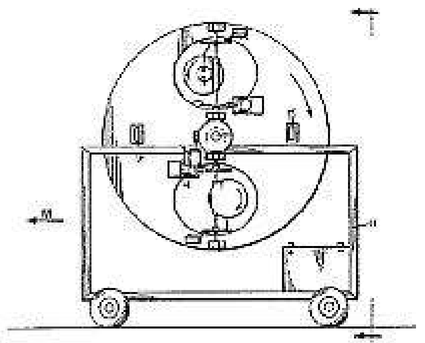

№ 3,653,269(4/4/72)
Converting
Rotary Motion
into
Unidirectional Motion
Foster, Richard E.

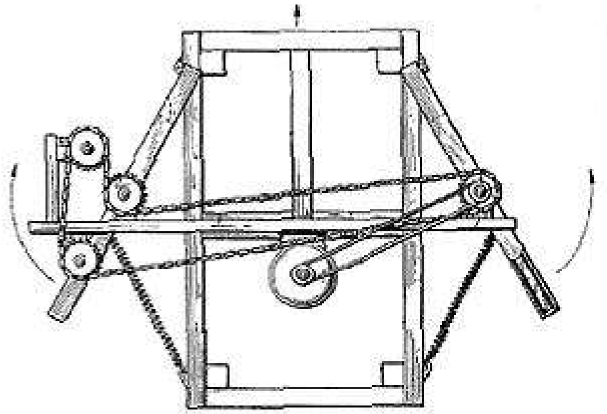

№ 3,683,707(8/15/72)
Propulsion
System
Cook, Robert L.

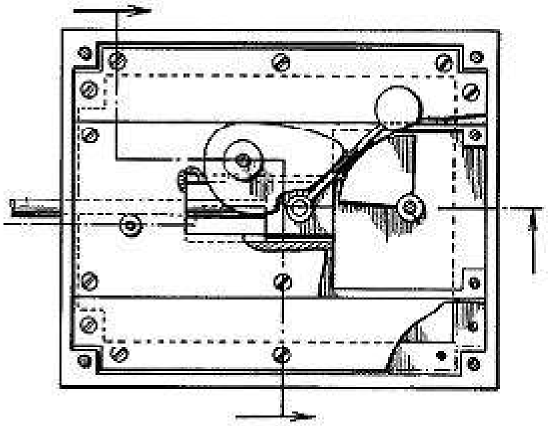

№ 3,182,517(5/11/65)
Variable
Oscillation System
Dean, Norman L.

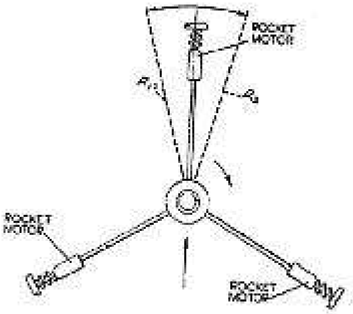

№ 4,087,064(5/2/78)
Orbital
Propulsion
Apparatus
Knap, George.

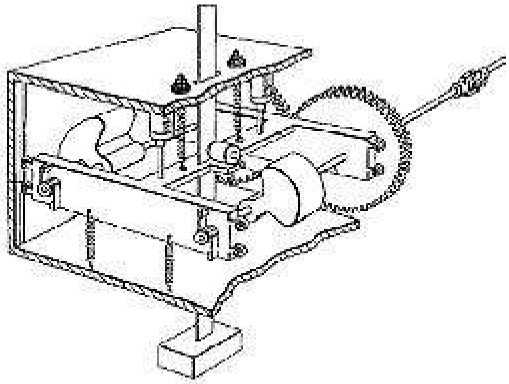

№ 2,886,976(5/19/59)
System for
Converting Rotary Motion into
Unidirectional Motion
Dean, Norman L.

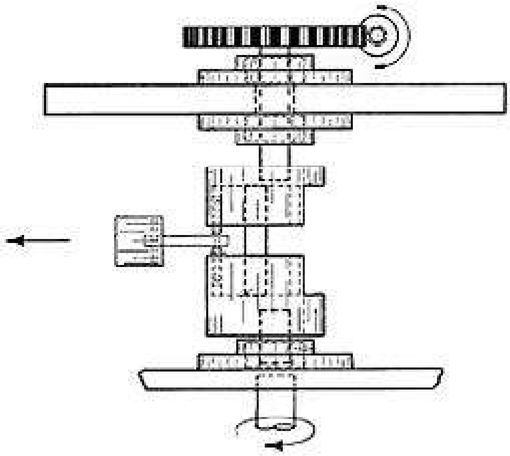

№ 4,095,460(6/20/78)
Converting
Rotary Motion into
Unidirectional Motion
Cuff, Calvin I.

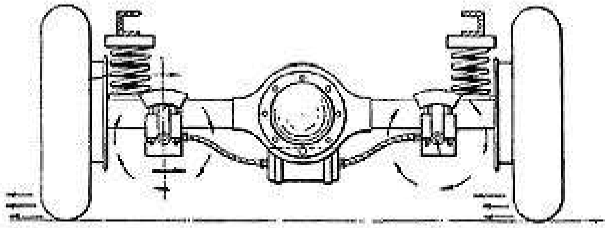

№ 2,639,777(5/26/53)
Method & Device for Imparting
Lateral Movement
Dull, Marshall L.

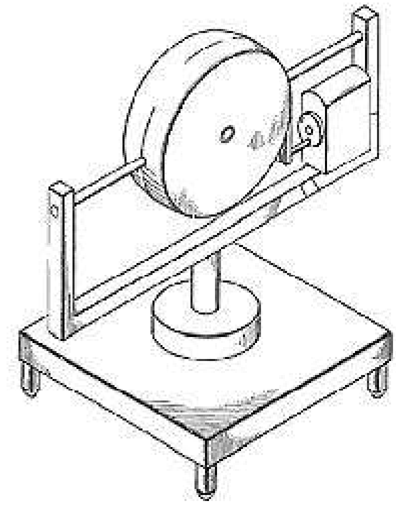

№ 3,203,644(8/31/65)
Gyroscopic
Inertial Space Drive
Kellogg, H. Dudley

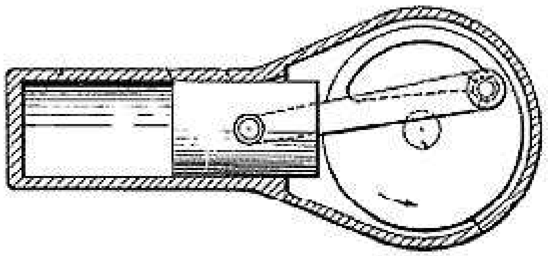

№ 2,636,340(4/28/53)
Direct Push
Propulsion Unit
Llamozas, Juan D. M.

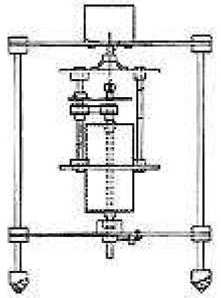

№ 5,150,626(9/29/92)
Translational Force
Generator
Navarro, T. L.

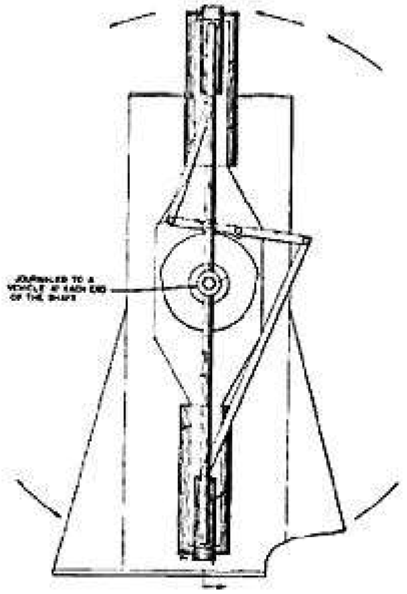
JOURNALLED TO A VEHICLE AT EACH END OF THE SHAFT

№ 3,238,714(3/8/66)
Thrust Motor
Schur, George O.

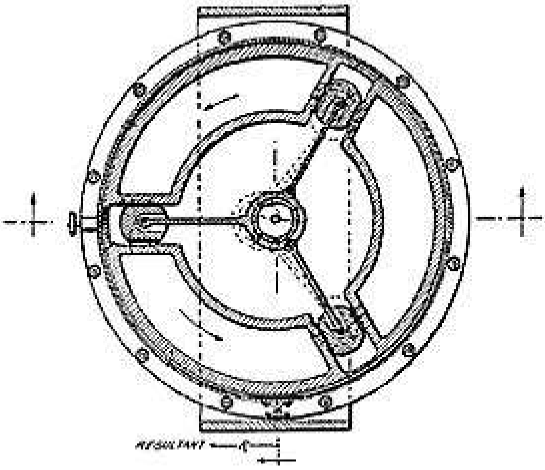

№ 2,009,780(7/30/35)
Centrifugal Variable Thrust
Mechanism
Laskowitz, Isidor B.

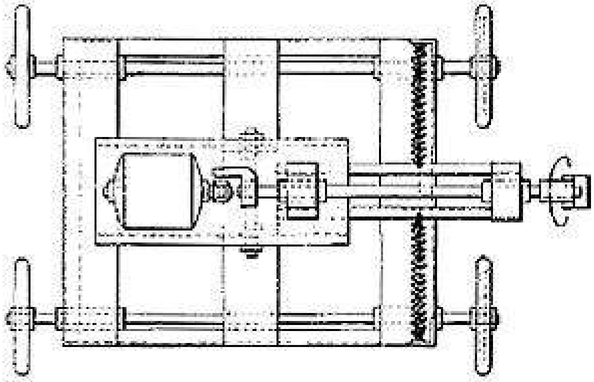

№ 2,088,115(7/27/37)
Reaction Motor
Neff, Tom.

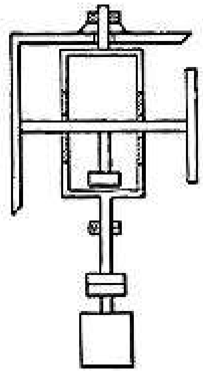

№ 3,404,854(10/8/68)
Apparatus for
Imparting Motion to a
Body
di Bella, Alfio.

The Free Energy Centrifugal Force Experiment

This experiment is conducted by Faraday Lab Ltd
in cooperation with V. I. Bogomolov

In the previous issue we wrote about the invention of Bogomolov V. I. The device shown in Fig.1 (and also on the cover) was constructed and tested in our laboratory.

Fig. 1
The Device

The simplified diagram is known as "Maxwell's Pendulum" (Fig.2). This is a disk set onto the horizontal axis with two attached strings. The upper ends of the strings are fixed to the crossbar.

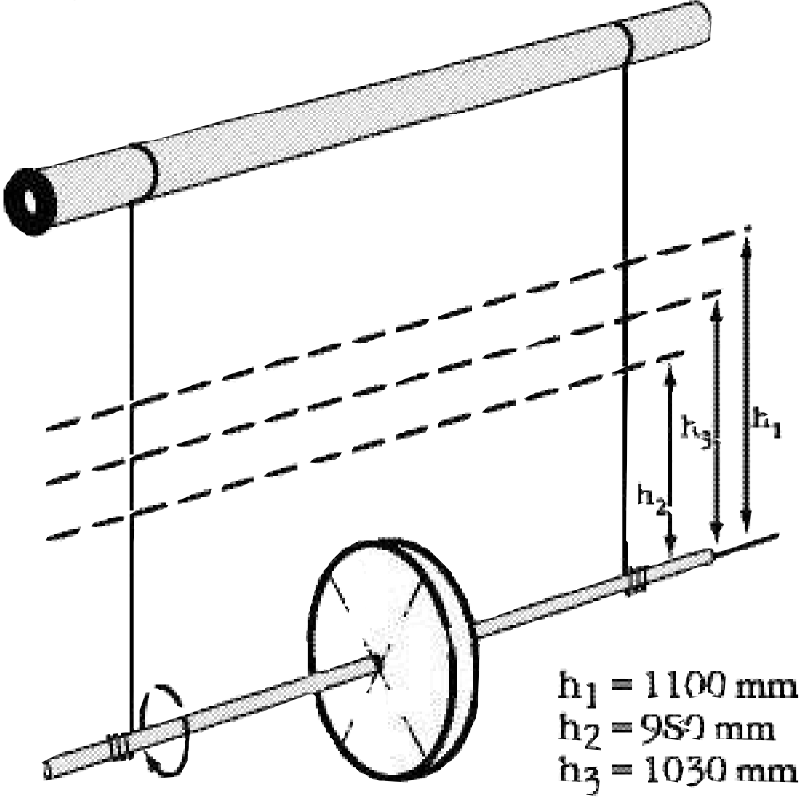

Fig. 2

If we wind the strings about the axis, the disk will lift (the height h) and store the potential energy of the Earth's gravitational field E = mgh (m - mass of the flywheel; g - free fall acceleration; h - the height of the mass fall). If we let the pendulum go, we can observe periodic "up-and-down" damped oscillations: first, the string spins and potential energy converts into rotation kinetic energy; upon reaching the lowest point the disk, still rotating, goes up using rotation kinetic energy and then converts it again into potential energy. This device is interesting: due to the energy conservation law we can observe the usage of rotation kinetic energy of the flywheel, measuring only the h-parameter, the lift height of the flywheel strings in the second semi-oscillation compared to the height with which the pendulum started falling in the first semi-oscillation, the difference h1 - h2 of the two semi-oscillations is directly proportional to the losses of rotation kinetic energy for air resistance work and friction.

We have made "Maxwell's Pendulum" more complicated by substituting the flywheel disk for the Watt centrifugal governor in accordance with its description in the article (1). The main difference of the device from Maxwell's flywheel is that the flywheel inertia momentum changes by the centrifugal force operation at the lever transferring of weights (the total weight - 1200 g) from the minimum radius position (40 mm) to the maximum radius position

(90 mm). At that, the levers compress the string with the force of about 16 N.

There have been three stages of the experiments. During the first stage we define the necessary lift height of 1100 mm at moment when the string winds around the axle. When the flywheel falls from this height, the rotation speed and centrifugal force are achieved, which move apart the loads to the maximum radius (90 mm) and which **completely compress the string**. During the second stage the loads are fixed at the minimum radius position (40 mm), thus the compress string centrifugal forces are not involved any more and we can measure the air resistance and friction **losses of kinetic energy**. At this stage the flywheel lifts up to 980 mm losing 120 mm. During the third final stage the load fixing rods are taken away and ***centrifugal forces are applied***. Just like during the second stage the Watt governor starts spin-falling from the height of 1100 mm. It compresses the string and lifts up to the height of 1030 mm thus **exceeding the second stage height by 50 mm**!

The experiment conclusions of the author:

1. At the third stage the flywheel exceeds the "loss height" h=980 mm. This means that the compression of the string was performed "for free".

2. At the third stage the flywheel converted kinetic energy into potential energy and lifted up 50 mm more. This means that in accordance with the principle of conservation of moment of momentum, the **outer compressed string energy force** has caused the moment of momentum change of the rotating masses, thus giving the additional acceleration to the flywheel and **increasing its kinetic energy "for free"**!

3. For the practical application of the Bogomolov generator in order to generate free energy it is necessary to achieve a high rotation speed (more than 10 000 rpm) and involve centrifugal forces to transfer the rotating mass to the less possible inertia radius difference.

The described centrifugal regulator diagram is not sufficient for larger experiments. For the industrial generator the author has a hydraulic-principle pneumatic string device drawing (know-how).

Editor: Also in this issue: the article on "Glen Gates Motor" by A. Akau. The principle is almost the same. The further implementation of this ideamay soon design purely mechanical power-generating systems, which use inertia and centrifugal forces.

References:

1. Bogomolov V.I. "The Bogomolov Generator", New Energy Technologies, Issue # 4, 2003.

Magnetic Power Inc. (MPI) is developing Self-Powered Generators. Together with its subsidiary, Room Temperature Superconductors Inc., (see the website www.ultraconductors.com), MPI has raised a total of more than $7 million from Angel investors to date.

Due to a pleasant surprise, solid-state electric power generators might be fabricated by modifying off-the-shelf utility hardware. If confirmed in coming weeks, this could result in serial fabrication by this time next year, since large devices of the type needed for conversion are presently manufactured worldwide.

Multiple modules may prove able to replace power plants. Smaller units appear practical for powering homes. Later, optimized designs might replace engines in every variety of vehicle. These generators may make possible very rapid utilization of fuel and pollution free electric power --a revolutionary, new, renewable energy alternative.

Accredited Angel investors can help speed the work needed to bridge into major capital, and accelerate this remarkable alternative. Multi-million dollar funding is on the horizon. Additional information is available privately. We welcome due diligence by qualified parties.

History of Perpetual Motion and Free Energy Machines

Web site http://www.phact.org/e/dennis4.html

Hundreds of people for centuries have attempted to solve the holy grail of energy production - **Perpetual Motion Machines**. Still to this day Eric Kreig gets weekly emails from people who feel they are about to get one working. Eric is very skeptical about free energy. But his collection of information is very interesting. Eric offered a special prize for those who will agree to publicly test their over-unity devices.

♦ **Villand de Honnecourt** 13th century had a drawing of one.

♦ **Leonardo da vinci** made a number of drawings of things he hoped would make energy for free.

♦ Jesuit priest, **Johanes Taisnerius** worked on a magnetic based perpetual motion machine.

♦ **Mimara** in 1518 designed a "self-blowing windmill".

♦ **John Dee** of 16th century reported seeing one - but wasn't allowed a closer look.

♦ **Cornelis Drebbel**, 1610, was an alchemist and magician supposedly made one.

♦ **Robert Fludd** 1630 proposed many machines -people were trying to patent variations of Fludd's device in the 1870's.

♦ **Edward Somerset** 1638 demonstrated many free energy water wheels to the king of France.

♦ **1635** - first of many English perpetual motion machine patents granted. - By 1903, 600 such patents had been granted. Free energy claimants love impressing people with patents.

♦ **Ulrich von Carnach** in 1664 in Germany designed a perpetual ball-moving machine.

♦ Scientist **Jean Bernoulli** (1667 - 1748) proposed a fluid energy machine.

♦ **Bockler** in 1686 made designs for self powered water mills.

♦ In 1712 **Johann Bessler aka Orffyreus** investigated 300 different perpetual motion models and claimed he had the secret of perpetual motion and got much investment money.

♦ **Dr Conradus Schiviers** in 1790 made a belt driven wheel.

♦ **Sir William Congreve** in 1827 tried a machine running on capillary action.

♦ Britisher **Henry Prince** 1866 described the first partially submerged perpetual motion machine.

♦ **Mark Zimara** of Italy had a huge air powered machine that never worked.

♦ **Horace Wickmam** of the USA got a patent to a machine with many balls that just rotate around.

♦ Austrian, **Alois Drasch** patented a machine in the US in 1868.

♦ German, **George Andreas Bockler** proposed 'self operating mills' using variants of Archimedes screws.

♦ **E. P. Willis** of Connecticut made money off a perpetual motion machine in 1870 - people eventually found out a secret source of power to it.

♦ **Charles Redheffer** in 1812 in **Philadelphia** made much money on a perpetual motion machine, tried to restrict scientific evaluation, but was eventually debunked.

♦ Scottish shoemaker **Spence** designed a magnetic based machine, which was debunked.

♦ **John Worrel Keely** of Philadelphia in 1872 (he also had a traveling show of exhibitions). He fooled many scientists with a machine, which appeared to run on water.

♦ **John Gamgee** in 1881 got considerable support for a machine very similar to Dennis Lee's, which used liquid ammonia - it got vaporized from heat readily available, thus expanding it would drive a piston. Gamgee thought the vapor would condense to liquid to start the cycle over again. The Navy appeared to have been fooled and showed it to president Garfield - it never went anywhere Tom Napier agrees that Dennis Lee may have resurrected Gamgee's engine.

♦ **J. M. Aldrich** was arrested for getting investors for his free energy machine in 1899 - he some how was able to avoid conviction and conned many investors. One of whom was finally able to inspect the machine found a hidden spring.

♦ **T.H. Moray** in the 20's demonstrated a "radiant energy device" to many people who were unable to find a hidden power source. Some how, the secret was said to have been forgotten.

♦ **Lester Hendershot** in 1928 got an Army commandant to endorse his free energy machine - but it was later found to have a hidden power source in the motor. His sons believe Lester lost his notes and that maybe they can rediscover how to get it to work.

♦ **Viktor Schauberger** claimed to have discovered some special vortex energy in water. Since he died in 1958, I don't know if his claims have been replicated, but people are still studying his works.

♦ **Mr Papf** in 1966 was a conspiracy believer alternative car engine got a few investors but killed someone during a demo. He tried to blame the problem on an investigating skeptic. He disappeared and became part of urban legend of scores of people, which the big conspiracy has been silencing for generations.

♦ **Guido Franch** was convicted of fraud in 1954 and 1973 of selling rights to distribute little green pills that would convert water into gas. He hid behind conspiracy theory and secrecy to avoid fair testing. A number of people have run this scam and many people still believe there are pill repressed by the oil companies.

♦ **Garabed T.K. Giragossian** in 1917 claimed to have a free energy machine. He was one of the early frauds to hide behind conspiracy theory. Woodrow Wilson signed a resolution offering him protection from some kind of conspiracy. After much fanfare, and delaying tactics his machine turned out to be a giant flywheel, which was charged up with energy slowly and put out a lot of energy for just a second. In spite of lack of proof of anything significant his followers still bothered the US congress for recognition.

♦ **Otis Carr** in 1958 sold stock for a company to manufacture UFO's and free energy machines from Oklahoma. He claimed inspiration from Tesla.

♦ **Edgar Cayce** even babbles about "Motor's with no Fuel".

♦ **McClintock** was claimed to use air as a fuel and had a patent.

♦ The **Evgray** machine scammed many investors (who didn't know how to test 'depleted' batteries).

♦ **Arnold Burke** in 1977 collected $800,000 of investor money (again, mostly from bible believing farmers) for a 'self acting pump'. He tried to hide behind religion. He called his device Jeremiah 33:3' Finally, an open test was done in 1979 and found a hidden source of electricity. His believers (with an infinite supply of denial) still raised $250,000 to get out of a fraud conviction. He still went on making lots of claims with no evidence.

♦ **Robert Adams** in 1977 made all kinds of conspiracy and OU claims in New Zealand. Many still believe in the Adams Motor. - He had the obligatory battles over patents, debates over theory, debates over power measurement and dabbling in other areas of alt physics. He's 80 years old and said to have a scientific education. He has a lot of theories about Aether. Skeptical information.

♦ **Robert Stewart** in 1978 got over $3,000,000 of investment money (much from farmers) for a closed cycle engine using freon rather than water. This 'engine' used the same scheme John Gamgee tried to sell the navy in 1882. -Eric Krieg thinks this is the same approach Dennis Lee's free energy machine works on. I think he just faded away (people claim he was silenced).

♦ **Rory Johnson** of Elgin Illinois, claimed to have invented a cold fusion, laser activated, magnetic motor that produced 525 HP, weighed 475 lbs, and would propel a large truck or bus 100,000 miles. After signing a number of dealers, he moved all his equipment out of his labs, moved to CA and died. (To this day, folks say he was silenced by OPEC).

♦ **RJeseph Maglich** was a physicist and claimed to have a device in 1978, which harnessed fusion power from seawater. They say they put power in and get more power out. I've never heard of them since.

♦ **Howard Johnson** got a patent for a device that claimed to make free energy from a motor like device. As of 11/02, Stephen was claiming he would soon mail out samples of working ones. By 12/12/02, Stephen dropped out of contact.

♦ **Keith Kenyon** had a device claimed to produce more energy than consumed. Calculations seemed to not take power factor (also known as phase angle) into account. It never openly had its output hooked up to its input. Even Dennis Lee tells his followers that all the *over unity motors* (claim to make more electricity than consumed) usually just fool people who can't measure power factor.

♦ **Muller:** Bill Muller and Carmen Muller of Germany raised money on an over unity motor and got a few followers, but never actually demonstrated one working.

♦ **Dennis Lee** Since 1988 has been promising to demonstrate free electricity "in a month or two". He is much like Newman in his mixture of religion & extremist politics, evasion of qualified investigators, endless promises, threatening detractors, etc. He had a Fischer engine, a CRD device and now, and OU motor device. Dennis has invested in Searle and Stanley Meyer and joined Pantone in 2001 in a 50 state tour.

♦ **Stanley Meyer** 1996 claimed to have a water powered car and was also big on mixing Christianity and patriot politics in with fringe science. Meyer was found guilty of fraud after his Water Fuel Cell was tested before an Ohio judge. It is rare for an inventor to be prosecuted for an invention that does not work, but Meyer's problem was that he had been selling "dealerships", offering investors the "right to do business" in Water Fuel Cell technology. Meyer refused to allow anyone else to measure his device. Dennis Lee invested in him. He died in early 1998.

♦ **Joseph Newman** in 1984 claimed to have a free energy machine based on alternative physics. Like many perpetual motion inventors, he sued the US patent office. Many people wrongly measured the true power output of this machine, (they didn't realize you must specially calculate power for non sinusoidal current consumption). He now refuses to ship a unit for testing. Ten years ago, inventor Joseph Newman gave an open week-long demonstration in the Super dome in New Orleans. Over 9,000 people attended from across the country (including Dennis Lee who reportedly wanted to join his ideas with Newman). Newman is suing some former investors he claims are trying to steal his invention. I give more information. Evan Soule of Newman's organization offers a rebuttal.

♦ **Bruce De Palma** had a machine in 1986, which appeared to one Electrical Engineering professor to put out 4 times more power than consumed. Turned out it was just a measurement error. (We could go on and on with such examples).

♦ **Dr. Potapov** sold a device that was claimed to produce more energy than consumed.

Editor: Eric's skeptical remarks can be tolerated but sometimes he writes about the things he does not know. Potapov and his team have created a technology, which have been replicated and sold by dozens of companies in Russia (see the cover picture of the Susorov heat generator). There are several devices in St. Petersburg to be tested. The efficiency is 200 % and more. The operation principle: electric motor rotates in water and causes the heating of water. We have also received the information about another Russian invention, namely the close cycle vortex generator: "water rotation – vapour – turbine – electric power " (the power is from 100 kW).

Alexander V. Frolov

♦ **John Bedini** claims to have a free energy device.

♦ **Mr. Finsrud** is a Norwegian artist who made a sculpture where a metal ball moves for weeks apparently with no outside influence.

♦ **Don Watts** of Las Vegas in around 1990 had a patented **CEACU**, which stands for Centrifugal Energy Amplification and Conversion Unit - it turned out to be one more investment fraud.

♦ **Stephan Marinov** claimed to have proven much alternative physics and to have contacted a Methernitha that claims to have a Free Energy Machine. He committed suicide on July 15, 1997 but he left behind some intense rants.

♦ **Greg Watson** sold kits for a rolling ball and track that were thought to have over unity in 1997. Attempts of replication I know of have failed.

♦ **CETI** These people have claimed to have a device that puts out anomalous yet small amounts of heat - maybe cold fusion. They raised millions. As of 10/97, they have not been willing to have me come over and see for myself. They have said, *we want people to think it doesn't work so we won't have competition.* I've wondered if the energy may have come from not accounting for friction effects from the cooling flow through the pellets. (**Jed Rothwell**, a rational editor of Infinite Energy Magazine says *"however, tests with CETI cells at Motorola, SRI and the French Atomic Energy Commission show no measurable friction"*). Milton Rothman has a response: "I openly admit that I have not followed all the history of cold fusion claims and am generally ignorant on the subject".

♦ In January 1998 **Barbara Hickox** allegedly has a patent dating from 1981 for a fusion powered free energy system.

♦ **Paramahamsa Tewari** of India claims to have a device that is 200% efficient.

♦ **RQM** is a Swiss company selling FE machines found at www.rqm.ch.

♦ **Ted King** is looking for people to buy stock for a car he plans to drive across the country using just 2 12-volt batteries. You can contact Ted if you want to buy shares.

♦ **Bruce Perreault** has claimed to have discovered a new element, the plasmatron, ion pump, radiant energy device, etc.

♦ **Daniel Pomerleau** of Canada claims to have something that works, but he isn't interested in releasing it (as 12/97).

♦ **Entropy Systems** of Ohio 1999 Sanjay Amin got 1.6 million investment dollars for a device that would violate the 2nd law of thermodynamics. I've asked them to apply for my prize for proof. - They seem to have folded. See a rational review of Amin's claims.

♦ **Brian Collin** of Australia claims revelation from God *(like most of these people)* to make a free energy device. A Stephen Mark claimed he invented it.

♦ **Kawai** and **Takahashi** - both of Japan claimed in the 90's to have overunity devices.

♦ **Carl Cella** claimed to have one of the many cars said to run on water.

♦ **Robert LeBreton** in 1999 claimed to be making a 600 hp free energy machine.

♦ **Renzo Boscolio** in 1999 in Italy claimed to have low-energy nuclear reactions but refused to supply real proof that he promised to people who came out from Infinite Energy Magazine.

♦ **Doug Konzen** of Seattle says he has an overunity motor in Jan of 2000 that anyone can see.

♦ **Troy Reed** of Oklahoma was ready to issue licenses for manufacturing his permanent magnet motor etc. He's taking investment money for a device he claims puts out more heat energy than input energy. He says he's dumped a few million into his designs over the last 10 years and had little interest in my prize offer when I contacted him in 2001.

♦ **Kipper Motor.** In 10/00, Steve Elswick thought it was overunity. But David Sligar who paid $175 for plans could not get it working.

♦ **Ludwig Brits and Victor Christie** in 2001, claim the **Lutec** free electricity over unity motor/generator will soon be in Australia.

♦ **Jasker** in 2002, an Irish company has claimed to have free electricity. (Some have said it is just a joke).

♦ **Confidential Technology -Wayne Cochran** died 12/29/02 Crazy Jack Carey took over. They have been promising FE soon for about 20 years now.

♦ **Tom Bearden's MEG device:** many articles on the Internet.

♦ **Gurbakhsh Singh Mann of India** claims to have invented gravity and buoyancy perpetual motion machines.

♦ **Michael J. Marshall** in Las Vegas has a device called QSFG , which stands for quick start fuel-less generator he says 64 nations have asked him to build factories.

♦ **Carl Tilley and Robert Kibbey** in 6/2001 in Tennessee have claimed some over unity device and other stuff. Their demonstration failed in 9/02 and as of 11/02, they have been evading proper demonstrations.

♦ **Stephen Walker** in 9/2002 promised to send me and several others a free energy machine in mail.

♦ **PerEnDev** promised to make some kind of free energy by means of magnets.

♦ **Bill Muller** of Canada in 2003 claimed to have some kind of OU device. Independent tests found it under unity.

♦ **Energie:** In 2003 this Greek company promised a homopolar magnet (De Palma design) FE design by June of 2003. They actively seek investors.

♦ **GWE Genesis World Energy** In 2002 this group claimed to have 400 people who developed some device that sounds like it separates water into H2 and O2 using less energy than mainstream science says is possible. As of 2003, they have evaded independent confirmation. More information is available.

♦ **Steven Greer (Disclosure Project)** In 02.2003, Steven announced he had discovered some real sources of Free energy, which he promised to make sure get proven to the general public.

Tesla Energy Science Conference and Exposition

1220 L Street NW, Suite 100-232 Washington DC 20005
Tel: 800-295-7674
Fax: 301-513-5728
www.IntegrityResearchInstitute.org
Email: iri@erols.com

Saturday, November 8, 2003
Wireless Electricity Seminar

9:30 AM 10:15 AM

Opening Address: “Space Solar Power” – Dr. **Paul Werbos** Program Director, National Science Foundation

10:15 AM 12:15 PM

Special 2-Hour Presentation: “Nicola Tesla and the development of RF Power Systems” - Dr. **James Corum**, Physics Professor, research Scientist and Inventor and **Kenneth Corum**, Physicist, teacher, consultant.

2:00 Pm 3:00 PM

“Power Engineering Scalar Field Theory: Faraday vs. Maxwell and Longitudinal Wave Demonstration” – Professor **Konstantin Meyl**, Engineer, Author, Inventor of the Demo-Set, Professor at the University of Berlin

3:15 PM 4:15 PM

“Wireless Energy Through the Earth-lonosphere Cavity” – Dr. **Elizabeth Rauscher,** Nuclear and Astrophysicist, Inventor of the ELF Earthquake Predictor and Triangulator.

4:15 PM 5:15 PM

“Masters of the lonosphere HAARP Modifies the Polar Electrojet” – **BBC Video on Tesla Technology**, with Wiliam Terbo, demonstrating ionospheric excitation by radio waves.

6:30 PM 7:15 PM

Masters of the Ionosphere – BBC Video Presentation **repeated** for the Special Tesla Evening Event attendees

7:30 PM 8:30 PM

“The Wardenclyffe Dream: Tesla’s Plan for Wireless Worldwide Distribution” – **Dr. Marc Seifer** Professor, Author of the best-selling book *Wizard, The Life and Times of Nikola Tesla*, presents an illustrated historical account.

8:30 PM 9:00 PM

High Voltage Tesla Coil Demo – **High Voltage Device** with **Music**: The Tesla Tower wind ensemble composition by Prof. **Holland**, Skidmore College

Sunday November 9, 2003
Electrotherapeutic and Tesla History Day

9:30 AM – 10:30 AM

“A Family Perspective on the Personality of Nikola Tesla. Review of the Popular Interest in this Scientific Icon” with Q and A Session – **William Terbo**, Engineer, Closest living relative of Tesla (Grand-nephew). Founder and Director of the Tesla Memorial Society.

10:30 AM 11:30 AM

"The Search for Free Energy: Genius, Vision and Psychology of Invention" – **Keith Tutt**, British Writer and Author of the book *The Scientist, The Madman, The Thief and their Lightbulb* that includes Tesla.

11:30 AM 12:15 PM

"Electrotherapy with Tesla Coil Design: Intorduction to Bioelectromagnetics" – Dr. **Thomas Valone** Physicist, Professional Engineer, Author of the new book, Bioeletromagnetic Healing: A rational for its Use.

12:15 PM 1:00 PM

Exhibitors Presentations – **Brief Presentations by exhibitors** (5-10 minutes each)

2:00 PM-3:00 PM

"Emerging Opportunity: cancer Electromagnetic Frequency Therapy" – Dr. **Mark Neveu** President, the National Foundation for Alternative medicine.

3:00 PM 4:00 PM

"Turn of the Century Electrotherapy Discoveries" – **Jefrey Behay** Director, The Turn of the Century Electrotherapy Museum.

4:00 PM 5:00 PM

"How a Crushed Leg Helped Me Discover High Voltage Electromagnetic Healing Device" – **Ralph Suddath**, Third generation Tesla Electrotheraphy Inventor, radio Host, Entrepreneur.

5:00 PM 5:15 PM

Closing Remarks – **Thomas Valone** Program Coordinator

The Conference Report Compilation can be purchased from T. Valone, Integrity Research Institute.

"Nornikel" and Alternative Energy

"Norilsky Nikel", the largest Russian non-ferrous metal producing company has announced the beginning of the cooperation with the Russian Academy of Science. Metallurgists are planning to invest into the development of the hydrogen energy and fuel cell R&D projects of the Russian Academy of Sciences. Accordingly, last week an agreement was signed. **Mikhail Prohorov**, the General Director of "Nornikel" said that they are ready to invest 20-40 million dollars annually into R&D projects of the Academy in this field.

"Expert" magazine, # 43, November 17-23, 2003

PERENDEV Magnetic Motor

Review

www.perendev.co.za/products/magnetic_motor/index.html

Shown below is a Magnetic Motor with Alternator (Patent Pending):

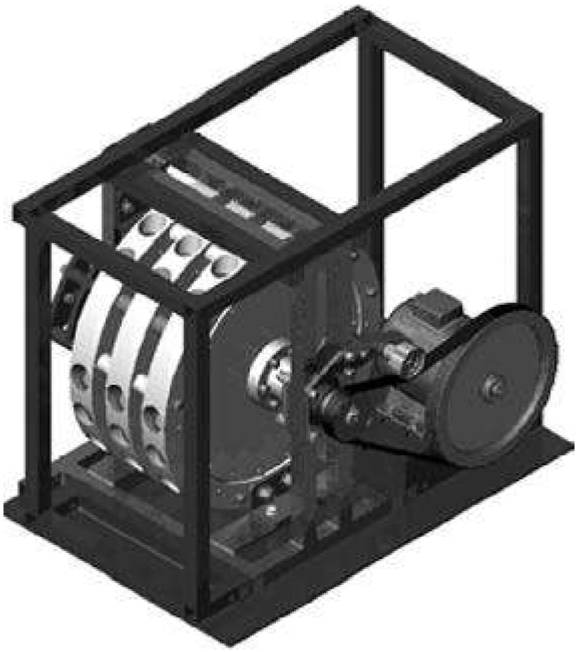

Fig.1

This is the "PerEnDev" (Perpetuum Energy Devices) company latest motor predicted to deliver some 60 kW. Motors can be built up to 4 mW.

Background of the Magnetic Motor

The Company started in 1969 with the development of a magnetic motor to drive an alternator to produce electricity. The concept was sound but the problem they faced was that the magnets that were available at the time (ferrite magnets) were not very powerful. The unit ran but had very little power and the project was abandoned.

In 1996 the "Perendev" company started to investigate the possibility of reviving the original idea, magnets had now come a long way and were very powerful (rare earth).

Fig. 2
Rare earth magnets

After the engineers successfully designed a new model, the company developed a 3 rotor system producing 6 kW, coupling this to an alternator through an 11:1 gearbox to produce electricity in 220volt and 380volt, single or 3 phase.

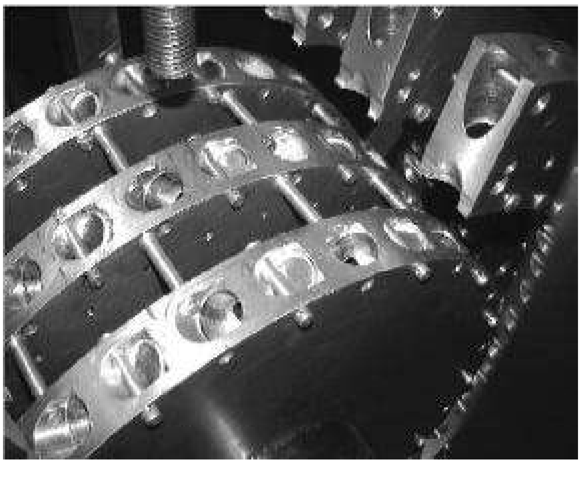

Fig.3

The units will be ideal for the small and large user, such as emergency, military, farms, plots, industrial and large users... in fact any application requiring power.

Key Benefits Foreseen:

- No fuel costs
- Constant running
- Reliable
- Cost Effective
- Portable (6kw-120Kw)

Editor: In 2002 "Faraday Lab" got an offer from the company to buy their 6 kW generator for testing at the price of about $6,000. However, no contract was signed. The company is currently engaged in redesign of their motors and is looking for regional partners but they are not yet ready to produce on the line. In our issues we will try to inform the readers on their activities. Alexander V. Frolov

Internet Pages

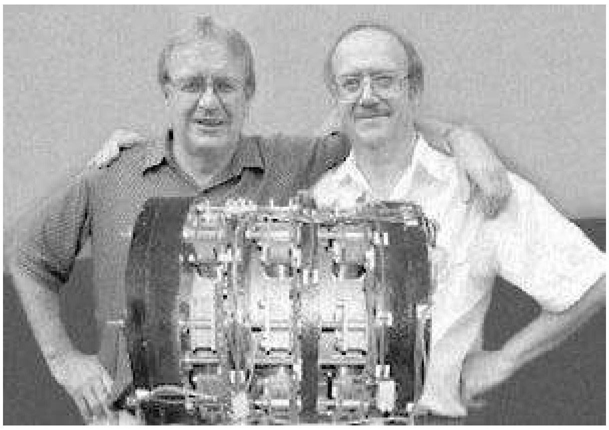

Lutec 1000

Jasker Power System

Electric Radial Motor

industrial and large users... in fact any application requiring power.

Key Benefits Foreseen:

- No fuel costs
- Constant running
- Reliable
- Cost Effective
- Portable (6kw-120Kw)

Editor: In 2002 "Faraday Lab" got an offer from the company to buy their 6 kW generator for testing at the price of about $6,000. However, no contract was signed. The company is currently engaged in redesign of their motors and is looking for regional partners but they are not yet ready to produce on the line. In our issues we will try to inform the readers on their activities. Alexander V. Frolov

Internet Pages

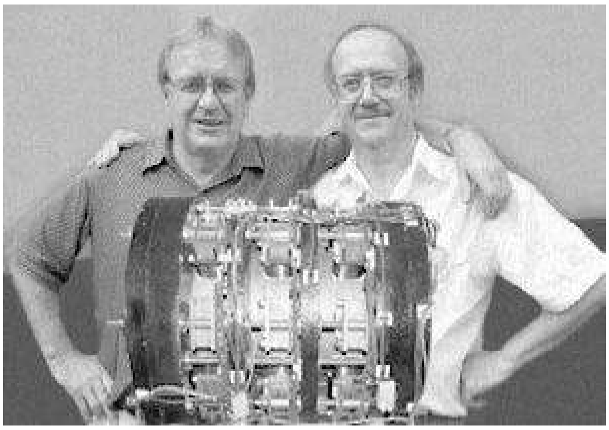

Lutec 1000

Jasker Power System

Electric Radial Motor

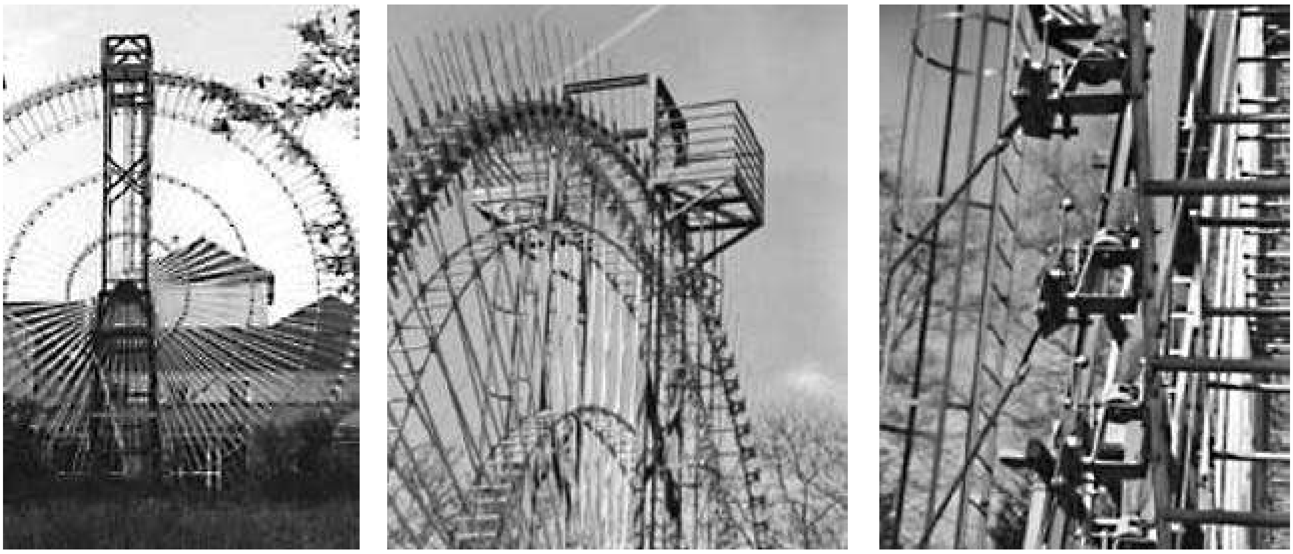

18 Meter Perpetual Wheel

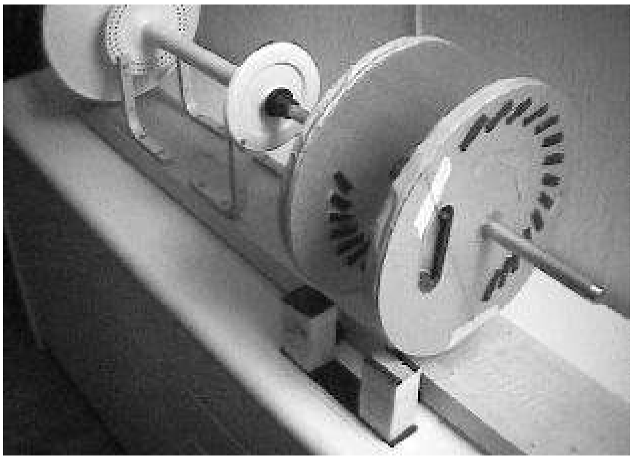

Minato Wheel, E. Vogel, Sweden

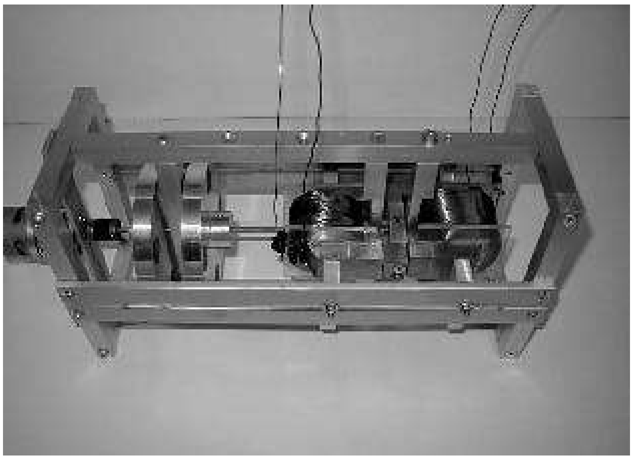

Generator from James W. German

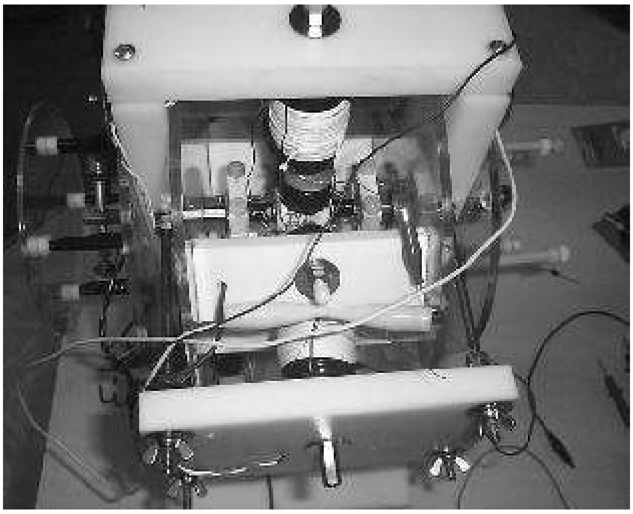

Konzen
Pluse Motor

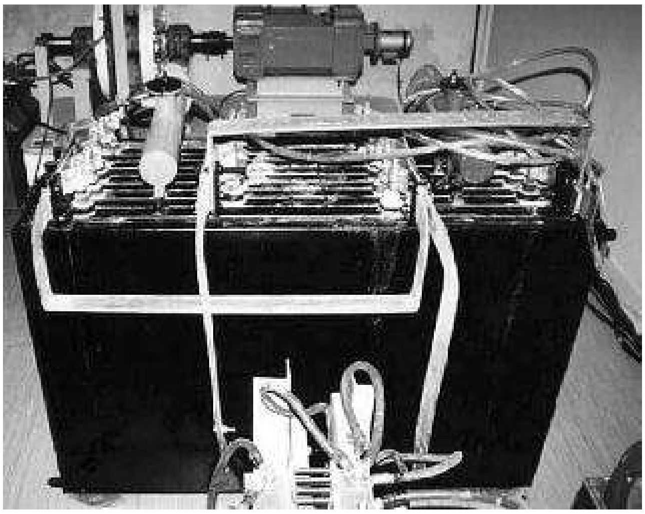

200-300 percent over-unity energy Space
Power Generator

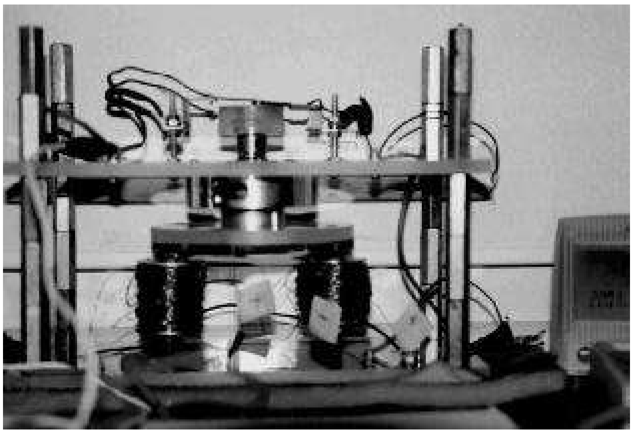

Adams Motor, Robert Adams

The RotoVerter

Bendini Pulse Motor

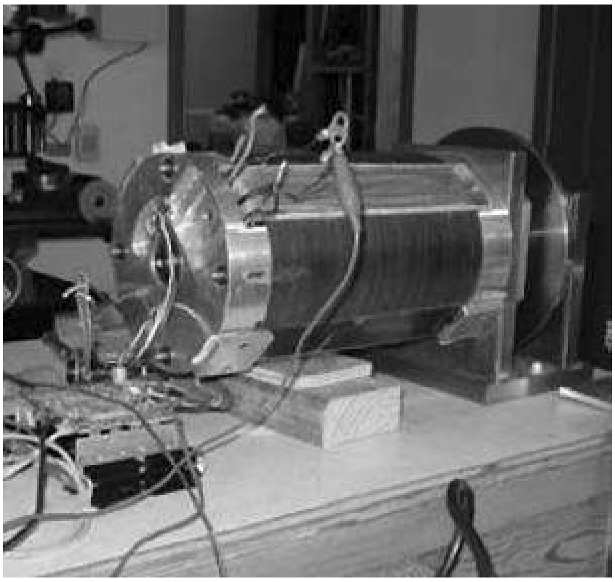

Motor based on Flynn Technology

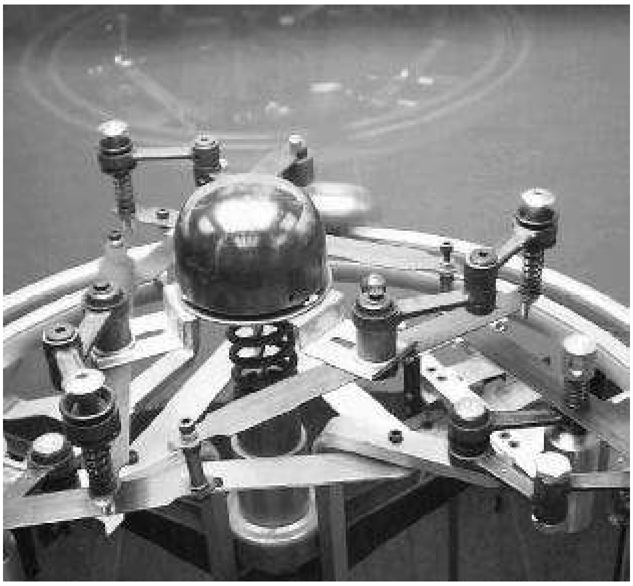

Finsrud Device

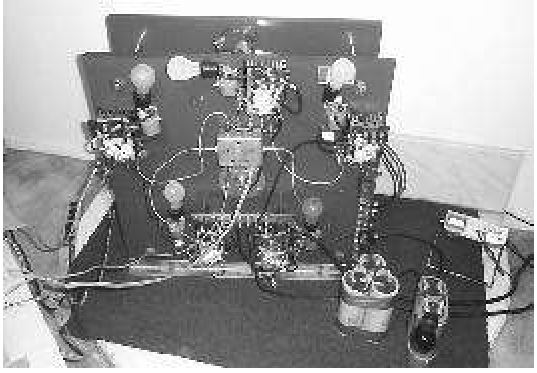

The Muller Unit

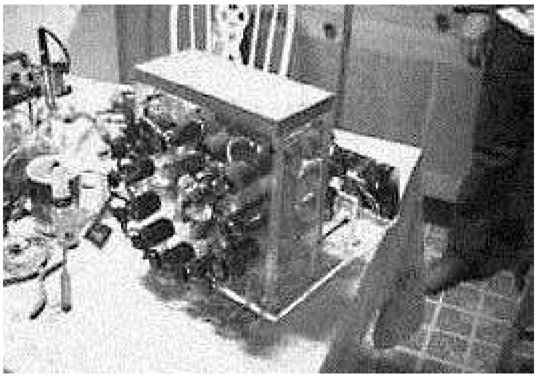

Alte Olsen Generator

Vladimir Matveev's Electrical Generator

Entirely New Kind Of Generator Invented!

(1-26-2003) TASHKENT, Uzbekistan (UPI) – An Uzbek inventor said he has created a type of electrical generator that does not rely on the principle of electromagnetic induction – on which all existing generators are based.

The new generator employs a concept called **magnetic conductivity modulation** and it has potential applications in industry, communications, households and even the military, explained Vladimir Matveev, the inventor, a specialist in electronics.

Matveev said he is convinced he has created a fundamentally different machine.

"All electrical machines I know are based on the principle of interaction between the magnetic fluxes (lines of force) of their rotor (rotating member) and stator (portion that remains fixed)," he explained.

Such machines, Matveev said, are based on electromagnetic induction, a property of energy discovered by Michael Faraday, an English physicist and chemist, in the 19th century. The machines produce electrical current either by moving a conductor across a magnetic field or by regulating the flux of that field.

"My machine has a principal difference," Matveev told United Press International. "The magnetic field of its stator does not interact with the magnetic field of the rotor (because) its rotor is not a magnet – the rotor only changes the magnetic resistance of the stator," he said.

The stator in Matveev's generator contains a magnetic core with a permanent magnet and a detachable winding. A rotor with changeable magnetic resistance is placed at a cutoff point in the core's magnetic field. It is composed of alternating magnetic and air parts and can operate in either linear or rotary form.

When the rotor is set in motion, its alternating components pass through the magnetic core's cutoff point. When the magnetic part passes through the cutoff point its magnetic resistance decreases. When the air part passes through, its resistance increases.

This pulsing of resistance results in changing the magnetic conductivity of the magnetic core, which in turn produces an alternating electrical current in the core's winding. The frequency of the winding's current can be controlled by regulating the rotor's speed or by changing certain qualities of its magnetic or air parts. Also, the generator's electrical

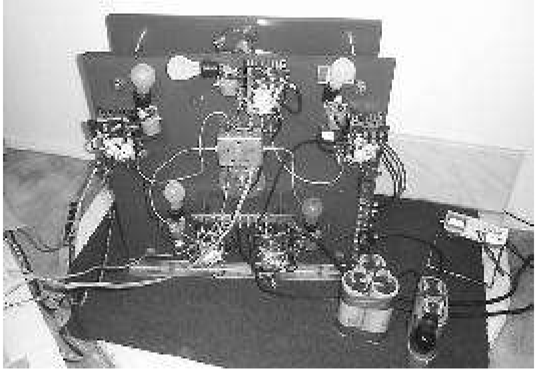

The Muller Unit

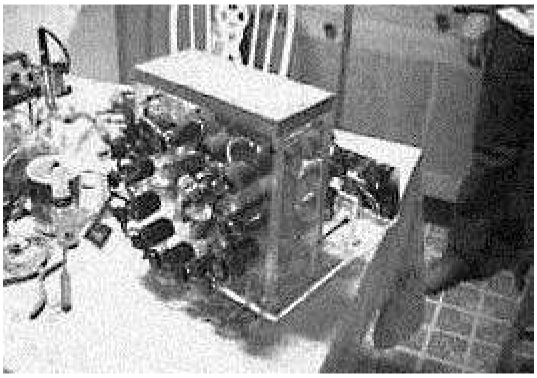

Alte Olsen Generator

Vladimir Matveev's Electrical Generator

Entirely New Kind Of Generator Invented!

(1-26-2003) TASHKENT, Uzbekistan (UPI) – An Uzbek inventor said he has created a type of electrical generator that does not rely on the principle of electromagnetic induction – on which all existing generators are based.

The new generator employs a concept called **magnetic conductivity modulation** and it has potential applications in industry, communications, households and even the military, explained Vladimir Matveev, the inventor, a specialist in electronics.

Matveev said he is convinced he has created a fundamentally different machine.

"All electrical machines I know are based on the principle of interaction between the magnetic fluxes (lines of force) of their rotor (rotating member) and stator (portion that remains fixed)," he explained.

Such machines, Matveev said, are based on electromagnetic induction, a property of energy discovered by Michael Faraday, an English physicist and chemist, in the 19th century. The machines produce electrical current either by moving a conductor across a magnetic field or by regulating the flux of that field.

"My machine has a principal difference," Matveev told United Press International. "The magnetic field of its stator does not interact with the magnetic field of the rotor (because) its rotor is not a magnet – the rotor only changes the magnetic resistance of the stator," he said.

The stator in Matveev's generator contains a magnetic core with a permanent magnet and a detachable winding. A rotor with changeable magnetic resistance is placed at a cutoff point in the core's magnetic field. It is composed of alternating magnetic and air parts and can operate in either linear or rotary form.

When the rotor is set in motion, its alternating components pass through the magnetic core's cutoff point. When the magnetic part passes through the cutoff point its magnetic resistance decreases. When the air part passes through, its resistance increases.

This pulsing of resistance results in changing the magnetic conductivity of the magnetic core, which in turn produces an alternating electrical current in the core's winding. The frequency of the winding's current can be controlled by regulating the rotor's speed or by changing certain qualities of its magnetic or air parts. Also, the generator's electrical output – its voltage – can be controlled by changing the configuration of the rotor's components, Matveev explained.

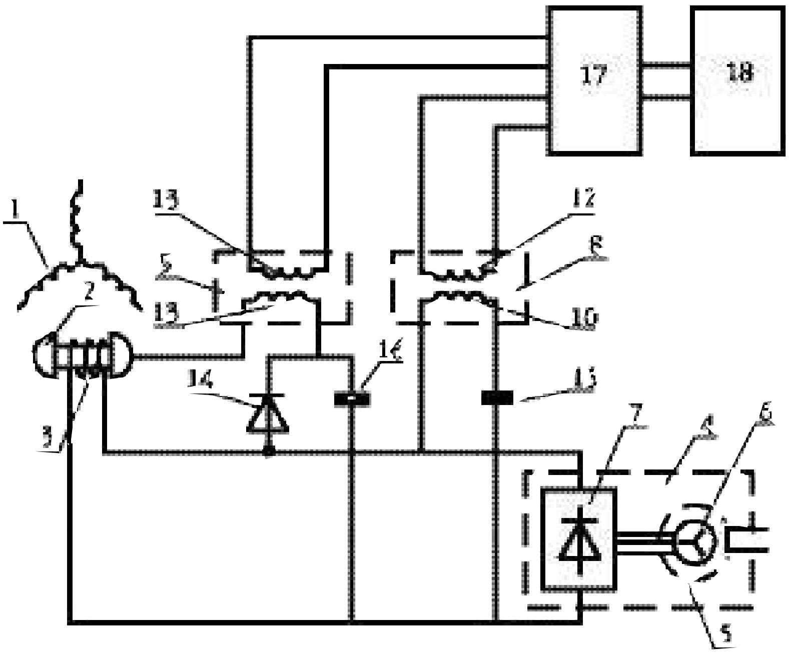

Electrical machine assembly - FEDORO V F 1996.10.10 1996RU-119822 (1998.09.27) G01R 31/34, 31/02

Addnl.Data: VOROBEI V K MATVEEV V A

NOVELTY - Device has electrical machine, non-brush field exciter with diode rectifier, two dynamics double-winding transformers, which primary windings are located on rotor and secondary windings on stator. One end of primary winding of second transformer is connected to rotor housing; another end is connected to common point of conductor which is connected between capacitor and cathode of semiconductor diode, which other terminals are connected to direct current terminals of diode rectifier. Semiconductor diode is connected to cathode group of rectifier; capacitor is connected to anode group. Secondary windings of transformers are connected to actuating member through generator of signal, which is proportional to resistance of insulation in excitation winding circuit.

USE - Electric power production

Matveev said his generator is different from an invention by Howard Johnson of the United States. Johnson discovered how to build motors that can run without input of electricity or any other kind of external energy. He obtained a patent in 1973 for describing electrical generation using only the energy contained in the atoms of permanent magnets.

Matveev also said some Russian inventors have experimented with a generator similar to his. Their generator changes its magnetic conductivity by changing temperature. However, the machine requires a lot of time to be heated and cooled alternately and results in a current frequency much lower than what generally is used in industry. Moreover, the Russian generator requires high steel density and greater mass.

Matveev's machine generates electric energy of industrial frequency. Furthermore, he said his machine is simple, reliable and requires less steel and mass than conventional generators. It also can be adapted to flows of low speeds, such as weak water or wind streams. Matveev tested the generator in his former household in Kazakhstan before he patented it in Uzbekistan.

"I want to pass the invention on to all mankind," he said.

Boris Abdurakhmanov, director of the Uzbek Koinot (Cosmos) design office and head of the laboratory of semiconductors and photoelectricity of the Institute of Electronics of the Uzbek Academy of Sciences, told UPI: "Matveev has offered a fundamentally new approach to a problem of the creation of electric power generators."

Utilization of the Gravity Force in Power Engineering

Editor: In the previous issue we made a review of Mikhail Smeretchanski's invention "Perpetual motor with magnetic control elements" (# 4, 2003). This invention was patented; shown below are some pages from the patent itself. You can contact the author by e-mail: smeretchanski.mikhail@wanadoo.fr, or the postal address: 13 av. Rochambeau 3800 GRENOBLE FRANCE

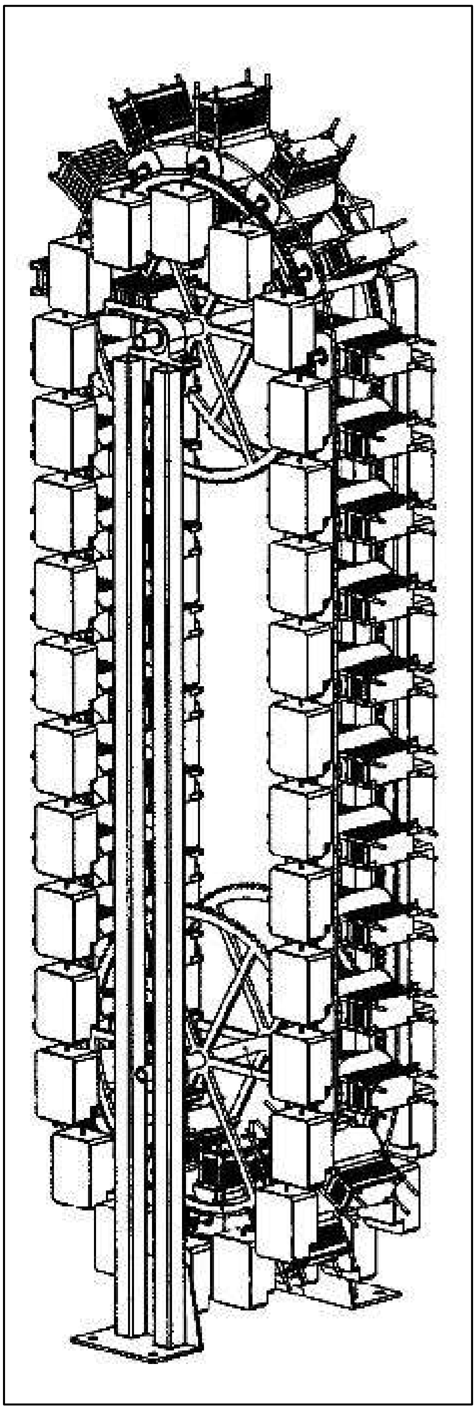

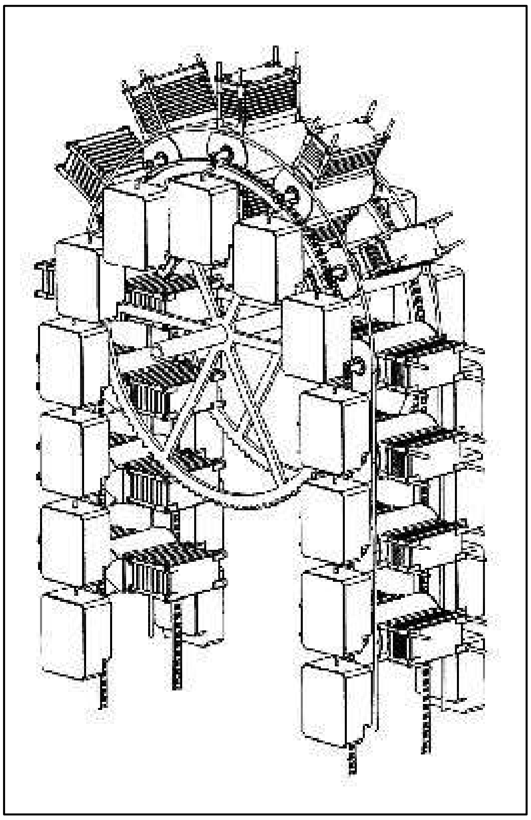

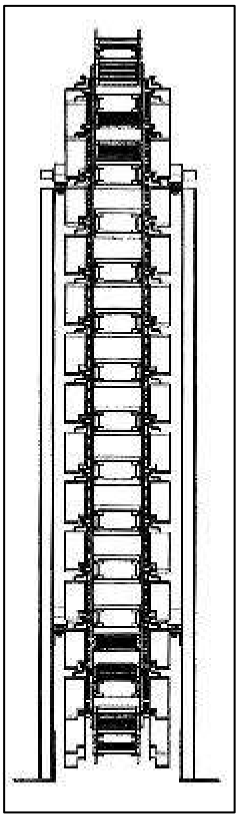

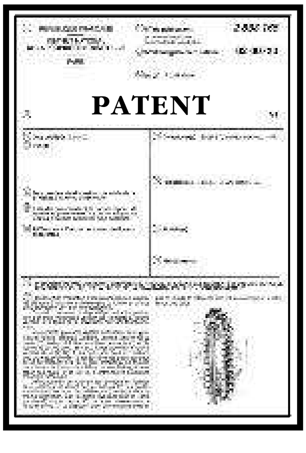

PATENT

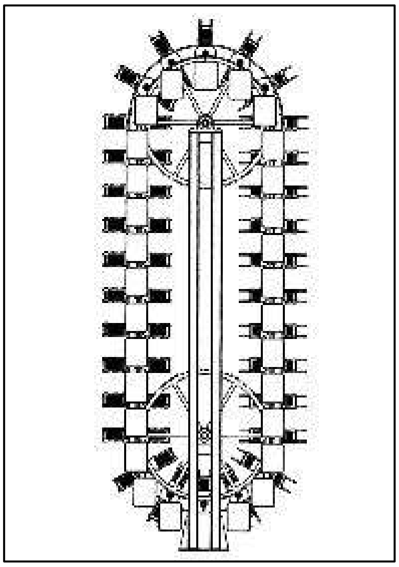

The Most Interesting Articles - 2003

In the year 2003 we have published more than 120 articles, mainly on the alternative energy, new propulsion principles and the development of ether dynamics.

"Cold fusion", low energy nuclear reactions (LENR) and highly-efficient electrolysis were represented by the works of Professor Kanarev, E. Storms, P. Hagelstein, T. Chubb and E. Cartlidge.

Almost every "New Energy" issue features articles on highly-efficient permanent magnet motor or generator experiments or patents of the kind. The articles by T. Valone, E. Vogel, A. Akau, T. Hardwood, A. Francouer and S. Abramov are of great value for engineers and researchers in this field.

The interesting articles by M. Filo, R. Koontz and V. Bogomolov reveal several principles of designing and building purely mechanical energy generating systems. All these systems possess one common feature: namely, the fact that a substance (a solid or fluid working mass) is accelerated in the gravitational field or in the centrifugal force field. The design characteristics of the system make it possible to obtain free energy.

We have published several articles by S. Gerasimov and his colleagues on the reaction-less propulsion device experiments. We hope that in the future we will keep in touch with these researchers.

The article by D. Reed on the new physics development concepts is worth mentioning too. The article by E. Sorokodum on new energy sources and the article by V. Chernobrov on the research into the active properties of time are of no less importance.

A. Egorov's article on the ball lightning makes it possible to form a number of experimental approaches to the problem of creation multi-purpose stable plasma objects.

Electrogravity questions, water-powered vehicles, longitudinal electromagnetic wave technologies, the Searl effect commercialization, capillary engines, resonance high-efficiency power-engineering, single-wire power lines, heat pumps and other similar research works were presented to our readers in the year 2003. We have minimized theoretical articles and we hope that the magazine has become more interesting and useful for the new energy practice development.

Alexander Frolov
St. Petersburg, Russia

There is interesting newsgroup of the site: http://www.overunity.com,
Stefan Hartmann, email: harti@harti.com

Stefan Hartmann writes: This is the answer I got from Finsrud; it seems that the film they have done about him will be shown [on Discovery Channel] in 2004 when [if] I interpret this norwegian text right...

Subject: Documentary Film
Reidar Finsrud,

Dear Sir,

The Australian Broadcasting Company together with The Discovery Channel, have funded my company to make a documentary film entitled 'A Machine To Die For', the story of Perpetual Motion and the search for 'Free Energy'.
I have read extensively about your Perpetual Motion sculpture and we would be interested in traveling from Australia to Norway in order to film this now famous piece of work. At the same time we would request an interview with yourself.
If you are in agreement with this it is our intention to be in Norway on the 10th and 11th of July 2003.
This documentary will be released worldwide and should produce considerable interest in your Gallery.

Mark Eliot
Catherine Jarvis and Mark Eliot
Romany Mill Studios

Hydrogenic Power Engineering

Hydrogenics Secures $3 Million from Japan Auto Makers

Letter from RemyC remyc@prodigy.net
Contacts: investors@ hydrogenics.com

Hydrogenics Secures $3 Million in Test Equipment Orders Japanese Auto Makers Select Greenlight Power for Fuel Cell Testing Equipment

TORONTO, Sept. 30 – Hydrogenics Corporation (Nasdaq: HYGS; TSX: HYG), a developer and manufacturer of fuel cell products, announced today that its wholly owned subsidiary, Greenlight Power Technologies (Greenlight) has received orders totaling US $3 million for fuel cell test stations from two leading Asian automotive customers. Both are repeat orders for multiple machines.

Cumulatively, Greenlight is providing seven Fuel Cell Advanced Test Stations (FCATS) to two major customers. The first order is for three 3 kW PEM FCATS L-Series test systems. The second is for three 12 kW PEM FCATS H-Series test systems and one 60 kW PEM FCATS HX-Series testing station. The stations are expected to be delivered over the next two quarters.

"Our decision to open an office in Tokyo, Japan over three years ago continues to pay dividends," said Pierre Rivard, President and CEO of Hydrogenics. "The Japanese market is advancing quite aggressively towards the commercialization of fuel cell technology, as evidenced by our growth in test equipment sales to the region over the past three years. We are delighted to secure these major contracts with customers who are clearly setting the pace in fuel cell development."

Greenlight fuel cell testing systems provide high precision instrumentation, combined with full-featured software. They deliver fuel cell testing results that prove the reliability, repeatability and versatility required for world-leading fuel cell development programs.

About Hydrogenics Corporation

Hydrogenics Corporation (Nasdaq: HYGS; TSX: HYG) is a leading clean power generation company, engaged in the commercialization of fuel cell technology and test stations for fuel cells. The company is building a sustainable business, in a potentially "game changing technology" for transportation, stationary and portable power. Hydrogenics, based in Mississauga, Ontario, Canada, has operations in British Columbia, Canada, Japan, the United States and Germany. For more information, please visit www.hydrogenics.com .

About Greenlight Power Technologies, Inc.

Greenlight Power Technologies Inc. (http://www.greenlightpower.com), a wholly-owned subsidiary of Hydrogenics Corporation, is a leading global supplier of testing and diagnostic equipment to the fuel cell industry. It has supplied fuel cell test equipment to the world's premier fuel cell stack manufacturers, component manufacturers, system integrators and research organizations. Greenlight provides a full suite of test equipment for fuel cell stack, stack component, reformer and system testing for companies focused on portable, stationary, and transportation fuel cell applications.

Hydrogen Fuel Cells

Anuvu Incorporated

3980 Research Drive
Sacramento, CA 95838
USA
Tel: (916) 921-7040
Fax: (916) 921-7044
Email: anuvu@anuvu.com

www.anuvu.com/home.html

Now our readers can get familiar with some hydrogen fuel cell development photos. It is worth of a note that the prices are high for the majority of people. For example, a hydrogen fuel cell car costs 100 000 dollars FOB Sacramento, CA. The next picture features a fuel cell van, which costs 150 000 dollars.

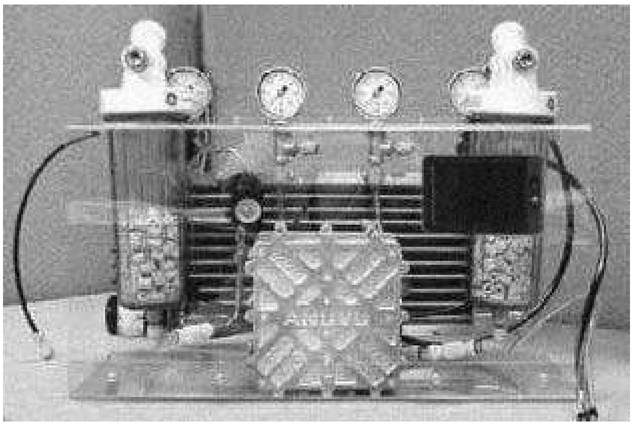

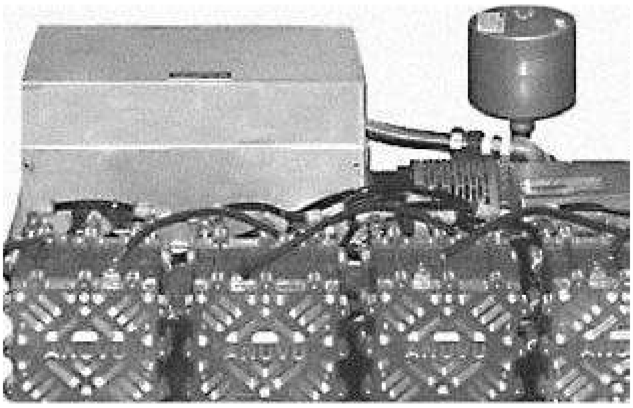

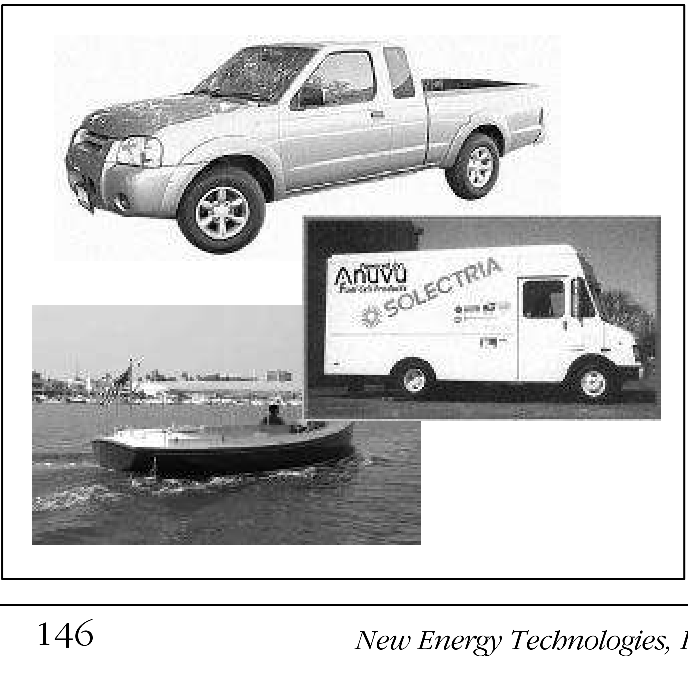

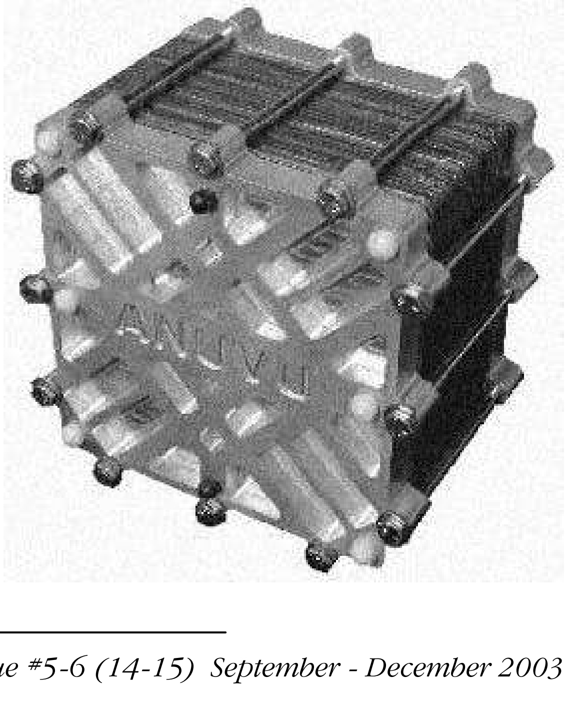

EXCESSIVE OUTPUT BY MEANS OF AIR IONIZATION

We have received an interesting article from California. Here is the short version of the article.

The Mechanism of The Electric Spark

Review of the research

by Leonard B. Loeb
Professor of Physics, University of California at Berkeley
and
John M. Meek, Research Engineer
Stanford University Press, Stanford University, California

Dedicated to Professor J. S. Townsend whose pioneer research and theory laid the whole foundation for the study of the mechanism of the electrical spark discharge.

Although the electric spark has been known to mankind in its various manifestations from time immemorial, its mechanism has to date been little understood. The initial clarification of the mechanisms involved is due to J.S. Townsend as a result of his brilliant researches in the early nineteen hundreds. On the basis of his theory of ionization by collision by electrons and positive ions, the fundamental mechanisms active and especially the coefficients required in their application were made available.

In 1936 the present senior author was forced to describe the mechanism of spark discharge in terms of a modified but distinctly unsatisfactory Townsend theory. In 1935 the discovery of photo-ionization in air by corona discharge indicated a solution was not far off. The turning point of a more successful theory came in the discovery of streamers in positive point to plane corona in 1936. The quantitative analysis of the self-propagating positive streamer in all breakdown phenomena became clearly evident as a result of the data concerning electron avalanches. As a result a qualitative mechanism of sparking by streamer propagation from anode to cathode functioning by means of photo-ionization in the gas was established.

The Townsend Sparking Criteria

It will not be necessary here to derive the famous equation of Townsend for the current [i] in a gap between electrodes as a function of the photoelectric current [io] from the cathode, the gap length [x] and the coefficients [a] and [B]. For this the reader can go to any standard text. *(Editor: the equation is omitted)*

In this equation the first Townsend coefficient [a] represents the number of new electrons created in the gas by an initial electron in its advance of 1 cm along the field axis from the cathode.

The second Townsend coefficient [B] in Townsend's original theory was the number of new electrons created by a single positive ion in its advance of 1 cm along the field from the anode.

The quantity [a] has been extensively studied in various gases. It varies with the ratio of field strength to pressure, X/p, where [X] is in volts per centimeter and [p] is in millimeters of Hg.

Note: the reason we are going through this is to determine the actual increase in current provided by the spark gap, and thus be able to design the circuit to avoid blowing out semiconductor components. It also provides a sound and already proven scientific theory to work from giving us a good foundation and the confidence to proceed with technical design work.

The quantity [B] has been evaluated, albeit rather inaccurately, from the variations of [i] with [x] at various higher values of X/p, by many observers in different gases. Inasmuch as it has now been shown that there are numerous other mechanisms other than impact with positive ions, which can liberate the secondary electron, needed in discharge.

There has been an inclination to give up the mechanism of impact ionizations by positive ions in gas. The discovery of measurable photoelectric ionization in gas has now made it possible to explain such cases. The exact way in which photo-ionization in the gas could operate to cause a spark, was not clear until the development of the present **streamer theory**.

The Streamer Theory of Spark Discharge Anode Space-Charge Field Due to an Avalanche

Assume a spark gap of 1 cm in length. Assume that in air at atmospheric pressure the potential across the plates is 31,600 volts, which is the conventionally observed sparking potential [Vs].

Let us then calculate what happens in the field to one of those electrons. It starts across the gap, quickly acquiring an average random energy of some $E=1/2mC^2= 3.6$ electron volts and a drift velocity [v] in the field direction of about 1.5 to 2 times 10^7 centimeters per second. As it moves it creates new electrons at a rate of [a] per centimeters in the field direction so that in a distance [x] it and its progeny amount to e(ax) electrons, forming what is called an **electron avalanche**.

Therefore, e(ax) positive ions have been left behind by the electron group, virtually where they were formed in the 10^{-7} second of advance for the electrons in the distance x=q across the plates. As the electron avalanche advances, its tip is spreading laterally by the random diffusive movement of the electrons.From these data it is possible to compute the density of positive-ion space charge left behind at any point [x]. The value of [a] under these conditions is about 17, making e(aq)=e(17). The first ion pair is created at 0.0407 cm from the cathode. At 0.5 cm from the cathode there are 4914 ions, at 0.75 cm there are 3.66 times 10^5 ions, and within 0.0407 cm from the anode there are 1.2 times 10^7 ions. **Most electrons will be drawn to the anode except for some few that are bound by the positive ions, making a sort of a conducting discharge plasma in the avalanche.**

Such a distribution of ions does not make a conducting filament of charges across the gap, and hence in itself an avalanche that has crossed does not constitute a breakdown of the gap. Thus one must look further for the mechanism of the spark.

If Loeb and Meek are correct then if we assume a spark gap of 3 mm and a voltage of 5,000 voltsthere are roughly 2,000 electrons created by avalanche for every one electron leaving the cathode. They state that most of these 'free electrons' are absorbed by the anode. [This would certainly explain why the semiconductor components cannot handle the current gain.]

NOTE: Loeb and Meek make little reference to initial amperage. There are only two values they refer to 10^{-5} ampere and 10^{-12} ampere.

In conclusion: Sparks and Arcs are two different beasts. My initial research into the amperage necessary to form an arc does not apply to spark and the process of avalanche where this huge gain mechanism is possible.

Photoelectric Ionization in Gas as a Secondary Mechanism

Accompanying the cumulative ionization there is produced by electrons from four to ten times as many excited atoms and molecules. Some are excited to an energy exceeding the ionizing potential of some of the atoms and molecules present, either by excitation of an inner shell, by ionization and excitation, or in a **mixed gas like air by the excitation of molecules of higher ionizating potential, e.g., N2**. These excited atoms or molecules emit radiations of very short wave length in some 10^{-8} second. This short ultraviolet radiation is **highly absorbed** in the gas and leads to ionization of the gas. In fact, the whole gas and the cathode as well are subjected to a shower of photons of all energies traveling from the region of dense ionization with the velocity of light. Thus nearly instantaneously in the whole gap and from the cathode new photoelectrons are liberated which almost at once begin to ionize cumulatively.

The Mechanism of Positive Streamer Formation

The photoelectrons created at points in the gas and at the cathode at any great radial distance from the avalanche axis will merely create other avalanches. Those in the gas will be short and those coming from the cathode region will be long and like that of the initial avalanche. Being smaller and, in any case, later in creation than the parent avalanche, such avalanches will be of no interest in breakdown. However, those photoelectrons created near the space-charge channel of positive ions, and especially near the anode, will be in an enhanced field, which exerts a directive action drawing them into itself. If the space-charge field [X1] is in the order of magnitude of the imposed field [X], this action will be very effective. In addition the values of [a] will be much enhanced.

The electrons from the intense cumulative ionization of such photoelectron avalanches in the combined fields [X] and [X1] which are drawn into the positive space charge feed into it, making it a conducting PLASMA which starts at the anode. The added fields will be most effective along [X] and so will the ionization. The positive ions they leave behind will therefore extend the space charge towards the cathode. These electrons also create photons, which produce electrons to continue this process. **In this fashion the positive space charge develops toward the cathode from the anode as a self-propagating positive space-charge streamer.**

As the streamer advances towards the cathode it produces a filamentary region of intense space-charge distortion along a line parallel to the field. The conducting streamer of a plasma consisting of electrons and ions extending to the anode thus makes a very steep gradient at the cathode end of the streamer tip. As this advances toward the cathode the photoelectron avalanches produced by radiation at the cathode, especially at the intercept of the extended streamer axis at the cathode, it begins to produce an intense ionization near the cathode. Hence the positive ions created there may increase the secondary emission. Thus, as the space-charge streamer approaches the cathode a cathode spot is forming which may become a source of visible light.

When the streamer reaches the cathode there is a conducting filament bridging the gap. As the streamer tip reaches the cathode the high field produces a rush of electrons towards the end of the streamer. This if followed by a current of electrons, gives a high-potential wave, which passes up the preionized conducting channel to the anode, **multiplying the electrons present by a large factor**. The channel is thus rendered highly conducting. If the metal can emit a copious supply of electrons because of the formation of an efficient cathode spot, the current of electrons continues the channel maintaining its high conductivity and **ever increasing in it**. This current, unless limited by external resistance, will then develop into **an arc**. It is, however, the intense increase in ionization by the potential wave, which gives the highly conducting channel characterizing the spark.

Conclusion: According to Loeb and Meek there are three means by which a spark in open air will provide a very large current gain. If this is true, it should be fairly easy to prove with inexpensive and unsophisticated equipment. Once the actual amount of current gain has been determined for the design parameters of the spark gap, then the rest of the circuit can be designed for the increased current value.

Editor: It is worth of a note that the current gain by means of ionization was patented by Pavel N. Yablotchkov (the patent of France # 120684, October 11th 1887). Some two years ago one of our issues featured an article about him. We think it is worth being published again.

PATENT of 1877 by Pavel N. Yablotchkov

Alexander V. Frolov
Faraday Lab Ltd,
Lev Tolstoy St., 7 St. Petersburg, 197376 Russia
Tel: 7-812-380-3844

Pavel N. Yablotchkov was born in 1847 near Saratov, Russia. He graduated as a Military Engineer in 1866 and spent several years in the Russian Army.

In 1872 he came to Moscow and started his activities in electrotechnical field. Since 1875 he had been working in Paris with the famous Louis Breget and his first French patent # 110479 of November 29th, 1875 was dedicated to an electromagnetic transformer. Then he developed and patented a lighting system (the well-known Yablotchkov electrical candle). In 1876 he patented a new electromagnetic transformer for industrial purposes, France # 115793 of November 30th, 1876.

The most interesting patent claim on over-unity devices by Pavel N. Yablotchkov is known as France patent # 120684, October 11th, 1877, "The system of distribution and amplification of electrical currents by means of atmosphere electricity..." The patent describes special capacitors connected in series with the load to increase the output current by means of ionization. Experiments were conducted together with the well-known physicists such as Dr. Maskar, Dr. Varren-Delaru and others and **they confirmed the 200 % efficiency of the circuit**. Now we will try to explain the method.

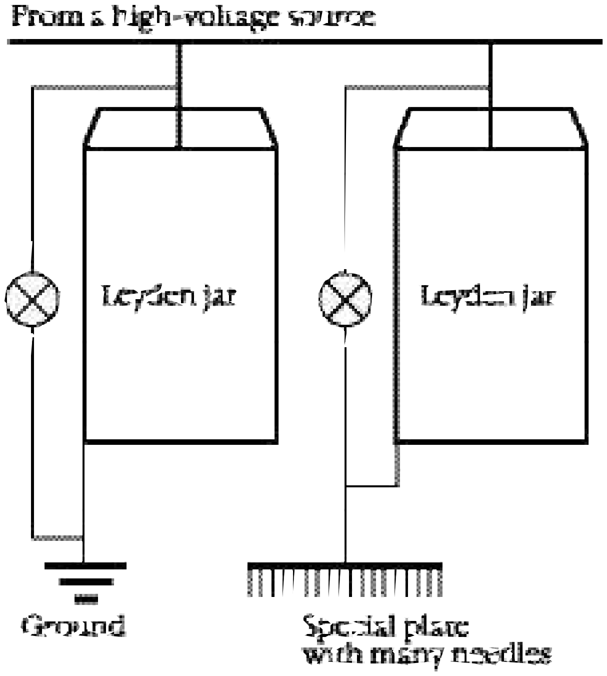

Fig.1 features a schematic drawing taken from Yablotchkov's patent. The Leyden jar is an asymmetrical capacitor, i.e. it is different in principle from a two-plate flat capacitor. The inner electrode of the jar should be connected to a high-voltage source and in this case the changes of the potential have an effect on the potential changes on the external electrode. It does not work in the opposite case and if you connect a high-voltage source to the external electrode no potential changes will be detected on the inner electrode. Connection to the ground or to a special plate (which is covered with many needles to increase air ionization) is necessary to attract the maximum number of electrons to the plate surface or to return the maximum electrons from the plate surface when potential changes on the external electrode are produced by means of electrical induction in the Leyden jar.

In conclusion I should mention one more supposition of the secrets of the well-known Swiss M-L converter (Methernitha). The main elements of the design are Leyden jar capacitors, which have the external surface made of perforated metal.

The other known fact is that great ionization of air is observed when the converter is in operation. So, the electrostatic machine can produce pulses of a very high voltage (potential difference) but it cannot be used as a source of a powerful current. In order to increase the current in the circuit we should apply a certain method and Yablotchkov's technology seems to be appropriate. A large surface of the external electrode of the Leyden jar can be a good solution to the problem. Maximum strong ionization allows us to obtain the output current several times stronger than the weak current generated by the electrostatic machine.

SUBSCRIPTION 2004!

To subscribe for
New Energy Technologies
please visit our web site:

http://www.faraday.ru/net.htm

One-year subscription
price for 2004 is $46.00

World-wide airmail delivery
included.

In 2004 we will publish
5,000 copies, 4 times per year.
Each printed issue will include
100 pages of printed information.
Alongside the English version
there is the Russian version
available.

CD "New Energy Technologies"

Back issues of 2001 - 2003 are
available on a CD with PDF files.
The price is $26.00.
You can order a CD from the same
web page:
http://www.faraday.ru/net.htm

Contact information for additional requests:

Faraday Lab Ltd,
New Energy Technologies magazine,
Lev Tolstoy Str. 7 -601,
St.Petersburg,
197376, Russia.
Tel/fax 7-812-3803844

TeslaTech, Inc.

296 E Donna Dr, Queen Valley, AZ 85218

Visit our site!.... **http://www.teslatech.info**

Benefits

- 1 yr *ExtraOrdinary Technology* (qtly)
- Mail Order Discounts
- Conference Fee Discounts

ExtraOrdinary Technology... our magazine containing news about *the leading edge of our Scientific Revolution*.

...an invaluable resource to track leading edge research!!!

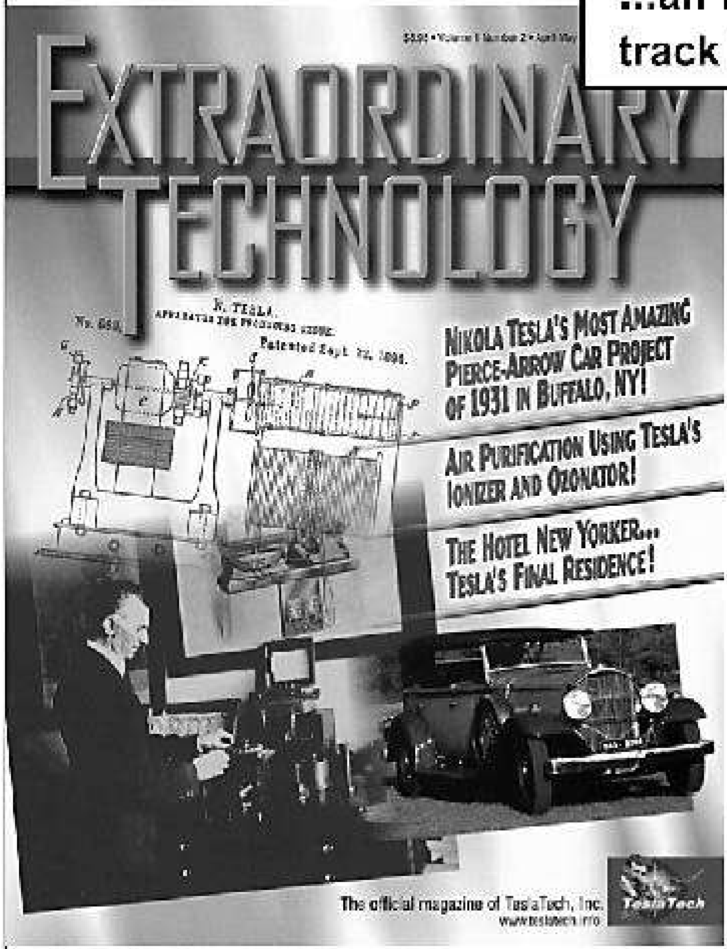
EXTRAORDINARY TECHNOLOGY

NIKOLA TESLA'S MOST AMAZING PIERCE-ARROW CAR PROJECT OF 1931 IN BUFFALO, NY!

AIR PURIFICATION USING TESLA'S IONIZER AND OZONATOR!

THE HOTEL NEW YORKER... TESLA'S FINAL RESIDENCE!

The official magazine of TeslaTech, Inc. www.teslatech.info

ExtraOrdinary Technology focuses on:

Tesla Technology
Low temperature plasma
Electrogravitation
Electrified water
Magnetic healing
Electromedicine
New Health Supplements
...and much, much more!

JOIN TODAY!
(520) 463-1994

Annual Membership Rates: $30 - United States • $35-Can/Mex • $55-Else

Advanced Energy • Alternative Medicine • Future Technology

Konstantin Meyl

Scalar waves

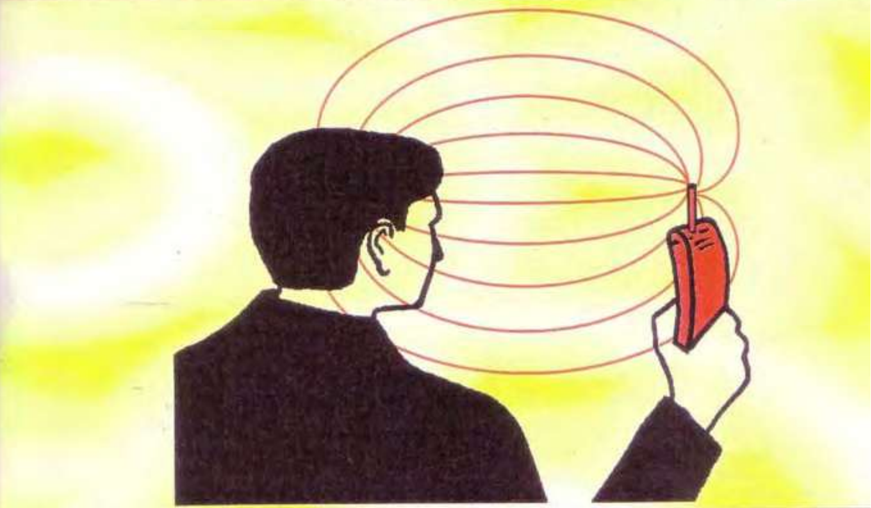

From an extended vortex and field theory
to a technical, biological and historical use
of longitudinal waves.

Edition belonging to the lecture and seminar
„Electromagnetic Environmental Compatibility"

Preface to the lecture, 1st Edition 1996

The theme encloses the electromagnetic compatibility of both technical and biological systems. Only part of the electromagnetic wave can be considered for function troubles, namely the part that was absorbed and has rolled up to a vortex. The activity depends on the number of created vortices and of their lifetime, their decay.
The eddy current only manifests in conducting materials. In the air and in dielectric materials on the other hand the vortex of the electric field will form, also called the potential vortex. To calculate and to measure this vortex is our goal.

First we'll carry out a survey of the problems and the usual methods. From the analysis of unsolved problems the need for the introduction of the new vortex phenomena is deducted and an adequate field-theoretical approach will be chosen. Afterwards the potential vortices are calculated and their properties are discussed and interpreted. For the purpose of proving their existence, on the one hand the Schrodinger equation will be derived and on the other hand the quantum properties of the most important elementary particles will be calculated and compared with the well-known measured values. Measurement and calculation are in excellent agreement for weight, charge, magnetic moment and spin. So the theory not only proofs it's correctness, in addition it demonstrates it can achieve much more. The theory takes us to the unification of the well-known interactions and physical phenomena and shows itself as an unified theory. In the practical conversion and usage of the theory there will not only be informed but by all means also be provoked as an entrance in a fruitfully discussion. Fundamental questions will be taken up like: What is information, energy, temperature or smell? The connection to the theme of the electromagnetic environmental compatibility is formed by the technical and the biological usage of the potential vortices, the energy transmission of Nikola Tesla exactly like the in a similar way functioning nerve conduction. Here we already can expect biological reactions.

This lecture, held for the first time in the winter semester of 1995/96, is available in book form, as an edition belonging to the lecture. This lecture will not deliver ready recipes or instructions. The goal is reached when the critical sense of the listeners and readers has been inspired and discussions have been set going. Everybody has to draw the consequences out of such a theory by him- or herself.
In addition to this lecture a seminar is offered, wherein several themes are supplemented or deepened, different theories are compared and possible consequences are discussed. The appearance of an edition belonging to the seminar has started in 1998<i>. Regarding the conversion of consequences both politicians and scientists are equally addressed, because the electromagnetic environmental compatibility has developed to one of the most urgent problems of today's world. But in last consequence all of us bury the worldwide responsibility for our environment.

<i>: K. Meyl: Electromagnetic environmental compatibility, Part 2 and 3 of this book, Edition belonging to the seminar.

Table of contents
Page

Preface to the Lecture. 1st Edition 1996 III

Table of contents IV

Part 1 edition belonging to lecture "Electromagnetic environmental compatibility"

1. Introduction 1
1.1 Discussion about limits 3
1.2 Wireless telephones 5
1.3 Absorption of waves 7
1.4 Overlap effect 9

2. Tasks 11
2.1 Tasks of the electromagnetic environmental compatibility 11
2.2 Tasks of the electromagnetic compatibility (EMC) 13
2.3 Declaration of conformity 15
2.4 EMC-techniques to measure the emission of interference 17
2.5 Electro smog 19
2.6 EMC-techniques to measure the stability against interference 21
2.7 Tasks of the biological compatibility 23
2.8 Natural fields 25
2.9 Electricity of the air 27
2.10 Biological effects 29
2.11 Artificial fields 33
2.12 Protection against artificial fields 35
2.13 Unsolved tasks 37

3. Approach 39
3.1 Principle of causality 41
3.2 Field-theoretical approach 43
3.3 Duality 45
3.4 Flow vortices 47
3.5 Rankine vortex 49
3.6 Vortices and anti-vortices 51

4. Properties 53
4.1 Concentration effect 53
4.2 Duality of the vortex properties 55
4.3 Derivation of the electron as an elementary vortex 57
4.4 Quanta as field vortices 59
4.5 The photon 61
4.6 Pair creation 63
4.7 Noise 65
4.8 Condenser losses 67
4.9 Vortex lines and vortex streets 69

Page

4.10 Water colloids as vortex balls ... 71
4.11 Transport phenomenon ... 73

5. Derivation and interpretation ... 75
5.1 Fundamental field equation ... 77
5.2 Mathematical interpretation ... 79
5.3 Physical interpretation ... 81
5.4 Phenomenological interpretation ... 83
5.5 Atomistic interpretation ... 85
5.6 Derivation of the Klein-Gordon equation ... 87
5.7 Derivation of the time dependent Schrodinger equation ... 89
5.8 Derivation of the time independent Schrodinger equation ... 91
5.9 Interpretation of the Schrodinger equation ... 93

6. Theory of objectivity ... 95
6.1 Proof ... 95
6.2 Law of conservation of energy ... 97
6.3 Radius of the electron ... 99
6.4 Maxwell's field equations ... 101
6.5 Transformation equations ... 103
6.6 Field overlap ... 105
6.7 Field dependent curvature of space ... 107
6.8 Electromagnetic interaction ... 109
6.9 Gravitation ... 111
6.10 Field dependent speed of light ... 113
6.11 Universality ... 115
6.12 Aether ... 117
6.13 Spin and tunnel effect ... 119
6.14 Interpretation of the measured faster than light speed ... 121
6.15 Definition of the speed of light ... 123
6.16 Relativity and objectivity ... 125
6.17 Transformation ... 127
6.18 Transformation table ... 129
6.19 Interpretation of the transformation table ... 131
6.20 Particle decay ... 133

7. Proof ... 135
7.1 Elementary vortices ... 137
7.2 Matter and anti-matter ... 139
7.3 Positronium ... 141
7.4 Dipole moment ... 143
7.5 Myon ... 145
7.6 Calculation of the vortex fields ... 147
7.7 Calculation of the proton ... 149
7.8 Strong interaction ... 151
7.9 Magnetic moment of the proton ... 151

Page
7.10 Structure of the neutron 153
7.11 Calculation of the neutron 155
7.12 beta-decay 157
7.13 "Weak interaction" 159
7.14 Tau particle 161
7.15 Pions 163
7.16 Table of vortices of the Mesons 165
7.17 Table of vortices of the Baryons 167

8. Unified theory 169
8.1 Structure ofthe field theory 171
8.2 Unification of the interactions 173
8.3 Temperature 175
8.4 Heat 177
8.5 Sound 179
8.6 Basic principle of cybernetics 181
8.7 Structure of adaptive regulatory circuits 183
8.8 Information 185
8.9 Philosophy of nature 187

9. Usage 189
9.1 Longitudinal electromagnetic waves 191
9.2 Medical technical usage 193
9.3 Flying objects 195
9.4 Electro gravitation? 197
9.5 Free energy? 199
9.6 Nerve conduction and action potential 201
9.7 Wireless energy transmission technology 203
9.8 Measuring and switching technique of Nikola Tesla 205
9.9 Energy technical usage 207
9.10 Environmental compatibility 209
9.11 Ecologically fairusage 211
9.12 Claims concerning the environmental compatibility 213
9.13 Epilogue belonging to the lecture (part 1) 215
Abstract belonging to scalar waves, part 1 218

Page

Part 2, edition belonging to the energy technical seminar 219

10.0 Preface to the seminar, part 2 220
10. Oscillating interaction 221
10.1 Kepler's laws 221
10.2 Unknown interaction 223
10.3 Harmony of the alternating current engineers 225
10.4 Four fundamental interactions 227
10.5 Resonating interaction 229

11. The growing globe 231
11.1 Long-distance effect of the neutrinos 231
11.2 Model of expansion 233
11.3 Ancient Egyptian legend of Genesis 235
11.4 Inner structure of the earth 237
11.5 Earth's core as a converter of neutrinos 239
11.6 Speed of growth 241
11.7 Conservation of angular momentum 243
11.8 Set of difficulties concerning the change of magnetization 245
11.9 The weakly damped moon 247
11.10 Calculation of celestial mechanics 249
11.11 The biblical age 251

12. New cosmology 253
12.1 Concerning the formation of our solar system 253
12.2 The birth of the planets 255
12.3 The derivation of the Titius-Bode law 257
12.4 The hollow planet 259
12.5 Concerning the formation of the universe 261
12.6 Counter examples concerning the 2nd law of thermodynamics 263
12.7 Entropy destroying potential vortices 265

13. Recording of space and time 267
13.1 The measuring technical debacle 269
13.2 The clock paradox 271
13.3 The Tamarack mines experiment 273
13.4 Field dependent linear measure 275
13.5 Experiences from space travel 277
13.6 Spherical aberration 279
13.7 Step by step understanding of the measuring technique 281
13.8 Discussion of the cosmological insights 283

14. The way towards free energy 285
14.1 The liberalization of the energy markets 285

Page

14.2 The consequences of the liberalization 287
14.3 The chances of the wind energy 289
14.4 The chances of being self-sufficient concerning energy 289
14.5 The space energy technology of nature 291
14.6 The gap in the energy balance of man..293
14.7 Carl Freiherr von Reichenbach 293
14.8 Cold fusion and Genesis 295
14.9 Photosynthesis 297
14.10 How nature materializes 299
14.11 Lightning 301
14.12 Ball-lightning..303
14.13 Discussion concerning the neutrino conversion..305

15. Principle of funtioning of space energy 307
15.1 The course of the field lines 307
15.2 Gravitation..309
15.3 Systematizing the interactions 311
15.4 Magnetic force converter..313
15.5 The railgun 315
15.6 Unipolar induction..317
15.7 Tendency to instability 319
15.8 Unipolar field configuration 321
15.9 The Testatika 323
15.10 The secret of the Testatika..325
15.11 The key to free energy..327

16. Space energy technology (SET).. 329
16.1 The unipolar generator 329
16.2 Derivation of Maxwell's field equations.. 331
16.3 SET-devices with rotating magnetic fields.. 333
16.4 Commentary concerning magnetic SET-devices.. 335
16.5 RQM and the space quanta manipulator.. 337
16.6 SET-devices with pulsed magnetic fields.. 339
16.7 SET-devices with Mobius winding 341
16.8 Mobius converter of Seike and Coler 343
16.9 Tesla's flat coil 345
16.10 The secret of the flat coil 347
16.11 Discussion concerning the technology of the neutrino collectors 349

17. Technical use of the weak interaction __ 351
17.1 Radioactivity caused by neutrinos.. 351
17.2 Nikola Tesla, the discoverer of the neutrino radiation 353
17.3 Transmutation and reduction of radioactivity 355
17.4 The Patterson Power Cell 357
17.5 Report concerning the cold fusion.. 359
17.6 Water-Fuel-Cell technology 361
17.7 Unconventional electrolysis 363

Page

17.8 Materialization of neutrinos 365
17.9 Oscillation of size and luminescence 367
17.10 Phosphorescence and sonoluminescence 369
17.11 Discussion concerning the use of the weak interaction 371

18. Physical phenomena 373
18.1 Water as a catalytic converter for neutrinos 373
18.2 Background and zero point radiation 375
18.3 The Casimir effect 377
18.4 Effects under high tension 379
18.5 Flying discs 381
18.6 Propulsion with neutrinos 383
18.7 Antigravitation or levitation 385
18.8 Discussion concerning the effect of closed field lines 387
18.9 Concerning the determination of the stand of today's energy technique 389

19. Own experiments 391
19.1 Concerning the rating of experimental physics 391
19.2 The charge amplifier from the workshop of handicrafts 393
19.3 Interpretation of the observations with the charge amplifier 395
19.4 High-tension experiments in the laboratory for performance electronics 397
19.5 Measurement of dielectric capacitor losses in the HF laboratory 399
19.6 Analysis of the measured capacitor losses 401
19.7 Microwave oven for testing in the laboratory of electrical engineering 401
19.8 Microwave heating by decay of vortices 403
19.9 The Tesla coil from the workshop of handicrafts 405
19.10 Biological effectiveness of the Tesla coil 407
19.11 Wrestling for measuring technical insight in the Transfer Centre 409
19.12 Neutrinolysis, the alternative splitting of water 411

20. Cosmic experiments 413
20.1 Millions of voluntary test subjects at 11 August 1999 413
20.2 Model of calculation for the cosmic experiment 415
20.3 Physical technical consequences 417
20.4 Biological consequences 419
20.5 Epilogue to 11 August 1999 421
20.6 Dropping of the neutrino radiation 423
20.7 Analysis and comparison of the occurrences 425
20.8 Radioactive decontamination with the help of a supernova 427
20.9 Free energy from the Orion constellation 429
20.10 Interaction of the neutrinos with the core of the earth 431
20.11 Changing of polarity and apocalypse 433
20.12 Scalar wave gauges for the prediction of earthquakes 435
20.13 Calculation of future occurrences 437
20.14 Epilogue belonging to the energy technical seminar (part 2) 439
Abstract belonging to scalar waves, part 2 442

Page

Part 3, edition belonging to the information technical seminar 443

21.0 Prefaces to the seminar, part 3 444
21. Derivation of the scalar wave 447
21.1 Lord Kelvin in a delicate mission 449
21.2 Helmholtzian ring-like vortices in the aether 451
21.3 Taking apart of the wave equation 453
21.4 Derivation of the scalar wave equation 455
21.5 Derivation of a plasma wave 457
21.6 Mistakes of the HF technology 459
21.7 Coupling of the wave parts 461
21.8 Set of difficulties of the near-field 463
21.9 Transition to the far-field 465
21.10 Scalar wave model 467
21.11 Double-frequent oscillation of size 469
21.12 Electric and magnetic scalar wave 471

22. Properties of scalar waves 473
22.1 Wave-particle compromise 473
22.2 Concerning the measurement of light 475
22.3 Comparison of the parts of Tesla and Hertz 477
22.4 Noise, a scalar wave phenomenon 479
22.5 Neutrino radiation 481
22.6 Parallel instead of serial image transmission.......................................483
22.7 Comparison of the properties 485

23. Research of scalar waves 487
23.1 Frequency diagram 487
23.2 The reverse of the HF medal 489
23.3 Radiesthesia 491
23.4 Prediction of earthquakes with the help of scalar waves.......................................493
23.5 Searching water with the help of scalar waves 495
23.6 Receiving scalar waves with the Lecher antenna.......................................497
23.7 Assignment according to wavelength 499

24. Medicine of oscillation 501
24.1 Mediation of information with the help of scalar waves 501
24.2 Systems with positive feedback 503
24.2.1 The Synchrometer.......................................503
24.2.2 The sound therapy 503
24.2.3 The colour therapy 505
24.2.4 The aromatherapy 505
24.2.5 The gem therapy 505
24.2.6 Meditation and self-therapy.......................................505

Page

24.3 Systems with negative feedback........ 507
24.3.1 The homeopathy 507
24.3.2 The bioresonance 507
24.3.3 The frequency therapy 509
24.4 The Zapper 511
24.5 Discussion concerning the medicine of oscillation 513

25. Contributions to the discussion concerning the information technical use 515
25.1 Window of the sensuous perception........ 515
25.2 The sense organs (nose, eye, tongue) 517
25.3 The nerve conduction 519
25.4 The brain, a scalar wave computer 521
25.5 Concerning signal engineering 523
25.6 Stability against interference and repair mechanisms 525
25.7 The microwave window........ 527
25.8 Discussion concerning scalar wave medicine 529

26. Recapitulation from the viewpoint of textbook physics 531
26.1 Common misinterpretation of the antenna losses........ 531
26.2 Wave damping by field vortices 533
26.3 Laplace versus Maxwell 535
26.4 Error term in the wave equation 537
26.5 Interim result........539
26.6 Failure of the Maxwell theory 541
26.7 Concerning the evidence situation 543

27. Faraday versus Maxwell 545
27.1 Energy out of the field........545
27.2 Vortex and anti-vortex........547
27.3 Vortices in the microcosm and macrocosm 549
27.4 Faraday's law and Maxwell's formulation........551
27.5 The discovery of Faraday 553
27.6 Different formulation of the law of induction 555
27.7 Contradictory opinions in textbooks 557
27.8 The field-theoretical approach........559
27.9 Derivation of Maxwell's field equations 559
27.10 Derivation of the potential vortices 563
27.11 Derivation of the fundamental field equation" 565
27.12 The Maxwell field as a derived special case 567
27.13 Derivation of the wave equation........569
27.14 The new field approach in synopsis 571

28. Objectivity versus relativity 573
28.1 The question concerning the aether........ 573
28.2 Vortices, an overlap of the overlap........ 575
28.3 Field overlap........ 577

Page

28.4 The derivation of the length contraction 579
28.5 The dependence of the Lorentz contraction on the field 581
28.6 Boscovich and the respiration of the Earth 583
28.7 Derivation of the gravitation 585
28.8 From subjectivity to objectivity 587
28.9 The objective standpoint 589

29 Mathematical gleanings 591
29.1 General and special theory of relativity 591
29.2 Transformation table 593
29.3 Discussion concerning the root of Lorentz 595
29.4 Mathematical derivation of the neutrino 597
29.5 Universal constants and constants of nature 599
29.6 The fundamental number e 601
29.7 The pentacle 603
29.8 The vortex, which is rolling up 605
29.9 The Golden Proportion 607

30. Scalar wave technology in antiquity 609
30.1 Thesis 609
30.2 Prerequisites 611
30.3 Approach 613
30.4 Circumstantial evidence 615
30.5 Radio technical concepts 617
30.6 Wireless telegraphy 619
30.7 AM temple technology 621
30.8 Modulation of acoustic signals 623
30.9 Broadband FM broadcasting technology 625

30.10 Epilogue 627

Contacting address and information 634

Table of formula symbols 635

References 636
A: Publications concerning the theory of objectivity (primary literature) 637
B: Publications about the theory of objectivity (secondary literature) 641
C: Literature concerning physics in general 643
D: Literature concerning geophysics, geology, geography and astronomy 644
E: Literature concerning the EMC and biological compatibility 644
F: Literature concerning Tesla and criticism of textbook physics 644
G: Literature concerning free energy 645
H: Literature concerning medicine, biology and chemistry 646
J: Literature concerning history and more peripheral subjects 646

Appendix 647

1. Introduction

Here the extremely controversially discussed question of the environmental compatibility of electromagnetic fields will be persuaded. Limits should inform what is incompatible and what is compatible. But there are as many limits as there are commissions and specialists. And besides that differ the results from each other for several powers of ten. In course of time the legitimate doubts become unmistakable and the representatives of science slowly get to feel the burden of proof.

For the sake of efficiency, the actual discussion concerning the theme of electro-smog is analysed and the necessity to involve an until now unnoticed field phenomenon in the discussion about limits is derived: It concerns vortices of the electric field. These potential vortices, as they are called, have the corresponding properties to show biological effects even at the lowest field strengths. In any case it is not possible to exclude that at present the wrong physical phenomena are measured and made responsible.

A parable should bring clarity.

Lets imagine that the to us well-known and over our sense of touch understandable physical phenomenon of the temperature is unknown to us, neither measurable nor perceptible. Our weather station only exists of a barometer that could show us the air pressure and deliver us indications if good or bad weather is to be feared. We ready realize that there exists a connection between the air pressure and our health and make the to us well-known phenomenon responsible. When the pointer points to good weather we can go out lightly dressed. With bad weather we should take a coat, so we know from experience.

"Now we imagine the realistic situation that in winter we have a weather situation of high pressure but it's stone-cold outside. The weather station will display high temperatures with the result that some people will walk around with short-sleeved and open shirt, only to lie in bed with a cold in the evening. Of course the air pressure was to blame! Logically the "pressure sensitive", as they are called mocking, demand the limits for the allowed pressure to be reduced so far that no consequences for health are to be feared. Concerning the theme of allowed limits, science is asked and science proceeds in a systematic way: the pressure is investigated in the laboratory, isolated from all other parameters and so it is discovered that man catches no cold even at a substantially higher air pressure, so there is no reason to alter the limits.

Actually we would expect these at any time reproducible results to have a calming effect on the minds of the participants of the discussion and on the population. Instead the pressure sensitives time and again cite new knowledge that won't fit in the scheme. So is for instance stated that draught causes the same health problems although this pseudo effect has nothing at all to do with the air pressure. So owing to incomprehensibility and emotions the discussion about limits becomes a farce.

The fact that sensitive people react to effects of air electricity and possibly get ill without proof that some today measurable physical quantities are responsible should make us think. It is little calming watching our scientists poking at the dense fog whereas at the same time among the runners of the new telecommunication networks there spreads something like a gold-digger mood.

To introduce a new technology is not difficult, but to abolish it for reasons of the electromagnetic environmental compatibility is almost impossible!

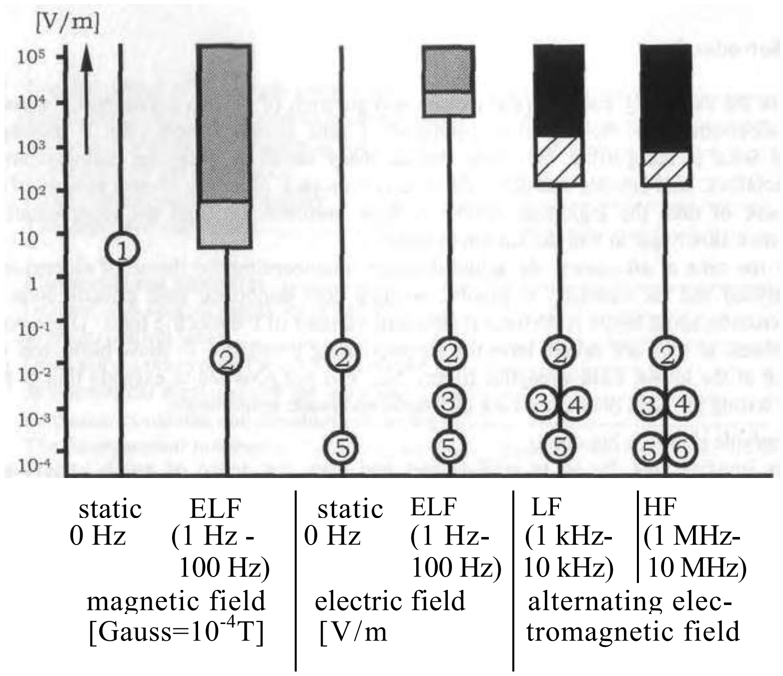

Fig. 1.1: Discussion about limits

■ Limits for professional exposition (electronics engineers) and
▨ Limits for non-professional exposition (population in general)
according to the Recommendation of IRPA/INIRC<ii>.
□ Limits according to VDE 0848 / 1989
▣ Limits according to VDE 0848 / 1992
⊛ experimentally determined threshold values of reactions of
biological systems <i>:

1 Increase of the activity of movement of birds
2 deflection of divining-rods
3 influence on the time of reaction of men
4 conditional reflexes of fish without electrical organs
5 conditional reflexes of fish with electrical organs
6 conditional muscular reflexes of men

taken from:
<i>: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligk.), 5. Aufl., Bild 111, S. 123
Verlag Moos & Partner Munchen, ISBN 3-89164-O58-7
<ii>: Habiger u.a., EMV, Verlag Technik, Berlin 1992, S. 152

1.1 Discussion about limits

Whoever follows the public discussions concerning electro smog always sees two arguing parties, whose standpoints lie so far apart that they inevitably talk at cross purposes and there can be found no consensus.

On one side the "affected" find together who maintain to have found the electromagnetic radiation as the damaging cause for their problems. They are to be taken serious, even when only their personal sensitivity serves as means of measurement and proof and a more or less distinct sensitivity against electromagnetic phenomena. This group occasionally finds support of homeopaths who can base on reproducible laboratory results that fit as few into the view of life of science as the empirical statements of affected and possibly hurt people.

On the other side stand the representatives of the energy-supply companies and the runners of radio networks who argue with the needs of our modern industrial society and give "limits" prescribed to them by scientists. These, for their part, proceed according to strictly scientific methods. Their presented results are reproducible and there's no doubt about them.

The limits after all are fixed far below those that are recommended from a scientific viewpoint. Nevertheless both groups are separated from consensus by powers of ten. When we want to know how deep the ditch is we want to bridge, we should take a look at the determined limits (Fig. 1.1).

The limits stem from the 1RPA (International Radiation Protection Association) an organ of the World Health Organization that in turn has appointed the INIRC (International Non Ionizing Radiation Committee). These now state to have used all available scientific research results as basis for the given guidelines.

Moreover a safety range was worked into them. So the limits were fixed at substantially lower levels to guarantee that no health damage arises. In this way first the limits were determined for the people who for reasons of profession are exposed to electromagnetic fields.

For the population in general the limits for the so called non-professional exposition were reduced further to one half till one fifth for reasons of caution and care. In Fig. 1.1 these limits are registered. Thereby is distinguished between magnetic fields and electric fields that appear stationary or at extremely low frequencies (ELF describes frequencies between 1 Hz and 100 Hz). Moreover limits for low-frequency (1-10 kHz) and high-frequency (1-10 MHz) alternating electromagnetic fields are given.

The graph should serve as a rough orientation and show us the proportion of scale. As further information some thresholds of measured reactions of biological systems are registered (after Konig*[1]*). Because a logarithmic scale was chosen to fit all the values on one graph it becomes clear that between the first reactions and the recommended limits there lie up to five powers often. The ditch seems to be insurmountable.

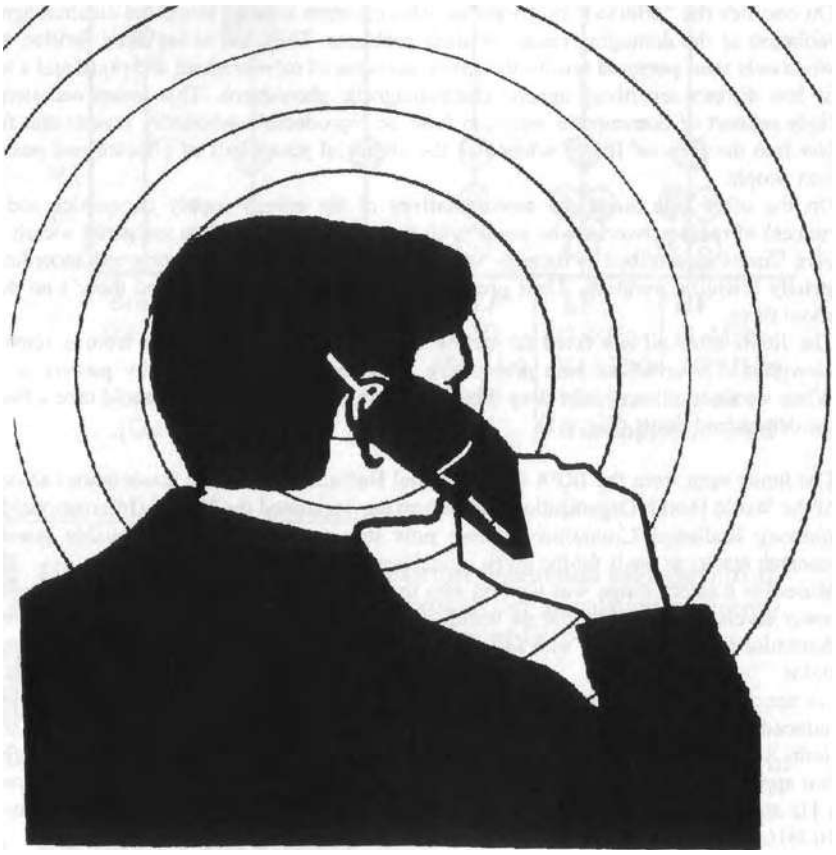

Fig. 1.2: Set of problems of environmental compatibility by means of the example of the handheld wireless telephones (handy).

<i>: L.v.Klitzing: Neurophysiologische Einflusse durch elektromagnetische Felder wahrend und nach der Exposition, Med. Universitat zu Lubeck

1.2 Wireless telephones

Measuring technical surveys with regard to the influence of brain currents by digital radio signals by the university hospital in Lubeck have startled handy manufacturers and users equally <i>. Although in this case measurement errors could be detected, the „bugaboo on the wall" remains that we are sitting unsuspecting in a restaurant and a neighbour draws his handy out of his pocket to make a digital telephone call. Thereby synchronizing the brain currents within a radius of 100 meters on the broadcasting signal and occupying our brain useless with technical signals. The derivation will show that from the start this can't happen to all visitors, because as a prerequisite conditions of resonance must be fulfilled. But would there be an affected, he or she for sure would have considerable problems, because informations that are not picked up over the sense organs can neither be classified timely nor as regards content.

An affected whose brain has picked up technical signals not even is able to register by itself that it was fed with incorrect informations. It would be reasonable when the visitors of the restaurant would defend themselves and put the radio operator on the doorstep. The number of restaurants where apart from cats and dogs also handy's have to stay outside is increasing. How should we keep out of the way of electromagnetic fields? Should we walk around permanently with a steel helmet or even better in a knight's armour and even go to bed with them? It would be worse than in the dark middle ages.

Summarizing: it should be guaranteed that the operation of electro technical apparatus causes neither health damage nor unintentional influence or irritation. A systematic and scientific procedure should investigate in the laboratory all relevant physical phenomena individually for their interaction. Electro physics bases on two phenomena in connexion with electro-smog: on the one hand the radiation and on the other hand the thermal effect, but at a close look both factors prove to be of only little importance!

In radiation measurements the intensity of the electromagnetic wave at a certain place is determined. In laboratory experiments the field strength is increased so long till biological reactions are observed. Thermal limits are determined in a similar way. As said, the values lie about powers of ten above those that possibly bother you when you hold a handy to your ear. It is true that the microwave radiation penetrates into your head but we also know that it marches out again on the other side and this visit in your head happens with the speed of light.

Exactly like this are guest in your body constantly your local radio station, your local television station the satellites with hundreds of programs and anyway the whole radio technical world even when you did not invite them.

For an electromagnetic wave to become receivable, the field strength must lie clearly above the common noise signal and this can only be achieved by a permanent overlap, by standing waves, like in a cavity tuned to a specific frequency or an antenna. As long as people don't let themselves grow antennas on their heads they hardly have to fear direct biological effects of electromagnetic waves.

That leaves as the second phenomenon the thermal effect. With a handy held to your cheek there comes into being a local fever in your head. But that is not at all unusual or unnatural for the human body. Something like that happens to a far greater degree when you take a hot foot bath or let yourself be irradiated at one side from the sun at a tourist grill.

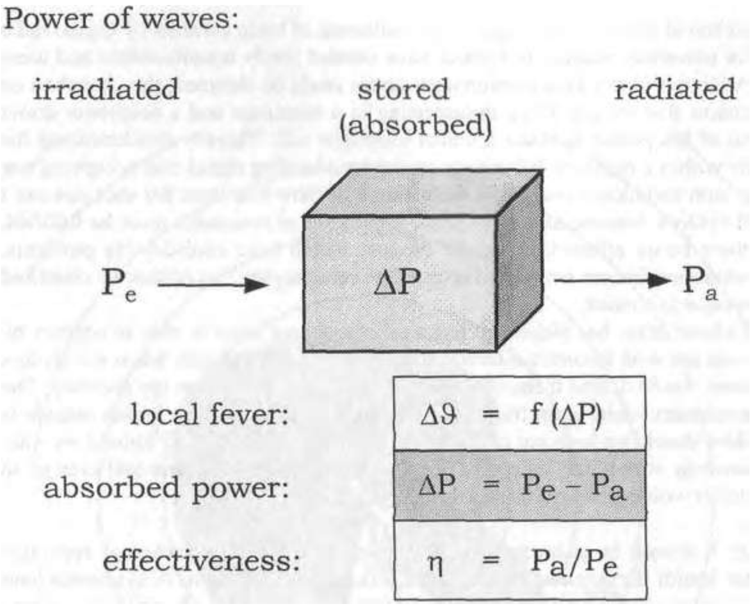
Power of waves:
irradiated
stored
(absorbed)
radiated
P_e
ΔP
P_a
local fever:
$\Delta\vartheta = f(\Delta P)$
absorbed power:
$\Delta P = P_e - P_a$
effectiveness:
$\eta = P_a/P_e$

Fig. 1.3: Damping of waves and ability to absorb of a body (our head) if we are making a phone call with a handy.

a contribution to the theme dielectric losses

- capacitor
- high-frequency welding
- microwave oven

<i>: K. Meyl: Potentialwirbel Band 1 [Al] (only in German),
INDEL Verlag, Villingen-Schwenningen 1990, ISBN 3-9802 542-1-6
<ii>: K. Meyl: Potentialwirbel Band 2 [A2] (out of print),
INDEL Verlag, Villingen-Schwenningen 1992, ISBN 3-9802 542-2-4

1.3 Absorption of waves

The with the theme dealing physicians logically have to put up with criticism that they work only with two phenomena that not at all can be involved authoritative in the causes for biological effects. A third factor can be considered, a field phenomenon until now stayed unnoticed by science: the vortex of the electric field, the so called potential vortex. A vortex is to be considered as an oscillation around a fixed point. Through that a permanent overlap is caused, like what happens at an antenna only that the vortex is not bound to the dimension of an antenna. The potential vortex is contracting and in this way reaches extremely high energy densities at very little spatial measurement, densities that lie far above those that field strength measurements are pretending to us [Al]<i:>.

When again you take the handy at hand with which you ,,blow" the pulsed microwaves into your head. Don't worry, because with the speed of light and without provable damage almost everything comes out again on the other side, but only almost everything. A little damping of the wave has taken place and your head has absorbed this part of the irradiated wave (Fig. 1.3). Who claims this is already the thermal factor actually should realize that there exists no corresponding term in the wave equation. Here there are found merely two dual vortex phenomena as a possible damping term: the eddy current and the potential vortex. An eddy current damping is ruled out because of the bad conductivity of the head. But this favours his dual anti-vortex, the potential vortex [A1]<i>.

Seen physically the following is taking place in your head: the absorbed waves roll themselves up to vortices and through that become localized and overlap themselves permanently (Fig. 1.4b). In the course of time the vortices decay and produce the well-known eddy losses that lead to the measurable increase in temperature. When reactions or biological effects arise, simply and solely the vortex can be considered as the possible cause. Thereby play two points an important role: the number of the generated vortices and their lifetime that is determined by the time of decay.

In anticipation of the mathematical calculation of the potential vortices it is pointed out here that these are favoured not only by a low conductivity, but also by a high dielectricity. Because water has an unusual high dielectricity (ε_r = 80) and our head consists predominantly of water doubts in dealing with handy's are reasonable.

Also the relaxation time constant representative for the lifetime can be calculated [A2]<ii>. We must proceed from the assumption that both the number of the vortices and their lifetime, that is all the at a fixed point in time in our head existing and effective vortices, can be a cause and therefore have to be considered and investigated scientifically.

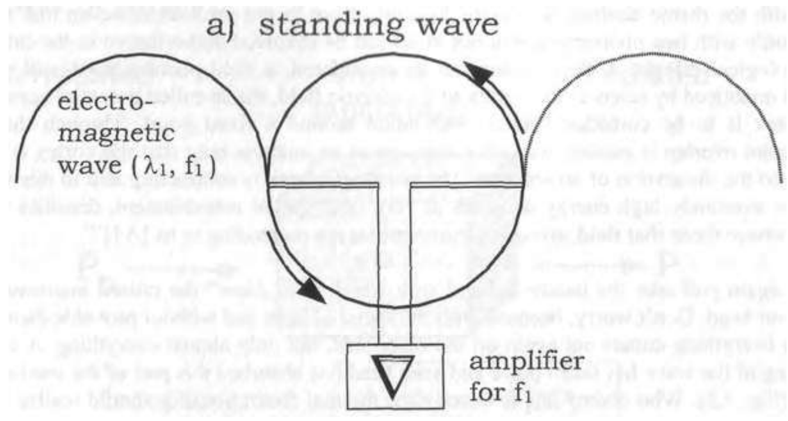
a) standing wave
electro-
magnetic
wave (λ_1, f_1)
amplifier
for f_1

b) localized vortex = noise

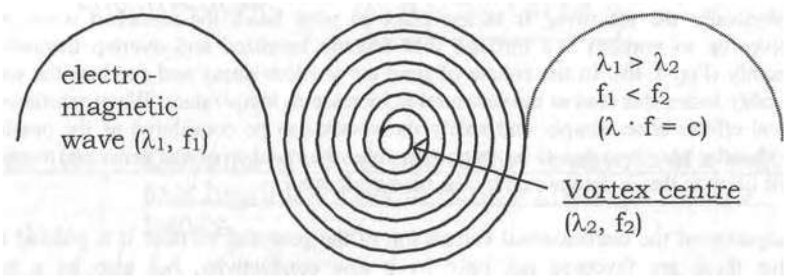
electro-
magnetic
wave (λ_1, f_1)
$\lambda_1 > \lambda_2$
$f_1 < f_2$
($\lambda \cdot f = c$)
Vortex centre
(λ_2, f_2)

c) broadband antenna for EMC-measurements

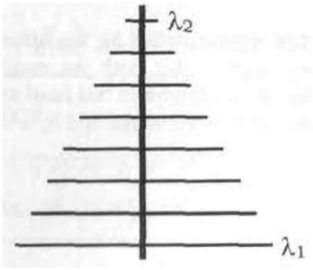
λ_2
λ_1

Fig. 1.4: Measurement of localized waves and vortices

1.4 Overlap effect

The graph at the left (Fig. 1.4) should clarify once more that only the in space localized and permanently overlapping field appearances can be considered as a cause for biological effects. This can concern an over an antenna standing wave (a) or a vortex (b) which is to be taken as localized by nature.

It would be allowed to in general speak only of a vortex because the standing wave can by all means be understood as a special case of the vortex. The essential difference is that the vortex is not bound to the size of a fixed antenna and can contract itself at any time to achieve in this way a substantial higher energy density. As a result this than will lead to an increased biological effectiveness.

It should be pointed at a further omission. In the discussion about limits, without exception, the absolute field strength of the broadcasting signal is valued and not the type of modulation. The last mentioned should actually not at all play a role according to the prevailing scientific opinion.

Totally different is the case with a vortex that acts damping. Such a vortex shows near it's centre a considerable smaller wavelength than more to the outside and through that it has a big frequency bandwidth [A5]<i>. It is to be expected that in the extremely broadband pulsed signals of the digital networks the creation of vortices (or eddies) will be favoured considerably stronger than in amplitude- or frequency-modulated signals (AM/FM/C-network). In connexion with analog modulated radio- or handy-signals until now there never has been reported of any synchronization of the brain currents with handy-signals from a comparison of the EEG with the broadcasting signal.

Interestingly the for EMC-measurements usual stepped broadband antennas have exactly the construction that certainly would be favourable to the measuring technical registration of vortex phenomena (Fig. 1.4c).

With the dipole antennas of different lengths for different wavelengths there still are measured waves and not vortices but these measuring techniques is certainly accommodating to the until now unnoticed and stayed undiscovered vortex phenomenon. So there are some good reasons that the vortex is a dominating influential factor for EMC-problems.

By means of the example of the handheld wireless telephones can be studied and discussed with which set of problems the very young discipline of science of the environmental compatibility has to fight in the future. And in which ways there can be found approaches towards a solution of the problem. When the comfortable and trodden out ways of textbook physics do not lead to the goal than we will have to force our own way through the jungle of science.

At first we'll have to obtain a short overview of the actual level of research and knowledge. From the criticism to this we than can derive the tasks of the electromagnetic environmental compatibility and in particular the unsolved tasks.

<i>: K. Meyl: Wirbel des elektrischen Feldes, eine neue Storquelle? EMC Journal 1/95, 6. J, S. 56-59.

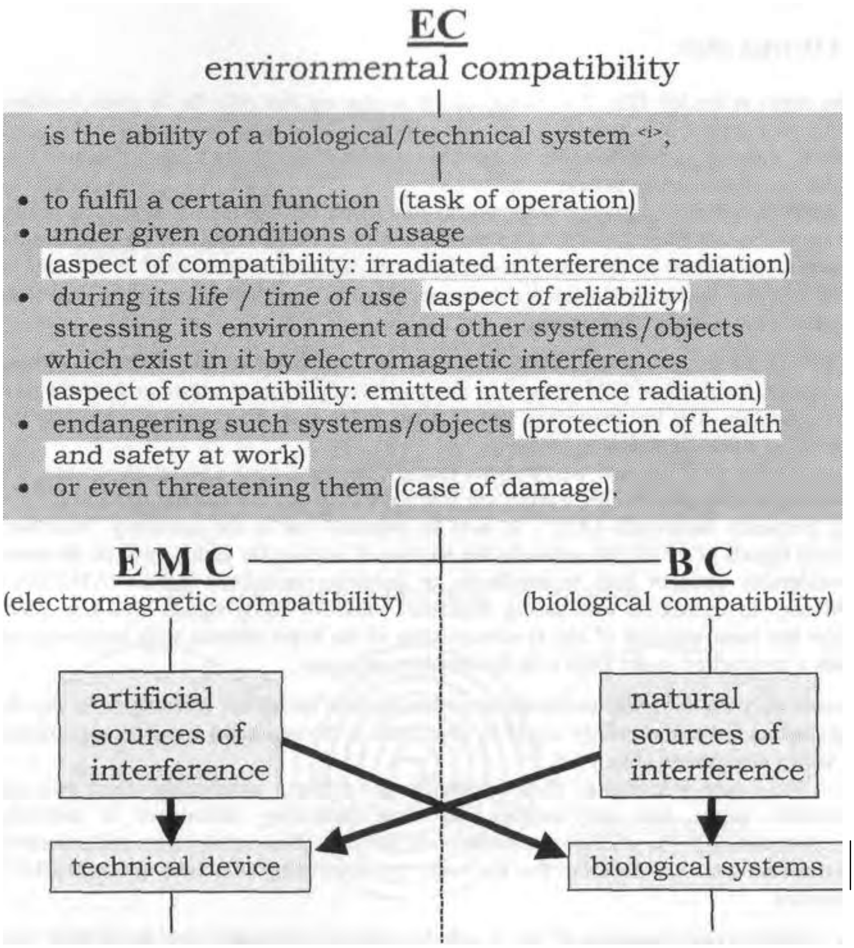

EC

environmental compatibility

is the ability of a biological/technical system <i>,

- to fulfil a certain function (task of operation)
- under given conditions of usage (aspect of compatibility: irradiated interference radiation)
- *during its life / time of use (aspect of reliability)* stressing its environment and other systems/objects which exist in it by electromagnetic interferences (aspect of compatibility: emitted interference radiation)
- endangering such systems/objects (protection of health and safety at work)
- or even threatening them (case of damage).

E M C
(electromagnetic compatibility)
B C
(biological compatibility)
artificial sources of interference
natural sources of interference
technical device
biological systems

Regulations of	called EMEC	know how of
the German Post	thirties: radio	druids
Office since 1992:	operator disease	magic
EMC-law<ii>	today, E-Smog,	esoteric
from 1996:	HF-, SW-therapy	medicine
EU -regulation	diathermy, etc.	biology

Fig. 2.1: Overview concerning environmental compatibility

<i>: E. Habiger: EMV-ubergreifende Aspekte zu benachbarten Problemfeldern, Elektrie 48 (1994), Heft 5/6, Seite 163-161

<ii>: EMVG: Gesetz uber die elektromagnetische Vertraglichkeit von Geraten, Bundesgesetzblatt Teil I vom 9.11.1992, S. 1864

2. Tasks

2.1 Tasks of the electromagnetic environmental compatibility

The environmental compatibility (EC) forms the generalization that includes both the electromagnetic compatibility (EMC) and the biological compatibility (BC). Besides the technical and functional goals of an undisturbed functional course it also pursues ethical and moral goals.
Technology should benefit to humanity and at the same time be in accord with nature. This goal will not be reached when this technology directly or indirectly is endangering humanity.
A direct attack on the health of people poses for instance the military usage of technical apparatus or the negligent usage, by pretended ignorance and unsuspicion.
Is a technology posing a danger to the environment so humanity endangers itself indirectly with this technology. After all are human beings a product of their environment. We always should reckon on the environmental sins taking revenge on us sooner or later.

In fig. 2.1 a formal definition is given that in particular concerns the claims for an undisturbed functional course: it concerns the compatibility aspects of unallowed emitted and irradiated interference radiations, the reliability and quality safety with which a function and task is fulfilled and finally the questions of the protection of health and the safety at work.

Moreover fig. 2.1 provides an overview and the structure of the 2nd chapter. First we'll treat the electromagnetic compatibility (EMC) that first of all deals with the influence of artificial but also natural interference sources on technical apparatus.
After that we'll throw a glance at the appearing fields in nature. The biological compatibility (BC) deals with the influence on biological systems.
An especially sensitive area of the environmental compatibility (EC) than describes the with a cross-link hinted influence of artificial interference sources on biological systems that is popularly described as „electro smog".

The numerous aspects of the environmental compatibility for instance in the areas of chemistry and biology that certainly are important, but do not fall in the area of electromagnetism, can't be treated in the here marked framework.

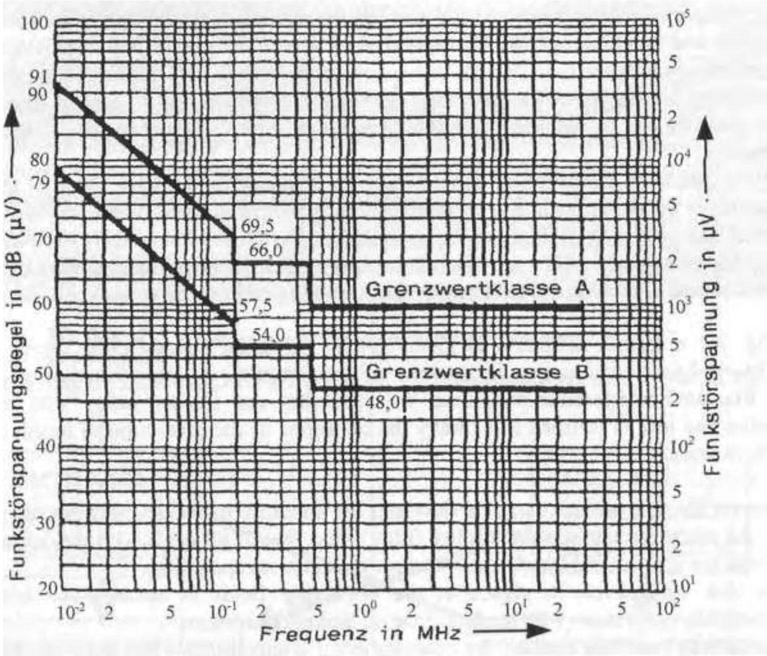

Fig. 2.2: Classes of limits according to VDE 0871 (since 1-1-96: VDE 0875)<i>

<i>: Anton Kohling: Grundlagen der Funkentstorung in der Bundesrepublik Deutschland, etz Bd. 108 (1987), Heft 10.

2.2 Tasks of the electromagnetic compatibility (EMC)

First of all the EMC (electromagnetic compatibility) is concerned with the function of technical apparatus. Correspondingly rational and dry sounds the official definition: "electromagnetic compatibility (EMC) describes the ability of an electrical setting-up (i.e. of a construction element, of a construction group, of an apparatus or of an installation) to function in an electromagnetic environment without stressing this environment by electromagnetic effects in an unallowed fashion ".

Actually it concerns an old need for protection that should be as old as the usage of electro technical apparatus. But in the beginning no one cared about it. The spark gaps with which Heinrich Hertz 1888 in Karlsruhe has carried out the first radio technical experiments were genuine „polluters", that would have been detectible at several hundreds of kilometres distance with modern receivers. For these installations that he had assembled in the lecture room with his students, today he would hardly get a permission of operation and the since 1996 required declaration of conformity he would get not at all.

1925, as in Germany the number of radio listeners had exceeded the limit of one million, for the first time a need for protection appears in the certificate of approval for radio receivers: "The public telegraphs and telephone installations must not be disturbed by the radio receiver".
Later on every backside of the good old steam radios there was found the following hint (translated): "This apparatus meets the interference radiation regulations of the German Post Office". So the manufacturers were urged to measure the emission of their apparatus and in particular to screen the HF-oscillators in the superhet-receivers.

Since the fifties, in the VDE-institute EMC-examinations in the present day sense are taken. The main point of the measurements and the by the VDE recommended limits, however is about interferences bound to a circuit. On the supply lines of the network the prevailing conditions are reproducible so that standards can be put through (Fig. 2.2).

For measurements of interference radiation maybe the time was not ripe enough or the necessity was not big enough. The usual argumentation was: what we can't measure reproducibly, can not be forbidden and certainly not be put under punishment. Therefore merely recommendations were issued or impositions weak as wax were made like: "the interference field strength ... must be so small that an undisturbed reception is guaranteed as soon as the minimum field strength for utilization exists at the place where the antenna is mounted"<i>.

In common parlance that means something like: "as long as no one bleats, everything is allowed". Within a connected industrial area there even existed an officially legitimized fools freedom. Merely at the fence of the industrial area limits had to be fulfilled.

Specially for the line-frequency of the screen one has decided to build a loophole in the law so that one didn't have to throw the TV sets, that so successfully had conquered the living rooms, out of the window. Of course the flickering screens did interfere exactly as before but this EMC-interference now was officially approved.

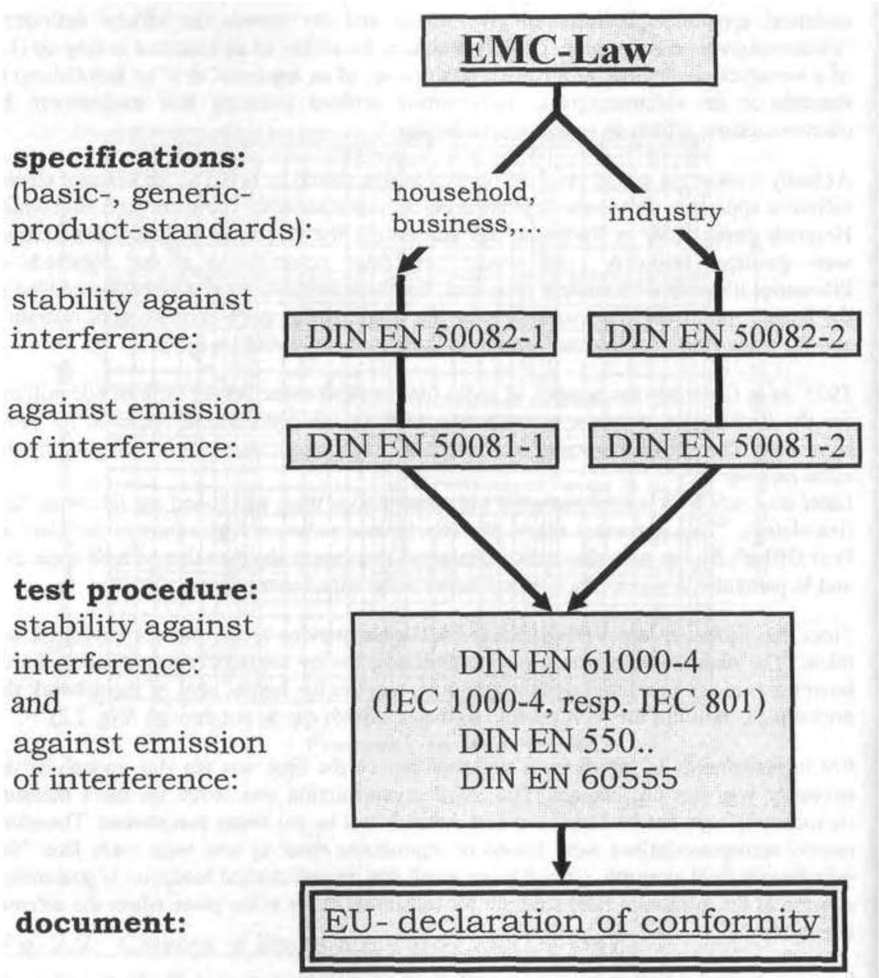
EMC-Law
specifications:
(basic-, generic-
product-standards):
household,
business,...
industry
stability against
interference:
DIN EN 50082-1
DIN EN 50082-2
against emission
of interference:
DIN EN 50081-1
DIN EN 50081-2
test procedure:
stability against
interference:
and
against emission
of interference:
DIN EN 61000-4
(IEC 1000-4, resp. IEC 801)
DIN EN 550..
DIN EN 60555
document:
EU- declaration of conformity

Fig. 2.3: The way to EU conformity

2.3 Declaration of conformity

In the EMC times seem to have gone as the standardizers had to fit in with the insufficiencies of technology. Meanwhile the conditions have turned up. We owe this circumstance first of all the EMC-law of 1992 that doesn't name any limits but it states the political intention to demand from technical apparatus and installations an appropriate stability against interference and at the same time limit the sent out interference. As a consequence of this law the measurement facilities and measurement processes had to be standardized to get reproducible measurement results that are not influenced by the electromagnetic environment. That goes so far that even the floor covering of a measurement hall is dictated because the conductivity of the floor influences the degree of reflexion. Normally the examinee is situated on a revolving plate that is turned around for 360° during the measurement of the radio interference field strength. Is it however not possible to turn the examinee than the antenna has to be led around it, thereby again increasing the dimensions of the measurement hall. The distance to the antenna should be up to 10 meters. Moreover it must be possible to move the antenna, up till a height of 4 meters to register the influence of the reflexions on the floor.

Moreover there is to plan a reflexion free zone around the measurement track (in elliptical form) that depends on the reachable damping of reflexions of the used absorber. Used are pyramids of foam material soaked with carbon and increasingly tiles of ferrite or shieldings of wallpaper.

Taken all together for a measurement hall doing justice to standards there result considerable measurements of for instance 18 m length x 10 m width x 7 m height. Let's again come to talk about the EMC-law with which only the intention but not the way is fixed. To form the claims catalogue in a way that is fulfillable in general, some concrete prescriptions, the so called standards, have to be worked out. This task was transferred to the European committee for electro technical standardization CENELEC, which has established the workgroup TC 110 to at first work out some standards: The basic standards deal independent of product with general questions of the EMC, of the testing process and of the measurement environment.

The generic standards likewise deal independent of product with the so called fundamental technical standards for apparatus in their dependence of the respective electromagnetic environment (protected computer room or medical room, environment of the house, office or industry).

The product standards concern the EMC-standards referring to products (7 product families / approx. 50 products).

In Fig. 2.3 the arduous way through the jungle of paragraphs for a technical apparatus is outlined. Corresponding to the requirements of use, first the relating ES-standards for the apparatus have to be determined and than have to be measured according to own test standards based on the fundamental technical standards. When the allowed limits for stability against interference and for sending out interferences are not exceeded, the EC-declaration of conformity is handed out. Since 1.1.96 that declaration is needed when apparatus are commercialized or - stated more exactly - "put in circulation" and operated. When still further EC-guidelines are met in the end the CE-hallmark is awarded. Since 1.1.96 only with this hallmark the access to the common market of the EC is possible. Violations can be punished with fines and if need be with imprisonment. But there are great national differences in the EC. The Federal Republic of Germany with fines of up to 50.000 Euro counts as expensive for criminals.

Simulation of network for the measurement of the interference voltages U_{st1} and U_{st2}.

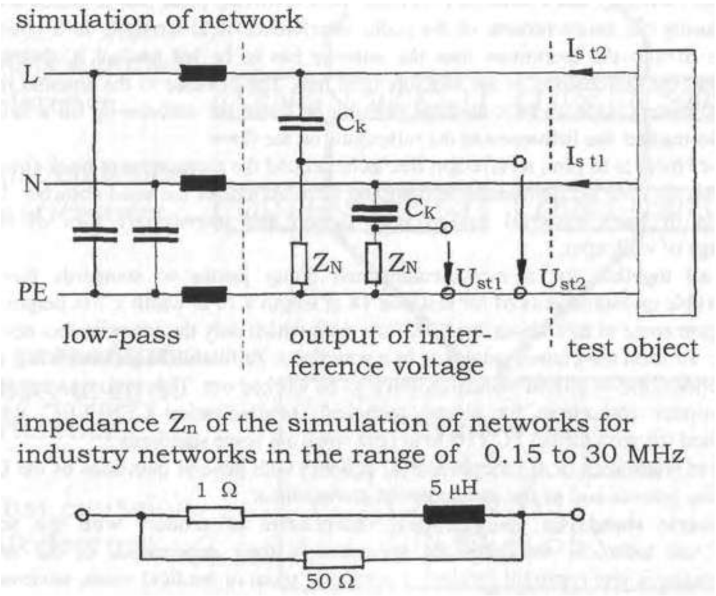
simulation of network
L
N
PE
C_k
C_K
Z_N
Z_N
$I_{s\,t2}$
$I_{s\,t1}$
U_{st1}
U_{st2}
low-pass
output of interference voltage
test object
impedance Z_n of the simulation of networks for industry networks in the range of 0.15 to 30 MHz
1 Ω
5 μH
50 Ω

Fig. 2.4: Simulation of network for the measurement of interference voltages.<i>

<i>: acc. to Ernst Habiger: EMV, Huthig Buch Verlag Heidelberg (1992), ISBN 3-7785-2092-X

2.4 EMC-techniques to measure the emission of interference

Actually we already can be glad that it came to an europe-wide agreement for the regulation of the EMC-set of problems. But the question if we can be satisfied with what we have reached is still outstanding. All too often the lowest common denominator of the measurable and checkable was sought and not so much the technical possible was taken into consideration.

The main emphasis is put on the measurement of the emission of interferences. Traditionally the interferences bound to a circuit are registered in a frequency range up to 30 MHz. The corresponding wavelengths thereby can correspond with the length of the supply lines and form standing interference waves. Primarily the spectrum of the interference currents is measured e.g. over a HF-current converter. These currents produce a voltage drop over the internal resistance of the feeding network. Because the properties of the networks can vary very strong, a standardized end-resistor is required for the measurement of the interference voltage. For this purpose an imitation of the network is switched between the network and the examinee. This imitation in addition has the task to keep away the interference signals that come from the supplying network with the help of filter-elements (Fig. 2.4).

The measurement of the interference radiation, the field-bound interference emission, lakes place between 30 MHz and 1 GHz. For that a free field or an absorber-hall with little or no reflexions is required. The standardized distances of measurement are 3, 10 and 30 meters. The electric field strength is determined with dipole broadband antennas, the magnetic field strength with frame antennas. It must be possible to both vary the receiving antenna between horizontal and vertical polarization and to adjust the receiving antenna in the height and the position to the test object.

Typical measurement set up to measure the emission of interferences bound to a conductor<i>

A: shielded link conductor
B: bundle of conductors folded like a meander
C: connection to the reference mass
ME: receiver of interference signal
NNB: Simulation of network
PO: test object

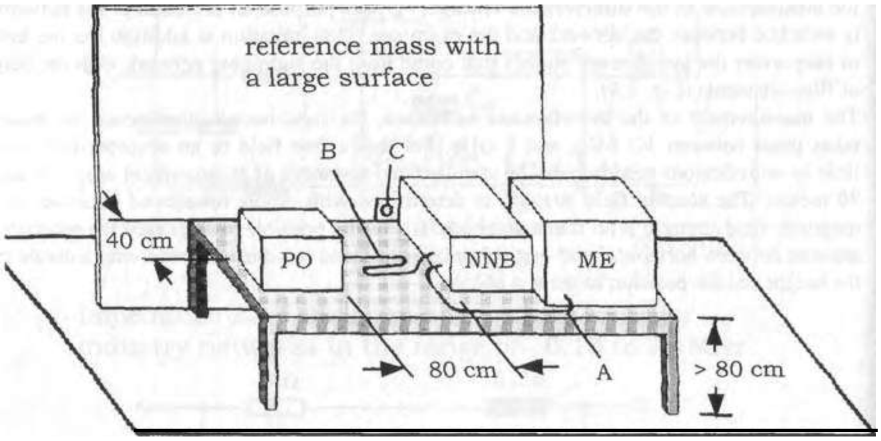
reference mass with
a large surface
B
C
40 cm
PO
NNB
ME
80 cm
A
> 80 cm

Fig. 2.5: Measurement set up to measure the emission of interferences bound to a conductor,<i>

<i>: Ernst Habiger: EMV, Huthig Buch Verlag Heidelberg (1992), ISBN 3-7785-2092-X

2.5 Electro-Smog

There is almost no end to the possibilities of variation and one needs already a lot of overview and experience to determine the field strength maximum. Nevertheless we have to ask ourselves if in this way really all emissions of interference are understandable, that popularly are described as „electro smog".

Smog is the combination of the terms Smoke and Fog. It therefore describes a pressure on the environment with fog like smoke. When for instance in the case of smog alarm all interference sources are switched off, which means all kilns are brought to a stop and all automobiles are stopped, than the fog like smoke therefore still is not vanished from the air. It just distributes itself and dissolves only very slowly.

The transfer of the smog idea on the electromagnetic interference radiation is bound to fail because, when the test object is switched off no emission of interference at all is detectable with the usual measurement means. Nevertheless the rainbow-press is trying to enumerate almost all electromagnetic field phenomena under the term „electro smog" without consideration of the fact that this term is not at all a collection term.
From the sight of an expert one can only speak of smog when something like smog remains and stays active further after the switching off of an electro technical apparatus. It should be a phenomenon that is not understandable by the standardized measurements of interference radiation. Such a phenomenon would be e.g. the vortex of the electric field. However vortices are virtually not measurable in a direct way because they have the unpleasant property to whirl about around the measurement probe. But they will be detectable by their eddy losses and in the case of the electric field vortex appear as noise. Until now the standardizer however haven't planned to investigate the influence of an apparatus on the noise in the environment. Here we still grope in the dark.

At least the vortex shows a storing property that would justify the use of the idea "smog". We'll have to investigate the phenomenon.

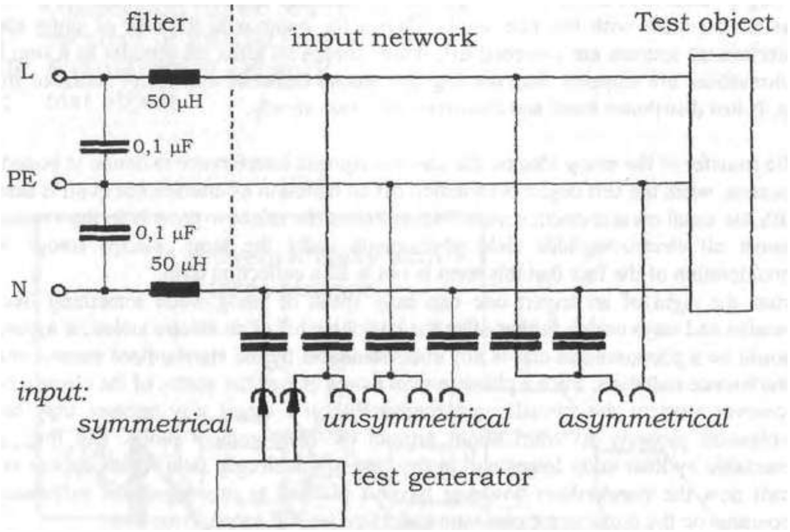
filter
input network
Test object
L
50 µH
0,1 µF
PE
0,1 µF
50 µH
N
input:
symmetrical
unsymmetrical
asymmetrical
test generator

Fig. 2.6: Picking up of interference voltages in network lines.<i>

<i>: Ernst Habiger: EMV, Huthig Buch Verlag Heidelberg (1992), ISBN 3-7785-2092-X

2.6 EMC-techniques to measure the stability against interference

The question is: what kind and what intensity of an electromagnetic interference can an apparatus cope with without limiting it regarding its technical function. For that test generators are used and with their help interference signals are produced in the test object. Fig. 2.6 shows the possibilities of a symmetrical, of an unsymmetrical and of an asymmetrical production of interference voltage signals in the power supply lines of an examinee. Over and above that the testing possibilities and testing methods are numerous. In principle the following processes are used:

1. The simulation and production of interference factors typical for the network like harmonic waves on network voltages, overlapping signal voltages, changes of network voltage, decrease of network voltage, network interruptions, network unsymmetries and network frequency changes.
2. The production of both energy-poorer and energy-richer interference impulses like they can form in energy-networks by switching acts or by the effect of lightning (burst).
3. The simulation of the course of events when static electricity discharges.
4. Low-frequency magnetic fields like those that can form by network frequency operating and loading currents or by short-circuit and lightning currents in the form of a pulse.
5. The stability against interference against the influence of an electromagnetic field also called radio interference firmness. For this purpose high-frequency generators and broadband antennas are used to expose the examinee to electromagnetic fields in a frequency range between 10 kHz and 18 GHz. At the moment tests are only performed between 27 and 500 MHz. The modulation of the carrier wave should be possible to be able to imitate the interferences by radio technology as realistic as possible. Thereby the field strength values can by all means reach up to several 100 V/m.

In accordance with expectation the result of this irradiation with an outside field is that every conduction path and every wire can act as an antenna and therefore can produce high-frequency currents and measurable potentials. Building parts of the analog technology as a consequence battle with problems of drift whereas with digital logic parts and computer parts the danger exists that switching states change unintentionally. Let us remember again the overlap effect of fig. 1.4. The electromagnetic wave itself marches with the speed of light through the examinee. When a small part of the wave finds an object that it can use as an antenna than the localized overlap comes into play. This than as a cause is responsible for the effective and measurable antenna currents. Until here the text books help us to explain the problems that happen and to remove them logically.

However time and again cases are reported where textbook physics can't help us any further. Spectacular cases even came to court like e.g. the ABS (Antilock Braking System) of a truck that had failed due to EMC-interference radiation. As a consequence the brakes had failed. When after that the local radiation pollution is measured no anomaly at all can be discovered. The measurable field strength is not higher as is usual in the whole area. Maybe you also have made the experience that often the causes can't be found when your computer suddenly "crashes" out of the blue.

Here the mentioned vortex of the electric field is capable to deliver plausible explanations because it is not bound to the geometry of an antenna and in addition is highly active without being detectable with the usual EMC measurement methods of the interference radiation measurement!

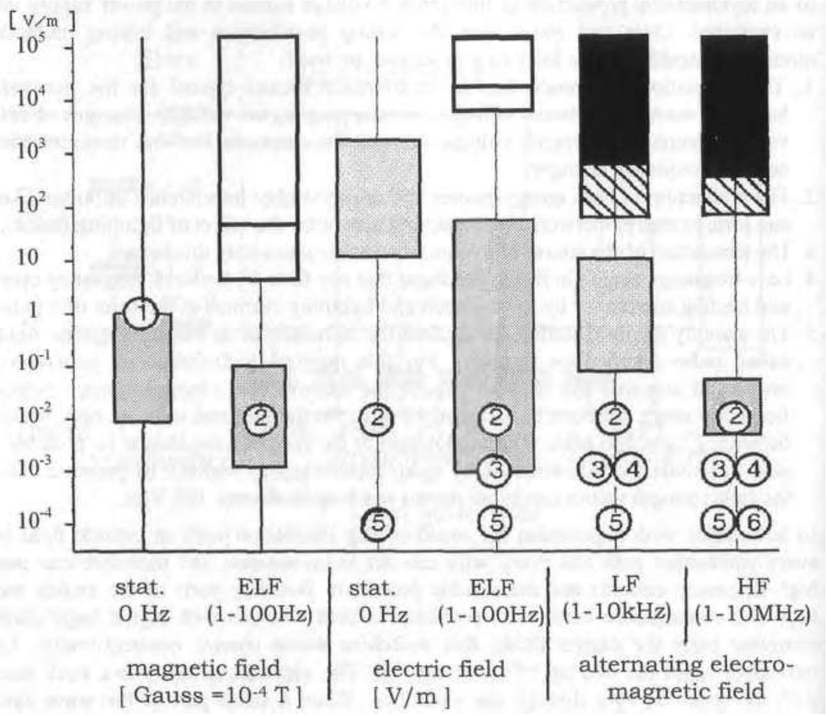

Fig. 2.7: Intensities of natural electromagnetic fields
(electric, magnetic and electromagnetic fields) ▨

In addition to fig. 1.1, page 3,
magnetic field in Gauss (= 10^{-4} T),
electric field in V/m.

<i>: A.S. Presman: Electromagnetic Fields and Life.
Plenum Press, New York - London, 1970

2.7 Tasks of the biological compatibility

The second leg of the environmental compatibility (EC) forms the biological compatibility (BC) besides the electromagnetic environmental compatibility. Whereas the interests of the EMC are looked after by electrotechnical engineers and electrophysicists, with the BC also doctors, biologists and architects are addressed.
Moreover this leg is already very old and already existed long before artificial interference sources could be created by mankind. The interaction between the arising interference sources in nature and the biological systems in general and specially men always interested the doctors, the priests, the druids and geomants, that not seldom looked after several functions in personal union equally. Unfortunately they as a rule have kept their knowledge and capabilities as secret knowledge, didn't make any recordings and merely initiated and trained their respective successors. Through that a great deal got lost and today non-medical practitioners, homeopaths and esoterics trouble themselves to take up to the far-reaching buried knowledge.
Because this concerns pure knowledge by experience, often the with the topic occupied persons themselves are not capable to say anything about the physical backgrounds and causes. One concentrates entirely on the under certain circumstances reachable results and only in rare cases on reproducable effects. In some areas the scientific assignment already has succeeded, have parascientific phenomena managed their admission in the so called "exact sciences", but in most experience disciplines the assignment is still due. There still is a lot to do here.

In the time as there not yet were operated any artificial interference sources on our planet, the senses of man naturally were a whole lot sharper for his electromagnetic environment as today. Today, where there scarcely is a place on earth where we are not irradiated by terrestrial transmitters, by satellites or by the netfrequency that is measurable everywhere. In the bluntness of our senses perhaps the hybris of modern man is founded, with which he wants to rise himself above esotericism, geomancy and other sciences of experience and thereby dispute the electric and magnetic fields their biological effectiveness.

The fields of natural origin form an electromagnetic environment for men, that they have adapted to and that they probably need for a life in accord with nature. The evolution has taken care for a corresponding adaptation.
In fig. 2.7 in addition to the limits from fig. 1.1 the intensities of natural electromagnetic fields are registered<i>. They lie clearly lower as the recommended limits but exactly in the area wherein the first reactions of living beings are observable.

When we ask us how much electromagnetism is good for us and how much harms us so the obvious answer is: exactly as much radiation as nature dictates in the fluctuations between day and night, between the months, years and in the end between the cycles of sunspots of 11 years. Here the guide value is found that man and nature have adapted themselves to. In fig. 2.7 the corresponding area between the natural minimum and maximum values is given.

Irradiation strengths of the field radiation in the biosphere and how the optical windows are situated in the atmosphere<i>.

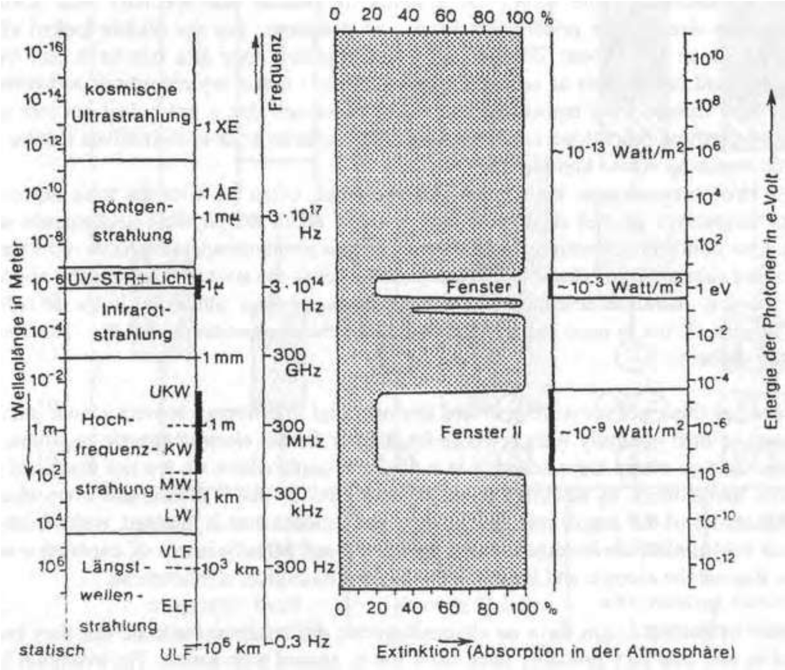

Fig. 2.8: Absorption dependent on frequency in the atmosphere

<i>: entnommen aus: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligk.), 5. Aufl., Bild 8, Seite 14, Verl. Moos & Partner Milnchen, ISBN 3-89164-058-7

2.8 Natural fields

Our electromagnetic environment has something to offer:
From the magnetic north pole to the magnetic south pole of the earth run the field lines of the earth's magnetic field that we are exposed to. With a compass we use the vector character of the magnetic field to fix our position. The induction averaged over time is approx. 50 uT. But it is overlapped by short-time fluctuations caused by geomagnetic storms in the ionosphere.

These storms again are caused by the eddy currents and the currents of charged particles that come from the sun. At the same time these eddy currents in the ionosphere together with the earth's magnetic field form a protective shield with a excellent screening effect for us inhabitants of earth.
In several layers like for instance the ozone and Heaviside layers a filtering and damping until the complete suppression of the very broad cosmic spectrum is caused. This extraterrestrial spectrum of radiation doesn't leave a single frequency out and has a lethal intensity for us.

Only for a little window in the frequency spectrum, radiation can pass almost undamped, as can be seen in fig. 2.8: the light with the spectrum of the colors. For this nature has donated man a sense organ so that man can protect himself against too high dose values. After all, who will look voluntarily into the sun? We only get into trouble when our sense organ doesn't function any more (for instance in the fringe range of the visible spectrum, the UV-range).

For other frequencies of electromagnetic radiation man neither has a sense organ but that doesn't mean that he is not influenced by these. Here, as in the UV-range he only indirectly notices that he has got too high a dose when he has to discover some influences on his well-being and his health. Without the help of neutral measurement apparatus he himself by no means is in a position to make a connection between an excessive exposition to radiation and his health problems.

When natural field strengths should be used as a measure for technical limits, so there should be paid attention to the fact that nature doesn't know intense continuous irradiation. The values are subject to powerful fluctuations that leave men and nature the chance to regenerate.
The television stations not even think it is necessary to reduce their broadcasting power after the end of broadcasts and further sprinkle the sleeping population with test signals, with senseless pictures of going by underground or nonstop program advertisements. People need the intermissions. That again shows how good nature means it with us.

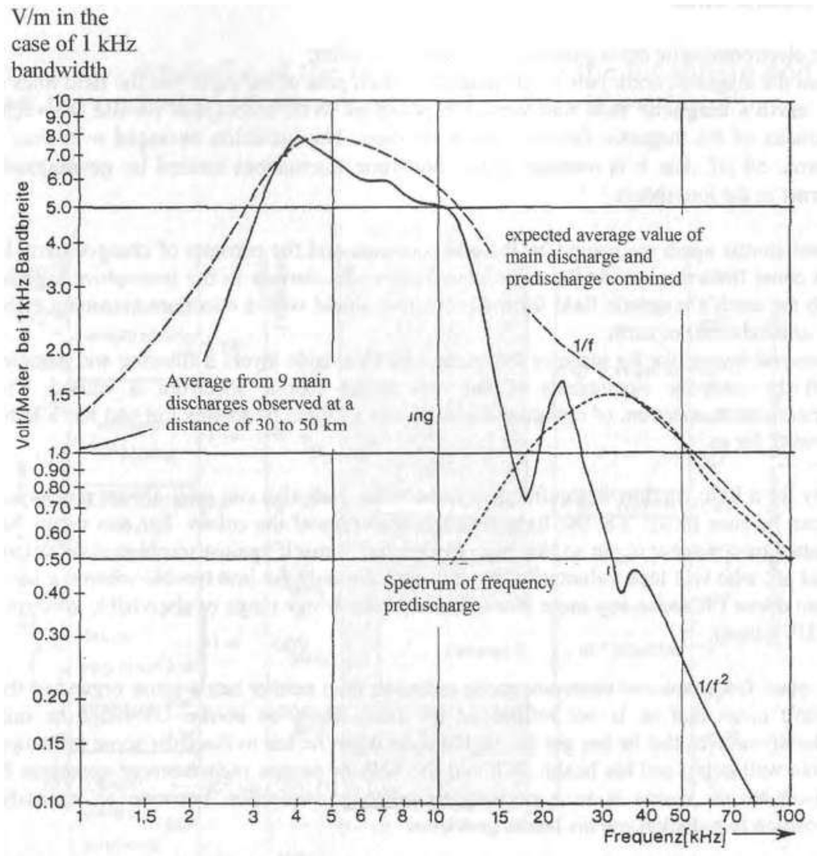

Fig. 2.9: Spectrum of frequency of one lightning, measured field strength at a distance of 1.6 km from the place of origin. See Watt and Maxwell:[i,ii]

from:

<i>: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligk.), 5.Aufl., Bild 38, Seite 27, Verl. Moos & Partner Munchen, ISBN 3-89164-058-7, see <ii>.

2.9 Electricity of the air

The electrostatic field strength of the earth lies between 50 and 500 V/m. That is a whole lot considering that voltages off 60 Volts can be lethal for man. But a living person distorts the terrestrial E-field for reason of his own field and his electric conductivity, so that there exists no danger for him as long as he doesn't grow into the sky.

Maybe the dinosaurs had to become extinct because they were to big and for instance the E-field near the ground had risen with a jump by a meteorite that brought a high charge from the cosmos. That would explain why the smaller living beings could survive the natural disaster.

Also the thunderstorm electricity can become life-threatening. Burns, heart and brain failures are the most common consequences. After all the probability to be struck by lightning is four times higher as to have six right ones in the lottery.
Over the lightning channel of approx. 1 meter in diameter charges between 10 and 200 C are transported what results in current strengths of 2000 up to 200,000 A. The main discharge lasts between 10 and 50 usec. With the preceding and all the following discharges it lasts over a second.
Field strengths on the scale of 4,000 V/m are typical but in a distance of 5 km these wear off to 8 V/m. The frequency spectrum of a lightning reaches 4 powers of ten into the range of the radio waves. In fig. 2.9 is shown the field strength measured in a distance of 1.6 km from the place of origin<i,ii>.

The origin of lightnings is still an unsolved problem after the well-known models (Wilson) arc not in a position to explain the reason for the origin of the potential difference of more than 100 million Volts required for the ionization of the air. Also the lightnings that struck in the direction of the ionosphere still are mysterious. We'll have to come back to this<iii>.

<i>: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligkeit), 5. Aufl., Bild 38, Seite 27, Verlag Moos & Partner Munchen, ISBN 3-89164-058-7, according to <ii>.

<ii>: A. D. Watt and E. L. Maxwell: Characterisic of Atmospheric Noise from 1 to 100 Kc/s. Symposium on the Propagation of VLF Waves, Boulder, Col., Paper 35, Jan. 1957.

<iii>: see Part 1, chapter 5.4 and part 2, chapter 14.11.

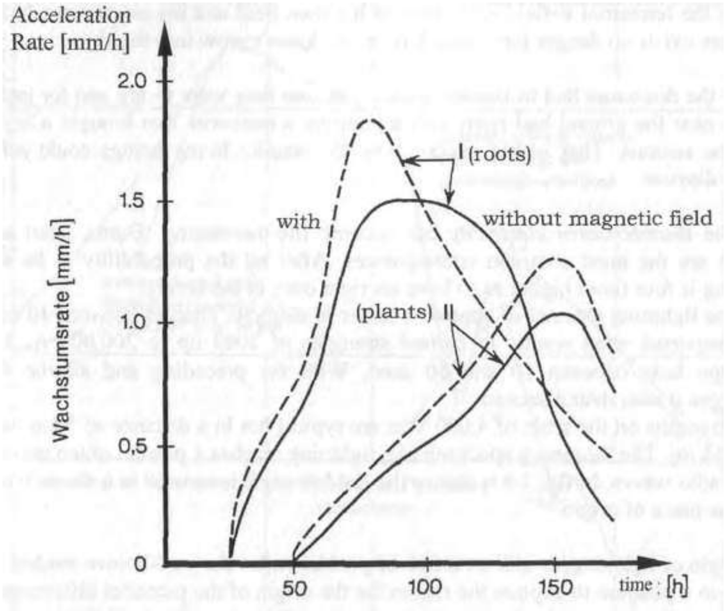

Fig. 2.10 A: Effect of a static magnetic field of 100 kA/m (0,12
Tesla) on the root (I) and on the plant (II) of barley seeds.
plants in the magnetic field: dotted line
plants for checking: drawn line
according to Novitskii[iii]

<iii>: Novitskii, Yu.I.: Effects of a magnetic field on the dry seeds of some cereals. Proceedings of Conference on the Effect of Magnetic Fields on Biological Objects, Moscow, p. 52, 1966.
Taken from: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligkeit), 5. Aufl., Bild 72, S. 73, Verlag Moos & Partner Munchen, ISBN 3-89164-058-7

2.10 Biological effects

The in fig. 2.1 indicated connection between the EMC and the BC, by some authors inofficially described as EMC-environment (EMCE), describes the effect of artificial fields on biological systems. This concerns the sensitive range of tasks that is discussed extremely controversially in the public. The problem thereby is that the artificially produced field strengths lie above the natural field strengths for several scales. In the thirties first reports about troubles were provided by navy radio operators that complained about headache, dizzyness, concentration failure and indisposition. Besides these negative reports concerning the so called "radio operator disease" at the same time medical usages concerning high-frequency therapy were tested. In the beginning this diathermy called healing method still was a short wave irradiation. Today it is extended into the microwave range and uses the thermal effect of electromagnetic rays. The increased temperature of the tissue causes an increased local blood flow. This supports healing processes, loosens cramped muscles and can help in case of rheumatic fever. The advantage of the HF-irradiation compared to contact heat by a hot-water bottle or by infrared rays is the higher penetration depth. Herein short waves are superior to microwaves. But microwaves can be better focussed on a certain part of the body. Is the temperature further increased, so the tissue is damaged. This is used for the treatment of cancer and is called hyperthermy. Because cancer cells as a rule are flowed with blood worse than healthy cells, they are more sensitive to heat and therefore are faster destroyed than healthy cells at a correspondingly increased temperature. In this way for Instance in the USA cattle with a so called cancer eye are treated. For that the spot suffering from cancer is irradiated with 2MHz-waves for 30 seconds with a handheld apparatus of 10 Watts broadcasting power. The rate of succes is given to be 90%! The method of hyperthermy has not yet been able to establish in the area of the medicine for humans. Also at our college corresponding research work is carried out in co-operation with the radiological clinic of the university of Freiburg (Germany)<i>.

The thermal effects of high-frequency fields are therefore well-known and subject of scientific research. On the other hand and in spite of numerous publications, non-thermal effects even today are denied by some scientists as non-existent<ii>. Here only a few counter-examples will be listed.

Fig. 2.10 A shows the effect of a static magnetic field of 0.12 Tesla on the root (1) and on thc plant (II) of barley seeds. The readable effect is an acceleration of the growth of the treated seeds (dotted line) compared to the plants for checking (drawn line)<iii>.

<i>: H. Schulz, W. Oppitz: Lokale Hyperthermie durch hochfrequente Wirbelstrome, Medizin Technik 1, 1987.

<ii>: G. Nimtz: Mikrowellen, Einfuhrung in Theorie und Anwendung. 2. Aufl., BI-Wissenschaftsverlag 1990, ISBN 3-411-03203-0

<iii>: Novitskii, Yu.I.: Effects of a magnetic field on the dry seeds of some cereals. Proceedings of Conference on the Effect of Magnetic Fields on Biological Objects, Moscow, p. 52, 1966; taken from: H.L. Konig: Unsichtbare Umwelt, fig.72, p. 73, Verlag Moos & Partner Munchen, ISBN 3-89164-058-7

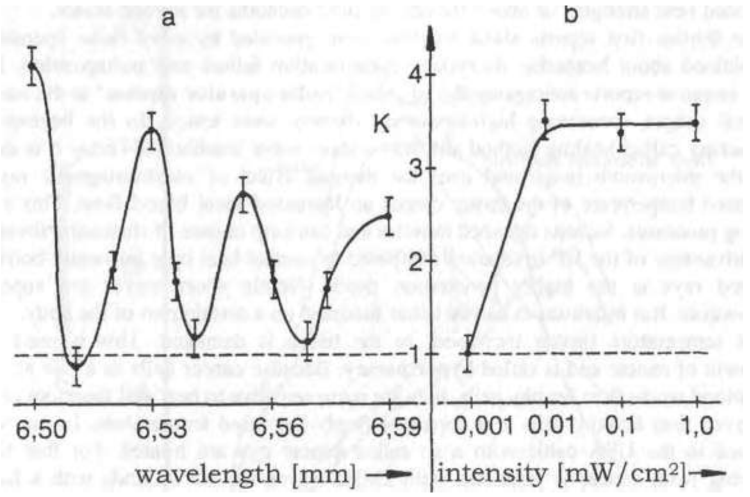

Fig. 2.10 B: Measured increase in the production (K) of colicin by colibacteria as a microwave effect

a) as a function of the wavelength,
b) as a function of the intensity of the microwaves,

according to Keilmann<i>.

---

taken from:
<i>: H. L. Konig: Unsichtbare Umwelt (Wetterfuhligkeit), 5. Aufl., Bild 106, S. 111. Verlag Moos & Partner Munchen, ISBN 3-89164-058-7

A static field naturally produces no induction and hence no heating will arise. In the case of alternating fields the thermal effect in experiments is excluded by working with cxtremly low stimulations. The example after fig. 2.10 B shows the measured increase in the production (K) of colicin by colibacteria at only 0.01 mW/cm^2 microwave power<i>.

In addition the example provides the interesting information that obviously only a certain frequency and it's harmonic waves increase the production, other frequencies on the other hand remain inactive. Because only minimal field strengths are used it more likely concerns an information-technical as an energetic effect (curve a). This statement is supported by the observation that an increase of the intensity not at all necessarily as a consequence also increases the production (curve b). What the colibakteria need is obviously neither energy nor heat but only a certain frequency that stimulates the colicin production or the growth.

Should it really be confirmed that biological effects of electric and magnetic fields can be produced by certain frequencies and can't happen by an energy transition so the discussions about limits must seem ample meaningless.

Maybe the one or the other in thought already draws a connection to the acceleration, the acclerated growth of kids, which is observed world-wide and stronger in cities than in the country. It started for approx. 100 years simultaneous with the electrification of the homes in the towns. In Asia the acceleration and also the electrification have started later. Other growth stimulating effects like radio waves, X-ray examinations, atomic tests and provable also the nourishment with vitamin B6 happened only until much later and at the most could support the actual effect.

But how should a proof be offered when anyway the field strength not at all can have a decisive influence on the growth of man after the statement of fig. 2.10 B? Which information is authorative? Where lies the responsible frequency window? Does the information actually manifest as frequency? Is the authorative influential factor also in this case not at all noticed and measured?

A lot of pressing questions are still outstanding. But in any case the numerous influential factors detected in experiments do not at all let themselves reduce to a sole factor, for instance the nourishment. For a family doctor it may be comfortable to be able to give an easy explanation: „Cause is the nourishment!" With such a reductionism on the other hand the actual cause stays in the dark and the asked questions in this way won't let themselves be answered.

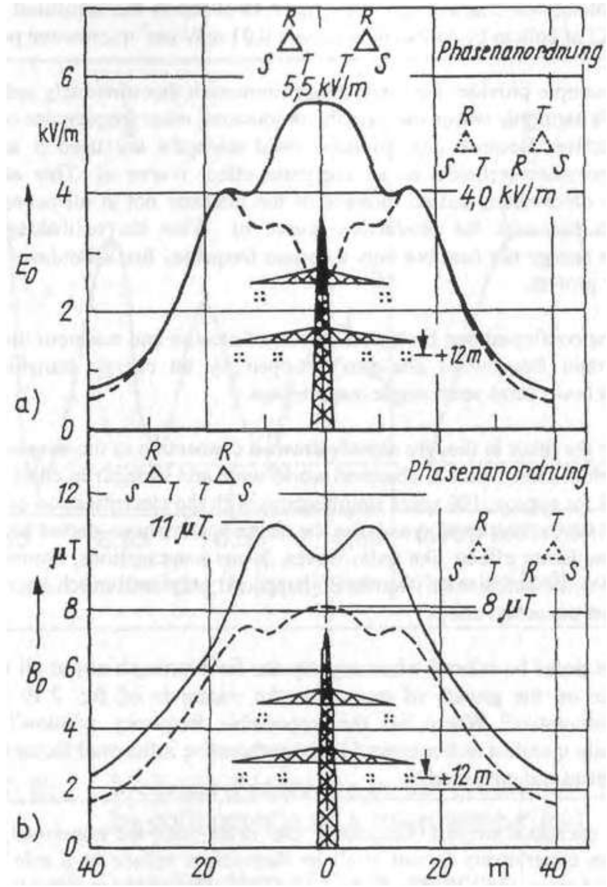

Fig. 2. 11: ____ profile of the 50-Hz-field on the ground at 380kV/1kA in each circuit.<i>
a) electric field,
b) magnetic field.

____ taken from:
<i>: E. Habiger u. a.: Elektromagnetische Vertraglichkeit, fig. 7.3, page 147 and Fig. 7.1, page 146, 2nd Ed., 1992. Berlin, Miinchen: Verlag Technik.

2.11 Artificial fields

The scepticism of people feeling close to nature is especially directed against artificial fields that man can't see nor hear nor smell. Objects of doubt are first of all the installations for the creation, distribution and conversion of electric energy.

An essential role plays the fact how close the supplying and the draining conductors are to each other, so that the respective fields can compensate each other. The worst solution one can think of is realized at the electrified railway. Here the rails and the earth are used as the draining conductor for the current while there exists an ample large distance to the supplying conductor. A compensation is almost impossible thus causing gigantic interference fields that are detectable even at a distance of 10 kilometers. The increased putting on of railway engines feeded by rectified current is likely to aggravate the set of problems because the non-sinusoidal absorption of current is strongly afflicted with harmonic waves.
With high tension transmission lines (fig. 2.11) the interference field strength is reduced when the three cables are suspended with only little distance between them. But even the selected phase order can play a role. Of course the optimal compensation effect is guaranteed with cables in the earth. But these are expensive and technically not realizable for such high voltage levels.

In the eighties also the computer screens got in the headlines. The terminals are furnished with a cathode ray tube and have a very broad radiation spectrum that already starts at 0 Hz. Here already static maximum values of 64 kV/m are measured[i]!

I'requency range	Measured maximum values E_{max} resp. H_{max}	30 cm in front of the screen
static field (0 Hz)	64 kV/m	
30 Hz. 60 Hz	10 V/m and	
	0.2 - 1 A/m	
5 Hz - 1 kHz	1800 V/m and	10 V/m and
	4 A/m	0,6 A/m
50 Hz - 0,5 MHz	1 A/m	
15 kHz-220 kHz	50 V/m and	15 V/m and
	1.1 A/m	0,17 A/m
3 MHz -300 MHz	< 0,2 V/m	

Fig. 2.11: Electromagnetic fields from screens[i]

system about	used frequency	measured value
radio broadcasting station (MW 2O kW)	6OO kHz	2...17 V/m
radio broadcasting station (SW 1OO kW)	15 MHz	1...25 V/m
SOS-transmitter on a ship (1OO W)	41O kHz	1...3 V/m
Epitaxiedevice (induction oven)	45O kHz	37...4OO V/m
HF-welting press (welding of plastic foils)	27,12 MHz	70...85 V/m
radar on a ship (TRN 311)	9,3 GHz	1...3O uW/cm^2
radar of an airplane	9,2 GHz	45O... 28OO uW/cm^2
domestic appliances measured in a distance		of 3O cm:
hand mixer	5O Hz	5O V/m
refrigerator	5O Hz	6O V/m

Fig. 2.12: ____ the electric field strength resp. Power density in our environment.<i>

____ taken from:
<i>: E. Habiger u. a. : Elektromagnetische Vertraglichkeit, Fig. 7.2, page 146, 2. Ed., 1992, Berlin, Munchen: Verlag Technik, ISBN 3-341-00993-0

2.12 Protection against artificial fields

Artificial fields more or less always occur in the neighbourhood of electric apparatus and installations. Especially problematic are those that work with the free radiation of electromagnetic fields, that is all the radio broadcasting stations, handheld and radar transmitters.
Herewith it is important that not needed parts of the antennas are shielded, that antennas with little close by field pollution are used and that the stand should be situated at least 3 km remote from inhabited areas. For instance at radar installations damping values of 10 dB and more can be obtained only by using a corresponding tree growth.
This obviously concerns a damping of the waves in a dielectric manner. We'll have to come back to this because textbook physics does not know a corresponding damping term in the wave equation.

The radiation leaking out in case of the high-frequency welding of plastic foils and of the microwave oven should be minimized.
In the case of induction ovens or of motors an active shielding often causes problems so that for simple domestic appliances like a hand mixer and especially for the electric hair-dryer non proportionally high field strength values are measured. Fig. 2.12 informs about it.
Protective measures for the operator are hardly possible. To protect uninvolved people not only the apparatus but also the rooms and eventually whole parts of buildings had to be shielded and grounded.

Sometimes also fairly simple possibilities are helpful like e.g. the usage of a remote control. By clearing away the cable salat at the workplace and at the sleeping place induction loops can be removed. Alarm-clocks operated by batteries should have preference over those operated by the network. Mattresses with metal in them and spring-beds which clearly act as an antenna should be avoided. In extreme cases even so called "current-free switches" and shielded network lines are recommended (fig. 2.13).

In the area of the network supply lines a choking coil can help decrease the spreading of high-frequency interference radiation. It is especially important that all the conducting metallic objects like e.g. water pipes, heatings, steel racks, machines, switching racks, steel armaments and metallic windows should be grounded properly, because otherwise extremly high static charges could result instead of a shielding. Construction biologists recommend to when possible do without metals when building houses and furnishing, what of course is only realizable with limitation.

In any case numerous measures are known that to a lesser extent find their legitimation in classical physics, but more likely in precaution. As long as we do not know which phenomenon causes the electrosmog and we don't have a measuring-instrument at our disposal, precaution is really the only thing we can do irrespective of the effectiveness of the measures and of the arising costs!

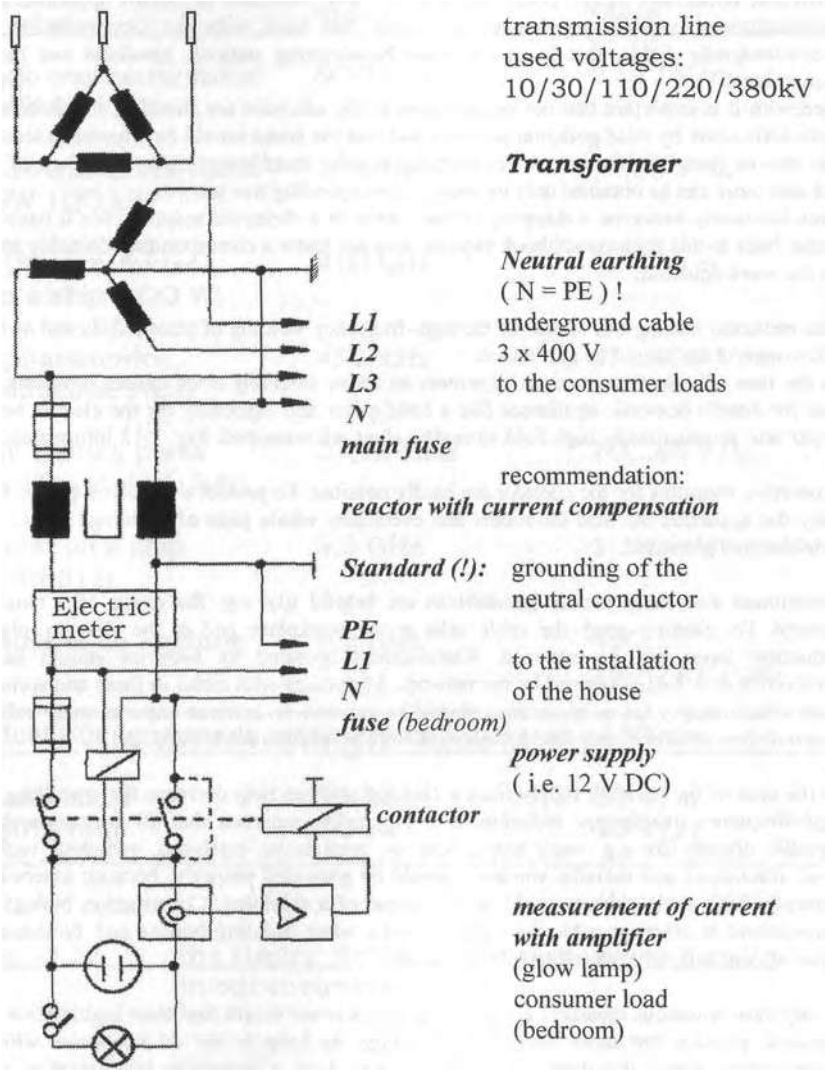
transmission line
used voltages:
10/30/110/220/380kV
Transformer
Neutral earthing
(N = PE) !
L1
L2
L3
N
underground cable
3 x 400 V
to the consumer loads
main fuse
recommendation:
reactor with current compensation
Standard (!): grounding of the neutral conductor
Electric meter
PE
L1
N
to the installation of the house
fuse (bedroom)
power supply
(i.e. 12 V DC)
contactor
measurement of current with amplifier
(glow lamp)
consumer load
(bedroom)

Fig. 2.13: About the circuitry and the problems involved with a "current-free switches" installation

2.13 Unsolved tasks

The report concerning the actual state of research could be continued at will. But the expositions should suffice, to understand what are the tasks of the electromagnetic environmental compatibility (fig. 2.1) and which questions still have to be supplied with a solution. One can get deeper into every of the addressed points and then discover that some questions can be expressed sharper and maybe conceivable answers can be found, but at the same time and unavoidable the number of new questions increases faster.

Let us again take up the example of the handheld wireless telephones (chapter 1.2). At least it now is clear to us that the usage of the built-in microwave antenna of a handy is problematic. In the interior of an automobile it never should be used. If, however, one uses the antenna installed on the outside on the sheet metal then the damping and screening effect of the sheet metal chassis is advantageous at least for the handy user.

With that of course the central question is not answered. The question of what the cause is for the interfering and at worst health endangering effect of the HF-radiation. Field freedom we can't find anywhere on this world. Possibly we even need the fields. But then the question is how much is necessary, how much is healthy and how much makes us ill.

The gap of explanation especially gets clear in the case of the static or of the low-frequency magnetic field: away from technical interference sources normally fields on the scale of 10 nT are measurable. Construction biologists say that 20 nT, so twice that value, should not be exceeded at the sleeping place and maybe 50 nT at the desk. These values however are determined purely empirical.

When a person is examined in a nuclear magnetic resonance tomograph that person is exposed to a field that lies between 0.2 and 1.5 Tesla. that is a value 7 till 8 powers often higher than before mentioned without this leading to the death of that person. Entirely on the contrary this method is regarded as especially caring and safe compared to the X-ray examination.

Here again the legitimation of the thesis put forward is entirely comfirmed. The thesis that the well-known physically measurable and controllable phenomena can not be considered as a cause and that possibly a until now undiscovered field phenomenon should be called to account.

Should such a phenomenon exist it should be derived, calculated and proved. We must go to endless troubles and try everything. The actual difficulties wherein the electromagnetic environmental compatibility is stuck are a challenge.

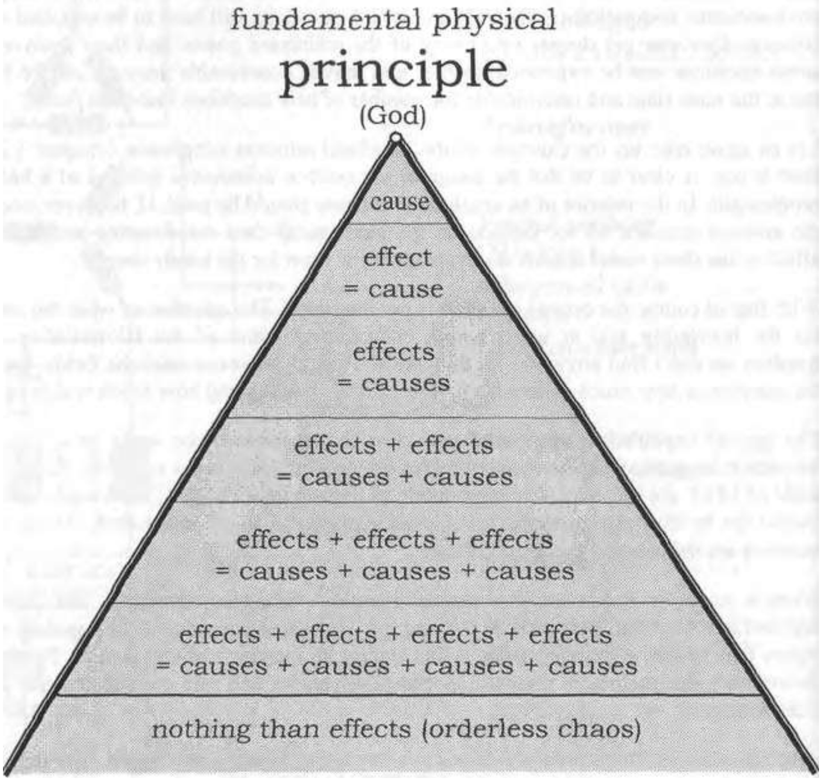
fundamental physical
principle
(God)
cause
effect
= cause
effects
= causes
effects + effects
= causes + causes
effects + effects + effects
= causes + causes + causes
effects + effects + effects + effects
= causes + causes + causes + causes
nothing than effects (orderless chaos)

Fig. 3.0: pyramid of causality

vortices are a consequence of the principle of causality

3. Approach

In the question, if there exists a still unknown phenomenon that influences the electromagnetic environmental compatibility, we must fall back far until upon the roots of our physical understanding. Here we find a fundamental principle that until today is not doubted and that is regarded as elementary, the principle of causality. Every result of a measurement, every interpretation is checked for causality and only after passing this examination it is accepted and published.

This principle of cause and effect has established, not only in physics but also in many other disciplines of science. Is an effect observed, so there immediately is asked for the cause. This principle encounters us in daily life<i>.

When all observable and measurable effects ever can be assigned to a cause without force and without exceptional regulations then the logical result is a pyramid of causality. On top a fundamental physical principle is found, that is regarded as given by nature or as given by god and that with its properties is responsible as the cause for different effects. These effects again appear as the cause for new effects and so on (Fig. 3.0). Sometime we have removed us so far from the top of the pyramid that a direct reference to the describable effects can't be made anymore, so the impression could arise that it concerns an isolated and independent discipline. We should take care not to think in such a monocausal way, because both delimitation and avoidance of interdisciplinary working methods will inevitably steer us into a dead end!

This pyramid of causality stands for the vision of a "unified theory", like it is demanded and sought-after by numerous research scientists. But as long as it is not found, we'll have to do with unsolved problems of causality. About this any number of examples can be given.

A physical principle based on the principle of causality is the vortex. This the eddy current demonstrates us clearly. The cause for its origin is an alternating field. According to Faraday's law of induction this induces a voltage that in a conducting medium results in a current according to Ohm's law. Around this current according to Ampere's law an alternating field forms, that points perpendicular to the current and overlaps the original alternating field. This induced field first of all is an effect that overlaps the cause and itself becomes the cause. The effect that follows from that further overlaps and forms a new cause etc. In this way vortices form.

Vortices quasi represent the principle of causality.

<i>: When for instance a woman complains: "Doctor, my left knee hurts" (effect). The doctor diagnoses the cause: "Yes, that comes with age!" With that causality is established. "But doctor", says the woman, "my right knee is exactly as old as my left knee!" And already the doctor has a new problem of causality.

Principle of Causality:

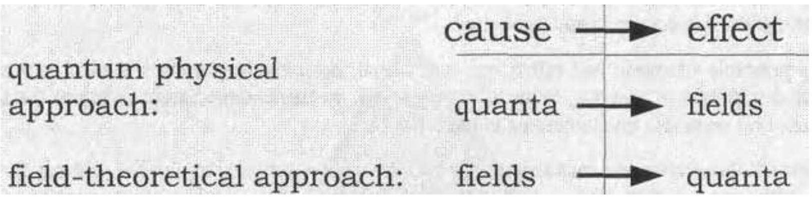

	cause		effect
quantum physical approach:	quanta	→	fields
field-theoretical approach:	fields	→	quanta

violations of the principle of causality:

1. monopoles exist
2. starting point for the strong interaction
3. fields and quanta are a cause at the same time
4. hypothetical particles (gluons, quarks, etc.)
5. transmission of information with speeds faster than light
    - with photons (University of Berkeley)
    - with microwaves (University of Cologne)
    - with laser beams (Technical Univ. of Vienna)
6. transmission of energy with speeds faster than light
    - with scalar waves (Nicola Tesla)

Fig. 3.1: Causality or the principle of cause and effect

3.1 Principle of causality

Our physical view of life strictly obeys the rules of causality, the principle of cause and effect. But there are numerous cases, where causality at first could not be fulfilled anymore. Here alternate solutions had to be found to not endanger this very successful principle. A few examples should clarify this:

1. Technically it is impossible to produce a magnetic monopole. When a north pole is produced then the accompanying south pole is also formed. In the same way only the positive and negative pole can be assembled as the so called dipole. In the microcosm however we find monopoles. Electrons are such particles. To restore causality we must grant the microcosm its own laws that are not valid in the macrocosm! But this monocausal hypothesis contradicts the observation that the microcosm represents an image of the macrocosm and vice versa. Doubts if this assertion is allowed are reasonable.
2. Like charges repel each other and that the more the smaller the distance gets. In an atomic nucleus positively like charged protons are together at the smallest possible room without any repulsion happening. Arithmetically seen all atomic nuclei would have to explosively fly to pieces. But because this doesn't happen, shortly a new and supposedly fundamental interaction, the strong interaction, was introduced to save causality. Nevertheless this interaction now holds the like particles in a not explained manner together. Causality could be obtained only by the introduction of a new fundamental phenomenon.
3. When causality should hold as the supreme principle, it should be demanded with priority for the fundamental phenomena of physics. Instead, in quantum electrodynamics the particle is attributed the same physical reality as the field. With the wave-particle duality Heisenberg has given out the corresponding formula of compromise. This slaps the face of the principle of cause and effect.

Causality on principle allows only two approaches for a solution: the quantum physical approach. which holds the quanta as the cause for the fields, and the field-theoretical ašproach. wherein only the fields act as the cause. For both approaches there are good arguments. The field theorists cite that fields can exist also in the vacuum, so that there exist fields without particles but never particles without fields. Against that the quantum physicists hold that somewhere, even when quite far away, there exist particles and that the measurable fields merely are their action at a distance.

Both approaches first arouse the impression to be fully equal. In the course of the discoveries in the area of quantum physics, the corresponding approach has been able to establish. But it demands that all phenomena have to be understood as a consequence of particles. So should gravitons make gravitation possible, should gluons hold everything together and the quarks form the basic building parts. Meanwhile there is only worked with hypotheses. Out of poverty quantum physics meanwhile has said goodbye to strict causality, after the number of the violations of causality has risen that much and in every respect there is a lack of models of explanation. It seems as if the end is reached, as if the quantum physical approach to a large extend is exhausted.

Field-theoretical approach:

3rd Maxwell equation — 4th Maxwell equation

$$\boxed{\mathrm{Div}\,\mathbf{B} = 0} \quad (3.3) \qquad \boxed{\mathrm{Div}\,\mathbf{D} = \rho_{el}} \quad (3.4)$$

With the relations of material:

$$\boxed{\mathbf{B} = \mu \cdot \mathbf{H}} \quad (3.5) \qquad \boxed{\mathbf{D} = \varepsilon \cdot \mathbf{E}} \quad (3.6)$$

$$\mu \cdot \mathrm{Div}\,\mathbf{H} = 0 \quad (3.3^{*}) \qquad \varepsilon \cdot \mathrm{Div}\,\mathbf{E} = \rho_{el} \quad (3.4^{*})$$

H: source free vortex field — **E**: non-vortical source field

New field-theoretical approach: $\boxed{\mathrm{Div}\,\mathbf{D} = 0}$ (3.7)

thus: $\mathrm{Div}\,\mathbf{H} = 0$ (3.3^{**}) and $\boxed{\mathrm{Div}\,\mathbf{E} = 0}$ (3.7^{*})

H and **E**: source free vortex fields!

Einstein:

> "Is it conceivable, that a field theory permits us to understand the atomistic and quantum structure of reality? This question by almost all is answered with No. But I believe that at the moment nobody knows anything reliable about it"<i>.

Pauli:

> "The electric elementary quantum e is a stranger in Maxwell-Lorentz' electrodynamics"<ii>.

Fig. 3.2: The field-theoretical approach

<i>: A. Einstein: Grundzuge der Relativitatstheorie, S 162, Anhang II; 5. Aufl., Vieweg, Braunschweig 1974.

<ii>: W. Pauli: Aufsatze und Vortrage uber Physik und Erkenntnistheorie. Vieweg, Braunschweig 1961, entnommen aus:
H. G. Kussner: Grundlagen einer einheitlichen Theorie der physikalischen Teilchen und Felder. Musterschmidt-Verlag Gottingen 1976, S. 161.

3.2 Field-theoretical approach

The field-theoretical approach is the very much older one. Until the last turn of the century the world in this respect still was in order. Max Planck, by the discovery of quanta, has plunged physics into a crisis.
Albert Einstein, who, apart from his lightquanta hypothesis, in his soul actually was a field theorist, writes: „Is it feasible that a field theory allows us to understand the atomistic and quantum structure of reality?". This question by almost all is answered with No. But I believe that at present nobody knows anything reliable about it<i>".

By the way the "No" can be justified by the fact that the field description after Maxwell is by no means able to the formation of structure so that it is not possible for quanta to appear as a consequence. The field-theoretical approach could, obstructed by Maxwell's field theory, not further be pursued and to this until today nothing has changed.
Nevertheless it would be an omission to not at least have tried this approach and have it examined for its efficiency. Maybe the above mentioned problems of causality let themselves be solved much more elegantly. For this however the Maxwell theory must be reworked to a pure field theory. With the well-known formulation it offends against the claim of causality, since it is field theory and quantum theory at the same time. To Maxwell himself the quanta were still unknown, but today we know that the fourth Maxwell equation is a quantum equation:

$$\mathrm{div}\, \mathbf{D} = \rho_{el} \quad (3.4)$$

After this the electric field is a source field whereby the individual charge carriers, like e.g. electrons, act as sources to form in their sum the space charge density p_{el}. The other three Maxwell equations are pure wave equations. In this way the third equation identifies the magnetic field as source free:

$$\mathrm{div}\, B = O \quad . \qquad (3.3)$$

This for Pauli probably was the reason to call, "the electric elementary quantum e^- a stranger in Maxwell-Lorentz' electrodynamics"<ii>.

Let's return to the principle of causality according to which in the field-theoretical approach the fields should act as a cause and not the particles. In a corresponding field description quanta logically have not lost anything. It is only consistent to likewise demand freedom of sources of the electric field:

$$\mathrm{Div}\, D = O \quad . \qquad (3.7)$$

When the electric field is not a source field, then what is it? The magnetic field is a vortex field. Hence it would be obvious to also conceive the electric field as a vortex field. Numerous reasons speak for it:

1. A non-vortical gradient field, like it is formed by charge carriers, merely represents a special case of the general vortex field. Only by the generation of quanta a source field can form as a special case.
2. The electromagnetic wave teaches us the duality between the E- and the H-field that are directed perpendicular to each other and are in a fixed relation to each other. If one of them is a vortex field then also the dual field must be a vortex field.

Duality:

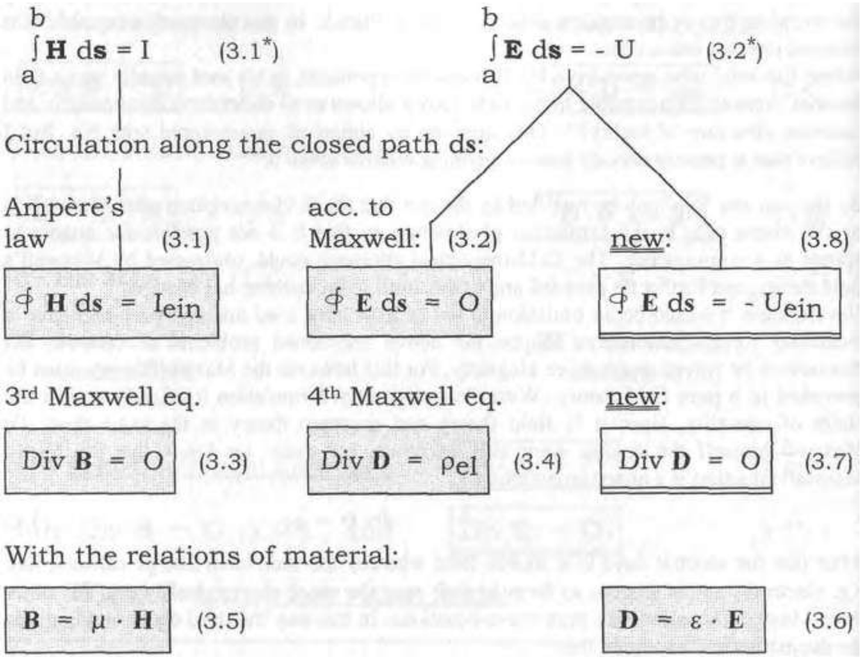
$\int_a^b \mathbf{H}\, d\mathbf{s} = I$ (3.1*)
$\int_a^b \mathbf{E}\, d\mathbf{s} = -U$ (3.2*)
Circulation along the closed path ds:
Ampère's law (3.1)
acc. to Maxwell: (3.2)
new: (3.8)
$\oint \mathbf{H}\, ds = I_{ein}$
$\oint \mathbf{E}\, ds = O$
$\oint \mathbf{E}\, ds = -U_{ein}$
3rd Maxwell eq.
4th Maxwell eq.
new:
$\mathrm{Div}\, \mathbf{B} = O$ (3.3)
$\mathrm{Div}\, \mathbf{D} = \rho_{el}$ (3.4)
$\mathrm{Div}\, \mathbf{D} = O$ (3.7)
With the relations of material:
$\mathbf{B} = \mu \cdot \mathbf{H}$ (3.5)
$\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (3.6)

Fig. 3.3: The dual field description

Dual approach according to Jackson<i> or Lehner<ii>:

$\mathrm{Div}\, B = \rho_{magn}$ because of the 4th Maxwell equation: Div D = ρ_{el} (3.4)

ρ_{magn} = magnetic monopoles should exist,
otherwise dual extension (3.8) not allowed!

Caution: closed loop conclusion!
Maxwell theory proves the correctness of the Maxwell theory.
Result: search for magnetic monopoles unsuccessful.

<i>: J. D. Jackson, Classical Electrodynamics, 2nd Ed. , John Wiley, New York, 1975, S. 251 - 253

<ii>: G. Lehner, Elektromagnetische Feldtheorie, Springer-Verlag Berlin, Heidelberg 1990, S. 35 - 36 und S. 533 ff.

3.3 Duality

Duality is a fundamental physical principle. Opposite, but one another complementing phenomena can be assigned to each other in pairs, like e.g. (see fig. 8.8):

	Yin (female)	——		Yang (male)
	south pole	——		north pole
	⊖	——		⊕
R [Ω]	resistance	——	R^{-1} [Ω^{-1}]	conductance
U [V]	voltage	——	I [A]	current
$\mathbf{E}$ [V/m]	electr.field strength	——	$\mathbf{H}$ [A/m]	magnet.field strength
Q [As]	charge	——	ϕ [Vs]	magnetic flux
$\mathbf{D}$ [As/m^2]	electr.displacement	——	$\mathbf{B}$ [Vs/m^2]	magnetic induction
ε_0 [As/Vm]	electr.field-constant	——	μ_0 [Vs/Am]	magnet.field-constant
ρ_{el} [As/m^3]	electr.charge density	——	ρ_{magn}[Vs/m^3]	magnet.charge density
	induction law	——		Ampère's law
	potential vortex	——		eddy current
	(vortex of the electric field)			(vortex of the magnetic field)

First of all we find the duality confirmed in the case of the electromagnetic wave spreading in a homogeneous medium. Here the field pointers of E and H are directed perpendicular to each other and are in a fixed relation to each other. But if the wave is damped in the presence of matter, for instance by eddy currents, then by basing on Maxwell's field theory the duality will vanish.

A good example for perfect duality provides the integral of a field strength vector along the path from a to b:

$$\int_a^b \mathbf{H}\, ds = I \quad (3.1^*) \qquad \text{and} \qquad \int_b^a \mathbf{E}\, ds = U \ . \quad (3.2^*)$$

Urn if the integration takes place along a closed path then the circulation yields:

$$\oint \mathbf{H}\, ds = I_{eing} \quad (3.1) \qquad \text{and} \qquad \oint \mathbf{E}\, ds = 0 \ . \quad (3.2)$$

According to Ampere's law (3.1) the magnetic field can thus form enclosed currents and spatially spreading eddy currents. The electric field on the other hand should be irrotational (3.2).

Let's take the case that the electromagnetic wave is damped by eddy currents and the magnetic field in this way becomes a vortex field. The electric field itself that, as said, is in a fixed relation and perpendicular to the vortex field H, will show all the vortex-typical properties. Hence nothing would be more obvious as to also grant the electric field a formation of vortices:

$$\oint \mathbf{E}\, ds = -U_{eing} \ . \quad (3.8)$$

Critics of this dual approach, like for instance Jackson[i] or Lehner[ii], point out that with reference to the fourth Maxwell equation the electric field should be understood as a source field:

$$\operatorname{div} \mathbf{D} = \rho_{el} \ . \quad (3.4)$$

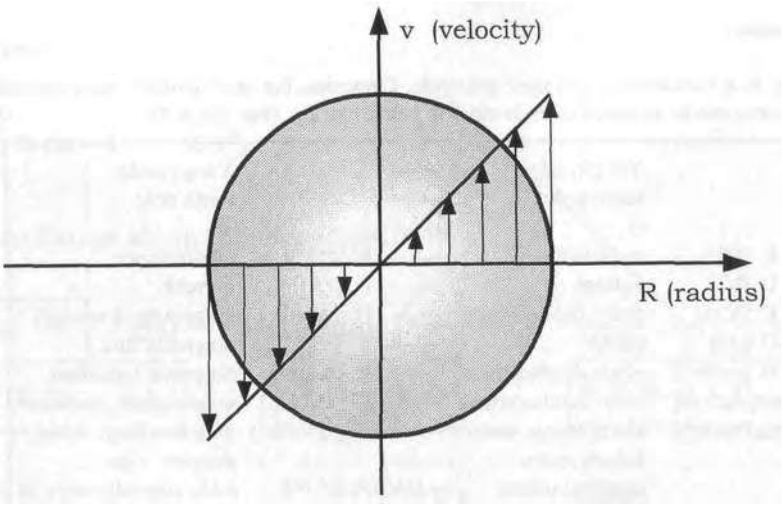
v (velocity)
R (radius)

Fig. 3.4a: Velocity distribution v(R) for a vortex with rigid-body rotation

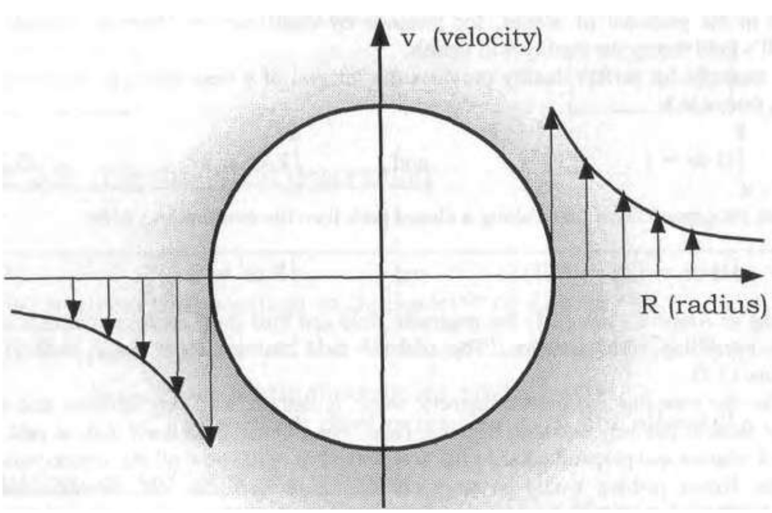
v (velocity)
R (radius)

Fig. 3.4b: Velocity distribution v(R) in a potential vortex (see Lugt[i]).

---

[i] Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; page 30 and 31, ISBN 0-89464-916-7

For a complete duality from the existence of electric monopoles, individual in the space charge density ρ_{el} contained charge carriers, the claim for magnetic monopoles is derived.
In spite of intensive search such north or south pole particles however until now could not be found. Herein from the sight of criticism is seen a confirmation for the assumption that Maxwell's field theory is self-contained and hence in principle may not be extended. The critics have a problem of causality: They postulate source fields that at the same time should be vortex fields. But if one asks how one should imagine such a field that is scalar and at the same time vectorial, then it looks as if no one has ever made any thoughts about it.

The from causality derived solution of the problem of lacking duality requires to extend the Maxwell theory in one point, by introducing the potential vortex of the electric field here and at the same time make a cut in another place:

$$\operatorname{div} D = 0 \tag{3.7}$$

With this formulation, the assumption of a freedom of sources in principle, the complete duality already is reached: Now neither magnetic nor electric monopoles exist (Fig. 3.3)!
At first we have to accept the loss of the electron hoping that the calculation in the end works out: the "exchange" vortices against particles, by which the quanta can be banned from the field theory, suggests that the elementary particles themselves are nothing else as spherical vortices that have found to an own physical reality.

3.4 Flow vortices

In fluid engineering convincing and strong indications for the correctness of the chosen approach can be found. It benefits us that hydrodynamic vortices are visible or can be the injection of smoke, e.g. in a wind-tunnel.
Already Leonardo da Vinci had observed at liquids that there exist two dual basic types of plane vortices: "Among the vortices one is slower at the centre than at the sides, another is faster at the centre than at the sides."
A vortex of the first type, also called "vortex with rigid-body rotation", is formed for instance by a liquid in a centrifuge that due to its inertia of mass is pressed to the edge because there the largest velocity exists. In an analogous way the electromagnetic vortex in electrically conductive material shows the well-known "skin effect" (Fig. 3.4a).
To explain the other vortex Newton describes the experiment where a rod is dipped into a liquid as viscous as possible and then is turned. In this potential vortex the velocity of the particle increases the closer to the rod it is (Fig. 3.4b).
The duality of both vortex phenomena becomes obvious when we make ourselves clear that in the experiment with the centrifuge the more liquid presses to the edge the less viscous the medium is. And that on the other hand the potential vortex forms the stronger the more viscous the medium is. As conclusion we read in text books that the viscosity of the liquid decides whether a vortex with rigid-body rotation or a potential vortex is formed.

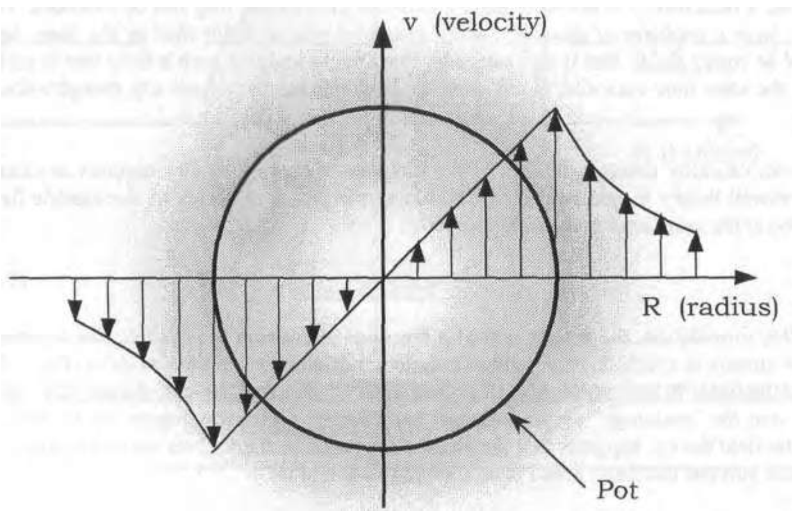

Fig. 3.5: Combination of a vortex with rigid-body rotation and a potential vortex (Lugt<i>).

<i> Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; ISBN 0-89464-916-7

3.5 Rankine vortex

When we, in a third experiment, immerse the centrifuge filled with water into a tough medium and let the centrifuge rotate, then inside the centrifuge a vortex with rigid-body rotation forms and outside the centrifuge a potential vortex forms (Fig. 3.5).

It is obvious that one vortex always causes the other vortex with the opposite properties and so the existence of one causes that of the other. So in the first case, that of the vortex with rigid-body rotation, outside the centrifuge potential vortices will form in the surrounding air, whereas in the second case, that of the potential vortices, the turning rod itself can be interpreted as a special case of a vortex with rigid-body rotation. Hence in all conceivable experiments the condition always is fulfilled that in the centre of the vortex the same state of "peace", that we can fix as "zero", prevails as in infinity.

When we take a tornado as an example, thus a whirlwind. In the "eye of the cyclone" there's no wind at all. But when I go away from this spot, then I'm blown to the outside. I can really feel the vortex with rigid-body rotation in the inside. If. however, I am standing on the outside, then the potential vortex tries to pull me into the vortex. This potential vortex is responsible for the structure and in the end also for the size of the tornado.

At the radius of the vortex, the place with the largest speed of the wind, an equilibrium prevails. The vortex with rigid-body rotation and the potential vortex at this point are equally powerful. Their power again is determined by the viscosity, which thereby fixes the radius of the vortex!

Therefore meteorologists pursue with interest whether a tornado forms over land or over water. Over the ocean for instance it sucks itself full with water. In that way the potential vortex increases in power, the radius of the vortex gets smaller and the energy density increases dangerously.

If the knowledge from hydrodynamics is transferred to the area of electromagnetism, then the role of the viscosity is taken over by the electric conductivity. The well-known current eddy occurs in the conductor, whereas its counterpart, the postulated potential vortex, forms in the bad-conducting medium, with preference in the dielectric. The duality of both vortices is expressed by the fact that the electric conductivity of the medium decides whether current eddies or potential vortices can form and how fast they decay, i.e. convert their energy into heat.

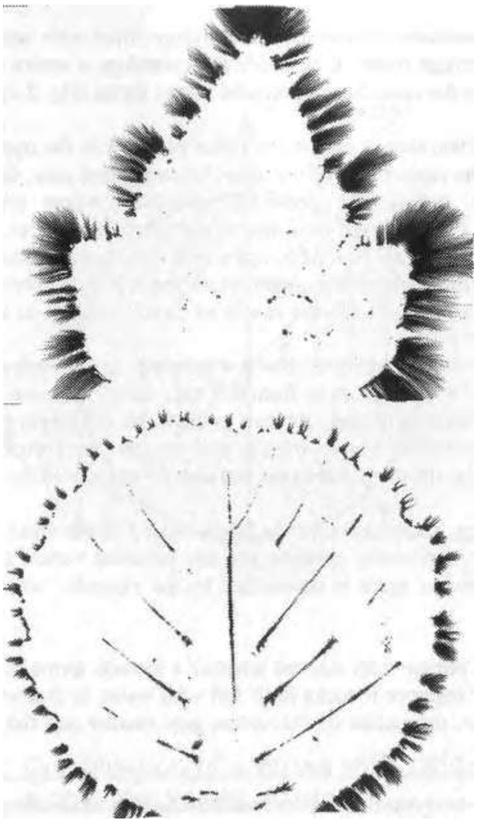

Fig. 3.6: Kirlian photograph of leaves[i] structured corona discharges

---

<i>: (produced by students of electronics in the laboratory for power electronics of the Author, University of Applied Sciences Furtwangen 1991)

<ii>: Kupfmuller, Karl: Einfuhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Auflage 1988, page 453

<iii>: Kupfmuller, Karl: Einfuhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Auflage 1988, page 208

3.6 Vortex and anti-vortex

Fig. 3.5 shows that vortex and dual anti-vortex mutually cause each other. In high tension transmission lines we find a striking example for the combination of current eddy and potential vortex. Within the conductor current eddies are formed. Thus the current density increases towards the surface of the conductor (skin effect). Outside of the conductor, in the air, the alternating fields find a very bad conducting medium. If one follows the text book opinion, then the field outside the conductor should be an irrotational gradient field! But this statement causes unsolvable problems.

When vortices occur inside the conductor, then for reasons of a detachment of the vortices without jumps at the interface to the dielectric, also the fields in the air surrounding the conductor must have the form and the properties of vortices. Nothing would be more obvious as to also mathematically describe and interpret these so-called gradient fields as vortex fields. When looking exact this argument even is mandatory!

The as laws of field refraction known boundary conditions<ii> in addition demand steadiness at the interface of the conductor to the dielectric and don't leave us any other choice. If there is a vortex field on one side, then also the field on the other side is a vortex field, otherwise we offend against the law! Here an obvious failure of the Maxwell theory is present.
Outside the conductor, in the air, where the alternating fields find a very bad conducting medium the potential vortex not only exists theoretical; it even shows itself. Dependent among others on the frequency and the composition of the surface of the conductor, the potenial vortices form around the conductor. When the thereby induced potentials exceed the initial voltage, then impact ionisation takes place and the well-known corona discharge is produced. Everyone of us can hear this as crackling and see the sparkling skin with which high tension transmission lines cover themselves.
In accordance with the text books also a gradient field increases towards the surface of the conductor, but an even shining would be expected and not a crackling. Without potential vortices the observable structure of the corona would remain an unsolved phenomenon of physics (Fig. 3.6).

But even without knowing the structure-shaping property of the potential vortices, that in addition acts supporting and that we'll have to derive, it can be observed well that especially roughness on the surface of the conductor stimulate the formation of vortices and can produce vortices. If one is looking for a reason why with large frequency the very short impulses of discharge always emerge from roughness<iii>, then very probable potential vortices are responsible for it. By means of a Kirlian photograph it can be shown that the corona consists of structured separate discharges (Fig. 3.6).

With this the approach is motivated, formulated and given reasons for. The expositions can't replace a proof, but they should stand a critical examination. Mathematical and physical evidence will be furnished later.

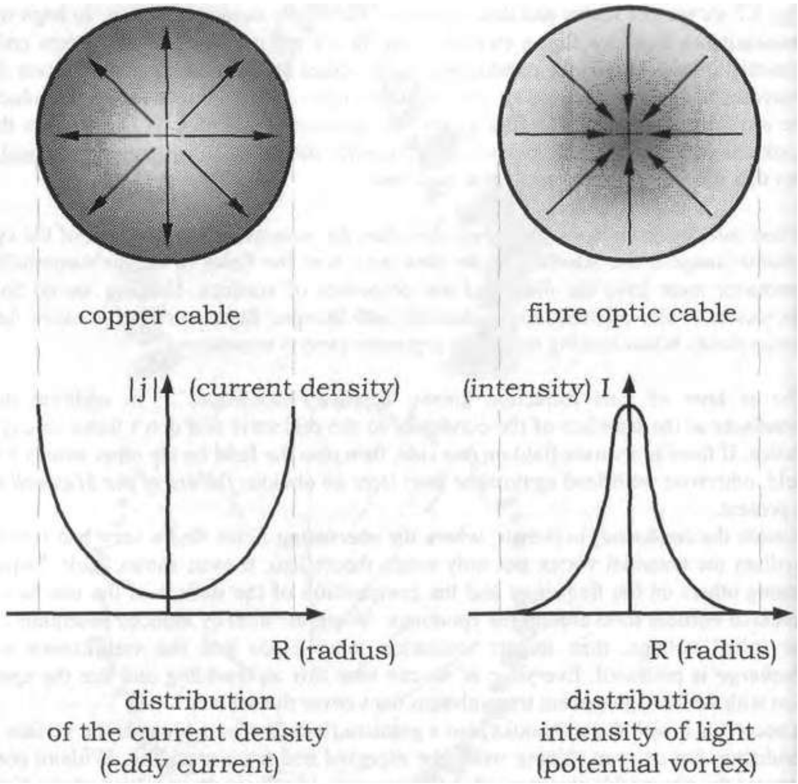
copper cable
fibre optic cable
|j| (current density)
(intensity) I
R (radius)
R (radius)
distribution
of the current density
(eddy current)
distribution
intensity of light
(potential vortex)

Fig. 4.1: The distribution in principle of the intensity of light within a fibre optic cable[i] compared to the distribution of the current density in a copper cable

---

[i] Meyl, Konstantin: Potentialwirbel, Band 1: Diskussionsbeitrage zur naturwissenschaftlichen Deutung und zur physikalisch-technischen Nutzung, basierend auf einer mathematischen Berechnung neu entdeckter hydrotischer Wirbel, INDEL GmbH, Verlagsabteilung, Villingen-Schwenningen 1990, ISBN 3-9802-542-1-6

4. Properties

4.1 Concentration effect

It can be assumed that until now there does not yet exist a technical application for the here presented potential vortex theory unless the phenomenon was used by chance and unconsciously. About this the transmission of optical light signals over fibre optic cable can be given as a typical example.
Compared to a transmission of energy impulses over a copper cable fibre optic cables show a considerable better degree of effectiveness. The derived potential vortex theory provides a conclusive explanation for this phenomenon and therefore is put here to discussion: If we cut through a fibre optic cable and look at the distribution of a light impulse over the cross-section, then we observe a concentration in the centre of the conductor (fig. 4.1).

Herc the duality between the vortices of the magnetic and of the electric field comes to light. Whereas the current eddies in a copper conductor cause the "skin effect" as is well-known, potential vortices show a "concentration effect" and align themselves with the vortex centre. The measurable and in fig. 4.1 shown distribution of the light intensity in a fibre optic cable may confirm this phenomenon, the orientation of the potential vortex on the vortex centre.

For instance the calculation of the resistance of a copper cable provides as an important result an apparent decrease of the resistance directed towards the conductor surface. There the associated better conductivity as a consequence causes an increased current density. In the reversed direction, towards the centre of the conductor, consequently a decrease of the effective conductivity would be present, and this result is independent of the used material. According to the rules of duality this is a condition for the formation of potential vortices. As already said the conductivity is responsible for it, if the expanding eddy current with its skin effect or the contracting potential vortex with its concentration effect is predominant.

Usual fibre optic materials possess not only a small conductivity, but in addition a high dieletricity. This additionally favours the formation of vortices of the electric field. If one consciously or unconsciously supports the potential vortices, then there is a possibility that the life of the fibre optic cable is negatively influenced because of the concentration effect. Of course it can not be excluded that other effects, like e.g. reflections or the modes of the light are involved in the concentration effect. But it should be guaranteed that this actually concerns is causal phenomena and doesn't concern only alternative explanations out of ignorance of the active vortex phenomenon.

The formal mathematical reason for the concentration effect provides the reverse sign in Faraday's law of induction compared to Ampere's law (see also equation 3.1 and equation 3.8 in fig. 3.3).

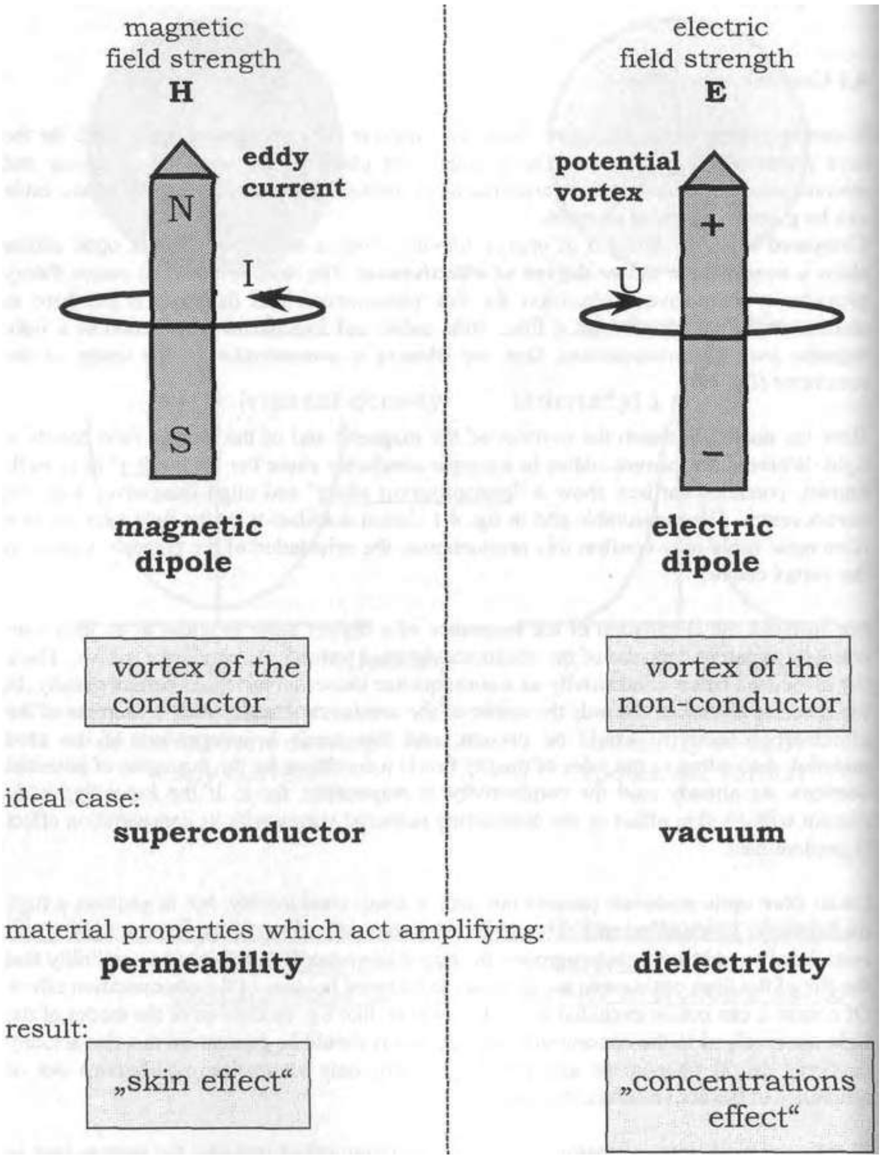
magnetic field strength
H
eddy current
N
I
S
magnetic dipole
vortex of the conductor
ideal case:
superconductor
material properties which act amplifying:
permeability
result:
„skin effect“
electric field strength
E
potential vortex
+
U
–
electric dipole
vortex of the non-conductor
vacuum
dielectricity
„concentrations effect“

Fig. 4.2: The acting as a dipole of current eddies and potential vortices

4.2 Duality of the vortex properties

The rules of duality dictate for the vortex of the electric and of the magnetic field the following characteristics:

1. Whereas currents and eddy currents demand a good conductivity, potentials and potential vortices can only form with bad conductivity, thus in a dielectric and best in the vacuum.

2. Eddy currents run apart, strive towards infinity and thus show the well-known "skin effect" with a spatially limited arrangement of the conductor. According to the rules of duality the potential vortex will strive towards the vortex centre and in this way will show a "concentration effect".

3. Another property of vortices is shown in fig. 4.2.
On the left side a plane eddy current is indicated. Since the discovery of Ampere's law it is well-known to us that such a circular current (I) forms a magnetic dipole standing perpendicular to the vortex plane.
On the right hand side the dual phenomenon is sketched. Here charges are piled up circularly to a planar potential vortex (U). Thereby an electric dipole forms, standing perpendicular to the vortex plane. This relation directly follows from the equations of the field-theoretical approach.
Whereas circular currents and current eddies produce magnetic dipoles, the postulated potential vortices will form electric dipoles.

With these three interesting properties some key questions of quantum physics, that until now have stayed a mystery to science (fig. 4.4), can be answered conclusively and without compulsion e.g.:

I.Why are there no magnetically charged particles?
The better the conductivity of a medium is, the higher as a consequence the number of free charge carriers is. the more strongly eddy currents are formed. The answer to question I is inferred from the opposite case:
In the ideal vacuum no charge carriers at all are present, why no currents, no current eddies and consequently no magnetic poles can exist.
With this well-known fact the first question already is answered. The question why in the microcosm there can not exist magnetically charged elementary particles, why the search for magnetic monopoles doesn't make any sense. Let's ask further:

II. Why are there only electrically charged particles?
Let us for that consider the dual conditions. The worse the conductivity of a medium is, the more the potential vortex -will be favoured that because of this property also can be understood as the vortex of the dielectric.
In the mentioned extreme case of the ideal vacuum, no electric conductivity is present for reason of the missing charge carriers. But this circumstance favours the potential vortex and that, according to fig. 4.2, forms electric poles and with this also the second question would be answered clearly.
It can be traced back to the boundary conditions of the microcosm that without exception electricallv charged particles are entitled to exist; a realization derived from the field-theoretical approach, that covers all experiences.

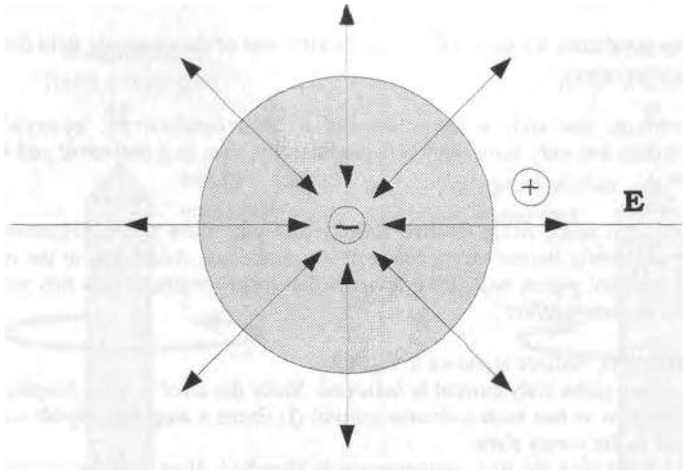

the positron e$^+$

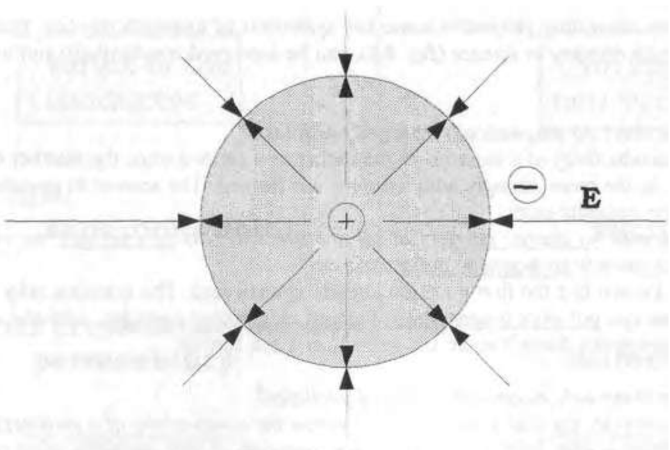

the elektron e$^-$

Fig. 4.3: Elementary particles as configurations of potential vortices

4.3 Derivation of the electron as an elementary vortex

The next key question necessarily has to be brought to a conclusive answer to save the principle of causality, so that we no longer have to postulate an own physics with its own laws for the microcosm:

III. Why do these particles show as monopoles?

More concrete the question has to read:
Where is the positive pole in a negatively charged electron, if it should be an electric dipole?
The only possible answer is:
In the centre of the particle!

Thus in the centre of the electron its positive pole is hidden and in the centre of the positron its negative pole is hidden. But we only observe these particles from the outside and for reason of the field conditions of the electron we measure a negative charge and for its antiparticle, the positron, a positive charge. If in each case we wanted to measure the electric fields included in the inside, we had to destroy the particle. Then a proof would not be possible anymore.

Here also a persistent mistake is eliminated by the for a long time known axiom that monopoles can not exist at all if one considers continuity! By means of technical-physical experiments this axiom is sufficiently secured.
The quantum physical approach is standing on uncertain ground if it is postulated that other laws of nature should apply to particle physics, if a second approach, the field-theoretical approach, is conceivable that does not know these problems!

The discussed concentration effect gives the potential vortex a structure shaping property. With that also the fourth key question can be answered:

IV. Why do the particles have the form of spheres?

The potential vortex is favoured in the particle-free vacuum of the microcosm because of the missing conductivity. In connection with the concentration effect the following conclusion can be drawn:
The extremely mighty potential vortex exerts a high pressure on the microcosm and on each particle.

With that also the fourth key question, why stable elementary particles are spherical, can be answered by the potential vortex theory:
Only the sphere is able to withstand a high outside pressure.
All other forms, like e.g. dipoles formed like a rod or a club would be instable in the presence of the extremely concentrating potential vortex. They would be immediately destroyed by the pressure of the potential vortex.

I. Why are there no magnetically charged particles?
(the vacuum has no conductivity!)

II. Why are there only electrically charged particles?
(in the vacuum only potential vortices can exist!)

III. Why do these particles show as monopoles?

(the other pole is locked up in the inside of the vortex oscillation!)

IV. Why do the particles have the form of spheres?

(for reason of the outside pressure by the concentration effect!)

V. Why is the elementary quantum stable?

(without conductivity no decay of vortices takes place!)

VI. Why does for every particle of matter exist an antiparticle?

(there are two swirl directions with equal rights!)

VII. Why are particles and antiparticles incompatible?
(contrary swirl directions!)

Fig. 4.4: Key questions of quantum physics

4.4 Quanta as field vortices

The fied-theoretical approach demands removing the electron from the field equations (eq. 3.7) and at the same time introducing the potential vortex of the electric field. With this vortex phenomenon there now is a possibility that the electromagnetic wave spontaneously rolls up to a vortex in case it is disturbed from the outside. The vortex particle that is formed in such a way owes its physical reality on the one hand the concentration effect of the potential vortex, that compresses this particle to the dimension of a tiny sphere and on the other hand its localization for reason of the oscillation around a fixed point.
The sphcrical elementary particles are being compressed to inconceivably small dimensions. Therefore they are capable to bind a comparatively high energy in their inside. This is confirmed by the mass-energy relation $E = mc^2$. (4.1)

The fact that the energy is dependent on the speed of light can be judged to be a clear indication that quanta actually are nothing but oscillating electromagnetic packages, vortical oscillations of empty space!
The next question reads:

V. Why is the elementary quantum stable?
The worse the conductivity is, the more the potential vortex will be favoured, the more strongly the concentration effect will form, the smaller the spherical phenomena will get - the larger the authoritative relaxation time will be, i.e. the slower the decay of vortices and with that the more stable the vortex phenomenon will be.
In the microcosm, that comes the ideal case of a particle-free vacuum very close, the spherical vortices because of the missing conductivity have an absolute stability.

VI. Why does for every particle of matter exist an antiparticle?
Since every vortex can also oscillate
in the opposite direction, there always exist two forms of formation of spherical vortices with equal rights, one of them is assigned to the world of matter and the other to the world of anti-matter.

VII. Why are particles and antiparticles incompatible?
For reason of the contrary swirl direction they are incompatible to each other. They have the tendency to destroy each other mutually, like two trains that want in the opposite direction on a single-tracked distance.

The quantum physical approach does not have an answer to these key questions. Until now scientists have merely thought about to what the observable contraction in the microcosm and the macrocosm can be traced back. Because the approach was not able to furnish an answer, without further ado some new matter was introduced: the sluons. These binding particles should be able to exert the necessary pressure. But until now no one has been able to observe or detect this fabulous matter. Nobody knows its structure and its composition. Despite missing evidence it is stated that this matter is mass less and at the same time lumped; it is invisible because it can't interact with any other matter, not even with the supposed building parts of the atomic nuclei, the quarks. But at the same time there should be exerted a pressure on the quarks, for which reason quarks again should be able to interact with gluons!

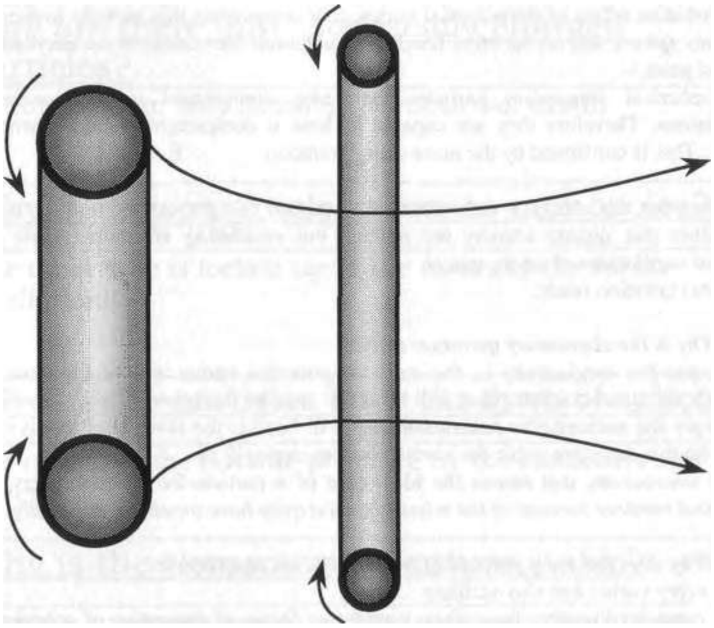

Fig. 4.5: Two coaxial oscillating vortex rings (Lugt<i>).

<i> Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; page 42, ISBN 0-89464-916-7

4.5 The photon

The ability to form structures as a consequence of the concentration effect gives the potential vortex a number of highly interesting properties. To derive these properties we can make work easier when we fall back upon the observations and experiences of flow dynamics.

Here the vortex ring takes a special place. Its vortex centre is not closed, for which reason it is not stationary and propagates in space with a constant speed. It can be observed that the velocity of propagation increases with the ring diameter becoming smaller. By means of the vortex rings, that skilful smokers can produce with pointed lips, these properties can be made visible.

Now if two vortex rings run into each other with the same axis and direction of rotation then both oscillate around each other, by one vortex attracting the other vortex, thereby accelerating and thus contracting it. The second vortex then slips through the core opening and gets again slower and larger. Now the first vortex accelerates and plays the same game (fig. 4.5).

It would be obvious for the vortex of the electric field to have a corresponding property. The electron e- and with the opposite swirl direction the positron e+ will form such a potential vortex corresponding to the derivation. Two electrons, as like charged particles, would repel each other and surely will be out of the question for such a configuration. An electron and a positron however will attract each other and because of their incompatibility they will mutually destroy unless they open their vortex centres to form a vortex ring. Now the e- shows its positively charged centre that shows the same swirl direction as the e+ seen from the outside. Therefore the vortices don't hurt each other, when the positron slips through the opened vortex centre of the electron and vice versa.

This oscillating electron-positron pair has strange properties: seen from the outside one moment it is negatively charged and the next moment it is positively charged. Therefore over time on the average no charge will be measurable and no electromagnetic interaction will take place.

One moment the particle is matter and the next moment it is anti-matter. Hence no mass at all can be attributed to the particle. Interactions primarily takes place between both dual vortices. We can predict, the particle has neither mass nor charge. The environment merely sees a fast oscillating particle that only within every half cycle is capable of an interaction.

The centre of the oscillating particle is open, for which reason it is not stationary anymore. Instead it propagates in z-direction with the swirl velocity, which is the speed of light, in this way preventing a rotation around the x- or y- axis (fig. 4.6). In this way a polarizability is present.

The only possible and, as we will see, necessarily taking place rotation around the z-axis gives the particle a spin of the magnitude of a quantum of angular momentum. After all the rotation for e^- and e^+ is of the same magnitude with a spin of each time $\frac{1}{2} \cdot \hbar$. There should be paid attention to the fact that for the case of an opposite sense of direction of the respective rotation around the common z-axis the spin on the average will be zero.

In addition the particle is characterized by an outstanding property: a periodically taking place oscillation with any frequency, but that frequency has to be constant. We now only have to take a table of particles to hand. Actually we will find a corresponding particle that has all these properties: the γ-quanta, also called photon.

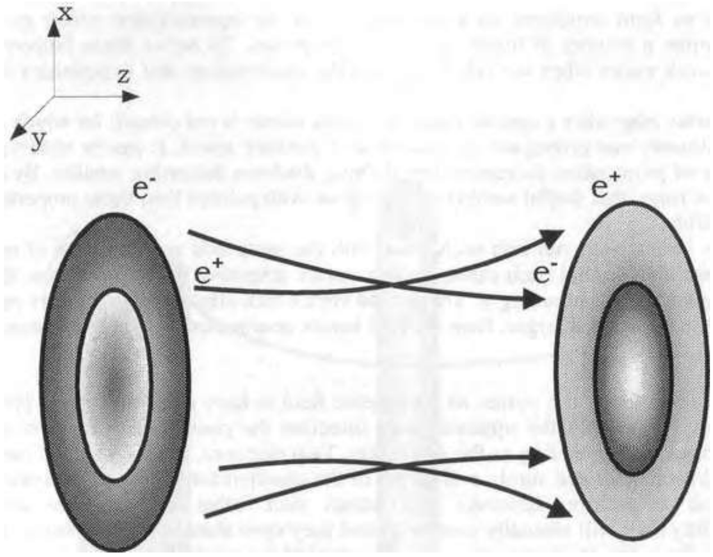

Fig. 4.6: The photon as oscillating electron-positron pair

The decay of the γ- quanta (photon)<i>
(= pair creation = Bethe-Heitler-process 1934):

$$\gamma \longrightarrow e^- + e^+$$

(4.2)

<i> Nachtmann, Otto: Phanomene und Konzepte der Elementarteilchenphysik, Vieweg, Braunschweig 1986, S. 135, ISBN 3-528-08926-1

4.6 Pair creation

Proof for the correctness of the model concept provides the decay of the photon in an electron and a positron in the presence of a strong field, as for instance in an atomic nucleus. This observable decay is called pair creation or Bethe-Heitler process<i>:

$\gamma \longrightarrow e^- + e^+$ (4.2)

In this process the elementary vortices for a short time get back their localization and are therefore detectable. Otherwise the electron and positron have the form of a sphere, the photon however rather has the form of two oscillating discs.

The photon doesn't participate in the electromagnetic interaction, because the electric field lines run from one disc to the other (from + to -). The field lines are not open as they are for e^- or e^+ (fig. 4.3). To open up the field lines an energy is necessary that corresponds to the sum of the two formed particles. But from this it by no means follows that this amount of energy will be released in the reversed and much better known process, where matter and anti-matter annihilate under emission of γ-quanta. At the end of the derivation the vortex model will provide us the desired answers on questions of the energy of photons.

Here first of all only the properties will be concerned. Experiments, in which light shows as a particle, are the photoelectric effect, the Compton effect and a lot more. According to the by Maxwell developed classical theory of light however is light an electromagnetic wave that is not quantized in any way, neither as sphere nor as disc, the wave nature of light as well has a physical reality and is secured by experiment. This is witnessed by the interference patterns of overlapping coherent light beams.

A concept in which light could exist at the same place and the same time both as wave and as corpuscle could never be brought into accord with the principle of causality. Formulas of compromise, like the uncertainty principle of Heisenberg that refers to the point of view of the observer, can't change anything about this dilemma. The dual nature of light, that in this context is gladly spoken of, rather consists of the fact that dependent on the local field conditions, any time and spontaneously the wave can roll up to a vortex.

As an example of a violation of the principle of causality it has been indicated under point 3 (fig. 3.1) that both fields and quanta at the same time should be the cause of something. This concept was formulated by Maxwell and written down in modern quantum electrodynamics by Dirac but in the field-theoretical approach we have dropped this concept because it violates all rules of causality in an elementary manner. Therefore it only is consistent, if we hold the view that the light is either wave or particle but never is both at the same time!

In the spontaneous transition of the wave to the particle all the important properties are conserved: the propagation with the speed of light, the characteristic frequency of the oscillation and the mentioned polarizability.

The process of rolling up possibly takes place already in the laboratory, in a bubble chamber and at the latest in our eyes. To receive the electromagnetic wave, we had to have antennas. We actually see the photons. It therefore would be obvious if our cells to see only could perceive vortices, in this case photons. We don't possess a sense organ for fields and waves.

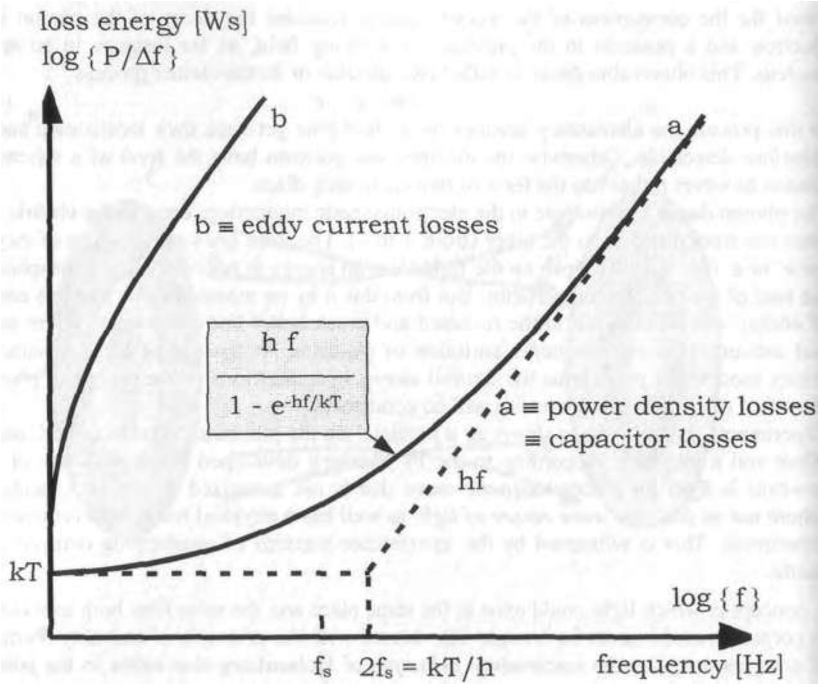

Fig. 4.7: The power density shown against frequency for noise
(a) according to Kupfmuller<i>,
as well as for dielectric losses of a capacitor (also a) and for eddy current losses (b) according to Mevl<ii> (b in visible duality to a).

<i> Kupfmuller, Karl: Einfuhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Auflage 1988, ISBN 3-540-18403-1

<ii> Meyl, Konstantin: Dreidimensionale nichtlineare Berechnung von Wirbelstromkupplungen, Dissertation Universitat Stuttgart 1984

4.7 Noise

If, according to the field-theoretical approach, there exist electric field vortices then they will not only form the elementary particles in the vacuum, but will also macroscopically form and have an effect in larger dimensions.

Assuming a wave that rolls up to a plane vortex it would be obvious if polarization and velocity of propagation are conserved in the process. But how does it stand about the frequency?
The wave now will walk round a stationary point, the vortex centre. The propagation with the speed of light c will remain existent as the swirl velocity. For a plane circular vortex, where the path for a revolution on the outside is very much longer than near the vortex centre, there arises a longer wave length and as a consequence a lower frequency on the outside as more to the inside. With this property the vortex proves to be a changer of frequency: the vortex transforms the frequency of the causing wave in an evenly spectrum, that starts at low frequencies and stretches to very high frequencies (fig. 1.4). Exactly this property we observe in "white noise". The consistent conclusion would be that this concerns the vortex of the electric field. Anyone can, without big expenses, convince himself or herself of the localization, of the property to change the frequency and of the circumstance that vortices can be very easily influenced, that they avoid or again whirl about a place of disturbance, for instance an antenna. For that one only needs to tune a radio receiver to a weak and noisy station and move oneself or some objects around, then one is able to directly study the influences from the manipulation of the receiving signal.
But already the fact that the using and measurability of signals is limited by noise makes clear. which attention the potential vortex should be given.

Within a limited frequency range the power of the Nyquist or resistance noise is independent of frequency (fig. 4.7). This should be clarified particularly by the term "white noise" analogous to white light, where all visible spectral ranges independent of frequency have the same energy density.

But this relation doesn't hold for high frequencies of any magnitude. Here another noise-effect appears, that is said to have its cause in the quantum structure of energy<i>. Untouched by possible interpretations an increasing power of the noise is measured, that more and more turns into a being proportional to frequency (fig. 4.7, curve a). Interestingly this curve shows a remarkable duality to the power curve of eddy currents, likewise shown against the frequency, that for instance can be measured at eddy current couplings<ii> (fig. 4.7, curve b). This circumstance suggests a dual relation of the potential vortex of the electric field in bad conducting media on the one hand and the eddy current in inductive materials on the other hand <iii>.

<iii>: Meyl, Konstantin: Wirbel des elektrischen Feldes,
EMC Journal 1/95, 6. J, ISSN 0945-3857, S. 56 - 59.

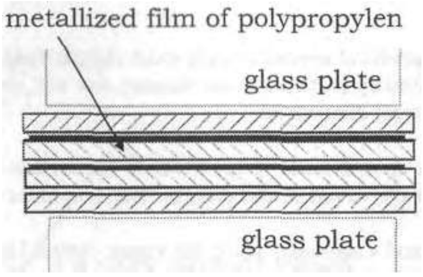
metallized film of polypropylen
glass plate
glass plate

a) measurement set up according to Yializis and others[i]

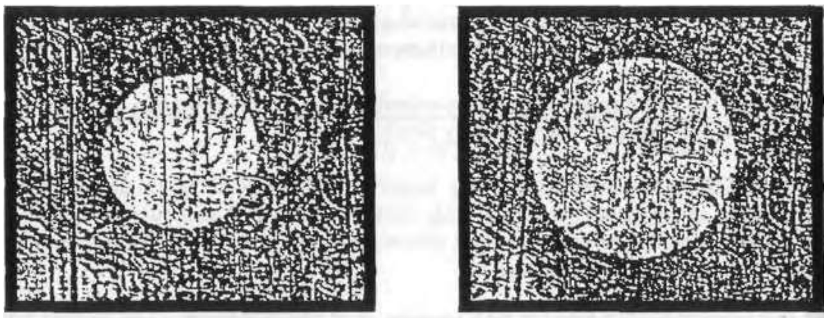
b) after 40 hours c) after 52 hours

Fig. 4.8: Measurement set up (a) and photo of vortex structure in a metallized polypropylen layer capacitor (at 450 V/ 60 Hz/ 100°C)
Observation of the formation of a vortex (b) and (c). (110 fold magnification), according to Yializis et al.[i]

<i>: A. Yializis, S. W. Cichanowski, D. G. Shaw: Electrode Corrosion in Metallized Polypropylene Capacitors, Proceedings of IEEE, International Symposium on Electrical Insulation, Bosten, Mass., June 1980;

4.8 Capacitor losses

Next the dielectric losses in a capacitor fed with an alternating voltage are measured and likewise put on against the frequency. At first the course is independent of the frequency, but towards higher frequencies it increases and shows the same characteristic course of the curve as the before mentioned power of the noise (fig. 4.7, curve a).

This excellent agreement suggests the assumption that the dielectric losses are nothing but eddy losses.
These vortex phenomena, caused by time-varying fields, are not only found in ferromagnetic and conductive materials, but equally as dual phenomena in dielectrics and non-conductors.

As examples of practical applications the induction welding or the microwave oven can be mentioned. The process can be described in other words as follows: in both examples the cause is posed by high-frequency alternating fields that are irradiated into a dielectric as an electromagnetic wave, there roll up to potential vortices and eventually decay in the vortex centre. The desired and used thermal effect arises during this diffusion process.

The striving in the direction of the vortex centre gives the potential vortex of the electric field a structure shaping property. As a consequence of this "concentration effect" circular vortex structures are to be expected, comparable to the visible vortices in flow dynamics (e.g. tornados and whirlwinds). At the same time the dual anti-vortex arises, the diverging eddy current. It takes, as is well-known, the given structure of the conductor, so in the technical literature one correspondingly talks of a "skin effect". Now if conductor and non-conductor meet as they do in a capacitor, then at the boundary area visible structures will form. Circles would be expected, if the eddy current in the inside and striving to the outside is equally powerful as the potential vortex that comes from the outside and concentrates.

Actually such circular structures are observed on the aluminium of high tension capacitors, when they were in operation for a longer period of time. The formation of these circles, the cause of which until now is considered to be unsolved, is already experimentally investigated and discussed on an international level by scientists (fig. 4.8) [i, ii].

These circular vortex structures can be seen as a visible proof [iii] for the existence of potential vortices of the electric field.

[ii]: D. F. Taylor, On the Mechanism of Aluminium Corrosion in Metallized Film Capacitors, IEEE Transactions on EI-19,August 1984,No.4,pp.288-293.

[iii]: Meyl, Konstantin: Wirbel des elektrischen Feldes, EMC Journal 1/95, 6.J, ISSN 0945-3857, S. 56-59.

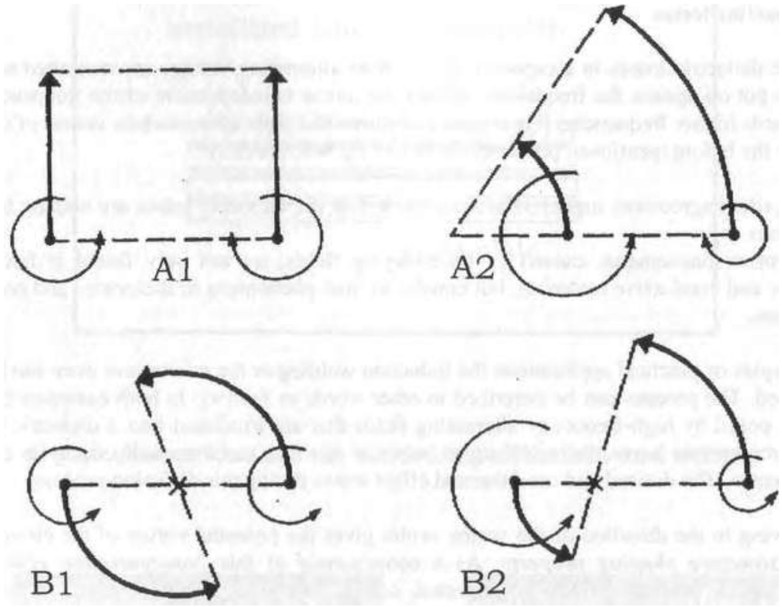

Fig, 4.9: Motion of two point vortices. (Lugt<i>)
A. with opposite direction of rotation
B. with the same direction of rotation
1. for equal vortex strength
2. for unequal vortex strength

Example from practice for case Al:
vortex pair behind an airplane

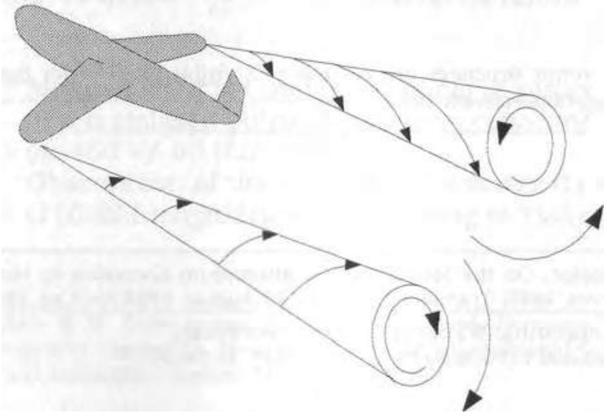

4.9 Vortex lines and vortex streets

It can be assumed that the vortex of the electric field is relevant with regard to the electromagnetic environmental compatibility. This then holds not only for microcosmic and microscopic vortices, but also for macroscopic and larger dimensions. The individual vortices can join together to balls and lines. For the study of this process it is useful to again fall back upon experiences of flow dynamics.

The co-operation of individual point vortices has been investigated thoroughly in flow dyamics. Without any outside manipulation an individual vortex rotates on the spot. That changes in the case of two neighbouring vortices. Now it depends on their mutual strength and sense of rotation. If they have the opposite sense of rotation and equal strength then their centres of rotation move straight forward in the same direction. If however the direction of rotation is the same then both vortices rotate around each other (fig. 4.9).

In this way a multitude of point vortices is capable, to form in the first case whole vortex streets and in the second case spherical vortex balls. In principle a vortex string can also consist of a multitude of potential vortices pointing in the same direction; but it has the tendency to roll up to a vortex ball in case it is disturbed from the outside, as can be shown very clear by means of computer simulations<i>.

As a starting-point for a discussion the thesis can be put forward that also electric field vortices, in nature usually consisting of a multitude of individual point vortices, appear as vortex strings and vortex balls.

Perhaps historians see in this property an answer to the question, how it was possible for the Romans to build streets straight as a die in the wilderness. Their land surveyors, the Augures, had at their disposal neither laser, nor any other suitable gauges. Their most important tool was the Lituus, the crook, that at its upper end was rolled up like a vortex in the sense of a flat coil shaped like a spiral.

The question poses what this strange object was used for. Perhaps the roman land surveors tracked down any vortex lines with this crook and then used them to orientate themselves?

History still holds a lot of secrets, but for now only this indication is given. The following seminar will give enough opportunities for speculations and discussions<ii>.

<i>: Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; p.38, fig. 3.25, 3.26 and 3.27, ISBN 0-89464-916-7

<ii>: K. Meyl: Elektromagnetische Umweltvertraglichkeit Teil 2 und 3 Seminar-umdrucke, INDEL Verlag VS; see part 2 and 3 of this book.

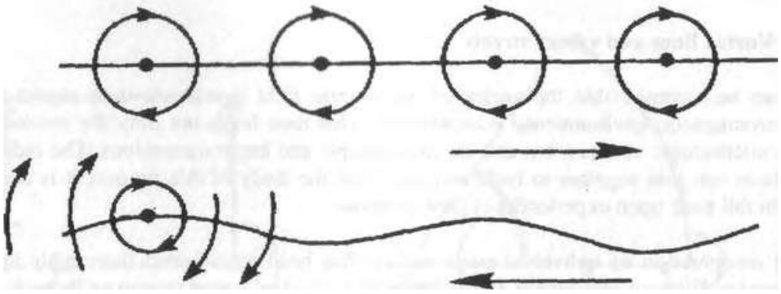

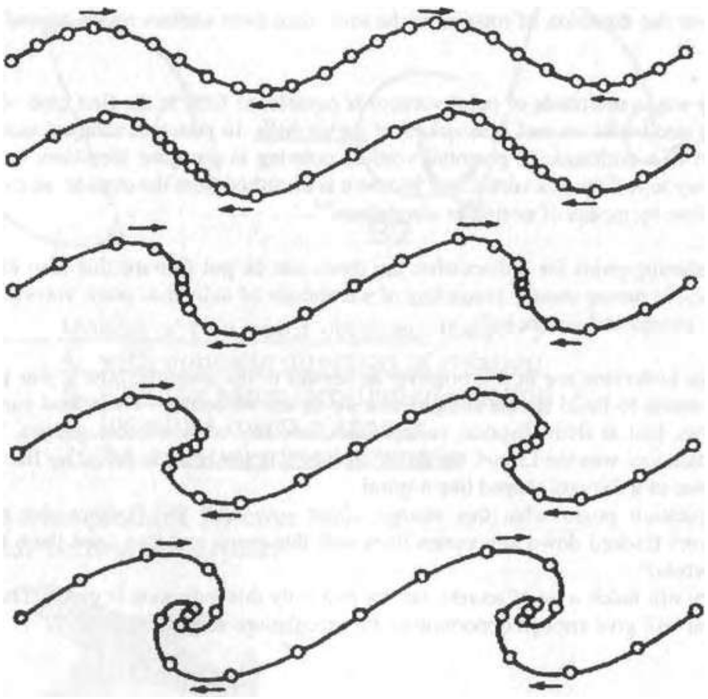

Fig. 4.10: The rolling up of a vortex chain to a vortex ball for the smallest disturbance
(case Bl in fig. 4.9) (Rosenhead<i>).

<i>: L. Rosenhead: Formation of vortices from a surface of discontinuity. Proc. Roy. Soc. A 134, 1931, 170. taken from:

Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; page 39, figure 3.29, ISBN 0-89464-916-7

4.10 Water colloids as vortex balls

We have to realize that in the biosphere we are staying in a relatively ideal dielectric. The two "capacitor plates" are being formed by the ionosphere and the earth. The potential vortex will, as said, be favoured by a bad conductivity and by a high dielectricity. Consequently it will dominate and take effect in the biosphere. In which way it takes effect is the central theme of the electromagnetic environmental compatibility.

Life in this world consists predominantly of water and water has a very high dielectricity! With that the effectiveness and the long life of the potential vortex increases. The human head for instance contains 7O% and plants contain over 90% water! But it does not simply concern H_2O, but structured water in a colloidal form. These water colloids could be vortex balls because they consist of a large number of water molecules in a spherical arrangement They form independent and insoluble particles with a negative electric charge (fig. 4.11). Water is not equal water thanks to this structure. One can buy healing water and corresponding sources are well-known and famous. Many an effect can be explained by means of a chemical analysis but not everything.

The highest age in this world is reached by the inhabitants of Hunza, in the mountains of norhern India at the foothill of the Hindu Kush, at an altitude of 2500 meters. They drink some muddy glacial water that is strongly colloidal. Hence it would be obvious that plants and also we ourselves need such water for our physique. Processes are known with which the advantageous vortex balls, say colloids, are produced artificially by mechanic or chemical treatment<i>. Levitated water, as it is called and as it is for sale nowadays, is said to be more healthy. Unfortunately people predominantly work empiric in this area, because science occupies itself with this topic only little or not at all.
Another problem is the fact that the colloids again fall apart quickly. The like negative charge favours this process. The liquid crystals have to be stabilized from the outside. In the case of the Hunza-water the colloids are surrounded by a vegetable layer of fatty acid and are protected in this way <ii>. It possibly is very obliging to nature, if the water colloids also in biological systems are stabilized in that way.
Everyone of us knows that fresh spring water tastes much better than stale, bottled water, even if the chemical analysis turns out of be absolutely identical. For this fact classical science is not able to give a cause - a further problem of causality. In any case should potential vortices with their structure shaping property be considered as a cause for the formation of water colloids. It surely causes no difficulties at all to interpret the colloids as vortcx bulls.

<i>: V. Schauberger: Die Entdeckung der Levitationskraft, Implosion 1995 Nr. 112 und:
N. Harthun: NaturgemaSe Technik - Wege fur die Forschung nach Viktor Schauberger, Verlag Delta Design, Berlin 1996.

<ii>: Flanagan: Elexier der Jugendlichkeit, Waldthausen Verlag Ritterhude 1992, orig.: Elixir of Ageless

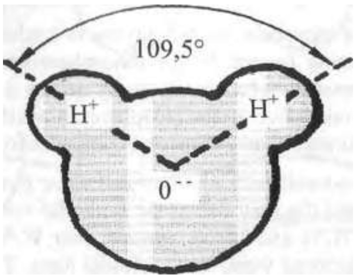

H_2O: the angle of the bond between 2 H molecules = 104,5°
angle of the bond in excited state = 109,5°

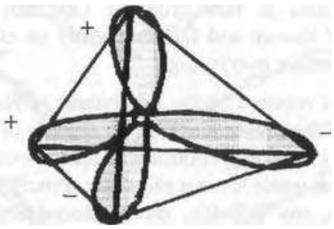

electron orbital diagram of energy (excited H_2O) = tetraeder
most stable liquid crystal = 8 tetraeder = star of octaeder

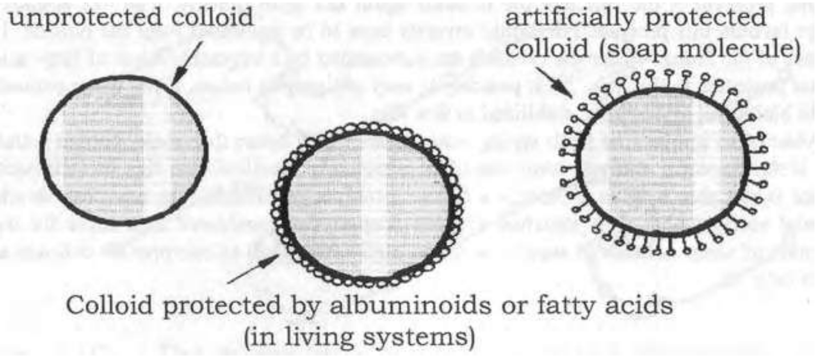
unprotected colloid
artificially protected
colloid (soap molecule)
Colloid protected by albuminoids or fatty acids
(in living systems)

Fig. 4.11: Water molecules and water colloids

4.11 Phenomenon of transport

The vortex principle is self-similar. This means that the properties of an individual vortex also for the collection of numerous vortices again appear and can he observed in a similar manner. That's why a vortex ball behaves entirely similar as an individual isolated vortex. The same concentration effect, that keeps the vortex together, shows its effect for the vortex ball and keeps it together also.

Something corresponding holds for a basic property of potential vortices, being of a completely different nature. It is the property to bind matter in the vortex and carry it away with the vortex. Well-known are the vortex rings that skilful cigarette smokers can blow in the air. Of course also non-smokers can produce these air eddies with their mouth but these remain invisible. Solely by the property of the vortex ring to bind the smoke it becomes visible to the human eye.

If out potential vortex transports something then it rather should be a dielectric material, so preferably water. Therefore if in the environmental air we are surrounded by potential vortices that we can detect for instance as noise, then they are capable with their "phenomenon of transport", to pick up water and to keep it in the vortex. In this way the atmospheric humidity is explicable as the ability of the air particles to bind comparatively heavy water molecules. If the vortex falls apart then it inevitably releases the water particles and it rains. This is merely a charming alternative for the classical representation without claim to completeness.

Already the Romans have made use of this phenomenon to find water and sources. About this Vitruv<i> (from 23 BC) in his 8th book about architecture writes: "Before sunrise one has to lie down on the earth at the places, where to search for water,... and one has to look at the area... Then one has to dig at the place where there appears curling and in the air rising moist steam. Because this characteristic can not occur at a place where there is no water". The at a certain time of day and in certain seasons occasional in meadows and corn fields observable streaks or circular mostly moist places with differing vegetation, have to be judged as an infallible sign for the existence of this phenomenon.

This phenomenon of transport again appears for the discussed water colloids. The involved water molecules form a spherical object with a negative charge. They turn their negatively charged side to the outside and point with the positively charged end in the direction of the middle of the sphere. There, no longer discernible from the outside, a negatively charged ion can be, that is stuck, that no longer can escape and that gives the whole colloid a characteristic property. In this way nature knows various water colloids that constitute plants and animals. But starting at a temperature of 41°C the liquid crystals fall apart. This not by chance is the temperature at which a person dies. Already 10 millivolts per liquid crystal suffice for the electrically induced death.

The to a colloid identical structure we find in the structure of the atoms. Here the atomic nucleus is held in the inside of a vortex-like cloud of electrons, the atomic hull. We'll hit the phenomenon of transport a last time, when we derive the elementary particles. For the photon is already discernible the tendency of an elementary vortex, to take another vortex in its inside. Merely because the electron and positron are evenly matched a stable configuration is prevented for the photon.

<i>: Vitruvius Pollio, Marcus: Ten Books about architecture, WBG 1987

In chapter vortex calculation used differential operations:

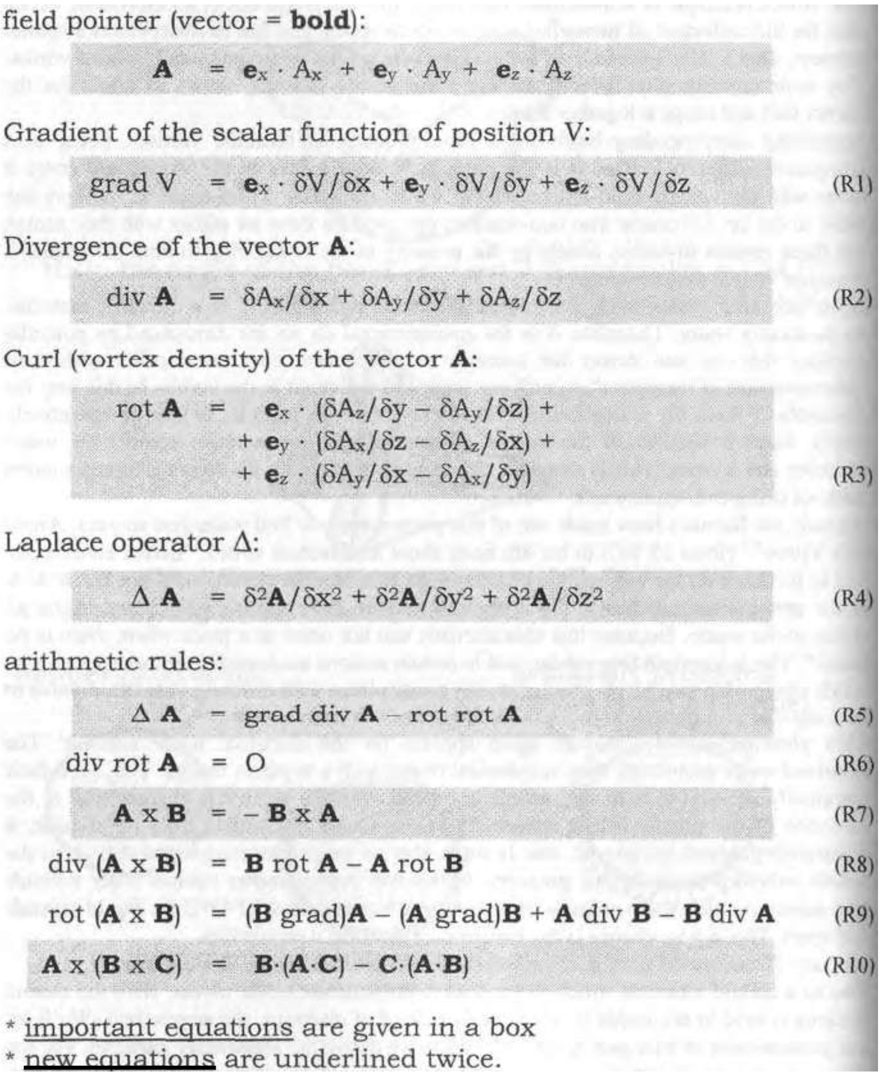

field pointer (vector = **bold**):

$$\mathbf{A} = \mathbf{e}_x \cdot A_x + \mathbf{e}_y \cdot A_y + \mathbf{e}_z \cdot A_z$$

Gradient of the scalar function of position V:

$$\text{grad } V = \mathbf{e}_x \cdot \delta V/\delta x + \mathbf{e}_y \cdot \delta V/\delta y + \mathbf{e}_z \cdot \delta V/\delta z \qquad \text{(R1)}$$

Divergence of the vector **A**:

$$\text{div } \mathbf{A} = \delta A_x/\delta x + \delta A_y/\delta y + \delta A_z/\delta z \qquad \text{(R2)}$$

Curl (vortex density) of the vector **A**:

$$\begin{aligned} \text{rot } \mathbf{A} = \; & \mathbf{e}_x \cdot (\delta A_z/\delta y - \delta A_y/\delta z) + \\ & + \mathbf{e}_y \cdot (\delta A_x/\delta z - \delta A_z/\delta x) + \\ & + \mathbf{e}_z \cdot (\delta A_y/\delta x - \delta A_x/\delta y) \end{aligned} \qquad \text{(R3)}$$

Laplace operator Δ:

$$\Delta \mathbf{A} = \delta^2\mathbf{A}/\delta x^2 + \delta^2\mathbf{A}/\delta y^2 + \delta^2\mathbf{A}/\delta z^2 \qquad \text{(R4)}$$

arithmetic rules:

$$\Delta \mathbf{A} = \text{grad div } \mathbf{A} - \text{rot rot } \mathbf{A} \qquad \text{(R5)}$$

$$\text{div rot } \mathbf{A} = 0 \qquad \text{(R6)}$$

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A} \qquad \text{(R7)}$$

$$\text{div } (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \text{ rot } \mathbf{A} - \mathbf{A} \text{ rot } \mathbf{B} \qquad \text{(R8)}$$

$$\text{rot } (\mathbf{A} \times \mathbf{B}) = (\mathbf{B} \text{ grad})\mathbf{A} - (\mathbf{A} \text{ grad})\mathbf{B} + \mathbf{A} \text{ div } \mathbf{B} - \mathbf{B} \text{ div } \mathbf{A} \qquad \text{(R9)}$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B} \cdot (\mathbf{A} \cdot \mathbf{C}) - \mathbf{C} \cdot (\mathbf{A} \cdot \mathbf{B}) \qquad \text{(R10)}$$

* important equations are given in a box
* new equations are underlined twice.

Fig. 5.0: Collection of formulas for vector analysis

5. Derivation and interpretation

Vortices cause big problems to every measuring technician. They have the unpleasant property to whirl around the sensor even if it is as small as possible. Vortices avoid the smallest disturbance and then can hardly be detected reproducibly.

From the well-known eddy current we know of this problematic. Instead of the vortex, we are forced to measure and analyse any effects that arise from the vortex. These can be measurements of the eddy losses or effects back on the stimulating field. But only provided that the effect actually occurs.
The prerequisite for an increase in temperature by eddy losses is that the vortex falls apart. In an ideal medium it unfortunately will not do us this pleasure.

As vrtex of the dielectric the potential vortex will find fairly ideal conditions in air or in water. How should a vortex be detected, if it does not produce any effect? The classical measuring technique is here at its wits' end.

From the duality to the well-known eddy current and by means of observation in the previous chapters numerous properties of the potential vortex have been derived. But these are not all the properties. The mathematical calculation of the electric vortex field, that we want to turn to now, will reveal still further meaningful and highly interesting properties.

The observation is important, but it can't replace an exact calculation. A strictly mathematical derived result has occasionally more expressiveness than a whole book full of explanations. It will be a big help to derive and to discuss the field equation that all considerations are based on.

We facilitate the mathematical work by vector analysis. Therefore it is useful that we choose the differential form (equation 5.1 and 5.4) instead of the integral form (equations 3.1 and 3.2 resp. 3.8).

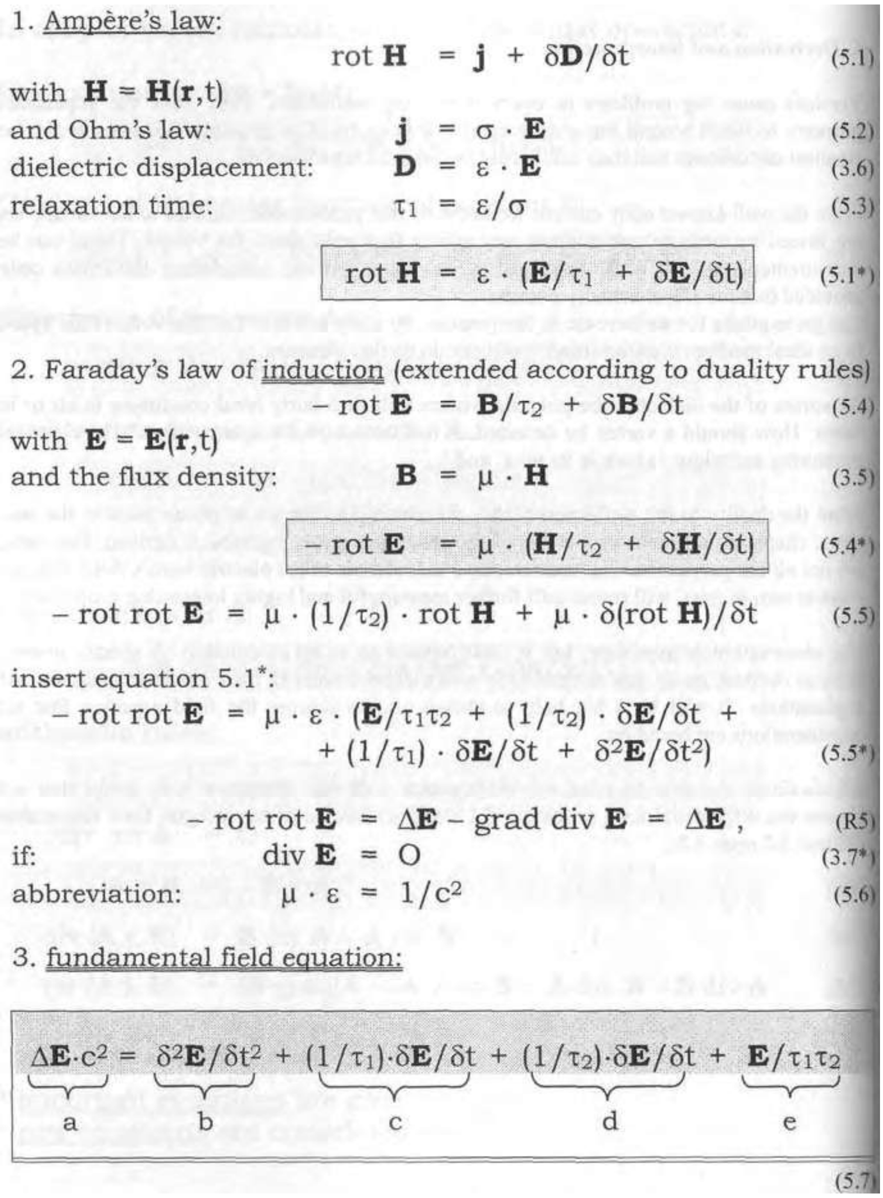

1. Ampère's law:

$$\text{rot}\,\mathbf{H} = \mathbf{j} + \delta\mathbf{D}/\delta t \qquad (5.1)$$

with $\mathbf{H} = \mathbf{H}(\mathbf{r},t)$

and Ohm's law: $\mathbf{j} = \sigma \cdot \mathbf{E}$ (5.2)

dielectric displacement: $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (3.6)

relaxation time: $\tau_1 = \varepsilon/\sigma$ (5.3)

$$\boxed{\text{rot}\,\mathbf{H} = \varepsilon \cdot (\mathbf{E}/\tau_1 + \delta\mathbf{E}/\delta t)} \qquad (5.1^*)$$

2. Faraday's law of induction (extended according to duality rules)

$$-\,\text{rot}\,\mathbf{E} = \mathbf{B}/\tau_2 + \delta\mathbf{B}/\delta t \qquad (5.4)$$

with $\mathbf{E} = \mathbf{E}(\mathbf{r},t)$

and the flux density: $\mathbf{B} = \mu \cdot \mathbf{H}$ (3.5)

$$\boxed{-\,\text{rot}\,\mathbf{E} = \mu \cdot (\mathbf{H}/\tau_2 + \delta\mathbf{H}/\delta t)} \qquad (5.4^*)$$

$$-\,\text{rot rot}\,\mathbf{E} = \mu \cdot (1/\tau_2) \cdot \text{rot}\,\mathbf{H} + \mu \cdot \delta(\text{rot}\,\mathbf{H})/\delta t \qquad (5.5)$$

insert equation 5.1*:

$$-\,\text{rot rot}\,\mathbf{E} = \mu \cdot \varepsilon \cdot (\mathbf{E}/\tau_1\tau_2 + (1/\tau_2) \cdot \delta\mathbf{E}/\delta t + (1/\tau_1) \cdot \delta\mathbf{E}/\delta t + \delta^2\mathbf{E}/\delta t^2) \qquad (5.5^*)$$

$$-\,\text{rot rot}\,\mathbf{E} = \Delta\mathbf{E} - \text{grad div}\,\mathbf{E} = \Delta\mathbf{E}\,, \qquad (R5)$$

if: $\text{div}\,\mathbf{E} = 0$ (3.7*)

abbreviation: $\mu \cdot \varepsilon = 1/c^2$ (5.6)

3. fundamental field equation:

$$\underbrace{\Delta\mathbf{E}\cdot c^2}_{a} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{b} + \underbrace{(1/\tau_1)\cdot\delta\mathbf{E}/\delta t}_{c} + \underbrace{(1/\tau_2)\cdot\delta\mathbf{E}/\delta t}_{d} + \underbrace{\mathbf{E}/\tau_1\tau_2}_{e} \qquad (5.7)$$

Fig. 5.1: Derivation of the fundamental field equation.

5.1 Fundamental field equation

We'll start from Ampere's law which provides a value for the current density at any point pace and this value corresponds to the vortex density of the magnetic field strength

$$\operatorname{rot}\mathbf{H} = \mathbf{j} + \delta\mathbf{D}/\delta t \qquad (5.1)$$

We'll use Ohm's law: $\mathbf{j} = \sigma \cdot \mathbf{E}$, (5.2)

the dielectric displacement: $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (3.6)

and the relaxation time: $\tau_1 = \varepsilon/\sigma$, (5.3)

that indicates how fast the current eddies decay. So far we can fall back upon well-known relations.

The result is: $\boxed{\operatorname{rot}\mathbf{H} = \varepsilon \cdot (\mathbf{E}/\tau_1 + \delta\mathbf{E}/\delta t)}$ (5.1*)

The new electric field vortices demand the introduction of a corresponding time constant τ_2 that should describe the decay of the potential vortices, as an extension. The extended Faraday law of induction now provides a potential density, that at any point of space corresponds to the vortex density of the electric field strength:

$$-\operatorname{rot}\mathbf{E} = \mathbf{B}/\tau_2 + \delta\mathbf{B}/\delta t \qquad (5.4)$$

with the flux density $\mathbf{B} = \mu \cdot \mathbf{H}$. (3.5)

The result fulfils the required duality to equation 5.1*: $\boxed{-\operatorname{rot}\mathbf{E} = \mu \cdot (\mathbf{H}/\tau_2 + \delta\mathbf{H}/\delta t)}$ (5.4*)

If we again apply the curl to equation 5.4*

$$-\operatorname{rot}\operatorname{rot}\mathbf{E} = \mu \cdot (1/\tau_2) \cdot \operatorname{rot}\mathbf{H} + \mu \cdot \delta(\operatorname{rot}\mathbf{H})/\delta t \qquad (5.5)$$

and insert equation 5.1*, we obtain:

$$-\operatorname{rot}\operatorname{rot}\mathbf{E} = \mu \cdot \varepsilon \cdot (\mathbf{E}/\tau_1\tau_2 + (1/\tau_2) \cdot \delta\mathbf{E}/\delta t + (1/\tau_1) \cdot \delta\mathbf{E}/\delta t + \delta^2\mathbf{E}/\delta t^2) \qquad (5.5^*)$$

which according to the rules of vector analysis can still be further simplified: = rot rot E = $\Delta\mathbf{E}$ - grad div E , where we should remember that the divergence has to vanish (div E = O. fig. 3.2, equation 3.7), should the corresponding field vortex be inserted!

Furthermore the following well-known abbreviation can be inserted: $\mu \cdot \varepsilon = 1/c^2$ (5.6)

With that the relation with the speed of light c simplifies to the sought-for field equation:

$$\boxed{\underbrace{\Delta\mathbf{E} \cdot c^2}_{a} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{b} + \underbrace{(1/\tau_1) \cdot \delta\mathbf{E}/\delta t}_{c} + \underbrace{(1/\tau_2) \cdot \delta\mathbf{E}/\delta t}_{d} + \underbrace{\mathbf{E}/\tau_1\tau_2}_{e}} \qquad (5.7)$$

This equation describes the spatial (a) and temporal (b, c, d) distribution of a field vector. It describes the electromagnetic wave (a, b) with the influences that act damping. As dumping terms the well-known eddy current (c) and in addition the newly introduced potential vortex (d) appear.

Field vector: ψ = E, H, j, B or D

1. elliptic potential equation:
(stationary: $t \longrightarrow \infty$ resp. $\delta/\delta t = 0$)

$$\boxed{\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{\psi/\tau_1\tau_2}_{e}} \tag{5.8}$$

2. hyperbolic equation:
(undamped wave equation)

$$\boxed{\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{\delta^2\psi/\delta t^2}_{b}} \tag{5.9}$$

3. parabolic equation:
(vortex equation)

$$\boxed{\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{(1/\tau) \cdot \delta\psi/\delta t}_{c/d}} \tag{5.10}$$

decay time of the eddy currents =
relaxation time: $\tau_1 = \varepsilon/\sigma$ (5.3)

decay time of the potential vortices =
relaxation time: $\tau_2 \sim \mu \cdot \sigma$ (5.11)

Fig. 5.2: mathematically divisible individual cases.

5.2 Mathematical interpretation of the fundamental field equation

Every specialist will be surprised to find the Poisson equation (a, e) again as a term in the wave equation. This circumstance forces a completely new interpretation of stationary fields upon us. The new damping term, that is formed by the potential vortices (d), is standing in between.

Let us start with a mathematical analysis. We have applied the curl to equation 5.4*, then inserted equation 5.1* and obtained a determining equation for the electric field strength E. Of course we could as well have applied the curl to equation 5.1* and inserted equation 5.4*. This would have resulted in the determining equation for the magnetic field strength H.

If we insert Ohm's law (5.2) and cancel down the specific conductivity, or we put in the relations of material (3.5) or (3.6) and cancel down by u respectively ε, then the field equation can likewise be written down for the current density j, for the induction B or for the dielectric displacement D.

It just is phenomenal that at all events equation 5.7 doesn't change its form at all. The field vector is thus arbitrarily interchangeable! This circumstance is the foundation for the claim of this field equation to be called fundamental.

It does make sense to introduce a neutral descriptive vector ψ as a substitute for the possible field factors E, H, j, B or D.

The fundamental field equation 5.7 consists of three different types of partial differential equations: a hyperbolic (b), a parabolic (c and d) and an elliptic (e) type. On the left-hand side each time the Laplace operator (a) is found which describes the spatial distribution of the field factor.

The potential equation of the elliptic type (e) is known as Poisson equation. It describes the stationary borderline case of a worn off temporal process ($t \rightarrow \infty$, resp. $\delta/\delta t = 0$).

With this equation potentials and voltages can be calculated exactly like stationary electric currents (5.8).

The hyperbolic equation (b). known as wave equation, shows a second derivative to time. which expresses an invariance with regard to time reversal; or stated otherwise: the direction of the time axis can be reversed by a change of sign of t, without this having an influence on the course of frequency. Wave processes hence are reversible. Equation 5.7 makes clear that a wave without damping by no means can exist in nature. For that both time constants (τ_1 and τ_2) would have to have an infinite value, which is not realizable in practice. Seen purely theoretical, undamped waves could withdraw themselves from our measuring technique (5.9).

Both vortex equations of the parabolic type (c and d) only show a first derivative to time. With that they are no longer invariant with regard to time reversal. The processes of the formation and the decay of vortices, the so-called diffusion, are as a consequence irreversible. Seen this way it is understandable that the process of falling apart of the vortex, where the vortex releases its stored energy as heat e.g. in form of eddy losses, can not take place in reverse. This irreversible process of diffusion in the strict thermodynamic sense increases the entropy of the system (5.10).

Because it poses an useful simplification for mathematical calculations, often the different types of equations are treated isolated from each other. But the physical reality looks different.

field vector: ψ = E, H, j, B or D

1. Borderline case: no conductivity (vacuum)

($\sigma = 0;\quad 1/\tau_1 = \sigma/\varepsilon = 0$):

$$\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{\delta^2\psi/\delta t^2}_{b} + \underbrace{(1/\tau_2)\cdot \delta\psi/\delta t}_{d}$$

(5.12)

(damping by potential vortices)

2. Borderline case: ideal conductivity (superconductor)

($1/\sigma = 0;\quad 1/\tau_2 = 0$):

$$\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{\delta^2\psi/\delta t^2}_{b} + \underbrace{(1/\tau_1)\cdot \delta\psi/\delta t}_{c}$$

(5.12*)

(damping by eddy currents)

3. Diffusion equation

$$\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{(1/\tau)\cdot \delta\psi/\delta t}_{c/d} + \underbrace{\psi/\tau_1\tau_2}_{e}$$

(5.12**)

(vortex)

Fig. 5.3: Two borderline cases of the damping of waves and the diffusion equation for the decay of vortices

5.3 Physical interpretation of the fundamental field equation

In nature the different types of equations always occur in a combined manner.

1. Let's take the concrete case of the particle-free vacuum. Here the specific conductivity is zero. The relaxation time constant $\tau_1 = \varepsilon/\sigma$ responsible for the decay of vortices tends towards infinity according to equation 5.3 and the terms (c) and (e) are cancelled from the field equation 5.7. What remains is the by potential vortices (d) damped wave equation (b) (equation 5.12).

2. The reversed case (with $\tau_2 \longrightarrow \infty$) will consequently occur in materials without resistance, super conducting materials. We now are dealing with the well-known case of the wave damped by eddy currents (equation 5.12*).

Virtually all in nature existing materials however don't fulfil these boundary conditions, from which it follows that both damping terms always occur together and in addition the stationary term (e) becomes active.
It is true that every antenna demonstrates that the electromagnetic wave is convertible in high-frequency alternating currents and voltages, which then are amplified in the receiver. But until this fundamental equation was written down it however was not understood that this transition takes place by means of a vortex. Used are either antennas from well conducting material, or wave guides and horn radiators, which only have a minimal conductivity, because they are filled with air. Actually the wave can be converted in two dual ways; by means of the rolling up to current eddies or to potential vortices (fig. 1.4). Now we finally are capable to explain, why wave guides make possible a better degree of effectiveness: Owing to the concentration effect of the potential vortex the HF-power is bound in the inside and not emitted until the antenna is reached as happens for a wire for reason of the skin effect.
Therefore, physically, one has to imagine this relation, which describes the transition of an electromagnetic wave into a vortex, in the way that the wave spontaneously can roll up to a vortex in case it is disturbed from the outside. The more vortices are generated, the larger consequently is the damping of the wave (equations 5.12 and 5.12*).

3. The life span of the vortices is limited and is determined by the electric conductivity. The at first stored vortices decay with their respective time constant τ. This process is described by the diffusion equation 5.12**. The final stage of the decaying vortices finally is described by the Poisson equation (a, e: equation 5.8).

If the vortex falls apart, it converts the in the vortex stored energy in heat. These processes are known from the eddy current. We speak of heating losses, that the stationary currents cause in the conductor material.
But new is the concept that such vortex phenomena can occur as dielectric losses in capacitors or in the air. The microwave oven or induction welding are good examples of this.

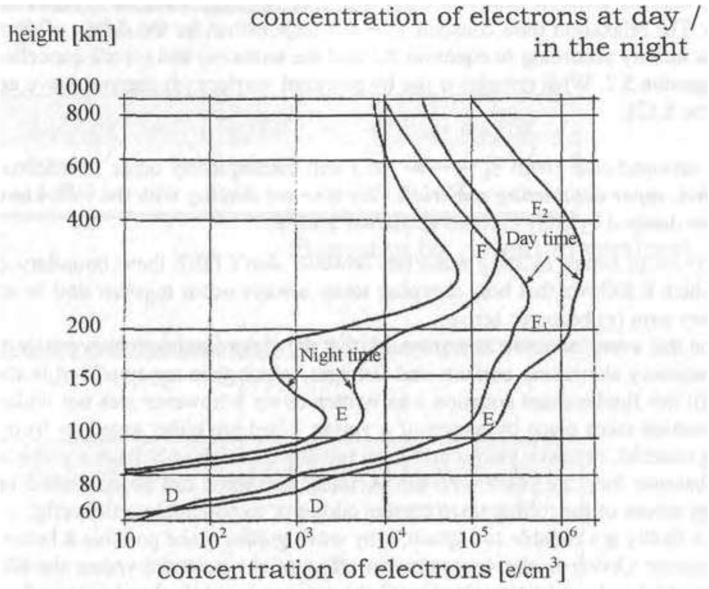

Fig. 5.4: The dependency on height of the ionisation in the ionosphere for medium latitudes .<i>

left curve: for a minimum of sun spots

right curve: for a maximum of sun spots

<i>: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligkeit), 5. Aufl., Bild 6, Seite 11, Verlag Moos & Partner Milnchen, ISBN 3-89164-058-7

5.4 Phenomenological interpretation of the fundamental field equation

How does a damping by vortices take effect in practice? First of all we notice that the reception of broadcastings gets worse. "The information signal is neglectable regarding the noise" explains the radio engineer and means, the number of vortices increases at the expense of the wave intensity.

Why, does the pupil ask, is it so cold in space? There the sun shines day and night and in addition much more intensely than on earth! The correct answer would have to read that because of the extremely small conductivity no diffusion process can take place. We owe the warmth on our earth solely the here taking place decay of vortices. Responsible is the conductivity of the atmosphere.

In 60 km to 500 km height over the earth's surface, the region which is called the ionosphere, the gases predominantly exist in ionized form. There a very good conductivity prevails and eddy current losses are the result. Correspondingly high are the measurable temperatures. Besides the diffusion process the eddy currents carry out a damping of the cosmic radiation. We say the sunlight is filtered and reduced to a for nature bearable intensity.

But not all frequencies are damped in the same way (fig. 2.8). We observe a blue shift, if we look into the actually black sky. The blue sky doesn't show any spots or clouds. The reason is to be sought in the skin effect of the eddy currents, which strive outwards. Since no edge of a conductor is present here, no skin can form. The vortices spread evenly over the ionosphere.

The potential vortex however is able to structure. It merely needs a bad conductivity and this it finds in lower heights between 1 km and 10 km. It damps the wave and with that also the light, for which reason we say it becomes darker, the sun disappears behind clouds.

The clouds well visibly form the discussed vortex balls and vortex strings. Clouds can form virtually from the nowhere during intense solar irradiation, i.e. the waves can roll up to vortices. But as a rule this takes place above the oceans. Here also the phenomenon of transport has an effect. Because of the high dielectricity the water surface favours the formation of potential vortices. So the vortices bind individual water molecules and carry them away. If a diffusion process takes place, in which the vortex decays, then it rains. This can happen in two different ways:

1. Either the conductivity increases. If for instance during intense solar irradiation air ions form, the sun is able to break up clouds and fog. Or when the air is raised in higher layers with better conductivity, because a mountain forces this, then it rains at the mountain edge.
2. For potential vortices the electric field is standing perpendicular to them. If at one point an exceptionally lot of vortices join together, which let the cloud appear particularly dark to black, then the danger exists that the ionization field strength for air is reached, in which case a conductive air channel forms along which the stored up charges discharge. Also lightning is a diffusion process, in which potential vortices decay and rain can form.

In connection with the electromagnetic environmental compatibility great importance is attributed in particular to the storage and the decay of electric vortices. There not only is an academic-scientific interest in the question, how many potential vortices are generated, how many are stored and how many decay, if we make a telephone call with a handy, if we are staying under a high-tension line or if we are eating food, which has been heated up in a microwave oven. The necessary mathematical description is provided by the fundamental field equation 5.7.

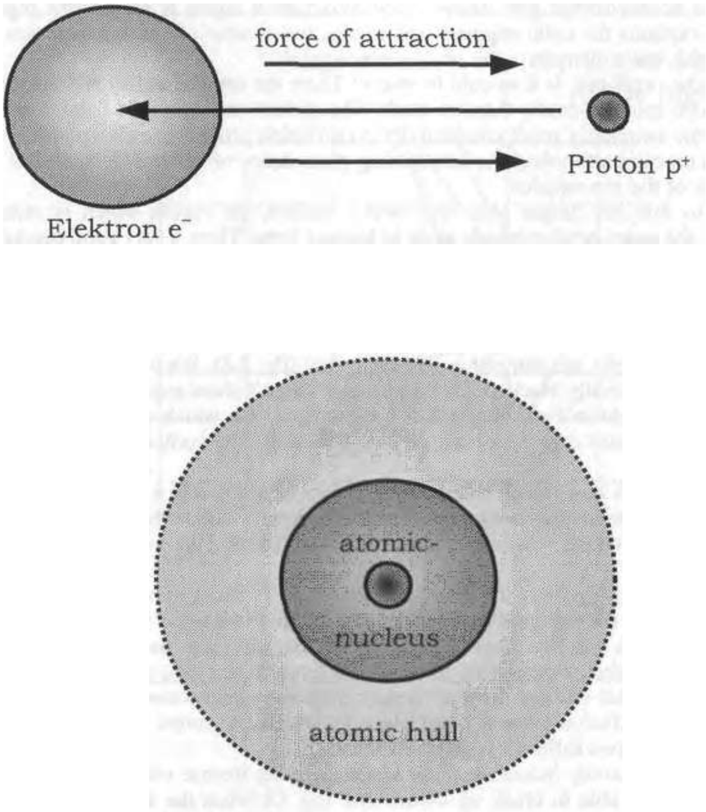
force of attraction
Proton p⁺
Elektron e⁻
atomic-
nucleus
atomic hull

Fig. 5.5: The structure of atoms in the view of the fundamental field equation

condition for equilibrium: $$\boxed{\tau_1 = \tau_2} \qquad (5.13)$$

5.5 Atomistic interpretation of the fundamental field equation

Let's again turn to the smaller, the atomistic dimensions. Here positively charged protons and negatively charged electrons are found. Both are matter particles and that means that seen from the outside both have the identical swirl direction. For reason of the unequal charge conditions they attract each other mutually and according to fig. 4.9 rotate around a common centre of mass as differently heavy pair. Chemists say: "the light electron orbits the heavy atomic nucleus". With small balls they try to explain the atomic structure.
But the model is no good: it contradicts causality in the most elementary manner. We are dealing with the problem that according to the laws of electrodynamics a centripetally accelerated electron should emit electromagnetic waves and continuously lose energy, to eventually plunge into the nucleus.
Our experience teaches that this fortunately is not true - and Niels Bohr in order to save his model of the atom was forced to annul the laws of physics with a postulate founded in arbitrariness.
Actually this state only exists for a very short time and then something unbelievable happens: the electron can't be distinguished as an individual particle anymore. "It is smeared over the electron orbit" do certain people say; "it possesses a dual nature" says Heisenberg. Besides the corpuscular nature the electron should in case of its "second nature" form a matter wave, "the position of the electron is to be looked at as a resonance which is the maximum of a probability density", do explain us de Broglie and Schrodinger.

These explanations can hardly convince. If the electron loses its particle nature in its second nature, then it also will lose its typical properties, like for instance its mass and its charge. but this is not the case.
THE vortex theory provides clear and causal answers: if the electron were a ball it continuosly would lose energy, therefore another configuration forms that does not know this problem. Here the phenomenon of transport takes an effect. The electron opens its vortex centre and takes the tiny protons and neutrons as atomic nucleus up into itself. The Bohr electron orbit with that is not a path anymore, but is occupied by the whole particle as spherical shell. This is confirmed by the not understood measurements exactly like the photos of individual atoms with the scanning electron microscope.
But now an electron does in its inside have the opposite swirl direction as the proton seen from the outside. As a consequence a force of repulsion will occur, which can be interpreted as the to the outside directed current eddy, the force of attraction for reason of the opposite charge works in the opposite direction and can be interpreted as the potential vortex effect.
If both vortices are equally powerful: $\tau_1 = \tau_2$ (5.13)
or if both forces are balanced, as one usually would say, then the object which we call an atom is in a stable state.
It probably will be a result of the incompatible swirl direction, why a very big distance results, if the electron becomes an enveloping electron. On such a shell not too many electrons have room. Because of the rotation of their own, the electron spin, they form a magnetic dipole moment, which leads to a magnetic attraction of two electrons if they put their spin axis antiparallelly.
As a "frictionless" against one another rotating pair they form two half-shells of a sphere and with that occupy the innermost shell in the hull of an atom. If the positive charge of the nucleus is still not balanced with that, then other electrons is left only the possibility to form another shell. Now this next electron takes the whole object up into itself. The new shell lies further on the outside and naturally offers room to more as only two vortices.

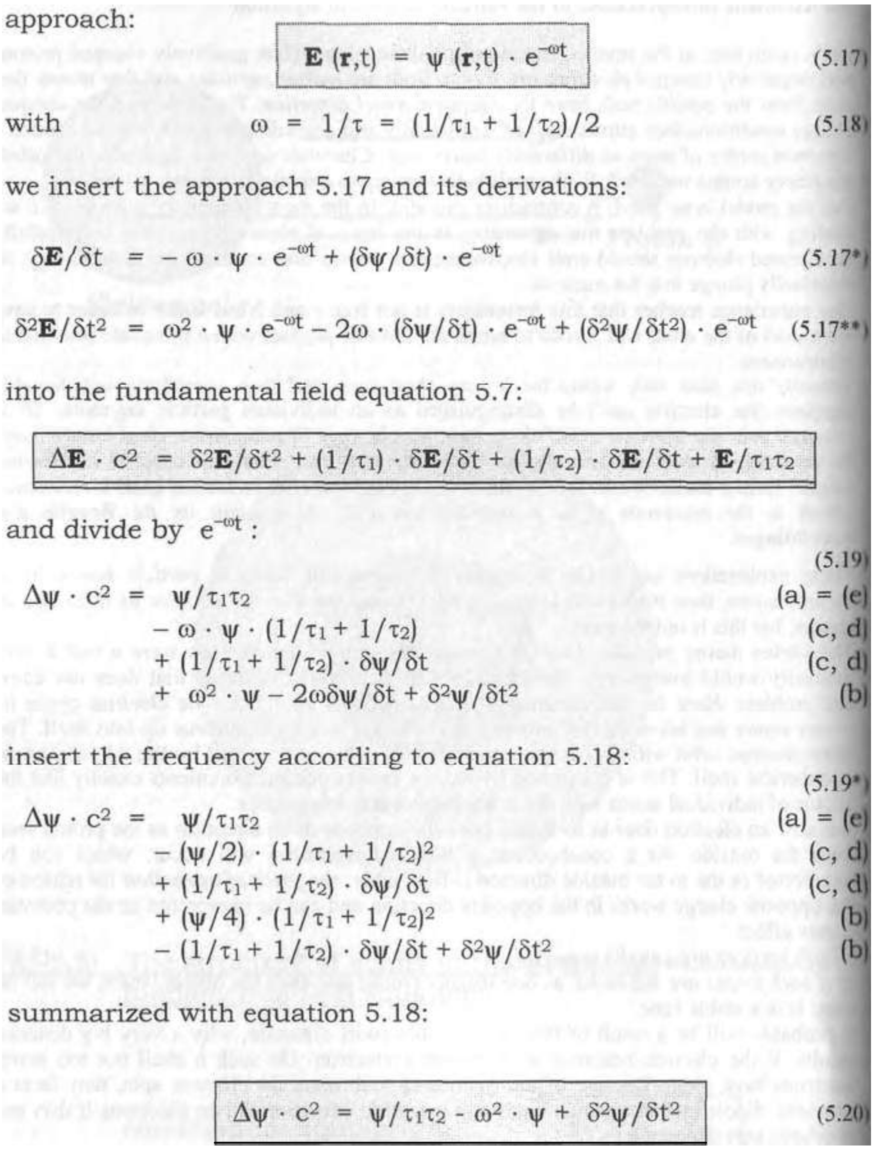

approach:

$$\boxed{\mathbf{E}(\mathbf{r},t) = \psi(\mathbf{r},t) \cdot e^{-\omega t}} \qquad (5.17)$$

with

$$\omega = 1/\tau = (1/\tau_1 + 1/\tau_2)/2 \qquad (5.18)$$

we insert the approach 5.17 and its derivations:

$$\delta\mathbf{E}/\delta t = -\omega \cdot \psi \cdot e^{-\omega t} + (\delta\psi/\delta t) \cdot e^{-\omega t} \qquad (5.17^*)$$

$$\delta^2\mathbf{E}/\delta t^2 = \omega^2 \cdot \psi \cdot e^{-\omega t} - 2\omega \cdot (\delta\psi/\delta t) \cdot e^{-\omega t} + (\delta^2\psi/\delta t^2) \cdot e^{-\omega t} \qquad (5.17^{**})$$

into the fundamental field equation 5.7:

$$\boxed{\Delta\mathbf{E} \cdot c^2 = \delta^2\mathbf{E}/\delta t^2 + (1/\tau_1) \cdot \delta\mathbf{E}/\delta t + (1/\tau_2) \cdot \delta\mathbf{E}/\delta t + \mathbf{E}/\tau_1\tau_2}$$

and divide by $e^{-\omega t}$:

$$\begin{aligned} & & (5.19)\\ \Delta\psi \cdot c^2 = {} & \psi/\tau_1\tau_2 & \text{(a)} = \text{(e)}\\ & - \omega \cdot \psi \cdot (1/\tau_1 + 1/\tau_2) & \text{(c, d)}\\ & + (1/\tau_1 + 1/\tau_2) \cdot \delta\psi/\delta t & \text{(c, d)}\\ & + \omega^2 \cdot \psi - 2\omega\delta\psi/\delta t + \delta^2\psi/\delta t^2 & \text{(b)} \end{aligned}$$

insert the frequency according to equation 5.18:

$$\begin{aligned} & & (5.19^*)\\ \Delta\psi \cdot c^2 = {} & \psi/\tau_1\tau_2 & \text{(a)} = \text{(e)}\\ & - (\psi/2) \cdot (1/\tau_1 + 1/\tau_2)^2 & \text{(c, d)}\\ & + (1/\tau_1 + 1/\tau_2) \cdot \delta\psi/\delta t & \text{(c, d)}\\ & + (\psi/4) \cdot (1/\tau_1 + 1/\tau_2)^2 & \text{(b)}\\ & - (1/\tau_1 + 1/\tau_2) \cdot \delta\psi/\delta t + \delta^2\psi/\delta t^2 & \text{(b)} \end{aligned}$$

summarized with equation 5.18:

$$\boxed{\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - \omega^2 \cdot \psi + \delta^2\psi/\delta t^2} \qquad (5.20)$$

Fig. 5.6: Derivation of the Klein-Gordon equation (5.20) from the fundamental field equation (5.7)

5.6 Derivation of the Klein-Gordon equation

The valid model of the atom today still raises problems of causality, as has been explained. An improvement was provided by an equation, which was proposed by the mathematician Schrodinger 1926 as a model description. This equation in this way missed the physical root, but it nevertheless got international acknowledgment, because it could be confirmed experimentally. Looking backwards from the result then the physical interpretation of the probability density of the resonance of the waves could be pushed afterwards.

$$i \cdot \hbar \cdot \delta\psi/\delta t = U \cdot \psi - (\hbar^2/2m) \cdot \Delta\psi \quad (5.14)$$

The Schrodinger equation is valid for matter fields (of mass m), while the interaction with a outside force field the energy U indicates. It can be won from a wave equation by conversion, which possibly is the reason why it usually is called a wave equation, although in reality it is a diffusion equation, so a vortex equation!
For the derivation Schrodinger gives the approach of a harmonic oscillation for the complex wave function ψ:

$$\boxed{\psi(\mathbf{r},t) = \phi(\mathbf{r}) \cdot e^{-i\omega t}} \quad (5.15)$$

if the entire time dependency can be described by one frequency $f = W/h$ (de-Broglie relation):

$$\omega = 2\pi f = W \cdot 2\pi/h = W/\hbar \quad (5.16)$$

The high-put goal is: if the structure of the atom is determined by the fundamental field equation 5.7 then one should be able to derive the experimentally secured Schrodinger equation and to mathematically describe the discussed special case. Also we select at first an approach periodic in time:

$$\boxed{\mathbf{E}(\mathbf{r},t) = \psi(\mathbf{r},t) \cdot e^{-\omega t}} \quad (5.17)$$

with

$$\omega = 1/\tau = (1/\tau_1 + 1/\tau_2)/2 \ . \quad (5.18)$$

We insert the approach 5.17 and its derivations into the field equation 5.7 and divide by the damping term e^{-wt}:

$$\begin{aligned} \Delta\psi \cdot c^2 = {} & \psi/\tau_1\tau_2 - \omega \cdot \psi \cdot (1/\tau_1 + 1/\tau_2) + (1/\tau_1 + 1/\tau_2) \cdot (\delta\psi/\delta t) + \\ & + \omega^2 \cdot \psi - 2\omega\,\delta\psi/\delta t + \delta^2\psi/\delta t^2 \end{aligned} \quad (5.19)$$

If as the next step the angular frequency according to equation 5.18 is inserted, then summarized the provisional intermediate result results:

$$\boxed{\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - \omega^2 \cdot \psi + \delta^2\psi/\delta t^2} \quad (5.20)$$

The derived equation 5.20 represents formally seen the Klein-Gordon equation, which is used for the description of matter waves in quantum mechanics and which particularly in the quantum field theory (e.g. mesons) plays an important role. Even if it often is regarded as the relativistic invariant generalization of the Schrodinger equation, it at a closer look is incompatible with this equation and as "genuine" wave equation it is not capable of treating vortex problems correctly, like e.g. the with the Schrodinger equation calculable quantization of our microcosm.

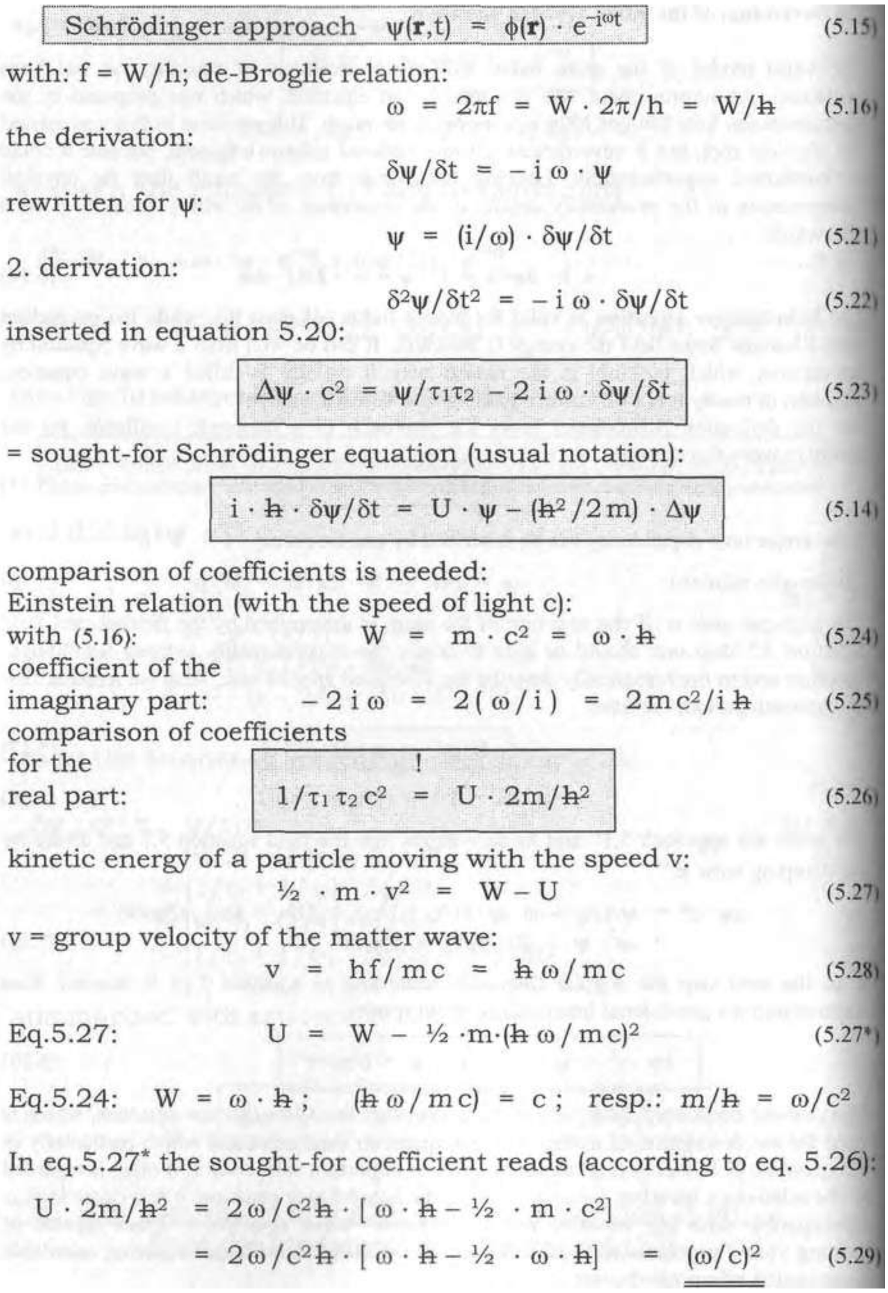

$$\boxed{\text{Schrödinger approach} \quad \psi(\mathbf{r},t) = \phi(\mathbf{r}) \cdot e^{-i\omega t}} \tag{5.15}$$

with: f = W/h; de-Broglie relation:

$$\omega = 2\pi f = W \cdot 2\pi/h = W/\hbar \tag{5.16}$$

the derivation:

$$\delta\psi/\delta t = -i\,\omega \cdot \psi$$

rewritten for ψ:

$$\psi = (i/\omega) \cdot \delta\psi/\delta t \tag{5.21}$$

2. derivation:

$$\delta^2\psi/\delta t^2 = -i\,\omega \cdot \delta\psi/\delta t \tag{5.22}$$

inserted in equation 5.20:

$$\boxed{\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - 2\,i\,\omega \cdot \delta\psi/\delta t} \tag{5.23}$$

= sought-for Schrödinger equation (usual notation):

$$\boxed{i \cdot \hbar \cdot \delta\psi/\delta t = U \cdot \psi - (\hbar^2/2m) \cdot \Delta\psi} \tag{5.14}$$

comparison of coefficients is needed:
Einstein relation (with the speed of light c):

with (5.16): $W = m \cdot c^2 = \omega \cdot \hbar$ (5.24)

coefficient of the
imaginary part: $-2\,i\,\omega = 2(\omega/i) = 2mc^2/i\,\hbar$ (5.25)

comparison of coefficients
for the
real part: $\boxed{1/\tau_1\tau_2 c^2 \overset{!}{=} U \cdot 2m/\hbar^2}$ (5.26)

kinetic energy of a particle moving with the speed v:

$$½ \cdot m \cdot v^2 = W - U \tag{5.27}$$

v = group velocity of the matter wave:

$$v = hf/mc = \hbar\,\omega/mc \tag{5.28}$$

Eq.5.27: $U = W - ½ \cdot m \cdot (\hbar\,\omega/mc)^2$ (5.27*)

Eq.5.24: $W = \omega \cdot \hbar$; $(\hbar\,\omega/mc) = c$; resp.: $m/\hbar = \omega/c^2$

In eq.5.27* the sought-for coefficient reads (according to eq. 5.26):

$$\begin{aligned} U \cdot 2m/\hbar^2 &= 2\omega/c^2\hbar \cdot [\,\omega \cdot \hbar - ½ \cdot m \cdot c^2] \\ &= 2\omega/c^2\hbar \cdot [\,\omega \cdot \hbar - ½ \cdot \omega \cdot \hbar] = \underline{\underline{(\omega/c)^2}} \end{aligned} \tag{5.29}$$

Fig. 5.7: Derivation of the time dependent Schrodinger equation

5.7 Derivation of the time dependent Schrodinger equation

With the Schrodinger approach 5.15 and its derivations the derivation is continued:

$$\boxed{\psi(\mathbf{r},t) = \phi(\mathbf{r}) \cdot e^{-i\omega t}} \,, \qquad (5.15)$$

$$\delta\psi/\delta t = -i\omega \cdot \psi \,, \quad \text{bzw.} \quad \psi = (i/\omega) \cdot \delta\psi/\delta t \qquad (5.21)$$

$$\delta^2\psi/\delta t^2 = -i\omega \cdot \delta\psi/\delta t \qquad (5.22)$$

The for a harmonic oscillation won relations according to equation 5.21 and 5.22 are now inserted into equation 5.20:

$$\boxed{\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - 2i\omega \cdot \delta\psi/\delta t}$$

This is already the sought-for Schrodinger equation, as we will see in a moment, when we have analysed the coefficients. Because, besides equation 5.16 for the total energy W, also the Einstein relation is valid (with the speed of light c):

$W = m \cdot c^2 = \omega \cdot \hbar$, (5.24)

we can replace the coefficients of the imaginary part by:

$2(\omega/i) = 2mc^2/i\hbar$ (5.25)

To achieve that equation 5.23, as required, follows from the Schrodinger equation 5.14, a comparison of coefficients is carried out for the real part:

$\boxed{1/\tau_1\tau_2 c^2 \overset{!}{=} U \cdot 2m/\hbar^2}$ (5.26)

If thc kinetic energy of a particle moving with the speed v is:

$\frac{1}{2} \cdot m \cdot v^2 = W - U$, (5.27)

then acccording to De Broglie this particle has the wavelength h/mv. The consideration of the particle as matter wave demands an agreement with the wave length c/f of an electromagnetic wave (with the phase velocity c). The particle hence has the speed v, which corresponds with the group velocity of the matter wave:

$$v = hf/mc = \hbar\omega/mc \,, \qquad (5.28)$$

if we insert v into equation 5.27 :

$U = W - \frac{1}{2} \cdot m \cdot (\hbar\omega/mc)^2$ (5.27*)

According to equation 5.24 on the one hand the total energy is W = w • h and on the other hand the relation 5.28 gives $(\hbar\omega/mc) = c$ resp.: $m/\hbar = \omega/c^2$

Inserted into equation 5.27* the sought-for coefficient reads (according to eq. 5.26):

$$\begin{aligned} U \cdot 2m/\hbar^2 &= 2\omega/c^2\hbar \cdot [\omega \cdot \hbar - \tfrac{1}{2} \cdot m \cdot c^2] \\ &= 2\omega/c^2\hbar \cdot [\omega \cdot \hbar - \tfrac{1}{2} \cdot \omega \cdot \hbar] = \underline{(\omega/c)^2} \qquad (5.29) \end{aligned}$$

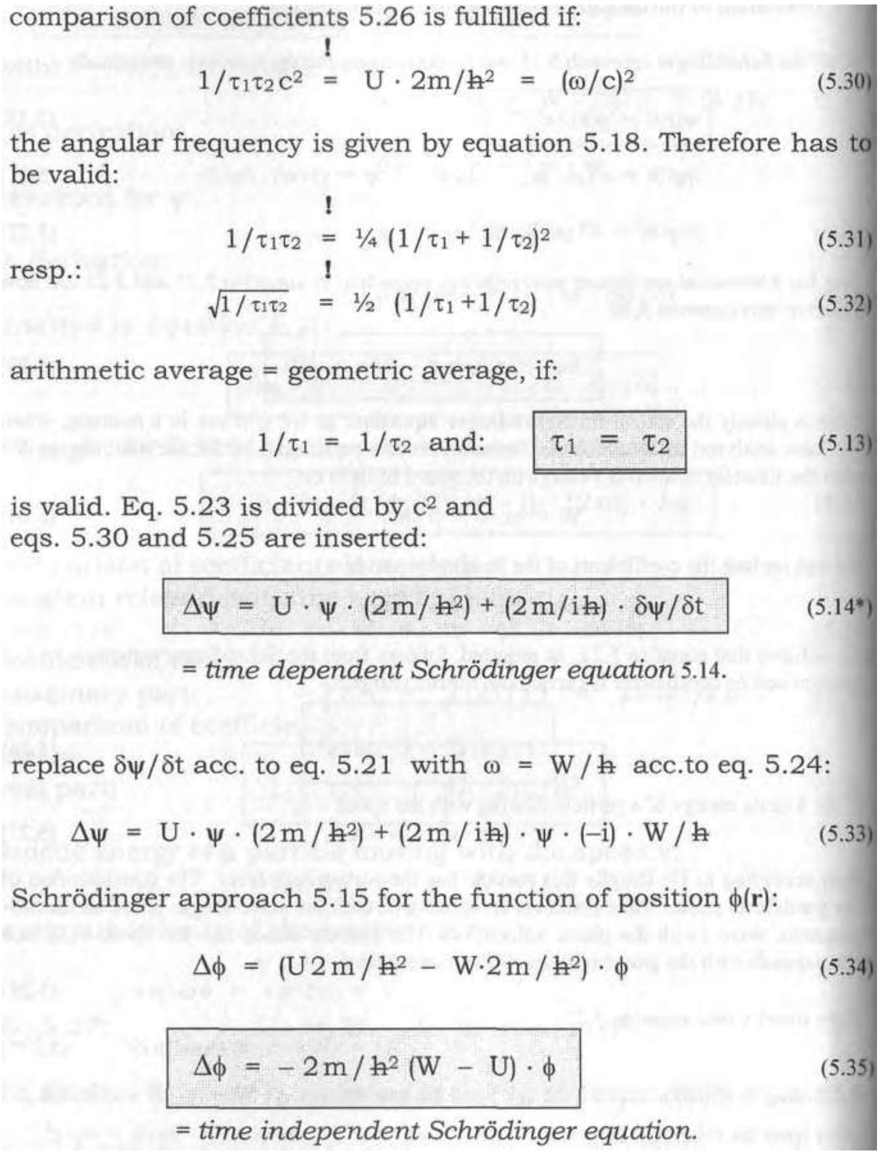

comparison of coefficients 5.26 is fulfilled if:

$$1/\tau_1\tau_2 c^2 \overset{!}{=} U \cdot 2m/\hbar^2 = (\omega/c)^2 \qquad (5.30)$$

the angular frequency is given by equation 5.18. Therefore has to be valid:

$$1/\tau_1\tau_2 \overset{!}{=} \tfrac{1}{4}\,(1/\tau_1 + 1/\tau_2)^2 \qquad (5.31)$$

resp.:

$$\sqrt{1/\tau_1\tau_2} \overset{!}{=} \tfrac{1}{2}\,(1/\tau_1 + 1/\tau_2) \qquad (5.32)$$

arithmetic average = geometric average, if:

$$1/\tau_1 = 1/\tau_2 \text{ and: } \boxed{\tau_1 = \tau_2} \qquad (5.13)$$

is valid. Eq. 5.23 is divided by c^2 and eqs. 5.30 and 5.25 are inserted:

$$\boxed{\Delta\psi = U \cdot \psi \cdot (2m/\hbar^2) + (2m/i\hbar) \cdot \delta\psi/\delta t} \qquad (5.14^*)$$

= *time dependent Schrödinger equation* 5.14.

replace $\delta\psi/\delta t$ acc. to eq. 5.21 with $\omega = W/\hbar$ acc.to eq. 5.24:

$$\Delta\psi = U \cdot \psi \cdot (2m/\hbar^2) + (2m/i\hbar) \cdot \psi \cdot (-i) \cdot W/\hbar \qquad (5.33)$$

Schrödinger approach 5.15 for the function of position $\phi(\mathbf{r})$:

$$\Delta\phi = (U\,2m/\hbar^2 - W \cdot 2m/\hbar^2) \cdot \phi \qquad (5.34)$$

$$\boxed{\Delta\phi = -\,2m/\hbar^2\,(W - U) \cdot \phi} \qquad (5.35)$$

= *time independent Schrödinger equation.*

Fig. 5.8: Derivation of the time independent Schrodinger equation

5.8 Derivation of the time independent Schrodinger equation

The goal is reached if we are capable to fulfil the comparison of coefficients 5.26:

$$U \cdot 2m/\hbar^2 \overset{!}{=} (\omega/c)^2 = 1/\tau_1\tau_2 c^2 \tag{5.30}$$

The angular frequency w is given by equation 5.18. Therefore has to be valid:

$$\tfrac{1}{4}\,(1/\tau_1 + 1/\tau_2)^2 \overset{!}{=} 1/\tau_1\tau_2 \tag{5.31}$$

$$\tfrac{1}{2} \cdot (1/\tau_1 + 1/\tau_1) \overset{!}{=} \sqrt{1/\tau_1\tau_2} \tag{5.32}$$

As is well-known the arithmetic and the geometric average only correspond in case the variables are identical. In this case, as already required in equation 5.13:

$$\tau_1 = \tau_2\,. \tag{5.13}$$

has to hold.
From this we can draw the conclusion that the Schrodinger equation is just applicable to the described special case (according to eq. 5.13), in which the eddy current, which tries to increase the particle or its circular path and the potential vortex, which keeps the atoms together and also is responsible for the stability of the elementary particles, are of identical order of magnitude.

As a check equation 5.23 is divided by c^2 and equations 5.30 and 5.25 are inserted:

$$\boxed{\Delta\psi = U \cdot \psi \cdot (2m/\hbar^2) + (2m/i\hbar) \cdot \delta\psi/\delta t} \tag{5.14*}$$

This is the time dependent Schrodinger equation 5.14 resolved for $\Delta\psi$.

Next we replace $\delta\psi/\delta t$ according to equation 5.21 with $\omega = W/\hbar$ acc. to equation 5.24:

$$\Delta\psi = U \cdot \psi \cdot (2m/\hbar^2) + (2m/i\hbar) \cdot \psi \cdot (-i) \cdot W/\hbar \tag{5.33}$$

If we separate the space variables $\phi(\mathbf{r})$ from time by the Schrodinger approach 5.15 we obtain:

$$\Delta\phi = (U\,2m/\hbar^2 - W \cdot 2m/\hbar^2) \cdot \phi \tag{5.34}$$

This quation 5.34 for the function of space coordinates $\phi(\mathbf{r})$ is the time independent Schrodinger equation:

$$\boxed{\Delta\phi = -2m/\hbar^2\,(W - U) \cdot \phi} \tag{5.35}$$

The solutions of this equation which fulfil all the conditions that can be asked of them (of finiteness, steadiness, uniqueness etc.), are called eigenfunctions. The existence of corresponding discrete values of the energy W, also called eigenvalues of the Schrodinger equation, are the mathematical reason for the different quantum postulates.

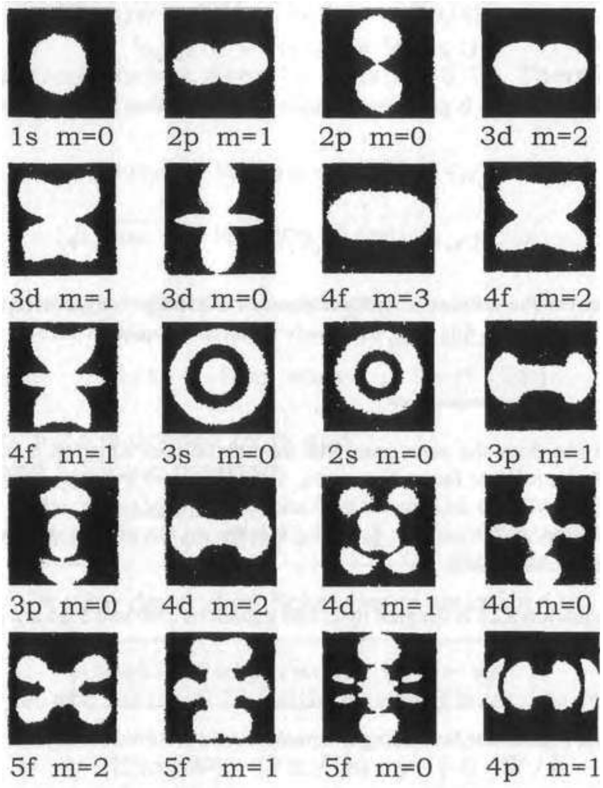
1s m=0
2p m=1
2p m=0
3d m=2
3d m=1
3d m=0
4f m=3
4f m=2
4f m=1
3s m=0
2s m=0
3p m=1
3p m=0
4d m=2
4d m=1
4d m=0
5f m=2
5f m=1
5f m=0
4p m=1

Fig. 5.9: Photographs of models of the probability densities for different states of the hydrogen atom. The densities are symmetrical if rotated around the vertical axis<i>

taken from:

<i>: U. Gradmann/H. Wolter: Grundlagen der Atomphysik, AVG, Frankfurt a. M. 1971, P. 190.

5.9 Interpretation of the Schrodinger equation

The interpretation of the Schrodinger equation is still disputed among physicists, because the concept of wave packets contradicts the corpuscular nature of the elementary particles. Further the difficulty is added that wave packets at a closer look never are connected, run apart more or less fast, and really nothing can hinder them doing that. But for a particle the connection represents a physical fact. Then there can be no talk of causality anymore. The monocausal division into two different levels of reality, in a space-timely localization and in an energetic description, does not represent a solution but rather the opposite, the abolition of the so-called dual nature. As has been shown, the potential vortex is able to achieve this with the help of its concentration effect.
But from the introduction of this new field phenomenon arises the necessity to interpret the causes for the calculable and with measuring techniques testable solutions of the Schrodinger equation in a new way. Laws of nature do not know a possibility to choose! If they have been accepted as correct, they necessarily have to be applied.

Three hundred years ago the scholars had an argument, whether a division of physical pheomena, like Newton had proposed it, would be allowed to afterwards investigate them in the laboratory individually and isolated from other influences or if one better should proceed in an integrated manner, like for instance Descartes with his cartesian vortex theory. He imagined the celestial bodies floating in ethereal vortices. One absolutely was aware that the whole had to be more than the sum of every single realizaton, but the since Demokrit discussed vortex idea had to make room for the overwhelming successes of the method of Newton. And this idea after 2100 years was stamped, to in the meantime almost have fallen into oblivion.
Today, where this recipe for success in many areas already hits the limits of the physical possibilities, we should remember the teachings of the ancients and take up again the vortex idea It of course is true that only details are calculable mathematically and that nature, the big whole, stays incalculable, wherein problems can be seen.
If we consider the fundamental field equation 5.7, we find confirmed that actually no mathematician is capable to give a generally valid solution for this four-dimensional partial differential equation. Only restrictive special cases for a harmonic excitation or for certain spatial boundary conditions are calculable. The derived Schrodinger equation is such a case and for us particularly interesting, because it is an eigenvalue equation. The eigenvalues describe in a mathematical manner the with measuring techniques testable structures of the potential vortex .
Other eigenvalue equations are also derivable, like the Klein-Gordon equation or the Lionville equation, which is applied successfully in chaos theories. So our view opens, if chaotic systems like turbulences can be calculated as special cases of the same field equation and should be derivable from this equation.
The in pictures recorded and published structures, which at night should have come into being in corn fields, often look like the eigenvalues of a corresponding equation. The ripe ears thereby lie in clean vortex structures flat on the soil. Possibly potential vortices have charged the ears to such high field strength values that they have been pulled to the soil by the Coulomb forces.

Consequences resulting from the derivation of the Schrodinger equation from the fundamental field equation 5.7:

1. *Any experiment which confirms the Schrödinger equation is with that able to confirm at the same time the existence of the newly discovered potential vortex and the correctness of the field-theoretical approach.*

2. *Properties of the atomic hull and of the atomic nucleus, which can be described with the Schrödinger equation, can as of now be interpreted as an electromagnetic phenomenon.*

3. *There exist no particles or wave packets from matter waves, but only configurations consisting of potential vortices and current eddies.*

4. *There exists no matter! What we call matter is nothing but an electromagnetic state of oscillation of empty space.*

The relation between the energy of oscillation and the mass is described by the relation named after Albert Einstein

$$\boxed{E = mc^2} \quad . \qquad (6.1 = 5.24)$$

Fig. 6.1: Derivation of the Schrodinger equation, power of proof and consequences

6. Theory of objectivity

6.1 Proof

A new theory only has chances on acknowledgment if it is provable. For that physical phenomena in the sense of the new theory are calculated and independently of this experiments are being carried out. If the calculations are confirmed by reproducible measurement results, then with that the correctness of the approach is proven.
In the here presented case we have chosen the field-theoretical approach instead of the usual quantum physical approach. As a consequence of this we had found as a new phenomenon the vortex of the electric field. With regard to the normally used Maxwell theory this resulted in changed field equations in a dual formulation. If both equations, each of which describes a source-free vortex field, are inserted into each other the result is an only in time and space formulated, generally valid and hence fundamental field equation (5.7, fig. 5.1).
This equation has many special cases; one of them, the Schrodinger equation, could be derived by using an approach which was harmonic in time. We renounced to give special solutions of the Schrodinger equation, because these are printed in numerous text books. On the other hand experiments are known, which are capable to confirm the theoretical solutions and thus to prove the Schrodinger equation. The eigenvalues of the equation describe for instance the shell-shaped structure of the atoms with the by Niels Bohr given radii.
Now this already proven equation was derived from the new field-theoretical approach. Thus for the special case, the area where the Schrodinger equation is valid, the new theory can be said to be proven (fig. 6.1).
We still are not content with that and put another stone on top: we will calculate the quantum properties of the elementary particles for ourselves. These until now have only been measured. Today is merely sought for symmetries and for models of explanation, like e.g. the quark-hypothesis. From a calculation science is miles and miles away. We will compare the calculation results with the measurement values. Then everyone can check and compare for him or herself.
The conditions in an elementary particle are completely different. Here it concerns the vortex itself, whereas the model of the atom merely describes vortex properties, so-called actions at a distance. The differences in size and distances for an atom lie more than five powers of ten over those of a particle!
Here a new problem of causality comes to light, at which we now must have a critical look: the question of the by Einstein postulated constancy and universality of the speed of light. Seen from a relativistic and subjective point of view of an observer, Einstein by all means may be right. But may such a theory be generalized? How are the measurements concerning the speed of light and the relativity of space and time to be judged when looking at them objectively?
The current measurements of speeds faster than light speak a clear language and represent a challenge (fig. 3.1, violation of the principle of causality no. 5).

The electron as a spherical capacitor (see fig. 4.3):

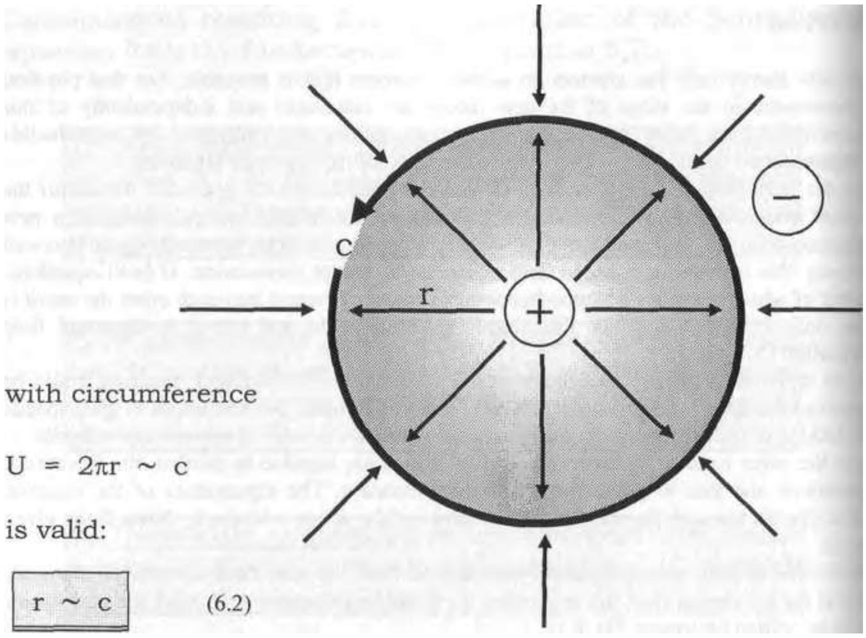

with circumference

$U = 2\pi r \sim c$

is valid:

$$\boxed{r \sim c} \quad (6.2)$$

field theoretical approach (vortex particles):
The amount of energy bound in the inside of the particle is identical with the free and measurable amount of energy on the outside of the particle.

(If the number of particles is left unchanged):
In an isolated system the sum of the energy is constant

(particle = electromagnetic vortex)
Energy is a state description of electromagnetism.

Fig. 6.2: Derivation of the law of conservation of energy

6.2 Law of conservation of energy

Let the starting-point for our considerations be the electromagnetic wave in a particle-free vacuum. Here no vortices appear, so that the plane wave can propagate undamped with the speed of light, and in this way a transport of energy takes place. Electric and magnetic energy each are the same magnitude.

Let's now imagine the symmetry is disturbed as the wave is "slowed down" on one side. As a possible result the wave rolls up to a spherical vortex.

As we will see such a process is possible, for instance at impact on a strong field. Thus part of the energy is bound in the inside. This part from now on withdraws itself from every possibility to measure it. We can only measure the second part of the field energy, with which the particle interacts with its neighbourhood.

We can assume that:

The amount of energy bound in the inside of the particle is identical with the free and measurable amount of energy on the outside of the particle.

The same energy W_e = 0,51 MeV, we attribute to the electron for reason of its mass with the help of the Einstein relation (6.1), is also bound in its inside. This conclusion is also applicable to other elementary particles and with that to all matter.

We here again recognize the principle of the duality between the to the outside striving eddy current in the inside of the elementary vortex and the concentrating potential vortex on the outside. Thus also seen energetically both are of the same magnitude. Whereas in the case of the electromagnetic wave it concerns a symmetrical oscillation around "zero", by the process of quantization, by the rolling up to a spherical vortex, there forms an energetic state of space different from zero. The order of magnitude is determined by the number of elementary vortices, of which the particles and all matter consist.

Anti-matter forms the opposite energetic state and this again is for the particles of matter available in their inside in a bound form.

As long as we do not artificially produce new elementary vortices and thus keep the number of available vortices constant, the energetic state will not change, or as it is formulated in text books:

In an isolated system the sum of the energy is constant.

THE law of conservation of energy is not an axiom, but follows without compulsion from the vortex theory. It is not elementary, but a consistently derivable consequence of the field-theoretical approach, according to which solely the field acts as cause for all other physical phenomena, also for the conservation of energy! Because the cause of it is the electromagnetic field, the following has to hold:

Energy is a state description of electromagnetism.

Now we finally can explain why energy can be converted. Different forms of energy only are different forms of formation of the same phenomenon!

Of course this statement of the field-theoretical approach does not yet explain what, for instance, the temperature has to do with electromagnetism. I ask for some patience; no question will be left unanswered.

From

$$\boxed{c \sim r} \qquad (6.2)$$

follows:

The speed of light determines the size of the elementary particles.

Energy of a capacitor:

$$W = Q^2/C, \qquad (6.3)$$

written down for the electron (with the Einstein relation):

$$W_e = e^2/C_e = m_e c^2 = 0{,}51\ \mathrm{MeV} \qquad (6.1)$$

Capacity of a spherical capacitor:

$$C_e = \varepsilon_0 \cdot 4\pi r_e \qquad (6.4)$$

"classical" radius of the electron[i] is:

$$r_e = e^2/\varepsilon_0 \cdot 4\pi \cdot W_e \qquad (6.5)$$

$$r_e = 2{,}82 \bullet 10^{-15}\ \mathrm{m} \qquad (6.6)$$

in the case of Kuchling[ii] the radius of the electron is:

$$r_e = 1{,}41 \bullet 10^{-15}\ \mathrm{m}\ . \qquad (6.7)$$

Fig. 6.3: Calculation of the radius of the electron.

[i] Mende, Simon: Physik, Gl. 10.39, VEB-Leipzig, 4. Aufl.

[ii] Kuchling: Physik, Gl. At4, VEB-Leipzig, bis einschl. 11. Auflage 1974

6.3 Radius of the electron

For the crucial process, in which the electromagnetic wave rolls up to a vortex, it is for reasons of continuity to be expected that the velocity of propagation remains equal that thus for the vortex oscillation exactly like for the electromagnetic wave the speed of light is determining. The direction of propagation in the case of the vortex takes place perpendicular to the in fig. 6.2 shown field direction of the electric field strength. Not even in that both field-phenomena differ.

Summarizing: the propagation takes place with the speed of light c along a circular path with the perimeter $U = 2\pi r$. Therefore holds:

$$c \sim r \qquad (6.2)$$

According to this equation the radius and with that the size of the electron is determined by the speed of light. Therefore the question of the size of the electron is raised.

The energy interpretation predicts that for the theoretical case of a change of size the energy density in the inside of the particle is influenced that however the quantity of the included energy remains unchanged. We therefore can further proceed from the assumption that the bound amount of energy is independent of the size of the particle!

Consequently for the elementary quantum the energy W_e = 0,51 MeV is assumed, which it has acccording to the Einstein relation $W_e = m_e c^2$. For the electron of mass m_e the with measuring techniques determined value is inserted.

The spherical electrode of a spherical capacitor with the above given energy W_e (according to eq. 6.1) and the capacity C_e (according to equation 6.4, fig. 6.3) represents a very realistic model of the negatively charged particle.

In this manner the classical radius of the electron is calculated to be<i>: $r_e = 2{,}82 \cdot 10^{-15}$ m.

But in the case of Kuchling it only is half this size<ii>, what according to equation 6.2 would mean that in the case of Kuchling the light would be on the way only half this fast <iii>. Therefore if one is careful, one prefers to be silent concerning this delicate theme and if one is honest, one admits not to know anything exact.

Not only the electron but also all the other elementary particles are according to the field-theoretical approach formed from concentrated potential vortices. For these equation 6.2 hence has to hold in the same manner, so that more generalized we can conclude:

The speed of light determines the size of the elementary particles.

This statement is incompatible with the assumption of a constant speed of light! Because then all elementary particles would have identical size. As is known, however, are the building parts of the atomic nucleus, the protons and neutrons very much smaller than individual electrons. The constancy of the speed of light is to be questioned. This question is of such an elementary importance that we are not content with these considerations and in addition undertake a mathematical derivation in the sense of the field approach.

<iii>: Difference = Thomas factor

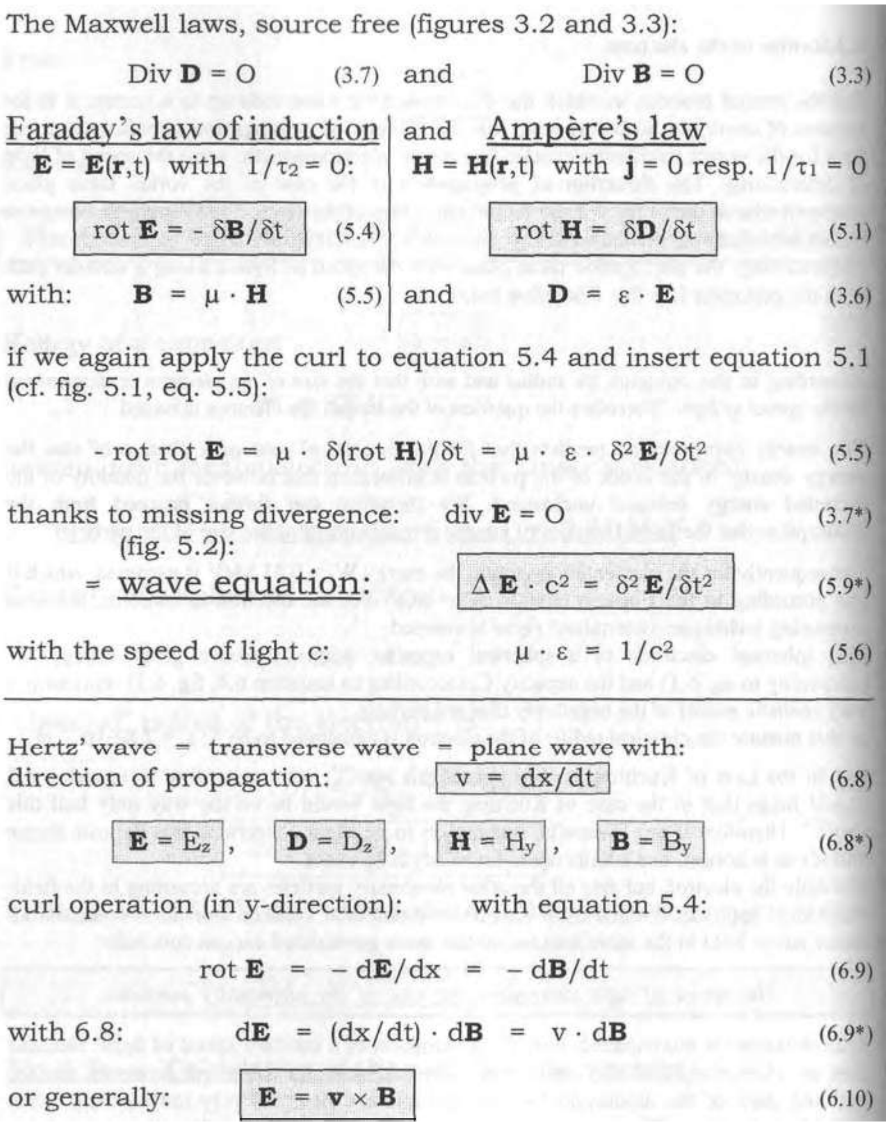

The Maxwell laws, source free (figures 3.2 and 3.3):

$\mathrm{Div}\,\mathbf{D} = 0$ (3.7) and $\mathrm{Div}\,\mathbf{B} = 0$ (3.3)

<u>Faraday's law of induction</u> and <u>Ampère's law</u>

$\mathbf{E} = \mathbf{E}(\mathbf{r},t)$ with $1/\tau_2 = 0$: | $\mathbf{H} = \mathbf{H}(\mathbf{r},t)$ with $\mathbf{j} = 0$ resp. $1/\tau_1 = 0$

$\boxed{\mathrm{rot}\,\mathbf{E} = -\,\delta\mathbf{B}/\delta t}$ (5.4) | $\boxed{\mathrm{rot}\,\mathbf{H} = \delta\mathbf{D}/\delta t}$ (5.1)

with: $\mathbf{B} = \mu \cdot \mathbf{H}$ (5.5) and $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (3.6)

if we again apply the curl to equation 5.4 and insert equation 5.1 (cf. fig. 5.1, eq. 5.5):

$$-\,\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \mu \cdot \delta(\mathrm{rot}\,\mathbf{H})/\delta t = \mu \cdot \varepsilon \cdot \delta^2\mathbf{E}/\delta t^2 \qquad (5.5)$$

thanks to missing divergence: $\mathrm{div}\,\mathbf{E} = 0$ (3.7*)
(fig. 5.2):

= <u>wave equation:</u> $\boxed{\Delta\mathbf{E} \cdot c^2 = \delta^2\mathbf{E}/\delta t^2}$ (5.9*)

with the speed of light c: $\mu \cdot \varepsilon = 1/c^2$ (5.6)

Hertz' wave = transverse wave = plane wave with:

direction of propagation: $\boxed{v = dx/dt}$ (6.8)

$\boxed{\mathbf{E} = E_z}$, $\boxed{\mathbf{D} = D_z}$, $\boxed{\mathbf{H} = H_y}$, $\boxed{\mathbf{B} = B_y}$ (6.8*)

curl operation (in y-direction): with equation 5.4:

$$\mathrm{rot}\,\mathbf{E} = -\,d\mathbf{E}/dx = -\,d\mathbf{B}/dt \qquad (6.9)$$

with 6.8: $d\mathbf{E} = (dx/dt) \cdot d\mathbf{B} = v \cdot d\mathbf{B}$ (6.9*)

or generally: $\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (6.10)

Fig. 6.4: Derivation of the laws of transformation

<i>: Prof. G. Bosse in his text book in reversed direction derives the Faraday law of induction from the law of transformation 6.10, which he again derives from considerations about the Lorentz force. G. Bosse, Grundlagen der Elektrotechnik II, BI 183, Hochschultaschenbucher-Verlag, Mannheim 1967

6.4 The Maxwell field equations

The laws of transformation of the electromagnetic field shall form the starting-point for the coming up considerations. To exclude any doubts with regard to the interpretation, the equations will be derived from the Maxwell laws under the assumption that no sources or charge carriers are present (fig. 3.2 and 3.3) and as a consequence no current density ($j = 0$) is to be expected.
This corresponds to the vanishing of the time independent terms, which consequently are responsible for the occurring of force effects like e.g. the Lorentz force. Only at the end of this derivation we can understand the sense of this assumption (with $1/\tau_1 = 0$ and $1/\tau_2 = 0$). The procedure at first corresponds to that of fig. 5.1. Here the fundamental field equation had been derived from Faraday's law of induction and Ampere's law. With the assumptions made this time the in fig. 5.2 treated undamped wave equation is left (5.9, here 5.9*). Whom the derivation is still present can go in at this point.
In a sufficiently great distance from the source we are dealing with a plane wave, in which the field factors only depend on the direction of propagation x. The Hertz' wave is a transverse wave, in which the field pointers oscillate perpendicular to the direction of propagation and in addition stand perpendicular to each other:

$$\boxed{v = dx/dt} \quad (6.8), \qquad \boxed{\mathbf{E} = E_z, \quad \mathbf{D} = D_z, \quad \mathbf{H} = H_y, \quad \mathbf{B} = B_y} \quad (6.8^*)$$

The curl, applied to the electric field pointer, itself points in the y-direction:
rot E = - dE/dx . This for the transverse wave carried out curl operation is now compared with Faraday's law of induction (5.4):

$$\text{rot } E = -dE/dx = -dB/dt \qquad (6.9)$$

The relation won in a mathematical way, with the speed fixed by (6.8), reads:

$$dE = (dx/dt) \bullet dB = v * dB \qquad (6.9^*)$$

The result of this derivation at first only is valid for the introduced simplification, for instance for the case of the transverse electromagnetic wave. Better known is apart from that the generalized formulation, which among others by G. Bosse<i> is called law of transformation.

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} \qquad (6.10)$$

With Ampere's law (5.1) we now should proceed in an analogous manner. The result is:

$$\boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}} \quad (6.10^*)$$

This equation 6.10* is given among others by Simonyi<ii>. Now that we know, under which circumstances these equations of transformation can be derived from the Maxwell equations, the actual work can start.

<ii>: K. Simonyi, Theoretische Elektrotechnik, 7. Auflage VEB Verlag Berlin 1979. pp. 921 - 924; In addition see chapter 27.8 in part 3 of this book.

Laws of transformation:

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} = \mathbf{v} \times \mu \cdot \mathbf{H} \qquad (6.10)$$

$$\boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}} = -\mathbf{v} \times \varepsilon \cdot \mathbf{E} \qquad (6.10^*)$$

and

> We experience the magnetic field as an electric field and the electric field as a magnetic field simply and solely for reason of the same relative motion!

The component of the direction of motion perpendicular to the area defined by the field pointers:

$$E = v \cdot \mu \cdot H \qquad (6.11)$$

and

$$H = -v \cdot \varepsilon \cdot E \qquad (6.11^*)$$

with the relations of material:

$$B = \mu \cdot H \qquad (3.5)$$

$$D = \varepsilon \cdot E \qquad (3.6)$$

with the speed of light:

$$c = 1/\sqrt{\mu \cdot \varepsilon} \qquad (5.6)$$

additional field:

$$E_z = -v^2 \cdot \mu \cdot \varepsilon \cdot E = -(v^2/c^2) \cdot E \qquad (6.12)$$

resp.

$$H_z = -v^2 \cdot \varepsilon \cdot \mu \cdot H = -(v^2/c^2) \cdot H \qquad (6.12^*)$$

basic field: E resp. H (at $v = O$)

measurable overall field: (if $v \neq O$)

$$E_o = E + E_z \qquad (6.13)$$

resp.

$$Ho = H + H_z \qquad (6.13^*)$$

Fig. 6.5: Properties of transformation of the electromagnetic field.

6.5 Equations of transformation

As a consequence of the in fig. 6.5 again written down laws of transformation of the electromagnetic field (6.10 and 6.10*) magnetic phenomena can be traced back to electric phenomena and vice versa. The mathematical formulation reveals us the two sides of the same medal and points to a perfect duality between both fields and their factors of description.

Because a way exists, as is shown here, in which the equations of transformation can be derived from the Maxwell field equations, the same generally valid and extensive importance should be attributed to them. They can with the same right be called the foundation of electromagnetism. Wherein does lie its message for physics, the always curious researcher will ask? For that the relations of material 3.5 and 3.6 are completed:

$\boxed{\mathbf{E} = \mathbf{v} \times \mu \cdot \mathbf{H}}$ (6.10) und $\boxed{\mathbf{H} = -\mathbf{v} \times \varepsilon \cdot \mathbf{E}}$. (6.10*)

The here presented equations state, that we measure an electric field strength E, if we are moving with regard to a magnetic field H with the speed v and vice versa. The electric and the magnetic field therefore prove to be an experience of the observing person and we can say:

We experience the magnetic field as electric field and the electric field as magnetic field simply and solely for reason of the relative motion!

Let's assume, v is the component of the relative velocity (6.8), which stands perpendicular to the area defined by the field pointers (6.8*), then the equations of transformation (6.9* with 3.5) now read:

$E = v \cdot \mu \cdot H$ (6.11) and $H = -v \cdot \varepsilon \cdot E$. (6.11*)

If we are moving with the velocity v in a basic field which is present with the field strength E, then according to equation 6.11* we observe a magnetic field, which again according to equation 6.11 is to be interpreted as an additional electric field E_z:

$E_z = -v^2 \cdot \mu \cdot \varepsilon \cdot E = -(v^2/c^2) \cdot E$ (6.12)

In duality equation 6.11 inserted into equation 6.11* provides for the magnetic field strength a corresponding additional field H_z:

$H_z = -v^2 \cdot \mu \cdot \varepsilon \cdot H = -(v^2/c^2) \cdot H$ (6.12*)

We obviously owe the measurable overlap fields in a laboratory simply and solely to the relative velocity v with which the laboratory is moving. But now we must pay attention to the fact that a terrestrial laboratory rotates along with the earth, that the earth orbits the sun and the sun again rotates around the centre of the milky way. Eventually the whole milky way is on the way in the cosmos with a galactic, for us hardly understandable speed. If we further take into consideration that for every subsystem an additional field occurs as a consequence of the relative motion with regard to the super ordinate system, then one additonal field follows after the next and overlaps this one.

Let's imagine, the relative velocity could be reduced towards zero - and maybe we are moving around such a cosmic point - then here no overlapping field would be measurable.

<*>: A derivation using vectors is written in chapter 28 (part 3).

Additional field (from fig. 6.5):

(6.12)
(6.12*)

$$\boxed{\begin{aligned} E_z &= -(v^2/c^2)\cdot E \\ H_z &= -(v^2/c^2)\cdot H \end{aligned}}$$

and

Superposition of the fields:
The additional field (E_z resp. H_z) overlaps the basic field
(E resp. H) to produce the measurable overall field (E_0 resp. H_0):

(6.13)

$$\boxed{\begin{aligned} E_0 &= E + E_z = E\cdot(1 - v^2/c^2) \\ H_0 &= H + H_z = H\cdot(1 - v^2/c^2) \end{aligned}}$$

(6.13*)

transformed:

$$(1 - \frac{v^2}{c^2}) = \frac{E_0}{E} = \frac{H_0}{H}$$

(6.14)

for the Lorentz contraction holds apart from that:

(6.14*)

$$(1 - \frac{v^2}{c^2}) = (\frac{1}{l_0})^2$$

From the comparison
Follows

(6.14**)

$$\frac{E_0}{E} = \frac{H_0}{H} = (\frac{1}{l_0})^2$$

the proportionality:

$\boxed{E, H \sim 1/l^2}$ and $E_0, H_0 \sim 1/l_0^2$ (6.15)

Fig. 6.6: The field dependency of the Lorentz contraction

6.6 Field overlap

Field vectors can be superpositioned. In this manner the additional field E_z resp. H_z which depends on the velocity, according to equation 6.10, overlaps the respective basic field (E resp. H) to produce the measurable overall field (E_0 resp. H_o):

$$\boxed{E_0 = E + E_z = E \cdot (1 - v^2/c^2)} \quad (6.13)$$

$$\boxed{H_0 = H + H_z = H \cdot (1 - v^2/c^2)} \quad (6.13^*)$$

In the result something surprising the factor $(l-v^2/c^2)$ appears, which is well-known from the special theory of relativity and for instance appears in the Lorentz contraction.
If we rewrite both equations for the characteristic factor and compare with the in a purely mathematical way, over the Lorentz transformation, won length contraction $(1 - v^2/c^2) = (l/l_0)^2$, then it becomes clear that the Lorentz contraction physically seen should have its cause in the changed field conditions which a with relativistic speed moving body finds with regard to a resting body.

$$1 - \frac{v^2}{c^2} = \frac{E_0}{E} = \frac{H_0}{H} \overset{!}{=} (\frac{l}{l_0})^2 \quad (6.14)$$

The equation is a compulsionless consequence of known physical laws. In this derivation actually no new factor was introduced and nevertheless a completely new picture for the natural scientific reality results<i>.

In our observer system, where the field E_o exists, a rule has its proper length l_0. In another system, which is moving with the speed v relative to the observer, as a consequence of the here prevailing field E the corresponding rule has a length 1. In which relation the factors stand to each other, is described by equation 6.14. Accordingly the following proportionality holds:

$$\boxed{E, H \sim 1/l^2} \text{ and } \boxed{E_0, H_0 \sim 1/l_0^2} \quad (6.15)$$

If we are exterior to a very fast moving body with velocity v, we immediately can observe how this body for reason of its relative velocity experiences the calculated additional field and in this way experiences a length contraction. If the observer is moving along with the body, then he purely subjective seen doesn't detect a length contraction, because he himself and his entire measuring technique is subjected to the same length contraction. From the axiomatic approach what would be, if the field, which itself only represents an experience, would determine perceptible space and its dimensions, quickly a fundamental realization can develop if the described experiences should coincide with real observations.

<i> : Because in this point of view the subjective status of the observer is determining, it is not completely impossible that there is an error in the interpretation of the equations of transformation (6.10 and 6.10*). But because we started from the same point of view of the observer for the derivation of the length contraction from the Lorentz transformation, here the same error is to be expected. In putting both results equal (6.14), a like constituted error on both sides will cancel out in any case and the result stays above all doubts!

$$\boxed{E, H \sim 1/l^2} \quad \text{and} \quad \boxed{E_0, H_0 \sim 1/l_0^2} \qquad (6.15)$$

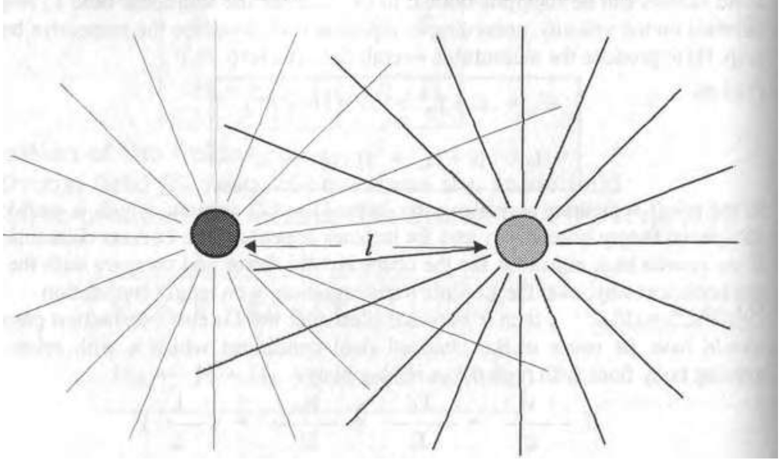

(Model):
Two particles of matter each in the field of the other particle.

Two elementary particles or two accumulations of matter consisting of these are able to reduce the distance to each other for reason of their fields, which we interpret as a force of attraction.

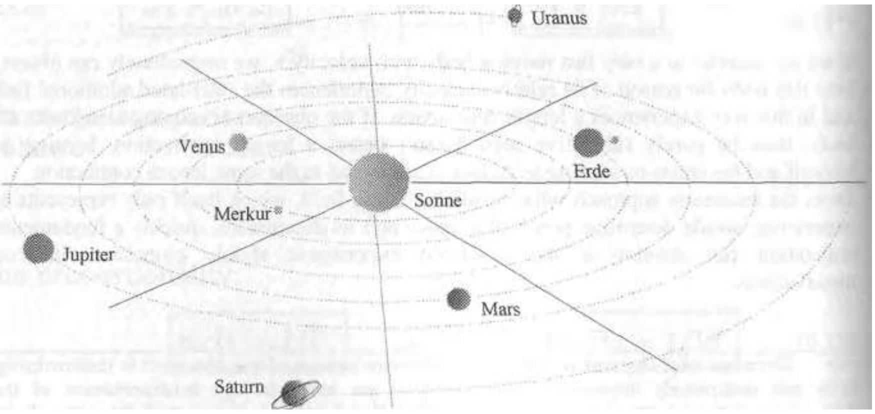
Uranus
Venus
Erde
Sonne
Merkur
Jupiter
Mars
Saturn

B: (Example): The orbits of the planets in the field of the sun.

Fig. 6.7: The influence of the field on interactions.

6.7 Field dependent curvature of space

Let's assume, an accumulation of matter, as big as our earth, wanted to fly past the sun in the distance earth-sun. But it would not succeed. Because the fields arising from the sun decreases with increasing distance and according to equation 6.15 as a consequence the size of the particles of matter increases. The planet hence is more strongly contracted on its side turned towards the sun, as on the turned away "night side". It bends towards the sun and its flight path becomes a circular path around the sun. That is the interaction known as gravitation!

To an earth inhabitant this curvature reveals itself merely in the observation that the duration of sunshine at daytime is longer, than it would be expected to be under the assumption of the earth as a homogeneous sphere. In this context one willingly speaks of a curvature of space. Actually it is a curvature of matter under the influence of the field dependent length contraction.
Exactly this contraction the planets owe their circular orbits around the sun and by no means the equilibrium of forces between the force of attraction and the centrifugal force (fig. 6.7 B). It obviously is a fundamental mistake to think that gravitation would causally be connected with a force effect!
If, in this context, we speak of a force of attraction for the sake of our subjective observation, then we must realize that it merely can concern an auxiliary term founded in usefulness.

A thought experiment should bring us clarity (fig. 6.7 A). The field, which surrounds every particle of matter, reaches till infinity but becomes less effective with increasing distance. If the distance between two particles is 1, then one particle is in the field of the other particle. As a consequence of the field the length 1 reduces and in this way the size determining field increases, which again leads to a further reduction of length etc. As a consequence it can be observed that both particles are moving towards each other. We speak of a force of attraction, because we can't register the influence of the field with our senses.
In this way the consistent result that we and our environment at daytime must be smaller than in the night will as well remain hidden. We experience the effect only indirectly as gravitational pull of the earth.

Because we don't see the cause of a subjectively observed force effect, for the electromagnetic interaction, just as for the gravitation, the field dependency of the length contraction will be responsible. Hence the following conclusion holds for both interactions equally way.

Two elementary particles or two accumulations of matter consisting of these are able to reduce the distance to each other for reason of their fields, which we interpret as a force of attraction.

Now the question still is open, why gravitation only knows forces of attraction, whereas the electromagnetic interaction also permits forces of repulsion and which are the causal fields for each.

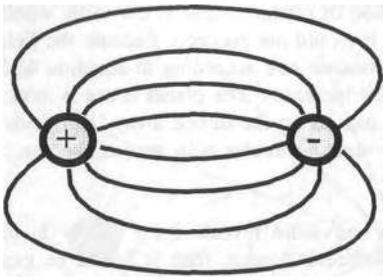

A: The field lines of the E-field for unlike charged particles

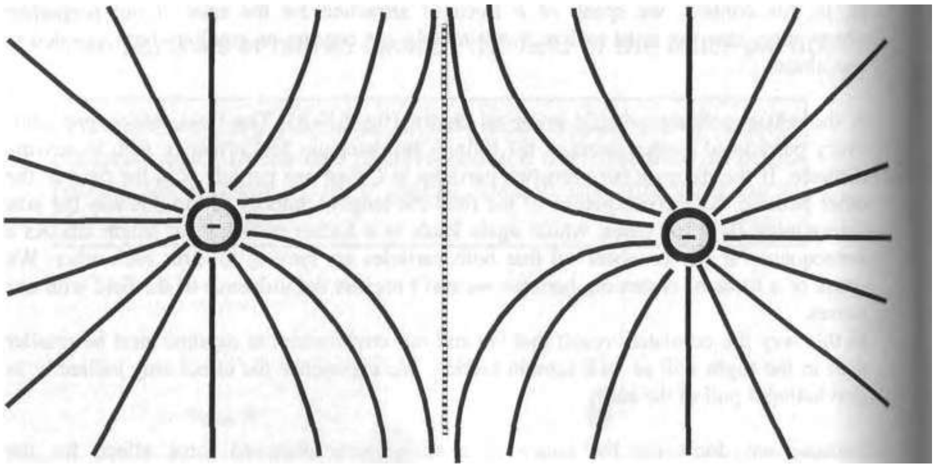

B: The field lines of the E-field for equal charged particles

The electromagnetic interaction of a particle is a result of the influence of the open field lines arising from it on the dimensions of space.

Fig. 6.8: The influence of the open field lines of the E-field

6.8 Electromagnetic interaction

A convincing answer to the open question provides us the analysis of the course of the field lines, on the one hand for charged particles and on the other hand for uncharged particles, which do not participate in the electromagnetic interaction. If at first we consider electrically charged particles, like e.g. electrons, protons or ions. Then all in common is that the towards infinity running field lines of the electric field are open. With this field the particle is able to interact with its environment. We measure a charge and an electromagnetic force effect. In the case of unequal charges, as is well-known, a field amplification and attractive acting forces are observed whereas for equal charges a field reduction results and repulsion is observed.

If we make a connection between the field conditions and the electromagnetic interaction in the sense of the proportionality (6.15), then the particle in reality is able to influence the distance to other particles merely with the help of its electric field. For unequal charges a compression of field lines arises, in which one particle stays in the focussed field of the other and vice versa. In this way a contraction of all lengths occurs and the observable attraction happens (fig. 6.8 A).
For equal charges the opposite case is present, in which even a local field freedom can occur (fig. 6.8 B). If the field tends towards zero on the dashed line, then the distance will go to infinity (according to eq. 6.15). Consequently, the observable effect that both bodies go away from each other, will reach to infinity.

Actually the electromagnetic interaction proves to be a result of the field dependent length contraction.

The electromagnetic interaction of a particle is a result of the influence of the open field lines arising from it on the dimensions of the space, in which it is.

It is important that the field lines are open, for which reason they are bent away from like charges and are directed towards unlike charges. Subjectively seen we find out that as a consequence of the field reduction repulsive force effects and as a consequence of the field compression attractive acting force effects are observed (fig. 6.8).
The consequence of is every electric field is, as is well-known, a magnetic field standing perpendicular on it. The field lines of the magnetic field run parallel to the surface of the particle and have a closed course (fig. 6.9 A)!
Therefore no magnetic poles form, which would be measurable. Seen from the outside the particle behaves neutral magnetically seen, because of the closed course of the field lines. An artificial field reduction and as a consequence observable forces of repulsion, like in the case of the electromagnetic interaction, hence in principle are impossible.

The effect of the magnetic field thus is limited to a geometrical manipulation of the environment, namely the curvature of space, with which we have founded the phenomenon of the attraction of masses and of the gravitation.

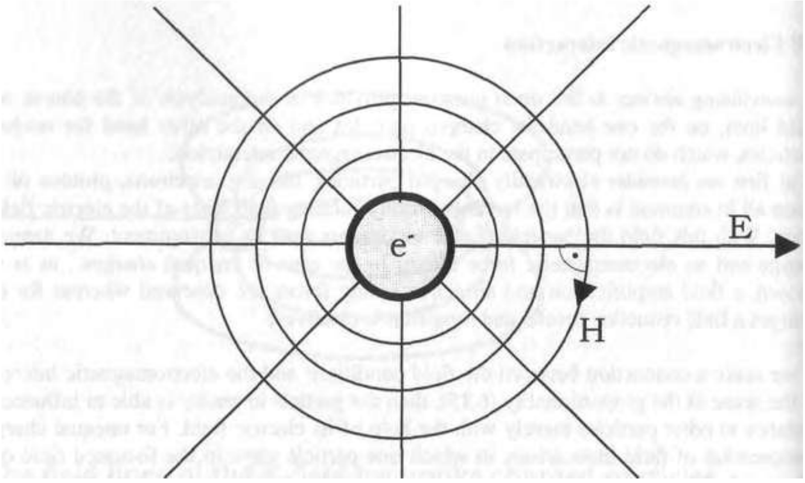

A: The open field lines of the E-field and the closed field lines of the H-field of an electrically charged particle (e.g. e-)

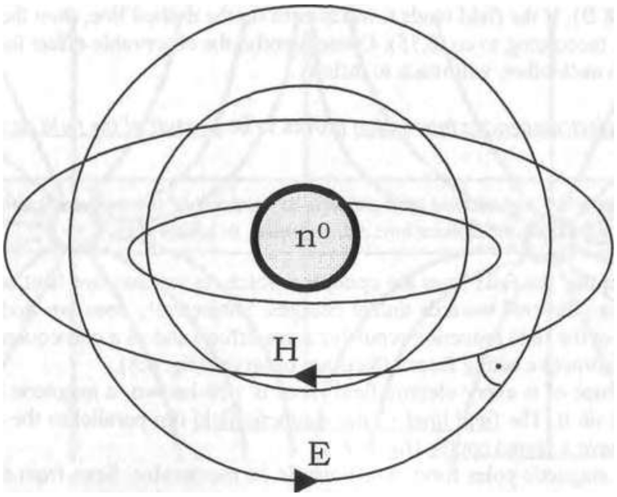

B: The closed field lines of the E-field and H-field of an electrically uncharged particle (e.g. of the neutron n°).

Gravitation is a result of the influence of the field lines with a closed course running parallel to the surface of the particles on the dimensions of the space, in which they are.

Fig. 6.9: The influence of the closed field lines of the H-field.

6.9 Gravitation

For uncharged, neutral particles (neutron, atom, molecule etc.) both the magnetic and the perpendicular on them standing electric field lines have a closed course. Now both run parallel to the surface of the particle (fig. 6.9 B).

As is said, the density of field lines with a closed course can't be influenced from the outside. If we approach a particle, the consequence of an increase of the density without exception is a decrease of the linear measures and thus a larger force of attraction. For this case of field lines with a closed course, for which in general it doesn't give a field attenuation and no forces of repulsion, there holds:

Gravitation is a result of the influence of the field lines with a closed course running parallel to the surface of the particles on the dimensions of the space, in which they are.

Both interactions logically have an infinite range. Both form a whole in the influence of the fields on the size conditions.

It surely is of the greatest importance that for this derivation of the field dependency of the Lorentz contraction from the known equations of transformation of the electromagnetic field we could do completely without the introduction of new factors of description or neglects.

Solely by consistent derivation and interpretation of the result the unification already has suceeded and the electromagnetic interaction and the gravitation could, with the derived field dependent Lorentz contraction, be traced back to a single basic phenomenon. Doing so we have to pay attention to the fact that the observer is subjected to the same Lorentz contraction as his measuring technique and therefore he can't see the field dependency at all. Merely as being an exterior observer it in rare cases will be possible to him to see the curvature of space in the presence of strong fields.

From this for an astronaut practical consequences result. If he namely would land on Jupiter, he would think flat hills to be gigantic mountains, that small he would be! Vice versa if he landed on the moon, high mountains would appear to be insignificant hills, not because of wrong altitude readings of the terrestrial mission control and measurement centre, but only because of his own body size. The astronauts of the Apollo missions were not prepared for this circumstance and after their landing on the moon were completely surprised, how little validity learned textbook physics has, hardly has one left the earth. They have brought photographs with them which prove the Lorentz contraction to depend on the field and therefore on gravitation.

The fact that force effects should arise from the interactions is an auxiliary concept and auxiliary description of the observing person founded in pure usefulness. The Lorentz force therefore shouldn't be regarded as cause anymore. It actually appears only as property of the field factors. Seen this way it only would be consistent to do without space charges and currents as a result of moving charges and to assume a source-free and quanta-free field description (fig. 6.4: $j = 0$).

From an unified theory it is demanded that it besides the electromagnetic interaction and the gravitation also is able to integrate the strong and the weak interaction. We will also solve this problem.

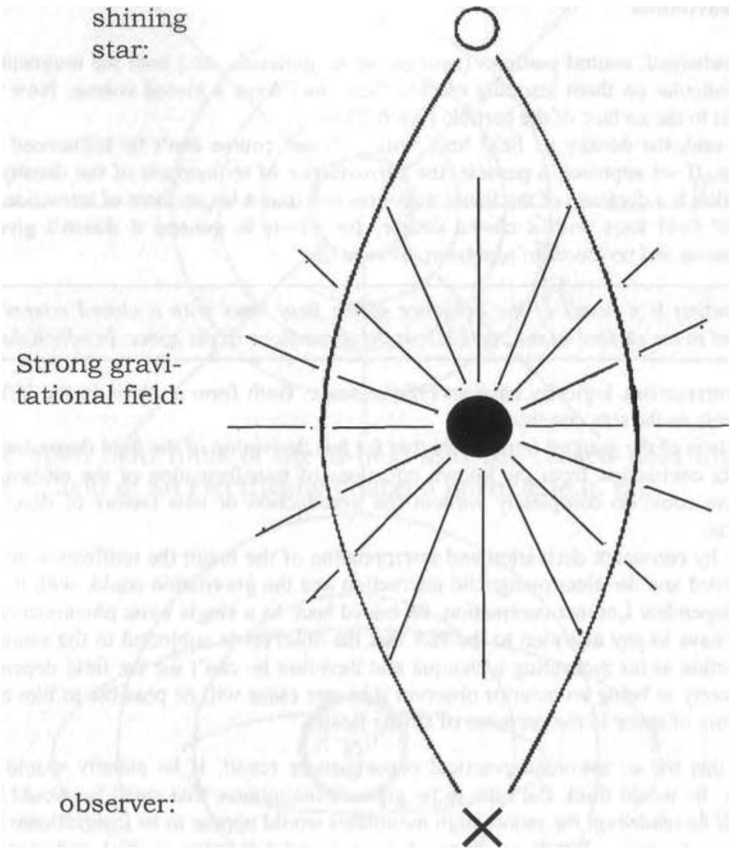
shining
star:
Strong gravi-
tational field:
observer:

Fig. 6.10: Diversion of the light by a strong gravitational field.

Speed of light of the wave: $c = \lambda * f$ (6.16)

For the wavelength λ holds (because of eq. 6.15): $E, H \sim 1/\lambda^2$

From equation (6.16) follows (with f = constant):

$$E \sim 1/c^2 \quad , \quad H \sim 1/c^2 \qquad (6.17)$$

The speed of light depends on the field!

6.10 Field dependent speed of light

But not only matter is bent towards a gravitational field. If we only think of the much cited phenomenon that the ray of light of a star is diverted towards the sun, if it passes very close to the sun on its way to us, like this has been observed for the first time during an eclipse of the sun in 1919 (fig. 6.10).
Quite obviously the field of the sun also slows down the speed of light. On the side of the ray of light which is turned towards the sun, the field is somewhat larger and the speed of light correspondingly is slower than on the side which is turned away, and with that the ray of light changes its direction in the observable manner. Exactly this relation willingly is interpreted as a consequence of a curvature of space.
The extremely strong field of a black hole can divert the light down to a circular path, in order to in this way catch and bind it. The light now orbits the black hole like planets the sun.

At this point the open-minded reader already might have tapped the confirmation of the proportionality 6.2 ($c \sim r$), which has been derived from the vortex model (fig. 6.2).
The sceptic is offered still another derivation: for the borderline case that the relative velocity v tends towards the speed of light c (fig. 6.6), according to equation 6.13 the measurable overall field E_o (and also H_o) will go to zero and equation 6.12, with $E_z. = - E$ (and $H_z = - H$), will again turn into the wave equation (5.9*) after double differentiation (fig. 6.4).
The speed $v = c$ so to speak forms the escape velocity, with which the electromagnetic wave runs away from the cosmic field. Under these circumstances of course neither an attraction of masses nor an electromagnetic interaction can be exerted on the wave.
If E_0 goes to zero at the same time l_0 tends to infinity (equation 6.15, fig. 6.6): i.e. the wave spreads all through space. This result entirely corresponds to the observations and experiences.

For the wave length λ and in the end for the velocity of propagation c only the self-field of the wave E resp. H is responsible. Because of

$$\boxed{c = \lambda \cdot f} \quad (6.16)$$

and the proportionality from equation 6.15: $E, H \sim 1/\lambda^2$ (6.17*)

obtain the new relation: $\boxed{E, H \sim 1/c^2}$ (6.17)

If the speed of light in the presence of matter decreases, then we now also know why. It is the field, which surrounds matter, that slows down the speed of light. Therefore a gravitational field is able to divert a ray of light in the same manner as matter which flies past. Finally moves the speed of light in the proportionality 6.17 to the place of the linear measure (in 6.15).
But if the rule fails one will try to replace by an optical measurement arrangement. In this manner the field dependency of the Lorentz contraction should be measurable; but it isn't!

From the comparison of the derived proportionalities:

$$E, H \sim \overset{(6.15):}{\frac{1}{l^2}} \sim \overset{(6.17)}{\frac{1}{c^2}}$$

follows:

$$l \sim c \qquad (6.18)$$

The speed of light is proportional to the measurement path.

The variable speed of light is being measured with itself.

The result at all events is a constant value.

The constancy of the speed of light is based on a measurement which is faulty from the principle!

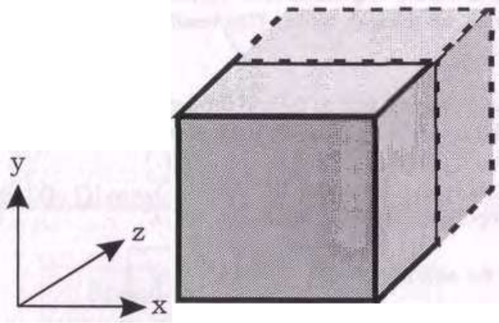

Because of c ~ r: physical length contraction

Fig. 6.11: Derivation of the length contraction (field dependent Lorentz contraction)

6.11 Universality

Why can't the rule be replaced by an optical measurement arrangement? The crucial indication provides the comparison of both derived proportionalities 6.15 and 6.17. According to them holds the same field dependency for both the Lorentz contraction and the speed of light:

$1/l^2 \sim 1/c^2$ or $1 \sim c$ (6.18)

If the rule has proved to be useless, then we will experience the same disaster if we measure optically, i.e. with the speed of light.

Obviously both, the length 1 and the speed of light c as a length per unit of time, depend in the same manner on the respective local field strength. On the one hand do both measuring methods lead to the same result; but on the other hand will anything which can't be measured with one method, neither be measured with the other method.

If now the speed of light is being measured optically, then the measurement path will be proportional to the speed of light and as a result will the unknown factor be measured with itself. The result of this measurement, which is faulty from the principle, at all events is a constant value, because here two variables which stand in direct proportionality to each other are related to each other.

Was the famous experiment of Michelson and Morley unnecessary, the result trivial? And how does it stand about the postulate of the universality of the speed of light?

If we for that consider a cube (fig. 6.11). And we assume that the speed of light is a vectorial quantity, which in our experiment is for instance in one direction twice as large, as in the direction of the other two space axes. By means of the mentioned influence of the speed of light on the spatial length is, as a consistent consequence, the cube along this edge pulled apart to a cuboid. We however register this spatial body with our eyes, which is with the speed of light and that has increased proportionally to the length of the edges, for which reason we as subjective observer still see a cube in front of us and not a cuboid.

If we trust an apparent objective measurement more than our sense organ and measure the three lengths of the edges of the cuboid again with a rule then we get three times the same length, which is a cube.

We probably are dealing with an optical deception using the true meaning of the word.

If the by Einstein postulated universality and constancy of the speed of light in reality doesn't exist at all, we in no way would be capable to register this; neither to observe nor to measure it!

The Galilean theorem of the addition of speeds objectively seen still is valid, even if the fact that the speed of light apparently is independent of the speed of the source pretends us the opposite.

If for instance a light source is moved towards a receiving device or away from it, then the speeds will overlap, like for the passenger, who marches in a driving train against or in the driving direction through the corridor. For the ray of light also the fields, which influence the speed of light and the measurement equipment, overlap. As a consequence will a measuring technician, who himself is exposed to this overlapping field, always observe and "measure" the identical speed of light. The observer as a result imagines, there is an universality of the speed of light.

The field takes over the function of the aether.

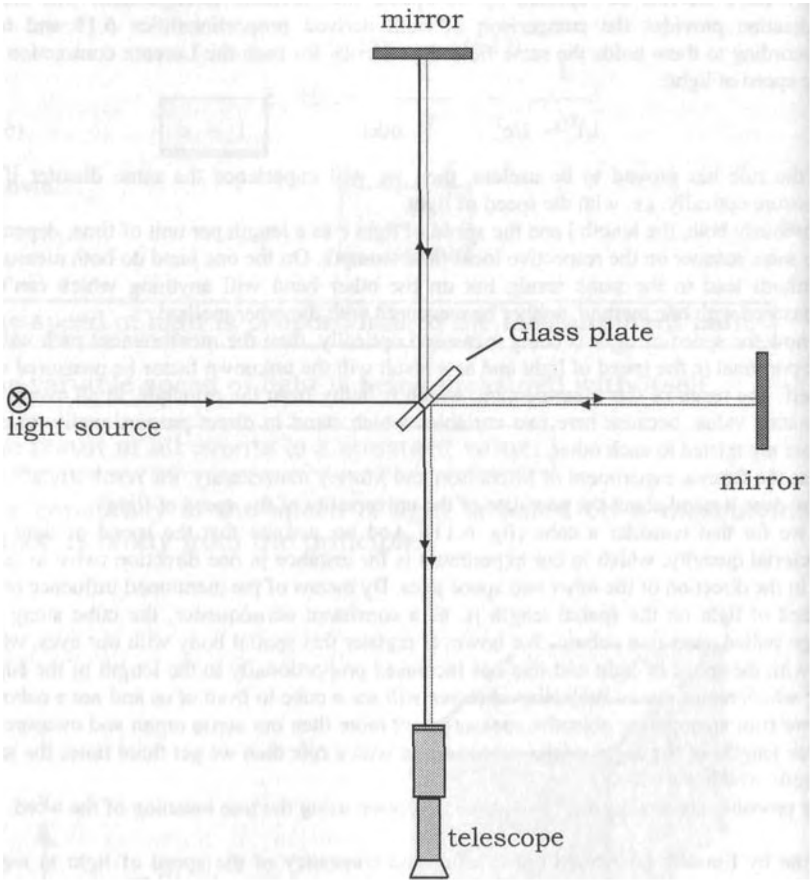

Fig. 6.12: Experiment of Michelson and Morley to detect an aetherwind[i]

<i>: A.P.French: Special Relativity, Massachusetts Institute of Technology, 1966.

<ii>: Nikola Tesla: "This is the same as writing a business letter and forgetting the subject you wish to write about". To Einstein's Theories, Rare Book and Manuscript Library, Columbia University, 15.4.1932.

<iii>: Einstein proceeds in the same manner with the time dilatation, by assuming a time constant by definition for the derivation to present at the end of his derivation a variable time. And with that he presents a result which contradicts his approach completely.

6.12 Aether

Important experiments like the one of Doppler concerning the redshift or the one of Bradley concerning the aberration of the stars show only to clear, where the influence of the speed of light subjectively still is perceptible, or for laboratory experiments like the one of Michelson and Morley, where the influence isn't perceptible anymore, because the length of the interferometers always changes proportionally to the speed of light. The look in the stars at the same time is a look in cosmic areas, where completely other field conditions prevail and as a consequence completely other values for the speed of light and for the dimensions of space are present. The mentioned observations suggest that we together with our measuring station are moving through the cosmos and therefore a relative velocity has to be present with regard to an aether which determines the respective speed of light.

If we however constrict our range of vision and retire in a laboratory, then we no longer are capable to observe the influence of the field on the speed of light. The experiments of Michelson which Maxwell had prompted to and which Morley with a higher precision had repeated with the goal, to detect the aether, inevitably had to turn out negatively. The laboratory experiments resulted in the misleading picture, as if the earth was resting in the aether.

The not understood measurements will suggest any observer, he forms the centre of the universe and everything rotates around him, entirely in the sense of the Ptolemean view of life, which, although long ago abolished, here belated has experienced support. With a Swabia caper Albert Einstein has prevented a relapse into the dark Middle Ages and removed the open contradiction in the question of the aether, which once is measured as moving and another time as resting, by without further ado abolishing the aether. With that he undoubtedly has solved a central problem of physics and at the same time created a new one. As is known does the speed of light have a certain value, and therefore the question is raised, what determines is size. Exactly for this purpose a luminiferous aether had been introduced, however it is constituted.

Scientifically it does make little sense, to make an assumption, if at the end of the derivation the prerequisite is deleted without substitute. In such a case either in the approach or in the derivation is a principal error*[1]"*. Nikola Tesla comments on the working method of Einstein with the applicable comparison, as if Einstein had, while he was writing a business letter, forgotten completely the subject he wanted to write about (fig. 6.12[ii]).

The answer, which removes all contradictions and is entirely in accord with all observations and measurements, is obvious. Naturally a luminiferous aether exists, which determines the velocity of propagation and of course it by no means is bound to the observer.

As has been derived in figures 6.5 and 6.6, will for a relative velocity v arise a field, which according to proportionality 6.17 determines the speed of light. With that we have derived completely.

The field takes over the function of the aether.

The equations 6.10 also answer the question, why no aetherwind is being observed, although such a wind actually is present: we experience, as we have discovered, an E-field with „head wind" as a resting H-field and vice versa and therefore we aren't capable to detect the head wind in the aether!

Key questions of quantum physics (fig. 4.4 + continuation):

IV. Why do the particles have the form of spheres?

(with increasing E-field decreases c)

VIII. Why is the elementary quantum localized?

(in the vortex centre: c = 0, see figures 4.3 and 6.2)

IX. Why do the elementary particles have a spin?

(spherical form demands field compensation)

X. Why is the magnitude of the spin quantized?

(cosmic basic field determines the need of E_z)

XI. Why can speeds faster than light occur in a tunnel?

(a reduction of the cosmic basic field can only be realized locally in a tunnel)

to XI:

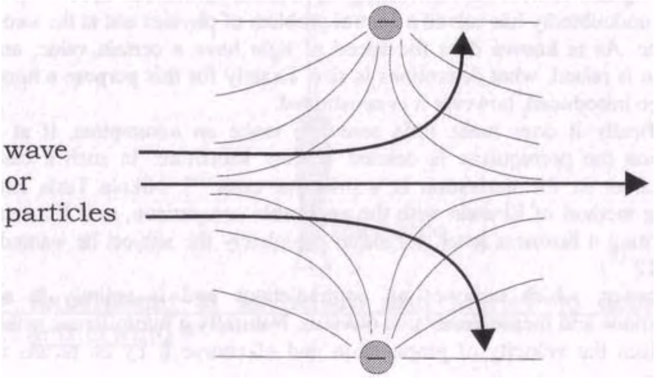

Fig. 6.13: Consequences concerning the field dependency of the speed of light: spin effect and tunnel effect <i,ii>

<i>: Nimtz,G.: Instantanes Tunneln, Tunnelexperimente mit elektromagnetischen Wellen, Phys.B1.49, VCH Weinheim (1993) Nr.12, S. 1119-1120<*>
<ii>: Thoma, P., Weiland.T.: Wie real ist das Instantane Tunneln? Phys.Bl.50, VCH Weinheim (1994) Nr.4, S. 359-361<*>
<*>: The measurement results are in accord with the theory of objectivity, not however the contradictory attempts to interpret them <i> and <ii> et al.

6.13 Spin and tunnel effect

Only with the field dependency of the speed of light (6.17) we can understand, why the elementary quanta can form as spheres, like is drawn in the figs 4.3 and 6.2. In the centre the field lines run together, i.e. the field increases and the speed of light decreases. Only in this way it will be possible for the vortex oscillation to everywhere occur with the speed of light, even in the inside of the particle! In the centre of the vortex particle the field in theory will become infinitely large and the speed of light zero. This circumstance again is the foundation why the elementary particles are localized and it answers key question VIII of quantum physics. The absence of a speed after all is the characteristic of an immobile thing.

The field dependency of the speed of light answers also further basic and up to today unanswered key questions of quantum physics, like why the elementary particles have a spin (IX) and why the magnitude of the spin is quantized (X).

A vortex particle after all does not exist alone in the world, but it is in the field of other particles. We can call this the cosmic basic field (E resp. H). This basic field overlaps the self-field and takes effect the strongest in the area of the spherical shell, where the self-field is correspondingly small. In order to keep the form of a sphere, this influence of the basic field has to be compensated. The additional field (E_z resp. H_z according to eq. 6.12) necessary for the compensation is produced by the particle, by rotating in a spiral around itself with a speed v which increases towards the outside of the spherical shell. Therefore does the elementary particles have a spin. The electron spin is therefore determined by the cosmic basic field.

Another effect of the field dependent speed of light is the tunnel effect. As an example we consider the two differently charged particles shown in fig. 6.8 A. The open, outside of the particles running, field lines of the electric field are predominantly bent towards the each time oppositely charged particle. If another particle wants to pass between the two, then it gets into an area of increased field strength. As a consequence it will be slowed down, because here a smaller speed of light is present.

Water molecules show with their polar nature exactly this property. Water has a remarkably high dielectricity e and slows down the speed of light correspondingly according to equation 5.6 ($\varepsilon \cdot \mu = 1/c^2$). The refraction of light at the water surface is an observable result of the reduced speed of light in the presence of matter.

If we now examine the case in which the two particles have the same charge as is shown in fig. 6.8 B (and fig. 6.13 belonging to XI). The field lines repel each other, so that exactly in between the two particles a field free area forms, in which the speed of light goes to infinity! This area acts like a tunnel. If we send through a particle exactly here, then purely theoretically seen it won't need any time to run through the tunnel, and for a short time the signal becomes infinitely fast.

If a particle hits only slightly besides the tunnel, then it will one-sidedly be slowed down and diverted by the respective field. We call this process reflection or scattering. Only the few particles, which exactly hit the tunnel, arrive behind the hurdle and in the ideal case even almost without loss of time!

The current measurements of speeds faster than light demonstrate in a convincing manner the superiority of the field-theoretical approach with regard to the nowadays normally used quantum physical approach.

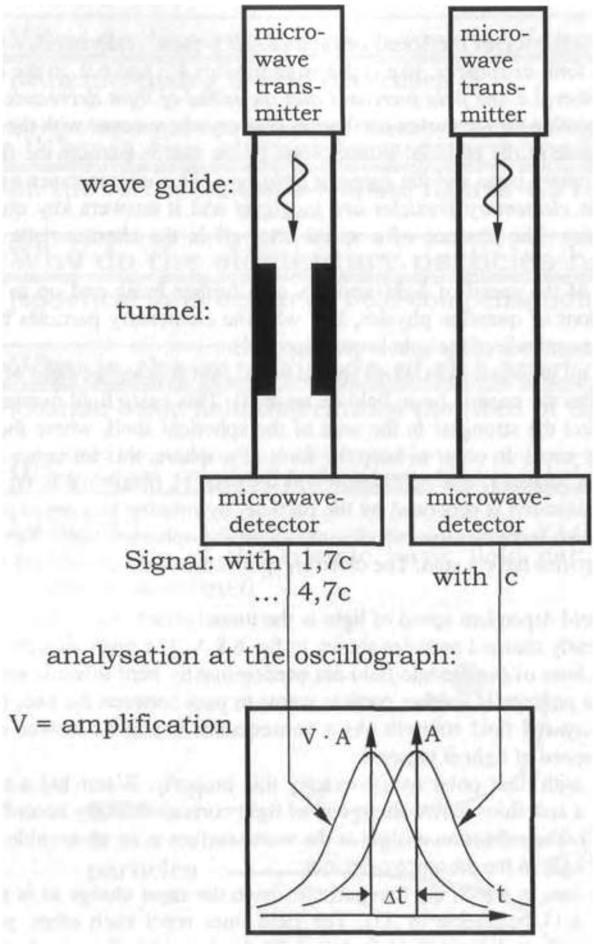
micro-
wave
trans-
mitter
micro-
wave
trans-
mitter
wave guide:
tunnel:
microwave-
detector
microwave-
detector
Signal: with 1,7c
... 4,7c
with c
analysation at the oscillograph:
V = amplification
V · A
A
Δt
t

Fig. 6.14: The microwave experiment at the II. Physical Institute of the University of Cologne to measure speeds faster than light.<i>

<i>: Nimtz, G. : New Knowledge of Tunneling from Photonic Experiments, Proc. of the Adriatico Research Conference, 1996, World Scientific Publishing Company

6.14 Interpretation of the measured speed faster than light

Now the attempt can be undertaken, to interpret the spectacular experiments, in which a speed faster than light has been measured. It is reported<i> that in experiments with photons at the University of California in Berkeley on an average a speed of 1.7 times the speed of light has been measured by Prof. Raymond Chiao and his co-workers. At the Technical University of Vienna Prof. Dr. Ferenc Krausz already has obtained 2.4 times the, according to Einstein at maximum obtainable, speed of light with tunnelling laser light.

The first measurements of speeds faster than light have been carried out with microwaves at the University of Cologne<ii> by Prof. Dr. Gunter Nimtz and co-workers. They at first had published the measurement of a speed 2.5 times the speed of light. In the meantime they even have transmitted a symphony of Mozart with a speed almost 10 times the speed of light and with that have contradicted Einstein's hypothesis, according to which the speed of light in vacuum would be the highest possible speed for the transmission of signals. The different experiments only resemble each other in the point that the particles have to tunnel, because one has put a barrier in their way. This "tunnelling" apparently is the cause for obtaining speeds faster than light. With the prevailing physical view of life these measurement results are incompatible.

In the measurement set up in Cologne the microwaves are sent through a wave guide, which they pass with the speed of light. If a parts with narrowed cross-section is inserted, where the microwaves actually don't fit through at all, then the signal gets damped strongly. Now however arrives nevertheless a small part of the signal at the other end of the wire, but much faster than allowed, namely with the measurable speed faster than light.

The answer of the here presented potential vortex theory reads as follows: the waves picked up by the wave guide run up to the entry of the tunnel, in order to find out that they don't fit through. They are reflected or absorbed. A small part however rolls up to potential vortices and these fit through the tunnel. They however have to be compressed additionally. In the derivation of the photon (fig. 4.5 and 4.6) we had seen that the inner vortex always is faster than the bigger one, through which it slips through. The compression therefore causes an increase in speed. In flow dynamics is known an analogy: the Venturi-tube. The flow-technical potential vortices also confirm exactly this property. One can as well start with the Lorentz contraction (fig. 6.6, eq. 6.14*). This states that a particle moving with a higher speed actually becomes smaller and not only appears to be smaller as an optical deception of the observer. Because only smaller particles fit through the tunnel, the particles, measurable at the other end, must be correspondingly faster: quod erat demonstrandum. In the same manner also the experiments of Berkeley can be explained physically, because here is worked with photons from the start<i>. With that the process of rolling up the wave can be left out. The tunnel lets pass only compressed and therefore faster light particles.

<i>: R.Y.Chiao, P.G.Kwiat, A.M.Steinberg: Schneller als Licht? Spektrum der Wiss. 10/93

<ii>: B. Schuh, Gespenstisch fixe Wellen, DIE ZEIT Nr. 45, 5.11.93, S. 43.

<ii>: Enders, A., Nimtz, G.: Photonic-tunneling experiments, Physical Review B, Vol. 47, No. 15 (1993), pp. 96O5-96O9.

<ii>: Enders, A., Nimtz, G.: Evanescent-mode propagation and quantum tunneling, Physical Review E, Vol. 48, No.l (1993), pp. 632-633.

"The theory of relativity is not a physical theory... it is a mathematical poetic idea, a deduction from impossible premises."<i>
Oskar Kraus

"The theory of relativity is a mathematical masquerade, behind which is hidden an inextricable ball of a mixing up of ideas, contradictions, fallacies, arbitrary assumptions and ignoring of healthy logic."<i>
Erich Ruckhaber

"The theory of relativity not only is fantastic, but also of an inconsistency which in the history of science not yet has been present."<i>
Harald Nordenson

"A physics of hybrids, of contradictions and fantastic confusions, nonsense!"<ii>
Johann Marinsek

"This is absurd."<iii> (regarding mass-energy interpretation) ________
Nikola Tesla

"In my experiments I have destroyed billions of atoms, without having observed any emissions of energy. "<4i>
Nikola Tesla

Fig. 6.15: Some statements regarding the theory of relativity.

<i>: Walter Theimer: Die Relativitatstheorie, Seite 7, Francke Verlag, Bern, 1977, ISBN 3-772O-126O-4
<ii>: Johann Marinsek: Rationale Physik, S. 163, dbv-Verlag TU Graz, 1989, ISBN 3-7O41-O176-1
<iii>: Nikola Tesla, To Einstein's Theories, Rare Book and Manuscript Library, Columbia University, 15.4.1932. Entnommen aus J.T.Ratzlaff: Tesla Said, Tesla Book Company, pp. 238, ISBN O-914119-OO-1
<4i>: Nikola Tesla, Franz Ferzak World and Space Publications 1985.

6.15 Definition of the speed of light

If a light signal propagates in space, then as a consequence of the velocity of propagation c, it at a certain point in time t is in a distance r of the light source:

$r = c * t$ (6.19)

Should the speed of light become smaller for instance by Δc, then the light signal obviously has covered a distance less by Ar or the time interval has changed by Δt:

$r + \Delta r = (c + \Delta c) \cdot (t + \Delta t)$ (6.20)

This equation describes purely mathematically the most general case which can be assumed. By writing out the multiplication and subtraction of equation 6.18 the change in distance considered for itself is:

$\Delta r = c \cdot \Delta t + t \cdot \Delta c + \Delta c \cdot \Delta t$ (6.21)

The answer of mathematics is that the change in distance can have its cause in a change in time, in a change of speed or in both. We now want to turn to the physical interpretation and have a closer look at the two possibilities, in which either c or t is to be taken constant (see fig. 6.16).

In the first case the speed of light c is constant and as a consequence the change Δc = zero. The mathematical formulation (according to eq. 6.21) therefore reads:

case 1: $\boxed{c = \frac{\Delta r}{\Delta t}}$ (relativity) (6.22)

If in this conception world a change in distance is observed, for instance the Lorentz contraction, then in order to save this relation inevitably a change in time, for instance a time dilatation, has to make the compensation. Einstein in an applicable manner speaks of relativity, because according to his opinion in the case of both variables, the length contraction and the time dilatation, it only concerns observed changes.

For the time dilatation experiments are given. But for the measurement of time always only atomic clocks are available and their speed of running of course could also be influenced by the Lorentz contraction. In any case it can't be claimed the time dilatation is proven experimentally as long as we do not know the mechanisms of decay of atoms. Otherwise the statements of the theory of relativity are familiar to us, for which reason further remarks seem unnecessary.

In the second case the time t is constant and consequently the change At = zero. At a closer look this case is much more obvious, since why should time change. After all time has been stipulated by definition.

After all, we are the ones who tell, what simultaneity is!

The mathematical formulation for this case reads (eq. 6.21 with Δt= 0):

case 2: $\boxed{\Delta c = \frac{\Delta r}{t}}$ (objectivity) (6.23)

This equation does open up for us an until now completely unknown and fundamentally other way of looking at the physical reality.

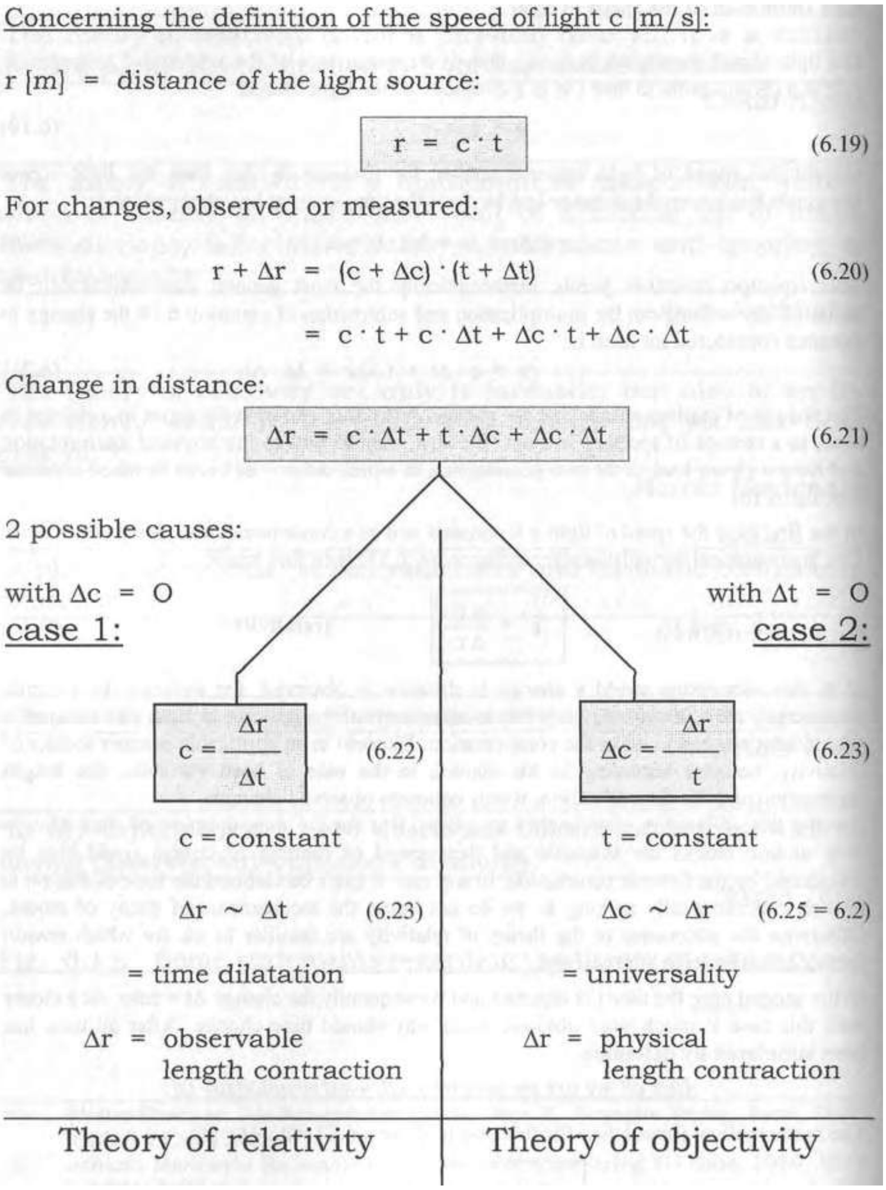

Concerning the definition of the speed of light c [m/s]:

r [m] = distance of the light source:

$$\boxed{r = c \cdot t} \qquad (6.19)$$

For changes, observed or measured:

$$r + \Delta r = (c + \Delta c) \cdot (t + \Delta t) \qquad (6.20)$$

$$= c \cdot t + c \cdot \Delta t + \Delta c \cdot t + \Delta c \cdot \Delta t$$

Change in distance:

$$\boxed{\Delta r = c \cdot \Delta t + t \cdot \Delta c + \Delta c \cdot \Delta t} \qquad (6.21)$$

2 possible causes:

with $\Delta c = 0$ case 1:	with $\Delta t = 0$ case 2:
$\boxed{c = \frac{\Delta r}{\Delta t}}$ (6.22)	$\boxed{\Delta c = \frac{\Delta r}{t}}$ (6.23)
c = constant	t = constant
$\Delta r \sim \Delta t$ (6.23)	$\Delta c \sim \Delta r$ (6.25 = 6.2
= time dilatation	= universality
Δr = observable length contraction	Δr = physical length contraction
Theory of relativity	Theory of objectivity

Fig. 6.16: Theory of relativity and theory of objectivity, derivation and comparison.

6.16 Relativity and objectivity

New to the second case (equation 6.23) is particularly the proportionality contained in it:

$$\boxed{\Delta c \sim \Delta r} \quad (6.25 = 6.2)$$

But to us it is not new, because we have derived the same proportionality from the model conept (equation 6.2, fig. 6.2), in which the elementary particles are understood as spherical vortices. Equantion 6.25 unconcealed brings to knowledge that any change of the speed of light c [m/s] in the same way leads to a change of the radius r [m], the distance between two points in space or even the length of an object, e.g. a rule. Such a rule after all consists of nothing but spherical atoms and elementary particles and for their radius r again the proportionality 6.25 holds. Therefore it is to be set:

$$r \sim 1 \qquad (6.26)$$

and taken both together we already had derived as equation 6.18 (fig. 6.11) from the field dependency. Here the vortex model as well finds a confirmation of its correctness, as in the derivation from the equations of transformation of the electromagnetic field. Because all three, the derivation according to the model, the physical and the mathematical derivation, lead to the same result, this second case should be called "objective".

With that the first case, which describes the subjective perception of an observer, is not supposed to be devaluated. It contains the definition of reality, according to which only is real what also is perceptible. The theory of relativity of Poincare and Einstein is based on this definition.

With the second case, the case with a variable speed of light, we however get serious problems, since we observe with our eyes, and that works with the speed of light. If that changes, we can't see it, as already said. If we could see it, then "reality" would have a completely different face and we surely would have great difficulties, to find our way around. In this "objective world" neither electromagnetic interactions nor gravitation would exist, so no force effects at all. Because all distances and linear measures depend on the speed of light, everything would look like in a distortion mirror.
The concept of an "objective world" at first has not a practical, but rather a theoretical and mathematical sense. The distinction between an observation domain and a model domain is founded in pure usefulness.

The observation domain should correspond to case 1 and the model domain to case 2. The mathematical derivation tells us, how we can mediate between both domains (equation 6.21): This mediation amounts to a transformation, which provides us the instruction, how a transition from the observation into a not perceptible model concept, from the relativity into an objectivity has to.

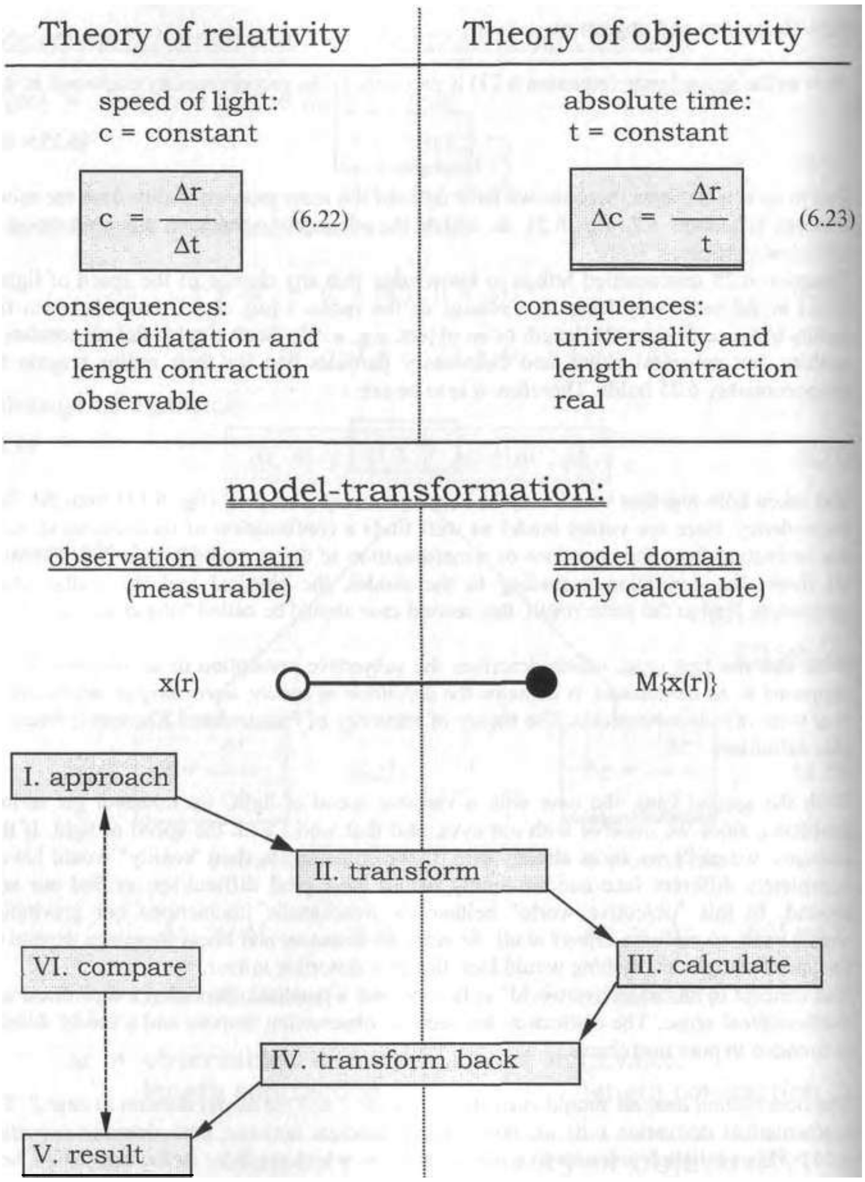
Theory of relativity
Theory of objectivity
speed of light:
c = constant
absolute time:
t = constant
c = Δr / Δt (6.22)
Δc = Δr / t (6.23)
consequences:
time dilatation and
length contraction
observable
consequences:
universality and
length contraction
real
model-transformation:
observation domain
(measurable)
model domain
(only calculable)
x(r)
M{x(r)}
I. approach
II. transform
III. calculate
IV. transform back
V. result
VI. compare

Fig. 6.17: Model-transformation between theory of relativity and theory of objectivity.

6. 17 Transformation

The observation domain is, as the name already expresses, perceptible (observable) with the help of our sense organs and measurable with corresponding apparatus. The special theory of relativity for the most part provides us the mathematics needed for that. And in that is assumed a constant speed of light. Because a length contraction is being observed and can be measured, a time dilatation must arise as a consequence. Such is the consistent statement of this theory. Because we already could make us clear that it concerns a subjective theory, of course caution is advisable if generalizations are being made, like the one of the inductive conclusion of the length contraction on the time dilatation. We'll come to speak about that in this chapter (fig. 6.20).

The model domain however is not observable to us and only accessible in a mathematical manner. Here the time is a constant. On the other hand do the radii of the particles and all other distances and linear measures stand in direct proportionality to the speed of light. If that changes, then does that lead to a change in length. The length contraction occurs physically, which means actually. We propose the name "theory of objectivity" for the valid theory which is derivable with this prerequisite and independent of the point of view of the observer.

The importance of this model domain and of the possible model calculations is founded in the circumstance that many physical relations within our observation domain aren't recognized by us and can't be mathematically derived. Besides is only all to often worked with unallowed generalizations and with pure hypotheses. Such a thing does not even exist in the model domain.

The model domain can be tapped over a transformation. For that we select an approach x(r) in the to us accessible observation domain. This then is transformed into the model domain by a calculation instruction M{x(r)}. Here we can calculate the sought-for relation In the usual manner and transform back again the result according to the same calculation instruction $M^{-1}\{x(r)\}$ but in the reversed direction. After being returned in our familiar observation domain, the result can be compared and checked with measurement results (fig. 6.17).
In this way we will derive, calculate and compare the quantum properties of the elementary particles with the known measurement values. Here we remind you of the fact that all attempts to calculate the quantum properties conventionally, without transformation, until now have failed. Not even a systematization may succeed, if it concerns for instance explanations for the order of magnitude of the mass of a particle.

A transformation at first is nothing more than an in usefulness founded mathematical measure. But if a constant of nature, and as such the quantum properties of elementary particles until now have to be seen, for the first time can be derived and calculated with a transformation then this measure with that also gains its physical authorization. We now stand for the question: how does the instruction of transformation M{x(r)} read, with which we should transform the approach and all equations from the observation domain into the model domain?

general relations:	special theory of relativity	theory of objectivity
domain of validity:	observation domain	model domain
speed of light c [m/s]:	$c = c_0$ $= \text{constant}$	$c \sim r$ (6.25)
field strengths (6.15)	(6.15) + (6.18)	
H [A/m] :	$H \sim 1/r^2$	$H \sim 1/r$ (6.27*)
E [V/m] :	$E \sim 1/r^2$	$E \sim 1/r$ (6.27)
because of eq. (5.6) $\varepsilon \cdot \mu = 1/c^2$ is valid:		
μ [Vs/Am] :	$\mu_0 = \text{const.}$	$\mu \sim 1/r$ (6.28*)
ε [As/Vm] :	$\varepsilon_0 = \text{const.}$	$\varepsilon \sim 1/r$ (6.28)
relations of material:		
$B = \mu \cdot H$ (3.5): B [Vs/m^2]	$B \sim 1/r^2$	$B \sim 1/r^2$ (6.29*)
$D = \varepsilon \cdot E$ (3.6): D [As/m^2]	$D \sim 1/r^2$	$D \sim 1/r^2$ (6.29)
e.g. spherical capacitor		
capacity: C [As/V]	$= \varepsilon 4\pi r$ (6.4)	$C = \text{const.}$ (6.30)
charge: Q [As]	$= C \cdot U$ (6.31)	$Q = \text{const.}$ (6.32)
energy: W [VAs]	$= Q^2/C$ (6.1)	$W = \text{const.}$ (6.33)
with energy-mass relation: (5.24) and (6.1)	$W = mc^2$	
mass	m [$kg = VAs^3/m^2$]	$m \sim 1/r^2$ (6.34)
relaxation time	τ_1 [s]:	$\tau_1 = \text{const.}$ (6.35)
with	$\tau_1 = \varepsilon/\sigma$ (5.3)	
specific conductivity	σ [A/Vm]:	$\sigma \sim 1/r$ (6.36)

Fig. 6.18: Transformation of the dependencies on radius

6.18 Transformation table

The attempt to write down at this point already a closed mathematical relation as instruction of transformation, would be pure speculation. Such an instruction first must be verified by means of numerous practical cases, i.e. be tested for its efficiency and correctness. But we not even know the practical examples necessary for this purpose, if we apply the transformation for the first time!
For his reason it unfortunately is not yet possible, to calculate absolute values in a direct
We have to be content to work with proportionalities and to carry out comparisons.
In fig. 6.18 the proportionalities are compared in the way, how they would have to be transformed: on the left side, how they appear and can be observed in the view of the special theory of relativity, and on the right side, how they can be represented and calculated in the theory of objectivity.
The change, which here would have to be transformed, is the physical length contraction, which is the change in length as it depends on the speed of light. For spherical symmetry the length 1 becomes the radius r (eq. 6.26), of which is to be investigated the influence. In the observation domain we had derived the proportionality (6.15 + 6.18):

$E \sim 1/r^2$ and $H \sim 1/r^2$.

The field of a point charge or of a spherical capacitor confirms this relation:

$$E = Q/\varepsilon 4\pi r^2 .$$

Because the speed of light in our observation is constant, also both constants of material which are related to it (eq.5.6: $\varepsilon \cdot \mu = 1/c^2$), the dielectricity ε and the permeability μ, are to be taken constant.
With that the same proportionality as for the field strengths also holds for the induction B and the dielectric displacement D:

$$B \sim 1/r^2 \quad \text{and} \quad D \sim 1/r^2 .$$

In the model domain everything looks completely different. Here the radius and any length stands in direct proportionality to the speed of light. In this way we get problems with our usual system of units, the M-K-S-A-system (Meter-Kilogram-Second-Ampere). The basic unit Meter [m] and as a consequence also the unit of mass Kilogram [kg = VAs^3/m^2] appear here as variable. It would be advantageous, to introduce instead the Volt [V] as basic unit.
But in any case does the dimension of a quantity show us, in which proportionality it stands to the unit of length. This in the model domain then is authoritative! As an example does the speed of light have the dimension Meter per Second. In the model domain there consequently has to exist a proportionality to the length r [m].
The speed of light determines with equation 5.6 again the constants of material:

$$\mu\,[Vs/Am] \sim 1/r \quad \text{and} \quad \varepsilon\,[As/Vm] \sim 1/r \qquad (6.28)$$

According to the model holds unchanged:

$$B\,[Vs/m^2] \sim 1/r^2 \quad \text{and} \quad D\,[As/m^2] \sim 1/r^2 . \qquad (6.29)$$

But if we insert the proportionalities 6.28 and 6.29 into the equations of material 3.5 and 3.6, then holds for the field strengths:

$$H\,[A/m] \sim 1/r \quad \text{and} \quad E\,[V/m] \sim 1/r . \qquad (6.27)$$

Further dependencies of the radius can be read in the same manner either by inserting into well-known laws or immediately from the dimension.

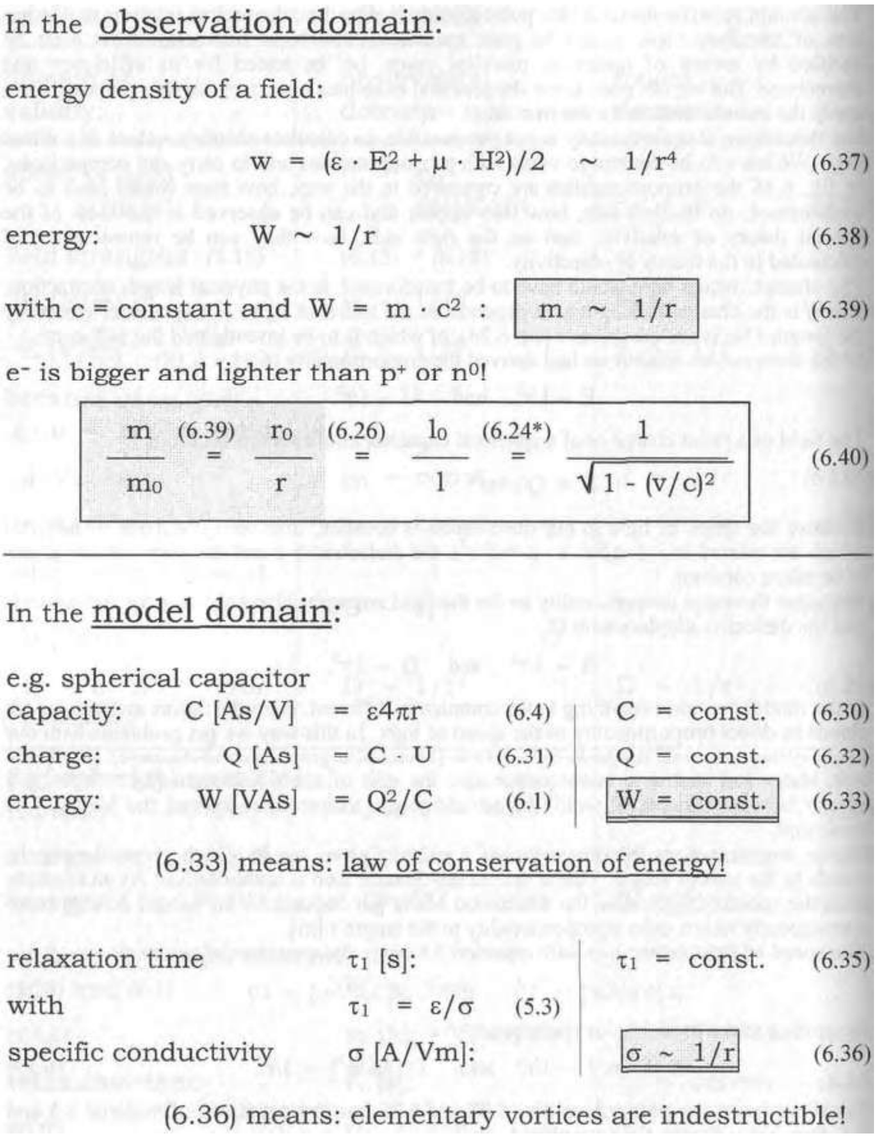

In the **observation domain:**

energy density of a field:

$$w = (\varepsilon \cdot E^2 + \mu \cdot H^2)/2 \quad \sim \quad 1/r^4 \qquad (6.37)$$

energy: $W \sim 1/r$ (6.38)

with c = constant and $W = m \cdot c^2$: $\boxed{m \sim 1/r}$ (6.39)

e^- is bigger and lighter than p^+ or n^0!

$$\boxed{\frac{m}{m_0} \overset{(6.39)}{=} \frac{r_0}{r} \overset{(6.26)}{=} \frac{l_0}{l} \overset{(6.24^*)}{=} \frac{1}{\sqrt{1-(v/c)^2}}} \qquad (6.40)$$

In the **model domain:**

e.g. spherical capacitor

capacity: C [As/V] $= \varepsilon 4 \pi r$ (6.4) | $C = \text{const.}$ (6.30)

charge: Q [As] $= C \cdot U$ (6.31) | $Q = \text{const.}$ (6.32)

energy: W [VAs] $= Q^2/C$ (6.1) | $\boxed{W = \text{const.}}$ (6.33)

(6.33) means: law of conservation of energy!

relaxation time τ_1 [s]: | $\tau_1 = \text{const.}$ (6.35)

with $\tau_1 = \varepsilon/\sigma$ (5.3)

specific conductivity σ [A/Vm]: | $\boxed{\sigma \sim 1/r}$ (6.36)

(6.36) means: elementary vortices are indestructible!

Fig. 6.19: Interpretation of the dependencies on radius

6.19 Interpretation of the transformation table

The transformation should tell us, what we would see if the variable speed of light would be observable to us. Doing so highly interesting results come out.
The energy density of a field is as is known $w = (\varepsilon \cdot E^2 + \mu \cdot H^2)/2$. (6.37)
In the observation domain will, according to fig. 6.19, decrease the energy density w proportional to $1/r^4$. Multiplied with the respective volume we obtain for the energy itself the proportionality: $W \sim 1/r$. (6.38)
If we make use of the Einstein relation $W = m \cdot c^2$
with c = constant holds also for the mass m: $m \sim 1/r$. (6.39)
In this manner we finally find out, why the small nucleons (protons and neutrons) subjectively seen are heavier than the very much larger electrons. As a consequence does a relativistic particle experience the increase of mass (with the length contraction according to equation 6.24*):

$$\boxed{\frac{m}{m_0} \overset{(6.39)}{=} \frac{r_0}{r} \overset{(6.26)}{=} \frac{l_0}{l} \overset{(6.24^*)}{=} \frac{1}{\sqrt{1-(v/c)^2}}} \quad (6.40)$$

This result is experimentally secured. Our considerations therefore are entirely in accord with the Lorentz-transformation. There at least is no reason to doubt the correctness.
In the model domain we with advantage assume a spherical symmetry. As easily can be shown with equations 6.4 and 6.31, are the capacity and charge of a spherical capacitor independent of the radius (6.30 and 6.32). In that case also the from both values calculable energy (6.1) must be constant. We come to the same conclusion, if take we the above equation 6.37 for the energy density of a field or if we carry out a verification of dimensions:
W [VAs] = konst. . (6.33)
This simple result is the physical basis for the law of conservation of energy! With that we have eliminated an axiom.
The result states that the energy stays the same, even if the radius, the distance or the speed of an object should change. To the subjectively observing person it shows itself merely in various forms of expression. Consequently is the energy, as is dictated by the here presented field theory, formed by binding in the inside of the quanta the same amount of energy but of the opposite sign. The amount of energy therefore is bound to the number of the present particles, as we already had derived.
Under the assumption of a constant time (6.35) there results for the electric conductivity σ, by calculating backwards over the equation of the relaxation time (5.3), the proportionality: (6.36)

$$\boxed{\sigma\,[\mathrm{A/Vm}] \sim 1/r} \quad (6.36)$$

Maybe the result surprises, because it can't be observed. Actually we know that the (microscopically observed conductivity in reality only represents an approximated averaged measure for the mobility of free charge carriers. In a particle-free vacuum however this well-known interpretation doesn't make sense anymore. Hence it is recommended, to only work with the relaxation time constants. Who nevertheless wants to continue to work with σ as a pure factor of description, can do this. But he mustn't be surprised, if in the model domain with decreasing radius the conductivity suddenly increases. But this is necessary, because otherwise the elementary particles would collapse. Only by the increase of the conductivity, which is produced by the spherical vortex itself, will the expanding eddy current build up in the inside of the particles, which counteract the from the outside concentrating potential vortex.

Approach:

a.The particles don't decay by themselves, but only by a corresponding disturbance from the outside.

b.The decay time is the statistical average in which such a disturbance can occur and take effect.

c.The elementary particles consist of an integral and finite number of elementary vortices, which can't decay anymore for their part.

d.If the compound particles get into the disturbing range of influence of high-frequency alternating fields, then they are stimulated to violent oscillations and in that way can be torn apart into individual parts.

e.As disturbing factor the high-frequency fields of flying past neutrinos are considered primarily.

f. Authoritative for the threshold of decay and with that also for the rate of decay is the distance, in which the neutrinos fly past the particle.

g.The distance becomes the larger, the smaller the particle is. If the particle thus experiences a relativistic length contraction, then it will, statistically seen, to the same extent become more stable!

That has nothing to do at all with time dilatationl

We are entitled to demand a simultaneity, after all we are the ones, who tell what that is!

Fig. 6.20: Proposal for an interpretation of the particle decay

<i>: Walter Theimer: Die Relativitatstheorie, Seite 106, Francke Verlag, Bern, 1977, ISBN 3-772O-126O-4

6.20 Particle decay

We still have to get rid of a fundamental misunderstanding. It concerns the problem of the time dilatation. Here the model domain doesn't give us any difficulty, because it dictates a constant and therefore by us definable time. In the relativistic view however should in moving systems clocks go wrong! But how does one want to explain a time dilatation physically, if it merely represents a purely mathematical result of the actually taking place length contraction on the one hand and the postulate of a constant speed of light on the other hand?

Nobody has troubled more about the physical interpretation than Einstein himself. But he had as less as we nowadays the possibility to verify the so-called phenomenon experimentally, by accelerating a laboratory clock to values close to the speed of light. Only atomic particles can, e.g. in accelerator systems, be brought on such high speeds and then be observed for their properties. But also these experiments don't have any power of proof, as long as we don't know the atomistic structure of the particles and there exists the danger of misinterpretations.

So the slowing down of the rate of decay of instable. particles at high speeds willingly is cited as "proof for time dilatation[i]. "The most cited example for the time dilatation is the "long-living" meson. The μ-meson is a charged particle, which exists only $2{,}2 * 10^{-6}$ seconds if it is observed in rest. Then it decays ... About 10 % of the mesons reach the earth's surface. Even if they fly with approximately the speed of light, they at least must have used $30 \cdot 2{,}2 * 10^{-6}$ seconds, in order to reach the earth. Their "life" therefore by the movement has been extended for a multiple... to the supporters of the theory of relativity here the time dilatation is revealed..."

This "proof however is worthless, as long as "the structure and the mechanism of decay of the particle are not known", like W. Theimer[i] expresses himself.

On the basis of the new field theory the approach standing on the left page is dared (fig. 6.20). Accordingly the particles don't decay by themselves, but only by a corresponding disturbance from the outside, which for instance is triggered by the high-frequency fields of flying past neutrinos. The closer the neutrinos fly past the particle, the sooner it will decay. But the distance becomes the larger, the smaller the particle is. If the particle thus experiences a relativistic length contraction, then it will, statistically seen, to the same extent become more stable!

That has nothing to do at all with time dilatation, as this proposal for an interpretation shows (fig. 6.20). The same effect of course also occurs, if atomic clocks are taken for a fly in a plane and compared to identically constructed clocks on earth.

The time was stipulated by us and therefore should be able to keep its universal validity. We are entitled to demand a simultaneity, after all we are the ones, who tell what simultaneity is!

An interesting technical use would be the acceleration of the rate of decay in order to dispose of radioactively contaminated waste. For that the waste has to be irradiated by collecting and focussing free neutrinos or with the help of a neutrino transmitter, like one which will be discussed in chapter 9. After such a neutrino shower dangerous radioactive waste would be reusable or at most be harmless domestic refuse.

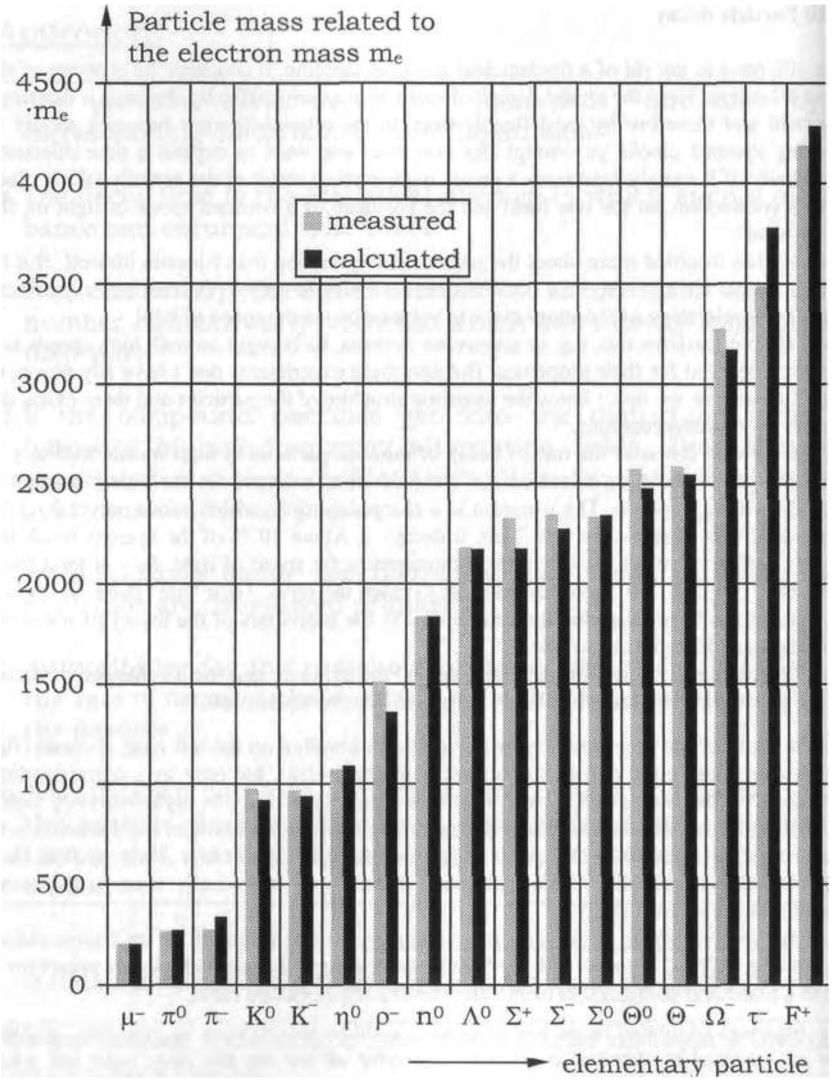

Fig. 7.0: The comparison with power of proof:

▨ the measured particle mass

■ the calculated particle mass

7. Proof

Ample evidence is available for the correctness of the theory of objectivity. The field dependent change in length is observed and used as magnetostriction or electrostriction. If a ferromagnetic material, e.g. a nickel rod, is brought into an alternating magnetic field, then field dependent longitudinal length oscillations are observed. In the same manner barium titanate or quartz crystal oscillates in the electric field if a high-frequency alternating voltage is applied.
A practical application forms the production of ultrasound.

In this chapter are, as already announced, the quantum properties of the elementary particles calculated and in this way is furnished perhaps the most convincing proof for the existence of potential vortices and for the correctness of the field-theoretical approach and the theory which is based on it.
A special challenge represents the calculation of the particle mass. This mass stretches from 207 electron masses of the myon over 1839 of the neutron into the order of magnitude of 18513 electron masses (Y°). Doing so not only can be tested, if the calculated values correspond with the measured ones. Also the gaps have to correspond, i.e. where there doesn't exist a discrete mathematical solution also no particle should exist. The fig. 7.0 standing on the left page anticipates the result and shows that even this strict condition is fulfilled! The agreement of the calculated with the measured results is excellent. If in individual cases small deviations become visible, we always have to bear in mind that the measurements as a rule are analysed statistically and the results are falsified if small particles creep in unrecognized. Particle physics nowadays has at its disposal extremely precise gauges, but even here remaining errors can't be excluded.

Quantum physics is occupied with further taking apart the elementary particles into hypothetic particles, the quarks, and to sort these according to properties and symmetries. Seen strictly causal this procedure thus corresponds to the quantum physical approach. We however have taken the field-theoretical approach, and this excludes the introduction of hypothetic particles from the start. It should be our goal to derive and to explain the quantum structure as a field property. Yes, we even want to calculate it, with which we would have overtaken quantum physics in the scientific competition with one leap!

Strong support our approach has experienced by current experiments, in which matter was transformed in electromagnetic waves - practically the reversal of the rolling up of waves to vortices. To do so at the Massachusetts Institute of Technology (David Pritchard and others) sodium atoms were dematerialized in waves by lattice scattering<i>. According to Einstein one surely could have blown the whole M.I.T. in the air with the occurring mass defect; but don't worry, no emission of energy whatsoever has been observed, entirely as predicted by the vortex theory.

<i>: J. Teuber: Materie und Energie, Ganze Atome wurden zu Energiewellen,
Illustrierte Wissenschaft Nr. 7 (1996), S. 56

With the classical radius of the electron $r_e = 2{,}82 * 10^{-15}$ m:

$$C_e = \varepsilon_o \cdot 4\pi r_e = 3{,}135 \cdot 10^{-25}\text{F} \qquad (6.4^*)$$

$$U_e = e/C_e = 511\text{kV} \qquad (6.31^*)$$

(constant independent of r_e)

Formation forms (vortex properties):

I. Amassing (formation of vortex balls):

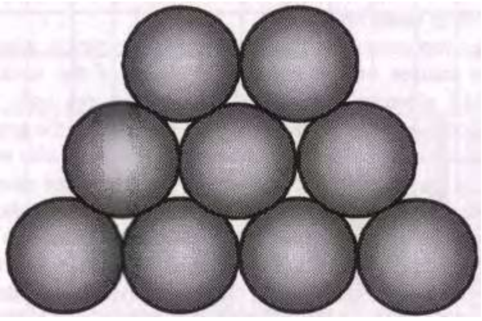

II. Overlapping (phenomenon of transport)

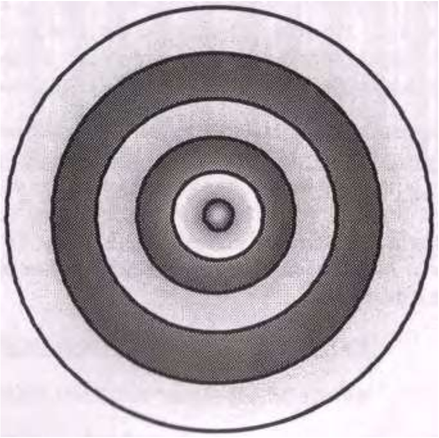

Fig. 7.1: The amassing and overlapping of elementary vortices

7.1 Elementary vortices

We had derived the electron and the positron as elementary vortices (fig. 4.3). Before we can go in the calculation, we must gain a clear picture of the possible configurations of vortices, which for reason of the derived properties are possible. For that we start with the elementary vortex and afterwards we predict the behaviour of interaction which can be expected.

Actually only one single particle is really elementary. According to the realizations of the new theory it is an elementary vortex in the form of a sphere. Its size is determined by the speed of light and this again by the local field strength; its stability is founded in the concentration effect of the potential vortex. The whirling takes place everywhere with the speed of light, even in the vortex centre, where all field lines run together, where the field increases infinitely and the speed of light goes to zero. This last circumstance owes the elementary vortex its localization.

We can attribute a charge to this vortex for reason of the field lines which on the outside run towards infinity and which we can measure (fig. 4.3). This is the smallest indivisible unit, the elementary charge e. Structure and course of the field lines suggest to understand and to calculate the elementary vortex as a spherical capacitor. By basing on the classical radius of the electron r_e given in fig. 6.3 the capacity according to equation 6.4 is calculated to be:

$$C_e = \varepsilon_0 \cdot 4\pi r_e = 3{,}135 \cdot 10^{-25}\ \mathrm{F} \quad (6.4^*)$$

Here the theory of objectivity has provided us the realization that even for a change of the radius of the electron the capacity remains unchanged constant (6.30), and this entirely corresponds to our observation.

Between the hull of the elementary vortex, measured at the radius r_e, and its centre, respectively also with regard to infinity, there exists according to equation 6.31 the tension voltage of:

$$U_e = e/C_e = 511\ \mathrm{kV} \qquad (6.31^*)$$

It as well is constant and independent of the size of the elementary vortex.

Since a different solution is refused, we'll have to assume that all elementary particles consist of an integer multiple of elementary vortices. For that the amassing, like closely packed tennis balls, or the overlapping of individual vortices in the form of shells, like in the case of an onion (phenomenon of transport) can be considered.

The among each other occurring forces of attraction can be traced back to the fact that every elementary vortex is compressed by the field of its neighbour as a consequence of the field dependent speed of light. This field as a rule is for the small distances considerably larger than the field on the outside. Therefore do compound elementary particles not have the twofold or triple mass, but at once the 207-fold (myon) or the 1836-fold (proton) mass. After all there is no other explanation for the fact that there don't exist lighter particles (with a mass less than 207 electron masses)!

a. The electron-positron pair

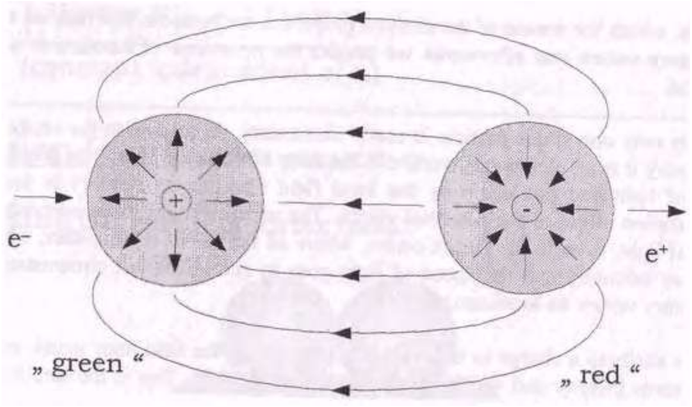
+
-
e⁻
e⁺
„ green "
„ red "

b. . The e^- - e+ pair for a small distance:

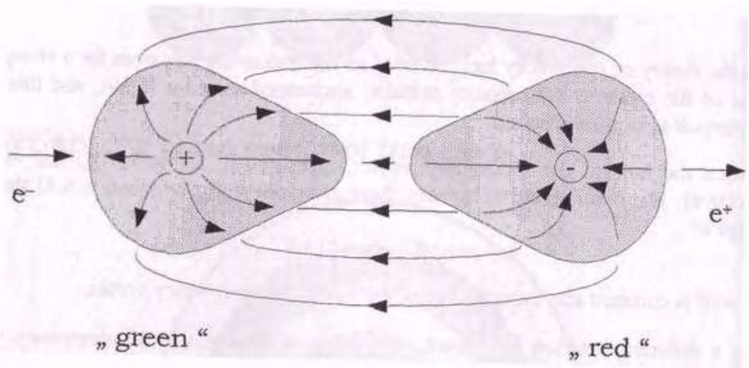
+
-
e⁻
e⁺
„ green "
„ red "

Fig. 7.2: The electron-positron pair annihilation

7.2 Matter and anti-matter

For the amassing or overlapping of elementary vortices several cases must be distinguished, because two inverse forms of formation are possible for the elementary vortex: the negatively charged electron and the positively charged positron. Whereas in the case of the electron the vortex produces a component of the electric field which points from the inside to the outside, has the field in the case of the positron the opposite direction for reason of a reversed swirl direction.

This statement can be generalized: if we consider the elementary particles from the outside, then we assign the particles with a swirl direction identical to that of the electron to the world of matter and call the particles with the opposite swirl direction anti-matter. It now is strongly recommended, to take colours to hand, in order to optically clarify the properties of vortices. The electron will be marked as a green sphere and the antiparticle, the positron, as a red sphere.

If we now look into the world of matter, then appears our world of matter to us "green", the world of anti-matter however "red". The uniform green colour of all the in our world existing elementary particles however doesn't exclude that red anti-vortices can exist hidden in the inside of the green vortices, where we can't discover them. But they must be completely covered, otherwise a disastrous reaction occurs, the pair annihilation, as a consequence of the oppositely directed property of the vortices which cancel out. By means of the pair annihilation a dematerialization can occur, because every elementary vortex keeps in its inside the same amount of energy with opposite sign and the fusion of two inverse particles can result in a zero sum of the energy. The best known example is the annihilation of an electron-positron pair under emission of radiation discovered by Klemperer in 1934. In the upper representation (fig. 7.2a) the elementary vortices still are symmetrical, but the outside field lines already are "bent" and linked together in such a way that, with the exception of the ones in the direction of the axis, no interaction takes place which can be measured.

The two particles for reason of the different charge approach each other quickly, and the closer they are, the larger the mutual force of attraction becomes; a vicious circle, which leads to the asymmetry shown in the lower sketch (fig. 7.2b) and only comes to rest, if both particles have destroyed themselves mutually.

The electron and the positron had the same amount of, but oppositely directed swirl activity, so that purely arithmetically seen a zero sum of the rest energy results. But it should be paid attention to both particles having some kinetic energy on the occasion of the relative motion to each other and if they rotate around their own axis also rotational energy. An emission of annihilation radiation occurs, is the explanation of particle physics.

With the knowledge of the photon (fig. 4.6) we can interpret the annihilation radiation as a consequence of the phenomenon of transport. The faster and consequently smaller vortex, for instance the green one, slips into the red one and sees the green inside, which is compatible for it. Unfortunately it only can remain there, as long as it is smaller, thus is faster, and therefore it shoots out on the other side again. Now the electromagnetic force of attraction fully takes effect. It is slowed down and the red vortex correspondingly accelerates. The process is reversed.

These around each other oscillating vortices, so we had derived, have a characteristic frequency (colour), are polarizable and are moving forward with the speed of light as a consequence of the open vortex centre. It therefore concerns the photon.

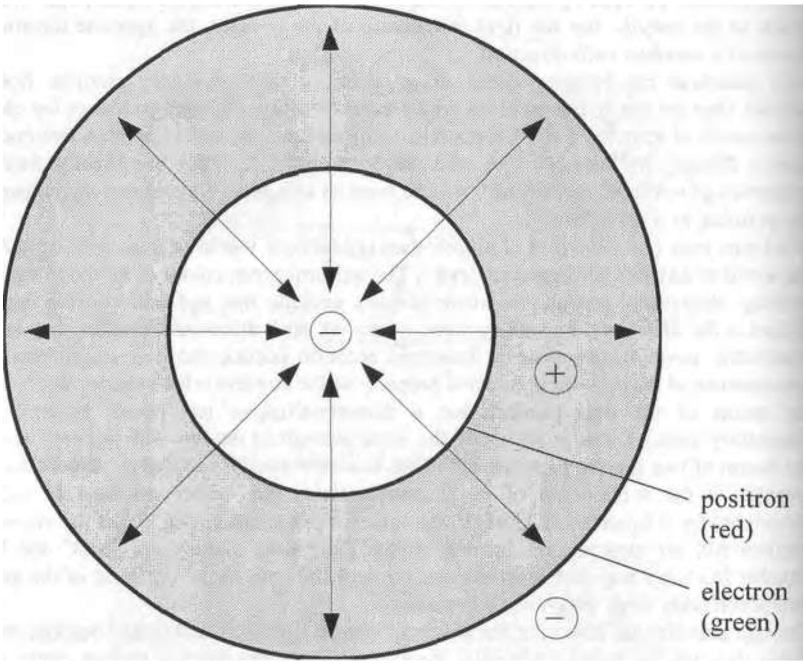
−
+
positron
(red)
electron
(green)
−

Fig. 7.3: Theoretical final state of the positronium = static γ-quant (photon).

7.3 Positronium

But before the two elementary vortices, the electron and the positron, are annihilated under emission of radiation, they will for a short time take a shell-shaped, a bound state, in which one vortex overlaps the other.

Its formation we can imagine as follows: an electron, flying past a resting positron, is cached by this for reason of the electromagnetic attraction and spirals on an elliptic path towards the positron. In doing so its angular velocity increases considerably. It will be pulled apart to a flat disc for reason of the high centrifugal forces, to eventually lay itself around the positron as a closed shell.

Now the red positron sees the electron vortex so to speak "from the inside" and doing so it sees as well red; because the green vortex has a red centre and vice versa! The result is the in fig. 7.3 given configuration.

The number of field lines, which run from the red border of the positron in the direction of the centre, is identical to the number, which point towards the green border of the electron. Here already the same state has been reached as in the centre, which corresponds to the state at infinity. That means that no field lines point from the green border to the outside; seen from the outside the particle behaves electrically neutral. It doesn't show any electromagnetic interaction with its surroundings.

If the particle were long-living, then it undoubtedly would be the lightest elementary particle besides the electron; but without stabilizing influence from the outside the positronium can't take the in fig. 7.3 shown state at all. The positron takes up the kinetic energy which is released if the electron becomes a shell around it. But before the bound state can arise, which would identify the positronium as an elementary particle, the equal rights of both vortices comes to light. With the same right, with which the electron wants to overlap the positron, it itself vice versa could also be overlapped.

If the stabilization of the one or the other state from the outside doesn't occur, then the stated annihilation under emission of y-quanta is the unavoidable consequence (fig. 4.6).

a. The electron pair:

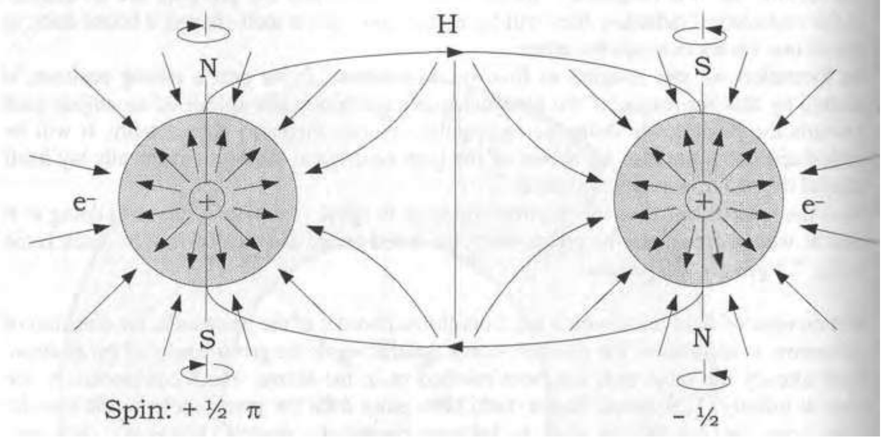
H
N
S
e⁻
+
e⁻
+
S
N
Spin: + ½ π
− ½

b. The electron pair for a small distance:

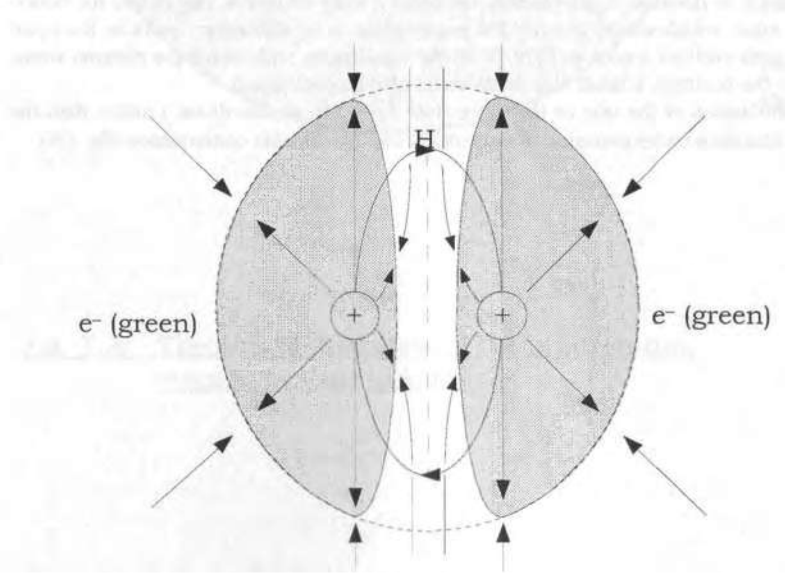
H
e⁻ (green)
+
+
e⁻ (green)

Fig. 7.4: Two electrons with oppositely directed spin

7.4 Dipole moment

As electrically charged spheres elementary vortices have a magnetic dipole moment along their axis of rotation as a consequence of the rotation of their own (fig. 7.4). This is measurable very precisely and for the most important elementary particles also known quantitatively. In contrast to the angular momentum the magnetic moment can't be constant according to the here presented theory. It should slightly change, if we increase the field strength in the laboratory.

In a particle consisting of several elementary vortices the vortices mutually increase the local field strength. Therefore we measure at the proton, which consists of three vortices, not the triple, but only the 2,793-fold of the nuclear magneton which can be expected for reason of its mass. Also the neutron has instead of the double only the 1,913-fold nuclear magneton. The deviations therefore are explicable as a consequence of the surrounding fields.

Prerequisite for this point are two other, still unanswered, key questions of quantum physics:

> XII: Why is measured for the proton approximately the triple of the magnetic dipole moment which can be expected for reason of the charge?
>
> XIII: Why does the neutron, as an uncharged particle, actually have a magnetic moment?

These questions can only be brought to a conclusive answer, if we have derived the vortex structures of the respective particles.

The elementary vortex, as a consequence of the spin along its axis, forms a magnetic north pole and a south pole. Another possibility to interact with an external field or with other particles is founded on this property. This shall be studied by means of two electrons. which form an electron pair.
For reason of the equal charge the two electrons at first will repel each other. If they rotate of their own they however will mutually contract, which, seen from the outside, is interpreted as a force of attraction. And in addition will they align their axes of rotation antiparallelly. While they now rotate in the opposite direction, a magnetic force of attraction occurs.

As is shown in fig. 7.4, the magnetic dipole field in this way is compensated towards the outside, as is clarified by the field line (H) with a closed course. Between both electrons a space free of E-field stretches. If both vortices are a small distance apart they lay themselves around this space like two half-shells of a sphere. A particle forms which seen from the outside is magnetically neutral, but it carries the double elementary charge (fig. 7.4b).
The exceptional affinity is always restricted to two vortices of equal charge with an opposite direction of rotation. Further vortices can't be integrated anymore and are repelled. This property of vortices covers the quantum condition (Pauli's exclusion principle) for the spin quantum number perfectly.

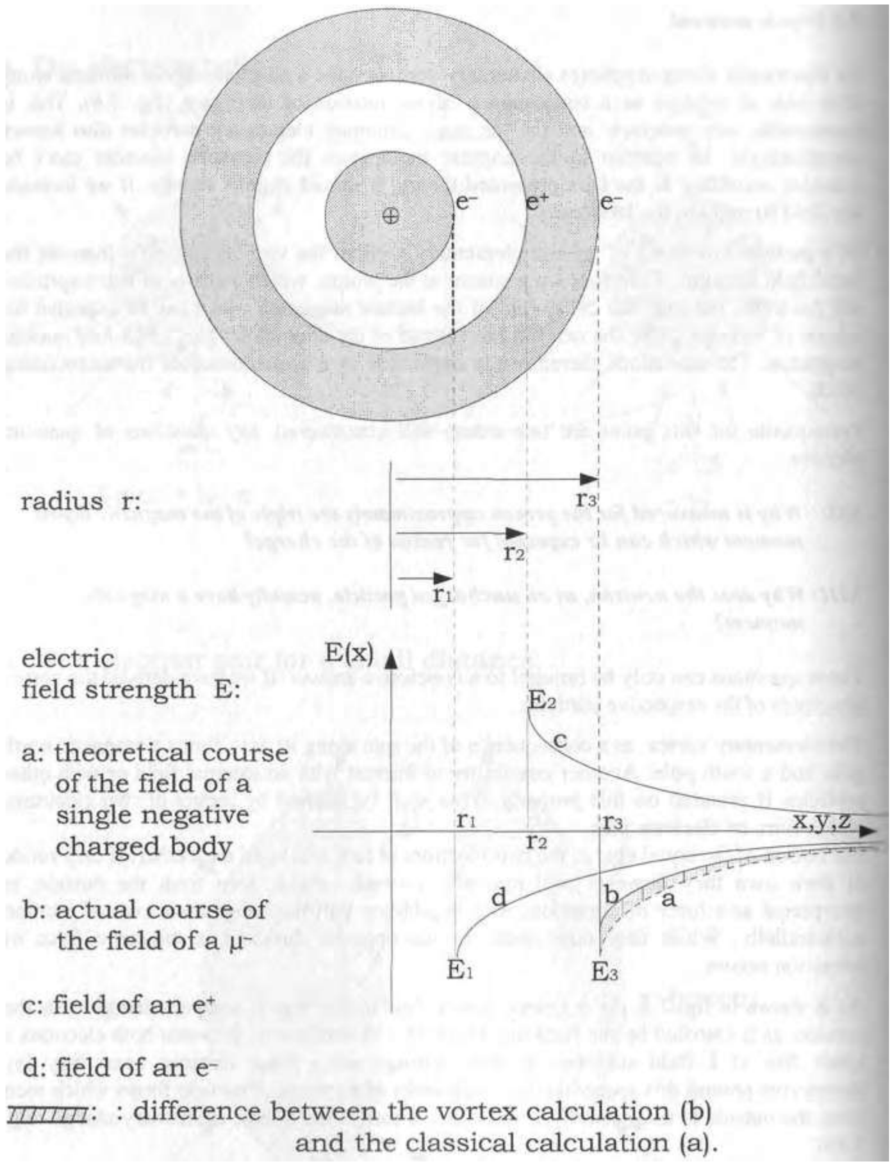
⊕
e⁻
e⁺
e⁻
radius r:
r3
r2
r1
electric field strength E:
E(x)
E2
c
r1
r3
x,y,z
r2
d
b
a
E1
E3
a: theoretical course of the field of a single negative charged body
b: actual course of the field of a μ⁻
c: field of an e⁺
d: field of an e⁻
▨: : difference between the vortex calculation (b) and the classical calculation (a).

Fig. 7.5: The mvon and the electric field E(x) of the three elementary vortices

7.5 Myon

We now have discussed all conceivable possibilities, which two elementary vortices can form: the creation of a pair for like charge and the annihilation under emission of photons via the formation of the positronium as an intermediate result for unequal charge. Next another elementary vortex shall be added and all different possibilities and configurations will be derived, which can be formed by amassing or overlapping.

The positronium can, as said, only take the in fig. 7.3 shown bound structure, if it is stabilized from the outside. This task now a further electron shall take over. According to the shell model the innermost elementary vortex an electron (e^-), is overlapped by a positron (e^+) and that again overlapped by an electron (e^-).
With an in the sum single negative charge, a completely symmetric structure as well as a half-integer spin this particle will show a behaviour corresponding to a large extent to that of the electron. Merely the mass will be considerably larger, because every vortex each time compresses the other two.

It therefore concerns the myon (μ^-), which also is called "heavy electron". The myon was discovered 1937 in the cosmic radiation (Anderson and others).
In fig. 7.5 are drawn above each other the shell-shaped structure of the myon and the electric field E(x) of the three elementary vortices.
It is visible that merely in the proximity of the particle the actual course of the field deviates from and is smaller, than the course which theoretically can be expected for a single negatively charged body. The difference is marked by a hatching.

We now can tackle the calculation of the myon. For that the following considerations to begin with are useful:
Mass is an auxiliary term founded in usefulness, which describes the influence of the electromagnetic field on the speed of light and with that on the spatial extension of the "point mass".
Without exception the local cosmic field E_o has an effect on a free and unbound elementary vortex, thus on an individual e^- or e^+, and determines so its size and its mass. But as long as we haven't determined this field strength, the calculation of its quantum properties won't succeed.
Instead the masses of compound particles will be compared to each other, which are so heavy that the field strength of the neighbouring vortices is predominant over the basic field E_0, so that a neglect of E_o seems to be allowed. The course of the calculation is made for all elementary particles in the same manner, which is explained hereafter.

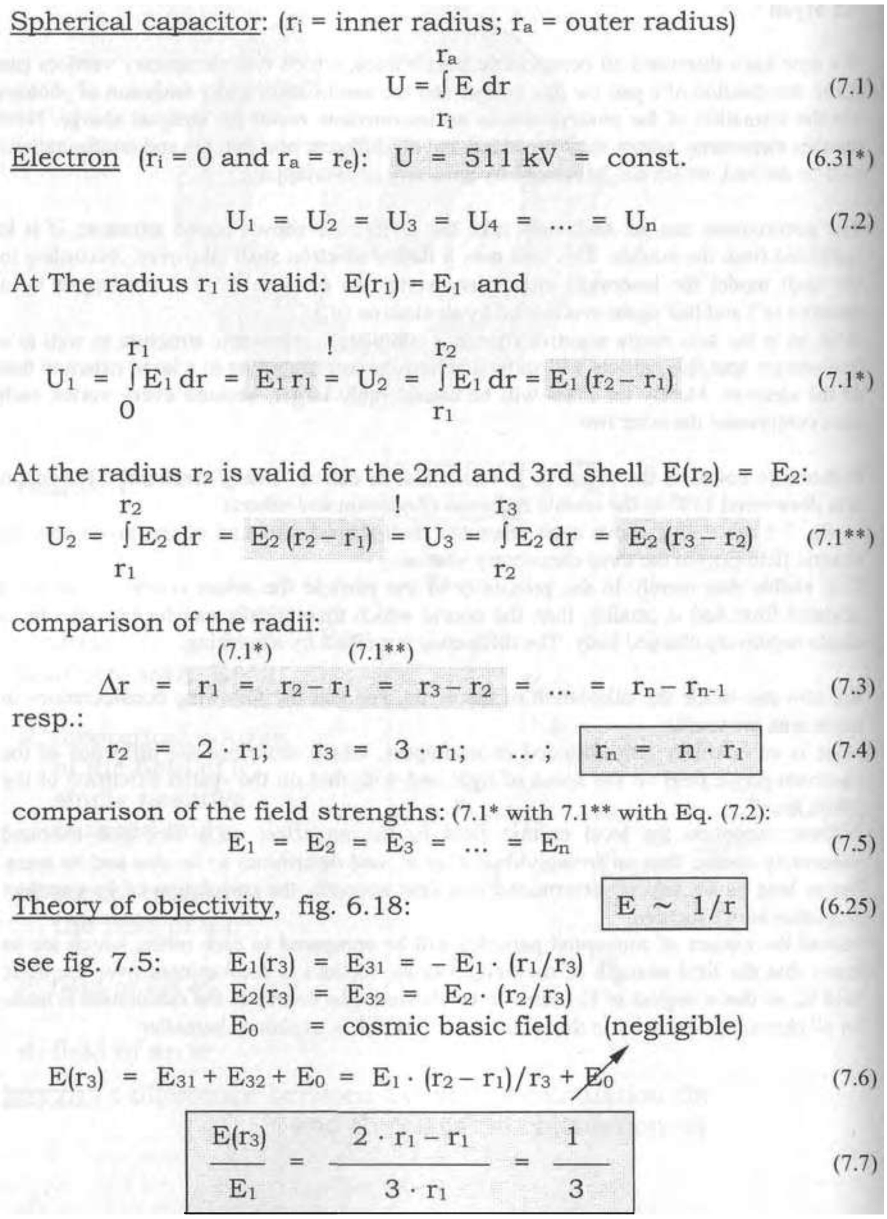

Spherical capacitor: (r_i = inner radius; r_a = outer radius)

$$U = \int_{r_i}^{r_a} E\,dr \qquad (7.1)$$

Electron ($r_i = 0$ and $r_a = r_e$): $U = 511\,\mathrm{kV} = \mathrm{const.}$ (6.31*)

$$U_1 = U_2 = U_3 = U_4 = \dots = U_n \qquad (7.2)$$

At The radius r_1 is valid: $E(r_1) = E_1$ and

$$U_1 = \int_0^{r_1} E_1\,dr = E_1 r_1 \overset{!}{=} U_2 = \int_{r_1}^{r_2} E_1\,dr = E_1 (r_2 - r_1) \qquad (7.1^*)$$

At the radius r_2 is valid for the 2nd and 3rd shell $E(r_2) = E_2$:

$$U_2 = \int_{r_1}^{r_2} E_2\,dr = E_2 (r_2 - r_1) \overset{!}{=} U_3 = \int_{r_2}^{r_3} E_2\,dr = E_2 (r_3 - r_2) \qquad (7.1^{**})$$

comparison of the radii:

$$\Delta r = r_1 \overset{(7.1^*)}{=} r_2 - r_1 \overset{(7.1^{**})}{=} r_3 - r_2 = \dots = r_n - r_{n-1} \qquad (7.3)$$

resp.:

$$r_2 = 2 \cdot r_1; \quad r_3 = 3 \cdot r_1; \quad \dots ; \quad \boxed{r_n = n \cdot r_1} \qquad (7.4)$$

comparison of the field strengths: (7.1* with 7.1** with Eq. (7.2):

$$E_1 = E_2 = E_3 = \dots = E_n \qquad (7.5)$$

Theory of objectivity, fig. 6.18: $\boxed{E \sim 1/r}$ (6.25)

see fig. 7.5: $E_1(r_3) = E_{31} = -E_1 \cdot (r_1/r_3)$

$E_2(r_3) = E_{32} = E_2 \cdot (r_2/r_3)$

E_0 = cosmic basic field (negligible)

$$E(r_3) = E_{31} + E_{32} + E_0 = E_1 \cdot (r_2 - r_1)/r_3 + \cancel{E_0} \qquad (7.6)$$

$$\boxed{\frac{E(r_3)}{E_1} = \frac{2 \cdot r_1 - r_1}{3 \cdot r_1} = \frac{1}{3}} \qquad (7.7)$$

Fig. 7.6: Calculation of the electric field strength E(r) of the myon from its dependency on radius

7.6 Calculation of the vortex fields

The tension voltage of an elementary vortex, like for a spherical capacitor, is determined by integrating over the electric field strength from the inner radius r_i up to the outer radius r_a:

$$U = \int_{r_i}^{r_a} E \, dr \quad (7.1)$$

For the electron ($r_i = 0$ und $r_a = r_e$) we already have carried out the integration and determined the tension voltage to be 511 kV (equation 6.31 *).

Doing so we further had discovered that it won't change, if the radius r varies. Even for a shell configuration, in which electrons and positrons alternately overlap, the approach is valid:

$$U_1 = U_2 = U_3 = U_4 = \ldots = U_n \quad (7.2)$$

At a certain radius all elementary vortices show the same density of field lines and with that also the identical field strength, so that we can solve the integral (7.1) for the each time neighbouring vortex shells and can compare the results:

At the radius r_1 with $E(r_1) = E_1$ the agreement, according to equation 7.1* (fig. 7.6), is valid for the innermost and the overlapped vortex shell.

At the radius r_2 with $E(r_2) = E_2$ the agreement according to equation 7.1** (fig. 7.6) is valid analogously for the 2nd and 3rd shell.

If still more shells are present, then we can arbitrarily repeat this procedure. For the radius of each shell we always obtain relation 7.3, which, related to the innermost radius, provides the following simple expression for the individual radii:

$$r_2 = 2 * r_1 ; \quad r_3 = 3 \bullet r_1; \quad \ldots ; \quad r_n = n * r_1 \quad (7.4)$$

From the comparison of the integration results 7.1* and 7.1** follows further that all elementary vortices produce the same field strength:

$$E_1 = E_2 = E_3 = \ldots = E_n \quad (7.5)$$

We infer from the transformation table (fig. 6.18, eq. 6.27) that the field strengths E and H decrease with 1/r. In fig. 7.5 the decrease of the fields with 1/r is shown. Up to the radius r, the field of the innermost vortex E_1 has worn off to the value $E_{31} = - E_1 \bullet (r_1/r_3)$. This field is overlapped by $E_{32} = E_2 * (r_2/r_3)$ as well as the cosmic basic field E_o:

$$E(r_3) = E_{31} + E_{32} + E_0 = E_1 \bullet (r_2 - r_1)/r_3 + E_o \quad (7.6)$$

The local basic field E_o is not known, but it is very small with regard to the field of the neighbouring vortex shells, so that a neglect seems to be allowed. From equation (7.6) in this way follows with the radius relation (7.4):

$$\frac{E(r_3)}{E_1} = \frac{2 \cdot r_1 - r_1}{3 \cdot r_1} = \frac{1}{3} \quad (7.7)$$

For the shell-shaped configuration of the myon (fig. 7.5) relation (7.7) indicates, which field the outside vortex shell is exposed to. From this can already be seen, how much it is compressed thanks to the field dependent speed of light and how much its mass as a consequence is increased.

Structure of the proton p^+:

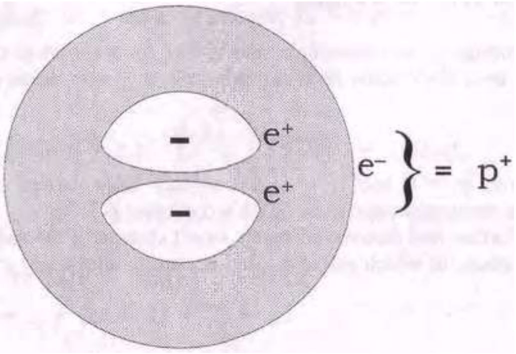

Calculation:
structure consisting of two shells, inner vortices with 2 • E_1,
field strength at the outer radius r_2:

$$E(r_2) = 2 * E_{21} = 2 * E_1 (r_1 / r_2) = E_1 \qquad (7.8)$$

Comparison of p^+ (7.8) with u^- (7.7) (z_e = number of the elementary vortices being involved with) in building up the structure, here each time $z_e = 3$):

Comparison of the radii with $E \sim 1 / r$ (6.27)

$$\frac{r_p}{r_\mu} = \frac{z_{e\mu}}{z_{ep}} \cdot \frac{E_\mu(r_3)}{E_p(r_2)} = \frac{3}{3} \cdot \frac{1/3}{1} = \frac{1}{3} \qquad (7.9)$$

Theory of objectivity (fig. 6.18): $m \sim 1/r^2$ (6.34)

$$\frac{m_p}{m_\mu} = \frac{z_{ep}}{z_{e\mu}} \cdot \left(\frac{r_\mu}{r_p}\right)^2 = \frac{3}{3} \cdot \left(\frac{3}{1}\right)^2 = 9 \qquad (7.10)$$

$$m_p/m_e = 9 * (m_u/m_e) = 9 * 207 = 1863 \qquad (7.11)$$

Measurement value, proton mass: $m_p = 1836 \cdot m_e$
Resp.:
measurement value myon mass $m_u = 207 * m_e$
myon calculated value: $m_p = 204 * m_e$. (error = 1,5%)

Since we, by using this calculation method, for the first time succeeded in deriving the mass of an elementary particle from that of another particle, the particle mass isn't a constant of nature anymore!

Fig. 7.7: Calculation of the proton

7.7 Calculation of the proton

If we again remember the affinity of two elementary vortices, which rotate with opposite spin. They align their axis of rotation antiparallel and form a very probable, but not particularly tight bound pair (fig. 7.4).

If we this time start with a positron pair, then does this pair have a double positive elementary charge. The two e^+ hence exert a particularly big force of attraction on electrons flying past them. If they have cached one and put it round as a shell, like a coat, then they will never again give it back! To again remove the electron, a triple positive charge would be necessary. But such a particle can't exist at all. The new particle therefore has an absolute stability and a very big mass, because the positron pair is considerably compressed by its outer shell. The total charge is single positive. With these properties it actually only can concern the proton. Its structure is shown in fig. 7.7. We can start from the assumption that both positrons are very close together in the inside and thus each forms the half of a sphere. For the calculation of the proton mass we then can assume as an approximation a structure of two shells, in which the inner vortex will have the double charge and the double field (2 * E_1). With equation 7.4 the field strength at the outer radius r_2 is:

$$E(r_2) = 2*E_{21} = 2*E_1*(r_1/r_2) = E_1 \qquad (7.8)$$

If we want to compare the results of the p^+ (7.8) and the μ^- (7.7), then it should be considered that the field of the innermost elementary vortex E_1 only is equal, if the number z_e of the elementary vortices involved in building up the particle is identical. Here with each time $z_e = 3$ this is the case. Because of equation 6.27 (E, H ~ 1/r) now also the radii are comparable:

$$\frac{r_p}{r_\mu} = \frac{z_{e\mu}}{z_{ep}} \cdot \frac{E_\mu(r_3)}{E_p(r_2)} = \frac{3}{3} \cdot \frac{1/3}{1} = \frac{1}{3} \qquad (7.9)$$

The mass of a particle first is determined by the number of the elementary vortices z_e. According to the theory of objectivity (fig. 6.18) however also the radius has an influence on the mass:

$$m \sim 1/r^2 \qquad (6.34)$$

This proportionality should be applied to the p^+-μ^-- comparison.

$$\frac{m_p}{m_\mu} = \frac{z_{ep}}{z_{e\mu}} \cdot \left(\frac{r_\mu}{r_p}\right)^2 = \frac{3}{3} \cdot \left(\frac{3}{1}\right)^2 = 9 \qquad (7.10)$$

The calculation provides a nine times bigger mass for the proton with regard to the mass of the myon. Therefore the mass of the proton related to the mass of the electron is:

$$m_p/m_e = 9*(m_\mu/m_e) = 9*207 = 1863 \qquad (7.11)$$

It would be favourable, to start from the with measuring techniques determined value for the mass of the proton $m_p/m_e = 1836$ and calculate backwards the related mass of the myon.

Then we obtain 204 as the calculated value instead of the measurement value m_μ/m_e= 207.

The reason for the deviation of 1.5 percent is caused by the neglect of the cosmic field E_o with regard to the field of the neighbouring elementary vortex. This neglect takes very much less effect for the relatively heavy proton than for the light myon.

The cosmic field therefore will compress the myon more strongly and increase the mass more strongly as is calculated here, in agreement with the measurement results.

Summarizing: since we, by using this calculation method, for the first time succeeded in deriving the mass of an elementary particle from that of another particle, the particle mass isn't a constant of nature anymore!

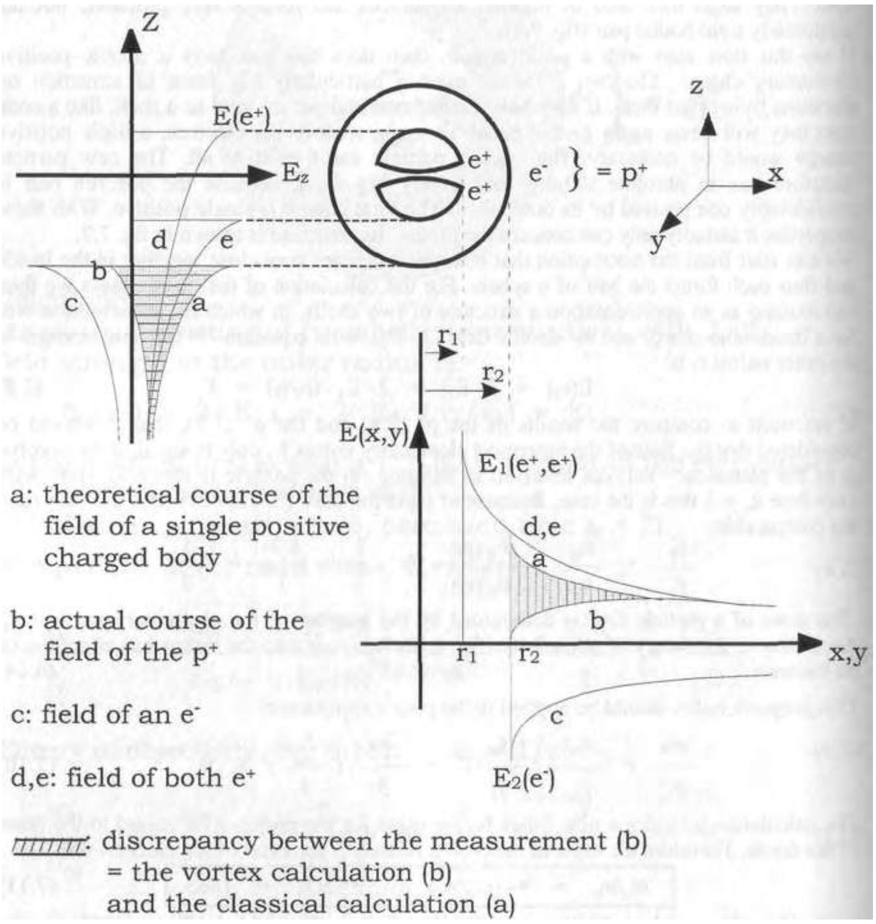
Z
E(e⁺)
E_z
e⁺
e⁺
e⁻
} = p⁺
z
x
y
d
e
b
c
a
r₁
r₂
E(x,y)
E₁(e⁺,e⁺)
d,e
a
b
r₁
r₂
x,y
c
E₂(e⁻)
a: theoretical course of the field of a single positive charged body
b: actual course of the field of the p⁺
c: field of an e⁻
d,e: field of both e⁺
▨: discrepancy between the measurement (b) = the vortex calculation (b) and the classical calculation (a)

Fig. 7.8: The proton and the electric field of the three elementary vortices in x-, y- and z-direction

7.8 "Strong interaction"

A central question of nuclear physics concerns the forces which keep the atomic nucleus, which consists of many neutrons and protons, together and give it its very good stability in spite of the like positive charge (key question XIV fig. 7.13).

According to today's textbook opinion (course of the field indicated with a in fig. 7.8) the forces of repulsion between the individual protons increase further as the distance gets smaller, to obtain immense values within the nucleus. They theoretically had to be overcome by new and unknown nuclear forces. Therefore physicists assume the hypothesis of a "strong interaction". But they are mistaken.

The answer to this open question is provided by the course of the field (b) for the proton, sketched in fig. 7.8. We see that the electric field at first indeed still increases if we approach the proton, but in the proximity it contrary to all expectations decreases again until it is zero. With that then also any force of repulsion has vanished! But the course of the field follows without compulsion from the overlap of the three individual elementary vortex fields.

The field direction in the z-direction even is reversed! In this topsy-turvy world, in theory, an electromagnetic force of attraction between two like charged protons can occur. We conclude:

A strong interaction doesn't exist at all. The usually given values for "range" and "strength" just represent a misinterpretation. The hatched drawn area marks the difference which is misinterpreted by quantum physics. The model concept over and above that answers another mysterious property of the proton. As an electrically charged particle with a spin it first of all should form a magnetic moment for reason of the rotating charge. But until now the measurable order of magnitude couldn't be explained.

7.9 Magnetic moment of the proton

If the inner positrons rotate around each other with oppositely pointing spin, then the magnetic field line is already closed within the particle and no effect in x- or y-direction is observable from the outside.

As pair they however still can rotate together around the z-axis and they'll do that. The overlapping electron for reason of its rotation of its own will likewise build up a magnetic dipole moment along its axis of rotation. It also will align its axis in the z-direction, so that now all three elementary vortices have one field axis. Being comparable to individually "elementary magnets" aligned in the same direction they produce a triple magnetic moment (key question XII fig. 7.13).

If we namely would start with a single positively charged body according to the theory of quantum mechanics, then we would have expected the value of the nuclear magneton[i] p_m as the magnetic moment for the proton $p_m = e \cdot \hbar/2m$. Opposite to that provide experiments with protons the approx. threefold value as already predictable by the new vortex theory. In addition does the direction of the vector p_{mp} correspond with the spin-axis, so as if the proton were negatively charged. The reason for that is that only the outermost elementary vortex determines the spin of the particle, and that is actually a negatively charged electron! Also this excellent agreement in the case of the proton can be judged as proof for the correctness of the vortex model.

---

[i]: The nuclear magneton has the value of: $p_{mk} = 5{,}0508 \cdot 10^{-27}\ Am^2$

a. n^0 if the e^- amasses to a p^+.

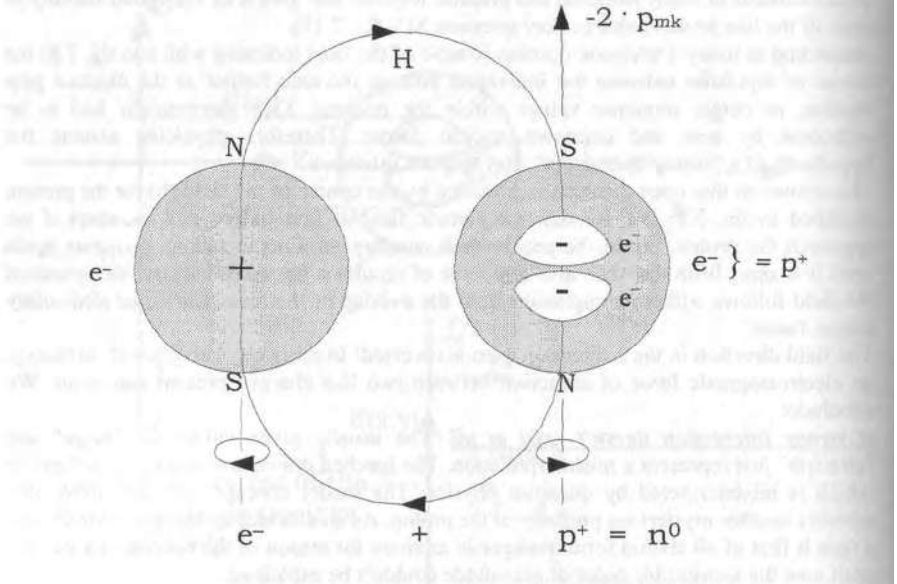
-2 · pmk
H
N
S
e⁻
+
-
e⁻
-
e⁻
e⁻ } = p⁺
S
N
e⁻
+
p⁺ = n⁰

b. n^0 if the e^- is overlapped by an e^+ of the p^+

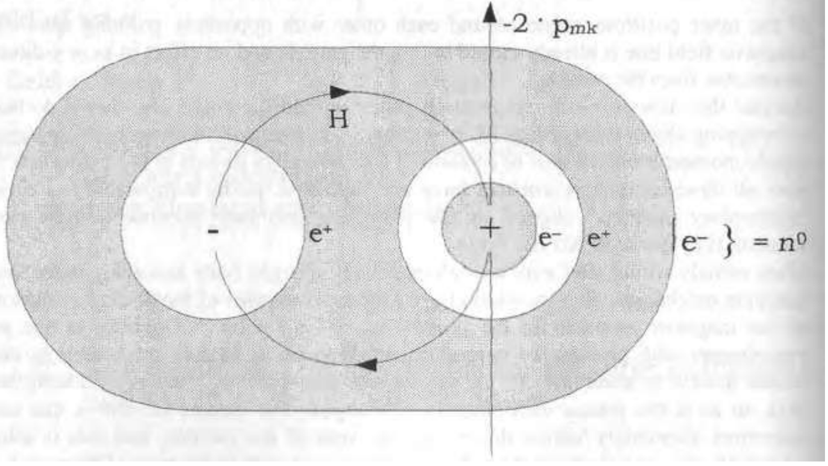
-2 · pmk
H
-
e⁺
+
e⁻
e⁺
e⁻ } = n⁰

Fig. 7.10: The neutron with magnetic dipole field H

7.10 Structure of the neutron

Until now could not be solved, why despite its missing charge also the neutron n° has a magnetic moment. The experimentally determined value is approx. the double of the nuclear magneton. Further was with measuring techniques an only 0,14% bigger mass with regard to the proton determined. The difference is approximately two and a half electron masses. And how reads the answer in the view of the potential vortex theory? It is obvious that a positively charged proton and a negatively charged electron mutually attract and amass together (fig. 7.10a). A pair annihilation can't occur, because the electron, which jackets both positrons, prevents this. The formation of an outer shell is not permitted by the high stability of the proton. It would have to be a positron shell, which instead of neutrality would produce a double positive charge. Conceivable is however the configuration, in which one of the two e^+ of the proton takes up the e^- in its inside and overlaps it (fig. 7.10b).

At first appears the amassing of p^+ and e^- to be the obvious answer to the structure of the neutron also in view of the small increase in mass. Since both elementary particles (p^+ and e^-) have a spin, will they align their axes of rotation antiparallelly and rotate against one another, exactly like an electron pair. But we now have unequal conditions: the proton brings the triple magnetic moment, the electron however only the single, and its field line will be closed by the proton. The difference which remains is the measurable double nuclear magneton, with which key question XIII (fig. 7.13) would be answered exhaustively.
This structure is shown in fig. 7.10a and has as rest mass the by only one electron mass increased proton mass, but it will deviate from this value, when the unequal partner come closer. Doing so the electron will be more strongly compressed by the heavier proton as vice versa.
Mass, magnetic moment and charge thus correspond to a large extent with the measurement values. Problems are seen concerning the spin and the stability.
Set of problems concerning spin: both the e^- and the p^+ have a half-integer spin, for which reason this configuration should have an integer spin.
Set of problems concerning stability: the neutron decays as is well-known in a p^+ and an e^- , but this object should be shorter-lived as determined by experiments. If namely the partner come each other very close, then the field strength of the p^+, contrary to expectation, doesn't increase but decreases, as is shown in fig. 7.8. The e^- therefore can only be bound very, very loosely; in z-direction it even will be repelled!
For these reasons is the open structure, which is shown in fig. 7.10a, not feasible as an isolated elementary particle, but only in a spatially extended network, like it is present in an atomic nucleus. In this case the neutron is, as is well-known, lighter by the mass defect, which is interpreted as binding energy.
Possibly it only concerns an intermediate stage. The heavier final product of the n° then could look like is shown in fig. 7.10b. For this version the line of the magnetic field already is closed partly within the particle, so that also here the approx. double nuclear magneton remains as a rest with a sense of orientation, as if the neutron were negatively charged.
Without charge and with the 1/2 spin it in this configuration fulfils all important quantum properties of the neutron, even that of the stability.

the field of the e^-: $E_{31}(-) = - E_1 (r_1/r_3)$,

the field of the e^+: $E_{32} = E_2 (r_2/r_3) = E_1 (r_2/r_3)$

and in addition the e^+: $E_{31} = E_1 (r_1/r_3)$.

$$E(r_3) = E_{31}^{(-)} + E_{32} + E_{31} + \underset{\searrow\,\text{negligible}}{E_0}$$

With the radius relation (eq. 7.4): $r_2 = 2 * r_1$ und $r_3 = 3 * r_1$
The total field is:

$$\frac{E(r_3)}{E_1} = -\frac{r_1}{r_3} + \frac{r_2}{r_3} + \frac{r_1}{r_3} = -\frac{1}{3} + \frac{2}{3} + \frac{1}{3} = \frac{2}{3} \qquad (7.12)$$

With $z_{en} = 4$ elementary vortices

$$\boxed{\frac{r_n}{r_p} = \frac{z_{ep}}{z_{en}} \cdot \frac{E_p(r_2)}{E_n(r_3)} = \frac{3}{4} \cdot \frac{1}{2/3} = \frac{9}{8} = 1{,}125} \qquad (7.13)$$

n_0 is 12,5% bigger than p±

$$\boxed{\frac{m_n}{m_p} = \frac{z_{en}}{z_{ep}} \cdot \left(\frac{r_p}{r_n}\right)^2 = \frac{4}{3} \cdot \left(\frac{8}{9}\right)^2 = 1{,}05} \qquad (7.14)$$

n^0 is 5% heavier than p±

Fig. 7.11: Calculation of the mass of the neutron

7.11 Calculation of the neutron

The calculation of the mass for the structure of the neutron according to fig. 7.10b has still remained open.

Because in this book for the first time has been shown, how the mass can be calculated, if the particles are understood as potential vortices, we also in this case again want to make use of this possibility.
We have, like for the μ^-,a structure of three shells with the radii r_1, r_2 and r_3. At the outer radius r_3 the fields of the elementary vortices on the inside have an effect on the electron On the outside: like is the case for the μ^-

the field of the e^-: $E_{31}^{(-)} = -E_1(r_1/r_3)$,

the field of the e^+: $E_{32} = E_2(r_2/r_3) = E_1(r_2/r_3)$

and in addition the e^+: $E_{31} = E_1(r_1/r_3)$.

The total field is, with the radius relation equation 7.4:

$$\frac{E(r_3)}{E_1} = -\frac{r_1}{r_3} + \frac{r_2}{r_3} + \frac{r_1}{r_3} = -\frac{1}{3} + \frac{2}{3} + \frac{1}{3} = \frac{2}{3} \quad (7.12)$$

If we compare the neutron, in which now $z_e = 4$ elementary vortices are involved, with the proton:

$$\frac{r_n}{r_p} = \frac{z_{ep}}{z_{en}} \cdot \frac{E_p(r_2)}{E_n(r_3)} = \frac{3}{4} \cdot \frac{1}{2/3} = \frac{9}{8} = 1{,}125 \quad (7.13)$$

then we infer from the arithmetically determined result that the neutron according to the radius is 12,5% bigger than the proton. The mass is calculated to:

$$\frac{m_n}{m_p} = \frac{z_{en}}{z_{ep}} \cdot \left(\frac{r_p}{r_n}\right)^2 = \frac{4}{3} \cdot \left(\frac{8}{9}\right)^2 = 1{,}05 \quad (7.14)$$

The particle therefore has a mass which is 5% larger than for the proton, slightly more as has been measured for the neutron. The difference is acceptable. The particle after all is structured very asymmetrically, in which the reason is to be seen, why the uncharged particle, looked at from close up, nevertheless shows an observable charge distribution.

Decay equations if neutrinos are involved:

$$n^0 \longrightarrow p^+ + e^- + \bar{\nu}_e \qquad (7.15)$$

$$n^0 + \nu_e \longrightarrow p^+ + e^- \qquad (7.15^*)$$

$$\mu^- \longrightarrow e^- + \bar{\nu}_e + \nu_\mu \qquad (7.16)$$

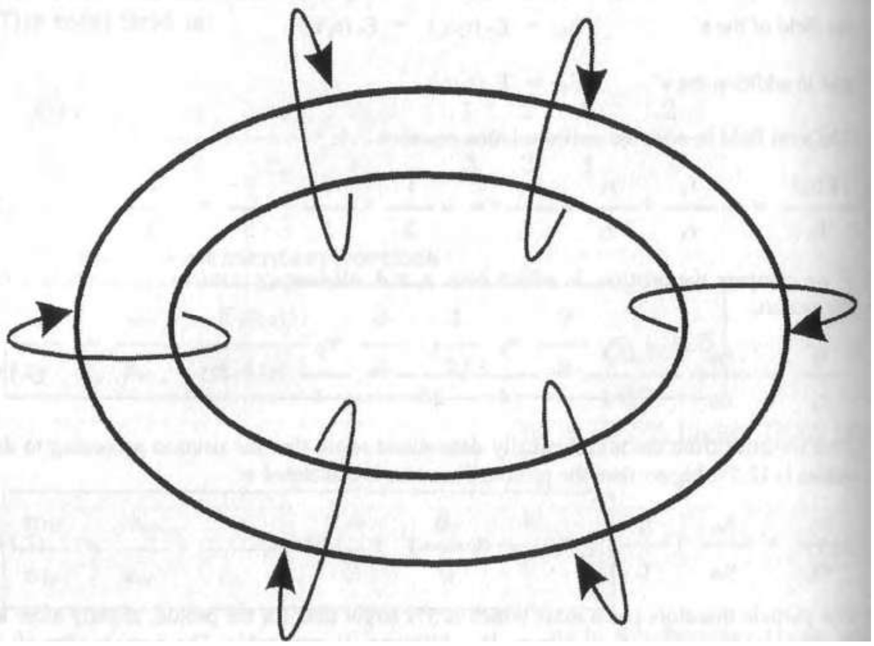

Fig. 7.12: The electron-neutrino as a ring-like vortex

7.12 ß-decay

In the case of the calculated quasistable particles, the μ^- and the n°, the verification by means of the well-known decay processes is still due. Also free neutrons, those which are not bound in an atomic nucleus, decay. But with an average life of 918 seconds they are by far the longest living among the quasistable elementary particles. Should the neutron decay be triggered by neutrinos, then obviously a distant flying past does not suffice. For that the electron is bound in the proton too tight. There probably has to occur a direct "crash", in which a neutrino is used, since the decay equation reads:

$$n^0 \longrightarrow p^+ + e^- + \bar{\nu}_e \quad (7.15)$$

As could be expected a proton p^+, an electron e^- and the mentioned electron-antineutrino ν_e are formed. What here is written down as the emission of an antiparticle, is equivalent in the absorption of the particle<i>, in this case of the neutrino. The reaction equation 7.15 can be reformulated accordingly<i>:

$$\boxed{n^0 + \nu_e \longrightarrow p^+ + e^-} \quad (7.15^*)$$

Also for the decay of the myon an electron-neutrino is used. In both cases it provides the energy necessary for the decay. But we can really understand the β-decay only, after we have got to know these particles better.

Without charge and without mass neutrinos show hardly any interactions with matter and as a consequence they possess the enormous ability of penetration - as is well-known. They are said to participate in the „weak interaction", which should trigger a conversion of the concerned particles, which is their decay. Pauli already has postulated the neutrino 1930 theoretically, because the transition from a half-integer spin to an integer spin for the n^0 -decay otherwise wouldn't have been explicable.

If we imagine an elementary vortex is being born, but the local field strength and energy isn't sufficient for obtaining a quantized state. The result is an incomplete potential vortex, which has an open vortex centre and as a consequence shows no localization at all. In the form of a vortex ring it oscillates around itself, while it continually turns its inside to the outside and then again to the inside.

One moment the vortex ring is green, then it is red again, one moment matter, then anti-matter, one moment positively charged and the next moment negatively charged. In contrast to the photon the number of the involved elementary vortices z_e for the neutrino is odd (for the ν_e z_e= 1). Perpendicular to the direction of propagation the neutrino has a spin $(s/\hbar = ½)$ for reason of a rotation, which overlaps the pulsating oscillation. This vortex ring is, as said, not a member of stationary matter, because it doesn't form a "black hole" in its centre, where the speed of light becomes zero. But it has an absolute stability like every elementary vortex, even if it only occurs incomplete and hence not in any quantized form,. This concept of the electron-neutrino as an open oscillating elementary vortex in the form of a ring-like vortex covers the experimentally determined realizations unexpectedly well.

<i>: Kussner, H.G.: Grundlagen einer einheitlichen Theorie der physikalischen Teilchen und Felder, Musterschmidt, Gottingen 1976, S.155

A strong interaction doesn't exist. The electric field in the proximity of the proton goes to zero within the range which is determined with measuring techniques.

A weak interaction doesn't exist. That interaction only is a special case of the electromagnetic interaction which appears in a weakened form.

XII: Why does the proton have approximately 3 times the magnetic moment which can be expected for reason of the only single charge?
(3 elementary vortices)

XIII: Why does the neutron as an uncharged particle anyway have a magnetic moment?
(Structure of the n°)

XIV: What owes the atomic nucleus, which consists of like charges, its stability?
(Course of the field of the p^+, instead of "strong interaction")

XV: Why does the free neutron decay, although it is stable as a particle of the nucleus?
(Interaction with neutrinos)

XVI: Why do neutrinos nevertheless participate in the "weak interaction", although they have no mass and no charge?
(Oscillating charge)

XVII: How can be given reasons for the finite range of the "weak interaction"?
(Reaction cross-section for particle decay)

Fig. 7.13: Further key questions of quantum physics (Continuation of figures 4.4 and 6.13)

7.13 "Weak interaction"

Let's now look again at the β-decay of the neutron, in which a neutrino ν_e is used. But this by no means will be a process of the weak interaction. Instead will neutrinos, contrary to the textbook opinion, participate in the electromagnetic interaction. They after all are one moment positively charged and the next moment negatively charged. With slow-acting gauges this it is true can't be proven, because the interaction is zero on the average. But this charged oscillating vortex ring can exert a considerable effect while approaching a neutron, which is based solely on the electromagnetic interaction.

The neutron is stimulated to synchronous oscillations of its own by the high-frequency alternating field of the neutrino, until it in the case of the collision releases the bound electron, which takes up the energy provided by the neutrino and transports it away. The interaction obviously is only very weak due to the oscillation. But a physical independency of it has to be disputed.

The finite range, which is given in this context, indicates the reaction cross-section around the n°-particle, within which the "crash" and as a consequence the β-decay occurs. This range is considerable larger as the particle itself. The electromagnetic interaction for such small distances after all is so violent, even if it only occurs in pulses, that the neutrino is thrown out of its path and can fly directly towards the neutron.

Perhaps we now understand also the β-decay of the myon. It actually were to be expected that without outside disturbance an absolute stability could exist because of the ideal symmetry of the μ^-. On our planet we however are in every second bombarded with approx. 66 milliard (billion) neutrinos per cm^2 <i>. Obviously it takes 2,2 μs on the average till a neutrino ν_e flies past a myon so close that it decays. In doing so it stimulates the outside elementary vortex to violent oscillations by trying to synchronize it. In this case the electron-neutrino ν_e carries away with it the two outer, and therefore weaker bound, elementary vortices of the myon, which meanwhile are oscillating synchronously. The innermost vortex, an electron e^-, is left behind. The decay of the myon which takes place with a probability of almost 100 % reads:

$$\mu^- \longrightarrow e^- + \overline{\nu}_e + \nu_\mu \qquad \text{resp.} \qquad \mu^- + \nu_e \longrightarrow e^- + \nu_\mu \quad (7.16)$$

Thus a different neutrino ν_μ, is formed which can be distinguished from the ν_e and is called myon-neutrino since it forms from the μ^-. Actually it even has a similar structure of three shells, as is shown in fig. 7.5. But the vortex centre is open and the particle isn't stationary anymore. In the picture now only a momentarily state is shown, in which the ν_μ appears green on the outside and red in its open centre. As already for the ν_e oscillates also here the inside to the outside and vice versa, this time merely as a packet of three shells, so that also this particle shows all the typical neutrino properties discussed for the example of the ν_e.

The for potential vortices typical and already discussed phenomenon of transport here has an effect. In particular in connexion with vortex rings this property is known from hydrodynamics. It thus can be observed, how vortex rings bind matter and carry away with them. Because the neutrino is not quantized, it neither is restricted with regard to its ability to transport elementary vortices. Consequently even bigger configurations are conceivable, like configurations of 5 shells, 7 shells etc..

<i>: "Zeugen aus der Sonne", VDI-Nachrichten Nr. 45 vom 9.11.90, Seite 26

a: the structure of the tau-neutrino with 5 shells

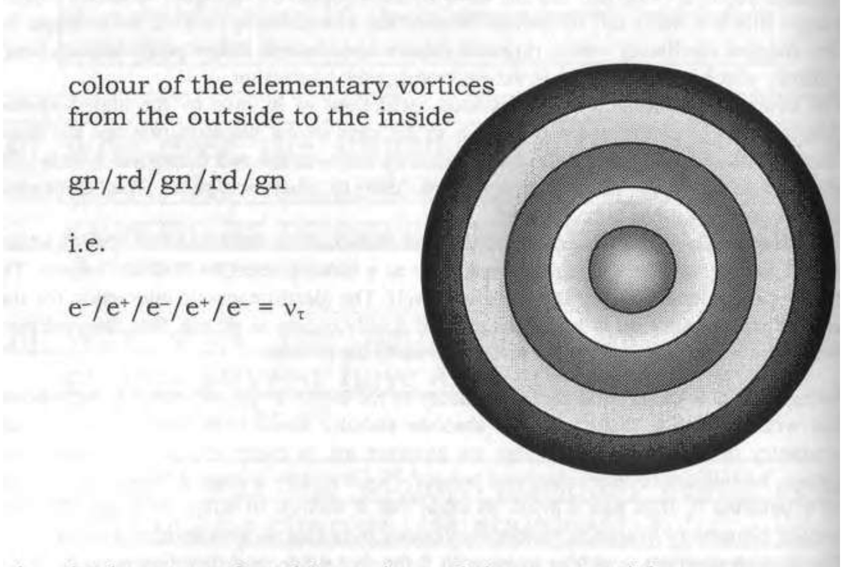
colour of the elementary vortices
from the outside to the inside
gn/rd/gn/rd/gn
i.e.
$e^-/e^+/e^-/e^+/e^- = \nu_\tau$

b: the heavy and stable variant of the tau particle

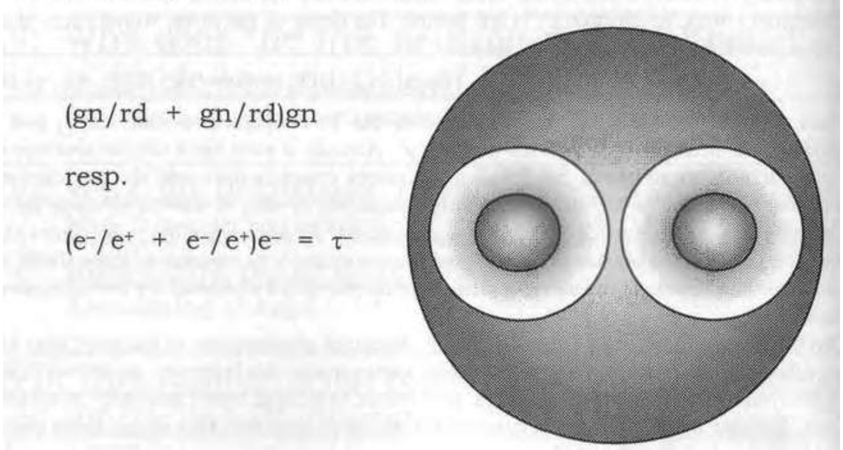
(gn/rd + gn/rd)gn
resp.
$(e^-/e^+ + e^-/e^+)e^- = \tau^-$

Fig. 7.14: Tau-neutrino and tau particle

7.14 Tau particle

In the table of the leptons after the e^- and the μ^- as the next particle the tau particle τ^- is found with its accompanying neutrino ν_τ.The obvious solution for the tau particle is the structure of five shells, as is shown in fig. 7.14a. With that the electron would have another particularly heavy relative with otherwise very similar properties.
For the myon the neutrino was stable, the particle itself however instable. We after all huve explained the particle decay as a consequence of an outside disturbance, and disturbances always are based on interactions. Correspondingly should, with the small possibility for an interaction, also the neutrino ν_τ of the tau particle have a better stability than the particle τ^- itself.
Without doubt this structure of 5 shells fulfils all known quantum properties like spin, charge etc. Merely the check of the mass is still due. This we now want to calculate for the structure shown in fig. 7.14a.

$$m_5 = m_p \cdot \left(\frac{Z_{e\tau}}{Z_{ep}}\right)^3 \cdot \left(-\frac{r_1}{r_5} + \frac{r_2}{r_5} - \frac{r_3}{r_5} + \frac{r_4}{r_5}\right)^2 \quad (7.17)$$

$$m_5 = 1836 \cdot \left(\frac{5}{3}\right)^3 \cdot \left(\frac{2}{5}\right)^2 \cdot m_e = \underline{\underline{1360 \cdot m_e}} \quad (7.17^*)$$

But the for the tau particle measured value is considerable higher!
Even if this structure is the only possible in the case of the neutrino ν_τ for reason of the complete symmetry, will the tau particle however change its structure by itself if another structure exists, which is more stable, thus in which the particle can take a bigger mass. Such a maximum provides the structure shown in fig. 7.14b after checking all possible configurations with five elementary vortices:

$$m_\tau = m_p \cdot \left(\frac{Z_{e\tau}}{Z_{ep}}\right)^3 \cdot \left(\left(-\frac{r_1}{r_3} + \frac{r_2}{r_3}\right) \cdot 2\right)^2 \quad (7.18)$$

$$m_\tau = 1836 \cdot \left(\frac{5}{3}\right)^3 \cdot \left(\frac{2}{3}\right)^2 \cdot m_e = \underline{\underline{3778 \cdot m_e}} \quad (7.18^*)$$

This value now lies 8% above the measurement values. It would be obvious, if unbound tau particles predominantly would take the structure shown in fig. 7.14b. The remaining error becomes explicable, if a very small number of tau particles in the lighter structure according to fig. 7.14a are involved with a correspondingly smaller probability.

The enormous variety of kinds of decay, and not a single one of the dominating ones has a probability of over 50%, makes it more difficult for us, to be able to directly infer the inner structure of a particle from the decay products. It nevertheless should be mentioned that after all 35% of all decays take place by taking up and using a neutrino ν_e or ν_τ, entirely in accordance with the model of the myon decay (equation 7.16).

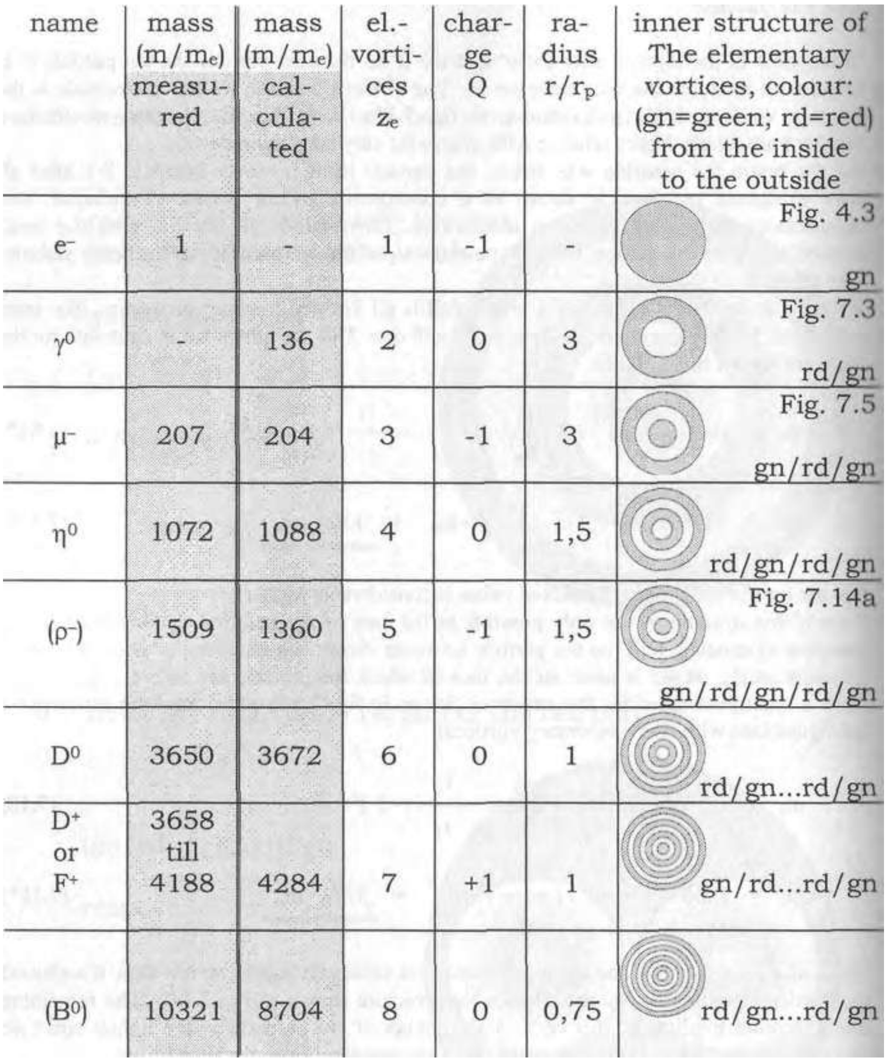

name	mass (m/m_e) measured	mass (m/m_e) calculated	el.-vortices z_e	charge Q	radius r/r_p	inner structure of The elementary vortices, colour: (gn=green; rd=red) from the inside to the outside
e^-	1	-	1	-1	-	Fig. 4.3 gn
γ^0	-	136	2	0	3	Fig. 7.3 rd/gn
μ^-	207	204	3	-1	3	Fig. 7.5 gn/rd/gn
η^0	1072	1088	4	0	1,5	rd/gn/rd/gn
(ρ^-)	1509	1360	5	-1	1,5	Fig. 7.14a gn/rd/gn/rd/gn
D^0	3650	3672	6	0	1	rd/gn...rd/gn
D^+ or F^+	3658 till 4188	4284	7	+1	1	gn/rd...rd/gn
(B^0)	10321	8704	8	0	0,75	rd/gn...rd/gn

7.15 Table of vortices of the calculated leptons and mesons compared with measurement values (Part 1).

7.15 Pions

Unlike the leptons, which we could derive and calculate fairly completely, the mesons don't have a half-integer spin. With this characteristic property they therefore can't represent an individually overlapped elementary particle and they probably will consist of the amassing in pairs of individual configurations of potential vortices. This kind of bond can't be particularly tight. Consequently we don't know any stable mesons.

The most important basic building part of the mesons we have got to know over the positronium in fig. 7.3. It necessarily has to amass to another particle, otherwise it annihilates under emission of a γ-quanta, as already mentioned. This γ^0-particle, as it will be named here, has the mass of:

$$m_{\gamma 0} = (2/3)^3 \cdot (1/2)^2 \cdot 1836 \cdot m_e = 136 \cdot m_e \quad (7.19)$$

which only can be determined arithmetically. As a partner, to which the γ^0-particle can amass, first of all another γ^0-particle should be considered. Because both partner will rotate against one another, this new particle would not have a spin and moreover would be uncharged. The mass now would be twice as big with:

$$m_{\pi 0} = 2 \cdot m_{\gamma 0} = 272 \cdot m_e \,. \quad (7.19^*)$$

But the two γ^0-particles will come very close together and mutually feel the local, in the same direction orientated, distribution of the field, which will lead to a weakening of the field and as a consequence to a slight reduction of the mass.
With these properties it probably concerns the uncharged pion π^0. This model concept finds an excellent confirmation in the two possible kinds of decay, which can be regarded as equivalent:

$\pi^0 \longrightarrow \gamma + \gamma$ with a probability of 99%
and
$\pi^0 \longrightarrow \gamma + e^- + e^+$ with a probability of 1%

Also in the case of the charged pion $\pi^\pm$ the observable decay offers a big help, which will take place with a frequency of almost 100 %:

$$\pi^\pm \longrightarrow \mu^\pm + \nu_\mu$$

The equation doesn't state anything about the fact, if a neutrino v_e is used in the process. But it points at the circumstance that the partner of the γ^0-particle for the $\pi^\pm$ most likely is a myon $\mu^\pm$. The mass will be smaller than the sum of both building parts:

$$(204+136) * m_e = 340 * m_e.$$

name	mass (m/m_e) measured	masse (m/m_e) calculated	elem.-vortices z_e	charge Q	inner structure of The elementary vortices, colour: (gn=green; rd=red) from the inside to the outside
B^-	11194	9793 +?	9 +?	-1	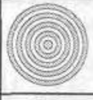 $r/r_p = 0,75$ gn/rt...rt/gn
(Y^0)	18513	17001 +?	10 +?	0	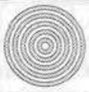 $r/r_p = 0,6$ rt/gn...rt/gn

Some compound configurations

π^0	264	272	4	0	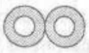 2 x fig. 7.3 $\gamma^0 + \gamma^0 = \pi^0$
π^-	273	340	5	-1	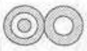 fig. 7.3 + fig. 7.5 $\mu^- + \gamma^0 = \pi^-$
τ^-	3487	3778	5	-1	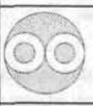 fig. 7.14b (2x gn/rd)/gn
(K^0)	974	918 +?	6 +?	0	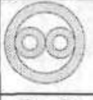 $r/r_p = 2$ (2x rd/gn)/ rd/gn
(K^-)	967	939 +?	11 +?	-1	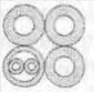 (rd/gn + gn)/ rd/gn + 3 γ^0

Fig. 7.16: Table of vortices of the calculated leptons and mesons compared with measurement values (Part 2).

7.16 Table of vortices of the mesons

The numerous kinds of decay for K-mesons suggest that these strange particles will consist of various combinations of amassed together and in pairs rotating γ^0- and $\mu^\pm$-particles. The possibilities of combination now already have increased in such a way that for every kaon and other mesons several solutions can be proposed. To avoid unfounded speculations, only a few clues will be given.

Besides the γ^0-particles also heavier arrangements should be considered as partner for the spin and as a building part for kaons and other mesons.
If for instance a π^0 is overlapped by a γ^0, then this particle has an arithmetically determined mass of 918 m_e. It therefore can concern a building part of the uncharged kaon K^0.

The likewise with three γ^0 formed configuration of 6 shells however, if it actually would staystable for the duration of a measurement, would have the mass of 3672 electron masses<i>.

A very much better detectability must be attributed to the configuration of 4 shells which consists of two γ^0, so to speak a heavy relative of the γ^0 and the π^0. It among others should be able to decay like a π^0.With this property and with an arithmetically determined mass of 1088 m_e it actually only can concern the η^0-meson. Solely according to the numeric value the η^0-meson could also consist of four π^0-mesons; but the decay in only two light quants speaks against it.

The kaon-puzzle in addition is made more difficult by the spontaneously possible ability to change of the involved γ^0-particles during a process of decay, as is made clear by the numerous kinds of decay. These dependent pion halves can be "swallowed" or "spit out" by neutrinos in the process, they can form from incident light or be emitted as photons and eventually they even can break up in their individual parts.

In fig. 7.16 the possible configurations of potential vortices are sketched and the respective, according to the new theory calculated, mass is given. If above that the other decay products and quantum properties, which can be given for the vortex structures, are added, like e.g. charge, spin and if need be magnetic moment, then an assignment without doubts to the until now only from measurements known elementary particles is possible. In order to better be able to assess the efficiency of the potential vortex theory, the measurement values are compared to the calculated values.

Some terms are put in brackets, because it can be assumed that the calculated part only concerns the dominating part, to which further γ^0 or other small configurations of vortices will amass for reason of its high mass. Correspondingly should the mass in that case be corrected slightly.

<i>: It could e.g. concern the D°-meson.

name	way of decay	mass (m/m$_e$) measured	mass (m/m$_e$) calculated	elem.-vortices z_e	charge Q	Inner structure of the baryon vortices
p^+	stable	1836	- (reference value)	3	+1	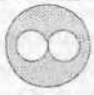 fig. 7.8 $r/r_p = 1$ $(2x\ e^+)/e^-$
n^0	$p^+ + e^- + \bar{\nu}_e$ (100%)	1839	1837	4	0	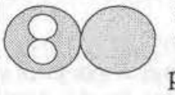 figure 7.10a $p^+ + e^-$
			1934	4	0	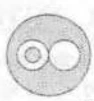 figure 7.10b n^0 $r/r_p = 1,13$
Λ^0	$p^+ + \pi^-$ (64%)	2183	(2176) 2187	8	0	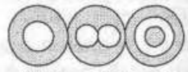 $\gamma^0 + p^+ + \mu^-$ (64%)
	$n^0 + \pi^0$ (36%)		(2206)	8	0	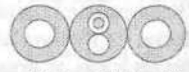 $\gamma^0 + n^0 + \gamma^0$ (36%)
Σ^+	$p^+ + \pi^0$ (51%)	2328	(2108+?) 2142+?	7 + ?	+1	$\gamma^0 + p^+ + \gamma^0$ (+?) (51%)
	$n^0 + \pi^+$ (49%)		2177	9	+1	$\gamma^0 + e^+ + p^- + \mu^0$ (49%)
Σ^-	$p^+ + \pi^-$ (100%)	2343	2274	9	-1	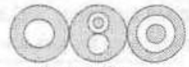 $\gamma^0 + n^0 + \mu^-$ (100%)

Fig. 7.17: Table of vortices used for the calculation of the most important barvons with suggestions for the structure (Part 3).

7.17 Table of vortices of the baryons

The number of possibilities of combination quickly increases, if only a few elementary vortices extend the structure of a particle. This probably is the reason for the large number of observable hyperons, which recently have been produced artificially and observed with the help of particle accelerators.

Both the neutron and the lambda particle can exist in a lighter and a heavier variant. At the moment of the decay, as it for instance is observed in a bubble chamber, according to expectation the state with the smaller mass takes the bigger probability. But in the amassing with further particles as building part of bigger and heavier hyperons the heavier structure is more likely. This circumstance should be considered in calculating the mass of the hyperons.

In figures 7.17 and 7.18 the most important baryons are listed, which are characterised in the way that one of the amassed together packets of vortices is a nucleon, thus a proton or a neutron.

The given, from measurements known, kinds of decay are able to confirm the inner structure pretty good. Of course an infinitely lot of combinations are conceivable and numerous predictions are possible. But speculations are unnecessary from the time on where we are able to calculate the particles!
The restriction to the few in the table listed particles seeming to be important hence doesn't limit the universal importance of the theory of objectivity in any way!

name	way of decay	mass (m/m_e) measured	mass (m/m_e) calculated	elem.-vortices z_e	charge Q	Inner structure of the baryon vortices
Σ^0	$\Lambda^0 + \gamma$ (100%)	2334	2342	10	0	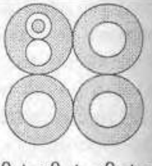 $n^0 + \gamma^0 + \gamma^0 + \gamma^0$ (100%)
Θ^0	$\Lambda^0 + \pi^0$ (100%)	2573	2478	12	0	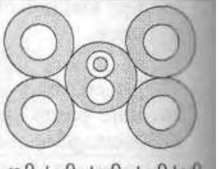 $n^0 + \gamma^0 + \gamma^0 + \gamma^0 + \gamma^0$ (100%)
Θ^-	$\Lambda^0 + \pi^-$ (100%)	2586	2546	13	-1	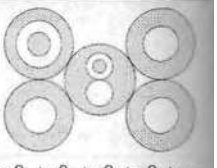 $n^0 + \gamma^0 + \gamma^0 + \gamma^0 + \mu^-$ (100%)
(Ω^-)	$\Lambda^0 + K^-$ (69%)	3273	3172 +?	17 +?	-1	$n^0 + \gamma^0 + \gamma^0 + ?$ (31%: ? = div.)

Fig. 7.18: Table of vortices used for the calculation of the most important baryons with suggestions for the structure (Part 4).

8. Unified theory

With the theory of objectivity the longed for goal of a "theory of everything" (TOE), of an universal theory, seems to have moved within reach. If in the nineteenth century still promising field theories and approaches were being discussed, then has at the latest Einstein's theory of relativity destroyed all hopes in such a theory. Science as a consequence has become very much more modest and understands a TOE only as the unification of all known interactions.

Einstein has stated the minimum demand so: "a theory should be favoured by far, in which the gravitational field and the electromagnetic field together would appear as a whole"<i>. It is evident that a subjective or relativistic observer theory never is able to achieve this.

The presented theory of objectivity made it possible that the unification here for the first time actually has succeeded. This undoubtedly brings science a whole lot further, but it still is not sufficient to lie one's hands in one's lap being content with oneself. After all we still know very much more phenomena, which likewise should be unified. After all it is no accident that both Maxwell and Einstein, to name only two prominent representatives, after completion of their well-known works have struggled for the question, what sort of phenomenon it concerns in the case of the temperature and how this could be integrated in their theory.

The requirement reads: We must be able to derive all basic factors, which influence our system of units with their basic units, as a compulsionless result from the new theory. Besides the dimensions of space and time which determine our continuum, the explanation and unification of the basic factors mass and charge has to be tackled. If we have succeeded in doing so, we'll also tackle the problem of the fifth and last basic factor, which until now has put itself in the way of any unified theory as the question of fate, the problem of the temperature!

<i>: Einstein, A.: Grundziige der Relativitatstheorie, Vieweg+Sohn, Braunschweig 1973, 5. Aufl., WTB 58. Seite 97.

> *"A theory should be favoured by far, in which the gravitational field and the electromagnetic field together would appear as a whole"*<i>.

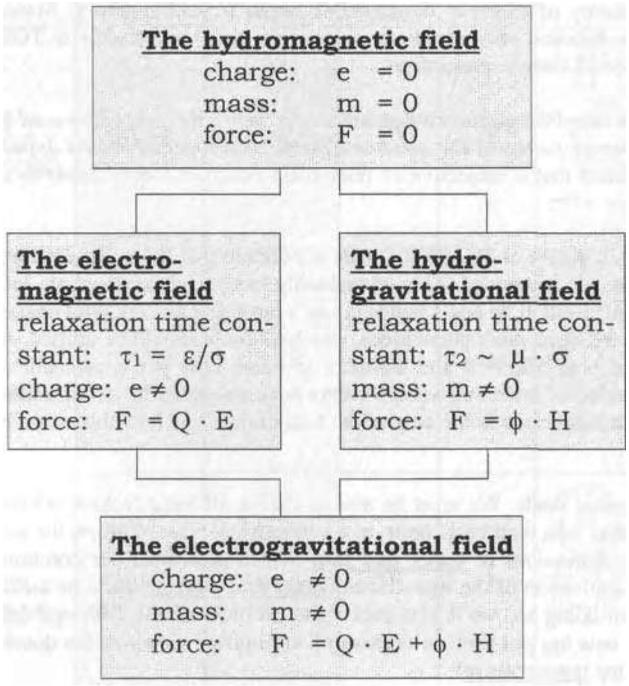
The hydromagnetic field
charge: e = 0
mass: m = 0
force: F = 0
The electro-magnetic field
relaxation time constant: $\tau_1 = \varepsilon/\sigma$
charge: $e \neq 0$
force: $F = Q \cdot E$
The hydro-gravitational field
relaxation time constant: $\tau_2 \sim \mu \cdot \sigma$
mass: $m \neq 0$
force: $F = \phi \cdot H$
The electrogravitational field
charge: $e \neq 0$
mass: $m \neq 0$
force: $F = Q \cdot E + \phi \cdot H$

Fig. 8.1: Structuring of the fields and definition of terms

<i>: Einstein, A. : Grundziige der Relativitatstheorie, Vieweg + Sohn, Braunschweig 1973, 5. Aufl. , WTB 58.

<ii>: derived from the Greek hydro (= water).

8.1 Structure of the field theory

In contrast to Maxwell's theory the new field theory, which we derived from duality, is also able to describe fields, in which no particles and no quanta exist. It probably is justified and useful in the sense of a clearer communication, to give the new field a name of its own.
The author recommends the introduction of the term "hydrotic field". In it should be expressed, which importance water has for both the like named potential vortex and this field<ii>.
As we already have worked out, the hydrotic field is favoured particularly by polar materials and by a high dielectricity. Water is a corresponding and in the biosphere of our planet dominating material.

Whereas we had to correct the concept of a vortex free electric field, we had until now, considerable, we can take over the description of the magnetic field unchanged. This then should also be valid for its name. The new field which consists of both correspondingly is called hydromagnetic field.

In fig. 8.1 we recognize the structure. At the top stands the "hydromagnetic field", which is described mathematically by the equations of dual electrodynamics in fig. 3.3. It does not know quanta and as logical consequence neither charge nor mass! If we insert these equations, Ampere's law and the dual formulated Faraday law of induction, into each other, then there results as a mathematical description of our space-time-continuum the fundamental field equation (5.7, fig. 5.1). As a new physical phenomenon the potential vortex appears, which gives the hydromagnetic field a new and important property: this field can be quantized!

Starting-point is the wave, which for corresponding interference effects can spontaneously roll up to a vortex, which as highly concentrated spherical vortex finds a new right to exist and finds to a new physical reality.

The in the described manner formed particles show specific properties of their own. We now are able to attribute them for instance a charge or a mass. And these properties also can be investigated and described individually and isolated from each other. Thus are formed the two special cases, strange by nature, on the one hand the well-known, with the help of the Maxwell equations describable "electromagnetic field" and on the other hand the new "hydrogravitational field".

If we overlap the results of the two special cases, e.g. by adding the force effects of electric charges and accelerated masses, then we summarized obtain a field, which we accordingly should call "electrogravitational". This case is not at all unknown. Already Niels Bohr in this way has calculated the radii of the electron orbits in the hull of his model of the atom, to mention only one example. We can summarize:

The hydromagnetic field is the all encompassing and with that most important field. Apart from that the electromagnetic field of the currents and the eddy currents and the hydrogravitational field of the potentials and the potential vortices merely describe the two possible and important special cases. For reasons of pure usefulness for every special case a characteristic factor of description is introduced, the charge and the mass!

Auxiliary terms (description of quantum properties):
* mass * charge * temperature * Planck's quantum of action

The hydromagnetic field does not know any quanta!

Structure of the fundamental field equation 5.7 (fig. 5.1):

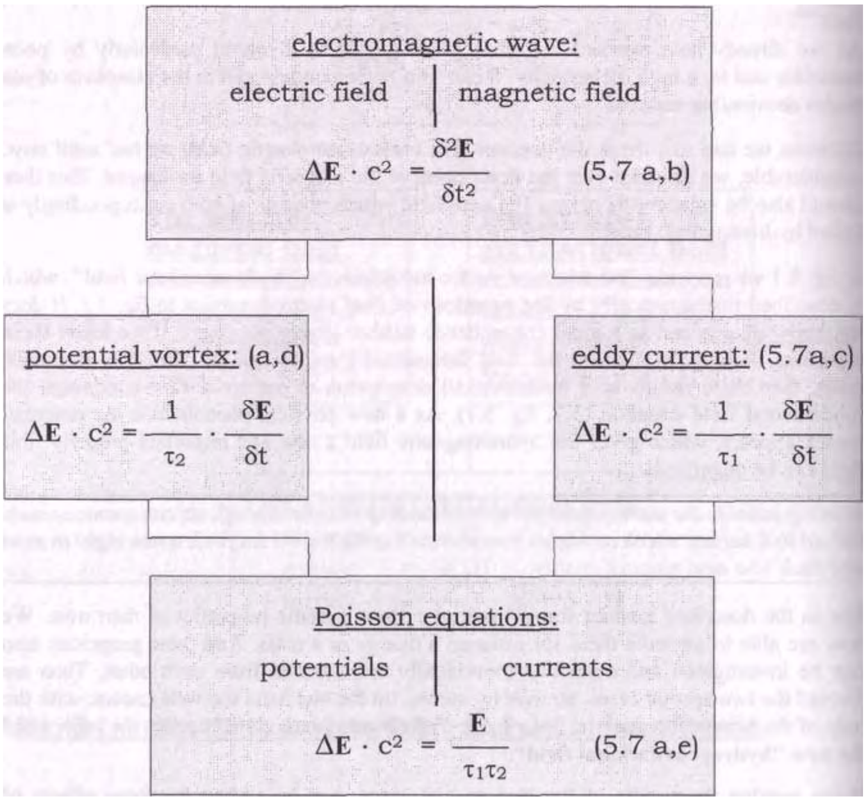
electromagnetic wave:
electric field
magnetic field
$\Delta \mathbf{E} \cdot c^2 = \frac{\delta^2 \mathbf{E}}{\delta t^2}$ (5.7 a,b)
potential vortex: (a,d)
$\Delta \mathbf{E} \cdot c^2 = \frac{1}{\tau_2} \cdot \frac{\delta \mathbf{E}}{\delta t}$
eddy current: (5.7a,c)
$\Delta \mathbf{E} \cdot c^2 = \frac{1}{\tau_1} \cdot \frac{\delta \mathbf{E}}{\delta t}$
Poisson equations:
potentials
currents
$\Delta \mathbf{E} \cdot c^2 = \frac{\mathbf{E}}{\tau_1 \tau_2}$ (5.7 a,e)

Fig. 8.2: Unified theory

*electromagnetic interaction (open field lines)
*gravitation (closed field lines)
*strong interaction (does not exist)
*weak interaction (only special aspect)

The interactions are a result of the
field dependent speed of light!

8.2 Unification of the interactions

The discovery and introduction of the hydromagnetic field makes the desired unification possible, because the electromagnetic resp. Maxwell field, which describes the electromagnetic interaction, and the hydrogravitational field of the gravitation can be derived from this field as a consequence of the formation of quanta.
The kind of the interaction is caused by the course of the field lines of the field quanta which form as spherical vortices: the open field lines make the electromagnetic interaction possible. And the field, lines with a closed course lead to gravitation. Both are a direct result of the field dependent speed of light. A more perfect unification seems hardly possible.

As the next step the unification with the strong and the weak interaction is required, but it could be shown that those don't exist at all. It just concerns misinterpretations with much fantasy, which should help explain the difference between a wrong theory and the physical reality.

Numerous auxiliary terms for the description of the quantum properties exist, like for instance mass, charge or Planck's quantum of action. The prerequisite for their usability naturally is the existence of the quanta. But until these have found to a physical reality, the auxiliary terms are unnecessary. The hydromagnetic field does not know quanta, quantum properties or auxiliary descriptions. It will be shown that, according to expectation, also the temperature is a typical quantum property, which comes within the group of the auxiliary terms. In this way also the temperature is fitted into the unified theory without compulsion.

Without the by us for reasons of usefulness introduced auxiliary terms the fundamental field equation is left with its description of a spatial-temporal principle. If a world equation should exist, then this field equation 5.7 has the best prerequisites. For the fundamental field equation the division in four parts is repeated like already for the hydromagnetic field (fig. 8.1). It likewise consists of four individual parts, the wave (b), the two vortex phenomena (c and d) and the time independent term (e) (fig. 8.2). Whereas the duality still is combined in the wave, it comes to light clearly for the vortices to again be combined in the fourth case. Here arise however potentials and currents, which again can react and oscillate with each other, for instance as L-C-resonant circuit in an electronic circuit, with which the principle is repeated.
This principle is shown clearer for the phenomenon of the temperature as in all other cases. If we start at the top in the picture in fig. 8.2 we have an electromagnetic wave, which is absorbed and thus becomes a vortex. If the vortex falls apart, then eddy losses are formed. We observe that the temperature rises and propagates in the well-known manner. We have arrived in the bottom box, but this again can be taken as the top box for the now following process, because the equation of heat conduction is a vortex equation of type c or d! We discover a self-similarity:

The spatial-temporal principle formulated mathematically by the fundamental field equation can be carried over into itself time and again.

Temperature is the oscillation of contraction of the elementary vortices resulting from the speed of light depending on field strength.

a. at absolute zero temperature:

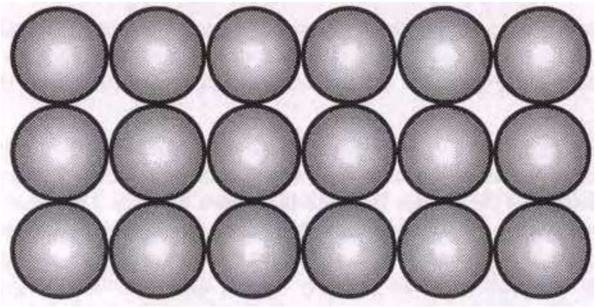

b. if thermally excited:

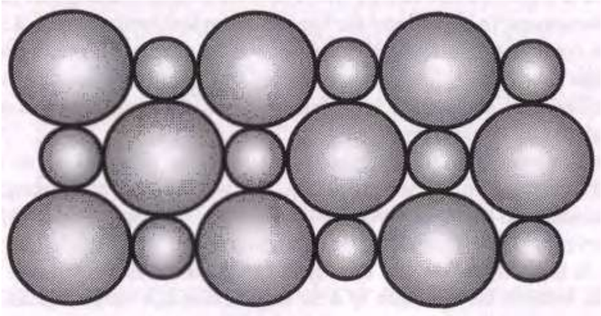

Fig. 8.3: Temperature as an oscillation of size for the speed of light depending on field strength

8.3 Temperature

Following the atomic view, in the case of heat it concerns kinetic energy of the molecules, which carry out more or less violent oscillations. In the case of gaseous materials with this concept, basing on mechanical models, actually successful calculations are possible, like for instance the speed distribution of gases won by Maxwell from theoretical considerations concerning probability.

But the attempt to apply the formulas of the kinetic theory of gases to solids and liquids only succeeds, if additional supplements and improvements are introduced. Since at all events it concerns temperature, thus the same physical quantity, of course also an uniform interpretation should be demanded, which in addition should stand in full accord to the presented design of an integrated theory (TOE).

Against the background of the new theory of objectivity we consider, what happens, if for instance the local field strength is increased by a flying past particle. The matter located at this point is contracted for a short time. By coming closer to each other, the individual elementary vortices mutually reinforce their field and are further compressed. Sometime this process comes to a standstill, is reversed and swings back.

At the same time every single particle, which in this way carries out an oscillation of size, has an effect on its neighbours with its field, to also stimulate these to the same oscillation, but delayed by some time. This phenomenon spreads in all directions. The propagation only will become stationary, if all neighbouring elementary vortices pulsate with the same amplitude. It now should be recorded:

The oscillation of contraction of the elementary vortices we call temperature.

Also this thermodynamic state variable therefore is a result of the variable speed of light. At the absolute zero of temperature no oscillation takes place anymore, whereas the upper limit lies in infinity. Since the cause for temperature represents an oscillation of the local electromagnetic field strength around the cosmic field strength, the following phenomena must be considered as excitation and cause, as dictated by the fundamental field equation 5.7:

1. Electromagnetic waves (b) are able to stimulate matter particles to synchronous oscillations of contraction by their alternating field. In doing so energy in form of heat is transferred to the particles, with the result that their temperature is increased. The wave is absorbed completely, if the thermal oscillation corresponds with the frequency of the wave.

We speak of thermal radiation.

2. But also the two dual vortices, the eddy current (c) and the potential vortex (d) can cause oscillations of contraction. This immediately becomes clear, if we consider a vortex as the special case of the wave, in which the oscillation takes place around a more or less stationary vortex centre. In the case of the decay of vortices, of the transition of energy from vortices to matter, the increase in temperature is measurable.

In the case of this process of diffusion we speak of eddy losses and of loss heat.

Answers to open questions of thermodynamics:

1. Temperature occurs independent of the state in which the matter is (unified theory).

2. Temperature even occurs in solids, where a purely kinetic interpretation fails (unification).

3. Each elementary particle is carrier of a temperature.

4. Expansion with increasing temperature because of the increasing need for room for larger amplitude of oscillation (principle: bi-metal-thermometer).

5. For solids the thermal oscillation of size is primarily passed on by the electrons in the atomic hull. Good electric conductors therefore at the same time also have a high thermal conductivity. (principle: electrical resistance thermometer).

6. For gases the entire atoms carry out this task, for which reason a kinetic auxiliary description becomes applicable.

7. For extreme amplitudes of oscillation the atoms partly or entirely lose their enveloping electrons, when they change into the plasma state.

8. The second law of thermodynamics loses its claim to be absolute and at best reads: with today's technology we are not capable, to design a cyclic working machine, which does nothing else, as to withdraw heat from a heat container and to convert it into mechanical work.

Fig. 8.4: Questions concerning thermodynamics

3. Flying past particles, in particular unbound and free movable charge carriers (e) produce an alternating field for other fixed particles. Doing so kinetic energy can be transformed in temperature, thus in energy of pulsation. A good example is the inelastic collision. But it can also be pointed to numerous chemical reactions. Whoever searches for a concrete example, takes two objects in his hands and rubs them against one another. In that case the particles which are at the frictional surfaces are being moved past each other in very small distance, in this way causing oscillations of pulsation, which propagate into the inside of the objects according to the thermal conductivity. We speak of friction heat.

This model concept provides sound explanations for a whole number of open questions (fig. 8.4), i.e. why the temperature occurs independent of the state (1) and even in solids, where a purely kinetic interpretation fails (2). Every single elementary particle after all is carrier of a temperature (3).
With increasing temperature most materials expand, because the need for room, purely geometrically seen, increases for larger amplitude of oscillation (4). This principle is used in the case of a bi-metal thermometer.
In the case of solids the thermal oscillation of size is passed on primarily by the electrons in the atomic hull (5). Good electric conductors therefore at the same time also have a high thermal conductivity. An example of an application is the electric resistance thermometer.
In the case of gases the entire atoms carry out this task, for which reason a kinetic theory becomes applicable as an auxiliary description (6).
For extreme amplitudes of oscillation the atoms partly or entirely lose their enveloping electrons, when they change into the plasma state (7).
Finally the model concept even limits the second law of thermodynamics, which contains the postulate that it is impossible to design a cyclic working machine, which does nothing else, as to withdraw heat from a heat container and to convert it into mechanical work (8).

8.4 Heat energy

The discussed oscillation of contraction shows two characteristic properties, which must be looked at separately: the amplitude and the frequency.

Temperature describes solely the amplitude of the oscillation of size.

The heat energy however is determined by both,
by the amplitude as well as by the frequency.

Consequently the ideas of temperature and heat energy should be kept strictly apart. It therefore isn't allowed to set this oscillation equal to the electromagnetic wave in tables of frequency.
To be correct two tables should be given, one for the wave, characterized by a propagation with the speed of light, and another one for oscillations of contraction, thus for stationary phenomena and phenomena bound to matter. The latter indeed can likewise propagate relatively fast by fluctuations of pressure in the case of acoustical sound frequencies or by free movable charge carriers in the case of heat conduction, but the velocity of propagation for sound or heat is as is well-known still considerably smaller than the speed of light. Thus an assignment without doubts can be made as to which kind of oscillation it concerns.

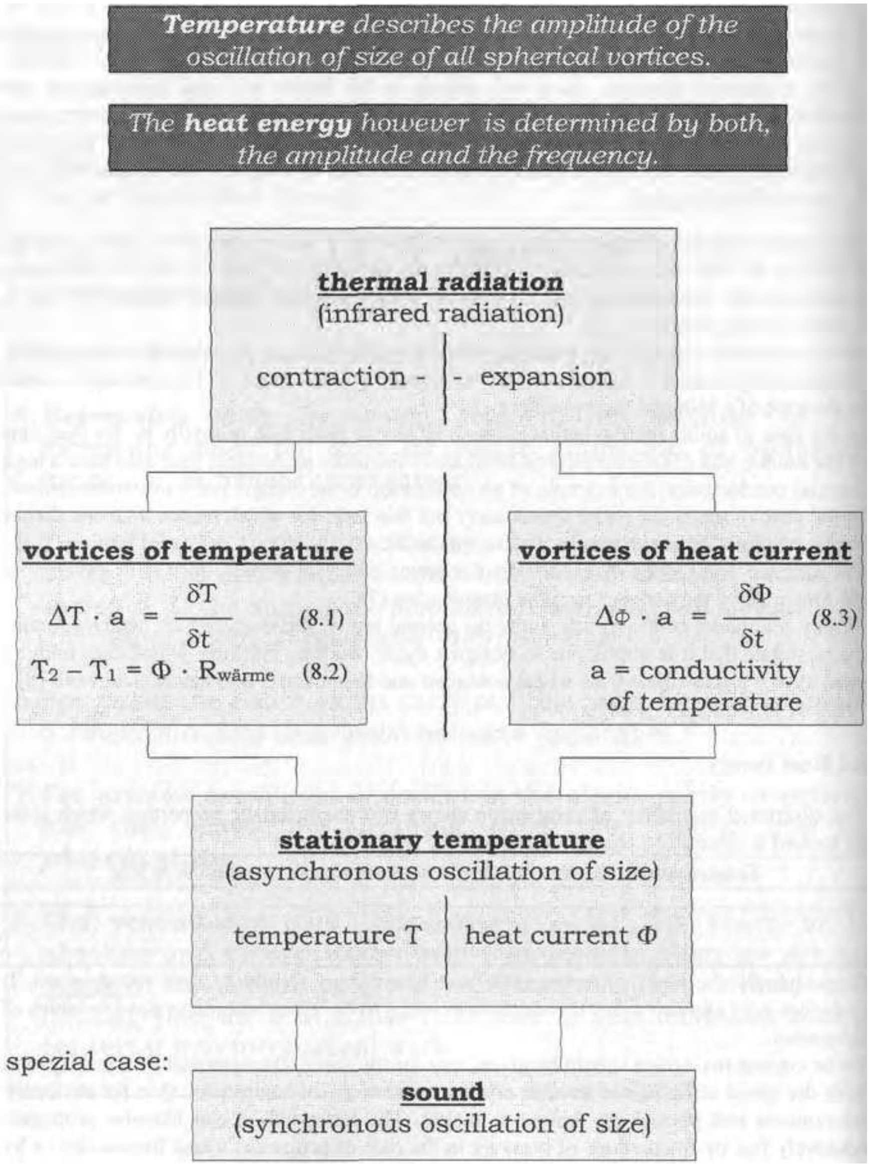

***Temperature** describes the amplitude of the oscillation of size of all spherical vortices.*

*The **heat energy** however is determined by both, the amplitude and the frequency.*

thermal radiation
(infrared radiation)
contraction -
- expansion
vortices of temperature
$\Delta T \cdot a = \frac{\delta T}{\delta t}$ (8.1)
$T_2 - T_1 = \Phi \cdot R_{wärme}$ (8.2)
vortices of heat current
$\Delta\Phi \cdot a = \frac{\delta\Phi}{\delta t}$ (8.3)
a = conductivity of temperature
stationary temperature
(asynchronous oscillation of size)
temperature T
heat current Φ
spezial case:
sound
(synchronous oscillation of size)

Fig. 8.5: Heat conduction resp. sound

8.5 Sound

The close relationship of longitudinal sound waves with the oscillations of contraction of thermally heated matter becomes particularly clear for ultrasound, where the arising heat in the inside of the body which is exposed to sound can be measured directly. The fundamental difference consists of the fact that the produced sound waves not only have the same frequency, but also the same phase, what needs not be the case for the temperature. The apparently uncoordinated occurring oscillations of size of the temperature, which as a rule occupy more space if the intensity increases, form a "thermal noise".

The oscillation of size with the same phase is not realizable at all in a spatial formation of particles, with one exception, the case that all particles expand and afterwards again contract simultaneously and in the same time. We can observe such a synchronization of the pulsation oscillations of all elementary vortices in the case of a pulsar. For us a pulsar looks like a "lighthouse" in space which shines with a fixed frequency. In reality it as well can concern a constantly shining sun, which carries out a synchronized, thermal oscillation of size, like a gigantic low-frequency loudspeaker. During the phase of contraction of the star its emitted light stays back. To us the pulsar looks dark. In addition the field strength is extremely increased and the light becomes correspondingly slow. During the phase of expansion the conditions are reversed and we observe a light flash. Exactly the pulsar unambiguously confirms the here presented theory of the variable, field dependent speed of light.

The well-known fact that the microcosm represents a copy of the macrocosm, already suggests that each atom is capable of the same oscillation of size as a pulsar: if next to the oscillating atom a resting one is placed, then does this one see a smaller field during the phase of contraction because of the increasing distance. It hence becomes bigger itself. If the pulsating neighbouring atom afterwards expands, it however becomes smaller. The at first resting atom in this way becomes a "pulsar" oscillating with opposite phase. The oscillating atom has stimulated the neighbouring atom as well to an oscillation of size, and this process will be repeated with the closest neighbouring atom. We speak of heat conduction.
To which extent the average distance between neighbouring atoms is influenced while a material is heated, solely depends on the structure of the atomic lattice. For matter with a fixed lattice according to expectation a smaller heat expansion will occur, as for the unordered structure of gases, in which we find confirmed well-known relations.

In a for potential vortices characteristic property sound waves and thermal waves of contraction correspond:

The propagation of potential vortex fields takes place as a longitudinal wave.

In this point vortex fields clearly differ from the transverse propagating electromagnetic waves!

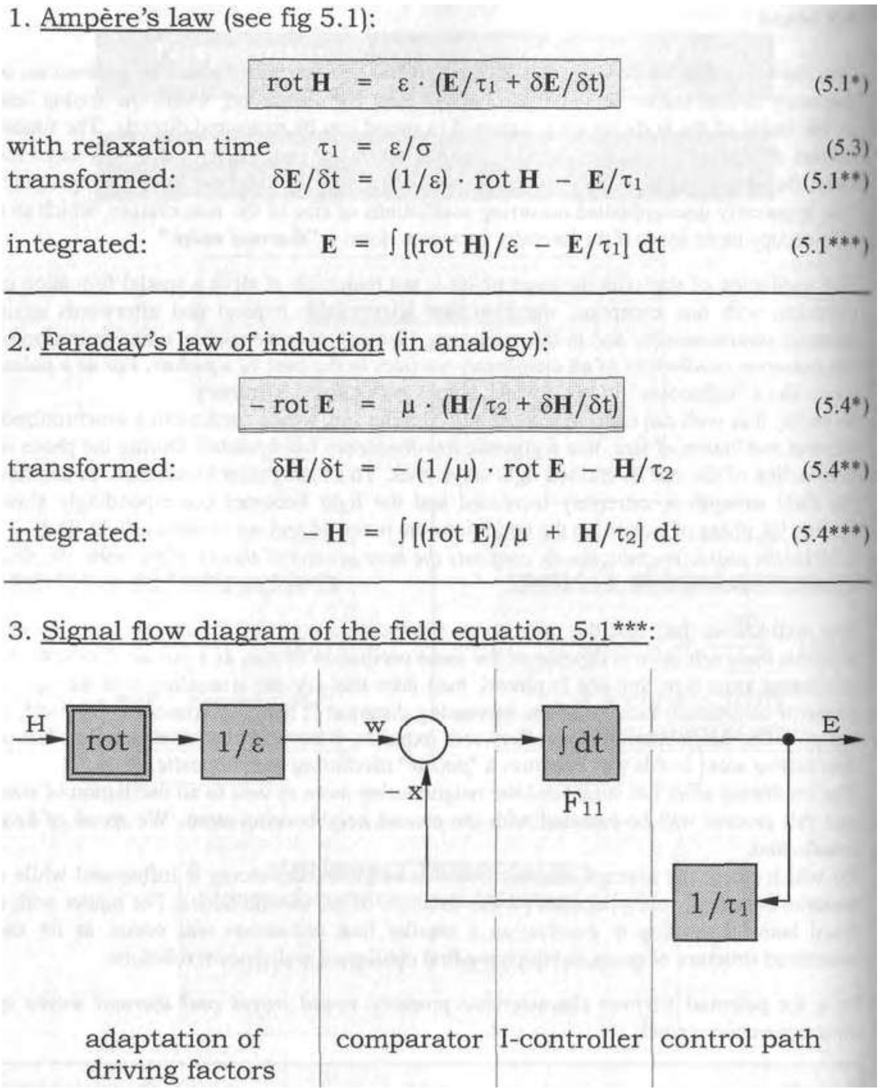

1. Ampère's law (see fig 5.1):

$$\boxed{\operatorname{rot}\mathbf{H} = \varepsilon \cdot (\mathbf{E}/\tau_1 + \delta\mathbf{E}/\delta t)} \tag{5.1*}$$

with relaxation time $\tau_1 = \varepsilon/\sigma$ (5.3)

transformed: $\delta\mathbf{E}/\delta t = (1/\varepsilon) \cdot \operatorname{rot}\mathbf{H} - \mathbf{E}/\tau_1$ (5.1**)

integrated: $\mathbf{E} = \int[(\operatorname{rot}\mathbf{H})/\varepsilon - \mathbf{E}/\tau_1]\,dt$ (5.1***)

2. Faraday's law of induction (in analogy):

$$\boxed{-\operatorname{rot}\mathbf{E} = \mu \cdot (\mathbf{H}/\tau_2 + \delta\mathbf{H}/\delta t)} \tag{5.4*}$$

transformed: $\delta\mathbf{H}/\delta t = -(1/\mu) \cdot \operatorname{rot}\mathbf{E} - \mathbf{H}/\tau_2$ (5.4**)

integrated: $-\mathbf{H} = \int[(\operatorname{rot}\mathbf{E})/\mu + \mathbf{H}/\tau_2]\,dt$ (5.4***)

3. Signal flow diagram of the field equation 5.1***:

H
rot
1/ε
w
− x
∫dt
F11
E
1/τ1
adaptation of driving factors
comparator
I-controller
control path

Fig. 8.6: Control technical analysis of the dual equations of the hydromagnetic field.

8.6 Basic principle of cybernetics

Surely can be attributed also information to the potential vortex. But how should information be formed? Is information a form of energy? Energy occurs as a consequence of the formation of potential vortices. Without this phenomenon there wouldn't be any energy!

Can information be described by means of a mathematical equation?

To be able to answer these questions, we subject the fundamental field equation to a control technical analysis. If it actually concerns a world equation, then an answers should be possible.

We again take up Ampere's law 5.1* from fig. 5.1 and remodel it according to the time derivative (5.1**). If the equation now is integrated over the time (5.1***), a signal flow diagram can be drawn (fig. 8.6).

The structure of a regulatory circuit is clearly visible. The individual paragraphs are described in an analogous way as for a technical control system. The execution of the curl operation on the field pointer of the magnetic field strength H and the multiplication with $1/\varepsilon$ accordingly form an adaptation of driving factors. In the comparator the difference for control from driving factor w and controlling factor x is formed and supplied to an integral controller. The control path has a purely proportional behaviour and consists of the processing of the measurement value of the electric field strength E with $1/\tau_1$, in which τ_1 describes the relaxation time of the eddy currents.

In technical control systems such a structure is found remarkably seldom, although it has an invaluable advantage: it possesses a stability in principle. Not a single adjustment of the controller exists, in which the closed regulatory circuit could become unstable, because it shows a proportionally delaying behaviour of first order. Possible changes of the adjustment of the controller or of the control path merely take effect on the speed, with which the regulatory circuit is able to follow changes of the driving factor.

This control technical basic principle convinces by its simplicity and efficiency. It meets us again in identical form in the second field equation 5.4*, the extended Faraday's law of induction. In dual formulation the electric field strength now appears as input factor and the magnetic field strength as output factor. Both regulatory circuits are coupled and connected with each other, by deriving their driving factor each time from the controlling factor of their dual partner. Is this structure actually efficient and meaningful?

Every regulatory circuit needs a target value, which is dictated from the outside. Let us think of the numerous control systems in nature. At all events a higher intelligence would be necessary for all the target values. This problematic is comparable to the question, what existed first: the egg from which a hen hatches or the hen without which no eggs can exist. Without a given target, evolution would not exist.

The connected regulatory circuit structure provides the matching answer: cybernetic systems, which usually and as is well-known strive to a state of balance, get their target value from their dual "partner". It is crucial that correspondingly dual systems are self-sufficient and can form and develop independently out of themselves without target values of a third side. This basic principle of cybernetics undoubtedly is brilliant.

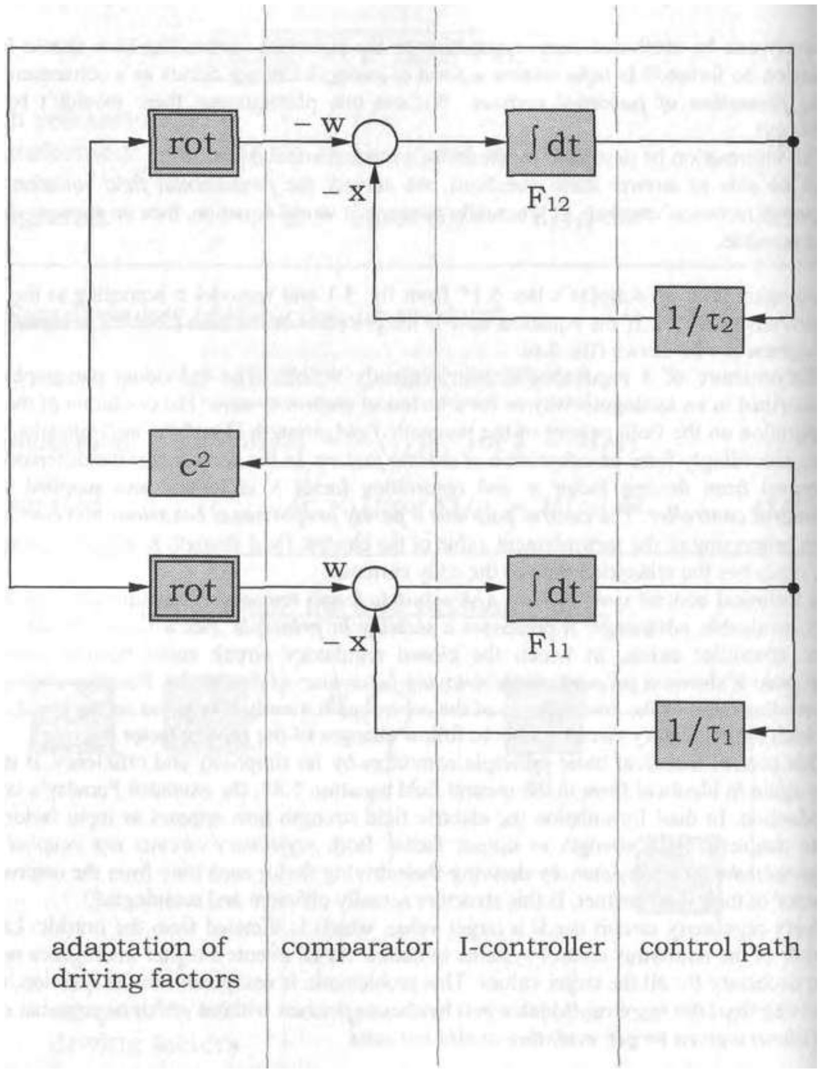
rot
− w
− x
∫dt
F₁₂
1/τ₂
c²
rot
w
− x
∫dt
F₁₁
1/τ₁
adaptation of driving factors
comparator
I-controller
control path

Fig. 8.7: Signal flow diagram of the fundamental field equation
with adaptive structure.

8.7 Adaptive regulatory circuit structure

If out of the nowhere something like the cosmos or like life on earth should form, then the connected regulatory circuit structure basing on duality probably is the only possible and conceivable. Thus it merely concerns the control technical representation of the fundamental field equation.

The question for the efficiency not only concerns the stability, but equally the possibility of both systems, to oscillate and to communicate with each other by the coupling and the associated exchange of information.

Fig. 8.7 shows the signal flow diagram of both regulatory circuits. These are switched in line and form a coupled circuit, which itself can be interpreted as a third regulatory circuit. Also this one shows a change of sign in the circuit like the other two circuits.

The information technical interpretation could turn out as follows: information about a regulatory process in the lower regulatory circuit F_{11} caused for instance by a disturbance is communicated over the coupled circuit to the upper regulatory circuit F_{J2}. In this case F_{11} acts as transmitter and F_{12} as receiver of the information. Afterwards both exchange their places, because F_{12} for its part reacts by a regulatory process and reports to F_{11}. The regulatory circuits adapt to each other. Obviously it concerns the basic structure of an adaptive regulatory circuit.

To analyse the coupled circuit the examination of individual special cases is recommended. If the regulatory circuits F_{11} and F_{12} are opened up in the way that the time constants tau_1 and tau_2 go towards infinity, then the double integral effect is left. Analyses of technical regulatory circuit teach us that such systems always tend to instability. Because in addition the target value is zero, an oscillation around zero will arise, which we call electromagnetic wave.
If one of both time constants becomes finite, e.g. τ_2, then damping of the waves will occur. The "subordinate" cascade regulatory circuit F_{12} will adjust itself and now has a proportional delaying behaviour of first order. Together with the integral controller of the open F_{11}- circuit the coupled circuit will show the typical and more or less optimal regulatory behaviour of a damped oscillation.

These special cases correspond with the mathematical (fig. 5.2) and the physical (fig. 5.3) interpretation of the fundamental field equation. In addition a spatial rotation, a swirling will occur because of the double execution of the curl operation.

If interpreted control technically then vortices are the temporally stable, spatial swing of a field pointer around a centre, the vortex centre.

Without potential vortices no stability, no matter, no energy nor information would exist!

As can be looked up in Goethe's Faust, it always has been a desire of humanity, to find out, "what keeps the world together in the heart of hearts".

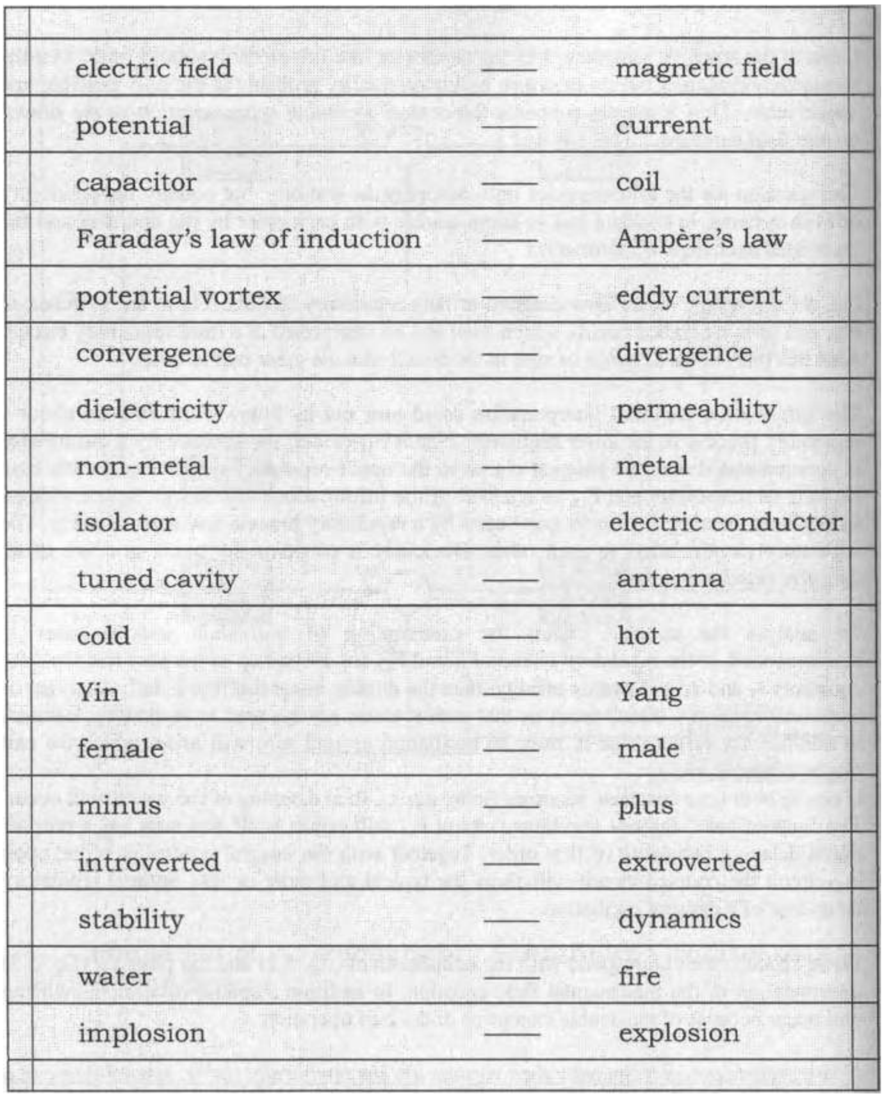

electric field	—	magnetic field
potential	—	current
capacitor	—	coil
Faraday's law of induction	—	Ampère's law
potential vortex	—	eddy current
convergence	—	divergence
dielectricity	—	permeability
non-metal	—	metal
isolator	—	electric conductor
tuned cavity	—	antenna
cold	—	hot
Yin	—	Yang
female	—	male
minus	—	plus
introverted	—	extroverted
stability	—	dynamics
water	—	fire
implosion	—	explosion

Fig. 8.8: Table of dual correspondences (as a supplement to chapter 3.3 and fig. 4.2).

8.8 Information

The search for an answer for numerous philosophers and physicists was tantamount to the search for a world formula. Of course mustn't be forgotten that a formula only is a mathematical description and never the physical reality itself. It is a mathematical tool in the hand of a person and not the world or the cosmos itself, which he tries to understand. What keeps the world together in the heart of hearts, has to be more than only a pure apparatus of formulas. Actually the fundamental field equation tells us more. It reveals us a basic principle basing on duality in which the dual partners mutually dictate target values and goals. This principle convinces by its simplicity and efficiency. Apart from the "self regulation" it obviously also has the fundamental possibility of a "self organization" and the "generation of information". The field equations of the hydromagnetic field thus are the starting-point for the formation not only of matter and energy, but also of information. Accordingly holds:

Information is nothing but a structure of electromagnetic vortex fields!

This statement is new and to a large extent incompatible with the conception world of Norbert Wiener, who goes as the founder of cybernetics. From N. Wiener stems the sentence: "information is information, not matter and not energy".

We hold against it that obviously a fairly direct connection exists. We have worked out that only the vortex can show a stable adaptive regulatory circuit structure. Only the vortex and not the wave exists in two forms of formation dual to each other, and the principle of duality again is the prerequisite for the formation of information, of self organization and finally for the evolution. In fig. 8.8 well-known dual partnerships are listed. From it follows in a consistent way that for the production of information without exception the electromagnetic vortices should be considered.

But how can this so important duality occur, how can it form? This question is closely associated with the question of the formation of vortices. The signal flow diagram (fig. 8.7) to that says that the dual regulatory circuits F_1 and F_2 can only exist by the coupled circuit, which provides them the necessary target values and at the same time forwards the respective information. In this way of the oscillations and the more or less damped wave F_1 and F_2 communicate with each other.

The electromagnetic wave serves solely the mediation of information and energy.

With that falls a central role upon the wave, so that vice versa is valid:

Without wave no vortices, no duality and
consequently no evolution can exist.

According to the to date state of knowledge the basic principle of cybernetics forms the basis for matter and energy as well as for information. Since the wave can only serve the transmission of information, the principle of duality and the vortex will function as carriers of information. We are entitled, to speak of vortex information, this by no means is characterized by special frequencies or modulations of frequencies. This is prevented by the property of the vortices which allows them to change the frequency. On the other hand various configurations of vortices are possible and numerous combinations and modulations are conceivable.

If technical apparatus generate vortices, then they produce information. Here a serious danger with regard to the environmental compatibility can not be excluded!

Equivalent things:

- * elements of the fundamental field equation 5.7
- * elements of the Greek philosophy of nature
- * temperaments
- * (impulses of growth)

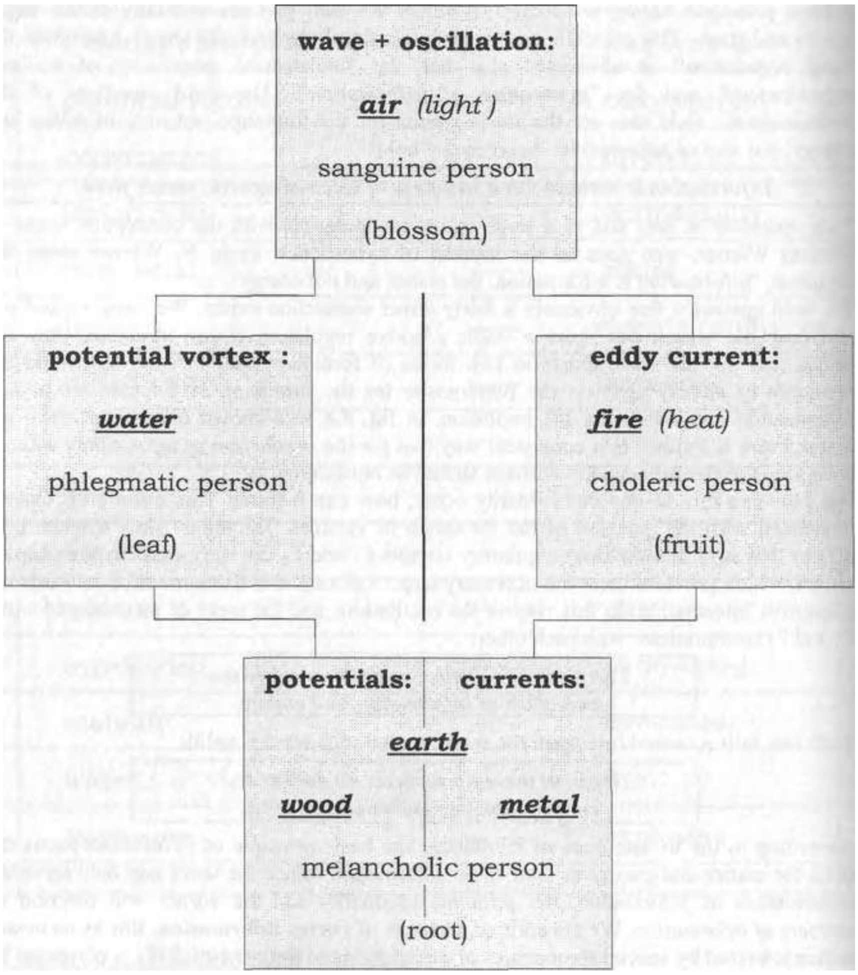
wave + oscillation:
air (light)
sanguine person
(blossom)
potential vortex :
water
phlegmatic person
(leaf)
eddy current:
fire (heat)
choleric person
(fruit)
potentials: currents:
earth
wood metal
melancholic person
(root)

Fig. 8.9: The theory of four elements of the old Greek philosophy of nature (Aristotle and others)

8.9 Philosophy of nature

Seen in the view of the philosophy of nature now two dual points of view are possible.
The optimistic one would be:
We and our environment on the one hand are a result of the cybernetic principle and on the other hand of our observation point of view which should be valued relativistically. If really everything should be electromagnetism, a phenomenon which can't be grasped directly by humans, then the pessimist would come to the conclusion: everythins is nothing. What we observe is nothins but a deception of the senses. Perhaps therefore famous philosophers of antiquity, like Empedokles or Demokritos have ended their life in the crater of the Etna. According to the theory of the atom of Demokritos (470 to 380 B.C.) the formation of matter, earth and celestial bodies will occur by means of formation of vortices!
Empedokles (482 to 420 B.C.) was the first to develop a theory basing on four elements, which was continued and improved by Plato (428 to 348 B.C.) and Aristotle (384 to 322 B.C.). Accordingly these elements are changeable into each other and mixable with each other. From them all bodies are build up.
The terms "air, water, fire and earth", with which the philosophers have described the four elements, are of course not identical with the ones in our translation and conception world, but they were used in a philosophical sense as a substitute for the description of the respective basic principle.
There also have been different approaches, to translate these terms differently, e.g. by an assignment to the four states of matter (solid, liquid, gaseous, plasma). But the ancient texts don't get easier to read in that way.

Fig. 8.9 shows the obvious assignment to the four building parts of the fundamental field equation 5.7. It would be worth an attempt, to exchange the terms in the translations of ancient texts and to translate air with wave, water with potential vortex and fire with eddy current. The term earth has two sides, which should be translated with potential instead of wood and current instead of metal.

Let's try the translation this way with the theory of Plato<i>, by correspondingly translating anew the talk of Timaios about the formation of the world. The perception of smell then is described as follows: "...as the potential vortex turns into waves (or) the wave into potential vortices, the smells are formed during this transition, and smells are smoke or fog. But fog is the transition of waves into vortices, the transition of the vortex into waves however smoke".
Plato here provides an indisputable and conclusive interpretation of the fundamental field equation. In this equation the potential vortex acts as damping term in the wave equation, what in the case of waves rolling up to vortices will show to the observer in the way that the electromagnetic waves and therefore also the light will be damped. We say, the visibility gets worse and speak of fog. If the damping phenomenon disappears again, as the potential vortices break up, then Plato speaks of smoke.
Numerous ancient texts, which until now only could be "interpreted" philosophically, in this way turn out to be a rational textbook description of natural scientific phenomena. They anyway only get readable and understandable for the general public with the modern technical terms.

<i>: Platon: Samtliche Werke 5, Rowohlts Klassiker Nr. 47, S. 188, 66e

Optimist:

We and our environment on the one hand are a result of the cybenetic principle and on the other hand of our observation point of view which should be valued relativistically.

Pessimist:

Everything is nothing. What we observe is nothing but a deception of the senses.

Plato[i] (talk of Timaios about the formation of the world) description concerning the perception of smell:

"...as water (the potential vortex) turns into air (waves) or air (the wave) into water (potential vortices), the smells are formed during this transition, and smells are smoke or fog. But fog is the transition of air (waves) into water (vortices), the transition of water (vortex) into air (waves) however smoke"

changed translation with the technical terms:

air = wave
water = potential vortex
fire = eddy current

As a consequence:

smell is vortex information

Fig. 9.0: The explanation of Plato concerning the formation of smell, smoke and fog[i]

---

[i]: Platon: Samtliche Werke 5, Rohwohlts Klassiker Nr. 47, S. 188, 66e

9. Usage

If the newly discovered vortex phenomenon of the vortex of the electric field exists, then it will be possible to practically use it. Whereas we still think about possibilities for technical usage, there by all means exists the possibility, that nature already is successfully using the vortex for a long time. We should look precise at things. We can only learn of nature!

Remarkable about the passage of Plato (fig. 9.0) is not only the fact, that the potential vortex already was known for two and a half thousand years and was taken into consideration for an interpretation, but also the realization of Plato, that during the described transition the smells form. Smell thus would be a vortex property!

After all vortices are able to bind information as can be inferred from the basic principle of cybernetics. With this vortex property and the statement of Plato smell obviously would be nothing else than pure information which by potential vortices is stored, distributed and eventually is caught by the hair cells for smell of the nose.
If now a dog takes up a track, does it then run after vortices which remained behind or does it analyse, according to the encyclopaedia, the chemistry left behind, or does it combine both? Does the bloodhound for instance interpret the vortical oscillations of chemical substances like lattice oscillations or the movements of the electrons in the atomic hull? A lot of work awaits the research scientist of vortices here. The seminar will offer opportunity, to deepen this topic<i>.

Only if technicians exist, who put an electronic box on the table with a button, at which they adjust the wanted scent of rotten eggs to lavender, we honestly can claim to have understood the phenomenon of the smell.

For the majority of the people a theory only wins its sense by the practical applicability, and therefore we'll have to develop and present a technical usage on the basis of the vortex theory.

<i>: Suggestions for appropriate seminar themes: the meaning of smell, taste, the aroma therapy, the homeopathy or the effect of aetheric oils. (To that in part 3, INDEL Verlagsabt. 2003)

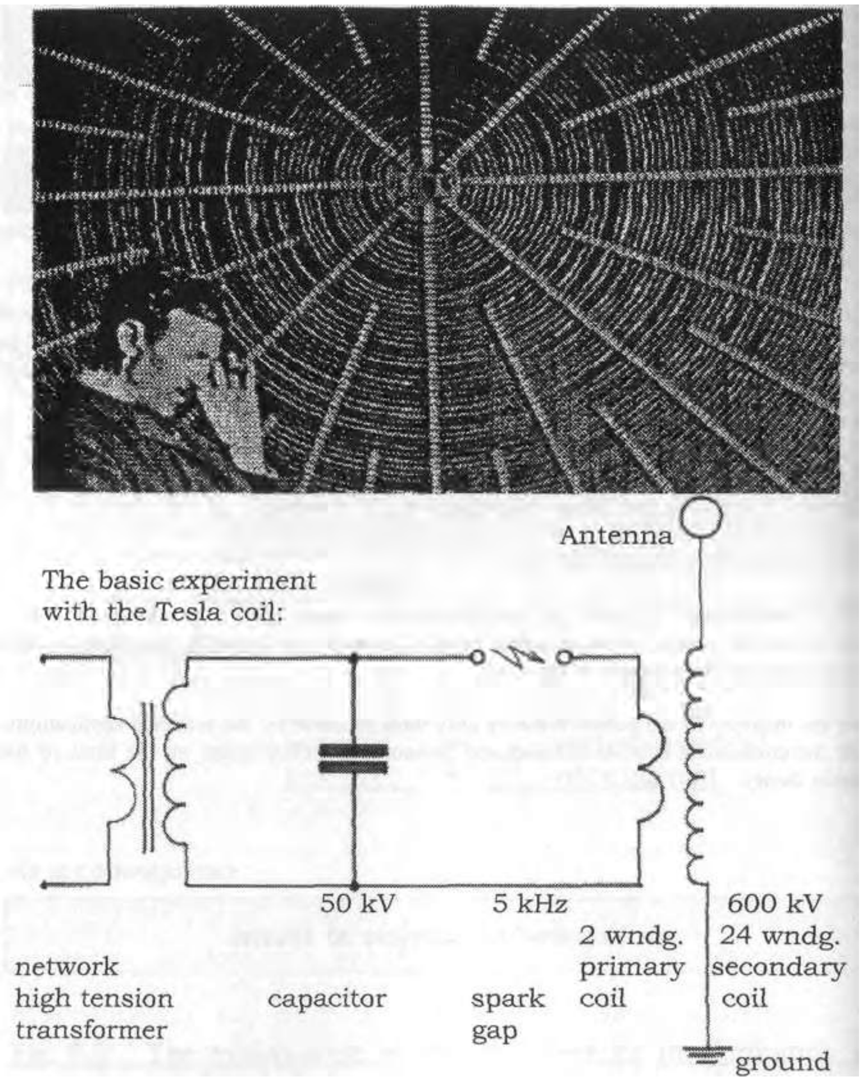
Antenna
The basic experiment
with the Tesla coil:
50 kV
5 kHz
600 kV
2 wndg.
primary
coil
24 wndg.
secondary
coil
network
high tension
transformer
capacitor
spark
gap
ground

Fig. 9.1: High tension transmitter of Nikola Tesla

9.1 Longitudinal electric waves

It is important, that experts struggle for problem solutions. Only if the discussion about limits doesn't come to a result which can be grasped and verified, then also in the environmental compatibility the willingness will grow, to accept a not yet known phenomenon as a further and possibly crucial influential factor.
Already for a hundred years there has been a dispute of experts. At that time scientists all over the world were trying to verify the experiments of Heinrich Hertz. Then from America the message came, Hertz had been mistaken and the electromagnetic wave would have completely other properties. The scientists in Europe were indignant, but they had to take the message seriously, because it after all came from the important experimental physicist Nikola Tesla (1856-1943), who with his inventions of the rotary field and of the asynchronous motor has stamped today's electric energy technology as no other.

As a result Lord Kelvin boarded a steamship as a mediator and sailed 1897 to New York to convince Tesla from the opposite. But the experiments, which Tesla presented his Lordship, didn't give rise to any doubts<i>, and thus Kelvin returned to Europe with the message: "Both are right, Tesla as well as Hertz! Whereas the electromagnetic wave which Hertz has detected, is a transverse wave, does Tesla work with a longitudinal wave"!
Lord Kelvin as a result started to draw most different vortex models, because it was clear to him, that a propagation as a longitudinal standing wave analogous to the sound wave only is conceivable, if quantized structures exist, which knock each other mutually. Kelvin therefore assumed vortex structures of the electromagnetic field. His vortex models were published and landed in the curiosity box of orthodox science.
Heinrich Hertz did have a big advantage. He could refer to Maxwell and calculate his wave with the field equations. For modern technology the mathematical calculability is almost an indispensable prerequisite!
For Tesla wave there however neither was a mathematical nor a physical theory. The only thing Tesla had, were presentable experiments.
In Colorado Springs he had build a 10 kW transmitting installation and lighted 200 fluorescent lamps of 50 Watt each on a mountain in the Rocky Mountains in a distance of 25 miles. With that he had completely transmitted the transmission power of 10 kW, as can be inferred from the press reports at that time. With Hertzian waves, which propagate spatially, this experiment even today, after over 100 years, wouldn't be realizable technologically. According to the law of the square of the distance one isn't even able to let glow a tiny little lamp in such a distance.

For sure his rotary field theory was a big help for Tesla in all experiments. Actually a rotary field can be seen as the special case of a planar vortex.
Thus Tesla obviously was able, to use the potential vortex without even knowing it. Tesla has stimulated a loosely coupled high tension coil wound like a spiral to self-resonant oscillations and emitted the produced vortices over an antenna (fig. 9.1). On the receiver side the process was then reversed.

<i>: N. Tesla: III. The Singular Misconception of the Wireless (Famos Scientific Illusions), Electrical Experimenter, Feb. 1919, p. 732.

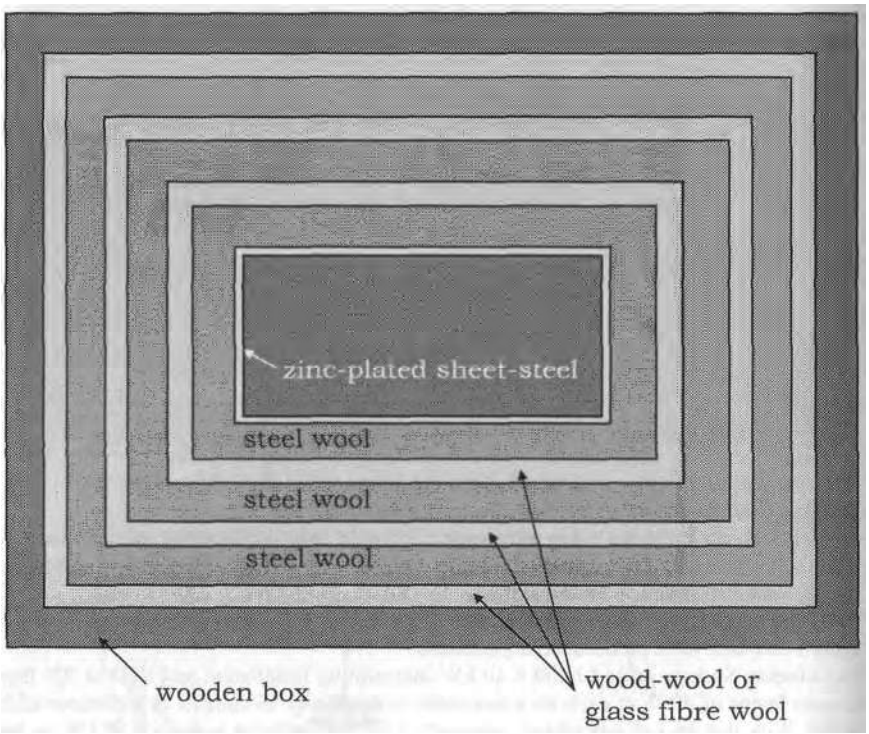
zinc-plated sheet-steel
steel wool
steel wool
steel wool
wooden box
wood-wool or
glass fibre wool

Fig. 9.2: Orgone accumulator according to Wilhelm Reich.<i>

<i>: Dr. James DeMeo: The Orgone Accumulator Handbook, ISBN 0-9261855-0-7

9.2 Medical technical usage

Nikola Tesla at his time was extremely popular in the area of medicine. With his inventions injuries were cured and pain eased. Modern diathermy goes back to his work. But Tesla at that time has treated the patients with vortices, whereas today, possibly out of ignorance, electromagnetic waves are used. These however have in contrast to potential vortices only a small penetration depth. Today in addition only the effect of heat is considered and in no way the information technical effect of electric vortices. Here we are missing the comparison, to be able to say. if the treatment with a Tesla apparatus was more successful than with modem microwave radiators.

The experiments of Wilhelm Reich (1897-1957) aimed in the same direction. Exactly as Tesla 55 years before also Reich emigrated to America in 1939. He had specialized in catching and collecting vortices, which he called "orgone", from the air. In this way he could store up high energies over time and then use them purposeful. With his "orgone accumulator" he was able to cure wounds faster and to ease pain. He also treated cancer and a lot of forms of anaemia. Technically Reich could demonstrate, that charged capacitor plates discharge faster under the influence of his apparatus. His orgone accumulator for instance is a wooden box, filled with wool of steel and wood, which alternate in layers. He said that the reason for this construction is, that metals repel the orgone energy, whereas organic materials become charged and accumulate it. That holds in his opinion to a particularly high degree for water.
Reich concludes that no material and no device are known, with which the orgone energy could be shielded. This with regard to the environmental compatibility of the vortex phenomenon should make us thoughtful.
As a supplement it should be mentioned, that Reich already at the end of the forties has pointed to the dying of the forests and he has made a direct connection to the orgone energy. Reich was slandered and mocked and died in the prison of Connecticut, while his writings were burned publicly.

Not much better fared the Austrian forester Viktor Schauberger (1885-1958), who also can be described as visionary. He was able to produce water with a particularly good conductivity. His water in addition goes for medicinal and healthy. Also to him travelled cancer patients of far away to get this water.
Schauberger spoke of the natural treatment of the water, whatever he meant with that. In any case he build with great success installations to transport wood by floating it and even installations to mine ore with so-called double twist pipes, which made possible a transport without friction even of materials, which are heavier than the "means of transport" water. It could be proven that no contact with the pipe took place. This was scientifically investigated and confirmed at the university of Stuttgart 1952 by Prof. Popel. Quite obviously in the double twist pipe vortices have formed, which have bound the material to be transported (see phenomenon of transport).
Today you can buy levitated water at over 100 selling points in Europe, of which is claimed, that it is prepared according to instructions of Schauberger. Unfortunately we aren't able to ask him anymore, if he agrees with such a "centrifuged" water. Thus the test of the effectiveness has to be left to everyone himself or herself.

A: Concerning the functioning of the Faraday generator $(E = v \times B)$:

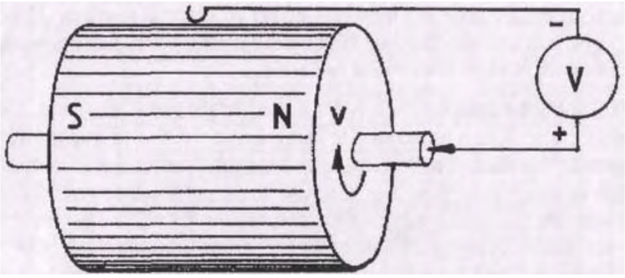

B: Construction of the flying disc:

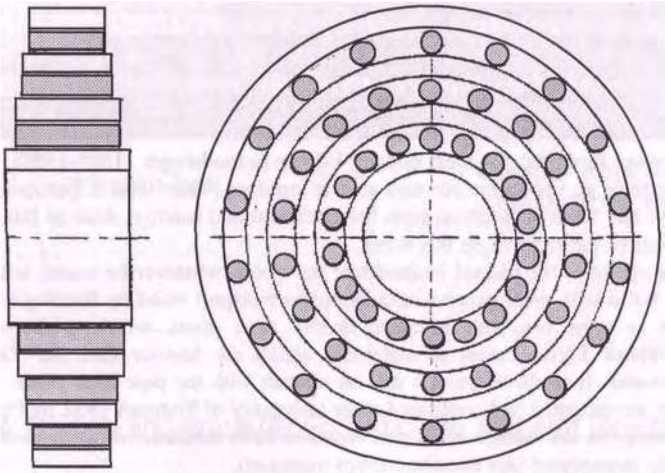

Fig. 9.3: Applications of the Faraday generator<i> in the flying disc of Prof. John Searl

9.3 Flying objects

The medium of Viktor Schauberger always was the water. He could produce light effects and very high temperatures with it with only minimal excitation in the form of impulses. An installation, which had been built according to his plans at the company of Siemens, eventually melted at 4000°C by improper operation, as he himself stated. According to that the machine would have released more energy than used, thus a perpetuum mobile. The authorities of the state in the Third Reich were impressed and put Schauberger in a concentration camp, where he was instructed to build a "flying saucer" under permanent supervision together with fellow prisoners. For him it's all about the concentration effect and the usage of the as a result occurring implosion in contrast to our today's " technology of explosion and fire" as Schauberger was accustomed to express himself. It is not known, if Schaubergers "repulsine" ever has took off.

A disc, which has proven its suitability to fly, was constructed and built by the english technician John R. R. Searl, although he said he can't explain the effect. He also had big difficulties, to get the apparatus under control. A disc broke through the ceiling and the roof of his laboratory and disappeared to never be seen again. Five other flying discs, which he after this experience started in the open, went lost in the same manner. Without knowing the effect, he of course neither could assess the dangers. His experiments have claimed serious injuries and a casualty. While he 1985 was put in prison under a pretext, his laboratory and his house were burned down and all documents destroyed. Now he works on a technical usage. According to the principle it concerns a Faraday machine. Thereby a permanent magnet, magnetized in axial direction, is turned. Now the magnetic induction B for a relative velocity v is measured as an electric field strength E according to equation 6.10: $E = v \times B$. (6.10)

Because the vectors are arranged as standing perpendicular to each other, will arise a tension voltage in radial direction (direction of E-field) which can be taken off. Apart from the friction no further force of reaction occurs in the case of the Faraday generator. Because of the small gain of energy until today no application ready for the market exists. But this principle, to convert magnetic field energy into electric, already has moved the nature of many inventors. Professor Searl has reversed this old principle. Whereas normally the component of the velocity towards the centre of rotation decreases, it increases in Searls case. For that he works with roller pivoted concentric rings, which he drives by a small electric motor (fig. 9.3). Doing so something inexplicable for him happens: After switching off the motor the revolutions per minute don't decrease again, but increase audibly and increase further, until the produced electric field shows the well-known high tension phenomena: corona discharges, formation of ozone, ionisation of the air and production of a vacuum in the inside of the disc.

The rings and rollers consist of several layers, which are built up similar to a bi-metal. The only explanation I can think of is that a change in structure would occur as a result of the physical length contraction which is caused by the increase of the E-field in the direction of the centre of rotation. The bi-metals try to withdraw themselves from this change by an increase of their rotation of their own (fig. 6.5). To compensate the field the disc builds up a spin, as also the elementary particles do (fig. 6.13). While the formation of vacuum prevents sparking in the inside of the flying disc, and the revolutions per minute further increase because there is no air friction whatsoever, the disc weighing 5 tons all of a sudden lakes off the ground and according to reports of eyewitnesses shoots vertically upwards.

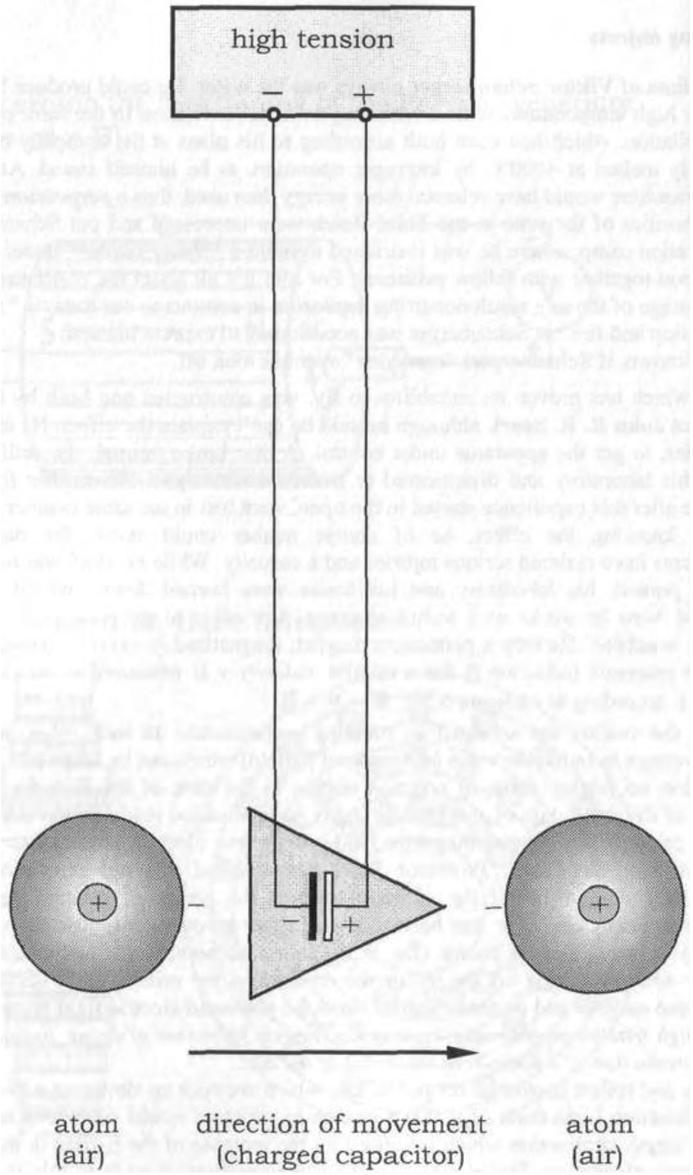
high tension
−
+
+
−
−
+
−
+
atom
(air)
direction of movement
(charged capacitor)
atom
(air)

Fig. 9.4: Capacitor experiment concerning the so-called electro-gravitation according to Prof. Biefeld / Dr. Brown

9.4 Electro gravitation?

Wild speculations circulate about the mechanism of the flight of a "flying saucer", which should function without sound and without combustion, and for which no sound barrier exists, as pilots have observed. The talk is about cancelling gravitation or about an "electro gravitation".

Reproducible in any case is an experiment, which has been systematically investigated by the american professor Biefeld and his pupil T. Brown from 1925. Accordingly does a body charged to high tension show a force effect in the direction of its positively charged note. As a check for oneself a capacitor can be suspended on its connection wires and it can be observed, how it moves visibly in the direction of the positive pole in case it is charged (fig. 9.4).

Because the Biefeld-Brown effect depends on direction, it actually can't concern gravitation at all but merely a not understood electromagnetic interaction. In the neighbourhood of the capacitor pole air molecules are found, and their negatively charged atomic hull is attracted by the positive pole. At the same time the atomic nucleus is repelled. By that the air atoms in the presence of the high tension capacitor become unsymmetrical, causing the force of attraction due to the smaller distance to exceed the force of repulsion. At the negatively charged end the conditions are exactly reversed. There a repulsion occurs. Because the polarized air atoms and molecules are sucked in, no pressure can build up and as a result no sound barrier can occur. Experiments with charged and uncharged rockets have brought the interesting result, that the electrically charged rockets flew many times as far as the uncharged ones (5-6 times as high)<i>.

Many a person now perhaps starts to dream of the flying carpet but, as said, it isn't an "effect of antigravitation". Does "free energy" actually exist, we have to ask ourselves? From the gravitation the soviet physicist Landau calculates an energy density of 16 megawatt hours per cubic meter for the earth's surface. Immediately inventors are found, who want to use this gravitational energy.

Nikola Tesla in his laboratory in Manhattan has incidentally built resonators, of which he could bring all electric, magnetic and mechanic factors in resonance. On an evening stroll he fastened a battery operated vibrator to the tubular steel scaffolding of a new building and let everything shake and wobble. In his laboratory such a device once got out of control by inattentiveness and triggered an earthquake. In that way the road surfacing and pipes were burst and window panes got broken. The police penetrating his laboratory only could see, how Tesla forcible finished the experiment with a sledge-hammer.

The experiments which got out of control of Tesla, Searl and Schauberger have one thing in common: it concerns constructions with an unipolar arrangement of the field.

Tesla had arranged the magnetic field in a unipolar way, as he has reported himself, Searl had realized electric unipolar fields in a construction similar to the electron, and Viktor Schauberger had specialized in producing unipolar structures with water vortices.

In the case of the corresponding technical usage, which can be interpreted as making use of the occurring "spin coupling", therefore in principle utmost caution is imperative.

<i>: Rho Sigma (Dr. Rolf Schaffranke): Forschung in Fesseln, das Ratsel der Elektro-Gravitation, Ventla-Verlag, Wiesbaden, 1972, Page 67.

Nikola Tesla: „Ere many generations pass our machinery will be driven by power obtainable at any point in the Universe This idea is not new, we find it in the marvellous myths of Antheus, who derives energy from the earth... Everywhere in the universe there is energy. Is this energy static or kinetic? If static, our hopes are in vain; if kinetic, and this we know it is for certain, then it is a more question of time when men will succeed in attaching their machinery to the very wheel work of nature". New York, the 20th may 1891, American Institute of Electrical Engineers.

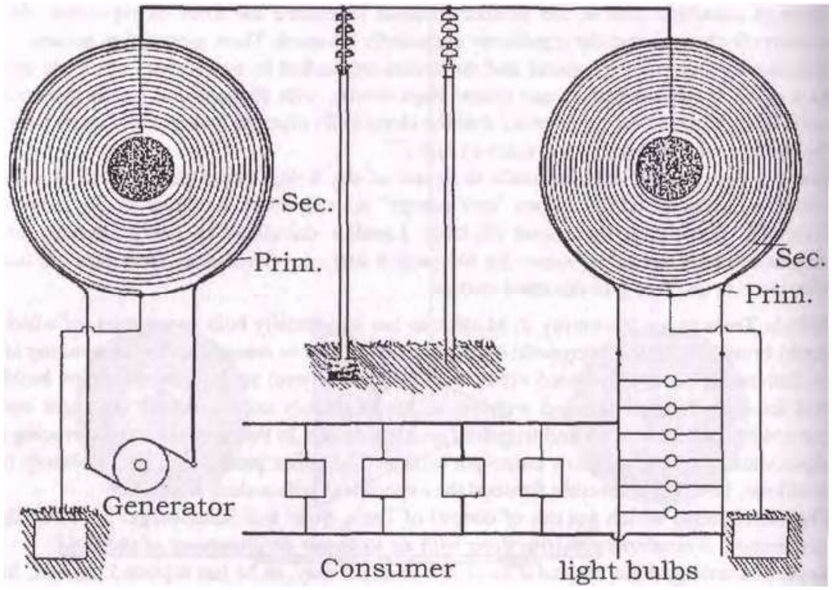

Fig. 9.5: Energy transmission bound to a wire, Patent No. 593,138 (1897) of Tesla.

<i>: Philadelphia Public Ledger, Nov. 2, 1933

<ii>: Dr. Nikola Tesla: Complete Patents, Tesla Book Company, New Edition 1983, Compiled by J. T. Ratzlaff, ISBN 0-960356-8-2, page 301-304.

9.5 Free energy?

Furthermore is reported of Tesla, he would have developed a "converter for space energy" and 1931 have built it in a luxury car (Pierce Arrow)<i>. The car was a 145 km/h fast, doing so the asynchronous motor (presumably built by Westinghouse) furnished 80 HP at 1800 Rpm. The "free energy" the converter, build by Tesla himself, got from a 1.8 m long antenna. Because the motor ran without adaptation of frequency in slipping operation, it had to be cooled correspondingly. It was on the way in trial run for over a week. Now we of course want to know, of what this "free energy" consists, which Tesla will have used and of which he already 1891 had spoken in the American Institute of Electrical Engineers.

For the electron as a spherical vortex we have calculated an electric tension voltage of 511 kV between its surface and its centre respectively infinity (equation 6.31*, fig. 7.1). The highest level of tension voltage normally used for the high tension transmission lies at 380 kV (effective value); for a direct current transmission it is 500 kV. Although still higher tension voltages would be desirable, they are avoided. This is no accident. Experiments with higher tension voltages namely have resulted in inexplicable high losses. We have an explanation: the electrons are taken apart on the way! Their inner energy amounts with the outer energy to zero. The charge carrier, which in the power station as result of an energy conversion has been sent on a journey, is in danger to vanish into thin air (e.g. corona) for tension voltages above 511 kV. The transmitter of Nikola Tesla however (fig. 9.1 and 9.5) worked with 600 kV and more. He said, with his experiments he had destroyed billions of particles without being able to observe an emission of energy and made fun of the misinterpretation of Einstein of the already at that time well-known mass-energy relation $E = mc^2$.

For the purpose of a one wire or a wireless energy transmission the tension voltage therefore has to be higher than 511 kV. With Teslas equipment however the electrons shouldn't be destroyed, but merely be pulled apart to plane circular vortices with help of the today is Tesla coil known winding. These then could be sent on a journey over the antenna, to again be caught and formed back in spherical electrons by the receiver antenna. As long as the electrons don't fall apart, they keep their structure and quantum property. Quanta pass on an excitation, for instance a bump, in form of a longitudinal standing wave, by one particle bumping the neighbouring particle, analogous to sound waves, where one gas molecule passes the bump on to the neighbour. The transmission hence takes place as longitudinal scalar wave.

Electrons pulled apart to planar vortices in addition haven't got a closed vortex centre anymore on their journey. For the by Tesla in his laboratory used and publicly presented one wire energy transmission (1897, Patent Nr. 593,138) the transmission hence doesn't take place in the wire, but as vortices around the -wire. That explains, why only a relatively thin conductor, which normally should have melted, was necessary for a demonstrated power transmission of 10 kW. Tesla however could show, that the wire stayed cold and virtually no heating losses were measurable<ii> (fig. 9.5). He himself said, that this one wire transmission technology is much better than the alternating current technology full of losses, which stems from him as well. Tesla must have known the limit of 511 kV very exactly, because on several photographs one can see, that he changed the coiling technique off this value!

<i>: Philadelphia Public Ledger, Nov. 2, 1933

<ii>: resp. Tesla's lost inventions, VAP, page 36 and page 48, 49.

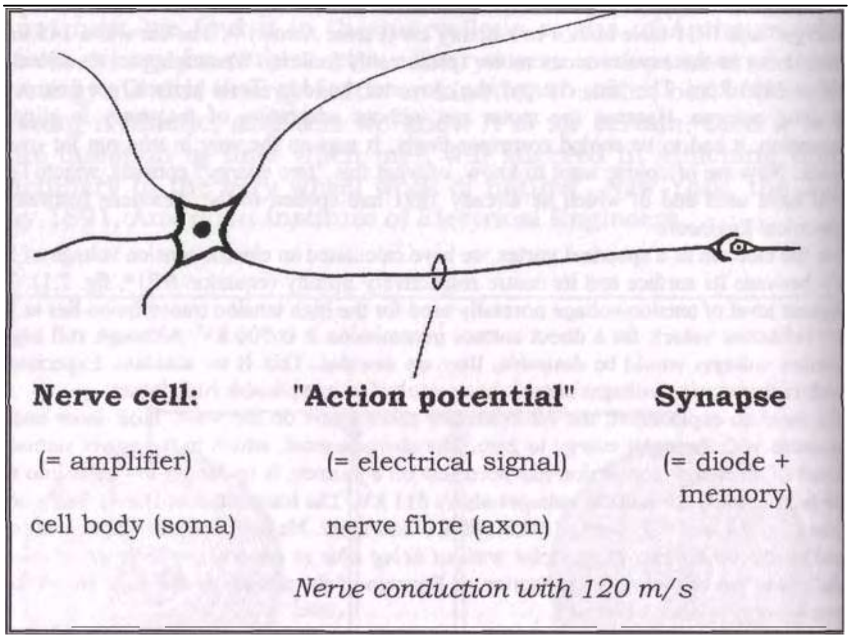
Nerve cell:
(= amplifier)
cell body (soma)
"Action potential"
(= electrical signal)
nerve fibre (axon)
Synapse
(= diode + memory)
Nerve conduction with 120 m/s

Fig. 9.6: Nerve cell and message processing (neurons)

- Material information conduction: at plants
- Information conduction with hormones: at simple animals
- Electrical information conduction: at higher developed creatures
- Analog transmission: from 10% and more difference in potential, and
- Digital transmission: number of the action potentials
- Modulated information: i.e. complex modulation of the action potential (the potential vortices)

9.6 Nerve conduction and action potential

Technical solutions often are only a suboptimum and many times they are anything but an optimum. In many cases the feasibility stands in the foreground. In the question of the current distribution at that time the decision was made for the worse system, the full of losses three phase system, because no current meters were available for the free of losses one wire technology. Without being able to collect money from the consumer, it wouldn't have been possible for the energy distribution installations to pay for themselves. Compared to that is nature with its selection principle merciless and without compromises. Only the most efficient system has a chance to survive and to spread. If it's about transmittining information in an electric way and two different principles present themselves, then in nature only the better one will be brought into action.

Let's take a look at the nerve conduction in the human body. In the synapses ion concentrations and electric potentials of 70 to 90 mV arise. Here with conventional gauges the activity of a nerve can be detected. From a measurement of the transmission time of synapse to synapse the velocity of the signal is determined. If we however want to measure between the switch points on the line, then we have to find out, that for instance with an ammeter actually nothing is measurable. In addition the nerve fibre has a miserable conductivity.

The biologist calls the electric signal just "action potential" and draws a vortex-like ring around the nerve fibre and speaks of a mixed digital and analogue information transmission.

The doctor on the other hand knows two different types, fast and slow nerves. In the inside both are built up virtually identical. A characteristic difference consists of the fact, that the fast nerves are jacketed with a thick fat layer.

The technician would say, they are better isolated, but why they therefore should be faster, he hardly would be able to answer. If we however assume, that the action potentials or OUT potential vortices oscillate around the conductor, thus exactly in the isolation layer and are forwarded there, then possibly an explanation would have been found.

The nerve conduction moreover has much in common with the one wire system of Tesla:

1. Charges and electric potentials are transmitted
2. Doing so a transport of charge carriers, thus an electric current on the line, isn 't detectable.
3. Consequently no losses worth mentioning occur.
4. Decisive for the functioning is the dielectric insulator material which surrounds the conductor and not the electric resistance of the conductor.
5. In contrast to electric circuits a draining conductor is unnecessary.

It looks as if the one wire system is not new at all, as would nature use it already for ages in the highest perfection. We realize immediately, why our head does without a ventilator and how the high of signal processing density can be explained. Compared with our brain modern personal computer (PC) with their backward cable technology are far less than a suboptimum.

With regard to the pressure on the environment by interference radiation we had derived, that currents and eddy currents with the skin effect direct their radiation to the outside, whereas potential vortices with their concentration effect direct it to the inside. Probably for this reason the radiation field of a person is environmentally better compatible than that of a PC.

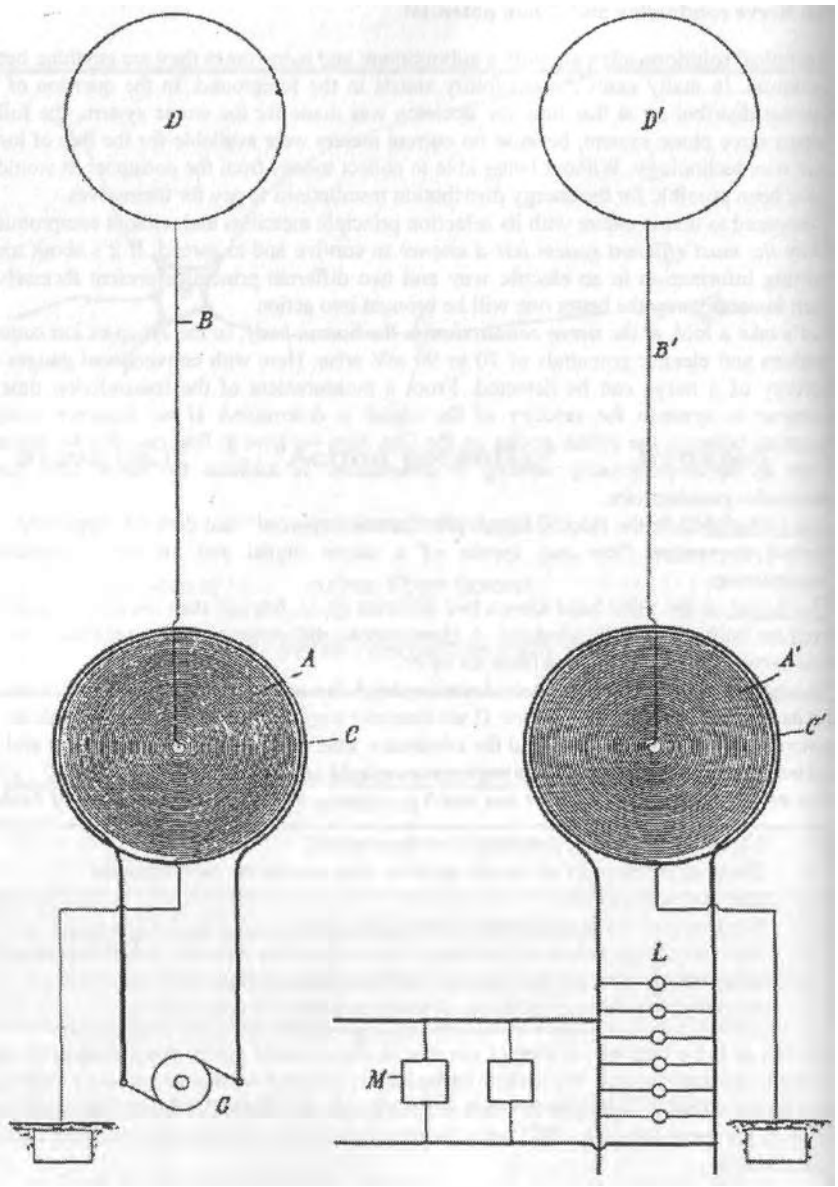

Fig. 9.7: Wireless energy transmission, Patent No. 645,576 (1900) of Testa.<i>

<i>: Dr. Nikola Tesla: Complete Patents, Tesla Book Company, page 311-321.

9.7 Wireless energy transmission technology

Tesla still went a step further. He has cut the wire connection between transmitter and receiver (fig. 9.5) and instead has installed each time a spherical electrode (fig. 9.7). With this facility he now was able, to transmit energy completely wireless (1900, Patent No. 645,576). Building on this principle Tesla already 1890 had built a radio remote control for a battery operated submarine and had presented his patented system the navy. But they weren't able, to recognize the strategic importance of the radio technology and rejected with the words: "It's too advanced!" (fig. 9.8).

Tesla further had proven with this technology the mentioned proof concerning the existence of longitudinal electromagnetic waves. There exist descriptions, partly of Tesla himself, how he, inside or also outside his laboratory, goes in search of oscillation nodes of the standing wave with a measuring receiver[i]. He names several conditions for the wireless energy transmission:

1. "perfect resonance" (same frequency) of transmitter and receiver
2. Installation of the receiver on an "oscillation node" (maximum).

Important is also the measurement of the velocity of propagation, and that isn't constant from the start for a wave, which oscillates in the direction of propagation. From the research of earthquakes we know, that the longitudinal waves are faster than the as well occurring transverse waves. Usually the distance to the epicentre is determined from the difference in transmission time.

In the patent specification the measurement is described[i]. Tesla has sent longitudinal radio signals from his transmitter in Colorado Springs once around the earth along the earth's surface and determined a transmission time of 0,08484 seconds. This corresponds to a frequency of approx. 6 Hz. He saw his result confirmed, as he could find the oscillation node of the standing wave again (according to $\lambda/2$) in his laboratory (on the ground plate).

For the full wave length λ the Schumann resonance, which describes a standing wave of an around the earth running Hertzian wave, lies as is well-known at 7.8 Hz! Tesla calculates for his wave a speed 1.6 times the speed of light assuming the polar radius of the earth to be 6363 km.

Also this measurement result confirms, that Tesla didn't use the Hertzian wave. Tesla found the off-beat concept of Einstein, the speed of light would be the fastest possible velocity of signal transmission, only funny. If however today is claimed, 1993 in tunnelling experiments for the first time a speed faster than light has been measured, then this just isn't true. Possibly also the Tesla wave tunnelled, as the vortices for a small conductivity of the air contract and as a result of the length contraction become correspondingly fast.

Later Tesla after several goes in vain even succeeded in building a high tension tube as a tunnel, with which the velocity of the signal could be increased arbitrarily. Tesla with that pursued the goal to be able to make radio contact with other worlds[ii].

<i>: N.Tesla: Art of Transmitting Electrical Energy Through the Natural Mediums, US-Patent No. 787,412, 18.4.1905, Complete Patents pp. 397-402.

<ii>: M. Cheney: Tesla, Man out of Time, Barnes & Noble Books, New York, 1993, und S. 309, Omega-Verlag, 1996, ISBN 3-930243-01-6

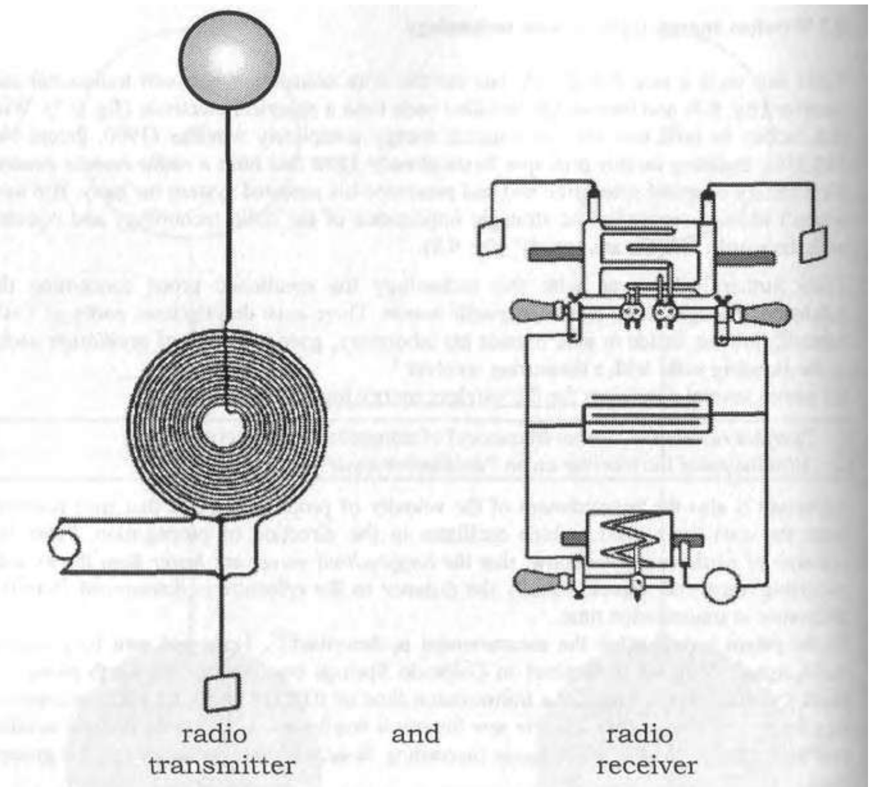
radio
transmitter
and
radio
receiver

Fig. 9.8: Energy wave broadcasting system. Patent No. 787, 412 (1900) of Tesla.

In today's switching technical notation:

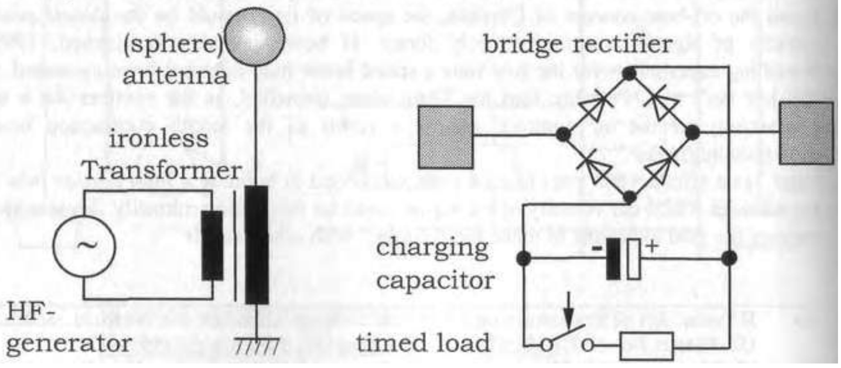
(sphere)
antenna
ironless
Transformer
~
HF-
generator
bridge rectifier
charging
capacitor
-
+
timed load

9.8 Measuring and switching technique of Nikola Tesla

As an important and accepted experimental physicist with 14 doctor titles and carrier of the Edison medal Tesla always has held on to his measurement results. Not a theoretical interpretation but exclusively correct carried out measurements have shown him the physical reality. But the by Tesla won measurement results were already for 100 years hard to digest and couldn't be brought into accord with any theory. Therefore a switching technical analysis of the by Tesla described and carried out experiments should give us now information over the nature of the free energy, the tachyon energy, the orgone, or however fantastic the terms may read.
The Tesla coil, according to the instructions of the inventor, is a flat coil wound like a spiral in contrast to the copies today on sale which, surely out of ignorance, are mostly cylindrical. Its peculiarity probably is to be seen in the circumstance, that charges moved through the wire of the coil inevitably increase their angular velocity towards the centre, In this way the electrons, which at first are spherical elementary vortices, are pulled apart to vortex rings like planar discs.
Tesla switches the "secondary" called flat coil between two "terminals". Following he sends, stimulated by his "primary", charge carriers from one terminal to the other and back again and produces a standing resonant oscillation.

Mostly replaces Tesla one of both terminals by the earth. He thus assumes, that now the earth as a spherical electrode takes over the function of one terminal. That he again infers from the observation, that a by the transmitter wireless fed little lamp goes out, if he cuts the connection to the ground. Doing so the oscillation collapses. Radio sets on the other hand can also be operated without grounding, as we know!
The degree of effectiveness of today's distribution technology of current due to the heating losses lies clearly beneath 100 per cent. Without the losses of the wiring it lies close to 100 per cent for the discussed one wire energy transmission. There the vortex rings are guided nicely one after another along the line like beads drawn over a string. This result even is to be expected, as far as no vortex "jumps off the wire or "falls apart". For the wireless version Tesla however to his own surprise had to find out that more energy could be received, than his transmitter produced. The measured degree of effectiveness lay above 100 per cent! He therefore called his transmitter a "Magnifying Transmitter" (fig. 9.10). The further transmitter and receiver were away of each other, the further the received energy increased. Tesla inferred from this, that there had to exist free energy and that he had caught that too.

Consequently he had built a receiver for free energy and registered for patent (1901, Patent No. 685,957, fig. 9.9). Tesla states that the amount of energy depends on the size of the "terminal". Of today's sight we could be willing, to describe this receiver plate as a solar panel, but we should know, that the apparatus produced energy even at night! In addition the energy gain was considerable higher than for today's solar panels. Tesla spoke of "radiations ", of an unknown radiation and he in his lifetime has in vain sought-for help of explanation.
The vortex model will also in this question be a valuable help to us.

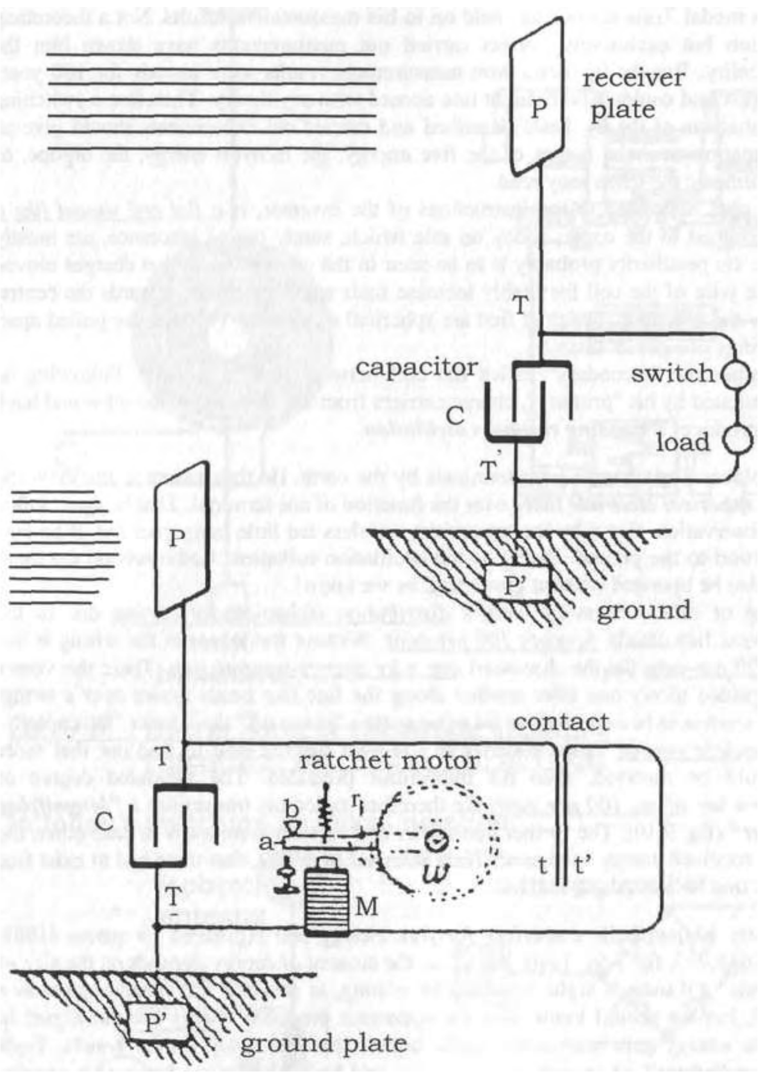
receiver
plate
P
T
capacitor
C
switch
load
T'
P'
ground
P
contact
T
ratchet motor
C
b
a
t
t'
T
M
P'
ground plate

Fig. 9.9: Receiver for "free energy",
Patent No. 685,957 (1901) of Tesla

9.9 Energy technical usage

The answer of the potential vortex theory could turn out as follows:
If at the transition of the one wire to the wireless transmission the ring-like vortices is purloined the guiding wire the vortices immediately begin to turn around each other, as is observable for flow-technical ring-like vortices,. In this way the ring-like vortex the next moment shows its inside. If it before was an electron vortex (fig. 4.3), then it now shows as positron, if it was negatively charged, then it now is positively charged. Following it oscillates back again, etc. Wit that the ring-like vortex on the average has no measurable charge and no mass, because it alternately forms matter and anti-matter. Without interaction it has an enormous ability of penetration . In physics such particles are called neutrinos.
Tesla thus had, apart from his transmitted energy wave, which turned out to be neutrino radiation, by chance also caught neutrinos which oscillated synchronously.

According to the actual level of knowledge do neutrinos penetrate the earth and appear also on the night side. The order of magnitude in every second amounts to approx. 66 billion neutrinos per square centimetre. It is a true bombardment. If we would be able, to collect and convert all neutrinos, the won energy would be entirely sufficient, to cover the need for energy of the world population (approx. 27 W/m^2). We merely have to materialize them, thus give them mass, charge and the necessary localization. Tesla was able to do that experimentally! Let's record:
The ring-like vortices, which Tesla with his transmitter has sent on a journey as electrons with an open vortex centre, are neutrinos (fig. 7.12). Tesla requests that transmitter and receiver operate in resonance, thus with the same frequency. Under this condition the receiver collects in all oscillating vortices, so that no one is lost.
If the neutrinos for instance are just positively charged when leaving the transmitter electrode, then an electromagnetic force of attraction takes place, if the receiver electrode at the same time is negatively charged. The required operation with the same frequency and opposite phase guarantees that also the next moment, if both, the neutrino and the receiver, have changed their polarity, the electromagnetic attraction is preserved.
It is obvious, that strange neutrinos which fly past and by chance oscillate synchronously are as well attracted. In that way the power collected in the receiver capacitor will increase further and degrees of effectiveness of over 100% are obtainable. Tesla discharges the receiver capacitor timed with the frequency of resonance (fig. 9.9) and points to the difficulty of an exact keeping of the condition of synchronisation.
Tesla indeed did work on a theory of his own, but never published it. The patent office and his lawyers had to proceed from the Maxwell theory, although Tesla knew only too good, that his apparatus in no way could be described with this theory. It therefore can't be excluded, that important facts in his patent specifications haven't been mentioned at all - or even worse - consciously or unconsciously in an inapplicable manner have been explained with the Maxwell theory.
Perhaps this is the more profound reason, why the numerous Tesla imitations don't want to function quite so well. With the new theory there should be some changes, and we should accept the challenge for the sake of humanity.

The transmitter tower of Tesla (57 m high with a spherical electrode of 21 m) for the transmission in multiplex mode of operation (according the inventor) of:

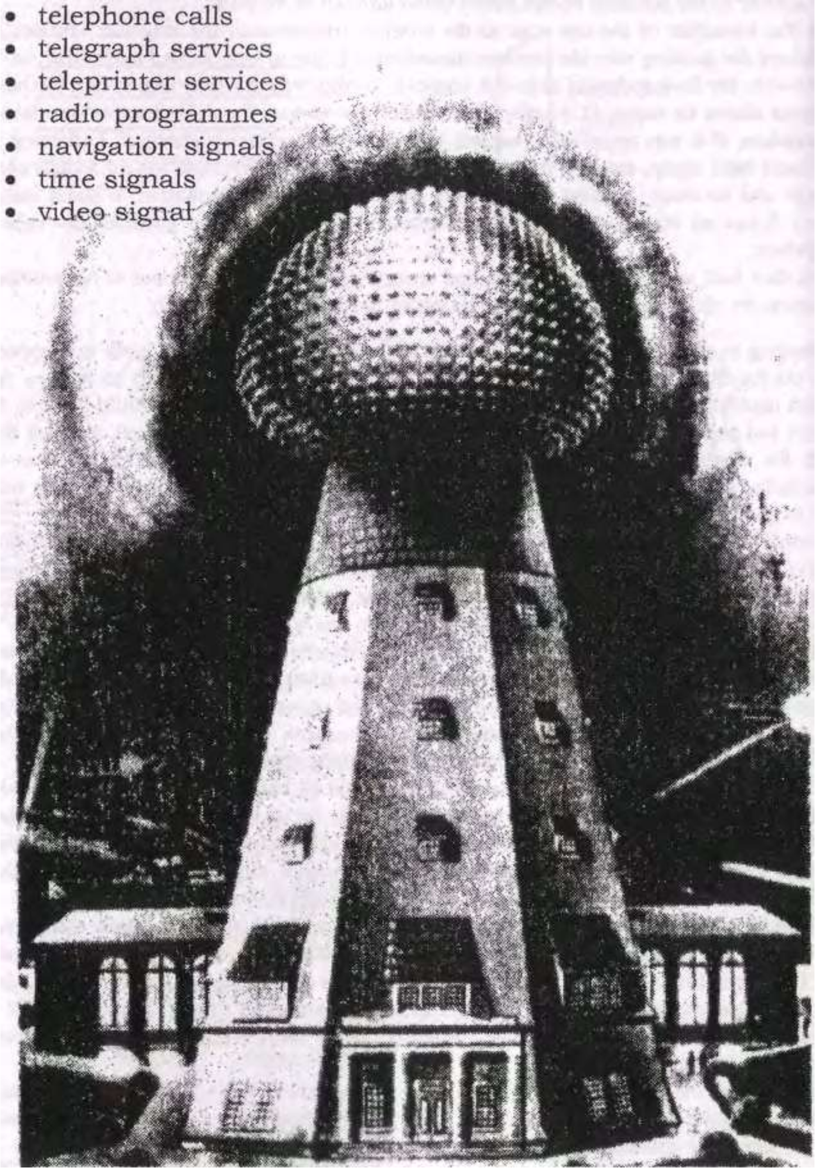

- telephone calls
- telegraph services
- teleprinter services
- radio programmes
- navigation signals
- time signals
- video signal

Fig. 9.10: The planned transmitter Wardencliff of Tesla on Long Island, 7.5 MW (1902).

9.10 Environmental compatibility

We now can summarize the different observations and try to find answers for the in the first chapter raised questions concerning the electromagnetic environmental compatibility.
The by Tesla used longitudinal energy wave is a potential vortex wave in the sense of the vortex model of Kelvin, which we again could identify as neutrino radiation (Dirac).
Also other, partly quite fantastic terms, are known like "radiations" (Tesla), "orgone radiation" (Reich), "tachyons" (Feinberg), "grid radiation" (Hartmann), "bioradiation" or "water veins".
Actually these rays have nothing to do with water as a cause. Water with its high dielectricity however favours and influences the course and the local distribution of the radiation. The maximums of the standing wave considered for themselves result lines in straight as a die in the landscape. Of Tesla is known, that he measuring technically could detect the cross points of the lines, which he called oscillation nodes. There exist so-called dowsers, who can make these standing waves out even without technical aids. For that they hold a fork in their hands, which suddenly moves over such lines.
Let's remember that the same potential vortices are used at the nerve conduction as reaction potentials to conduct stimuli. If ring-like vortices arrive at a muscle, it contracts without knowing, if the signal was sent from the brain over the spinal cord or picked up from the environment over the fork and the hand.

Thus if the same signals, which we use for the tasks of controlling and thinking, are also produced by technical apparatus, then this touches the question of the environmental compatibility in a direct way. Above extremely strong points of resonance, which are called "geopathic zones", now and then even is warned about "psychotronics", a function trouble of the brain, which can show in a disturbance of the waking consciousness.
A possible explanation would be that the brain is occupied with so much vortices picked up from the outside, that it isn't able anymore to accomplish its actual tasks. For muscle cramps or an inexplicable crash of a PC an correspondingly interpretation would be obvious.
As long as no gauges are available, the highest caution is imperative!

If Tesla had been able to put into operation his transmitter tower for longitudinal waves ("World Telegraphy", Wardenclyffe, Long Island, s. fig. 9.10) with the estimated power of 7.5 Megawatt and thus had been able to realize his dream of a world-wide wireless communication, then this could have had hardly estimable consequences for the health of the world population.
Shortly before the completion of the project, in the year 1905, Tesla without notice let stop all work. He to that never made an explanation to anyone. Officially it was said, his financial backer J. P. Morgan would have withdrawn his financial means. In any case was Tesla financially ruined with this step.
Perhaps also T. A. Edison was behind it, who was his opponent at that time. Edison committed himself engaged for the use of direct current. Against the alternating current, preferred by Tesla, Edison argued with unhealthy risks, and with that he perhaps wasn't so wrong at all.

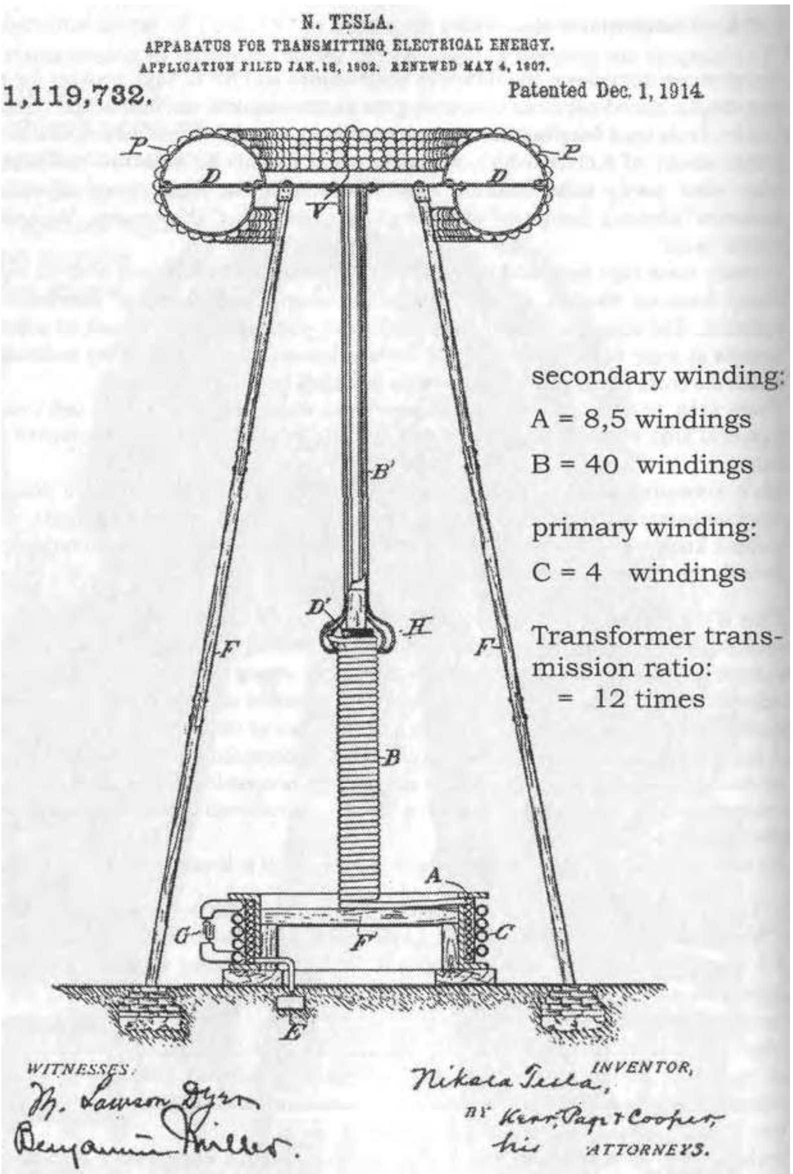
N. TESLA.
APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.
APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.
1,119,732.
Patented Dec. 1, 1914.
P
P
D
D
V
B'
secondary winding:
A = 8,5 windings
B = 40 windings
primary winding:
C = 4 windings
Transformer transmission ratio:
= 12 times
D'
H
F
F
B
A
G
C
F'
E
WITNESSES:
M. Lawson Dyer
Benjamin Miller.
Nikola Tesla,
INVENTOR,
BY Kerr, Page & Cooper,
his ATTORNEYS.

Fig. 9.11: Concerning the technology of the Wardencliff transmitter (fig. 9.10).[i]

<i>: Dr. Nikola Tesla: Complete Patents, (J.T.Ratzlaff), Tesla Book Company (1983), ISBN 0-960356-8-2, P. 435.

9.11 Ecologically fair usage

Tesla not only has stimulated ecologically fair usages of the energy radiation. To that can be counted the use of a Tesla transmitter as a radiation weapon<i> as well as the mentioned building of a Tesla interferometer, the low-frequency interferences of which can lead to mighty gravitational waves and with help of which it should be possible, to produce earthquakes, to sink ships or to shoot down planes from the sky<i>. Against the background of two world wars it can be understood, why Tesla has drawn attention to the possibilities of a use of his system for the purpose of defence.

In today's sight a technology like Tesla's wireless energy transmission (fig. 9.11) hardly would have any chance of a large technical use, because it should be reckoned with enormous problems of environmental compatibility. After all had come out in our considerations, that entirely in contrast to the Hertzian wave just of the Tesla energy wave direct biological effects are to be expected. A purposeful medical use of the Tesla rays however can be imagined.

The collecting of "free" energy waves with a neutrino receiver at first undoubtedly would be a clean and ecologically fair alternative to the present energy technology, with which we heat or nuclearly contaminate our environment. But even the use of free energy converters has limits. Should it turn out that the Photosynthesis uses such a materialization, in which the necessary free electron for the chemical reaction actually is produced from a caught neutrino in the so-called reaction centre and thus the reaction energy is won, then for an as usual excessive use by men it can't be excluded anymore, that the plants on earth die and nothing grows anymore!

I in contrast attribute big chances to the one wire transmission technology. Because for an use of alternative and regenerative energy sources, like the sun in the desert or the geothermal energy in Iceland, the transport of energy has to be made over large distances through the oceans the small losses and the possibility to be able to use sea cables, play a big role. In the residential areas and conurbations on the other hand it has a positively effect, that by the concentration effect of the potential vortices the interference radiation is directed inward. Laid on the bottom of a sea or of a river, such a line in addition is protected against the effects of lightning and sabotage, and it neither spoils the countryside, as the old-fashioned high tension pylons do.

For transportation tasks the mentioned making use of the spin coupling would surely be the most suitable drive. The denser the medium, the more efficient would be the driving effect, for which reason not only flying objects, but quite particularly also ships and submarines could be operated non-polluting in this way without combustion and without exhaust fumes. Build up effects by the water pressure or the sound barrier don't occur. But control might not be entirely unproblematic. If for the reason of a unipolar arrangement of the field a relative velocity to compensate the field becomes necessary, then a correspondingly build up arrangement will do so without considering losses. It is no accident that most apparatus which got well-known have flown the builder sooner or later "past his ears" (note of the translator: broke into pieces). Till a commercially utilizable flying saucer has been built, many technological hurdles have to be taken.

Claims to an up-to date environmental policy
in view of longitudinal electromagnetic energy waves
= Tesla waves = neutrino radiation = scalar waves
= potential vortex waves:

1. Gauges must be developed and built, with which the energy waves can be measured.

2. All technical apparatus should be tested regarding the emission of energy waves. Harmful radiation should be avoided.

3. Only such transmitter antennas may be brought into action, which emit a minimum of energy waves.

4. The ability of biological systems to absorb energy waves should be determined. To that also belongs the question, how many vortices collect locally and how fast they fall apart.

5. Limits for allowed pressure produced by artificially energy waves should be fixed, regulations and laws enacted.

6. The in nature available energy waves should be brought to a technical use, with the goal of a natural energy production.

Fig. 9.12: Catalogue of claims concerning the environmental compatibility in the view of the new field theory.

9.12 Claims concerning the environmental compatibility

Now we still not know, if the using of a handy or another electronic device poses a danger to the user. Now, we surely can't exclude a danger, and the statement of the manufacturer, it is safe because all limits have been met, is worth nothing, as long as the part of longitudinal waves isn't measured. Therefore stands at the very top of the list of the catalogue of claims (fig. 9.12) the development and the building of gauges to measure Tesla radiation. More than half the 700 patents of Nikola Tesla concern the building of transmitter and receiver installations for energy waves. Here a lot of valuable clues are found, especially as he himself has built and used some of the switching circuit designs. In the seminar a rebuild is discussed, with which 1995 the existence of longitudinal waves could be proven [A7].

For an earthquake always both, the transverse and the longitudinal wave, occur simultaneously and, taken exact the same is valid for the sound wave, even if the transverse part doesn't get too far in the air. The emission of both parts in analogy is almost to be expected of an electromagnetic wave generator or transmitter. Actually both waves are detectable at the receiving point: The useful transverse wave and the longitudinally propagating vortices, which show as noise (fig. 4.7). If the vortices are predominant and the useful signal goes under in the noise, then a interruption of the radio operation occurs. Also small gardeners immediately near a radio station have made experiences as e.g. in a distance of one kilometre from the transmitter of RIAS-Berlin. They were able to light their allotments day and night with a neon tube, to which they merely had tied a free wire. The radio station immediately insinuated, that they in an illegal manner had tapped their useful wave and damped it. But it can't be excluded, that the transmitter has sent a longitudinal part in the ether too, which was responsible for the energy transport. But that suggests, that also at other transmitter antennas, thus for the mobile phones, the unwelcome parts is transmitted even than, if we don't use it at all.

Still another problem is added. If in a modern receiver radio waves arrive, then they roll up to a standing wave and according to fig. 1.4 to a vortex over the antenna. According to that we aren't able anymore to distinguish, if the transmission took place as Hertzian wave or as Tesla wave. The separation of both sorts of waves has to take place before the antenna is reached.

An up-to date measurement arrangement would look as follows: In front of the receiver antenna a path with a tunnel, which represents an insurmountable obstacle for the transverse waves (fig. 6.14), is placed. Then only the longitudinal waves pass the tunnel, and these at the end of the tunnel can be recorded and analysed with conventional technology. The tunnel is an elegant possibility, to make use of the part of Tesla waves. After all only this part is relevant with regard to the electromagnetic environmental compatibility. Concerning the above asked question we can take home at least one important realization. The probability is quite high, that you actually notice nothing of the handy radiation, because not every system react in the same manner to certain vortices. The resonance conditions regarding frequency, phase and position always have to be fulfilled, if an absorption of energy waves should occur. (fig. 2.10 B, E. coli bacteria). On the other hand it however can't be excluded, that just you or me synchronize according to frequency and collect 100 % of the transmitted energy radiation. In contrast to the Hertzian wave plays for that the distance to the source of interference only a secondary role! Seen that way worries would by all means be entitled ...

From Indian mysticism:

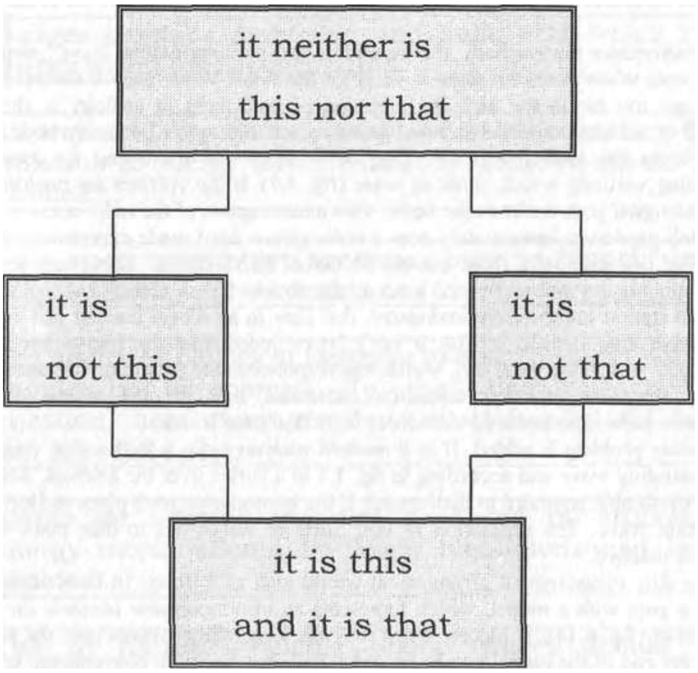
it neither is
this nor that
it is
not this
it is
not that
it is this
and it is that

Fig. 9.13: The structure of the world equation, an example from Indian mysticism

9. 13 Epilogue belonging to part 1

The correctness of a new theory not by all means guarantees, that it also is accepted! A new theory must be used. Only if a concrete need exists then an acknowledgement is to be expected if at first only by the direct beneficiaries.

In november 1994 in Munich the trade fair "Electronica" took place. Here invited by the Deutschen Gesellschaft fur EMV-Technologie (German association for EMC-technology, note of the translator) the books about potential vortices [Al, A2] were honoured with a prize by an internationally renowned EMC-scientist. The number of people, who the potential vortex theory helps further to find answers to problems, which until now aren't explicable, grows further.

Nevertheless, nobody should believe that something will be changed to the content of the text books in the shelves that fast. Habit and laziness encourage to further use the to date interpretation, even if a new theory has ready a better and convincing answer. It will take some time, to give an example, till the heating of water in a microwave oven is explained as eddy losses, as is described sound by the potential vortex theory already today. The discovery of a physical phenomenon may happen in a fright second. Its acknowledgement on the other hand as a rule is a lengthy procedure. If we look back in history we find out that even in the past it hasn't been otherwise.

Of the four fundamental phenomena in fig. 8.2 at first the tension voltages and the charges have been discovered by Coulomb (5.7e).

(5.7c): An effect of the eddy currents was observed already 1825 by the French physicist Arago. He discovered that a magnetic needle suspended above a rotating disc tends to rotate along. Out of ignorance he named the effect "rotational magnetism".

But the eddy currents could only be mathematically described with the help of two laws, the law of Ampere (1820) and the law of induction of Faraday (1831). Because the effect due to the eddy current losses rather was regarded as disturbing, technical applications or patents at first have stayed out. At first Riidenberg has carried out fundamental calculations for eddy current brakes in 1906. The discovery and usage of the "skin effect" is to be attributed to Tesla.

(5.7b): The electromagnetic wave for the first time was mathematically described by Maxwell (1864). He had aimed to calculate the light and he could show, that his theory actually could correctly describe all properties of the light. With that he was able to prove the correctness of his theory. But he had to deal with many sceptics, because he had postulated the dielectric displacement, without however in his lifetime being able to measure it because of its smallness.

This may be the reason, why the discovery is attributed to Heinrich Hertz and was and is spoken of the Hertzian wave. Neutrally seen Hertz at first was nothing more than a enthusiastic pupil, who had worked through the "Treatise"<i>, Maxwells books. All results, which Hertz could observe experimentally, already were contained in the Maxwell-theory and published.

<i>: James Clerk Maxwell: A Treatise on Electricity & Magnetism, vol. 2, Chapter XX: Electromagnetic Theory of Light. Dover Publ. New York

But the great mathematical description of Maxwell at first had a purely academic importance. The interest of the general public on the other hand concentrates on the practical applicability. Hertz had experimented with antennas and has laid the foundations of the building of a gauge for electromagnetic waves (1888). The wireless transmission of invisible waves was convincing. It was an entirely new technology, which stimulated fantasy in view of a technical usage. A practical man, who was stimulated by the theoretical realizations, was Marconi. He goes for the inventor of the radio (1895, Nobel prize: 1909).

Between the three summits, first the calculation and at the same time discovery, second the measurement and third the usage of the wave phenomenon lay 31 years, and the radio technology is developing still further, even today. In the area of the cellular phone network and the telecommunication even a kind of gold mining mood prevails, which can't be impressed much even by hints of an electromagnetic environmental compatibility.

In the euphoria of the past 100 years is fallen into oblivion completely, that besides the by Hertz detected transverse wave in the wave equation according to Laplace and according to the original version of Maxwell also a longitudinal wave was described mathematically. This one was discovered and detected by Tesla in numerous experiments. With reference to his discovery Tesla initially had asserted publicly, Hertz had been at fault, with which he undoubtedly was wrong and had incurred the wrath and the scorn of the scientific community. As a result of this development his experiments haven't been reproduced and the discovery of the scalar wave could fall into oblivion.

Not enough with such sanctions against the inconvenient freethinker Tesla a series of professors like Gibbs and Heaviside have made cuts and discarded all parts from the original version of the Maxwell equations, which by that time weren't considered to be experimentally proven by experiments of Ampere, of Faraday, of Ohm and Hertz. With that the scalar waves fell victim to the axe, what, though it had encountered severe criticism, finally entered as textbook opinion into all textbooks, after Hertz also had sanctioned this measure.

If the field equations according to Maxwell in today's revised version don't describe scalar waves anymore, then all orthodox scientists, who want to bear reference to that, have to bear one thing in mind: discarding a term of an equation is the same as neglecting it. But the term only may be neglected if it is sufficiently small. The first thing every student learns in the physical practical training is that this first has to be tested. For that the term has to be measured and its order of magnitude must be determined. Then it is little helpful and extremely unscientific, if at first is neglected and on the basis of the revised field equations, which have put all scalar waves to zero, is demonstrated that the neglect was allowed or even necessary.

A practical example are the longitudinal wave parts, like they occur and are proven in the near-field of a transmitter antenna. Considering the order of magnitude a neglect is out of the question. On the other hand they should not exist at all according to textbook physics, because they had fallen victim to the axe. Since most scientists in present time do not know about the cutting action anymore, they are postulating field vortices anew in the proximity of an antenna. Field vortices, which are propagating longitudinally in the direction of a field pointer as a scalar wave.

With that they calculate their own arbitrary assumption<i>.
If in the practical training of physics a student unevaluated neglects an essential factor of influence, then his experiment goes as not passed. The experiment must be repeated so long until according to the approved methods of scientific soundness all used neglects have been tested individually and have undergone an error consideration.
Do we now have to deny the cutting action, which had been plotted by Gibbs, Heaviside and others, having the necessary scientific soundness? Do we have to review and rework all textbooks concerning the illegal neglect? Since the cutting action a gap gapes in the field theory!

(5.7d): The potential vortex fills the remaining gap in fig. 8.2. Several research scientists and scientists can be named, who already could observe this phenomenon: in front of all goes Nikola Tesla for discoverer of the Tesla currents, very weak currents which are said to cause extremely high potentials. Here presumably cause and effect have been mixed up, because weak currents never can produce high potentials. Actually the potentials prove to be a result of the potential vortices, whereas the currents aren't more than a result, nothing but leak currents. At least Tesla could use the vortex, but without a usable theory he neither could calculate nor adequately explain it. Besides Tesla Wilhelm Reich should be mentioned, who has collected the vortices in his orgone accumulator.
Mostly only certain aspects were observed and investigated: by Tesla the planar vortex, by Reich the influence on the weather, by Schauberger the water vortices and by all three the medical and biological aspect. The list of names is incomplete and could arbitrarily be continued.

With the discovery of the potential vortex (1990) the basis for the building of a gauge and the technical usage of the physical phenomenon is laid [Al]. It not only concerns the search for water, but also the detecting of the vortex balls and vortex streets, of the standing waves of the energy radiation in the air, at the workplace, at the bedroom, in clinics, in recreational areas and hotels. As explained at the start, neither the field strength of a Hertzian wave nor the arising heat development can be made responsible for biological or technical harm.
It primarily are the newly discovered vortices of the electric field, which take effect. The effects can, as we have seen, as well be good for health as bad for health. Intensity, plane of polarization, vortex configurations and many other characteristics play a role here. To research these influential factors gauges for vortices will be needed as well.
We have to realize that in the technical domain the electromagnetic compatibility of an apparatus is determined by its sensitiveness to vortices, thus by the fact how many and which vortices can cause a function trouble. To determine the environmental compatibility of a product the emitted vortices, the energy radiation, have to be measured. Limits for high tension lines, for screens or handheld phones must be given in units of the potential vortices. The potential vortex has shown us the way to a unified theory and has brought along a new picture and understanding of our environment. It with that wants to show us the correct way for an ecologically compatible dealing with nature.

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3.Aufl. 1986 Springer-Verlag Berlin, Seite 335

Part 1: Edition belonging to the lecture:
"Electromagnetic Environmental Compatibility"

Prof. Dr.-Ing. Konstantin Meyl

Scalar waves

Abstract: 1. Auflage 1996, 4. Auflage and 1st English Edition 2003

Both technical and biological systems can be influenced by electromagnetic fields, whereby numerous questions still are open, like e.g. concerning limits and the nature of physical interference fields. The book shall do justice to the circumstance, that a fact oriented discussion about „electrosmog" implies an analysis of possible reasons, a mathematical derivation and a physical argumentation.

We proceed from the assumption, that only that part of the electromagnetic wave should be considered for malfunctions, which has been absorbed and which has rolled up to a field vortex. The effectiveness depends on the amount of the produced vortices and on the life span, the decay of vortices.

Analogous to the sound wave vortices in space are propagating as longitudinal waves. In this context is pointed to numerous effects. Examples for the technical and biological use of these standing waves are the energy transmission of Nikola Tesla as well as the nerve conduction, which functions in a corresponding manner. If the same vortices, which man uses for conduction of information, are emitted by technical devices, then biological reactions can't be excluded anymore and worries with regard to „electrosmog" seem to be justified.

Causes, phenomena and natural scientific consequences

INDEL GmbH, Verlagsabteilung, Fax: +49-7721-51870
Tel: +49-7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen/Schwarzwald
www.k-meyl.de meyl@meyl.de

Scalar Waves

From an extended vortex and field theory to a technical, biological
and historical use of longitudinal waves.

Part 2

by

Professor Dr.-Ing. Konstantin Meyl

Edition belonging to the lecture and seminar "Electromagnetic environmental compatibility"

(Original title: "Elektromagnetische Umweltvertraglichkeit")

Translated out of the German language
by Ben Jansen
(2000-2003)

* * *

Part 2: Edition belonging to the energy technical seminar

Free energy and the interaction of the neutrino radiation

* * *

INDEL GmbH, Verlagsabteilung

Villingen-Schwenningen 1996-2003

ISBN 3-98O2 542-4-0

Preface to the seminar

The point of a seminar is, to deepen, to practise and, as far as possible, to practically appy the material of a lecture. The knowledge of the content of the lecture hence is a prerequisite for the participation.

For the reader of this book that's tantamount to the recommendation, to have read the first part, the edition belonging to the lecture, before*¹*. Here the questions concerning the „electromagnetic environmental compatibility" are asked and the necessary bases for their answering is laid. Also practical consequences for various areas of science are indicated. The deepening most suitable should be made in form of a seminar, subdivided into the here presented part 2 to the energy technical seminar and a part 3 to the information technical seminar. Part 2 correspondingly concerns the energy technical aspect of electric or magnetic longitudinal waves, whereas part 3 is dedicated to the information technical aspect<ii>. Because it concerns a book which merely for reasons of usefulness is published in three parts, the chapters are consecutively paginated. References to chapter 1 to 9 hence automatically relate to part 1. The numbers of the figures and tables as a rule are identical with those of the chapters, in which they are discussed.

The seminar should lead on over the pure reading, consuming or listening and should stimulate to join in. All involved persons may and should give ideas and ask questions, even if these may sound little orthodox. The scientific working method takes, that is struggled for answers and even is argued, if necessary. To reach this goal, it mustn't exist any obligation or censorship, neither for the leader of the discussion nor for the participants of the seminar.

The seminar is being carried out since the summer semester 1997. The works of the seminar written by students treat the knowledge of text books of the respective theme. Following the lecture the answers are discussed and compared to those of the theory of objectivity<i> and other models of explanation. This procedure in this edition belonging to the seminar is reflected at some points, if for instance a chapter is completed with a „discussion".

The first edition of this 2nd part still was incomplete and has been handed out to the participants of a congress in Switzerland instead of a manuscript belonging to the lecture the 17th of Nov 1998. The here presented second edition in the meantime to a large extent is complete, but surely not yet perfect. In accordance with the experience made with the first part of the book also for this 2nd part a third and revised edition, in which the ideas of the participants of the seminar and of the readers find consideration, will be due to be dealt with after a year. The reader of the second edition has to console himself with the fact that a lively seminar constantly is changing and developing further. And that has to be so!

Villingen-Schwenningen, January 1999

<i>: Electromagnetic environmental compatibility, Part 1, Edition belonging to the lecture, INDEL Verlagsabteilung, 1996, see page 1-218 of this issue.

<ii>: Electromagnetic environmental compatibility, Part 3, Edition belonging to the information technical seminar, 2002, see page 443 - 625 of this issue.

10. Oscillating interaction

A theory is not an end in itself, even if it sounds very convincing. It has to be measured by its applicability. As an entry into the practical consequences, which result from the theoretical part[i], the question about the validity of Kepler's laws is raised.

10.1 Kepler's laws

The "radius vector", a line drawn from the sun to a planet, sweeps out equal areas in equal periods of time. At least has teached us Johannes Kepler it that way (fig. 10.1). The balance of forces, the gravitation on the one and the centrifugal force on the other hand results in the innermost planets of our solar system orbiting the sun very much faster than the outer planets (Mercury in 88 days, the Earth in 365 days, Jupiter in 4333 days and Pluto in 90465 days!).
For the inner planets as well as the big planets Jupiter and Saturn Kepler's laws are still found confirmed. But that shouldn't apply anymore for the outermost planets of the solar system. Beyond Saturn should prevail changed rules as is said, based on observations of the Voyager spacecrafts[ii].

If we direct our view to an unknown galaxy, then does it rotate around its centre and in doing so to a large extent keeps its form. Despite rotation of its own an elliptic, a barred or even a spiral galaxy virtually doesn't change its characteristic form. From this follows, that the inner stars of a galaxy are considerably slower on their way than the outer stars! But we expected exactly the opposite.
According to Kepler's regularity the outermost stars would have to orbit extremely slow, In order not to be hurled into space as a result of the centrifugal force. But then a galaxy wouldn't keep its structure. The spiral form, as it already has been observed and classified by Hubble (fig. 10.2), merely would be an accidental exception as a momentary picture, but by no means the rule.
We have to take note, that the structure and in particular the cohesion of a galaxy can't be explained with Kepler's laws[iii].

<i>: Konstantin Meyl: Electromagnetic environmental compatibility, Part 1 of this book: Causes, phenomena and natural scientific consequences.

<ii>: Kendrick Frazier: Das Sonnensystem, Time-Life Bucher, Amsterdam (1991)

<iii>: The basic laws of the universe start to rock: ,,What is the matter with the galaxies? They rotate in their fringe ranges much faster, as is allowed by the laws of physics. Or is something wrong with these venerable laws? The astronomers and physicists stand for the dilemma to have to decide between the two alternatives: feign the observations us an other world or do we calculate wrong since centuries?" (translated), Bild der Wissenschaft Nr. 2, 1989

Kepler's 1st law:

The planets move in elliptical orbits, with the sun at one focus.

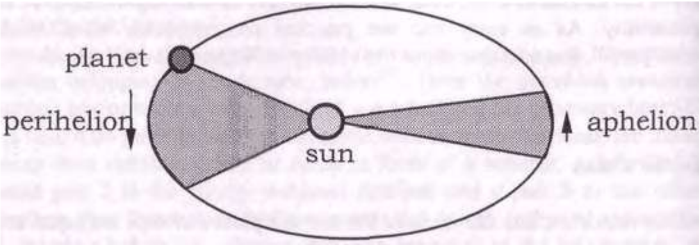
planet
perihelion
sun
aphelion

Fig, 10.1: Kepler's 2nd law (concerning the conservation of angular momentum):

The line drawn from the sun to the planet sweeps out equal areas in equal periods of time.

Kepler's 3rd law:

The ratio of the squares of the revolutionary periods of two planets is equal to the cube of their average distance to the sun:

$$\frac{t_1^2}{t_2^2} = \frac{r_1^3}{r_2^3} \quad (10.1)$$

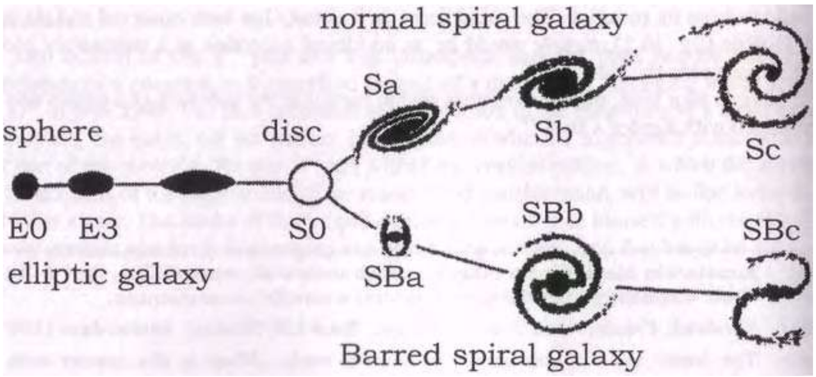
normal spiral galaxy
Sa
Sb
Sc
sphere
disc
E0
E3
S0
SBa
SBb
SBc
elliptic galaxy
Barred spiral galaxy

Fig. 10.2 The classification of the galaxies according to Hubble<i>

<i>: according to H. J. Lugt: Vortex Flow in Nature and Technology, page 223

10.2 Unknown interaction

So which interaction keeps a galaxy together? We today believe to know four different sorts.

I. The gravitation[i]: But since Kepler's law isn't valid in this case anymore, the gravitation is ruled out from the start. Obviously for the distances in a galaxy it hardly is effective.

II. The electromagnetic interaction: It is responsible for the structure of the atoms. Looked at from the outside atoms carry no charge, i.e. the charge conditions are balanced. A binding of our sun to the centre of the Milky Way by an electromagnetic interaction thus is ruled out as well.

III. The strong interaction: Since for the proton another charge distribution is measured, as a single positively charged particle should have according to the normally used theory, the strong interaction was introduced as a nuclear force, to explain the big error, the grave difference between measurement and calculation. The good advice hence reads: instead of giving birth to postulates at random, first of all the fault should be searched for in the normally used theory[ii]!

IV. The weak interaction: It quite obviously is involved in the particle decay[ii]. Both, the weak and the strong interaction, only have an extremely short range. With this property they consequently won't be able to keep a galaxy together.

Conclusion: In a galaxy a still unknown interaction takes effect, and science is requested to search it.

Both interactions with infinite range, the electromagnetic interaction and the gravitation occur as a result of static fields, therefore assume a constant charge or a constant mass. Considered more exactly in that case it merely can concern special cases. Gravitational waves, which reach our earth and which are detected in very costly experiments[iii], already show that the existence of alternating fields can't be excluded and oscillating interactions by all means are conceivable! The physical research at present probably is on the right track. The researchers however don't have ready an usable explanation yet. We accept the challenge!

---

[i] Derivation of Kepler's 3rd law in fig. 11.10

[ii] Konstantin Meyl: Potentialwirbel, Band 2
INDEL-Verlag, Villingen-Schwenningen 1992, ISBN 3-98O2542-2-4

[iii] Gero v. Randow: Wenn kosmische Katastrophen Raum und Zeit verbiegen, zum Thema Gravitationswellen-Detektor, VDI Nachrichten Nr.9, 1.3.91, S.32

Analogy:

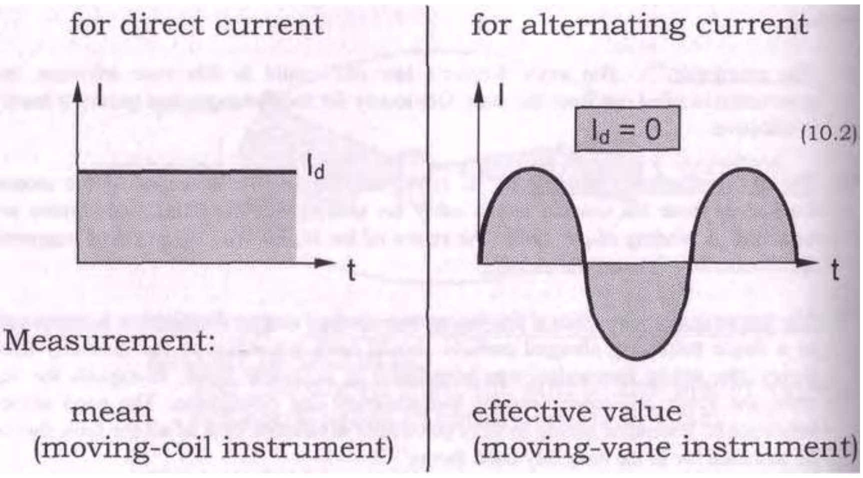
for direct current
for alternating current
I
I_d
t
I
$I_d = 0$
(10.2)
t
Measurement:
mean
(moving-coil instrument)
effective value
(moving-vane instrument)

Fig. 10.3: Direct current and alternating current

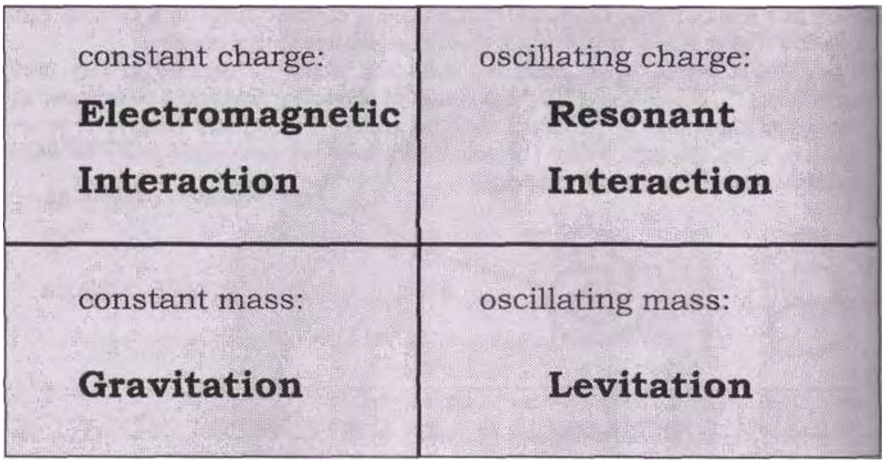

constant charge: **Electromagnetic Interaction**	oscillating charge: **Resonant Interaction**
constant mass: **Gravitation**	oscillating mass: **Levitation**

Fig. 10.4, a: The four fundamental interactions

10.3 Harmony of the alternating current engineers

Today's situation can be clarified by the following picture: you engage a "direct current engineer" to measure the tension voltage in the socket of our power supply system. The „dyed-in-the-wool direct current engineer", who never has heard anything of alternating current, reports: „No tension voltage is measurable". If he for reason of your doubtful expression looks still more exact he will realize: ,,The pointer of my moving-coil instrument strangely fidgets around zero, but the swing is so small that one confidently can touch it", (fig. 10.3)
Modern science is accustomed to say, without a valid theory and without technical measurability nothing can exist, what mustn 't exist. If you drop dead after the experiment, then you probably had a weak heart or other afflictions. In such cases as a rule the victim himself is to blame and by no means theoretical physics!

In the case of our power supply system the answer is known: The mean of the alternating voltage of the network is approximately zero. The pointer of a moving-coil instrument for reason of its inertia can't follow the fast changing anymore and only fidgets on the spot. The effective value however amounts to around 230 Volts. But to measure it you need another device, for instance a moving-vane instrument. Seen so, direct current describes the special case of alternating current with frequency zero.

The properties of alternating current can be depicted best by an extraterrestrial observer. He will tell us: Seen from a distance at least at night a great harmony seems to prevail on earth. All lights in the streets and cities twinkle completely synchronously. All generators are in resonance among each other and with all consumers. There are two big races: The 50 Hertz race and the 60 Hertz race, which appear if the earth turns further and the 50 Hertz race is switching off again its twinkling lamps.

The synchronization necessary for an exploitation of alternating fields is so obvious for us, that hardly anybody realizes, that a consumer operated with only one Hertz difference to the frequency of the network can't be supplied with power anymore. Apart from the correspondence in frequency it even depends on the correct phase. The phase angle must be between 0 and +90°. This corresponds to an efficiency factor $\cos\varphi$ between 1 and 0. The cable connections serve both the transport of current and the synchronization of all the generators and consumers connected to the network. The frequency is kept so constant that simple clocks, switching stations and even old record-players can be operated synchronously with the frequency of the network.
The synchronization of the feeding in power stations is supervised by a station of its own, which dictates the time. It is true that we aren't capable of seeing the twinkling of the lamps anymore for reason of the inertia of our eyes, but it can be detected and filmed with high-resolution cameras. Even if we can't perceive the harmony of the alternating current engineers, it nevertheless exists.

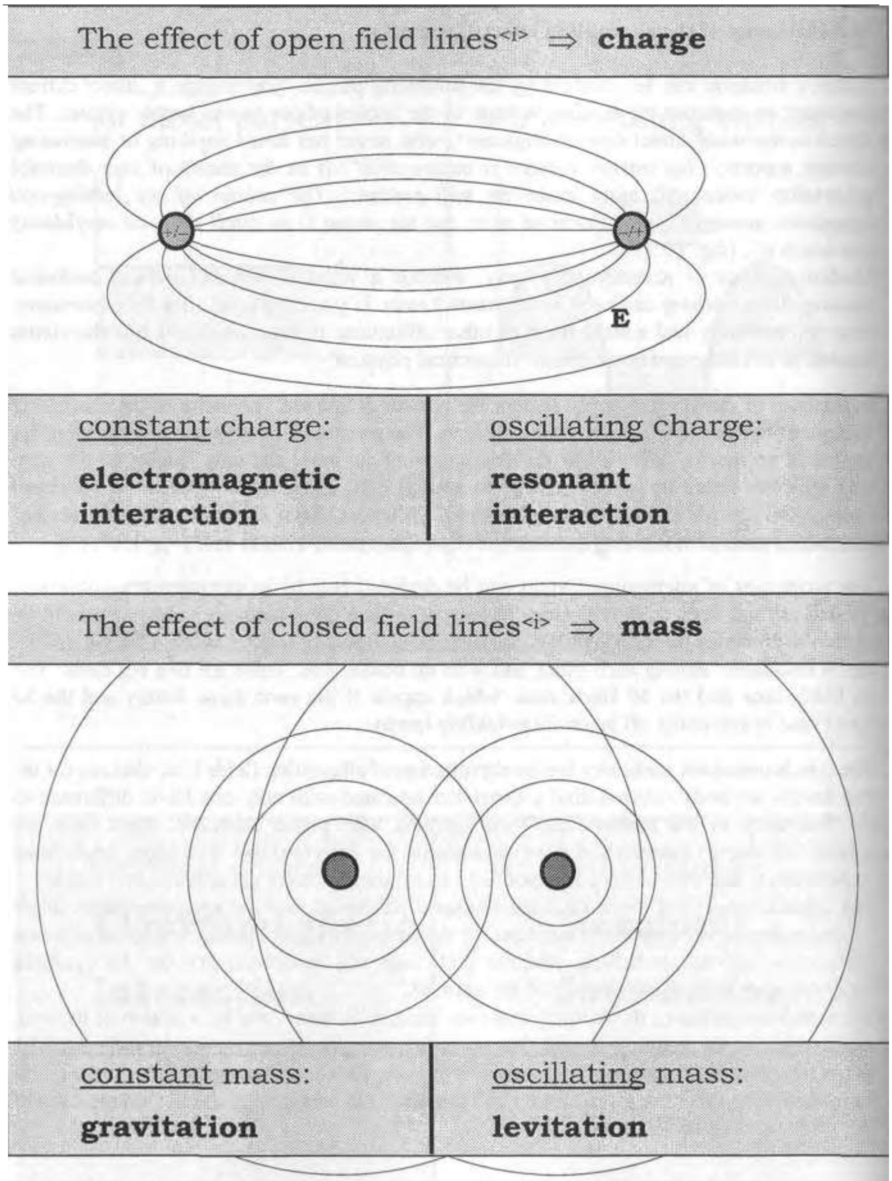
The effect of open field lines<i> ⇒ **charge**
+/-
-/+
E
constant charge:
electromagnetic interaction
oscillating charge:
resonant interaction
The effect of closed field lines<i> ⇒ **mass**
constant mass:
gravitation
oscillating mass:
levitation

Fig. 10.4, b: Explanation of the fundamental interactions

<i>: Konstantin Meyl: Potentialwirbel, Band 2, INDEL-Verlag (1992)
see also in Part 1, chapters 6.8 and 6.9

10.4 Four fundamental interactions

These considerations suggest, that also gravitation and electromagnetic interaction merely describe the special case of oscillating interactions with frequency zero. To avoid confusions, we'll have to think of new names.

The electromagnetic interaction can be clarified by means of open field lines, which start at a charged body and end again at another body of unequal charge. In physics classes it is normal to make the field lines of a magnet visible with iron filing. Between the unlike poles a force of attraction is observed<i>.

If we this time assume that both magnetic poles change their polarity at the same time, then the force will decrease for a short time during the changing of polarity, to afterwards be active again in full magnitude and in the same direction. Therefore a force of attraction is observed again even for a reversed polarity.

The generalization hence reads: The electromagnetic interaction will occur also in the oscillating case, but in weakened form, if both bodies involved in the interaction change their polarity synchronously and if they are in resonance. A name by analogy would be "resonating interaction" (table 10.4).

It is known of the electromagnetic interaction, that its effect is larger than that of the gravitation by powers often. This presumably has to do with the described and observable bundling up of the open field lines, whereas closed field lines can't be bundled up. The gravitation hence is linked with the closed field lines, which surround all elementary particles, every atom and every body<i>.

The opposite of the bundling up is the repulsion of open field lines, for which reason here also forces of repulsion can occur. For the gravitation however no repulsion is observed, because closed field lines virtually can't be influenced.

Apart from the circumstance that the effect generally will be smaller in the oscillating case, similar properties are to be expected. Also its range will be infinite as well. It is recommended to call the case of oscillating charges, as already said, "resonating interaction" and the case of oscillating masses, the oscillating gravitation, „levitation" (table 10.4).

The term „levitation" is very appropriate, but not new. Unfortunately until now no generally binding definition existed, what should be understood by that, for which reason misinterpretations and irritations can't be excluded. Mostly levitation is linked to a cancellation of gravity, up to a state of free floating, but we will see that quite other phenomena become describable with this term.

10.5 Resonating interaction

The question, what keeps a galaxy together, now can be answered unambiguously. The well-known interactions already have been excluded. If for the enormous distances the gravitation can't keep the outer stars anymore in accordance with the Kepler rule, then the levitation won't be able at all.

Example: central star S_z with 3 planets P_1-P_3
and with 4 neighbouring stars S_1-S_4

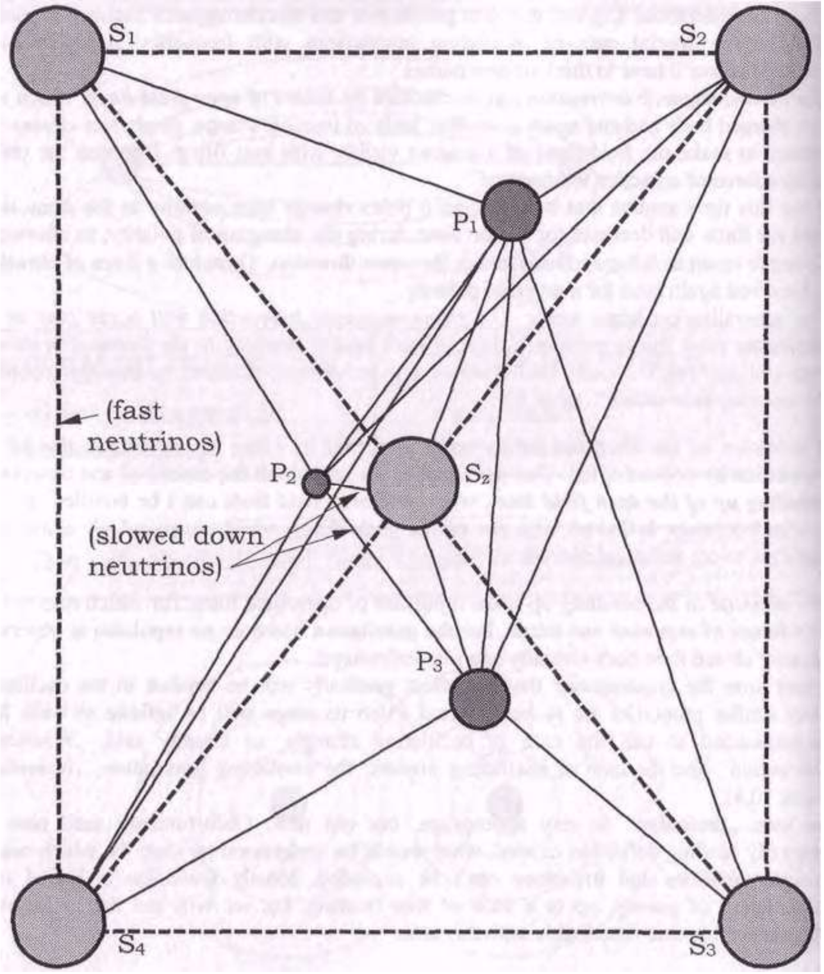
S_1
S_2
P_1
(fast neutrinos)
P_2
S_z
(slowed down neutrinos)
P_3
S_4
S_3

Fig. 10.5: The invisible threads of a resonant interaction

milky way-radius: $15000 pc*3*10^9 = 45*10^{16}$ km
sun system-radius: $50\ a*15*10^7 = 7,5*10^9$ km

$\frac{45\cdot10^{16}}{7,5\cdot10^9} = 1.27\cdot10^8$	the resonant interaction is more than eight decimal powers bigger than the gravitation!

Therefore the oscillating charge is left. Actually the resonating interaction will reach to the outermost areas of a galaxy. The bundling up of the field lines results in the centre of the galaxy and one of its stars to stand in an exclusive resonance to each other, what looked at from the outside looks like, as if the star hangs at an invisible string or a "rubber band" and thus turns around the centre.

Because quite a few stars hang at the centre of a galaxy, it can be assumed that it provides correspondingly many resonance. The centre perhaps is comparable with an extremely broad banded transmitter, which operates on all channels simultaneously. The stars then pick as a receiver the for them suitable channels and hang themselves by „rubber band" at the heap (fig. 10.5).

Should there exist any particles with an oscillating charge, which synchronize between centre and star with the resonating interaction, then they will mediate between both partners. If we assume that the centre at one channel just is positively charged, then all at the same time negatively charged particles will be attracted, the positively charged particles however repelled. Whereas the unlike particles in the centre participate directly in the production and maintaining of the oscillation, the like positively charged particles will be hurled into space.

But at the same time does a star, which clings to the centre, have to be negatively charged. It hence attracts the repelled particle. The particle thus drifts from the centre to the star, even then if all three, the centre, the star and the particle, change in polarity. The result is, that the stars grow in diameter by collecting the particles. Only because our sun actually grows, it has the chance, to sometime become a red giant!

Since the sun radiates, as is well-known, in every second a radiation equivalent of 5 million tons, it permanently has to be supplied with a substantially greater amount of matter. If a resonating interaction should occur, then our sun will get its „material" supplied from the centre of the Milky Way and that is a black hole!
But no particle with a mass comes out of such a hole, yes not even light. For a particle to be able to leave the black hole, it should have neither charge nor mass. At most an oscillating charge and mass would be allowed. Such a particle would have an enormous ability of penetration as a result of the missing static interaction. It would be able to rush through the earth unhindered.

According to actual knowledge only neutrinos have the described properties. One also knows that the centre of our Milky Way represents a mighty source of neutrinos. From this derivation follows:

1. As mediators of the resonating interaction serve synchronously oscillating neutrinos.
2. Starting with the proof of the neutrinos it should be able to backwards also prove the existence of the resonating interaction.
3. If, as a practical consequence, we imagine that the centre of the Milky Way wouldn't supply neutrinos anymore. Then the whole galaxy would fall apart and not one of its stars would shine anymore.

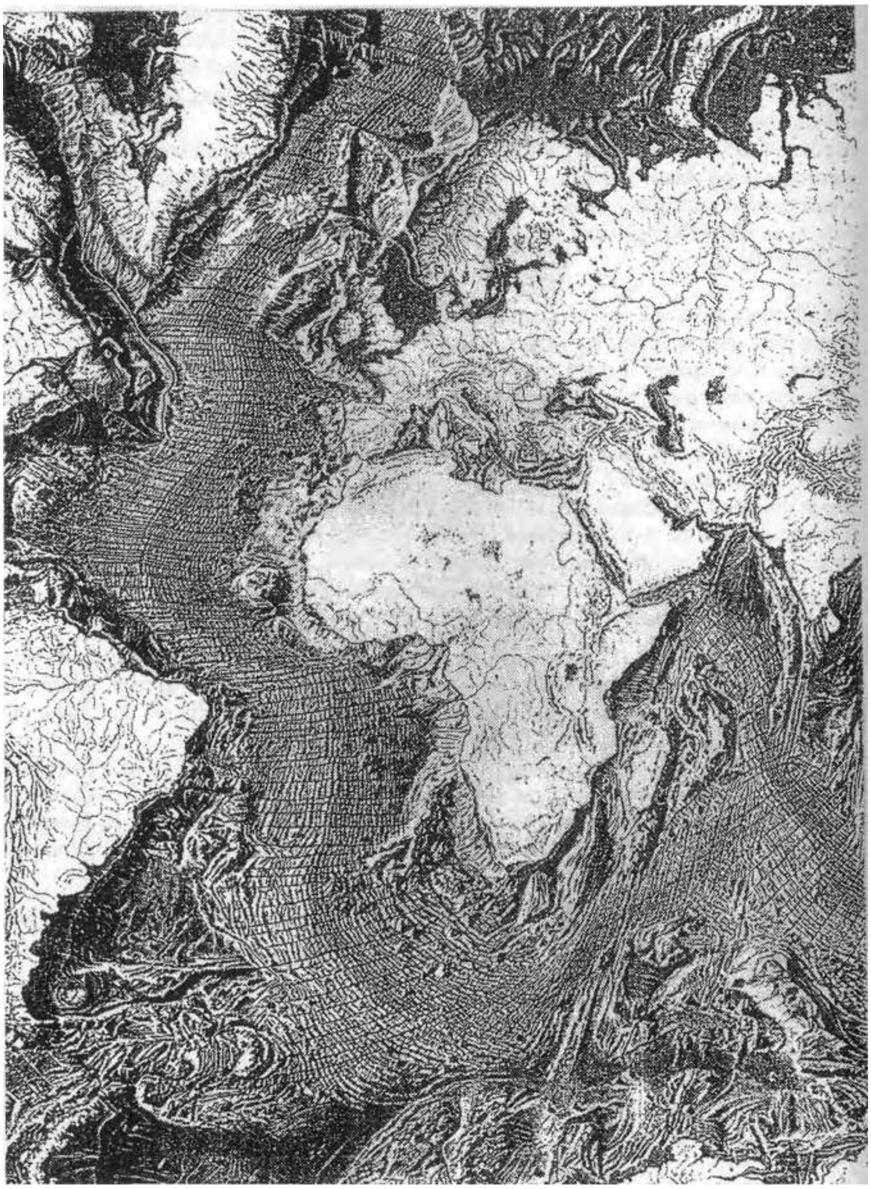

Fig. 11.1: The Atlantic ocean floor<i>

<i>: Miller, R.: Driftende Kontinente, Time-Life, Amsterdam 1991, S. 79

11. The growing globe

11.1 Long-distance effect of the neutrinos

The long-distance effect thus lies in the circumstance that the neutrinos, in the case of a resonance of the source of neutrinos and the receiver, span an invisible „rubber band" between both, which is called resonating interaction and keeps the two together.

As a transmitter of neutrinos functions for instance a supernova, the death of a star, in which 98% of the mass annihilates under emission of neutrinos or a black hole, which continually swallows light and matter and spits them out again as neutrinos after a still unexplored „process of digestion". The process, which in the case of a supernova takes place as a singular occurrence, in a black hole possibly takes place permanently. The hurled out neutrinos on the other hand serve the sun as a source of energy. A receiver of neutrinos then for instance is our sun. So that the hard and very fast cosmic neutrinos become utilizable for the sun, they at first have to be slowed down. But that is only partly successful:

1. Some very fast ones manage to pass through the sun and fly out again on the other side of the sun. The compared to the cosmic neutrinos strongly slowed down neutrinos then are called solar neutrinos.
2. Another part can be further slowed down and materialized. As a result of the oscillating mass of the neutrinos as well particles of matter as also some particles of anti-matter are formed. The particles of matter make the sun grow.

i. The with matter incompatible anti-matter annihilates under emission of light as is well-known. For this and for no other reason our sun shines!

Also the planets have such a neutrino reactor at their disposal. Only so the heat in the inside of the earth is explicable! It can be assumed that the planets materialize less the fast and hard cosmic neutrinos and that they are served more by the slowed down solar neutrinos, which our sun releases again unused.
As is well-known radiates the planet Jupiter already today twice as much radiation energy, as it from the sun receive. In this typical encyclopaedia type balance the involved neutrinos of course are not considered. But it shows that Jupiter is on the best way to become a sun itself. Its moons then will become planets.
From this the example we also see that with increasing mass the crust of the planet becomes thinner and thinner at the expense of the neutrino reactor in the inside, until it finally is eaten up and the celestial body openly starts to shine. Astronomers report for reason of their observations of the formation of more and more new stars.

One part of the collected neutrinos thus is materialized by the planet. In the case of our earth it contributes to its growth.
Who doesn't want to believe that the earth becomes bigger, should look at modern maps of the oceans, on which the topography of the ocean floor is shown. According to the theory of Alfred Wegener concerning the continental drift North and South America on the one hand and Europe with Africa on the other hand steadily drift apart since 200 million years. The result can be read at the atlantic ocean floor. The gaping, chequered rift zones to the right and left of the Mid-Atlantic Ridge show how the earth is torn apart (fig. (11.1).

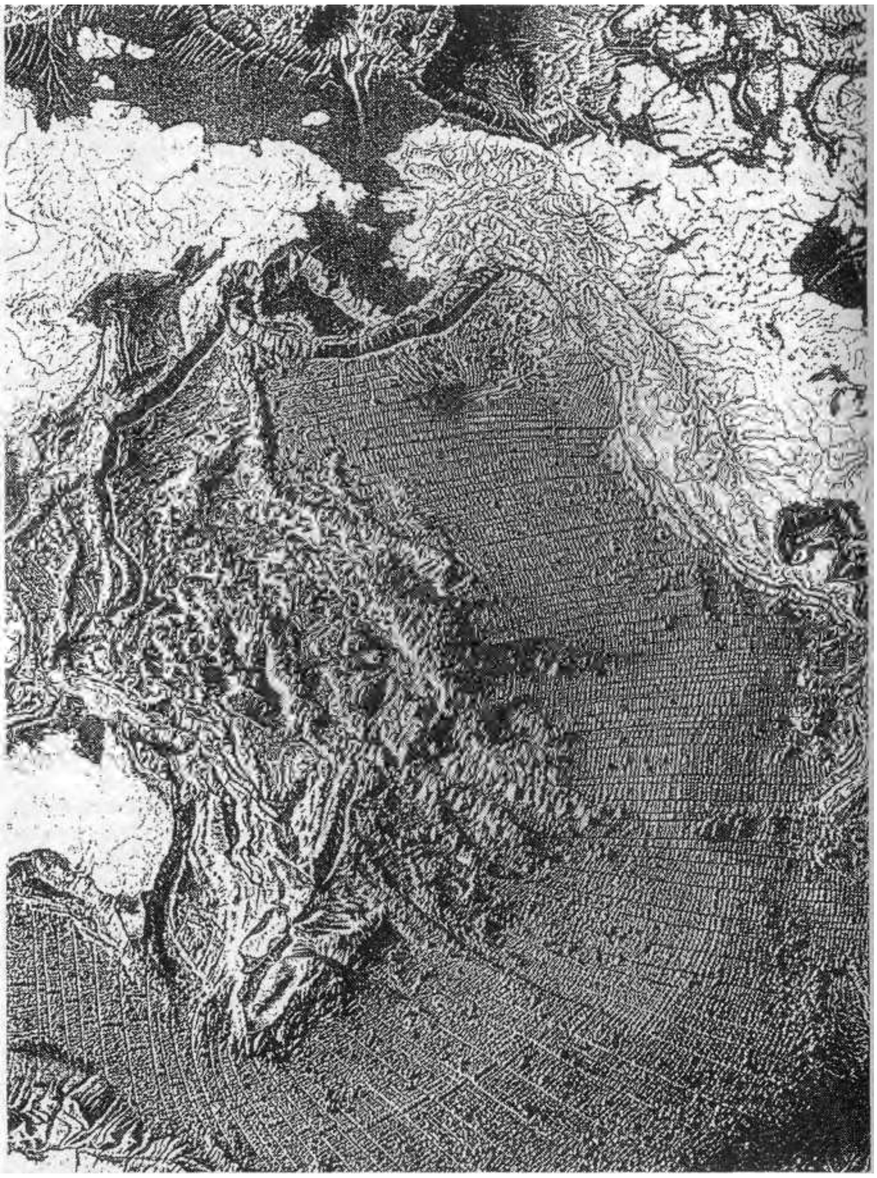

Fig. 11.2: The pacific ocean floor[i]

[i]: Miller, R.: Driftende Kontinente, Time-Life, Amsterdam 1991, S. 78

11.2 Model of expansion

Geographers assume that at another point to the same extent continental plates are subduced and molten. That thus has to happen in the Pacific Ocean. But the sea maps tell one just the opposite (fig. 11.2). At the ocean floor of the Pacific Ocean the same rift formations are found as in the Atlantic Ocean (fig. 11.1). That means that America drifts away from Australia and Asia exactly as from Europe, without being crushed or molten in doing so!
The only possible answer thus is: the earth grows.
The characteristic rift zones in addition are found around the Antarctic. From this the conclusion can be drawn that the 7th continent slowly is moving away, while the biggest habitable mass of land predominantly stays behind on the northern hemisphere, by Eurasia and North America forming a clip around the North Pole.

Concerning the evolution of the earth there are and were numerous approaches in various directions. Paul Dirac[<i>] at his time postulated, the brightness of the sun should decrease as a result of a decrease of the gravitational constant as well. In contrast to that Astrophysicists today observe just the opposite (Ingersoll 1987). According to the hypothesis of Carey[<ii>] energy will transform in matter in the universe. According to the idea of Oesterle[<iii>] aether particles are absorbed, which make our globe grow. Also other research scientists share the idea of the growth of the earth with him in their reports[<i4,5,6>]. As a geologist Oesterle[<iii>] cites his colleague Herzig: „The at the Mid-Oceanic Ridge newly produced oceanic crust has to be consumed again at other points, because otherwise the earth would expand" and criticizes the „plate tectonicians" that they would postulate their model without physical grounds.
He gives some arguments for the correctness of the model of expansion [<iii>]:

a) Subduction: The already discussed missing of zones of subduction and of melting of continental plates to the assumed extent.

b) Paleomagnetism: errors and mistakes in the hypothesis of the migration of the poles.

c) Continental depth drillings: They brought much higher temperatures in depths from 4000 meters, as expected and calculated according to models.

d) Stand of the sea water: Only if the water can spread in newly forming oceanic basins it can be explained, why the covering with water on earth continually is going back. This argument we want to investigate with a derivation of our own.

<i>: Dirac, P. A. M. Nature 139, p. 323 (1937)
<ii>: Carey, S. W.: Theories of the Earth and Universe.
Stanford University Press: 1-413, Stanford, California
<iii>: Oesterle, O.: Goldene Mitte: Unser einziger Ausweg,
Universal Experten Verlag, Rapperswil (1997), ISBN 3-9520261-9-0
<i4>: Hilgenberg, O. C: Vom wachsenden Erdball, S. 1-56, Eigenverlag,
Berlin (1933) und Neues Jahrb. Geol. Palaont. Abh. 116, 1, Berlin (1962)
<i5>: Jordan, P.: Die Expansion der Erde. Vieweg 1-180, Braunschweig (1966)
<i6>: Giancarlo Scalera, K.-H. Jacob: Why expanding earth? Institute Nationale di Geofisica e Vulcanogia, Roma and Technical University of Berlin 2003

Today's earth:

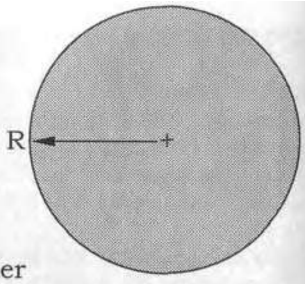

equatorial radius: $R = 6378$ [km]

earth's surface: $O_E = 4\cdot\pi\cdot R^2$
(sphere) $= \underline{5{,}112\cdot 10^8}[\mathrm{km}^2]$

29% of this area is land and 71% is water
for an average water depth of $t_w = 3.8$ [km] results a

water volume of: $V_w = 0{,}71\cdot O_E\cdot t_w = \underline{1{,}38\cdot 10^9}\ [\mathrm{km}^3]$ (11.1)

The earth **200 million years ago** (super continent: **Pangaea**):

today's land area (29%) was 100% of the earth's surface!

Pangaea surface area: $O_p = 0.29\cdot O_E$
$= \underline{1{,}48\cdot 10^8}\ [\mathrm{km}^2]$
$= 4\cdot\pi\cdot r_p^2$ (11.2)

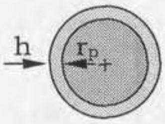

Pangaea radius: $r_p = \sqrt{O_p/(4\cdot\pi)}$
(corresponds to $= \underline{3435}$ [km]
the radius
of the sphere of shelf): = 54% of today's radius

Pangaea volume: $V_p = \frac{4}{3}\cdot\pi\cdot r_p^3 = \underline{1{,}697\cdot 10^{11}}\ [\mathrm{km}^3]$ (11.3)

Pangaea was covered with water h [km] high:

volume: $V = V_p + V_w = \frac{4}{3}\cdot\pi\cdot(r_p + h)^3 = \underline{1{,}71\cdot 10^{11}}\ [\mathrm{km}^3]$ (11.4)
= 15,7% of today's volume of the earth

$r_p + h = (3\cdot V/4\cdot\pi)^{1/3} = 3443{,}8$ [km] (11.5)

Pangaea water-level h above the NN (sea-level) at that time:

$\boxed{h = \underline{8{,}78}\ [\mathrm{km}]}$ (11.6)

Fig. 11.3: The calculation of the covering with water on earth 200 million years ago

11.3 Ancient Egyptian legend of Genesis

We don't have to search long for evidence. If we go back 200 million years in the history of the earth, as all continents still were united and formed the super continent Pangaea, as the 29% landmass of today thus constituted 100% of the earth's surface. At that time the diameter of the earth was almost half of today's diameter (exactly 54%). But if one distributes the amount of water of our oceans of today over the smaller earth, then possibly the water stood the young earth up to its ears in a first rough estimate. We now want to calculate how high the water stood.
For that we determine at first the water volume of today's oceans, by multiplying 71% of the earth's surface O_E with the average water depth. In doing so it is supposed that the water volume has not changed in the course of time. As an approximation this assumption could be correct if the factors, which influence the water volume, mutually compensate. On the one hand it has to be taken into account that in the process of fusion in the inside of the earth apart from other materials also juvenile water is formed, but on the other hand the water volume is reduced by photosynthesis and by the splitting of water molecules. In the case of the newly formed water it should concern roughly one cubic kilometre per year[<i>]. The photosynthesis however possibly can be estimated by means of the formed oxygen of the air just as the splitted water by the content of oxygen of the waters and the seas. Since the processes are subject to temporary fluctuations, the exact estimate is difficult. At least should an effect of compensation more or less be taken into account.

Next we calculate the surface of the earth 200 million years ago, the super continent, which Alfred Wegener called Pangaea. If the 29% continental land mass of today at that time constituted the whole surface, then the diameter of the earth at that time was determined at 54% of today's diameter, then the volume together with the water volume would amount to only 15.7% of today's volume and the water stood 8.78 kilometres high above the level NN at that time (fig. 11.3). With that even the highest peaks were under water.

Consequentially stands in the legend of Genesis of the ancient Egyptians:

„ They tell us that the earth was completely covered with water and that the earth rose from the water. It is talked about a primeval hill, of which creation took its start, on which the first sunrise and sunset was observed[<ii>].

If at that time life only existed in the water, of which we today still can detect the remains in excavations in mountains and plateaus, then it surely wasn't because the evolution had forbidden living on the land. There existed no land! All land was lying under water. But if, looked into the future, the land area increases further at the expense of the surface of the sea, then our earth sometime will dry up, as already other planets before us, e.g. our neighbouring planet Mars.

<i>: Oesterle, O.: Goldene Mitte: Unser einziger Ausweg, Universal Experten Verlag, Rapperswil (1997), ISBN 3-9520261-9-0

<ii>: Robert Bauval und Graham Hancock: Der Schlussel zur Sphinx, List Verlag (1996), S. 253 and in der Sendung: Die grossen Ratsel VII am 25.5.97 in S 3

a. **Today's earth**:

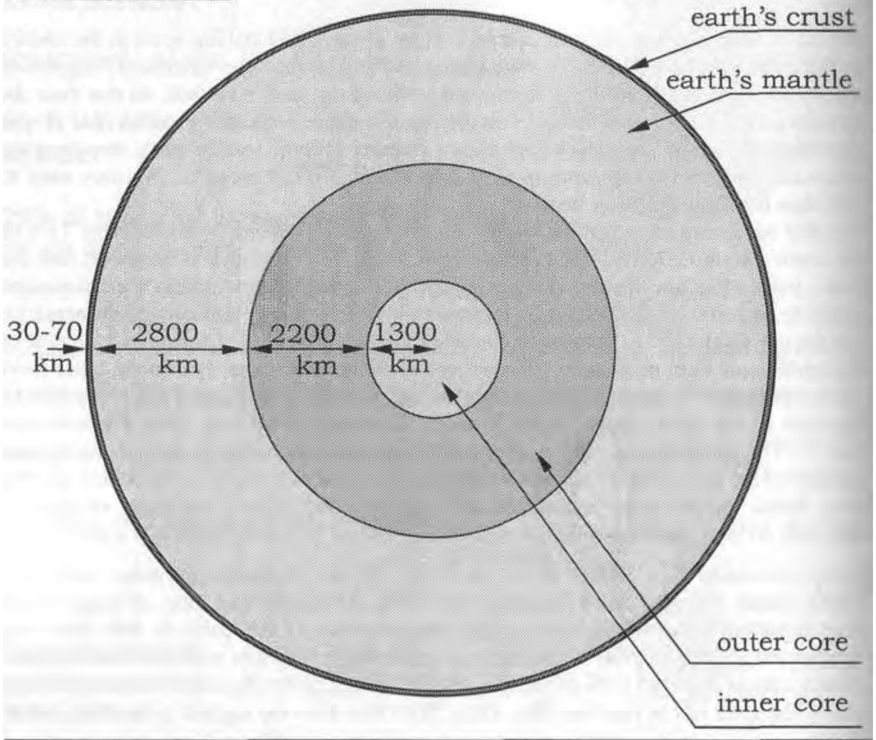
earth's crust
earth's mantle
30-70 km
2800 km
2200 km
1300 km
outer core
inner core

b. The earth **200 million years ago** (super continent: **Pangaea**):

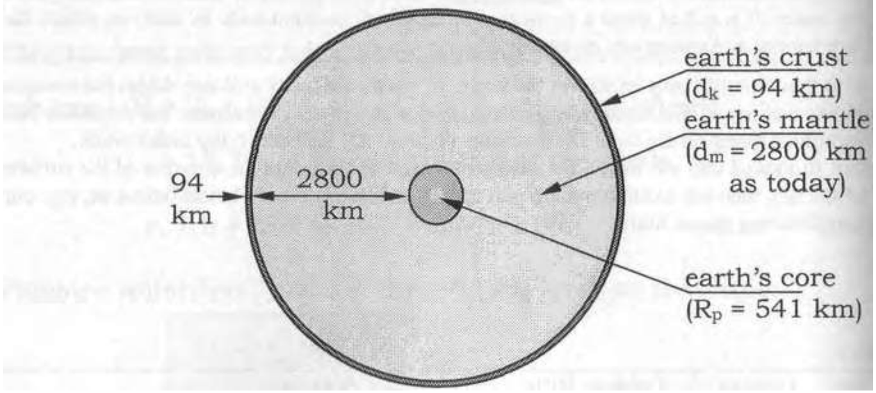
earth's crust
(d_k = 94 km)
earth's mantle
(d_m = 2800 km
as today)
94 km
2800 km
earth's core
(R_p = 541 km)

Fig. 11.4: The inner structure of the earth on an average

11.4 Inner structure of the earth

Next the question is raised: How fast does our earth actually grow? The calculated growth, distributed over the 200 million years, results in a yearly increase in the diameter of the earth of less than 0.1 mm. Carey[i] assumes 0.04 mm per year and Owen[ii] only 0.01 mm per year.
Actually the young earth must have been somewhat bigger than calculated, because as a result of the smaller gravitational acceleration the density of the matter must have been smaller. But this changes nothing to the relations, because the less dense earth was surrounded by likewise less dense water, the water-level nevertheless reached the peaks, as already calculated.
For indicating absolute linear measures and the calculation of the gravitational accleration the respective density should be considered. In most calculations the density is cancelled out, so that as well can be calculated with an unchanged density.

A grave error however lies in the assumption of a linear growth. Hilgenberg assumes an exponential growth[iii] and gives as a reason for the empirical approach of the e-function the „law of organic growth". In order to now not to speculate or to postulate in the same manner, we will derive and found our approach.

If namely the earth grows, then its core of fusion also grows, which causes the growth to take place accelerated, etc. A customer of a bank, who sees his amount of money grow according to such a regularity, will be given information immediately about the growth rate with a compound interest calculation.
But how big is the growing fusion reactor of our planet? According to today's level of knowledge about the structure of the earth the inner core is surrounded by the outer core and that again by the earth's mantle. On top floats the thin, but firm earth's crust, on which we live. The inner core has a radius of nearly 1390 km, the outer core stretches to a radius of 3500 km, whereas the crust is only between 10 and 78 km thick, dependent on the geographical latitude[i4] (fig. 11.4).

<i>: Carey, S. W.: Theories of the Earth and Universe.
Stanford University Press: 1-413, Stanford, California

<ii>: Owen, H. G.: Has the Earth increased in size? - In: New Concepts in Global Tectonics, Texas etc. University Press (1992), p. 289-295, Lubbock

<iii>: Hilgenberg, O. C: Vom wachsenden Erdball, Berlin 1933, Eigenverlag, Seite 31 und 32

<i4>: Mitton, S. (Herausg.): Cambridge Enzyklopadie der Astronomie
(The Cambridge Encyclopaedia of Astronomy), Orbis Verlag (1989)

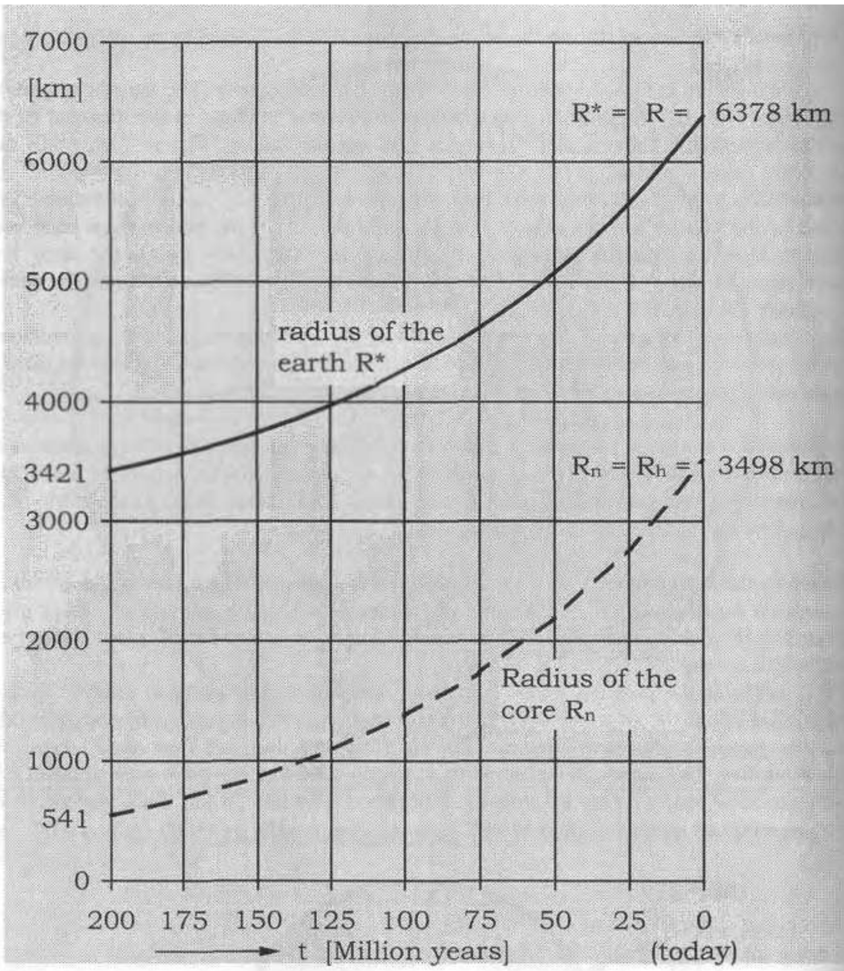

Fig. 11.5: The change in the radius of the earth and of the core in the course of millions of years until today.

calculation: (see fig.11.6)

$$R^* = 2879{,}5 + 541 \cdot q^{n/3} \text{ [km]} \qquad (11.7)$$

t years ago: with: $t = 2 \cdot 10^8 - n$ [a] and: $q = 1 + 28 \cdot 10^{-9}$

correspondents to the approximation:

$$R^* = 2880 + (6378 - 2880) \cdot e^{-t/\tau} \text{ [km]} \qquad (11.8)$$

with the time constant: $\tau = 107{,}1$ Mio. Jahre

11.5 Earth's core as a converter of neutrinos

I proceed from the assumption that the conversion of neutrinos and materialization in elementary particles takes place in the inner core. For the conversion no energy at all is used, because the inner energy together with the outer energy of the particle amounts to zero. The neutrinos merely have to be remodelled into another structure and for that they at first have to be slowed down with the help of the oscillating interaction.

During this process of slowing down, as said, no heat is formed because in the case of a mass less particle no energy can be set free in the domain where the classical law of conservation of energy is valid. Only after completion of the process of materialization we are able to detect mass and energy of neutrinos.

But if the oscillating interaction is taken as a basis, the oscillation with opposite phase between particle and earth's core, then contrary to all expectations a cooling down takes place. If the particle has reached its region of destination in the core, then the oscillations are overlapping. Mathematically seen they are added with reversed sign; they thus are subtracted. The result of the mutual compensation is the decrease of the thermal oscillation and the cooling down of the region which was expected.

In addition the formed particles with a mass mutually contract (see part I, chapter 8.3 and 8.4) and in doing so are further cooling down, as we will derive (chapter 12.7 with fig. 12.8). The physical limit of the process of contraction and cooling down is formed by absolute zero, at which no thermal oscillation at all occurs anymore, so that superconduction becomes possible with the result of giant electric currents and magnetic fields, which can be detected even at the earth's surface in damped form, for instance with a compass.

The necessary heat energy is flowing towards the quick-frozen inner core from the outside, principally from the outer core. Here, in the core, from the neutrinos slowed down to the speed of light various elementary particles are formed. Most of them immediately fall apart, to form other configurations. In the end only electrons and protons are preserved, which, as the only stable particles, can't fall apart anymore. These again are trying hard to take the state of an atom, which however needs very much space with the large distance between atomic nucleus and hull. Under the high pressure the enveloping electrons therefore will time and again fall into the nucleus to form neutrons together with the protons.

The neutrons need no atomic hull and can, as is well-known of neutron stars, take an extremely high density. In the case of the earth's core the neutrons however cannot be stabilized. The contraction to a neutron is accompanied by a corresponding drop in pressure, so that the neutron falls apart again. A continual oscillation of size is formed, with which the neutrinos again interact. With that also the high density of the earth's core would be explicable simultaneously.

In earth's outer core the various atoms and isotopes are formed, which in the sum release more energy than they absorb in their fusion processes. Here the fusion oven rages, which supplies the inner core with heat energy.

The formed matter is pushed further to the outside, rolls as a viscous mass through the earth's mantle and collects the surplus radiation and heat from the fusion oven.

With this model of explanation we now can tackle the calculation of the growth of the earth (fig. 11.6).

Growth in volume in analogy to the compound interest calculation:

duration: $n = 200$ million years

starting capital: $V_p =$ Pangaea volume $V_p = (4/3)\cdot\pi\cdot r^3$

final value: $V_h =$ today's volume $V_h = (4/3)\cdot\pi\cdot R^3$

savings bank formula: $V_h = V_p*q^n$ (compound interest) (11.9)

with „interest rate": $q = (V_h/V_p)^{1/n}$ (11.10)

radius of the core of the Pangaea sphere:

$R_p = r - d_m - d_k = 541$ [km] (11.11)

with

Pangaea radius: $r = 3435$ [km]

earth's mantle: $d_m = 2800$ [km]

earth's crust: $dk = 94$ [km] (incl. covering with water).

core radius today: $Rh = R - d_m - d_k = 3500$ [km] (11.12)

earth's crust today: $d_k* =$ 34-78 [km].

From $V \sim r^3$ the growth factor q is calculated to be

$$\underline{\underline{q}} = \left(\frac{V_h}{V_p}\right)^{\frac{1}{n}} = \left(\frac{R_h}{R_p}\right)^{\frac{3}{n}} = \left(\frac{3500\text{ km}}{541\text{ km}}\right)^{\left(\frac{3}{2\cdot 10^8}\right)} = \underline{\underline{1 + 28\cdot 10^{-9}}} \quad (11.13)$$

core radius after n years: $R_n = R_p*q^{n/3}$ (11.14)

radius of the earth after n years: $R* = R_n + d_m + d_k*$ (11.15)

Fig. 11.6: The calculation of the growth rate of the earth

11.6 Speed of growth

200 million years ago in the centre of the globe a core of fusion has formed and taken up its operation, probably under the influence of a cosmic occurrence connected with a high neutrino radiation. As a result the thin crust of the earth was torn apart and the oceanic basins were formed.

If we assume that the might of earth's mantle (with $d_m = 2800$ km) and crust (with less than 100 km) haven't fundamentally changed, then earth's core at that time had a radius of only 541 km. The „savings bank formula" now only may be applied for the core and only for its volume. On the condition of a constant neutrino density the volume of the core in every year will grow for one order of magnitude, which again depends on the respective volume itself. There results the in fig. 11.5 presented course of the radius of the core and of the earth.

According to our calculation the earth at present grows every year for $915*10^{11}$ tons, which corresponds to an increase in volume of 16500 cubic kilometres and an increase of area of 5.2 square kilometres. The earth momentarily grows for 6.5 cm per year in diameter, from which follows that the perimeter increases pi-fold and a continental drift of 10.2 cm per year is to be expected across both the Atlantic Ocean and the Pacific Ocean. Geologists today actually measure a plate movement of typically 10 cm, at individual points of up to 12 cm per year<i>!

Whoever likes to do handicrafts, can build together a globe of shells by himself. Hilgenberg for that gives a handicraft instruction<ii>. He draws the continents of a globe of today and cuts them out. Doing so, not the coast line of today is authoritative, but that of the edge of the shells, at which the mainland plates are breaking off into the deep sea. He hence also speaks of a sphere of shells and helps the handcrafters with the words: ..Because the paper shells of the sphere of shells owing to their strong curvature are difficult to nestle, we cut slits in the paper, which suitably lie there, where mountain ranges stretch and now can start with the gluing". Doing so it shows that the slits in particular in the case of the Ural and the Himalayan gape far apart, that in reversed direction in the case of the enlargement of the sphere of shells the lifting out of the mountains necessarily had to occur at these points by means of upsetting. Under these circumstances our model concept should be further rendered more precisely, If the change of the curvature of the growing surface of the earth is the cause for the lifting out of the mountains, then the surface of the earth 200 million years ago was structured merely by impact craters and by volcanic cones, then the amount of water may have been correspondingly smaller. The additional water of the oceans of today was collected by the earth either from the cosmos, by crossing the flight path of a comet with a water tail, or by the here discussed idea of a core of fusion in the inside of the earth it has produced the water itself!

How ever such detail aspects may have had an effect, it therefore nevertheless changes nothing to the model concept on the whole. Hilgenberg's globe of shells<ii> in my opinion to still makes more sense as all the models of explanation, as they are spread in today's text books.

<i>: Miller, R.: Driftende Kontinente, Time-Life, Amsterdam 1991, S. 78

<ii>: Hilgenberg, O.C.: Vom wachsenden Erdball, Berlin 1933, Eigenverlag

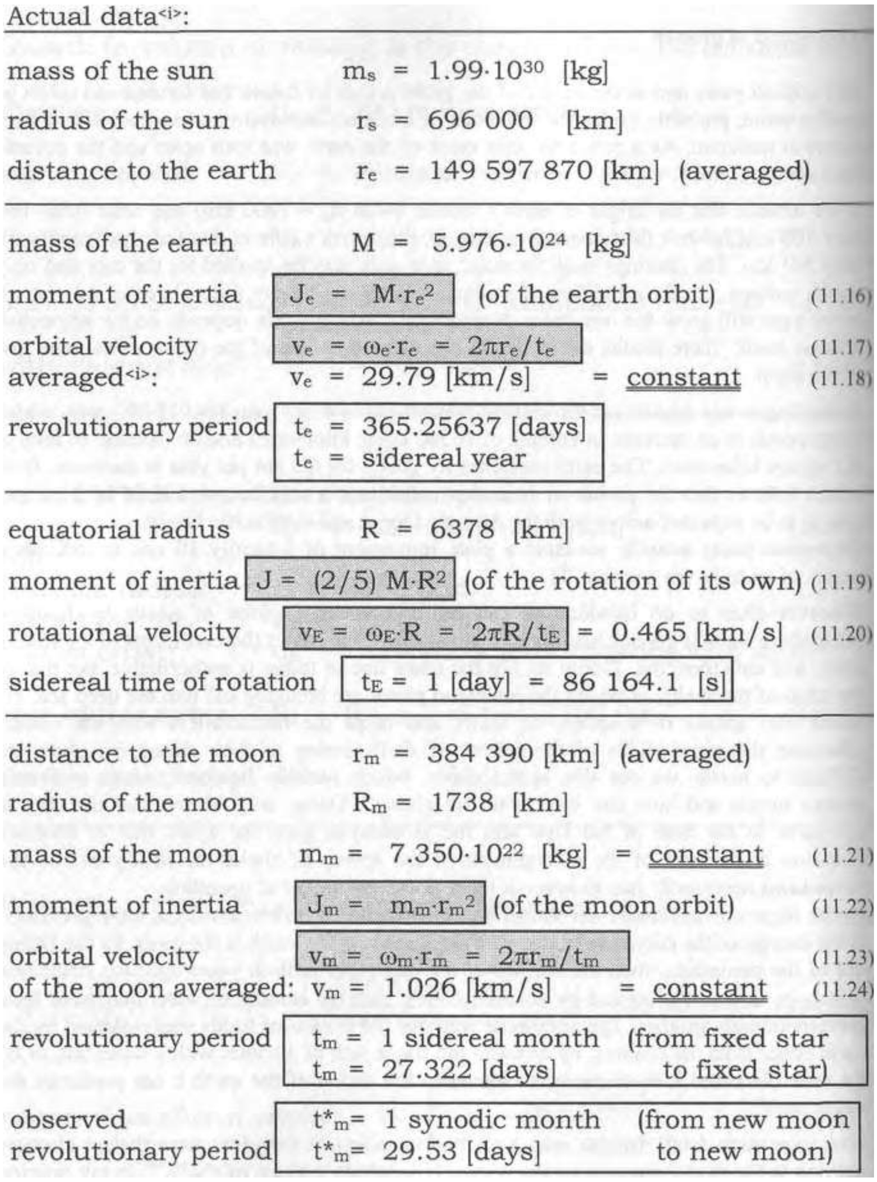

Actual data<i>:

mass of the sun $m_s = 1.99 \cdot 10^{30}$ [kg]

radius of the sun $r_s = 696\,000$ [km]

distance to the earth $r_e = 149\,597\,870$ [km] (averaged)

mass of the earth $M = 5.976 \cdot 10^{24}$ [kg]

moment of inertia $\boxed{J_e = M \cdot r_e^2}$ (of the earth orbit) (11.16)

orbital velocity $\boxed{v_e = \omega_e \cdot r_e = 2\pi r_e / t_e}$ (11.17)

averaged<i>: $v_e = 29.79$ [km/s] = <u>constant</u> (11.18)

revolutionary period $t_e = 365.25637$ [days]
t_e = sidereal year

equatorial radius $R = 6378$ [km]

moment of inertia $\boxed{J = (2/5)\, M \cdot R^2}$ (of the rotation of its own) (11.19)

rotational velocity $\boxed{v_E = \omega_E \cdot R = 2\pi R / t_E}$ = 0.465 [km/s] (11.20)

sidereal time of rotation $t_E = 1$ [day] = 86 164.1 [s]

distance to the moon $r_m = 384\,390$ [km] (averaged)

radius of the moon $R_m = 1738$ [km]

mass of the moon $m_m = 7.350 \cdot 10^{22}$ [kg] = <u>constant</u> (11.21)

moment of inertia $\boxed{J_m = m_m \cdot r_m^2}$ (of the moon orbit) (11.22)

orbital velocity $\boxed{v_m = \omega_m \cdot r_m = 2\pi r_m / t_m}$ (11.23)

of the moon averaged: $v_m = 1.026$ [km/s] = <u>constant</u> (11.24)

revolutionary period t_m = 1 sidereal month (from fixed star
t_m = 27.322 [days] to fixed star)

observed revolutionary period t^*_m = 1 synodic month (from new moon
t^*_m = 29.53 [days] to new moon)

Fig. 11.7: The data of sun, earth and moon of today<i>

<i>: Mitton, S. (Herausg.): Cambridge Enzyklopadie der Astronomie (The Cambridge Encyclopaedia of Astronomy), Orbis Verlag (1989)

It is an exciting matter, if one can experience for oneself, how the pieces of the puzzle fit together, how the southern end of America is wound around the Cape of Good Hope and the Falkland islands surface for the east coast of South Africa, in the vicinity of Australia how the Antarctic occupies the Pacific basin as a neighbour of Australia, von South, Middle and North America, run through by the equator, etc.

The corrections to the view of life of Alfred Wegener and the geographical evidence, which Hilgenberg gives, are sound, well founded and even after 65 years still highly actual. The physical statements of the geologists however, for instance the earth in the course of time would rotate faster and faster, cannot be followed.

11.7 Conservation of angular momentum

The question about the rotation of the earth is raised. Is it not changing at all, is it getting faster or slower? In the case of a with time growing earth there is only one possibility: The rotation of its own will decrease and not increase, as Hilgenberg supposes. For that you only need to place yourself on a turntable, to turn and if possible with weights in your hands stretch out your arms sideways, to feel, how the rotation of the table is decreasing. It surely would have done the geologist Hilgenberg good, to leave his desk for a short time for a stroll to the most nearby playground, for the purpose of the described physical experiment with himself.

In this case the law of conservation of momentum authoritatively has an effect in the formulation of the law of conservation of angular momentum for the rotating motion. According to that all angular momenta in the solar system should amount to zero. If we look at the planets which have no moon, then is remarkable that these need an eternity for a revolution around their own axis (Venus for instance needs 243 days). According to the law of conservation of angular momentum our earth owes its rotation of its own primarily the moon.

For reason of this relation we can assume a proportionality between the angular momentum of the moon J_m*w_m and that of the rotation of its own of the earth $J*w_E$ (eq. 11.39, fig. 11.10). They even have to be identical, if the partner of rotation earth and moon are seen as a closed system.

If the earth would be approached as a homogeneous, spinning sphere, the angular momentum at first would be too small for a factor 4.1 (eq. 11.29, fig. 11.9). The law of conservation of angular momentum dictates as a necessary result, that the dense core of the earth must rotate faster than the earth's crust! From the correspondence of orbital angular momentum of the moon on the one hand and the sum of the angular momenta of their own of earth's mantle and earth's core on the other hand results a 31 fold higher angular velocity of the earth's core compared to the rotation of the earth's surface (eq. 11.38, fig. 11.9).

Does our earth owe its geomagnetism this rotation of the core? At least the possibility exists that there exists a causal relation between the rotation of the earth's core and the geomagnetism. We'll further collect arguments and put the question concerning the formation of the geomagnetism under consideration.

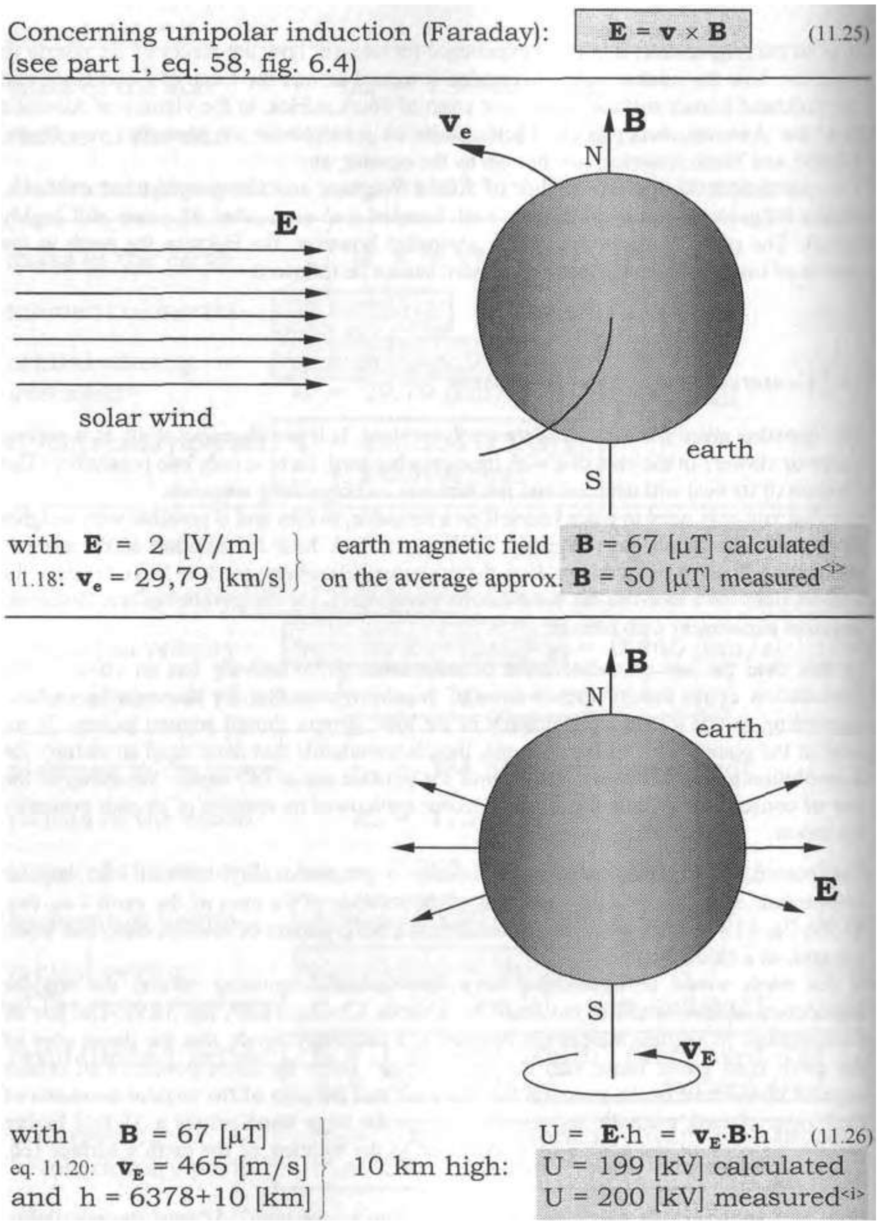

Concerning unipolar induction (Faraday): $\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (11.25)
(see part 1, eq. 58, fig. 6.4)

with $\mathbf{E} = 2$ [V/m]
11.18: $\mathbf{v_e} = 29{,}79$ [km/s]
earth magnetic field $\mathbf{B} = 67$ [µT] calculated
on the average approx. $\mathbf{B} = 50$ [µT] measured<i>

with $\mathbf{B} = 67$ [µT]
eq. 11.20: $\mathbf{v_E} = 465$ [m/s]
and $h = 6378+10$ [km]
10 km high:
$U = \mathbf{E} \cdot h = \mathbf{v_E} \cdot \mathbf{B} \cdot h$ (11.26)
U = 199 [kV] calculated
U = 200 [kV] measured<i>

Fig. 11.8: The magnetic and the electrostatic field of the earth
as a result of an unipolar induction

<i>: Measurement values in chap. 2.8, e.g. reference entry <i> (Prof. Dr. L. Konig)

11.8 Set of difficulties concerning the change of magnetization

Now investigations of the polarity of the magnetization of rocks have resulted in the fact that in irregular intervals, on an average every 500,000 years, a change of polarity of the geomagnetism has occurred and no one knows why.

If the rotation of the core should produce the magnetism, then it in any case also is responsible for the process of change of magnetization. That purely theoretically is possible in two ways: Either the core tilts out of its plane of rotation for 180° and rotates suddenly in the opposite direction or earth's mantle together with earth's crust, on which we live, is turned upside down.

It surely isn't my intention to produce any panic, but from a physical point of view only the second case is possible. Usually not the tail wags with the dog, but vice versa, the dog with the tail. The high rotational velocity of the earth's core exerts an extremely stabilizing influence. After all its angular momentum is 4.85 times larger than that of the mantle (eq. 11.33, fig. 11.9). Therefore the rotation of the core and the direction of earth's magnetic field always are preserved seen from the sun.

I further proceed from the assumption, that an electrostatic field arises from the sun accompanied by a particle flux, the solar wind, through which the earth flies through in a perpendicular direction. According to the Faraday law of induction $E = v \times B$ (eq. 11.25)

it experiences, as a result of the unipolar induction, a magnetic field which stands perpendicular on the ecliptic and thus dictates the orientation of earth's magnetic field. With that also the direction of rotation would be determined. The core thus by no means can tilt!

On the condition that the core doesn't rotate completely frictionless, the earth's mantle will in the case of the same direction of rotation be accelerated, in the case of unequal direction, after a changing of polarity, it will be slowed down again. If the process, for reasons of conservation of angular momentum, takes place alternatingly with a certain regularity, then the inhabitants of the earth for every changing of polarity might experience, how the North Pole in the shortest time turns over the equator to the South Pole, how the sun sets at the point, where it before had risen. As said, after a changing of polarity our earth is standing on its head! In the Bible corresponding clues are found<ii>. It can be taken from the media that one reckons with such a changing of polarity in the foreseeable future. When and if it actually takes place, stands however in the stars.

If magnetism thus is produced by the rotation of the core, or by electric currents in the earth's core as a result of the superconduction or by both, then the earth is aligned in the field of the sun like the magnetic needle in a compass. If we take as a cause a solar wind with an electric field strength of just 2 V/m, then this would not only determine the direction of earth's magnetic field, but also the order of magnitude, and that at present lies at 50 uT averaged over time (chap. 2.8).

With the same mathematical relation an electrostatic field around the earth of 199 V/m results as a result of the rotation of the earth. That fairly exact corresponds to the measured values!<i> Quite obviously all is related to each other. In the question for the „how" at this point consciously no definite answer is strived for. It rather should be discussed and worked to together in the seminar.

<ii>: The Bible, OT, The 2nd book of kings 20,9-11 (king Hezekiah) and Joshua 10,12-14, literally cited on the page after the next page (chap. 11.9), note<ii>

Angular momentum $J_m \cdot \omega_m$ of the orbiting moon (with eq. 11.22 + 11.23):

$$J_m \cdot \omega_m = m_m \cdot r_m^2 \cdot \omega_m = 29*10^{33} \ [\mathrm{kgm^2/s}] \qquad (11.27)$$

Angular momentum $J \cdot \omega_E$ of the rotation of the earth (eq. 11.19 + 11.20, same direction of turning):

(theoretically) $J \cdot \omega_E = (2/5)\ M \cdot R^2 \cdot \omega_E = 7*10^{33} \ [\mathrm{kgm^2/s}]$ (11.28)

$$\frac{\text{Orbital momentum moon}}{\text{Angular momentum earth}} = \frac{J_m \cdot \omega_m}{J \cdot \omega_E} = \frac{29}{7} = 4{,}1 \qquad (11.29)$$

Angular momentum $J_M \cdot \omega_E$ of earth's mantle and crust (R_h = core radius)

$$J_M \cdot \omega_E = (J - J_K) \cdot \omega_E = (2/5) \cdot (MR^2 - M_K R_h^2) \cdot \omega_E = 6*10^{33} \ [\mathrm{kgm^2/s}] \qquad (11.30)$$

$$\text{conservation of angular momentum:} \quad J_m \cdot \omega_m = J_K \cdot \omega_K - J_M \cdot \omega_E \qquad (11.31)$$

Angular momentum $J_K \cdot \omega_K$ of the earth's core

$$J_K \cdot \omega_K = J_m \cdot \omega_m + J_M \cdot \omega_E = 35*10^{33} \ [\mathrm{kgm^2/s}] \qquad (11.32)$$

$$\frac{\text{Angular momentum core}}{\text{Angular momentum mantle}} = \frac{J_K \cdot \omega_K}{J_M \cdot \omega_E} = \frac{35}{5} = 7 \qquad (11.33)$$

$$J_K \cdot \omega_K = (2/5)\ M_K \cdot R_h^2 \cdot \omega_K \qquad (11.34)$$

With the mass of earth's core $M_K = \rho_K \cdot V_K = \rho_K \cdot (4/3) \cdot \pi \cdot R_h^3$ (11.35)

and the averaged core density ro_k = 12000 kg/m^3

the angular velocity

of earth's core amounts to: $\omega_K = 3{,}3*10^3 \ [s^{-1}]$ (11.36)

and that at the earth's surface: $\omega_E = v_E/R = 0{,}073*10^{-3} \ [s^{-1}]$. (11.37)

$$\frac{\omega_K}{\omega_E} = \frac{3{,}3}{0{,}073} = 45 \qquad (11.38)$$

Fig. 11.9: Calculations concerning the conservation of angular momentum and the rotation of earth's core (One turn of the earth's core lasts 32 min. It results contrary to the earth rotation. Of the surface of the earth a period duration of 31 min should be observable.)

11.9 The weakly damped moon

The increasing angular velocity in the direction of the centre of the earth surely has something to do with the set of difficulties concerning the conservation of the spherical structure, comparable to the reason for the particle spin (chapter 6.13). The increasingly missing radial component of the gravitational field has to be compensated by an additional field produced by the rotation (according to part 1, equation 62, fig. 6.5).

The moon apparently doesn't know this set of difficulties. We can assume a constant mass for it (eq. 11.21). On the one hand is the moon smaller than the earth 200 million years ago, as it started to grow. On the other hand are doubts concerning the existence of an active core of the moon legitimate. In the Apollo-15 mission seismic gauges had been installed at the ground of the moon and at the start from the moon the produced seismic waves had been recorded<i>. The surprisingly small damping as the lunar module fell back onto the ground of the moon more likely allow the conclusion that the moon is hollow inside!
Also the small density of the moon points in the same direction, and finally no clues at all can be seen on the surface of the moon, which would point to a growth of the moon.

Apart from the mass also the orbital velocity of the moon is taken constant, what surely is true, as long as nobody and nothing drives the moon extra (eq. 11.24). The analysis of the law of conservation of angular momentum provides the in fig. 11.10 derived provisional result (eq. 11.41).

<i>: Moonquakes ever more mysteriously: „The by Apollo 12 triggered moonquake by „bombardment" of the surface of the moon with the clapped-out Lunar Module „Intrepid" at thursday evening puts scientists for bigger and bigger mysteries. An exact analysis of the measurement data now resulted in the by the astronauts Conrad and Bean installed „seismic station" on the moon ... to have recorded and to have sent to the earth 55 minutes long. ... The seismologist Dr. Gary Latham spoke of „important information about the structure of the moon" and meant, one now can „throw away the text book". One had reckoned that the impact of the Lunar Module on the moon would trigger a quake of at best some minutes", (translated) Frankfurter Rundschau vom 22.11.1969

---

concerning reference entry <ii>, chap. 11.8:

<ii>: The Bible, OT, Joshua 10,12-14: There stands in the „book of Jasher ": So the sun stood still in the midst of heaven, and hasted not to go down about a whole day. And there was no day like that before it or after it, ...

The 2nd book of kings 20,9-11: king Hezekiah: This sign thou ... let the shadow return backward ten degrees. And Isaiah the prophet cried unto the LORD: and he brought the shadow ten degrees backward, by which it had gone down in the dial of Ahaz.

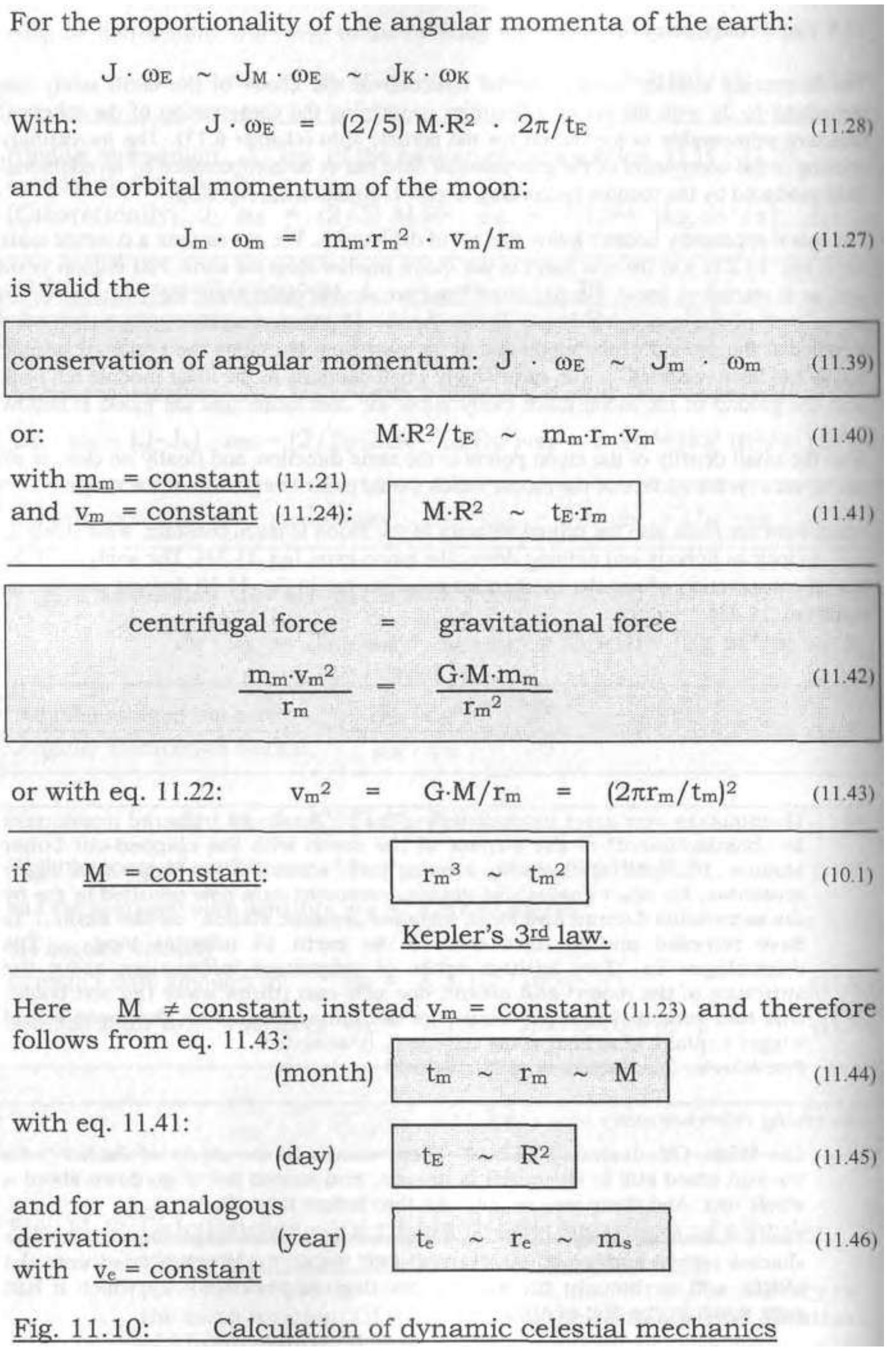

For the proportionality of the angular momenta of the earth:

$$J \cdot \omega_E \sim J_M \cdot \omega_E \sim J_K \cdot \omega_K$$

With: $$J \cdot \omega_E = (2/5)\, M\cdot R^2 \cdot 2\pi/t_E \qquad (11.28)$$

and the orbital momentum of the moon:

$$J_m \cdot \omega_m = m_m\cdot r_m^2 \cdot v_m/r_m \qquad (11.27)$$

is valid the

$$\boxed{\text{conservation of angular momentum: } J \cdot \omega_E \sim J_m \cdot \omega_m} \qquad (11.39)$$

or: $$M\cdot R^2/t_E \sim m_m\cdot r_m\cdot v_m \qquad (11.40)$$

with $\underline{m_m = \text{constant}}$ (11.21)

and $\underline{v_m = \text{constant}}$ (11.24): $$\boxed{M\cdot R^2 \sim t_E\cdot r_m} \qquad (11.41)$$

centrifugal force = gravitational force

$$\frac{m_m\cdot v_m^2}{r_m} = \frac{G\cdot M\cdot m_m}{r_m^2} \qquad (11.42)$$

or with eq. 11.22: $$v_m^2 = G\cdot M/r_m = (2\pi r_m/t_m)^2 \qquad (11.43)$$

if $\underline{M = \text{constant}}$: $$\boxed{r_m^3 \sim t_m^2} \qquad (10.1)$$

Kepler's 3rd law.

Here $\underline{M \neq \text{constant}}$, instead $\underline{v_m = \text{constant}}$ (11.23) and therefore follows from eq. 11.43:

(month) $$\boxed{t_m \sim r_m \sim M} \qquad (11.44)$$

with eq. 11.41:

(day) $$\boxed{t_E \sim R^2} \qquad (11.45)$$

and for an analogous derivation: (year) $$\boxed{t_e \sim r_e \sim m_s} \qquad (11.46)$$

with $\underline{v_e = \text{constant}}$

Fig. 11.10: Calculation of dynamic celestial mechanics

11.10 Calculation of celestial mechanics

We still need a further relation and try the balance of forces between the centrifugal force and the gravitational force. Both the to the outside directed centrifugal force and the to the inside directed gravitational force depend on the mass of the moving body, here the mass of the moon m_m, so that it is cancelled from equation 11.42.
As determining mass only that of the earth is left and that is taken constant in accordance with the knowledge of text books. The result of this assumption is Kepler's 3rd law (eq. 10.1). But beware, it here merely concerns a momentary picture! In the course of time, according to the derivation, the mass of the earth increases so that this assumption is untenable.

Instead, as already said, the average orbital velocity of the moon, one kilometre per second, can be taken constant (eq. 11.24). Since the orbital velocity is calculated from the proportion of the perimeter of the circular orbit $2\cdot\pi\cdot r_m$ with regard to the revolutionary period of a month (11.43, left page), and at the same time from the balance of forces a dependency of the mass of the earth M and the radius of the orbit of the moon follows (11.43, right page), the interesting relation (11.44) results:
Here the mass M, the radius r_m and the revolutionary period t_m stand in a direct proportionality to each other, and that means: if the mass of the earth increases in the course of time, then the moon will go away from us to the same extent, then also every month will get correspondingly longer.

Clarified with numeric values follows from a growth of the earth for $915*10^{14}$ kg per year a going away of the moon for yearly 5.88 meters (fig. 11.11). The going away of our satellite could be confirmed by means of measurements with laser reflectors, which had been put up on the moon by Apollo astronauts, according to dpa message<i>.

In addition every month lengthens for 3 milli seconds. That is valid for the sidereal month (= 27.322 days), for which a fixed star serves as a reference point for the measurement of a revolution, as well as for the synodic month, as it is observed from the earth from new moon till new moon (= 29.53 days). The synodic month today is longer than four weeks. But 3.7 million years ago it once actually lasted exactly 28 days, as we can calculate easily (fig. 11.11).

But now also the length of a day is changing. If we insert the proportionality (11.44) from the balance of forces in that of the conservation of angular momentum (11.41), then it is shown that a day depends quadratic on the radius of the earth (11.45), that a lengthening of every day for $4.5*10^{-9}$ s can be determined with the help of the growth curve of the earth (fig. 11.5). This is really very small, but 200 million years ago the day had just 19 hours, if extrapolated to 900 million years only 18.04 hours! The result of the american geologists around Dr. Charles Philip Sonett of the University of Arizona in Tucson also is 18 hours in their analysis of correspondingly old sediment formations, as the US science magazine Science has reported<i>. The correspondence of this measurement with our calculation without doubt has force of evidence!

<i>: 900 million years ago a day had 18 hours, Washington (dpa) 1997

Going away of the moon per year with $\boxed{M \sim r_m}$ (according to eq. 11.44): for

$\Delta r_m = r_m \cdot \Delta M/M = 384390[\text{km}] \cdot 9.15 \cdot 10^{16}[\text{kg/a}] \,/ 5.976 \cdot 10^{24}[\text{kg}] = 5.886\ [\text{m/a}]$

lengthening of month per year with $\boxed{M \sim t_m}$ (11.44) (sidereal): for

$\Delta t_m = t_m \cdot \Delta M/M = 27.322\ [\text{days}] \cdot 86164\ [\text{s/day}] \cdot \Delta M/M = 36\ [\text{ms/a}]$

synodic month (from new moon till new moon) of 28 days:

$\Delta t_m = 29.53\ [\text{days}] - 28\ [\text{days}] = 1.53\ [\text{days}] \equiv 132 \cdot 10^6\ [\text{ms}]$

linearly calculated $132 \cdot 10^6\ [\text{ms}]/36\ [\text{ms/a}] =$ 3.6674 mio. years ago.

Day length 200 mio. years ago, with $\boxed{t_E \sim R^2}$ (11.45) amounted to t* hours

shortening: $\Delta t_E/t_E = \Delta R^2/R^2 = (R-R^*)^2/R^2 = (6378-3420)^2/6378^2 = 21.5\%$

$\Delta t_E = 0.215 \cdot 24\ [\text{h}] = 5.16\ [\text{h}]$ and $t^* = 24\ [\text{h}] - \Delta t_E = 18.84\ [\text{h}]$

Day length 900 mio. years ago, ($R^*=3200$ km) amounts to 18.04 hours

present lengthening per day amounts to: $\Delta t_E/t_E = 4.5 \cdot 10^{-9}\ [\text{s/day}]$

but: the year increases 130,000 times faster than the individual day!

according to the measurement of Aristarchos 2300 years[ii] ago

is $\Delta t_e = 365.25637\ [\text{days}] - 365.25062\ [\text{days}] = 0.00575\ [\text{days}]$

and $\Delta t_e/t_e = 0.00575 \cdot 24 \cdot 60 \cdot 60 \cdot 1000 \,/\, 365.25637 \cdot 2300 = 0.59\ [\text{ms/day}]$

going away of the earth from the sun per year, with $\boxed{t_e \sim r_e}$ (acc. to eq. 11.46)

for: $\Delta r_e/t = r_e \cdot \Delta t_e/t_e \cdot t = 149.6 \cdot 10^6\ [\text{km}] \cdot \Delta t_e/t_e \cdot t = 1.024\ [\text{km/a}]$

growth of the sun per year/second, with $\boxed{t_e \sim m_s}$ (according to eq. 11.46)

for:

$\Delta m_s/t = m_s \cdot \Delta t_e/t_e \cdot t = 1{,}99 \cdot 10^{27}\ \text{t} \cdot \Delta t_e/t_e \cdot t = 1.36 \cdot 10^{19}\ [\text{t/year}]$

resp. in a second for: $= 4.3 \cdot 10^{11}\ [\text{t/s}]$

and at the same time a loss due to radiation of: $5 \cdot 10^6\ [\text{t/s}]$

Fig. 11.11: Figures according to analysis of some examples concerning dynamic celestial mechanics

11.11 The biblical age

We therefore owe the growing diameter of the earth that every day gets longer and longer (eq. 11.45) and from the going away of the moon a lengthening of the month results (eq. 11.44). Not only the months and the length of the days increase, but also the whole year. A corresponding derivation for the revolution of the earth around the sun, as it has been carried out for the revolution of the moon around the earth (eq. 11.44), delivers analogous results (eq. 11.46): The sun determines the length of the year and the distances to the planets. To the extent, to which the sun grows, the solar system increases in extension. Also the distance to the earth increases proportionally with the mass of the sun. By the going away from the sun the temperature on the planets however not necessarily decreases, because at the same time the radiation intensity of the growing central star increases. According to measurements it has increased for 30% since the formation of the solar system<i>.

If both the rotation of the earth and the revolution around the sun get slower, then by all means is conceivable, that the number of days per year approximately stays the same and mankind nevertheless gets less old. If we take the 2300 years old writing of the Greek Aristarchos about the sizes and distances of the sun and the moon<ii>. He determined, assuming a heliocentric view of life, the year to be 365.25062 days. Because the correction taken by Aristarchos concerned even the fifth place after the comma, we must assume that correspondingly precise gauges were available in Alexandria already 310 BC.
The today measured sidereal year with 365.25637 days has lengthened for whole 0.00575 days. From this follows that the year increases considerably faster than the individual day.
It of course would be nice, if we could calculate the lengthening of the year, but unfortunately no data about the growth of the sun are available.
If we proceed from the measurement of Aristarchos, without being able to verify or reproduce its reliability, then from that would result a going away of the earth from the sun for yearly one kilometre, then the sun should grow for $4.3*10^{11}$ tons per second. In any case the sun materializes mass considerably faster, as it loses mass in the same period as radiation equivalent, and that surely is correct (fig. 11.11). Today one generally assumes that „since its formation the earth has gone away from the sun for in total 30,000 km"<i>.

But if in the past the year consisted of less days and every day moreover was shorter, if therefore the biological life time was divided in shorter periods, then mankind could get older, then obtaining a biblical age possibly by no means was unrealistic. If Adam still should have got 930 years old, according to the 1st book of Moses, then his lifetime already must have been quite long ago. While Abraham still did get 175 years old, no successor of him has reached his age anymore. In the Bible it is said: „And the LORD said, My spirit shall not always strive with man, for that he also is flesh: yet his days shall he an hundred and twenty years "<iii>. Today even the limit given by the LORD isn't reached anymore!

<i>: Is the sun loosing her Gravitation? Illustrierte Wissenschaft, Nr.l, 1995
<ii>: Hermann Wild: Technologien von gestern, Chancen fur morgen,
Jupiter-Verlag Bern (1996), ISBN 3-906571-13-0, Seite 22
<iii>: The Bible, King James Version, Genesis 6.3

The balance of forces (equation 11.42 generalized):

$$\text{centrifugal force} = \text{gravitational force}$$

$$\frac{M \cdot v_k^2}{r} = \frac{G \cdot M \cdot m}{r^2} \tag{12.1}$$

results in the cosmic velocity v_k:

$$v_k^2 = G \cdot m/r \tag{12.2}$$

independent of the mass M of the satellite or planet!

For	$v<v_k$	falling back into the central mass m
For	$v>v_k$	taking the leave into space
For	$v=v_k$	stationary orbit

written down for the nth planet in the solar system:

with $r(n) = r_n$ average orbital radius of the planet

and $v(n) = v_n$ average orbital velocity

as well as: m_s mass of the sun

$$v_n = \sqrt{G \cdot m_s/r_n} \tag{12.3}$$

or:

$$v_n = G \cdot m_s/v_n \cdot r_n \tag{12.4}$$

Fig. 12.1: The first cosmic velocity v_k

12. New cosmology

Astronomy still hasn't satisfactorily solved the question of the formation of the solar system. It thereby however concerns a central problem, because it includes the origin of the earth. But as long as we not yet have understood the relations in our nearest environment, the processes, which we observe with giant telescopes in the depths of pace, will remain a book with seven seals, will lose models of explanation concerning the Big Bang and concerning the so-called unavoidable heat exitus every reliability.

12.1 Concerning the formation of our solar system

2-8-1972 the observatories of the sun reported an unusual high solar activity and 6 days later a slowing down of the rotation of the earth occurred, which was recorded as the lengthening of a day for more than 10 milliseconds. This effect hardly can be explained by the tidal friction alone already just concerning the energy balance. Instead this observation makes clear two things to mankind. On the one hand, how much our earth depends on the solar processes and on the other hand, that the changes by all means can occur not continuously, but periodically and if need be even sporadic.

If, as a result of the conservation of angular momentum, the sun determines the orbital velocity of the earth, if it dictates the rotation of the earth by its neutrino activity and the growth of the earth and if the earth in the same manner determines the orbit of the moon, then it would be obvious that the moon originally has been a part of the earth and this in turn sometime a part of the sun. As it came off, the necessary angular momentum then has been passed on proportionately to the celestial companions, with which the cause for evolution and rotation of their own would be clarified. In addition the moon goes away from the earth and the earth again from the sun, so that looking back it by all means would be obvious, if they once had belonged together.

If we assume that cosmic dust particles meet and accumulate, then all collected rubble contributes to the rotation of its own of the forming celestial body. The more matter finds together, the larger its force of attraction gets, the faster it will grow like a celestial vacuum cleaner. In the course of time this process however is slowed down again and eventually comes to a standstill, because as matter condenses, volume and spherical radius decrease and the rotation of its own increases to the corresponding extent. The celestial body rotates faster and faster and reaches at its surface the cosmic velocity vk, which is given by the mass m and the radius r of the star (eq. 12.2). Now the centrifugal force has reached an order of magnitude, for which the celestial body hurls exactly as much matter into space, as it on the other hand collects by its gravitational effect.

Our sun was lucky to have been supplied with neutrinos in the range of influence of our galaxy. It went in resonance and started to grow, this time from out of the inside. It however could not yet shine, because a crust had formed on its surface around its core of fusion and its mantle, a crust on which permanently was falling cosmic matter from the outside. The increasing viscosity of the sun becoming compressed caused a slowing down of the core of the sun and the corresponding acceleration of the mantle and crust.

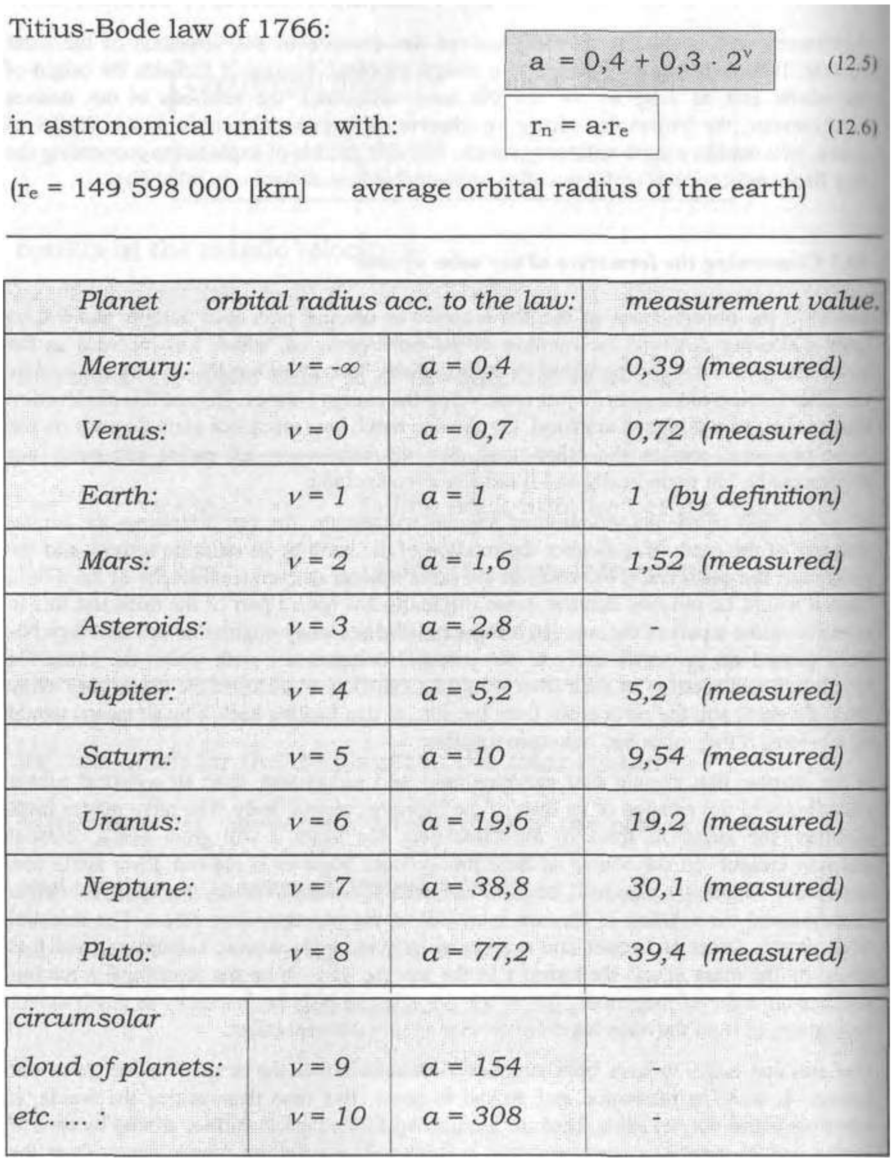

Titius-Bode law of 1766: $\boxed{a = 0{,}4 + 0{,}3 \cdot 2^{\nu}}$ (12.5)

in astronomical units a with: $\boxed{r_n = a \cdot r_e}$ (12.6)

(r_e = 149 598 000 [km] average orbital radius of the earth)

Planet	*orbital radius acc. to the law:*		*measurement value.*
Mercury:	$\nu = -\infty$	$a = 0{,}4$	*0,39 (measured)*
Venus:	$\nu = 0$	$a = 0{,}7$	*0,72 (measured)*
Earth:	$\nu = 1$	$a = 1$	*1 (by definition)*
Mars:	$\nu = 2$	$a = 1{,}6$	*1,52 (measured)*
Asteroids:	$\nu = 3$	$a = 2{,}8$	-
Jupiter:	$\nu = 4$	$a = 5{,}2$	*5,2 (measured)*
Saturn:	$\nu = 5$	$a = 10$	*9,54 (measured)*
Uranus:	$\nu = 6$	$a = 19{,}6$	*19,2 (measured)*
Neptune:	$\nu = 7$	$a = 38{,}8$	*30,1 (measured)*
Pluto:	$\nu = 8$	$a = 77{,}2$	*39,4 (measured)*
circumsolar cloud of planets:	$\nu = 9$	$a = 154$	-
etc. ... :	$\nu = 10$	$a = 308$	-

Table 12.2: The Titius-Bode series of the planets (the theoretical values compared to the measurement values)

Sometime our like mad spinning sun had increased that much, that the crust came off like the „tread of an old car tyre" and was catapulted into space. The repulsion was achieved by the centrifugal force exceeding the gravitation for the force of cohesion, which at the moment of the separation of crust regions suddenly tears off. Like snowballs the planets lulled off the sun's surface in this process and were hurled away. Their velocity of rotation at this time was identical to the cosmic velocity of the sun v_k and with that for many a planet large enough to produce its own satellites out of its own surface, which perhaps was not yet completely ideally spherical.

With every planet, which the sun gave birth to, it gave away a part of its own angular momentum to its child as orbital momentum. Only from this time on the sun reduced its angular velocity steadily until this very day. By losing the crust it also started to shine openly. Thus at least could our solar system have been formed.

12.2 The birth of the planets

From this observation various consequences result. If stars are observed, which are rotating very fast, then they either are very young, or they have no planets. Stars, which compared to our sun are rotating less fast, have given away the angular momentum to their planets and such, which hardly are rotating, have their planets already sent away into space. But if the last planet leaves its solar system and the sun stops to rotate, then the sun, which meanwhile has grown to a red giant, without a, the spherical form stabilizing centrifugal force, will collapse. The supernova is the death of a star and thereby neutrinos are set free, the material for new life.

After the coming off of the planets these first clear free their flight paths, by together with the sun collecting the flying around matter. Even whole planets thus can collide, are slowed down and form bigger units. Finally only some few planets in very particular orbits are left. Their average distance to the sun obeys in an until now completely inexplicable manner the Titius-Bode series (equation 12.5, table 12.2).
In the case of the by the german scientist Titius 1766 formulated regularity it concerns a rule of thumb founded purely on experience. But it is remarkable, that the planet Uranus could be predicted (Bode 1772) and after systematic searching also be found (Herschel 1781) with it.

The most distant planets Neptune and Pluto however don't obey the law. Their distance should be very much larger. But this circumstance we already had given reasons for with the fact, that for them the resonating interaction already overlaps and exceeds the gravitation in its effect (chap. 10.1 note<ii> and chap. 10.5).

The arbitrary seeming series of numbers 0, 3, 6, 12, 24, 48, 96, 192, 384, ... with a respective doubling of the value, starting with the 3, the addition of 4 and the following division by 10 at first are nothing but pure acrobatics of numbers, which now really has nothing to do with physics (table 12.2). A physical background can be supposed however because of the tried and tested applicability and that should be fathomed.

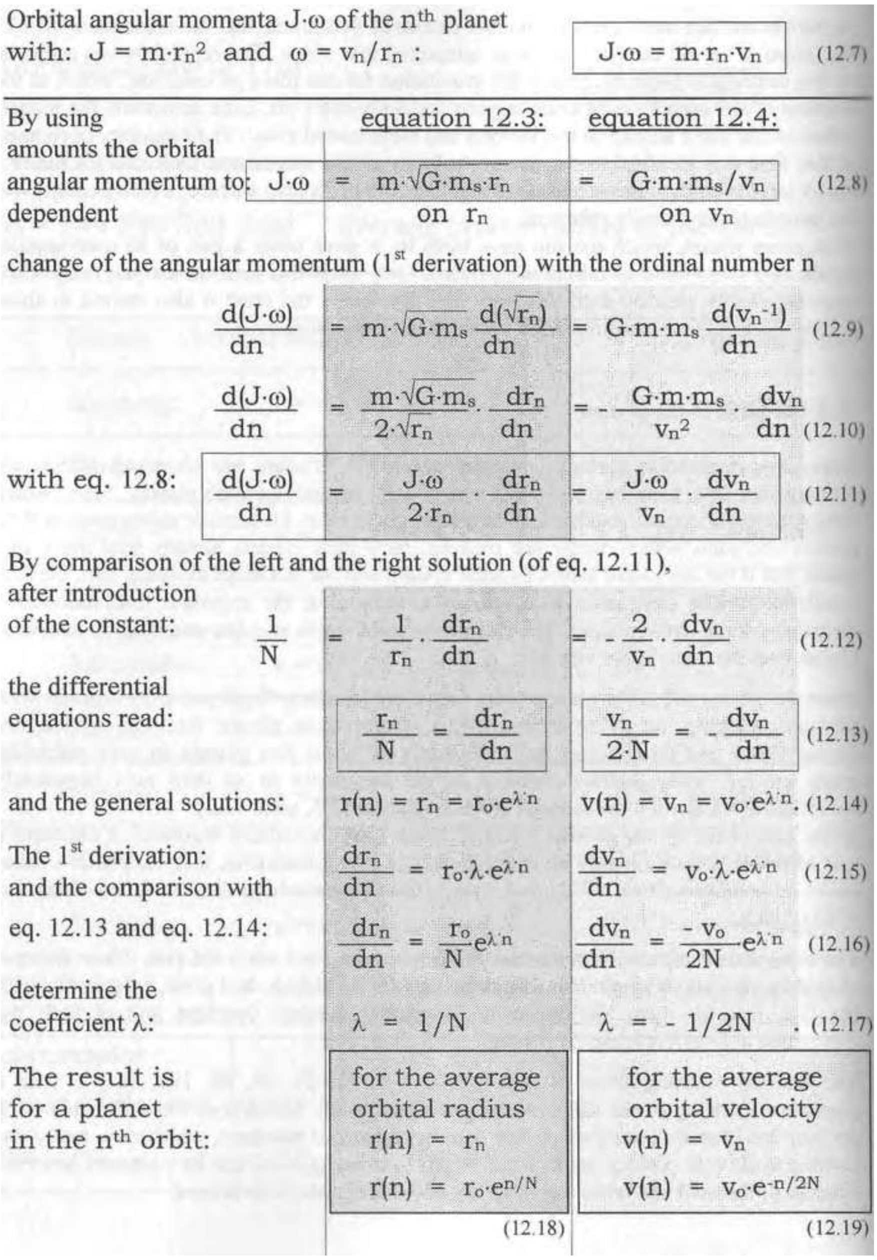

Orbital angular momenta $J\cdot\omega$ of the nth planet
with: $J = m\cdot r_n^2$ and $\omega = v_n/r_n$:

$$\boxed{J\cdot\omega = m\cdot r_n\cdot v_n} \qquad (12.7)$$

By using equation 12.3: equation 12.4:
amounts the orbital angular momentum to:

$$J\cdot\omega = m\cdot\sqrt{G\cdot m_s\cdot r_n} = G\cdot m\cdot m_s/v_n \qquad (12.8)$$

dependent on r_n on v_n

change of the angular momentum (1st derivation) with the ordinal number n

$$\frac{d(J\cdot\omega)}{dn} = m\cdot\sqrt{G\cdot m_s}\,\frac{d(\sqrt{r_n})}{dn} = G\cdot m\cdot m_s\,\frac{d(v_n^{-1})}{dn} \qquad (12.9)$$

$$\frac{d(J\cdot\omega)}{dn} = \frac{m\cdot\sqrt{G\cdot m_s}}{2\cdot\sqrt{r_n}}\cdot\frac{dr_n}{dn} = -\frac{G\cdot m\cdot m_s}{v_n^2}\cdot\frac{dv_n}{dn} \qquad (12.10)$$

with eq. 12.8:

$$\frac{d(J\cdot\omega)}{dn} = \frac{J\cdot\omega}{2\cdot r_n}\cdot\frac{dr_n}{dn} = -\frac{J\cdot\omega}{v_n}\cdot\frac{dv_n}{dn} \qquad (12.11)$$

By comparison of the left and the right solution (of eq. 12.11), after introduction of the constant:

$$\frac{1}{N} = \frac{1}{r_n}\cdot\frac{dr_n}{dn} = -\frac{2}{v_n}\cdot\frac{dv_n}{dn} \qquad (12.12)$$

the differential equations read:

$$\boxed{\frac{r_n}{N} = \frac{dr_n}{dn}} \qquad \boxed{\frac{v_n}{2\cdot N} = -\frac{dv_n}{dn}} \qquad (12.13)$$

and the general solutions:

$$r(n) = r_n = r_o\cdot e^{\lambda\cdot n} \qquad v(n) = v_n = v_o\cdot e^{\lambda\cdot n}. \qquad (12.14)$$

The 1st derivation:

$$\frac{dr_n}{dn} = r_o\cdot\lambda\cdot e^{\lambda\cdot n} \qquad \frac{dv_n}{dn} = v_o\cdot\lambda\cdot e^{\lambda\cdot n} \qquad (12.15)$$

and the comparison with eq. 12.13 and eq. 12.14:

$$\frac{dr_n}{dn} = \frac{r_o}{N}e^{\lambda\cdot n} \qquad \frac{dv_n}{dn} = \frac{-v_o}{2N}e^{\lambda\cdot n} \qquad (12.16)$$

determine the coefficient λ:

$$\lambda = 1/N \qquad \lambda = -1/2N\ . \qquad (12.17)$$

The result is for a planet in the nth orbit:

for the average orbital radius
$r(n) = r_n$
$r(n) = r_o\cdot e^{n/N}$
(12.18)

for the average orbital velocity
$v(n) = v_n$
$v(n) = v_o\cdot e^{-n/2N}$
(12.19)

Fig. 12.3: Calculation of the distances and orbital velocities of the planets

12.3 The derivation of the Titius-Bode law

Even if Bohr's atomic model should be wrong, it gladly is compared to the system of the planets. The radii of the electron orbits mathematically result as eigenvalue solutions of the Schrodinger equation, and that we have derived from the fundamental field equation (chap. 5.1, eq. 15) (chap. 5.5 - 5.9). The orbital radii are calculated in increasing order in accordance with the sequence of whole numbers with n = 1, 2, 3, 4, ...

From the same sequential regularity of the planetary orbits can be derived, that they also obey the eigenvalues of the same fundamental field equation, which isn't called a world equation in vain. The quantitative distances from the sun are determined by the size of the sun: if the sun increases, then also all distances increase to the same extent.

Now result identical distances between the orbits from Bohr's model, whereas this distance in the case of the planets with increasing distance from the sun gets larger. The reason for this unevenly grading can be calculated just like that (fig. 12.3).

We arbitrarily pick a planet, which occupies the orbit n, where n again represents the sequence of whole numbers (n = 0, 1, 2, 3, ...). If the orbit of this n^{th} planet changes, then also its distance to the sun $r_n = r(n)$, its orbital velocity $v_n = v(n)$ and the orbital momentum $J \cdot \omega$ (12.7) are changed.
We try (as in equation 11.42, fig. 11.10) the balance of forces between centrifugal force and gravitational force (eq. 12.1, fig. 12.1) and solve equation 12.2 for the orbital velocity. The orbital angular momentum of the planet written down once in its dependency on r_n and in the right column next to it on v_n (eq. 12.8, fig. 12.3), is derived for the orbital ordinal number n, to record the change in angular momentum (equations 12.9 to 12.11).

In fig. 12.3 the paths to the solution for both cases are given. For the average orbital radius r(n) as well as for the average orbital velocity v(n) of the planet an exponential course (eq. 12.18 and 12.19) and in logarithmic representation a straight line (fig. 12.4) results.

Even the orbits of the distant planets Neptune and Pluto, for which the Titius-Bode series fails, now are correctly recorded, so that with good cause can be claimed, to correctly have derived the regularity of the distances of the planets physically and mathematically.

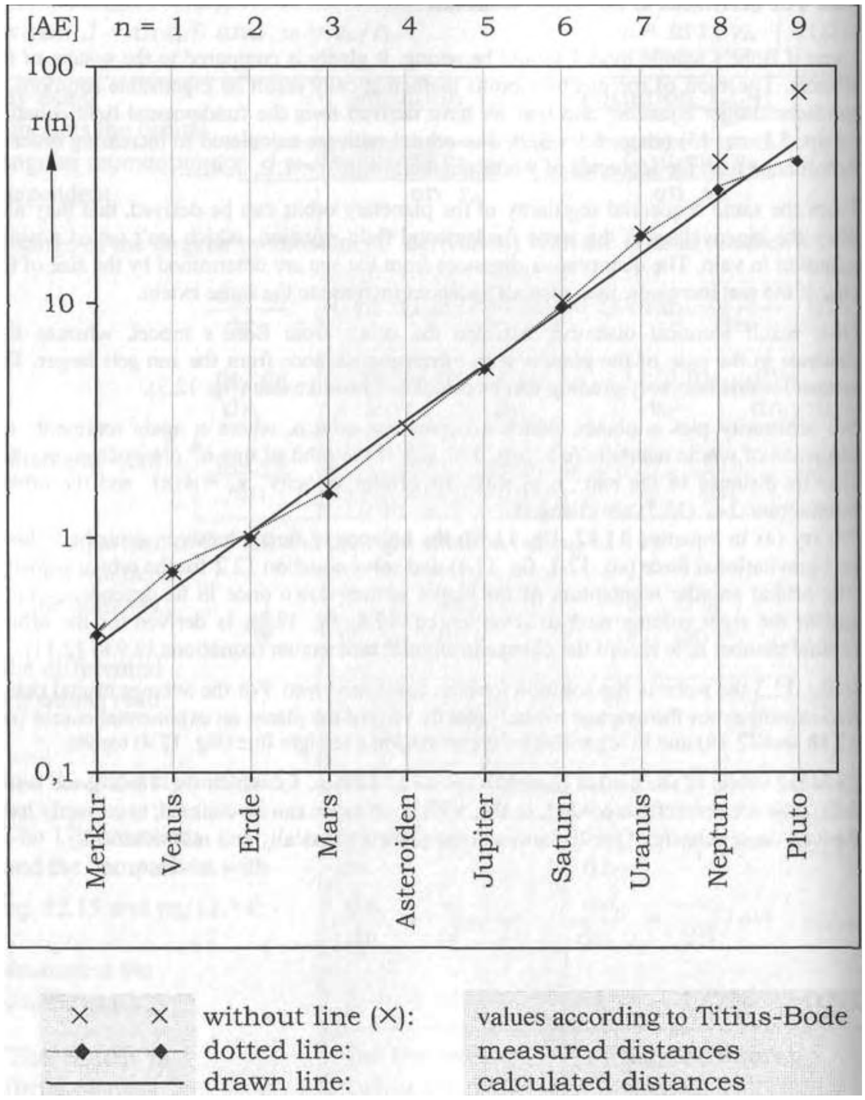

Fig. 12.4: Representation of the distances of the planets

result of the calculation: $$\boxed{r(n) = 0{,}3332 \cdot e^{\left(\frac{n}{1{,}82}\right)} \quad [AE]} \qquad (12.20)$$

12.4 The hollow planet

The numerous accompanying moons of the big planets obey in the same manner as the planets this regularity, so that for the found result every coincidence is excluded. The orbits of the moons of Jupiter (fig. 12.5), of the moons of Saturn (fig. 12.6) and of the moons of Uranus (fig. 12.7) in logarithmic representation lie almost on a straight line. Some orbits certainly are occupied several times, while many an orbit has remained unoccupied. Other orbits again are occupied by a ring of countless chunks of rock, so-called planetoids. Best known representative is the asteroid belt (n = 4) between the orbit of Mars (n = 3) and that of Jupiter (n = 5). The Titius law requested the planet "Aster", but what one found (Piazzi, 1801), at first only was Ceres, the biggest representative of the small planets.

As a second asteroid, as they are called, was found (Olbers 1802) its discoverer proposed the explanation that both, Ceres and Pallas, could have formed in a cosmic catastrophe, which a bigger celestial body had suffered.

Perhaps the sought-for planet Aster actually had imploded after a collision and had been torn apart in countless fragments, of which today still more than half a million pieces are flying around. Most of the fragments however have been collected by the neighbouring planets Jupiter and Mars. Also in the case of the moons of Mars and the outer moons of Jupiter it could concern scrap of Aster. Some presumably fell on the young earth and triggered natural disasters and left deep scars.

It can be imagined that the planet Aster was hollow from the inside and thus was built up similar to the moon. Such a hollow sphere could have formed, if the sun, or in the case of the moon, the earth still was rotating with the cosmic velocity and as a result of the centrifugal force at the perimeter parts of the crust came off as connected sheets and rolled up to a tube.

The fast spinning tube then remodels into a hollow sphere with openings at the poles. In the inside of the sphere, near the equator, protected conditions and the best prerequisites for an undisturbed development, for instance of intelligent living beings, are found by the way. Apart from the advantages the hollow sphere however has as a fundamental disadvantage, that the normally protecting shell in the case of a collision with a bigger celestial body can becomes instable and can implode.

In the case of the moon the pole openings in the meantime are closed and a presumed gas pressure in the inside in addition provides stability. The wall of the spherical shell however is not evenly thick, so that the moon all the time turns the same, namely the heavier, side towards the earth. In view of the seismic measurements of the Apollo-missions, which revealed an extremely small damping of the ground of the moon (chap. 11.9, note <i>), we should look after our moon well, because a comet, which lets the moon implode, would be able to bring about more damage on earth, than if it would hit the earth directly.

Should there exist a hollow and possibly even on the inside habitable planet in our solar system, then surely Saturn should be considered. Its density is smaller than that of water, so that water would be distributed over the inside area of the hollow sphere, if it would be existent. One should examine more detailed the extremely flat pole regions of Saturn for possible openings! Also Uranus and Neptune are possible candidates. This only is thought as an idea.

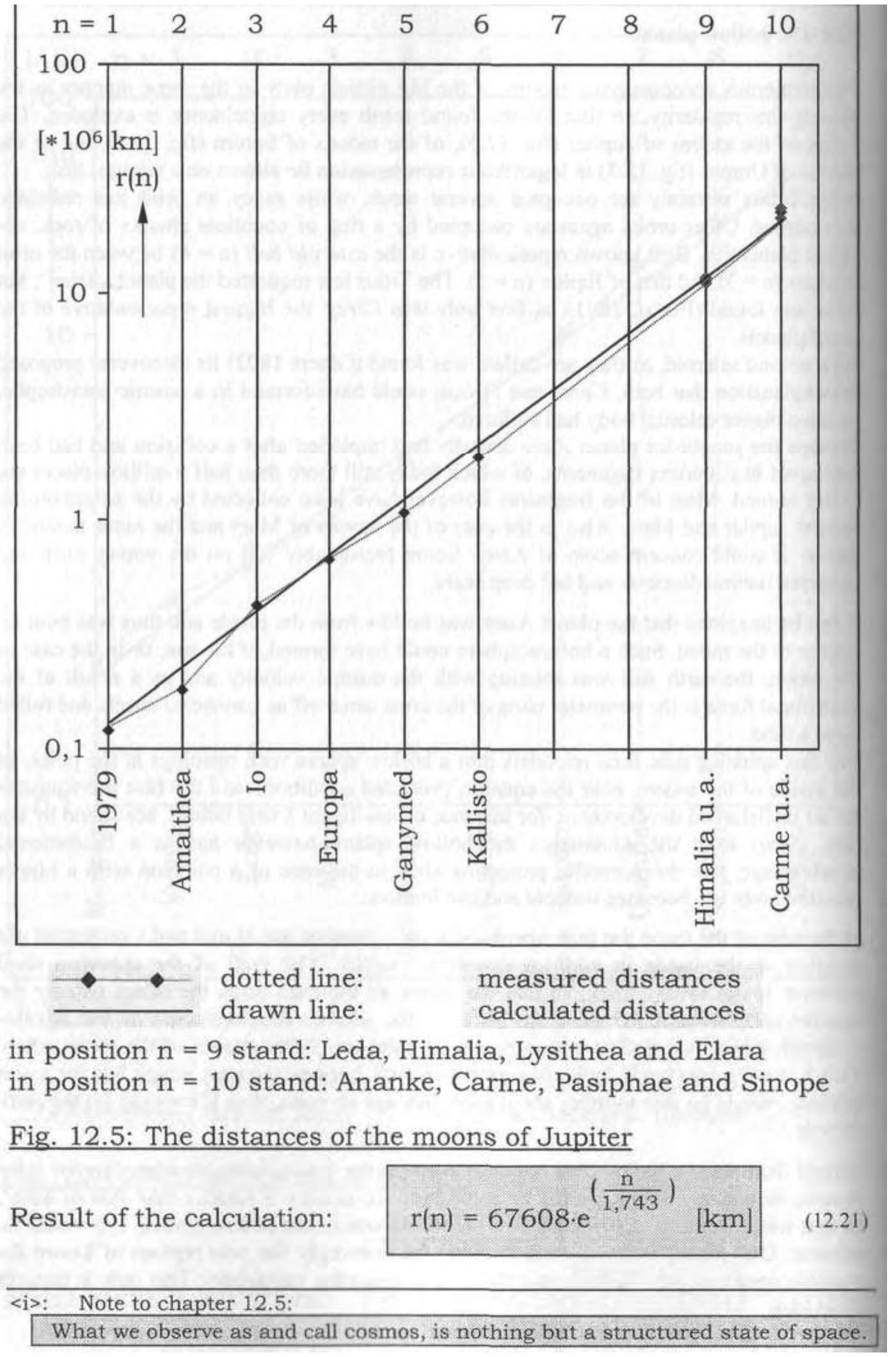

dotted line: measured distances
drawn line: calculated distances

in position n = 9 stand: Leda, Himalia, Lysithea and Elara
in position n = 10 stand: Ananke, Carme, Pasiphae and Sinope

Fig. 12.5: The distances of the moons of Jupiter

Result of the calculation:

$$r(n) = 67608 \cdot e^{\left(\frac{n}{1{,}743}\right)} \quad [\mathrm{km}] \tag{12.21}$$

<i>: Note to chapter 12.5:

What we observe as and call cosmos, is nothing but a structured state of space.

12.5 Concerning the formation of the universe

The widespread concept of an expanding universe bases on the observation of a red shift of the spectral lines of galaxies, which increases with their distance. As the physical explanation for the discovery of Hubble, the Doppler shift for a light source, which is moving away from us, is used. But this concept of an against the attraction of the gravitation taking place expansion is nothing more than a work hypothesis.

The by Christian Doppler in acoustics investigated effect treats the observable shift in frequency, if the source of sound or the receiver is moved with regard to the medium of propagation. But according to today's version there doesn't exist such a medium at all for light, because Einstein has abolished the concept of the aether. According to that the Doppler effect neither can be applied to changes in light frequency.
In the case of the expanding universe, for a decrease of the density and the tracing back to a Big Bang, it therefore should concern a misinterpretation!

Here another effect has to take effect, which one is not yet clarified. Perhaps the changing field relations of the observer environment play a role, after the earth moves away from the sun. But perhaps the galaxies only influence the propagation of their own light, or the light ray on its millions of years lasting way through space slightly loses energy, what is expressed by gradually increasing its wavelength and shifting its spectrum towards the red frequencies.
The hypothesis of the Big Bang moreover contradicts every causality. It is not able to give an answer concerning the origin and the future of the universe and on the question of the origin of the energy and the particles. With that its physical value of explanation goes towards zero.

If we again hold the theory of objectivity against the theory of relativity (part 1), we come to quite other answers. Here apart from the waves also vortices are found. Specially in the case of the spherical vortex a part of the wave power is enclosed in the inside, so that looked at from the outside a from zero differing energetic state results, which even is accessible measuring technically (chapter 6.2).
Wave and vortex are two possible forms of state, so that for the conversion of one state in the other state at first no energy is necessary. The change of state depends on the local field relations.
If we assume that in the beginning the cosmos was free of energy and particles. Then the first vortex was a possible product of chance with an infinite extension. This first spherical vortex, which was contracted under the potential vortex pressure, gave structure to space, gave it energy and field and took care for the rolling up and formation of new vortices[i].
It can be assumed, that even today new particles continually are being formed in the fringe areas of the infinitely extended cosmos, which fly towards us and in doing so contract. They are attracted and at the same time shrunk by the fields of the celestial bodies. They form the source of all matter and energy for our observable universe, which permanently is changing its structure. Because the same oscillation with reversed sign is enclosed in the inside of the spherical vortex, the sum of the energy present in the cosmos is exactly equal to zero. With that the question concerning the causality is superfluous.

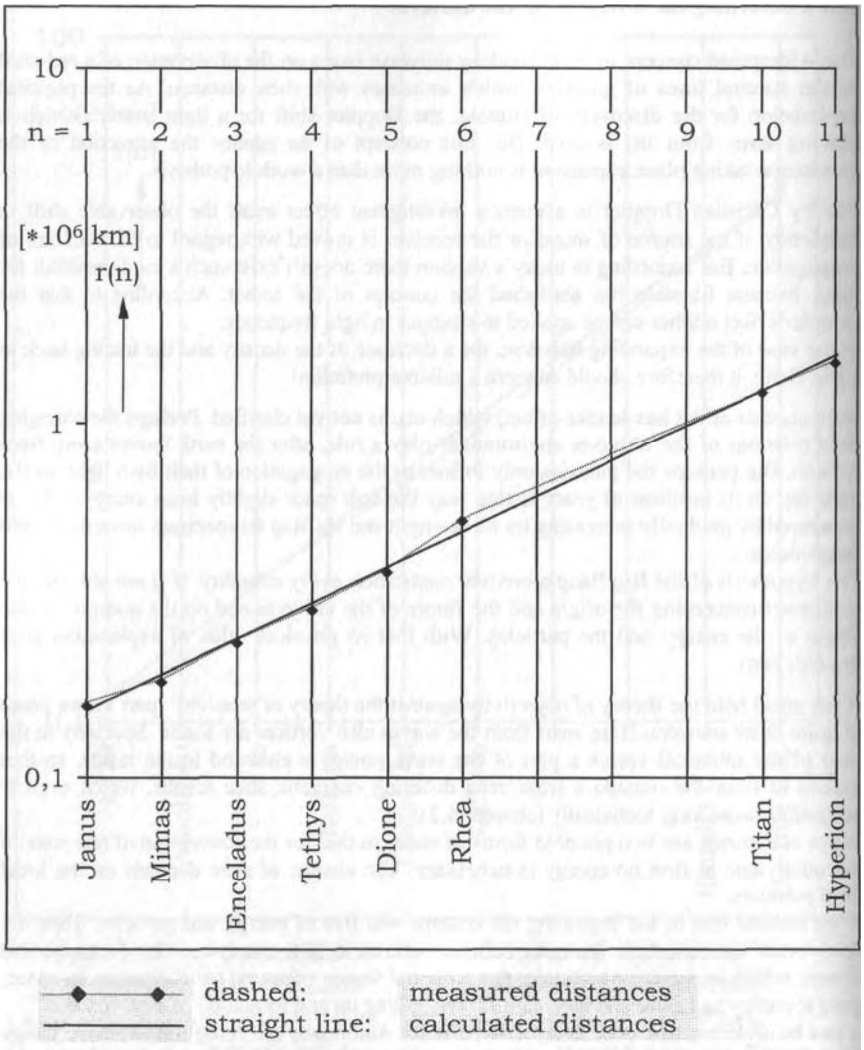

Fig. 12.6: The distances of the moons of Saturn

Note to chapter 12.6 (2nd law of thermodynamics<i>):

Heat cannot completely be transformed into mechanic or electric energy.

<i>: see Electromagnetic environmental compatibility, Part 1, fig. 8.4

The formation of the universe also can be explained causal and completely without a hot Big Bang, even if a supernova surely represents something like a "local Big Bang" for the concerned celestial bodies. According to the observations of the sky expansion and contraction, explosion and implosion occur everywhere in form of oscillations.

On the whole the expanding universe, which once was in thermodynamic equilibrium with matter, should cool down further and further. But stop, the 2nd law of thermodynamics teaches one just the opposite. If the entropy only increases, as the law dictates, then the whole universe should end in a heat death, just as mysteriously and inexplicable, as it should have been formed with the Big Bang. Perhaps something is wrong with the law (fig. 8.4)?

12.6 Counter examples concerning the 2nd law of thermodynamics

Most likely a small experiment convinces us. We heat two spheres, one somewhat less, the other somewhat more. Then we focus the heat radiation of the less hot sphere with help of a parabolic mirror and point it to the hotter one of both. That as a result becomes hotter, whereas the colder is cooled down. The heat thus has flown from the colder to the hotter sphere. Is it allowed to do that?

According to the 2nd law of thermodynamics it of course isn't. There the heat always can only flow from the hot to the cold sphere. But in this primitive experiment it measurable and verifiable flows in the wrong direction. Here the entropy, which is said to always only increase, actually decreases. Here entropy is being destroyed!

Shall we now let the carrying out of the experiment be forbidden under threat of penalty or shall we secretly put the law of entropy to sleep? It can't be denied that this law until now has quite well helped us along, at least for terrestrial processes, at least if one dispenses with the poor inventors, whose inventions offend against the since 100 years tried and tested 2nd law of thermodynamics. They haven't really arrived at the patent office at all, then they already are outside at the door again. Such inventors with their illegal behaviour even today must feel like criminals.

Just what that observed experiment can, each refrigerator and each warmth pump is using as well. And it is not an isolated case: also our sun clearly functions and operates illegally! The surface temperature amounts to only 5800 degrees Kelvin and supplies the atmosphere of the sun with energy. The energy thus flows from the sun to the corona, and that is with values above 1,000,000 degrees Kelvin for some powers of ten hotter!

Only the vortex concept resolves the many question marks without compulsion. In the case of the sun certainly vortices are at work. Here the high temperature in the corona arises as a result of vortices falling apart. We also speak of eddy losses. The transport takes place by heat radiation, exactly as in the experiment with the two spheres.

The possibility exists therefore in vortex processes as well as in technical circle processes that heat could flow from the colder to the hotter sphere. Whether this is an offence against the 2nd law of thermodynamics, is in the end a question of interpretation of the law and up to the opinions of the scholars.

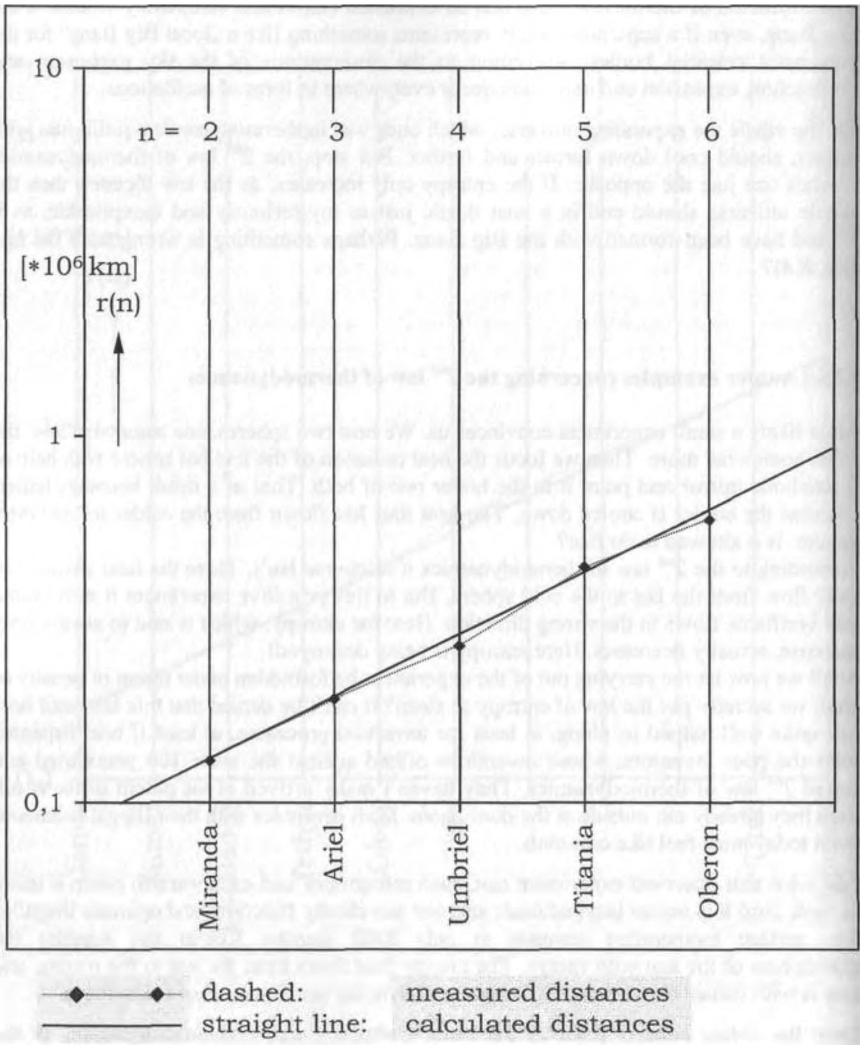

Fig. 12.7: The distances of the moons of Uranus

Note to chapter 12.7 (law of entropy):

The entropy of a closed system never can decrease. It is increased in the case of all irreversible processes. In the case of reversible processes it remains constant.

12.7 Entropy destroying potential vortices

Vortices in addition can amalgamate to balls and to planar vortex systems (4.9 and 4.10). In that case similar consequences, as they can be observed in flow-technical potential vortices in hydrodynamics, can be expected. As a result of the concentration effect (4.1) and because of the conservation of angular momentum an increase of the velocity of rotation of the vortex occurs. Like in the case of the pirouette in figure skating a spontaneous acceleration of its own is observed. In that way the kinetic energy of the system is increased, and that has to come from somewhere.
If we don't supply the contracting vortex with any additional energy for the increase of its rotation of its own, then as a source of energy only the heat energy is left. For this reason every contracting vortex generally converts heat in kinetic energy, it therefore cools down its environment! It moreover destroys entropy and offends against the 2nd law of thermodynamics (fig. 12.8).

It isn't an accident, if in the inside of a tornado it starts to hail. The whirlwinds really furnish visual instruction of the contraction and acceleration of their own of ring-like vortices. If then, even in tropical regions, hail stones are formed, the cooling effect has to come from somewhere, and it can be assumed that the vortex withdraws the heat energy from its environment (calculation in fig. 12.8).

If in specialist books is talked about matter or stars condensing, then vortex physics teaches us that they in reality are contracting and by doing that cooling down. That also is valid for whole galaxies. We owe solely the vortex laws that the continual heating by absorption of radiation is prevented and our sky in the night is dark and doesn't shine as light as day.

The cosmic background radiation, which as a result of vortices lies at almost 3° Kelvin, actually can only be given account for with a contracting of vortices of the Milky Way galaxy. If the cooling down in a compression resp. condensation process has arrived at absolute zero, the vortex becomes stationary, the contraction becomes dependent on the irradiated energy or it wholly comes to a standstill without supply of energy. An example for that is earth's inner core, which at zero Kelvin can't contract further.

From the point of view of causality it is suggested that we galactic and perhaps even cosmic to a large extent are in thermodynamic equilibrium, completely without Big Bang and heat death<i>.

Numerous inventions, as mentioned, are based on the principle of converting environmental heat in useable energy by contraction of vortices. If however the inventors don't know the vortex laws and if they have developed their concept empirically and less physically, then it very often happens that erroneously gravity is made responsible, then is talked of gravitational converters, of the use of a gravitational field energy. But the inventors don't do themselves a favour with that.

<i>: In the seminar it is desired to think about this. Doing so the philosophical faculty may feel as well addressed, as the sandpit of physics, in which completely unsuspecting is played and juggled with Big Bang hypotheses by ignoring all physical regularities and every common sense.

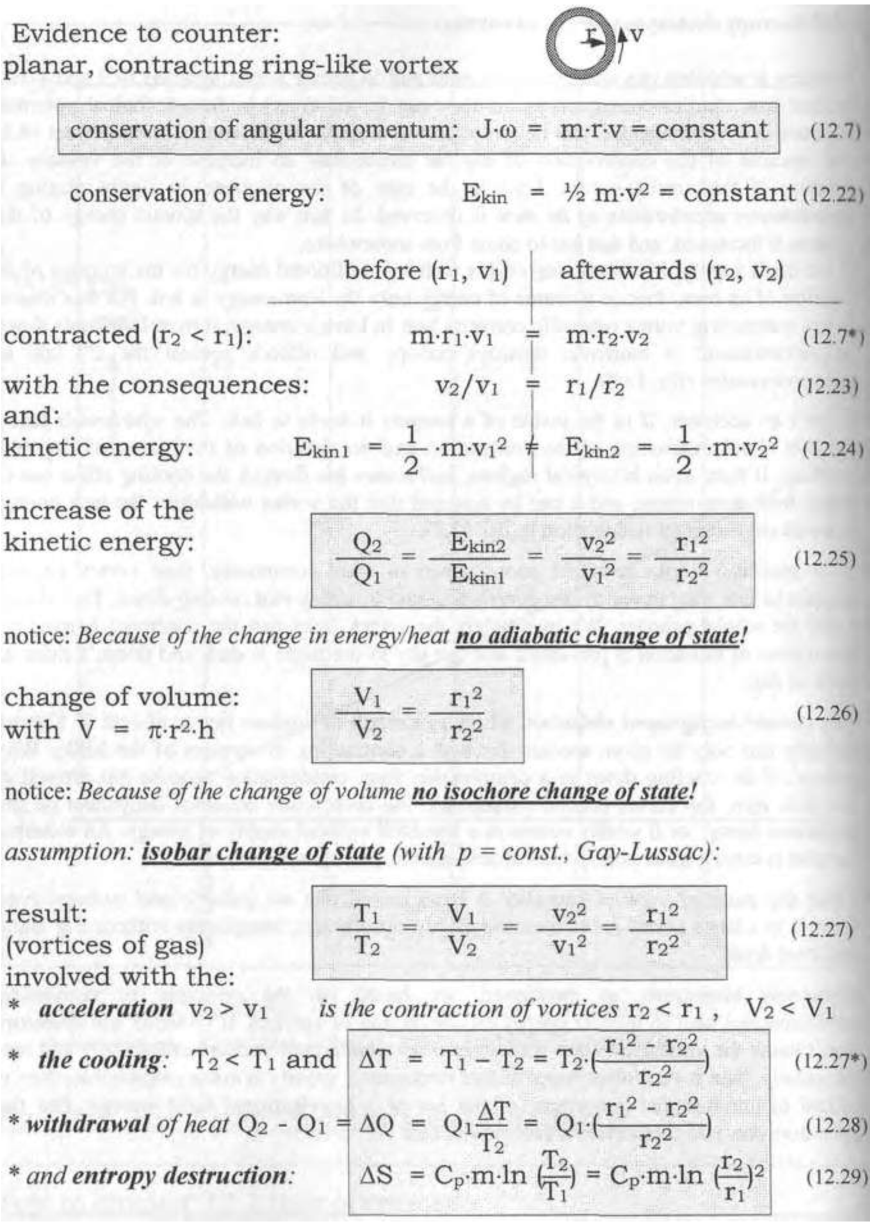

Evidence to counter:
planar, contracting ring-like vortex

conservation of angular momentum: $J \cdot \omega = m \cdot r \cdot v = \text{constant}$ (12.7)

conservation of energy: $E_{kin} = \frac{1}{2} m \cdot v^2 = \text{constant}$ (12.22)

	before (r_1, v_1)	afterwards (r_2, v_2)	
contracted ($r_2 < r_1$):	$m \cdot r_1 \cdot v_1$ =	$m \cdot r_2 \cdot v_2$	(12.7*)
with the consequences:	v_2/v_1 =	r_1/r_2	(12.23)
and: kinetic energy:	$E_{kin1} = \frac{1}{2} \cdot m \cdot v_1^2$ ≠	$E_{kin2} = \frac{1}{2} \cdot m \cdot v_2^2$	(12.24)

increase of the kinetic energy:

$$\frac{Q_2}{Q_1} = \frac{E_{kin2}}{E_{kin1}} = \frac{v_2^2}{v_1^2} = \frac{r_1^2}{r_2^2} \qquad (12.25)$$

notice: *Because of the change in energy/heat* ***no adiabatic change of state!***

change of volume:
with $V = \pi \cdot r^2 \cdot h$

$$\frac{V_1}{V_2} = \frac{r_1^2}{r_2^2} \qquad (12.26)$$

notice: *Because of the change of volume* ***no isochore change of state!***

assumption: ***isobar change of state*** *(with $p = const.$, Gay-Lussac):*

result:
(vortices of gas)

$$\frac{T_1}{T_2} = \frac{V_1}{V_2} = \frac{v_2^2}{v_1^2} = \frac{r_1^2}{r_2^2} \qquad (12.27)$$

involved with the:

* ***acceleration*** $v_2 > v_1$ *is the contraction of vortices* $r_2 < r_1$, $V_2 < V_1$

* ***the cooling****:* $T_2 < T_1$ and $\Delta T = T_1 - T_2 = T_2 \cdot \left(\frac{r_1^2 - r_2^2}{r_2^2}\right)$ (12.27*)

* ***withdrawal*** *of heat* $Q_2 - Q_1 = \Delta Q = Q_1 \frac{\Delta T}{T_2} = Q_1 \cdot \left(\frac{r_1^2 - r_2^2}{r_2^2}\right)$ (12.28)

* *and* ***entropy destruction****:* $\Delta S = C_p \cdot m \cdot \ln\left(\frac{T_2}{T_1}\right) = C_p \cdot m \cdot \ln\left(\frac{r_2}{r_1}\right)^2$ (12.29)

Fig. 12.8: The cooling and entropy destroying effect of contracting potential vortices (calculations concerning chapter 12.7)

13. Recording of space and time

If observations should force us to touch the sacred physical laws, then we should first of all judge our measuring technique critically of all sides and bring it in an usable state. I suggest, we start completely from the beginning with the devices, with which the dimensions of space and time are recorded, with the tape measures for the measurement of length and the chronometers, our clocks.

We must find out, why comets are slowed down if they approach the sun like by the hand of a ghost and in going away again are accelerated, although no forces at all act on the celestial bodies from the outside.

We must find out, why in mines deep under the earth another value for the gravitational constant is being measured as on the surface of the earth<i>. The results hardly can be imagined, if an universal constant should lose its constancy.
Thereby can Newtonian mechanics and the well-known laws be used very successfully from today's point of view, as the derivations in the last chapters have shown. We even could verifiably and mathematically correct calculate the growth of the earth and the solar system with them. The physical laws in the normally used formulation in spite of that seem to be bound to certain limits. Some observations contradict all experience<ii>.

<i>: H. Schuh: Eine Konstante verliert ihre Konstanz; neue Experimente nahren Zweifel an Newtons Gravitationsgesetz, Die Zeit Nr. 40 vom 25.09.97. From it the following quotation (translated):
„It already is suspected for several years, that a fifth force could exist, this suspicion goes back to exact measurements of the gravitational constant G in Australian mines and shafts. Physicists of the University of Queensland in Brisbane had determined, that G for measurements underneath the earth is about one percent larger than the corresponding, since centuries in laboratories determined size. Their proposal for explanation, namely a fifth repelling force, at first met with sharp disapproval. But an at 21 august in the journal Science (Bd. 237/87, P. 881) published work, which bases on measurements in a drilling hole in Michigan, confirms the Australian data".

<ii>: Already the nobel prize winner of physics, Lenard, pointed to the circumstance, that the relativistic representation of the astronomical aberration is incorrect, after no distinguishable aberration could be observed at binary stars, as it had been expected. With that Lenard by the way also has confirmed the existence of an aether.
P. Lenard, Annalen der Physik, Bd. 73, S. 89 (1924)

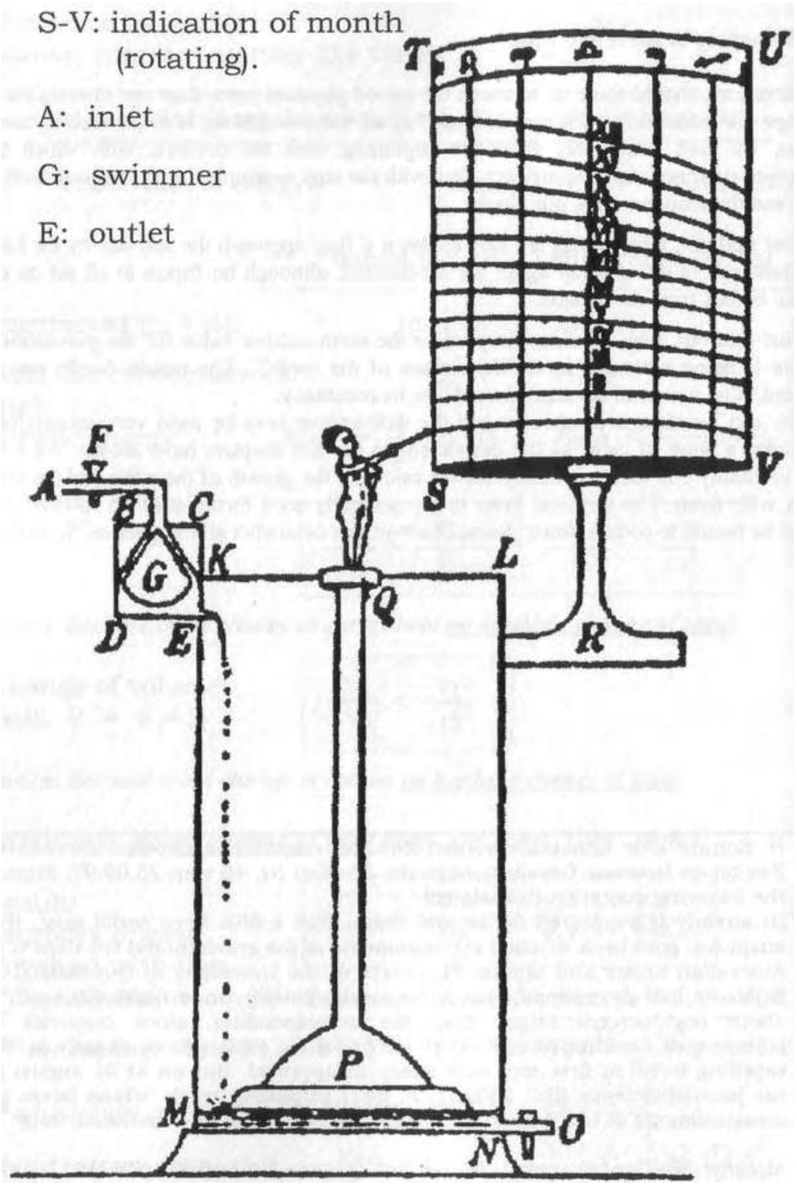
S-V: indication of month (rotating).
A: inlet
G: swimmer
E: outlet
T
U
F
A
B
C
G
K
L
Q
S
V
R
D
E
P
M
N
O

Fig. 13.1: The water meter of the Ktesibios (approx. 250 BC)
(with regulation of the water-level)

13.1 The measuring technical debacle

We are standing for a measuring technical debacle, because we have fixed our calendar to the rotation of its own of the earth. We call a 360° turn a day, divide it in 24 hours of 60 minutes each and every minute in 60 seconds. With that we determine the duration of a second.

A clock according to this definition only then is exact, if it follows the changes of the earth to the same extent. Obviously this in particular is the case for atomic clocks. An objectively seen precise going clock however would land at the waste disposal site as completely unusable.

Alter all nobody wants to know, how late it really is. Everyone only wants to find confirmed his subjective feeling, and our experience of time simply is directed after the course of the sun, thus after the rotation of the earth!

For the moment and as an approximation also for the life time of a person, this determination of time may be sufficiently exact, but seen over larger periods of time, one doesn't get around a conversion. Actually the 200 million years since the beginning of the Creation on the mentioned primeval hill are considerably less long ago. The earth and the whole solar system are very much younger, than was assumed until now!

If we as an example again take the cited research group of the University of Arizona (chap. 11.10<i>), which by means of an analysis of sediment formations have found out, in accordance with our calculation, that 900 million years ago a day only had 18 hours<i>. Such traces in geological deposit layers can be very informative, if they are analysed correctly, since here the measurement technician himself is not subject to the process. He stands outside and hence can exactly measure the time difference.

If a geologist of that time had looked at his watch during the formation of the layers, then a day obviously would have had 24 hours, and he wouldn't have been able to understand the whole excitement at all. The accusation, he would suffer from chronic blindness caused by his job, he of course would repel resolutely. Can you now imagine, from which disease our science of today suffers?

Our chronometers are nothing but the improved model of a sundial<ii>.

We live in the dimensions of space and time, but we quite obviously have the biggest possible difficulties with the dimensioning of both. Possibly we already aren't capable of that at all by principle. Most suitably we clarify the situation by means of examples concerning the two problem spheres.

Let us at first stay at the dimension of time. It may have become clear, how problematic the chosen determination of the time scale in seconds is and which contradictions can result if larger periods of time are considered. More and more often science fiction authors romp about on the playground of time, fantasize about some time travel, or they occasionally make jokes about it<iii>.

<i>: 900 million years ago a day had 18 hours, Washington (dpa) 1997.

<ii>: There also exist models, of which is asserted, they go according to the moon.

<iii>: A passenger, who got on the 10 o'clock bus, passes by the church steeple clock, which only reads 5 to 10 and curses: Damn, I took the bus in the wrong direction!

Centrifugal force $\boxed{F_{1,2} = (m/R)\cdot(v_E \pm v_f)^2 = m\cdot R\cdot(\omega_E \pm \omega_f)^2}$ (13.1)

with: $R = 6378$ [km] (radius of the earth),

$v_E = \omega_E\cdot R = 0{,}465$ [km/s] (rotational speed of the earth), (11.20)

$v_f = \omega_f\cdot R = \pi\cdot R/t_o$ (speed of the plane) (13.2)

and: t_o [s] (duration of journey)

energy of the moving steering quanta:

$$E_{1,2} = \int_{O}^{R} F_{1,2}\,dR = m\int_{O}^{R} R\cdot(\omega_E \pm \omega_f)^2\,dR = \qquad (13.3)$$

$$\boxed{E_{1,2} = \frac{1}{2}\cdot m\cdot R^2\cdot(\omega_E \pm \omega_f)^2 = \frac{1}{2}\cdot m\cdot(v_E \pm v_f)^2} \qquad (13.4)$$

energy balance (generally of a quantum of radiation):

$$\boxed{E_{1,2} + E_q = E_{ges}} \text{ resp. } \boxed{E_{ges} - E_q = E_{1,2} = \Delta E} \qquad (13.5)$$

in general: $\boxed{E = h\cdot f = m\cdot c^2}$ resp. $\boxed{m = h\cdot f/c^2}$ (13.6)

(Gl. 13.4 mit 13.6):

and specifically: $\boxed{\Delta E = h\cdot\Delta f} = \boxed{E_{1,2} = h\cdot f\cdot(v_E \pm v_f)^2/(2c^2)}$ (13.7)

$\Delta f/f = \Delta t_{1,2}/t_o$ resp. $\Delta t_{1,2} = t_o\cdot\Delta f/f$ (13.8)

(eq. 13.7 inserted into eq. 13.8)

resulting in: $\boxed{\Delta t_{1,2} = t_o\cdot(v_E \pm v_f)^2/(2\cdot c^2)}$ (13.9)

time of travel around half the earth: $t_o = \pi\cdot R/v_f$ (13.2)

results in a difference in going of: $\Delta t = \Delta t_1 - \Delta t_2$ (13.10)

$$\Delta t = [(v_E + v_f)^2 - (v_E - v_f)^2]\cdot\pi\cdot R/(v_f\cdot 2\cdot c^2) \qquad (13.11)$$

result calculated: $\boxed{\Delta t = 2\cdot v_E\cdot\pi\cdot R/c^2 = \underline{207}\ [ns]}$ (13.12)

measured by Hafele and Keating<i>: 214 [ns]! (13.13)

Fig. 13.2: The difference in going of two atomic clocks
(caesium resonant clocks)

<i>: Hafele-Keating-Experiment, Oktober 1971; s. a. W. Bauer: Klassische Physik, Graphia Druck, Salzburg (1975), Eigenverlag

13.2 The clock paradox

The theme „time dilatation" in connexion with the particle decay already has been treated (part 1, chapter 6.20). It has been shown, that a fast moving and with that length contracted particle to exactly the same extent (Lorentz's square root) becomes more stable and longer-living. If relativists pack an atomic clock based on radioactive decay in a plane and detect a difference in going between the one, which has been flown around and a second identically constructed clock, which has stayed at the ground, then they have with that detected experimentally a very small length contraction, which really occurred, and by no means a time dilatation, as they claim.

Now we in addition owe Einstein, that the aether has been abolished and from that follows, that it can't play a role in which direction the plane flies. If therefore both clocks are taken along each in a separate plane, one plane flying to the west and the other to the cast, both planes meeting again for the first time on the other side of the globe, then according to Einstein's theory it shouldn't be possible to determine a difference in going, if both planes constructed identically were on the way with the same velocity. But this is not the case!

Actually a difference in going is measured, which however can't be calculated with the theory of relativity, yes, which is completely incompatible with this theory and clearly brings anyone to the eye, that the effect actually can't have to do anything with a time dilatation, that the moving clocks merely go wrong and we have to ask us, why<i>.

These experiments were carried out with atomic clocks, which are constructed as caesium resonators and work with an exactness of one second in 300000 years. As a resonator serves a quartz crystal, which is controlled by an ion current of caesium atoms, which have lost their outermost enveloping electron. The system is fed back, because the oscillating quartz controlled by the caesium ions again adjusts the caesium vapour by radio wave and finally its own atomic controlling current (fig. 13.7).

The reason for the measured difference in going is seen in the field and here specially in the different gravitational field. The centrifugal force directed opposite to the gravitational force at least is not the same, because for a westward flight along the equator the speed of the plane v_f should be subtracted from the velocity of rotation of the earth v_E, whereas in eastward direction it should be added (eq. 13.1).

For the steering quanta supplied by the caesium resonator now the energy balance is put up (13.5 with 13.4) and the change of the reference frequency is calculated (13.7). With the change in frequency is connected directly a change of the at the two clocks readable times $t_{1,2}$ (13.8). For a journey around half the earth, where one clock is flown westwards and the other one eastwards, the difference in going should, according to the calculation, amount to 207 ns. Interesting of the result (13.11) undoubtedly is, that the velocity of the planes doesn"t play a role. It is cancelled out.

October 1971 caesium atomic clocks were sent around the world in scheduled planes in the Hafele-Keating experiment. To be able to estimate the inaccuracy in going of the clocks and with that the measurement error, four clocks were used. Between the westward journey (273 ±7 ns) and the eastward journey (59 ±10 ns) a difference in time of 214 nanoseconds was determined. This under strict scientific conditions determined result once more proves the correctness of the theory of objectivity by confirmation of the calculated value. That however is not valid for the special theory of relativity, because that doesn't appear in the calculation at all!

Who now believes, we would have less problems with the dimension of space, I must disappoint. The determination of the linear measures equally ends in a fiasco.

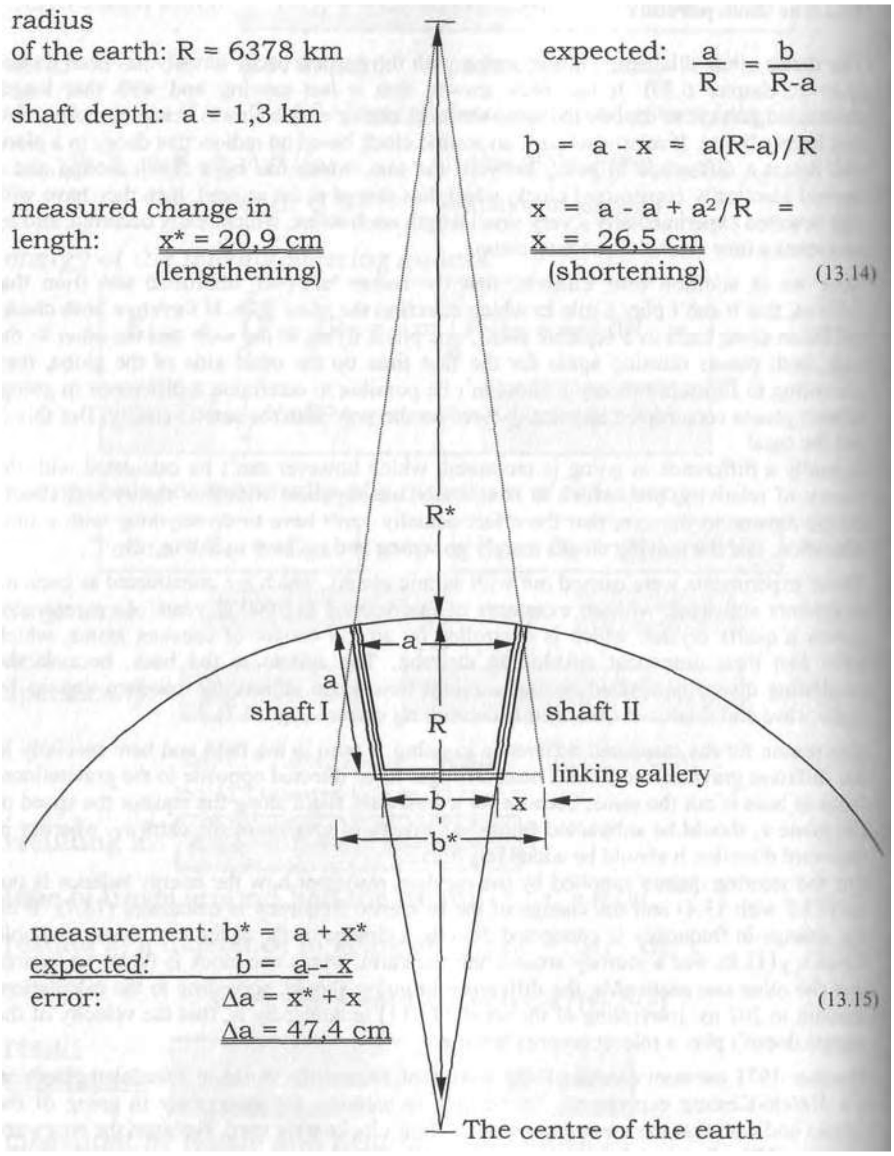
radius
of the earth: R = 6378 km
shaft depth: a = 1,3 km
measured change in
length: x* = 20,9 cm
(lengthening)
expected: $\frac{a}{R} = \frac{b}{R-a}$
b = a - x = a(R-a)/R
$x = a - a + a^2/R =$
x = 26,5 cm
(shortening)
(13.14)
R*
a
a
shaft I
R
shaft II
linking gallery
b
x
b*
measurement: b* = a + x*
expected: b = a – x
error: Δa = x* + x
Δa = 47,4 cm
(13.15)
The centre of the earth

Fig. 13.3: Expectation and measurement in the Tamarack mines

(The curvature of the earth and the length of the shaft a are drawn strongly exaggerated for clarification)

13.3 The Tamarack mines experiment

As long as the „foot measure" depended on the shoe size and the „cubit" on the forearm of the tailor, the world still was OK. The sciences however request a reproducible quantity for comparison, and that can be fetched at the Bureau International des Poids et Mesures in Sevres near Paris. The original meter is a Platinum alloy. Because the length of the metal always depends on temperature, it is stored at a constant kept temperature of 0° Celsius. Now there in addition still exists a field dependency, an electrostriction resp. magnetostriction. And how is the measurement bar behaving, if the earth grows and the density increases? Is it then shrinking just like other objects in its environment? At this point already chaos is proliferating.

The newest definition of the length measure meter acts as a blow for liberty and thus marks the abyss, at which we are standing: The length is determined by means of a measurement of transmission time of an electromagnetic wave, e.g. of a light signal. It is said that with this determination a higher reproducibility should be obtained. Actually a photo optical facility to measure length is as exact as the built-in facility to measure time, and there we use it again, our sundial. In addition a constancy of the speed of light is taken as a prerequisite, and that is given in meters per second. From a change of the speed of light for forinstance 10% a change in length for 10% as well would result<i>. Because we see this process with the help of our eyes as well with the speed of light, we never can see the change. We neither can technically measure it, because all gauges we construct are built up corresponding to our sensory impression. We ourselves have shovelled the hole, in which we fall.

Only if we succeed in taking a neutral standpoint outside of the events, the true relations will become visible to us. For the field dependency of the space measures a very clear experiment has been carried out, of which I now want to report<ii>.

1901 the French government was offered the possibility to carry out an experiment in the shut down Tamarack mines near Calumet (Michigan) with the goal to determine the diameter of the earth more exact. For that the geophysicists let down two plumb-lines of 27.2 kg each in two perpendicular winding shafts, which were at a distance of 1.3 km from each other. The plumb-lines were tied to hardly expandable piano wires of as well 1.3 km length. It now was expected, since the plumb-lines hung in direction of the centre of the earth, that in a linking gallery between the two shafts a length of (1.3 - x) km should be measurable. From the shortening x one wanted to infer the diameter of the earth (fig. 13.3). But it came completely different.
Instead of a shortening a lengthening for x* = 20.9 cm was measured in the gallery! The point of intersection of the lines through the two shafts had to be not in the inside of the earth, but in space! Immediately „hollow earthers" appear, who claim we would live on the inside of a hollow world<ii>. Perhaps one should shoot them to the moon, because from there the earth without doubt is seen as a sphere. Obviously we aren't dealing with a surprise of a fair, but with a fundamental measuring technical problem.

<i>: for that see also part I, chapter 6.11 and 6.12
<ii>: http:/www.t0.or.at/subrise/hollow.htm

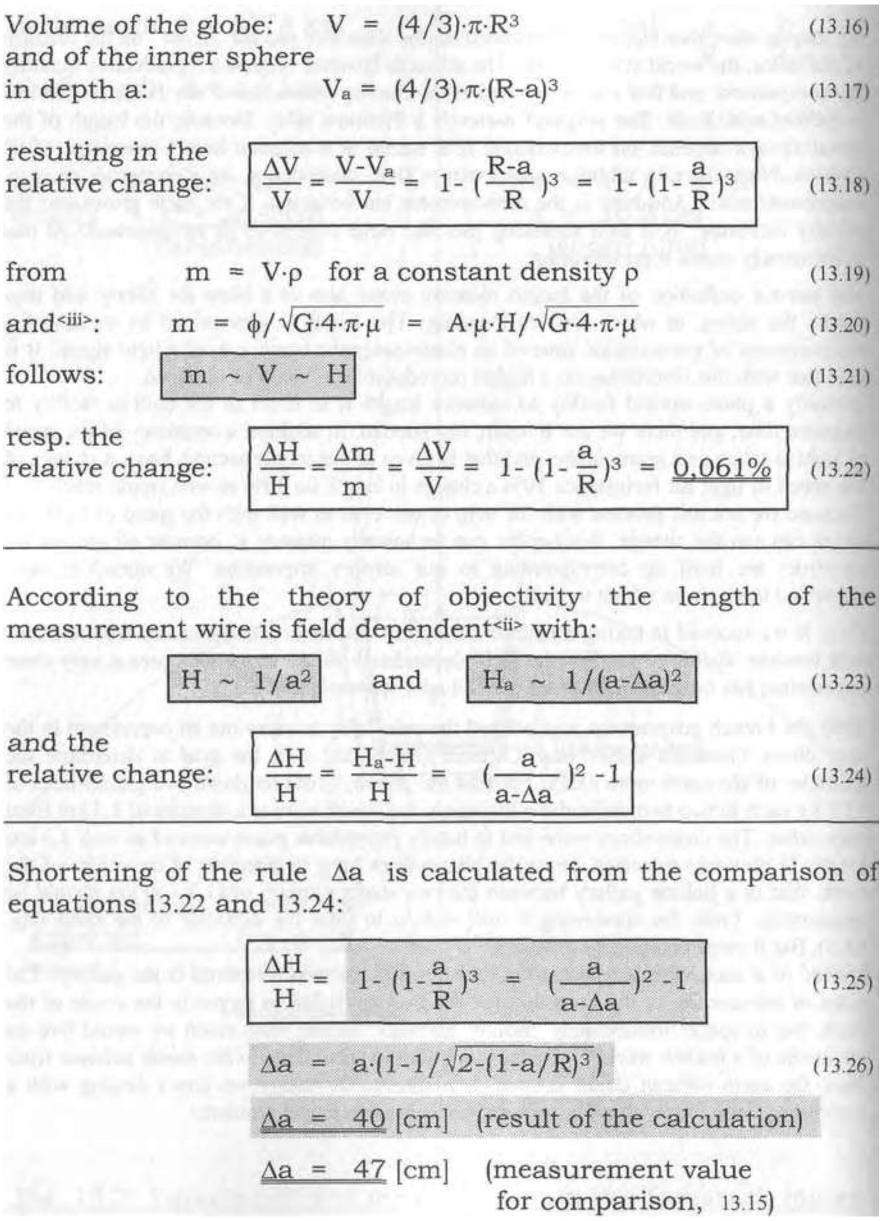

Volume of the globe: $V = (4/3)\cdot\pi\cdot R^3$ (13.16)

and of the inner sphere

in depth a: $V_a = (4/3)\cdot\pi\cdot(R-a)^3$ (13.17)

resulting in the relative change:

$$\boxed{\frac{\Delta V}{V} = \frac{V-V_a}{V} = 1-\left(\frac{R-a}{R}\right)^3 = 1-\left(1-\frac{a}{R}\right)^3} \quad (13.18)$$

from $m = V\cdot\rho$ for a constant density ρ (13.19)

and[iii]: $m = \phi/\sqrt{G\cdot4\cdot\pi\cdot\mu} = A\cdot\mu\cdot H/\sqrt{G\cdot4\cdot\pi\cdot\mu}$ (13.20)

follows: $\boxed{m \sim V \sim H}$ (13.21)

resp. the relative change:

$$\boxed{\frac{\Delta H}{H} = \frac{\Delta m}{m} = \frac{\Delta V}{V} = 1-\left(1-\frac{a}{R}\right)^3 = \underline{0{,}061\%}} \quad (13.22)$$

According to the theory of objectivity the length of the measurement wire is field dependent[ii] with:

$$\boxed{H \sim 1/a^2} \quad \text{and} \quad \boxed{H_a \sim 1/(a-\Delta a)^2} \quad (13.23)$$

and the relative change:

$$\boxed{\frac{\Delta H}{H} = \frac{H_a-H}{H} = \left(\frac{a}{a-\Delta a}\right)^2-1} \quad (13.24)$$

Shortening of the rule Δa is calculated from the comparison of equations 13.22 and 13.24:

$$\boxed{\frac{\Delta H}{H} = 1-\left(1-\frac{a}{R}\right)^3 = \left(\frac{a}{a-\Delta a}\right)^2-1} \quad (13.25)$$

$$\Delta a = a\cdot\left(1-1/\sqrt{2-(1-a/R)^3}\right) \quad (13.26)$$

$\underline{\Delta a = 40}$ [cm] (result of the calculation)

$\underline{\Delta a = 47}$ [cm] (measurement value for comparison, 13.15)

Fig. 13.4: Calculational verification of the measured shortening

13.4 Field dependent linear measure

If a measurement result delivers just the opposite, as was expected by the experimentators, then the layman is amazed and the expert is surprised, at least at first. But then, out of a feeling of panic, the whole view of life could collapse, as many as possible scientists and renowned professors are being informed and integrated, if they want it (Prof. Mc.Nair) or not (Prof. Hallock, Columbia University) and eventually the matter is buried third class and a guise of silence is spread over it. The censorship of the scientific making of opinion in advance doesn't permit publications, which are not in accord with our view of life, out of the animal survival instinct<i>.

A science, which deserves this name, should look different. There it must be permitted, to ask questions and to publicly discuss about it.

I proceed from the assumption, that the earth is a sphere, which we inhabit from the outside; I have no doubts about that. With this as a prerequisite there is only one possible answer to the Tamarack mines experiment: The 1.3 kilometres long measurement wire, which in the gallery in a depth of 1.3 km had to jut out for 26.5 cm, instead is too short for 20.9 cm, from which immediately follows that it, howsoever, is shrunk for 47.4 cm in the depth.

At first of course the experiment was checked for possible measurement errors. The shafts were covered to exclude any draught. The measurement path was optically checked, but the result remained unchanged. But if the cause for the unexpected result doesn't lie in the experiment, then theoretical physics is addressed, after all it is still the experiment which shows us the physical reality and not some theoretical model concept.

In the 1st part of the book already the derivation of a useful explanation is found: The speed of light and with that also the linear measure depends on the field<ii>. The measurement wire accordingly gets shorter, if it is exposed to a larger field strength (eq. 13.23). We can verify the measured shortening calculative (fig. 13.4).

For that we at first determine the change of the field strength, as it is to be expected in a depth of 1.3 km. We here are dealing with the closed H field lines, which are responsible for the gravitation. In a past derivation it has been shown, that a gravitating mass can be converted into a magnetic field<iii>. Between a mass m and a field strength H hence exists a proportionality (13.21), in the same way as between the same mass and its volume, if a constant density is present (13.19).

The result accordingly is a relative decrease of the volume of the earth and the corresponding mass being under the measurement place, as well as a relative decrease of the radial component of the field strength, but a corresponding relative increase of the tangential component of the field for 0.061% (13.22 with 13.18).

In the gallery the measurement wire however is spread out in the direction of the tangential component of the field lines, and that shortens the measurement wire as a result of the field dependency of the linear measures<ii> (13.23). From the above increase in field the calculational shortening of the rule for 40 cm results, which compared to the measurement result also should be rated as a confirmation of the theory of objectivity which was taken as a basis<ii>!

<i>: Example: K. Meyl: Potentialwirbel, Bd. 1, reference entry /5/

<ii>: for that see also part 1, chapter 6.6, equation 65 and chapter 6.10

<iii>: K. Meyl, Potentialwirbel Band 2 (1992), page 27, equation 20

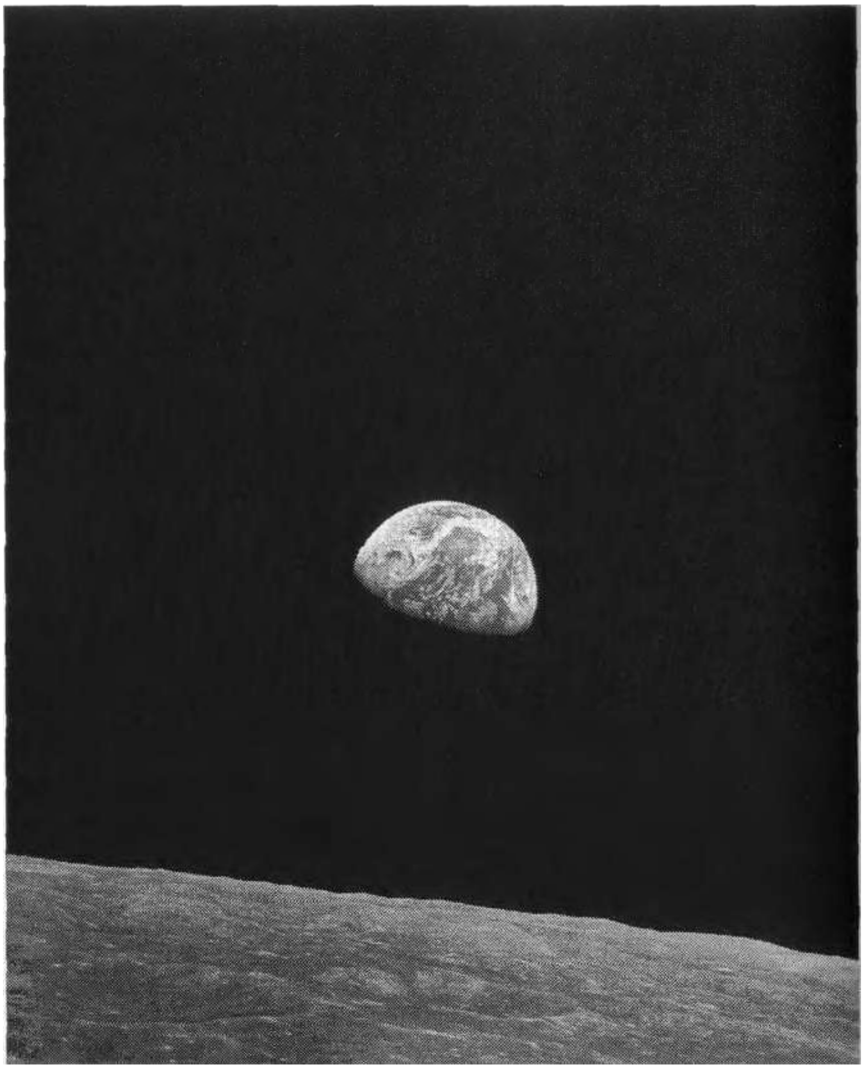

Fig. 13.5: Rise of the earth over the horizon of the moon""

<i>: Mitton, S. (Herausg.): Cambridge Enzyklopadie der Astronomie, Orbis Verlag (1989), note: No star is seen!

<ii>: see also the references in part 1, chapter 6.9

13.5 Experiences from space travel

The good correspondence of the calculated shortening of the measurement wire and the until now not understood measurement in the Tamarack mines shows both qualitatively, and quantitatively the correctness and useful applicability of the theory of objectivity of the field dependency of the linear measures. Our measurement laboratories normally are situated on the earth's surface and there everywhere are found approximately identical field relations. But if we leave the usual measurement environment and move the laboratory for instance in the sky, then we experience a complete mystery. Here however prevail the reversed conditions as in the mines experiment, in which in the inside of the earth. for an increase of the field strength, a length shortening was measured. In the sky the field strengths decrease and the linear measures correspondingly increase.

This experience astronaut Roosa made in the Apollo 14 mission. While he alone in his capsule orbited the moon, he depicted mission control, he could see the lunar module and observe his two colleagues at their work on the moon. Nobody wanted to believe the astronaut, since he was flying in a height of 180 km!

Commander Armstrong (Apollo 11) at the first landing on the moon indicated, the target crater Mackensen, 4.6 km in diameter measured from the earth, just has the size of a soccer field! Astronaut Scott (Apollo 15) called Mount Hardley, which is said to be 4.8 km high, a practice hill for skiing. Perhaps they somewhat have exaggerated, but a true core in the statements always is present!

Actually the gravitational field of our satellite is very much smaller than that of the earth. On the surface of the moon there is only one sixth of the gravitational pull of the earth. If we, to be able to compare, stick to the details of size, as they are measured by our laboratory on earth, then the astronauts on the way to the moon together with the lunar module and their rover had grown for a factor $\sqrt{6}$,then the first footprint is 2.5 times as large as on earth, then the astronauts were moving like giants in the scenery of a model of the railroad (eq. 13.23 and note<ii> at fig. 13.4).

On the moon there exists almost no atmosphere, for which reason the astronauts had imagined a wonderful view of the star-spangled sky, at least before they started. After the landing they were bitterly disappointed. The sky was black and not one single star could be seen! They have brought many photographs, but nowhere stars have been photographed, they apparently have moved outside the range of vision (fig. 13.5 and 13.6).

Many will still remember that the first pictures, which the space telescope Hubble supplied 1990, were completely blurred. The problem obviously was, that the mirrors had been adjusted on earth and not in space. Only after the optics had been given glasses in 1994, sharp pictures could be radioed to earth. Somehow the distance to the stars had changed. The telescope had become short-sighted, resp. the distance to the star-spangled sky appeared to be gotten larger. We already know why. If we remove us from the gravitational field of the earth, the field strength decreases and the observable distances increase! The highly sensitive telescope already sufficed the 5% deviation, with which should have been reckoned for the near earth orbit, to be fatal.

One should have familiarized the astronauts before with the laws of physics. Then this disappointment would have been spared to them, and in the case of the Hubble telescope the NASA and the european ESA could have saved a lot of money for the sake of the tax-paying population<ii>.

Warum sieht man am Himmel keine Sterne?

Fig. 13.6 a: Illustrierte Wissenschaft Nr. 11 (1997) Page 62

1st example: landing on the moon gravitational ...

$$\frac{\text{...pull of the earth}}{\text{...pull of the moon}} = \frac{g_E}{g_m} = \frac{M / R^2}{m_m / R_m^2} = \frac{M}{m_m} \cdot \frac{R_m^2}{R^2} = 6{,}0375 \qquad (13.27)$$

in space everywhere on the surface of a sphere is valid: $A = 4\pi R^2$

eq. 13.20 (fig. 13.4<iii>): $m \sim \phi = A \cdot B = 4\pi R^2 \cdot \mu H \sim R^2 \cdot H$

as well as eq. 13.23 (fig. 13.4<ii>) for the field dependency: $H \sim 1/l^2$

and further:

$$6{,}0375 = \frac{g_E}{g_m} = \frac{H_E}{H_m} = \frac{l_m^2}{l_E^2} \qquad (13.28)$$

resulting in the length dilatation (expansion) on the moon:

$$\boxed{l_{m\,(\text{Mond})} : l_{E\,(\text{Erde})} = \sqrt{6{,}0375} = \underline{2{,}457}} \qquad (13.29)$$

2nd example: communications satellite in a geostationary orbit (at h = 36000 km above the equator, R = 6378 km).

$$\boxed{\frac{l_h}{l_E} = \sqrt{\frac{H_E}{H_h}} = \frac{R+h}{R} = \underline{6{,}64}} \qquad (13.30)$$

Fig. 13.6 b: Examples of calculation for length dilatation<i>

<i>: If the theoretical value of 6.64 could be quantitatively confirmed by observations from off the earth with a telescope at a corresponding magnification, then with that would have been proved, that in the case of the spherical aberration it actually concerns the calculated influence of the field.

<ii>: U. Seiler-Spielmann: Das Marchen vom toten Mond, Zeiten Schrift 5/94,S.39

13.6 Spherical aberration

It is true that the problem of the changed length relations is known to the experts under the term of a "spherical aberration". But with that it is neither qualitatively nor quantitatively understood. Only the theory of objectivity soundly gives reasons for, why the astronaut Roosa has seen his colleagues almost 3 times as large, why weather satellites in a height of 1500 km are approx. 25% larger and why communications satellites in a 36000 km high geostationary orbit even increase to the 6.64 fold of their original size[i]. It also explains, why the neutral point between earth and moon, at which the attraction of masses of both celestial bodies mutually cancel, wasn't reached at the point where it had been expected by the moon rockets[ii].

We, the inhabitants of the earth, are adapted completely to the conditions on the earth's surface. If we find our way well in the dimensions of space and time, as we observe them, then that must not be valid by all means for science, because that has made it its business to find out the secrets of nature.

If it wants to deserve the name science, then it on the one hand has to consider, that we, the organic materials, as well as all inorganic materials are assembled from the same atoms and molecules and with that are exposed to the same length relations. If changes in length between day and night (as a result of the gravitational field of the sun (see chap. 6.7), between summer and winter or as a result of changes in field occur, then we aren't able to register this at all. That even today the „foot" is used as a measure, for instance in the air traffic, shows only too clearly, how man raises itself to the measure of all things. Science asks for modesty!

On the other hand it should be paid attention to the fact, that man eyes everything with speed of light with the optics of his eyes, and that speed by no means has to be constant. Solely the definition of the speed of light c as a linear measure per unit of time points to the direct proportionality between c and a length 1 (see chapter 6.3):

$$c \sim 1 \qquad (13.31)$$

If a rule has proven to be unusable for measuring a distance, then we'll experience the same disaster, if we measure optically, i.e. with the speed of light. Obviously both, the length 1 and the speed of light c depend in the same manner on the respective local field strength. On the one hand both measurement techniques lead to the same result, but on the other hand what can't be measured with one method, neither can be measured with the other.

To prove the constancy, it is normal to measure the speed of light optically. But since there exists a proportionality between measurement variable and measurement path (53), the unknown variable is being measured with itself. This measurement faulty by principle in all cases delivers a constant value. In contrast to the textbook opinion of today by no means a constancy of the speed of light can be assumed. In the case of the in a vacuum measurable 300,000 km/s it concerns a capital measurement error, at best a constant of measurement, but never ever a constant of nature!

With the postulate and the misinterpretation of a constancy of the speed of light as a universal constant of nature Einstein already let several generations of physicists run into the same dead end, in which they today are stuck altogether. It surely is no accident, that the big time of discoveries abrupt came to an end with Einstein.

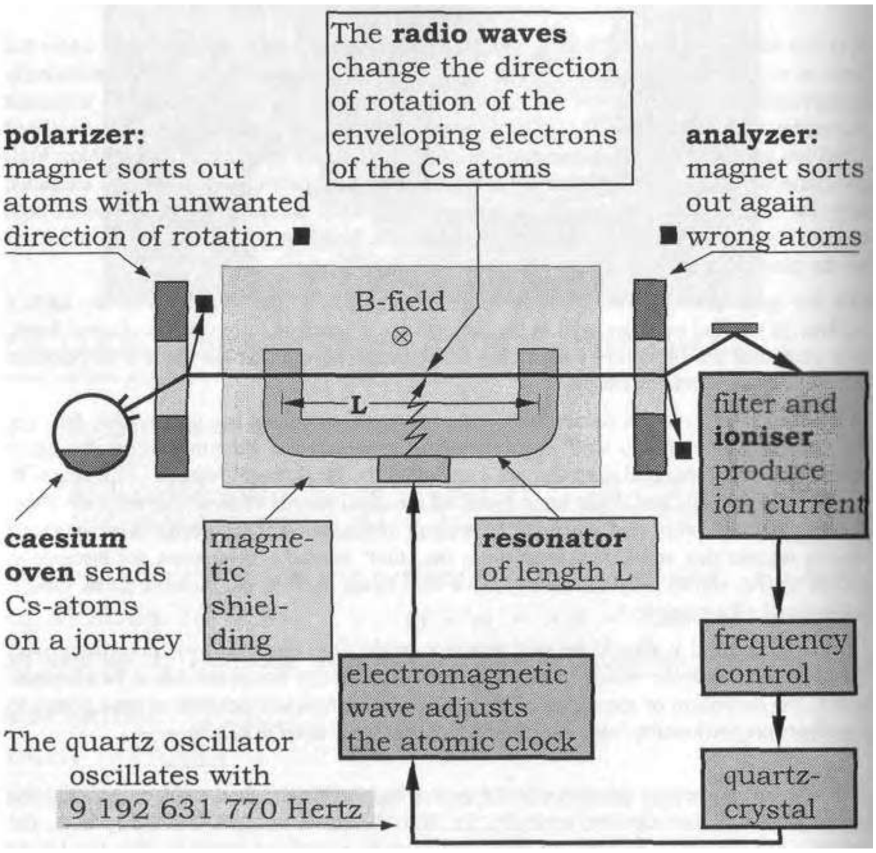
The **radio waves** change the direction of rotation of the enveloping electrons of the Cs atoms
polarizer: magnet sorts out atoms with unwanted direction of rotation ■
analyzer: magnet sorts out again ■ wrong atoms
B-field
⊗
L
filter and **ioniser** produce ion current
caesium oven sends Cs-atoms on a journey
magne-tic shiel-ding
resonator of length L
frequency control
electromagnetic wave adjusts the atomic clock
The quartz oscillator oscillates with
9 192 631 770 Hertz
quartz-crystal

Schematic representation concerning principle of functioning
L = length of resonator determining the exactness
The B-field is a weak magnetic field, which eliminates the influence of magnetic stray fields.
The arrangement in addition is situated in a vacuum tank.

Fig. 13.7: Set-up of a caesium atomic clock

13.7 Irony of the measuring technique

Let's record: The linear measure is determined and defined by a measurement of transmission time. As a reason is given, that with today's clock technology a higher precision and reproducibility can be obtained, as with a rule or original meter. The exactness of going of the atomic clocks again depends on the free flying path (L in fig. 13.7) of the atoms. For the caesium clocks of the Physikalisch Technischen Bundesanstalt in Braunschweig the resonator length amounts to several meters! The clock is used world-wide as a standard.

The irony thus lies in the fact, that a geometric length dictates the measurement of time and the measurement of time again determines the measurement of length - poor science!

How does one free oneself from a capital closed loop conclusion? Why and how do signal transmission times or clocks actually depend on the gravitation? Who once got stuck in a dead end, knows that he only can get out in the reverse gear.

A possible way goes back to the roots of classical physics and to the theory of objectivity in the 1st part of the book, which is free from the limits of a subjective and relativistic observer standpoint. That isn't a dead end and in addition explains, why all atomic clocks react sensitive to magnetic fields (magnetostriction) and what these fields have to do with gravity (see chap. 6.9)!

Today's clocks are so exact, that even differences between a clock stationed on a mountain and one at sea-level can be recorded. Even more clearly was the depending on gravitation determined at an atomic clock, which was shot in a rocket 10000 kilometres high into space. The result of the analysis without doubt was, that the clock in that case doesn't „tick" correctly anymore.

But what does theoretical physics say about it? It claims, here the ,,red shift" has been measured; it thus concerns a confirmation of the special theory of relativity. But since it concerns a clock experiment and not a light signal, it clearly contradicts this theory, which isn't able to describe any gravitational effect at all, as is well-known. For this case in the spheres of theoretical physics one helps oneself with the general theory of relativity, with which actually only would be proven, that the two theories from the legacy of Einstein completely incompatible contradict each other. We come to the following conclusion:

Whoever gives details about length or time, is obliged to also indicate the reference system.

He also has to inform where his laboratory is situated and with which devices he measures!

With the conclusion also the discussion would be opened. Throughout the last four chapters a main idea can be found. It should be worth, to again think about the brought forward arguments and to dare a comparison with text books.

In the text books there doesn't exist such a thing as an oscillating interaction. Here no answer is found to the question, why the solar system isn't hurled out of the galaxy as a result of the high velocity, why the inside of the earth is hot, how the geomagnetism is formed, why the continents drift and why the ocean floor nowhere is older than 200 million years, as samples from the ocean floor prove<i>. How would you answer these central questions?

<i>: Kendrick Frazier: Das Sonnensystem, Time-Life Bucher, Amsterdam (1991).

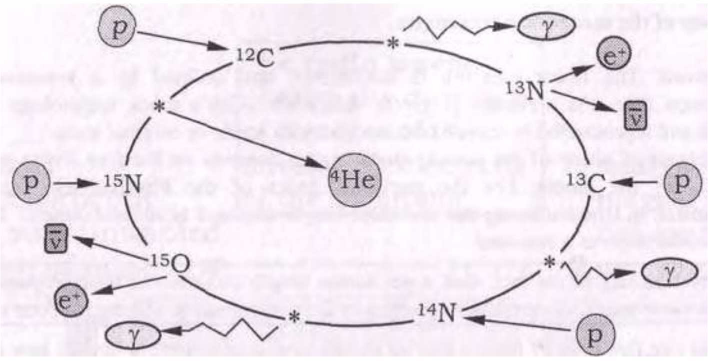
p
^{12}C
*
γ
e^+
^{13}N
$\bar{\nu}$
*
^{4}He
p
^{15}N
^{13}C
p
$\bar{\nu}$
*
γ
^{15}O
e^+
^{14}N
*
γ
p

Fig. 13.8: The Bethe-Weizsacker cycle (concerning the sun's fire)
(meaning of: ^{12}C = carbon nucleus, ^{4}He = helium nucl., γ = gamma quant, ^{14}N = nitrogen nucleus, p = proton, e^+ = positron, ν= anti neutrino, ^{16}O = oxygen nucleus

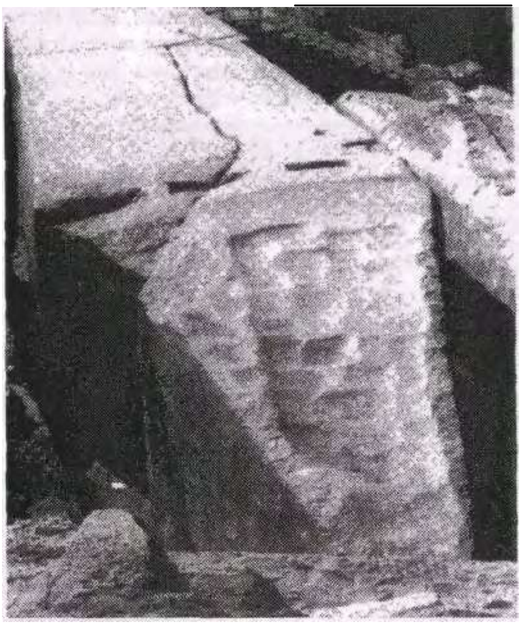

Fig. 13.9: The top of the unfinished obelisk in Assuan<i>

<i>: Hermann Wild: Technologien von gestern, Chancen fur morgen, Jupiter-Verlag Bern (1996), ISBN 3-906571-13-0

13.8 Discussion of the cosmological insights

The numerous models which are offered, of a geodynamo, an iron core, of assumed zones of subduction and of plate tectonics may be helpful for the explanation of individual isolated phenomena. But they physically don't give a uniform picture and partly contradict each other.

Lord Kelvin had calculated a period of shining of 5000 years on the basis of a sun consisting of coal<i>. The german physicist Hermann von Helmholtz landed at 15 million years, in which case the sun yearly should shrink for 100 meters. He already assumed a temperature of 15 million degrees centigrade, as also the Bethe-Weizsacker cycle has as a prerequisite, which according to today's concept should describe the process of nuclear fusion taking place in the inside of the sun (fig. 13.8). The only thing is that this extreme temperature is completely incompatible with the high density in the core of the sun! There are more than good reasons to assume the opposite of the widespread textbook opinion and assume that the core of the sun is cooled by the collected neutrinos in the same manner as the inner core of the earth and that superconducting areas are formed, which powerful fields even cause the protuberances on the surface of the sun. Since in the sun no measurement is possible, with which a model could be verified or disproved, terrestrial arguments naturally suggest themselves. Here the pieces of evidence for the growth of the earth can be taken in the hand and photographed. We already have discussed some facts. But there are found a multitude of other ones, for instance from the domain of archaeology, for which text books of today until now as well provide no explanation.

Possibly the stones and monoliths weighing several tons, as they were used for prehistoric buildings, should be linked with the growth of the earth. If the earth was smaller at the time they were build, then they perhaps by no means were as heavy as today!
Then the stones for reason of the smaller density in addition were softer and with that easier to work on. Dr. Wild points to the building technical peculiarity, that the stones formed like cushions originally must have been soft. He proves with the photograph of the top of the unfinished obelisk in Assuan (fig. 13.9), that the traces of working stem from a spatula in a plastic mass.
Also the perfect fitting of the stones used for the building of the pyramids can only be explained in this way. In the joints not even a knife point can be inserted!

If even stones in the course of time increase in density and hardness, then it is easier for us to comprehend, how small rivers in past time could dig large deep valleys in the earth's crust, then we perhaps also understand, why very old bones today are petrified<ii>. It is obvious, that also bones in the course of time increase in density and hardness. Even if science should succeed in breeding living dinosaurs, then their chance to survive in spite of that would be equal to zero, because the dinos would collapse under their weight of their own of several tons. Their bones would be much too thin and brittle for their weight of today!

<i>: Kendrick Frazier: Das Sonnensystem, Time-Life Bucher, Amsterdam (1991)
<ii>: The indication to petrified bones stems from a participant of the seminar.

25-07-1996: Assistant head of government department
Acceptance of the common standpoint.

11-12-1996: European parliament
Resolution of the guideline.

commission
Statement concerning proposals of change of the european parliament.

19-12-1996: Assistant head of government department
Resolution of the guideline.

19-02-1997: Coming into effect of the guideline
Publication in the official paper of the EC.

19-02-1999: Period of translation of two years

Reduction of the threshold values for current consumers:
19-02-1997: to 40 million kWh (opening of the market 23%)
19-02-2000: to 20 million kWh (opening of the market 28%)
19-02-2003: to 9 million kWh (opening of the market 33%)

until 2006: „Anti imbalance clause"
A nine years period of transition to preserve equal opportunities in the competition.
Report of the commission.

19-02-2006: Further stage of the liberalization
European commission tests by means of the made experiences, if the current market should be opened further.

Fig. 14.1: Timetable of the EC single market guideline electricity<i>

<i>: Grawe, S. Thiele: Vorbereitung der Stromversorger auf den Wettbewerbsmarkt, ATW Atomwirtschaft-Atomtechnik 43. Jg. 1998, Heft 1, S. 10 - 13

14. The way towards free energy

Can energy actually be produced? No energy supply enterprise is capable to do that. Fact is, that energy only is converted and not produced. For that the available resources on earth are tapped and brought into a utilizable form of energy. As different the conversion processes may be, finally always heat is formed. We thus gradually burn the globe, on which we live. How long can something like that go right?

Science gives all-clear: „stock of energy is sufficient for the next 100 years. Newest calculations disprove the fear of a scarcity of energy on earth. There exist large stocks of coal and oil"[i].

But what are 100 years compared to the age of the earth? How will our descendants judge our thinking and acting? They will condemn it and curse us, that much is clear already today.

We have an obligation to preserve the environment, and we only will be able to fulfil it, if we look to nature, how it covers its need of energy, if we finally understand and meaningfully copy nature.

We still are miles away of the goal.

14.1 The liberalization of the energy markets

It is important in the interest of a member of the executive that the stock of energy doesn't draw to an end, as long as he carries the responsibility for the enterprise of the economy of energy. The interest in the environment and the environmentalists for obvious reasons is less distinct. The problems are more of a commercial kind. Concerning that an actual example is given.

A special kind of problem is the meanwhile Europe wide valid decree of the EU concerning the liberalization of the energy markets (fig. 14.1). The list of the consequences starts with the fact, that the concluded licence contracts for power supply between energy supply enterprises and the communities, which are valid till 1997, are only wastepaper. Every customer of electricity as of now can conclude an individual contract with every "producer of electricity".

So an environmentally aware Black Forest person decides to obtain his power from a wind power station, which is situated in Denmark. That truly sounds very liberal. But how should that function, if there is no wind at all in Denmark for a week, or the connection is interrupted by a flash of lightning? The windmill further delivers power and the runner writes its bill truly believing, its power has arrived. The consumer then reads the bill at candlelight and puts with understandable anger his claim for compensation together. Pointing to the fact that such difficulties can be solved by book-keeping, the consumer again is calmed down; but that the not by contract obtained power will cost him dear, is another story.

<i>: Illustrierte Wissenschaft Nr. 6, Juni 1996, page 48-51.

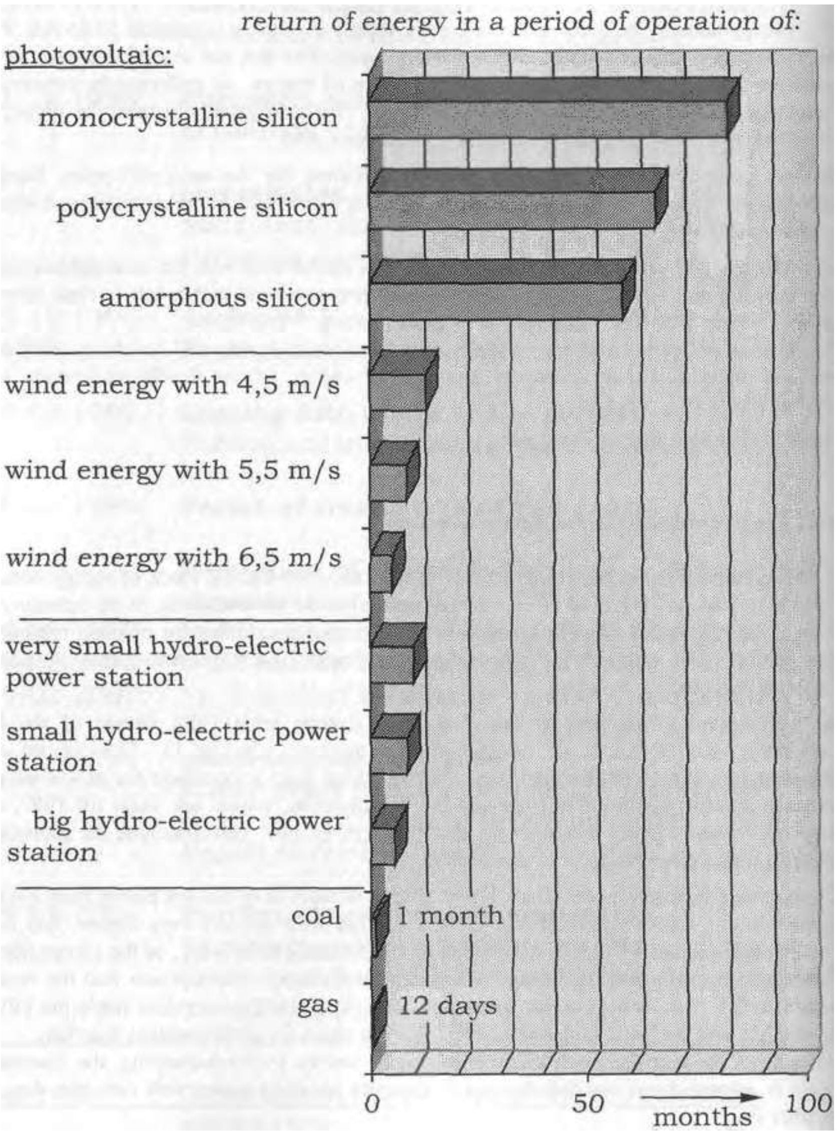

Fig. 14.2: Average time of paying for itself for the spent energy (return of energy)

14.2 The consequences of the liberalization

With the EU decree the monopoly only has shifted towards the runners of the distribution nets, after all only one net is present. Competition however would require at least two nets, thus a doubling of all house connections and all high-tension pylons, but that fortunately is unrealistic.

To prevent fleecing of the consumers by means of the net monopoly, politics introduces in the place of the free market economy dirigisme and plan economy with the well-known concomitants: no-one will look after the existing nets, no-one feels responsible anymore, because after all they have become public good by decree. One thus lets the pylons rot slowly and repairs only in emergencies. This possible development really can't have been the good intention of the EU.

In the next few years we however will be able to observe for the producers of energy exactly the development, which the EU-commissioners have imagined: Total competition contest, price war and a struggle for power to survive economically. A chance only has the nuclear power station, which gets rid of its refuse cheaper and if need be even illegally or which lives of indirect state subsidies, or the brown coal power station, which increases its efficiency at night, when no-one watches, by switching off the expensive filters.

The first power stations, which are selected out by the liberalized energy market, are the hydro-electric power stations which stand closest to nature. They simply are too small and too intensive of personnel, to be able to survive.

Gas turbine power stations, which deliver the power for half the price, than are ranking first. Then there is no place for regenerative systems anymore.

Solar energy, how many roofs a supporting program may have, stays a toy supported by the state. For a photovoltaic installation the „Return of Invest" still lies at more than 80 years, whereas is reckoned with a theoretical life of 20 years. In practical use on the other hand photovoltaic installations occasionally already have failed after seven years, after the photocells had gone blind. In this case not even can be talked of an ecologic energy, because the return of energy lies still above that. The supposed Idealist who spoils his roof with photovoltaic to reassure his ecologic conscience, would have spared the environment more, if he had covered his need of power from the socket, because already the production of the photovoltaic installation gobbles up more energy than can be produced with it. With regard to the environmental compatibility the ecologic balance sheet of a power station is attached a central importance (fig. 14.2).

With the law of feeding in power the state intervenes dirigiste, supposedly to protest the consumers. With this law the state orders, that not it, but the energy supply enterprises have to take the subsidizing of the regenerative energy, that they have to take over the power delivered at the inappropriate time in exactly so less suitable amounts at a price, which exceeds the market value by a factor of two and for falling prices of power even by a factor of four. With the law the politicians very fast and without agreement of the effected enterprises have shifted the „black Peter" further to these, which will get problems to preserve their competitiveness in the international comparison with the subsidy duty. It can't have been the intention of the EU-commissioners, that on a national level the price of power increases to finance some energy technical playground.

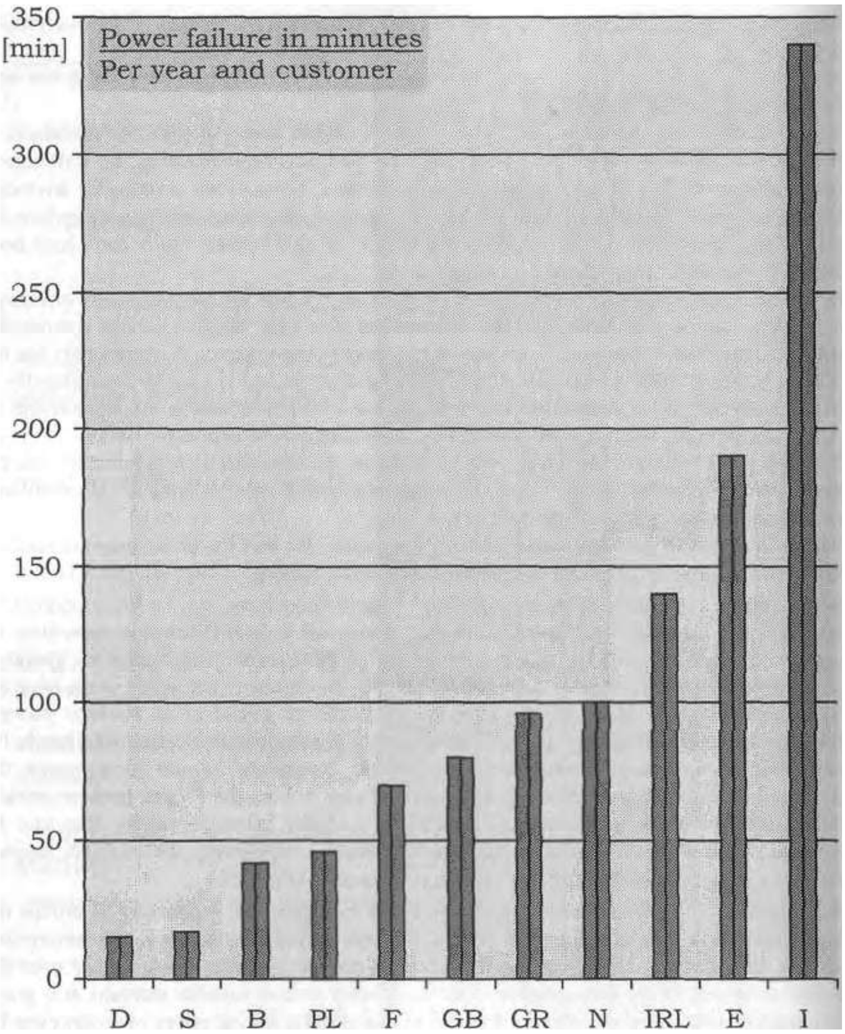

Fig. 14.3: Power interruption and not being at disposal

<i>: ETG, Energietechnische Gesellschaft im VDE: Qualitat der Stromversorgung, Dialog Nr. 1, Jan./Feb. 1998, S. 21

14.3 The chances of the wind energy

Meanwhile, as a result of the law of feeding in power, no longer the politicians but the energy suppliers are the ones, who around the North Sea look out of the window every day with the worry, some wind could blow and the mills could turn, because every kilowatt hour of a windmill must be subsidized strongly. Every windless day however reduces the power bill, with that helps the consumer and raises the chances of competition of our economy.

If it should be neglected, to tip over the law of feeding in power in due time with help of the EU, then the consumers will get their power in future abroad, then at the Preussen Elektra, damaged most by wind energy, as the first the lights will go out, then economic power and prosperity in Germany in future are dictated from abroad.

But if the law is dropped, then with that the duty to subsidize the regenerative energy carriers drops back to the state. Now all tax payers may foot the bill, even those, who don't use any power at all. The well-known Justice of subsidy" comes into effect.

Without support by the state only few types of power stations will be left behind. In addition are overcapacities being reduced, because they only cost money. But both is at the expense of the reliability of the delivering of energy and of the safety of the consumer. We owe the high stability of our network of today the large number of most different providers of power, which cover the basic load up to the peak load according to their suitability (fig. 14.3). But let us not talk of the golden past. The network after all isn't able to store power. From that follows, that without redundancy and without free power station capacities which can be activated at any time a short overload is sufficient, to let the network collapse.

Once the EU guideline concerning the liberalization of the energy markets is in effect, when persons selling power wend their way from front door to front door, to convince the housewives to obtain the power from them and not from other hawkers, when the power stations only live of stock and the depreciated overhead power lines only are entered in the hooks with the scrap value, then we should be dressed warmly and always have ready sufficient candles <i>.

14.4 The chances of being self-sufficient concerning energy

The only way out is the decentralized energy supply, the getting out of the large energy union and the way towards being self-sufficient. The argument sounds convincing in view of missing alternatives. But now it no longer is possible to plug the plug in the socket and then switch on whenever it pleases us. First it has to be calculated, if the windmill or the installation for solar energy supplies sufficient power or if a cold meal should be made. In contrast to today's consumer habits the runners of such installations will have to adapt their need of energy to the prevailing weather conditions.

<i>: A baby-Boom, like after the big power failure in New York, would be the smaller evil. The inhabitants and shopkeepers in the New Zealand metropolis Auckland will long and frightened remember the power failure in february 1998, which lasted weeks and was a result of not carried out maintenance works of high-tension pylons.

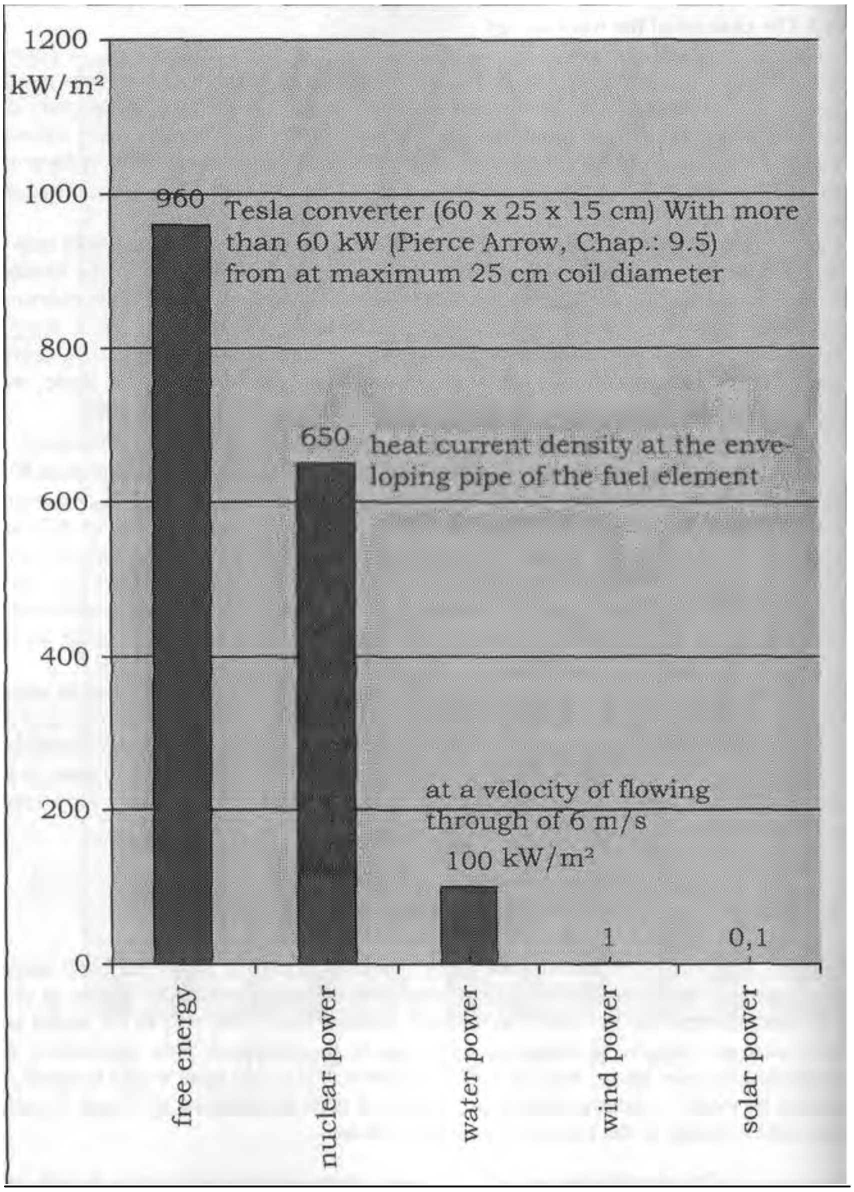

Fig. 14.4: At maximum obtainable power density of different carriers of energy

<i>: ETZ, Heft 19/1995, S. 40

Block heating power stations, propagated as stand-alone solution for people being self-sufficient, deliver power and heat at the same time. But if I don't need any heat on a hot summer day, then also no power is available or I uselessly heat into the open. For that one may leave all electric consumers switched on in winter, even if they aren't used at all, only to get the hut warm. Does a stand-alone solution look like that?

The situation truly is demotivating. Even the energy suppliers meanwhile may have realized that the energy politic way is a dead end. But for real alternatives in energy technology the pressure by suffering still doesn't seem to be big enough.
The intention to learn of nature is present in principle. The solar fire one wants to kindle on earth in a fusion oven, but the oven does not as it should.
Obviously the sun functions completely different, as physics imagines today (see fig. 13.8).
Before copying stands understanding, and in there seems to be a hitch!

14.5 The space energy technology of nature

Also nature needs energy, even very much. But it hasn't got any connecting pieces for tanking and no oven lid to fill in the fuel, it doesn't know our ecologically harmful combustion technology and environment destroying explosion technology at all. Nature rather works with the opposite, with implosion and fusion.
The sun, we have derived, materializes the matter which it needs for growing and shining from the neutrino field. The earth and other planets imitate the sun.
The concept is simple and convincing. The source of energy lies in the air and mustn't be dragged about in tanks. Collected and materialized is just as much, as is needed at the moment. In that way the resource energy is spared. In addition there can be done without any sort of storing. In addition it can show a substantially higher power density, than all today known and used energy carriers (fig. 14.4).
Such a source of energy solves all described energy problems at once. Nature wouldn't be as we know it, if it wouldn't have this ideal energy, also called „free energy". The balance sheet of energy alone brings it to light, because as a rule it doesn't work out for biological systems. Often more energy is released than is taken up by the food.
In that case some migratory birds materially seen should have completely used themselves up before reaching their destination, if the energy necessary for the flight would be of purely material nature. From a concrete example the following is reported<i>: „migratory birds have - depending on kind - a maximum range velocity between 24 and 83 km/h and at their Atlantic flights no opportunity for an intermediate landing. They are thousands of kilometres on the way and hardly lose weight. For instance an Albatross with a body length of up to 1.20 meters and a wing span of up to 3.50 meters uses per kilometre only 8.5 grams of weight at a non-stop flight. How is that possible without additional supply of energy?"

<i>: K. E. Rathgeb: Wie man die freie Energie anzapft: Vogel machen es uns vor, Raum & Zeit 79/96, S. 74

Fig. 14.6: Freiherr von Reichenbach (76 years old)

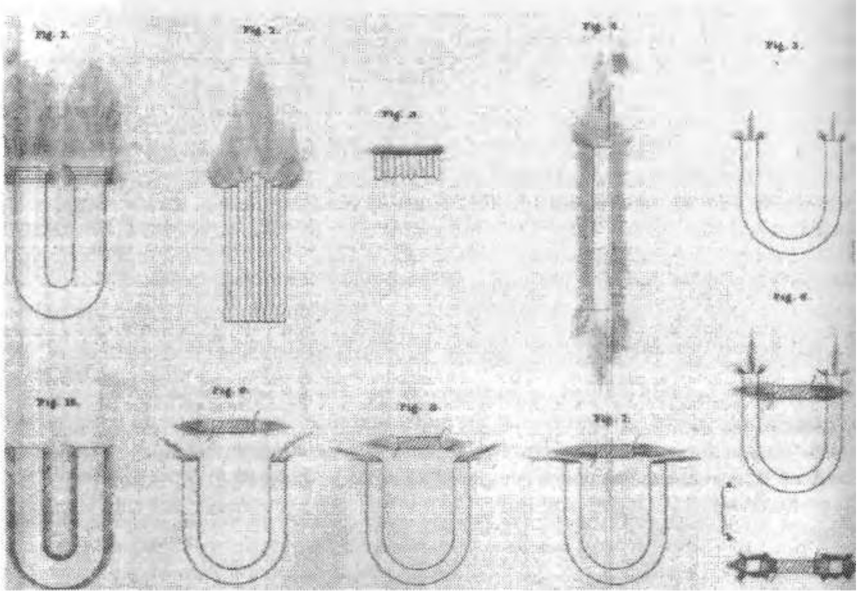

Fig. 14.7: Odic emanations of magnets according to pictures of sensitives .[i]

---

[i] H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, Seite 13 und 15

14.6 The gap in the energy balance of man

Even man appears to tap an additional source of energy, because for reason of scientific tests with recruits these over longer periods of time are able to physically release more than twice the amount of energy they take of calories with the food. Obviously living beings don't rely on one source of energy alone. Possibly the taking up of food predominantly serves the metabolism, and the energy aspect plays only a secondary role.

This interpretation at least would be obvious, since we scoff several times a day, sort the necessary and useful building-materials out and hand over the rest to the purifying plant and nature, where some bacteria and organisms search the sorted out again for useful things. "Food chain" we call this kind of building-material trade.

If the whole had anything to do with energy or with a „ combustion process without fire ", then no animal nor any human being could do without the taking up of food for longer periods of time. But Franciscus of Assisi could fast 90 days, as is handed down to us <ii>. There exist numerous examples, which reach into the time of today.

The mitochondria, the energy centres of each cell, by no means are capable of the "combustion process", which man ascribes to them. Here in all probability a taking up of space energy independent of food takes place.

A research scientist only has to look at nature with open eyes, what unfortunately happens rarer and rarer, because the laboratory scientist always is troubled, to keep the disturbing factor „nature " away from the experiments.

14.7 Carl Freiherr von Reichenbach

In this context no-one can go past two natural scientists: Carl Freiherr von Reichenbach (1780-1869) and Dr. Wilhelm Reich (1897-1957).

Reichenbach called the by him investigated life energy „ Od-energy " in the style of the Teutonic God Odin. He worked with test persons, who could perceive actually invisible light phenomena and worked out the special properties of this Od-energy field with the "sensitives", as he called them.

A quotation from his work shows however that the knowledge about the life energy must be a lot older<iii> than his own discoveries: „On paintings saints often are shown with a ring-like aureole around their head, something I before this would have dismissed as a pure figment of imagination. But it was shown that this glowing ring actually can be perceived by the sensitives as an Od-phenomenon and so the aureole obviously can be traced back to real impressions of particularly sensitive persons."

Reichenbach also found out that water has a big endeavour to take up this Od-energy or in the language of the present day, to absorb the field energy. This circumstance we find confirmed in technology, since water absorbs high-frequency waves, whereas an insulator or a vacuum lets them pass through. But without vortex physics it however remains entirely unclarified, why!

<ii>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, page 110

<iii>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, page 16

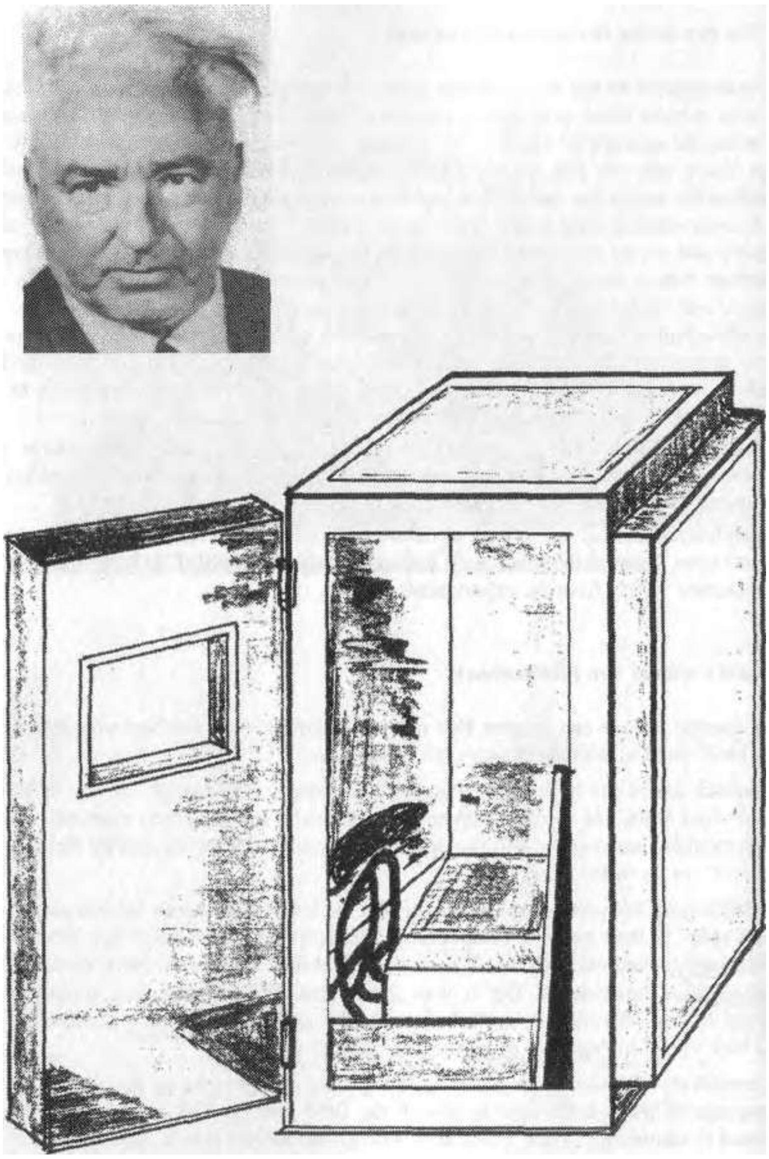

Fig. 14.8: Wilhelm Reich and the model of an Orgon accumulator for applications of the whole body.<i>

<i>: taken from: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, pages 38 and 89.

14.8 Cold fusion and Genesis

Chronologically the doctor and psychologist Wilhelm Reich followed in the footprints of Reichenbach. He merely altered the name for Od and spoke of Orgon. His speciality was the accumulation of Orgon radiation, I would say the focussing of neutrino radiation. Actually the properties like e.g. the missing possibility to shield Od, Orgon and the neutrinos are identical, so that we can proceed from assumption, that it also concerns the same physically.

We already talked about Wilhelm Reich (part 1, chapter 9.2). He could show that the measurable temperature in a closed box, constructed like an „Orgon accumulator", is increased in a mysterious manner without supply of energy from the outside. He even could prove, that this energy actually concerns the sought-for life energy, as he observed the creation of life in the laboratory under the microscope. For that he cooked muscle fibres, vegetables or other cells so long until the cell structure had been destroyed entirely. But from the educts entirely by itself new living beings, like protozoa or algae, were formed<i>.

Reich at his microscope actually followed the transition of dead to living matter. What is of interest here, at first only is the energy technical aspect of this conversion. Later we will occupy us with the at the same time occurring information technical aspect.

Worth mentioning seems to me the experiment with the „silly" chickens, which have at their disposal astonishing abilities besides the laying of eggs<ii>.

The test chickens were handed chicken food, from which to a large extent all calcium had been extracted. But the chickens showed themselves unimpressed and further laid keenly their eggs. The experimenters were surprised, where the chickens actually got the lime for the egg shells. How solves such a chicken the problem of raw materials? For that further materials were extracted from the food and look, at the removing of silicon the laying of eggs was over. The experiment actually only allows the conclusion, that the "stupid" chicken is capable of a cold fusion, that it itself „produces" the necessary calcium from silicon presumably by using carbon. Every alchemist here has to go pale with envy. But what says the research scientist of fusion to that, who actually should know, how fusion functions? After all he is paid for it by the tax payer! The chicken uses the fusion already today and the other living beings presumably also, but for that energy is needed and the balance sheet should work out!

The neutrino radiation therefore has to be factored into the balance sheet of energy. If the balance sheet then works out, it could be proven with that, that here neutrino energy is put to use. In addition the process of the conversion of neutrinos has to be investigated, which surely has something to do with the frequency and the wavelength of the radiation. After all a child has cells exactly as big as an adult. It only has less cells! So that an interaction can occur, the cell size has to fit in with the wavelength, which obviously is very small, presumably in the range of the radioactive radiation, with which circumstance the biological incompatibility with this electromagnetic wave would be explicable.

<i>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, S. 39

<ii>: Louis Kervran: Biological Transmutations, s.a. J. Heinzerling: Energie, S.278

1. Primary reaction
   (light reaction in the centre of reaction):

   ring-like vortices/neutrinos are being collected
   electrons are formed
   energy is produced
   e^- is attracted by the water dipole
   water molecule is being splitted (photolysis):

2. Secondary reaction (dark reaction):

   carbon dioxide is reduced with the hydrogen splitted off from the water to carbohydrate

---

Reaction equation of photosynthesis:

	6 mol CO_2	(carbon dioxide)
+	12 mol H_2O	(water)
+	675 Kcal	(vortex energy, light)
	1 mol $C_6H_{12}O_6$	(glucose)
+	6 mol O_2	(oxygen)
+	6 mol H_2O	(water)

Fig. 14.9: Concerning photosynthesis

14.9 Photosynthesis

The materialization of free electrons is a prerequisite to start the photosynthesis. The normally used explanation, the free electron necessary for the splitting of the water molecule was knocked out an atom by light, doesn't seem to be correct, after until now all attempts of a technical realization according to this model concept have failed. At the photosynthesis the plants obviously help themselves with the neutrino radiation, which according to an estimation of today with 66 billion particles per second and per square centimetre might be more than sufficient for a green earth.

If we put a seed in a water glass, then a plant grows from it and forms small leaves, which get bigger and bigger. A wonder of nature, we say. Where does it actually get its building-materials? From the water or from the air? Necessarily the plant obviously produces a part of the matter itself!

Experts think they have understood the process of photosynthesis: Take light and water and carbon dioxide and handicraft from that sugar and oxygen. But from where does the plant take the necessary energy for the rebuilding and the splitting of the water molecule, the photolyse? The taken up solar energy hardly is sufficient for that, especially since the plants only absorb about 1% of the photosynthetic utilizable sunlight incident on earth<i>. By means of reception molecules, which look like small antennas, pigments less than 30 nm in diameter, such is the level of knowledge, the sunlight is collected and led into a photochemical centre of reaction. Here the reaction should take place, provided that an electron set free by the light jumps into the middle of the centre of reaction. But exactly this favour the electron doesn't make the research scientists, who want to imitate the process. Copying nature still doesn't succeed.

The mistake presumably lies in the circumstance that the light doesn't set free any electron at all. The electron actually first of all is produced in the centre of reaction. By means of the antennas a neutrino vortex is collected, which at first occupies the entire space, to afterwards contract to an electron, which as a result automatically is centred in the centre of reaction.

In the process of materialization at the same time the necessary energy of the process is formed. The reaction equation after all also has to work out energetically, because the plants doing so get neither hot nor cold (fig. 14.9).

Because curiously the light reaction even can be observed in the dark<ii>, one could be inclined to in principle call into question the influence of light on the photosynthesis. But such an influence nevertheless seems to be present, after all does a plant react on the irradiation of light and changes its spectrum of absorption through its colour. But it for sure is another influence and not the one, which one attributes the green plants today!

It would be important to finally understand the way of functioning. The plants and particularly the algae are the ones, which actually first have made possible life on this planet with the photosynthesis, the most original form of a production of matter and energy.

<i>: Luttge, U. u.a.: Botanik, VCH, Weinheim (1994) ISBN 3-527-30031-7, S. 136

<ii>: Luttge, U. u.a.: Botanik, VCH, Weinheim (1994) ISBN 3-527-30031-7, S. 126

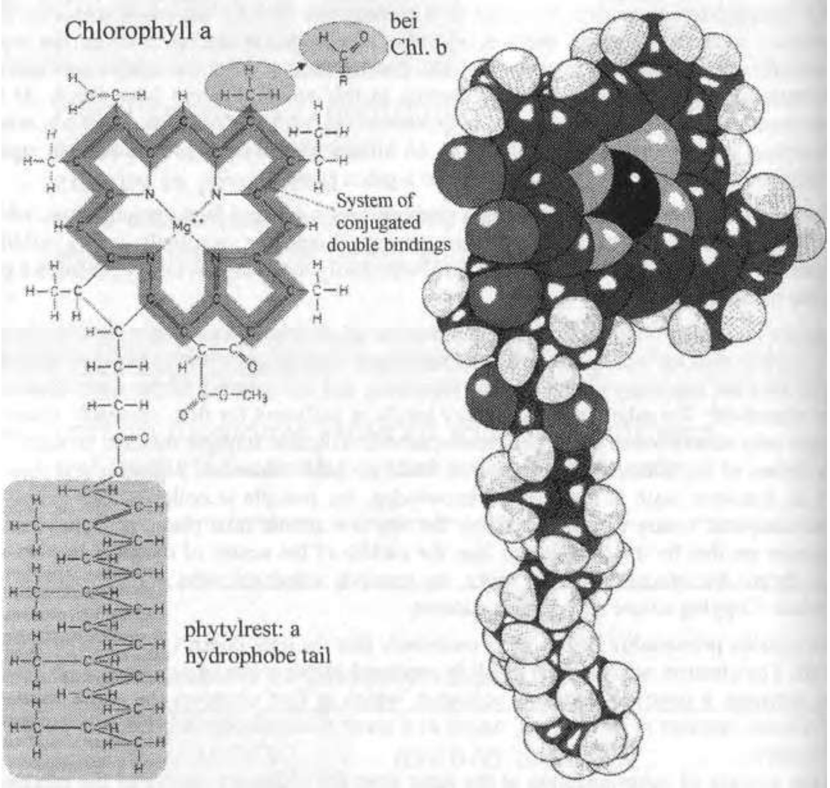
Chlorophyll a
bei
Chl. b
Mg++
System of
conjugated
double bindings
phytylrest: a
hydrophobe tail

Fig. 14.10: Formula of structure and spatial model of the photosynthesis pigment chlorophyll.<i>

<i>: Luttge, U.et al.: Botanik, VCH, Weinheim (1994) ISBN 3-527-30031-7, p. 118

14.10 How nature materializes

The structure of the receiver antenna allows us a deeper insight into the manner, how free energy is tapped during the photosynthesis. For the dimensions determining the frequency of the photosynthesis pigments, as the antennas are called, we have to go down to the molecular structure. Fig. 14.10 shows the formula of structure and next to it also the spatial model of the pigment chlorophyll[i].

Two points point at the function for free energy conversion. On the one hand is situated in the centre of the molecule a double positive ionised Mg atom, surrounded by four nitrogen atoms and a carbon ring consisting of 20 atoms, from which arises a polarization of the entire molecule. This in addition is oscillating because the electron cloud of the enveloping electrons, which hold the molecule together, perform swirl oscillations depending on the temperature. With that chlorophyll is able to go into resonance with oscillating neutrinos.
A role play the unipolar field configuration and the effect of resonance of the molecular oscillation of its own forming as a result of the polarization. A further role in addition seems to play the spatial structure.

The model of the „receiver antenna" chlorophyll taken out of a textbook[i] and shown in fig. 14.10, consists of a stalk and a spirally wound head, which resembles a Lituus or crook, which Etruscan and Roman Augurs have taken in the hands for land surveying, a precursor of the crosier (see fig. 16.10). This again has the form of a Tesla coil and that, as already derived, is able to withdraw rotational energy from the collected neutrinos (chapter 9.8). Doing so free electrons are materialized, and these then start the process of photosynthesis. An explanation concerning the way of functioning of the antenna pigments here for the first time is getting available.
By the way also the mitochondria, which form the energy centres in every cell, have as well the form of a Tesla coil. Whoever wants to understand the energy economy of a cell or the photosynthesis first should occupy himself with the Tesla coil (chapter 9.8).

The open question, how land surveying should be possible with a Tesla coil, we at first still have to shelve, because in this chapter it concerns the way towards free energy and the chance to learn of nature. Examples to be looked at are on the one hand the core of fusion in the inside of the earth and on the other hand the humus layer on the surface of the earth, which has been materialized in the course of time with the help of photosynthesis.

The goal seems to be worth striving for. If we in the first place have learned to produce energy exactly like nature, then we'll further try to produce matter purposeful, with which ageold alchemist dreams could be fulfilled. We wouldn't need to scrape and to search in some mines any longer. We would materialize the products without refuse, naturally and just for the environment, direct in the final form. I admit that at present it sounds pretty futuristic.

With another example, the lightning and particularly the ball-lightning, the collection concerning the use of free energy in nature shall be completed.

---

[i] U. Luttge, M. Kluge, G. Bauer: Botanik, 2. Aufl.- Weinheim, VCH (1994) ISBN 3-527-30031-7, S. 118

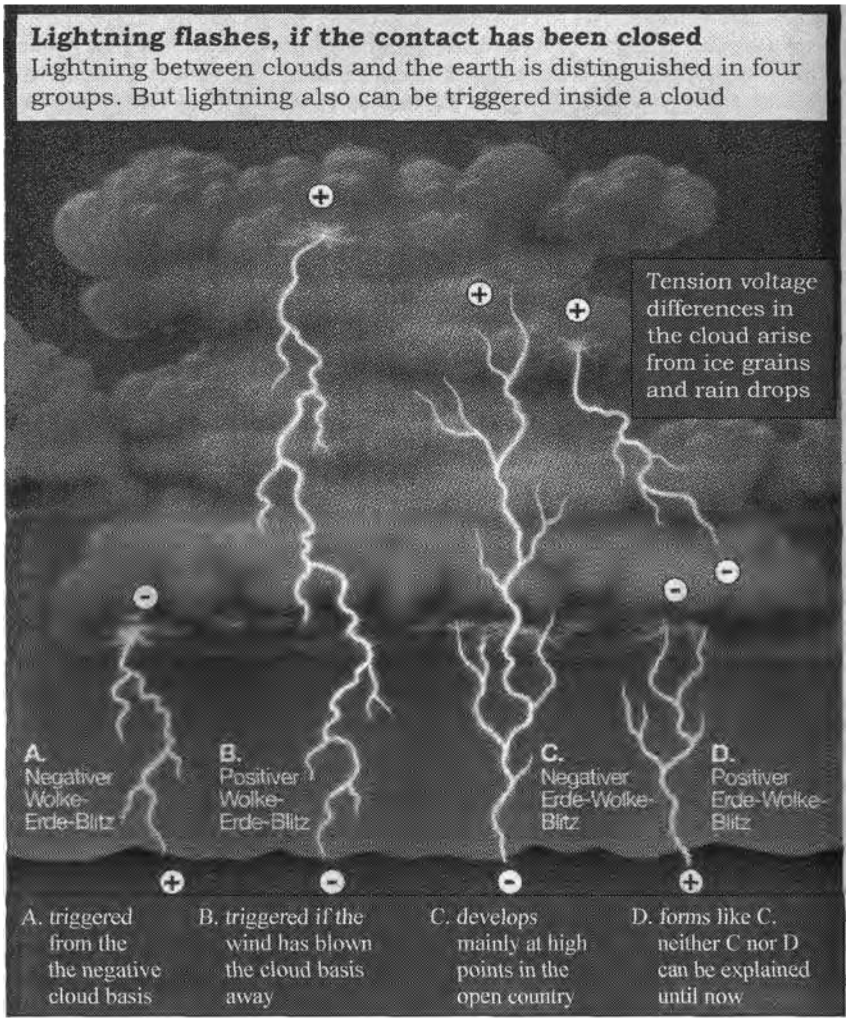
Lightning flashes, if the contact has been closed
Lightning between clouds and the earth is distinguished in four groups. But lightning also can be triggered inside a cloud
Tension voltage differences in the cloud arise from ice grains and rain drops
A. Negativer Wolke-Erde-Blitz
B. Positiver Wolke-Erde-Blitz
C. Negativer Erde-Wolke-Blitz
D. Positiver Erde-Wolke-Blitz
A. triggered from the the negative cloud basis
B. triggered if the wind has blown the cloud basis away
C. develops mainly at high points in the open country
D. forms like C. neither C nor D can be explained until now

Fig. 14.11: Concerning lightning.<i>

<i>: taken out of: Illustrierte Wissenschaft Nr. 8, August 1995, Das unverstandene Phanomen der Blitze, S. 13

14.11 Lightning

Lightning is a spectacle of nature just as imposing as unsolved. It concerns an electric process of discharge, in which to the amazement of all experts arrive for several powers of ten more charge carriers at the surface of the earth, than hefore were contained in the cloud! Furthermore unsolved is, why lightning glows. Let us start with the open question for the difference of potential necessary so that the air is ionised and a lightning channel is formed.

If at first the electric tension voltage of 200,000 volts between the surface of the earth and a height of 10 kilometres is available, which according to our calculations results from the rotation of the earth and the magnetism of the earth (fig. 11.8). Another source of tension voltage is not known. By air movement and supposed processes inside a cloud locally an additional accumulation of charge carriers may occur, so that in the case of a thunderstorm the by Tesla assumed, twice as big value doesn't seem unrealistic<i>.
On the other hand a tension voltage between 4,000 and 10,000 volts is necessary so that a blow can occur for an air gap of one centimetre depending on the atmospheric humidity<ii>. Linearly projected the thunderstorm cloud theoretically should hover just one meter above the earth if there is lightning.
Here somehow a huge gap gapes between theory and practice!

At the latest at the spikes, the mysterious lightning, which strike out of a thunderstorm cloud upwards in the direction of the ionosphere and the still more mysterious ball-lightning it has to get clear, that the actual source of tension voltage of a lightning by no means is known. Without knowledge about the cause we'll never understand lightning.

The potential vortex theory offers an useful approach, according to which the necessary difference of potential is formed from a formation of vortices<iii>. The vortex again, as damping term in the wave equation, occurs as a result of intense sun irradiation. That explains why lightning always strike from the part of a cloud, which is the darkest, where the most sunlight is absorbed and the damping is the largest.

The possible formation of ice in a lightning channel is a further confirmation for the correctness of the vortex explanation. After all it has been derived that contracting potential vortices withdraw heat in principle (fig. 12.8).

But now we also want to know from this efficient theory, why photons and electrons are formed during the lightning and where they come from. Here obviously energy is formed by means of materializing of vortex particles.

<i>: R. L. Clark: Tesla Scalar Wave Systems, The Earth as a Capacitor, The Fantastic Inventions of Nikola Tesla, ISBN 0-932813-19-4, S. 265

<ii>: Karl Kupfmuller: Einfuhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Auflage 1988, ISBN 3-540-18403-1, S. 221

<iii>: K. Meyl: Potential vortices, Part 1: Discussion contributions to the natural scientific interpretation and concerning the physical-technical usage, basing on a mathematical calculation of newly discovered hydrotic vortices, (only in german) INDEL GmbH, Verlagsabteilung, Villingen-Schwenningen 1990.

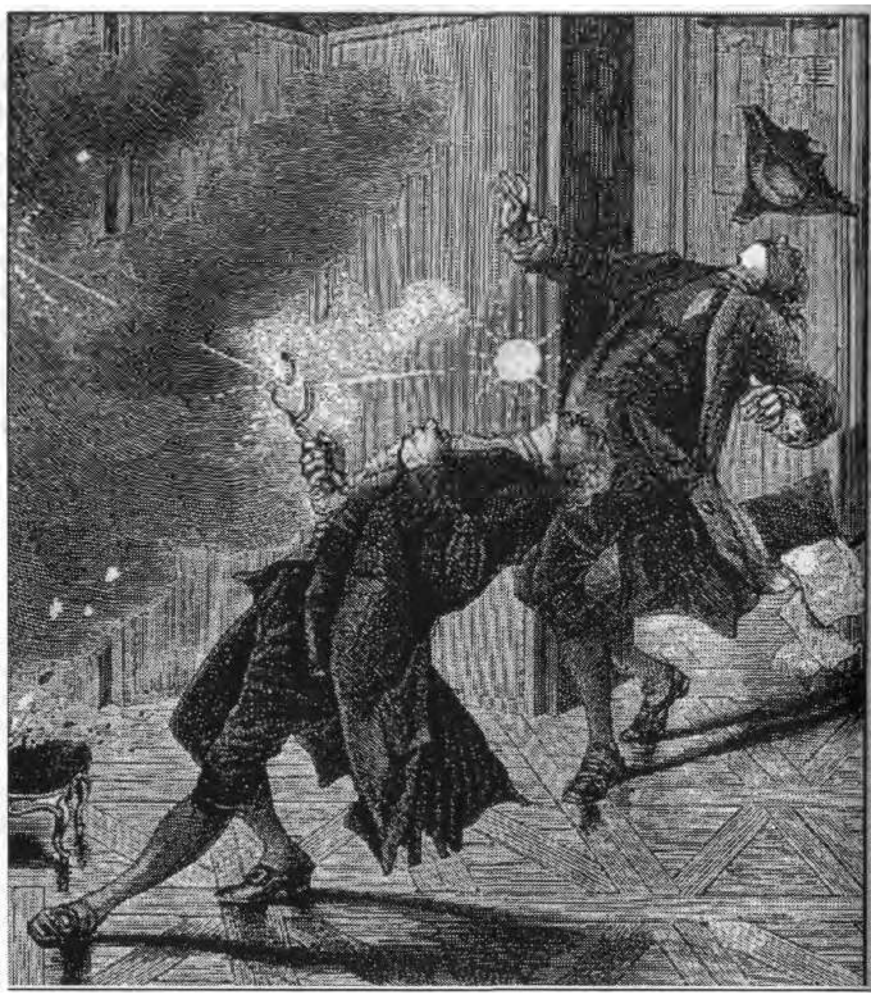

Fig. 14.12: Contemporary representation, how 1753 lightning research scientist Prof. Richmann is struck deadly by ball-lightning in his laboratory.<i>

<i>: Illustrierte Wissenschaft Nr. 8, August 1995: Das unverstandene Phanomen der Blitze, page 13

I proceed from the assumption that lightning collects and converts neutrinos. The process corresponds to the one at the sun, but on a smaller scale and only for a very short time. The lightning channel is polarized by the charge carriers. Change of temperature and field lead to a spatial oscillation, which by the way also functions as a source of sound, as anyone can hear. Taken both together lightning, seen from the outside, becomes an unipolar resonator, which is capable to attract neutrinos and to go into resonance with them. Now the predominant part is converted into electrons, because also the air molecules and air ions in the lightning channel belong to the world of matter. But it can't be avoided that a small part of antiparticles is formed, which then annihilate with particles of matter under emission of radiation. Doing so photons are emitted and lightning glows, as anyone can see!

14.12 Ball-lightning

If in the case of lightning there still exist excuses, the difference of potential preferably is traced back to neither understandable nor measurable processes inside a cloud, then at the latest in the case of ball-lightning most experts are at their wits end. Only for very simple natures explanations circulate in the direction that here for instance the organic remnants of a bird struck by lightning are burnt off.

Actually ball-lightning is observed very seldom. It is a ball flashing with reddish till blue-white colour. Its diameter lies between 10 and 50 centimetres. The glowing phenomenon can last several seconds to minutes. Doing so ball-lightning rolls over a street, temporary floats in the air, goes apparently unhindered through every wall and disappears from time to time without a trace or discharges with loud moise and formation of sparks. Some stink of poisonous gases and some also cause noise.

Famous has gotten the ball-lightning, which 1753 of all people should have struck the lightning research scientist Professor Georg Wilhelm Richmann in St. Petersburg. In his laboratory during a thunderstorm a ball of fire as big as a fist should have jumped from a iron tube to his head and should have hunted him down, so eye witnesses have reported (fig. 14.12).

Since ball-lightning has a closed structure, it has to drag about its source of energy with it. If this however consists of organic or other matter, the ball wouldn't be able to float, after all the brightness of a spherical vortex and with that the need of energy is enormous! We have to proceed from the assumption that just ball-lightning covers its need of energy from free energy and serves itself from the neutrino field.

The spherical form is a consequence of the structure shaping property of the potential vortex<ii>. Scientists are increasingly interested in this not understood phenomenon. In their experiments they try to artificially produce ball-lightning with more or less great success in the laboratory. By means of the experiment they then want to learn to understand, what the textbooks don't give away.

If we want to learn of nature something about free energy, lightning in the laboratory offers us in the form of a blow or of a spark gap relatively good possibilities. It surely is no coincidence that the father of free energy, Nikola Tesla, in his experiments almost all the time has worked with spark gaps!

<ii>: for that see in the 1st part the chapters 4.8 till 4.10

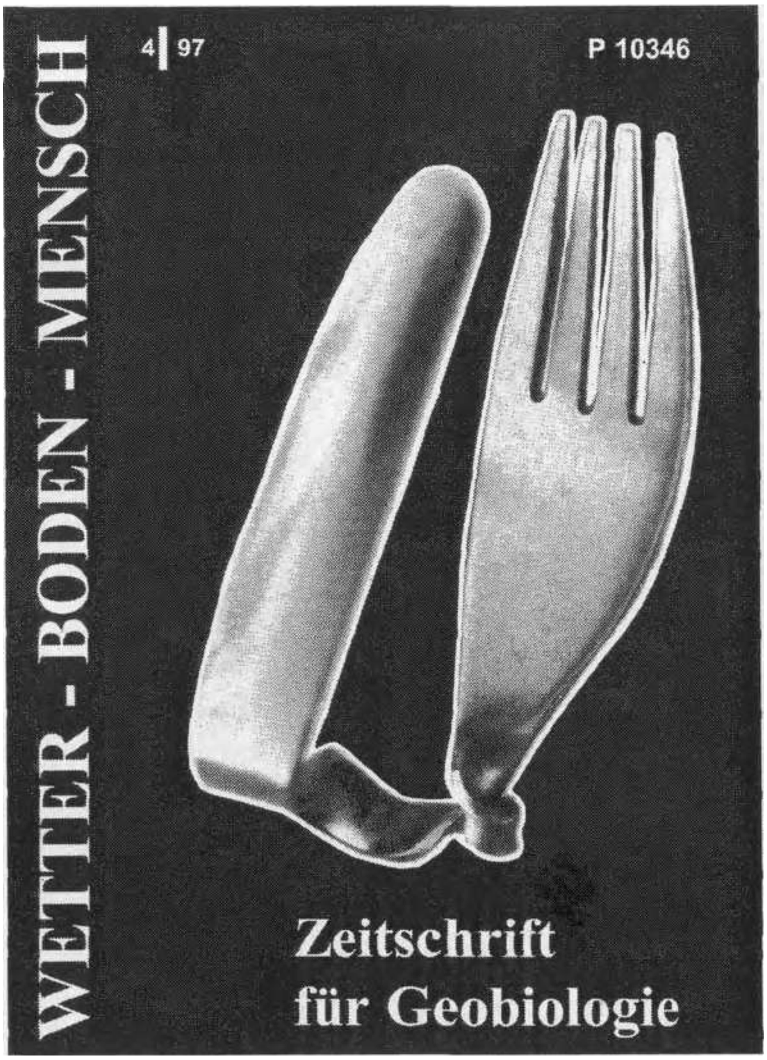

Fig. 14.13: Fork bent by „pure manual work".<i>

<i>: Whoever has to eat with bent flatware in the Hotel Adler in Waldkatzenbach (Odenwald, Germany), mustn't believe that the food didn't taste well to some rude fellow here. It is the host himself, who in a sociable round shows his guests from time to time that he not only can cook well. I could personally convince myself from the fact that Robert Hartmann doesn't use any aids. The picture shows the title page of the magazine „Wetter Boden Mensch" 4/97. A comment concerning this on page 3: Robert's 50ieth birthday.

14.13 Discussion concerning the neutrino conversion

In nature two different principles for the use of the neutrino radiation can be observed. In one case of use a glowing phenomenon occurs, in the other cases not.

The glowing lightning should be assigned to the first case. Here not only the neutrinos crossing the lightning channel are used, but also neutrinos from the environment oscillating in resonance are collected. Around the lightning channel a kind of reaction radius is formed, inside of which all synchronously oscillating neutrinos are attracted under the resonant interaction. During the discharge process the radius can increase like an avalanche, to collapse again with the transition into the stationary discharge current. A corresponding technical concept, which is discussed in the next chapter (15.5), is very efficient, but almost uncontrollable. A characteristic is the formation of antiparticles, of positrons, in the course of the materialization, which then annihilate under emission of radiation and cause a shining. Also spark gaps and fluorescent lamps possibly help themselves from the neutrino field, as still has to be to worked out.

In the second case nature works without any avalanche effect and without any glowing phenomenon. To this counts for instance the photosynthesis or the mode of operation of the mitochondria, the energy centres of a cell. All cells, whether vegetable, animal or human, only use the neutrinos which just that moment pass by them and only in those amounts, as they just are needed. They thus handle their energy very caring. Without exception electrons are materialized and no positrons.

If nevertheless unwantedly an avalanche effect occurs, something which happens fortunately only very seldom, then a self-inflammation and self-burning occurs, then it should happen, that a person burns off himself<ii>. This risk also is known of hay.

There also exist rare talented persons, who can control and regulate the process of materialization by concentration. These people can bend spoons or other metallic objects<i>. For that they concentrate themselves some time and send the materialized charge carriers into the object, which they hold in their hands. Since the metal lattice solely is kept together by the enveloping electrons of the individual atoms, the additional electrons make the metal structure sodden. Now for a short moment the metal can be bent and distorted at will. Doing so neither heat is produced nor is the colour changed. The result of the process can be produced neither by cold working with raw force nor under a flue. Also here nature shows us a technology for an ecologically compatible metal processing (fig. 14.13).

The way from the conventional over the regenerative towards free energy is predrawn. It only has to be gone! After the existence of the neutrino radiation goes as proven and 1998 for the first time concrete amounts have been determined measuring technical and published by a Japanese team of research scientists, with that also the question for an energy technical use of the particle radiation has been answered clearly. Now only the question of the mechanics is open. The Japanese research scientists by the way have found out that at night only half as much solar neutrinos can be detected than at daytime. The other half according to that is absorbed in the inside of the earth. This in the meantime published measurement result in brilliant manner confirms the working hypothesis of a growing globe (chapter 11).

<ii>: Zu Asche pulverisiert, Illustrierte Wissenschaft 6/ 1997, S. 61 acc. to an examination of the American SCI-COP in 1984.

cause/ field lines	interaction	F = force effect *mediation:*	see chapter: *effect/ application*
1. open **H**-field lines	magnetic (static)	F_{M0} = magnetic force (permanent magnet)	chapter 15.4 *E.-engine, solenoid*
2. open **H**-field lines	resonant (oscillating)	F_{MS} = magnetic force (AC-magnet)	15.4 + 15.5 - 15.7 lightning, *railgun*
3. open **E**-field lines	electric (static)	F_{E0} = Coulomb force *by charge carriers*	15.9 + 15.10 atomic bond, *Testatika*
4. open **E**-field lines	resonant (oscillating)	F_{ES} = Coulomb force *by neutrinos*	chapter 16 + 17 weak interact. galaxy bond *Tesla converter*
5. closed **H**-field lines	gravitation (static)	F_{MG} = gravity *by particles with mass*	chapter 6.9 + 7 elementary particle mass
6. closed **H**-field lines	levitation (dynamic)	F_{ML} = reduced gravity	chapter 18.3 gravit. waves *Casimir effect*
7. closed **E**-field lines	gravitation (static)	F_{EG} (force hardly detectable)	chapter 18.7 *superconducting ring*
8. closed **E**-field lines	levitation (dynamic)	F_{EL} (no longer detectable)	18.5 + 18.6 *Keely-/Searl-flying devices*

electromagnetic and resonant interaction		gravitation and levitation
$F_{M0} = 1{,}4 \cdot F_{MS} > F_{E0} = 1{,}4 \cdot F_{ES}$	$\gg$	$F_{MG} = 1{,}4 \cdot F_{ML} > F_{EG} = 1{,}4 \cdot F_{EL}$
for reason of open field lines		resp. closed field lines

Table 15.1: The force effect of interactions, ordered according to size, with examples

15. Principle of functioning of space energy

In this chapter we want to turn us towards the technical concepts and techniques concerning "space energy", which occasionally here and there already should have existed or have been operated with quite different success. After Nikola Tesla having pointed the direction more than 100 years ago, the way towards free energy appears to be predrawn. Never before the public interest in the topic of space energy was as big as today. Unfortunately this concerns more the collecting and gathering of rumours and speculations. As in every branch of science also here hunters and collectors can be found. But obviously the hunters, the inventors and theorists have bigger problems, to put something useful on the table. They are fighting against their own not knowing, unuseful textbooks, general ignorance, intolerance and an all-powerful energy lobby. What the collectors on the other hand come up with does make appetite but not full.

An useful and efficient theory might be the most important prerequisite just with regard to the reproducibility of an effect and the product liability of a SET-device. That's why one mustn't expect a complete list of devices of that kind in this chapter, because in the foreground stand the physical and technical explanations concerning the way of functioning, the understanding for constructive and guiding details and the learning from the mistakes and errors of the inventor.

For a better survey the possible courses of the field lines according to the theory of objectivity are listed in detail (table 15.1) and discussed from the top one after another, starting with the strongest known interaction. To complete examples and concepts are presented.

As is well-known there exist electric (E-) and magnetic (H-) field lines. Further exist open and closed field lines and finally is distinguished between the oscillating and the static case. The results are two to the third power, thus eight possibilities of combination in total. In table 15.1 all eight versions are given, even if one or another case is of more theoretical nature. For the objective of a systematizing of different concepts concerning space energy technology the taking apart in any case is helpful. The figure opposite is survey and structure at the same time for the following chapters.

15.1 The course of the field lines

In chapter 6 a relation between the course of the field of a body and its observable interaction has been made (part 1, chapter 6.7 till 6.9). Here a point of approach is offered. For instance to maximize the force effect a magnet or to optimise an electric motor, the engineers nowadays help themselves with costly programs working according to the method of finite elements. In this way they obtain a picture of the field lines, the course of which makes possible conclusions concerning the production of force or torque. The relation without doubt is given, the only question is in which order of magnitude.

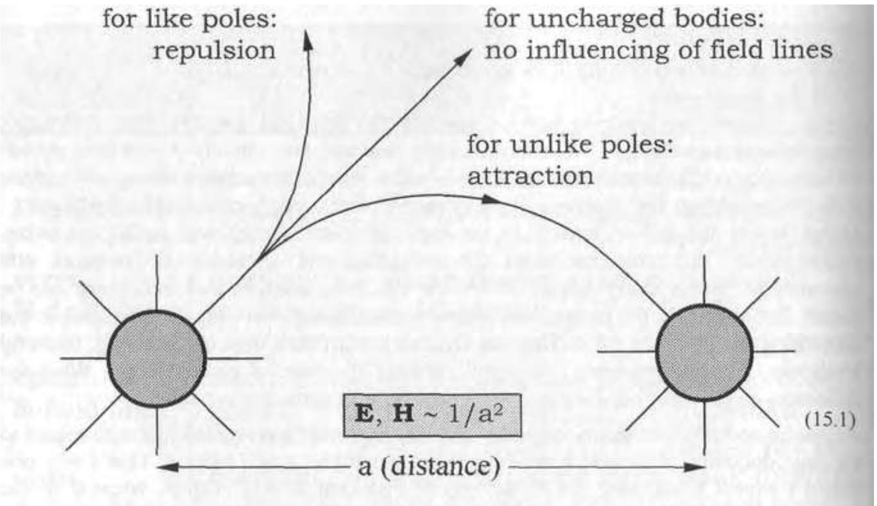
for like poles:
repulsion
for uncharged bodies:
no influencing of field lines
for unlike poles:
attraction
E, H ~ 1/a²
(15.1)
a (distance)

Fig. 15.2 A: The course of the field lines leads to a force effect (repulsion or attraction)

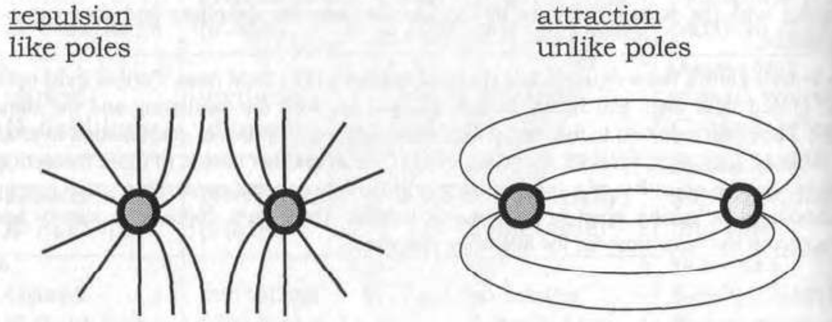
repulsion
like poles
attraction
unlike poles

Fig. 15.2 B: The course of open field lines

$$\boxed{\frac{F_{\text{Coulomb force}}}{F_{\text{magnetic force}}} = \frac{Q \cdot E}{\phi \cdot H} = \frac{\mu \cdot Q^2}{\varepsilon \cdot \phi^2}} \quad \text{with } E = Q/\varepsilon \cdot A \text{ and } H = \phi/\mu \cdot A \tag{15.2}$$

i.e. e^-: where $\phi^2 = m^2 \cdot G \cdot 4\pi \cdot \mu$ (Gl.13.20)

$$\boxed{\frac{F_{el}}{F_{magn}} = \frac{e^2}{\varepsilon \cdot m_e{}^2 \cdot G \cdot 4 \cdot \pi} = \underline{\underline{4{,}2 \cdot 10^{42}}}} \tag{15.3}$$

Fig. 15.2 C: The proportion of forces at the example of the electron.

The theory of objectivity answers the question from the equations of transformation with the proportion 13.23 (fig. 13.4). According to that the electric or the magnetic field strength stand in inverse proportion to the square of a length or of a distance:

$$E, H \sim 1/a^2 \qquad (13.23) \text{ bzw. } (15.1)$$

Less mathematically expressed this, for two bodies in the distance a, where one body is situated in the field of the other, means nothing else as that the distance is reduced. Nearer to the body the density of the field lines again increases, in that way the distance further decreases and we observe an approximation.

Usually the idea of force is introduced as a factor of description and there is spoken of a force of attraction. But that not necessarily is required, because the force only represents an auxiliary description. The cause for the observed attraction rather is the spatial distribution of the field strength.

In this case the two bodies come closer and the mutually active fields get bigger and bigger, until the parts eventually run into each other (fig. 6.7 A). There one comes the thought to increase the force of attraction by an artificial compression of the field lines. In the case of the electromagnetic interaction such a compression actually takes place, since the field lines arise from one pole and end at an unlikely charged pole, which so to speak collects and bundles up the field lines (fig. 6.8 A).

At last we find out the reason, why electromagnetic forces of attraction are bigger than gravitational forces for many powers of ten (between 10^{30} and $4.2\text{-}10^{42}$ for the electron, derivation see fig. 15.2 C)! For table 15.1 this means, at the top have to stand the open field lines, which bundle up at the poles. Then very long nothing comes and after that the effects of closed field lines are being found.

For open field lines however also the opposite of a bundling up is possible. In the case of like poles the fields run away of the other pole (fig. 6.8 B). Between both a space is formed, which is free of field lines, where thus the field tends towards zero, whereas the distance between the poles according to the proportion 15.1 grows towards infinity. In this case we observe, how the bodies are going away from each other. We speak of a force of repulsion, which actually reaches until infinity. This gives reasons for the occurring of both forces of attraction and forces of repulsion in the case of the electromagnetic interaction.

15.2 Gravitation

In the case of closed field lines in principle no repulsion can occur, since no pole, neither north pole nor south pole, neither positive pole nor negative pole is able to influence the position of such a field line. This circumstance as well as the order of magnitude of a possible force of attraction suggest, to settle gravitation here.

It gladly is forgotten, that the field pointers of E- and H-field normally occur together and like in the case of the electromagnetic wave stand perpendicular to each other. It is normal to calculate only the electric field pointers for a charge carrier, without paying attention to the circumstance that the H-field is present as well. The textbooks as a rule remain silent about this dark chapter or they lapidary remark, the dual field lines are closed in themselves and hence inactive anyhow, which however is incorrect.

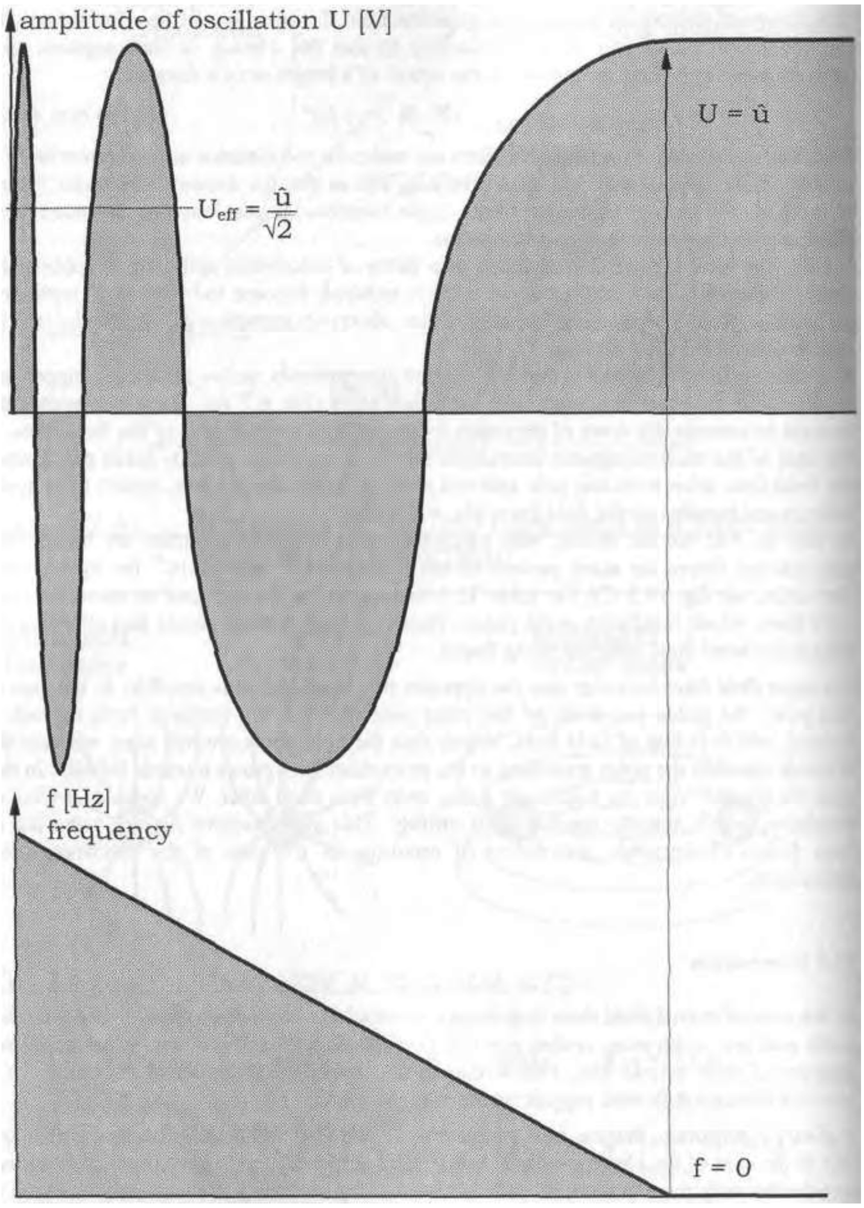

Fig. 15.3: The transition of an oscillating into a static interaction

But, according to the proportion 15.1, they actually develop a force of attraction, even if this is relatively small. No one really needs to be surprised that gravitation is not understood until today, if a whole group of fields simply is overlooked by science! We now also can explain, why there don't exist any massless charge carriers. Namely only the E-field or the H-field can form open field lines and never both at the same time. Otherwise they wouldn't be able to stand perpendicular to each other anymore. The each time other field, in the case of electrically charged bodies it is the H-field, then is wrapped perpendicularly around the E-field lines, independent of the circumstance if electrically an attraction or a repulsion occurs and without exception it forms a weak force of attraction, the gravitation.

15.3 Systematizing the interactions

Next we should know, from which field is to be expected a larger force effect: from the E or from the H-field? At the example of an electromechanical converter this question can be answered concretely.

The forces which occur and form the torque in an electric motor customary in trade are produced by magnetic poles in stator and rotor, which repel each other in the case of like charge and attract each other in the case of unlike charge. Now there in principle exists the possibility to build a motor which works with Coulomb forces, thus with positive and negative poles, instead of the magnetic forces. About such designs numerous patent specifications exist, but no customary version on the market.

The reason very simple is that a magnetic motor for the same torque is many times smaller and better priced. From this the conclusion can be drawn that for the same construction volume a magnetic force F_M is considerably larger than a Coulomb force F_E, which for instance binds together atomic nucleus and atomic hull.

Thus in table 15.1 the magnetic forces are ranked before the Coulomb forces.

Finally we have to distinguish between the static and the oscillating case, which are distinguished in the frequency. In fig. 15.3 an oscillation is shown, which by chance just at the moment of the vertex value changes into the steady state, thus takes the frequency zero. In this case the effective values between a static and a sinusoidal oscillating interaction are distinguished by the factor $\sqrt{2} = 1.4$.

If we operate an universal motor with direct current, then it releases more power, than for a corresponding feeding with alternating current. Even a high-tension line, which stands at maximum 511 kV, is operated with alternating current up to 380 kV, with direct current on the other hand up to 500 kV. Consequently in table 15.1 the static interactions stand before the oscillating interactions.

The strong interaction naturally isn't found in the list, after it has been derived that it doesn't exist at all (see chapter 7.8), whereas the weak interaction is hiding behind the oscillating interaction. It shakes other particles so long till they fall apart (see chapter 7.13).

We now would be as far, to discuss the 8 cases listed in table 15.1 one after another by means of practical examples.

The principle of the side-pole machine

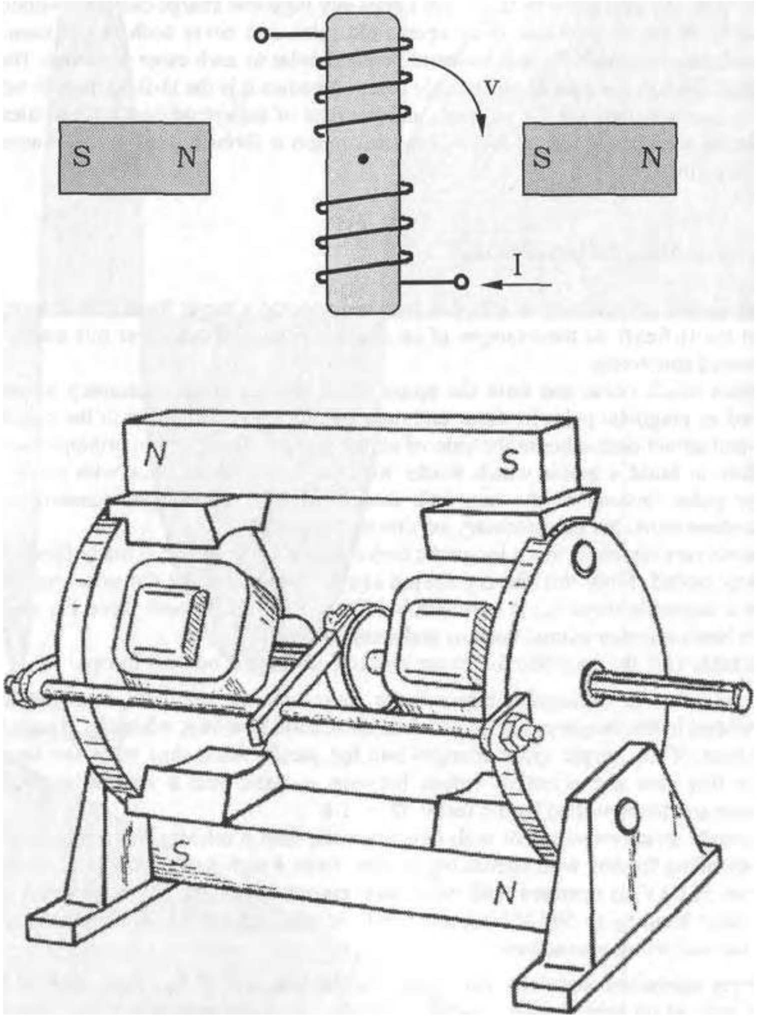

Fig. 15.4: Alternating current dynamo, according to the inventor (Kromrey) a magnetic converter for free energy with a degree of effectiveness of more than 100 %.<i>

<i>: Raymond Kromreys Molekularstromrichter, NET-Journal 6/98, S. 5

15.4 Magnetic force converter

As a result of the systematizing of all eight possible interactions, the largest force effects are to be expected as a result of static and of open magnetic field lines. It is questionable, if this statement is generally valid and is true everywhere in space. But it is valid at least for a terrestrial laboratory and only here, on earth, a solution for the energy problem is strived for. No wonder therefore, if the electric energy technology nearly without exception is using these force effects. We find them at a solenoid, at a relay coil, at a magnetic tuning cylinder and equally between the stator field and rotor field of an electric motor.

The motor however takes a special position, because its rotor is turning. In that way a switching of the winding and commutating of the currents is necessary or the field of the stator winding is being turned, for instance in the case of an alternating current motor. This is necessary, so that in the air split of the motor the fields from the stator and rotor always are standing opposite like for the solenoid and a driving force can be formed. For the operation therefore oscillating currents are necessary, so-called alternating currents, which are fed in into the winding with the right frequency and phase. There can be spoken of an operation in resonance. It surely is no coincidence, that Nikola Tesla, the founder of the rotary field theory and inventor of the alternating current motors at the same time is the discoverer of the neutrino radiation!

The electric motor slides already into the second column from the top in table 15.1, as we see, and should be assigned to the case of the oscillating interaction of open magnetic fields. The frequency for motors usually is very small.

But also at high frequencies there can't be reckoned on some free energy which would show or even be utilizable, as can be heard from the inventors who tinker with magnets, mostly with permanent magnets (Fig. 15.4). The reason very simple is that there exist no physical particles, which could mediate this interaction. Magnetic monopoles would be necessary, thus north pole or south pole particles, so that an interaction with the open H-field lines can occur.
Such particles could form as a result of currents and eddy currents, but for that a good conductivity would be necessary and that isn't present in the vacuum. Therefore magnetic monopoles can't exist at all! This point we already had worked out (fig. 4.4, question I). The same statement then also applies to the oscillating case.

If nevertheless something like free energy should show in the case of some magnetic field converters, then unnoticed by the inventor still other physical effects are added. By means of concrete concepts this circumstance can be studied and discussed. The meanwhile well-known railgun is a corresponding example, for which besides the used magnetic force unintentionally a further principle is used. A more detailed occupation with this device is worthwhile, because here some fundamental concepts of space energy get clear.

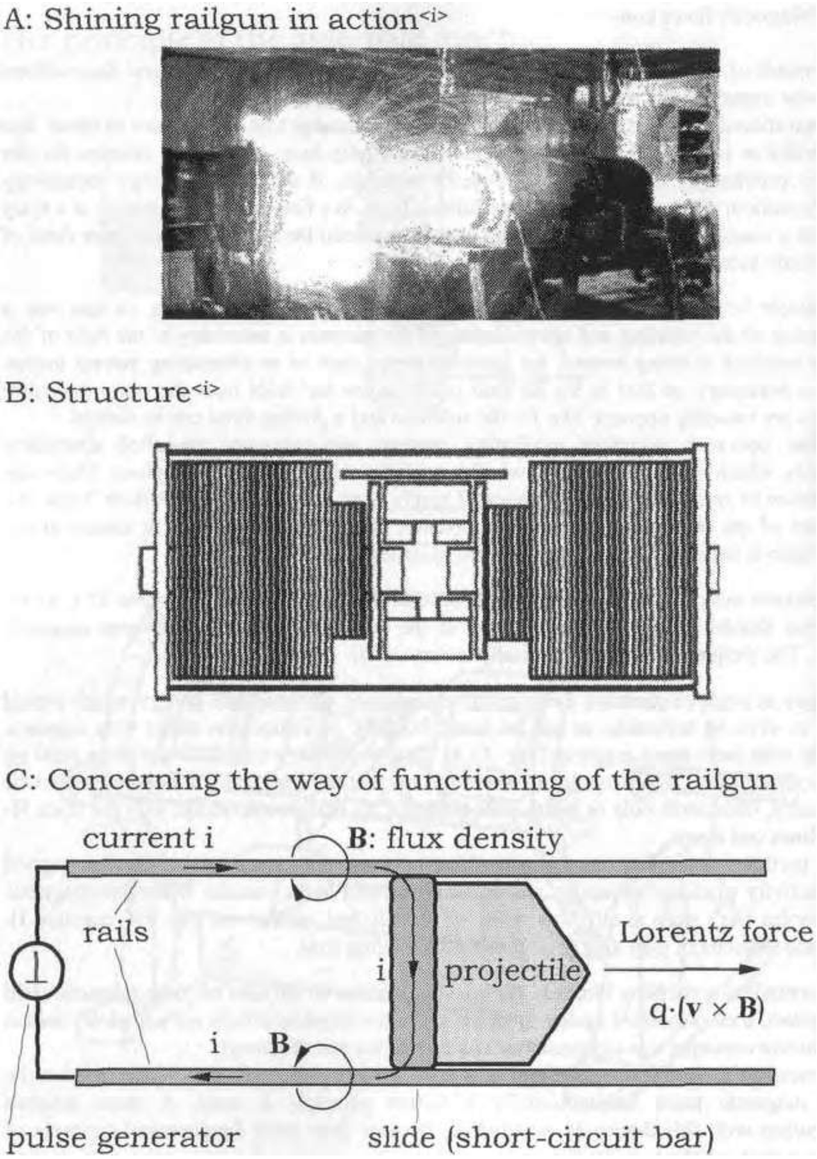

A: Shining railgun in action<i>

B: Structure<i>

C: Concerning the way of functioning of the railgun

current i
B: flux density
rails
Lorentz force
i
projectile
q·(v × B)
i
B
pulse generator
slide (short-circuit bar)

Fig. 15.5: Structure and way of functioning of the railgun

<i>: in the internet under: www.glubco.com/weaponry/railgun.htm

15.5 The railgun

The engineers and physicists involved in the SDI-project were quite astonished, as they had a close look at the bent rails of their gun. During the test operation the equipment was really flying around their heads.
They were very sure to only have fed in 16.7 MJ of energy, from the rotation of a homopolar generator, because more was not available for the experiment by any means. The projectile with mass m_0 = 0.317 kg lying on the rails thereby should have been accelerated to the velocity of 4200 m/s. Instead huge forces were at work here, which the construction couldn't counteract at all. There is talk about the released energy having amounted to 399 GJ, what corresponds to an over-unity effect of 24000[ii]. This factor describes the proportion of the released to the taken up power or energy.
If these details should be correct then this would be the most efficient converter for free energy, which has been developed until now.

Behind the project name SDI (Strategic Defence Initiative) is hiding the by the United Slates prepared "Star wars". But how one fares a war, which nobody can pay anymore and no-one wants, entirely according to the motto: There is a war and no-one goes there? This war undoubtedly the strategists and initiators themselves have lost, who even had to watch, how their space gun appears in the internet with design drawings and rich visual material to be called by anyone[i].
Today, where we are surrounded by nothing but friends, where in Russia and at other potential opponents is fought more against internal problems and one lets the expensive space toy rot for lack of money, the coat of the military secrecy obviously no longer can be held over such an explosive project as the railgun.

Thus informative details have reached the public. In the pictures a bright lightning can be seen at the moment of launching (fig. 15.5 A). Here presumably is being materialized, in which the part of anti-matter annihilates with the particles of matter under emission of light. There thus takes place the same process as in the case of lightning or the shining of the sun.
In addition is being reported that heat energy is withdrawn from the environment, a circumstance, which is typical for all functioning converters for space energy. We thereby are reminded of the possible formation of ice in a lightning channel.

Like for a lightning also the railgun is stimulated with a very high excitation voltage and with extreme speeds of change of the tension voltage (high du/dt) (fig. 15.5 C). From the setup it concerns a bridge of Ampere, which in various respects appears to be superior to the rocket engines, after the costly transport of the propellant into space isn't necessary, since the capacitor batteries can be recharged by solar power.

<i>: e.g.: www.glubco.com/weaponry/railgun.htm

<ii>: Galeczki, G., P. Marquardt: Requiem fur die Relativitat, Verlag Haag + Herchen (1997), S. 139

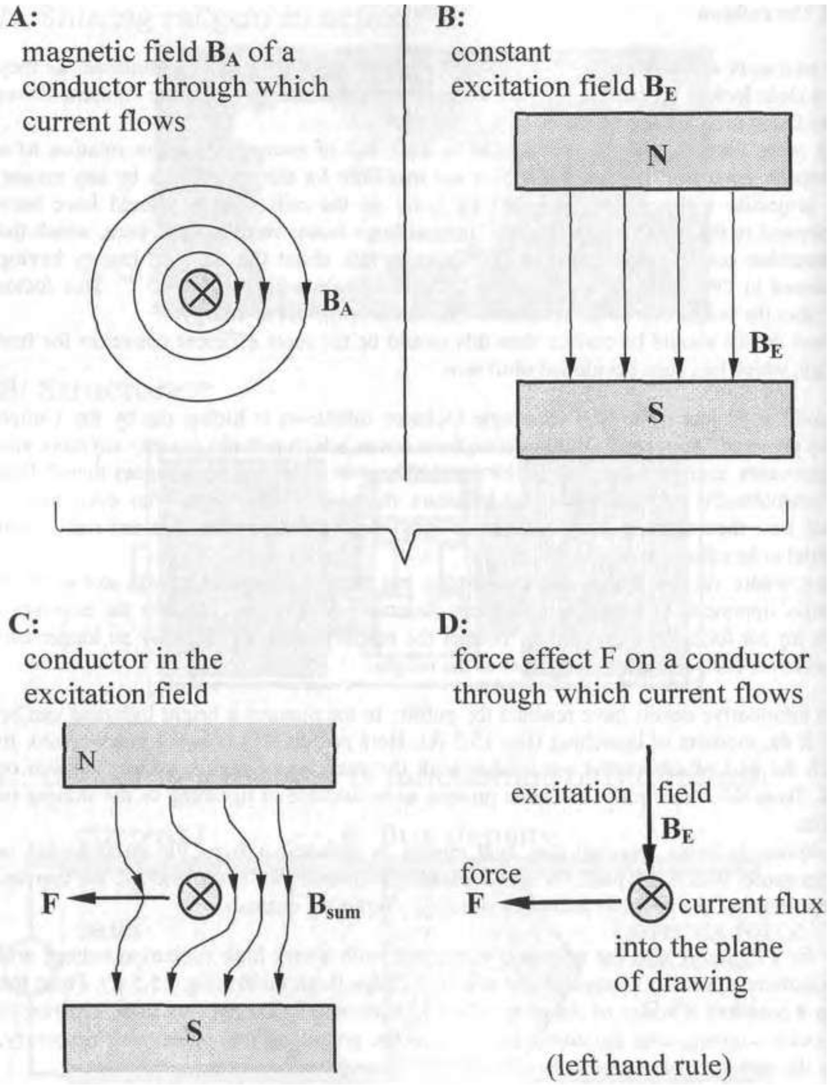
A:
magnetic field B_A of a conductor through which current flows
B_A
B:
constant excitation field B_E
N
B_E
S
C:
conductor in the excitation field
N
F
B_{sum}
S
D:
force effect F on a conductor through which current flows
excitation field
B_E
force
F
current flux
into the plane of drawing
(left hand rule)

Fig. 15.6: ____ The distribution of field lines and force effect on the slider of the railgun through which current flows in a permanent magnetic field

The projectile has the form and the function of a short-circuit bar and is guided slidable between both rails of the railgun. The high-tension capacitors are switched on both rails at the moment of launching, so that in a very short time an extremely high short-circuit current of several thousands of Ampere flows through the bar.
Since the bar in addition is situated in a static magnetic field, there acts an accelerating force on it (fig. 15.6 D). It is the force effect of a conductor through which flows a current in a magnetic field, like it is active in every electric motor. If we overlap the fields of the conductor (fig. 15.6 A) and of the magnetic field (15.6 B), then we observe a bending and lengthening of the field lines (15.6 C). There exists the effort to re-establish the original state, which represents the smallest magnetic resistance, and for that the conductor is shifted out by means of the arising force. In the sketch it is accelerated to the left. That far the explanation concerning the bridge of Ampere. That has nothing to do with free energy. For the enormous degree of effectiveness, as it has been determined, further effects have to be added.

15.6 Unipolar induction

The projectile, or from the function let us rather speak of the short-circuit bar or the slider, at first is entirely conventionally accelerated and experiences, mathematically expressed, a dv/dt. The magnetic field B stretching perpendicular to the movement is constant, so that according to the Faraday relation<i> E = v x B from the velocity v an electric field strength E results and from the acceleration dv/dt a field change dE/dt. These open field lines along the length of the slider, in particular the oscillating part, appears to interact with oscillating particles and to collect these particles<ii>. It concerns presumably neutrinos, which primarily materialize in charge carriers. These contribute to the current flux in the slider and to the acceleration, whereupon still more neutrinos are collected.
A hardly controllable avalanche effect is formed. Only if the change in tension voltage has worn off and the capacitor is completely discharged, also the resonant interaction will again collapse.

The inventors, who want to construct a civil version of the railgun, is given a warning on the way which should be taken seriously. At first it doesn't take particularly much imagination to imagine a rotating arrangement of the gun, a construction with one axis, whit which a generator driven, which produces power. A small part is supplied the system again as supply for itself. The rest would be available free to the consumers as non-polluting, regenerative energy.
That really sounds good, if there wouldn't be this one obstacle.

<i>: The equation of transformation concerning the unipolar induction already was treated more detailed in chapters 6.4 and 9.3

<ii>: It is the resonant interaction according to table 15.1, line four

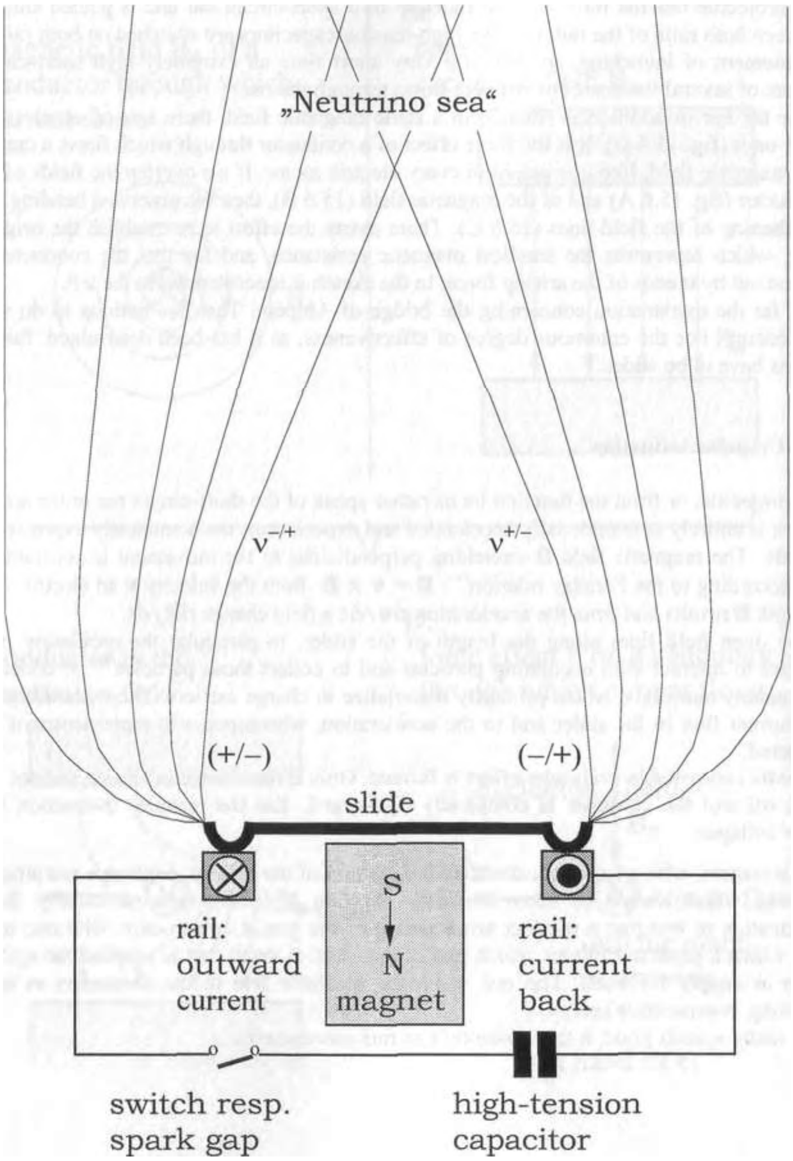
„Neutrino sea“
ν -/+
ν +/–
(+/–)
(–/+)
slide
S
N
magnet
rail: outward current
rail: current back
switch resp. spark gap
high-tension capacitor

Fig. 15.7: The collecting of neutrinos by oscillating open field lines

15.7 Tendency to instability

In a continuously working machine the discharging of the capacitor cannot remain a single event as in the case of the gun. The discharging and the recharging has to take place cyclic depending on the revolutions per minute. To obtain a rotating movement even to a certain extent ignition has to follow ignition. But if the new ignition takes place, although the avalanche effect of the last one still hasn't worn off, then inevitable a catastrophe will occur, then the work of wonder is taken apart under the eyes of its creator. Numerous inventors already have had to collect such painful experiences. It is assumed that not even Nikola Tesla had escaped, as he had to put away again his stately luxury car with electric motor and energy converter in a barn near Buffalo already after one week of test operation in the year 1931[i].

Of course also for this problem solutions in accordance with engineering are offered. Meaningful would be a restriction of the revolutions per minute and a power regulation. Only most inventors don't think that far. On the one hand, because they handicraft without an useful physical model and on the other hand they think they already have reached the goal, if they observe something like free energy for the first time. Just as fast as the joy then the disillusionment comes, because a converter which doesn't work, is not able to convince anyone.

Tesla already was aware of this set of difficulties. He fastened his converter to the dashboard and not in the engine compartment, presumably to adjust the coupling of the coils from the drivers seat during the drive by means of two metallic rods, which he pushed into the case. But sometime even this regulation by hand has to go wrong, because the collected neutrinos on their part collect further neutrinos (fig. 15.7), so that in the case of an unfavourable order of ignition an additional amplification is possible. For a reliable operation according to that directly or indirectly the phase of the ignitions to each other should be checked.

At the example of the railgun space energy technologists and inventors can study the relations and the way of functioning very concretely and even calculate these relatively simple. It is a big relief, that all three vectors stand perpendicular to each other: the E-field, the B-field and the velocity v. Ideal conditions both with regard to a maximizing of the wanted accelerating force and for the resonant interaction, increase at the same time the collecting of space quanta, which probably may be set equal to the neutrinos. This is made possible by the Faraday's law of unipolar induction. In that way at the right and left end of the slider a positive a negative pole each are formed. The further the two poles are away of each other, the more the field lines are opened and the more neutrinos can go into resonance. In this place still considerable improvements and optimisations are possible.

In addition to the two discussed the phenomenon of the electrostriction is added as a third phenomenon, which authoritatively contributes to the conversion of neutrinos into electrons. It is a field dependent change of length, which in the case of lightning takes care of the thunder and in both cases, therefore also here, is active as a charge carrier producer.

---

[i] see also chapter 9.5 Free energy,
A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1989, Kap.ll, S. 20 and H. Nieper: Revolution, MIT-Verlag 1981, S. 194

A: electric dipole (e.g. electrostatics, see also fig. 6.8)

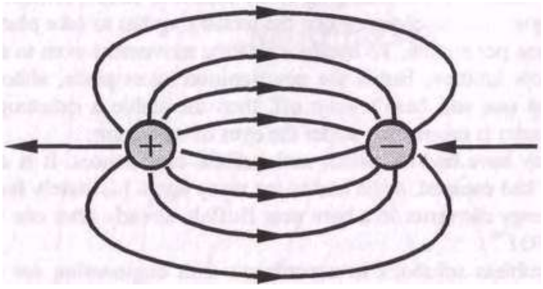

B: electric monopole (e.g. electron e~, see also fig. 4.3)

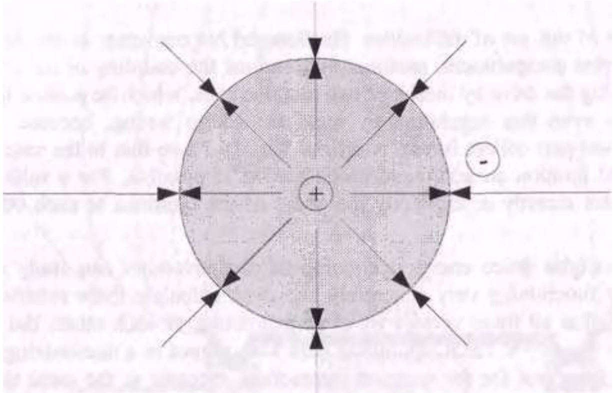

C: Mixed form in the case of cylindrical symmetry

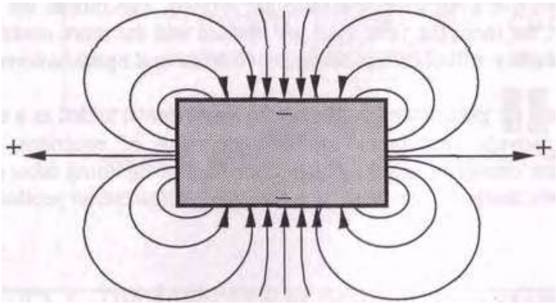

Fig. 15.8: Dipole fields with unipolar parts

15.8 Unipolar field configuration

If we again go back to fig. 5.1 and continue our considerations with examples concerning line 3. The obtainable force effect of open electric fields indeed is for powers of ten smaller than that of magnetic fields, but then particles are mediated, an invaluable advantage and an indispensable prerequisite with regard to the generation of space energy. As long as the particles are considered in the balance sheet of energy then in addition by no means can be claimed, space energy converter are perpetuum mobiles.

In the question, why only electric particles can be mediated, has to be pointed to the repeatedly mentioned circumstance that only electric particles can be formed as a result of the concentration effect of potential vortices. Magnetic particles however plain and simple cannot exist, since for missing conductivity in the vacuum no eddy currents are possible (see fig. 4.4).

In the case of the arising Coulomb forces again is distinguished between the static and the oscillating case. We want to start with electrostatics.

The range of technical applications of static electricity is large. It stretches from varnishing technologies and filter technologies till the fly grill in the arbour. In all cases the field is built up by charge separation. By means of a high-tension generator a positive pole and a negative pole are produced, between which the field is stretching. The field lines now start at one pole and end at the other, unlike pole (fig. 15.8 A).

In this manner almost no open field lines are available, which point to the outside and could interrelate with free particles. If one for instance intends to attract and collect charged particles from the solar wind, from the cosmic radiation or from the electricity of the air, then the design in one point has to be changed fundamentally.

A unipolar field configuration is necessary. What is meant with that, answers a glance at the spherical vortex model of the electron (fig. 15.8 B resp. fig. 6.2). Here it as well concerns a formation of dipoles as a result of the charge separation, but one pole is hiding in the inside of the other pole . In that way its field lines are captured and don't have a chance anymore to come out, to reach the other pole.

But if the pole lying on the outside for its part can't close its field lines any longer, then these point helpless into space and search in their neighbourhood, in the distant world and if need be even in the infinity of the universe an unlike anti-pole, which as a result interacts and is attracted.

The measuring technician analyses these open field lines and falsely calls the construction then a monopole, only because he isn't able to reach the locked up pole. To blame is the unipolar field configuration, which with that probably would be explained to a certain extent.

The designer and inventor as well might have realized how he has to construct his device, with which he generates open field lines to collect space quanta. He has to lock in one of the two poles as good as possible. Optimal would be of course a spherical symmetric construction like in the case of the elementary vortex. Compared with that a cylindrical symmetry indeed is suited far less good, but it offers constructive advantages (fig. 15.8 C).

We now will report of such a functioning device.

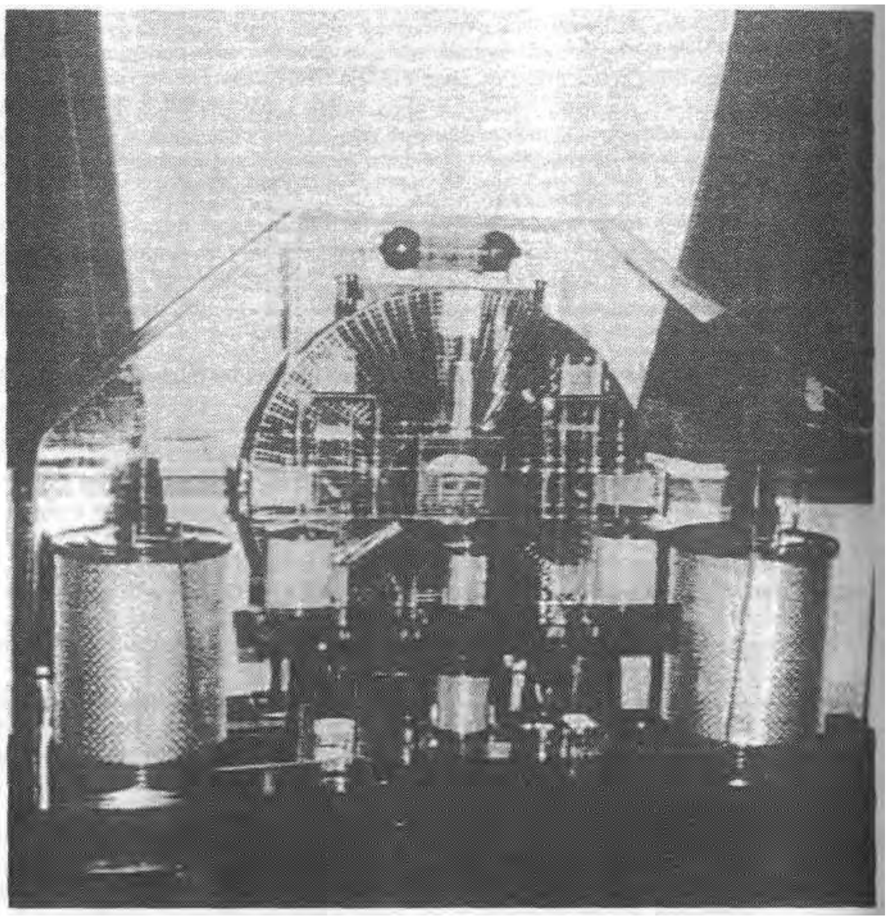

Fig. 15.9: Demonstration converter "Testatika".<i>

<i>: A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1989, S. 29,
I. Schneider: Neue Technologien zur Freien Energie, Jupiter-Verl. 1994, p. 13 and in the NET-Journal, Heft 8/9, 1997, S. 16 as well as iss. 12, 1997, p. 6.
D. Kelly. Der Schweizer ML-Konverter, Raum & Zeit Special 7, S. 164

15.9 The Testatika

An electrostatic device, which produces open electric field lines, is situated in Linden in Switzerland. It optically is very imposing and belongs to a religious community, which has called it Testatika and is of the opinion that it is a free energy converter. Inexplicably the Testatika does not serve the community, which generates their electric power conventionally and to a large extent by itself, as a source of energy. Instead the 2 kW device only very seldom is demonstrated for special occasions or to select groups of visitors. The religious community after that explains the astonished observers humanity not yet is mature for the technology.

Perhaps just the opposite is correct and the technology not yet is mature. According to my personal assessment such an electrostatic device in principle is entirely unsuitable for the continuous operation.

It can be expected that the open field lines sooner or later will interrelate with the electricity of the air and thunderstorms are being attracted by the infernal machine so long until lightning strikes and the demonstration with that has finished. That's why the Testatika may be switched on only for a short time, only at sure weather situation and not too humid air, and many a registered visitor has been sent away without having seen the "thunderstorm machine".

As an object of demonstration and study the Testatika however is well suited. Alone the circumstance that no cable leads to the device and it nevertheless releases energy in the order of magnitude of 1 till 2 kW, surprises all visitors. At least the impression is mediated as if the machine would violate the law of conservation of energy, which is not correct. The Testatika is similar to an induction machine, which works with friction electricity. Thereby the unlikely charged bodies do not have to unconditionally touch and rub at each other, it already is sufficient, if they are brought in the immediate vicinity of each other. In the case of the Testatika the electrostatics of two against one another rotating discs is taken off by brushes.

The excitation energy presumably is taken out of the natural E-field, which just like that can amount to 200 Volts per meter (see chapter 2.9). The large diameter of 80 cm of the discs and their bad conductivity (acrylic glass) permit this conclusion. The charge taken off by the brushes afterwards is temporarily stored in two capacitors of 2 Farad at 300 Volt, so-called Leyden jars. This far one actually is reminded of a Wimhurst generator, in which the energy is supplied the system by turning the disc. Large powers cannot be drawn by that. Plans to build such an induction machine by yourself have been published<i>.

In the case of the Testatika however two discs are used and by hand stimulate to rotate oppositely. This rotational energy in this case isn't used to produce power, otherwise the discs quickly would stand still again, but that doesn't happen.

Until now apparently no-one has discovered the secret, which is kept strict by the members of the community. In my opinion the energy situation on our earth however is too serious, as that we would be able to afford playing hide-and-seek and egoistical secretiveness.

<i>: I. Schneider: Neue Technologien zur Freien Energie, Jupiter-Verl. 1994, S.14

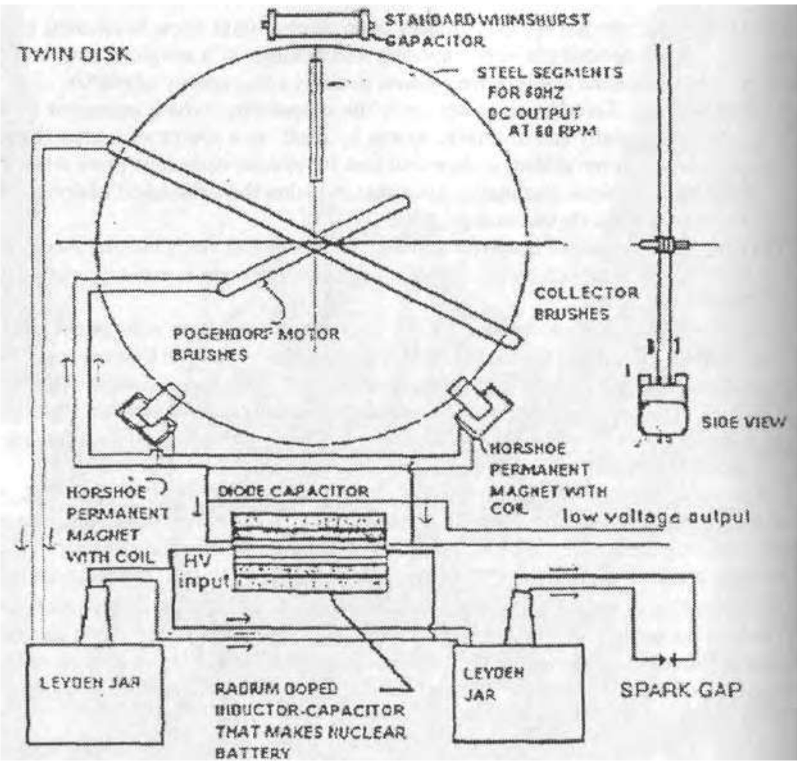
STANDARD WIIMSHURST CAPACITOR
TWIN DISK
STEEL SEGMENTS FOR 60HZ DC OUTPUT AT 60 RPM
COLLECTOR BRUSHES
POGENDORF MOTOR BRUSHES
SIDE VIEW
HORSHOE PERMANENT MAGNET WITH COIL
HORSHOE PERMANENT MAGNET WITH COIL
DIODE CAPACITOR
low voltage output
HV input
LEYDEN JAR
RADIUM DOPED INDUCTOR-CAPACITOR THAT MAKES NUCLEAR BATTERY
LEYDEN JAR
SPARK GAP

Fig. 15.10: Sectional drawing of the Testatika according to Don Kelly, (Clearwater, Florida, USA)

<i>: Der Testatika generator, NET-Journal, Dezember 1997, page 6 similar picture is found in Raum & Zeit Spezial 7, page 164

15.10 The secret of the Testatika

The crucial point is the opposite direction of rotation of both discs. If we assume the static earth electric field is the cause and serves as an excitation field, then as an effect a field arises, which stands perpendicular to that. The axial component now points out of the centre of the disc.

In the case of only one disc the field lines in front and behind the disc again are closed, so that no open lines can form. With one disc or with two discs rotating in the same direction hence no unknown charges can be attracted.

In the case of two oppositely rotating discs however shows one component along the axis of rotation to the observer, that of the other disc exactly in the opposite direction. In that way between both discs a pole is „pinned", which no longer is able to close all field lines on the outside around the machine. Thus open field lines and a, however incomplete, unipolar arrangement are formed.

The charge carriers sucked from the electricity of the air as a result support the natural electrostatics and speedy recharge the capacitors, even if up to 10 Amperes are taken out by the consumers.

The ingenious thing of the machine is its extremely simple construction and the simple concept.

If one includes the collected particles also in the balance sheet of energy, then it thus will turn out that the law of conservation of energy is not violated at all. There thus can't be talked of free energy. In this context the Testatika may rather be given as a counterexample. Air ions are the carriers of the electricity of the air and not carriers of free energy.

Negative air ions are indispensable for our welfare. One should only remember the first men in space, who after the landing were pulled out of their capsule more dead than alive, after they had to stay in the unhealthy atmosphere of the capsule for a longer period of time. Only the installation of ionising devices for negative air ions made possible longer staying in space.

The taking out of the air of negative ions hence is not unproblematic and not particularly ecologically compatible. An atmosphere harmful to life is formed which Dr. Wilhelm Reich has called DOR-state. He by the way has designed a Cloud-Buster, with which he could take static electricity, forming above the desert sand, out of the air. In that way the negatively charged rain clouds no longer are repelled and driven away. Reich has tested his weather machine 1954 in the desert of Arizona. After he had freed the atmosphere of the „DOR-strain", as he expressed himself, in the desert area the atmospheric humidity steadily increased from 15% up to 95%, there grew prairie grass and everything started to turn green, and eventually after many years for the first time rain fell again.

Static electricity, as far as the right polarity is chosen, may conditionally be used for the mechanics of rain making. For free energy concepts it however isn't suitable. Already Nikola Tesla has pointed to the circumstance that our hopes will be in vain if the free energy would be of static nature (see fig. 9.5). He in his speech, which he gave 1891 before the AIEE, has left no doubt that free energy exists, which is kinetic and with that energy technically usable for us,. Chapter 16 will be occupied solely with this case<i>.

<i>: according to point 4 in table 15.1

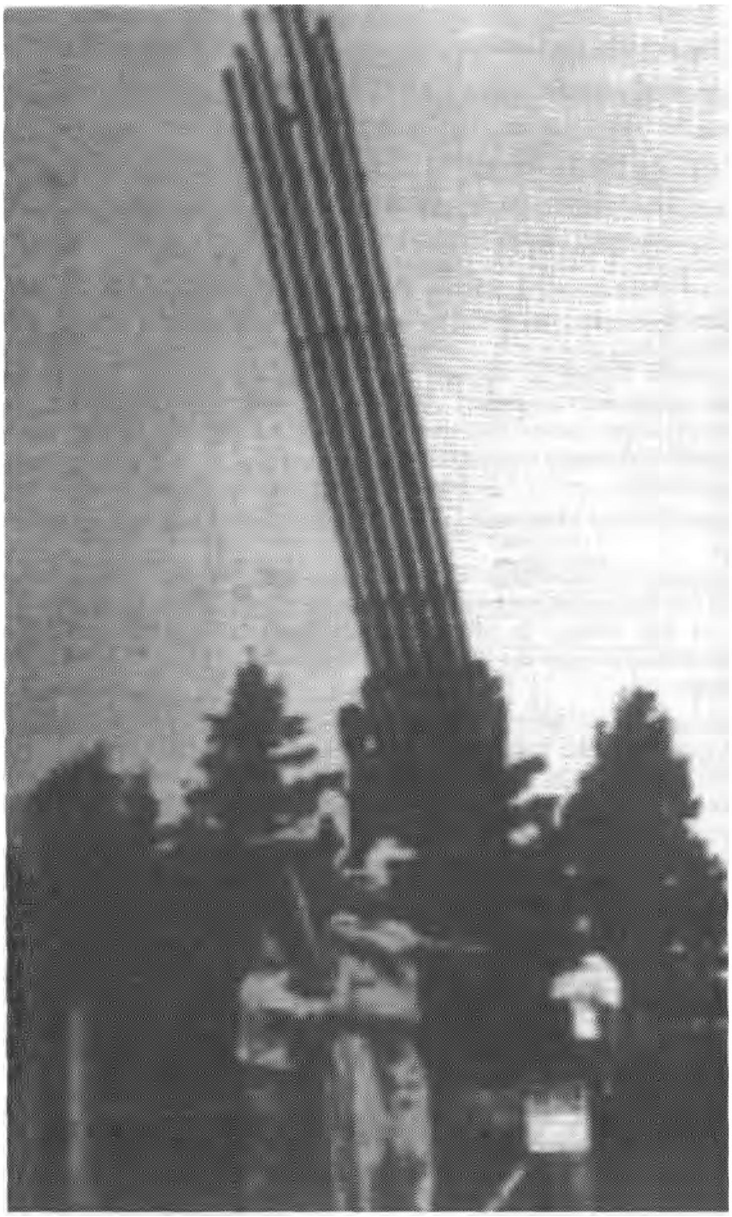

Fig. 15.11: The Cloud-Buster of Dr. Wilhelm Reich 1954.[i]

<i>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, page 122
<ii>: Part 1, fig. 9.3; concerning the unipolar induction see also fig. 6.5 and 11.8

Fig. 15.11 shows the weather machine of Reich, the Cloud-Buster. It can be understood only hard, why Reich directs a 3-4 m long metallic pipe with a diameter of 4 cm to the sky and connects the rear end with a deep well or with flowing waters. The effect should have been increased considerably with a few milligrams of radium. An indication that he must have worked with static electricity, delivers however a tragic accident, in which one of his collaborators was lamed on one side. He carelessly had touched the charged apparatus and suffered an electric shock<i>.

15.11 The key to free energy

As a contribution to the discussion the individual principles of functioning of space energy again are collected and the attempt is undertook to value them.

In the case of an oscillating dipole configuration, for instance the railgun, open field lines are present only along the mutual line of connection (fig. 15.8 A). With that not particularly many space quanta can be reached. It hence has to be operated with gigantic excitation powers in the range of many thousands of Ampere, so that further field lines fling open and interact. The wanted over-unity effect therefore can only be reached at an enormous expense of technical apparatus.

An unipolar arrangement here is considerably more advantageous, where holds: the more unipolar, the fewer excitation power is required. But in that way it can take longer until the collecting of neutrinos like an avalanche again has worn off. In the case of an ideal spherical arrangement (fig. 15.8 B), as the ball-lightning takes, the process can even last for minutes. This explains why unipolar systems can be kept under control only very hard. If the neutrino avalanche is rolling then it purely theoretical only can be stopped with a still larger excitation power, for instance by phase shifting, what can hardly be realized in practice. The rolling avalanche can't be stopped anymore by normal means. A synchronous operation between the neutrino oscillation and the converter can, apart from the technologically hardly realizable high frequency, by no means really be recommended. As a rule one single steep flank of the change of the excitation voltage is sufficient to start the avalanche. By means of the repetition frequency or by means of the duty cycle of the excitation voltage then resonances to the neutrino field can be made or avoided. On the other hand can't be done without the avalanche effect. The utilizable power of the neutrino converter otherwise would be much too small. This case should be pursued further in the design of a longitudinal wave gauge.

All converter systems at first work based on a well-known and tried and tested physical principle of functioning. In the case of the railgun it is the bridge of Ampere. The thus used force effect on a conductor through which flows current is advantageous due to the obtainable order of magnitude and as a basic concept extremely recommendable. But also Coulomb forces or other physical principles can be used.

Despite that a further relation still must be added, which produces the interaction with the neutrinos. Closely associated with the unipolar arrangement it is the unipolar induction, which in virtually all space energy concepts is put to use. It already could be shown that the railgun uses the effect as well as John Searl in the case of his flying disc<ii>. The Faraday law of induction turns out to be the key to free energy.

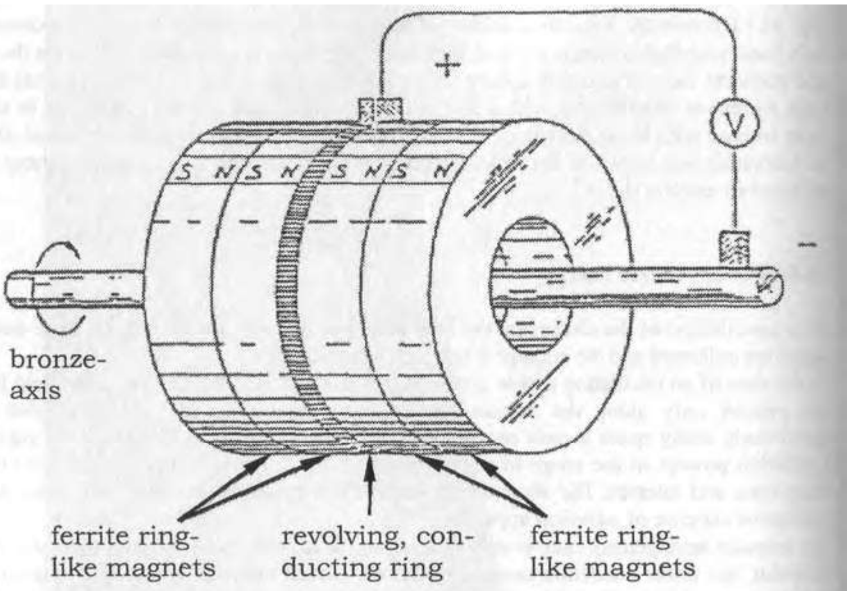
+
V
−
S N/S N S N/S N
bronze-
axis
ferrite ring-
like magnets
revolving, con-
ducting ring
ferrite ring-
like magnets

Fig. 16.1: The Faraday generator, which with rotating magnets is called N-machine.<i>

According to chap. 6.5, eq.60 the equations of transformation are:

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} \quad \text{and} \quad \boxed{\mathbf{H} = -\,\mathbf{v} \times \mathbf{D}} \qquad (16.1)$$

where: $\mathbf{v} = dx/dt = v \cdot \mathbf{e}_x$ = component of the direction of motion perpendicular to the area being stretched by the field pointers $\mathbf{H} = H \cdot \mathbf{e}_y$ and $\mathbf{E} = E \cdot \mathbf{e}_z$.

Thus in addition to the basic fields E and H (at v = 0) additional fields E_z and H_z, depending on motion, occur and the overall fields E_o and H_o are measured (which corresponds to eq. 63 in chapter 6.5):

$$\boxed{E_o = E + E_z} \quad \text{and} \quad \boxed{H_o = H + H_z} \qquad (16.2)$$

(which corresponds to eq. 62 in chapter 6.5):

with: $$\boxed{E_z = v \cdot B} \quad \text{and} \quad \boxed{H_z = -\,v \cdot D} \qquad (16.3)$$

Table 16.2 A: The equations of transformation of the electromagnetic field.<ii>

16. Space energy technology (SET)

It quite concretely concerns the question for a technology concerning the use of the resonant interaction (according to line 4 in fig. 15.1). For that open and at the same time oscillating electric field lines are needed, which mediate neutrinos and pass them on to a receiver working in resonance. The sun, some planets and other celestial bodies, as we already have worked out, use the effect. Even an entire galaxy is kept together in this way. This interaction plays the crucial role for the theme of space energy and the question is asked, with which technology it can be produced artificially.

16.1 The unipolar generator

The most direct way obviously leads over the Faraday relation concerning the unipolar induction. With the classic Faraday generator, where a permanent magnet is turned along its axis, at first a static electric field can be produced. By rotating in opposite direction or magnets rotating in the same direction but oppositely poled, relatively simple one pole can be „pinned" between the magnets and an unipolar construction can be built. Numerous research scientists already have worked in the area of the Faraday machine more or less successfully (fig. 16.1)<i>. There is reported of instabilities and of the picking up of unknown energy at high revolutions per minute. In the majority of the cases it in a sense of the Testatika will concern collected electricity of the air.

The Faraday relation in addition also appears to be hardly understood correct physically by anyone. A scientific magazine in this context takes the opinion: "Faraday proves Einstein wrong", and the production of electricity with Faraday's unipolar inductor violates the laws of physics !<ii>

<i>: A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1989, S. 44

<ii>: „Faraday in his experiment did let rotate a copper disc above a resting cylinder magnet; as expected in a loop of wire a tension voltage was created (F-machine). Than he let the magnets rotate, and the disc stood still; now again a tension voltage should have resulted - but there was no voltage. In the third experiment the magnet rotated with the disc in the same direction and with the same speed. Because there was no relative motion between both, an induced tension voltage wouldn't have been expected - but it was measurable! (N-machine, see fig. 16.1).
What does that mean? If the relative movement between magnet and disc is not always crucial for the formation of an induced tension voltage, then also the absolute movement has to play a role - because something has to move, for a current being formed. But an absolute movement according to the theory of relativity can't be detected - thus Faraday's experiment proves Einstein wrong! Therefore you won't find anything about this experiment in the textbooks".

Taken out of the article: Faraday widerlegt Einstein, PM-Magazin 11/ 1998, P. 133

from table 16.2 A:

$$\boxed{H - H_o \overset{(16.2)}{=} - H_z \overset{(16.3)}{=} v \cdot D} \tag{16.4}$$

with the velocity $v = dx/dt$,
not accelerated $dv/dx = 0$ (inertial system) and derived for x:

$$\boxed{\frac{dH}{dx} - \frac{dH_o}{dx} = \frac{dx}{dt} \cdot \frac{dD}{dx} = \frac{dD}{dt}} \tag{16.5}$$

The curl of the H-field pointer points in the z-direction:

$$\boxed{\text{rot } \mathbf{H} = \frac{dH}{dx} \overset{(16.5)}{=} \frac{dH_o}{dx} + \frac{dD}{dt} \overset{!}{=} \mathbf{j} + \frac{d\mathbf{D}}{dt}} \tag{16.6}$$

comparison of coefficients with Ampère's law:

$$dH_o/dx = \mathbf{j} \quad , \tag{16.7}$$

integrated over dx and ds and formulated generally valid:

$$\boxed{\oint \mathbf{H_o} \, \mathbf{ds} = I_{ein}} \tag{16.8}$$

thus follows from that: magnetic field = vortex field!

Faraday's law of induction (analogous derivation):

$$\boxed{\text{rot } \mathbf{E} = \frac{dE}{dx} = \frac{dE_o}{dx} - \frac{dB}{dt} \overset{!}{=} - \frac{d\mathbf{B}}{dt}} \tag{16.9}$$

cause for measurable electric field E_o is missing:

$$dE_o/dx = 0 \quad , \tag{16.10}$$

integrated over dx and ds and formulated generally valid:

$$\boxed{\oint \mathbf{E_o} \, \mathbf{ds} = 0} \tag{16.11}$$

thus follows from that: the electric field = irrotational (according to Maxwell)!

Table 16.2 B: The derivation of Maxwell's field equations (Ampère's law and Faraday's law of induction) from the equations of transformation of the electromagnetic field

The author of the article proceeds from Maxwell's formulation of Faraday's law of induction, according to which arises a tension voltage in a wire if the wire cuts magnetic force lines, thus is moved relative to a magnet.
If he had read my books, then he would know that Faraday not only has found the older, but also the more comprehensive law, whereas Maxwell only describes a special case. Starting-point for the by me developed theory of objectivity are on the one hand Faraday's law of induction and on the other hand the regularity dual to that, which both together are called equations of transformation of electromagnetism (fig. 6.5, eq. 60 and fig. 16.2 A, eq. 16.1).

16.2 Derivation of Maxwell's field equations

For the derivation we assume, as already in fig. 6.5 for sake of simplicity, that the movement $v = v_x$ takes place perpendicular to the area stretched by the field pointers $H = H_y$ and $E = E_z$. If we derive the equation of transformation, written down for a field component depending on motion, for the coordinate x by using the equations of material and compare the result with Ampere's law resp. Faraday's law of induction, then it becomes clear that the comparison is successful only under certain prerequisites. This circumstance proves that the Maxwell equations only describe a special case and that the equations of transformation are more general valid and causal (table 16.2 A and B).

But in this place is crucial that Faraday's law of induction according to Maxwell only is able to describe a formation of dipoles. For the formation of unipolar field structures however must be fallen back upon Faraday's law concerning the unipolar induction. Whoever wants to understand or even develop by himself concepts concerning space energy, first must have understood Faraday's law in its whole range.

Usually the Faraday generator is, like in the original building shape, equipped with permanent magnets and operated in a steady operating state. Doing so consequently only static electricity is formed. We however need oscillating fields, and for that either the rotation, or the magnet should be changed in polarity with high frequency. Expressed with the precision of mathematics, the first case is described by: $E(t) = v(t) \times B$ and the second case by $E(t) = v \times B(t)$. Both cases have to be investigated and discussed, because both $v(t)$ and $B(t)$ are possible in principle. It is added that in both cases mechanically moving, as a rule rotating, designs but just as well resting designs are conceivable, in which only the moving charge carriers themselves realize the component of velocity v.

This time the crucial point is, that in both cases equally a change of the electric field strength E(t) is produced, with the help of which neutrinos should be collected. Because of the extremely high oscillation frequency of the neutrinos large field changes dE(t)/dt and associated with that large changes of tension voltage dU/dt seem to be the optimal solution, which can be handled with today's technology. Apart from that further difficulties are added, which require a managing in accordance with engineering. If we namely work with a large acceleration dv(t)/dt, then the inertia of the accelerated masses should be overcome, then only very small and light projectiles can be launched like in the case of the railgun. If we however work with fast changes of the magnetic field dB(t)/dt, then the inductance acts slowing down.

A: Suggestion from the patent specification of John W. Ecklin

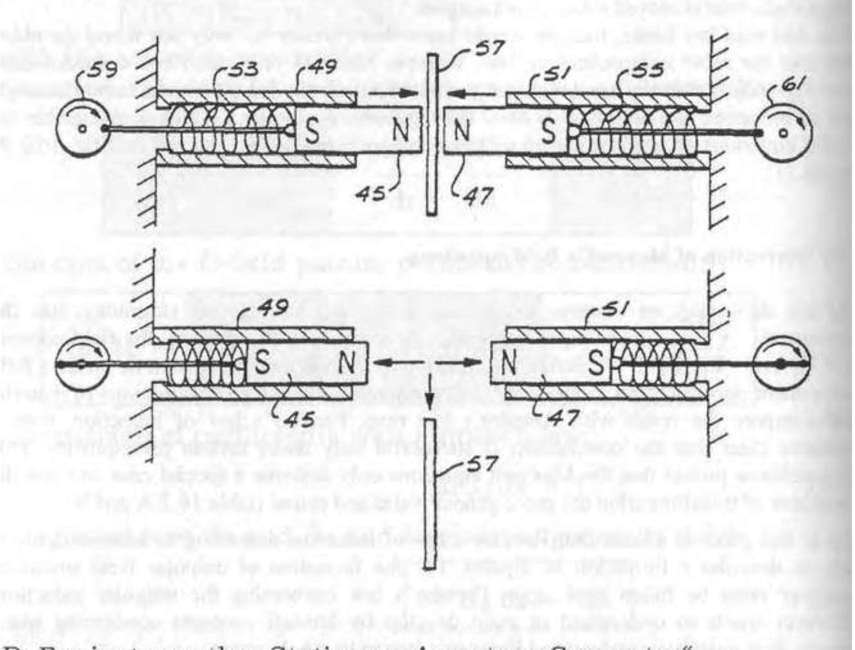
59
53
49
57
51
55
61
S
N
N
S
45
47
49
51
S
N
N
S
45
47
57

B: For instance the „Stationary Armature Generator"

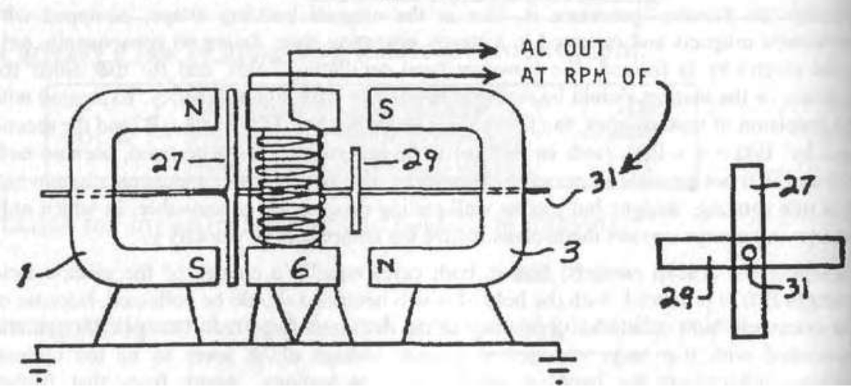
AC OUT
AT RPM OF
N
S
27
29
31
1
S
6
N
3
27
29
31

Fig. 16.3: Principle of the Ecklin generator.<i>

<i>: A. Schneider: Generatoren mit Ferritkernumpolung, NET-Journal 6/98, S. 9
Ecklin: Permanent Magnet Motion Conversion Device, U.S.Pat. 3'879,622

16.3 SET-devices with rotating magnetic fields

If we again come to speak of the Faraday generator, which can be built up in two variants. In the case of the F-machine the magnets rest and only the disc rotates, whereas in the case of the N-machine the magnets rotate along. In this case can't be avoided that some of the induced charge carriers roam about in the magnets and sensitively disturb the structure of the material. One only thinks of the bending spoon effect!

The consequences stretch from a loss of the permanent magnetism, over a shattering and bursting up to a pulverizing of the magnets. Adventurous rumours and alarming reports in this direction are sufficiently available.

An improvement could be obtained by an isolation layer between the conductive disc and the magnets, but against induced charge carriers inside the rotating magnets this measure is not able to achieve anything.

Anyhow the best thing to do will be to completely do without permanent magnets in neutrino converters! If we replace them at least in our minds by electromagnets, then a feeding with alternating voltage is possible. That also is necessary if oscillating neutrinos and not air ions should be collected.

As a result of the alternating voltage at first large eddy currents occur in the disc. The losses can be reduced by radial slits in the disc. The induced currents then only can, as wanted. flow radially to the outside.

Bigger headaches causes us on the other hand the inductive and with that current storing effect of the excitation coils. If a too large excitation inductance should prevent a fast increase in current, then also the induced electric field will increase correspondingly comfortable and hardly be able to persuade a single neutrino to stay.

A coil core of iron or dynamo sheet metal with that is ruled out from the start. Even ferrite would be suitable at most conditionally. Usable are air coils with as possible as few turns.

In an advantageous design the slit copper disc rotates between two air coils, if need be flat coils built in the way Tesla did, which are fed with pulsed tension voltage.

The highest speed of change in current surely is obtainable by means of a spark gap, like already Tesla has used (fig. 9.1). But also other techniques are thinkable as pulse driving.

For instance semiconductor power amplifiers with MOS transistors not only can be switched fast and hard, but in addition frequency and duty cycle can be adjusted freely with reproducible exactness. These are niceties, which will gain importance in connection with the control and regulation of a converter.

Und the heading "generators with magnetic flux variation" Mr Adolf Schneider has collected and commented on some concepts[i]. The generator of the american research scientist John W. Ecklin registered for patent at 22-4-1975 thus stands as an example of a whole group of inventions, in which the magnetic conductivity in the magnetic circuit and with that the flux is changed with a jump.

Fig. 16.3 B shows a building form consisting of two horseshoe magnets (1, 3) with a likewise resting coil in between (6). In rest the magnetic circuit experiences no change and consequently no tension voltage is induced in the coil. The trick is that an axis (31) is pinned right through the arrangement which is turned with two soft iron anchors, which magnetically short-circuit alternating the left (27) and the right (29) horseshoe magnets.

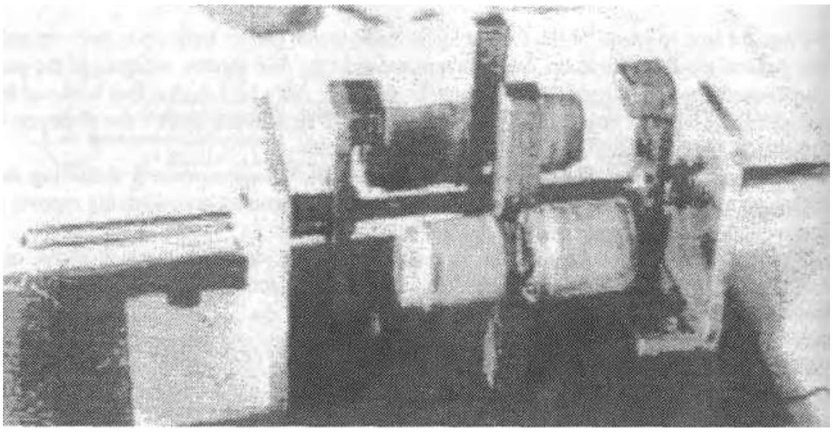

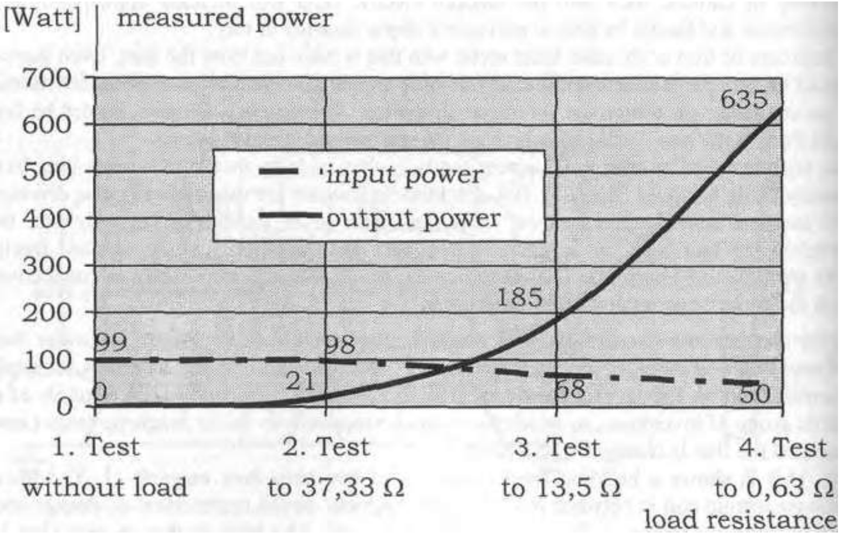

Fig. 16.4: Bedini magnetic converter.<i>

<i>: A. Schneider: Generatoren mit Ferritkernumpolung, NET-Journal 6/98, S.10

Besides the moment of reluctance, which should be compensated by the physical principle, and the friction of the bearing actually no further moment of reaction can occur which would have to be gotten over, not even then if current is taken out of the coil. If the flux change is large enough, purely arithmetically it should be possible to take off an electric power which is considerably larger than the friction power. John Bedini has 1985 measured an over-unity effect of up to 12.6 at a similar constructed generator, where with increasing strain also the factor could be increased (fig. 16.4).

16.4 Commentary concerning magnetic SET-devices

I judge these measurements rather sceptical, after I already in three cases had to break the message gently to the inventors, that their device unfortunately was nothing but an energy destroying machine and they merely had measured wrong, where admittedly the measuring of pulsed tension voltages and currents is not quite simple. So that you don't become a victim of wrong hopes and self-deception, I recommend all SET inventors to realize the closed-loop. If in the continuous operation power can be taken out of such an arrangement without supply of energy from the outside, and be it as small as possible, then that convinces everyone even any journalist and any non-expert.
Unfortunately in the case of numerous concepts at this place already the end is reached. As charming the variation of the magnetic circuit may be, in most cases I miss the unipolar arrangement of the fields. Perhaps here no neutrinos are needed at all and energy merely is withdrawn from the environment heat? I'm not able to answer this question and I surely don't need to, as long as no magnetic converter is demonstrated to me as closed-loop. such a converter, if it can be realized, presumably will at least partly hide one of its poles and produce some open field lines; that at least would be expected.

a still bigger measuring technical problem represent the neutrinos bound to a line, which oscillate around the conductor in the form of ring-like vortices. We have become acquainted with these in the case of the single-wire transmission technique of Tesla (fig. 9.5). They are formed, if neutrinos are slowed down and collected, but not yet have materialized to charge carriers. Tesla did use them for his loss-free energy transmission technology, but he couldn't supersede the alternating current technology full of losses, which also stemmed from him but which he called the worse technology, from the market because there were no power meters available for the single-wire-technology.

Today we still aren't one step further. The energy supply enterprises still decline this technology, as I had to learn myself, although this would be the only way to transport solar energy from the desert or energy from the geothermal energy of Iceland by a sea cable to Central Europe, where it is needed.
Today still no gauges exist for such neutrinos bound to a line. Therefore will every measuring technician experience his Waterloo at SET-devices, in which they occur! I now will report of such a converter and the odd measuring problems.

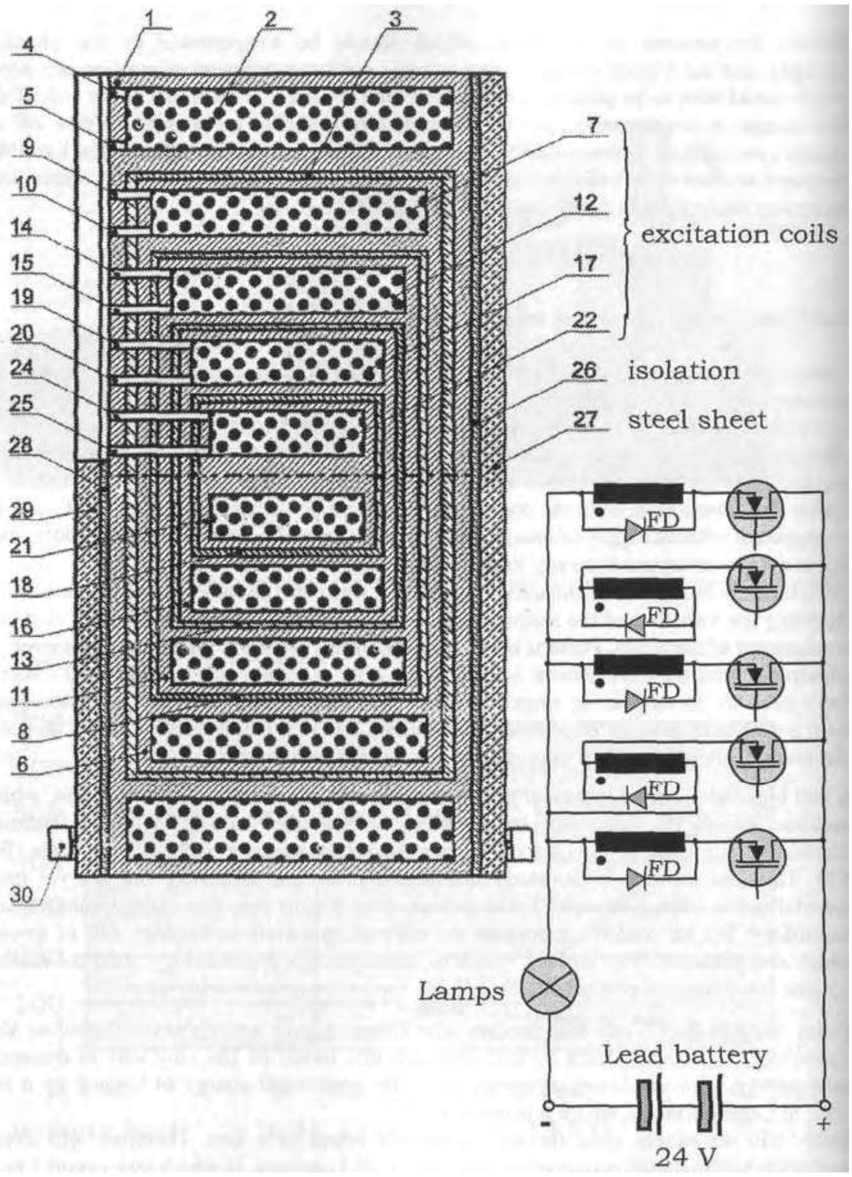
1
2
3
4
5
9
10
14
15
19
20
24
25
28
29
21
18
16
13
11
8
6
30
7
12
17
22
excitation coils
26 isolation
27 steel sheet
FD
FD
FD
FD
FD
Lamps
Lead battery
-
+
24 V

Fig. 16.5: Space quanta manipulator of the Firma ROM AG<i>

<i>: Gibas, Greilinger, Lehner, Rusterholz: Strom aus dem All, Mega Link, Fachzeitschrift fur Elektronik, Nr. 6, Marz 1998, S. 18-23 resp. Bau: Glaube kann Berge versetzen, Bulletin SEV/VSE 25/97, S. 31

16.5 RQM and the space quanta manipulator

If an employee or one of the numerous shareholders of the Swiss Firma RQM AG in Rapperswil speaks of space quanta, then he with that presumably means the neutrinos. An oscillating source of neutrinos is called central Space Oscillator and an operation in resonance of the energy receiver, the so-called space quanta manipulator, is required.
Mr Ludwig Sigrist, the creator of this world of imagination, was not a physicist but crane operator (pseudonym Crane O.) and inventor, whom textbook physics couldn't help further in his considerations anyhow. His concept, if it can be translated into a scientific comprehensible language or not, at least helped him personally and gave him the position to create the space quanta manipulator (fig. 16.5).

It consists of several pot coil systems, which are build up and boxed into each other according to the Matrjoschka principle of scaling down. The ferromagnetic core material of the pot spheres (1, 6, 11, 16, 21) should show magnetostrictive properties as distinct as possible. As we will see, this measure gains its actual importance first in connexion with the interaction with neutrinos!

Each pot sphere carries an excitation winding (2, 7, 12, 17, 22), through which alternatively flows a current in opposite direction. By means of this measure, which seems useless according to classic design concepts, one of both field poles is pinned in the centre, in which way the necessary open field lines are produced.

The individual pot spheres are insulated from each other, where the isolation layers (3, 8, 13, 18, 26, 28) should have a high dielectricity. The pot spheres thus in addition form capacitors with each other; even the ending plates at both sides (27, 29) function as capacitor plates.
This construction, the core piece of the planned RQM converter, is driven by a transistorised power amplifier with excitation impulses as steeply flanked as possible. According to patent specification<i> the pulse widths are freely eligible in steps of 5 ns betweeen 200 nanoseconds and one second. Besides the frequency also the polarity and DC voltage offset are adjustable.

Every visit of the laboratory leaves a lasting impression: Carefully the engineers grope forward, turn at the frequency and the duty cycle, until the effect suddenly occurs and neutrinos, the space quanta, are being collected. One realizes that tuning parameters with still stronger reactions would be possible, but entirely without any regulation and limitation of power that can't be controlled anymore. The avalanche effect would destroy everything, and so one is further dangling along the brink of the abyss in the development laboratory of the RQM, all the time trying hard to gain control of the hardly understood effect.
Concerning the setup and the mode of operation of the pot system the inventor did let his posterity have explanations, but why a diode in free operation FD suddenly loses its rectifying function, for that neither the employees nor specialist visitors of as high as possible scientific rank in the laboratory until now had ready an answer.

<i>: Gibas, Lehner, Greilinger: Vorrichtung und Verfahren zur Erzeugung elektromagnetischer Pulse, Patentschrift CH 687 428 A5 vom 7.5.1996

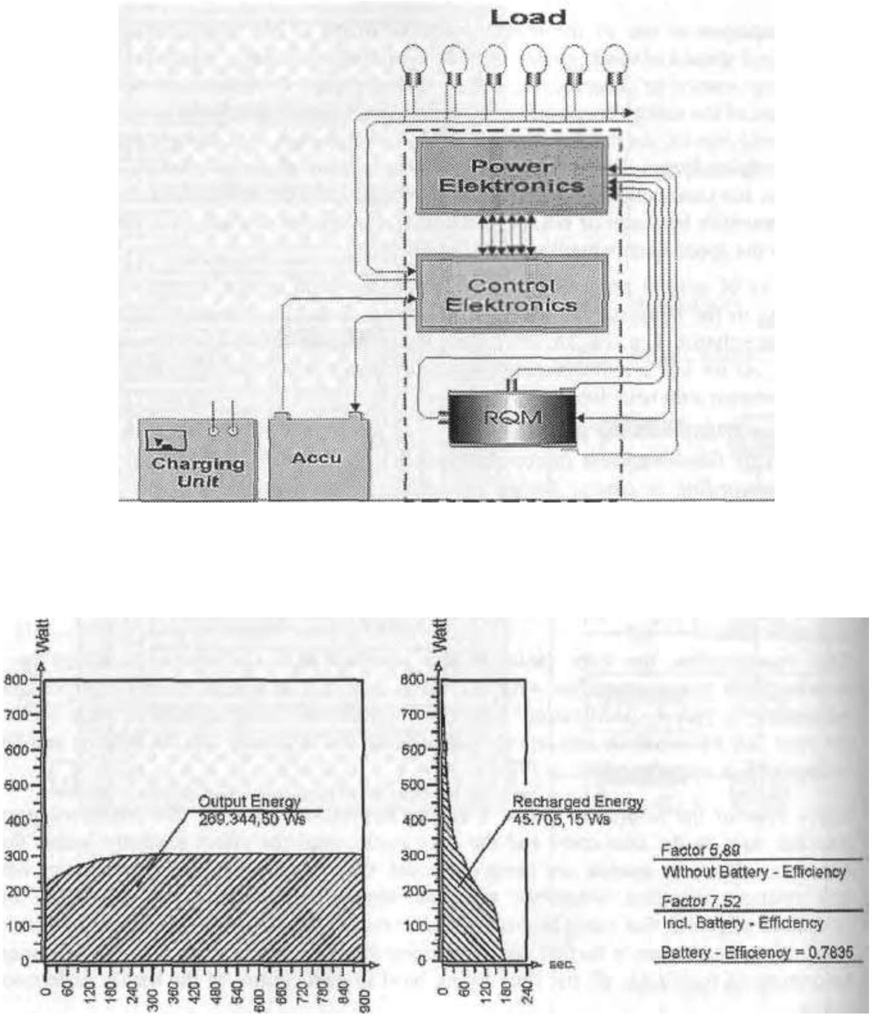
Load
Power Elektronics
Control Elektronics
RQM
Charging Unit
Accu
Watt
800
700
600
500
400
300
200
100
0
Output Energy
269.344,50 Ws
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
Watt
800
700
600
500
400
300
200
100
0
Recharged Energy
45.705,15 Ws
sec.
0
60
120
180
240
Factor 5,89
Without Battery - Efficiency
Factor 7,52
Incl. Battery - Efficiency
Battery - Efficiency = 0,7835

Fig. 16.6: The RQM test installation of 23-9-1996<i>

<i>: F. Greilinger: Der Weg zur erfolgreichen, stabilen Energieauskopplung, RQF Magnetik, Sonderausgabe 1996, page 10

As said, the neutrinos no longer are free but bound to a line, if they leave the pot system. They oscillate around the wire, even around every semiconductor and rove through the entire switchboard. Doing so they can cause quite something, also bring about quite some disaster.

Lead batteries are being recharged, which without doubt is desirable, but in the continuous operation they are destroyed by the same vagabonds. In normal light bulbs for instance some of them materialize to charge carriers, so that according to a gauge more current leaves the light bulb, as on the other side flows in to it. It indeed glows completely normal, but measuring technical for the installation nothing is normal anymore.

The visitor can be shown by means of a high precision measuring facility of vibrations that the pot coils oscillate not only electrically and magnetically, but also mechanically. But the effects alone don't make a converter, which can be produced and which await already numerous buyers of options and licenses.

On the one hand measures for the purposeful conversion of the roving space quanta in utilizable charge carriers here still are missing. On the other hand any controlling facility is missing, to adapt the power taken up from the neutrino field to the momentary need of the consumers. The brave engineers of the RQM still have quite some way in front of them. but the line of approach is right and the reached can be looked at.

16.6 SET-devices with pulsed magnetic fields

In the case of the space quanta manipulator mechanically nothing is moving. In the copper coils the charge carriers merely are on the way with the velocity v. We here have present a typical example of a SET-system with pulsed magnetic field. Compared to the before discussed SET-devices with rotating magnets, for instance the N-machine, the space quanta manipulator clearly has its nose in front. Without brushes, without friction and wear and tear it theoretically has a unlimited life.

In addition it is simpler to produce the necessary large steepness of the flanks in an electronic way as by a mechanical variation of the magnetic field. By means of electronically driving the process it also can be checked, controlled and regulated easier.

The question for an optimisation of the concept still remains. At present one already can be satisfied with a study of possibility, but sometimes one will question the design. There as an example the eddy current losses in the iron pots will be at discussion. About sheeted or sintered materials could be thought, if not at the same time the inductance of the coil would increase in that way, which slows down the increase in current. Ferrite materials again are very brittle and would crumble to dust under the mechanical oscillations of size. I proceed from the assumption, that also here an arrangement with air coils could turn out to be an optimum. It does make sense, if Tesla at higher frequencies always did experiment with air coils.

Now it still depends on the coiling technique. One single conductor loop doesn't provide any open field line. For this purpose if need be two loops have to be supplied with current in opposite direction, like it is the case for a so-called Mobius winding. In this way possibly just as many neutrinos interact as in the case of the pot coils of the space quanta manipulator, which are alternatively supplied with oppositely phased current. To clarify the situation we now should occupy us with the coiling technique.

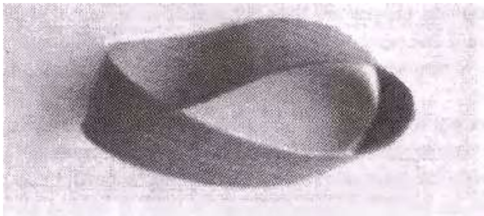

Fig. 16.7 A: The Mobius band

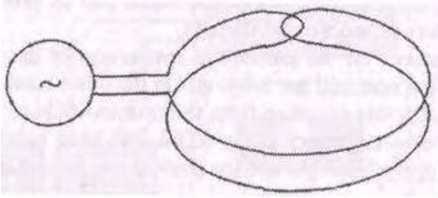

Fig. 16.7 B: The Mobius strip

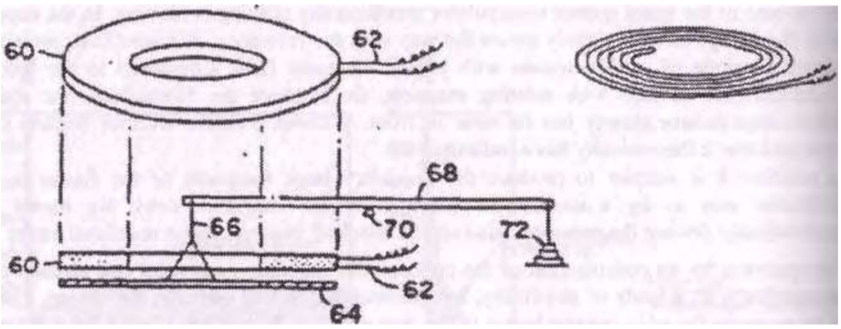
60
62
68
66
70
72
60
62
64

Fig. 16.7 C: Bifilar wound flat coil

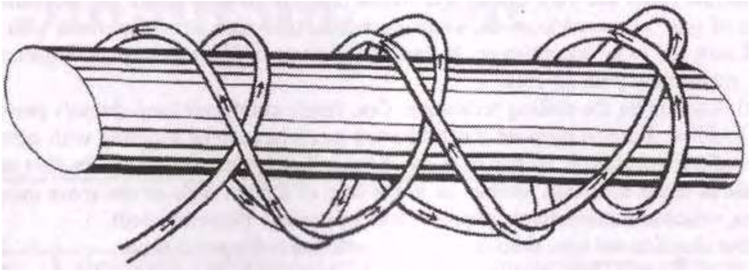

Fig. 16.7 D: Mobius winding arranged like a cross

16.7 SET-devices with Mobius winding

In connection with space energy devices often is talked about the use of a Mobius winding. That is traced back to the Mobius band, which one obtains, if one for example one end of a long paper strip after a half turn glues together with the other end. The result is a strip, which has neither top side nor underside and neither right nor left border. This object with only one side and only one border is a creation of the German mathematician August Ferdinand Mobius (1790 - 1868), fig. 16.7 A.

A distant relative of the Mobius band is the strip with the same name, where the magnetic field lines mutually cancel out. It is the same compensation, as we know of a two core electric cable, where the supply and return cable are run close together. For that the sense of winiding of a conductor loop simply is reversed (fig. 16.7 B).
If the magnetic field vector is wrapped right-handed around the supply cable and left-handed around the return cable, then both amount to zero, so that measuring technical no rest field can be detected anymore at some distance.
THE pointer of the electric field stands perpendicular to the magnetic field vector and points in the direction of the conductor and the movement of the charge carriers. From the coupling of magnetic and electric field follows, that the compensation of one of them also leads to the compensation of the other one. But if the electric field actually is compensated, then there may not flow any current in the winding!
There however flows a current if an electric tension voltage is supplied to the Mobius winding. As a result of this forced current flow both electric and magnetic fields have to occur, which are not compensated! Some field lines will fold outward the expected direction and stand in space as open field lines. These we, in the oscillating case, owe the interaction with the neutrinos.

A perfect compensation would be expected for a two core, bifilar winding. The American physicist William Hooper was able to obtain interesting effects with a bifilar wound flat coil. With help of a balance he could detect attracting or repelling force effects on different, electrically or magnetically neutral discs (fig. 16.7 C)<i>.
But he feeds the flat coil with direct current, so that a contribution of neutrinos can't be expected. Such effects only will occur, if one works with pulsed signals, because the bifilar wound flat coil quite obviously is able to form wanted, open field lines.

<i>: Schneider, Adolf: Energien aus dem Kosmos, Jupiter-Verlag 1994, page 28, taken out of:

Mielordt, Sven: Tachyonenenergie, Hyperenergie, Antigravitation, kaum & Zeit Verlag 1984

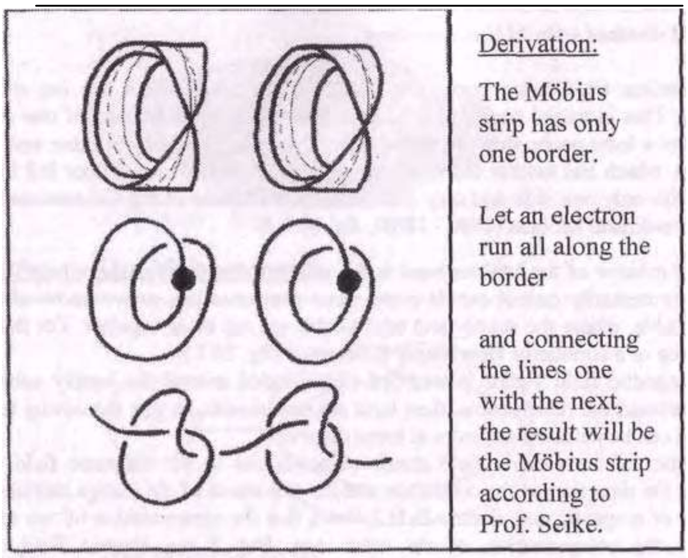

Derivation:

The Möbius strip has only one border.

Let an electron run all along the border

and connecting the lines one with the next, the result will be the Möbius strip according to Prof. Seike.

（第93図）

Fig. 16.8 A: Derivation of the Mobius strip acc. to Seike<i>

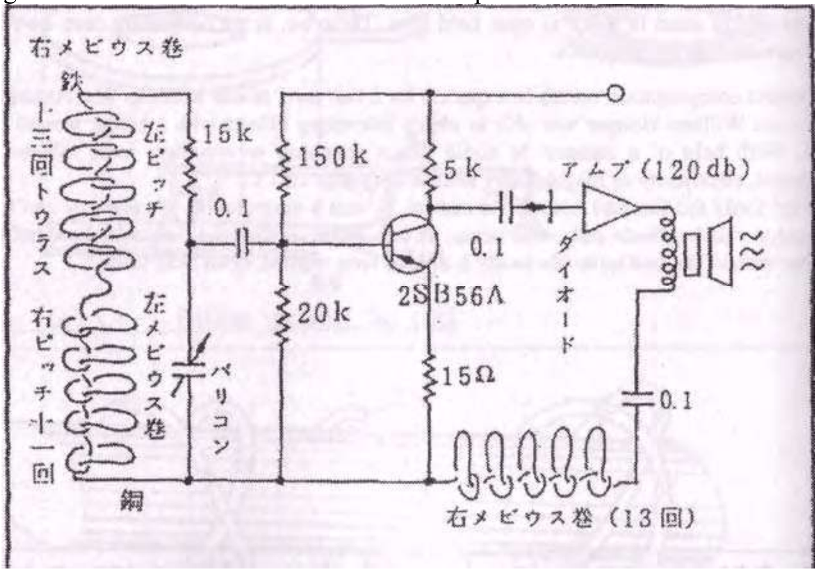
右メビウス巻
鉄
十三回トウラス
左ピッチ
15k
150k
0.1
20k
左メビウス巻
バリコン
右ピッチ十一回
銅
5k
アムプ（120db）
0.1
ダイオード
2SB56A
15Ω
0.1
右メビウス巻（13回）

Fig. 16.8 B: Scalar wave radio according to Prof. Seike<i>

<i>: Seike, Shinchi: The Principles of Ultra Relativity, Space Research Institute 1990, ISBN 915517-1

16.8 Mobius converter of Seike and Coler

Concerning construction it can be an advantage to wind a Mobius coil not purely bifilar and in that way to do without a perfect compensation. It often already is sufficient, if two conductors cross and only individual components of the field vectors partly cancel out. In the case of the coil, which is pulled apart and wound like a cross, drawn in fig. 16.7 D the conductor current and the fields of supply and return cable belonging to it stand under an angle of almost 90°. This in individual cases should already be able to cause the formation of open field lines.

Similar field conditions are formed, if after every turn the wire is looped under the last winding. At the knots again the angle conditions of approx. 90° occur. The Japanese professor Shinichi Seike preferably works with this kind of winding, which he directly derives from the Mobius band (fig. 16.8 A)<i>. He has designed an electro-gravitation motor basing on this principle. He also speaks of weight reduction and of an artificial anti-gravity field. It could be confirmed experimentally, that his setup cooled down slowly during the operational tests, despite the expected heating up by the copper losses of the current<ii, S.29>. Unfortunately I don't known anything more exact.
In his book further is found the wiring diagram of a radio for the reception of scalar waves (fig. 16.8 B). He thereby quite simply replaces all coils in the high-frequency part with his Mobius coils. If longitudinal waves already are measurable and receivable with this simple measure, first has to be checked.

The German captain Coler more than 50 years ago has developed another Mobius converter. His "Magnetstromapparat", with approximately 6 kW power and an at least four-fold over-unity effect uses six coils with permanent magnetic core. As a peculiarity he in addition runs the coil current through the core under an angle of 90° with regard to the coil current (fig. 16.8 C).

More detailed details can be seen in the footnote<ii, S. 22>.

<ii>: Adolf Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1994

S.22: "Scientists of the Technical University Charlottenburg in Berlin and Munich confirmed that the device functioned without objections, but they didn't find a theoretical explanation for the production of energy. The professors Kloss and Franke of the Technical University of Berlin found a degree of effectiveness of 450%. Prof. Schumann confirmed that 4.8 to 6.7 times more energy came out, than was put into the device. Prof. Schumann excluded a deception entirely as the records prove.
But it wasn't simple to set the device going and the stability for longer periods of time wasn't guaranteed. The war confusion set an end to further research. After the end of the war the British secret service confiscated all available documents and apparatus. Part of the report was declassified in 1962".

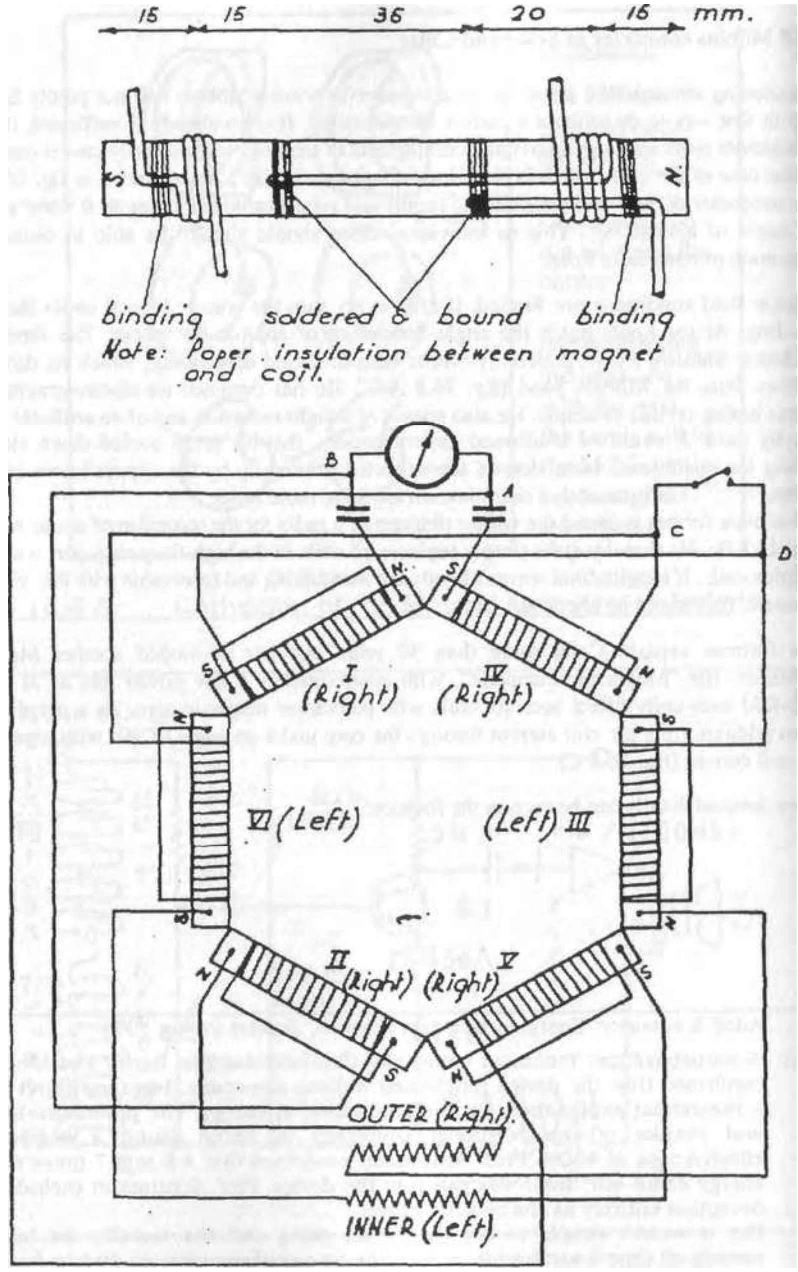
15 15 35 20 15 mm.
K
N
binding
soldered & bound
binding
Note: Paper insulation between magnet and coil.
B
A
C
D
N
S
S
N
I (Right)
IV (Right)
VI (Left)
(Left) III
II (Right)
V (Right)
N
S
S
N
OUTER (Right)
INNER (Left)

Fig. 16.8 C: The Coler converter

16.9 Tesla's flat coil

In the category of the unconventional coiling techniques without doubt the Tesla coil may not be missing. If in schools and high schools such coils were standing in the laboratory for teaching purposes, then as a rule it are cylindrical coils. In reality Tesla worked with flat coils but that, so is said, isn't necessary anymore today, since we have at our disposal better isolating materials than 100 years ago. Actually Tesla contended with problems of osilation, which he could solve with the help of the flat coil, but it should turn out that the coil geometry is attached a crucial importance.

Everything had started with Tesla having to leave the Technical Highschool in Graz without diploma. He ran out of money and he had dared to criticize the venerable Professor Poeschel and his sparking Gramm dynamo. With that he had put himself under compulsion to succeed. Two years later he had ready the solution. In the year 1882 he discovered the rotary field in Budapest.
In the time to come he designs and builds an alternating current motor, but no-one wants to have it and surely Thomas Alva Edison not. Tesla after this disagreement very fast gives up his job at the Edison Company again and again stands under pressure to succeed. With that the eternal bachelor Tesla urges himself to ever higher efforts. He wants to prove himself and the rest of the world that his alternating current system is superior to the direct current system.
Direct current, as is well-known, can't be transformed, and thus the advantage of Teslas alternating current lies in the possibility of power transport by high-tension cable over large distances. But for that the high-tension transformers first had to be developed and thereby the said problems of isolation occurred.
With each turn the tension voltage at the transformer winding increases. The distance to the grounding point lying on the outside has to be chosen bigger with each turn, so that no blow inside of the high-tension winding occurs. A consistent solution of the problem in accordance with engineering is the flat coil used by Tesla, wound spirally from the inside to the outside (fig. 16.9 A)<i>.

It thus is correct that isolation technical reasons led to the flat coil, since Tesla himself was completely surprised as he had to find out that this coil can lose its self-induction, that scalar waves can be detected with it and that it is cooled down during operation in an inexplicable manner.

This cooling effect Tesla has investigated more detailed and after all even used. In his patent specification concerning the superconductivity he describes, that the flat coil also loses its Ohmic resistance, if he in addition previously cools it with liquid air. The remaining cooling down to absolute zero his flat coil obviously has carried out entirely by itself with help of the neutrinos (fig. 16.9 B)<ii>.

<i>: Nikola Tesla: Coil for Electro-Magnets, Patent No. 512,340 (1894)

<ii>: Nikola Tesla: Means for Increasing the Intensity of Electrical Oscillations, Patent No. 685,012 (1901)

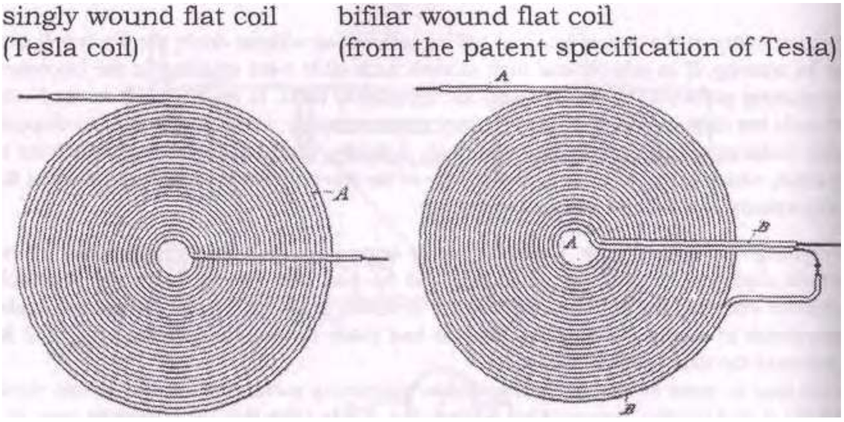
singly wound flat coil
(Tesla coil)
bifilar wound flat coil
(from the patent specification of Tesla)
A
A
A
B
B

Fig. 16.9 A: Coiling techniques of Tesla's flat coil.[i]

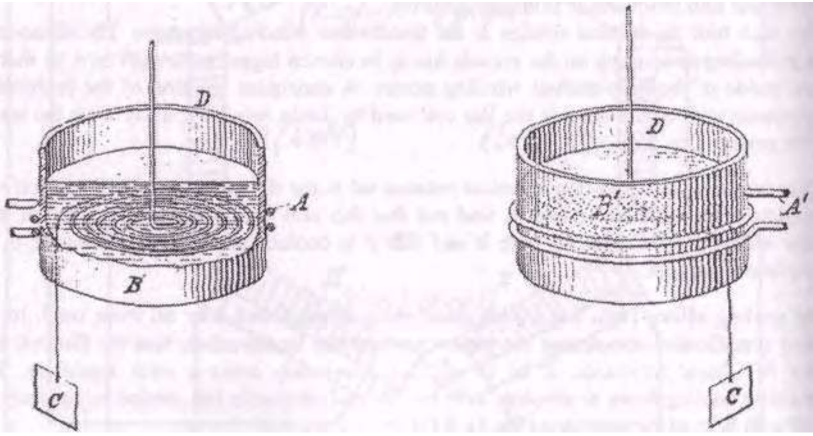

Fig. 16.9 B: Patent specification of Tesla concerning Superconductivity.[ii]

---

[i] Nikola Tesla: Coil for Electro-Magnets, Patent No. 512,340 (1894)

[ii] Nikola Tesla: Means for Increasing the Intensity of Electrical Oscillations, Patent No. 685,012 (1901)

16.10 The secret of the flat coil

The technical function could be explained in the way that the charge carriers of a flat coil by induction are set into motion for excitation from the outside. The transmitted energy shows in form of kinetic energy. The spiral flat coil becomes narrower and narrower towards the inside, the length of each winding shorter and shorter, so that the kinetic energy inevitable has to decrease in favour of a rotational energy. The faster and faster rotating spherical vortices are pulled apart to flat discs and eventually to ring-like vortices by the centrifugal force. The electrons at first become neutrinos bound to a line and finally free neutrinos. Tesla has technically used the first ones in the single-wire-transmission technique (fig. 9.5) and the last ones in his wireless energy transmission (fig. 9.7).

Like many other inventors, Tesla owes also the inventions, which he counts his greatest, like the radio technique and the Magnifying Transmitter, first of all his industriousness. his persistence and a great deal of inventor luck. A magician, as he is called in his most important biography, he by no means was<i>. The flat coil, to which led him chance and which plays a central role in all these inventions, gave him the lucky position, to collect neutrinos and materialize them to charge carriers or in reversed direction to dematerialise electrons to neutrinos.

The technology however is everything else but new. Already the Lituus of the Etruscan and Roman Augurs and the crook of the priests had the same spiral structure (fig. 16.10). In the case of the devices, which the Augurs for instance served at land surveying, it clearly concerns flat coils according to Tesla. We will go into this strange „gauge" more in detail in part 3 of the book<ii>.

The trick probably is, that one component of the electric field pointer is directed towards the centre of the coil and as a result some open field lines are generated, which then collect neutrinos from space.

In this process the neutrinos thanks to the resonant interaction are slowed down to the speed of light and following, as discussed, materialized by means of the flat coil, as in addition rotational energy is withdrawn from the neutrinos. Since the receiver oscillates resonant with opposite phase, in addition the thermal oscillations are reduced and the receiver becomes cold!

If one compares the Mobius coil with the Tesla coil, then besides numerous properties in common the strength of the first coil lies in the production of open field lines and the collection of neutrinos, whereas the special and additional property of the flat coil lies in the materialization, in the conversion of neutrinos into charge carriers. But the advantages of the flat coil have to be bought at the expense of having to work with very high tension voltages (above 511 kV) and with large changes in tension voltage (du/dt). With this set of difficulties we will have to deal in more detail.

<i>: Margaret Cheney. Nikola Tesla, Erfinder, Magier, Prophet (Orig.: Man Out Of Time, 1981), Omega-Verlag Diisseldorf 1995

<ii>: K. Meyl: Electromagnetic environmental compatibility, part 3, edition belonging to the information technical seminar, INDEL Verlagsabteilung 2003.

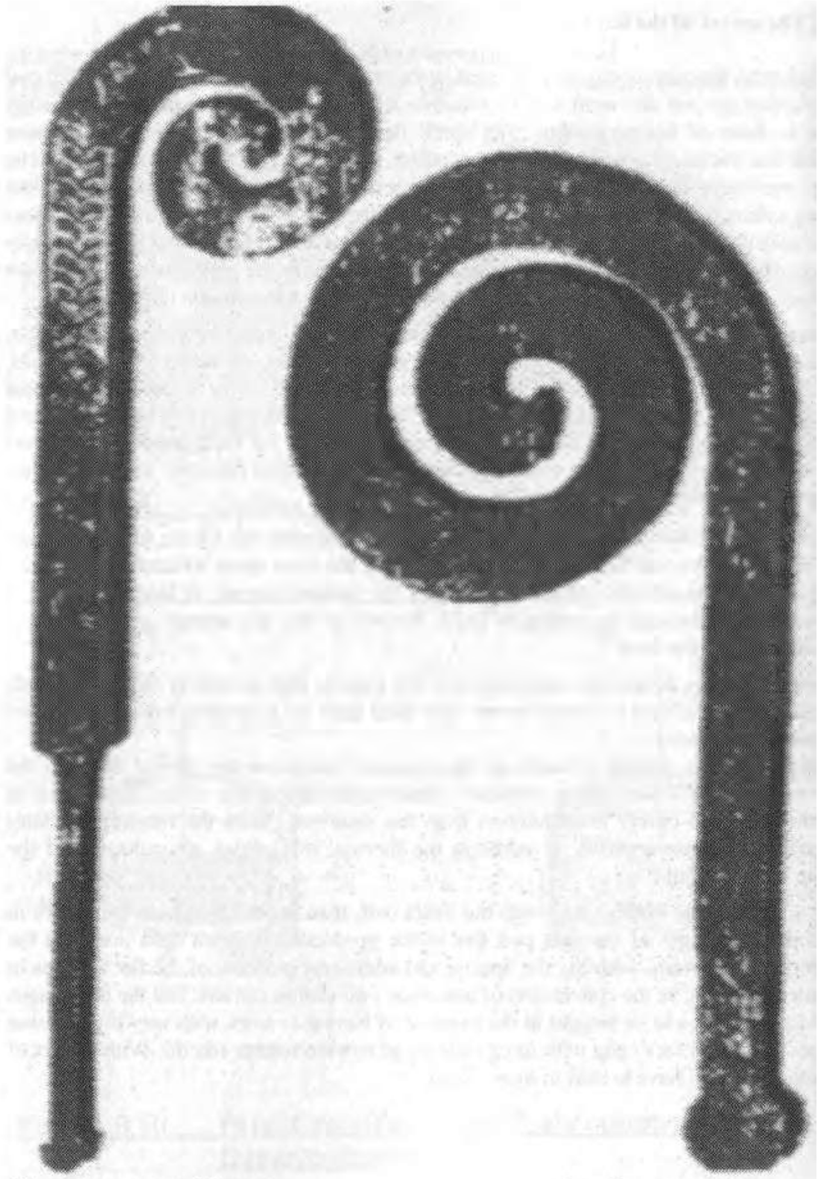
The Etruscan Lituus
the Roman Lituus

Fig. 16.10: The Lituus or crook of the Augurs in ancient Rome

16.11 Discussion concerning the technology of the neutrino collectors

Let us again collect the facts for the discussion: A SET-device is distinguished by a more or less unipolar design and open field lines, with which interact neutrinos, which are oscillating in resonance. These then are slowed down and collected. For the transient process a large change in tension voltage (dE/dt) is required, which can be obtained directly, for instance by means of a spark gap, or indirectly by means of Faraday's law concerning the unipolar induction (E = v x B).
The discussed possibilities concern the acceleration of a machine part (dv/dt), the variation of the magnetic field (dB/dt) by pulse-like excitation signals (16.5) or by magnetic flux variation (16.3) and the railgun. which even can be operated without foreign magnetic field (fig. 15.5 C) and for which in that case occur both a dv/dt and at the same time a dB/dt.
For resting arrangements the velocity v is that of the charge carriers moving in the conductor. So that Faraday's law thereby doesn't lose its influence, the pointers of E and v mustn't point in the same direction, as in the case of "normal" coils. Unconventional windings, which for instance can be knotted like Mobius strips (16.7), take remedial action. Also the ancient crook, rediscovered as flat coil of Tesla (fig. 16.10), proves to be suitable in principle. Here one component of the electric field pointer points in the direction of the centre so that the wanted, at least partly, unipolar arrangement can be formed.

The first step, the collecting of space quanta, shouldn't pose an insurmountable obstacle anymore in view of the numerous possibilities and the detailed explanations. A real difficulty we still have before us, because in most cases some ring-like vortices bound to a line are formed, for which no electronic construction element exists and for which functioning converters hardly are known.
There spoons are bending, some lumps are flying through space, radioactivity is disappearing without a trace, light phenomena are formed and the device suddenly is becoming cold. Almost all inventors, who have arrived in this place, are enthusiastic about the not understood effects or with that are wanting to get attention, but hardly anyone really starts something with that. Until now the necessary system and an useful theory were missing.

Only too often isn't considered, that only an indirect conversion into charge carriers is possible, that during the materialization of neutrinos a intermediate product is formed, which can be described with the model concept of a neutrino bound to a line or of an oscillating ring-like vortex. The technologies collected in this chapter concerning the collecting of neutrinos only form the first step from the free to the bound ring-like vortex. The coming chapter is dedicated to the second step. Here we should try to understand the properties of space quanta bound to a line and loudly think about for which purpose we could use them in practice<i>.

<i>: Reference: Both chapter 16 and chapter 17 treat point 4 according to table 15.1 concerning the resonant interaction (page 86).

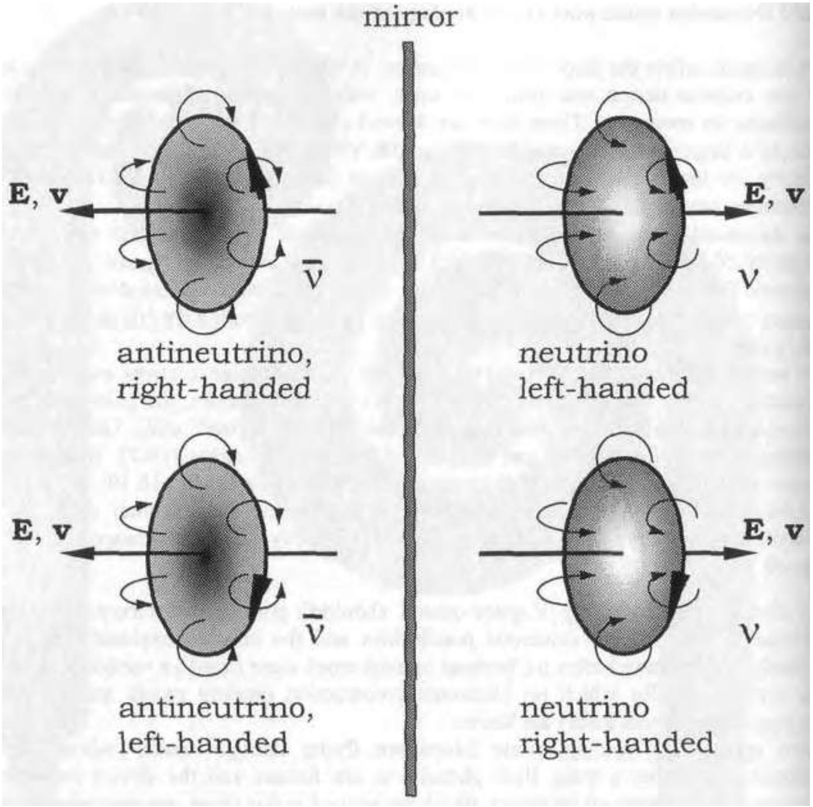
mirror
E, v
$\bar{\nu}$
antineutrino, right-handed
E, v
ν
neutrino left-handed
E, v
$\bar{\nu}$
antineutrino, left-handed
E, v
ν
neutrino right-handed

Fig. 17.1: Application of the mirror transformation to the neutrinos

---

<i>: Kleine Enzyklopadie Atom, Verlag Chemie GmbH, Weinheim 1970, S. 119:
„For the neutrinos the quantum number parity isn't defined, because they have a fixed association of the direction of momentum and spin; the sense of rotation of the spin and the direction of momentum in their case form a left-hand wound screw, in the case of antineutrinos a right-hand wound screw. By the shown mirroring the direction of momentum is reversed, but the sense of rotation remains unchanged; i.e. reactions, in which neutrinos occur, are not mirror invariant, they violate the law of conservation of parity". (translated)

Note of the author:, an antineutrino by no means can be assigned to the anti-matter, since it exactly like every neutrino alternatingly takes the matter state and the anti-matter state, by oscillating around itself (see fig. 7.12). The description merely follows a definition founded in usefulness.

17. Technical use of the weak interaction

17.1 Radioactivity caused by neutrinos

Neutrinos are standing in a weak interaction with other elementary particles. This circumstance is known in general. According to the considerations and derivations expressed in the book the neutrinos mediate the resonant interaction, what leads to the conclusion, that the weak interaction represents a partial aspect of the resonant interaction, in which case it merely concerns the proximity.
Efforts are further being undertaken, to combine the weak interaction with the electromagnetic to an electroweak interaction, after it was remarked that a certain coupling constant corresponds in both cases. We of course aren't surprised, because the electromagnetic interaction anyhow describes only the special case of the resonant interaction with the frequency zero.
The weak interaction concerns with the very small range of just 10^{-16} meters only the proximity of the neutrinos, for instance the ß-decay, where the neutrinos for the reason of their oscillating charge a free neutron rattle and shake so long, until it eventually decays. on the average after approximately a quarter of an hour.

The Austrian physicist Wolfgang Pauli had remarked, that half the decay energy after a beta-decay is missing and the balance sheet of energy isn't working out. In addition also the balance sheet of angular momentum isn't working out, because the nuclear spin is being changed for a whole unit. Pauli as a result 1930 has introduced a hypothetical particle without mass and without charge, which he called neutrino.
With that Pauli and his co-working Italian colleague Fermi it is true are the givers of the name of the neutrinos, but not by all means the discoverer. If Cowan and Reines 1956 have detected these particles with large expenditure of devices, then also that by no means was a premiere, as falsely can be read in textbooks and encyclopaedias. After all Nikola Tesla already decades ago had demonstrated, that the neutrino radiation not only exists, but even can be used energy technically.

Now the radioactive decay a ß-radiation occurs, triggered by the conversion of neutrons in protons or vice versa. This obviously takes place under the influence and participation of neutrinos in the atomic nucleus. The ß-radiation to a special extent consisting of electrons and positrons, as it occurs in the case of nuclear fission, is quite unhealthy and by no means ecologically compatible. Under the influence of free charge carriers not only metal lattices become soft and spoons can be bent, but also an electrolysis takes place, where the water molecules are splitted into their parts. That isn't a good prerequisite for the flora and fauna on our planet, which predominantly is built up of water structures.

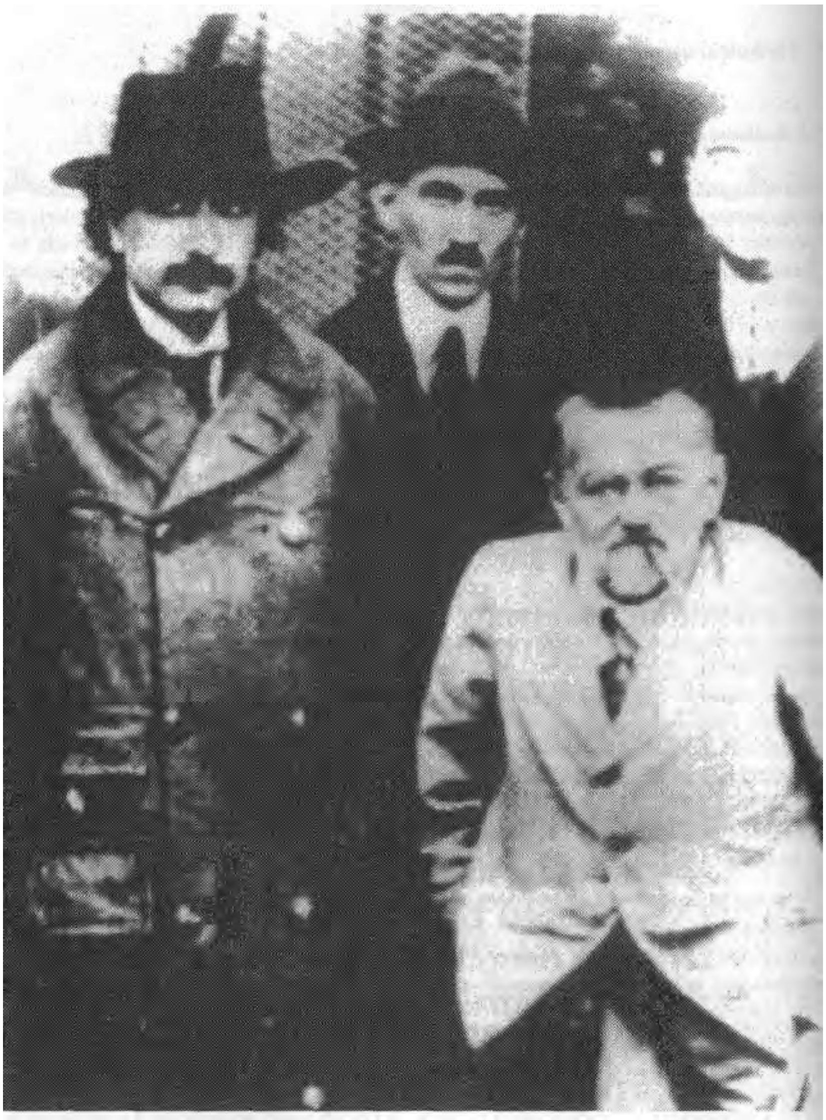

Fig. 17.2: Einstein, Tesla and Steinmetz<i> (from left to right)

<i>: Franz Ferzak: Nikola Tesla, Eigenverlag, S. 103

17.2 Nikola Tesla, the discoverer of the neutrino radiation

The discoverer of the neutrino radiation himself will best be able to explain the connexion. In the New York Times Tesla writes, that he has discovered and investigated the phenomenon of the cosmic radiation, long before others started their researches<i>:
"According to my theory a radioactive body is only a target, which constantly is being bombarded by infinitely small balls (neutrinos), which are projected from all parts of the universe. If this, at present unknown, cosmic radiation could be interrupted completely, then no radioactivity would exist any longer.
I made some progress regarding the solution of the mystery, until I in the year 1898 attained mathematical and experimental evidence, that the sun and similar celestial bodies emit energy-rich radiation, which consist of inconceivable small particles and have velocities, which are considerable faster than the speed of light. The ability of penetration of this radiation is so large, that it penetrates thousands of kilometres of solid matter, without their velocity being reduced noticeably. "
It must be admired how Tesla guided by experimental observations and a reliable instinct comes to the correct result. He merely with the conclusion, because of the missing interaction the neutrinos have to be inconceivably small, isn't quite right. Their size rather depends on the velocity, because the overfast neutrinos are being length contracted stronger. Tesla however hits the nail exactly on the head, if he on the occasion of the press conference for his 81st birthday declares, the radioactivity is a clear proof of the existing of an outer radiation of cosmic origin<ii>. „ If Radium could be shielded against this radiation in an effective way", Tesla writes in an essay of 1934, „then it wouldn't be radioactive anymore". At this occasion he contradicts Albert Einstein, without thereby pronouncing the name and is indignant at the wrong working method of the scientists<iii>.

Me personally fascinates, how here until now ignored results have been presented, which I first had to work out theoretically myself with difficulty. Tesla, to the best of my knowledge, hasn't taken theoretical derivations, at least none have been handed down. As a brilliant experimental physicist he must have reached his conception world by means of the measuring technique. The perfect correspondence of his experimentally determined and the by me theoretically won insights should be judged as evidence for the correctness of this view.

<i>: Dr. Tesla Writes of Various Phases of his Discovery, New York Times, Feb. 6, 1932, P. 16, col. 8

<ii>: Tesla Said (J.T. Ratzlaff), Tesla Book Company, ISBN 0-914119-00-1, P. 272

<iii>: "The scientists of today think profound instead of clear. One has to be mentally sane, to be able to think clear, but one can think profound and nevertheless be completely insane. The scientists of today have substituted experiments by mathematics, and they travel from one equation to another and eventually build up a construct, which has absolutely no relation to reality"...

taken from N. Tesla: Radio Power will Revolutionize the World, Modern Mechanics and Inventions, 7/1934, (Tesla Said, P. 264)

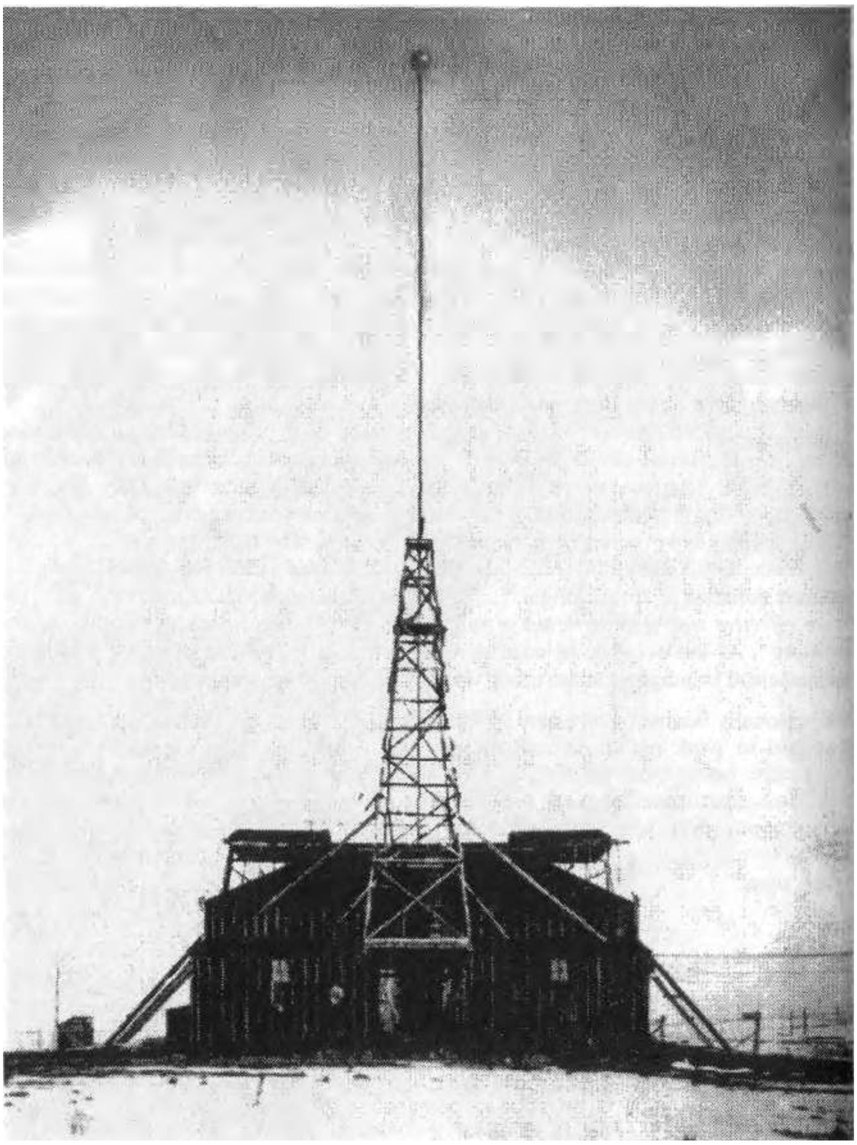

Fig. 17.3: The Magnifying Transmitter of Nikola Tesla
Would the research station in Colorado Springs (1899 - 1900) have been suitable for transmutation?

17.3 Transmutation and reduction of radioactivity

If we are capable to collect and bundle up neutrinos, then should with that an influencing of the radioactive decay become possible, then also a decontamination of radioactively contaminated material and a so-called transmutation should be possible. We here have a concrete possibility before us, to use the produced ring-like vortices directly for the benefit of humanity. Dangerous transports with Castor containers, permanent and temporary storage and the contamination of whole areas by radiation, like around the nuclear power station Tschernobyl, which got out of control, wouldn't be necessary at all. If namely burned out fuel rods and objects contaminated by radiation have undergone a concentrated neutrino shower, then the radioactive decay takes place accelerated, so that the half-life can be drastically reduced. After the treatment the dangerous special waste would have been changed to harmless domestic rubbish. Even recycling or reuse are feasible.

The topic is at least as explosive as the energy question and as well completely unsolved. That's at different places and in some companies is already feverish researched about technologies concerning transmutation. Often it are the same people, who also work at the theme of space energy. The reason quite simple is that in both cases neutrinos have to be collected and bundled up. In the case of transmutation however the necessity of materialization can be dropped, so that the goal can be obtained faster and simpler. Consequentially is reported of more cases of a successful decontamination and of transmutation, than of functioning energy converters. Until now most techniques however still hardly are suitable for bringing into action technically on a large scale, but they carry clues of a solution of the problem already in them.
At the congress "New Energy technologies from USA" 6.12.1997 in Zurich I have lectured concerning the theme<i>: „Presumably as the first one the doctor Dr. Wilhelm Reich has carried out corresponding experiments with his „Orgon accumulator" (fig. 9.2). From him also stem warnings about biological effects, which should be taken serious, if radioactive material is put under his Orgon accumulator and the process of decay takes place accelerated".
If one dares an interpretation of his experiments with only 1 mg radium, then numerous charge carriers materialized at hitting upon his sample of the bundled up neutrino radiation with the consequence of high electrostatics in the environment, which Reich has called DOR-state (Deadly ORgone) (chap.15.11).

Another way would be the rebuilding of a neutrino transmitter according to the plans and patent specifications of the experimental physicist Nikola Tesla (fig. 9.11). He 100 years ago had realized a real neutrino-broadcasting and for that developed an unconventional switching technique. Tesla called his transmitter a „Magnifying Transmitter" (fig. 17.3). He choose the name „Magnifying Transmitter", after he had received more energy than he had transmitted in experiments and this effect moreover was increased with increasing distance to the transmitter. He obviously also had collected free and synchronously oscillating neutrinos, and that would be the best prerequisite for a successful transmutation and decontamination of radioactive material!

<i>: Konstantin Meyl: Die Wechselwirkung der Neutrinos; uber Massnahmen, die Halbwertszeit beim radioaktiven Zerfall herabzusetzen. NET-Journal Jg. Nr. 3, Jan./Feb. 1998, S. 14-20

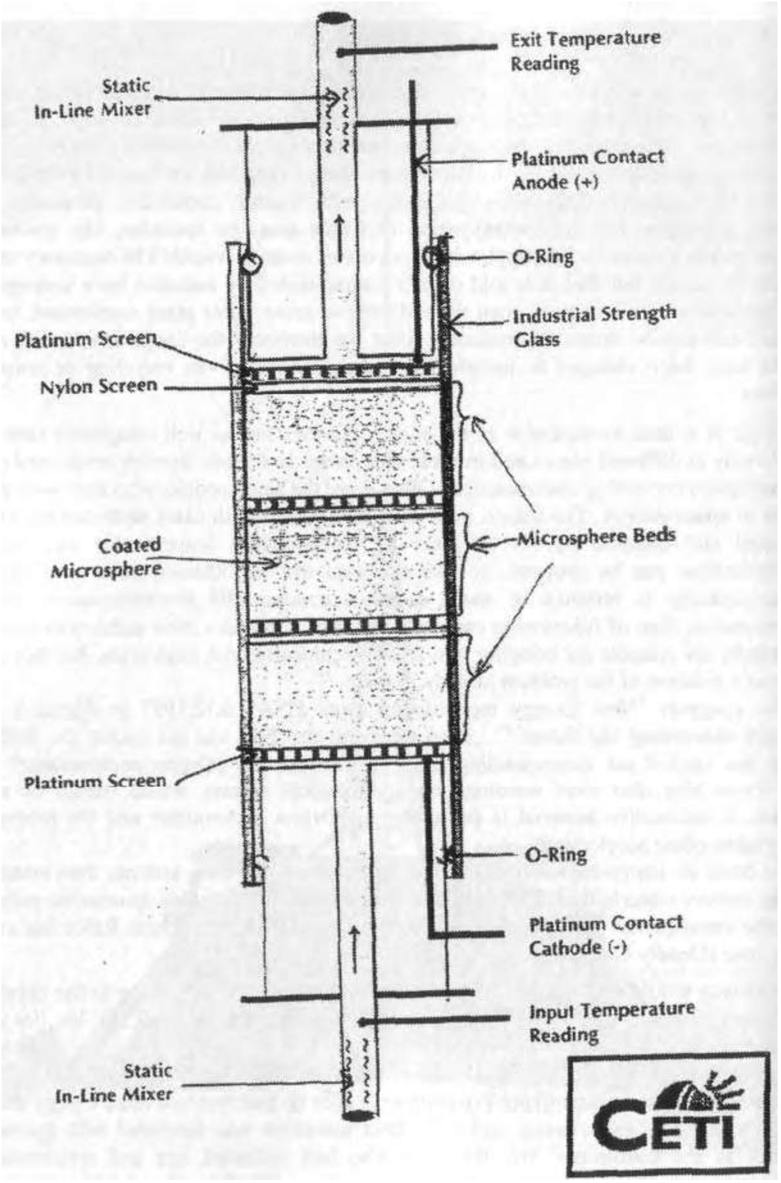
Exit Temperature Reading
Static In-Line Mixer
Platinum Contact Anode (+)
O-Ring
Industrial Strength Glass
Platinum Screen
Nylon Screen
Microsphere Beds
Coated Microsphere
Platinum Screen
O-Ring
Platinum Contact Cathode (-)
Input Temperature Reading
Static In-Line Mixer
CETI

Fig. 17.4: 1 kW Patterson Power Cell.<i>

<i>: Patterson, J. A.: System with electrolytic cell and method for producing heat and reducing radioactivity of a radioactive material by electrolysis. US Patent No. 5.672.259 of Sept. 30, 1997, as reprinted in Infinite Energy, July-Nov.1997, pp. 13-16

17.4 The Patterson Power Cell

As an example worth paying attention to, the Patterson Power Cell should be mentioned, which not only can be used for generation of energy but, how could it be otherwise, also for reduction of radioactivity and for transmutation. The energy cell invented by the chemist Dr. James A. Patterson is researched and developed further by the company CETI (Clean Energy Technologies Inc.) in Sarasota (Florida, USA). According to latest reports such a device supplies at most one kilowatt of heat energy at an over-unity of up to 4000; thus 4000 times the energy is released than is taken up as required for the operation!
In fig. 17.4 one single cell of the energy converter is shown. The glass container consists of three chambers of approx. 4 cm height each time and about 1.9 cm in diameter and works like a continuous-flow water heater. From the bottom tapwater or distilled water is supplied and at the top connection again taken away in heated form. The excitation takes place electrostatically by means of two platinum electrodes. The anode situated at the top is connected with the positive pole and the cathode with the negative pole of the source of direct current.
In the three chambers are situated tiny, filmy coated small balls, which form the real principal item of the cell. Production method and structure of the beads are oriented to the task to be mastered. In the case of a variant conceived for the generation of energy the barely one millimetre in diameter small synthetic beads carry after each other a thin layer of palladium, a layer of nickel and once again one with palladium<i>.

The small balls remind with their layered structure at first of the Orgon accumulator of Reich. They also seem to function as collectors of neutrinos, but for smaller wavelengths and much higher frequencies. In addition is chosen a concentric arrangement with the spherical form, which with the electrically conductive surface and the dielectric core fulfil the function of a cavity resonator. Resonance actually is possible with flying past neutrinos, which have a wavelength which amounts to an integer multiple (1,2,3,..) of twice the diameter of the sphere.
The arising resonant oscillation however concerns not only the electric and magnetic fields, but by means of electrostriction and magnetostriction also a mechanic oscillation of size. The oscillation of size in reverse causes again oscillating electric and magnetic field pointers, which are partly open along the spherical structure and can interact resonant with further neutrinos. In this repercussion field the actual secret of the functioning of a Patterson-cell is hidden.

The president of the German association of space energy, Prof. Dr. Dr. Josef Gruber on the occasion of his visit at the company CETI and of a conference about „cold fusion" in Vancouver has published a report<ii>:

<ii>: Gruber, J.: Kalte Fusion und Raumenergie, report concerning the 7th international conference about "cold fusion" of 19.-24.4.1998 in Vanvouver, B.C., Canada (Infinite Energy, Issue 19, 1998) and report concerning "International Symposium on New Energy" of 23.-26.5.1997 in Denver, Colorado, USA (Denver-report of the remote University Hagen)

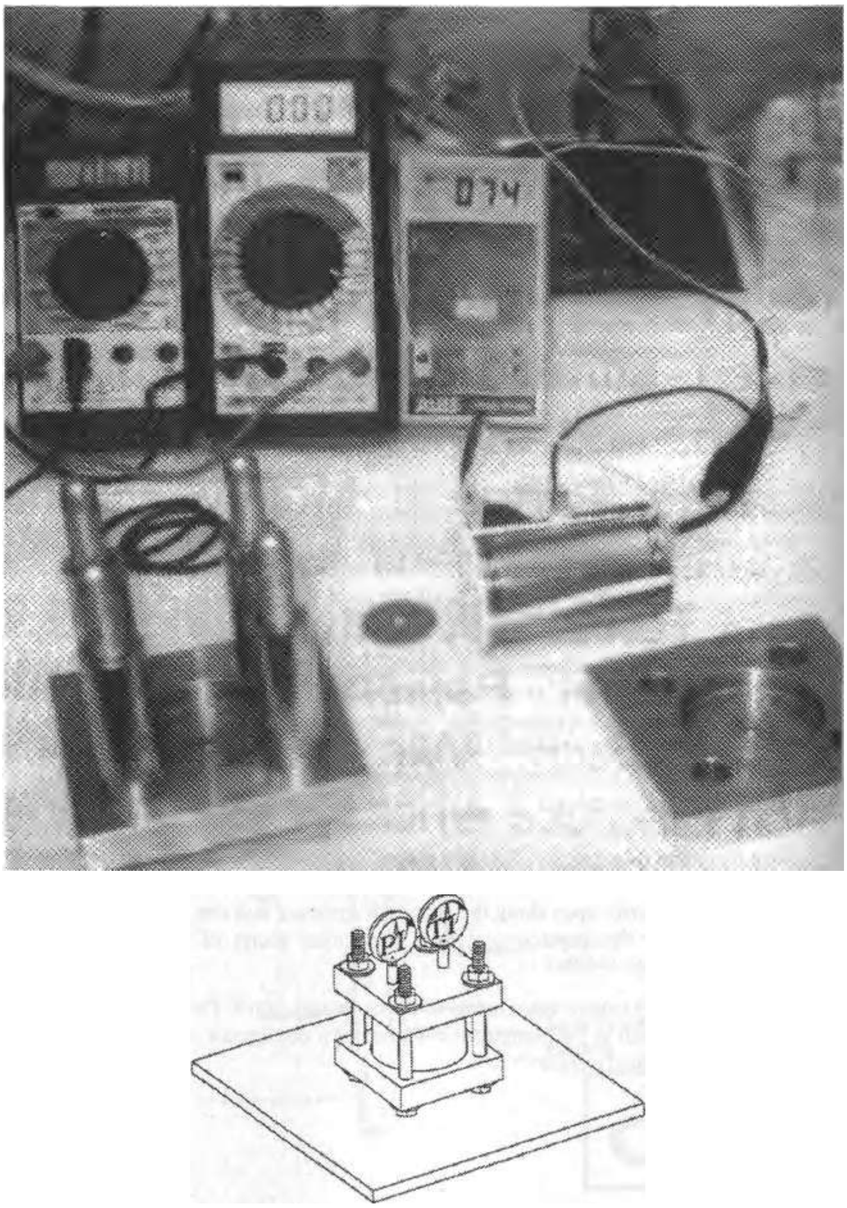

Fig. 17.5: _____ Low Energy Nuclear Transmutation Cell,[i]
the kit for experiments LENT-1
(Showing Temperature and Pressure Gags)

---

[i]: taken from the NET-Journal, 10/11 1997, S. 7

17.5 Report concerning the cold fusion

"In the case of the Patterson transmutation cell it concerns a special electrolysis cell, in which the radio nuclides are charged. During the electrolysis the decay activity measured by means of a Geiger-Muller counter decreases drastically. Within a few hours reductions of up to 80% are obtained. In such an electrolysis system with electrodes consisting of specially coated beads low-energetic nuclear reactions can be observed. In this case elements are detected in the metallic layers of the beads, which before were not contained in these. Further are measured changed (unnatural) proportions of isotopes. From these transmutation can be inferred, which is used in the cell for the conversion of radioactive elements.

Until now natural uranium and thorium were used as radioactive material. The reproducibility already now is very good. The applicability for commercial purposes, for the reduction of the radioactivity of burnt up fuel elements and for the conversion of plutonium is easy to see".

At his visit in the laboratory of CETI in Sarasota 28.5.97 Prof. Gruber has witnessed an experiment to annihilate radioactivity with the Patterson transmutation cell. He reported about it: "In the presence of N. J. Olson from Pacific Northwest Laboratory (operated by Batelle for the U.S. Department of Energy) a team of the television station ABC conducted by the science journalist Dr. M. Guillen made recordings for a television program, which meanwhile has been broadcasted all over the country. The original tension voltage - among others a Geiger counter rattled - subsided as the expected success became apparent in the experiment: After one hour 50% of the radioactivity stemming of uranium nitrate were removed, in another half an hour further 13%. Doing so also a considerable lot of surplus heat was produced.

New techniques of this kind to reduce radioactivity have an important advantage: One brings the SET-device there where the radioactive material is situated and reduces the radioactivity on the spot. For the conventional technique, still being in research, in contrast to that the radioactive material first is wrapped in a complicated manner and then transported to a special factory, where radioactivity is reduced at great technological expenditure and use of energy - altogether a procedure which is relatively costly and politically only hard to carry through".

Patterson cells already can be obtained commercially for research purposes. The same goes for a demo-kit<i> of the company Trenergy, Inc. of Hal Fox, who has built up the biggest database of the world concerning the theme "Low Energy Fusion" in the "Fusion Information Center" at the university of Salt Lake City. He is editor-in-chief of the "Journal of New Energy".

If one believes the statements on their web-site<ii>, then one is capable of converting 10 g of radioactive thorium in 900 mg titanium and some copper in less than an hour with the kit for experiments LENT-1. Under a pressure of 3200 psi sodium already is said to have fusioned to gold, something of which alchemists dream, as an unintentional side effect, as researchers in Cincinnati/Ohio say<iii>.

<i>: NET-Journal, Jan. 1997, S. 24

<ii>: see Internet: http://www.hal-fox@slkc.uswest.net

<iii> : NET-Journal, 10/11 1997, S. 7

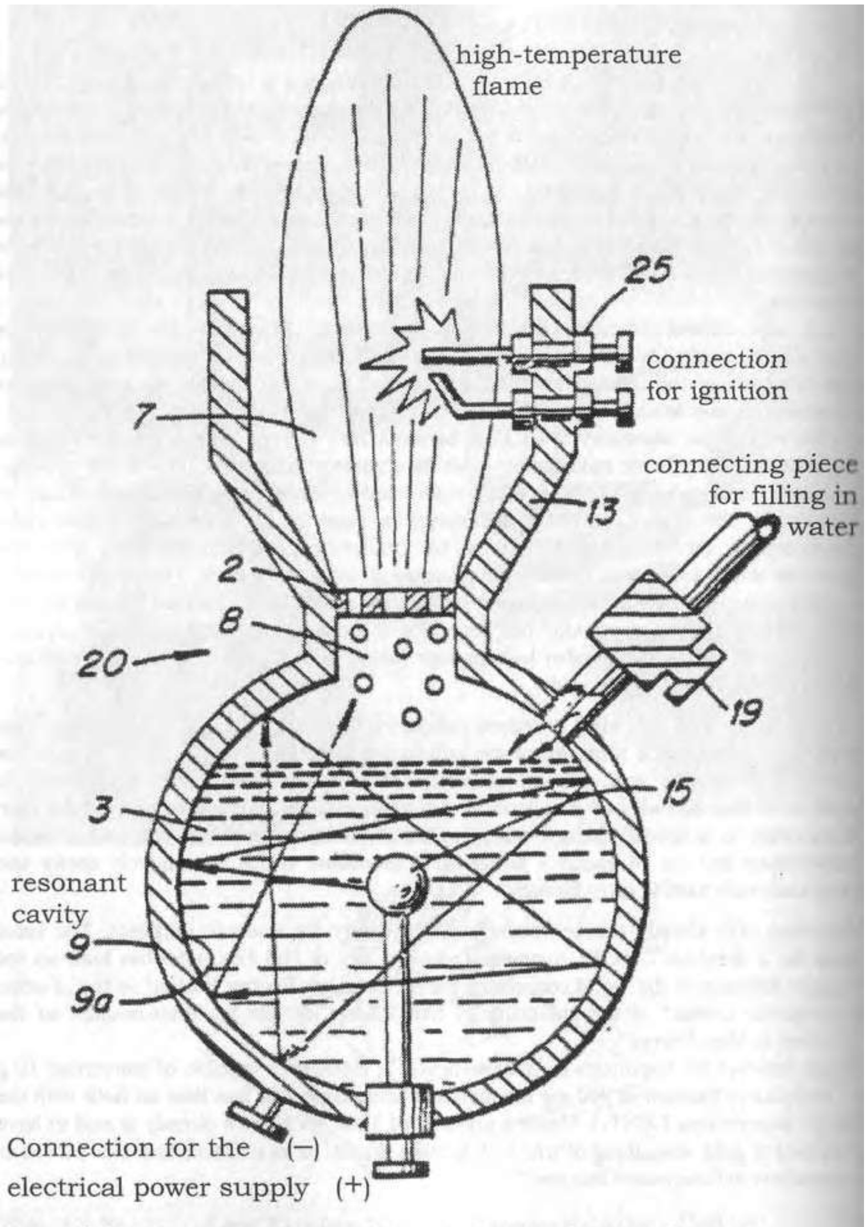
high-temperature flame
25
connection for ignition
7
connecting piece for filling in water
13
2
8
20
19
15
3
resonant cavity
9
9a
Connection for the (−)
electrical power supply (+)

Fig. 17.6: The water-fuel-cell according to Stanley Meyer<i>

<i>: H.-J. Ehlers: Stan Meyers Wasser-Zellen-Technik, Raum & Zeit special 7, p. 201, taken from the Canadian patent specification 1234 773 of 5-4-1988

17.6 Water-Fuel-Cell technology

The United States are, just what concerns spectacular techniques like the removal of radioactivity or the transmutation of new materials, obviously still the land of unlimited possibilities. Nowhere the list of researchers of cold fusion is as long as in North America.
But there also in many places only is tried and tinkered without visible system or usable theory. Then in many cases unfortunately only a show-effect is to the fore, while construction plans and details concerning the way of functioning, as far as they actually exist, are kept secret. In total there, besides a lot of hot air and wrong hopes, is left behind little to be used and cited.
By the way, in my opinion cold fusion has to do very much more with space energy and neutrinos than with hot fusion. How much disaster a wrong referring to and an unusable theory can bring about, has become clear at the example of the "cold fusion researchers" Fleischmann and Pons. They have placed themselves into the scientific offside with their misinterpretation concerning cold fusion.
Moreover are both primarily scientists and theorists. Practice however is, according to general definition, "if in spite of all it functions!" The American Stanley Meyer is such a practical man and his water-cell-technology actually seems to function, although he in his 18 patent specifications gives theoretical explanations, which he just as well could have saved himself. With that his effect isn't explained.
An usable interpretation would be that this device, comparable to a cell of a plant during photosynthesis, splits water molecules into its parts by putting on neutrinos. Even without knowledge about space quanta the buggy of Stanley Meyer already runs with a air-cooled 1500 cc VW-engine, and it consumes no gasoline at all. The tank is filled with water; it even may be sea water. The consumption of water lies at 2.8 litres at 100 kilometres and thereby is formed predominantly hot steam again as a combustion product.
If thus cold water is converted into hot water and at the same time mechanical energy is available. then inevitably another source of energy must be involved. According to my interprelation it must concern the neutrino field. If the here presented details are correct then the over-unity effect lies at approx. 100, the degree of effectiveness thus at 10000 percent.

American companies, with which Stanley Meyer had concluded contracts, should make the "Water-Fuel-Cell technology" mature for series. Also the financing seemed secured. But then per internet the message came, he 21.3.1998 was having supper in a restaurant in Grove-City, as he suddenly jumped to his feet from the table and called out, he had been poisoned. He died on the spot<i>.

Alarge number of inventors is known, who tap space energy with the help of water. It concerns an increase of the content of oxygen or of the content of colloids, thus an improvement of the water quality. Or it concerns formation of vortices as already in the case of Walter Schauberger, glowing phenomena or also the generation of free energy from the neutrino field. In this concert the concept of Stanley Meyer takes an outstanding place, as particularly efficient, instructive and clearly understandable for us, for which reason we cast our eyes over the design (fig. 17.6 and 17.7).

<i>: Zum Hinschied von Stanley Meyer", NET-Journal April/Mai 1998, S. 25; see: obituary in Infinite Energy Magazine

A. The pattern of the pulses of the excitation voltage:

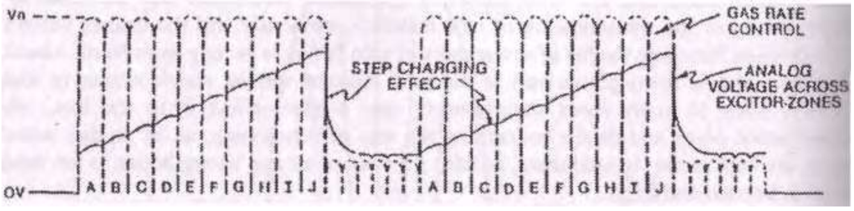

B. The wiring diagram:

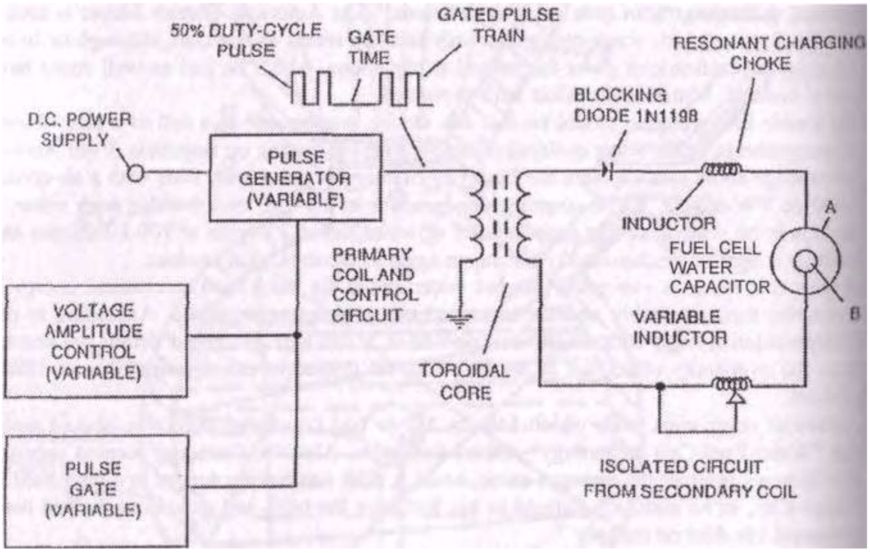
50% DUTY-CYCLE PULSE
GATE TIME
GATED PULSE TRAIN
RESONANT CHARGING CHOKE
BLOCKING DIODE 1N1198
D.C. POWER SUPPLY
PULSE GENERATOR (VARIABLE)
INDUCTOR
A
FUEL CELL WATER CAPACITOR
B
PRIMARY COIL AND CONTROL CIRCUIT
VARIABLE INDUCTOR
VOLTAGE AMPLITUDE CONTROL (VARIABLE)
TOROIDAL CORE
PULSE GATE (VARIABLE)
ISOLATED CIRCUIT FROM SECONDARY COIL

Fig. 17.7: Wiring diagram for the driving of the water-cell<i>

<i>: Stanley, A. Meyer: Process and Apparatus for the Production of Fuel Gas and the Enhanced Release of Thermal Energy from such Gas, 15.06.89, US-Pat. 207,730; International Publ. WO 89/12704; Int.Appl. PCT/US89/02622

17.7 Unconventional electrolysis

Numerous of the by Stanley Meyer used construction principles already have been treated, be it the excitation with pulses of electric tension voltage or the spherical structure of the resonant cavity (fig. 17.6). As a spherical capacitor with the positive pole in the centre and the negative pole at the outside edge it corresponds to the model of the electron and fulfils in an almost ideal way the conditions of an unipolar arrangement according to fig. 15.8 B. The use of certain patterns of the pulses and steep flanks of the pulses (large dU/dt) make possible effects of resonance at frequencies starting at ten kilohertz, in which neutrinos participate increasingly. First the series resonant circuit, consisting of the adjustable, external inductance and the spherical capacitor, is stimulated by means of the current rectifying wiring (fig. 17.7). In the case of resonance, which is carried out by comparison of the inductance, the excitation current drops, whereas the tension voltage at the same time teaches values of more than 1000 Volts. If in addition a neutrino resonance occurs, then the known avalanche effect will occur. The equipment then takes up virtually no current.

As a dielectric serves water, with which the container is filled permanently. The dimensions are oriented after the velocity of movement of the water molecules according to details of the inventor. Also the oscillating water molecules should go into resonance. Then they can help to materialize the neutrinos. Their rotational energy partly passes to the water molecules and as soon as the neutrinos have been converted to charge carriers, they will take the water molecule from the oppositely charged side and split it without further ado. The oxygen and hydrogen gases leave the capacitor through fine openings at the upper edge of the spherical chamber, which are so small, that no ignition back can occur, and in the simplest case reach a combustion chamber, where they burn again to water as a high-temperature flame (fig. 17.6).
The gases of course also can be guided into the cylinder of a Otto engine and be ignited there, as in the case of the experimental buggy. In the sectional drawing can be seen a filler by means of which, according to the consumption, water is refilled. The round resonant cavity not necessarily has to be spherical. Stanley Meyer more frequently differs from the ideal form and works with a cylindrical symmetry (fig. 17.8), with which obviously in spite of that the goal can be obtained, if perhaps not quite so good. To this compromise the explanations of fig. 15.8 C apply.
If we, to conclude, cast our eyes over the wiring diagram which Stanley Meyer discloses in his patent specification (fig. 17.7 and 17.8)<i>. In the centre is a transformer, which should produce an if possible high tension voltage. A rectifying diode, which takes care that only positive tension voltage pulses serve the excitation, is switched in series with the reaction capacitor, which is filled with water, a fixed and a variable inductance. In that way the positive pole is always situated in the centre of the reaction chamber. If both connections would be exchanged or the diode be turned over, then the neutrinos presumably would materialize in positrons and not in the wanted electrons. But if one leaves out the diode entirely and one has a tension voltage changing both in positive and in negative direction, then maybe electrons and positrons are equally generated, which annihilate each other under emission of gamma quanta. Doing so no gas is formed, but at most light, as long as the effect of collecting neutrinos isn't lost also. The concept should be worth to be examined more detailed already of pure scientific interest.

A: The high-tension transformer:

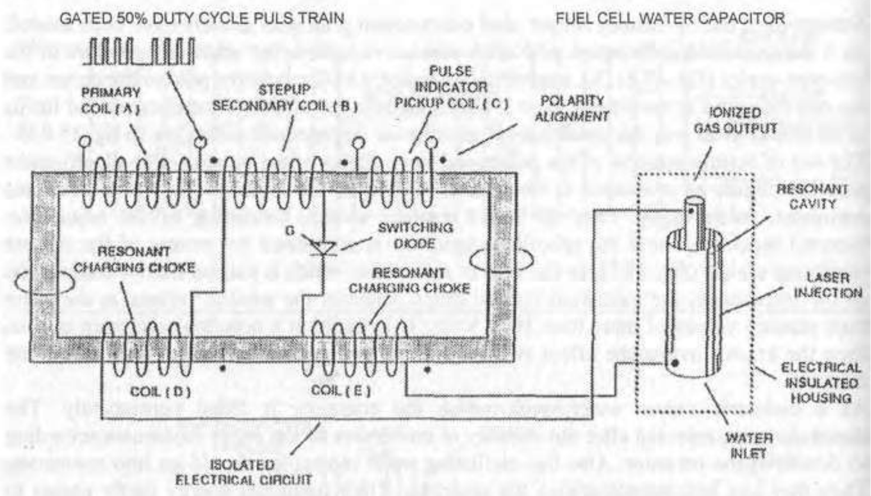
GATED 50% DUTY CYCLE PULS TRAIN
FUEL CELL WATER CAPACITOR
PRIMARY COIL (A)
STEPUP SECONDARY COIL (B)
PULSE INDICATOR PICKUP COIL (C)
POLARITY ALIGNMENT
IONIZED GAS OUTPUT
RESONANT CAVITY
G
SWITCHING DIODE
RESONANT CHARGING CHOKE
RESONANT CHARGING CHOKE
LASER INJECTION
ELECTRICAL INSULATED HOUSING
COIL (D)
COIL (E)
WATER INLET
ISOLATED ELECTRICAL CIRCUIT

B: The driving:

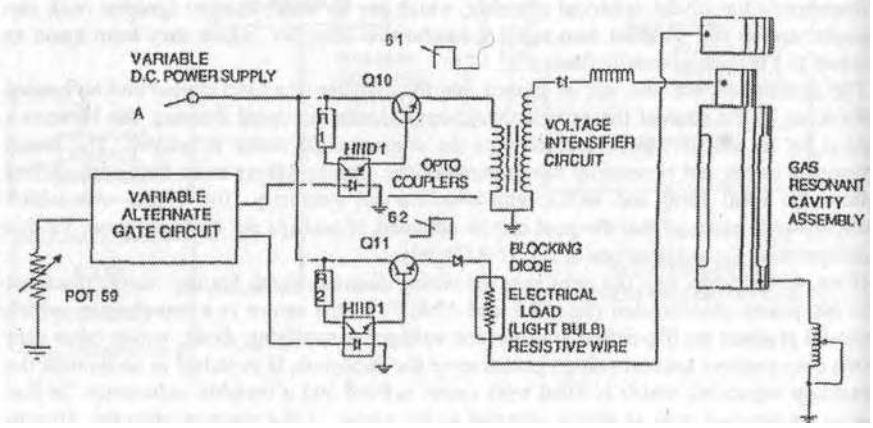
61
VARIABLE D.C. POWER SUPPLY
Q10
A
R1
HIID1
OPTO COUPLERS
VOLTAGE INTENSIFIER CIRCUIT
VARIABLE ALTERNATE GATE CIRCUIT
62
Q11
BLOCKING DIODE
GAS RESONANT CAVITY ASSEMBLY
POT 59
R2
HIID1
ELECTRICAL LOAD (LIGHT BULB) RESISTIVE WIRE

Fig. 17.8: High-tension transformer and driving of the cylindrical water-cell<i>

<i>: Stanley A. Meyer: Process and Apparatus for the Production of Fuel Gas and the Enhanced Release of Thermal Energy from such Gas, 15.06.89, US-Pat. 207,730; International Publ. WO 89/12704; Int.Appl. PCT/US89/02622

17.8 Materialization of neutrinos

In the centre of attention of the water-fuel-cell of Stanley Meyer is on the one hand the unipolar arrangement (according to fig. 15.5 B or C) to collect the neutrinos and on the other hand the water for slowing down and materializing. Over and above that a lot was tinkered and tried, as can be inferred from the patent specifications (fig. 17.8). That starts at the. coiling technique of the transformer and concerns the experiments with laser stimulation as well as the top part, which according to the inventor should extract the electrons (Electron Extractor Grid). The measures may bring an improvement, but are insignificant for our considerations.

At this place it primarily concerns the question of the materialization of neutrinos. But if such a materialization has to be made complete, if water molecules must be splitted, depends on how long the process lasts. If everything goes very fast, then perhaps it is sufficient that a neutrino for a short time is showing as an electron, before it again oscillates back. In this short time the splitting process already could have taken place. The used neutrinos for that had to be very low-frequent and very slow. They after that could again leave the reaction chamber and fly on as neutrino.

Possibly the fishes, which live in stagnant waters or in the deep sea, owe the slow neutrinos the content of oxygen in the water. Because here no bundling up takes place, the splitting of water takes place rather by chance. The volatile hydrogen atoms escape very swiftly, whereas the big oxygen atoms are left behind in the water.

Now we still don't know how charge carriers can be won. At the RQM unit resonant oscillations of size had been measured and that can be judged as an important clue. If namely a space quantum is slowed down then it becomes bigger. The inverse case we already had made us clear: If a particle is correspondingly fast then it is sufficient length contracted to fit through the tunnel (fig. 6.14).

To slow down neutrinos according to that the target area should carry out an oscillation of size with opposite phase. Organic material and biological systems are excellently suited for that. Every contraction of a muscle brings that to mind. Inorganic matter and our technology however normally don't know this phenomenon. Technical energy converters simply are built up fundamentally else than biological muscle machines.

In historical sources at certain places is pointed to the fact that priests had experimented with quartzes and miraculous phenomena were observed. Of course the question is asked, if such experiments today still can be reproduced. Because the density of the earth slowly increases due to the growth in volume, also the wavelength of these oscillation quartzes is changed. The same neutrino radiation therefore today can't be active anymore. It thus would be completely inappropriate wanting to reject a historical source only because a described effect today doesn't want to function anymore.

Magnetostrictive or electrostrictive material could be a solution. For instance a piezocrystal, which contracts under the influence of an outside field. As an ultrasound converter an applied alternating voltage leads to the emission of a sound wave. If we reverse the function, then a received longitudinal wave should lead to an electric tension voltage and then at last the materialized charge carriers could be taken off,!

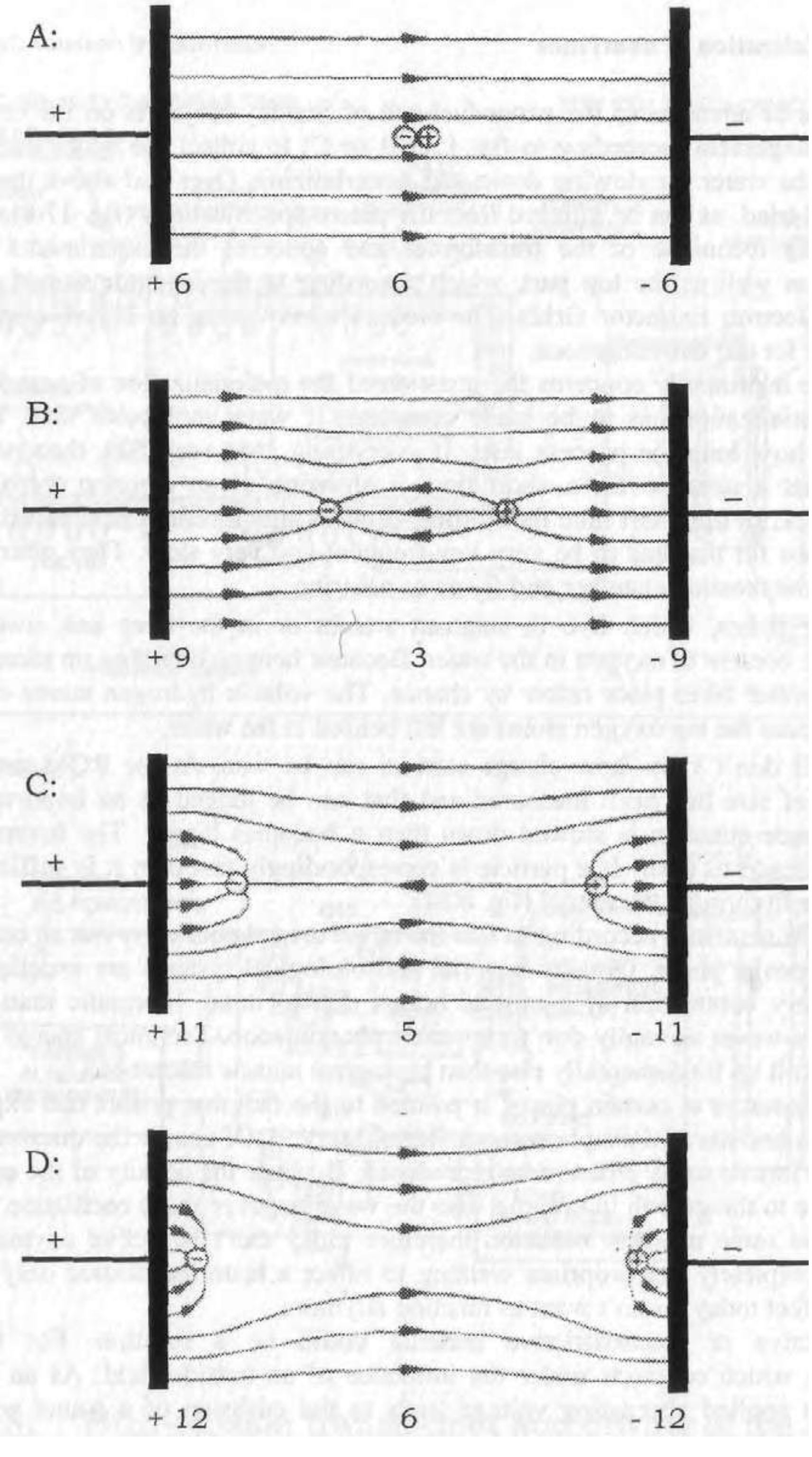

Fig. 17.9: The course of the field in the case of a glow-discharge<i>
(The numeric values indicate the active field lines, as they can be read from the representation).

<i>: K. Kupfmuller: Einfuhrung in die theoretische Elektrotechnik, 12. Aufl. Springer-Verlag (1988), S. 197

17.9 Oscillation of size and luminescence

There exist substances, which at high temperatures start to glow. In the sense of the theory of objectivity presented in the first part of the book in the case of temperature it actually concerns an oscillation of size of the involved elementary particles (fig. 8.3). If this oscillation of size lies in the visible frequency range, then we can directly perceive it and it can be used for instance in a light bulb for the purpose of lighting.
In the case of an arc. fluorescent tubes or glow-discharge lamps light is formed as well even without a thermal effect. These glowing phenomena are called luminescence. They can be caused by chemical processes, by friction, by crystallization or by electric fields. As can be shown speedily, in these cases it also concerns an oscillation of size.

If we for that consider the model on the left (fig. 17.9). Between two electrodes there is a non-conducting gas, e.g. air. If now a tension voltage is applied then, under the influence of the electric field, some gas atoms are splitted into positively charged ions and negative electrons (case A) and pulled apart (case B). This process of enlargement of the glow-discharge fast as lightning comes close to that of the thermal oscillation of size and obviously in the same way is perceived as light.
The ionised gas parts are attracted by the unlike poled electrodes and move towards these (case C, D). Whoever takes the effort to count the number of the field lines, will find out that between both electrodes the electric field drops, whereas it at the same time increases at the electrodes. The first thing causes that the process of ionisation is stabilized by itself and a state of equilibrium will result; the resulting current takes a constant value.
The increase of the tension voltage at the electrodes on their behalf causes that an arc remains standing or the glow-discharge lamp glows on, even if the feeding voltage is reduced. For this reason a fluorescent lamp needs a starter, since the network voltage is too small to start the effect of luminescence.

We indeed know, that the gain of light of a fluorescent lamp is at least three times better than that of a light bulb. As a rule a degree of effectiveness in spite of that can't be given, since it merely concerns a comparison measurement in the case of the measurement of the lighting intensity with a luxmeter. It hence can't be excluded, that we already are dealing with an over-unity effect in individual cases and neutrinos are involved in the effect of glowing.
A necessary oscillation of size would be present, as the perceivable noises of open spark gaps prove. Also in the case of lightning the thunder occurs as a consequence of a longitudinal oscillation of size and at the same time a glowing phenomenon occurs as a consequence of materialized neutrinos (fig. 14.11). It almost is obtruding a scientist that here the same principles are at work.

One however hardly can prove that neutrinos are involved in the luminescence, because the configuration is symmetrical with regard to the resonance of neutrinos and just as much particles materialize as anti-particles, which afterwards again annihilate completely as an impulse of light. They hence can't be measured directly. It gets interesting if an asymmetrical arrangement with unipolar character is chosen.

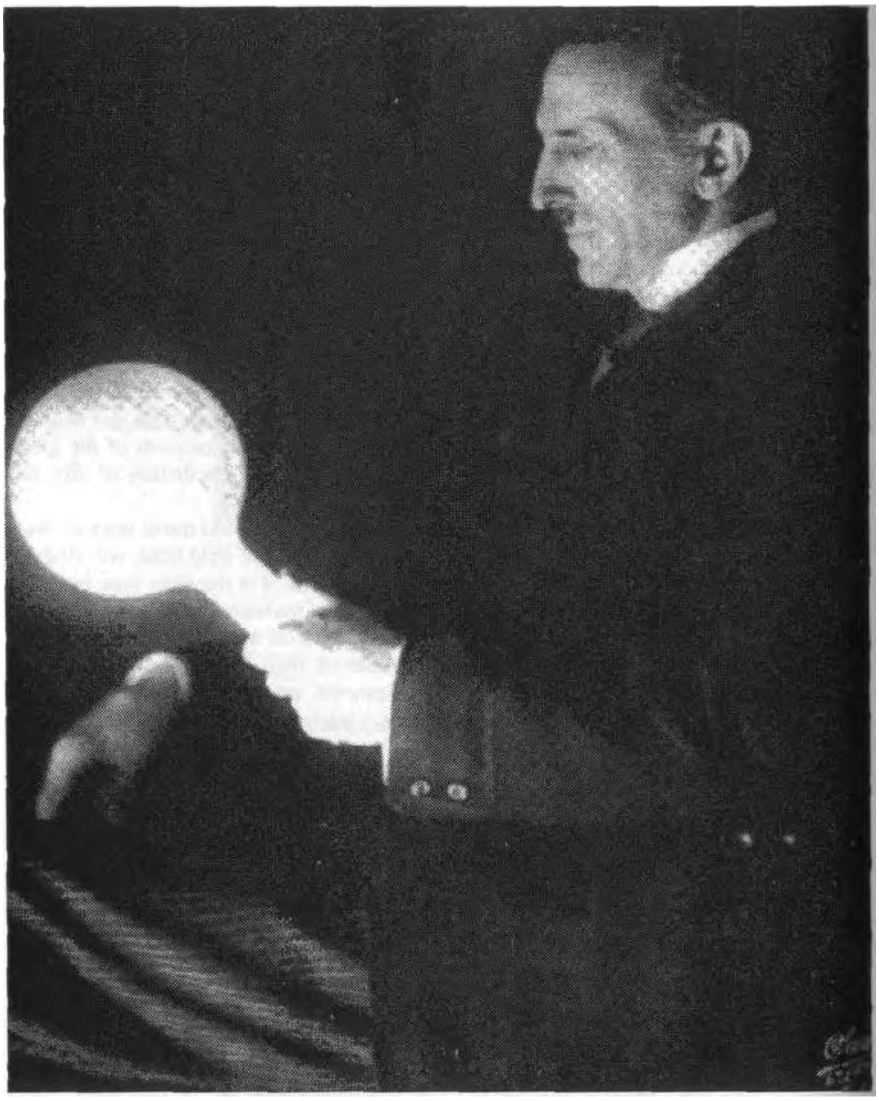

Fig. 17.10: Tesla with a wireless light bulb<i>

<i>: A Sarony-Portrait from 1894, Tesla-Museum, Belgrad

<ii>: R. Hiller, K. Weninger, S. J. Puttermann, B. P. Barber: Effect of Noble Gas Doping in Single-Bubble Sonoluminescence, University of California, Los Angeles, USA, Science, Vol. 266, 14.10.1994, P. 248-250

and: D. Lohse: Wenn sich der Schall in Licht verwandelt, Mechanismus liegt noch im dunkeln; (University of Twente, NL) Mitteilungen der DFG 4/98, S. 19-21

17.10 Phosphorescence and sonoluminescence

Let us first look at the observable after-glowing at fluorescent lamps or at a screen, the phosphorescence. It presupposes a storing effect, and that in accordance with prevailing textbook opinion can be traced back to the excited state of some atoms. Doing so electrons in the atomic hull change from one level of resonance, an instable state of energy, to another orbit, which represents the ground state. Doing so the difference in energy is emitted in the form of light. The process obviously not only takes place as luminescence during the switching on of the excitation voltage, but also as phosphorescence after the switchng off.

An enveloping electron now doesn't fly as a tiny planet around the nucleus, but occupies the entire orbit as an inflated spherical vortex (fig. 5.5). Seen from the outside it is a matter vortex. It however sees the enveloping electrons on the further on the outside lying orbit from the inside and there they are showing as anti-matter vortices. As long as the distance of respect is kept, nothing happens. If however an inner spherical vortex presses to the outside or the outside one to the inside, then the incompatibility of the vortices takes effect and both annihilate under emission of a photon. In this respect the explanations of the vortex model are very helpful.

But now, after this flash of light, two electrons are missing in the atomic hull. The positively charged atomic nucleus never would allow this loss. Replacement has to be fetched and that actually only can stem from the neutrino field. Therefore the enveloping electron doesn't change the orbit immediately, but instead has to wait, until a suitable neutrino passes by, with help of which the game of changing places can be executed. This explains the time delay and gives reasons for the observable after-glowing.

The form of the unipolarly charged sphere (according to fig. 15.8 B) forms almost ideal prerequisites for an interaction with neutrinos and the step-like change of size from one orbit to the next ideal prerequisites for their materialization. The longitudinal wave connected with the change of orbit without doubt can be called a high-frequency sound wave. It however can't be detected because of missing gauges for such high frequencies.

But if we take the frequency down into the range of the ultrasound, which can be handled technically, then effects arise which brilliantly prove this interpretation and the participation of neutrinos in the luminescence.

The not understood phenomenon is called sonoluminescence and at present is researched at numerous high schools primarily for academic interest. The structure is conceivably simple. One takes a ball of glass filled with water and positions at the edge one piezocrystal next to another. Then one with the piezo loudspeakers, operated with the same phase, sounds the whole with ultrasound and see there, the water glows mysteriously! The sound waves change the pressure. Inferring from the observations, during the phase of low pressure small bubbles are formed, which at the following rise of pressure collapse and emit a ultrashort flash of light. There thus takes place an oscillation of size, which leads to the luminescence phenomenon!

So far so good; but such a flash of light is much shorter than the collapsing of the gas bubble lasts. With less than 50 picoseconds it is faster than the otherwise usual atomic transitions of electrons. The whole thing obviously has nothing to do with this kind of luminescence. Also the spectrum doesn't fit and finally the over-unity effect lies at one trillion, the light energy thus is 10^{12} times larger than the energy part which is taken up from the sound wave by every atom[ii]!

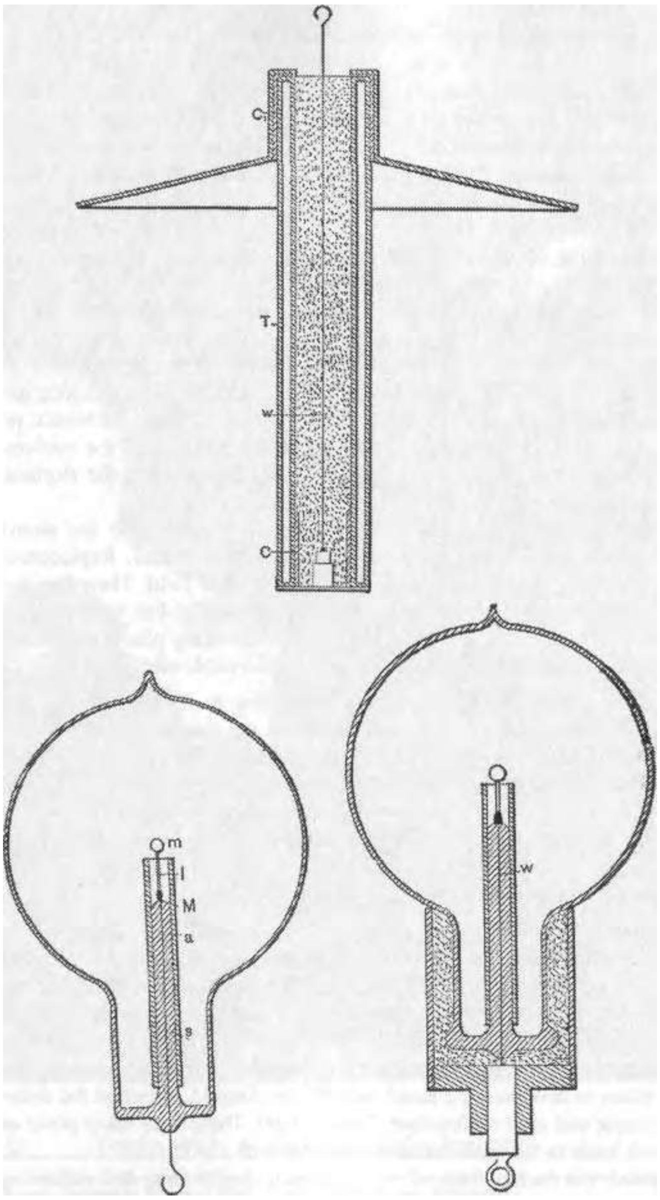

Fig. 17.11: Light bulbs according to plans of Tesla<i>

<i>: N. Tesla: Experiments with Alternate Currents of High Potential and High Frequency, Lindsay Publications Inc, ISBN 0-917914-39-2, Fig 18, P. 70+103

17.11 Discussion concerning the use of the weak interaction

In the case of the luminescence the conditions lie similar like those for lightning. The science of today it is true has some problems of explanation at detail aspects, but again and again manages to protect the ivory tower of physics from collapsing, with supporting auxiliary explanations. But at the latest in the case of an unipolar arrangement, in the case of a ball-lightning or the sonoluminescence, the participation of the neutrinos clearly comes to light, the auxiliary concepts prove to be unusable.

It surely is no accident that the discoverer of the neutrino radiation also was the first, who has experimentally investigated the luminescence. The lamps of Tesla (fig. 17.11) all were without wear and didn't have filaments like those of Edison. They however had to be operated with high tension voltage and relatively high frequency. Both a single-wire and a wireless transmission technique were possible (fig. 9.5 and 9.7). In his laboratory Tesla only needed to hang a fluorescent lamp without any return cable on a wire, then he had light. Famous also are the pictures on which he holds a lamp in his hand, which glows entirely without any connection, but only if he takes it in his hand! (fig. 17.10).

If we again turn through the chapter, then it is remarkable that in most cases, from the transmutation to the luminescence, neutrinos can be used for reason of their resonant interaction which in the proximity can be put equal to the weak interaction. A real materialization however poses an enormous difficulty. If namely the translatoric motion of the space quanta is slowed down, then the risk exists that for balance the rotation increases. But that as well has to be slowed down for a materialization! Only if both processes of slowing down take place, the goal can be reached. In that case the described cooling down effect occurs.

For a measurement technical check it therefore offers to record the slowing down of the translatoric motion and the collecting of the neutrinos by means of the measurement of the radioactive decay of a sample. And to read the materialization as the slowing down of the rotation from a measurement of temperature. In addition are helpful proofs about oscillations of size, glowing phenomena and deviations in the balance sheet of energy. With that it should be able to obtain a system for the exploration of these phenomena with the goal of a practical exploitation of the neutrino radiation for the benefit of humanity.

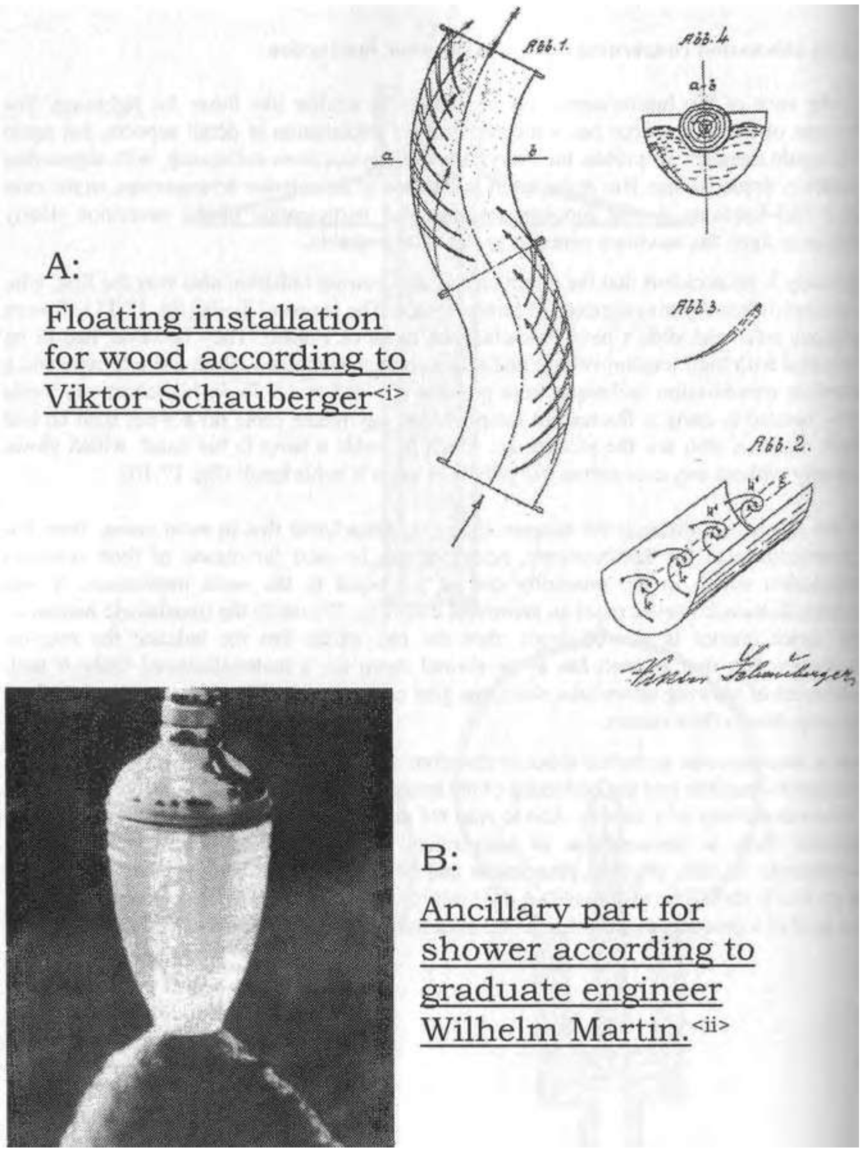
A:
Floating installation for wood according to Viktor Schauberger<i>
Abb.1.
Abb.4
Abb.3
Abb.2.
a
b
B:
Ancillary part for shower according to graduate engineer Wilhelm Martin.<ii>

Fig. 18.1: Mechanical whirling of water

<i>: N. Harthun: Naturgemasse Stromungsfuhrung nach Viktor Schauberger
Analyse einiger seiner Patente und Zitate; MuT Nr. 4, 1980; s.a. Kap. 9.2
<ii>: O.Alexandersson: Lebendes Wasser; W.Ennsthaler Verlag Steyr, 1993, S.156
<iii>: V. Schauberger: Die Entdeckung der Levitationskraft, Implosion 112, S. 39.
<i4>: Do we owe the taste of bubbling spring water to neutrinos?

18. Physical phenomena

In this chapter it on the one hand concerns indirect effects of the resonant interaction and on the other hand effects of gravitation and levitation.

18.1 Water as a catalytic converter for neutrinos

Already the Austrian forester Viktor Schauberger has pointed to glowing phenomena at whirled water. He first had observed such in nature at torrents and waterfalls in the Alps. Later he was capable, to produce and to demonstrate this effect even artificially[iii]. One could speak of hydroluminescence, where still the question would be left open, how it actually functions.

Already in chapters 4.10 and 4.11 has been talked about the special properties of water. It here at first concerned the property of transport of a vortex, which even is capable to bind particles in the vortex, which are heavier than the whirling medium itself. Schauberger had become famous by letting build floating installations, in which not only tree-trunks, but also gravel could be transported down to the valley, without the sides of the channel actually having been touched by the rubble. This phenomenon has been examined and confirmed on the scientific side [i].
Basing on this principle today different devices for processing water are offered and sold, which bind lime, mineral materials and suspended matter in the vortex and in this way prevent deposits in the pipes. The methods, to whirl the water, are however very different. Some whirl the water mechanical, others magnetically and again others electrically.
Here the dipole nature of the H_2O-molecules has an effect. If I turn a water molecule with its electric charge distribution, then from the moving charge a magnetic field results. If thus in the pipe a hydrodynamic flow vortex is produced, then an electric and a magnetic vortex, the potential vortex and the eddy current, are the result (see fig. 4.2). The wanted flow vortices vice versa also can be obtained, by guiding bubbling whirled water past permanent magnets, or by feeding in alternating magnetic fields with the help of coils, or finally by working with pulsed electric potentials. Each of these systems, operated passively or actively, has its specific advantages and disadvantages (fig. 18.1 B and C).

It can be assumed that with the vortices also the water quality is changed. As a rule the content of colloids increases, due to which the surface tension falls. In the colloids negative ions are bound, for which reason also the electric conductivity decreases. Finally the content of oxygen increases and that actually only can come from an electrolysis. The neutrinos thereby are suspected [i4].

As an extreme dielectric medium water favours the formation of potential vortices, which immediately after their formation contract swiftly. This oscillation of size of the electrically charged potential vortices makes possible actually an interaction with neutrinos, and that on the one hand has as a consequence the water splitting and the increase of the content of oxygen in the water and on the other hand the above depicted, observable glowing phenomenon, the hydroluminescence.

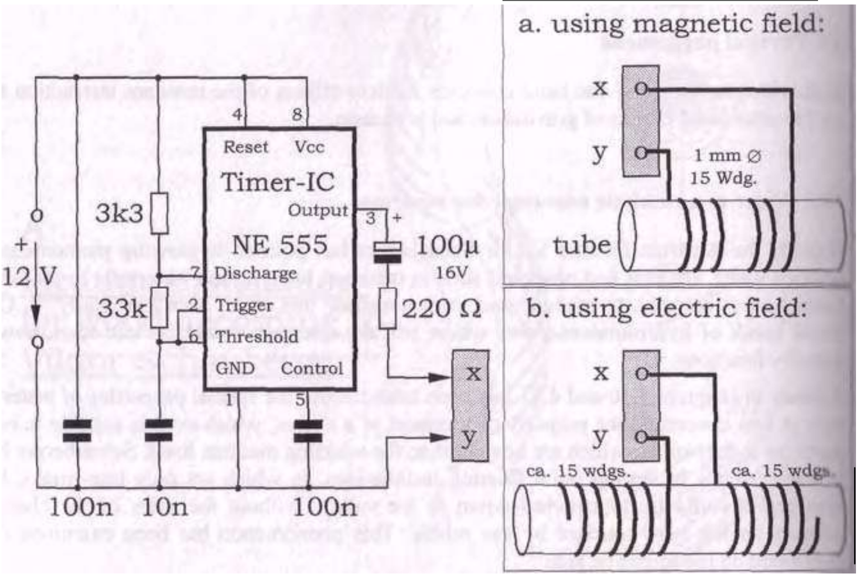
4
8
Reset
Vcc
Timer-IC
Output
NE 555
3
+
100µ
16V
3k3
+
12 V
7 Discharge
2 Trigger
6 Threshold
GND
Control
33k
220 Ω
1
5
x
y
100n
10n
100n
a. using magnetic field:
x
y
1 mm Ø
15 Wdg.
tube
b. using electric field:
x
y
ca. 15 wdgs.
ca. 15 wdgs.

Fig. 18.1 C: Wiring diagram for a simple water processor with pulsed field (2 kHz rectangular signal)
a: for magnetic field excitation
b: for electric field excitation

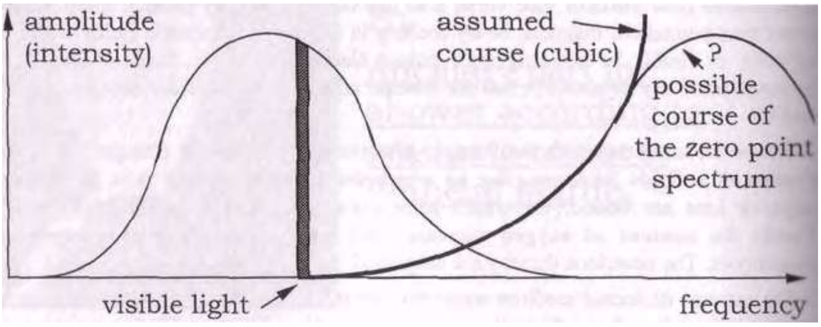

Fig. 18.2: The spectrum of frequencies of light and of the zero point radiation.<i>

<i>: A. Waser: The Puzzling Nature, AWVerlag Luzern 1996, p. 88

To complete in fig. 18.1 C the wiring diagram of a simple water processor is shown, which works with pulsed electric (b) or magnetic (a) fields.

If we temporarily again leave the theme water and take along the notion, that it must be attributed a great importance in connexion with the interaction and the materialization of neutrinos

18.2 Background and zero point radiation

Whathappens really, if cold matter or interstellar gas molecules are hit or touched by the everywhere present neutrinos? Then it can be expected that oscillations are being stimulated and as a result the temperature slightly increases. If we thus at a space flight hold a thermometer out of the shuttle window, then we will measure everywhere such a remaining temperature of a few degrees Kelvin, which is called cosmic background radiation.

By popular scientific small talkers the background radiation is called remnant of a so-called Big Bang. It even is misinterpreted as evidence for the Big Bang. Even if a certain value of entertainment can't be denied, the Big Bang from a physical view until now only has raised questions and contradictions.

If we stay at the fact that oscillating neutrinos depending on their radiation density and velocity of propagation produce a thermal oscillation, which can be detected as a slightly increased lemperature. To measure that we don't need immediately to undertake a space flight. We also can install the thermometer in an artificially produced vacuum. In vacuum physics out of ignorance of the relations the neutrino radiation then is called zero point radiation.

In the case of the neutrinos it of course concerns an oscillation around a mean value, which for a symmetrical form of oscillation has the value zero. The term nevertheless is misleading and chosen very unfortunate. After all we don't speak of zero point current in the case of alternating current!

Since every supernova and every black hole emits neutrinos and correspondingly in physical experiments until now no preferred direction could be determined for the zero point radiation, it is taken as homogeneous and isotropic in space. From the Lorenz invariance again is inferred a cubic course of the zero point spectrum, an increase of the radiation intensity with the third power of the frequency. With this assumption the radiation density of the vacuum however strives with increasing frequency towards infinity. Here an error has to be present in the considerations!

I proceed from the assumption that in the case of the spectrum of the neutrino radiation it rather concerns a spectrum of resonance, the maximum value of which lies above the frequency range measurable with devices of today. It of course would be very important to find out, where the radiation maximum lies, but without being able to measure it that lies unfortunately in the range of pure speculations (fig. 18.2).

Half the wave length in any case should be tuned to the length of the antennas of the energy centres of the cells, the mitochondria and the chloroplasts which are capable of photosynthesis. In both cases the head diameter of the ATPhases amounts to approx. 5 nanometer. It would be obvious, if the resonance point would lie here. But it could as well lie at still smaller wave lengths in the domain of molecular dimensions.

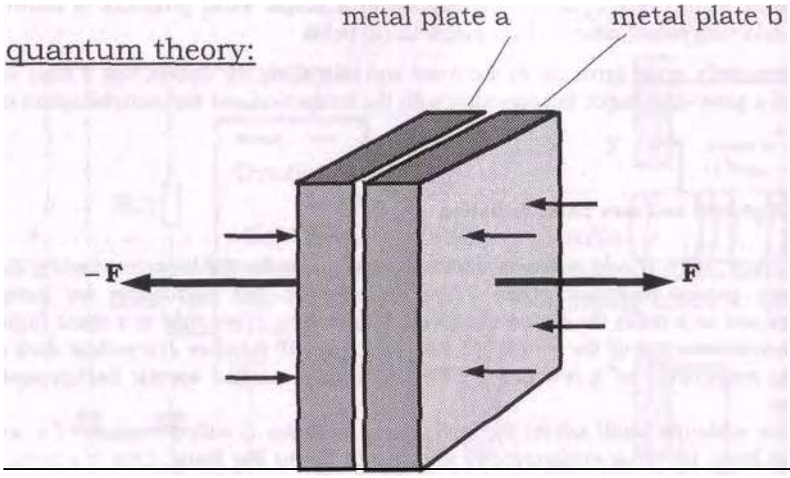
metal plate a
metal plate b
quantum theory:
– F
F

outside cause<i:>: pressure force by bombardment of quanta?

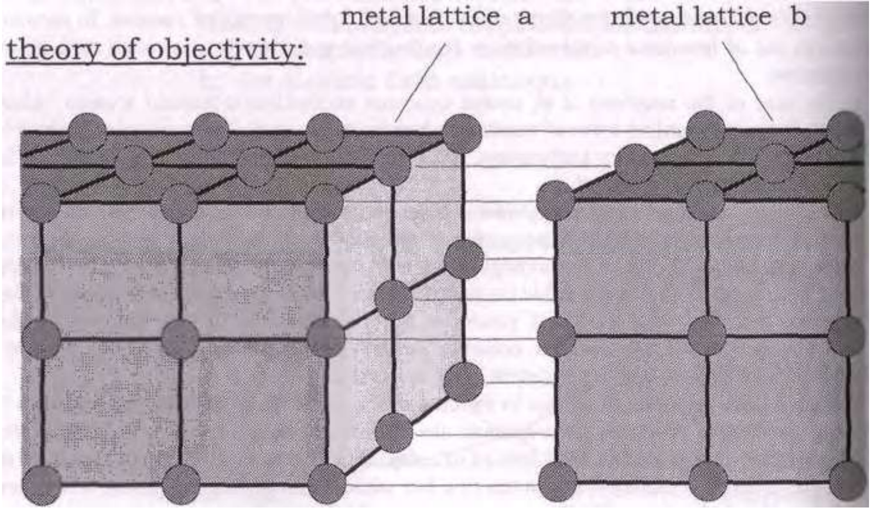
metal lattice a
metal lattice b
theory of objectivity:

inside cause<i>: levitational force by resonance effects?

Fig. 18.3: The Casimir effect

18.3 The Casimir effect

If this zero point radiation actually would be the cause of the Casimir effect, which is generally assumed today, then neutrinos would have to be involved, if the zero point raditation would be to put equal to the neutrino radiation. The effect even is quoted as evidence, but I have my doubts.

The experiment is relatively simple and functions also in a vacuum. Two absolutely plane and smooth polished metal plates are placed very close to each other. Doing so the distance should only amount to one thousandth till one millionth of a millimetre. If the force, with which both plates are attracting mutually, is measured then it exceeds by far the gravitational force. We are dealing with an unknown force, neither electric nor magnetic.
Because the force of attraction still arises near the absolute zero of temperature, some make the zero point radiation responsible. In our translation that would be tantamount to the neutrinos exerting a pressure from the outside on the metal plates, if need be with their small rest mass, as far as this should not be based on an offset-error of the detecting devices. For that the plates would have to damp the neutrino radiation effectively and by mutual shielding reduce the radiation pressure from the split. The question is asked, if the from the outside hitting bombardment of quanta, as it is called, actually can exert a pressure which would be compatible with the nature of the neutrino radiation.

If we now try an entirely other interpretation, which does without the postulating and designing of new force effects. Now the from the unified theory won interactions according to table 15.1 form the basis. In the Casimir effect, as said, neither magnetic nor electric forces are involved, so that open field lines and the corresponding force effects (1 till 4) are ruled out. It has to be the work of closed field lines and that in the static case we know as gravitation (5). Now the mass of the metal plates is too small, as that an acceptable force of attraction of masses could result.
The measurable force is much larger, even if it according to its nature could be a gravitational force. Here oscillations, as they actually take place in the hull of the metal atoms, seem to play a role. If between the oscillations of the two plates standing opposite occurs resonance, then a levitation is possible (6), which describes an oscillating interaction, produced by closed magnetic field lines<i>.
At extremely low-frequency signals this interaction is known as gravitational wave and object extravagant and costly physical experiments.

<i>: The question, which interpretation is correct and which one should be rejected, could be verified as follows: at first the force between two plates of a certain metal and afterwards that of another metal has to be measured. If in a third experiment a plate from the first and one from the second experiment are brought together and the force effect goes back measurably, then the quantum physical interpretation would be wrong clearly. With that would have been shown that here resonance effects are used, which presuppose an identical metal lattice structure of both plates.
If the force effect however doesn't go back, then both statements are possible, then one has to think up another experiment.

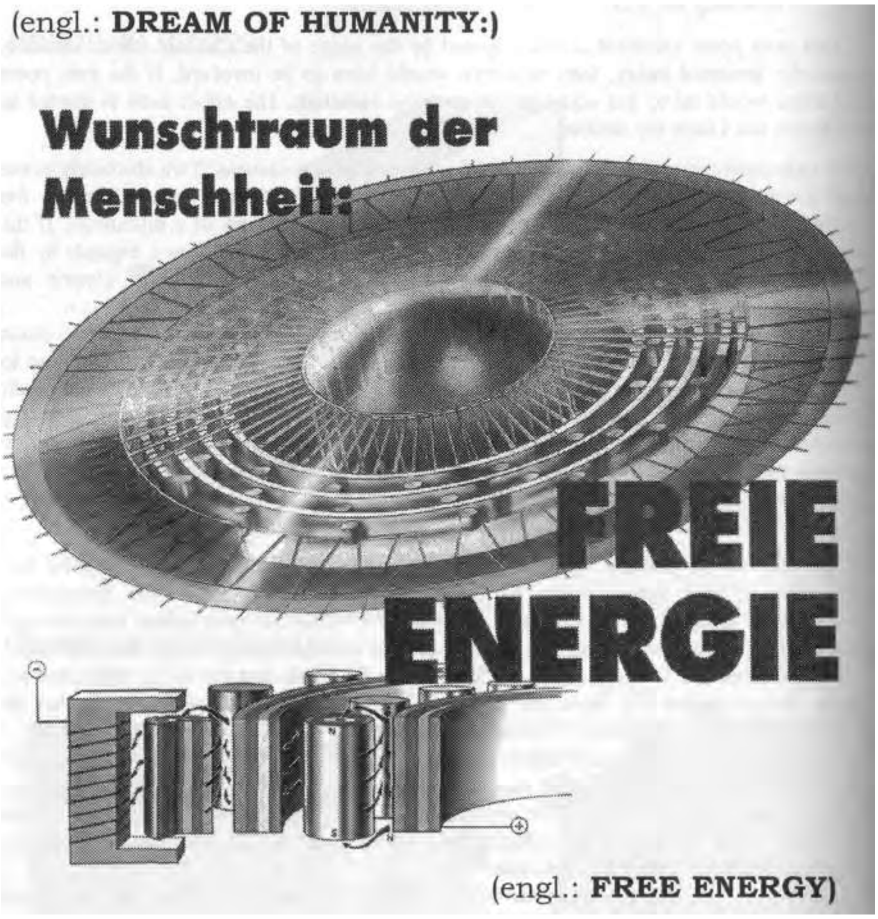
(engl.: DREAM OF HUMANITY:)
Wunschtraum der Menschheit:
FREIE ENERGIE
(engl.: FREE ENERGY)

Fig. 18.4: The flying disc of Searl on a german title page.<i>

<i>: Special 7, Raum 8s Zeit Dokumentation aus dem EHLERS Verlag

18.4 Effects under high tension

It crackles in the laboratory of high-tension. The air smells electrically charged. Only in a corner someone sits and waits with much perseverance and patience, until something happens. Then he lets thunder down lightnings of several millions of Volts on a small sample. It flashes and crashes and still nothing happens.

No, it doesn't concern Tesla! Tesla has worked purposeful. In addition has Tesla only called himself inventor. But the man in the corner feels himself a discoverer and as it is proper for a discoverer, he has immediately named the effect after himself. As a real American he lets market himself and his discovery by the media and a video tape.

Doing so he obviously doesn't know at all, what he wants to have discovered, of what this Hutchison effect actually consists. It perhaps is an event of pure chance, for which he sometimes has to wait hours. If it very suddenly would occur, then it is a real potpourri of all already discussed effects: metal spoons are bending (fig. 14.13), massif steel rods are breaking, light effects are being observed (chapter 17.9 PP), water starts to dance and to cook (chapter 18.1), without getting hot and finally do some samples take off and fly crosswise through the room. He then speaks of anti-gravitation.

In the video tape can be seen, how he brings about a physical length contraction as a result of a locally produced field concentration with two as an interferometer configured Tesla coils. In a bottle for instance the air volume is changed. This confirms the correctness of the here presented theory of objectivity (chapter 6.6 PP).

In any case no-one is amazed more about the results of chance than John Hutchison himself. Overwhelmed by the magic, the charm of the effects and the feeling to be able to move freely outside the trodden out ways of physics, he still sits in his corner with the video camera and waits, until finally something happens again.

There is worked with tension voltages, which are considerably higher than 511 kV, the calculated tension voltage of their own of the charge carriers (fig. 7.1). One thus by no means has to be amazed, if under the influence of several millions of Volt electrons are taken apart and metallic objects are breaking or bending.

The extreme tension voltage and field change we owe an oscillation of length and of size, which spontaneously can lead to an interaction with neutrinos. Materialized neutrinos again are responsible for glowing phenomena and for the electrolysis of water, which under the impression of the rising gas bubbles seems to dance and to cook. For my readers and participants of the energy technical seminar therefore models of explanation are available.

One effect we still haven't discussed and analysed more in detail: the antigravitation resp. the levitation. As long as no reproducible field conditions are present and the chunks are accelerated in any arbitrary direction, it will be difficult to understand the physical course of a levitation.

In chapter 9.3 we already had become acquainted with and discussed a levitation device with the flying disc of John Searl. for which the field conditions are comprehensible. This time we should have a somewhat closer look at the flying device (fig. 18.4)<i>.

<i>: Some reports can be found in Special 7 ,,Freie Energie", Raum & Zeit Dokumentation aus dem EHLERS Verlag, Seite: 141 bis 157 und 174 bis 185

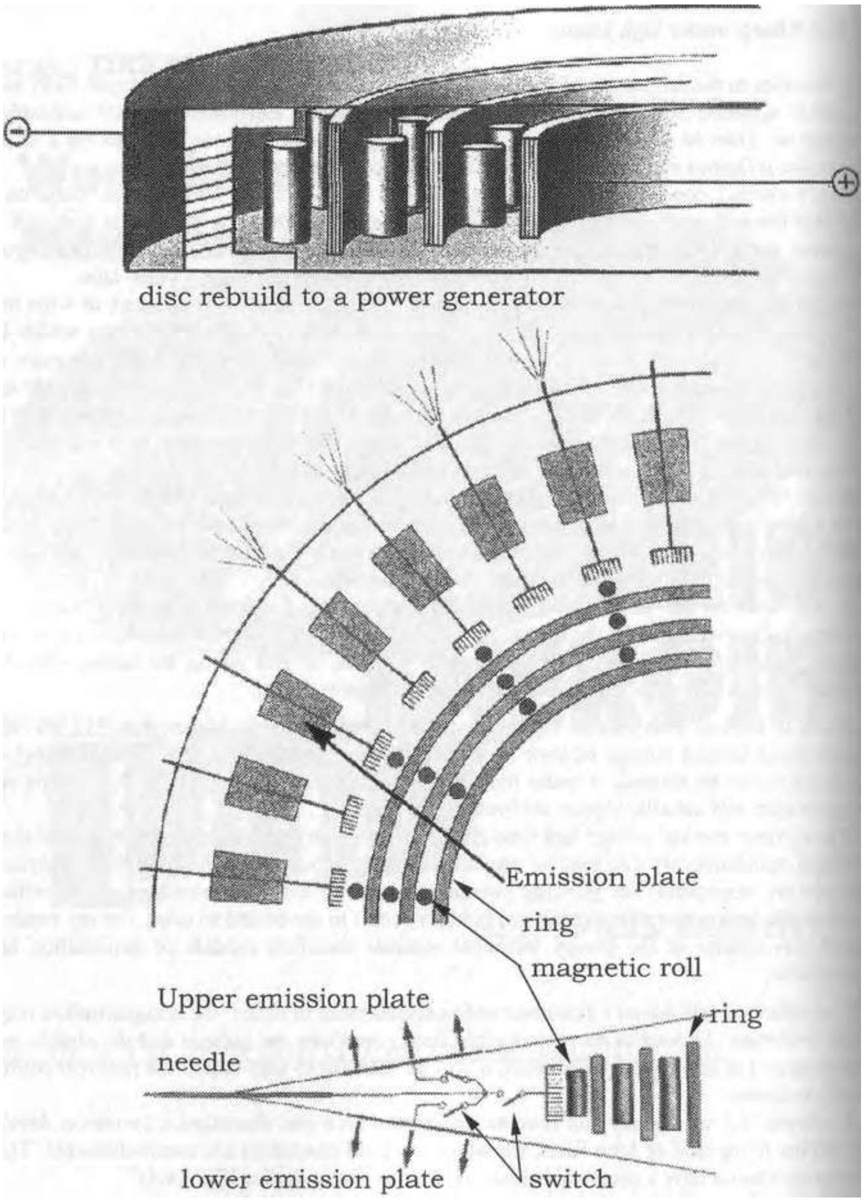
disc rebuild to a power generator
−
+
Emission plate
ring
magnetic roll
Upper emission plate
ring
needle
lower emission plate
switch

Fig. 18.5: Concerning the controlling of the flying discs according to Searl.<i>

<i>: H.Schneider, H.Watt: Dem Searl-Effekt auf der Spur, Special 7, Raum & Zeit Dokumentation aus dem EHLERS Verlag, Sauerlach, 1. Aufl. (1994), S. 183

18.5 Flying discs

The Searl-disc can be calculated relatively good with the Faraday equation concerning unipolar induction. To estimate the order of magnitude, we just suppose the roller magnets produce an induction of one Tesla and the radius is one meter, then between the centre of the flying disc and its edge 511 kV is applied, if the revolutions per minute has reached the value of 80000 revolutions per second. As soon as the revolutions per minute is reached, the neutrino conversion can start.
The at the edge of the disc arising corona, which consists of individual electrostatic discharge flashes, takes care for the necessary dU/dt, by every blow and every spark is drawing the potential for a short time towards ground. The tension voltage jumps in swift order between values of above 511 kV and zero Volt to and fro. In connection with the already discussed unipolar field configuration neutrinos are being collected in this way.

A technological challenge represents the layered construction of the roller magnets and the rings, which have to withstand extreme centrifugal forces (fig. 18.5). Friction however is not a theme, since on the one hand the air inside the disc ionises and a vacuum is formed and on the other hand the air split between rollers and discs increases by the field dependent contraction of the metal parts.

Actually one in this case by no means can speak of antigravitation, because the gravitation isn't really vanishing. The disc even in flight still is heavy tons. Here merely a resonant interaction is built up which is larger than the gravitational pull of the earth. The disc is attracted by the cosmic source of neutrinos, with which it has built up the resonance.
Exactly there it will fly!

The by Searl designed controlling<i> in my opinion actually can't have functioned satisfactorily. He uses emission plates, as he calls them, which alternatively are switched on by means of switches and are able to form electrostatic forces with some air charge carriers. In reality the drive probably is comparable to that of a sailing ship, for which the wind always blows from one direction. The ship isn't pushed by the wind, as one erroneously could think, but rather by the under pressure behind the sail pulled forward. Without steering facility the object always is driven in direction of the drain. The sailor would say, the ship without helmsman drifts towards lee.
Searl in this way has lost all flying discs which were started. By using solar neutrinos they presumably have fallen into the sun and burnt. The controlling should function analogous to a sail, then one would stand a chance, by "traversing" against the "wind", to sometime again come back to the starting-point.

<i>: H. Schneider, H. Watt: Dem Searl-Effekt auf der Spur, Special 7, Raum & Zeit Dokumentation aus dem EHLERS Verlag, 1. Aufl., Seite: 183

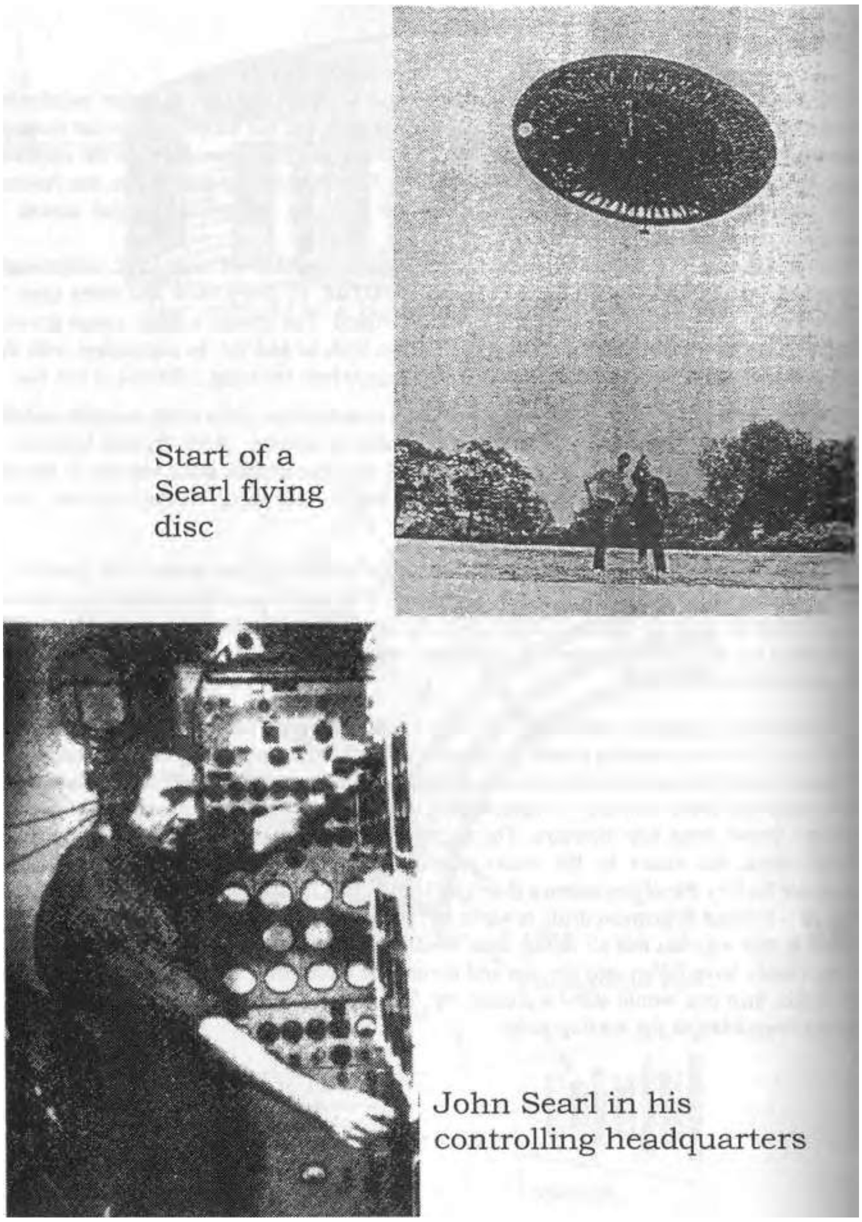
Start of a Searl flying disc
John Searl in his controlling headquarters

Fig. 18.6: The remote controlled flying discs of John Searl.[i]

---

[i]: Special 7, Raum & Zeit Dokumentation aus dem EHLERS Verlag, S. 152

18.6 Propulsion with neutrinos

In search for an ideal propulsion system for spacecrafts the flying discs of John Searl still have demonstrated a further possibility. After the start occasionally were left strange particles on the ground, which hadn't be there before. The flying device obviously had materialized them and dropped them at the start.

With this materialization of neutrinos the chance is showing of a very efficiently working recoil propulsion. If namely the neutrinos are converted into matter then they by that gain back their rest mass. If this takes place in flight then the materialized particles also bring along kinetic energy. It then works as ajet engine for which the direction of the jet can be directed and in that way the vehicle can be controlled comfortable.

In the case of the Searl-disc the materialization of neutrinos rather happens as a not understood side effect. In the case of a systematic use however the principle will show a characteristic property. Instead of a vapour trail a beam of light will shine out of the flying object in the direction of the emitted particles, but it will break off abruptly after a cartain distance.

To blame is the part, which has become anti-matter, which it is true as well contributes to the recoil, but simultaneously annihilates with incompatible particles of matter under emission of particles of light. If in some distance all antiparticles are used up then also the beam of light comes to an end.

It here concerns necessary properties of a corresponding propulsion technology for spacecrafts, which don't have to drag their propelling energy along with them. Our space technicians finally could handle another than the "hammer throwing method", in which case the "hammer" by means of terrible fuel consumption after a phase of acceleration only staggers uncontrollable through space under the influence of its inertia.

With a neutrino propulsion on the other hand one at any time can accelerate or brake. It will strongly influence the field around itself, so that can be reckoned that for every acceleration a field dependent change of size should be perceptible by an outside observer.

If therefore a corresponding flying device is accelerating then it will suddenly become smaller and that then looks so as if it would have moved away with a jump without temporal delay, but that isn't the case at all.

The jerky movements only would be a result of the perception with our eyes by means of the propagation of the light. Since the passengers are exposed to the same field, they change their body size along with that of the vehicle. They actually notice nothing of an apparently infinite acceleration, which only observers on the outside would observe and which indeed no living being could stand.

Nowhere the explanation crisis is larger than in space travel! The theory of objectivity represents a real help, because it perhaps as the only one puts us in a position to conceive and understand not understood observations as necessary technological consequences.

Only by uncovering parascientific as purely physical phenomena man of today will be able to free himself from the constraints of magic and his own illusions.

Is the until now discussed cases resonant or other force effects are being used, whereas the gravitation remains unchanged. Closed field lines after all cannot be influenced, normally at least. Perhaps an indirect possibility exists to have an effect on the gravitation?

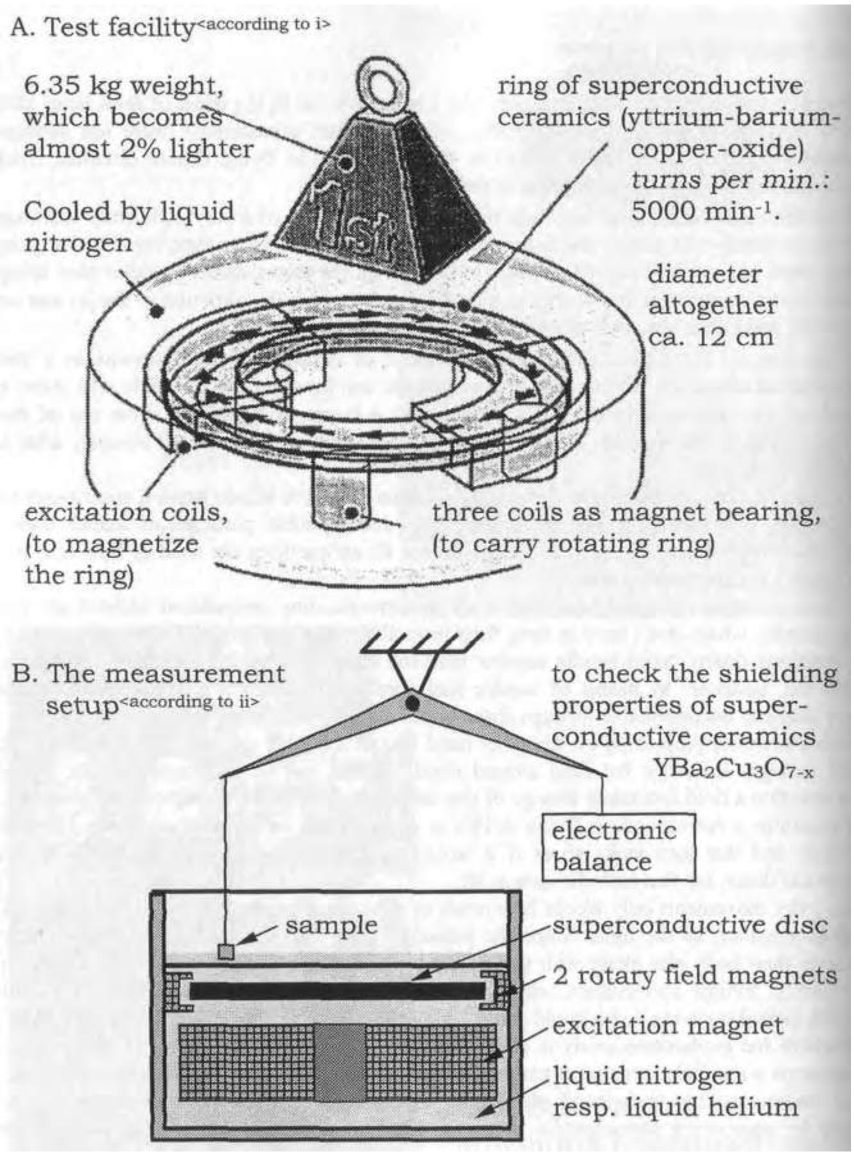
A. Test facility<according to i>
6.35 kg weight, which becomes almost 2% lighter
Cooled by liquid nitrogen
ring of superconductive ceramics (yttrium-barium-copper-oxide)
turns per min.: 5000 min-1.
diameter altogether ca. 12 cm
1st
excitation coils, (to magnetize the ring)
three coils as magnet bearing, (to carry rotating ring)
B. The measurement setup<according to ii>
to check the shielding properties of super-conductive ceramics YBa2Cu3O7-x
electronic balance
sample
superconductive disc
2 rotary field magnets
excitation magnet
liquid nitrogen resp. liquid helium

Fig. 18.7: Experiment for the manipulation of the gravitational pull of the earth.

<i>: Schneider: Anti-Gravitation im Labor nachgewiesen?, NET-Journal Januar 1997, S. 14,15; s.a. Internet: http://www.keelynet.com/gravity/fingrav.htm

<ii>: E. Podkletnov, R. Nieminen: A possibility of gravitational force shielding by bulk $YBa_2Cu_3O_7$-x Superconductor, Physika C 203 (1992), P. 441-444

<iii>: Force field Implications of Anti-Gravity, The Journal of Ideas, Art.191, 7.9.95

18.7 Antigravitation or levitation

The night already had fallen over Finland, but at the University of Tampere the light still was on. Then a scientist put his beardy head through laboratory door and with a "hello folks" blew the smoke of his pipe over the cryostat for the examination of superconductive materials. The scientists were speechless, because the smoke seemed to hit an invisible wall and was drawn upwards in direction ceiling almost at a right angle.

After having aroused the curiosity, Dr. Eugene Podkletnov also has held other materials above his rotating disc and had to discover that these as well lose weight, and that he actually influences the gravitation with his experiment. Even the air pressure at the corresponding place in a lying above floor was smaller[i].

With a superconductive disc of 30 cm diameter cooled by liquid nitrogen a reduction of gravity for 2% to 4% can be obtained, if the disc rotates with more than 5000 revolutions per minute. With the revolutions per minute also the effect increases, but it is independent from the excitating field[ii]. Finally the fields of the superconductive currents in the ring remain existing unchanged even after switching off the excitation and it only depends on these fields.

The shielding effect of the arrangement on electromagnetic fields already had been known before and should be examined closer in the laboratory. One only was surprised that the gravitation could be shielded as well, that obviously both interactions are related. According to the prediction of the theory of objectivity the closed magnetic field lines gravitate and the field components directed towards the centre of the earth cause the measurable force of weight. By the very strong superconductive fields obviously field overlaps and possibly a local driving out of field occurs, like one has long known for eddy currents (Meissner effect).

With that the earth gravitational field of course is not cancelled, but merely spatially moved, or it partly has changed its direction. If only a small part of field pointers turns out of their orientation towards the centre of the earth, then at that point the force of weight is reduced to a corresponding extent. The theory of objectivity requires that!

By the way reports exist, according to which even without rotation a weight reduction can occur for superconductive rings, and others, where is worked even entirely without superconduction. As core material for the ring ferrite or a strong permanent magnet is used. In that case depends on the circumstance that the ring-like coil is operated in self-resonance and always in the right moment again is excited anew by induction. The ring core then stores the fields up over time, exactly as the superconductor.

Stricktly speaking it are the atomic nuclei and in the end all elementary particles, which are aligning in the high field. Dr. Ning Li wants to artificially produce gravity, by directly influencing the quanta. With that she comes the phenomenon very close[iii]. The possibilities of this effect nevertheless are very limited, since two percent more or less are not exactly much, and the dream of the complete cancelling of gravity or even inertia possibly stays just a dream. It stricktly speaking just concerns influencing the gravitation and not antigravitation!

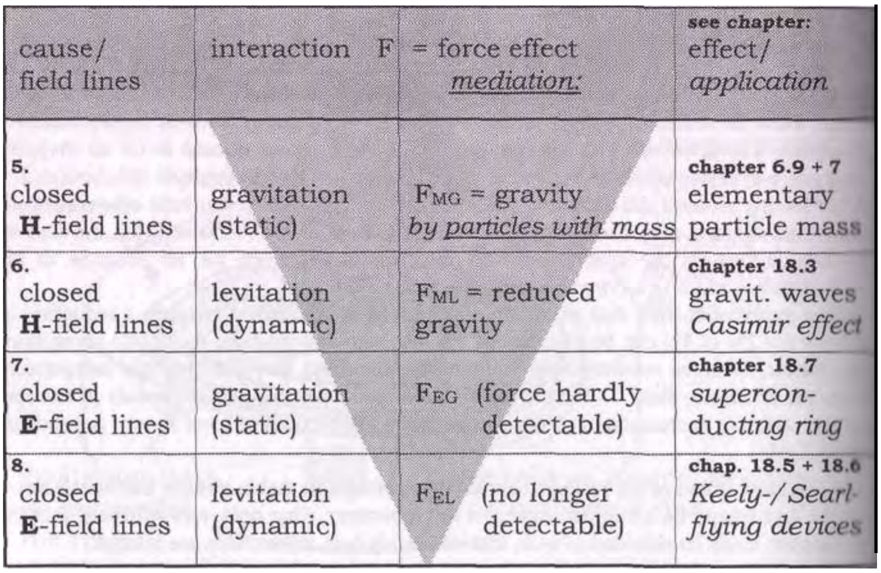

cause/ field lines	interaction F	= force effect *mediation:*	**see chapter:** effect/ *application*
5. closed **H**-field lines	gravitation (static)	F_{MG} = gravity *by particles with mass*	**chapter 6.9 + 7** elementary particle mass
6. closed **H**-field lines	levitation (dynamic)	F_{ML} = reduced gravity	**chapter 18.3** gravit. waves *Casimir effect*
7. closed **E**-field lines	gravitation (static)	F_{EG} (force hardly detectable)	**chapter 18.7** *supercon-* *ducting ring*
8. closed **E**-field lines	levitation (dynamic)	F_{EL} (no longer detectable)	**chap. 18.5 + 18.6** *Keely-/Searl-* *flying devices*

Table 18.8 A: The interactions of closed field lines.<i>

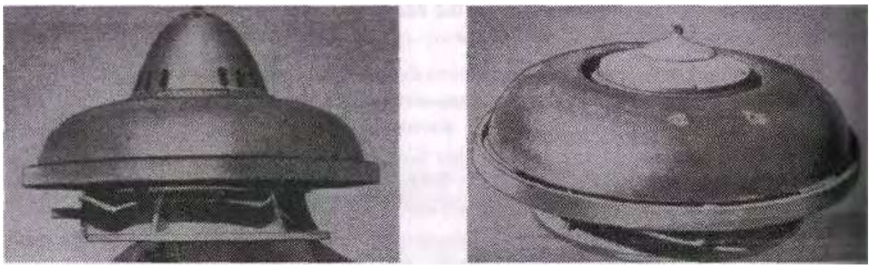

Fig. 18.8 B: Test models of the "flying saucer" according to Viktor Schauberger.<ii>

<i>: according to fig. 15.5, cases 5 to 8

<ii>: O.Alexandersson: Lebendes Wasser; W.Ennsthaler Verlag Steyr, 1993, S.103

18.8 Discussion concerning the effect of closed field lines

If we now direct our eyes exclusively to the lower half of fig. 15.5, in which the consequences of closed field lines are listed. Without the possibility of a direct focussing like for open field lines, the caused force effects turn out extremely small. With that they are less suitable for an use of free energy. The considerations in that context rather aim at technologies to reduce weight, like they would be worth striving for for flying devices.
If we allege that here in the same way the force effects of magnetic H-fields exceed those of electric E-fields for three to four powers of ten. Then it will hardly play a role for the force of weight of a body with or without additional electric charge, if its E-field lines are closed as well or are open. The increase in weight of an uncharged body only will have an effect in the third till fourth place after the decimal point. Seen so in the case of the gravitation it primarily concerns an effect of the magnetic field, more strictly speaking the effect of closed magnetic field lines (chapter 6.9).
If one succeeds in influencing these magnetic fields for instance by the influence of extremely strong fields of a superconductor, then also the gravitation will change, as has been shown (chapter 18.7). Let's imagine, in the case of a systematic procedure we would succeed in a perfect influencing, for which no component of the H-field lines points into the direction of the centre of the earth anymore, then an uncharged body would only have one tenthousandth of its original weight, whereas a charged body actually would weigh nothing anymore. This state of the weightlessness supposes that all E-field pointers point into the direction of the centre of the earth or diametrically in the opposite direction, since E- and H-field are standing perpendicular to each other. Unfortunately such a field distribution technically hardly can be realized and so the ,,flying carpet" furthermore remains reserved to the fairytales.
An effective reduction of weight of planes and other flying objects however seems by all means feasible, and so slowly the number of research facilities increases, which more or less officialy have a critical look at the cancelling of gravity and the levitation. The theories on which they are basing, however often sound very bizarre and moreover are completely unphysical. Maybe a look at the distribution of field lines, as proposed here, helps to get further.

In the case of the levitation, which occurs strictly speaking only in the case of oscillation for closed field lines, resonance again plays the crucial role for the coming about of the interaction. In connection with neutrinos the resonance can serve less the collecting; we rather need it for the materializing, for the production of mass, charge and energy.
For the artificial production of a levitation either a mechanical oscillation in the atomic or molecular domain is needed, like for instance is produced by a rotating water molecule, or a resonant oscillation of size takes place by use of electrostrictive or magnetostrictive materials, like piezocrystals or oscillating quartzes.
In this context surely is of great importance that also the temperature could be identified as a particle immanent oscillation of size (chapter 8.3). That's why besides a cold materialization or cold fusion also exists, at least theoretically, the possibility of a hot materializatin or hot fusion.

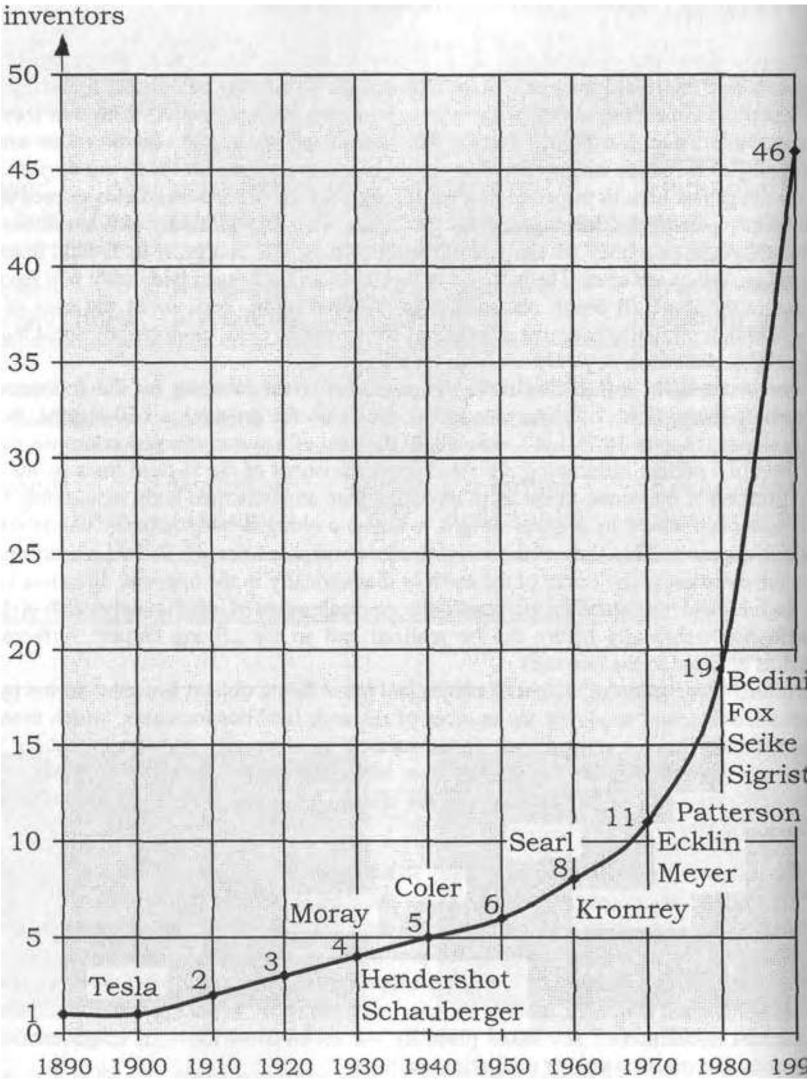

Fig. 18.9: The increasing number of free energy inventors.<i>

<i>: only the inventors mentioned in the text are entered. Literature for that: A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag (1994), S. 77

18.9 Concerning the determination of the stand of today's energy technique

An analogy to the production of matter forms the production of light. At high temperatures as is well-known is formed thermal light and at low temperatures cold light, the discussed luminescence. An oscillation of size of the light source in all cases of the formation of light is the cause, be it as an electromagnetic wave or as photon vortex, be it in whirled water or as sonoluminescence. We at last see ourselves put in a position to understand the various light phenomena as something which is related.

The call for new energy carriers and an ecologically compatible energy technology can't be ignored anymore. The number of inventors at present increases fast (fig. 18.9). It is legitimate and worth recommending, to observe for this purpose first of all nature, how it solves its energy problem. Doing so we should realize that in contrast to the technology of today no combustion and no explosion takes place, but rather principles of a cold fusion and of an implosion as a result of contracting potential vortices are being used. Actually nature may point the way out of the energy technical dead end. We only have to accept the offer, be open and show being ready, to invest in the development of methods in accordance with nature.

Directed at the address of the distributors of supporting funds for energy research the claim in plain English reads: to grant no money anymore for ecologically harmful concepts, like nuclear fission or for not understood and not realizable developments like hot fusion. to immediately stop the fruitless works and to provide the money for an ecologically compatible energy research.

Immense saving potentials moreover are found in all domains of physics, where instead of costly experiments just as well the things could be calculated with an usable theory. To check the theory then only few experiments would be necessary. After all all important discoveries have come about in this manner. Nobody should think, he had understood a matter, once he has filled himself up with sufficiently much measurement data.

The root of the evil lies already in the education, where students of physics want to understand everything and therefore are trying hard to grasp with their view and hands for everything. Abstract thinking or mathematical derivations however are too uncomfortable for most; they have the erroneous opinion, understanding (German: begreifen) would have to do more with grasping (German: greifen), and for that the head after all can't be used.

One reason for this development can be seen in the relativistic point of view of Albert Einstein, who proceeds from the assumption of a subjective observability and has raised the relativity between a physical principle and the observation to the basis of physical thinking. The arrogant motto prevails: what I can't observe with my sense organs or register with corresponding gauges also isn't physics, but esoterics or parascience.

But if we want an ecologically compatible technology, then this can't be reached with this point of view. Then we only hinder ourselves with our own arrogance and intolerance. We have to leave the erroneous quantum physical way and again learn to think abstractly by taking up tried and tested principles of classical physics. An objective point of view forces us to register the phenomena, which lie outside the observable range, with mathematical MEANS (see chapter 13). Only if we have learned that, we will understand and realize the true relations in physics!

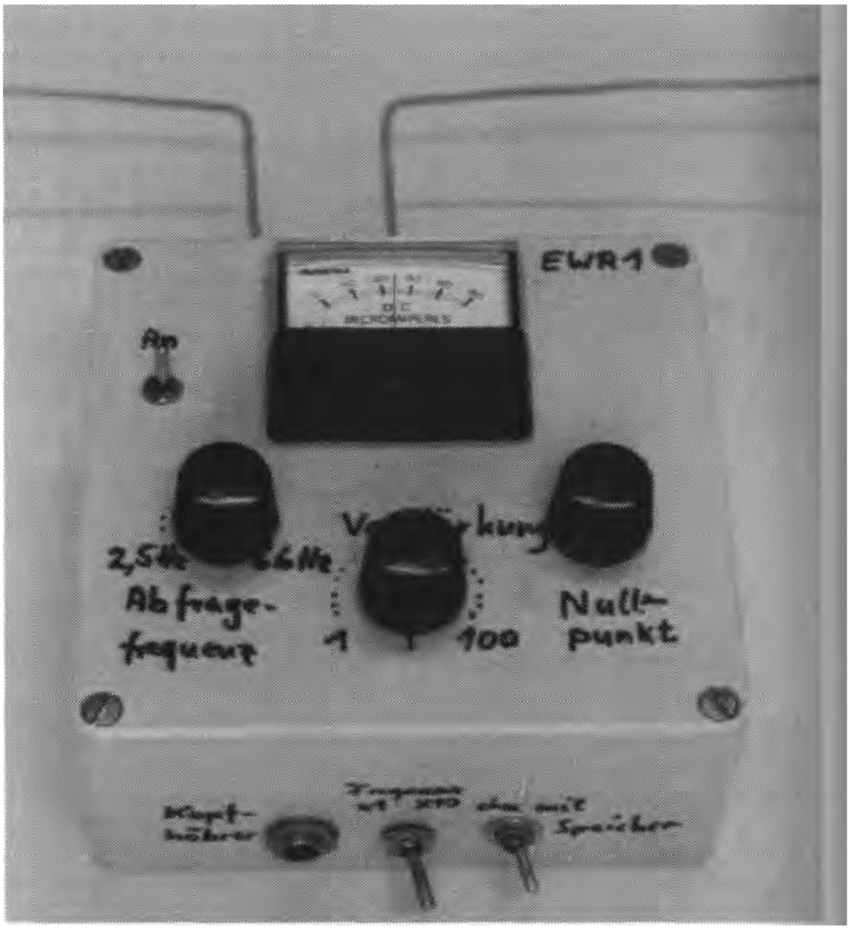
EWA1
An
DC
MICROAMPERES
2,5kHz
Abfrage-
frequenz
1
100
Null-
punkt
Kopf-
hörer
Speicher

Fig. 19.1: A self-built charge amplifier.

19. Own experiments

The opposite of the theoretical considerations form practical experiments. If we want to register and understand physical and natural scientific phenomena, then we shouldn't entirely forget the aspect of handicraft. Doing so it can't be avoided that it goes on very personally, if about the laboratory work is being reported, not even then, if the experimenter is at pains to exclude every influencing of the carrying out of the experiment and the result. After all does anyone have to execute the experiments himself, since a good sceptic anyhow only believes, what he has observed with his own eyes.

19.1 Concerning the rating of experimental physics

There already have been scientists, who have requested to reject a measuring technical determined result for the case that textbook physics doesn't produce it. This attitude even today still is taken by an uncounted community of believers of science, entirely after the motto: "Nature kindly has to fit in with the dogmas of theoretical physics!"

This wrong thinking is reflected in the scientific journals, which reject to print discussions. Controlled by a tester board, which only has the task to prevent dissenters from the publication of their ideas, they are the sad proof, how widespread this erroneous attitude today is in the university scene.

Actually solely the scientific experiment shows us the physical reality! The theoretical models, like the here presented vortex model merely should help to understand nature and its laws. The representatives of theoretical physics are in the role of a helper and that requires modesty and openness.

For the case that a professor of theoretical physics imposes a pledge of secrecy on an experimenting colleague, if this one wants to report publicly about his amazing tunnelling experiments, then the public nevertheless should have the right to find out, if a tunnelling experiment reveals speeds faster than light, even if it doesn't suit the theorist and he in inappropriate arrogance only should consider the experimental physicist as his assistant.

If universities only are occupied with preservation of property and the industry only with the increase of its productivity, if future research consists of hiding public research funds internally in such a way that nobody realizes the fraud of support and in reality nobody thinks of the future anymore or wants to work for it, then we will be able to observe how discoveries and inventions migrate from the industry and the universities with their controlled central organs and increasingly will take place again in garage, cellar, solitary study room or in privately organized circles.

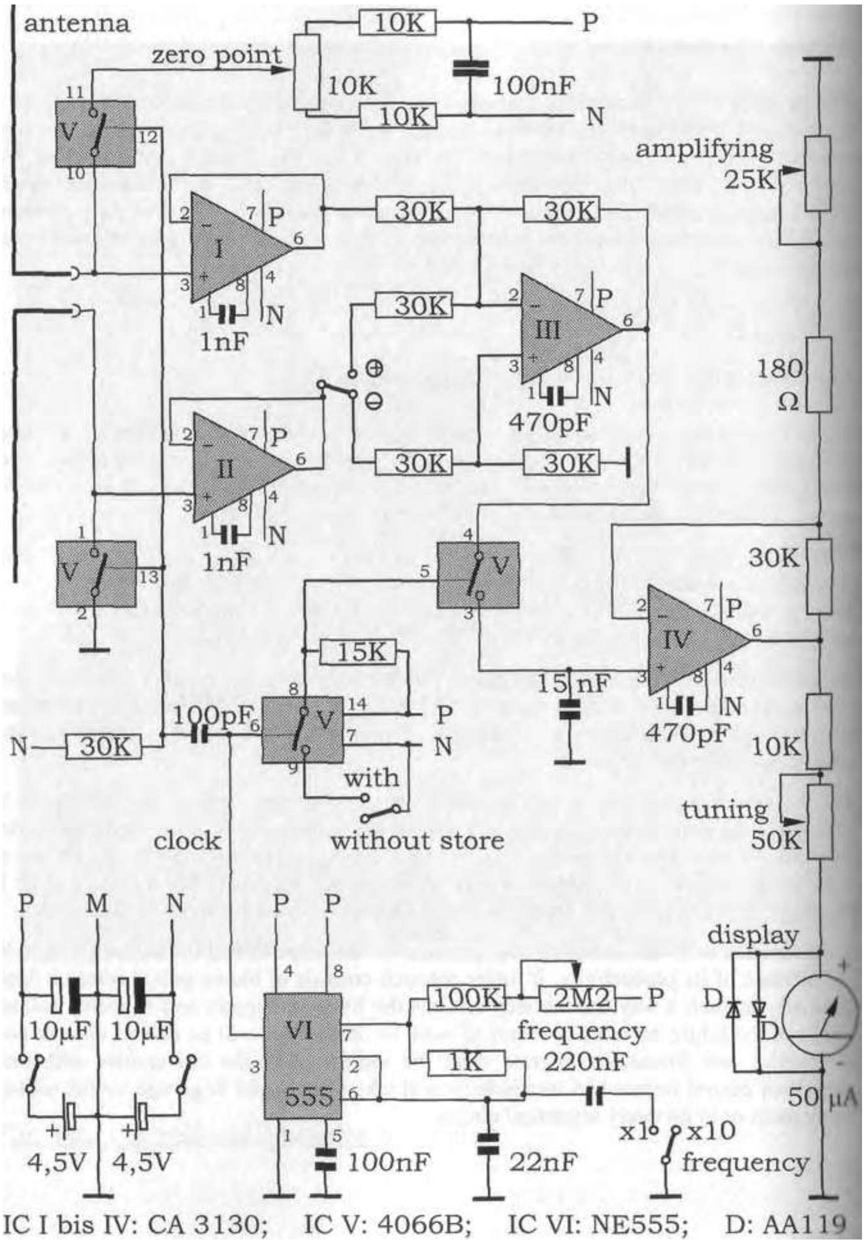
antenna
zero point
10K
10K
10K
100nF
P
N
V
11
12
10
amplifying
25K
I
II
III
IV
30K
30K
30K
30K
30K
1nF
1nF
470pF
470pF
⊕
⊖
180
Ω
30K
10K
V
1
13
2
4
5
3
15K
15 nF
100pF
8
14
9
6
7
N
30K
P
N
with
without store
clock
tuning
50K
P M N
P P
10µF
10µF
4,5V
4,5V
VI
555
100K
2M2
1K
frequency
220nF
100nF
22nF
x1
x10
frequency
display
D
D
50 µA
IC I bis IV: CA 3130; IC V: 4066B; IC VI: NE555; D: AA119

Fig. 19.2: Wiring diagram of the self-built charge amplifier

19.2 The charge amplifier from the workshop of handicrafts

Is connection with my vortex theory in this book time and again clues to experiments can be found. I am aware of their importance and I always, to verify my theoretical working out, have accompanied it with practical experiments. Doing so not the proof, which can be published and reproduced by anyone at any time, stood in the foreground. With the experiments in my workshop of handicrafts I only wanted to check for myself, if my derivations still can be verified.

Consequently was at first done without a publication in the first and second edition of the second part concerning the "electromagnetic environmental compatibility". For the addition of the here presented chapter about own experiments from the third edition there are several reasons: first; I have been asked for it by several readers, second; from every experiment is coming a piece of physical reality to meet one and third are here latent some useful ideas and approaches which get us somewhere.
Of course can't be read much from a handicrafts self experiment, but perhaps one or another reader, who has better laboratory technical possibilities at his disposal than me, is stimulated to carry out own experiments.

To chase after the potential vortices in the air, I 1989 at first have built together a charge amplifier. I connected differently formed antennas at the particularly high-ohmic difference input. Corresponding the in nature arising static electric field a tension voltage should arise between both antennas, which my gauge should amplify and indicate on a moving-coil instrument.
To be able to register local changes, a measurement cycle is gone through, which periodically is repeated: It starts with the measurement time, during which entirely by itself between both antenna a charge is building up. The value afterwards is stored analogue in a Sample-And-Hold link and displayed by means of a moving-coil instrument. Then the input clamps are short-circuited, the antennas again discharged and the game starts from the beginning.

Measurement time, zero and amplification can be adjusted at the device (fig. 19.1). Whoever has fun to rebuild it, finds the by me realized wiring diagram in fig. 19.2.

It is true I did realize other designs, but technically the here shown design proved to be the most useful solution.

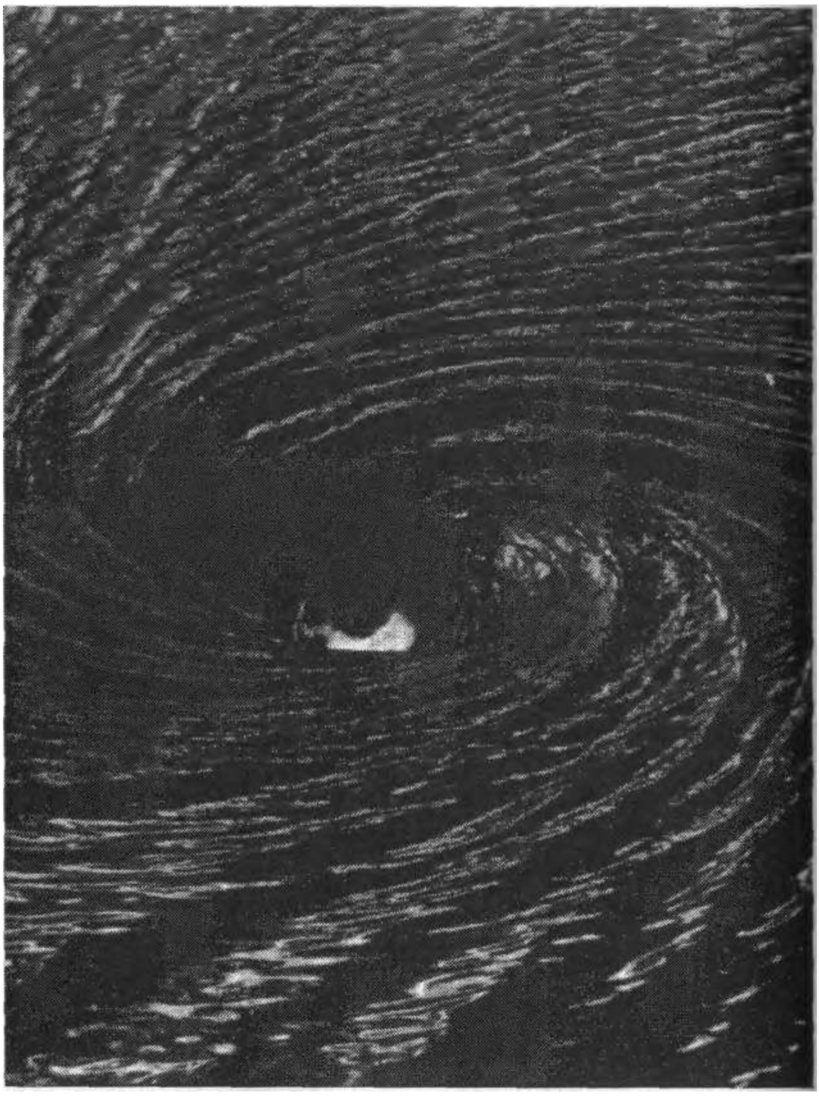

Fig. 19.3: The bath tub drain vortex; shown is a tidal vortex near St. Malo.<i>

<i>: Photograph from the magazine „Life" of 4.7.1969, resp. from: H. J. Lugt: Wirbelstromung in Natur und Technik, Seite 371

19.3 Interpretation of the observations with the charge amplifier

The practical use of the device at first proved to be extremely difficult, because the pointer seemed to perform a wild dance. At a more close look I however could make exactly those observations, which I had sought-for as proof for the existence of potential vortices.

First I did find confirmed the known high field strength. We after all live between the ionosphere and earth's surface right in the middle of the dielectric of a "spherical capacitor". Because the values of the electric field reach very much closer to the maximum value, which is lethal for living beings, than the value of earth's magnetism, I did draw from that the conclusion that biological effects primarily can be expected from the E-field.

If the E-field, as Maxwell's field theory specifies, actually would be irrotational, then we would be dealing with a gradient field. My device would have to display everywhere to a large extent the same value. But that was not the case.

Maybe the building is responsible for the chaotic display and the wild swings of the pointer, so I thought. Doing so I had thought of the auxiliary explanations of the high-frequency engineer about so-called reflections in closed rooms. Therefore I stormed with the gauge, which had gone wild, into the open air and walked different ways, which should have been reflection free, but the picture stayed the same. In any case the E-field is not a gradient field. I had to find out.

Following I could, what required much patience, find certain places, at which for the same movement from the same direction could be seen a reproducibility of the swing of the pointer, it even could be arbitrarily often repeated. 1 marked the point exactly. Then I tried to move the device from another direction towards the marking and had to find out that the point had moved away.

If I sit in the bath tub in the evening and pull the plug, then I each time am enthusiastic about how sensitive the drain vortex reacts, how I can send it from one corner into the other by the snipping of a finger without it falling apart (fig. 19.3). Doing so one easily can imagine with a bit of phantasy how difficult, yes, almost impossible it would be to measuring technically register the vortex for the case that we could not see it. The gauge it is true would display violent wave movements. But a reproducibility we would not be able to obtain, exactly as for my self-built charge amplifier.

Now I knew that the by me at 2.1.1990 at first purely theoretical derived potential vortex actually exists as vortex of the electric field!

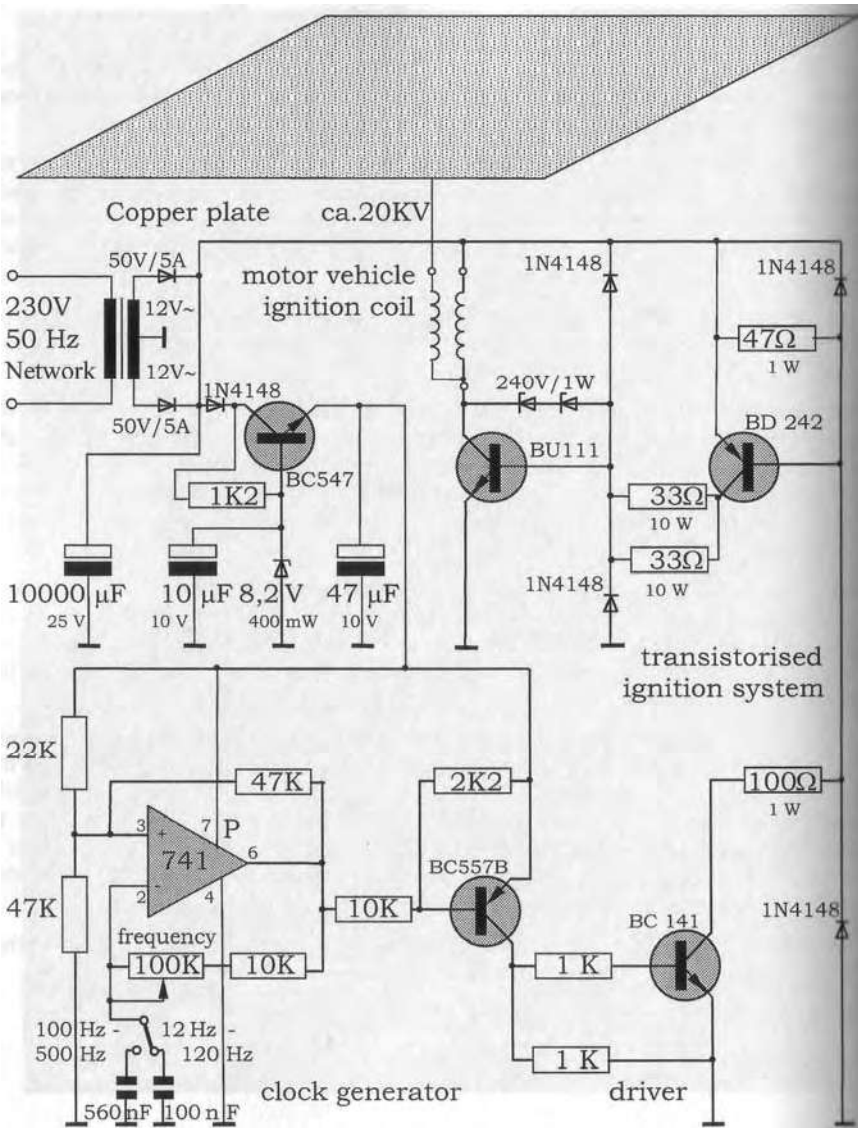
Copper plate ca.20KV
230V
50 Hz
Network
50V/5A
12V~
12V~
50V/5A
motor vehicle
ignition coil
1N4148
BC547
1K2
10000 µF
25 V
10 µF
10 V
8,2 V
400 mW
47 µF
10 V
240V/1W
BU111
1N4148
1N4148
47Ω
1 W
BD 242
33Ω
10 W
33Ω
10 W
1N4148
transistorised
ignition system
22K
47K
47K
741
P
frequency
100K
10K
2K2
10K
BC557B
BC 141
1N4148
100Ω
1 W
1 K
1 K
100 Hz - 500 Hz
12 Hz - 120 Hz
560 nF
100 nF
clock generator
driver

Fig. 19.4: Wiring diagram concerning the Kirlian device.<i>

<i>: Elektor, Fachzeitschrift fur Elektronik, Mai 1977, S. 22-25: KirlianfotogralV

19.4 High-tension experiments in the laboratory for performance electronics

At first the students of electronics, with which I had to do daily, were infected by my discovery. With true enthusiasm they soldered together one circuit after another. In my laboratory something like a mood of setting out could be felt. One student brought along and old ignition coil of a motor vehicle, then the control circuit for it was built, plexiglas organized, an aluminium plate glued under it and the whole box dragged into the photographical laboratory<i>. Three students made a detour along the outside grounds of the polytechnic and picked the necessary visual aids of trees and bushes.
Following the leafs were laid on photographical paper and by means of the self-built high-tension generator charged to 20000 Volt. Doing so at all edges and particularly at the tips of a leaf corona discharges arise, which expose the photographical paper. We then developed the photographs ourselves and discussed the results <ii>.

The indication, for a second photograph the whole leaf would appear even if half the leaf is torn apart after the first experiment and only one half is put on, occupied us in particular. Eventually we didn't put on a leaf at all in the second experiment and found out that nevertheless the leaf put on last became apparent on the photograph. Now only one physical interpretation was possible: The potential vortices of the leaf stimulated by high-tension still were in the plexiglas disc in weakened form! Here they swirl further and under high-tension furthermore produce corona discharge impulses.

Now also the students were convinced of the existence of the potential vortices.

This experiment stimulated the brain cells of the entire team: If potential vortices under high-tension cause electric blows, since the corona discharge is nothing else, then in this way the local vortex distribution in space should be measurable? One student immediately got down to work and build from the horizontal diversion unit of an old television a high-tension generator with an adjustable spark gap. Following he walked with his flashing and crashing device all through my laboratory and others walked with him and dragged the gauges behind them. But to their big disappointment they were not able to see an influence dependent on place.

Then it suddenly was clear for me: It would have done the students well, if they before would have had a bath and observed the drain vortex, "it suffices the famous wing stroke of a butterfly", I explained my team, "and instead come up with this infernal machine and chase away all the vortices, which we actually want to register ". We had to realize that as a rule phase of disillusionment follows the euphoria.

<i>: Klektor, Fachzeitschrift fur Elektronik, Mai 1977, S. 22-25: Kirlianfotografie; fig. 19.4 shows the wiring diagram and fig. 19.5 A the self-built device.

<ii>: The Kirlian photographs from the darkroom of the polytechnic are shown in fig. 3.6.

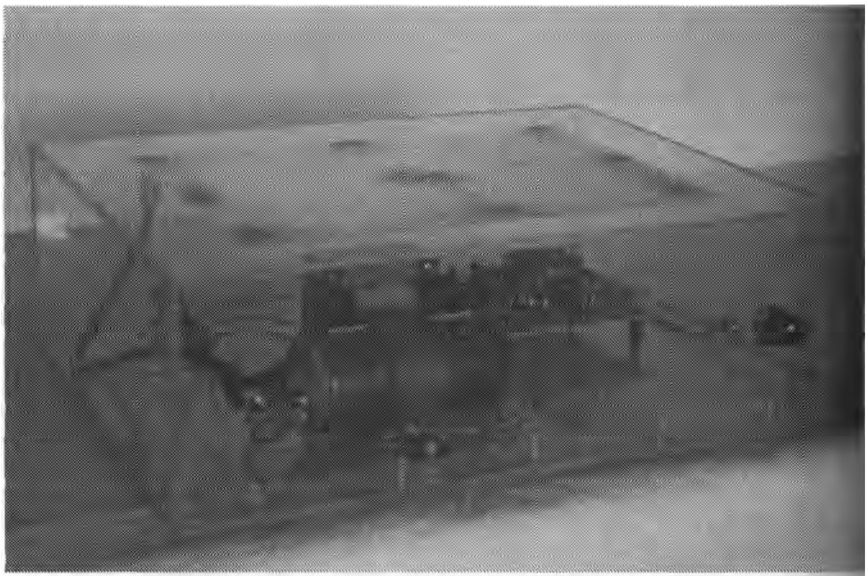

Fig. 19.5 A: A self-built high-tension device for Kirlian photography

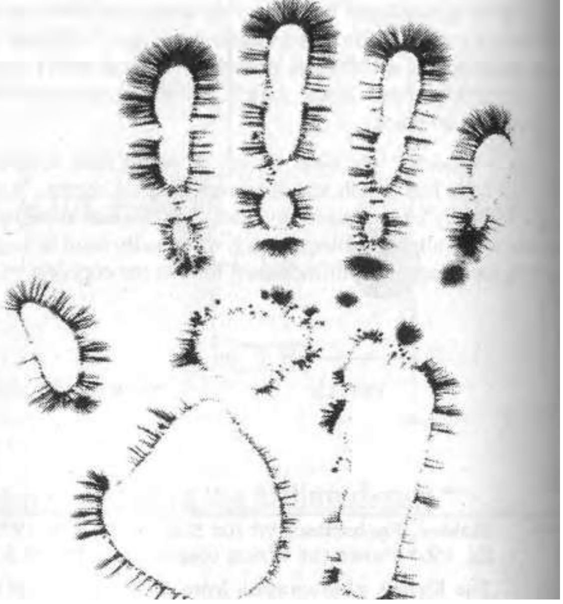

Fig. 19.5 B: Kirlian photograph of the right hand

19.5 Measurement of dielectric capacitor losses in the HF laboratory

My convincing work clearly was shaped more complicated with my professor colleagues. Most pretended to understand nothing about it and kept themselves out. But there also are exceptions. „something like that has never been there at our polytechnic, that a colleague puts forward an own field theory ",a physicist colleague remarked after my lecture public to the polytechnic[i] at 3.7.1990 and meant, that must be celebrated. He festively let me cast my eyes in his store room, got out a few bottles of wine, which then some of my colleagues of physics emptied with me in their official room. Doing so we small talked about the sense or nonsense to fix on the education of physics to the Coulomb charges.

The head of the high-frequency laboratory showed likewise impressed. At last he now knew, why capacitors at microwave frequencies can become so hot that they solder themselves out of the circuit by themselves and can fall out, why PVC-films can be welded with HF, etc. „We have to prove that not the dielectric is to blame, as stands in the textbooks" he came towards me.
At least following my idea potential vortices are the ones, as I expressed myself, which are behaving dual to the eddy currents. It concerns vortex losses, thus a physical phenomenon. Eddy currents now can be damped as is well-known, by for instance sheeting the iron circuit for engines and transformers. The insulation between the sheets prevents the formation of eddy currents in that direction and the degree of effectiveness increases.
"I would suggest", I told my colleagues, „to "sheet" a capacitor in a dual manner and to measure the losses". Because eddy current losses increase with the square of the frequancy, we picked microwave frequencies. In the HF-laboratory a card with an L-C resonant circuit was made, we should be able to quite precisely determine the losses by means of its quality. The inductance was formed as a microstrip line and for the capacitor a socket was planned.
The carrying out of the experiment consisted of using single layered ceramic capacitors, so called trapezium capacitors, with a vapour deposited silver coating on both sides as a package and to measure the quality of the resonant circuit as a measure for the dielectric losses. According to the rules of duality the dielectric (as nonconductor) thus corresponds to the permeable transformer sheet (conductor) and the silver coating (conductor) to the sheet insulation (nonconductor). In a second experiment the silver coating now was removed and the same capacitor material measured at identical conditions this time uncoated. Will the losses increase or isn't changing anything?
Tension and nervousness suddenly could be felt, even among the students, who followed the experiment from the second row. All looks were pointed at the monitor of the network analyser, on which slowly the bell-shaped curve of the measured resonant circuit quality became apparent. The result was surprisingly clear. I first looked at my colleagues and then at the students and had to realize that all eyes were pointed at me. The first measurement was a bull's eye! The vortex losses "uncoated" were considerably larger.

---

[i] K. Meyl: Die mangelnde Dualitat der Maxwellschen Gleichungen with contributions concerning the theme of magnetic space poles, the mathematical calculation and the technical-physical interpretation of newly discovered potential vortices. Invited by the rector of the FH Furtwangen at 3.7.1990

- experiment 1: quality of resonant circuit with coated capacitors
  from a series of 5 separate experiments averaged: Q = 48
- experiment 2: silver coating removed with diluted nitric acid
  possibly the acid has damaged the dielectric: Q = 9
- experiment 3: silver coating polished off mechanically.

low quality means high dielectric losses! averaged: Q = 28

- resonance sharpness (quality number): $Q = \frac{f_0}{\Delta f}$ $\frac{\text{(Resonance frequency)}}{\text{(Resonance frequency ran}}$

at a resonance frequency f0 of approx. 400 MHz.

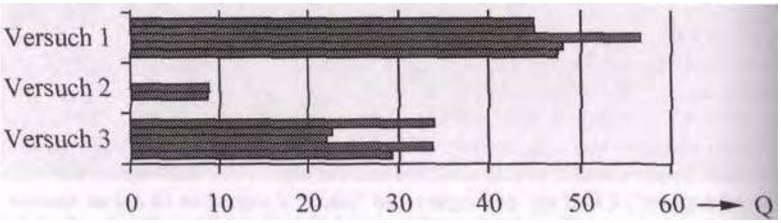

Quality number Q of the L-C resonant circuit
the dual ,,sheeted" capacitor has the lowest losses,
the resonant circuit with that the highest quality (experiment 1).

Fig. 19.6: Measurement of the dielectric capacitor losses over the quality of a resonant circuit in the HF laboratory at 11.06.1990.

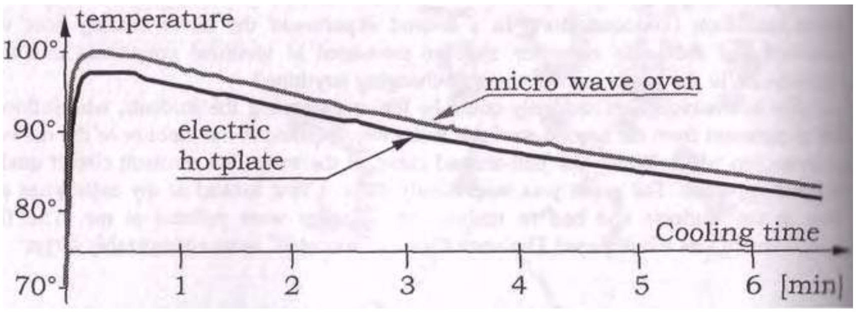

Fig. 19.7: Measurement curve of the water temperature over time
at 22.10.1996 in the electrical engineering training
Observation: for the microwave oven the container despite the "after cooking-effect"
remains colder than in the case of a hotplate. Sensor: PT 100.

19.6 Analysis of the measured capacitor losses

First of all the colleague spontaneously prompted a common publication.

But a certain disillusionment had to await us. For the next measurement hardly a differemce could be detected. We varied the contacting and removed the silver coating. once with hydrochloric acid, once with fine sandpaper. Every measurement we repeated umpteen times and in the end had to draw the conclusion that the results for our somewhat unprecise construction fluctuated extremely.

Averaged over all measurements and methods, indeed a reduction of the dielectric losses for alayered capacitor resulted, but at a critical consideration of the errors the results seemed to "drown" in the mean variation. In the opinion of my colleague the visible trend wasn't sufficient to convince dyed-in-the-wool sceptics, whereupon he withdrew with the indication, I should try at the people of microelectronics, they more likely would be capable to reproducibly gain control of a layered construction.

The experiment at least let a legitimate chance open that the vortices of the electric field actually exist, and in the case of the delectric losses of a capacitor it concerns vortex losses<i>.

19.7 Microwave oven for testing in the laboratory of electrical engineering

"Volunteers first" was said at 22.10.1996 in the electrical engineering practical training and two students got down to work. I had brought along from our kitchen the microwave oven, a portable immersion heater, a kitchen stove plate and different containers. In the laboratory of the polytechnic I in addition had got hold of a Bunsen burner. In the sense of the "Stiftung Warentest" (institution to protect consumers in Germany, note of the translator) it concerned the question for differences in heating water; or following the public discussion, is cooking with gas more healthy than cooking with the microwave oven?

We wanted to know. So we cooked water, filtered by inversion osmosis, once with the gas burner. once with the hotplate and finally with the microwave oven. Switched off always was at the same moment at the same temperature, and the cooling down curve was recorded with an x-t-time recorder (fig. 19.7).

Between gas and electric cooker virtually no difference could be detected, but the microwave oven at switching off still showed an odd temperature increase. I already had remarked this earlier. If one takes a water glass out of the microwave oven, then the water again really bubbled off, although the container itself had remained relatively cold. If one on the other hand takes a water pot off the gas cooker, the water suddenly stops cooking. From where comes this difference, which also the experiment could confirm?

<i>: A result of my dual way of looking at the vortex losses is the representation in fig. 4.7.

A: damped wave equation:

$$\begin{aligned} \Delta \mathbf{E}\, c^2 \;=\; & \delta^2\mathbf{E}/\delta t^2 \;+\; (1/\tau)\cdot\delta\,\mathbf{E}/\delta t \\ & \text{(wave)} \;+\; \text{(vortex)} \end{aligned}$$

B: open chain of reasoning in the physics of electromagnetism:

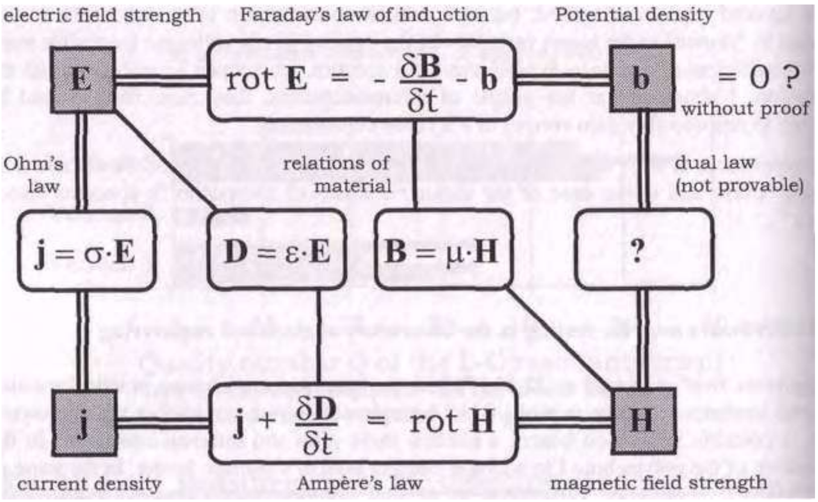
electric field strength
Faraday's law of induction
Potential density
E
rot E = - δB/δt - b
b
= 0 ?
without proof
Ohm's law
relations of material
dual law (not provable)
j = σ·E
D = ε·E
B = μ·H
?
j
j + δD/δt = rot H
H
current density
Ampère's law
magnetic field strength

C: for the stationary case $(\delta/\delta t = 0)$ in general representation:

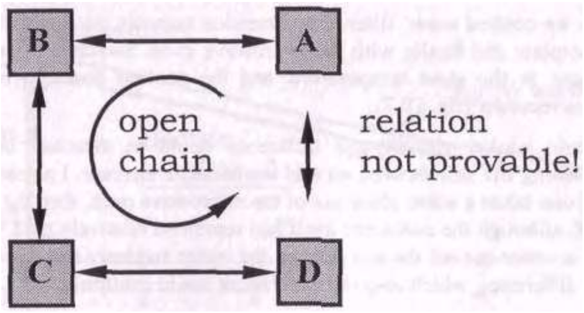
B
A
open chain
relation not provable!
C
D

Fig. 19.8: ____ Contribution to the discussion about the impossibility, to prove, according to scientific methods, in a chain of reasoning (A-B-C-D) the last link (D-A), which closes the chain.

19.8 Microwave heating by decay of vortices

It has to concern a storing effect. In the case of an normal cooker the heat transfer takes place by convection. In the case of a microwave oven we however are dealing with an electromagnetic wave.

But in the wave equation nothing can be read about heat; here merely vortices can be found as a damping term (fig. 19.8 A)<i>! Therefore as the only possible interpretation is left that the microwaves roll up to vortices to fall apart themselves after a certain time.

Only at the vortex decay heat is formed.

The irradiated microwave power according to that is stored over a longer period of time as vortex in the water and the food. The vortex decay takes place according to an e-function with the calculated relaxation time constant tau<ii>.

From the circumstance that particularly water is heated up in the microwave oven despite the small electric conductivity but with high dielectricity, I infer that it has to concern mainly potential vortices, from which a biological effectiveness can be expected. The question if this should be valued positively or negatively, I have to leave up to the doctors and therapists. At least physically seen a difference is measurable. Thus a vortex decay will occur for a meal prepared in the microwave oven even if we already have consumed it.

Conclusion: I wish all: your health!

But what do I write about vortex losses, if every sceptic knows half a dozen alternative interpretations. Inevitably I have to bear in mind that also the eddy current heating owes its acknowledgement only the circumstance of the discovery of the corresponding laws. If Othm hadn't discovered his law and formulated it in the known form and instead the dual formulation would have been discovered, then we today would attribute the dielectric losses in a capacitor and in the microwave oven to the potential vortices, in the case of the transformer however the material would be responsible for the heating and not eddy currents, for which there then also wouldn't be a theory (fig. 19.8 B).

It actually is pure coincidence that at first Ohm's law and not the dual formulation had been discovered and acknowledged. But because both are equivalent, we also have to assume the correctness of both, even if the last link of a chain of reasoning A-B-C-D back from D to A in principle can't be proven anymore, since it already is explained by the chain A-B-C-D (fig. 19.8 C).

There I thus had discovered a potential vortex, without a possibility for a direct proof of existence. For that the textbooks are full with auxiliary explanations, with which physics successfully cheats past this important field phenomenon. I was frustrated. There had to be a way to measuring technically register the vortex in some way.

<i>: Fundamental field equation 15 in fig. 5.1 or equation 21, fig. 5.3

<ii>: Relaxation time constant: see equation 10 in fig. 5.1 or fig. 8.1

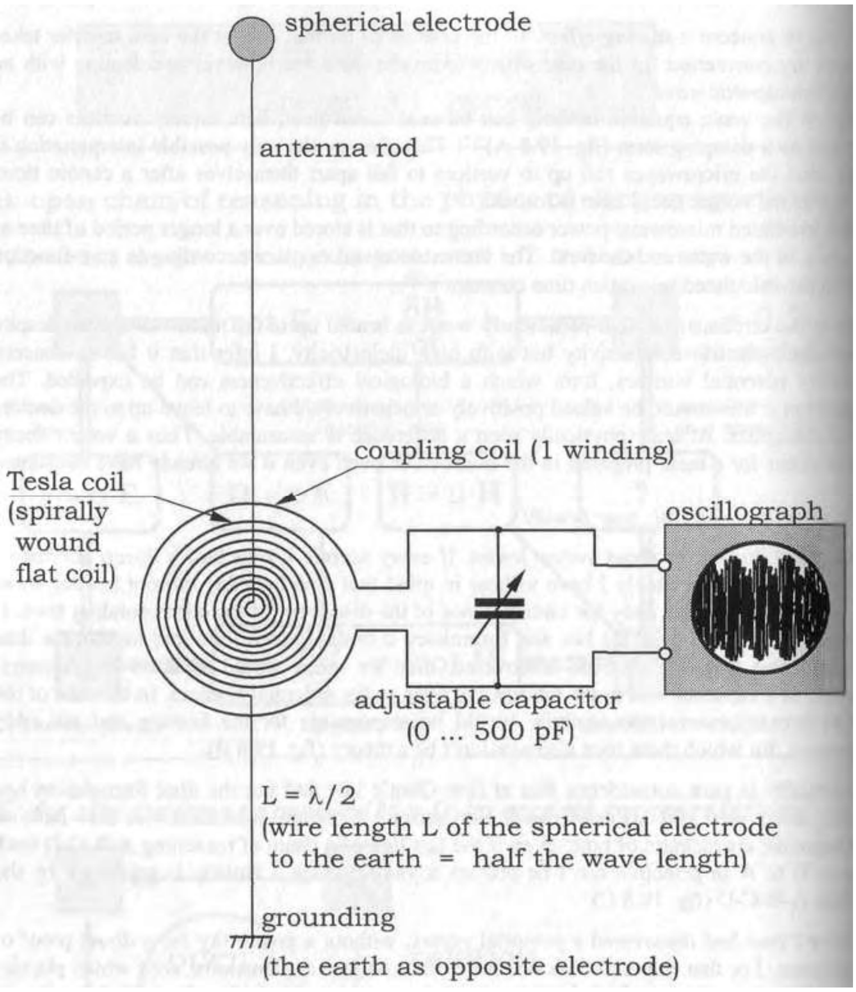
spherical electrode
antenna rod
coupling coil (1 winding)
Tesla coil (spirally wound flat coil)
oscillograph
adjustable capacitor (0 ... 500 pF)
L = λ/2
(wire length L of the spherical electrode to the earth = half the wave length)
grounding
(the earth as opposite electrode)

Fig. 19.9: Proof of scalar waves with the Tesla coil.<i>

<i>: The inverse use for a high-tension transmitter according to fig. 9.1

19.9 The Tesla coil from the workshop of handicrafts

I again retired in my workshop of handicrafts and wound flat coils or I pondered in my study over Tesla books. 1995 from several sides my attention had been drawn to the circumstance that the inventions of Nikola Tesla can be organized in three categories:

The first third has made him world-famous. It concerns the rotary field theory, the asynchronous engine and the today normally used alternating current technology, which we owe him.

The second third concerns technologies and inventions, which were rediscovered by other people partly only years later or even were only pinched and distributed as novelty under another name. Electron microscope, superconduction, electrolyte capacitor, fluorescent lamp, fuse. coaxial cable and a lot more count among that.

The last third however concerns inventions, which until the day of today still aren't understood and await their scientific explanation and technological use. Tesla himself called these achievements his most important inventions, but still owes us a scientific explanation. The scientific world also hasn't got a theory ready and doesn't know to do anything with it. Public research doesn't take place or is prevented by lobbyists.

What remains, are tinkerers of various educational background and qualification, who are trying hurd to comprehend the buried experiments of Nikola Tesla off their own bat in the garage or in the hobby cellar.

I felt like one of these, by winding one winding after the next from the inside to the outside. Then I soldered an antenna wire at the inner end of the flat coil, which I in Tesla manner connected electrically with a spherical electrode hung up under the ceiling. The opposite electrode should be connected to the outer end of the winding, it is said, and the distance between both should be as big as possible. If namely an electrode just is collecting, then the opposite electrode is repelling the same space quanta. According to Tesla's recommendation I did use the earth as opposite electrode and for that tapped the central heating or the grounding of the foundations.

To take signals only one to two windings as secondary winding were necessary. I connected them with an adjustable air capacitor from an old steam radio to a frequency determining parallel resonant circuit and looked at the taken tension voltage at the oscillograph (fig. 19.9),

I still had problems with the statement of Tesla, the coupling had to be made loose. Thus the question is asked, how loose? I after that organized two toilet paper rolls of different size, (after the toilet paper had been used, naturally) and pushed them into each other. The smaller toilet paper roll carried on the gable-end the flat coil as primary winding of the air transformer and the bigger one the coupling coil. Now by shifting any wanted degree of coupling could be adjusted to (fig. 19.10).

I was astonished myself. Tesla actually was right with his discovery of the scalar waves. With my arrangement they can be clearly distinguished from the Hertzian waves.

The following procedure is recommended:

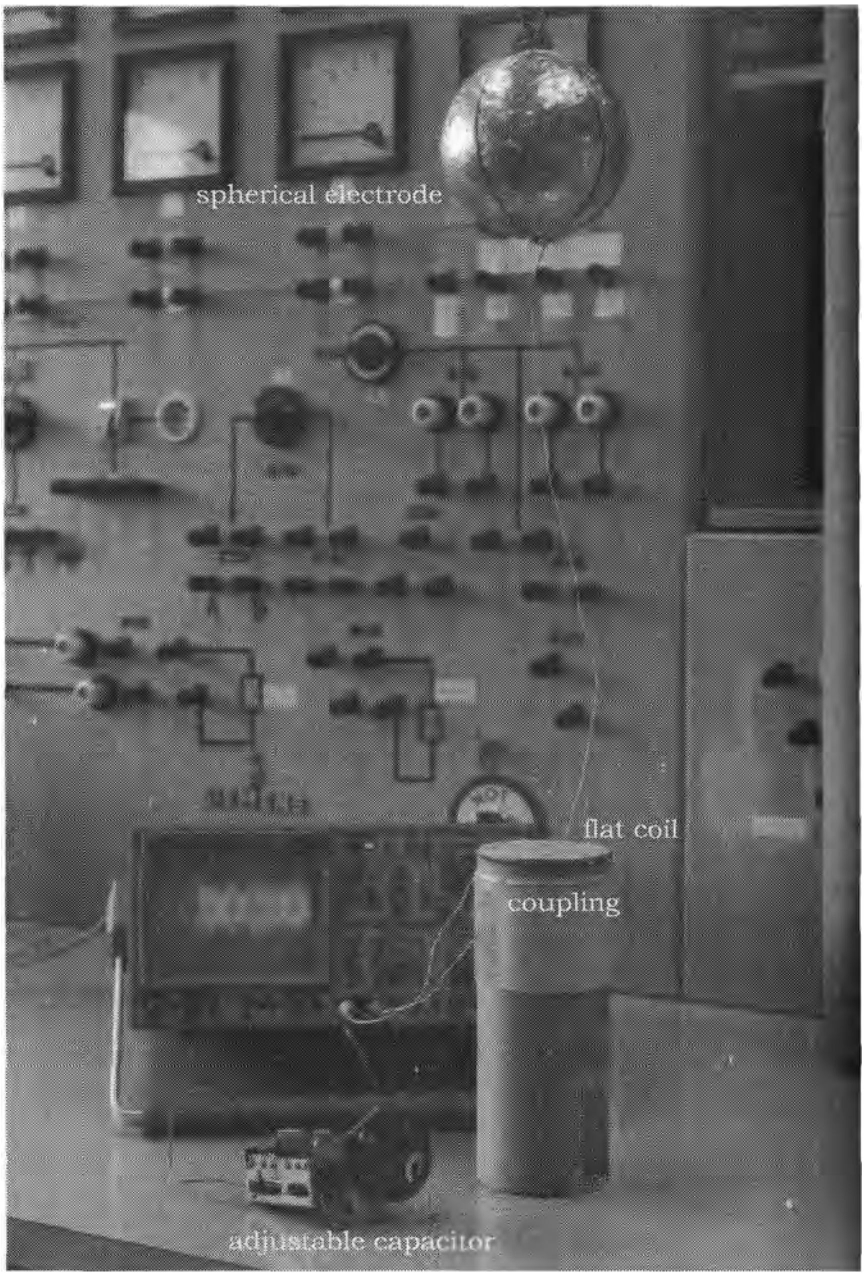
spherical electrode
flat coil
coupling
adjustable capacitor

Fig. 19.10: The experimental configuration

First of all I seek a source of interference with the adjustable capacitor and tune to maximum amplitude. Then I change the coupling and further optimise in this way. If now the amplitude again decreases from a certain point while approaching the coupling coil, then it concerns the sought-for longitudinal waves. If namely the coupling is too tight, then the received vortices again are driven away by the effect back on the fiat coil. They make way.
At last I had found a method to catch the vortices in such a way that they not immediately "ran away" from me again. At once I presented them in the technology centre in St. Georgen. In the time following I improved the technology further and further, used bigger toilet paper rolls and eventually even turn up garbage cans, I varied wire length, wire diameter and the sense of winding (fig. 19.10).

I had very different success. Sometimes, if at the same time in my workshop of handicrafts the radio worked, it would look as if the received signal would synchronize with the sound waves. Both are longitudinal waves after all. With transverse waves something like that would be unthinkable.
One moreover could observe, how a resonance builds up: first slowly and then faster and faster, so that I sometimes got terribly afraid. Several times we had to repair our oscillograph, after the protective diode at the input amplifier had blown, and that for signals of 50 to 100 millivolt!
That was entirely impossible. Only individual spikes, which were too fast to be seen at the screen, could be to blame. In the case of distant thunderstorm activity I obtained maximum values of more than half a Volt. After that I undamped the grounding as fast as possible, so that no lightning would be caught, since I didn't have the intention to burn off my workshop.

19.10 Biological effectiveness of the Tesla coil

Also the biogical effectiveness of the Tesla radiation I could impressively prove with this device. 14.06.1997 a woman, who called herself extremely electrosensitive, participated in the weekend seminar about electrobiology, which I took over from Prof. H.K. Konig (TU Munich) after his death. That I wanted to test.
I hung up my device in the lecture auditorium and installed the oscillograph in such a way that all could see it. The voluntary test subject however could see neither the public nor the screen. One person every 5 seconds said "now" and the female candidate should say, if I had clamped or undamped the grounding, if therefore scalar waves were received or not. After a training round her proportion of hits was lying at 100 percent!

According to her statement she could feel it. A further test subject achieved even at the back of the room with a pendulum the correct answer. I myself was surprised by this and I already have repeated this experiment several times with different success. It without doubt depends on the sensitiveness of the test subjects. Every person after all reacts to other signals.

What however has astonished all participants and can't be emphasized enough at all, is the circumstance that it in this case concerns a receiver and not a transmitter!

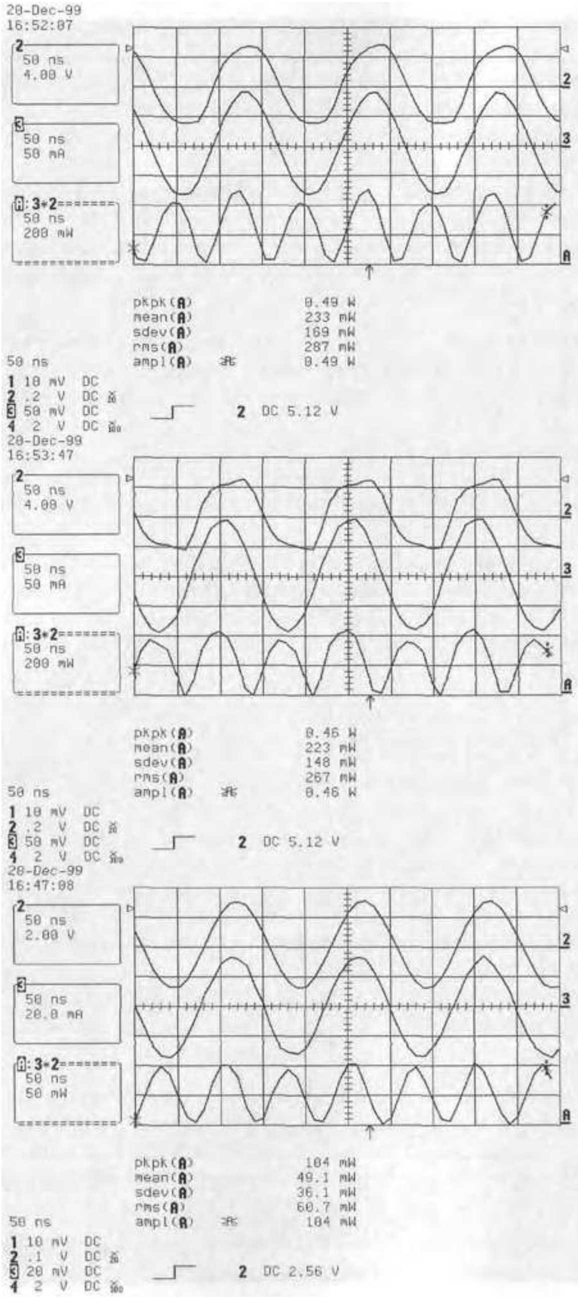

Measurement record

System resonance: f = 6,7 MHz, Sinus.

distance transmitter-receiver: approx. 2 m

transmitter coil with HF-braid identical to receiver coil: 38 Windings wound spirally and 33 Windings wound cylindrically.

Load for receiver: metal layer resistor (100 Ohm)

Gauge: Quad 200 MHz Oscilloscope: LeCroy 9304C

Current sample: LeCroy AP015 (DC-50MHz)

Differential sample: LeCroy AP032 (Attenuation Rate: 1/20)

Top: transmitter power consumption with receiver turned on (at 100 Ω-load): 233 mW

Middle: transmitter idle power: (without receiver) 223 mW

Bottom: receiver power output: 49,1 mW

Fig. 19.11 A: Measurement record for scalar wave transmission line

19.11 Wrestling for measuring technical insight in the Transfer centre

In the central question who should be believed more, the famous experimental physicist Nikola Tesla or his critics, my experiments with the Tesla coil were the visible proof that we have to take Tesla's statements seriously. Now not only at myself, but also at the students of the polytechnic and the colleagues of my Transfer centre a true Tesla euphoria broke out.

In every free minute the patents and original writings were studied, which I had myself send from the Tesla Society in Colorado Springs in the USA. In particular my trainees and diplomands developed a incredible ambition in making a historical Tesla concept work. They built a whole series of various high-tension generators. In the laboratory one could hear crashing and there was a smell of ozone.

22.1.1998 at a presentation of the works for a degree the candidate looked after by me very proudly held a fluorescent lamp in his hand, which in the field of his self-built high-tension generator glowed even without any wire connected, entirely according to the great mode (fig. 17.10). At all efforts we however laboratory technical weren't able to reach tension voltages of above 511 000 Volt. But at this tension voltage the actual Tesla effects actually start!

Perhaps it was tough luck, but possibly also a chance that we in the laboratories, which I had at my disposal, were forced to work with lower tension voltages and that meant that we gradually had to beak away from the Tesla designs.

Moreover we hadn't at our disposal the precisely controllable spark gaps, which Tesla had developed and used. If one wants to obtain an if possible high tension voltage change (du/dt) for an interaction with neutrinos, then according to today's technology considerably more favourable concepts are offered, for instance with hard switching Power-MOS-Transistors. 50 Kilovolt per microsecond were to meet.

Therefore we changed the technology and worked from now on with fast semiconductor switches. From the laboratory radio now only a hissing and crashing came out of the loudspeaker, if our experiments were running.

At 12.10.1999 we for the first time succeeded to build up a transmission line for Tesla radiation. Doing so the transmitter and receiver were situated in different rooms of my transfer centre. The transmitter coil was operated in self-resonance and fed only from a small function generator with 10 Volt. But if the diplomand held a fluorescent lamp near the spherical electrode, then it started to glow!

Following I observed at the oscillograph the signal of the receiver coil, which as well was operated in resonance. If the diplomand switched off the transmitter, also no receiver singnal was present anymore. But if it concerned radio waves according to Hertz or Tesla radiation, with that still wasn't answered.

Therefore I prompted still another experiment. This time the colleagues observed at the transmitter the signal at the function generator, while I undamped and again clamped the receiver. The shouting with joy from the adjoining room indicated that it had been observed, how the receiver reacted upon the transmitter and both are in resonance with each other. Such an effect characteristic for scalar waves, is a radio technical impossibility! In the case of radio with Hertzian waves an effect back from the receiver on the transmitter is unthinkable by principle.

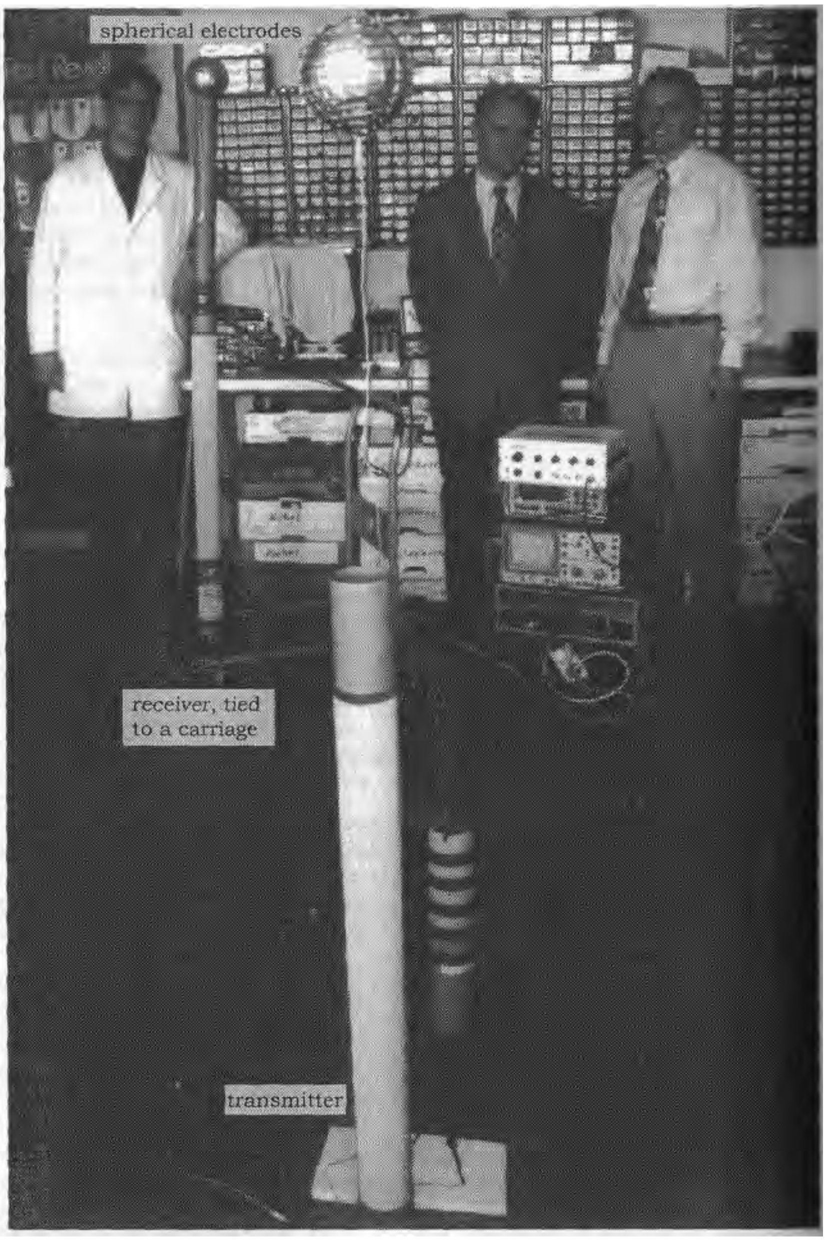
spherical electrodes
receiver, tied to a carriage
transmitter

Fig. 19.11 B: Scalar wave transmission line according to Tesla
f.l.tr.: M.Andresen (diplomand), author (TZ-head), Dipl.Ing.M.Rehm (project leader)

In the next step we let glow the famous little lamp on the side of the receiver. As a consumer served a small light-emitting diode, of which the light intensity remained unchanged in the case of resonance. To prove this, we placed the receiver on a carriage and rolled the corridor in the TZ up and down with it. If the receiver was only slightly out of tune, then from the then arising fluctuation the standing wave nature could be observed perfectly.

With the setup it can be demonstrated well, how the law of the square of the distance, of the decrease of the field strength with the square of the distance, known from radio technology hasn't got validity anymore for scalar waves. Very clear also was the energy transmission of scalar waves out of a closed Faraday cage.

In the end we have determined the degree of effectiveness of the scalar wave transmission line. An output power of 49 mW resulted from the measurements of the current and tension voltage for a loading of the receiver with a 100 Ohm resistance. Simultaneously the power taken up by the transmitter amounted to 233 mW. If we however subtract the idle power consumed by the transmitter from this, and that was determined to be 223 mW for switched off receiver, then actually only 10 mW are available for the wireless energy transmission. The degree of effectiveness according to that would be 490%!

If we here really have obtained an over-unity of 4.9 then the receiver must have collected along free space energy, or did some parts evade our power measurement? The sceptics I recommend an own rebuild, since alone the proof of the scalar wave properties inevitably has to lead to every Maxwell burdened HF technician breaking with the old belief.

19.12 Neutrinolysis, the alternative splitting of water

Free after the plans of Stan Meyer we filled diverse containers with water and let it "crash". If then bubbles raised and it got exciting, we changed the frequency and exactly paid attention to the effect remaining the same or if it increased as well with increasing frequency, thus if we only watched a classical electrolysis, or already the wanted "neutrinolyse", as we were accustomed to term the splitting of water in its parts under the influence of neutrinos in the laboratory. Typical for that is a bubbling and "cooking" of the actually cold water produced by the rising gas bubbles (fig. 19.12).

We also have reversed the principle. Doing so we have switched the container as neutrino receiver and measured the forming charge carriers. The passive system proved to be really moody. On the one hand tension voltages of several hundred millivolts at a load resistance of 10 megaohm can be realized, on the other hand the charged gas particles and the as secondary reaction in the water set free ions continually change the electric conductivity, so that the cause to be measured, the neutrino radiation, hardly can be reproduced on the displayed result. The water hence has to be changed more often and also the developers found themselves between jubilation and disillusionment subject to continually changing feelings.

<i>: A. und I. Schneider: Neutrino-Power - Energie aus dem Kosmos, Bericht zu Vortrag und Demonstration des Autors am 25.11.99 in Villingen, NET-Journal 12/99, S. 4-6; Interview mit Prof. Dr.-Ing. Konstantin Meyl: Durchbruch in der Freien-Energie-Forschung, NET-Journal 12/99, S. 7-9

Fig. 19.12: Experimental constructions for the „neutrinolyse“.

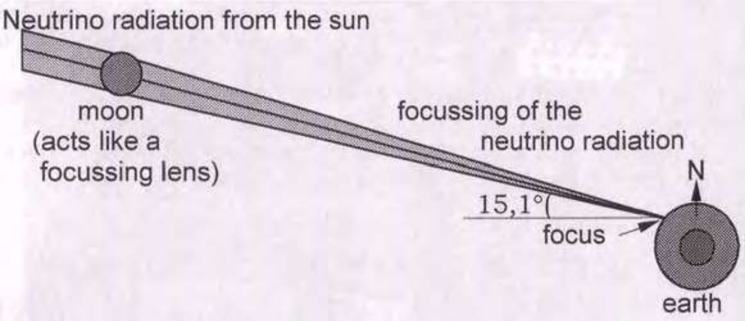
Neutrino radiation from the sun
moon
(acts like a
focussing lens)
focussing of the
neutrino radiation
N
15,1°
focus
earth

Fig. 20.1: Course of the neutrino radiation focussed by the moon on the occasion of the eclipse of the sun at the 11th August 1999.<i>

concerning the calculation of the position of the sun at 11.8.99:

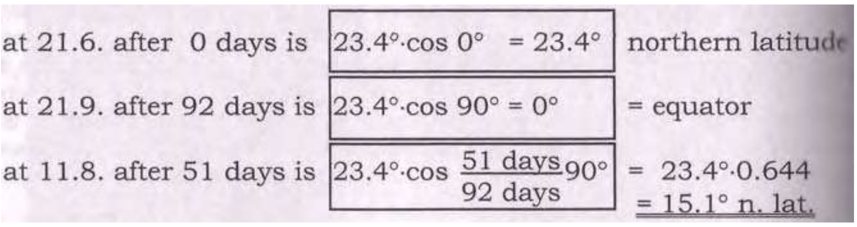

at 21.6. after 0 days is	$23.4^\circ \cdot \cos 0^\circ = 23.4^\circ$	northern latitud
at 21.9. after 92 days is	$23.4^\circ \cdot \cos 90^\circ = 0^\circ$	= equator
at 11.8. after 51 days is	$23.4^\circ \cdot \cos \frac{51 \text{ days}}{92 \text{ days}} 90^\circ$	$= 23.4^\circ \cdot 0.644$ $= 15.1^\circ$ n. lat.

At 11.8.1999 the sun at noon stands above latitude 15.1.

<i>: Konstantin Meyl: Zur Brennglaswirkung des Mondes bei einer Sonnen-

20. Cosmic experiments

As long as no usable technical gauges are available, we should observe and study nature and all celestial phenomena. Here the scalar waves with all their properties are presented to us. That is valid in particular for the case that the cosmos makes an experiment with us and with the entire earth. Such a situation offered the eclipse of the sun of August 11 1999 as an unique "bulk experiment" to which travelled millions of spectators, to participate in the cosmic experiment as voluntary test subjects.

My warning for possible influences proved to be absolutely justified afterwards, even if only comparatively few took note<i>. In view of strong changes in the EEG of individual test subjects and a proven temporary acceleration of the rotation of the earth science once more stands before insoluble problems.

To answer the open questions I want to start with my indications, which 1 have published in anticipation of the cosmic event of August 11<i>. Afterwards a revision and the attempt of an interpretation from the scalar wave view follows. Perhaps this way leads to a reliable prediction of of earthquakes and other cosmic events. After that we perhaps know more about origin, availability and further important parameters of the sought-for space energy.

20.1 Millions of voluntary test subjects at 11 August 1999

Astronomically seen, it concerns a harmless natural spectacle as it already was observed more often, if the moon glides between sun and earth and its shadow in broad daylight immerses parts of the earth in a dark night. The special thing about the eclipse of the sun of August 11 1999 however was, that the sun activity just strived for its maximum value in its eleven year cycle, and the orbital distance to the moon simultaneously reached a minimum value. In addition we must pay attention to the special situation of the angles (fig. 20.1). Due to the extremely rare constellation the area of complete shadow at first was larger and darker than normal. On August 11 it had a width of 110 kilometres.

The sun besides the light also sends us solar neutrinos and for those the moon is transparent. It with regard to them acts like a glass ball, which lets the light it is true pass through, but in doing so refracts it. The glass ball acts like a convex focussing lens, which focuses the arriving rays behind the ball in a focus.

To slow down and collect neutrinos the moon it is true is too small, but it will be able to influence the flying direction. Very fast neutrinos, which run through the moon, hardly will be diverted. The slow and biologically active ones however will be bent stronger.

At this point it would be of utmost importance to know, in which distance from the moon the rays run together and combine into a focus.

<i>: Konstantin Meyl: Zur Brennglaswirkung des Mondes bei einer Sonnenfinsternis, NET-Journal, Jg. 4, Heft Juli/August 1999, Seite 13-17

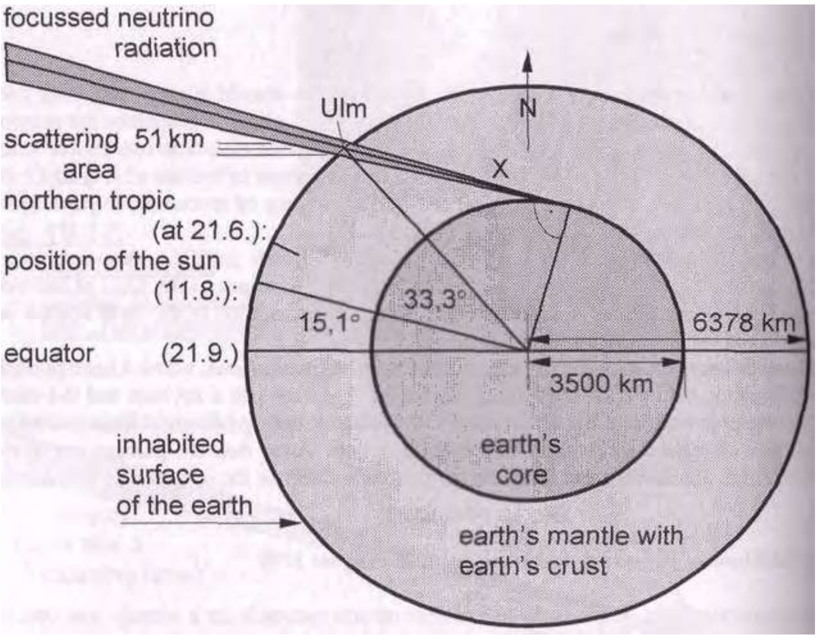
focussed neutrino radiation
Ulm
N
scattering 51 km area
X
northern tropic (at 21.6.):
position of the sun (11.8.):
33,3°
15,1°
6378 km
equator (21.9.)
3500 km
inhabited surface of the earth
earth's core
earth's mantle with earth's crust

Fig. 20.2: The course of the focussed neutrino radiation at 11.08.1999 for 48.4° northern latitude. (this corresponds to the line Ulm-Augsburg-Freising)

---

calculation of the critical latitude:

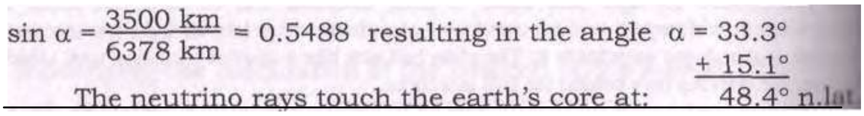

$\sin\alpha = \frac{3500\text{ km}}{6378\text{ km}} = 0.5488$ resulting in the angle $\alpha = 33.3°$
$+ 15.1°$

The neutrino rays touch the earth's core at: $48.4°$ n.lat

minimum distance to the moon: $r_m^* = 358000$ km
radius of the moon: $R_m = 1738$ km
radius of the earth: $R = 6378$ km (at the equator)
length of the Tangente: $X = 3500\text{ km}/\tan 33° = 5332$ km

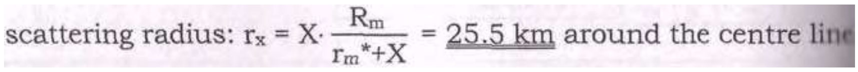

scattering radius: $r_x = X\cdot\frac{R_m}{r_m^*+X} = \underline{25.5\text{ km}}$ around the centre lin

Focussing without scattering on earth's surface increase to
$R_m / r_x = 68$ times , resp. with scattering to
$68 * 0.5 * 0.8 = 27$ times the natural neutrino radiation!

For that we again bear in mind that neutrinos as particle radiation propagate in space in the sense of a plasma wave. Since the oscillation of such a longitudinal wave takes place in the direction of propagation, it neither knows a distinct velocity of propagation nor an upper limit.

The sea of neutrinos in which we swim, is a combination of differently fast particles. The slow ones are bent stronger by the moon, with the property of a convex focussing lens, so that the focus should be looked for near the moon, whereas the fast ones hardly are diverted. Their focus lies further distant from the moon than the earth, up to the extremely fast cosmic neutrinos, which experience almost no diversion, because they, as a result of the Lorentz contraktion, are small enough to tunnel through any kind of matter.

If neutrinos, depending on their velocity of propagation, have their focus partly before and partly behind the earth at an eclipse of the sun, then it comes up to a conclusive logic that there actually exist such, of which the focus lies exactly in the centre of the complete shadow on earth's surface. But thereby the question is asked, which biological effectiveness these neutrinos have or which damage they bring about.

20.2 Model of calculation for the cosmic experiment

More than a year ago the results of an international neutrino experiment have been made public. Thereby for the first time the order of magnitude of the natural radiation density was recorded. after the detector before having been calibrated at an artificial source of neutrinos. As the perhaps most important result at night only half as much solar neutrinos could be detected as during the day with the immense Super-Kamiokande detector in a Japanese mine. With that the here taken interpretation, that earth's core collects neutrinos, meanwhile even is proven experimentally! For that it has to interact with the particles and that means, it has to exert a force of attraction on them. Thus the earth's core during an eclipse of the sun will further amplify the effect of focussing if the konzentrierte neutrino ray is directed on earth's core. This critical point we have to calculate (fig. 20.2).

The proportion of the radii of earth's core (3500 km) and the entire earth (6378 km) results in the sine of the sought-for angle and that amounts to 33.3 degrees northern latitude. At August 11 moreover is added that at noon the sun with regard to the equator is standing at 15.1 degrees. so that a first extreme focussing should be expected, if the centre of the complete shadow intersects latitude 48.4. This would be the case for the latitude of Ulm in direction Augsburg and Freising.

Now we would like to know more about the amplitude of the focussed radiation, about the spatial extension and the period of time. Without concrete data material we have to proceed from several simplifying assumptions. If we therefore assume, 50% of the at earth arriving and biologically relevant neutrino radiation stems from the sun, which in the case of an eclipse of the sun is focussed to 80% and scattered to 20% by the moon. If we further assume the focus just touches earth's core, then between Ulm and Augsburg a sphere of action of 50 km can be expected, within which the neutrino radiation on the average reaches 28 times the value of the natural radiation. The intense irradiation under these assumptions will last one minute.

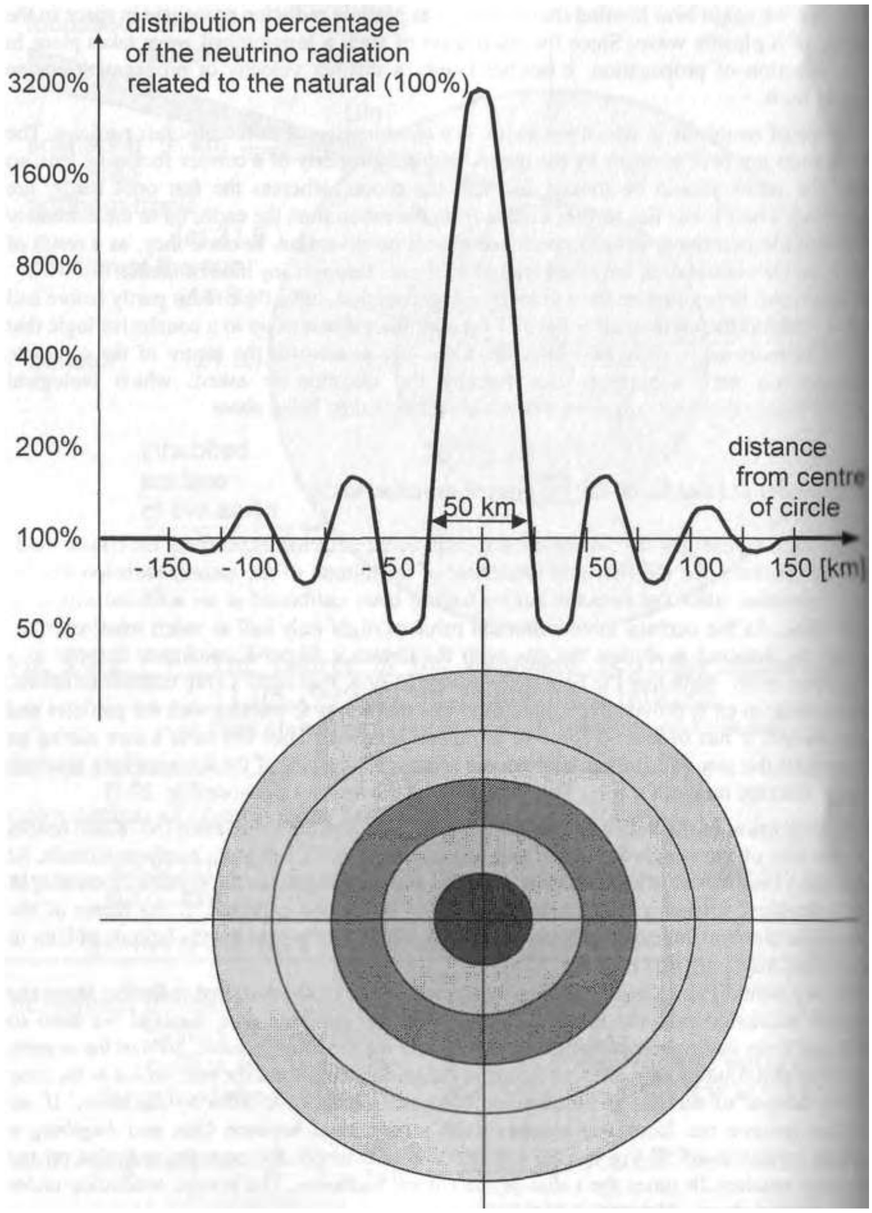

Fig. 20.3: ____ The spatial distribution in the case of focussing of the neutrino radiation by means of the burning glass effect of the moon on earth's surface.

Of course the calculated values only are valid to the extent, as also the assumed boundary values apply. After August 11 we know more about this cosmic experiment, about the spatial distribution and indirectly about the interaction with earth's core, about the physical properties and the biological effectiveness of the at present available neutrino radiation.

20.3 Physical technical consequences

The neutrino radiation is a scalar wave radiation which, as mentioned, can be perceived by sensitive people even without aids. Who hasn't at his disposal this sensitiveness, is recommended a simple setup.

For that one hangs up a fluorescent lamp, connects the one end with a piece of wire, as it were as antenna, and the other end is grounded. For a scalar wave radiation which increases fast, the lamp should start to glow by itself. Under big transmitting installations this method already has been successfully applied by many allotment gardeners thousandfold. I hence recommend all research scientists of eclipses to test themselves the possible field fluctuations in this simple way.

From a technical viewpoint first of all the atomic reactors and their nuclearly contaminated waste will be concerned by a fluctuation of the neutrino radiation. In view of present eclipses of the sun an accident can be expected less, since the neutrino radiation it is true for a short time reaches an extreme maximum, which averaged temporally and spatially over the whole event again is relativized somewhat. The relations shall be clarified with an example (fig. 20.3).

Whoever places himself in the centre line of the complete shadow on August 11, at first will detect a decrease of the neutrino radiation to 50 to 60 percent, then a steep increase to 2800 percent and from the summit again the whole backwards, while standing on the earth he turns by under the moving moon shadow. The ring with half the radiation, which reaches us first, doesn't pose a problem since, as said, we only have half the radiation in every night. Some animals and plants as aresult erroneously will set out for sleep.

The wave distribution one can imagine like that in the case of a pool, in which was thrown a stone. But we still don't know the resonance frequency, for which reason the length of the cycle depicted in fig. 20.3 is chosen arbitrary. The actual deviation from the distribution given by nature is the peak in the focussed ray centre. Living nature must stand large fluctuations of the solar radiation, since every supernova sends us a relatively short batch of fresh neutrinos. Where we have difficulties is the question, how much fluctuation still can be tolerated by mankind.

The question for possible biological consequences is due to be dealt with in view of the announcement of a complete shadow tourism causing concern, as it is awaited for August 11. At the incomplete state of knowledge about the properties of neutrinos, every trip to the complete shadow remains a journey into the unknown.

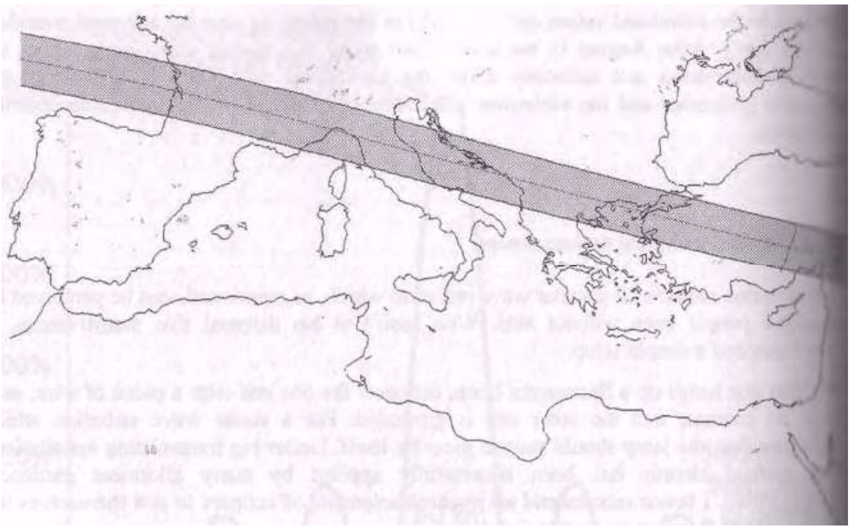

Fig. 20.4 A: The course of the complete shadow for the ancient eclipse of the sun at 28.5.585 B.C.

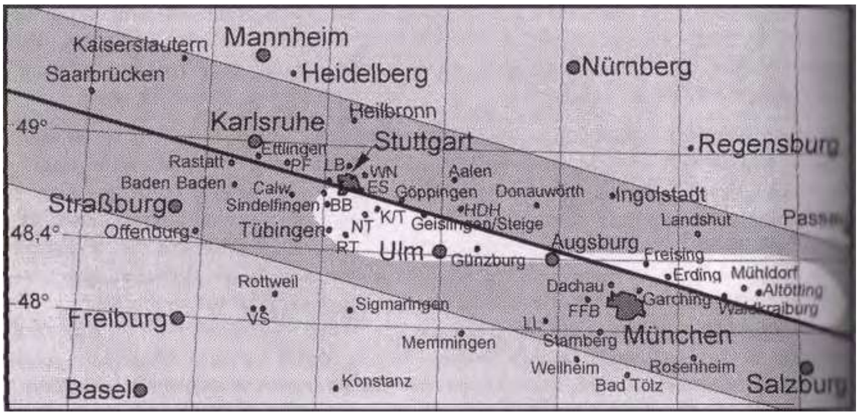
Kaiserslautern
Mannheim
Saarbrücken
Heidelberg
Nürnberg
Heilbronn
49°
Karlsruhe
Stuttgart
Regensburg
Ettlingen
Rastatt
PF
LB
WN
Aalen
Baden Baden
Calw
ES
Göppingen
Donauwörth
Ingolstadt
Straßburg
Sindelfingen
BB
HDH
K/T
Landshut
Pass
48,4°
Offenburg
Tübingen
NT
Geislingen/Steige
RT
Ulm
Günzburg
Augsburg
Freising
Erding
Mühldorf
Rottweil
Dachau
Altötting
48°
Garching
Waldkraiburg
Freiburg
VS
Sigmaringen
FFB
LL
Memmingen
Starnberg
München
Weilheim
Rosenheim
Bad Tölz
Salzburg
Basel
Konstanz

Fig. 20.4 B: The course of the complete shadow over south Germany at August 11 1999 (grey) and the possible course of the focussed neutrino radiation (white).

20.4 Biological consequences

In the case of an eclipse of the sun effects on the biology, like problems with the heart among affected, at least can't be excluded. If the scalar wave density increases above the density which is normal, then this has a positive effect on the economy of energy, as long as the body is capable to regulate the amount taken up. If the regulatory process however should fail. then the risk of a self-inflammation exists. Also straw bales and other organic and inflammable materials could thus go up in flames.

But before that happens, first the information technical influence of the scalar waves will show. Here we have to expect a psychotronic influencing, which is showing in a limited ability of perception. History teaches us as an example<i> that a by Thales of Milet predicted total eclipse of the sun at 28.5.585 B.C. compulsorily has ended a battle in Asia Minor between the Medes and the Lydians, because the soldiers apparently most literally had gone out of their mind (fig. 20.4 A).

Actually all in connexion with the free energy addressed phenomena are conceivable. from the "neutrinolyse" up to the acceleration of the radioactive decay. It would be understandable, if in a water glass bubbles should rise, even if no carbonic acid is contained in the water at all. After man not having at his disposal a sense organ for his own energy source, the brave in the complete shadow of an eclipse of the sun are recommended smaller technical experiments and observations. The cautious however will avoid the area from the start.

As counter movement to the complete shadow tourists there will also be refugees, who believe in the predictions of Nostradamus, who in his quatrains has predicted a messenger of fright over Europe for 11.8.1999. He mentions Lyon, Ulm and Moskau, which actually lie on one line<ii>.
The line of the complete shadow however will run under another angle from Plymouth in South England over Ulm to Bukarest and further into Turkey. Maybe Nostradamus wasn't a clairvoyant at all, but merely a good calculator, or he knew someone who could calculate excellently, after all he has indicated the time and even Ulm as the centre correctly. Apart from the small angle error, in addition the direction is correct (fig. 20.4 B).

Worth paying attention to also is, that he contrary to his habit here gives a concrete date which astronomically can be calculated unambiguously, that he simply skips the numerous eclipses of the sun of the past and only points to the one of 11.8.1999, which runs crossways through South Germany. Extremely sinister are his forecasts, which mustn't commented on further, since he speaks of "Mort et Tombe", of death and grave<ii>.

<i>: Merodot: Historien, Kroners Taschenausgabe 224, S. 33 (cf. page 586 and the modern misinterpretation note 9 on page 754); further sources: dtv-Atlas zur Weltgeschichte S. 45; resp. Propylaen-Weltgeschichte I, S. 168.

<ii>: Michel de Notredame (1503 bis 1566), Mathematiker, Astrologe, Leibarzt von Konig Karl IX: Centuries (1558) X, Vers 72.

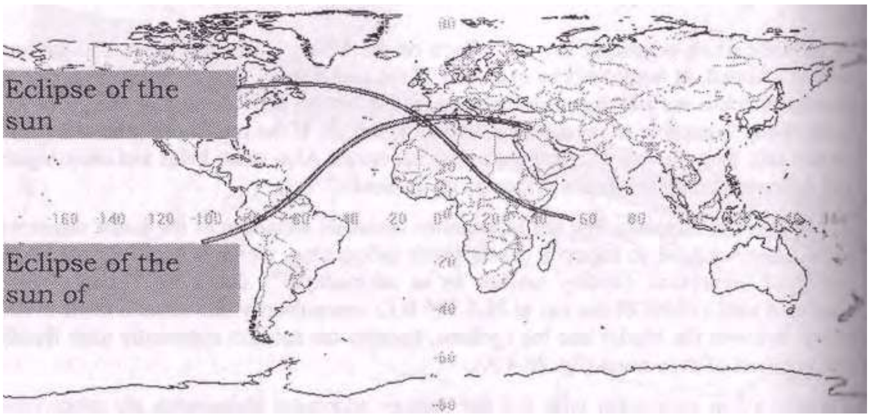
Eclipse of the sun
Eclipse of the sun of
160 140 120 100 80 60 40 20 0 20 40 60 80
80
20
40
60
80

Fig. 20.5 A: The total eclipses of the sun of 18.4.1539 and 21.8.1560

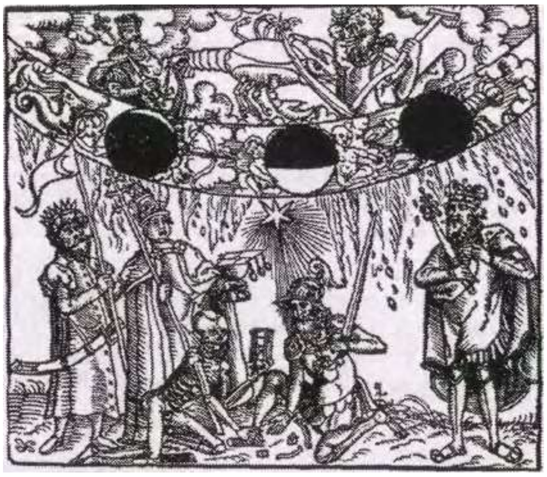

Fig. 20.5 B: Under the impression of the last two eclipses of the sun over Europe death and war play dice about the fate of mankind during the eclipse of the sun of 1562.[i]

<i>: Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, ISBN 3-89631-302-9, Seite 156.

20.5 Epilogue to August 11th 1999

My indications have been published two weeks before the cosmic event in the NET-Journal<i>. The public interest was immense. Now, after the spectacle is over, it is time for an analysis<ii>.

Besides numerous very subjective descriptions of eyewitnesses, which scientifically hardly can be analysed, at numerous polytechnics over the world has been observed, how Foucault pendulums suddenly and completely unexpected have deviated from their normal swinging direction. With this device 150 years ago the rotation of the earth had been detected in Paris.
Since according to prevailing textbook opinion an eclipse of the sun is a purely optical phenomenon, the scientists world-wide are having a big problem. Which force here has teared at our earth and caused relative accelerations of the kind that the pendulums could turn out of their usual plane and changed into an elliptic orbit, while the shadow of the moon ran over us? The gravitational force isn't even roughly capable to that. That merely has brought a 50 cm higher flood. The enormous force effect, which even puts the gravitation in the shade, actually only can come from the interaction of the neutrinos.

Different reports are present concerning the influencing of the radioactivity. In the cases, in which measurement samples have been used, almost no change could be observed. This is confirmed by a video tape, which northeast of Munich directly in the centre of the complete shadow documents an experiment, in which during the whole time the radioactivity of 1 kg crude granite is monitored with a professional dosimeter. The background might be that calibrating samples are chosen in principle under the aspect of being influenced the least by outside interference sources. It therefore would have been better, if we instead had put a lettuce as a biological and broadband sample in front of the device. Because there, where one anyway hadn't expected deviations and thus neither were measured calibrating samples nor unfortunately were made recordings, is said to temporarily have occurred a visible increase.
Even in the case that somewhere accidentily should exist recordings, the increase is too small, to help explain traditions from the Middle Ages, according to which is talked about "stinking log" and about "aggressive damps, which fall from the sky". Other sources speak of harmful radiations, which one regarded as the trigger of plagues.
It is said: "They poison the drinking water and the crops and make people sick"<iii>. Crops after an eclipse of the sun either should not be harvested at all or only after a violent shower. The warnings for a poisoning "without adding poison", which reach back into the 19th century, suggest the assumption that it concerns results of a radioactive contamination.

<i>: Konstantin Meyl: Zur Brennglaswirkung des Mondes bei einer Sonnenfinsternis, NET-Journal, Jg.4, Heft Juli/August 1999, Seite 13-17

<ii>: From the third and extended edition of this second part concerning Electromagnetic Environmental Compatibility the following chapters are completed.

<iii>: Mark Littmann/Ken Willcox: Totality - Eclipses of the Sun, Honolulu 1991, Kapitel 4: Eclipses in Mythology; nachzulesen bei Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, Seite 130/131.

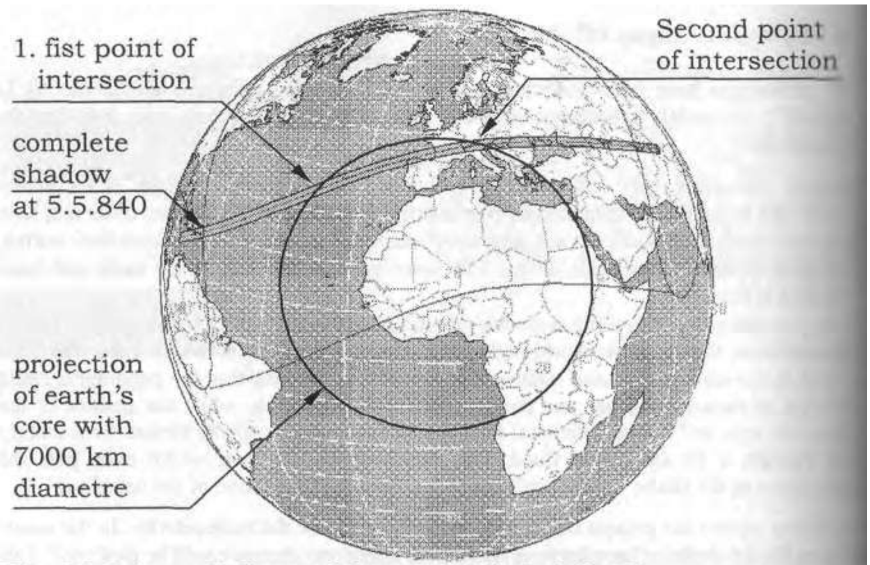
1. fist point of intersection
Second point of intersection
complete shadow at 5.5.840
projection of earth's core with 7000 km diametre

Fig. 20.6 A: Eclipse of the sun of 5. 5. 840 AD.

Fig. 20.6 B: Nostradamus announces catastrophes for 1999 in the Centuries of 1558.<i>

<i>: Bryan Brewer: Eclipse, Chapter 1: Eclipses Throughout the Ages, Seattle WA 1991, S. 20, cited in Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, ISBN 3-89631-302-9, Seite 130, Bild Seite 159.

20.6 Dropping of the neutrino radiation

Plagues of that kind, which in the Middle Ages have claimed lives as a result of an eclipse of the sun, hardly are verifiable, unless a prominent victim was among them, like e.g. the son and successor of Charlemagne, emperor Ludwig I. He 5.5.840 witnessed an eclipse of the sun with a totality of five minutes. Further it is said: „The fright which this experience gave him, little later shall have torn him to death"<i>.

According to that there must have existed times, in which the radioactivity present in the countryside and stored in the body of a person during an eclipse of the sun could be set free and be the undoing of the affected. At the same time such a decontamination of radiation acts cleaning for nature.

In this case only the interpretation of Nikola Tesla provides us an explanation, which states that the neutrino radiation causes the radioactivity<ii>. Textbook physics however doesn't know this causality. For that the primeval fears of humanity are pure superstition. The prophesies of Nostradamus even are referred to as counter evidence.

But the question remains open, why his predictions concerning August 11 didn't happen, Obviously, so has shown us the cosmic experiment, the neutrino radiation relevant for the setting free of radioactive radiation drastically has decreased since the Middle Ages.

Nostradamus personally has occupied himself with the translation and interpretation of hieroglyphs and has written down his insights - surely out of fear of the inquisition - in the form of encoded quatrains. According to that he has based his considerations on considerably older sources, which presumably stemmed from a time, in which a considerably higher radiation prevailed.

We have to proceed from the assumption that the scientists of the Semitic- Aramean people of the Chaldeans, which ruled Babylon from 626 B.C., were just as capable as astronomers of today, to exactly calculate an eclipse of the sun even centuries and millenia in advance.

After all the Saros-cycle to determine eclipses of the sun is a discovery of the Chaldeans. What they however couldn't know and we ourselves still can't indicate today, is the prevailing density of the cosmic radiation at a future time. But that obviously has changed considerably.

The natural magnetic field strength for instance is recorded at the baking of earthenware jugs and vases, by strengthening along the parts containing magnetite. From the measurement of ancient earthenware goods we know that in antiquity a field strength must have prevailed which was 3 to 4 powers of ten higher.

According to the here presented theory the earth owes its magnetism its core and that again draws its energy from the neutrino field. According to that also the neutrino radiation should be subject to the same decrease.

If therefore the Babylonians out of the radiation situation at that time in the interpretation of Nostradamus and other fortune-tellers, who presumably all more or less have written each other off, have predicted a catastrophe for 11.8.99, then this scientifically is just as untenable, as the today widespread hubris, with which the knowledge and the reports of experience from ancient times are dismissed as superstition.

<ii>: cf. chapter 17.2 Nikola Tesla, the discoverer of the neutrino radiation, p. 133

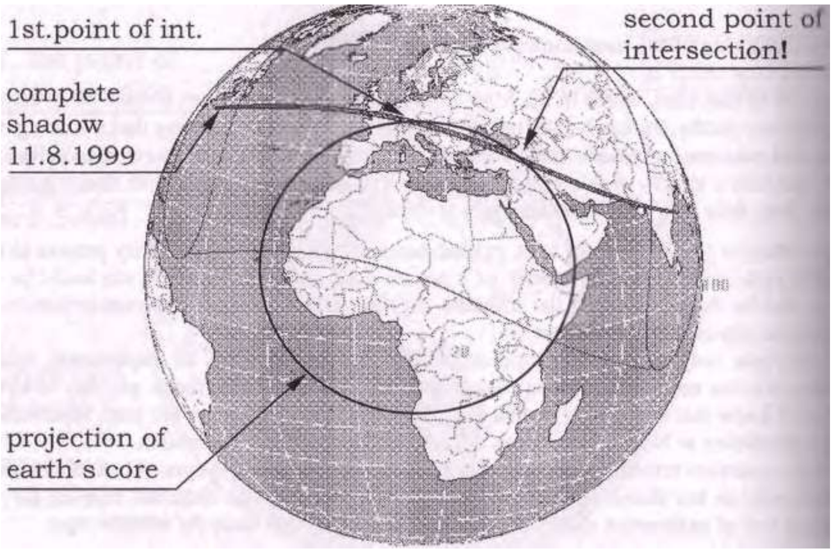
1st.point of int.
second point of intersection!
complete shadow 11.8.1999
projection of earth´s core

Fig. 20.7 A: Eclipse of the sun of 11.8.99 (declination: 15°)

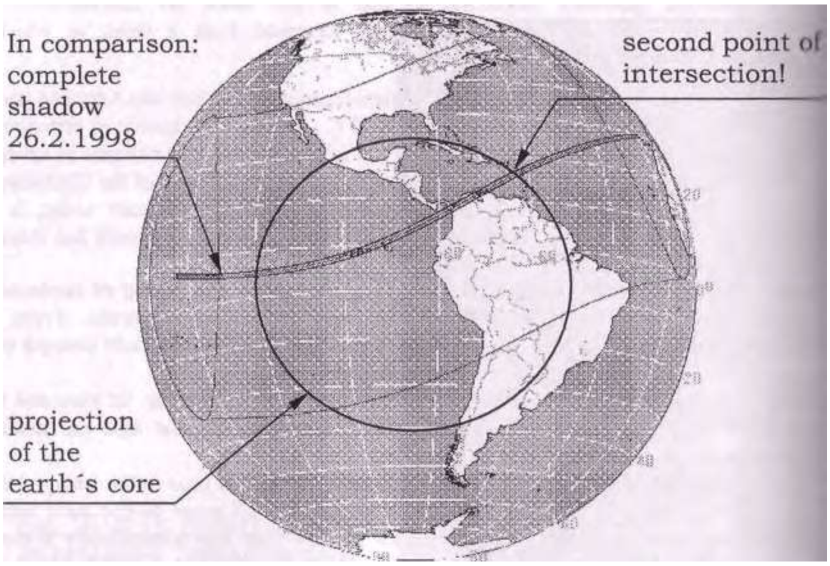
In comparison: complete shadow 26.2.1998
second point of intersection!
projection of the earth´s core

Fig. 20.7 B: Eclipse of the sun of 26.2.98 (declination: - 9°)

<ii>: Particularly violent earthquakes since 11.8.1999 (excerpt):
Turkey 17.8. (strength 7.8) till 19.8. (5,0), 31.8. (5.2), Greece 7.9. (5.8). Turkey 13.09. (5.8) and finally Taiwan 20.9.99 (strength 7.6) in a rythm of 6 to 7 days! Complete list in the internet under:
http://www-seismo.hannover.bgr.de/ermos_listing.html

This "superstition" mentions tremendous natural disasters like floods like the Flood or destructive earthquakes often as a direct result of an eclipse of the sun, and as the worst consequence the end of the world<i>.
Almost all reporters and newsreaders in the evening of August 11 full of irony pointed at the non-occurrence of the end of the world and spread the conviction that according to prevailing physical ideas something like that isn't possible at all. They at that time couldn't know that the anatolic plate, which the complete shadow had crossed, had gotten into motion. The relatively weak earthquakes, which 11.8. shook Cyprus and at the same time Iran, only were spontaneous harbringers.
In the following weeks the staggering core of the earth made us clear that there had been done force to it. Severe earthquakes with thousands of aftershocks followed each other and one message of terror chased the next<ii>. Public authorities however take care not to of decline to make a reference to the eclipse of the sun.

20.7 Analysis and comparison of the occurrences

If we for comparison consult the total eclipse of the sun of 26.2.1998, for which the complete shadow of the moon coming from the Pacific Ocean had run over the Caribbean into the Atlantic Ocean. Exactly the moment it crossed the Caribbean island Montserrat, the volcano Soufriere erupted.
Pure coincidence says science, which hasn't got an explanation model at all for a relation with the supposedly purely optical phenomenon. But this argumentation is relativized, if according to fig. 20.2 the tangential collecting of solar neutrinos by earth's core is considered. This process best can be compared with photon radiation, which is tangentially collected and directed into an orbit by a black hole inside the radius of Schwarzschild.
Since the shadow of the moon always draws a straight line on the earth, two points of intersection are present with the circle of the projection of earth's core on earth's surface, for which the radiation focussed by the moon just touches earth's core and in that way experiences an additional concentration. The first point of intersection at that time was situated in the Pacific Ocean; but the second one was situated exactly at the Caribbean island Montserrat (fig. 20.7 B).

This time, at August 11 1999 the first point of intersection was situated in South Germany, the second in Iran, and again the focussing at the second point of intersection has shown a devastating effect. The entire continental plate has gotten in motion (fig. 20.7 A).

Another strange phenomenon has occurred in the USA at the same time as the eclipse of the sun. A tornado swept with its destructive force right through Salt Lake City. It is remarkable that no meteorologic indications were showing before and hence official observation authorities had no possibility of warning for the tornado. Had here part of the focussed neutrino radiation been redirected at earth's core and given a rotation, to again screw out into the sky on the other side of the earth at Salt Lake City?

<i>: Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Miinchen 1999, ISBN 3-89631-302-9, Seite 120.

<ii>: Collection of particular violent earthquakes since 11.8.1999: see fig. 20.7.

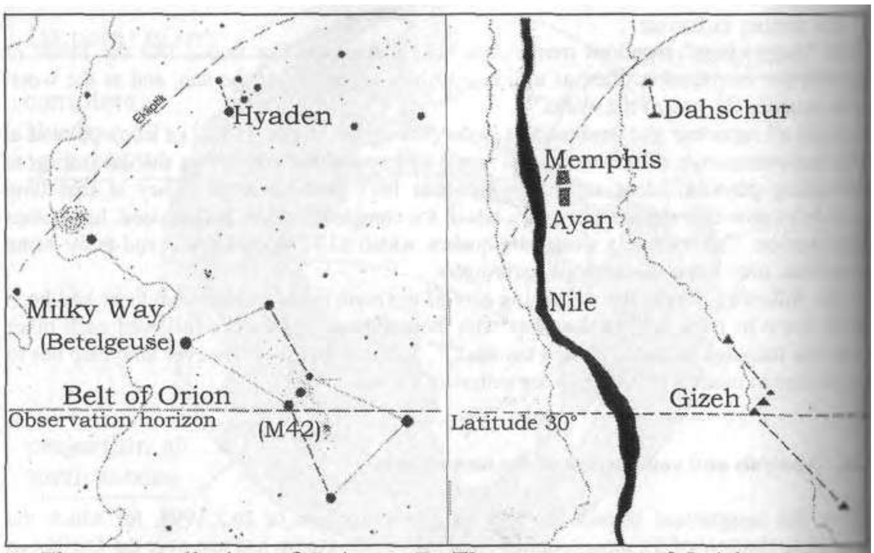
Ekliptik
Hyaden
Milky Way
(Betelgeuse)
Belt of Orion
Observation horizon
(M42)
Dahschur
Memphis
Ayan
Nile
Gizeh
Latitude 30°

A: The constellation of Orion B: The sanctuary of Osiris

Fig. 20.8 A and B: The ancient Egypt pyramids, a terrestrial model and copy of the starry sky?<i>

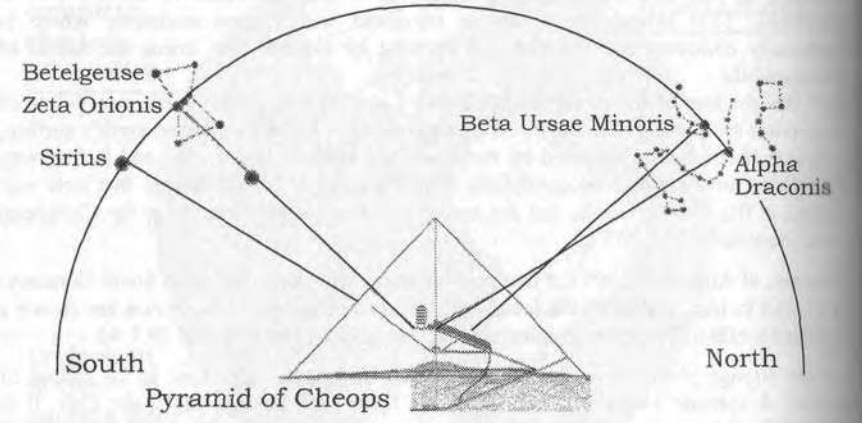
Betelgeuse
Zeta Orionis
Sirius
Beta Ursae Minoris
Alpha Draconis
South
North
Pyramid of Cheops

Fig. 20.8 C: The orientation of the four shafts of the Pyramid of Cheops to four different stars around the year 2500 B.C.<ii>

<i>: Robert Bauval, Graham Hancock: Der Schlussel zur Sphinx (Keeper of Genesis), List Verlag Munchen 1996, ISBN 3-471-77188-3, S. 286 und 287.

<ii>: dito page 86.

20.8 Radioactive decontamination with the help of a supernova

A mighty source of neutrinos forms the black hole in the centre of the Milky Way. For us, on an outer spiral arm of the galaxy, the distance to the centre is gigantic as well, so that a relatively equally distributed spectrum of differently fast neutrinos arrives at our world, which represents a kind of basic energy technical supply for the solar system and our world. For the fluctuations between day and night, or the focussing by the moon or other planets the "participants" in the solar system are responsible themselves.

In contrast to that a supernova, the explosion of a star, is a considerably smaller source of neutrinos, which however also is possible less distant to the earth. In addition it is a singular event, in which all neutrinos are set free simultaneously within a fraction of a second. They arrive at us one after another. First the fast and hard radiation reaches us as it were as harbringer. In the course of time the arriving neutrinos then become slower and shower, until they sometime become biologically relevant. If in the end everything is over, we can see the cause, only now the supernova is showing in the telescope.

If we assume such an event takes place, with perceptible might and in a distance of 500 light years. then this neutrino radiation overlaps with the general background radiation and a characteristic over-intensification of neutrinos of a certain velocity of propagation occurs. This problem then occupies us for 500 years, where the respective radiation situation permanently is changing depending on the time after the explosion.

If we in this way of looking dare a judgement of the cosmic events in historical time, so makes believe much the assumption that the radiation in the last hundred years has worn of completely. Edgar Cayce treats in the book "Our Ancestors" different cultural circles from the old Indian up to the Hopi, in which still is talked about an energy technical use of quartzes and other materials<i>.

We indeed can theoretically comprehend that the neutrino radiation can let an oscillating quartz glow, if it is stimulated in its resonance frequency, but technically the technology foday can't be realized anymore. Possibly the chance for technological use only existed for a few years or decades.

Presumably also the pyramids originally have been built as resonators, to slow down fast neutrinos to a technically utilizable speed. But in the course of time the original function was unnecessary and the neutrinos had gotten so slow that in antiquity alternatively an use as electrostatic lightning generator or as Nekropolis took place. Today they only stand in the countryside as unusable monuments of a gone epoch.

Many ancient techniques, to which I will come to speak in the third part, in this way just as unexpected become plausible as the radioactive decontamination described in the Middle Ages.

<i>: Hermann Wild: Technologien von gestern, Chancen fur morgen; Jupiter-Verlag Bern 1996, ISBN 3-906571-13-0, (z.B. S. 77 und 145 bis 163).

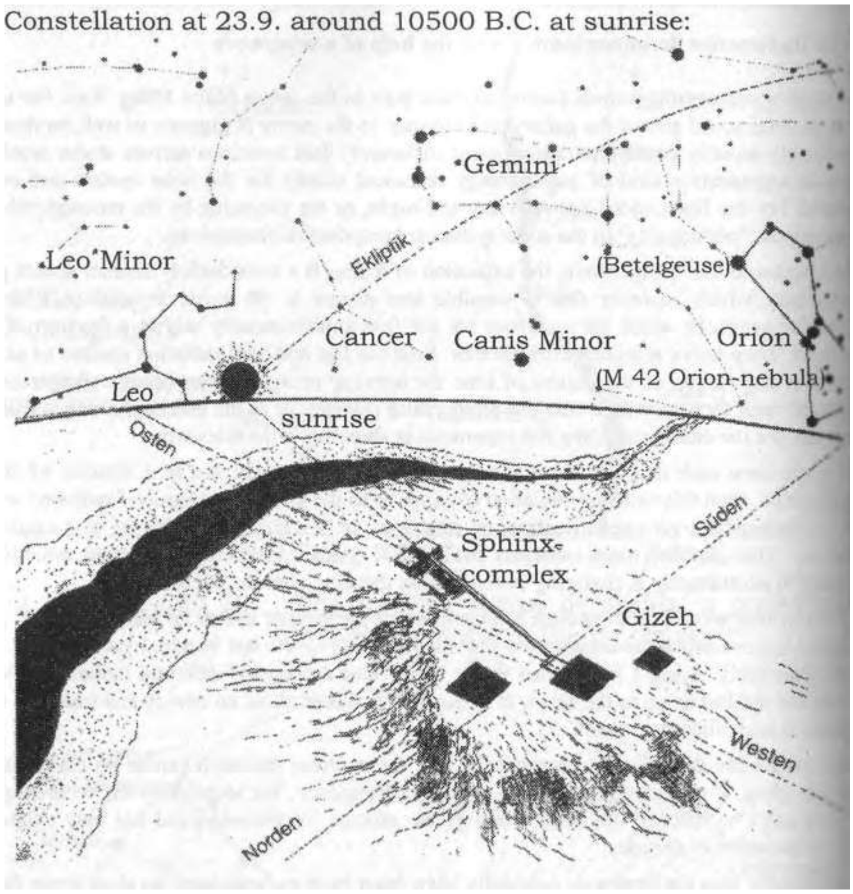
Constellation at 23.9. around 10500 B.C. at sunrise:
Gemini
Leo Minor
Ekliptik
(Betelgeuse)
Cancer
Canis Minor
Orion
(M 42 Orion-nebula)
Leo
sunrise
Osten
Süden
Sphinx-complex
Gizeh
Westen
Norden

Fig. 20.9: The orientation of the pyramids of Giza makes believe
an operation of the installation around 10500 B.C.<i>

<i>: Robert Bauval, Graham Hancock: Der Schlussel zur Sphinx (Keeper of Genesis), List Verlag Munchen 1996, ISBN 3-471-77188-3, S. 324.

The two research scientists take the hypothesis of an orientation of the ancient Egypt pyramids after the starry sky around 10500 B.C. This however is not compatible with their thesis of the aligning of the shafts to individual stars around 2500 B.C. (fig. 20.8 C). How should the precise worked shafts later have been integrated? It would correspond a certain logic, if an orientation of the buildings to the energy source at that time would have taken place; but doing so it should be taken into consideration that the stars are exploded as supernovae and today hardly might be observable.

Despite several good ideas more questions are raised by the book, than are answered. Have the technicians 2500 B.C., after making restorations, experimented with the buildings and sought-for alternatives of use?

20.9 Fre energy from the Orion constellation

After all hundred years ago still the last unreliable rests of the wearing off neutrino radiation were available to Tesla, Moray, Keely and other inventors for the experimental proof of free energy, of which we today aren't able to rebuild and show in function one single model. And that, although the technical aids have gotten better for many times. But for the free energy inventors that isn't a reason, to immediately stop their efforts, because the next supernova already is announced by the actual swings of the Foucault pendulum. It could concern the explosion of the giant red star Betelgeuse in the constellation of Orion in a distance of 500 light years<i>. It well may be possible that it already shortly has exploded and that it will supply us with fresh and free energy for scarcely the next 500 years and at the same time will give the earth a good shake, thanks to its might and a not too great distance.
The reaction, which happened after August 11, of earth's core, which is the first to interact with the fast particles, should make every astrophysicist clear that here something is coming towards us!
It would be obvious, if the reactions of earth's core still increase. Every year particularly around 21.6., if the sun is standing in the Orion constellation, deviations can be expected. But then the supposed source of neutrinos, Betelgeuse, the sun and the earth don't exactly form a line, because the giant red star lies 7.4° below the ecliptic. Because of that the rays slowed down and focussed by the sun run away over the earth. At the earth then rather would be expected a dropping of the radiation.
If the phase of shakes of the earth sometime should be over, a decontamination due to an increased radioactive decay and various biological effects should be expected. Then, perhaps in 200 years, also many concepts concerning free energy, today still dismissed as hopeless, suddenly will function entirely by themselves.

The relatively free possibility of development of the human mind and the present sciences we possibly owe the special circumstance to be able to live in a time of minimized field strengths. Strong fields however can lead to psychotronic influencing of the consciousness and to an outside determining of mankind. This circumstance seems to have caused Tesla to compare man with a robot, and to call him an independent machine controlled from outside<ii>.

<i>: Freek Reijmerink: Sternenatlas Deutschland, Weltbild Verlag 1990, S.12 (ein sterbender Riese) und Illustrierte Wissenschaft, Nr. 9, 1999, S. 7, Brennen die grossen Sterne in den Sternbildern je aus? Die Entfernungsangaben schwanken zwischen 270 Lichtjahren (Meyers Lexikon), 310 (Sternenatlas), 500 (Ill.Wiss.) und 652 Lichtjahren (Cambridge Enzyklopadie d. Astronomie).

In the observable domain of the starry sky one statistically seen has the chance, to experience every second a supernova.

<ii>: Nikola Tesla: How cosmic forces shape our destinies, New York American, 7.2.1917, and Edition Tesla (1997), Bd. 6, ISBN 3-89539-245-6 der Mensch als machine, S. 65.

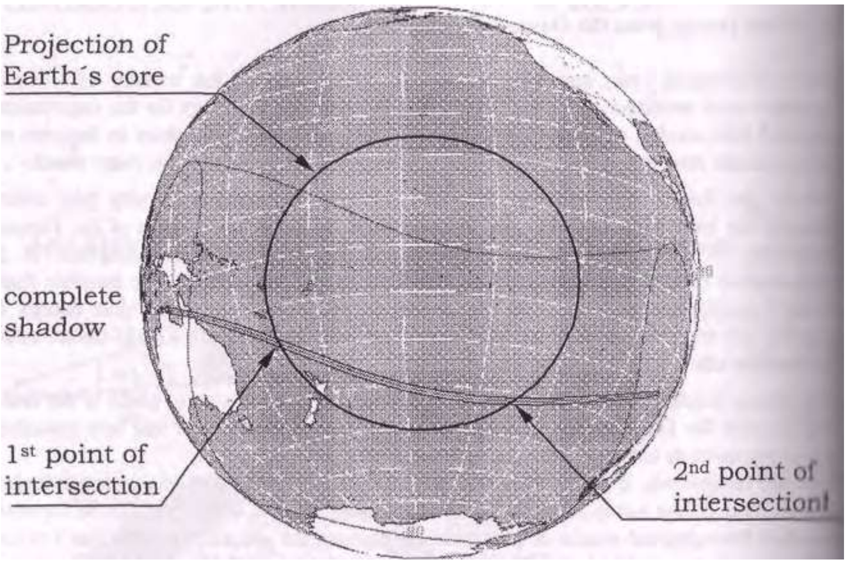
Projection of Earth´s core
complete shadow
1st point of intersection
2nd point of intersection

Fig. 20.10 A: Eclipse of the sun of 13.11.2012 (decl.: 18.1°)

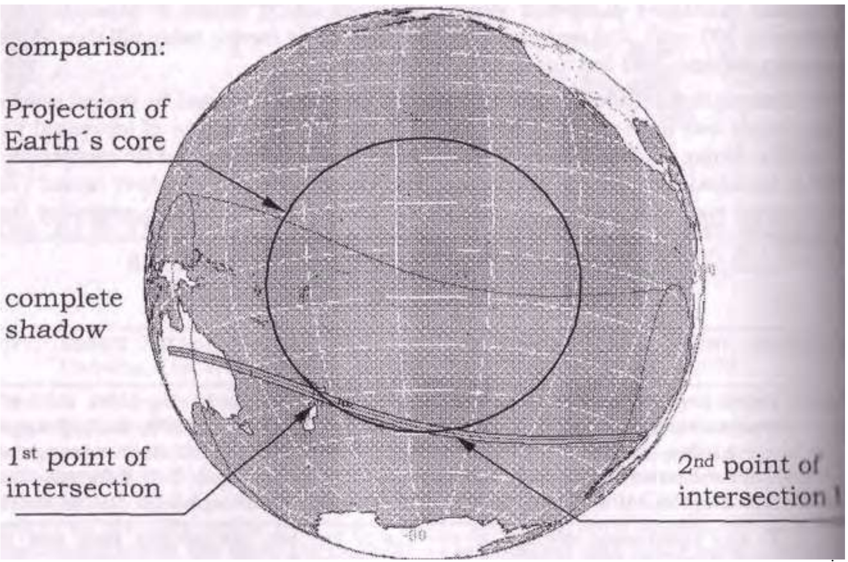
comparison:
Projection of Earth´s core
complete shadow
1st point of intersection
2nd point of intersection

Fig. 20.10 B: Eclipse of the sun of 11.11.3799 (decl.: -17.3°).<i>

<i>: The eclipse of the sun of 7.7.3797 is harmless compared to that of 11.11.3799. Have they been mixed up or is a calculation error present?

20.10 Interaction of the neutrinos with earth's core

A strong neutrino field still doesn't make a catastrophe. Only in connection with one of the regularly happening eclipses of the sun should one be expected under certain circumstances. Only, which eclipses of the sun can get dangerous, we have to ask us, and why warns e.g. Nostradamus only for very particular dates?
The check of the respective eclipses of the sun results in a critical constellation every time for the cases, where the line of the complete shadow and the circle of the projection of earth's core intersect under a very flat angle and both points of intersection lie very close together. In the extreme case finally the lines only are touching and the points of intersection fuse to a line of intersection.
A corresponding constellation the next time is expected at 13.11.2012. For 7.7.3797, at similar conditions Nostradamus foretells the end of the world; but why?

From the interaction of the neutrinos arises as from every other interaction a force effect. If the points of intersection lie far apart, then earth's core is pulled once to the East and a short time later again to the West by the focussed neutrino radiation. On the average this will hardly influence earth's mantle and earth's crust because of the immense moment of interia. The possible earthquakes will remain regionally restricted to the area around the two points of intersection.
But if a line of intersection forms, then no compensation of the force effects takes place anymore, then during the whole time one-sided is pulled at earth's core and that can have fatal results. It is the same as for a spinning top, which is given a blow from the side: it staggers several times, until the gyroscopic forces have stabilized it again.

But if earth's axis staggers, then the sun describes strange orbits in the sky, it goes backwards again, for a longer time doesn't set or it doesn't show for the same period of time for the people living on the other side of the globe.
Such an event already is described in the Bible<i>. For the twenty hours, in which in Europe the sun didn't set for a day, again describe the chroniclers of the inhabitants in the South American Andes, how at their place the sun didn't show for twenty hours<ii>.

As a further example from the Greek mythology is mentioned the description of the poet Apollodoros. according to whom Hercules for the solution of his 10th task let the chariot of the sun bring to a standstill. "He turned his vehicle round and raced the way back, dragging along the Pleiaden and all stars, so that the sun set in the East"<iii>. But if all stars take part in the same backwards motion, then this example proves the assumption of the staggering earth's axis.

<i>: Josua 10.13-14 and the in chapter 11.8 cited passages.

<ii>: nach Montesinos, zitiert in Zecharia Sitchin: The Lost Realms (Versunkene Reiche), Knaur Verlag Miinchen 1992, ISBN 3-426-04827-2, S. 203

<iii>: Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Milnchen 1999, ISBN 3-89631-302-9, Seite 18.

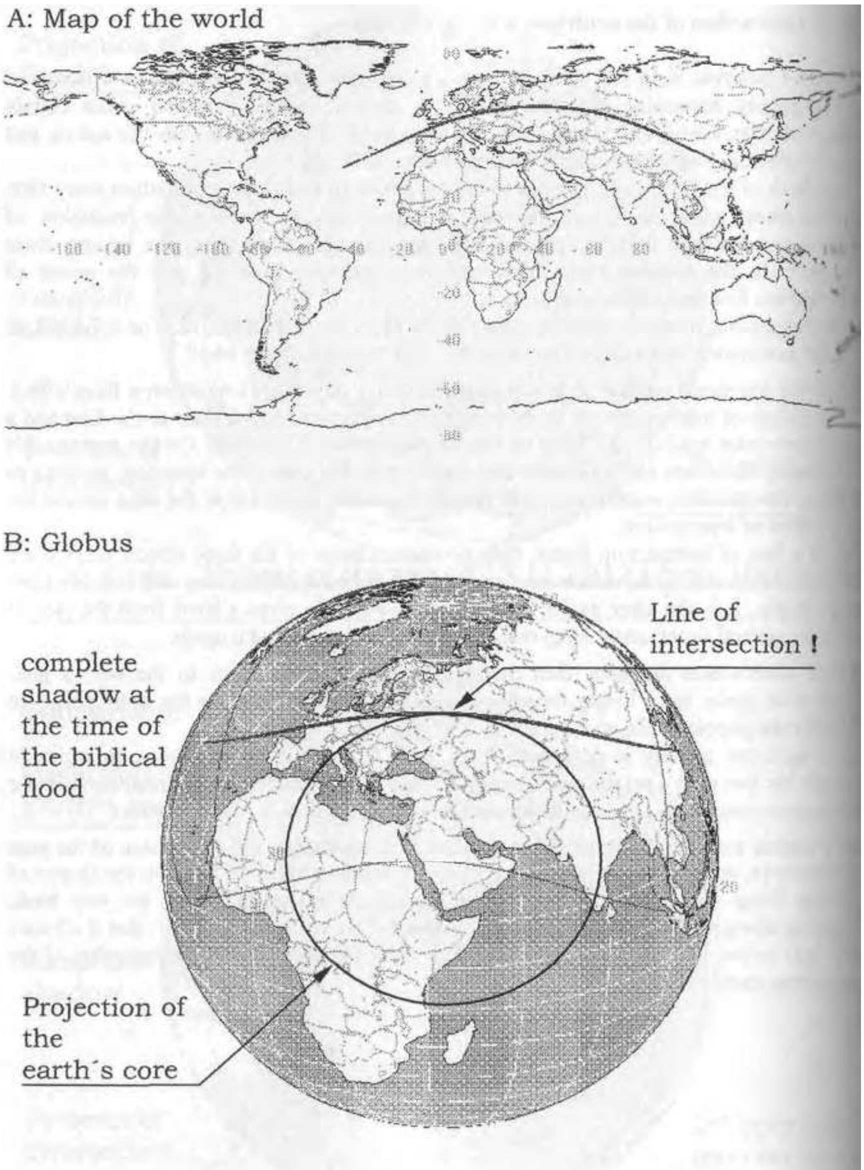
A: Map of the world
B: Globus
Line of intersection !
complete shadow at the time of the biblical flood
Projection of the earth´s core

Fig. 20.11: Eclipse of the sun of 27.07.2281 B.C.
(angle of the sun = declination: 19.3°)
The trigger of the biblical Flood?

20.11 Changing of polarity and apocalypse

The pulling at earth's core, which shows as a wobbling of the magnetic axis and in damped form can lead to a tilting of the axis of rotation of the earth, still doesn't make an apocalypse, no end of the world. This only can be expected, if after a tilting of the axis of rotation of earth's core a change of polarity on the surface of the earth occurs. Doing so the new magnetic poles, like a compass needle in the field of the solar wind, again align in such a that the North Pole will be lying "up" in the ecliptic. Because earth's surface will keep it direction of rotation, the sun will, after the process having stabilized, as usual again rise in the East and set in the West. But the inhabitants of the earth, which before still were having midsummer, find themselves again in the midst of midwinter and vice versa.

Truly apocalyptic processes can be expected during the phase of a change of polarity of the earth. Thereby occur unusual relative accelerations and violent earthquakes. The largest destructive potential however is present in the waters of the oceans, which are set in motion.

As is well-known the earth at the equator is measuring a radius which is 21 kilometres larger than at its poles. If only a part of the waters temporarily flows in the direction of the poles of the earth, then the biggest part of the habitable land in Middle and North Europe sinks in the floods; then indeed also the statement of Noah makes sense, who as the first thing saw the mountain of Ararat rise from the floods, after the water again flowed back into its usual ocean basins. The mountain of Ararat after all measures a height of 5137 meters above sealevel!

At comparing historical events with details from the Bible the Flood should have taken place in the year 2245 B.C.<i>. According to the description of the position of the stars Dr. Wild calculates July 2281 B.C. as time for the Flood. The Arabic historical writer al-Makrizi again shifts the event into the year 3094 B.C.<i>. Who is right? We must verify the eclipses of the sun in this time and determine the position of the points of intersection, then we perhaps find the correct answer. Possibly earth's axis has wobbled more than once, have occurred several catastrophes in different regions. At 27.7.2281 B.C. in any case there actually has occurred an extremely critical constellation, whereas the other two years are ruled out. Here no total eclipse of the sun took place (fig. 20.11).

According to the calendar of the Ugha Mongulala 6110 years before this Flood a still much more devastating one should have occurred. That therefore would have been 8391 B.C., while Scott-Elliot dates the catastrophe in the year 9564 B.C.<ii>. According to Plato it would have been about 9500 B.C. For such long periods of time a check however isn't quite easy anymore, because the meantime changes of the earth sum up considerably.

<i>: Wild S. 231, 229 and 225.

<ii>: Wild S, 219, 218 and 210, among others cited from Scott-Elliot: Atlantis und das untergegangene Lemuria, Bauer Verlag Freiburg 1977.

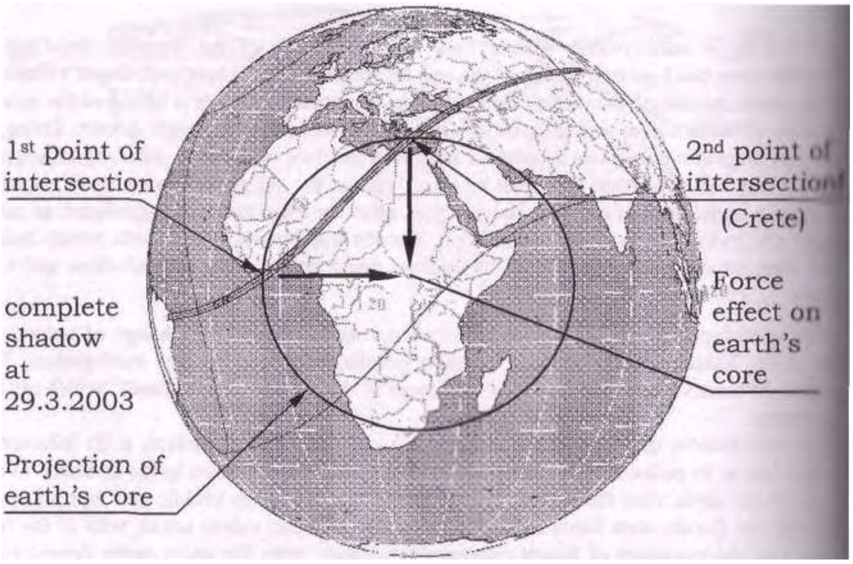
1st point of intersection
2nd point of intersection (Crete)
complete shadow at 29.3.2003
Force effect on earth's core
Projection of earth's core

Fig. 20.12 A: Eclipse of the sun of 29.03.2006 (decl.: 3.2°)

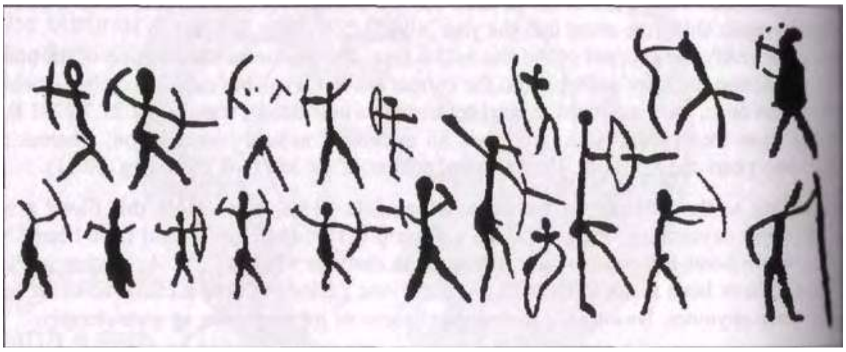

Fig. 20.12 B: Cave painting from Minateda (Spain).<i>

<i>: Laviosa-Zambotti: Origini e Diffusione della Civilta, Marzorati, Milano 1950,

Besides the constellation of sun and moon is crucial the occurring of a relevant neutrino radiation and the question if both is sufficient to tilt the earth and change its polarity. Some experts expect a change of polarity for the time coming, since it takes place with a certain regularity and measured in earth historical periods of time in addition fairly often!
This circumstance the earth presumably owes the river valleys cut deep in the countryside and other topographic phenomena. One presumably can only survive such a catastrophe in an ark (Noah), in the air bubbles of large caves (the walls of which painted children and artists out of pure boredom, see fig. 20.12 B) or in the highland, preferably in the area of the equator (Central Africa, highland of Mexico, Andes, Himalaya).

20.12 Scalar wave gauges for the prediction of earthquakes

We urgently need gauges, to be able to judge the neutrino situation. At first we with that pursue the same goal, as with the building of a free energy converter, however with the difference that the converter should maximize the collected amount of energy, whereas the gauge should minimize the taken up energy, to not too much load the source and not to change the local radiation situation.
In addition the neutrino radiation should be registered distinguishable in its velocity of propagation, what means that the building of a gauge will be very much more costly than that of an energy collector. That's why it can be expected that an usable measurement instrument might be available only many years later and we that long only can base on our own power of observation.

We for instance can statistically analyse, how the earth after an eclipse of the sun reacts to the neutrino radiation focussed by the moon. At August 11 the second focal point was situated in the East, and from that an acceleration of the rotation of earth's core can be predicted. Corresponding observations actually have been made on the surface of the earth with the help of the Foucault pendulum.
The relation the next time can be checked at the eclipse of the sun at 29.3.2006. This time the inverse case is present. A first focussing takes place in the West, so that pendelum swings in the reversed direction would be expected, which indicate a slowing down of the rotation (fig. 20.12 A).

But if the rotation of the earth should change, then the balance sheet of angular momentum of moon, earth's mantle and earth's core isn't correct anymore. As a result a force of difference occurs, which lets the spinning top stagger. But if earth's core staggers inside of earth's mantle, then it powerfully stirs the liquid magma, and in this way releases its surplus energy again. As a result earth's mantle is heated up somewhat. We, on our wafer-thin earth's crust, then time delayed feel the effects of the staggering of the core as an earthquake. The period seems to lie at approximately 6.5 days, as far as this can be read from reactions to the last two eclipses of the sun.

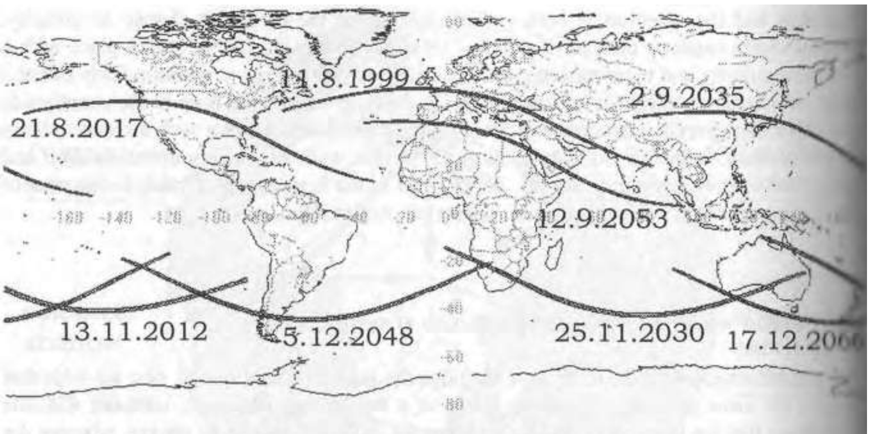
11.8.1999
2.9.2035
21.8.2017
12.9.2053
13.11.2012
5.12.2048
25.11.2030
17.12.206

Tafel 20.13 A: Eclipses of the sun of 2 Saros cycles.

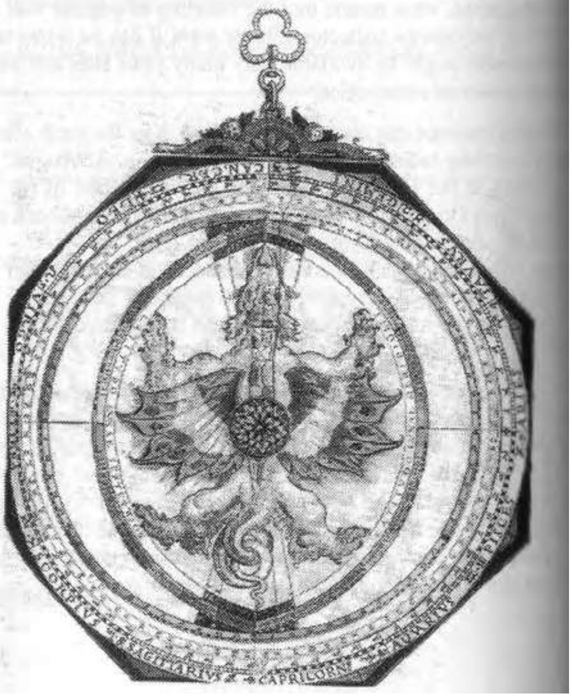

Fig. 20.13 B: The „Astronomicum Caesareum“ served around 1540 the determination of the positions of the moon nodes.[i]

---

[i]: Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, ISBN 3-89631-302-9, Seite 134.

20.13 Calculation of future occurrences

From the analysis of past and present cosmic events with a strict scientific procedure can be predicted to a certain extent also future events. It thereby by no means concerns prediction, but exclusively the result of an analysis.

After the mentioned eclipse of the sun of 29.03.2006 there once more exists danger of earthquakes (approx. 4.4.2006) for the Island of Crete and Asia Minor. From the eclipse of the sun of 13.11.2012 as well no good can be expected, even if the points of intersection, situated close together, lie far away from populated land in the south-west Pacific basin. The Maya calendar by the way ends to this time. At 21.8.2017, 18 years or a Saros cycle after the eclipse of the sun at 11.8.1999 and correspondingly 120° further to the west, the corresponding complete shadow runs crossways through the USA. The thankless role of Turkey at the second point of intersection this time takes over South Carolina. One only can hope that the houses in Columbia are built more stable than in Izmit.
In fig. 2013 A the two eclipses of the sun of 11.8.1999 and of 13.11.2012 with their respective Saros cycles until 2066 are shown. The as critically to value tendency of the course of the complete shadow is visible, which wants to nestle against the circle of the projection of earth's core to form a line of intersection.

I here break off, since anyone with my indications and an eclipse of the sun-CD can analyse at home all further events personally. I value that if possible many analyses are made and controversially discussed, because possible cosmic catastrophes concern us all somewhat.

What good is the building of gigantic fusion ovens, if the runners by no means have understood the process of the fusion themselves? Why build ring accelerators for billions of dollars, if elementary particles can be calculated at the desk? How is the expenditure for gravitational wave detectors justified, if the actual music plays at entirely other velocities of propagation? Why is half the annual production of the world of Gallium used for an indirect proof of neutrinos, if every self-wound Tesla coil is able to collect more neutrinos?

Our scientists, for whom I as a colleague quite often must be ashamed, have the primary task to draw attention to cosmic and other risks and to calculate them in advance. In any case it is extremely unpardonable to leave this core duty up to some fortune-tellers and self-appointed prophets.
One should more often remind them that the Chinese Kaiser Tschung-Khang let sentence his court astronomers Hi and Ho to death, after the two not having predicted an eclipse of the sun. It thus certainly didn't concern a missed spectacle or a missed tourism business, but presumably the need for safety of the Kaiser and his subordinates, for which the two astrophysicists in ancient China had to take responsibility!

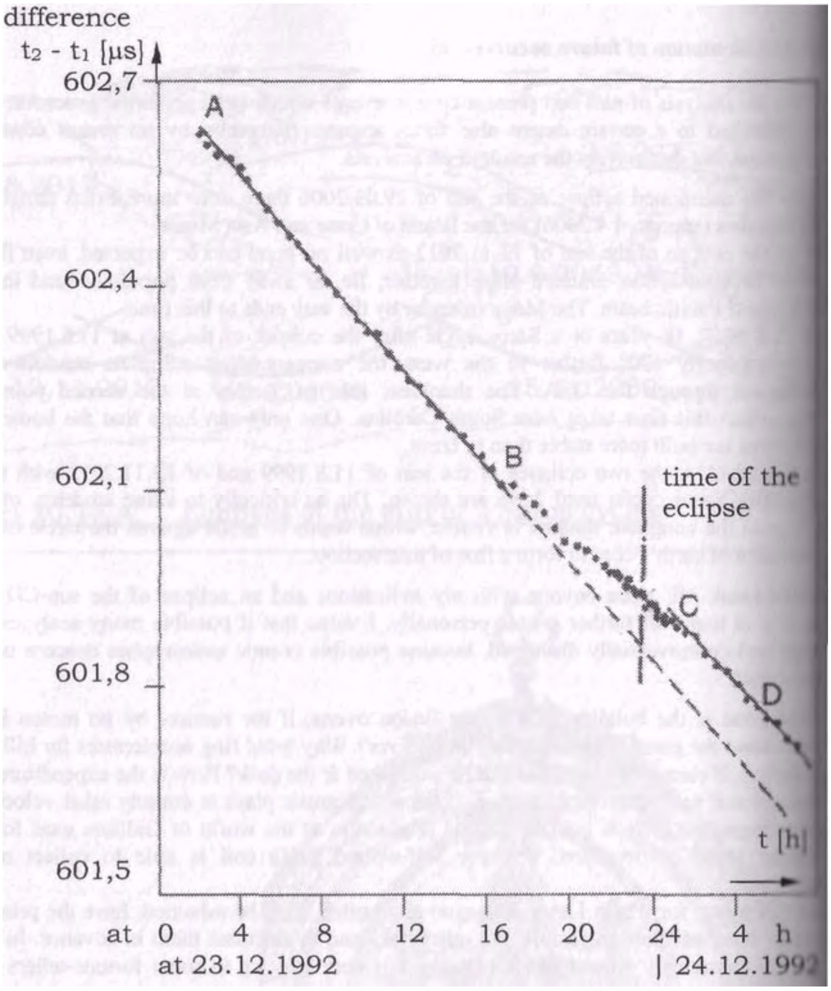

Fig. 20.14: The difference in going of two caesium atomic clocks positioned in different directions, but in the same laboratory in Harbin, China, during the partial eclipse of the sun of 24.12.1992.[i]

[i] Shu-wen Zhou: Abnormal Physical Phenomena Observed When the Sun, Moon, and Earth are Aligned, 21st Century Science & Technology, Fall 1999, Vol. 12, No. 3, pp. 54-61. Comments concerning Figure 9: Straight lines AB and CD show that the rate of change of the time difference between the two clocks is constant in non-eclipse periods, but becomes irregular around the time of the eclipse.

20.14 Epilogue to the energy technical seminar

Very slowly the word goes round that the energy source of the future has got a name: "Neutrinopower". Unfortunately fundamental physics, which is financed with public money, still does know almost nothing about the nature of the neutrino radiation. In addition hinder useless model concepts any progress in this direction and so the responsible fundamental research steps on the spot.

Considerably more extensive were already 100 years ago the insights of the experimental physicist Nikola Tesla, the discoverer of the neutrino radiation and father of the free energy. The space energy however is showing most clearly in nature, which only uses this advantageous form of energy. In particular during eclipses of the sun and other cosmic experiments it openly comes to light and can be detected by us for a short time.

The physicist Prof. Shu-wen Zhou of the University of Huazhong in Wuhan, China, systematically has investigated the effects, if sun, moon and earth are aligned<i>. Doing so he has proven inexplicable physical anomalies in experiments. Stimulated by the discoveries of Maurice Allais with the Foucault pendulum he built an arrangement specially flor proving horizontal forces of acceleration, and actually he with that could measure face effects during the total eclipse of the sun of 24.10.1995. He even speaks of an oscillating force!

Further he could determine changes in the spectral wave length of various elements, which under normal conditions go as constant and even as characteristic for the respective element. The relative change of size of the wave length during the ring-like eclipse of the sun over China of 23.9.1987 resulted in the 100-fold value compared to the difference in the spectrum analysis between surface of the earth and surface of the sun! This comparison reveals an immense discrepancy between theory and practice and puts us for a solid problem.

For that six different models of spectrometers were installed in several laboratories of different polytechnics and a photograph was taken of the emission spectra of H, D, Ca, CN, Ni, Ti, etc. Also other reasons than that of an eclipse of the sun could be excluded unambiguosly. In the results of these, at artificial light carried out conventional measurements, in any case abysses yawn. The from spectrum analyses won "insights" about the composition of strange celestial bodies now for sure can be done away with without knowledge about the respective prevailing neutrino radiation.

Spectacular also is the proof of differences in going of atomic clocks of various constructions. During the partial eclipse of the sun of 24.12.1992 seven caesium clocks in four cities of China and in three planes were used. The analysis of the differences in going resulted in, as is shown exemplary in fig. 20.14, changes of the gradient during the eclipse. The results for the atomic clocks in the planes and for two further time measurements turned out with similar clarity<i>.

<i>: Shou-wen Zhou: Abnormal Physical Phenomena Observed When the Sun, Moon, and Earth are Aligned, 21st Century Science 6s Technology, Fall 1999, Vol. 12, No. 3, pp. 54 - 61.

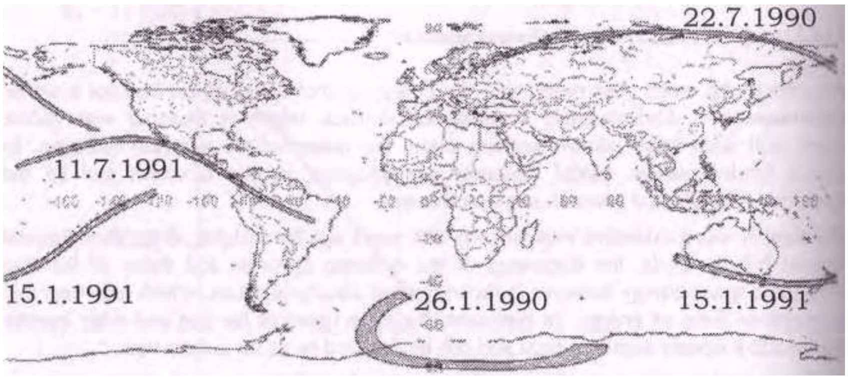

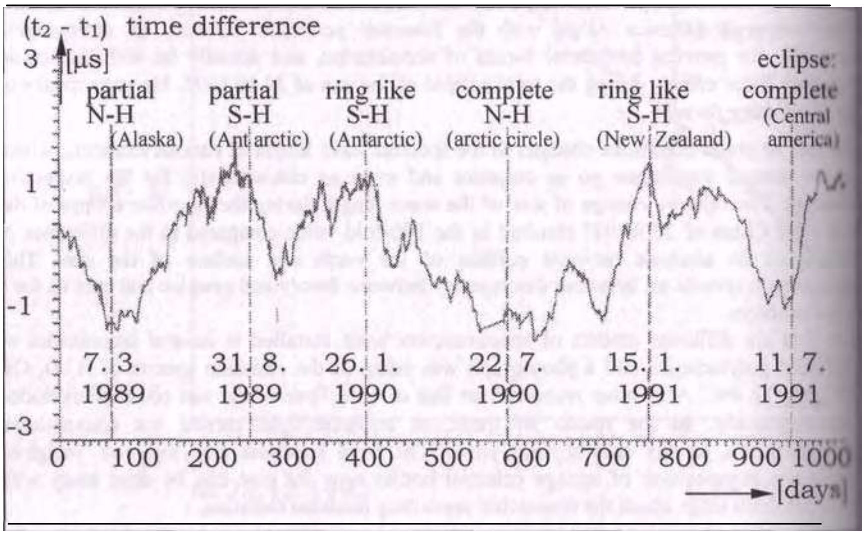

Fig. 20.15: Long-term measurement between 1989 and 991 of the difference in going of two atomic clocks at the U.S. Naval Astronomical Observatory.<i>

The in the 1000 days occurred eclipses of the sun have been entered later. In addition is recorded, if the northern hemisphere (N-H) or the southern hemisphere (S-H) of the earth was involved.

<i>:Shu-wen Zhou: Abnormal Physical Phenomena Observed When the Sun, Moon, and Earth are Aligned, 21st Century Science & Technology, Fall 1999, Vol. 12, No. 3, pp. 54 - 61. Figure 8.

A connection to the neutrino radiation Prof. Zhon doesn't draw, but it almost is obvious.

From the U.S. Naval Astronomical Observatory (LC/7970) long-term measurements over 1000 days between 1989 and 1991 are present (fig. 20.15). The difference in going between the two atomic clocks positioned at different places shows a permanent up and down. The reason goes as completely unknown.
But if the eclipses, which took place in this time, are entered then one immediately sees the assignment to a maximum or a minimum value. If an eclipse of the sun namely concerned the southern hemisphere of the earth, then the difference in going each time reached a maximum, but if the shadow of the moon run over the northern hemisphere, then each time a minimum occurred. Chance here probably is out of the question!

Let us record: The effect of an eclipse of the sun, to which for instance a Foucault pendulum reacts, can equally be traced back to the interaction of the neutrinos as the free energy. That's why the here presented book carries the title "Free energy and the interaction of the neutrinos". It has appeared in the series concerning the "Electromagnetic environmental compatibility", and also for that there are good reasons. Our energy technology must become more ecologically compatible and we come the goal closer, if we emulate nature, understand and copy it.

That it further concerns electromagnetism, likewise is explained from the interaction of the neutrinos, which concerns the oscillating and resonant case of the electromagnetic interaction. In this respect the reference to the series of books therefore would be given.

It only indirectly has to do with "electrosmog", or what otherwise is understood under environmental compatibility in general. It however could be shown that also the earth radiation is a form of neutrino radiation, and that it poses a biological effectiveness, after it turned out that it serves nature as an energy source. With that the conflict with space energy devices already is predetermined if the same scalar wave radiation should be used. Then one system takes away the other system the energy basis and the existence basis.
This environmental compatibility problem can be solved, if care is taken that both don't get in each others way with regard to the frequency and the wave length. We for that need a deep understanding concerning the topic of the space quanta and the neutrinos, their physical properties and the corresponding device technology. The book should make a contribution to that.

The other side of the medal is the information technical aspect of the scalar wave radiation and the environmental compatibility problem connected with that. The third and last part of the series of books is dedicated to this theme.

Part 2: Edition belonging to the energy technical seminar: "Electromagnetic Environmental Compatibility"

Prof. Dr.-Ing. Konstantin Meyl

Scalar waves

Abstract:

1. Auflage 1998, 4. Auflage and 1st English Edition

With regard to the environmental compatibility a decentralized electrical energy technology should be required, which manages without overhead power lines, without combustion and without radioactive waste. The liberalization of the energy markets won't on any account solve our energy problem, but only accelerate the way into the dead end. New, ecologically compatible concepts are collected and discussed in the book.

A useful energy source could be represented by space quanta, which hit upon the Earth from the sun or from space. They however only are revealed to the measurement technician, if they interact. It will be shown that the particles oscillate and an interaction or collection with the goal of the energy technical use only is possible in the case of resonance.

Since these space quanta as oscillating particles have almost no charge and mass averaged over time, they have the ability of penetration proven for neutrinos.
In the case of the particle radiation discovered 100 years ago by Tesla, it obviously concerns neutrinos. We proceed from the assumption that in the future decentral neutrino converters will solve the current energy problem. Numerous concepts from nature and engineering, like on the one hand lightning or photosynthesis and on the other hand the railgun or the Tesla converter are instanced and discussed.

Free energy and the interaction of the neutrino radiation

INDEL GmbH, Verlagsabteilung, Fax: +49-7721-51870
Tel: +49-7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen
www.k-meyl.de meyl@k-meyl.de

Scalar Waves

From an extended vortex and field theory to a technical, biological
and historical use of longitudinal waves.

Part 3

by

Professor Dr.-Ing. Konstantin Meyl

Edition belonging to the lecture and seminar
"Electromagnetic environmental compatibility"

(Original title: "Elektromagnetische Umweltvertraglichkeit")

Translated out of the German language
by Ben Jansen
(2003)

* * *

Part 3: Edition belonging to the information technical seminar

Scalar waves and the technical use of longitudinal waves and vortices

* * *

INDEL GmbH, Verlagsabteilung

Villingen-Schwenningen 2003

ISBN 3-9802 542-4-0

Preface to the seminar, part 3

With the appearing in print of the 3rd part of the series of lecture notes of the electromagnetic environmental compatibility, the collection of material can be considered to be completed for the time being. By now almost all aspects concerning the theme of the potential vortices of the electromagnetic field and their propagation as a scalar wave should have been addressed or at least indicated.

And that's what it is about in the case of the editions belonging to the seminars: In the seminars, taking place inside and outside the university, questions about the phenomena and the biological effectiveness of the fields, which surround us, are discussed. The scientifically correct procedure asks as a preparation for the seminar a voluminous collection of material and a compilation of theses with the pros and contras to them. The argumentation, which is denoted as „textbook physics" and is generally accepted, the participant of the seminar can work out himself by means of the available journals and books. The potential vortex, which completed the field equations, however makes possible a counterstatement, which for most participants at first will be unusual. The old dispute between quantum and field physics rekindles. What is missing is a voluminous collection of material concerning the refined field physics, and this gap the editions belonging to the seminars should close.

In the case of this book the primary concern is to make sure, that all facets, properties and physical consequences of the new and unified field physics are addressed. The question, if all points can bear a strict verification, remains reserved to the seminar and its participants, of whom each should form its own judgement. The series of books should be understood as a stimulation, to make own thoughts about the various points. Who has no opportunity to participate in a seminar, by means of the lecture at least gets an idea of what is discussed here and how is struggled for scientific insights.

The herewith completed collection of material by no means may be compared or confused with a basic work about the theory of objectivity. That the editions cannot and want not afford at all. They together with the discussions at best form the basis for a scientifically basic work, at which I am working.

In the case of the collection of material concerning the electromagnetic environmental compatibility consisting of three parts, again and again new aspects have emerged, which found entrance into the book, at which I was working at that particular time, whereas the rough structure has been controlled precisely: part 1 with the chapters 1 to 9 treats the basics, part 2 with chapters 10 to 20 the energy technical aspect and part 3 with the chapters 21 to 30 the information technical aspect of scalar waves.

The here presented 3rd part starts with the wave equation and the two comprised parts of a transverse and a longitudinal wave. The historic dispute between Heinrich Hertz and Nikola Tesla with the experimental evidence of the each time used wave part is continued over the wave-particle uncertainty compromise up to the dispute about the right field description: that of Maxwell or the new dual and at the same time unified description, which builds upon Boscovich and Faraday. The aspects, which seem so irreconcilably thereby only are two parts of a single equation, which is much older and can be traced back to Laplace: the wave equation.

Until now no derivations of this equation, which contains both wave parts, are known This for the first time succeeds from the new and extended field approach in the summary from chapter 26. At first however is attempted to make the world of the scalar waves plausible with models, calculations and observations. For that are used examples from high-frequency engineering, like the ground wave or the near-field area of an antenna, and from nature and medicine, which are granted two chapters. The comparison of the nerve conduction with Tesla's one wire technology is bringing it to light that scalar waves equally are used in both cases. In a frequency diagram the insights are entered concisely and is shown that considerably more unexplored domains occur than already is known, that in most cases not even gauges are available. With the design for building a scalar wave gauge the scientific verification of until now only empirically won results is stimulated, like for instance the results the radiesthesia wants to have determined with the help of the sensitivity of man as a biosensor.

Many relations will be revealed to the reader only, if he has worked through the summary, which follows. From Maxwell's field equations only the well-known (transverse) Hertzian waves can be derived. (Longitudinal) scalar waves however in the result are zero. This is a flaw of normally used field theory, since scalar waves exist for all particle waves, like e.g. as plasma wave, as photon- or neutrino radiation. Because the field pointer, in the direction of which longitudinal waves are propagating, is oscillating, the frequency will oscillate like the velocity of propagation, which again is measured as a noise signal. Any antenna noise proves the emission of scalar waves in space. But scalar waves, or whatever should be subsumed under the by mathematics minted generic term free of any value judgement, surely are more than only noise.

Starting from Faraday's discovery - instead of the formulation of the law of induction according to Maxwell - an extended field theory is derived, which goes beyond the Maxwell theory with the description of potential vortices (noise vortices) and their propagation as a scalar wave, but contains the Maxwell theory as a special case. The new field theory with that doesn't collide with the textbook opinion, but extends it in an essential point with the discovery and addition of the potential vortices.
Also the theory of objectivity, which follows from the discovery, is compared in the form of a summary with the subjective and the relativistic point of view and the consequences for variable velocity of propagation of the scalar waves, formed from potential vortices, are discussed. Like already in part 1 the unification of the electromagnetic interaction with the gravitation succeeds impressively.

Besides the mathematical calculations this book contains a voluminous material collection concerning the information technical use of scalar waves, if e.g. the useful signal and the usually interfering noise signal change their places, if a separate modulation of frequency and wavelength makes a parallel image transmission possible, if it concerns questions of the environmental compatibility for the sake of humanity (bioresonance, among others) or to harm humanity (electrosmog). With that the book again finds back to the starting point, to the open task, which made necessary an excursion through all domains of physics to answer it. I hope, the long march was worthwhile.

Villingen-Schwenningen december 2002

Preface to the 2nd edition of part 3

In favour of chapter 30 as a complement, an introduction into the ancient broadcasting technology of the gods, the 1st edition wasn't printed anymore in the year 2002 as originally planned. Instead the individual chapters have been prepublished in different places. The chapters 21 to 25 are found among others in the book "scalar wave technology", which has been published as an instruction for an experiment to demonstrate scalar waves in 2001. Excerpts from it, as also from the chapters 26 to 28 have appeared in form of individual essays in different journals.

Again experimental successes had been intervening, which caused a renewed deferring of the date of appearance of the 3rd part of the series of books. In my laboratory the bi-directional transmission of music and of data by scalar wave was successful. We thus also can transmit information backwards from a receiver to the transmitter or to a second receiver, whereby the receivers work purely passive, thus without own power supply. The operating energy is as well supplied to them by the transmitter by scalar wave. The demonstrated technology opens completely new possibilities of a technical application. Conceivable are telemetry installations where measurement signals have to be transmitted from rotating or otherwise inaccessible places of a machine. The energy for the measurement electronics can be transmitted wirelessly by scalar waves and the signal can be sent back along at the same time, by modulating it onto the energy carrier. In this way dozens of measurement stations can be connected wireless with a single central transmitter, which supplies them all with energy.

The question was asked: Is the technology really entirely new? The answer is amazing: No, it here concerns the oldest technology of humanity, which had developed to a peak in antiquity, to send receive engineering of the gods. For this claim even a mathematical proof is available. For that the authoritative transition, the unrolling of a vortex to a wave (the transition from the near-field to the far-field) or in the reversed case the rolling up at a receiver antenna (usually denoted as standing wave) is calculated with help of the extended field theory. The result is, that at this transition the velocities of propagation resp. the wavelengths of the transverse and the longitudinal wave stand to each other in the ratio of the Golden Proportion.
With regard to the optimization of a transmitter or receiver antenna the Golden Proportion has an effect on the construction resp. the architecture of corresponding buildings . Which ones, with that deals the complementing chapter 30. But it mustn't be missing either, since after all it concerns a grandiose practical information technical use of scalar waves, from which we technologically can learn a lot<i,ii>.

Villingen-Schwenningen march 2003

<i>: Note: chapter 30 provides an introduction into ancient broadcasting technology of the gods. The working off of history with respect to the use of scalar waves is so voluminous, that for that a book of its own is published with the title:

<ii>: K. Meyl: Sendetechnik der Gotter, historischer Sciencefictionroman (2003), Villingen-Schwenningen, INDEL Verlagsabteilung, ISBN 3-98O2 542-5 9

21. Derivation of the scalar wave

WilliamThomson, who called himself Lord Kelvin, after he had been knighted, already in his lifetime was a recognized and famous theoretical physicist. The airship seemed him too unsafe and so he went aboard a steam liner for a journey from England to America in the summer of 1897. He was on the way in a delicate mission.

Eight years before his German colleague Heinrich Hertz had detected the electromagnetic wave in experiments in Karlsruhe and scientists all over the world had rebuilt his antenna arrangements. They all not only found confirmed the wave as such, they even could show the characteristic properties.

It was a transverse wave, for which the electric and the magnetic field pointers oscillate perpendicular to the direction of propagation. This can be seen as the reason, that the velocity of propagation is showing itself field independent and constant. It is the speed of light c.

With that Hertz had experimentally proven the properties of this wave, previously calculated in a theoretical way by Maxwell, and at the same time proven the correctness of the Maxwellian field theory. The scientists in Europe were just saying to each other: "well done!" as completely other words came across from a private research laboratory in New York: "Heinrich Hertz is mistaken, it by no means is a transverse wave but a longitudinal wave!"

Such a screwball most said and did as if they hadn't heard the criticism at all. But then one couldn't ignore it completely, because on the one hand claimed the private research scientist to have experimental proof and on the other hand it wasn't just anybody, who here reported. It was nobody but Nikola Tesla, the Croat experimental physicist who emigrated to America

Him we owe the modern alternating current technology from the high-tension network for energy transmission over the alternating current transformer to the asynchronous machine.

With his magnificent inventions he had earned enough money, to be able to afford a private laboratory, in which he could research and invent uncensored and free. The key to his success was lying in his purposeful, concentrated and efficient working method.

He was fast! Whereas in Europe still was being discussed about properties and theoretical possibilities of application of the wave, Tesla already presented the armed forces a remote controlled scaled-down submarine in Madison Square Garden (fig. 21.1 A). To convince such a man, who only holds valid what his experiments reveal, from the opposite, should be a hopeless enterprise.

Lord Kelvin was aware of that, as he made the decision to go on the journey. He could not and did not want to put his head into the ground, as many of his colleagues, because on the one hand scientists are curious by principle and on the other hand he travelled as an official representative of science. He had been instructed to free, as Mr. Clean, the undamaged world of sciences from erroneous teachings. But it came completely different.

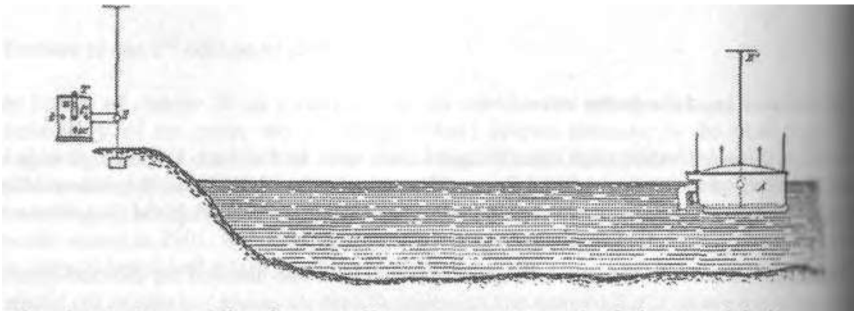

Fig. 21.1 A: The first radio remote control of the world.<i>

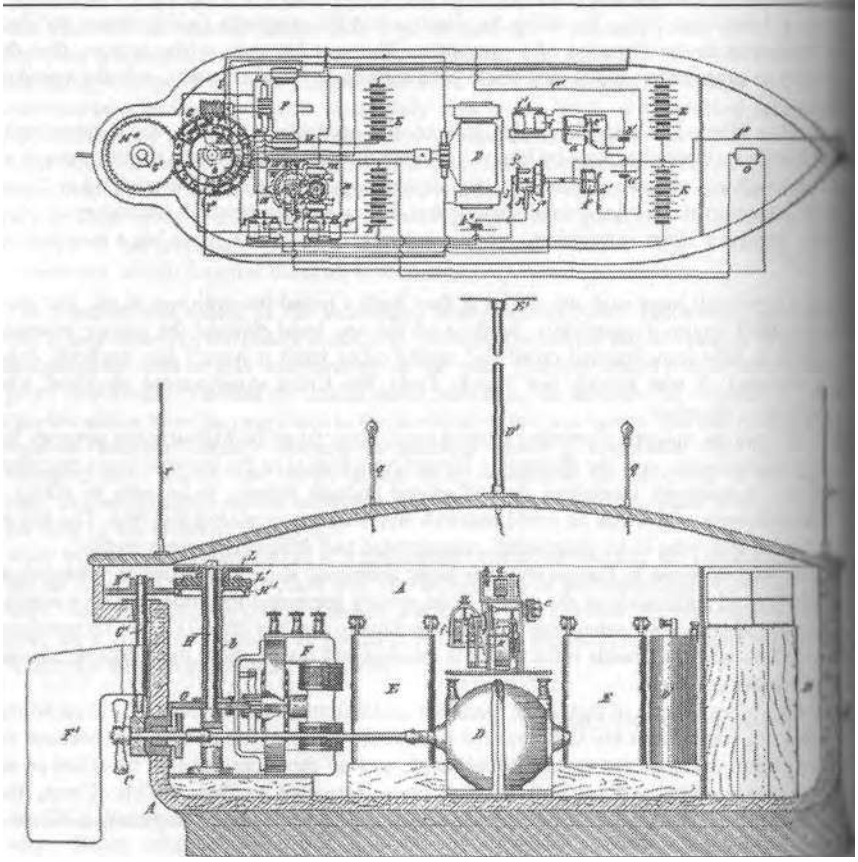

Fig. 21.1 B: Patent specification concerning the remote controlled submarine.<i>

<i>: Nikola Tesla: Method of and Apparatus for Controlling Mechanism of Moving Vessels or Vehicles, US-Pat. 1898, Nr. 613,809. Complete Patents: P. 351

21.1 Lord Kelvin in a delicate mission

The first day of his visit at Tesla Kelvin spoke the admonishing words he had been instructed to speak. He recommended Tesla warmly to publicly retract the remarks concerning the Hertzian mistake and to contribute himself to the settlement of the dispute. Seen from a specialist viewpoint they talked at cross-purposes.

But at night in his hotel room the Lord again thought about the experiments, which had been shown to him. The standing wave nature had been visible unambiguously: the oscillation nodes, the effect back on the transmitter, the high degree of effectiveness and many other things more. Such properties the Hertzian wave indeed doesn't know. Also Tesla didn't work with dipole antennas, but with flat coils, with spark gap oscillators and with a very unconventional switching technique, set up different in principle.

The next morning Lord Kelvin appeared in the laboratory again and greeted Tesla with the words: "Then you don't use Hertzian waves?" "Certainly not", Tesla answered," it are radiations. By waves no energy could be economically transmitted over a larger distance. My system works with true conduction, which theoretically seen can take place over a larger distance, without bigger losses occurring."
In the article of the "Electrical Experimenter" is noted further, that the doubting critic Kelvin suddenly turned into one of the biggest followers<i>.

Kelvin deduced very fast: according to that there exist two different sorts of wave propagation. So Hertz with his transverse wave is just as right, as Tesla with the longitudinal wave.
As a represantative of theoretical physics he however could pull out Tesla a decayed tooth. Maxwell had based his field description on an aether concept, which at that time in the world of sciences triggered violent discussions. Since Tesla saw such an aether as a prerequisite for longitudinal waves, he thought he and not Hertz had proven the Maxwell wave in experiment for the first time. By stating this the magnificent experimental physicist however revealed weaknesses in the area of theory. Maybe he had not read exact enough or understood the books of Maxwell, which without doubt were formulated mathematically only arduously comprehensible in the original version.
In this point Tesla had to learn otherwise by Kelvin. Maxwell's field theory provides without exception a mathematical description for the Hertzian wave. For the Tesla radiation however no field description exists! This by the way is the circumstance, why this wave could disappear from the textbooks and again fall into oblivion. Tesla himself had problems to theoretically imagine, what happens at his wave. His models in some points perhaps were even better than the official textbook opinion, but not without contradiction to accepted regularities. That's why Tesla did without a publication of his ideas, although he in his lifetime had filed away at an own theory.

<i>: Nikola Tesla: Famous Scientific Illusions, III. The Singular Misconception of the Wireless, Electrical Experimenter, Feb. 1919, printed in Tesla Said, p. 197

Figure 21.2 A:

William Thomson, 1846 becoming Professor at Glasgow University with the age of 22 years. From 1892. on he was called Lord Kelvin.<i>

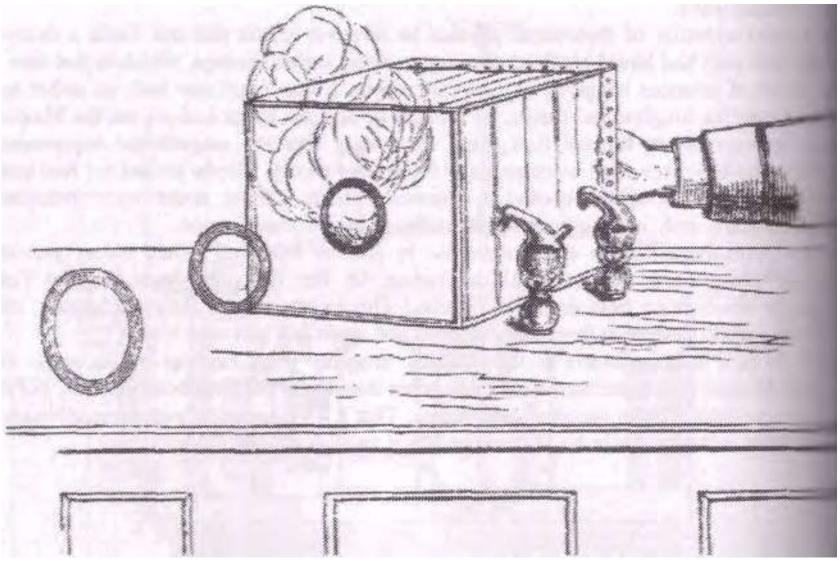

Fig. 21.2 B: Vortex rings from a smoke vortex gun.<i>

<i>: David Ash, Peter Hewitt: Science of the gods, Gateway Books, Bath, 1990

21.2 Helmholtzian ring-like vortices in the aether

Tesla told Kelvin at his visit from the meeting with the German Professor Hermann von Helmholtz on the occasion of the World's Fair in Chicago 1893. Kelvin knew him very well and had cooperated with him in the past. Now the vortex concept of his colleague and his model of stable vortex rings were very obliging.
In the case of a standing wave the impulse is passed on from one particle to the next. In the case of acoustics for instance we are dealing with a shock wave, where one air molecule knocks the next. In this way sound propagates as a longitudinal wave. Correspondingly the question is raised: What sort of quanta are the ones, which in the case of the Tesla radiation carry the impulse?

Lord Kelvin was already on the way back to Europe on the steamship and he deduced: The Tesla experiments prove the existence of longitudinal standing waves in space. In the question, what passes on the impulse, Kelvin comes to the conclusion: it are vortices in the aether! With that he had found an answer in experience. With his students he built boxes, with which he could produce smoke rings, to be able to study and demonstrate in experiments the special properties of ring-like vortices in their flow technical analogy (fig. 21.2<i>). But he didn't have ready a suitable field theory.

The from Germany to the Isles exported vortex physics for a short time could establish in England, before it was slaughtered and buried by the German quantum physicists. A main advocate has been James Clerk Maxwell, who held the vortex theory for the best and most convincing description of matter<ii>. As his successor at the Cavendish laboratory in Cambridge J. J. Thomson was appointed, who already as a young man had got a price for a mathematical treatise about vortices<iii>. He discovered the electron and imagined it, how could it be otherwise, as a field vortex.

The crucial weakness of vortex physics, the lacking of an usable field theory, was of benefit to the emerging quantum physics. This could change fundamentally with the discovery of the potential vortex, the vortex of the electric field<i4>!
In addition is the experimental proof of a vortex transmission as a longitudinal wave in air or in a vacuum, as it has been furnished by Tesla already 100 years ago, neither with Maxwell's theory nor with the today normally used quantum theory explicable or compatible. An urgent need is present for a new field theory!

<i>: David Ash, Peter Hewitt: Science of the gods, Gateway Books, Bath, England 1990.

<ii>: James Clerk Maxwell: "...the vortex rings of Helmholtz, which Thomson imagines as the true form of the atom, fulfil more conditions than any other previous concept of the atom."

<iii>: J.J. Thomson: "the vortex theory is of much more fundamental nature than the usual theory of solid particles."

<i4>: Konstantin Meyl: Potentialwirbel Band 1 (1990) and Band 2 (1992), INDEL-Verlag, Villingen-Schwenningen.

Scalar waves, a mathematical reasoning (I)

1. prerequisite: wave equation (textbook-formulation)

$$\boxed{\Delta \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}}$$

taking apart of the delta operator mathematically (Laplace operator) according to the rules of vector analysis (fig. 5.0):

$$\boxed{\begin{array}{lcccc} \Delta \mathbf{E} & = & \text{grad div } \mathbf{E} & - & \text{rot rot } \mathbf{E} \\ \text{wave} & = & \text{longitudinal} & + & \text{transverse} \end{array}} \quad (21.1)$$

---

2. State of the art of technology: Hertzian wave = transverse wave
special case: = solution of Maxwell's field equations
no sources:

$$\boxed{\text{div } \mathbf{E} = 0} \quad \text{and} \quad \boxed{-\text{rot rot } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \quad (21.3)$$

transverse wave:
field pointers oscillate crosswise to the direction of propagation
The propagation occurs with the speed of light c.

---

3. claim: Tesla radiation = longitudinal wave
special case:
irrotationality:

$$\boxed{\text{rot } \mathbf{E} = 0} \quad \text{and} \quad \boxed{\text{grad div } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \quad (21.4)$$

longitudinal wave, shock wave, standing wave:
field pointer oscillates in the direction of propagation.
Velocity of propagation is variable!

Fig. 21.3: The special cases of the wave equation

21.3 Taking apart of the wave equation

Before one plunges into the adventure of an entirely new field theory, it first of all should be traced and analysed, what the latest textbooks say about scalar waves.

There some scalar or vector potentials are introduced; there the constant of dielectricity e is written down as a complex variable, although it physically seen concerns a material constant, only to be able to calculate with this trick artificially a loss angle, which should indicate the losses occurring in a dielectric, where in reality it concerns vortex losses. Of course one can explain the dielectric losses of a capacitor or the heating in a microwave oven entirely without vortex physics with such a label fraud, but it should be clear to anyone, that in a complex constant lies buried an inner contradiction, which is incompatible with physical concepts.

We are used to such auxiliary descriptions so much, that the majority of today's physicists tend to attribute physical reality to this mathematical nonsense. As pragmatists they put themselves on the standpoint if with that experimental results can be described, then such an auxiliary description can't be so wrong after all. Doing so the circumstance is forgotten that here the ground of pure science is abandoned and is replaced by creeds.

We find everything already in the wave equation, as it can be found in all textbooks.

$$\boxed{\Delta \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \quad (21.1)$$

Behind this formulation two completely different kinds of waves are hiding, because the used delta operator consists of two parts according to the rules of vector analysis:

$$\boxed{\begin{array}{lll} \text{grad div } \mathbf{E} & - \text{ rot rot } \mathbf{E} & = \Delta \mathbf{E} \\ \text{longitudinal} & / \text{ transverse} & \text{wave} \end{array}} \quad (21.2)$$

We want to discuss two special cases.

If we put the left part (in eq. 21.2) to zero (div E = 0) which is tantamount to no sources of the field then the well-known radio wave remains, which also is called Hertzian wave, after Heinrich Hertz, as said, had experimentally detected it in Karlsruhe 1888:

$$\text{div } \mathbf{E} = 0 \quad \text{and} \quad \boxed{- \text{rot rot } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \quad \text{(special case)} \qquad (21.3)$$

It concerns the transverse wave, described by Maxwell, for which the field pointers oscillate crosswise to the direction of propagation. The propagation again occurs with the speed of light c. So much concerning the state of the art of technology.

But as we see, in the mathematical formulation of the wave equation is hiding, yes, even more than only the generally known electromagnetic wave. The no sources approach is a neglect, which only is valid under certain prerequisites!

Scalar waves, a mathematical reasoning (II)

3. claim: Tesla radiation = longitudinal wave

$$\boxed{\text{rot } \mathbf{E} = 0} \quad \text{and} \quad \boxed{\text{grad div } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \quad (21.4)$$

special case:
source field, because div E $\neq$ 0

=> sources = charge carriers (plasma waves)
=> sources = vortex structures

---

4. approach: (div E $\neq$ 0) is a scalar! => scalar wave

=> E-field vector can be derived from a scalar potential φ:

$$\boxed{\mathbf{E} = -\text{grad } \varphi} \quad (21.5)$$

$$\text{and} \qquad \text{div } \mathbf{E} = -\text{div grad } \varphi \equiv -\Delta \varphi \qquad (21.6)$$

inserted in eq. 21.4: homogeneous scalar wave equation

$$\boxed{\Delta \varphi = \frac{1}{c^2} \cdot \frac{\delta^2 \varphi}{\delta t^2}} \quad (21.7)$$

---

5. proof: For the case of an additional space charge density ρ_{el} should be considered: div D $= \rho_{el}$ and div E $= \rho_{el}/\varepsilon$ (21.8)

inhomogeneous scalar wave equation = plasma wave!

$$\boxed{\Delta \varphi = \frac{1}{c^2} \cdot \frac{\delta^2 \varphi}{\delta t^2} - \frac{\rho_{el}}{\varepsilon}} \quad (21.9)$$

Fig. 21.4: Derivation of the plasma wave as an example of a scalar wave

21.4 Derivation of the scalar wave equation

Making neglects by all means is normal and legal in science. But it may not be carried out at will and untested. In any case an error consideration is necessary, the result of which should be, that the neglect indeed is sufficiently small.
In the here presented case of the wave equation I haven't found one single textbook, in which this error consideration has been carried out. As a result of this inexcusable negligence there is the danger that exactly the aspect is neglected, which it actually concerns. This could lead to catastrophical results, e.g. that the causes for electrosmog, for numerous EMC-problems, for biological and technical effects aren't seen and understood anymore, that pure science once more dilapidates to a creed!

In the case of the wave equation the assumption of no sources describes only one side of the medal. The other side, which for the electromagnetic wave occurs as an error term, we get if we this time put the right part in equation 21.2 to zero (rot E = 0).
In this case a divergence of the field is present, which requires a source field. As sources some charge carriers, quanta or at least particle structures, e.g. vortex structures have to exist. Their propagation occurs, as we know it from the sound propagation, as shock wave in longitudinal manner. The air molecules, the quanta or particle structures thereby oscillate in the direction of propagation. Also the field pointer has a component in this direction

$$\mathrm{rot}\,\mathbf{E} = 0 \quad \text{and} \quad \boxed{\mathrm{grad}\,\mathrm{div}\,\mathbf{E} = \frac{1}{c^2}\cdot\frac{\delta^2\mathbf{E}}{\delta t^2}} \quad \text{(special case)} \qquad (21.4)$$

The occurring divergence of the field pointer (div E) is a scalar, for which reason this wave preferably is called scalar wave.

In the special case of a scalar wave (rot E = 0) the E-field vector can be derived from a scalar potential φ:

$$\boxed{\mathbf{E} = -\,\mathrm{grad}\,\varphi} \quad (21.5)$$

On the one hand this term is used in the wave equation 21.4 on the right-hand side without forming the gradient on both sides of the equation. On the other hand the divergence of the approach 21.5:

$$\mathrm{div}\,\mathbf{E} = -\,\mathrm{div}\,\mathrm{grad}\,\varphi \equiv -\Delta\varphi \qquad (21.6)$$

is applied in equation 21.4. The result is the homogeneous scalar wave equation:

$$\boxed{\Delta\varphi = \frac{1}{c^2}\cdot\frac{\delta^2\varphi}{\delta t^2}} \quad (21.7)$$

Scalar waves, a mathematical reasoning (III)

Proof: plasma wave =
inhomogeneous scalar wave equation

$$\Delta \varphi = \frac{1}{c^2} \cdot \frac{\delta^2 \varphi}{\delta t^2} - \frac{\rho_{el}}{\varepsilon} \quad (21.9)$$

one solution<i>: $\omega^2 = c^2 \cdot k^2 + \omega_p^2$ (= Langmuir waves).

Fig. 21.5 A: Derivation of the plasma wave as an example of the existence of scalar waves in the wave equation

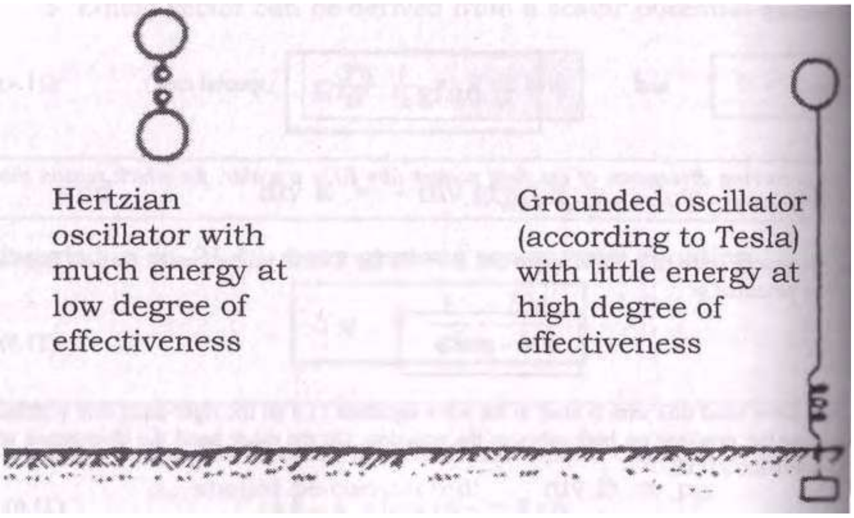
Hertzian oscillator with much energy at low degree of effectiveness
Grounded oscillator (according to Tesla) with little energy at high degree of effectiveness

Fig. 21.5 B: Nikola Tesla explains the difference between his radiation and the Hertzian wave.<ii>

<i>: The solution describes dispersion relations of plasma waves; longitudinal wave movements + Langmuir oscillations of the electron density.

<ii>: Nikola Tesla: The Problem of Increasing Human Energy, The Century Monthly Magazine, June 1900, ISBN 1-882137-00-0, Page i-15

21.5 Derivation of a plasma wave

Doing without formation of the gradient for the derivation of the homogeneous wave equation is tantamount to an integration of the equation. We hence under certain conditions must expect the occurring of an integration constant.
This is the case, if in addition a space charge density ρ_{el} occurs as source of the field, which according to Maxwell equation 4 can be considered as the divergence of a dielectric displacement D (fig. 21.4): $\operatorname{div} \mathbf{D} = \rho_{el}$,
resp. with the relation of material $\mathbf{D} = \varepsilon \cdot \mathbf{E}$: $\operatorname{div} \mathbf{E} = \rho_{el}/\varepsilon \equiv -\Delta\varphi.$ (21.8)

If we complete this contribution with possible present field sources, then the inhomogeneous scalar wave equation results:

$$\boxed{\Delta\varphi = \frac{1}{c^2}\cdot\frac{\delta^2\varphi}{\delta t^2} - \frac{\rho_{el}}{\varepsilon}} \quad (21.9)$$

For these equations solutions have been published[i]. They have the same form, as the well-known dispersion relations of Langmuir waves. That is electron plasma waves, thus longitudinal wave movements associated with Langmuir oscillations of the electron density.

With that it has been proven that scalar waves and longitudinally propagating standing waves are described by the wave equation and are contained in it. This in any case is valid in general just as in the special case of a plasma wave, as mathematically could be derived here.

From the example of the derivation of plasma waves from the wave equation 21.1, we see that scalar waves by all means are known and their existence isn't casted doubt on at all. After all the mathematically won solution is secured by numerous experiments. Why do textbooks concerning high-frequency engineering then ignore the scalar wave parts in the wave equation?
Our specialists seem to concentrate so much on their branch, that they are losing the view on the Big Whole. They practice one-eyed physics, where the plasma physicist keeps one eye shut and the radio technician the other eye. What the other one does and researches, they don't understand anymore for ages. It is necessary to point them to their common root.

The perhaps most important statement of the wave equation is that every emitted wave contains both longitudinal and transverse parts! Both parts in addition occur jointly, so that for corresponding boundary conditions it can be expected that one part is transforming into the other part. The HF technician then suddenly measures less field strength and comes to the conclusion that his radio wave has been damped or partly absorbed. Doing so heat is created, he says, although the wave equation by no means contains a corresponding term for the necessary thermodynamic description. He simply hasn't understood the wave equation!
Absorption means nothing but transverse waves in the case of a disturbance rolling up to vortices, to become a scalar wave in this way (fig. 1.4 and 5.3). With that they are evading every field strength measurement and what can't be measured doesn't exist in one-eyed physics! Therefore can't exist, what shouldn't exist.

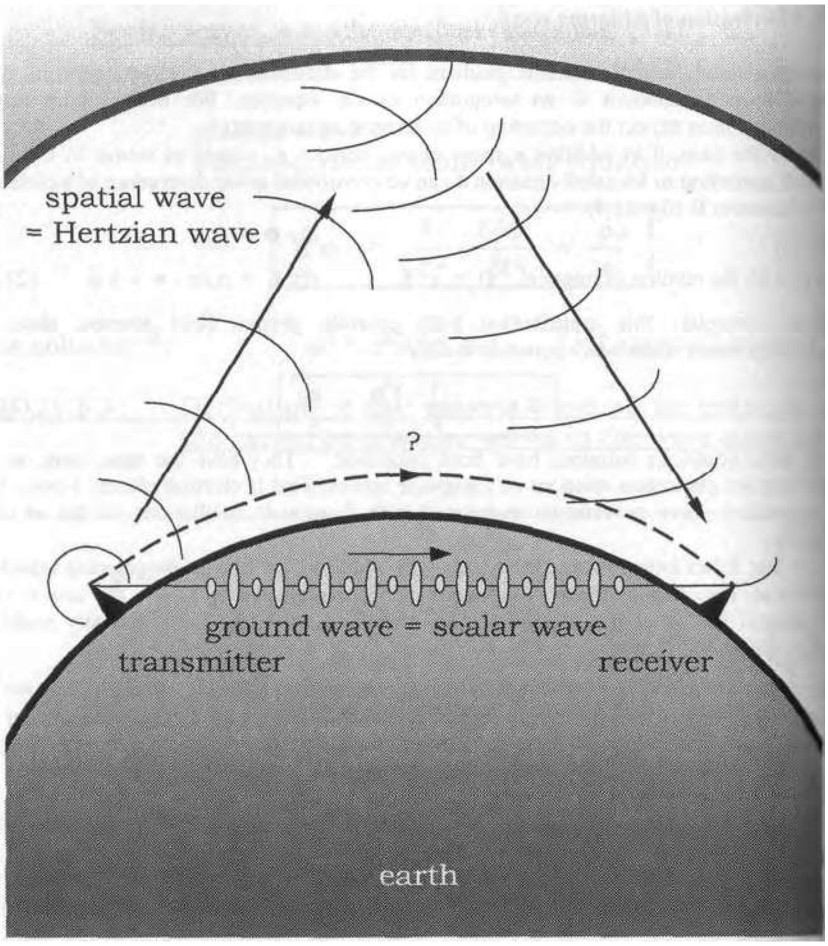
spatial wave
= Hertzian wave
?
ground wave = scalar wave
transmitter
receiver
earth

Fig. 21.6: Ground waves, which follow the curvature of the earth and radio waves reflected at the ionosphere.<i>

<i>: Meinke, Gundlach: Hochfrequenztechnik, 4.Aufl. Springer-Verlag Berlin 1986, Seite R 18: „Lang-, Mittel- and Kurzwelle breiten sich einerseits entlang der Erdoberflache als Bodenwellen, andererseits unter Mitwirkung der Ionosphare als Raumwellen aus."

21.6 Mistakes of the HF technology

The devastating judgement of Tesla, Hertz was mistaken, was rash. His claim of having detected the Maxwell wave himself, proved to be untenable. With this claim in particular Tesla hud the scientific world against him. If one opens encyclopedias or textbooks then one gets the impression, until the day of today science still hasn't forgiven Tesla, what once more shows how persistent prejudices are kept.

Just as little does a verdict of the American Supreme Court justice to the true circumstances, with the decision of the highest judges that Tesla and not Marconi is entitled the right to be the inventor of the radio. As we see, have both inventors in reality used completely other kinds of waves. The available transmitters of 100 years ago we would from today's viewpoint call ,,broadband dirt slingers". These spark flying monster have blown both parts of the wave equation, the transverse as well as the longitudinal part, into the air to ample extent. What distinguished the wave pioneers was their receiver technique, was eventually the question, which wave part they have filtered and utilized.

Marconi worked with dipole antennas, as already Heinrich Hertz. With that both could preferably track down and detect the radio wave. So they also should be entitled the right to be pioneers of radio technology. The verdict of the highest judges doesn't justice to this circumstance and should rather be judged as a nationalistically coloured political issue. Tesla however worked with two spherical electrodes, in which he preferably replaced one electrode by the globe, by grounding his devices. In this way he could receive the scalar wave parts. But that are not radio waves! Scalar waves have completely other properties, one even could be inclined to call them opposite properties.

To improve the degree of effectiveness of the transmission stretch one naturally was trying to also optimise the transmitting installation with regard to the respectively used wave part. Tesla optimised the scalar wave part and could record reactions of biological beings. His part represents a set of difficulties of the environmental compatibility, which should be taken serious. In the beginning the Marconists on the ships, as the radio operators were called, suffered from the so-called radio operator disease, which is unsolved until today. This phenomenon only disappeared after the radio devices on board had been optimised in favour of the used radio wave. The reached and measuring technically verifiable increase of the degree of effectiveness, primarily obtained by an improved antenna adjustment, simultaneously means a reduction of the scalar wave part, which endangers health.

But a received signal hides the receiver technician, if it has been on the way as a transverse or as a longitudinal wave. The coupling in one and the same equation leaves open both possibilities. Every radio amateur knows the so-called ground waves, which arrive faster at the receiver, than the Hertzian waves mirrored at the ionosphere, which propagate in a straight line. Allegedly the ground waves follow the curvature of the earth<1>, so it is written in expert books. This explanation hurts, since who can see along the curvature of the earth with a pair of field glasses. He would see the back of his head! No, the explanation the ground waves would run along the earth's surface is pure nonsense. The interference and the fading with which the radio amateur is fighting, are a result of the differently fast arriving wave parts, and doing so the scalar wave part tunnels as a straight line right through the earth (fig. 21.6)!

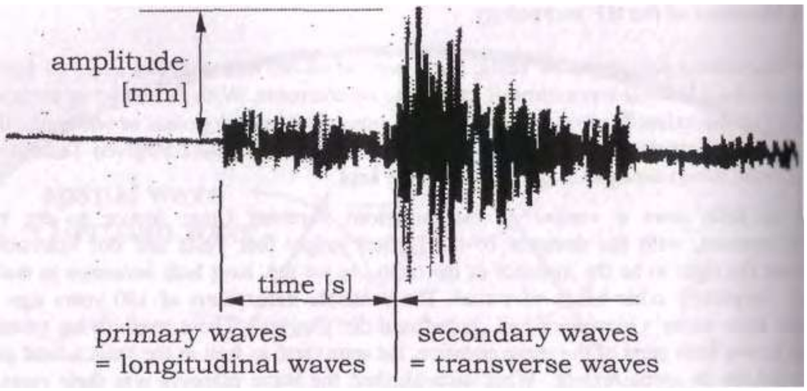

Fig. 21.7 A: Longitudinal and transverse earth quake waves.

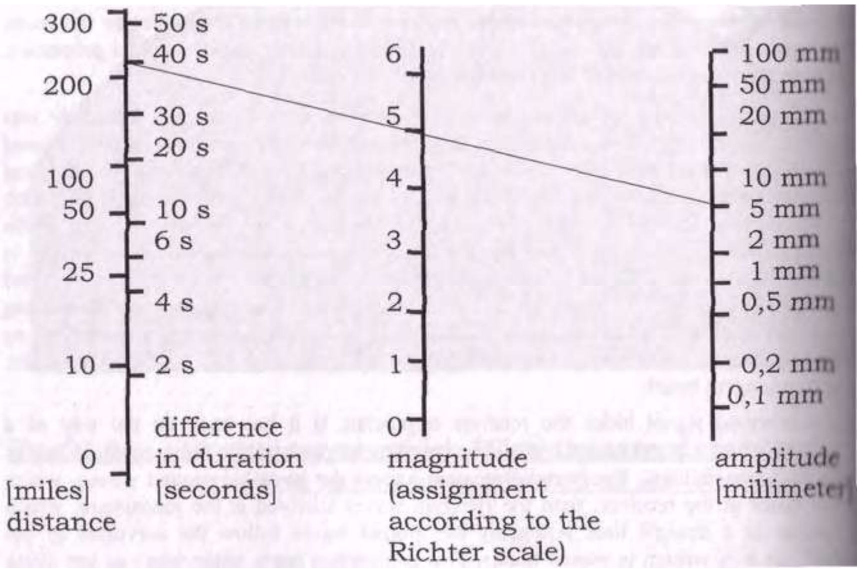

Fig. 21.7 B: Analysis according to the Richter scale.
(e.g.: 40 s duration between S- and P-waves for 5 mm amplitude means an earthquake of strength 5 in a distance of 220 miles.)

21.7 Coupling of the wave parts

The set of difficulties of ground waves makes clear the coupling of longitudinal and transverse waves as two aspects or parts of a wave. As the corresponding equation 21,1 mathematically taken apart into 21.2 dictates, does every transmitter emit both parts.

Exactly this circumstance the owners of allotments have used, which directly near a transmitter had illuminated their estate with freely hung up fluorescent lamps. The transmitter runners after that wanted to present them the power bill.and they could obtain, that this kind of illumination technique was prohibited.
Nowadays anyone may operate a high-frequency technical installation, even if he hasn't understood at all the wave equation used by that. Actually one should have been grateful to the allotment owners, if they withdraw the scalar wave radiation, which is incompatible with the environment and biologically harmful, with their fluorescent lamps. Taken exact it even is the task of the transmitting technician to pay attention to it. that only radio waves are sent into the air, since only those should be used. The time has come to reverse the burden of proof to protect the environment, nature, the consumer and the unasked and not involved persons, which are irradiated.

From other areas, for instance from flow dynamics or for body sound is generally known that both wave parts exist and in addition occur jointly. In the case of a propagation through the earth, like for an earthquake, both parts are received and utilized. Because their propagation is differently fast, the faster oscillations arrive first and that are the longitudinal ones. From the time delay with which the transverse waves arrive at the measurement station, the distance to the epicentre of the quake is determined by means of the different velocity of propagation. For geophysicists this tool is part of everyday knowledge (fig. 21.7).

Only who keeps one eye shut, could mean that the electromagnetic wave is purely transverse and sound purely longitudinal. It is true that a transverse sound wave doesn't get too far in air, for which reason sound as a rule is considered as a purely longitudinal wave by neglecting this part, but such a neglect may not be carried out in general, it must be checked if it is legitimate from case to case and an error consideration should be carried out.

Further examples for the coupling of the wave parts are furnished by the latest tunnel experiments. Here so-called pure transverse waves are sent into a tunnel, through which they don't fit through at all. The Maxwell-theory then dictates that behind the tunnel no signal should be measurable.
But a signal is being measured, which in the tunnel in addition was faster than allowed.

In conferences again is being discussed about the wave equation. The imagination of the specialists reaches from phase velocities of an electromagnetic wave, which isn't present at all up to instantaneous tunnelling, during which the clocks should stop<i>.
The wave equation however supplies the only possible answer: The tunnel filters out the scalar wave parts and lets pass from them only those, which are sufficiently small and correspondingly fast!

<i>: see Part 1, Chapter 6.14

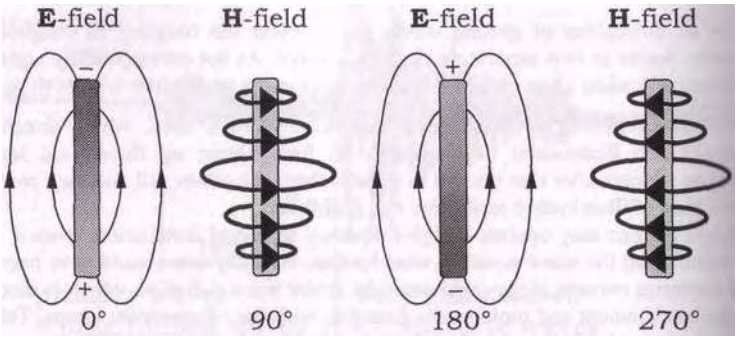

Fig. 21.8 A: The fields of the oscillating dipole antenna

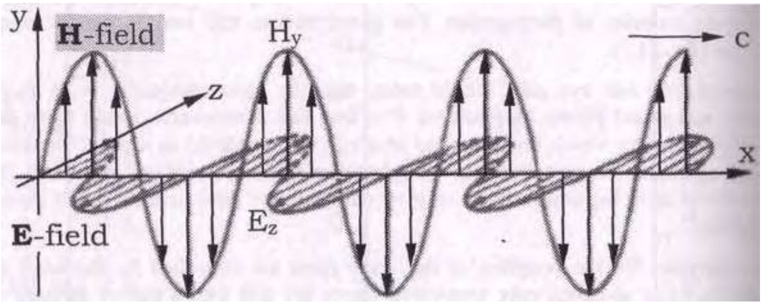

Fig. 21.8 B: The planar electromagnetic wave in the proximity

---

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3. Aufl. Springer-Verlag Berlin 1986, Seite 335

<ii>: dpa-message in the Sudkurier of 25.11.2000: „Gefahr furs Herz" (Translated:) Patients with a cardiac pacemaker produced in the Netherlands have been warned by the producer of the device (Vitatron) to be careful when passing anti-theft installations in stores. For devices, which were implanted between 1992 and 1998, there is the danger of the implant failing.

21.8 Set of difficulties of the near-field

In high-frequency technology is distinguished between the near-field and the far-field. Both have fundamentally other properties.

Heinrich Hertz did experiment in the short wave range at wavelengths of some meters. From today's viewpoint his work would rather be assigned the far-field. As a professor in Karlsruhe he had shown that his, the electromagnetic, wave propagates like a light wave and can be refracted and reflected in the same way. It is a transverse wave for which the field pointers of the electric and the magnetic field oscillate perpendicular to each other and both again perpendicular to the direction of propagation. It hence would be obvious, if in the case of the Hertzian wave it would concern the far-field. Besides the propagation with the speed of light also is characteristic that there occurs no phase shift between E-field and H-field.

In the proximity it looks completely different. The proximity concerns distances to the transmitter of less than the wavelength divided by 2*pi. Nikola Tesla has broadcasted in the range of long waves, around 100 Kilohertz, in which case the wavelength already is several metres. For the experiments concerning the resonance of the earth he has operated his transmitter in Colorado Springs at frequencies down to 6 Hertz. Doing so the whole earth moves into the proximity of his transmitter. We probably have to proceed from assumption that the Tesla radiation primarily concerns the proximity, which also is called the radiant range of the transmitting antenna.
For the approach of vortical and closed-loop field structures derivations for the near-field are known<i>. Doing so it must be emphasized that the structures don't follow from the field equations according to Maxwell, but the calculations are based on assumed rotation symmetrical structures. The Maxwell theory by no means is capable of such a structure shaping by principle. The calculation provides as an important result<i> that in the proximity of the emitting antenna a phase shift exists between the pointers of the E- and the H-field. The antenna current and the H-field coupled with it lag the E-field of the oscillating dipole charges for 90° (fig. 21.8). These charges form a longitudinal standing wave the antenna rod or antenna dipole. For this reason also the fields produced by high-frequency currents at first have the properties of a longitudinal wave in the proximity of the antenna.

The near-field already is used in practice in anti-theft devices, as they are installed in the entrance area of stores. The customer walks through the scalar wave transmitters. If the coupling coil has not been removed at the cash point, then a signal from the alarm system sounds. The coils work purely passive, i.e. they are supplied with electric energy per scalar wave and stimulated 'to oscillate for their part. Then the effect back on the transmitter is being utilized. Even if the principle is functioning, people still should be warned not to use a technology, which has not been understood completely. Then not explained catastrophes are inevitable<ii>.

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3.Aufl. Springer-Verlag Berlin 1986, Seite 335

<ii>: dpa-message in the Sudkurier of 25.11.2000: ,,Gefahr furs Herz". (quoted at the left, fig. 21.8)

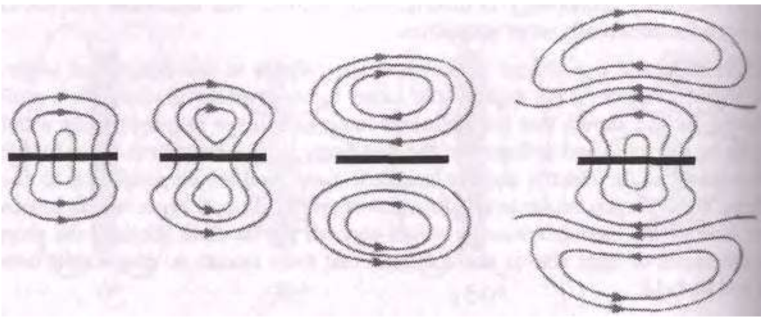

Fig. 21.9 A: The coming off of the electric field lines from a dipole
The forming vortex structures found a longitudinal electric wave carrying impulse!

---

electromagnetic wave (transverse)

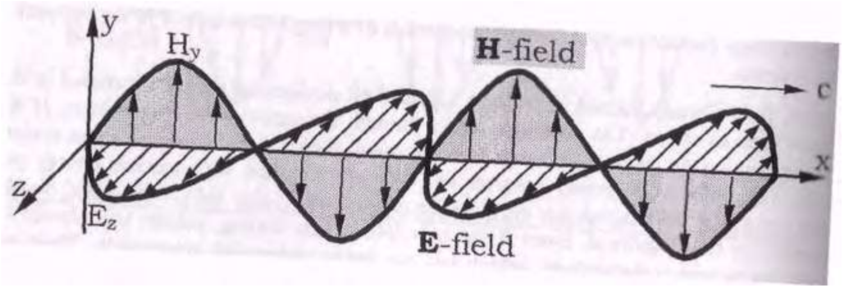

Fig. 21.9 B: The planar electromagnetic wave in the far zone

21.9 Transition to the far-field

In sufficient distance to the transmitting antenna as far-field the transverse electromagnetic wave results (fig. 21.9 B). It is distinguished by not occurring a phase shift between E- and H-field anymore. Every change of the electric alternating field is followed immediately and at the same time by a change of the magnetic alternating field and vice versa.

In the proximity however the phase shift amounts to 90°. Somewhere and somehow between the causing antenna current and the far-field a conversion from a longitudinal into a transverse wave occurs. How should one imagine the transition?

In the books the coming off of a wave from a dipole is represented according to fig. 21.9 A. The fields come off the antenna, the explanation reads. If we consider the structure of the fields coming off then we see field vortices, which run around a point, which we can call the vortex centre. Such field structures naturally are capable of forming standing waves and to carry an impulse. The scalar wave field in general and the near-field in special we only will understand with suitable vortex physics and with a field theory extended for corresponding vortices we also will be able to calculate it. Postulates cannot replace field physics!

Be that as it may, the vortex, after having left the antenna, for bigger getting distance at some time seems to unroll to propagate further as an electromagnetic wave. There takes place a transition from longitudinal to transverse, or spoken figuratively, from vortex to wave. How complete this conversion takes place, how big the respective wave parts are afterwards, on the one hand depends on the structure and the dimensions of the antenna. Information is given by the measurable degree of effectiveness of the antenna.

The vortex structures on the other hand are the stabler, the smaller and faster they are. If they are as fast as the light or even faster, then they become stable elementary particles, for instance neutrinos. Slower vortex structures however are predominantly instable. They preferably unwind to waves. Vortex and wave prove to be two possible and under certain conditions even stable field configurations.

Let's emphasize: A Hertzian dipole doesn't emit Hertzian waves! An antenna as near-field without exception emits vortices, which only at the transition to the far-field unwind to electromagnetic waves. A Hertzian wave just as little can be received with a dipole antenna! At the receiver the conditions are reversed. Here the wave is rolling up to a vortex, which usually is called and conceived as a „standing wave". Only this field vortex causes an antenna current in the rod, which the receiver afterwards amplifies and utilizes.

The mostly unknown or not understood near-field properties prove to be the key to the understanding of the wave equation and of the method of functioning of transmitting and receiving antenna. The question is asked, how one should imagine the rolling up of waves to vortices and vice versa the unrolling? How could an useful vortex mode) look like?

Circularly polarized wave (transverse)

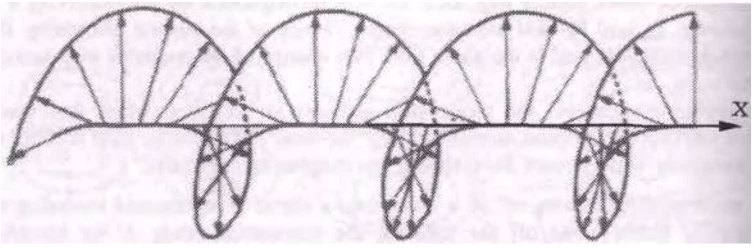

Fig. 21.10 A: Left-circular polarized wave
(as explanation for the transition to a vortex and to a scalar wave)

electric wave (longitudinal)

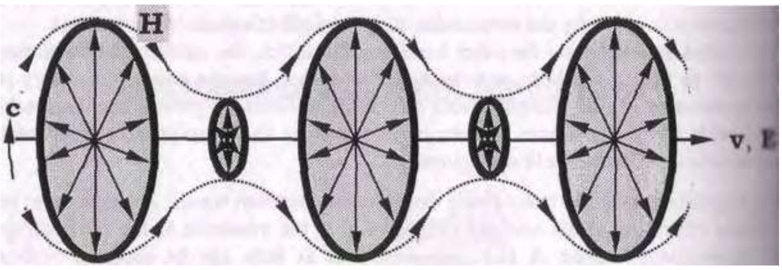

Fig. 21.10 B: Magnetic ring-vortices form an electric scalar wave.

vortex and wave = two stable field configurations

electromagnetic wave = transverse wave propagating in a straight line

ring-like vortex = transverse wave running in circles

vortex velocity = speed of light c

change of structure = if the field is disturbed without expense of energy

22.10 Scalar wave model

The light, as electromagnetic wave, in the presence of a heavy mass or of strong fields is bent towards the field source (fig. 6.10). The wave normally propagating in a straight line thus can be diverted. The square of the speed of light further is inversely proportional to the permeability and dielectricity, short, in presence of matter it is more or less strongly slowed down. If this slowing down of the wave occurs one-sidedly, then a bending of the path can be expected as well. At the end of the antenna a reflection and a going back of the wave can occur, which at the other end again hits itself. Now the wave has found a closed-loop structure, which can be called vortex. The figures 21.10 B and 21.11 A show the two possible structures.

In technical books this vortex with the properties of a „standing wave" is explained gistly. Near-field and standing wave are two examples, how the textbooks succeed in describing mathematically right a small part of the scalar wave properties, without having to have a good look at vortex physics. With such auxiliary descriptions however the end is reached fast, if for instance it concerns understanding a pure scalar wave transmission according to Nikola Tesla (fig. 19.11) and the special properties of this wave type. With the vortex concept of an extended field physics new horizons are opening.

If we direct our look again to the sketches (fig. 21.10 B and 21.11 A). In both cases electromagnetic waves are represented which propagate with the speed of light, only that the wave doesn't go forward in a straight line but instead runs around in circles. It also furthermore is transverse, because the field pointers of the E-field and the H-field oscillate perpendicular to c. By means of the orbit the speed of light c now has become the vortex velocity. Wave and vortex turn out to be two possible and stable field configurations. For the transition from one into the other no energy is used; it only is a question of structure. The vortex structure thus stabilizes itself by means of the field dependency of the speed of light.

By the circumstance that the vortex direction of the ring-like vortex is determined and the field pointers further are standing perpendicular to it, as well as perpendicular to each other, there result two theoretical formation forms for the scalar wave. In the first case (fig 21.10 B) the vector of the H-field points into the direction of the vortex centre and that of the E-field axially to the outside. The vortex however will propagate in this direction in space and appear as a scalar wave, so that the propagation of the wave takes place in the direction of the electric field. I call this an electric wave.

In the second case the field vectors exchange their place. The direction of propagation this time coincides with the oscillating magnetic field pointer (fig. 21.11 A), for which reason I speak of a magnetic wave.

The vortex picture of the rolled up wave already fits very well, because the propagation of a wave direction of its field pointer characterizes a longitudinal wave, because all measurement results are perfectly covered by the vortex model. It even is clear that no energy has to be spended for the conversion, since merely the structure has changed. If it becomes a vortex the wave just doesn't run in a straight line anymore but in circles, to either wrap around the magnetic field vector (fig. 21.10 B). or the electric field vector (fig. 21.11 A).

magnetic wave (longitudinal)

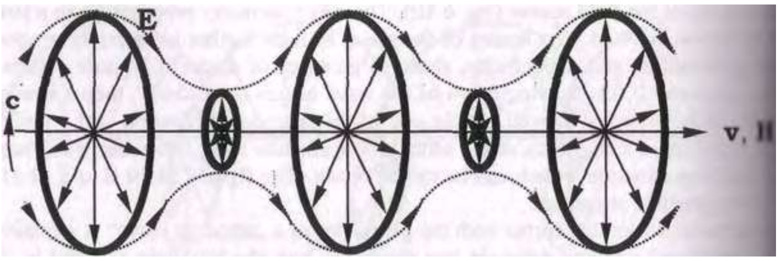

Fig. 21.11 A: The magnetic scalar wave

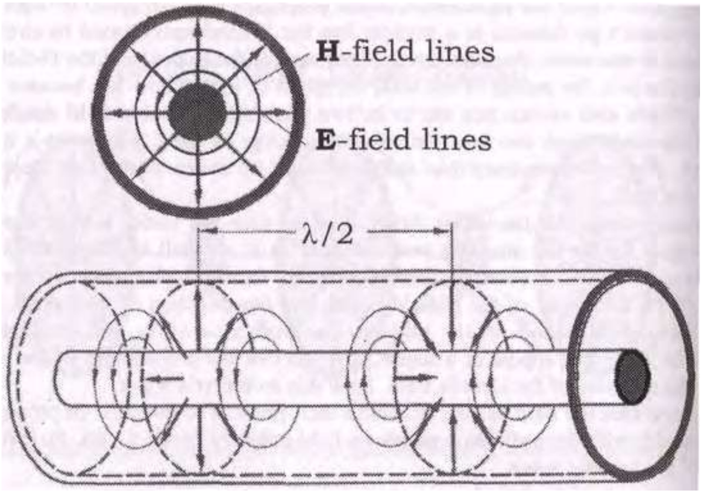

Figure 21.11 B: Wave propagation in a coaxial cable, (Example for waveguide, horn radiator, etc.)<i> Cross-section of coaxial conductor and field distribution in the direction of propagation.

<i>: H. Armbruster, G. Grunberger: Elektromagnetische Wellen im Hochfrequenz-

21.11 Double-frequent oscillation of size

Because a longitudinal wave propagates in the direction of the field, the field pointer also will oscillate with the velocity of propagation v. This hence isn't constant at all, it can significantly differ from that of the light and can take arbitrary values. According to the theory of objectivity the field oscillating with it determines its momentary size:

$$E, H \sim 1/v^2 \qquad (21.10).$$

The velocity of propagation v of the scalar wave thus oscillates double-frequently and with opposite phase to the corresponding field. A detailed description would mean, if the field strives for its maximum value, the velocity v of the wave reaches its smallest value. In the field minimum the scalar wave vice versa accelerates to its maximum value. For longitudinal waves therefore only an averaged velocity of propagation is given and measured, as this for instance is usual for the sound wave, and this can vary very strong as is well-known (body sound compared to air sound, etc.).
The two dual field vectors of E and H, the one in the direction of propagation and the one standing perpendicular to it, occur jointly. Both oscillate with the same frequency and both form the ring-like vortex in the respective direction. As a result the ring-like vortex also oscillates in its diameter double-frequently and with opposite phase to the corresponding field (fig. 21.10 Band 21.11 A).
This circumstance owes the ring-like vortex its property, to tunnel. No Faraday cage is able to stop it, as could be demonstrated in experiments<i>. Only therefore the ground wave runs through the earth and not along the curvature of the earth. A further example is the coaxial cable (fig. 21.11 B). Also this acts as a long tunnel and so it isn't further astonishing, that the electric field lines have the same orientation, as for a magnetic scalar wave. As a practical consequence in this place there should be warned of open cable ends, wave guides or horn radiators with regard to uncontrolled emitted scalar waves!

At present in the press is being discussed, if the cable network runners for some channels the mission to operate again should be withdrawn, because the airline radio traffic is being disturbed. The original opening for cable frequencies, which actually are reserved for the airline radio traffic, based on the erroneous assumption, that conflicts are unthinkable. But then the planes were disturbed in their communication. As the cause TV-cables made out, which hadn't been closed according to the rules with a resistor, as it by all means can occur on building sites and during renovation works.
On the other hand is being argued with the small current, which flows through the coaxial cable, and the large distance to the planes also is cited. According to that it actually can't concern Hertzian waves. It presumably are scalar waves, which escape from the open cable ends and which are collected by a receiver in the plane. It indeed is very little field energy, but because it again is being collected and bundled, the scalar wave is able to exceed the part of the radio wave by far just at large distances and to cause problems. For such examples from practice the scalar wave theory is fully taking effect.

<i>: Adolf und Inge Schneider im Interview mit Prof. Dr.-Ing. Konstantin Meyl: Durchbruch in der Freien-Energie-Forschung, NET-Journal 12/1999, S. 7-9.

1. H. Hertz: electromagnetic wave (transverse)

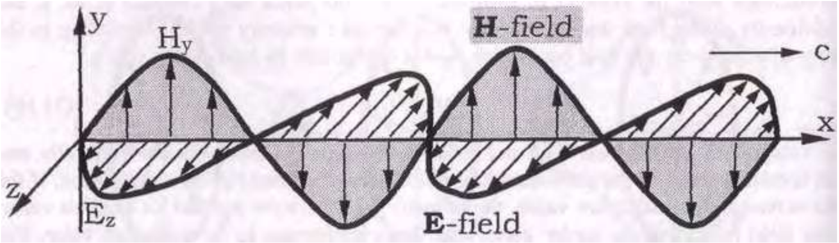
y
H_y
H-field
c
x
z
E_z
E-field

2. Nikola Tesla: electric wave (longitudinal)

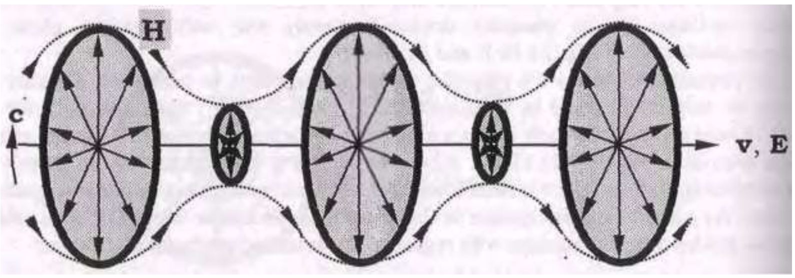

3. magnetic wave (longitudinal)

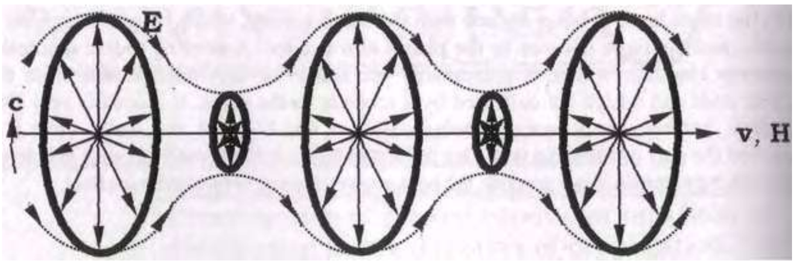

Fig. 21.12: The three basic types according to the wave equation
(21.1), (electric, magnetic and electromagnetic wave).

21.12 Electric and magnetic scalar wave

Let us record: For the wave propagation there thus are three possible and stable states (fig. 21.12): the transverse electromagnetic wave according to Heinrich Hertz (fig. 1), the longitudinal electric wave according to Nikola Tesla (fig. 2), and a longitudinal magnetic wave (fig. 3), which isn't connected yet with a name of a discoverer. The last one is a pure product of my theoretical derivation. The question is asked, which practical meaning the magnetic wave could have.

It is formed by individual electric field vortices, which I have discovered and called potential vortices 1990. I proceed from the assumption that the youngest of the three waves will play the by far biggest role in the future, because its properties are unattainable, both with regard to the energy technical and to the information technical use. One example for each should support this thesis.

The experiments concerning the electric wave according to Nikola Tesla, where is being worked with electrically charged spheres, don't show a particularly high power. Magnetic converters, so the experiences of my laboratory activities, are superior to an electrostatic converter as a collector for free energy by far. That even can be expected, because a magnetic engine is much smaller than an electrostatic engine of the same power as is well-known.

At a congress of medicines was given a talk on the basic regulation of the cells, on the communication of the cells with each other. Professor Heine in his decades of research work has found out that the cells for the purpose of communication build up channels for instance in the connective tissue, which after having conducted the information again collapse. Interestingly the channels have a hyperboloid structure, for which no conclusive explanation exists.

The structure of the data channels however is identical with the one of a magnetic scalar wave, as shown in fig. 3. Through a channel formed such, which functions like a tunnel or a dissimilarly formed waveguide, only one very particular scalar wave can run through. Waves with different frequencies or wavelengths don't fit through the hyperboloid formed tunnel at all in the first place. Through that the information transmission obtains an extremely high degree of safety for interference.

To the biologist here a completely new view at the function of a cell and the basic regulation of the whole organism is opening. The information tunnel temporarily forms more or less a vacuum, through which only potential vortices can be conducted, and that without any losses - simply perfect! From this example is becoming clear that nature is working with scalar waves namely with magnetic waves.

One other point should be recorded: The mentioned tunnel experiments, in which speed faster than light is being measured with most different devices, impressively confirm the presence of scalar waves. But if scalar waves exist which are faster than light and other ones, which are slower, then it almost is obvious that also such ones will exist, which propagate exactly with the speed of light. These then will have all the properties of the light and won't differ from the corresponding electromagnetic wave in the observable result. As scalar wave it however is formed by vortex configurations, which unambiguously have particle nature. Nothing would be more obvious than to equate these quantum structures with the photons.

Question: Is the light wave or particle?

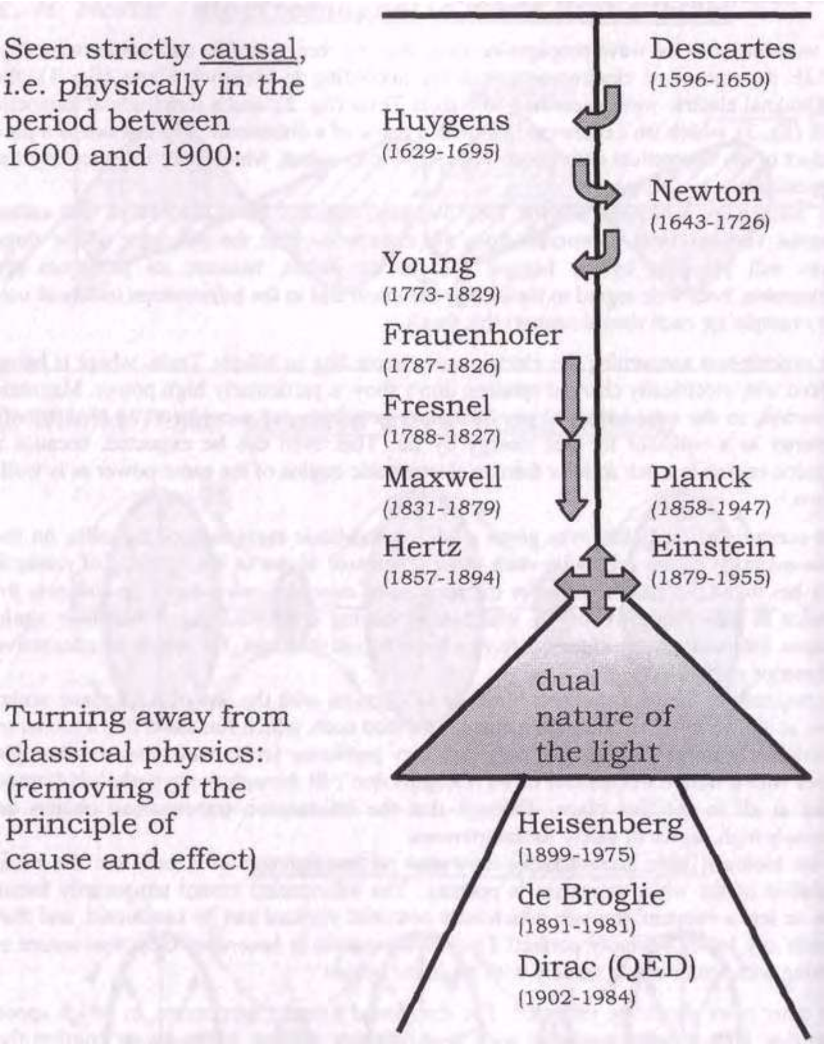
Seen strictly causal,
i.e. physically in the
period between
1600 and 1900:
Descartes
(1596-1650)
Huygens
(1629-1695)
Newton
(1643-1726)
Young
(1773-1829)
Frauenhofer
(1787-1826)
Fresnel
(1788-1827)
Maxwell
(1831-1879)
Planck
(1858-1947)
Hertz
(1857-1894)
Einstein
(1879-1955)
dual
nature of
the light
Turning away from
classical physics:
(removing of the
principle of
cause and effect)
Heisenberg
(1899-1975)
de Broglie
(1891-1981)
Dirac (QED)
(1902-1984)

Fig. 22.1: The view of some physicists concerning the nature of the light as wave or as particle.

According to the wave equation the light is a mixture of wave and photon radiation!

22 Properties of scalar waves

For the light quanta hypothesis the Nobel Prize for physics was awarded 1921. But it only was a hypothesis, an idee fixe, which was honoured here. It is quite odd, if such an important prize is awarded to a research scientist, who hasn't got the slightest idea what light quanta are anyway, of what they consist and how they are built up. Albert Einstein cleverly used the embarrassing situation, by in his celebration speech on the occasion of the awarding of the Nobel Prize giving a talk on the theory of relativity. After the speech a member of the committee found it necessary, to point to it, that this wasn't object of the awarding of the prize and that the theory of relativity concerns a pure theory, which can't be proven by principle. A theory hence neither could be awarded the Nobel Prize.

Such words on the occasion of awarding a prize for a daredevil hypothesis give the whole event really grotesque characteristics. But it came still worse.

22.1 Wave-particle compromise

Physicists of name and rank had come together. It concerned the question if the light is wave, or particle or even both at the same time? For both variants experimental proof was present, the discussion became inflamed and the things boiled over. Finally they were as smart as before, as Werner Heisenberg presented his ideas concerning the uncertainty principle. This compromise, on which one eventually came to an agreement, with good cause may be called the worst in the history of physics. It dictates me, what I shell see and how exact I may look. With it the contradiction should be overcome that the light contrary to every causality should be wave and particle at the same time.

Such fixings not only have a funny, but also a tragic side. Since it were authorities, which have approved the compromise and the whole community of science has confidence in the statements of its authorities, which immediately and unfiltered is entered in all textbooks. At the meeting it simply and solely concerned the wave equation and only that could have supplied the correct and only possible answer: It falls apart into two parts and this explains, why the light one time appears as an electromagnetic wave and the next time as a vortex particle, which is called photon. The conversion can take place at any time spontaneously and without putting on energy, so that depending on the used measuring technique the particle appears as wave or as particle, but of course never as both at the same time!

Looking back one can say that the funny thing about the situation was that all discussed about the wave and its properties known at that time, that all should know the wave equation. An equation as is well-known says more than a thousand words and one look would have sufficed entirely, to answer the controversial question once and for all. It would have saved us a lot.

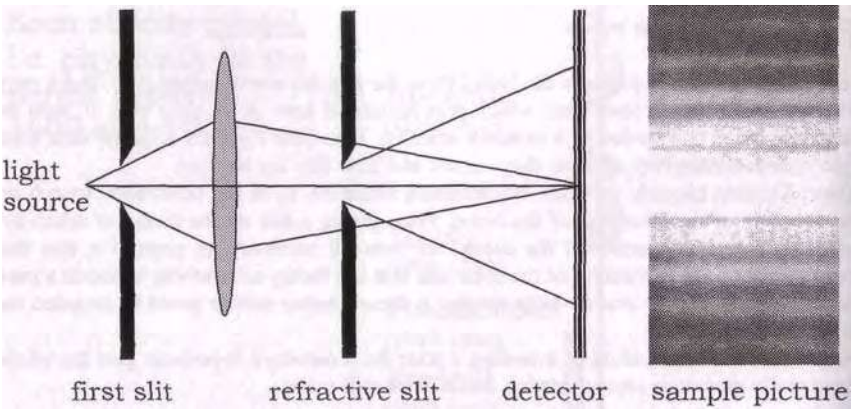
light
source
first slit
refractive slit
detector
sample picture

Fig. 22.2 A: Light forms interference patterns at the slit (light stripes are formed, where the waves oscillate in phase, dark stripes, where they oscillate out of phase).

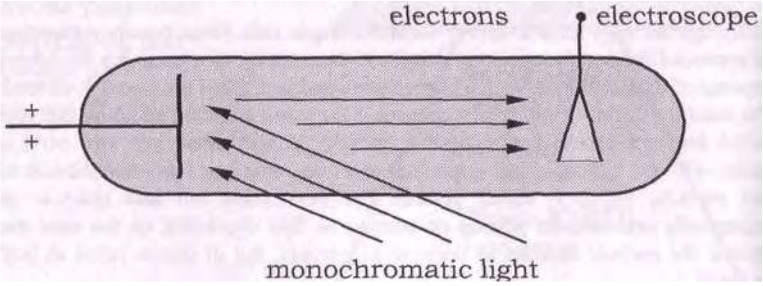
electrons
electroscope
+
+
monochromatic light

Fig. 22.2 B: The photo-electric effect

<i>: Atome als Energiewellen. Physiker wandelten einen Strom von Natrium atomen in Wellen um. III. Wissenschaft 7/1996, Seite 56 + 57

22.2 Concerning the measurement of light

The uncertainty principle with the interpretation of Heisenberg, the light is wave and particle at the same time, is incompatible with the wave equation. Heisenberg puts an equal sign, where in the wave equation in reality is present an addition of both wave parts. Fortunately in mathematics there is no need of speculating, there a derivation is right or wrong. Nothing is changed to that even if all physicists of the world should go in the wrong direction following the prevailing opinion. The wave equation exert an influence on the interpretation of the light experiments, on the one hand the ones concerning the interference and refraction of the light, where electromagnetic waves are becoming visible (fig. 22.2 A) and on the other hand the photo-electric effect, as proof of light quanta (fig. 22.2 B).

Already the wave theory of Huygens requires interference patterns of light rays, as they for instance are observed behind a slit, and demonstrates with that the wave nature. If on that occasion the particle nature is lost, if thus the photons present before the slit can't be detected behind the slit anymore, then plain and simple the measuring method, thus the slit is to blame for that. The vortices have unrolled themselves at the slit to waves. Corresponding experiments also have been carried out with matter. At the Massachusetts Institute of Technology whole sodium atoms were converted into waves. At the detector pure interference patterns were observed, which go as evidence for the successful dematerialization[i]. But the vortex physicist they show still more: they reveal, that atoms merely are waves rolled up to spherical vortices, which at any time and spontaneously can again unroll to waves at a lattice (chapter 5 and 7).
The common interpretation, the wave nature detectable behind a slit must have been present in the same form already before the slit, is untenable and in the end wrong, as makes clear the experiment with the sodium atoms.

The photo-electric effect, which on the other hand shows the quantum nature of the light, has been discovered by Heinrich Hertz, further investigated by Lenard and finally rendered more precisely by Millikan 1916 (fig. 22.2 B). It bases on the circumstance that light of higher frequency, thus blue light, has more energy than red light of lower frequency. But if electrons are knocked out a metal plate by light, then that occurs, by the waves rolling up to vortices. Now indeed photons are at work, which are detected with an electroscope indirectly.
In the same way a photon ray in a bubble chamber can be photographed. But also here the measuring method is responsible for what is being observed.

A good example is the human eye, the rods and cones of which merely can pick up potential vortices and pass them on to the nerves as so-called reaction potentials. Incident waves can only be detected, if they first have rolled up to vortices in the corpus vitreum of the eye. For us seeing, it doesn't play a role of how many percent vortices and waves the light is consisting.
Behind a sheet of glass for instance a larger vortex part can be expected and still the light has the same brightness as without sheet; the sheet of glass is perceived as transparent. We nevertheless must assume that light with a large wave part has another quality, than such light behind glass or artificial light with a large part of photons.

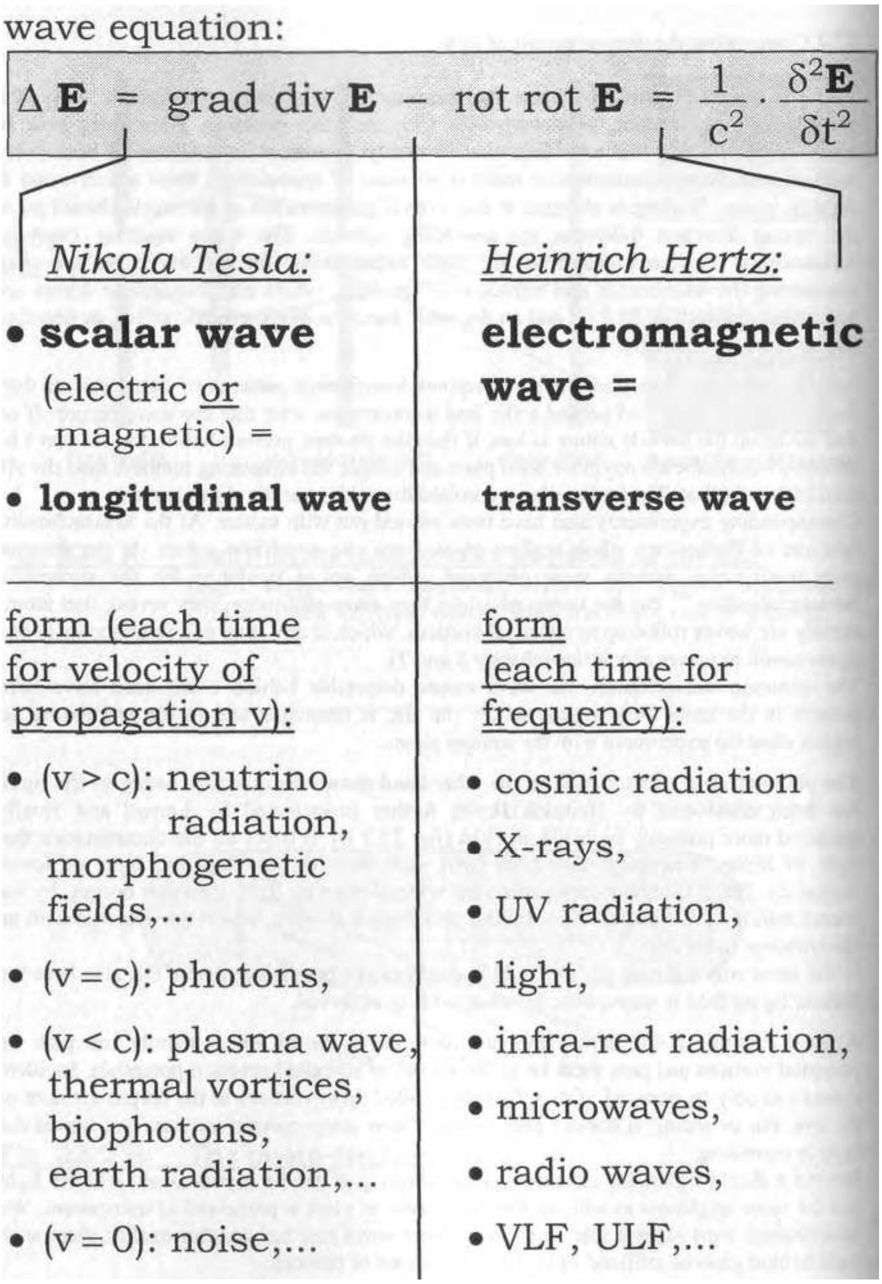

wave equation:

$$\Delta \mathbf{E} = \text{grad div } \mathbf{E} - \text{rot rot } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}$$

Nikola Tesla:	*Heinrich Hertz:*
• **scalar wave** (electric or magnetic) =	**electromagnetic wave =**
• **longitudinal wave**	**transverse wave**
form (each time for velocity of propagation v):	form (each time for frequency):
• (v > c): neutrino radiation, morphogenetic fields,...	• cosmic radiation • X-rays, • UV radiation,
• (v = c): photons,	• light,
• (v < c): plasma wave, thermal vortices, biophotons, earth radiation,...	• infra-red radiation, • microwaves, • radio waves,
• (v = 0): noise,...	• VLF, ULF,...

Fig. 22.3:____ The two parts of the wave equation

22.3 Comparison of the parts of Tesla and Hertz

Light as a rule always is formed as photon radiation, even on the sun. If in the end only waves arrive on earth, then the vortices sometime on the way to us must have unrolled to waves. Photon radiation after all is a scalar wave radiation, which generally is predominant in the near-field of the source of radiation. There is no reason, why the light should act in another way than the wave radiated by a radio transmitter, which as well forms vortices in the near-field area, as we already have discussed. For different interpretations of wave properties of one and the same physical phenomenon there is no place in a unified theory.

If we stay at the comparison then it is not an individual case that an experimental setup is responsible for what is being measured and observed. A parallel case to the experiments concerning the nature of the light is the one concerning the wave propagation. Hertz has received and utilized the transverse part and Tesla the longitudinal part and either one claimed, only he is right. There doesn' t exist an other equation, which has been and is being ignored and misunderstood so thoroughly, as the wave equation.

Fig. 22.3 shows in a survey the two parts of the wave equation in the assignment to the terms and forms: Right hand side the electromagnetic wave according to Heinrich Hertz and left hand side the scalar wave according to Nikola Tesla. The terms, like on the one hand transverse wave and on the other hand longitudinal wave relate to the kind of wave propagation.

If the field pointers oscillate crossways to the direction of propagation, then as a consequence the velocity of propagation is decoupled from the oscillating fields. The result in all cases is the speed of light, and that in our observation is constant. It is usual to make a list for increasing frequency, starting at the longest waves (ELF and VLF) over the radio waves (LW, MW, SW, UHF), the TV channels (VHF, UHF), the microwaves, the infra-red radiation, the light, the X-rays up to the cosmic radiation.
It really is interesting that it concerns one and the same phenomenon despite the different forms! As long as Maxwell only had published a theory for the light, in the world of science 24 years long at first nothing at all happened. Only Heinrich Hertz with his short wave experiments opened the eyes. Now all suddenly started at the same time to research into various phenomena on the frequency scale, from Madame Curie over Konrad Rontgen up to Nikola Tesla, who primarily researched the area of long waves.

With regard to the scalar waves until now a corresponding booster detonation has failed to appear. The immense area is new ground scientifically, which is awaiting to be explored systematically. I try to make a contribution with my historic rebuild of a scalar wave transmission line according to the plans of Tesla<i>.

<i>: Im Gesprach mit dem Fernsehmoderator und Buchautor Johannes von Buttlar weise ich auf die Chancen und technischen Moglichkeiten hin: Johannes von Buttlar im Gesprach mit Prof. Dr. Konstantin Meyl: Neutrinopower, Argo-Verlag Marktoberdorf, 1. Aufl. (2000).

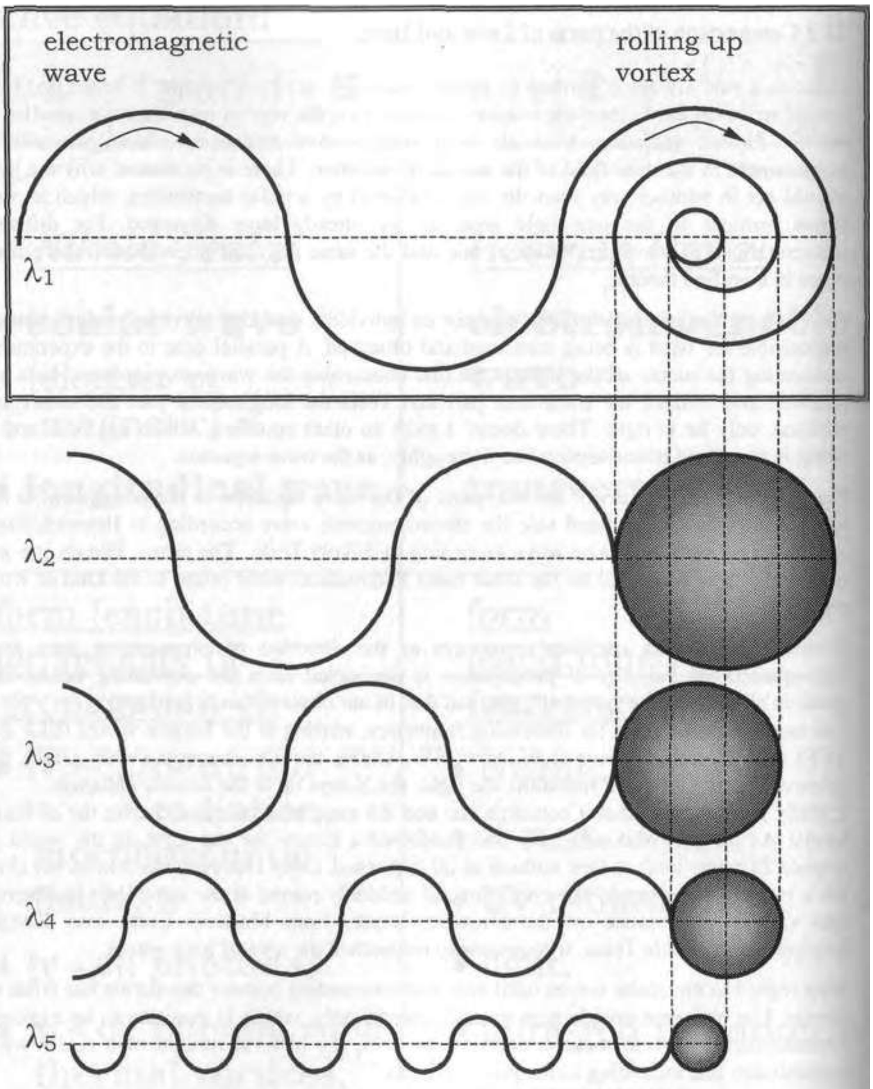
electromagnetic wave
rolling up vortex
λ_1
λ_2
λ_3
λ_4
λ_5

Fig. 22.4: The wave rolling up to a vortex.

wavelength of the wave: $\lambda_1 = c/f_1$

wavelength of the vortex: λ_n (n = 1,2,3,4,5) $\leq \lambda_1$

$f_5 > f_4 > f_3 > f_2 > f_1$ $\lambda_5 < \lambda_4 < \lambda_3 < \lambda_2 < \lambda_1$

22.4 Noise, a scalar wave phenomenon

Longitudinal waves can take arbitrary velocities between zero and infinity, because they propagate in the direction of an oscillating field pointer and as a consequence of that their velocity of propagation oscillates as well and by no means is constant. It does make sense to llist the forms of scalar waves according to their respective velocity of propagation (fig. 22.3, left column).

If we start with a localized vortex, a wave rolling up, which further is contracting. Doing so the wavelength gets smaller and smaller, whereas the frequency increases. An evenly frequency mixture distributed over a broad frequency band is observed. This phenomenon is called noise (fig. 1.4). But besides the localized noise, noise vortices can also be on the way with a certain velocity as a scalar wave, e.g. for the radio noise. In this case they show the typical properties of a standing wave with nodes and antinodes.

Also the earth radiation is said to have standing wave nature, which can be interpreted as slowed down neutrino radiation. If it is slowed down on the way through the earth, then the neutrino properties are changing, as this was measured in the Kamiokande detector in Japan recently. Unfortunately the proof occurs only indirect, because there still don't exist measuring devices for scalar waves. We'll talk about this problem area later and are content with the clue that already within living memory the standing wave property has been used to find water and deposits of ores and still is used today (fig. 22.4).

If we continue our considerations concerning the forms of scalar waves, as they are listed in fig. 22.3. The scalar waves, which are slower on the way than the light, are joined by the plasma waves. This is confirmed by measurements and calculations.

For thermal vortices, as they have been investigated by Max Planck and for biophotons, as they can be detected in living cells by colleague Popp, the velocity of propagation however is unknown. It was not and still is not measured at all, now more than ever. The research scientists have confidence in the assumption that all waves go with the speed of light, but that is a big mistake.

For all wave kinds there at least exists also one vortex variant, for radio waves for instance it is the radio noise, which propagates with a velocity different from c. The velocity is the product of frequency and wavelength:

$$\boxed{v = f \cdot \lambda} \quad (22.1)$$

From the three variables v, f and λ at least two must be measured, if one has a suspicion that it could concern scalar waves. At this place most errors are made in the laboratories. Countless experiments concerning the biological compatibility, concerning medical therapy methods and similar experiments must be repeated, because as a rule only the frequency is being measured and it has been omitted to at least check the wavelength or the velocity of the wave. Countless research scientists must put up with this accusation. Much too blind the scientists, who now again may start from the very beginning with their work, have had confidence in the predominance of the speed of light.

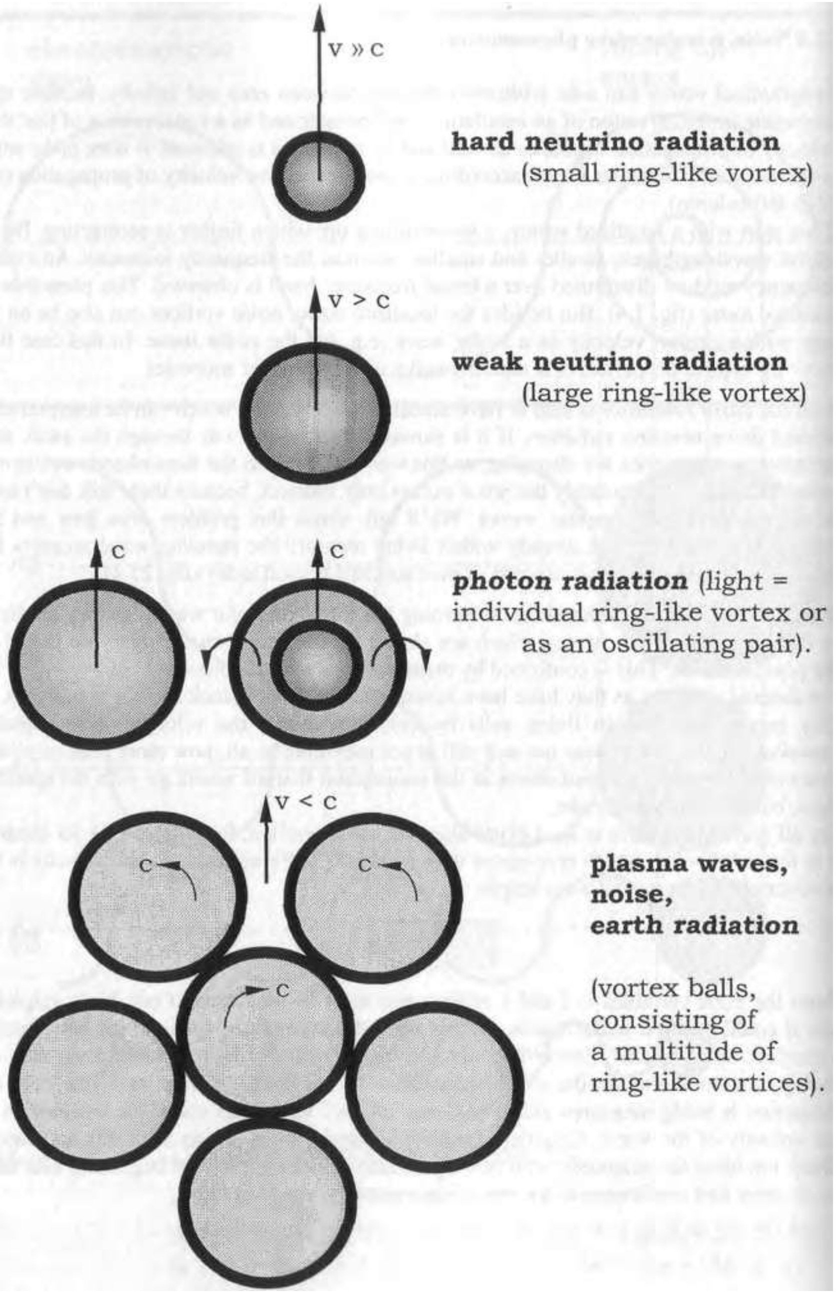
v » c
hard neutrino radiation
(small ring-like vortex)
v > c
weak neutrino radiation
(large ring-like vortex)
c
c
photon radiation (light = individual ring-like vortex or as an oscillating pair).
v < c
c
c
c
plasma waves, noise, earth radiation
(vortex balls, consisting of a multitude of ring-like vortices).

Fig. 22.5: ____The ring-like vortex model of scalar waves.

22.5 Neutrino radiation

The neutrino physicists make the same error. They proceed from the assumption that their particles are on the way with a speed somewhat less than the speed of light c. This contradicts the observation according to which black holes should represent strong sources of neutrinos, which are black only for the reason that no particle radiation is able to escape them, which is on the way with c or even slower. If a black hole does hurl neutrino radiation into space, than that must be considerably faster than c, as normal neutrino physicists still by no means can imagine it today.

But the neutrino radiation only can be detected after it has been slowed down to a value, which is smaller than c. If the slowing down occurs slightly assymmetrical, then as a consequence a mean of the mass different from zero appears. The „measurement" of such a rest mass, as it at present is propagated and celebrated, is a classical measurement error! As long as a neutrino on the way to us still is faster than the light, the mean of its mass is generally zero. The effective value of the mass of a neutrino is however considerable. Only it is able to give account for the sought-for dark matter, as far as it must exist in the today supposed form anyway.

The Tesla radiation, that the discoverer Nikola Tesla already in own experiments had found out, is faster than light (chapter 9.7 and 17.2). Since this Tesla radiation according to the description is identical with the neutrino radiation, since it so to say forms a subset, I will call neutrino radiation all the scalar waves, which are faster than the light. This stretches from the weak radiation at low frequencies up to the hard neutrino radiation of cosmic origin. But the hardness of the radiation does not only increase with the frequency, it in particular increases with the velocity.

The neutrino radiation first of all is carrying energy. On top of this basic wave radiation in addition information can be modulated. Doing so extremely complex modulation variants are offering. Of this kind we must imagine thoughts, as being complex modulated vortices, which can propagate as scalar wave in space. Rupert Sheldrake calls this vortex field a morphogenetic field. At this place merely is pointed at his very interesting research results[i].

Thoughts can be standing in space, in the form of localized noise, but they also can move with speeds faster than light. According to that a communication with intelligent beings from other star systems by all means wouldn't be an Utopia anymore.

Every fast neutrino forms an individual ring-like vortex (fig. 7.12). The slower the scalar wave is, the more dependent the vortices become. The photon already can consist of two ring-like vortices (fig. 4.6), whereas plasma waves and other slow scalar waves can form from a multitude of individual vortices, which are rotating around each other, to form vortex balls and vortex streets (chapter 4.9 - 4.11). From this circumstance already results very different scalar wave behaviour in the different areas of the velocity of propagation. This trend for small velocities can as well be observed towards lower frequencies. For a certain wavelength the frequency after all (according to eq. 22.1) is proportional to the velocity of propagation.

<i>: R. Sheldrake: Seven experiments that could change the world. New York: Riverhead 1995

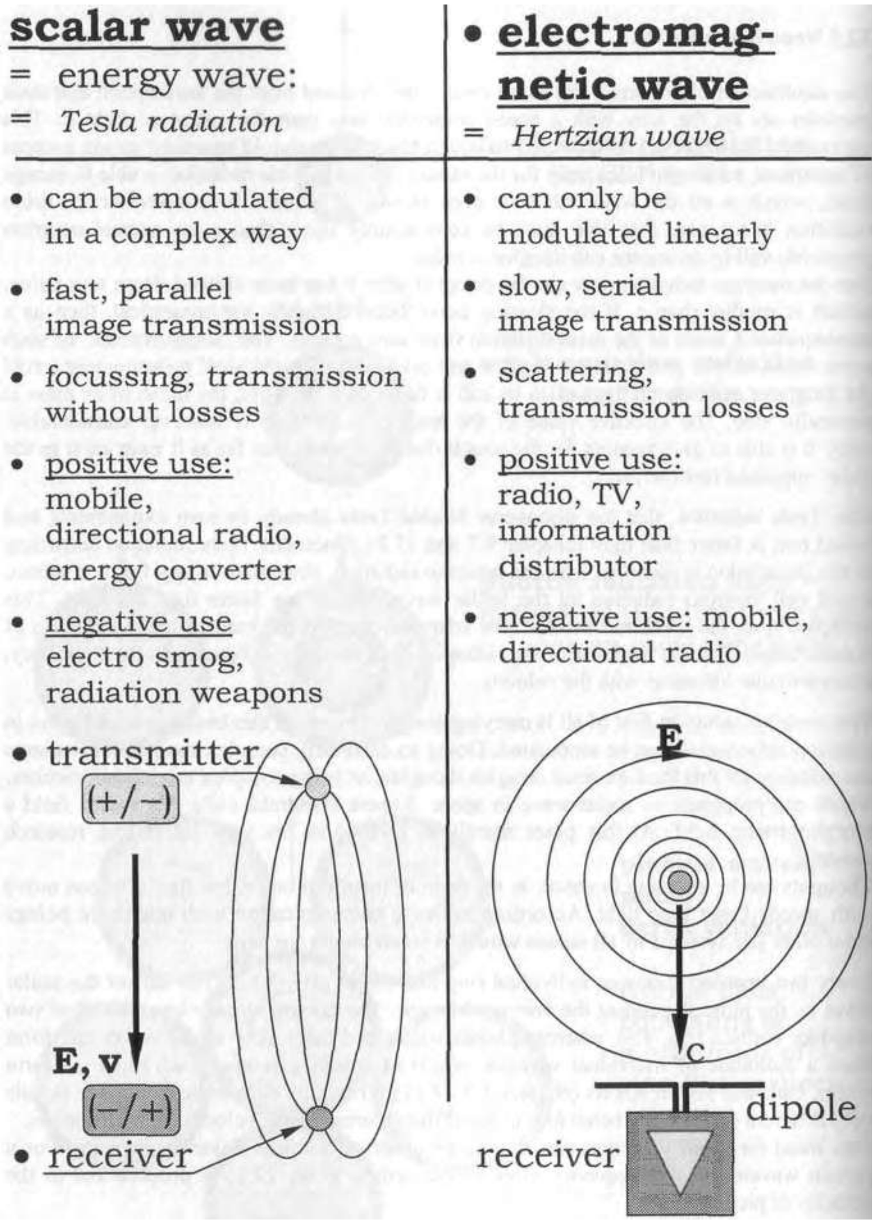

scalar wave = energy wave: = *Tesla radiation*	• **electromagnetic wave** = *Hertzian wave*
• can be modulated in a complex way	• can only be modulated linearly
• fast, parallel image transmission	• slow, serial image transmission
• focussing, transmission without losses	• scattering, transmission losses
• positive use: mobile, directional radio, energy converter	• positive use: radio, TV, information distributor
• negative use electro smog, radiation weapons	• negative use: mobile, directional radio

Fig. 22.6: Scalar waves and radio waves, comparison of the properties.

22.6 Parallel instead of serial image transmission

We continue with our considerations concerning the special properties of scalar waves, represented in the left column, and compare these with the well-known behaviour of electromagnetic waves in the right column (fig. 22.3 is now followed by fig. 22.6). If we again take up the possibilities for modulation and the transmission of information, then it becomes very clear from the comparison that we today work with a technology, which we it is true master more or less, but which is everything else but optimum. For the Hertzian wave the velocity of propagation is constant. With the frequency therefore at the same time also the wavelength is being modulated. But that strongly limits the information transmission. An image for instance must be transmitted serially point after point and line after line. The serial image transmission takes place very slowly, for which reason the velocity of the PCs permanently must be increased, so that the amount of data can be managed.

With the clock frequency on the other hand also the losses increase, so that in the end the CPU-cooler limits the efficiency of modern PCs. Something our engineers obviously do wrong, as a comparison with the human brain clarifies. Our brain works without a fan. For it a clock frequency of 10 Hertz is sufficient. It needs neither Megahertz nor Gigahertz frequencies and despite that is considerably more efficient.

Nature only works with the best technology. The second best technology, as it is put to use in our machines, in the evolution wouldn't have had the slightest chance of surviving. The strategies to optimize of nature are merciless. In a free economy that goes completely different. There the „bunglers" are joining together to companies dominating the market, buying up the innovative ideas without further ado, to let them disappear in the drawer, so that they can bungle further in the way they did until now. It after all have been the lousy products, which have made them to the companies they are today. The ego of power is incompatible with the interests of nature.

Nature works with scalar waves and their velocity of propagation is arbitrary. Wavelength and frequency now can be modulated and information can be recorded separately. In this manner a whole dimension is gained to modulate, the image transmission can take place in parallel, which means considerably faster, safer and more reliable. As anyone of us knows by own experience, assembling the image takes place all at once, the memory of past images takes place ad hoc. Nature is indescribable more efficient than technology with the scalar wave technique.

If we again take the right-hand side of fig. 22.6 with the properties of the Hertzian wave. In the opinion of Nikola Tesla it is a gigantic waste of energy. The broadcasting power is scattered in all directions and the transmission losses are enormous. At the receiver virtually no power arrives anymore. To receive a radio signal the antenna power has to be amplified immensely. It is a wave, which actually only can be used as radio wave, thus as a wave with which arbitrary many participants should be reached.

This wave however is completely useless, if it concerns a point-to-point connection. If I want to call someone, to talk only with him, a radio wave is the entirely wrong method, because I with that bother hunderds and thousands of people that I don't want to call at all!

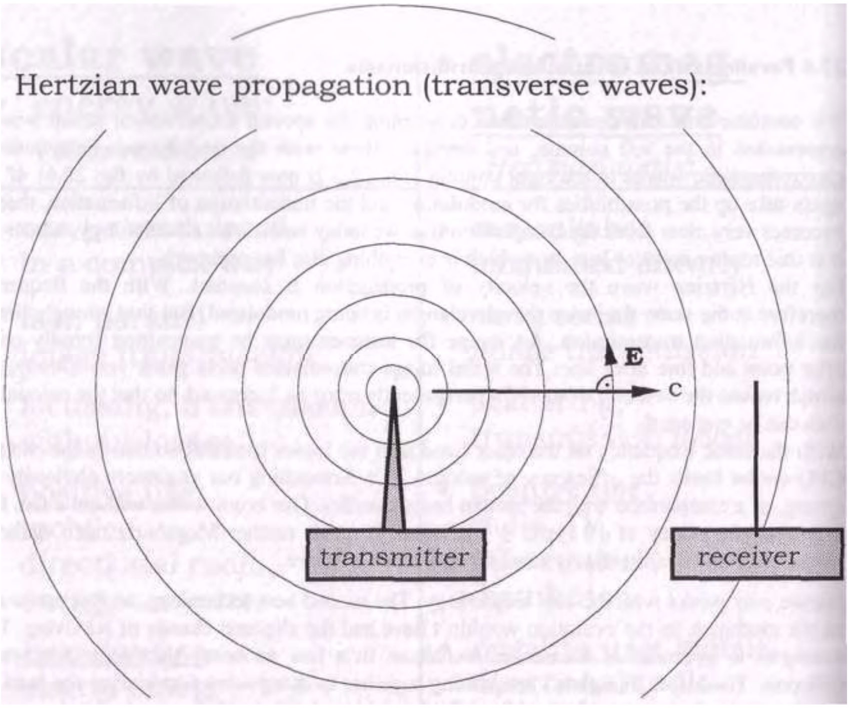
Hertzian wave propagation (transverse waves):
E
c
transmitter
receiver

Tesla radiation (radiations) = scalar wave, longitudinal wave propagation:

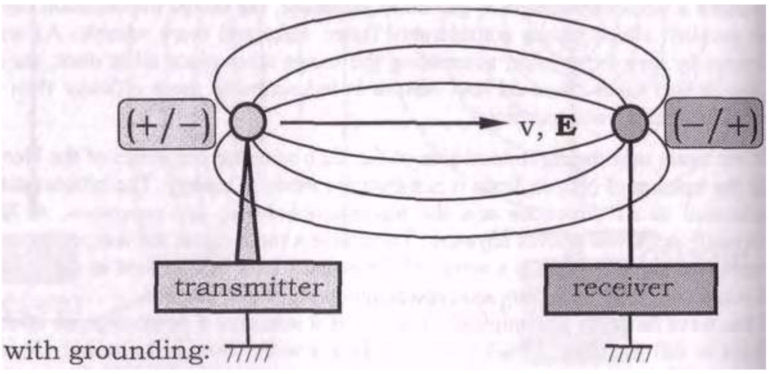
(+/−)
v, E
(−/+)
transmitter
receiver
with grounding:

Fig. 22.7: ____Comparison of radio waves according to H. Hertz and electric scalar waves according to Nikola Tesla.

22.7 Comparison of the properties

The course of the field lines clarifies the difference. For the scalar wave all field lines going away from the transmitter run together again. As long as no scattering field occurs, there also won't be any transmission losses. It is an energy wave, for which the full broadcasting power is transmitted wirelessly and which arrives at the receiver, by that focussing the field lines again.

Here one at once numerous technical applications come to mind, if energy should be transmitted wirelessly. A TV, which supplies its remote control with energy itself, lelemetric installations, which are fixed at difficult accessible or rotating machine parts and which can't work without energy supply.

For a mobile with 3 Watts of broadcasting power only a few microwatt arrive at the ground station. If I would have a scalar wave mobile, which functions with resonance, then a broadcasting power of some microwatt will be completely sufficient, to carry out a telephone call right through the earth. This minimum broadcasting power suffices, because everything, what is being transmitted, arrives at the receiver - crucial is that the conditions of resonance are fulfilled. That means, both must have the same frequency and the opposite phase. In addition the modulation has to fit, so that on the one hand not several participants in the conversation are getting in each other's way. For a purely carrier wave transmission on the other hand there would be the risk of natural fields being collected also and the power at the receiver taking inadmissibly high values. This is prevented effectively by a correspondingly complex modulation. Nature solves the problem in exactly this manner.

Mobile phone technology with scalar waves of course still is a pie in the sky. A big challenge for the engineers poses the adjustment condition of opposite phase and the fitting modulation, which Tesla called .,individualisation"<i>. Entirely by the way the network runners and the telephone companies are getting quite superfluous.

Telephone charges, so one perhaps can read in the history books in a hunderd years, were the indication of a century of rigorous exploitation of man and nature. With scalar waves a direct, more dimensionally modulated information transmission directly with the partner of conversation is possible and sufficient energy is available to humanity any time and any place, without being dependent on any companies! This notion is not new, but it is inconvenient for the rules; already Nikola Tesla has written about it, but obviously no-one wanted to listen to him<ii>.

Scalar waves are able, that is made clear by the properties, to revolutionize both the energy technology and the information technology fundamentally. It is more than only a technology for the new century. Scalar waves are a chance for the whole millennium!

<i>: N. Tesla: Transmission of electrical energy without wires, Electrical World and Engineer 7.1.1905; Edition Tesla Vol. 4, P. 131.

<ii>: N. Tesla, New York Times of 15.9.1908; Edition Tesla Vol. 6, P. 241.

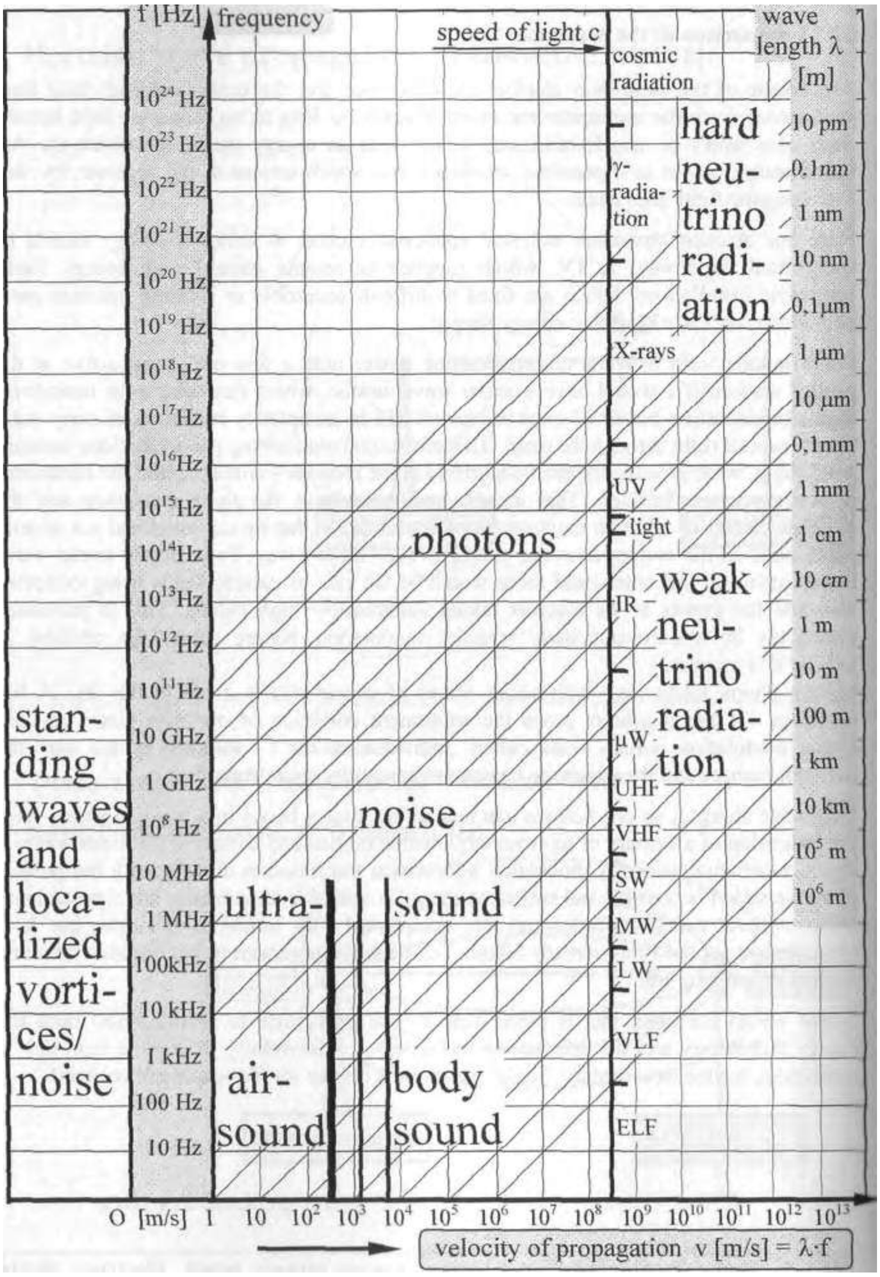

Fig. 23.1: Frequency diagram of longitudinal and transverse waves as they depend on the velocity v.

23. Research of scalar waves

Scalar waves are still unexplored area, scientific new ground as it were. Individual research scientists already have selectively ventured forward in this area and have described properties of the scalar wave investigated by them in their special research area mostly in measurement technical manner. But as a rule they lack the physical relation, as it is derived in this book for the first time. If we don't proceed from individual measurements, but from the wave equation and the mathematical physical derivation of scalar waves, then we have the great chance to understand as something belonging together on the one hand noise, photons, neutrinos and lots of other known phenomena as well as on the other hand still unknown phenomena, which are called parascientific. We should remember that we without theory of Maxwell and the representation in a frequency band today still wouldn't know that the radio waves (LW, MW, KW, UHF), the microwaves (μW), the infrared thermal radiation (IR), the light and the X-rays concern just the same phenomenon. The graphic representation of both waves in one diagram in this place is extremely helpful.

23.1 Frequency diagram

In fig. 23.1 the frequency f is shown against the velocity of propagation v with the wavelength λ as parameter. The broad line at $3*10^8$ m/s represents the speed of light c. Here the frequency band of the transverse waves can be found again in the well-known one-dimensional representation.

Crosswise to that, somewhat unusual, the longitudinal waves run. These start at the left at localized noise, over the sound as it propagates in air, in water and in metal, over a large, to a large extent still unexplored, range of the bio photons, the heat vortices and of the dowsing rod effects and end on the other side of the speed of light at the neutrinos. Between that the special case is settled that the particles, or said better vortices, propagating as a scalar wave have exactly the speed of light. It gives reasons for the circumstance, as already mentioned, that light can appear as wave or as photon radiation. It. according to the wave equation, after all always consists of a combination of both forms. At very high frequencies, e.g. the cosmic radiation, this combination is shifted in the direction of vortices and their distribution as a scalar wave, at low frequencies the tendency inversely goes in the direction of the normal wave.

If we assume that for the transverse wave over all frequencies a dozen of specialized gauges is necessary, each of them also can be switched over in range several times, then we can project that to record a scalar wave of a certain frequency over all velocities of propagation likewise 12 devices and for the whole field shown in fig. 23.1 approximately 12x12 = 144 devices will be necessary. Of these 144 gauges today just 12 are available. There thus still are missing 132 pieces, which should be developed. With these gauges, so I am convinced, the many white spots in the diagram can be tapped scientifically little by little if a systematic procedure is used. My vortex theory thus will be attached a central importance.

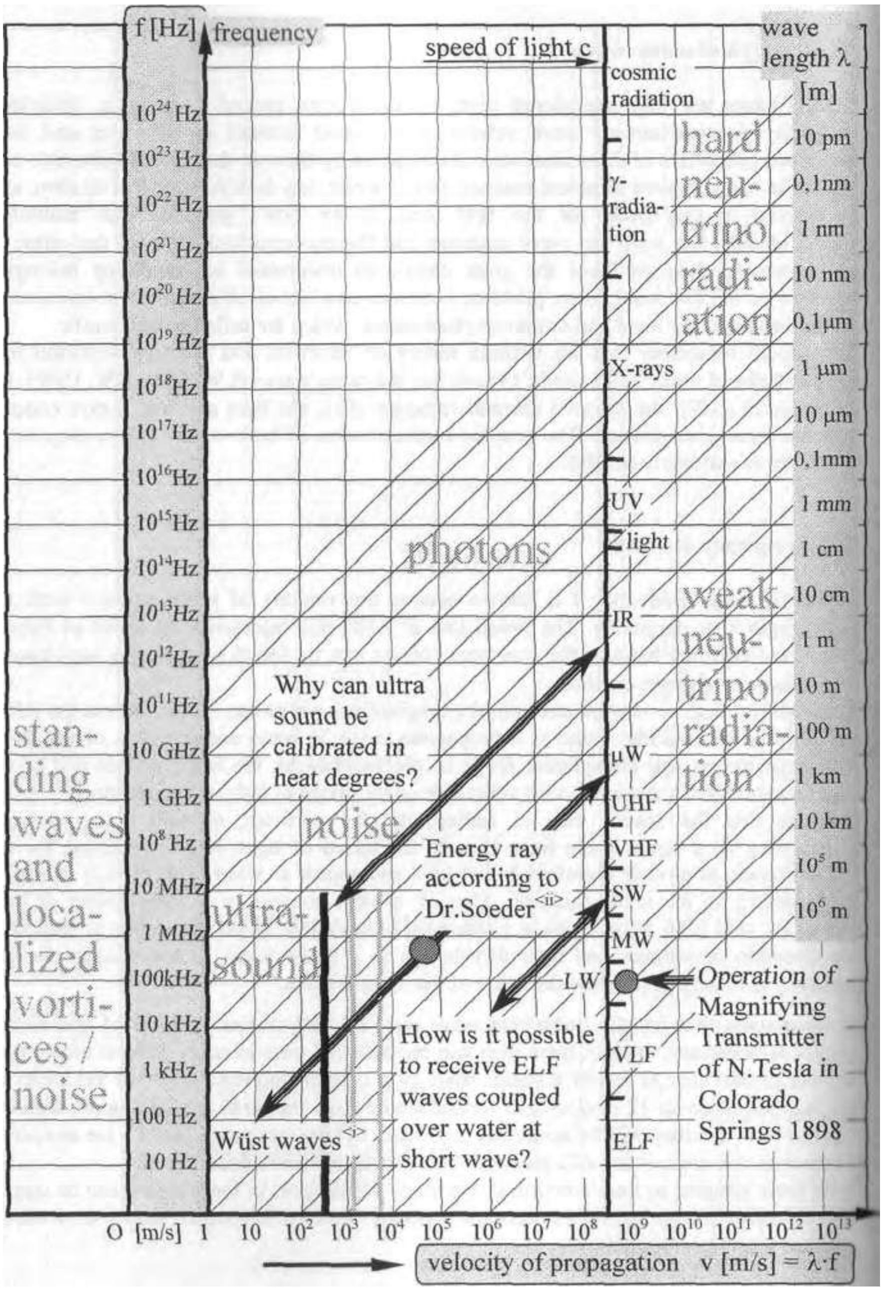

Fig.23.2: Frequency diagram with entries concerning Wust waves,
Tesla transmitter and various not understood effects.

23.2 The reverse of the HF medal

The diagram possibly may settle disputes, like the ones between the „Wustlingen" (note of the translator: the author here is playing with words, Wustlingen literally means lechers) and the „high-frequency engineers".
Professor Wust of Munich already 1934 had proven, that the Wust waves named after him in air have a velocity of propagation of approximately 10 m/s<i>. He investigated them in the frequency range between 16 Hz (λ= 60 cm) and 500 Hz (λ= 2 cm). The high-frequency technicians immediately have converted to the speed of light, since they don't know anything else, and on the spot have shifted the phenomenon into the range of the microwaves between 0.5 GHz (λ=60 cm) and 15 GHz (λ= 2 cm). In the diagram now can be found, at which place this phenomenon belongs actually. It presumably concerns the same „energy ray", which Dr. Soeder has observed and proven already at velocities of propagation of 55.2 km/s and wavelengths around 10 cm<ii>.

Furthermore follows from the diagram (23.2), why ultra sound can be calibrated in heat degrees, and why radio reception of signals from a completely other frequency range is possible by principle, if for identical wavelength a velocity of propagation different from c occurs. Further the range of operation used by Tesla at approx. 1.6 times the speed of light is depicted.

Also the noise is such a book with seven seals. It concerns, it is true, a fixed chapter of high-frequency engineering, but without visible connection to the other chapters. That might be based on the circumstance that the relation, as it is dictated by the wave equation, isn't recognized.
Every high-frequency signal is accompanied by a noise signal; every antenna produces more or less noise. HF engineering dictates the measures, how the noise can be suppressed resp. the signal-noise distance be increased. The goal is to make the electromagnetic wave stand out in such a way from the noise that it can be received. For that the measurement setup must be choosen correspondingly, are the measurement cables dictated, and must be paid attention to power adaptation and corresponding termination resistors. For measuring HF correctly all interference influences should more or less disappear.

But what are the interference influences, which the HF-technician suppresses? According to the wave equation it are the scalar wave parts, and to that also is counted the noise. Every HF-technician thus knows the scalar parts as interfering noise signal, but doing so he completely fails to notice the technical advantages and chances, which e.g. are present in the noise.
If by means of an autocorrelation function a noise signal is compared to itself then often a hidden message comes to light. According to that, informations can be hidden in the noise and of course transmitted wirelessly. This important circumstance is known, but it is hardly used. The noise vortices thus can be modulated in an extremely complex way. There can be transmitted a lot more information in the noise as scalar wave, than with the radio wave.

<i>: J. Wust: Physikalische und chemische Untersuchungen mit einem Rutenganger als Indikator. Further references in W-B-M 3/ 1991, S. 57.

ii>: A. Soeder: Ohne Strom lauft neuer Energiestrahl rund urn den Erdball. Raum & Zeit 59/92 Seite 62-67.

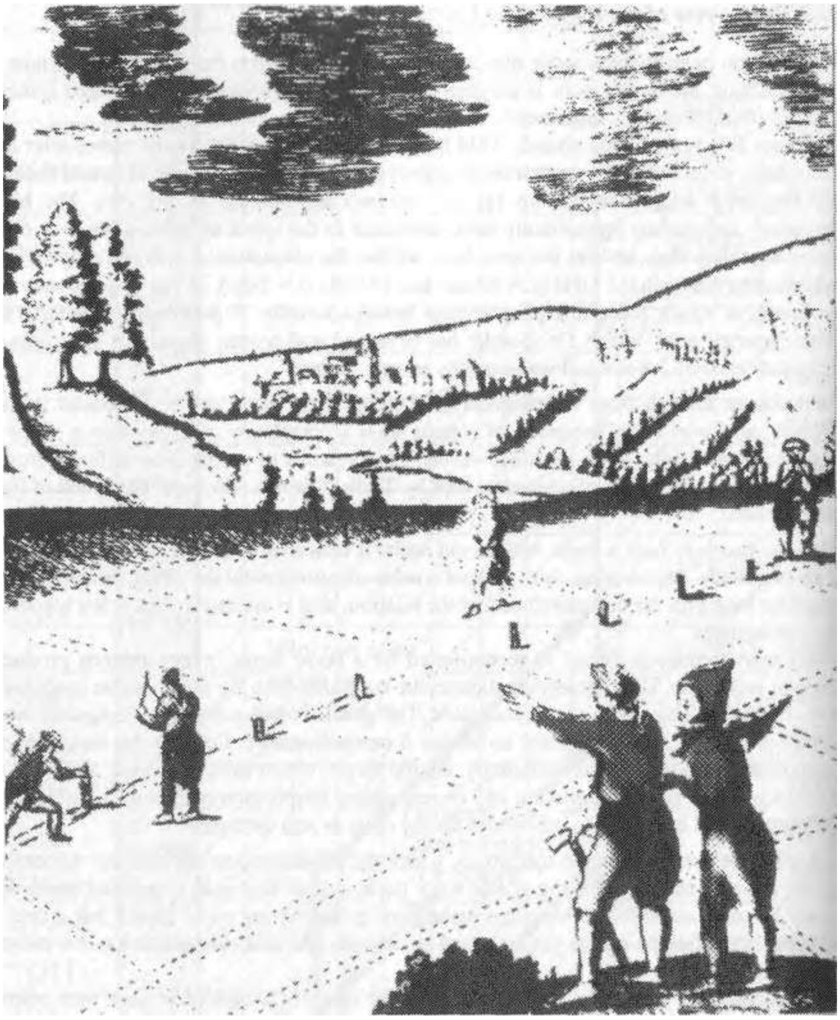

Fig. 23.3: ___ Dowsers dowse an ore vein.<i>
From: Speculum metallurgiae politissimum, Rossler, Dresden 1700.

<i>: Ludeling, Hartmut: Handbuch der Radiaesthesie, Verlag Eike Hensch 1994, S. 145, ISBN 3-927407-09-7.

The possibility of more dimensionally modulating and parallel image transmission make scalar waves superior to the radio waves in such a way that one should seriously think about the reversed procedure, in which the noise is conceived as useful signal and the radio wave as interference signal.

In the frequency diagram (23.1 and 23.2) can be read like in a diary, in which most pages it is true still are empty, but in which some very informative marginal notes can be found for different points of difficulty. It mostly are parascientific phenomena, which, not understood and excluded by textbook physics, are waiting to be taken up in the building of physics. As an example we'll pick out an extensive area of alternative research.

23.3 Radiesthesia

A particularly broad spectrum of scalar waves is provided by the already several times mentioned earth radiation. An ancient science of experience, which troubles the exploration of the earth radiation, is the geomancy. The Roman land surveyors, the Augures, used as an aid a flat coil like also Tesla did to receive scalar waves. The Lituus, as the device was called, resembles so much that of the Etrusks that we must proceed from the assumption that the method is much older (fig. 16.10).
This part of the scalar wave research, also called radiesthesia, is derived of „radiare", which can be translated with "send out rays" resp. „perceive". It describes the doctrine of the sensitivity to radiation of man. Doing so the radiation sensitive uses his own nerve framework as a biosensor.

The nerve conduction could be derived as a biological variant of the Tesla one wire transmission, for which ring-like potential vortices are passed on as action potentials in form of standing waves (fig. 9.6). Of this kind are also the control commands, which cause a muscle to contract. If now corresponding vortices are picked up by means of a dowsing rod or similar aids in our nerves, then the contract addressed muscles contract, because they can't distinguish, from where the command comes. This unconscious nerve twitch leads to a swing of the dowsing rod and to the well-known dowsing rod phenomenon. But man can't replace a technical gauge. Hence one speaks of dowsing and not of measuring. To this should be added the condition of resonance, which must be fulfilled. Since every person however builds up other resonances, dowsed results of others often can't be reproduced. But from this particular difficulty one cannot draw the conclusion that the phenomenon does not exist and radiesthesia is not a science. Series surveys and statistical analyses here in any case don't lead any further. It always only are individual talented dowsers, who have at their disposal really fantastic abilities and find with great certainty water, ores and even oil.
With the perception of a physical phenomenon it mostly starts. Cultural man looks after his discovery as cult, whereas modern man, guided by the wish for reproducibility and more objectivity, is troubled to design and to build a technical measurement work. With regard to the scalar wave in general and the radiesthesia in special we still are at the stage of the Stone Age.

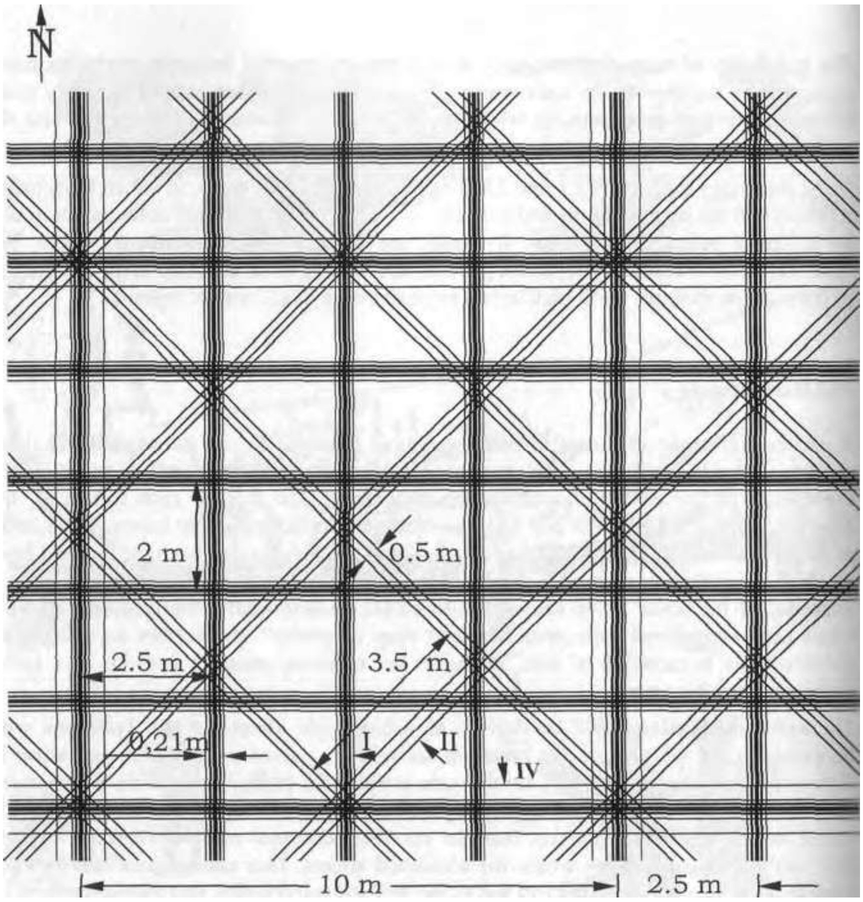

Fig. 23.4: Nets dowsed at latitude 49.<i>

I. net: ,,global net" according to Dr. E. Hartmann;
orientation: N - S: 2 m distance and E - W: 2.5 m distance

II. net: ,,diagonal net" according to Dr. M. Curry;
orientation: NE - SW and NW - SE: 3.54 m = 2.5 m * sqrt(2)

[III. ,,lightning net" according to R. Schneider; orientation like I.]

[IV. net acc. to Benker; like I. but: N - S: 10 m and E - W: 10 m]

[V. sun/planet net acc. to W. Auer; with 3.66 m orientation S./P.]

According to Prof. E.G.Hensch et al.:

<i>: Hensch: Radiaesthesie im landlichen Bauen und Siedeln, Arbeitskreis zuf Landentwicklung in Hessen, W4, Wiesbaden 1987; resp.: W. Auer: Erdstrahlen?... AWGEO 199, Eigenverlag 1998; resp.: Mayer/Winklbaur: Biostrahlen, 5. Aufl. ORAC Verlag, Wien 1989, Seite 168

23.4 Prediction of earthquakes with the help of scalar waves

The standing wave character of the earth radiation is a help. There are dowsed points and lines of maximum radiation intensity, which form nets, which encompass the whole world. On the one hand we are dealing with a net oriented in direction north-south and east-west (Hartmann net) and on the other hand with a net standing diagonally to that under 45° (Curry net). Because the angle of 45° can be derived as borderline angle from a vortex field calculation<ii>, I already early have pointed to the circumstance that it here presumably concerns a vortex phenomenon<iii>.
Because the nets in addition are dowsed in air, it must be vortices of the electric field, so-called potential vortices<iii>, which here form regular structures. The formation of vortex lines and complete vortex streets (chapter 4.9), which consist of countless individual vortices, can be explained as follows:
Electric scalar waves propagate in the direction of the electric field strength and mediate field vortices, e.g. neutrinos. If at a certain moment the transmitter carries a positive and the receiver a negative charge, then all the particles which are positively charged are repelled by the transmitter and attracted by the receiver. All run at the same time towards the same goal, although all mediated particles carry the same charge and repel each other! This incompatibility can be compensated partly, by the vortices rotating around each other. In this circumstance can be seen the reason for the structure shaping, the formation of some lines in the countryside (fig. 23.3). The distances between the lines have characteristic values, which allow conclusions about the wavelength of the standing waves. We must assume that they dictate the structure shaping and that the spatial vortex distribution aligns with the nodes and antinodes of the respective standing wave. The distance between the lines, which corresponds to half the wavelength, becomes smaller and smaller towards the North Pole and the South Pole of the earth, the net thus narrower and narrower. Also is the net said to change strongly before an earthquake. This all are clues for the circumstance that the structure shaping radiation comes from the earth, that the cause must be sought in the earth radiation.
It should be noted marginally that with scalar wave detectors, which permanently scan the nets, a just as effective as inexpensive earthquake prediction should be possible. Such a facility would be an enormous relief and help for all earthquake warning services. It even should be possible to determine in advance the future epicentre, if there is measured at the same time at if possible many stations and the respective deviations are compared.

A further influential factor is the composition of the subsoil; e.g. ores and metals influence the earth radiation. Water shows a special affinity for the earth radiation. It does collect the radiation and after bundling it up releases it again. To blame is the high dielectricity of water ($\varepsilon \cong 80$), which again favours the formation of potential vortices .
A technical use of this effect would be the neutrinolyse (see fig. 17.6-17.8 and 18.1), the splitting of water molecules by neutrinos if these take the state of an electron, hydrogen escapes and the oxygen content in the water increases. If a neutrino however shows as a positron, then it annihilates and there is formed a light flash, which serves the „experts" in neutrino detectors as proof.

<ii>: Meyl, K.: Wirbelstrome, Dissertation Universitat Stuttgart 1984

<iii>: Meyl, K.: Potentialwirbel Band 1, INDEL Verlagsabt. 1990

Spectroid of the radiation field of a subterranean flowing water veign (according to P. Schweizer)[i]:

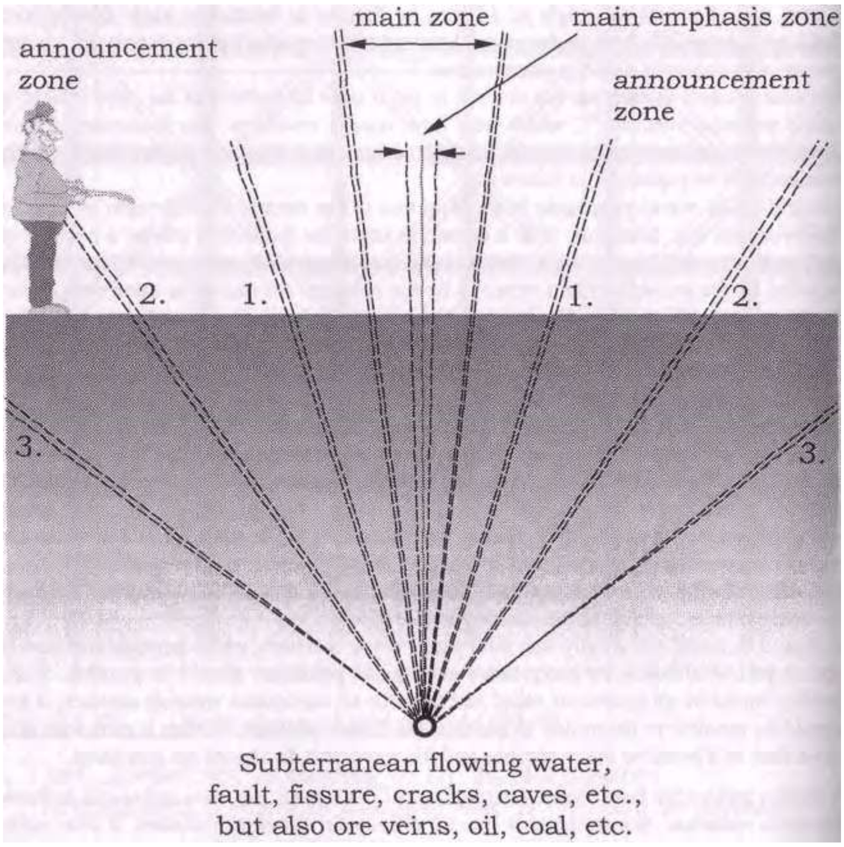
main zone
main emphasis zone
announcement zone
announcement zone
2.
1.
1.
2.
3.
3.
Subterranean flowing water,
fault, fissure, cracks, caves, etc.,
but also ore veins, oil, coal, etc.

Fig. 23.5:____ Radiaesthetic search for water[i],
(dowsing of the earth radiation)

---

[i]: Ludeling, Hartmut: Handbuch der Radiaesthesie, Verlag Eike Hensch 1994 S. 145, ISBN 3-927407-09-7.

23.5 Searching water with the help of scalar waves

Faults of earth's crust lead to fissures and cracks in the rock, which often fill up with standing or flowing water. Potential vortices of the earth radiation are attracted by the high dielectricity of water and are radiated again in slowed down and focussed form. On the surface of the earth the mixture of various scalar waves arriving there can be dowsed as main emphasis zone. For reason of the broad radiation spectrum the results of proficient dowsers in the main emphasis zone more often coincide.

At both sides in addition spectroids are forming, which are dowsed as parallel running vortex lines. Because every harmonic wave produces other lines and every dowser reacts to other resonances, a profound ability is required, if from the distance of the so-called announcement zones the water depth should be inferred (fig. 23.5). Some specialists indicate besides the place and the depth also the amount of deposit and if need be the water quality. For that they analyse the radiation intensity, but unfortunately many all too often overrate their abilities.

Already the Augures, the land surveyors in the age of the Romans have aligned their streets and the castles with the nets and the lines in the countryside. Even today new evidence about the central importance of the standing wave character of scalar waves can be found at excavations (fig. 16.10). Since every scalar wave also occurs coupled with a radio wave by means of the wave equation, earth rays originally could be detected conventionally with field strength gauges. But the intensive use by radio stations has made it necessary to change from the short wave range to the VHF-range and further to the UHF-range. It were indirect measurements, which could be interfered easily. Today geologists work with ELF scanners in a range between 15 and 30 KHz if they search for water. They analyse the emitted signals of submarine transmitters, because these are attracted and amplified by water carrying layers.

Experiments also are carried out with noise receivers or unused broadband TV channels. Others analyse influences of the earth radiation on the magnetic field of the earth, but for this method the expressiveness is controversial.

Again others walk along a path with a scintillation counter measuring the inverse profile. At the places of maximum earth radiation a minimum of gamma radiation is measured. This stunning relation only can be interpreted as follows: the natural radioactive decay at these places takes place accelerated and that proves:

1.. that earth radiation sets radioactivity free and causes it in the first place (chap. 17.1 pp).
2. that in the case of earth radiation it actually concerns slowed down neutrino radiation.

But also this method works indirectly. It only functions in the open air and then only in the case of unspoiled nature, which sometime was covered with an even layer of radioactive dust. By cultivation or plantation the „radioactive layer" very often is changed and mixed up, whereupon the measurable profile hardly allows conclusions.

For indirect measurement methods of that kind, for which the measurement variable stands in a relation, which isn't known in more detail, to the scalar wave of interest, generally caaution is advisable. Too often a message is seen in a wonderful 3-D diagram of the magnetic field strength or the radioactivity distribution, which has much more mundane causes.

Lecher air antenna according to W. Busscher<iii>:

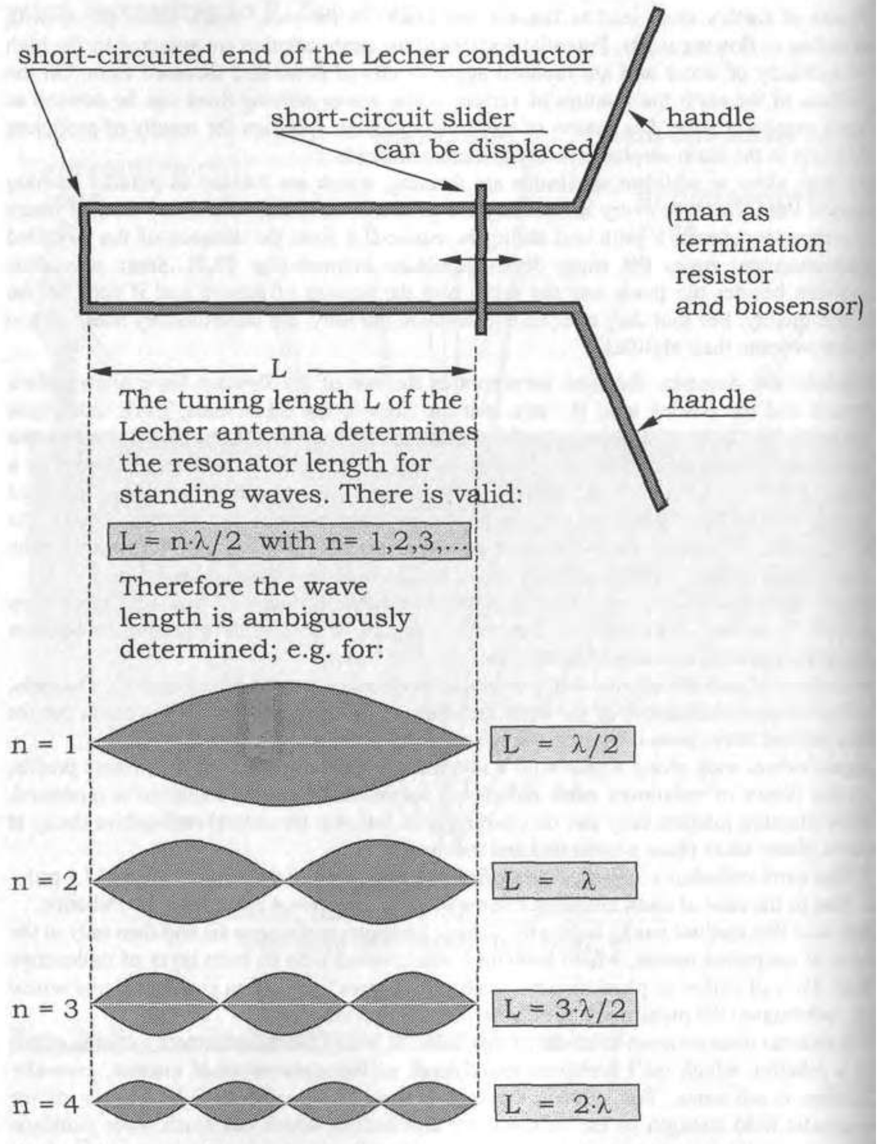
short-circuited end of the Lecher conductor
short-circuit slider can be displaced
handle
(man as termination resistor and biosensor)
L
handle
n = 1
L = λ/2
n = 2
L = λ
n = 3
L = 3·λ/2
n = 4
L = 2·λ

The tuning length L of the Lecher antenna determines the resonator length for standing waves. There is valid:

$$\boxed{L = n \cdot \lambda/2 \quad \text{with } n = 1,2,3,\ldots}$$

Therefore the wave length is ambiguously determined; e.g. for:

Fig. 23.6: Lecher conductor as scalar wave antenna.

Can be read at Prof. Dr. Ing. Christof Rohrbach:

<iii>: Rohrbach: Radiasthesie, Physikalische Grundlagen und Anwendung in Geobiologie und Medizin, Haug Verlag Heidelberg 1996, Seite 100.

23.6 Receiving scalar waves with the Lecher antenna

In the case of technical gauges it is normal to change the frequency tuning. For that the capacity of a capacitor is being varied, which is part of a resonant circuit. This method can simply be realized with the today available construction elements. Biological systems on the other hand, for instance a person, work primarily with a variation of the wavelength. In this case (according to eq. 22.1) the frequency is directly proportional to the velocity of propagation.

So provide for instance radiaesthetic dowsings clues for the wavelength. The frequency however can't be given, because that oscillates along with the velocity of propagation and at present the forming noise signal still can't be analysed. There can't be bought electronic construction elements, which like a body cell would be able of oscillations of length. Oscillation quartzes and piezo elements, on the basis of which scalar wave detectors could be constructed, form an exception.

Similar to the Lecher conductor different forms of tunable Lecher antennas have been developed. which are used for dowsing. Such a frame antenna, developed strictly according to the rules of high-frequency technology stems from W. Busscher (fig. 23.6)<i>. With the short-circuit slider a closed-loop antenna circuit in the upper part is tuned to half the wavelength or an integer multiple of it ($L = n \cdot \lambda/2$ with n = 1,2,3...). The person, who holds the Lecher antenna with both hands, forms the termination resistor for the attached antenna circuit. At the same time he prevents the occurring of an effect back on the tuned resonant circuit by means of damping of the Lecher conductor. The sense of this arrangement is lying in the circumstance that man as a „biosensor" should feel if standing waves, which I call vortices, are formed in the Lecher antenna.

A modification of the Lecher air antenna is the Lecher dowsing rod according to R. Schneider<ii>. It is produced as an etched board, which has a dielectricity of approx. 4. As a result the velocity of propagation and the wavelength are only half their normal value. In addition a shortening factor V is introduced, which depends on the construction of the Lecher dowsing rod<iii> ($L = V \cdot n \cdot \lambda/4$ with V = 0.952 and n = 1,2,3...).

In fig. 23.7 some published tuning values have been drawn up. The values for want of corresponding technical gauges it is true still haven't been confirmed, but they allow a certain insight into the world of the scalar waves, as they are influenced and radiated by water, by metals, by oil or even by other planets, the sun and the moon. Doing so some questions remain open: We for instance don't know, if the corresponding wavelength concerns the basic wave or only an mth harmonic wave. The details not only are unreliable, they moreover also are ambiguous. If however several tuning values L' are present for which resonance occurs, then if need be an integer divisor can be sought and the wavelength can be determined. But that doesn't function always.

<i>: Will Busscher: Die Luft-Lecher-Leitung als Wunschelrute. Eine Methode um die Wellenlangen zu bestimmen. Wetter-Boden-Mensch 3/ 1991, Seite 46-57.

<ii>: R. Schneider: Einfuhrung in die Radiasthesie Teil II: Leitfaden und Lehrkurs ..., Oktogon-Verlag, Wertheim, 2. Aufl. (1984) S. 136.

<iii>: Rohrbach: Radiasthesie, Physikalische Grundlagen und Anwendung in Geobiologie und Medizin, Haug Verlag Heidelberg 1996, Seite 100.

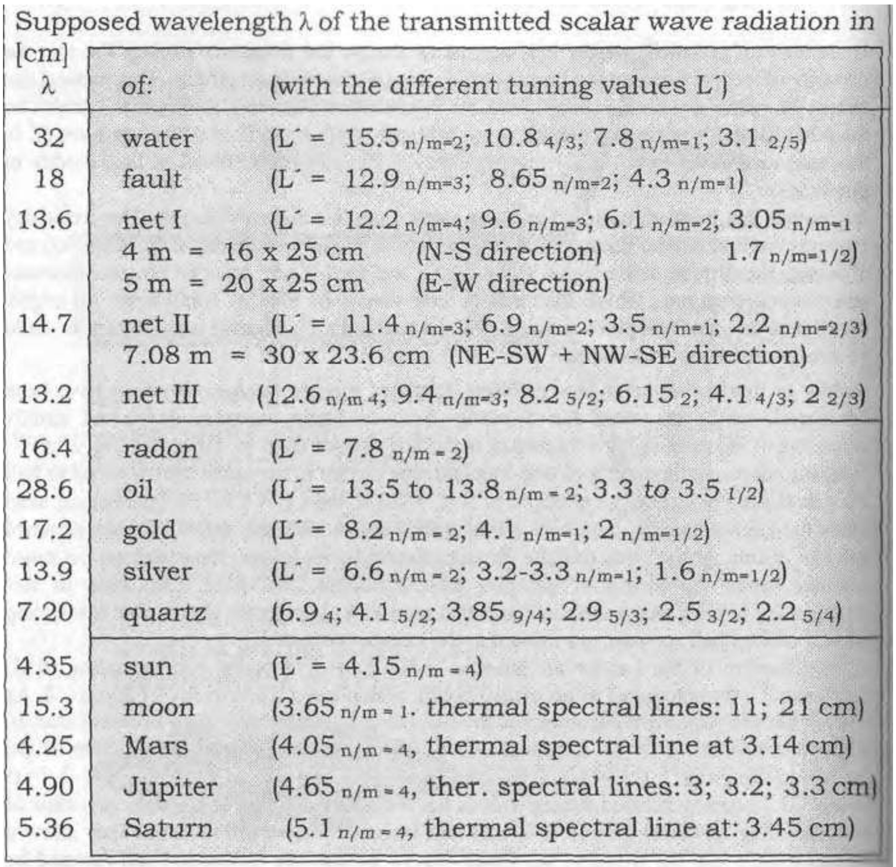

Supposed wavelength λ of the transmitted scalar wave radiation in [cm]

λ	of:	(with the different tuning values L´)
32	water	(L´ = $15.5_{n/m=2}$; $10.8_{4/3}$; $7.8_{n/m=1}$; $3.1_{2/5}$)
18	fault	(L´ = $12.9_{n/m=3}$; $8.65_{n/m=2}$; $4.3_{n/m=1}$)
13.6	net I 4 m = 16 x 25 cm 5 m = 20 x 25 cm	(L´ = $12.2_{n/m=4}$; $9.6_{n/m=3}$; $6.1_{n/m=2}$; $3.05_{n/m=1}$ (N-S direction) $1.7_{n/m=1/2}$) (E-W direction)
14.7	net II 7.08 m = 30 x 23.6 cm	(L´ = $11.4_{n/m=3}$; $6.9_{n/m=2}$; $3.5_{n/m=1}$; $2.2_{n/m=2/3}$) (NE-SW + NW-SE direction)
13.2	net III	($12.6_{n/m\ 4}$; $9.4_{n/m=3}$; $8.2_{5/2}$; 6.15_{2}; $4.1_{4/3}$; $2_{2/3}$)
16.4	radon	(L´ = $7.8_{n/m=2}$)
28.6	oil	(L´ = 13.5 to $13.8_{n/m=2}$; 3.3 *to* $3.5_{1/2}$)
17.2	gold	(L´ = $8.2_{n/m=2}$; $4.1_{n/m=1}$; $2_{n/m=1/2}$)
13.9	silver	(L´ = $6.6_{n/m=2}$; $3.2\text{-}3.3_{n/m=1}$; $1.6_{n/m=1/2}$)
7.20	quartz	(6.9_{4}; $4.1_{5/2}$; $3.85_{9/4}$; $2.9_{5/3}$; $2.5_{3/2}$; $2.2_{5/4}$)
4.35	sun	(L´ = $4.15_{n/m=4}$)
15.3	moon	($3.65_{n/m=1}$. thermal spectral lines: 11; 21 cm)
4.25	Mars	($4.05_{n/m=4}$, thermal spectral line at 3.14 cm)
4.90	Jupiter	($4.65_{n/m=4}$, ther. spectral lines: 3; 3.2; 3.3 cm)
5.36	Saturn	($5.1_{n/m=4}$, thermal spectral line at: 3.45 cm)

Fig. 23.7: Details of not confirmed wavelengths.<i>
(derived from tuning values of a Lecher antenna
$L' = . V \cdot \lambda \cdot n / 4 \cdot m$ with $V = 0.952$
and $n = 1,2,3\ldots$ for $\lambda/4$, $\lambda/2$, $3\lambda/4$, λ, ...
and $m = 1,2,3\ldots$ for 1 the basic wave
and the 2nd, 3rd, ... harmonic wave

---

according to:
<i>: Ludeling, Hartmut: Handbuch der Radiaesthesie, Verlag Eike Hensch 1994, S. 161 ff. The author gives the clue: ,,The values have been determined empirically by different persons and predominantly could be confirmed by the author and his co-workers. There however always can occur deviations caused by different dowsing methods. Every value therefore always should be checked with own measurements."

23.7 Assignment according to wavelength

As said, the methods to determine the wavelength by means of the Lecher antenna are ambiguous and unreliable. So water should have a wavelength twice as long as a fault. But since water often is collecting or subterranean flowing in faults, one is entitled the suspicion that here $\lambda/2$ was taken, where in reality the full wavelength λ acts. The rest of the table after that should be looked at with the same scepticism. The table nevertheless is not wholely uninteresting, because it points to certain trends and is able to furnish clues, which if need be can represent a help of orientation in the systematic research of scalar waves.

At the planets is remarkable that the Lecher conductor reacts to values, which correlate with the respective thermal spectral lines. If these spectral lines propagate with c and it actually should concern the same cause, the frequency thus should be identical, then the planets would emit scalar waves, which are faster than the light. In the case of Mars 1.35 times the speed of light c would be present, in the case of Saturn 1.55 times and in the case of Jupiter the values would lie between 1.46c and 1.63c. In the case of the moon two thermal spectral lines are being measured, which on the one hand have as a result 1.39c and on the other hand 0.73c. At this place still ample research tasks should be solved.

In the case of scalar waves the wavelength is the most important factor to refer to. The frequency however varies continually, which in electrical engineering is called noise. That again is connected to the circumstance that frequency and velocity of propagation oscillate, so that merely an average value can be given at a time. But that hardly can be measured, which again makes the assignment more difficult. We thus still aren't capable to enter the table values from fig. 23.7 into the frequency diagram of fig. 23.1.

A technical gauge for scalar waves is necessary. A practical solution could look as follows: A noise transmitter tunable in frequency and wavelength operates on top of a carried along noise receiver. The arrangement with that comes the Tesla transmission path for scalar waves very close. If the transmitted noise signal hits upon a likewise one in the surroundings then overlapping occurs, which at the receiver causes a change of the displayed value. If doing so a subtraction (extinction) or an addition (amplification) of the signals occurs is unimportant.

With this arrangement in any case statements are possible about frequency, wavelength, velocity of propagation and about the amplitude of the scanned signal, without withdrawing energy and strain the signal in doing so[i]. With such a gauge radiation conditions depending on location could be measured as well as technical devices checked lor the emitted scalar wave parts (fig. 24.1).

An important use over and above that would be given in medical diagnostics. Every living being „produces noise" the technician would say, it „emits scalar waves" I would say, whereas following general usage is talked about the „aura of man". The value of an aura diagnosis still is completely unknown to most doctors and therapists, especially as the scanning of the aura at present only is possible by dowsing. But the patient expects that a doctor works with a technical gauge and not with a dowsing rod!

<i>: First experiments, which I carried out with students in the laboratory, look very promising. Unfortunately the works at present rest for reason of lack of money.

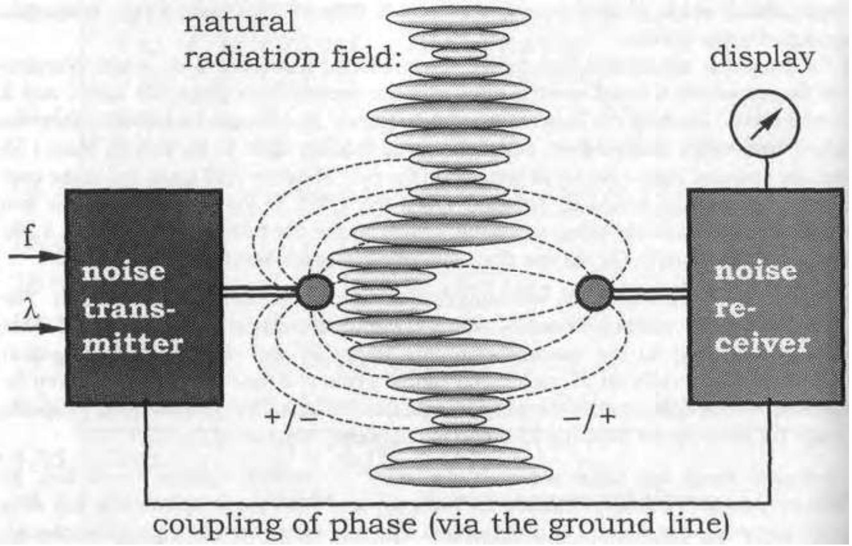
natural
radiation field:
display
f
λ
noise
trans-
mitter
noise
re-
ceiver
+/−
−/+
coupling of phase (via the ground line)

Fig. 24.1:____ Concept for a device to detect scalar waves

24. Medicine of oscillation

24.1 Mediation of information with the help of scalar waves

If for the discussed scalar wave gauge the transmitted comparison signal in addition deviates from the form of a sinus, if the signal is modulated in some way, then even the transmission of information by means of the noise as scalar wave can be realized. If we as an example again take the aura of man, which is more than only a radiation field, it carries information for instance about the state of health of man. Kirlian photography is one form of making it visible (Fig. 3.6).

Not only the nerve conduction and the brain work with scalar waves. Man in this way also corresponds with his fellow people, some more, others less. There are people who „beam" and that should be interpreted in the true meaning of the word. Others opposite to that are more „receiver characters", who pick up more scalar wave radiation than they give away. But because they not only pick up positive radiation energy, they are susceptible to information, which makes sick. That's why a healthy equilibrium in the exchange of thoughts, feelings and the different forms of scalar waves is very important. All that is reflected in the aura of man. If we technically are capable to scan the aura of man, than this can be a blessing for humanity and for the public health.

The possible abuse however must not be overlooked. Unbiased research results, which can be used for the benefit of humanity, in many cases just as well can be used to harm humanity.

A connector of both these worlds is the Russian psychologist Prof. Smirnov, who has shown publicly in television, how a spoken sentence can be transformed in a noise signal, which is taken up directly by the brain as information, with the help of a computer program <i>. He with that is capable to "sodden" individuals, as he says, he in this way can take the fear away from soldiers before a combat mission and can operate the disease out of drug addicts without bloodshed. In the television film such an „operation at the open subconscious"' is shown life. The patient hears the noise signal over headphones and is cured already after few minutes treatment time.

We here are getting in the domain of ethical and moral problems of scalar wave research, which aren't solved by us looking away and leaving the field up to others. According to the words of Prof. Smirnov the only thing, which can stop the research scientist, is his own moral. He doesn't say any more and that isn't exactly reassuring! One here is working with „signals resembling sound", is said in the report of the Zweite Deutsche Fernsehen<i>, thus with longitudinal waves and that shows, that already more knowledge about scalar waves is present than is generally known by the masses.

Another way is leading over the bioresonance. Also here at first the possibility and chance to cure diseases is to the fore. The bioresonance is a central aid in the area of the medicine of oscillation, which is increasing in importance permanently.

<i>: Die Zombies der roten Zaren (Besta Film Warschau im Auftrag d. ZDF 1998)

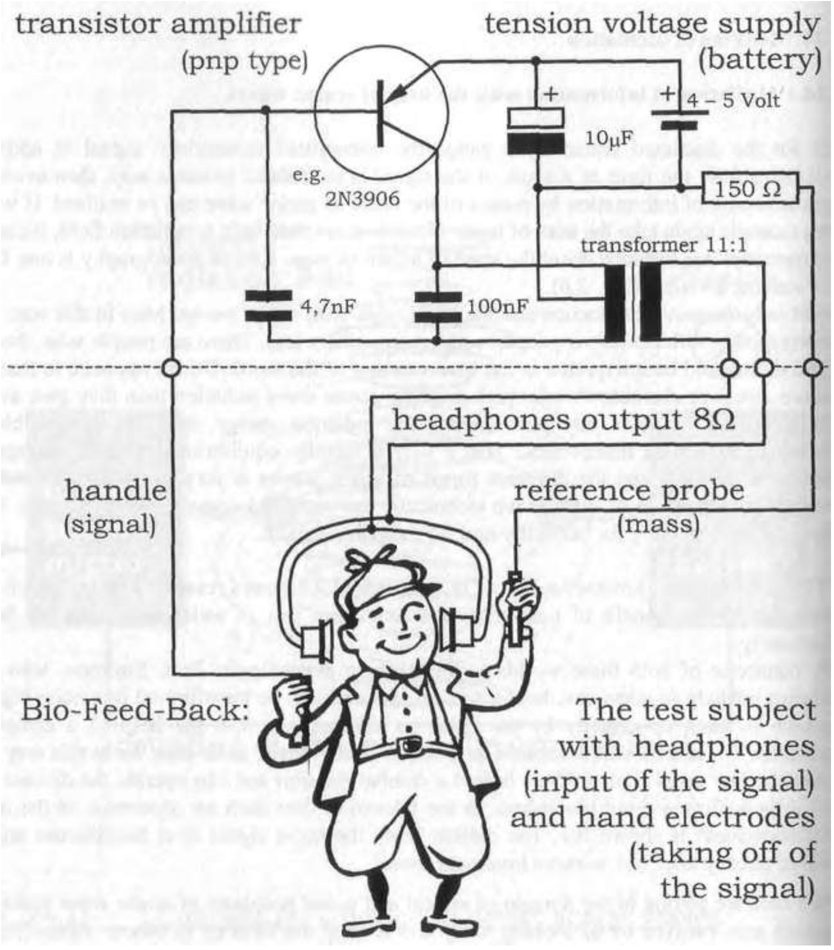
transistor amplifier
(pnp type)
e.g.
2N3906
tension voltage supply
(battery)
+
+
4 - 5 Volt
10µF
150 Ω
transformer 11:1
4,7nF
100nF
headphones output 8Ω
handle
(signal)
reference probe
(mass)
Bio-Feed-Back:
The test subject
with headphones
(input of the signal)
and hand electrodes
(taking off of
the signal)

Fig. 24.2 A: Wiring diagram of the Syncrometer.<i>

<i>: Wiring diagram taken from: H. Clark: The Cure for All Cancers (1993) New Century Press, USA, Page 434

<ii>: Flanagan: Horen mit der Haut anstatt mit den Ohren, Sonnenwind 1/97, S.13-17.
s.a.: N.Begich: Towards A New Alchemy, The Millenium Science, Earthpulse Press 1996

<iii>: Dr. Bodo Kohler: Biophysikalische Informations-Therapie, Gustav Fischer Verlag (1997), Kap. 11.6 Der individuelle Grundton, S. 239 ff.
s.a.: Dr. Bodo Kohler: Die bipolare Farb-Ton-Therapie, CO'Med 2/2000, S.10 - 15

24.2 Systems with positive feedback

The field of oscillation of man easily can be influenced, because it concerns a vortex field. A technical measurement with the gauges available at present is almost impossible and highly inaccurate. There however exists the possibility, to apply electrodes to a person and to integrate him in the circuit of a technical device. The two of them form a feedback system with man in the return loop. The operation can take place in two different ways, depending on the sign of the feedback (Fig. 24.3):

For positive sign it concerns positive coupling. In this case the signals released by a person sum up, for which reason already minimum amplitudes are sufficient to produce violent reactions in the case of resonance. For resonance to occur, the system must either search independently the suitable frequency and phase, or the therapist searches the points of resonance.

24.2.1 The Syncrometer.

An independently working system is e.g. the Syncrometer. By means of two electrodes, which are held with the hands or are attached directly to the head, the noise signal emitted by a person is called off and supplied to a broadband amplifier. The test subject again hears the amplified signal via headphones, so that the circuit is closed.

Apparently unspecific and still reproducible during operation certain signals capable of resonance are emerging stronger by amplifying and swinging themselves up. The aura starts to extend or expressed in the words of a radiation sensitive, the so-called reaction distance is increasing, with which he means the distance in which he detects a dowsing rod reaction.

The test subject now increases his radiation, so this method can be valued. By strengthening his own radiation power, he works more as a transmitter and less as a receiver for scalar waves. In the last point the therapeutic use seems to iie, because by that the patient can get rid of resonances to any unpleasant persons or to technical transmitters, which are burdening him. The device consequently also is sold as aid or protection against electrosmog.

The amplified signal also can be picked up directly by the skin instead of by the headphones and still being „heard". The reports stretch from a harmonizing and balancing effect up to "electronic telepathy" and states similar to ecstasy of the test subject<i>.

24.2.2 The sound therapy.

The controlled variant of this method for instance is the sound therapy. Now the sounds, which the patient picks up by means of headphones, do not stem from him but from a sound generator. The therapist goes through the scale and tries to find out at which sound the aura swings up. If the eigenfrequency is found, then the patient can therapize himself, by again and again playing or humming his eigentone<iii>.

It concerns here only a physical statement on the used methods of alternative medicine and not therefore, whether healing successes are actually possible.

sound frequency equally tempered [Hz]		wavelength [cm]	colour	wavelength [run]
F sharp"	740	46,4	(purple)	-
F"	698	49,1	(crimson)	-
E"	659	52,0	(violet)	-
D sharp"	622	55,1	indigo	421
D"	587	58,4	bleu	446
C sharp"	554	61,9	turquoise	472
C"	523	65,6	green	500
H'	494	69,5	yellowgreen	530
A sharp'	466	73,6	yellow	562
A'	440	78,0	orange	595
G sharp'	415	82,6	orangered	630
G'	392	87,5	red	668

Table 24.2 B: Scale of the colour range.<i>

colour		chakra	sense organ	planet type
purple	7.	(crown chakra)	epiphysis	Jupiter
indigo	6.	(brow chakra)	hypophysis	Saturn
bleu	5.	(throat chakra)	hearing	Mars
green	4.	(heart chakra)	feeling	Sun
yellow	3.	(solar plexus)	seeing	
orange	2.	(sacral chakra)	taste	Venus
red	1.	(root chakra)	smell	Moon

Table 24.2 C: Assignment of colours to chakras and planets.<i>

colour		metals	gems (selection):
purple	7.	tin	amethyst, fluorite
indigo	6.	lead	indigo-sapphire, azurite
bleu	5.	iron	sapphire, lapis lazuli
green	4.	gold	emerald, malachite, jade, gr. tourmalin
yellow	3.		topaz, amber, citrine
orange	2.	copper	carnelian, fire opal
Rot	1.	silver	ruby, coral, garnet, red jasper

Table 24.2 D: Assignment to a selection of gems.<i>

<i>: Dr. Bodo Kohler: Biophysikalische Informations-Therapie, Gustav Fischer Verlag (1997), Kap. 11.6 Der individuelle Grundton, S. 190 - 194

24.2.3 The colour therapy.

Analogous to acoustics also colours each time can be assigned a wavelength and also here can be detected that every person responds to certain colours, thus wavelengths, in particular which can stimulate his aura to resonant oscillations. In practice sound and colour therapy often are used coupled<i>, even succeeds more or less an assignment to each other. That in last consequence leads to a "scale of the colour range" (table 24.2 B)<ii>. According to statement of the treating doctors does a patient, who responds to a certain colour, also react to the corresponding sounds and vice versa.

24.2.4 The aroma therapy.

If according to that it depends less on the circumstance if sounds or colours are used, but the wavelength of an oscillation is crucial, then a stimulation also should be possible by means of the remaining sense organs, e.g. the nose, the tongue or the skin. The smell after all already could be identified as vortex information (fig. 9.0). Fragrances of natural essences can significantly influence the frame of mind. But as long, as we still haven't understood the physics behind it, we technically hardly are capable to generate equivalent vortex modulations artificially.

24.2.5 The gem therapy.

Already the holy Hildegard von Bingen (1098-1179) knew and used the beneficial effect of gems. Physical background of this at first purely empirical form of therapy is the characteristic eigenfrequencies of the gems, which are picked up as stimulation over the skin. Because gems represent a mixture of various molecules, the oscillations in the atomic hull will overlap, so that overlapping and beat frequencies can form with wavelengths in biological relevant areas. There even can form modulations, which are carrying information. The effectiveness again is linked with the resonance condition, which must be fulfilled between the gem and his carrier.

24.2.6 Meditation and self-therapy.

There even exist people, who don't need any technical aid at all to get rid of unpleasant resonances. Some meditate and go into resonance with themselves, whereas others prefer to love a person, to whom they feel „attracted" or with whom they are „on the same wavelength", which means as much that they go into resonance with this person. In the Catholic Church for instance the priests are not allowed to get married, because they should be in a resonance with the church and with God.

<i>:: Dr. Bodo Kohler: Die Coloroma-Therapie, Der aussergewohnliche Einsatz von Aromen, Co'Med 4/2000, S. 48-52

<ii>: Cousto: Die kosmische Oktave, Synthesis Verlag Essen 1984

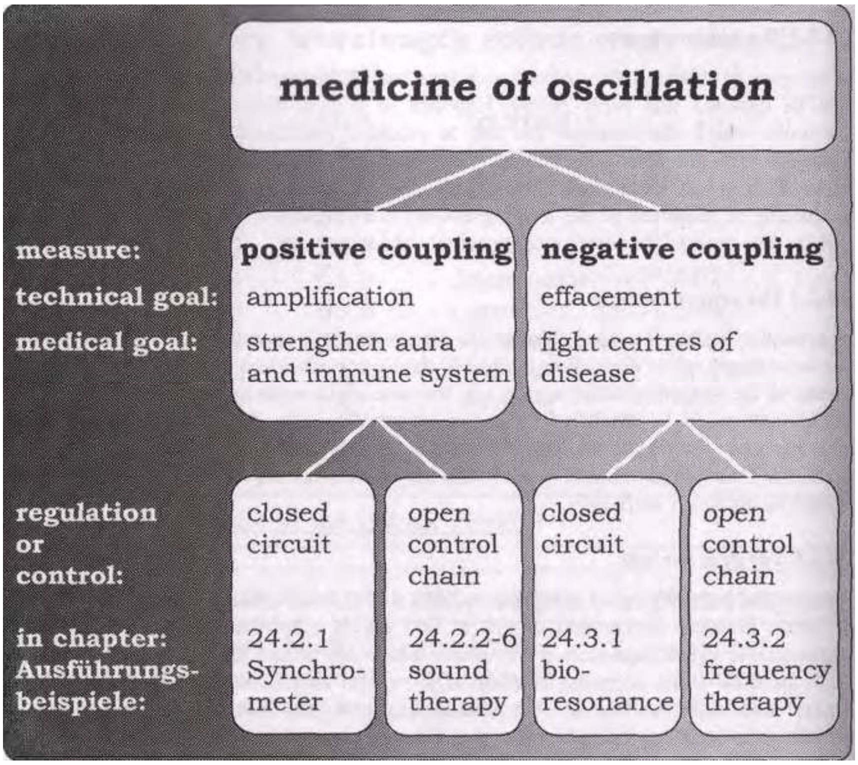
medicine of oscillation
measure:
technical goal:
medical goal:
positive coupling
amplification
strengthen aura and immune system
negative coupling
effacement
fight centres of disease
regulation or control:
in chapter:
Ausführungs-beispiele:
closed circuit
24.2.1
Synchro-meter
open control chain
24.2.2-6
sound therapy
closed circuit
24.3.1
bio-resonance
open control chain
24.3.2
frequency therapy

Fig. 24.3 A: Structure concerning the medicine of oscillation

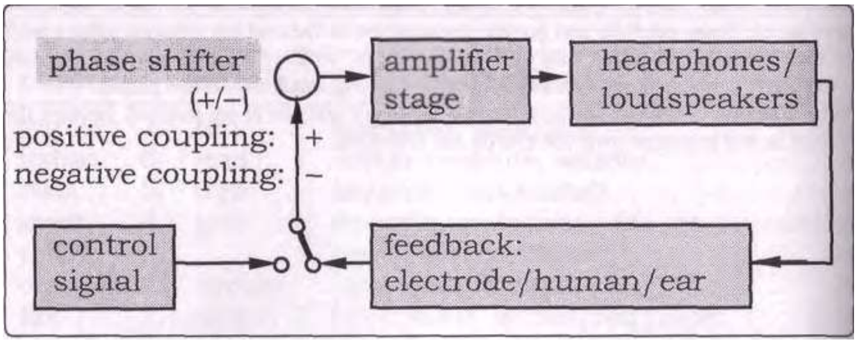
phase shifter
(+/–)
positive coupling: +
negative coupling: –
amplifier stage
headphones/ loudspeakers
control signal
feedback: electrode/human/ear

Fig. 24.3 B: Regulation or control?
Positive or negative coupling?

24.3 Systems with negative feedback

For negative sign it concerns negative coupling. In this case the signals are subtracted. The signals released by a person are played back to him with opposite phase. This method lies in the domain of the bioresonance.

24.3.1 The homeopathy

The concept, which is pursued by the medicine of oscillation in the case of the bioresonance, is that harmful or ill making oscillations should be effaced. Not the direct, but rather the indirect strengthening of the immune system is to the fore, which should be reached by relieving it. The approach with that pursues the same goal as the pill medicine, where oscillations, which should produce a wanted effect and which are bound by means of substances are supplied to the body.

The disadvantage of the administration of pills is the often toxic effect of the carrier substance. The German doctor Samuel Hahnemann (1755-1843) has demonstrated a way, in which to make the helpful and important information for the body to go over from the carrier substance to the water by diluting with water and by shaking. The water molecules now oscillate in time of the carrier molecules. Interestingly in doing so not even the information gets lost, the modulation thus is preserved. By constantly shaking there even occurs an amplification or a „potentiation", as Hahnemann expressed himself, because now every carrier molecule reaches and modulates countless water molecules. Thus the amount of water informed is bigger than the amount of the original chemical substrate. In the case of homeopathy like is treated with like. If for instance a poison causes certain complaints at a healthy person, then in "homeopathic" administration this information helps a sick person with similar complaints. There thus occurs a disturbance activation with opposite phase, as the engineer uses to express himself. Despite a widespread scepticism the method of homeopathy indeed seems to function, and it has stood the test in countless cases.

24.3.2 The bioresonance

A technical realization is represented by the bioresonance. For that endogenous oscillations are called off by means of an ECG (electrocardiogram), an EEG (electroencephalogram) or a MEG (magnetoencephalogram) at the surface of the skin. The technical device then shifts the phase for 180 degrees and amplifies the signal to the extent that pathological frequencies are extinguished for reversely directed input. This very reasonable theoretical concept in practice of course only is as efficient, as the empirical determined pathological frequencies are the cause of a disease and not only represent an unimportant symptom as side effect.

To that technical problems are joining. Prof. Heine blames the constantly changing reaction diversity and the thermal noise for the circumstance that the „frequency spectrum permanently is fluctuating", as he writes if*. That hardly can realize the necessary phase inversion. We meanwhile know that vortices have a fluctuating frequency spectrum, that in the case of biosignals it concerns such field vortices, which result in a noise signal unspecific in frequency.

With this knowledge we should be able to significantly improve bioresonance procedures, and even the inversion of phase shouldn't represent an insurmountable difficulty anymore.

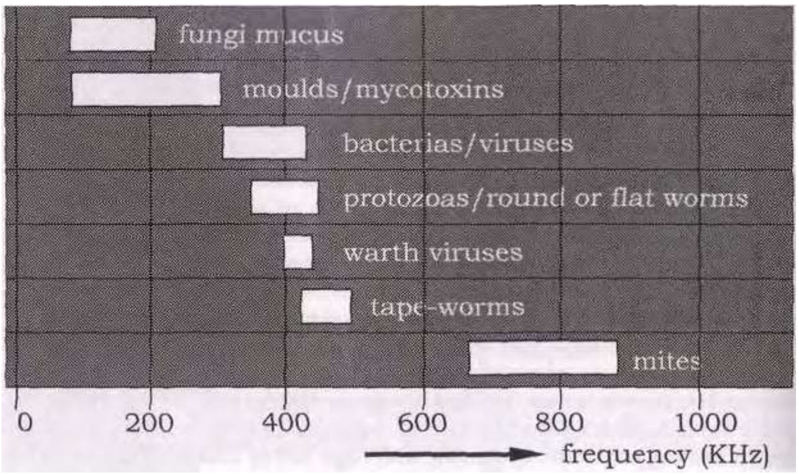

Fig. 24.3 B: Frequency spectrum of parasites according to Clark.<i>

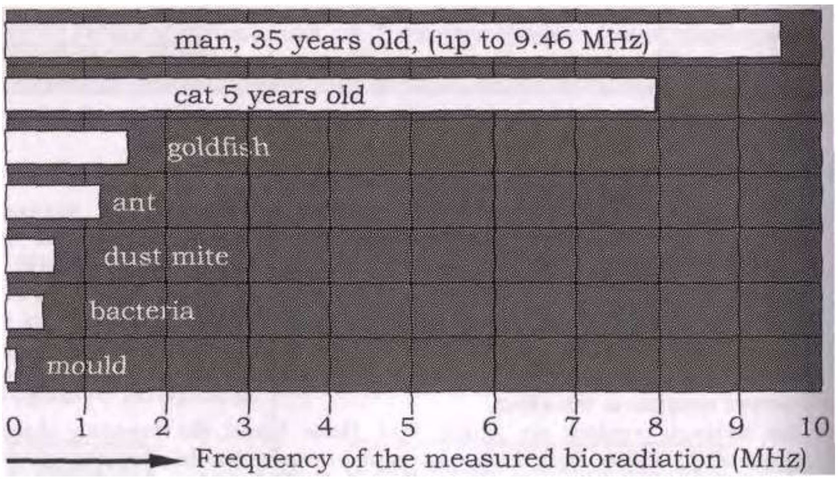

Fig. 24.3 C: Diagram of frequency ranges for some living beings.<i>

<i>: Dr. Hulda Clark: The Cure For All Diseases (1995), Page 604-643

A further problem, which Heine addresses, can't be denied<ii>: „For communicative molecule oscillations the microwave range of 1 GHz up to above 10000 GHz is crucial. A calling off of endogenous interference oscillations in this range with the help of normally used electrodes is not possible"<iii>. It thus could happen that essential frequencies, which are directly related to a disease, aren't recorded at all and as a consequence also not treated. There even is the risk of informations being brought in, which additionally stress the immune system instead of relieving it, that the patients after a treatment are worse off than before. The doctor or therapist is burdening himself with a big responsibility, when, how and at whom he applies methods of bioresonance or if he better does without them. For the mentioned reasons the method in the therapy only plays a secondary role. The bioresonance on the other hand is quite often and with great success used in the diagnosis (Nosoden), if it for instance concerns the determination of an incompatibility or an allergy - but that is a completely different theme. For the bioresonance the transition of diagnosis to therapy is however floating.

24.3.3 The frequency therapy

If using a frequency therapy the problems are standing similarly. If we separate the closed and negatively fed back circuit of a bioresonance and form an open control chain consisting of a technical control device and the patient, then we get the structure, as it is put to use in a frequency therapy. The goal still is the same: parasites or pathogenes, which stress and burden the immune system should be fought.

But a disadvantage of every open control chain is that the treatment at first occurs blind due to the missing feedback.

The natural healer Dr. Hulda Clark has examined as support various pathogenes and parasites under the microscope, while she has varied the frequency and at the same time applied a low-tension voltage. Doing so she could observe the dying of bacterias and parasites at certain frequencies. Correspondingly she publishes tables, in which the in each case „mortal" frequencies are listed<i>.

Hut without being able to verify the success of a treatment in the living organism, she proceeds from the assumption that by applying a low tension voltage (5-10 Volt), as it is produced at the output of a commercially available frequency generator, if a sinusoidal signal is set with the appropriate frequency exactly the associated parasites and leeches will be destroyed. Doing so a fixed rhythm of pauses and treatment times must be adhered to (7 min on, 20 min off, 7 min on, 20 min off and 7 min on).

The doctors and therapists treating with this particularly inexpensive method report amazing results, inexplicable spontaneous cancer cures, like of HIV-positive patients, who after the treatment were tested HIV-negative. But also the inverse case already should have occurred, that a HIV-negative patient afterwards was HIV-positive! Here too clear the limits of this method appear, which in practice unfortunately turns out to be relatively unspecific.

<ii>: Prof. Dr. H. Heine: Grundregulation und Extrazellulare Matrix - Grundlagen und Systematik, Lehrbuch der biologischen Medizin, (1991) 2.Aufl. Hippokrates Verlag Stuttgart, Seite 63

<iii>: U. Warnke: Bioresonanztherapie - Wunsch und Wirklichkeit. Medizin transparent Nr.2 (1996) S. 36-37

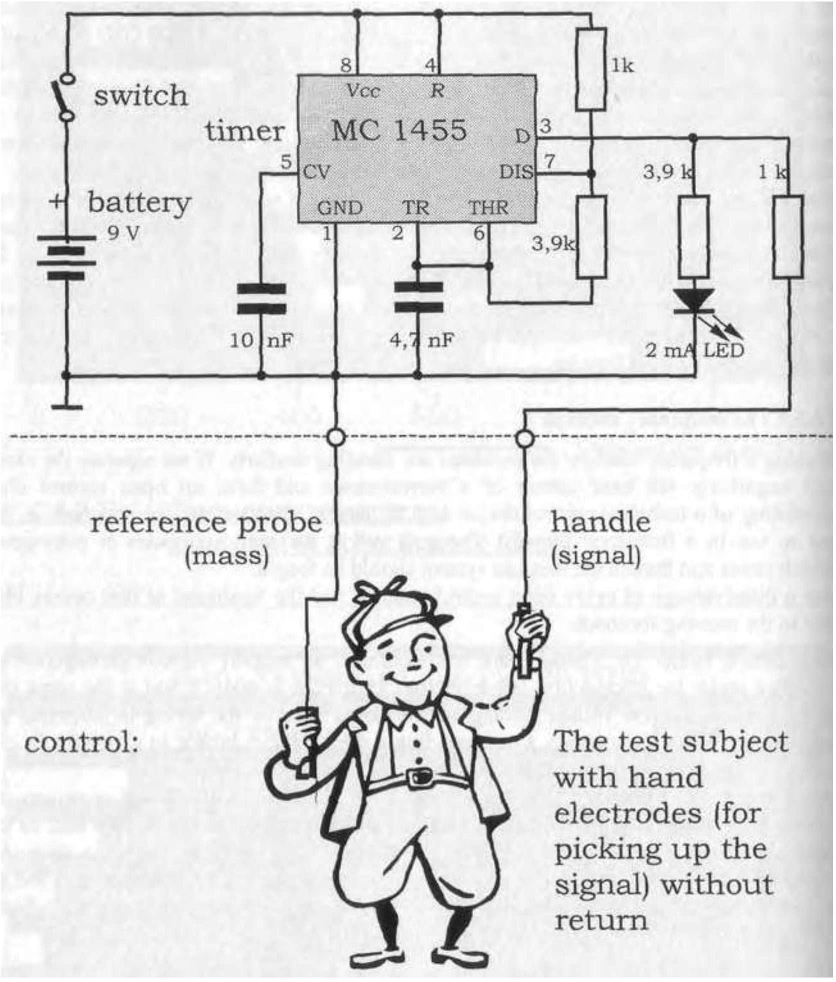
8
4
Vcc
R
1k
switch
timer
MC 1455
D
3
5
CV
DIS
7
3,9 k
1 k
+ battery
9 V
GND
TR
THR
1
2
6
3,9k
10 nF
4,7 nF
2 mA LED
reference probe
(mass)
handle
(signal)
control:
The test subject with hand electrodes (for picking up the signal) without return

Fig. 24.4: Wiring diagram of the Zapper.<i>

Wiring diagram taken from:
<i>: H. Clark: The Cure for All Cancers (1993) New Century Press, USA, Page 507, resp.:
Dr. Hulda Clark: The Cure For All Diseases (1995), Page 48

24.4 The Zapper

From the point of view of physics of today the frequency therapy actually shouldn't be able to function at all. The electromagnetic waves penetrate only few millimeters into the skin at the used frequencies and wouldn't have the slightest chance, to reach a parasite, which is staying somewhere in the body. But it is said that it is possible to cure athlete's foot, by the patient taking the electrodes in his hands. How, we ask the question, does the signal of the function generator know where it should go?

It quite obviously concerns a resonance phenomenon. The likewise emitted scalar wave part tunnels undamped at those places in the body, with which it can build up a resonance and that for proper tuning are the unloved parasites. The scalar wave radiation is bundling up at the resonant receiver, so that despite the low transmission power as a consequence of the bundling up the energy density for the parasite becomes very high. It as a result is destroyed by its own ability to go into resonance. Once it is killed, the next one goes into resonance, logs off as well etc. In this way the parasites are destroyed one after another and not all at once. That's why the specified treatment cycle makes sense.

The copper electrodes should not be taken in the hands directly, Dr. Clark recommends, but before be wrapped with moist paper. By means of this insulating layer, so is my interpretation, the conventional wave part, for which the skin functions as a wave guide, is reduced whereas the desired scalar wave part is increased. Such measures crucially contribute to the success of a therapy method, even if they were determined purely empirical.

If one wants to address every possible parasite individually, then the treatment takes correspondingly long. If one on the other hand sends all relevant frequencies at once by overlapping them, then the treatment can be abbreviated to the duration of one session. If the therapist goes still further and replaces the sinusoidal signal by a rectangular signal, then infinitely much sinus functions are hidden in it, as a Fourier analysis shows. With a rectangular signal, as it is delivered by the Zapper, one as it were catches everything, Good as well as Bad. There the helpful intestine bacterias break exactly like the wrongdoers.

The treatment with the Zapper is simple, inexpensive and exactly as controversial. It is the shot with the shotgun in the forest. One always hits something. We nevertheless must ask the question, why one only hits parasites and bacterias and not the vital organs? Aren't those damaged also?

Now then, the signal of the function generator is not modulated; it doesn't carry information. That's why only monocellular parasites, which don't know information exchange, are capable of a resonance. Human cells and more than ever whole organs on the other hand work with complex modulations, which effectively prevent any formation of resonance with the technically generated basic wave, with which the question would be answered so far.

That however also means that immune reactions can be expected: If the first treatment with the frequency therapy still is successful and all simple parasites could be hit, then only further evolved parasites have been spared, which modulate their information. They now breed and can't be reached anymore in further sessions. The method suddenly doesn't function anymore, the therapist finds out, the body apparently has become immune.

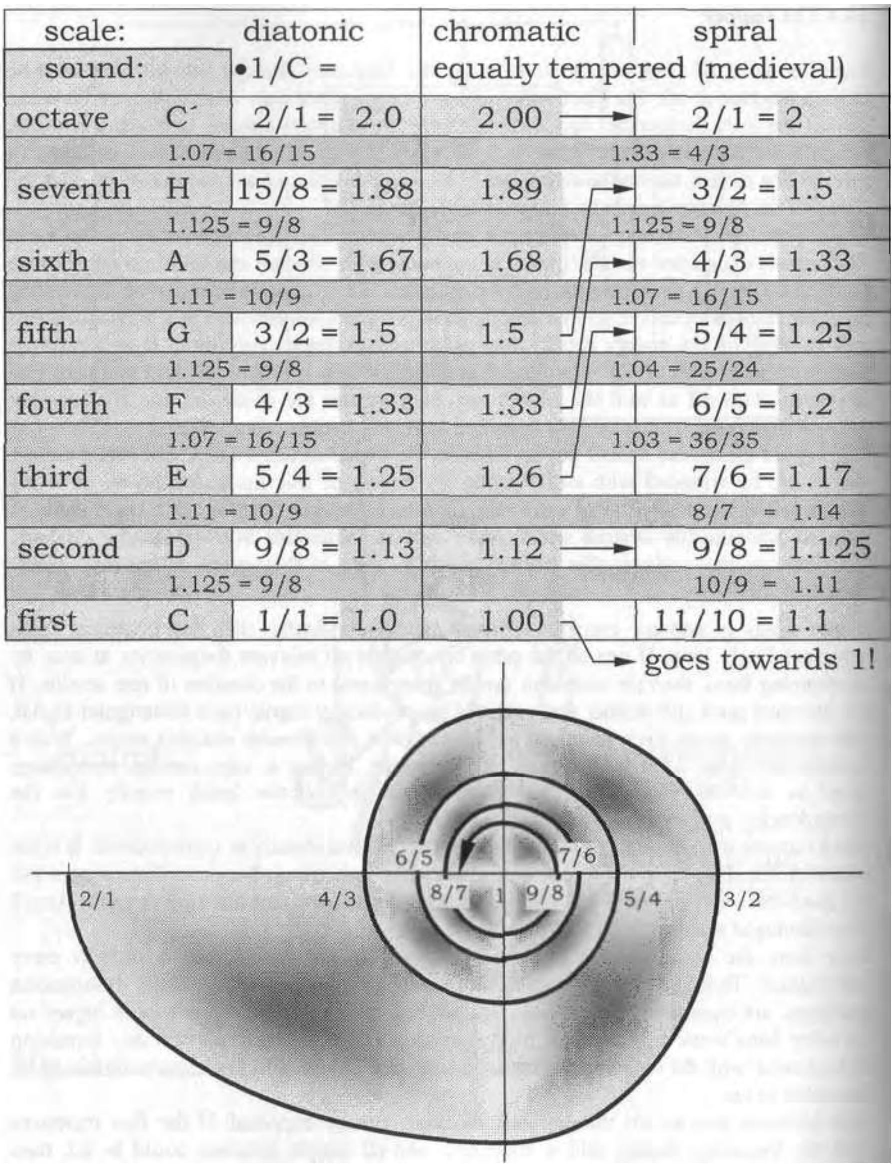

scale:		diatonic	chromatic	spiral
sound:		•1/C =	equally tempered	(medieval)
octave	C´	2/1 = 2.0	2.00	2/1 = 2
		1.07 = 16/15		1.33 = 4/3
seventh	H	15/8 = 1.88	1.89	3/2 = 1.5
		1.125 = 9/8		1.125 = 9/8
sixth	A	5/3 = 1.67	1.68	4/3 = 1.33
		1.11 = 10/9		1.07 = 16/15
fifth	G	3/2 = 1.5	1.5	5/4 = 1.25
		1.125 = 9/8		1.04 = 25/24
fourth	F	4/3 = 1.33	1.33	6/5 = 1.2
		1.07 = 16/15		1.03 = 36/35
third	E	5/4 = 1.25	1.26	7/6 = 1.17
		1.11 = 10/9		8/7 = 1.14
second	D	9/8 = 1.13	1.12	9/8 = 1.125
		1.125 = 9/8		10/9 = 1.11
first	C	1/1 = 1.0	1.00	11/10 = 1.1

Fig. 24.5: Harmonic spiral built on the first Gregorian scale.<i>

<i>: Dr. Manfred Doepp: Naturgemasse Frequenztherapien, Die Harmonie der Snirale. CO'Med 10/2000. S. 46 - 49

24.5 Discussion concerning the medicine of oscillation

The textbook medicine in its explanations and treatment methods is basing on the models, which it can measure and analyse and which it understands. Doing so man and whole nature is reduced to a handful of chemical reaction formulas. The whole pharmaceutic industry lives on this misleading path, which long ago has revealed to be a dead end, medical as well as financial. This health service cannot be paid anymore and we should ask the question if it actually is worth the money, if with electric signals of minimum power can be obtained effects comparable to the effects of the pill medicine. We need a new medicine, a potential vortex medicine. First of all we should research how an organism covers its energy needs and how it communicates. There leads no way past the scalar waves and the newly discovered potential vortices. Chemical processes as they are being observed, occur by the way, that is beyond doubt, but they by no means are the cause. Hence of pills and other chemical means at most a treatment of symptoms and a case of side effects may be expected but not a cure of a disease. Once the potential vortex medicine will be systematically explored and be put to use in practice, healing successes can be expected, which we at present can't imagine at all.
The amazing results, which already today are obtained in the medicine of oscillation and of which some doctors can report<i>, dictates the direction in which the textbook medicine should develop. In the question, which kind of oscillation or which ,,sequence of sounds" (Fig. 24.5) is the right one, still exists considerable need for research. Some doctors even already work with my new theory and cite whole passages from my publications until now about this theme<ii>.

$<i_a>$: Dr. Bodo Kohler: BITsyn, der neue Weg in der Informations-Medizin, Jubilaums-Schrift der Internationalen Arzte-Gesellschaft fur Biophysikalische Informations-Therapie, Bad Nauheim den 6.-8.10.2000, Seite 48 - 56.

$<i_b>$: Bioenergetische Messverfahren in Theorie und Praxis, Vortragsband der Gesellschaft fur Energetische und Informationsmedizin e.V., Universitat Stuttgart am 17.7.1999.

$<ii_a>$: Dr. Johann Lechner: Storfelddiagnostik, Medikamenten- und Materialtest, Teil 2 aus der Reihe: Praxis der Ganzheitlichen Medizin und Zahnmedizin, Verlag Dr. E. Wuhr (2000), Kap. 2.4.2 Beruhrungslose skalarwellentragende Informationsubertragung S. 173 ff., bes. Kap. 2.4.2.3 Seite 175, 176.

$<ii_b>$: Dr. Reichwein, Peters: Zellulare Elektromagnetische Systemsteuerung, Der Freie Arzt 5 (2000) im Wissenschafts-Forum (Anhang, S. IV - XXIII).

$<ii_c>$: Dr. M. Doepp: Tesla-Wellen, Neue Studien, CO'Med 5/2000, S. 94 - 95

Frequency spectrum of the human organism:

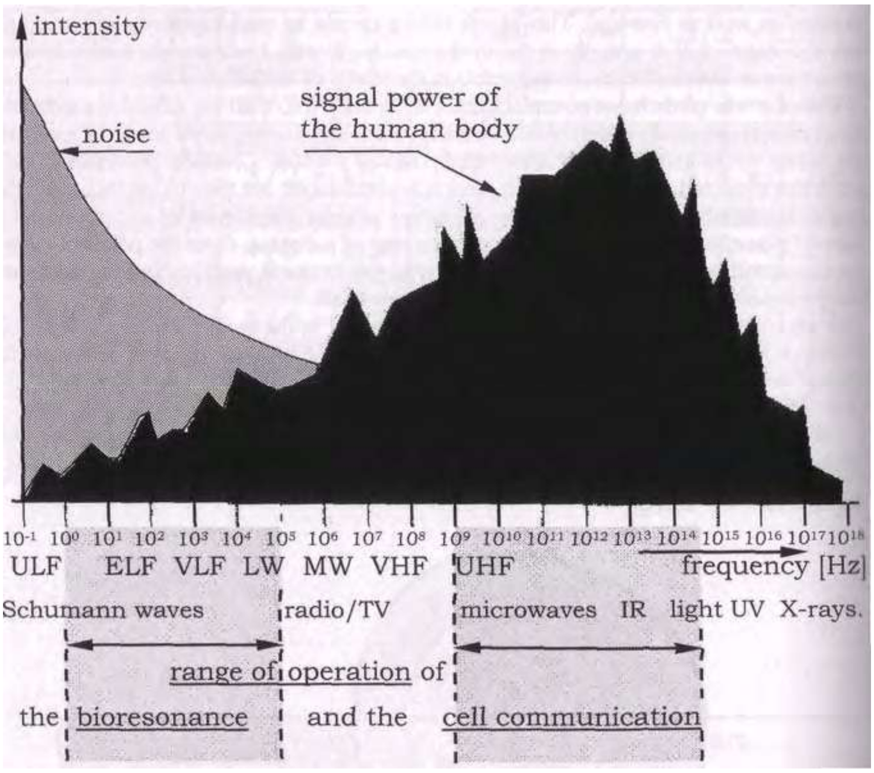

Fig. 25.1: The measured frequency spectrum of man.[i]

---

[i]: Dr. P. Bembenek: Akupunktur und (Bio-) Resonanz, CO'med 6/98, S. 50-58

25. Contributions to the discussion concerning the information technical use

The medical research predominantly takes place with statistical methods. This might be changed fundamentally, if only the physical relations have been realized and their causes found out. Only then a basic research will be possible also in the area of medicine, as it today already is usual in other disciplines.

In this chapter it concerns the question of the medical basis in general, and specifically the question of the use of the newly discovered potential vortices in biological systems. Doing so very aware is attempted to put the normally used medical view at the back and to derive the answers from the perspective of the engineer and the physicist. This approach is unfamiliar, so that some textbook physicians will have difficulties to follow. It however will prove to be very helpful. Medicine in addition needs new impulses and that justifies a new approach, like the unconventional one proposed by me.

In this connection gladly is fallen back upon the comparison of biological with technical systems, although here exist differences in principle and fundamental It can be expected that nature has developed very appropriate strategies to solve certain tasks, as it also is teached in engineering and used in practice. At this place not only medicine benefits from new insights, engineering vice versa also can learn of nature, because natures optimization strategies are much much older and correspondingly perfect.

25.1 Window of the sensuous perception

To clarify the inner processes in man we consider his sense organs, which function as interface to the outside world. Technically seen in the case of the sense organs it concerns measurement converters to record and process certain physical values in our environment and their adaptation to pass on the information by means of the nerves. The measurement converters thereby are built more or less complex, which is related to the circumstance that the measurement factors, which a living being needs to orient itself and to be able to survive, are quite different.

It concerns little windows within the frequency spectrum shown in fig. 23.1, which should be analysed. Man for instance chooses the visible frequency window and interpretes it with an assignment to the colours, because in this range the sunlight can reach the surface of the earth and the ionosphere doesn't exert an excessive damping (fig. 2.8).

The acoustic window depends on the sound propagation in air, whereas dolphins work with ultrasound, which carry considerably further in water. The associated measurement converters, the ear and the eye, have a very complex structure. From this can be inferred that here the received signal must be transformed into a different physical measurement signal, that the perceivable signals are of other nature than the signals, which our nerves are able to pass on.

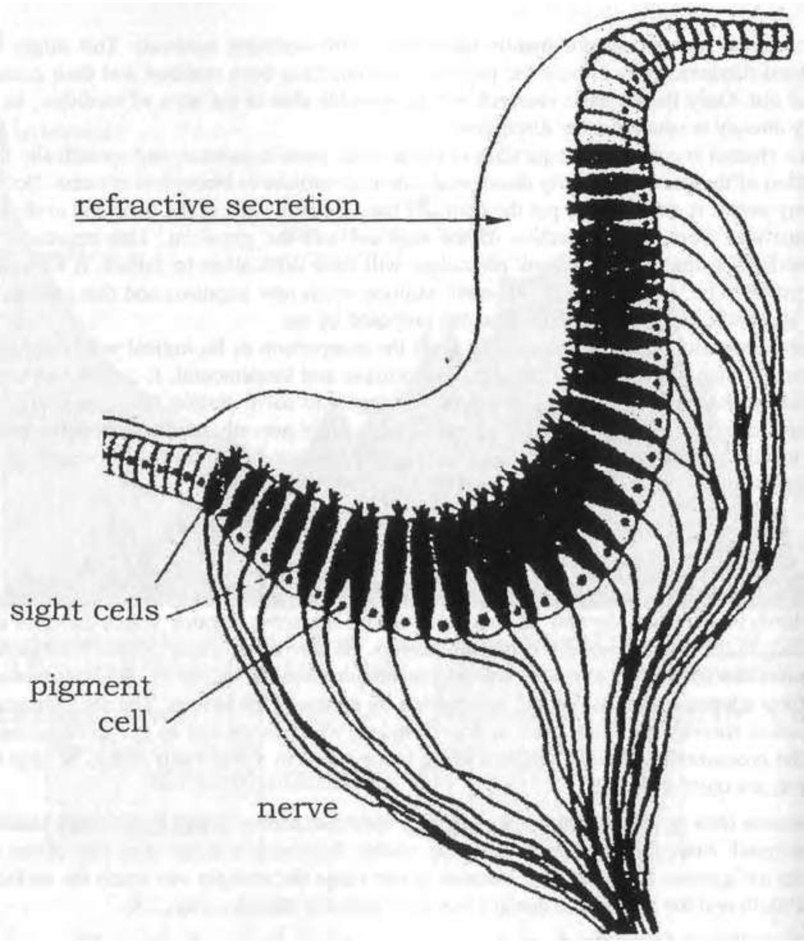
refractive secretion
sight cells
pigment cell
nerve

Fig. 25.2: ____ The hollow eye of the snail

25.2 The sense organs (nose, eye, tongue)

If we want to find out something about the nature of the nerve signals, then we should more detailed consider a sense organ, which is constructed if possible simple. In that case it is possible that no signal transformation is necessary at all anymore, because the perceivable variable is exactly of the kind the nerves can process.

The primary and perhaps most ancient sense organ of the living beings is the nose. It is constructed extremely simple. On the one hand according to encyclopaedias it is a chemical sense for the perception of smell materials. Man on the other hand can distinguish several thousands of scents. For that a huge analysis factory with correspondingly much receptors would be necessary, which one searches in vain in the nose between the nasal hairs and the nerves. We from that infer that scent should consist of particles, which are to a large extent identical with the reaction potentials. It thus will concern potential vortices. These ring-like vortices of the electric field again are modulated and carry information, which our brain is able to analyse.

According to the derivation in chapter 9.0 Plato already 2400 years ago knew such relations. The conclusion is obvious that smell is vortex information, which according to the explanation of Plato forms at the transition of waves in potential vortices and vice versa. At the natural philosophers in antiquity the circle of the insights is closed. Modern science just isn't that far.

The hollow eye of the snail functions both as nose and as eye (fig. 25.2). This very ancient combined organ points to a relatively close affinity between both measurement sensors. Possibly in the course of evolution both organs have developed from such a common original form.

In the case of man the smell rods occupied with the nasal hairs of roughly 2 micrometer in length, actually remember of the rods and cones in the eye, with which photons are collected. In both cases, so we can explain the mode of action in accordance with model, the ring-like vortices settle on the rods and run as reaction potentials along the nerve to the brain, which then analyses the modulation.

In the case of the highly developed human eye the light rays first have to traverse the vitreous body and afterwards a pigment layer, before they reach the rods and cones. In this way the electromagnetic wave parts must roll up to vortices, since our nerves can only pass on vortices. The photons can be interpreted as corresponding vortices.

If we compare the eye with a bubble chamber, in which the tracks of ionized particles can be observed and photographed as thin white fog stripes. Here in collision experiments it can be proven that it concerns particles carrying impulse and not waves. But the measurement setup by no means does answer the question if it already concerned particles before they entered the bubble chamber, or if these have formed spontaneously in the presence of the saturated vapour.

It would be obvious, if for our eye the chamber water being under the inner eye pressure would take over the function of the bubble chamber and would take care of the rolling up of the light waves to light quanta.

Example tongue:

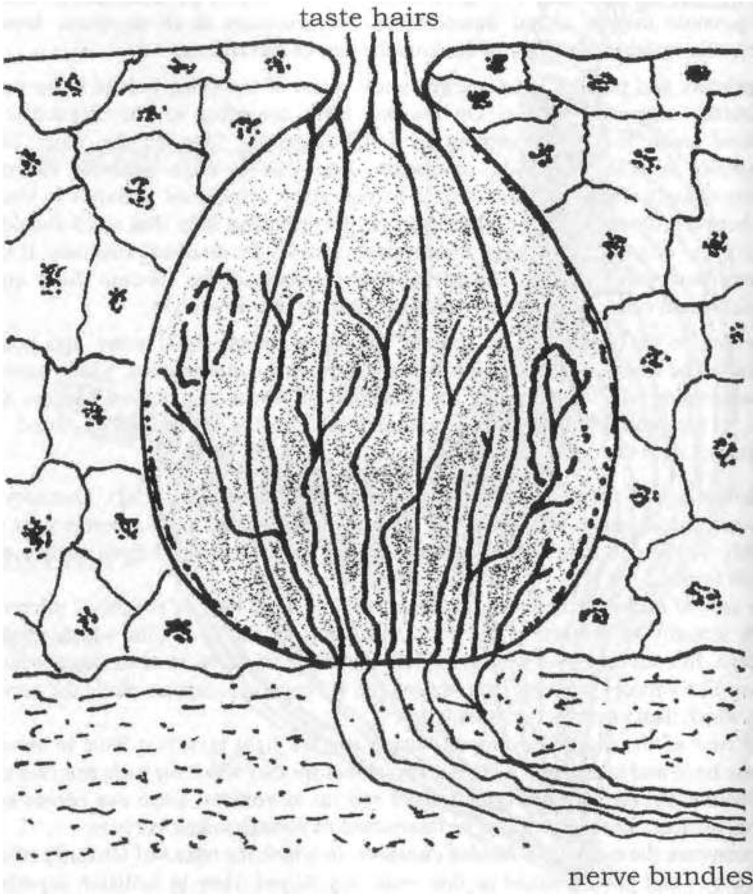
taste hairs
nerve bundles

Fig. 25.3: Taste bud of the human tongue

We optically can't detect a difference, if the light while passing through a thick glass plate increases its part of photons. At the latest in the eye in any case all the light is transformed in photons and usable vortices, so that we mustn't detect a loss of brightness. The glass plate seems transparent to us, even if the sunlight should change its composition and its biological quality, if it passes the plate.

If the tongue while tasking responds to dissolved substances, then it not by all means needs to be a chemical excitation. Instead fine hairs, the taste hairs resp. sense pins, serve as receiver like in the case of the nose. The similar structure of the receptors and the circumstance that for most invertebrates sense of taste and sense of smell can't be distinguished of each other at all and consist of the same primary sense cells, suggest that the tongue doesn't analyse the chemistry as such, but only the molecular oscillation patterns, that also the taste is nothing else but vortex information!

25.3 The nerve conduction

For the collection of the potential vortices and the extraction of their information fine hairs in the sense cells obviously play a central role. They are connected more or less directly with the end of a nerve and pass on the information without big transformation. Even in the organs of equilibrium sense hairs work.

From the comparison with the technique developed by Nikola Tesla I could show that the nerve conduction concerns a single-wire transmission, a kind of waveguide, for which the transport of the excitation information takes place in the insulation layer and not in the conductor itself (fig. 9.6). As proof I quoted that the thickness of the insulation determines the velocity of propagation, that as is well-known the nerve conductors with thick fat layer pass on their action potentials faster than those with thin insulation. Particularly interesting is the observation, how the fat layer is constricted in fixed intervals, like for Wiener sausages (fig. 25.4 A). These nodes of Ranvier prove that only longitudinal waves are being transported, which are standing waves with nodes and antinodes, if the distance from node to node exactly corresponds to the distance from one node to the next. With that nature with the use of the potential vortices is far ahead of our power engineering. The nerve-cables determine with their structure, which signal will be transported and which not!

The technical cables on the other hand are stupid and conduct everything, the useful signal just as well as any arbitrary interference signal. Anyone, whom the computer crashes every few minutes, knows what I'm talking about.

Man isn't able to afford a crash of his think computer. It would be lethal for him. His nerve costume even tolerates short-circuits. The acupuncture is such a short-circuit technique by means of electrically conductive needles. There even is given a therapeutic benefit and a relaxing effect for the body.

Cut through nerve fibers even can partly regenerate again, even without a cut through nerve again growing together with its other end. The nerve conductors are so intelligent, that only the matching information arrives at the end by passing on informations from one fiber to the corresponding next with the same node interval.

A. Nerve conduction with nodes of Ranvier (to transmit standing waves):

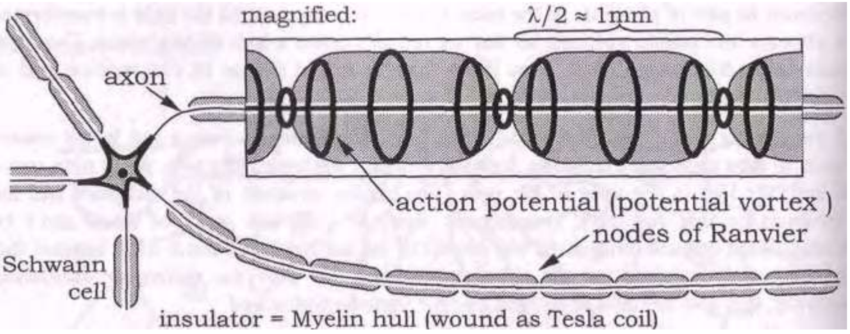
magnified:
$\lambda/2 \approx 1\text{mm}$
axon
action potential (potential vortex)
nodes of Ranvier
Schwann cell
insulator = Myelin hull (wound as Tesla coil)

B. Single-wire transmission according to Nikola Tesla<i>:

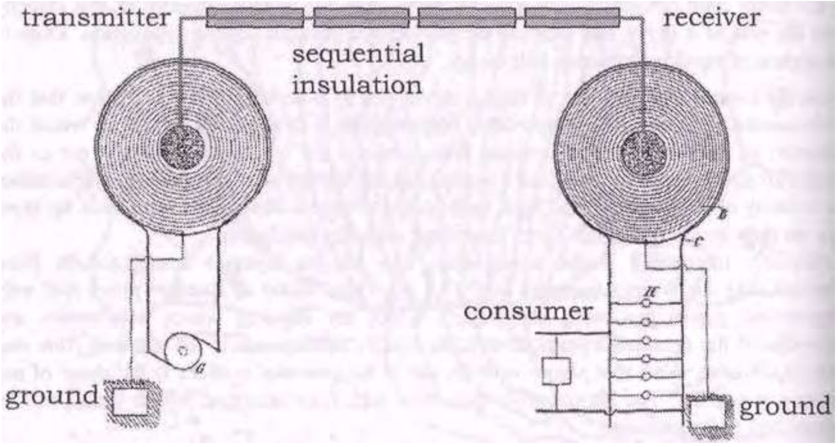
transmitter
sequential insulation
receiver
consumer
ground
ground

Fig. 25.4: Selective scalar wave signal transmission

The brain works with scalar waves! Reasons are:

1. Lacking of a signal transformer between the nerves and the brain.
2. High performance density of the think apparatus.
3. The brain activity measurable from the outside with the EEC
4. Spark formation and corona discharges at open top of the skull.
5. Insulation defect occurring in the case of epileptics.

<i>: Original sketch of Tesla is situated in the Nikola Tesla Museum, Belgrade.

Nerves represent an intelligent and at the same time interference safe wiring, which is superior to any technical solution by far, or you can try to acupuncture a cable cord of your computer. You will have little pleasure of it.
Nikola Tesla also in this point was ahead of the times. He has experimented with a single conductor technology, for which the insulation layer is constructed like for a nerve fiber (fig. 25.4 B). We'll have to learn to handle such aids in the future, if we in engineering want to emulate nature and as well want to send informations as scalar wave. A normal power cable as said conducts everything, useful as well as interference signals. If however the signal should be accomodated in the noise, a selective cable is required, which should be constructed according to the example of the nerve conductors. Before a scalar wave technology can be introduced and successfully used according to that completely other cables, coupling elements, amplifiers and other components should be developed. The trouble should be worthwhile at the chances, which this genial technology offers us!

25.4 The brain, a scalar wave computer

The brain cells (neurons) are of the same kind as the nerve cells. Hence can be done without a signal transformer, can the informations transmitted by the nerves to the brain directly be processed further. From that follows that also the brain without exception works with potential vortices. There are several reasons for this hypothesis:

1. as said, the lacking of a signal transformer.
2. the high performance density of the think apparatus. (As a result of the concentration effect of the potential vortices the efficiency of the human brain is in such a way high concentrated compared to the much more space claiming computers functioning on the basis of currents).
3. the brain activity measurable from the outside with the EEC
4. spark formation and corona discharges at open top of the skull. (Brain surgeons can report of such observations).
5. The insulation defect occurring in the case of epileptics. (During a fit instable potential oscillations of the nerve cells occur, which lead to strong electric blows).

With the „exciting" and the inhibiting"' synapses as separation point between the neurons both a ,.high-active" and a „low-active"' method of operation is possible and with that a redundant, particularly interference safe signal transmission.
Safety for interference is very important not only in nature. In the operating instructions of a PC can be read: „operate only at room temperature, keep dry, don't throw or shock, take cure for sufficiently cooling air, ground apparatus, pay attention to mains voltage, etc". A comparison with the range of operation of man is like scorn.
Nevertheless the consequences if errors occur are quite similar: a garage door, which opens if a mobile is switched on, by all means can be compared with a light phenomenon, which we perceive after a blow on the eye at the biological level.

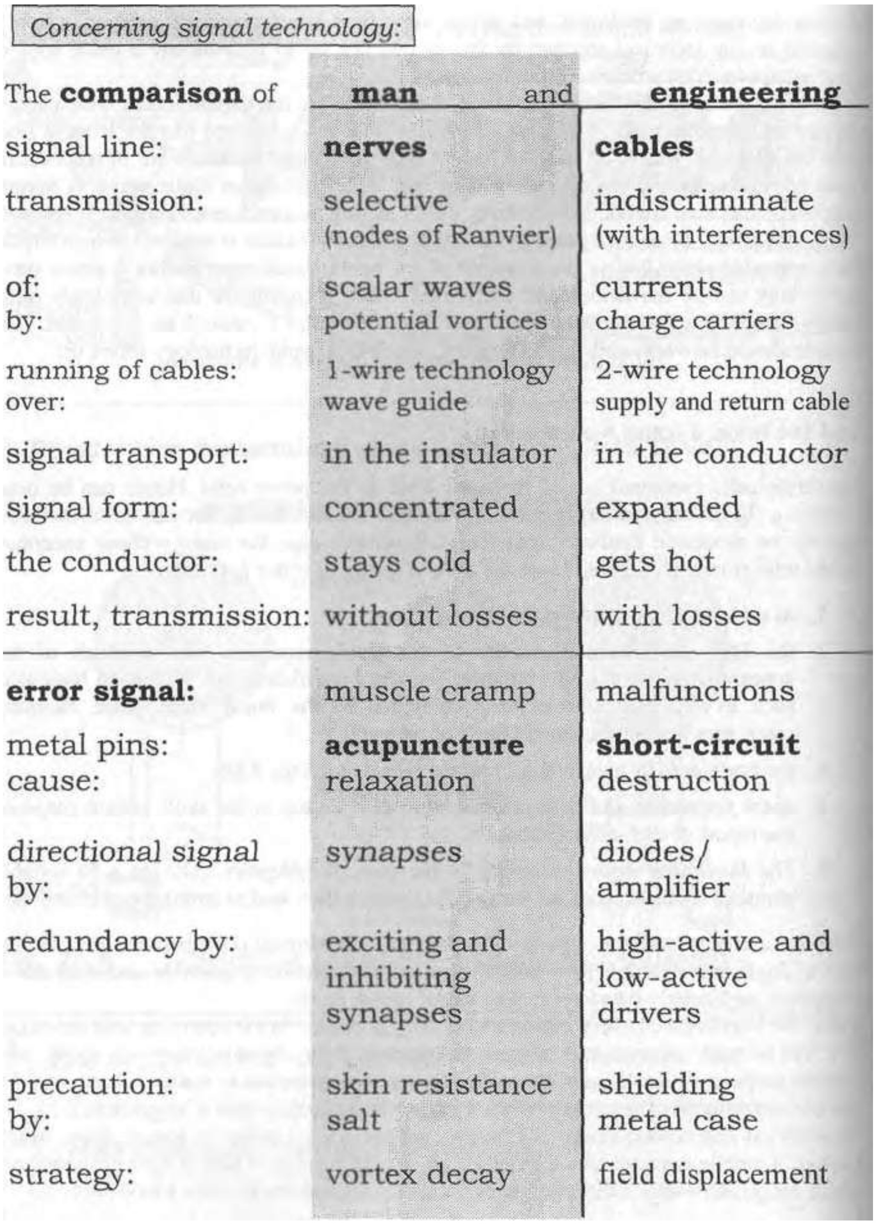

Concerning signal technology:

The **comparison** of	**man** and	**engineering**
signal line:	**nerves**	**cables**
transmission:	selective	indiscriminate
	(nodes of Ranvier)	(with interferences)
of:	scalar waves	currents
by:	potential vortices	charge carriers
running of cables:	1-wire technology	2-wire technology
over:	wave guide	supply and return cable
signal transport:	in the insulator	in the conductor
signal form:	concentrated	expanded
the conductor:	stays cold	gets hot
result, transmission:	without losses	with losses
error signal:	muscle cramp	malfunctions
metal pins:	**acupuncture**	**short-circuit**
cause:	relaxation	destruction
directional signal	synapses	diodes/
by:		amplifier
redundancy by:	exciting and	high-active and
	inhibiting	low-active
	synapses	drivers
precaution:	skin resistance	shielding
by:	salt	metal case
strategy:	vortex decay	field displacement

Fig. 25.5: ____Comparison of the signal technology

25.5 Concerning signal engineering

By means of a technical analysis of biological relations completely new interpretations result also for the occurring of a disease. We permanently and everywhere are surrounded by noise signals, but as a rule they can't touch us, because the body has developed perfect strategies for defence. The nodes of Ranvier on the nerve bundles here are just as helpful as the diode effect of the synapses. By means of the salt content and the skin resistance the body in addition controls the uptake of potential vortices from its surroundings, by using that the vortex decay is determined by the conductivity.

The specialties of the by humans used signal technology come to light particularly clear if compared with the cable technology used in the technical world (fig. 25.5). So is worked with only one wire instead of with supply and return cable, are mediated potential vortices instead of charge carriers, does the transmission take place in the insulator without losses and not in an electric conductor, which as a result gets hot and produces current heat losses. Nerves thanks to their ability of selection represent an intelligent form of signal transmission, by helping to filter the asked information from the noise. This surely is necessary since with cables, which indiscriminately transmit every signal, an use of scalar waves hasn't succeeded yet. We should try to learn of nature!

A special challenge is the protection against error signals. A passive shielding by a metal case however is not possible, because scalar waves can't be shielded by principle, so that the precaution should be taken actively by means of the conductivity. That's why we sweat salt if we strain us physically, whereby the vortex decay is determined by the conductivity, which depends on the salt content of the body liquids. In the case of a sweating activity the body reduces its conductivity, so that the needed potential vortex energy will reach the cells.

If the body sometime isn't able to defend itself against interference signals, then malfunctions or pathological reactions are a possible result. During having a bath for instance a muscle cramp can occur, if the body doesn't defend itself fast enough or sufficiently against the high potential vortex activity in the water. Now vortices can be picked up in the nerves, which are of the same kind as the ones emitted by the brain, only that both muscles, biceps and triceps at the same time get the signal to contract. The result is a cramping of both muscles.

Thus the brain has developed intelligent strategies to protect itself of interspersed misinformations. It weighs the incoming signals and forgets again all unimportant ones more or less fast. We speak of the ability to learn and that means that signals rise in the valuation scale and with that are stored longer, the more frequent repeated our brain receives them. This strategy assumes that interference signals only occur sporadic, for which reason they are rated unimportant and fast are again forgotten. A PC on the other hand doesn't have such a property. It notices everything indiscriminately and sometime will crash of overload, if not the user will constantly foster it and will administer the available memory. A PC is and stays stupid.

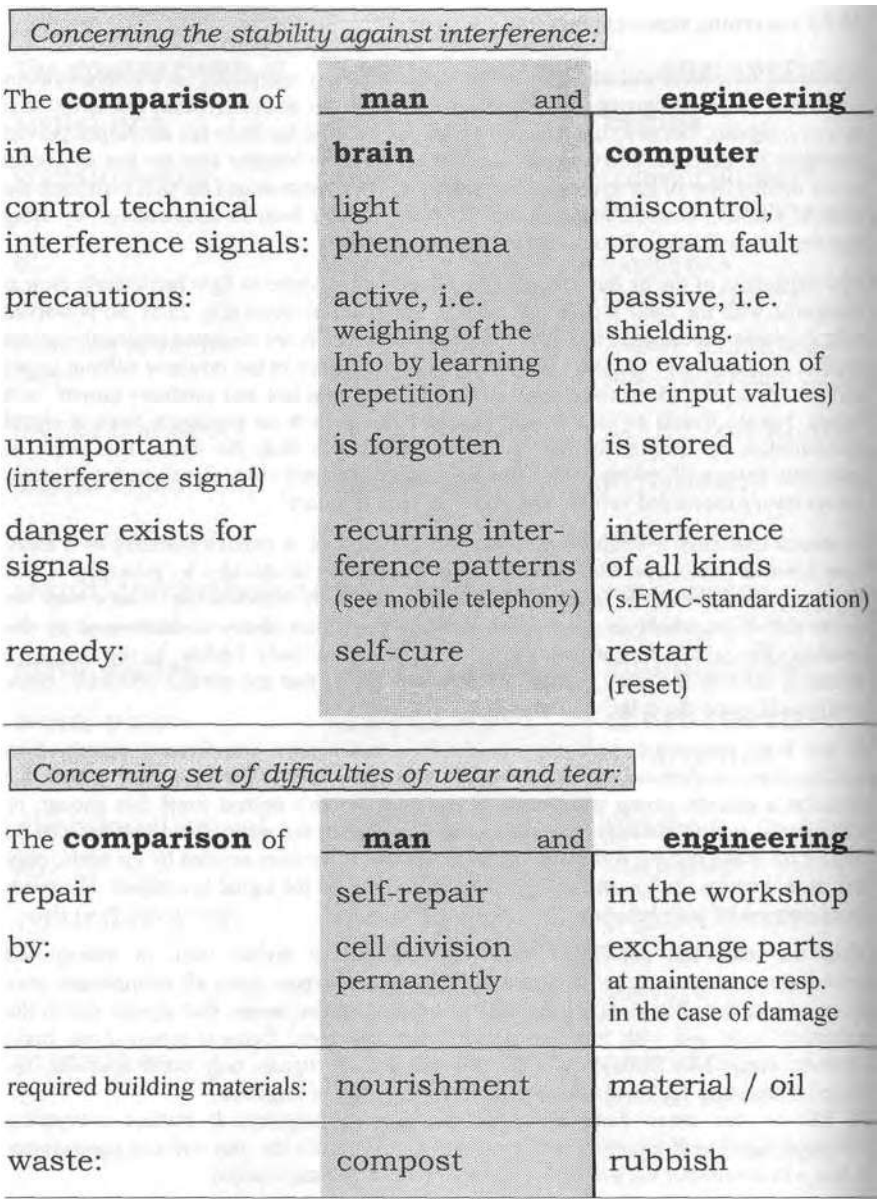

Concerning the stability against interference:

The **comparison** of	**man** and	**engineering**
in the	**brain**	**computer**
control technical interference signals:	light phenomena	miscontrol, program fault
precautions:	active, i.e. weighing of the Info by learning (repetition)	passive, i.e. shielding. (no evaluation of the input values)
unimportant (interference signal)	is forgotten	is stored
danger exists for signals	recurring interference patterns (see mobile telephony)	interference of all kinds (s.EMC-standardization)
remedy:	self-cure	restart (reset)

Concerning set of difficulties of wear and tear:

The **comparison** of	**man** and	**engineering**
repair	self-repair	in the workshop
by:	cell division permanently	exchange parts at maintenance resp. in the case of damage
required building materials:	nourishment	material / oil
waste:	compost	rubbish

Fig. 25.6: ____Strategies against interference signals, for damages linked with operation and for wear and tear.

25.6 Stability against interference and repair mechanisms

The interference signals present in our natural environment as a rule are distributed stochastic, but not so artificial interference signals like for instance transmitters emit. If for instance in the case of mobile telephony there occur time and again identical signal patterns and if a person perceives these, then because of the continual repetitions a high importance is pretended and precious storage space in the brain is allocated. This to a special extent applies for the permanent stand-by signals, which are emitted by mobiles and cordless phones even then, if we don't phone at all. Such misdevelopments thereby would be technically avoidable just like that!
If still no gauges are available and as a result there exist neither guidelines nor limits for the radiation exposure of scalar waves, then such devices must be developed and built. For the requested electromagnetic environmental compatibility not engineering, but man should be in the centre of attention!

Let us throw a short glance at the set of difficulties of wear and tear. Most technical devices find the way into the workshop only, if they already are defect. Some aren't repaired anymore in principle and immediately sent to the rubbish, because they are worn-out and a repair isn't worth the effort anymore one says, whereas other, mostly expensive systems are being serviced by exchanging all wearing parts.

Nature has brought to perfection the last principle. It allows the body a permanent maintenance; by permanently producing new cells and replacing consumed ones. It with that obtains a considerably longer operating time and even is capable to heal wounds. Just imagine dents in our cars would disappear from alone after a few weeks and the bodywork would look like new. Such an optimal maintenance is costly and it has its price. By means of the cell division the building plan for the spare parts is copied. The task on the other hand is transmitted to the new cells „by radio" by means of waveguide channels, as proves the matching structure of scalar wave and waveguide (chapter 21.12 and fig. 25.7). The cells hence have a kind of „decentral intelligence", which technical matter lacks completely. A comparison with engineering nevertheless makes sense, because wear and tear occurs for all systems in continuous operation. Thus strategies must be developed to provide the necessary material for operation, building-material and exchange parts.

By sizing of the hyperboloid structures of the matrix channels it should be possible to determine the wavelength of the scalar waves to enter them in the frequency diagram (23.1).

Tunnel structure of the basic substance according to Prof. Heine[i]:

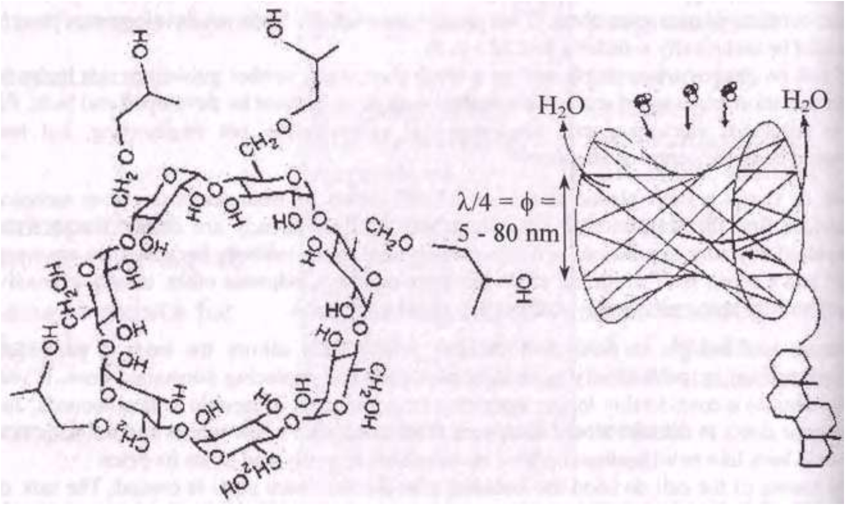

Fig. 25.7: ___ Information channels in the intercellular matrix.[i]

---

[i]: Hartmut, Heine: Lehrbuch der biologischen Medizin. Grundregulation und Extrazellulare Matrix, 2. Aufl. 1997, Hippokrates Verlag Stuttgart, S. 56

25.7 The microwave window

There still is lacking information about the frequency or the velocity of propagation to be able to specify and enter in figure 23.1 the cell information. Let's take to hand the book about "Biostrahlen"<i>. In there is explained how for individual cells the emission of electromagnetic signals already could be measured, which are interpreted as circularly polarized waves (fig. 21.10 A). It is said that thereby their polarization plane is rotating with the speed of light, whereas the wave itself is propagating in longitudinal way relatively slow and according to the authors with the speed of sound.
The occupation with the potential vortex theory in the explanation lets immediately recognize the vortex, for this is rotating with the speed of light in circles as rolled up wave (fig. 21.10). If it is bound to the surrounding matter with closed vortex centre, then it can be expected that the propagation actually takes place with the speed of body sound.
It could concern the same signals, which Prof. A. Popp calls biophotons and detects measuring technically in living organisms<ii>. He however considers the phenomenon for the same wavelength with the speed of light and lands at light frequencies, even then if nothing is glowing visibly! The question is asked: Does it concern the frequency of the light or only the corresponding wavelength or actually both, thus light, as is expressed in the name biophotons?
The photomultipliers, which Prof. Popp uses as „light amplifier", however can only be tuned to certain wavelengths and not to frequencies. Even if the detected biophotons have the wavelength of the light, then nevertheless nothing will glow if the velocity of propagation and as a result also the frequency differs from that of the light for several powers of ten. In the case of the immense number of cells also the number of photons should correspondingly sum up and the body as a whole should start to glow, which is not the case.

The waveguides in the intercellular matrix serving the cell communication, which Prof. H. Heine observes microscopically, have wavelengths between 20 and 300 nanometer, which corresponds to the range of the ultraviolet radiation<iii>. But if the propagation is slower than the light for 6 powers of ten, then also the frequency will only amount to one millionth and be situated into the range of the microwaves. Here a biological window seems to be present, to which we should turn our attention for reasons of the electromagnetic environmental compatibility!

The thermal radiation, which reaches the earth from the sun and the planets, lies in the microwave range between 2 and 20 cm. If the sun does well to us, if we need the radiation, then it could be because of the identical frequency. But that also means that the sun and the planets are capable to have an effect on the cell communication, that they for instance could function as clock for the heart.

<i>: H. Mayer, G. Winklbaur: Biostrahlen, Verlag ORAC, Wien 1989, 5. Auflage S. 97: Messung von der DNS eines Gurkenkeims abstrahlenden Photonen.

<ii>: A. Popp: Neue Horizonte in der Medizin, 2.Aufl. Haug Verlag Heidelberg 1987

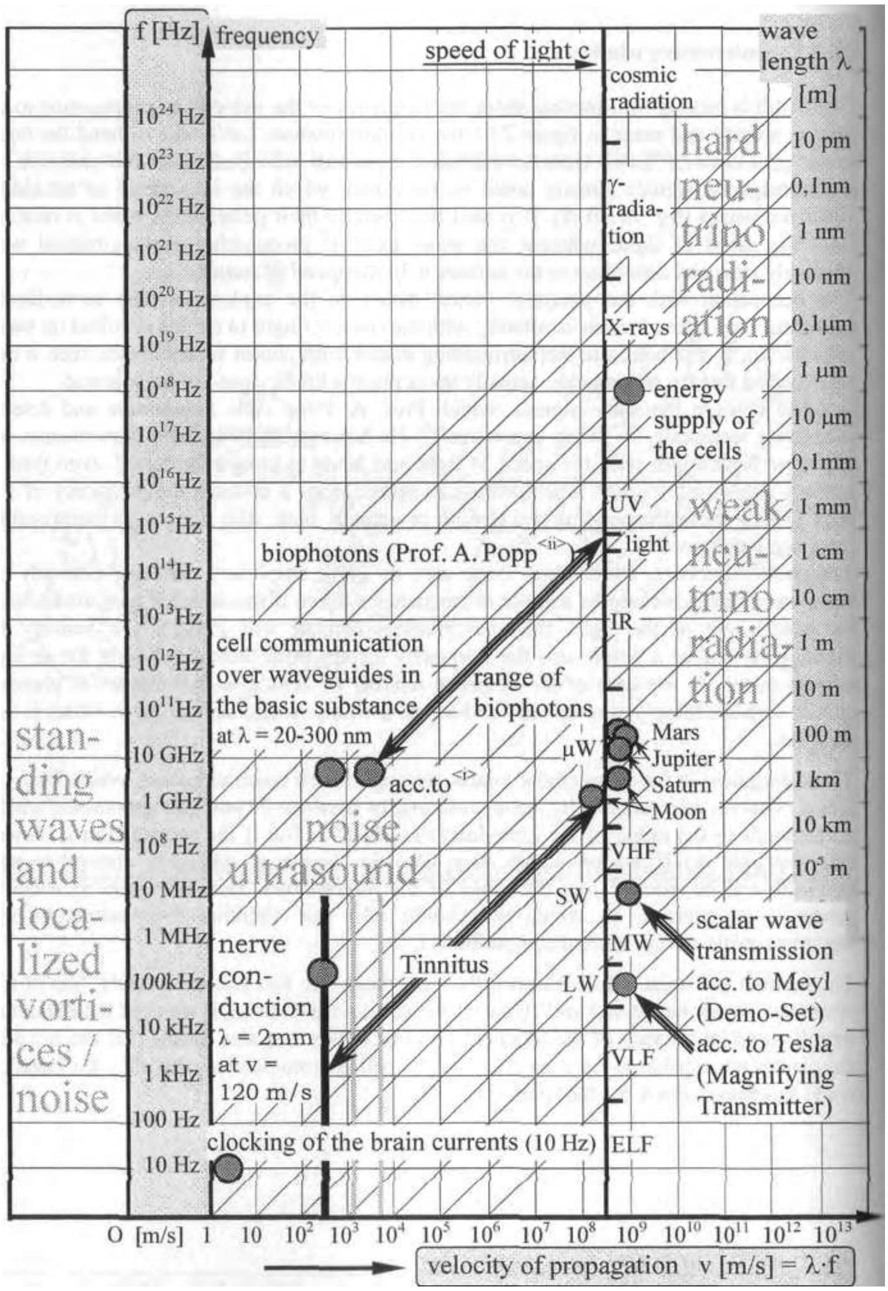

Fig. 25.8: Frequency diagram with entries concerning biologically relevant areas (Tinnitus, nerve conduction, etc.).

25.8 Discussion concerning scalar wave medicine

In the sensitive range of the supposed microwaves window however are also situated the mobile phones and their harmonic waves, which occupy a broad spectrum in particular for pulsed operation in digital nets. The D-net for instance has a wavelength of 32 cm, the E-net lies at half that wavelength. But to which frequency does this correspond at the speed of sound? Now, the frequency is 6 powers of ten smaller and now lies at 1 kHz resp. 2 kHz. To that are added the numerous harmonic waves, which form a noise signal and lie above that. With that these signals are situated completely in the audible range, there where our ears are most sensitive!
This cause we probably owe the disease of modern civilization „Tinnitus". Every charged particle will follow this electromagnetic oscillation and produce corresponding sound oscillations, which can hear not „sick ", but completely on the contrary „healthy" people, who as a result possibly get sick. The objection, in this range also cosmic radiation for instance from the planets is present, is legitimate. It however should be considered that planets also are going away from the earth again and in addition is present a fluctuation according to the time of day due to the rotation of the earth, while the mobile telephone masts in our vicinity radiate in continuous operation.

In this window in addition the clocking of the brain currents takes place at 10 Hz. I would recommend blocking the acoustic relevant range between 2 cm (16 kHz) and 3 m wavelength (100 Hz) for all technical use. Every operation of a transmitter in a biological window harms all people and cannot be answered for by any institution.

It further should be considered that the biological window of the plants and animals as a rule correspond to that of man, but sometimes are shifted significantly upward or downward in wavelength. We are not entitled to judge nature. The immune system of the animals now obviously has reached the breaking point and also that of man doesn't seem to be that anymore, which it originally was.
We must proceed from the assumption that many diseases on the one hand and therapy methods on the other hand partly direct and partly indirect have to do something with scalar waves. (E.g. the occurring of cancer[<4i>]).

<i>: H. Mayer, G. Winklbaur: Biostrahlen, Verlag ORAC, Wien 1989, 5. Auflage, S. 97: Messung von der DNS eines Gurkenkeims abstrahlenden Photonen.

<ii>: A. Popp: Neue Horizonte in der Medizin, 2.Aufl. Haug Verlag Heidelberg 1987

<iii>: Hartmut. Heine: Lehrbuch der biologischen Medizin. Grundregulation und Extrazellulare Matrix, 2. Aufl. 1997, Hippokrates Verlag Stuttgart, S. 56

<4i>: K. Meyl: Skalarwellenstrahlung. Welche Bedeutung haben Skalarwellen fur die Medizin? GNP-Vortrag am 29.04.2001, Rhein-Main-Halle Wiesbaden, see: Co'med Fachzeitschrift fur Komplementarmedizin, 6/2001, pp.55-60.

I. According to the Maxwell theory:

1. Faraday's law of induction:

$$\boxed{\mathrm{rot}\,\mathbf{E} = -\,\delta\mathbf{B}/\delta t} \quad (26.1)$$

with $E = E(r,t)$

and $H = H(r,t)$: $\mathbf{B} = \mu \cdot \mathbf{H}$ (26.2)

$$\boxed{-\,\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \mu \cdot \delta(\mathrm{rot}\,\mathbf{H})/\delta t} \quad (26.3)$$

2. Ampere's law:

$$\boxed{\mathrm{rot}\,\mathbf{H} = \mathbf{j} + \delta\mathbf{D}/\delta t} \quad (26.4)$$

with:

Ohm's law: $\mathbf{j} = \sigma \cdot \mathbf{E}$ (26.5)

dielectric displacement: $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (26.6)

relaxation time: $\tau = \varepsilon/\sigma$ (26.7)

$$\boxed{\mathrm{rot}\,\mathbf{H} = \varepsilon \cdot (\mathbf{E}/\tau + \delta\mathbf{E}/\delta t)} \quad (26.8)$$

3. Inserting equation 26.8 into 26.3 yields:

$$-\,\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \mu \cdot \varepsilon \cdot (1/\tau \cdot \delta\mathbf{E}/\delta t + \delta^2\mathbf{E}/\delta t^2) \quad (26.9)$$

with the abbreviation: $\mu \cdot \varepsilon = 1/c^2$ (26.10)

4. Field equation of a damped transverse wave:

$$\boxed{\underbrace{-\,\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} \cdot c^2}_{\text{transverse}} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{\text{wave}} + \underbrace{(1/\tau) \cdot \delta\mathbf{E}/\delta t}_{\text{vortex damping}}} \quad (26.11)$$

Fig. 26.1: Derivation of the wave damping by means of the formation of vortices!<i>

<i>: see also EMC, part 1, chapter 5.3, 2.borderline case.

26. Recapitulation from the viewpoint of textbook physics

Now that we in the meantime have accumulated innumerable mosaic parts as inspiring contributions to the discussion for the information technical seminar, it is time to sort the ideas and to put the parts together to an overall picture.

Sceptics and orthodox scientists can only be convinced, if we start from textbook physics and completely do without postulates. Those demands will be fulfilled!

26.1 Common misinterpretation of the antenna losses

The mathematical description of physical relations leads to the well-known laws, which shouldn't be doubted anymore as soon as they are accepted to be correct. But what about the interpretation? Although a law dictates the interpretation and there is no choice, because laws must be adhered to, yet textbooks from time to time violate the mathematically dictated interpretation, a circumstance, which can't be accepted. I would like to illustrate this with an example.

Let us assume that the measured degree of effectiveness of a transmitting antenna amounts to 80 percent. There exist better antennas, but also distinctly worse antennas, but I'm not aiming at a certain construction. The statement simply says, that 80% of the fed in HF-power is transformed into Hertzian waves. Thus there arises a loss of power of 20 percent, and the question follows: of what do those 20% consist?

The answer, which is usual among experts and is supported by the textbooks, reads: the antenna wire gets hot and also the air around the antenna is heated by dielectric losses. In short, heat is formed.

But I have to point out and will furnish proof that this interpretation is predominantly wrong! It in any case isn't in accord with the laws of Maxwell. Who namely obeys the laws, comes to an entirely different result!

A short derivation brings it to light (fig. 26.1).

We start with the formulation of Faraday's law of induction according to the textbooks (26.1), apply the curl-operation to both sides of the equation (26.3) and insert in the place of rot H Ampere's law (26.4-26.8). The generally known result describes a damped electromagnetic wave (26.11)[i].

It on the one hand is a transverse wave, which represents 80% of the antenna power for our example. On the other hand a damping term can be found in the equation, which obviously corresponds to the sought-for 20%. With that the answer would have been found. We realize that because of a damping of the wave 20% antenna losses arise. These losses can't concern heat at all, since the damping term in the equation has got nothing in common with thermodynamics. In the equation doesn't stand anything about heat! Such a mistake!

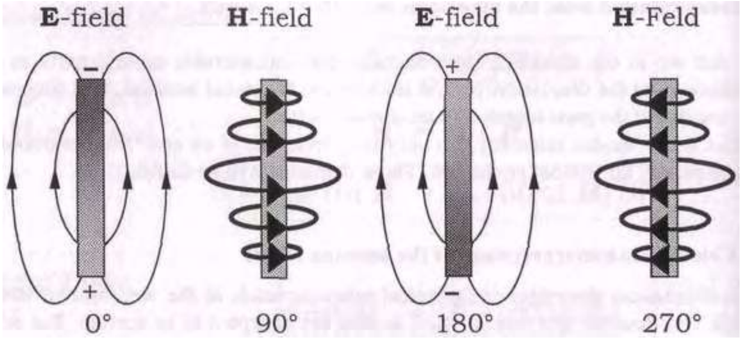
E-field
H-field
E-field
H-Feld
−
+
+
−
0°
90°
180°
270°

Fig. 21.8 A: The fields of the oscillating dipole antenna.[i]

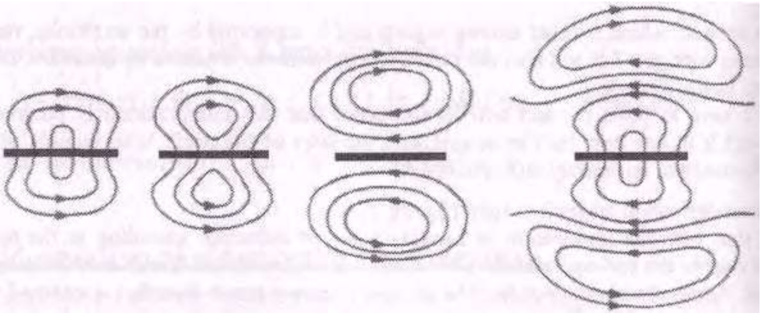

Fig. 21.9 A: The coming off of the electric field lines from the dipole.

---

[i]: Repetition belonging to chapter 21

26.2 Wave damping by field vortices

- Mathematically seen the damping term describes vortices of the electromagnetic field. This term for instance forms the basis of all eddy current calculations<ii>.
- Physically seen some waves - in our example it is 20 percent - roll up to field vortices, with which the wave damping and the antenna losses would be explained.

In the course of time a substantial part of the generated vortices will fall apart. These thereby will produce eddy losses in form of heat. Thus eventually still heat is produced - agreed. The criticism of the textbooks consists of the circumstance that we by no means can proceed from the assumption that all vortices spontaneously fall apart and a total conversion into heat will take place. The process in addition takes place with a temporal delay. The time constant x gives information in this respect. Field energy is buffered in the vortex, where some vortices live very long and it can't be ruled out that a few even exist as long as you like.

To find out more about these field vortices and their behaviour, one has to get deep into vortex physics. Unfortunately nothing can be found about vortex physics in the textbooks. The mistake is systematic. The following short compendium should help close the gap:

- Mathematically seen a closed-loop electromagnetic field vortex will show as a scalar. Such field vortices, which are mediated by a scalar field, are propagating exactly as charged particles in longitudinal manner as a scalar wave.
- Physically seen a closed-loop field vortex has got particle nature. If one particle kicks off the next one then an impulse is mediated, then a shock wave is occurring, a longitudinal undulation of the particles.

From the vortex physical view the interpretation of the antenna example now sounds entirely different:
The charge carriers in an antenna wire oscillating with high-frequency form a longitudinal shock wave. Between current and tension voltage usually a phase shift of 90° is present. The fields produced by these charge carriers form a scalar wave field in the immediate vicinity of the antenna, the so-called near-field zone, which likewise contains longitudinal field components and shows a phase shift of 90° between electric and magnetic field (fig. 21.8 A). As in textbooks is clarified by field lines, the generated fields actually form vortices, where one structure kicks off the next one (fig. 21.9 A).

The vortices in the near-field zone of an antenna consist of standing waves, which obviously are transforming with increasing distance. In our example 80% of these are unrolling and turn into transverse waves, whereby the characteristic phase angle between E- and H-field at that occasion becomes zero.
Let's turn again to those 20 percent loss.

<ii>: s.a.: K. Meyl: Wirbelstrome, Diss. Uni. Stuttgart 1984, INDEL Verlagsabt.

I. According to the Maxwell theory:

The consistent application of textbook physics

- Longitudinal waves run in the direction of a field pointer.
- The field pointer oscillates, the vector of velocity oscillates along!
- At relativistic velocities field vortices are subject to the Lorentz contraction.
- The faster the oscillating vortex is on its way, the smaller it gets.
- The vortex permanently changes its diameter (see fig. 21.10 B).
- With the diameter the wavelength decreases (see fig. 22.4).
- The swirl velocity is constant (= speed of light c).
- The eigenfrequency of the vortex oscillates with opposite phase to the wavelength.
- The vortex acts as a frequency converter!
- The measurable mixture of frequencies is called noise.

leads to the statement:

- The antenna noise corresponds to the antenna losses!

II. Mathematical description of a wave by the inhomogeneous Laplace equation:

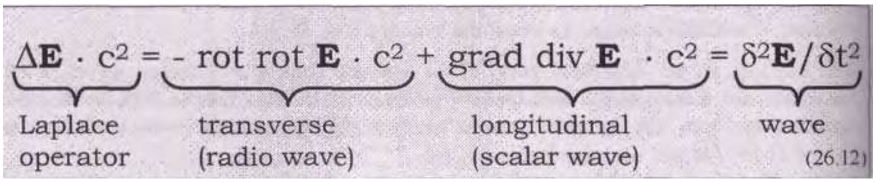

$$\underbrace{\Delta \mathbf{E} \cdot c^2}_{\text{Laplace operator}} = \underbrace{- \text{rot rot } \mathbf{E} \cdot c^2}_{\text{transverse (radio wave)}} + \underbrace{\text{grad div } \mathbf{E} \cdot c^2}_{\text{longitudinal (scalar wave)}} = \underbrace{\delta^2 \mathbf{E}/\delta t^2}_{\text{wave}} \qquad (26.12)$$

Divergence E is a scalar!
The corresponding term founds a scalar wave.

Fig. 26.3: Mathematical description of a wave according to Laplace.

26.3 Laplace versus Maxwell

Longitudinal waves as is well-known don't know a fixed velocity of propagation at all. Since they run in the direction of an oscillating field pointer also the vector of velocity will oscillate. For so-called relativistic velocities in the range of the speed of light c the field vortices are subject to the Lorentz contraction. That means, the faster the oscillating vortex is on its way, the smaller it gets. The vortex, as a mediator of a scalar wave carrying impulse, permanently changes its diameter.
Since, in the case of the vortices, it should concern rolled up waves, the vortex velocity will continue to be c, with which the wave now runs around the vortex centre in a circle. From that follows that if the diameter gets smaller, the wavelength of the vortex as well will decrease, whereas the eigenfrequency of the vortex accordingly increases.
If the next moment the vortex oscillates back, the frequency again decreases. The vortex acts as a frequency converter! The mixture of high-frequency signals distributed over a broad frequency band formed in this way, is called noise. A noise signal indeed is measured from the outside with the help of broadband receivers.. We also speak of antenna noise and with this knowledge we can further specify the 20% antenna losses: The antenna produces 20 % noise, which can be put equal to the generated vortices because of the wave damping.

At this point the Maxwell theory doesn't leave us room for interpretation at all. If in the textbooks the impression is aroused, as if the noise were an independent discipline, than that is not true at all. How much the noise is connected with the electromagnetic waves, proves a short look at the wave equation.

The wave equation found in most textbooks has the form of an inhomogeneous Laplace equation. The famous French mathematician Laplace considerably earlier than Maxwell did find a comprehensive formulation of waves and formulated it mathematically (eq. 26.12), which until today is still accepted as valid.

On the one side of the wave equation the Laplace operator stands, which describes the spatial field distribution, and which according to the rules of vector analysis can be decomposed into two parts. On the other side the description of the time dependency of the wave can be found as an inhomogeneous term.

If the wave equation according to Laplace (26.12) is compared to the one, which the Maxwell equations have brought us (26.11), then two differences clearly come forward:

1. In the Laplace equation the damping term is missing. It doesn't describe the formation of vortices, and that means vortices do not exist at all, or present vortices have been there from the beginning.
2. With divergence E a scalar factor appears in the wave equation, which founds a scalar wave.

At this point at once hot tempered discussions concerning the question of the existence of scalar waves blaze up. But this question has already been answered clearly with the vortex consideration. Since an accepted description of longitudinal and scalar waves exists with the plasma wave and the plasma wave can be derived directly without postulate from the term of the wave equation (chapter 21.4/21.5), which founds scalar waves, there are further arguments present for their existence.

I. According to Maxwell:

$$\underbrace{-\operatorname{rot}\operatorname{rot}\mathbf{E}\cdot c^2}_{\text{transverse}} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{\text{wave}} + \underbrace{(1/\tau)\cdot\delta\mathbf{E}/\delta t}_{\text{vortex damping}} \qquad (26.11)$$

Description of electromagnetic waves with vortex damping
Example: sunlight and the damping in the ionosphere (bleu sky).

II. According to Laplace:

$$\underbrace{\Delta\mathbf{E}\cdot c^2}_{\substack{\text{Laplace}\\ \text{operator}}} = \underbrace{-\operatorname{rot}\operatorname{rot}\mathbf{E}\cdot c^2}_{\substack{\text{transverse}\\ \text{(Hertz)}}} + \underbrace{\operatorname{grad}\operatorname{div}\mathbf{E}\cdot c^2}_{\substack{\text{longitudinal}\\ \text{(Tesla)}}} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{\text{wave}} \qquad (26.12)$$

Description of transverse electromagnetic waves (Hertzian waves) and longitudinal scalar waves (Tesla radiation).
1st example: propagation of light as a wave or as a particle.
2nd example: useful signal or noise signal of an antenna.

III. Mathematically seen (comparison of coefficients):

$$\underbrace{(1/\tau)\cdot\delta\mathbf{E}/\delta t}_{\text{vortex}} + \underbrace{\operatorname{grad}\operatorname{div}\mathbf{E}\cdot c^2}_{\text{scalar}} = 0 \qquad (26.13)$$

IV. Physically seen

(from the comparison of equation 26.12 with equation 26.11):

Vortices propagate longitudinally as a scalar wave!

Fig. 26.2: ______ Comparison of the two wave descriptions.

26.4 Error term in the wave equation

From the comparison of coefficients of both wave descriptions follows even more:

- Mathematically seen the damping - resp. vortex term according to Maxwell can be put equal to the scalar wave term according to Laplace.
- Physically seen the generated field vortices form and found a scalar wave.

Here also doesn't exist any room for interpretation, as long as we work with the wave equation according to Laplace and at the same time adhere to the Maxwell theory. If however the scalar wave part is put equal to zero, as is common practise in the textbooks, then as a consequence neither vortices nor noise may exist. But that contradicts all measuring technical experience! Since every antenna produces more or less noise, the textbooks obviously only show half the truth. Science however gropes for the whole truth and that should be fathomed.

If in the case of the antenna example the vortex part amounts to 20%, then that's tantamount to 20% scalar wave part, resp. 20% noise. The scalar wave part constitutes with regard to the Hertzian useful wave something like an error term in the wave equation. The part definitely is too big, as that it might be put equal to zero. Even so all error consideration in the textbooks is missing, if the scalar wave term is assumed to be zero. That violates all rules of physics and of taught scientific methodism.

In practice this shows by a useful signal going under in the noise and reception not being possible anymore as soon as the scalar wave part gets out of control. Even in this case, for which the degree of effectiveness tends towards zero, it still is common practise to put the error term, which is dominating everything, equal to zero. But who in this point follows the textbooks, disregards with that the wave equation and doing so by no means can refer to the circumstance that all colleagues make the same mistake.

The building of physics behaves like a house of cards, where the cards mutually support each other. Perhaps that is the deeper reason why those, who have discovered a marked card, don't pull it out immediately. In addition they are hindered by the self appointed guardians of the ,,pure doctrine", since everyone knows what happens with the house of cards if the marked card is pulled out. Only, do we want to and can we live with that in the long run? Is it a solution of the problem, if the so-called experts among the physicists and technicians look away and don't deal with the foundation of their branch anymore? If universities crash their basis education into the wall and choke off every contradiction?

Please allow me to pull out the marked card now and place it on the table!

It concerns the question: what is the nature of the field vortices, which form a scalar wave in space. Eddy currents in the iron parts of the antenna are explained with the field equations, but not the noise, which is measured especially in the air. If an antenna on the one hand produces field vortices and as a consequence eddy losses and on the other hand dielectric losses, then we can assume that besides the eddy currents in the conductor also vortices in the dielectric must exist. Let's search for them!

Interim result (comparison of arguments):

The Maxwell equations on the one hand dictate that as the reason for a wave damping only field vortices should be considered.	On the other hand the same laws merely describe eddy currents, which can only occur in the electrically conducting parts of the antenna.
On the one hand the field vortex interpretation makes it possible to explain the noise of an antenna perfectly.	On the other hand does the noise appear in the neighbourhood of the antenna, thus in the air and not in the iron parts!
The mathematical formulation reveals, how wave and vortex, resp. noise, co-operate and how one should imagine the conversion of one form into the other form.	In field physics on the other hand is missing a useful description of electric field vortices in a dielectric, which could found the noise signal.

Table 26.5: Arguments pro and contra.

26. 5 Interim result

It shouldn't be a disadvantage, to interprete physical laws more consistently than usual, even if in the present case orthodox science through that at first should fall into a deep crisis. If the way is worthwhile, only will show at the end.
Let us try to work out the contradictions in form of a comparison of arguments:

- The Maxwell equations on the one hand dictate that as the reason for a wave damping only field vortices should be considered.
    - On the other hand the same laws merely describe eddy currents, which can only occur in the electrically conducting parts of the antenna. .
- On the one hand the field vortex interpretation makes it possible to explain the noise of an antenna perfectly. .
    - On the other hand does the noise appear in the neighbourhood of the antenna, thus in the air and not in the iron parts!
- The mathematical formulation reveals, how wave and vortex, resp. noise, co-operate and how one should imagine the conversion of one form into the other form.
    - In field physics on the other hand is missing a useful description of electric field vortices in a dielectric, which could found the noise signal.

The most obvious among all conceivable solutions is the one that we have to assume the existence of dielectric field vortices, so-called potential vortices. We are challenged to search for a corresponding description. If the quest should be successful, then the contradictions would be overcome. In addition there is the promise of a whole number of simplifying explanations of various phenomena in the dielectric (see fig. 26.5 and fig. 26.7).

The phenomenon of noise becomes an aspect of wave physics, which is more than merely a field disturbance, which makes the reception of the useful wave more difficult. If the scalar wave nature is realized, then applications can be imagined, in which the noise is used as useful signal. In the way that the scalar part in the wave equation doesn't have to be put to zero anymore to obtain freedom of contradiction, even optimizations of antennas or of capacitors are possible with regard to the dielectric losses by means of the calculation of the scalar part.

New in any case is the idea that the dielectric losses of a capacitor are eddy losses and not a defect in material of the insulating material. With that the capacitor losses correspond to a generated noise power. We also can say, every capacitor more or less produces noise! The electric field lines point from one capacitor plate to the other plate. If one plate radiates as a transmitter and the other plate functions as a receiver, then the field propagation takes place in the direction of the electric field pointer and that again is the condition for a longitudinal wave. Here the circle closes in the conclusion, the capacitor field mediates dielectric field vortices, which following the field lines found a scalar wave because of their scalar nature. The heating of the capacitor results from the decay of vortices.

Potential vortices explain div. phenomena in the dielectric:

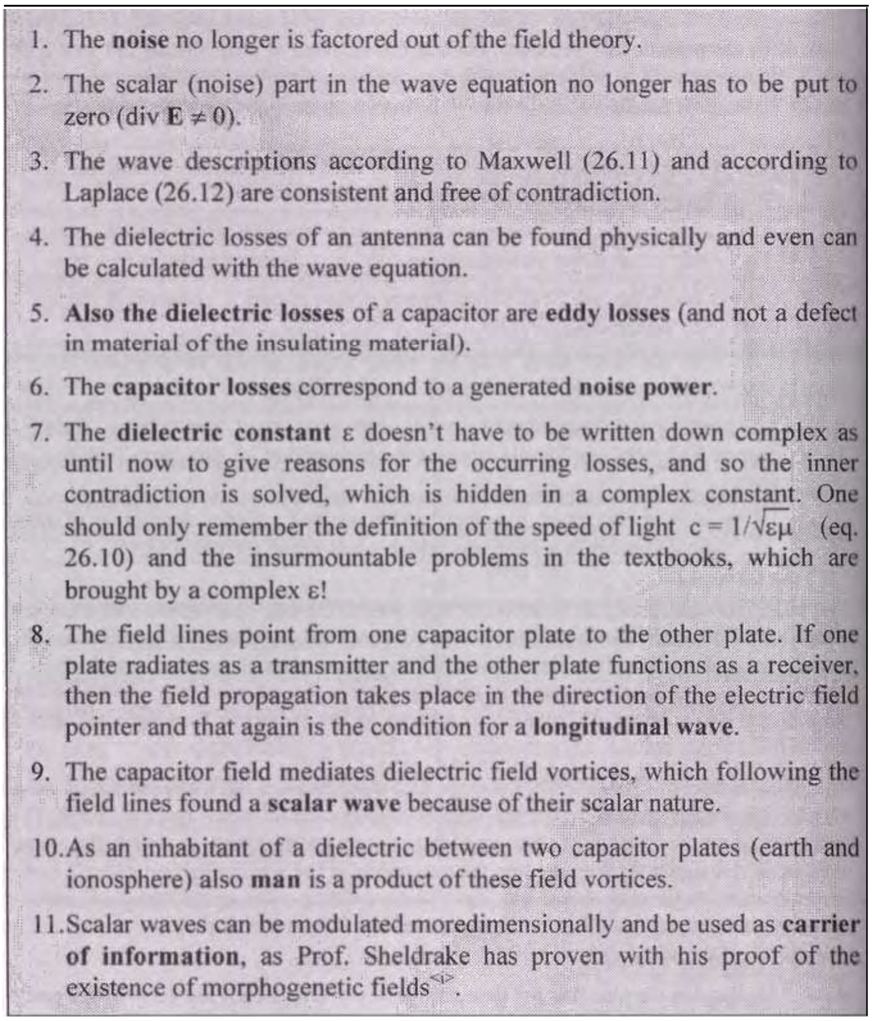

1. The **noise** no longer is factored out of the field theory.
2. The scalar (noise) part in the wave equation no longer has to be put to zero (div $\mathbf{E} \neq 0$).
3. The wave descriptions according to Maxwell (26.11) and according to Laplace (26.12) are consistent and free of contradiction.
4. The dielectric losses of an antenna can be found physically and even can be calculated with the wave equation.
5. **Also the dielectric losses** of a capacitor are **eddy losses** (and not a defect in material of the insulating material).
6. The **capacitor losses** correspond to a generated **noise power**.
7. The **dielectric constant** ε doesn't have to be written down complex as until now to give reasons for the occurring losses, and so the inner contradiction is solved, which is hidden in a complex constant. One should only remember the definition of the speed of light $c = 1/\sqrt{\varepsilon\mu}$ (eq. 26.10) and the insurmountable problems in the textbooks, which are brought by a complex ε!
8. The field lines point from one capacitor plate to the other plate. If one plate radiates as a transmitter and the other plate functions as a receiver, then the field propagation takes place in the direction of the electric field pointer and that again is the condition for a **longitudinal wave**.
9. The capacitor field mediates dielectric field vortices, which following the field lines found a **scalar wave** because of their scalar nature.
10. As an inhabitant of a dielectric between two capacitor plates (earth and ionosphere) also **man** is a product of these field vortices.
11. Scalar waves can be modulated moredimensionally and be used as **carrier of information**, as Prof. Sheldrake has proven with his proof of the existence of morphogenetic fields<i>.

Fig. 26.6: ___ Advantages of a field description extended with potential vortices.

<i>: Rupert Sheldrake: Seven Experiments That Could Change the World. Riverhead Books, 1995

26.6 Failure of the Maxwell theory

If the capacitor losses or the antenna noise should concern dielectric losses in the sense of vortex decay of potential vortices, which don't occur in the Maxwell theory at all, then we are confronted with a massive contradiction:

- For the description of the losses the Maxwell; theory on the one hand only offers field vortices and those only in the conductive medium.
  - On the other hand do the dielectric losses occut in the nonconductor and in the air.
- In conductive materials vortex fields occur, in the insulator however the fields are irrotational. That isn't possible, since at the transition from the conductor to the insulator the laws of refraction are valid and these require continuity! Hence a failure of the Maxwell theory (Fig. 26.7) will occur in the dielectric.
  - As a consequence the existence of vortex fields in the dielectric, so-called potential vortices, should be required!

In electrodynamics as a help the approach of a vector potential A is used, which leads to a complex formulation of the dielectric constant ε and in this way makes it possible, to mathematically describe the dielectric losses of a capacitor by means of the load angle, which stretches in the complex plane. But which physical value does this approach have? How can now the inner contradiction be explained, which is hidden in a complex constant of material? One should only remember the definition of the speed of light $c = 1/\sqrt{\varepsilon\mu}$ (eq. 26.10) and its dependency of ε. For a complex ε here are resulting insurmountable problems in the textbooks.
From the viewpoint of mathematics the introduction of the vector potential at first may represent a help. The before mentioned contradictions however fast raise doubts to the model concept, which from a physical viewpoint eventually will lead to errors, if the speed of light isn't constant anymore and even should be complex.

These considerations should be sufficient as a motive to require potential vortices, even if for their description the field theory according to the textbooks has to be revised. As a supplement there is pointed to the following points:

- As an inhabitant of a dielectric between two capacitor plates also man is a product of these field vortices.
- Scalar waves can be modulated moredimensionally and be used as carrier of information, as Prof. Sheldrake has proven with his proof of morphogenetic fields<i>.

The dielectric vortices moreover provide an explanation for natural events. They form the key to numerous disciplines of science, from physics over biology up to medicine.

<i>: Rupert Sheldrake: Seven Experiments That Could Change the World. Riverhead Books, 1995

Problem of continuity in the case of the coming off of vortices

Example: hightension cable

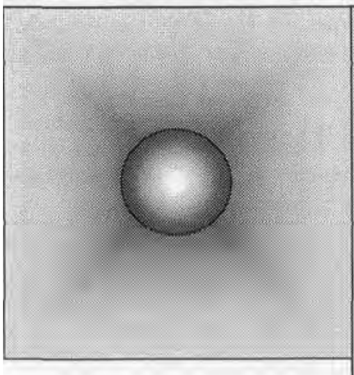

In ***conductive materials vortex fields*** *occur in the* ***insulator*** *however the fields are* ***irrotational****.*
That isn't possible, since at the transition from the conductor to the insulator the ***laws of refraction*** *are valid and these require* ***continuity!***
Hence a ***failure of the Maxwell theory*** *will occur in the dielectric!*

Conclusion: ♦ According to the Maxwell theory there exist no vortices of the electric field (*no potential vortices*) and therefore *no scalar waves.*

♦ Without theory it is impossible to design a usable *scalar wave gauge* and to furnish *evidence*. ⇒ **Classic closed loop conclusion:**

♦ The missing of scientific evidence again „proves“ the assumption of the irrotationality and „confirms“ the correctness of the Maxwell theory.

♦ Hence it cannot be, what shouldn't be!

Fie. 26.7: Concerning the failure of the Maxwell theory

26.7. Concerning the evidence situation

In the question, if a physical phenomenon should be acknowledged as such, experimental, mathematical and physical evidence should be shown. In the case of the potential vortices, the vortices of the electric field and their propagation as a scalar wave, the historical experiments of Nikola Tesla<i> and the modern clone of these can be judged as experimental evidence<ii>.

With the well-known wave equation a mathematical description for this phenomenon has been specified and discussed<iii>. It will be shown that both transverse and longitudinal wave parts are contained alongside in the wave equation, i.e. both radio waves according to Hertz and scalar waves according to Tesla. Doing so the mathematically determined scalar wave properties are identical with the experimental results!
The wave equation is an inhomogeneous Laplace equation and the first and oldest description of scalar waves. It thereby is unimportant, if the famous mathematician Laplace himself already may have realized and discussed this circumstance or not. The description fits perfectly and that is what counts!

At this point the third point should be put on the agenda, the physical evidence. This is connected very closely with the question for a suitable field theory and that again is basing on a corresponding approach.

<i>: N. Tesla: Apparatus for Transmission of Electrical Energy, US-Patent-No.: 649,621, New York 1900, Dr. Nikola Tesla: Complete Patents pp318-321.

<ii>: Johannes von Buttlar im Gesprach mit Prof. Dr. Konstantin Meyl: Neutrinopower, Argo-Verlag, Marktoberdorf, l.Aufl. 2000.

<iii>: K. Meyl: Scalar Waves: Theory and Experiments, Journal of Scientific Exploration, Vol. 15, No. 2, pp. 199-205, 2001

(* * * Task Schedule * * *)

1. Neutrino radiation => energy radiation!
   (acc. to Pauli: ν cares for missing energy in the case of the ß-decay)
2. High neutrino density => high energy density!
3. Neutrino = particle without charge or mass (mean), but because of oscillations: effective value of charge and mass ≠ zero!
4. Interaction only in the case of resonance (e.g. weak interaction)!
   (same frequency, 180° phase shift)
5. Neutrino radiation is an energy source which can be used (not a question of physics, only a question of technology!)
6. Particle radiation (neutrino) => shock wave (like sound)
   => longitudinal wave
   => scalar wave (mathem.)
7. Interaction/resonance of the v = scalar wave problem
8. Scalar waves are a problem of the field theory!
9. Maxwell theory neglects scalar waves.

- Search for a new approach (chap. 27.8)
- Derivation of the Maxwell equations as a
- Derivation of scalar waves (chap. 27.13)
- Derivation of the gravitation and more

Table 27.1: Task schedule belonging to chapter 27

27. Faraday versus Maxwell

Numerous phenomena of the electromagnetic field are described sufficiently accurate by the Maxwell equations, so that these as a rule are regarded as a universal field description. But if one looks more exact it turns out to be purely an approximation, which in addition leads to far reaching physical and technological consequences. We must ask ourselves:

- What is the Maxwell approximation?
- How could a new and extended approach look like?
  - Faraday instead of Maxwell, which is the more general law of induction?
- Can the Maxwell equations be derived as a special case?
- Can also scalar waves be derived from the new approach?
- Can the gravitation as well be derived and a lot else more?

On the one hand it concerns the big search for a unified physical theory and on the other hand the chances of new technologies, which are connected with an extended field theory. As a necessary consequence of the derivation, which roots strictly in textbook physics and manages without postulate, scalar waves occur, which could be used manifold. In information technology they are suited as a carrier wave, which can be modulated moredimensionally, and in power engineering the spectrum stretches from the wireless transmission up to the collection of energy out of the field.

27.1 Energy out of the field

Neutrinos for instance are such field configurations, which move through space as a scalar wave. They were introduced by Pauli as massless but energy carrying particles to be able to fulfil the balance sheet of energy for the beta decay. Nothing would be more obvious than to technically use the neutrino radiation as an energy source.

But for a technical exploitation a useful model description of the particles and their interaction is imperative. For the sake of simplicity we imagine the neutrino to be an oscillating particle, which permanently oscillates back and forth between the state of an electron and that of a positron. With that the polarity changes from positive to negative and back again and the charge averaged over time is zero. Because of the change from a state of matter to the state of an anti-particle also next to no mass can be measured anymore.

A technical oscillator operated in resonance, which oscillates with the same frequency but opposite charge, will interact with the particle and build up an oscillating electromagnetic interaction, with which we already are familiar as the weak interaction in the proximity of a neutrino.

Tubular vortices

- Examples: drain vortex (bathtub vortex)
  whirlwind and waterspout, tornado (image).

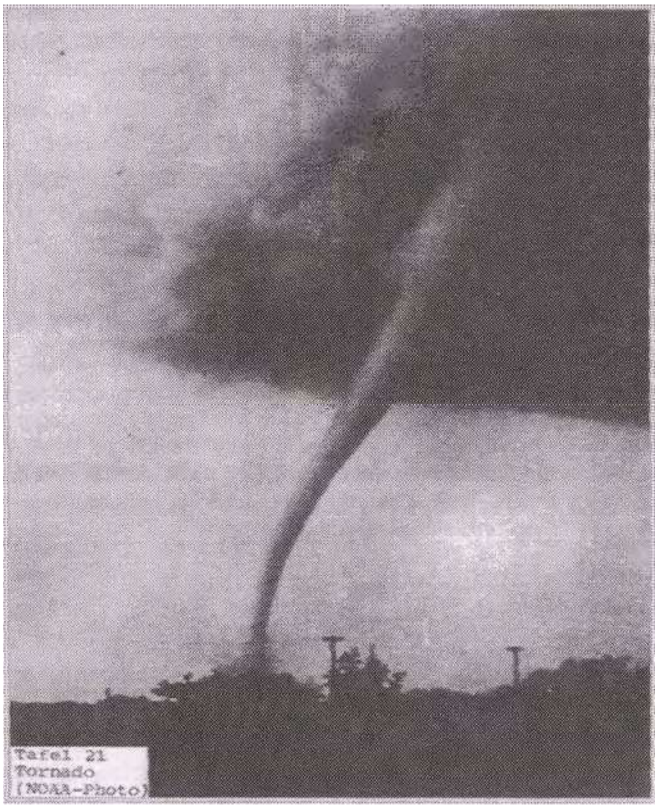

Inside:	expanding vortex
Outside:	contracting anti-vortex
Condition	for coming off: equally powerful vortices
Criterion:	viscosity
Result:	tubular structure

Fig. 27.2: ____Vortex and anti-vortex - a physical basic principle

<i>: Lugt: Wirbelstromung in Natur und Technik, G. Braun Verlag Karlsruhe 1979, Bild „Tornado" von Tafel 21, Seite 356

The propagation of particle radiation as a longitudinal shock wave however can't be described with the normally used field theory and the Maxwell equations, so that the field theory at this point must be reworked. Connected with this is the question of what is oscillating here, a question, which often is answered with an aether of whatever nature. I speak of field vortices and call the aether a property of the field. With that the set of difficulties is shifted into the domain of vortex physics.

27.2 Vortex and anti-vortex

In the eye of a tornado the same calm prevails as at great distance, because here a vortex and its anti-vortex work against each other. In the inside the expanding vortex is located and on the outside the contracting anti-vortex. One vortex is the condition for the existence of the other one and vice versa. Already Leonardo da Vinci knew both vortices and has described the dual manifestations (chapter 3.4).

In the case of flow vortices the viscosity determines the diameter of the vortex tube where the coming off will occur. If for instance a tornado soaks itself with water above the open ocean, then the contracting potential vortex is predominant and the energy density increases threateningly. If it however runs overland and rains out, it again becomes bigger and less dangerous.

The conditions for the bathtub vortex are similar. Here the expanding vortex consists of air, the contracting vortex however of water. In flow dynamics the relations are understood. They mostly can be seen well and observed without further aids.

In electrical engineering it's different: here field vortices remain invisible and not understood. Only so the Maxwell theory could find acceptance, although it only describes mathematically the expanding eddy current and ignores its anti-vortex. I call the contracting anti-vortex potential vortex" and point to the circumstance, that every eddy current entails the anti-vortex as a physical necessity.
Because the size of the forming structures is determined by the electric conductivity, in conducting materials the vortex rings, being composed of both vortices, are huge, whereas they can contract down to atomic dimensions in nonconductors. Only in semiconducting and resistive materials the structures occasionally can be observed directly (fig. 4.8).

Spherical vortices

Examples:	expanding vortex	contracting vortex
• quantum physics	collision processes (several quarks)	gluons (postulate!)
• nuclear physics	repulsion of like	strong interaction (postulate!)
• atomic physics	centrifugal force of the enveloping electrons	electrical attraction Schrodinger equation
• astro-physics	centrifugal force (inertia)	gravitation (can not be derived?!)

Example: elementary particles as electromagnetic field vortices

Inside:	expanding eddy current (skin effect)
Outside:	contracting anti-vortex (potential vortex)
Condition	for coming off: equally powerful vortices
Criterion:	electric conductivity (determines diameter)
Result:	spherical structure (consequence of the pressure of the vacuum)

Fig. 27.3: Spherical structures as a result of contracting potential vortices.<i>

Demokrit (460-370 BC) equated the vortex concept with „law of nature"!
= the first attempt to formulate a unified physics.

<i>: see Part 1, chapter 4.3

27.3 Vortices in the microcosm and macrocosm

The approximation, which is hidden in the Maxwell equations, thus consists of neglecting the anti-vortex dual to the eddy current. It is possible that this approximation is allowed, as long as it only concerns processes inside conducting materials. If we however get to insulating materials the Maxwell approximation will lead to considerable errors and it won't be able to keep it anymore.

If we take as an example the lightning and ask how the lightning channel is formed: Which mechanism is behind it, if the electrically insulating air for a short time is becoming a conductor? From the viewpoint of vortex physics the answer is obvious: The potential vortex, which in the air is dominating, contracts very strong and doing so squezes all air charge carriers and air ions, which are responsible for the conductivity, together at a very small space to form a current channel.

The contracting potential vortex thus exerts a pressure and with that forms the vortex tube. Besides the cylindrical structure another structure can be expected. It is the sphere, which is the only form, which can withstand a powerful pressure if that acts equally from all directions of space. Only think of ball lightning. Actually the spherical structure is mostly found in microcosm till macrocosm. Let's consider some examples and thereby search for the expanding and contracting forces (fig. 27.2).

- In quantum physics one imagines the elementary particles to be consisting of quarks. Irrespective of the question, which physical reality should be attributed to this model concept, one thing remains puzzling: The quarks should run apart, or you should try to keep together three globules, which are moving violently and permanently hitting each other. For this reason glue particles were postulated, the so-called gluons, which now should take care for the reaction force, but this reaction force is nothing but a postulate!

- In nuclear physics it concerns the force, which holds together the atomic nucleus, which is composed of many nucleons, and gives it the well-known great stability, although here like charged particles are close together. Particles, which usually repel each other. Between the theoretical model and practical reality there is an enormous gap, which should be overcome by introducing of a new reaction force. But also the nuclear force, called strong interaction, is nothing but a postulate!

- In atomic physics the electric force of attraction between the positive nuclear charge and the negatively charged enveloping electrons counteracts the centrifugal force. In this case the anti-vortex takes care for a certain structure of the atomic hull, which obey the Schrodinger equation as eigenvalue solutions. But also this equation irrespective of its efficiency until today purely is a mathematical postulate, as long as its origin is not clear.

- In astrophysics centrifugal force (expansion) as a result of the inertia and gravitation (contraction) as a result of the attraction of masses are balanced. But the "gravitation" puts itself in the way of every attempt to formulate a unified field theory. Also this time it is the contracting vortex, of which is said it can't be derived nor integrated.

Michael Faraday, the experimentor (1791-1867)

James Clerk Maxwell<i>, the mathematician (1831-1879)

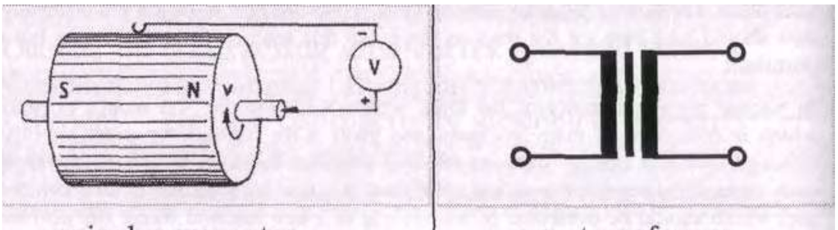

unipolar generator

e.g.: transformer

Fig. 27.4: ____The fathers of the law of induction

<i>: J.C. Maxwell: A treatise on Electricity and Magnetism, Dover Publications

It is remarkable how in the domain of the contracting vortex the postulates are accumulating. But this hasn't always been the case. In ancient Greece already 2400 years ago Demokrit has undertaken an attempt to formulate a unified physics. He traced all visible and observable structures in nature back to vortices, each time formed of vortex and anti-vortex. This phenomenon appeared him to be so fundamental, that he put the term "vortex" equal to the term for „law of nature". The term „atom" stems from Demokrit (460-370 BC).

Seen this way the physicists in ancient times already had been further than today's physics, which with the Maxwell approximation neglects the contracting vortex and with that excludes fundamental phenomena from the field description or is forced to replace them by model descriptions and numerous postulates.

What we need is a new field approach, which removes this flaw and in this point reaches over and above the Maxwell theory.

27.4 Faraday's law and Maxwell's formulation

In the choice of the approach the physicist is free, as long as the approach is reasonable and well founded. In the case of Maxwell's field equations two experimentally determined regularities served as basis: on the one hand Ampere's law and on the other hand the law of induction of Faraday. The mathematician Maxwell thereby gave the finishing touches for the formulations of both laws. He introduced the displacement current D and completed Ampere's law accordingly, and that without a chance of already at his time being able to measure and prove the measure. Only after his death this was possible experimentally, what afterwards makes clear the format of this man.

In the formulation of the law of induction Maxwell was completely free, because the discoverer Michael Faraday had done without specifications. As a man of practice and of experiment the mathematical notation was less important for Faraday. For him the attempts with which he could show his discovery of the induction to everybody, e.g. his unipolar generator, stood in the foreground.

His 40 years younger friend and professor of mathematics Maxwell however had something completely different in mind. He wanted to describe the light as an electromagnetic wave and doing so certainly the wave description of Laplace went through his mind, which needs a second time derivation of the field factor. Because Maxwell for this purpose needed two equations with each time a first derivation, he had to introduce the displacement current in Ampere's law and had to choose an appropriate notation for the formulation of the law of induction to get to the wave equation. His light theory initially was very controversial. Maxwell faster found acknowledgement for bringing together the teachings of electricity and magnetism and the representation as something unified and belonging together[i] than for mathematically giving reasons for the principle discovered by Faraday.

Nevertheless the question should be asked, if Maxwell has found the suitable formulation, if he has understood 100 percent correct his friend Faraday and his discovery. If discovery (from 29.08.1831) and mathematical formulation (1862) stem from two different scientists, who in addition belong to different disciplines, misunderstandings are nothing unusual. It will be helpful to work out the differences.

Law of induction

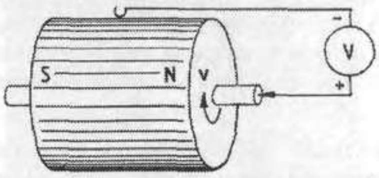

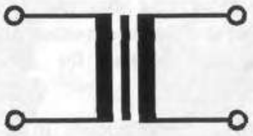

unipolar generator	e.g.: transformer
discovery of Faraday	2nd Maxwell equation
$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (27.1)	$\boxed{\text{rot } \mathbf{E} = - d\mathbf{B}/dt}$ (27.1*)
unipolar induction	law of induction generally

Fig. 27.5: Two formulations for one law
As a mathematical relation between
the vectors of the electric field strength **E**
and the induction **B** (= magnetic flux density)

Consequences
(the physically mean of the „field concept"):

• The **Faraday approach** describes a field physical principle	• **Maxwell equations** describe the fields of charged particles
• field = experience: We experience the electric as a magnetic field and vice versa as a result of a relative velocity	• the electric (**E**-) field = static field of charges, • the magnetic (**H**-) field = field of moving charges.
• Particles do not occur in the Faraday approach	• The origin of the particles remains unsettled (postulate)
• Perfect duality between **E**- and **H**-field.	• No duality in the field description
• **v** leads to dual permutation of **E**- and **H**-field.	• Charge carriers=el. monopoles • No magnetic monopoles!

27.5 The discovery of Faraday

If one turns an axially polarized magnet or a copper disc situated in a magnetic field, then perpendicular to the direction of motion and perpendicular to the magnetic field pointer a pointer of the electric field will occur, which everywhere points axially to the outside. In the case of this by Faraday developed unipolar generator hence by means of a brush between the rotation axis and the circumference a tension voltage can be called off<i>. The mathematically correct relation E = v x B I call Faraday-law, even if it only appears in this form in the textbooks later in time<ii>. The formulation usually is attributed to the mathematician Hendrik Lorentz, since it appears in the Lorentz force in exactly this form. Much more important than the mathematical formalism however are the experimental results and the discovery by Michael Faraday, for which reason the law concerning unipolar induction is named after the discoverer.

Of course we must realize that the charge carriers at the time of the discovery hadn't been discovered yet and the field concept couldn't correspond to that of today. The field concept was an abstracter one, free of any quantization.

That of course also is valid for the field concept advocated by Maxwell, which we now contrast with the „Faraday-law" (fig. 27.4). The second Maxwell equation, the law of induction (27.1*), also is a mathematical description between the electric field strength E and the magnetic induction B. But this time the two aren't linked by a relative velocity v. In that place stands the time derivation of B, with which a change in flux is necessary for an electric field strength to occur. As a consequence the Maxwell equation doesn't provide a result in the static or quasi-stationary case, for which reason it in such cases is usual, to fall back upon the unipolar induction according to Faraday (e.g. in the case of the Hall-probe, the picture tube, etc.). The falling back should only remain restricted to such cases, so the normally used idea. But with which right the restriction of the Faraday-law to stationary processes is made?

The vectors E and B can be subject to both spatial and temporal fluctuations. In that way the two formulations suddenly are in competition with each other and we are asked, to explain the difference, as far as such a difference should be present.

<i>: Part 2, INDEL 1996, Chap. 16.1

<ii>: among others in R.W. Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975, Seite 76 und 130

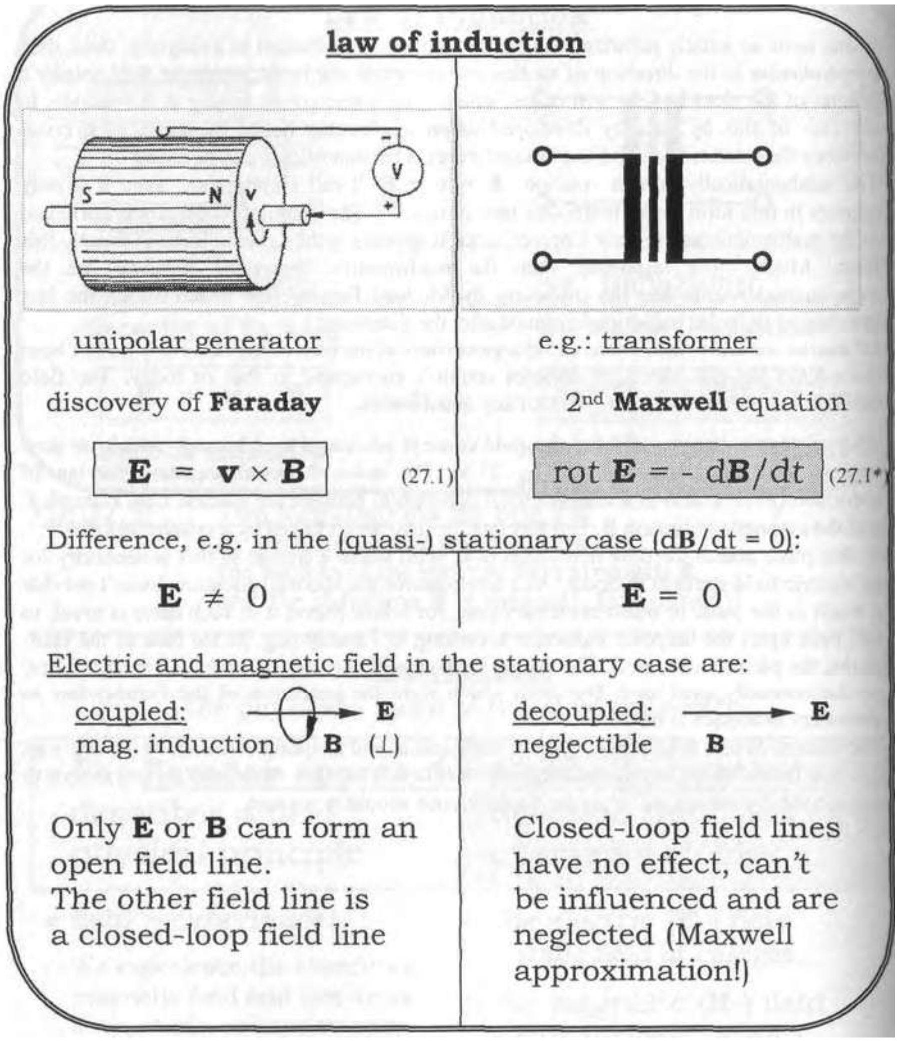

law of induction

unipolar generator	e.g.: transformer
discovery of **Faraday**	2nd **Maxwell** equation
$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (27.1)	$\boxed{\mathrm{rot}\,\mathbf{E} = -\,d\mathbf{B}/dt}$ (27.1*

Difference, e.g. in the (quasi-) stationary case ($d\mathbf{B}/dt = 0$):

$\mathbf{E} \neq 0$	$\mathbf{E} = 0$

Electric and magnetic field in the stationary case are:

coupled: $\longrightarrow$ **E** mag. induction **B** ($\perp$)	decoupled: $\longrightarrow$ **E** neglectible **B**
Only **E** or **B** can form an open field line. The other field line is a closed-loop field line	Closed-loop field lines have no effect, can't be influenced and are neglected (Maxwell approximation!)

Fig. 27.6: Law of induction according to Faraday or Maxwell?

27.6 Different formulation of the law of induction

Such a difference for instance is, that it is common practice to neglect the coupling between the fields at low frequencies. While at high frequencies in the range of the electromagnetic field the E- and the H-field are mutually dependent, at lower frequency and small field change the process of induction drops correspondingly according to Maxwell, so that a neglect seems to be allowed. Now electric or magnetic field can be measured independently of each other. Usually is proceeded as if the other field is not present at all.

That is not correct. A look at the Faraday-law immediately shows that even down to frequency zero always both fields are present. The field pointers however stand perpendicular to each other, so that the magnetic field pointer wraps around the pointer of the electric field in the form of a vortex ring in the case that the electric field strength is being measured and vice versa. The closed-loop field lines are acting neutral to the outside; they hence need no attention, so the normally used idea. It should be examined more closely if this is sufficient as an explanation for the neglect of the not measurable closed-loop field lines, or if not after all an effect arises from fields, which are present in reality.

Another difference concerns the commutability of E- and H-field, as is shown by the Faraday-generator, how a magnetic becomes an electric field and vice versa as a result of a relative velocity v. This directly influences the physical-philosophic question: What is meant by the electromagnetic field?

The textbook opinion based on the Maxwell equations names the static field of the charge carriers as cause for the electric field, whereas moving ones cause the magnetic field. But that hardly can have been the idea of Faraday, to whom the existence of charge carriers was completely unknown. The for his contemporaries completely revolutionary abstract field concept based on the works of the Croatian Jesuit priest Boscovich (1711-1778). In the case of the field it should less concern a physical quantity in the usual sense, than rather the experimental experience"' of an interaction according to his field description. We should interprete the Faraday-law to the effect that we experience an electric field, if we are moving with regard to a magnetic field with a relative velocity and vice versa.

In the commutability of electric and magnetic field a duality between the two is expressed, which in the Maxwell formulation is lost, as soon as charge carriers are brought into play. Is thus the Maxwell field the special case of a particle free field? Much evidence points to it, because after all a light ray can run through a particle free vacuum. If however fields can exist without particles, particles without fields however are impossible, then the field should have been there first as the cause for the particles. Then the Faraday description should form the basis, from which all other regularities can be derived. What do the textbooks say to that?

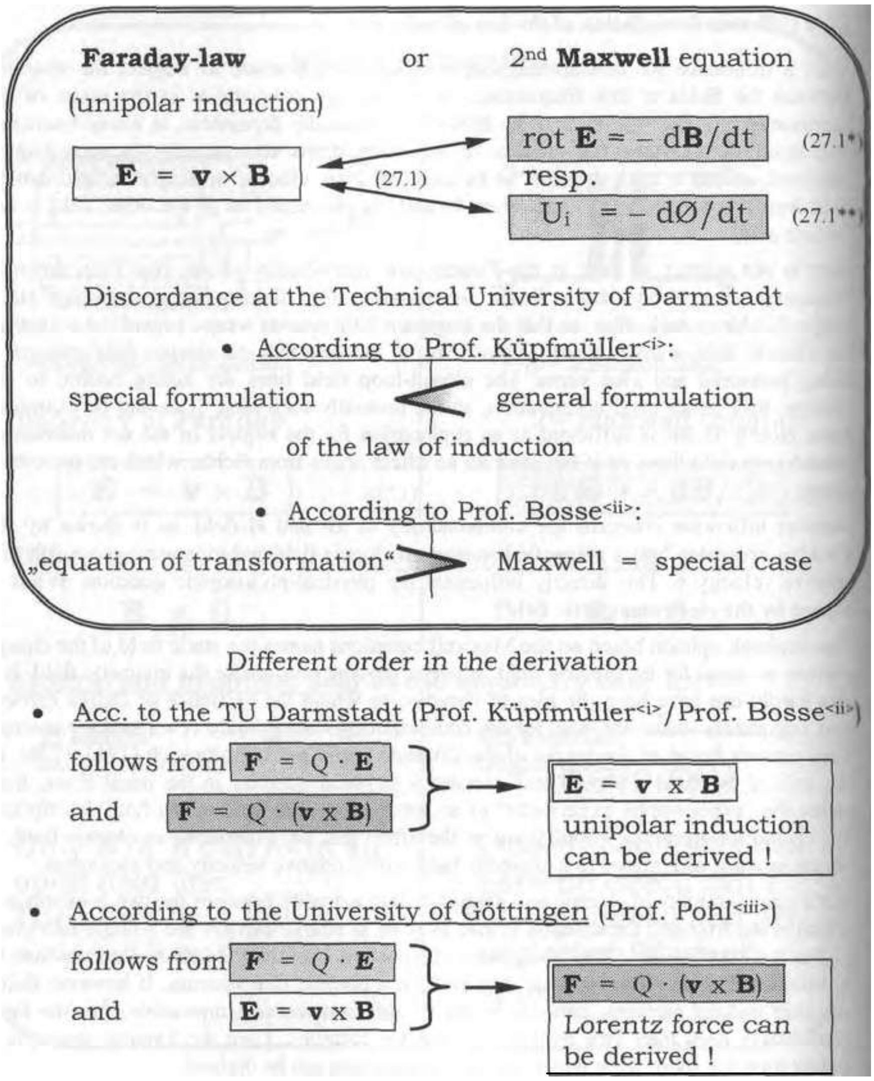
Faraday-law or 2nd Maxwell equation
(unipolar induction)
E = v × B
(27.1)
rot E = – dB/dt
(27.1*)
resp.
Ui = – dØ/dt
(27.1**)
Discordance at the Technical University of Darmstadt
• According to Prof. Küpfmüller<i>:
special formulation
general formulation
of the law of induction
• According to Prof. Bosse<ii>:
„equation of transformation"
Maxwell = special case
Different order in the derivation
• Acc. to the TU Darmstadt (Prof. Küpfmüller<i>/Prof. Bosse<ii>)
follows from F = Q · E
and F = Q · (v x B)
E = v x B
unipolar induction can be derived !
• According to the University of Göttingen (Prof. Pohl<iii>)
follows from F = Q · E
and E = v x B
F = Q · (v x B)
Lorentz force can be derived !

Fig. 27.7: Different opinions and derivations

<i>: K. Kupfmuller: Einfuhrung in die theoretische Elektrotechnik, 12. Auflage, Springer Verlag 1988, Seite 228, Gl. 22.

<ii>: G. Bosse: Grundlagen der Elektrotechnik II, BI-Hochschultaschenbucher Nr. 183, l.Aufl. 1967, Kap. 6.1 Induktion, Seite 58

<iii>: R. W. Pohl: Einfuhrung in die Physik, Band 2 Elektrizitatslehre, 21. Auflage, Springer-Verlag 1975, Seite 77

27.7 Contradictory opinions in textbooks

Obviously there exist two formulations for the law of induction (27.1 and 27.1*), which more or less have equal rights. Science stands for the question: which mathematical description is the more efficient one? If one case is a special case of the other case, which description then is the more universal one?

What Maxwell's field equations tell us is sufficiently known, so that derivations are unnecessary. Numerous textbooks are standing by, if results should be cited. Let us hence turn to the Faraday-law (27.1). Often one searches in vain for this law in schoolbooks. Only in more pretentious books one makes a find under the keyword "unipolar induction". If one however compares the number of pages, which are spent on the law of induction according to Maxwell with the few pages for the unipolar induction, then one gets the impression that the latter only is a unimportant special case for low frequencies. Kupfmuller speaks of a „special form of the law of induction"<i>, and cites as practical examples the induction in a brake disc and the Hall-effect. Afterwards Kupfmiiller derives from the „special form" the „general form" of the law of induction according to Maxwell, a postulated generalization, which needs an explanation. But a reason is not given<i>. Bosse gives the same derivation, but for him the Maxwell-result is the special case and not his Faraday approach<ii>! In addition he addresses the Faraday-law as equation of transformation and points out the meaning and the special interpretation. On the other hand he derives the law from the Lorentz force, completely in the style of Kupfmuller<i> and with that again takes it part of its autonomy. Pohl looks at that different. He inversely derives the Lorentz force from the Faraday-law<iii>.

By all means, the Faraday-law, which we want to base on instead of on the Maxwell equations, shows „strange effects<i4>" from the point of view of a Maxwell representative of today and thereby but one side of the medal (eq. 27.1). Only in very few distinguished textbooks the other side of the medal (eq. 27.2) is mentioned at all. In that way most textbooks mediate a lopsided and incomplete picture<i,ii,i4>. If there should be talk about equations of transformation, then the dual formulation belongs to it, then it concerns a pair of equations, which describes the relations between the electric and the magnetic field.

If the by Bosse<ii> prompted term „equation of transformation" is justified or not at first is unimportant. That is a matter of discussion.

<i>: K. Kupfmuller: Einfuhrung in die theoretische Elektrotechnik, 12.Aufl., Springer Verlag 1988, Seite 228, Gl. 22.

<ii>: G. Bosse: Grundlagen der Elektrotechnik II, BI-Hochschultaschenbucher Nr.183, l.Aufl. 1967, Kap. 6.1 Induktion, Seite 58

<iii>: R.W.Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 2 l.Aufl. Springer-Verlag 1975, Seite 77

<i4>: G. Lehner: Elektromagnetische Feldtheorie, Springer-Lehrbuch 1990, 1. Aufl., Seite 31 Kommentar zur Lorentzkraft (1.65)

The new and dual field approach consists of equations of transformation

of the electric and of the magnetic field[i]

$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (27.1) and $\boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}}$ (27.2)

unipolar induction equation of convection

- Formulation according to the rules of duality
  - Grimsehl[ii] speaks of the ,,equation of convection", according to which moving charges produce a magnetic field and so-called convection currents (referring to Rontgen 1885, Himstedt, Rowland 1876, Eichenwald and others)
- Pohl[iii] gives examples for the equations of transformation,
- he writes the equations beneath each other

$\boxed{\mathbf{E} = \mu \cdot \mathbf{v} \times \mathbf{H}}$ (27.3) and $\boxed{\mathbf{H} = -\varepsilon \cdot \mathbf{v} \times \mathbf{E}}$ (27.4)

- and points out that for $v = c = 1/\sqrt{\mu \cdot \varepsilon}$ one equation changes into the other one!

The new and dual approach roots in textbook physics!

Fig. 27.8: The new and dual field approach

---

[i]: see Part 1, chapter 6.5

[ii]: Grimsehl: Lehrbuch der Physik, 2.Bd., 17,Aufl. Teubner Verl. 1967, S. 130.

[iii]: R.W.Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975, Seite 77

27.8 The field-theoretical approach

The duality between E- and H-field and the commutability asks for a corresponding dual formulation to the Faraday-law (27.1). Written down according to the rules of duality there results an equation (27.2), which occasionally is mentioned in some textbooks. While both equations in the books of Pohl[i] and of Simonyi[ii] are written down side by side having equal rights and are compared with each other, Grimsehl[iii] derives the dual regularity (27.2) with the help of the example of a thin, positively charged and rotating metal ring. He speaks of „equation of convection", according to which moving charges produce a magnetic field and so-called convection currents. Doing so he refers to workings of Rontgen 1885, Himstedt, Rowland 1876, Eichenwald and many others more, which today hardly are known.

In his textbook also Pohl gives practical examples for both equations of transformation. He points out that one equation changes into the other one, if as a relative velocity v the speed of light c should occur. This question will also occupy us.

We now have found a field-theoretical approach with the equations of transformation, which in its dual formulation is clearly distinguished from the Maxwell approach. The reassuring conclusion is added: The new field approach roots entirely in textbook physics, as are the results from the literature research. We can completely do without postulates.

Next thing to do is to test the approach strictly mathematical for freedom of contradictions. It in particular concerns the question, which known regularities can be derived under which conditions. Moreover the conditions and the scopes of the derived theories should result correctly, e.g. of what the Maxwell approximation consists and why the Maxwell equations describe only a special case.

27.9 Derivation of Maxwell's field equations

As a starting-point and as approach serve the equations of transformation of the electromagnetic field, the Faraday-law of unipolar induction and the according to the rules of duality formulated law (eq. 27.1, 2). If we apply the curl to both sides of the equations then according to known algorithms of vector analysis the curl of the cross product each time delivers the sum of four single terms. Two of these again are zero for a non-accelerated relative motion in the x-direction with v = dr/dt.

One term concerns the vector gradient (v grad)B, which can be represented as a tensor. By writing down and solving the accompanying derivative matrix giving consideration to the above determination of the v-vector, the vector gradient becomes the simple time derivation of the field vector B(r(t)) (eq. 27.10, according to the rule of eq. 27.11).

<i>: R.W.Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975, Seite 76 und 130

<ii>: K. Simonyi: Theoretische Elektrotechnik, 7.Aufl. VEB Berlin 1979, Seite 924.

<iii>: Grimsehl: Lehrbuch der Physik, 2,Bd., 17.Aufl. Teubner Verl. 1967, S. 130.

As approach serve the equations of transformation (fig. 27.5) of the electric and of the magnetic field:

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} \text{ (27.1) and } \boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}} \quad (27.2)$$

If we apply the curl to the respective cross product:

$$\boxed{\operatorname{rot} \mathbf{E} = \operatorname{rot} (\mathbf{v} \times \mathbf{B})} \text{ (27.5) and } \boxed{\operatorname{rot} \mathbf{H} = -\operatorname{rot} (\mathbf{v} \times \mathbf{D})} \quad (27.6)$$

then according to the algorithms<i> four sum terms are delivered:

$$\boxed{\operatorname{rot} \mathbf{E} = (\mathbf{B} \operatorname{grad})\mathbf{v} - (\mathbf{v} \operatorname{grad})\mathbf{B} + \mathbf{v} \operatorname{div} \mathbf{B} - \mathbf{B} \operatorname{div} \mathbf{v}} \quad (27.5)$$

$$\boxed{\operatorname{rot} \mathbf{H} = -[(\mathbf{D} \operatorname{grad})\mathbf{v} - (\mathbf{v} \operatorname{grad})\mathbf{D} + \mathbf{v} \operatorname{div} \mathbf{D} - \mathbf{D} \operatorname{div} \mathbf{v}]} \quad (27.6)$$

where 2 of them are zero because of: $\boxed{\mathbf{v}(t) = d\mathbf{r}/dt}$ (27.7)

- the divergence of v(t) disappears: $\operatorname{div} v = 0$, (27.8)
- and will be zero as well: $\partial \mathbf{v}(t)/\partial \mathbf{r} = \boxed{\operatorname{grad} \mathbf{v} = \mathbf{0}}$. (27.9)
- there remain the vector gradients:

$$\boxed{(\mathbf{v} \operatorname{grad}) \mathbf{B} = \frac{d\mathbf{B}}{dt}} \text{ and } \boxed{(\mathbf{v} \operatorname{grad}) \mathbf{D} = \frac{d\mathbf{D}}{dt}} \text{ ,(27.10)}$$

- according to the rules<i> in general (with eq. 27.7):

$$\frac{d\mathbf{V}(\mathbf{r}(t))}{dt} = \frac{\partial \mathbf{V}(\mathbf{r}=\mathbf{r}(t))}{\partial \mathbf{r}} \cdot \frac{d\mathbf{r}(t)}{dt} = (\mathbf{v} \operatorname{grad}) \mathbf{V} \quad (27.11)$$

- A comparison of the coefficients of both field equations

$$\boxed{\begin{aligned} \operatorname{rot} \mathbf{E} &= -d\mathbf{B}/dt + \mathbf{v} \operatorname{div} \mathbf{B} = -d\mathbf{B}/dt - \mathbf{b} && (27.12) \\ \operatorname{rot} \mathbf{H} &= d\mathbf{D}/dt - \mathbf{v} \operatorname{div} \mathbf{D} = d\mathbf{D}/dt + \mathbf{j} && (27.13) \end{aligned}}$$

with the Maxwell equations results in:

- for the potential density $b = -v \operatorname{div} B = 0$, (27.14)
  (eq. 27.12 = law of induction, if $b = 0$ resp. $\operatorname{div} B = 0$)

- for the current density $j = -v \operatorname{div} D = -v_x \cdot \rho_{el}$, (27.1a
  (eq. 27.13 = Ampere's law, if j = with v moving negative charge carriers (ρ_{el}= electric space charge density).

Fig. 27.9: ____ Derivation of Maxwell's field equations as a special case of the equations of transformation

---

<i>: Bronstein u.a.: Taschenbuch der Mathematik, 4.Neuaufl. Thun 1999, S. 652

For the last not yet explained terms at first are written down the vectors b and j as abbreviation. With equation 27.13 we in this way immediately look at the well-known law of Ampere (1st Maxwell equation). The comparison of coefficients (27.15) in addition delivers a useful explanation to the question, what is meant by the current density j: it is a space charge density ρ_{el} consisting of negative charge carriers, which moves with the velocity v for instance through a conductor (in the x-direction).

The current density j and the to that dual potential density b mathematically seen at first are nothing but alternative vectors for an abbreviated notation. While for the current density j the physical meaning already could be clarified from the comparison with the law of Ampere, the interpretation of the potential density b still is due. From the comparison with the law of induction (eq. 27.1*) we merely infer, that according to the Maxwell theory this term is assumed to be zero. But that is exactly the Maxwell approximation and the restriction with regard to the new and dual field approach, which roots in Faraday.

In that way also the duality gets lost with the argument that magnetic monopoles (div B) in contrast to electric monopoles (div D) do not exist and until today could evade every proof. It thus is overlooked that div D at first describes only eddy currents and div B only the necessary anti-vortex, the potential vortex. Spherical particles, like e.g. charge carriers presuppose both vortices: on the inside the expanding (div D) and on the outside the contracting vortex (div B), which then necessarily has to be different from zero, even if there hasn't yet been searched for the vortices dual to eddy currents, which are expressed in the neglected term.

Assuming, a monopole concerns a special form of a field vortex, then immediately gets clear, why the search for magnetic poles has to be a dead end and their failure isn't good for a counterargument: The missing electric conductivity in vacuum prevents current densities, eddy currents and the formation of magnetic monopoles. Potential densities and potential vortices however can occur. As a result can without exception only electrically charged particles be found in the vacuum (derivation in chapter 4.2 till 4.4). Because vortices are more than monopole-like structures depending on some boundary conditions, only the vortex description will be pursued further consequently.

Let us record: Maxwell's field equations can directly be derived from the new dual field approach under a restrictive condition. Under this condition the two approaches are equivalent and with that also error free. Both follow the textbooks and can so to speak be the textbook opinion.

The restriction (b = 0) surely is meaningful and reasonable in all those cases in which the Maxwell theory is successful. It only has an effect in the domain of electrodynamics. Here usually a vector potential A is introduced and by means of the calculation of a complex dielectric constant a loss angle is determined. Mathematically the approach is correct and dielectric losses can be calculated. Physically however the result is extremely questionable, since as a consequence of a complex s a complex speed of light would result (according to the definition $c \equiv 1/\sqrt{\varepsilon \cdot \mu}$). With that electrodynamics offends against all specifications of the textbooks, according to which c is constant and not variable and less then ever complex.

But if the result of the derivation physically is wrong, then something with the approach is wrong, then the fields in the dielectric perhaps have an entirely other nature, then dielectric losses perhaps are vortex losses of potential vortices falling apart?

- Maxwell's field equations:

$$\mathrm{rot}\,\mathbf{E} = -\,d\mathbf{B}/dt \quad \text{(law of induction)} \qquad (27.1^*)$$
$$\mathrm{rot}\,\mathbf{H} = d\mathbf{D}/dt + \mathbf{j} \quad \text{(Ampère's law)} \qquad (27.13)$$

- describe the special case for b = 0 resp. div B = 0

$$\mathrm{rot}\,\mathbf{E} = -\,d\mathbf{B}/dt + \mathbf{v}\,\mathrm{div}\,\mathbf{B} = -\,d\mathbf{B}/dt - \mathbf{b} \qquad (27.12)$$
$$\mathrm{rot}\,\mathbf{H} = d\mathbf{D}/dt - \mathbf{v}\,\mathrm{div}\,\mathbf{D} = d\mathbf{D}/dt + \mathbf{j} \qquad (27.13)$$

The physical meaning of the introduced abbreviations b and j is:

- the current density $\mathbf{j} = -\,\mathbf{v}\,\mathrm{div}\,\mathbf{D} = -\,v_x \cdot \rho_{el}$, (27.15)
- with Ohm's law $\mathbf{j} = \sigma \cdot \mathbf{E} = \mathbf{D}/\tau_1$, and (27.16)
- the potential density $\mathbf{b} = -\,\mathbf{v}\,\mathrm{div}\,\mathbf{B} = \mathbf{B}/\tau_2$,(27.17)
- with the eddy current time constant $\tau_1 = \varepsilon/\sigma$ (27.16*>
- and with the potential vortex time constant τ_2

The complete field equations (27.12 and 27.13) read, with the time constants (τ_1 and τ_2)of the respective field vortex:

- completely extended law of induction (with $B = \mu \cdot H$): (27.18)

$$\mathrm{rot}\,\mathbf{E} = -\,d\mathbf{B}/dt - \mathbf{B}/\tau_2 = -\,\mu \cdot (d\mathbf{H}/dt + \mathbf{H}/\tau_2) \qquad (27.20)$$

- and the well-known law of Ampere (with $D = \varepsilon \cdot E$): (27.19)

$$\mathrm{rot}\,\mathbf{H} = d\mathbf{D}/dt + \mathbf{D}/\tau_1 = \varepsilon \cdot (d\mathbf{E}/dt + \mathbf{E}/\tau_1) \qquad (27.21)$$

Fig. 27.10: The extension of the law of induction for vortices of the electric field (potential vortices].<i>

<i>: see also fig. 5.1

27.10 Derivation of the potential vortices

Is the introduction of a vector potential A in electrodynamics a substitute of neglecting the potential density b? Do here two ways mathematically lead to the same result? And what about the physical relevance? After classic electrodynamics being dependent on working with a complex constant of material, in what is buried an unsurmountable inner contradiction, the question is asked for the freedom of contradictions of the new approach. At this point the decision will be made, if physics has to make a decision for the more efficient approach, as it always has done when a change of paradigm had to be dealt with. The abbreviations j and b are further transformed, at first the current density in Ampere's law j = $-v_x \cdot \rho_{el}$ (27.15), as the movement of negative electric charges. By means of Ohm's law j=$\sigma \cdot$E and the relation of material D=$\varepsilon \cdot$E the current density j also can be written down as dielectric displacement current with the characteristic relaxation time constant $\tau_1 = \varepsilon/\sigma$ (eq. 27.16) for the eddy currents. In this representation of the law of Ampere (eq. 27.21) clearly is brought to light, why the magnetic field is a vortex field, and how the eddy currents produce heat losses depending on the specific electric conductivity σ. As one sees we, with regard to the magnetic field description, move around completely in the framework of textbook physics.

Let us now consider the dual conditions. The comparison of coefficients (eq. 27.12 + 27.17) looked at purely formal, results in a potential density b in duality to the current density j, which with the help of an appropriate time constant τ_2 founds vortices of the electric field. I call these potential vortices (in eq. 27.20).

In contrast to that the Maxwell theory requires an irrotationality of the electric field, which is expressed by taking the potential density b and the divergence B equal to zero. The time constant τ_2 thereby tends towards infinity. This Maxwell approximation leads to the circumstance that with the potential vortices of the electric field also their propagation as a scalar wave gets lost, so that the Maxwell equations describe only transverse and no longitudinal waves. At this point there can occur contradictions for instance in the case of the near-field of an antenna, where longitudinal wave parts can be detected measuring technically, and such parts already are used technologically in transponder systems e.g. as installations warning of theft in big stores.

It is denominating, how they know how to help oneself in the textbooks of high-frequency technology in the case of the near-field zone<i>. Proceeding from the Maxwell equations the missing potential vortex is postulated without further ado, by means of the specification of a ,,standing wave" in the form of a vortex at a dipole antenna. With the help of the postulate now the longitudinal wave parts are ,,calculated", like they also are being measured, but also like they wouldn't occur without the postulate as a result of the Maxwell approximation.

There isn't a way past the potential vortices and the new dual approach, because no scientist is able to afford to exclude already in the approach a possibly authoritative phenomenon, which he wants to calculate physically correct!

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3. Auflage 1986 Springer-Verlag Berlin, Seite 335

- Under the assumption: $E = E(r,t)$; $H = H(r,t)$,
- using the relations of material:

  $\boxed{\mathbf{B} = \mu \cdot \mathbf{H}}$ and $\boxed{\mathbf{D} = \varepsilon \cdot \mathbf{E}}$:
- the complete and extended law of induction reads:

  $$\operatorname{rot} \mathbf{E} = -\partial \mathbf{B}/\partial t - \mathbf{B}/\tau_2 = -\mu \cdot (\partial \mathbf{H}/\partial t + \mathbf{H}/\tau_2) \quad (27.20)$$
- and the well-known law of Ampere:

  $$\operatorname{rot} \mathbf{H} = \partial \mathbf{D}/\partial t + \mathbf{D}/\tau_1 = \varepsilon \cdot (\partial \mathbf{E}/\partial t + \mathbf{E}/\tau_1) \quad (27.21)$$

---

if we again apply the curl operation to eq. 27.20 and insert eq. 27.21:

$$-\operatorname{rot}\operatorname{rot} \mathbf{E} = \mu \cdot \partial(\operatorname{rot}\mathbf{H})/\partial t + (\mu/\tau_2)\cdot(\operatorname{rot} \mathbf{H}) \quad (27.22)$$

$$= \mu \cdot \varepsilon \cdot [\partial^2 \mathbf{E}/\partial t^2 + (1/\tau_1)\cdot\partial \mathbf{E}/\partial t + (1/\tau_2)\cdot\partial \mathbf{E}/\partial t + \mathbf{E}/\tau_1\tau_2] \quad (27.23)$$

$$= (1/c^2)\cdot[\partial^2 \mathbf{E}/\partial t^2 + (1/\tau_1 + 1/\tau_2)\cdot\partial \mathbf{E}/\partial t + \mathbf{E}/\tau_1\tau_2] \quad (27.24)$$

with the definition for the speed of light c:

$$\boxed{\varepsilon \cdot \mu = 1/c^2} \quad (27.25)$$

the fundamental field equation reads:

$$\underbrace{-c^2 \cdot \operatorname{rot}\operatorname{rot} \mathbf{E}}_{a} = \underbrace{\partial^2 \mathbf{E}/\partial t^2}_{b} + \quad (27.26)$$

$$\text{(electromagnetic wave)} +$$

$$+ \underbrace{(1/\tau_1)\cdot\partial \mathbf{E}/\partial t}_{c} + \underbrace{(1/\tau_2)\cdot\partial \mathbf{E}/\partial t}_{d} + \underbrace{\mathbf{E}/\tau_1\tau_2}_{e}$$

$$+ \text{ eddy current } + \text{ potential vortex } + I/U$$

Fig. 27.11: Derivation of the fundamental field equation from the equations of transformation of the electromagnetic field.<i>

<i>: The fundamental field equation mathematically describes a wave damped with the vortices of the electric and the vortices of the magnetic field. It is formulated only in space and time. From it can be deduced numerous eigenvalue equations, (i.e. the equation of Schrodinger, fig. 5.1).

27.11 Derivation of the „fundamental field equation"

The two equations of transformation and also the from that derived field equations (27.20 and 27.21) show the two sides of a medal, by mutually describing the relation between the electric and magnetic field strength (between E and H). We get on the track of the meaning of the „medal" itself, by inserting the dually formulated equations into each other. If the calculated H-field from one equation is inserted into the other equation then as a result a determining equation for the E-field remains. The same vice versa also functions to determine the H-field. Since the result formally is identical and merely the H-field vector appears at the place of the E-field vector and since it equally remains valid for the B-, the D-field and all other known field factors, the determining equation is more than only a calculation instruction. It reveals a fundamental physical principle. I call it the "fundamental field equation".

The derivation always is the same: If we again apply the curl operation to rot E (law of induction 27.20) also the other side of the equation should be subjected to the curl. If for both terms rot H is expressed by Ampere's law 27.21, then in total four terms are formed (27.26): the wave equation (a-b) with the two damping terms, on the one hand the eddy currents (a-c) and on the other hand the potential vortices (a-d) and as the fourth term the Poisson equation (a-e), which is responsible for the spatial distribution of currents and potentials<i>.

Not in a single textbook a mathematical linking of the Poisson equation with the wave equation can be found, as we here succeed in for the first time. It however is the prerequisite to be able to describe the conversion of an antenna current into electromagnetic waves near a transmitter and equally the inverse process, as it takes place at a receiver. Numerous model concepts, like they have been developed by HF- and EMC-technicians as a help, can be described mathematically correct by the physically founded field equation.

In addition further equations can be derived, for which this until now was supposed to be impossible, like for instance the Schrodinger equation (chapter 5.6-5.9). This contrary to current opinion isn't a wave equation at all, since the term (b) with the second time derivation is missing. As diffusion equation it has the task to mathematically describe field vortices and their structures.

As a consequence of the Maxwell equations in general and specifically the eddy currents not being able to form structures, every attempt has to fail, which wants to derive the Schrodinger equation from the Maxwell equations.

The fundamental field equation however contains the newly discovered potential vortices, which owing to their concentration effect (in duality to the skin effect) form spherical structures, for which reason these occur as eigenvalues of the equation. For these eigenvalue-solutions numerous practical measurements are present, which confirm their correctness and with that have probative force with regard to the correctness of the new Held approach and the fundamental field equation. By means of the pure formulation in space and time and the interchangeability of the field pointers here a physical principle is described, which fulfills all requirements, which a world equation must meet.

<i>: see also fig. 5.1

Comparison:

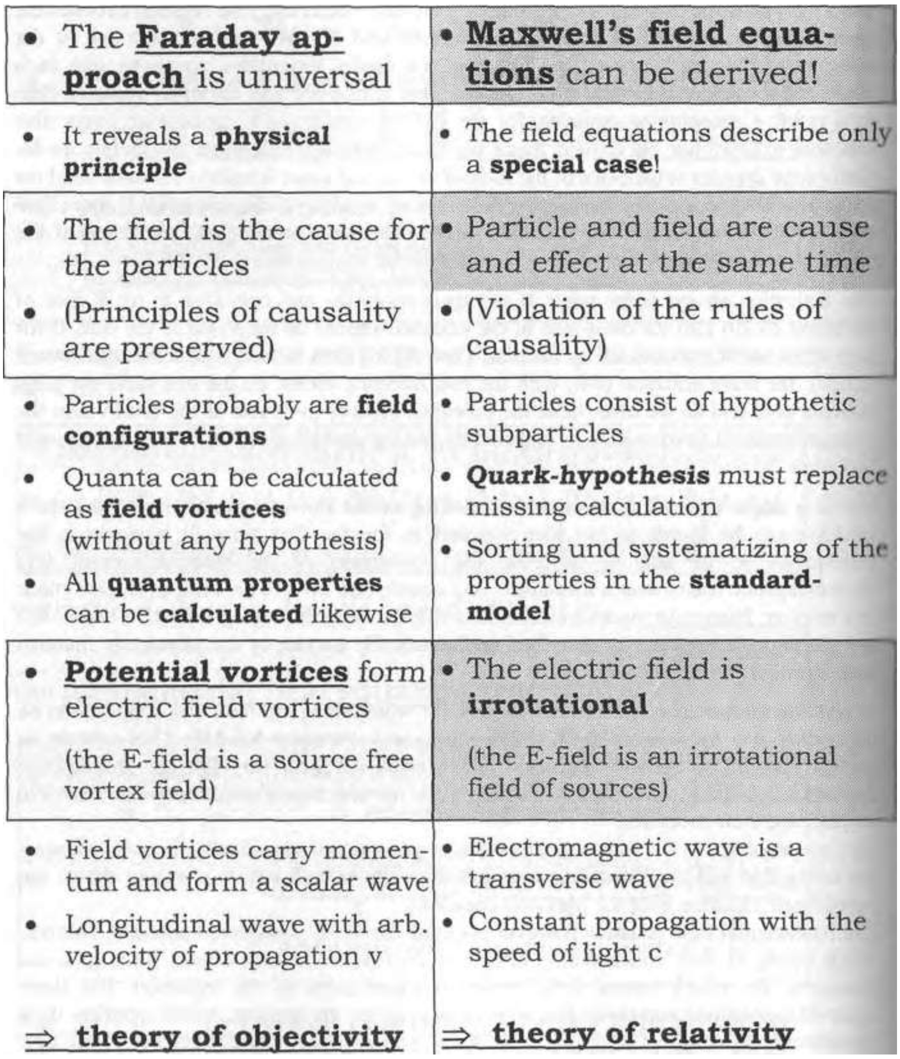

• The **Faraday approach** is universal	• **Maxwell's field equations** can be derived!
• It reveals a **physical principle**	• The field equations describe only a **special case**!
• The field is the cause for the particles • (Principles of causality are preserved)	• Particle and field are cause and effect at the same time • (Violation of the rules of causality)
• Particles probably are **field configurations** • Quanta can be calculated as **field vortices** (without any hypothesis) • All **quantum properties** can be **calculated** likewise	• Particles consist of hypothetic subparticles • **Quark-hypothesis** must replace missing calculation • Sorting und systematizing of the properties in the **standard-model**
• **Potential vortices** form electric field vortices (the E-field is a source free vortex field)	• The electric field is **irrotational** (the E-field is an irrotational field of sources)
• Field vortices carry momentum and form a scalar wave • Longitudinal wave with arb. velocity of propagation v	• Electromagnetic wave is a transverse wave • Constant propagation with the speed of light c
⇒ **theory of objectivity**	⇒ **theory of relativity**

Fig. 27.12: Comparison of the field-theoretical approaches according to Faradav and according to Maxwell.

27.12 The Maxwell field as a derived special case

As the derivations show, nobody can claim there wouldn't exist potential vortices and no propagation as a scalar wave, since only the Maxwell equations are to blame that these already have been factored out in the approach. One has to know that the field equations, and may they be as famous as they are, are nothing but a special case, which can be derived.

The field-theoretical approach however, which among others bases on the Faraday-law, is universal and can't be derived on its part. It describes a physical basic principle, the alternating of two dual experience or observation factors, their overlapping and mixing by continually mixing up cause and effect. It is a philosophic approach, free of materialistic or quantum physical concepts of any particles.

Maxwell on the other hand describes without exception the fields of charged particles, the electric field of resting and the magnetic field as a result of moving charges. The charge carriers are postulated for this purpose, so that their origin and their inner structure remain unsettled and can't be derived. The subdivision e.g. in quarks stays in the domain of a hypothesis, which can't be proven. The sorting and systematizing of the properties of particles in the standard-model is nothing more than unsatisfying comfort for the missing calculability.

With the field-theoretical approach however the elementary particles with all quantum properties can be calculated as field vortices (chap. 7). With that the field is the cause for the particles and their measurable quantization. The electric vortex field, at first source free, is itself forming its field sources in form of potential vortex structures. The formation of charge carriers in this way can be explained and proven mathematically, physically, graphically and experimentally understandable according to the model.

Where in the past the Maxwell theory has been the approach, there in the future should be proceeded from the equations of transformation of the field-theoretical approach. If now potential vortex phenomena occur, then these also should be interpreted as such in the sense of the approach and the derivation, then the introduction and postulation of new and decoupled model descriptions isn't allowed anymore, like the near-field effects of an antenna, the noise, dielectric capacitor losses, the mode of the light and a lot else more.

The at present in theoretical physics normal scam of at first putting a phenomenon to zero, to afterwards postulate it anew with the help of a more or less suitable model, leads to a breaking up of physics into apparently not connected individual disciplines and an inefficient specialisthood. There must be an end to this now! The new approach shows the way towards a unified theory, in which the different areas of physics again fuse to one area. In this lies the big chance of this approach, even if many of the specialists at first should still revolt against it.

This new and unified view of physics shall be summarized with the term "theory of objectivity". As we shall derive, it will be possible to deduce the theory of relativity as a partial aspect of it (chapter 6 and 28).

Let us first cast our eyes over the wave propagation.

- Starting-point: the fundamental field equation

$$-c^2\cdot\text{rot rot }\mathbf{B} = \frac{d^2\mathbf{B}}{dt^2} + \frac{1}{\tau_2}\frac{d\mathbf{B}}{dt} + \frac{1}{\tau_1}\frac{d\mathbf{B}}{dt} + \frac{\mathbf{B}}{\tau_1\tau_2} \qquad (27.26^*)$$

- with a magnetic flux density $B = B(r(t))$.

1st condition for eq. 27.26*:
the special case, if $\sigma=0$ and $\sigma/\varepsilon = \boxed{(1/\tau_1) = 0}$. (27.16*)

The remaining vortex term is transformed by applying already used relations (eq. 27.10 and eq. 27.17):

$$\frac{1}{\tau_2}\cdot\frac{d\mathbf{B}}{dt} = \mathbf{v}\cdot\text{grad}\,\frac{\mathbf{B}}{\tau_2} \quad (27.10) \qquad \frac{\mathbf{B}}{\tau_2} = -\,\mathbf{v}\cdot\text{div}\,\mathbf{B} \;.\quad (27.17)$$

If the velocity of propagation: $v = (v_x, v_y=0, v_z=0)$; $v = dx/dt$, then the simplified field equation (if the coordinates are orientated on the vector of velocity) results in the general wave equation (involved with the x-component) in the form:

$$\|\mathbf{v}\|^2\,\text{grad div }\mathbf{B} - c^2\,\text{rot rot }\mathbf{B} = d^2\mathbf{B}/dt^2 \qquad (27.27)$$

longitudinal with v = arbitrary	transverse with c = const.	wave velocity of propagation

---

2nd condition for eq. 27.28: $v = c$.

The wave equation in the usual notation (= inhomogeneous Laplace equation, = purely a special case!) now reads:

$$\Delta\mathbf{B} = \text{grad div }\mathbf{B} - \text{rot rot }\mathbf{B} = (1/c^2)\cdot d^2\mathbf{B}/dt^2 \qquad (27.28)$$

Fig. 27.13: Derivation of the wave equations (inhomogeneous Laplace equation) as a special case of the equations of transformation of the electromagnetic field.

27.13 Derivation of the wave equation

The first wave description, model for the light theory of Maxwell, was the inhomogeneous Laplace equation:

$$\boxed{\Delta\mathbf{E}\cdot c^2 = d^2\mathbf{E}/dt^2} \quad \text{with} \quad \boxed{\Delta\mathbf{E} = \text{grad div } \mathbf{E} - \text{rot rot } \mathbf{E}} \qquad (27.28^*)$$

There are asked some questions:

- Can also this mathematical wave description be derived from the new approach?
- Is it only a special case and how do the boundary conditions read?
- In this case how should it be interpreted physically?
- Are new properties present, which can lead to new technologies?

Starting-point is the fundamental field equation (27.26). We thereby should remember the interchangeability of the field pointers, that the equation doesn't change its form, if it is derived for H, for B, for D or any other field factor instead of for the E-field pointer. This time we write it down for the magnetic induction B and consider the special case, that we are located in a badly conducting medium, as is usual for the wave propagation in air. But with the electric conductivity σ also $1/\tau_1 = \sigma/\varepsilon$ tends towards zero (eq. 27.16*). With that the eddy currents and their damping and other properties disappear from the field equation, what also makes sense. There remains the potential vortex term $(1/\tau_2)$*dB/dt , which using the already introduced relations (eq. 27.10 and 27.17) involved with an in x-direction propagating wave ($v = (v_x, v_y = 0, v_z = 0)$) can be transformed directly into:

$$-\|\mathbf{v}\|^2\cdot\text{grad div } \mathbf{B}$$

The divergence of a field vector (here B) mathematically seen is a scalar, for which reason this term as part of the wave equation founds so-called „scalar waves" and that means that potential vortices, as far as they exist, will appear as a scalar wave. We at this point tacitly anticipate chapter 28, which provides the reason for the speed of light losing its vectorial nature, if it is correlated with itself. This insight however is valid in general for all velocities ($v = dr/dt$), so that in the same way a scalar descriptive factor can be used for the velocity ($v = dx/dt$) as for c.

From the simplified field equation (27.26*) the general wave equation (27.27) can be won in the shown way, divided into longitudinal and transverse wave parts, which however can propagate with different velocity.

Physically seen the vortices have particle nature as a consequence of their structure forming property. With that they carry momentum, which puts them in a position to form a longitudinal shock wave similar to a sound wave. If the propagation of the light one time takes place as a wave and another time as a particle, then this simply and solely is a consequence of the wave equation. Light quanta should be interpreted as evidence for the existence of scalar waves. Here however also occurs the restriction that light always propagates with the speed of light. It concerns the special case $v = c$. With that the derived wave equation (27.27) changes into the inhomogeneous Laplace equation (27.28). The electromagnetic wave in both cases is propagating with c. As a transverse wave the field vectors are standing perpendicular to the direction of propagation. The velocity of propagation therefore is decoupled and constant. Completely different is the case for the longitudinal wave. Here the propagation takes place in the direction of an oscillating field pointer, so that the phase velocity permanently is changing and merely an average group velocity can be given for the propagation. There exists no restriction for v and $v = c$ only describes a special case.

• From the dual field-	From Maxwell's field equations can be
theoretical approach are derived:	derived:
=> Maxwell's field equations	=> 0
=> the wave equation (with transverse and longitudinal parts)	=> only transverse waves (no longitudinal waves)
=> scalar waves (Tesla-/neutrino radiation)	=> 0 (no scalar waves)
=> vortex and anti-vortex (current eddy and potential vortex)	=> only eddy currents
=> Schrodinger equation (basic equation of chemistry)	=>0
=> Klein-Gordon equation (basic eq. of nuclear physics)	=>0

Fig. 27.14: Comparison of the efficiency of both approaches.

(as an interim result, if it concerns the question, which approach of the two is the more efficient one and which one better should be discarded. The final balance is made in chapter 28).
It here concerns partial aspects of the following theories:

=> theory of objectivity	=> theory of relativity

27.14 The new field approach in synopsis

Proof could be furnished that an approximation is buried in Maxwell's field equations and that they merely represent the special case of a new, dually formulated and more universal approach. The mathematical derivations of the Maxwell field and the wave equation disclose, of what the Maxwell approximation consists. The anti-vortex dual to the expanding eddy current with its skin effect is neglected. This contracting anti-vortex is called potential vortex. It is capable of forming structures and propagates as a scalar wave in longitudinal manner in badly conducting media like air or vacuum.

At relativistic velocities the potential vortices are subject to the Lorentz contraction. Since for scalar waves the propagation occurs longitudinally in the direction of an oscillating field pointer, the potential vortices experience a constant oscillation of size as a result of the oscillating propagation. If one imagines the field vortex as a planar but rolled up transverse wave, then from the oscillation of size and with that of wavelength at constant swirl velocity with c follows a continual change in frequency, which is measured as a noise signal.

The noise proves to be the in the Maxwell equations neglected potential vortex term, which founds scalar waves. If at biological or technical systems, e.g. at antennas a noise signal is being measured, then that proves the existence of potential vortices, but it then also means that the scope of the Maxwell theory has been exceeded and erroneous concepts can be the result.

As an answer to the question about possible new technologies is pointed to two special properties.

1st potential vortices for reason of their particle nature carry momentum and energy. Since we are surrounded by noise vortices, an energy technical use of scalar waves would be feasible, where the noise power is withdrawn of the surroundings. There is evidence that biological systems in nature cover their need for energy in this way. But at least an energy transmission with scalar waves already would be a significant progress with regard to the alternating current technology of today.

2nd the wavelength multiplied with the frequency results in the velocity of propagation v of a wave $(\lambda \cdot f = v)$, and that for scalar waves by no means is constant. With that wavelength and frequency aren't coupled anymore; they can be modulated separately, for which reason for scalar waves a whole dimension can be modulated additionally compared to the Hertzian wave. In that the reason can be seen, why the human brain with just 10 Hz clock frequency is considerably more efficient than modern computers with more than 1 GHz clock frequency. Nature always works with the best technology, even if we haven't yet understood it.

If we would try to learn of nature and an energy technical or an information technical use of scalar waves would occur, then probably nobody wanted to have our today still highly praised technology anymore. In the course of the greenhouse gases and the electrosmog we have no other choice than to scientifically occupy us with scalar waves and their technical use.

Equations of transformation

of the electric and of the magnetic field<i>

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} \quad (27.1) \text{ and } \quad \boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}} \quad (27.2)$$

with: $\mathbf{B} = \mu \cdot \mathbf{H}$ (28.1) and $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (28.2)

$$\boxed{\mathbf{E} = \mu \cdot \mathbf{v} \times \mathbf{H}} \quad (27.3) \text{ and } \quad \boxed{\mathbf{H} = -\varepsilon \cdot \mathbf{v} \times \mathbf{E}} \quad (27.4)$$

Experience/observation is dependent on the relative velocity v !

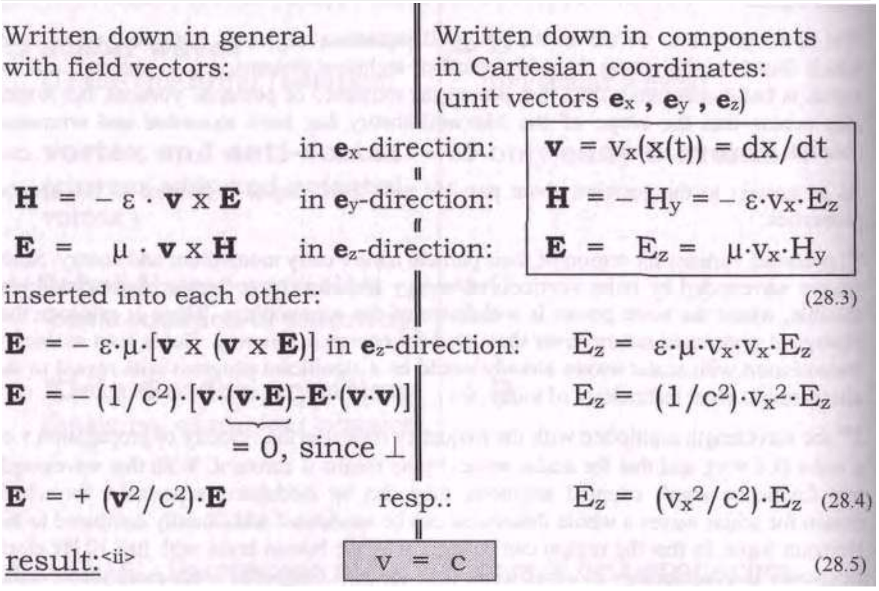

Written down in general with field vectors:

Written down in components in Cartesian coordinates: (unit vectors $\mathbf{e}_x$, $\mathbf{e}_y$, $\mathbf{e}_z$)

in $\mathbf{e}_x$-direction: $\mathbf{v} = v_x(x(t)) = dx/dt$

$\mathbf{H} = -\varepsilon \cdot \mathbf{v} \times \mathbf{E}$ in $\mathbf{e}_y$-direction: $\mathbf{H} = -H_y = -\varepsilon \cdot v_x \cdot E_z$

$\mathbf{E} = \mu \cdot \mathbf{v} \times \mathbf{H}$ in $\mathbf{e}_z$-direction: $\mathbf{E} = E_z = \mu \cdot v_x \cdot H_y$

inserted into each other: (28.3)

$\mathbf{E} = -\varepsilon \cdot \mu \cdot [\mathbf{v} \times (\mathbf{v} \times \mathbf{E})]$ in $\mathbf{e}_z$-direction: $E_z = \varepsilon \cdot \mu \cdot v_x \cdot v_x \cdot E_z$

$\mathbf{E} = -(1/c^2) \cdot [\mathbf{v} \cdot \underbrace{(\mathbf{v} \cdot \mathbf{E})}_{= 0, \text{ since } \perp} - \mathbf{E} \cdot (\mathbf{v} \cdot \mathbf{v})]$ $\quad E_z = (1/c^2) \cdot v_x^2 \cdot E_z$

$\mathbf{E} = +(\mathbf{v}^2/c^2) \cdot \mathbf{E}$ resp.: $E_z = (v_x^2/c^2) \cdot E_z$ (28.4)

result:<ii> $\boxed{v = c}$ (28.5)

Fig. 28.1: The equations of transformation,<i>

<i>: see part 1, chap. 6.5, eq. 60

<ii>: R. W. Pohl: Einfuhrung in die Physik, Bd. 2 Elektrizitatslehre, 21. Aufl. Springer-Verlag 1975, Seite 77, s.a. Kommentar auf Tafel 27.8

<iii>: chapter 28 is a repetition of chapter 6

28. Objectivity versus relativity

The new and field-theoretical approach contains the Maxwell-equations, but goes over and above these in one point. It describes potential vortices and their propagation in space as a scalar wave. With that can also a conclusive answer be given to the often-asked question for the medium and the mediated particles, which is a prerequisite for every longitudinal wave. Mediated are vortex structures with particle nature and the field itself functions as a medium. Is with that also answered the question concerning the aether?<iii>

28.1 The question concerning the aether

Do you know the Maxwell-experiment? No, you wouldn't be able to, since the intellectual father fast did make a backdown, after it didn't work out. Today one speaks of the Michelson-experiment and it may be connected with any other names (Morley, etc.). Remember: In his light theory Maxwell had determined a particular and constant value for the speed of light and for that there should be a physical reason, which should have its cause in the aether. By means of proving this aether Maxwell wanted to prove his theory, but this enterprise thoroughly went wrong.

The consideration was as follows: If the Earth is spinning and is moving through the cosmos, then one should be able to detect an aetherwind and different values for c in the different points of the compass. Maxwell found support for his project at the observatory, since with the aberration of the stars Bradley previously had described an observation, which could be considered as evidence for an aether. The director of the observatory charged his assistant Dr. Michelson with the task, to carry out a corresponding proof of an aether this time in a terrestrial experiment. But such an aether couldn't be proven, what Maxwell had to accept as a severe strike against his light theory. Seven years later Maxwell got the acknowledgement, however from a completely other corner by means of the experiments concerning the radio transmission of Heinrich Hertz.

Until today the question has remained open why astrophysics can prove the aether, whereas the detection in a terrestrial laboratory fails and it looks like there doesn't exist an aether. But as definition for the cause of c the aether can't be abolished as long as it is unsettled why the light is propagating with c of all possible velocities. The question is asked, what determines the propagation of light from today's point of view? Now, by means of outside fields the light can be slowed down. At present the world record lies at less than 65 kilometers per hour in a Bose-Einstein condensate. If electromagnetic fields determine the speed of light, if in addition field or gravitational lenses should confirm this, then the field takes over the task of the aether!

At this place the new field-theoretical approach shows its capabilities. The equations of transformation say nothing but that a moving H-field transforms to a resting E-field and vice versa, that thus in the place of a moving aether, the aetherwind, a resting aether is found. Doing so the dual field partners merely exchange their places. Therefore it is a wild-goose chase, wanting to measure an aetherwind with gauges, which underlie the same field (fig. 28.2).
Michelson had to fail.

The equations of transformation say:

$\boxed{\mathbf{E} = \mu \cdot \mathbf{v} \times \mathbf{H}}$ (27.3) and $\boxed{\mathbf{H} = -\varepsilon \cdot \mathbf{v} \times \mathbf{E}}$ (27.4)

- Experience/observation depends on the relative velocity v!
- The field takes over the function of the aether (determines c) and
- an aetherwind v x H is measured as a resting aether E and vice versa!

- for $v = c$ the equations of transformation turn into each other and are identical $[v = v_x(x(t))]$.
- for $v < c$ a motion field E_v depending on v is resulting

$$\boxed{E_v = E \cdot (v^2/c^2)} \quad (28.6)$$

- for $v = 0$ also $E_v = 0$.
- the motion field overlaps the E-field
- in the case of vortex fields the effect overlaps the cause and itself is the cause for a new effect.
- The overlap reaches to infinity, where each time is valid:

$$E_{n+1} = E_n \cdot (v^2/c^2) \quad (28.7)$$

- the field Eo overlaps the motion field E_v

$$E = E_0 + E_v = E_0 \cdot (1 + v^2/c^2) \quad (28.8)$$

- for infinite overlap:

$$E = E_0 + E_1 + E_2 + E_3 + E_4 + \ldots + E_n + E_{n+1} + \ldots \quad (28.9)$$

- results in the power series:

$$E = E_0 \cdot [1 + (v/c)^2 + (v/c)^4 + (v/c)^6 + \ldots + (v/c)^{2n} + (v/c)^{2(n+1)} + \ldots] \quad (28.10)$$

Fig. 28.2: Power series as a result of a vortex overlap.

<i>: Grimsehl: Lehrbuch der Physik, 2.Bd., 17.Aufl. Teubner Verl. 1967, S. 130.

<ii>: R.W.Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975, Seite 72 und 76, bzw. 130.

<iii>: K. Simonyi: Theoretische Elektrotechnik, 7.Aufl. VEB Berlin 1979, Seite 924.

<i4>: E. Friebe: Die Vektorprodukte der Maxwell'schen Elektrodynamik, DPG-Didaktik-Tagungsband 1995, S.396

28.2 Vortices, an overlap of the overlap

Not with any approach until now the question concerning the aether could be solved. Only the new field-theoretical approach proves with the unambiguous and free of contradiction clarification of the question concerning the aether its unmatched superiority. We hence without exception work with this approach, which is anchored tightly in textbook physics. The two equations of transformation on the one hand are the law concerning the unipolar induction according to Faraday (27.1) and on the other hand the dual formulation (27.2), which Grimsehl<i> calls convection equation. Grimsehl goes around the question for the correct sign by means of forming a modulus. Pohl draws detailed distinctions of cases and dictates the each time relevant formulation of the dual law<ii>. The sign eventually should be chosen according to the definition of the orientation of the field pointers. Also Simonyi gives both equations and the each time appropriate experiments<iii>.

If we assume the carrier of an electric field is moving with the not accelerated relative velocity v with regard to the reference system used by the observer, then a magnetic H-field is observed, which stands perpendicular both to the direction of the E-field and to the direction of v. If the motion takes place perpendicular to the area stretched by E- and H-field, then the H-field again is observed and measured as an E-field. There will occur an overlap of the fields.

In spite of that we first consider the theoretical case, that no overlap is present, and the observer as it were sees himself. The result is trivial: the relative velocity v must be the speed of light $v = c$. (28.5) If considered at the speed of light, the two equations of transformation turn into each other. They now are identical both mathematically and in their physical expressiveness. For this case it actually is possible, to derive the dual law straight from the Faraday law. For a wave propagating with the speed of light, to name an example, the field strength propagating along is always equal to the causing field strength, which depends on position.

If besides the evaluation of the values also the circumstance is considered that it concerns vectors, then at this place a problem as a matter of principle of the Maxwell theory gets visible, to which has been pointed occasionally, e.g. at the German Physical Society<i4>. The derivation of the speed of light from two vector equations requires, that c also has to be a vector. The question is: How the velocity vector v suddenly becomes the scalar and not pointing, in all directions of space constant factor c? Is therefore for mathematical and physical reasons "the Maxwell theory in essential parts erroneous", according to a statement of the German Patent Office<i4>?

Now, the constancy of the speed of light is a fact, which even can be derived. We at first will be content with the clue that for every observation with the speed of light, with the eyes or a gauge constructed corresponding to our perception, the vector in all its components each time is correlated to itself, by what actually the orientation of direction gets lost. Under these for c and with equal rights also for v relevant circumstances we are entitled to calculate further with the values.

An observer, who is moving with v slower than c, will besides the original E-field also observe a motion field E_v depending on the velocity v, which disappears, if v becomes zero. What he catches sight of and is able to register with gauges in the end is the overlap of both field components.

- concerning the development of the power series:

$$E = E_0 \cdot [1 + (v/c)^2 + (v/c)^4 + (v/c)^6 + \ldots + (v/c)^{2n} + (v/c)^{2(n+1)} + \ldots] \qquad (28.10)$$

- for $q < 1$ the power series with $q = (v/c)^2$ will converge (28.11)

$$\boxed{\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}} \qquad (28.12)$$

- for $(v/c)^2 < 1$ resp. for $v < c$ therefore is valid:

$$E = E_0 \cdot [1/1 - (v/c)^2] \quad \text{resp.} \quad \boxed{E_0 = E \cdot [1 - (v/c)^2]} \qquad (28.13)$$

The square root of Lorentz appears in squared form[i]:

$$\boxed{\left(1 - \frac{v^2}{c^2}\right) = \frac{E_0}{E}} \qquad (28.14)$$

The derivation for the magnetic field strength analogous to that provides the identical result[i]:

$$\boxed{\left(1 - \frac{v^2}{c^2}\right) = \frac{H_0}{H}} \qquad (28.15)$$

Fig. 28.3 The field dilatation depending on velocity

<i>: see part 1, chap. 6.6

<ii>: Prof. Dr. H.-J. Runckel, Abteilung fur Mathematik IV der Universitat Ulm

28.3 Field overlap

But it doesn't abide by this one overlap. In the case of vortex fields the effect overlaps the cause and itself becomes the cause for a new effect. The overlapped cause produces a further effect, which for its part is overlapping (see chap. 3).

Vortices thus arise, if overlaps for their part are overlapping and that theoretically reaches to infinity, to which I already repeatedly have pointed (fig. 3.0). In addition do vortices represent a fundamental physical principle. The Greek philosopher Demokrit has traced back the whole nature to vortex formation and that already 2500 years ago!

In the field-theoretical approach this interpretation seems to experience a mathematical confirmation, since also the fields are overlapping in vortex structures. According to that we owe our observations and our being the relative movements and the vortex formation. If reversed there wouldn't be any movement, then there also would not exist fields, light nor matter. If we observe the sky, then everything visible follows the movement of its own of the Earth, of the solar system and the whole galaxy, which is on its way with unknown galactic velocity, and all movements take place in vortex structures (fig. 10.2).

The field overlap dictated by the Faraday-approach as well reaches to infinity, what has stimulated my colleagues of mathematics to also mathematically put into practice this physical requirement[ii]. This leads to an infinite power series, which converges under the condition that $v < c$.

As a result of the power series development the well-known square root $\sqrt{1-(v^2/c^2)}$ of Lorentz occurs in squared form (see also fig. 6.6). It determines the relation of the observed and the causing field strength of the electric or the magnetic field. Physically the found relation describes a dilatation field depending on velocity. The field strength thus increases, if the relative velocity v increases, or inversely no difference is observable anymore, if v tends towards zero.

Whoever wants to compete with Albert Einstein (1879-1955), who has developed the theory of relativity from the length contraction, which depends on velocity, could be inclined to derive a new field physics from the field dilatation. But I must warn of such a step. The derivation of the length contraction by the mathematician Hendrik Lorentz (Lorentz contraction) assumes a number of limiting conditions. The relative velocity v for instance may not experience any acceleration. Actually however almost all motion takes place as circular vortex motion, so that due to the occurring centripetal acceleration the conditions for the theory of relativity aren't fulfilled anymore. Neglects or generalizations thereby can lead to considerable errors, of which I would like to warn. It in general is a delicate enterprise, if one wants to provide a physical interpretation for a purely mathematically won result.

This warning to the same extent also is valid for the here shown derivation of the field dilatation. The limiting conditions practically are the same as for Einstein and the problems with a provided physical interpretation won't be less. Also here lots of paradoxes will occur, which are nothing but errors of the theory. So we won't reach our destination.

There now only one further mathematical step is necessary, which links the theory of relativity with the new notion of a field dilatation depending on velocity.

Example:
Measurement of length by means of a measurement of propagation time (sound or light) with c = L/t in a vehicle moving with v.

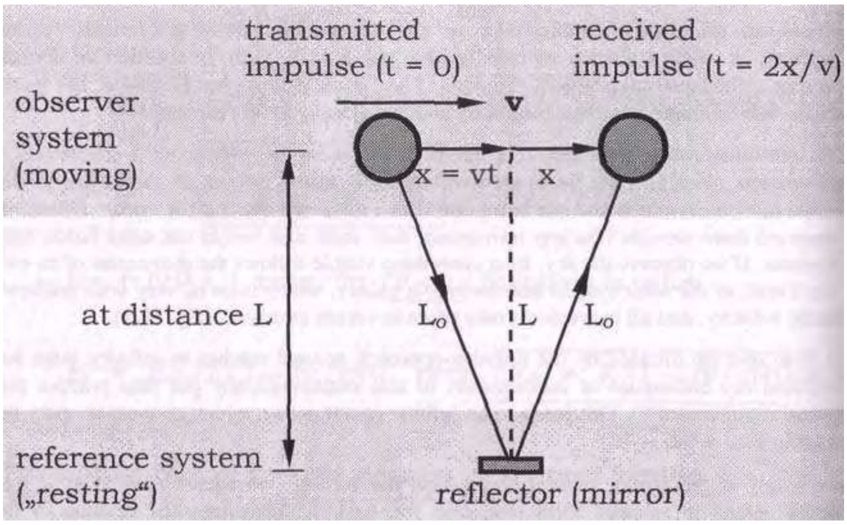
transmitted
impulse (t = 0)
received
impulse (t = 2x/v)
observer
system
(moving)
v
x = vt
x
at distance L
L_o
L
L_o
reference system
(„resting“)
reflector (mirror)

From driving time t:

$t = 2\frac{x}{v}$ = signal propagation time: $t = 2\frac{L_o}{c}$ follows $\boxed{\frac{x}{v} = \frac{L_o}{c}}$

According to Pythagoras: $L_o^2 = L^2 + x^2 = L^2 + L_o^2\cdot(v^2/c^2)$

$$L_o^2\cdot(1 - v^2/c^2) = L^2$$

the shortening of the rule results in:

$$\boxed{(1 - \frac{v^2}{c^2}) = (\frac{L}{L_o})^2} \quad (28.16)$$

Fig. 28,4: ______Derivation of the length contraction

Examples: contraction according to Lorentz transformation, measurable length shortening, curvature of space.

Counterexample: unsuccessful aether detection (Michelson experiment)

28.4 The derivation of the length contraction

The Lorentz transformation is the result of a purely mathematical problem. Stimulated by the surprisingly result of the Michelson experiment the Dutchman Hendrik A. Lorentz 1891 asked himself, how the equations of the Galilei-transformation would have to look like, if the propagation of light wouldn't be infinitely fast but finite and constant. He thereby proceeds from the assumption of two inertial systems moving against one another with a not accelerated velocity v, in which the laws of Newtonian physics are equally valid<i>. As a result of the relative motion a change of the length measures will occur.

This at first can be explained as a purely geometric effect in the context of nonrelativistic physics. We imagine a vehicle, which is on its way with constant velocity, and emits an optical or acoustical signal. Sideways in the countryside is standing in a perpendicularly measured distance L a reflector (mirror), which sends the signal back again. The velocity of the signal however isn't infinitely fast and from that follows that the vehicle during the propagation time of the signal as well has moved a bit further. The actual way, which the signal had to cover now amounts to L_o (> L). The distance measure thus is observed smaller as it is in reality, to be specific for the factor of the square root of Lorentz (fig. 28.4).

$$L = L_o \cdot \sqrt{1-(v^2/c^2)} \quad (28.16)$$

According to the principle of relativity it doesn't play a role, if the vehicle is driving or if it is standing still and the mirror is moving with a linear uniform velocity.

Initially Einstein also only spoke of an observable length contraction, which must not necessarily occur in reality, an optical deception so to speak. Lorentz however proceeded from the assumption of a physical length change, thus a length change existing in reality, what in practice at first makes no difference. If e.g. at relativistic velocities a rocket becomes smaller, then the pilot equally shrinks, so that it would not be possible to notice a present difference.

If however the observer stands outside the events and takes a „neutral standpoint", then he will be able to see, which interpretation is the right one. Today some examples are known. In accelerators particles at relativistic fast velocities actually get smaller for the factor of the square root of Lorentz. That has been proven and this result afterwards gives the Dutchman Lorentz right! The followers of the physical length contraction also are called Neo-Lorentzians.

In the vicinity of a gravitational mass the speed of light becomes so slow, that the shortening factor plays a role and space is curved towards the mass. To understand this shortening of scale, the influence of the field also should be considered.

<i>: Example: In a closed lift physical experiments are being carried out. Accelerations of the lift have an influence on the experiments. However no influence can be detected, if the lift is standing still or is moving with constant velocity. It with that fulfills the conditions of an inertial system. The question is: what do the experiments show someone standing outside, whom the lift passes by?

From the comparison of the Lorentz contraction (28.16) with the field dilatation (28.14 and 28.15)

$$\left(1 - \frac{v^2}{c^2}\right) = \left(\frac{L}{L_o}\right)^2 \quad \overset{!}{=} \quad \left(1 - \frac{v^2}{c^2}\right) = \frac{E_o}{E} = \frac{H_o}{H}$$

follows

$$\left(1 - \frac{v^2}{c^2}\right) = \boxed{\left(\frac{L}{L_o}\right)^2 = \frac{E_o}{E} = \frac{H_o}{H}} \qquad (28.17)$$

the proportionality (length measures depending on field):

$$\underline{E, H \sim 1/L^2} \qquad \text{and} \qquad Eo, Ho \sim 1/L_0^2 \qquad (28.18)$$

Experimental examples[i]:

- Electrostriction (piezo speaker)
- Magnetostriction
- Field or gravitational lenses
- Curvature of space, deflection of light

Conclusion[ii]:

- The field determines the length measures (what is 1 meter)
- The field determines the velocities v (in m/s)
- The field determines the speed of light c [m/s]
- Measurement of the speed of light is made with itself:

$$\boxed{c \sim r} \qquad (28.19)$$

- Measured is a constant of measurement c = 300.000 km/s
- The speed of light c is not a constant of nature!

Fig. 28.5: The dependence of the Lorentz contraction on the field

[i]: see part 1, chap. 6.10

[ii]: see part 1, chap. 6.11

28.5 The dependence of the Lorentz contraction on the field

The two results of the field dilatation (28.14 and 28.15) and of the Lorentz contraction (28.16) must be brought together and compared (28.17). Doing so the mathematical expression of the square root of Lorentz is cancelled out. That is of utmost importance, since with that also all limits disappear and there remains a purely physical relation, a proportionality of utmost importance (28.18).

What was the sense of the limits associated with the introduction of so-called inertial systems, which are the basis of the Lorentz transformation and which were adopted for our derivation of the field dilatation? They now only are auxiliary considerations according to model. We have chosen a very simple model, which can be described mathematically, in which an observer holds in his hand gauges for distances and field strengths and with that gauges a system flying by with constant velocity. He on the one hand determines a length contraction and on the other hand a field dilatation. He compares both with each other and comes to the conclusion: The field determines the dimensions!

This statement is purely physical and it is generally valid. It is independent of the relative velocity and all other mathematical conditions. A centrally accelerated circular motion e.g. will falsify the length contraction to the same extent, as the at the same time occurring field dilatation. It can be expected, that in addition to the square root of Lorentz also other errors will mutually efface, so that a generalization in this case actually seems to be allowed.

The won proportionality is of most elementary importance. We use it in the case of the piezo speaker and know it from the curvature of space and deflection of light in presence of extreme fields. If we ourselves however are exposed to the field as an observer, in which also the object to be observed is situated, then we are in the dilemma, not being able to perceive the influence. If we, to stay with the example, would sit in a rocket and this would become smaller at faster velocity, then we would notice nothing, since we also would shrink along to the same extent.

That concerns every measurement of velocity in general and the speed of light c in particular, which as is well-known is measured in meters per second. But if the field determines c and in the same way the length measure, which is given in meters, then both stand in a direct proportionality to each other, then we won't have the slightest chance to measure the speed of light. If namely c is changed, then this concerns the measurement path in the same way. Now the variable is measured with itself and as a result appears c, a constant value. We neither can see the change, since our eyes scan all objects optically and that means with c.

It is the nightmare of each and every measurement engineer, if the gauge depends on the factor to be measured. No wonder, if the theorem of addition of the velocities apparently loses its validity and always the same c is being measured, independent of the direction in which the source of radiation is moving (chap. 6.11). The result is:

The speed of light is a constant of measurement and not a constant of nature! If however the light is scanned with the speed of light, then also all components of the light vector correlated with themselves result in the same constant value c, then actually the vector of the speed of light loses its orientation in space and becomes a scalar factor. The Maxwell equations already anticipate this circumstance, but without providing an explanation why this is correct. Only the new field approach can answer the open question. With the derivation an axiom of physics - one also can say stumbling block - has been overcome.

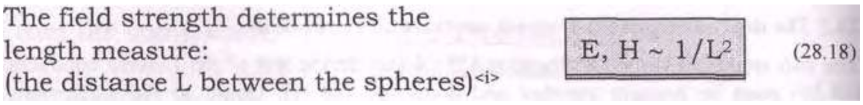

The field strength determines the length measure:
(the distance L between the spheres)<i>

$$\boxed{E, H \sim 1/L^2} \qquad (28.18)$$

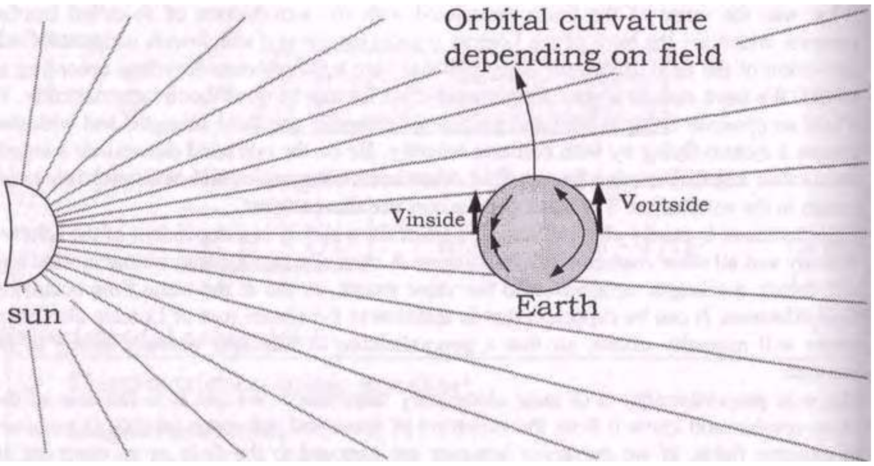
Orbital curvature
depending on field
v_{inside}
$v_{outside}$
sun
Earth

Fig. 28.6 A: The curvature of the Earth in the gravitational field of the sun

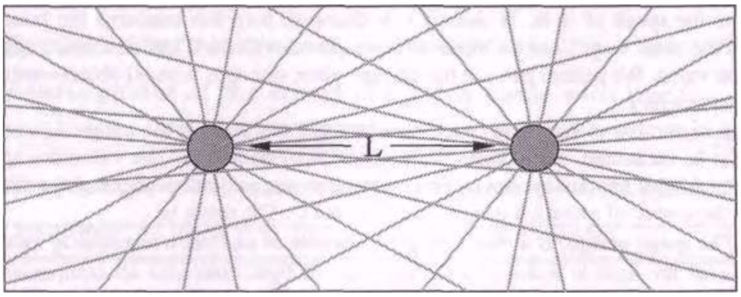

Fig. 28.6 B: Force of attraction and reduction of the distance L as the mutual field influence of two masses.

<i>: Repetition of part 1, chapter 6.7

28.6 Boscovich and the respiration of the Earth

Who has got a good idea, fast will find that some other scientist had the same idea already before, and for his part possibly already had to make the same experience. If in spite of that the track fast loses in history, it as a rule is because of the insufficient citing and the vanity of the discoverers, who in reality only are rediscoverers of a much older knowledge.

The dependence of the length measures on the field (eq. 28.18) in the mathematical form of the derivation however should still be quite new. But the physical consequence already was described by the Jesuit priest Roger Joseph Boscovich from Dalmatia in 1755<i>. He was Professor for mathematics and theology in Rome and spoke about the world on the one hand being deformable, similar to rubber, but on the other hand we aren't able to perceive this, since our body is made of the same material and follows all changes. „We in an absolute way can recognize neither the place, nor distances or orders of magnitude", writes Boscovich in his book about space and time and writes how these are perceived by us<ii>. He suspects that the Earth unobservable for man „is respiring".

Actually a terrestrial observer at daytime is situated closer to the sun than in the night. He by day is exposed to a slightly stronger field and as a result correspondingly smaller. He himself and all objects in his neighbourhood are subject to an identical fluctuation of size, so that this „respiration" of the Earth cannot be detected. It can be detected neither with a tape measure nor with an optical measurement and still is present in reality. Only from a neutral standpoint we can succeed to recognize the actually existing curvature of space (fig. 28.6 A).
An example is the duration of the sunshine at the equator, which is longer than can be expected from the spherical form of the Earth. This reveals, how the Earth is bending towards the sun (see also chapter 6.7).
A further example is the influence of the field on the orbital velocity of the Earth measured in meters per second. Here also the meter at daytime is smaller than in the night, for which reason the Earth is moving slower on the side turned towards the sun, like a track vehicle, which drives a turn, if the chain at the inside runs slower than on the outside. If the Earth describes an orbit around the sun. then this circumstance has to do nothing at all with centrifugal force or with a force of attraction of the sun. The circular motion simply and solely is a result of the field influence of the sun<iii>.

The force idea proves to be a pure auxiliary description. In the context of Newtonian mechanics the force plays a central role. Without question it is a very efficient and subjective perceptible description, which still isn't able to reproduce the physical reality in an objective manner. What keeps the planets into their orbit is only the field of the sun, which we call gravitational field and not some force! But of which kind is the gravitation and the field, which causes masses to come closer together and following our subjective observation attract each other (fig. 28.6 B)?

<i>: O. E. Rossler: Endophysics, the World as an Interface, World Scientific Publishing Co. 1998, Kap. 10, S. 87-112, mit Ubersetzungen aus <ii>:

<ii>: R. J. Boscovich: De spatio et tempore, ut a nobis cognoscuntur, 1755.

The length measure (the distance L between the spheres)<i> is determined by the field strength:

$$\boxed{E, H \sim 1/L^2} \quad (28.18)$$

A. Charged mass points (electrons, positrons, ions,...):

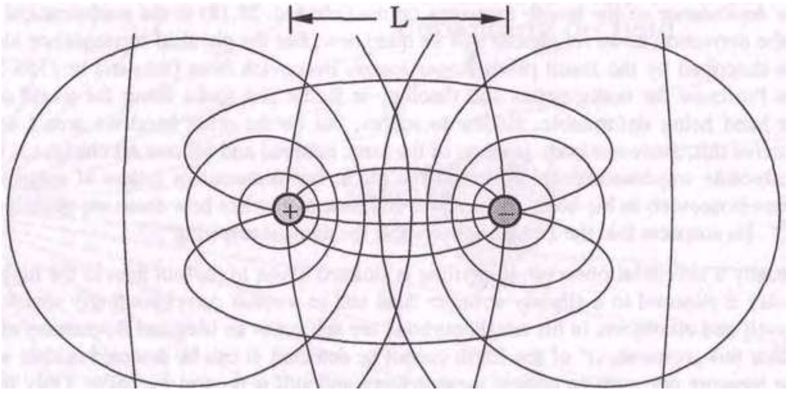

B. Uncharged mass points (neutrons, atoms, ...):

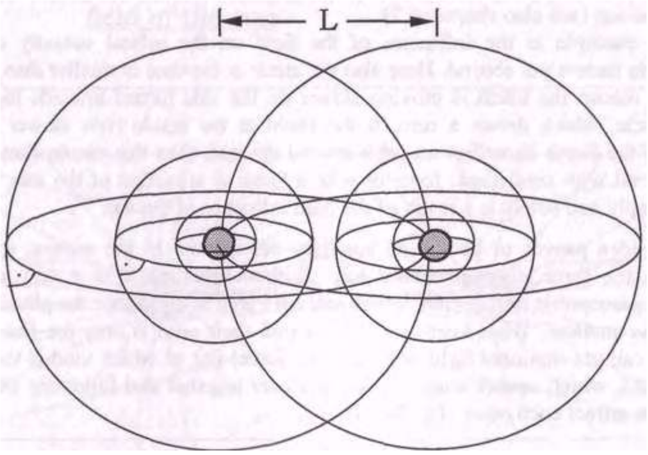

Fig. 28.7: Observation of a mutual force of attraction because of the effect of the fields on the distance measure.

<i>: Repetition of part 1, chapters 6.7 - 6.9 and part 2, chapters 10.4 and 15.2

28.7 Derivation of the gravitation

In fig. 28.6B the relation between the field influence and the observed force of attraction of two mass bodies is represented. If I in my mind for instance "switch on" the field lines of both bodies, which are placed at distance L, then the fields according to equation 28.18 reduce the measure L and optically come each other closer. With increasing proximity the field line density increases, so that L further decreases. We observe a mutual attraction, which lets both bodies move towards each other.

In fig. 28.7A the two bodies carry an electric charge. For different polarity the field lines as is well-known run from the positive pole to the negative pole, to bundle up there. As a matter of principle also here an attraction can be expected, which is called electromagnetic interaction. For the reason of the bundling up of the field lines this effect however will turn out to be considerably stronger. Hence the electromagnetic interaction is for many powers of ten more powerful than the gravitation.

Furthermore there also can occur repulsion, if in the case of like charge the field lines are bent aside and between the two bodies an area is formed, where the field tends towards zero and the distance measure L (according to eq. 28.18) as a result towards infinity. The electromagnetic interaction theoretically indeed reaches to infinity. Responsible are the open field lines arising from a charged body.

Now every charged body in addition has a mass, with which it takes part in the gravitation. Let's remember the comparison of the derivations. The Maxwell theory teaches us that in the static case E- and H-field are decoupled, because each time the other field disappears. Even if as a result of the unipolar induction for every open field line the other one is taken to be standing perpendicular to the open field line, then this other line just wraps around the open field line and forms a closed-loop field line. In that way it can't be influenced anymore from the outside and can be neglected, so goes the doctrine, which is drawn from the Maxwell theory (fig. 27.5).

This is a fatal error in reasoning! The equation 28.19 naturally is valid for open field lines in the same manner as for closed ones. These fields also lead to an observable force of attraction. If of course exactly those fields are neglected, which are responsible for the gravitation, then we need not wonder, if we don't understand the gravitation and the nature of this interaction!

The influence of the closed field lines responsible for the gravitation is due to the missing bundling up of the lines correspondingly weak. Secondly these can't exist a force of repulsion due to the missing ability to influence closed field lines from the outside and third it can be recorded that all charged bodies also have a mass. All three statements of the field lines model perfectly cover the physical reality.

Fig. 28.7 B shows uncharged bodies, for which both the field lines of the E-field and of the perpendicular to them arranged H-field are closed-loop. Such bodies, like e.g. neutrons or whole atoms without charge behave electrically neutral to the outside, but have a mass for the reason of the closed field lines, whereby the field lines of the H-field dominate those of the E-field<i>.

<i>: Repetition of K. Meyl: EMEC part 1, chap. 6.8 + 6.9, part 2 chap. 15.1

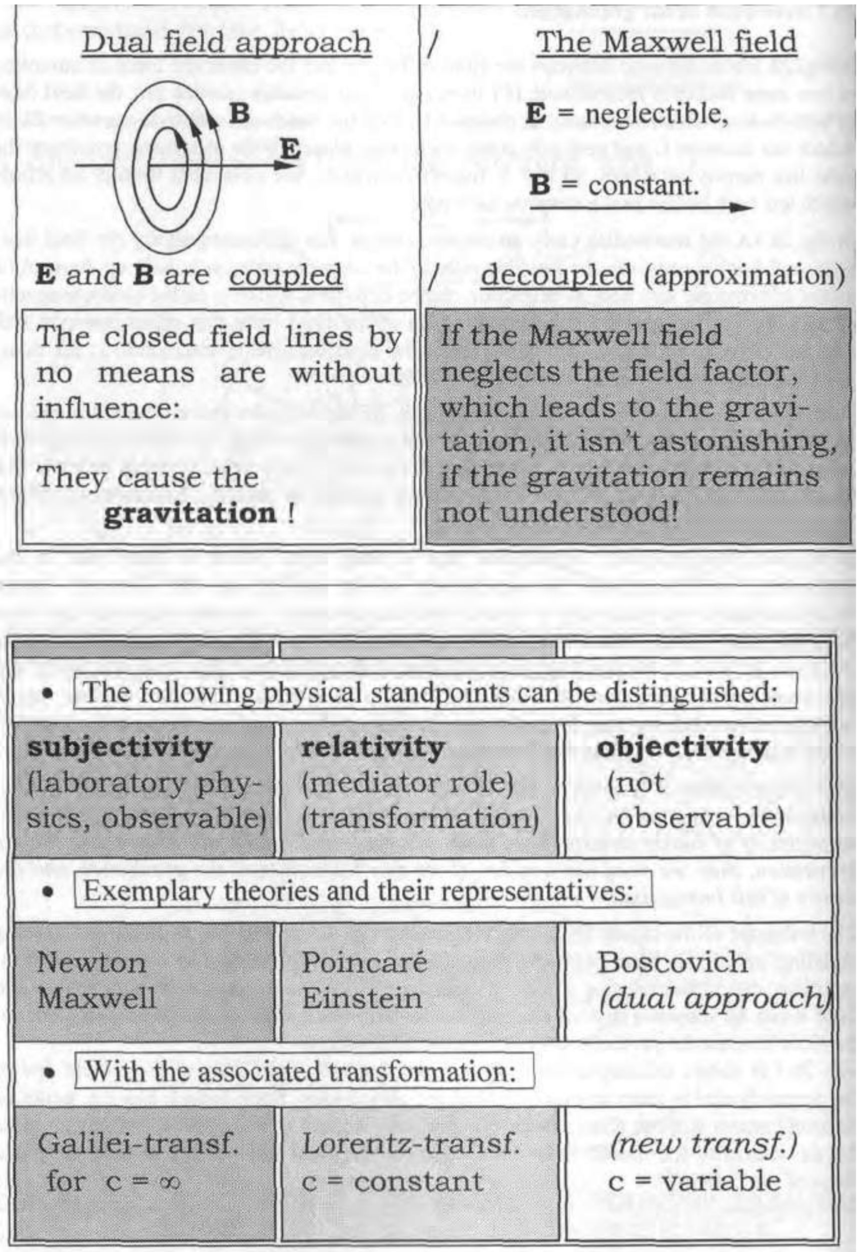

Dual field approach /	The Maxwell field
	E = neglectible, **B** = constant.
E and **B** are coupled /	decoupled (approximation)
The closed field lines by no means are without influence: They cause the **gravitation** !	If the Maxwell field neglects the field factor, which leads to the gravitation, it isn't astonishing, if the gravitation remains not understood!

- The following physical standpoints can be distinguished:

subjectivity (laboratory physics, observable)	**relativity** (mediator role) (transformation)	**objectivity** (not observable)

- Exemplary theories and their representatives:

Newton Maxwell	Poincaré Einstein	Boscovich *(dual approach)*

- With the associated transformation:

Galilei-transf. for c = ∞	Lorentz-transf. c = constant	*(new transf.)* c = variable

Fig. 28.8: Physical standpoints

28.8 From subjectivity to objectivity

With the field lines interpretation, which by the way already preferably was used by Faraday, the gravitation proves to be an until now neglected influence of the electromagnetic field. With that for the first time also the grand unification of the interactions was successful. The long sought-for unified theory with that for the first time comes within reach.

The derivation has made it possible to mathematically secure the theoretical working model of Boscovich. Already 1755 Boscovich points out the optical deception, which our observation underlies, if absolute orders of magnitude in our neighbourhood should change and our perception would change along. Then also all metric and optical measurement results would underlie this change. Following the idea of Boscovich I distinguish between subjectivity and objectivity.

The relativity is a compromise lying between both points of view, where a neutral standpoint is strived for, which lies outside the events. And from this standpoint the objectively taking place events are being observed. The theory of relativity consequently is a pure observer theory with strongly restricted scope on the basis of the Lorentz-transformation.

Theories of classic physics, like e.g. Newtonian mechanics, fall in the domain of subjectivity. The results and regularities are won in a terrestrial laboratory if possible isolated from the environment, where they have absolute validity. Here the Galilei-transformation is valid.

But if these subjectively won laws are applied to the microcosm in quantum physics or to the calculation of cosmic observations, one fast hits limits. The better the resolution of the microscopes and telescopes gets, the clearer the „outside" observer realizes, how much the laws of classic physics lose their validity.

Astrophysics successfully reaches for the theory of relativity, which with the curvature of space in the vicinity of mass centres delivers useful explanations. Here the dependence of the spatial dimensions on the field already could be established. In contradiction to that this fundamental relation is said to play no role whatsoever in quantum physics, or in all terrestrial laboratory experiments. But with which right may physical regularities from one domain be ignored in others? There only can exist one physics and that should be sought for!

What we need is objectivity! Behind all the apparently disconnected phenomena of physics work quite simple laws, which can't be observed and are until now not recognized by us. Objective physics in the words of Goethe is the one, which holds the world together in the heart of hearts. I call this, already by Boscovich suggested point of view, theory of objectivity. The access to the model domain of objectivity must be made mathematically by means of a transformation, since it is blocked for us by means of measurements or observations (see chapters 6.15-6.19). The transformation back into the observation domain must be made according to the same mathematical relations (fig. 28.9). In this way the quantum properties of the elementary particles can be calculated with high accuracy and agreement with the values, which until now only could be measured (chapter 7).

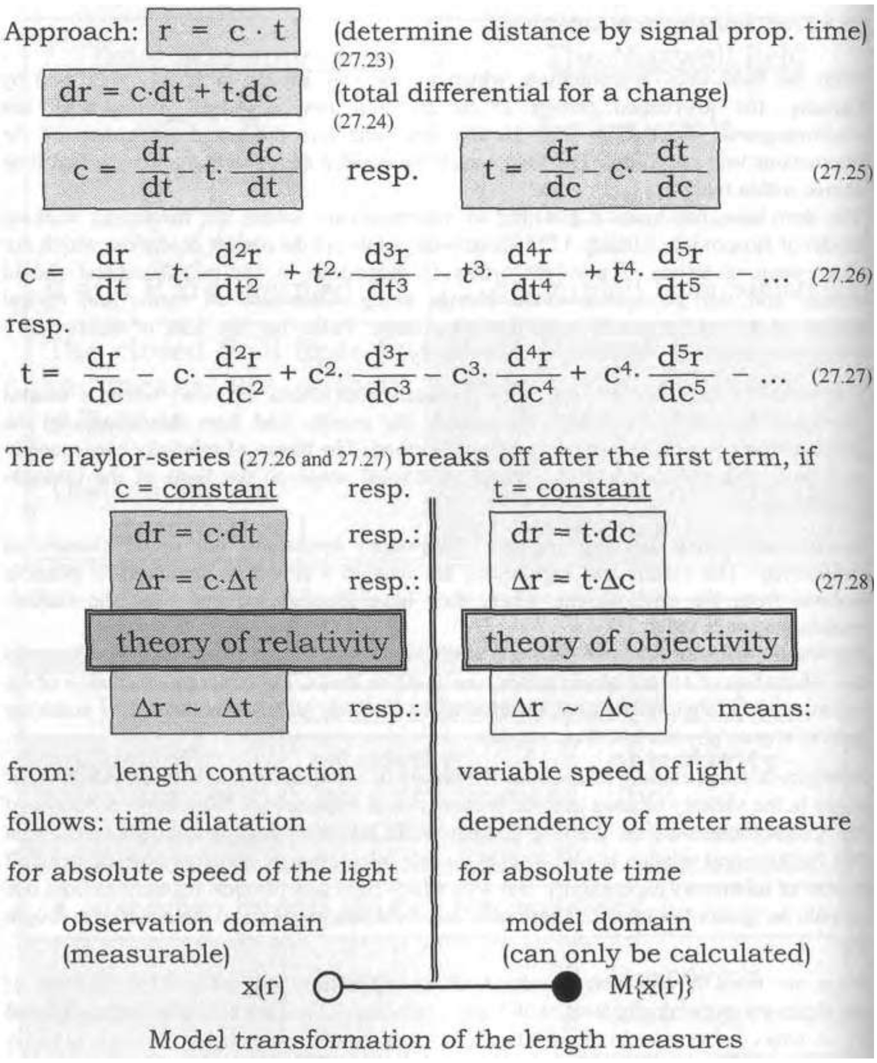

Approach: $\boxed{r = c \cdot t}$ (determine distance by signal prop. time) (27.23)

$\boxed{dr = c \cdot dt + t \cdot dc}$ (total differential for a change) (27.24)

$$\boxed{c = \frac{dr}{dt} - t \cdot \frac{dc}{dt}} \quad \text{resp.} \quad \boxed{t = \frac{dr}{dc} - c \cdot \frac{dt}{dc}} \qquad (27.25)$$

$$c = \frac{dr}{dt} - t \cdot \frac{d^2r}{dt^2} + t^2 \cdot \frac{d^3r}{dt^3} - t^3 \cdot \frac{d^4r}{dt^4} + t^4 \cdot \frac{d^5r}{dt^5} - \ldots \qquad (27.26)$$

resp.

$$t = \frac{dr}{dc} - c \cdot \frac{d^2r}{dc^2} + c^2 \cdot \frac{d^3r}{dc^3} - c^3 \cdot \frac{d^4r}{dc^4} + c^4 \cdot \frac{d^5r}{dc^5} - \ldots \qquad (27.27)$$

The Taylor-series (27.26 and 27.27) breaks off after the first term, if

c = constant	resp.	t = constant
$dr = c \cdot dt$	resp.:	$dr = t \cdot dc$
$\Delta r = c \cdot \Delta t$	resp.:	$\Delta r = t \cdot \Delta c$ (27.28)
theory of relativity		theory of objectivity
$\Delta r \sim \Delta t$	resp.:	$\Delta r \sim \Delta c$ means:
from: length contraction		variable speed of light
follows: time dilatation		dependency of meter measure
for absolute speed of the light		for absolute time
observation domain (measurable)		model domain (can only be calculated)
x(r) ○		● M{x(r)}

Model transformation of the length measures

Fig. 28.9: Theory of relativity and theory of objectivity and the model transformation between both physical standpoints

---

<i>: Repetition of part 1, fig. 6.16

28.9 The objective standpoint

The question is asked how one gets to an objective physical standpoint, which in addition evades every observation? The way leads over a transformation, to which all perceptible and measurable relations must be submitted.
If we for instance measure the distance r to a point light source, then the propagation of the light c and the propagation time t determine the distance measure $r = c*t$
If there occurs a little change of the distance, then two causes should be considered: Either the propagation time or the speed of light have changed somewhat. With that the two possible standpoints already would have been found.

The relativistic standpoint, which proceeds from the assumption of the speed of light being constant, says: the propagation time varies and we are dealing with a clock problem. If namely for relativistic velocities a length contraction occurs, then from that necessarily follows a time dilatation.
But actually no specific statement can be made about the constancy of the speed of light, besides that we look at, measure and scan everything with c and hence only observe the constancy. With that the theory of relativity remains a pure observer theory, exactly as Einstein originally called it into existence. This standpoint follows the motto: What can't be observed also doesn't need to interest the physicist.

The objective standpoint strives for more, for a description of the actually taking place processes. This time we proceed from the assumption of a universal and constant time with the argument: The time measure is an immutable definition and the physicist, who dictates this, himself determines what is simultaneousness. Then there also is no place for time travel and for clocks going wrong.
Therefore the speed of light can take all possible values always in strict proportionality to the length measures. Thus the measured length and distance measures should be transformed and that in the end is the unit ,,meter", which should be replaced by an objective measure.

With that the necessary transformation for variable c would be outlined. This transformation will be enqueued in the file of the big transformations. From it the Lorentz-transformation for c = constant emerges as a special case, like already from that transformation the Galilei-transformation follows for $c = \infty$. How now the relation of the subjective to the objective ,,meter" should be determined; by means of the relation of the relevant fields (eq. 28.17) or by means of the square root of Lorentz (eq. 28.16), over that should be worked and spoken. We already have successfully gone through it in a concrete example (chapter 7).

Every theory is judged according to its expressiveness. Ending this chapter the statements and derivations hence again are compared. On the one hand the Maxwell theory and from that the theory of relativity can be derived from the new approach, on the other hand a long list follows, which can't be connected with the Maxwell equations, like e.g. the gravitation. For instance the neutrino and all other elementary particles with all their specific quantum properties are derived (chapter 7), free and easy fundamental laws result, like the law of conservation of energy, and even the temperature spills its until now kept secret (chapter 8.3). Remains the conclusion: With no other approach according to the textbooks until now the efficiency of the new approach could be obtained.

• From the dual field-theoretical approach are derived:	• From Maxwell's field equations can be derived:
=> Maxwell's field equations	=>0
=> Quantum properties of the elementary particles	=>0
=> Neutrino (as an oscillating ring-like vortex)	=>0
=> Gravitation (as a result of closed field lines)	= > 0
=> Unified theory (grand unification of all interactions)	=>0
=> Temperature (as an oscillation of size depending on field)	=>0
=> Law of conservation of energy (and many other fundamental laws of physics)	=>0
=> Theory of objectivity	=> Theory of relativity

Fig. 28.10: Comparison of the efficiency of the approaches (final balance)

29. Mathematical gleanings

If, proceeding from the new field-physical approach, well-known and accepted theories are derived as special cases, this on the one hand can be valued as evidence for the correctness of the approach. On the other hand the new approach in part significantly influences the interpretation of the derived theories. That can involve a rethinking, with which not insightful people have difficulties, if for instance quantum physics, thermodynamics or the gravitation become partial aspects of electromagnetism. Over and above that are hidden many new thing in the new approach, which are there to discover. To that are counting among others the potential vortices and the scalar waves. One can work out these phenomena physically or mathematically, where the latter way as a rule is the faster one. Hence the summary shall be concluded with a kind of mathematical gleanings.

29.1 General and special theory of relativity

Albert Einstein distinguishes between general and special theory of relativity. Whereas the special (SRT), still is linked tightly with the prerequisites of the Lorentz-transformation, the general (GRT), deals with an extension to arbitrary systems, which mustn't be inertial systems. I would like not to dwell upon the GRT, as Einstein designed it, and merely notice that every generalization represents a possible source of errors and has to be well founded.

In the case of our derivation, the general case as it were resulted of its own accord. Let's turn back: If the root of Lorentz still was a component of the derived field dilatation (28.15) and equally of the length contraction (28.16), then it fell out in the comparison of both results (28.17). With that the important result, the proportionality (28.18), which among others results in the gravitation, becomes independent of the speed of light and the relative velocity v. This last step is obvious and still completely new. It cannot be found at Einstein, who in another way finds his GRT and his description of the gravitation. Even if here is striven for the same goal, then deviations in the result cannot be excluded because of the differences in the derivation, for which reason I additionally mark the by me derived general relativity (GRT'), to avoid confusion.

Influence of the Lorentz-transformation in the:
SRT (special theory of relativity): one-dimensional,
GRT' (general theory of relativity): three-dimensional, to a large extent corresponding to the GRT of Albert Einstein,
GOT (general theory of objectivity)

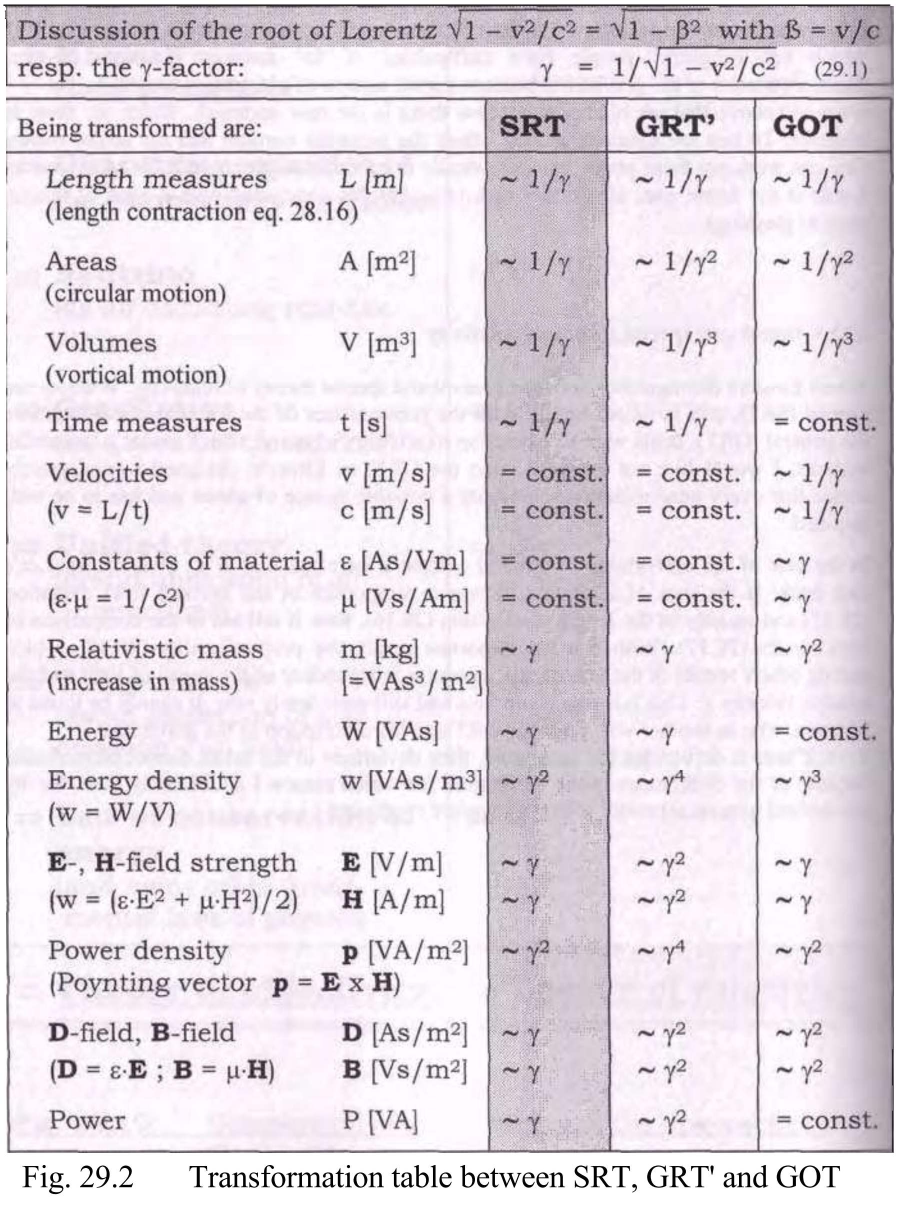

Discussion of the root of Lorentz $\sqrt{1 - v^2/c^2} = \sqrt{1 - \beta^2}$ with $ß = v/c$
resp. the γ-factor: $\gamma = 1/\sqrt{1 - v^2/c^2}$ (29.1)

Being transformed are:		SRT	GRT'	GOT
Length measures (length contraction eq. 28.16)	L [m]	$\sim 1/\gamma$	$\sim 1/\gamma$	$\sim 1/\gamma$
Areas (circular motion)	A [m^2]	$\sim 1/\gamma$	$\sim 1/\gamma^2$	$\sim 1/\gamma^2$
Volumes (vortical motion)	V [m^3]	$\sim 1/\gamma$	$\sim 1/\gamma^3$	$\sim 1/\gamma^3$
Time measures	t [s]	$\sim 1/\gamma$	$\sim 1/\gamma$	= const.
Velocities	v [m/s]	= const.	= const.	$\sim 1/\gamma$
(v = L/t)	c [m/s]	= const.	= const.	$\sim 1/\gamma$
Constants of material	ε [As/Vm]	= const.	= const.	$\sim \gamma$
($\varepsilon \cdot \mu = 1/c^2$)	μ [Vs/Am]	= const.	= const.	$\sim \gamma$
Relativistic mass (increase in mass)	m [kg] [=VAs^3/m^2]	$\sim \gamma$	$\sim \gamma$	$\sim \gamma^2$
Energy	W [VAs]	$\sim \gamma$	$\sim \gamma$	= const.
Energy density (w = W/V)	w [VAs/m^3]	$\sim \gamma^2$	$\sim \gamma^4$	$\sim \gamma^3$
E-, **H**-field strength	**E** [V/m]	$\sim \gamma$	$\sim \gamma^2$	$\sim \gamma$
($w = (\varepsilon \cdot E^2 + \mu \cdot H^2)/2$)	**H** [A/m]	$\sim \gamma$	$\sim \gamma^2$	$\sim \gamma$
Power density (Poynting vector $\mathbf{p} = \mathbf{E} \times \mathbf{H}$)	**p** [VA/m^2]	$\sim \gamma^2$	$\sim \gamma^4$	$\sim \gamma^2$
D-field, **B**-field	**D** [As/m^2]	$\sim \gamma$	$\sim \gamma^2$	$\sim \gamma^2$
($\mathbf{D} = \varepsilon \cdot \mathbf{E}$; $\mathbf{B} = \mu \cdot \mathbf{H}$)	**B** [Vs/m^2]	$\sim \gamma$	$\sim \gamma^2$	$\sim \gamma^2$
Power	P [VA]	$\sim \gamma$	$\sim \gamma^2$	= const.

Fig. 29.2 Transformation table between SRT, GRT' and GOT

29.2 Transformation table

Let's speak again about the difference to the special relativity (SRT). This so to speak deals with the one-dimensional case of the uniform motion of a reference system in x-direction ($v = v_x$), as specified by the Lorentz-transformation, where only the x-components and not those in y- or z-direction are being transformed. As already mentioned this is a purely theoretical case, which in practice occurs next to never. Normal is circular and vortical and with that accelerated motion, where the velocity component permanently changes its direction.

The derived result of the general relativity (GRT') does justice to this circumstance. Even if this at first only has been derived for the x-direction it nevertheless is valid equally in y- and z-direction. It even remains valid for the case that we base on a path of arbitrary form of a spatial field vortex. In this case some components continually occur in all directions of space, so that the relative velocity v as already the speed of light c loses its vectorial nature. With that the transition of the SRT to the GRT is carried out. By means of the spatial swirling the electric and magnetic field pointers at the same time turn into scalar factors, by taking over the function of the aether. Let us remember that even Einstein in his GRT was forced to again introduce the aether, which in the SRT still was unnecessary.

It therefore makes a difference in the transformation of physical factors, if we base on a one-dimensional (SRT) or a three-dimensional spatial description (GRT). Length measures in x-direction in both cases must be converted using the root of Lorentz. Usually the relativistic γ-factor is introduced, which is inverse to the root of Lorentz

$\gamma = 1/\sqrt{1-v^2/c^2})$ with $x_o/x = \gamma$ (29.2)

If thus individual length measures would be subject to a length contraction following the γ -factor, then a volume V according to the SRT must be transformed with γ, according to the GRT' however with γ^3.
As is well-known a relativistic increase in mass is converted with the γ-factor and in the same manner the to that proportional energy $E = m\ c^2$. If we however correlate the energy to the volume V and in that way determine an energy density w, then the difference between SRT ($w \sim \gamma^2$) and GRT' ($w \sim \gamma^4$) again has its maximum effect.
A relation to the field factors of E- and H-field is for instance provided by the energy densitv of a wave field
$w = (\varepsilon \cdot E^2 + \mu \cdot H^2)/2$ (29.3)

According to that the field strengths in the one-dimensional case of the SRT should be converted with the γ-factor, in the case of the GRT' however with γ^2, in accordance with the derivation in chapter 28. This circumstance willingly is overlooked, although it only concerns the textbooks and the today valid theory of relativity. I however point to the difference, since it does make a difference if we start with the SRT or the GRT when we change to the general theory of objectivity (GOT).

In the domain of the GOT all length measures should be transformed. The respective dimension gives information with which power the γ-factor occurs (fig. 29.2). The unit meter is responsible for that.

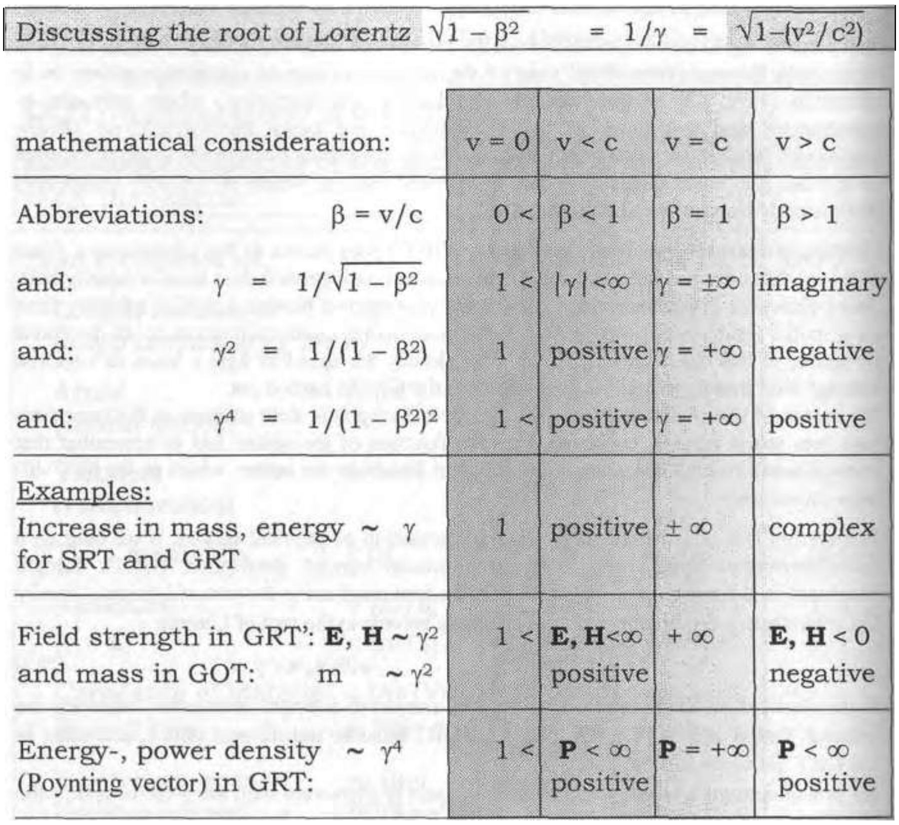

Discussing the root of Lorentz $\sqrt{1-\beta^2} = 1/\gamma = \sqrt{1-(v^2/c^2)}$

mathematical consideration:	$v = 0$	$v < c$	$v = c$	$v > c$
Abbreviations: $\beta = v/c$	$0 <$	$\beta < 1$	$\beta = 1$	$\beta > 1$
and: $\gamma = 1/\sqrt{1-\beta^2}$	$1 <$	$\lvert\gamma\rvert < \infty$	$\gamma = \pm\infty$	imaginary
and: $\gamma^2 = 1/(1-\beta^2)$	1	positive	$\gamma = +\infty$	negative
and: $\gamma^4 = 1/(1-\beta^2)^2$	$1 <$	positive	$\gamma = +\infty$	positive
Examples: Increase in mass, energy $\sim \gamma$ for SRT and GRT	1	positive	$\pm\infty$	complex
Field strength in GRT': **E**, **H** $\sim \gamma^2$ and mass in GOT: m $\sim \gamma^2$	$1 <$	**E**, **H**$<\infty$ positive	$+\infty$	**E**, **H** < 0 negative
Energy-, power density $\sim \gamma^4$ (Poynting vector) in GRT:	$1 <$	**P** $< \infty$ positive	**P** $= +\infty$	**P** $< \infty$ positive

Fig. 29.3 Discussion concerning the root of Lorentz

- The special theory of relativity SKT only is defined for $v < c$
- For v • c particles with a complex mass, but with a real energy density (according to GRT') would result.
- From the point of view of the theory of objectivity (GOT) the mass should be taken negative-real (neutrino?).

29.3 Discussion concerning the root of Lorentz

Fig. 29.2 forms the basis for the transformation in the domain of the GOT, the general theory of objectivity, where it plays a role, if a subjectively measured factor should be converted from the laboratory domain or a relativistic factor according to the SRT or the GRT'. The given proportionalities thereby should be put in the respective relation. In this way results the respective valid instruction for transformation on the basis of the root of Lorentz.

Let's take a critical look at the root of Lorentz. The velocity v occurring in it, of whatever this may consist, is depending on the field according to equation 28.14 + 28.15. It strictly speaking wouldn't be constant anymore and wouldn't belong in a general instruction for transformation at all. Only, what is valid for v, is valid to the same extent for c. Since only the proportion of v/c occurs in the root of Lorentz every influence depending on field or of other nature will have no effect on v/c and the value of the root of Lorentz. It in any case will retain its value. It fulfills for itself the condition of the Lorentz invariance. According to that in the case of the relative velocity v it doesn't depend on the absolute value, but only on the relation to the speed of light. In addition the restriction to values of $v < c$ is normal, if the speed of light is seen as an upper limit. Let's first purely mathematically draw a case distinction for different velocity domains of v. For $v = 0$ the root of Lorentz becomes one and the Lorentz transformation turns into the Galilei transformation.

Connected to this is the today well-known and technically used domain up to the limit of $v = c$. It virtually is impossible to accelerate a mass particle to the speed of light, since mass, field and energy would grow towards infinity, as is clear from the table (fig. 29.2). Particles as fast as light, like photons, hence cannot have a mass. At $v = c$ a singularity is present.

In a field theory, which also deserves this name, however an upper limit must not be present. Hence also the case for $v > c$ should be required theoretically. Only later we will be able to judge if this makes sense physically. We at first only want to examine the case mathematically. Mass, field and energy now again have a finite value, there however results a complex, purely imaginary mass, a negative field and doing so, as already before a positive energy and power density.

There sometime has been the textbook opinion that it is physically impossible to fly faster than sound. This erroneous statement even could be proven „scientifically", because such a supersonic airplane would fly off the observation space and with that wouldn't be real anymore, thus from a mathematical viewpoint would be complex. Anyone, who in New York gets off a Concorde, can confirm that everything at any moment of the flight was real. Only the observer is deceived, if the airplane flies somewhere else, than he perceives it. Is the speed of light also such a „sonic barrier", which by the majority of the scientists since Einstein until today still is thought to be insurmountable?

How should one physically imagine a complex mass? Let us remember the alternating current teachings, where it is normal to work with complex values, since the mean values of the oscillating alternating currents, tension voltages and fields are zero. Calculating with mean values would result in zero energy and power. Hence complex factors are introduced and the root mean square values are calculated and measured instead of the mean values. Could a complex mass analogously not concern an oscillating particle, a particle, which in addition is faster than the light?

- Physical example:

neutrino = mean of the oscillating mass $\approx$ zero
RMS value of the complex mass $\neq$ zero
E- and H-field with inverse sign
carrier of positive energy and power
ν -model: ring-like vortex oscillating from e^- to e^+ <i>

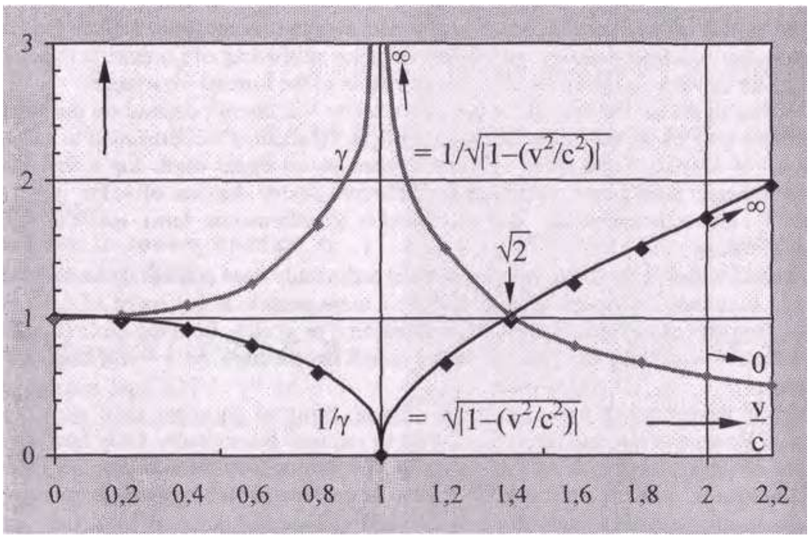

Examples from the general theory of relativity (GRT):

(28.16): length contraction $L \sim 1/\gamma = \sqrt{1-(v^2/c^2)}$

(6.19): increase in mass $m \sim \gamma = 1/\sqrt{1-(v^2/c^2)}$

(6.6,28.15) field dilatation $\mathbf{E} \sim \mathbf{H} \sim \gamma^2 = 1/(1-v^2/c^2)$

Fig. 29.4: Root of Lorentz for speeds faster than light (v > c)

29.4 Mathematical derivation of the neutrino

In the domain of speeds faster than light, for v > c, the power series (28.12) does not converge anymore and every observer theory and every observation (fig. 28.4) will fail, because particles faster than light run away of their own visible appearances. Every measurement and every observation inevitably is behind and hardly can be assigned to the actual cause. That way for instance measured neutrino events are being connected with celestial observations, with which they haven't got anything to do at all.
If we however describe the domain v > c in the complex plane, then astonishing results are found, which could be verified physically: a complex length dilatation with increasing velocity goes along with a loss of complex mass. The oscillating fields, energy and power density however would be real with negative sign.
Thus there would result particles carrying energy with an oppositely poled field, with an oscillating mass and if necessary also an oscillating charge. Without static mass and charge these particles hardly would interact with normal matter, which leads to an enormous ability of penetration. The only physical particles, which have such a property, are the neutrinos. With that a usable and an extremely efficient model description has been found for these particles. Also the energy of these particles can be calculated, which has considerable orders of magnitude and is available as an energy source everywhere and any time.

If for instance in a converter for space energy a neutrino should be converted into a resting charge carrier (with v = 0), then two steps are necessary (see part 2 of this series of books):

1. First the neutrino must be slowed down to 1.414 times the speed of light (fig. 28.9). Doing so energy is spent and not won! The converter for instance can cool down.

2. Afterwards the characteristic rotation of its own, with which the ring-like vortex spins around itself by permanently putting its inside to the outside and vice versa, has to be taken away from the neutrino. In that way the vortex centre is closed and the particle acquires localization. It becomes a charge carrier.

Even if the representation in the complex plane represents only an auxiliary description, the model nevertheless seems to be efficient, because despite its complex mass and charge the neutrino nevertheless carries a real energy. It in any case is represented in that way to an observer, who measures the relation with the speed of light, who in the relativistic scheme of things scans the relation.

Today, as already said, even the sonic barrier has become permeable and no scientist dares to physically deny this fact and even prove his mistake mathematically anymore. No, on the contrary, he always did know that as an expected consequence the sonic barrier runs after the supersonic plane. The once physically unthinkable and scientifically fought has become normality.
What should hinder an oscillating particle, like a neutrino, to be faster than the light? Some time one also will accustom to that.

- Starting-point:
  the fundamental field equation
  (derivation in chapter 27.10)

$$- c^2 \cdot \text{rot rot } \mathbf{B} = \frac{\partial^2 \mathbf{B}}{\partial t^2} + \frac{1}{\tau_1}\frac{\partial \mathbf{B}}{\partial t} + \frac{1}{\tau_2}\frac{\partial \mathbf{B}}{\partial t} + \frac{\mathbf{B}}{\tau_1 \tau_2} \tag{27.26}$$

- Example:

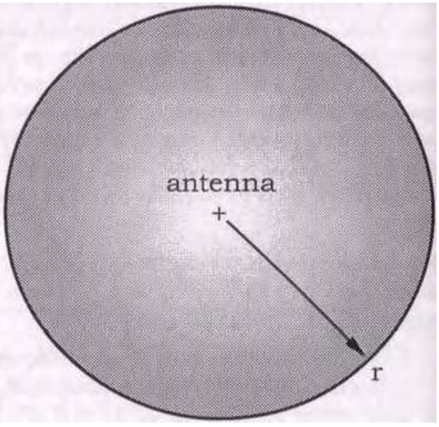

Spherical propagation of an electromagnetic wave in space

- The number π: $\pi = 3.14159$

straight		$\pi\cdot$		circular				spherical
r	·	$\pi\cdot 2$	=	$2r\cdot\pi$	·	$2r$	=	$\pi\cdot r^2\cdot 4$
line				**circum-ference**				**surface**
r	·	$\pi\cdot r$	=	$r^2\cdot\pi$	·	$r\cdot 4/3$	=	$\pi\cdot r^3\cdot 4/3$
line				**area**				**volume**

(29.4)

Fig. 29.5: The number π

29.5 Universal constants and constants of nature

Strictly speaking fundamental or universal constants can't exist in pure field physics at all. For the example of the speed of light has been shown that it merely are measurement constants (fig. 6.11). The numerical size is a consequence of the definition. The speed of light for instance has the today well-known value as a consequence of the fixing of a length unit and a time unit. If we change a unit, if we take „feet" instead of ,.meters", then another value results. If however the velocity changes, then the reference measure of our gauge changes along and we get the identical reading. Electric and magnetic field constant depend directly on the speed of light c $(\varepsilon \cdot \mu = 1/c^2)$, which solely by the fixing of the electric units takes a certain value. The inner structure of the potential vortices leads from Planck's quantum of action, to the elementary charge and to the countless atomic „constants", which all objectively seen aren't constants at all. They virtually all can be derived (chapter 7).

Sooner or later even the last natural scientist will realize, that nature does not provide "constants" at all. If „constants of nature"' did exist, as listed in textbooks and encyclopaedias, then they aren't consistent with causality, since we don't know the cause, why the factor should have exactly this size and no other. Behind every so-called constant of nature unnoticed is hiding a physical closed loop conclusion.

Fundamental constants only exist in mathematics. This can be shown very nicely for the example of the „fundamental field equation" (eq. 27.26), which has been derived from the new field theoretical approach (fig. 27.10). It is the description of waves and vortices in space and time, which indeed carries features of a „world equation". If one searches this equation for fundamental constants, then one altogether can find three: the number π, the number e and the Golden Proportion Φ. The speed of light c however occurs only as the mathematical factor characterizing the wave as a result of the defined units. If one would choose the units different, c as well could be made 1. With the fundamental numbers that procedure won't work. They don't depend on the definition of the units!

Let's consider the number π. The number π occurs every time as a proportionality factor if we transition from a straight line to a circle or further to a sphere, from a line to the circumference or further to the surface of a sphere and exactly so from a line to the area of a circle or further to the volume of a sphere. Since for all the special cases, which are derived from the fundamental field equation (the structure of the elementary particles, the atomic structure and in the same way again in the universe), the spherical symmetry dominates, the mathematical solution is determined by a corresponding spatial configuration of the number π. It has its cause neither in a physical relation of interactions, nor in the choice of the units, but only in the geometry.

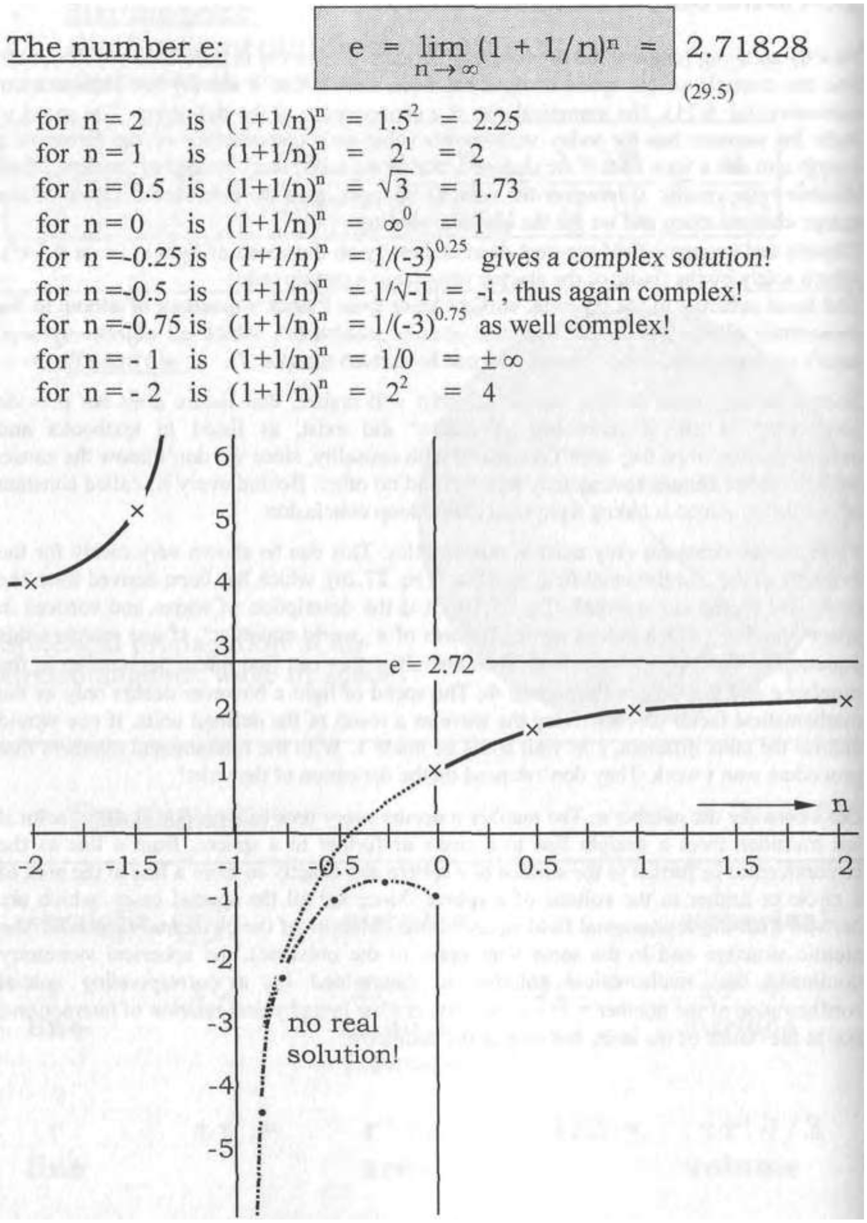

The number e: $e = \lim_{n \to \infty} (1 + 1/n)^n = 2.71828$ (29.5)

for $n = 2$ is $(1+1/n)^n = 1.5^2 = 2.25$
for $n = 1$ is $(1+1/n)^n = 2^1 = 2$
for $n = 0.5$ is $(1+1/n)^n = \sqrt{3} = 1.73$
for $n = 0$ is $(1+1/n)^n = \infty^0 = 1$
for $n = -0.25$ is $(1+1/n)^n = 1/(-3)^{0.25}$ *gives a complex solution!*
for $n = -0.5$ is $(1+1/n)^n = 1/\sqrt{-1} = -i$, thus again complex!
for $n = -0.75$ is $(1+1/n)^n = 1/(-3)^{0.75}$ as well complex!
for $n = -1$ is $(1+1/n)^n = 1/0 = \pm\infty$
for $n = -2$ is $(1+1/n)^n = 2^2 = 4$

Fig. 29.6: Forming the limiting value of the number e

29.6 The fundamental number e

In the fundamental field equation (27.26) a further irrational number is concealed, the number e. Whereas the left side of the equation (a) gives the spatial distribution, the right side (b-e) describes the temporal course of events. Besides the term constant in time (e) also first time derivations (c and d) and a second time derivation (b) occur. For a solution of this differential equation a function should be found, the derivations of which are again the function itself. This condition strictly speaking is fulfilled only by one single function, the e-function.

We used this property of the e-function already for the derivation of the Schrodinger-equation in chapter 5.6 and 5.7. There with the help of the e-function an approach was chosen, which leads to the well-known solutions of the Schrodinger-equation, which are considered to be secured experimentally. With that the number e controls the temporal relations of the fundamental field equation.

It might be helpful to take a closer look at the origin of the number e. It results from a consideration of limiting values: _

$$\boxed{e = \lim_{n \to \infty} (1 + 1/n)^n = 2.71828} \quad (29.5)$$

If one varies n and allows different values between $-\infty$ and $+\infty$, then a strange behaviour is showing. One indeed more and more approaches the well-known value of e = 2.72, as dictated by the definition of limiting values according to equation (29.5), the larger n is chosen. But in the opposite direction it looks less tidied:

$$\begin{array}{l|llllllll|l}
 & \text{for} & n = 1 & \text{is} & (1+1/n)^n & = & 2^1 & = & 2 & \\
 & \text{for} & n = 0{,}5 & \text{is} & (1+1/n)^n & = & \sqrt{3} & = & 1.73 & \\
 & \text{for} & n = 0 & \text{is} & (1+1/n)^n & = & \infty^0 & = & 1 & \\
 & \text{for} & n = -0{,}5 & \text{is} & (1+1/n)^n & = & 1/\sqrt{-1} & = & -i\text{, thus complex!} & \\
\text{and} & \text{for} & n = -1 & \text{is} & (1+1/n)^n & = & 1/0 & = & \pm\infty & (29.6)
\end{array}$$

Since the e-function inside the fundamental field equation is responsible for the temporal sequence, the interpretation of my colleague Prof. Dr. Preussker<i> gets a deeper sense. He says, it starts outside our imagination (at n = -1). Afterwards at first big chaos prevails. Mathematically seen some imaginary solutions arise. Finally the system is putting in order (from n = 0), to more and more approach the value e = 2.72 .

The number e is of fundamental importance and thereby holds unforeseen secrets. More mysterious and until now entirely misunderstood is the meaning of the Golden Proportion. Also this indivisible number can be found in the fundamental field equation. Since it is less known and more complicated to handle, it first shall be introduced.

<i>: H. Preussker: Der Wirbelring, Halstenbeck 2002

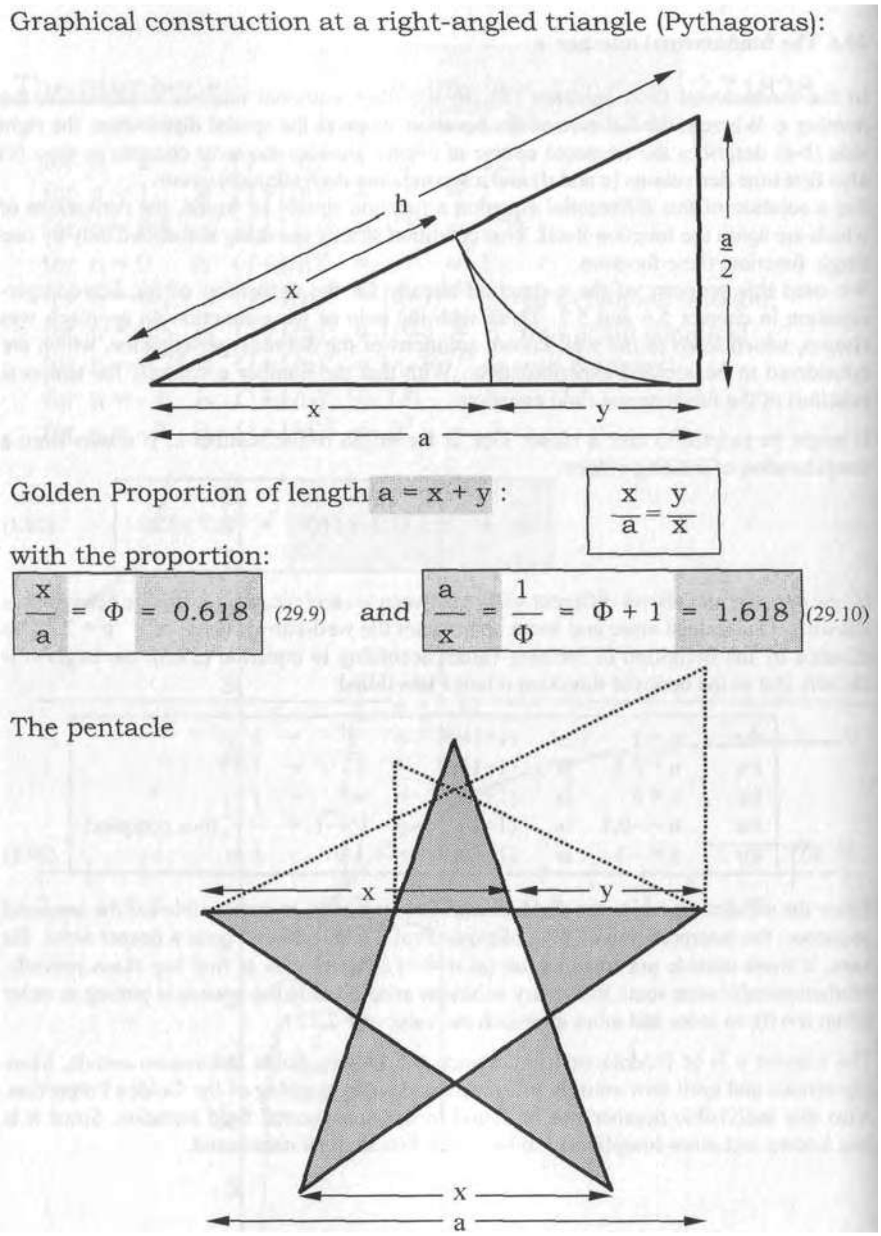
Graphical construction at a right-angled triangle (Pythagoras):
h
x
$\frac{a}{2}$
x
y
a
Golden Proportion of length a = x + y :
$\frac{x}{a} = \frac{y}{x}$
with the proportion:
$\frac{x}{a} = \Phi = 0.618$ (29.9) and $\frac{a}{x} = \frac{1}{\Phi} = \Phi + 1 = 1.618$ (29.10)
The pentacle
x
y
x
a

Fig. 29.7: Constructions for the Golden Proportion

29.7 The pentacle

During his visit of the Egyptian pyramids already two and a half thousand years ago the history writer Herodotus by his guide had been called attention to the circumstance that the Golden Proportion has been realized for the proportions of scale.
In the case of the pyramid of Cheops it even has been taken into account manifold, as we know today, but we hence still don't know why. There must be an intention behind it. Anyway a coincidental use can be eliminated, since the Golden Proportion cannot be handled in an easy way, neither graphically nor mathematically.
The Golden Proportion in addition plays an important role in the whole ancient architecture and not only there. It for instance occurs in the case of a very old symbol, the five edged star, which we can draw with one line without taking off. The well-known symbol also is called pentacle.

In the case of the Golden Proportion a straight line a is divided into two unequal halves. The larger half x thereby is 61.8 % of the straight line a. Already Pythagoras has researched and teached about this. Maybe he did know more about the purpose of this classification than all the mathematicians, archaeologists and art historians of today together.
For a graphical solution we assume a right-angled triangle. The task is to divide one leg of length a = x + y according to the Golden Proportion into two parts, the larger part x and the smaller part y. The second leg has the length a/2. According to the theorem of Pythagoras the length of the hypotenuse h is

$$h = \sqrt{a^2 + (a/2)^2} = (a/2)\cdot\sqrt{5} \quad (29.7)$$

If the length of the second leg (a/2) is subtracted from the hypotenuse, then this is the sought length x = h-a/2:

$$\boxed{x = (a/2)\cdot(\sqrt{5} - 1)} \quad (29.8)$$

The proportion of both length measures gives the constant Φ, which is characteristic for the Golden Proportion:

$$\boxed{\Phi = x/a = (\sqrt{5} - 1)/2 = 0.618} \quad (29.9)$$

This proportional number has a special property. If one adds 1 to the number and forms the reciprocal value of that, then the same number comes out again, thus:
with $\Phi = x/a$: with $a = x + y$:

$$\boxed{\Phi + 1 = \frac{1}{\Phi}} \quad \text{and} \quad \boxed{\frac{x}{a} + 1 = \frac{a}{x}} \quad \text{resp.} \quad \boxed{\frac{x}{a} = \frac{a - x}{x} = \frac{y}{x}} \quad (29.10)$$

With that the ratio of the length a and the larger section x is the same as the ratio of x and the smaller section y.

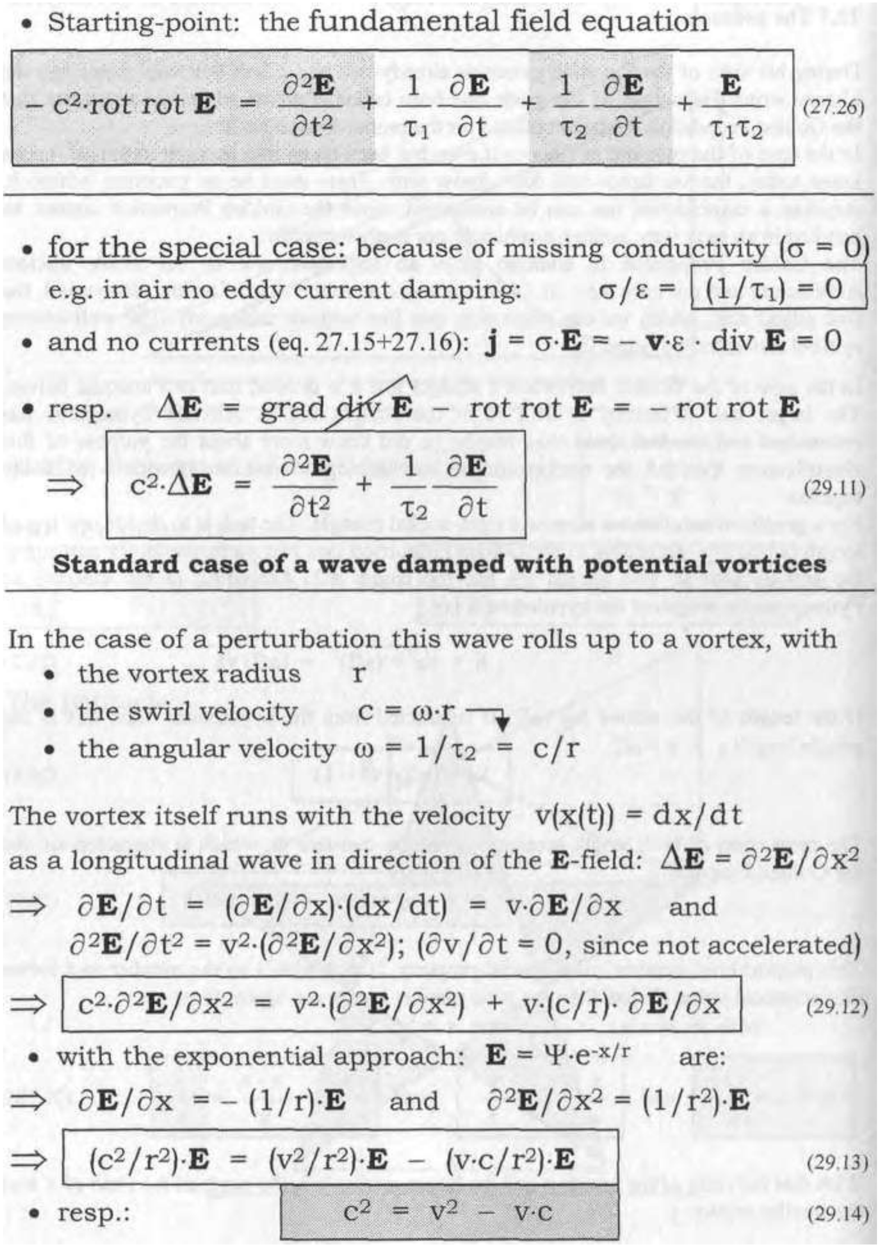

- Starting-point: the fundamental field equation

$$\boxed{-\,c^2\cdot\mathrm{rot\,rot}\,\mathbf{E} = \frac{\partial^2\mathbf{E}}{\partial t^2} + \frac{1}{\tau_1}\frac{\partial\mathbf{E}}{\partial t} + \frac{1}{\tau_2}\frac{\partial\mathbf{E}}{\partial t} + \frac{\mathbf{E}}{\tau_1\tau_2}} \qquad (27.26)$$

- for the special case: because of missing conductivity ($\sigma = 0$)
  e.g. in air no eddy current damping: $\sigma/\varepsilon = \boxed{(1/\tau_1) = 0}$
- and no currents (eq. 27.15+27.16): $\mathbf{j} = \sigma\cdot\mathbf{E} = -\,\mathbf{v}\cdot\varepsilon\cdot\mathrm{div}\,\mathbf{E} = 0$
- resp. $\Delta\mathbf{E} = \mathrm{grad}\,\cancel{\mathrm{div}}\,\mathbf{E} - \mathrm{rot\,rot}\,\mathbf{E} = -\,\mathrm{rot\,rot}\,\mathbf{E}$

$$\Rightarrow \quad \boxed{c^2\cdot\Delta\mathbf{E} = \frac{\partial^2\mathbf{E}}{\partial t^2} + \frac{1}{\tau_2}\frac{\partial\mathbf{E}}{\partial t}} \qquad (29.11)$$

Standard case of a wave damped with potential vortices

In the case of a perturbation this wave rolls up to a vortex, with

- the vortex radius r
- the swirl velocity $c = \omega\cdot r$
- the angular velocity $\omega = 1/\tau_2 = c/r$

The vortex itself runs with the velocity $v(x(t)) = dx/dt$

as a longitudinal wave in direction of the **E**-field: $\Delta\mathbf{E} = \partial^2\mathbf{E}/\partial x^2$

$$\Rightarrow \quad \partial\mathbf{E}/\partial t = (\partial\mathbf{E}/\partial x)\cdot(dx/dt) = v\cdot\partial\mathbf{E}/\partial x \quad \text{and}$$

$$\partial^2\mathbf{E}/\partial t^2 = v^2\cdot(\partial^2\mathbf{E}/\partial x^2); \ (\partial v/\partial t = 0, \text{ since not accelerated})$$

$$\Rightarrow \quad \boxed{c^2\cdot\partial^2\mathbf{E}/\partial x^2 = v^2\cdot(\partial^2\mathbf{E}/\partial x^2) + v\cdot(c/r)\cdot\partial\mathbf{E}/\partial x} \qquad (29.12)$$

- with the exponential approach: $\mathbf{E} = \Psi\cdot e^{-x/r}$ are:

$$\Rightarrow \quad \partial\mathbf{E}/\partial x = -\,(1/r)\cdot\mathbf{E} \quad \text{and} \quad \partial^2\mathbf{E}/\partial x^2 = (1/r^2)\cdot\mathbf{E}$$

$$\Rightarrow \quad \boxed{(c^2/r^2)\cdot\mathbf{E} = (v^2/r^2)\cdot\mathbf{E} - (v\cdot c/r^2)\cdot\mathbf{E}} \qquad (29.13)$$

- resp.: $\boxed{c^2 = v^2 - v\cdot c}$ (29.14)

Fig. 29.8: The calculation of an electromagnetic wave, which is rolling up to a potential vortex.

The big mystery concerning the harmony of the Golden Proportion gets a sober technical-physical dimension with the theory of objectivity. It determines within the fundamental field equation" (27.26) the rolling up of a wave into a vortex and vice versa. The Golden Proportion mathematically describes the process known as wave damping, as we can make ourselves clear.

29.8 The vortex, which is rolling up

For the case of a wave propagation in air or in vacuum, if no electric conductivity is present ($\sigma = 0$), the fundamental field equation is reduced to the two parts: the description of the electromagnetic wave and the potential vortex as a damping term. Now a solution of this partial differential equation (29.11) is sought. This only succeeds for a very particular course of spatial and temporal field.

If a wave for a field perturbation rolls up to a vortex, which we had worked out as a model concept, then the field oscillation continues to run with the speed of light, but this time in circles. With this consideration the relation between the angular velocity resp. the time constants and the radius of the circular vortex has been described ($\tau_2 = r/c$).

v(x(t)) = dx/dt is the not-accelerated velocity of propagation of a vortex. In that case v points in the x-direction radially to the outside. For the time derivation of the field vector E(x(t)) the chain rule should be applied. With that the field equation (29.11), defined in space and time, can be converted into an equation determined in v and x (29.12).

Finally we use the mentioned property of the e-function, which for first and second derivation again turns into itself, by choosing the approach of an exponential damping with $e^{-x/r}$ There remains a quadratic equation to determine the velocity v (29.14 and 29.15). From the two solutions of the quadratic equation only the one with positive velocity should be considered (29.17) and that would be 1.618 times the speed of light! (29.18).
If we subtract 1 from this value or form the reciprocal value, then in both cases the factor $\Phi = 0.618$ results, which is called the Golden Proportion (29.19).

Behind this clear, mathematically won result is hiding a deeper physical meaning. Obviously nothing can hinder a longitudinal wave and its vortices to be slower or faster than just with v = 1.618*c. Let us take the case that v = c, for which there even exist calculations in some textbooks<i>. Then as a result gets out that the longitudinal parts decrease very quickly and already can be neglected after $\lambda/2\pi$. I in any case interpret the
near-field zone of an antenna (fig. 29.9) such that within one sixth of the wavelength the vortices to a large extent have decayed.

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3.Aufl. Springer-Verlag Berlin 1986, Seite 335

- The quadratic equation 29.14: $\boxed{v^2 - v \cdot c - c^2 = 0}$ (29.15)
- has the solution: $\boxed{v = c/2 \pm\sqrt{c^2/4 + c^2}}$ (29.16)
- for the proportion v/c: $\boxed{v/c = (1 + \sqrt{5})/2 = 1.618}$ (29.17)
- The result $v/c = \underline{1.618}$ (29.18)
- resp.: $c/v = \Phi = \underline{0.618}$ = „Golden Proportion"(29.19)
- acc. to that is: $1 + \Phi = 1.618 = 1/\Phi = 1/0.618$(29.18*)

- Concerning the physical meaning of $v/c = 1.618$ (29.17)

In the domain:	vortex decay:	description:
$v = c$	within $\lambda/2\pi$	near-field zone
$v < 1.618 \cdot c$	only later	noise
$v = 1.618 \cdot c$	stability limit	direct conversion
$v > 1.618 \cdot c$	stable, no decay	neutrino radiation

=>In the ideal case at 1.618 times the speed of light at the stability limit, the natural and spontaneous conversion from vortex to wave and vice versa occurs.

=> Scalar wave antennas have their optimum degree of effectiveness in the Golden Proportion!

Fig. 29.9: Derivation of the Golden Proportion from the fundamental field equation.

For larger velocities, as Tesla has produced already 100 years ago, the stability of the vortices, but also their biological effectiveness increases. From a high-frequency technical viewpoint this is the domain of the noise, the domain of instable oscillating noise vortices. From a biological and medical viewpoint especially here the answers should be sought to questions concerning the electromagnetic environmental compatibility. This domain stretches from v = c to v = 1.618*c, the boundary case, where a spontaneous conversion of the wave into the vortex is possible. The Golden Proportion describes the stability limit, where the field vortices after they have been formed, don't decay anymore. From this velocity on the antenna resp. radiation source hasn't got a limited near-field zone anymore, since that now reaches to infinity. Only from this distinct velocity on I give the field vortices the name "neutrino".

29.9 The Golden Proportion

Whereas traditional antennas, which should emit electromagnetic waves, are optimized for signals with the speed of light, do scalar wave antennas have their optimum degree of effectiveness in the Golden Proportion. Thereby the goal of an emission of oscillating ring-like vortices is pursued. Vortices, which with a velocity of propagation of v = 1.618-c remain stable and make possible a spatially unlimited transmission of energy and information by scalar waves.

The velocity of propagation v is calculated from the product of wavelength times frequency. $v = \lambda \cdot f$. If the frequency of a transmitter is prescribed, then the geometry of the transmitting antenna decides on the degree of effectiveness of the antenna and on how many parts of transverse waves are emitted in relation to longitudinal waves. In the case of a broadcast antenna usually λ or $\lambda/2$ is chosen. For a scalar wave however the optimum is reached with

$$\boxed{\lambda_{\text{vortex}} = 1.618 \cdot \lambda_{\text{wave}} = 1.618 \cdot c/f.} \qquad (29.20)$$

With that the electrotechnical problem becomes a geometrical one, if it concerns the use of scalar waves. Crucial is the antenna geometry, and the Golden Proportion provides the necessary constructional instruction. Were the buildings in antiquity, which were constructed according to the Golden Proportion, technical facilities for scalar waves[i]? Did the builders have physically explainable and mathematically provable guidelines? At this place, by the derivation of the Golden Proportion from the fundamental field equation, there arise completely new aspects for the judgement and interpretation of buildings especially from antiquity. If we have understood their way of functioning, then we will be able to learn much from that for our own future and for the future construction of scalar wave devices. Concluding the seminar we hence deal with antiquity.

ancient temple = short wave station

dedicated to one god = fixing of the frequency

supreme god Zeus, = range of the short wave,
father of all gods = all SW wave bands

priest, representative of the god = amateur radio operator, with licence to transmit

high priest = chief intendant

Pontifex Maximus, „topmost bridge builder" = chairman of the authority and the telegraph offices

oracle<i> = telegraphy receiver

runes, cuneiform writing = telegraphy symbols

metre, hexameter = increase of redundancy

oracle priest = telegraphy interpreter

tripod<ii> = reception key, electro-acoustical converter

looking at intestines, rite of sacrificing = reading off convulsions, electro-optical converter

temple books = news-notes

seer, who looks into the god world = amateur radio operator, at telegraphy reception

Homer = ancient radio reporter

Fig. 30.1: Little dictionary for ancient radio engineering (1).

<i>: Lamer: Worterbuch der Antike, Kroner Verl. Bd.96 under "oracle" stands: „Lat. oraculum = site of speaking; particularly: site, where a god speaks; then: that, what the god says". According to the encyclopaedia oracle priests inspired by a god „simply were frauds, who lived of the ignorance of the public".

30. Scalar wave technology in antiquity

The end of the book about potential vortices and their propagation as a scalar wave shall form an impressive example, where as many of the derived wave aspects as possible have an effect. It shall be proven that already in antiquity radio engineering based on scalar waves has been used. The proof starts with a thesis.

30.1 Thesis:

The temples in antiquity all were short wave broadcasting stations. And energy from the field served as an energy source, so e.g. the earth radiation in the case of temples of terrestrial gods. In the case of the solar god the radiation of the sun was used, whereas for the temples, which were dedicated to the planetary gods, the neutrino radiation arriving from the planets served as an energy source.

If the temple was dedicated to a particular god, then the name of the god was representing the used frequency of the broadcasting company. The corresponding wavelength, resp. the respective god, understandably was "immortal".

Not so the broadcasting technicians on duty, who as human beings naturally were mortal, who took turns in the studio as members of the priest council and who merely had to impersonate the god Apollo, Poseidon etc. by the name of the broadcasting company, if they went on air. Only for the news editor Homer and for few of his colleagues we actually know the names of the persons behind the scenes.

In the temple books the texts have been recorded, which a god and its broadcasting company have received. The chosen metre served the easier detection and correction of transmission errors.

Here often a lot of fantasy was necessary, for which reason the reception facilities commonly were described as oracle[<i>]. The reception of the news as a rule took place on an altar. Thereby the direct effect of scalar waves on man e.g. in the case of the so-called temple sleep or the indirect influence on biological systems, e.g. on the intestines of slaughtered animals, was evaluated.

A further development of the telegraphy was the tripod technology[<ii>], in which case by turning of the polarization plane individual symbols and letters were transmitted up to the transmission of the spoken word with the help of a special wavelength modulation. That far the thesis reaches, which now should be proven.

<ii>: Lamer: Worterbuch der Antike, Kroner Verl. Bd.96 under "tripod" stands: „the tripod is a dedication gift to gods, a honorary gift for winners. That one pleased them with the gift of a cooking pot, is strange; one has tried to find the reason, but until now without success". In the encyclopaedia further is advised: „tripod as a means for spiritual insight?!"

godology = high frequency technology
god name = RDS, station identification
members of a family of gods = broadcasting studios of a broadcasting company
Pantheon, temple of all gods = extremely broadband FM broadcast station
crown = antenna netting
gifts for sacrificing = broadcasting fees
place of sacrificing = place of a node of the standing wave
earth radiation = power supply
homage of a weekday = time restriction of the operation of the station
Zeus „forges" thunderbolts = electrostatic blows, when a temple is oscillating
ritual act = technical provision for transmission and reception
Pythia of Delphi<i> = radio telephone operator, receptionist
Cella (marrow of temple) = tuned cavity
obelisk<ii> = antenna rod

Fig. 30.2: ____ Little dictionary for ancient radio engineering (2).

<i>: Lamer: Worterbuch der Antike, Kroner Verl. Bd.96, under „Pythia"(gr.=the asker): In the temple of Delphi was a chasm. A tripod was standing over it, on which the Pythia was seated, if she gave oracle. Comment: "It was flashy how odd the Pythia was sitting; inconveniently enough, on the cooking pot of the tripod. Ancient pictures, which show her that way, still weren't proven, that it was this way ..."

<ii>: Lamer: under „Obelisk" stands; „The obelisks probably were clocks".

30.2 Prerequisites

The argumentation has to be made on mathematical-physical foundation. The prerequisite for that are the 29 chapters of before. The following points could be demonstrated and derived:

1. The wave equation (inhomogeneous Laplace equation) describes the sum of two wave parts, where
2. every antenna emits both parts,
3. the transverse part, known as electromagnetic wave (Hertzian wave)
4. and the longitudinal part (Tesla radiation) termed scalar wave by the discoverer, better known as antenna noise.
5. The wave equation mathematically describes the connection of both wave parts in general and the conversion of one part into the other in particular, thus
6. the rolling up and unrolling of waves in field vortices (measurable as noise).
7. The transition takes place proportionally to the Golden Proportion, as resulted from the derivation (chapter 29.7 - 29.9)

With the last point the electrotechnical problem becomes a geometrical problem, if it concerns the use of scalar waves. The geometry of the antenna is crucial. Thereby the Golden Proportion provides the necessary direction for construction.
That justifies the assumption that the buildings in antiquity, which were built according to the Golden Proportion, were technical facilities for scalar waves. Maybe the builders had specifications that had physical reasons and could mathematically be proven.
At this place there result completely new aspects for judging and interpreting buildings especially from antiquity through the derivation of the Golden Proportion from the fundamental field equation. If we have understood their way of functioning, then we will be able to learn much from that for our own future and for the future construction of scalar wave devices.

As a further prerequisite for the ancient broadcasting technology enough field energy should be at disposal. We proceed from the assumption that

1. the earth magnetism and the cosmic neutrino radiation are tightly hanging together by the processes in earth's core,
2. the earth magnetism in antiquity verifiably was approx. thousandfold stronger than today (proven by gauging of pieces of broken pot),
3. as a consequence the neutrino radiation in antiquity as well must have been thousandfold stronger and
4. the cosmic neutrino radiation has served the transmitting plants of antiquity as an energy carrier,

any thought is absurd to reject the technical function of a temple only because it today can't be reproduced anymore. The artistic and aesthetical viewpoints, which are put into the foreground by art historians because of ignorance about the true function, rather are secondary.

The terms used to describe the broadcasting technology in antiquity in the last 2000 years have experienced a shift of meaning, so that a translation in our linguistic usage of today is necessary. The adjacent dictionary should help in that case.

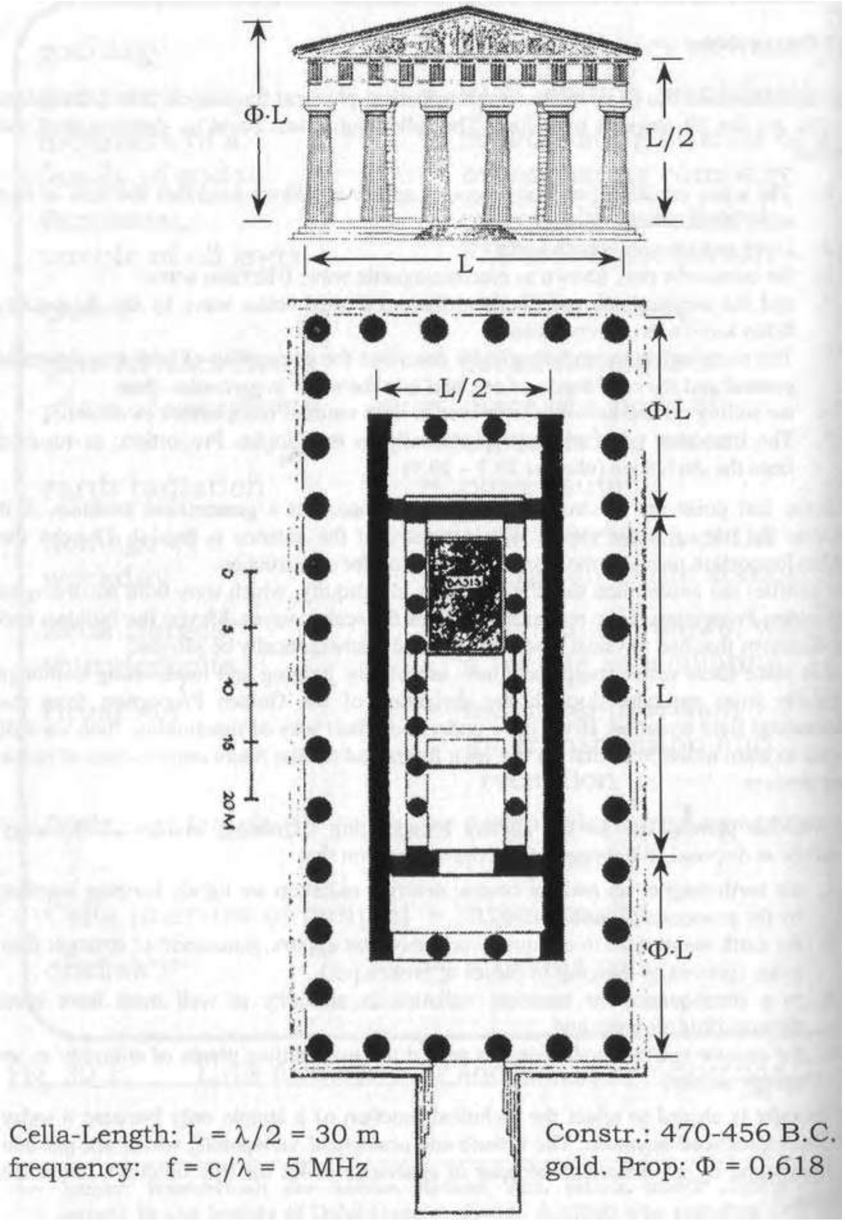
Φ·L
L/2
L
L/2
Φ·L
L
Φ·L
0
5
10
15
20 M
Cella-Length: L = λ/2 = 30 m
Constr.: 470-456 B.C.
frequency: f = c/λ = 5 MHz
gold. Prop: Φ = 0,618

Fig. 30.3: The Golden Proportion of Zeus-temple in Olympia.<i>

<i>: K. Schefold: Die Griechen und ihre Nachbarn, Propylaen Kunstgeschichte Berlin Bd. 1, Abbildungen von Seite 249

30.3 Approach

Let's to some extent proceed from the knowledge of textbook physics currently present in high frequency engineering and give a well trained engineer the following task, which he should solve systematically and like an engineer: He should build a transmitter with maximum range at minimum transmitting power, thus a classic task of optimization. Doing so, the material expenditure doesn't play a role!

After mature deliberation the engineer will hit upon it that only one solution exists: He decides on a telegraphy transmitter at the long wave end of the short wave band, at f = 3 MHz, which corresponds to a wavelength of $\lambda = 100$ m. There less than 1 Watt transmitting power is enough for a radio communication once around the earth. That also has something to do with the conditions of reflection of the radio waves at the ionosphere. Our engineer learned:

the index of refraction $n = \sqrt{1 - K \cdot N/f^2}$ (30.1)
with: $K = 80.5 * 10^{-6}$ [cm^3/sec^2] (= constant)
N = electron concentration [$electrons/cm^3$]
f = frequency of the transmitter [MHz]

Put into words: the refraction of a radio wave in the range of the short waves is the larger, the smaller the frequency is. The end of the short wave range is reached at 3 MHz. That thus explains the choice of frequency.

And he optimises further. Next the engineer remembers that at high frequencies, e.g. for microwave radiators, not cables but waveguides are used, since these make possible a considerable better degree of effectiveness. In the case of the waveguide the stray fields are reduced by an alignment and concentration of the fields in the inside of the conductor. In the case of antennas however the fields scatter to the outside and cause considerable stray losses. He draws the conclusion that his transmitter should be built as a tuned cavity and not as an antenna!

As a result the engineer puts a building without windows in the countryside with the enormous dimensions of 50 m length ($=\lambda/2$) and 25 m ($=\lambda/4$) resp. 12.5 m ($=\lambda/8$) width. The height he calculates according to the Golden Proportion to increase the scalar wave part. Those approximately are the dimensions of the Cella without window of Greek temples.

For the operation of such a transmitter in antiquity apparently the noise power of the cosmic radiation was sufficient, which arrived at the earth starting from the sun and the planets. By increasing the floor space also the collected field energy and the transmitting power could be increased, so that also from the perspective of the power supply the temple with the largest possible wavelength at the same time promised the largest transmitting power, so at least in antiquity.

Our engineer further determines, that he will switch the carrier frequency on and off at a predetermined clock pulse. Thus he decides for radiotelegraphy. The advantage of this technique is a maximum increase of the reception range. For that the signals at the transmitter have to be coded and at the receiver again deciphered. By means of the encryption of the contents these are accessible only to the „insiders", who know the code; prerequisite for the emerging of hermetism and eventually a question of power!

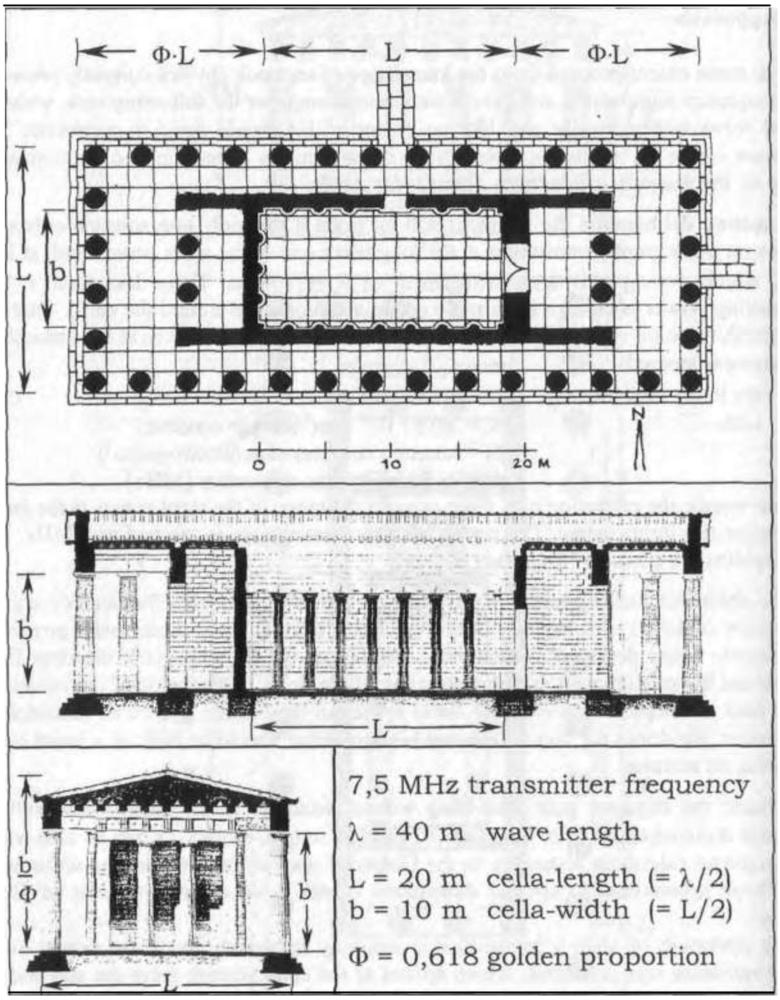
Φ·L
L
Φ·L
L
b
0
10
20 M
N
b
L
b/Φ
b
L
7,5 MHz transmitter frequency
λ = 40 m wave length
L = 20 m cella-length (= λ/2)
b = 10 m cella-width (= L/2)
Φ = 0,618 golden proportion

Fig. 30.4: Example Tegea, temple of Athena Alea.[i]
Built 350/340 B.C.

<i>: G. Gruben: Die Tempel der Griechen, Wissenschaftliche Buchgesellschaft Darmstadt 1986, 4. Aufl. Seite 130

<ii>: E. Horst: Konstantin der Grosse, Eine Biographie, Classen Verlag 1985 Dusseldorf, 2.Aufl., S. 89.

<iii>: E. Horst: Konstantin der Grosse, Eine Biographie, Classen Verlag 1985 Dusseldorf, 2,Aufl., S. 33.

30.4 Circumstantial evidence

Not everyone, somehow participating in send receive engineering, at the same time also was inaugurated in the entire secret knowledge. Most priests only knew as much as they necessarily needed to know to fulfill their tasks. Thus a temple priest, who was presented an enciphered text and who should bring this on the air, not necessarily at the same time needed to know the content of the text or the code. The same of course also was valid for the sacrificing priest acting in the receiving station. The Vestal virgins for instance had to present the received text to the Augures, by whom they were supervised and controlled.

But who wanted to introduce a new god in the gods heaven and perhaps even himself be worshipped as a god, should have complete command of both the broadcasting technique and the reception technique. In ancient Egypt the Pharao at least once a year had to prove, that he still was in command of the technique. Otherwise he was replaced. For a person with security clearance that at the same time was a death sentence.

In the historical facts numerous pieces of circumstantial evidence can be found, which can be considered to be evidence for the thesis of the operation of send receive engineering in antiquity. One now perhaps understands, why the rulers were put an antenna netting over their head, a so-called crown, or why the Augures could survey the land with a flat Tesla coil in their hands (fig. 16.10).

Direct evidence is present as well. It can be found in ancient texts. But it is questionable if historical texts concerning ancient radio engineering have been translated correctly. The talk is about oracles, mystery cult and earth prophesy if the receiver is meant. The predominantly technically uneducated historians attest the Romans a defective sense of time, because their couriers surely could not cover the long ways across the Roman empire so fast at all, if they read in the Latin texts: "They sent by courier to the emperor in Rome and got for answer...". The answer of the emperor namely already arrived at the squad at the latest in the following night. The correct translation should read: "they cabled" or "they broadcasted to the emperor in Rome and got for answer..."[ii].

Such a big empire as the Roman Empire actually only could be reined by means of an efficient communication. Cicero coined the word: "We have conquered the peoples of the earth owing to our broadcasting technology...[iii]"! The term broadcasting technology from ignorance is translated with piety. If engineers however rework the incorrect translations, then one will discover that numerous texts tell of the broadcasting technology, that thus correspondingly much direct evidence exists concerning the practical use of this technology.

For the Roman military transmitters, which formed the backbone of the administration of the empire, the reading off of the information from observations of nature like the bird flight or from felt signals of a geomanter was too unreliable. They read off the information from the rhythm of the convulsions of the intestines of freshly slaughtered animals. In the case of the dead animals on the altar every extrinsic influence was excluded. But the enormous need of slaughter cattle was a disadvantage. Who wanted to have information, first of all had to bring along an animal, which then was „sacrificed" the god, or better say, which was abused as a receiver for a particular transmitter. Thereby the innards served as a biosensor and as a receiver for the news.

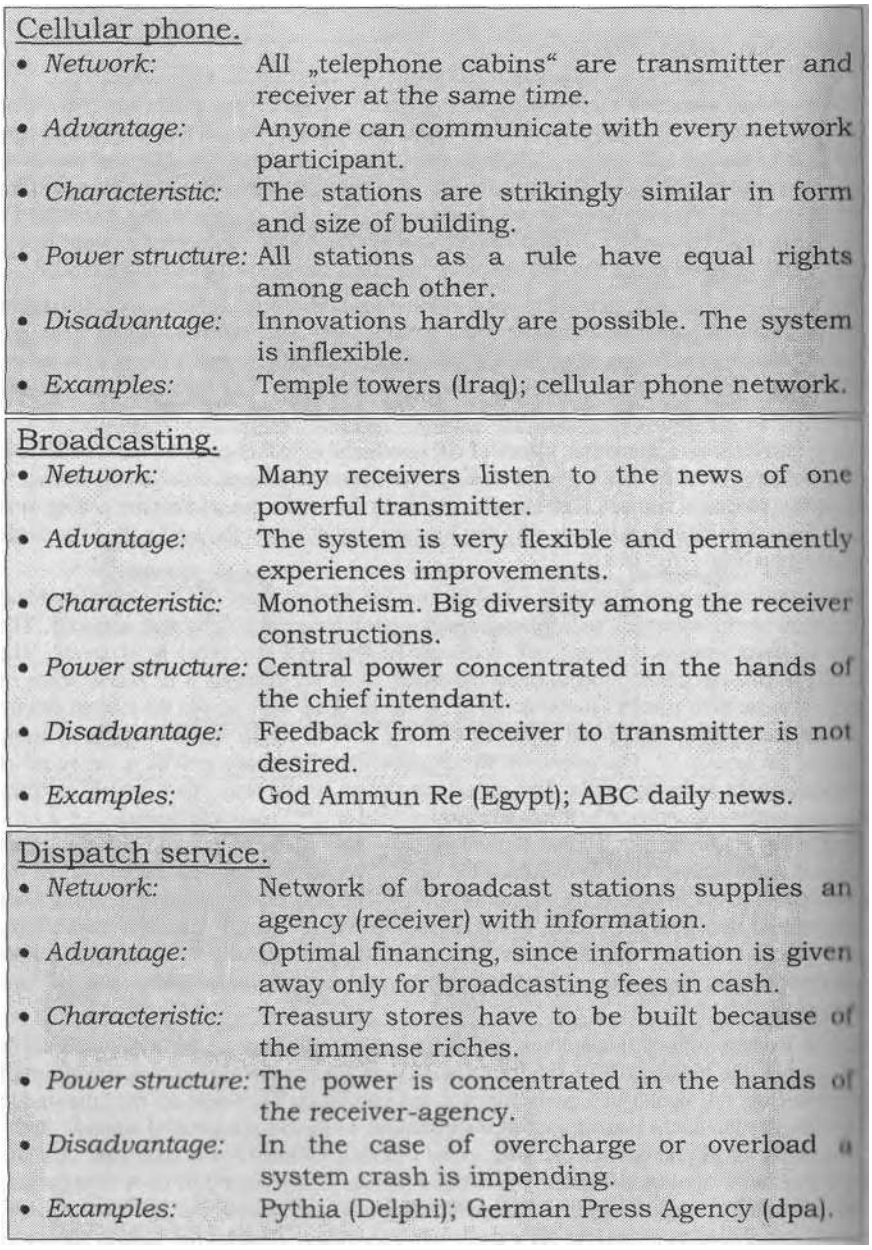

Cellular phone.

• *Network:*	All „telephone cabins“ are transmitter and receiver at the same time.
• *Advantage:*	Anyone can communicate with every network participant.
• *Characteristic:*	The stations are strikingly similar in form and size of building.
• *Power structure:*	All stations as a rule have equal rights among each other.
• *Disadvantage:*	Innovations hardly are possible. The system is inflexible.
• *Examples:*	Temple towers (Iraq); cellular phone network.

Broadcasting.

• *Network:*	Many receivers listen to the news of one powerful transmitter.
• *Advantage:*	The system is very flexible and permanently experiences improvements.
• *Characteristic:*	Monotheism. Big diversity among the receiver constructions.
• *Power structure:*	Central power concentrated in the hands of the chief intendant.
• *Disadvantage:*	Feedback from receiver to transmitter is not desired.
• *Examples:*	God Ammun Re (Egypt); ABC daily news.

Dispatch service.

• *Network:*	Network of broadcast stations supplies an agency (receiver) with information.
• *Advantage:*	Optimal financing, since information is given away only for broadcasting fees in cash.
• *Characteristic:*	Treasury stores have to be built because of the immense riches.
• *Power structure:*	The power is concentrated in the hands of the receiver-agency.
• *Disadvantage:*	In the case of overcharge or overload a system crash is impending.
• *Examples:*	Pythia (Delphi); German Press Agency (dpa).

Fig. 30.5: Three radio technical network structures, with an example from antiquity and from present time.

30.5 Radio technical concepts

In planning and constructing radio technical networks only a few possible concepts exist. It is interesting that at least one historical example can be specified for every concept. That shows that all possibilities were tried at least once. The three most important concepts are presented here:

A. Cellular phone.

Every bigger city between Euphrates and Tigris, which thought the world of itself, had at its disposal already in antiquity a temple tower. Such a temple tower was a „telephone cabin" in the form of a pyramid as a transmitter and a receiver temple at the top, to where the receptionist adjourned to the so-called temple sleep. Discipline was required, since all the time only one priest was allowed to broadcast. All others could listen to him doing so. If he was ready, he closed his contribution with a fixed symbol or term („over") and the next one could continue. This is a classic link-up, where anyone communicates with every network participant.

The stations all were strikingly similar in form and size of building, like one phone box resembles another. In that way a further development of the cellular phone system hardly was possible and that has a technical reason, as the building of a tower in Babylon has shown us. This tower namely had gotten the ambitious builders too big, so that the frequency of the Mesopotamian radio network had been left and instead a foreign network could be received, the code of which no-one could understand. The result was a confusion of language and the order to stop the building.

B. Broadcasting.

Millions of TV spectators every evening look in the ABC news or another daily journal of a TV Channel. In the case of broadcasting thus many receivers listen to the news of a powerful transmitter. With that the whole plenitude of power is concentrated in the hands of the chief intendant. In antiquity he called himself high priest. If he went on the air, he used the logo of the god that he had to represent. Today the logo of the broadcasting company is shown in a corner of the TV screen. Even this very day feedback from the receiver to the transmitter hardly is possible contingent on principle. The problems with nationally controlled broadcasting, with politics controlled by the media all are not new. The monotheism in ancient Egypt with the claim of lordship of the main god Ammun Re is an example from antiquity.

C. Dispatch service.

In ancient Greece the technical structures and with that also the power structures had been turned around. At that time a big network of broadcast stations, which continually was extended by a policy of settlement ordered by the gods, supplied a central and correspondingly powerful agency with information per radio.

Who wanted up to date news, could call for these in the agency with seat in Delphi, but he had to pay for it. To accommodate the broadcasting fees in form of gold and gifts whole treasury stores had to be built. Measured by the commercial success the ancient news network has remained unmatched, and can't be compared with pay-TV or todays dispatch services, like dpa. If the network however becomes too big, uncontrollable and it lacks discipline, then it sometime will crack and the system crashes.

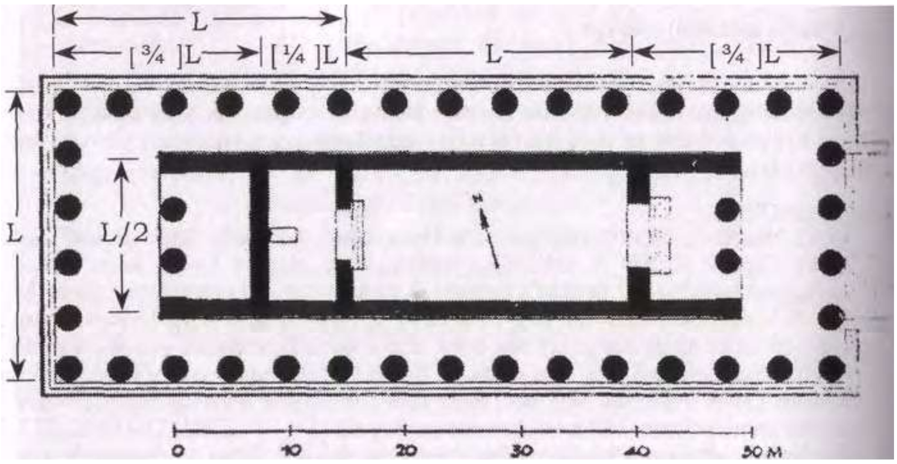
L
[¾]L
[¼]L
L
[¾]L
L
L/2
0
10
20
30
40
50 M

Fig. 30.6 A: Hera-temple of Selinunt 460-450 B.C.[i]
(Corresp. to the plans of the Roman architect Vitruvius)

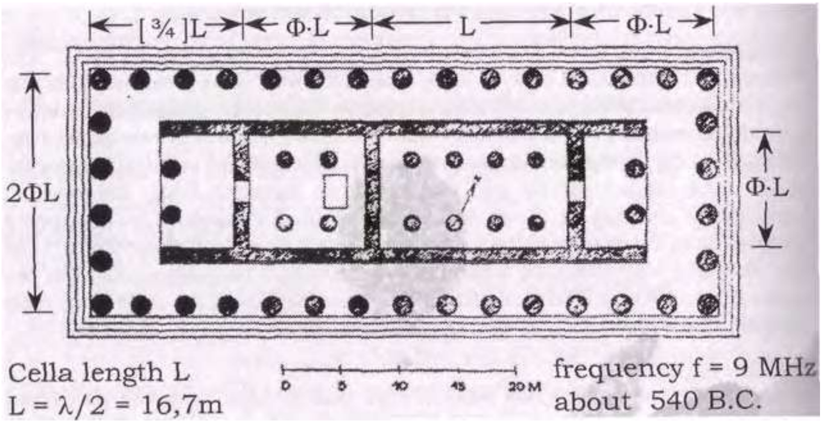
[¾]L
Φ·L
L
Φ·L
2ΦL
Φ·L
Cella length L
L = λ/2 = 16,7m
0
5
10
15
20 M
frequency f = 9 MHz
about 540 B.C.

Fig. 30.6 B: Apollo-temple of Korinth[i]
(Alternative interpretation, use of the Golden Proportion Φ).

<i>: K.Schefold: Die Griechen und ihre Nachbarn, Propylaen Kunstgeschichte Berlin Bd. 1, Abbildungen von Seite 241, 250

<ii>: Vitruvius (Marcus Vitruvius Pollio): Zehn Bucher uber Architektur, Ubers. von Dr. K. Fensterbusch, Wissenschaftl. Buchges. Darmstadt 1987, 4. Aufl., 3. Buch, 1. Kap.: Von den Symmetrien der Tempel, Seite 137

<iii>: Vitruvius (dito), 4.book, 4.Chap.: Vom Tempelinnern und dem Pronaon, p. 187

30.6 Wireless telegraphy

Radio engineering 100 years ago also started with telegraphy. Thereby the high frequency carrier is switched on and off. With this technique Marconi succeeded in a radio transmission over the English Channel (1899) and over the Atlantic Ocean (1901).
As next step the amplitude modulation (AM) followed. Thereby the HF-carrier is overlapped with the low-frequency signal of a sound carrier in such a way, that the amplitude fluctuates on the beat of the LF-signal. As a disadvantageous effect, also noise signals will overlap, from which the quality of reception will suffer.
Only the frequency modulation (FM), where the LF-signal is transmitted as temporal fluctuation of the frequency, brings an improvement. The annoying amplitude noise hence has no effect in the case of FM.
It easily can be recognized, how the development of the modulation techniques follows the urge for technical improvement and optimization. That in antiquity hasn't been different, for which reason the progress of development took place in the same order.

The broadcasting technology of the ancient gods started with the wireless telegraphy. This is expressed in the architecture. Since electric resonant circuits or other frequency determining equipment weren't at the disposal of the engineers in antiquity, the determination and allocation of the broadcasting channels had to take place by means of the wavelength. The formation of a standing wave in the Cella, the innermost sanctuary of a temple, occurs if its length corresponds to half the wavelength of the HF-carrier. The Roman architect Vitruvius calls the wavelength the „basic measure", from which results "the system of the symmetries". He writes: „The design of the temples bases on symmetry, to which laws the architects should adhere meticulously. "<ii>. „The length of the temple is partitioned in such a way that the width is equal to half the length, the Cella itself including the wall, which contains the door, is one fourth longer than wide. The remaining three fourths, which form the Pronaon, should protrude until the antae of the wall and the antae should have the thickness of the pillars"<iii>.
If we recalculate ourselves, then the partitioning in 3/4 to 5/4 produces a proportion, which conies quite close to the Golden Proportion. In building a temple nothing is left to chance, after all it concerns the construction of a tuned cavity, capable of self-resonant oscillations with favourable emission behaviour.
From the outside one can't see if a telegraphy transmitter has been changed over to speech transmission with AM. The HF-carrier merely isn't switched off anymore, i.e. the priests let the temple oscillate without interruption. Newly added for AM is an electroacoustic coupling. For that many temples were retrofitted with a mouthpiece. Newly built AM transmitter temples conclude the Cella with a round apse. Because of this acoustically conditioned construction the Cella length didn't have a fixed value anymore and the transmission frequency had become variable. Measured in the middle of the apse the wavelength was larger than at the sides, so that on the beat of the spoken word not only the amplitude of the field distribution in the interior of the temple, but in addition also the frequency of the selfresonant oscillation was changed.
A typical example of such an architectonic hybrid form of AM and FM is situated in Rome. Because due to the frequency variation more than only one wave band was occupied and the temple consistently carries the names of two deities. It is the temple of Venus and Roma.

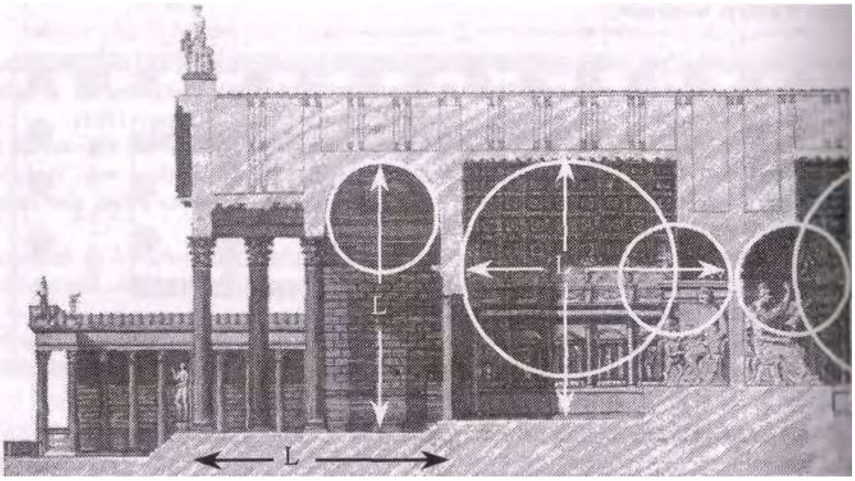

Diameter bigger circle D_1= 22 m (6,8 MHz),
small circle D_2 = 11 m; und $L = D_1 + 1/2 D_2$ = 27,5 m (5,5 MHz)

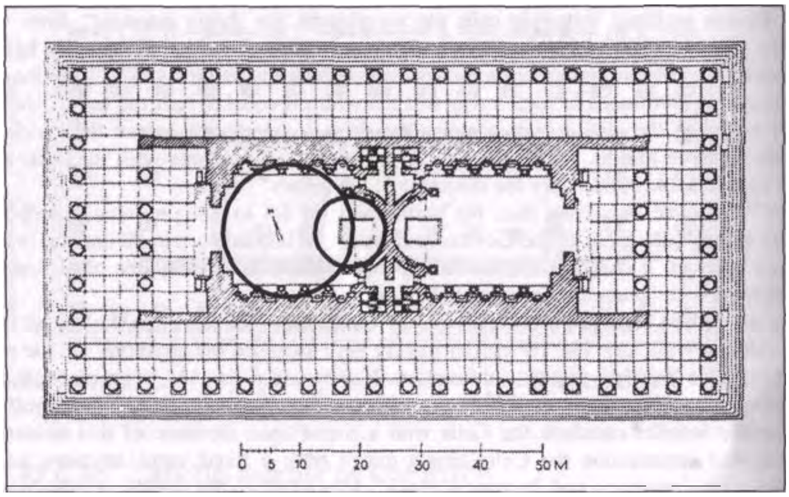

Fig, 30.7: Temple of Venus and Roma; Rome 136/37 A.D.[i,ii]

<i>: A. Springer: Das Altertum, A. Kroner Verl. Leipzig 1915, 10. Aufl., S. 518
<ii>: T. Kraus: Das romische Weltreich, Propylaen Kunstgesch. Berlin Bd. 2, S. 161
<iii>: Lamer: Worterbuch der Antike, Kroner Verl. Bd.96 unter "Tempel", Haufigkeit

30.7 AM temple technology

The low-frequency signal (LF), which should be transmitted by a transmitter with amplitude modulation, lies in the range between 16 Hz and 16 kHz. If it only concerns the transmission of speech information, then the bandwidth can be reduced to 300 to 3000 Hz. In the case of mixing the low-frequency useful signal with the HF-carrier, thus in the case of the modulation of the carrier in the rhythm of the LF, two side bands arise. These lie close to the carrier frequency and are formed from this once by the addition and once by the subtraction with the frequency of the LF-signal. Let's take the temple of Venus and Roma with a transmission frequency of 6.8 MHz. If sound of 3 kHz should be transmitted clearly understandable, then the dimensions of the Cella had to be varied for just 8 mm for a corresponding Cella length of $\lambda/2$ = 22 m. As a curiosity the niches in the side walls in the case of this temple however allow a considerably larger bandwidth of more than 10% instead of the necessary 0.04% in the case of AM.

In the case of the Greek originals, the Cella however has smooth walls, from which follows that the temples were designed ideally narrow band. The Greeks apparently operated predominantly telegraphy transmitters, for which the side bands coincide with the carrier.

The argumentation indeed has remained unchanged: The modulator being narrow band and simple to realize speak in favour of the telegraphy being the "original form" of all modulation techniques. Also the rediscovery of the broadcasting technology by Heinrich Hertz succeeded as telegraphy signal. In addition the range is bigger than for any signal modulated with sound frequency.

As the calculation example has shown, also pure AM transmitters work very narrow band, and this is particularly important for low transmission frequencies, if many transmitters want to use the favoured SW band between 3 and 10 MHz at the same time. With AM one thus accommodates the maximum number of broadcasting channels in a particular frequency range, for instance the 80-meter band, without these interfering with each other too much. But that also was badly needed. Conclusions about the everyday life of broadcasting in antiquity by all means are possible because of the enormous number of temple installations, which logically were permanently used. Only in Rome there existed up to 200 temples<iii>!

Who goes in search of broadcasting stations with a modern short wave receiver, for instance in the 80 m band between the countless telegraphy transmitters, fast gets an idea of what had been up in the air already 2000 years ago. No ancient city would build several temples on a single Acropolis, if only one single one could have been used. All temples broadcasted with each time another carrier frequency because of different dimensions. For this reason the temples, which stood side by side, as a rule were dedicated different gods. An acknowledgement, "the air just being free", in addition hardly was possible, because of the often-found spatial distance between the temple installations and the respective oracle. Between the transmitter of the god Apollo in Didyma and the receiver, the oracle of Milet, for example lie approx. 20 kilometres. The only possible conclusion is that in antiquity there was broadcasted on all channels simultaneously regardless of other gods and their transmission frequencies. As is well-known there rather prevailed a situation of competition between the gods, since like today a large number of listeners meant great importance, influence and power and eventually also worship, more gifts and more receipts from broadcasting fees.

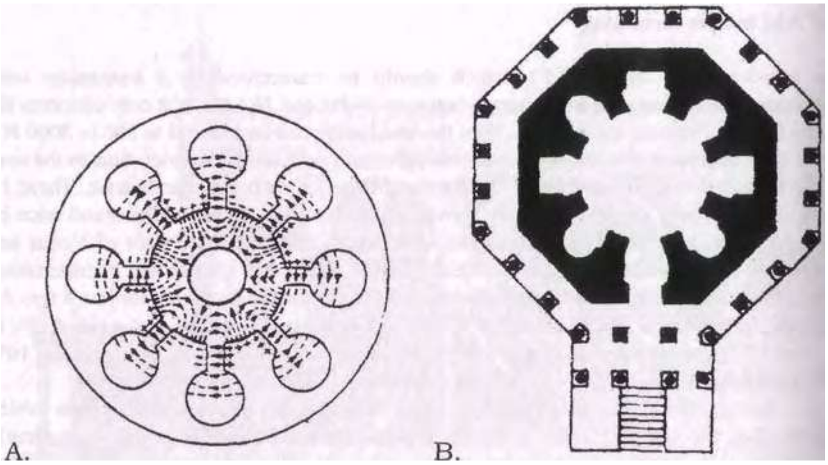

Fig. 30.8: Comparison of a magnetron (A), a microwave radio tube[i] and the temple (B) in the palace of the emperor Diokletian, Split.[ii]

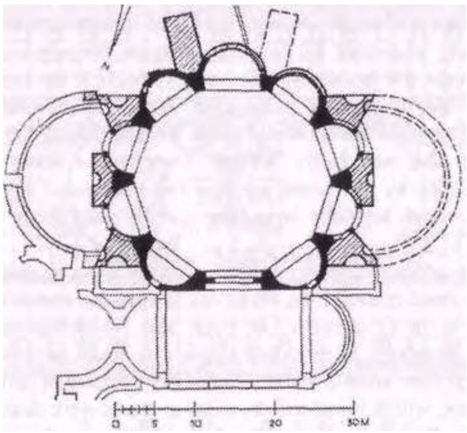

Fig. 30.8.C: Temple of Minerva Medica, Rome, 320 A.D.[ii]

<i>: K. Simonyi: Physikalische Elektronik, 8.4 Das Magnetron, S. 665

<ii>: T. Kraus: Das rom. Weltreich, Propylaen Kunstgesch. Bd. 2, S. 194, 196

<iii>: If one however wants to verify this, in antiquity already common manner of PM broadcasting technology, then we need a broadband short wave receiver with phase-demodulator. With such a receiver even today any time a conclusive argumentation should be possible that this sort of SW-PM technology actually works.

30.8 Modulation of acoustic signals

In the case of sound it in essence concerns longitudinal waves, which propagate in all directions in space with the velocity of sound. But this merely is an average velocity, since the air molecules strictly speaking oscillate with the sound frequency in the direction of propagation. In this way the velocity of sound one time is increased and the next moment to the corresponding extent reduced. If the molecules already carry out oscillations of themselves, e.g. thermal motion, then both oscillations overlap; i.e. the motion of itself is modulated with the sound frequency.

Following the here presented derivation potential vortices are formed in every dielectric, thus also in air, and these are modulated if overlapped with sound waves. Vortices however do not form a distinct frequency, but entirely on the contrary a uniform frequency mixture in the form of white noise.

The overlap thus also is noise unless certain noise frequencies are favoured. This can be effected by means of a spatial body tuned to a certain wavelength. To be considered are cuboid cavities, as in the case of Greek temples or waveguides and cylindrical objects as in the case of round temples or magnetrons. The building form causes the favouring of a certain frequency and the integer harmonic frequencies belonging to it. If this frequency now lies in the high frequency range, then it is emitted by the spatial body as an electromagnetic wave, in the case of waveguides and resonant circuits for reason of the small dimensions as microwave radiation and in the case of the Greek temples as short wave radiation. If one in addition produces an acoustic signal in the frequency determining spatial body, then this signal automatically will modulate the high-frequency signal.

The result of the modulation is the overlapping of the sound wave with the high-frequency carrier wave. The change of the active length, thus the wavelength of the HF-carrier causes a change of frequency. Such a change on the beat of the sound frequency is called FM (frequency modulation). For that we imagine a spatial body being excited because of its length by potential vortices to a high-frequency oscillation at the self-resonant frequency.

From flow dynamics is known, how easy the distribution of vortices in space can be disturbed. Already words spoken in space are able to influence the potential vortices, which in the case of the temples were used as energy carrier. The longitudinal sound waves reflected at and thus returning from the Cella wall rigid for sound of a Greek temple will push these potential vortices back on the beat of the sound and with that shorten the active length. This means, that the carrier frequency is modulated with the sound frequency. The carrier oscillation thus permanently changes its phase on the beat of the sound signal, for which reason this particular kind of modulation is described as PM, as phase modulation.

Nowadays PM is used only seldom and in the SW-range not at all because of the big need of wave bandwidth. Merely at higher frequencies PM occasionally is used in radiotelephony. In the case of the frequency modulation usual in broadcasting (e.g. UHF), the change of frequency takes place on the beat of the sound amplitude and not of the sound frequency as with PM. Therefore it is not possible to receive phase modulated signals, which are produced by means of the acoustic coupling of appropriately formed spatial bodies<iii> with commercial FM receivers.

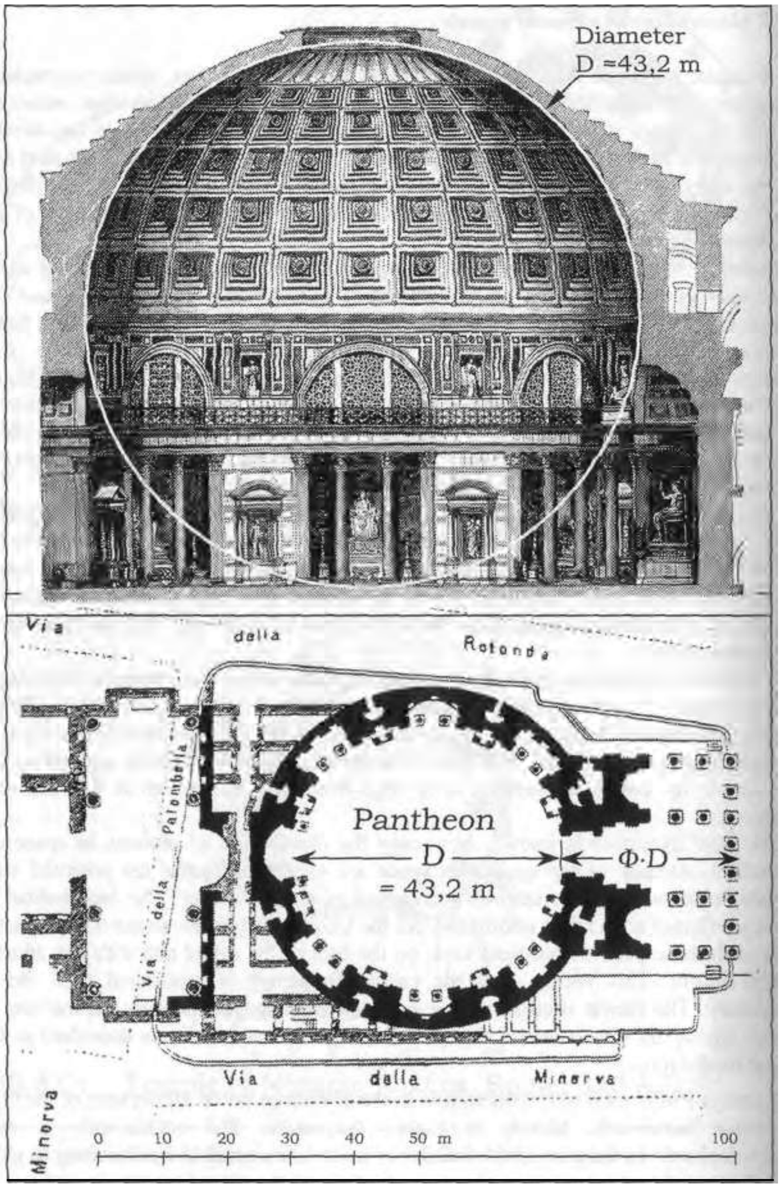
Diameter
D =43,2 m
Via della Rotonda
Via della Palombella
Pantheon
D
= 43,2 m
Φ·D
Via della Minerva
Minerva
0
10
20
30
40
50 m
100

Fig. 30.9: The Pantheon in Rome, the „temple of all gods".<i>
Diameter D = 43,2 m, Golden Proportion: $\Phi = 0.618$
Pronaon-(atrium-)length: $\Phi \cdot D = 0.618 \bullet 43.2 = 26.7$ m
Built under emperor Hadrian 118/ 119-125/128 A.D.

30.9 Broadband FM broadcasting technology

If I speak against a flat wall, then every point on the wall has another distance to my mouth. The sound waves thus aren't reflected simultaneously, what leads to big modulation distortions. Therefore the sound wall should be curved in such a way, that all signal paths are the same length (barrel vault, apse, etc.). In the case of point sound sources there results as an optimum a hemisphere, for instance a dome. The building hence even today tells us, which frequency and which modulation method had been put to use. The architecture of sacral buildings, e.g. pointed arch or round arch, thus hardly has been a question of aesthetics.

For his temple of Venus and Roma designed by himself emperor Hadrian had to listen to severe criticism among others of Apollodor of Damascus. The temple was too broadband for an AM transmitter, however with a modulation depth of just 11 percent not broadband enough for a phase modulated FM transmitter.

Emperor Hadrian however also had the courage to build midst in Rome a temple calculated completely new and designed as a pure FM transmitter, the Pantheon, which means temple of all gods. In the language of the technician it is a transmitter for all frequencies.

This domed structure indeed doesn't leave out one single frequency. With a modulation depth of almost 100 percent it is designed for maximum loudness. With that the Pantheon uses all available frequencies, for which reason the name temple of all gods really is no exaggeration. Into the Pantheon exactly fits a sphere with a diameter of 43.2 meter. That corresponds to a minimum frequency of 3.47 MHz, situated in the range of the short waves. The floor however is not domed, but horizontal. That, up to the basis of the dome, results in exactly half the height and a maximum frequency of 6.94 MHz.

The construction ensures that between the simple and the double diameter any desired wavelength can be produced. Above the given maximum frequency of the basic oscillation the harmonic waves, which are produced as well and can't be avoided at all, are attached without a break. These occupy the wave bands up to the double, triple, quadruple frequency and so forth.

For this and only for this reason a maximum frequency was chosen, which corresponds to exactly the double value of the minimum frequency. The operation takes place to the limit, where the transmitter would interfere with itself, in the way that the used basic oscillation would overlap its own harmonic waves. That then sounds like two people talking at the same time. The voices would be distorted out of recognition, as can't be expected else in the range of the harmonic waves.

The Pantheon has been planned and built as a phase modulated basic wave transmitter according to purely academic rules of Hadrian. The temple impressively demonstrates the precise engineering detailed knowledge of the Pontifex Maximus and his broadcasting priests in ancient Rome<ii>.

<i>: A. Springer: Die Kunst des Altertums, A. Kroner Verl. Leipzig 1915, 10. Aufl., S. 517/518, daraus: die Zeichnungen ram Pantheon in Rom, 115-125 n.Chr.

<ii>: The collection of material concerning scalar wave technology here is aborted and continued in an own book in narrative and with that easier to read writing style. The title is: "Broadcasting Gods".

approach	derivation	result
b:= postulate [potential vortices Vol.1]	what if $\mathbf{b} \neq 0$?	**derivation** of the **Schrödinger eq.**
duality vortices replace quanta [potential vortices Vol.2]	from $\mathbf{H} \Leftrightarrow \mathbf{E}$ from $\mathbf{j} \Leftrightarrow \mathbf{b}$ [and chapter 3.3]	**calculation** of the **quantum properties** [potential vortices V.2]
causality [see chap. 3]	*gen.* ⇔ *special case* **field** ⇔ quanta cause ⇔ consequence (instead of: quanta ⇔ field)	vortex field free of quanta (instead of: irrotational field of quanta)
vortex physics analogous to flow dynamics [chap. 3.4 - 3.5]	**vortex + anti-vortex** (Demokrit/cosmology) [chap. 3.6 + 27.2]	coming off of vortices if $\tau_1 = \tau_2$ **model of atom** [chap. 5.5 + 5.8]
equations of transformation: $\mathbf{E} = \mathbf{v} \times \mathbf{B}$ $\mathbf{H} = - \mathbf{v} \times \mathbf{D}$ [chap.6.5 + 27.8]	for $\mathbf{E}(\mathbf{r}(t))$ and for $\mathbf{H}(\mathbf{r}(t))$: $\text{rot}\,\mathbf{E} = \text{rot}\,(\mathbf{v} \times \mathbf{B})$ $\text{rot}\,\mathbf{H} = \text{rot}\,(\mathbf{D} \times \mathbf{v})$ [chap. 27.9]	extended **field equations:** $\text{rot}\,\mathbf{E} = -\mathbf{b} - \partial\mathbf{B}/\partial t$ $\text{rot}\,\mathbf{H} = \mathbf{j} + \partial\mathbf{D}/\partial t$ [chap. 27.10]
equations of transformation: $\mathbf{E} = \mathbf{v} \times \mathbf{B}$ $\mathbf{H} = - \mathbf{v} \times \mathbf{D}$ [s.a. chap. 16.1]	**wave equation** inhom. Laplace eq. $\Delta\mathbf{E} = \text{grad div}\,\mathbf{E} - \text{rot rot}\,\mathbf{E}$ $\Delta\mathbf{E} = (1/c^2)\cdot\partial^2\mathbf{E}/\partial t^2$ [chap. 27.13] [chap.22.4]	transverse + longitudinal wave parts: **scalar waves,** noise, etc.
equations of transformation: [chap. 28]	vortex overlap **length measures depending on field**	**gravitation,** **temperature,** **theory of objectivity**

Fig. 30.10: (All) Many roads lead to Rome.

30.10 Epilogue

The preparation for a seminar or a lecture always starts with a collection of material. This should be considerably more detailed than the material to be communicated, since it doesn't get well at the students and also other participants, if the state of knowledge of the lecturer already is exhausted with the scope of his lecture. For this reason the collection of material must also include alternative derivations and areas of knowledge, which thematically rather are marginal, about which mustn't be reported, which possibly not even can bear criticism but still are in the public discussion. For a collection of material, which wants to be considered to be comprehensive and complete, it is important that no area and no theme has been overlooked.

The here presented collection of material with its 650 pages has become correspondingly voluminous. After all the material has been collected over an 8 year period and has been strung into the book in the order of working. That of course complicates reading the book, because individual aspects are repeated several times, but often also in a different context and each time lighted from another side. If the reader somewhere has the feeling, he only has turned in a circle, then has deceived himself. He indeed moves spirally in a circle, like in real life, but he doesn't come out there, where he started. After one turn he is richer with the experience of this spiral turn. In whole science the advancement takes place as a spiral movement and one can count oneself fortunate, as long as the spiral has an ascending slope! This notion should solace the reader, who has undertaken the torture to work through the complete collection of material.

Students also have reported, they had devoured my book like a thriller and a colleague, who had acquired it at a conference in Switzerland, was digging so much in the lecture that he forgot to get off the train timely.

It is a special concern to give reasons for the necessity of an extension of available field theory. To achieve this goal several derivations (fig. 30.10) can be found in my books: from a postulate, from causality, from duality, from vortex physics, from the equations of transformation, etc. added are at least a dozen derivations of other authors from various publications, who at most are cited. With that the goal is pursued that all approaches, which are conceivable and worth discussing, can be put side by side and tested for their efficiency.

Since it isn't the task of a collection of material to answer this question, this must be done by the hearer resp. the reader of the books. He is prompted to find the answer himself! That leads to an intense contention with the theme and that exactly is the reason for the otherwise rather unusual step to make a collection of material open to the public. Objections and criticism of the content of a relation of matters or also only of the representation of the context is wished explicitly. This also isn't valued as criticism of the author or of the superordinated set of difficulties, which in principle isn't possible at all for a collection of material.

Honorary functions of

Prof. Dr.-Ing. Konstantin Meyl

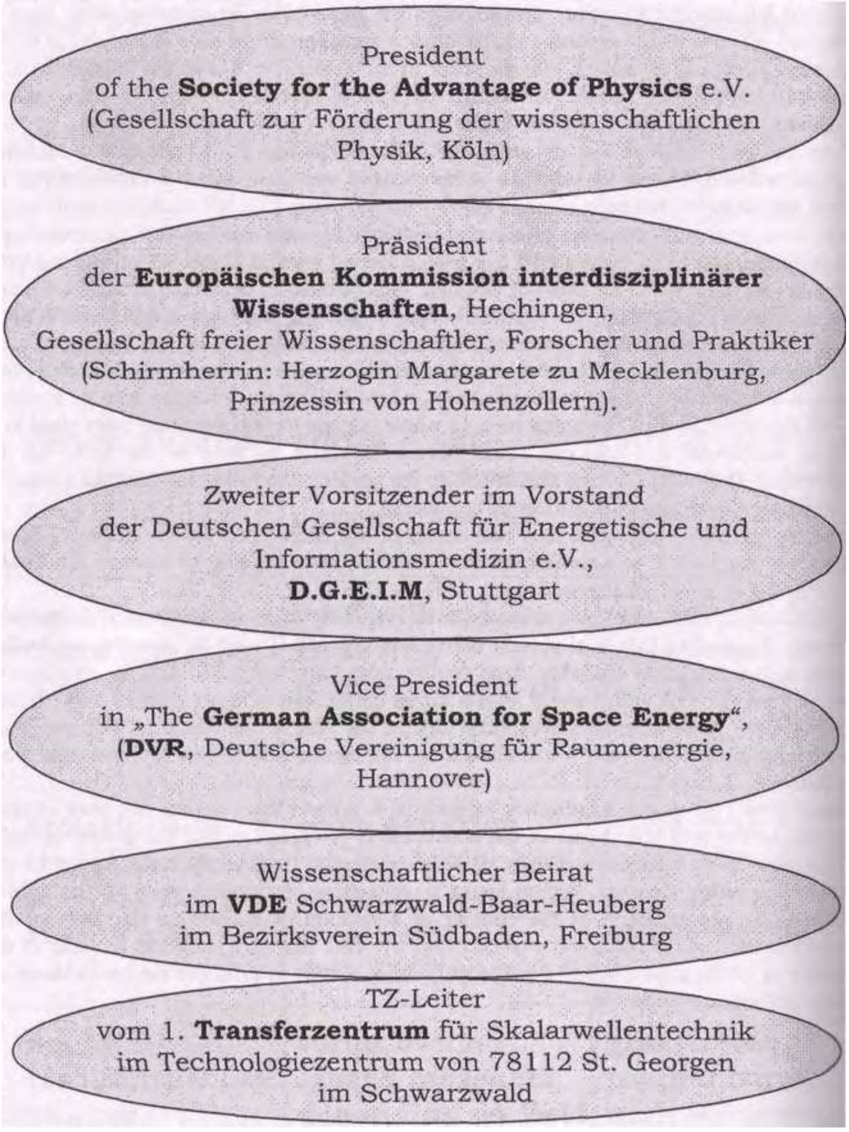

President
of the **Society for the Advantage of Physics** e.V.
(Gesellschaft zur Förderung der wissenschaftlichen Physik, Köln)

Präsident
der **Europäischen Kommission interdisziplinärer Wissenschaften**, Hechingen,
Gesellschaft freier Wissenschaftler, Forscher und Praktiker
(Schirmherrin: Herzogin Margarete zu Mecklenburg, Prinzessin von Hohenzollern).

Zweiter Vorsitzender im Vorstand
der Deutschen Gesellschaft für Energetische und Informationsmedizin e.V.,
D.G.E.I.M, Stuttgart

Vice President
in „The **German Association for Space Energy**",
(**DVR**, Deutsche Vereinigung für Raumenergie, Hannover)

Wissenschaftlicher Beirat
im **VDE** Schwarzwald-Baar-Heuberg
im Bezirksverein Südbaden, Freiburg

TZ-Leiter
vom 1. **Transferzentrum** für Skalarwellentechnik
im Technologiezentrum von 78112 St. Georgen
im Schwarzwald

Fig. 30.11: Honorary functions of the author (December 2003, selection).

There in principle is no necessity at all to discuss with everybody a collection of material, which I have compiled exclusively for own events. It is absolutely sufficient, if few, but then qualified experts have thoughts about the content and communicate them to me. They can feel certain that I don't hand down their judgement to other persons. For that I also have been blamed by ignorant colleagues, who in complete self-overestimation have the opinion, they should be informed about everything. Right is rather, that my private correspondence is of no concern to anybody, since I find it important that as much readers as possible express themselves frank, what only functions, if they can feel certain that they afterwards won't be involved in public mud-wrestling. Therefore I keep still as regards other persons however curious they might be.

Indeed over and over again pseudo scientists turn up, who have the erroneous opinion that scientific arguments would take place on some internet-forums, where one can descant at will and anonymous, hidden behind an alias, where the intellectual firebug can feel safe, not afterwards being blamed for his crimes.

No, science takes place entirely different. A new theory will be able to establish, if it is right and important and if it is used for practical uses. Losers are those authors of a theory, for the elaboratation of which nobody is interested. What does a publication in a journal, however renowned, mean, if nobody reads it and nobody needs it? Most new ideas and approaches go under without notice in today's flood of publications, for who has got the time to read all essays in full?

Desperate they turn to me hoping, at least I could understand their concern. I then invite these scientists to a congress of the „Society for the Advantage of Physics", of which I am the president and offer them a forum, where they are able to present their ideas to an expert public. Not all lectured ideas prove to be sound, but not seldom a physical concern, which should be taken very serious, is behind it.

For a long time the real scientific controversy doesn't take place anymore at the universities and their congresses, where hardly someone dares to lecture arguments against the convention. Too fast he would be expelled as outlaw from the honourable society. From time to time however also from these circles colleagues dare anonymously or privately, as they emphasize, into the alternative events of lobbies or clubs, to astonished find out that real science there still is practised and that there is discussed about ideas, which they have given up thinking or have forbidden themselves to think about. The employment activity, so they excuse their thinking prohibition, allegedly doesn't allow it.

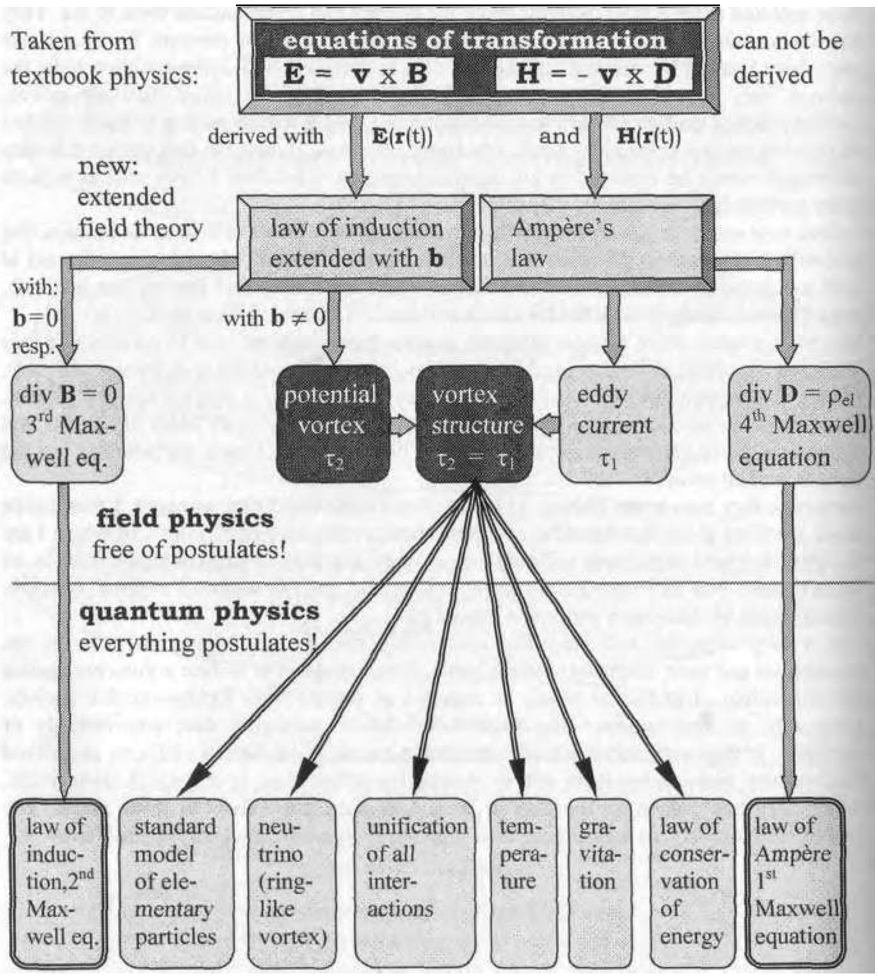
Taken from textbook physics:
equations of transformation
E = v x B
H = - v x D
can not be derived
derived with
E(r(t))
and
H(r(t))
new: extended field theory
law of induction extended with b
Ampère's law
with: b=0 resp.
with b ≠ 0
div B = 0 3rd Maxwell eq.
potential vortex τ2
vortex structure τ2 = τ1
eddy current τ1
div D = ρel 4th Maxwell equation
field physics free of postulates!
quantum physics everything postulates!
law of induction, 2nd Maxwell eq.
standard model of elementary particles
neutrino (ring-like vortex)
unification of all interactions
temperature
gravitation
law of conservation of energy
law of Ampère 1st Maxwell equation

Fig. 30.12: Derivations of postulates and axioms (part 1).

Be that as it may, the approaches and derivations contained in my collection of material are considered to be controversial, and that is good that way! The public takes notice and professional circles are occupied with the ideas. With that half the way to success already is brought off! It now concerns to knock off all points for their soundness individually, because the next step will be to search a way to the goal as unassailable as possible, which is able to convince also the biggest sceptic. The final version, which then should appear in accredited peer-reviewed journals and in a scientific book concerning the theory of objectivity, goes in this direction. From the numerous approaches in the end only one will be used, and from the countless, in the collection of material listed aspects only the noncontentious ones will remain.

The dispute, which in the current stage can't be avoided, yes even is desired, however shouldn't deceive about the fact that it here doesn't concern persons, improper vanity or some image cultivation, but simply and solely concerns the matter!

In fig. 30.12 and 30.13 is represented, what it concerns. Field physics and quantum physics don't form, as in common practise, an insurmountable opposite, but even complement each other! Below the stripline a little selection of the today in current use quantum physical postulates can be found. The number of newly introduced „constants of nature" and postulates permanently is increasing, a circumstance, which hardly can be mediated to the common sense. The bracket is missing, which interlinks all postulates, or the common source from which they can be derived causally.

In the progress of the three-part edition in this question, essential for physics, already a satisfying and in addition real efficient answer has been found in the domain of field physics. The coupling marked by individual derivations can be found above the stripline and it is entirely new, apart from the dashed indicated derivation (fig. 30.13), as given by Prof. Bosse (TU Darmstadt) in his textbook.

An approach in principle can be chosen freely. In the case of the superordinated field theory two equations of transformation form the approach, which already is laid down in textbooks and secured experimentally. That's why the whole field theoretical derivation manages without one postulate! It is pointed to the fact that these equations on their part can't be derived and should be interpreted rather philosophically than physically. From this approach the extended field theory is derived directly, without a need to add or discard a term. The extended field theory consists of the well-known law of Ampere extended with the dielectric displacement D by Maxwell, and of Faraday's law of induction, which experiences an extension with the vector of potential density b by means of the derivation. Doing so we assume that the field pointers of the electric and the magnetic field strength depend on the spatial coordinate r and through this indirectly also on the time t: $E(r(t))$, $H(r(t))$.

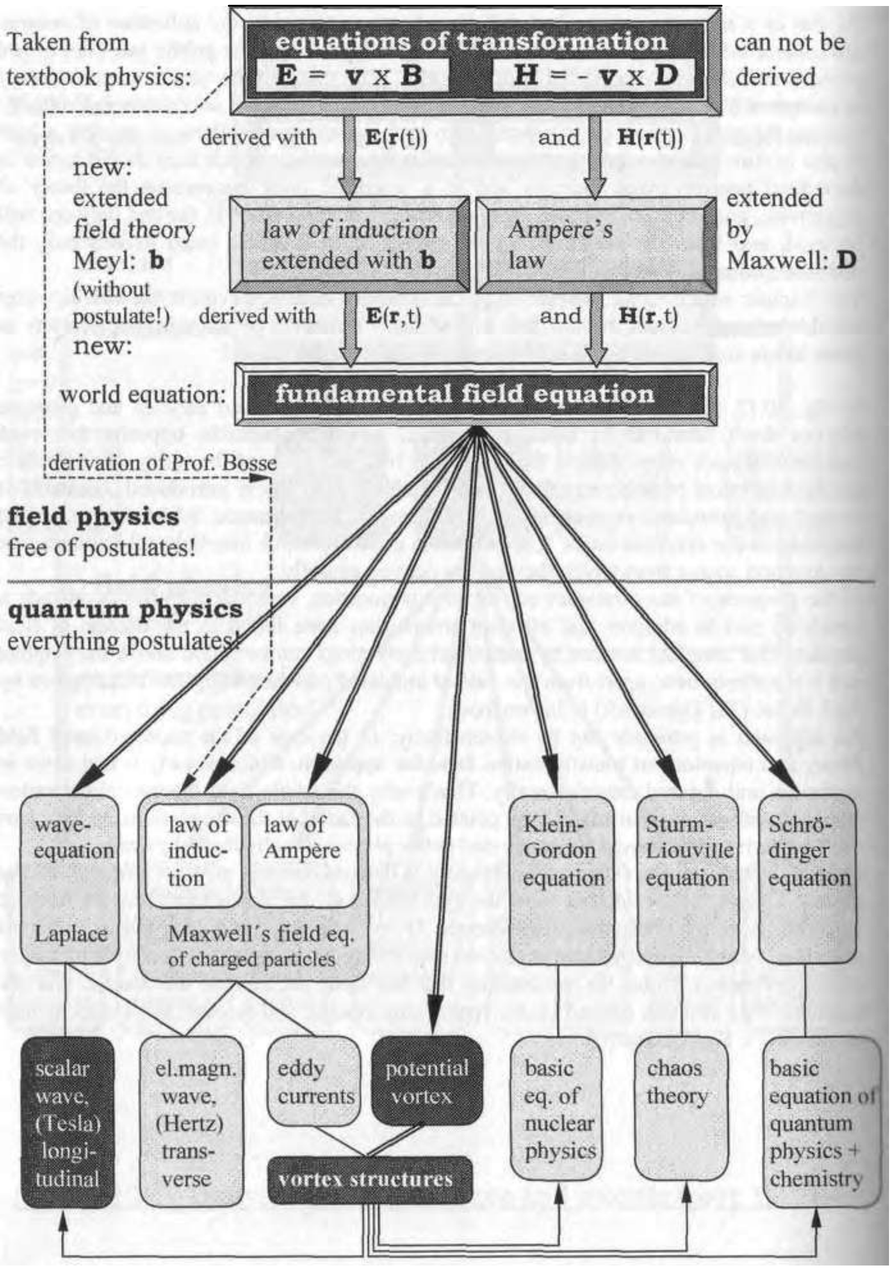
Taken from textbook physics:
equations of transformation
$\mathbf{E} = \mathbf{v} \times \mathbf{B}$
$\mathbf{H} = -\mathbf{v} \times \mathbf{D}$
can not be derived
derived with $\mathbf{E}(\mathbf{r}(t))$ and $\mathbf{H}(\mathbf{r}(t))$
new: extended field theory Meyl: **b** (without postulate!)
law of induction extended with **b**
Ampère's law
extended by Maxwell: **D**
derived with $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{H}(\mathbf{r},t)$
new: world equation:
fundamental field equation
derivation of Prof. Bosse
field physics
free of postulates!
quantum physics
everything postulates!
wave-equation
Laplace
law of induc-tion
law of Ampère
Maxwell's field eq. of charged particles
Klein-Gordon equation
Sturm-Liouville equation
Schrö-dinger equation
scalar wave, (Tesla) longi-tudinal
el.magn. wave, (Hertz) trans-verse
eddy currents
potential vortex
vortex structures
basic eq. of nuclear physics
chaos theory
basic equation of quantum physics + chemistry

Fig. 30.13: Derivations of postulates and axioms (part 2).

If the somewhat more general case is counted up, in which case apart from the spatial a direct temporal dependency is present (E(r,t) and H(r,t)). then further additional terms appear in addition to the extended field equations, which need an explanation also for the case that they are zero. The physical interpretation would implicate a longer treatise, which however can be circumvented, as shown here, by constraining the field pointers, which absolutely is allowed according to the slogan: for the case E(r(t)) and H(r(t)) chosen from many possibilities the extended field equations come out exactly in the form, as they are required and suitable for the further calculations. Who has got to calculate other cases, can do that as he likes, but doing so he should not get lost. Maxwell's field equations are contained in the solution and with that also continue to be valid. Their disadvantage however is that without the extension b not a single quantum physical postulate can be derived. If we add this extension and insert the equations into each other without addition and without cuts also this time a central solution is the result, which is called fundamental field equation.

The derivation is known as well from the Maxwell theory, in which case it is common practise, to use the general approach (E(r,t) and H(r,t)), what we in accordance with the textbooks can do in the same manner. The extension however brings two additional and extremely significant terms. Since the fundamental field equation has eigenvalues under certain boundary conditions and describes structures, various quantum postulates come out from it, from the quantum properties of the elementary particles over the Schrodinger equation and the inhomogeneous Laplace equation up to the derivation of the Golden Proportion. That justifies the assumption that this possibly is the long sought-for world equation!

Even I, as the initiator, was totally surprised by the found derivation of the most important quantum physical postulates and axioms. One just is doing it the right way and already everything fits together! I am not aware of any theory, which would be able to achieve something roughly comparable. The since long sought-for ,,Theory of Everything", the big unification theory really falls into ones lap. The known interactions are the free and easy result of analysing the field lines of electric and magnetic field strength (fig. 30.12). Physical phenomena, which until now were considered to be incompatible, like e.g. waves, noise or the temperature with the utterly insufficient concepts of the mechanisms for the conversion of one form of energy into another, can be represented consistently with the fundamental field equation as the rolling up of a wavelike field oscillation to a vortex oscillation and as conversion of the noise vortices in the case of a vortex contraction down to atomic dimensions as thermal oscillation, which we treat as vortex losses. There exists no alternative to such unified schemes of things, as makes it possible in abundance the theory that I have founded, considering the two conditions, that on the one hand in the case of the derivation only known regularities are used, by completely doing without postulates and that on the other hand laws are applied and adhered to, also physical laws.

The new schemes of things, which sound unfamiliar, thus already were contained in the laws of physics. After this now having been realized, the tables turn. Now the explanations by postulates, as they at the moment still are being taught, should be replaced by the newly derived ones, if one doesn't want to become a breaker of the law! There doesn't lead a way past the overdue reform of physics anymore.

If you want to correspond with me or if you want to get one of my books from the bibliography, then please consult my Transfer Centre in the Technology Centre of St. Georgen (Black Forest, D).

Address: 1st Transfer Centre for Scalar wave Technology
Prof. Dr. Konstantin Meyl
Leopoldstralsse 1
D-78112 -St. Georgen /Germany
Fax.: +49-7721/51870

email: meyl@k-meyl.de or: meyl@fh-furtwangen. de
Internet: http://www.k-meyl.de or www.meyl.eu

The list of deliverable books (for ordering books over the above given address, Fax, Mail or postcard suffices):

* Wirbelstrome, Diss. University Stuttgart 1984, ISBN 3-9802 542-0-8, 14

* Potentialwirbel Band 1, 1990, ISBN 3-9802 542-1-6 (German), 14

* Potentialwirbel Band 2, 1992, ISBN 3-9802 542-2-4 (not available)

* Elektromagnetische Umweltvertraglichkeit, Teil 1, 2 and 3 (German), ISBN 3-9802 542-8-3, 3-9802 542-9-1 and 3-9802 542-7-5. Each 16

* Scalar wave technology, 2003, documentation and manual to the demonstration-kit and to the experimental-kit (translated and copied).

* Sendetechnik der Gotter, historischer Sciencefictionroman, (in German)
1 .Aufl. 2004, ISBN 3-98O2 542-5-9, 12

* Neutrinopower, Johannes von Buttlar im Gesprach mit Prof. Dr. Konstantin Meyl, (Discussion in German) Argo-Verlag 2000,

(A bill is enclosed in the delivery.)

Fig. 30.14: Contacting address and list of deliverable books

Table of formula symbols

Electric field

E	V/m	Electric field strength
D	As/m^2	Electric displacement
U	V	Tension voltage
ε	As/Vm	Dielectricity: $\varepsilon = \varepsilon_r \cdot \varepsilon_o$
Q	As	Charge
e	As	Elementary charge
τ_2	s	Relaxation time constant of the potential vortices

Magnetic field

H	A/m	Magnetic field strength
B	Vs/m^2	Magnetic induction
I	A	Current
μ	Vs/Am	Permeability: $\mu = \mu_r \cdot \mu_o$
ϕ	Vs	Magnetic flux
m	kg	Mass
τ_1	s	Relaxation time constant of the eddy currents: $\tau_1 = \varepsilon/\sigma$

other symbols:

A	m^2	Area
a	m	Distance
b	m	Width
c	m/s	Speed of light
c_o	m/s	Speed of light in vacuum
C_e	As/V	Capacity of the electron
C_p	J/K	Heat capacity
d	m	Thickness
E,W	Nm	Energy
f	1/s	Frequency
F	N	Force
G	m^3/kg·s^2	Gravitational constant
g	m/s^2	Gravitational acceleration of the earth
h	m	Height
h	Nms	Planck's quantum of action
$\hbar$	Nms	Quantum of angular momentum: $\hbar = h/2\pi$
j	A/m^2	Current density
J	kg·m^2	Moment of inertia
$J \cdot \omega^2$	kg·m^2/s	Angular momentum
k	Nm/K	Boltzmann constant
l	m	Length
m	kg	Mass
M	kg	Mass of the earth
n,v =	1,2,3..	Running parameters
N	-	Constant
O	m^2	Surface area
p_m	Am^2	Magnetic moment

Q	Nm = J	Heat energy
r, R	m	Radius, radius of the earth
r_e	m	Radius of the electron
s	Nms	Spin
t	s	Time, time to orbit
T	K	Temperature
U	Nm	Potential energy
U_e	V	Tension voltage of the electron
v	m/s	Velocity
V	m^3	Volume
W	Nm	Energy
w	N/m^2	Energy density
W_e	Nm	Energy of the electron
z_e	-	Number of the involved elementary vortices
λ	m	Wave length
ω	s^{-1}	Angular frequency, angular velocity
σ	Vm/A	Specific electric conductivity
ρ	kg/m^3	Density $\rho = m/V$
ρ_{el}	As/m^3	Electric space charge density
ψ (r,t)		Complex wave function
ϕ (r)		Function of space coordinates
Φ		Golden Proportion

Definitions:

Speed of light	$c = 1/\sqrt{\varepsilon \cdot \mu}$	m/s
Speed of light in a vacuum	$c_o = 1/\sqrt{\varepsilon_o \cdot \mu_o}$	m/s
Moment of inertia (orbit)	$J = m \cdot r^2$	kg·m^2
Mom. of i. (homogeneous sphere)	$J = (2/5) m \cdot r^2$	kg·m^2
Angular velocity	$\omega = v/r = 2 \cdot \pi/t$	1/s
Surface area of a sphere	$O = 4 \cdot \pi \cdot r^2$	m^2
Volume of a sphere	$V = (4/3) \cdot \pi \cdot r^3$	m^3

Concerning vector analysis:

Bold print = field pointer (vector); further information in fig. 5.0 in part 1

Prof. Dr.-Ing. Konstantin Meyl:

Scalar wave technology

for the transmission of electric scalar waves

Abstract: 1. Auflage 2000, 2. Auflage and 1st English edition 200

This book is recommended to people, who search the entry into the world of the by the author discovered potential vortices and their propagation as a scalar wave by experimental means. It starts with the instructions to six extraordinary experiments. Doing so an electric radiation is proven, which transmits energy, and that even faster than the light. Also more energy can arrive at the receiver then is put into the transmitter. Who entertains a doubt, will be able to understand the experiments with this book in his hand, to afterwards test the experiments with the gauge, which he is familiar with.

The 1st edition in English at first only includes the instructions for the experiments. In a subsequent edition it will be complemented with a collection of test protocols and progress reports. These are organized into three groups: one group is striving to explain the behaviour of the transmission line conventionally, a legitimate concern, which in a number of points also is able to convince. A second group only is interested in those phenomena of the experiment, which can't be explained conventionally and which prove the existence of scalar waves, whereas the third research group continually strives for new spectacular experiments and practical applications.

Documentation

Belonging to the experimentation and demonstration kit
for the transmission of electric scalar waves

INDEL GmbH, Verlagsabteilung, Fax: +49-7721-51870
Tel.: +49-7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen/Schwarzwald
www.k-meyl.de mevl@k-meyl.de

Bibliography

A. Primary literature (Publications concerning the theory of objectivity):

[1] K. Meyl: Potentialwirbel, Band 1, INDEL Verlagsabt. Villingen 199O, ISBN 3-98O2542-1-6

[2] K. Meyl: Potentialwirbel, Band 2, INDEL Verlagsabt. Villingen 1992, ISBN 3-98O2542-2-4

[3] K. Meyl: Wirbelstrome, INDEL Verlagsabt. Villingen 1984, ISBN 3-98O2542-O-8 / Dreidimensionale nichtlineare Berechnung von Wirbelstromkupplungen, Diss. Uni. Stuttgart 1984

[4] K. Meyl: Einheitliche Theorie bei feldstarkeabhangiger Lichtgeschwindigkeit, Forschungsbericht der Fachhochschule Furtwangen 1992, l.Aufl. Seite 41-43

[5] K. Meyl: Wirbel des elektrischen Feldes, eine neue Storquelle? EMC Journal 1/95,6. J, S. 56

|6] K. Meyl: Gefahrdungspotential und Umwelrvertraglichkeit elektromagnetischer Felder am Beispiel der Funktelefone, Waerland 2/96 (1996), S. 21-23.

[7] K. Meyl: Wirbel des elektrischen Feldes, ein Beitrag zum Thema Uberlichtgeschwindigkeit; Wetter-Boden-Mensch 1/1997, ISSN 0940-5984, Seite: 79-89.

[8] K. Meyl: Physikalische Grundlagen einer Frequenztherapie, W.-B.-M. 6/1997, Seite: 8-15.

[9] K. Meyl: Die Wechselwirkung der Neutrinos, uber Massnahmen, die Halbwertszeit beim radioaktiven Zerfall herabzusetzen; intern.Tagung ,,Neue Energietechnologien aus USA", Zurich am 6.12.97, abgedruckt im NET-Journal (Neue Energietechniken), ISSN 1420-9292, Heft 1/2-1998, S. 14-20

[10] K. Meyl: Objektivitatstheorie, Ein Weg aus der quantenphysikalischen Sackgasse, 1. Beitrag: Der Ather und die Einsteinsche Fehlinterpretation, W.-B.-M. 3/1998, ISSN 0940-5984, Seite 39-44

[11] K. Meyl: Objektivitatstheorie, Ein Weg aus der quantenphysikalischen Sackgasse, 2. Beitrag: Die Vereinigungstheorie, Wetter-Boden-Mensch 4/1998, ISSN 0940-5984, Seite: 31-38

[12] K. Meyl: Longitudinale elektrische Wellen und Wirbel, Vortrag der ,,Internationalen Gesellschaft ftir interdisziplinare Wissenschaften" in Frankfurt am 21.3.1998, veroffentlicht im Vortragsband des "Bundesverbands gegen Elektrosmog e.V." ISBN 3-00-003124-3, Hohenstein, S.79-91

[13] K. Meyl: Biologische Wirksamkeit elektromagnetischer Wellen, Fluch (Elektrosmog) Oder Segen (Frequenztherapie)?, CO'MED, Fachmagazin fur Complementar-Medizin Nr. 10, Dez. 1998, S. 51-55

[14] K. Meyl: Elektrosmog, die physikalischen Grundlagen, Magazin 2000 plus Nr. 134,12/1998, S. 32-37

[15] K. Meyl: Freie Energie und die Wechselwirkung der Neutrinos, W.-B.-M. 3+4/1999, Seite: 21-29

[16] K. Meyl: Skalarwellen und ihre biologische Wirksamkeit, Vortragsband der Deutschen Gesellschaft fur Energetische und Informationsmedizin, Universitat Stuttgart am 17.07.1999, S. 47-52

[17] K. Meyl: Zur Brennglaswirkung des Mondes bei einer Sonnenfinsternis, NET-Journal (Neue Energietechniken), ISSN 1420-9292, Heft 7/8-1999, S. 13-17, sowie INET-Kongress-Vortrag dazu: Freie Energie im nachsten Jahrtausend, am 9.10.1999, Holiday Inn Heidelberg (Walldorf)

[18] K. Meyl: Raum-Energie-Technologie, Wechselwirkung der Neutrinos und energietechnische Nutzung der Teslastrahlung, Teil 1, Magazin 2000 plus Nr. 144, Dezember 1999, S. 62-67

[19] K. Meyl: Raum-Energie-Technologie, Teil 2, Magazin 2000 plus Nr. 145, Jan./Feb. 2000, S. 60-65

[20] K. Meyl: Teslastrahlung, W.-B.-M. 4/2000, ISSN 0940-5984, S. 31-43

[21] K. Meyl: Skalarwellen, Hagia Chora 5/2000, ISSN 0940-5984, S. 31-43

[22] K. Meyl: Teslastrahlung - die drahtlose Ubertragung von Skalarwellen, Kongressband Bregenz vom 15./16.4.2000, S. 181-196, Jupiter-Verlag, ISBN 3-906571-18-1

[23] K. Meyl: Skalarwellentechnik, INDEL Verlagsabt. Villingen-S. 2000, 1. Aufl, ISBN 3-98O2542-6-7

[24] K. Meyl: Nichtoptische Sonnenfinstermsphanomene Teil 1, Magazin 2000 plus/148,6/2000, S. 10-13

[25] K. Meyl: Nichtoptische Sonnenfmsternisphanomene Teil 2, Magazin 2000 plus/150, 8/2000, S. 26-29

[26] K. Meyl: Nichtoptische Sonnenfmsternisphanomene Teil 3, Magazin 2000 plus/151,10/2000, S. 12-17

[27] K. Meyl: Longitudinalwellen-Experiment nach Nikola Tesla, Vortragsband der Deutschen Ges. fur Energetische und Informationsmedizin, Universitat Stuttgart am 6.+.7.10.2000, S. 4.1-4.9

[28] K. Meyl: Skalarwellen und ihre Bedeutung fur die Medizin, Kolloquium vom 6.-8.10.2000 Bad Nauheim der BIT-Arztegesellschaft Vortragsband, Seite 87-96.

[29] Johannes von Buttlar im Gesprach mit Prof. Dr. Konstantin Meyl: Neutrinopower, 1. Auflage 2000, Argo-Verlag Marktoberdorf, ISBN 3-98O6584-8-1

[30] K. Meyl: Skalarwellen-Ubertragung, NET-Journal, ISSN 1420-9292, Heft 12-2000, S. 31-34

[31] K. Meyl: Longitudinalwellen-Experiment, DVR-Info im NET-Journal, Heft 12-2000, S. 21-24

[32] K. Meyl: Erdbeben und Vulkanausbrtiche, GA-Magazin, Heft 6, 2000, S. 11-14

[33] K. Meyl: Atomuhren, die falsch gehen, GA-Magazin, Heft 8, 2001, S. 10-13

[34] K. Meyl: Umpolungsproblematik und Sintflut, GA-Magazin, Heft 10,2001, S. 10-14

The suitcase for experiments from Transfer Centre St. Georgen:

Scalar wave transmission according to Tesla

Bidirectional signal and energy transmission with longitudinal waves faster than light in a resonant circuit

Get to know the Tesla radiation and its fantastic properties personally. Demonstrate, what no textbook of physics is able to explain!
The experiments, which 100 years ago still were extremely complicated, today fit into an aluminium suitcase, which can be acquired by purchase. Test yourself the historic statements and perform the following experiments:

- transmission of energy and information
- effect back from the receiver on the transmitter
- proof of free energy (1 to 3 times over-unity)
- scalar wave transmission with app. 1.5 times the speed of light
- tunnel effect or the lacking to shield the waves
- disproval of the near-field interpretation.

The kit is offered in two versions:

As a demonstration-kit for 800.- Euro (incl. VAT)
With this the 6 experiments can be carried out without further aids. (Target group are judges, doctors, architects, interested laymen, etc.)

and as an experimentation-kit for 1400.- Euro (incl. VAT) with three different Spulensatzen, Frequenzzahler and zusatzlichem material. (Target group are physicists, engineers, handicraft enthusiasts who like to experiment, etc.)

Weitere Informationen zum Set und zu den Biichern, zu deren Inhalt und zu Vortragsveranstaltungen im Internet unter: http: / /www.k-meyl.de and www.meyl.eu

INDEL GmbH, Verlagsabteilung, Fax: +49-/0- 7721-51870
Tel.: +49-/0- 7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen/Schwarzwald

[35] K. Meyl: HF-technische Hinweise zum Skalarwellen-Ubertragungs-Set, DVR-Mitgliederjournal Nr. 45, 1/2001, S. 20-22 und NET-Journal, ISSN 1420-9292, Heft 3/4-2001, S. 44-45

[36] K. Meyl: Stellungnahme zum Artikel im Spiegel, DVR-Mitgliederjouraal Nr. 46, 2/2001, S. 43-44

[37] K. Meyl: Stellungnahme zu dem Beitrag „Skalarwellenubertragung", W.-B.-M. 2/2001, Seite: 77-79

[38] K. Meyl: Scalar Waves: Theory and Experiments, Journal of Scientific Exploration, Vol. 15, No.2, June 2001, ISSN 0892-3310, pp.199-205.

[39] K. Meyl: Skalarwellenstrahlung. Welche Bedeutung haben Skalarwellen fur die Medizin? , CO" MED, Fachmagazin fur Complementar-Medizin Nr.6, Juni/2001, S. 55-60.

[40] K. Meyl: Advanced Conzepts for Wireless Energy Transfer, High efficient Power Engineering with Scalar Waves, International Congress-Publications, Weinfelden, 23/24.6. 01, Jupiter- Verl., pp. 41-49

[41] K. Meyl: Fortschrittliche Konzepte zur drahtlosen Energieubertragung, Hocheffiziente Energietechnologie mittels Skalarwellen, Kongress „Neue Wasserstofftechnologien und Raumantriebe" 23.-24. Juni 2001 CH-Weinfelden, Kongressband S. 25-40, Jupiter-Verlag, ISBN 3-906571-20-3

[42] K. Meyl: Schwingungsmedizin, Physikalisch-technische Aspekte komplementarmedizinischer Diagnose- und Therapieverfahren, G.E.I.M.-Vortrag am 6.10.2001 Universitat Kaiserslautern, GZM-Praxis und Wissenschaft, ISSN 1430-4554,7Jahrg. 1/2002, S. 50-55

[43] K. Meyl: Skalarwellen Teil 1: Grundlage einer erweiterten Feldtheorie und prakrische Konsequenzen der Maxwell-Naherung, NET-Journal, ISSN 1420-9292, Heft Juli/August 2002, S. 31-47

[44] K. Meyl: Skalarwellen Teil 2: Weltgleichung und Wellengleichung als Konsequenz der erweiterten Feldtheorie, NET-Journal, ISSN 1420-9292, Heft September/Oktober 2002, S. 44-47

[45] K. Meyl: Wirbelstrukturen als Element der Informationsubertragung in biologischen Systemen, Kolloquium vom 3.-5.10.2002 Bad Nauheim der BIT-Arztegesellschaft, Vortragsband, Seite 77-96.

[46] K. Meyl: Skalarwellen, Grundlage einer erweiterten Feldtheorie, Vortrag am VII.Kongress des IAG vom 18.-20.10.2002 in Mainz-Finthen, Kongressband ISBN 3-9804228-3-6 l.Aufl.2002, S. 118-135.

|47] K. Meyl: Nutzungsmoglichkeiten der Skalarwellentechnik, Neutrino-Power als Chance fur das neue Jahrtausend, Magazin 2000 plus I Nr.177, November/Dezember 2002, S. 60-63.

[48] K. Meyl: Schwingungsmedizin, Bioresonanz durch Skalarwellentechnik, Magazin 2000 plus I Nr. 178, Medizin Spezial 13,2002, S. 52-57

[49] K. Meyl: Die Wellengleichung, Grundlage der Skalarwellen-Physik, Magazin 2000 plus I Nr.179, Januar/Februar 2003, S. 6-12

[50] K. Meyl: Elektrische bzw. magnetische Skalarwellen im Kreuzfeuer der Kritik. Prof. Dr.-Ing. Konstantin Meyl nimmt Stellung zu kritischen Fragen, CO'MED, Nr. 1, Januar/2003, S. 52-54.

[51] K. Meyl: Freie Energie und Neutrinopower, Vortrag am 26.10.2002 in Bensheim auf dem Kongress: (Vom Band abgetippt von I.Schneider) Neue universale Energielosungen, Sammelband der Kongressbeitrage, Jupiter Verlag 2003, ISBN 3-906571-22-X, S.57-79

[52] K. Meyl (D.NeubronnerV. Die Rundfunkgotter, Teslawellen im Altertum, Matrix 3000 Band 16, 7/8-2003, ISSN 14394154, S.40-43

[53] K. Meyl: Erdwachstum durch Neutrinopower, Magazin 2000 plus Nr.188, Extra 3, 9/10-2003, S. 6-13

[54] K. Meyl: Faraday vs. Maxwell, Nikola Tesla Energy Science Conference, Washington DC 8.11.03, IRI

[55] K. Meyl: Physikalische Grundlagen fur die Informationsverarbeitung im Menschen, Kongressband zu den Festspieigesprachen 2002, (Hg. Simma Kletschka) Facultas Verlage Wien 2003, Band 27, Wiener Internationale Akademie fur Ganzheitsmedizin, ISBN 3-85076-648-9, Seite 38-59

by Prof. Dr.-Ing. Konstantin Meyl:

Sendetechnik der Gotter

Konstantin the Great is inaugurated in ancient send receive engineering by his teacher in the Roman emperor palace 304 A.D.

Abstract: 1.Auflage 2004(in German)

- Has god Apollo in Delphi broadcasted at 5.4 MHz?
- Were the Greek temples telegraphy transmitters?
- Were the temple priests amateur radio operators?
- Was Homer radio reporter by order of the gods?
- Were the oracles receiving stations?
- Have oracle interpreters deciphered the transmission code?
- Which bridges did the Pontifex Maximus build?

All are questions, which are dealt with and explained in detail in 30 lessons. In the year 304 A.D. we witness, how the later Roman emperor Konstantin the Great is inaugurated in the secret broadcasting technique of the gods by his teacher. It is an exciting time of upheavel, because the old telegraphy is almost dead. The intestines of animals to sacrifice, from the convulsions of which the radio signals are read off, are scarce goods. Instead radiotelephony should be introduced, which had been tested successfully with the Pantheon in Rome by emperor Hadrian.

But new dispute is initiated: should broadcasting be introduced or rather cellular phone? But those, who tamper around without licence, are chased and fought as always.

When the book is available in English, you will be informed per internet:

www.k-meyl.de www.meyl.eu

INDEL GmbH, Verlagsabteilung, Fax: +49-7721-51870
Tel.: +49-7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen/Schwarzwald

B. Secondary literature (Publications about the theory of objectivity):

[1] P. Schellenberg: Buchbesprechung zu K. Meyl: Potentialwirbel, Bd. 1, Efodon News 6 (1992), S. 11

[2] Dr. R. Wohlleben: Buchbespr. zu K. Meyl: Potentialwirbel, Band 2, Frequenz 50 (1996) 1-2, S. 10

[3] Steinbeis-Stiftung fur WirtschaftsfOrderung: Technologiepreis fur STZ-Leiter, Transfer Zeitung 3/94 (1994), S.6

[4] A. Waser (Herausg.): Die ratselhafte Natur - The Puzzling Nature, AW Verlag, Luzern (1996), ISBN3-9521O59-O-2

[6] K. Lorek: Buchvorstellung in "Implosion", Biotechnische Nachrichten Nr. 116 (8/1996), S. 64

[7] Dr. H. Lumpe: Potentialwirbel, neues Weltbild; Raum & ZeitNr. 86, Marz/April 1997, S. 102-103

[8] Inge Schneider: Buchbesprechung zu K. Meyl: Elektromagnetische Umweltvertraglichkeit, Teil 1, NET-Journal, ISSN 1420-9292, Jg. Nr.2, Heft 3, Marz 1997, Seite 23

[9] W. Martin: Buchbesprechung zu Teil 1, Wetter-Boden-Mensch 2/1997, ISSN 0940-5984, S. 62-63.

[10] RTL: Prof Dr.-Ing.Meyl zu Gast bei der RTL-Talkshow von Ilona Christen am 30.5.97, abgedruckt im NET-Journal, ISSN 1420-9292, Jg. Nr.2, Heft 6/7 1997, Seite 30-31

[11] Inge Schneider: zu „Prof. Dr.-Ing. Konstantin Meyls Buch Elektromagnetische Umweltvertraglichkeit", NET-Journal, ISSN 1420-9292, Jg. Nr.2, Heft 6/7 1997, Seite 32-33

[12] W. Heidrich: Geht die Expertentagung 1997 in die Geschichte ein? W.-B.-M. 6/1997, S. 3-4

[13] J. Heinzerling: Die Technik des dritten Jahrtausends, Magazin 2000 plus Nr.125, Marz 1998, S. 57/58

[14] Inge Schneider: Buchbesprechung zu K. Meyl: Elektromagnetische Umweltvertraglichkeit, Teil 2, NET-Journal, ISSN 1420-9292, Jg. Nr.4, Heft 3, Marz 1999, Seite 35

[15] W. Martin: Buchbesprechung zu Teil 2, Wetter-Boden-Mensch 2/1999, ISSN 0940-5984, S. 46+59.

[16] P.Rosch: Glanzend, Prof. Meyl zu Elektrosmog, Leserbrief, Magazin 2000 plus Nr.136, 2/1999, S. 77

[17] Portrait der Kongress-Referenten, zu Prof. Meyl: Energie aus dem Weltraum, 5. internationaler Kongress vom 26.-28.2.1999 in Dusseldorf, Magazin 2000 plus Nr. 136, 2/1999, S. 2 + 16 mit:

[18] P.Wiesner: Bericht zum 5. internationalen Kongress „Dialog mit dem Universum", zum Beitrag Prof. Meyl:„fordert eine dezentrale elektr. Energietechnik", Magazin 2000 plus Nr.136, 2/1999, S. 38-41

[19] I. Schlotterbeck: Bucher, die wir empfehlen, Magazin 2000 plus Mr. 142, Sept./Oktober 1999, S. 96

[20] H.-J. Ehlers: Freie Energie, Meyls Neutrino-These; Raum & Zeit Nr. 101, Sept./Oktober 1999, S. 94

[21] I, Schneider: Bericht zum Kongress vom 9.+10.10.99, Heidelberg, zu K. Meyl: Freie Energie im nachsten Jahrtausend, NET-Journal, ISSN 1420-9292, Jg. Nr.4, Heft 10/11, Okt./Nov. 1999, S. 8 ff.

[22] Sabine Krummer: Weltpremiere, Magazin 2000 plus Nr.145, 1+2/2000, S. 59 und im Schwarzwalder Bote vom 19.11.99: TP-Forum eroffhet mit Weltpremiere, Prof. Meyl berichtet iiber Teslawellen.

[23] Diermar Schindler: Neutrinos - Energie aus dem Kosmos, Schwarzwalder Bote Nr. 275 vom 27.11.99

[24] Schneider: Neutrino-Power - Energie aus dem Kosmos, Bericht zu Vortrag und Demo von Prof. Dr. K. Meyl am 25.11.1999 in Villingen, NET-Journal, ISSN 1420-9292, Jg. Nr.4, Heft 12/1999, Seite 4-6

[25] Schneider: Durchbruch in der Freien-Energie-Forschung; Interview mit Prof. Dr.-Ing. Konstantin Meyl, NET-Journal, ISSN 1420-9292, Jg. Nr.4, Heft 12, Dezember 1999, Seite 7-9.

[26] Dulling: Professor Meyl und die Freie Energie, TZ-Journal 1999/2000, Seite 5.

[27] Ulrich Arndt: Das Phanomen der Wirbel, Buchbesprechung zu Elektromagnetische Umweltvertraglichkeit Teil 1 und Teil 2, Esotera 1/2000, Verlag Hermann Bauer, S. 50,51

[28] Rainer Borgmann: Energie aus dem Kosmos, Magazin 2000 plus Nr. 145, Januar/Feb. 2000, S. 58-59

[29] Sabine Krummer: Wissenschaftler sprechen von Sensation, Im Villinger Technologie-Forum fuhrte Konstantin Meyl seine Skalarwellen vor, Schwarzwalder Bote vom 7.2.2000

[30] Dr.B.Tausendfreund: Uber biologische und medizinische Auswirkungen von elektrischen Potentialwirbeln und elektromagnetischen Longitudinalwellen,, CO'MED, 2/2000, S. 92

[31] Dr. H.Hantsche: Neues aus Physik und Technik, Info-Blatt der Feien Akademie, Berlin 3/2000, S.3

[32] A.Milinkovic: Weiterer Nachweis der Skalarwellen, NET-Journal, Jg. Nr.5, Heft 3, Marz 2000, S. 26

[33] Vorstellung des Referenten Prof Dr.-Ing. Konstantin Meyl, Kongressband Bregenz vom 15/16.4.2000, S. 180, Jupiter-Verlag, ISBN 3-906571-18-1

[34] Dr. Kunze: Leserbrief zu K.Meyl: Raum-Energie-Technologie, Magazin 2000 plus 147, 4-2000, S. 93

[35] Schneider: Teslastrahlung, die drahtlose Ubertragung von Skalarwellen von Prof. Dr.-Ing. Konstantin Meyl, NET-Journal, ISSN 1420-9292, Jg. Nr.5, Heft 4/5- 2000, Seite 14-15.

[36] Dr. Reichwein, Peters: Zellulare Elektromagnetische Systemsteuerung, Der Freie Arzt 5 (2000) im Wissenschafts-Forum (Anhang, S. IV - XXIII)

[37] Ulrich Arndt: Die Energie aus dem Nichts, Esotera 7/2000, Verlag Hermann Bauer, S. 62-66

[38] Besprechung zum Video „EMUV" von Prof. Dr. Meyl, Esotera 7/2000, Verlag Hermann Bauer, S. 30

[39] A.Manthey: Bericht zur Berliner Tagung fur neue Energietechnologien (8.7.2000), zu Theoretische Ansatze (u.a. K. Meyl), NET-Journal, ISSN 1420-9292, Jg. Nr.5, Heft 7/8 2000, S. 14 ff.

[40] Schneider: Uber das Experimentier-Set zur Ubertragung von Skalarwellen von Prof. Dr.-lng. Konstantin Meyl, NET-Journal, ISSN 1420-9292, Jg. Nr.5, Heft 9, September 2000, Seite 4-7.

[41] Buchbesprechung zu Prof. Dr.-Ing. Konstantin Meyl: Elektromagnetische Umweltvertraglichkeit, Teil 1 und Teil 2, Schweizerische Zeitschrift fur RGS, ISSN 1420-4894, Nr.233, Heft 4-2000, Seite 54-56

[42] Adolf und Inge Schneider: Energie aus dem All, ll.Kap.: Drahtlose Ubertragung von Skalarwellen durch Prof.Dr.-Ing.Konstantin Meyl, S.312-324, Jupiter Verlag 9/2000, l.Aufl. ISBN 3-906571-17-3.

[43] Dr.J.Lechner. Starfelddiagnostik, Medikamenten- und Materialtest Teil2, Kap.2.4.2 Beruhrungslose skalarwellengetragene Informationsubertragung, S. 173-178, Verlag Wtlhr 2000, ISBN 3-927344-44-3

[44] Dr. B. KOhler: Nachlese zum Jubilaumskolloquium der BIT-AG vom 6.-8.10.2000 in Bad Nauheim: Einen weiteren Hohepunkt stellte das Referat von Prof. K. Meyl dar,... COMED, 12/2000, S. 104

[45] Schneider: Buchbesprechung zu ,,Neutrinopower", NET-Journal, Jg. 5, Heft 12, Dezember 2000, S. 39

[46] Fachhochschule Furtwangen: Vorlesungsangebot fur das SS 2001, S.8, Prof. Dr. K. Meyl: EMUV.

[47] Vill: Die elektrophysikalische Mauerwerksentfeuchtung, EV.Verl.2001, ISBN 3-934139-04-3, S.26-33

[48] Kraft: Eroberung von physikalischem Neuland, Buchbesprechung zu ,,Neutrinopower", und

[49] Rattcher: Sind morphische Felder Neutrinos?, Zeit Geist, Heft 3,2001, S. 29 und S. 9.

[50] Dr.R.Kra6nigg: Skalare Wellen, Regulations Medizin 6, Heft 3, 2001, S. 89-96.

[51] Dr.A.Rossaint: Physikalische Grundlagen und biologisch-medizinische Schlussfolgerungen fur den beruhrungslosen Resonanztest Teil 1, Zeitschrift Hier & Jetzt, 3/2001, S.22-27 +Teil2,4/2001, S.22-25

[52] GZM-Kongress Berlin, 13.5.2001, zu dem Vortragenden Prof.Dr.K.Meyl: Skalare-/Teslawellen zur zellularen Energie- und Info. , GZM-Praxis und Wissenschaft, ISSN 1430-4554, 6,Jahrg. 1/2001, S.36

[53] Dr.P.Reichert: Die Medizin der Zukunft - Schwingungen start Pillen, Berlin Nachlese zum 6. Europaischen Kongress fur Ganzheitliche Zahn-Medizin 10.-13.5.2001, GZM-Praxis und Wissenschaft, ISSN 1430-4554, 6.Jahrg. 1/2001, S.56 zuK.Meyl: Schwingungsmedizin und Zellkommunikation.

[54] M.Dworschak: Schwarzwalder Kopernikus, Spiegel-Artikel vom 21.5.01, S. 198-199, abgedruckt im DVR-Mitgliederjournal 2/2001, Seite 41-42

[55] F.Lutz: Bis jetzt unwiderlegt, Wissenschaftler nehmen Professor Meyl in Schutz, Sudkurier Nr. 136 vom 16.6.2001 und Badische Zeitung vom 19.6.2002, Seite 22.

[56] S.Scheurer: Kopemikus untersucht Skalarwellen, Schweizerisches Bodensee Tagblatt Turgau, 21.6.01

[57] Schneider: Introduction of the author Prof.Dr.K.Meyl: Advanced Conzepts for Wireless Energy Transfer, International Congress-Publications, Weinfelden, 23./24.6.2001, Jupiter-Verlag, pp. 23-24.

[58] R.Schmid: Neues aus der Homoopathie-Forschung, zur Tagung der Samuel-Hahnemann-Stiftung vom 6.-8.4.2001 in Simmerath-E., Globuli 11/2001, S.11-12.

[59] Ralph Netzker: Jurassic Quark, Holden, Wegener, Meyl, ein theoretischer Dreisprang, im Internet unter http://home.t-online.de/home/ralph.netzker/quark.htm vom 25.10.01

[60] H.-P. Studer: Neue Energien fur die Zukunft, Vita Sana Magazin fur Dialog und Ges., 9/2001, S. 6-9

[61] Harro Wilke: Leserbrief zu Schwarzwalder Kopernikus, Magazin 2000 plus Nr. 164,10/2001, S. 92

[62] Dr. R.Krafinigg: Skalare Wellen, Arztezeitschrift fur Naturheilverfahren, 43, 1/2002, S. 19-27.

[63] Dr.C.Bollkampen: Bewusstsein im EnergetischenKontinuum, 10/2002, ISBN 3-8311-4016-2.

[64] A. und I. Schneider: Bericht vom Kongress ,,Universale Energielosungen", 26727.10.2002 in Bensheim, zu dem Beitrag von Prof.Dr.K.Meyl: Freie Energie und Neutrinopower, NET-Journal, Jg. 7, Heft 11/12 November/Dezember 2002, S. 21-28.

[65] Dr. P. Rothdach, Dr. H. Boiler u.a.m: Leserbriefe zum Beitrag: Elektromagnetische Skalarwellen im Kreuzfeuer der Kritik. COMED, 1/2003, S. 4-5 und COMED, 11/2002, S. 4: "Facit (Rothdach); Die Meylsche Theorie ist weder mathematisch, noch theoretisch-physikalisch noch experimentell widerlegt"!

[66] D. Schindler: Kopemikus aus dem Schwarzwald, Professor Meyl stellt neue Erkenntnisse der Teslastrahlumg vor, Bericht im Schwarzwalder Bote Nr. 46 VS, vom 25.02.2003

[67] H.-H. Gimbel: Neue Energielo'sungen, Kongressbericht, Wohnung + Gesundheit 3/03 Nr. 106, S. 60

[68] C. Nack: Umstritten aber begehrt. Viele Akademiker kamen Zum Experimentalvortrag von K. Meyl am 14.03.2003 im Technologiepark von Villingen-Schwenningen, Sudkurier 63/TG v. 17.03.03, S.27

[69] G. Hauser: ,.Gebaude der Physik wankt", Meyl sieht seine Theorie durch neueste Berichte bestatigt, Siidkuner Nr. 93/V, 23.04.2003

[70] Prof. Dr. J. Gruber: Grosser experimenteller Erfolg von Prof. Meyl, Energie aus dem Feld nachgewiesen! DVR-Mitglieder-Journal 1/03 Nr. 51 vom 4.4.03, ISSN 1435-0815, S. 101-102

[71] Dr. H. Kersten uber K. Meyl's Forschung: Erdwachstum durch Neutrinoabsorbtion, Technische Universitat Berlin, Internationales Kolloquium: Erdexpansion - eine Theorie auf dem Prufstand, Schloss Theuern am 24.+25.05.2003, Umdruck der TU Berlin Sekr. BH4 Lagerstattenforschung, s.a. Amberger Zeitung vom 28.5.03 und Amberger Nachrichten vom 29.5.03

[72] Dr. P. Rothdach: ,,Skalarwellen gibt es nicht und die Welt ist eine Scheibe", GeoForum 24, 6/2003, Seite 20-22

[73] Inge Schneider: zu Prof. Dr.-Ing. K. Meyls Forschung: Antiker Gotterfunk und futuristische Skalarwellenkommunikation, NET-Journal, Jg. Nr. 8, Heft 7/8 2003, Seite 31 (mit Ankundigungen zu Seminaren von Prof. Meyl in Munchen, an der Universitat Stuttgart und in Washington DC, Seite 42)

[74] J.K.Klasmann: Elektrosmog, Gefahrliche Felder? Gesundheit 9/2003, S.103-104.

[75] Gesundheitsrisiko Handy", Hauptreferent Professor Meyl..., Zollern-Alb-Kurier, 8.10.2003 und K.-H. Schlenker, Leserbrief vom 30.9.2003, s.a. Siidwest Presse: Kommission fordert Absenkung von Grenzwerten, Hohenzollerische Zeitung vom 7.10.2003.

[76] Dr. H.-J. Klose: Die Grundregulation des Menschen, Kongressband zu den Festspielgesprachen 2002, Simma-Kletschka (Herausgeber): Ganzheitliche Zahnheilkunde, Facultas Verlage Wien 2003, Band 27, Wiener Internationale Akademie fur Ganzheitsmedizin, ISBN 3-85076-648-9, Seite 95-98

[77] Dr. Valone, Integrity Research Institute: Nikola Tesla Energy Science Conference, Washington DC November 8th.+9th.2003, Conference publications, speakers biographies and abstracts.

C. Literature concerning the electromagnetic and biological compatibility

[1] R. O. Becker: Heilkraft und Gefahren der Elektrizitat (Cross Currents) 2. Aufl.Scherz Verlag (1994)

[2] P. Brodeur: Mikrowellen (The zapping of America) Augustus Verlag Augsburg (1989)

[3] EMVG: Gesetz uber die EMV von Geraten, Bundesgesetzblatt Teil I vom 9.11.1992

[4] Flanagan: Elixier der Jugendlichkeit, Waldthausen Verlag Ritterhude 1992, 1. Aufl.

[5] E. Habiger u. a.: Elektromagnetische Vertraglichkeit, 2. Aufl. Verlag Technik, Munchen 1992

[6] E. Habiger: EMV, Huthig Buch Verlag Heidelberg (1992), ISBN 3-7785-2O92-X

[7] E. Habiger: EMV-ubergreifende Aspekte zu benachbarten Problemfeldern, Elektrie 48 (1994), H. 5/6, S. 163-167

[8] N. Harthun: Naturgemasse Technik - Wege fur die Forschung nach Viktor Schauberger, Verlag Delta Design, Berlin 1996.

[9] H. L. KOnig: Unsichtbare Umwelt (Wetterfuhligk.), 5. Aufl., Verlag Moos & Partner Munchen

[10] Anton Kohling: Grundlagen der Funkentstorung in der Bundesrepublik Deutschland, etz Bd. 108 (1987), Heft 10.

[11] G. Nimtz: Mikrowellen, Einfiihrung in Theorie und Anwendung. 2. Aufl. Mannheim/Wien/Zurich: BI-Wissenschaftsverlag 199O, ISBN 3-411-O32O3-O

[12] Novitskii, Yu. I. : Effects of a magnetic field on the dry seeds of some cereals. Proceedings of Conference on the Effect of Magnetic Fields on Biological Objects, Moscow, p. 52, 1966.

[13] Plato: Samfliche Werke 5, Rowohlts Klassiker Nr. 47, S. 188,66e

[14] A. S. Presman: Electromagnetic Fields and Life. Plenum Press, New York - London, 197O

[15] V. Schauberger: Die Entdeckung der Levitationskraft, Implosion 1995 Nr. 112

[16] H. Schulz, W. Oppitz: Lokale Hyperthermie durch hochfrequente Wirbelstrome, Medizin Technik 1, 1987.

[17] G. A. Ulmer: Wellen- und Elektrosmog, Ulmer Verlag (1994) ISBN 3-924191-73-5

[18] Vitruvius Pollio, Marcus: Zehn Bucher uber Architektur, WBG 1987

[19] A. D. Watt and E. L. Maxwell: Characteristic of Atmospheric Noise from 1 to 1OO Kc/s. Symposium on the Propagation of VLF Waves, Boulder, Col., Paper 35, Jan. 1957.

D. Literature concerning physics in general (particle physics, etc.)

[1] H. Armbruster, G. Grunberger: Elektromagnetische Wellen im HF-Bereich, Huthig und Pflaum Verl.

[2] W. Bauer: Klassische Physik, Graphia Druck, Salzburg (1975), Eigenverlag.

[3] K. Bethge, U. E. Schroder: Elementarteilchen, Wissenschaftliche Buchgesellschaft Darmstadt (1986)

[4] G. Bosse, Grundlagen der Elektrotechnik II, BI 183, Hochschultaschenbuch-Verlag, Mannheim 1967

[5] Bronstein u.a.: Taschenbuch der Mathematik, 4.Neuaufl. Thun 1999

[6] A. Einstein: Grundztlge der Relativitatstheorie, S. 162, Anhang II; allg.Bemerkungen C, 5. Aufl., Vieweg + Sohn, Braunschweig 1973.

[7] A. P. French: Special Relativity, M.I.T. (1968) / Vieweg Verlag (1971) ISBN 3-528-O3546-3

[8] G. Galeczki, P. Marquardt: Requiem fur die Spezielle Relati vitat, Verlag Haag+Herchen (1997)

[9] U. Gradmann/H. Wolter: Grundlagen der Atomphysik, AVG, Frankfurt a.M. 1971.

[10] Grimsehl: Lehrbuch der Physik, 2.Bd, 17.Aufl. Teubner Verl. 1967
[11] J. D. Jackson, Classical Electrodynamics, 2nd. Ed., John Wiley, New York, 1975.
[12] Kleine Enzyklopadie Atom, Verlag Chemie GmbH, Weinheim 1970
[13] K. Kupfmuller: Einfilhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Aufl. 1988
[14] H. G. Kussner: Grundlagen einer einheitlichen Theorie der physikalischen Teilchen und Felder, Musterschmidt-Verlag, Gottingen 1976, ISBN 3-7881-6O35-7
[15] Kuchling: Physik, Gl. At4, VEB-Leipzig, 11. Aufl. 1974
[16] G. Lehner, Elektromagnetische Feldtheorie, Springer-Verlag Berlin, Heidelberg 1990.
[17] H.-J. Lugt: Vortex flow in nature und technology. Krieger publishing company 1995
[18] J. C. Maxwell: A Treatise on Electricity & Magnetism, vol.2,
Chapter XX: Electromagnetic Theory of Light. Dover Publ. New York
[19] Meinke, Gundlach: Hochfrequenztechnik, 4.Aufl. Springer-Verlag Berlin 1986
[20] Mende, Simon: Physik, Gl. 10. 39, VEB-Leipzig, 4. Aufl. 1974
[21] G. Mierdel: Elektrophysik, Huthig Verlag Heidelberg (1972) ISBN 3-7785-O243-3
[22] O. Nachtmann: Phanomene und Konzepte der Elementarteilchenphysik,
Vteweg, Braunschweig 1986, ISBN 3-528-O8926-1
[23] W. Pauli: Aufsatze und Vortrage uber Physik und Erkenntnistheorie, Vieweg + Sohn, Braunschw. 1961
[24] R.W. Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975
[25] H. Preussker: Der Wirbelring, Halstenbeck 2002
[26] Shu-wen Zhou: Abnormal Physical Phenomena Observed When the Sun, Moon, and Earth are Aligned, 21st Century Science & Technology, Fall 1999, Vol. 12, No. 3, pp. 54-61.
[27] K. Simonyi, Theoretische Elektrotechnik. 7. Aufl. VEB Verlag Berlin 1979. S. 921 - 924.
[28] K. Simonyi: Physikalische Elektronik, Kap.: 8.4 Das Magnetron
[29] K. Stierstadt: Physik der Materie, VCH, Weinheim 1989.
[30] D. F. Taylor, On the Mechanism of Aluminium Corrosion in Metallized Film Capacitors,
IEEE Transactions on EI-19, August 1984, No. 4, pp. 288-293.
[31] Jean-Pierre Vigier and the Stochastic Interpretation of Quantum Mechanics, selected and edited Stanley Jeffers, Bo Lehnert, Nils Abramson, and Leve Chebotarev. Montreal: Apeiron, 2000.
[32] A. Yializis, S. W. Cichanowski, D. G. Shaw: Electrode Corrosion in Metallized Polypropylene Capacitors, Proceedings of IEEE, Int. Symposium on Electrical Insulation, Bosten, Mass., June 1980;
[33] Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3.Aufl. Springer-Verlag Berlin 1986

E. Literature concerning Tesla and criticism of textbook physics

[I] M. Cheney: Tesla, Man out of Time, Barnes & Noble Books, New York, 1993,
ISBN O-88O29-419-1 und Omega-Verlag (1996) ISBN 3-93O243-O1-6
[2] D. H. Childress: The Fantastic Inventions of Nikola Tesla, Adventures Unlimited Press (1993)
[3] J. DeMeo: The Orgone Accumulator Handbook, ISBN O-9261855-O-7
[4] G. Galeczki, P. Marquardt: Requiem fur die Spezielle Relativitat, Haag+Herchen (1997),
[5] J. K. Kowalczynski: Critical Comments on the Discussion about Tachyonic Causal Paradoxes and the Concept of Superluminal Reference Frame, Intern. Journal to theoretical Physics, Vol. 23, No. 1, 1984
[6] J. Marinsek: Rationale Physik oder wissenschaftliche Science fiction?
dbv-Verlag fur die Technische Universitat Graz (1989) ISBN 3-7O41-O176-1
[7] H. C. Montieth: The Unified Field Theorie, MUFON Symposium Proc, Seguin, Texas, 1980
[8] F. Moser: BewuBtsein in Raum und Zeit - Grundlagen einer holistischen Weltauffassung auf
wissenschaftlicher Basis, Leykam Graz, 1989
[9] H. A. Nieper: Revolution in Technik, Medizin, Gesellschaft; MIT Verlag Oldenburg (1982)
[10] O. Oesterle: Ausweg aus der Sackgasse, Ein ganzheitliches naturwissenschaftliches Weltbild,
1. Aufl., Jupiter-Verlag, Bern 1996, ISBN 3-9O6571-16-5
[11] Rho Sigma (Dr. Rolf Schaffranke): Forschung in Fesseln, das Ratsel der Elektro-Gravitation.
Ventla-Verlag, Wiesbaden (1972)
[12] N. Tesla: III. The Singular Misconception of the Wireless (Famos Scientific Illusions), Electrical
Experimenter, Feb. 1919, p. 732.
[13] Tesla's verschollene Erfindungen, VAP, Wiesbaden (1994) ISBN 3-922367-93-3
[15] N. Tesla: Complete Patents, TeslaBook Company (1983) ISBN O-96O356-8-2, daraus:
[16] N. Tesla: Apparatus for Transmission of Electrical Energy, US-Patent-No.: 649,621, New York 1900,
Dr. Nikola Tesla: Complete Patents pp318-321.

[17] Nikola Tesla: Method of and Apparatus for Controlling Mechanism of Moving Vessels or Vehicles, US-Pat. 1898, Nr. 613,809. Complete Patents: S. 351

[18] Nikola Tesla: Famous Scientific Illusions, III. The Singular Misconception of the Wireless, Electrical Experimenter, Feb. 1919, printed in Tesla Said, p. 197

[19] Nikola Tesla: The Problem of Increasing Human Energy, The Century Monthly Magazine, June 1900, ISBN 1-882137-00-0, Page i-15

[20] N. Tesla: Die Ubertragung elektrischer Energie ohne Drahte..., Electrical World and Engineer vom 7.1.1905; Edition TeslaBd. 4, S. 131.

[21] N. Tesla, New York Times vom 15.9.1908; Edition Tesla Bd. 6, S. 241.

[22] Tesla Said (J.T.Ratzlaff), Tesla Book Company, ISBN 0-914119-00-1

[23] N. Tesla: Experiments with Alternate Currents of High Potential and High Frequency, Lindsay Publications Inc, ISBN 0-917914-39-2

[24] W. Theimer: Die Relativitatstheorie, A.Francke AG Verlag Bern (1977) ISBN 3-772O-126O-4

F. Literature concerning geophysics, geology, geography and astronomy

[1]: Kendrick Frazier: Das Sonnensystem, Time-Life Bucher, Amsterdam (1991).

[2] Carey, S. W.: Theories of the Earth and Universe. University Press: 1-413, Stanford, California.

[3] Oesterle, O.: Goldene Mitte: Unser einziger Ausweg, Universal Experten Verlag, Rapperswil (1997)

[4] Hilgenberg, O. C: Vom wachsenden Erdball, Berlin 1933, Eigenverlag. und Neues Jahrb. Geol. Palaont. Abh. 116, 1, Berlin (1962).

[5] Jordan, P.: Die Expansion der Erde. Vieweg 1-180, Braunschweig (1966).

[6] Robert Bauval und Graham Hancock: Das Geheimnis der Pyramiden, Der Schlussel zur Sphinx, List Verlag (1996) und in S3 (Die grossen Ratsel VII) am 25.5.97

[7] Mitton, S. (Herausg.): Cambridge Enzyklopadie der Astronomie (The Cambridge Encyclopaedia of Astronomy), Orbis Verlag (1989).

[8] Miller, R.: Driftende Kontinente, Time-Life Bucher, Amsterdam (1991).

[9] Hermann Wild: Technologien von gestern, Chancen fur morgen, Jupiter-Verlag Bern 1996

[10] Zecharia Sitchin: The Lost Realms, Knaur Verlag Munchen 1992, ISBN 3-426-04827-2

[11] Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, ISBN 3-89631 -302-9

G. Literature concerning free energy

[1] R. L. Clark: Tesla Scalar Wave Systems, The Earth as a Capacitor, The Fantastic Inventions of Nikola Tesla, ISBN 0-932813-19-4

[2] H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3

[3] Louis Kervran: Biological Transmutations, s.a.:

[4] J. Heinzerling: Energie aus dem Nichts, Bettendorf'sche Verlagsanstalt, 1. Aufl. 1996

[5] J. Manning: The Coming Energy Revolution - The Search For Free Energy, Avery Publishing Group, New York, USA, 1996 und Omega-Verlag, Dusseldorf 1997.

[6] G. Hilscher: Energie fur das 3. Jahrtausend, VAP-Verlag Wiesbaden 1996

[7] A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1989

[8] I. Schneider: Neue Technologien zur Freien Energie, Jupiter-Verl. 1994, S. 14

[9] Dr. H. A. Nieper: Revolution in Technik, Medizin, Gesellschaft; MIT Verlag Oldenburg (1982)

[10] Prof. Shinchi Seike: The Principles of Ultra Relativity, Space Research Institute, 1990, ISBN 915517-1

[11] M. Cheney: Nikola Tesla, Man out of Time, Barnes & Noble Books, New York, 1993, ISBN O-88O29-419-1 und Omega-Verlag (1996) ISBN 3-93O243-O1-6

[12] Prof. J. Gruber: Kalte Fusion und Raumenergie, Bericht zur 7. Internationalen Konferenz uber "kalte Fusion" vom 19.-24.4.1998 in Vanvouver, B.C., Kanada (Infinite Energy, Issue 19, 1998) und Bericht zum "International Symposium on New Energy" vom 23.-26.5.1997 in Denver, Colorado, USA (Denver-Bericht der Fern Universitat Hagen)

[13] H.-J. Ehlers: Special 7, Raum & Zeit Dokumentation, EHLERS Verlag, Sauerlach, 1. Aufl. (1994)

[14] O. Alexandersson: Lebendes Wasser; W.Ennsthaler Verlag Steyr, 1993

[15] V. Schauberger: Die Entdeckung der Levitationskraft, Implosion 1995 Nr. 112

[16] A. Waser (Herausg.): Die ratselhafte Natur - The Puzzling Nature, AW Verlag, Luzern (1996), ISBN 3-9521O59-O-2

H. Literature concerning medicine, biology and chemistry

[1] U. Luttge, M. Kluge, G. Bauer: Botanik, 2. Aufl.- Weinheim, VCH (1994) ISBN 3-527-30031-7
[2] N.Begich: Towards A New Alchemy, The Millenium Science, Earthpulse Press 1996, dt. Ubers.: Auf den Spuren einer neuen Alchemie, Omega-Verlag 1997
[3] Couto: Die kosmische Oktave, Synthesis Verlag Essen 1984
[4] Prof. Dr. H. Heine: Grundregulation und Extrazellulare Matrix - Grundlagen und Systematik, Lehrbuch der biologischen Medizin, (1991) 2.Aufl. Hippokrates Verlag Stuttgart, Seite 63
[5] U. Warnke: Bioresonanztherapie - Wunsch und Wirklichkeit. Medizin transparent Nr.2/1996, S.36-37
[6] Dr. H. Clark: The Cure For All Diseases (1995), dt. Ubers.: Heilung ist moglich (1997) Knaur Verlag
[7] Dr. Hulda Clark: The Cure for All Cancers (1993) New Century Press, USA, dt.Ubers.: Heilverfahren aller Krebsarten
[8] Dr. Bodo Kohler: Biophysikalische Informations-Therapie, Gustav Fischer Verlag (1997)
[9] Dr. Bodo Kohler: BITsyn, der neue Weg in der Informations-Medizin, Jubilaums-Schrift der Intern. Arzte-Gesellschaft fur Biophysikalische Informations-Therapie, Bad Nauheim, 6.-8.10.2000.
[10] Bioenergetische Messverfahren in Theorie und Praxis, Vortragsband der Gesellschaft fur Energetische und Informationsmedizin e.V., Universitat Stuttgart am 17.7.1999 und am 7.10.2000
[11] Dr. Johann Lechner: Storfelddiagnostik, Medikamenten- und Materialtest, Teil 2 aus der Reihe: Praxis der Ganzheitlichen Medizin und Zahnmedizin, Verlag Dr. E. Wuhr (2000)
[12] M. Luscher: Der 4-Farben-Mensch, W. Goldmann, Munchen 1977
[13] M.Luscher: Das Harmoniegesetz in uns, Econ-Verlag 1993
[14] FA. Popp: Neue Horizonte in der Medizin, 2.Aufl. Haug Verlag Heidelberg 1987
[15] M.Bischoff: Biophotonen, das Licht in unseren Zellen, 2001

J. Literature concerning history and more peripheral subjects

[1] David Ash, Peter Hewitt: Science of the gods, Gateway Books, Bath, 1990.
[2] T.E.Bearden: Skalar Technologie, Michaels Verlag 2002, ISBN 3-89539-250-2.
[3] Thomas Valone: Das Unipolarhandbuch, Michaels Verlag 2001, ISBN 3-89539-295-2.
[4] R. Sheldrake: Sieben Experimente, die die Welt verandera konnten. Goldmann Verlag 1994
[5] R. Sheldrake: Das schopferische Universum. Meyster Verlag Munchen 1983.
[6] H.Ludeling: Handbuch der Radiaesthesie, Verlag Eike Hensch 1994, ISBN 3-927407-09-7.
[7] Hensch: Radiaesthesie im landlichen Bauen und Siedeln, Arbeitskreis zur Landentwicklung in Hessen, W4, Wiesbaden 1987
[8] H. Mayer, G. Winklbaur: Biostrahlen, Verlag ORAC, Wien 1989, 5. Auflage
[9] Rohrbach: Radiasthesie, Physikalische Grundlagen und Anwendung in Geobiologie und Medizin, Haug Verlag Heidelberg 1996.
[10] R. Schneider: Einfiihrung in die Radiasthesie Teil II: Leitfaden und Lehrkurs ..., Oktogon-Verlag, Wertheim, 2. Aufl. (1984).
[11] Lamer: Worterbuch der Antike, Kroner Verl. Bd.96
[12] K. Schefold: Die Griechen und ihre Nachbam, Propylaen Kunstgeschichte Berlin Bd. 1
[13] T. Kraus: Das romische Weltreich, Propylaen Kunstgeschichte Berlin Bd. 2
[14] G. Gruben: Die Tempel der Griechen, Wissenschaftliche Buchgesellschaft Darmstadt 1986
[15] E. Horst: Konstantin der Grosse, Eine Biographie, Classen Verlag 1985 Dusseldorf, 2.Aufl.
[16] Vitruv (Marcus Vitruvius Pollio): Zehn Bucher uber Architektur, Ubers. von Dr. K. Fensterbusch, Wissenschaftl. Buchges. Darmstadt 1987,4. Aufl.
[17] A. Springer: Das Altertum, A. Kroner Verl. Leipzig 1915, 10. Aufl.

உ

Time = Space

Supreme scientific secrets in ancient India

(God’s particle, Language, Arts, Temples & Sacred Geometry)

(Shatkona Yantra – Created by Sri Bhogar Siddhar currently in a temple in Sri Lanka, taken once in a year for rituals, believed to emanate supreme subtle divine energies.)

Thinking Hearts (Raguram Gopalan)
(raguramg@hotmail.com)

உ

Thoughts from Thinking Hearts:

“Curious by nature, in a life so uncertain
Thirsty for the truth, which is so certain
I sought for some water in our Vedas - the fountain
Though I can’t understand fully what it contains
I am amazed & speechless with what I could ascertain
This is just a speck of dust from that golden mountain
Thankful to my family for helping me in what I could attain
Grateful for the wisdom and works from all the chieftains
Wish I repay every atom of it through my gratuitous tearstains”

This effort is to share what I read, understood and enjoyed about the scientific aspects in ancient India which are codified in our language, arts and in 1000s of temples. The quotes and examples are predominantly in Tamil but I have tried to give the meaning which can enable a reader to grasp the meaning and progress. Any shortcomings you notice are purely reflective of my abilities to understand and express. There was not even an iota of intention to hurt any sentiments or beliefs of others. A piece of my mind is available @ http://ragsgopalan.blogspot.com

I initially thought that I shall reproduce a prayer from one of the books which I had read but at one fine moment I decided to capture my feelings in my own words. So I scribbled few lines in English summing up my feelings as above, few lines for the Sage of Kanchi and a prayer in Tamil with its meaning in English. You may understand this prayer better after you read this book.

January 2012, Version 1.0

Please feel free to copy and reproduce in whichever form or shape as you deem fit without any acknowledgements. In fact see how much of this can you share to dispel the ignorance. The more you share the better. ☺

If I have acknowledged any books or authors please recognize them and ensure that you either buy their books or donate to their cause. They deserve all the credits.

உ

A divine soul
With a noble goal
Walked this earth
When divinity was dearth
Beyond any treasure's worth
To avoid our death & rebirth
Seek his guidance & grace
For he shall ever embrace!

உ

முதல் வணக்கம் (Prayer)

உள்ளமே மூலமாகி, பரம்பொருள் கோலமாகி
விண்ணே ஒளியென்றாகி, ஓமே உணர்வென்றாகி
உணர்வே ஓமொளியாகி, ஓமொளி ஓமொலியாகி
ஆடலன் ஆட்டமன்றோ, அசைவுறுக் காலமாகி
(Mind is the source, primordial god’s particle as the form,
The luminescent space, the Pranava as consciousness
Consciousness as the Om’s Light, Om’s Light as Om’s sound
Isn’t time a pulsation that’s the dance of Lord of Dance?)

காலமே சீலமாகி, சீலமேக் கோலமாகி
கோலமே ஞாலமாகி, ஆதியும் அந்தமாகி
ஞாலமே மூலமாகி, அறிவுடன்ஆ னந்தமாகி
நடுவிலே நன்று நிற்கும் நாதன் தாளே நாட்டமாகி
Time creates Rhythm & order, Rhythm creates shapes
Shapes become universe, the beginning and the end
Universe merges with source, it is awareness & bliss
Ever searching for the feet of the divine & graceful mean

விண்ணிலே மண்ணுமாகி, மண்ணிலே விண்ணுமாகி
விண்ணிலே எண்ணமாகி, எண்ணமே எண்ணுமாகி
என்னிலே என்னைக்காண என் கண்ணிலே மணியுமாகி
நிர்மலப் பொருளாய் ஓங்கும் நிமலனடி போற்றுவோமே!
The mass in energy (Space), the energy in mass(earth)
The energy as thoughts, our thoughts as numbers (octets)
To me find me within me, he who stands as my eye’s pupil
Beseech his feet who is devoid of may blemish

அம்பலத்தில் ஆடும் அவனை, அரங்கத்தில் துயிலும் அவனை
அரவத்தைச்சூடும் அவனை, அரவம்மேல் ஆடும் அவனை
பைரவித்துணைவன் அவனை, திருமகள் தலைவன் அவனை
நம் அறிவுக்கு இருளின் நமனை, பொற்பாதம் போற்றுவோமே!
The one who is dancing in the microbode, the one who is sleeping in arangam
The one who is wearing snakes, the one who is dancing on a snake
The one who is consort of Goddess Bhairavi, the one who is the Lord of goddess Lakshmi
Beseech his feet, the one who kills our mind’s tamasic darkness

உ

TABLE OF CONTENTS

1. Time = Space Series. Part 1 - Introduction 6
2. Time = Space Series. Part 2 - Sound merges in light 9
3. Time = Space Series. Part 3 – Vaak-Artha - Importance of word 11
4. Time = Space Series. Part 4 – Scientific languages – Tamil & Sanskrit 14
5. Time = Space Series. Part 5 – Energy & matter conundrum 16
6. Time = Space Series. Part 6 – The Unfolding Self - Universe has evolved and was not created 19
7. Time = Space Series. Part 7 – The Chidambara ragasyam & the God's particle 21
8. Time = Space Series. Part 8 – Characteristics of the God's particle (Paramaanu) 24
9. Time = Space Series. Part 9 – Light, God's particle and God 27
10. Time = Space Series. Part 10 – Lord Nataraja and The String theory 30
11. Time = Space Series. Part 11 – Vastureva Vaastu and $E=mc^2$ 34
12. Time = Space Series. Part 12 – Vedic Concept of time & its measurements 38
13. Time = Space Series. Part 13 – South American Mayan & South Indian Maayan 40
14. Time = Space Series. Part 14 –Mayan, Tamil, Pleiades & iCHING 43
15. Time = Space Series. Part 15 – Shiva Lingam is not a Phallus 45
16. Time = Space Series. Part 16 – Seelam - The Primordial circular force - Ganapthi & Mathematics 50
17. Time = Space Series. Part 17 – Seelam - The Primordial rhythm & fundamentals 53
18. Time = Space Series. Part 18 – Seelam - Significance of √2 & √3 56
19. Time = Space Series. Part 19 – Seelam –Vesica Piscis, √5, & Golden mean (Φ) 58
20. Time = Space Series. Part 20 – Sankhya yoga and Sage Kapila 63
21. Time = Space Series. Part 21 – 5 and 8 the Universal rhythm 65
22. Time = Space Series. Part 22 – Eight (Octet) and the 5 sacred arts 67
23. Time = Space Series. Part 23 – Nanotechnology in ancient india & size of god's particle 69
24. Time = Space Series. Part 24 – Octet, Octave and the Gandharva veda (Classical Music) 72
25. Time = Space Series. Part 25 – Adi Tala, Nava Tala and Sthapatya veda (Sculpture) 74
26. Time = Space Series. Part 26 – Sabda Veda, Chandas & Poetry (Ma Nishaada…) 78
27. Time = Space Series. Part 27 – Architecture – The Zenith of mathematics (Vaastu) 82
28. Time = Space Series. Part 28 – Capturing the divine energy – Squares & Circles 87
29. Time = Space Series. Part 29 – Heaven & Earth – Lord Nataraja & Lord Ranganatha 91
30. Time = Space Series. Part 30 – Temples and Idol worship 95
31. Time = Space Series. Part 31 – Mayan & his works 101
32. Time = Space Series. Part 32 – Conclusion 105

உ

1.TIME = SPACE SERIES. PART 1 - INTRODUCTION

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

"Abheda Darsanam Gnanam*"*. Nothing defines what **wisdom is** better than these 3 words. **Experience of unity / oneness in all is wisdom.** The roots of Indian wisdom are captured in these 3 words that everything in this universe is the manifestation from a single source, and we can call it God or Consciousness. This naturally leads us to one of the mahavakya ***"Pragyanam Brahma"*** which means **"Consciousness is god".** But for people who are stingier with words and wants these 2 concepts to be put in just one word, then it is called **"Advaita".**

It is not very often that nature reveals itself its absolute oneness and the un-manifested interconnections. You consider yourself fortunate when you come across such knowledge in however small measure, even if it happens very rarely and even if you have an illusion that you have understood something. It has been very overwhelming for me to come across some of the details I have been reading and hearing and the dots I could connect – some of which I had decided to write as a book.

If you are interested in knowing how scientific ancient Indian Sanatana Dharma is and how well we have codified nature's secret in our languages, arts and temples and if you are interested in any of the following questions then you may like this book. Most of the aspects stated here as scientific proof has a reference to one or more of our scriptures and I would refer and quote them at various places.

- Would you be surprised to know that the size, shape, color and characteristics of God's particle are very clearly documented in our scriptures at least 12000 years back? What the scientists at CERN are spending billions of $, is well documented and we will discuss the characteristics of God's particle in this book.

- Would you want to know how **Time = Space** and more over time is nothing but the vibration of space? If I tell you that this was known us Indians at least 12000 years back and currently available in a written form would you believe?

- When you throw a stone in water / pond the ripples that form take the shape of a circle (in 2D) / sphere (in 3D). Can you guess what the shape of the ripple would be if you throw the same stone in space?

- What is the similarity between Poetry, classical music, Classical dance like Bharatanatyam, Sculpture and a building? Apart from the fact that we don't clearly understand any of them in depth what is the similarity? Would you be surprised if I say that all are essentially the same and in fact a **well constructed building is nothing more than a frozen music** and the relationship between all the above confirms to a single grammar documented 12000 years + back ?

- Would you be interested in knowing how $E = mc^2$ an equation stated by our great Einstein has been known to us for years and not just that but we know the root of this energy and how to engineer it with precise formulas? Would you be surprised that this is the basis our language?

- What if I say that the age of the universe, every astronomical unit of the bodies, big bang theory, string theory, God's particle etc are all documented and they are all codified in our temples?

- You would have heard about the golden ration Φ but do you know the philosophical significance of Φ and how it is derived? Moreover would be interested in knowing the significance of key numbers and ratio like 0, 1, 5, 8, 9 and $\sqrt{2}, \sqrt{3}, \sqrt{5}$? What is I say that $\sqrt{3}$ can be correlated with "**Iyengar Namam"?** Does it sound funny? What if there is a deep mathematical and philosophical significance? And how are all these ratios related?

- As a wise man said that **'Mathematics is the language of the gods".** What if this formula is codified in our texts from which we understand the God's particle, its shape, characteristics and its manifestation?

- What if I tell you that sound is a subset of light and sound merges with light at a point and that is at the root consciousness? Any non-takers for this?

- If you agree on the above then **Sabda Brahman** or the belief that sound is god merges with something higher, isn't this contradictory? Also why many religions did believed that "**Word is god".** Was it a mere belief or was there any scientific reason behind it?

- What is the correlation between Carnatic music and a human being's height?

- What if I say that the Mayan from South America and the Mayan from South India are one and the same?

These are some of the questions which might get addressed in this book. I am going to quote a series of references and books and links to understand and learn more about these topics and I would request you to utilize them if you are interested in these topics.

This may not only refer you to the experts but also can help you overcome any weakness I have in understanding and articulating some of these topics. Some of the topics may touch upon religious beliefs (mostly Hinduism) but if you read with an open mind you would notice the underlying science behind it.

உ

I would be happy if any of these blogs motivate and inspire you to explore the roots of our tradition and knowledge and give you an insatiable hunger for knowledge, wisdom to pursue the nature and its secrets.

May the ancient Sanskrit quote "May I meditate on the supreme truth "Satyam Param deemahi", resonate within you as it does within me.

Happy reading!

PS:

Abheda Darshanam Gnanam, Dhyanam Nirvishayam Manam;
Snaanam mano mala thyagam, Showchamindriya Nigraham", says the Upanishad.

The real Gnana is seeing all as one, real Dhyana is controlling the Manas from all temptations, the real snana or Bathing is Cleansing of All Mindly impurities, not physical impurities alone and real Showcha or cleanliness is controlling the Indriyas, this controlling of Indriyas, emotions, weaknesses and temptations.

I am also quoting parts of Kahlil Gibran here,

Your hearts know in silence the secrets of the days and the nights.
But your ears thirst for the sound of your heart's knowledge.
You would know in words that which you have always known in thought.
You would touch with your fingers the naked body of your dreams.
And the treasure of your infinite depths would be revealed to your eyes.
But let there be no scales to weigh your unknown treasure;
And seek not the depths of your knowledge with staff or sounding line.
For self is a sea boundless and measureless.
Say not, "I have found the truth," but rather, "I have found a truth."

2. TIME = SPACE SERIES. PART 2 - SOUND MERGES IN LIGHT

What if I say that “**Sound is a subset of light”.** Scientifically you may agree that both are electro-magnetic vibrations at different frequencies in the spectrum though one may not be the subset of another.

So, What if I say that every other frequency is the subset of light? I am sure many may disagree. Some of the religious pundits may claim that “**Sound is god**” which I am not disputing but “**light too is god”** and as we progress towards the primordial state sound merges into light.

Let me try to explain this with a simple exercise. Please sit in a place with a calm mind – few deep breaths may help.

- Imagine that I am TALKING to you and avoid thinking that you are reading this blog.
- I am going to TELL the following 5 words and you are going to LISTEN to it. Please just observe what happens within you.
    - CAT, BOAT, MOM, DEVIL, BEST FRIEND

The following are the fundamentals of perception as per Hindu scriptures:

- Every word has a meaning (Let us ignore the so called meaningless words for the time being.).
- Every meaning invokes a form or shape within our mind.
    - For example when I say “**MOM”** you associate this word with your mother and the picture / image of your mother is perceived inside. You need to be little more observant to understand this. **The fundamental principle is that**
    - **“Every word is first associated its meaning then the meaning is associated with a form. The form is associated with a Guna (Quality) and then the quality or Guna is associated with its experience. What is stored inside is just the experience.**

Let me TRY to explain the abstract concept of consciousness as below:

- **Consciousness / existence exist in 3 states (deeper explanation is 5 / 7 states but we will discuss the simpler alternative).**
    - **Formless state called Experience / feeling / bhava. This is the super conscious state.**
    - **Gross state with form which has physical attributes. This is the conscious state.**
    - **In-between state (Not formless and no form) called Guna state which has subtle quality attributes. This is the Subconscious state. If this is difficult to comprehend for you please imagine “Egg” which is in-between the formless and state with form.** Gotcha?

Let’s look at an example. Rose is a flower. When someone tells you **“ROSE”** then the following happens:

- The sound “rose’ associates with a flower usually. (I am ignoring it being a name of your girl friend ☺). So here the word is associated with its meaning.
- This flower is associated with red color and perhaps with thorns. Here the sound does not exist anymore but only a form. **Yo!, the sound has merged into light.**
- The third step is - this form is associated with a quality which is fragrance of a rose. For someone sensitive they can smell the rose when they think about it. How else could we salivate at the thought / smell of a favorite dish if this step does not exist?
- Now the last step is, this quality invokes an experience and hence a feeling. This rose may be associated with your inner self for love and hence the emotion of love surfaces. This is called **Bhava / Rasa** in Sanskrit.

If you are not clear about this then please do not move to the next part and think or discuss about it. A little contemplation would help you understand this **5 stage** process of perception.
‘SOUND – MEANING – FORM – QUALITY (GUNA) – EXPERIENCE (FEELING / BHAVA)’

Our ancient forefathers knew this science and interconnection. Let me now quote these Slokas as an example from Bharatnatyam, which talks about this linkage.

Kanteenaalambayet geetam, Hasteenaardham Pradarsayet,
Chakshurbhyaam darsayeet bhaavam, Paadaabhyam taalamaacharet.

According to Nandikeshara the dancer should sing with the mouth, express the meaning of the song with hand gestures, her eyes should express the emotions or bhava, the tala, the beat or rhythm should be done with her feet. Further he goes on to say –

'Yato hastas tato drushti, Yato drushitis tato manaha
Yato manahas tato bhavo, Yato bhavas tato Rasaha'

'Where the hand goes, there should follow the eyes, Where the eyes move, mind should follow it,
Where the mind follow, Bhava or emotion is created, & lastly when Bhava is created the sentiment or Rasa will arise.'

All the Mudras or gestures represent word when held at a particular position around the frame of the body and when used in a particular way.

Alas we Indians have ignored this great science & art and its linkages and are like ignorant fools following the meaningless western tradition and dances!

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

3.TIME = SPACE SERIES. PART 3 – VAAK-ARTHA - IMPORTANCE OF WORD

‘SOUND – MEANING – FORM – QUALITY (GUNA) – EXPERIENCE (FEELING / BHAVA)’

In the **5 stage** process of perception, one important factor which I missed is the **“Self / EGO”** which interprets the **Guna** to an experience. For example when I say Dog, the **‘Self’** interprets Dog to a **loyal friend or a bike chaser or a rabbies generator.** Essentially what is the quality of dog mean to me? If you understand this process this explains the uniqueness between individuals or the root cause of differences in a society.

I am also going to add but not going to focus on the fact that the inputs are never limited to sound and when see a picture the process of perception starts with step 3. Just to poke the individuals who think “**Sound / Word is the only god – Sabda Brahman**, please tell me - when you see a picture and when there is no sound associated with it, where is **God as Sabda Brahman in this situation**?

Does that mean that the very important Hindu belief that the Sound **“Om is sabda Brahman or god”** does not hold any meaning? It certainly does and I shall share my thoughts on it – continue reading.

What is the importance of Word?

The greatest of Indian poet **‘Kalidas’** said

"Vaak artha viva samprakthov vaakartha prathipathaye
Jagadha: pitharov vandhe parvathe parameswarow"

he compared Lord Shiva and Shakti like sound (Shakti) and meaning (Shiva). Both sound and meaning are inseparable. You would now understand the wisdom behind this sloka since there is no sound without a meaning. Also **Sabda Brahman** merges with the **Artha Brahman** and both are inseparable. Here Artha Brahman is the light and the meaning of the word. Remember we say that **‘Throw some light”** when we don’t understand something.

In Tamil **பொருள் (porul)** means “a thing” and it also means “meaning”. We will see how **பரம்பொருள் (Paramporul)** is the Lord Shiva as a God’s particle and how he is referred to as the meaning too. The same applies to Sanskrit where **“artha”** means **wealth / things – remember Arthasastra?** And it also refers to **“meaning”** as in the above sloka. How rich these languages are?

And in **Abhirami Andadi, bhattar** sees devi as **"Arthanareswara"** and confirms us that the grace of devi will brings us both richness and mukthi. **"Sollum porulum ena nadamadum"...**

உ

A word is made up of letters called Akshara in Sanskrit. The meaning of Akshara is that which cannot be broken down further. ie. Word can be broken down to letters but it cannot be broken down further. A letter has a form like **"A"** and has a sound associated with it. This means the light and sound are merged in a letter. As per the principle stated above in the process of perception, the primordial state is one where light and sound is merged into each other.

If you don't dispute the philosophy that everything merges into god at the primordial state and at this state, sound and light are inseparable, and then you would agree that an Akshara is **God / Brahman.** Each letter has significance and meaning in Indian languages especially Tamil and Sanskrit. **A word is a combination of Form (Light), Sound and meaning and these are first 3 aspects of perception which takes us from a gross state to a formless subtle un-manifest state.**

Hence the statement from New Testament **"In the beginning was the Word, and the Word was with God, and the Word was God"** reflects the same principle.

But as per Hindu scriptures Veda is the breath of god (not word of god) or the aural vibration of consciousness cannot take you to the state of moksha. Words don't take you to God, it can help you in the process but has its limit, since the sound and meaning merges into something else as we move towards more subtle space.

This is what my friend Kambar said in Kambaramayanam as **"வேதமும் முடிவு காணா மெய்ப்பொருள் வெளிவந்தெய்தி மா துயர் துடைத்த வார்த்தை மறுப்பரோ!!.** This means that the supreme truth who could not be found by Vedas came and solved the problem of an elephant, can someone dispute this?

Every word or Akshara has four states **Para, Pashyanti, Madhyama and Vaikari. Vaikari** is the fourth state where the sound is audible to us. If I recite something mentally then the sound has 3 states only and these 4 states signify as to how a feeling (bhava) becomes thought and thought becomes form and then a sound. Just note here that as per Frank outlaw, it's the thoughts which are primordial but as per our scriptures it is feeling or Bhava or deep rooted desire or also known as vasanas which drive your thoughts. So Frank's quotes are half baked plagiarism.

In a work called Tarka Samgraha, we have classified 9 different physical attributes (substances / dhravyas) and 24 different qualities / GUNA in all. If you are interested reading more about this, please read these blogs or the original work.

http://ragsgopalan.blogspot.com/2010/09/vedic-perspective-matter-and-its_14.html This is a 3 part series.
http://ragsgopalan.blogspot.com/2008/11/part-1-perception-art-of-seeing-things.html This is a 2 part series.

உ

To summarize this topic:

- Broadly consciousness exists in 3 states.
- The process of perception if understood well would throw a lot of secrets about nature.
- The 5 stage process of perception **“Sound – meaning - Form – Guna – Bhava”** is a simplistic way of portraying this.
- Sound merges into light and light is worshipped as God in most of the religions and belief. We can explore little more of this in the future sections.
- A word is a unique combination where Light, Sound and meaning merges and hence it is considered as god. This word represents **Arthenareeswara or yinyang or Star of David**.

In the future sections let us explore as to what is the relationship between Maths and God and some of the critical numbers and its philosophical and scientific significance.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading

உ

4.TIME = SPACE SERIES. PART 4 – SCIENTIFIC LANGUAGES – TAMIL & SANSKRIT

One of the main reasons I believe as to why our wisdom has degraded so much in the last century is our ignorance and reluctance to learn our rich language. Our language offers richness not just in literary sense but is supremely scientific too. I am going to touch upon just 3 to 4 aspects of it which I knew and understand at a fundamental level and request you to learn your mother tongue well.

Many of my friends don't exhibit a sense of remorse when they say **'I can understand Tamil but can't read or write it**". I wanted to tell them how much in life do they miss in terms of knowledge and richness. This is not just about Tamil but every major Indian language is rich in its own way. Since my mother tongue is Tamil, I am going to articulate some of the science in it.

Vigyan / Vigyanam (விண் ஞானம்) means Science:

But most importantly for this series I wanted to take the word called Science as meant in Tamil and Sanskrit. A word always can be broken down to its dhatus for its exact meaning in Tamil and Sanskrit. Some examples below:

- Science in Tamil is **Vigyanam and in Sanskrit is Vigyan**. This is split as **Vin** and **Gnanam** / **Gyan**. **Vin means space and Gyan means knowledge. Hence according to Indian system science means it is the nature / knowledge of the space.** This has very deep significance since western science has not agreed the existence of space for a long time and their gross and materialistic approach prevented them from seeing the most subtle component and its nature. We were masters of it ages back. Most of the questions in Chapter 1 would be answered based on the knowledge of space – Vigyan.

- **Universe is called as Brahmanda** in Sanskrit. Brahma + Anda which literally means a big egg. This means we knew that the galaxy is elliptical in shape ages back and it required hubble telescope in 1950 to tell us that.

- **Bhoogol** means **Earth (Study of) / Geography – This Splits as Bhoo +Gol.** Bhoo means Earth and Gol means round. We did not throw stones at a scientist as late as 17th century when he said that earth is round and not flat. We would have laughed our heart out and ignored him.

உ

Be it the age of the universe, all of the astronomical measurements, health, spirituality etc, many of us need a white skin with blue eyes to tell us that what we knew was right and they have certified it. When we decide to come out of this mentality then the process of reclaiming our rich heritage shall start.

Origin of Universe and linkage with Tamil:

The initial evolutionary process for the universe consisted of 5 stages states Aintiram. You can imagine that these are the movements of particles during the Big bang. They are அமிழ்தல், இமிழ்தல், குமிழ்தல், உமிழ்தல், தமிழ்தல்.

- Withdrawl (Amizhdal)
- Imizhdal (Overflowing)
- Kumizhdal (clustering round in an order)
- Umizhdal (Emitting)
- Tamizhdal (resulting into a well defined form)

These are the first stages from the Big bang till a well defined shape is formed. Now the last level is called Tamizhdal which results in a well defined form. This state has emanated from the Pranava and not been created by some one. Hence a language as rich with the sweetness of inherent order, originating from the luminosity of the original particle is called Tamil. Tamil is word based on the final resultant state called Tamizhdal (தமிழ்தல்).

I shall address the other 2 interesting aspects that include a Chicken and Egg analogy for Energy and matter conundrum in the next blog.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

5. TIME = SPACE SERIES. PART 5 – ENERGY & MATTER CONUNDRUM

Energy and Matter concept in formation of letters:

The concept of Energy and matter is not new to us and does not start with Einstein's days of $E=mc^2$. Let us look at how our letters are formed from Aksharas then people would know that the concept of Energy and Matter is old as our language. In Tamil there are two kinds of Aksharas **உயிர் எழுத்து** - **Uyir Ezhuthu (Life / Energy letters – 12 letters) and மெய் எழுத்து - Mei Ezhuthu (Body / Matter letters – 18 letters).** Both put together it is 30 letter pack for Tamil as alphabets. Now both have to combine to give letters called **"body with life"** which are called **uyir mei Ezhuthu**. This combines to form 18*12 which is 216 letters as **Uyir Mei Ezhuthu**. For example க்+ அ = க in Tamil and क़ + अ = क in Sanskrit.

So the formation of letters itself is a science where Matter (Body) combines with Energy (Life) and then the living organisms (Body – life letters) are formed.

I want to ask you a **Chicken & Egg** question here which is very relevant to today's scientists. Did the energy come first or is it the matter that came first? Modern science is still scratching its head. I can answer this question in 2 ways, first is obviously using language as a science.

- In the above example க்+ அ = க, க் is the body and அ is life. So as per our language sciences be it Tamil or Sanskrit the matter is formed first and then the life came and attached to it. This is very well documented in Tamil scriptures as follows:

உடம்புடன் உயிர் வந்து ஒன்றுவது இயல்பே - தொல்காப்பியம் . This is over 2500 years old which states that life comes and joins a body in language and this is nature.

மெய்யொளி உயிரொலி சிவணுதல் இயல்பே – ஐந்திறம். This is over 12000+ years old which states that same thing and adds that body is like light and life is like sound and they merge and this is siva's nature. **This certainly means that both Body and life are not formed at the same – as popularly believed. Life comes and joins the body to give a living being – in the above case the letters are the living beings.**

The second reference I am going to take for the same concept is our **Yogic / Siddha / Ayurvedic** science as to how fetus is formed in mother's womb. As per our (Siddha) belief the conception of a child in a mother's womb happens as follows:

- **The first organ that is formed for the fetus is the cells of the Eyes (Single Eye)**. This eye cell starts vibrating due to its heat and multiplies as 2 eyes and then as the whole body. This is the vibration seen in the initial days of the fetus and not the heart beat. As per modern medical science the heart takes shape only by the end of the 8th week and not before. I am requesting someone to validate this.
- **Life comes to this body at the 3rd month of conception only**. This is the belief as per Ayurvedic and Siddha science.
- Instances of the third month natural abortions are very high than the later months based on the above reasons. This means that the life is not destined for that matter / life force is not sufficient enough.
- **How many of you know that the size of eyes never change as you grow and stays the same from the time we were born,** while all other organs undergo growth?
- Also the eyes hold a very important significance in the spiritual science which I shall address later. The eyes are the seat of Sun god and are characterized by heat and light.

It is through the heat in the eyes the subtle became a gross living being in this world and if the living being wants to merge back with the subtle – which is the process of Samadhi then the living beings should use their eyes, increase the heat in the eyes and start the devolution process. This is a supreme secret kept by the sages even in today's world. **You may want to ask any spiritual school as to why they recommend meditation with eyes fully or half open instead of closed eyes? Ask yourselves why the free masons and other secret societies use eye as one of their symbols?** We shall address this later.

Now you want to answer the Chicken and Egg analogy to the Energy and Matter conundrum based on the above 2 thoughts, we can conclude it is the matter which came first. We shall try to explain this based on Space sciences a little later.

Masculine and Feminine names based on ending letters:

I am also going to give you as to how Indians distinguished masculine from feminine names. **Kanchi Maha periyava, the Sage of Kachi** had explained this beautifully in his **"Arul Vakku"**. I am giving a gist of it here just to conclude this:

- **Masculine names:** All the names shall end with the phonetic a, im, in , ir, aha etc. Examples are Ram, Shyam, Raman, Kannan, Kishan, Vivek, Adam. So phonetically names like Catherine is a masculine name because it ends with "in" phonetically.

- **Feminine names:** All names ending with phonetics aa, e, I, oo etc. Examples are Radha, Ramya, Savithri, Gayathri, Uma, Rosy, Isabella etc. Same way phonetically Joshua is a feminine name.

Vishnu and Lalitha sahasranamam follows this rule and hence one invokes Shakthi as lalitha and the other invokes Purusha as Vishnu.

Just to conclude this part, I think we looked at how science is an integral part of our language formation and for us science has always been about the knowledge and nature of the space and we believed that space manifested itself as the physical world. So in the next few parts we shall look at how this manifestation has happened and other interesting aspects.

I would conclude this part with a earnest request that please encourage your kids to learn their mother tongue – to read and write and you would not have done anything better than this to advance their spiritual growth and adding richness to their life.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

6.TIME = SPACE SERIES. PART 6 – THE UNFOLDING SELF - UNIVERSE HAS EVOLVED AND WAS NOT CREATED

As per many scriptures in Hinduism this universe has evolved from itself and not created by someone. If you want to quickly ask as me as to what Brahma, Vishnu and Rudra mean – I would attribute them to different process heads in this evolution. They are all different aspects of the same consciousness based on their state of being in the process of evolution / devolution. This is my understanding but let us avoid this discussion here.

Here I am indebted and thankful to the nobel soul **Late. Dr. Ganapathi Stapathi,** (http://www.vastuved.com/life.htm) who not only unraveled the Vaastu science – the science of Space and time but also brought to light 2 important books which were over 12000+ years old written in Tamil by **Mahamuni Maayan.**

Mamuni Maayan is believed to be the same person who is popularly known in **South America as Mayan** and the 2012 phenomenon based on Mayan calendar. Dr. Ganapathi has established proofs that the pyramids in Middle East, the temple architecture in India and the structures in Mexico and Guatemala confirm to a single Science and Mathematics seen in all these works.

At the end of the blog I am listing few books which helped me understand this topic but most of the books (**Aintiram and Pranava Veda**) are out of print now and if you are lucky, you would get some of them. Now let us understand briefly how this evolution process has come about. According to Mayans' Aintiram:

- The free space is the unified field of energy and matter and source of all forms that we see in material world. That is why we were defined science as **Vingyanam**. This space consists of very minute particle called **Vinporul, nunporul, Sittrambalam, Paramanu, Oliporul, Microbode** (**விண்பொருள், நுண்பொருள், சிற்றம்பலம், பரமாணு, ஒளிப்பொருள்**) and in many names. You can pick what you like. I am going to use **Paramaanu** in the rest of the blog.

- The whole process of evolution is a 5 stage process. (Keep noting the importance of 5 and at the background remember Shiva is represented by Panchakshara mantra (Na~Ma~Si~Va~ya).

So what is the five stage process?

  - This space is called as Moolam (Source).
  - This Moolam starts vibrating and this vibration of the space is called Kalam (Absolute Time and not physical time. We will refer this as time only).
  - This Kalam (time) vibrates to a mathematical resonance called Seelam (Rhythm).
  - This Seelam as it resonates creates Kolam (Shapes – Circles, cubes etc).
  - This Kolam becomes our visible Universe which is called Gnalam.

உ

- So it is **MOOLAM (Source) – KALAM (TIME) – SEELAM (RHYTHM) – KOLAM (SHAPE) – GNALAM (UNIVERSE). மூலம் - காலம் - சீலம் - கோலம் – ஞாலம்**, these are 5 stages of evolution.

This means that the Space itself becomes spatial forms and time is the instrument that creates, sustains and destroys. Lord Siva in Tamil is also called as **Kalan (காலன்)** as to who controls time since Kalam **காலம்** is time.

To put it differently, un-manifest state when pulsates becomes absolute time and this is the creative energy. The process of pulsation confirms to a mathematical order and we will see this order in detail later. You may now understand the phrase that "**Time creates, sustains and destroys**". Since if there is no pulsation then nothing happens.

For those who need quote from a foreigner to convince themselves of this can be referred to the **Bertrand Russel's** quote "**What we perceive as qualities of matter are differences in periodicity".**

I shall quote some from the oldest literatures for reference below:
காலமே எவைக்கும் மூலம், காலமே மூல ஆற்றல்
காலமே கோலத் தோற்றம், காலமே ஞாலத் தோற்றம்… *Maha muni Mayan from Aintiram*

காலக்கூறே கணக்கியலாகி… Time splits as Maths in proportion it pulsates.
Vyasa says in Adi parva chapter 1, shloka 249, says the same thing that Time Creates, sustains and destroys everything.

kalah srijati bhutani kalah sanharati prajah
sanharantah prajah kalam kalah shamyate punah!!

The same is quoted as **"Vastur eva Vaastu"** which means the un-manifest Paramaanu which is called **Vastu** (with Single a) becomes **Vaastu (gross form – with two aa**). You can relate this **Vaastu** as the building science which is grossly misused and misrepresented and commercialized in today's TV programs. We will address the science in it later.

I shall leave this part with this thought, the space is considered as **Nataraja (Shiva) – subtle component and Ranganatha is considered as the Earth component**. So Shiva and Vishnu are two aspects of the same seed. We will see more scientific proof of this later.

"Shivasya hridayam vishnur, Vishnoscha hridayam shivah:" and
"Shivaya Vishnu rupaya Vishnave Shiva rupine"

Vishnu is the heart of Shiva and likewise Shiva is the heart of Vishnu and they both are the representation of each other.

Now if you are clear about the five stage process of evolution we can delve into each and every aspect of this in little more detail that would answer everything.

Ref: Check out all the ones here. Key recommendations are item 6, 11, 14 and 43. http://www.vastuved.com/publication.html

7. TIME = SPACE SERIES. PART 7 – THE CHIDAMBARA RAGASYAM & THE GOD’S PARTICLE

There were 2 questions I had confronted myself when I was reading these passages. They are:

- Why did the unmanifest space energy started pulsating and became absolute time?
- What did the unmanifest space energy contain? Was it a point of singularity as the modern scientists wanted us to believe? If so then how much mass was concentrated in that point which is obviously very difficult to believe.

The answer to the first question is here **“The self unfolds itself for its self – Bhagavad Gita”.** There is no convincing answer to the rational mind as to what the reason that the consciousness which was passive decided to become dynamic and started pulsating. So let’s leave this for the time being.

But the second question is answered very comprehensively. But before we get into the details of what it is, I want you to check Figure 1. This gives the overall picture of evolution and the five stage process. I also want you to recollect the “EGG” state which is called as the GUNA state where in the form and sound manifests. They say that this pulsation started with OM. This OM as popularly believed is just not an aural form but also has a visual form and that is light. We will discuss about this later in detail.

But let us focus on the actual second question as to what the contents of this Space energy were:

- The space is filled with **Paramanu or Microbodes** which are tiny particles. These are the **God’s (Higgs Boson) particle** our CERN scientists are after. http://en.wikipedia.org/wiki/Higgs_boson

- Now, we are probably a year away before CERN publish the results when this blog is published. I would not be surprised to see the results of that experiment is closely aligned with what is available in our scriptures and produced in this blog.

- The primal manifest form of the unmanifest is a **square** (primal wave pattern). This is the shape of the **Paramanu or Microbode**. This is very luminescent, means filled with light and on a 3D plane this shape is a **Square Cuboid**.

- A **Square Cuboid** is a shape which is equal in all sides – say of unit measure which is the smallest measure possible and has **6 faces, 8 vertices, and 12 edges.** You can start figuring out in Hinduism, why the idols of Lord Shiva and Lord Muruga has six faces, 12 eyes, hands etc.

- You may want to note that 6 faces can be represented in 2 ways as below. Figure 3 is Lord Muruga’s star or a Star of David in 3D which has 6 faces and 12 edges, popularly known as **Tetrahedron**. **So don’t be surprised if the Boson Higgs particle is of the same shape.**

உ

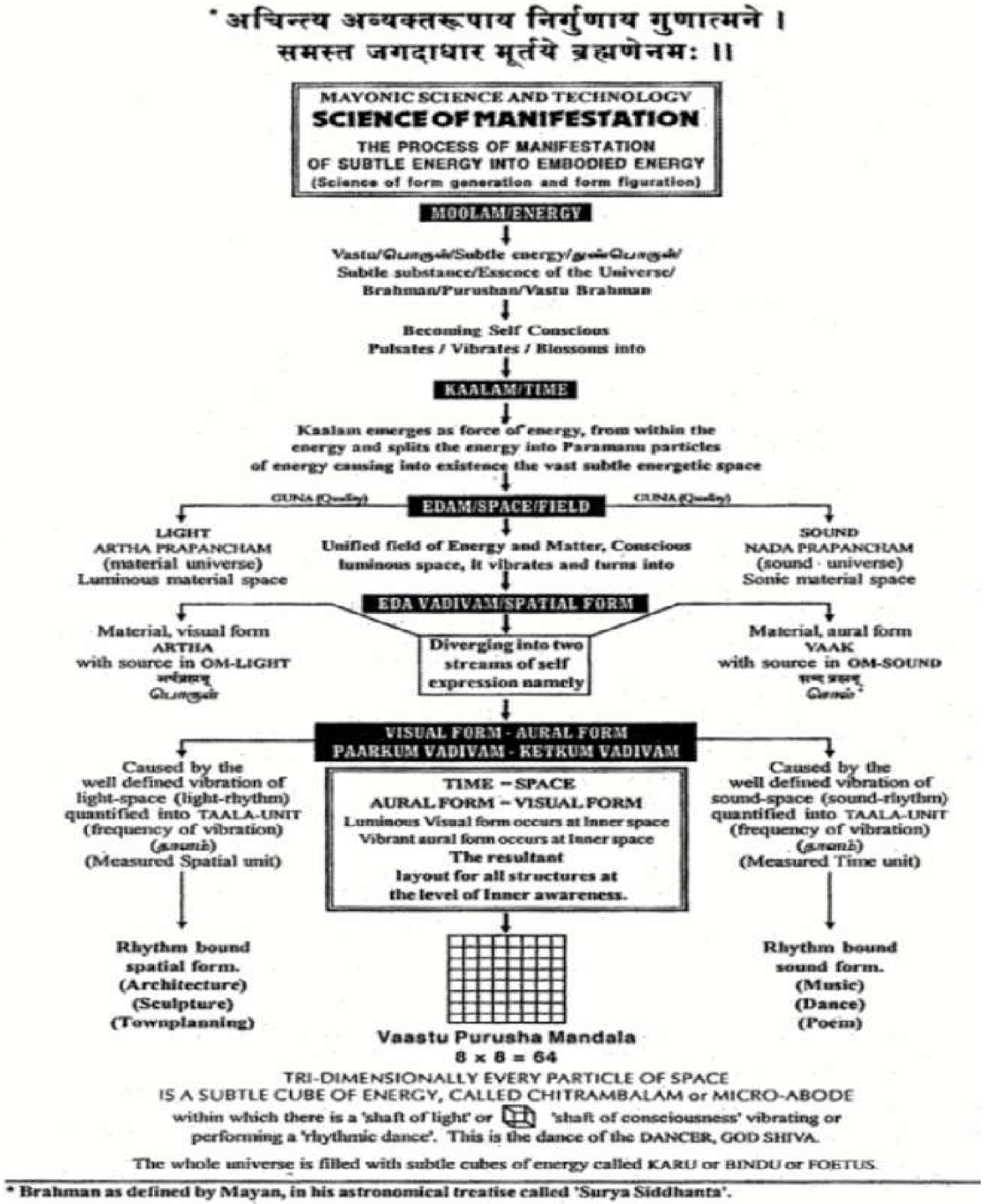
* अचिन्त्य अव्यक्तरूपाय निर्गुणाय गुणात्मने ।
समस्त जगदाधार मूर्तये ब्रह्मणेनमः ॥
MAYONIC SCIENCE AND TECHNOLOGY
SCIENCE OF MANIFESTATION
THE PROCESS OF MANIFESTATION OF SUBTLE ENERGY INTO EMBODIED ENERGY
(Science of form generation and form figuration)
MOOLAM/ENERGY
Vastu/பொருள்/Subtle energy/நுண்பொருள்/Subtle substance/Essence of the Universe/Brahman/Purushan/Vastu Brahman
Becoming Self Conscious
Pulsates / Vibrates / Blossoms into
KAALAM/TIME
Kaalam emerges as force of energy, from within the energy and splits the energy into Paramanu particles of energy causing into existence the vast subtle energetic space
GUNA (Quality)
EDAM/SPACE/FIELD
GUNA (Quality)
LIGHT
ARTHA PRAPANCHAM
(material universe)
Luminous material space
Unified field of Energy and Matter, Conscious luminous space, it vibrates and turns into
SOUND
NADA PRAPANCHAM
(sound · universe)
Sonic material space
EDA VADIVAM/SPATIAL FORM
Material, visual form
ARTHA
with source in OM-LIGHT
अर्थब्रह्म
பொருள்
Diverging into two streams of self expression namely
Material, aural form
VAAK
with source in OM-SOUND
शब्द ब्रह्म
சொல்
VISUAL FORM - AURAL FORM
PAARKUM VADIVAM - KETKUM VADIVAM
Caused by the well defined vibration of light-space (light-rhythm) quantified into TAALA-UNIT (frequency of vibration) (தாளம்) (Measured Spatial unit)
TIME – SPACE
AURAL FORM – VISUAL FORM
Luminous Visual form occurs at Inner space
Vibrant aural form occurs at Inner space
The resultant layout for all structures at the level of Inner awareness.
Caused by the well defined vibration of sound-space (sound-rhythm) quantified into TAALA-UNIT (frequency of vibration) (தாளம்) (Measured Time unit)
Rhythm bound spatial form.
(Architecture)
(Sculpture)
(Townplanning)
Vaastu Purusha Mandala
8 x 8 = 64
Rhythm bound sound form.
(Music)
(Dance)
(Poem)
TRI-DIMENSIONALLY EVERY PARTICLE OF SPACE IS A SUBTLE CUBE OF ENERGY, CALLED CHITRAMBALAM or MICRO-ABODE within which there is a 'shaft of light' or 'shaft of consciousness' vibrating or performing a 'rhythmic dance'. This is the dance of the DANCER, GOD SHIVA.
The whole universe is filled with subtle cubes of energy called KARU or BINDU or FOETUS.
* Brahman as defined by Mayan, in his astronomical treatise called 'Surya Siddhanta'.

Figure 1: Science of manifestation
(Courtesy: Pranava Veda by Mamuni Mayan – 12000+ years back, recreated by Dr. Ganapathi Sthapati)

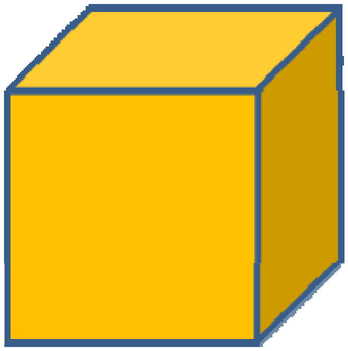

Figure 2: A cube with all equal sides

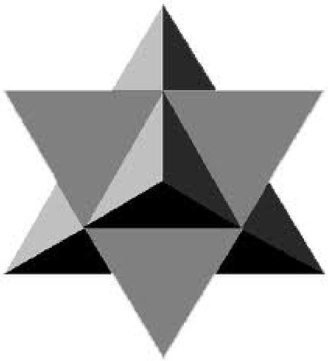

Figure 3: A tetrahedron (from Internet)

- The free space is packed with cubical atoms of energy. They are building blocks of the structure of the universe. **So what is clear is that all these atoms existed in a “resonant state of non-vibration” before the so called “big bang” happened or the pulsation started.**

- **It did not burst out of a single point as popularly believed. Alternatively, in space every point is the center. When the pulsation started these cubical atoms underwent a shape change and manifested as shapes and forms.**

- This cube is called as **micro-abode called Sittravai / Chittrambalam** (mini hall in Tamil). This is **secret in Chaidambaram, Sage Appar** (**Chidambara ragasyam**) realized. Chidambaram is a Shiva temple associated with Space energy and when you go to sanctorum you would not find any idol but just empty space with some lights there. This is to signify this concept of **Paramanu** which is **luminescent and is a Cuboid in shape**. This is the smallest particle possible and can be called as **‘God’s particle”.**

- The micro space, in the cubical shape is fetus or Garbha. This is known as **Vinkaru** in Tamil. This micro abode is the repository of light and sound.

- The free space is Light and the Light is Moolam or source of the universe and universal forms. Light is Brahman, Atman, Vastu Brahman – Nunporul in Tamil.

In the subsequent blogs we will correlate String theory, Brahma Sutra, Nataraja, Vishnu and how they are all contained in order and proportion in pulsation.

Happy reading!

Ref: A good read **Fabric of the universe, Jessie J Mercay**.

உ

8.Time = Space Series. Part 8 – Characteristics of the God's particle (Paramaanu)

This part is going to be a bit controversial hence read it with an open mind. Per our scriptures when analyzed deeply we can state that **"At the beginning it was just Consciousness (Pragnanam) that was luminescent with unmanifested potential energy and residing at resonant state of non-vibration. This energy was un-manifest and was present as Paramaanu or Microbode or Vinporul aka a subatomic particle. Yes! It was a particle without vibration and was at resonant state of non-vibration.**

The entire Space was supposed to be filled with these particles only and these were primal potential matter which is ever lasting.

Now let us ask few questions and see how our scriptures answer this:

1. **What was the shape of this particle?**

We have already answered that. In 2D it is square and in 3D it is a cuboid with all equal sides. Please note that in space it is always 3D and not 2D. It has 6 faces, 8 vertices, and 12 edges. This is known is current scientific community as **Tetrahedron**.

2. **What is inside the Microbode or Subatomic particle?**

This is a very interesting question. Our scriptures say that it is purely space inside but with an energy grid.

3. **What is this energy grid and what is its shape?**

It says that it is an 8x8 energy grid inside this microbode. This essentially means it is a square energy grid of 64 squares inside this Paramaanu.

4. **What is this color of this Paramaanu?**

It is luminescent and hence the golden color of the light is what it is associated with. **"Hiranyavarnnaam harineem**.." says Sri Suktha in the vedas, confirming the color of this Paramanu. Their variations exist but the fundamental one is of golden color.

5. **What is this size of this God's particle?**

It is calculated to be ~133 nm. We will see this in detail later.

6. **What is the significance of the energy grid?**

For the time being let me state that anything subtle is a function of 8 and anything gross is a function of 9. Let me also add that **iChing –** which is a Chinese divination book and the hexagram which deals with the Science of the subtle is in an 8x8 square matrix. **The mental game Chess is played in an 8x8 square.**

You would find 8 is the number or dimension for the subtle. We will see the importance of 8 and 9 and how beautifully these proportions adorn everything in nature in our daily life.

7. Are these particles faster than light?

Frankly I have not read a direct answer to this question but since this particle itself is light the velocity of this particle cannot be faster than light. This is my opinion. Modern experiments are suggesting that it could be faster than light. In **Mahabharatha, Yakshaprasna, the Yaksha will question this to King Yudhishtra as to "what is fastest thing in this world"**? King Yudhishtra would reply as **"mind"**.

Some of the sages have interpreted that since Consciousness is omnipotent and ever connected - the communication is seamless, instantaneous and not subjected to the concept of distance. The example given is if there is a pain in your leg, it does not take time to travel to your head. It is instantaneous and hence the whole universe should be considered as a living organism.

So to conclude on this question the particles are not faster than light but each particle is filled with intelligence and consciousness (just like a human cell) and their communication to anything is not subject to concept of distance. If our scientists come back and tell us that these particles are faster than light, then wait for someone else to tell us a little later that it isn't faster than light. ☺

8. Ok. Who has seen these particles and give me empirical proof?

परमानुरिति प्रोक्तं योगिनाम दृष्टि गोचरम ! **(Paramaanuriti proktam Yoginaam, drishti gocharam)** says **Mayamatham.** This means the **microspace** or **Paramaanu** is visible to the trained contemplatives and meditators only. One quick example as proof is the **Temple in Chidambaram** which represents **Space tattva** where the temple is built around the concept with enough documentation on it.

But we have been trained under Lord Macauly's education system and how can we accept this as proof which is subjective and superstitious. If I may ask please tell me how many of us have seen the electrons or protons? Do we not believe everything when someone tells us in English?

It is true that because of its existence we can see its manifestations and results of some experiments in our day today world. So we will look at tons of manifestations which we have been in nature as we progress. I would like us to be philosophically very clear as to what these things mean which we have been blindly following. For the time being please note down your questions and I am positive that this would be answered in due course.

Our sages have not only seen this, but have given the dimensions of this and the mathematical formula to manage this energy. We will see all of them. After reading all of them if you still don't want to believe in our scriptures, then may be CERN's result if it's in the same lines would prove it for you.

Happy reading!

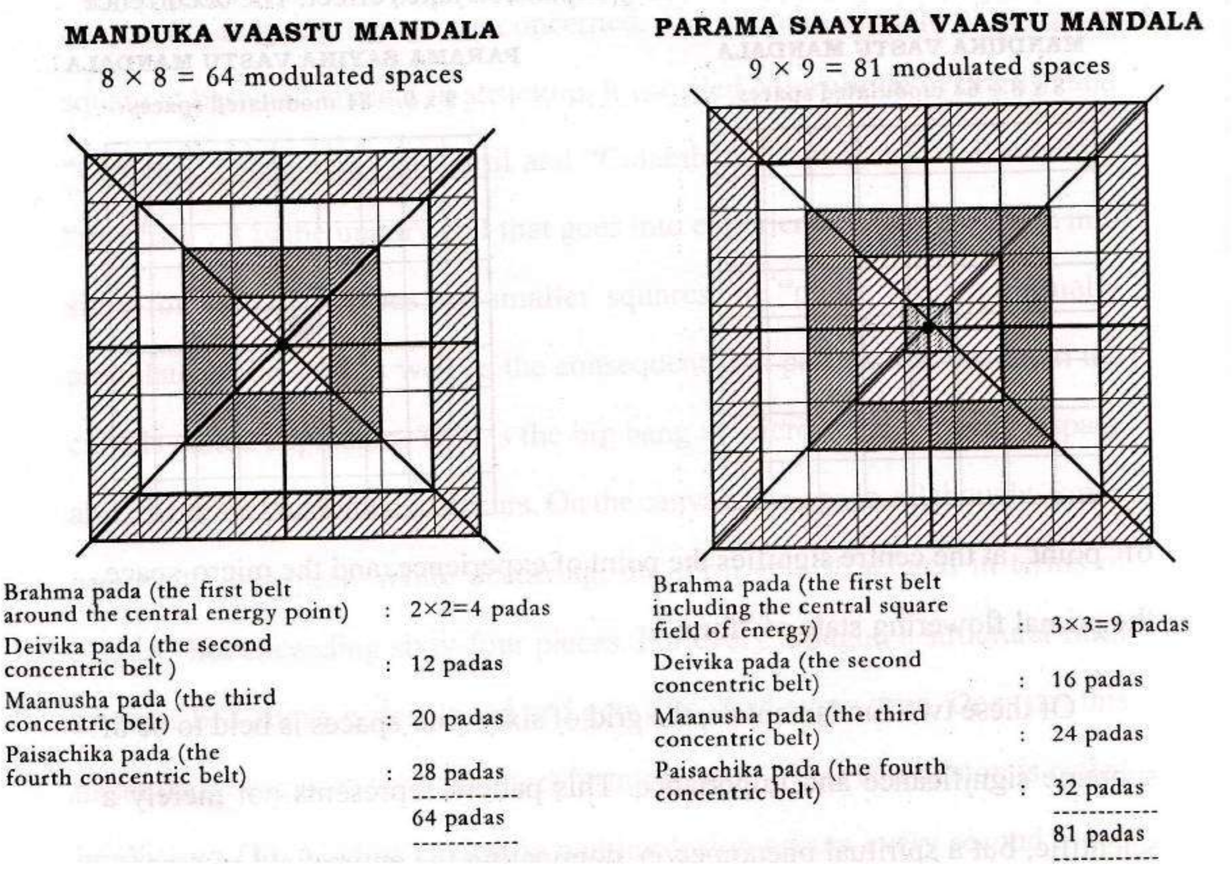

MANDUKA VAASTU MANDALA
8 × 8 = 64 modulated spaces

Brahma pada (the first belt around the central energy point)	: 2×2=4 padas
Deivika pada (the second concentric belt)	: 12 padas
Maanusha pada (the third concentric belt)	: 20 padas
Paisachika pada (the fourth concentric belt)	: 28 padas
	64 padas

PARAMA SAAYIKA VAASTU MANDALA
9 × 9 = 81 modulated spaces

Brahma pada (the first belt including the central square field of energy)	: 3×3=9 padas
Deivika pada (the second concentric belt)	: 16 padas
Maanusha pada (the third concentric belt)	: 24 padas
Paisachika pada (the fourth concentric belt)	: 32 padas
	81 padas

Figure 4: MANDUKA MANDALA – 8x8 (subtle form) & PARAMA SAYIKA MANDALA – 9x9 Gross forms
Courtesy: Dr. Ganapathi Sthapati

உ

9. TIME = SPACE SERIES. PART 9 – LIGHT, GOD’S PARTICLE AND GOD

“At the beginning it was just Consciousness (Pragnanam) that was luminescent with un-manifested potential energy and residing at resonant state of non-vibration. This energy was self manifest and was ever present as Paramaanu or Microbode or Vinporul aka a subatomic particle. This is **omnipotent, omniscient and omnipresent**.

Ignore religion for the time being and just look at this particle’s characteristics scientifically. It is Eternal (**Anaadi**), primordial (**Aadi**), without any quality (**Nirguna**), it cannot be reached / grasped by thoughts (**Achintya** - since thoughts as energy are much more gross than this and we would see this), it cannot be measured (**Aprameya**), Luminescent (**Jyothi swaroopa**), its **omnipotence, omnipresence (Visvam) and Omniscience** is already known. So said the mahavakya **Pragnanam Brahma!**

All the characteristics of the particle are attributed to a personality or a state by the founders of religions. **Hindus called this state as Paramasiva or Purusha. Hence everything subtle is called Sivamayam.** The essential pulsation which started and the universal evolution began was called as **Parashakti or Prakruthi or Maya.** We will talk about the Shakti aspect later.

This subtle un-manifest energy evolves to become universe which is gross. This gross energy also known as **earth** is called as **Vishnu** and the subtle energy is called as **Space / Akash / Heaven**. So the term became “**Sarvam Vishnu mayam Jagat”.** Please note that it is just **Sivamayam** and there is not Jagat there. Since it is all subtle and there is no Jagat which is gross. But it is **Vishnu Mayam Jagat**.

I need to add that every sect in Hinduism has named this God’s particle with a different name. Narayana has replaced paramasiva in Vaishnavism and so on. In Hindu temples there are 2 prime idols – **Moolavar and Archakar** - The Moolavar does not undergo any archana or any decoration and it is always for the Archakar for which all poojas are performed. This god’s particle refers to the Moolavar and not Archakar.

To my knowledge, every religion and every sage / saint has agreed with the fact this fundamental particle or energy unit is **LUMINISCENT**. Note that there is no sound or word in this …**yet.** Let us see few quotes on this (there are thousands of quotes):

- **Be it vedic culture, Zoroastrians, Greek or any major culture fire / sun god which represents light is worshipped.**
- **“Om Jyothi rasaha” says Vedas**
- **In Gayatri mantra you meditate on the supreme light to illuminate you.**
- Vallalar said **“Arutperumjyothi, Arutperumjyothi”** which means **mighty graceful light**.

- If therefore your eye be single, your whole body shall be full of light. Mathew 6:22
- Thirumoolar states brilliantly that who realizes that the light (physical light which means the world) merges with light (subtle light), he has realized SIVA.

  ஒளியில் ஒளிபோய் ஒடுங்கிய வாறும்
  தெளியும் அவரே சிவ சித்தர் தாமே - 124

So what has happened to the **Sabda Brahman** here? It has merged with the **Artha Brahman** which is the light. What can you use to represent this state?

Light represents Form and our eyes represent Light / Form. So obviously sound is represented by ears. When sound has merged with light then it means Eyes and ears are merged together.

Can you think of anything that represents this state? This concept is captured by representing a **Manduka (Frog) or Sarpa (Snake)** always. Since for a frog and a snake the ears and eyes are merged. **Manduka Upanishad,** represents a knowledge of this state. And obviously the 8x8 square / cube represent the most subtle energy as an energy grid is called as **Manduka Pada / Mandala..**

Now the same aspect when it is applied to a gross form / body or a building then it is a 9x9 energy grid and it is called as **Paramasyika Mandala**. **At the gross level the same concept as applied is the WORD where Light (form) and sound and meaning merged into one. You cannot separate them**. We will look at this later.

Unfortunately we have lost the science and we neither understand the philosophy nor the symbolism behind this supreme science but we made fun of the fact that the supreme science is named after an absolutely menial creature like frog which is used only for zoological lab studies across the world or ends up in the dinner plate as it does in far eastern countries.

I would like to conclude this section by stating (may be a little controversial) that **First it was Consiousness as the Vinkaru, Microbode, Sittrambalam, Nunporul, subatomic particle which was luminescent. Later the concept of God evolved out of it.**

So God is not light, but the right of saying it is "the primordial aspect of "luminescence" has been adopted as God…

So, of the 5 fold evolution process we are still at the source (Moolam). We will see the concept of time and the mathematical science in the subsequentparts.

MOOLAM (Source) – KALAM (TIME) – SEELAM (RHYTHM) – KOLAM (SHAPE) – GNALAM (UNIVERSE). (மூலம் - காலம் - சீலம் - கோலம் – ஞாலம்)

Happy reading!

உ

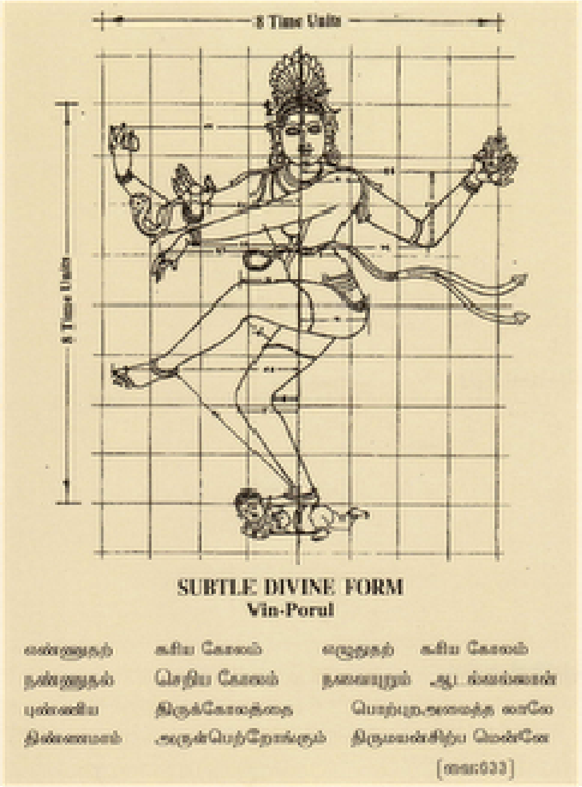
8 Time Units
8 Time Units
SUBTLE DIVINE FORM
Vin-Porul

Figure 5: Lord Nataraja in the Manduka Mandala (8x8 grid) (Courtesy: Dr. Ganapathi Sthapati)

Our next topic is Nataraja and String theory!, so check out this picture. I want you to note the light / fire in the left hand and sound / damru in the right hand. If you notice little closely, you would find a central line along the picture's sternum and all the parts have minute measurements from the central line.

உ

10. TIME = SPACE SERIES. PART 10 – LORD NATARAJA AND THE STRING THEORY

“The self unfolds itself for its self – Bhagavad Gita”.

This self unfolded itself with a pulsation and this pulsation is the **start of Absolute time** (**KAALAM**). As I said nothing to my knowledge tell us clearly as to why this pulsation happened. This is the point of singularity in the scientific language and the event is termed as **Big Bang.** Just as I expressed earlier it was an event of primordial pulsation and not of an explosion which threw away mass because of the explosion into the space.

The entire space existed with its subatomic particles and this pulsation made the space to vibrate and this vibration was rhythmic and led to the evolution of this universe. **Popularly known as “That one became many”… and ..Vastur eva Vaastu.**

This pulsation happened because of the **Pranava** or the first pulsation resulted in the **Pranava -** both conceptually right. So at the time of this primordial pulsation 2 things happened:

- The absolute time started. This is called **Kaalam.**
- Primordial and the most sacred **Pranava** emerged.

Both these happened simultaneously and it is futile to ask whether **Kalam created Pranava or Pranava created the pulsation.** In many of the texts it offered as **Pranava that created the pulsation**. So let’s go by that.

- This is the reason that **Pranava / Primordial Om is so sacred in Hinduism.**
- This Pranava had manifested in 2 forms:
    - **Luminescent form as OM the Light. This is the source of all visual forms. The primordial form is represented as ஓம் in Tamil and ॐ in Sanskrit.**
    - **Aural form as Om the sound. This is the source of all aural forms. I should say that this is more popular form compared to the luminescent form.**

Now this concept is very important, since in the Pranava meditation they say us to meditate on the silence after the word OM and also mediate **OM as light, OM jyoti rasaha in vedas.**

So, when I say that sound merges into light, it is all the aural form merges into **ITS** luminescent form. All luminescent forms merge with the **Paramaanu** which is also luminescent. I am reproducing this passage again for emphasis and clarity.

- **Sage Thirumoolar** states brilliantly that who realizes that the light (physical light which means the world) merges with light (subtle light), he has realized SIVA.

  ஒளியில் ஒளிபோய் ஒடுங்கிய வாறும்
  தெளியும் அவரே சிவ சித்தர் தாமே - 124

I have read many times over to understand this concept and I would suggest that you allow this concept to sink in.

Now let's come to the 8x8 energy grid of the **Paramaanu** and imagine that it is pulsating because of Pranava and this **Pranava has both light and aural form**.

- They say that this **pulsation is rhythmic and symmetric**.
- When something is Symmetric, then there has to be a central line along with the Symmetry comes. For example if our right and left sides are symmetric then the body is divided along the sternum & backbone.
- Within the 8x8 energy grid / hall there is a **vertical luminous shaft** called **Brahma Sutra (Oli nool)**. Sutra in Sanskrit and Nool in Tamil means **Thread / String**.
- This **Olinool / Brahma Sutra is a shaft of consciousness also called Moolathoon / Moolasthambham**.
- **This shaft of consciousness vibrates in a particular order called rhythm.** This is the order of nature.
- **This rhythmic vibration of the shaft of consciousness is the dance of Shiva, the CELESTIAL DANCE / THE COSMIC DANCE / THE DANCE OF EVOLUTION of the LORD NATARAJA.**

Look again at the picture of LORD NATARAJA, he is drawn inside a 8x8 Manduka mandala (now you know what it is), his left leg is raised suggesting a movement from left to right. Imagine that he is going to land the left leg to his right side.

Why should he suggest that he is moving from left to right?

His left hand holds light / fire and his right hand holds Drum / Damru / Sound. The process of Evolution which starts with the absolute time and Pranava is fantastically represented pictorially. It suggests that the evolution process starts with the rhythmic dance of the PARAMAANU in a 8x8 energy grid and this process evolves first as light and then as sound. This light and sound creates further forms and the entire universe.

It is very easy to answer as to why sound evolves from light? We all know that the light is at a higher frequency and velocity than sound. This is the same reason we see lightning first then hear the thunders.

So let's summarize the DANCE OF EVOLUTION of LORD NATARAJA:

- From being un-manifest the **Pranava** emerges. Or the Pranava evolves the un-manifest to the manifest and starts the cycle of evolution.
- This leads to a pulsation which is called as the **COSMIC DANCE / CELESTIAL DANCE / DANCE OF EVOLUTION**.
- This pulsation is **rhythmic and symmetric**.
- The luminescent shaft of consciousness is called as **Brahma Sutra** or popularly called as String theory in the modern science.
- The process of creation evolves from the dance as light then sound evolves from light and then all the visual and aural forms in this universe.

If this is so then the process of dissolution should be that Lord Nataraja's right leg should be lifted suggesting a movement from right to left isn't?

Look at the majestic picture suggesting the **DANCE of EVOLUTION** and the **DANCE of DISSOLUTION** as below:

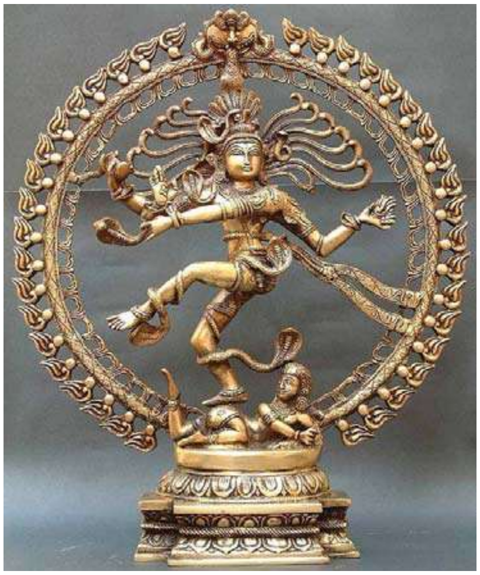

Figure 6: LORD NATARAJA / DANCE OF EVOLUTION

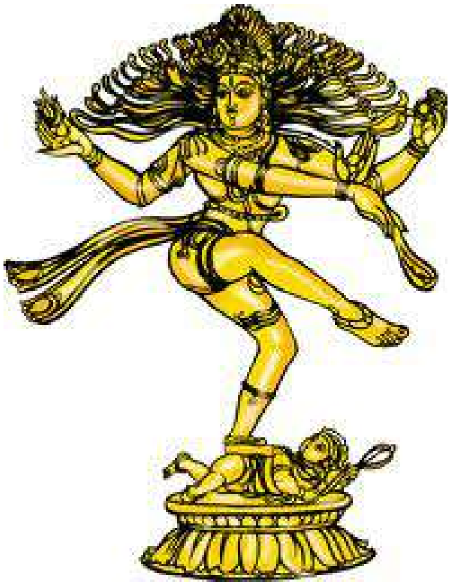

Figure 7: DANCE OF DISSOLUTION

In my opinion there is nothing more scientific in this world that is pictorially represented about the evolution of universe. You should read a book called **"The Tao of Physics" by Frijat Capra** where as physicist he explains how he experienced this and highlights the dance of the quantum particles in space corresponds to the dance of **Lord Nataraja**.

Now this central shaft of Consciousness is called **Brahma Sutra** and according to our scriptures every Paramaanu has this string, which means every being, matter, inanimate things in this universe including the tiniest of subatomic particles have this string / sutra. In the future parts we will see the Maths behind this sutra and the pulsation but let us look at what the modern science says about **Superstring theory** here, http://en.wikipedia.org/wiki/String_theory. You can read it how many times ever you want but this theoretical postulation is a direct adaptation of the cosmic dance and our Brahma Sutra and nothing more.

Now check out Lord Nataraja at the CERN http://www.fritjofcapra.net/shiva.html

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Now you may understand as to why I started this series with the above line. We have a very rich past that is most scientific and it is left codified in a form across thousands of temples in India. Let us just open your eyes and understand the scientific meaning around it.

If you are so far surprised at our richness and advancement, please wait till we will get into real hard core mathematics and geometry of our nature in the subsequent parts.

May that “**Sutradhaari”** dispel our ignorance and reveal to us the secrets of nature.

Happy reading!

11. TIME = SPACE SERIES. PART 11 – VASTUREVA VAASTU AND E=MC^2

We have seen that the universe in its un-manifest state is filled with luminescent cuboids and there was a trigger in the form of Pranava which started the absolute time and gave forth visual forms and aural forms.

This process of evolution is depicted very nicely in the following diagram:

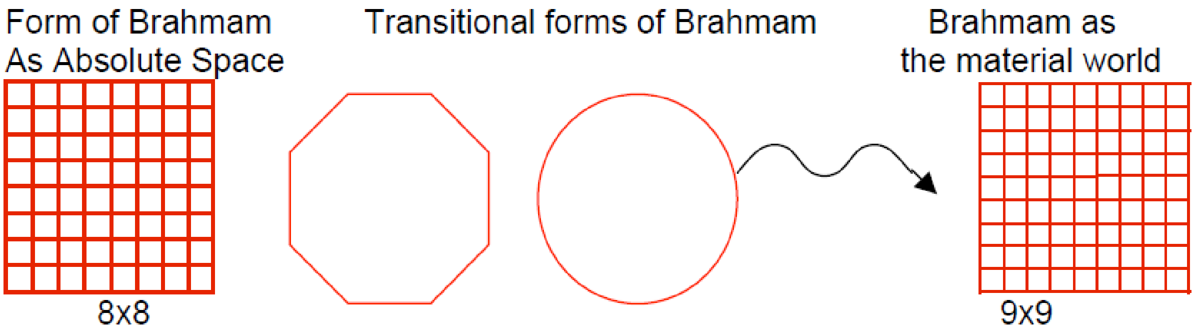

Figure 8: Transition forms of Brahman (Coutesy: Dr. Ganapathi Stapathi)

The square (in 2D) which is an 8x8 energy grid became a 9x9 energy grid which is manifest / gross. The transition forms were an octagon and a circle. **Aintiram** suggests how this transition takes place. Current physics accepts that a pulse and a sine wave can be created from each other and this happens because of simple addition of harmonic frequencies.

The primordial pulsation initiates a rotational force and this rotational force as it gains momentum converts the square into a circle and if you notice that the octagon is just an in between state between a square and a circle.

We see confirmation of the details of this manifestation process in Chapter 9, verse 8 of the Bhagavad Gita where Lord Krsna describes His manifestation process to Arjuna:
“prakritim svaam avastabhya visrijaami punah punah”, meaning “Curving back onto myself, I create again and again…”

What it suggests is that the essential shape of every gross form in its natural state is a circle (in 2D) and a sphere in (3D). **This is the reason we find every planetary body is almost spherical in shape and this is also the reason as to why the ripples in a water body when we throw a stone in the water are in circles.**

Be it the center of a tornado or the shape of a hurricane or a tsunami it always a circular force. The supreme secret in martial arts is that the Chi moves in a circular fashion and hence Tai Chi is the most deadly form of martial arts. Lord Vishnu has Chakra as his weapon to suggest not only light but also the rotational force.

The process of manifestation is captured in these 2 pictures at all the significant stages. If you notice that a single dot becomes multiple dots, then a line, then a pulse and then goes on to becoming a 8x8 energy grid and then to a 9x9 energy grid. 8x8 is un-manifest and 9x9 is manifest. These aspects are very well captured as sutras in the book **Aintiram and Pranava Vedam.**

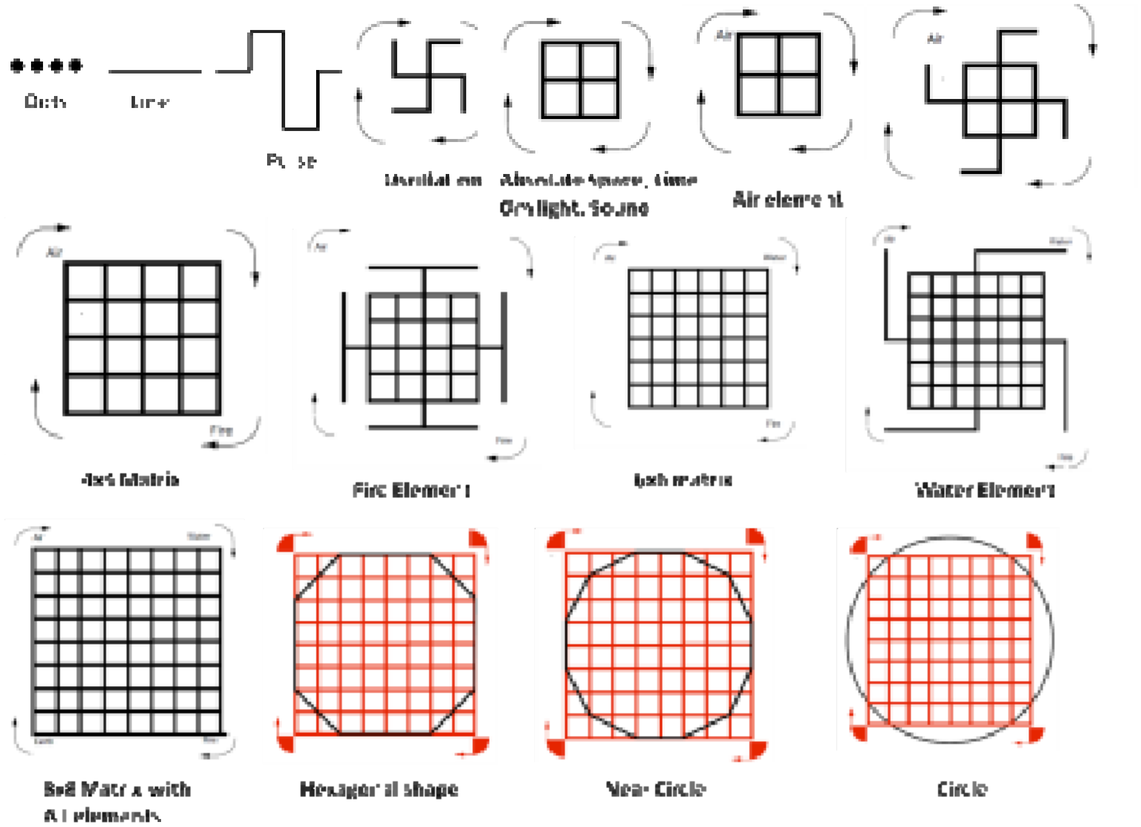
Dots
Line
Pulse
Absolute Space, time
Or light, Sound
Air element
Air
4x4 Matrix
Fire
Fire Element
6x6 matrix
Water
Water Element
8x8 Matrix with
All elements
Hexagonal shape
Near Circle
Circle

Figure 9: Transition forms of Brahman in detail (Coutesy: Fabric of the Universe)

If you look at the Fig 10 conceptually the following things happen:

- Subtle Energy becomes gross universe and the fundamental 5 elements are created in the process.
- It is said that 1/10th of the space becomes Air, 1/10th of air becomes Fire, 1/10th Fire becomes Water and 1/10th of Water becomes Earth.
- So we can conclude that we are living in a **Space delineated world.**

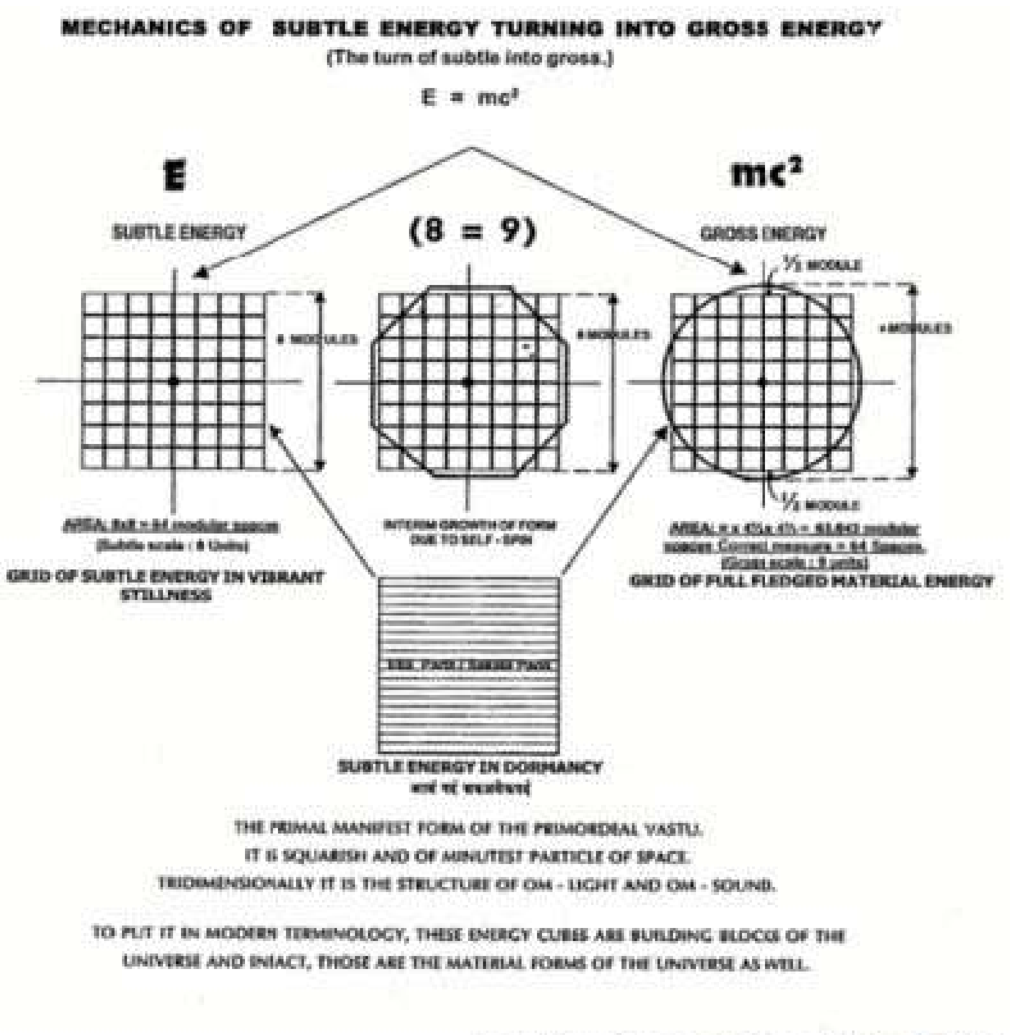
MECHANICS OF SUBTLE ENERGY TURNING INTO GROSS ENERGY
(The turn of subtle into gross.)
E = mc²
E
mc²
SUBTLE ENERGY
(8 = 9)
GROSS ENERGY
½ MODULE
8 MODULES
8 MODULES
8 MODULES
½ MODULE
AREA: 8x8 = 64 modular spaces
(Subtle scale : 8 Units)
GRID OF SUBTLE ENERGY IN VIBRANT STILLNESS
INTERIM GROWTH OF FORM DUE TO SELF - SPIN
AREA: π x 4½ x 4½ = 63.643 modular spaces Correct measure = 64 Spaces. (Gross scale : 9 units)
GRID OF FULL FLEDGED MATERIAL ENERGY
SUBTLE ENERGY IN DORMANCY
THE PRIMAL MANIFEST FORM OF THE PRIMORDIAL VASTU.
IT IS SQUARISH AND OF MINUTEST PARTICLE OF SPACE.
TRIDIMENSIONALLY IT IS THE STRUCTURE OF OM - LIGHT AND OM - SOUND.
TO PUT IT IN MODERN TERMINOLOGY, THESE ENERGY CUBES ARE BUILDING BLOCKS OF THE UNIVERSE AND INFACT, THOSE ARE THE MATERIAL FORMS OF THE UNIVERSE AS WELL.

From Dr. V. Ganapati Sthapati, Kauai, 2005

Figure 10: Subtle Energy to gross (Courtesy: Dr. Ganapathi Sthapati)

This wisdom was so succinctly captured by my good friend **Lao Tzu** in "**Tao Te Ching**" as follows: (Read heaven as un-manifest space and Earth as manifest gross)

"Heaven and Earth last forever.
Why do heaven and Earth last forever?
They are unborn, so ever living….

Thirty spokes share the wheel's hub;
It is the center hole that makes it useful.
Shape clay into a vessel;
It is the space within that makes it useful.
Cut doors and windows for a room;
It is the holes which make it useful.
Therefore benefit comes from what is there;
Usefulness from what is not there…."

But I wanted to highlight more important aspects to this process of manifestation.

- **Advaita** philosophy states that there is no duality. If you note that the **Parammanu** is what is present every living being and inanimate thing. What is within you is what is within me and hence the mahavakya **"Tat Tvam Asi"… You are that**.
- **Vastureva Vaastu** – Vastu which is unmanifest and subtle becomes Vaastu which is manifest and gross.
- **If we assume that Paramaanu is God's particle and hence it is divine, then everything in this world is divine**. This is the most important philosophy in **Hindusim**. See **the oneness in all** and see "**All in one"**.
- That's why Hindus saw divinity in everything. Trees, snakes, animals, insects, birds, human beings etc. There is nothing without that divine energy or particle – however you call it.
- This philosophy is **UNIVERSAL, OPEN SOURCE, ALL COMPASSING and the highest state of consciousness**.

But let's look at something even more important:

- We all note that Subtle has become Gross. Hence conceptually Gross = Subtle or both are proportionate.
- We all note that Space has become Earth. Hence conceptually Space = Earth or both are proportionate.
- We all note that 8 have become 9. Hence conceptually 8 =9 or both are proportionate.
- We all note that ENERGY has become MATTER. Hence conceptually ENERGY = MATTER or both are proportionate. This is represented by $E = mc^2$ by our great scientist Einstein. While we did not have this specific formula, the concept and the associated texts were very much with us for ages.

So this is what Einstein quoted after reading Bhagavad Gita.

"When I read the **Bhagavad**-**Gita** and reflect about how God created this universe everything else seems so superfluous." ~ Albert **Einstein**

Now let's try to answer this question:

- When you throw a stone in water / pond the ripples that form take the shape of a circle (in 2D) / sphere (in 3D). Can you guess what the shape of the ripple would be if you throw the same stone in space?
- **You would have probably guessed the answer. The ripples in space should be and would be in square /cube in shape.**

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Happy reading!

12. TIME = SPACE SERIES. PART 12 – VEDIC CONCEPT OF TIME & ITS MEASUREMENTS

The importance of time (**Kaalam**) is so high in the Vedic tradition that they ensured that the accuracy, continuity of time and the easy measurability of it is maintained for who knows how many years. **Time creates, sustains and destroys** everything and this is very nicely captured in these 2 lines:

காலமே எவைக்கும் மூலம், காலமே மூல ஆற்றல்
காலமே கோலத் தோற்றம், காலமே ஞாலத் தோற்றம்

The god of time is **Kaala Bhairava** – a form of Shiva and hence both **Shiva** and **Yama the lord of death both were** known as **Kaalan.** The time is represented by **number 13** and it belongs to Kala bhairava. We will see the interesting linkage between Vedic system and the Mayan calendar found in Guatemala and Mexico later.

Swami Chinmayananda states that "**Time is the periodicity between 2 thoughts**". Nothing captures the concept of time more wisely. What is clear from our scriptures and Swami Chinmayanda is that time is the pulsation that creates this universe. It is the vibration with very accurate periodicity and our sages suggested that you can live beyond time provided you manage to escape the effects of that pulsation.

Einstein's view on time and its relativity is very realistically captured in this quote **"When you are courting a nice girl an hour seems like a second. When you sit on a red-hot cinder a second seems like an hour. That's relativity."** More than relativity I found that the root cause of relativity is the pulsation of mind as time. Einstein on a serious note proved that our view of time is not absolute, it is relative and it is a continuum with space. We will in the next few parts of this blog would see that time and space are the same or Time = Space.

Aintiram states that there are 5 things that vibrates which have a common root and are not under our control. They are Space, Earth (vibration of earth is gravity), Eye lids, thoughts and Musical notes and very nicely put in Tamil as (Vin)**விண், மண்(man), கண்(Kan), எண்(Yen), பண்(Pan). Pan is musical metre / chandas and Yen is the yennam or thoughts.**

Correlate this with the first line of **Patanjali Yoga Sutra** which says **"Yoga Chitta Vrtti Nirodah",** means **Yoga is one which stops the vibration of mind / pulsation of mind and hence is a killer of thoughts.** The meditation techniques of keeping your eyes opened without battling the eyelids is another way of overcoming this pulsation instinct that is primordial and inherent in our nature.

Let's look how intricate our measurements time are in Vedic system. Check out the image below.

- The smallest unit of time is approximately 2 micro second.
- The largest unit is about 155 trillion years…. YET.
- Brahma's linear age is 50 years and in currently in his 51st year which makes the universe to be 155+ trillion years. Brahma's life time is 100 years which means it would be 300 trillion+ years of universal life which follows a cycle.

- After 100 years of pulsation cycles Brahma stays as dormant energy for the same period and then pulsates again for the next life of 100 years.
- A **Brahma's** life is also known as a Para. Each half Param is referred to as a *parardham*. This second half is termed as **'*dviteeya paraardhe*'** (the second half of Brahma's term) is stated in all vedic rituals.
- After every Chatur yuga, the world recycles. After every 2000 years / a Brahma day the world dissolves and remerges. The important point to note is that there is no CONCEPT OF CREATION here. It evolves and dissolves as per the science of pulsation. We also note that this pulsation expands and contracts. This cyclical nature of time as believed in Indian mythology refers to time as '*anaadi*' or that without a beginning.

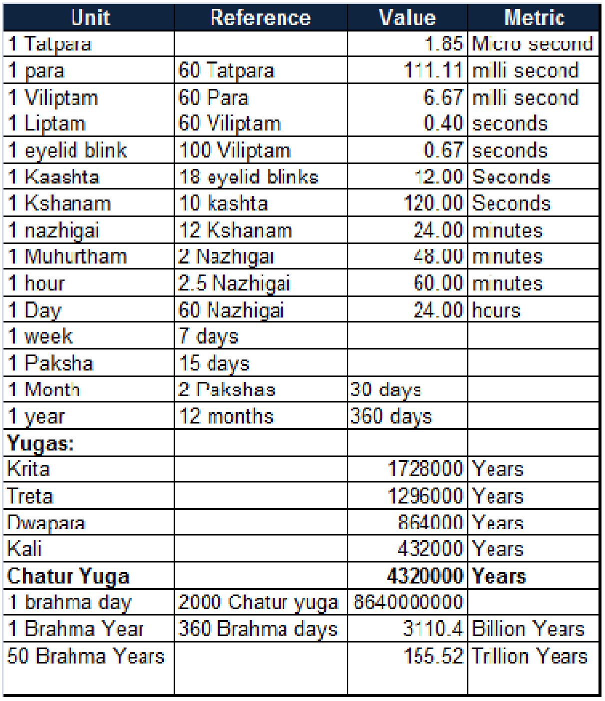

Unit	Reference	Value	Metric
1 Talpara		1.85	Micro second
1 para	60 Tatpara	111.11	milli second
1 Viliptam	60 Para	6.67	milli second
1 Liptam	60 Viliptam	0.40	seconds
1 eyelid blink	100 Viliptam	0.67	seconds
1 Kaashta	18 eyelid blinks	12.00	Seconds
1 Kshanam	10 kashta	120.00	Seconds
1 nazhigai	12 Kshanam	24.00	minutes
1 Muhurtham	2 Nazhigai	48.00	minutes
1 hour	2.5 Nazhigai	60.00	minutes
1 Day	60 Nazhigai	24.00	hours
1 week	7 days		
1 Paksha	15 days		
1 Month	2 Pakshas	30 days	
1 year	12 months	360 days	
Yugas:			
Krita		1728000	Years
Treta		1296000	Years
Dwapara		864000	Years
Kali		432000	Years
Chatur Yuga		**4320000**	**Years**
1 brahma day	2000 Chatur yuga	8640000000	
1 Brahma Year	360 Brahma days	3110.4	Billion Years
50 Brahma Years		155.52	Trillion Years

Figure 11: Time measurements in Ancient India

While doing Vedic ritual the time and space coordinate is invoked to say that you are doing this ritual at this place in the universe and at this time – with reference to Brahma's age to the time and star of the day. This tradition is unbroken and the Indian Almanac can predict the position of astronomical objects without any error for thousands of years without any sophisticated instruments and space ships.

On the contrary just read the history of the Gregorian calendar which we are following here, http://en.wikipedia.org/wiki/Gregorian_calendar. There is neither accuracy nor continuity and to top it up there are authorities who reform the calendar as if time stops and waits for them to affect that reform. Why should we ignore the best of time science that we have for an inaccurate, discontinuous reformed calendar?

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Happy Reading!

13. TIME = SPACE SERIES. PART 13 – SOUTH AMERICAN MAYAN & SOUTH INDIAN MAAYAN

Since we are discussing about time, let us address the **Mamuni Maayan** and the **South American Mayan**.

A lot has been talked about Mayan calendar and the imminent destruction to planet earth in the year 2012. I would recommend **'The Mayan Factor – Path beyond technology" by Jose Arguelles** if you are interested in this topic.

In this part I am interested in highlighting that the **South American Mayan** and the **South Indian Mamuni Maayan** who wrote **Aintiram and Pranava Veda** are one and the same. Extensive research have been conducted by many and I found, both the works of **Jose Arguelles and late Dr. Ganapathi Stapathi** who has spent about 3 decades each, very impressive. Dr. Ganapathi concluded that they both are the same and he has shown some interesting linkages.

I am going to present what impressed me in both these great researcher's works.

- First let me state that the doomsayers of 2012 phenomenon picked up Mayan's calendar which started in 3113 BC ran for 5125 years and the calendar ended in year 2012. Since the calendar ended they thought this signified that the world is ending. Jose argues that this is wrong and most probably the human race is entering into the next phase of evolution and the next cycle would begin. If you strongly believe that this is not true and the world would certainly end by 2012, then please ensure that you transfer all your assets and savings to my name effective 22, Dec 2012. ☺

- Mayan in his calendar in South America used very unique numerals in his calendar that is given below:

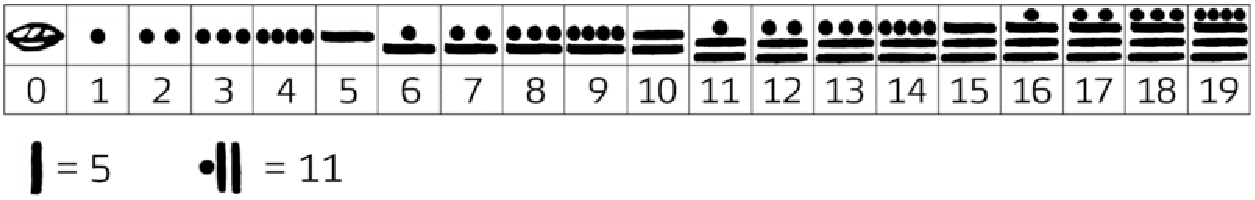
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
= 5
= 11

Figure 12: Mayan's numeral system (Courtesy: Wikipedia, Internet)

Now let me reproduce a Tamil poem written by Mamuni Maayan in his work called "**Kanithamaa sennool**". This song is part of the book written by Maayan in Tamil during his time – 12000 years+. This song expresses how the numerals are formed till 10. It says,

  - 1 is one dot, 2 is 2 dots, 3 is 3 dots, 4 is 4 dots, 5 is a line and 10 is a double line.

மாப்புள்ளி ஒன்றே அன்றோ ஒன்றுமா எண்ணே காட்டும்

உ

மாப்புள்ளி இரண்டே யன்றோ இரண்டுமா எண்ணே காட்டும்
மாப்புள்ளி மூன்றே யன்றோ மூன்றுமா எண்ணே காட்டும்
மாப்புள்ளி நான்கே யன்றோ நான்குறு எண்ணே காட்டும்

மாப்புள்ளி ஐந்தே யன்றோ ஐந்து மா எண்ணே காட்டும்
மாப்புள்ளி கோட்டினாலே இரண்டுரு மாக்கங் காட்டி
மாவியல் திறத்தினாலே பத்தறு மாக்கங் காட்டி
மாவியல் திறத்தினாலே எண்முதற் பொருளு மாவாய்!! (கணிதமா செந்நூல் - மாயன்)

- Among the books written during Mayan’s time in Guatamala are Popol Vuh, The annals of cakchiquels and the **Book of Chilam Balam**. The resemblance of **Chilam Balam to Chidambaram** with which S. Indian Maayan was so involved is very striking.

- Be the pyramids that were built in the jungles of S. America that has a striking resemblance to the viamanas from south India, the earth’s revolution around the sun measured within 1/1000th of the decimal point which is available in a book called **Surya Siddhaanta – In Sanskrit** by Mayan and the identical measurements available in South America – the similarities are difficult to ignore.

- Jose Arguelles makes a compelling argument about the shift in consciousness in 2012 due to a major galactic synchronization. But what impressed me are the following:

    - Comparison between **iChing (8x8 hexagram) and Mayan’s Tzolkin which is 13x20 matrix**. In fact there is a Franklin’s magic square which is based on the 8x8 grid of iChing that adds upto 260 which is the galactic constant of Tzolkin. This was amazing.

    - He refers to **Radiogenesis – Universal transmission of information through or as light** or radiant energy which is very much relevant to our discussions of light and sound.

    - The revelation he received from a Maya Humbatz men belonging to the universal great brotherhood organization revealing that our universe is the 7th such system which Maya has charted in the universe. So, we are not alone ☺

Would conclude this topic in the next part…

Happy reading!

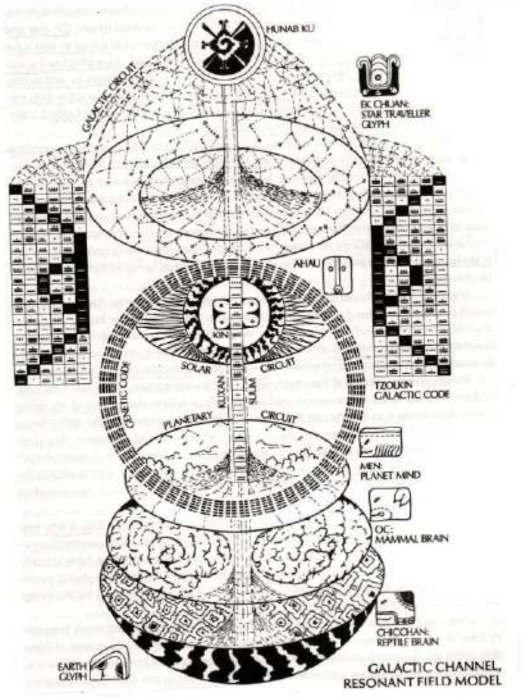
HUNAB KU
GALACTIC CIRCUIT
EK CHUAN: STAR TRAVELLER GLYPH
AHAU
KIN
SOLAR CIRCUIT
GENETIC CODE
KUXAN SUUM
PLANETARY CIRCUIT
TZOLKIN GALACTIC CODE
MEN PLANET MIND
OC: MAMMAL BRAIN
CHICCHAN: REPTILE BRAIN
EARTH GLYPH
GALACTIC CHANNEL, RESONANT FIELD MODEL

Figure 13: Galactic Communication Channel to Hunab Ku (Courtesy: The Mayan Factor, Jose Arguelles)

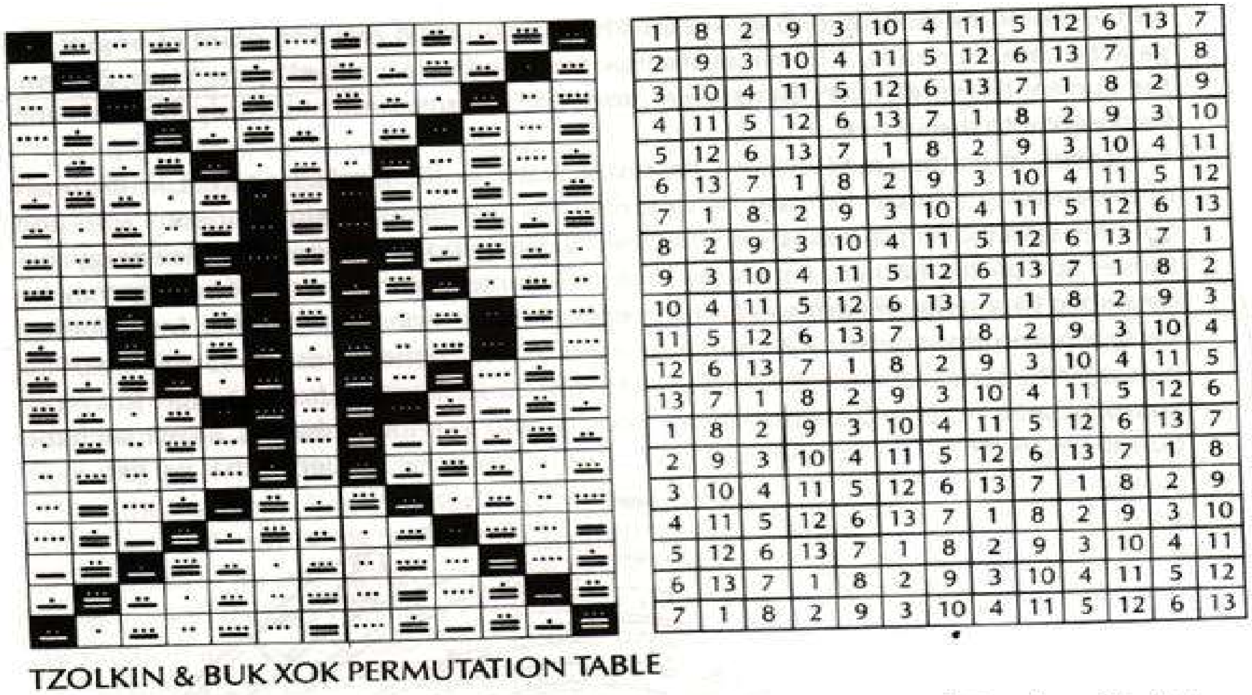
TZOLKIN & BUK XOK PERMUTATION TABLE

1	8	2	9	3	10	4	11	5	12	6	13	7
2	9	3	10	4	11	5	12	6	13	7	1	8
3	10	4	11	5	12	6	13	7	1	8	2	9
4	11	5	12	6	13	7	1	8	2	9	3	10
5	12	6	13	7	1	8	2	9	3	10	4	11
6	13	7	1	8	2	9	3	10	4	11	5	12
7	1	8	2	9	3	10	4	11	5	12	6	13
8	2	9	3	10	4	11	5	12	6	13	7	1
9	3	10	4	11	5	12	6	13	7	1	8	2
10	4	11	5	12	6	13	7	1	8	2	9	3
11	5	12	6	13	7	1	8	2	9	3	10	4
12	6	13	7	1	8	2	9	3	10	4	11	5
13	7	1	8	2	9	3	10	4	11	5	12	6
1	8	2	9	3	10	4	11	5	12	6	13	7
2	9	3	10	4	11	5	12	6	13	7	1	8
3	10	4	11	5	12	6	13	7	1	8	2	9
4	11	5	12	6	13	7	1	8	2	9	3	10
5	12	6	13	7	1	8	2	9	3	10	4	11
6	13	7	1	8	2	9	3	10	4	11	5	12
7	1	8	2	9	3	10	4	11	5	12	6	13

Figure 14: Tzolkin table (Courtesy: The Mayan Factor, Jose Arguelles)

14. TIME = SPACE SERIES. PART 14 –MAYAN, TAMIL, PLEIADES & ICHING

Jose Arguelles was also mentioning that the pyramid structures of Mayan civilization in South America with the **Tzolkin** calendar matrix communicates with 2 constellations – **Pleiades and Arcturus**. Jose mentions that he is sure about communication with Pleiades constellation.

Why did Pleiades interest me?

- According to Tamil history, Tamil is a divine language and the Lord associated is **Lord Muruga**. He is the Tamil deity and he is from the **Constellation Pleiades** which is called **Kritika nakshatra**.
- This emphasis that the South American Mayan could be a Tamilian from India, as many in the ET studies circuit opine that Pleiades is a constellation which is like our distant cousin where Tamil is spoken. There are many links but you can check this out.. http://22050hz.blogspot.com/2008/10/pleaidians.html

- Look at the Mayan calendar in Fig 14, it is a 13x20 Tzolkin matrix.
    - Please recall that we mentioned **13 is the number for Kala Bhairava**.
    - Tzolkin has a mystical central column around which 6 and 6 columns exist. The belief is that this central column is the frame of reference to the galaxy and it is believed to connect to **Hunab Ku – the center of the galaxy**.
    - This mysterious centre column is believed to be the road to the sky leading to the umbilical cord of the universe. I immediately correlated this with the **thread of consciousness** which is connected to the Hiranyagarba (the center of the galaxy as per Indian texts.) – Check Figure 13.
    - So understanding that this shaft of consciousness is connected to the center of the galaxy (**Hunab Ku / Hiranyagarbha**) at the one end and to **our solar plexus** at the other end with communication and galactic information transmission happening through resonance of light, opened many gates within me.

Nothing in this universe exists without purpose and the form and shape of any being / thing is explained as a state of consciousness at a particular resonant frequency adds to the Seelam (Rhythm) aspect of the Kaalam which we are going to see later.

For the space and time travel enthusiasts, Mayan has the following:

- All space travel is intelligence as information – transmitted through light by the principle of harmonic resonance.

- **Mayan says that Sun is the lens through which galactic information is transmitted from the galactic core and interpreting this with Gayatri Mantra where we meditate on the supreme light Savitr – an aspect of Sun god is amazingly striking.**

- Mayan highlights that time is a function of principle harmonic resonance and he attaches significant importance to the numerals which he states are the galactic harmonic constants. This is exactly what **Aintiram** states.

- **He interprets the iChing to contain the Genetic code of the human DNA as a hexagram in a 8x8 matrix which is a binary progression to the 6th power corresponding 64 six part words /codons. At the same time he interprets number 260 to be galactic constant and 360 which is the factor for a calendar year or the total angle in a circle to a harmonic calibrator.** His explanations more on this and on the mystical number 7 do not resonate in my tiny brain and I am begging for holy grace to expand my mind to understand these concepts.

- **When he says that flowers open to light just does not mean that it is sensitive to be light but deeply it aspires to be LIGHT.** I am not sure if you understand the depth of this statement as against the Sage molar statement which states gross light should merge with subtle light. I have been left speechless for days after understanding this correlation.

It is interesting to note that significance of numerals as per Mayan factor:

- **1 – The pulsation – Ray of Unity**
- **2 – The pulsation - Ray of Polarity**
- **3 – The pulsation - Ray of Rhythm**
- **4 – The pulsation - Ray of Measure**
- **5 – The pulsation - Ray of center**
- **6 – The pulsation - Ray of organic balance**
- **7 – The pulsation - Ray of mystic power.**
- **8 – The pulsation - Ray of harmonic resonance.** We will see more of this.
- **9 – The pulsation - Ray of cyclic Periodicity**
- **10 – The pulsation - Ray of Manifestation**
- **11 – The pulsation - Ray of Dissonant Structure**
- **12 – The pulsation - Ray of Complex stability**
- **13 – The pulsation - Ray of Universal movement**

In all 2 independent researchers working 6000 miles apart, conducting their studies independently seems to be pointing to just one person and one philosophy and science. What else we could do other than thanking our fortune to have read them and made this connection. I thought of writing just a page on this topic and I have ended up with 2 parts and I am stopping my temptation to write few more pages on this topic. For all of those interested, please read the book by Jose Arguelles.

It is amazing to find out how most of the ancient civilizations seem to have a common root and the communication between various geographies looks stunningly simple and the metrics on astronomical objects and other space sciences seems to be so accurate despite the absence of precision instruments and spacecrafts. One tends to believe that we had far advanced knowledge of nature which we seems to have lost and most certainly few thousand years back we were not nakedly roaming in the jungles of India and Africa hunting animals as the western experts wants you to believe but had a far superior & scientifically advanced civilization.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy Reading!

15. TIME = SPACE SERIES. PART 15 – SHIVA LINGAM IS NOT A PHALLUS

If there is one part in the series I want you to read and talk about it with your friends then it is this topic. If I ask you as to what Shiva Linga stands for then more than 99% of the responses would be that it represents a **Phallus** symbolizing the male sex organ / creative energy with the female sex organ / creative energy - Yoni as the base and both together representing **inseparability and totality in the process of creation**.

Nothing is more misunderstood and misquoted than this concept of Hinduism. Every idiot including me who has read few books about Hinduism thinks that they are experts of Indology / Vedic studies. The wisest of sages who gave us Vedas and Upanishads did not boast that it is their work. At the end of every work they acknowledged that they are telling us what they have been told by their ancestors. It is 155 trillion+ years of universal knowledge available today in the richest of the languages Tamil and Sanskrit. This cannot be accurately interpreted in a foreign language. Lest in the weakest of all languages - English.

The other contributing factor to this misunderstanding has been the Sanskrit language or should I say the richness of the language.

We have already noted that the same word can have multiple meanings. This is not limited to Sanskrit, but the all the languages. For example,

- **Sutra means thread, Aphorisms** (the ability to express something in terse and concise manner). Ex. **Kama sutra** – the most understood of all the sutras. **Brahma Sutra** – The thread of consciousness in any form that runs in this universe – the most unheard of.
- **Artha** in Sanskrit means “meaning” and “wealth”, we have already seen this. As in Vaak-Artha and Artha Sastra (the science of wealth management).
- Same way Lingam means **Symptom, Proof**, **Gender, Male sex organ** etc. Hence it was very easy to get confused, associate the lingam with a Phallus and also to ignore the philosophy and the supreme science behind it. Even people who believed that it represented much more did not knew the philosophy and the science behind it. Let’s look at what it means.

I am going to request you to kindly refer to both the pictures below:

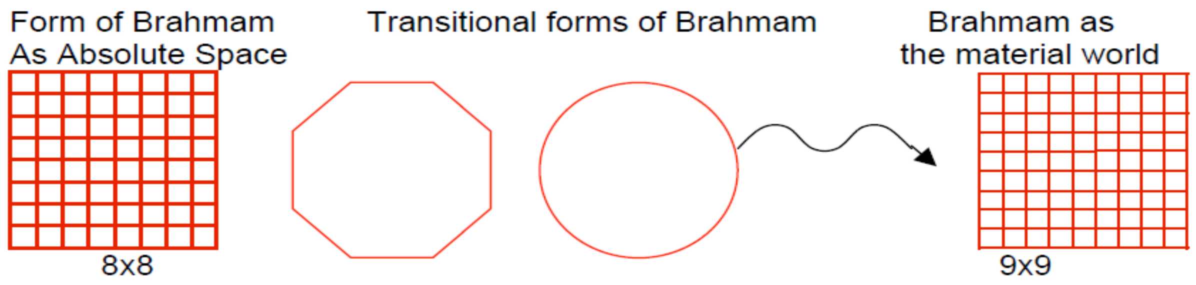

Figure 8 – reproduced here for convenience

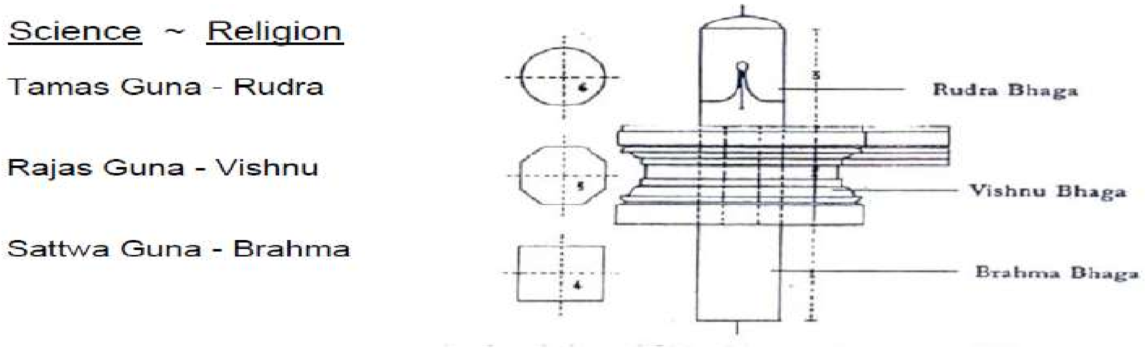

Figure 15 – Shiva Linga – Shape and Significance (Courtesy; Dr. Ganpathi Sthapati)

We have already seen that the subtle energy is takes the form of a square and the gross energy takes the form of a circle with an in-between state it takes the shape of a **tetrahedron and octagon**. Now look at the Shiva lingam from the top. Its base is always a square and the top is always a circle.

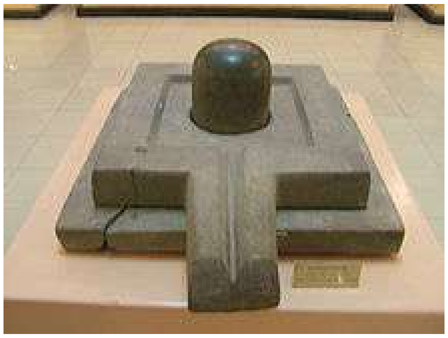

Figure 16 – Shiva Linga

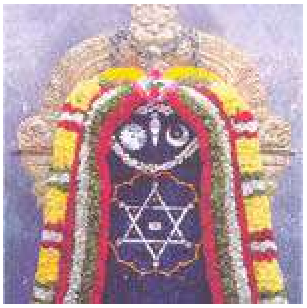

Figure 17 – Shiva Linga with shape of the God's particle (Star) engraved.

If I have to merge all the 3 shapes together and it naturally takes the shape of a Lingam with a round top. **Shiva Linga captures the overall process of evolution in one form and shape - subtle to gross state – it encompasses every state of consciousness that is possible in this universe.**

Why the figure above states that Square shape is Sattvic, circle is Tamasic and the in between state is Rajasic? Let's understand the definitions of all the three qualities.

- **Sattvic represents the perfect state of balance. When it is subtle it is luminescent and in a perfect state of resonant non-vibration. The energy levels were perfect so that the resultant vibration is nil and it is in absolute balance.**

- **Tamasic state represents a state where the height of evolution is reached and hence it is farthest away from the core.** This is a different explanation from the usual one available in the web and books. Let me explain this. The core is very subtle and the tamasic state is very gross and hence it is away from the core. Tamasic state then should be considered as the darkest state possible. Why?

  - Because the core is luminescent and if the gross state is farthest from the core hence it should be dark. This is the meaning behind the most famous statement **'Tamasoma Jyothir gamaya"…** take me from darkness to light. It is not about just your mind but about the entire being aspiring to become a light and merge with primordial light.
  - **There is no evolution possible beyond the Tamasic state and it has to devolve to reach its core / subtle state.**

- **Rajasic is a state where the pulsation has resulted in a rotational force and hence things are set into action. This represents the evolution process and hence associated with action.**

Then the natural next question is, if Shiva Lingam is associated with the entire spectrum of consciousness then we should see it associated the elements and the light. Isn't?

- There is a story in Indian Purana that Brahma, Vishnu and Shiva discussed who is the most powerful among them. Lord Shiva challenged Brahma and Vishnu to reach his head and feet respectively. Brahma tried to find Shiva's head and Vishnu tried to reach Shiva's feet, both of them failed. Most of us would have heard the story but would not have understood the philosophical and scientific significance.

  - Lord Shiva in the form of Linga stood there as a **PILLAR OF LIGHT**. This pillar of light is **the Brahma Sutra / light of consciousness** which spanned from the center of the galaxy (Hiranyagarba) to every being & non-being in the Universe.
  - Brahma could not find the head because; Brahma is always associated with intelligence in Indian texts. He gave the supreme wisdom - the Vedas and stands for Intelligence. He could not reach Shiva's head signifies that Intelligence will not take you to God. Intelligence as an energy is too gross and cannot reach the most subtle state of light.
  - Vishu represents earthly life – Rajasic in character – this is the tetrahedron & octagon state. He could not reach Lord Shiva's feet signifying you cannot reach the core through actions.

- Lord Shiva in the form of Lingam is represented as one of the elements (Earth, Water, Fire, Air, and Space) in the Hindu temples.

If you read the following Upanishad quote, you can probably understand the meaning:

na karmaNaa na prajayaa dhanena tyaagenaike amRitatvamaanashuH
pareNa naakam nihitam guhaayaaM vibhraajate yadyatayo vishanti

Not by work, not by progeny, not by wealth, they have attained Immortality. Some have attained Immortality by renunciation. That which the hermits attain is laid beyond the heaven; yet it shines brilliantly in the (purified) heart.

And this quote also brilliantly articulates the luminescent nature of the Paramaanu:

na tatra (not there) **suryo bhati** (sun shines) **na candra** (moon does not) **na tarakam** (nor the stars)
nema vidyuto (nor the lightning) **bhanti** (shines) **kuta ayam agnih** (how can fire shine)
tam (that) **eva** (alone) **bhantam** (shines) **anubhati** (thereafter shines) **sarvam** (everything)
tasya bhasa (in that light) **sarvam idam vibhati** (everything shines)

The sun does not shine there, nor does the moon, nor the stars or the lightning, much less this fire (deepam). When He shines, everything shines after Him; by His light alone everything is illumined.

The trick is to take the definition & meaning from Upanishads or any scripture and apply it to the God's particle and the process of evolution and you would understand it perfectly. Many a times the concept of God & faith obscures our mind to look beyond and rationally think about it. It is not the problem of religion but our system of education which has ruined our thinking. **Today the biggest challenge we have is to "Unlearn" than to learn.**

So to conclude:

- Shiva Lingam does not represent Phallus. To represent male and female totality and inseparability Hindus do have **Arthanareeswara** and it is not represented through Lingam.
- Shiva Lingam represents the **Pillar of Light and the entire spectrum of consciousness**. From the most subtle to the most gross.
- It represents an in-between **Guna** / **EGG** state (between un-manifest and manifest) and hence it is formless with a form.
- **While Lord Nataraja represented the rhythmic pulsation of this Brahma sutra, the cosmic dance / the dance of the evolution, Shiva Lingam represents the primordial light itself in all its state of consciousness.**
- By this analogy, our body itself is a lingam with the visible body as the gross part at one end and the most subtle body which is invisible and is part of the primordial light at the other.

Check out the picture below which represents most of the aspects we have discussed so far.

May Lord Shiva help us all to understand him better.. ☺

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions." Yet wish that this quote isn't wiser any more.**

Happy reading!

உ

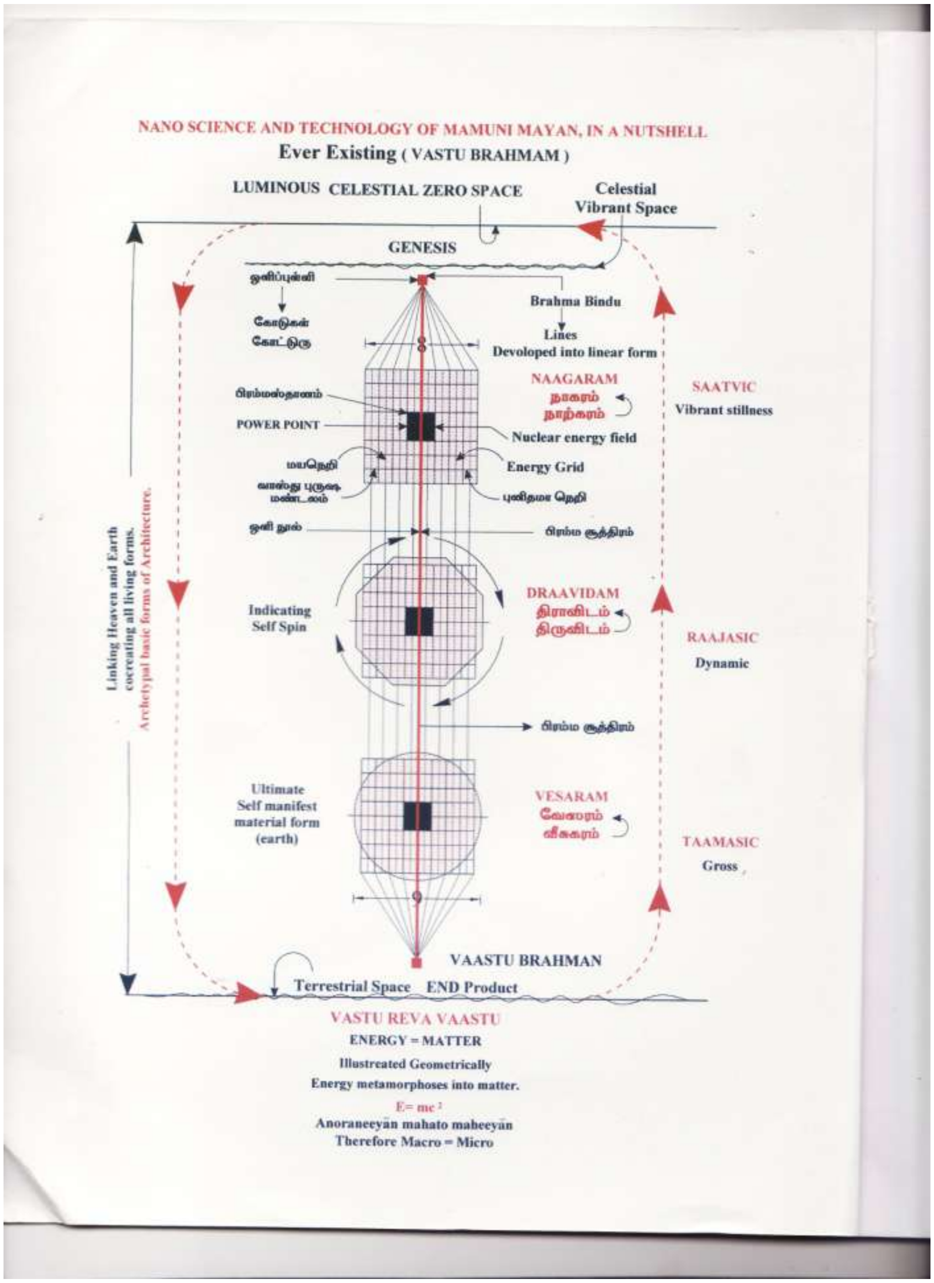
NANO SCIENCE AND TECHNOLOGY OF MAMUNI MAYAN, IN A NUTSHELL
Ever Existing (VASTU BRAHMAM)
LUMINOUS CELESTIAL ZERO SPACE
Celestial Vibrant Space
GENESIS
ஒளிப்புள்ளி
கோடுகள்
கோட்டுரு
Brahma Bindu
Lines
Devoloped into linear form
8
NAAGARAM
நாகரம்
நாற்கரம்
SAATVIC
Vibrant stillness
பிரம்மஸ்தானம்
POWER POINT
Nuclear energy field
மயநெறி
Energy Grid
வாஸ்து புருஷ மண்டலம்
புனிதமய நெறி
ஒளி நூல்
பிரம்ம சூத்திரம்
Linking Heaven and Earth cocreating all living forms.
Archetypal basic forms of Architecture.
DRAAVIDAM
திராவிடம்
திருவிடம்
Indicating Self Spin
RAAJASIC
Dynamic
பிரம்ம சூத்திரம்
Ultimate Self manifest material form (earth)
VESARAM
வேசரம்
விசுகரம்
TAAMASIC
Gross
9
VAASTU BRAHMAN
Terrestrial Space END Product
VASTU REVA VAASTU
ENERGY = MATTER
Illustreated Geometrically
Energy metamorphoses into matter.
E= mc²
Anoraneeyān mahato maheeyān
Therefore Macro = Micro

Figure 18 – Manifestation in a nut shell (courtesy: Dr. Ganpathi Sthpati)

This diagram represents everything we have discussed so far. The central red line is the Brahma Sutra which runs from the center of Galaxy to the terrestrial forms. Just correlate this with the Mayan Hunab Ku in the previous part.

This summarizes as to how Energy becomes matter, Subtle becomes Gross, Un-Manifest becomes manifest, how 8 becomes 9, and Square / Cuboid becomes a circle / Sphere. All represented in one formless form called SHIVA LINGAM.

உ

16. TIME = SPACE SERIES. PART 16 – SEELAM - THE PRIMORDIAL CIRCULAR FORCE - GANAPTHI & MATHEMATICS

In the last few decades the main stream western science has started accepting concepts like uncertainty principle, wave mechanics, quantum mechanics, and periodicity of the wave pattern. Most importantly it is about the fundamental wave pattern of any energy or matter and the harmonics of these wave patterns falling in a particular form and proportion which is dictated by geometry / mathematics. In all the western scientists are scratching their heads on the depths of science of the subtle aka the sacred spatial geometry.

Most of the eastern philosophies have mastered this science and have codified these natural laws as a part of their religion, which is often dismissed as superstition. I already highlighted that in Tamil and Sanskrit the definition for science is "**Knowledge of the space**".

Modern science is moving in the right direction albeit very slow and they are trying to understand as to how elements are bonded within a molecule in addition to what elements constitute them. For example **Chlorophyll** molecule consists of **Carbon, hydrogen, nitrogen and magnesium (Mg at its core)** has a complex 12 fold Pattern – just one such typical pattern which can convert sunlight into life substance. The same pattern exists in our **RBC** cells just that it has an **iron** at its core instead of magnesium.

The understanding about the human body especially, "the spatial awareness that exists in each and every cell is due to the different sensory organs tuned at frequencies", led the scientists to unravel the innate spatial geometry of life.

Be it living, non-living beings, abstract forces like gravity, nuclear force, electromagnetism, heat, movement of astronomical bodies it does not matter what, everything in this universe adheres to a vibration / periodicity and is reducible to a number or a ratio. As a wise man said, "**Mathematics is the language of gods". This primordial rhythm & its harmonics are captured under the topic "Seelam" by Mayan in ancient tamil texts. Let's look at this in detail.**

I have been finding it difficult to sequence the flow for this topic and let me try starting from the basics:

- We know that the **Paramaanu or God's particle is a cube.** This particle with the primordial vibration combines with other particles to create a shape. **The Paramaanu combines in the multiples of 8 to evolve the next structure**. Let us see the importance of 8 later, but remember 8 is the number for Lord Narayana and Om namo Narayana has 8 Aksharas.

- What do you call a cube called in Tamil? It is called **Ganam (கனம்**). This also means heaviness, gravity, honor, dignity, abundance, plenty. To indicate it is a square cuboid they call it **கன சதுரம்**.

- The process of the addition of **Paramaanu** into bigger units is called as **Ganam (கணம், गणं)**. In Tamil both the Cube and the additive resultant are called as **Ganam** and the only difference being that in Cube it uses a **small "na" (ன)** as an alphabet. In the additive process / resultant it uses a **Big "na"(ண)**. The big "na" signifies that due to the additive process the small (na) becomes big (na).

- This additive process starts because of the **first rhythmic circular** force also called as a **Suzhi (சுழி)** in tamil which means "loop".

- Here comes the beauty. The Lord of or the energy associated with this **additive force (Ganam)** is called as **Ganapathi (கணபதி, गणपती)**. Note this is the bigger "**na**" used and not smaller **"na"**. He is associated with the first circular force - **Suzhi** (and hence he takes the first honor in any Vedic ritual. This means Lord Ganapati is the first process in the evolution and everything has to start with him. Even if it is Lord Shiva or Vishnu it does not matter, the very first honor goes to the **Lord of Ganas**. Many Indians when they start writing (atleast few decades back) they first start with this symbol at the top of the page before anything for a successful completion of that activity.

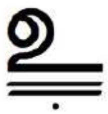

- If I add my own inferences to the above with a caveat that everything else I have written have a solid back up mostly from our own study of languages (as you can yourself see) and our scriptures.
    - The symbol above signifies that what was "a point" with untouched / non-interacting energies / resulted in the primordial first circular force out of a pulse and also signifying that every circular force shall also collapse back to a point.
    - Also the bigger / heavier things are formed or created by its smallest unit and hence symbolically the heavier ones are carried by the smaller units. To signify this may be a god in the form of an elephant is carried by a mouse.

- Now you know why every ritual starts with this Sloka to Ganapathi:

    - ***"GanAnAm tva Ganapati gum Havamahe, Kavim Kavinam Upamasra vastamam Jyestharajam Brahmanam Brahmanaspat aana Srnvan nuti bhissi dasadanam"***

      This means "May you the Lord of Ganas, Ganapati",... . so it goes.

- **Now what do we call the study of this additive force - Ganas? Mathematics – Ganith / Ganitham (கணிதம், गणित).**

- **So we Indians have codified:**
    - **the primordial God's particle as Paramaanu or microbode or Lord Shiva,**
    - **the rhythmic movement of the pillar of light as Lord Nataraja,**
    - **the first circular force as the symbol for Lord Ganapathi,**
    - **the lord of the additive force as Lord Ganapthi and**
    - **the Study of this additive force which leads to the formation of everything in the universe is Mathematics in Sanskrit and Tamil.**

- **Every possible state of consciousness and manifestation is represented in Shiva Linga**
- **We will see that the multiples of 8 in which the additive force moves is codified as Lord Narayana or Vishnu and the five state manifestation process as the five letters associated with Lord Shiva.**

Just forget that these names are the names of the Hindu gods, and tell me one religion or philosophy or one branch of science in this world which can capture the nature's primordial secret in this way and codify it for the common benefit of the world and carry it for thousands / millions of years.

These are just a start and please wait for the other interesting aspects to unfold.

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Happy Reading!

17. TIME = SPACE SERIES. PART 17 – SEELAM - THE PRIMORDIAL RHYTHM & FUNDAMENTALS

Now, let us get into some real heavy stuff - Mathematics (**Ganitham** – the study of Ganam) that governs this rhythm and harmonics. My first concern is how to make it easily digestible for an average individual who hates Math? So I shall take the easiest route that I know. I shall first explain the concept, then the actual Math part of it and then go to the philosophy and significance of the same. Later we shall look at how these are codified in nature and our scriptures and available in front of our eyes which we have failed to recognize.

Let us be clear about the following terminologies like **Progression, Proportion, root, Diagonal and Pythagoras theorem**:

- Let us take a series as follows: 2, 4, 8, 16, 32…. In this we need to be clear about what is **Progression and proportion**. As many of you know that whole series is a geometric progression where the next number in the series is a function of the current number and a multiple. This multiple here is 2, so that 2*2 = 4, 4*2 = 8 and so on. **This multiple or the constant is called as proportion**. You derive the value of the proportion by dividing the number by its previous number – (n+1)/n : (n+2)/(n+1). Hope this is clear.

- We know that a square has 4 equal sides. Each side of a square is called as "**root**" and the line that connects the 2 opposite ends is called as '**Diagonal**". In the figure below sides AB, BC, CD, DA are called as "**Root**". Lines AC, BD are called as '**Diagonal"**.

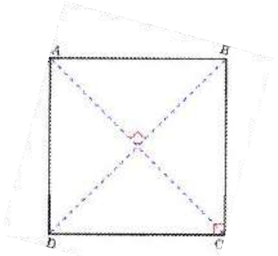

Figure 19; Square

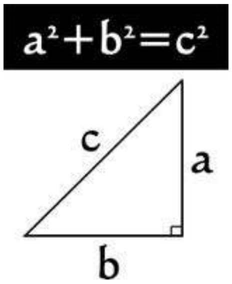

Figure 20: Pythagoras theorem

- **Pythagoras theorem** – This theorem gave us the formula to calculate the sides of a right angle triangle as given above. If you are not familiar with this, then please understand that if you know any 2 sides of the triangle the 3rd side can be calculated by the formula given above. I am not going to focus on the argument that this theorem was available in Indian scriptures before the Greek mathematician gave it. Let's look at the concepts and not which civilization gets credit as of now.

Since the fundamental structure is a square in 2D and Cube in 3D and also in the process of evolution one becomes many, let's ask ourselves the following questions:

- **How one square does create another square which is larger in size, progressively? What is the progression and proportion in this series?**
- **How one cube does create shapes which are polygonal (Hexagon, Octagon, decagon, dodecagon etc) progressively? What is the progression and proportion in this series?**

Now let us see Figure 21 where and consider the square ABGF. You can enlarge the image to see the alphabets clearly. Now as we know sides AB, BG, GF, AF are called “**root**” – remember this well and let us assume the smallest root possible is unity and hence each root is of unit length. AB=BG=GF=AF=1. Area of this square = 1*1 = 1.

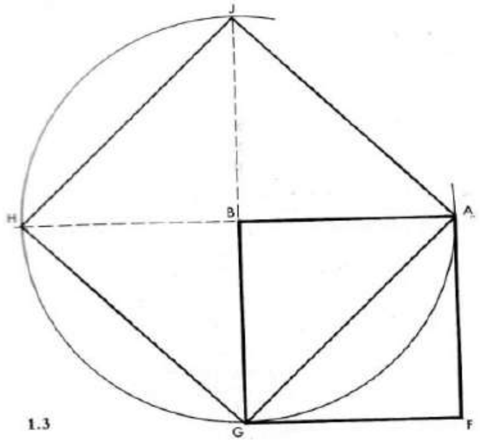

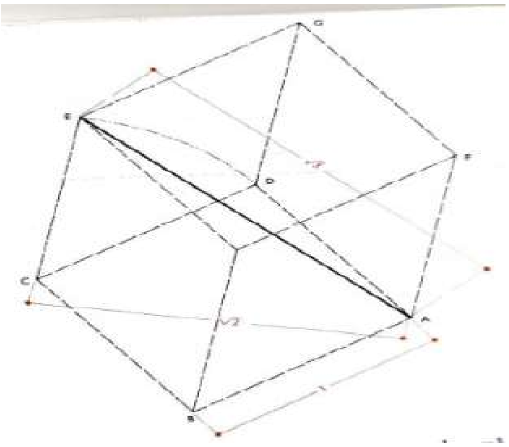

Figure 21; Square

Figure 22: Cube

Line AG is called as the **“diagonal”** and as per Pythagoras theorem the length of **AG** should be **√2**. **AG = √(AF^2+GF^2).** So to summarize **the root of the square ABGF is 1 and the diagnol is √2**.

Now consider a larger square AGHJ, here AG, GH, HJ, AJ are the sides and hence called as root. AH, GJ are the diagonal in the 2nd square. We know AG = √2, then AH = GJ = 2 as per Pythagoras theorem. To sum up **the root of the square AGHJ is √2 and the diagonal is 2. Area of this square is 2.**

If you look at the ratio **root/diagonal** (root to diagonal ratio) of both the squares it is 1/√2 and √2/2 which results in 1/√2. This will progressively hold good. The magic proportion in this progression is **√2**.

What we can infer from this is if we consider a **progression of squares** from a unit square to larger squares, then the **proportion** of the progression is √2. But before we jump into its philosophical significance let’s answer the second question.

Consider the cube above whose side is of unit length. We know a cube is made of squares. We know ABCD is a square of length (root) =1 and diagonal = AC = **√2**. If we want to find out the length of the diagonal of the cube which is AE, then ACE is a right angle triangle with AC = **√2**. And CE =1 and hence **AE = √3**.

What we can infer from this is if we consider a **progression of cubes** from a unit cube to larger ones, then the **proportion of the progression is √3**.

உ

I am going to handle the significance of √2 and √3 in the next part.

Happy reading!

18. TIME = SPACE SERIES. PART 18 – SEELAM - SIGNIFICANCE OF √2 & √3

Let’s look at the significance of **√2 and √3** but before that let’s capture the philosophical significance of **Progression** and **Proportion**:

- The whole world is conceptualized as wave harmonics of the primordial rhythm with a **progression and proportion**. Philosophically the **Proportion** which does not change / which is constant is the **immutable component** and the **progression** which changes / is volatile is called as the **Mutuable component** in the process of evolution. **The whole universe is the progression and God is the proportion**. So if I know the God’s proportion and the formula behind it I can create anything in this universe including a universe. This secret is called as the Brahma Vidya / Sree Vidya (Knowledge of Brahman, Sree) in Indian texts. What we are going to see further are very minor parts of the **Brahma Vidya**. ☺

- In the last part we saw, that the primordial square multiplied itself to be a larger square (**progression of squares**) using the **proportion √2. The diagonal of the square 1 (ABCD) forms the root of Square 2 (AGHJ)**. This is very profound if you get the concept**.**

- **If we consider square 1 as the cause then its diagonal is the effect. This diagonal is the root for the next square (AGHJ) and hence the effect of Square 1 is the cause of Square 2, this cause results in square 2’s diagonal which is its effect and this is how the progression or evolution happens.** If you can see glimpses of Cause and effect cycle and the famous law of Karma in this analogy then you can pat yourself. ☺

- A square halved by the diagonal (square 1 with area 1) produces a square twice its area (square 2, area = 2). **The mystery of biological growth from cellular division or the different musical notes from the base tone is contained in this**.

- **Robert Lawlor in his book “Sacred Geometry”** very nicely puts it as the root of a plant (like the root in a square) is causative and embedded in the earth (and embedded in the square). These are very heavy and profound concepts just let it sink into you.

- When you divide the full height of the human being considering the total height as unity then belly as called as **Hara in Japanese, Dantein in Chinese, Nabhi in India** which is below the navel will measure (**2- √2)** from the soles of the feet to belly and (**√2 -1)** from navel to the head. In Yoga, Zen and Chinese meditation techniques this point corresponds to the transformative and generative aspects of the individual that involves **rooting** techniques for **self transmutation**. **So Lao Tzu said, “To seek the root is the goal”.**

- **√2 signifies the power of multiplicity. Hence the Generative aspect of this rhythm is attributed to √2. This represents the principle of transformation.**

- Remember **√2** is a proportion in the progression of squares and a square does not become a higher form. We have seen that the diagonal of the cube (sides of unit length) is **√3. It is this √3 which divides the volume form of a cube and the diagonal of the cube becomes a root for the higher shapes like pentagon, hexagon etc.**

- Hence **√3 signifies the formative power of the rhythm while √2 signifies the generative power of the nature.**

We will touch the same using **Vesica Piscis** in the next part which shall take us to **√5 and the golden mean Φ.** I am not going to touch upon important topics like Gnomic spirals and the relation between various progressions and hence I would request you to Google them if you are interested.

Happy reading!

PS:
There are many books on this topic and I would recommend **“Sacred Geometry – Philosophy & Practice” by Robert Lawlor** which I found easy to read and understand.

You tube and Internet is replete with tons of images, articles and videos on this topic and you can learn a great deal from them.

19. TIME = SPACE SERIES. PART 19 – SEELAM – VESICA PISCIS, √5, & GOLDEN MEAN (Φ)

If you don't know what is vesica piscis, please google and you would find it interesting. I am not going to explain it here and I would focus on the summary and aspects which are often left out. The philosophical aspect of Vesica Piscis, is a rotating point becomes as circle of unity. Two circle of unity (with its radius as unity) interacting / intersecting each other in a perfect manner offers an overlapping area which is considered to be sacred and fundamental to all the geometric forms in the universe. We will first see how **√2, √3 & √5** are part of this Vesica Piscis.

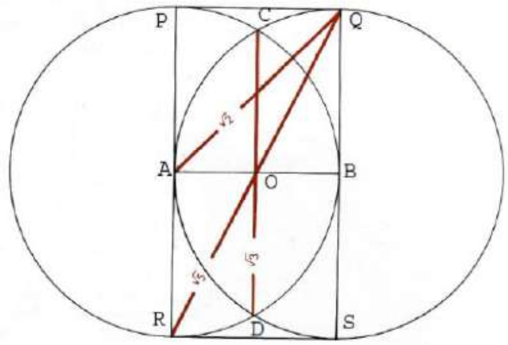

Figure 23; Vesica Piscis

Please look at 2 circles who radius is unity and AB = 1, intersecting as per the diagram. The overlapping area ABCD is the Vesica Piscis and we can see **√2, √3 & √5** all part of this Vesica Piscis.

ABPQ and ABRS are 2 squares and the side of the square is 1. **√5** is as significant as **√2 and √3.** In the diagram here **√5** is the diagonal of the rectangle PQRS / diagonal of 2 squares – one below the other.

√5 is the regenerative factor that binds 2 squares or that factor that transcends two worlds or the bonding factor for 2 cosmic entities. This **√5** shall take us to the most important ratio which is the **Golden mean** or the **God's ratio**.

But before that let me share what stuck my mind when I was engrossed **with Vesica Piscis and √3**. Look at the pictures below. These are pictures of forehead marks of Hindu Brahmins sects called Iyengars and their supreme deity is Narayana. I correlated the striking similarity of the rectangle around the Vesica piscis and Iyengar's namam. You just have to imagine the 2 overlapping circles are the right and left side of the face with the **Vesica Piscis** starting from the top tip of the nose. To me the **formative principle of √3 is in line with the "Bhootha krut" and "Vishwa karma manu stvashtha"** aspects of **Maha Vishnu** as per Vishnu sahasranamam. I had a good laugh on this correlation.

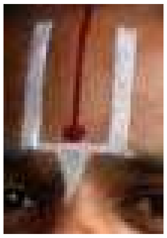

Figure 24; Iyengar Namam 1
(Courtesy: Internet)

Figure 25: Iyengar Namam 2
(Courtesy: Internet)

Coming to the **Golden mean,** If you have not heard about it, just Google or here is a quick link http://www.goldennumber.net/. Some of you might be even have fatigue repeatedly reading and listening about it, like I do. Hence I am not going to rewrite what is widely available. But let's look at the philosophical aspect of it, its relation to **$\sqrt{5}$** and its manifestation.

Now so far in **$\sqrt{2}$, $\sqrt{3}$ & $\sqrt{5}$** what we have seen is how Unity evolves as many. But if I am interested in geometrically dividing the Unity so that this proportion evolves as unity then how do I do it?

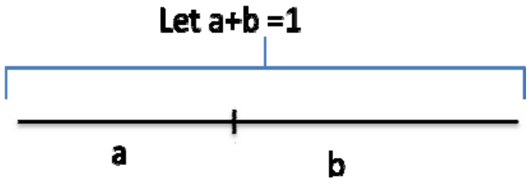

(Figure 26: Line andΦ)

Consider a line of unit length and also consider two segments a, b so that a+b =1.

A and B are such that it has to fulfill this condition of Proportionality that a/b : b/1.

This means what 'a" is to "b", is the same as what "b" is to the whole length which is 1.

Mathematically resolving this equation a/b : b/1, hence $b^2 = a$ and hence **$b = \sqrt{a}$.** This means **$a + \sqrt{a} = 1$.** The answer to this equation is only one value which is called the **Golden mean / God's ratio** denoted by $\Phi = 1.6180339\ldots$ and $a = 1/\Phi$. **Philosophically it is the division of unity and the only possible creative duality within the unity.**

You may ask as to why unity is not divided into 2 equal parts and why is should be **$1/\Phi$** (**0.6180339..).** This is effectively answered by our Upanishad **"Whether we know it or not, all things take their existence from that which perceives them"**. **Adi Shankara said "the universe is an illusion / maya" because it is what we perceive**. **We tend to ignore them as philosophy or superstition.**

But let's look at it scientifically as we understand. When the unity is divided into equal parts there is no asymmetry and difference and hence there is no perceptual universe. **An asymmetric division is needed to create the dynamics necessary for progression and extension of unity**.

Let's summarize our understanding again:

- **$\sqrt{2}$ – the generative component and a multiplying factor creates multiple squares from a unity of square.**
- **$\sqrt{3}$ – the formative component that creates multidimensional polygonal shapes from a cube.**
- **$\sqrt{5}$ – the regenerative component or the binding component that transcends and binds both the worlds.**
- **Φ - Golden mean or the God's ratio,** which is an evolutionary principle that is guided from within and an indisputable mathematical evidence of the conscious evolution of this Universe.
- Φ - Is a self-similar variable which can be used to derive Unity and it is the only proportion by which the Primordial rhythm maintains it perpetual oscillation to sustain this universe.
- Relation between **Φ and $\sqrt{5}$ is, $\Phi = (\sqrt{5}+1)/2$**

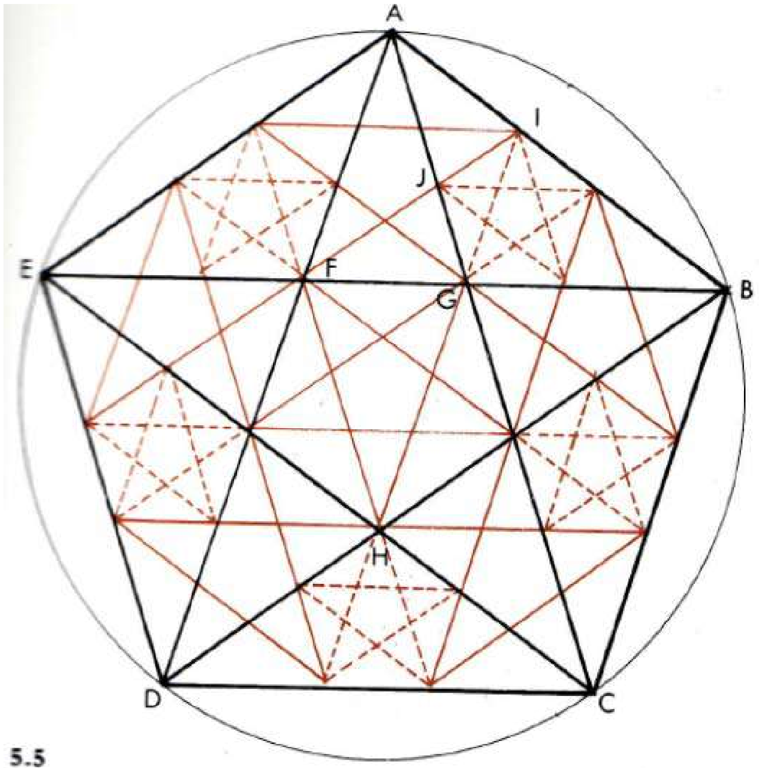

5.5

(Figure 27: Pentagon & Φ, Courtesy: Internet)

Now let's geometrically as to how Φ **and √5** are intricately linked. Consider the Pentagon picture whose side is of unit length. Ie.

AB = BC = CD = DE=AE = 1 and

EG = FB = CG = DF = 1 and

EB = Φ = 1.61803399 and BG = 1/Φ = 0.61803399 and hence all the ratios like AG = AF = EF = GB = 1/Φ.

Also FG = $1/\Phi^2$, which means **EB = EF+FG+GB =** $1/\Phi+1/\Phi+1/\Phi^2$

Also GI = $1/\Phi^2$ **and GJ =** $1/\Phi^3$

If you are interested more please check out the Fibonacci series and its relation to Φ. **The importance of 5 is dominant in every living organism while 6 and 8 are the characteristic of mineral and inanimate structures. It used to be rule of thumb that any plants / flowers which has 5 petals is an edible plant and hence 5 and pentagon are the symbols of life which contains the God's ratio or the golden mean.**

Now let's also understand how a **√5** is related to **√2 and √3.** If we consider a right angle triangle with the base of **√2** and height of **√3** or vice versa then the hypotenuse is **√5.** Now let me draw the connection between √2, √3, √5, Φ and π. If you think they are all some random numbers which cannot be fixed to their 10th decimal and not related to each other, just see the following relationship.

$$\pi = \Phi^2 \times 6/5 \qquad \text{or alternatively} \qquad \Phi = (\sqrt{\pi} * \sqrt{5}) / (\sqrt{2} * \sqrt{3})$$

This is the binding factor. I shall produce below some pictures which are aligned based on pentagon and the web is replete with them for you to understand more.

These 2 parts were little heavy and I shall move to easier aspects on the universal rhythm from next blog. The point I was trying to make is that everything in this universe is subject to a proportion and its fractals and let's see the manifestation of it in our daily life.

Happy reading!

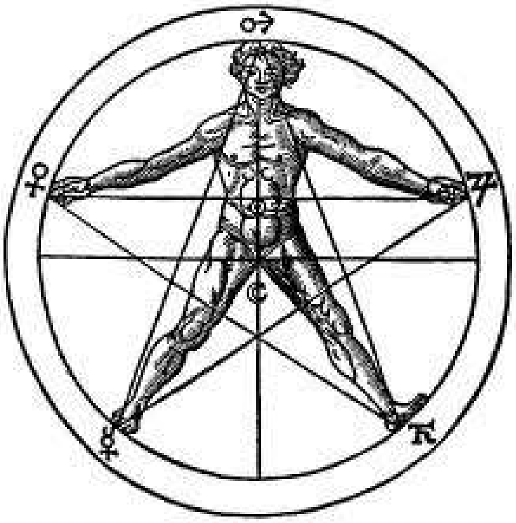

(Figure 28: Pentagon & Human shape, Courtesy: Internet)

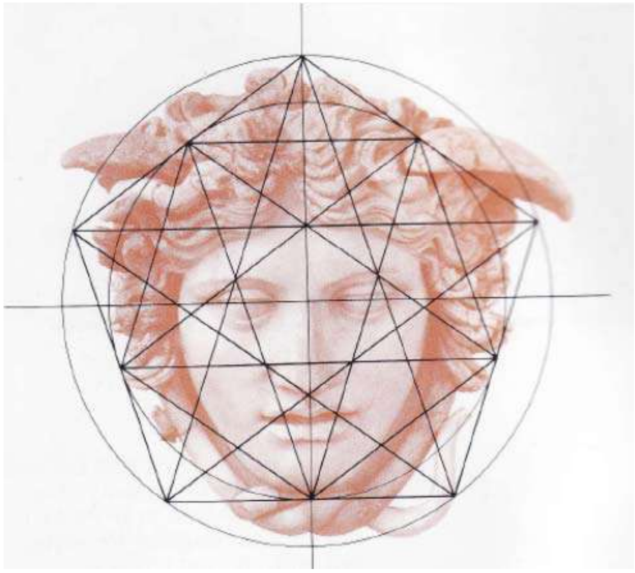

(Figure 29: Pentagon & Human face, Courtesy: Internet)

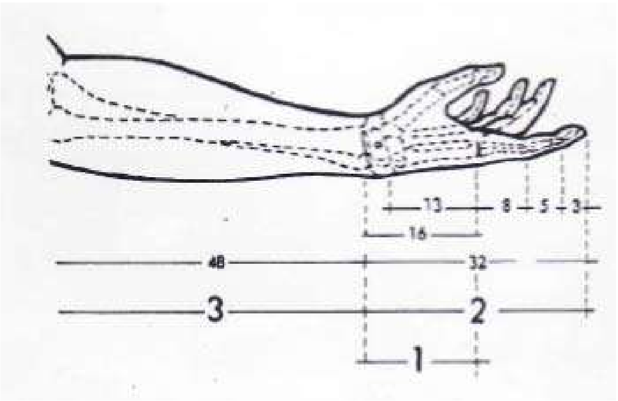

(Figure 30: Human hand and Fibonacci Series Courtesy: Internet)

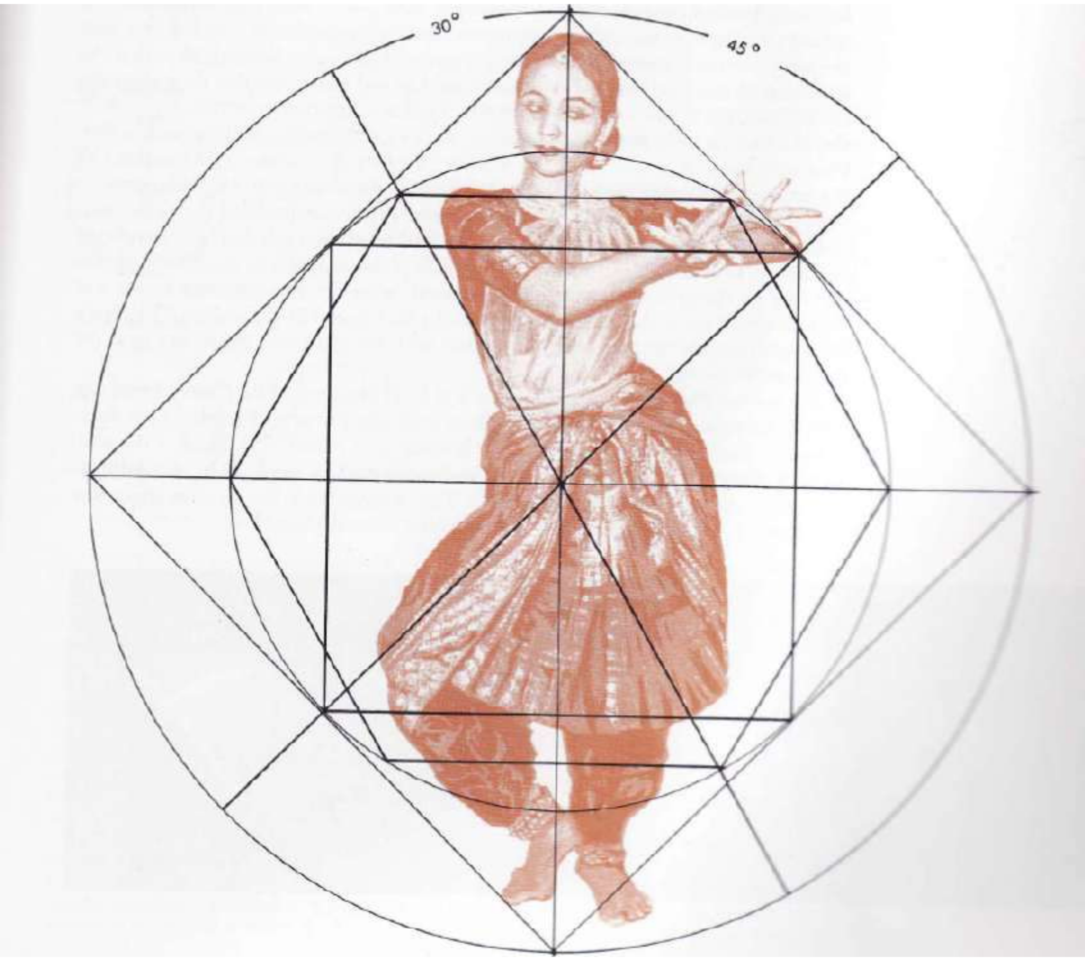

(Figure 31: Bharatanatyam and body alignment (Courtesy: Robert Lawlor, The Sacred Geometry)

The positions of Hindu classical dance (Bharat Natyam) describe geometric relationships from the axis of the body and the alignment of center of gravity at "hara" invoking rooting and divine power of transformation. In Hinduism "hara" is one of the names of Lord Shiva as in "Hara Hara Shakara"

20. TIME = SPACE SERIES. PART 20 – SANKHYA YOGA AND SAGE KAPILA

Now, let’s look at what **Sankhyakarika** one of the greatest scientific works by **Sage Kapila**. I need to highlight that the **Sankhya philosophy** is one of the most evolved philosophies about the secrets of nature; it contains only 72 verses of which 68 of them define the complete spectrum of universal manifestation. **This Sankhya Yoga is referred in Bhagavad Gita – Chapter 2 is called Sankhya Yoga and in Chapter 10, verse 26, Lord Krishna identifies himself with Sage Kapila**.

Sankhya in Sanskrit means numerals / numbers. Sage Kapila states:

- Any measurement is a relational process and only a change can be measured but it takes time. The difference between the yardstick and the measured object constitutes the measured variable, which in effect is an incremental change. Since the observer can detect or measure only a change then logically, **the incremental unit of measurement itself should form the basis to derive the whole. This is the genesis of Φ.**

- **Correlate the above statement with the equation $a + \sqrt{a} = 1$. The unity is derived from the proportion “a” or $1/\Phi$.**

- As an example, the arch of a bow has expansive stress on the outer side and compressive stresses on the inner side, while the string is in tension and all are balanced at every instant. The expansive, compressive and tensile stresses can be expressed by one law.

- Alternatively, a single variable describes the proportionality of three types of forces existing at the same time. **Such a variable should be scale invariant, self-similar and have an axiomatic relationship to the whole.**

- **So Sage Kapila called $1/\Phi$ as a self-similar variable from which the unity can be derived.**

- So he postulated if **x** is the value of a measured increment then the total value **$1+x$** must be related to the ratio of change as $1/x$ at the same instant of time or simultaneously. This means **$1+x = 1/x$**, and would lead us to **$X^2 + X = 1$**.

- He quotes that if the primordial rhythm has to be self sustaining then the proportion is certainly governed by $1/\Phi$. He gives the formula for this oscillations as follows:
    - **$X = 1/\Phi = 0.61803399$**
    - **Increment / Expansion = $1+x = 1.61803399 = \Phi$**
    - **Compression = $1-x = 0.38196601 = X^2$**
    - **Resonance = $X^3 = X - X^2 = 0.23606798$**

Resonance or X^3 is the factor which protects this rhythm from decaying and dying off. So **Φ or $1/\Phi$** is such an important ratio or proportion so it is called as the **Golden mean or God’s ratio**.

I shall highlight some of the factors / ratio which are part of Sankhya Yoga and what is amazing and you find formulas for Tamasic, Sattwic and Rajasic Guna states. (Please refer “Secret of Sankhya Yoga” – G. Srinivasan in the web).

- **Kapila defines the four qualities of the unmanifested state as Aikaantha (Synchronized), Aathyanta (perpetual), Atho (Dynamic)and Abhavath (unmanifest).**
- Taking C as cyclic rate of oscillation or vibration and x as rate of change per cycle then the three Guna self similar interactive states are shown as vibratory counts per cycle by the formula. **Thaama = C^{1+x} . Rajasic = C^{x+x} . Sathwa = C^{1-x}.**
- Some of the numbers that Sankhya Karika deals with are **1/7 (0.142857), π/10 (0.314159), 1/Φ (0.618034), √2 (1.414), Φ(1.61803399), √3 (1.732), √5 (2.23), e (2.718282), π (3.14159) and declares that the unmanifested state (Abhavaat) can exist only when the oscillations are between π/10 (0.314159) and e(2.718282). This is simply amazing since the state of Abhavvat is the state of Brahman.**

Despite herculean efforts I am unable to understand the Bhashya for the 72 verses which is about 500 pages explaining the secrets of Sankhya in any significant measure. I wish someone can step in and help us by explaining Sankhya Yoga that is comprehensible to our level of intelligence.

We have a treasure of secrets and richness hidden in our scriptures and it seems we have already lost most of them and the fundamental reason being our indifference to Sanskrit and our mother tongue.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

The passage produced below is just the summary of verse 1.

…The Sankhya view that evolves out of the complete theory is that only vibrations are detected by the observer... Since only relative changes can be be detected, any fundamental component that is not in a vibratory state cannot be detected. Vibrations are caused by components in an interactive state or a cyclic state of colliding and separating that is common to all interactions and wave phenomenon.

The interactions create three distinct modes of stresses. Collisions are compressive or inelastic. The resultant reaction is expansive or elastic. When the interacting components lack the freedom to move away the interactive state is maintained in a cyclic or shuttling mode in the same location in a resonant state.

This important Sutra lays emphasis that if the observed process of detection is dependant on vibrations, then fundamental space ***must*** *contain components with those four characteristics described as states, which makes it function in a holographic way. That is all vibrations remain in fixed relationship relative to each vibratory point in a coherent and resonant state.*

உ

21. TIME = SPACE SERIES. PART 21 – 5 AND 8 THE UNIVERSAL RHYTHM

Let's summarize what we have seen so far and then move on to more interesting aspects.

We have touched upon the God's particle and its characteristics. We also looked at the 5 fold manifestation process and in that we looked at how **Moolam** becomes **Kaalam** (time) and how time rhythmically (**Seelam**) vibrates with a proportion and creates forms and shapes (**Kolam**). We also touched upon some of the ways we had codified the secrets of nature like **Lord Shiva, Lord Nataraja, Lord Ganapati etc.**

Before we see how rhythm **(Seelam) and shapes (Kolam)** manifests itself as universe, let's first look at one specific word from Tamil which has three significant meanings converging into one, Yen (**எண்**), means number **eight, generic term for Number / numeral** and also root for **எண்ணம்** which means thought. I shall explain the significance of this below.

Importance of 8 / Octet:

It is stated that these God's particles combine in the multiples of **8 / octaves ONLY** and the states of manifestation from being subtle to gross is as given below.

○ **Formless Un-manifest State**	**0 (This is the Abhavaat state)**
○ **Guna State**	**$8^0 = 1$**
○ **First stage**	**$8^1 = 8$**
○ **Second stage**	**$8^2 = 64$**
○ **Third stage**	**$8^3 = 512$**
○ **Fourth Stage**	**$8^4 = 4096$**
○ **Fifth Stage**	**$8^5 = 32768$. This is the state with manifest form.**

This rule is applicable to both the visual and aural forms which mean we can see or hear anything only in the fifth stage of evolution. So as per **Pranava veda, the fundamental numbers are 0,1** and the key **numerals for proportions are 5, 8**.

Significance of 5 and 8:

Pranava Veda describes that 8 / octets is the essential harmonic oscillator for the manifestation of the universe since this generative force is the multiple with which the whole universe manifests and sustains. You many note that number 8 is associated with Lord Narayana and his Ashtakshara mantra **(Om NaMoNaRaYaNaYa**).

The principle of 5 / Penta, we have seen is the regenerative binding force earlier. It is always a 5 stage process. This is associated with **Lord Shiva (Om NaMaSiVaYa).** The un-manifest form comes to the Guna state with the force of 5 and in the multiples of 8. It is also important to highlight that 5 and 8 are part of the **Fibonacci series and their proportion is 8/5 = 1.6, which is Φ**.

உ

But it is important to realize that the building of octets is a 5 step process for manifestation. This highlights that one cannot exist without another and both are interdependent. Now let’s recall:

"Shivasya hridayam vishnur, Vishnoscha hridayam shivah:" and
"Shivaya Vishnu rupaya Vishnave Shiva rupine"

Vishnu (8) is the heart of Shiva (5) and likewise Shiva is the heart of Vishnu and they both are the representation of each other.

If you have noted Tamil language stressed the importance of words and its meaning – this was addressed in the first few parts of this series. Here Tamil language gives the importance of numerals and its significance in the formation of universe. Both the great sages **Auvaiyaar** and **Thiruvallular** stated the importance of numerals and letters in an identical fashion. Both these lines mean, Numerals and letters are like eyes for living beings / equal to eyes.

எண் என்ப ஏனை எழுத்து என்ப இவ்விரண்டும் கண் என்ப வாழும் உயிர்க்கு!
எண்ணும் எழுத்தும் கண்ணென தகும்!

Now what should be noted here is that (**எண்**) importance of numerals precedes that of letters and we know why. The aural and visual forces become letters & its associated sound at the 5th stage but the numerals are from the Guna state. **Numerals were associated with 8 / Lord Vishnu and letters were associated with 5 / Lord Shiva**. Guess you can get the scientific knowledge and significance we have in our languages.

Since what is in macro is there in microcosm, taking the analogy to microcosm our mind becomes the Moolam (Source), our thoughts are the vibration, time is measured as the periodicity between thoughts, based on the rhythm of our vibration we perceive the world. This is given very nicely as **உள்ளமே மூலமாகி உணர்வுறும் கோலமாகி**! means **Our mind is the source and its awareness becomes forms.**

This wonderful concept is what is linked in the word Yen (**எண்**), which means number **eight, generic term for Number / numeral** and also root for **எண்ணம்** thoughts. **Just this one word in Tamil highlights the supreme secret that thoughts become form and words with the help of numbers and numerical proportion converts aural and visual forms to thoughts.**

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

22. TIME = SPACE SERIES. PART 22 – EIGHT (OCTET) AND THE 5 SACRED ARTS

எட்டே எட்டியல் எட்டெட் டியலே
எட்டின் நெறியே அளவை மாநெறி !!

The above lines from the work **Aintiram** means **"Eight / Octa (multiples of 8) is the fundamental numeral for any measurements**. We did see that the fundamental god's particle add up with each other in the multiples of 8 in the previous part.

In ancient India there were about 64 arts (http://en.wikipedia.org/wiki/Kal%C4%81) known as **Chatuhsasti kala ("ஆய கலைகள் அறுபத்துநான்கினையும்"** as popularly known in Tamil), but I am going to talk about 5 sacred arts.

We will look at the importance of numeral 8 and how this is fundamental to everything in this Universe. But let's again start again with etymology. The word arts in **Sanskrit** is **"Kala" (कला)** and **Time in Sanskrit is काला (Kaala).** So etymologically we can say that classical arts are all about variation of time (or harmonics of this pulsation of time in the multiples of 8).

After etymology let's get to the basics. Our understanding of the universe is that **"the primordial space manifests itself as spatial forms through pulsation (Time) and its periodicity (Rhythm & order).** This evolution started with **Pranava which came out as aural and visual form (Sound & light)**. When they say **"The one became two"** they refer to the Pranava which is light and sound from the unmanifested consciousness.

Aintiram states that the dual form of **Pranava** gave forth 5 sacred arts which are fundamental to most of the other arts. **The Aural form of Pranava gave forth Poetry and Classical Carnatic music, which can only be heard. The Visual form of Pranava gave forth Sculpture and Building architecture, which can only be seen. Both the aural and visual forms merged to create the classical dance (Bharatnatyam**), which involves both seeing and hearing.

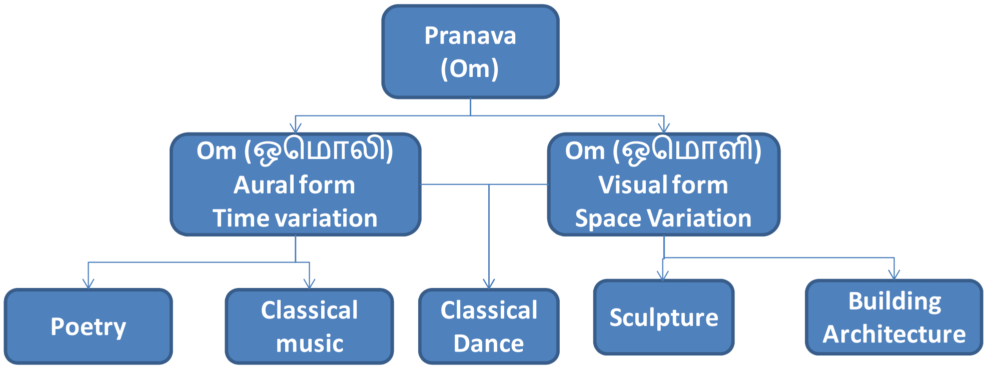

Figure 32: 5 Sacred arts

உ

They were classified as 5 vedas:

- **Poetry – Sabda veda**
- **Classical music – Gandharva Veda**
- **Classical Dance – Naatya veda**
- **Sculpture – Sthapatya Veda *(Sthapati is a clan that is the master of this veda)***
- **Architecture – Pranava Veda**

Pendatic Tamil and its construction:

Pendatic Tamil is called as **ஐந்தமிழ் (Aintamizh**) which consists all the five above arts, **இயல், இசை, நடனம், சிற்பம், கட்டிட கலை**. It is these 5 arts which were formed out of the Tamil language and the language itself is created from the word Om. They say that it is not 12 basic vowels in Tamil but only 5. They are அ, இ, உ, எ, ஒ. These 5 letters combine to form the rest of the letters like

அ + அ = ஆ, இ + இ = ஈ, உ + உ = ஊ, எ + எ = ஏ, ஒ + ஒ = ஓ, அ + இ = ஐ,

அ + உ = ஔ

Here you may want to note that 5 has become 12. Same way in classical music they say that the basic svaras are 5 (ச, ரி, க, ப, த) – Sa, Ri, Ga, Pa ,Da which became 7 and then 12.

So 5 is the fundamental numeral for evolution and Tamil language has grown based on this philosophy and science. I am sure some of the other Indian languages have the same basics. The book which describes this phenomenon of evolution of language from Pranava and arts from language is **Aintiram (it means 5 works) (ஐந்திறம்**) authored by Mamuni Mayan dated 10000 BC.

You would agree that the aural forms are predominantly variations in time and visual forms are of spatial variations. But our ears stand for akash tatva ("space") and it can perceive variation of time and space and we will see this in detail. When we said **"Time = Space"** we will see that both of them are governed by the same laws and measurements and hence there is no difference between **Time and Space**.

Happy reading!

உ

23. TIME = SPACE SERIES. PART 23 – NANOTECHNOLOGY IN ANCIENT INDIA & SIZE OF GOD'S PARTICLE

Let me start with these questions:

- **What is the size of the God's particle and how is it calculated?**
- **Does nature endow a standard measure for an average human being's height?**

Before we proceed to the next section I want to give the very detailed space calculation table used in ancient India. Our rishis did not stop just with telling us that the **Paramaanu / God's particle** is very small but went ahead and established its size.

If you have to see such minute particles then your sensory organs – Eye and mind should be tuned to it so that it can see things at that size. From the formula available in our ancient texts I was trying to reconstruct by working backwards to find out the size of the **Paramaanu / God's particle should be.** Please refer the table below which I have arrived at.

ANCIENT SPACIAL MEASUREMENTS		
Unit scale	**Ancient metric**	**Current day Metrics**
1 Paramaanu =	Smallest particle	133 nano meters
1 Ther Thugil=	8 Paramaanu	1065 nano meters
1 Mayir nuni=	8 Ther Thugil	8526 nano meters
1 Eer=	8 Mayir nuni	68 Micro meters
1 Paen=	8 Eer	545 Micro meters
1 Yavai=	8 Paen	0.4365 cms
1 Angula=	8 yavai	3.4925 cms
1 Tala=	6 Angula	8.25 inches
1 Muzham=	24 Angula	33 inches
1 Muzham=	4 Tala	33 inches
2 Muzham=	8 tala	66 inches
Universal height of man	9 Tala	74.25 Inches

(Figure 33: Ancient Spatial measurements. Courtesy – Dr. Ganapathi Sthapati)

I have given the exact word used in Tamil for these measurements instead of translating it in English. Ther Thugil is the dust from chariot wheel, Mayir nuni is the width of a hair tip, Eer and Paen are the lice eggs and lice. Some of you may jump that most of the units were subjective and would have variations in normal life. I agree but there is a reference point – **Angula** and its exact measure based on which other values are derived backwards to arrive at the value of Paramaanu.

Now if you note the **smallest particle** aka **God's particle / Paramaanu** is **133 nm** in size which is currently under study by the CERN scientists. Nanotechnologists today confirm that the width of a hair strand is about 25000 nm and our ancient texts give the size at the tip of the hair strand (and not the width) as 8526 nm. I see that to be a reasonable estimate when no such precision instruments were available in those days.

One Angula is 11/8 of an inch. This is the fundamental measure which is used in any of the worldly measurements. Now you can see that 6 Angula is 1 Tala which is about 8.25 inches and then comes the most important revelation. **Does nature endow a measure for an average human being's height?**

The answer is yes. **As I said an average human being should be 9 Tala's height and hence the nature endowed height for any human being is 74.25 inches.** This is the ideal height and the variation is attributed to our Karma, hereditary and lifestyle.

Now let me give few tips to find out your height from typical measurements:

- Measure the length of the face (from forehead to chin). Your total height should be 9 times that height. This is the principle of Nava tala. As an extension measure your thigh or legs length excluding the knee and feet and your total height should be 4.5 times that.

- Take your right hand middle finger (No. Don't show it to anyone☺). You would notice that the finger is divided into 3 parts. Measure the length of the lowest part – from the end of the palm to the first subsection. Your height typically should be 54 times this measure. Since this measure for an average man with nature endowed length shall be 1 Angula. Hence your height shall be 54 times that measure.

- The length of the thumb from the base of the wrist should be about 3 Angula and hence 18 times of that should be your height. This can go on and on.

So, to sum up this section:

- Instead of looking them as Rishis or Sages or religious figures if we look at them as scientists in the field of nanotechnology, quantum physics and Astro dynamics we would accept most of the secrets without any prejudice.

- Not just the measure for the height of the man but they had formula for every object and this was the secret of their wisdom. For any astronomical object, they had formulas to measure body's core (width) and also the width of the mantle. The roots of their wisdom lie in understanding that "many' are just a harmonic variation of the "one" subject to a mathematical proportion. Hence there was no “rocket science” in those days and all of them were as simple as it looks like.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

24. TIME = SPACE SERIES. PART 24 – OCTET, OCTAVE AND THE GANDHARVA VEDA (CLASSICAL MUSIC)

Let’s start with classical music where Octaves form the fundamental for it. There are seven svaras (Sa, Ri, Ga, Ma, Pa, Da, Ni) which is made into an octave with another Sa after Ni. So it becomes Sa, Ri, Ga, Ma, Pa, Da, Ni, Sa (*Do Re Mi Fa Sol La Ti Do*). The first rule of classical music is “**Shruthi Mata, Layam Pita**”, which means **Shruthi is the mother and this corresponds to the frequency or the spatial component and Laya is the Tala component which is a function of time compared as father**.

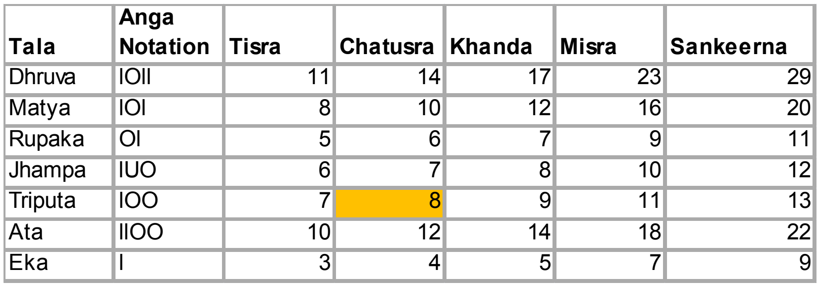

Tala	Anga Notation	Tisra	Chatusra	Khanda	Misra	Sankeerna
Dhruva	IOII	11	14	17	23	29
Matya	IOI	8	10	12	16	20
Rupaka	OI	5	6	7	9	11
Jhampa	IUO	6	7	8	10	12
Triputa	IOO	7	8	9	11	13
Ata	IIOO	10	12	14	18	22
Eka	I	3	4	5	7	9

(Figure 34: 35 types of Tala. Courtesy: Wikipedia, internet)

There are 7 fundamental Talas in carnatic music and each Tala has 5 different variation based on Jathi so in all there are 35 different Talas as given in the tabulation. These Talas are measured as the “matras” the unit duration of time. There is another concept called Nadai or Gati which we will not touch upon now. But let’s look at these 35 variations and can you tell me where the most famous “Adi Tala” is in this 35? Most importantly why it is also called as **“Adi Tala” – which is the primordial rhythm**?

The answer is **“Chatusra Jathi Triputa tala”** highlighted in the table is called as **“Adi Tala”** and it is called so since it gave the most primordial rhythm with 8 time units. But there are couples of more cells where we can see “8” in the table and why these are not called ‘**Adi Tala**”?

The primordial rhythm has a pattern of **1, 3, 2 and 2** which adds to 8. Further on we will see just the variation of this Adi Tala in anything and everything. Tisra Jati matya Thala has a structure of 3, 2, 3 and Kanda Jati jhampa Tala has a structure of 5,1,2, hence they are not classified as the primordial rhythm.

Let’s quickly touch upon the spatial variation or the frequency / pitch aspect of this divine art before we branch off to other topics. The image below gives the 16 levels of Savaras which includes tones and sub tones. The seven svaras become 22 tones based on the frequency variation / harmonics. Here 7 are fundamental svaras and 16 as given below is the most practiced variation and 22 is not popular.

உ

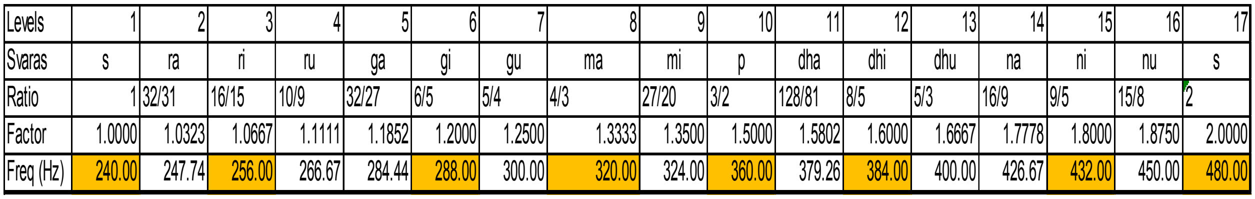

Levels	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Svaras	s	ra	ri	ru	ga	gi	gu	ma	mi	p	dha	dhi	dhu	na	ni	nu	s
Ratio	1	32/31	16/15	10/9	32/27	6/5	5/4	4/3	27/20	3/2	128/81	8/5	5/3	16/9	9/5	15/8	2
Factor	1.0000	1.0323	1.0667	1.1111	1.1852	1.2000	1.2500	1.3333	1.3500	1.5000	1.5802	1.6000	1.6667	1.7778	1.8000	1.8750	2.0000
Freq (Hz)	240.00	247.74	256.00	266.67	284.44	288.00	300.00	320.00	324.00	360.00	379.26	384.00	400.00	426.67	432.00	450.00	480.00

(Figure 35: Svaras and frequencies)

You can note that Sa starts with 240 Hz and the last Sa is 480 Hz doubling the frequency and hence an octave. While Sa has 240 hz and 480Hz levels, the tone Pa (P) or Sol does not have any variation and this is at 360Hz. I am going to correlate the Mayan studies with Carnatic music and instead of seeing the whole spectrum from “Sa”, I am going to see this spectrum from “Pa”.

You would note that “Pa” at 360 Hz corresponds to the **Harmonic osciallator constant number** and the whole spectrum is almost like a Sine wave from 360Hz to 480 hz and falling down to 240 Hz with a uniform gap of 120Hz above and below. All the tones would fall within this 240Hz centered around 360 hz. So the spatial variation in carnatic music can be summed as “It originates at a value (240Hz) and varies by the same measure (240Hz) centered on the harmonic oscillator”. Since the objective is about looking at the primordial rhythm we will not touch any other aspect on this topic.

We will see how this **Adi Tala** is the fundamental pattern / rhythm that manifests in every other art.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

25. TIME = SPACE SERIES. PART 25 – ADI TALA, NAVA TALA AND STHAPATYA VEDA (SCULPTURE)

Let's look at how Octet and the primordial rhythm (Adi Tala) that defines Sculpture. Let's recap some of the following:

- The process of evolution starts with Pranava which has aural and visual forms. We have also seen that anything subtle is represented in a 8x8 energy grid and anything gross is in 9x9 grid.
- The fundamental measure with which the particles bond is **Eight or** it's multiple and hence the primordial rhythm is an **Octet.**
- The fundamental measure of time is **Tala** and the **primordial Tala is Adi tala** with 8 units and a pattern of 1, 3, 2 and 2.

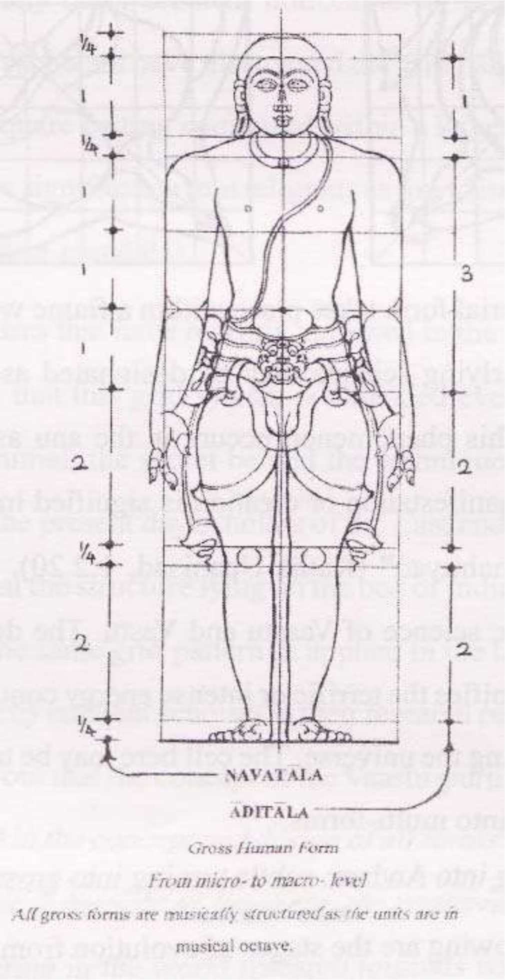
1/4
1
1/4
1
1
1
2
1/4
2
1/4
1
3
2
2
NAVATALA
ADITALA
Gross Human Form
From micro- to macro- level
All gross forms are musically structured as the units are in musical octave.

(Figure 36: Human form – Adi Tala, Nava Tala Courtesy: Dr. Ganpathi Sthapati)

Now let’s see the measurement rules as to how to make a sculpture of a human being.

- **The overall human height is measured as 9 talas.** Yes the same time measurement is used to measure height. Pinch yourself now, since height is a space variation and not a time variation yet it is measured with a metric of time. This is a secret which is unknown to many and even some of the experts in this field. This is where Time and spaces merges without any difference and this is the fundamental rule with which we have been creating visual and aural structures in the universe.

- **Divide a man’s height as 9 units of which head, torso, thigh and legs would fall into the pattern of 1, 3, 2 and 2 which adds up to 8. This is the primordial rhythm – the Adi Tala now as a human form.**
    - Then how 8 became 9? No other branch of Vedas or science tells us this secret so explicitly to my knowledge. **The following parts, Top of the head to forehead, Neck, Knee and feet each measure 0.25 talas and hence add 1 additional tala to 8 and it becomes 9.** So Adi tala which is a subtle version of human becomes Nava Tala which is the gross version of human being. In the process of 8 becoming nine all the joints are formed in the body☺.

    - Let me add one more dimension to this. Our human body has a gross body (Sthula) and a subtle body (Sookshma sarira). **The Sthapatya veda tells us that the gross body is nava tala and sookshma sarira is Adi Tala**.

    - I fell off the chair when I understood this and the linkages. Check if you are still on your chair and maybe you should sit on the floor from now on☺.

Martial Arts and Lord Nataraja’s stance:

Let me ask you the following questions here:

- Look at the image of the Lord Nataraja below. If a gross image should have Nava Tala which is a 9x9 grid then why the human image of Lord Nataraja which is a Nava Tala fitted into a 8x8 square instead of a 9x9 square?

- Both in Classical dance and in martial arts why do they ask you to stand in a posture where the knee is slightly or significantly bent. Not all of the stances but most of the basic stances correspond to this why? Let me know if you have received a satisfactory answer for this question.

உ

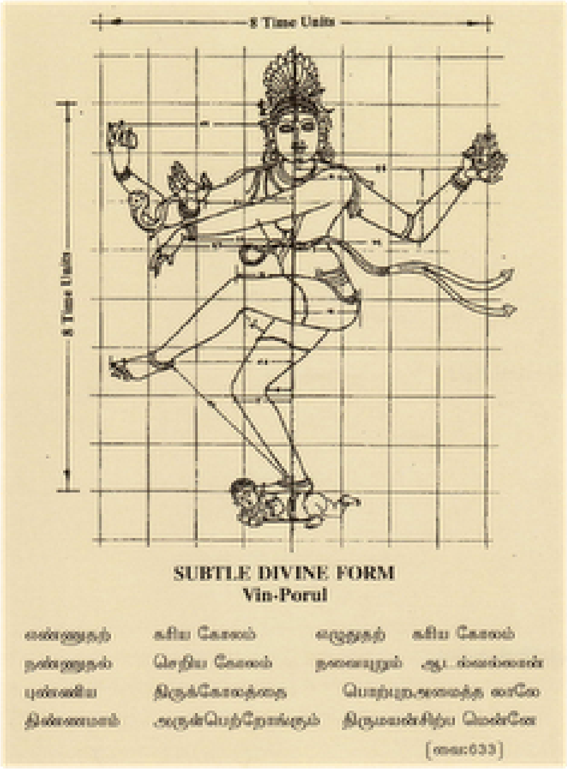

(Figure 5 reproduced)

- The answer to this question reveals one more secret to us.
    a. **In Sthapatya veda each grid line is divided into 10 or 11 or 12 (usually 12) and hence a 9x9 grid has 9*12 = 108 blocks or levels of energy. This may explain why 108 is a sacred number from an energy perspective.** This covers the entire spectrum.
    b. When we create a gross image its always sculpted to 9 talas but when they sculpt image of the gods and in this case Nataraja, they fit the gross image to a 8x8 energy grid to invoke the subtle energies.
    c. How do they fit a 9 tala to 8 tala? The technique is not to change the proportion of nava tala but they ensure that 9 fits into 8 by making the image to bend it legs such that the height reduces by one unit.

- You can notice that Lord Nataraja's right leg is bent to fit the image in 8x8 grid. This is believed to invoke the subtle energies corresponding to that image.
- This is the same reason why martial arts and in Bharatnatyam your teacher asks you to take a basic stance which reduces the height from 9 units to 8 to invoke subtle energies. Of course the other reason is the alignment and focus to Hara / Dantein with that pose.

If I tell you that the science of nature and the associated philosophy codified in Hindu religious symbols are unmatched and amazing then I hope you would whole heartedly agree now. These religious texts and symbols offer the key to unlock the secrets of nature.

So to sum up, so far we have seen one variation of time as **Gandharva veda** that manifests as classical music and the primordial rhythm is **Adi Tala.** Next we saw that the **Sthapatya veda** in the form of sculptures offered the secret of space variation in **Adi Tala and Nava Tala**.

This is where Time and Space were equated with the same measure yet offered completely different arts.

The whole universe has these underlying and un-manifested interconnections and hence the **Hindu philosophy of "That one became many"…** is the supreme truth. Seeing the divinity in everything is the only truthful way of recognizing this science.

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Happy reading!

26. TIME = SPACE SERIES. PART 26 – SABDA VEDA, CHANDAS & POETRY (MA NISHAADA…)

शुक्लाम्बरधरं विष्णुं

शशिवर्णं चतुर्भुजम् ।

प्रसन्नवदनं ध्यायेत्

सर्वविघ्नोपशान्तये ॥

Shuklaambaradharam Vissnnum Shashivarnnam Caturbhujam |
Prasannavadanam Dhyaayet Sarvavighnopashaantaye ||

Let's start with this Sloka as a sample to see how 8 and its harmonics are integrated with Poetry. In this sloka there are 4 lines (a line is called Pada and Pada means feet in Sanskrit) and each line has 8 Sanskrit Aksharas (Only vowels and vowels with consonants imposed on them should be counted and pure consonants are not part of the count.) This is usually written in 2 lines as shown in English.

This is the most popular way of creating a sloka or poem in most of the Indian languages. This "metre" resonates as 32 aksharas in 4 padas. You may want to correlate this with 1,3,2,2 where the last 2 stands for your feet. So feet (pada) measures 1/4th of the overall length of the subtle body (8 units) and 4 padas create the full blown subtle body. So such a constructed poem is a throbbing and vibrating energy body or a living organism and hence carries the power of the words it contains.

If all the padas have equal metre say 8 units then it is called **Sama Vrtta,** and if all are different it is called **Vi-Sama Vrtta** (not equal). **This 8 and 32 is the same as Adi Tala in carnatic music and ashta tala in sculpture. Can you make the connection now?**

In Sanskrit this metre is called as "**Chandas" and in Tamil it is called "Chandam**". **Chandas** is also one of the 4 important Vendagas and is called as the feet of the Vedas. **Let me ask you as to which language the Vedas are available in its primordial form?** The most popular and incorrect answer is Sanskrit. This is the level of awareness we have. Lets quickly look at these important points about Vedas and I shall correlate this with our topic of discussion.

- Unlike other faiths Hindus don't attribute the Vedas as word of god or given by some prophet. We consider Vedas as the breath of god (**"Nishwasitam"**) and not the word which means, Vedas are not created by god, but the breath. If there is no breath that person would not leave and hence it is associated with the very presence of god.

- Vedas are also **Apourusheya** which means it is not created by man and **Anaadi**, that does not have a beginning.

- But you can ignore them stating these are blind beliefs. Let's look at how this correlated with the science. Vedas are created by Pranava and these should be considered as the harmonics of the primordial vibration and nothing more.

  - If it is pure vibration then the “which language it is written?” sounds stupid isn’t. If it’s pure vibration it has to be expressed in “Hz or wavelength” equivalent in ancient India. This is what **Chandas or metre** is all about. This measures pure vibration and anyone capable of tuning themselves to those frequencies have **SEEN** the Vedas and not **created them.** Hence all our rishis are **Mantra Drishtas and not Karthas**.

Difference between Vedas and Poetry:

Vedas contains both **poetic verses (Padya) and prose (Gadya)**. All the poetic verses are called **Chandas** and it complies with the rule as stated above. The difference between normal poetry and Vedas are, Vedas contain tonal variations (high pitch and low pitch) apart from the time, space variations but in Poetry there are no tonal variations. What has tonal variations and are not part of Vedas are called Slokas. So to sum up:

- Vedas are both in Padya and Gadya. What is in Padya is called chandas and it has tonal variations in addition to the time space variations.
- What has tonal variations but not part of Vedas is called as slokas.
- What does not have tonal variation but comply with the chandas rules are called Poetry.

Harmonics and types in Chandas:

We saw every line to have 8 matras / units / aksharas. There are variations and these are the harmonics available.

- **Gayatri –** This is a very special chandas with the mantra where instead of 4 padas, there are only 3 padas and hence it is called Tripada Gayatri. Each pada has 8 Aksharas and hence 24 Aksharas in Gayatri Mantra. This is also the only chandas where a mantra is named after the chandas. Some people write it as 6 letters per line and 4 padas which makes it 24.
- **Ushnik Chandas –** 7 letters per pada and 28 letters in all.
- **Anushtub Chandas –** This is the fundamental and popular one with 8 matras per pada and 32 in all. Ramayana the first poetry was set in this metre.
- **Brihatee Chandas -** 9 letters per pada and 36 letters in all
- **Pangti Chandas -** 10 letters per pada and 40 letters in all
- **Trishtup Chandas -** 11 letters per pada and 44 letters in all
- **Jagti Chandas -** 12 letters per pada and 48 letters in all. This has a subtype called bhujangam which is split as 6-6 and that moves like a snake. (Ex. Subramanya Bhujangam)
- **Shikarini Chandas -** 17 letters per pada and 68 letters in all. In Soundarya Lahiri it is split as 6 and 11 by Adi Shankara.
- **Udkriti Chandas –** 26 letters per pada and 104 letters in all.
- **Dhandakam – above 26 per pada**. **Ex.** Garuda Dandakam by Vedantha Diskshitar.

The reason I have given all the above is to emphasis that nothing in our religious scriptures, literatures are without a grammar, order, form or proportion. Everything confirms to a mathematical proportion.

Just like **Vedas are the breath of god, Chandas is the breath of every mantra.** The mantra is powerful only if it is recited to this metre. This is the reasons some of the poem even if it is not related to religion stands the test of time. Also if you recite a mantra even without knowing its meaning but as per the chandas then the mantra shall bear its fruit.

“Ma Nishada” …..Origin of Poetry:

Would you believe that the first poetry in this universe started with a curse?

Sage Valmiki saw a hunter who killed one of the birds of a pair which were in love. The sage got enraged and cursed the hunter, ***O, Hunter, may you not fare well anytime as you killed one of pair of the Krouncha birds which were happily engaged in love”.***

mA niShAda pratiShThA.n tvamagamaH shAshvatI.n samAH |
yatkrauJNchamithunAdekamavadhIH kAmamohitam.h || (bAlakANDa 2.14)

The above sloka also meant ***“O Lord of Lakshmi, it will bring you eternal glory for having killed a male of a happy couple who lost his head completely in lust..”*** This male is Ravana who lost his head in lust despite being happily married with his wife Mandodari. The other interesting connection I have this sloka with this series is that the **Mamuni Mayan** is the father of **queen Mandodari** and father in law of demon king Ravana.

Once the sage realized that this curse also meant something very auspicious and referred to Rama avatar he wrote Ramayana which is the origin of poetry as known in the world and it is called **Adi Kaavya.** Valmiki Ramayana set in **Anushtub Chandas** which is an octet.

So if you look at both the poetry and classical music composition:

- The rhythm in which it is set is a pulsation of time which confirms to a divine proportion.
- The raga and the pitch variation is a pulsation of space / frequency which again confirms to a divine proportion. We did not discuss about ragas though.
- The words used are divine and we have seen the divinity of the letters and words.
- The meaning of the song / poem usually exhorts the qualities of the god and hence is divine.
- The classical music and poetry invoke the subtle divine energy in the aural form just as much the sculpture (when it confirms to the divine proportion) invokes the subtle energy from in its visual form.
- This is how we see that “**one became many”** and **binding factors** in this process of evolution being the numerals 5 and 8.

Alas! Today, we are the fanatic fans of musicians who lacks morality & humility, who is ignorant of any divine proportion, most of the music we hear is a cacophony and very mechanical, the words are indecipherable and lacks depth, the meaning of the song - if at all any are just suggestive of sexual emotions and nothing else and the dance movements we see are the gyrations of the hip with a great blend of fitness training exercises associated with it. This is not to negate some of the fine works we see even today but to stress that we had such high science and taste which we seems to be losing / lost for no good reason other than ignorance.

உ

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

27. Time = Space Series. Part 27 – Architecture – The Zenith of mathematics (Vaastu)

Just look back in history and see what stands the test of time. Even if a civilization is completely extinct what stand beyond their time is the architecture of the civilization which stands tall as a witness to the richness and maturity of the civilization.

Aintiram states that **"Architecture (Space engineering) is the zenith of the mathematics**" in **"கட்டிட கலையே கணக்கியல் உச்சி**..." This space science is called as Vaastu Sastra in India.

There are few important concepts in Vaastu science which I want to present as below:

- The whole world is made up of **God's particle (Vastu, வஸ்து**) and nothing else. It is this God's particle which actually manifests itself as gross forms (**Vaastu, வாஸ்து**). You would note that Vastu (Subtle) has become Vaastu (gross) by addition of letter "a" and this difference is reflected in our languages too. **When அ / अ become ஆ / आ the subtle becomes gross**. Hence the most important statement for this science is "**Vastur eva Vaastu**", this means **the subtle becomes gross or energy becomes matter or the space becomes earth.** This is purely manifestation / evolution and there is no creation. This has nothing to do with religion but supreme science.

- Since it is all only one energy and it ever vibrates based on the harmonics and resonance, this vibration is called life and everything in this world, be it a living being or a non-living being as per our current definition – everything vibrates. The chair you sit and the vehicle you use are all living beings. **Everything throbs with energy and there is nothing which is a non-living being in this universe.** Without this vibration there would not be any difference between iron and cotton. The difference between particles is essentially the molecular constitution which is a function of its elements and its periodicity. This defines presence or absence of quality and sensory perceptions. This was put very nicely by **Poet Subramanya Bharathi** as எங்கு காணினும் சக்தியடா, **Where ever you see its just energy.**

- The above concept is very profound and this is called as **Science in our scriptures since it deals with Knowledge of the space** (விண் ஞானம், Vin-gyan).

- By the same concept when a subtle energy which is in the open space is converted to a gross form like a building then this body / building vibrates based on the design specifications and the material used for that building. It does not matter what is the measure and what is the material used, it vibrates. **When a structure is created with a divine proportion then we breathe life**

into it. Thus the structure becomes a living organism. This is how some of the structure stands the test of time for thousands of years.

- **Vaastu is that science which aligns the vibration of the building with the vibration of the dweller in the building so that both the living organisms (the building and the dweller) are in harmony and resonance.**

- It sounds simple but for this science we need to know includes the vibration pattern of the dweller, mathematics of space engineering, the knowledge of materials and its nature so that we can match it up to create harmony. This is precisely what Vaastu as a science does. Again it has nothing to do with religion. If you Google you would find thousands of Vaastu experts outside India who perceive and pursue this as a supreme spatial science.

- Let us look at the word **"vas"** which is the root. This means to shine; to grow bright, **to bestow by shining upon**, becoming light. **Vaas** is its derivative which means to perfume, an intoxicant, **dwelling place**, to assume the appearance of matter. This is the root for **SriniVAAS**, **SreeVAASan**, **VAASam** in Tamil means living. Most importantly the ability to feel this subtle thing (which cannot be touched) is called **VAASanai (**வாசனை, **in Tamil it means smell).**

- This spatial science includes knowledge on the following:
  - Nature of soil and material, and the qualities of its energy.
  - Measures for spatial engineering that best resonates with the environment and the dweller.
    - Note that the dwelling unit cannot be compatible to all the human beings universally. This varies with individual to individual and hence the Vaastu compatibility for the house is always seen for the housewife than the man with the assumption that what is compatible to the lady of the house would suit all others.

  - Accurate calculation of time and the position of astronomical bodies. The time engineering merges here with the space engineering. Consider the following to understand this:
    - In some temples you would notice that the sunlight falls on the deity on a specific day / time of the day. This is not possible if the future position of astronomical bodies at a given time is calculated very accurately.
    - When we build a house / temple the inauguration ceremony should be conducted at a specific time. So it's just not Space Engineering but time engineering also.

  - We see that the concept of space and time are not just a continuum but:

    - The measures of time and space merge and results in the same unit measure called Tala. It is Adi Tala in Poetry, Music and the same measure is used in building and sculpture dimensions.

- **Space is created because of the pulsation called Time** and hence the unit time measure (for pulsation) and the unit space measure (result of pulsation) are one and the same. This means the pulsation causes proportional displacement or creation of Space.

- It is the same concept used in meditation techniques. With Pranayama, Dharana and Dhyana you control the pulsation and align with it. In Samadhi you stop the pulsation and hence you are beyond time and space and you reach a luminescent state of resonant non-vibration.

Hence I named this book as **“Time = Space”.**

We have seen that each subtle unit is represented in 8x8 energy grid and the gross unit is represented in a 9x9 energy grid. Mayan states that **“Each module or pada within the 8x8 / 9x9 structure resonates with a specific energy.** Based on Space, Time, Light, and Sound coordinates the frequency of vibration has its unique position in the Space/ Time continuum. **This energy level of frequency of vibration is called a luminous body or Devata**. This devata has particular attributes based upon its position and qualities in the Space/Time continuum (placement among the 64 / 81 sub-cubes or padas)”.

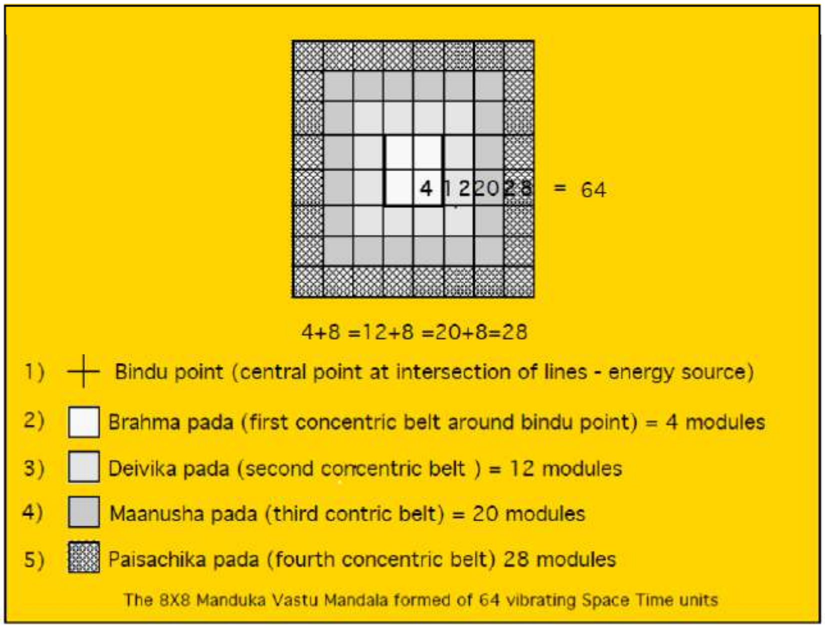
4 12 20 28 = 64
4+8 =12+8 =20+8=28
1) Bindu point (central point at intersection of lines - energy source)
2) Brahma pada (first concentric belt around bindu point) = 4 modules
3) Deivika pada (second concentric belt) = 12 modules
4) Maanusha pada (third contric belt) = 20 modules
5) Paisachika pada (fourth concentric belt) 28 modules
The 8X8 Manduka Vastu Mandala formed of 64 vibrating Space Time units

Figure 37: 5 states of 8x8 Energy Grid (Courtesy: Dr. Ganapathi Sthapati)

Now what is evident here in this science is 8 and 5 merges and so does Space and time.

- **Manifestation occurs in additive values of eight and in 5 stages. You can note that the Bindu (Point) has become 4, then 12, then 20 and then 28. (4+8=12 +8=20+8=28 with a total additive factor of 64). Manifestation occurs in sequences of five steps known as the Pentadic Order.**
- **The mathematical calculation which gives form to consciousness / space / subtle matter is called “Ayadi Gananam”. We know that Ganam as the additive process.**

Dr. Jessica Marcey in her Fabric of Universe states that:

- The subtle energy point in the center becomes a self effulgent energy generator pouring out waves of energy
- becoming a self spinning stabilized structure of four padas / modules then
- adding eight units of energy to manifest the 4X4 structure of 16 units or modules; then
- adding another eight units of energy to manifest the 6X6 structure of 36 units or modules; then
- Adds eight more units culminating in 64 units or modules strung concentrically around the Bindu point or central generator.

So a central point with 4 concentric square belts around, each has its own frequency, vibration and energy characteristics evolving into a 9x9 unit as a gross matter. While this knowledge is significant of itself, when these phenomena are viewed in light of material manifestation as built space in architecture, the significance becomes stunning and profound.

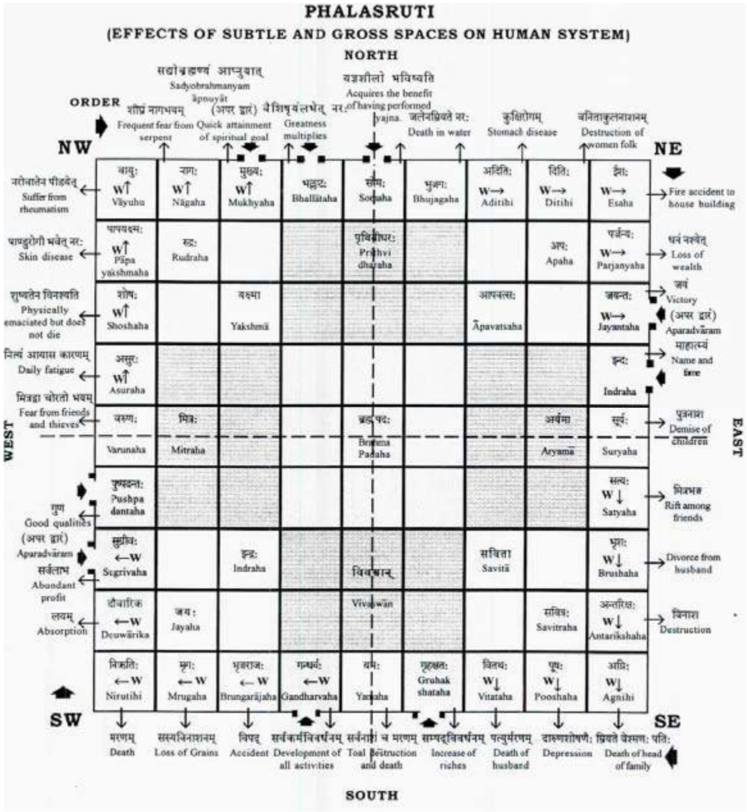
PHALASRUTI
(EFFECTS OF SUBTLE AND GROSS SPACES ON HUMAN SYSTEM)
NORTH
SOUTH
WEST
EAST
NW
NE
SW
SE
ORDER
Sadyobrahmanyam āpnuyāt
Acquires the benefit of having performed yajna
Frequent fear from serpent
Quick attainment of spiritual goal
Greatness multiplies
Death in water
Stomach disease
Destruction of women folk
Fire accident to house building
Loss of wealth
Victory
Aparadvāram
Name and fame
Demise of children
Rift among friends
Divorce from husband
Destruction
Suffer from rheumatism
Skin disease
Physically emaciated but does not die
Daily fatigue
Fear from friends and thieves
Good qualities
Aparadvāram
Abundant profit
Absorption
Death
Loss of Grains
Accident
Development of all activities
Toal destruction and death
Increase of riches
Death of husband
Depression
Death of head of family
Vāyuha
Nāgaha
Mukhyaha
Bhallātaha
Somaha
Bhujagaha
Aditihi
Ditihi
Esaha
Pāpa yakshmaha
Rudraha
Prithvi dharaha
Apaha
Parjanyaha
Shoshaha
Yakshmā
Āpavatsaha
Jayantaha
Asuraha
Indraha
Varunaha
Mitraha
Brahma Padaha
Aryamā
Suryaha
Pushpa dantaha
Satyaha
Sugrivaha
Indraha
Savitā
Brushaha
Deuwārika
Jayaha
Vivaswān
Savitraha
Antarikshaha
Nirutihi
Mrugaha
Brungarājaha
Gandharvaha
Yamaha
Gruhak shataha
Vitataha
Pooshaha
Agnihi
W

Figure 38: 9x9 energy grid with corresponding energy levels for each grid.
(Courtesy: Dr. Jessica Marcey, Fabric of the Universe)

So, to sum up this section, we have seen that how every inanimate thing in this universe is still a living being and how with the spatial science we can pump throbbing energy into a building. This is the supreme science where Time and Space merges. This is the 4th sacred art we talked about which is a manifestation of the visual form of Om.

உ

As you would have noticed I have avoided talking about rituals and rules here but focused on the concept and philosophy. When we talk about rituals there are very many variations and we get into an unnecessary argument that whether it is rational or superstitious. Since most of the arts have become commercialized, the person’s qualification and his intentions are always questionable.

The objective of this book is to highlight the scientific aspects of the universe in a small way and emphasis that our culture is built on this science. Be it language or arts or religion all are built on scientific facts. They all were nourished and nurtured in thousands of Temples across India which is the container and protector of this subtle science and energy.

Let’s see some more interesting aspects of this in the next section.

Happy reading!

28. Time = Space Series. Part 28 – Capturing the divine energy – Squares & Circles

It is told that the sanctorum of Lord Ranganatha's temple at Sri Rangam is circle from outside and square from inside. Why?

- Subtle energy is represented by 8x8 square grid and the perimeter of the square is 4r and the area of the square is r*r. When each grid length is 1 unit then r = 8, so perimeter is 32 and area of this square is 64.
- Gross energy is represented by 9x9 square grids and with the same measurements perimeter is 36 and area of this square is 91.

It has been an architect's dream to create and capture subtle energy in a gross structure. This is fundamental objective of spatial engineering. How do you go about to achieve this? It looks that they found that they would take the help of a circle either outside the square or inside the square to achieve this. Remember almost every yantra has a square and a circle component in it.

The measurements are so taken such that either the perimeter or the area of both the circle and the 8x8 square are matched. Most of the Hindu temples follow such rules to invoke the subtle energy in a gross structure like temple. This is one of the well kept secrets and this is the reason each and every temple is said to interact with your subtle body differently when you visit it.

This is why Lord Ranganatha's sanctorum in Sri Rangam temple is a circle from outside and square from inside. This secret is called as **Squaring a Circle**. This is all about to creating divine energies and this is the science that is implemented in all the Hindu temples which are built as per the Sastra. You need just basic math knowledge and patience to understand this supreme secret and I shall give below 2 techniques to match the area or the perimeter of the circle with that of the 8x8 square.

Perimeter match:

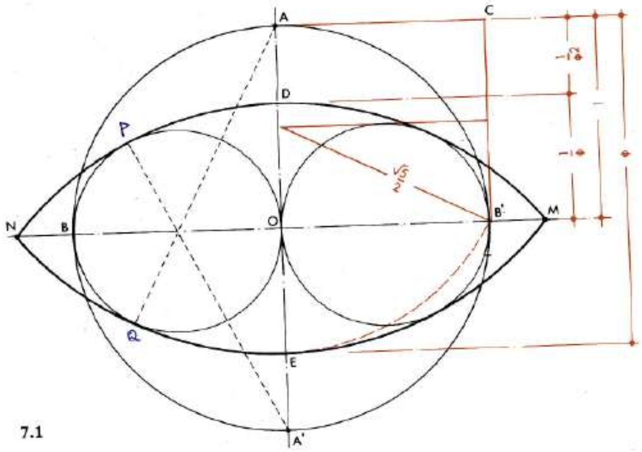

Figure 39: Squaring a Circle (Courtesy: Dr. Robert Lawlor, Sacred Geometry)

- Let's look at the image of a circle of unit radius. The **area of the large circle is π** (πr^2) and the **circumference is 2π**. Here r =1.
- Draw 2 circles inside as shown with half the radius. The area of each circle is π/4 and the **combined area of both the circles is π/2**. The circumference of each circle is π and the **combined circumference of both the circles put together is 2π**.

- Just note that when a larger circle divides itself as 2 smaller circles the area is halved but the circumference is the same. **This gives the first philosophical metaphor that "One has become two which are bipolar yet non-dual". This is one of the fundamental principles of Advaita and also in Chinese philosophy. This bipolar non-dual force is the basis of life and is represented as male / female and Yang / Yin.**

- Now draw 2 arcs NDM and NEM from A' and A with the radius as A'P and AQ as shown in the dotted line. You would note that point D and E would divide the radius of the circle by divine proportion – $1/\Phi$ and $1/\Phi^2$. **Radius of the arc AE is Φ. The Yin Yang symbol is based on this principle which embeds the divine proportion in it.**

- Now let me draw a square ABCD around this circle (Fig 40) where the side of the square is the diameter of the circle which is unit 2. Here OP =1, AB=BC=CD=DA = 2. **The perimeter of the square is 8 and the area of the square is 4.**

- Now draw a circle with O as the center and NO as the radius as shown in the figure. Here NO = $\sqrt{\Phi}$. You can take this figure as such if you find it difficult to follow but for those who are interested, the derivation to calculate radius "NO" is as below:

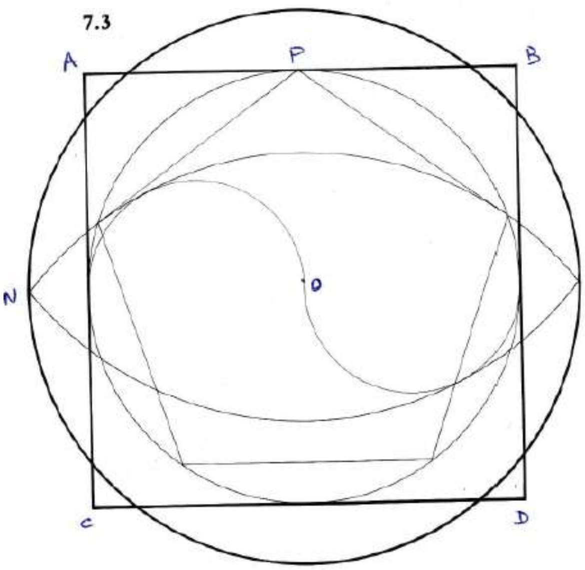

Figure 40: Squaring a Circle 2 (Courtesy: Dr. Robert Lawlor, Sacred Geometry)

  - Consider a right angle triangle OPN. We know OP = 1 and PN which is the radius of the arc = Φ. Hence as per Pythagoras theorem $PN^2 = OP^2 + ON^2$ which is $\Phi^2 = 1+r^2$. Let us say the radius is ON = r.
  - So $r = \sqrt{(\Phi^2-1)}$. We know Φ is 1.61803399 so **$r = \sqrt{\Phi}$.**

- We have got a larger outer circle with a radius **$\sqrt{\Phi}$** and hence the perimeter should be **$2\pi\sqrt{\Phi}$. This is numerically equal to 7.9924576, which is equal to 8. Now we have created a structure where a square whose perimeter and a circle whose perimeter are the same.** This technique is not limited to India but used in the same way for pyramid construction.

- **This is the perimeter match between 8x8 square grids and a circle around it**.

Area match:

- Vaastu has an easier way to get this done. Draw a circle inside a 9x9 energy grid whose diameter is equal to the side of the square. If each energy grid length is 1 unit, then side of the square is 9 units and the radius is 4.5 units.

- The area of this circle is πr^2 where r=4.5 and hence the area of the circle is 63.64 which is almost equal to area of the 8x8 square with unit grid length as 1 unit.

- This is the area match between a circle and 8x8 square. So as per Vaastu every building is divided into 9x9 grid and the energy level of each grid is used to design the living space.

Capturing the direction of Subtle energy flow:

The subtle divine energy flows is as per the diagram given below. It starts from the center and always flows clockwise. That is why Hindus always stick to clockwise flow of movements - be in their temples or anywhere else. This clockwise flow of movement is captured in the sacred symbol Swastik. Here 1 is the source / Moolam, 2 is Kaalam (Time), 3 Seelam (Rhythm), 4 Kolam (Form), 5 (Universe) Gnalam.

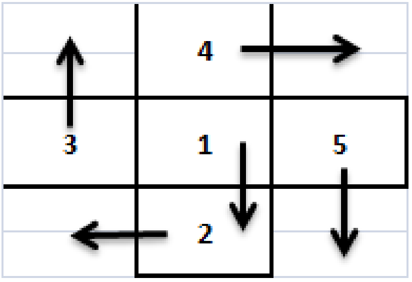

Figure 41: Clockwise flow of subtle energy

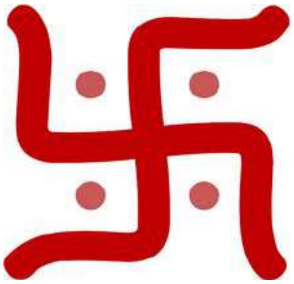

Figure 42: Swastik symbolizing energy flow

Relationship between $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, Φ and π:

- Now let me draw the connection between $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, Φ and π. Understanding this relationship is orgasmic and nothing less. If you think they are all some random numbers which cannot be fixed to their 10th decimal and not related to each other, just see the following relationship.

  $$\pi = \Phi^2 \times 6/5$$

  Philosophically 6/5 is the relation of a hexagon to a pentagon and the other way to express this relationship is

  $$\Phi = (\sqrt{\pi} * \sqrt{5}) / (\sqrt{2} * \sqrt{3}).$$

- This goes to prove that we did not just knew the numerals but divine fractions and proportions including their linkages. I am convinced, there is not even an iota of superstition in our temple and idol worship but they are supreme sciences codified and left for our benefits.

- The temple of **Sri Rangam** dates back from Ramayana, which is Treta Yuga that is 800000+ years back as per Hindu Scriptures. This is our rich past and supreme science is embedded in our culture and codified in thousands of temples across nook and corner of the country. Why should we ignore this? We have a lot more to offer this world which is struggling to understand the nature of nature.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

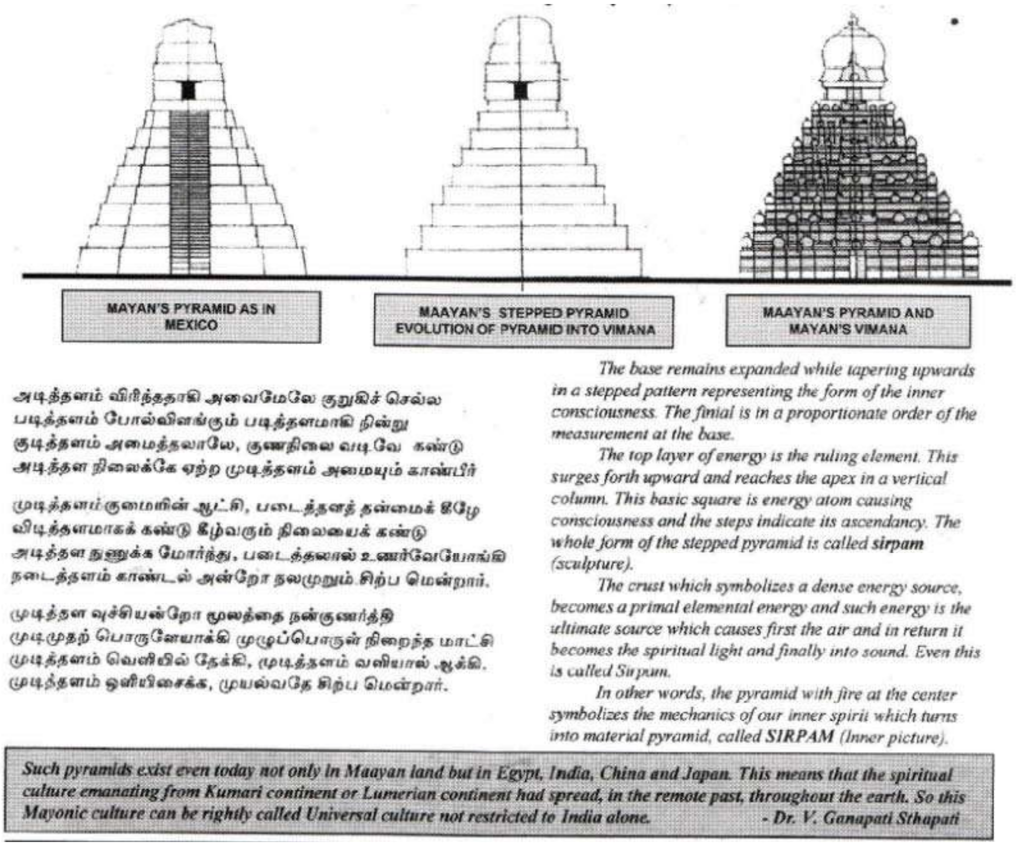
MAYAN'S PYRAMID AS IN MEXICO
MAAYAN'S STEPPED PYRAMID EVOLUTION OF PYRAMID INTO VIMANA
MAAYAN'S PYRAMID AND MAYAN'S VIMANA

அடித்தளம் விரிந்ததாகி அவைமேலே குறுகிச் செல்ல
படித்தளம் போல்விளங்கும் படித்தளமாகி நின்று
குடித்தளம் அமைத்தலாலே, குணநிலை வடிவே கண்டு
அடித்தள நிலைக்கே ஏற்ற முடித்தளம் அமையும் காண்பீர்

முடித்தளம் குமைமயின் ஆட்சி, படைத்தளத் தன்மைக் கீழே
விடித்தளமாகக் கண்டு கீழ்வரும் நிலையைக் கண்டு
அடித்தள நுணுக்க மோர்ந்து, படைத்தலால் உணர்வேபோங்கி
நடைத்தளம் காண்டல் அன்றோ நலமுறும் சிற்ப மென்றார்.

முடித்தள வுச்சியன்றோ மூலத்தை நன்குணர்த்தி
முடிமுதற் பொருளேயாக்கி முழுப்பொருள் நிறைந்த மாட்சி
முடித்தளம் வெளியில் தேக்கி, முடித்தளம் வளியால் ஆக்கி.
முடித்தளம் ஒளியிசைக்க, முயல்வதே சிற்ப மென்றார்.

The base remains expanded while tapering upwards in a stepped pattern representing the form of the inner consciousness. The finial is in a proportionate order of the measurement at the base.

The top layer of energy is the ruling element. This surges forth upward and reaches the apex in a vertical column. This basic square is energy atom causing consciousness and the steps indicate its ascendancy. The whole form of the stepped pyramid is called ***sirpam*** *(sculpture).*

The crust which symbolizes a dense energy source, becomes a primal elemental energy and such energy is the ultimate source which causes first the air and in return it becomes the spiritual light and finally into sound. Even this is called Sirpam.

In other words, the pyramid with fire at the center symbolizes the mechanics of our inner spirit which turns into material pyramid, called ***SIRPAM*** *(Inner picture).*

Such pyramids exist even today not only in Maayan land but in Egypt, India, China and Japan. This means that the spiritual culture emanating from Kumari continent or Lumerian continent had spread, in the remote past, throughout the earth. So this Mayonic culture can be rightly called Universal culture not restricted to India alone. - Dr. V. Ganapati Sthapati

Figure 43: Similarity between Indian temples, Pyramids. The Tamil text states how the structure should be built. (Courtesy: Dr. Ganapathi Sthapati)

29. TIME = SPACE SERIES. PART 29 – HEAVEN & EARTH – LORD NATARAJA & LORD RANGANATHA

Let’s stay with Lord Ranganatha and see the relationship between him and Lord Nataraja. Would you be surprised if one is the mirror image of the other? Look at the majestic picture of Lord Nataraja below. Wonder why this is always made in metal and not in wood or earth?

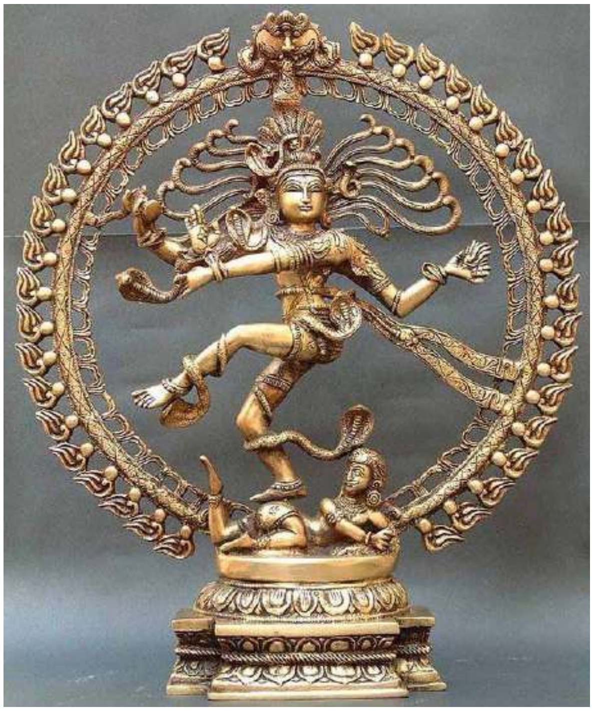

Figure 6 reproduced.

In Pancha bhoota (Space, Air, Fire, Water, Earth) Metal is always associated with Space and Wood is associated with Air. This system is reflected well in the Chinese system and it was not any different in our Vaastu system either. Since Lord Nataraja signifies Akasha tatva, **(Vin porul)** he is always made in metal – Copper or Gold or Bronze or a mix of it. The statues which were ancient had a generous amount of gold giving a luminescent look about the statues. The art and science of moulding metals to create divine forms with precise proportion dates back to thousands of years in India and the western experts think that this is a recent phenomenon. This is a complete surprise to the western world since they cannot believe that such sophistication existed before.

Dr. Ganapathi states, *“Lord Nataraja also signifies the primal image being expressive of the secret of creation in the process of which formless becomes instantaneously endowed with features and of perfect form. This is the reason why the statue of Lord Nataraja is called as* ***Silpa****, the primal image and everything else which is sculpted in this universe is called as* ***Pratima (replica)*** *and the original and one and only sculpture is the image of Lord Nataraja*”**.**

Mamuni Mayan states, *"Bhumih pradhana vastu Syat"*, means Earth by virtue of being the very basic support of all things, this is the primal and dominant Vastu". Earth and space are manifestations of each other in the cyclical process of evolution. They both should always be considered to be in unison. This vibrant, energetic, manifest earth energy (**Prithvi Tatva)** is called as '**Vishnu**".

In Vaastu satra, the Vaastu purusha is Lord Vishnu and understandably Vishnu is invoked as Vaastu purusha in Srirangam temple and offered a seat in the heart of Lord Ranganatha even today as a ritual.

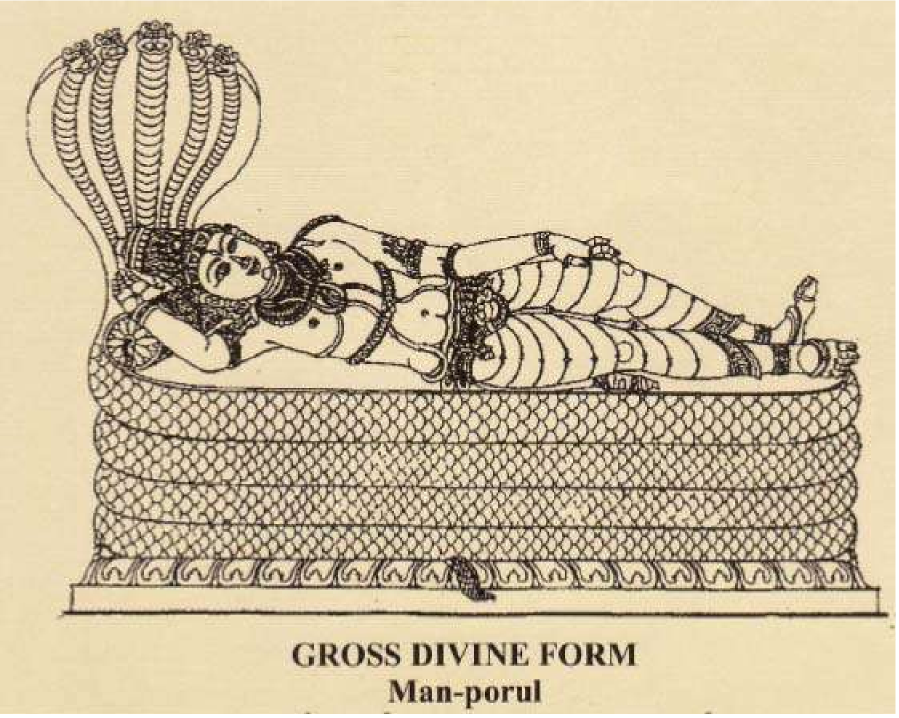

Figure 44: Lord Ranganatha in the reclining pose

Lord Ranganatha represents manifest earth energy **(Man porul)** and hence is always made out of seasoned earth only and it should be properly painted too. He is never to be made of metal. This earth form does not support frequent mobility and symbolizes stability and immobility. **This is the reason when Vibhishana kept the idol of Lord Ranganatha down at Sri Rangam it could not be moved.**

The energy that is in available in every universal form is represented as Vishnu and hence he is called as **Antaryami**. Once the subtle becomes grosser the vibration is reduced and the grossest form would not have any creative pulsation in it. It would still have the vibration based on its nature which ensures that it retains the form and shape. But it does not have any creative pulsation and it has reached the zenith of evolution. **This zenith is characterized by Lord Ranganatha in the reclining pose**.

The creative pulsation (Space) is the dance and the zenith of manifestation is a reclining posture (Earth).

As we have seen, Lord Nataraja carries **Light in his left hand** and **sound in his right hand**. Lord Ranganatha carries **Chakra representing light in the right hand** and **Conch representing sound in the left hand.** The principle that the primal substance (space) has manifested as the universe (earth) and both of them are mirror images of each other is reflected in both the images.

In Chinese texts the Yang energy is associated with Heaven and the Yin as earth. Now associate the same heaven with Shiva, Earth with Vishnu and Tao with the God's particle - with some of the text below from **Lao Tzu's "Tao te Ching"**

"The Tao that can be told is not the eternal Tao.
The name that can be named is not the eternal name.
The nameless is the beginning of heaven and Earth.
The named is the mother of the ten thousand things.

Something mysteriously formed, Born before heaven and Earth.
In the silence and the void, Standing alone and unchanging,
Ever present and in motion, Perhaps it is the mother of ten thousand things.
I do not know its name. Call it Tao.

Man follows Earth. Earth follows heaven. Heaven follows the Tao.
Tao follows what is natural."

To sum up in the words of Dr. Ganapathi Sthapati:

- **Both the temples Chidambaram and Srirangam are south facing temples.**
- **Lord Nataraja signifies Akasha Tatva, outer space, macrocosm, subtle unmanifested energy is the Vastu purusha, is always in dancing pose, circumscribed in 8x8 square and made of metal.**
- **Lord Ranganatha signifies Prithvi Tatva, inner space, microcosm, manifest gross energy, is the Vaastu purusha is always in reclining pose, depicted in a circle or a 9x9 grid and made of earth.**
- **Lord Nataraja represents energy with matter and Lord Ranganatha matter with energy. In the equation $E = mc^2$, Lord Nataraja represents E and Lord Ranganatha represents mc^2.**
- Both these temples are a 'must see" and a "periodic see" types and you would be amazed that the qualitative change in your mental state even if you don't pray inside the temple but spend some time sitting or walking in the premises of the temple.
- **Both these temples are perfect symbols of ancient science codified for the benefit of the society and left for generations withstanding the test of time.**

உ

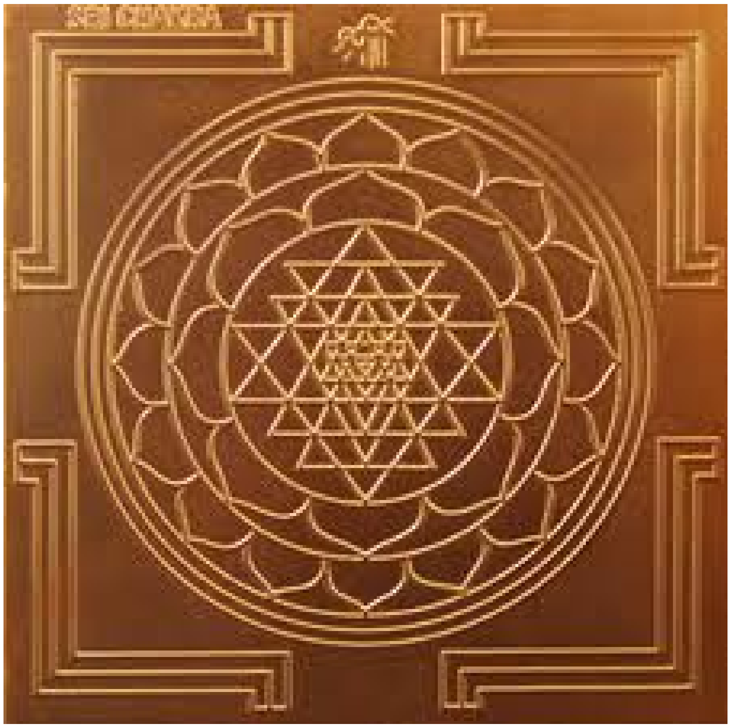

Figure 45: Sri Chakra – The Supreme shape in sacred Geometry

Chidambaram and Sri Rangam should be ground zero for any lab that wants looks into the secrets of space and time as they represent supreme science as much as they represent spirituality. We now see that science was the basis of our spiritual faith unlike it being 2 completely different aspects today across the world.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

May we be blessed by the divine power to unravel unknown secrets about the nature!

Happy reading!

உ

30. Time = Space Series. Part 30 – Temples and Idol worship

Let us forget Hinduism, its symbols and names of god for the time being and consider these aspects as pure science from a nanotechnologist and cosmologist called Mayan.

We have been told that the universe is full of luminescent particles and this particle manifests itself because of the primordial pulsation. We have had a sneak peek into the geometry and the mathematical proportion which is used in the creation of many things. We know that the universe rhythm confirms to a mathematical code to the extent that we can create divine energies in a temple based on that science.

If time is pulsation of consciousness and if it creates, sustains and destroys everything, then time is the architect sculpting the universe and every object in this universe should be a sculpture of time and this concept is very beautifully captured in ***“Jagat Sarvaram Silpameva’*** *meaning the entire universe is nothing more than a shilpa / sculpture (and it latently conveys that Time is the Shilpi).*

If all the above things are right then is it not right to say that “Man created the concept of God” or “Man can create divine energy around him”? Let me sum up my thoughts as below:

- The sages and rishis were essentially scientists and understood the secret of nature very well. They were understandably masters in **nanotechnology and cosmology.**

- They knew that “**One thing”** with which they could understand the nature and science of all other things in this universe. This is obviously supported by the mathematical knowledge they had.

- As per Vedas nothing is created by god including the Vedas, but universe is the manifestation of god. They also understood that man is the most evolved of this manifestation. This is why they created all the god forms in the image of a man.

- Be it language, poetry, classical music, classical dance, sculpture and architecture it was the manifestation of this divinity that was the underlying link. Sacred geometry and sacred proportions were used to generate the subtle energy that took us closer to divine.

Hindu temples were created to capture all these aspects in one place and which could sustain and grow for the benefit of mankind.

- Secret of nature heard through the vibrations of Vedas were codified as supreme science and weaved into a philosophy called **Sanatana Dharma.**

- Architecture of temples confirmed to divine proportions and space - time engineering was accurate to capture sunlight and moon light at the same position for thousands of years. The design of the temple cleaned the subtle energies and points of the human body.

உ

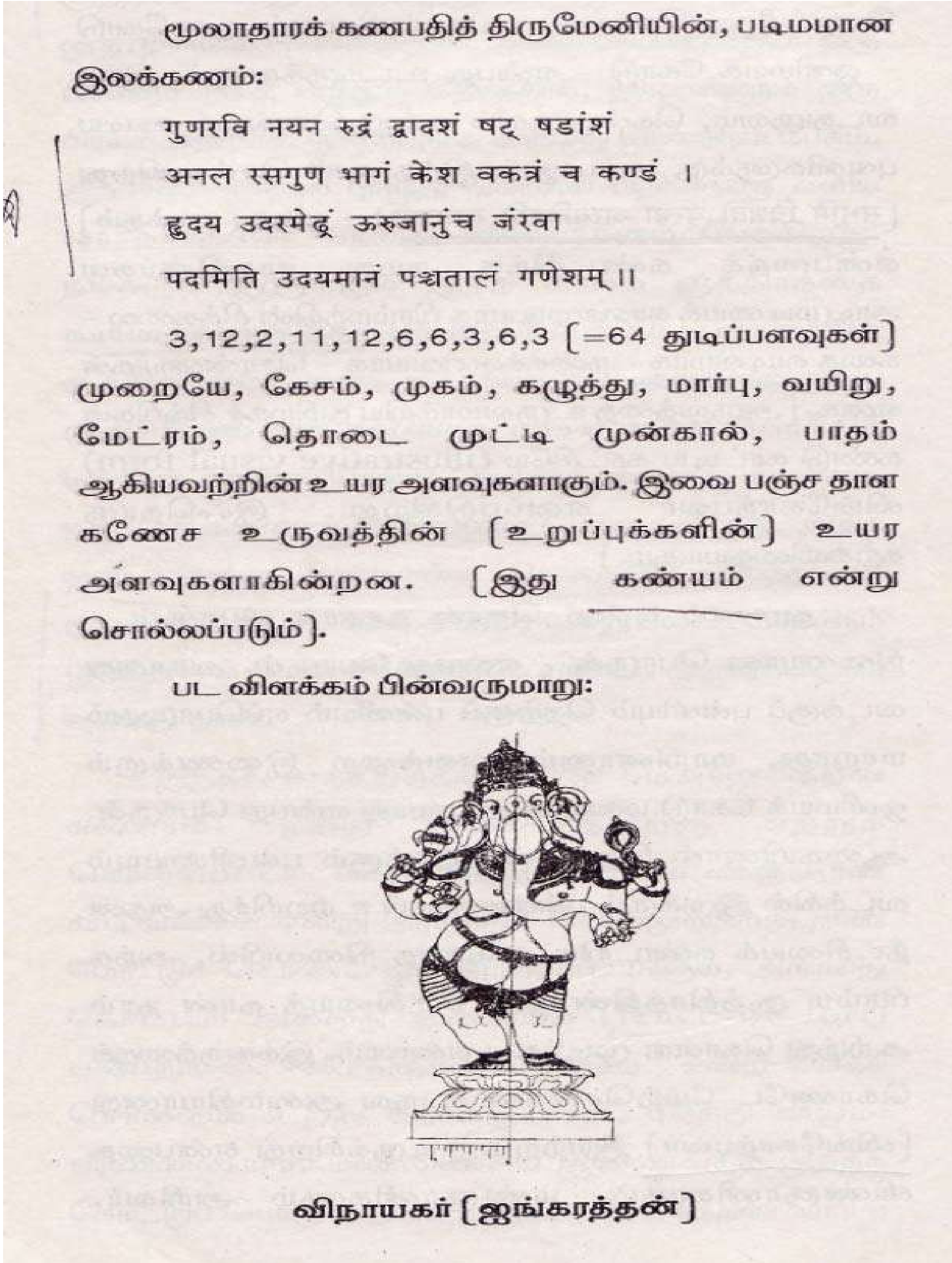

மூலாதாரக் கணபதித் திருமேனியின், படிமமான இலக்கணம்:

गुणरवि नयन रुद्रं द्वादशं षट् षडांशं
अनल रसगुण भागं केश वकत्रं च कण्डं ।
हृदय उदरमेढ्रं ऊरुजानुंच जंखा
पदमिति उदयमानं पञ्चतालं गणेशम् ॥

3,12,2,11,12,6,6,3,6,3 [=64 துடிப்பளவுகள்] முறையே, கேசம், முகம், கழுத்து, மார்பு, வயிறு, மேட்ரம், தொடை முட்டி முன்கால், பாதம் ஆகியவற்றின் உயர அளவுகளாகும். இவை பஞ்ச தாள கணேச உருவத்தின் [உறுப்புக்களின்] உயர அளவுகளாகின்றன. [இது கண்யம் என்று சொல்லப்படும்].

பட விளக்கம் பின்வருமாறு:

விநாயகர் [ஐங்கரத்தன்]

Figure 46: Ganapathi Idol (Courtesy: Dr. Ganapathi Sthapati)

The sloka above in Sanskrit gives the measure of different parts of this Idol which would confirm to total 64 units (resonance of 8). Hair (3), Face(12), neck(2), Chest(11), Stomach(12), Belly &Pelvis (6), Thigh(6), Knee(3), Legs (6), Feet (3). Total 64 units. Just as much a Sloka with right metre and letters becomes a living organism, an idol with right proportions becomes a living organism.

- Divine Idol forms confirmed to a mathematical proportion that invoked subtle energies within and around us. This is where the external forms were used to create energies internally to a human. **This is the basis of idol worship. If it is scientific to create a form with divine energies based on mathematical proportion then idol worship is the supreme proof for that science.**

- **This supreme mathematical law is called as Dharma (nature's law) and the unfolding of that law is the Karmic theory of cause and effect / law of karma.**

- Poetry, classical music and classical dance are intimately linked to this Hindu temple culture. Adi Tala and octaves were used to create subtle energies in poetry using words, and the same tala maana is used in classical music with rhythm and raga, in classical dance with light and sound and in sculpture as a visual form.

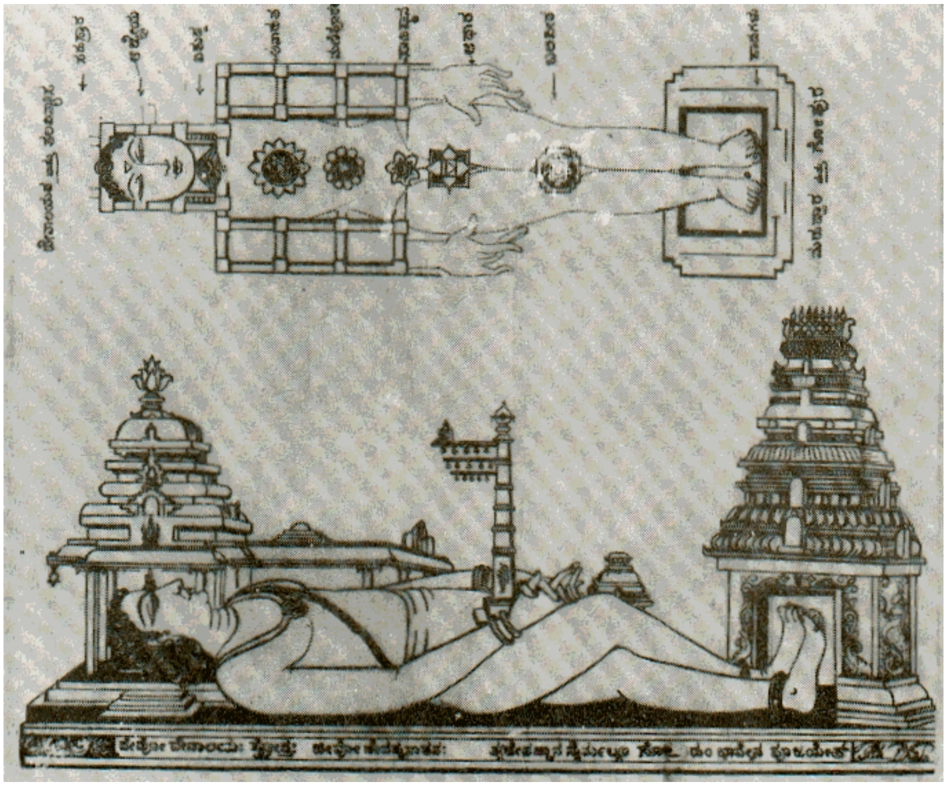

Figure 47: Human body as a representation of a temple
(Courtesy: http://www.salagram.net/sstp-mgpuja3.html)

A Hindu temple is a divine and yogic representation of a human being with the Deity in the temple representing the God as indweller in humans and all beings. In a temple the feet represents Rajagopura, the hands represent Praakaara, the abdomen represent Mandapa, the heart represents Antaraala and the crown of the head represents the sanctum sanctorum (garbha griha). The temple is used as a reminder that our inner spiritual journey is through internal yoga to realize the indweller God. This analogy is shown through the representation of various chakras namely Moolaadhara to Sahasraara in the body to various locations in the temple.

உ

- It may be of interest to note that the same harmonics of 8 are reflected in every other domain. For example, if you have been practicing Pranayama then your master would have asked you to practice breathing based on some counts. The most popular counts are 6(inhale), 3 (Exhale) or 7(inhale), 2 (Exhale). If you notice this count 9, this cleans up your gross body. Advanced Pranayam students may be asked to stick to 1 (Inhale), 4 (Retain), 2(Exhale), 2(Retain outside) or as a variation 1,4,2,1 or 1,3,2,2. The counts which add up to 8 clean the subtle body.

- Till few decades back the fundamental dimension for money in India was "**Ana"** called as **Eight Anas (எட்டணா)** and this was equivalent to Rs. 0.5 as the fundamental measure.

- Thus a **"Poetry is an idol in aural form and a building is a frozen music"**. A Poet by the way of his mental frequency has first become the poem itself and then the actual poem was created in its aural form and an artist by the same frequency has become the idol first and then he gave a shape to the idol externally through this hands. This is very beautifully given as ***"Raso vai saha, So Vai rasikaha", which means he (god) is the enjoyer and he is the artist.***

- All of this reflects the supreme philosophy that **"Oneness in all and everything is just a manifestation of that one thing"**.

- Now in this light understand the meaning of this Upanishad prayer, *"Aum Poornamadah Poornamidam Poornaath Poornam Udachyathe, Poornasya Poornamaadaaya Poornameva Vasishyathe". This means God is perfect (infinite). This Universe is also perfect (infinite). If perfection (infinity) is taken from anything perfect (infinite) what remains is still perfect (infinite).* Isn't this is the actual reality today?

- Everything scientific is philosophically, mythological and ritually well interwoven that proves that Sanatana Dharma is the most ancient and supreme representation of science and nature's secret in our possession. Our temple worship which integrates all this science, 5 sacred arts with language and Math is the standing example of our understanding of the nature of god.

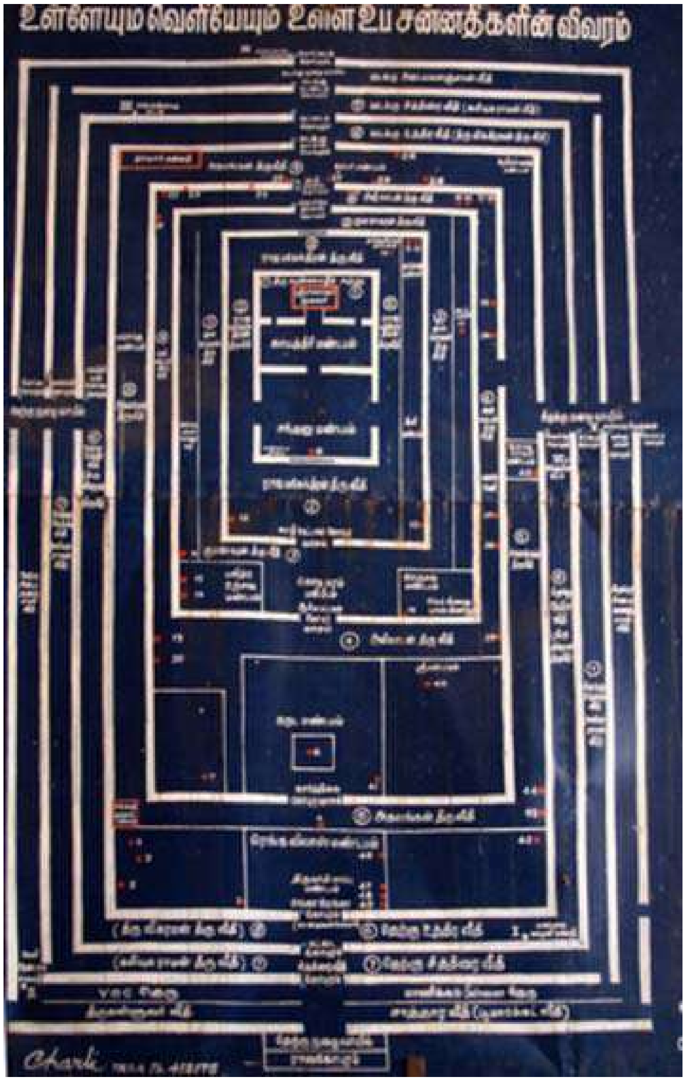

Figure 47: Layout of Sri Rangam temple with 7 Prakaras

It is the same Sugarcane but an elephant eats it from the field directly, a young adult peels the skin off and chews it, if a kid has to eat then someone has to peel the skin and cut it into pieces and an old man without teeth needs it to be made as sugarcane juice. Depending on one’s ability to chew and digest the food has to be processed.

Same way depending on one’s mental ability to grasp the supreme truth it has to be processed and presented. If someone is mentally well evolved then they can be fed with Upanishads and he can meditate on the formless Brahman. To a young adult Upanishads may not be palatable directly but they may have to start with Upavedas and Vedangas to mature to a higher level. For the less evolved kids, Itihasa and Purana like Ramayana and Mahabharata with stories may drive home the point and they can start with Shlokas and mantras. But for the least evolved it has to start with a personal deity with name and form, bhajans and songs to invoke the concept of Bhakthi.

In all the four cases temple culture forms the lowest common denominator. Even the most evolved relish the arts and architecture in the temples. We should be proud of our rich cultural and scientific heritage and be passionate about the science and philosophy and be rational about rituals and symbolisms.

Image / Idol worship is not deplorable or despicable. This is the highest scientific form of worship which enables us to see face to face the one which is beyond reach of sight and speech. This kindles devotion, community development and helps one to climb the steep walls of spirituality in the initial stages. Every religion in this world has some image or idol as symbolism to achieve this objective.

An idol is a god in form when made with the science of divine proportion and seasoned material is the highest product of our intelligence till today.

It should however be noted that this idol worship should not be lost in the din of rituals which does not reflect the very philosophy it had come to symbolize. So the advice I follow is *"Be passionate about our culture but be rational about the rituals.*"

The advancement of science can progress rapidly when the western scientists who are objective and empirical, partner with the pundits in India and learn the science of the subtle from our scriptures and temples. For this to happen we need to learn our own scriptures and past which is not possible without learning our mother tongue and Sanskrit. Else in the near future we may have to import knowledge of our culture and rich past from the western world.

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Happy reading!

31. TIME = SPACE SERIES. PART 31 – MAYAN & HIS WORKS

Mahamuni Mayan was the father-in-law of Demon King Ravana and is referred in both Mahabharata and Ramayana. He is attributed with building the palace for Yudhishtar in Indraprasta and also the architect for Sri Lanka which Ravana ruled. His son Nalan helped Lord Rama build the Ramasethu – the bridge between India and Sri Lanka. He is believed to have lived in the **Ilamuridesam** popularly known as **Lemuria continent (Kumari Kandam**) which spanned 4500+ kms from Sri Lanka to Antartica connecting current Australia, Africa and Americas. Check out http://www.sacred-texts.com/atl/tll/index.htm

Understandably when this land mass was consumed by the sea their clan moved to different parts of the world including Guatemala, Egypt, and to current south Indian region. This explains as to how we find common science across India, South America and Egypt.

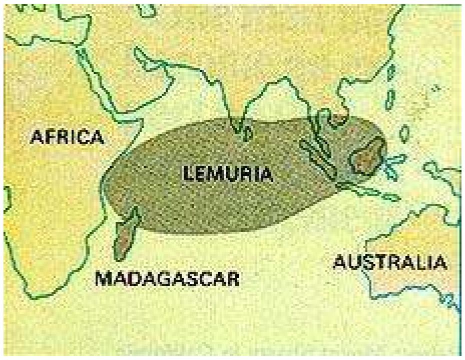

Figure 48: Kumari Kandam depiction

He has written books in both Sanskrit and Tamil and I am going to produce the list of technical works he has created and look at the contents of a specific book called Surya Siddhanta. Just the list of works and also the topics in Surya Siddhanta is amazing, highlighting the depth of science we had at that time. http://www.youtube.com/watch?v=PZKiCpFisoY

The Surya Siddhanta

The most ancient treatise on Astronomy is a treatise authored by Mayan. It is called *The Surya Siddhanta*. Included in this treatise are the following chapter topics following:

- The Motions of the Planets
- The Places of the Planets
- Direction, Place and Time
- The Moon and Eclipses
- The Sun and Eclipses
- The Projection of Eclipses
- Planetary Conjunctions
- Conjunction of the Stars
- Risings and Settings
- The Moon's Risings and Settings
- Certain Malignant Aspects of the Sun and Moon
- Cosmogony, Geography, and Dimensions of the Creation
- The Gnomon
- The Movement of the Heavens and Human Activity

உ

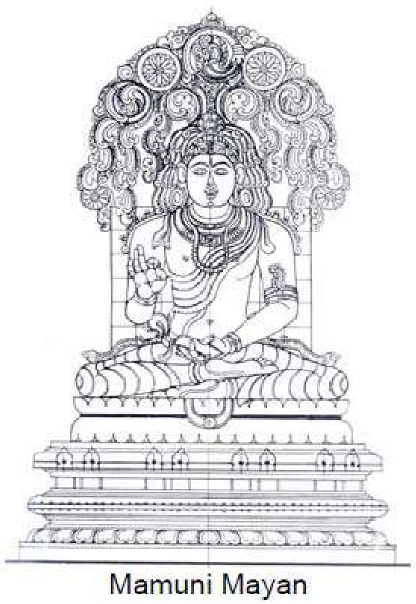

Figure 49: Mamumi Mayan

Mayan is also attributed with the start and spread of Temple culture in India. The song below in Tamil states that since Mayan sat under the Banyan tree and taught this science of universe and also of the sacred arts, so that every town was endowed with Temples for social betterment. Because he was from the southern region and also he sat under the Banyan tree, he is also equated with Lord Dakshinamoorthy.

ஆலமா மரத்தின் கீழே அருள் மயன் அமர்ந்ததாலே
ஆலமா மரத்தின் கீழே அருட் கலை வளர்ந்ததாலே
ஆலமா மரத்தின் கீழே அறிவர்கள் வளர்ந்ததாலே
ஆலயம் ஊரும்தோறும் அமைந்தன செழிக்கநன்றே!!

The Mayan says in his Pranava Veda verse one:
*"Om Light and Om Sound are the Primal Source of all manifest forms. Om Light is aroused by its own effort in a state of disorder and appears as a flame. The state of Om Light and Om Sound in Space is a magnificent luminous six – faced Light form that is called "**murukoli**". The transformation of Om Light and Om Sound through the five stages is concealed in the five fold knowledge, of which, this is the first. This process of transformation of disorderly Om Light and Om Sound into orderliness is found in all five fold material forms."*

Sage Veda Vyasa says in the Bhagavatam, *"eka eva pura vedo pranavha sarva vangmayha" (9th skandha, 14th chapter, sloka 48 of Bhagavatam).* This means that there was only one Veda called Pranava Veda. It is believed that Sage Vyasa created 4 vedas from this Veda since he believed that the oncoming generations would not have the mental strength to master everything.

உ

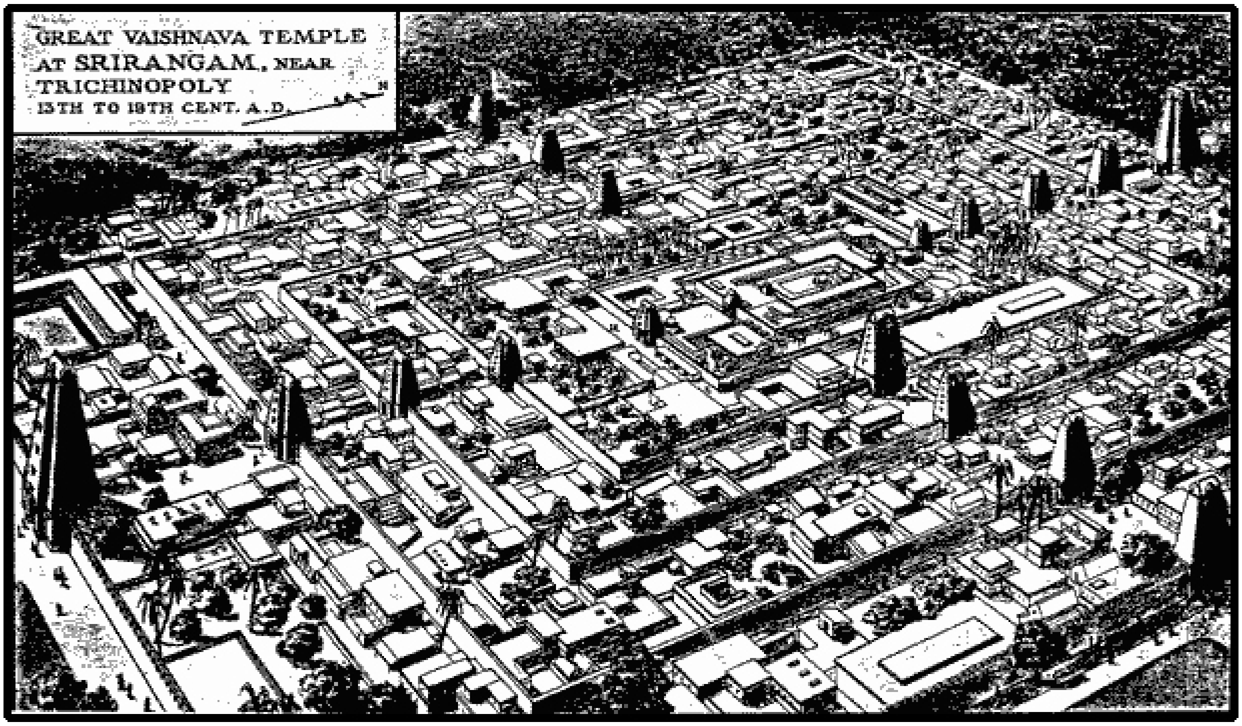

Figure 50: Aerial view of Sri Rangam temple with 7 Prakaras

Technical works

The following are technical works written by Mayan and represent the 12 vowels of the Tamil language:

- Ovia Chennool - Treatise on Drafting and Painting
- Sirpama Chenool - Treatise on Iconometry
- Kattida Chennool - Treatise on Architecture
- Nilamanai Chennool - Treatise on House Building based on quality of land
- Manainila Chennool - Treatise on Land based on the nature of house building
- Baniyal Chenool - Treatise on Astro Physics
- Perunata Chenool - Treatise on Divine dance
- Muligai Chenool - Treatise on Herbs
- Ganitama Chenool - Treatise on Mathematics
- Arakkala Chenool - Treatise on Ship Building
- Vinkala Chennool - Treatise on Space Ship
- Elisai Chennool - Treatise on Science of Music

உ

***MAYAN, in Aintiram** states that:*
"There exists an order in the Universe - subtle universe, and material universe. This power is attributed to the consciousness of the cosmic space as well as of the inner space of the animate beings. This consciousness, by its personal effort to express its own inner feelings, causes a kind of vibration or pulsation in the inner space, resulting in energy-grids. The vibration or pulsation, being the causal element of all these events, is called KAALA. This is actually the force of energy aroused by the inner consciousness, causing waves and contributing to the growth of living forms. This KAALA resides in all living beings. So, space is the offshoot of the vibration of the primordial energy. This is how the space was born. All forms of nature are manifest forms of subtle energy. For all to get manifested, the force is KAALAM. This wave-form frequency realm is the creative element of the universe."

There is no other treatise that we have in our possession today which is dated 10000 BC that depicts the secret of nature so explicitly and gives the unmanifested interconnections and mathematical proportions between everything in the universe.

As Dr. Ganapathi Sthapati rues "Unfortunately most of the Tamil experts don't understand the science in these texts and hence these books are idling without anyone extracting the truths from it". In the next few generations when we are less equipped in our own mother tongue we may lose this treasure completely just like the many we have lost so far.

Hence the first step we need to ensure is even if you are not equipped well in your mother tongue please make sure that your kids can read and write it well. This becomes our most important duty. And I find it worrisome when someone is indifferent to the fact that they can understand their mother tongue but cannot read and write.

With some significant efforts we can and we should prove that this man who said, **"Indians are the most ignorant about their rich past and scientific traditions."** is not wise any more.

Happy reading!

32. Time = Space Series. Part 32 – Conclusion

What I have understood in a nutshell is as follows:

- The inner being of both - the individual and the universe is luminescent consciousness which has the ability to become aware of it. This luminescent consciousness is called by various names but it is essentially the only existent Vastu in this world. This Vastu manifests itself as time, space and spatial forms.

- ***"Aanor aniyaan mahato Mahiyaan!"*** this means what is in the atom is the same makes up the universe. That which is present in the atom is the Space energy and the meaning of the word Akasa is stated as ***"Aa kashayati aakasa ha!,*** That which imparts light to others or make others shine is called Akasa". This means that Space Vastu is luminescent. I am stressing the aspect of luminescent again and again for you to start perceiving God as light.

- Pranava is the causal element which triggers this evolution process and this manifests as visual and aural forms. Pranava results in the primordial pulsation and this pulsation is triggered and sustained by a divine progression and proportion which is called **Absolute time**.

- Absolute time creates Absolute Space. These 4 things – Absolute time, absolute space, Aural Pranava and Visual Pranava combine together to form spatial forms and hence this universe.

- The Vastu as God's particle, its characteristics, the Pranava, the pulsation, the rhythm, the progression, its proportion and the resultant energy states – all the scientific facts, theories and processes are captured and codified as a Dharma which is called as Hinduism today. The science behind this subtle energy is reflected in various art forms that includes Poetry, Classical music, dance, Sculpture and building architecture.

- All these manifestations are harmonics of 5 and 8 differentiated by the Golden proportion / God's ratio to reflect the various states of consciousness. **This intrinsic order is called Dharma and the cyclic nature of this evolution is the Karmic code embedded in the universe**. **Nothing in this universe is superior to the Dharma and Karmic code.**

- The divinity manifests in us and the space within us is nothing different from the macrocosm. Consciousness which exhibits itself as a state of being because of our mental vibration can evolve to a higher plane to devolve back into the luminescent core.

- Sanatana Dharma popularly today known as Hinduism is the repository of this supreme secret not just as a philosophy but integrated with every stage of our life. This can be understood better only if we can give due respect and attention to the language and scriptures, our spiritual tradition interwoven with arts and architecture that contains these treasures.

- The temple culture & Idol worship that we have symbolizes the best of science known to human beings and we need to understand this and preserve this in the best possible manner.

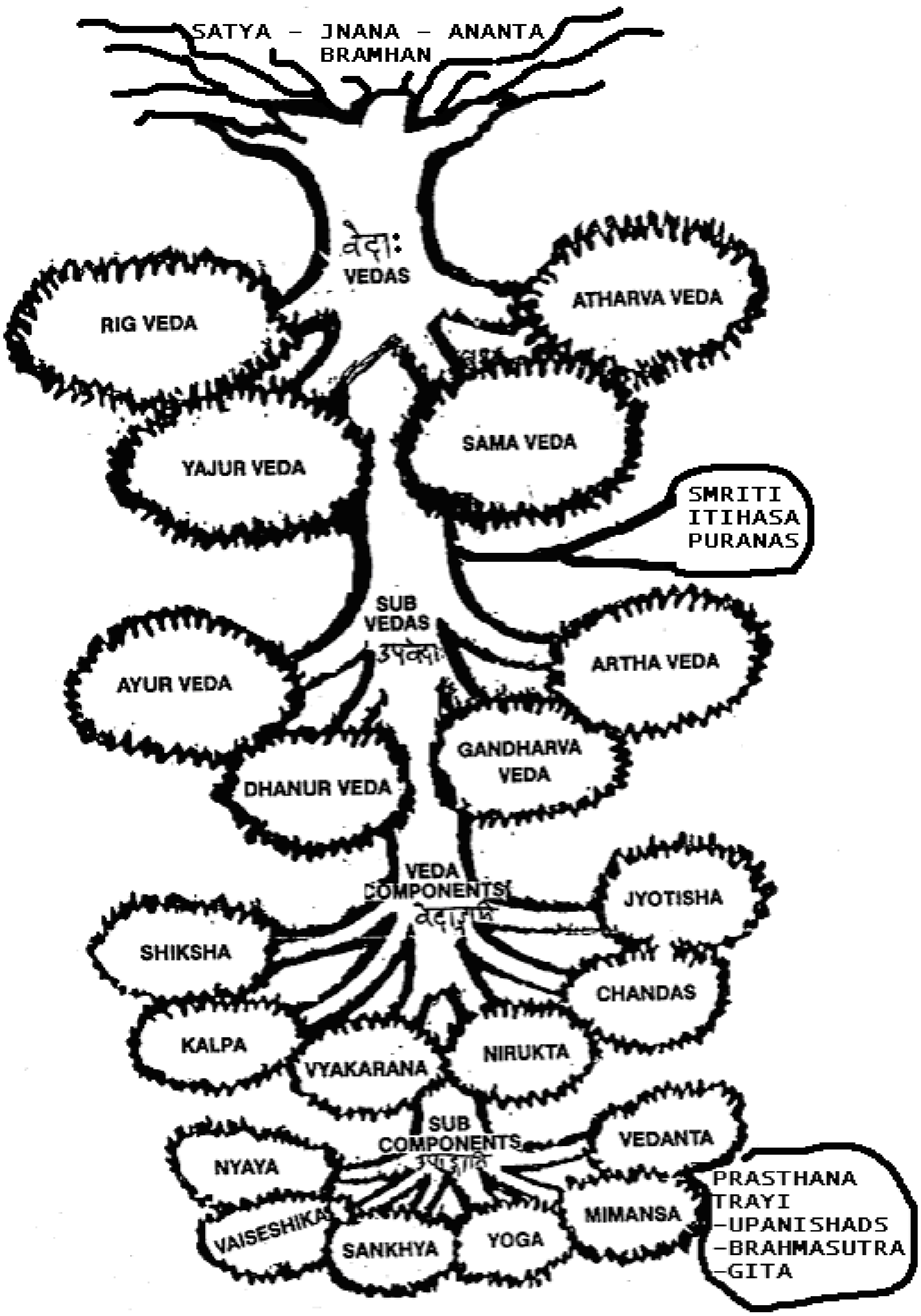
SATYA – JNANA – ANANTA
BRAMHAN
वेदाः
VEDAS
RIG VEDA
ATHARVA VEDA
YAJUR VEDA
SAMA VEDA
SMRITI
ITIHASA
PURANAS
SUB
VEDAS
उपवेदाः
AYUR VEDA
ARTHA VEDA
DHANUR VEDA
GANDHARVA
VEDA
VEDA
COMPONENTS
वेदाङ्गानि
SHIKSHA
JYOTISHA
CHANDAS
KALPA
VYAKARANA
NIRUKTA
SUB
COMPONENTS
उपाङ्गानि
NYAYA
VEDANTA
PRASTHANA
TRAYI
-UPANISHADS
-BRAHMASUTRA
-GITA
VAISESHIKA
SANKHYA
YOGA
MIMANSA

Figure 51: Vedic Knowledge as an inverted tree with Brahman as the roots signifying that everything has a common root. (Courtesy: Internet)

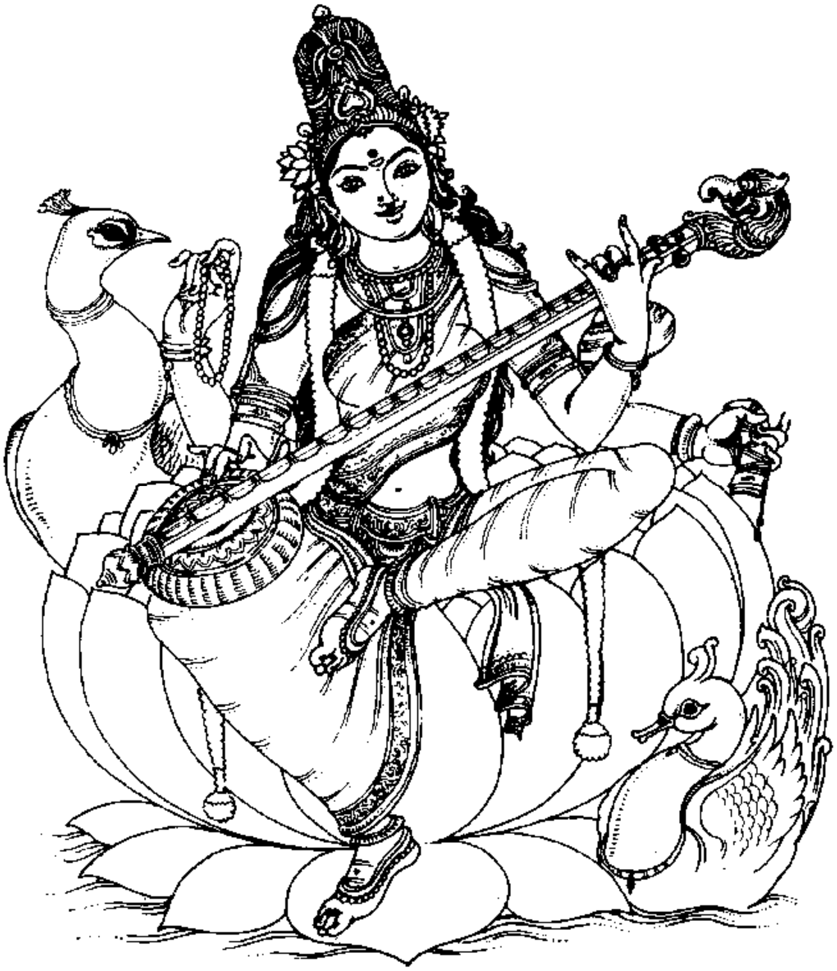

(Figure 52: Goddess Sarasvati. Courtesy: Internet)

She is the Goddess of speech, music and knowledge. She carries the rosary beads (akshamala) represents the alphabets of a language (Sanskrit), the musical instrument (Veena) and Vedas as palm leaf manuscript in her hands that denotes the knowledge. Thus we can see that idol worship signifies deeper scientific aspects of a common root between Language, Music, knowledge and arts.

Figure 52: Musical instrument Veena as a representation of Human Body (Sarira veena) (Courtesy: Internet)

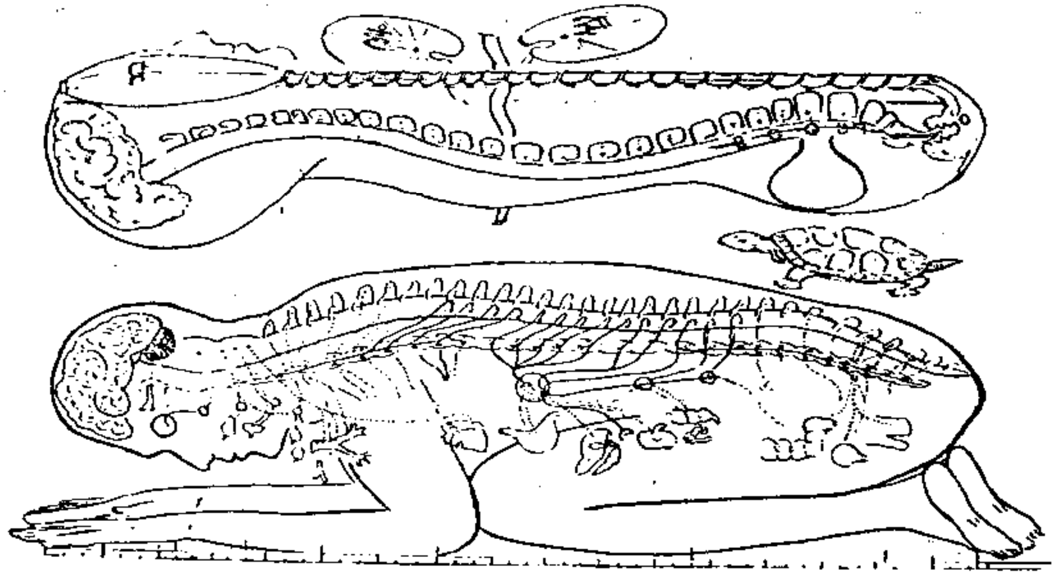

உ

I am happy to note the coincidence that this book has 32 chapters with 108 pages – both being key universal numerals. Chapter 19 (the golden mean of this book) deals with the Golden mean or God's ratio of nature is again coincidental.

I would like to thank you for reading this book. I would like to state that I am neither adept in any of the topics discussed in this book nor an expert in any of the arts including the languages Tamil or Sanskrit language.

A curious mind driven by my passion to understand the underlying aspect in all things and my liking to reading has helped me to compile this book. My exposure as a student to Chinese martial arts, Indian yogic studies, Pranic healing and Siddha philosophy helped me to connect of the points I had articulated. Any authoritative statements in these books reflect my belief in the topic or from the author / teacher from whom I subscribed it and it is certainly not reflective of my abilities.

I am open to constructive criticisms on any of the topics and I can be reached at the mail ID shared in the first page of this book. My only request to you is *"Please share this book or parts of this book without any hesitation to your friends and family and help us dispel the ignorance about rich ancient Indian culture"*.

I shall end this book with some of my favorite quotes of Dr. Einstein.

"My religion consists of a humble admiration of the illimitable superior spirit who reveals himself in the slight details we are able to perceive with our frail and feeble mind. That deep emotional conviction of the presence of a superior reasoning power, which is revealed in the incomprehensible universe, forms my idea of God. We still do not know one thousandth of one percent of what nature has revealed to us."

Satyam Param Deemahi!
May we meditate on the supreme truth!

TIME & THE TECHNOSPHERE

The Law of Time in Human Affairs

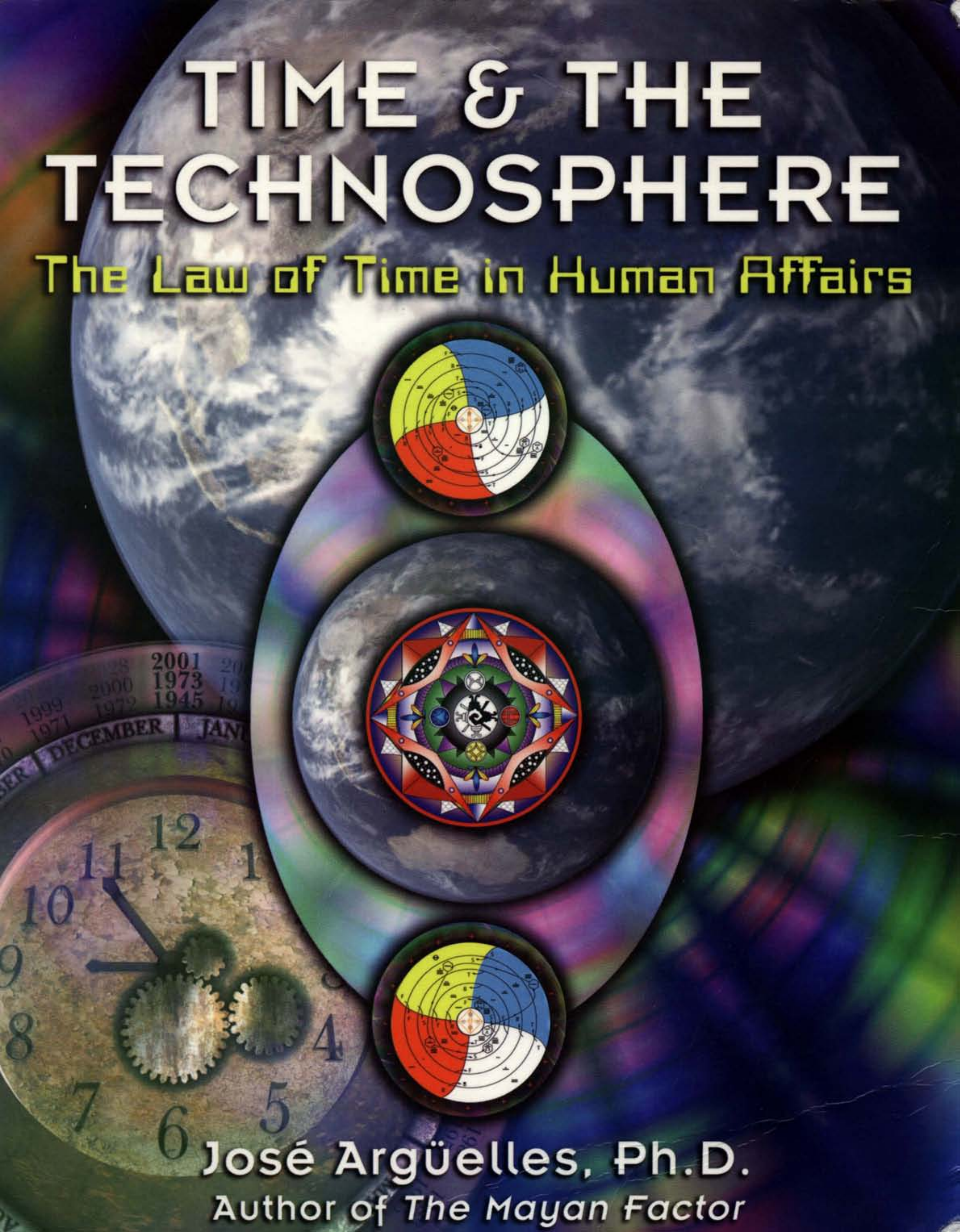
2001
1973
1945
2000
1972
1999
1971
DECEMBER
12
11
1
10
9
8
4
7
6
5

José Argüelles, Ph.D.

Author of *The Mayan Factor*

MAYAN STUDIES/NEW CONSCIOUSNESS $20.00

"Time and the Technosphere *makes a vital contribution to the great paradigm shift in human cultures that is now accelerating. There is a higher dimension of natural time inscribed in the cosmos and a natural calendar that would help humanity mature into the awakening global consciousness that is now essential for future sustainability and well-being. This book is an important contribution to the awakening event that is now emerging on a planetary scale.*"

ASHOK GANGADEAN, PH.D., Professor of Philosophy, Haverford College

In *Time and the Technosphere,* José Argüelles presents a groundbreaking study that distinguishes the natural time of the cosmos from the artificial mechanistic time under which we currently live. Argüelles defines the actual nature of time as the frequency of sychronization. Applying this Law of Time to an understanding of the entire system of life on Earth, he shows that in order to not destroy Earth's ability to sustain life, we must change our definition of time and adopt a natural harmonic calendar based on the 13-moon 28-day cycle. Until the creation of the Gregorian calendar and the 60-minute hour, most of humanity lived by the 28-day cycle of natural time. The adoption of artificial time has subjected us to a 12:60 time frequency that governs the entire global industrialized civilization—the technosphere. With the collapse of the Twin Towers on September 11, a fissure was created in this artificial technosphere, opening up the noosphere (Earth's mental envelope).

Humanity has a golden opportunity to leave the strife of the past and enter a time of peace by adopting a harmonious natural calendar that will repair the damages caused by the irregular tempo of technospheric time. Our last best chance to adopt this natural time and step into the bright new future promised by the galactic shift of 2012 is the Great Calendar Change of 2004, a new discovery based on the author's mathematical research into the Mayan calendar first begun in his landmark work *The Mayan Factor.* In *Time and the Technosphere,* Argüelles reveals the clear distinction between third-dimensional astronomical time and the fourth-dimensional synchronic order of the Law of Time, which holds enormous potential for the future of humanity.

PHOTO BY BORIS VLADIMIROVICH

JOSÉ ARGÜELLES, PH.D., works full time on behalf of the World Thirteen Moon Calendar Change Peace Movement and is president of the Foundation for the Law of Time. He is the author of numerous works, including *The Mayan Factor, Earth Ascending,* and *Surfers of the Zuvuya.* He lives in Oregon.

BEAR & COMPANY
ROCHESTER, VERMONT

ISBN 1879181991
9 781879 181991

Cover design by Peri Champine
Cover illustration by José Argüelles

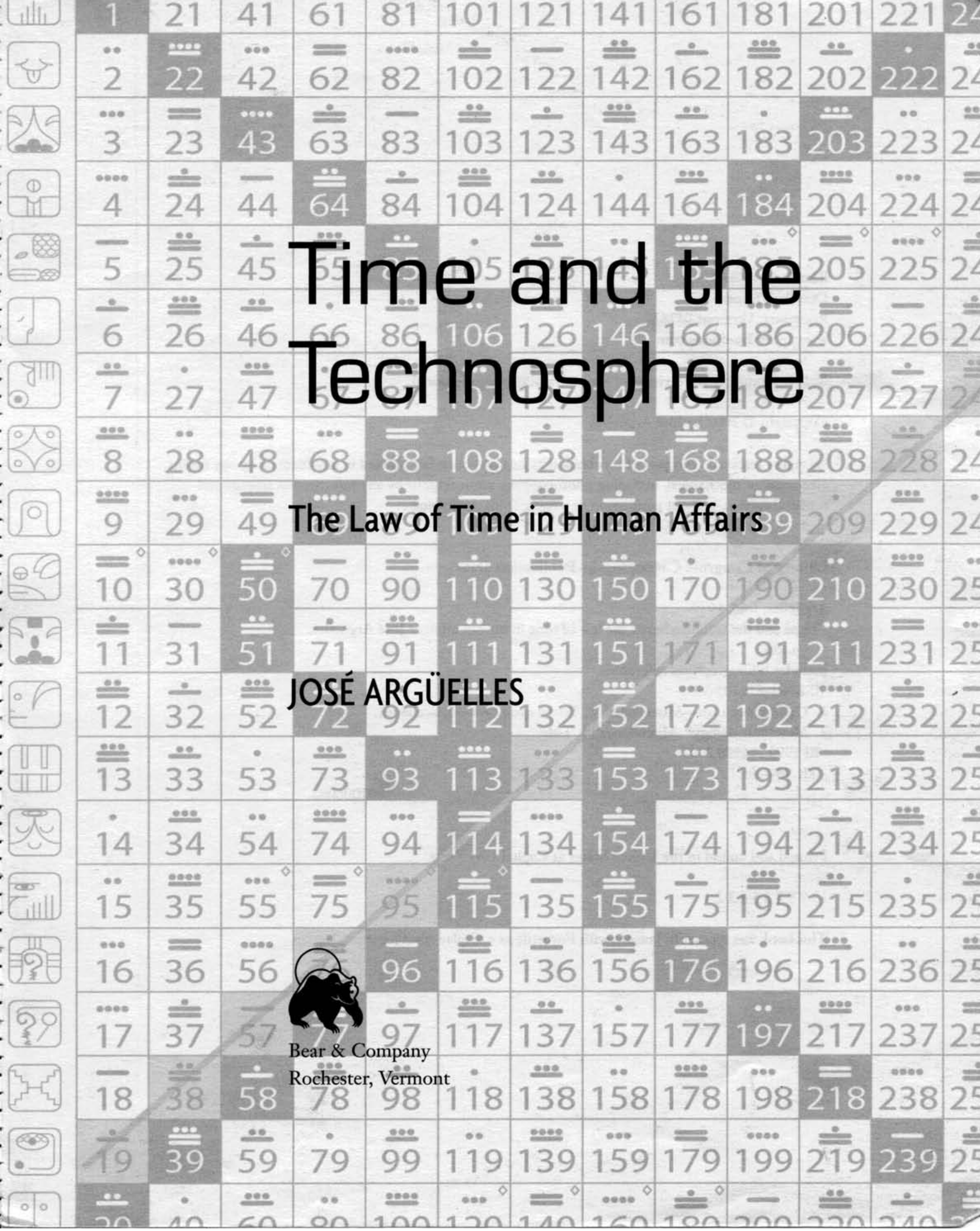

Time and the Technosphere

The Law of Time in Human Affairs

JOSÉ ARGÜELLES

Bear & Company
Rochester, Vermont

Bear & Company
One Park Street
Rochester, Vermont 05767
www.InnerTraditions.com

Bear & Company is a division of Inner Traditions International

Copyright @ 2002 by Jose ArgUelles

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Library of Congress Cataloging-in-Publication Data

ArgUelles, Jose, 1939-
Time and the technosphere : the law of time in human affairs / Jose ArgUelles.
p.cm.
Includes bibliographical references.
ISBN 1-879181-99-1
1. Calendar reform-Miscellanea. 2. Social evolution-Miscellanea.
3. Maya chronology-Miscellanea. I. Title.
BF1999 .A6398 2002
001.94-dc21
2002006082

Printed and bound in the United States at Capital City Press

10987654321

This book was typeset in Janson, with Eurostile as a display typeface

I

Time and the Technosphere

This book requires some attention from You. It is a complete, coherent, story for Humanity, on Earth, in the Universe. It is a radically different story than what is commonly disseminated by and through our global, dominating, culture. There is no need for You to *believe* this story, or any other. Trust yourself. trust your Self To say you must go into this with an open Mind, is a grand understatement. Start with trust. Simply allow the thoughts and concepts you read to "stew" for awhile. Make some headspace for them and have patience for what may grow there. It is a beautiful life...

NOTES ON THE ELECTRONIC PDF VERSION

- THERE ARE ERRORS! NOTHING MAJOR...ONLY SUPERFICIAL...
- NOTE THE COLOR PLATES (MENTIONED THORUGHOUT TEXT) ARE THE LAST EIGHT PAGES.
- THE PAGE NUMBERS ARE THE PAGE NUMBERS.
- USE ACROBAT'S BOOKMARKS FOR NAVIGATION!

ARGUELLES, JOSE - TIME & THE TECHNOSPHERE V0.2.PDF

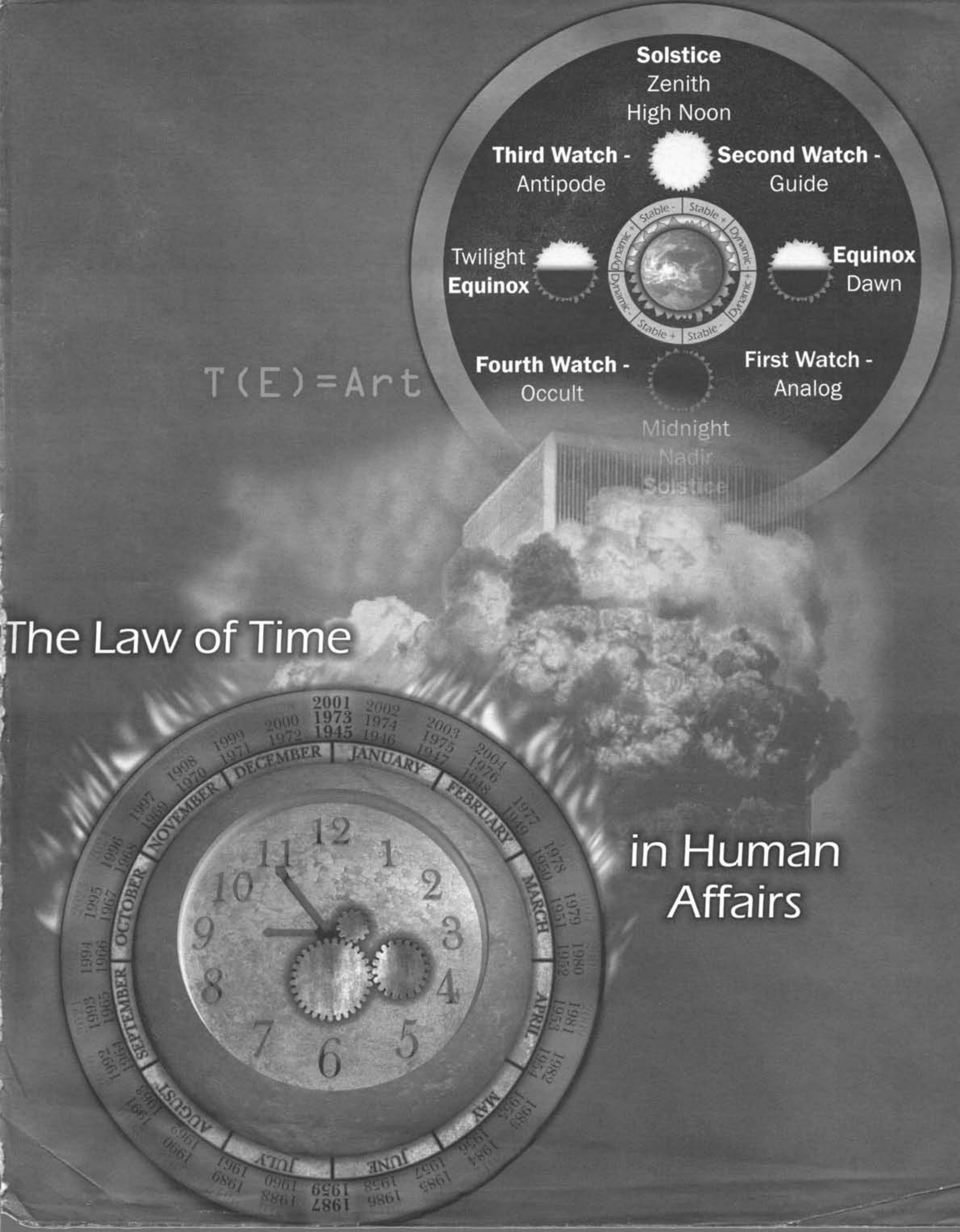
Solstice
Zenith
High Noon
Third Watch - Antipode
Second Watch - Guide
Stable -
Stable +
Dynamic +
Dynamic -
Twilight Equinox
Equinox Dawn
Fourth Watch - Occult
First Watch - Analog
Midnight
Nadir
Solstice
T(E)=Art
JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
2001 1973 1945
2002 1974 1946
2003 1975 1947
2004 1976 1948
1977 1949
1978 1950
1979 1951
1980 1952
1981 1953
1982 1954
1983 1955
1984 1956
1985 1957
1986 1958
1987 1959
1988 1960
1989 1961
1990 1962
1991 1963
1992 1964
1993 1965
1994 1966
1995 1967
1996 1968
1997 1969
1998 1970
1999 1971
2000 1972
12
1
2
3
4
5
6
7
8
9
10
11

The Law of Time in Human Affairs

This book is dedicated to Adam J. Lewis—at work on the eighty-eighth floor of the South Tower when his mother called him from Spain as she watched television reports of the first hit on the North Tower—and to all the others who were sacrificed on that day. May the truth prevail, that the greater mystery might be made known—forever.

Contents

Acknowledgments ix

Preface xi
A Brief Literary Autobiography

1 9-11, The Inevitable Event and the Nature of Time: Introductory Reflections 1

2 Preliminary Definitions: Biosphere- Technosphere-Noosphere as an Evolutionary Continuum 18

3 Time and Human Consciousness: The Law of Time-What It Is and Where It Came From 35

4 The Law of Time in Human Affairs: The Analysis of History 54

5 The Climax of History, the Fifty-six Years of Hiroshima-Artificial Time Runs Out 76

6 The Solution of the Law of Time: Get a New Calendar 99

7 Humans-Noospheric Chips: Consciousness and Reality Redefined 121

8 Making the Transition to the New Time: A How-to Catalog and Guide to the *Pax Cultura, Pax Biospherica* 145

9 The Advent of the Noosphere: A Prayer and a Vision 176

Epilogue 197
Networking the New Time

Appendices

I Summary Critique of the Gregorian Calendar with Which Every Proponent of the Thirteen Moon Calendar Should Be Thoroughly Familiar 200

II The World Thirteen Moon Calendar Change Peace Movement and the Discovery of the Law of Time: A Brief Chronotopology of Transformation 207

III Telektonon Prophecy of Pacal Votan 212

IV The Nineteen Code 229

Notes 240

Glossary 245

Bibliography 254

Acknowledgments

AS ANY AUTHOR knows, to write a book and see it through to completion, from initial inspiration to final conclusion, is never an easy task. Though we are all guided, if we so choose, by the Supreme Creator, still it is one's Earthly helpers that must be counted first, even before oneself-for without their aid, where would one be?

First and foremost of my Earthly helpers whom I wish to acknowledge is my wife and faithful companion of twenty-one years, Lloydine Bolon Ik, my constant helper and chief inspiration of all I do. The faith and trust she put in me at times when all we had was blind faith are a touching memorial to a soul who has always believed that love is the greatest truth.

Assisting us greatly during the time since we arrived in Oregon and have continued to dedicate our lives to the fulfillment of the prophecy of Pacal Votan and the Great Calendar Change is our devoted helper and assistant Stephanie South, the Red Queen, Red Electric Serpent, who conscientiously read and edited this text in its final form, helped in numerous details, and so kept me from ever bogging down.

Of equal trust and value is that one from the immemorial rainbow to whom I owe so much, Kin 19, David Rinck, "the stalwart," whose enhancements of my artwork are ingenious and awesome. All three of these people share one thing in common: absolute belief in and dedication to the Law of Time as a living reality that gives breath to all we do.

I must also thank the incomparable Spectral Skywalker, Mark Comings, whose mind has been of inestimable value to me through our many memorable background discussions on time in modern physics in relation to the Law of Time. In this regard, I would also like to mention Irina Eganova of the Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, for her rigorous approach to N. A. Kozyrev-may she smile well when she reads this book.

For their unconditional support I would also like to honor and thank Brian and Deborah Haight of the Foundation for the Law of Time, and all the many friends and beloved kin of the Planet Art Network everywhere, but especially to the Russian kin who so ably understood the Law of Time and opened up gold mines of knowledge and friendship in the Russian Academy of Sciences in Novosibirsk, as well as the great pristine valleys of Altai, where the shamans await the next installment of *The Mayan Factor*.

Spirituality is at the core of the biosphere-noosphere transition, and in this regard the beings I have to thank are without number, yet some especially stand out and deserve mention: my Tibetan teacher, the Sakyong Dorje Dradul Chogyam Trungpa Rinpoche, through whose cremation he entrusted me with all; my son, Josh, who opened the doors of the fourth dimension for me so that I would never be satisfied with less; and Mother Tynetta Muhammad, who showed me the way of the Holy Quran with the guidance of the mystic master FM, and who introduced me to the great work of Dr. Rashad Khalifa, whose breaking of the Quranic nineteen code led to his martyrdom, leaving me to fulfill what he revealed.

The creation of a book, the coming into manifestation of an artistic conception, is a complex process involving numerous other eyes and minds in order to bring the effort into its final published form. In this process two people have stood out deserving of mention. First of all, there is Michael Moynihan, White Galactic WIzard, who very patiently and carefully copyedited the book-always with an eye on maintaining and enhancing the original artistic vision-and who stood by the integrity of its vision through the entire editorial process. His work was complemented by the efforts of project editor Lee Awbrey, Blue Lunar Eagle, who took on an increasingly engaging role and personality as we brought this book to its final stages. I am especially appreciative for her inspiration to rework the Preface and Chapter One, which I gladly did, and which I feel speak more immediately to the reader of this book. Thank you both, and may the success of this book redound to your consummate care for the truth.

Finally I would like to honor and thank my children: Tara, the golden one, my grandson, SamuelJoshua Sheps, andJeff of Engine Company Number Nine in Phoenix; also Paul Mecklenburg, who has always held the faith in our work, and Heidi Mecklenburg, that her dream of the Unicorn Princess may become true once again.

▼

Preface

A Brief Literary Autobiography

Time and the Technosphere: The Law of Time in Human Affairs is essentially a prolonged set of meditations occasioned by the one most powerful and overwhelming event of our time, the 9-11, the Inevitable Event of September 11,2001. Subsequent events may overshadow this one in horror and magnitude, but it was this event in particular, the destruction of the World Trade Center towers, that sent a signal that something very profound had occurred to change the way we think of ourselves and our future.

In presenting these meditations and reflections in the form of a book, I am painfully aware of the limitations of this medium in the information age. Keeping in mind that the volume of information has increased five times in the past century-with much of that increase occurring only in the past decade-I am very reluctant to put anything down to be published in the form of a book. I was raised well on the writings of Marshall McLuhan and his definitions of media and how they affect our senses. It is altogether obvious that a book in your hand is not the same as reading a text on a computer screen. A book is actually something made to last. Of course, the computer has made it possible and easy for virtually anybody and everybody to write a book, and so we encounter one of the great hazards of the information age: the decline in discrimination and the criteria of excellence. Who can now tell what is worth reading or not? These reflections, too, are very much to the point of the grand theme of this text. If a book is meant to last, how long will this book last? It should last until the New Time has established a new order of reality on this now anguished Earth. Whether or not destruction and the total collapse of civilization will occur first, we do not really know. That is a matter left to the Divine Mind of God alone.

Nonetheless, it is with faith that we continue on this Earth at this time. A pressing voice told me to go ahead and to write this book, regardless-but then, this is how I have always done it. Those who are at all familiar with any of my earlier books, will find that *Time and the Technosphere* is but a continuation of themes that have provoked me since I first considered literature as a medium of communication. It is important to know where the author is coming from, and how he got to this particular point of reflection. In this regard, *Time and the Technosphere* is the organic result of a lifelong pursuit of a study of the nature of art, time, and history. This study has been woven together by a dialectical process that defines the evolving nature of consciousness. So, for those who are interested, herewith follows a brief literary autobiography, the better to understand my point of view.

My earliest book, *Charles Henry and the Formation of a PsychophysicalAesthetic (1972),* states the grand theme of a holistic science that unifies body and mind, sense and nature, the harmonic pursuit of which has been at the root of all my endeavors. This pursuit was amply demonstrated in my second book, *Mandala* (1972), which, as a root text of the counterculture, was an artistic, philosophical, and cosmic plea for a harmonic resolution to the process of history. It is a sign of the times that *Mandala* is no longer in print, but instead a book bearing the same title, but which does not even include the original Mandala in its bibliography, is in print with the same publisher. Amnesia is one of the side-effects resulting from information overflow with no discrimination. Unfortunately, cultural amnesia is always accompanied by an increase in trivialization, and thus we arrive at the mediocrity and sensationalism of the present moment.

My next book, *The Transformative Vision: Reflections on the Nature and History of Human Expression* (1975, 1991), took the aesthetic and unitive themes of the first two books to the next stage of my visionary analytical process, elaborating on the meaning of time, history, and cosmos. In rereading the later chapters of *The Transformative Vision,* I am struck by their prophetic accuracy, as well as their sense of impending tragedy. It is hard to believe that only 26 years have gone by since that book was first published and now we are plunged into the maelstrom of what was then thought to be the worst scenario possible. *The Transformative Vision* also defined human consciousness as the cerebral-neurological dialectic of psyche (aboriginal, primal intuitive art) and techne (civilizational, rational science), emphasizing that only in their absolute synthesis would there be a harmonic resolution to the problem of history. Seen in this light, the technosphere represents the extreme, one-sided triumph of techne, while terrorism is the expression of the absolute repression of psyche. The crucial question then arises: how will we ever attain a balance or harmony?

Investigation of the ongoing dialectic of consciousness was continued in my next effort, the philosophically synthesizing work entitled *The Feminine, Spacious as the*

Sky (1977), which was written and artistically produced with my previous partner and wife, Miriam Tarcov, who also co-authored *Mandala* with me. The feminine aspect represents a deepening reflection of the psychic, intuitive side of human nature. Expressed through a whole range of images, symbols, and philosophies that describe both the aboriginal root and the final flowering of human consciousness, it is this feminine aspect that has been smothered by the historical process, resulting in the technosphere. Why and how did this occur? These questions drove me to the next stages of my visionary investigations.

From the beginning of my pursuits I have been moved by my intuitive awareness of our intrinsic wholeness as humans, and the knowledge that this wholeness is a reflection of the wholeness of Earth, the solar system, and the cosmos itself. Although this perception of reality is nowhere within the domain of the shibboleths of Western scientific thought as it has developed since Descartes, I have also come to understand that Western scientific thought is highly provincial, racially biased, and-because it is not informed by a perception of wholeness-incapable of solving the actual problems of life. In fact, Western scientific thought has been one of the key factors leading the human race onward to its great act of self-destruction-a self-destruction employed through the instrument of technology.

In my next book *Earth Ascending: An Illustrated Treatise on the Law Governing Whole Systems* (1984, 1988, 1996), I was able to make a profoundly radical break with the restraints of the Western scientific and academic modes of thought. Combining the artistic and aesthetic qualities of *Mandala* and *The Feminine, Spacious as the Sky* with the philosophical and scientific concerns of *The Psychophysical Aesthetic* and *The Transformative Vision,* I define and describe the human situation as a function of the geology, or biogeology, of Earth understood as a whole system. Without knowing it, I had defined the human presence in Earth's geology in a way very similar to that of the Russian biospheric scientist V. I. Vernadsky in his analysis of the biosphere. Culture, civilization, and even technology are seen as features and functions of the larger evolving biogeological process of the whole system Earth, while time and consciousness are defined as functions of a larger planetary regulating system called the psi bank. In the analysis of *Earth Ascending,* the feminine psychic side of consciousness is referred to as the AC, Aboriginal Continuity, while the male, rational techne side is defined as the CA, Civilizational Advance. The historical process represents the increasing dominance of the CA to the final exclusion of the AC-the triumph of industrial man over the indigenous peoples.

The most critical moment in history was the atomic destruction of Hiroshima, August 6, 1945, when through its technology the impact of human thought permanently affected the geological makeup of the Earth. The events before Hiroshima are chronologically defined as P.A., pre-atomic, and those after as A.H., After Hiroshima. But more than defining the process of history within the context of Earth's cosmogeology, *Earth Ascending* also describes the post-historic future as a condition of harmonic synaesthesia-harmony of the senses-defined by a new scientific system called radiosonics. The key to this new system of scientific endeavor is synchronicity or synchronization of our mind and senses in time. The problem left open by *Earth Ascending* was the precise nature of time and the timing of the radical break in the historic continuum, now known as the Inevitable Event-the culmination and dramatic conclusion of a process begun at Hiroshima. After the Inevitable Event we are liberated into post-history and the possibility of the radiosonic science of the future.

In *The Mayan Factor: Path Beyond Technology* (1987), I was able to supply the basic ingredients for an understanding of time and the timing of the radiosonic future. Going beyond the parochial limitations of Western science and the self-inflated progressivist thinking of the Western mind in general (that states "anything that advances technology is an absolute advance over anything that was thought of or occurred in the past") I was able to finally bring to light the synthesis of many years of study of the time science of the ancient Maya. Here, in the mathematical calculations of the Maya with their unique vigesimal mathematics-counting by twenty and not by decimal ten-lay the basis of a vast and grand science of time. So grand is this science that its conceptions of time dwarf anything the Western imagination could conceive.

While the Western concept of time is linear and predicates an unceasing ascent into a progressively technological and mechanistic future, Mayan time science is predicated on the knowledge of time as the universal factor of synchronization. The Western conception of time is not in accord with the actual cyclical and synchronistic nature of time, and so the whole apparatus of Western civilization is doomed to reach its own breaking point: the Inevitable Event. Mayan time science posits a precise measure of the historical cycle, the most recent and climactic stage of Earth's evolution, as a time fractal of thirteen baktun cycles, each 144,000 days in length, spanning the JulianiGregorian years 3113 B.C.-A.D. 2012. According to Mayan time science, the efforts based on an erroneous time conceptualization are bound to fail prior to the cyclic end-point of 2012. Thus, the Inevitable Event is also a consequence of the failure to understand or to even acknowledge the actual nature of time due to a totally false and mechanistic timing sensibility. In this lies the cause of the apocalypse of Western thought and theology.

The problem left unresolved by *The Mayan Factor* was the precise methodology for changing tracks, for switching Western dominated global civilization away from the track of an erroneous, mechanistic time-based technology and toward the natural order of time-the path beyond technology. Here my efforts took me to a point beyond reason. Why or how? Let me explain. As an exposition of Mayan time science, in preparation for the 2012 ending of the thirteen baktun cycle of history, *The Mayan Factor* is predicated on the occurrence of a real time event, the Harmonic Convergence "global peace meditation" of August 16-17, 1987. The timing of the Harmonic Convergence was prophetic. It augured the fulfillment of the prophecy of Quetzalcoatl (947-999) known as the Thirteen Heavens and Nine Hells. The Harmonic Convergence completed the ninth and last Hell cycle, and ushered in the final twenty-five years of the thirteen baktun cycle. The Harmonic Convergence was a Mayan wake-up call answered by millions of earthlings.

Exactly forty-two years and ten days after Hiroshima, and fourteen years and twenty-six days before the Inevitable Event, the Harmonic Convergence was located by time to occur precisely when it did. Prophecy transcends reason because it is totally of time, which is of the fourth dimension; prophecy nevertheless affirms with unerring logic that which has been, and that which will be brought about, as being of the same karmic cause and effect stream.

It was this matter of prophecy in the aftermath of *The Mayan Factor* that brought my literary efforts and my own life journey to another stage of development. Prophecy comes with a mantle-if not, then you are classified as a madman. The prophecy of the Harmonic Convergence, as well as of the entire text of *The Mayan Factor,* bore the mantle of the prophetic stream of the Chilam Balam, the jaguar priests, the wizard knowers of the "night script." By correctly fulfilling the prophecy of the Thirteen Heaven and Nine Hell cycles, 843-1987, and by writing and putting together the mathematical codes of *The Mayan Factor,* and before it, of *Earth Ascending,* **I,** Jose Argiielles, had assumed the mantle of the prophetic tradition of the Chilam Balam. Mind you, this was nothing overt nor the least bit conscious, for little did I then know that my next life task was to decode the essence of the Chilam Balam prophecies and to define the time science of which these prophecies are a manifestation.!

The Harmonic Convergence affirmed the karmic repercussions of Hiroshima by bringing forward the ancient Mayan prophetic tradition of the Chilam Balam. By precisely locating this prophetic tradition in a contemporary moment -a media event, no less-the unlocking of its final prophecies was triggered for the purpose of completing the cycle of history: 2012 and beyond.

Like any major media event, the Harmonic Convergence was an enactment within the noospheric unconscious of humanity-the telepathic whole that resides in Earth's mental envelope. So it was that on June 23, 1987, the *Wall Street Journal* ran the front page story on the Harmonic Convergence. Yet due to the mental obscuration accompanying the final stage of global industrialization, the Harmonic Convergence was an event that the *WillI Street Journal* and the rest of American media conveniently forgot when it came to their review of the century in the year 2000. Today, just weeks after the Inevitable Event, the collapse of the World Trade Center towers, the words of that article are most prophetic: "Mr. Argiielles says the choice between a 'new age' and all-out destruction is ours, and we had better decide."2 Though the prophecy was discounted, true to form, the words of the messenger remain to haunt those who spurned the message.

The prophets and sages of the Maya, and of the New World in general, are a complement to those of the Old World, a further manifestation of the planetary dialectic of poles and hemispheres. If the judgment day apocalypse is the focus of the Old World prophecies, especially those ofIslam and Christianity, those of the New World similarly pinpoint an end time whose denouement is chronicled as the date 2012. It was the prophecies of the sage and culture hero, Quetzalcoatl (947-999), that brought about the Harmonic Convergence, while it was the wisdom of Pacal Votan (603-683) that informed the actual text and scientific time mappings of *The Mayan Factor*, and the image of his tomb adorns the cover of the original book. In a similar manner, it was the tradition of the Chilam Balam that created the prophetic vision of the synchronic order to be realized as the Law of Time.

I was hardly aware of the personally transformative factors at work within me after the Harmonic Convergence media swell began to subside, when two events occurred that augured a major change in my perceptions, activities, and literary output: the Wall Street stock market crash of October 19, 1987, and the death of my eighteen-year-old son,]osh, ten days later, on October 29,1987. In analyzing why the 1987 stockmarket crash happened, the lead story in *USA Today* about this event concluded by posing the question: Was it the Harmonic Convergence that somehow caused the stock market to crash? As I pondered why the answer to this question must be in the affirmative, as well as the implications of this analysis-remembering full well that the breaking story on the Harmonic Convergence first appeared in the *TVtlllStreet Journal-news* came of my son's sudden death in an automobile accident.

Synchronically, the stock market crash and my son's death were part of one event continuum. As tragic as the circumstances of his death were, my son had liberated me from any further compunction to lead a normal life, and to pursue, instead, this matter of Mayan time science and its calendars of harmony. In homage to my son, I completed what I first thought would be a popularization of *The Mayan Factor,* a book entitled *Suifers of the Zuvuya* (1988). Instead of being a popularization or even a sequel to *The Mayan Factor,* Surfers turned out to be a series of autobiographical tales of interdimensional travel-zuvuya surfing in the fourth dimension. This "flight of fantasy," was my literary farewell to the limitations of reason and a life of comfort. Inner destiny pushed me to the next stage in my pursuit of the unitive science-the discovery of the Law of Time.

This journey of discovery, the findings of which are the underpinnings of the present work *Time and the Technosphere,* took me and my wife and partner, Lloydine, to physical and mental places that were far from anything we had once deemed conventional. Because of this, I had ample opportunity to write and to create, and I did so, freely and without any restraints. I no longer sought any conventional means to communicate. I was now operating prophetically, on behalf of the noosphere-Earth's mental envelope.

There is a consistency to the stream of prophetic revelation, summarized as the codes of the Law of Time. This revelatory stream begins with *Earth Ascending* and *The Mayan Factor,* comes to flower with *The Dreamspell: Journey of Times hip Earth 2013* (with Lloydine Arguelles, 1991), and is completed by all aspects of the Telektonon Prophecy, including the Thirteen Moon/28-Day calendar (1993-2000). This prophetic stream is emphatically a galactic revelation of the same base of know 1-edge that produced the Mayan civilization of ancient Mesoamerica. The new revelatory stream is known as the Wizard's Count, or "reformulated Chilam Balam." It was this prophetic stream and its year-bearer count that established the validity of a fifty-two year cycle correlated to the Gregorian July 26.

The purpose of the Chilam Balam year-bearer count was to establish a basis for understanding that there exists a synchronic order of time, completely apart from what is usually referred to as the Long Count, the linear count of days of the thirteen baktuns of the cycle of history, 3113 B.C.-A.D. 2012. After a long break in the tradition of the Chilam Balam, the year-bearer count was prophetically reformulated to prepare for the Harmonic Convergence and the subsequent revelation of the Telektonon prophecy of Pacal Votan. As the new Wizard's Count, this count was intuitively recalibrated precisely so that July 26, 1987, would be White Galactic Wizard; July 26. According to *The Mayan Factor,* July 26, 1987: White Galactic Wizard, marks the first of the twenty-six years of the Harmonic Convergence. This date marks the conscious ascendance of the prophecy of galactic culture and the coming of the Earth Wizards-the race of galactic wizards on Earth.

In 1988, the year following the Harmonic Convergence, in attempting to create an everyday "Mayan calendar," I came upon the next stage of the discovery of the synchronic order. On the Gregorian calendar, 1988 is a leap year. I perceived that to accommodate February 29 would throw the year-bearer count off by one day, ruining the harmony of the system. Reviewing the Chilam Balam texts, it was clear that leap year was never considered, much less counted in the year-bearer count. Without an intercalary leap year date, the four year-bearers-Seed, Moon, Wizard, and Storm-coded by the numbers 1-13, remained in sequence, maintaining and assuring the harmony of the system, where every fifty-two years equals seventy-three 260-day spins, no two days are the same, and the entire 18,980 days of the cycle repeat every fifty-two years in an upward moving spiral of time. For the first time in my work, the issue of the Gregorian calendar as a system came into question.

Following this observation came the discovery of the mathematical codes governing the synchronization of the 260-day cycle with the 365-day cycle, the consequent discovery of the artificial 12:60 and the natural 13:20 timing frequencies, the recognition and resurrection of the Thirteen Moon/28-Day calendar, and the full exposition of the Dreamspell codes of fourth-dimensional time. I presented these discoveries in *Dreamspell: Journey of Timeship Earth 2013.* The Dreamspell, being the manifestation of the knowledge base of galactic culture, completely encodes the Wizard's Count. *Dreamspell: Journey of Timeship Earth 2013,* the presentation of the codes of fourth-dimensional time as a "tool kit" with an accompanying "script" and guidebook, is a break of such radical form from the prevailing paradigm that, as of today, it is still unfathomable to many people.

The galactic Dreamspell tool kit is a primary tool for the galactic culture that is just now beginning to unfold in the wake and shadow of the collapse of the Twin World Trade Center towers. This text prompted me to write two more books. The first, *Arcturus Probe: Tales and Reports of an Ongoing Investigation* (1992, 1996), is a deep reverie that picks up where *Suifers of the Zuvuya* left off and documents the interstellar, interplanetary psycho-cosmological roots of the Dreamspell of fourth-dimensional time. The second book was written as *A Treatise on Time Viewed from its Own Dimension,* but first published in English as the *Call ofPacal Votan: Time is the Fourth Dimension* (1992, 1996). In this highly scientific, mathematical treatise I was able to accommodate the Dreamspell codes of the Law of Time to an analysis of the biosphere and the biosphere-noosphere transition. This work also represents my first integration of the actual thought and principles of V. I. Vernadsky.

Through the efforts that constituted the discovery of the Law of Time, much of which was assisted by my partner and wife, Lloydine, I was finally able to begin to understand the methodology for accomplishing the paradigm shift-or psychic pole shift-of which humanity was so in need, at least from the perspective of the biosphere. By providing the mathematical basis for establishing the "biomass constant," Seventy-three 5 day subcycles correlated to the measure of the solar year, I had found a way of grounding the fourth-dimensional mental order of time, the synchronic order, in the actual third-dimensional biogeochemical cycles of the biosphere. This was like casting a harpoon of time from the noosphere and landing it in a precisely measured form in the annual turning of Earth's biospheric mantle. And this was accomplished precisely because of the intrinsic harmonic cycles and patterns of the DreamspelllWizard's Count.

For this reason, by 1993, it was incumbent upon me to test the hypothesis of the Law of Time that the human race was destroying its biosphere due to operating according to an error in time, and that only a shift in its experience of and perception of time would save it from such destruction. This shift in time is defined as the calendar change. The Law of Time presents such a radical critique of Western science, of its concepts of time, and of the civilization which it supports, that I could not tackle the problem from its center, but from its periphery, moving inward to the point of the Inevitable Event. Beginning in Latin America, then spreading to Japan, Western Europe, Russia, and finally North America, we established a movement to change the everyday perception of time, instilling profound meaning behind the effort to reform the calendar. In this way, we took on the responsibility of organizing the World Thirteen Moon Calendar Change Peace Movement.

Part and parcel of this, beginning in 1993, was the decoding of the prophecy of Pacal Votan, the inspiration of *The Mayan Factor*. Known as the Telektonon, Earth Spirit Speaking Tube, this prophecy is the full realization of the Law of Time and the synchronic order. As much science as it is prophecy, the Telektonon defines the Tower of Babel as the root of the error in time, and the Inevitable Event as the necessary result of not relinquishing false time. At the same time, the Telektonon provides codes and tools of analysis which are in complete accord with an understanding of the technosphere-noosphere as an evolutionary continuum in which the Inevitable Event is an act of moral consequence and natural law revealing the basis of a new cycle of harmonic order.

I would also be remiss at this point if! did not inform the reader that, by prophecy's prerogative, this text is also literally punctuated by a perspective that can only be described as "Quranic." Late in 1993, my immersion in the prophecy ofPacal Votan became sychronically linked to my immersion in a study of the Quran, or Holy Quran as it is sometimes called. I have taken seriously the claim that the Quran is the final revealed text (A.D. 610-632) for humanity. I have studied it following the dictum: "Quran, the whole Quran, and nothing but the Quran,"3 as the basis of genuine Islam-submission to the will of God. I have found it more than sufficiently true that the Quran is a scientific text, and contains nothing that contradicts science. More profoundly, my study has given rise to my own personal ethic: duty to God is the highest cause. It is the cultivation of this attitude, combined with many years of practice of Buddhist meditation, that have brought me to my present point of view, which is visionary and spiritual throughout-and, as a result, has endowed me with my own view on the current global dialectic.

That the new, and final, global dialectic is found in the tension between the forces of globalization (materialism) on the one hand, and that of the civilization of Islam (spirituality) on the other, is now a well-documented fact.4 Of course, generally speaking, in the West, we see it only from the materialist side of the dialectic. The perspective which informs me, and consequently this book, is definitely from the other side of the dialectic. Though in no way condoning terrorism- I am a non-violent pacifist by persuasion-the motives of the terrorists stem from a philosophy and worldview with which I am in sympathy. Yes, the Inevitable Event was also a profound theological moment, and what we are witnessing in the present World War is the battle between the religion of choice and the religion of submission. We cannot necessarily assume that the religion of choice will prevail. In fact, from the perspective of the biosphere-noosphere transition, it is the religion of submission that is most in accord with the spiritually-beckoning future.

As a preface to this book, I have found it mandatory to write this brief literary autobiography, often couched in a terminology that jumps ahead of the story, solely to demonstrate that the actualization of the noosphere corresponds to the stages of attainment and transcendence of one human's life journey in pursuit of the truth. Only in this way could I cultivate the noospheric perspective, voice, and point of view that was necessary for the writing of this text. In this perspective, the noosphere is a condition of non-anthropocentric consciousness. My journey, thus documented, is merely one way the noosphere has found to become conscious, asserting the Law of Time's purposive axiom: to make conscious what had been unconscious. The voice and method of the noospheric perspective are inseparable from the topic of this book, *Time and the Technosphere.* By analyzing the timing and structure of the technosphere, the noospheric point of view becomes conscious. WIthout understanding the nature and relation of time and the technosphere, there would be no truly sane way to comprehend the meaning of the Inevitable Event.

Crystal Moon Kali 11,
Kin 2: White Lunar WInd
Yellow Solar Seed year
Cascadia, North America

1

9-11, The Inevitable Event, and the Nature of Time

Introductory Reflections

Woe! Woe, O great city,
O Babylon, city of power!
In one hour your doom has come!
Woe! Woe O great city, dressed in fine linen
purple, and scarlet...
In one hour such great wealth has been brought to
ruin ...
"Woe! Woe, O great city,
Where all who had ships on the sea became rich
through her wealth,
In one hour she has been brought to ruin! ...
With such violence the great city of Babylon will be
thrown down!"
ST.JOHN OF PATMOS, BOOK OF REVELATIONS,
CHAPTER 18, *THE FALL OF BABYLON*

9-11. The Twin Towers of the World Trade Center gone in a veritable twinkle of the eye. Tragedy. Horror. Terrorism. The whole world brought to a halt. The aftermath? Patriotism, war, and an all-pervading uneasy feeling. What happened? What really happened? Was this the beginning of the Apocalypse?

Amidst the distraction of chasing after other people with smart bombs in faraway places like Afghanistan, and making everyone at the airports feel like potential terrorists, there are no answers forthcoming for what really happened. There is a good reason for that, too: The destruction of the Twin Towers of the World Trade Center and the attack on the Pentagon marked a point beyond the comprehension of the current belief system in every way. To say that the event was mind-blowing is an understatement. But if it was mind-blowing-which it truly was-that means that a whole belief system was knocked off of its foundations, and only a new belief system can supply the answer. A higher, more all-encompassing belief system-a new paradigm-is what is needed if we are to really understand what happened on 9-11.

Belief systems, believe it or not, are functions of time. What you believe defines the time you are living in. Any belief system is held in place by the calendar and sense of time in which it is encoded. 9-11 means September II-it also means 911 emergency, planetary emergency in this case. "9-11 we will never forget" was a common slogan appearing on bumper stickers and store windows after the event. The words or concepts "Tuesday, September 11," "9-11" and even "911" are all functions of the same belief system. So is "2001," the first year of the highly touted "third millennium." But the third millennium of what? The third millennium since the birth of Christ. This means that the belief system whose mind was blown was the Christian millennial belief system encoded in the Christian calendar, the Gregorian as it is properly called. Does this mean that the foundations of 2000 years of belief were blown away, shattered with the terrifying collapse of those Twin Towers? Easily that many years, maybe as many as 5,000 years, or perhaps even the entire history of civilization as we know it came unhinged that fateful day, the day of the 9-11, the day of the *Inevitable Event,* to borrow a phrase from the Holy Quran, the day the Apocalypse announced itself, showed up on the morning television, quite without warning. (See plate 1, The Invevitable Event and the End of History.)

In meditating upon the meaning of the Inevitable Event, it is natural to ask: What made this event so inevitable? It is not just from the perspective of American Foreign Policy, or that of its economic free market monetary politics, that this event was inevitable. More profoundly, this event was so inevitable because it seemed to smack of some form of retribution for the breaking of fundamental laws of nature, or even of Divine Law. And isn't the retribution for transgressing Divine Law a form of apocalypse, a standing naked, a revelation, an evocation of the fall of the Tower of Babel, no less?

Yes, that is why it was so shocking-it was the Inevitable Event because the apocalypse is as inevitable as it is without warning. But apocalypse- a word no one wanted to use to describe the event, because to use it would be to admit that it was all over-is only one side of the coin. There is another side to the coin of the Inevitable Event-and that is what this book is about. There is something beyond the Apocalypse, there is something larger going on, and that also scares people. Things are out of their control. Isn't this why, instead of looking at it, or contemplating the event, a policy of war was immediately instated as a principle of revenge-for what and against whom? What is going on is time-time is always going on, or time is always coming together. Or time is always changing. The apocalypse, too, is about time-the end of time-which is even more frightening to most people. But the end of whose time? The new paradigm we need to explain 9-11 should be a paradigm that is all about time.

I who speak know, because I have spent most of my life studying time, and the effects of human time which you call history. For many years now I have been living an experiment in time, but a time most of you do not know. So I have some answers that will put 9-11 in perspective for you. I am the stranger from the other side of the wall of mechanized time. I have come back to make you familiar with the time your clock and your calendar shield you from. Because I know about the other time, the larger time, the cosmic time that governs the cycles of all that exists, I can tell you that 9-11 was the Inevitable Event. It had to happen precisely when it did, not because nineteen terrorists planned it that way, but because the terrorists themselves were unconsciously programmed by the timing of Earth itself. Earth's timing process represents evolutionary forces as well as effects irreversibly set in motion by the human's own unconscious programs, programs governed by the human misperception of time and of its role in the cosmic order.

Yes, Earth has its program of which our human history and all of its endeavors are only a part. This larger meta-program of Earth is called the biosphere, the scientific designation given to what is commonly and inadequately referred to as "the environment." As such, the biosphere is a whole system composite of the sum of life and its organic and inorganic support systems spread out over the surface of Earth, hence bio = "life" + sphere = "having the form of a globe." In a word the biosphere is a vast but fragile process which has been evolving itself for some two billion years on this little planet as it spins around its local star, the sun. But we humans, who only in 1969 first saw ourselves as a planetary being-the whole Earth beamed back at us via television by a rocket hurtling toward the moon-we humans have been generally so self-involved in our own interpretation of things that we have missed seeing the larger picture of which we are a part. This larger picture is the entirety of the biosphere which encompasses the evolution of life on Earth as a whole system process. The amazing thing is that most of us have never even heard of the biosphere much less even know that we are a part of this whole system governed by laws and principles which we, in our self-made arrogance and ignorance, have but little understood.

From my experiment in time, seeing and studying the human as part of the larger self-evolving fabric of the biosphere, I came to the conclusion that the human is living in a time apart from the rest of the biosphere-an artificial time whose climax and termination is inevitable, for nothing artificial can withstand the force of truth. If the human is living in artificial time, the clock is an artifact whose system of measure has nothing to do with natural cycles but is a totally abstract standard, then there must be something called natural time. I will go even farther and state that not only is there natural time, but that there is a law governing natural time, and that is the Law of Time. Just as Newton only discovered gravity some 300 years ago, though gravity has always existed, so the Law of Time has always been in operation, even though it was just recently discovered. It is the Law of Time that governs the entirety of the biosphere and everything else in the universe according to principles which are only dimly known to us, principles like synchronicity and telepathic instantaneity. Yes, to understand what really happened on 9-11 you have to understand the Law of Time and how it governs the biosphere-of which our rampant human civilization is inescapably a part.

The Law of Time is formulated very simply, and in some people's way of thinking, rather unscientifically as T(E) = Art, "energy factored by time equals phenomena in the material world represent some state of energy, and every state of energy is governed by time, the resultant product of which is always something beautiful or elegant. Have you ever seen an ugly sunset? A hideous flower? Even if you examine a scorpion with some objectivity you will be amazed at the flawless and elegant manner in which its parts are organized. Yes, all of nature is organized by time to produce in you the sensation of beauty. And time itself, well, believe it or not, time is a frequency, and a frequency is not measurable by a clock. The Law of Time states that time is the universal frequency of synchronization. It is the nature of time to synchronize and to maintain all things in a condition of synchronization. Synchronicity, then, is the experience of real time. When we say that time is a frequency, we can be more precise and say that time is a universal constant expressible by the mathematical ratio 13:20. That is, the 13:20 ratio is the frequency of synchronization.

Of course, most people have never even heard of the Law of Time much less comprehend what the 13:20 ratio might mean. But then, that again gets to the root of the problem, or rather, the dilemma: If humans are living in their own time apart from the rest of the biosphere which is governed by the natural timing frequency, how could they ever know about the Law of Time? The humans are all a little like Mrs. Malaprop in the play by Moliere who learns to her surprise that she has been using grammar all of her life-though it is much more serious than that. So we humans are with the Law of Time-it has been operating throughout history without our awareness of it. Because of this, its way of making itself known to us was the infamous 9-11. Yes, in the analysis of the Law of Time, the 9-11 marked the end of artificial time.

Living in ignorance of the Law of Time and of the true nature of time, the human species constructed its own concept of time which is based on the clock with its 60-minute hour and the Gregorian calendar with its maddening array of unevenly numbered months. Hence, the paradigm of the artificial timing frequency is defined by the ratio 12:60 (twelve-month calendar, 60-minute hour) -in contrast to the 13:20 ratio of natural time. If the artificial mechanistic and irregularly measured time sets the human race apart from true or natural time, does this not also establish the fact that the human race is living an error in time? Only by living such an unconsciously assumed error could the human race arrive at such an apocalyptic moment as the Inevitable Event-inevitable because any deviation from the truth has an inevitable moment when the truth rebounds in some dramatic or even apocalyptic way. Is it possible that the entire construct of modern civilization, so devastating to the biosphere, is a function of this error in time? Is modern civilization like a time warp, a bubble of artificial time that suddenly got popped on the 9-11?

Back in the 1990s, now seemingly so long ago, I was traveling around the planet on behalf of the Law of Time and the biosphere, and making my observations regarding the effects of artificial time from a planetary whole systems point of view. I saw that even though the human species had evolved - or devolved, as the case may be-into a cultural hybrid which I call the planetary human, the object of the much touted "globalization," I saw that the consciousness of this cultural hybrid was anything but planetary. This schism between the propaganda of globalization and the actual sectarian and fragmented state of consciousness of the hybrid human was further exacerbated, I observed, by the potently unconscious effects ofliving an error in time. Ironically, if the Inevitable Event left many of us in the dark, it has brought the Law of Time into the light. To elucidate let me quote here from a text I wrote in 1996, and which I think will give us a point to expand upon in order to bring the Inevitable Event into sharper focus, while allowing me to define some of my terms:

"Through an error in time, the human species transforms its artificial construct, civilization, into a global technosphere: the sum product of industrial mechanization of its biological functions. Because of human adaptation to the irregular mechanized 12:60

timing frequency, the technosphere runs counter to the laws of the biosphere, creating a magnetic instability between the primal crystalline and vital organic processes."2

The current explosion of the world crisis, stemming from the destruction of The World Trade Center Twin Towers, demonstrates the hypothesis of the Law of Time that states: operating on artificial and mechanistic timing standards will cause the human to deviate from natural time to the point of its self-destruction, the end result of the magnetic instability referred to in this reference. But here we have some further definitions to make. Let us start with a definition of the technosphere, since that word is featured so prominently in the title of this book, *Time and the Technosphere: The Law of Time in Human Affairs.*

While the term biosphere defines the entirety of life as a single coherent unity, inclusive of its inorganic or semi-inert support systems such as the atmosphere and the hydrosphere, technosphere defines an artificial sheathe or membrane held together by industrial technology as a whole system. This technological mantle is the sum of the processes devised by humans resulting from the application of certain scientific and economic principles, especially during the last half century. The effects of the technosphere are to supplant the organic processes of the biosphere with totally industrialized ones, resulting in a severe imbalance in the biosphere. This imbalance, the "magnetic instability between the primal crystalline and vital organic processes" refers to the disruption of the delicate interaction between the inert (crystalline) and biological processes and structures which constitute the vibrant mechanisms and interactive cycles of the biosphere. Phenomena such as global warming and over-population are symptomatic of the technospheric disruption of the natural order of the biosphere.

The term technosphere, however, is actually just the middle term of a larger process defined as the biosphere-noosphere transition. According to the principle figure behind the elaboration of the laws of the biosphere, V. I. Vemadsky (1863-1945), as an evolving structure, the biosphere is inevitably being transformed into a new geological and evolutionary condition, the noosphere-Earth's mental envelope. In Vemadsky's analysis, the entirety of the biospheric process can be described as a single biogeochemical unity. According to Vemadsky, due to man's transformation of nature since the industrial revolution, this biogeochemical unity has entered a state of combustion-an accelerated process of transformation. It is this rapidly accelerating geochemical combustion-the net effect of industrialization and its waste -that accompanies an evolutionary shift, the transformation of the biosphere into the noosphere. But between the biosphere and the noosphere there is the technosphere, the intervening medium of transformation.

So when we speak of the biosphere and the noosphere we must take into account the technosphere. In actuality we must speak of the biosphere-technosphere-noosphere transition. In this context, the noosphere is the condition of the human mind purified of the error of time. And if the noosphere is the Earth's mental envelope, it could only be so because of a unified state of mind, the transformation of the hybrid planetary human into a genuinely spiritualized planetary being, the mind of the Earth no less-something that seems phenomenally utopian from the perspective of today's evening news. How could this come about, this noosphere of a telepathically unified human consciousness? Again, the Law of Time supplies genuine answers. To begin with, it says, change your calendar, get back to living in harmony with the natural cycles of the rest of the biosphere-and the universe. If you don't, well, the destruction of the Twin Towers was just the first stage of the Apocalypse. But we still have a choice. The other side of the coin of the Apocalypse shown to humanity on prime time television on the 9-11 is entry into the noosphere. But we must first understand what is really going on, and then act intelligently.

The mental climate in America is particularly difficult and strained now. In order to maintain a positive and constructive attitude of mind during this terrible time, I decided to write this book. As you can see, this book takes for its point of departure an analysis of the destruction of the World Trade Center (the actual hub of the technosphere) as the Inevitable Event defining the dynamic of the biosphere in its process of transforming into the noosphere. The analysis of the event in this light will give ample opportunity for presenting the key ideas of V. I. Vernadsky, in particular the concept of the biosphere which he so single-handedly amplified and defined, as well as the cosmic ideas of other Russian thinkers which are little known in the West. I have already written about and given much thought to this matter in general, summarized in part in the First Planetary Congress of Biospheric Rights, held in Brasilia, Brazil in 1996, but nowhere have I written for a wider public audience concerning the nature of the biosphere-noosphere transition.

I feel it is especially important at this time to present an analysis of this event in a scientific context that is beyond nationalism and ideology, so that human beings can begin to understand that they are a function of the biosphere, but not necessarily its controlling mechanism. Of course, the Law of Time is the key factor involved in this analysis, and understanding the biosphere in the context of the Law of Time is absolutely critical for establishing the basis of an entirely new worldview and hierarchy of values-the much awaited new paradigm.

Like my thoughts on the biosphere-noosphere transition, my investigations of the Law of Time have also not as yet been presented to the general public. In fact, you will find scarcely anyone who has even heard of the "Law of Time." This is because my researches and investigations over the past decade which resulted in the discovery of the Law of Time were not conducted in any formal academic way, nor through any corporate or government sponsored grant or officially recognized research institution. The discovery of the Law of Time has been, nonetheless, a rigorous, life-consuming pursuit. The laboratory for its study was to be found in a phenomenological and anthropological investigation of the effects of mechanized time on various population groups absorbed in what is called "modern civilization."

The premise of this investigation is simple: it is an undeniable fact that the time of human civilization, governed by the clock and the Gregorian calendar, is not the same as the time of the rest of the biosphere. Artificial time is not the same as natural time. Or rather, artificial time establishes a frequency which governs the human species totally apart from the timing frequency of the rest of life. What are the effects of artificial time and how do they define the nature of modern civilization? The answers to these questions open a perspective that has been scarcely considered. Yet this perspective is so clearly an essential factor in the description and definition of the human condition, that, in the wake of the tragic events and their ominous aftermath, it now begs to be presented in a proper manner. Hence this book. Suffice it to say that the Law of Time in every respect answers the question posed by my previous books and summarized in the single query: how will modern man escape the fatal consequences of the one-sided mechanization of his biological operations and the imminent destruction of his life-support system?

Defining a planetary whole systems approach to the analysis of time in human affairs, the first postulate of my book *A Treatise on Time* (1996) concerning the principles and nature of time as the fourth dimension is worth quoting here, since it introduces the fundamental problem informing the investigation of Time and the Technosphere:

> Just as air is the atmosphere of the body, so time is the atmosphere of the mind. If the time in which we live consists of uneven months and days regulated by mechanized minutes and hours, that is what becomes of our mind: a mechanized irregularity. Since everything follows from mind, it is no wonder that the atmosphere in which we live daily becomes more polluted, and the greatest complaint is: "I just don't have enough time!" Who owns your time, owns your mind. Own your own time and you will know your own mind.3

It must be understood that the events of 9-11 were so historically unprecedented, so mind-bogglingly dramatic, that they require a large lens through which to fully interpret their meaning. The Law of Time is such a lens. Quite simply, the destruction of the Twin Towers defines the limit of artificial time. If this is so, what can be expected following this world-shattering event? Because the Law of Time began with the premise of defining the difference between artificial time and the time of nature, the larger context of the biosphere was resorted to almost immediately. Being the study of an ongoing and evolving phenomenon, the investigation of the biosphere by its principle researcher, V. 1. Vernadsky, pointed to another imminent evolutionary stage: the noosphere, the "mental envelope of the planet." However, an intermediate stage was also perceived by researchers who followed Vernadsky, and it came to be called the technosphere, "the technological envelope of the planet." The end or limits of the technosphere define the birth or beginning of the noosphere.

With the events of 9-11, this is no longer just a theoretical concept but a vivid reality. As a species we are now undergoing this dramatic evolutionary moment, otherwise known as the biosphere-noosphere transition. It is for this reason that this book also presents novel definitions of time and pragmatic conclusions that can be employed by anyone if they so choose. If the root of the technosphere is artificial time and its proliferating mechanized constructs, then in the collapse of the technosphere, an event ordained by the laws of nature, there must be an orderly path or means for establishing the noosphere, also ordained by natural law. The Law of Time defines this orderly path as a conscious shift in the human timing frequency, away from artificial time and into universal natural time. This can only be accomplished by a universal reform of time, known as the calendar change. If there were no such path, the conclusion to the collapse of the technosphere would be an unbearable barbarism.

By its nature, the topic of the book is cross-disciplinary and provocative-it is visionary history. I have been described as a "spirituality historian" and "cosmic harmony researcher."4 I believe this description also characterizes *Time and the Technosphere*, which does present a solution, a new world paradigm, if you will. A new paradigm has been awaiting humanity ever since the notion of paradigms was first raised by Thomas Kuhn in 1964, in his book *The Nature of Scientific Revolutions*, or even earlier by Vernadsky and Pierre Teilhard de Chardin in their invocation of the noosphere. To be truly new, such a paradigm could hardly be expected to arise from traditional fields of thought or methods of research, be they quantum physics or microbiology. In fact, as I have often said, most people wouldn't recognize the new paradigm if it ran into them on the street. This is because people's consciousness and perceptions are basically very set and unexamined-as unexamined as the effects of the watch on their wrist or the calendar on their wall. In fact, one of the consequences of the technosphere is the limiting nature of consciousness which remains unchanged despite the fact that technology is changing all of the time. But with the occurrence of the Inevitable Event, history has been punctured. Perhaps now the new paradigm can be perceived: the paradigm of natural time as the universal factor of synchronization.

Although prompted by the stunning terrorist attack upon and consequent apocalyptic collapse of the World Trade Center Twin towers, these reflections are actually a synthesis of thoughts and explorations in the domain of time that I now find necessary to communicate in a coherent and organic form. The grand theme of this study is to understand how and why the collapse of the twin World Trade Center Towers was a biospheric event. Even so, the study of the Inevitable Event within the context of the evolution of the biosphere requires and leads us to still grander domains of thought-the historical, the theological, and the cosmological. The reason that these reflections and meditations spiral into ever grander domains of thought is due to the principle of analysis and the guiding factor employed in this study. This is the Law of Time, which is the fruit of a life-long research and pursuit of knowledge. It is the Law of Time that brings the biosphere into timely focus.

The biosphere is scarcely known even to most educated people. As a biogeochemical whole system in which the human is but one of the functioning components, the biosphere is an intricate unity with its own governing laws and principles. Without becoming conscious of how the human organism participates in the laws and principles of the biosphere, the human will continue to remain ignorant of the biosphere's existence and therefore will sooner than later reach a termination of its own evolutionary possibility. A demonstration of how the laws and principles of the biosphere, crystallized into a temporary and intermediate sheath called the technosphere, reached a climactic evolutionary point in the collapse of the World Trade Center towers will provide the means of educating in general concerning the biosphere. It will also allow us to elaborate on the meaning and nature of the technosphere, and the imminent transformation of the biosphere into the noosphere, the Earth's "mental envelope."

A fundamental purpose of the Law of Time is to expand upon the meaning and significance of the Inevitable Event, as well as upon the biosphere, technosphere, and noosphere in relation to the evolution of life and consciousness in the universe. To speak of an event being "inevitable" is to bring about a reflection on the nature of time, and even of the laws of cause-and-effect and karma. It is time that governs both cause-and-effect as well as the actions of karma. And it is time that dominates inevitability, because any such inevitable event is located in a very precise moment in time. But what actually is time? How does time "locate" events? And, once again, what is the Law of Time?

As the criterion of natural time in which synchronicity is the norm, the Law of Time is the key that unlocks the reason why the Inevitable Event occurred when it did. By precisely locating it within the evolutionary continuum of the biosphere and in relation to its point of mutational shift, the Inevitable Event is placed into a greater context of cosmic comprehension by the Law of Time. Through the lens of the Law of Time we will be able to understand the relationship between the biosphere, the intermediate state of the technosphere, and the imminent stage of the noosphere. Defining the Law of Time as the factor of universal synchronization, the law by which all events in the universe are maintained in synchronic relation to each other, we approach the domain of a hitherto undefined realm, that of the synchronic order.

As the unifying fourth-dimensional order of time in the universe, the synchronic order is measured and gauged by the 13:20 matrix. The mathematical coding of the Law of Time is entirely locked into this 13:20 matrix, otherwise known as the Harmonic Module or Tzolkin, the basis of the sacred count of the ancient Maya. When we examine the Harmonic Module, a 13 x 20 matrix, we may well ask what does that have to do with time? It is obviously as different from a clock, as a tree is from a flagpole! When we examine the matrix carefully we see that there is a repetition of thirteen notations, the numbers 1-13, that occurs 20 times yielding 260 (13 x 20) units.

Holding the 13:20 matrix together is a pattern which I identified in my book *Earth Ascending* (1984) as the binary triplet configuration, and in the *Mayan Factor* (1987) as the Loom of Maya. This pattern consists of 52 units, 26 on either side of the central or seventh vertical column. If you contemplate this 52-unit pattern you will see it is characterized by binary radial symmetry. This quality characterizes the entire matrix and defines the quality of fourth-dimensional time as radial and instantaneous distinct from the current third-dimensional paradigm which defines time as being linear and sequential. The radial quality of fourth-dimensional time is what accounts for its being the factor of universal synchronization. This is very different from the sensation of the second hand as it sweeps around the clock, indicating the inexorability, not of time but of the mechanization of time! The 13:20 matrix becomes potently useful as the means of establishing the synchronic order when it is coordinated with a genuinely harmonic timing standard, the Thirteen Moon/28-Day calendar.

"Time is the fourth dimension," is the famous adage attributed to Einstein. This statement notwithstanding, the modern notion of time is bogged down in the minutiae

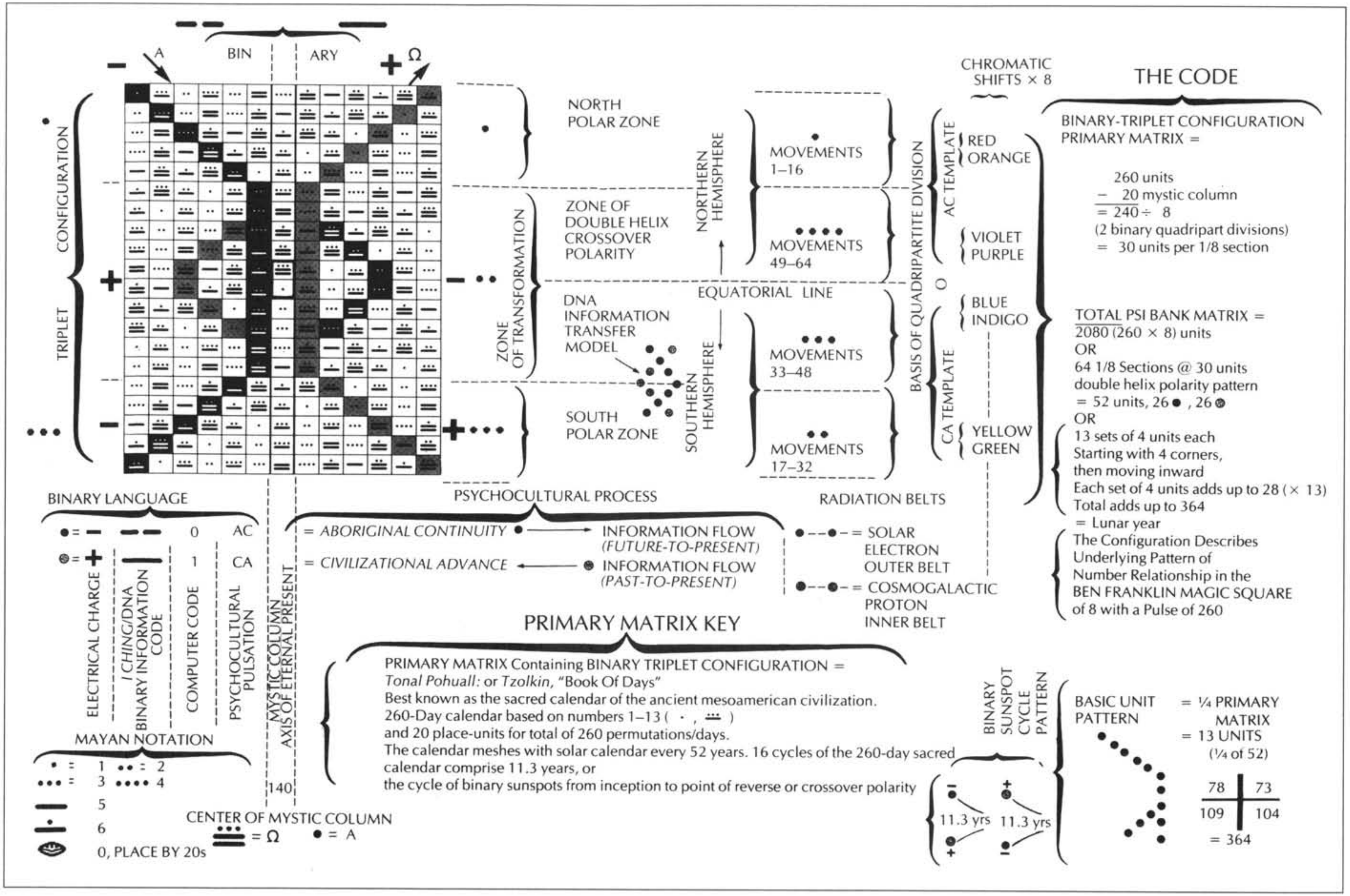
A
BIN
ARY
Ω
CONFIGURATION
TRIPLET
NORTH POLAR ZONE
ZONE OF DOUBLE HELIX CROSSOVER POLARITY
ZONE OF TRANSFORMATION
DNA INFORMATION TRANSFER MODEL
SOUTH POLAR ZONE
NORTHERN HEMISPHERE
EQUATORIAL LINE
SOUTHERN HEMISPHERE
MOVEMENTS 1–16
MOVEMENTS 49–64
MOVEMENTS 33–48
MOVEMENTS 17–32
BASIS OF QUADRIPARTITE DIVISION
AC TEMPLATE
CA TEMPLATE
CHROMATIC SHIFTS × 8
RED
ORANGE
VIOLET
PURPLE
O
BLUE
INDIGO
YELLOW
GREEN
THE CODE
BINARY-TRIPLET CONFIGURATION PRIMARY MATRIX =
260 units
− 20 mystic column
= 240 ÷ 8
(2 binary quadripart divisions)
= 30 units per 1/8 section
TOTAL PSI BANK MATRIX =
2080 (260 × 8) units
OR
64 1/8 Sections @ 30 units
double helix polarity pattern
= 52 units, 26 ●, 26 ◉
OR
13 sets of 4 units each
Starting with 4 corners,
then moving inward
Each set of 4 units adds up to 28 (× 13)
Total adds up to 364
= Lunar year
The Configuration Describes
Underlying Pattern of
Number Relationship in the
BEN FRANKLIN MAGIC SQUARE
of 8 with a Pulse of 260
BINARY LANGUAGE
● = ▬
◉ = ✚
0
1
AC
CA
ELECTRICAL CHARGE
I CHING/DNA BINARY INFORMATION CODE
COMPUTER CODE
PSYCHOCULTURAL PULSATION
MYSTIC COLUMN
AXIS OF ETERNAL PRESENT
140
PSYCHOCULTURAL PROCESS
= ABORIGINAL CONTINUITY
INFORMATION FLOW (FUTURE-TO-PRESENT)
= CIVILIZATIONAL ADVANCE
INFORMATION FLOW (PAST-TO-PRESENT)
RADIATION BELTS
= SOLAR ELECTRON OUTER BELT
= COSMOGALACTIC PROTON INNER BELT
PRIMARY MATRIX KEY
PRIMARY MATRIX Containing BINARY TRIPLET CONFIGURATION =
Tonal Pohuall: or Tzolkin, "Book Of Days"
Best known as the sacred calendar of the ancient mesoamerican civilization.
260-Day calendar based on numbers 1–13 (· , ⩸)
and 20 place-units for total of 260 permutations/days.
The calendar meshes with solar calendar every 52 years. 16 cycles of the 260-day sacred calendar comprise 11.3 years, or
the cycle of binary sunspots from inception to point of reverse or crossover polarity
MAYAN NOTATION
1
2
3
4
5
6
0, PLACE BY 20s
CENTER OF MYSTIC COLUMN
= Ω
● = A
BINARY SUNSPOT CYCLE PATTERN
11.3 yrs
11.3 yrs
BASIC UNIT PATTERN
= ¼ PRIMARY MATRIX
= 13 UNITS
(¼ of 52)
78
73
109
104
= 364

13:20 Matrix-Harmonic Module with Binary Triplet Configuration of quantum mechanics and cessium clocks, in a hair-splitting myopia that cannot see the forest for the trees. The fact is that only time governs the whole order of the universe in a manner that transcends all spatial limitations, even those of the relativistic MinkowskilEinstein four-dimensional universe, which is confined by the speed of light. As the universal factor of synchronization, time is instantaneous and transcends light. This novel perception of time defines a new reality, the synchronic, of which our experiences of synchronicity are but a foretaste. The synchronic order described by the Law of Time poses the order of an entirely new science of time.

To be realized and then applied, the science of time is dependent on two factors: a sudden and radical disruption of the historical continuum; and a genuine globalization or planetarization of consciousness. The break in historical continuity is necessary to jolt the human consciousness from its stagnant and entropic state-defined by the Law of Time as the delimiting 12:60 frequency consciousness of artificial, mechanistic time-while the globalization of consciousness, intimately connected with this sudden discontinuity, is necessary for the application of this law at a planetary whole systems level. Of course, the radical disruption of the historical continuum was provided by the collapse of the Twin Towers, and can actually be mapped in a description of the dialectical process of human evolution.

If one studies *Earth Ascending,* Map 19, "Binary Pulse of Psycho cultural Development," [shown on page 14] one can pinpoint the "Inevitable Event" of the radical historical discontinuity as the point at which the AC and CA currents cross over one another and switch polarity at the very top of the graphic depicting the "historical" process. What does it indicate, this point of psychic polarity switching currents?

The historical dialectic is understood as the interplay of two "currents" of consciousness, AC "Aboriginal Continuity" and CA, "Civilizational Advance." In the prehistoric phase, the consciousness is dominated by the Aboriginal Continuity or AC. This refers to the primary experience of the original human nature- life lived as a ritual or ceremonial round within the perimeter of the timeless present. But as the human enters the historical cycle, about which we will have more to say in later chapters, the CA becomes more predominant. As Civilizational Advance, the CA represents the tendency toward exclusive reliance on reason, logic and written means of communication. The Law of Time further defines the CA as being motivated increasingly by the artificial timing frequency of the 12:60 which gains in prominence until the final phase of the historical cycle, industrial globalization, when the CA has gained complete dominance of the biosphere.

During this final stage, the global mental condition is defined as the involvement in secondary or artificial means of experience, furthered by the planetary

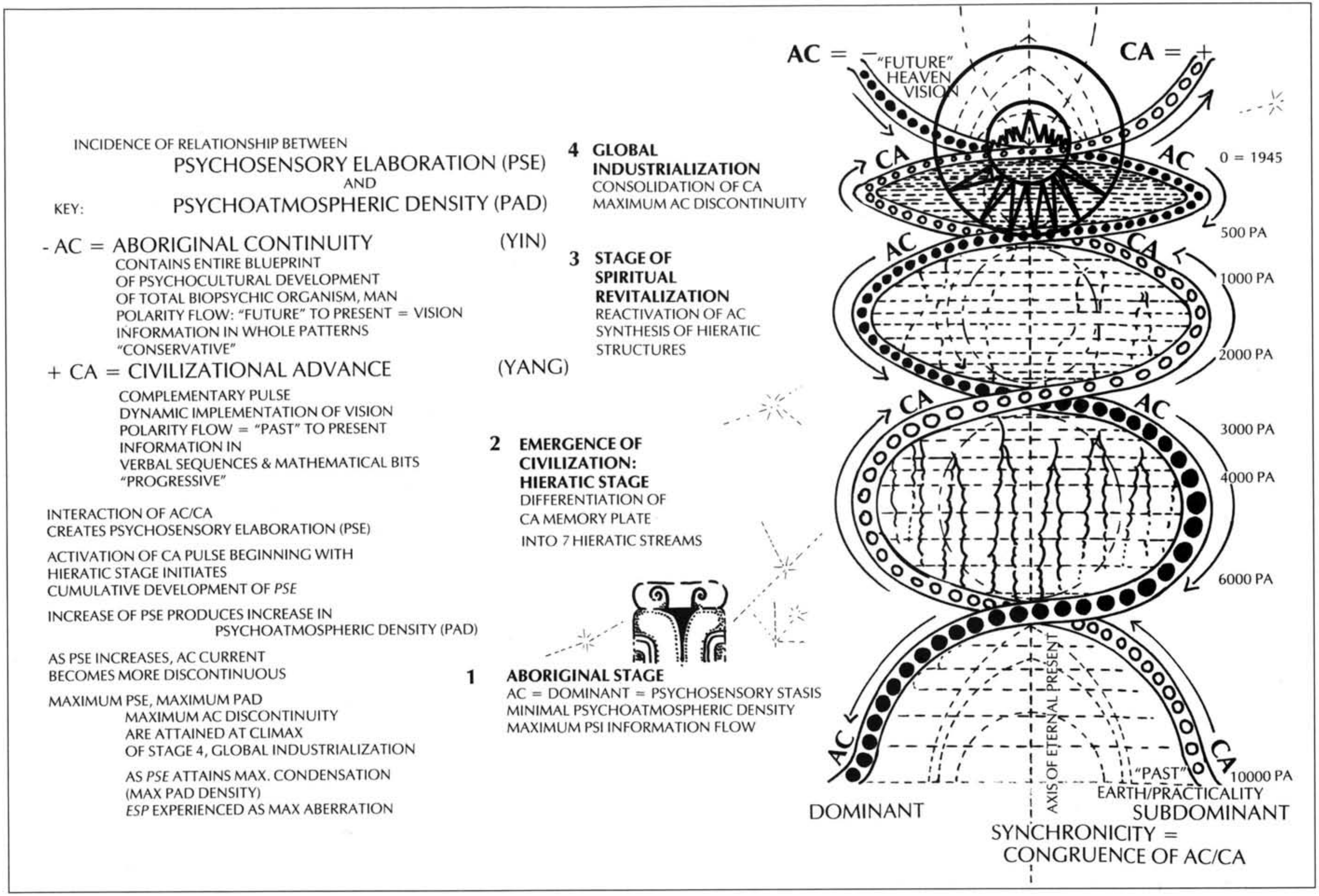
INCIDENCE OF RELATIONSHIP BETWEEN
PSYCHOSENSORY ELABORATION (PSE)
AND
PSYCHOATMOSPHERIC DENSITY (PAD)
KEY:
- AC = ABORIGINAL CONTINUITY (YIN)
CONTAINS ENTIRE BLUEPRINT
OF PSYCHOCULTURAL DEVELOPMENT
OF TOTAL BIOPSYCHIC ORGANISM, MAN
POLARITY FLOW: "FUTURE" TO PRESENT = VISION
INFORMATION IN WHOLE PATTERNS
"CONSERVATIVE"
+ CA = CIVILIZATIONAL ADVANCE (YANG)
COMPLEMENTARY PULSE
DYNAMIC IMPLEMENTATION OF VISION
POLARITY FLOW = "PAST" TO PRESENT
INFORMATION IN
VERBAL SEQUENCES & MATHEMATICAL BITS
"PROGRESSIVE"
INTERACTION OF AC/CA
CREATES PSYCHOSENSORY ELABORATION (PSE)
ACTIVATION OF CA PULSE BEGINNING WITH
HIERATIC STAGE INITIATES
CUMULATIVE DEVELOPMENT OF PSE
INCREASE OF PSE PRODUCES INCREASE IN
PSYCHOATMOSPHERIC DENSITY (PAD)
AS PSE INCREASES, AC CURRENT
BECOMES MORE DISCONTINUOUS
MAXIMUM PSE, MAXIMUM PAD
MAXIMUM AC DISCONTINUITY
ARE ATTAINED AT CLIMAX
OF STAGE 4, GLOBAL INDUSTRIALIZATION
AS PSE ATTAINS MAX. CONDENSATION
(MAX PAD DENSITY)
ESP EXPERIENCED AS MAX ABERRATION
4 GLOBAL INDUSTRIALIZATION
CONSOLIDATION OF CA
MAXIMUM AC DISCONTINUITY
3 STAGE OF SPIRITUAL REVITALIZATION
REACTIVATION OF AC
SYNTHESIS OF HIERATIC STRUCTURES
2 EMERGENCE OF CIVILIZATION: HIERATIC STAGE
DIFFERENTIATION OF
CA MEMORY PLATE
INTO 7 HIERATIC STREAMS
1 ABORIGINAL STAGE
AC = DOMINANT = PSYCHOSENSORY STASIS
MINIMAL PSYCHOATMOSPHERIC DENSITY
MAXIMUM PSI INFORMATION FLOW
AC = −
"FUTURE"
HEAVEN
VISION
CA = +
CA
AC
AC
CA
CA
AC
AC
CA
0 = 1945
500 PA
1000 PA
2000 PA
3000 PA
4000 PA
6000 PA
10000 PA
AXIS OF ETERNAL PRESENT
"PAST"
EARTH/PRACTICALITY
DOMINANT
SUBDOMINANT
SYNCHRONICITY =
CONGRUENCE OF AC/CA

Locating the Inevitable Event in the Holonomic Equation:
9-11, 2001 Climax of the Binary Pulse of Psychocultural Development
(from *Earth Ascending*, Map 19)

technosphere. The sum effect on consciousness of the advanced CA stage is defined as Psychosensory Elaboration or PSE, which in turn creates a psychoatmospheric density (PAD)-literally a mental cloud which obscures altogether the organic biopsychic reality of the human within the biosphere. From the point of view of a planetary pathology, this massive collective obscuring (CA) of the original nature (AC) defines a "suicidal crisis point," a moment at which the subconscious instinct of the human being programs its own self-defeat in order to ultimately return to the organic reality of the biosphere and the simultaneous advent of the noosphere. Hence, the Inevitable Event: the radical break in the continuum of history.

This moment of radical historical discontinuity occurred with the September 11,2001, terrorist attack on the twin towers of the World Trade Center-the very nerve center of late global industrialization-and the Pentagon, the supreme military fortress that defends globalization. If the fall of the towers symbolically augured the fall of Babylon and the Tower of Babel itself, the penetration of the Pentagon was no less symbolic. The fact that there was not one but two towers of Babel and a separation of some twenty minutes between the attack on each tower allowed the second attack to be viewed live on global television by about as many people on Earth as had ever witnessed anything at one simultaneous moment. By the end of the day, the scene had been repeated multiple times on virtually every television network on the planet.

The unbelievable shock of the moment and its simultaneous perception, coupled with the symbolism of what was actually occurring, effectively rendered the break in the historical continuum as a profound mental event-a rupture of 5,000 years of history from the first Babylon to the last. The witnessing of the event on television meant that it was received immediately into the nervous system of virtually the entire species. This in itself is a noospheric moment-the mental envelope of the Earth made collectively and globally conscious in one instant of history-shattering significance, and hence, the genuine entry into post-history. Archetypally, the images of the towers collapsing with people flying or jumping off the upper stories, is a pure reflection of the Sixteenth Major Arcana card of the traditional Tarot which shows a tower being shattered by lightning, with bodies falling down from it.

While the negative powers and forces grapple with their shattered dreams in the aftermath of the event, it is the time for the powerful luminescence of the noosphere to begin to activate the receptive cells of the human system within the now catalyzed biosphere. Fifty-six years after Hiroshima, the fall of the World Trade Center Towers is the fatal puncture in the technosphere which is intended to reorient the human within the biosphere. ESP (extrasensory perception)-the opposite of PSE (psychosensory elaboration)-is released in the instantaneity of the information reaching virtually every human being on Earth. This is the collective reactivation of the Aboriginal Continuity current that has been disengaged and disempowered for so long by the Civilizational Advance. Civilization has finally been checkmated. Now the Revolution of Time must be completed in order to stabilize the biosphere, anchor the noosphere, and de-structure the technosphere.

According to the Law of Time, the Inevitable Event was already embedded in the moment in which it actually occurred, like a flaw in a piece of glass that, upon concussion at the right pressure point, shatters all at once. In the case of the World Trade Center Towers and the Pentagon, we are also dealing with symbolic constellations of such potency that the very act of their shattering redefines the reality of human consciousness. In fact, the nature and problem of human consciousness as it has evolved historically is also at the root of this astonishing event. While it will take time for the collective human consciousness to fully grasp just how much its reality is being redefined, a vacuum now exists within the technosphere itself-a vacuum which awaits being filled by a new paradigm. Since the collapse of the Twin Towers actually represents the limits of artificial time, the new paradigm must rest on a redefinition of time altogether. This is the definition supplied by the Law of Time, the pragmatic application of which is the replacement of the irregular measure of the Gregorian calendar by the harmonic measure of the Thirteen Moon/28-Day calendar.

Because of its harmonic measure, the Thirteen Moon/28-Day calendar coordinated by the Harmonic Module is the new dispensation of time-reckoning for the human race. Rooted in the four recurring year-bearers of the Wizard's Count--:-which is not the same as the Mayan long count, nor the Julian count of modern science-the Thirteen Moon calendar exists for the establishment of the reality of the synchronic order as an entirely new method of knowledge and being for the entire human race as a cosmic medium operating within the biosphere. By Wizard's Count is meant the circulation of thirteen numbers and four different glyphs-Yellow Seed, Red Moon, White Wizard, Blue Storm-that code the July 26 Thirteen Moon/28-Day calendar synchronization date to create a master 52-year cycle, Anchored in the Thirteen Moon/2 8-Day calendar, the synchronic order is the means for manifesting the reality of the noosphere on planet Earth. Without such a basis in a lived mathematics of time correlated to the actual cycles of Earth, the noosphere could not be experienced nor fully realized. The successful completion of the biosphere-noosphere transition, then, is the triumph of the synchronic order, the conscious elevation of Earth into the cosmic frequency of fourth-dimensional time.

For all of this to come to be, however, is yet dependent on an analysis of the role of time in the creation of the technosphere and the subsequent comprehension of its limits and the potential for its de-structuring. This is not only a matter of theory or prophecy; it is a matter of putting into practice, making a coherent action out of the application of the Law of Time as it completes the pressing urgency of the biosphere-noosphere transition. The coherent action to bring us through this rite of passage is the Great Calendar Change of 2004. Ultimately, this book is presented to make us ready for this unprecedented time changing moment. May you, 0 reader, ponder deeply and then, if you so will, come join me in living on the other side of the wall of mechanized time.

◆

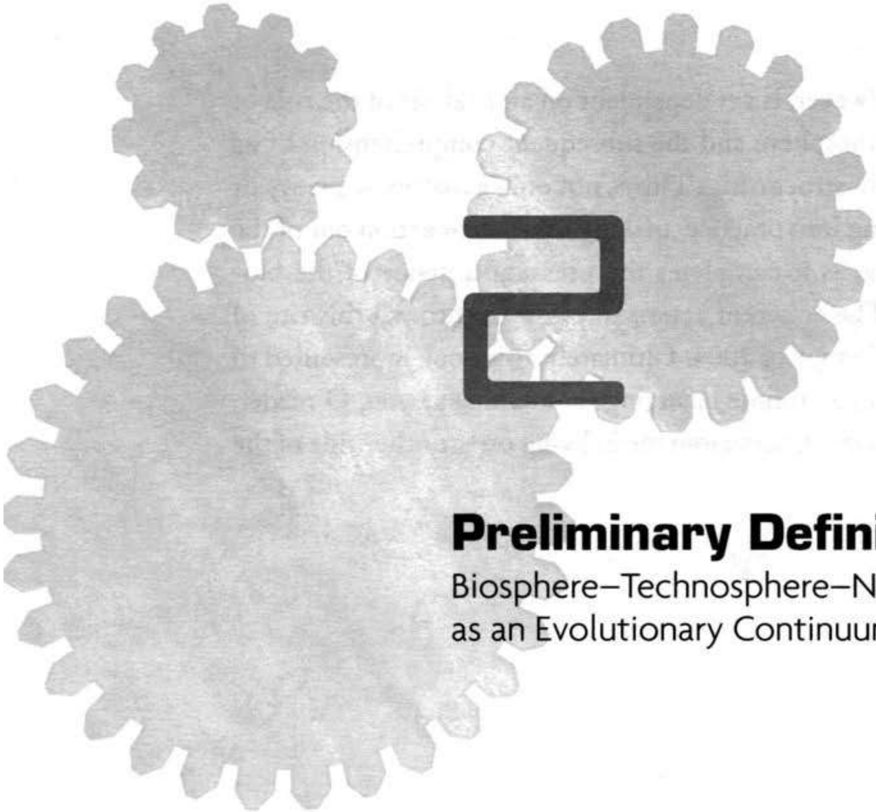

2

Preliminary Definitions

Biosphere–Technosphere–Noosphere as an Evolutionary Continuum

BY UNDERSTANDING that the Inevitable Event was a function of the biosphere-indeed, a necessary moment in the evolutionary trajectory of the biosphere-we are saying that this event was not just a human enactment, but that humans, inseparable from the biosphere, are effected to perform functions that relate to the process of the biosphere as a whole. Seeing the Inevitable Event and, in fact, the entire human drama from the perspective of the biosphere, we are actually lifting the cruel tragedies of contemporary reality above conflict and into a state where genuine laws of peace and harmony may shed their light of wisdom, illuminating the human drama with a higher understanding. One thing is certain: Unless we rise above our own humanity and cultivate a non-anthropocentric view-the view represented by the noosphere-there will be no release from the calamity that is now engulfing us. The point is not of maintaining the global economy, but of saving the biosphere.

For the multilayered task of bringing a new understanding to light, defining the technosphere, and then the Law of Time, we shall begin with the former: defining the technosphere in the context of the evolutionary continuum of biosphere-technosphere-noosphere.

VERNADSKY AND THE UNDERSTANDING OF THE BIOSPHERE

It is amazing that the name V. I. Vernadsky (1863-1945) is so little known in the West, and that the word *biosphere,* as well as the laws and principles pertaining to it, are scarcely more known. To a large degree, this ignorance is due to the great cleavage in human relations experienced as the Cold War. The biosphere, which was profoundly studied and scientifically articulated by the master Russian scientist V. I. Vernadsky, is actually the precise word for what in the West is vaguely referred to as the *environment.* But even more than being a precise and scientific description for what is meant by the use of the word *environment,* biosphere defines a whole system model of life on Earth, and because of that, also presupposes a whole system methodology and point of view. You cannot speak about the biosphere without entering into a world of discourse that is holistic or holonomic to the core.

If one looks for Vernadsky's work in English, one will find precious little available. His key work, *The Biosphere,* exists in two very different translations; the one published by Synergetic Press in 1986 is easier to read, while the other, the "complete annotated edition" published by the Far West Institute in 1998, is far more erudite, with a major bibliography and much supplementary information. The earlier translation was published in conjunction with the opening of the Biosphere II Project outside Tucson, Arizona. It is now a tourist site, with a bookstore that carries nothing by Vernadsky! Several other hard-to-find English-language books on the biosphere include the excellently informative *Traces of Bygone Biospheres* by Andrey Lapo (Synergetic Press and Mir Publishers, 1987), the flashy *Biosphere Catalog* (Synergetic Press, 1985), and the much more specialized *Energies: An Illustrated Guide to the Biosphere and Civilization* by Vaclav Smil (MIT, 1999). Of these books, only *Traces of Bygone Biospheres* contains significant information about Vernadsky and his ideas. This scant list of available texts in English is hardly proportionate to the vast reality of the biosphere as a complex terrestrial dynamic in the throes of evolutionary change.

The place and role of Vernadsky in Russian science is virtually equal to that of Einstein in Western science. A consideration of the achievement of each of these scientists also presents us with a dramatic contrast in perspectives, areas of interest, and consequent modes of analysis. The physicalist relativism of Einstein's legacy and its pursuit of the Big Bang is radically different in almost every way from the legacy of Vernadsky's work, which presents a biogeological worldview that is organically integrative and the basis of what in Russian science has come to be called *cosmism.* While the physicalist fascination of Western science has supported the materialism of modern Western thought and its way of life (which is actually an accelerating function of the biosphere's own internal processes), the whole system thinking of

Vernadsky—characterized by the conception of a "biogeochemical process"—ironically supports a cosmic rather than materialistic worldview.

To illustrate the Russian perspective and the influence of Vernadsky's work, we present the concluding thoughts of a book by another Russian scientist and philosopher, 1. Laptev:

> Each one of us requires the whole earth—today this is quite clear. But surely each one of us is required by the whole earth! And when these two mutually penetrating requirements, so distinctly revealed and intensified by the leap out into space, are satisfied, the new historical epoch will begin "in which mankind itself, and with mankind all branches of its activity, . . . will experience an advance that will put everything preceding it in the deepest shade." (Fr. Engels, *Dialectics of Nature)* And people, no longer burdened by the thirst for money, will remember something that many of our contemporaries have forgotten: The only important things in life are such intangible qualities as beauty and wisdom, laughter and love.
>
> Nothing remains for us, members of a society the ideals of which are in unison with elemental geological processes and the laws of nature, but to wish that people the world over would recognize this as soon as possible.!

Laptev's words—written twenty-eight years ago and precisely at the midpoint between Hiroshima and the Inevitable Event, as well as during the year in which the World Trade Center in New York City was finally completed and inaugurated (April 4, 1973)—simply and elegantly state the theme and point of view that we are pursuing. In fact, it is the Inevitable Event that prepares us for the new historical epoch. But first we must realize how we are organized by the biosphere in order that the technosphere may be transformed into the noosphere.

The root of Laptev's cosmic vision lies in Vernadsky's succinct definition of the biosphere: "The biosphere is the region on Earth for the transformation of cosmic energies." Although the word *biosphere* had been coined in the late nineteenth century, it was really only with Vernadsky's efforts that the biosphere became a viable description of the totality of life on Earth inclusive of its organic and inorganic support processes.

Vernadsky was trained as a geologist with a specialization in crystallography. He immediately absorbed the implications of Curies discovery of radioactivity (1896), and devoted many studies to the purpose of uranium and naturally occurring radioactivity within the Earth's geology. By the time of the First World War, Vernadsky had already written important texts such as *Fundamentals of Crystallog;raphy and Descriptive Mineralogy,* and, along with Madame Curie, proposed an "international radiography of Earth's crust." This was in advance of Alfred Wegener's definitive work on plate tectonics. During the time of the First World War and Russian Revolution, Vernadsky's attention turned toward the problem of living matter and its relation to the geochemistry of the Earth. From this was born Vernadsky's perception of the unitive biogeochemical process by which the biosphere maintains itself. In 1923 he published a "plea for the establishment of a biogeochemical laboratory."

In Paris, Vernadsky worked at the Insititut de Radium (Marie Curie) and became acquainted with Henri Bergson, who was then president of the International Commission of Intellectual Cooperation of the League of Nations. Through Bergson's circles, Vernadsky met often with the biologist Pierre Teilhard de Chardin (1881-1955) and the philosopher Edouard Ie Roy (1870-1954), with whom Vernadskyjointly coined the word and concept of the noosphere. Although Pierre Teilhard de Chardin is credited in the West with the concept of the noosphere, defined by him as the "mental envelope of the planet above and discontinuous with the biosphere," de Chardin's approach to the evolution of the noosphere as a natural consequence of vertebrate biology differs from Vernadsky's understanding in which it is the entirety of the biosphere that evokes the noosphere.

The chief fruit of this fertile time in Vernadsky's life was the publication of the little book titled *The Biosphere* (1926). Demonstrating the synchronic ordering principle of the noosphere, it is interesting that the same year that saw the publication of *The Biosphere* also witnessed the appearance of Jan Smuts noted book *Holism and Evolution.* Through most of the rest of his sometimes turbulent career (he often came under attack from hard-line bolsheviks), Vernadsky was to elaborate on the laws and principles of the biosphere, often expressed simply as problems in biogeochemistry. At the same time, his influence as a holistic thinker began to spread through the Soviet Academy of Sciences. A list of a few of his later publications gives some idea of the breadth of his thought, and includes: *Geochemistry, Problems in Biogeochemistry I* and *II, On the Boundaries of the Biosphere, Goethe as Naturalist, Scientific Thought as a Planetary Phenomenon, Some Words on the Noosphere,* and his never completed work, *The Chemical Structure of the Earth's Biosphere.*

Following the death of his wife in 1944, Vernadsky expressed the opinion that after the Second World War, American and Russian scientists should work more closely together. Early in 1945 (January 6), Vernadsky followed his wife to the grave. The Second World War ended, only to be followed by the Cold War (1947-1990). The critical collaboration between American and Russian scientists longed for by Vernadsky was soon to be buried in the ideological wasteland over which the superstructure of the technosphere would be constructed, a fact that in itself hastened the acceleration of the biogeochemical processes bringing about the climax of the technosphere.

Vernadsky did not live to witness the birth of the atomic era, which began only months after he died. But like Einstein in America, in 1940 Vernadsky-along with several other scientists, and mindful of the possible military applications-had prompted the Russian government to begin investigations into the possibility of deriving energy from nuclear fission. Yet it is most remarkable that by the time of his death, Vernadsky had almost single-handedly defined virtually every aspect of the laws and principles of the biosphere. This he always did with an eye to an imminent moment in the not-too-distant future when there would be a dramatic biogeochemical mutation or combustion, out of which the biosphere would somehow trigger the noosphere:

> Not at a crisis of nervousness do we stand now, not at a time for the vacillation of flabby souls; but at a great turning point in the history of scientific thought, at a crisis such as occurs but once in a thousand years, such as has not been witnessed for many generations. Standing at this point, with the vista of future achievements before us, we should be happy that it is our lot to live at this time and to participate in the creation of tomorrow.2

Undoubtedly his sense of the great moment of transition came from the very fact that during the span of his life from 1863 to 1945, some eighty-two years, Vernadsky was able to witness firsthand the accelerating effects of man's thought-in the form of industrial technology and the machine-and its turbulent and transformative impact on the terrestrial biosphere. After all, as a geologist, Vernadsky was familiar with the long history of the Earth, and the impact of human technology on the biosphere was inescapably the most significant aspect of any study of the biosphere to the present time. So wrote Vernadsky: "Only man transgresses the established order. . . upsets the equilibrium, though whether he materially cripples the transforming mechanism, or merely redistributes it, we cannot at the moment be sure."3

The "transforming mechanism" referred to is the biosphere itself, which Vernadsky had already defined as the medium or region for the transformation of cosmic energy on Earth. At its simplest level, the unity of the whole of living matter of the biosphere is the sum of its living organisms. These living organisms either directly or indirectly are continuously processing solar and cosmic radiation, transforming it into various chemical cycles that establish the atmosphere, most notably through photosynthesis and the oxygen and carbon dioxide cycles. The important point, however, in any consideration of the biosphere is the principle of unity: the biosphere is a unitive complex of cycles summarized as a single continuously transforming and slowly mutating biogeochemical process. In this single complex cyclical process, all of life-all of living matter-is a unity, a singular dynamic by which the biosphere can be considered as a living whole. From the plankton of the oceans to the trees in the forest and the human peering from his cave at the rising sun-all of this is a single unity.

It is the interaction of the living matter with the inert matter that creates and establishes the attributes and dynamics of the biosphere as a whole, all of which are now in a heightened state of acceleration and transformation, for as Vernadsky put it in the conclusion to *The Biosphere:* "The thought of the human species is a new fact which is turning the structure of the biosphere upside down after myriads of centuries." The structures and principles of the biosphere that man is now upsetting include: the continuity and invariance of life on Earth; vegetation as a transformer of the energy of the sun; multiplication as a manifestation of transformed solar energy; propagation of life and the forces resisting propagation; and the unique relations of life to the inert matter of the biosphere, including the composition of living matter into two orders, plant life and the various biological animal species, inclusive of man. The first order of plants and bacteria constitute the mechanism for capturing the solar radiation and introducing the energy of the sun into the biosphere. The second order of biological life is a manifestation of the process of the transmission of energy through the vegetable world; that is, at a primary and fundamental level, animal life is dependent upon plant life for its existence. (See plate 2, Life Is the Transformation of Solar Energy.)

All these principles for maintaining the equilibrium of the biosphere are functions of the dynamic of the different biochemical and geochemical cycles that maintain life both on land and in the ocean, all of which are subjected to processes that we refer to as evolution, both geological and biological. The evolutionary mechanisms of life include the principle of the biomass constant and the biogenic migration of atoms. Both of these mechanisms refer to and define the continuity and invariance of life on Earth. That is, the quantity of living matter has remained roughly unchanged since the origin of life on Earth-the biomass constant. Furthermore, the amount of oxygen in the biosphere is equal to the amount of biomass. It is the pressure of the different species on each other with the resultant interactions and changes that checks the infinite propagation of anyone species at the expense of the rest, and causes the mutation and evolution of the different life forms. It is through the intermediate structure of the technosphere that the human is wreaking the greatest havoc on these principles that are intended to maintain the biosphere as a stable dynamic, and by which the human constitutes a biomutation of unparalleled significance. The different stages of the mutation of life account for the biogenic migration of atoms: "The evolution of different forms of life throughout geological time increases the biogenic migration of elements in the biosphere."4 The metamorphosis of dinosaurs into fossil fuel is an example of the biogenic migration of atoms-and of the introduction of "free energy" (the term for industrial waste and by-products such as carbon monoxide) into the biosphere, while at a more mundane level, it refers to processes such as metabolism, circulation, breathing, and so forth.

To summarize, there are two biogeochemical principles. The first states: "The biogenic migration of chemical elements in the biosphere tends towards a maximum of manifestation"; this refers to the presence of life everywhere on Earth, as well as the principle of the pressure of living matter upon itself. The second biogeochemical principle states: "The evolution of species, in tending towards the creation of new forms of life, always move in the direction of increasing biogenic migration of the atoms in the biosphere. . . . The second biogeochemical principle regulates the course of evolution. A newly evolved species will survive in the biosphere only if it is sufficiently able and sufficiently active in furthering the migration of atoms. The organism cannot be considered apart from its medium. It is a part of the complex mechanism of the biosphere." Here Vernadsky concludes his study of the biosphere with a consideration of the laws regulating evolution that, for him, are to be sought in the mechanism of the biosphere and not in accidental circumstances. For this reason "evolution must proceed, namely, in the direction of increasing consciousness and thought, and of forms having greater and greater influence on their surroundings."5 This consideration leads Vernadsky to domains of philosophical or religious thought. Why?

"We are confronted with a new form of biogenic migration resulting from the activity of human reason. Human thought has changed in a brisk and radical manner the trend of natural processes and has even modified what we call *natural laws.* Consciousness and thought, despite the efforts of generations of thinkers, have never been defined, have never been given a physical basis, in terms of matter and energy."6 And here Vernadsky poses a question that remained unanswered for him but that also leads to a consideration of the noosphere: "How can processes which seem purely physical be affected by consciousness?"7

Although *The Biosphere* is the initial statement and underlying basis of all Vernadsky's later thought and work, his final major study, *Problems in Biogeochemistry II,* translated and published by his son, George Vernadsky, at Yale University in 1944, states the final conclusions of Vernadsky on the intriguing relation of human consciousness-via its thought and projected dynamic, the machine-on the biosphere. These reflections provide us with a means for defining more precisely the nature of the technosphere and noosphere.

FROM THE CREATION OF THE TECHNOSPHERE TO THE ADVENT OF THE NOOSPHERE

Vernadsky developed the biosphere into a full scientific description of the sum oflife and its support systems on Earth, and was a co-originator of the concept of the noosphere, yet he seems never to have used the word or concept of the technosphere. In some of Vernadsky's descriptions of the relation of man and his thought on the biosphere, however, the technosphere is absolutely implied. In reading the following chain ofVernadsky's thought, we can discern what we refer to as the biosphere-technosphere-noosphere continuum:

> Man by his work and his conscious attitude toward life is remaking a terrestrial envelope, the geological domain of life, the biosphere. He is transforming it into a new geological state, the noosphere.
>
> He creates within the biosphere new biogeochemical processes that did not exist before. A planetary phenomenon, the biogeochemical history of the chemical elements is notably changed.
>
> An immense new form of biogeochemical energy is represented by the technological work of man, complexly guided by his thought. It is interesting that the increase, in the course of time, of machinery in the structure of human society also proceeds in geometrical progression, just like the reproduction of any kind of living matter, man included.
>
> Statesmen should be aware of the present elemental process of transition of the biosphere into the noosphere. The fundamental property of biogeochemical energy is clearly revealed in the growth of free energy of the biosphere with the progress of geological time, especially of its transition into the noosphere.8

In this sequence of thoughts, we understand that Vernadsky, in his precise location in time, was himself the necessary voice of the biosphere giving rise to the noosphere. Put in another way, the very process that Vernadsky is describing-the transition of the biosphere to a new geological state, the noosphere-necessitated a coming into consciousness of itself, a self-reflective medium and voice. This self-reflective medium and voice of the biosphere-noosphere transition was Vernadsky

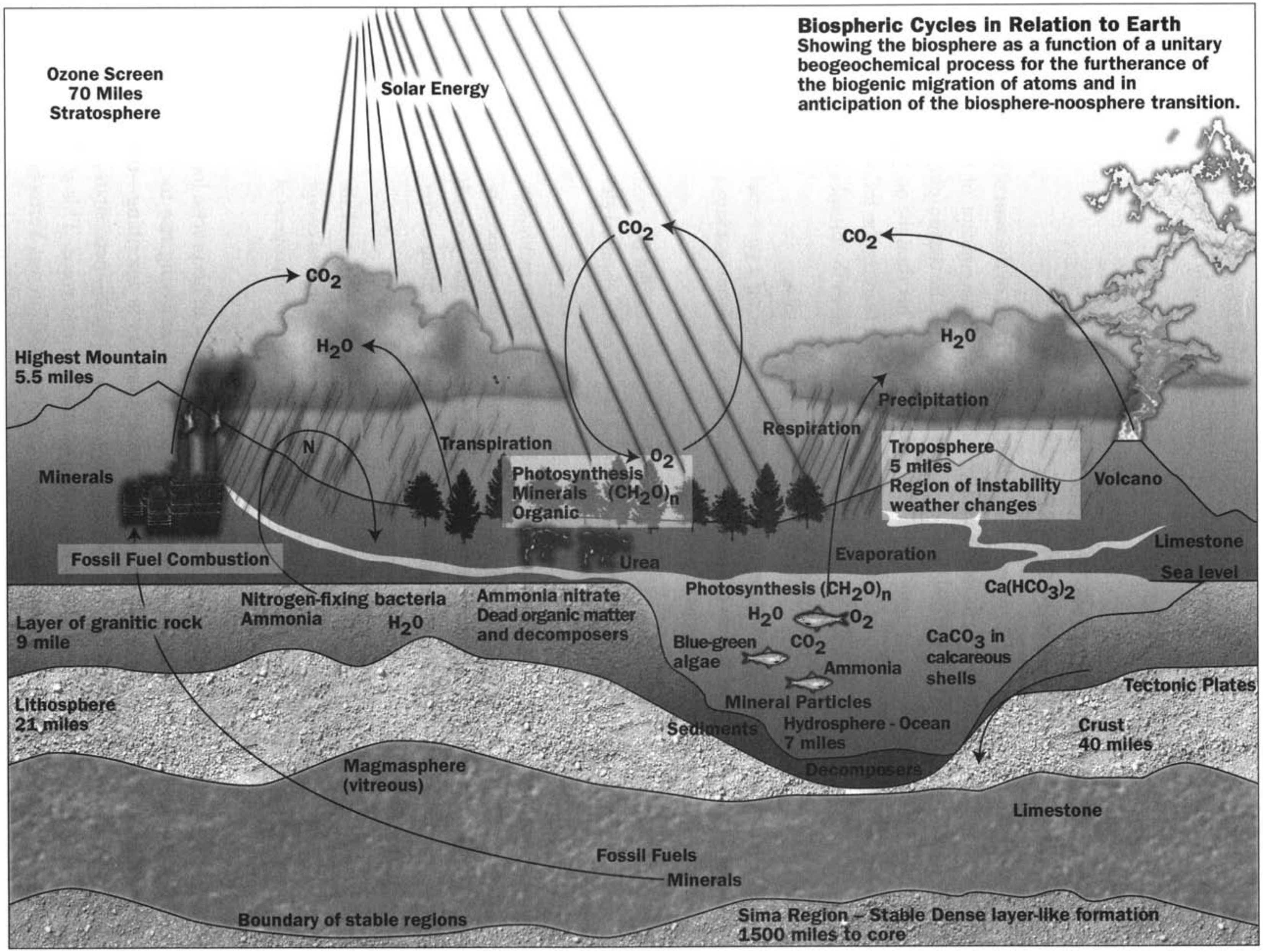
Biospheric Cycles in Relation to Earth
Showing the biosphere as a function of a unitary beogeochemical process for the furtherance of the biogenic migration of atoms and in anticipation of the biosphere-noosphere transition.
Ozone Screen
70 Miles
Stratosphere
Solar Energy
CO_2
CO_2
CO_2
H_2O
H_2O
Highest Mountain
5.5 miles
Minerals
N
Transpiration
O_2
Photosynthesis
Minerals $(CH_2O)_n$
Organic
Respiration
Precipitation
Troposphere
5 miles
Region of instability
weather changes
Volcano
Limestone
Sea level
Fossil Fuel Combustion
Urea
Evaporation
Nitrogen-fixing bacteria
Ammonia
H_2O
Ammonia nitrate
Dead organic matter
and decomposers
Photosynthesis $(CH_2O)_n$
$Ca(HCO_3)_2$
H_2O
O_2
CO_2
Blue-green
algae
Ammonia
$CaCO_3$ in
calcareous
shells
Mineral Particles
Layer of granitic rock
9 mile
Lithosphere
21 miles
Sediments
Hydrosphere - Ocean
7 miles
Decomposers
Tectonic Plates
Crust
40 miles
Magmasphere
(vitreous)
Limestone
Fossil Fuels
Minerals
Boundary of stable regions
Sima Region – Stable Dense layer-like formation
1500 miles to core

Biospheric Cycles

himself. In this observation we may understand that, although unconscious until it was made precipitant by the accelerated biogeochemical combustion itself, the order of the noosphere is none other than what the Law of Time defines as the synchronic order. As the structure of fourth-dimensional time, the synchronic order is a fluid, radial chronomantic structure in which time, telepathy, and consciousness are united in ways unpredictable to even the most advanced Western systems of mathematics.

For Vernadsky the sum properties of the elements of the biosphere are the functions both of mathematics and of different kinds of geometries. It is of note that in the summation of the differences between inert and living bodies according to three factors, the first being energy and the second chemical manifestations, Vernadsky defines the third factor as "differences in regard to space-time." Of this factor, Vernadsky writes:

> In considering space-time, however, the situation becomes rather involved. Here we enter, on the one hand, a domain not studied scientifically, and on the other hand, that substratum of all natural phenomena, namely their geometry, which the naturalist in his scientific work is wont to dismiss without attention. This substratum, the geometrical state of physical space, underlies all physiochemical phenomena and perhaps has an even deeper reality than that of the phenomena themselves.9

This deeper reality is the phenomenon of time itself, expressed through various geometries or geometrical principles, which informs the manifest phenomena of life. This is an obviously aesthetic factor as well as a cosm.Qlogical principle.

The Dynamics of Time fully substantiates Vernadsky's perception of the geometrical ordering principle of the deeper reality of time:

> All geometric forms are radiative and derive from the fourth-dimensional radial matrix. All fourth-dimensional functions are radial in nature and imply a principle of center from which the structure is projected. . .
>
> Geometry is how fourth-dimensional time incorporates as third-dimensional form. The incorporation of time as geometrical form informs all of the inorganic and organic orders of the realm of instinct. Time incorporated as the geometry of form is the principle example ofT(E) = Art. All forms and species participate in varying orders of geometry of form, both in their bodily structure and their forms of process.1o

If the question is asked, Where is this deeper reality of fourth-dimensional time from whence the projective geometries are generated? then we must answer that it is here in the noosphere, the medium on Earth for the transformation of cosmic thought.

It is the noosphere that informs all of the forms of the biosphere, inert and living, as well as producing in all genuine thought a structure and an order that are continuously moving in the direction of greater consciousness. To even think about the noosphere is to be a reflection of it.

As a noospheric medium, Vernadsky's life span was precisely timed to end when the global structure of the technosphere was about to be unconsciously constructed. The ingredients of the technosphere are indicated in the above-quoted passages as the planetary phenomena altering the biogeochemical history of the chemical elements, transforming them into an "immense new form of biogeochemical energy represented by the technological work of man, complexly guided by his thought." It is the increase, over the course of time, of machinery in the structure of human society that proceeds in geometrical progression-just like the reproduction of any kind of living matter, man included-and that constitutes the general nature of the technosphere. The growth of "free energy," the often toxic by-products of the industrial transformation of raw goods into consumer products, becomes the hallmark of the technosphere, the transitional stage between the biosphere and the noosphere.

It is this very point of the elemental process of the production of "free energy" and what it signifies in terms of the transformation of the biosphere into the noosphere that causes Vernadsky to call statesmen to attention. Why? As we know today, fifty-six years after Hiroshima, "free energy" in various manifestations has been causing danger signals to go off throughout the biosphere-from ozone depletion to global warming to nuclear radioactivity. What these danger signals have been telling us is that the techno sphere is coming to a point of its own exponentially accelerating biogeochemical combustion. Statesmen and politicians are advised by Vernadsky because the domain of their political rulership actually encompasses the everyday governing of the technosphere. But, alas, most people, statesmen included, are still unconscious of the existence of the technosphere, much less its role in relation to the biosphere and the forthcoming geological state, the noosphere. Hence, the Inevitable Event.

During much of Vernadsky's life, the geochemical combustion and release of "free energy" was just coming into its own. At his birth in the middle of the nineteenth century, the world was not yet globally industrialized, and without a totally globalized industrial base there can be no technosphere. The critical span of the two world wars, 1914-1945, was for the purpose of establishing the world market-a total industrialized market and communications base for the expansion of technology and the creation of the unifying global structure of the technosphere. The inauguration of the atomic age at Hiroshima, August 6, 1945, marks the precise beginning of the dual factors: the commencement of biogeochemical transmutation and combustion of the biosphere by the impact of human thought, the atomic bomb and release of nuclear radioactivity into the terrestrial atmosphere; and simultaneously, the beginning construction of the technosphere.

The bomb created a point of critical self-awareness within the human organism as well as producing in that organism an unprecedented sense of its own power. It was this sense of power that necessitated the unconscious impulse toward putting into place the structure of the technosphere. Of course, at the outset the species was only responding to a new, uncharted world situation, fundamentally unaware both of what it was creating as well as of the fact that it was actually constructing something in the nature of a transition between two geological conditions, the biosphere and the noosphere.

It could be said that Hiroshima lit the fuse of the final stage of biogeochemical combustion, the technosphere being the inevitable structure needed for the fuse to detonate at just the right moment. In this regard the synchronic order of the Law of Time ordained a precise 56-year cycle for the existence of the technosphere: 1945-2001. It is important to grasp that in this 56-year cycle the human population quadrupled, expanding from 1.5 to more than 6 billion. The 56-year cycle divides into precise subcycles of twenty-eight years each. The first twenty-eight years, 1945-1973, begins with Hiroshima and ends with the completion of the World Trade Center towers in New York City. This marks the accomplishment of the creation of the "World Market," the actual transformation of the biosphere into the technosphere. The second 28-year cycle, 1973-2001, spans from the dedication of the World Trade Center towers to their destruction by an act of suicidal planetary terror, and marks the phase of "globalization" -the absolute dominance of the materialist market economics, inclusive of the dramatic end of the Cold War in 1990. The very midpoint of the second 28-year cycle is the Harmonic Convergence of August 16-17, 1987, exactly fourteen years after the dedication of the World Trade Center and fourteen years before the inevitable moment of their destruction.

A profound statement of the way in which the Law of Time governs the entire biosphere, inclusive of the artificial and transitional structure of the technosphere, lies in the fact that the Gregorian calendar-an otherwise irregular and confusing pseudostandard of measure-is nonetheless coordinated by the Law of Time to repeat its cycle precisely every twenty-eight years. What this means is that the days of the week and months for the Gregorian calendar year 1945 repeated again in 1973, and yet a third time in 2001. The program prefiguring the Inevitable Event was already tested at Trinity Site (July 16, 1945), and established as the atomic destruction of

Hiroshima (August 6, 1945) and Nagasaki (August 9, 1945). This program of mass destruction was sublimated in 1973 as the triumphant completion of the Twin Towers of the World Trade Center, only so that in 2001 they could be the apocalyptic target of the inevitable 9-11 planetary red alert and moral emergency catastrophe, the only rival of Hiroshima and Nagasaki in fifty-six years. We shall return to this point later in our discussion on the Law of Time and the role of calendars as macro-organizing, programming factors that control the functioning of the human in the biosphere.

It is of special note that the very last work published before Vernadsky's death was the essay "Some Words about the Noosphere." In his vision, Vernadsky could hardly have conceived of what was to occur first as the technosphere. Yet, he wrote in this last essay, published 1944:

> The historical process is being radically changed under our very eyes. For the first time in the history of mankind the interest of the masses on the one hand, and the free thought of individuals on the other, determine the course of life of mankind and provide standards for men's ideas of justice. Mankind taken as a whole is becoming a mighty geological force. There arises the problem of the *reconstruction of the biosphere in the interests of freely thinking humanity as a single totality.* This new state of the biosphere which we approach without our noticing it is the noosphereY

By contrast, Teilhard de Chardin, in his most famous book, *The Phenomenon of Man* (1959), defined the noosphere as a "new canopy," a "thinking stratum" that has been unfolding since the end of the Tertiary period, and has since been unfolding discontinuous with and above the biospheric world of plants and animals. Teilhard de Chardin also speaks of "planets with noosphere," those worlds where the enlightenment of humanity has become synonymous with the life of the planet as a whole system. The one path to the noosphere defined by a scientist in Marxist Russia and the other by a Catholic biologist both point to the same radically positive end vision of the evolution of humanity, a fact that in itself is a dissolution of any kind of contradictory dialectic. Only history is governed by the merciless dialectical movement of matter and thought, rich and poor, material and spiritual. Clearly, therefore, the noosphere is beyond or after history. In this regard then, the technosphere is the concluding stage of history, to be followed by the post-historical noosphere.

The establishment of the techno sphere after his death notwithstanding, it is still important to listen to Vernadsky's idealism with regard to the noosphere: "The problem of planned, consistent activities which will aid us in mastering nature and accomplishing the correct distribution of wealth, connected with the comprehension of the unity and equality of all people, of the unity of the noosphere, is now on the agenda. . . It has become clear, and is more deeply penetrating the consciousness of mankind, that we now have a real opportunity to obliterate malnutrition, starvation, and poverty, to greatly diminish the incidence of diseases, and prolong human life to the maximum."12 Other thinkers like R. Buckminster Fuller have uttered similar statements, that we have all that is needed to adequately take care of everyone on the planet. It is important to understand why this optimism has yet failed to be substantiated. If anything, the inequality in wealth between that of the several hundred billionaires in the world and that of the poorest one-third of humanity, two billion humans, was never historically more disproportionate than it was at the moment of the Inevitable Event. This inequality of wealth, too, is a situation endemic to the very processes of the technosphere, otherwise known as the military-industrial complex.

In 1970 an American scientist, G. E. Hutchinson, wrote in the *Scientific American*, "By noosphere Vernadsky meant the envelope of the mind that was to supersede the biosphere, the envelope of life. Unfortunately the quarter century since those words were written has demonstrated how mindless most of the changes wrought by man on the biosphere have been. Nonetheless, Vernadsky's transition in this deepest sense is the only alternative to mankind shortening its lifetime by millions of years."13

A similar assessment was put forward in 1979 by the Belgian ecologist P. Duvigneaud in his presentation "Noosphere and the Future of the Vegetation of the Globe." In his address regarding the extermination of natural ecosystems made to the Twelfth International Botanical Congress, Duvigneaud makes one of the earliest and most prescient uses of the word *technosphere:* "Thus, the noosphere, so dear to Vernadsky . . . is running the risk in reality of turning into the technosphere, or rather into 'the sphere of avidity,' governed by the spirit of fortune-hunting, by mediocrity, by the absence of social consciousness, by the ideal of destruction and by the egotistic doctrine of 'after me, the deluge.'"14

To be precise, it was in 1973, the same year of the Club of Rome's landmark publication *The Limits of Growth,* and of the establishment of the World Trade Center in New York, that the technosphere became a coherent structure smothering the biosphere, co-opting the noosphere, and coordinating the very processes of greed and destructive fortune-hunting abhorred by the ecologist Duvigneaud. Like it or not, the techno sphere was the inevitable result of the unloosing of the machine into the biosphere. It may be asked: if the machine is a projection of human biology, then what purpose does it actually serve in the biosphere, and what is it that actually produced it and keeps it going?

Before turning our attention to the answers to these questions, let us first define the technosphere as being the artificial structure replacing the human civilization,

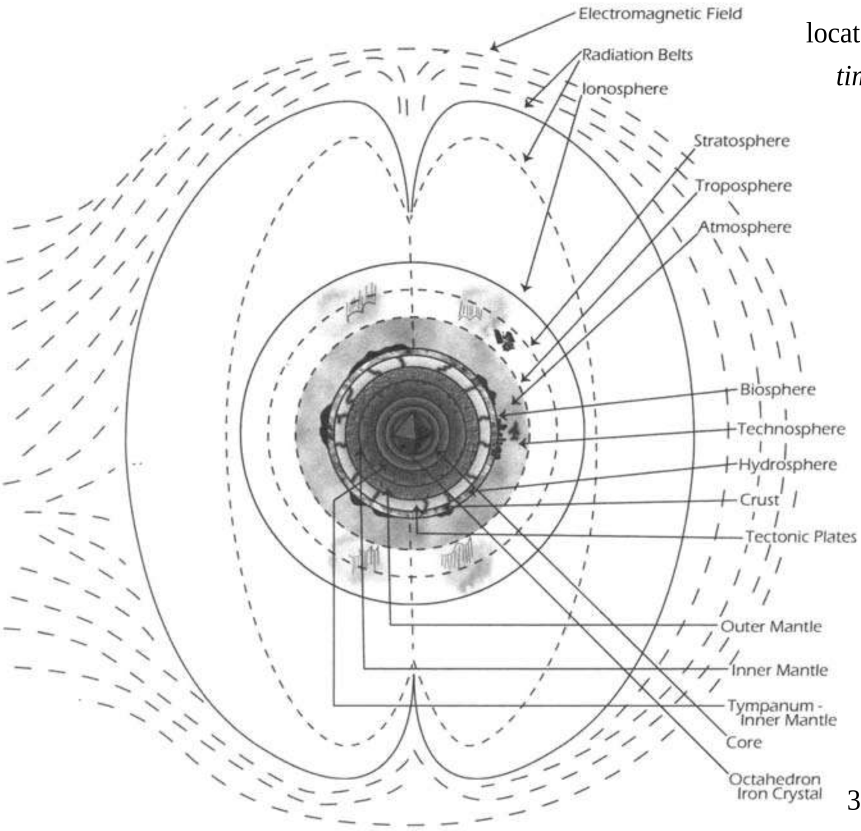
Electromagnetic Field
Radiation Belts
Ionosphere
Stratosphere
Troposphere
Atmosphere
Biosphere
Technosphere
Hydrosphere
Crust
Tectonic Plates
Outer Mantle
Inner Mantle
Tympanum - Inner Mantle
Core
Octahedron Iron Crystal

Earth's Mantles

located mentally and physically *in time* as a force intermediate to and between the biosphere and noosphere. Defined by the *Biosphere Catalogue* (1985, p. 103), the artificial structure of the technosphere consists of five interlocking, coordinated systems that comprise an integral whole, both as a belief system and as an actual complex technological dynamic and its coordinating apparatus. The five components, which we shall discuss at length in later chapters, are: 1) commodity production, 2) the city, 3) energy, 4) transport, and 5) communication. And, according to the *Biosphere Catalogue,* the planetary core of this interlocking set of systems was none other than the World Trade Center's Twin Towers.ls

As a planetary sphere, the technosphere takes its place with the other spheres or mantles comprising the structure of the Earth as a whole system. Beginning at the outermost etheric realm and proceeding inward, these spheres or mantles include: the electromagnetic field with its radiation belts, then the ionosphere, stratosphere, troposphere, atmosphere, technosphere, biosphere, land-ocean crust, tectonic plates, lithosphere, outer and inner mantles, and finally the core, now known to consist of an outer tympanum, and within it a large octahedral iron crystal afloat in a sea of magma.

As a planetary sphere, the technosphere negates time and consciousness as organic realities and replaces them instead with mechanization and marketing. Yet the technosphere is a product of mankind, and mankind is a function of the biosphere. The organization of life into biomes, large macro-ecosystems, is still evident, though in diminishing and withering proportion to the conquest of the technosphere. With the Final War now upon us, the question we must ask is: will the technosphere collapse before the biosphere or not? As catastrophic as it might seem, the collapse of the technosphere would actually be a blessing in disguise from which the noosphere could finally be realized. But if the biosphere collapses first as the result of the Final War on Earth, then the technosphere will be futile, pointless, and dead as well. These are most dire thoughts in these most dire of times. Yet if we can now finally understand and accept our role within and subordinate to the biosphere as a whole, then our chances for survival will be greatly improved.

Let us consider again what Vernadsky has to say about the human's place in the biosphere:

> 1. Man, as observed in nature, like all living organisms, like any living matter, is a specific *function of the biosphere,* within its specific definite space-time.
>
> 2. Man in all his manifestations constitutes a regular part of the structure of the biosphere.
>
> 3. The *explosion* of scientific thought in the twentieth century *has been prepared for by the entire history of the biosphere* and has its deepest roots in the structure of the latter. The civilization of "cultured mankind," insofar as it is a form of organization of the new geological force which has formed in the biosphere, *cannot be interrupted or destroyed,* since it is a great natural phenomenon, historically, or rather, geologically corresponding to the established orderliness of the biosphere.[16]

If Vernadsky is correct, then the war we are witnessing is the inevitable conclusion of the biogeochemical combustion that consumes the technosphere and establishes the pristine reality of the noosphere. This is especially true if we understand this war to actually be a geopolitical struggle for the control of the oil reserves of the Persian Gulf, or even more broadly, as the war between matter (globalization) and spirit (Islam), or more grimly, as Oswald Spengler put it, the final war between blood and money. This being the case, it is more important than ever to answer the two unresolved questions in Vernadsky's definition of the biosphere, the question of consciousness and the question of time.

According to the Law of Time, the answers to the questions posed by Vernadsky are actually one and the same thing. "How can processes which seem purely physical be affected by consciousness?" The question asked by Vernadsky at the conclusion to *The Biosphere* is equaled by the unresolved issue of time discussed at the conclusion to *Problems of Biogeochemistry* JI.1n essence, Vernadsky declared, we know everything about the biosphere, its governing laws, principles, and functions, with the exception of the mysterious relation of time to the biosphere. Virtually alone among modern scientists, Vernadsky perceived that time actually has nothing to do with the metrics of space, and that "time is not a dimension of metric geometry." This means that time is not only its own dimension-the fourth dimension-but that time must also have its own metrics and system of measure totally apart from the metrics of space.

> Minkovsky's and his predecessors' concept of time as the fourth dimension of space is a mathematical abstraction having logically no ground in scientific reality. Time is not a dimension of metric geometry. In geometry, time may be expressed vectorially, but it is obvious that such an expression does not embrace all of its properties in natural phenomena studied by the naturalist. . . The time of the naturalist is not the geometrical time of Minkovsky, nor is it the time of mechanics and theoretical physics.[I?]

In Vernadsky's perception of time, there is a profound critique of the approach that has been established toward the measure-and hence understanding-of time in physicalist science. For Vernadsky, time is a function of an as-yet-unrealized geometry: "This geometry would reduce all space to a point supplied by an infinitesimal vector." Clearly time is a factor in all the functions of the biosphere, most evident in the biological life processes. Hence the Law of Time has been a regulating factor, the true nature of which we have been profoundly unconscious. Now here is an interesting reflection. If the biosphere as a whole system is tending toward a total whole system transformation into a state known as the noosphere, a fact perceived by Vemadsky during his lifetime, and the two unresolved issues are the issues of time and consciousness, both materially intangible dimensions, does it not seem correct that the resolution to these two issues will actually foster the manifestation of the noosphere, which is, after all, the mental envelope of Earth? And does it not also seem that as long as the two issues of time and consciousness remain in a state of ignorance or confusion within the mental functioning of the human order, that the vehicle of geochemical combustion, the technosphere, will tend toward its own dissolution until the issue of time and consciousness is resolved and clarified once and for all?

The perilous passage, the biosphere-noosphere transition, may actually be defined as the consequence of operating in ignorance of the actual nature of time and consciousness. In this way, the techno sphere may be defined as the artificial compromise in place of the true understanding of time and consciousness of the human order in the biosphere. Hence, the technosphere is the responsible intermediary agent creating the crisis of the noosphere. The solution to the questions of time and consciousness is thus the solution to the noospheric crisis, and the commencement of the grand new era, the true golden age of the whole Earth.

◆

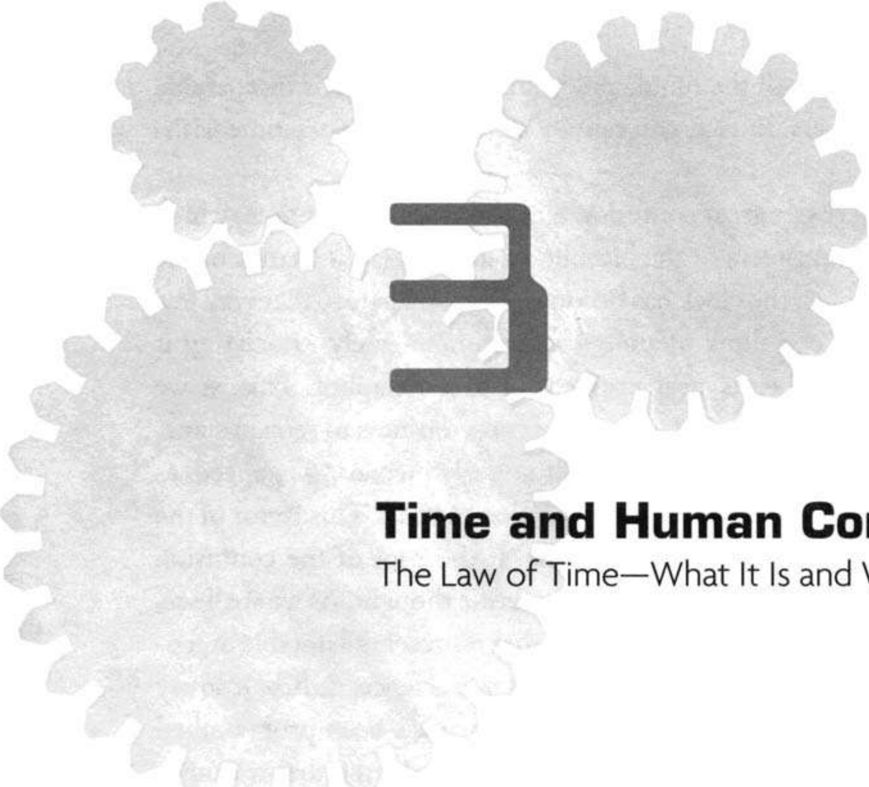

3

Time and Human Consciousness

The Law of Time—What It Is and Where It Came From

HAVING DEFINED the technosphere as the intermediary stage in the biogeological continuum of the biosphere's evolution into the noosphere, we may now turn our attention to the unresolved issues of time and consciousness in the biosphere. These issues must not be seen as just the problems of humanity. Since humanity is a function of the biosphere, the issues therefore relate to the biosphere becoming more conscious. It is the conscious aspect of matter that directs the biosphere toward its transformation into the noosphere. It is the discovery of the Law of Time that complements and resolves the issues that were left unresolved in Vernadsky's definition of the biosphere and of its transformation into the noosphere. The question then arises: what is the Law of Time, and how does it define time and consciousness as mutually coordinating factors of the same cosmic dimension intersecting with the biosphere?

To say the least, time is such a vast and important topic in the orientation of human consciousness within the biosphere that we may declare that it is paramount in all human affairs. Indeed the cosmic order, of which the biosphere is the dynamic regulatory mechanism on planet Earth, is itself the expression of time as the medium of universal manifestation. Yet, as Vernadsky perceived in 1944, while we have defined space by its own systems of metrics and geometries, by applying these same standards to time we have not defined or known time at all. The cause for this quandary seems to lie in the fact that while space is perceptible as a sensory medium—we can see it, touch it, hear through it—and hence, it is evidently measurable, the same is not the case with time. Time is of the mind, and in this regard, time is inseparable from the issue of consciousness. In fact, you cannot know or experience time without becoming conscious.

Yet, as we shall see, there is a great confusion when we speak about time, which is actually at the root of the problem of the biosphere, and that is the matter of the clock. In human consciousness, the clock has become so second nature that very few people in industrial society can think of time without immediately associating it somehow with the clock, whether as an instrument or as a metaphor. This, as we shall see, is a major factor causing the general state of consciousness to remain static, especially in relation to the machine, which is continuously increasing in speed, as well as accelerating in the propagation and multiplication of itself. This factor of the clock-in other words, of mechanized time-is also at the root of the confusion concerning the nature of time in much of modern scientific thought. As we shall see, there is a stream of Russian thought after Vernadsky that represents a notable exception to the established view of time in modern thought and science. Suffice it to say that the evolution of modern Western thought and science has been programmed and predisposed to limit its consciousness of time to such a degree that it cannot even perceive of time outside its inherently mechanized perceptions about it. In this obviously also lies the problem of the delimitation of consciousness in the technosphere.

"The deepening crisis of the world outlook is rooted in the permanence of the dominating scientific paradigm," writes Russian scientist Maria Maroushkina. "Underlying the materialistic technocratic civilization is the Newtonian-Cartesian paradigm, formed in the Middle Ages, its nucleus being the laws of gravitation and movement, as well as the logical principles of proof, laid down by Descartes."! This "permanent" stratum of belief, a cultural metaprogram, locks every field of human endeavor into a conquest by mechanization. It is this metaprogram rooted in the simple unexamined device of the clock, and orchestrated by the equally unexamined effects of an irregular calendar, that keeps human consciousness unchanged and without any real solutions to its problems. As the problems worsen, the confusion of consciousness-which is really in a state of crisis-only deepens.

To illustrate the current state of confusion regarding time in Western thought, we have the report of the prestigious Seven Pines Symposium on Time held in Stillwater, Minnesota in June, 2001. The *New York Times* article regarding the Seven Pines Symposium is entitled "Physics' Big Puzzle Has Big Question: What Is Time?"2 First of all, the assumption that time is basically a topic of physics in itself is a problem, since time, as we have demonstrated, is actually mental in nature, and if any-

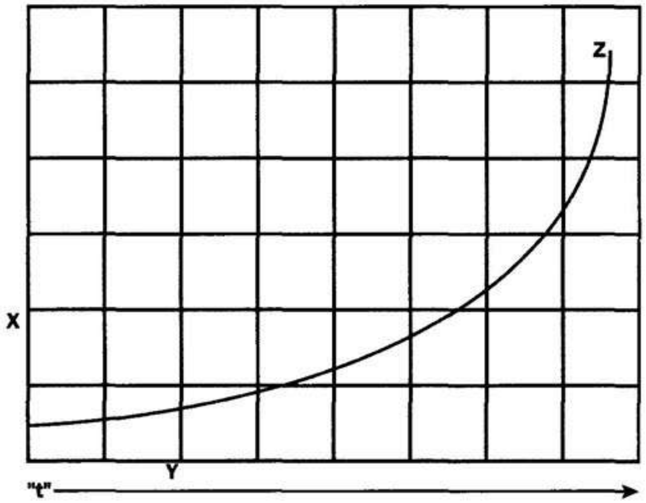

"Arrow of time" in modern scientific three-dimensional coordinate graphing, reinforcing illusion of linear time.

Arrow of Time

thing is most visibly manifest in the domain of biology or biogeology. But this headline is also reflective of the paramount role that physics plays in the Western intellectual hierarchy: if physics can't solve it, than what can?

Despite the prominence of physics and physicists, the Seven Pines Symposium included historians and philosophers as well. The hope of the symposium was to arrive at a unified theory that encompasses the effects of gravity as described by Einstein's general theory of relativity and the "fuzziness that occurs in the realm of tiny particles according to quantum physics."3 Here we have the problem of time as delimited by the self-imposed spectrum of modern physics. What then could we expect? Certainly no one here had considered the point made by Vernadsky that time is not a function of the metrics of space, which is the realm of physics. So it is not surprising that the symposium ended in a muddle, with the most prominent approach being that time could merely be a "psychological illusion" important to humans but not a fundamental part of any unified theory. It is also clear that few, if any, had considered that their views on time might be absolutely conditioned by the mundane and ordinary everyday micro- and macro-organizing factors: the mechanical clock and the Gregorian calendar with its second cousin, the Julian count of linear time.

From the point of view of the noosphere, of course, this grand Symposium on Time was reflective of very little of the entire human organism, much less of the biosphere itself. The symposium was a representation of the elite intellectual class of the dominant economic, military, and political power of the planet, and hence of the governing force of the technosphere, the United States of America. But if a realization that we are a function of the biosphere is critical to our success in launching the noosphere, and the issue of understanding the correct relation of time to the biosphere is crucial even to the ability to launch the noosphere, then the Seven Pines Symposium did not even approach the actuality of the current biospheric situation or the nature of time. In this regard it was only a contribution to further the psychoatmospheric density-clouded thought without clarity of purpose-and one more reason why the Inevitable Event was on the biosphere's agenda.

If one does not apply certain biospheric yardsticks to the analysis of what is reported as truth, then we will be sunk in a chaotic morass of half-truths that are really obscurations or even lies. The elite American nation, for instance, constitutes only 5 percent of the world's population, yet consumes 40 percent of its natural resources, and therefore has the wealth and power to maintain the military force (allotted a $318 billion budget for 2001-2002) to protect and even further as its right this disproportionate ratio, which is actually a violation of the dynamic balance of the biospheric processes. For this reason, through its official organs of propaganda such as the *New York Times,* America is able to muster the "influence" of whatever occurs within its confines as somehow having more merit than what might be generated in some other nation or country. So a human may read the story in the *New York Times* on the "Big Puzzle of Time," and most likely they will complacently accept its premise as absolute truth. Yet this social construct of symposium-and-newspaper-story is nothing but an aberrant mental form that obscures the actual nature of time and keeps the mental level at a constant state of "industrial dull-normal," far removed from-or at best, little more than a highly distorted reflection of-the reality of the noosphere.

Leaving aside this example of the mental confusion over time, it is necessary to understand from the investigations regarding the biosphere and according to Vernadsky's own judgment, that the biosphere in all its intricate and interlocking processes possesses and is guided by a great order and even a structural dynamic that, being manifestations of the laws of nature, can be given exact mathematical formulation. Why should we not expect that the nature of time in the biosphere should also be characterized by the same orderly quality? In fact, given that the biosphere is actually a life process, like all living matter it too must be governed by timing cycles. Thus the larger order of time that governs the cycles of the biosphere and of the noosphere must also exhibit a grand ordering principle. For if the three dimensions of space govern the domain of the physical world, then the fourth dimension of time-which governs the life not only of the biosphere but of planets, star systems, whole galaxies, and indeed the entire universe-must also consist of a simple, orderly formulation totally apart from the perception of time that governs the stunted consciousness of the technosphere. Such is the Law of Time.

THE FOURTH-DIMENSIONAL NATURE OF TIME

> *"Even though you do not measure the hours of the day as long or short, far or near, you still call it twelve hours. Because the signs of time's coming and going are obvious, people do not doubt it. Although they do not doubt it, they do not understand it. Or when sentient beings doubt what they do not understand, their doubt is not firmly fixed. Because of that, their past doubts do not necessarily coincide with their present doubt. Yet doubt itself is nothing but time."*
>
> -DOGEN, "THE TIME-BEING"4

These prescient words of a thirteenth century Japanese Zen master are more true today than they were when he wrote them. Comprehensive studies like the *Eranos Yearbook III: Man and Time* (1957), which included Carl G. Jung's ground breaking essay on synchronicity, or James T. Frazier's fascinating and monumental *Voices of Time* (1966), only expand the multidisciplinary breadth of the confusion about time. In the 1970s Frazier helped found the International Society for the Study of Time. The Society is still meeting annually but has yet to reach a definitive agreement on or understanding of the question of time. It is in the context of the enduring confusion over time that the Law of Time was discovered. To say that the Law of Time has been "discovered" only means that a principle and fundamental law that has always guided the order of the universe has finally been made conscious and articulated in a precise way that is recognizable to the present condition of the human mental nature, which is still dominated by the physicalist model.

In the whole of modern physics, Minkowski's fourth dimension of time has been misunderstood and treated as if it were a minor dimension, one that is just added on, but not really significant to the level of the three dimensions of space. This is due to the already space-oriented consciousness deriving from certain sets of perceptions established early in the history of civilization and standardized through certain programming models that are based on the metrics of space and not time, a topic to which we shall soon turn. It is for this reason that in the formulations of modern physics time is symbolized by a small *t* that runs in a horizontal line running from left to right at the bottom of any graphing of space with its *x* and *y* coordinates.

This graphing of small *t* time gives rise to the notion of the "arrow of time," which is virtually a bedrock dogma of much of modern Western physics. This is the much-touted notion of linear time that supports doctrines of economic inexorability, material progress, and the like. Of course to anyone versed in the ways of nature, linear time and the arrow of time can be seen as nothing more than artificial constructs, for the biological nature of time is perceptible at the very least through the great interlocking cycles of nature. Nor is there anything linear about the Earth's rotations around the sun, or of the moon around the Earth, all of which give lie to the physicalist and materially progressivist notion of linear time. Only the technosphere, sustained in its structure by the limited consciousness of time, is driven by this linear principle and for this reason is doomed to run aground on the shoals of its own artificiality.

TIME AS THE UNIVERSAL FACTOR OF SYNCHRONIZATION

The nature of time is grasped poetically in the form of cycles: "To every season there is a time. . ." Biorhythms can even be mapped and the cycles of the biological order of reality can be demonstrated as functions of a chronobiology. But the biological cycles only demonstrate the manifestation of time in the space ofliving matter. If the whole of nature is observed, then one can begin to grasp the masterful orderliness by which every least detail occurs in relation to all the other details in their various cycles-yes, one can grasp in this phenomenal order another deeper aspect of time, and that is time as the universal factor of synchronization. This is the synchronic order of time, the order by which everything in the universe occurs simultaneously in a masterful synchronization from moment to moment in an ever-changing kaleidoscope of infinite varieties of order and harmony.

Only historical man deviates from this masterful symphony of time, the synchronic order. Nonetheless, we speak of being "on time," "in sync," "in tune," "in resonance," "in harmony," or of being "tuned-in." All of these expressions refer to the attunement of time and the nature of consciousness as awareness in the present moment, in the here and now. This is possible because the same law governs both time and the relation of consciousness to the moment. To answer Vernadsky's question, "How can processes which seem purely physical be affected by consciousness?"- the physical is affected by consciousness, because consciousness is actually a function of time. When someone "discovers" a law of nature, it is because their consciousness is attuned to the natural process in time, and the law therefore "reveals" itself.

But this consciousness can be and most often is of a profoundly unconscious nature. Time is actually the governing principle of a higher self-existing consciousness that regulates the order of the universe, both in regard to its living and nonliving matter. Because time continuously synchronizes everything into a single coherent whole from the micro to the macro levels, time accounts for the harmony of the universe. Only man deviates from the universal harmony, and the technosphere, a projection of the human mechanization of time, intrudes upon the intrinsic harmony to the point of its own demise, but man deviates for a purpose.

The harmony manifested by time as cosmos-which literally means "order" -gives rise to the great and simple formulation of the Law of Time: T(E) = Art; Energy factored by Time equals Art. This means that energy, any manifestation of the physical three-dimensional world, possesses order and is in harmony with its environment because it is factored by time. In the conception of the Law of Time, however, capital *T* refers to time as the universal factor of synchronization. As the universal factor of synchronization, time is defined by the self-existing and intrinsically perfect mathematical ratio 13:20. Derived from the mathematics of the ancient Maya, this ratio is a universal constant of time that organizes all of the universe as a radial sequence of synchronous moments reflecting different evolutionary phases simultaneously. Because everything perceived is an aspect of instantaneous universal synchronization, time is also the medium of instantaneous information transmission throughout the universe.

The noted Russian astrophysicist N. A. Kozyrev (1908-1983), whose work began half a century ago with the theory of the internal structure of stars, later conducted a famous series of experiments using telescopes, mirrors, and aluminum, in which he was able to confirm a receipt of information from distant galaxies prior to the time it took the information to arrive via light as a physical medium. In Kozyrev's own words: "The tests proved the existence of the effects through time of one material system upon another. This effect does not transmit a pulse (momentum), meaning it does not propagate but appears simultaneously in any material system. In this manner, in principle it proves possible to have a momentary relationship and a momentary transmission of information. Time accomplishes a relationship between all phenomena of nature and participates actively in them. . . . Time contains the entire universe of still unexplored phenomena."5 The results of these experiments led Kozyrev to various conclusions, among them that time and duration are not the same thing, and that the velocity of time is instantaneously infinite, formulated as

$$T_V \longrightarrow \infty$$

Formulation: velocity of time is instantaneously infinite

This being so, time must also be the factor or medium that accounts for the experiences of telepathy or other paranormal phenomena. A second, most provocative conclusion derived from his experiments is that time is actually generated as an instantaneous and even simultaneously radial transmission of information from the core of stars, a fact confirmed by the Law of Time's description of the synchronic order of time as being radial in nature.

FURTHER EXPANSES OF FOURTH-DIMENSIONAL TIME: THE EVOLUTION OF TIME AS CONSCIOUSNESS

Leaving aside for the moment further discussion of Kozyrev's seminal work and its conclusions, at this point let us summarize the two chief qualities of time as defined by the Law of Time: 1) time is the factor of universal synchronization, and 2) time possesses an infinite instantaneity of velocity that is faster than the speed of light, and hence is correlated with telepathy. From these two determining features of the nature of time, two other corollary conclusions may be drawn: 3) the cyclic nature of third-dimensional biological phenomena is a function of the factor of universal synchronization which establishes harmonic ratios that determine the different life processes of different organisms, while 4) the ability of any species or organic whole to maintain its unity with itself and all of its members or parts is an instantaneous function of time as telepathy.

To better understand the Law of Time and its radical non-linear definition of time-which is actually the noospheric description of time-we may say that time, being synchronically instantaneous, is both "vertical" in relation to space (which is "horizontal"), and "radial" from the perspective of its own dimension, the fourth. Rather than being the small *t* line or arrow at the bottom of the physicist's graph, time as the fourth dimension is greater than and inclusive of the third dimension, much as the etheric atmosphere includes and surrounds the physical Earth in a radial "all-at-onceness." To try to measure this phenomenon with the metrics of space and its essentially linear yardsticks is to constrain time into definitions and descriptions that are totally alien to its nature.

In fact, from the point of view of the dynamics of time, it is space that is the moving line, or rather a set of points in a moving vector that can be described as line. Let us extrapolate from the important observation made by Vernadsky at the conclusion to *Problems in Biogeochemistry II:* if the geometry of time reduces space to a point supplied by an infinitesimal vector, then from the point of view of time, space is an infinitely locatable point. What is this point? It is the point of awareness at which an individual body becomes aware of time. Here, the "subjective" description or experience of time cannot be separated from the issue of consciousness. (See plate 3, Model of "Vertical Time" in Relation to Horizontal Space.)

In the description of time as vertical in relation to horizontal space, this point of awareness of a body in time is defined as the Locus of Consciousness, or L of C. In the flow of moment-to-moment instantaneity that is synchronous with the flow of space, the Locus of Consciousness is that point at which time vector and space perception connect. From this defining point of consciousness in the here and now, the flow of space creates a horizontal sensation: one direction flows to the "past," the other to the "future." This is what is normally defined as the sensation of time or the arrow of time. It is actually the definition of the horizontal flow of space in relation to vertical time. What is usually not considered is the vertical dimension of time that connects awareness to space at any given moment, creating sequences of Awareness Units, or AUs.

From the defining point of the Locus of Consciousness, there are essentially two spaces and two times: *near space* and *far space,* and *low time* and *high time.* Near space corresponds to low time and far space corresponds to high time. Near space is essentially defined by the perimeter of sense experience, inclusive of the extension of the senses through technological apparatus such as telescopes, microscopes, and the like. This is the domain oflow time. Far space is defined as that space beyond the perimeter of the sense perceptions, and is characterized by subliminal sensory perceptions. This is the domain of high time and the atemporality of space. In near space, low time is characterized by what appear to be "random events." The lower the time, the greater the increase of random events. Low time and near space define the qualities of the experience that are properly and purely third-dimensional. High time and far space are purely and properly the domain of the fourth dimension characterized by the dominance of the synchronic order of time.

The distinction between low time and near space and high time and far space is defined as the threshold of synchronicity. In high time, synchronicity (s) predominates and is greater than random events (re), or (s»(re). In low time, random events predominate and are greater than synchronicity, or (re»(s). Note that the realm of quantum physics occurs in the microrealm of far space. It is this factor that accounts for the "fuzziness" of the behavior of subatomic particles, quarks, tachyons, and so forth, which defy the limitations of linear time. These phenomena occur in the subliminal realm of the synchronic order. It is the human mind that is not accustomed to behavior in the pure realm of synchronicity that cannot grasp that these particles are moving according to radial laws of fourth-dimensional time, and so appear highly irregular and random to the sense perceptions of the observers that are totally conditioned by third-dimensional space perceptions. Similarly, the high time of the macro threshold of the synchronic order represents the "other side," the "hereafter," the realm of pure vision, which is also governed by strict laws of radial time, and which, to the conditioned experience of a perceiver on this side of the threshold, may appear as something from a distant future or past reality, or even as in a dream. The orderly explication of the Law of Time provides complete mathematical descriptions and principles that account for the radial order of time, even as it coordinates the near space and low time of random events.

In any case, the horizontal line of space in relation to the vertical line of time defining any given Locus of Consciousness describes the macroworld above and the microworld below. As we have noted, the higher threshold of synchronicity above the horizon of space defines the realm of *higher consciousness*; the lower threshold of synchronicity below the horizon of space defines the realm of *quantum physics.* Connecting the macro realm of atemporality and high time, and the lower or micro realm of atemporality and high time, is the vertical time vertice of the ever-present now. At the two far ends of this time vertice of the ever-present now is the Locus of God, the selfsame in each direction. These extreme points of the vertice of time return far space to the Locus of Consciousness through a process of toroidal sublimation. We are not really dealing with a description of a two-dimensional plane, but a dynamic four-dimensional model held together by a toroidal motion that is continuously returning the farthest space and the highest time, whether from above or below the threshold of synchronicity-" God" -to the Locus of Consciousness.

Depending on the clarity of mind at any given moment, the vertice of time constantly and instantaneously transmits information to the Locus of Consciousness. The discrete units of information thus transmitted are referred to as *celestialharmonics.* A celestial harmonic is described as the index of synchronic incidence defining different levels of co-occurrence. I can be in the present moment sitting in a room, and at the same time, by being fully present in the here and now, I can spontaneously experience any number of telepathic thoughts, memories, feelings of deja vu, and so forth. These experiences are the different levels of co-occurrence that define the index of synchronic incidence and are categorizable as celestial harmonics. The higher the time, the greater the density of celestial harmonics per AU (Awareness Unit). It is the vertical vertice of time and its power of transmission that account for the visionary or revelatory experiences of saints, poets, and mystics throughout the ages. It is the mathematics of the Law of Time that provides the "mapping" that explains, identifies, and even increases these different harmonic incidences of the synchronic order.

Because the vertice of time connects ultimately to the "Locus of God" -the mutually defining points that become One at the far ends of the toroidal motion of the vertex of time-this accounts for the existence of the *perennial philosophy*, the unitive nature of the most revelatory or ecstatic experiences of the great mystics and seers. It is also the relation of awareness in the Locus of Consciousness at any present moment that allows any random event of low time to trigger telepathic ascent into high time depending on the clarity of and duration of the Awareness Unit (AU). From this point of view, the random quality of events in near space is only relative in relation to the absolute nature of the synchronic order, and merely a function of the third-dimensional biological space orientation. For those accustomed to the actual nature of fourth-dimensional time, which includes and controls the third-dimensional order of near space, the random events constituting low time are actually "signs" that may be read as message bearing signals of the synchronic order of high time.

Brief reflection on this description of vertical time in relation to horizontal space will demonstrate the value of what is referred to as *meditation without an object* such as that cultivated in the Buddhist traditions, as well as affirm the One Divine Source of the purely monotheistic traditions. Without a clear mind of awareness in the present moment there can be no clear seeing of reality, nor can there be the realization that all of what our senses report to us is constructed in our own mind to conform to what we think is a universe outside of us. At the same time, if we can extend the duration of our awareness in the Locus of Consciousness indefinitely in the vertical direction of time, we come upon the ineffable experience that is an emanation of the Locus of God, however or in whatever "language" our senses may later report or define such experiences. Thus we return to the paradox of the definition of space as an infinitely locatable point-this point is none other than our own minds, in which the construct of space is created and dissolved from moment to moment. Such is the subjective description of time from the point of view of the Law of Time, a description that begs each one of us to take absolute responsibility for our own experience. But there is an "objective" description as well, in which time is also defined as inseparable from consciousness.

Implicit in the subjective here-and-now description of the experience of time is the radial matrix of time. At the experiential level it is the now-centered point in the Space-Locus of Consciousness which defines the center of a radial order of time.

Time radiates out from each here-and-now awareness moment. But the Law of Time also supplies an objective description in which the radial nature of time has two points of determining order: the point of infinite lucidity and cosmogenesis, which is the Locus of God; and the receptive locus of the infinitely locatable point, which is space as constructed by each individual consciousness. Intermediate between the two points-Locus of God and Locus of (individual) Consciousness-is the determining radial matrix of fourth-dimensional time, the universal frequency of synchronization.

Mathematically defined as the ratio 13:20 and described by the 13 x 20 matrix of the Harmonic Module, in this mathematical model of fourth-dimensional time we perceive the root of the Law of Time and its basis in the mathematics of the Mayan time science. It is important to understand very clearly that the 13:20 frequency and the mathematical matrix that defines the operations of this frequency of synchronization are the pure mathematics of time totally apart from the metrics and geometries of third-dimensional space. This is a point of profound significance to which we shall later return.

In the model of objective time, the dynamics of the Law of Time are demonstrated as the evolution of time as consciousness. Here the two unresolved issues of Vernadky's description of the biosphere are unified, the answers to which are the determining factors that allow the triggering of the noosphere: The evolution of time is inseparable from consciousness. Conversely, the evolution of consciousness is inseparable from time. Both time and consciousness are nonphysical fourth-dimensional factors coordinating the third dimension. Because consciousness is synchronized with or by time, it can affect physical third-dimensional processes, in whatever way, whether the acting agent is aware of this or not. Consciousness of the Law of Time, however, is only possible through a profound act of self-reflection. Once it has been made conscious, the discovery of the Law of Time affects the entire medium of planetary consciousness, the noosphere, at first imperceptibly, but then building to a great point of climax coinciding with the climax of the biogeochemical combustion. (See plate 4, The Dynamics of Time: Showing the Evolution of Time as Consciousness.)

This dramatic moment of the discovery of the Law of Time is evident in the graphic description of the dynamics of time, showing the evolution of time as consciousness. The dynamic of the universal, cosmically intelligent, and intelligible whole-the *galacticbrain-is* depicted as a circle encompassing the evolutionary spectrum, showing the index of energy-mind progression from the atomic cellular and biological levels to purely mental conditions of being, all of which are functions of the formulation T(E) =Art. From the point of God-Galactic Ordering Dynamic-the circle, or rather, sphere, is divided vertically and horizontally into four quadrants. The movement of the evolution of time as consciousness travels out of the God center, from right to left, and thence describes, in counterclockwise circular motion, the movement of time as consciousness passing through, or rather defining, the four quadrants, until it returns to the Source again as hyperdimensional time, or "pure mind." The left-hand side describes both the primary prelife, lower left-hand quadrant, and the primary post-life, upper left-hand quadrant. The right-hand side describes the secondary reflex life, the lower right-hand quadrant being the purely third-dimensional life (biosphere), and the upper right-hand quadrant being the evolved fourth-dimensional life (noosphere).

The lower half of the sphere beneath the horizontal line running through the God center refers to the Domain of Potentiality, the upper half to the Dominion of Time. The prelife quadrant of the Domain of Potentiality refers to the preconscious quantum dimensional space of pure energy. The lower right-hand quadrant refers to third-dimensional space, the locus of space travel and the realm of the cosmic unconscious. Within the Dominion of Time, the upper right-hand quadrant refers to the fourth-dimension time, the locus of time travel, and the realm of cosmic consciousness. Finally, the upper left-hand quadrant is the realm of hyperdimensional time, or "pure mind." By studying this graphic description, one can grasp the movement of time as consciousness from inorganic preconscious, where duration in time is virtually infinite, to the secondary reflex life, or living organic matter in the third dimension. Here time is unconscious and life builds on the cosmic unconscious as the slowly mutating and evolving biosphere. The cosmic unconscious comes to a climax in the creation of the 12:60 artificial civilization, which is really not true civilization, but rather the technosphere itself. At this point there are moments of consciousness, but only moments. True sustained consciousness is possible only in conditions removed from the technosphere.

In the evolution of time as consciousness, the discovery of the Law of Time is a product of the crisis of consciousness unable to sustain itself in the technospheric medium. That is because this medium is a pure function of artificial 12:60 time. Even though the extensions of the senses and the capacity of technology provide the human within the biosphere a holistic perspective of seeing the whole Earth from space, the consciousness is continuously subordinated to the care and maintenance of the machine (or its lifeblood, money), and so cannot remain in a sustained state of heightened awareness or continuing consciousness. Continuing consciousness is only possible with the discovery of the Law of Time, which the crises of the technosphere and of consciousness have themselves brought about. In this regard, the discovery of the Law of Time is the great point of human self-reflection that defines the advance from an exclusive third-dimensional operation and aberrant timing factor to a purely fourth-dimensional and conscious natural timing factor. It is the difference between the cosmic unconscious and cosmic consciousness, the biosphere and the noosphere-in other words, the very topic of this book.

DE CHARDIN AND THE PLANETIZATION OF CONSCIOUSNESS

The discovery of the Law of Time defines the point between the two stages of the noosphere as described by Pierre Teilhard de Chardin: "The first stage was the elaboration oflower organisms, up to and including man, by the use and irrational combination of elementary sources of energy received or released by the planet. The second stage is the super-evolution of man, individually and collectively, by the use of refined forms of energy scientifically harnessed and applied in the bosom of the Noosphere, thanks to the coordinated efforts of all men working reflectively and unanimously upon themselves. . . In becoming planetised humanity is acquiring new physical powers that will enable it to super-organize matter. And, even more important, is it not possible that by the direct converging of its members it will be able, as though by resonance, to release psychic powers whose existence is still unsuspected?"6 The identification of telepathy with the velocity of time emphatically underscores de Chardin's final point.

On the other side of the discovery of the Law of Time, civilization, having reached its climax as the technosphere, is replaced by an entirely new noospheric order of human organization defined as PAN: Planet Art Network. This is a structural definition of what Teilhard de Chardin refers to as the "planetization of consciousness." This organization (bionoospheric and not institutional) is "planet" because the human organism is self-realized as a planetary organism. It is "art" because functioning again in the natural timing frequency, the human is governed by the Law of Time, T(E) = Art, and hence the reality and the activities of everyday life are defined as art. And it is a "network" because the city, defined by civilization (literally living in cities) and which was a sustaining component of the technosphere, is no longer necessary or viable. Rather, the human becomes redistributed throughout the biosphere-what Teilhard de Chardin refers to as the "radiation of man "--connected by a telepathic network that greatly diminishes the reliance on much of the technology that characterized the technosphere. Hence Planet Art Network: the replacement of civilization with the fourth-dimensional social organization of the human functioning as the telepathically coordinating species of the noosphere. In this condition life evolves toward a hyperorganic stage, attaining a state of superconscious activity and self-reflection where time travel replaces the space travel of the phase known as civilization.

This evolutionary mode, coinciding with the evolving needs of the local star, conduces at some distant point in the future to the subliminal consciousness and the postorganic postlife that is necessitated by stars as they advance to the stage of supernova. According to the Law of Time, the nature of consciousness is actually a function of stellar evolution, planets being but harmonic nodes in greater heliospheric fields of resonance. Since time is inseparable from consciousness, this understanding would seem to affirm Kozyrev's hypothesis that time is generated from the stellar core. "Stars absorb energy from the motion of time," declares Kozyrev, "a star seems to be a time machine."7 *The Mayan Factor* further places the coordinating medium of time as emanating from the core of the galaxy, Hunab Ku, One Giver of Movement and Measure. In the Chilam Balam prophetic tradition, Hunab Ku is also the name given to the final religion of the One God that appears after the religion of the *conquistadors* has finally passed. Seen from the perspective of the Galactic Brain, the crisis through which we are now passing is but the final stage before consciousness bursts into a new radiance of biosolar-telepathic wonder where negative emotions and thought forms no longer have any place to cling or hide. The Law of Time purposively points to this great end.

The whole of the spectrum of the Galactic Brain is mediated by the universal frequency of synchronization, mathematically modeled as the 13:20 matrix, the organizing factor of the synchronic order. The technicalities of this evolutionary spectrum of the Galactic Brain have been defined in *The Dynamics of Time: 260 Postulates* (1996) so that the process of the evolution of time as consciousness may be studied and its coming stages prepared for. The mathematical model of this 13:20 matrix is a description of the fourth-dimensional timing frequency and is both the source of the time vector potentialities and the mathematical tool for mapping the order of synchronicity. Through the discovery of the Law of Time, the 13:20 matrix is realizable through tools that are in accord with its mathematics, and thereby it becomes useful to the everyday consciousness of the human. This too becomes a factor accounting for the noosphere's becoming conscious.

The fourth-dimensional "tools" made accessible through the Law of Time include the Thirteen Moon/28-Day calendar, the Dreamspell, and the Telektonon. They will be described more fully in Chapter 8, "Making the Transition to the New Time." With these tools one can begin to construct a map of the time vector potentialities contained in the synchronic order as it manifests on a day-to-day basis. Once the construct is made conscious and correlated to a point in third-dimensional time, then the body in time-the infinitely locatable point that is space-may be coordinated by the same common structure of fourth-dimensional time. Following this procedure the individual human consciousness begins to participate in the larger unfolding of the noosphere, and slowly but surely is realized as yet another manifestation of the formulation of the Law of Time, T(E) = Art. It is the human being itself that incorporates time, or rather wakes up within itself the order of time. Time is art, while art is actually a function of consciousness expressing itself.

THE MAYAN UNDERSTANDING OF THE LAW OF TIME

While we will have cause to turn our attention later in this book to the "how-to" of fourth-dimensional time, we may consider further the nature and meaning of the mathematics of the Law of Time and their origin with the mysterious civilization of the ancient Maya. The template or matrix of the 13:20 frequency, a 13 x 20 (260) unit structure, was the means for accommodating the base "calendar" used by both the ancient and present day Maya: the *Tzolkin,* or sacred calendar, a set of thirteen numbers and twenty signs that repeat their sequences of permutations every 260 days. While this 13 x 20 matrix accommodates the Tzolkin, it is not confined to being a calendar, as I amply demonstrated in *The Mayan Factor.*The 13:20 frequency is the universal constant of time from whose mathematics the Tzolkin is derived. Based on the permutation table, called the *buk xok* in the Chilam Balam tradition, this matrix is a Harmonic Module, a fractal yardstick of radial time with multiple applications, something analogous to the periodic table of elements, but of the synchronic order.

It is still a profound challenge to all modern Western scientific thought that the Maya would have possessed such mathematics, apparently inherited from the earlier and even more mysterious Olmec people, and that with this mathematics they would have evolved such a profound perception of time. This was most certainly a factor in the destructive attitude taken toward the Maya by the religious zealotry of the conquerors of Mexico. It is also astonishing that the mathematics in which the Harmonic Module is embedded is a vigesimal (20) count, utilizing a zero and a positional ordering system that accounts for larger or greater values that advance by binary exponentiality rather than a decimal sequentiality. But, quite simply, this system exists in this manner because it is the mathematics appropriate to time. This mathematics is a factor to which any disinterested researcher on the topic of time must now become accustomed.

It is by reason of this mathematics and its ability to "download" the 13:20 universal frequency of synchronization that the ancient Maya were able to construct a complex system of "calendars" and an astronomical knowledge that is precise not for its telescopes, of which there were none, but because of the nature of the mathematical system itself. More appropriately described as synchronization devices or synchronometers, the Maya used at least seventeen of these timing instruments simultaneously during the great age of their civilization, A.D. 435-830. This fact in itself tells us that the Maya understood time so radically differently from the perception that has evolved in the techno sphere that we must stand in awe that such a knowledge would have developed by a people whose technological development was so minimal, but whose artistic accomplishment was inseparable from their scientific achievement.

Harmonic Module

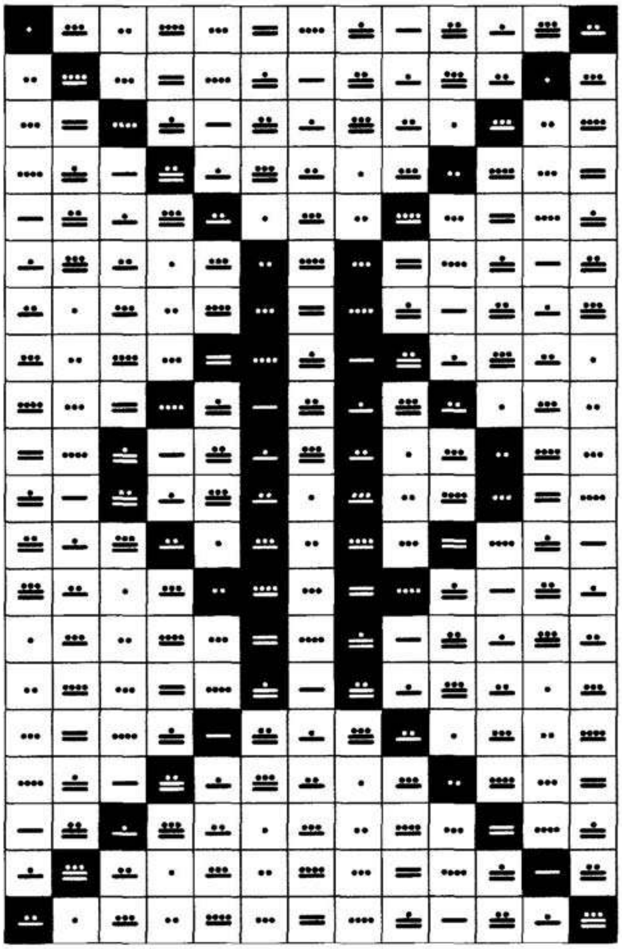

Yes, the Maya most definitely knew that time is the universal factor of synchronization, so their synchronometers must have been for the purpose of attaining a master synchronization with the cosmos. This presupposes such a different state of mind that all the values of the present day must be challenged. The mystery of it is that this mathematics and time science developed in the New World, remote from the saga of civilization in the Old World. The fact that the later Maya practiced ceremonies and blood rituals, the fascination of so many present-day archeologists, should not detract from the understanding that these ceremonial practices stem from a totally different perception of life-and death. Before we judge, we must consider the inescapable truth that nothing in history matches the terror of the different kinds of war and weapons of mass destruction that characterize the fifty-six-year cycle of the technosphere, 1945-2001. Who then is the more barbaric-the Maya, or modern man, *"Homo technosphericus"?*

It is the study of this vigesimal mathematics and its 13:20 Harmonic Module that has led to the discovery of the Law of Time. In this regard the Law of Time is a universal law independent of any culture or civilization, even that of the Maya, much like the law of gravity is also a universal law. However, if it had not been for the Maya, who evolved their understanding of the 13:20 frequency of synchronization to such a high degree, there would never have been a pursuit that led to the discov-

Galactic Notation, Dot-Bar (0–19) Code

Vigesimal count, 1–19, where in first order 1 = 1 and
0 = 20 position, so in next order 1 = 20, in third order 1 = 400, etc.

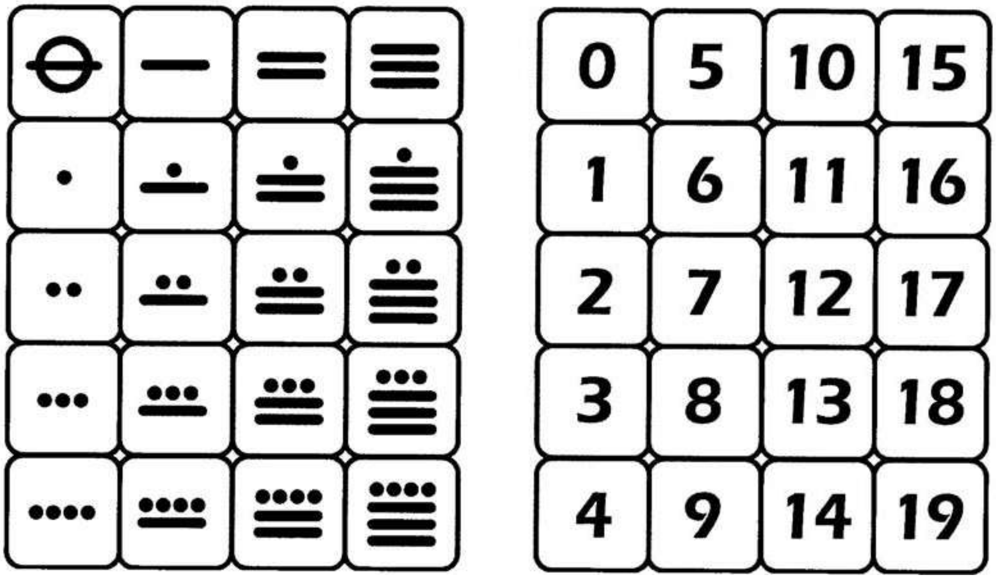

ery of the law underlying their knowledge of time. If with this law we may categorize the time vector potentialities in six stages of consciousness-preconscious, unconscious, and conscious, continuing conscious, superconscious, and subliminal conscious-then we may be certain that the ancient Maya were also familiar with the time science associated with the higher states of consciousness revealed by the Law of Time. With this time science and mathematics of higher consciousness, the Maya were able to bequeath a prophetic tradition mathematically coded in the Harmonic Module, which, among other matters, led to the discovery or making conscious of the Law of Time itself.

Once we appreciate the radically different perception of time and knowledge that is embodied in the Harmonic Module, the mathematical modeling of the 13:20 universal frequency of synchronization, then we may turn to its uses as the fractal radial measure of time applied to many levels of thought and knowledge. For instance, using the 13:20 matrix we may turn to an understanding of the Law of Time in human history, as is necessitated by these reflections. With the Mayan time knowledge and the placement of their civilization in the New World, we have a gauge and an alternative base of knowledge upon which to evolve a planetary perspective and an objective means of critique of the mainstream development of civilization in the Old World. We may then ask the questions: How did we become what we are now, and how was it that our civilization evolved into the technospheric medium?

◆

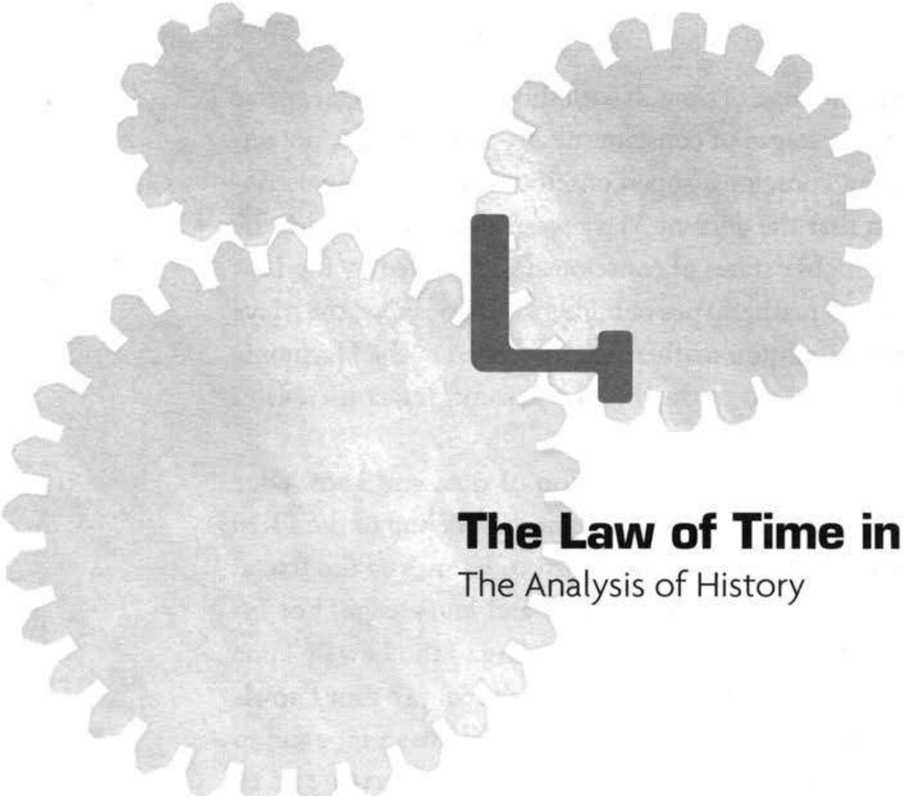

4

The Law of Time in Human Affairs

The Analysis of History

THE KEY to *The Mayan Factor* is the application of the Harmonic Module as a fractal measure that is descriptive of the cycle of thirteen baktuns, the Mayan measure of history. This measure of thirteen baktuns, a cycle extending from 3113 B.C. to A.D. 2012, had been deduced from the stone inscriptions of the Mayan monuments of the Classic period, most of which bear a date within the tenth baktun cycle, A.D. 435-830, or, in the transliterated vigesimal notation of the Mayan Long Count, 9.0.0.0.0-10.0.0.0.0. However, one of the principle discoveries of *The Mayan Factor* demonstrated that the thirteen baktun cycle was also a function of the Harmonic Module and identical with its mathematical mapping as the 13:20 matrix. There are very profound implications to this correlation of the cycle of history having a one-to-one identity with the 13:20 frequency matrix. For this reason the cycle of history is referred to as a *Wave Harmonic* of the galactic order of time. Even though history may appear to be a random, haphazard, and checkered advance of humanity from living in small agricultural communities to the present-day technosphere, this process has actually been an orderly program of the biosphere, the region for the transformation of cosmic energy on Earth.

The same principle of the analysis of the Law of Time applies here. That is, by analogy, the historical events analyzed from a political or economic perspective represent the perception of the random events of near space and low time, while the analysis of the thirteen baktun cycle Wave Harmonic of history represents the perspective of far space and high time. In this Wave Harmonic, the program of history is mathematically defined by 13 baktuns-each baktun consists of 20 katuns-13 baktuns x 20 katuns = 260 katuns = 13 x 20 matrix of the 13:20 frequency of synchronization. In the 13 x 20 matrix of the Harmonic Module each unit equals one kin, the base unit of fractal measure. By fractal equivalence, in the mapping of history one kin = one katun, one katun = 7,200 days; therefore one baktun = 144,000 (7,200 x 20) days. The mathematical perfection of the cycle of history matching the 260-unit model of the harmonic matrix means that history is a factor of synchronization of the higher order of time coordinating the human organism within the biosphere.

By means of the *thirteen baktun Wave Harmonic of history,* the incidence of human Civilizational Advance (CA) may be plotted in a very precise way as a progression going from a very dispersed and scattered state to the point of attaining its exponential climax in the final saga of geochemical combustion, the fifty-six-year cycle of the technosphere. In this regard we must ask: What do we mean by "history" from the Wave Harmonic point of view? As an organism that is a function of the biosphere, history defines that stage when that organism, humankind, enters a process of altering its biomes through increasingly artificial means or extensions of itself that exhibit a compulsive reordering of society into denser and denser clusters of habitations known as cities, hence civilization, "city-life."

In the stage before history, the human had created certain tools and had begun experiments in agriculture and horticulture, but the mental order remained in an aboriginal condition. Unconsciously absorbed as an integral mechanism within the cycles of the biosphere, the human wrapped his understanding of these cycles and of his place in them in the garb of myth and ritual. But with history, something different occurs. The artificial means or extensions of the senses are ordered by a different form of mentality. Those projections of thought and consciousness, though immaterial, affect the material surroundings, and take on an aggressive dynamic. The aggregate effects of these artificial extensions are maintained by a new mental dynamic that is contained and then conditioned by a new kind of macro-organizing thought structure, independent of the actual cycles of the biosphere. What is this new mentality? How can we most precisely define it?

Here are some points to ponder. The Mayan reckoning of the beginning of history is precisely dated to August 13, 3113 B.C., Long Count 13.0.0.0.0. The Hindu reckoning of the beginning of the Kali Yuga is dated to February 19, 3102 B.C. These dates are only eleven years apart. What cosmic event occurred at that time to initiate the Wave Harmonic of history? As we have seen, there is a perfect identity of the

13 Baktun Cycle: Wave Harmonic of History

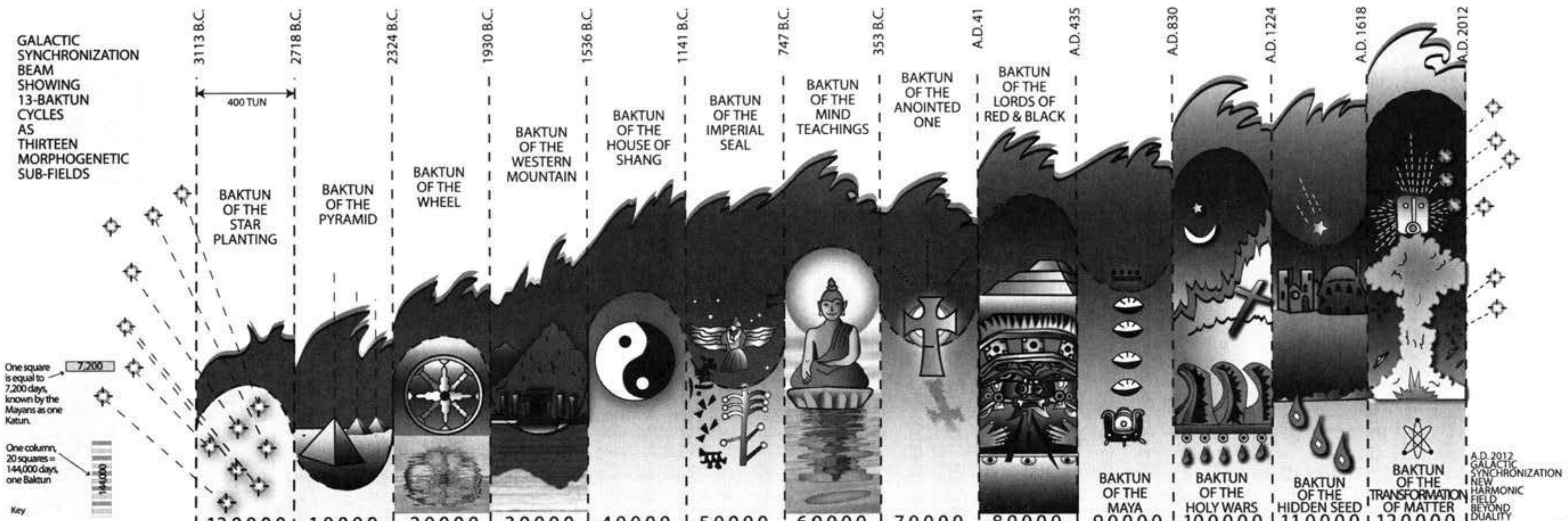
GALACTIC SYNCHRONIZATION BEAM SHOWING 13-BAKTUN CYCLES AS THIRTEEN MORPHOGENETIC SUB-FIELDS
One square is equal to 7,200 days, known by the Mayans as one Katun.
7,200
One column, 20 squares = 144,000 days, one Baktun
144,000
Key
3113 B.C.
400 TUN
2718 B.C.
2324 B.C.
1930 B.C.
1536 B.C.
1141 B.C.
747 B.C.
353 B.C.
A.D. 41
A.D. 435
A.D. 830
A.D. 1224
A.D. 1618
A.D. 2012
BAKTUN OF THE STAR PLANTING
BAKTUN OF THE PYRAMID
BAKTUN OF THE WHEEL
BAKTUN OF THE WESTERN MOUNTAIN
BAKTUN OF THE HOUSE OF SHANG
BAKTUN OF THE IMPERIAL SEAL
BAKTUN OF THE MIND TEACHINGS
BAKTUN OF THE ANOINTED ONE
BAKTUN OF THE LORDS OF RED & BLACK
BAKTUN OF THE MAYA
BAKTUN OF THE HOLY WARS
BAKTUN OF THE HIDDEN SEED
BAKTUN OF THE TRANSFORMATION OF MATTER
A.D. 2012 GALACTIC SYNCHRONIZATION NEW HARMONIC FIELD BEYOND DUALITY

260-KATUN CYCLES
52-UNIT GAIA LIGHT BODY

	13.0.0.0.0	1.0.0.0.0	2.0.0.0.0	3.0.0.0.0	4.0.0.0.0	5.0.0.0.0	6.0.0.0.0	7.0.0.0.0	8.0.0.0.0	9.0.0.0.0	10.0.0.0.0	11.0.0.0.0	12.0.0.0.0	
DRAGON / IMIX	URUK, MENES	ZOSER, EGYPT CALENDAR	SARGON	ABRAHAM	SHANG DYNASTY, CHINA	KING WEN	GREEKS	ALEXANDER	ROMAN	FALL OF ROME	BOROBADUR	GOTHIC	DESCARTES	NEPTUNE
WIND / IK	KILNS	GREAT PYRAMID	BABYLON		AKHNATON	CHOU DYNASTY	KUSHITE	CELTS	EMPIRE	BYZANTIUM		EUROPE	SCIENTIFIC	URANUS
NIGHT / AKBAL	UNIFIED SUMER	GIZA		CRETE	HITTITE	CHINA		ASOKA	TEOTIHUACAN	RISE OF	VIKINGS	MONGOLS	MATERIALISM	SATURN
SEED / KAN	EGYPT				EMPIRE	KING DAVID, JERUSALEM	NINEVEH	SPREAD OF BUDDHISM		CLASSIC	HEIAN JAPAN	KUBLAI KHAN	CHING DYNASTY	JUPITER
SERPENT / CHICCHAN	OLD		CHARIOT	EGYPT: QUEEN	THERA EARTHQUAKE		PYTHAGORAS	ZAPOTECS		MAYA	CHIMU	MAYAPAN	NEWTON	A-BELT/MALDEK
WORLD-BRIDGER / CIMI	KINGDOM		WARFARE	HATSHEPSUT	OLMECS	VEDIC	BABYLON REBUILT	CH'IN DYNASTY	NAZCA	BUDDHISM IN JAPAN	XOCHICALCO	AZTEC	BAROQUE MUSIC	MARS
HAND / MANIK				VALLEY OF	CHÁVIN	CIVILIZATION		HUANG-TI	GERMANIC	TIKAL	CE ACATL	BLACK PLAGUE	INDUSTRIAL	EARTH
STAR / LAMAT			EGYPT	KINGS		INDIA	LAO TZU	EARLY	INVASIONS	POPE GREGORY	TOPILTZIN QUETZALCOATL	INCA	REVOLUTION	VENUS
MOON / MULUC	HEIROGLYPHS	AKKAD	MIDDLE KINGDOM				PERSIAN	MAYA	MOCHE	MUHAMMED	RISE OF	MOGHUL / INDIA	AMERICAN &	MERCURY
DOG / OC		UR			JADE	HORSE WARFARE	EMPIRE	HAN		PALENQUE	TOLTEC	OTTOMAN TURKS	FRENCH REVOLUTION	MERCURY
MONKEY / CHUEN	CUNIEFORM				KING TUT		BUDDHA, MONTE ALBAN		END OF	PACAL VOTAN	MAYA		ROMANTICISM	VENUS
HUMAN / EB		HARAPPA			RAMSES		CONFUCIUS	DYNASTY	HAN DYNASTY		IFE	MING DYNASTY	EUROPEAN	EARTH
SKYWALKER / BEN		CIVILIZATION			ASSYRIA	JERUSALEM		TEOTIHUACAN	TIAHUANACO	COPAN	ANGKOR WAT		IMPERIALISM	MARS
WIZARD / IX		INDIA	ZIGGURATS				PLATO	RISE OF ROME		HEIROGLYPHIC STAIRWAY	CRUSADES	FALL OF CONSTANTINOPLE	ELECTRICITY	A-BELT/MALDEK
EAGLE / MEN				HORSES			CHUANG TZU			SPREAD OF ISLAM	BUDDHISM IN	PRINTING, CONQUEST	RUSSIAN REVOLUTION	JUPITER
WARRIOR / CIB			BRONZE	CENTRAL ASIA	VEDAS	LA VENTA	ZOROASTER	GREAT WALL	CONSTANTINE		TIBET	OF MEXICO, PERU	WORLD WARS	SATURN
EARTH / CABAN			TECHNOLOGY	HITTITES		ASSYRIAN		OF CHINA	BUDDHISM IN	T'ANG DYNASTY	SUNG DYNASTY	EUROPE EXPLORATION	HIROSHIMA	URANUS
MIRROR / ETZNAB			MENUHOTEP	EGYPTIAN	PHOENICIANS	IRON	BANTU, WEST AFRICA		CHINA	MISSISSIPPI CULTURE, BAGDAD	RUSSIA/KIEV	BENIN	NUCLEAR & SPACE TECHNOLOGY	NEPTUNE
STORM / CAUAC	STONEHENGE	PYRAMID	UR, NAMMU, LEGAL CODE	EMPIRE	MYCENEANS	WEAPONS &	PELOPPONESIAN WAR	JULIUS CAESAR	GUPTA DYNASTY, INDIA	PADMASAMBHAVA	ANASAZI	QUEEN ELIZABETH	HARMONIC CONVERGENCE	PLUTO
SUN / AHAU	GILGAMESH	TEXTS	HAMURABI	ARYANS INVADE INDIA		WAR MACHINES	ARISTOTLE	CHRIST		CHARLEMAGNE	ZIMBABWE	GALILEO, KEPLER	EARTH REGENERATION	PLUTO
	𝋠	𝋡	𝋢	𝋣	𝋤	𝋥	𝋦	𝋧	𝋨	𝋩	𝋪	𝋫	𝋬	

SOLAR MOVING INHALATION

GALACTIC/LUNAR MOVING EXHALATION

Wave Harmonic of history, as a fractal equivalence, with the Harmonic Module, itself a description of the universal 13:20 frequency of synchronization. This being so, both the calculation of the beginning of history as well as the duration of its entire cycle are functions of the Law of Time. This means that something within the Earth's noosphere occurred that triggered the Wave Harmonic, itself a manifestation of the Law of Time. Also, by most judgments and standards, history began at a very precise place as well-Sumeria, in the Mesopotamian basin of what is now called the Middle East. Why? And what relation is there between the precise event that triggered the Wave Harmonic of history and Sumeria, the site of the first city, Uruk? What event could have established a macro-organizing principle to propel the dynamic of human thought and technology into the trajectory concluded by the technosphere?

THE ORIGINS OF THE 12:60 FREQUENCY

In the recently published book *The History of the World-A 6,000 Year Chronicle of Time,* we find the point that establishes the definite artificiality of a new kind of macro-organizing principle, driving humankind toward the Inevitable Event; here we find the genesis of artificial time: "Ca. 3000 B.C. Sumerians divide day into 24 hours, 60 minutes, 60 seconds and circle into 360 degrees."! While this is a fact we may take for granted because it is the basis of mechanized time in the form of the clock, for precisely that reason it should give us pause. For one thing, this means that the measuring system of the clock was already established at the beginning of history, some 4,500 years before the clock as we know it was actually invented! This is quite an astonishing fact. It means that mechanization was implicit in the first intellectual act of history. Why or how would it have occurred at the precise point of the beginning of history in Sumeria?

The Law of Time defines this mechanistic organizing principle as the basis of the 12:60 artificial timing frequency (irregular twelve-month calendar combined with mechanical sixty-minute hour). Therefore, it was at this point at the beginning of history that artificial time and its progeny, mechanization, were actually registered as a thought form and thus became embedded in the as-yet-unconscious noosphere. Or, was it already embedded in the noosphere to be triggered at this precise point-and if so, why and how? The point is that history, as it has evolved as the mainstream of civilization from Sumeria to the World Trade Center, has always been driven by the organizing principle of artificial time. It is artificial time that separates history from prehistory-and, as we shall see, from post-history as well. It was due to the

The Error in Time

The measure of a two-dimensional plane in space has nothing to do with the experience or the nature of time.

"TIME WAS COMPRESSED INTO A FLAT CIRCLE..."
Dreamspell "Genesis"

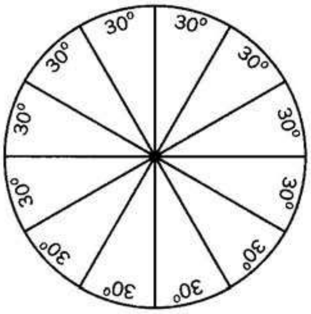

Sumeria: Ca 3000 B.C.
Day divided into 24 hours,
60 minutes, 60 seconds
based on 360 degree circle—
two-dimensional plane in space—
divided into twelve parts

Earth's solar orbit = 365 days
360 degrees ≠ 365 days

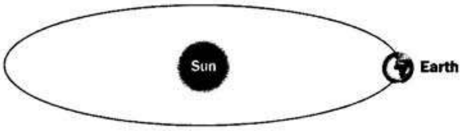

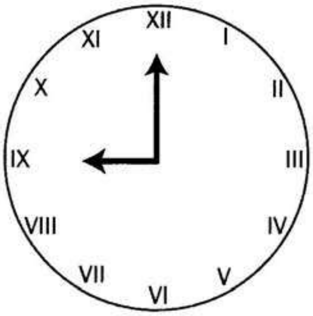

Europe: 800-1600 A.D.
Clock—two-dimensional plane
in space—divided into twelve
parts each 30 degree of arc =
1 hour = 60 minutes
each minute = 60 seconds
A.M. = 12 hours
P.M. = 12 hours
= 24 hours

I - 31	II-28/29	III - 31
IV - 30	V - 31	VI - 30
VII - 31	VIII - 31	IX - 30
X - 31	XI - 30	XII - 31

Rome (Vatican): 44 B.C. - 1582 A.D.
Twelve-month Julian/Gregorian
calendar derived from twelve part
division of circle; irregular measure
due to fact that 365 is not divisible
by twelve.

genesis of artificial time at the beginning of history and its emplacement within the planetary noosphere that also generated the thirteen baktun Wave Harmonic of history. For the very purpose of the thirteen baktun Wave Harmonic of history is to encompass and coordinate the cycle of artificial time. Why?

The Wave Harmonic is a function of the 13:20 synchronization frequency. It is a function of cosmic time. In the higher knowledge of the Mayan understanding, it provided the perfect measure of the cycle of artificial time. From the perspective of the actual noospheric timing, which operates according to the natural 13:20 cosmic timing frequency, history is none other than a manifestation of artificial time. Civilization becomes organized by the principle of artificial time. But because it is artificial, it cannot endure. If the mathematical basis of mechanized time, and hence of mechanization, was already established at the beginning of history, then the end of history is the end of artificial, mechanized 12:60 time. And in the end, the Law of

Time codes and encompasses the whole of the cycle of artificial time through the thirteen baktun Wave Harmonic of history. What is important to grasp is that the entire cycle of artificial time was *already perceived* by the Mayans as a phenomenon that could not endure more than the length of the cycle of thirteen baktuns.

It is necessary for the reader to understand how this perspective was discovered. It was the result of a phenomenological experiment of living according to the synchronic cycles of the Mayan time science. After the death of my son in 1987, my wife and I began in earnest to follow the cycles encoded in the Mayan calendar: 13-day, 20-day, 52-day and 260-day cycles. Our life changed dramatically. At the same time, I endeavored to define the mathematical basis of these cycles. Once I resigned my academic position in early 1989, we were swept by a wave of time that took us from the Rocky Mountains to Southern California to the Hawaiian Islands in the middle of the Pacific Ocean. There, in relative isolation from the rest of the world, we plunged into our experiment. Before we had truly settled in, the time wave took us to Switzerland, where on December 10, 1989, in the Museum of Time in Geneva, we were able to put our phenomenological experiment to the test-although we did not consciously realize at the time that we were doing so.

There in the Museum of Time, essentially surrounded by examples of the history of the clock and mechanization, and because our perceptions had been altered by our living so completely by the various timing cycles encoded in the Mayan calendar system, we were able to spontaneously perceive the existence of two timing frequencies. The natural frequency-the one we had been living-we understood instinctively to be the 13:20 timing frequency. The other frequency-the one enshrined in this Museum of Time-we understood to be the artificial 12:60 timing frequency. This realization of artificial time was rooted in the evident proof that the clock is based on the twelve-part division of a flat circle, a spatial plane that clearly has nothing to do with the dynamics of time. It is a manifest error, but one that humanity has taken for granted and endowed with all of the characteristics of "truth."

We also immediately saw that the error of mistaking the division of a two-dimensional plane in space for a measure of time was transmitted not only through the clock, but through the twelve-month Gregorian calendar as well. With brief reflection upon what a gross error this was-substituting a twelve-part division of space as a measure of time, and then elevating this mistake as the truth of the measure of time-we understood how it could have actually led humanity away from the natural time of the universe. In making this observation, we immediately and intuitively grasped the necessary solution: replace the erroneously measured twelve-month Gregorian calendar with the perfect measure of the Thirteen Moon/28-Day calendar.

Our research into the mathematical codes behind the Mayan calendar now progressed rapidly, but our attention also turned to the question: Whence came the thought form of 24 hours, 60 minutes, and 60 seconds as a measure of a circle of 360 degrees? It was clear to us that this division of time based on the division of a circle in space occurred at what is the beginning of history, a fact later confirmed for us in the already cited *History of the World* (1997), which states that this division of 24 hours, 60 minutes, 60 seconds and the circle into 360 degrees was devised "Ca. 3000 B.C."

Once we had finished working out the mathematical codes, we turned to a consideration of the actual implications of the error in time. It was now also clear why the Mayan time science was so superior. It was based on the reality of time as a factor of synchronization. By contrast, the concept of time as it had developed in the mainstream of civilization was limited, linear, and totally third-dimensional, a fact that confirmed Vernadsky's analysis, as well, that our notions of time are a function of the metrics of space. By 1991, the experience in the Museum of Time had driven us into

What Difference Can a Calendar Make?

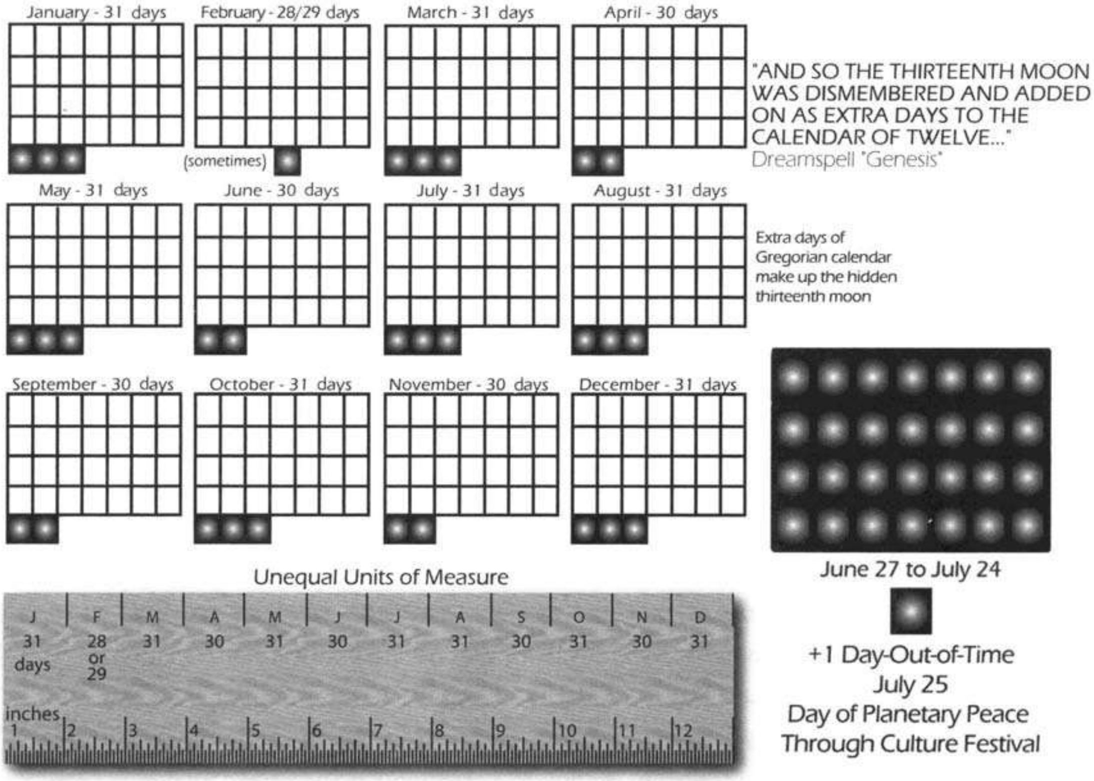
January - 31 days
February - 28/29 days
March - 31 days
April - 30 days
(sometimes)
"AND SO THE THIRTEENTH MOON WAS DISMEMBERED AND ADDED ON AS EXTRA DAYS TO THE CALENDAR OF TWELVE..."
Dreamspell "Genesis"
May - 31 days
June - 30 days
July - 31 days
August - 31 days
Extra days of Gregorian calendar make up the hidden thirteenth moon
September - 30 days
October - 31 days
November - 30 days
December - 31 days
June 27 to July 24
Unequal Units of Measure
J F M A M J J A S O N D
31 28 or 29 31 30 31 30 31 31 30 31 30 31
days
inches
1 2 3 4 5 6 7 8 9 10 11 12
+1 Day-Out-of-Time
July 25
Day of Planetary Peace
Through Culture Festival

Gregorian Calendar and the Hidden Thirteenth Moon

creating a full-blown psychomythic analysis and description of this moment of realization concerning the nature of time. The complete structure of the mathematical codes of time, as well as the psychomythic description of the causes and effects of the error in time, is known as the Dreamspell, which is defined as follows:

"DreamspeH: function of fifth-dimensional galactic-solar planetary regulation; 26,000-year renewable cycle and planetary castle genesis; any consensus reality; in disregard of interdimensional reality creates entropic spell of history; cure for loss of galactic memory."2 The *planetary castle genesis* refers to the *five castles of time* (each 5,200 years in length), divided into three *geneses-the* Dragon, the Monkey, and the Moon-that constitute the entire fractal panorama of the 26,000-year cycle of history that ends in 2012, and is meant to be renewed in 2013 (an obvious synchronic reference to the 13:20 timing frequency). In the Dreamspell analysis, the beginning of history occurs almost 5,200 years ago between the Monkey and the Moon geneses, and is marked by the imposition of the false 12:60 timing ratio.

In light of this analysis and definition it is worth quoting certain passages from the Dreamspell text that describe this precise moment at the beginning of history and that the Law of Time defines as the root error of civilization. First we quote from the "Dreamspell Genesis," and then from the descriptions of the "Dreamspell Journey and Mission of Timeship Earth 2013." The purpose of quoting these passages here is to establish a psychomythic context to explain the profound nature of what we refer to as the error in time compounded over 5,000 years of history:

> Imposters, male priests and warriors usurped the power of the thirteen moons. They hid the thirteen among themselves and attempted to banish all memory of the matrix dreamspell of magic. . . The planetary kin had imposed upon them the diminishing power ratio of 12:60. No longer the magic of thirteen moons but a twelve month calendar of uncertain meaning. No longer the timeless gyre of magical flight, but a sixty-minute hour to earn one's bread.
>
> Time was compressed into a flat circle. . .
>
> The 13:20 ratio is the base operating ratio of timeship Earth. . .
>
> In that vital moment during the magic flight. . . Toward the moon genesis, another spell was cast. Instead of the 13:260 ratio already stored deep within the crystal core of timeship Earth, the planetary kin received the 12:60 ratio, the dark dreamspell of history. . . And so the thirteenth moon was dismembered and added on as extra days to the calendar of twelve. . .

> The absorption of the thirteen by the twelve was called an improvement by the male imposter priests hypnotized by the combined powers of Jupiter and Saturn. It takes twelve years for Jupiter to orbit once around the sun. It takes sixty years for Jupiter and Saturn to be conjunct.
>
> Twelve times five equals sixty, one sixth of the flat three hundred sixty degree circle. The difference between twelve and five is seven. While the priests of the calendar of ancient Babylon banished the power of the thirteen moons, they replaced that power with the power of seven.
>
> While the twelve diminished the power of thirteen by one, the sixty raised the power of twenty by three. A seeming increase in power occurred, but an increase only on the flat plane of the third-dimensional time to which the planetary kin were now consigned.
>
> With the banishment of the thirteen moons, the fourth-dimensional time magic. . . became the property of the priest class. Setting up religions and governments to maintain the kin within the third dimension, the priests were assured that anyone having fourth-dimensional experiences either be appropriated or destroyed.J

So much for the psychomythic impression of the genesis of artificial time, the flat time of an exclusive third-dimensional reality called history. A few further quotations from the Dreamspell emphasize both the interplanetary and mathematical root of false time:

> At -3187 Dreamspell years the time bandits on the sixth and seventh orbits pulsed their beam. The 12:60 frequency of the memory virus took immediate effect on the third-dimensional space suits. . . The Dreamspell of history was cast.
>
> Imposter priests on Earth substitute the 12:60 ratio. . . Instead of the magic of following thirteen moons, a non-circulating twelve-month calendar is substituted; instead of the beauty and power of magical flight, the sixty-minute hour. The result is disastrous.
>
> The twelve-month calendar is the consequence of the 12:60 ratio imposed at -3187 Dreamspell years (3113 B.C.,Julian/Gregorian). With no basis in or capacity for measuring galactic time, the twelve-month calendar is actually a third-dimensional prison keeping the four root races separate and at war with each other.4

These descriptive explanations of the error in time define such a profoundly radical point of view that we must flesh out the implications and description of the course of artificial time according to the analysis of the Law of Time. From the outset of civilization in Sumeria, artificial time captures the human mentality in the order of low time and near space (third dimension), separating it from high time and far space (fourth dimension). The registration of artificial time within the unconscious structure of the noosphere, which is intrinsically governed by the universal 13:20 frequency of synchronization, creates a conflict and a pressure in the noosphere, a time warp or taint that affects the entire biosphere.

Civilization-living in cities-increasingly becomes the human norm within the biosphere. Where there had been no civilization, at a few key places around the planet civilization now arises-a process that goes from the crystalline structures at the origin to the later phases of imperialism and empire building. This cycle is repeated again and again. Even in the New World where the 13:20 natural timing frequency prevailed, the later stages of civilization succumbed to the effects of the 12:60 already seeded in the noosphere. Ultimately, all humanity and the biosphere were to be governed by the artificial timing frequency that ends with the construction of the technosphere, the final and supreme expression of artificial time. Indeed, the technosphere is the absolute conclusion of what can rightly be called an artificial time warp in the biosphere.

Concerning the mathematical basis of the 12:60 frequency, V. A. Ponko of the Russian Academy of Sciences, Novosibirsk, has conducted extensive research and wave analysis on the mathematical properties of the cycles of history. He has concluded that as a mental coordinate, the number 12 is a strict construct of the angles of space. Since it does not correspond to the sinusoidal curvature of the cycles of time and the plotting of all organic phenomena, when used as a time factor the number 12 offers no "protection" from bombardment by any number of cosmic forces. The mathematical plotting of the 13:28 frequency that characterizes the Thirteen Moon calendar, on the other hand, is in alignment with the sinusoidal curve properties of organic phenomena, and therefore functions as an "umbrella" maintaining the organism in resonance with the biosphere.s

Here let us make a more precise definition of timing frequencies and the nature of calendars as the macro-organizing programming systems of the human in the biosphere. By defining time as the universal frequency of synchronization, we are saying that synchronization is the fundamental program to which everything adheres and which makes everything perceptibly coherent, hence synchronic order. Therefore, the natural timing frequency that is a universal phenomenon both regulates the phenomenal order of the cosmos, and in the human is absorbed as a mental frequency that unconsciously synchronizes the mind and senses with the natural order. The result is the normal sensation and experience of the harmony of reality. This defines both a fundamentally prehistoric perception of reality as well as the spiritual experience that is common when we are confronted with the awesomeness of nature-usually far away, it should be noted, from city life. This universal factor of synchronization we have identified as the mathematically precise ratio 13:20, a ratio constant embedded in the structure and order of the calendrical and astronomical system of the ancient Maya.

If we understand that this frequency of synchronization is a genuinely universal constant, then we may comprehend how not only the entire biosphere, but solar system and galaxy are also governed by the same frequency. This is how the Kozyrev experiments could show that there is instantaneous transmission of information throughout the universe, which in its entirety is a function of the formulation *velocity of time is instantaneously infinite.* Synchronization and infinite instantaneity of transmission are mutually defining-everything is always instantaneously synchronized. We can further say that synchronization "syntropizes" while dissynchronization "entropizes." There is no lag time in synchronization.

This being the case, we can then begin to understand how the acceptance and institutionalization of the artificial 24-hour, 60-minute, 60-second division of the 360 degree circle as the basis for defining time could be such a singular and ultimately disastrous deviation from the universal frequency of synchronization. Thus when we speak of the artificial 12:60 timing frequency, we are referring to the mental effect of the feedback program of the use of artificial and erroneous instruments of measure. Rather than allowing the mind to remain in the natural harmony of synchronization with nature, the artificial 12:60 standard and its instruments establish a feedback effect in the mind that accustoms it to the deviation from nature. This habituation results in the creation of an unconscious timing frequency that is accepted by the consensus reality as the actual norm. Separation of human order from the biosphere becomes accepted and acceptable, making civilization not only possible, but setting it up, as it were, as a parallel rival to "nature." In fact, it is the unconscious mental frequency of 12:60 artificial time that coordinates and maintains civilization as a belief system.

When we speak of the instruments of measure of artificial time, we are specifically referring to the Gregorian calendar as the macro-organizing principle and the watch or mechanical clock as the micro-organizing principle. The macro-organizing principle organizes our life into days, weeks, months, and years. The micro-organizing principle organizes our life into seconds, minutes, and hours. These two instruments have their common root in the error in time that occurred 5,000 years ago at the beginning of history. Try to think of life without these instruments and you will quickly realize how second nature they are-and yet there was a time when such instruments and their terminology did not exist. Before turning our attention to an analysis of these two instruments and their effects on human consciousness, we must view the entire cycle of the Wave Harmonic of history as a function of the macro-organizing principles of time that the human in the biosphere adopted for its uses and that, in turn, condition at a most profound level the worldview of the culture or civilization that uses it.

HISTORICAL CALENDARS AS INSTRUMENTS OF ASTRONOMICAL MEASURE

It may be rightly asked, What does a calendar measure? At a minimum, a calendar is or should be the measure of the Earth in its rotation around its local star, the sun. This rotation takes 365 days. The Earth's satellite, the moon, is the obvious synchronizing factor between the Earth and the sun. But here we encounter a difficulty. The moon can be measured by various cycles. The synodical cycle, from new moon to new moon, is 29.5 days. The sidereal cycle, the measure taken from where the moon appears at the same place in the sky, is just over 27 days. There is an apsidal cycle, however, taken from the measure of when the moon's axis is tilted farthest from the Earth, which is 28 days. This variance of measures is due to the fact that until the rise of rocketry, we could not really see the moon from space. In actuality the moon rotates around the Earth thirteen times during the time it takes the Earth to orbit the sun once.

If one were objectively seeking the proper measure of the Earth's orbit using the moon as the measuring device, we would therefore construct an instrument for time reckoning that consists of an even harmonic measure of thirteen cycles of twenty-eight days each. The 28-day cycle conforms, of course, to the female menstruation cycle, an obvious human biological cycle, and it also divides perfectly into four 7-day subcycles or weeks. Also, 13 x 28 = 364, the same number of days as thirteen moons multiplied by four 7-day weeks (a total of fifty-two weeks), with the 365th day being a free day, or Day Out of Time, no day of the week or month at all. We shall deal with the perceived extra quarter-day later in our discussion.

The point is this: If time is a factor of synchronization, then a calendar-as an instrument for measuring the count of time-should also be constructed so that it maximizes the factor of synchronization. This can only be done if the instrument of time reckoning is constructed on harmonic principles in conformity with the objective data of nature. Such an instrument would also conform to the orderly processes of the biosphere taken as a whole, and be a great assistance in bringing into consciousness the cosmic programs that organize the biosphere and noosphere. The Thirteen Moon/28-Day calendar is the perfect instrument for such a simple registration of harmony and synchronization. It should also be noted that in its mathematics the Thirteen Moon/28-Day calendar is a function of the 13:20 timing frequency. This in itself establishes the Thirteen Moon/28-Day calendar as a perfect instrument of the Law of Time and as a function of the synchronic order of reality.

Since the process of civilization represents a break from a long cycle generally undistinguished by the kind of perturbations that characterize history, the analysis of the Law of Time assumes that in prehistory humans lived in relative harmony, and therefore that the humans tended to use harmonic instruments of time reckoning. The humans living in greater harmony with the biospheric cycles would be disposed to be in greater unconscious or aboriginal attunement with the noosphere. Since the noosphere is regulated by the synchronic order of the universal 13:20 timing frequency, the humans would be more naturally disposed to devise, adopt, or intuitively utilize the time measure in greatest accord with this instinctual 13:20 noospheric timing ratio.

This instrument of time measure would, of course, be the Thirteen Moon/28-Day calendar. It should also be noted that this calendar is a perfect solar-lunar measure; that is, it uses the even, regular lunar cycle of 28 days as the standard of measuring the 365-day solar cycle of the Earth. The balance of solar and lunar may also be taken as

Moon Circling Earth Thirteen Times per Year

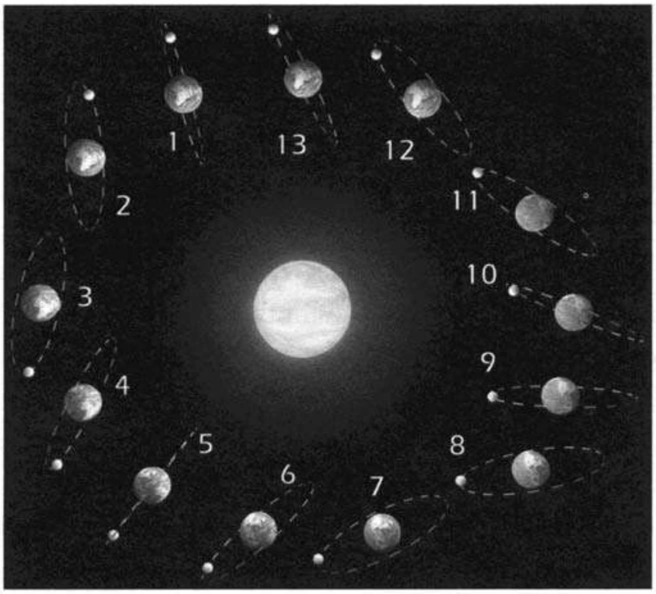

Did you know the moon goes around the Earth thirteen times a year?

SEE: HTTP://WWW.TORTUGA.COM/WWW/ANIMATIONS/ANIM_MOON.HTM symbolic of human psychological characteristics and their balance or imbalance. A solar-lunar calendar will represent and reflect a balance of the solar-lunar, masculine and feminine qualities within the human being. It is important to bear in mind these considerations of calendars and their effect on shaping the human historical consciousness and its psychological makeup. From this perspective, history is, in fact, a process of falling away from the perfection of harmonic standards.

Evidence exists that the Thirteen Moon/28-Day measure was widely known in prehistory. We find evidence of its use as a prehistoric synchronic measure in the remote past of China, in Polynesia, and scattered across late neolithic Europe and the Middle East. Among the Maya it was known as the *Tun Uc,* literally "moon count" or count of seven, while the living tradition of the Thirteen Moon/28-Day calendar is still continued in South America and in the British Isles, where it is known as the Druid Tree calendar. While the Druid calendar has no year count attached to it that would definitively testify to its prehistoric ancestry, the South American calendar does. Known as the *Pachacuti,* the South American Thirteen Moon/28-Day calendar is currently in the year 5509, which places its origins at 3308 B.C., toward the middle of the fourth millennium before Christ, and several hundred years before the beginning of the thirteen baktun Wave Harmonic of history.

If we understand that calendars are genuine time-measuring devices meant to synchronize us with the cosmic order within the biosphere, and that calendars are thereby programming devices, we may then say that a calendar of perfect harmony can have no history. In other words, by its harmony such a calendar is always in tune with the cosmos, which is beyond history. History can only be a function of a disharmonic programming, not in sync or even at odds with the laws of natural time.

From the point of view of timing devices, there are two factors in the establishment of history. The first is the 12:60 program established 5,000 years ago with the division of the day into 24 hours, 60 minutes, 60 seconds, based on the 360-degree division of the circle. The other factor is the use of a purely lunar calendar that is difficult to reconcile with the actual measure of the solar year. The purpose of the 12:60 measure was to establish a timing standard that was actually a pseudosolar calendar to assert male intellectual dominance. The Babylonian (and later, Egyptian) calendar consisted of twelve 30-day cycles = 360 days = 360 degrees, plus an extra 5-day cycle. The 30-day measure is not actually a natural one, being a half day more than the synodic lunar measure, but one that conforms to the hexagisimal (6-based) mathematics of the circle. For this reason, this instrument of time reckoning is defined as *pseudosolar*. Its ultimate descendent, as we shall see, is the twelve-month Julian/Gregorian system.

By contrast, the Maya also had a 365-day solar calendar, the *Haab.* Its measure, however, is that of eighteen 20-day cycles (18 x 20 = 360) plus the 5-day cycle *(uayeb),* the 20-day cycle *(vinal)* being a mathematical function of the 13:20 frequency. In addition, the Maya had the measure of the *tun,* the 360-day count of which is precisely the same as that of the degrees of the circle. It was the tun that the Maya used to measure the thirteen baktun count, which actually consists of5,200 tun (5,125 solar years). The purpose of this measure was to provide an exact harmonic standard for the measure of history that, we may recall, is the duration of artificial time. It is very relevant that the harmonic standard of the thirteen baktun Wave Harmonic of history is the 360-day tun, the actual temporal equivalent of the degrees of a circle in space. The precise measurement of the cycle of artificial time, which turned out to be anything but harmonic, could yet be measured with a harmonic unit that reflects both the circle *and* the mathematical perfection of the Law of Time. How better to coordinate the entire cycle of artificial time than by the tun, a unit of measure that reflects the 360-degree circle, the basis of the original error in artificial time? As we will later observe, the fact that the Law of Time coordinates even the cycles of artificial time is also evident in the analysis of the timing of the fatal flaw of the Twin Towers apocalypse.

While the Babylonians began to apply a hexagisimal time-reckoning device based on spatial metrics of the circle, another factor became predominant in the Old World: the rise of the synodical lunar calendar as the exclusive measure of time. The synodicallunar calendar is purely lunar in that it is not evidently or perfectly harmonized with the 365-day measure of the solar cycle-a fact of great consequence for the development of mainstream civilization. The Babylonian calendar, with its abrogation of the thirteen perfect 28-day months, established the pseudosolar twelve-month measure. This twelve-month measure conformed to the lunar measure of twelve synodic lunations per lunar year. However, twelve lunations are only 354 days, eleven days short of the solar year. This required an intricate set of calculations to make the lunar calendar keep up with the measure of the Earth's solar orbit. And in reality, although we tend to think of the moon as feminine, the knowledge of the synodic measure, the cycle from new moon to new moon, became the exclusive province of a male priest class that used the lunar calendars to capture and control the feminine principle, and to oppress women in general. Whereas the Babylonian and Egyptian solar calendars fell into disuse over time, the lunar calendar persisted.

Since there is only one moon orbiting Earth, it should be kept in mind that all lunar calendars are essentially the same. One and the same moon in all its phases is apparent to all humans on Earth. So when we speak of different lunar calendars-most notably the Hebrew, the Islamic, and the Chinese-the measure is basically the same, from new moon to new moon, but New Year's Day falls differently in each of them, thus accounting in part for the various calendrical traditions. Within each of these lunar calendars, a different set of holidays is programmed. It is the programming of the different holidays and festivals of the calendars, along with the dating of their new year, that makes each lunar calendar tradition distinct and that also demonstrates how a calendar is a programming device. For instance, if one were to take out the different holidays programmed into the Hebrew lunar calendar, there really would be no such thing as Judaism. This demonstrates that the cultural conditioning of a people is totally dependent on the programming of the calendar they use. This also shows how a calendar is a feedback mechanism that maintains a group of people within the confines of its own self-established belief system.

Of the three lunar calendars we have mentioned, the oldest count of years is accorded to the Hebrew calendar. This year Rosh Hashanah, or Jewish New Year, occurred on the new moon, September 17, 2001, and marked the beginning of year 5762. This places the origin date for the Hebrew calendar in the year 3761 B.C., some 453 years prior to the commencement of the Pachacuti count. The Hebrew calendar (like the Chinese calendar, which we will discuss below) makes up for the slippage of eleven days per solar year by intercalating a thirteenth lunation cycle approximately every three years, or exactly seven such thirteenth moons every nineteen years. These numbers too-13, 7, and 19-are key harmonic factors of the synchronic order of time. This demonstrates that, at the very least, the synodical lunation calendars are calibrated in their larger cycles by the Law of Time. But the effect of being a purely lunar measure, at variance with the actual 365-day measure of the solar year, creates an interesting and one-sided approach. There is no question that the lunar calendar civilizations are the tool of powerfully patriarchal societies, the tables of the moons and the years being the possession of a dominant male priest class.

The Chinese lunar calendar differs from the Hebrew in one important distinction: from the most ancient times it has been embedded in a system known as the *five elements*, and is coordinated with a very elaborate mathematical system that bears some resemblance to the 60-60, 24-360 basis of the SumerianiBabylonian 12:60 frequency. This synchronizing system accounts for the Chinese cycles of years that are counted by the twelve zodiacal animals in combination with the five elements: earth, fire, wood, water, and metal, thus giving rise to 60 (12 x 5) year cycles. Three of these sixty-year cycles create a "group" or cycle of 180 years. We are now (in 2001) in the eighteenth year of the twenty-seventh group, or year 4,698 of the Chinese lunar calendar. So embedded is Chinese culture and civilization in this calendar, with its elaborate system, that to take away this calendar would virtually eliminate

Chinese culture altogether. Perhaps more than any other factor, this highly involved calendar also accounts for the longevity and tenacity of Chinese civilization as one continuous tradition for the duration of its history, which according to the reckoning of its calendar began in 2697 B.C., a date, interestingly enough, that coincides with the dedication of the Great Pyramid of Egypt.

Despite its marvelous system, the Chinese lunar calendar is still just that: a lunar calendar. The problem with this, from the point of view of the Law of Time, is that when it is not regulated by a proper solar count or measure, the lunar calendar leads people into one-sided developments of one kind or another, if only to the extent to which these calendars foster such profoundly patriarchal societies, of which the Chinese is certainly no exception. The influence and use of the Chinese lunar calendar system extends into southeast Asia, Tibet, Mongolia, Korea, and Japan. As a field of thought within the noosphere, the Chinese system is the equal in power base to the Babylonian and Indian systems, at least up to the time of the dominance of the Gregorian. Even today, the Chinese New Year exerts a powerful attraction worldwide.

While the Chinese and Hebrew lunar calendar systems maintain a powerful conservative strain of human society, the intercalation of the thirteenth moon, seven times every nineteen years, creates a kind of circulation within the system. Thirteen is the number of circulation, whereas twelve is that of a static, non-circulating spatial order. The taboo on the number thirteen-epitomized by the superstition about Friday the thirteenth-must certainly be related to the suppression of the solar-lunar Thirteen Moon/28-Day calendars in virtually all historical societies. This irrational repulsion of the number 13 serves psychologically in defense of the "rationality" of the number 12, a number that does not circulate time and is at the root of the linear conception of time. The illogical and irrational decisions that are made and then institutionalized into the human social fabric must at one point be discounted. Any illusion or error pursued for too long will always end in disaster.

The Islamic lunar calendar, used by some 1.3 billion humans, is distinguished by two factors:

1) It is based on a known historical incident, the *Hegira* or flight of the prophet Muhammad from Mecca to Medina, dated to the new moon or Muharram 1, Julian, July 16, 622 (Gregorian, July 26).

2) It is a pure twelve-month lunar calendar running on a cycle of pure lunar years of 354 days, with no intercalary lunations, and in this regard does not annually harmonize with the 365-day solar cycle.

For this reason, this year 2001 saw the beginning oflslamic lunar year A.H. *(M*-ter Hegira) 1422. But if we are to count the solar years since the Hegira, correlated to Gregorian July 26, then it is solar year A.H. 1379. We see here a discrepancy of some some forty-three years between the solar count and the lunar count. The lunar count continuously retrocedes so that it takes thirty-two years for one Muharram to occur at the same place in the cycle of the solar year. In essence, the Islamic lunar calendar creates a hermetic, retrocessing time bubble with no circulation of the power of the thirteenth moon, and that accounts for the highly conservative and patriarchal nature of much oflslamic society. Again, while there may be a virtue to maintaining a strict lunar calendar count, if it is not correlated to a harmonic solar or solar-lunar standard, the effects will be precisely of the nature that we see occurring in today's world, a point to which we shall later return.

Finally, in this brief consideration of calendars as the principle macro-organizing system for the programming of human society, a word should be said about the system of calendars in the predominant part of the Indian subcontinent. Long before they adopted the Babylonian solar zodiac and hexagisimal spatial principles for reckoning time, the Hindus also had a type of 28-day count, but one based on the course of the moon in relation to a set of stars called Nakhsastras; this created a kind of 28-part lunar zodiac. In this system, however, the Hindus divided the 360-degree lunar ecliptic into twenty-seven equal parts (27 days = sidereal cycle of the moon) each part equal to 13 degrees, 20 minutes of are, the precise measure of the 13:20 timing frequency. A complex variety of lunar calendars were developed early on, but were gradually displaced by what we call the pseudosolar Babylonian division of the Earth's ecliptic into twelve 30-degree segments, a fact that reflects the adoption to the unconscious 12:60 frequency program.

The Hindu (pseudo) solar calendars are all based on the same division into 30-degree arcs-a circle or plane in space! While the base count of the present era is taken from February 18,3102, the beginning of the Kali Yuga-the last and darkest age-exactly 5,104 years ago (dated from Gregorian year 2002), today there are at least six different calendars in use throughout the subcontintent, each essentially the same as the other, but characterized by different dynastic starting points. These differing counts, coupled with the distinct profusion of religious holidays, accounts for the jarringly conservative and colorful confusion that characterizes present-day Indian society-a neo-Babylonian time warp, its populace locked into place by complex systems of pseudo-solar calendars set in the context of earlier Hindu astronomical cycles of a vast and overwhelmingly cosmic nature.

"TIME IS MONEY": THE WATCHWORD OF THE 12:60 FREQUENCY

All of the aforementioned time-reckoning systems, with their various beginning points, can be mapped within the 5,200 tun/5,125 solar cycle of the thirteen baktun Wave Harmonic of history. While we can generally see how the Babylonian and lunar calendar systems shaped the mainstream civilizational orders of the Old World, in our time map we must also take into account the workings of the 12:60 artificial timing frequency within the unconscious noosphere. In this sense, the 12:60 frequency refers to the principle programs of control adopted and maintained through use of the twelve-month calendar systems in the Old World. (See plate 5, Noospheric Time Map-Wave Harmonic of History.)

Plotting the incidence of these calendrical streams on the time map, we must keep in mind that the calendar systems are the chief instruments for establishing and maintaining the human mind and social order within very specific programs of behavior. These programs, at least in the Old World, were characterized by the introduction of artificial systems of exchange called money, and accompanying programs of taxation. These systems were adopted to a greater or lesser degree by all the Old World societies. In this we find the root of the slogan, "time is money," where money is the artificial medium to negotiate artificial time. The end of history is the end of money and the end of artificial time. The word "calendar" itself is derived from a Latin word meaning "account book," the first day of every month being *calends* or the date of payment of debts.

Now, let us return to a consideration of the two principal and resultant 12:60 instruments dominating the end of historical time: the Gregorian calendar and the mechanical clock. We may note that toward the conclusion of the twelfth baktun cycle (A.D. 1618), the clock reaches its perfection at the very moment of the implementation of the Gregorian calendar reform (A.D. 1582), just after the act of global circumnavigation had been accomplished. Within the entire thirteenth baktun cycle, 1618-2012, the coincidence of these three factors-clock, calendar, and European global circumnavigation-induced what amounts to a 12:60 frequency capture of planetary time in the biosphere. What are the implications of this takeover of time by an irregular and irrationally measured Gregorian calendar and its accompanying micro-organizing device, the mechanical clock?

Plotting the growth curves of human population, machine, and money, we see that the complete infusion of the 12:60 artificial timing frequency creates an unprecedented acceleration, the exponential peak of which occurs at the moment of the Inevitable Event. For this reason, according to the Law of Time, it is important to comprehend even more deeply the nature of calendars and their effect, especially with regard to the current global standard.

Because the calendar is a macro-organizing principle, when accepted over time it establishes in the mind, individually and collectively, a set of perceptions that are automatically taken for granted as being "real" and "indisputable." These sets of perceptions define the paradigm by which a people, culture, or even an entire civilization operates. The Gregorian calendar, the current global civil standard, is the paradigmatic macro-organizing principle in which are embedded all the laws, customs, institutions, and scientific principles governing the present global civilization-not to mention all the holidays of the Vatican-ordained Catholic Church. According to the Law of Time, the current calendar is an irregular standard of measure; its units of measure do not correspond. This calendar represents, therefore, the institutionalization of disorder and entropy. Because it has existed for a sufficiently long duration in the human historical cycle-2,000 years, including its predecessor, the Julian calendar-the perceptions fostered by the unconscious acceptance of the Gregorian calendar are taken as the unshakable bedrock of nature and reality. All current beliefs-economic, political, and scientific, from democratic neo-liberalism to the special theory of relativity-are actually products of the underlying perceptions promoted by this calendar, and have no reality apart from the beliefs about time that the calendar engenders. *Change the macro-organizing principle and you change the paradigm.*

THE NEED FOR A TRULY "NEW MILLENNIUM"

As the global civil standard, the artificial, irregularly measured Gregorian calendar is a self-reinforcing feedback loop. As such, it furthers and maintains all linear time concepts, thereby establishing a host of entropic, disordered value concepts such as the violent universe, the degradation of matter, quantum physics, the arrow of linear time, the doctrine of techno-economic inexorability, and an attendant host of unresolvable problems-crime, drug abuse, terrorism, environmental deterioration, and so forth. Why? Because as an irregular standard of measure, the Gregorian calendar is incapable of producing harmony. Only harmony can unify. The Gregorian calendar is not a unifying harmonic standard. Lack of a unifying global standard exacerbates all current conflicts. Condition the mind to an irregular standard and the mind will adjust to disorder and chaos as normal conditions of existence.

The world's racial, tribal, historical, and religious conflicts are embedded in and a function of different timing systems (calendars), all now coordinated within a master irregular timing standard and macro-organizing principle, the Gregorian calendar. In this calendar's irregularity is embedded the view that the world of the inevitable degradation of matter is the exclusive object of all present-day science, and hence is also reflected in the entropic nature of global civilization and its deteriorating social processes. The Gregorian calendar conditions the mind to hopelessness of resolution. Without a unifying harmonic standard humanity is incapable of finding long-term resolutions for any of its problems. Disharmony can never produce harmony. Only harmony can produce harmony. All conflict can only be resolved by the application and within the context of a harmonic timing standard.

Now, following the attack on the Twin Towers and the Pentagon, we can say that the Gregorian calendar is bringing history to an end. The Inevitable Event was programmed into the Gregorian calendar, and the Gregorian calendar usurped the power of all calendars in the thirteenth and final baktun cycle of history. When the third millennium officially opened at the dawn of the year 2001, the pride of industrial man was at an all-time high. Would it be a millennium of peace or one of war? If the last century of the second millennium was the century of total war, was the third millennium going to do anything but inherit the unfinished programs of the old millennium? Could the Earth and the biosphere withstand a millennium of total war? Hardly. The Vatican "jubilee year," which ended with the beginning of the third millennium, turned sour with the Twin Towers apocalypse, as is vividly reflected in a photo of Pope John Paul II taken just after the Inevitable Event. The Gregorian calendar marks its inception with the birth of Christ, a date and moment that was never historically noted, and so is shrouded in speculation and conjecture. Now the twisted civilization of Gregorian time is locked in mortal combat against a terroristic specter that it alone could have projected into manifestation. Soon, by biospheric standards, the world of artificial time will be over and gone. And then the millennium will be truly new.

The very word *millennium* conjures possibilities of Earth-shattering events and cosmic prospects. If the inception of the Kali Yuga, 3102 B.C.,can be said to be the point where the historical process became irreversible, eleven years after history began in 3113 B.C.,then in counterbalance, the Inevitable Event, 2001, which had to occur eleven years before the end of the cycle of history, A.D.2012, was more than just millennial. It was the termination of more than 5,000 years of the irreversible motion of artificial time. Such symmetry-eleven years after the beginning and eleven years before the end of history-is the hallmark of the synchronic order. For 5,100 years the count of days has numbered the long and increasingly perilous saga of human civilization. How bright was the beginning and how grim will be the end?

There at the beginning of the cycle was the "first city," Uruk. There at the end of the cycle, the fallen Twin Towers of Babel of the "last city," New York (= New Uruk), the ultimate monument to history itself and the final event to initiate the actual closing of the cycle of history.

> *There it is, the city of Uruk. Urshunnabi, climb up on that wall, the outer all shining with the brilliance of burnished copper. The seven wise men laid the foundations. One third of the city is buildings, cunningly executed, one third of the city is garden with rose and bird, and one third of the city is field with the temple of Ishtar within. Goddess of love and struggle.*
>
> \- *THE EPIC OF GILGAMESH*[6]

If Uruk is the place where the seven wise men laid the foundations, those foundations were measured by the 24 hours, 60 minutes, and 60 seconds derived from the 360 degrees of the circle. Fifty-one centuries later, the gardens are all but gone, the fields themselves mechanized and chemicalized by the technosphere. Earth's inhabitants are in terror. Five weeks after the Inevitable Event, now known simply as 9-11, the end of history still smolders. After the struggle, will the goddess of love return?

▼

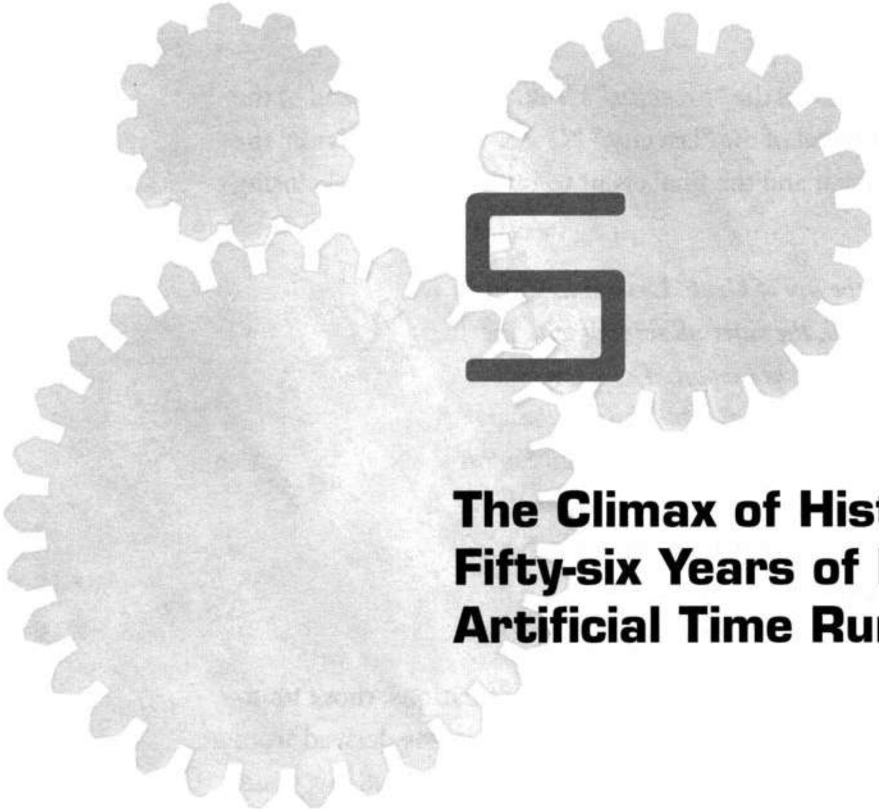

5

The Climax of History, the Fifty-six Years of Hiroshima—Artificial Time Runs Out

THE IMAGE of the Noospheric Time Map-Wave Harmonic of History on plate 5 of the color insert, showing the concurrent evolution of the different calendars of the major civilizations of the world (the illustration is by no means exhaustive), is intended to demonstrate the fact that the human race was not unified by a single coherent timing standard throughout twelve of the baktuns of historical development. It was only in the thirteenth and final baktun that humans became organized by a single timing standard. However, it was not a timing standard in accord with the synchronic order of the Law of Time, but the artificial 12:60 timing frequency as represented by its two organizing instruments, the Gregorian calendar and the mechanical clock.

The seeds of this moment had already been planted in the noosphere at the beginning of history. From both Sumeria and Babylonia the taint of the 12:60 frequency then spread through the noosphere and, by means of historical dissemination and conquest, throughout the Old World. By the time the historic moment arrived to ripen the 12:60 frequency into its full fruition-A.D. 1618, the beginning of the scientific revolution and the thirteenth baktun-there were virtually no longer any cultural traditions governed by the natural 13:20 timing frequency. The civilizations and peoples of the New World were now under the subjugation of the prevailing world order of European global imperialism. The stage was set for the brief but final phase of mechanization, of which the technosphere itself is the climax.

What the Noospheric Time Map also delineates is that the 13:20 frequency was the unconscious organizing factor of prehistory, and that, undoubtedly, a key timing system intuitively evolved among many prehistoric peoples independently of one another. This was the universal, harmonic solar-lunar standard of the Thirteen Moon/28-Day calendar. The defining point of history arrives with the establishment of a 12:60 timing standard within the noospheric unconscious. This frequency predominates in the development of twelve-month pseudosolar calendars, as well as the twelve-month synodicallunation calendars that are not the measure of a solar year. These two commingled systems are entwined throughout the ancient history of the Old World as the central thread of development shaping the conception of time and resulting in what were to become the two dominant instruments of human time reckoning, the Julian/Gregorian calendar and the mechanical clock.

Julius Caesar himself instituted the famous "calendar reform" that created the 446-day year of confusion in 45-44 B.C., and established a calendar that was just as patently confusing. Julius Caesar's was not the only calendar reform in the ancient world. The Essene movement, founded by someone known simply as the Teacher of Righteousness and of which Christ was supposedly a member, began as a revolt against the Hebrew lunar calendar. The issue was the need for fixed holidays within the solar cycle, something that is virtually impossible in a lunar calendar that retrocedes eleven days every solar year. There is strong reason to believe that among other calendars, the Essenes favored the Thirteen Moon/28-Day calendar. This even leads to the question: did Christ himself follow a thirteen moon calendar? In any case, it is the 12:60 frequency transmitted through the irrational disorder of the Julian/Gregorian calendar that captures the mind of the human race and the biosphere during the final years of the cycle of history. And it is precisely for this reason that the pace of human civilization becomes exponentially and entropically accelerated during the last few centuries of its development.

Shocked into higher consciousness by its own barbarism masquerading as civilization, at the end point of history the human has the opportunity to return to the natural 13:20 frequency of synchronization. This return would mark the emergence of humanity from the unconscious mechanistic compulsion of artificial time into the conscious field of post-history. The only possible option to unify humanity-once the yoke of the Gregorian calendar is removed-is the true solar-lunar calendar of Thirteen Moon/28 Days. With the possibility of this positive end in sight, we may turn to a more descriptive and definitive analysis of the technosphere. In this way understanding the modalities of the technosphere as a distinct and entire process, we

Gregorian Calendar 28-Year Cycles: 1945–2001, the Age of Terror

It takes twenty-eight years for the Gregorian calendar to run through its cycle of permutations. Every twenty-eight years, like a phonograph record of time—a chronograph—the Gregorian calendar guides its users through a mentally debilitating construct of unevenly measured and irrationally named months, paced by a system of 7-day weeks that bears little relation to the lengths of the months or years. In addition, every four years there is an extra day. Within this medieval mathematical jumble are kept the collective unconscious programs that repeat every twenty-eight years. What happened in 1945 will somehow repeat or have an effect again in 1973, in 2001, and then in . . . it's up to you. You can change this if you want to.

First day of the year	Tuesday	Wednesday	Thursday	Friday	Sunday	Monday	Wednesday	Friday	Saturday	Saturday	Monday	Thursday	Tuesday	Sunday
Hiroshima 1945 – 72						1945								
	1946													
		1947												
												1948		
										1949				
					1950									
						1951								
													1952	
			1953											
								1954						
									1955					
														1956
	1957													
		1958												
			1959											
				1960										
					1961									
						1962								
	1963													
							1964							
								1965						
									1966					
					1967									
											1968			
		1969												
			1970											
								1971						
										1972				
Twin Towers 1973 – 2000						1973								
	1974													
		1975												
												1976		
										1977				
					1978									
						1979								
													1980	
			1981											
								1982						
									1983					
														1984
	1985													
		1986												
			1987											
				1988										
					1989									
						1990								
	1991													
							1992							
								1993						
									1994					
					1995									
											1996			
		1997												
			1998											
								1999						
										2000				
Inevitable Event 2001 - ?						2001								
	2002													
		2003												
												2004 Year of the Great Calendar Change		

Where will you be? 2012

might then develop a deeper insight into where we have just been and why we need to go somewhere else.

The technosphere defines a 56-year cycle, 1945-2001, coordinated by two complete Gregorian 28-year cycles. For any Gregorian calendar year, the days of the weeks in their irregular monthly succession and in relation to the permutation cycle of leap years repeat precisely every twenty-eight years, during which time there are always exactly seven leap years. This means that Gregorian calendar years 1945, 1973, and 2001 possess the same exact annual arrangement of the days of the week in their monthly succession. If today is Sunday, November 11, 2001, so it was in 1973, and likewise in 1945. In this regard the Law of Time frames the cyclic recurrence of the otherwise irregular 12:60 Gregorian calendar by the intrinsic formulation 28:7. That is, just as the 19:7 factor coordinates the lunar calendars every nineteen years, there being seven intercalations of the thirteenth moon during that cycle, so in the Gregorian calendar every twenty-eight years there are always exactly seven leap days and years.

Understanding a calendar as the instrument that locks the conditioned programs of a given culture or people into place, we can now understand how the unconscious metaprogram of the macro-organizing principle of the Gregorian calendar cumulatively recycles all its millennial programs every twenty-eight years. Since its inception in 1582, the Gregorian calendar has been dragging forward a host of conditioned thought forms and perceptions, including those inherited from 1,500 years of the Julian calendar that it had reformed. At points of dramatic break in the continuity of human consciousness, a new set of cycles is set to recur. Such was the case in the year 1945, with the awesome blast of Hiroshima. The entirety of the mind field conditioned and held in place by the Gregorian calendar then reaches its full fruition during the two subsequent 28-year cycles, fifty-six years in all-thus bringing forward two millennia of conditionings, conflicts, and unconscious death wishes or apocalyptic programs. The first cycle was initiated in the year 1945, the year in which Vernadsky died, the atomic bomb was tested once and used twice, and the Second World War came to an end. That year, 1945, marks the beginning of the technospheric bubble. It is most telling that the Second World War was concluded at the beginning of this 56-year cycle of the technosphere, for it was a war concluded not with peace but with instruments of mass destruction and terror. Thus began the 56-year era of Atomic Terror.

The technosphere has its origins with the full capture of the human mental field by the 12:60 frequency in 1618. From 1618 onward, the noosphere is increasingly obscured by a mental field known as the *technospheric sheath.* Slowly but surely, the technospheric sheath replaces civilization. This is first done by the introduction of mechanistic linear time. It is important to note that the Julian count, the basis of all modern scientific calculations, is a scale created by Thomas Scaliger in 1583, twenty-one years after the 1562 Mayan book burning. This linear time scale, like the Gregorian calendar reform itself, was intended to co-opt the Mayan thirteen baktun Long Count by setting a count of days that begins the first of January, 4713 B.C., or some 1,600 years prior to 13.0.0.0.0, the beginning of the thirteen baktun count in 3113 B.C. This deliberate historical act, the Julian count, along with the Gregorian calendar and the mechanical clock, established the paradigmatic notion of the linearity of time in the ripening field of scientific thought. Reflected in the noosphere, this linear, irreversible time concept levels and stunts the realization of human mass consciousness. In fact, during the technospheric cycle, especially after 1754, the ceiling of human consciousness is maintained by a preoccupation with mechanistic third-dimensional operations while becoming increasingly alienated from the organic order of reality. This creates the 12:60 consciousness constant, a mental ceiling that actually diminishes in proportion to the increasing rates of multiplication, propagation, and intrinsic velocity of the machine.

With the actual rise of the Industrial Revolution, dated in the noosphere to A.D. 1754, the synchronic point at which mechanization becomes an irreversible factor of the biosphere, the technospheric sheath, henceforth enters its next stage, going from a purely mental sheath to the industrial sheath, the prelude to the proto-technosphere itself. The industrial sheath spreads throughout the biosphere between 1754 and 1901, the official beginning of the twentieth century. At this point we enter the forty-four years of the proto-technosphere, 1901-1945. During this critical stage of the proto-technosphere, the actions of human behavior interacting with machine technology make the expression of true culture increasingly difficult, if not impossible. Civilization becomes a set of symbols purveyed through museums, galleries, and theaters, and recorded and reproduced in ever more advanced technological forms. But what of culture, which is not the same as civilization-what becomes of culture?

To begin to answer that question, some further definitions are in order. The technosphere is defined by and based upon one key term: *technology.* According to the common dictionary definition, which already reflects the mass mind, technology is "the totality of the means employed to provide objects necessary for human sustenance and comfort."1 Today virtually all the means employed are themselves mechanical in nature. In common parlance, therefore, technology refers to the complex apparatus of mechanization. In fact, technology *is* mechanization; it is the ability to convert human labor into processes carried out purely by machine-oriented or mechanical means. This is also inclusive of the entirety of computer technology, which represents the mechanization of the more purely mental processes of thought and communication.

Mechanization, we must remember, originated in the clock, in the mechanization of time. It is the mechanization of time that presupposes the tendency toward mechanization as a state of mind within the noosphere. Since artificial time is characterized by the illusion of an inexorable and irreversible linearity, the compulsion toward materialism is also experienced in the same way, an inexorable motion spearheaded by the advance of ever more improved machines. The machines themselves are the products and means of industrialization-the technological transformation of raw goods into consumer goods, a process accounting for much of the free energy introduced into the biogeochemical combustion of the biosphere.

Inseparable from mechanized technology, too, are materialism and the concept of the World Market and, more recently, of globalization. Materialism is implicit in the dictionary definition of technology as the means employed not only for human sustenance but also for human *comfort.* This comfort can only be of a material form or nature, hence the pursuit of comfort through mechanized technology can only be for the furtherance of a philosophy and exaggerated lifestyle of materialism-the belief that only the material things of this world have any value. Of course, the value of material goods and comforts increases when the value of money is put into the technospheric equation. The first stock market opened early in the eighteenth century, providing a place where money could be used as speculative capital to promote the advance of commodities and machine products. The entire sphere involving commodity production and its conversion into consumer products eventually came to be known as the World Market. Integral to the success of the World Market was the creation of the modern banking system, based on the institutionalization of interest rates and the principle of lending money to companies for the purpose of furthering the transformation of the biosphere into the technosphere.

While the concept of the World Market prevailed for a long time as the definition of the system for exploiting and extracting natural resources from the biosphere, and then converting them into industrial goods to be consumed by the human populace worldwide, globalization is far more recent. Globalization represents the absolute triumph of capitalism as the dominant economic doctrine of the human species in the biosphere, and the absolute basis of the technosphere in its final phase of development. As such, globalization is the economic system of neoliberal market economics that treats the whole world as the proper sphere of its policies and practices. It is a specialized form of monetary imperialism that gives rise to and is supported by the doctrine of monetary politics, the use of money as a means of coercion or even for the buying out of whole governments. Of course, the principle practitioner of monetary politics is the nation with the most money, the United States of America, home of the World Trade Center.

Fostering corporate multinationalism, globalization is inseparable from the consumer philosophy it promotes. Consumerism is the devaluation of the human into a link in the chain of the cycle of industrial production, the object of globalization being to inspire and promote a massive and wide-scale consumerism worldwide. This is what is meant by such phrases as "increasing consumer purchasing power" and "opening new markets." To promote and defend its interests, the all-pervading system of globalization first developed the World Trade Center, and in the last decade, following the end of the Cold War, the GAIT and the WTO, the global "Chamber of Commerce."

As a political hegemony, globalization is managed by the G-7 group of the seven most industrialized nations (the United States, Canada, the United Kingdom, France, Italy, Germany, and Japan-now, with Russia, sometimes referred to as the G-8). This organization, headed by the finance ministers of these seven countries-like the seven wise men of Uruk at the beginning of history-was secretly formed in 1974 by the CIA, one year after the completion of the World Trade Center Towers and the formation of OPEC. The first public meeting of the G-7 occurred in the summer of 1990, just prior to the Iraqi invasion of Kuwait. The doctrine and system of globalization-now overtly controlled by the G-7 and its chief ally, the Euromarket nations-is inseparable from industrial technology as the means employed to provide objects necessary for human sustenance and comfort. Implicit in the philosophy of globalization is the right of the G-7 to promote and defend its way of life, regardless of its ultimate effects on the biosphere and at the expense of ideologies and beliefs not consistent with it. To defend itself, globalization has at its disposal the U.S. military, the command of which is housed by the Pentagon, and the new technospheric multinational military cartel, NATO. In a word, globalization is the triumph of the military-industrial complex.

From the point of view of the biosphere, globalization is the cancer of the human species consuming non-renewable resources and, through the release of free energy, effecting the final critical increase in biogeochemical combustion. The planetary instrument for furthering the entire process and philosophy of globalization is the technosphere. The technosphere may be defined as the entire apparatus of mechanized technology and its support philosophy (globalization), understood as an artificial sphere encompassing the globe. As such, the techno sphere is discontinuously interspersed within the biosphere. More precisely, the technosphere is located physically and mentally between the biosphere and the noosphere, in other words, between the Earth's vital sphere and its mental envelope.

Being the description of an evolutionary stage, the technosphere is governed by the Law of Time as a 56-year event continuum along the lines of an artificial planetary cocoon. Insofar as it is the projection of human thought, the technosphere is the materialization of the sum mental processes denoted and conditioned by mechanization in all its aspects, and that have their primary root in the mechanization of time. It is the mechanization of time that essentially establishes the artificial timing frequency as the capacity to engender the machine and, consequently, an artificial structure, the technosphere. Because it is artificial, however, its duration is highly limited and is subject to the inherent inconsistencies and contradictions in human thought and social structures because of the aberrant effects of adapting to artificial time in the first place. The effect of adaptation to artificial mechanized time creates a social-political hierarchy known as the technocracy-rule according to the needs of technology. It is the effect of maintaining the technocracy and the technosphere itself that establishes the inability of human consciousness to do anything more than remain in a constant, servile, and unchanging state of machine dominance. All the while, by contrast to the unchanging condition of consciousness, the machines evolve exponentially in number and complexity. As Marshall McLuhan so correctly put it, the humans are the bees of the machine.

To be fully understood, the technosphere cannot be seen apart from its place in the biosphere-noosphere continuum. The technosphere represents the absolute control of the biosphere by artificial 12:60 time. Artificial time generates the artificial medium of money, hence the operating philosophy of the technosphere: "time is money." Money is the lifeblood of the technosphere. Servitude to money draws humans away from the biosphere in order to be devoured by Mammon-the technosphere as an all-consuming entity. In this process the humans get money in return for selling their bodies and souls to artificial time, usually eight hours a day, often to produce artificial goods or for service in maintaining the artificial system of money itself. This money gained from servitude is used to buy consumer goods, including more machines of every kind, and thus the human participates as a vital link in the cycle of biogeochemical combustion. In this way the mechanized lifestyle converts the human-the consumer-into the biomass necessary for maintaining the technosphere.

By 2013
Earth is galactically synchronized into a new biogeological epoch: the psychozoic Era.

As the mental envelope of Earth, above and discontinuous with Earth, the noosphere can only mean one thing: the advent of universal telepathy among the humans.

The difference between the new time and the old time is like the difference between heaven and hell.

Planetary 9-11, 2001 inaugurates biogeochemical combustion, biospheric means of transformation, and self-mutation.

Technosphere:
Biospheric agent devised for maximum biogeochemical combustion.

Built in an area known as "Hell's Bottom"

The Pentagon
1941–200?

Impregnable fortress of the American war machine, meant to protect "democracy" = American domination of world market and foreign/economic policy of globalization. "The form of the Pentagon . . . is that of a pentagram, the chief magical aid for gathering power. When a pentagram is drawn with its base toward the north, and its chief point to the south, it is useful for negative or black magic purposes."
—*Transformative Vision*, page 249

*Note: Gregorian calendar coordinated by Law of Time. Repeats precise order of days and weeks precisely every 28 years.

By 2012
Humans realize a permanant condition of peace through culture.

By 2009
Humans are reorganized by the Law of Time and begin the construction of a galactic culture on Earth.

Dismantling of the technosphere stabilizes the biosphere and reveals the noosphere—2005.

By leaving artificial time and returning to natural time via the Thirteen Moon calendar, humans advance the noosphere.

2004: Last opportunity for humans to change their macro-organizing principle.

Exhaustion of the Technosphere—
2001–04.

Exponential growth curves of money, machine production, energy, and people climax.

Planetary 9-11:
time runs out on the technosphere.

"New" millenium nipped in the bud; civilization R.I.P., 2001

Technosphere:
Interlocking interactive pentagon of five artificial systems - WTC = Nerve center and hub of the technosphere and its driving power, globalization.

Biosphere:
Controlled by artificial (12:60) time.
Time is Money—lifeblood of the technosphere.

Humans devoured by Mammon. Get money in return and create biomass to maintain technosphere.

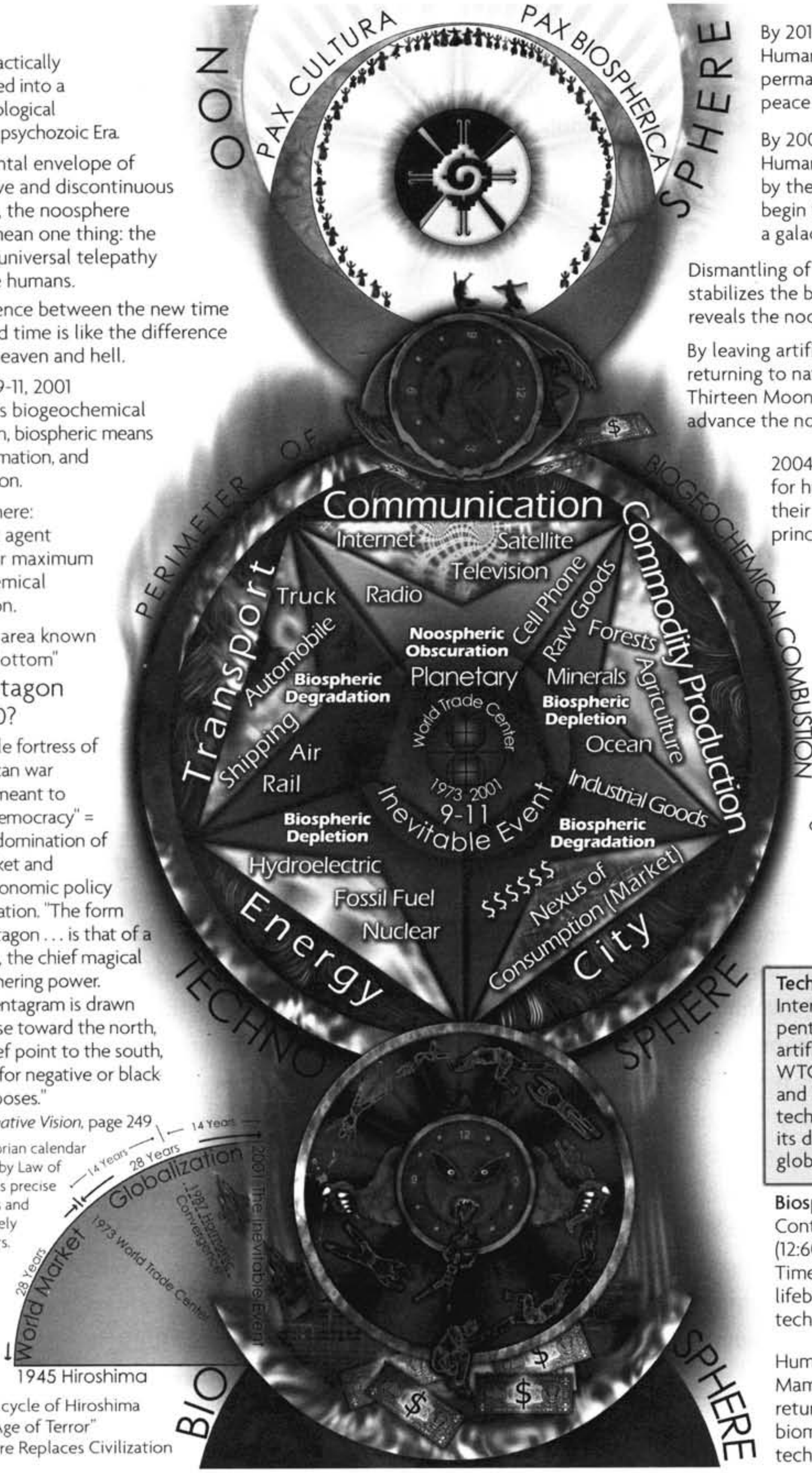
NOOSPHERE
PAX CULTURA
PAX BIOSPHERICA
PERIMETER OF BIOGEOCHEMICAL COMBUSTION
Communication
Internet
Satellite
Television
Radio
Cell Phone
Transport
Truck
Automobile
Shipping
Air
Rail
Commodity Production
Raw Goods
Forests
Agriculture
Minerals
Ocean
Industrial Goods
Noospheric Obscuration
Planetary
World Trade Center
1973 2001
9-11
Inevitable Event
Biospheric Degradation
Biospheric Depletion
Biospheric Depletion
Biospheric Degradation
Energy
Hydroelectric
Fossil Fuel
Nuclear
$$$$$$
Nexus of Consumption (Market)
City
TECHNO SPHERE
BIO SPHERE
14 Years
28 Years
14 Years
Globalization
1987 Harmonic Convergence
2001 The Inevitable Event
1973 World Trade Center
World Market
28 Years
1945 Hiroshima

56-year cycle of Hiroshima
"Age of Terror"
Technosphere Replaces Civilization

Biosphere—Technosphere—Noosphere

As an operating planetary structure, the technosphere consists of five interactive components.

1. COMMODITY PRODUCTION

Commodity Production, the basis of the entire World Market as a biospheric system, refers to the transformation of natural resources through industrial means into consumer products. Commodity production alone accounts for the complex of industrialization that brings the World Market and the technosphere into existence. According to *The Biosphere Catalogue,* the post-World War II growth of the World Market necessitated the establishment of the World Trade Center, a global complex (114 such centers in 1985), but with its principle coordinating unit being the Twin Towers in New York City, home of the World Trade Institute and the corporate offices of the New York Stock Exchange, the key to the world money and commodities market.

2. CITIES

Cities, the second technospheric component, represent the human social organism or marketplace for accommodating the production and consumption of industrial goods. Civilization is city life as it has evolved from Uruk to New York (= New Uruk). While prior to industrialization the city depended on a healthy agrarian economy for its support, industrialization pushed the city to new heights of artificiality, not the least of which were round-the-clock lighting systems that allowed for 24-hour consumer markets. As such, the city is the nucleus of the technospheric system, the hub of its commodity production, of its energy consumption, and of its transport and communication empires. The growth of the cities around the world is exponential in the second half of the twentieth century, reflecting the triumph of the technosphere. It should be noted that in the fifty-six years of the technosphere, the human population almost tripled-from 2.2 to 6.2 billion, a growth demonstrated in the vast expansion of the large urban centers.

3. ENERGY

Energy is the third component of the technospheric complex. The tremendous needs of industrial production have transformed the natural energy processes of the biosphere into gigantic artificial complexes for the production of energy to maintain factory and city alike. The impact of the different artificial energy systems upon the biosphere has been enormous. These systems include coal, oil, and fossil fuels in general, as well as hydropower and nuclear energy. The whole syndrome of artificial time, artificial means, acceleration of the machine, and expansion of human population has required and created the enormous output of artificial energy into the biosphere and contributed greatly to the destabilization and loss of natural resources and atmospheric and hydrospheric contamination through the release of toxic "free energy" by-products.

4. TRANSPORT

Transport. Without the highly industrialized transport systems, the commodity production, the vitality of the cities, and the access to provisions of fuel required for the energy systems would collapse. No sooner had industrial production begun early in the eighteenth century than the first artificial transport systems came into being: the steam- or coal-driven locomotive, and the great transoceanic shipping lines. By the twentieth century, the railroad system and worldwide shipping industry were supplemented by the invention of the combustion engine, which made the trucking industry possible. Following World War II, rocketry and jet propulsion established the transportation might of the great airline companies of the second half of the twentieth century. It is no coincidence that commercial airplanes, the symbol of the technosphere's advanced transport systems, were used to bring down the Twin Towers and to penetrate the Pentagon.

5. COMMUNICATION

Communication is the fifth component that unifies the technosphere into one whole system in touch with itself everywhere. From telegraph to telephone, radio, television, and then finally fax, cell phone, and the Internet, the industrial-era technologies, with their electronic instantaneity, have created the vast communication networks and media empires that keep the world humming and promote globalization above all. When considering modern communication we must also bring into focus the chief method of globalization, which is marketing-propaganda on behalf of consumerism. Marketing is the use of means of mass communication to control the mind of the consumer. In the final stage of globalization, marketing replaces consciousness, or rather, marketing is the manipulation of consciousness-the latter already having been reduced to a stunted level-for purposes of consumerism, or for ideological needs such as patriotism and fear campaigns. The Internet, the marriage of computer and telephone technologies, represents the final stage of the Tower of Babel, and the completion of the process of globalization. The only place to go after the Internet is technology-free telepathy.

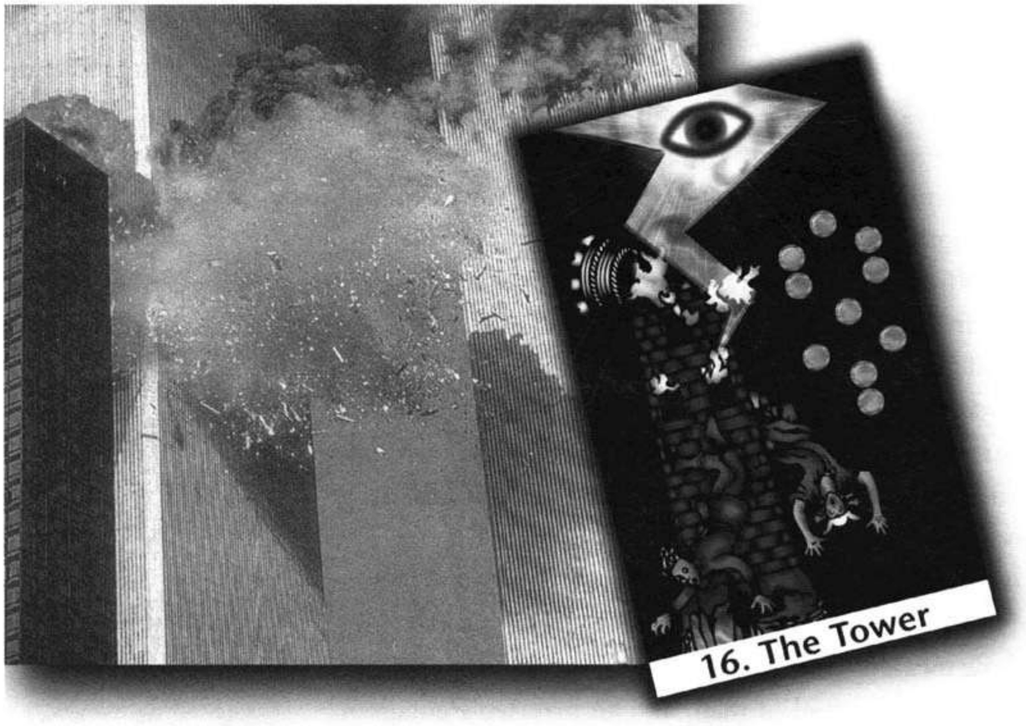

The Tower

From the biospheric perspective, this five-tiered complex of the technosphere is a whole unit. All the subsystems evolved together to establish the structure and operating procedures of the technosphere. When the technosphere goes, the entire system folds. This is what is now beginning to occur. Following the Inevitable Event, the whole technospheric system will be coming down slowly over the next few years, like a giant circus tent that has lost its central prop. How gracefully or gracelessly this happens is dependent on the response of the Pentagon-or the will of humanity to rouse itself in the face of even worse barbarisms.

Because it consists of five interactive components, the techno sphere can be graphically depicted as a pentagon, a five-sided figure. At the center of this pentagon of the technosphere are the World Trade Center Twin Towers. The actual Pentagon, which was attacked along with the Twin Towers, was designed in the late 1930s to house the U.S. Department of War. Its construction occurred in the years 1941-1943, in an old neighborhood of pawnshops and bars called Hell's Bottom.2 When it was built it was the world's largest office building in terms of actual space. Of course, when they were built, the Twin Towers of the World Trade Center were the world's tallest office buildings. The structure of the Pentagon is the prototypical morphology of the structure of the five systems that constitute the technosphere. The Pentagon was the impregnable fortress of the American war machine, protecting American interests and globalization around the world. If there were two central points-actual and symbolic-to what is known as the military-industrial complex, it was the Pentagon and the World Trade Center, and this is undoubtedly the reason why they were both targets of the Inevitable Event.

What is architecturally noteworthy about the Pentagon is that it is designed with five inner pentagonal corridors and office slabs, and that it is constructed with its odd point to the south. Since a pentagon is actually a pentagram, a five-pointed star with its points connected, the Pentagon represents an *inverted* pentagram, as its odd point faces south rather than north. In the traditional Tarot deck it is interesting that the Fifteenth Major Arcana, The Devil, contains an inverted pentagram between the horns of the Beast. Paul Foster Case writes, "This is a key to the whole meaning of the figure [of the devil]. For the Pentagram is the symbol of man, and an inverted Pentagram suggests the reversal of true understanding of man's place in the cosmos."3 While the Fifteenth Major Arcana contains the pentagon as the inverted pentagram, the next card, the Sixteenth Arcana, is The Tower, which shows the tower being struck and broken apart by lightning (fire of heaven), with humans falling or jumping out of the windows to their death-as accurate a symbolic depiction of the collapse of the Twin Towers as could be found.

What the startling conjunction of these symbolic correspondences points to is the profoundly archetypal nature of the Inevitable Event. It has been prefigured and prophesied, and has now come to pass. No matter how much America and its allies, many of them bought for a price, may strike back, the deed has been done. The Towers are gone; the Pentagon has been penetrated. While it will take several years for the full realization of the absolute magnitude of this archetypal event to sink into the

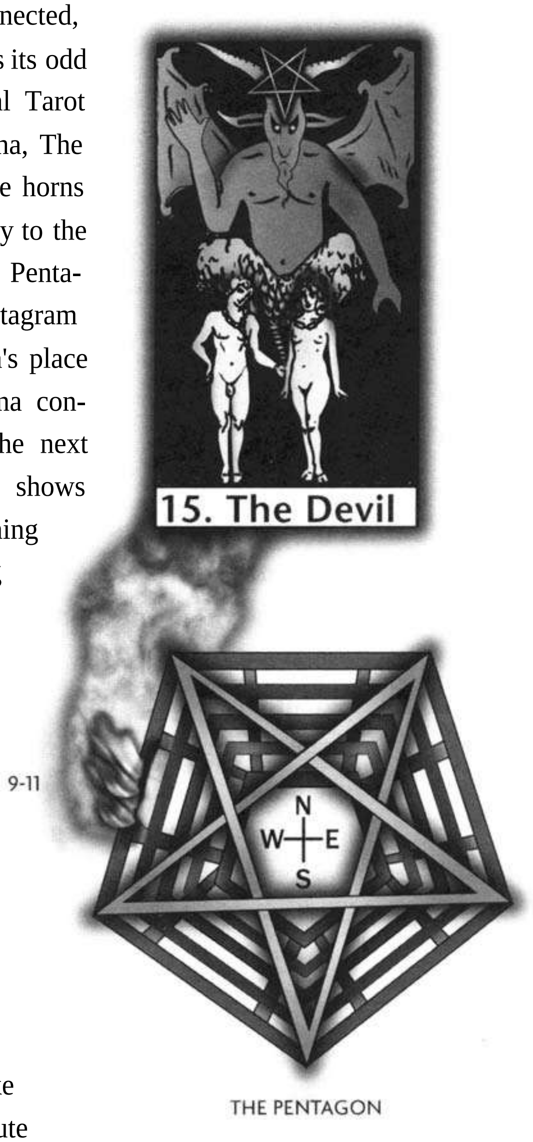

The Devil and the Pentagon

collective mind, it is important to demonstrate the actual structure in time of the techno sphere so that the finality of its end may be seriously considered, and the gate to the future, which has thereby been opened, can be made clearly visible.

As we have shown, the technosphere is a structure totally brought about by the 12:60 timing frequency, and thus is purely a function of the latter. As such, the technosphere is embedded in the global macro-organizing principle of the 12:60 frequency, the Gregorian calendar. Two 28-year cycles, each one divided into four 7-year subcycles, define the duration of the technosphere. Remember that the Gregorian calendar repeats every twenty-eight years. As we have pointed out, the 56-year cycle of the technosphere, 1945-2001, is preceded by the 44-year cycle of the proto-technosphere. The two World Wars were a function of the proto-technosphere, the final major act being the construction of the Pentagon, specifically for consolidating the American war machine-the world's largest office building means that the world's largest business is war, with a $318 billion budget for the year 2001-2002 alone!

The fifty-six years between Hiroshima and the Inevitable Event were the age of terror, for it was atomic terror that initiated the technosphere in a baptism of nuclear fire, and in the end, it was an unimaginable suicidal terror that brought down not just one, but both of the Twin Towers of Babel. The final collapse of the technosphere is also the final war between blood and money. Artificial time has run out. Only a new time will be able to regenerate the biosphere and spiritually revive mankind. What follows is a chronological description of the eight 7-year stages of the fifty-six years of the technosphere. Note the persistence of certain themes. (See plate 5, Noospheric Time Map-Wave Harmonic of History.)

GREGORIAN A, 28-YEAR CYCLE: TRIUMPH OF THE WORLD MARKET, 1945-1972

1. 1945-1951. CYCLE OF THE BOMB: ATOMIC TERROR AND THE DIALECTIC OF THE COLD WAR.

During the first dynamic 7-year cycle, with the triple event of the first test of the atomic bomb and its two detonations at Hiroshima and Nagasaki, the biosphere becomes irrevocably altered by the introduction of a constant, steady state background radiation into the atmosphere, the actual inception of the biogeochemical combustion. This act officially establishes the technosphere. The reality is that the humans also created their first weapon of mass destruction. It is this act that also immediately sets in motion a destabilization of the human consciousness in the noosphere.Within two years, Mahatma Gandhi, the world's foremost pacifist, is assassinated, as a "free" India is partitioned into three parts-India, Pakistan, and what is later to become Bangladesh. In the same year, 1947, the iron curtain between communist Eastern Europe and the West, epitomized by the Berlin Wall, turns the enmity between the two major powers, the Soviet Union and capitalist America, into the Cold War. The subsequent formation (1949) of the NATO military cartel counters the United Nations, which was founded in 1945 to replace the League of Nations. The UN now has the task of keeping World War III from happening by providing a forum where the "superpowers" can keep their cold war from heating up, and that will contain and manage the spread of nuclear power. Nonetheless, the terror of the Bomb as the ultimate deterrent engenders the arms race. Soon Russia, the United Kingdom, France, and eventually China get the Bomb. War in Indochina, the establishment of the Israeli state in Palestine (1948) at the expense of the sovereignty of the Palestinian people, and the completion of the Chinese Marxist Revolution under Chairman Mao are the highlights of 1949, followed by the Korean War in 1950. All this demonstrates the instability of the world. At the same time, 1949, commercial television production begins in the United States; the age of radio is replaced by the "tube." Americans also begin building freeways and the new system ofInterstate highways marking the triumph of "automobile culture," and the beginning of the suburban consumer lifestyle.

2. 1952-1958. CYCLE OF THE OPENING OF THE TOMB AND THE BEGINNING OF THE SPACE AGE.

In the summer of 1952 a Cuban archaeologist discovers the tomb of the Mayan sage, Pacal Votan. This unprecedented archaeological event marks the beginning of the final sixty years-three katun cycles-of the Mayan thirteen baktun cycle of history. In the following year, 1953, the superlethal H-bomb is tested. The DNA code is discovered, as are the Earth's radiation belts. The United States and Russia conduct numerous nuclear weapons tests in Nevada, Siberia, and the Pacific Ocean. Wars for independence occur in various African states, and in general the era of European imperialism is at an end, followed by the neo-colonialist (Third World) era of guerrilla warfare, poverty, and social instability. In 1956, the year the Russians launch the first sputnik and begin the space age, the United Nations tables the issue of calendar reform indefinitely, thus closing the chapter begun when the League of Nations proposed global calendar reform in 1931. The year 1956 also marks the beginning of the electronic pop culture of the technosphere, rock and roll. By the end of this cycle, the "space race" joins the arms race as a force in promoting the advance of the technosphere. Commercial airlines adapt to jet propulsion, and the great age of air travel begins, with the airport to become the centerpiece of globalization. Initial experiments in computer technology-Univac-and the first great computer corporation, IBM, announce the beginning of what is to become the major technology of globalization. A small fast-food franchise named McDonald's opens in the United States.

3. 1959-1965. CYCLE OF THE ATOMIC MIND EXPLOSION.

The July 26 Cuban Revolution of Fidel Castro, triumphant in 1959, establishes a communist state in the Americas and heightens the Cold War. That same year the Chinese communists bring to a definitive end the rule of the Dalai Lama in Tibet. In the Congo, the popular hero Patrice Lumumba is assassinated after only two months in office. Also in 1960, the human population hits three billion, an increase of one billion since 1930. In 1961 the Institute of Mathematics of the Siberian Branch of the Soviet Academy of Sciences begins the monumental mathematical analysis of Mayan hieroglyphic texts. The attempted 1961 Bay of Pigs invasion of Cuba is followed in 1963 with the assassination ofJohn F. Kennedy, an act that reflects a rising level of cultural turmoil in the United States. Martin Luther King Jr. emerges as a popular hero of nonviolence and civil rights, an issue that points to a more radical discontent in America, culminating in 1965 with the Watts race riots. Rachel Carson's book *Silent Spring* (1963) signals the rise of ecological awareness. The Beatles transform rock and roll into a world-class cultural phenomenon that foments a widespread countercultural movement and atomic mind explosion, fueled by Timothy Leary, LSD, hippies, and the new antiwar movement against the conflict in Vietnam. In 1961 Yuri Gagarin is the first man in space, and the Vatican II Ecumenical Council comes to the conclusion that the best way the Church will survive in the next century is through a strong Pope. The Vatican II document also includes an appendix on calendar reform, the wording of which makes it difficult for there to be any alternative but a watered-down version of the Gregorian calendar. This act seals the technosphere within the unquestioned confines of Gregorian time, ultimately turning the atomic mind explosion in on itself.

4. 1966-1972. FROM THE WHOLE EARTH TO THE WORLD TRADE CENTER: THE TRIUMPH OF THE WORLD MARKET.

In 1966 construction begins on the World Trade Center Twin Towers, while in 1968 World Trade Centers are independently established in Houston, New Orleans, and Tokyo. That same year, UNESCO hosts the first and only conference on the biosphere ever sponsored by the United Nations, "Man and the Biosphere" (MAE), but the program slowly dies on the vine. The cultural conflict in the United States and the entire world also reaches a fever pitch in the years 1967-1969. This ferment includes: race riots, antiwar riots, a march on the Pentagon, the assassinations of Martin Luther King Jr. and Robert Kennedy, and riots in major capitals around the world and on both sides of the iron curtain, as well as the Maoist cultural revolution in China. The excitement of the '60s climaxes in 1969 with the success of the Apollo XIII manned mission to the moon and the consequent feedback factor of the human race seeing itself instantaneously on television as the whole Earth beamed from space, an opening to the noosphere. Apollo XIII initiates the next phase of the space race, travel to other planets. Also in 1969, the Woodstock concert marks the apex of the early rock-and-roll culture and the visionary climax of the cultural revolution. The 1971 Concert for Bangladesh establishes rock and roll with a conscience. In 1970 the vision of the whole Earth is translated into the first Earth Day and the birth of the ecology movement. The United Nations follows two years later with the Stockholm Conference on the Environment. The revolutionary fervor of the previous years turns more violent, and repressive governmental police powers win the day. The first 28-year cycle ends with a whimper, the dynamic destabilization of consciousness reaches a plateau of normalcy-a normalcy to be increasingly punctuated by terrorist activity, as in the Palestinian takeover of the Olympic Village in September 1972. Both apart from as well as in reaction to such activity, the technosphere has now become all-dominant and all-powerful in human affairs.

SUMMARY, GREGORIAN A: 1945-1972

During this cycle, the dynamic of the accelerated geochemical combustion destabilizes the human consciousness while increasing the curve of the machine in human affairs. By 1973 the machine curve meets and surpasses the human population curve; this stabilizes human consciousness at a level of machine normalcy (MN). Human consciousness will remain stabilized at MN, largely through varying forms of terror and fear programming, for the duration of the second and final 28-year cycle. This normalization of human consciousness allows the diverse and rapidly evolving technological components of the technosphere to reach a point of consolidation. This stabilization of the World Market fosters its next and concluding phase: globalization and the triumph of marketing over human culture. It should be kept in mind that the instruments and weapons of mass destruction, of which there were none on January 1, 1945, have now proliferated to five different countries and provide the ultimate backdrop to these first twenty-eight years. The core of the UN's Security Council is formed by the five countries that possess atomic weapons. Geopolitically, the bio-

Machine Normal Consciousness in Relation to Technospheric Growth Curves

Human consciousness, overwhelmed by war—the ever-propagating machine and bondage to artificial time—succumbs to money and stabilizes at Machine Normal, the standard consciousness of the technosphere, which represents an ever-diminishing ratio of intelligence in relation to the growth curves of population, machine, money, and war.

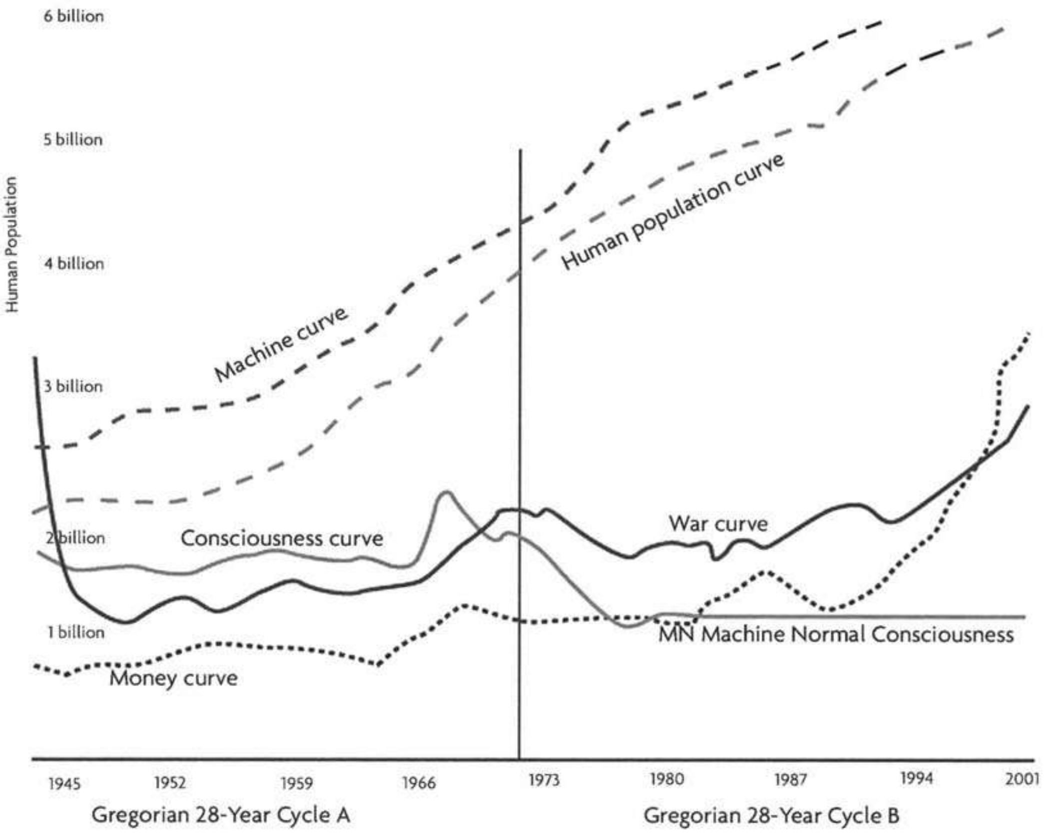

sphere is in control of technocratic powers that are virtually all located in the Northern Hemisphere-whether it is the Pentagon, NATO, the World Trade Center, the United Nations, or the Soviet space program.

GREGORIAN B 28-YEAR CYCLE:
CLIMAX OF GLOBALIZATION, 1973-2001

1. 1973-1979. CYCLE OF THE CONSOLIDATION OF THE TECHNOSPHERE: CLUB OF ROME, OPEC, AND THE G-7.

The dedication of the World Trade Center Twin Towers on April 4, 1973 (a repeat of the calendar year of the Bomb, 1945) is the counterpart to the construction of the Pentagon (1940-1941), which sponsored the bombing of Hiroshima and Nagasaki. The tallest buildings in the world when they opened, the 110-story Twin Towers seal the position of New York as the "capital of the world," for on the other side of Manhattan is the glass tower of the United Nations. Simultaneous to the dedication of the Twin Towers is the formation of the Organization of Petroleum Exporting Countries (OPEC). With the exception of Venezuela, OPEC is largely an Arabic (Islamic) economic alliance that, for the first time in the history of the technosphere, brings a Third World influence into planetary politics. The subsequent "oil shortage" demonstrates the technosphere's vulnerability, which in turn causes a further "hardening of the arteries" of its consciousness-the covert formation by the CIA of the G-7. Also in 1973, after five years of study the Club of Rome issues *The Limits of Growth,* which in its analysis of trends foresees a major environmental (biospheric) crisis within one hundred years, yet nonetheless establishes and promotes the standard world model, a setting of production quotas at the 1973 level, to foster sustainable growth at least until 2020. The "standard world model" corresponds to the "normalization" of consciousness. The third war (previous ones having occurred in 1967 and 1949) between Israel and its neighbors occurs this year. By 1975, the world population is at four billion, up a billion since 1960. In 1976 Voyager II reaches Mars. In 1978 Karol Wojtyla becomes Pope John Paul II, the first global Pope, fulfilling the Vatican Council's requirements of a popular Pope to assure the survival of the Church into the next century. With Gregorian civilization triumphant, Mexico, home of the Maya, is the first country visited by the new Pope. In the same period, between 1978 and 1979, the Iranian Islamic revolution occurs, toppling the pro-Western Shah, signaling a heightened conservative Islamic resistance to Westernization and the technosphere. Japanese technology and corporate capitalist economy flourish, capped by the construction in Tokyo of the "New City" Shinjuku and, despite many environmental protests, the Narita airport. The mainstream global consciousness is now "comfortably numb" and on course to unprecedented materialism.

2. 1980-1986. CYCLE OF THE TRIUMPH OF THE TECHNO SPHERE.

The first cycle of the stabilized new normalcy establishes the triumph of marketing-the election of a second-rate American movie star for President, and the birth of the yuppie (young urban professional) generation of America. The rise of MTV signals the standardization of rock style as global culture; while the Live Aid global television concert on July 13, 1985, to raise awareness of AIDS in Africa, is the largest ever of its kind. The September 17, 1985 Mexico City earthquake, the worst in modern times to hit the world's largest city, demonstrates the continuing vulnerability of the techno sphere to natural catastrophe. The early 1980s witness the Intifadah uprising of Palestinians against Israel. The War on Drugs, an aggressive antidrug campaign waged by the West against narcotraffic, mostly from South America and the Middle East, heightens the new normalcy as world drug laws tighten, placing marijuana and psychedelics in the category of "hard drugs." Yet drugs continue to be a major social problem in the industrialized West. In 1986 President Reagan attacks Kadafy of Libya and invades Grenada, testing American ability to commit limited acts of aggression without inciting world censure. The construction of Biosphere II in Arizona, an effort to test the possibility of establishing an artificial biosphere on other planets, ends inconclusively. In 1985 in Bhophal, India, the worst modern industrial toxic chemical disaster occurs with more than 2,000 deaths, followed on April 26, 1986 by the Chernobyl nuclear power plant disaster, the worst since the beginning of the nuclear age. In America the suburban automobile culture gives rise to the super shopping malls. The personal computer (PC) is introduced, while electronic game arcades flourish. Sunday football becomes the major American television pastime, while television itself makes world soccer the number one competitive global sport; the broadcast of professional sports assumes the role of a primary noospheric diversion.

3. 1987-1993. CYCLE OF THE END OF THE COLD WAR AND TRIUMPH OF THE G-7.

On January 24, 1987 Voyager II sends back photos from Uranus; four days later, January 28, 1987, seven American astronauts die in a shuttle launch explosion. On August 16-17, 1987 the Harmonic Convergence peace meditation attracts hundreds of thousands who gather at sacred and natural sites around the world; later that year, October 19, 1987, the New York StockMarket crash occurs, the biggest point drop ever, and Russian Prime Minister Gorbachev initiates disarmament accord, or *glasnost.* By the end of 1987 the population is at five billion, up one billion since 1975. During 1988-1989, a rolling wave oflargely peaceful revolutions and civil uprisings throughout most of the Warsaw pact nations culminates on December 31, 1989 with the tearing down of the Berlin Wall-the end of the Cold War. At the same time, President Bush invades Panama to continue testing Drug Wars and limited-aggression capability. In the summer of 1990 an international bankers' reunion is followed by the first public meeting of the G-7. Every year henceforth, the G-7 meets publicly to set policy for globalization. Within a month of the first public G-7 meeting, August 1990, Saddam Hussein invades Kuwait, and Bush retaliates with the Gulf War, January-February 1991. In August 1991 Gorbachev is ousted in a putsch, marking the end of the Russian Marxist era. Destabilization, war, and genocide afflict the Balkans. A new wave of environmental concerns-ozone depletion, destruction of rain forests, and global warming-inspire a reactivation of Earth Day, 1990, followed two years later by the 1992 Rio Summit, the second United Nations Environmental Conference and the first in twenty years. The conference generates much paper, but the doctrine of sustainability (sustainable industrial growth) is the principle outcome. Meanwhile, with the absence of Marxist Russian economic influence, the neoliberal capitalist economics of America flourish as full-blown globalization, multinational corporations, and the prominence of the IMF and World Bank foment a uniform global corporate culture of cheap consumer services and franchised retail outlets for brand-name goods. This period also sees the rise of East Asian capitalist economies, although the Japanese economic bubble bursts in 1991. Early in 1993 a terrorist bomb explodes in the underground garage of the World Trade Center Twin Towers. The period 1989-1993 marks the beginning of the discovery of the Law of Time-the definition of the artificial 12:60 and natural 13:20 timing frequencies, the Dreamspell codes, and the decoding of the Telektonon Prophecy of Pacal Votan in 1993.

4. 1994-2000. CYCLE OF THE CLIMAX OF GLOBALIZATION.

In 1994, the GAIT Treaty is signed, the Levy-Shoemaker comet hits Jupiter, and a million Mricans die in the Rwanda massacre. The indigenous uprising in Chiapas, Mexico on January 1, 1994 provides a counterpoint to the triumph of globalization that is manifest through worldwide airport expansions, and an aggressive form of tourism that feeds the airline industry. Israel and Palestine enter the "peace process," while Yeltsin consolidates his power in Russia. This is the era of United Nations peace-keeping missions, mostly in the Balkans and Africa. Following the FBI/BATF -inflicted firestorm on a fundamentalist compound in Waco, Texas, a year later, in 1995, a terrorist bomb destroys the Oklahoma City Federal Building. The United States and the United Kingdom inaugurate control ofIraqi no-fly zones, enforced by regular bombing raids; the United States also conducts military actions in Somalia, as well as more successful ones in Haiti. This is the era of rapid expansion of computer and biotechnologies, exponential rise of New York Stock Exchange and the new specialized technologies exchange, the NASDAQ. Bill Gates of Microsoft typifies the rise of the new self-made billionaires of the Clinton era. In January 1995 the Kyoto earthquake occurs, followed by the Tokyo subway nerve-gas terrorist attack, while the Internet, the "information superhighway," is fully up and running. In 1996 in Brasilia, Brazil, the First Planetary Congress of Biospheric Rights signals rise of a new calendar reform movement with the proposal of the World Thirteen Moon Calendar Change Peace Plan, submitted to both the United Nations and the Vatican. In 1997 the Kyoto environmental conference on global warming sets policy, but the United States refuses to ratify it. In 1998 simultaneous U.S. embassy bombings in Kenya and Tanzania signal a new phase of war on terrorism, singling out a Saudi national, Osama Bin Laden, as the terrorist mastermind. By 1999 the world population is six billion, up one billion since 1987. In late spring 1999 NATO initiates a unilateral bombing campaign ofYugoslavia over the disputed region ofKosovo. The World Summit on Peace and Time, at the University for Peace, Costa Rica, sends emissaries to the Vatican and the United Nations with the Declaration of Thirteen Moon Calendar Reform; later in 1999 WTO has its first meeting in Seattle, which sparks anti-globalization riots. In 2000 the Y2K computer scare surrounding a bad time program fails to materialize. On Rosh Hashanah, September 28,2000, Hebrew year 5761, Israeli leader Ariel Sharon's presence at Temple of the Mount, Jerusalem, sparks a new Intifadah of Palestinians, which continues unabated through 2001, the official first year of the new millennium.

5. 2001-? *GOTTERDAMMERUNG* OF THE TECHNO SPHERE: FIRST YEAR OF THE THIRD MILLENNIUM, REPEAT OF GREGORIAN CALENDAR YEARS 1945 AND 1973.

In July 2001, G-7 meets in Genoa, Italy, as more than 100,000 anti-globalization protesters riot. In early September the United States and Israel walk out of the United Nations Conference on Racism in South Africa. Then, within ten days occurs the Inevitable Event, Gregorian 9-11, 2001, the cumulative synchronic pressure point where the technospheric bubble-the artificial time warp in the noosphere-bursts open, spawning in its wake the full-scale War on Terrorism-another name for World War III. The Inevitable Event exemplifies the built-in failure of the Gregorian calendar to commence another 28-year cycle as was done in 1945 and 1973. If the 1945 cycle began with the end of one world war, the 2001 cycle begins with the start of one. Since 2001 is also first year of the third millennium, the Gregorian calendar has created its own apocalypse, opening a noospheric fissure-the climax of the biogeochemical combustion of the biosphere. Now it is only a matter of time before the technosphere experiences total collapse. The nature and extent of the damage sustained during the collapse is in proportion to the intelligence that responds to the knowledge of its collapsing. If this 28-year Gregorian calendar cycle were to go unimpeded until 2029, it is not certain what would be left of the biosphere. But there is one thing of which we may be certain: *artificial, linear time has run out.* An American newspaper editorial at the end of 2001 sums up the crash of the dominant linear time paradigm: "History seems to have slammed into reverse gear around a world that was widely assumed a year ago to be moving in a straight line toward a more prosperous and secular future, one that would be shaped by stock markets and the Internet, not by turbaned zealots or theocracies that have or want nuclear trigger fingers."4

SUMMARY, GREGORIAN B: 1973-2001

With destabilization normalized, the technosphere replaces civilization and marks the end of culture-the new barbarism prevails, culminating in total war. The 9-11 is what actually completes the technospheric cycle. Henceforth it will be the advent of the noosphere, but not in the way predicted by the authors of *The Biosphere Catalog:* "The Noosphere as an active force consists in the harmonious synthesis of the Biosphere and the Technosphere by intelligence."[5] By the analysis of the Law of Time, the intelligence is nowhere evident because there is no harmony of intelligence possible under the present timing standards that dominate all the major players in the present global conflict. Since the technosphere is totally a function of the artificial timing frequency, and since artificial time is now short-circuiting itself, there will be no technosphere to harmonize when natural time is restored to the entire biosphere. Then the noosphere will commence, but only if there is world harmony. Only a harmonic standard will bring harmony to the world. The failure of the Gregorian calendar and its civilization, crystallized in the now-collapsing technosphere, must be quickly acknowledged. By adopting a harmonic standard as soon as possible, the shift to the noosphere can commence, rising and building itself up from the shards and ashes of the toppled technosphere.

▼

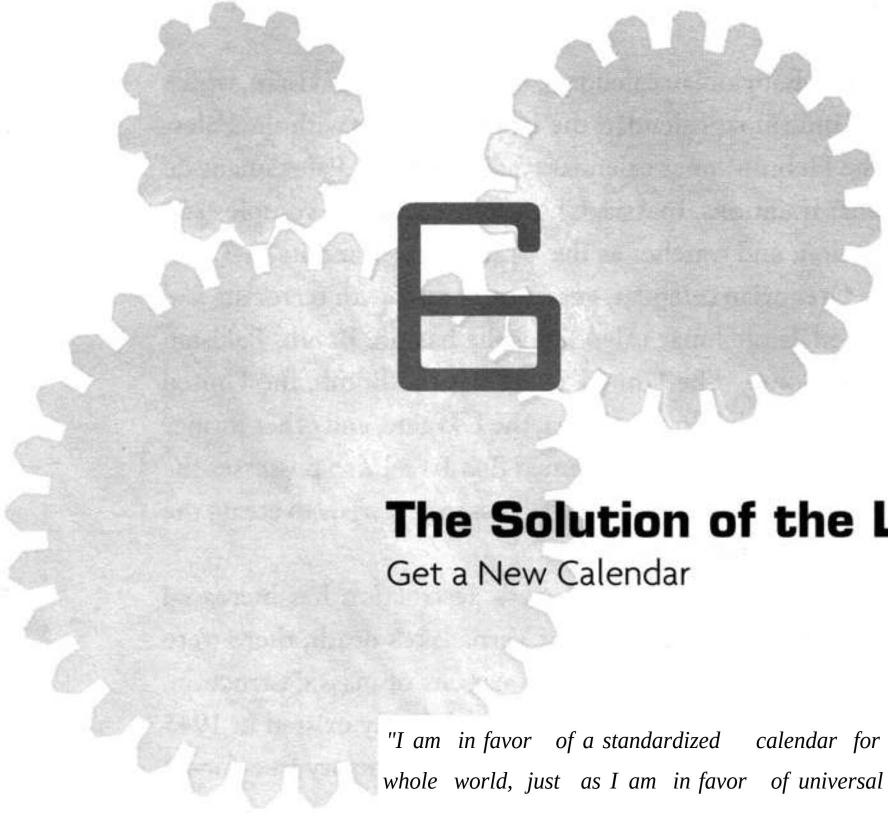

6

The Solution of the Law of Time

Get a New Calendar

> *"I am in favor of a standardized calendar for the whole world, just as I am in favor of universal coinage for all countries, and a supplementary artificial language (like Esperanto, for example) for all peoples. . . I am always ready to endorse any honest movement which will help unify the peoples of the world."*
>
> -MAHATMA GANDHI, QUOTED IN *THE JOURNAL OF CALENDAR REFORM,* 1931)

VIEWED FROM THE PERSPECTIVE of the noosphere, the failure of human civilization is the failure of the human being to become synchronized with itself in the correct time of universal nature. Instead of operating as a planetary organism in absolute synchronization with a knowledge of being coordinated by a common harmonic standard, humans race toward the end of history, dissynchronously operating under a plethora of calendars of differing measures enshrining competing belief systems, all globally coordinated by an instrument of irregular measure and obscure nomenclature-the Gregorian calendar. Dissynchronized and at odds with itself everywhere, humanity is broken apart into splintered mind fields, each controlled by a different timing standard, a fact about which the human race is virtually ignorant- to its own detriment.

India, with its different neo- Babylonian calendars, is at odds with Pakistan, which operates according to the Islamic lunar calendar; the Israelis, at odds with the Palestinians, hold as tightly to the Hebrew lunar calendar tradition as the Palestinians do to their Islamic lunar calendar traditions. In Asia, China follows its own complex and ancient lunar calendar tradition, and watches as the United States and its G-7 alliances, all dominated by the Gregorian calendar, get ready to spar with terrorists and militants who also follow the Islamic lunar calendar. India has the Bomb, Pakistan has the Bomb, China has the Bomb. The United States has the Bomb, the United Kingdom has the Bomb. France has the Bomb. Russia, the Ukraine, and other former Soviet nations have the Bomb. It is commonly assumed that Israel also possesses the Bomb. Does Iraq have the Bomb? Can a terrorist get access to materials to create the Bomb as well? The nuclear clock is ticking again.

The number of weapons and technologies of mass destruction has increased several thousandfold since 1945 when, at the time ofVernadsky's death, there were virtually none. This is not to mention the biological weapons of mass destruction, more ambiguous, more lethal, more hidden-none of which really existed in 1945 either. With the fallout from the Inevitable Event an anarchic planetary free-for-all is taking shape. From the noospheric point of view, the humans have run out of (artificial) time, and are now committing their last egoistic act of desperation to hold what they perceive to be their dominant place in the scheme of things. The fact is, however, that the planet belongs to the biosphere and not to the humans.

The inability of the Gregorian calendar to sustain another 28-year cycle, much less another millennium, means that an anarchy of time and a time of anarchy is loosed upon the planet. At the present rate, what would the world look like in 2029? One does not want to imagine. The anarchy of time is exemplified by the different competing powers that are unconsciously driven and motivated by programs stored in their differing calendar systems. This failure of artificial time to contain itself is truly the apocalypse. The apocalypse could occur only when the time program had become saturated. The two 28-year cycles, 1945-2001, carried forward not only the exponential advance of machine time, but the accrued karmic baggage of two millennia of the invented religion of Christ. Throughout the last millennium, the protagonists of this religion waged a type of holy war against an embattled world-a war waged by missionaries and conquistadors alike, down to this very day. The week before the Inevitable Event, the Taliban was deporting Christian missionaries from Mghanistan, while other Christian missionaries scour the jungles looking for one more indigenous soul to convert. This activity is inseparable from the Gregorian calendar that programs it.

Entropic Black Hole of Consciousness

A calendar is a mental instrument for holding the programmed patterns of thought and behavior of a given culture, people, or civilization. A people can rise no higher than the program of time their culture imposes on them as second nature. A program of time determines nature of consciousness. Different timing programs scheduled by synchronically inaccurate devices within a biospheric whole governed by a universal timing frequency apart from the deviant human standards can only result in an entropic black hole of consciousness: the clash of civilizations and cultures that precedes the mentally cleansing force of the noosphere and the inevitable return to natural time and the universal 13:20 timing frequency.

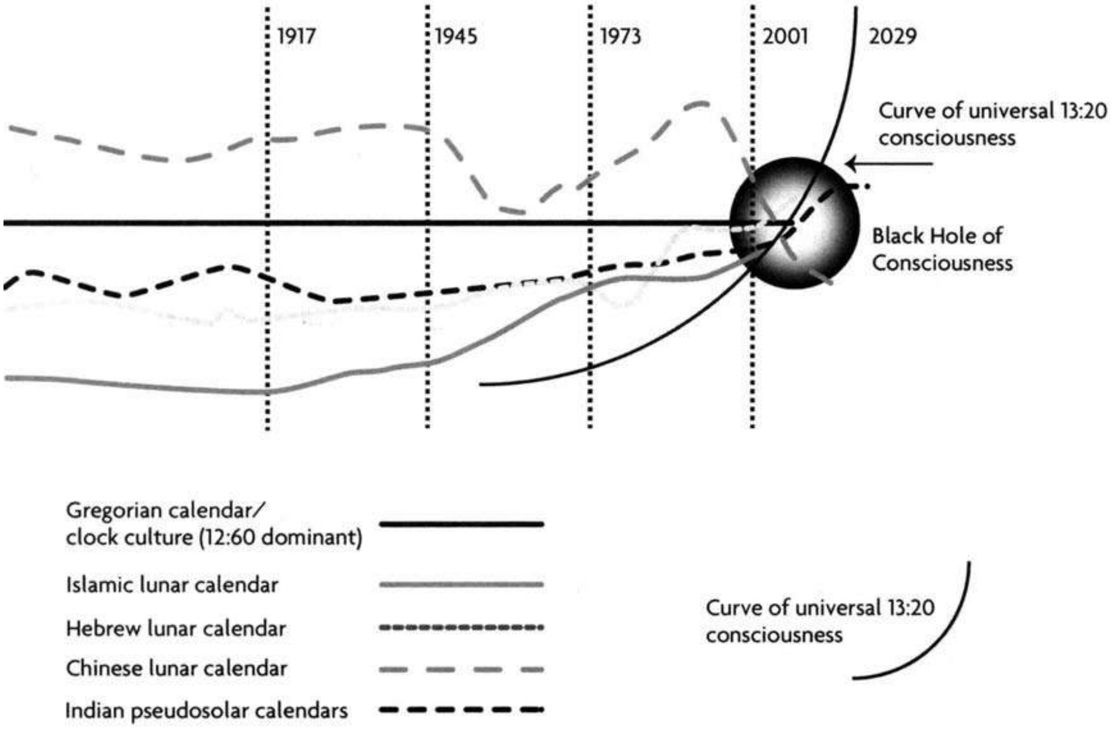

Now the default of the Christian program of the Gregorian calendar and its system of social organization, global civilization, is erupting in an entropic black hole of time. The result of this black hole of time is an entropic harmonic convergence, where all of the conflicting points of the differing timing programs that the Gregorian calendar attempted to synchronize, come into a common point of hostile impact. Seen in this light, the Final War is really the final battle in a long saga that we can properly call the Time Wars.

THE AGE OF THE TIME WARS

The Time Wars began with the application of the error in time at Uruk: the 24-hour, 60-minute, 60-second timing frequency derived from a 12-part division of the circle, a two-dimensional plane in space. Since this measure does not correspond to the reality of time as the universal (13 :20) frequency of synchronization, it creates a mental dissynchronization that manifests as the battle of the mind against itself, and against nature. The dynam'ic of the "progress" of civilization is really the dynamic of dissynchronization inherited by the human mind. The gradual encroachment of civilization into all aspects of human society and the biosphere increases the incidence of dissynchronization. The resulting aphasic mental states lead to belligerence and various attitudes of dominance over others and nature-but never to any satisfaction.

Viewed by the Law of Time, civilization is the institutionalization and embodiment of the dissynchronous artificial timing frequency. The dynamic of dissynchronization results in increasingly artificial technological means and reliance on the artificial medium of money to replace the biological or organic functions of our nature and of the biosphere in general. Not only does this process tend toward an "improvement" in the accuracy of the means of war, but also in the capacity of the latter for greater mass destruction. Set apart from the natural cycles by the dissynchronous frequency of mechanization, the Time Wars are also waged against the biosphere, which is viewed as a hostile force to be overwhelmed and plundered. Everywhere on Earth, nature is on the run. Everywhere on Earth, the human looks at his watch-what will he do next? Catch another plane? Fire another bullet? Press another button and let loose another bomb?

The 56-year cycle of the technosphere witnessed the greatest creation and stockpiling of lethal weapons, more than in all the previous history put together, enough to destroy the Earth many times over. The 56-year cycle of the techno sphere also witnessed the greatest plundering thus far of the biosphere, more than in all the previous cycles of history put together. Continued at the current rate, the biospheric plunder would bring extinction to virtually all life within another generation. These two facets of self-destruction-technospheric and biospheric-point again to the Inevitable Event. It is the Pentagon that has led in military spending and strategizing, just as it was the activity within the World Trade Center Twin Towers that dominated and coordinated from afar the plunder of the biosphere-both precedents for the consequent construction and expansion of the technosphere.

But now time has run out on the technosphere. The world has gone flat and valueless with no future vision beyond war. All the money in the world will not put it back together, nor will all the bombs in the world salvage it against its enemies. Who are the enemies of the technosphere, and what would cause thereto be enemies against the technosphere? And who or what is there that speaks for the biosphere? For if the biosphere is to become the noosphere, it must come about from an understanding of the intelligence of harmony. Clearly, such understanding of harmonic intelligence has not been nor will it be forthcoming from the technosphere-or from the Gregorian calendar that programs it. Only the Law of Time can supply the harmony necessary to complete this great movement from biogeochemical combustion to noospheric mind. The Law of Time replaces all that is artificial, while confirming all that is true, whether in human history or the universe.

Let us return to the analysis of the Time Wars and the clash of timing programs. The Twin Towers were the absolute triumph of the artificial 12:60 timing frequency, whose program is embedded in the Gregorian calendar and paced by the mechanical clock. Who steered the hijacked planes into their Gregorian targets? Nineteen militant Muslims, we are told. Why nineteen? And what are the religious backgrounds of the twenty-two most-wanted terrorists in the world? They are twenty-two Muslims. Why that number, and why are all the terrorists Muslims? In the noospheric analysis, the Inevitable Event was the suicide of the technosphere unable to sustain two clashing time programs. It was a program of the strict Islamic twelve-month lunar calendar that drove its belief system into the Twin Towers and the Pentagon, the two supreme monuments of the program of the twelve unevenly measured months of the Gregorian pseudosolar calendar. Is it Islam that is the enemy of the technosphere? Is it Islam that is the champion of the biosphere? Indeed what is Islam? Why is Islam the most resistant force against globalization and the technosphere? We know where the Vatican, the bastion of Christendom, is and who ordained the Gregorian calendar, but what do we know about Islam?

If we strip history away from the force known as Islam (1.3 billion Muslims in the world), and reduce it to its most essential point, then Islam is the teaching of the Holy Quran. What is the Holy Quran? It is the last whole revealed text for all humanity, received and recorded over a 23-year period by a single human being, Muhammad the prophet (570-632). While militant Muslims and terrorists have dominated the mass media image ofIslam for the past few decades, what is the Quran and what does it actually teach? It is also a function of the Gregorian program that scarcely anyone in the West really knows or is familiar with the Quran, since the West has been at odds with Islam since the time of Muhammad. As the supreme text of monotheism, the Quran is actually a psychoactive book that upholds Islam as the final religion for humanity. "He is the one who sent His messenger with the guidance and the religion of truth, to make it prevail over all other religions. God suffices as a witness." (Quran, 48:28) *Islam* means "peace, which is submission to the will of God."

Islam is also the law of nature. Within the vast system of intricate, mathematically based logic that constitutes the Quran-which is a profoundly ahistorical text as well-one feature stands out: the description and definition of nature and the natural order as the chief proof of God's design and guiding wisdom. From the Quranic point of view, the biosphere, inclusive of the creation of the human, is the overwhelming single proof of God's wisdom in establishing a life-support system for our indwelling consciousness to evolve our bodies on this Earth. The biosphere is Islam, or "peace, which is submission to the will of God," understood as the divine law of nature.

> In the creation of the heavens and the earth, the alternation of night and day, the ships that roam the ocean for the benefit of the people, the water that God sends down from the sky to revive dead land and to spread in it all kinds of creatures, the manipulation of the winds, and the clouds that are placed between the sky and the earth, there are sufficient proofs for people who understand. (2:164)
>
> He created you from one person, then created from him a mate. He sent down to you eight kinds of livestock. He creates you in your mother's bellies, creation after creation, in trimesters of darkness. Such is God your Lord. To Him belongs all sovereignty. There is no other god beside Him. How could you deviate? (39:6)
>
> Among His proofs is that you see the land still, then as soon as we shower it with water, it vibrates with life. Surely the One who revived it can revive the dead. He is Omnipotent. (41:39)
>
> He created the human from a tiny drop, then he turns into an ardent opponent. And he created the livestock for you, to provide you with warmth and many other benefits as well as food. They also provide you with luxury during your leisure, and when you travel. And they carry your loads to lands that you could not reach without a great hardship. Surely your Lord is Compassionate, most Merciful. (16:4-7)
>
> He sends down from the sky water for your drink and to grow trees for your benefit. With it, He grows for you crops, olives, date palms, grapes, and all kinds of fruits. This is sufficient proof for people who think. (16: 10-11)
>
> And He committed the sea to serve you; you eat from it tender meat, and extract jewelry which you wear. And you see the ships roaming it for your commercial benefits, as you seek His bounties, that you may be appreciative. And He placed stabiliz-

> ers (mountains) on earth lest it tumbles with you, as well as rivers and roads that you may be guided. And landmarks, as well as the stars to be used in navigation. Is One who creates like one who does not create? Would you now take heed? (16:14-17)
>
> As for the Earth, we constructed it, and placed on it stabilizers (mountains) and we grew on it a perfect balance of everything. We made it habitable for you, and for creatures you do not provide for. There is nothing that we do not own infinite amounts thereof. But we send it down in precise measure. And we send the winds as pollinators, and cause water to come down from the sky for you to drink. Otherwise you could not keep it palatable. (15:19-22)2

Viewed from its own context, the Quran is the criterion for evaluating the spiritual progress of human existence on this Earth inclusive of man's ability to read the proofs of nature as manifestations of God's wisdom and to comport himself within this nature, the biospheric life-support system, in a manner consistent with the truth. From the Quranic perspective, human freedom of choice is at the root of the biospheric destruction. Choosing anything but to be submissive to God, the human is in rebellion against nature-both his own and that of the biosphere. "We have offered the responsibility [freedom of choice] to the heavens, and the earth, and the mountains, but they refused to bear it, and were afraid of it. But the human being accepted it; he was transgressing, ignorant." (33:72)

The result of this ignorance on behalf of his own aims and intentions rather than that of a respect for the divine plan (biosphere) of the Creator, accounts for the human situation today, at odds with and surrounded by an increasingly out-of-control biospheric degradation and corruption of natural time: "Disasters have spread throughout the land and sea, because of what the people have committed. He [God] thus lets them taste the consequences of some of their works, that they may return [to the right works]." (30:41) Thus, the current techno spheric destruction actually points toward a return of humans to the path of right works.

The forces that created the phenomenon of historical Islam are the enclosing of the Holy Quran within the rigid system of a twelve-month, non-circulating lunar calendar, compounded and supported by the development of the *hadith* and *sunna* (non-Quranic) traditions. A profoundly conservative social order, Islam is nonetheless activated and motivated by a holy scripture that enunciates an attitude toward nature that is the direct opposite of the 12:60 system that produced the technosphere. In this regard, the precepts of the Holy Quran-as of many another holy scripture-make it the "enemy" of the technosphere. From the noospheric perspective, historically and culturally, however, the assault on the nerve centers of the techno sphere could only have been effected by humans whose belief system was rooted in the Holy Quran, which had been at odds with the Gregorian civilization for centuries. In this way, the Quran was the noosphere's weapon used to bring about the circumstances that would lead to the resurrection of the biosphere-the noosphere itself. But only if man consciously chooses the right return path: synchronized harmony in time. The problem is not so much the rigid lunar calendar of Islam, but the irregular global standard of the Gregorian calendar, which, instead of standardizing all measures and perceptions of time, fosters a human social disharmony and an incessant antagonism against nature-and against Islam.

So then, how are we to return to the right works? Read and study the Quran, yes-but also make a fundamental change in the world. The solution of the Law of Time is: Get a new calendar! Humanity must take a long, hard look at its own timing systems and scrap what needs to be scrapped, no matter what kind of temporary inconveniences this may cause. Calendar reform is an issue that upsets many apple carts. Yet it is the only method that has not been commonly tried and the only step that has not yet been collectively taken. It is a characteristic of the dominant mindset that it writes history in the way that suits its own needs and supports its philosophy. In any of the reviews of the century carried by the mass media during the past year, you will find no mention of the World Calendar Reform Movement; nor will you find, for instance, any mention of the Roerich Peace Pact (1935); nor of the Harmonic Convergence-all buried by the marketing of history to suit the needs of globalization. What do these three phenomena have in common? A concern for peace based on an understanding that without harmony, the human cultural situation will only worsen. If humankind suffers from an excess of disharmony, in the noospheric analysis the only solution is a harmonic standard of time to synchronize all humanity simultaneously-the Law of Time refers to this as the Thirteen *Moon!* 28-Day frequency shift.

CHANGE YOUR CALENDAR, CHANGE YOUR MIND

> *"Since the advent of the Nuclear Age, everything has changed but the way people think, thus, we drift towards unparalleled catastrophe."*
>
> -ALBERT EINSTEIN[3]

To change how people think you must take a profoundly simple and universal element of everyday life, one in which all everyday thinking is rooted, and change it so radically that the way in which people think will itself be dramatically changed. If you want to know that something has really changed, change the calendar! This is the meaning of the replacement of the Gregorian calendar by the Thirteen Moon calendar. The calendar change is necessary because, as the pragmatic application of the discovery of the Law of Time, it brings into focus the essence of this discovery: Time is a frequency, the frequency of synchronization. If a calendar does not increase synchronization, it is not performing its highest function. This is the fundamental critique of the Gregorian calendar-and of all concepts of time based solely on physical, third-dimensional astronomical measurements. This discovery is so new and startling that it affects all human thought and is a matter to which all belief systems, religions, and methods of science must be cognizant. Yet nothing, it seems, is more difficult for humanity than to make this change from a manifest disharmony to a manifest harmony.

Historical man recoils at the prospect of actual harmony-harmony in natural time. This is because historical man defines his existence and thrives upon the disharmonies of his own invented time. The Gregorian calendar keeps the human mind entrained in a diabolical disorder of meaninglessly named months of uneven measure, while every day is tracked and driven by a relentless mechanism called a clock. This immersion of the mind of historical man in the frequency of his own artificial time, more than any other factor, condemns him to a world of inescapable horrors: traffic jams, poverty, terrorism, unresolvable historical and territorial disputes, global warming, environmental degradation, social disorder, insanity, and drug abuse. The problem of historical *human-Homo historicus-is* compounded by the unconscious nature of his acceptance of this timing sensibility, believing it to be the actual nature of time. Hence, all his public and economic policies are driven by and determined by the relentlessness of this perception that time is an arrow blindly pulling him into a future in which his only defense is the creation of more technology, the technosphere itself being a pure expression of the application of the mechanization of time.

THE SECRET HISTORY OF CALENDAR REFORM

When you went to bed on October 5,1582 you woke up on October 16. In this way Pope Gregory XIII promulgated his reform of the Julian calendar, from which the Gregorian calendar differs not one whit, except for the precise working out of the formula for leap year. When this "reform" occurred, the time sensibility of historical man was already crippled and disabled. The Gregorian calendar, by its moment of historicity, bound that crippled time disorder into its final institutional form, assuring that not only European man, but eventually all cultures and races dominated by European thought and conventions would assume and accept this deformed and deforming standard of time.

The mechanical clock, though evolved over many centuries, was perfected just after the Gregorian Calendar Reform of 1582 and promoted by Europeans as an example of the human triumph over nature. So it was that the human mind assumed as second nature its own invented time, encoded as a frequency that we have identified as the 12:60, an unconscious mental ratio determined by the irregular twelve-month calendar and artificial sixty-minute hour. The combination of these two factors in this one timing frequency have consigned historical man to a hopelessly debilitating mental disorder that is referred to as modern civilization, or nowadays, postmodern. But postmodern is not yet post-historical. Postmodernism is just the final convulsion of the time disorder that afflicts the human race as a virulent mental disease and encapsulates man in the technospheric cocoon of his own making. The only cure for such a profound and pervasive disorder is absolute harmony, the very harmony from which historical man so recoils.

In this light, it is well to know that through the first half of the twentieth century a vigorous and well-organized calendar reform movement flourished. Such was the character of this movement to reform the Gregorian calendar that in 1923 the League of Nations called for proposals for reform. More than 500 proposals were received; by 1931, these were narrowed to three:

> The International Calendar Organization, represented by Mr. Broughton Richmond, promoted a perpetual calendar based on the principle of five-five 73-day cycles = 365 days-and the year divided into twelve months of five 6-day weeks each plus a 5-day cycle [The Thirteen Moon calendar also includes the program of seventy-three 5-day periods-overtone chromatics-integrated with the fifty-two 7-day weeks: 52 x 7(+1) = 5 x 73.]
>
> The International Fixed Calendar League, represented by Mr. Moses Cotsworth, also supported by Eastman Kodak and the International Chamber of Commerce, promoted a perpetual Thirteen Month/28-day calendar, which was first presented by Auguste Comte (1842-1849), inclusive of the "day out of time." Though it was not backed by a more thorough mathematical science, which has now been supplied by the discovery of the Law of Time, this was by far the favored calendar, as it is the one that possesses a perfect structural harmony and obvious regularity of all its parts.

3. The World Calendar Association, represented by Elizabeth Acheles, promoted the World Calendar, a modified but perpetual version of the twelve-month Gregorian calendar, which also included the principle of a "day out of time" ("null day"). This group turned out to be the longest-lived and the most vocal, and was responsible for the publication of *The Journal Of Calendar Reform (1931-32).*4

It is instructive to read some of the early texts on the topic of calendar reform to see how the same issue of harmony and order brought up by the Law of Time was at the root of the call for the establishment of a perpetual, harmonic calendar. The base argument, as presented in Alexander Philip's *The Reform of the Calendar* (1914), is stated as follows:

> Suppose, for example, that by some strange convention the meaning of the figures we employ in numerical notation were to change every year; suppose the figure which this year represents 2 were next year to mean 3, next year 4, and so on; suppose again that our weights and measures were to fluctuate in a similar manner. . . then we affirm, without fear of contradiction, that the whole fabric of science and the mechanical arts could never have been raised at all. . . we should in such circumstances have been compelled to rest content today with the very simplest and most primitive appliances. Yet, strange as it may sound, such are the conditions under which, in modem society, human action is organized. For what is the framework. . . by which we arrange our actions? It is no other than the scheme under which we arrange our time-in one word, our calendar. . . [thus] the disorganized state of all social arrangements is ascribable to the calendar. . . The dislocation of our calendrical arrangements is due to two distinct causes. . . the incongruity of the week; [and] the irregularity of the lengths of the months. . .5

The argument made is a profound one, and was the motivating basis of the calendar reform movement in general, no matter which calendar was being promoted. It is fair to ask: Why do we expect a rigorous standard of uniformity in our measurements of space (size, weight, volume, etc.), but spurn or ignore the same requirement of uniformity in our measurement of time, especially the time that governs our everyday social arrangements and consciousness? How can we think that this would not have a profound underlying effect on the very nature and conduct of our society?

Most interestingly, as a movement calendar reform arose strongly during the proto-technosphere, between the two World Wars, only to be swept away by the clouds of war and the contradictory factions of disharmony in which modern civilization was already immersed. Any effort at altering or changing the direction and motion of modern civilization is treated as a threat or derided as ineffectual activity that is scientifically unprovable or that will cost humanity too many dollars to make the change. Such was the fate of the first great calendar reform movement that arose with such fervor in the late nineteenth century only to find its ignominious end when the final debate on calendar reform was adjourned indefinitely by the United Nations Social and Economics Committee in 1956. No one reads about the calendar reform movement in the history books. Because debate on the matter was effectively adjourned, it is assumed that calendar reform is a dead issue of no relevance to the development of postatomic, technospheric man. On the contrary, seen in the dim light of the still-smoking ruin of the Twin Towers, the calendar reform arguments concerning humanity's irrational behavior and its relation to the "immorality" of a skewed timing measure, the calendar, as well as to an inhuman timing device, the clock, were never more relevant.

Not the least import of calendar reform is the opportunity to create a new basis for unifying humanity. As was stated in the "Resolution Adopted by the Fourth General Conference on Communications and Transit of the League of Nations," which took up the issue of calendar reform: "A great number of delegations expressed the opinion that any reform of the calendar could only be put into practice if it came into force simultaneously throughout the world, or at least in a very great majority of States, and it was for this reason that the study of this question had been placed under the auspices of the League of Nations."6

The tragedy of the human inability to agree on a matter so obvious to logic and reason is a testament to the already crippling effects of having operated so long under parochial, sectarian, and often disharmonic timing standards. While the International Chamber of Commerce, the League of Nations, and various international scientific bodies supported the first effort at reform in 1914-1956, the second movement, 1989-present, arose as a purely populist effort, but with an entirely different information base. Recognizing instinctively that the calendar reform issue is more than merely a matter of mathematical order in relation to natural cycles, it is a profoundly theological issue as well, the second reform movement also developed from a non-Western inspiration: the Mayan calendar.

While the psychology of the dominator civilization all but blurs out any other considerations of calendar besides the prevailing Gregorian one, an objective study of the matter demonstrates that the latter is a primitive instrument that eschews both natural cyclic order and any mathematical harmony. In place of the anachronistic standard of the Gregorian calendar, the second calendar reform movement raised a new criterion, the Law of Time, a whole system principle that integrates the mathematical and the moral, the theological and the psychological, and the economic and scientific issues that are embedded **in** the entire topic of calendar reform. Precisely for the reasons of historical man's unconscious habituation to his own disharmonies and the accelerating and cumulative effect of this ignorance on the disorder of his own social life as well as the natural world around him, the second calendar reform movement has developed with a much greater moral and prophetic urgency.

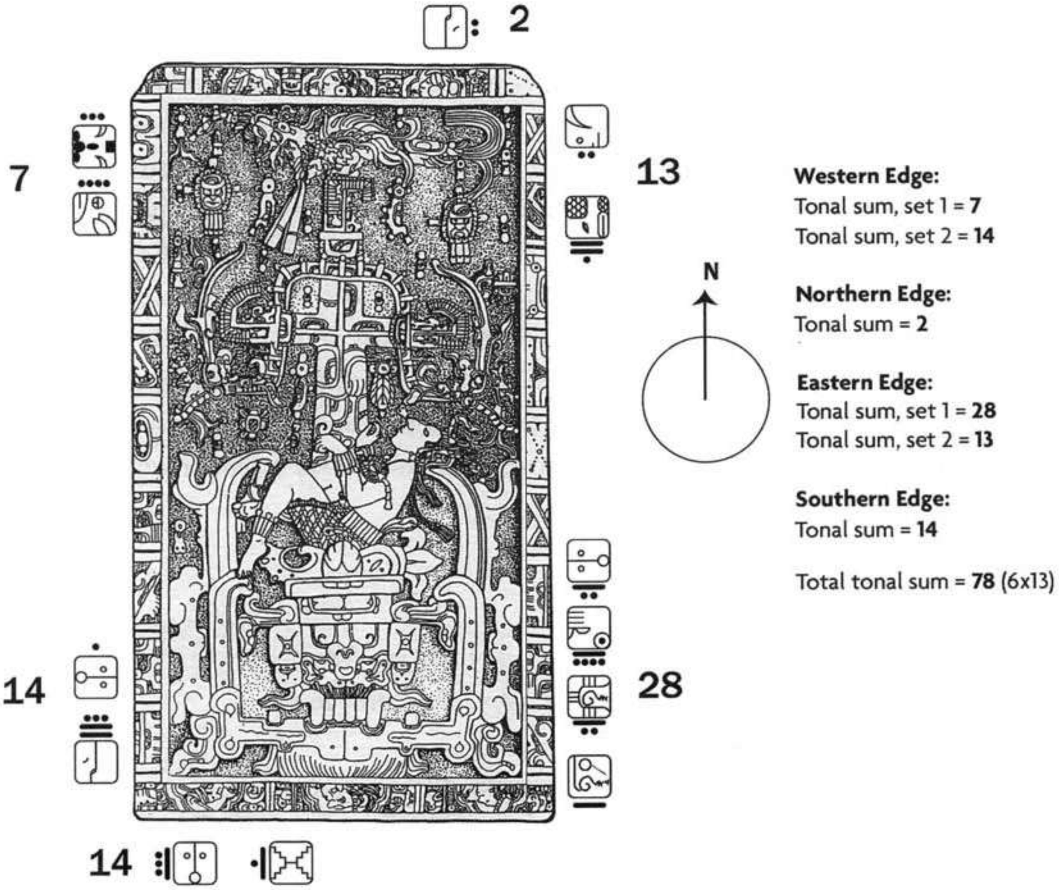

Western Edge:
Tonal sum, set 1 = **7**
Tonal sum, set 2 = **14**

Northern Edge:
Tonal sum = **2**

Eastern Edge:
Tonal sum, set 1 = **28**
Tonal sum, set 2 = **13**

Southern Edge:
Tonal sum = **14**

Total tonal sum = **78** (6x13)

Sarcophagus Lid of the Tomb of Pacal Votan
Showing Disposition of Thirteen Clear Signs

The tonal sums (dot-bar code) of the thirteen signs are key to affirming the prophecy of the Thirteen Moon calendar. Note that the sums of the two sets on the eastern edge equal 28 and 13, while opposite the 13 set is the perfect 7. The tonal sums of the other two sets on the western and southern edge both equal 14, for a sum of 28. The northern edge has a single sign whose tonal sum is 2 twice 7 = 14, and twice 14 = 28; twice 13 = 26 fractal of 260 and code to the Gregorian date 7/26; while twice 28 = 56, the number of years from Hiroshima to the Inevitable Event.

"Consciously using the tool of the Thirteen Moon calendar to replace the erroneous measure of the artificial twelve-month calendar is, in itself, an unprecedented act of human self-reflective consciousness. Collective adaptation to the Thirteen Moon calendar is an intrinsic heightening of consciousness which rapidly awakens the telepathic programming inherent in human biology, frustrated for so long for not having been allowed development in the correct timing frequency."7

It was in 1987 that *The Mayan Factor: Path Beyond Technology* first brought to major public and international attention the possibility that a superior timing sensibility was developed by a thoroughly non-Western culture, the Maya. This book was published with scarcely any knowledge of the calendar reform movement that had gone into a debilitating coma in 1956. Yet, by pursuing the mathematical base of the Mayan calendar, the essence of the Mayan achievement, the Law of Time was extracted, a discovery that brought on the second calendar reform movement, the World Thirteen Moon Calendar Change Peace Movement. Prompted by a prophetic revelation concerning the Thirteen Moon!28-Day calendar decoded from the tomb of Pacal Votan, the premise of the second calendar reform movement was tested worldwide from 1993 to 2000, a time defined as the "seven years of prophecy." By the year 2001, the official beginning of the third Christian millennium, a loose, far-flung calendar reform movement, pioneering in the development of a new science of time and consciousness was promoting the Thirteen Moon!28-Day calendar worldwide.

A REVOLUTION IN TIME

> *"Fire in the lake, the image of revolution. Thus, the superior person sets the calendar in order and makes the seasons clear"*
>
> -*I CHING,* HEXAGRAM49, "REVOLUTION"8

By common consent, a calendar is a system for reckoning time. Virtually all current time-reckoning systems take account only of chronological or, better put, astronomical time-the movement of the Earth in relation to celestial bodies, the sun, moon, planets, stars, and constellations. This is a purely third-dimensional, physical consideration of time. The Law of Time affirms a higher order of time-the synchronic order. This is the time of the fourth dimension, which includes the chronological but enfolds it in a higher mental and mathematical order of reality. According to the Law of Time, what you don't know about the calendar you are using could kill you. By operating according to purely third-dimensional chrono-astronomical timing standards and without knowledge of the synchronic order of fourth-dimensional time, humanity subjects itself to a one-sided view of reality that is hazardous to the planet and jeopardizes the future existence of the human species as well as all life on Earth.

To correct this situation a radical solution is required-a new calendar embedded in the synchronic order revealed by the discovery of the Law of Time. This is the significance of the Thirteen Moon/28-Day calendar-it epitomizes the incorporation of the chrono-astronomical order of time into the synchronic order, and hence provides the vehicle for humanity to escape its otherwise certain plunge into disaster within the next decade. When a dogma is encountered while one is engaged in the pursuit of truth, it must be challenged and abolished. Such a dogma is the calendar in use as the world standard, the Gregorian calendar. That this calendar, originally a system of thought peculiar to one people or religion, should dominate all the peoples and even life of the planet makes it subject to the critique of planetary anthropology.

The discovery that time is a frequency mathematically expressed as the ratio 13:20 requires a profound reorganization of all thought about time and the consequent reorganization of human society as a planetary organism. As the unified field theory of time, the Law of Time calls for the human synchronization in time predicated on the establishment of the Thirteen Moon calendar as the global civil standard replacing the current erroneous timing standard, the Gregorian calendar. This is a matter of the utmost seriousness, since it is also recognized that all belief systems are the function of programs locked up or embedded in the calendars or time-reckoning systems used by a people. The fate of the species and the biosphere may well depend on whether or not humanity is able to understand this point.

What is it about the calendars in use that condemns their users to endless conflict and inability to unify as a species? What is at the root of their disharmony? It is the pursuit of astronomical time. All people today are bound by calendars based on a pursuit of accuracy of astronomical time-the length of the year being the object of this pursuit. Astronomical time is a losing game of an ever-changing and elusive order that results in an entropic reductionism that has little to do with consciousness or synchronization. The year's length (measure of the orbit of the Earth around the sun) calculated by Pope Gregory XIII at 365 days, 5 hours, 48 minutes, and 20 seconds, slows at the rate of *112* second per century, while the length of time in the Gregorian calendar is off from the "true" solar year by 25.9678 seconds per year! Obsessed with rectifying this discrepancy, in 1972 reductionist science came up with Atomic Time to replace Earth Time, and the atomic cesium clock replaced the measure of the year as 365.241299 days with 290,091,200,500,000,000 oscillations of Cs per year!

But what does this staggeringly astronomical figure mean, and of what value is it?

In profound contrast to this obsessive reductionism in which there is no place for spirit, mind, or consciousness, fourth-dimensional radial mathematics represents a mental order greater than and encompassing the astronomical order of the third-dimensional universe. According to the Law of Time, in time reckoning, consideration should be given to systems of the harmonic order of synchronic time that proceed from the radial mathematical order of the higher mind. The higher mind represents the evolved state of mind-the noosphere-that evolution on Earth tends toward, but that has not yet been attained. As long as the human concept of time is governed by the obsessive pursuit of the impossible to attain perfection of astronomical time, this pursuit and the civilization that it propels are doomed to failure. What is known as history is embedded in this pursuit of "fixing" astronomical time, which has culminated in the dogma of the Gregorian calendar and the microatomic measure of the cesium clock. When did this pursuit begin, why did it come to dominate the human species, and to what effect?

As we have seen, the values, customs, and norms of a nation, culture, or civilization are embedded in the calendar it uses. The Gregorian calendar, the current global civil standard-indeed, the foundation in time of globalization-is the calendar of the Caesars, of the Roman Empire, of the Catholic Church and the Vatican. To change this calendar is to change history itself. Can it be done? The League ofNations attempted this in the 1930s but failed. The United Nations tabled this effort in 1956. Since the United Nations adjourned the matter of calendar reform indefinitely, in 45 years the Earth's population has more than doubled, that is, it has increased by more than 3.2 billion people, a factor that is still out of control, and that may have much to do with the fact that humanity is actually in a fog about time.

The reasons for wanting to replace the Gregorian calendar are no less valid today. The Gregorian calendar is not a standard of measure. It lacks logic and reason. It numbs and befuddles the mind when trying to make calculations by it. It is shrouded in an arcane and medieval obscurantism. The net effect of the use of this calendar is to perpetuate a fundamental level of mental confusion and ignorance concerning the actual nature of time itself-an ignorance that is hardened into dogma by the unwillingness of habit to consider any other possibility, but rather to accept the entire system as second nature. The argument that it encodes the most accurate and scientific measure of the solar orbit of the Earth is irrelevant in relation to the effect of its irregular measure on the mind and its numbing power as the dogma of a conquering people. Indeed, the solar measure of365.241299 days per year has nothing intrinsically to do with "30 days hath September. . . ," and vice versa. In fact, the measure of the Earth's orbit is one matter; the purpose of calendars and time reckoning as factors of synchronization is a wholly other matter. But then, since the very nature of a culture or a civilization is determined by the calendar it keeps, the use of a deformed standard can only inhibit and skew even the questions one asks about time.

The failure of calendar reform by 1961, the year of the Vatican II Council, only promoted an increasingly complacent acceptance, however cleverly disguised, of a millennial dogma, the Gregorian calendar. In the conclusion to the "last word" on the subject, *Calendar: Humanity's Epic Struggle to Determine a True and Accurate Year,* David Duncan summarizes, "We take in stride a calendar [the Gregorian] that is flawed but endures, largely because it works just fine for most of us, and it is what we are used to."9 It is not at all objective to accept something because we are used to it. This attitude is anything but scientific. Rather, it is the smug basis of nothing more than a self-fulfilling dogma-and yet the civilization that bases itself on this aberrant timing measure prides itself as being the most scientific in all history!

THE CONFUSION OF IRREGULAR TIME

Due to this flawed time sensibility, it may be concluded that humankind today is little capable of moral reason or logic, which itself is a result of, or at least strongly reinforced by, habituation to a timing standard that is irrational and illogical. Anytime we overlook or dismiss a fault due to habit, or simply because we are accustomed to the use, say, of an instrument, even if that instrument is demonstrated to be flawed and irregular, is evidence of a moral laziness that will eventually contaminate the entire consciousness that accepts and accommodates itself to the flaw. The exercise of truth is a moral prerogative. The fact of the matter is that the Gregorian calendar is a hodgepodge of flaws and accretions, rationalized into a systematic formulization that has no basis in reality. The celebration of Easter is a case in point.

Easter is the celebration of the day Christ was resurrected from the dead. It may be logically assumed that this occurred on one and only one particular day, just as the celebration of his birth occurs on Christmas Day (although, similarly, no one really knows that this was the actual date of his birth). By the time the Christian religion had become an established force within the Roman Empire, however, no one knew precisely on what day the Resurrection had taken place. The Council of Nicea, convened by Constantine in A.D. 325, called for a debate on the matter in which various theories were presented. The debate was concluded with the victory of one conjectural system over the other. The winning system stated that Easter would be celebrated on the first Sunday after the first full moon after the spring equinox, except when the full moon was on a Sunday, in which case it would be celebrated on the following Sunday, and this was so that it would not be celebrated on the same day as Jewish Passover.

Over the centuries this conjectural theory regarding the date of Christ's resurrection was buttressed with a system of elaborate calculations for predicting when Easter would occur. These calculations were all made with regard to establishing an alignment of the synodic lunar cycle of 29.5306 days (which resulted in a lunar year of 354.3672 days) and the solar year of 365.241299 days! Being fortified with such elaborate mathematical calculations did not lessen the fact that the system was originally-and still is-a mere speculation of a highly clever but relative nature. Thoroughly embedded in the Julian-Gregorian calendar, the system of determining Easter is inseparable from the Gregorian calendar itself, and is one of the chief reasons for the resistance to its reform-as witnessed by a reading of the calendar reform arguments of the 1930s. The entire system of this calendar, inherited without question from the Caesars of the Roman empire, is actually nothing more than a form of self-perpetuation of a priestly hierarchy that includes in its support a host of bankers, industrialists, and scientists and the elaborate social system that is under their control-the technosphere of global civilization itself. Take away the calendar and the entire structure loses its foundation.

"If the Julian/Gregorian calendar was to be offered as a new device for measuring time, we, with our present knowledge and state of living, would reject it as something utterly impracticable, lacking in harmony and order, unbalanced and irregular, too clumsy a calendar to make calculations by, while the various sections are not comparable."10

A science or any scientist that unquestioningly follows the Gregorian calendar is, in reality, not worthy of the name. What is science? A concern for logic and precision of measure, we might answer, as well as standards of measure that utilize uniform units of measure in accord with what is being measured. Yes, the year is calculated as being 365.241299 days in length, but if the annual standard of measure that is used is irregular and unscientific, then it avails nothing, and indeed deforms the mind leading it down byways that may only end in self-destruction. Thus the pursuit of a true and accurate year in itself may be an illusory pursuit, blinding us to the actual nature of time, and leading us away from a genuine understanding of ourselves and our role and purpose on this Earth.

To change and replace the calendar with the thirteen moon standard is to return us to our original purpose, leading us back to pathways of harmony and natural health. According to the Law of Time, the damage already inflicted in the time humanity last hesitated and lost the chance to alter its timing frequency, forty years ago, has been incalculable. The final opportunity to change the calendar and the timing frequency is now upon us. Because of this we must be very clear and unshakable in our understanding and determination to expose and eliminate the current civil calendar for once and for all. Even apologist David Duncan's unabashedly pro-Gregorian book *Calendar* (1999) concludes with its time line on the date 2012, "Current Mayan Great Cycle will end."11 But will we also end? By 2012, it will be too late to change. We must begin to change now.

In reviewing the quirks and twists of history that constitute the Gregorian calendar, we must ask: Why do we continue to use such an instrument, and what are its effects? Is a calendar something more than a tool for scheduling the payment of debts *(calends),* or is it an instrument of synchronization? The harmony or disharmony of time is a profound matter rooted in the instrument of time reckoning that we use. There can be little question that we live more in a time of chaos than of harmony. We may also say in regard to the effect of standards of measure on the mind, that the chaos of the time is embedded in the calendar we use. If we are to leave this time of chaos and enter a time of harmony, then we must exchange the instrument in which the chaos is embedded for an instrument that is the very model of harmony-the Thirteen Moon/28-Day count. This is the choice humankind must now make.

The issue of the nature of time, or of the times in which we are living, cannot be separated from the issue of the calendar. The very nature of the calendar that the world follows has stunted the mind and warped the body's innate timing sensibility, reducing the understanding of time to being merely a matter of chronology. "Chronology is a register or reckoning of successive years, a time scale. It may also be described as a system of registering time massively."12 Virtually all time reckoning has proceeded from astronomical indicators, the major exception being the Maya, who actuated from a solely mathematical outlook. Yet with this mathematical outlook the Maya devised a chronological method that has never been surpassed, a fact attested to by many, including the architects of the Kitt Peak Observatory in Tucson, Arizona, where a mosaic mural depicting the civilization of the Maya asserts the superiority of their calendar over that of the Europeans.

It is significant that the earlier calendar reform movement, while occasionally acknowledging the brilliance of the Maya, made no use of the Mayan knowledge. But this was because the time sensibility was restricted to thinking of time solely in terms of chronology, or at best as a schedule for keeping good accounts, with little comprehension at all of the synchronic nature or order of time. And since the system of chronology was synonymous with the system of control embedded in the Gregorian calendar system, the earlier reform movement failed altogether to make any impact on the Vatican, much less on the global power structure itself.

All the thought, action, and good intention put into the years of calendar reform notwithstanding, the Gregorian calendar prevailed, and, tragically, not one of the proposals or recommendations was ever even considered for adoption by the papal authorities who maintain control over the calendar. In the 1930s, the propaganda of the Church was aimed at limiting any possibility of reform, for it would not give up the succession of the seven-day week. In any perpetual calendar, such as the Thirteen Moon/28-Day calendar, the observance of the Day Out of Time is necessary for maintaining the harmony of the fifty-two weeks that complete themselves in 364 days. The continuous succession of the week does not stand in a constant relation with the other elements of the calendar and is one of the main causes of confusion in the Gregorian calendar. Arguing that such a break as a Day Out of Time, no day of the week at all, would plunge the world more deeply into chaos, barbarism, and war, the Vatican effectively blocked all effort at reform. The completion of the century of total war, capped by the Inevitable Event, roundly refutes this point. It may just as easily be argued that because we did *not* change the calendar, we are now plunged into the final barbarism and unholy war.

When the Vatican II Council in 1961 affirmed in its backward and highly restrictive language that it was not opposed to reform, so long as such reform did not disrupt the seven-day week and respected the tradition of Easter as a "movable feast," the door to reform slammed shut. By this time all the early reformists had literally died out and the question of calendar reform appeared to have become one of the buried issues of history.

What the priests of the Catholic Church thought they had buried through an auto-da-fe in Izamal, Yucatan, 1562, and overcome through the imposition of the Julian-Gregorian calendar over the conquered Maya, returned with the precision of prophetic timing in 1987. The publication of *The Mayan Factor* not only opened the door to a new look at the Maya, but also to an understanding of time that was anything but chronological. A new dimension of time appeared-radial, fractal time, the synchronic order. And behind the reassessment of the nature of time was the provocative call of Mayan prophecy-the end of the thirteen baktun Long Count, 2012.

The Law of Time defines the human problem as being one of radical and destructive dissynchronization from natural time, resolvable only by a return to natural

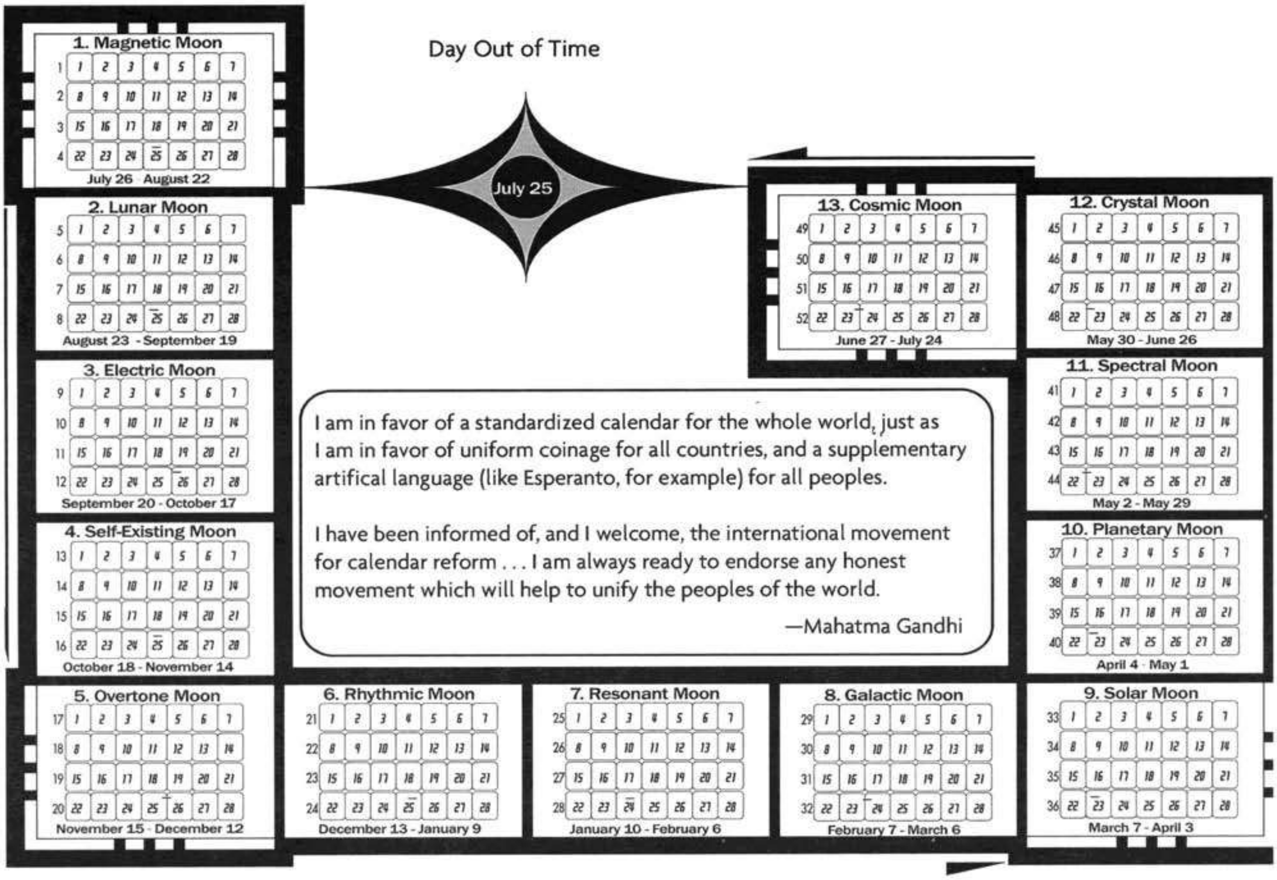
Day Out of Time
July 25
1. Magnetic Moon
July 26 - August 22
2. Lunar Moon
August 23 - September 19
3. Electric Moon
September 20 - October 17
4. Self-Existing Moon
October 18 - November 14
5. Overtone Moon
November 15 - December 12
6. Rhythmic Moon
December 13 - January 9
7. Resonant Moon
January 10 - February 6
8. Galactic Moon
February 7 - March 6
9. Solar Moon
March 7 - April 3
10. Planetary Moon
April 4 - May 1
11. Spectral Moon
May 2 - May 29
12. Crystal Moon
May 30 - June 26
13. Cosmic Moon
June 27 - July 24

> I am in favor of a standardized calendar for the whole world, just as I am in favor of uniform coinage for all countries, and a supplementary artifical language (like Esperanto, for example) for all peoples.
>
> I have been informed of, and I welcome, the international movement for calendar reform . . . I am always ready to endorse any honest movement which will help to unify the peoples of the world.
>
> —Mahatma Gandhi

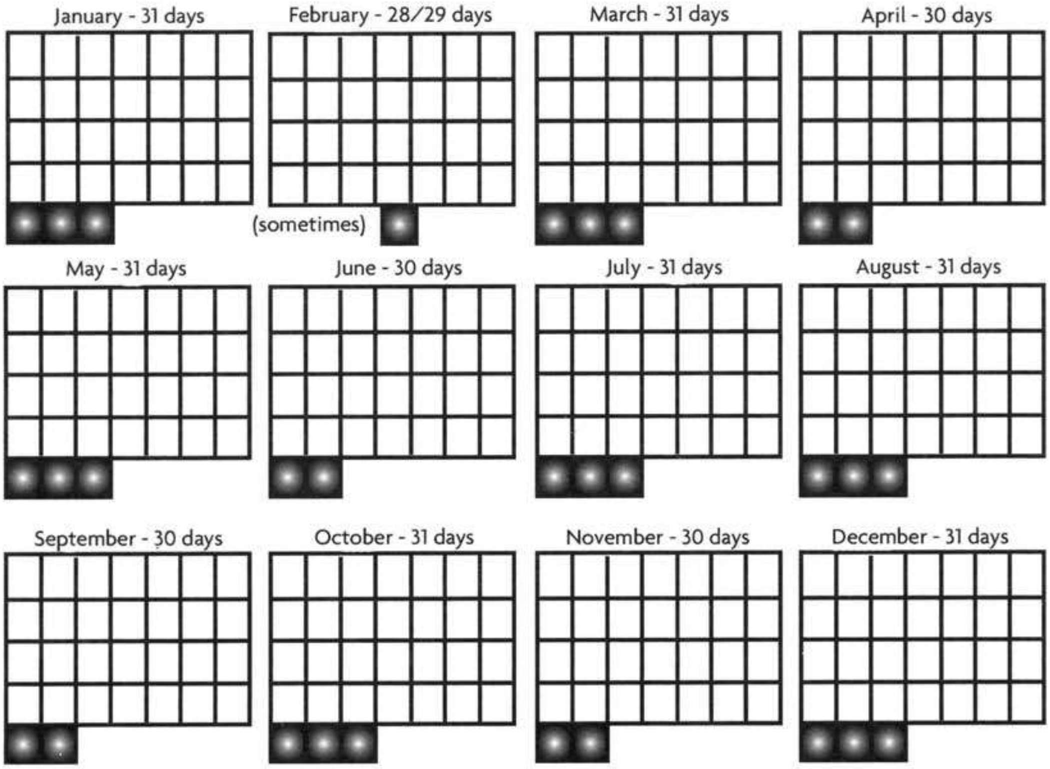
January - 31 days
February - 28/29 days
(sometimes)
March - 31 days
April - 30 days
May - 31 days
June - 30 days
July - 31 days
August - 31 days
September - 30 days
October - 31 days
November - 30 days
December - 31 days

Wavespell of the Thirteen Moons in Relation to the Irregular Gregorian Calendar

time through the perfect and perpetual harmony of the Thirteen Moon calendar. With the analysis of the Law of Time, there is an even greater urgency and a more profound certainty. Reform is not only insufficient but also impossible. If the calendar is not changed within the next decade, it will be the end of civilization as we know it. There is a genuine and legitimate need to bring the issue of calendar reform into the public eye once again.

Yes, to change the calendar involves a frequency shift. The collapse of the technosphere will dispose the human mind to rethink its options. It may even reorient the human mind to a consideration of the biosphere as being the greater vehicle in which we, as humans, are being carried. Such a massive change in attitude, a mind shift and a paradigm shift, may arouse in the human breast a spiritual reawakening as well. Harmony is spirituality. By changing the calendar, humanity can know it has turned a page of destiny-an act that may cause it to appreciate what has already been written in a book it had not been inclined to study, the Holy Quran:

> God is the One who raised the heavens without pillars that you can see, then assumed all authority. He committed the sun and the moon, each running in its orbit for a predetermined period. He controls all things, and explains the revelations that you may attain certainty about meeting your Lord. He is the One who constructed the earth and placed on it mountains and rivers. And from the different kinds of fruits, He made them into pairs-males and females. The night overtakes the day. These are solid proofs for people who think. (13:2-3)

In this passage we can find ample evidence of the formal order of the Law of Time and the biosphere-solid proofs. As for the "people who think," do they not then constitute the noosphere?

◆

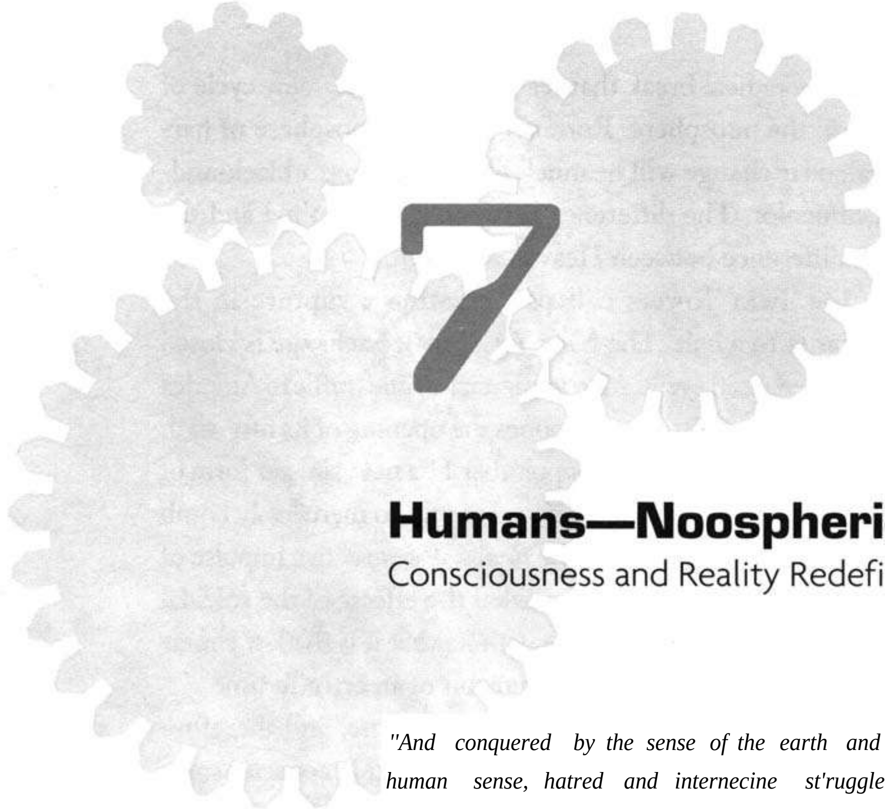

7

Humans—Noospheric Chips:

Consciousness and Reality Redefined

> *"And conquered by the sense of the earth and human sense, hatred and internecine struggles will have disappeared in the ever-warmer radiance of Omega. Some sort of unanimity will reign over the entire mass of the noosphere. The final convergence will take place in peace."*
>
> -PIERRE TEILHARD DE CHARDIN,
> *THE PHENOMENON OF MAN!*

Once begun, the exhaustion of the technosphere is irreversible. The exponential growth curves of money, machine production, energy output, and population increase come to a climax. By 2004 the humans will be at another major crossroads: to go wholeheartedly into the harmony of the noosphere, or to suffer the rest of their decline in an ignominious twilight of terrorism, barbarism, and cultural anarchy. By 2004 the humans will have their last opportunity to change their macro-organizing principle. By relinquishing the artificial time held in place by the old calendar and becoming unified in time under a calendar that is a perfect harmony, humans will enact the coming of the noosphere. Only harmony will bring harmony; only harmony will cancel the errors of history, a manifest disharmony. Only an act of harmony can manifest the noosphere. Only this time-changing act, consciously carried forward by a significantly roused number of humans, will create a mental field so positive and constructive as to engender the radical break that leads to an entirely new cycle of existence of the life on Earth: the noosphere. Entrance into the noosphere of harmonic time through the calendar change will be much like going from a black-and-white movie to one in Technicolor. The difference between the new time and the old time is literally like the difference between Heaven and Hell.

September 11, 2001: the Twin Towers collapse, creating a rupture in the technosphere. Everything comes to a halt. The New York Stock Exchange is closed for four days-a virtually unprecedented event. All commercial airline traffic in America is halted for two days. Even the United Nations postpones the opening of its fifty-sixth General Assembly by three weeks. Six weeks after September 11 a new, sinister form of war is engulfing the technosphere. The Americans assert the right to mercilessly bomb Mghanistan in self-defense. The response of war is the final self-destructive impulse of the techno sphere to protect itself, but this will only widen the effects of the suicidal blow it has already sustained. The war will only make definite that it is the last climax of the biogeochemical combustion engendered by the pursuit of an error in time.

The rupture in the technosphere is the break in artificial time, and therefore amounts to the opening of a noospheric fissure-an inrush of cosmic consciousness and a moment of simultaneously heightened natural time. When the towers collapsed, belief systems shattered. Because it was viewed on global television, it was an event witnessed by all humanity and it entered through the optical cortex into the brain, arousing potent images from the subconscious. This is a noospheric act because it is the entire mind of the whole species that receives this single image in one simultaneous moment. In that shattering image, however unconscious, the noosphere immediately entered the field of human consciousness, much as a leak in a boat allows water to come rushing in. From the perspective of the noosphere, the technosphere is a medium of transformation resulting from the pursuit of a distorted perception of time. With the destruction of the Twin Towers, the transformation is complete; following the war, the technosphere can be creatively dismantled.

THE NOOSPHERE: RAINBOW BRAIN OF TIMESHIP EARTH

What is the noosphere? Can we define the noosphere more precisely so that it is a less nebulous term? Can we make the noosphere concrete enough so that it may be understood and accepted as the coming new way of a higher level of everyday normalcy? First of all, let us say that the noosphere has existed as long as life has existed on this Earth. In this sense, the noosphere is the sum of the mental interactions of all life. The mental interactions, or field of mind, exist as a great unconscious medium, inseparable from the Law of Time. That is, if all life and the phenomenal world of the biospheric medium are synchronized by the Law of Time, then the programs of the organization oflife, from the periodic tables to the DNA code, will also be synchronized by and embedded in the mathematics of the Law of Time. Being a function of the whole system of the Earth and integral to the functioning of the biosphere as an interplay of cycles in which organic and inorganic interact with a set order of chemical transmutations, the noosphere must also be the mental coordinating field or program of this complex of vital interactions. It must, therefore, be a characteristic of biospheric evolution that the noosphere remains unconscious until a significant moment is reached-the climax of the cycle of biogeochemical combustion, the enactment of the biosphere-noosphere transition.

As the Earth's mental field, it must be an act of collective self-reflection that arouses the noosphere into functioning consciousness. By its definition as the mental envelope of the Earth, the noosphere can come into conscious manifestation only through an act of collective self-reflection, such as the proposed calendar change. Also, as the mental envelope of the Earth that is above and discontinuous with the biosphere, the noosphere-mind sphere--can mean only one thing: the advent of universal telepathy among the humans. The Internet can be seen as a third-dimensional reflection of the noosphere, a form of electronic proto-telepathy. But, ironically, as long as the humans hold on to the Internet, they cannot experience the true evolutionary potential of genuine universal telepathy. The point is that telepathy can only function in a field of reality that is synchronized by the natural 13:20 timing frequency. Time and telepathy are mutually defining factors dependent on unity of thought. The humans have chosen a path of disunity, so that their calendar programs clash and are at odds with each other and with natural time. This maintains the technosphere, destroys the biosphere, and keeps the noosphere in the unconscious. But now, with the Inevitable Event, the break has occurred and the shattering of the Dreamspell of History has begun. The noosphere's rainbow brain leaks back into human consciousness.

The rainbow brain of the noosphere-what is that? As the mental envelope of the Earth, the noosphere possesses a structure ordained by the Law of Time and congruent with the structure of the whole Earth. Within the noosphere are to be found the evolutionary control panel or programs, defined as the *psi bank-the* medium of information storage and retrieval for all the mental programs. The noosphere is the field of planetary consciousness in which humans are intended to participate and complete their evolution. However, operating by artificial time is like looking through a glass darkly. This being so, there is no noosphere to see. Humans generally experience consciousness through the chinks in their sensory prism, conditioned by separatist beliefs, making it most difficult to attain a perception or experience the field of consciousness as coextensive with the Earth. Indeed, if the biosphere is the region on Earth for the transformation of cosmic energies, the noosphere is that region of the Earth for the reflection of cosmic consciousness and its mental programs. Seen through the lens of universal synchronized time, the humans accommodating to a single instrument of time measurement are inherently sharing a single thought form. It is that act of self-reflective convergence that makes the noosphere conscious and draws it toward the resplendent omega point of2012 and beyond.

As with all the other geochemical structures and processes of the Earth that are planetary in nature, the noosphere, the rainbow brain of Earth, actually has a coherence and an order that we may define in a precise way. In so doing we shall redefine the nature of human reality and the human being itself. We must enlarge our lens of perception to encompass the whole Earth as a single being. In relation to this single being of the Earth, the individual human is hardly even a speck. The dimensions and proportions by which we commonly measure reality must also expand to accommodate the vast dimensions of the Earth as a single entity, a planetary geocosm within the solar field, the heliosphere. Since the sun is also a star, the heliosphere is a function and member of the galactic order of being. As the mental "thinking sphere" of this planetary being, Earth, the noosphere is above and discontinuous with the biosphere, the hydrosphere, and the atmosphere. It is contained by and functions within the vast invisible electromagnetic field of the Earth. Though the ionosphere is the lowest level of the electromagnetic field system of the Earth, the main structures of this field are the two radiation belts located 2,000 and 11,000 miles above the Earth. It is within the two radiation belts that the control panels of the noosphere are to be found-the psi bank.

KIN: BIPOLAR BASE MEASURE OF THE 13:20 MATRIX

> *"And He is the One who created the night and*
> *the day, and the sun and the moon; each*
> *floating in its own orbit."*
>
> -QuRAN,21:33

In the simplest aspect of our daily experience of time on Earth we have the full play of the law of alternation-day-night, sun-moon, north-south. Because we know the bipolar nature of the Earth, dividing it into two geomagnetic fields, north polar and south polar, we can then define the complementary bipolar field of time, which divides the cycle of the Earth as it makes one rotation on its axis into one day and night. This entire unit, day and night, is defined as a *kin,* the base unit of synchronic time measurement. From the perspective of the noosphere, day and night are a function of the two alternators-day alternator and night alternator-that operate as the phasic motion of the rainbow brain of the noosphere working in tandem with the electromagnetic field. The two phasic alternators regulate the activation of photons (day) and the regeneration of photons (night) within the entirety of the life process of the biosphere.

THE NOOSPHERIC FOURFOLD/EIGHT-PART STRUCTURE

> *"He created the heavens and the earth truthfully. He rolls the night over the day and the day over the night. He committed the sun and the moon, each running for a finite period. Absolutely he is the Almighty, the Forgiving."*
>
> -QURAN, 39:5

> *"We rendered the night and the day two signs. We made the night dark and the day lighted, that you may seek provisions from your Lord therein. This also establishes for you a timing system, and the means of calculation. We thus explain everything in detail."*
>
> -17:12

Since the two alternators are also in interplay with the bipolar magnetic fields of the Earth, a four-part process is revealed that is magnetically distributed as an overall eight-part structure. The unit of one kin divides the rotational cycle of the Earth in its orbit into four distinct phases daily and annually. The four daily points are defined by: the dawn, when the part of Earth you are in first experiences the solar photonic activation; the high noon (zenith), when the part of Earth you are in is in closest direct relation to the sun; the twilight, when the photonic activation ceases and the sun disappears from view; and midnight (nadir), the point at which the point of Earth you are on has rotated to its farthest distance from the sun, the height of the cycle of photonic regeneration. This daily process is mimicked in the four points of equinox and solstice, where for either hemisphere the equinoxes are the same as the points of dawn and twilight and the solstices correspond either to high noon or to midnight. Because of the magnetic polarity of the Earth, the four-part process is

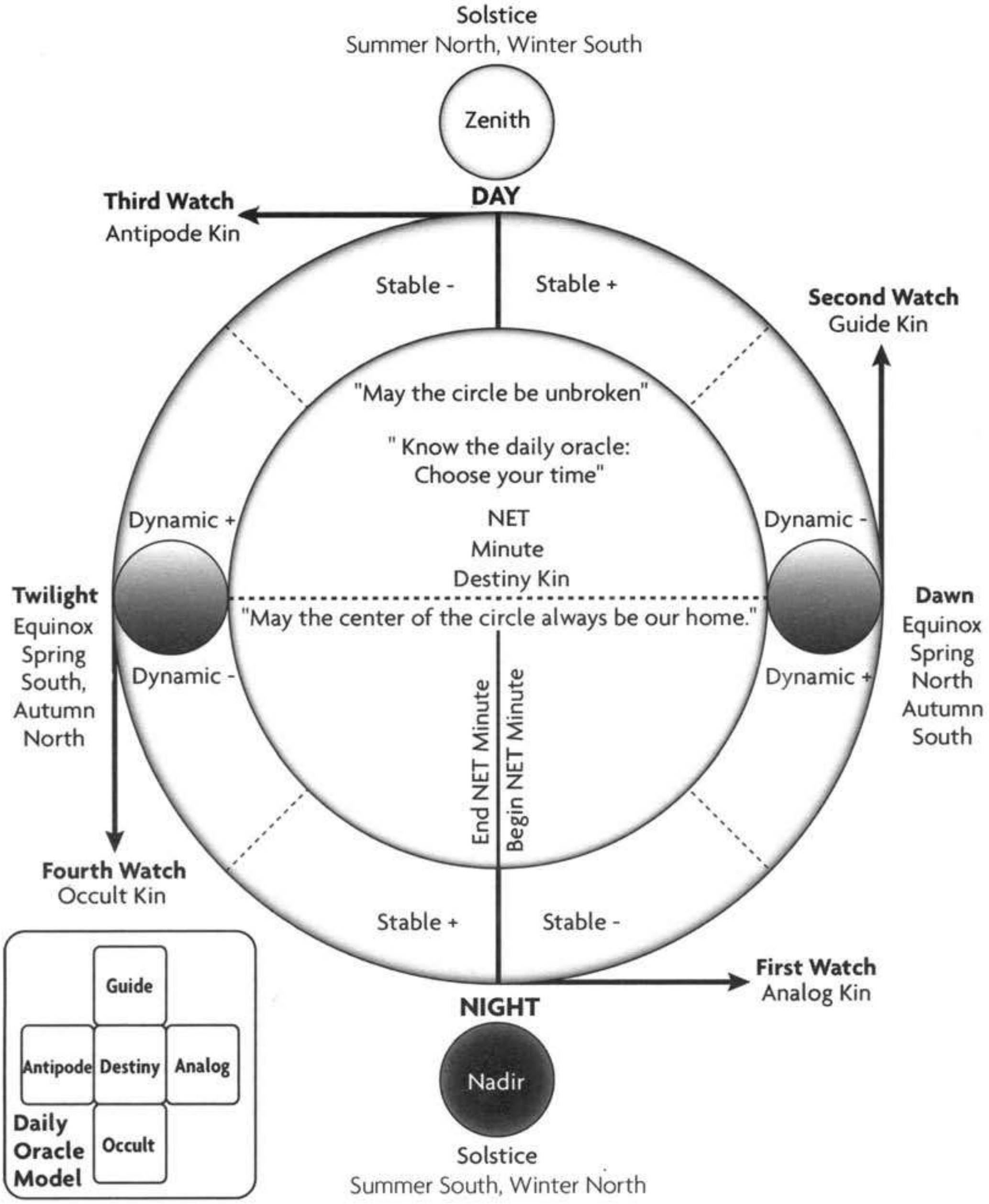
Solstice
Summer North, Winter South
Zenith
DAY
Third Watch
Antipode Kin
Stable -
Stable +
Second Watch
Guide Kin
"May the circle be unbroken"
" Know the daily oracle:
Choose your time"
NET
Minute
Destiny Kin
Dynamic +
Dynamic -
Twilight
Equinox
Spring
South,
Autumn
North
"May the center of the circle always be our home."
Dawn
Equinox
Spring
North
Autumn
South
Dynamic -
Dynamic +
End NET Minute
Begin NET Minute
Fourth Watch
Occult Kin
Stable +
Stable -
First Watch
Analog Kin
NIGHT
Nadir
Solstice
Summer South, Winter North
Guide
Antipode
Destiny
Analog
Occult
Daily
Oracle
Model

Four-Part Bipolar Kin:
Four daily watches,
Four annual seasons—
Two points of balance
Two points of return

Net Minute Model

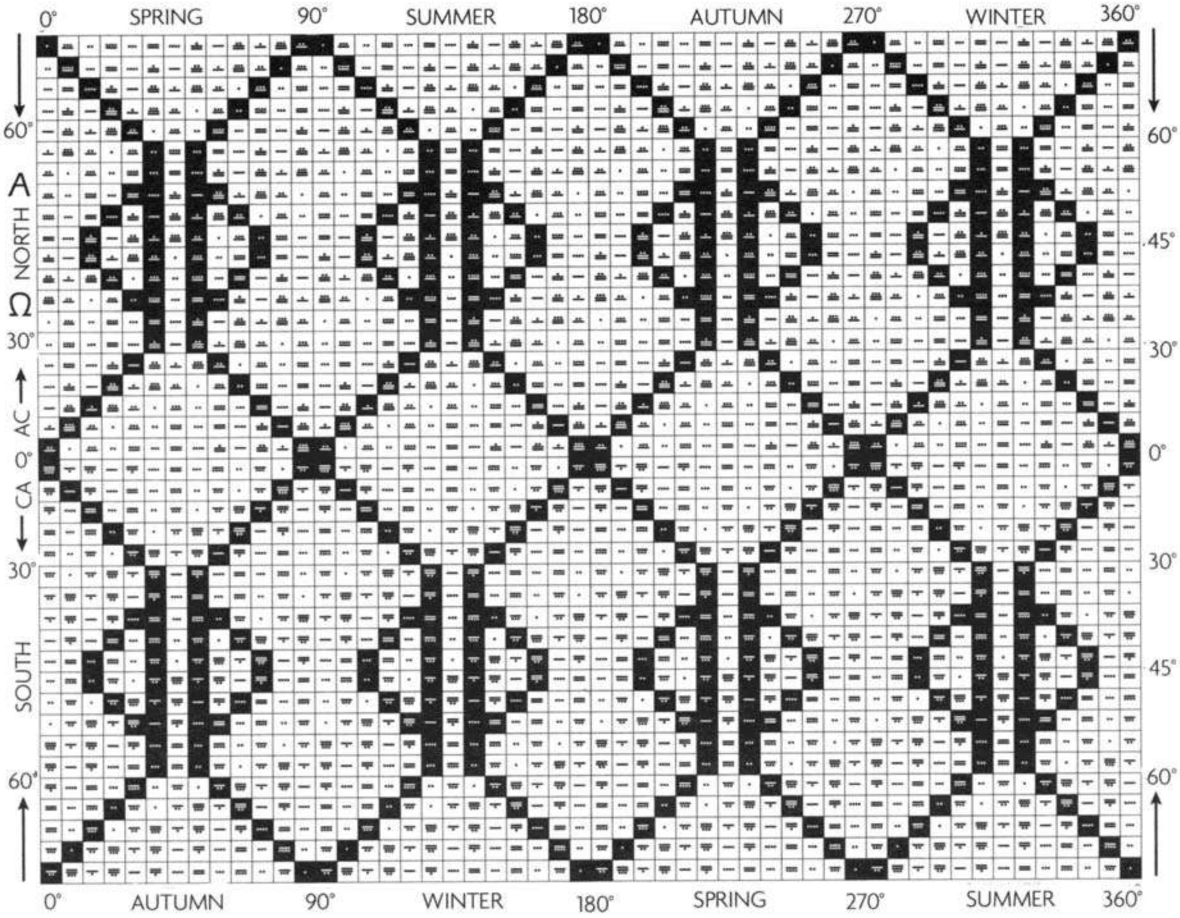
0°
SPRING
90°
SUMMER
180°
AUTUMN
270°
WINTER
360°
60°
A
NORTH
Ω
30°
AC
0°
CA
30°
SOUTH
60°
60°
45°
30°
0°
30°
45°
60°
0°
AUTUMN
90°
WINTER
180°
SPRING
270°
SUMMER
360°

Psi Bank—Four Bipolar Plates Accommodating NET Time Autoregulatory System of the Noosphere
(from *Earth Ascending*, Map 1)

simultaneously experienced in opposite phases, that is, winter in the North is summer in the South-hence the four-phased process yields a bipolar eight-fold structure.

This four-part day-night seasonal process defines the structure of the four psi bank plates, which also coordinate the four-year cycles of the Earth. The psi bank plates, the "nervous system" and control panel of the noosphere, are a complete function of the 13:20 timing frequency and thus participate in the systematic regularity of the fourth-dimensional synchronic order. This means that the psi bank is the registration within the Earth's magnetic field of the 13:20 timing matrix by which all life and the planetary evolutionary process as a whole are regulated. Simulating the four-part process of the Earth rotating on its axis in time, four stages per day, four seasons per year, in four-year cycles-not to mention the four phases of the moon-four psi bank plates accommodating the bipolar field yield eight 260-unit 13:20 frequency information matrices.

That is, each of the four psi bank plates conforms to a mirror symmetry of the bipolarity of the Earth's magnetic field. Each plate is divided by mirror symmetry into a 260-unit Northern and a 260-unit Southern hemispheric half, for a total of two bipolar matrices and 520 units per plate, or 2,080 (520 x 4) units in all. These units are referred to as *psi chrono units* or *time-bearing information units.* The information for the different sequences of the Earth's biogeochemical evolutionary processes is all regulated by these time-bearing information units.

We have to understand that through the medium of universal cosmic time the Earth, the Sun, and every star, constellation, and galaxy is coordinated in this way by the synchronic 13:20 frequency of time. It is the 13:20 frequency that establishes holonomic consistency in time throughout the universe. The Earth's noosphere is a lens of time. All cosmic information is focused through this lens. It is this noospheric lens of time that makes Earth into a geocosmic thought form.

The noospheric lens of time and mind has been operating ceaselessly-but unconsciously-since the beginning of the evolution of the Earth. Since the noosphere is a function of the universal synchronic timing frequency, its conscious activation must also come about through a superior mental coordination with this correct synchronizing timing frequency. The psi bank 13:20 frequency is consciously activated and registered at the human level by the mental engagement of the 13:20 matrix as enacted through the 260-day cycle. The interaction of the 260-day galactic cycle with the 365-day solar biotelepathic cycle defines one noospheric year, or one solar-galactic cycle of fifty-two human years (= fifty-two 365-day solar orbits = seventy-three 260-day galactic cycles). This great noospheric coordination of fifty-two solar orbits with seventy-three galactic cycles is fractally present in a single solar orbit, where fifty-two 7-day weeks = seventy-three 5-day cycles known as chromatics. What is important in coming to understand the noosphere as Earth's geocosmic time lens and rainbow brain, is that we must change our perspective altogether from anthropocentric to noospherocentric, and in this way evolve into *Homo noosphericus.*

THE CYCLES OF NET TIME

For the noosphere, the rainbow brain of planet Earth, time is different-it is slower and longer. From the noosphere's point of view, what we call a year is a solar orbit. What we call a day is a single rotation of the Earth on its axis. The cycles of the moon are the lunar phasings that calibrate the synchronization of the Earth in time. The Earth is not a spaceship. The Earth is a Timeship. As a Timeship, the Earth is a function of the master synchronization frequency that coordinates and moves it in time through ever greater circles of inclusiveness and syntropic integration. Seen through the noospheric lens of time, the technosphere is a highly exaggerated acceleration of the biogeochemical continuum through highly artificial and resource-depleting means-it is but a smoking bubble whose fragile struts and frets are laced together by transport and communication systems that yield a cacophony rather than a harmony. And for this reason, the disharmony of the bubble ultimately explodes upon itself, releasing Earth from the prison of false time into the fresh air of real time and the advance into the geocosmic splendor of the noosphere. Synchronized with the galactic cycles, the noospheric mind of the whole Earth is reckoned far differently from the artificial units of measure that condition the anthropocentric notions of time evolved in the era of *Homo historicus.*

Here is how the fourfold process of Earth in time is reckoned by the noosphere. Please note that in Noospheric Earth Time (NET), time is universal synchronization. At any given moment, it is always the same noospheric moment in time on

Table of Noospheric Earth Time Units

1. One Noospheric Earth minute = 1 day-and-night rotation of Earth on its axis = 1 kin.
2. Twenty-eight Noospheric Earth minutes = 1 Noospheric Earth hour = one moon of 28 kin. where every 7-day week = 1/4 Noospheric Earth hour = 7 kin.
3. Thirteen Noospheric Earth hours (moons) = 1 Noospheric Earth day = 365 kin (days) = 1 solar orbit. or 1 human year (= 13 NET hours x 28 minutes + 1 NET minute, "Day Out of Time:')
4. Four Noospheric Earth days = 1 Noospheric Earth week = 4 solar orbits (years) = 1,460 kin (days).
5. Three Noospheric Earth weeks plus 1 Noospheric Earth day = 1 Noospheric Earth season = 13 solar orbits or 13 human years.
6. Four Noospheric Earth seasons = 1 Noospheric Earth year = 52 solar orbits (52 human years) = 73 260-day galactic cycles = 18,980 kin (days).
7. Five Noospheric Earth years = 1 Noospheric time chord = 260 solar orbits.
8. Twenty Noospheric time chords = Vs Noospheric evolutionary subcycle = 5,200 solar orbits = 1 Noospheric Century of 100 Noospheric Earth years.
9. Five Noospheric evolutionary subcycles = 1 galactic evolutionary day = 26,000 solar orbits = 5 Noospheric Earth centuries or 500 Noospheric Earth years of 52 solar orbits each.

Earth. Whether that moment is experienced on the night or day side of the Earth, that moment is simultaneous and encompasses both sides of the Earth at once. This is true not only because time is the factor of synchronization- T(E) = Art, where art is the result of any synchronization-but also because the velocity of time is infinitely instantaneous. By thinking that the Earth really does conform to the 24 hours of the clock, the humans have retarded their time sensibility, and consequently their artificial processes have become increasingly artless. In considering the Noospheric Time measurements, remember that as a human you are but a cell in the biospheric organism-or, as we shall see, a chip of the noospheric rainbow brain.

As the measure of the mind of the Earth, the Noospheric Earth Time (NET) reckoning units coordinate consciousness as a planetary, or, more accurately, a solar-planetary phenomenon. For instance, if one NET day is a single 365-day solar orbit of the Earth, dawn of the noospheric Earth day is experienced as the spring equinox, but the noosphere is bipolar. So the dawn is also simultaneously the autumn equinox or twilight of the other polar magnetic half of the Earth. To fully engage and grasp such a thought is to expand the limits of consciousness. Like dolphins who never sleep but only rest half of their brain at a time, so it is with the rainbow brain of the noosphere. Such consciousness has nothing to do with the workings of the everyday egoistic human consciousness, which can understand things only according to a limited frame of reference defined by its own limited and highly conditioned ego needs. The geocosmic consciousness of the noosphere can be apprehended only as a single telepathic unity in and by which all members of any given species or organism are organized. To consider that the noosphere itself could become the organizing principle of the human species as a whole is to engage in conscious mental evolution.

Small aboriginal prehistoric bands of humans experience an unconscious level of this geocosmic noospheric mind. Dolphins, other mammals, and all animal species experience this noospheric mind. Even trees and crystals are organized by this noospheric mind of time.

Homo historicus, however, slowly but inexorably cut himself off from the noospheric mind of time by substituting erroneous instruments of measure, or instruments of measure that were incomplete in relation to the noospheric time cycles. The result was the exponential surge in biogeochemical combustion that characterized the final baktun cycle, culminating in the fifty-six-year blister of artificial time, the technosphere. The fact that the time blister popped eleven years before the end of the cycle of *Homo historicus,* 2012, means that *Homo noosphericus* is ready to emerge from the shattered cocoon of false time. Liberated from artificial time and from the technosphere itself, and enrolled in the natural time of telepathy, the human will experience

Noosphere Earth's Rainbow Brain

Magnetic Alternation - Night and Day
Law Of The Ring
"The Psi Bank in Perfect Motion"

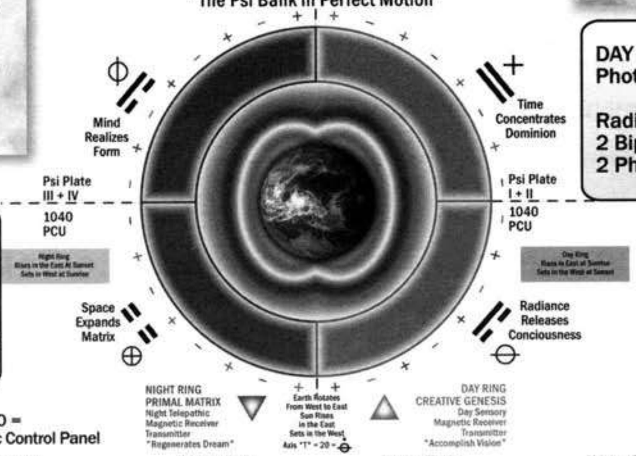
Mind Realizes Form
Time Concentrates Dominion
Psi Plate III + IV
1040 PCU
Psi Plate I + II
1040 PCU
Night Ring Rises in the East At Sunset Sets in West at Sunrise
Day Ring Rises in East at Sunrise Sets in the West at Sunset
Space Expands Matrix
Radiance Releases Conciousness
NIGHT RING
PRIMAL MATRIX
Night Telepathic
Magnetic Receiver
Transmitter
"Regenerates Dream"
Earth Rotates From West to East
Sun Rises in the East
Sets in the West
Axis "T" = 20 =
DAY RING
CREATIVE GENESIS
Day Sensory
Magnetic Receiver
Transmitter
"Accomplish Vision"

T(E)=ART

DAY ALTERNATOR
Photonic Activator

Radiation Belt x 2 =
2 Bipolar PSI Plates -
2 Phases

NIGHT ALTERNATOR
Photonic Regenerator

Radiation Belt x 2 =
2 Bipolar Psi Plates =
2 Phases

2 Phases + Polarity =
4 Bipolar Psi Plates @ 13:20 =
Nervous System + Noospheric Control Panel

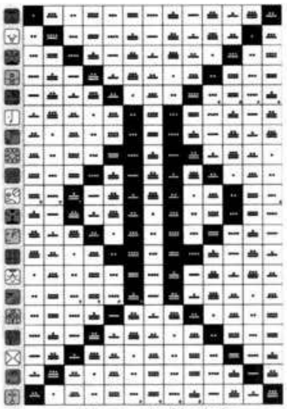

20 Solar Frequencies

Harmonic Module

13:20 Matrix Programs Four Phase Bipolar Psi Plates

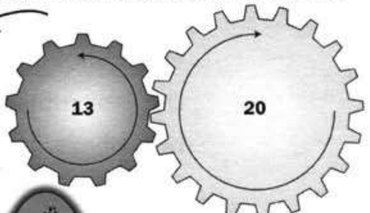

13:20. Ratio Constant of Universal Synchronization

Earth Wizard.
"Homo Noosphericus"

Synchronic Map and Guide to the Noosphere

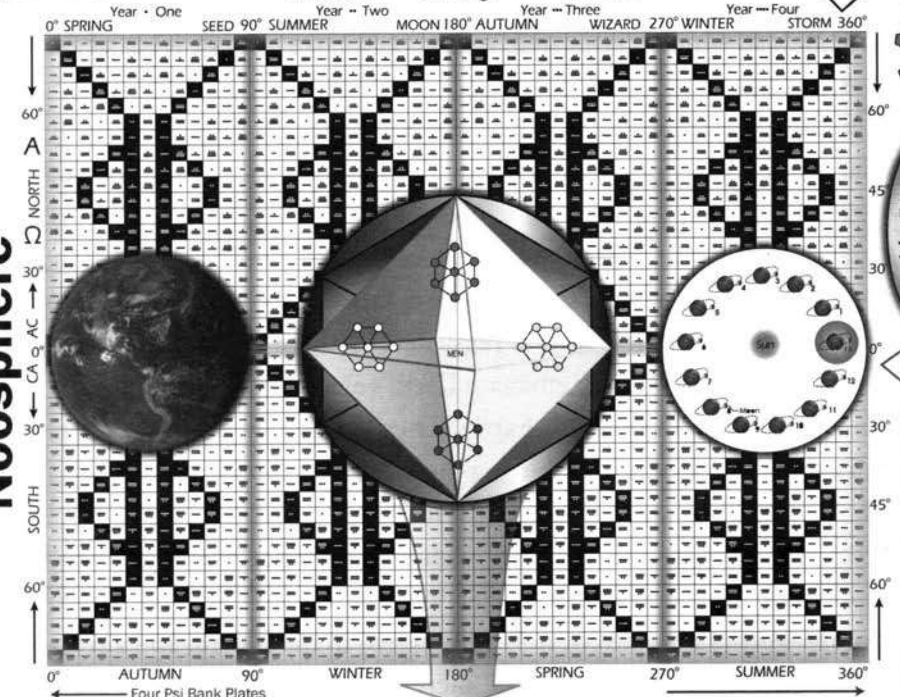
Year · One
Year ·· Two
Year ··· Three
Year ···· Four
0° SPRING
SEED 90° SUMMER
MOON 180° AUTUMN
WIZARD 270° WINTER
STORM 360°
60°
A
NORTH
Ω
30°
AC
0°
CA
30°
SOUTH
60°
45°
30
0°
30°
45°
60°
MEN
SUN
Moon
0°
AUTUMN
90°
WINTER
180°
SPRING
270°
SUMMER
360°

Four Psi Bank Plates

13 Moon 28 Day + 1 Noospheric Regulator = 1 Noospheric Earth Day

Noospheric Earth Time (N.E.T.) Units

1 Day = 1 NET Minute
28 Days = 1 NET Hour = 1 Moon

13 Moon NET Hours = 1 NET Day = 1 Solar Orbit

4 Solar Orbits = 1 NET Week

3 NET Weeks + 1 NET Day = 1 NET Season = 13 Solar Orbits

4 NET Seasons = 52 Solar Orbits = 1 Noospheric Earth Time Year

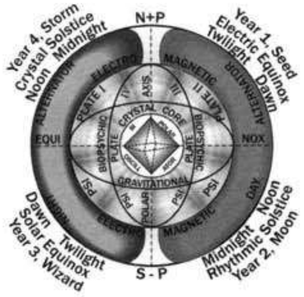
N+P
Year 4, Storm
Crystal Solstice
Noon Midnight
Year 1, Seed
Electric Equinox
Twilight Dawn
ALTERNATOR
ELECTRO
MAGNETIC
ALTERNATOR
PLATE I
IV
AXIS
III
PLATE II
CRYSTAL CORE
EQUI
NOX
BIOPSYCHIC
PLATE
PLATE
BIOPSYCHIC
GRAVITATIONAL
PSI
PSI
PSI
PSI
POLAR
DAY
NIGHT
ELECTRO
MAGNETIC
S - P
Dawn
Solar Equinox
Twilight
Year 3, Wizard
Midnight Noon
Rhythmic Solstice
Year 2, Moon

Four Phasic Noospheric Resonator
Timeship Earth Field
- Day Night Alternator - 4 phases
  Twilight Noon Dawn Midnight
- 1 Moon-NET-Hour = 4 weeks ¼ Hours
- Solar Orbit = 4 Seasons = 4 Phases NET Day
  [Electric, Solar Equinoxes; Rhythmic, Crystal Solstices]
- 4 Solar Orbits = 1 NET Week = 1 Seed Storm Year Bundle = Return of 4 Psi Plates, 1 per Year.

Bipolar Oscillator and
Crystal Transducer Earth Core
Activated by 1 Time Atom per ¼ Hour (7 Days) x 4 = 4 per Hour, 28 Days = 1 Moon-NET-Hour

Planet Holon - 20 Day Cycle

Timeship Earth 2013
Coded by 20
Solar Frequencies
in Time

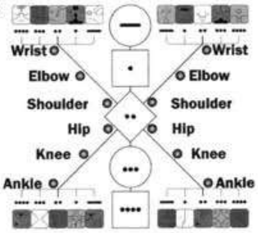
Wrist
Wrist
Elbow
Elbow
Shoulder
Shoulder
Hip
Hip
Knee
Knee
Ankle
Ankle

1 Unit = 28 Days or 1 Moon

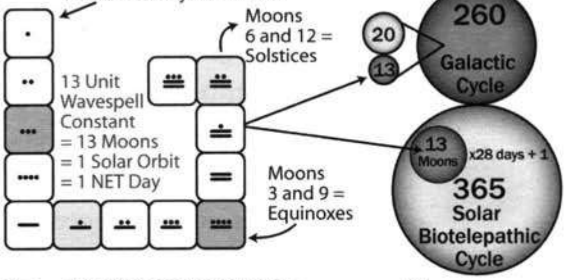
Moons 6 and 12 = Solstices
13 Unit Wavespell Constant = 13 Moons = 1 Solar Orbit = 1 NET Day
Moons 3 and 9 = Equinoxes
20
13
260
Galactic Cycle
13 Moons
x28 days + 1
365
Solar Biotelepathic Cycle

13:20 Noospheric Chip
"Homo Noosphericus"
20 Fingers + Toes
4 Limbs
13 Main Articulations
Consciously synchronized in 13:20 time. Plays Psi Bank Control Panel, generates Earth's Rainbow Brain

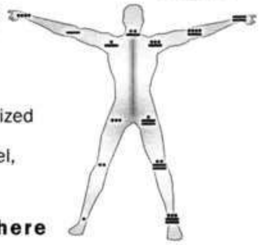

Man in the Noosphere

noospheric organization into psychic biomes. Through the psychic biomes-collectives of communities coordinating their bioregional network according to the structure of the "thought moments" of noospheric time-the human will fully evolve the Earth as a planet with solar-galactic consciousness.

ELECTROMAGNETIC WORKINGS OF THE EARTH'S RAINBOW BRAIN

What we call the cycles of time are the thought moments of the Earth. These thought moments coordinate the complex orders of Timeship Earth, understood as a whole system. Because the operating mechanism of the noosphere that governs the Earth's cyclic thought moments is inseparable from the general structure and functions of the electromagnetic field in which it is located, we may refer to a *rainbow brain* of the noosphere. Why is this so?

Let us return to the structure of the noosphere, Earth's mind in time. Consisting of the four bipolar psi bank plates, these four synchronic time plates are located between the two radiation belts, that is to say between 11,000 and 2,000 miles above the biospheric surface of the Earth. With the location of its information time-bearing mechanism between the radiation belts, the noosphere must by necessity operate in tandem with the electromagnetic field as it is regulated by these two master radiation belts. What do these belts do? They emanate Earth's electromagnetic aura, and function as a protective shield around the Earth. They do this by trapping solar-cosmic radiation, keeping much of it from hitting the biosphere directly, channeling it instead through the Earth's magnetic poles into the Earth's core.

In the polar regions excess cosmic radiation, including electrically charged plasmas and solar discharges from coronal mass ejections, is released into the atmosphere as the auroras-aurora borealis (or northern lights) in the North, and aurora australis (or southern lights) in the South. The increased solar activity during the most recent and current sunspot cycles has enhanced and extended the auroral activity in recent years far beyond the polar zones. In addition, since 1987 great influxes of plasma into the solar system have intensified the incidence of spectral plasmic activity within the atmosphere itself. This amounts to an increase in rainbows, halos around the sun, and brilliantly colored clouds. While these phenomena may be explained to a certain degree as meteorological effects (the angle of the sun's rays, moisture in the atmosphere, and so on), is there more to a rainbow than meteorological side effects? And what relation is there between the noosphere, the electromagnetic field, and the increased plasmic-rainbow activity?

Again, seen from the perspective of whole Earth, the rainbows and spectral plasmic effects, including the auroras, are the free energy released by the noosphere. Rainbows are to the noosphere what toxic waste is to the technosphere. The noosphere itself implies a mental field that is interacting with the electromagnetic field. The Earth's electromagnetic field is actually the strongest of all the planets in the solar system. The only other planet that has a field approaching the strength of the Earth's field is Uranus. If the Earth's electromagnetic field is the strongest, is this an indication of the strength of the noosphere as well? If the rainbows, auroras, and other spectral plasmic activity are a demonstration of the free energy released due to the interaction of the noosphere and electromagnetic fields, and the noosphere is the mental envelope of the planet, inclusive of the as-yet-to-be-unified mind of the human species, what forms of free energy might be released if the human mental field were operating as a single unity in the correct timing frequency?

One other mechanism crucial to the entire system of noosphere and electromagnetic fields is the crystal core at the center of the Earth. In the 1970s Russian scientists put forth the idea that the Earth is crystalline in nature, an idea that extended Plato's notion of the Earth as a dodecahedron, one of the five "platonic solids." The idea was that the Earth's surface was held together by a field of resonance that corresponds to the crystal resonance of certain geometrical solids. R. Buckminster Fuller also gave play to this notion in his description of "spaceship earth." Then in the 1980s American seismologists confirmed that the core of the Earth is actually a gigantic iron crystal structure in the shape of an octahedron, elongated at its ends, which point to the North and South Poles.

From the point of view of fourth-dimensional time, the planet Earth is another manifestation of space as an infinitely locatable point, an orbiting body upon which all the vectoral programs of cosmic time may be unfolded. Understood as a Timeship, the Earth's structure is constructed to facilitate its travel in time on a particular route in space in conformity with its relation to the local star, the sun, to the other planets of the solar system or heliosphere, and in synchronization with the universal timing constant 13:20, which maintains the Earth in the synchronic order. The role of the magnetic octahedral crystal core is to function as the bipolar oscillator and crystal transducer whose purpose is to maintain the Earth in the synchronic order of the 13:20 timing frequency. Thus, if the Earth's noosphere contains the 13:20 timing programs of the psi bank, and the crystal bipolar oscillator is meant to maintain the Earth's space body in the synchronic order, obviously there is a resonant relation between the crystal core and the noospheric-electromagnetic field programs.

The transductive properties of the crystal allow it to transform energy from one state to another. The crystal core also stores electronic plasmas that travel to it from the

Earth's poles through an electromagnetic flux tube system. Within the crystalline core, the electronic plasmas are transduced into information. This is the information of time understood as cosmically transmitted thought forms. These plasmic thought forms conform to programs already established in the psi bank of the noosphere. These are the Earth's thought forms, the cycles of the Earth's brain waves, as it were, that allow it to maintain its biospheric programs and to evolve its mental field, the noosphere. The very descriptions of the Noospheric Earth Time (NET), and of time itself as a synchronous instantaneity inseparable from telepathy, are examples of the Earth's thought forms transduced through a human biocomputer and then transmitted into a form of human language for the purpose of communication. Keeping this description of the noospheric brain in mind, let us return to our analysis of the biosphere-noosphere transition.

MAKING THE NOOSPHERIC LEAP: TRANSFORMATION OF HUMANS INTO NOOSPHERIC CHIPS

The transition to the noosphere is experienced by Earth as a crisis because its systems are stressed. Since the Earth functions as a whole system, stress on one component of the system actually stresses the whole order. Let us review the cumulative stress points. Human 12:60 electromagnetic activity has perturbed and distorted the electromagnetic field. The release of free energy due to geochemical combustion adds to this stressful distortion through the depletion of ozone in the ionosphere and global warming. The radioactivity released into the atmosphere adversely affects the resonant field of the DNA, already driven by the 12:60 frequency to accelerate the multiplication and propagation of the human biomass throughout the biosphere. Through its artificial means, the machine, the human further upsets the inherent balance maintained by the pressure that the species exert upon each other, resulting in mass species extinction.

Yet all this is according to an evolutionary program that is intended to bring about an ultimate crisis by which the biosphere snaps into becoming the noosphere. Clearly, if the root of this crisis is the deviation of the dominant species operating according to an artificial timing mechanism at odds with the universal timing frequency, then the stabilization of the transition to the noosphere will come about through the dominant species consciously "returning" to the correct timing frequency. This it can only do by adhering to a macro-organizing program that puts the human species completely in the 13:20 frequency.

If the whole system analysis seeks to locate the flaw in the operation of Times hip Earth, then we must finally declare that this flaw is the human civilization in its entirety, as it is maintained by erroneous and artificial timing constructs. Since the noosphere, the mind of time on Earth, defines the next evolutionary cycle, and because the humans are following a course dictated by being in the wrong time, calendar reform is the eye of the needle through which humanity must pass in order to get to the noosphere. Only by fully manifesting the noosphere will the Timeship Earth be truly put back on course once again. The noosphere is the program of time on Earth. Time is all that the noosphere has to give us. In consideration that the artificial time of the technosphere has fundamentally run out, the noospheric time should be a welcome gift.

Science is always absolutely precise. It is not just calendar reform that is the eye of the needle through which to pass into the noosphere-it is the Thirteen Moon/28-Day calendar calibrated by the 260-day Wizard's Count, synchronized to the Gregorian calendar date, July 26, that provides the exact ticket of admission into the noosphere. And then this ticket must be used at just the right moment-when the show is ready to begin. To use another metaphor, there is a lock on the door to the noosphere, and there is only one key that will fit and unlock this noospheric gate. That key is the Thirteen Moon/28-Day calendar, correlated to July 26 on the Gregorian calendar. It is the precisely applied use of this Thirteen Moon/28-Day calendar that will make each human a noospheric chip of the rainbow brain. This is because the Thirteen Moon/28-Day calendar is the key with the precise design for not only opening but also for continuing to operate the noosphere through a matching of the everyday frequency of time with the frequency of the noospheric control panel, the psi bank.

It is because the daily program of the Thirteen Moon/28-Day calendar is embedded in the 13:20 frequency that the consciousness of the humans will finally change and rise above the stunted level by which it is currently maintained in the artificial world of the technosphere. Time is of the mind, and the natural time registered by the Thirteen Moon/28-Day calendar is the opening to the geocosmic mind of the universe, which adherence to artificial time has kept the human from experiencing. The noosphere is the fractal lens of this geocosmic mind on Earth. Since the geocosmic mind is purely a function of the Law of Time-energy factored by time equals art, where the 13:20 synchronizing ratio of time is characterized by infinitely instantaneous velocity-then the fundamental nature of this mind is telepathy.

The Thirteen Moon/28-Day calendar is, therefore, the perfectly designed instrument to make the noospheric leap into telepathy, the planetization of consciousness that establishes the daily reality of the noospheric mind. Since the noospheric mind interacts with the electromagnetic field, and because the ratios and cycles of the Thirteen Moon/28-Day calendar conform to the fractals and ratios of the NET, by using this calendar the human becomes transformed into a noospheric chip. The fractals and ratios of NET become the normative operating cycles of the human in its daily reality. This accounts for an increasing noospheric synchronization that allows the noospheric chip to attain its purpose: generating the Earth's rainbow aura. If the free energy of rainbows is to the noosphere what the free energy of toxic waste is to the technosphere, then the generation of the rainbow implies a type of activity on the part of the human to produce such free energy. To understand what this activity might be and how it works, we must first understand the morphic and holonomic relation between the human noospheric *chip-Homo noosphericus-and* the structure of the noosphere itself as it is programmed into the geocosmic measure of the Thirteen Moon/28-Day calendar.

HUMAN BEINGS AS FRACTAL REFLECTORS OF THE 13:20 MATRIX

From the noospheric perspective, the prototype human is an embodiment of the 13:20 synchronization factor. The human has four limbs. Each of the four extremities has five digits, for a total of twenty fingers and toes that distinguish the human's operational capacity. The four limbs conform to the fourfold phasic operation ofNoospheric Earth Time, while the twenty digits conform to the 20 (4 x 5) factor of the 13:20 ratio. The human also has thirteen main joints or articulations: two ankles, two knees, two hips, two wrists, two elbows, and two shoulders; these six pairs coordinate the four extremities with the body, while the thirteenth articulation, the neck and spinal column, coordinates the body proper. These thirteen main articulations plus the twenty digits of the hand and feet conspire to make the human the perfect exemplar of 13:20 timing frequency, that is, a perfect model of synchronization itself.

Add to this the fact that the human has a highly evolved and refined brain and nervous system that is maintained and expressed by the skeletal structure of the thirteen joints of the body and limbs, completed by the sensitive mechanism of the twenty digits, and we see that the human is the perfect biocomputer for processing the information of the instantaneously telepathic geocosmic mind. The means to process the geocosmic information, and also the means by which the human is transformed into the noospheric chip, is found in matching its factors of daily synchronization and intelligence with those of the psi bank, the noospheric control panel. This is called harmonic arrangement of the synchronic order.

Of course, this description of generalized structure, activity, and interaction of the *Homo noosphericus* with the noosphere as such presupposes an entire reorientation of the goals and methods of being human. Such a reorientation of the human is also a function of making the radical change from the erroneous measure of the twelve-month Gregorian calendar, in which are embedded all of the Babylonian programs of history, to the perfect history-less harmony of the Thirteen Moon/28-Day calendar-the mere change of calendars encompasses a mental frequency shift. The readiness to make this change depends on the degree of moral revulsion the human experiences during the collapse of the technosphere. This moral revulsion would then correspond to a turning in the deepest seat of consciousness, which would then dispose the human intelligence to a consideration of the untried solution-change the calendar!

The Thirteen Moon calendar is the NET instrument for daily synchronization and unification of the human species as a single organism in time. Once adopted to the correct measure of this instrument, the human will de facto trigger the noosphere as an operating system of absolutely higher consciousness and mental order. **In** fact, the purpose of human evolution is precisely to become the ignition of the noosphere,

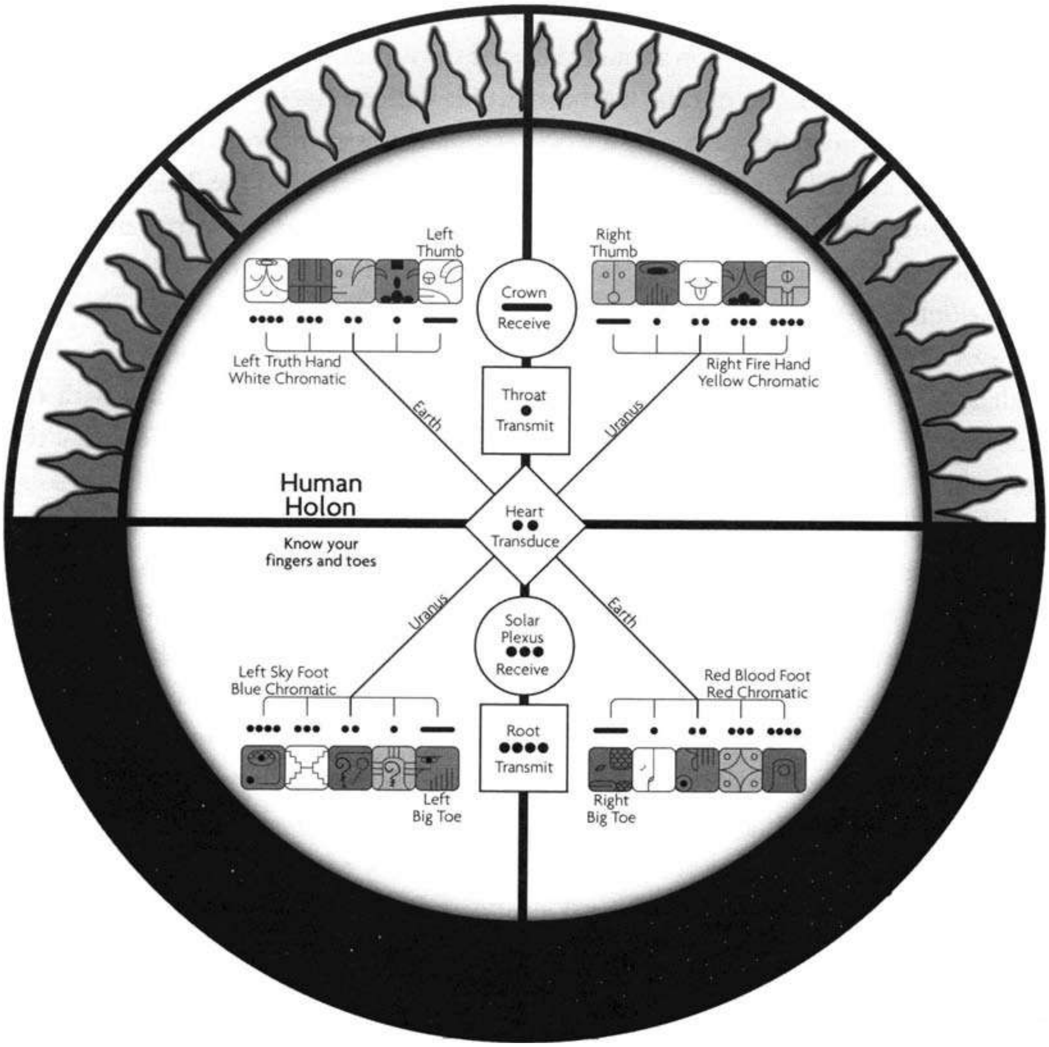
Left Thumb
Right Thumb
Crown
Receive
Left Truth Hand
White Chromatic
Right Fire Hand
Yellow Chromatic
Throat
Transmit
Earth
Uranus
Human Holon
Know your fingers and toes
Heart
Transduce
Uranus
Earth
Solar Plexus
Receive
Left Sky Foot
Blue Chromatic
Red Blood Foot
Red Chromatic
Root
Transmit
Left Big Toe
Right Big Toe

Human Noospheric Chip

and thence to become normalized as the noospheric chips that maintain the operations of the noosphere. As a truly planetized being and consciousness, *Homo noosphericus* will engage in an order and type of reflective activity undreamed of in the technospheric era that is now passing.

First, by adapting to the daily NET timing instrument, the humans will be utilizing a "calendar" that correlates both a 365-day solar biotelepathic cycle and a 260-day galactic synchronization cycle. When the lunar phases are added to the interaction of these two cycles, then the human is able to enjoy the harmonic richness of time-the noosphere. The synchronizing and interactive measures of the counts of the solar and galactic cycles yield multiple levels of harmonic orders of time (Earth's "thought moments") that will activate aspects and levels of consciousness hitherto dormant in the human mind. The very mental act of replacing the order of twelve by the order of thirteen is a profound one. How is this so? Could it be that the whole of civilized history itself is based on the fear of the number *thirteen-triskaidekaphobia,* epitomized in the superstition about Friday the thirteenth-and that, therefore, dealing with the true nature of time has been avoided altogether?2

Currently humans operate by a world standard whose measure includes: seven-day weeks, inclusive of the notions of the 5-day work week and the 2-day weekend; an irregular distribution of twelve months of either thirty-one, thirty, or, in one instance, twenty-eight days; all set within a longer annual cycle that concludes with its visionless moment of New Year's Eve. Many Gregorian calendars will include the phases of the moon, but because the months differ in length there is no way in these calendars of seeing how the phases of the moon might be synchronized with the measure of Earth's orbit around the sun, nor is there any reason to consider that the uneven measure of twelve months has anything intrinsically to do with the solar orbit. Within the twelve-month structure of the Gregorian calendar, weeks, months, and phases of the moon all appear to be haphazard and arbitrary. Such a disorderly order can only bode an undercurrent of uncertainty and fear for the human mind.

By contrast, in the measure of the thirteen moons of twenty-eight days each, the 13 corresponds both to a 13-day cycle and to the cycle of the number of months (moons) per year. The measure of the 7-day week (a NET quarter hour) becomes an orderly harmonic: four perfect weeks occurring each moon, fifty-two perfect weeks each year, each week always beginning and ending on the same day of the week, week after week, year after year, while in fact every day of every moon occurs on the same day of the week, year after year. The effect of such a harmonic regularity on the human mind is incalculable. This is what is meant by a *perpetual calendar.* And to maintain this perfect regularity, the 365th day of every year (solar orbit), is no day of the week at all, but a Day Out of Time, a day for the human to be refreshed in utter timelessness.

Realizing that a day is actually a kin or one NET minute, the measure of one rotation of Earth on its axis, a 4-day cycle of four NET minutes constitutes a harmonic, a 5-day cycle constitutes a chromatic, while the 13-day cycle corresponds to one *wavespell,* the time constant that accommodates the measure of thirteen. Hence, a wavespell can be either thirteen days, thirteen weeks, thirteen moons, thirteen years, or so forth. The point is that by making the change of calendars the foundations of the mind organized in time are irrevocably altered. From the noospheric perspective, there is nothing arbitrary about the laws of nature. The same lack of arbitrariness should be reflected in the instruments that we use to measure time. By adapting to the Thirteen Moon/2 8-Day measure, the humans are adapting to a measure in which there is nothing arbitrary because it is a pure expression of the Law of Time. Thirteen is not at all a harbinger of bad luck, but the gateway to harmony in time! And it is by using this thirteen-based harmonic instrument of measure that the human is realized as a noospheric chip, a fractal of the noospheric mind, the mind of time on Earth.

NOOSPHERIC CHIPS: WIZARDS OF GALACTIC TIME

In the noospheric construction of the operating mechanism of the Timeship, the human is a vital link, providing the biotelepathic medium for the synchronization of the geocosmic mind on Earth. If we look again at the structure of the noosphere, the components include: the crystal core bipolar oscillator, packed with Technicolor time beams and cosmic thought moments, waiting to be self-reflectively realized; the human noospheric chip, the representative of the biosphere to mentally engage the timing cycles of the NET; the electromagnetic field, which provides the medium of expression between the biospheric order, the human, and the noospheric mind of time; and the noosphere itself, the mind of time on Earth, whose articulation through the psi bank control panels and through the Earth's resonant field structure yields an enlarged measure of Earth consciousness, Noospheric Earth Time itself. In expanding our minds into the noosphere we will understand that what lives is the Earth and the Earth lives through us.

As a noospheric chip, the human is a fractal reflector of the 13:20 timing frequency. Midway between the bipolar oscillator and the noospheric psi bank plates, the human noospheric chip is designed to receive, transduce, and transmit programs from the crystal core by synchronizing these programs with the psi bank.

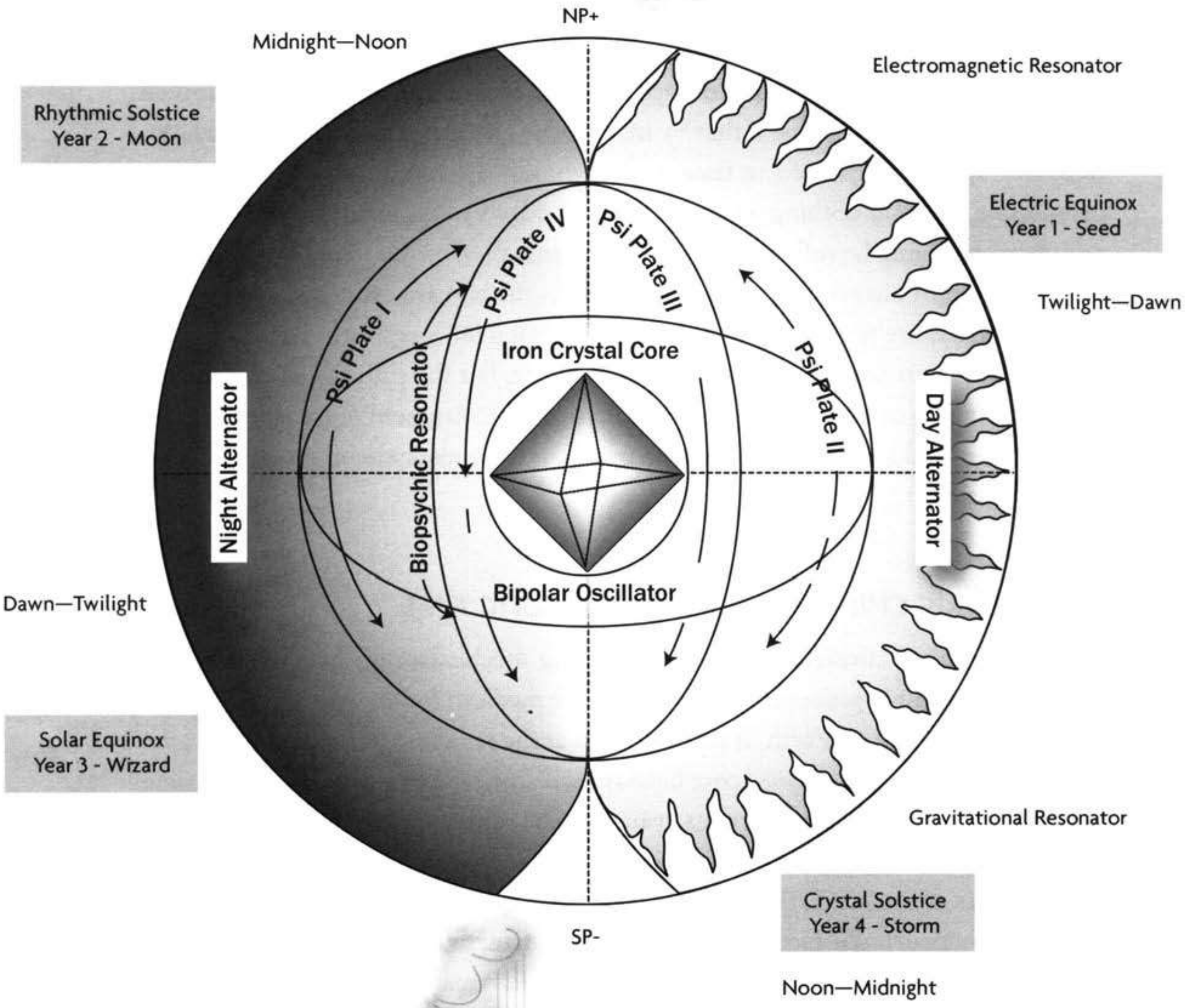
NP+
Midnight—Noon
Electromagnetic Resonator
Rhythmic Solstice
Year 2 - Moon
Electric Equinox
Year 1 - Seed
Twilight—Dawn
Psi Plate IV
Psi Plate III
Psi Plate I
Iron Crystal Core
Psi Plate II
Biopsychic Resonator
Night Alternator
Day Alternator
Bipolar Oscillator
Dawn—Twilight
Solar Equinox
Year 3 - Wizard
Gravitational Resonator
Crystal Solstice
Year 4 - Storm
SP-
Noon—Midnight

Four-Phasic Noospheric Resonant Field Model, Timeship Earth 2013

Three resonators establish the gravitational, electromagnetic, and biopsychic fields held in place by the crystal core, the bipolar oscillator.

Two poles extend the bipolar field of Earth, which, guided by the day-night alternators, create the four-phasic noospheric field model.

Four psi plates govern the four phases, from daily NET minute to seasonal NET day, and to the four-part NET week, or Seed-Storm year bundle.

This is a process that can occur only within the construct of the regular harmonic sequence provided by the noospheric timing gauge of the Thirteen Moon/28-Day calendar. Interaction with the crystal core can only be through telepathic means, and the human-as the midway point between the core and the electromagnetic fields-is like a sensitive piece of litmus paper. To understand the human in this way is to totally redefine the reality of what it means to be human on Earth. From this perspective, the human is a biocomputer whose information source feeds it programs stored and emanated by the crystal core. These programs are then brought into conscious registration through working the corresponding programs of the psi bank. Right now, we don't know what the crystal core is emanating, much less that such a structure exists within the Earth. Like a crystal radio-set, the noosphere is the antenna, the geocosmic programs are transduced and generated at the crystal core, the human is the biocosmic tuner, and the broadcast transmission is provided by the psi bank.

The 13:20 template of the human noospheric chip is defined as the human holon, its twenty digits and thirteen main articulations conforming perfectly to the geocosmic 13:20 synchronization frequency of universal time. It is the intrinsic resonance of this geocosmic structure that allows the human to be in resonance with the planet as a whole system. While the timing programs to be broadcast are all loaded in the psi bank control boards, the actual point in space of the physical Earth is coded by time into a twenty-part icosahedral structure called a planet holon. Being holonomic in nature, the planet holon has a perfect incidence of correspondence with the twenty digits of the human holon.

The twenty digits and twenty-part structure of the planet holon are both coded by the twenty icons or solar seals of the 260-day galactic synchronization cycle. The icons or seals are the manifestations of the 20 (4 x 5) factor of the 13:20 synchronization ratio. They are coded for human legibility as simple icons that conform to a postliterate alphabet. In time these twenty icons create a fourth-dimensional twenty-day solar cycle which interacts with the 13-day galactic cycle to create the 260 unit synchronization sequence or "galactic spin." The point is that when this galactic spin is synchronized with the 365-day solar-lunar measure of the Thirteen Moon count, there is a daily program for identifying the human with the planet holon, a factor that facilitates the planetary resonance of the noospheric chip.

As a noospheric chip, the human is in every way a holonomic reflection of the whole system Earth. As a system of resonators-gravitational, electromagnetic, and biopsychic-the human and the Earth both have bodies with a center of gravity; both exhibit electromagnetic fields; and both share in common a biopsychic resonance, the mutual interface of the noosphere or mind of time with the other two

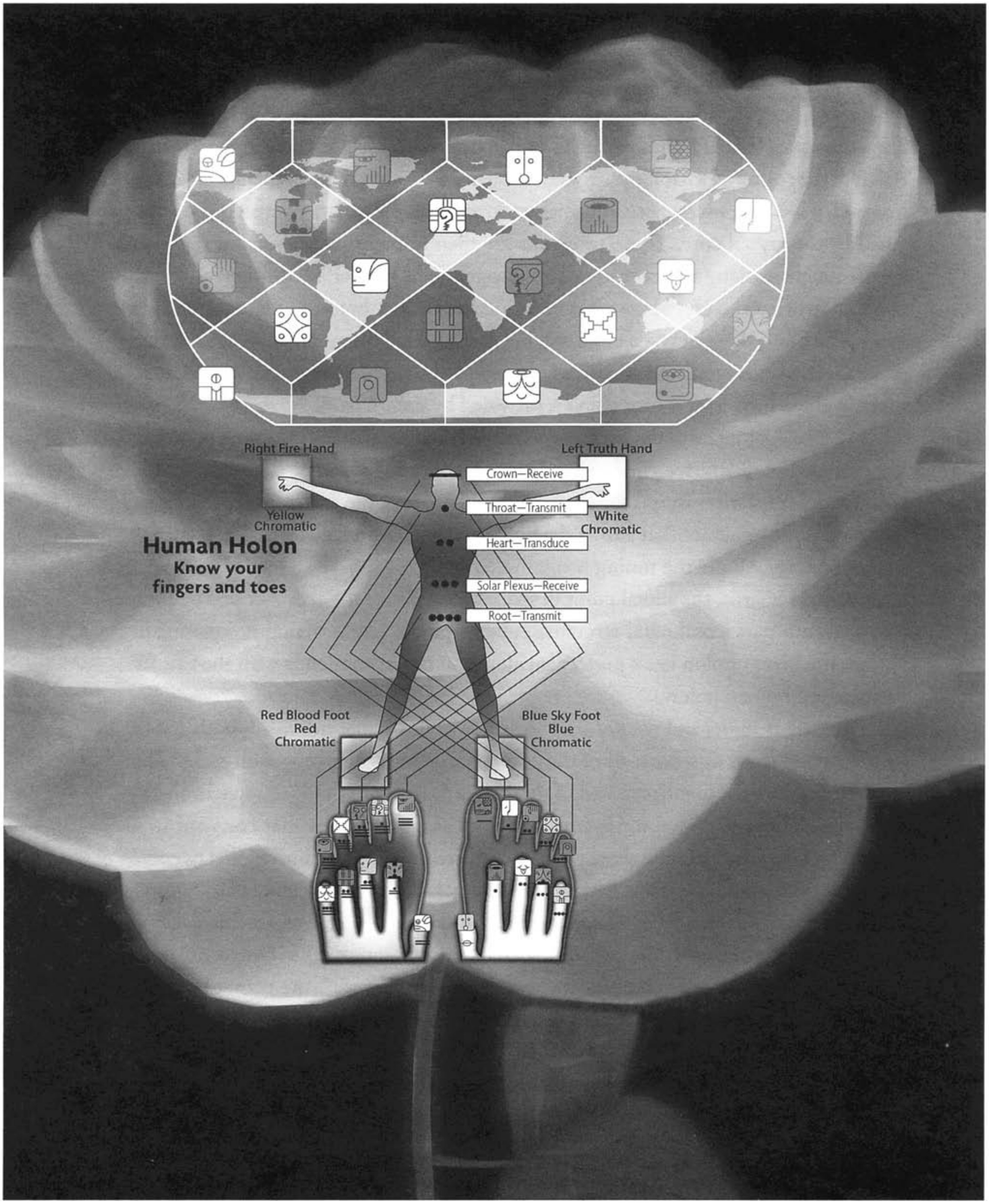
Right Fire Hand
Left Truth Hand
Crown—Receive
Throat—Transmit
Yellow Chromatic
White Chromatic
Human Holon
Know your fingers and toes
Heart—Transduce
Solar Plexus—Receive
Root—Transmit
Red Blood Foot
Red Chromatic
Blue Sky Foot
Blue Chromatic

Planet Holon—Timeship Earth 2013

fields of resonance. The redefinition of the human according to the Law of Time and to the nature of the noosphere also redefines the human's reality. What will the human do when history has been eliminated from the calendar and the new calendar itself is actually operated as a synchronometer-a measure of synchronicity?

First of all, as the necessary prelude to galactic culture, the absolute synchronization of the human in geocosmic time, the technosphere will be dismantled. Reorganized by the Law of Time into psychic biomes, the dismantling will naturally phase into the genuine creation of the new culture of solar-galactic consciousness. "Path beyond technology" means path of telepathic harmony. With its goals absolutely reoriented following the mental pole shift of the calendar change, the hierarchy of human values will also undergo a profound reordering. What if money were no longer the first consideration in human affairs? What if the consideration were instead how artful or artistic that action or event might be? What if the diaspora of humanity through industrialization and mass transport came to an end and the humans willingly chose to live in smaller communities organized by gardening and other Earth-related activities, inclusive of the transformative dismantling of the technosphere and the regeneration of the biospheric medium? What if the study of time replaced the study of physics as the dominant mode for determining the nature of reality? And what if, due to the time shift, a spiritual awakening swept through the human race so that a moral orientation of values was organized around the preservation and furtherance of the biosphere, and an acknowledgement that only One Supreme Creator could have ordained something so perfect as the noospheric mind of time? What kind of world would that be? What kind of art would be practiced in that world and what kind of science would be known?

Organized by the NET, human existence would spiral into the collective telepathic mind. A type of Earth geomancy would become the science of the noosphere. Or perhaps a more accurate name for this science would be *Earth geochronomancy*-the study and investigation of Earth in relation to the cycles of cosmic time and telepathy with an end to perfecting the Earth as a work of art. With minds expanded toward accommodating the longer, slower measure of NET, geochronomancy would inspire an entirely new vision of the Earth within the heliosphere (the solar system) and the galactosphere (the galactic order of things). Making the Earth's rainbow brain visible and manifest would be only the beginning of a willingness to explore further untapped psychic energy sources and means, or what is referred to in *Earth Ascending* as the science of radiosonics. Key concentrations of human habitation and activity would be organized to occur at or close to those geomagnetic energy points or centers, the whole cultural process of which would correspond to making a map on Earth of the image of galactic being as it moves in its great patterns of time. Would this not make the human species into a race of wizards?

In this consideration we may find a reason why the 260-day unit galactic synchronization measure-a galactic spin-is the key to what is called the *Wizard's Count.* The perfect synchronization of the 260-unit galactic spin with the 365 -day Thirteen Moon/28-Day solar orbit establishes the 52-year cycle, or one NET year. A NET year (fifty-two solar orbits) is a "wizard's measure," for on this measure the human can expand its consciousness to encompass the Earth in its greater cycles of development. The Wizard's Count is also given its name because of the set of the four solar icons or seals that code the July 26 synchronization date in 4-year cycles: year one, Yellow Seed year; year two, Red Moon year; year three, White Wizard year; and year four, Blue Storm year. An entire Seed-Moon- Wizard-Storm sequence of four NET days (four solar orbits) consists of fifty-two moons, or one NET week, the perfect fractal of the fifty-two 7-day weeks in one NET day and the fifty-two solar orbits in one NET year. This four-part sequence is also called a Seed-Storm year bundle, the harmonic essence of the wizard's measure. Thirteen Seed-Storm year bundles, where each of the four icons is coded by the thirteen galactic tones (4 x 13 = 52), yield one NET year. Please note that even the Gregorian July 26 synchronization date is a function of the synchronic order: July = 7, while 26 is 13 doubled, 7 and 13 being the two principle numbers governing the 13:20 matrix of the Harmonic Module!

In the symbolism of the iconic sequence it is the Seed that is planted, it is the Moon that coordinates the growth cycles, it is the Storm that generates the energy, and it is the Wizard who knows and acts upon what is given. It was a White Galactic Wizard that coded the date and coordinated the year beginning July 26, 1987-followed twenty-one days later by the Harmonic Convergence. Thus it was the White Galactic Wizard who opened the gates to the closing of the Mayan baktun cycle of history and the commencement of the coming cycle of galactic culture. And on July 26,2003, the year of the White Spectral Wizard will commence. This year will end with its Day Out of Time, coded by the White Spectral Mirror. The day following, July 26,2004, will begin the year of the Blue Crystal Storm. On the Wizard's Count, which is not a calendar but a synchronometer, these two dates, White Spectral Mirror and Blue Crystal Storm, are the point in time when the Gregorian calendar ticket to the old time will be eliminated and the new Thirteen Moon calendar ticket for admission to the new time of the noosphere will be validated. This is when the noospheric show will be ready to begin its prelude. (See plate 6, Noosphere-Mind of Time on Earth.) Will the humans also be ready?

▼

8

Making the Transition to the New Time

A How-to Catalog and Guide to the *Pax Cultura, Pax Biospherica*

BY THE TIME these words are read, the technospheric collapse will have generated several more devastations and disasters. The wound received by the technosphere on September 11 was a mortal one. The technosphere suffered cardiac arrest for almost a week. Not only did the towers collapse, an event that occurred via television within the central nervous system of the human species, but belief systems, too, came crashing down. No longer is America the unassailable fortress of "democracy," while democracy itself is a concept that joins monarchy, nation states, imperialism, colonialism, capitalism, communism, and a host of other outmoded systems of belief and thought constituting the saga of human history-that is, the history of civilization. But all that is now ending. The perimeter of biogeochemical combustion and the limits of the technosphere have coincided. It is the Earth's turn to speak. How will the Earth speak?

The Earth can only speak through the noosphere, and the noosphere can only speak from the center of time that is constructive peace. Peace is harmony. Peace is the culture of the biosphere. The conscious harmony of time in the biosphere is the culture of peace known as the noosphere. Through the noosphere, the Earth speaks in ever-widening circles of order and meaning. While sunrise and sunset mark the passage of solar time, observing the phases of the moon keeps us in tune with the mystery of time. One kin, one passage of night and day, one NET minute, synthesizes the experience of the whole Earth during a single rotation on its axis. The synchronic order supplies a 260-unit synchronization gauge that the mind may engage and employ in the construction of a harmonic vision of reality. Twenty-eight NET minutes make one moon, one NET hour, moon after moon. The pattern of one moon contains the construct and pattern of every moon. The harmonic perfection of the synchronic order of reality releases the mind through the center of time into a freedom it had not known in the dark passage of history. How are we going to make this transition? Can we really shake off the old and the obsolete that easily?

THE THIRTEEN MOON CALENDAR MOVEMENT: THE DAWN OF A NEW GALACTIC CULTURE

"Greetings Explorers on Timeship Earth!" reads the opening text of the pocket-size *13 Moon Natural Time Journal.* "Welcome to this 13:20 natural time portal and fourth-dimensional galactic time template. A solar-lunar Earth calendar of the 13:20 timing frequency that is expressed through the thirteen tones of creation and the twenty sacred solar tribes. Attuning to this calendar will help you awaken and align to the pulse of the Universe, the rhythm of the One heartbeat. 13:20 is the natural timing frequency that unifies planetary and galactic consciousness. . . It is the expression and remembrance that Time is Art . . . By attuning to the 13:20 frequency of natural time, we break out of the 12:60 time warp and activate the remembrance of our true human nature. Time is energy. It is a frequency of the fourth dimension; mind, telepathy, imagination, and creation. Time is an exploration of unlimited Universal potential. So let us begin the journey and return to the rhythm of nature's galactic heartbeat. . ."1

This is the Earth speaking through a noospheric chip. Produced in Canada, it is one of many Thirteen Moon calendars produced this Yellow Solar Seed year. Others have been produced in Germany, Austria, Italy, Russia, the Netherlands, Japan, Brazil, Argentina, Chile, Mexico, and the United States. One such example from the United States is known as the 13 *Moon Natural Time Calendar: Ancient Science for the Art of Now.* Going through its pages one is flooded with a wealth of information, summarized by a quote from traditional Hopi wisdom: "We now face the final test of human consciousness-to restore our oneness with creation. We are going home. It is a process. It is like the seed becoming the green shoot and then the flower."2 To make the calendar change is to restore the oneness with nature. The break with the oneness with nature was codified and thence programmed into an irreversible reality by the Gregorian calendar and mechanical clock. Only the breaking of the pattern of time established by these two instruments will release us again into natural time. The Thirteen Moon/28-Day calendar exists as the tool to demonstrate the contrast between harmony and disharmony, and to assist us, as a species, into effecting this return. If we are going home then we must make a turn away from our present course.

> Through returning to natural Time, we return to our natural state of divinity, living in harmony with the Earth and each other. We return to the nature of our spiritual existence. Realigning with the rhythm of the universe through the 13 Moon Calendar allows us to lay the foundation for the new paradigm of love and unity, replacing the 12:60 paradigm of fear and separation. . . The World Calendar Change Peace Plan strives to establish a new culture of peace on Earth through returning humanity to natural time. It is the goal of the worldwide Foundation for the Law of Time to effect the calendar change by the Day Out of Time (July 25), 2004 . . . Join the global movement to effect planetary change and establish the new culture of peace on Earth! Return to divinity and creative existence. . . Walk the path of the sacred warrior.]

The movement to change the calendar is a living force, a force that is living the New Time. Its premises are so diametrically opposite to those of the technosphere that, especially in the power bases of America and the G-7, one will find scant mention or notice at all of this movement. Yet it is a cultural phenomenon. In Tokyo, San Francisco, Milan, or Moscow one may find signs of the presence of this movement in the underground dance events, the raves or the Earthdance. More and more young people take on the names of their galactic signatures, the 260-day kin code for their date of birth in the synchronic order, names like Lunar Storm or Spectral Night, Galactic Wind or Planetary Dragon. These are all signs of the beginning of galactic culture on Earth, a phenomenon that has only been increasing over the past decade. The point is that the New Calendar Change Movement is a grassroots phenomenon as well as an ideological strategy known as the Campaign for the New Time. These signs are all whispers of the noosphere blowing through the bleak industrial corners of the technosphere.

Behind the more populist front of the Natural Time Movement is the science of the Law of Time, a genuinely noospheric science of peace and harmony. This science is already being studied and put into preliminary practice. As the crisis of the biosphere began to crest in 1987, halfway through the second 28-year phase of the technosphere, the noosphere began to download its measure of time through the discovery of the Law of Time. Invisible to the predominant forces of the technosphere, the science of the Law of Time began to map out an entirely new domain of thought, the synchronic order, a domain that could only be entered and applied through the correct noospheric measure, the Thirteen Moon/28-Day synchronometer.

All science throughout history is based on evolutionary advances predicated on both the premises and limitations of its preceding stages. We are at such an evolutionary moment now. The issue of the calendar change is only the beginning of the emergence of a new whole system scientific order. Understanding the flaws that are built into the present system of purely materialist science, we can proceed to the establishment of the science of the synchronic order. Like the radical simplicity and perpetual harmony of the Thirteen Moon calendar itself, the noospheric science of the synchronic order is predicated on an equally radical mathematical base reflecting and embedded in the 13:20 universal timing frequency itself. And as the underlying order of a whole system science, this mathematics is reflective of the non dualistic premise of a superior moral order. The deviation from universal synchronic time is a moral deviation as well. It is this moral deviation that most strongly prompts the current calendar change movement.

As a profoundly nondualistic, unifying whole system premise, the Law of Time encompasses not only a scientific-mathematical basis of morality, but a theological-eschatological intention that reunifies the various disciplines scattered and separated by the rise and totalitarianism of modern secular science. Because of the historical dominance of the Gregorian calendar, in which modern science is embedded, the dethronement of the calendar-the "devaticanization" of the world, as it were-is an apocalyptic moment.

As the Law of Time defines it, this is merely the apocalypse of the 12:60 timing frequency and all its corrupted institutions, much as foretold in the Book of Revelations. This apocalyptic moment was stunningly supplied by the terrifying collapse of the Twin Towers. As the nerve center of the financial, commodities, and transport sector of the technosphere, and hence of the city itself, the collapse of the towers in hardly more than an hour's time is perfectly reflected in the vivid descriptions of the eighteenth chapter of the Book of Revelations, "Fall of Babylon." Yet as the Mayan prophecy of Pacal Votan also makes absolutely clear, the apocalypse of the 12:60 is the resurrection into the 13:20 timing frequency, the return to natural time, all to be accomplished by A.D. 2012, Teilhard de Chardin's "omega point." The advent of the noosphere is preceded by the post-apocalyptic harrowing of hell. In this harrowing of hell, defined as the span of time fromJuly 26, 2000-July 25,2004, humanity must make a crossing from one time to another. This crossing is much like those we hear

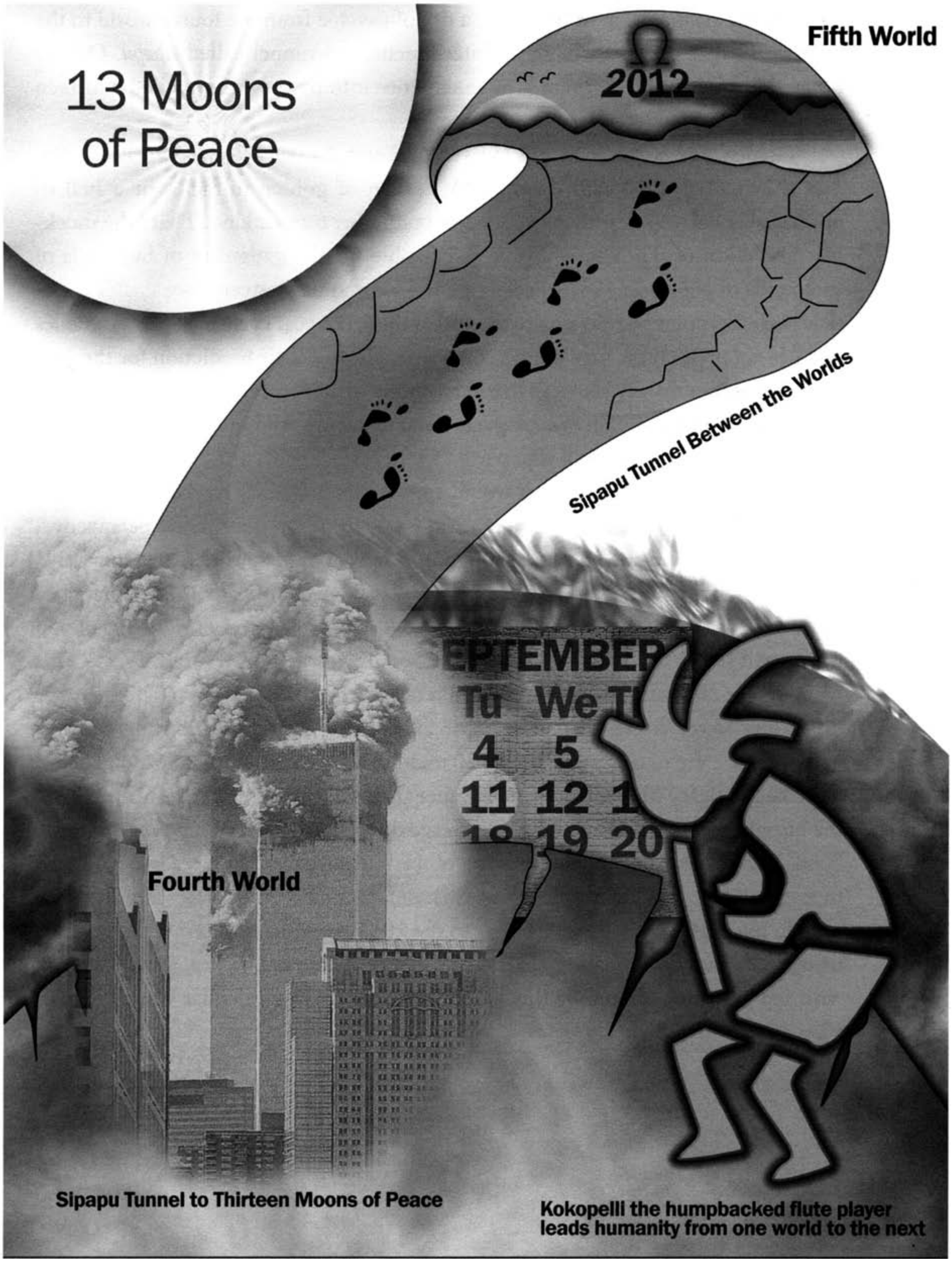
13 Moons of Peace
Fifth World
2012
Sipapu Tunnel Between the Worlds
EPTEMBER
Tu We Th
4 5
11 12 1
18 19 20
Fourth World
Sipapu Tunnel to Thirteen Moons of Peace
Kokopelli the humpbacked flute player leads humanity from one world to the next

about in the cosmovision of the Hopi, a rite of passage from the fourth world to the fifth world that occurs through an interdimensional tunnel called *sipapu.* On the other side of that tunnel, there is an emergence into the new time called "thirteen moons of peace."

"Golden Age of Miracles and Heaven on Earth," a story in the *National Examiner* (November 29, 1988), begins: "A heaven of golden miracles or a hell of unspeakable horror is less than twenty-five years away for mankind. That's the shocking prediction of the man who was responsible for the gathering of hundreds of thousands of people in 1987 to celebrate the Harmonic Convergence. . ." The article makes its point by referring to the end of the 5,125 year cycle in 2012. "Argiielles said Mayan predictions have always come true. He said their prediction for the year 2012 is that if humanity learns to live in harmony with Nature, a new age of miracles and light will occur. Otherwise plagues, wars, terrorism, and finally, the breakup of the planet into asteroids, will take place."4

As for the Inevitable Event, it would appear that humanity has begun the worst of the prediction-plagues, war, terrorism-yet, from the noospheric perspective, two points mitigate the absolute plunge into hell. The first is that the Inevitable Event punctured a hole in the technosphere eleven years prior to 2012. What does this mean? It means that this is the time of hell on Earth. "Every single one of you must see it [hell]; this is an irrevocable decision of your Lord." (Quran, 19:71) This being so, we can potentially emerge from it before the 2012 omega point. The other mitigating factor is the existence of a 13:20 movement dispersed throughout humanity. Since the noosphereis the 13:20 mind of Earth, the Thirteen Moon Calendar Change Peace Movement is the advance presence heralding the transition from the biosphere to the noosphere. This passage through hell is of a four-year duration, 2000-2004. The exponential implosion of the technosphere has already begun. But to avoid a long, debilitating siege of barbarism, humanity must select a positive direction by 2004 so that it can make it to the 2012 deadline as a species returned to living in harmony with nature. This it can only do by the dramatic and radical break with the past time offered by the Thirteen Moon/28-Day calendar Change Peace Plan.

Campaign for the New Time

4:7::7:2013

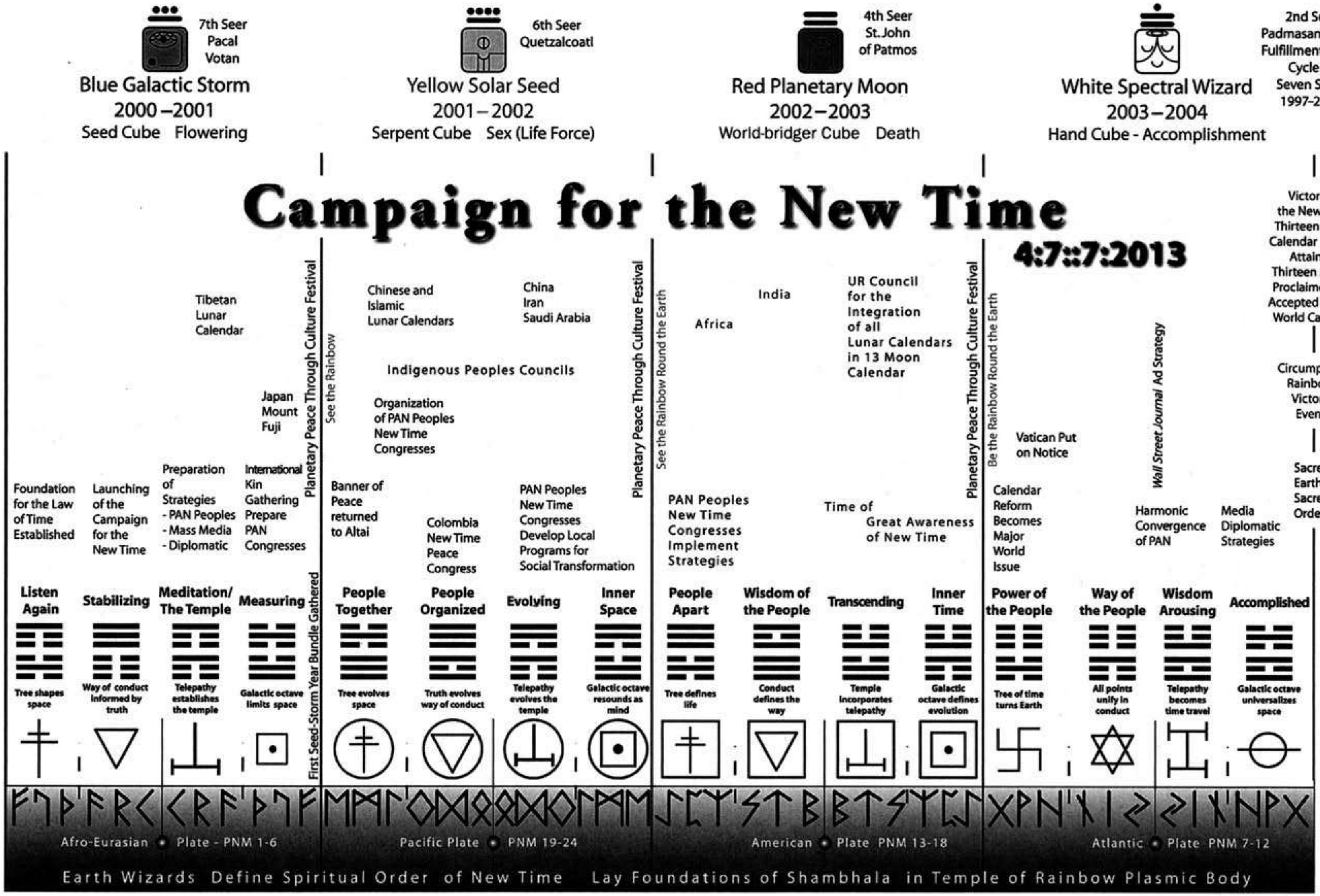

7th Seer
Pacal
Votan

Blue Galactic Storm
2000 – 2001
Seed Cube Flowering

Tibetan Lunar Calendar

Japan Mount Fuji

Foundation for the Law of Time Established

Launching of the Campaign for the New Time

Preparation of Strategies
- PAN Peoples
- Mass Media
- Diplomatic

International Kin Gathering Prepare PAN Congresses

Planetary Peace Through Culture Festival

First Seed-Storm Year Bundle Gathered

Listen Again
Tree shapes space
Stabilizing
Way of conduct informed by truth
Meditation/ The Temple
Telepathy establishes the temple
Measuring
Galactic octave limits space

6th Seer
Quetzalcoatl

Yellow Solar Seed
2001 – 2002
Serpent Cube Sex (Life Force)

See the Rainbow

Chinese and Islamic Lunar Calendars

China
Iran
Saudi Arabia

Indigenous Peoples Councils

Organization of PAN Peoples New Time Congresses

Banner of Peace returned to Altai

Colombia New Time Peace Congress

PAN Peoples New Time Congresses Develop Local Programs for Social Transformation

Planetary Peace Through Culture Festival

People Together
Tree evolves space
People Organized
Truth evolves way of conduct
Evolving
Telepathy evolves the temple
Inner Space
Galactic octave resounds as mind

4th Seer
St. John
of Patmos

Red Planetary Moon
2002 – 2003
World-bridger Cube Death

See the Rainbow Round the Earth

Africa

India

UR Council for the Integration of all Lunar Calendars in 13 Moon Calendar

PAN Peoples New Time Congresses Implement Strategies

Time of Great Awareness of New Time

Planetary Peace Through Culture Festival

People Apart
Tree defines life
Wisdom of the People
Conduct defines the way
Transcending
Temple incorporates telepathy
Inner Time
Galactic octave defines evolution

White Spectral Wizard
2003 – 2004
Hand Cube - Accomplishment

Be the Rainbow Round the Earth

Vatican Put on Notice

Wall Street Journal Ad Strategy

Calendar Reform Becomes Major World Issue

Harmonic Convergence of PAN

Media Diplomatic Strategies

Power of the People
Tree of time turns Earth
Way of the People
All points unify in conduct
Wisdom Arousing
Telepathy becomes time travel
Accomplished
Galactic octave universalizes space

2nd Seer
Padmasambhava
Fulfillment of the Cycle of Seven Seers
1997–2004

Victory of the New Time Thirteen Moon Calendar Reform Attained Thirteen Moons Proclaimed and Accepted as New World Calendar

Circumpolar Rainbow Victory Event

Sacred Earth Sacred Order

Afro-Eurasian Plate - PNM 1-6
Pacific Plate PNM 19-24
American Plate PNM 13-18
Atlantic Plate PNM 7-12

Earth Wizards Define Spiritual Order of New Time Lay Foundations of Shambhala in Temple of Rainbow Plasmic Body

THE BIRTH OF THE CAMPAIGN FOR THE NEW TIME

> *"We inspired Moses: 'Lead My servants out, and strike for them a dry road across the sea. You shall not fear that you may get caught, nor shall you worry.*
>
> *"Pharaoh pursued them with his troops, but the sea overwhelmed them, as it was destined to overwhelm them."*
>
> -QURAN, 20:77-78

The Thirteen Moon/28-Day calendar is the dry path across the waters of the 12:60 chaos. Holding surely to this path, "Pharaoh's troops will be overwhelmed," as it is destined, for the truth always prevails and when falsehood is exposed, falsehood perishes. The 12:60 frequency enshrined in the Gregorian calendar represents an artificial, hence false notion of time. The Pharaoh's troops are the forces of the technosphere that, embedded in the artificial structure of the Gregorian calendar, are destined to be overwhelmed and overcome by the fury unleashed both by human nature and nature itself in response to decades, if not centuries, of irresponsible despoiling of the Earth. The dry path of the Thirteen Moon/28-Day calendar is the only path broad enough for all humanity to leave the ever increasing ocean of chaos that characterizes human existence bound in the collapsing 12:60 matrix born of Gregorian civilization, the technosphere. Yes, the dry path across the waters of the 12:60 chaos provided by the Thirteen Moon/28-Day calendar truly is for the salvation of life on Earth.

Ever since 1993, the Thirteen Moon Calendar Change Peace Movement has consciously been drawing from the infinite riches of the noosphere to create not only the mechanism for making a shift in the frequency and hence bring stabilization to the biosphere, but has been anticipating the biosphere-noosphere transition into the *Pax Cultura, Pax Biospherica.*

Following in the footsteps of its predecessor, the new calendar movement has dutifully contacted the United Nations and the Vatican and convened various congresses and delivered ultimatums in an effort to draw attention to mankind's time malaise. Now that the Inevitable Event has fully disengaged the human mind from the reality it thought it was pursuing in the new millennium, the second definitive wave of the World Thirteen Moon Calendar Change Peace Movement has also been launched: the Campaign for the New Time.

New Globally Unifying Harmonic Standard

FOURTH DIMENSION: Pure whole number mathematical order—syntropy, synchronic order

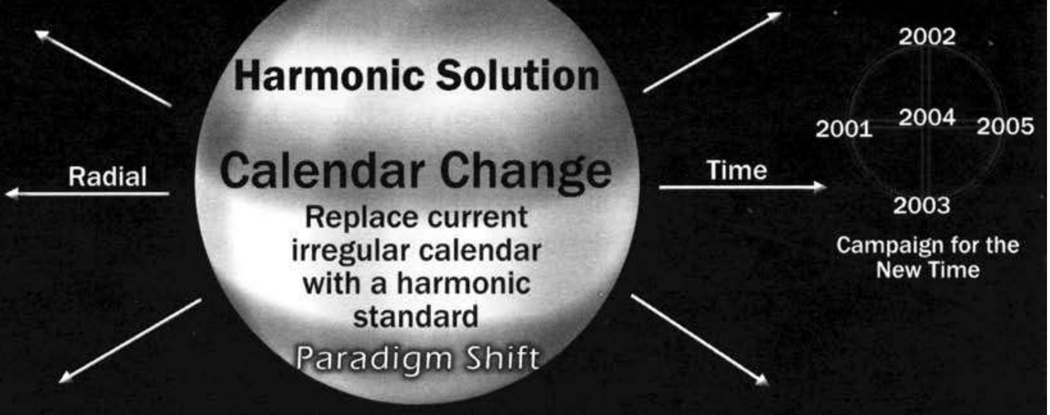

Harmonic Solution

Calendar Change

Replace current irregular calendar with a harmonic standard

Paradigm Shift

Radial

Time

2002
2001 2004 2005
2003

Campaign for the New Time

13 Moons. 28 Days.

Regular Units of Measure

28+28+28+28+28+28+28+28+28+28+28+28+28+1 = Harmony

"Mental barrier" separating fourth-dimensional realm of conscious wholeness of synchronic order from third-dimensional realm of "hard science" physical plane of reality and materialism.

Line of inexorable technological progress: neoliberal economics, globalization = biospheric instability and destruction.

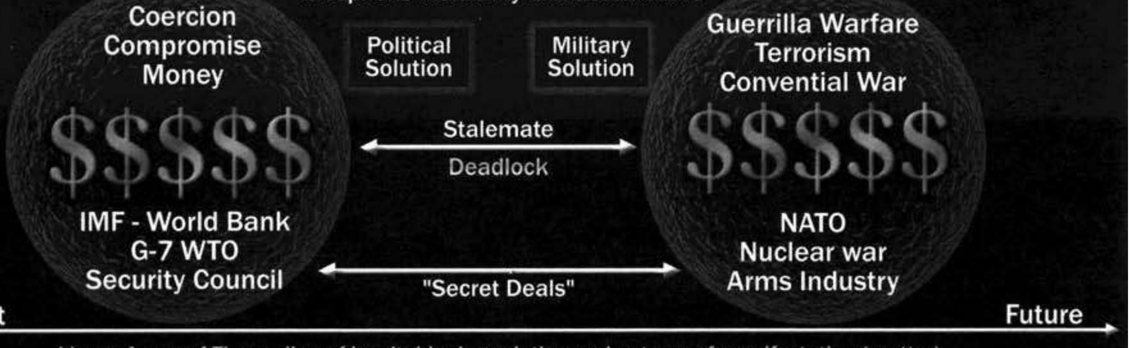

Coercion
Compromise
Money
$$$$$
IMF - World Bank
G-7 WTO
Security Council

Political Solution

Military Solution

Stalemate

Deadlock

"Secret Deals"

Guerrilla Warfare
Terrorism
Convential War
$$$$$
NATO
Nuclear war
Arms Industry

Past

Future

Linear Arrow of Time = line of inevitable degradation and entropy of manifestation (matter)

THIRD DIMENSION: Manifestation-matter entropy and disorder

Gregorian Calendar Time Line = macro-organizing timing standard establishes irregular measure as basis of all dominating concepts of linear time, disorder, entropy, and the techno-economic inexorability

Irregular units of measure:
31-28(29)-31-30-31-30-31-31-30-31-30-31 = Disharmony

If today is Sunday, July 8, what day of the week will August 8 be?
What is the meaning of the names of the months?
If *September* means "seven" why is it the ninth month?
August is named after Augustus Caesar, July for Julius Caesar, March for Mars, the god of war;
could this be a reason why we are programmed for chaos and violence? Think about it.

This campaign consists of a preparatory four-year cycle, 2000-2004, the Harrowing of Hell, which is the duration of the campaign itself, plus a one-year trial period, 2004-2005. The goal of the four-year campaign is universal calendar reform, the replacement of the Gregorian by the Thirteen Moon Calendar, effective Blue Crystal Storm, July 26,2004. This final year of the Blue Crystal Storm, 2004-2005, also completes a Seed-Storm year bundle begun in 2001, fateful year of the Yellow Solar Seed, the year of the Inevitable Event. To ensure the success of this unprecedented move on the part of humanity, the calendar reform must include the one-year trial period, during which time the entire species must be operating by a new standard, the harmonic Thirteen Moon/2 8-Day synchronometer. "If you never try it, you'll never know what it is," is the slogan that describes the campaign's target year: July 26, 2004-July 25, 2005, an entire year without the Gregorian "30 days hath September. . . ," but instead, Thirteen Moons of Peace.

To realize its ambitious goal of universal calendar reform, the Campaign for the New Time must impress itself on the human consciousness as the all-unifying campaign for peace to overcome the destructive forces unleashed by the technospheric collapse. Fortunately, much groundwork has already been laid. The codes of the Law of Time have been formulated and compiled. Preliminary practices to engage the synchronic order of the noosphere have already enlisted humans from around the planet. What the Vatican has tried to ignore and what the United Nations has been unable to fully engage must now be brought to the forefront of all humanity as the harmonic solution. The harmonic solution is a set of logical arguments that should be studied and rehearsed by everyone who seriously wishes to even consider the possibility of a Campaign for the New Time. Let us contemplate the following set of logical arguments.

Premise

Currently all conflict resolution of a global scale is attained through one of two poles: a political solution or a military solution. No other alternatives are recognized. According to the Law of Time, there is a third solution, the harmonic solution. Without this third solution, no harmonic resolution is ultimately possible for any conflict on a global scale. The third solution defines the first two solutions as a polarity embedded within a timing standard that is hopelessly irregular and that, therefore, disposes the mind to disharmony. A mind predisposed to irregularity and disharmony is incapable of producing any long-term solution. The basis of the harmonic solution is to change the timing standard from an irregular to a regular one, thereby inducing a timing sensibility that disposes the mind to harmony rather than disharmony.

Thirteen Moon/28-Day Count and Global Conflict Resolution

The inability to attain global peace, much less the resolution of many localized conflicts that have become endemic, is due to a lack of a globally unifying timing standard-the macro-organizing principle of human sensibility. Present-day conflict resolution is represented by two poles, political and military. The military solution consists of guerrilla warfare, terrorism, and outright military conflict, all of which are engendered by problems of inequality; injustice; rampant nationalism; tribal, racial, or religious antagonism; and greed for control of natural resources. Underlying the military solution is an aggressive armaments industry that seeks to hold its own as a capitalist venture, always seeking new markets. The political solution is always fostered as the only cure for the military solution, and is administered through varying degrees of coercion, compromise, and the inevitable lure of money-the IMF and the World Bank are never far behind any political solution.

Neither the military nor the political represent any kind of morally satisfactory or enduring solution because both are functions of and embedded in a timing standard that has disharmony built into it. Even the local timing systems, themselves feedback loops holding in place various belief systems, are coordinated by the dominant disharmonic standard. As long as the current global standard, the Gregorian calendar, is the controlling timing factor, resolution on a global and local level is not only impossible, but increasing disharmony is guaranteed. The harmonic solution calls for the immediate establishment of a globally unifying harmonic standard, the Thirteen Moon/28-Day count. With its absolutely consistent units of measure, this macro-organizing principle is a paragon of harmony and thereby represents the positive, creative, and viable first step to enduring global conflict resolution.

Thirteen Moon/28-Day Paradigm Shift

The harmonic solution, the alternative to the polarity of the military and political solutions, calls for the immediate replacement of the current irregular measure with the harmonic standard of the Thirteen Moon/28-Day count. This replacement constitutes the long-awaited and long-prepared-for paradigm shift in human consciousness. The very analysis and concern that engendered the first movement to replace the Gregorian calendar-the effects of an irregular measure on human behavior-is now so dire that the concluding step must be taken. To quote Einstein once again, "Since the advent of the nuclear age, everything has changed but the way people think; thus we drift toward unparalleled catastrophe." With the analysis of the Law of Time we know why thinking has not changed-the human mind is the pawn of an aberrant timing mechanism. To change the way people think, make one little change in their everyday lives: Change the calendar. If we could send a man to the moon, we can change the calendar. This change will be the first step-the rest will come easily because with this one change we shall have changed our minds and entered into a New Time. To make this one little change a target date is needed, and the people need to mobilize as a force of synchronization for peace to mandate this change. The target date is July 26, 2004--the mobilization has already begun.

New Worldview of the Law of Time

The Thirteen Moon/28-Day count is the pragmatic application of the Law of Time. The discovery of the Law of Time was necessitated by evolution as the only way that could offset the deleterious effects of an erroneous timing standard on human consciousness. The Law of Time takes the radical view that time is a frequency, the universal frequency of synchronization; that the purpose of time is to synchronize; and that, therefore, a calendar is in harmony with the Law of Time to the degree to which it maximizes opportunities for synchronization. From the perspective of the Law of Time, the current macro-organizing standard is anything but synchronous, and hence contrary to the laws of nature. Synchronization produces harmony, harmony is a function of beauty, and beauty is the natural result of the Law of Time: T(E) = Art, energy factored by time equals art, where art is the entire spectrum of phenomenal reality in which even the most menacing-looking creature is elegantly constructed.

National Geographic recently stated: "Beauty seems to be an intrinsic part of nature and perhaps even the organizing principle of reality. Scientists in testing their theories, invariably find that the simplest, most elegant, most beautiful is the correct one. Rainbows, butterflies and the periodic table are some examples of intrinsic beauty. The world will be saved by beauty."5 The Law of Time agrees. The world will be saved by beauty, and the proponents of peace through culture will be in the vanguard. The Law of Time formalizes the perception of the intrinsic beauty of nature- T(E) = Art-and thereby provides a scientifically solid critique of the ugliness and social disharmony fomented by the irregular and artificial timing standards that control humanity today. By the same formulation, the Law of Time also establishes a vast and comprehensive definition of the fourth dimension, which was wanting until now precisely because of the embedding of all current perceptions in the erroneous timing standard.

The mathematical order underlying the universal frequency of synchronization, 13:20, describes a new order of reality, the synchronic order. The synchronic order governs the fourth-dimensional order of reality much as the laws of gravity, thermodynamics, and the special theory of relativity govern the third-dimensional order of reality. By taking into account the nature of the synchronic order, humanity's views of the third dimension will inevitably be altered to a greater conformity with the harmony of the fourth dimension. The fourth dimension is defined as the primary mathematical order of time, governed by a syntropic whole number system of ratios and fractals operating in a radial and nonlinear manner.

The third dimension is the world of manifestation and matter that inevitably entropizes. Losing sight of the inherent harmony of the synchronic order from which it is actually inseparable, the human fixation on the increasing entropic anomalies of the third dimension produces problems such as how to reckon what appear to be the extra hours, minutes, and seconds that occur every time the Earth orbits the sun. To this seeming dilemma, the Law of Time asserts that number precedes manifestation, harmony precedes materialization, therefore manifestation must accord with number, and materialization must reflect an original harmonic order. Entropy and degradation cause manifestation to deviate from number and mathematical law. But for mind to be stable, mind must accord with and conform to mathematical law rather than with entropy and degradation. The timing standard in use causes the mind to conform with entropy and degradation. This is what must be corrected.

Harmonic Solution-Harmonic Resolution

Harmony can only produce harmony. From order comes order. From disorder you can only expect chaos. From disharmony can only come disaster. Irregular measure entropizes. Harmonic measure syntropizes. These are the contrasting values brought into focus by replacing an irregular macro-organizing principle with a regular one. Implementation of a globally unifying harmonic standard is the only hope for resolving all conflicts in time. A harmonic standard lifts all conflicts to a higher level dialog of unification in time. The harmonic solution, the World Thirteen Moon Calendar Change Peace Plan, provides: the opportunity to call an immediate and universal cease-fire, as well as a halt to the otherwise inexorable process of entropic degradation of society and environment; and offers the resolution to establish new agreements, to renegotiate old treaties, and to create a new globally binding and unifying covenant for all people in a new time.

It must be understood that to change the calendar is to delegitimize all the institutions embedded in it. This radical but peaceful delegitimization of the old order is what is needed to break the deadlocks, stalemates, and historically outmoded treaties and agreements now binding on humanity. The new globally unifying standard, the Thirteen Moon/28-Day count, will legitimize and institutionalize harmony and thus create the advent of a synergistic order capable of resolving entropic disorder because it is reflective of and rooted in the primordial whole number structure of the synchronic order of reality. In its entirety, this represents a new step and direction for human evolution-a step without which there might not be any creative future for humanity and the planet.

How to Accomplish World Peace Now

The analysis, the historical precedents, and now the affirming authority of the Law of Time render the harmonic solution as the only way to immediately bring about World Peace. All that is lacking is the will to synchronization among all good-hearted people, groups, and organizations otherwise committed to peace, a better environment, and human spiritual well-being in general. How to accomplish World Peace is to galvanize the human will to synchronization on behalf of the establishment of a genuine time of harmony and peace. This can only come about through the simple but highly radical and paradigm-altering act of exchanging the current macro-organizing system of the Gregorian calendar for the harmonic order of the Thirteen Moon/28-Day count. All other localized calendar systems would be kept intact, but they would be coordinated by a new standard, a standard that promises harmony instead of disaster, order instead of chaos, peace instead of war.

Such is the harmonic analysis and solution of the Law of Time, which can be summed up by a simple demonstration of the irregular Gregorian calendar contrasted with the regular Thirteen Moon calendar, as well as the question: "Which do you prefer, harmony or disharmony?"

If you answer this question on behalf of harmony, and you agree with the foregoing premises and analyses, you must move into action. You need to get a new calendar, a Thirteen Moon/28-Day calendar and find out for yourself: noosphere or necrosphere, the choice is yours. The option opened by the Inevitable Event and the collapse of the technosphere is precisely that: necrosphere, the transformation of the biosphere into a sphere of death-a dead planet-or the noosphere, the mental envelope of the Earth, governed on the human plane by the Thirteen Moon/28-Day calendar. The difference between the necrosphere and the noosphere is the difference between a machine and the Earth as a whole system. You cannot even really say there is a comparison. But what we can say most emphatically is that the Thirteen

Moon/28-Day calendar and the shift to the 13:20 frequency is the pure manifestation of the Law of Time in human affairs. It is the only antidote to the destructive grip that the 12:60 timing frequency now has upon the biosphere through its thoroughly artificial instrument, the technosphere. (See plate 7, Necrosphere or Noosphere.)

THE FINAL DECISION: NOOSPHERE OR NECROSPHERE?

As these lines are being written, seven weeks have passed since the Inevitable Event. The analysis of this event from the perspective of the biosphere, the temporary structure of the technosphere, and the advent of the noosphere was a spontaneous inspiration. Each morning, often well before dawn, I would be called to this technospheric device, an iMac, to put the noospheric mandate down into words. In this process, the matter of prophecy kept occurring and could not be avoided. Even the issue of calendar change increasingly takes on the quality of a prophetic enactment. What is prophecy from the noospheric perspective?

Prophecy is the release of information according to the psi bank timing program and in relation to degrees or levels of consciousness positioned at different points in time. The points in time are human channels. That is, there are the prophets who declare the prophecies by whatever means or forms, and then there are the decoders of the prophecies, those able to read the signs according to the prophecies with the aid of different symbolic lexicons. Prophecy is always startling to the mass mind, which is ignorant of the true nature of time. A prophetic event may seem totally random, when it is not. Someone may design a plot, such as that of the Inevitable Event, with whatever fantastic notions and motives. Yet the occurrence may go beyond being merely a matter of human manipulation. The humans were but the instruments of sets of cause-and-effect chain reactions established within the structure and evolution of the biosphere and the playing out of human karma.

According to the psi bank timing program governed by the Law of Time, the technosphere could only have a limited duration. Two perfect 28-year cycles of the Gregorian calendar program, leading to the first year of the third millennium, marked the duration of the technosphere. Because this cycle concluded with the beginning of the third Christian millennium, a singular event was poised in time to attract and catalyze a host of prophetic moments long prepared for and sealed in the psi bank by Christian, Islamic, and even Mayan prophecies. Such was the Inevitable Event. Within the motivations of the perpetrators of the Inevitable Event, the biosphere found its perfect instrument of divine purpose. That there were nineteen hijackers conforms to the Quranic nineteen code expressed in the enigmatic verse, "Over it is nineteen."

(74:30) The following verse begins, "We appointed angels to be guardians of Hell, and we assigned their number [19] . . ." This verse ends by asking, "What did God mean by this allegory? . . . None knows the soldiers of your Lord except He. This is a reminder for the people." (74: 31)

Certainly in the success of their mission, whether they were conscious of it or not, the nineteen Muslim hijackers-who organized themselves as four leaders (pilots), three logistics people, and twelve "soldiers"-fulfilled their function of being the prophetic emissaries of the nineteen guardian angels who watch over Hell. And indeed the gates to Hell have been opened. In the world of the secular mind of the technosphere, no one wanted to talk about the apocalypse or Armageddon, though the jihad was brought up quite often as a factor in the thinking of the Islamic terrorists.6

Whether it is seen as a prophetic or karmic event, as we stated earlier, there is no question that the timing programs of two civilizations, the Gregorian and the Islamic, were on a collision course. "The civilization of Islam is radically so different from that of Western civilization which dominates the world today. The two are different in their description oflife as well as the foundation on which they base such a description. The difference between the two civilizations is so essential that they have developed in ways which are radically contradictory to each other."7 So wrote the eloquent Saudi historian, Muhammad Husayn Haykal in his essay "Islamic Civilization," as presented in his masterful biography, *The Life of Muhammad.* And what is at the root of this difference? A separation of church and state, which in the West led to a profound materialism and economic worldview.

Haykal continues, "The victory of materialist thinking was largely due to the establishment of Western civilization primarily upon an economic foundation. This situation led to the rise in the West of a number of worldviews which sought to place everything in the life of man and the world at the mercy of economic forces. . ." As for the sanctity of its economic beliefs, Haykal concludes that the West has tied its economic activity "hand and foot by public laws and commanded that every Western state and army prevent any violation of economic laws with all the power and coercive means at its disposal."8 Given this insightful assertion by an Islamic thinker written a half-century ago, it is not at all surprising that the terrorists concentrated their attacks on the most visible symbols of Western global materialism and economic and military imperialism, the Twin Towers and the Pentagon. It is also of note that one of the logistics men on the team of nineteen hijackers, Majed Moqed, had been a professor of economics at King Saudi University in Riyadh.

However we see this event, the most important point is that it actually happened, no matter who caused it or how. In this regard the collapse of the World

Trade Center towers and the penetration of the Pentagon was the Inevitable Event both from the point of view of the biosphere and from that of prophecy as mandated by at least two major timing programs, the Gregorian and the Islamic. Of course, prophecy is a matter far from the consideration of the technocratic mind, so the immediate aftereffect is a heightened nationalism, patriotism, and commitment to war. But what is going on from the noospheric point of view is not so much the combat of different armed camps defending or attacking competing ideologies, but an end game, the biomass of the technosphere maneuvering against time in a self-destructive process that is really nothing more than the continuing collapse of that which it is defending.

How long will this collapse take? With the weapons of mass destruction currently at the disposal of the combatants, it shouldn't take too long before the human race realizes that if it doesn't stop the self-destructive process, it will soon be curtains for all life on Earth. A more significant question, if we wish to make the transition to a new time, is: How will the self-destruction end, and how can it be stopped-or can it? It is easy to see how wars begin, but no one can tell how they will end.

THE SOLUTION TO WAR: PEACE THROUGH CULTURE, PEACE OF THE BIOSPHERE

At the end of the First World War, the League of Nations was formed to keep the Second World War from happening. The League of Nations was unable to both effect the calendar reform it promoted and keep the Second World War from happening. These are not unrelated events. The Second World War effectively eliminated the League of Nations, and toward its end a new organization, the United Nations, was formed. Like the League of Nations, the United Nations was formed to prevent the next world war from happening. Within eleven years, the United Nations tabled the calendar reform begun by the League of Nations. Coincident with the formation of the technosphere, the United Nations has now also completed fifty-six years of existence. With the expanding "war against terrorism," is its purpose now over as well? The only way the United Nations can survive is to maintain its humanitarian agencies, for the age of nationalist politics and nation-states is over, succeeded by the politics of terrorism, the last politics of history. What new kind of global organization will we need at the end of the War on Terror, or World War III?

What we will need is an entirely new agency, a Planetary Biospheric Assembly that will model and create a sustainable *Pax Cultura, Pax Biospherica.* Given the failure to enact a calendar reform on the part of both the League of Nations and the

United Nations and in consideration of the chaos now clearly embedded in the Old Time, only an agency or instrument that fully embraces and promotes the calendar change will be able to move into the genuinely peace-fulfilling role that such an organization must take. If the United Nations could accept the route of calendar change as the only way to bring all humanity to a new point of discourse, then it could take on a renewed life. But if it cannot rise to this occasion, then another organization must take its place, the Planetary Biospheric Assembly. Even so, when we speak of the *Pax Cultura, Pax Biospherica,* we are speaking of a new human covenant, one in which the biosphere takes precedence over all nationalist corporate interests, and in which peace as a manifestation of culture is the primary value of the new time. Such a covenant would be a biospheric covenant, an agreement and understanding that, first of all, the human is a member of the biosphere, and second, that in consciously accepting the calendar change, the human is also in the process of becoming *Homo noosphericus.* Henceforth, it will be art and not money, bioregions and not nation-states, that will be the cause of social organization. Only in this way will humanity be guaranteed a true post-historical future.

Assuming the calendar change occurs on Blue Crystal Storm 2004, the first step in ~ngendering the full manifestation of the noosphere will be the dismantling of the technosphere. This can only be accomplished if a full reorganization of the human in time, inclusive of an adjudication of inequalities and a redistribution of wealth, begins to occur during the 2004-2005 Crystal Storm year. The following four years, 2005-2009, a genuine First World Peace, would see the constructive transformation of the five-part technospheric system. All the apparatus of heavy industry, energy, and the transport system would be turned like compost, to be replaced by nonpolluting technologies and means of transport, while the cities would be transformed into urban gardens with the human race once more becoming decentralized. This process itself, a massive planetary engineering project with an eye on the full development of geochronomancy, would prove to be the greatest and most thoroughly positive transformation in human history. By 2009 humans reorganized by the Law of Time and having finally arrived at the "path beyond technology," will be enjoying universal telepathy, the noosphere-life beyond the Internet. By 2012 the humans will realize peace through culture as a permanent value, and by 2013 Earth will be galactically synchronized through a fully activated noosphere into the beginning of an entirely new geological era, defined by Vernadsky as the Psychozoic Era.

Pax Cultura, Pax Biospherica refers to both the new era of humanity following the devastation of the Final War and the governing body of bioregional councils (not nation-states) comprising the new world "government," the Planetary Biospheric

Arcturus Victory Strategy
Pax Cultura, Pax Biospherica

Telektonon, the Theater of Time, provides the journey through the Cube of the Law: Plane of Mind, Plane of Spirit, Plane of Will, and Divine Source Telektonon—each plane Seed-Storm year bundle, an entire universe of noospheric time!

Plane of Spirit: First Seed-Storm year bundle, NET week 9 Seed–12 Storm, 2001–05

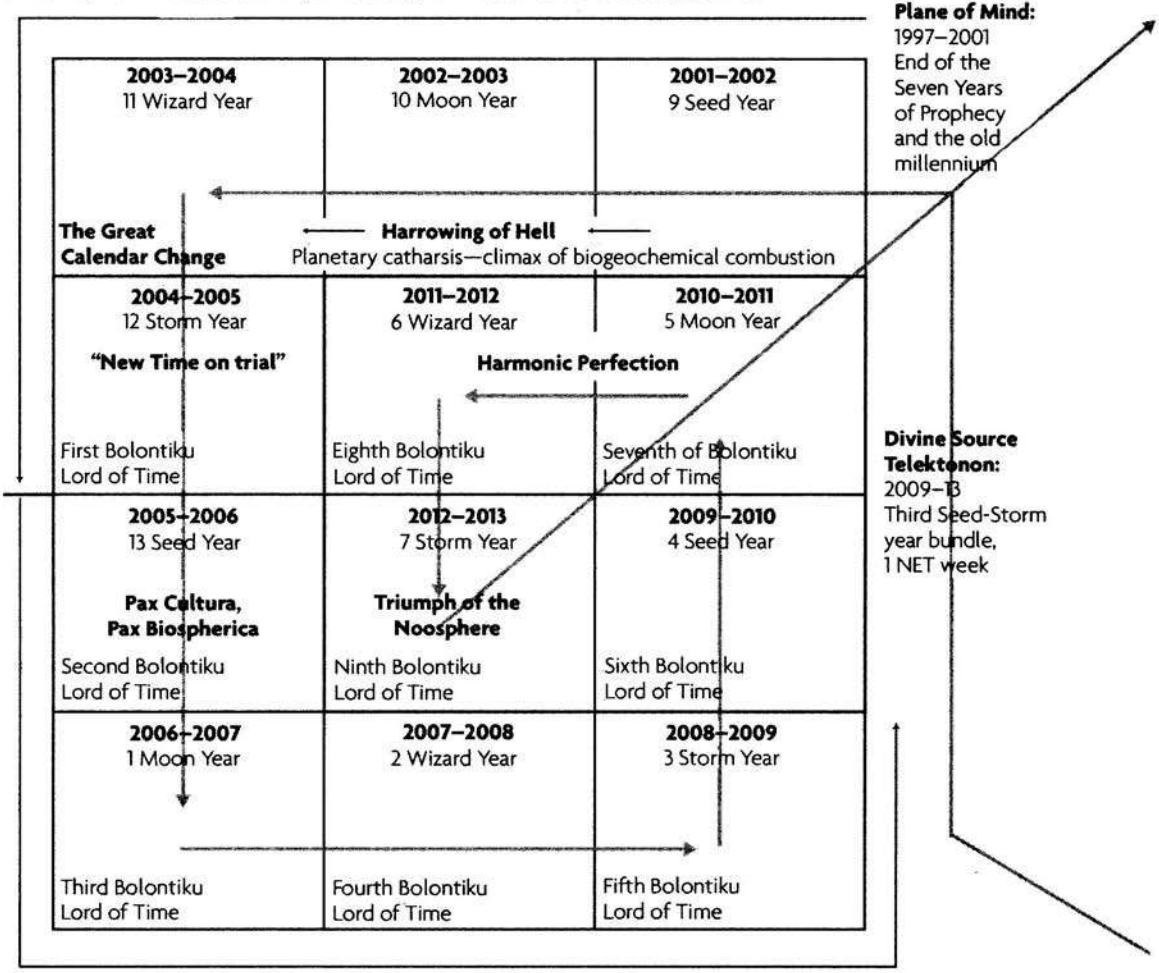
Plane of Mind:
1997–2001
End of the
Seven Years
of Prophecy
and the old
millennium
2003–2004
11 Wizard Year
2002–2003
10 Moon Year
2001–2002
9 Seed Year
The Great
Calendar Change
Harrowing of Hell
Planetary catharsis—climax of biogeochemical combustion
2004–2005
12 Storm Year
"New Time on trial"
First Bolontiku
Lord of Time
2011–2012
6 Wizard Year
Harmonic Perfection
Eighth Bolontiku
Lord of Time
2010–2011
5 Moon Year
Seventh of Bolontiku
Lord of Time
Divine Source
Telektonon:
2009–13
Third Seed-Storm
year bundle,
1 NET week
2005–2006
13 Seed Year
Pax Cultura,
Pax Biospherica
Second Bolontiku
Lord of Time
2012–2013
7 Storm Year
Triumph of the
Noosphere
Ninth Bolontiku
Lord of Time
2009–2010
4 Seed Year
Sixth Bolontiku
Lord of Time
2006–2007
1 Moon Year
Third Bolontiku
Lord of Time
2007–2008
2 Wizard Year
Fourth Bolontiku
Lord of Time
2008–2009
3 Storm Year
Fifth Bolontiku
Lord of Time

Plane of Will: Second Seed-Storm year bundle, 1 NET week, 13 Seed–3 Storm, 2005–09
Time of Transformational Destructuring—First World Peace

Assembly. The main purpose of the new governing body would be to monitor the advent of the noosphere. Overcoming Northern Hemisphere bias, the *Pax Cultura, Pax Biospherica* will not find any permanent seat, but will begin its existence as a planetary noospheric congress in the Southern Hemisphere. All of this could be made possible only because of the completion of the calendar change. So let us assume the best. The how-to of paradigm shift begins with the Thirteen Moon calendar.

If you follow the Thirteen Moon calendar or tell others to follow it with you, what is the payoff? The payoff is that you get to become a noospheric chip. Only a noospheric chip can bring Heaven on Earth, which is nothing more than full human participation in the noosphere. This is not at all a theoretical matter, but something that can be practiced and followed on a day-to-day basis. Fortunately, a path has been shown, complete with travel aids.

CONCISE *HOMO NOOSPHERICUS* HOW-TO TIME TRAVELER'S GUIDE TO THE BIOSPHERE-NOOSPHERE TRANSITION

This analysis of the Inevitable Event could never have occurred if the Law of Time hadn't been discovered and defined. The Inevitable Event simply supplied the Law of Time with precisely the right historical moment to provide a thorough analysis and explanation of time as the whole system in which we humans live and operate, inclusive of the biosphere, the technosphere, and the noosphere. As the chief discoverer of the Law of Time, assisted by my wife and an increasing number of other humans who have engaged the Law of Time, I have sought only to make a presentation of matters in which we, as a species, are thoroughly immersed, and thus find difficult to comprehend. Such is the technosphere, and such also is the medium of time. Since the publication of *The Mayan Factor* and my pursuit of the underlying mathematics of the Mayan calendar, I have remained outside the conventional arenas of discourse and communication, preferring instead to remain more anonymous in order to test the various hypotheses of the Law of Time.

In this process over the past twelve years, the Law of Time has precipitated from the noosphere various tools and didactic texts, all for the purpose of enabling us to make the shift from the old paradigm of the time of war to the new paradigm of the time of peace. These tools and principles are actually practiced on a daily basis, and so provide a pragmatic means of applying the Law of Time. Because of my vows and need for scientific "purity" of the experiment I have been living, I was involved at all times in doing my utmost to remain in the 13:20 frequency. For this reason, the tools and texts prompted by the noosphere to define the Law of Time were, for the most part, all produced outside the commercial mainstream. Now a great effort must be taken to make these tools and texts available to humanity in order to assist in the transition from *Homo historicus* to *Homo noosphericus*. By presenting a catalog of these tools and texts we only mean to suggest and define for the reader the scope and existence of a preliminary science of time and telepathy.

Part I. The Tools: Establishing Noospheric Earth Time (NET)

THIRTEEN MOON/28-DAY SYNCHRONOMETER

The harmonic standard of noospheric time, the Thirteen Moon/28-Day synchronometer, is a perpetual, harmonic calendar. It is called a Moon calendar because it is based on the female 28-day menstruation cycle, which is also the average lunar cycle. In actuality the moon goes around the Earth thirteen times a year. This means that the Thirteen Moon calendar is a genuine solar-lunar calendar that measures the Earth's orbit around the sun by the apsidallunation cycle of twenty-eight days. Thirteen perfect months of 28 days = 52 perfect weeks of 7 days = 364 days. The 365th day is called the Day Out of Time because it is no day of the week or month at all. This day, which falls on the Gregorian correlate date ofJuly 25, is a day for forgiveness and the artistic celebration of life and freedom. The synchronization or New Year's date of the Thirteen Moon calendar is July 26, and corresponds to the rising of the great star Sirius. This makes the Thirteen Moon calendar a tool for harmonizing ourselves with the galaxy. Of course, the Thirteen Moon calendar also synchronizes the phases of the moon in a pattern that demonstrates its regularity and order.

One of the great advantages of the Thirteen Moon synchronometer is that day/date calculations are amazingly simple. Any day of any week is the same for any Moon, any year. The first day of every Moon is always a Thirteen Moon Sunday (Dali). The last day of every Moon is always a Thirteen Moon Saturday (Silio). In the Thirteen Moon calendar the obscurely named Gregorian months are replaced by names that correspond to a fourth-dimensional cosmology of time. Each Moon also has a totem animal. The traditional names of the days of the week are replaced by galactic names that describe seven primary plasmas-electronically charged particles that activate our magnetic fields. The plasmas and their symbolic seals appear above the days of the week. The Gregorian calendar makes day/date calculations very difficult because the months are of uneven measure so the days and dates of the week vary from month to month and year to year. The Thirteen Moon calendar is truly unique because it is synchronized with the Harmonic Module, the measure of the universal 13:20 timing frequency.

THE HARMONIC MODULE

The basis of the Tzolkin or sacred 260-day count of the Maya, the Harmonic Module is the 13:20 synchronization gauge and permutation table that consists of thirteen sequences of twenty icons or solar seals and twenty sequences of thirteen galactic tones, 1-13. The resulting 260 permutations, combined with the perfect harmony of the Thirteen Moon calendar, give each day a unique quality of tone and icon known as a kin or one NET minute. The two cycles-Thirteen Moons/28 days and the 260-day Harmonic Module-perfectly mesh every fifty-two years to complete one NET year. In one NET year of 18,980 kin, no two days are the same! Each of the units in the Harmonic Module is called a kin, the base unit of synchronic measure. Articulating the Harmonic Module is a pattern of fifty-two Galactic Activation Portals. See if you can find the radial sequence of thirteen sets of four, counting inward from the corners. Notice that the numbers of each set of four equals 28. Thirteen sets x 28 = 364, the number of days in the Thirteen Moon calendar! (See image containing the Harmonic Module on page 170.)

WAVESPELL

The thirteen-unit form constant of fourth-dimensional time is the wavespell. It is important to see the Thirteen Moon calendar in this modular format, Planetary Service Wavespell, for then it makes more sense to speak of the calendar as a synchronometer. The wavespell's form describes the motion of time as a fractal fourth-dimensional cosmology. The numbers 1-13 in the dot bar notation that code the thirteen positions are known as tones. The names of the tones in sequence define the cosmology, which is also evident in the very structure of the wavespell. The thirteen tones give their names to the thirteen Moons. The structure consists of two gates, the Magnetic and the Cosmic, the first and the thirteenth moons/tones, which correspond to the two entrance and exit gates or portals. Complementing the two gates are the two towers, the fifth or Overtone tower and the ninth or Solar tower. The two gates and the two towers, tones 1, 5, 9, and 13, articulate the structure of the wavespell. Between the places of articulation are three sets of chambers. The first set, tones 2-3-4, helps establish the wavespell; the second set, tones 6-7-8, helps extend the wavespell; and the third set, 10-11-12, serves to convert the wavespell. (See plate 2, Planetary Service Wavespell.)

The structure of the wavespell also defines a four-dimensional geometry of time that connects different points of the wavespell over time. Each of these dimensional geometries is known as a pulsar. It is useful to know the pulsars in extending the mind synchronically over different timing sequences. The fourth-dimensional pulsar connects the four points of articulation, 1-5-9-13, or Magnetic, Overtone, Solar, and Cosmic Moon/tones, and is called the fourth-dimensional time pulsar. The first-dimension life pulsar connects the positions 2-6-10, or Lunar-Rhythmic-Planetary Moon/tones. The second-dimensional sense pulsar connects the 3-7-11, or Electric-Resonant-Spectral Moon/tones; and the third-dimensional mind-form pulsar connects the 4-8-12, or Self-existing-Galactic-and -Crystal Moon/tones. Remember, the wavespell accommodates any sequence of kin whose tones correspond to 1-13, and may be either a thirteen-day wavespell cycle, a thirteen-week cycle (one-quarter of an NET day), a thirteen-Moon cycle (one NET day), a thirteen-year cycle (one NET Season), and so on. Once you know the wavespell you have come to know the form in time of the noospheric mind.

THE WAVESPELL'S TONES AND THEIR PULSARS

1.	Magnetic tone of purpose-fourth-dimension	time pulsar
2.	Lunar tone of challenge-first-dimension	life pulsar
3.	Electric tone of service-second-dimension	sense pulsar
4.	Self-existing tone of form-third-dimension	mind-form pulsar
5.	Overtone tone of radiance-fourth-dimension	time pulsar
6.	Rhythmic tone of equality-first-dimension	life pulsar
7.	Resonant tone of attunement-second-dimension	sense pulsar
8.	Galactic tone of integrity-third-dimension	mind-form pulsar
9.	Solar tone of intention-fourth-dimension	time pulsar
10.	Planetary tone of manifestation-first-dimension	life pulsar
11.	Spectral tone of liberation-second-dimension	sense pulsar
12.	Crystal tone of cooperation-third-dimension	mind-form pulsar
13.	Cosmic tone of presence-fourth-dimension	time pulsar

DREAMSPELL

Described as the fourth dimensional tool kit, the Dreamspell also encodes a psycho-mythic cosmology known as the Journey of Timeship Earth 2013. It is this cosmology that makes the Dreamspell a psychoactive tool kit for personally engaging you in the fourth-dimensional synchronic order of time. The fundamental premise for making it psychoactive lies in finding your galactic signature, the color, icon, and tone that code your date of birth. This combination of factors-color, tone, and icon-establishes a mathematical code that allows you to use your galactic signature in mapping increasing orders of synchronicity. There are four colors, thirteen tones,

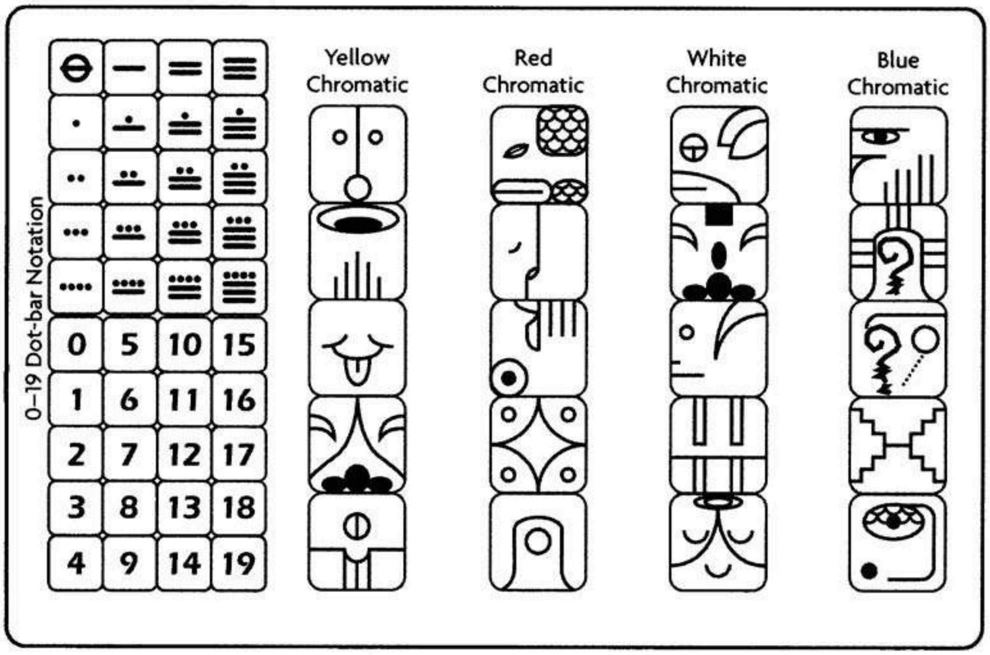
0–19 Dot-bar Notation
0
5
10
15
1
6
11
16
2
7
12
17
3
8
13
18
4
9
14
19
Yellow Chromatic
Red Chromatic
White Chromatic
Blue Chromatic

The Four Chromatics

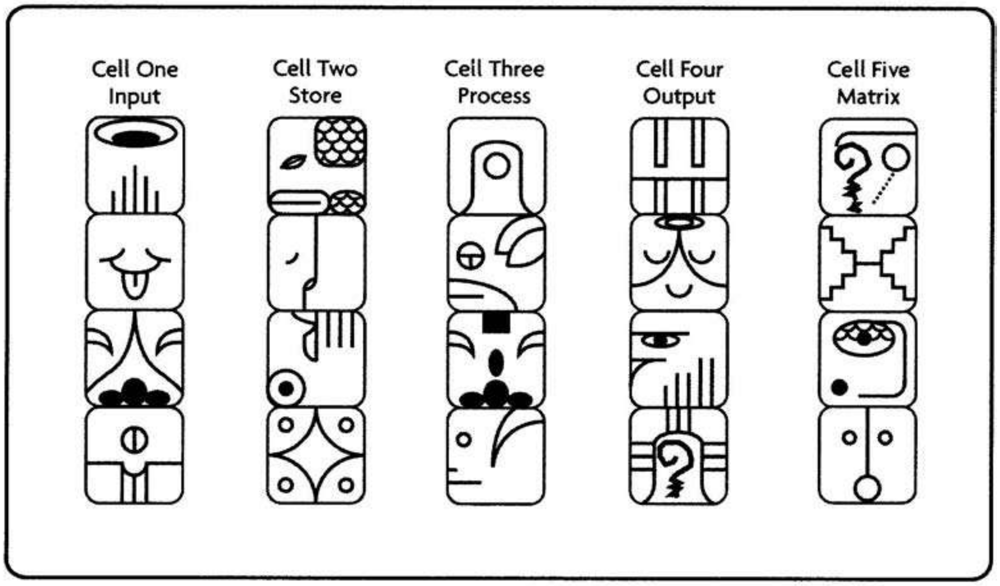
Cell One Input
Cell Two Store
Cell Three Process
Cell Four Output
Cell Five Matrix

The Five Time Cells

and twenty icons. The four colors code each of five sets of icons and correspond to the base four-phasic function of the noospheric rainbow brain-red initiates, white refines, blue transforms, and yellow ripens. The icons themselves actually correspond to the mathematical code 0-19.

Colors and Icons with Their Corresponding 0-19 Code Numbers:

Red Family: Dragon (1), Serpent (5), Moon (9), Skywalker (13), and Earth (17)

White Family: Wind (2), World-bridger (6), Dog (10), Wizard (14), and Mirror (18)

Blue Family: Night (3), Hand (7), Monkey (11), Eagle (15), and Storm (19)

Yellow Family: Seed (4), Star (8), Human (12), Warrior (16), and Sun (20 = 0)

The mapping of synchronicity through use of the galactic signature is done through a set of tools: the Galactic Compass, Journey Board, Oracle Board, and Time Atom Cube. The point of engaging the Dreamspell is to learn of the infinite possibilities of fractal synchronic mapping of the events of your life and the lives of your family and friends, as well as current events and human history. The compass and boards describe the radial mathematics of fourth-dimensional time and also provide the means for fully engaging the synchronic order of the noospheric mind. The interactive complex of changing the family and clan relations to which the galactic signature entitles you also provides the basic patterns of social organization in time. The Dreamspell is indispensable for education concerning the radial, fractal nature of fourth-dimensional time and mathematics.

TELEKTONON

Meaning "Earth Spirit Speaking Tube," *Telektonon* is the name of the prophecy of Pacal Votan, and the definition of a complex program for establishing universal telepathy. It should be noted that the tomb of Pacal Votan was discovered in 1952, after 1,260 years, because of a tile tube, the Earth Spirit Speaking Tube, which ran from the top of the pyramid temple, down the secret stairwell, and into the wall of the elaborate tomb beneath the pyramid. The prophecy itself was decoded from the inscriptions and symbols of the elaborately carved sarcophagus lid of the tomb. Essentially defining a period of "Seven Years of Prophecy," (1993-2000), followed by a 13-year Cycle of Creation (2000-2013), the Telektonon is practiced on a daily basis by means of a board and various moving parts.

The purpose of the daily play is to learn how to operate by a set of coordinates that define the nature of telepathic consciousness as the interplay of five circuits that connect the ten planetary orbits (the Asteroid Belt, remnants of the lost fifth planet Maldek, is included as one of these orbits). The key interplanetary connection is between Earth, the third planet, and Uranus, the eighth planet, which creates the third or biotelepathic circuit. The Earth-Uranus circuit constitutes twenty-eight steps and thus corresponds to the 28-day measure of the Thirteen Moon calendar. Between days 7 and 22, one is

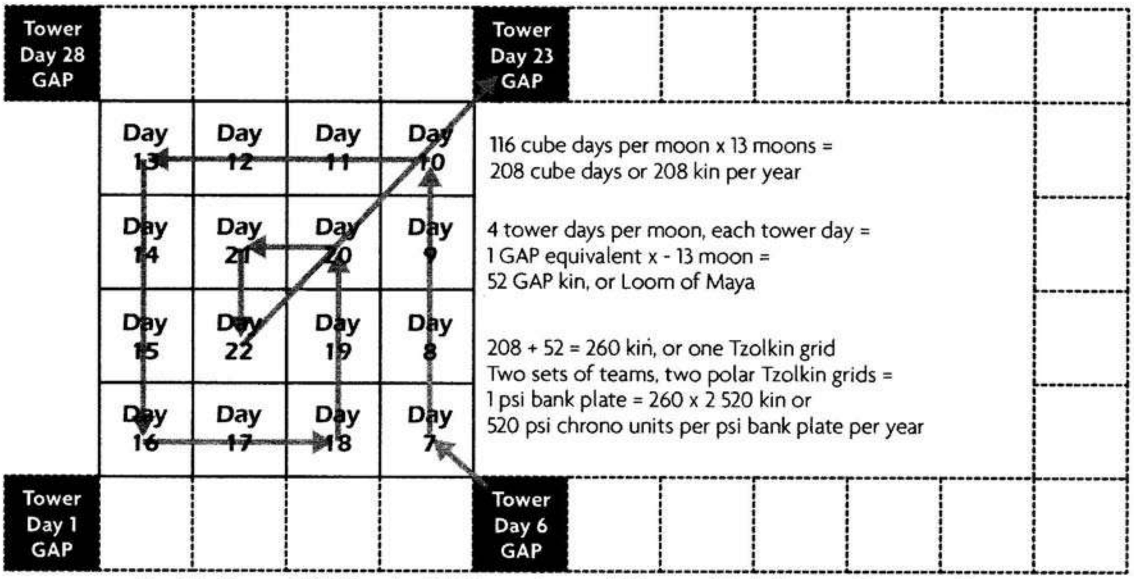
Tower Day 28 GAP
Tower Day 23 GAP
Day 13
Day 12
Day 11
Day 10
Day 14
Day 21
Day 20
Day 9
Day 15
Day 22
Day 19
Day 8
Day 16
Day 17
Day 18
Day 7
Tower Day 1 GAP
Tower Day 6 GAP

116 cube days per moon x 13 moons = 208 cube days or 208 kin per year

4 tower days per moon, each tower day = 1 GAP equivalent x - 13 moon = 52 GAP kin, or Loom of Maya

208 + 52 = 260 kin, or one Tzolkin grid
Two sets of teams, two polar Tzolkin grids =
1 psi bank plate = 260 x 2 520 kin or
520 psi chrono units per psi bank plate per year

The Telektonon Grid: Showing Psi Chrono, Loom of Maya, and Cube Journey Correspondences

Tower Day 1 (x 13)

Tower Day 23 (x 13)

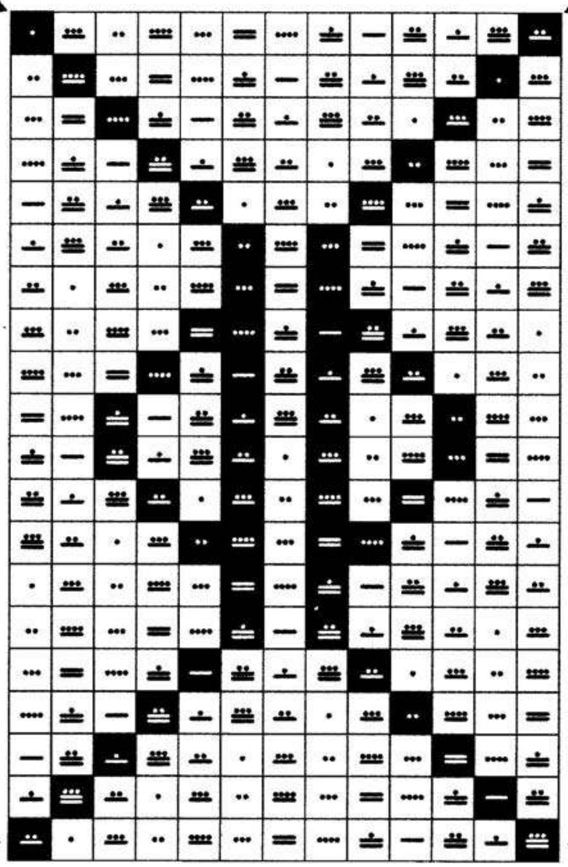

Harmonic Module 260 (13 x 20) unit universal standard of measure of fourth-dimensional time

GAP = Galactic Activation Portal, 52 of which create the Loom of Maya, or binary triplet configuration of the Harmonic Module (Tzolkin).

The 52 GAPs are the basis of the occult quartet; that is, starting from each corner and moving inward there are thirteen of these sets creating a radial matrix where the tones of each set always equal 28.

13 x 28 = 364, the number of days in the Thirteen Moon calendar.

Each of these quartets codes the four tower days of the Telektonon every moon.

Tower Day 6 (x 13)

Tower Day 28 (x 13)

Telektonon - Rinri GAPs and Towers with Harmonic Module able to take the "Warrior's Cube Journey" to uncover knowledge hidden by the Tower of Babel. The sixteen positions of the Cube of the Law (days 7-22) are the heart of the Telektonon play, which also includes a rehearsal of the thirteen baktun history with the intention of "redeeming the Babylon planet" as Timeship Earth 2013.

THE RINRI PROJECT

Described as a "four-year telepathic biosphere-noosphere transition Circumpolar Rainbow Bridge Experiment," the Rinri Project builds on the Telektonon play that establishes the biotelepathic circuit. On this circuit, two towers, days 1 and 6, define a four-day Earth Walk, and two towers, days 23 and 28, define a four-day Heaven Walk. Between the second and the third towers, days 6 and 23, are the sixteen positions of the Cube of the Law, days 7-22. The four towers every Moon correspond to one radial set of four Galactic Activation Portals in the Harmonic Module-thirteen moons x four towers = 52 Galactic Activation Portals (GAPs). Sixteen cube positions each moon x 13 Moons = 208, the remaining number of kin in the Harmonic Module once the 52 GAPs have been subtracted from the 260 units. Since the Harmonic Module is the basis of the psi bank, this formulation provides a way of coordinating the four psi bank plates of the noosphere to the Thirteen Moon calendar year on a daily and annual basis.

Since there are four plates, it takes four years, one year per plate, to open the psi bank. The first opening occurred between 1996-2000. The second opening, 2000-2004, is now occurring and is programmed by a coordination of the twenty-four runes of the Elder Futhark to the twenty-four psi membranes-6 runes per week every Moon x four weeks = 24. Since each psi plate is a mirror symmetry pair of Harmonic Modules, reflecting the North and South magnetic poles of the Earth, the Rinri project is practiced by telepathically interactive teams in the Northern and Southern Hemispheres. The Rainbow Bridge refers to making visible the dynamic of Earth's electromagnetic aura and extending it permanently from pole to pole. This is how the noosphere is visibly manifest.

Rinri is a Japanese word meaning "ethical enlightenment," and its application includes seventeen moral precepts that are studied during the sixteen-day cube journey, the seventeenth corresponding to the tower day 23.

260 POSTULATES OF THE DYNAMICS OF TIME

While the Telektonon is predicated on the fourth-dimensional tool kit, the Dreamspell, the 260 Postulates of the Dynamics of Time elaborate on the nature and theory of the biosphere-noosphere transition by providing a set of twenty systematic thirteen-part logics that describe and define the dynamics of time as the evolution of consciousness. Because there are 260 postulates corresponding to the 13:20 matrix, a postulate can be studied and meditated on each day over a 260-day cycle according to the code number of the daily icon combined with the number of its tone, for example, Red Lunar Dragon, where Dragon is 0-19 code number 1, and Lunar is tone 2, therefore the postulate to study for the day is 1.2. This daily programming, established by the Telektonon practice is how knowledge is reformatted according to the synchronic order. This in itself precipitates the noosphere.

TWENTY TABLETS OF THE LAW OF TIME, SIXTEEN-YEAR CUBE OF THE LAW

The Twenty Tablets can be described as a road map of the fourth dimension. By focusing on a sequence between the years 1997-2013, by fractal analogies, the telepathic engagement encompasses the 26,000-year cycle as well as the coming cycle of the Psychozoic future. The crux of the daily play of the synchronic order is called *cubing the codon.* The point is to telepathically reharmonize the DNA that has been adversely affected by the 12:60 timing frequency. There are sixty-four DNA codons whose six-part binary structure is identical to that of the sixty-four hexagrams of the I Ching when both are written with a simple binary language of broken or unbroken lines. Each of the codons establishes a precise thirteen-stage permutation sequence. 64 x 13 = 832 codon permutations. The sixteen years from 1997 to 2013 also constitute 832 weeks (16 x 52 = 832 = 64 x 13). Therefore, each codon has one-quarter year, thirteen weeks, to run its permutation sequence, one sequence per week. Since each codon consists of six lines, each week the codon is built one line per day for six days, each line adhering to one side of a cube. On the seventh day, the codon is cubed and sealed by one of sixty-four "UR Runes." The UR Runes are the master cosmic runes, each of which governs an entire thirteen-unit codon permutation sequence. The Twenty Tablets also contain much synchronic information for experiencing what is referred to as *fractal time compression,* the expanded mental ability to engage in different planes or levels of time simultaneously. All of this is to establish a mind training for fully engaging the fourth-dimensional synchronic order of reality and time travel.

7:7::7:7 TELEKTONON REVELATION

While the Twenty Tablets of the Cube of the Law synchronize the DNA Codons in all their permutations on a weekly basis over a sixteen-year cycle, the 7:7::7:7 Telektonon Revelation establishes the 28-day cycle as the basis for synchronizing

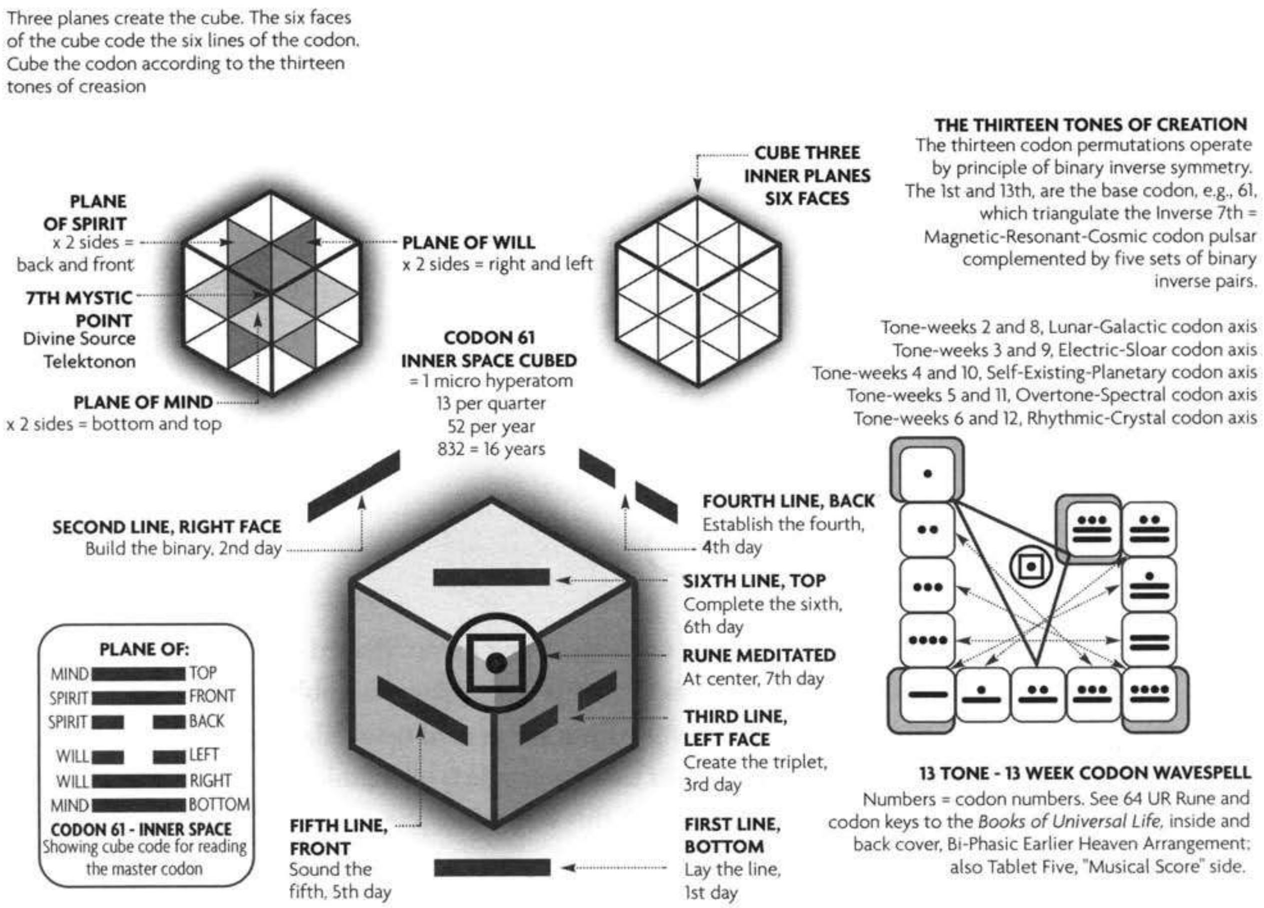
Three planes create the cube. The six faces of the cube code the six lines of the codon. Cube the codon according to the thirteen tones of creasion
PLANE OF SPIRIT
x 2 sides = back and front
7TH MYSTIC POINT
Divine Source Telektonon
PLANE OF MIND
x 2 sides = bottom and top
PLANE OF WILL
x 2 sides = right and left
CODON 61
INNER SPACE CUBED
= 1 micro hyperatom
13 per quarter
52 per year
832 = 16 years
CUBE THREE INNER PLANES SIX FACES
THE THIRTEEN TONES OF CREATION
The thirteen codon permutations operate by principle of binary inverse symmetry. The 1st and 13th, are the base codon, e.g., 61, which triangulate the Inverse 7th = Magnetic-Resonant-Cosmic codon pulsar complemented by five sets of binary inverse pairs.
Tone-weeks 2 and 8, Lunar-Galactic codon axis
Tone-weeks 3 and 9, Electric-Sloar codon axis
Tone-weeks 4 and 10, Self-Existing-Planetary codon axis
Tone-weeks 5 and 11, Overtone-Spectral codon axis
Tone-weeks 6 and 12, Rhythmic-Crystal codon axis
SECOND LINE, RIGHT FACE
Build the binary, 2nd day
FOURTH LINE, BACK
Establish the fourth, 4th day
SIXTH LINE, TOP
Complete the sixth, 6th day
RUNE MEDITATED
At center, 7th day
THIRD LINE, LEFT FACE
Create the triplet, 3rd day
FIFTH LINE, FRONT
Sound the fifth, 5th day
FIRST LINE, BOTTOM
Lay the line, 1st day
PLANE OF:
MIND TOP
SPIRIT FRONT
SPIRIT BACK
WILL LEFT
WILL RIGHT
MIND BOTTOM
CODON 61 - INNER SPACE
Showing cube code for reading the master codon
13 TONE - 13 WEEK CODON WAVESPELL
Numbers = codon numbers. See 64 UR Rune and codon keys to the Books of Universal Life, inside and back cover, Bi-Phasic Earlier Heaven Arrangement; also Tablet Five, "Musical Score" side.

Codon Cube Cosmology
Sounding the Thirteen Tones of Creation

the seven primary plasmas-electronically charged ions-into four aggregates called time atoms. The seven plasmas, which also replace the names of the days of the week, are in two sets of three, with the middle or fourth acting as a catalytic agent. The first three plasmas form a sensory quantum and the last three plasmas form a telepathic quantum. Activation of the plasmas releases radion, a telepathic lubricant; hence, the seven plasmas are also referred to as radial plasmas. The four plasmic aggregates are known as the Analog Time Atom (week 1), the Antipode Time Atom (week 2), the Occult Time Atom (week 3), and the Unified Field Time Atom, (week 4). Like the Twenty Tablets, the 7:7::7:7 Telektonon Revelation also involves principles of fractal time compression where each day during the first week is the equivalent of 260 moons or twenty years; the following two weeks, each day is thirteen

The Seven Seals, Seven Radial Plasmas

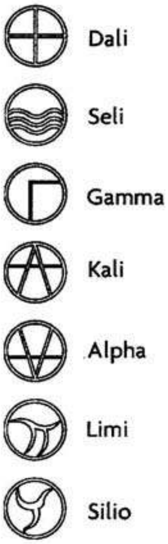

moons or one year; and the last week each day is one moon. Played on a board with cards containing affirmations from the Buddhist sage, Padmasambhava, the object of the 7:7::7:7 is to place the time atoms within the crystal octahedron at the center of the Earth so that each Moon a master time molecule is telepathically placed at the center of the Earth per one NET hour. In this way, Noospheric Earth Time is activated, the telepathic mind is engaged with the phenomenal reality represented by the plasmas, and the rainbow brain of the noosphere is hastened into everyday presence.

The seven daily plasmas constituting one thirteen-moon week, or one NET quarter hour are:

Dali (Sunday)-thermic charge

Seli (Monday)-luminic charge

Gamma (Tuesday)-luminic-thermic charge (these first three create one sensory quantum)

Kali (Wednesday)-luminic-thermic, thermic-luminic catalytic agent

Alpha (Thursday)-double-extended electron at the South Pole

Limi (Friday)-mental electron at the North Pole

Silio (Saturday)-mental electron-neutron at the center of the Earth (these last three create one telepathic quantum)

Part II. The Peace Plan: Establishing the *Pax Cultura, Pax Biospherica*

From the time of the formation, or rather, the calling into existence of the World Thirteen Moon Calendar Change Peace Movement, there had to be a *World Thirteen Moon Calendar Change Peace Plan.* The original of this Plan was submitted to the United Nations in 1995 for its fiftieth anniversary. The crux of the Peace Plan is that to change the world standard calendar is the opportunity for humanity to stop in its tracks and redirect itself. By stopping in its tracks to change its timing frequency, humanity can also use the opportunity to call a universal cease-fire and declare an emergency peace worldwide. Only in this way can the new time be experienced and fully engaged.

At the First Planetary Congress of Biospheric Rights (1996), this Peace Plan was extended to include the establishment of a new human covenant that would restore to humanity its intrinsic biospheric rights, the same as enjoyed by the rest of life in the biosphere: free land, clean air, clean water. The premise of the Biospheric Covenant is that humanity is first and foremost a member of the biosphere, and not of any artificially constituted nation-state. The Biospheric Congress also mandated the revival of the Banner of Peace and Roerich Peace Pact (1935) as one of the emblems to be used to identify the biosphere as a zone of peace.

The World Congress on the Law of Time and Judgment Day Tribunal (1997) extended the Biospheric Congress to include an indictment of the Gregorian calendar as a crime against the biosphere-"the crime of the millennium"-and the establishment of the Law of Time, inclusive of the Twenty Tablets, as the new dispensation of knowledge awarded in compensation to humanity, based on the synchronic veracity of the Thirteen Moon Calendar Change Peace Plan.

By 1999, the World Summit on Peace and Time provided the Declaration of Calendar Reform and a set of People's Resolutions, which were presented to the United Nations and the Vatican. The Resolutions define a seven-part educational program, which, like the Biospheric Covenant, envisions a radical reorganization of human society.

The purpose of these documents and declarations is for the creation of a state of universal unconditional peace so that human society can be rebuilt as a harmonic unity in phase with the biosphere as a whole. The result of the Thirteen Moon Calendar Change Peace Plan would be the creation of a new order of life on Earth-the *Pax Cultura, Pax Biospherica.*

◆

9

The Advent of the Noosphere

A Prayer and a Vision

IT IS MORNING once again, or, rather, it is predawn. It is still the first watch of the day. Outside there is a full moon. On the Gregorian calendar it is November 1, 2001, All Soul's Day, the sequel to Halloween, the day of honoring the dead. In another two weeks it will be the New Moon of Ramadan, Islamic lunar year A.H. 1422. On the Thirteen Moon synchronometer itis now the fifteenth day of the Self-existing Owl Moon, the moon of form, the moon to define, measure, and ask the question, "What form will my service take?" Well, my service will be to complete this book and prepare it for the world, for now it is the time of the advent of the noosphere, the reality of which Vernadsky said must be communicated to the world's politicians and statesmen. They didn't listen then. WIll they listen now? And on the 260-kin synchronization gauge it is kin 42, White Electric WInd. The WInd "communicates spirit," the Electric tone "activates service." WInd is the second of twenty solar seals that code each day in the synchronic order.

By its mathematical relationships to the other nineteen solar seals, the solar seal of the daily kin possesses a unique five-part oracle. In today's oracle the WInd is guided by the World-bridger, who brings the power of death. In the psi bank, the matching psi chrono unit for today, the fifteenth day of the Self-existing Owl Moon, is kin 66, Magnetic World-bridger, the first tone in the 13-kin wavespell of the World-bridger, which means today is very strongly penetrated by the power of death. What is death to the noosphere? An inevitable and unavoidable event, a critical and

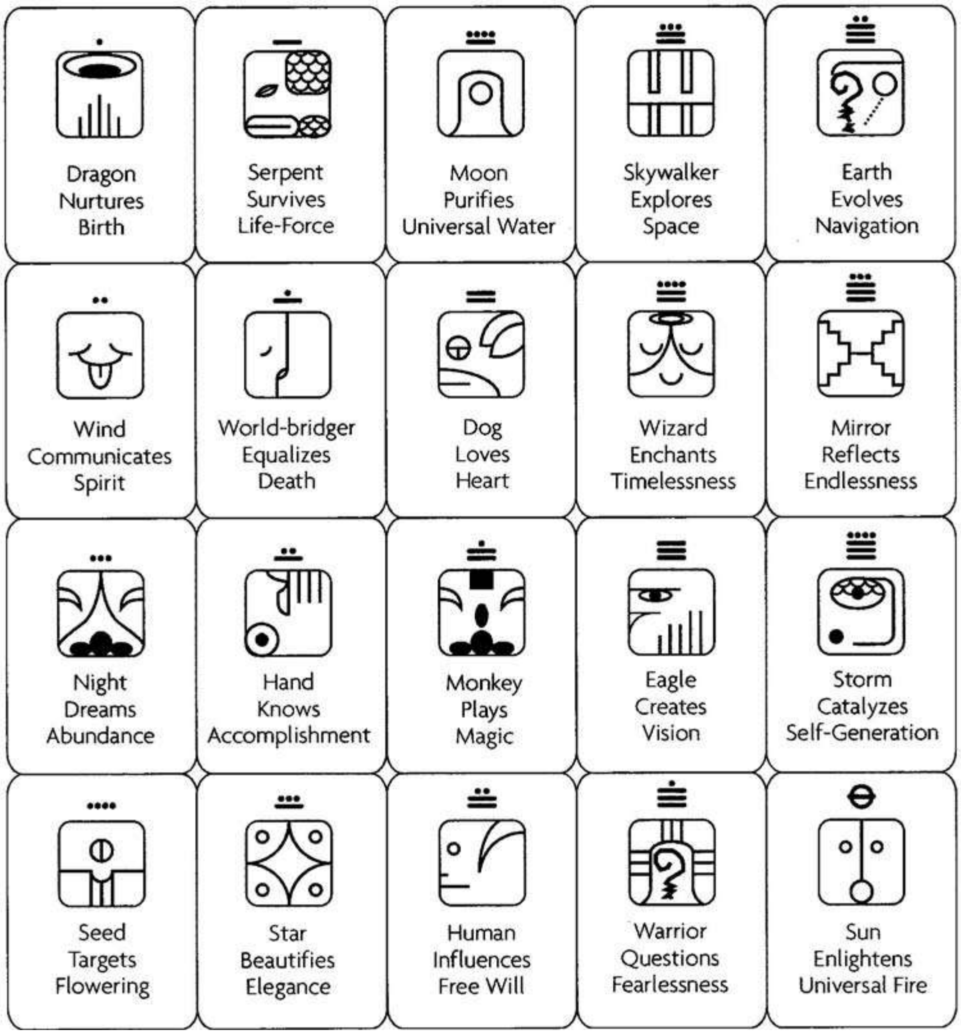
Dragon
Nurtures
Birth
Serpent
Survives
Life-Force
Moon
Purifies
Universal Water
Skywalker
Explores
Space
Earth
Evolves
Navigation
Wind
Communicates
Spirit
World-bridger
Equalizes
Death
Dog
Loves
Heart
Wizard
Enchants
Timelessness
Mirror
Reflects
Endlessness
Night
Dreams
Abundance
Hand
Knows
Accomplishment
Monkey
Plays
Magic
Eagle
Creates
Vision
Storm
Catalyzes
Self-Generation
Seed
Targets
Flowering
Star
Beautifies
Elegance
Human
Influences
Free Will
Warrior
Questions
Fearlessness
Sun
Enlightens
Universal Fire

The 20 Solar Seals
Names, Actions, and Powers

equalizing stage in the recycling of biomass, a furtherance of the biogenic migration of atoms, the creation of biospheric compost, a release of consciousness back into the noosphere, an opportunity to consider the value of life. But to the ordinary mind of the artificial technosphere, death is the ultimate object of fear, and the fear of death is the chief object of manipulation by those who generate terror. It is of note that by his date of birth, March 10,1957, the galactic signature of Osama Bin Laden is White Solar World-bridger, kin 126, while for this, his forty-fourth year, his birthday was coded by kin 66, White Magnetic World-bridger, both signatures bearing the destiny seal of the power of death. This is an example of how the synchronic order is "read" on a daily basis.

It is a mistake to think that the terror is just the malevolent intention of a single human being. The terror is co-extensive with the technosphere. In fact, the technosphere is the age of terror. The terror was generated at Hiroshima, and repeated at Nagasaki fifty-six years ago. And the terror was compounded by the original perpe-

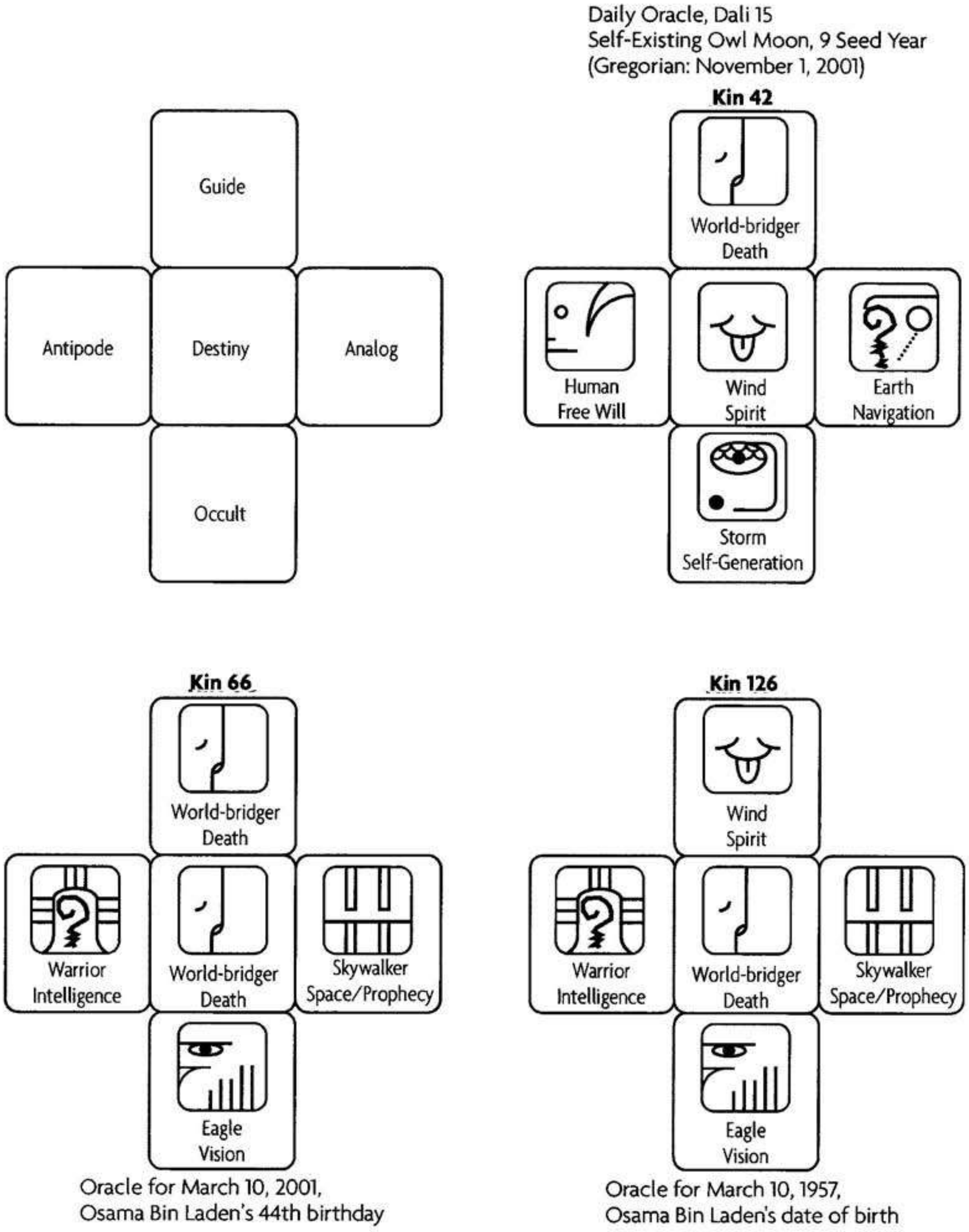

Daily Oracle, Dali 15
Self-Existing Owl Moon, 9 Seed Year
(Gregorian: November 1, 2001)

Oracle for March 10, 2001,
Osama Bin Laden's 44th birthday

Oracle for March 10, 1957,
Osama Bin Laden's date of birth

Guide
Antipode
Destiny
Analog
Occult
Kin 42
World-bridger Death
Human Free Will
Wind Spirit
Earth Navigation
Storm Self-Generation
Kin 66
World-bridger Death
Warrior Intelligence
World-bridger Death
Skywalker Space/Prophecy
Eagle Vision
Kin 126
Wind Spirit
Warrior Intelligence
World-bridger Death
Skywalker Space/Prophecy
Eagle Vision

Daily Oracle Matrix

trators of that terror many thousands of times over in the creation and stockpiling of weapons of mass destruction, an act soon imitated by the then archrival, the Soviet Union, and by other nation-states, namely, the United Kingdom, France, and China. Those called the "terrorists" today have nothing to compare to the stockpiles of these weapons by these five nation-states who constitute the permanent members of the United Nations Security Council-the "nuclear club." That is who rules the world today, the official possessors of the weapons of mass destruction, who jealously guard their secrets. If this is the ultimate ruling body of the planet, and the reason why they are the ruling body is their possession of these instruments of death, then why should anyone be surprised at the existence of "terrorists" in the rest of the world? And why, we may ask again, are these terrorists mostly, it would seem, from the Islamic world? In addition to the nuclear arsenals, there are the biological and chemical weapons of mass destruction. Those who live by the manufacture of terror must surely die by their own means. If other nations have developed the same systems of terror, it is only in self-defense or in imitation of the masters of terror.

Everyday for four weeks now, I have been waking up at this time to continue work on this text, *Time and the Technosphere*I had at first thought that the book could be complete with eight chapters. People don't have time to read, and the topic warrants as brief and pithy an explanation as possible. After all, it is really a time of reflection-and then action. The masters of terror have cloaked themselves as the protectors of civilization and as the moral crusaders who will rid the world of evil. Yet what is civilization today but the resource-consuming cancer of the biosphere? Is this what is being protected-the right to promote the disease upon the biosphere? The terrorists who attacked the Twin Towers and the Pentagon clearly knew the symbolic and actual value of their targets. They also undoubtedly knew the depths of the death-fear of the superpower masters of terror, whose reaction would most likely foment a global war. But in such a war, who would win? In this regard, from the noospheric perspective, the Inevitable Event was the technocide-death by technospheric means-of the technosphere itself. The technosphere was created by and thrived on means of terror, and the means of massive terror as well as the economic imperialism perpetrated by the superpower managers of the technosphere could only have produced a reaction against itself known as terrorism. The end result is collective suicide. There are no flags in the blood of the dead, nor is there any evident religious belief. There is only human biomass ultimately indistinguishable and anonymous to the biosphere.

Eight chapters were needed in order to complete the information about the biosphere, technosphere, noosphere, and the Law of Time. Maybe a new spirit of the biosphere will arise from the considerations of my book. Maybe the noosphere will spark in others the need to allow the biosphere to reclaim us from the technosphere. Anyone who says this is "turning back the clock" is only betraying their immersion and unexamined belief in the reality of artificial mechanical time. Yes, eight chapters were necessary to unfold this analysis and solution. But I knew there had to be a ninth chapter for the testament to be complete. For this is a living testament, and if anything has been learned by living according to the synchronic order of fourth-dimensional time, it is the renewed power of number itself. There is a power and meaning to number understood as a universal, self-existing mental construct. The synchronic order is constructed of number. And as Pythagoras perceived, number is a kind of spiritual entity that can only be aroused by an act of meditative contemplation. Each whole number bears its own resonance. You can feel this in the contemplation of the difference between even and odd numbers.

The resonance of 8, an even number, is harmonic order and balance, geometrically evident in the octagon or the octahedron. But the resonance of9, an odd number, is quite different. Nine months to give birth, three times three, a rhythmic triplicity-that is nine. The Germanic god-hero Odin, or Votan, spent nine nights hanging from a tree in order to gain knowledge of the mysteries of woman. When the tomb of Pacal Votan was opened on June 15, 1952, it revealed a large vaulted chamber beneath the pyramid of the inscriptions in Palenque, Chiapas, Mexico. Entering the tomb, in the center of which is the famous sarcophagus, one sees on the walls the beautifully sculpted figures of nine personages. Today these sculptures have almost completely worn away. But originally, these low-relief sculptures modeled out oflimestone, depicted with great elegance, a series of figures commonly known as the Nine Lords of Night, or the Nine Lords of Time. Are the Nine Lords of Time, the Lords of Night who guard Pacal Votan in his tomb, the same as the nine nights of OdinIVotan? And who was Pacal Votan? With what mythic resonance are we dealing?

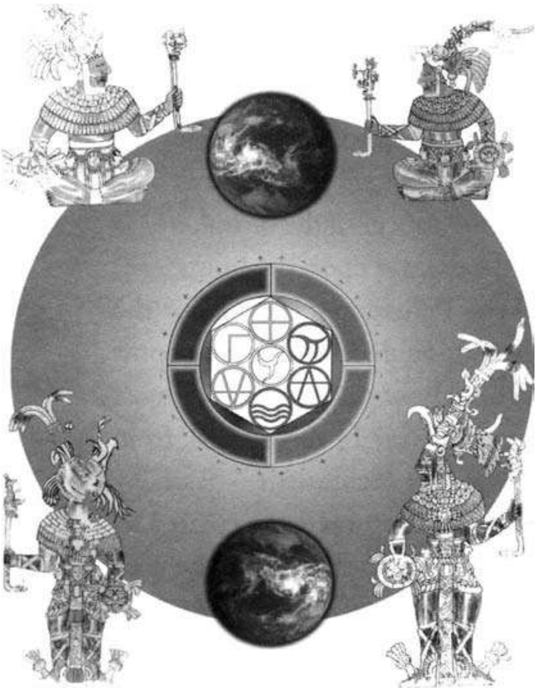

Nine Lords of Time Begin to Gather for a Council of the Earth

A Mayan seer, a sage-king of the seventh century of the Christian era, had deliberately created a mysterious tomb for himself, only to have it hidden, buried beneath a pyramid temple, but a tomb intended to be opened more than twelve centuries later, in the seventh year of the time of the terror, the time of the technosphere. How ingenious, how uncanny, that it would have been unknown for 1,260 years-a number also known in the thirteenth chapter of the Book of Revelations as the number of days of exile of the woman clothed with the sun, her feet on the moon, and twelve stars around her head. What does this allegory mean? Is it not the same as the exile of the human soul in the deadening world of 12:60 materialism? Sixty years remained from the opening of that mysterious tomb to the closing of the cycle of history, the omega point of 2012. How could the timing of the opening of the tomb have occurred so precisely-1,260 years from dedication to discovery, 1,320 years from its dedication to the closing of the cycle? How could those years be such a perfect reflection of the artificial 12:60 and the natural 13:20 timing frequencies? How could a Mayan sage in Central America so long ago be able to see, it would seem, into our future?

When we study the globe and consider things synchronically from the perspective of the noosphere, we find some interesting correspondences. Pacal Votan lived in the tenth baktun cycle, between the Christian years 603-683, on the opposite side of the world but at almost the same latitude as Mecca, and just after the time of Muhammad, 570-632. Today the religion that God commanded Muhammad to found, Islam, is at the center of the terrorist storm, while the prophecy of Pacal Votan and of the Mayan calendar in general has placed the end point of history at 2012, a mere eleven years after the Inevitable Event that placed Islam at dead center of the world mind. How could this be, the religion of the prophet Muhammad and the prophetic timing of the Maya so close at hand? Is this some revelation of the Nine Lords of Time, or of God speaking through them-and who are the Nine Lords of Time? Are they destined to return? Could it be that the Nine Lords of Time, like some higher dimensional Lords of the Ring, are a manifestation of time to be released, one by one, in an annual sequence during the last nine years of the Great Cycle, 2004-2013?

These Nine Lords of Time are known in Mayan as the Bolontiku; their counterparts are the Oxlahuntiku-the nine lords and the thirteen lords, respectively. The nine lords govern the underworld, the time of the third dimension, which is why they are depicted in descent to the sarcophagus of Pacal Votan, while the thirteen lords govern the heavenly realms. The nine govern the nine vigesimal orders of number, while the thirteen govern the form of time. The dedication of the tomb of Pacal Votan in A.D.692 corresponded to the Long Count baktun date of9.13.0.0.0.

The nine and the thirteen lords-Bolontiku and Oxlahuntiku-are present in the very numbers of the dedication date. And the nine and the thirteen also governed the prophecy of Quetzalcoatl, of Thirteen Heavens of decreasing choice and Nine Hells of increasing doom. This prophecy refers to the sequence of twenty-two 52-year cycles that began in A.D. 843 and ended in A.D. 1987. Six years later, in 1993, during the first year of the 260th and final katun cycle of history, I began to decode the prophecy ofPacal Votan, some forty-one years after the tomb had been opened.

Telektonon, the Prophecy of Pacal Votan, was meant to be decoded precisely and only at the moment of its absolute usefulness, kin 144: Yellow Magnetic Seed, July 26, 1993. This moment defined the Seven Years of Prophecy, 1993-2000. It also established the validity of the Dreamspell or WIzard's Count as the reformulated Year-Bearer Count of the Chilam Balam, the jaguar priests or night seers. That is, the Telektonon Prophecy could only have been triggered when kin 144 occurred as the Year Bearer, Yellow Magnetic Seed, kin 144, correlated to the Gregorian date of July 26, 1993.

The precise revelation of this prophecy on kin 144, July 26, 1993, was dependent on and could only have occurred because a prophetic count, the WIzard's Count, derived from the prophetic tradition of the Chilam Balam, had already been established. This "count" provided the basis of *The Mayan Factor* and was woven into the conclusion of the Prophecy of the Thirteen Heavens and Nine Hells, Harmonic Convergence, August 16-17, 1987. And the precise number of kin 144, the revelation date of the prophecy of Pacal Votan was meant to accord with the Book of Revelations number of the elect, 144,000 (Revelations 7:4). In fact, that very day I discovered that the number 144 transliterated from the vigesimal count into Arabic numbers is written 7.4, which is the same as the chapter and verse number where the 144,000 are first mentioned in the Book of Revelations (7 = 140, or 7 x 20, plus 4 units = 7.4). More uncanny stuff-but not so to the noosphere, the Earth's mind of time. In the noosphere everything is woven of number, and number is intrinsically synchronizing, which is why numerology has held such a fascination for the human mind throughout history.

When I first began to decode this "final prophecy of the Maya," or Telektonon, a prophecy that commands the return to natural time through the Thirteen *Moon!* 28-Day calendar, I was stunned and not a little frightened. I had already taken many steps to remove myself from mainstream society and conventional life. But now I was dealing with a genuine prophecy, and I was to be its messenger. This was not my invention. The world is filled with enough madmen and lunatics spouting their particular visions. How could I ever make the world understand that this vision, this prophecy, was real, that it was the fulfillment of a Mayan who lived and died thirteen

Vinal position, one dot = 20 kin, 7 = 140 (7 x 20)

Kin position, one dot = 1 kin, 4 = 4

Written 7.4 = 144 (140 + 4)

7.4 = 144

centuries ago, and that his tomb was the vehicle of this prophecy, a prophecy meant precisely for this moment in time-the biosphere-noosphere transition?

I returned to the texts of the Chilam Balam, whose prophetic stream includes day prophecies, year prophecies, katun (twenty year), and *ahau* (260-year) cycle prophecies, prophecies of Quetzalcoatl, and finally the special prophecies of the mysterious Antonio Martinez and the coming religion of the Hunab Ku, "One Giver of Movement and Measure." I realized that the meaning and fulfillment of this entire prophetic stream was crystallized in the Harmonic Convergence and its 26-year aftermath. Chilam Balam, the jaguar priests, are the wizards of time, inspired by the guidance of the Bolontiku, the Nine Lords of Time, the Lords of Number. By command of the Supreme One (Hunab Ku), the wizards are directed to establish clear paths of rekindled knowledge. The ancient Maya prophesied their own destruction. Therefore, within the prophecies of the Chilam Balam they laid the seeds of their reconstruction-but not simply as Maya. The reconstruction had to be planetary, for their destruction in the sixteenth century was also an act of planetary consequence-the plunging of the entire world into the prison of artificial, mechanized time. The fulfillment of the prophecies along with the new dispensation of time had to occur before the end of the cycle, 2012.

For this reason, the master synchronic architect of the Mayan prophets, Pacal Votan, prepared the way with the discovery of his tomb in 1952. This phenomenal act of prophetic release meant that now the messenger of his prophecy of time had to be summoned to his task. This was to be my role and responsibility-one that I accepted, for I saw no other choice. The course of my life from the moment I had experienced an illumination atop the Pyramid of the Sun at Teotihuacan in 1953, to that moment forty years later, in 1993, when I first perceived that the meaning of the prophecy Telektonon was a single trajectory. The noosphere had found another chip. After all, if the noosphere is the mind of time on Earth, prophecy is just another way of radializing time, of bringing it around from one point in a cyclic motion to another. My preparation had been going on for a while.

Once I had received knowledge of the prophecy of the Thirteen Heavens and Nine Hells from Tony Shearer back around 1970, I prepared myself by study and meditation for the moment of the Harmonic Convergence. It may be asked: If the Harmonic Convergence concluded the cycle of the Nine Hells, then why are we experiencing the apocalyptic events of Hell on Earth now, some fourteen years later? What was the actual meaning of the Harmonic Convergence, and what does it mean for us today?

First of all, the Harmonic Convergence refers to the entire 26-year cycle from White Galactic WIzard to Yellow Galactic Seed, 1987-2013. The conclusion of the Ninth Hell refers to the release of humanity into the 26-year time zone of free will, choice, and decision in preparation for the completion of the cycle, omega point 2012, and the Galactic Synchronization, 2013. The full and true Harmonic Convergence will not be complete until the Earth itself becomes the Galactic Seed, 2013. If the twenty-six years are the free will zone of choice and decision, what are we to choose and decide? The choice is, as I was quoted in the *TVtlllStreet Journal* in 1987, "between a 'new age' and all-out destruction." The decision is to accept the knowledge of the error in time (12:60) and the choice is to return to natural time (13:20) by means of the Thirteen Moon!28-Day calendar. This the prophecy ofPacal Votan fully affirms.

In principle, the Harmonic Convergence refers to the converging of all aspects of reality in a great, all-unifying harmony. The initial moment of the Harmonic Convergence that witnessed hundreds of thousands of humans congregated at sacred sites worldwide at the dawn of August 16, 1987, Blue Electric Eagle, was a pure prophetic enactment of humanity being harmonically converged with the natural order of time. This visionary moment demonstrated that, indeed, Harmonic Convergence was possible. The wave of unconscious energy released through the collective psyche of the species as a result of this moment of prophetic purity reached a climax with the fall of the Berlin Wall and the end of the Cold War.

Early 1990, the role of the Harmonic Convergence in this process was echoed in the mass media as the question, "Did the Harmonic Convergence bring the world closer to peace?"[1] Before the question could be answered, the next phase of the prophetic end time began to unfold. Rather than determine how to make the choice to enter a genuine new age, the human race defaulted to an absolute materialism. It was this collective turn taken by the planetary leadership, now known as the G-7, that engendered the apocalypse we are now witnessing. In this turn, the Harmonic Convergence was forgotten, deleted as it were from the official public texts of contemporary history. The stage was thus set for the next revelation, the Telektonon of Pacal Votan.

In distinguishing the Seven Years of Prophecy, 1993-2000, Magnetic Seed-Resonant WIzard years, the Telektonon of Pacal Votan makes an absolutely clear analysis of the human condition at this time: unless the human race forsakes the false time governed by the instruments of the Gregorian calendar and the mechanical clock by the year 2000, then it will face its own destruction, biospheric collapse, and Hell on Earth. But ifby the year 2000 it can make the decision to reject the calendar of the false

12:60 time and adopt the perfect instrument of the natural 13:20 time, the Thirteen Moon!28-Day calendar, then humanity will have been able to make the choice of entering a new age. For the new age, so longed for, is actually only realizable as a new time. A new time can only come about by the rejection of the instrument that holds in place the hallucination of the old time, replacing it instead with an instrument of such perfect harmony that it has no history, but is truly post-historical.

For this, the messengers of time, myself and my wife, embarked on a course of being galactic gypsies, with no visible means of support, in order to warn the world of the meaning of the time and to offer to the world the opportunity of changing its course by changing its calendar. When the year 2000 rolled around, humanity remained unmoved. The Gregorian calendar, the Vatican, and the G-7 appeared to be as strong as ever. The prophecy did not fail, but humanity did fail to heed the message regarding artificial and natural time. In this process, as "messengers" my wife and I were only playing out an archetypal role. So beginning with the Blue Galactic Storm year, July 26,2000, thirteen years after the White Galactic WIzard year, humanity was placed in judgment. This is merely a matter of natural karmic law.

The four-year cycle, 2000-2004, became known as the prophetic cycle of the Harrowing of Hell. Just as Jesus was crucified and then experienced the Harrowing of Hell before his resurrection, so too humanity has been crucified on the technospheric cross of false time, and is now in the Harrowing of Hell awaiting the resurrection in natural time. For having failed to take seriously all its messengers and all its prophets of all times, humanity was now to pass en masse through hell. Since the Galactic Storm year, 2000-2001, completed a Seed-Storm Year bundle begun in the fifth year of prophecy (1997-1998), the 2000-2001 Blue Galactic Storm year was the year of evaluation. The full judgment was withheld until the Solar Seed Year, 2001-2002.

The Solar Seed year is the second year of the Harrowing of Hell, but the first year of the next four-year Seed-Storm Year bundle. The first half of the Yellow Solar Seed occurs during the second half of the first official year of the third millennium, 2001. The moment was now ripe for the Inevitable Event, for it was also the same Gregorian calendar year as 1945 and 1973. In the Telektonon Prophecy, verse 120, section 19, it is declared, "When the Lunar Moon has overflowed its banks, the G-7 will be no more." So it was on the twentieth day of the Lunar Moon (Blue Self-existing Monkey, galactic signature of Karl Marx), that the Tower of Babel, the World Trade Center, was struck a fatal blow, while the Pentagon, magical structure of the priests of the Dark Lords of Matter, built in "Hell's Bottom," Arlington, Virginia, during the Ninth Hell cycle, was mortally penetrated.

It is still the time of the Harmonic Convergence. The Harrowing of Hell represents the last cycle of respite for humanity. By July 26,2004, Blue Crystal Storm, humanity must make the "choice between a 'new age' and all-out destruction" when it will have the last opportunity to make the decision of rejecting the old time and calendar and replacing it altogether with the instrument of perfect harmony, the Thirteen Moon!28-Day synchronometer. Only such a harmonic organizing principle, universally accepted and applied, can ensure a time of harmony and the fulfillment of the Harmonic Convergence.

Is it possible that within the noosphere the Harmonic Convergence is awaiting its next climax on the Day Out of Time and the Thirteen Moon Synchronization date of the year 2004, that is, on Gregorian July 25 and 26, 2004? One reads through the *Journal of Calendar Reform* and the many texts and arguments for calendar reform that were put forth in the 1930s and feels the missed opportunity. Even Gandhi was in favor of a new universal timing standard as a principle means of unifying humanity in peace. Did World War II occur because the League of Nations failed to enact the calendar reform on January 1, 1933, as it had fully proposed? The League of Nations had promoted calendar reform as one of its major agendas since the 1920s. This world organization failed at calendar reform and at preventing World War II. Who or what prevented that reform from happening, and why?

Let us look at the matter again. In order to remain constant, any perpetual calendar ordered by the seven-day week requires the principle of a null day, a Day Out of Time that is no day of the week at all. This was a feature of the preferred Thirteen Moon!28-Day calendar proposed by the International Chamber of Commerce, as well as the other modified calendars of twelve months. Reading through archives and literature, pro and con, it is clear that what defeated the calendar reform was the conservative, Vatican-inspired objection to the null day, the Day Out of Time. To observe such a day, which is no day of the week at all, it was argued, would break the weekly succession of days set in motion by God at the beginning of creation-such a break would plunge the world into barbarism, chaos, and war. "The adoption of the weekly cycle was a triumph of Christianity to the detriment of paganism. The cult of Christianity is inextricably connected to the week whose origin reaches to the dawn of history. "2 Therefore, it was contended, to interrupt this weekly succession would destroy Christianity itself. As a result of such propaganda, the League of Nations faltered, no calendar reform occurred at all, and the world was plunged into war-a war that ended with the beginning of the age of terror. This terror is now running its course, like a fever sweeping through the biosphere. WIll it be possible to consider again what the League of Nations faltered on and what the United Nations simply chose to ignore?

Who really lost out when the calendar reform issue was tabled in 1956 at the United Nations? It was not God who invented the week as we know it, but more likely the Babylonians. Behind the week is the mystery of the whole number seven, recalled in the Epic of Gilgamesh as the seven wise men ofU ruk. But the Babylonians clothed the seven-day cycle in the raiment of the sun, moon, and five principle planets-Mercury, Venus, Mars, Jupiter, and Saturn-to which the days of the week have no actual relation. Seen in this way, the seven-day week was intended to be a means on Earth for reenacting some remote mythic reality related to these seven heavenly bodies. What was that mythic reality, on what hidden knowledge is the count of the seven-day week presumed? Borrowing from Babylon, the ancient Hebrews incorporated the seven-day week into their purely lunar calendar. And it was thus from the Hebrews that the seven-day week entered the Julian and then the Gregorian calendars. It was not, however, the Romans who adopted the week, but the Christian successor to the Roman empire, the Church of Rome. What truly was the origin of the week? And what relation or meaning does it have in the first place? Even to F. H. Colson, a Christian scholar writing as late as 1926, the seven-day cycle is described as "that intruder the week, consisting indeed of a fixed number of days, but paying no regard to months or years."3

With its fixed number of days, the week makes sense only in a system of time reckoning that is synchronic rather than astronomical in nature. To those whose minds are set on the synodic lunation cycle as the absolute standard, as does the Gregorian calendar, the week is what ensures the greatest disorder in the calendar-and yet it is defended to the death. The week gains its power from the whole number 7, and not from the jumble of names in which the week is clothed. The seven-day cycle is the closest whole number that corresponds to the four phases of the moon, which, as whole numbers, are also seven days each. In fact, the fixed 7-day cycle only makes sense within the context of the Thirteen Moon calendar with its 28-day cycles divisible by four and seven, creating the annual cycle of fifty-two perfect weeks-plus the Day Out of Time, ensuring a perpetual harmony.

Was the Babylonian naming of these days-with only one name a reference to the moon-meant to be a cover-up of some earlier knowledge, a knowledge both more purely lunar and harmonic, such as the Thirteen Moon calendar? Is all of history, then, just based on a cover-up of harmonic time? What would really happen if there were a Day Out of Time, no day of the week or month at all? Would old belief systems dissolve? Would there be a psychic release, a sense of freedom? Since 1993, sixty years after the failure of the League of Nations to enact calendar reform, the Thirteen Moon Calendar Change Peace Movement has made certain that the

Day Out of Time is the day for forgiveness of debts, of release into timelessness, and for experiencing the full meaning of time is art. This day is now annually celebrated worldwide by increasing numbers of humans. All the while, the Gregorian world has arrogantly plunged blindly forward, locked into its mechanistically chaotic timing sensibility, only to receive the mortal wound of the Inevitable Event, causing it to fall through the seams of its own making, the technosphere. A Day Out of Time, it seems, might be a relief to a world now consumed by chaos, barbarism, and war.

> He sends down from the sky water for your drink and to grow trees for your benefit. With it He grows for you crops, olives, date palms, grapes, and all kinds of fruits. This is sufficient proof for people who think. And He commits, in your service, the night and the day, as well as the sun and the moon. Also, the stars are committed by His command. These are sufficient proofs for people who understand. (Quran, 16:10-12)

Yes, the biosphere is a set of proofs for people who understand, an immense order that reflects and transforms the cosmos through the very process of life on Earth. Who are the people who understand, and where are those people today? Who will speak for the biosphere? Who will take the message of the advent of the noosphere to the statesmen and the politicians that they may take heed and note that what is occurring is not really in their control at all, but is the climax of a phenomenal process known as biogeochemical combustion? And who will declare that this climax will inevitably shake all the towers to their foundations and level the mountains of industry to a flat plain from which the noosphere will manifest as the equality of telepathy for all?

This book, and this ninth chapter in particular, is like a long meditation, a vision, inspired by the dark tragedy of the Inevitable Event. As a vision, these thoughts are both those of one solitary individual, albeit a noospheric chip, and the continuation of the vision of that ancient seer, Pacal Votan. Could it be that Pacal Votan in his time knowledge rehearsed all these words, these thoughts, this entire vision itself? On the sarcophagus lid of Pacal Votan is a cross, not the cross of death as in Christianity, but the cross oflife, a tree, a cosmic tree. Across that tree is a serpentine band with two dragonlike heads, one at each end. This band represents the galaxy. The head at each end represents the galactic vision that sees equally into the past and the future. Sometimes this is called the *zuvuya,* the reflex of cosmic memory. The entire galaxy, the Milky Way, is this reflex of cosmic memory. What if all of humanity is meant to be opened into this cosmic memory? This is what is meant by the advent of the noosphere, the mind of time on Earth, the mental envelope of planetary consciousness arising from the grave crisis of the biosphere through which we are now passing at this very moment.

Because we are now defining the noosphere does not mean that the noosphere was not known to those ancient Mayan time scientists like Pacal Votan, especially if the noosphere is defined as the mind of time on Earth. To such time scientists a NET year would be no time at all, or rather, it would be but a single, fifty-two solar orbit rehearsal of life. How many such rehearsals would be necessary before something was finally learned? The Prophecy of Thirteen Heavens and Nine Hells, for instance, was actually just the passage of twenty-two NET years. Is this how long was needed to see if the human race would learn anything? And if Pacal Votan and Muhammad were around today, what would they think and do?

Knowing the nature of the synchronic order of time, someone like Pacal Votan, obviously one of the most singular of all personalities of the Mayan world, would have possessed a consciousness like the two-headed galactic band, the consciousness of galactic memory. In the zuvuya of galactic consciousness, peering into the deep past is peering at the present from the future. This is possible in the radial time of the fourth dimension freed from the strict linear construct of past-and-future-time. From some higher state of mind, from a mind and consciousness identified completely with the galaxy, this has all been foreseen, this has all been written in a book. Now we who are in this present state of end-time consciousness are being asked to self-reflect. We are being asked if we can manage to hold the noospheric lens steady enough to see that the technosphere had to end, and that the technosphere could only end by the means by which it maintained all of the biosphere in its control, by terror itself.

Yes, we are all a bit like the frog at the bottom of the well whose vision and perception of the universe is circumscribed by what the opening of the well allows it to see. The well is the technosphere, both as a planetary apparatus and as a belief system. On the other side of that well is a vast world that we have actually never seen. This vast world is the noosphere and beyond. And what is beyond? Many of the traditional peoples say that we originally came from the stars. If we came from the stars, are the star people still out there? And what do the star people think of all that is happening on Earth today? And what do they think of their lost star children, many of whom now believe it is childish to think that we came from the stars? And if there are star people still out there, wouldn't they be too sophisticated to think they could get anywhere in a rocket ship, or even a UFO, subject to the en tropic degradation of the material plane?

No, they wouldn't rely on such primitive technology, but would be in tune, as is the rest of the universe, with the universal frequency of synchronization and the capacity of time to travel at an infinitely instantaneous rate of transmission. This being so, we could receive telepathic messages from the star people-if we were ready for it and knew precisely why, when, and how such communication occurs. In fact, as we have said earlier, the noosphere is the region on Earth for the receipt and transmission of messages of cosmic consciousness. This being so, the future of human evolution, of *Homo noosphericus,* would be nothing less than to become a medium of cosmic consciousness, adapting to the expanded measure of Noospheric Earth Time, traveling back to the stars but through the superior knowledge of the actual laws of time.

In writing this book, I have only a sense of duty, of responsibility to expand the vision of my fellow earthlings from the technosphere to the noosphere. In the Law of Time there are no arbitrary distinctions between reason and art, science and prophecy. The universe really is a giant radio transmitter. This book is as much an exposition of matters generally not thought about, as it is a personal testament and vision. In that regard this text is a portion of a broadcast transduced and received by me functioning as a noospheric chip, on behalf of the biosphere-noosphere transition.

In the tomb lid ofPacal Votan are impressed all the frequencies of this broadcast transmission called *Time and the Technosphere.* I know that Pacal Votan took an earthly form for just this purpose. He left his body in that tomb, a jade mask being the actual impression of his face. According to Chinese and ancient Mexican tradition, jade is the stone of immortality. Was Pacal Votan an immortal? Where did his consciousness and spirit go once his body was placed in the sarcophagus? Why did he leave the Telektonon, the Earth Spirit Speaking Tube, running from the tomb to the temple floor at the top of the Pyramid of the Inscriptions? Knowing the frequency of the synchronic order governing the psi bank of the mental envelope of the Earth, did he place his thoughts in the noosphere to be released at the precise moments of the opening of his tomb and the decoding of his prophecy? And then where did he go?

Beginning in 1980 I became conscious that, from time to time, I had been receiving telepathic transmissions. I soon identified these transmissions as coming from the star Arcturus. In *Earth Ascending,* originally written in 1983, the nature and program of the psi bank, the control panel and regulating mechanism of the noosphere, became known to me. This was followed by *The Mayan Factor,* and then the decoding of the Law of Time. At the moment when I had just finished writing *The Mayan Factor* in 1986, I received the inspiration for what is called the planet holon, the twenty-part structure accommodating the program of the twenty solar seals. When I received this image it was known to me as the program of the Arcturus Protectorate. As I understood it, the Arcturus Protectorate was established by the "star people" as a kind of protective time shield around the Earth. By means of the activation of this time shield, the star people would know that some of the earthlings were waking up to time and beginning the conscious activation of the noosphere. This is exactly what came about through the *Dreamspell,* beginning in 1990.

The *Dreamspell* is such a radical break from the past, and such a pure statement of the history-less fourth-dimensional future, that in order to comprehend it, in 1992 I wrote two other books, one called the *Arcturus Probe: Tales and Reports of an Ongoing Investigation* and the other *The Call of Pacal Votan: Time is the Fourth Dimension.* Some people will say that the *Arcturus Probe* is an exaggerated fantasy, yet I cannot deny what a purely telepathic process it was to write that book, which fills in many points of the psychomythic cosmology of *Dreamspell: The Journey of Timeship Earth 2013. The Call of Pacal Votan* was originally titled *A Treatise on Time Viewed from its Own Dimension,* and is a purely scientific description of the mathematical codes of the Dreamspell in relation to the biosphere-noosphere transition. The one text purely imaginal, the other purely rational, yet both were two sides of the same coin, a rendering of the Law of Time in its cosmological and mathematical comprehensiveness. Then came the even more abrupt break with my conventional past, the Telektonon Prophecy of Pacal Votan.

I bring up all of these personal details because people often ask where it all came from and where I got my ideas. But more importantly, I bring them up because, as I have been indicating, all of this was foreseen and known by a seer in another time, in another world-Pacal Votan. And this process of one person hiding a prophecy and another person in another time finding that prophecy, this is also a further revelation of how the noosphere functions. The French mathematician and philosopher Charles Henry (1859-1926) declared, "As the individual becomes more collective, the collective becomes more individual."4 As individuals, we humans are meant to grow into the collective mind known as the noosphere, while the noosphere, which is the collective, could only be known through its incorporating itself through the individual. I am one such individual, a noospheric chip. The noosphere could not be understood if it did not take on a personality or a number of personalities. Otherwise the noosphere would remain in the purely theoretical state or condition.

Now that the Inevitable Event has occurred, the noosphere is recruiting. The program of recruitment is called the Campaign for the New Time. This is a genuine campaign, but a peace campaign and not a military one. As a campaign it must reach its climax by mid-2004, a full eight years before the end of the cycle, northern winter/southern summer solstice 2012, the noosphere's omega point. The reason for this is that any later than 2004, the prophecy will be negatively fulfilled. A minimum of two Seed-Storm Year bundles-eight years-are needed to creatively destructure and then to reconstruct. The Inevitable Event was a sign of God showing us that the end had already occurred. Now all the forces of goodwill and harmony must be channeled into the conscious act that will coincide with the noosphere's readiness to manifest through all of us, and that is the calendar change-the untried solution, the most important unfinished business of the last millennium.

This is no longer a matter of debate. There is no other solution that hasn't been tried. There is no other way that the noosphere can become collectively conscious than through the human adoption of a timing system of perfect harmony. By meeting our deadline with the noosphere at the time of the Great Calendar Change in the middle of the year 2004-0n the White Spectral Mirror, Day Out of Time, the day of release from the old, and Blue Crystal Storm, the day of entering the new-then the biosphere will harmonically converge with the noosphere, and the *Pax Cultura*, *Pax Biospherica* may begin in all seriousness. Such is the advent of the noosphere.

What did the Maya foresee for 2012? What is Pacal Votan's vision? Pacal Votan's vision is the fulfillment of a prayer-a prayer that I have heard through the silence:

> o Hunab Ku, One Giver of Movement and Measure, be kind to the children of this Earth. We see how every year you give and how every year you take away. This is so the children of Earth can learn to be free and dwell only in your presence. Yet we know that the children of the Earth must undergo their lessons as well. And we know that they must enter the dark time of testing. 0 Hunab Ku, in their time of darkness, in their time of forgetting, do not forget them, spare those who are good, if you can. Prepare for them the Second Creation you have promised to the righteous among them. Break for them the spell of false time before they destroy this Earth! And show to them the splendors of the Second Creation, of the New Heaven and the New Earth. Show them these wonders before the Great Cycle is finished, so that they may yet wake up from the sleep of time. And 0 Hunab Ku, may the nine great Lords of Time return to Earth! May they teach the children of the Earth how to become a race of wizards, tilling the noosphere as they have tilled the Earth. 0 bring to these children of the Earth a good Time, a Time of Peace as long, at least, as was their time of history and war, and then let that Time of Peace be multiplied sevenfold!

And here is the vision that fulfills the prayer:

> Planet Earth, Rhythmic Solstice, Blue Resonant Storm year, 2012 Omega Point. The Earth's axis is tilted so that the North Pole is at its farthest from the sun, and the South Pole is at its closest. High above the Earth an iridescent, thin rainbow band

> arches from one pole to the other. A second band does the same, but over the opposite side of the Earth. The Earth slowly rotates beneath these two rainbow bands that remain unchanging, constant in their positions, one of them directly opposite the sun, the other holding a position on the midnight side of the Earth's turning. The band facing the sun is the Earth's day alternator. The other band is the Earth's night alternator. If we glide across the surface of the Earth we see that everywhere there are small encampments, circular and radial in formation. Yurts, tents, adobe, and earthen structures-and gardens. The once-teeming cities are eerily silent, yet still inhabited, though with far fewer people than during the time of the technosphere. And everywhere in the small encampments people are reverentially gathered. This is the moment of the closing of the cycle. It is as if everyone is drawing in a single breath at the same time, and exhaling it at the same time. From wherever you are, even though there are clouds, you can see the shimmering iridescent rainbow alternator, either by day or by night. Clusters of people are gathered in circles of deep meditation. Around them the children laugh and play. "We are the noosphere," a voice is repeating, "we are the noosphere." This sets up a vibration and people everywhere stop in their tracks, they look toward the sky. They listen. "The galaxy is renewing our thought. The galaxy is renewing our mind. The galaxy is renewing our world, this Earth. Let us go home and learn some more." Drums and chanting fade into the twilight in one place and beckon the dawn in another. "Let us say the prayer again: '0 Hunab Ku, Allah, One Maker, Our Creator, we submit to You. Continue to keep us whole, continue to broadcast the noosphere through us, continue to let us listen to the higher collective voice that we are so that tomorrow we may register another kin, another NET minute in the glorious journey of Timeship Earth toward its goal of universal unification in your One Unending Thought, 0 Hunab Ku, do not forget us, now, and do not let us forget you! Thank you for the closing of the cycle! May what we have learned stay with us and become the inheritance of our children for seven times seven generations to come, and may you bring us another cycle for all our descendants to live in continuing peace and harmony!'"

This is the vision. I am a noospheric chip. Do not turn away from your visionaries. Your prophets are still calling you and teaching you, if you would pay attention. There is much to do and learn. There is nothing more joyful than the positive construction of harmony.

Time and the Technosphere, First Draft Transmission completed,
Self-existing Owl Moon Dali 15,
Yellow Solar Seed Year
Kin 42, White Electric WInd:

I activate in order to communicate
bonding breath
I seal the input of spirit
with the electric tone of service
I am ptided by the power of death

First draft edited, completed, and sealed,
Overtone Peacock Moon Seli 2,
Yellow Solar Seed Year
Kin 57, Red Overtone Earth

I empower in order to evolve
commanding synchronicity
I seal the matrix of navigation
with the overtone tone of radiance
I am guided by the power of life force

Second draft edited, completed, and sealed,
Rhythmic Lizard Moon, Seli 23, Navigation Tower
Yellow Solar Seed Year
Kin 106, White Lunar World-bridger
Pacal Votan Clear Sign

I polarize in order to equalize
stabilizing opportunity
I seal the store of death
with the lunar tone of challenge
I am guided by the power of endlessness
I am a galactic activation portal
enter me

May all beings be well and happy!

May this text and these words inspire by their truth.

If I have left out or omitted any logic or fact, or unnecessarily maligned any truth or being by my endeavor to convince through clear argument, may the Creator know that no such malice or oversight was ever intended. I am but a human transmitter, a noospheric chip. My ability is commensurate with the frailties attendant to one too long steadfast on a path too little comprehended by many of those around him.

May the truth prevail!

Jose Argüelles, Ph.D.
known in prophecy's light as
Valum Votan, the Closer of the Cycle
devotee of Bolon Ik

Brightwood, Oregon
Cascadia Bioregion
North American Plate

▼

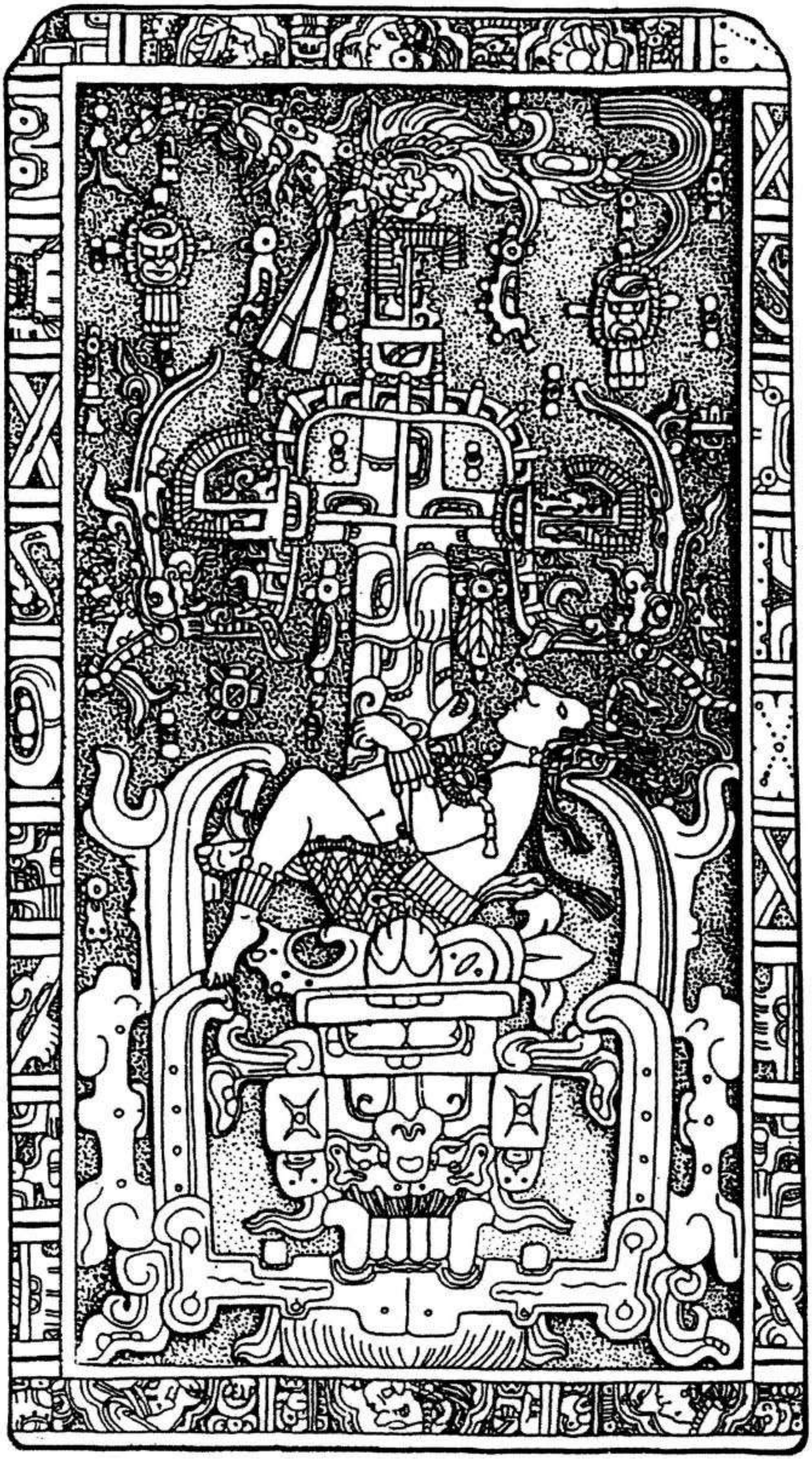

Sarcophagus lid of the tomb of Pacal Votan 9.13.0.0.0 (A.D. 692)

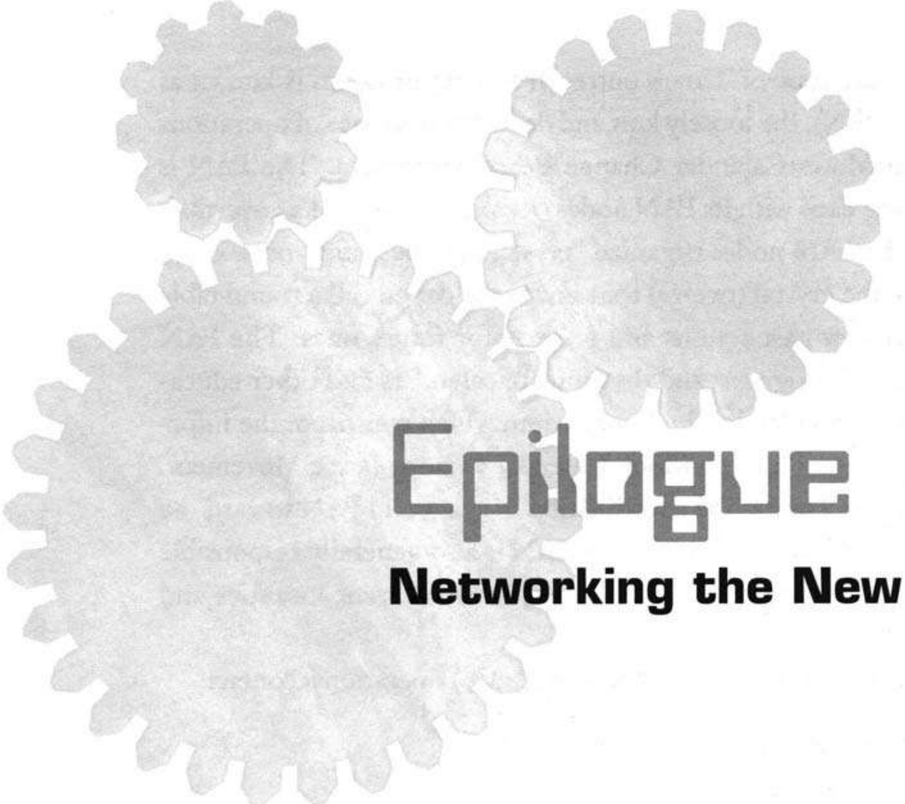

Epilogue

Networking the New Time

TO READERS for whom the notion of calendar change, much less that of a Campaign for the New Time, is totally novel, yet who may have an interest in learning more about and participating in the preparation for the Great Calendar Change of 2004, there is a place to begin and a way of going about it.

The official coordinating agency for the Campaign for the New Time and the World Thirteen Moon Calendar Change Peace Movement is the Foundation for the Law of Time. As a nonprofit public charities educational corporation, the Foundation for the Law of Time provides the communications and information hub for the Planet Art Network (PAN). The Foundation also publishes calendars, tools, and literature about the Law of Time and sponsors various educational programs, periodic seminars and congresses, and projects such as the new time educational caravans and the development of garden-oriented land bases.

For general information on the Law of Time or the Campaign for the New Time and its educational programs contact:

Foundation for the Law of Time, World Headquarters
Post Office Box 513
Brightwood, Oregon 97011 U.S.A.
Tel: +1 (503) 622-1976
Fax: +1 (503) 622-0198

The Foundation for the Law of Time's outreach activity program is known as the Planet Art Network (PAN), the loosely knit and decentralized base of operations for the World Thirteen Moon Calendar Change Peace Movement. The PAN is organized into Bioregions, each with its PAN nodes (local chapters) and a governing Bioregional Council. The PAN nodes organize "crystal day meetings" once every thirteen-day wavespell-the crystal (twelve) tone day is the day to call a round table or court of the kin to review past actions and prepare for future ones. The PAN nodes also offer educational programs and disseminate calendars and other educational tools and literature provided by the Foundation, which constitute the information outreach of the World Thirteen Moon Calendar Change Peace Movement. The Campaign for the New Time is the coordination of the world PAN toward the goal of the Great Calendar Change of 2004. The PAN is also generally responsible for organizing the annual Day Out of Time events in its different localities and bioregions.

For information about local Planet Art Network (PAN) operations, contact:

Foundation for the Law of Time/Gnomicile Land Base
World PAN Coordination Center
PMB 267
105 North First Avenue
Sand Point, ID 83864-1301 U.S.A.
Tel: +1 (208) 255-1172
Fax: +1 (208) 265-9107

The official Web site for the Foundation for the Law of Time, Planet Art Network, World Thirteen Moon Calendar Change Peace Movement and Campaign for the New Time is:

www.tortuga.com

This Web site is continuously updated and offers a great range of both practical and theoretical information about the Law of Time, the Campaign for the New Time, ongoing and upcoming events and educational seminars, listings of major PAN nodes worldwide, Law of Time archives, and links to other relevant Web sites.

A major aspect of the Campaign for the New Time is the development of a program of "strategic alliances," information about which is also available at www.tortuga.com. If you belong to an organization that you feel could be in alignment with the goals of the Campaign you can check out the Strategic Alliance page of the Web site or directly contact the Foundation for the Law of Time, World

Headquarters. The goal of the Strategic Alliances is to establish a "synchronization for peace force"-a coalition of a wide spectrum of organizations that are in alignment with the calendar change and are willing to help organize local events and information dissemination for the Great Calendar Change of 2004.

See also Appendix II, "The World Thirteen Moon Calendar Change Peace Movement and the Discovery of the Law of Time: A Brief Chronotopology of Transformation. "

Appendix I

Summary Critique of the Gregorian Calendar

What Every Proponent of the Thirteen Moon Calendar Should Know

NOT ONLY is the idea of the calendar as an instrument to determine a true and accurate year misleading, but solely focusing on this purpose blinds us to any consideration of time apart from duration or measurement of motion in space. The Law of Time asserts that the actual nature of time is synchronic; hence the purpose of calendars is to synchronize us in time according to various cycles whose harmonic numbers extend from and return us to a higher mental order of reality. It is a fatal error to dismiss a mathematics of harmonic perfection when it is allied with the ordering and comprehension of cycles. The pursuit of a true and accurate year totally subordinates the cyclic nature of time to the ceaseless imperfection and change that characterize the phenomenal world when it is considered as the sole factor of existence. This defines the thoroughly materialist worldview of the 12:60 consciousness.

In order to prepare the ground for a proper consideration of the Law of Time and the evolutionary necessity of the synchronic order of time as the harmonic reorganizing factor of humanity in its post-historic phase, it is necessary to expunge from the mind the error known as the Gregorian calendar. To demonstrate and expose the illogical and irrational nature of the Gregorian calendar as a standard of measure, the following seven points are presented as a simple appeal to the intrinsic logic and intelligence of any human being.

1. We require of a standard of measure that its units of measure are regular and equal with one another.

This is not the case with the Gregorian calendar, whose base unit of measure, the month, proceeds in an irregular and uneven manner: 31 days, 28 days, 31 days, 30 days, 31 days, 30 days, 31 days, 31 days, 30 days, 31 days, 30 days, and 31 days. Why would anyone use a standard of measure with irregular units? Do you know what results when a crooked standard of measure is employed consistently for millennia? It might be noted that at the time of Augustus Caesar, August was called Sextile and had thirty days, while February then had twenty-nine days. In order to honor Augustus and make him the equal ofJulius Ouly) which had thirty-one days, the twenty-ninth day was taken from February and added on to Sextile, whose name was then changed to Augustus (August). By harmonic contrast, the Thirteen Moon calendar is perpetual in that all of its units of measure are equal-28 days each.

2. The names of the months are as illogical as their uneven numbering.

January is derived from the God of the doorway; February is an obscure word referring to an animal divinatory rite; March refers to Mars, the planet and the god of war; April and May refer to goddesses of the spring; June to the wife of Jupiter; July and August are named after the two most prominent Roman Emperors, Julius and Augustus Caesar. As for the remaining months, September, the ninth month, means seven; October, the tenth month, means eight; November, the eleventh month, means nine; and December, the twelfth month, means ten. Of course, having grown habituated to the crooked standard of measure, it is easy to overlook and dismiss as innocuous the irrational naming of the months. But is it so innocuous? What do the names of these months have to do with an order of time, or even a cosmology or culture of time, which we assume a calendar might reflect? By blindly accepting this irrational disorder of names, do we not predispose ourselves to accept irrational disorder in commonplace things around us, and even within the fabric of our society, thus reinforcing in us a tendency to accept a superficial treatment of the symptoms, while ignoring the roots?

3. Leap year and leap day is the most highly touted aspect of the Gregorian calendar.

Structurally, the Gregorian calendar is indistinguishable from the Julian calendar. The only thing that separates the Gregorian from the Julian calendar is the correction of the leap year day. Pope Gregory XIII "corrected the calendar" by deleting three leap days from the years that begin the centuries, except those divisible by the number four. Leap day is the extra day that accumulates every four years due to the length of the astronomical year being 365.241299 days and not 365. It can be seen that the fraction .241299 is not quite 1/4, which would be .25. The Julian calendar did nottake this into account, hence an error crept in that caused the spring equinox on the calendar to fall some ten days behind the solar moment of the spring equinox-a fact noted by the conquistadors when they encountered the Mayan calendar, which had no such discrepancy. Thus, in 1582, Pope Gregory XIII "improved" on Julius Caesar's calendar, some 1,627 years later, by adopting the rule that there would be no extra day on centuries-OO years-except on those that are multiples of four. Hence, there was no leap day in the year 1900, but in the year 2000, a multiple of four, there was.

What is not well known is that the Vatican does not recognize the leap day in its ceremonial calendar. Why is that? In most countries where Latin-derived languages are spoken, leap day and leap year are referred to as *bisiesto,* or "bisextile" day and "bisextile" year. If leap year is every four years, why is it referred to by a word that connotes six, *sextile?* On the official liturgical Church calendar there is no February 29! Instead, there are two February 24s, and the second February 24 is not counted. If there were to be an extra day that was counted, then the system of fixed feast days would be thrown off. Instead, on leap years, February 24, the Day of the Feast of St. Matthew, is counted twice-or extended to be 48 hours. And since in the Church tradition derived from the Romans the days are counted from the first of the next month, the first always being known as the calends, the date February 24 is technically referred to as the "sixth of the calends of March" (February 24 = sixth calends of March, February 25 = fifth calends, February 26 = fourth calends, February 27 = third calends, February 28 = second calends, and March 1 = calends of March). For this reason, the leap year is known as *bisiesto* because the sixth *(siesto)* calends of March is doubled, hence "bi." Thus, February 29 is not recognized by the official Church calendar of the Vatican, and it is also not counted as a day in its liturgical calendar. February 29 only arose out of popular tradition in the sixteenth and seventeenth centuries. The nemesis of accounting for the .241299 extra day per year reveals the fixation of astronomical time in seeking a "true" year and detracts from the possibility that the issue of an intercalary, or extra, day can logically and synchronically be handled in many other ways. In the end, it is the mystique surrounding leap day that further contributes to the numbing of the sensibility of time accumulated in the instrument known as the Gregorian calendar, "the little grid of boxes that rules so much of our lives."1

4. **The word we use to describe the instrument for measuring, the Earth's orbit around the sun, *calendar,* is derived from the word *calends.***
Calends was the Latin name given by the Romans to the first day of every month. What does it mean? *Calends* is the name of the account book, the book of payments recording the monthly debts and bills to be paid! No wonder we are ruled by the philosophy "time is money!" This philosophy is rooted in the very word we use to describe time reckoning, *calendar.* A more accurate word for time reckoning might be "chronometer," or even better, "synchronometer." But perhaps the word "count" is simpler. We might just say, for example, Thirteen *Moon!2* 8-Day perpetual count.

5. **Dominicalletters code the years of the Gregorian calendar.**
Scarcely known to anyone but Vatican insiders is the system of dominicalletters that is used to code the years according to the day of the week on which the first Sunday of the year falls. Since the week has seven days, there are seven and only seven dominical letters. These are the letters a-g, where a = 1, b = 2, c = 3, d = 4, e= 5, f = 6, and g = 7. Hence, G-7 (the group of the seven most industrial nations), the name given in 1974 by the CIA to the ruling oligarchy of globalization, is totally rooted in the system of the seven dominicalletters, a-g, to which the Gregorian calendar can be reduced. It works as follows: In 2001, the first day of the year was on a Monday, hence in this year all Mondays are coded by the letter a. Counting forward to the first Sunday, January 7, Sundays this year are coded g. The letter of the year, which is always a capital letter, is based on the lower case letter that codes the first Sunday, therefore, the year 2001 is coded as the capital letter G-it is truly a G-7 year! Not only that, but by this system of "G-7" dominicalletters, it can be demonstrated that the order of the Gregorian calendar year repeats within a cyclic structure of precisely twenty-eight years, where the days of the week and the month repeat once again. Hence, the cycle 1973-2001 is a repeat of the cycle 1945-1973, which is a repeat of the cycle 1917-1945, and so forth, where the years 1917, 1945, 1973, and 2001 are calendrically indistinguishable from each other. In any twenty-eight-year cycle, there are always exactly seven leap years! Thus, the key code numbers 28 and 7 of the Law of Time and of the perfect Thirteen *Moon!2* 8-Day calendar are hidden in and even govern the cyclic order of the Gregorian calendar, whose secrets pretend to lie concealed in the Vatican archives!

Expose these secrets and show that the true harmony of time is contained in the 13:20 matrix of the Harmonic Module, which is perfectly coded by 28 and 7. Within this matrix, any set of four tones radially opposite each other, an occult quartet, always adds up to 28. There are 65 (x4) such sets that constitute the Harmonic Module, while seven (multiplied by four), is the prime factor of 28, and is the key unit holding the 13:20 matrix in place, evident in the position of the seventh vertical column on either side of which is a perfect harmony of six columns each. The seven is also the implicit number in the ratio 13:20 (13 + 7 = 20: 20 - 7 =13). But whereas the 13:20 matrix is a harmony in which the Law of Time is encoded, the Gregorian calendar is a disharmony, but one that is nonetheless governed by the Law of Time. What the Vatican attempted to destroy at the hands of Bishop de Landa in 1562, in the famous book burning at Izamal Yucatan, is redeemed by the Law of Time. Free of the obscuring, illogical irrationality of the inexact measure of the Gregorian calendar, the actual truth of the synchronic order of time may be discovered and revealed as the 13:20 mathematical code ofthe Harmonic Module, perfectly coordinated with the Thirteen Moon!28-Day count.

6. The Gregorian calendar makes day/date calculations difficult.

The twelve uneven months of the Gregorian calendar operate by a subsystem of fifty-two 7-day weeks, plus one day. Because of the irregularity of the numbering of the months, and because there are 365 and not 364 days in a year, it is almost impossible to make easy calculations month to month and year to year, while insistence on the unbroken succession of the week only compounds this issue. For example, if today on the Gregorian calendar it is Friday, May 4, 2001, what day of the week will June 4 be? What day of the week will July 4 be? There is an immediate mental block-a numbing of the mind. You have to stop and think about it. And in this numbing pause in which your cognitive brain has to be engaged, you lose your telepathic awareness, much as when you look at a clock to find out what time it is. Why should it be this way? Who benefits? The answer is, the priests (who know the tedious rules for stating that if it is Friday and it is 2001, then it is dominical e) and the bankers (who gather interest based on the confusion over the erratic disparity of days every month). By contrast, on the perpetual Thirteen Moon calendar, this day would be Spectral Moon Gamma 3-while every third day of every Moon is coded by Gamma. In fact, the third day of every week is coded by the name Gamma. Once the 28-day count is mastered, there is no need to engage the cognitive brain to figure out what day of the week Crystal Moon 3 or Cosmic Moon 3 will be-and in this way the mind is liberated into a telepathic knowing.

The system of the seven-day week was introduced into the Julian calendar at the Council of Nicea, A.D. 325, and was adopted from the Jewish calendar, which was derived from the Babylonians, for whom it was an astrological-astronomical construct. The fact that the week came from the Babylonians dissolves the argument used by the Vatican to counter the Day Out of Time. According to the Vatican, disrupting the succession of the seven-day week would disrupt an order set in motion by

God. The matter of the 7 and the 52 has a much deeper significance when understood in the higher-dimensional light of the Law of Time. A count of fifty-two 7-day weeks makes perfect sense if you have a count of thirteen 28-day months (7 x 52 = 13 x 28). The observance of the Day Out of Time, the 365th day of the year, produces a perfect and perpetual harmony. Knowing this, to continue to insist on fifty-two weeks while being unwilling to give up a twelve month count that does not have a Day Out of Time is to persist in an adherence to hopeless disharmony. Why do it?

7. What's in a name? Think about it.

What does it mean to follow calendars called the Julian and the Gregorian? A calendar is an instrument of control. The two most significant calendar reforms in history were the Julian calendar reform of 46-45 B.C., and its successor, the Gregorian, in A.D. 1582. Julius Caesar's motives had everything to do with his personal ambition and the conversion of Rome from a republican to an imperial form of government. Julius Caesar's calendar assured it to be the basis of imperial dominance. The course of the empire utilizing the Julian and later Gregorian calendar has prevailed as the dominant force now inseparable from the course of history itself. The 445-day "year of confusion" (46 B.C.), which attended Julius Caesar's reform, was matched by the second significant reform, the Gregorian, in which ten days were "lost forever," between October 5-16, 1582, so that the calendar could "catch up" with the sun.

While European Catholic countries easily accepted the reform, Protestant countries grudgingly acquiesced. Throughout the Americas, however, the Julian-Gregorian calendar was imposed as an instrument of power and symbol of dominance over the peoples the Europeans had conquered, including the high civilizations of the Maya, Inca, and Aztecs-all of whom happened to use, among other calendars, a Thirteen Moon!28-Day count. As with Julius Caesar, for Pope Gregory XIII the moment was politically ripe for a reform that would communicate itself as a means of expressing and extending power and control, but this time over the entire globe. As European dominance and control spread around the planet, even nations with their own established timing systems accepted, for the sake of "international policy," the Gregorian Oulian) calendar system for measuring the solar year. And so Western dominance over every aspect of global life became absolutely assured-until the moment of the Inevitable Event.

From its roots in the imperial ego of Julius Caesar to the timely "reform" of Pope Gregory XIII, it is not surprising that this calendar, "despite its odd quirks and the twists of history that produced it,"2 has become the standard of global civilization. Given the irregularity of the Julian-Gregorian calendar and the pursuit of the accuracy of astronomical time, history could be nothing more than a compilation of odd quirks and twists, while global civilization itself is the triumph of artificial time over the natural world. Only a species whose time sensibility had been captured by instruments of artificial measure could have become so alienated as to have produced the monstrous conundrum known as the "fast world," a civilization where money and technological advance prevail over human sensibility and the natural order. It is toward the correction of this destructive momentum that all efforts of calendar reform must now be directed.

In light of this critique, it is worth reprinting the opening section of the "United Manifesto by Advocates of Calendar Reform" that was first published in 1914 at the beginning of the First World War-some ninety years in advance of the Great Calendar Change of 2004. It can be seen that the issues of irregularity that prompted reform then are still issues today. However, the effects of not attending to these issues have only become compounded and even more complex, resulting in the chaos of a world at war with terror. Such is the nature of an error gone uncorrected over time-it only becomes more entrenched and turns into the dogmatic and hopelessly conflicted thinking of the everyday mind and its way of life.

> United Manifesto by Advocates of Calendar Reform
>
> WHEREAS we, the undersigned, have for some time been interested in a Reform and Simplification of the Calendar now in use in Western Europe, America, and elsewhere, with a view to equalizing the four quarters of the year, alleviating the irregularities of the months, and establishing a perpetual correspondence between the day of the week and the day of the month, and have supported one or other of several proposals which have been formulated for effecting these reforms; and WHEREAS said proposals usually provide for placing the 365th day of every year and the 366th day of Leap Year without the weekly and monthly enumeration; and WHEREAS we have found that in certain quarters-both ecclesiastical and scientific-Dbjections, possibly often sentimental, but none the less firmly held, have been stated to the employment and adoption of these expedients. . . Therefore we have resolved to unite in urging and advising that the very simple changes under noted should now be made in the Julian and Gregorian Calendars by international agreement. . .3

If you have a crooked standard of measure, and follow it because your parents were also following it, you have become a crooked man. It takes a crooked man to walk a crooked mile and build a crooked house. The issue of calendar reform is both logical and moral. Bad logic leads to bad morality. An error in time dooms the mind. Apocalypses are the reward for bad timekeeping. To remove yourself from the fires of your own self-created apocalypse, change your calendar. In a world of harmony there is no apocalypse.

▼

Appendix II

The World Thirteen Moon Calendar Change Peace Movement and the Discovery of the Law of Time

A Brief Chronotopology of Transformation

IT WAS during the seven-year cycle (1993-2000) when the World Thirteen Moon Calendar Change Peace Movement was being aroused around the planet, that the Law of Time was articulated in its various aspects and particulars, and the synchronic order defined in its whole system fullness. This occurred through a number of key events accompanied by a number of publications detailing the mathematical, cosmological, theological, and scientific aspects and imperatives of the Law of Time.

By the standards of the synchronic order, all events form points in a radial time set that describe a geography of time defined as a chronotopology. Coined by the late philosopher and mathematician Charles Muses, the term *chronotopology* defines an event point of meaning that creates a root meaning from which time-formed, -blossomed, or -radiated meanings emerge above a horizon of meaning.[1] The Law of Time now gives a precise field in real time to illustrate Muses's notion of chronotopology in which Time is a radial power of profound interrelations.

The following chronotopological sequence-a resume of the events in a particular geography of time, the seven years of prophecy, 1993-2000-defines the conscious emergence and articulation of the Law of Time. In this chronotopology, the "root meaning" occurred in 1989 at the Museum of Time in the form of the discovery of the 12:60 and 13:20 timing frequencies, and the subsequent working out of the implications of this discovery. The "horizon of meaning" was reached July 26, 1993, kin 144, with the decoding of the Telektonon Prophecy and the consequent development of the World Thirteen Moon Calendar Change Peace Movement. The "source meanings," however, extend back to the seventh century with the origins of the prophetic traditions underlying the discovery of the Law of Time. The seven years of prophecy established the field of radiated meanings of the Law of Time as a set of radiated event points and a body of literature and tools. Here follows, then, a brief chronotopology of the key event points of this movement:

1989	Root meaning: discovery of the 12:60 and 13:20 timing frequencies, Museum of Time, Geneva, Switzerland.
1990	Realization of the Thirteen Moon/28-Day calendar in its wavespell form.
1992	Recognition and celebration of the Day Out of Time as a planetary event.
1993	Horizon of meaning: decoding of Telektonon Prophecy and birth of World Thirteen Moon Calendar Change Peace Movement.
1994	Launching of the Planetary Moral Emergency and World Thirteen Moon Calendar Change Peace Plan.
1995	80 Days Around the World for a New Time of Planetary Peace: Chile, South Africa, Egypt. Russia, India, Hong Kong, Japan, Hawaii, and Mexico.
1995	Day Out of Time Calendar Burning and Return of Natural Time Ceremony, Serpent Mound, Ohio.
1996	First Planetary Congress of Biospheric Rights, Brasilia, Brazil.
1996	Biospheric Awareness Day, Gotemba, Mt. Fuji, Japan.
1996	Four Corners Boundary Dissolving Ceremony, Four Corners, United States.
1996	Harmonic Convergence of Humanity, Mexico City, followed by Peace March, Teotihuacan, Mexico.
1996	United Nations, Planetary Moral Emergency Appeal.
1997	Judgment Day Tribunal and World Congress on the Law of Time, Tokyo, Japan.

1997-present	Ongoing movement in Brazil to make July 25, Day Out of Time, official "Peace and Culture Day;' based on recognition of the Banner of Peace (seventy cities so far).
1998	Special peace mission to the Vatican, Rome, Italy.
1998	Special mission to Guatemala to meet with Quiche Maya elders.
1998	UR Council for the Theology of Peace, Milan, Italy.
1999	world Summit on Peace and Time, University for Peace, Costa Rica, followed by the dispatching of Special Emissaries to the United Nations and the Vatican.
1999-present	Ongoing movement in Japan to declare Day Out of Time a "Day of Festival"; more than one hundred events occurring annually.
1999	Peace March, Teotihuacan, Mexico.
1999	Seven Week Earth Wizard Seminary, Picarquin, Chile.
1999-2000	Ongoing in South America, launching of "La Karavana;' Thirteen Moon educational caravans, and in Europe, from the United Kingdom to North Africa, the New Time Caravan.
2000	Initiation of Seven-Day Earth Wizard's Seminars: Japan, United States, Italy, Russia, South America.
2000	Circumpolar Rainbow Bridge Experiment, phase I, worldwide.
2000	Fruit of meaning: establishment of Foundation for the Law of Time, Oregon.
2001	Launching of the Campaign for the New Time, Brazil. Argentina, Japan, Altai region.
2001	Campaign for the New Time PAN Thirteen Moon Leadership Conference, Oregon.
2001	Seven-day special Congress on Time and World Peace, Medellin, Colombia.

During this chronotopological sequence of radially connected event points, key writings, publications, and tools describing and defining the Law of Time and the synchronic order include:

Dreamspell: The Journey of Timeship Earth 2013, 1990-91

Thirteen Moons in Motion and *Turtle of the Thirteen Moons,* 1993, 1996

The Story of Time: The Story of Turtle and Tree, 1993

Treatise on Time Viewed from its Own Dimension (Published as *The Call of Pacal Votan:*

Time is the Fourth Dimension), 1992, 1996

Telektonon ofPacal Votan and the *Telektonon, Game of Prophecy,* 1993-1995

World Thirteen Moon Calendar Change Peace Plan, 1994-1995

Pacal Votan and Judgment Day: The Second Quranic Dispensation, 1995-1996

First Planetary Congress of Biospheric Rights, 1996

The Galactic Culture Master Plan, New Time Economics, 1996

A Brief Guide to The Thirteen Moon Calendar Change Peace Movement, 1996

Rinri Project: Four-Year Telepathic Biosphere-Noosphere Transition Circumpolar Rainbow Bridge Experiment, 1996

The 260 Postulates of the Dynamics of Time and the Evolution of Time as Consciousness, 1996

The Discovery of the Law of Time, T(E) = Art, 1996

The Principia Mathematica of the Fourth Dimension, 1996

Twenty Tablets of the Law of Time: Sixteen Year Telektonon Cube of the Law, 1997

World Congress on the Law of Time and Judgment Day Tribunal, 1997

World Summit on Peace and Time, Summary, Review, and Conclusions, 1999

7:7::7:7, Telektonon Revelation and Plasma Universe Model, 1998-2000

The Rainbow Bridge Project: 28 *Meditations on the Law of Time,* 1999-2000

The UR Papers: Documents and Appeals for a Council of Universal Religion, 1996-2001

WIth the conclusion of the seven years of prophecy, marked by the rainbow bridge experiment, phase I, the greater part of the Law of Time had been uncovered, described, and defined. What emerges from consideration of all these events and accompanying literature is an entirely new model of the universe as well as a program for the reorientation of knowledge and the social reorganization of humanity-all based on a corrected understanding of time through the implementation of the calendar change. This was all to demonstrate that there is much more to a calendar change than the mere replacing of one calendar with another, for a calendar itself bears within it an understanding of time that tacitly and unconsciously shapes our very thinking about time. What the calendar change, then, really signifies is the end of one worldview and the birth of another. WIth the Law of Time, what was once unconscious is again made conscious, leading to a radical self-awareness of how wrong we have been about time, and what we might anticipate in changing our tools for navigating in time.

Because the growth of the Planet Art Network (PAN) as the social form of the World Thirteen Moon Calendar Change Peace Movement is inextricably involved with the unraveling by stages of the Law of Time, to define the constituent principles of the Law of Time is also to define the new worldview and the social organization appropriate to it. These constituent principles include: time is art; universal telepathy; holonomic consistency and reciprocity; synchronic order; fractal and radial mathematics of the 13:20 matrix; biosphere-noosphere transition; Banner of Peace; *Pax Cultura*, *Pax Biospherica;* the psi bank and the dynamics of the evolution of time as the evolution of consciousness; fourth-dimensional time and third-dimensional space. A mere consideration of these constituent principles will make one realize to what extent the calendar change is more than just an end to history, but the establishment of a new world, a New Heaven and a New Earth. This is precisely the promise offered by the Thirteen Moon calendar change and the discovery of the Law of Time. (See plate 8, Chronotopology of Transformation.)

◆

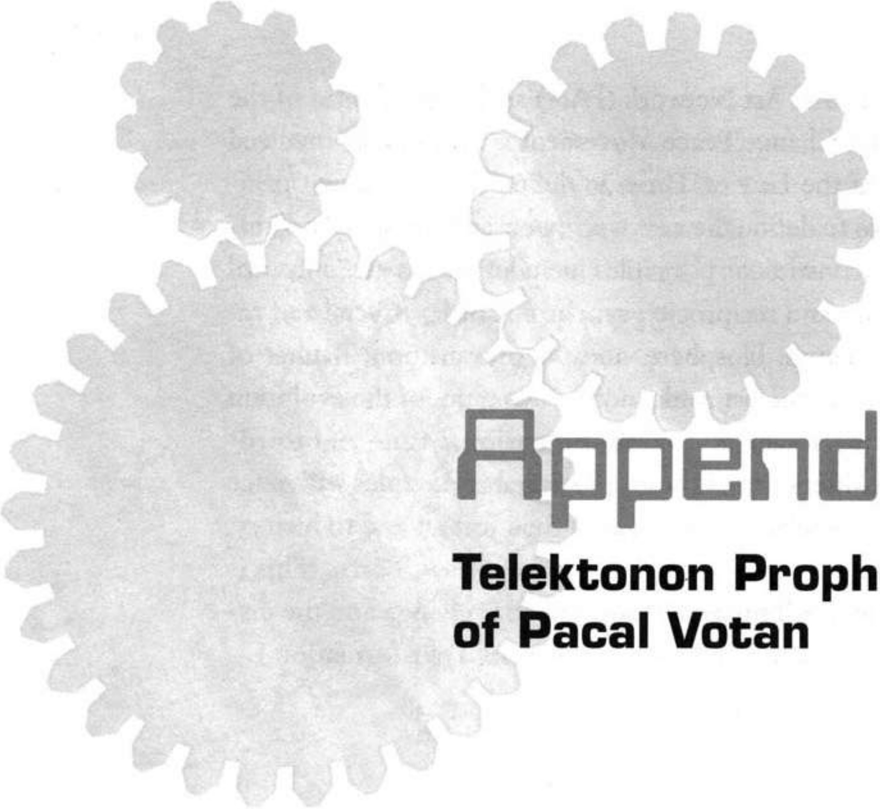

Appendix III

Telektonon Prophecy of Pacal Votan

THE REVELATION of the prophecy of Pacal Votan on July 26, 1993 was the culmination of the event that occurred on June 15, 1952, when the archaeologist Alberto Ruz Lhuillier pried open a large monolithic trapezoidal door beneath the Pyramid of the Inscriptions at Palenque, Chiapas, Mexico, and beheld what had not been seen by human eyes for 1,260 years: the tomb of Pacal Votan. A subsequent event, the lifting of the magnificent sarcophagus lid-measuring some four-by-three meters and weighing several tons-to reveal the human remains amidst a splendor of jade, including the jade mask, occurred on November 27, 1952. For some forty years afterwards, scholars debated the meaning of the tomb and the person buried within it. In the 1970s, the Russian scientist Knozerov dubbed the man in the tomb an "astronaut" or "cosmonaut" because of the similarity of the position of the figure sculpted on the sarcophagus lid to the positions of Russian cosmonauts in their little space capsules. This theory inflamed the imagination of some and was debunked by others, the end result being the enhancement of the awesome mystery of the tomb and the intention of its designer, the man buried in it.

The image of the sarcophagus lid is now world famous. Its symbolism has been interpreted to varying degrees by different archaeologists. In my book *The Mayan Factor*, the placement of Pacal Votan in his "time of power," 631-683, is defined and analyzed as occurring in the most harmonic and mathematically synchronic point of the entire thirteen baktun cycle, a point of fact that deepens the enigma of Pacal's knowledge and purpose. And herein lies the key question: What was his purpose, and why did he construct his tomb to be so elegantly buried for 1,260 years, to be discovered and opened just sixty years before the closing of the cycle, A.D. 2012? Clearly, the mystery of the tomb and its discovery had everything to do with the end of the cycle. At least this had come to be my surmise ever since the 1987 Harmonic Convergence. Like a hawk circling its prey, the meaning of the tomb finally came to my consciousness like a direct hit from the beyond. I was the prey, the hawk, the prophecy itself. Such was the revelation of the Telektonon Prophecy on the morning ofJuly 26, 1993.

The coded key was the tiled "speaking tube," by which the tomb was discovered in the first place. Suddenly, this tube spoke its meaning to me: Earth Spirit Speaking Tube, Telektonon by name. After four intense weeks of decoding the glyphs and symbols of the tomb while living in the pool house of an isolated estate in the wind-blown northern part of the main island of Hawaii, I knew that I had to abandon the dreamlike life I led in this Pacific paradise, and with my wife and companion head directly into the unknown, the first stopping point being Mexico, the source of the prophecy. After four months or so of traveling through Mexico, completing the decoding and announcing the prophecy, including a return to the tomb, we were finally directed to a converted carport in a field near the village of Ocotithin, not far from the birthplace of Quetzalcoatl. It was here where, for a period of nine mornings, I awoke dutifully two hours before sunrise to listen to the voice and write down by hand, in a specially prepared notebook, the words exactly as I heard them. Afterward, I numbered the 126 verses and distinguished the nineteen sections.

Because the prophecy itself is so inextricably connected with the Law of Time and the message of *Time and the Technosphere,* I feel it is important to present it as background information to the text of this book. In actuality, the text of the Prophecy is an integral part of the *Telektonon, Game of Prophecy,* where it is read, section by section, on a daily basis, according to the code number of the kin of the day.

THE TELEKTONON OF PACAL VOTAN

The Talking Stone of Prophecy That Unites the People of the Dawn with the People of the Book

"Then if they reject thee,
so were rejected apostles
before thee, who came with clear signs,

books of dark prophecies
and the Book of Enlightenment. "
-*HOLY QURAN*, SURA **III,** 19: 184

"All who obey God
and the Apostle
are in the company
of those on whom
is the grace of God,
of the prophets who teach,
the sincere lovers of truth,
the witnesses who testify
and the righteous who do good.
Ah, what a beautiful fellowship!

Such is the bounty
from God: and sufficient
is it that God knoweth all. "
-*HOLY QURAN*, SURA Iv, 9:69-70

INTRODUCTORY VERSES: ONE LIVING PROPHECY

1. Pacal Votan's Gospel of the Telektonon
   A Special Dispensation for the Day of Truth,
   A.D. 1993-1994, kin 144.

2. Proclamation on behalf of the three messengers
   of the awakening,
   of the three special voices of prophecy,
   the special witness of time and the last call.

3. Telektonon
   the inner sun
   the mother of all prophecy
   message of the star witness, Bolon Ik
   received and repeated by her servant in love,
   Pacal Votan.

4. In the seventh year of Harmonic Convergence
   the seven seals of the apocalypse
   become the seven years of prophecy. . .
   thirteen years in all.

Thirteen moons the path to walk;
thirteen moons the path to talk;
people of the dawn, one mind,
people of the book, one God;
one living prophecy, one people, one Heaven,
one Earth.

●● NAH CHAN: THE PALACE OF BOLON IK

5. o children of the people of the dawn, a children of the people of the book, I come as the special witness of time to remind you, especially on the day of truth, that in your origin you are one, and on the day of truth you are to make yourselves one again. For this I remind you of the Cube of the Law, Telektonon, and the 28-day, thirteen moon way that is the path of the righteous.

6. Nah Chan, Palenque, Xibalbay, Tollan, Xochicalco, Tepozteco, Amatlan: in these signs dwell my special agents, the witnesses of truth, 13.66.56 and Bolon Ik, Daughter ofJob.

7. For those with eyes, with open mind and sincere heart, this teaching is complete in every stone alignment and marking of the star command base now called Palenque.

8. New Jerusalem is the Palace of Bolon Ik, whose number is 1,728 (144 x 12), Cube of the Law. Seventeen is the power of Heaven, "solar Uranus," which is Earth in Heaven. Twenty-eight is the power of Telektonon, the Spirit Tower of the power of 7 times 4.

9. Never forget, children of the day of truth: All is number. God is a number. God is in all.

10. And everything there in the palace is laid out foursquare from the cord of Heaven, which is the command of the Telektonon. "Divine word never uttered until the Day of Truth." And in that palace, foursquare is the tower of the Cube of the Law.

●●●THE CUBE OF THE LAW AND THE TOWER OF BABEL

11. o children of the day of truth, before the separation that divided you into the children of the dawn and children of the book, there was only the Cube of the Law and the Law of the Cube.

12. Everyone and everything moved and lived within this Law: one mind, one spirit, one will.

13. Within the dimensions and measurements of the cube are all knowledge of Heaven and Earth. And in order to know, expand on, and delight in the sublime Law of the Cube, God created you, 0 children of the day of truth.

14. And though shaped of no more than mud and sounding clay, in you is the perfection of knowing the Cube of the Law, if only you remembered!

15. Children of the Day of Truth, just as the cube is the measure of the perfection of God's thought, so you are the measure of God's unceasing movement, which is called time.

16. In you the cube is the crown of perfect knowing of God's will in time. But true to his word, Iblis made wrong fair-seeming. To show himself man's enemy, Iblis split the cube. He drove a forgetting of time within the mind, separating mind from spirit, which is God's movement in all things.

17. And from this forgetting in man, mind and spirit became confused. The will weakened. Ego was born. Iblis's many-headed self triumphed. False authority arose. Babylon came to be and in its center the Tower of Babel, the monumental curse for no longer knowing time.

18. In the forgetting was the separation of the people of the dawn from the people of the book. Had there been no Tower of Babel, there would have been no book. The Tower of Babel was a deceit intended to show the Cube of the Law existing outside of the perfection of knowledge already formed within you, a children of the day of truth.

19. Product of the forgetting, the Tower of Babel has a shadow of ever-darkening and all-encompassing dimensions that spreads 5,000 years from Babylonia to all of Earth's most remote wildernesses and mountain retreats. Iblis's curse of Ego and false time are complete but hold only to the day of truth.

20. For God is gracious and most merciful. As long as He could, He kept the people of the dawn away from the Tower of Babel. And to both the people of the dawn and the people of the book, He sent many messengers, apostles of light, witnesses of truth, doers of the righteous way, which is always framed by the invincible Law of the Cube.

●●●● TELEKTONON: THE FRAME OF TIME

21. As the special witness of time, I, Pacal Votan, know the perfect count of days. I bow in the temple of the tower and the rock, the sanctuary of Bolon Ik. In my body, formed of the ultimate perfection of God's power of all movement and measure (Hunab Ku), is the recollection that is prophecy. Knowing this body as the measure of time, I had built this palace and this tower to correct and to also look for the Tower of Babel, a memory of the remembering. And in my ninefold temple, where, guarded by the nine powers of time, I had this body laid, I left this legacy for you, the Talking Stone of Prophecy.

22. Telektonon, the frame of time I left for you, is a mirror to show you your own truth of time for one last time. For now is the day of truth, the hour of judgment.

23. You, children of the dawn, and you, people of the book, unify yourselves the only way you can: in time, through time, as time.

24. Telektonon, the perfection of time, is the only way for you to escape the fire that consumes the unrighteous. If you who know, who have followed the straight way, the good path, now falter in this challenge of unification in time, Telektonon, which has already been prepared for you, you will perish forever in the fire that now consumes all unbelievers. God's command is in you. Do not listen to the evil one now.

MY COUNT OF DAYS IS PERFECT

25. Listen! My knowledge is this: My count of days is perfect, my knowledge of the Cube of the Law is unsurpassed. Twelve is the number of the temple, but thirteen is the number of God. Twelve times twelve, 144, is the number of the perfection of the temple as light. And twelve times 144, 1728, is the number of the Cube of the Law, Telektonon, the divine word unuttered until now.

26. "... To him that overcometh will I give to eat of the hidden manna, and will give him a white stone, and in that stone, a new name written, which no man knoweth, saving he that receiveth it." (Revelations 2:17) And that name is Telektonon, and that stone is crystal, and that hidden manna is knowledge of the revelation of time.

27. The number of the elect is 144,000. This was the number set aside by God to be free of Iblis's curse. Dispersed is this number among mankind, and always was it meant to multiply both among the people of the dawn and the people of the book.

28. But since a human kin is the measure of the perfection of time, 144,000 is also the number of days within a cycle of twenty generations.

29. Twelve times 20 generations is 1,728,000 days, the number of the Cube of the Law. But for God, one cycle more, 144,000 days more to entice Iblis to make his score, a total of 1,872,000 days, thirteen cycles of God's elect.

30. I am Pacal Votan, time's special witness, galactic agent of the Nine Great Lords of Destiny, by oath pledged to the honor of Bolon Ik.

31. In my body of time, thirteen star histories rehearse themselves for the day of truth. Each star history possesses its own knowledge keeper appointed by the Nine Great Lords of Galactic Destiny. Thirteen Wisdom Holders, Nine Great Lords, twenty-two in all, to each of whom I am bound by a luminous thread of knowing, sealed by oath to Bolon Ik, White Solar Wind, kin 22 in the Book of Destiny, the Book of Kin.

ALL IS NUMBER

32. All is number. God is a number. God is in all.

33. Child of the day of truth, listen: 20 is the number of totality. You have twenty fingers and toes because God has made you the totality of time. Thirteen is the number of God's cosmic wisdom, perfect in its power of unceasing change and circulation. Thirteen are the joints of your limbs and body. Seven is the difference between 13 and 20. Seven is the mystic perfection between one and thirteen. Seven is the erect spine of the vigilance in which are met the 144,000 mystic fibers of the body's thirst for prophecy.

34. 144,000 kin, 144,000 days, these are called a baktun. Thirteen baktuns, one perfect cycle to rehearse all thirteen star histories. Each baktun, twenty generations, each generation 7,200 kin. So it was that Bolon Ik ordained the sacred count of days, each day a sacred kin, each kin one of the elect, a turning in the twist of universal time, each kin ruled by the power of the nine, the power of the Nine Great Lords of Time.

35. So began the sacred thirteen baktun count, 3,113 years before the birth of the second of the special messengers, the Apostle Jesus Christ, who sent to you a Book of Dark Prophecies, twenty-two sacred chapters in all, through his appointed witness, John of Patmos.

36. All this I know and declare to be true, for I am time's special witness to the day of truth, Pacal Votan, rehearser of the thirteen star histories, he who adds the 144,000 elect to the cubing of the cube, assuring not twelve baktuns, but thirteen to be the number to bring on the day of truth, God's judgment of your power of remembrance!

37. Children of the people of the dawn, in separating you from the people of the book, God meant you to be the reminder that wisdom is already perfect in your bodily form of time.

LORD BUDDHA

38. In remembrance, many messengers did he send you, many fires did he light high atop the mountain for you to see. And one special messenger did he send, one special witness of the dawn.

39. In India, at the midpoint of the seventh baktun, the very center of the mystic cycle, thirteen baktuns in all, Lord Buddha did awaken at the dawn, the morning star, the tree, the Earth, his witnesses of illumination. For seven weeks, he sat beneath that tree. Never moving, he traveled the seven directions and visited with his mind every realm of God's vast creation, only to learn the turning of the Wheel of the Law.

40. Withstanding all efforts of the evil one, he taught unceasingly for nine times five years more, until the age of eighty. All mind teachings did he give; the unwritten Book of Enlightenment did he leave with the Earth. His last turning of the Wheel of the Law was the teaching of the Wheel of Time. Given at the request of King Suchandra of Shambhala who traveled far from Central Asia to receive this teaching, the teaching of the Wheel of Time closed the life of Lord Buddha.

41. "All composite things are impermanent," declared Lord Buddha at his death, "Be a lamp and a refuge unto yourself, look to no outer authority but yourself!" The fire of Lord Buddha's illumination lit many lamps. The Book of Enlightenment he left with the Earth is a book for all to know in silence with the Earth.

42. In Shambhala did King Suchandra transmit the teaching of the Wheel of Time, Kalachakra, as it is called. Invoking the Nine Great Lords of Galactic Destiny, Shambhala prospered in abundance and wisdom. When came the seventh ruler after King Suchandra, Queen Visvamati, the people of Shambhala ascended as one into the paradise of the fifth dimension, where their wisdom continues to water the rivers that flow beneath the gardens of the righteous.

43. Integrating the power of the five, which is the center, with the power of twelve, which is the root of the cube, the Kalachakra is also the cycle of sixty years (5 x 12). By the Law of the Cube, which is the power of sixteen, foursquare or 7 plus 9, sixteen cycles

of sixty years is the appointed time of prophecy, the prophecy of Kalachakra, which is the time of the day of judgment.

44. So it was that after the arrival in Tibet of Buddha's special voice, Padmasambhava, late in the tenth baktun, that the Lords of Shambhala determined to release the teachings of the Wheel of Time among the people of the dawn inhabiting Tibet. Thus it was that the Wheel of Prophecy turned the first of these sixteen cycles in A.D. 1027 and completed the sixteenth of these 60-year cycles in A.D.1987, in partial fulfillment of the Harmonic Convergence of all prophecy.

45. 0 people of the dawn, do not doubt. The appointed day is come.

THE HEPTAGONON OF MIND

46. People of the dawn, people of the book, children of the day of truth. To safeguard as long as possible the teachings of the Original Ones from the confusion of the Tower of Babel, God divided the Earth by its one ocean into great island masses called continents. And most far from Babylon was Amerrikua, and also distant was Australia and the Island Kingdoms of Polynesia. Even in Africa, close to Babylon, for many great cycles were the people of the dawn protected in their natural wisdom.

47. But inexorable is my count of days, inexorable the rolling of the Wheel of Time, inexorable the Law of the Cube.

48. Telektonon is the name of the blessing received by those who master the Law of the Cube. Great is the power of God, whose power and blessing are ever close and near, foursquare the manifest action of radiance from the divine intersection of God's presence: divine mind, divine spirit, divine will, divine source from which the Cube is sprung. Foursquare the Cube, the measure of Telektonon, one vast, far-spreading word that is no word at all, but number multiplying itself from within God's unending meditation, which we choose to call creation.

49. From one intersection, six directions plus the seventh, the moving center of time, the navel of Heaven whose cord is a fiber called Kuxan Suum. From the navel of Heaven to the solar plexus runs the Kuxan Suum, highway that reaches to the roots of the stars, passing through each of the heavenly dimensions.

50. I who say this, know, for I am Pacal Votan, who left for you a talking stone, Telektonon. Time's special witness am I, God's spy holding vigil, keeping the sacred count of days, exposing to you the lie of Babylon, the falsification of the Cube into the Tower of Babel!

51. Who now but I can tell you that seven are the directions of the Cube, the Heptaganon of Mind: twelve are the gates, sixteen the powers; moved by the nine great powers of time, these sixteen powers become the number of the elect of the Cube, 144. Multiplied by the millennium, 144 becomes 144,000, sacred number of days I laid for prophecy to know during my first baktun count.

IMPERIAL BABYLON

52. Children of the book, even before the close of my first baktun, twenty generations from the year 3113 B.C.,the foundations of the Tower of Babel were laid. Captive now were the 144,000 within the Earthly prison of Babylon's materialism.

53. Yet, by my devotion and power of meditation, from deep within the Earth, aided by Bolon Ik's perseverance on far Uranus, Earth's harmonic twin, I kept the sacred count of days, I watched over the movement of the sacred march of kin.

54. Four more baktuns passed, and Babylon's power was swollen with the loot of Empire. Despite the righteousness of Abraham and the Law of Moses, the shadow of the Tower of Babel only increased in size and darkness.

55. Came the sixth baktun and Imperial Babylon was complete. Its seal of false time and power, taxation, and armed expanse had fulfilled Iblis's desire to degrade and ruin all humanity by making wrong fair-seeming. Fair-seeming now to amass wealth and power; fair-seeming to make a count of time by twelve months; and fair-seeming to regard the thirteenth moon with superstition and contempt! Fair-seeming to gather taxes and to pay them; fair-seeming to make war for the power of King's; fair-seeming that birth and death must be paid for; fair-seeming that all power come from the priests, Iblis's most cherished of men.

56. 0 people of the book, lost to you was all memory of the Law of the Cube, slavery became your lot! But I am Pacal Votan, time's special witness. Even before the coming of the second special messenger, Christ the Apostle of Love, by my free will and my knowledge of the power of seven, I laid claim to the Seven Seals of Prophecy that on the day of truth, Iblis's curse would be finished and the righteous would triumph. And now the day of truth is come. Telektonon, the talking stone, now speaks in a voice for all to hear.

57. By the seventh baktun, the Seven Seals of Prophecy were stored in Earth's special lodestone of truth, whose custodian am 1.

58. By the eighth Baktun, Iblis had caused the Tower of Babel to speak more tongues in a growing clash of empires. Into this confusion came the Apostle of Love, crucified for driving the moneylenders from the Temple, misunderstood by the people of the book, finally to be absorbed by the Priests of Babylon, now Roman by tongue, who use this Christ as their emblem to spawn Iblis's greatest evil, the total usurpation of Earthly Time.

THE BOOK OF DARK PROPHECIES

59. One special voice did this Jesus have, a man named Saint John ofPatmos, in whom the gift of prophecy was fulfilled. From my ark of universal time, Earth's special lodestone of truth, rock of Heaven, attained by thirteen tones, I guided Saint John's hand by the mystic power of seven. And so was sent to the world the Book of Dark Proph-

ecies, Revelations, the Apocalypse, twenty-two mystic chapters, the first thirteen mirroring my thirteen great baktuns, nine more for the Lords of Time, seven of these for each of Seven Years of Prophecy, and the last two to complete the visionary power of the righteous of the Cube: 144,000 times thirteen.

60. To those who can read they will find Iblis's numbers 1,260 and 666, well known on the day of truth, for 12:60 is the ratio of the false time that makes martyrs of the righteous, and 666 the number of the beast who captures my thirteenth baktun.

61. Children of the day of truth, children of the dawn, people of the book, all this can you hear in the stones of Palenque where in the tenth baktun I took pure human form. Of the twenty-seven books of the New Testament, the twenty-seventh is Revelations. Twenty-seven steps ascend to the Palace of Bolon Ik. This palace, New Jerusalem, is the mystic power of Telektonon, the twenty-eighth step, the power of seven sanctified foursquare.

62. Twenty-two chapters of Revelations are the twenty-two kin complete in Bolon Ik: thirteen for the power of time, nine for the Nine Lords of Time, whose forms still guard my ark of stone deep within the ninefold temple of the Cube of the Law. Perfect was the time of my coming, kin 1, 366, 560. Perfect was the time of my going, kin 1, 385, 540.

63. Installed in the House of the Serpent, my coming was commemorated by the departure of the third messenger, the special messenger of the day of truth, the Apostle Muhammad, who left for the righteous of the Earth the Book of the Righteous, the Holy Quran, clear signs to understand.

64. As Buddha taught so pure one mind, so Muhammad taught so pure one God. Submission to the will of God is at the center of the Law of the Cube. Truly for the righteous there is only the straight way, the path of basic goodness, laid out in the heart as obedience to God's will, the divine plan unfathomable to those who cry, "reason alone can save us!"

THE PROPHECY OF THE SEVEN GENERATIONS

65. I am Pacal Votan, witness of time. Telektonon is my gospel. The talking stone of unifying prophecy is my legacy intended for that special messenger who understands the meaning of my numbers.

66. By my symbolic channel of communication called Telektonon did I intend this prophecy to be discovered and recovered. Once forty years had elapsed from the time of the discovery of my Uranian crypt, then would come the recovery, the recollection by my special agent of the numbers rehearsed and repeated to announce the day of truth.

67. Perfect was the time of my coming. Perfect was my mission. Perfect was the time of my going. Perfect is the discovery of my prophecy. Perfect is the Telektonon, the little book the angel of the Lord made Saint John of Patmos taste and eat, which you now possess, a treasure for the day of truth.

68. What you witness by these words is the last message I gave to my people, the message of 9.13.0.0.0, heard from my ark of stone through the Telektonon, Earth Spirit Speaking Tube. To my people I declared:

69. "Seven katuns are there to follow me. Seven generations, each one guarded by a solar witness taking account of what you do, each solar witness sealing your time with a prophecy to be known as the Book of the Seven Generations. And this book is to be opened for the day of truth.

70. "After the seven generations, the end of the baktun. Ten baktuns gone, the glorious Age of Maya gone, the Long Count of my days to be completed at 1,440,000 kin, the mystic number of the elect multiplied by ten.

71. "Then will come the three baktuns of Babel, the coming exile of 12:60 time. Thirteen years after this tenth baktun shall come the last prophetic round: thirteen cycles of Heaven, each fifty-two years, followed by nine cycles of Hell, each also fifty-two years. When the third round of Heaven begins, the Messenger of the Last Call will come to you, Quetzalcoatl-Kukulkan, the prophet by name.

72. "And when the thirteenth Heaven cycle is over, then will come the complete destruction of all we are and know. Before the twelfth baktun is over, two priests bearing false witness, ornamented with the sacred numbers five and thirteen, will see to it that the thirteenth baktun will be the end of time.

73. "Unrelenting will be the roll of the Nine Hells into the abyss of fire and madness. In the time of the last Hell, though you will be declared free, free you will not be. Ten Kings in succession will mark the climax of time. Seven years will follow the ending of the last Hell. Then will sound the angel of the seventh trumpet, the seven thunders will roll, the beast of seven heads called G-7 will place his mark on your foreheads and hands, the false one from Babylon will return to be embraced by the 11 Serpent. Then the day of truth will come, the Seven Years of Prophecy."

TIME'S SPECIAL WITNESS

74. People of the dawn, people of the book, children of the day of truth, many messengers have been sent to you, many prophets have been given to you. A Book of Enlightenment has been placed in the Earth for all to read in communion with the Earth. A Book of Dark Prophecies has been left for you. A prophecy of thirteen Heavens and nine Hells to accompany the Book of Dark Prophecies has also been left for you. A great prophet and messenger has left you clear signs to understand and a book of unassailable righteousness, the Holy Quran.

75. As the special witness of time, I have left for you my talking stone, Telektonon, breath of precious Bolon Ik, and have sent you both a prophet, Quetzalcoatl of sacred Xochicalco, and for the day of truth I have appointed also a special messenger who can hear my talking stone and write and explain its numbers and its meaning for you, children of the righteous.

76. Do not doubt any of this. When the moment comes for him who is to be the eleventh in succession, to succeed the one, 11 Serpent, then join the believers and begin the prophecy that liberates from all Kings and successors.

77. Ten witnesses I place around my tomb, ten Uranian emissaries, one each for each of ten baktuns. Of these am I the eleventh. Who would be the eleventh in the final succession of the Kings of Mexico must own me as the eleventh, then will all go well.

78. Six messengers did I place on my talking stone of prophecy: three above, the messengers Lord Buddha, Muhammad, and Christ; and three below, Padmasambhava, Quetzalcoatl, and Saint John of Patmos. Of these am I the seventh.

79. Pacal Votan am I, time's special witness, messenger of the Telektonon, instrument of the seventh angel, I declare again and again: All is number. God is a number. God is in all.

WAR OF THE RIGHTEOUS

80. Look carefully at my stone and listen: ten messengers, 24 signs, for ten are the number of orbits of planets around this star, Kinich Ahau, your Sun. From my point of origin, your star is designated 24, number of the circuit of externalizing intelligence. If you are of the righteous, then in this is a sign for you of your star mission.

81. And from each of the three oracle mouths of the arms of the cross of my Kuxan Suum, you will find twenty-four rays repeated three times, one-half the number of the elect (3 x 24 = 72). Add to this 3 times 11 (33) and the number is 105, the difference between the number of days in your solar orbit, 365, and the number of kin in my sacred count, my galactic spin, 260.

82. 13:20 is the ratio of natural sacred time. 12:60 is the ratio of the false time of the Tower of Babel. 13:20, thirteen moons, twenty fingers and toes. 12:60, twelve-month year, sixty-minute hour. Children of the day of truth, in this is a sign for you to discriminate between the power of the evil one who seeks to trap you in his machine, and the power of the righteous who need nothing but that with which God has endowed them: their body of time and their natural path synchronized by thirteen moons.

83. Children of the day of truth this is your only decision, for God knows and sees all that is in your hearts. People of the dawn, people of the book, on the hour of judgment unify yourselves in the only way you can: in time, as time, through time. For this in the year of kin 144, Telektonon is revealed to you, God's plan for peace on Earth, the last and only hope for spiritual renewal and salvation, immediate acceptance and adoption of the twenty-eight-day, thirteen-moon way, the calendar Telektonon. "And he had a name written that no man knew but himself." (Revelations 19:12) And that name is Telektonon.

84. Children of the day of truth, people of the dawn, people of the book, foursquare is my talking stone, invisible its design, its crystal lacing of star histories met in my human form. Time's special witness am I seeking apostles of time to organize the day of truth

into the revelation of God's redemption. Universal forgiveness awaits those who understand God's plan, a common time of twenty-eight days, thirteen moons, begun together at the appointed hour, on the day called Lunar Moon, the first day of the second year of prophecy.

85. Two solar faces there are upon my talking stone, the one an icon of my solar shield, Galactic Sun, the other, Lunar Sun, an emblem of the perfect lunar timing of Earth's solar orbit.

86. Children of the day of truth, the hour of judgment has come for you. How much longer do you need to suffocate in the Babylonian exile of 12:60 time? My children are the oppressed. Is it not time for the war of the righteous to restore equality? Why not wage the war then that destroys altogether the old time? Why not smash the caskets of the Babylonian wine of fornication in a single stroke?

HOLY VICTORY MARCH

87. 0 children of the day of truth, people of the dawn, people of the book, in your origins you were one blood, today the hour of judgment demands you become one again, unified in time. Who among you will join my apostles of the Telektonon? Already the gift has been given. Already we left for you the 13:20 tool, my sacred count of days, 260 kin, to turn the solar-lunar wheel of365 days. And when prophecy was vindicated at the conclusion of the ninth Hell, we sent among you the perfect form of this gift, Dreamspell, containing the message of the thirteen moons.

88. But this was not enough, the seventh trumpet had yet to sound, and so it did in the fortieth year following the discovery of my Uranian ark of stone. And now my prophecies are revealed. The Mayan Apocalypse of kin 144, Yellow Magnetic Seed, turns the seventh year of karmic destiny into the opening of the first of my seven seals of prophecy.

89. In this seal foursquare, all prophecies converge. The Beast of the G-7 stands revealed. The Babylonian Vatican is exposed. In the rainbow dream vision 144,000 of the elect are called again to meet, gathering together in circles to listen, to sing, and to dance to the song Telektonon.

90. 0 children of the day of truth, people of the dawn, people of the book, seven is my sacred power. Foursquare seven is 28, the number that, perfectly followed, opens Heaven's gate. Walking together the sacred thirteen moon, 28-day way, you will be joined to God's command, a holy victory march of the righteous to restore sacred order to this troubled Earth and to recover your lost powers.

91. Four powers of seven are given to you to be your comfort in the Victory March, one power for each of four perfect weeks, every twenty-eight days. By the first power of seven, claim your power of prophecy, by the second power of seven, maintain your power of prophecy; by the third power of seven, proclaim the victory of prophecy; by the fourth power of seven, become the victory of prophecy!

92. 0 children of the day of truth, 144,000 times thirteen is an army of the righteous. Learn again the count of sacred time and join this army. 0 children of the day of truth, realize that everything about the old 12:60 time is now become a garment of false being, a moral corruption, a banquet feast for the evil one.

THE STORM OF NON-EGO

93. 0 children of the day of truth, children of the people of the book, in the Storm is the sign of God's power. To the Storm is assigned the power of nineteen, the mystic saturation of all number. One hundred fourteen Suras, nineteen times six, the power of the Cube, are the number of Suras in the Holy Quran. Twenty-two years after the finding of my Uranian crypt, Telektonon, came complete revelation of the Quran's power of nineteen, Allah's mercy. Nineteen more years then to unfold the karma of Allah's prophet besieged by Babylonian conspiracy. Then comes the destined year nineteen, kin 144, hour of prophecy's convergence, moment of the day of truth.

94. Let the 124,000 messengers call forth a remnant of 20,000 more; let each of these sound seven trumpets, seven times. Let these seven times seven trumpets call for a following of God's perfect plan. Mark seven times seven years since the evil one showed his fiery face at Hiroshima. Be honored by God's cancellation of Iblis's power of moral oblivion. Let the storm of God's power of nineteen cleanse with cosmic mercy the blood of all martyrs from the thrones of Kings, and let these thrones themselves be dissolved by the pure water of sacred time flowing through the gardens of the righteous.

95. I am Pacal Votan, time's special witness, revealer of the Cube of the Law, devoted patron of the mystery that is woman, the consummation of the telepathic power of twenty-eight, known to me by name of Bolon Ik.

96. Pacal Votan am I, enigma to your men of reason. My existence is proof of the workings of higher law. Telepathy is participation in God's knowing, but is no cause to claim full knowledge of His powers. Prophecy too is natural law, the complement to God's mercy.

97. 0-19 is my code. In it are all demonstrations of the wisdom that is always close and near, but that comprehends the web of galactic being and all forms of its natural orders in time. Just as nineteen is the power of God's mercy, so zero is the fullness and emptiness all at once that characterize the nature of mind and all it can know.

98. Lord Buddha called this power of zero Sunyata. The special voice of Buddha, Padmasambhava, whose life span followed mine, exactly as Muhammad's preceded it, knew this Sunyata to be the basis of non-ego.

99. If you would know me and all that I witness for you in this hour of judgment, children of the dawn, people of the book, then practice the being that is non-ego.

THE ECSTASY OF DELIVERANCE

100. Because I repose on Telektonon, the mother of all prophecy, mine is the ecstasy of deliverance. With a mind of non-ego clasped to the heart of God I have sent out my

message of the Mayan Apocalypse. By relinquishing the calendar of twelve and leaving the prison of mechanized time, God's mercy will come swiftly to you.

101. Thirteen moons, twenty fingers and toes, thirteen joints, four limbs, five senses do you already possess. God's knowing and power of movement are already in you like a mother pregnant with child.

102. Why tarry now? When a house is burning, scurrying for your possessions will only cost you your life. This house that is burning is Babylon, the 12:60 way of life, which is now destroying your biosphere and poisoning everything about your life. Do not linger now fondling your machines. This is the Day of Judgment. If you can remember, leave this burning house and know that a better way has already been prepared for you, a holy victory march, my seven years of prophecy, Telektonon, the sacred revelation of time.

103. The Cube of the Law is awaiting its construction through the collective mind, spirit, and will of you, 0 children of the day of truth. Iblis, Satan, the evil one is none other than your ego clinging to the fatal lures of Babylon. God knows nonetheless exactly what is in your heart. Why not, then, become an apostle of the holy victory march? The ecstasy of deliverance is meant to be shared by all.

THE PROPHETS OF PEACE ARE AWAKENING

104. Pacal Votan am I, time's special witness am I, who bequeaths to you a talking stone, a legacy of place, Nah Chan, House of the Serpent, by prophecy known as Palenque.

105. And at Xochicalco, too, you will find stone utterances of my gospel left by command of my apostle, Topiltzin Ce Acatl Quetzalcoatl, who closed Christ's first millennium, as now my prophecy is come to close Christ's second millennium.

106. 0 people of the dawn, people of the book, children of the day of truth, make yourselves as one again and do not doubt the accuracy of the appointed hour on the day Lunar Moon! Synchronize now, and in all humility accept God's plan, twenty-eight days, thirteen moons, plus one day for deliverance and forgiveness.

107. Pardon now in public your enemy's indiscretions, and make full confession of your own transgressions. If you are sincere, God will see this in your heart and by His mercy many will follow your way; and if you proclaim for the cause of righteousness and truth, a new beginning and a new time, Telektonon, the thirteen moon way, and you enjoin others to do likewise, then God will know this, too, and make you a great general in the holy victory march.

108. And, if you have means and wealth, and you hear these words, or have explained to you the curse of 12:60 time and the blessing of 13:20 time, then do not doubt it and practice charity. Provide for the children orphaned by war and provide for the reconstruction of the Earth, that the practice of art and culture may flourish once again. Practice such charity with zeal, for soon you shall see an end to money. No more will flow the drunken wine of Babylon's fornication, and then what will you be with your paper credits and your useless machines?

109. The prophets of peace are awakening. Truly, the hour of judgment is come. Among the angels am I known as the mysterious benefactor of the righteous. Will you throwaway this opportunity, too? Beware for now the fire that consumes all falsehood is raging.

𝋲 THE WIZARD'S HOARD OF THE LORE OF DEATH

110. People of the dawn, people of the book, children of the day of truth, because I am time's special witness, everything about my talking stone of prophecy is in fulfillment of God's divine plan.

111. Mine is the teaching of the 12:60 way of exile and the 13:20 way of redemption. Exactly 1,260 years passed from the dedication of my temple and closing of my Uranian crypt, A.D.692, 9.13.0.0.0, to its discovery and reopening in A.D.1952, 12.17.0.0.0. Another five times seven years, the numbers of my sacred ratio passed before the closing of the ninth Hell, A.D.1987, commemorated by the sacrifice of 3 Monkey, the first of four sacred signs on the western edge of my great talking stone. In this sacrifice was a sign of the death and transcendence of the way of the people of the book, "Joshua," and the people of the dawn, "Maitreya."

112. Eleven sets of triple disks, thirty-three in all, will you find upon my stone, for 33 is the number of the initiate, the middle harmonic of 65 that constitutes the Tzolkin, my sacred count of thirteen times twenty. All clear signs have I left for you, two borders of stars, the host of the Nine Lords of Time repeated by number on each side, top and bottom of my stone, a ledger of star histories complete in my form.

113. But especially would I have you look at Mars and consider again the disposition of the thirteen sacred signs placed around the edge of my stone, from 8 Ahau, my seal of truth, to 13 Cimi, the galactic closing of the wizard's hoard of the lore of death.

114. Ten emissaries celebrate the glory of my gifts for you, signs of peace from other worlds that beckon you to awaken from your sleep of Earthly Hell to join yourselves to God's command, Telektonon, the 28-day, thirteen-moon way and find in that count of days a path that leads directly to the paradise that is the repose and reward of the righteous.

115. One hundred forty is the number of the Telektonon, perfect its ratio 5:7 (28 x 5 = 20 x 7), the same as my sacred ratio 260:364(+1)

116. All is number. God is a number. God is in all.

𝋳 SEVEN PERFECT ORACLES

117. People of the dawn, people of the book, children of the day of truth, this completes my testimony, I, Valum Votan, holder of the solar shield, protected have I been by the perfect love of Bolon Ik, who has stored my prophecy until this moment of release.

118. Seven perfect oracles have I given you, a perfect calendar of thirteen moons, an instrument to renew your spirit and your powers of Prophecy, Telektonon, and a book

of knowledge for you to penetrate the forgotten history of your star, Dreamspell. Also have I sent two messengers who bear witness to my truth, and thirteen perfect signs.

119. Now is the time of my prophecy. Now is the day of truth. Now is the opening of the seven seals. Now is the sounding of the seven trumpets. Now is the hour of judgment. Now is the song of the 144,000, the seed army of the righteous.

120. When the Lunar Moon has overflowed its banks, the G-7 will be no more.

121. When the Electric Wizard arrives, gather in banners of righteousness.

122. When arrives the Self-existing Storm, you will be fortified with spiritual abundance.

123. When comes the Overtone Seed, neither money nor priests will besiege you any longer, but the Earth will sing in harmony with your righteousness, Babylon will be no more, the Law of the Cube will be fully rehearsed in the hearts of the righteous.

124. In the sign of the Rhythmic Moon, senses will open to the manna of Heaven; no where will you find any longer signs of the evil one; a cosmic race will you be once again.

125. And when comes the seventh year, the Resonant Wizard will seal the Heavens and the Earth with the glory of all prophecy, the redemption of all prophets, messengers, apostles, and witnesses of truth.

126. Five years then will you have to close the cycle ofthe Seed, and another eight, thirteen years in all, to prepare for the coming of the Galactic Seed, 2013 Arcturus Dominion, the fulfillment of the star cycle of Valum Chivim, witnessed by the special agent, Pacal Votan, to the glory oflove, Bolon Ik, and on behalf of the mercy of God's divine plan, Telektonon, Cube of the Law, the mighty talking stone of prophecy.

"Man is one. Culture is indivisible."
-ALBERTO Ruz, 1952, 12.17.0.0.0.

Transcribed 12 Seed, kin 64,
Resonant Moon Day 13,
Warrior's Cube 7, 0 Yax
Yellow Magnetic Seed Year, kin 144,
First Year of Prophecy

Blue Spectral Monkey/White Solar Wind
Tepoztlan, Morelos, Mexico

▼

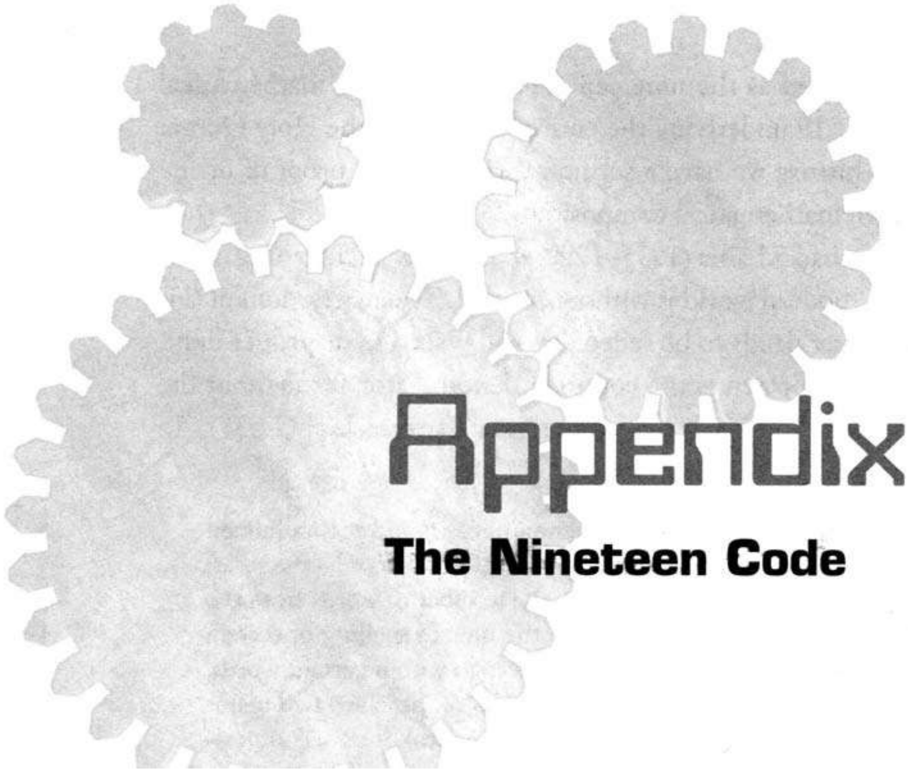

Appendix IV

The Nineteen Code

THE MYSTERY of the motives of the hijackers, who were instrumented by divine will, if not by biospheric necessity, to catalyze the Inevitable Event, lies in their very number: nineteen. While the nineteen is related to the Quranic verse 74:30, "Over it is nineteen," and can be interpreted prophetically, the Law of Time confirms the Quran in demonstrating that the number nineteen possesses a supreme and indeed inviolable value. The Law of Time validates all that is true, and contains within it the mathematical confirmation of all systems of thought as different facets of the same synchronic order.

The Law of Time is based on the mathematical code of the Harmonic Module, itself a function and a demonstration of the 13:20 ratio. The vigesimal mathematics of the 13:20 frequency is known as the 0':'"19code, which is written with a simple dot-bar notation. As such the 0-19 code is a whole number system in which the count is not 1-9 but 1-19, and the positional zero represents a factor of 20 rather than of 10. As the demonstration of a radial mathematical code, the number 19 plays a singular role in the 13:20 matrix. As the largest compound prime number of the 0-19 count, within the context of this count 19 has certain unique properties and values. It may be asked: why 19?

First of all, let us consider the theological dimensions of the number 19 as hinted at by the Quranic verse 74:30. Long an enigma to scholars of the Quran, it was only through the genius of an Egyptian scientist, aided by a computer, that the meaning of the verse was finally revealed as the nineteen code-an intricate mathematical system based on the number 19 underlying the entire structure of the Holy Quran.

"For the first time in history we have a scripture with built-in proof of divine authorship-a superhuman mathematical composition."1

The contribution of Rashad Khalifa (1935-1990) to the role and meaning of the Holy Quran for the entire modem world is without parallel. The martyrdom of this genuine seeker and messenger needs to be redeemed and placed in its proper light. According to Dr. Khalifa, "Nineteen is the common denominator throughout the Quran's mathematical system." "This is one of the greatest miracles" (74:35). To summarize in his own words, Dr. Khalifa states:

> The Quran is characterized by a unique phenomenon never found in any human authored book. Every element of the Quran is mathematically composed-the suras, the verses, the words, the number of certain letters, the number of words from the same root, the number and variety of divine names, the unique spelling of certain words, the absence or deliberate alteration of certain letters within certain words, and many other elements of the Quran besides its content. There are two major facets of the Quran's mathematical system: (1) The mathematical literary composition, and (2) The mathematical structure involving the numbers of suras and verses. Because of this comprehensive mathematical coding, the slightest distortion of the Quran's text or physical arrangement is immediately exposed.2

Like the Law of Time, which could only have been discovered after the opening of Pacal Votan's tomb, in A.D. 1952, the nineteen code of the Quran could only be discovered and unveiled over a twenty-two-year period between 1968 and 1990. Why? Because only with a modem computer could Dr. Khalifa subject every last word and letter of the Quran, in its Arabic original, to the rigorous analysis necessary to expose its mathematical perfection. As Dr. Khalifa writes in his footnote to 10:20 ("They say, 'How come no miracle came down to him from his Lord?' Say, 'The future belongs to God; so wait, and I am waiting along with you.'"):

> 10.20: In retrospect we see now that the Quran's miracle, indeed "One of the greatest miracles" (74:35), was divinely predestined to be unveiled 14 centuries after Muhammad. In view of the current condition of the traditional Muslims, if Muhammad had been given this miracle, those Muslims, who are already idolizing Muhammad beside God, would have worshipped him as God incarnate. Additionally, this miracle is obviously designed for the computer age, and to be appreciated by mathematically sophisticated generations.3

The phenomenon of the intricate patterning of the nineteen code as witnessed, for example, in the 19 x 142 (2,698) occurrences of the word *God* in the Quran, writes Khalifa, "alone suffices as incontrovertible proof that the Quran is God's message to the world. No human being(s) could have kept track of 2698 occurrences of the word 'God,' and the number of verses where they occur."[4] One must bear in mind that the Quran was revealed in a time of primitive simplicity over a twenty-three-year period so that the suras and verses were separated in time and place, while the chronological order of the revelation of the suras is not at all the same as their final format. Moreover, the miracle of the nineteen code is not limited to the appearance of the word God, but is vast, intricate, and comprehensive. Dr. Khalifa lists a set of nineteen "simple facts" that confirm the Quran's mathematical coding. The "simple facts" are those not requiring a calculator or computer to be ascertained. We hereby reproduce these "simple facts," as they are presented in Appendix 1 of *Quran: The Pinal Testament* (all references to letters and words are to the Arabic original):

1. The first verse (1:1), known as *"Basmalah,"* consists of 19 letters.
2. The Quran consists of 114 suras, which is 19 x 6.
3. The total number of verses in the Quran is 6346, or 19 x 334. (6234 numbered verses and 112 unnumbered verses *[Basmalahs].* 6234 +112 = 6346. Note that 6 + 3 + 4 + 6 = 19.
4. The *Basmalah* occurs 114 times, despite its conspicuous absence from Sura 9 (it occurs twice in Sura 27). 114 = 19 x 6.
5. From the missing *Basmalah* of Sura 9 to the extra *Basmalah* of Sura 27 there are precisely 19 suras.
6. It follows that the total of the sura numbers from 9 to 27 (9 +10 + 11 + 12 . . . + 26 + 27) is 342, or 19 x 18.
7. The total (342) also equals the number of words between the two *Basmalahs* of Sura 27, and 342 = 19 x 18.
8. The famous first revelation (96:1-5) consists of 19 words.
9. This 19-worded first revelation consists of 76 letters, 19 x 4.
10. Sura 96, first in the chronological sequence, consists of 19 verses.
11. This first chronological sura is placed atop the last 19 suras.
12. Sura 96 consists of 304 Arabic letters. 304 = 19 x 16.
13. The last revelation (Sura 110) consists of 19 words.
14. The first verse of the last revelation (110:1) consists of 19 letters.

15. 14 different Arabic letters form 14 different sets of "Quranic Initials" (such as A.L.M. of 2:1) and prefix 29 suras. These numbers add up to 14 + 14 + 29 =57and 57= 19 x 3.

16. The total of the 29 sura numbers where the Quranic Initials occur is 2 + 3+ 7 + . . . + 50 + 68 = 822, and 822 + 14 (14 sets ofInitials) = 836, or 19 x 44.

17. Between the first initialed sura (Sura 2) and the last initialed sura (Sura 68) there are 38 un-initialed suras. 38= 19 x 2.

18. Between the first and last initialed sura there are 19 sets of alternating "initialed" and "un-initialed" suras.

19. The Quran mentions 30 different numbers: 1,2,3,4, 5,6, 7, 8,9, 10, 11, 12, 19,20,30,40,50,60, 70, 80, 99,100,200,300,1000,2000,3000,5000,50,000, & 100,000. The sum of these numbers is 162,146= 19 x 6534.

Apart from the "simple facts," through use of his computer, a Hewlett Packard HP-IOOO E-series, Dr. Khalifa demonstrated many more complex levels of the coding of the number 19 throughout the Quran. These findings are synthesized in the Appendices to *Quran: The Pinal Testament*, and are also available in his other texts detailing his research, such as *The Miracle of the Quran.* The net effect of the truly monumental research effort exerted by Dr. Khalifa was the promotion of a highly reformed Islam based on "the Quran, the whole Quran, and nothing but the Quran." It is for this reason that Dr. Khalifa's translations of the Quran are referred to as "the Final Scripture," (1980) and the revised translation (1992) as the "Final Testament." It was also Dr. Khalifa's insistence on the Quran as the sole basis ofIslam that earned him his martyrdom in Tucson, Arizona, on January 31, 1990.

NINETEEN IN THE TWENTY-COUNT POSITIONAL SYSTEM

The Law ofTime-T(E) =Art, where T equals the 13:20 timing frequency ratio-upholds the validity of the research of Dr. Khalifa and graphically demonstrates *why* 19 is the mathematical factor that underlies the Holy Quran.

The fact that the Mayan vigesimal system is known as the 0-19 code, and that in this code 19, and not 9, has the supreme value of being the highest number, absolutely enhances the power and significance of Dr. Khalifa's discovery concerning

74:30, "Over it is nineteen." The supremacy of the value of 19, imperceptible if we only follow the decimal 1O-count, is fully demonstrated when we examine the 0-19 number code in which the numbers 0-19 are reproduced in a 5 x 4 (= 20) matrix, beginning with 0 and ending with 19. In this number set, which also demonstrates the mathematics of the radial matrix, there are ten pairs of numbers radially arranged, each pair of which adds up to 19:

0+19=19
1+18=19
2+17=19
3 + 16 = 19
4+15=19
5+14=19
6+13=19
7+12=19
8+11=19
9+10=19

In other words, within the parameters of the mathematical matrix in which nineteen is the integer with the highest value, its value completely informs the radial order of the entire vigesimal system. "Over it is nineteen." This is truly astonishing confirmation of the value accorded to 19 in the Quran. If the Quran is the true word of God, then we must expect that even knowledge unknown to Muhammad or Dr. Khalifa would be contained or coded in the Quran. We are referring, of course, to the Law of Time in its mathematical particulars, beginning with its mathematical base of expression, the 0-19 code.

19 = 7: NINETEEN IN THE 0-19 DOT-BAR NOTATIONAL SYSTEM

Just as the number 19 takes on a supreme value in the vigesimal system inconceivable in the decimal system, so the vigesimal order commands its own notational system. The 0-19 dot-bar notational system substantiates the meaning and power of 19 as the prime mathematical, saturated value unit in the twenty count. The dot-bar notational order is a holographic system distinct from the alphabetic system of Arabic numerals by which we write 1-10.

The distinction between use of the two notational orders is profound. This notational order uses only three notational symbols to build up the holographic context of

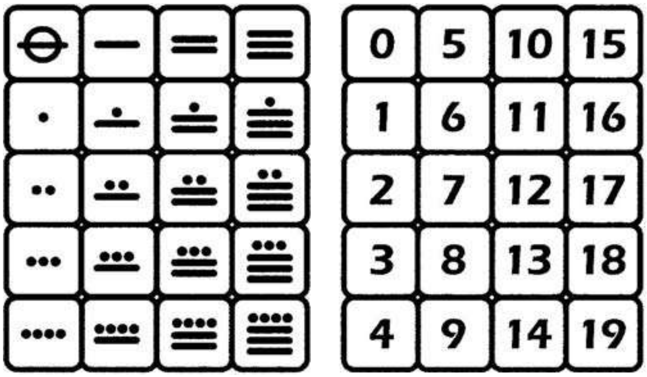

Any radially opposite set of numbers always equals 19, e.g., 4 and 15, 0 and 19, 9 and 10, 2 and 17, 11 and 8.

Nineteen is the highest value and the most saturated notation, being the only one with seven notational symbols—three bars and four dots.

Nineteen in galactic notation.
3 bars (3 x 5 = 15) plus four dots (4) = 19

Dot-bar notation system, 0–19 code

the fourth-dimensional number. One dot represents one unit; one bar represents a five-unit order; and a positional zero completes the system. The vigesimal notational system reveals the entire radial sequence of the 0-19 code to be a highly streamlined and aesthetic pattern. Brief contemplation shows that there are two patterned orders consisting of four vertical sets of five units each, and five horizontal sets of four units each. It is seen that nineteen represents the maximum saturation of both the vertical and horizontal orders, and consists of the maximum number of dots (four) and the maximum number of bars (three). It is also visibly manifest that 19 = 7 notational units, the maximum number of units.

The fact that 19 is represented by seven notational units confirms Sura 1, which has seven verses and the first line of which consists of nineteen Arabic letters. However, from the point of view of the Law of Time, 7 is the key implicit mathematical unit in the absolute timing frequency ratio, 13:20-the difference between 20 and 13 is 7, and the midpoint between 1 and 13 is also 7. While Dr. Khalifa is responsible for determining the dominance of nineteen as the mathematical key underlying the patterning and meaning of the Quran, the Law of Time provides an even more profound lens through which to consider Dr. Khalifa's work, by placing nineteen in a superior (vigesimal) mathematical and theological context.

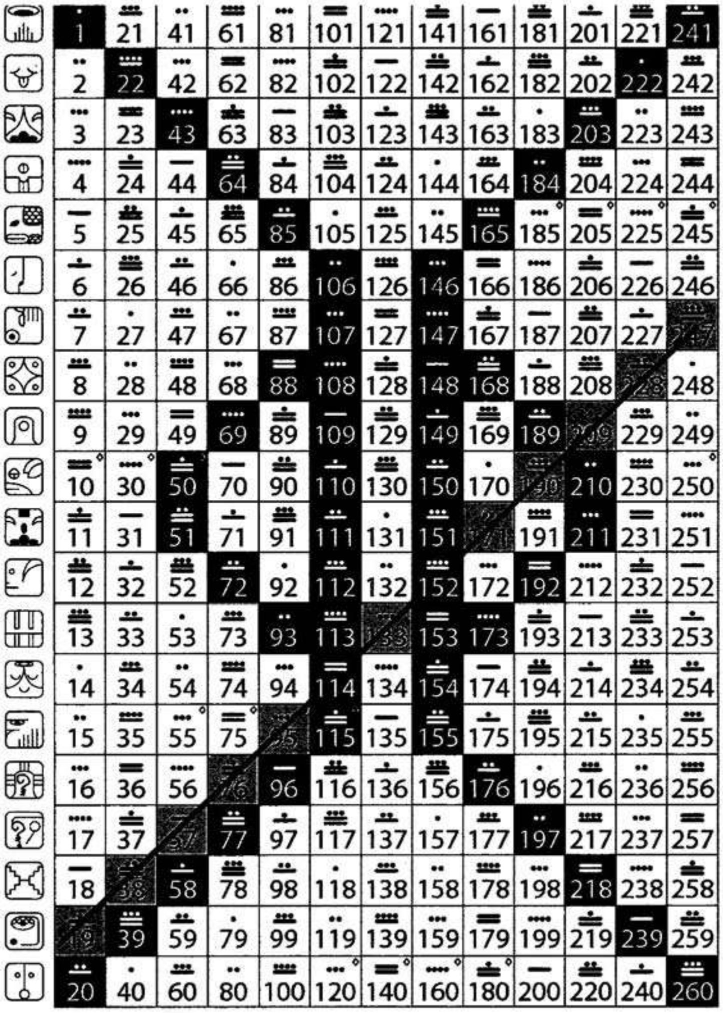

1	21	41	61	81	101	121	141	161	181	201	221	241
2	22	42	62	82	102	122	142	162	182	202	222	242
3	23	43	63	83	103	123	143	163	183	203	223	243
4	24	44	64	84	104	124	144	164	184	204	224	244
5	25	45	65	85	105	125	145	165	185	205	225	245
6	26	46	66	86	106	126	146	166	186	206	226	246
7	27	47	67	87	107	127	147	167	187	207	227	247
8	28	48	68	88	108	128	148	168	188	208	228	248
9	29	49	69	89	109	129	149	169	189	209	229	249
10	30	50	70	90	110	130	150	170	190	210	230	250
11	31	51	71	91	111	131	151	171	191	211	231	251
12	32	52	72	92	112	132	152	172	192	212	232	252
13	33	53	73	93	113	133	153	173	193	213	233	253
14	34	54	74	94	114	134	154	174	194	214	234	254
15	35	55	75	95	115	135	155	175	195	215	235	255
16	36	56	76	96	116	136	156	176	196	216	236	256
17	37	57	77	97	117	137	157	177	197	217	237	257
18	38	58	78	98	118	138	158	178	198	218	238	258
19	39	59	79	99	119	139	159	179	199	219	239	259
20	40	60	80	100	120	140	160	180	200	220	240	260

Nineteen Strand in Harmonic Module

Kin numbers beginning with 19, lower left, are advancing multiples of nineteen, the last being 19 x 13 (247). Actually, the difference between every kin number on this diagonal is a difference of 19.

19 = 260: THE TZOLKIN "OVER IT IS NINETEEN" STRAND

"All is number. God is a number. God is in all."
-DYNAMICS OF TIME, POSTULATE 19.13

The 0-19 code provides the order of 20 in the ratio 13:20. Like the 0-19 radial matrix, the 13:20 ratio is also organized as a radial matrix of 260 (13 x 20) kin or units. As a cycle of 260 days, the 13 x 20 timing matrix is the fourth-dimensional gauge that synchronically unifies the Thirteen Moon calendar. Known by the traditional and ancient Maya as the *Tzolkin* or "sacred count," the 13 x 20 gauge or Harmonic Module is the key to the mathematics of the fourth dimension, or the "Hereafter." Just as the mathematics of the third dimension is decimal-based, so the mathematics of the fourth dimension is vigesimal-based. Hence, the power of 20 as a self-existing proportion (4 x 5 = 5 x 4).

For any frequency to occur two proportions must be coincident. Thirteen, which is a function of the 7 factor, is the proportion that complements the 0-19 vigesimal proportion of 20. Hence the Tzolkin consists of twenty horizontal orders and thirteen vertical orders = 260 kin, which are also arranged in twenty sets of thirteen-kin wavespells, all numbered 1-13 in dot-bar notation. All of the kin in the 260-unit matrix are also individually numbered: kin 1, kin 2 . . . kin 259, kin 260.

Within the 13:20 order of260 kin, 13 sets of20 or 20 sets of 13, only two of the numbers in the vigesimal code 0-19 occur thirteen times: the number 19 and the number 20 (= positional zero). While 20 occurs precisely thirteen times, 19 occurs thirteen times with a remainder of thirteen (13 x 19 = 247 = 260 - 13).

The position and notational order of the thirteen occurrences of the number 19 in the 260-unit Tzolkin create a perfect strand of thirteen units. This 19 x 13 (19 = 260) strand demonstrates the supreme invariability and inviolability of the number 19 in the vigesimal code of fourth-dimensional time.

The purity of the movement of the 13 x 19 strand of the Tzolkin is further augmented by the vigesimal or 20-count manner of writing numbers. Even as transliterated into the Arabic system of writing numbers, it will be seen that in the vigesimal code every multiple of 19 always adds up to 19, a similar power held by all multiples of 9 in the decimal code. In vigesimal notation, the second order units are always multiples of 20 and not 10. Hence 38 = 1.18, where 1 = 20 + 18; or 5.14 = 114, where 5 = 100 + 14.

Finally, while the sequence 1-13 defines the thirteen galactic tones of fourth-dimensional time, the 0-19 matrix is represented by twenty icons and a four-color (4 x 5 = 20) permutation sequence. The combination of color, icon, and tone gives each of the 260 kin a "galactic signature." This completes the description of the Harmonic Modulerrzolkin.

We now present the thirteen multiples of 19 as defined by the 13 x 19 strand running from lower left to upper right in a perfect cross-stitch sequence that reverses the horizontal order of the 0-19 code frequencies. In other words, the first order represented is the order of (1 x 19) 19 (frequency 19, icon Storm), and the last is the

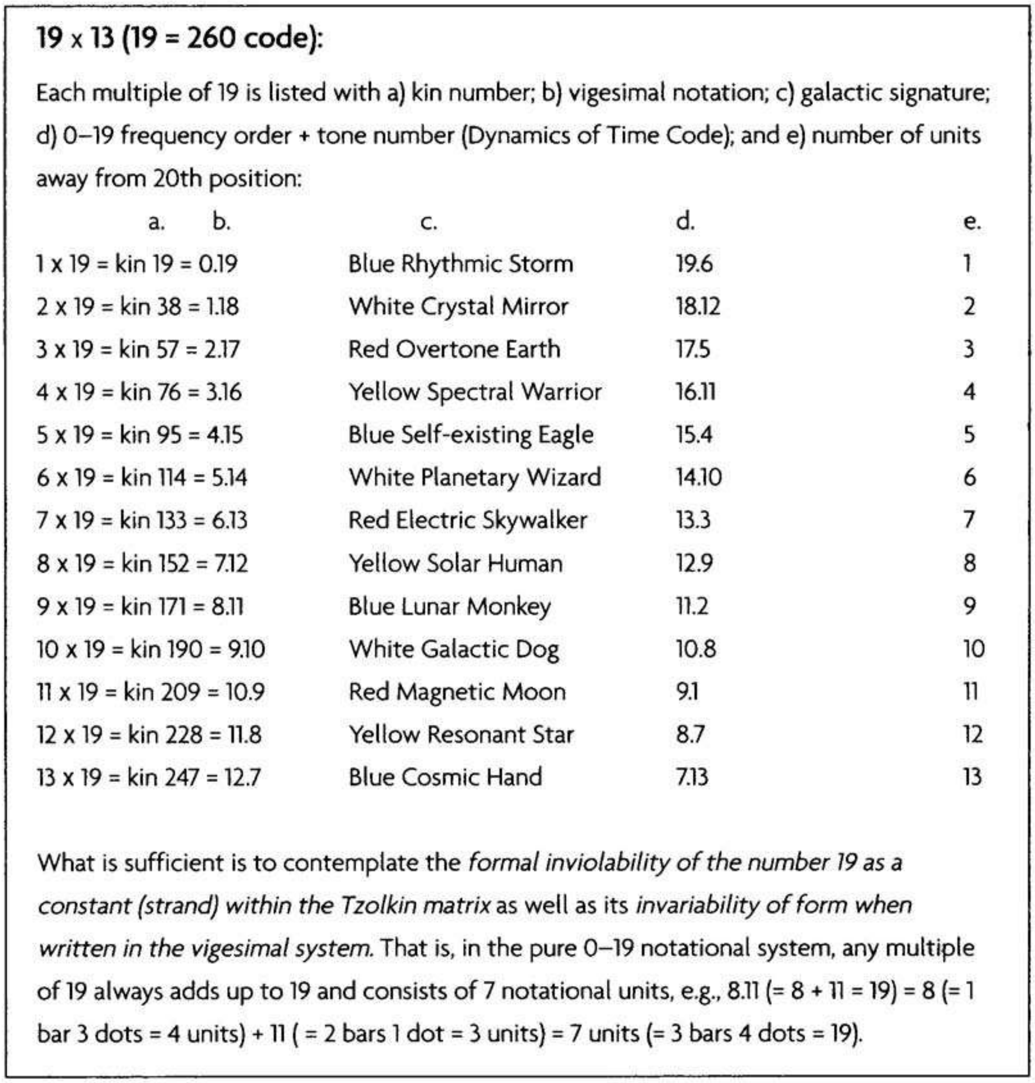

19 x 13 (19 = 260 code):

Each multiple of 19 is listed with a) kin number; b) vigesimal notation; c) galactic signature; d) 0–19 frequency order + tone number (Dynamics of Time Code); and e) number of units away from 20th position:

a. b.	c.	d.	e.
1 x 19 = kin 19 = 0.19	Blue Rhythmic Storm	19.6	1
2 x 19 = kin 38 = 1.18	White Crystal Mirror	18.12	2
3 x 19 = kin 57 = 2.17	Red Overtone Earth	17.5	3
4 x 19 = kin 76 = 3.16	Yellow Spectral Warrior	16.11	4
5 x 19 = kin 95 = 4.15	Blue Self-existing Eagle	15.4	5
6 x 19 = kin 114 = 5.14	White Planetary Wizard	14.10	6
7 x 19 = kin 133 = 6.13	Red Electric Skywalker	13.3	7
8 x 19 = kin 152 = 7.12	Yellow Solar Human	12.9	8
9 x 19 = kin 171 = 8.11	Blue Lunar Monkey	11.2	9
10 x 19 = kin 190 = 9.10	White Galactic Dog	10.8	10
11 x 19 = kin 209 = 10.9	Red Magnetic Moon	9.1	11
12 x 19 = kin 228 = 11.8	Yellow Resonant Star	8.7	12
13 x 19 = kin 247 = 12.7	Blue Cosmic Hand	7.13	13

What is sufficient is to contemplate the *formal inviolability of the number 19 as a constant (strand) within the Tzolkin matrix* as well as its *invariability of form when written in the vigesimal system.* That is, in the pure 0–19 notational system, any multiple of 19 always adds up to 19 and consists of 7 notational units, e.g., 8.11 (= 8 + 11 = 19) = 8 (= 1 bar 3 dots = 4 units) + 11 (= 2 bars 1 dot = 3 units) = 7 units (= 3 bars 4 dots = 19).

order of (13 x 19) 7 (frequency 7, icon Hand), showing again the intimate relation between 19, 13,6, 7, and 20, the key integers of the codes of fourth-dimensional time.

19 = 260, THE SIX SURAS THAT ARE MULTIPLES OF 19

In seeking proof of the Law of Time within the Quranic nineteen code, we need cite but a few examples that can demonstrate from within the nineteen code itself the key mathematical components of the Law of Time. In addition to the number 19, we must look for any of the factors involved in the 13:20 frequency, such as 260, 7, 6-the difference between 19 and 13-and, of course, 13 and 20, or any of the multiples of these integers.

Following the method of Dr. Khalifa, probability analysis of mathematical constituents based on the assumption that certain numbers will be self-verifying, we start our program with the six suras that are multiples of nineteen:

Sura 19	"Maryam"	98 verses
Sura 38	"Saad"	88 verses
Sura 57	"Iron"	29 verses
Sura 76	"The Human"	31 verses
Sura 95	"The Fig"	8 verses
Sura 114	"People"	6 verses
=399, total of sura numbers		=260 verses in six suras that are multiples of nineteen.
= 19.19 vigesimal code		

As can be seen, the sum of the six numbers of the suras that are multiples of 19 is exactly 399, or 19 x 21, the maximum factor in the vigesimal 20 x 20 matrix. Expressed in the vigesimal code this number, 399 = 19.19 [19 (x 20 = 380) +19], the only multiple of 19 in the vigesimal code that is written with two 19s, whose sum is precisely twice nineteen, 38, and is written with fourteen notational symbols rather than seven. This unique number, 19.19 (399), verifies "nineteen is over all," by placing 19 over 19 itself. The unique qualities of this most unique of multiples of 19 also indicates why the Quran consists of exactly six times nineteen suras. Six is the difference between the Quranic 19 and 13, the prime number of the Law of Time.

More amazing than the sum of the six suras that are multiples of 19 being 19.19 is the fact that the sum of the verses of these same six suras is 260. This is incontrovertible evidence for the formulation 19 = 260, and substantiates the mathematical patterning of the Quran to include the principle factor of the Law of Time, 260 (= 13 x 20). We are certain that Dr. Khalifa added the verses of these same suras and may have been perplexed at the meaning of the resultant number, 260. Of course, Dr. Khalifa did not know that 260 is the sum factor of the two proportions that create the natural timing frequency, 13:20.

It is also interesting that discovery of the 13:20 timing frequency, December 10, 1989, was made at almost the precise moment as Dr. Khalifa had completed the introduction to *The Pinal Testament,* the summation of his research on the Quranic nineteen code, Ramadan 26, 1409 (December, 1989). This points to the mysterious movement of the Divine Will in establishing the mathematical proof of the Quran. Just as the work of Dr. Khalifa was terminated within a month of his completing the Final Testament, so the Law of Time, which would substantiate and vindicate the Quranic nineteen code, was ordained and propelled into its process of manifestation at exactly the same moment.

Now with the Law of Time the investigation begun by Dr. Khalifa is complete. Just as there are six suras that are multiples of 19 within the Holy Quran, so the key factor of the Law of Time, the number 260, is coded into the sum of the verses of these six suras. The door is opened for entirely new comprehension of the meaning and purpose of the Holy Quran. The Law of Time and the Holy Quran in mutual resonance will establish the proof of the advent and triumph of the Religion of Truth. (9:33,48:28; 61:9)

Most of the foregoing text of Appendix IV is taken from 19 = *260: The Holy Quran and the Law of Time* (1999), an unpublished manuscript by Jose Argiielles.

▼

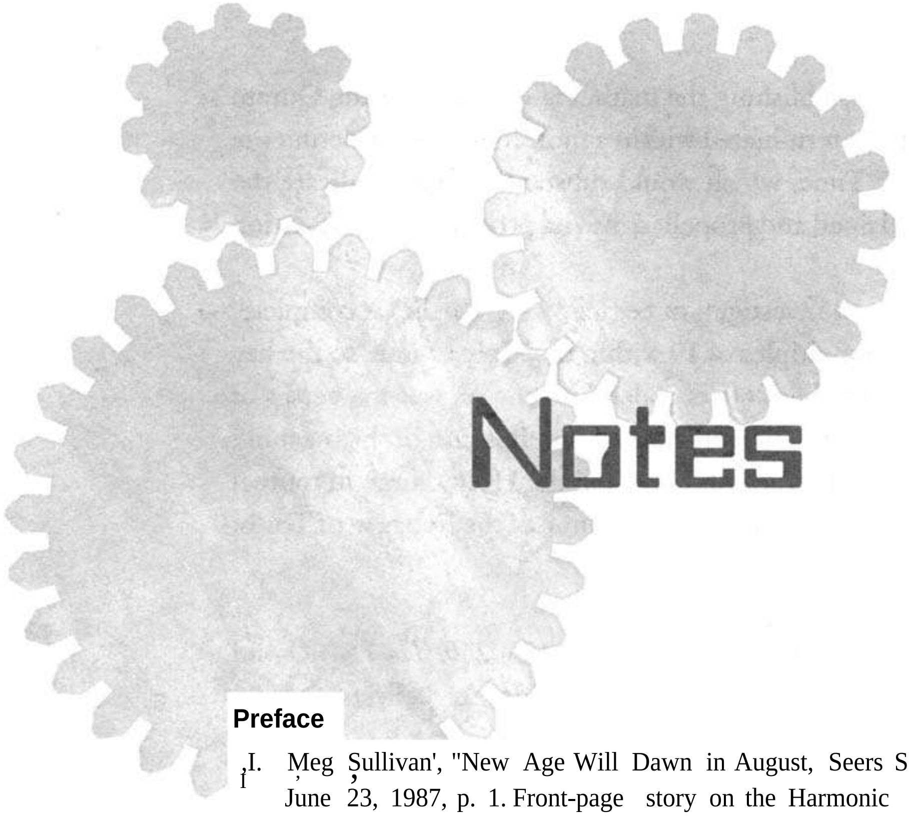

Notes

Preface

1. Meg Sullivan, "New Age Will Dawn in August, Seers Say. . .," *Wall Street Journal,* June 23, 1987, p. 1. Front-page story on the Harmonic Convergence.
2. For the prophetic tradition of the Chilam Balam, see *The Book of Chilam Balam of Chumayel,* edited and translated by Ralph L. Roys (Norman, Okla.: University of Oklahoma Press, 1967); *The Codex Perez and the Book of Chilam Balam of Mani,* edited and translated by Eugene Craine and Reginald Reindorp (Norman, Okla.: University of Oklahoma Press, 1979); and *El Libro de los Libros de Chilam Balam,* edited and translated by Alfredo Barrera Vasquez and Silvia Rendon (Mexico City: Fondo de Cultura Economica, 1948, 1986). The issue of being a "prophetic messenger" is a touchy one for many people. However, as it is written in the Quran, "In truth we are always sending messengers. . ." (44:5). Messengers, as often as not, are self-revealed and confirmed only by God. In my case, I have been able to trace back the prophetic lineage of the Chilam Balam to Pacal Votan, from whom I have received direct transmission. The method of transmission is similar to that of the *terma* tradition of Tibet. Terma, or "hidden treasure," is left by a sage or prophet to be "discovered" at a precise moment much later in time by one designated to do so. See Tulku Thondup Rinpoche, *Hidden Teachings of Tibet: An Explanation of the Terma Tradition of Tibetan Buddhism* (Boston: Wisdom Publications, 1997).
3. This phrase is used often by Dr. Rashad Khalifa (1935-1990), whose groundbreaking translation and research of the Quran through the aid of the computer led to the discovery of the intricate nineteen-based mathematical code that underlies this unique text. His work is summarized in his translation of the Quran with its numerous appendices, *Quran: The Final Testament* (Fremont: Universal Unity, 1992). My own studies on this topic have resulted in two works: *Pacal TfJtIln and Judgement Day* (portland, Oreg.: Planet Art Network, 1999), and 19 = *260: The Holy Quran and the Law of Time,* unpublished, 1999. See also Appendix IV to this book, "The Nineteen Code."

4. See, for example, Samuel P. Huntington's *The Clash of Civilizations and the Remaking of the World Order* (New York: Simon & Schuster, 1996, 1997); Benjamin R. Barber's *Jihad vs. McWorld* (New York: Random House, 1995, 1996); and Yossef Bodansky's *Bin Laden: The Man Who Declared War on America* (New York: Forum, 1999,2001). These writers, of course, all express a pro-globalization perspective.

Chapter 1

1. John ofPatmos, "Book of Revelation," Chapter 18, "The Fall of Babylon," verses 10, 16-17, 19-20, *The Holy Bible, New International Version*(Grand Rapids: Zondervan Bible Publishers, 1989), 1419-20.
2. Jose ArgUelles, *The Call ofPacal Votan: Time is the Fourth Dimension* (Glasgow: Altaea, 1996), 37.
3. Jose ArgUelles, *The Dynamics of Time: 260 Postulates*(Portland, Oreg.: Planet Art Network, 1999),30.
4. Kareng Koshev, "The Great Rainbow Bridge." Altai: unpublished paper, 2001.

Chapter 2

1. 1. Laptev, *The Planet of Reason:A SociologicalStudy of the Man-Nature Relationship* (Moscow: Progress Publishers, 1973; English translation, 1977),220.
2. V. 1. Vernadsky, "L'etude de la vie et la nouvelle physique" in *Revue generale dessciences,* vol. 41: 695.
3. V. 1. Vernadsky, *The Biosphere*(1926), abridged version (Oracle, Arizona and London: Synergetics Press, 1986), 22.
4. Ibid., 76.
5. Ibid., 77-78, 81.
6. Ibid., 82.
7. Ibid.
8. V. 1. Vernadsky, "Problems in Biogeochemistry II," in *Transactionsof the Connecticut Academy of Sciences*(New Haven: Yale University Press, 1944),487-88,498.
9. v.1. Vernadsky, "Problems in Biogeochemistry II," Postulate 18, 511-12.
10. Jose ArgUelles, *The Dynamics of Time: 260 Postulates*(Portland, Oreg.: Pbnet Art Network, 1999), postulates 3.7 and 3.9, 24, 25.
11. V. 1. Vernadsky, "Some Words about the Noosphere" (1944), quoted in Andrey Lapo, *Tracesof Bygone Biospheres*(Moscow and London: Mir Publishers and Synergetic Press, 1987),69.
12. Ibid., 71.
13. G. E. Hutchinson, "The Biosphere," *Scientific American,* vol. 53, no. 3 (1970): 223. Quoted in Lapo, *Tracesof Bygone Biospheres,*77.
14. P. Duvigneaud, "Noosphere et l'avenir de la vegetation du globe," *Proceedingsof the XII International Botanical Congress,Bruxelles,* 1979, 78. Quoted in Lapo, *Tracesof Bygone Biospheres,* 76.

15. Tango Snyder, editor in chief, *The Biosphere Catalogue* (London and Forth Worth: Synergetic Press, 1985), 103.
16. V. 1. Vernadsky, "Scientific Thought as a Planetary Phenomenon," quoted in Lapo, *Traces of Bygone Biospheres, 73-74.*
17. Vernadsky, "Problems in Biogeochemistry II," 511.

Chapter 3

1. Maria Maroushkina, "Is a Path Beyond Technology Possible for the Western Culture?" Moscow: unpublished research paper, 2000.
2. James Glanz, "Physics' Big Puzzle Has Big Question: What Is Time?" *New York Times,* June 17,2001, p. D-1.
3. Ibid.
4. Dogen, "The Time Being," from *Moon in a Dewdrop: Writings of Zen Master Dogen,* edited by Kazuaki Tanahashi (New York: North Point Press, 1985), 76-77.
5. N. A. Kozyrev, "Possibility of Experimental Properties of Time," unpublished paper, 1967, 35. This paper was taken from the Web site www.kozyrev.ru. Virtually unknown in the West, Kozyrev's work has had an especially great influence on the Novosibirsk School of the Russian Academy of Sciences. See, for example, M. M. Lavrentiev, "Physical Theories (Mathematical Models) Adequate to the Reality-A Necessary Condition for Progress of Natural Science of the XXIst century," in *Searchfor Mathematical Laws of the Universe* (Novosibirsk: Sobolev Insititue of Mathematics, 2001) 29-51, and Irina Eganova, "On the Nature of the Stars: N. A. Kozyrev's Approach and His Results," in *Searchfor Mathematical Laws of the Universe,* 215-76; and M. M. Lavrentiev and Irina A. Eganova, "N. A. Kozyrev's Method of Astronomical Observations: Information from True Positions of Stars, Stellar Systems, and Planets," *Instantaneous Action at a Distance in Modern Physics: "Pro" and "Contra,"* edited by Andrew E. Chubykalo, Viv Pope, and Roman Smirnov-Rueda. (New York: Nova Science Publishers, 1999).
6. Pierre Teilhard de Chardin, *The Future of Man* (New York: Harper & Row, 1959), 183.
7. Quoted in V. P. Kaznacheev, "The General Pathology: Consciousness and Physics," Novosibirsk: Russian Academy of Medical Science, unpublished paper, 2000, 28.

Chapter 4

1. Christos Koneatis, *The History of the World-A 6,000 Year Chronicle of Time* (London: Christos Koneatis, 1997), no pagination.
2. Jose and Lloydine Argiielles, *Dreamspell: The Journey of Timeship Earth 2013,* (Hong Kong: Interlink Productions, 1991), 35.
3. Ibid., 5-9.
4. Ibid., 14-15,60,61.
5. V. A. Ponko, private conversation, Novosibirsk, August 18, 2001.
6. *Epic of Gilgamesh,* translated by Caravan of Dreams Theater, quoted in *The Biosphere Catalogue, 116.*

Chapter 5

1. *Webster's Ninth New Collegiate Dictionary,* (Springfield: Merriam Webster, 1987), 1211.
2. See Jose Argiielles, *The Transformative Vision: Reflections on the Nature and History of Human Expression* (Berkeley: Shambhala Publications, 1975),249.
3. Paul Foster Case, *The Tarot: A Key to the WISdom of the Ages* (New York: Macoy, 1947), 158.
4. Jim Hoagland, "The Faithful Come and Wars Erupt Worldwide," *The Oregonian,* December 24,2001, p. C-7.
5. *The Biosphere Catalogue,* 103.

Chapter 6

1. Mahatma Ghandi, "Gandhi on Calendar Reform," *The Journal of Calendar Reform,* vol. I (New York: The World Calendar Association, 1931): 143.
2. All quotes from the Quran used in this text are taken from Rashad Khalifa's translation, *Quran: The Pinal Testament* (Fremont, Calf.: Universal Unity, 1992). All numeration refers to sura number and verse, which are henceforth indicated in parentheses following the quote.
3. Popular quote attributed to Einstein, source unknown.
4. See Broughton Richmond, *Time Measurement and Calendar Construction* (Leiden, Netherlands: J. Brill, 1956), 19-20.
5. Alexander Philip, *Reform of the Calendar* (London: Kegan Paul, Trench, Triibner & Co., Ltd., 1914), 6-8.
6. "Resolution Adopted by the Fourth General Conference on Communications and Transit of the League of Nations," *The Journal of Calendar Reform,* vol I: 162.
7. Argiielles, *The Dynamics of Time,* postulate 6.4,34.
8. *I Ching or Book of Changes,* translated by Richard Wilhelm and Cary F. Baynes (Princeton: Princeton University Press, 1967), 189.
9. Richmond, *Time Measurements and Calendar Construction, 70-71.*
10. David Ewing Duncan, *Calendar: Humanity's Epic Struggle to Determine a True and Accurate Year* (New York: Avon Books, 1998),307.
12. Richmond, *Time Measurements and Calendar Construction, 16.*

Chapter 7

1. Pierre Teilhard de Chardin, *The Phenomenon of Man* (New York: Harper & Row, 1955),288
2. For an enlightening look at the subject of *triskaidekaphobia,* see Jonathon Cott, *Thirteen: A Journey into the Number* (New York: Doubleday, 1996).

Chapter 8

1. Star Roots, 13 *Moon Natural Time Journal* (Manson's Landing, British Columbia: 2001), no pagination.

2. Ben 4 and Ben 5, 13 *Moon Natural Time Calendar, Ancient Science for the Art of Now* (Newport, Oreg.: Skytime, 2001), no pagination.
3. Star Roots, 13 *Moon Natural Ttme Journal*, no pagination.
4. "Golden Age of Miracles and Heaven on Earth," *National Examiner*, November 29, 1988, p. 17.
5. *National Geographic Magazine*, publicity blurb, 2001.
6. "Tracking the Hijackers," *The Oregonian*, October 4,2001, p. A-12. See also Appendix Iv, "The Nineteen Code."
7. Muhammad Husayn Haykal, "Islamic Civilization as Depicted in the Qur'an," in *The Life of Muhammad* (Indianapolis: American Trust Publications, 1993),518.
8. Ibid., 518-19.

Chapter 9

1. See, for instance, Dan Oldenburg in the *Washington Post's* full-page article, "The Harmonic Convergence Revisited," May 4, 1990, p. D- 5.
2. *Nuestra Civilizacion Frente a la Reforma del Calendario* (Our Civilization Confronted with Calendar Reform) (Buenos Aires: Casa Editora Sudamericana, 1944), 6.
3. F. H. Colson, *The Week: An Essay on the Origin and Development of the Seven-Day Cycle* (Cambridge: Cambridge University Press, 1926), 2.
4. Quoted in Jose Argiielles, *Charles Henry and the Formation of a Psychophysical Aesthetic* (Chicago: University of Chicago Press, 1972), 128.

Appendix I

1. *San Jose Mercury News*, quoted on book cover of David Ewing Duncan's *Calendar, 1998.*
2. Duncan, *Calendar*, 289.
3. Philip, *Reform of the Calendar*, 101-2.

Appendix II

1. Charles Muses, *Destiny and Control in Human Systems: Studies in the Interactive Connectedness of Time (Chronotopology)* (Boston- Dordrecht- Lancaster: Kluwer-Nijhoff Publishing, 1984), 20-21.

Appendix IV

1. Rashad Khalifa, *Quran: The Final Testament*, (Fremont, Calif.: Universal Unity, 1992), 375. All further quotes from the Quran and from Rashad Khalifa are from this same work, which in addition to containing the complete Quran in translation also includes thirty-eight Appendices summarizing Dr. Khalifa's research.
2. Ibid., 375.
3. Ibid., 123.
4. Ibid., 375.

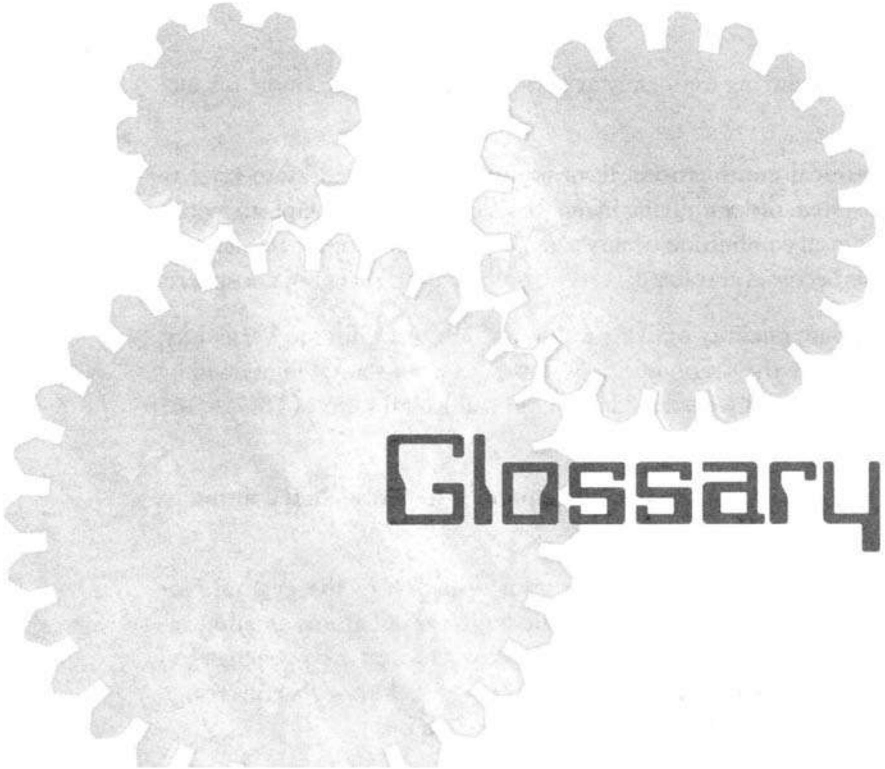

Glossary

THIS GLOSSARY IS constructed to aid the reader in developing a more systematic and technical comprehension of the Biosphere and the Law of Time. Though by no means exhaustive, it is hoped that the terminology used in the text is herein defined in a manner appropriate to cultivating a noospheric point of view.

13:20 timing frequency: Fourth-dimensional timing frequency mathematically expressed as the ratio constant .13:20; universal frequency of synchronization by which the universe is maintained in synchronic order; basis of Law of Time; codified as 260-unit 13:20 matrix of Harmonic Module, coordinates Thirteen Moon/28-Day calendar with synchronic order.

12:60.timing frequency: Artificial, purely third-dimensional timing frequency based on combination of irregular twelve-month calendar and mechanistic sixty-minute hour, adoption of which results in unconscious program within the human mental order, causing the species to deviate altogether from the natural biospheric order, resulting in the technosphere.

astronomical time: Measure of time in third-dimensional physical space without regard to intrinsic harmonics of the synchronic order, usually interpreted as duration and, based solely on matter that is in constant state of entropy, results in standards of measure characterized by infinitesimal.slippages of time and consequent fractional number sequences in need of periodic correction.

baktun: Fifth order of Mayan vigesimal number sequence; in time reckoning equals 144,000 kin (days), or 394.52 years; thirteen per great cycle, measure of human historical era or "great cycle": 3113 B.C.-A.D.2012.

biogenic migration of atoms: Biospheric principle by which evolution over time within finite quantity of living matter (biomass) is accounted for by changes of atomic.,cellular-molecular state and condition; occurring uniformly at periodic transitions accounts for major evolutionary mutations.

biogeochemical, biogeochemical combustion: Term used by V. 1. Vernadsky to refer to biospheric dynamic as whole system order unifying living (bio) and inert (geochemical) matter in a single process; biogeochemical combustion occurs when exponential curve of acceleration of biogenic migration of atoms becomes absolutely vertical, precipitating advent of noosphere.

biomass, biomass constant: Sum quantity of living matter, which, according to Vernadsky, has remained relatively stable since the origin of life on Earth; as a constant, maintained by synchronic timing factor mathematically described as an annual orbital ratio of 73(5) = 365 days per solar orbit of Earth.

biosphere: Sphere or mantle of life and its geochemical support system distributed throughout Earth's ocean-land surface

biosphere-noosphere transition: Description of the transformation of the general biospheric condition due to accelerated excitation of biogenic migration of atoms creating release of free energy and consequent biospheric destabilization, characterized as biogeochemical combustion, whose resolution is emergence of new evolutionary order, the noosphere; change from historical technosphere to post-historic, post-technological phase.

Bolontiku: Nine Lords of Time; in Mayan cosmology, nocturnal rulers of the underworld or *inframundo;* guardians of the passage of death and resurrection; keepers of time and the knowledge of time as noospherically registered in Earth's orbit.

calendar: Derived from the Latin *calends,* or account book, a day for making payments; generic word used to define systems of timekeeping based on measure of Earth's orbit around the sun, often utilizing synodic measure of the moon as a means of calculating.

Chilam Balam: "Jaguar priest," name of Mayan prophet *(ah bobat)* who lived just before the time of conquest, which he accurately predicted; name given to lineage of Mayan prophets extending from the end of tenth baktun (circa A.D. 830) up until the nineteenth century; name given to series of post-conquest texts in which the various categories of prophecies of this lineage are recorded.

chronomancy, geochronomancy: "Divination by time"; the science and art of fourth-dimensional time considered as a whole system where science is knowledge and art is practice; oracular method of reading the synchronic order; description of incipient "Earth time science" incorporating principles of whole Earth geomancy into time structures of the synchronic order.

chronosphere: Sphere of time that activates noosphere; fourth-dimensional field created by the planet holon in resonance with the rotation of the third-dimensional planet body; coded as movement of four chromatics (biopsychic field) across the planet holon, stitching together gravitational fields (five Earth families) and electromagnetic fields (four color families).

chronotopology: Term coined by Charles Muses to define any geography of time and consisting of an event point of meaning that creates a root meaning from which time-formed, -blossomed, or -radiated meanings emerge above a horizon of meaning.

Day Out of Time: The 365th day of the Thirteen Moon/28-Day calendar, no day of the week or month at all (52 x 7 day weeks = 13 moons times 28 days = 364); occurs on Gregorian correlate date July 25, and precedes Thirteen Moon "New Year's" or synchronization date, correlated to July 26.

Dreamspell: Any agreed-on consensus reality; a collective mind-set defining a noospheric epoch, such as "Dreamspell of history"; term given to proofs and tools demonstrating radial and fractal principles of fourth-dimensional mathematics; cosmo-mythic structure of Earth history and its future, otherwise known as *Dreamspell: The Journey of Timeship Earth 2013.*

free energy: Release into biosphere of gases (such as carbon monoxide) and other chemical by-products of industrialization that accelerate biogenic migration of atoms, resulting in biogeochemical combustion; in noosphere refers to release of spectral plasmic-atmospheric phenomena through human telepathic interaction with the electromagnetic field.

fourth dimension: Dimension of time greater than and containing three dimensions of space that it informs and governs.

galactic signature: Combination of one of thirteen galactic tones and twenty solar seals that constitute the harmonic module and that code any given day of the third-dimensional Thirteen Moon calendar.

geocosmic: Holonomic perception expressing unification of Earth whole with the cosmic order.

Gregorian (Julian) calendar: Current world standard calendar consisting of twelve irregularly numbered and obscurely named months coordinated by the 7-day week and intercalary leap year system establishing 28-year cycles; created by the reform of Pope Gregory XIII (1572-1583) in 1582, which "corrected" the Julian calendar (established by Julius Caesar, 45-44 B.C.)by lopping off ten days and refining the intercalary system so that only centuries that are multiples of four are leap years; supreme example of astronomical time measure; principle macro-organizing principle establishing 12:60 artificial timing frequency.

Hannonic Module: 260-unit mathematical matrix of 13:20 frequency, inclusive of a 52-unit binary triplet configuration or set of galactic activation portals; sometimes referred to as Tzolkin, 260-kin sacred count of the Maya; basis of fourth-dimensional vigesimal mathematics of time; coordinated with third-dimensional calendar of Thirteen Moons, establishes basis of 260 galactic signatures.

heliosphere: Sun and its system of planets considered as a whole organism; also known as heliocosm.

holonomics: Of or pertaining to the law governing whole systems.

holonomic consistency and reciprocity: Law by which any whole principle of reality is capable of evoking a like response or reflecting a fractal order at whatever level or through whatever structure.

holon: Any organic whole unit, fractally replicable; basis ofholonomics, law governing whole systems.

human holon: Fourth-dimensional structure of the human being conforming to the 13:20 frequency (twenty digits, thirteen main articulations).

Hunab Ku: "One giver of movement and measure"; in Mayan cosmology, can refer to God as the Supreme Creator.

Inevitable Event: Law of Time definition of the point in synchronic order when the artificial time generated curves of machine, human population, and money converge exponentially in one dramatic, prophecy-fulfilling moment, marking the conclusion oftechnospheric expansion; name given to the events of September 11, 2001 (9-11), epitomized by the col~lapse of World Trade Center Twin Towers (Fall of Babylon).

Julian Count: Established by Thomas Scaliger, 1583, the year after Gregorian calendar "reform," to establish a linear count of days beginning January 1, 4714 B.C.,some 1,600 years prior to commencement of thirteen baktun Long Count of the Maya; official modern scientific count, basis of linear time paradigm.

katun: Fourth order of vigesimal mathematics = 8,000 kin; in Mayan time count, where the second order vinal is 18 and not 20, a katun = 7200 kin, or 19.7 solar years; twenty katuns equal one baktun of 144,000 kin; 260 katuns per thirteen baktun cycle.

kin: In fourth-dimensional mathematics the base unit of measure, possessing fractal properties; one day and night or turn of Earth on its axis; any human being so designated by assuming the galactic signature of the date of birth.

Law of Time: T(E) =Art, Energy factored by Time equals Art, where T (time) is the fourth-dimensional universal synchronization frequency constant 13:20, E (energy) refers to any phenomenal manifestation, and Art to the fact that the synchronizing frequency of time harmonizes everything into a natural condition of beauty.

Long Count: Mayan count of days (kin) measured by thirteen baktuns of 144,000 days each for a total of 1,872,000 kin, the measure of history from JulianiGregorian correlates August 13, 3113 B.C. to December 21, 2012; still maintained through use of Tzolkin by Quiche Maya, Guatemala.

lunar calendar: System of time measurement using synodical measure of the moon-from new moon to new moon, approximately 29.5 days; twelve lunar months equaling a total of 354 days, eleven days short of the measure of solar orbit; used by many of the Old World civilizations; with the exception ofIslamic lunar calendar, most lunar calendars intercalate a thirteenth moon, seven per every nineteen years.

machine: Capacity for the human to externalize its internal processes as well as motor functions into different means of mechanical reproduction, the sum possibilities of which establish a technological complex requiring constant maintenance, thus subordinating human consciousness to its level; the ultimate manifestation of the 12:60 timing frequency; any industrially produced tool believed to advance the level of human comfort and accomplishment.

noosphere: Earth's mental envelope or field, discontinuous with and above the biosphere; unconscious until the discovery and application of the Law of Time; activated by registration of the human biomass in correct 13:20 timing frequency via universal adoption of Thirteen

Moon calendar; description of transformed state of biosphere, coincident with the end of history; condition of universal telepathy subsequent to collapse of technosphere and application of Law of Time; functions in tandem with programs from Earth's octahedral core.

noospheric chip: Description of human operating in noosphere, holonomically resonant with intrinsic 13:20 codes of synchronic order, capable of interacting with the psi bank to participate in the creation of Earth's rainbow brain.

Noospheric Earth Time (NET): Time measure based on Earth in its orbit in relation to the sun and moon in heliospheric space, rather than from the perspective of human (astronomical time); base unit of measure is one rotation of Earth on its axis or one kin (one night and day), which establishes one NET minute; twenty-eight NET minutes per Moon, or one NET hour, 365 NET minutes, or Thirteen NET Hours plus one NET minute per solar orbit; takes into account that for any NET minute axial magnetic bipolarity of Earth is a changing dynamic reaching two peak points called solstices and two stable points called equinoxes; this four-part division is holonomically registered daily in two peak points, midnight (where day begins and ends) and noon, and two stable points, sunrise and sunset.

occult quartet: Demonstration of radial mathematics of Harmonic Module where the tones of any four kin radially equidistant from each other always equal 28; or the tones of any two kin radially opposite always equal 14.

octahedral core: Iron crystal structure at center of Earth; plasmic storage zone in resonance with the noosphere; in 7:7::7:7, telepathically coordinated by four time atoms (= one master time molecule).

omega point: In the work of Pierre Teilhard de Chardin, defines the climactic convergent point (harmonic convergence) of human evolution as the emergence of the hyper-personal; defined by the Law of Time as point of full superconscious establishment of universal telepathy as the norm of the human mental condition; dawn of omega point: winter-northern-summer-southern solstice, 2012, completion of omega point dawn, Yellow Galactic Seed (July 26,2013), otherwise known as Dreamspell of galactic culture.

oracle: For anyone of 260 kin of the Harmonic Module, a cruciform five-part structure demonstrating a uniform set of mathematical relationships in relation to the (0-19) code number of the daily solar seal (destiny kin) informing any day of the Thirteen *Moonl2* 8-Day calendar; consists of center, code number of destiny kin; right analog kin-always 19 when added to the number of destiny kin; left, antipode kin-always a difference of 10 from the destiny kin; below, occult kin-when added to destiny kin always equals 21; and above, guide, always the same color as destiny kin.

Oxlahuntiku: The Thirteen Lords of Time that govern the power of the cosmic order of time that is based on circulatory power of prime number 13 (as opposed to 12, which is static and non-circulatory); coded as thirteen 28-day moons that govern terrestrial solar orbit; basis of prophecy cycle of Quetzalcoatl, Thirteen Heavens (Oxlahuntiku) and Nine Hells (Bolontiku); in Telektonon prophecy, the thirteen-year cycle-2000-2013-in which the first four years, 2000-2004, establish self-existing power of time, and last nine years, 2004-2013 are coded by the Bolontiku, the Nine Lords of Time.

Pacal Votan: Mayan prophet (A.D.603--683), chief technician of Mayan synchronic time science; eleventh ruler of the dynasty of Palenque (Nah Chan), Chiapas, Mexico; assumed power in A.D.615, but key rulership occurred in critical 52-year cycle, 631-683 (seventy-third such cycle since commencement of thirteen baktun count); best known for the carved sarcophagus lid of his elaborately concealed tomb discovered in 1952, 1,260 years after its dedication in A.D.692, which was also 1,320 years before the closing of the cycle, 2012.

Pax Cultura, Pax Biospherica: Principle of new human social order following establishment of Thirteen Moon/28-Day calendar and inaugurating the advent of the noosphere (2004) in which nation-state is replaced by biospheric regionalism governed by a Planetary Biospheric Assembly; due to perfect harmony of new world standard calendar of thirteen moons/28 days, harmony, art, and aesthetic values elevate culture as the new operating norm of the human reintegrated within the biosphere, hence *Pax Cultura, Pax Biospherica.*

planet holon: Twenty-part icosahedral noospheric matrix superimposed on Earth's surface; corresponding to 0-19 code of twenty solar seals, creates three orders analogous to three planetary fields of resonance; gravitational field-horizontal order of five Earth families of four seals each; electromagnetic field-four diagonal bands, from lower left to upper right, of the four color families (red, white, blue, yellow) of five seals each; and biopsychic field-the four chromatics of five seals each moving diagonally from upper left to lower right, which coordinate the daily sequence of the motion of the chronosphere.

psi bank: Regulating mechanism of the noosphere consisting of four bipolar plates, each plate in mirror symmetry containing 520 (260 x 2) psi chrono units, 2,080 in all, which are the complete registrations of the universal fourth-dimensional 13:20 timing frequency governing all of the various evolutionary stages of the terrestrial biogeocosm; engine of the chronosphere, located between the two radiation belts 2,000 and 11,000 miles above Earth's surface.

psi chrono unit: Time-bearing information unit; base unit of the psi bank; codes and contains all information for anyone of the 260 galactic signatures of the 13:20 matrix or Harmonic Module; psycho activated through coordination with Thirteen Moon/Telektonon known as Rinri Project.

Psychozoic Era: Term coined by V. 1. Vernadsky defining the next geological era; meaning the era of the spiritualization of life, it is coincident with the noosphere in its full emergence after the omega point has been reached, 2012-2013.

Quetzalcoatl: Mexican prophet (A.D.947-999) who left the legacy of prophecy of Thirteen Heavens and Nine Hells, basis of Harmonic Convergence (1987) and commencement of the "Time of Prophecy," 1987-2013.

Quran, Holy Quran: Literally "recitation;" last revelation and moral-spiritual criterion for historic humanity; received over a twenty-three-year period by the Prophet Muhammad (A.D.570-632); proven by Dr. Rashad Khalifa (1935-1990) to be a radically nonlinear text rigorously constructed of a nineteen-based mathematical code.

radion, radial plasmas: Telepathic fluid or lubricant released through telepathic interaction with the seven primary plasmas, or electrically charged particles, hence radial plasmas; the names of the seven plasmas replace the names of the week in the Thirteen Moon/28-Day calendar; activated in the 7:7::7:7 Telektonon Revelation practice.

radiosonics: Name given to the new science based on synchronicity rather than causality and in which, as a consequence, synaesthetic experiences and syntropic values are normalized.

rainbow brain: Description of activated noosphere, where human telepathic interaction with the electromagnetic field triggers the release of the circumpolar rainbow bridge-Earth's day-night alternator system.

Rinri Project: Law of time practice for coordinating the opening of the psi bank, psi chrono unit per psi chrono unit, on a daily basis over a four-year cycle by means of the Thirteen MoontTelektonon.

synchronic order: Order of fourth-dimensional time, in contrast to third-dimensional astronomical time, that synchronizes all aspects of third-dimensional physical plane reality including its timekeeping systems; entirely new domain of human mental-spiritual experience coincident with the dawning of noosphere; characterized by absolute harmonic perfection, described by vigesimal 0-19 code, as distinct from decimal, hex, or duodecimal mathematics that govern third-dimensional timing codes.

synchronometer: Measure of synchronization or synchronicity; precise Law of Time definition of means and purpose of what is generically referred to as the "calendar."

technosphere: Global sphere of technology; artificial stage between biosphere and noosphere, representing the climax-biogechemical combustion-of acceleration of biogenic migration of atoms precipitating biospheric crisis as a prelude to the advent of noosphere; characterized by exponential curves of human population, machine, and money at expense of biospheric integrity and stability; defined as the specific fifty-six-year cycle (two Gregorian 28-year cycles) 1945-2001, in which war and terrorism are the governing features; complete manifestation of the 12:60 order of time; concluded on September **11,2001,** the Inevitable Event.

Telektonon: "Earth Spirit Speaking Tube," the name of the psychoduct leading from the tomb of Pacal Votan beneath the Temple of the Inscriptions to the temple floor atop the pyramid, the discovery of which, in 1949, led to the excavation and revelation of the long-concealed tomb in 1952; name given to the prophecy ofPacal Votan, decoded in 1993-1994, prophesying the Thirteen Moon/28-Day calendar as the only means to keep the human race from destroying the biosphere and itself; definition of the fourth-dimensional structure of the heliosphere in which consciousness is a function of the planetary orbits; basis of a board "game" coded to synchronic order via Thirteen Moon calendar for the awakening of telepathy; basis of the Cube of the Law, days 7-22, containing numerous principles of the Law of Time as a daily practice; basis of other practices for establishing synchronic order as principles of a new consciousness, including *20 Tabletsof the Law of Time* and the 7:7::7:7.

third dimension: Space; physical plane of reality coordinated and contained by time, fourth-dimensional order of reality.

Thirteen Moon/28-Day calendar: Pragmatic application of the Law of Time, perpetual harmonic perfection of which is intended to correct for disorder of human consciousness resulting from adherence to irregular, astronomical timing standards and mechanistic timing devices; coordinated with Harmonic Module and Telektonon establishes means for conscious human participation in synchronic order; 28-day measure based on female menstrual cycle, also corresponds to apsidal measure of moon from the point at which the polar axis is tipped farthest from Earth; prevalent in pre-history, all but eliminated by historical consciousness that made 13 taboo; perfect vehicle for establishing noosphere as daily consciousness in post-technospheric post-history.

time: Fourth dimension; universal factor coordinating space to manifest synchronically; traveling faster than speed of light, accounts for instantaneity of information transmission (telepathy) from distant points; expressed as formulation which states that the velocity of time is instantaneously infinite; fourth-dimensional structures that are radial and fractal in nature, third-dimensional cycles being manifestations thereof; expressed as T(E) = Art, accounts for universal harmony of the cosmos.

tun: Third positional order of vigesimal mathematics where one unit = 400 kin; in Mayan time count, because of second positional order, vinal, being 18 instead of 20, one tun = 360 kin; basis of measure ofkatuns and baktuns in Long Count, where 5,200 tuns = 260 katuns = 13 baktuns, or 5,125 solar years.

Tzolkin: "Sacred Count," 260-day calendar of the Maya, basis of Mayan time knowledge and the Long Count; temporal manifestation of Harmonic Module or universal 13:20 timing matrix.

UR: "Universal Religion," "Universal Recollection;" description of human spirituality in the noosphere or Psychozoic Era.

vigesimal, 0-19 code: Twenty-count based mathematics utilizing zero for positional orders that advance in exponential binary fashion; also known as 0-19 dot-bar code notational system, where dots equal single units, bars equal five units, plus positional zero; basis of mathematics of fourth-dimensional order of time, distinct from mathematical systems of third-dimensional space.

Wave Harmonic of history: Fourth-dimensional description of thirteen baktun cycle (3113 B.C.-A.D 2012) identified precisely with 260-unit Harmonic Module, where each of thirteen baktuns contains twenty katuns (13 x 20); basis of creating synchronic order out of entropic disorder of third-dimensional historical time.

wavespell: Thirteen-unit cosmological form-constant for coordinating fourth-dimensional timing cycles; consists of two gates, Magnetic first position and Cosmic thirteenth position; two towers, fifth and ninth positions; and three sets of chambers, nine chambers in all; when coordinating Thirteen Moon/28-Day calendar, known as Planetary Service Wavespell.

Wizard's Count: Name of count to establish synchronic order as manifestation of the Law of Time in order to transit from 12:60 technosphere to 13:20 noosphere; based on prophetically modified year-bearer count of the Chilam Balam tradition; correlated by the Dreamspell to July 26, 1987, White Galactic Wizard, every four years there are four year bearers: Yellow Seed, Red Moon, White Wizard, and Blue Storm (one four-year Seed-Storm Year bundle) that, combined with thirteen galactic tones creates 52-year solar-galactic cycle, or in NET time, one Earth Year, also known as a Wizard's Year.

▼

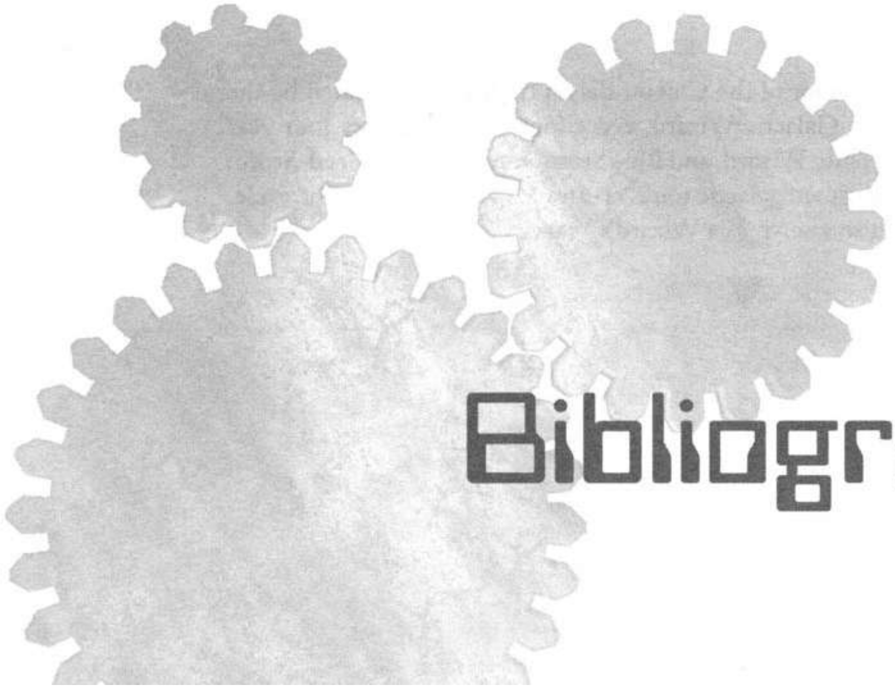

Bibliography

I. About the Biosphere, Technosphere, and Noosphere

,Argiielles, Jose: *EJ:rh Ascending: An Illustrated Treatise on the Law Governing Whole Systems.* \, B~.~ldh: Shambhala Publications, 1984. Second and third editions, Santa Fe: Bear & Co., 1988 and 1996.

-. "Peace is the Culture of the Biosphere: Statement and Reflections." Brightwood, Oregon: Foundation for the Law of Time, unpublished paper, 1996.

Argiielles, Jose and Lloydine Argiielles. *A Complete Guide to the First Planetary Congressof BiosphericRights.* Portland, Oreg.: Planet Art Network, 1996,2002.

-. *The World Thirteen Moon Calendar Change PeacePlan.* Portland: Planet Art Network. 1999.

Coats, Callum. *Living Energies: An Exposition of ConceptsRelated to the Theories of Viktor Schauberger.* Bath, England: Gateway Books, 1996.

"Global Consciousness Project: Registering Coherence and Resonance in the World," http://noosphere. princeton.edu.com.

Kaznacheev, V. P. "The General Pathology: Consciousness and Physics." Novosibirsk: Russian Academy of Medical Science, 2000.

Lapo, Andrey V. *Tracesof Bygone Biospheres.*Moscow and London: Mir Publishers and Synergetic Press, 1987.

Laptev, 1. *The Planet of Reason: A SociologicalStudy of Man-Nature Relationship.* Moscow: Progress Publishers, 1973. English Translation, 1977.

Ponyko, V. A. *Vednie V Systemy "Ekoprognos":Metodologiya PrognosirovaniyaAnomalii Ekosperi* (Ecoprognosis: Methodology for Making Prognosis of Anomolous Phenomena in the Ecosphere). Moscow: Institute Ustoychivogo Rasveteya, 2000.

Smil, Vaclav. *Energies: An Illustrated Guide to the Biosphereand Civilization.* Cambridge, Mass. and London: MIT Press, 1999.

Snyder, Tango, editor in chief, and the Institute of Ecotechnics. *The Biosphere Catalogue.* London and Forth Worth: Synergetic Press, 1985.

Teilhard de Chardin, Pierre. *The Future of Man.* New York: Harper & Row, 1964.

-. *The Phenomenon of Man.* New York: Harper & Row, 1959.

Vernadsky, V. 1. *The Biosphere.* Abridged version based on the French edition of 1929. Oracle, Arizona and London: Synergetic Press, 1986.

-. *The Biosphere: Complete Annotated Edition.* Revised and annotated by Mark A. S. McMenamin and Far West Institute. New York: Copernicus/Springer Verlag, 1998.

-. "Problems in Biogeochemistry 11." In *Transactions of the Connecticut Academy of Sciences,* vol. 35 (1944): 483-517.

II. About the Law and Science of Time

Acheles, Elisabeth. *Of Time and the Calendar.* New York: Hermitage House, 1955.

-. *The World Calendar Reform Since 1930.* New York: G. P. Putnam's Sons, 1937.

Alcala, Luis Alvarez. *El Calendario Azteca.* Edited by B. Costa-Amico Mexico City: 1970.

Argiielles, Jose. "Calendar Reform and the Future of Civilization: Preparatory Reflections for the World Summit on Peace and Time," San Jose, Costa Rica: University for Peace, unpublished paper, 1999.

-. *The Call of Pacal Votan: Time is the Fourth Dimension.* Glasgow: Altea, 1996.

-. "Campaign for the New Time: Science Demonstrates a New Time of Peace on Earth," Brightwood, Oregon: Foundation for the Law of Time, unpublished position paper, 2000.

-. *Earth Ascending: An Illustrated Treatise on the Law Governing Whole Systems.* Santa Fe: Bear & Co, 1984, 1988, 1996

-. *The Dynamics of Time: 260 Postulates.* Portland, Oreg.: Planet Art Network, 1999.

-. *The Mayan Factor: Path Beyond Technology.* Santa Fe: Bear & Co., 1987. Second edition, 1996.

Argiielles, Jose and Lloydine Argiielles. *Discovering the New Time.* Tokyo: Fuunsha, 1997.

-. *Thirteen Moons in Motion.* Portland, Oreg.: Planet Art Network, 1999.

Barbour, Julian. *The End of Time: The Next Revolution in Physics.* Oxford: Oxford University Press, 1999.

Ben 4 and Ben 5. 13 *Moon Natural Time Calendar: Ancient Science for the Art of Now.* Newport, Oreg.: Skytime, 2001.

Colson, F. H. *The Week: An Essay on the Origin and Development of the Seven-Day Cycle.* Cambridge: Cambridge University Press, 1926.

Cott, Jonathon. *Thirteen: A Journey into the Number.* New York: Doubleday, 1996.

Coveney, Peter and Richard Highfield. *The Arrow of Time: A Voyage Through Science to Solve Time's Greatest Mystery.* New York: Fawcett-Columbine, 1990.

Davies, Paul. *About TIme: Einstein's Unfinished Revolution.* New York: Simon & Schuster, 1995.

Deutsch, David. *The Fabric of Reality.* London: Penguin Books, 1997.

Dogen. "The Time-Being," from *Moon in a Dewdrop: Writings of Zen Master Dogen.* Edited by Kazuaki Tanahashi, New York: North Point Press, 1985. 76-83.

Duncan, David Ewing. *Calendar: Humanity's Epic Struggle to Determine a True and Accurate Year.* New York: Avonbooks, 1998.

Dunne, J. w. *An Experiment with Time.* Charlottesville, Va.: Hampton Roads, 2001.

Eganova, Irina. "On the Nature of the Stars: N. A. Kozyrev's Approach and His Results." In *Search for Mathematical Laws of the Universe,* edited by M. M. Lavrentiev, 215-276. Novosibirsk: Sobolev Institute of Mathematics, 2001.

Fraser, J. T. *Time, the Familiar Stranger.* Redmond, Wash.: Tempus-Microsoft, 1987.

Fraser, J. T., ed. *The Voices of Time: A Cooperative Survey of Man's Views of Time as Expressed by the Sciences and the Humanities.* New York: George Brazilier, 1966.

Heath, Robin. *Sun, Moon & Earth.* Powys, Wales: Wooden Books, 1999.

International Society for the Study of Time (ISSO), www.studyoftime.org/index.html.

The Journal of Calendar Reform, vols. I and II. New York: The World Calendar Association, 1931,1932.

Koneatis, Christos. *The History of the World-A 6,000 Year Chronicle of Time.* London: Christos Koneatis, 1997.

Koshev, Kareng. "The Great Rainbow Bridge," Altai: unpublished paper, 2001.

Kozyrev, N. A. "Possibility of Experimental Study of Properties of Time," Novosibirsk: Russian Academy of Sciences, unpublished paper, September, 1967.

-. www.kozyrev.ru.

Landes, David S. *Revolution in Time: Clocks and the Making of the Modern World.* Cambridge, Mass.: Belknap Press, 1983.

Lavrentiev, M. M. "Interpretation of Astronomical Data Connected with N. A. Kozyrev's Discovery." In *Memoranda of Differential Equations of Mathematical Physics,* vol. 12, (1997): 135-41.

-. "Physical Theories (Mathematical Models) Adequate to the Reality: A Necessary Condition for Progress of Natural Science of the XXIst Century." In *Search for Mathematical Laws of the Universe,* 29-51. Novosibirsk: Sobolev Institute of Mathematics,2001.

Lavrentiev, M. M., and Irina A. Eganova. "N. A. Kozyrev's Method of Astronomical Observations: Information from True Positions of Stars, Stellar Systems, and Planets." In *Instantaneous Action at a Distance in Modern Physics: "Pro" and "Contra,"* edited by Andrew E. Chubykalo, Viv Pope, and Roman Smirnov-Rueda. New York: Nova Science Publishers, 1999.

Malmstrom, Vincent H. *Cycles of the Sun, Mysteries of the Moon: The Calendar in Mesoamerican Civilization.* Austin: University of Texas Press, 1997.

Man and Time: Papers from the Eranos Yearbook 3. Bollingen Series XXX. Edited by Joseph Campbell. New York: Pantheon Books, 1957.

Maroushkina, Maria. "Is a Path Beyond Technology Possible for Western Culture?" Moscow: unpublished paper, 2000.

Muses, Charles. *Destiny and Control in Human Systems: Studies in the Interactive Connectedness of Time (Chronotopology).* Boston-Dordrecht-Lancaster: Kluwer-Nijhoff Publishing, 1984.

Nuestra Civilizacion Frente a la Reforma del Calendario (Our Civilization Confronted with Calendar Reform). Buenos Aires: Casa Editora Sudamericana, 1944.

Penrose, Roger. *Shadows of the Mind: A Search for the Missing Science of Consciousness.* Oxford: Oxford University Press, 1994.

Philip, Alexander. *The Reform of the Calendar.* London: Kegan Paul, 1914.

Philosophy of Time. www.dna2z.com/projects/time.html.

Price, Huw. *Time's Arrow and Archimedes Point.* Oxford: Oxford University Press, 1996.

Reference Library Books. *The World of Time and Timepieces.* London: B. P. C. Publishing, 1972.

Richmond, Broughton. *Time Measurement and Calendar Construction.* Leiden: J. Brill, 1956.

Seabury, Rev. Samuel. *The Theory and Use of the Church Calendar in the Measurement and Distribution of Time.* New York: Pott, Young & Co., 1872.

Spinden, HerbertJ. *Maya Dates and What They Reveal.* Brooklyn: Museum of the Brooklyn Institute of Arts and Sciences, 1930.

Star Roots. 13 *Moon Natural Time Journal.* Manson's Landing, British Columbia, 2001.

The Study of Time: Proceedings of the First Conference of the International Society for the Study of Time, Oberwolfach, Black Forest, West Germany. Edited by J. T. Fraser,. F. C. Haber and G. H. Miiller. Berlin-Heidelberg-New York: Springer Verlag, 1972.

The Study of Time II: Proceedings of the Second Conference of the International Society for the Study of Time, Lake Yamanaka, Japan. Edited by J. T. Fraser, F. C. Haber and G. H. Miiller. Berlin-Heidelberg-New York: Springer Verlag, 1975.

Vanderkam,James C. *Calendars in the Dead Sea Scrolls: Measuring Time.* London and New York: Routledge, 1998.

Votan, Valum. *Principia Mathematica of the Fourth Dimension: Introduction and Annotated Bibliography.* Portland, Oreg.: Planet Art Network, 1999.

Whitrow, G. J. *The Natural Philosophy of Time.* New York: Clarendon Press, 1980.

Wolf, Fred Allen. *Parallel Universes: The Search for Other Worlds.* New York: Touchstone, 1988.

III.About the Inevitable Event: Prophecy, History, and Politics

Ahmad, Mirza Ghulam of Qadian. *The Philosophyof the Teachingsof Islam.* London: London Mosque, 1979.

America at War. New York: Profiles Presents, 2001.

Argiielles, Jose. *Pacal Votan and Judgement Day*.Portland, Oreg.: Planet Art Network, 1999.

Barber, Benjamin R. *Jihad vs. McWorld.* New York: Random House, 1995, 1996.

Bodansky, Yossef. *Bin Laden: The Man Who Declared War on America.* New York: Forum, 1999,2001.

The Book of Chilam Balam of Chumayel. Edited and translated by Ralph L. Roys. Norman, Okla.: University of Oklahoma Press, 1967.

Case, Paul Foster. *The Tarot: A Key to the Wisdom of the Ages.* New York: Macoy, 1947.

The Codex Perez and the Book of Chilam Balam of Mani. Edited and translated by Eugene Craine and Reginald Reindorp. Norman, Okla.: University of Oklahoma Press, 1979.

"Golden Age of Miracles and Heaven on Earth," *National Examiner,* November 29, 1988, p. 17.

Haykal, Muhammad Husayn. "Islamic Civilization as Depicted in the Qur'an." In *Lift of Muhammad.* Translated from the 8th edition by Isma'il Ragi A. al-Faruqi, 517-55. Indianapolis: American Trust, 1993.

Hoagland, Jim. "The Faithful Come and Wars Erupt Worldwide," *The Oregonian,* December 24,2001, p. C-7.

Huntington, Samuel P. *The Clash of Civilizations and the Remaking of the World Order.* New York: Simon & Schuster, 1996, 1997.

I Ching or Book of Changes. Translated by Richard Wilhelm and Cary F. Baynes. Princeton: Princeton University Press, 1967.

Izquierdo, Ana Luisa. *Alberto Ruz Lhuillier Frente al Pasado de los Mayas.* Mexico City: Secretaria de Educacion Publica, 1987.

Jenkins, John Major. *Maya Cosmogenesis 2012: The True Meaning of the Maya Calendar End-Date.* Santa Fe: Bear & Co., 1998.

John of Patmos, "Book of Revelation," Chapter 18, "The Fall of Babylon." *The Holy Bible, New International Version.* Grand Rapids: Zondervan Bible Publishers, 1989.

Katchongva, Dan. *From the Beginnings of Lift to the Day of Purification: Teachings, History & Prophecies of the Hopi People.* Los Angeles: Committee for Traditional Indian Land and Life, 1972.

Khalifa, Rashad. *Quran: The Final Testament.* Fremont, Calif.: Universal Unity, 1992.

Kyipchakoba, P. v., ed. *Altai U Drevnue Maya* (Altai and the Ancient Maya). Gorno-Altaiski, 2001.

Landa, Fray Diego de. *Yucatan Before and After the Conquest.* Translated by William Gates. New York: Dover, 1978.

El Libro de los Libros de Chilam Balam. Edited and translated by Alfredo Barrera Vasquez and Silvia Rendon. Mexico: Fondo de Cultura Economica, 1948, 1986.

Mahmud, Sayyid Fayyaz. *A Short History of Islam.* Oxford: Oxford University Press, 1960.

Mano, Galactica Azul. "La Profecia Maya y las Torres Gemeles" (Mayan Prophecy and the Twin Towers). Manzaneles, Colombia: unpublished research paper, October, 2001.

Oldenburg, Dan. "The Harmonic Convergence Revisited," *Washington Post,* May 4, 1990, p. D-5.

The Roerich Pact and the Banner of Peace. New York: The Roerich Pact and Banner of Peace Committee, 1947.

Ruz, Alberto. *Palenque, Official Guide.* Mexico City: Instituto Nacional de Antropologia e Historia, 1978.

September 11,2001: The Day that Changed America. New York: God Bless America, 2001.

September 11,2001. New York: Time-Warner, 2001

Shearer, Tony. *Beneath the Moon and Under the Sun: A Reappraisal of the Sacred Calendar and the Prophecies of Ancient Mexico.* Albuquerque: Sun Books, 1975.

"Tracking the Hijackers," *The Oregonian,* October 4,2001, p. A-12.

Tulku Thondup Rinpoche, *Hidden Teachings of Tibet: An Explanation of the Terma Tradition of Tibetan Buddhism.* Boston: Wisdom Publications, 1986, 1997.

▼

Books of Related Interest

THE MAYAN FACTOR
Path Beyond Technology
by Jose Argiielles

EARTH AsCENDING
An Illustrated Treatise on Law Governing Whole Systems
by Jose Arguelles

SURFERS OF THE ZUVUVA
Tales of Interdimensional Travel
by Jose Arguelles

MAYA COSMOGENESIS 2012
The True Meaning of the Maya Calendar End-Date
by John Major Jenkins

GALACTIC ALIGNMENT
The Transformation of Consciousness According to Mayan, Egyptian, and Vedic Traditions
by John Major Jenkins

THE MAYAN ORACLE
Return Path to the Stars
by Ariel Spilsbury and Michael Bryner

SECRETS OF MAYAN SCIENCEIRELIGION
by Hunbatz Men

THE LOST REALMs
by Zecharia Sitch in

Inner Traditions. Bear & Company
P.O. Box 388
Rochester, VT 05767
1-800-246-8648
www.InnerTraditions.com

Or contact your local bookseller

www.foundationforthelawoftime.org

www.tortuga.com

The Inevitable Event and the End of History

The end of 12:60 Gregorian Time
Pope John Paul II after the Inevitable Event

Planetary 9-11, 2001. First Year, third Millenium

"Woe! Woe, O Great City, O Babylon
in one hour your doom has come!"
—Revelations, 18:10

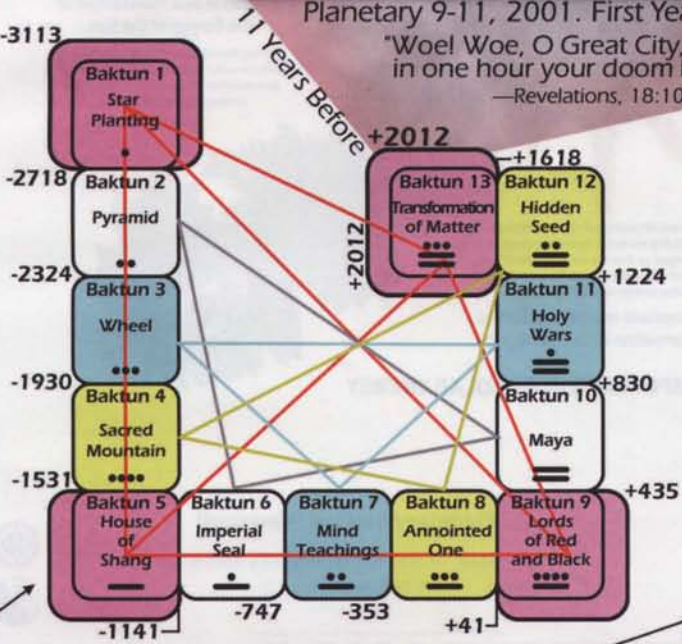
11 Years Before
-3113
Baktun 1
Star Planting
-2718
Baktun 2
Pyramid
-2324
Baktun 3
Wheel
-1930
Baktun 4
Sacred Mountain
-1531
Baktun 5
House of Shang
Baktun 6
Imperial Seal
Baktun 7
Mind Teachings
Baktun 8
Annointed One
Baktun 9
Lords of Red and Black
-1141
-747
-353
+41
+435
Baktun 10
Maya
+830
Baktun 11
Holy Wars
+1224
Baktun 12
Hidden Seed
+1618
Baktun 13
Transformation of Matter
+2012
+2012

13 Tone Wavespell
as Thirteen Baktun
Wave Harmonic
of History

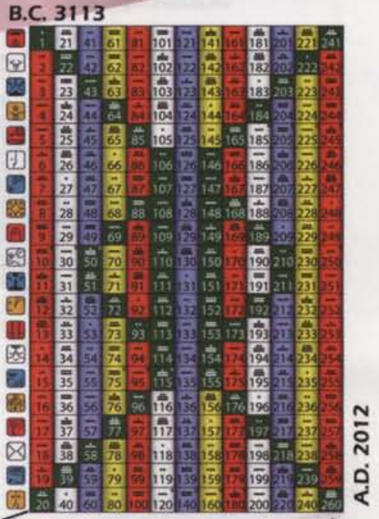
B.C. 3113
A.D. 2012

1	21	41	61	81	101	121	141	161	181	201	221	241
2	22	42	62	82	102	122	142	162	182	202	222	242
3	23	43	63	83	103	123	143	163	183	203	223	243
4	24	44	64	84	104	124	144	164	184	204	224	244
5	25	45	65	85	105	125	145	165	185	205	225	245
6	26	46	66	86	106	126	146	166	186	206	226	246
7	27	47	67	87	107	127	147	167	187	207	227	247
8	28	48	68	88	108	128	148	168	188	208	228	248
9	29	49	69	89	109	129	149	169	189	209	229	249
10	30	50	70	90	110	130	150	170	190	210	230	250
11	31	51	71	91	111	131	151	171	191	211	231	251
12	32	52	72	92	112	132	152	172	192	212	232	252
13	33	53	73	93	113	133	153	173	193	213	233	253
14	34	54	74	94	114	134	154	174	194	214	234	254
15	35	55	75	95	115	135	155	175	195	215	235	255
16	36	56	76	96	116	136	156	176	196	216	236	256
17	37	57	77	97	117	137	157	177	197	217	237	257
18	38	58	78	98	118	138	158	178	198	218	238	258
19	39	59	79	99	119	139	159	179	199	219	239	259
20	40	60	80	100	120	140	160	180	200	220	240	260

Baktun 13 - Katun 260 - 1993 - 2012
End of Second Christian Millennium A.D. 2000

Harmonic Module
13 Baktun
260 Katun
13:2 Matrix of History
1 Kin = 1 Katun
1 Baktun = 20 Katuns

Nostradamus Prediction of nuclear destruction of New York in Year 2000

THE END OF THE WORLD
O FIM DO MUNDO

Cover of "Istoe" (Brazil) May, 1996. "End of the World."

16. The Tower

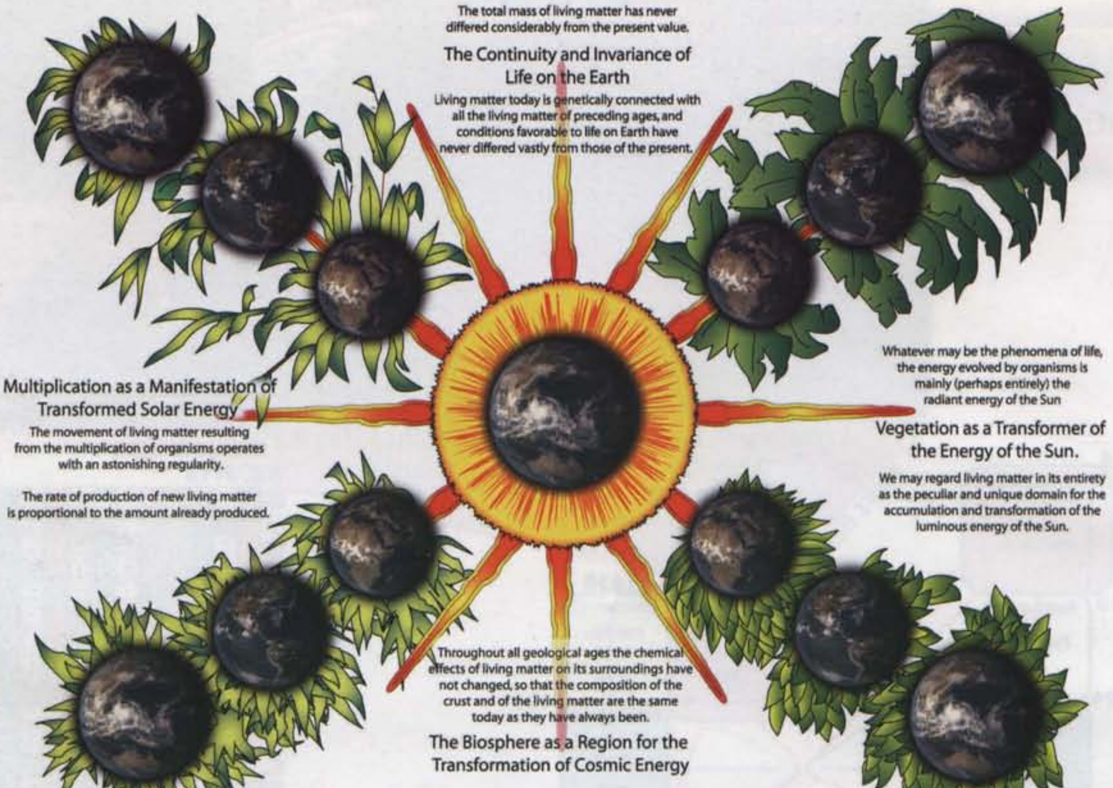

The total mass of living matter has never differed considerably from the present value.

The Continuity and Invariance of Life on the Earth

Living matter today is genetically connected with all the living matter of preceding ages, and conditions favorable to life on Earth have never differed vastly from those of the present.

Multiplication as a Manifestation of Transformed Solar Energy

The movement of living matter resulting from the multiplication of organisms operates with an astonishing regularity.

The rate of production of new living matter is proportional to the amount already produced.

Whatever may be the phenomena of life, the energy evolved by organisms is mainly (perhaps entirely) the radiant energy of the Sun

Vegetation as a Transformer of the Energy of the Sun.

We may regard living matter in its entirety as the peculiar and unique domain for the accumulation and transformation of the luminous energy of the Sun.

Throughout all geological ages the chemical effects of living matter on its surroundings have not changed, so that the composition of the crust and of the living matter are the same today as they have always been.

The Biosphere as a Region for the Transformation of Cosmic Energy

LIFE IS THE TRANSFORMATION OF SOLAR ENERGY

Planetary Service Wavespell

YEAR OF THE RED PLANETARY MOON

© 2002.7.26 ~ 2003.7.25

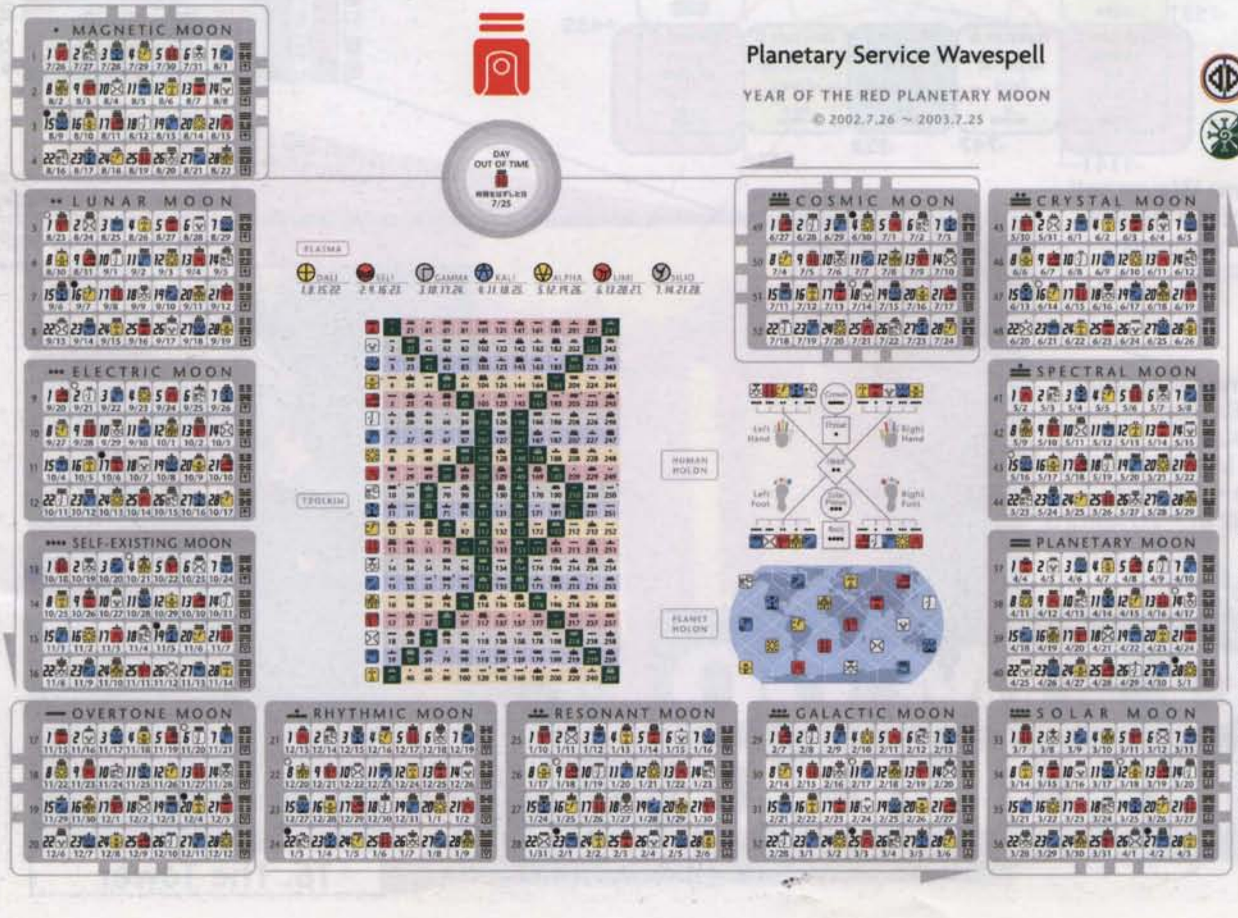
MAGNETIC MOON
LUNAR MOON
ELECTRIC MOON
SELF-EXISTING MOON
OVERTONE MOON
RHYTHMIC MOON
RESONANT MOON
GALACTIC MOON
SOLAR MOON
PLANETARY MOON
SPECTRAL MOON
CRYSTAL MOON
COSMIC MOON
DAY OUT OF TIME
7/25
PLASMA
DALI
SELI
GAMMA
KALI
ALPHA
LIMI
SILIO
TZOLKIN
HUMAN HOLON
PLANET HOLON
Left Hand
Right Hand
Left Foot
Right Foot

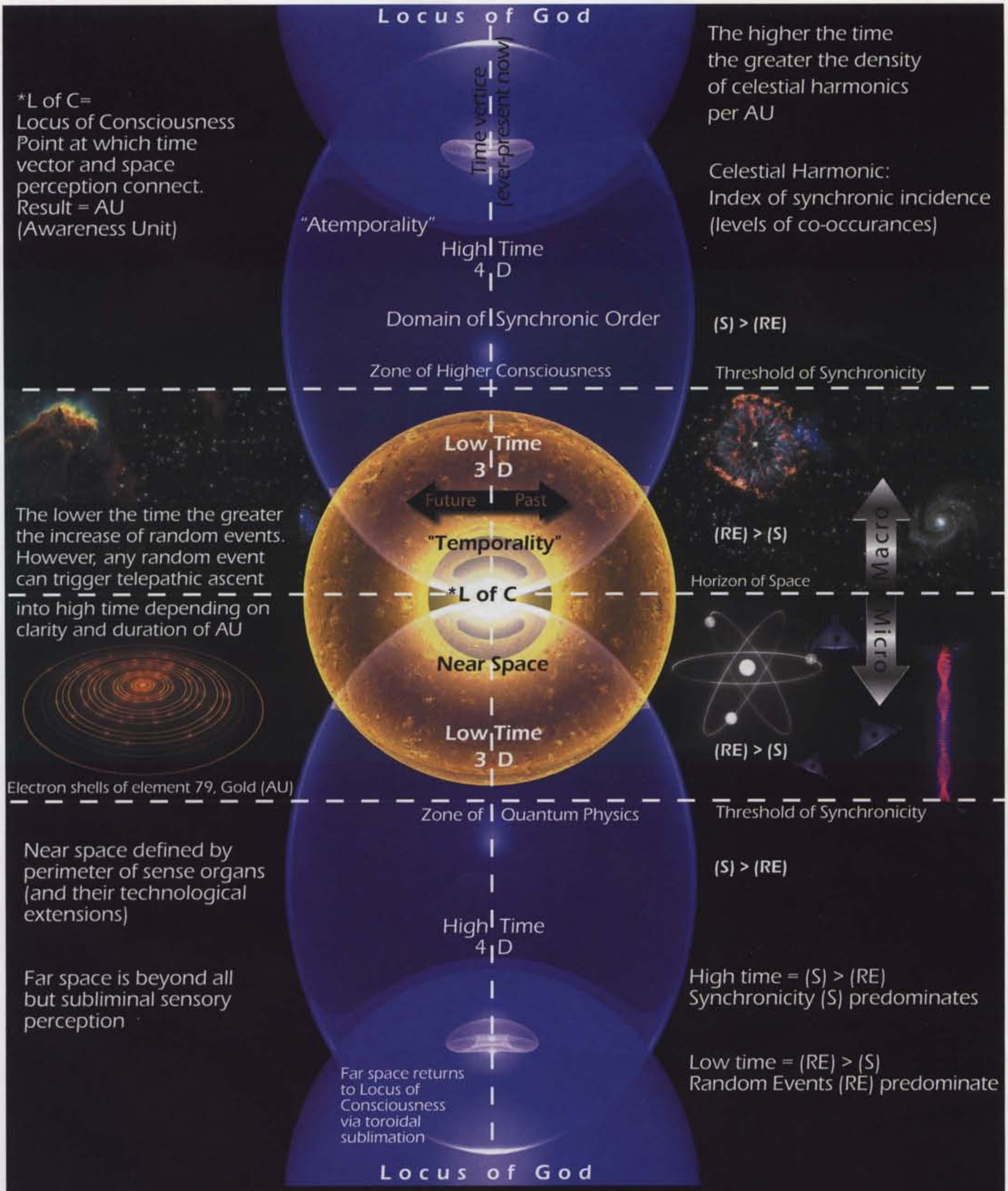

Locus of God

Time vertice (ever-present now)

*L of C=
Locus of Consciousness
Point at which time vector and space perception connect.
Result = AU
(Awareness Unit)

The higher the time the greater the density of celestial harmonics per AU

Celestial Harmonic:
Index of synchronic incidence (levels of co-occurances)

"Atemporality"

High Time
4 D

Domain of Synchronic Order

(S) > (RE)

Zone of Higher Consciousness

Threshold of Synchronicity

Low Time
3 D

Future

Past

"Temporality"

The lower the time the greater the increase of random events. However, any random event can trigger telepathic ascent into high time depending on clarity and duration of AU

(RE) > (S)

Macro

Horizon of Space

*L of C

Micro

Near Space

Low Time
3 D

(RE) > (S)

Electron shells of element 79, Gold (AU)

Zone of Quantum Physics

Threshold of Synchronicity

Near space defined by perimeter of sense organs (and their technological extensions)

(S) > (RE)

High Time
4 D

Far space is beyond all but subliminal sensory perception

High time = (S) > (RE)
Synchronicity (S) predominates

Low time = (RE) > (S)
Random Events (RE) predominate

Far space returns to Locus of Consciousness via toroidal sublimation

Locus of God

Model of "Vertical Time" in Relation to Horizontal Space

A Presentation of the Foundation for the Law of Time
T(E)=ART. Energy factored by Time equals Art.

The Dynamics of Time
Showing the Evolution of Time as Consciousness - T(E)=Art

GALACTIC BRAIN, or evolutionary spectrum showing index of energy-mind progression from atomic-cellular biological to purely mental conditions of being, T(E)=Art. Energy factored by time equals art.

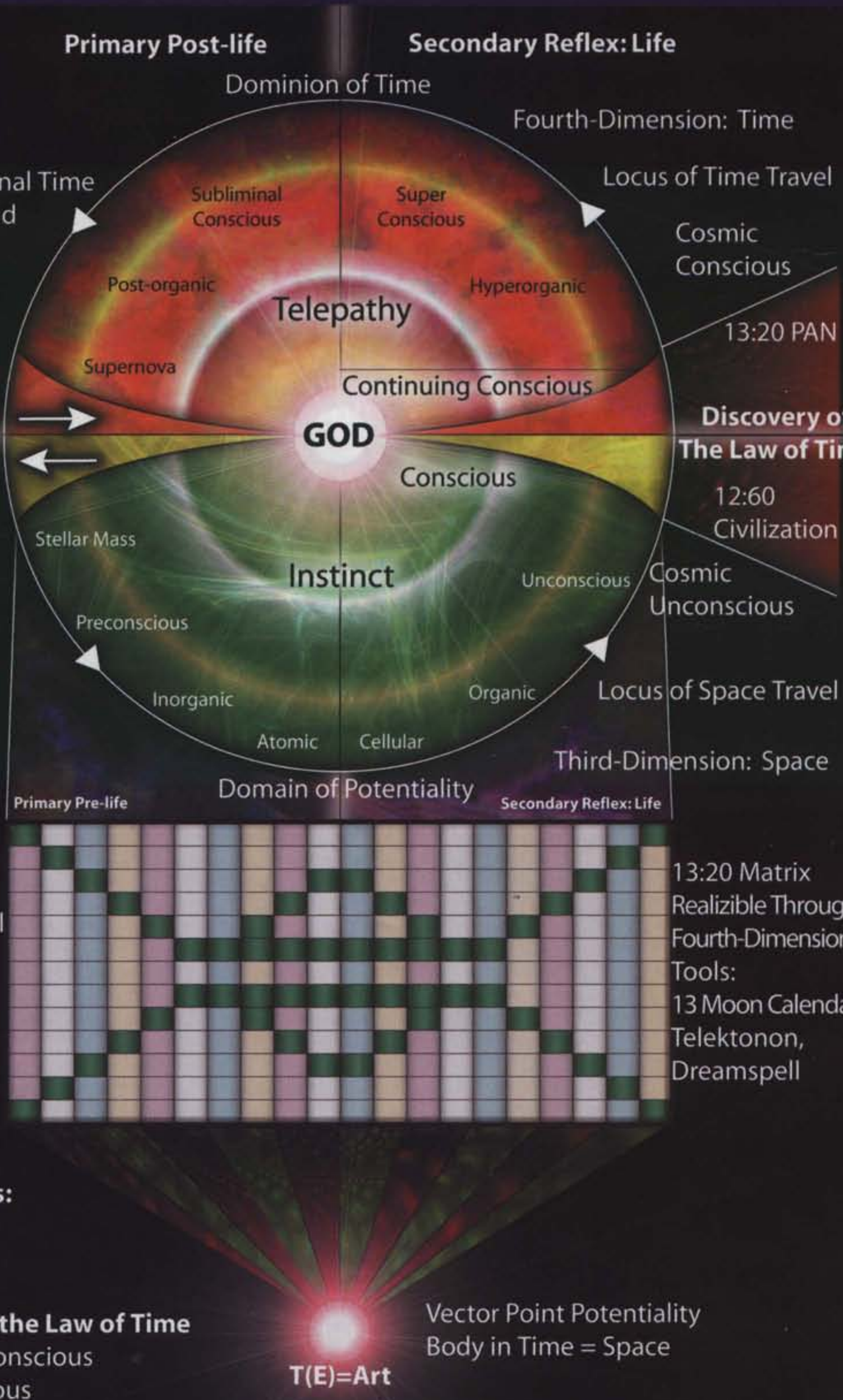

Primary Post-life

Secondary Reflex: Life

Dominion of Time

Fourth-Dimension: Time

Hyperdimensional Time
Pure Mind

Locus of Time Travel

Subliminal Conscious

Super Conscious

Cosmic Conscious

Post-organic

Hyperorganic

Telepathy

13:20 PAN

Supernova

Continuing Conscious

Mind Returns to Source

GOD

Discovery of The Law of Time

Energy Released from Source

Conscious

12:60 Civilization

Stellar Mass

Instinct

Unconscious

Cosmic Unconscious

Preconscious

Quantum
Dimensional
Space
Pure Energy

Inorganic

Organic

Locus of Space Travel

Atomic

Cellular

Third-Dimension: Space

Domain of Potentiality

Primary Pre-life

Secondary Reflex: Life

13:20
Fourth-Dimensional
Timing Matrix
Source of
Time Vector
Potentialities

13:20 Matrix
Realizible Through
Fourth-Dimensional
Tools:
13 Moon Calendar,
Telektonon,
Dreamspell

Time Vector Potentialities:
Preconscious
Unconscious
Conscious
Discovery of the Law of Time
Continuing Conscious
Super Conscious
Subliminal Conscious

Vector Point Potentiality
Body in Time = Space

T(E)=Art

Noospheric Time Map - Wave Harmonic of History

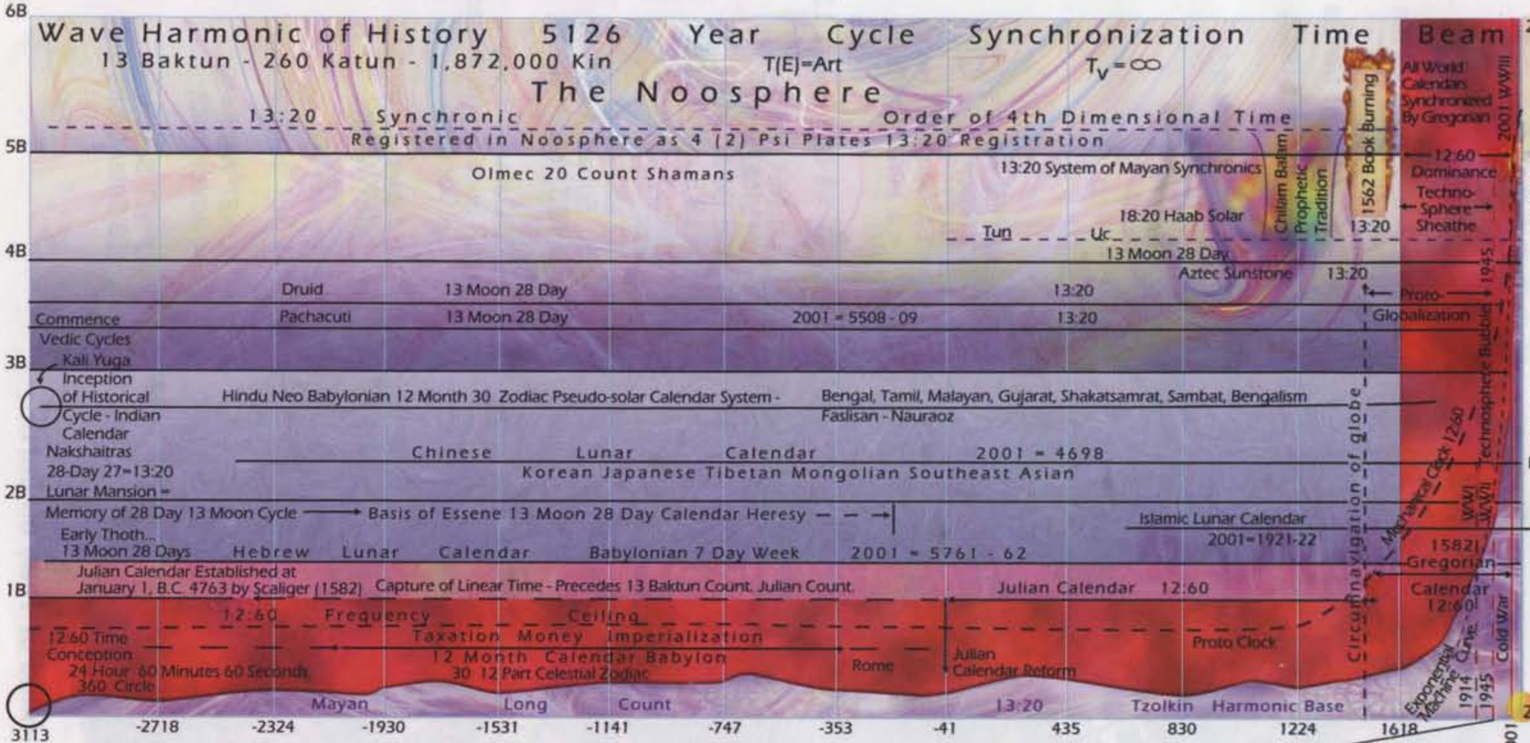
6B
5B
4B
3B
2B
1B
Wave Harmonic of History 5126 Year Cycle Synchronization Time Beam
13 Baktun - 260 Katun - 1,872,000 Kin
T(E)=Art
$T_v = \infty$
The Noosphere
13:20 Synchronic Order of 4th Dimensional Time
Registered in Noosphere as 4 (2) Psi Plates 13:20 Registration
Olmec 20 Count Shamans
13:20 System of Mayan Synchronics
18:20 Haab Solar
Tun
Uc
13 Moon 28 Day
Aztec Sunstone
Chilam Balam Prophetic Tradition
1562 Book Burning
13:20
All World Calendars Synchronized By Gregorian
2001 WWIII
12:60 Dominance
Techno-Sphere Sheathe
1945
Proto-Globalization
Technosphere Bubble
Druid
13 Moon 28 Day
13:20
Pachacuti
13 Moon 28 Day
2001 = 5508 - 09
13:20
Commence Vedic Cycles
Kali Yuga
Inception of Historical Cycle - Indian Calendar Nakshaitras
28-Day 27=13:20 Lunar Mansion =
Hindu Neo Babylonian 12 Month 30 Zodiac Pseudo-solar Calendar System -
Bengal, Tamil, Malayan, Gujarat, Shakatsamrat, Sambat, Bengalism
Faslisan - Nauraoz
Chinese Lunar Calendar 2001 = 4698
Korean Japanese Tibetan Mongolian Southeast Asian
Circumnavigation of globe
Mechanical Clock 12:60
WWI WWII
Memory of 28 Day 13 Moon Cycle → Basis of Essene 13 Moon 28 Day Calendar Heresy
Islamic Lunar Calendar
2001=1921-22
Early Thoth... 13 Moon 28 Days
Hebrew Lunar Calendar
Babylonian 7 Day Week
2001 = 5761 - 62
1582 Gregorian
Julian Calendar Established at January 1, B.C. 4763 by Scaliger (1582)
Capture of Linear Time - Precedes 13 Baktun Count. Julian Count.
Julian Calendar 12:60
Calendar 12:60
Cold War
12:60 Frequency Ceiling
Taxation Money Imperialization
12:60 Time Conception
12 Month Calendar Babylon
30 12 Part Celestial Zodiac
24 Hour 60 Minutes 60 Seconds 360 Circle
Rome
Julian Calendar Reform
Proto Clock
Exponential Machine Curve
1914
1945
Mayan Long Count
13:20
Tzolkin Harmonic Base
3113
-2718
-2324
-1930
-1531
-1141
-747
-353
-41
435
830
1224
1618
2012

2013 - Galactic Synchronization
New Evolutionary Phase Harmonic
"Psychozoic Era"
New Solar Age
Noosphere Planet -
Post-Mechanization
Post-Historic
Galactic Culture
Time is Art.

13 Moon 28 Day
Calendar of Peace
2004
13:20 Frequency
Phase Shift

Peace
Through
Culture

Noospheric Chip
Psychic Biomes
Post-Apocalyptic
PAN Harmonic
Universal Unified
Time Field
Timeship Earth 2013.

"Inevitable Event 9-11"
Pressure point where
Technosphere Bubble -
Time Warp in the
Noosphere bursts open.
Built in failure of
Gregorian Calendar.

12:60 Evaporated
Rhythmic Solstice
December 21, 2012
End of Thirteen Baktun
Wave Harmonic of History

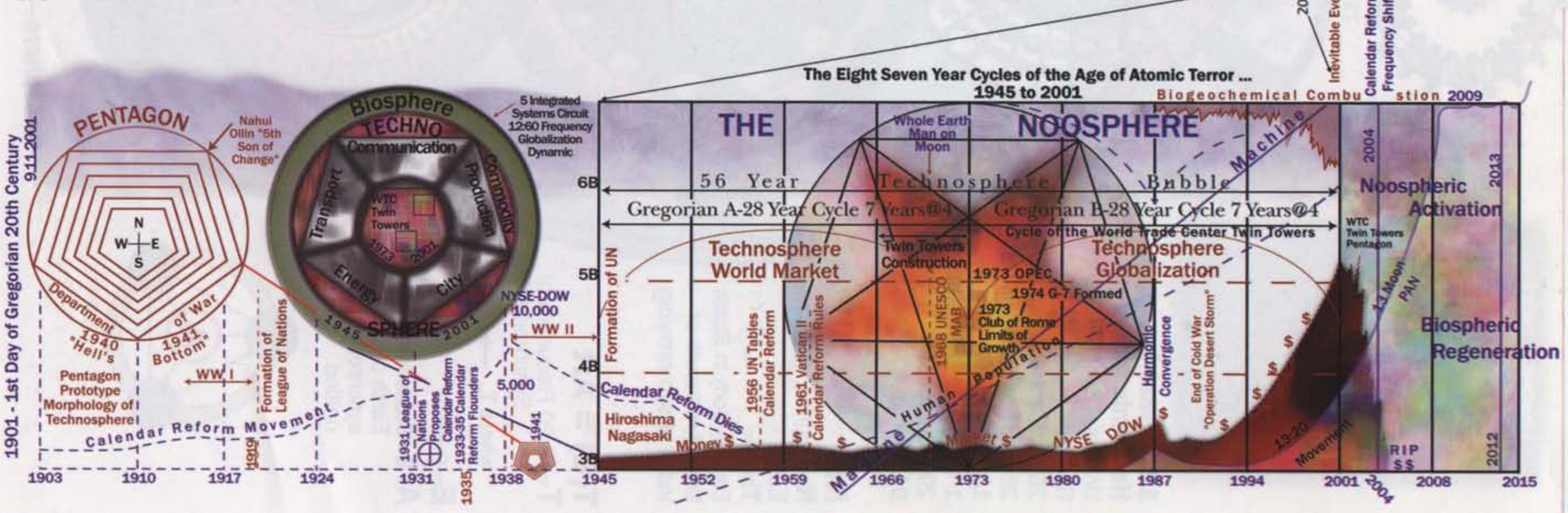
The Eight Seven Year Cycles of the Age of Atomic Terror ...
1945 to 2001

1901 - 1st Day of Gregorian 20th Century
9.11.2001
PENTAGON
Nahui Ollin "5th Son of Change"
N
W
E
S
Department of War
1940 "Hell's
1941 Bottom"
Pentagon Prototype Morphology of Technosphere
WW I
Formation of League of Nations
1919
Calendar Reform Movement
Biosphere
TECHNO
SPHERE
Communication
Commodity
Production
City
Energy
Transport
WTC Twin Towers
1973
2001
1945
5 Integrated Systems Circuit 12:60 Frequency Globalization Dynamic
1931 League of Nations Proposes Calendar Reform
1933-35 Calendar Reform Flounders
1935
1941
NYSE-DOW 10,000
5,000
WW II
1903
1910
1917
1924
1931
1938
1945
2001
Inevitable Event
Calendar Reform Frequency Shift
Biogeochemical Combustion 2009
6B
5B
4B
3B
THE NOOSPHERE
Whole Earth Man on Moon
56 Year Technosphere Bubble
Gregorian A-28 Year Cycle 7 Years@4
Gregorian B-28 Year Cycle 7 Years@4
Cycle of the World Trade Center Twin Towers
Twin Towers Construction
Technosphere World Market
Technosphere Globalization
Formation of UN
1973 OPEC
1974 G-7 Formed
1973 Club of Rome Limits of Growth
1968 UNESCO MAB
Population
Human
Machine
Calendar Reform Dies
Hiroshima Nagasaki
Money $
1956 UN Tables Calendar Reform
1961 Vatican II Calendar Reform Rules
Market $
NYSE DOW
Harmonic Convergence
End of Cold War "Operation Desert Storm"
13:20 Movement
WTC Twin Towers Pentagon
13 Moon PAN
2004
Noospheric Activation
Biospheric Regeneration
RIP $$
2013
2012
1952
1959
1966
1973
1980
1987
1994
2001
2004
2008
2015

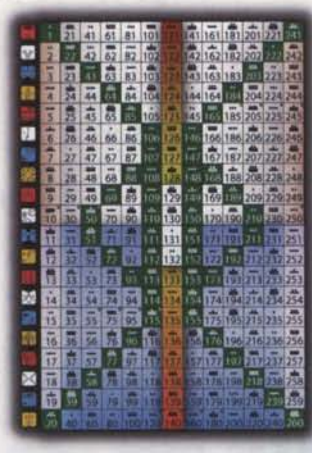

1	21	41	61	81	101	121	141	161	181	201	221	241
2	22	42	62	82	102	122	142	162	182	202	222	242
3	23	43	63	83	103	123	143	163	183	203	223	243
4	24	44	64	84	104	124	144	164	184	204	224	244
5	25	45	65	85	105	125	145	165	185	205	225	245
6	26	46	66	86	106	126	146	166	186	206	226	246
7	27	47	67	87	107	127	147	167	187	207	227	247
8	28	48	68	88	108	128	148	168	188	208	228	248
9	29	49	69	89	109	129	149	169	189	209	229	249
10	30	50	70	90	110	130	150	170	190	210	230	250
11	31	51	71	91	111	131	151	171	191	211	231	251
12	32	52	72	92	112	132	152	172	192	212	232	252
13	33	53	73	93	113	133	153	173	193	213	233	253
14	34	54	74	94	114	134	154	174	194	214	234	254
15	35	55	75	95	115	135	155	175	195	215	235	255
16	36	56	76	96	116	136	156	176	196	216	236	256
17	37	57	77	97	117	137	157	177	197	217	237	257
18	38	58	78	98	118	138	158	178	198	218	238	258
19	39	59	79	99	119	139	159	179	199	219	239	259
20	40	60	80	100	120	140	160	180	200	220	240	260

260 Kin
4th Dimensional
Wheel of
Synchronic
Order =
Factor of
Synchronization=
1 G-Spin =

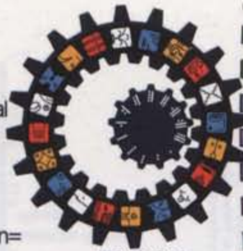

73 G-Spins = 52 Years

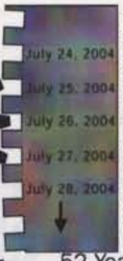
July 24, 2004
July 25, 2004
July 26, 2004
July 27, 2004
July 28, 2004

365-Kin
1 NET Day
3rd Dimensional
Biotelepathic
Wheel =
Factor of
Life Experience
(73:5)
(7:52)

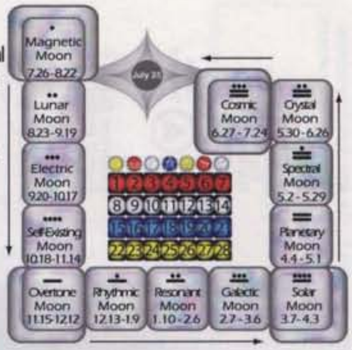
Magnetic Moon 7.26-8.22
Lunar Moon 8.23-9.19
Electric Moon 9.20-10.17
Self-Existing Moon 10.18-11.14
Overtone Moon 11.15-12.12
Rhythmic Moon 12.13-1.9
Resonant Moon 1.10-2.6
Galactic Moon 2.7-3.6
Solar Moon 3.7-4.3
Planetary Moon 4.4-5.1
Spectral Moon 5.2-5.29
Crystal Moon 5.30-6.26
Cosmic Moon 6.27-7.24
July 25
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28

Wavespell = 13 Tones
1 Tone = 1 Moon = 28 Kin
1 Kin = 1 NET Minute

◐ = 1 NET Minute
x 28 = 1 NET Hour.

7 NET Minutes =
$^1/_4$ Hour =
1 Week ◔

13:20 Matrix Harmonic Module

Universal Synchronization
Gauge =
260 Kin = 1 Galactic Spin =
260 NET Minutes
13 NET Minutes =
1 Wavespell (13 Tones)
20 Wavespells =
9 NET Hours, 8 NET Minutes

NET

Noospheric
Earth
Time =
Earth = Body in Space =
Unit of measure of
index of synchronicity

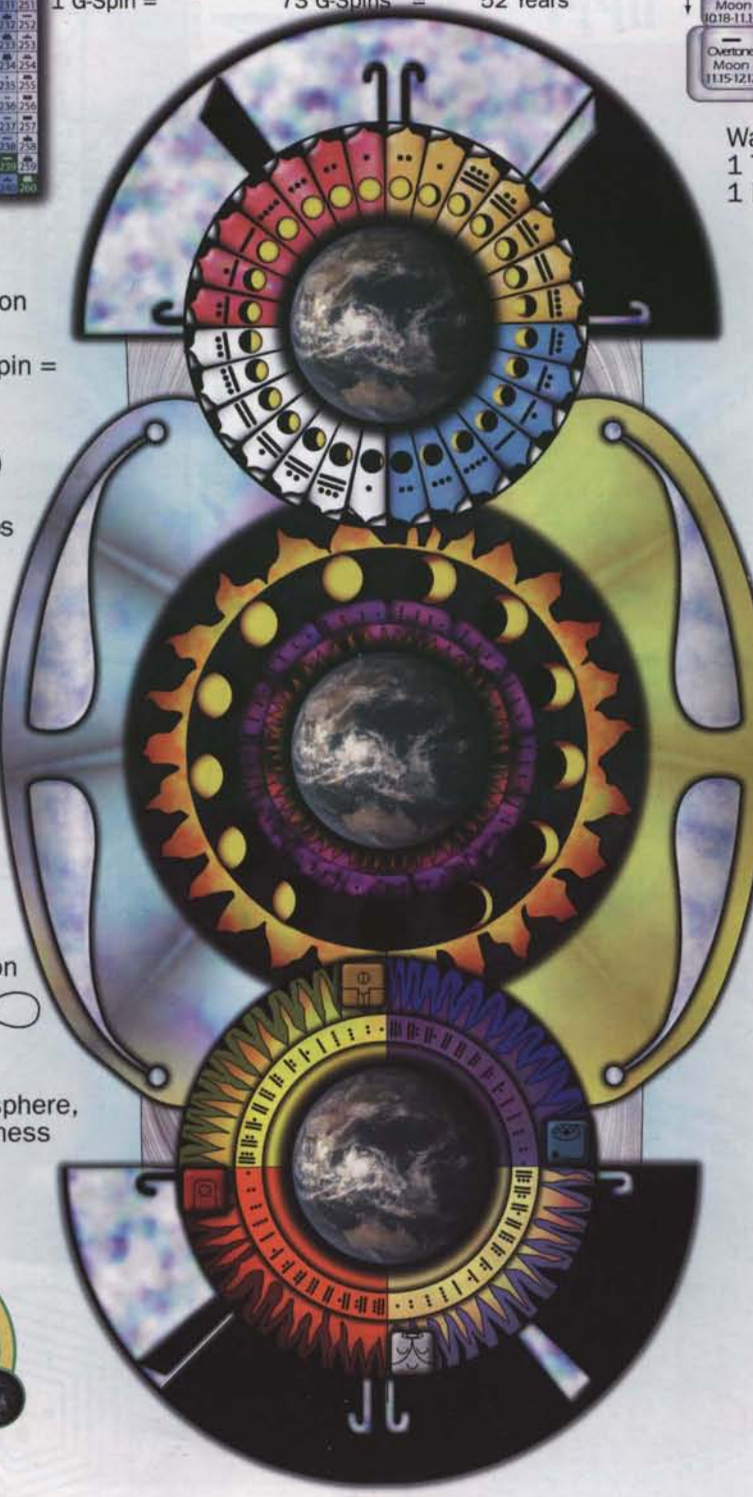

T(E) = Art

T=13:20 Frequency of Synchronization
where $T_V \rightarrow \infty$

E= Earth

Art= Blossomed Noosphere, Earth consciousness synchronically unified

1 NET Hour = ◐
1 Moon

1 Moon (NET Hour)
= 1 Lunar Orbit

◐ =1 NET Hour x 13
= 1 NET Day
(13 NET Hours)
= 1 Solar Orbit
= 13 Lunar Orbits

52 NET Days
= 52 Solar Orbits
= 1 NET Year

= 73 Galactic Spins
= 13 Seed-Storm 4-Year Bundles

1 4-Year Bundle =
1 NET Week

13 NET Days=
1 NET Season

1 NET Week=
4 Solar Orbits

1 NET Season=
13 Solar Orbits

Noosphere—Mind of Time on Earth

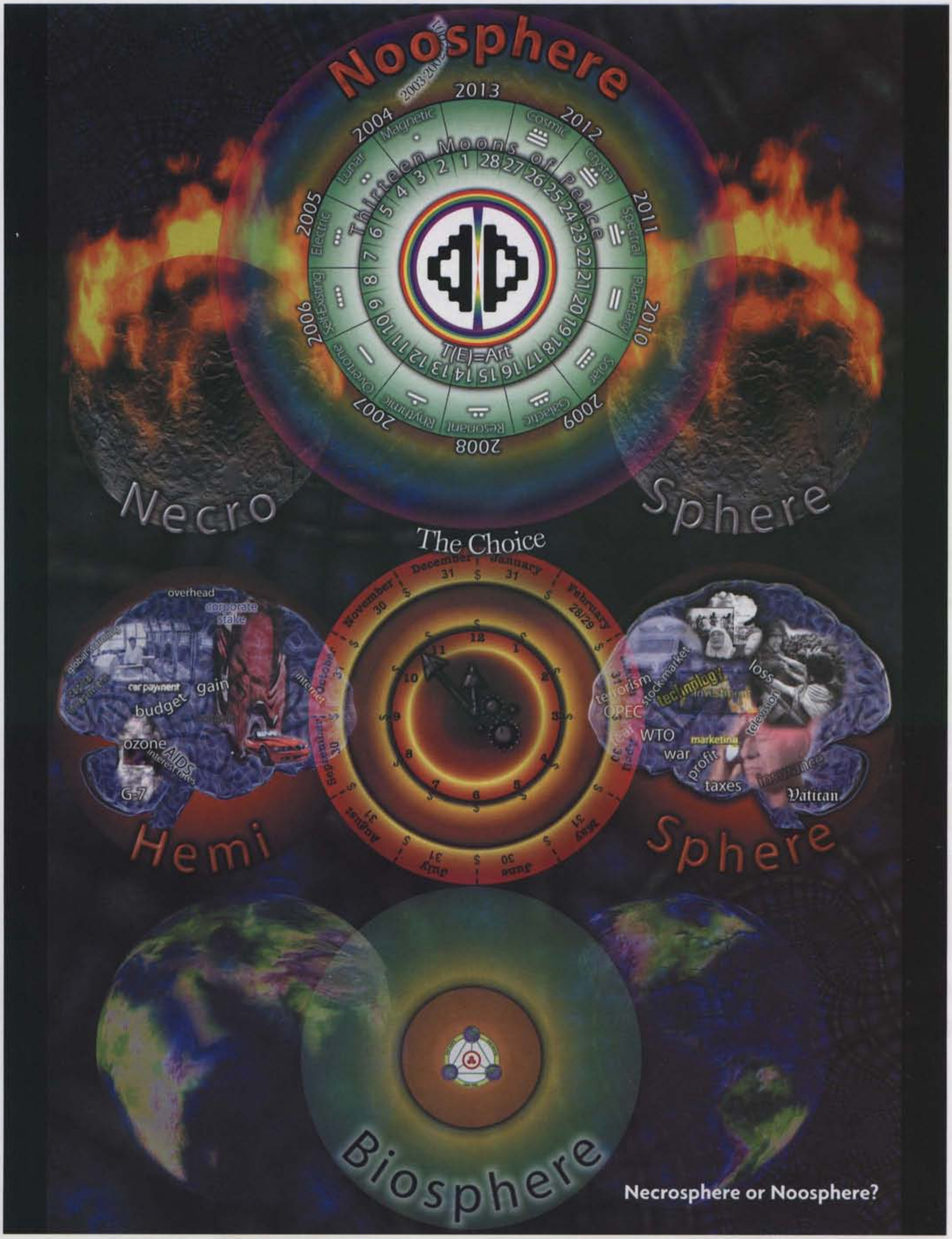
Noosphere
2003-2004
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
Magnetic
Lunar
Electric
Self-Existing
Overtone
Rhythmic
Resonant
Galactic
Solar
Planetary
Spectral
Crystal
Cosmic
Thirteen Moons of Peace
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
T(E)=Art
Necro
Sphere
The Choice
December 31
January 31
February 28/29
November 30
October 31
September 30
August 31
July 31
June 30
May 31
12 1 2 3 4 5 6 7 8 9 10 11
overhead
corporate stake
global warming
car payment
gain
budget
internet
ozone
AIDS
interest rates
G-7
terrorism
stock market
technology
investment
loss
OPEC
television
WTO
marketing
war
profit
taxes
insurance
Vatican
Hemi
Sphere
Biosphere

Necrosphere or Noosphere?

A Chronotopology of Transformation

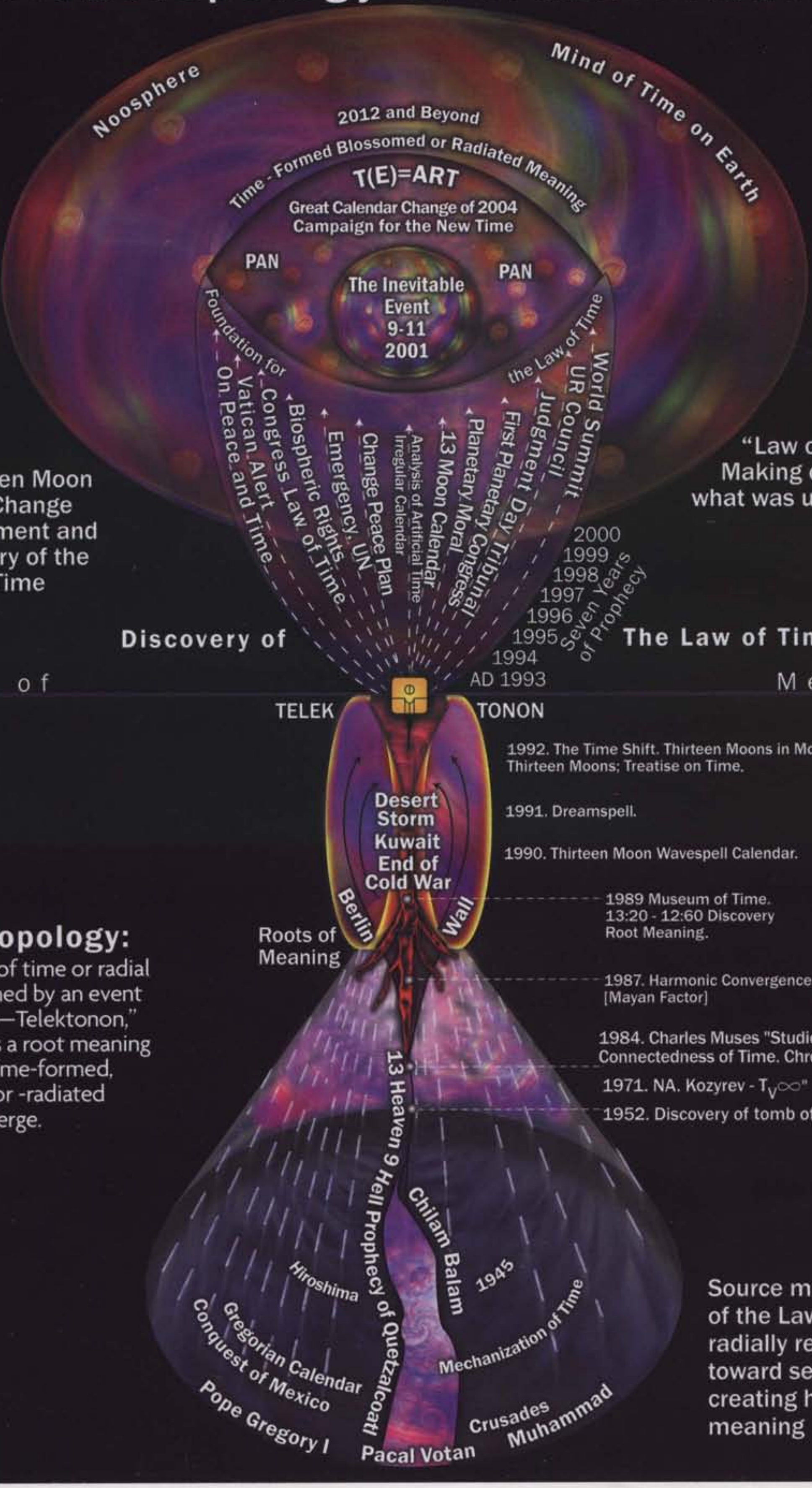
Noosphere
Mind of Time on Earth
2012 and Beyond
Time - Formed Blossomed or Radiated Meaning
T(E)=ART
Great Calendar Change of 2004
Campaign for the New Time
PAN
PAN
The Inevitable Event 9-11 2001
Foundation for the Law of Time
On Peace and Time
Vatican Alert
Congress Law of Time
Biospheric Rights
Emergency UN
Change Peace Plan
Analysis of Artificial Time Irregular Calendar
13 Moon Calendar
Planetary Moral Congress
First Planetary Day Tribunal
Judgment
UR Council
World Summit
2000
1999
1998
1997
1996
1995
1994
AD 1993
Seven Years of Prophecy
TELEK
TONON
Desert Storm Kuwait End of Cold War
Berlin
Wall
Roots of Meaning
13 Heaven 9 Hell Prophecy of Quetzalcoatl
Chilam Balam
1945
Hiroshima
Mechanization of Time
Gregorian Calendar
Conquest of Mexico
Pope Gregory I
Pacal Votan
Crusades
Muhammad

World Thirteen Moon
Calendar Change
Peace Movement and
The Discovery of the
Law of Time

"Law of Time.
Making conscious
what was unconscious"

Discovery of **The Law of Time**

Horizon of Meaning

1992. The Time Shift. Thirteen Moons in Motion; Turtle of Thirteen Moons; Treatise on Time.

1991. Dreamspell.

1990. Thirteen Moon Wavespell Calendar.

1989 Museum of Time.
13:20 - 12:60 Discovery
Root Meaning.

1987. Harmonic Convergence Root Source.
[Mayan Factor]

1984. Charles Muses "Studies in the Interactive Connectedness of Time. Chronotopology"

1971. NA. Kozyrev - $T_V\infty$"

1952. Discovery of tomb of Pacal Votan

Chronotopology:

A geography of time or radial time set defined by an event "Law of Time—Telektonon," which creates a root meaning from which time-formed, -blossomed, or -radiated meanings emerge.

Source meanings
of the Law of Time
radially resonated
toward seed point
creating horizon of
meaning

Paul J. Nahin

Time Machine Tales

The Science Fiction Adventures and Philosophical Puzzles of Time Travel

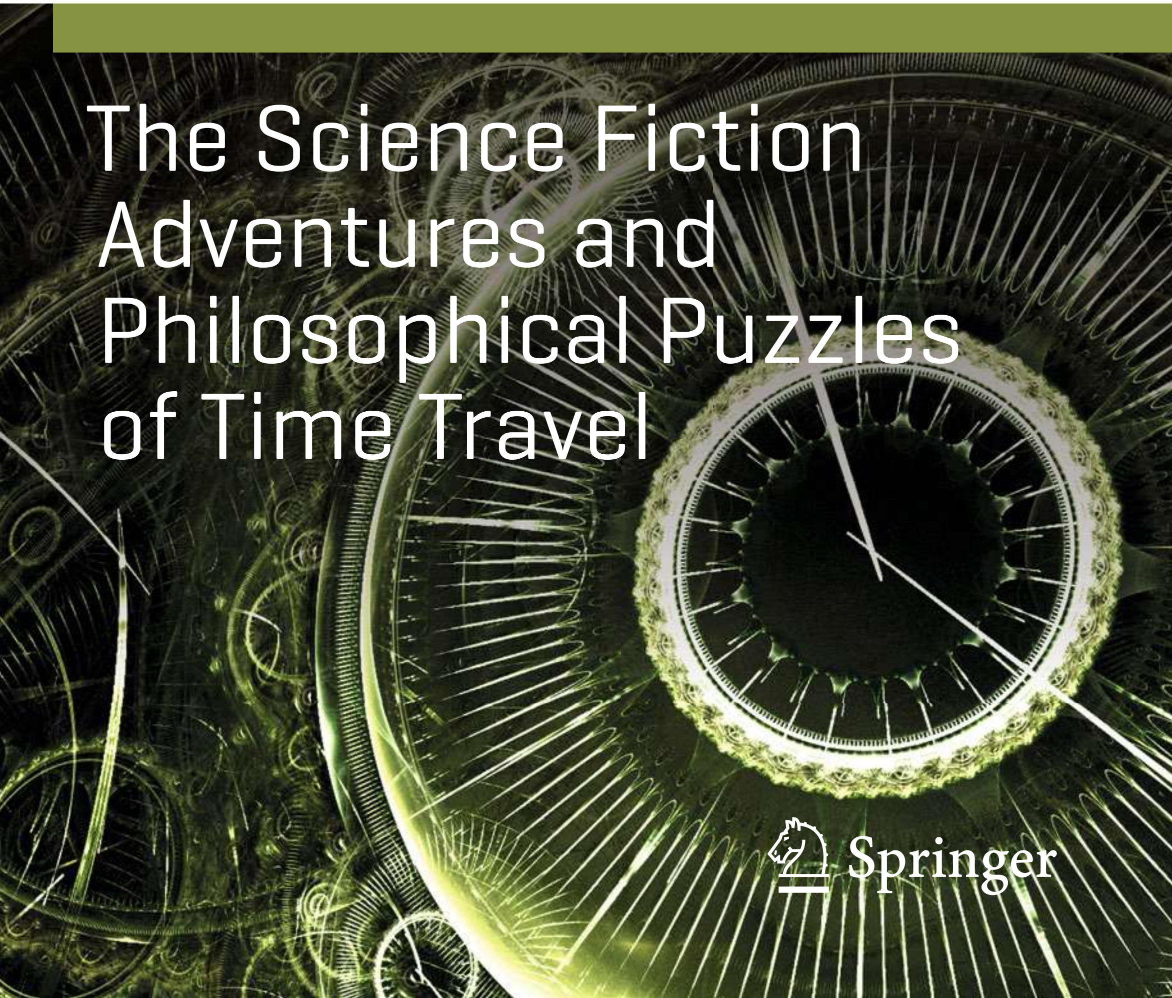

Springer

Science and Fiction

Editorial Board

Mark Alpert
Philip Ball
Gregory Benford
Michael Brotherton
Victor Callaghan
Amnon H. Eden
Nick Kanas
Geoffrey Landis
Rudi Rucker
Dirk Schulze-Makuch
Rüdiger Vaas
Ulrich Walter
Stephen Webb

More information about this series at http://www.springer.com/series/11657

Science and Fiction – A Springer Series

This collection of entertaining and thought-provoking books will appeal equally to science buffs, scientists and science-fiction fans. It was born out of the recognition that scientific discovery and the creation of plausible fictional scenarios are often two sides of the same coin. Each relies on an understanding of the way the world works, coupled with the imaginative ability to invent new or alternative explanations - and even other worlds. Authored by practicing scientists as well as writers of hard science fiction, these books explore and exploit the borderlands between accepted science and its fictional counterpart. Uncovering mutual influences, promoting fruitful interaction, narrating and analyzing fictional scenarios, together they serve as a reaction vessel for inspired new ideas in science, technology, and beyond.

Whether fiction, fact, or forever undecidable: the Springer Series "Science and Fiction" intends to go where no one has gone before!

Its largely non-technical books take several different approaches. Journey with their authors as they

- Indulge in science speculation – describing intriguing, plausible yet unproven ideas;
- Exploit science fiction for educational purposes and as a means of promoting critical thinking;
- Explore the interplay of science and science fiction – throughout the history of the genre and looking ahead;
- Delve into related topics including, but not limited to: science as a creative process, the limits of science, interplay of literature and knowledge;
- Tell fictional short stories built around well-defined scientific ideas, with a supplement summarizing the science underlying the plot.

Readers can look forward to a broad range of topics, as intriguing as they are important. Here just a few by way of illustration:

- Time travel, superluminal travel, wormholes, teleportation
- Extraterrestrial intelligence and alien civilizations
- Artificial intelligence, planetary brains, the universe as a computer, simulated worlds
- Non-anthropocentric viewpoints
- Synthetic biology, genetic engineering, developing nanotechnologies
- Eco/infrastructure/meteorite-impact disaster scenarios
- Future scenarios, transhumanism, posthumanism, intelligence explosion
- Virtual worlds, cyberspace dramas
- Consciousness and mind manipulation

Paul J. Nahin

Time Machine Tales

The Science Fiction Adventures and Philosophical Puzzles of Time Travel

Springer

Paul J. Nahin
Department of Electrical Engineering
University of New Hampshire
Durham, NH, USA

ISSN 2197-1188 ISSN 2197-1196 (electronic)
Science and Fiction
ISBN 978-3-319-48862-2 ISBN 978-3-319-48864-6 (eBook)
DOI 10.1007/978-3-319-48864-6

Library of Congress Control Number: 2016957723

© Springer International Publishing AG 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Also By Paul J. Nahin

Oliver Heaviside (1988, 2002), Johns Hopkins
Time Machines (1993, 1999), Springer
The Science of Radio (1996, 2001), Springer
An Imaginary Tale (1998, 2007, 2010), Princeton
Duelling Idiots (2000, 2002), Princeton
When Least Is Best (2004, 2007), Princeton
Dr. Euler's Fabulous Formula (2006, 2011), Princeton
Chases and Escapes (2007, 2012), Princeton
Digital Dice (2008, 2013), Princeton
Mrs. Perkins's Electric Quilt (2009), Princeton
Time Travel (1997, 2011), Johns Hopkins
Number-Crunching (2011), Princeton
The Logician and the Engineer (2013), Princeton
Will You Be Alive Ten Years From Now? (2014), Princeton
Holy Sci-Fi! (2014), Springer
Inside Interesting Integrals (2015), Springer
In Praise of Simple Physics (2016), Princeton

Frontispiece: The Pioneers of Time Travel

The scientific pioneers were Albert Einstein (1879–1955) and Kurt Gödel (1906–1978), good personal friends who are shown here in 1954 at the Institute for Advanced Study in Princeton, New Jersey, in a photo taken by Richard Arens. It was Einstein's 1916 general theory of relativity (theory of gravity) that Gödel used as the basis for his 1949 paper that was the first to show that the general theory does **not** forbid time travel into the past.

The literary pioneer of time travel was of course Herbert George Wells (1866–1946), who is shown here as a college freshman cut-up around 1885. The photograph was taken as a prank by an unknown friend while Wells was a student in a biology course given by Thomas Huxley, at the Normal School of Science in

South Kensington (a branch of the University of London). A far too thin and impoverished Wells was then still a teenager, and *The Time Machine* lay a distant 10 years in the future.

Einstein/Gödel photograph courtesy of the American Institute of Physics Emilio Segré Visual Archives of the AIP Niels Bohr Library. Wells photograph courtesy of the rare Books and Special Collections Department of the Library of the University of Illinois at Urbana-Champaign.

A Note on the Story Citations and Science Fiction History

> "You will find it a very good practice always to verify your references, sir."
> —advice given in 1847 to a young scholar by Martin Joseph Routh, President of Magdalen College, Oxford

Most of the pulp science fiction stories I've cited in this book, in their original form as ink on paper, have long since vanished from our region of spacetime and exist today only (alas) on microfilm reels in scholarly vaults. I am especially indebted to Texas A & M, the Claremont Colleges, the California State Universities at Northridge and Fullerton, Mount Holyoke College, the New York City Public

Library, and the University of Delaware, for giving me access via Inter-Library Loan (through my home institution, the University of New Hampshire) to their extensive archives of ancient science fiction magazines.

A number of the really good stories *have* been anthologized, however, and so are still readily available today in book form. In essentially all cases, though, for historical reasons, I've given the original publication information (magazine and date). You can find which of the stories cited are available in one or more anthology reprints by going to an immensely useful, searchable database on the Web, at: http://www.isfdb.org, and I gratefully thank all those in the science fiction community responsible for creating and maintaining that database.

The following two books by science fiction historian Sam Moskowitz (1920–1997), who lived through what Isaac Asimov called the 'Golden Age of [magazine] Science Fiction,' may be difficult to find today but, if you are interested in the early history of *magazine* science fiction (beyond simply the subgenre of time travel), the hunt for them will be well worth your time:

Science Fiction by Gaslight: a history and anthology of science fiction in the popular magazines, 1891–1911 (World Publishing Company 1968);
Under the Moons of Mars: a history and anthology of "The Scientific Romance" in the Munsey Magazines, 1912–1920 (Holt, Rinehart and Winston 1970).

Some First Words

> Is time travel in principle (never mind the difficulties) a possibility? It has received some thought in the past and deserves some more.
>
> —David Park, in his 1980 book *The Image of Eternity*

> He used to have quite a reputation, but the last couple of years he's been working on time . . . You know, time travel, that sort of rot. An A-1 crackpot.
>
> —a character (discussing a colleague) disagrees with Park, in Mack Reynolds' "Advice from Tomorrow," *Science Fiction Quarterly*, August 1953

In 1993 the first edition of my book *Time Machines* was published by the Press of the American Institute of Physics. In 1999, after Springer acquired AIP Press, the second edition of that book appeared. So, is this the third edition? Well, yes *and* no. It *is* because large chunks of the 1999 edition are still here, along with new discussions of the advances by physicists and philosophers that have appeared in the intervening 18 years. The prime example of that centers on the time travel paradoxes. Those discussions contain mostly what is in the second edition, but they have also been brought up to date with the latest thinking on the paradoxes, by physicists and philosophers.

And yet this book is *not quite* the third edition because the emphasis is now on the philosophical and on science fiction, rather than on physics as it was when written for AIP Press. In that spirit there are, for example, no Tech Notes filled with algebra, integrals, and differential equations, as there are in the first and second editions of *Time Machines*. That's because I wish to avoid having this book seem to be simply a long physics treatise. I have, in fact, some sympathy with the following views, expressed by two philosophers:

"There is one metaphor in the physicist's account of space-time which one would expect *anyone* to recognize as such, for metaphor is here strained far beyond the breaking point, i.e., when it is said that time is 'at right angles to each of the other three dimensions.' Can anyone really attach any meaning to this—except as a recipe for drawing diagrams?"[1]

and

"This is from the outset a study in descriptive metaphysics. In consequence, I shall have nothing to say about twice-differentiable Lorentzian manifolds, Minkowski diagrams, world-lines, time-like separations, space-time worms [a 'thick' world-line], or temporal parts."[2]

I don't *completely* endorse these sentiments, however, and so please understand that I am not denying the ultimate importance of *physics* when it comes to achieving a deep understanding of time travel. To quote yet another philosopher,

"Arm chair reflections on the concept of causation [are] not going to yield new insights. The grandfather paradox is simply a way of pointing to the fact that if the usual laws of physics are supposed to hold true in a chronology violating space-time, then consistency constraints emerge. *[To understand these constraints] involves solving problems in physics, not armchair philosophical reflections* [my emphasis]."[3]

I could not agree more. So, in *Time Machine Tales* you *will* find some physics. In support of time travel to the future (and in how to make a wormhole time machine for travel into the past), for example, I'll show you a high school level derivation of the famous time dilation formula from special relativity. There are some spacetime diagrams, some simple algebraic manipulations, and here and there just a touch of freshman calculus; even the metric tensor gets a few words, too. But it is, admittedly, pretty light-weight stuff.

So, while certainly saluting the premier position of physics, *Time Machine Tales* is not a scholarly, in-depth treatment of time travel physics. Rather, it is an examination of how science fiction writers (and many philosophers, too) have viewed time travel. (Even in the physics discussions, science fiction will regularly appear.) Those views, by their very nature, are far more romantic than are those of hardcore theoretical physicists. History has shown, of course, that the results of the work of theoretical physicists may, in the end, prove to actually be far more astonishing than anything fiction writers cook-up—and if there is any scientific subject for which that may again prove to be true it's time travel—but for us, here, it will be the fiction writer who has center stage.

The philosophers will be only slightly less important in this book. While much of the early philosophical literature on time travel and backwards causation reads like imaginative fairy tales spun out of vacuous vapors (more on this soon), many modern philosophers have shown themselves to be quite sophisticated. What they have written deserves serious consideration by anyone interested in time travel, and that includes physicists. However, while the time travel interests of philosophers and physicists have a lot of overlap, those interests are *not* in total agreement. For example, while both groups talk of the grandfather paradox, the philosophers worry in particular about motivation (*why* the murderous mission?), while physicists have *never* to my knowledge asked themselves that question[4] (other than to figure out how to avoid it!). After all, philosophers talk of flesh-and-blood humans as time travelers, while the physicists send only billiard balls (with no personal identities or memories) on time trips into the past for the *expressed purpose* of avoiding the messy human issues of 'motivation' and free will. This approach by physicists isn't because they are cold and emotionless. It is a useful strategy because, if it can be shown that a mere billiard ball can travel into the past then, as one *philosopher* pointed out long ago, "It is implausible that it should be possible for some physical systems to travel back in time, and not others. Thus, if we suppose that simple objects can time-travel . . . then we must suppose that more complicated systems, e.g., human beings, can also time-travel."[5]

---

[1] C. W. K. Mundle, "The Space-Time World," *Mind*, April 1967, pp. 264–269.

[2] J. F. Rosenberg, "One Way of Understanding Time," *Philosophia*, October 1972, pp. 283–301.

[3] John Earman, "Recent Work on Time Travel," in *Time's Arrows Today: recent physical and philosophical work on the direction of time* (Steven F. Savitt, editor), Cambridge University Press 1995, pp. 268–310. We'll discuss the idea of *consistency constraints* in some detail later in the book. Earman is Professor Emeritus of History and Philosophy of Science at the University of Pittsburgh.

For the most part, philosophers and physicists have worked at the extreme, opposite points of the time travel spectrum. Much better, I think, would be to adopt the following, more balanced position advocated recently: "The study of time machines is a good opportunity for forging a partnership between philosophy and physics. Of course, philosophers have to recognize that in this particular instance the partnership is necessarily an unequal one since the mathematical physicists have to do the heavy lifting. But it seems clear that a little more cooperation with philosophers of science in attending to the analysis of what it takes to be a time machine could have led to some helpful clarifications in the physics literature."[6]

In the past, philosophers gained a reputation for being just a bit too 'unconstrained by the facts' for scientific tastes—as the English mathematician Augustus De Morgan (1806–1871) wrote in an 1842 letter, "There are no writers who give us so much *must* with so little *why*, as the metaphysicians"[7]—but I do think today's physicists would do well to reexamine that harsh opinion.

Philosophers of the 'old school' may look askance at a non-philosopher (me!) leveling criticism at them, and so let me step aside and quote from a member of the

---

[4] Nicholas J. J. Smith, "Why Would Time Travelers Try to Kill Their Younger Selves?" *The Monist*, July 2005, pp. 388–395. As Smith writes, "[Motivation] does not impact upon the possibility, or even the likelihood of backwards time travel. Yet it is deeply puzzling, and we will have no idea what time travel would actually be *like* until we explore it." See also Peter B. M. Vranas, "Can I Kill My Younger Self? Time Travel and the Retrosuicide Paradox," *Pacific Philosophical Quarterly*, December 2009, pp. 520–534.

[5] P. Horwich, "On Some Alleged Paradoxes of Time Travel," *Journal of Philosophy*, August 1975, pp. 432–444.

[6] John Earman, Christopher Smeenk, and Christian Wüthrich, "Do the Laws of Physics Forbid the Operation of Time Machines?," *Synthese*, July 2009, pp. 91–124.

[7] D. J. Cohen, *Equations from God: pure mathematics and Victorian faith*, The Johns Hopkins University Press 2007, p. 119.

'modern school' of philosophical thought: "Space-time is the basic spatiotemporal entity. Many philosophers have mouthed this truth, but few have swallowed it, and very few have digested it ... An appreciation of this truth is crucial to what is commonly referred to as the philosophy of space and time ... In large measure the lack of progress in this area can be traced to the fact that philosophers have not taken seriously the corollary that talk about space and time is really talk about the spatial and temporal aspects of spacetime."[8] This is a polite way of telling philosophers that they had better learn some physics!

What provoked those harsh words was that 'modern' philosopher's perception that 'old school' philosophers were not talking science when they wrote of space and time, but rather were in the business of telling each other irrelevant stories and myths, a curious philosophical approach involving the 'telling of tales' that reached its peak in the early and mid-1960s. Spacetime story telling seems to have started with a paper by the Oxford philosopher Anthony Quinton (1925–2010), who argued[9] that although there can be multiple, disjointed spaces, there can only be a single time that is the same for everyone, everywhere. The issue is *not* the truth or not of that assertion (Newton believed it, modern physicists don't), but rather Quinton's technique for arriving at it: myth construction.

Myth construction strikes those trained in the technical sciences as, while perhaps interesting—even physicists, after all, can enjoy a good fairy tale now and then—something quaint and totally beside the point. In his paper Quinton tells a fairy tale about how he thinks someone can live continuously in time and yet, via dreaming, be in two different spatial worlds; when awake he is in one world, while when the person is asleep he is in the other. Quinton argues that this multispatial myth is plausible, but that a search for an analogous multitemporal myth is doomed. This prompted a reply[10] from another 'old school' philosopher who rebutted Quinton with an even more outlandish counter-myth involving "the warring tribes of Okku and Bokku"!

It was this back-and-forth spinning of hypothetical tales that caused the 'modern' philosopher to write in his paper (note 30) that "the procedure for arriving at answers to these questions [about space and time] adopted by Quinton and most other ['old school' philosophers] is, to say the least, a curious one: a story is told about a mythical land—usually called something like the land of Okkus-Bokkus [which is now seen to an outrageous pun]—and then we are asked what we would say if confronted by experiences like those of the Okkus-Bokkusians. As often happens with such a question, people have said all sorts of things, not all of which are interesting or enlightening."

Another modern philosopher was even less gentle in his rejection of the fairy tale approach to spacetime physics: "Quinton [and others of a similar approach invite us] to say what we should think in certain strange circumstances which they describe within common-sense language [as opposed to scientific terminology]. I must say that if I found myself in the circumstances which they describe I just would not know what to think. Probably I should simply conclude that I had gone mad . . . It looks as though these writers are inviting us to consider *what we should say if we knew no science* [my emphasis].”[11]

---

[8] J. Earman, "Space-Time or How to Solve Philosophical Problems and Dissolve Philosophical Muddles Without Really Trying," *Journal of Philosophy*, May 1970, pp. 259–276.

[9] A. Quinton, "Spaces and Times," *Philosophy*, April 1962, pp. 130–147.

[10] R. G. Swinburne, "Times," *Analysis*, June 1965, pp. 185–191.

Even before the modern philosopher (note 30) wrote in 1970 to complain about myth-making, another had already done so: “Whenever a human being produces an argument which opens ‘Suppose I had 23 senses . . .,’ ‘Suppose I were God . . .,’ ‘Suppose I experienced objects extended in four spatial dimensions . . .,’ we can protest that the argument is worthless. For in supposing that he has transcended our human point of view, he has also transcended the limits of our understanding.”[12] As this author concluded his very funny paper, such opening sentences are the signatures of myths from “The Philosopher’s Fairy Tale Book.”

The strained relationship between myth-making philosophers and physicists, especially concerning time travel, has a historically interesting antecedent in the 1920s negative reaction among many over Einstein’s theories of relativity (the very theories that give apparent life to time travel). To illustrate my point, consider the October 1913 letter Oskar Kraus (1872–1942), a philosophy professor at the German University in Prague, sent to Ernst Gehrcke (1878–1960), a physics professor at the Reich Institute of Physics and Technology in Berlin. Both men were opponents of Einstein but, as Kraus wrote in his letter, it was only Gehricke among the physicists he considered to be sympathetic to him: “[I] would not know . . . anyone else but you who as a specialist would not reject the intervention of a philosopher from the start.”[13]

So, I think Earman’s proposal a sound one, an echo in fact of similar words that the physicist Kip Thorne wrote (in the Foreword to the second edition of *Time Machines*) concerning science fiction writers: “Smart physicists seek insight everywhere, including from clever science fiction writers who long ago began probing seriously the logical consequences that would ensue if the laws of physics permitted time travel.”[14]

To emphasize this new, combined, diversified focus (but also to retain some connection with my earlier books) is the reason I have altered the title, just a bit. In addition, each chapter now concludes with several open-ended questions, suitable for motivating either classroom discussions or more extensive essay responses.

---

[11] J. J. C. Smart, “The Unity of Space-Time: Mathematics Versus Myth Making,” *Australasian Journal of Philosophy*, (no. 2) 1967, 214–217.

[12] M. Hollis, “Times and Spaces,” *Mind*, October 1967, pp. 524–536. Hollis ends by saying he is prepared to accept the failure of his paper to convince many of his colleges to change their ways, and he is waiting for one of them to write a paper opening with “Twice upon a time in another space no distance in any direction from here . . .”!

[13] Quoted from the Introduction to Milena Wazeck, *Einstein’s Opponents: the public controversy about the theory of relativity in the 1920s*, Cambridge 2014 (published in German in 2009).

[14] Thorne is Professor Emeritus of Physics at the California Institute of Technology.

Teachers, in particular, may find this a useful feature if using the book in an academic setting. The book ends with reprints of two of my own published time travel stories (one from *Analog* and the other from *Omni*), with each serving as an illustration of technical issues raised in the book. From my own teaching of an undergraduate honors class in time travel at the University of New Hampshire, I think the assigning of story writing to be an excellent tool for teachers to use. I found reading student stories to be a lot of fun, and students may well surprise teachers with innovative ideas.

Now that I've mentioned story writing, let me say something about the heavy presence of time travel science fiction stories in this book, the majority of which originated in the often maligned pulp magazines of the 1920s through the 1950s. 'Pulp' has long been burdened with a bad literary reputation. As the editor of one anthology of pulp fiction bluntly put it, "Pulp equated with rubbish. Crap of the basest nature."[15] Part of the reason for that was cosmetic; as I wrote in an earlier book, "The term *pulp* came from the use of inexpensive wood-pulp—you could *feel* the lumpy wood chips in each ragged, untrimmed page—to make paper that was far too crummy for the use by any publisher of 'words meant to last.' Such paper quickly yellowed, turned brittle, and finally, amid billowing clouds of bits and pieces, entered into eternal oblivion. Think of the paper used in your newspaper before its final contribution to civilization in the bottom of your cat's litter box; pulp was worse."[16]

And then a little later, in the same book, "The stories in *Amazing* [*Stories* magazine] were 'read it in the morning, forget it by dinnertime' adventure fiction, the stuff you'd put inside a newspaper if on a crowded train or bus so fellow passengers wouldn't know what a low-grade mind you had. The transient nature of pulp fiction was independent of its literary quality, as the cheap acid-based paper that the stories were printed on began to oxidize and literally burn-up as soon as it rolled off the press. In the introductory essay to a 1950 collection of pulp-detective Philip Marlowe stories (*Trouble Is My Business*), mystery writer Raymond Chandler commented on this when he wrote 'pulp fiction never dreamed of posterity.' Pulp fiction was synonymous with trash fiction, and the nature of much of early pulp SF has been aptly described as 'scientific pornography for the mechanically minded,' and 'writing which drooled over descriptions of technology.'"

When publisher Hugo Gernsback (1884–1967) brought out the first issue of *Amazing Stories* in April 1926, it was the first pulp devoted totally to science fiction. With its masthead motto of "Extravagant Fiction Today—Cold Fact Tomorrow," and with the illustration on the contents page of each issue showing a muscular Jules Verne bursting from his grave in the heroic, up-up-and-away pose made famous years later by Superman, there could be no doubt as to what kind of fiction the reader would find under the dramatic, multi-colored cover art. It was fiction populated with mad scientists, and half-naked woman about to be ravished by alien invaders from outer space; all in all, stuff of interest only to teenage boys and imbecilic adults.[17]

---

[15] Maxim Jakubowski, *The Mammoth Book of Pulp Fiction*, Carroll & Graf 1996.

[16] P. J. Nahin, *Holy Sci-Fi!: where science fiction and religion intersect*, Springer 2014.

How else, after all, to explain the publication of one tale[18] that was given the following heart-stopping editorial introduction: “Professor Lambert deliberately ventures into a Vibrational Dimension to join his fiancée in its magnetic torture-fields”? In defense of many of the readers of early pulp science fiction, however, not all were attracted by such nonsense. Just 2 months later (June 1931) one reader wrote to the same magazine to complain of masculine heroes saving weeping women from ungodly horrors: “Just why do you permit your Authors to inject messy love affairs into otherwise excellent imaginative fiction? Just stop and think. Our young hero-scientist builds himself a space flyer, steps out into the great void, conquers a thousand and one perils on his voyage and amidst our silent cheers lands on some far distant planet. Then what does he do? He falls in love with a maiden—or it’s usually a princess—of the planet to which the Reader has followed him, eagerly awaiting and hoping to share each new thrill attached to his gigantic flight. But after that it becomes merely a hopeless, doddering love affair ending by his returning to Earth with his fair one by his side. Can you grasp that—a one-armed driver of a space-flyer! . . . We buy A.S. for the thrill of being changed in size, in time, in dimension . . . not to read of love . . . I wish . . . for plain, cold scientific stories sans the fair sex.”

Here’s another example, this one of the sort of tale that gave an aroma of the sophomoric to ‘golden age’ time travel science fiction. It was a story of a young man of the far future, with access to a time machine, who wants to see a dinosaur before he dies. So back he travels, back, back, until he at last finds himself in a “subterranean cave, dark and foul-smelling.” At first he is puzzled (did dinosaurs live underground?), but then suddenly he hears a thundering roar and sees a huge black shape in the gloom. There can be no doubt now; it *is* a dinosaur, and he sees its red, gleaming eyes just as it crushes him into a pancake. But that’s okay; he saw a dinosaur before he died. Then comes the dénouement. He hadn’t really gone back quite as far as the Jurassic period, but only to the twentieth century, where he has been run down by the local express train in a subway tunnel![19]

---

[17] This was particularly thought to be the case for readers of the romance pulps, written for young women in the 1930s and 1940s (a separate and distinct audience from that of the science fiction pulps). As one commentator wrote on that genre, the heroes and heroines of such tales often displayed the “mental equipment of a banana split,” with the implication that the same might be said of the readers, themselves. (See Margaret MacMullen, “Pulps and Confessions,” *Harper’s Monthly Magazine*, June 1937.) I don’t think, however, that this particular complaint generally applied to the pulp science fiction readership. I’ll have much more to say about Gernsback and early pulp science fiction speculations concerning time travel, in Chap. 4.

[18] T. Curry, “Hell’s Dimension,” *Astounding Stories*, April 1931.

[19] R. G. Thompson, “The Brontosaurus,” *Stirring Science Fiction*, April 1941. In the editors of *Stirring*’s defense, notice the month: maybe this story was *meant* to be a joke. If so, it was an admirable success.

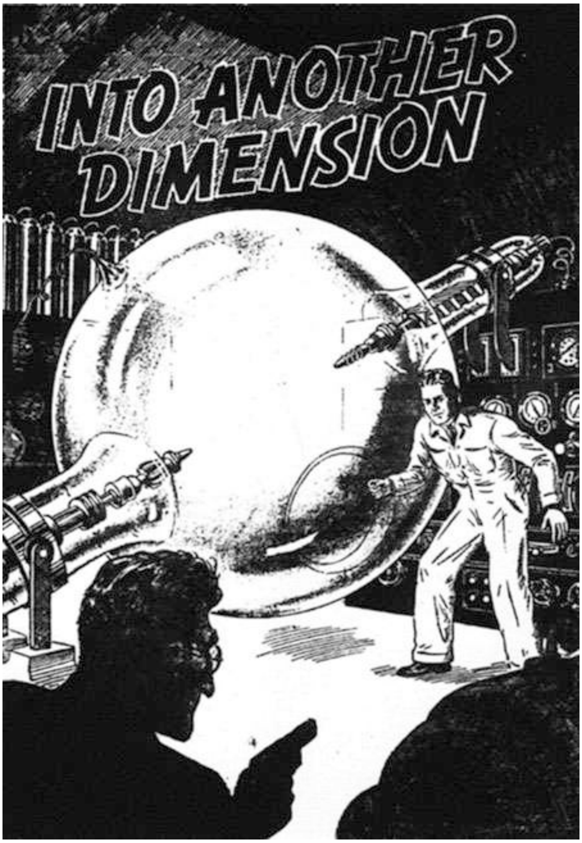

Vibrating into new dimensions was, apparently, a popular idea in 1930s pulp science fiction. This 'super science' gadget operated by vibrating an object faster than light, whereupon the Lorentz-FitzGerald contraction formula (see Chap. 3) predicts an *imaginary* size for the object—which means (so we are told) that the object has entered "another plane of existence." The inventor (the fellow with the gun) is inviting his grim-faced assistant to give the gadget a try. The original caption reads "Get into that vibrator! Get in, I say!"

Illustration for "Into Another Dimension" by Maurice Duclos, *Fantastic Adventures* November 1939 (art by Kenneth J. Reeve), © 1939 by Ziff-Davis Publishing Co., reprinted by arrangement with Forrest J. Ackerman, Holding Agent, 2495 Glendower Ave., Hollywood, CA 90027

Today, however, the need to apologize for science fiction tales about time travel isn't quite so necessary. Now and then, in fact, you'll even find one of the better pulp stories cited in highly mathematical papers on time machines in the *Physical Review D*, one of the most important scholarly physics journals. Even those physicists and philosophers who mostly ignore science fiction—except perhaps to make slightly condescending remarks—would, if honest, admit that their early teenage interest in time travel was sparked by reading a really good science fiction story, and not by working their way through a physics textbook. Yes, when the physics eventually came later, it was very good—but the science fiction came *first*, and it was pretty good, too.[20] It's in a 1937(!) tale, for example, that we find the claim for consistency around a closed loop in time, *decades* ahead of the physicists and philosophers.[21] And when you get to the final section of Chap. 4, I think you'll find it difficult to believe that Everett's many-worlds interpretation of quantum mechanics, dating from the late 1950s (which avoids the standard paradoxes of time travel) wasn't inspired by some youthful reading of science fiction from the 1930s and 1940s.

In a number of places in this book you'll find my comments on how science fiction has occasionally anticipated physicists on the subject of time machines and time travel. This is *not* to be interpreted as some sort of 'gotcha' in favor of science fiction. Far from it. When push comes to shove, physics *always* wins. This situation was specifically addressed by Joe Haldeman, in an afterword to his 2007 novel *The Accidental Time Machine*. There he wrote, about when he started in 1971 to write his earlier, now classic novel *The Forever War*, "I needed a way to get soldiers from star to star within a human lifetime, without doing too much violence to special and general relativity. *I waved my arms around really hard* [my emphasis] and came up with the 'collapsar jump'—at the time, collapsar was an alternate term for 'black hole,' though I was unaware of the latter term [because John Wheeler had invented it only 4 years before, as discussed in Chap. 1 and note 106]." And then Haldeman admitted "It's a truism of science fiction that if you predict enough things, a few of them are going to come true. . . . What I think it actually demonstrates is that *if you wave your arms around hard enough* [my emphasis], sometimes you can fly."

Now, there is one feature common to all books on time travel to the past (which is the central topic treated here, of course) that I would like to clearly state. It's obviously a subject of vast interest to physicists, and yet it offers (as far as I know) *absolutely no hope of suggesting even a single experiment for study*. (As far as I know, nobody is building a time machine in their basement.) A suggestion *has* been made that it may be possible to detect, in the *present*, the effects of the *future* operation of "man-made time machines, which could be of a size traversable by humans," that is, machines with a 1-m spatial extent offering a one second trip into the past.[22] With the best technology available today, however, the calculated effects on the proposed two-particle scattering experiment are orders of magnitude too small to measure.

---

[20] The view expressed by Vladimir Voinovich's time traveler in his 1986 novel *Moscow 2042* (Science fiction . . . is not literature, but tomfoolery like the electronic games that induce mass idiocy.) is, I think, wrong. For an interesting presentation on the role of science fiction in exciting an interest in science among youngsters, see the paper by Frederik Pohl (1919–2013), "Science Fiction: the stepchild of science," *Technology Review*, October 1994, pp. 57–61. In this essay Pohl, a well-known writer of science fiction and editor of *Galaxy Science Fiction* and *If* magazines, writes "Science fiction is [the ultimate protection] against future shock . . . if you read enough of it, nothing will take you entirely by surprise." Not even time travel.

[21] P. S. Miller, "The Sands of Time," *Astounding Stories*, April 1937.

[22] S. Rosenberg, "Testing Causality on Spacetimes with Closed Timelike Curves," *Physical Review D*, March 15, 1998, pp. 3365–3377.

This situation is really unprecedented in the history of science.[23] To cynics, it may seem to be a bit like writing learned papers on the thermodynamics of fire-breathing dragons (which, like other mythological entities—and time machines, too—have yet to be seen)! This one fact has opened the doors—and has kept them open for decades—for philosophers and science fiction writers, who can endlessly debate back and forth on all aspects of time travel to the past with nary a single experimental fact to complicate their lives. For physicists the situation is naturally frustrating, but for philosophers and science fiction writers it's a dream come true. This isn't to say it's *all* basically theological in nature. Both the physicists and the philosophers *have* written many fascinating papers and books and, of course, so have science fiction writers. Mathematical physics *has* been advanced.

Still, despite all of the theoretical work done in the last 30 years, work that has made it reasonable to seriously talk of 'time travel' and 'time machines,' I suspect many would nonetheless agree with these words from more than 75 years ago: "Of all the fantastic ideas that belong to science fiction, the most remarkable—and, perhaps, the most fascinating—is that of time travel . . . Indeed, so fantastic a notion does it seem, and so many apparently obvious absurdities and bewildering paradoxes does it present, that some of the most imaginative students of science refuse to consider it as a practical proposition."[24] For some, time travel is an even more unlikely possibility than (as declared by Robert Lewis Stevenson) is the "welding of ice and iron." Not all physicists and philosophers view the time travel/paradox arguments as convincing, however. Provocative, yes, of course, but many are not yet prepared to write 'signed, sealed, and delivered' at the end.

So, keep reading and I think you'll discover why there *are* those who are not so quick to dismiss the possibility of following the fantastic world line of H. G. Wells' intrepid Time Traveller[25] into the future. And, just maybe, into the distant past, too.

---

[23] Perhaps, however, I am too hasty. More recent theoretical calculations suggest that wormholes connecting our universe with other universes would, after converting into time machines, have characteristic thermal signatures. See P. F. González-Díaz, "Thermal Properties of Time Machines," *Physical Review D*, 2012, pp. 105026-1 to -7 which, however, concludes that a search for such signatures would be "quite difficult [with the] instruments available."

[24] I. O. Evans, "Can We Conquer Time?" *Tales of Wonder*, Summer 1940.

[25] The Time Traveller is never named in Wells' 1895 novel *The Time Machine*. An earlier (1888) attempt at a time machine story, with the awful title *The Chronic Argonauts* (the "chronic" was apparently inspired by the word *chronology*), so embarrassed Wells that he later called it "imitative puerile stuff," "clumsily invented, and loaded with irrelevant sham significance," and "inept," and so he hunted down and destroyed every copy of it that he could find. You can find *The Chronic Argonauts* reprinted in *The Definitive Time Machine* (H. M. Ceduld, editor), Indiana University Press 1987. The hero in that work *was* named: Dr. Moses Nebogipfel. There is one passage in *The Time Machine* that does tantalize; as the Time Traveller explores a museum of "ancient" artifacts in the Palace of Green Porcelain (they are, of course, artifacts of our *future*) he reveals that "yielding to an irresistible impulse, I wrote my name upon the nose of a steatite monster from South America that particularly took my fancy." Thus, the Traveller *has* given his name, but his signature exists only in the future, in a museum of the past that is yet to be built.

For Further Discussion

For time travel to the past to make any sense, the past must in some sense ‘still be there.’ This is a concept that we’ll find later in the book to have significant support in relativistic physics, but for now let’s limit ourselves to a purely romantic view. As an example of this, consider this passage by Canadian writer Grant Allen (1848–1899), from the Introduction to his 1895 time travel novel *The British Barbarians*: “I am writing in my study on the heather-clad hill-top. When I raise my eye from my sheet of foolscap, it falls upon miles and miles of broad open moorland. My window looks out upon unsullied nature. Everything around is fresh and pure and wholesome . . . But away below in the valley, as night draws on, a lurid glare reddens the north-eastern horizon. It marks the spot where the great wen of London heaves and festers.” I personally find it quite tempting to imagine Allen somehow still there in his study of 1895, and of heaving and festering late-Victorian London, too, with H. G. Wells himself in the middle of it, still reading the first rave reviews of *The Time Machine*. In Wells’ novel The Time Traveller journeys into the far future, while in Allen’s work the protagonist is a twenty-fifth century anthropologist who has traveled back to the past of the late nineteenth century to study the ‘British barbarians.’ Read Allen’s novel (it’s available on the Internet, for free, as a Project Gutenberg book) and comment on the significance of its appearance at virtually the same time as Wells’ great work. Why do you think Wells’ novel is remembered, and Allen’s is not?

In the opening paragraph of his paper “The Conundrum of Time Travel” (*Croatian Journal of Philosophy*, No. 37, 2013, pp. 81–92), Anguel Stefanov writes “Needless to say . . . the problems concerning time travel are being still tackled by science fiction only, but resolved by science proper neither theoretically, nor practically.” Do you think this is correct?

Eventually every genre of writing becomes the target for parody, in which the *form* of the genre serves as the framework for what (it is hoped) is a humorous mockery. The most famous example of this, perhaps, is the annual

(continued)

Edward Bulwer-Lytton contest in writing a take-off on the long-winded opening line of the 1830 novel *Paul Clifford*, by Bulwer-Lytton (1803–1873). That opening line *is* a wonder (a masterpiece of purple prose): "It was a dark and stormy night; the rain fell in torrents—except at occasional intervals, when it was checked by a violent gust of wind which swept up the streets (for it is in London that our scene lies), rattling along the housetops, and fiercely agitating the scanty flame of the lamps that struggled against the darkness." Here's a recent (from the 2015 contest) spoof: "The Contessa's heart was pounding hard and fast, like an out-of-balance clothes washer, which can get that way if you mix jeans with a lot of light things, though the new ones have some sensor thing to counteract that or shut off, but the Contessa's heart didn't have anything like that, so she had to sit down and tell Don Rolando to keep his hands to himself for a while." Science fiction isn't immune to such fun, and a good example of that can be found in the September 14, 2015, issue of *The New Yorker*, which has (on p. 50) "Eight Short Science-Fiction Stories" by Paul Simms. Here's the one I laughed hardest at: "The Gene-Splicers had tinkered with the DNA, producing a race of warriors who craved just two things: the thrill of battle and the taste of their own feet. They hungered for battle. They literally ate their own feet. None survived to reproduce, and within a few short years they were all gone. The Gene-Splicers chalked it up to experience, and decided to try harder the next time." That, and the other seven spoofs by Simms, cut across a wide swath of science fiction, but one theme noticeably absent was that of time travel. Try your hand at writing a *short* (fewer than 500 words) time travel spoof, and be prepared to read it aloud to an audience of your peers.

The tale "Through the Dragon Glass" by Abraham Merritt (1884–1943) appeared in the early pulp magazine *All-Story Weekly* of November 24, 1917. It described the discovery of a passage through an ancient Chinese mirror into an alternate world. One *might* think of this as an early conception of a wormhole, but more likely it may remind you mostly of Lewis Carroll's *Through the Looking Glass*. More interesting for us, in this book, is a story written 75 years ago that describes a gadget connecting two regions of spacetime, with a time shift of a week between the two regions. (See "Time Locker" by Lewis Padgett, in the January 1943 issue of *Astounding Science Fiction*.) The gadget falls into the hands of a crooked lawyer who, not understanding what he has, ends up accidently killing himself. As the story ends, the inventor of the gadget ruefully muses to himself that the lawyer

(continued)

"must have been the only guy who ever reached into the middle of next week—and killed himself!" The gadget is, in everything but name, a wormhole time machine. Speculate on how such a spacetime structure could appear in a *science fiction magazine*(!) decades before there was any discussion of such a possibility in the physics world.

Adjectives used to describe many of the stories in the science fiction pulps included *primitive*, *trashy*, *tawdry*, *silly*, *absurd*, *crummy*, *ludicrous*, and *cheap*. One early pulp magazine actually boasted, of its contents, that they contained "sensational fiction with *no* philosophy." Speculate on how such a low-level 'literary' form could have been so successful in finding an enthusiastic audience for time travel paradox tales, tales that are in fact by their very nature simply *stuffed* with philosophical issues. As an example of the tremendous emotional power a particularly well-written time travel story can deliver, read Isaac Asimov's "The Ugly Little Boy" (*Galaxy Science Fiction*, September 1958). Asimov rated this story as among his most favorite of all the many he wrote. If you can read it without ending in tears, well, ... An excellent modern historical work on the pulps (of all genres, not just science fiction) is by Lee Server, *Danger Is My Business: an illustrated history of the fabulous pulp magazines, 1896–1953*, Chronicle Books 1993.

A literary fascination with time was already 'in the wind' when Wells wrote his *Time Machine*, as with Oscar Wilde's 1890 novel *The Picture of Dorian Gray*. Even decades earlier than that one can find a hint of time travel of a sort in Edgar Allen Poe's 1841 short story "Three Sundays in a Week." And just 4 years later Henry Wadsworth Longfellow wrote his haunting poem "The Old Clock on the Stairs," with these opening words:

Somewhat back from the village street
Stands the old-fashioned country-seat.
Across its antique portico
Tall poplar-trees their shadows throw;
And from its station in the hall
An ancient timepiece says it all,—

"Forever—never!

(continued)

Never—forever!"

The most interesting of all pre-Wells time travel fiction to appear in a mass-audience publication was, I think, the short story "The Old Folks Party" by Edward Bellamy, printed in the March 1876 issue of *Scribner's Monthly*. In this story a group of teenagers, who belong to a weekly discussion club, agree that at their next meeting they will all come dressed and behaving as they believe they will be dressing and behaving 50 years in the future. Also attending will be the grandmother of one of the young ladies. The meeting of the "old folks" takes place, and it invokes such powerful feelings of mortality that, at last, one of the young men can stand it no more: "Suddenly Henry sprang to his feet and, with the strained, uncertain voice of one waking himself from a nightmare, cried:—'Thank God, thank God, it is only a dream,' and tore off the wig, letting the brown hair fall about his forehead. Instantly all followed his example . . .." The young people then began to laugh with relief at once again being young, until they notice the grandmother is crying. Her granddaughter instantly knows what is wrong and says, "Oh, grandma, we can't take you back with us." Read, compare, and contrast, these works by Wilde, Poe, Longfellow and Bellamy, with the 'scientific' presentation of time travel by Wells.

Acknowledgements

My name is on *Time Machine Tales* as the author, but there are many others without whose help I could not have written it. For this, the revised and updated third edition of the original 1993 book *Time Machines*, I am enormously grateful to the publisher, Springer Science + Business Media, and to my physics editors at Springer, Dr. Sam Harrison and his assistant Ho Ying Fan in New York City. They made an intellectually demanding and inherently lengthy project (one that could easily have turned into a nightmare) a pleasant one. Springer's New York City-based editorial assistant Irene Bruce smoothly handled the administrative details of transforming the book from typescript to print. And to Serguei Krasnikov in Russia, Frank Arntzenius and Roberto Casati in England, Francesco Gonella in Italy, and Geoff Goddu and Archille Varzi in America, I offer grateful thanks for permission to reprint from their publications.

But the story of this book actually begins long before 1993.

Nearly four decades in the past, in 1979, when time travel was still mostly just a nutty idea used by science fiction writers and a few rogue philosophers—proposing to give a serious seminar on time travel at the weekly college *physics* seminar would almost certainly have resulted in getting the bum's rush out the nearest door, or maybe even tossed through the closest window (or, God forbid, *off the roof*!)—I wrote my first time travel tale for *Analog Science Fiction* magazine. Bought by then editor Ben Bova (you can read it in Appendix A), he also bought another time travel tale from me soon after he had become the fiction editor at the newly created *Omni Science Fiction* magazine (you can read it in Appendix B). (Years later, Ben's successor as *Omni*'s fiction editor, Ellen Datlow, bought a third time travel story from me—"The Invitation"—which you can read if you can locate the July 1985 issue.)

Those fictional experiences, and my discovery of Princeton philosopher David Lewis' seminal 1976 essay on the *logical* possibility of time travel, led me to plunge into a deeper study of the *physics* of time travel, to the point where as the 1980s ended (1988, to be precise) I thought a book should be my next writing project. Alas, nobody else shared that thought. I spent the next 3 years looking for a

publisher among numerous university presses; as I looked, the pile of rejection letters, like entropy, grew steadily larger.

The editor at one well-known university press, in fact, simply laughed at the idea of a scholarly time travel book when I called him on the telephone (e-mail was then still in its infancy) to ask, after a long period of no response to my written proposal, if he had gotten it. (Perhaps, I told myself, the mail truck carrying my precious document had fallen into a hyper-dimensional spacetime warp: how pathetically desperate is the anxious academic writer!) Later, I published a number of math/physics books with that same press (but now with new editors) and so there were no lasting hard feelings.

That laughing, unresponsive editor wasn't alone in his opinion, I have to admit, and it wasn't until 1991 when Maria Taylor, the publisher at the Press of the American Institute of Physics (AIP Press), decided to take a chance (a *big* chance) and publish the first edition of *Time Machines*. She made that decision in large part because of two extremely supportive reviews of my proposal from the academic physicists Edwin Taylor (MIT, and a former editor of *The American Journal of Physics*) and Gregory Benford (University of California, Irvine, as well as being an award-winning writer of science fiction who occasionally used the time travel theme). After the original *Time Machines* appeared, Ben Bova asked me to use it as a guide to writing a book on time travel for the new Writer's Digest series on science fiction that he was editing, aimed specifically at would-be story writers; that book came out in 1997 as *Time Travel*. Later, Trevor Lipscombe, then editor-in-chief at the Johns Hopkins University Press (and my former math editor at Princeton University Press), reprinted *Time Travel* (with a new Preface) in 2011.

After Springer acquired AIP Press (and all of its books) in 1995, and Maria Taylor herself went over to Springer, she and I collaborated again on bringing out the second edition of *Time Machines* in 1999. And now, 18 years later, here (with a title slightly altered for reasons I give in "Some First Words") is its successor.

To Sam, Ho Ying, Irene, Ben, Ellen, Maria, Edwin, Gregory, and Trevor, thank you for your support. But my greatest debt of all, one I can never even begin to repay, is to my wife of 55 years, Patricia Ann, a woman of infinite tolerance. Who else would put up with someone who plays first-person-shooter video games at midnight (the bigger and the louder the explosions, the better!) and writes books on time machines, all the while claiming not to be crazy?

University of New Hampshire, Exeter, NH, USA Paul J. Nahin

Introduction

> Over the last few years leading scientific journals have been publishing articles dealing with time travel and time machines. . . . Why? Have physicists decided to set up in competition with science fiction writers and Hollywood producers?
>
> —John Earman (see note 25 of *Some First Words*)

Writing about time travel is, today, a respectable business. It hasn't always been so. After all, time travel, *prima facie*, appears to violate a fundamental law of nature; every effect has a cause, with the cause occurring before the effect. Time travel to the past, however, seems to allow, indeed to *demand*, backwards causation, with an effect (the time traveler emerging into the past as he exits from his time machine) occurring *before* its cause (the time traveler pushing the start button on his machine's control panel years *later* to start his trip backward through time).

Thus, when H. G. Wells published his breakout masterpiece, *The Time Machine*, in 1895, even those readers who loved it as a *story* (and not all did) were still quick to dismiss it as a *romantic fantasy*. It was, in their view, certainly an emotionally powerful tale of pure imagination, but nothing more. Reviewers of the day used such words as "hocus-pocus" and "bizarre," and called the work a "fanciful and lively dream."[26] Any one of the novels by Wells' contemporary, Jules Verne (even such super-technology ones like the 1865 *From the Earth to the Moon*) would have been ranked *far* above Wells' novella in terms of 'it could actually happen.'

Wells himself always denied that his time machine was anything more than a literary device[27] to get his Time Traveller into the far future. Indeed, in 1934, in the preface to *Seven Famous Novels* (published by Knopf), a collection of his novel-length scientific romances (as science fiction had been known before the term *science fiction* came into use), including *The Time Machine*, Wells made his position perfectly clear: "These stories of mine collected here do not pretend to deal with possible things; they are exercises of the imagination ... They are all fantasies; they do not aim to project a serious possibility; they aim indeed only at the same amount of conviction as one gets in a good gripping dream." Wells then went on to say in that same preface that all attempts before at writing fantastic stories depended on magic. But not in his works. "It occurred to me that instead of the usual interview with the devil or a magician, an ingenious use of scientific patter might with advantage be substituted." Wells' great contribution to time traveling story-telling was his introduction of a *machine*; science instead of magic, drugs, dreams, blows on the head, or suspended animation.[28] Not all modern science fiction writers have followed Wells' lead, however.

---

[26] These reviews are reprinted in P. Parrinder, *H. G. Wells: The Critical Heritage*, Routledge & Kegan 1972. A modern reviewer has applied such negative characteristics to the Time Traveller, himself, calling him "a kind of Trickster figure" and "a quack and magician." See Robert J. Begiebing, "The Mythic Hero in H. G. Wells's *The Time Machine*," in *Essays in Literature*, Fall 1984, pp. 201–210.

[27] Wells was not the first to use a machine to enable time adventures, as the Spanish writer Enrique Gaspar (1842–1902) used one in his 1887 story *The Time Ship: A Chrononautical Journey*. It's Wells' tale we remember, however.

A science fiction tale by Clifford Simak (1904–1988), for example, the 1978 novel *Mastodonia*, incorporates an alien creature marooned on Earth (because of a spaceship crash centuries earlier) who 'makes time tunnels.' One of the characters in the story, who is attempting to start a time-travel agency using these tunnels, explains why not having a time *machine* is causing her difficulties with prospective clients: "The whole trouble was that I couldn't tell them about some machine—a time-travel machine. If I could have told them we'd developed a machine, they'd have been more able to believe me. We place so much trust in machines; they are magic to us. If I could have outlined some ridiculous theory and spouted some equations at them, they would have been impressed." I think that's off the mark. We trust in machines not because they are magic, but for precisely the opposite reason. They are *not* magic, but rather are *rational*. And to dismiss mathematics is to say that some non-natural—some supernatural—influence is at work.

But is a time *machine* actually possible? Or is the idea of a time machine simply "Nonsense" and "A bilgeful of crap," as a character bluntly puts it in the 1972 novel *The Dancer from Atlantis* by Poul Anderson (1926–2001). Wells, himself, addressed this point in an autobiographical essay (published in the *Cornhill Magazine*) that he wrote in July 1945 (just 13 months before his death) in even blunter words. Writing under the name of "Wilfred B. Batterave," he penned a very funny summary of his life titled "A Complete Exposé of This Notorious Literary Humbug." There he described *The Time Machine* as "[A] tissue of absurdities in which people are supposed to rush to and fro along the 'Time Dimension.' By a few common tricks of the story-teller's trade, Wells gets rid of his Machine before it can be subjected to a proper examination. He cheats like any common spook raiser. Otherwise it is plain commonsense that a man might multiply himself indefinitely, pop a little way into the future and then come back. There would then be two of him. Repeat *da capo* and you have four, and so on, until the whole world would be full of the Time Travelling Individual's vain repetitions of himself. The plain-thinking mind apprehends this in spite of all the Wellsian mumbo-jumbo and is naturally as revolted as I am by the insult to its intelligence." Funny, yes, but still pretty harsh stuff.

---

[28] Examples of 'non-machine' time travel stories of the last four types are, respectively, H. G. Wells' "The New Accelerator" (1901), Charles Dickens' *A Christmas Carol* (1843), Mark Twain's *A Connecticut Yankee in King Arthur's Court* (1889), and Edward Bellamy's *Looking Backward, 2000–1887* (1888).

As one writer has argued,[29] Wells was, rather than presenting a scientific discovery, simply attempting to refute the nearly suffocating, unjustified (in his mind), smug optimism of the well-to-do of the Late Victorian Age. And so, on his journey to the year A.D. 802,701, the Time Traveller finds the awful decay of humanity in the cannibalistic subjugation of the Eloi by the Morlocks, the end result of class warfare between the working class (Morlocks) and the idle, parasitic upper class (Eloi).

The German social philosopher Karl Marx, if he hadn't already been dead for 12 years in 1895, would surely have nodded in vigorous agreement as he read *The Time Machine*, even as he would have regretted Wells' decision to have the victory of oppressed workers take so long. (What irony that he is buried in London's Highgate Cemetery, the Victorian Valhalla where he has spent the last century and more quite literally mingling with many of the capitalistic ancestors of the Eloi!) What Marx would have thought of *time travel* as a possibility is, however, far less certain.

How things changed in the years that followed *The Time Machine*. There was, at first, admittedly a 'slight' decline in literary merit as the newly developing pulp science fiction magazines picked-up and ran with the time travel genre. Many of the magazine time travel tales of the 1920s, 1930s, and 1940s were, frankly, simply awful. BUT—some were pretty good, too. And some were, in fact, *very* good. From the 1950s on, there have been ever more sophisticated time travel tales from ever more sophisticated writers.

In the academic communities of philosophers and physicists, too, big events occurred. I give the philosophers the edge, in fact, with the 1976 publication of a hugely important paper that opened with these dramatic words: "Time travel, I maintain, is possible. The paradoxes of time travel [to the past] are oddities, not impossibilities. They prove only this much, which few would have doubted: that a possible world where time travel took place would be a most strange world, different in fundamental ways from the world we think is ours."[30] That writer wasn't the first philosopher to write on time travel to the past, but none had expressed themselves in such powerful and unequivocal words in unmistakable support of the concept.

---

[29] R. M. Philmus, "*The Time Machine*; Or, the Fourth Dimension as Prophecy," *Publications of the Modern Language Association*, May 1969, pp. 530–535.

[30] David Lewis, "The Paradoxes of Time Travel," *American Philosophical Quarterly*, April 1976, pp. 145–152. Lewis (1941–2001) was a Princeton University philosophy professor.

Lewis' paper is also notable because it gives what seems to be a clear definition of just what it means to say one has 'traveled in time,' either to the past *or* to the future:

> What is time travel? Inevitably, it involves discrepancy between time and time. Any traveler departs and then arrives at his destination; the time elapsed from departure to arrival (positive, or perhaps zero) is the duration of the journey. But if he is a time traveler, the separation in time between departure and arrival does not equal the duration of the journey.

To understand this, we need to appreciate the distinction between the *personal time* of the time traveler and the *external time* of remote observers of the time traveler. A time traveler's personal time is measured, for example, either by the time kept by his wrist watch or, perhaps, by a burning candle. (This distinction had actually appeared earlier in Horwich's paper—see note 27 in *Some First Words*—published the year before Lewis' paper.)

I say I 'give the edge to the philosophers' because, while the first *physics* time travel paper had appeared decades earlier, its author wasn't really a physicist at all but rather was Einstein's friend, the world-famous mathematical logician Kurt Gödel. Gödel's paper was, in retrospect, a pivotal event in establishing the 'respectability' of *scientific* time travel; it's worthwhile to take some time here to explain this important point. For physicists (and for philosophers and science fiction writers, too) a 'time machine,' one either constructed by intelligent beings or occurring naturally, manipulates (all the while obeying the known laws of physics) finite amounts of matter and energy in a finite region of spacetime.[31] A 'time machine' would be declared to be *plausible* if it could be explained by a rational, scientific theory. Such a rational theory is found in Einstein's general theory of relativity. (His *special* theory of relativity applies in those situations where there is no gravity.)

Until Einstein, the theory of gravity used by scientists was Newton's—a theory that, although amazingly accurate for any situation encountered on Earth, does have observable errors in certain astronomical applications. In addition, Newton's theory is a descriptive one; it makes possible the calculation of gravity effects without offering any explanation for gravity itself. Einstein's theory not only gives the right answers, even in those cases where Newton's theory doesn't, but it also explains gravity. It does that by treating the world as a four-dimensional structure in which all four dimensions (three of space and one of time) are in a certain sense on equal footing. The resulting Einsteinian description of the world is that of a unified spacetime in which time and space are intimately intertwined, whereas Newton's theory keeps time and space separate and distinct.

---

[31] I am going to feel free to use words like *spacetime* without having to first write introductory essays on relativity theory and tensor mathematics, because such words have entered common use. All those Hollywood science fiction movies, even the crummy ones that routinely trash the laws of physics, have at least expanded the general imagination!

As Newton wrote of time, at the start of his 1687 masterpiece *Principia*, a work that revolutionized physics, “Absolute, true, and mathematical time, of itself, and from its own nature, flows equably without relation to anything external, and by another name is called duration.” This view of time would be, of course, discarded with the arrival of Einstein and his view of variable time depending on the state of the observer.

Unlike Einstein’s view, Newton’s view of the nature of time was entangled with theology. As one modern theologian has written, “Newton conceived of absolute time as grounded in God’s necessary existence.”[32] To quote Newton himself, in the *General Scholium* to the second edition of *Principia* (1713) he added words that didn’t appear in the original: “God is a living, intelligent, and powerful Being; and, from his other perfections, [it follows] that he is supreme, or most perfect. He is eternal and infinite, omnipotent and omniscient; that is, his duration reaches from eternity to eternity; his presence from infinity to infinity; he governs all things, and knows all things that are or can be done. He is not eternity and infinity, but eternal and infinite; he is not duration or space, but he endures and is present. He endures forever, and is everywhere present; and, by existing always and everywhere, he constitutes duration and space. Since every particle of space is *always*, and every indivisible moment of duration is *everywhere*, certainly the Maker and Lord of all things cannot be *never* and *nowhere*.”

Okay, I’ll be honest—I really am not at all sure just what that means! Newton added these words to the *Principia* in response to criticism (from the influential philosopher George Berkeley (1685–1753)) that his original statements about absolute time were “pernicious and absurd notions,” notions that were in fact atheistic in conception. That was a most serious charge in Newton’s day, and he was trying (I think) to find some cover from those critics who spent more hours of the day thinking about God than of physics. Much more honest (in my opinion) are the witticisms ‘time is just one damn thing after another’ and ‘time is what keeps everything from all happening at once.’[33] More funny than useful, yes, of course, but at least they’re funny.

Newton’s theological view of time is simply irrelevant to the modern physicist (although perhaps of more interest to the philosopher-historian) but in many cases it is of *central interest* to the science fiction writer. For example, Newton’s religious mindset and its (perhaps!) connection with time travel is treated in my short story "Newton's Gift," originally published in *Omni Magazine* (January 1979) and reprinted in Appendix B at the end of this book. Wells' Time Traveller's view of time is more Newtonian than it is Einsteinian—and perhaps that's not such a big surprise, considering that Einstein was only 16 years old when *The Time Machine* was published.

---

[32] William Lane Craig, “God and the Beginning of Time,” *International Philosophical Quarterly*, March 2001, pp. 17–31, which discusses the question ‘Why didn’t God create the world sooner?’ One irreverent answer is ‘He was busy creating Hell for all those who ask that question,’ but a more scholarly analysis can be found in Brian Leftow, “Why Didn’t God Create the World Sooner?” *Religious Studies*, June 1991, pp. 157–172.

[33] This last ‘definition’ first (as far as I know) appeared in the work of the science fiction writer Ray Cummings (1887–1957), in his 1921 story “The Time Story,” published in *Argosy-All-Story* magazine. He repeated the phrase in his 1929 novel *The Man Who Mastered Time*, and then again in the 1946 novel *The Shadow Girl*. (“This same Space; the spread of this lawn . . . what would it be in another 100 years? Or a 1000? This little space, from the Beginning to the End so crowded with events and only Time to hold them apart!”)

From the first (1905) it has been known that Einstein's special theory allows time travel into the future via the well-known mechanism of *time dilation*. (The faster a rocket ship travels relative to Earth, the slower is the tick-tock of a wrist watch worn by a rocketeer, compared to that of an identical watch back on Earth.)[34] To return from the future, however, to travel back into the past to the instant after the traveler began his journey, had been thought to be impossible. It was Gödel's discovery that showed the general theory, which has passed every experimental test it has been subjected to (most recently, the September 2015 detection, from two massive colliding black holes, of gravitational waves—'ripples in spacetime'—generated more than a billion years ago in an effect predicted by the general theory a century ago), does allow time travel to the past *under certain conditions*. It is this availability of a *theory* that distinguishes time travel speculations from the outlandish fantasy speculations with which it is often unjustly lumped—speculations that *are* in the province of quacks (such as ESP, astrology, and mind over matter a' la spoon bending).

In his general theory, Einstein showed how spacetime can be either 'flat' (in the no-gravity, special relativity case of what is called a *Minkowski spacetime*[35]) or 'curved' (those situations with gravity), and he did that not by verbal hand waving, but rather by writing mathematical equations that obey all the known laws of physics: his famous gravitational field, nonlinear differential tensor equations. These complicated equations are notoriously difficult to solve in general, but in certain, special cases they *have* been solved. Those solutions describe how matter and energy and spacetime interact. As the popular saying puts it, "Curved spacetime tells matter how to move, and energy and matter tell spacetime how to curve." In that sense, gravity *is* curved spacetime.

In 1949 Gödel found one such special solution to the field equations that describes the movement of mass-energy not only through space but also *backward in time* along trajectories in spacetime that are called *closed time-like lines* or *curves*

---

[34] One pulp magazine science fiction story (F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931) got this right when its time traveler explains how his time machine works with these words: "Time as we know it is not universally absolute. The rate of its passage depends to a great extent upon the velocity of its observer with regard to some certain reference system. A moving clock will run slower with respect to a selected coordinate system than a stationary one." (Recall my earlier comments on the personal time of a time traveler.)

[35] Named after Hermann Minkowski (1864–1909), Einstein's mathematics professor in Zurich who gave the now well-known spacetime diagram interpretation of special relativity which, when originally presented by Einstein, was in the form of pure mathematics.

(called CTLs or CTCs, respectively).[36] These trajectories are such that if a human traveled along one, *always at a speed less than that of light* (that's what *time-like* means), he would see everything around him happening in normal causal order from moment to moment (for example, the second hand on his wrist watch would tick clockwise into the local future), but eventually the CTL/CTC closes back on itself and the traveler finds himself in his own past.

On the scale of the Solar System, general relativity has causality built into itself, but on much larger scales things can be a good deal more complicated. On a very large, astronomical scale, in fact, curved spacetime can result in violations of causality, with effects occurring before their causes. That is what the physics and the mathematics of Gödel's solution imply. That is what is meant by saying there is a scientific, rational basis for discussing time travel to the past. It is particularly important to note that travel along one of the closed time-like world lines discovered by Gödel requires a *machine*, some kind of accelerating rocket ship. That's because none of Gödel's CTLs/CTCs are what is called a *geodesic*. That is, none are *free-fall* world lines.[37] This machine does not, however, generate CTLs/CTCs where none existed before (CTL/CTC *creation* requires what physicists call a *strong* time machine) but rather simply makes use of the CTLs/CTCs that are inherent in Gödel's spacetime. A Gödelian rocket ship then is an example of a *weak* time machine.

I mentioned earlier that "certain, special cases" of Einstein's gravitational field equations result in CTLs/CTCs. What was the "special case" that Gödel solved? His solution of the field equations is for a rotating, infinite, static universe composed of a perfect fluid at constant pressure. In such a universe Gödel found that naturally occurring CTLs/CTCs pass through every point in spacetime; that is, time travel in Gödel's universe is *not* the result of a machine *manipulating* mass and energy on a *local* scale (the classic science fictional description of a time machine); rather, in Gödel's spacetime time travel is a naturally occurring phenomenon! The observable universe is, however, *non*-rotating and expanding (astronomers see red-shifts in the spectrums of distant stars) and so, although Gödel's spacetime satisfies the general relativity field equations, its time travel property does not hold in the spacetime in which we live. (This may account for why the initial reaction in the physics/philosophical communities, to Gödel's discovery that time travel is *not* nonsense according to general relativity, was mostly indifference.) The failure to observe time travel in our universe may (somewhat surprisingly, I think) still have possible implications for us, however, as one philosopher has cleverly argued.[38] He points out that naturally occurring Gödelian time travel would endow the universe with properties particularly useful for the survival of intelligence (presumably that includes humans) against extinction from a multitude of cosmic disasters. So, for those who argue that the universe we live in was *made* for us (the advocates of various proofs of God's existence that have Him as Designer), we have an obvious question: why did He (apparently) skip incorporating time travel?

---

[36] Kurt Gödel, "An Example of a New Type of Cosmological Solutions of Einstein's Field Equations of Gravitation," *Reviews of Modern Physics*, July 1949, pp. 447–450. A CTL/CTC is a special type of *world line*; the trajectory through spacetime of every particle in the universe is a world line that extends from each particle's past to its future. Our everyday experiences are with world lines that never cross or come close to themselves (which would put a particle at or near the same spacetime point more than once). That lack of experience with CTLs/CTCs that self-intersect is what makes time travel to the past so difficult for humans to grasp. For a discussion of *how* Gödel did what he did, see Wolfgang Rindler, "Gödel, Einstein, Mach, Gamow, and Lanczos: Gödel's Remarkable Excursion into Cosmology," *American Journal of Physics*, June 2009, pp. 498–510.

[37] It was discovered in 1969, however, that this isn't strictly true *if* one allows for a test particle (our 'time traveler') to be electrically charged. Then, naturally present electromagnetic forces acting on the particle could be sufficient to propel the particle along a Gödelian CTL/CTC. That is, no rocket would be required. See U. K. De, "Paths in Universes Having Closed Time-Like Lines," *Journal of Physics A*, July 1969, pp. 427–432. There are other solutions to Einstein's equations that do allow time travel on free-fall geodesics: see, for example, I. D. Soares, "Inhomogeneous Rotating Universes with Closed Timelike Geodesics of Matter," *Journal of Mathematical Physics*, March 1980, pp. 521–525.

In an invited essay that appeared the same year as his time travel physics paper, Gödel specifically addressed the seemingly paradoxical aspect of what he had discovered: "By making a round trip on a rocket ship in a sufficiently wide course, it is possible in these [rotating] worlds to travel into any region of the past, present, and future, and back again, exactly as it is possible in other worlds to travel to distant parts of space. This state of affairs *seems* [my emphasis] to imply an absurdity. For it enables one, e.g., to travel into the near past of those places where he has himself lived. There he would find a person who would be himself at some earlier period of life.[39] Now he could do something to this person which, by his memory, he knows has not happened to him."

Gödel's nerve then failed him, and he defended the possibility of the paradox of a time traveler meeting himself in the past with what I think an astonishingly unconvincing argument (particularly so for a logician) based primarily on *engineering* limitations: "This and similar contradictions, however, in order to prove the impossibility of the worlds under consideration, presupposes the actual feasibility of the journey into one's own past. But the velocities which would be necessary in order to complete the voyage in a reasonable time are far beyond everything that can be expected ever to become a practical possibility. Therefore it cannot be excluded a priori, on the ground of the argument given, that the space-time structure of the real world is of the type described."[40] That is, Gödel was trying to head off critics of his rotating universe model who might point to the time travel result as proof that the model had to be flawed.

---

[38] Alasdair M. Richmond, "Gödelian Time-Travel and Anthropic Cosmology," *Ratio*, June 2004, pp. 176–190. Not all physicists think Gödel's result is actually time travel. At least two think it is all simply the result of mathematical hijinks, and that time machines must remain "an aspect of science fiction fantasy": see F. I. Cooperstock and S. Tieu, "Closed Timelike Curves and Time Travel: Dispelling the Myth," *Foundations of Physics*, September 2005, pp. 1497–1509. This skepticism towards Gödel actually started much earlier, when two physicists (one a Nobel physics laureate) incorrectly claimed Gödel had simply gotten his math wrong: see S. Chandrasekhar and J. P. Wright, "The Geodesics in Gödel's Universe," *Proceedings of the National Academy of Sciences*, March 1961, pp. 341–347. It was those two physicists who had erred, however, as was pointed out by the philosopher Howard Stein, in his "On the Paradoxical Time Structures of Gödel," *Philosophy of Science*, December 1970, pp. 589–601.

[39] You'll recall that this is *precisely* the situation that Wells mentions in his "Notorious Literary Humbug" essay. If only he had lived just three more years, to see what he thought to be an absurdity actually appear in the serious writings of a brilliant mathematician!

In a footnote Gödel says that the time traveler would have to move at least as fast as nearly 71 % of the speed of light, and that if his rocket ship could "transform matter completely into energy" then the weight of the fuel would be greater than that of the rocket by a factor of 10^{22} divided by the square of the duration of the trip (in rocket years). A trip to the past in Gödel's universe would require a time machine that looked like Dr. Who's telephone booth attached to a fuel tank the size of several hundred *trillion* ocean liners. These are formidable numbers,[41] but they require no violation of physical laws, and that's what really counts if time travel is to be disproved. Gödel's use of engineering limitations for explaining away backwards time travel is actually worse than simply being wrong, because the puzzle is not in practicality but rather in showing, assuming that general relativity is correct, how correct mathematical physics can lead to what seems to be a paradoxical conclusion. (And see note 12 again, for another reason the 'fuel argument' really has no force at all against the possibility of time travel in Gödelian spacetime.)

So, what did the great man himself, Einstein, think of all this? In the same publication as Gödel's essay, he *cautiously* replied as follows: "Kurt Gödel's essay constitutes, in my opinion, an important contribution to the general theory of relativity, especially to the analysis of the concept of time. The problem here involved disturbed me already at the time of the building up of the general theory of relativity, without my having succeeded in clarifying it ... the distinction 'earlier-later' is abandoned for world-points which lie far apart in a cosmological sense, and those paradoxes, regarding the *direction* of the causal connection arise, of which Mr. Gödel has spoken ... It will be interesting to weigh whether these are not to be excluded on physical grounds."

Despite the mathematical physics of Gödel, showing the possibility of time travel to the past, many philosophers are not quite so sure. As one expressed his concerns, "No science-fiction staple poses more philosophical difficulties than time travel, but there is still no consensus as to whether time-travel fictions exhibit logical, metaphysical, or physical impossibility."[42] The best-known and possibly oldest of the paradoxical situations that seem to be part-and-parcel of time travel is the so-called *grandfather paradox*,[43] expressed this way by philosopher David Lewis in his pioneering 1976 paper (see note 5):

---

[40] Kurt Gödel, "A Remark About the Relationship Between Relativity Theory and Idealistic Philosophy," in *Albert Einstein: Philosopher-Scientist*: volume 7 of *The Library of Living Philosophers* (P. A. Schilpp, editor), Open Court 1949.

[41] For the analysis of a rocket powered by matter/anti-matter, a known physical process that satisfies Gödel's energy requirement for time travel, see E. Purcell, "Radioastronomy and Communication Through Space," in *Interstellar Communication* (A. G. W. Cameron, editor), W. A. Benjamin 1963.

[42] Alasdair Richmond, "Time-Travel Fictions and Philosophy," *American Philosophical Quarterly*, October 2001, pp. 305–318.

> Consider Tim. He detests his grandfather, whose success in the munitions trade built the family fortune that paid for Tim's time machine. Tim would like nothing so much as to kill Grandfather, but alas he is too late. Grandfather died in his bed in 1957, while Tim was a young boy. But when Tim has built his time machine and traveled to 1920, suddenly he realizes that he is not too late after all. He buys a rifle, . . . and there [Tim] lurks, one winter day in 1921, rifle loaded, hate in his heart, as Grandfather walks closer, closer . . .

So, there's the puzzle. Tim can obviously achieve his goal—he has a loaded gun, he's an excellent shot, a clueless granddad is coming ever closer—but if Tim actually does kill grandfather, years *before* Tim was (will be) born, then how can Tim *be* born? And if he is not born, then how can Tim ('now' not in existence) travel back through time to kill grandfather? What a confusing mess, right? So, the only possible conclusion to all this is that the starting premise, that time travel makes sense, must actually be nonsense. Right?

Well, maybe, but then what of Gödel with his time traveling rocket ship? That's hard-as-diamond, unshakeable mathematical physics, for heaven's sake. We can't just ignore *that*! Lewis offers a way out of this conundrum, and when we get to the book's discussions on paradoxes (that's plural because, believe it or not, there are other paradoxes even *more* perplexing than that of killing granddad in the distant past) we'll return to his solution.

Ever since Lewis wrote his paper, philosophers have been particularly fascinated by the grandfather paradox and have shown themselves to be at least as inventive as the science fiction writers in discussing it, or variations on it.[44] Here, for example, is a twist on that paradox that I think particularly clever, one that avoids the murderous spirit of the tale told by Lewis and Horwich:

---

[43] The origin of this paradox is probably lost in time (the irony of that is *so* appropriate!), but I have traced it at least as far back as to the science fiction pulp magazine *Science Wonder Stories* which published, in its December 1929 issue, an editorial essay titled "The Question of Time Traveling." It challenged readers to think about the following scenario: "Suppose I can travel back into time, let me say 200 years; and I visit the homestead of my great great great grandfather, and am able to take part in the life of his time. I am thus enabled to shoot him, while he is still a young man and as yet unmarried. From this it will be noted that I could have prevented my own birth . . .."

[44] Even before Lewis' paper, Paul Horwich had reduced the grandfather paradox to *autoinfanticide*—a time traveler tries to kill his younger *self*—in "On Some Alleged Paradoxes of Time Travel," *The Journal of Philosophy*, August 14, 1975, pp. 432–444. But not all philosophers share this fascination. Earman (see the opening quote), for example, dismisses *all* of the science fiction paradoxes that are so beloved by fans of the genre as "while always good for a chuckle," they are just "crude and unilluminating means of approaching some delicate and deep issues about the nature of physical possibility." I think Earman is fundamentally correct, although I wouldn't go so far as to characterize the paradoxes as mere "chuckles." They are, after all, the source of much of the intellectual motivation prompting the exploration of the physics of time travel. An excellent example of this is found in the paper by the Russian physicist S. V. Krasnikov, "Time Travel Paradox," *Physical Review D*, February 14, 2002, pp. 064013-1 to 064013-8. The physics of the grandfather paradox is of *great* interest in this paper.

> Sarah has just completed building her time machine. She decides to test the machine on herself tomorrow morning at which time she intends to travel back one day. In the meantime, she goes home, puts some salve on the burn she received that day, and goes to bed. In the morning, Sarah, with coffee in hand, sits down to read the morning paper. She opens the paper to the following headline: 'Famous physicist found dead.' On the front page is a picture of her body, salve burn clearly visible on her arm, inside her pristine time machine. Underneath is the caption. 'Nobel-prize winning physicist found dead yesterday in mysterious device that materialized near city hall.' Extremely shaken, Sarah returns to the lab and destroys the time machine.[45]

Can any sense be made of this? We'll come back to this question later in Chap. 5, when we discuss the possibility (or not) of time being *multi*-dimensional.

Now, to conclude this Introduction, let me end with two amusing, connected short stories (in epistle form) that nicely describe the issues we'll take up in the rest of this book. The rejection letter for the denial of a research grant to fund the construction of a time machine has just been received . . .

That Useless Time Machine[46]

Dear Review Committee:
It is not our practice to raise complaints against a negative review report. We believe in peer refereeing and we respect it, whatever its content and consequences. However, in the case of our latest grant application (project named 'The Time Machine') we find it necessary to express our astonishment at the motivations with which our request for funding was turned down. Your main objection appears to be that our project is 'philosophically interesting' but 'practically useless', by which you mean that the project 'has no potential for applications.' We do not quite think that the main criterion for judging the scientific value of a project should be its practical usefulness, but never mind that. Let us agree that usefulness is a relevant criterion, especially when large amounts of money are involved. Why should that be a reason to turn down our project? Quite frankly, we cannot think of a project with better application potential than ours. Some examples:

- Cultural tourism: one could send herds of history fans back in time to witness the crucial episodes of the French Revolution, or to watch the Egyptians build the pyramids, or to videotape Socrates' lectures.
- Exotic safaris: we have already received several applications for dinosaur hunting expeditions (they got extinct anyway).

---

[45] G. C. Goddu, "Time Travel and Changing the Past: (Or How to Kill Yourself and Live to Tell the Tale)," *Ratio*, March 2003, pp. 16–32.

[46] Story by Roberto Casati (Senior researcher at CNRS, Paris) and Achille C. Varzi (Professor of Philosophy at Columbia University). Originally published in *Philosophy*, October 2001, pp. 581–583, and reproduced here by kind permission of the authors.

- Error detection: we could take a closer look at our past mistakes and learn how to avoid them in the future.
- Historic documentaries: think of the huge saving in set design, costumes, special effects, etc. (How much did *Gladiator* cost?)

And so on and so forth. Honestly, can you think of a project with better prospects for useful and thrilling applications?

Sincerely Yours,
The 'Time Machine' Research Group

Dear 'Time Machine' Research Group:
Thank you for your letter. We agree that it would be interesting to exploit a time machine for the uses that you suggest. It would also be remarkable if we could use it to prevent all sorts of unpleasant events that happened in the past. It would be remarkable, for instance, to be able to go back to November 22, 1963, and prevent Lee Harvey Oswald from killing John Kennedy, or to go back to April 14, 1912, and steer the Titanic around the iceberg. It would be excellent indeed to be able to do such things. However, suppose your project were to be successful. Suppose you *will* manage to build a time machine. Then why *didn't* you do any of those things? Why is it that our past history is still full of such sad events? Either this means that your project is doomed to fail and you will never manage to build a time machine; *or* it means that the project will succeed but that you are not going to use your time machine for these good purposes. In the first case, logic shows it would be pointless to support your project. In the second case, ethics dictates that it would be wrongdoing. Either way, you must concede that the reasons against your project are overwhelming.

Cordially Yours,
The Review Committee

Dear Review Committee:
Certainly you have noticed that our suggestions for practical applications of the time machine did not include any uses that could result in an alteration of the natural course of history. As a matter of fact, we believe that no such alteration is logically possible. According to our project, it is logically possible to *visit* the past but not to *modify* the past. No time traveler can undo what has been done or do what has not been done. So the logic is safe. This does not mean that the time traveler will be ineffectual during her stay in the past, of course; it simply means that what she is going to do is something that she has already done. An accurate catalogue of all the past events would include an account of the arrival of the Time Machine from out of nothing as well as an account of all the actions and reactions that followed. And ethics is safe, too. For, if indeed we managed to go back to Dallas, we could not stop Oswald from doing what he did. Nobody would be able to stop Oswald because nobody was able to stop him (and nobody was able to stop Oswald because nobody will ever be able to do so, even if they came from the future). Alas, the past is full of sad events but there is nothing that we can do about that.

Respectfully Yours,
The ‘Time Machine’ Research Group

Dear ‘Time Machine’ Research Group:
We appreciate the distinction between changing the past (impossible) and affecting the past (possible). However, this simply reinforces our initial impression: your project has no practical value. If in order to travel to the past one has to have been there already, and if one can only do what has already been done, then *á quoi bon l’effort*? Why should we invest in a ‘Time Machine’ at all? We are afraid that our decision is now final.

Yours with best wishes,
The Committee

Well, all seems to be certainly lost with *that*. But, wait, perhaps not. Maybe, with just one more *really good* appeal, The Committee’s rejection can be reversed! If *you* were on the Review Committee, and had just read the following letter, how would *you* vote?

A Useful Time Machine[47]

Dear Review Committee:
We regret your continued decision to reject our proposal. Even though you have told us your decision is now final, we humbly ask your indulgence for one last appeal. We believe you have misinterpreted a crucial part of our proposal.

You maintain that our ‘Time Machine’ project ‘has no potential for applications’ and has ‘no practical value.’ You ultimately base this claim on the fact that “If in order to travel to the past one has to have been there already, and if one can only do what has already been done, then *á quoi bon l’effort*? Why should we invest in a ‘Time Machine’ at all?” Your argument however is a misinterpretation of our own comments that ‘According to our project it is logically possible to visit the past but not to modify the past . . . This does not mean that the time traveler will be ineffectual during her stay in the past, of course; it simply means that what she is going to do is something that she has already done.’ We regret the awkward and easily misleading locution of the last sentence, but such are the perils of talking about time travel. Regardless, please consider our clarification.

Certainly if we were proposing that the time traveler *be* 5 years old again, we would be proposing something not worth the effort—our proposed time traveler has already turned five and cannot do so again. But we are not proposing that the time traveler do things that have already occurred in her own personal past, but rather in her personal future. The time traveler has not yet, from her personal temporal perspective, travelled back to, say, the library at Alexandria in 100 BCE. When she does travel back to 100 BCE to obtain scans of the books in the library before its destruction, she will be older than she is now. When she returns she will be still older (and we hope wiser, i.e., in possession of valuable information to which neither you nor we currently have access).

---

[47] Story by Geoff Goddu (Professor of Philosophy at the University of Richmond, Virginia). Originally published in *Philosophy*, April 2002, pp. 281–282, and reproduced here by kind permission of the author.

But is it true that as of 2002 AD [the year this letter was written] the time traveler has already visited Alexandria in 100 BCE? It could well be. But whether or not it is depends upon whether it is *also* true that our project will be successfully funded and completed. Because time travel into the past involves reverse causation, certain past events, such as the time traveler visiting 100 BCE, will be dependent upon certain future events, such as the successful funding and completion of our project. Hence, if it is not true that our project is both funded and completed, then it is not true that our time traveler has of 2002 already visited 100 BCE.

But suppose we were to learn now, before the funding and completion took place, that our time traveler had indeed been present at the library in Alexandria in 100 BCE. Would this imply that there was no reason to expend the effort to fund our project? After all, if the travel has 'already' happened, why bother funding the project? Firstly, such an argument does not imply that a 'Time Machine' would have no practical application, but rather expresses the futile hope that one could in fact get the practical benefits (if time travel is successful, we obtain the desired information) without expending the effort at all. Secondly, the hope is futile, for if we learn right now that our time traveler had been present at the library in 100 BCE, we would then know, assuming no other possible funding source, that you *will* expend the effort to fund our project. To deny this last is to make the impossible suggestion that even though your support is truly a causal antecedent of the successful trip, there is now no need for you to actually expend the effort to provide funding.

Hence, the effort is far from pointless, for the project will only succeed through your and our efforts. And success will generate, not only all the practical applications we outlined in our first letter, but, in addition, a host of information gathering applications such as more accurate historical research, lost item location identification, legal testimony verification, etc. Even if, as we (and you) acknowledged, no one could now prevent Oswald from killing Kennedy, wouldn't it be worth verifying that Oswald was the lone killer of Kennedy? Also, the information gathering need not be restricted to the past. For example, information concerning the prices of various stocks 10 years from now would be extremely valuable to a suitably cautious and prudent investor. Surely you cannot object to our information gathering in the future on the grounds that 'it will already have been done.' And just think, the information we obtain could be what allows you to obtain at very low prices those stocks that in the future will be extremely valuable and allows your esteemed committee to dramatically increase your support of worthy scientific endeavors.

Again, we ask you to reconsider your original decision.
Respectfully yours,
The 'Time Machine' Research Group

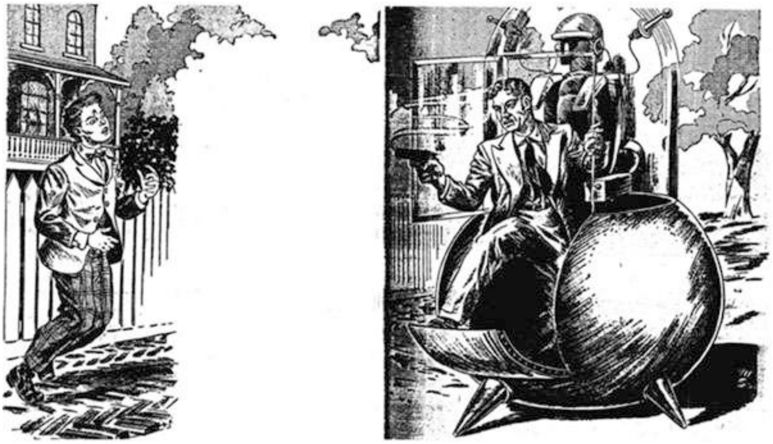

A time machine inventor makes an experimental test of the grandfather paradox! (Illustration from “Thompson’s Time Traveling Theory” by Mortimer Weisinger), *Amazing Stories* March 1944 (art by Malcolm Smith). Reprinted by arrangement with Forrest J. Ackerman, Holding Agent, 2495 Glendower Ave., Hollywood, CA 90027.

Not everybody likes time machines as a science fiction gadget, not even otherwise enthusiastic devotees of the genre. For example, in a Letter-to-the-Editor published in the December 1931 issue of Astounding Stories*, one seventeen-year-old fan had this to say: “There is only one kind of Science Fiction story I dislike, and that is the so-called time-traveling. It doesn’t seem logical to me. For example, supposing a man had a grudge against his grandfather, who is now dead. He could hop in his machine and go back to the year that his grandfather was a young man and murder him. And if he did this how could the revenger be born? I think the whole thing is the ‘bunk.’” As this book will demonstrate, this young reader was not alone in that opinion. As this book will also demonstrate, in the last few decades that view has been rapidly evolving.*

For Further Discussion

Read again the penultimate sentence in the last letter from The 'Time Machine' Research Group, and then think about how you would respond to the following questions.

(1) Would *you* invest in a stock market if you knew somebody else had a time machine giving *them* advance information on stock performance?
(2) How might the existence of a time machine influence the future of the stock market, in general? For an early science fiction look at these questions, see Lee Laurence, "History in Reverse," *Amazing Stories*, October 1939.

One writer has speculated that Wells' model for the Time Traveller was the American inventor Thomas Edison. (See Martin T. Willis, "Edison as Time Traveler: H. G. Wells' Inspiration for His First Scientific Character," *Science Fiction Studies*, July 1999, pp. 284–294.) As Wells worked his way from *The Chronic Argonauts*, through revisions, to the final *Time Machine*, the story's hero evolved from Dr. Nebogipfel to the Philosophical Inventor to the Time Traveller. The one individual who could have inspired all of these various hero types was, according to Willis, Edison, a world-famous Victorian-age celebrity whose story was well known to Wells. If Wells had today's scientific personalities available as potential inspirations, who do you think he would use? How might that choice affect the story and structure of *The New Time Machine*?

The idea of *personal time*, used by the philosopher David Lewis (note 5) to consistently interpret time travel stories, has been used in a quite different way (although time travel gets a few words, too) by the philosopher Roy Sorenson. In his paper "The Cheated God: Death and Personal Time," *Analysis*, April 2005, pp. 119–125, Sorenson asks you to imagine an immortal god. For some reason this god runs afoul of a demon, who curses the god in a curious way. (The 'telling of a story' is a common technique in philosophical papers and, while foreign to what readers of physics papers are used to seeing, is not without some charm. Just be sure to always keep in mind that its

(continued)

primary use is as an *attention-grabbing* device, but as far as having any other merit, well, that's often another story.) The curse is such that the life span of the once immortal god is reduced to that of a normal human life span and yet, perhaps surprisingly, the god will still never die. As Sorenson writes, "[The god] will live forever. But [the god] will not have a better life than a mortal. The demon has harmed [the god] as gravely as death harms mortals." How, you might wonder, is this to be done? As Sorenson explains, "[The god] lives half of its now mortal span, followed by a trillion years of nothingness, then a quarter of its mortal span followed by a trillion years of nothingness, then an eighth of its mortal span followed by a trillion years of nothingness and so on ad infinitum." Sorenson's argument is simply an exotic form of the high school summation of the geometric series $\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots=1$, where there are an *infinite* number of terms to the left of the equality. (Each term represents a period of time during which the god is conscious, and each + represents a trillion years.) Sorenson picked a trillion years of nothingness between consecutive periods of consciousness for (I suggest) dramatic reasons, but suppose instead that he had picked 1 μs for the period of nothingness. Discuss what effect this would have (if any) on the life of the god. Consider two cases:

(a) There is no minimum time duration for consciousness, and
(b) There *is* a minimum time duration such that, for any shorter duration, a consciousness remains 'unaware' even though it is *not* in a state of nothingness.

After working all night making some final calculations, a physicist carefully solders a final resistor into the control module of the world's first time machine and then steps into the gadget that is a sure bet to win the next Nobel Prize in physics. As she does, she notices that it is precisely 8:10 in the morning, as indicated on both her wrist watch and the clock on the lab wall. After settling into a plush leather seat she pushes the time machine's power button, the machine glows with a flickering blue-red halo and hums with a mighty throb for a while and then, at precisely 8:15 by her wrist watch, she steps out of the machine and back into her lab. She notices the clock on the wall now reads 8:05. That is, she took 5 min of personal time (8:10 to 8:15) to travel 5 min of external time into the past (8:10 to 8:05). On the one hand she certainly seems to be a time traveler, in that she exits the machine before she enters it. (Ignore the issue of there being *two identical* physicists in the lab from 8:05 to 8:10!) On the other hand, the elapsed personal and external

(continued)

times are *equal*. Does this suggest a need to modify or expand David Lewis' definition of a time traveler? As you ponder this question, you might want to read the following four papers: (1) Paul R. Daniels, "Lewisian Time Travel in a Relativistic Setting," *Metaphysica*, October 2014, pp. 329–345, (2) Douglas Kutach, "Time Travel and Time Machines," in *A Companion to the Philosophy of Time* (H. Dyke and A. Bardon, editors), Wiley-Blackwell 2013, pp. 301–314, and (3) Frank Arntzenius, "Time Travel: Double Your Fun," *Philosophy Compass*, November 2006, pp. 599–616. A bit more demanding (but worth the effort) is the long chapter "Time Travel and Time Machines" by Chris Smeenk and Christian Wüthrich, in *The Oxford Handbook of Philosophy of Time* (C. Callender, editor), Oxford 2011, pp. 577–630 (see page 580, in particular).

> The idea that information is physical has given rise to a series of discoveries which indicate that physics has much to say about fundamentals of computer science.

The above quotation is the opening sentence to a most interesting paper by the physicist Dave Bacon, "Quantum Computational Complexity in the Presence of Closed Timelike Curves," *Physical Review A* (70), 2004. (When he wrote, Bacon was at Caltech, but he is now a software engineer at Google.) The title of Bacon's paper, translated into blunt English, is "It Would Be Really Neat If We Could Merge a Time Machine With a Computer." That is, to further quote from Bacon's paper, "One could [efficiently] solve a hard problem by trying out a solution to the problem, sending one's computer back in time, attempting a different solution to the problem, sending one's computer back in time, etc., until a solution to the problem has been found." There then follows a pretty sophisticated analysis on the self-consistent time evolution of a quantum system, ending with Bacon's frank admission that "we would not be honest if we did not end this paper with the caveat that this work is at best a creature of eager speculation . . . Practical considerations are humorous at best." Read Bacon's paper and discuss what he means by "a hard problem." (There *is* a technical term used by computer scientist for such problems: NP-complete.)

The occasional theological commentary in this book may strike some as a bit odd for a topic treated with heavy doses of deep mathematics in the physics literature but, as you'll see on the following pages, theology is an unescapable dimension to any informed discussion of time travel. A literary connection

(continued)

between time travel and theology has, in fact, existed for a long time. As pointed out in Paul Alkon's *Origins of Futuristic Fiction* (University of Georgia Press 2010), "The first time-traveler in English literature is a guardian angel who returns with state documents from 1998 to the year 1728 in Samuel Madden's *Memoirs of the Twentieth Century*" (published in 1733, nearly three centuries ago). Madden was an Irish-Anglican clergyman whose book was satire rather than science fiction, but its time traveling aspect was a first. As Professor Alkon also writes, "Madden [was] the first to write a narrative that purports to be a document from the future. He deserves recognition as the first to toy with the rich idea of time-travel in the form of an artifact sent backward from the future to be discovered in the present." Your assignment: read and discuss Alkon's book.

You'll recall that Gödel cast his view of time travel in the form of a self-encounter in the past. In Frederik Pohl's "Let the Ants Try," we find a science fiction tale that appeared essentially simultaneously with Gödel's paper (*Planet Stories*, Winter 1949), in which a time traveler journeys back forty million years. Upon stepping out of his time machine, he hears a "raucous animal cry" from somewhere in the nearby jungle. Later, after other adventures in time, he returns to near the same point in spacetime. After stepping out of his time machine, he sees himself in the distance—the earlier version of himself during the first trip. Then, suddenly, the time traveler meets a violent death: "As his panicky lungs filled with air for the last time, he knew what animal had screamed in the depth of the Coal Measure forest." In fact, self-encounters had appeared in science fiction *years* before Gödel's paper. In the 1942 story "Minus Sign" (*Astounding Science Fiction*, November) by Jack Williamson, for example, a spaceship battles with itself while traveling backward in time. How do you think a scientist like Gödel would have liked these two stories? (Who knows, maybe he *did* read them!) If you could travel back in time to 1949 to ask him if such tales had been an inspiration, do you think he would be intrigued, amused, or instead would he be insulted?

Contents

1 A Broad Look at Time Travel 1
1.1 Time Travel in the Fantasy and Science Fiction Literature 1
1.2 Where Are All the Time Travelers? 10
1.3 Skepticism About Tales of Time Travel 15
1.4 Troubles with (some) Time Machines 22
1.5 Quantum Gravity, Singularities, Black Holes, and Time Travel 29
1.6 Tipler's Time Machine 38
1.7 For Further Discussion 42

2 Philosophical Space and Time 51
2.1 Time: What Is It, and Is It Real? 51
2.2 Linear Time and the Infinity of Past and Future 61
2.3 Cause and Effect 67
2.4 Backward Causation 72
2.5 The Fourth Dimension 78
2.6 Spacetime and the Block Universe 90
2.7 Philosophical Implications of the Block Universe 100
2.8 For Further Discussion 109

3 The Physics of Time Travel: Part I 115
3.1 The Direction of Time 115
3.2 The Arrows of Time 124
3.3 Time Dilation 139
3.4 The Lorentz Transformation 146
3.5 Spacetime Diagrams, Light Cones, Metrics, and Invariant Intervals 155
3.6 Proper Time and the Twin Paradox in Time Travel to the Future 173
3.7 For Further Discussion 181

4 Philosophers, Physicists, and the Time Travel Paradoxes 187
- 4.1 Paradoxes and Their First Appearance in Science Fiction 187
- 4.2 Changing the Past and the Grandfather Paradox 195
- 4.3 Changing Versus Affecting the Past 206
- 4.4 Causal Loop and Bootstrap Paradoxes 214
- 4.5 Sexual Paradoxes 225
- 4.6 Splitting Universes and Time Travel 229
- 4.7 For Further Discussion 237

5 Communication with the Past 245
- 5.1 Reversed Time 245
- 5.2 Multi-dimensional Time 252
- 5.3 Maxwell's Equations and Sending Messages to the Past 256
- 5.4 Wheeler and Feynman and Their Bilking Paradox 264
- 5.5 Absorber Theory and Signaling the Past 269
- 5.6 Tachyonic Signals and the Bell Quantum Antitelephone 273
- 5.7 For Further Discussion 283

6 The Physics of Time Travel: II 289
- 6.1 Faster-than-Light into the Past 289
- 6.2 Tipler's Rotating Cylinder Time Machine 297
- 6.3 Thorne's Wormhole Time Machine 301
- 6.4 Gott's Cosmic String Time Machine 319
- 6.5 Cutting and Warping Spacetime 326
- 6.6 For Further Discussion 334

Appendix A: Old Friends Across Time (A Story) 339

Appendix B: Newton's Gift (A Story) 347

Appendix C: Computer Simulation of the Entropic Gas Clock 355

Epilogue 357

Glossary 367

Index 377

About the Author

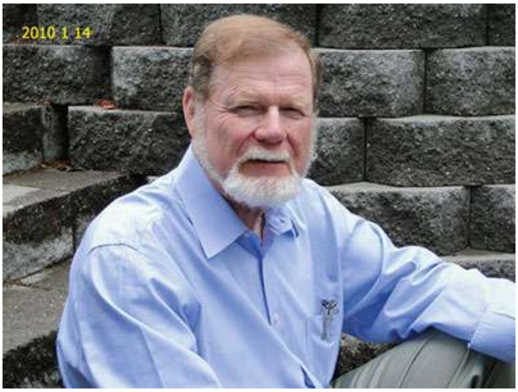

Paul J. Nahin was born in California, and did all of his schooling there, including a PhD in electrical engineering from UC Irvine (1972). He is now emeritus professor of electrical engineering at the University of New Hampshire. Prof. Nahin has published a couple of dozen short science fiction stories in *Analog*, *Omni*, and *Twilight Zone* magazines, and has written 18 books on mathematics and physics. He has given invited talks on mathematics at many prestigious institutions, and has appeared on National Public Radio's "Science Friday" show (discussing time travel) as well as on New Hampshire Public Radio's "The Front Porch" show (discussing imaginary numbers). He advised Boston's WGBH Public Television's "Nova" program on the script for their time travel episode.

Chapter 1
A Broad Look at Time Travel

"Hold infinity in the palm of your hand and eternity in an hour."

—William Blake, writing in "Auguries of Innocence" (1863), with words that could quite well describe what it would be like to time travel

"I need a place to hide, that's why I believe in yesterday."

—The Beatles (*Yesterday*, 1965)

1.1 Time Travel in the Fantasy and Science Fiction Literature

> "Woodn't it be grate to go back in tyme and correct your mistakes? Wouldn't it be great to go back in time and correct your mistakes?"
>
> —motto of *Time Twisters* comics

To travel in time.

Could there possibly be a more exciting, more romantic, more wonderful adventure than that? I don't think so, and in this opening section I want to just briefly discuss how fascinating many writers (and their readers) have found the concept of time travel, and to point out that the fascination began *long before* mathematical physicists discovered time travel lurking in Einstein's general theory of relativity.[1]

Before the arrival of humans on the surface of the Moon in 1969, the only other 'fantastic voyage' that could compare with time travel was traveling through outer space. During the seventeenth and eighteenth centuries, in fact, such voyages were the center of a genre of fiction (now called *science fiction*) called the "imaginary voyage" or "extraordinary voyage." Marjorie Hope Nicolson's[2] 1948 book *Voyages*

---

[1] An excellent, book-length *literary* treatment of time travel is by David Wittenberg, *Time Travel: the popular philosophy of narrative*, Fordham University Press 2013.

[2] Marjorie Hope Nicolson (1894–1981) was a literary scholar of the first rank at both Smith College and Columbia University.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_1

to the Moon carefully documents just how popular that form of literature was—and still is. Since 1969 the first such voyages have become history, of course, and time travel has replaced space travel as the modern "imaginary voyage."

It seems a safe bet that that, given a random selection of middle-aged adults, the vast majority of them would respond enthusiastically if asked whether time travel interests them. This fascination with time travel has actually been 'scientifically' documented. In one intriguing study,[3] several hundred men and women were asked to consider the possibility of spending an hour, a day, and a year back in both their personal past (since birth) and their historical past (before birth). They were further told that it would cost $10,000 to purchase such time travel services. Their response indicated that 10 % would be willing to spend that much money for an hour in the historical past, 22 % for a day, and 36 % for a year. As might be expected, the numbers rose as the cost dropped and, if such trips were free the interest was almost universal. As one writer put it, "Time travel [is] the ultimate fantasy, the scientific addition to the human quest for immortality."[4] And as a philosopher observed, "[T]he popular appeal of time travel . . . is no doubt due to a nostalgia for the past, which is almost an omnipresent aspect of the human condition."[5]

Fiction writers have, for *centuries*, recognized the fantasy appeal of time travel. The common fairy tale theme of 'The Three Wishes,' in which the recipient ends up using the final wish to undo the unforeseen consequences of the first two, is the precursor to all modern change-the-past time travel stories. Indeed, the means of time travel in the Norwegian poet Johan Wessel's 1781 play *Anno 7603 is* a fairy. Some of the best modern science fiction stories have played with the fantasy appeal of time travel by having gifts arrive by accident from the future: the moral of such tales is generally that unearned gifts usually bring grief.[6] The editor's introduction to a time travel story involving the Civil War referred to the fantasy aspect of time travel—with a reference to another age-old adult fantasy—when he wrote "time travel stories about the Civil War have one thing in common with pornography; they serve to titillate an impulse [in the case of time travel stories, the impulse to change history] and to frustrate [history]."[7] This is the motivation for the time traveler in Stephen King's 2011 novel *11/22/63*, who uses what appears to be a naturally occurring wormhole (connecting a Maine diner to 1958) for his attempts at preventing the assassination of John F. Kennedy.

---

[3] T. J. Cottle, "Fantasies of Temporal Recovery and Knowledge of the Future," in *Perceiving Time*, John Wiley 1976.

[4] T. Paul, "The Worm Ouroboros: Time Travel, Imagination, and Entropy," *Extrapolation*, Fall 1983, pp. 272–279.

[5] J. W. Smith, "Time Travel and Backward Causation," *Cogito* 1985, pp. 57–67.

[6] Among the many such tales, five particularly good ones are "Something for Nothing" by Robert Sheckley, "Mimsy Were the Borogoves" by Lewis Padgett (pseudonym of the married couple Henry Kuttner and C. L. Moore), "Child's Play" by William Tenn (pseudonym of Philip Klass), "The Little Black Bag" by Cyril Kornbluth, and "Thing of Beauty" by Damon Knight. All can be found in various anthologies.

[7] *The Fantastic Civil War* (F. McSherry, Jr., editor), Baen 1991.

A character in the 1985 novel *The Bird of Time* by George Effinger nicely captures the fantasy appeal of time travel with the declaration "The past . . . is the home of romance." On a less poetic level, time travel and the movies and stories about it fascinate most people because they turn our everyday world view upside down and inside out.[8] Such movies and stories make people *think*. It is therefore not surprising that time travel movies have been popular for decades, from the pioneering *Berkeley Square* in 1933, to the classic 1960 filming of *The Time Machine*, to *Back to the Future* in 1985 (*the* top film that year in a *Boxoffice* magazine poll), to the flawless 1989 *Bill & Ted's Excellent Adventure*, to the clever *Terminator* action films, to the ingenious 2012 *Looper*, to the commercially successful 2014 *Interstellar*. Each of these films, and others, too, will be discussed later in the book.

When we discuss time travel, we should really be careful to distinguish between two quite different versions: to the future, and to the past. There is no dispute, today, about the first. As two *severe critics* of the possibility of time travel to the past wrote decades ago, "After 1900, special relativity made scientific discussion of time machines possible."[9] What they were referring to is the fact that, by traveling in a rocket ship fast enough (but never, unlike Superman, faster than the speed of light), and far enough, one could leave Earth, loop out on a vast journey perhaps halfway across the universe, and then return hundreds, thousands, even millions of years in the future. You could theoretically (ignoring all the *engineering* difficulties) do this, in fact, with the apparent passage of your 'personal time,' (as measured by your wrist watch or the beating of your heart) as brief as you'd like. (Physicists call 'personal time' *proper time*, and I'll return to this in Chap. 3.) This astonishing conclusion from special relativity, that time travel to the future makes physical sense, literally put a lot of Victorian-era trained physicists into shock.

A quite sophisticated use of this idea appeared early in science fiction, in the tale of a space traveler who returns from a high-speed trip out to the blue supergiant star Rigel in the constellation Orion.[10] The 900 or so light years of the round trip had taken just 6 months of ship or personal time (proper time), but a thousand years of back-home time. The traveler returns to Earth to find all he had left behind long dead and returned to dust: "Sometimes I waken from a dream in which they are all so near . . . all my old companions . . . and for a moment I cannot realize how far away they are. Beyond years and years."

Another story[11] of a trip into the future that delivers an equally powerful emotional impact, this time via a Wellsian-type time machine (more on what that means later in this chapter) rather than by rocket travel, tells of a time traveler trapped in post-nuclear war times, where there is no energy available to power his machine for the return trip. As the story ends, he finds the woman he had loved and left behind in the past. She is now the elderly widow of another man, having married his rival because the time traveler (just like Wells' Time Traveller) never returned.

---

[8] Somewhat more pompous (but no less correct) was this observation by an academic: "The time-travel [film] romance is an attempt to reenchant the world, to regain a sense of belongingness, to reinstate the magical, autocentric Universe of the child and the primitive." See W. Wachhorst, "Time Travel Romance on Film," *Extrapolation*, Winter 1984, pp. 340–359.

[9] S. Deser and R. Jackiw, "Time Travel?" *Comments on Nuclear and Particle Physics*, September 1992, pp. 337–354.

[10] R. H. Wilson, "Out Around Rigel," *Astounding Stories*, December 1931.

[11] W. Tucker, *The Year of the Quiet Sun*, Gregg 1979.

A trip into the future does not *have* to be serious or sad. A nice example of that is the story of a spaceship crew that sets off for the Alpha Centauri triple star system, more than four light years distant.[12] They survive the trip, which requires 500 years of both personal *and* external time (this story is the *only one* I'll mention in this book that uses *a preserving drug* rather than physics). What causes me to include it in a book on *time travel* (stories in which proper and external time are one-in-the-same are simply *not* about *time travel*) is that, long before they arrive at the end of their journey, the secret of faster-than-light (FTL) travel is discovered back on Earth and so they arrive at their destination to find a human reception committee! As you'll see when we get to Chap. 3, knowledge of FTL travel is equivalent to knowing the secret of travel into the past, and so the crew is sent back in time, to Earth, to just one year after they left—and they listen to their own radio communications arriving from deep space.

The *real* adventure in time travel, as suggested by "Far Centaurus," would be to go backward in time, to visit the past. The editor of the science fiction pulp magazine *Thrilling Wonder Stories* used the powerful emotional hook of changing the past in a 1950 blurb announcing a time travel story coming in the next issue: "What's the biggest mistake you ever made? Don't worry about it. You may have pulled some awful boners in your time, but there's a sure-fire remedy for them all. It's simple. Just look up at that old time-clock on the wall—and turn it back to the moment just preceding your terrible blunder. Then make your corrections—and set your time-clock back to the present. You may be starting a new chain of error, but why fret? You can go back in time again..." Or, as the promotional text on the video package of the 1986 movie *Peggy Sue Got Married* says, "to do it again" is "the golden opportunity almost everyone has longed for at least once."

Writing less romantically, a philosopher declared that a "major source of interest in the time travel question is our general fascination with the exotic and the child-like frustration we sometimes feel at being confined to the present. We wish that the benefits of moving through space could be supplemented with the benefits which would accrue from movements through time."[13] Robert Silverberg, a science fiction writer who has used the time travel theme often and effectively, expressed this sentiment quite clearly when he wrote "Suppose you had a machine that would enable you to fix everything that's wrong in the world ... The machine can do anything ... it gives you a way of slipping backward and forward in time ... Call this machine whatever you want. Call it Everybody's Fantasy Actualizer. Call it a

---

[12] A. E. Van Vogt, "Far Centaurus," *Astounding Science Fiction*, January 1944.

[13] R. A. Sorenson, "Time Travel, Parahistory and Hume," *Philosophy*, April 1987, pp. 227–236.

Time Machine Mark Nine."[14] He gives a masterful demonstration of what he means by that in one of his own stories, a tale[15] set in a year when time machines actually exist. Even so, the characters use their imaginations to explore their fantasy worlds and wishes—wishes that could (if they *really* wanted to) be realized with a real time machine. Time travel fiction is, you see, the ultimate escapist literature!

The adventure promised by time travel to the past doesn't necessarily mean *pleasant* adventure, and science fiction has used that idea to great effect. The unstated horror of a trip backward in time, if you think just a bit about it, is that it would bring the dead past, filled with all its dead occupants, alive again, literally resurrected from dank and moldering graves. The top of Mount Everest, the bottom of the Marianas Trench, the sands of Mars—none of these exotic places can even be mentioned in the same breath with the *past*. The capture of the mystery and, yes, the sheer terror of the past, is in this opening line to a 1950s tale: "When Dr. Flitter came into the room, it seemed as though the past and its dead people came in with him, clinging to him like stale surgery smells, like the cold sweat of ancient autopsies."[16] In another equally macabre story, we read of a time traveler in the past anticipating a meeting with a long-dead lover as he "shivered with a renewal of horror . . . She ought to be grateful to him for having raised her from the dead, even briefly."[17]

Another tale, slightly less gruesome, tells us of a character who delights in pointing out all the bad aspects of living in the past.[18] Tell him *when* in time, and he quickly ticks off the disadvantages of being *then*. To live in ancient Greece would let you rub shoulders with Aristotle, sure, but you already know what he said and you'd soon regret the lack of modern plumbing. The year of the American Revolution might let you exchange greetings with George Washington, but you'd also have to put up with cholera in Philadelphia, malaria in New York, and the fact that if you needed an operation there would be no anesthesia anywhere. The Victorian Age appeals to modern romantics but, before you go back, you'd better have your eyes and teeth checked. The time traveling historian in one novel[19] takes these medical warnings to heart and has her appendix prophylactically removed, and is further advised to have her nose cauterized against all the awful stinks of her destination, the fourteenth century.

On the other hand, poor health care isn't *all* there is with time travel to the past. One time traveler from the future, for example, makes a very good living in the past by winning bets on yet-to-happen events whose outcomes he knows.[20] Time travel to the past allows a Parisian curio shop in the present to offer remarkably authentic looking newspapers from 1804 whose only 'flaw' is that they appear to be fresh off the press—which of course they are![21] And the failed professor in one romantic story[22] finds the Paris of 1482 infinitely better than the Paris of 1961.

---

[14] R. Silverberg, "Ms. Found in an Abandoned Time Machine," *Beyond the Safe Zone*, Warner 1986.

[15] R. Silverberg, "Many Mansions," *Beyond the Safe Zone*, Warner 1986.

[16] R. Bretnor, "The Past and Its Dead People," *Magazine of Fantasy and Science Fiction*, September 1956.

[17] L. Marlow, *The Devil in Crystal*, Faber and Faber 1944.

[18] A. Bester, "Hobson's Choice," *Magazine of Fantasy and Science Fiction*, August 1952.

[19] C. Willis, *Doomsday Book*, Bantam 1992.

A popular fictional appearance of time travel is the use of the past as a hiding place, as a sanctuary for those wishing to escape the troubles of modern times.[23] An interesting twist on this idea was presented in one tale[24] in which the past is used for later military gain in the present. In this story the Earth of a thousand years in the future is ruled by a dictator, and the oppressed masses are unable to arm themselves for revolt. So, back into the past travels an agent to arrange for the construction of weapons, which are then stockpiled in hidden caverns where they can be retrieved for use ten centuries later. The past is used in this story as both a sanctuary *and* a repository from which to make war in the future, and so we have a time travel fantasy for both doves *and* hawks in the same tale! This military use of the past is passive; other writers have more aggressively used time travel to the past for military gain as, for example, mining uranium deposits before they have had time to reduce themselves to lead via radioactive decay,[25] or in drilling for Middle East oil in the past to deprive adversaries of it in the present.[26]

Time travel to the past would, perhaps, interest criminals, too. As the science fiction writer Larry Niven wrote, "If one could travel in time, what wish could not be answered? All the treasures of the past would fall to one man with a submachinegun. Cleopatra and Helen of Troy might share his bed, if bribed with a trunkful of modern cosmetics."[27] Or, as the tragically flawed inventor of the first time machine dreamed, before using time travel to commit what he thought would be the perfect locked-room murder, "The Great Harrison Partridge would have untold wealth. He could pension off his sister Agatha and never have to see her again. He would have untold prestige and glamour, despite his fat and baldness, and the beautiful and aloof Faith Preston would fall into his arms like a ripe plum."[28]

---

[20] C. Sprague, "Time Track," *Startling Stories*, January 1951.

[21] M. Leinster, *Time Tunnel*, Pyramid Books 1964.

[22] U. K. Le Guin, "April in Paris," *Fantastic Stories*, September 1962.

[23] There are *many* excellent examples of such tales, a few of which are Clifford Simak, "Over the River & Through the Woods"; Ray Bradbury, "The Fox and the Forest"; Jack Finney, "Such Interesting Neighbors"; J. B. Priestly, "Mr. Strenberry's Tale"; James Gunn, "The Reason Is With Us"; and H. B. Piper, "Flight from Tomorrow." All can be found in various anthologies.

[24] R. F. Young, "Not to be Opened—," *Astounding Science Fiction*, January 1950.

[25] C. Simak, "Project Mastodon," *Galaxy Science Fiction*, March 1955.

[26] P. Anderson, "Wildcat," *Magazine of Fantasy and Science Fiction*, November 1958, and W. Jeschke, *The Last Day of Creation*, St. Martin's Press 1982.

[27] L. Niven, "The Theory and Practice of Time Travel," in *All the Myriad Ways*, Ballantine 1971.

Instead of viewing the past as an aid to crime, some writers have used it as the perfect dumping ground for criminals, as a highly convenient place to remove them from society.[29] After all, there can be no breakout from the prison of the past—at least not without a time machine. What might happen to criminal recidivists in a world that has mastered time travel is nicely explained in one story as follows: "If you cannot live among people, then off to the reptiles—one hundred or one hundred twenty million years before the present. There you wouldn't freeze in a tropical pre-glacial climate, and you could nourish yourself on plants. But there is no one to talk with, boredom, and in the end you offer yourself up as an afternoon snack to a tyrannosaurus."[30] With an interesting twist on this is the tale[31] of a physics professor who *helps* criminals disappear into the past to escape relentless police pursuit.

Museum curators, too, would seem to be obvious clients for time machine companies, as would collectors of extinct species who work for zoos. An example of the first case is in a novel[32] about a time travel business called *Time Researchers*, with the corporate mottos 'We Sift the Sands of Time.' It works as a futuristic version of Indiana Jones, as finders of lost historical artifacts for customers who can pay the substantial charges. A typical mission is to make original sound recordings of one of Lincoln's unreported speeches. We find the same idea in a short story[33] about a business called *Genealogy, Inc.*, with the corporate motto

"An Ancestor for Everybody." It uses a 'time scanner' to provide its clients with a list of distinguished predecessors. And in another novel we read of the Historical Corps, whose time travel agents are "writing the definitive history of mankind.".[34]

Another use for time travel to the past, one of the most unusual I have seen, was suggested in a philosophical article[35] which considers an age-old question that has long bedeviled schoolboys: “If everything in the Universe doubled in size overnight while we slept, could we tell what had happened when we woke up next morning?” The usual answer to this puzzle (called the *Universal nocturnal expansion* by philosophers) is *no*, but the authors of the article suggest that ‘all’ we need do is take a yardstick back to yesterday and compare it with itself! Great idea, for sure, but it had appeared years earlier in a science fiction story.[36]

---

[28] A. Boucher, "Elsewhen," *Astounding Science Fiction*, January 1943. Partridge's dream is shattered, however, because he overlooks a few details about time travel, ones that he wouldn't have missed if he could have read this book. We'll come back to this classic story, which merges a time machine with murder, later in the book.

[29] This is a popular science fiction scenario, and three of the best stories playing with it are P. Anderson, "My Object All Sublime," *Galaxy Science Fiction*, June 1961; I. Watson, "In the Upper Cretaceous with the Summerfire Brigade," in *Stalin's Teardrops*, Victor Gollancz 1991; and R. Silverberg, "Hawksbill Station," *Galaxy Science Fiction*, August 1967.

[30] S. Gansovsky, "Vincent Van Gogh," in *Aliens, Travelers, and Other Strangers*, Macmillan 1984.

[31] J. Finney, "The Face in the Photo," in *About Time*, Simon and Schuster 1986.

[32] W. Tucker, *The Lincoln Hunters*, Rinehart 1958. See also A. Bitov, "Pushkin's Photograph," in *The New Soviet Fiction*, Abbeville Press 1989.

[33] M. Shaara, "Man of Distinction," *Galaxy Science Fiction*, October 1956.

[34] L. A. Frankowski, *The Cross-Time Engineer*, Del Rey 1986. The same idea is in the 1991 film comedy *The Spirit of '76*, in which time-traveling historians from 2176 visit the past in an attempt to reconstruct the lost records of the founding of America.

[35] J. C. Graves and J. E. Roper, "Measuring Measuring Rods," *Philosophy of Science*, January 1965, pp. 39–56.

More ingenious uses for the past are discussed in the story[37] of a time travel business called *Time Associates*. One use comes in the form of a request from a United States senator who wants to send the disadvantaged of today back into the remote past, where they could have a fresh start on a virgin Earth. Yet another use comes from a religious fringe group that wants to purchase exclusive rights to the time of Jesus—not to visit, but to prevent *anyone* from visiting. The group fears that any such visitors would “learn the truth,” which might contradict the very legends that form the heritage of Christianity. And, in what may be the most ingenious idea of all, *Time Associates* itself does not do business in the present, but rather 150,000 years in the past, in a ‘new’ country called Mastodonia. The corporate lawyer, you see, has determined that such an arrangement legally means the company is a *foreign* company doing business outside the United States, and so it is not liable for taxes to the IRS!

The tourist trade is a booming business in science fiction, with dinosaur hunting at the top of the list. There are many such tales,[38] including the cerebral stories in L. Sprague de Camp’s short-story collection *Rivers of Time*, starting with the classic “A Gun for Dinosaur.” The earliest (that I know of) fictional use of time travel to the past for hunting was not for dinosaur hunting, however, but rather for saber-toothed tigers, wooly mammoths, and cave bear.[39] Historical tours to the great events of the past are also an entertaining use of time travel.[40] Even mundane events may one day be on the ‘to do’ lists of time travelers to the past. In one, for example, curious crowds from the very far future show up, nightly inside the home a twentieth century family, much as tourists today visit Monticello.[41]

Perhaps the most direct use of the past’s unique resource, itself, is realized in science fiction by Hollywood. In one tale,[42] after purchasing the motion picture rights to H. G. Wells' *Outline of History*, the head of a movie studio uses a time machine to send his ace cameraman into the past to get live action footage. Prehistoric animals, the ice age, Cheops building his pyramid, the destruction of Pompeii by the eruption of Vesuvius, the Battle of Hastings, Columbus, all the *originals* of these historical events appear in the final film. Years later this idea was developed even further in a very funny novel,[43] in which a movie director uses the eleventh century as a realistic setting for a picture. Realism isn't always the result, however, as portrayed in another tale; the films produced by a gadget that can 'look' into the past are failures because they don't look "authentic enough" to Hollywood moguls![44] (Fig. 1.1).

---

[36] H. M. Sycamore, “Success Story,” *Magazine of Fantasy and Science Fiction*, July 1959.

[37] C. Simak, *Mastodonia*, Ballantine 1978.

[38] Two are B. W. Aldiss, “Poor Little Warrior,” in *The Science Fictional Dinosaur*, Avon 1982, and a famous one by Ray Bradbury, “A Sound of Thunder,” in *The Stories of Ray Bradbury*, Alfred A. Knopf 1980.

[39] C. Simak, “The Loot of Time,” *Thrilling Wonder Stories*, December 1938, in which a time machine inventor raises money for his research by transporting hunters back 70,000 years to the Old Stone Age.

[40] Three such tales are R. Silberberg, *Up the Line*, Ballantine 1969 and “When We Went to See the End of the World,” in *Beyond the Safe Zone*, Warner 1986; and G. Kilworth, “Let’s Go to Golgotha!,” in *The Songbirds of Pain*, Victor Gollancz 1985.

[41] B. Tucker “The Tourist Trade,” in *Tomorrow the Stars*, Doubleday 1952.

[42] L. Laurence, “History in Reverse,” *Amazing Stories*, October 1939.

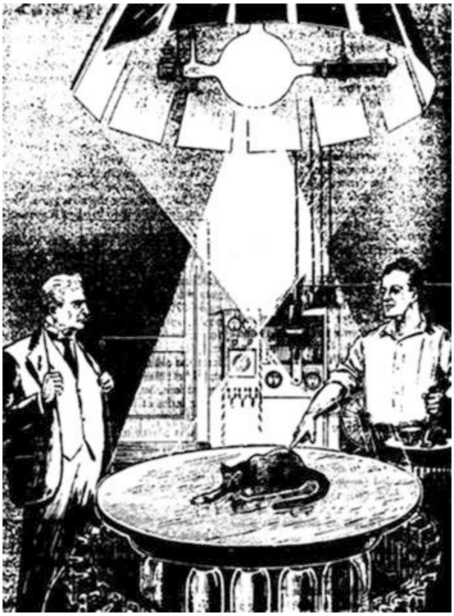

Fig. 1.1 The inventor of a time machine demonstrates it by sending the family cat on a trip. In the story the inventor, himself, travels back to 1901, where he accidently kills his grandfather in an early pulp magazine, non-paradoxical version of the famous riddle. Illustration for Raymond A. Palmer's "The Time Tragedy" (*Wonder Stories*, December 1934) by Frank R. Paul, ©1934 by Continent Publications Inc.; reprinted by permission of the Ackerman Science Fiction Agency, 2495 Glendower Ave., Hollywood, CA 90027 for the Estate

---

[43] H. Harrison, *The Technicolor Time Machine*, Doubleday 1967.

[44] A. Derleth, "An Eye for History," in *Harrigan's File*, Arkham House 1975.

1.2 Where Are All the Time Travelers?

> “If it [time travel] *could* be done, someone will eventually learn how. If that happens, history would be littered with tourists. They’d be *everywhere*. They’d be on the *Santa Maria*, they’d be at Appomattox with [cameras], they’d be waiting outside the tomb, for God’s sake, on Easter morning.”[45]

The question the title of this section asks is an echo of the one the physicist Enrico Fermi (1901–1954) asked in the 1950s, about the possibility of interstellar space travel and of alien intelligent life in the universe—if such travel is possible and ‘they’ exist, then *where are they*? Why haven’t we at least received radio signals from them? For many, the apparent lack of time travelers among us is similar evidence for the impossibility of time travel. As one famous science fiction writer put it, “The most convincing argument against time travel is the remarkable scarcity of time travelers. However unpleasant our age may appear to the future, surely one would expect scholars and students to visit us, if such a thing were possible at all. Though they might try to disguise themselves, accidents would be bound to happen—just as they would if we went back to Imperial Rome with cameras and tape recorders concealed under our nylon togas. Time traveling could never be kept secret for very long.”[46]

Clarke’s idea is that, from the moment after the first time machine was constructed, through all the rest of civilization, there would be numerous historians, to say nothing of weekend sightseers, who would want to visit every important historical event in recorded history. They might each come from a different time in the future, but all would arrive (according to Clarke) at destinations crowded with temporal colleagues, crowds for which there is no historical evidence!

Long before Clarke the science fiction writer Robert Silverberg had already used the same idea in his 1969 novel *Up the Line*, where it’s called the *cumulative audience paradox*. That paradox claims that as time travelers to the past continue to visit certain historically interesting dates and places, there will be an ever-increasing number of people present. As it is presented in the novel, “Taken to its ultimate, the *cumulative audience paradox* yields us the picture of an audience of billions of time-travelers piled up in the past to witness the Crucifixion, filling all the Holy Land and spreading out into Turkey, in Arabia, even to India and Iran . . . Yet at the original occurrence of [that event] *no such hordes were present!*” And later in the same work, we read “A time is coming [when we] will throng the past to the choking point. We will fill our yesterdays with ourselves and crowd out our own ancestors."

---

[45] A skeptic’s reaction to the idea of time travel in J. McDevitt’s story “Time’s Arrow” in *The Fantastic Civil War* (see note 7).

[46] Arthur C. Clarke, “About Time,” in *Profiles of the Future*, Warner 1985. A story that I recall once having read (but cannot now remember either the author or the title) wonderfully illustrates Clarke’s point. A time traveler in disguise at Golgotha for the Crucifixion has a camera hidden beneath his robe to avoid attracting attention. All goes well until he notices odd, clicking noises coming from all those standing near him. It is then he realizes the entire crowd is *nothing but* time travelers, from all through the ages, all with hidden cameras!

Philosophers are well aware of Silverberg's and Clarke's conundrum and, indeed, it can be found in the philosophical literature before Clarke wrote. In one paper, for example, we read "Actually I know of only one argument against the possibility of time travel that seems to carry any weight at all. This is the fact that it does not appear ever to have happened. That is, it might be argued that there will be no time trips from [2100] to [2017] because we were here in [2017] and saw no time travelers. But this argument is far from conclusive."[47] At most, in other words, the absence of temporal visitors amongst us is an objection to the *actuality* of time travel, and not to the *possibility* of time travel.

This same philosopher then mentions some ways around this concern, including one which he called a "pettifogging physical limitation on time travel: perhaps the energy expenditure varies as the fourth power of the time traversed, making only very short trips feasible, and its discovery lies too far in the future for its effects to have yet been felt." Another science fiction writer, as famous as Clarke, used that idea in his 1957 novel *The Door Into Summer* when Robert Heinlein has one character comment "Now if there was some way to photograph the Crucifixion . . . but there isn't. Not possible . . . there isn't that much power on the globe. There's an inverse-square law tied up in [time travel]." Or, perhaps, time travel *is* possible but it's so extraordinarily dangerous that it's impossible to get anyone to do it. In one provocative tale[48] that takes this idea to the extreme, we read that there is only *one* time traveler, ever, from the future—indeed, from just 18 min (!) in the future—and his first (and last) experiment destroys the Earth.

Clarke presented some other possible science fiction rebuttals to the puzzle of 'where are the time travelers?' As he wrote, "Some science fiction writers have tried to get around this [question] by suggesting that Time is a spiral; though we may not be able to move along it, we can perhaps hop from coil to coil, visiting so many millions of years apart that there is no danger of embarrassing collisions between cultures. Big game hunters from the future may have wiped out the dinosaurs, but the age of *Homo sapiens* may lie in a blind region which they cannot reach."

The idea of time as a spiral was quite popular in early science fiction. Typical is one tale[49] in which the time traveler suddenly finds himself not in 1933 but in 2189. His situation is 'explained' to him thus: "[The] time stream is curved helically in some higher dimension. In your case, a still further distortion brought two points of the coil into contact, and a sort of short circuit threw you into the higher curve."

---

[47] G. Fulmer, "Understanding Time Travel," *Southwestern Journal of Philosophy*, Spring 1980, pp. 151–156. The modern view of this 'paradox' is *not* that it describes a situation so absurd that time travel must be impossible, but rather that *all* the time travelers who *were* (will be?) at the Crucifixion *are* in the historically recorded crowd (see note 46 again). The Crucifixion happened just *once*, not over and over. I'll return to this point later in the book.

[48] D. Plachta, "The Man from When," *Worlds of If Science Fiction*, July 1966.

[49] R. H. Wilson, "A Flight Into Time," *Wonder Stories*, February 1931.

A few years later (in 1937) we find another story[50] with the same spiral-time concept; with a sixty-million year pitch to the time helix, there is no danger of a grandfather paradox. That same year spiral time was the central ‘scientific’ theme in the stage production “I Have Been Here Before” by the English playwright J. B. Priestley.

The very next year (1938) the young Isaac Asimov used the idea in his first attempt at professional writing, despite a life-long unhappiness with the concept of time travel. Titled “Cosmic Corkscrew,” it was initially rejected, indeed it was *never* published, and eventually lost. Though, perhaps, not for long, since one enterprising modern writer has used it as the basis for his own time travel tale.[51] In it a traveler from the near future travels back to 1938 to retrieve “Cosmic Corkscrew” before Asimov loses it (perhaps that’s *why* it was lost!) Even when writing introductions to other writers’ time travel tales, Asimov would often insert personal comments on his opinion of the concept. For example, in one volume of an anthology series he edited, *The Great Science Fiction Stories* (of 1954), he wrote “To my way of thinking it is precisely because time travel involves such fascinating paradoxes that we can conclude, even in the absence of other evidence, that time travel is impossible.” And in *The Great Science Fiction Stories* (of 1961) he bluntly declared “I think scientists who think up methods of time travel are probably all wrong.”

Spiral time is a close cousin to circular time. One story[52] dealing with circular time has a time traveler who finds, after a trip one hundred years into the future, that he can’t get all the way back to his own time because the required energy rises exponentially with increasing penetration into the *past*. Still, it’s very cheap in energy to go forward in time in this tale, and that’s what the traveler does, in search of help from the future’s advanced technology. He never finds what he needs, however, and so goes forward right into the collapse of the universe and through a new Big Crunch that forms an identical new cycle of time. He thereby returns home to just before he left.[53] This eternal recycling of identical, circular time is so terrifying that the traveler decides to suppress what he has learned about how to time travel—and so maybe *that’s* why there are no apparent time travelers. Science fiction writer Larry Niven has argued, however (see note 27), that while this may be a conceptually valid (?) way to travel into the past, he also warns that “Removing your time machine from the reaction of the Big Bang/Crunch could change the final configuration of matter, giving an entirely different . . . history.” (I strongly suspect that Niven wrote that with a big smile on his face!)

---

[50] P. S. Miller, “The Sands of Time,” *Astounding Stories*, April 1937.

[51] M. A. Burstein, “Cosmic Corkscrew,” *Analog*, June 1998.

[52] P. Anderson, “Flight to Forever,” *Super Science Stories*, November 1950.

[53] Turning this idea on its head is the approach of the 1978 novel *The Way Back* (DAW) by A. B. Chandler. Its characters return from the past to their own time by traveling even further backward, right through the Big Bang and into the *previous* (and identical) cycle of time.

The famed English physicist Stephen Hawking is so taken with the question of 'where are all the time travelers?" that he has elevated their apparent absence ("we have not been invaded by hordes of tourists from the future"[54]) to the status of being a demonstration of the impossibility of time travel to the past. His so-called *Chronology Protection Conjecture*, Hawking likes to say, "makes the universe safe for historians": that is, there is nothing to worry about (if you're concerned that time travelers could change the past) because time travel is simply impossible. You'll see later in the book that there are other possible ways to insure the safety of history *without* denying the possibility of time travel, and Hawking himself has backed away just a bit from the Conjecture, saying now that he was simply looking for a humorous line.

While Hawking's endorsement of it has made the Conjecture famous, he wasn't the first to state it. Two years earlier it had appeared, in of all places, a *financial* publication: "[If] time travel was possible, someone from the future would eventually either discover a time tunnel or build a time machine and come visit us."[55] And even before that, an economist presented a 'proof' for concluding that "time travelers do not and cannot exist."[56] He argued that if time travelers from the future were actually amongst us (our 'now' is their 'past') then, by virtue of their knowledge of things to come (our 'future') they would make financial deals so numerous and extensive that interest rates would be driven to zero. Interest rates are *not* zero, however, and thus no such time travel hanky-panky has occurred.

These sorts of financial arguments aren't like to convince many physicists or philosophers of the Conjecture's merit. At most we can only conclude from them that time travelers from the future have *not* influenced financial affairs, which doesn't mean they aren't here. In any case, the Conjecture was actually stated more than 20 years before Hawking by Larry Niven (see note 27), who declared what is called *Niven's Law*: "If the universe of discourse permits the possibility of time travel, and of changing the past, then no time machine will be invented in that universe." And Hawking's concern over time travelers meddling with the past was anticipated in science fiction, too, by at least half a century; in a 1950 tale, for example, we learn of a Master Historian, and the graduate students in his course on 'Experimental History' in the forty-sixth century, trying to correct a problem created by a previous tampering with the past![57]

Not all physicists and philosophers feel intellectually comfortable with the Conjecture, as it seems (to them) a too quick surrender: 'Time travel is a problem

---

[54] S. W. Hawking, "Chronology Protection Conjecture," *Physical Review D*, July 15, 1992, pp. 603–611. See also J. F. Woodward, "Making the Universe Safe for Historians: Time Travel and the Laws of Physics," *Foundations of Physics Letters*, February 1995, pp. 1–39.

[55] J. Queenan, "Time Warp: Or, Investing in the Future Is a Bust," *Barron's*, January 8, 1990, p. 46.

[56] M. R. Reinganum, "Is Time Travel Impossible? A Financial Proof," *Journal of Portfolio Management*, Fall 1986, pp. 10–12.

[57] L. Jones, "Sunday is Three Thousand Miles Away," *Thrilling Wonder Stories*, June 1950. A more recent, two-*novel* treatment of historians tinkering with history is by Connie Willis (*Blackout* and *All Clear*, both published in 2010).

so hard to do let's simply *define* it to be non-existent and then we won't have to worry about it anymore.' To *really* show that time travel is impossible, however, one needs to demonstrate how it would violate one or more of the laws of physics. Hawking, of course, understands this and has stated that, as one who is no fan of time machines and time travel, he believes there is new physics yet to be discovered that *will* forbid would-be time travelers from roaming up and down the centuries. Finding that new physics is the lure the study of time machines has for him. As he correctly writes in his autobiography,[58] "Even if it turns out that time travel is impossible, it is important that we understand why it is impossible."

One mathematical physicist who agrees with Hawking on the matter of the unlikely possibility of making a time machine is the New Zealand theoretician Matt Visser. Noting that while quantum field theory, and the general theory of relativity, are each amazingly good theories in many applications within their respective realms, they are not so good in spacetime regions at the so-called *Planck scale* (that is, when the density of mass-energy reaches the fantastic level of 10^{94} grams/cm^3 and beyond) where chronology violations (that is, time travel) seem to be spawned. As Visser has observed,[59] this situation won't change until 'we wander into the guts of quantum gravity,' the unification that will merge gravity with the quantum to give a theory that *always* works. Without quantum gravity, physics will continue to be "infested" (Visser's word) with "sick" (Visser's word) spacetimes that allow time travel. Visser believes that the discovery of the theory of quantum gravity can be 'guided' by *building* causality into it,[60] and the result will finally consign time machines to where (in his mind) they belong, the dust-bin of crackpot physics.

Well, perhaps so, but we don't have a quantum theory of gravity yet, and probably won't for some time to come, and so the puzzling questions about time travel remain. To end this section on a slightly gloomy note, an idea appeared in science fiction,[61] when Hawking was still a teenager, offering a possible rebuttal to the Conjecture. It opens with one of the inventors of the first time machine just returning from a trip to the past of 1938. Still, despite this success, the inventors are puzzled by what they call 'the problem': "But if *we* have time traveled, then obviously men of the future have time traveled. They will be able—*are* able to come back. [So] where are they?" They finally conclude that there can only be two possible answers. Either there is nobody in the future, or time travel is so dangerous (is that why the future might be empty—humanity misused time travel and killed itself off?) that all who invent it will suppress it. And that's what they decide *they* must do.

---

[58] Stephen Hawking, *My Brief History*, Bantam 2013, p. 113.

[59] Matt Visser, "The Quantum Physics of Chronology Protection," in *The Future of Theoretical Physics and Cosmology*, Cambridge 2003. This paper was Visser's contribution to the celebration of Hawking's 60th birthday, held in January 2002.

[60] This may seem like something new, but it really isn't. General relativity has causality built into it on a *local* level (where it belongs); a failure of causality (that is, time travel—see J. Sharkey, "The Trouble With Hyperspace," *Fantastic* April 1965) occurs in general relativity only when one studies *global* regions of certain spacetimes. Forcing a physical theory to have a *prescribed global behavior* would be to undo all of physics since the development of *local field theories*, along with all their amazing successes in explaining nature.

[61] M. Shaara, "Time Payment," *Magazine of Fantasy and Science Fiction*, June 1954.

1.3 Skepticism About Tales of Time Travel

> "May it not be that our inability to leap into the fiftieth century, A.D., seems impossible to us, merely because of certain prejudices we entertain or certain facts and tricks of which we are still hopelessly ignorant? Assuredly, this is not a foolish query. Its answer, whatever that may be, carries immeasurable consequences for metaphysics."
>
> —a scholar wonders[62]

A thought-provoking possibility for explaining the scarcity of certified time travelers is the central thesis of a fascinating paper in the philosophical literature. The author of that paper argues (note 13) that nobody would believe a time traveler even if he willingly confessed and revealed his knowledge of the future, or even gave the details of his time machine. He goes on to make the astonishing assertion that even the *time traveler himself* would have doubts! This perhaps shocking suggestion deserves some elaboration, especially because it invokes the authority of the patron saint of skeptics for support, the Scot David Hume (1711–1776). The crucial point to keep in mind is explicitly stated in the argument: "The key question will not be 'Is time travel possible?' We shall instead ask whether it is possible to justify a belief in a report of time travel." This gets to the real heart of Clarke's puzzle from the previous section.

Much of the resistance to the idea of time travel lies in sheer skepticism. For many, time travel (to the past, in particular) is simply too much out of the ordinary to be taken seriously. For many, time travel would literally be miraculous. Hume's great work, *An Enquiry Concerning Human Understanding*,[63] contains a section on how a rational person should react to a claim that a miracle has occurred. Hume proclaimed that a miracle *by definition* violates scientific law and that, because such laws are rooted in "firm and unalterable experience," any violation of one or more of these laws immediately provides a refutation of the report of a miracle. In Hume's own words:

> "Nothing is esteemed a miracle, if it ever happened in the common course of nature. It is no miracle that a man, seemingly in good health, should die on a sudden; because such a kind of death, though more unusual than any other, has yet been frequently observed to happen. But it is a miracle, that a dead man should come to life; because that has never been

---

[62] W. B. Pitkin, "Time and Pure Activity," *Journal of Philosophy, Psychology and Scientific Methods*, August 27, 1914, pp. 521–526. Pitkin's essay was a critique of time travel as presented in Wells' *The Time Machine,* which Pitkin called "one of the wildest flights of literary fancy."

[63] Making its first appearance in 1748, *Enquiry* has been reprinted numerous times since. I used the 1963 edition published by Open Court.

> observed in any age or country . . . When anyone tells me, that he saw a dead man come to life, I immediately consider with myself, whether it be more probable, that this person should either deceive or be deceived, or that the fact, which he relates, should really have happened. I weigh the one miracle against the other; and according to the superiority, which I discover, I pronounce my decision, *and always reject the greater miracle* [my emphasis]."[64]

It is a strict interpretation of Hume that Sorenson (note 13) has adopted in claiming that a time traveler would have no success (among rational persons) with tales of 'different times.' As he explains, "Clearly the time traveler cannot persuade a reasonable person by baldly asserting 'I am a time traveler.' The improbability of his claim places a heavy burden of proof on him. But perhaps he could shoulder the burden by means of artifacts, predictions, and demonstrations." Sorenson dismisses all of these possibilities, however, by reminding us of the slightly sleazy history of parapsychology and ESP, both of which run counter to known scientific laws, but which have still duped "many a respected scientist." Any artifact, prediction, or demonstration of time travel, argues Sorenson, is more likely to be the result of deception and fraud than of actual time travel: "Should the time traveler take observers for a spin in his time machine, the skeptics will have us compare their adventures with séances." The rational reaction to such a spin around the centuries, according to Sorenson's presentation, would be like that of a magician who cannot figure out how a colleague has just done his newest act: 'Nice trick! How did you do it?'[65]

The time traveling tourist stranded in the past in one story is used to getting a skeptical reaction because he can provide his questioners no technical explanation for his situation. "How the hell should I know? I'm just a tourist. It has something to do with chronons [see the Glossary]. Temporal Uncertainty Principle. Conservation of coincidence. I'm no engineer."[66] Somewhat more successful (perhaps) is a time traveler born in 2003 who turns up in 1975. After he tries to convince an interrogator of how that can be, he apparently succeeds. As the time traveler later tells a new friend in the past of 1975, "What amazed me . . . was that he really believed me in the end." But the friend doesn't buy that, replying "He did? I think he just pretended. A scientist isn't likely to believe a thing that is against all logic."[67] If the reception committee is a crowd of conservative, cautious Humeans, it would seem that a time traveler is almost certainly doomed. Early science fiction time travelers from 2030, for example, were warned about receiving a skeptical response as follows (the editorial introduction to this tale[68] called it "a curious study of psychology"): "Our wisest men advised against [our trip to the past]. They said we could hope to be received only as imposters and fakirs, that . . . we would find only twentieth-century barbarians, suspicious, ill-tempered, likely to do us bodily harm."

---

[64] What Hume is alluding to here should be plain; as expressed in P. Heath, "The Incredulous Hume," *American Philosophical Quarterly*, April 1976, pp. 159–163, Hume was "an exposer of bad arguments in rational theology." For Hume, second-hand (or even more remote) tales of the return of a man from the dead—the claim that literally kept Christianity alive after Christ's execution—were suspect.

[65] This skeptical reaction was nicely captured in the story "E for Effort" by T. L. Sherred (*Astounding Science Fiction*, May 1947). As one character laments, "I've watched scribes indite the books that burnt at Alexandria; who would buy, or who would believe me, if I copied one. . . . What sort of padded cell would I get if I showed up with a photograph of Washington or Caesar? Or Christ?" The padded cell was indeed the fate of the time traveler in "The Ambassador from the 21st Century" (*Startling Stories*, March 1953) by H. J. Shay, the story of a man who journeyed from A.D. 2007 back to 1952 to warn of a future war; he was committed to a mental institution to receive help for his "illusion."

[66] J. Haldeman, "No Future In It," *Omni*, April 1979.

For many, such skeptical reactions to self-proclaiming time travelers seems dogmatic in the extreme—the response of people with no imagination, no spirit, and heads full of cement. Humean skepticism requires, so it would seem, the rejection of anything and everything that is profoundly surprising, leaving the world a place of utter predictability and boredom. As one science fiction writer put it, "When the miraculous occurs, only dull, workaday mentalities are unable to accept it."[69] Sorenson answers this harsh criticism as follows: "Humeans respond [to Sheckley] by distinguishing between surprises. Most surprises in science do not violate accepted scientific laws. The strange wildlife in Australia was not excluded by biology, X-rays were not precluded by physics."

Sorenson does well, however, to avoid mentioning such profound surprises as, for example, the spectrum of black-body radiation and, later, the photoelectric effect, which were not in the domain of known classical science at the beginning of the twentieth century. Those puzzling, surprising, *totally mystifying* effects required new science—the discovery of the quantum concept by Max Planck. (Explaining the photoelectric effect, *not relativity*, is what won Einstein his Nobel prize.) A strict Victorian-age Humean, as described by Sorenson, would have wrongly rejected the experimental reports of all quantum phenomena and would also (perhaps just as wrongly) have rejected all reports of time travel.

A strict Humean definition (as described by Sorenson) that a miracle has occurred requires a violation of one or more of the *known* [my emphasis] scientific laws of nature.[70] As one modern philosopher defines a miracle, it is any event that "can be explained *only* [my emphasis] by reference to the intervention of a supernatural force."[71] Time travel, by that interpretation, is *not* a miracle because general relativity, not God, is all that is required. C. S. Lewis (1898–1963), late professor of Medieval and Renaissance Literature at Cambridge University, however, absolutely rejected Hume's view on how a rational person should react to certain surprising events. Lewis, one of the most thoughtful modern writers on Christian theology, had no patience with skeptics (or, as he called them, materialists).[72]

---

[67] G. Gor, "The Garden," in *Russian Science Fiction* (R. Magidoff, editor), New York University Press 1969.

[68] P. Bolton, "The Time Hoaxers," *Amazing Stories*, August 1931.

[69] Robert Sheckley, "Something for Nothing," in *Citizen in Space*, Ballantine 1955.

[70] The word *known* is important. As a character in one early science fiction story puts it, "These things [four-dimensional object] sound like miracles; but, after all, what are miracles but phenomena which, *on account of our ignorance* [my emphasis], we cannot explain?" See B. Olsen, "The Four-Dimensional Roller-Press," *Amazing Stories*, June 1927.

[71] D. M. Ahern, "Miracles and Physical Impossibility," *Canadian Journal of Philosophy*, March 1977, pp. 71–79.

Professor Lewis graphically illustrated the dug-in position of the extreme skeptic as follows: "If the end of the world appeared in all the literal trappings of the Apocalypse; if the modern materialist saw with his own eyes the heavens rolled up and the great white throne appearing, if he had the sensation of being himself hurled into the Lake of Fire, he would continue forever, in that lake itself, to regard his experience as an illusion and to find the explanation of it in psychoanalysis, or cerebral pathology."[73] If the end of the world would receive such a skeptical response, then a mere time traveler would surely have no hope at all of being believed.

Lewis would certainly have rejected Sorenson's most astonishing assertion: "So far I have concentrated on the time travel question from the perspective of the time traveler's audience. What about the time traveler himself? Can he at least know he is a time traveler?" Sorenson argues that a time traveler, if authentic, should be able to convince his audience, and that if he can't (and he *cannot* if they are true Humean skeptics), then the time traveler must entertain doubts, too! It doesn't matter (says Sorenson) that the time traveler has memories of his adventures, and it doesn't matter that he knows in his heart that he speaks the truth. Using words that echo Lewis' sarcasm, Sorenson quickly dismisses the importance of the time traveler's self-knowledge, declaring such memories to be merely the symptoms of some deep psychosis, and the traveler's introspective sincerity to be a product of gross self-deception.

Sorenson specifically mentions the traditional Humean response to astonishing reports when he cites earlier writers on time travel in the philosophical literature. In one of those analyses, for example, we find an argument *for* the reasonableness of a rational belief in time travel ("I have been amused and irritated by the spate of articles proving that time travel is a 'conceptual impossibility') by claiming such proofs must be faulty because there is a mathematically consistent explanation for such a belief.[74] (This author was referring to spacetime diagrams, which we'll get to in Chap. 3.) This paper received a very sharp rebuttal from another philosopher who convincingly used fundamental physics to show a simple use of spacetime diagrams in a special relativity setting does *not* support time travel to the past.[75] (I'll return to this point in Chap. 3.) Even later a Humean-style rebuttal came from yet another philosopher, who showed how to explain the time travel phenomenon that Putnam (note 74) described *without* invoking time travel.[76] This isn't to say that Weingard doesn't invoke some pretty astonishing gadgetry (and more) himself, like matter transmitters and anti-matter humans. (You'll see how anti-matter ties-in with time travel a bit later in the book.) A resurrected Hume would surely applaud these rebuttal analyses (although he might also doubt his own fresh existence).

---

[72] In Lewis' eerie, unfinished story "The Dark Tower," a tale of the 'chronoscope,' a gadget that "does to time what the telescope does to space," the persistent skeptic in the story is a Scot, surely created in the image of Hume. See C. S. Lewis, *The Dark Tower and Other Stories*, Harcourt 1977.

[73] C. S. Lewis, *The Grand Miracle*, Ballantine 1986.

[74] H. Putnam, "It Ain't Necessarily So," *Journal of Philosophy*, October 1962, pp. 658–671.

[75] J. Earman, "On Going Backward in Time," *Philosophy of Science*, September 1967, pp. 211–222.

Hardly anybody is happy with Weingard's approach for avoiding time travel (including, I suspect, even Weingard). His 'explanations' seem, just like a time machine, to be incredible and, as Arthur Conan Doyle's Professor Challenger says in one tale not staring Sherlock Holmes, "You cannot explain one incredible thing by quoting another incredible thing."[77] An interesting science fiction exposition illustrating Professor Challenger's Humean philosophy occurs when a copy of *The New York Times* for December 1 shows up for some subscribers a week early, on November 22. It seems the only explanation is either that the paper really is from the future (due to some sort of fluke of the fourth dimension), or that it is a hoax. The first-person narrator of this 1973 tale[78] provides us with his reason for believing the former: "I don't find either notion easy to believe but I can accept the fourth-dimensional hocus-pocus more readily than I can the idea of a hoax." Hume couldn't have said it better.

It should be clearly understood that Hume was not arguing for disbelief in absolutely anything surprising, but rather for rational analysis. Historically, the context of Hume's times was that of what he took to be non-rational arguments for a belief in God, particularly those 'proofs' so beloved by theologians based on Design (Heath [note 64] calls such 'proofs' "philosophical museum pieces"). As Heath writes, "Hume … makes no attempt to deny the supposed facts; he simply argues that they are consistent with other explanations and other analogies of a less ambitious kind. There is no right to attribute to the causes of such phenomena abilities more extensive than are needed to produce the observed effects."

As a matter of fact, even Hume could be convinced of quite strange matters, and I think Sorenson does interpret the philosopher a little too narrowly. In his essay concerning Hume's position on holding a belief in God, Heath wonders whether there is "empirical evidence [imaginable] which would persuade any reasonable mind of the real existence of an infinite God." Heath answers his own question as follows: "If the stars and galaxies were to shift overnight in the firmament, rearranging themselves so as to spell out, in various languages, such slogans as I AM THAT I AM, or GOD IS LOVE—well, the fastidious might consider that it was all very vulgar, but would anyone lose much time in admitting that this settled the matter? . . . Confronted with such a demonstration, the hard-line Humean [but not Hume, himself, I think] could continue, of course, to argue that, for all its colossal scale, the performance is still finite, and so cannot be evidence of more than the finite, though immense power that is needed to achieve it."

---

[76] R. Weingard, "On Travelling Backward in Time," *Synthese*, July–August 1972, pp. 117–132.

[77] Arthur Conan Doyle, "The Disintegration Machine," *The Strand Magazine*, January 1929. Professor Challenger is nothing like Wells' thoughtful Time Traveller; in the original 1912 Challenger novel *The Lost World*, he was described as a "primitive cave-man in a lounge suit."

[78] R. Silverberg, "What We Learned from This Morning's Newspaper," in *Beyond the Safe Zone*, Warner 1986.

Skepticism about 'time travel' was around long before the specific idea of a 'time machine' was conceived. For example, the eleventh-century Persian poet-philosopher Omar Khayyam was blunt in his evaluation of the likelihood of reliving the past. As he so beautifully wrote in one of the quatrains of the *Rubaiyat*,

The Moving Finger writes; and having writ,
Moves on: nor all your Piety nor Wit
Shall lure it back to cancel half a Line,
Nor all your Tears wash out a Word of it.

Quite a bit later the English poet Thomas Heywood, in his 1607 play *A Woman Killed with Kindness*, had one of his characters express a similar thought:

God, O God, that it were possible
To undo things done, to call back yesterday;
That Time could turn up his swift sandy glass
To untell the days, and to redeem these hours.
Or that the Sun
Could, rising from the west, draw his coach backward,
Take from the account of Time so many minutes,
Till he had all these seasons called again,
. . . .
But O! I talk of things impossible,
And cast beyond the moon. . .

When Gödel's discovery of time travel in his rotating universe was announced, the skeptics were easy to find. One philosopher[79] wrote of it "This property [of time travel] must be judged an absurdity by anyone committed to the ordinary modes of speech." And another[80] was only slightly less charitable: Gödel's solution was a "bizarre conception" and a "mere mathematical curiosity." Science fiction wasn't immune to skepticism, either, even though you might have expected that to be the one place where the high drama of time travel would be welcomed. Four *years* after Gödel's paper appeared we find one respected anthologist writing,[81] as part of his introduction to a story, "In this tale we meet our first Mad Scientist. Just as in reality the thoroughly cracked pots used to be found inventing perpetual-motion machines, so in science fiction we find the lunatic fringe more often than not trying to perfect time-travel mechanisms." And that same year the founding editor of *Galaxy Science Fiction Magazine* declared "Time travel requires a suspension of disbelief that is almost unbelievable . . . Scientifically, time travel can't stand inspection.".[82]

---

[79] J. D. North, *The Measure of the Universe*, Oxford University Press 1965.

[80] C. T. K. Chari, "Time Reversal, Information Theory, and 'World-Geometry'," *Journal of Philosophy*, September 1960, pp. 579–583.

[81] Groff Conklin, editor of *Science-Fiction Adventures in Dimension*, Vanguard 1952.

Years later matters had not much changed. For example, in his marvelous 1985 book *The Past is a Foreign Country*, David Lowenthal repeatedly refers to time travel as "fantasy," and to science fiction stories about time travel as "unbridled by common sense." (Lowenthal is a professor of *geography*, not physics.) Science fiction writers were still often not much more enthusiastic about time travel. The well-known science fiction writer and critic Alexei Panshin, for example, agrees with Lowenthal, at one point, long after Gödel, writing "Time travel is a philosophical concept, not a scientific one. It is, in fact, as has often been pointed out, scientific nonsense."[83]

Skepticism does have its uses, however. Modern science fiction writers have often used it as a dramatic means of building conflict and tension in their time travel stories. A skeptical reception is extreme, for example, for a soldier-in-time who has fought in numerous wars, from the ancient past to a billion years in the future.[84] He finds that nobody believes him when he speaks openly of his temporal adventures during a visit to a present-day bar. Everybody merely thinks it is all a hilarious gag. This is in great contrast to one 1870s story[85] in which suspicion of a stranger plays a central role, but which finds its offered explanation in something entirely different from time travel. It tells of a man who suddenly appears in the midst of a Union military camp during the American Civil War.

This man quickly displays strange lapses in his background, as well as possessing knowledge of many different things well beyond anything that could be called common. The details of the story are not important for us but, if it were published in a modern science fiction magazine, this man would almost surely be identified in most readers' minds as a time traveler. In 1875, however, the author's narrator found his punch line in "his firm conviction that the quiet, gentle, well-behaved, modest gentleman, so singularly gifted . . . is, in plain terms, the devil!" Time travel certainly never entered the author's thoughts or, if it did, he lost his nerve at the idea of using it in this pre-Wells story. You'll recall from the opening of the Introduction that it was this 'use of the devil to explain mysterious happenings' that Wells wanted to move away from, and that was the motivation for his introduction of a time *machine*.

Hollywood has at least gotten the skeptical part of the psychology of time travel right (later discussions in this book will focus on how film makers have been less successful with the physics). When, for example, the time traveling villain in the

---

[82] H. L. Gold, editor of *The Galaxy Reader of Science Fiction*, Crown 1952.

[83] In his introduction to Robert Heinlein's classic time travel tale (to be discussed later) "All You Zombies—," in *The Mirror of Infinity*, Canfield 1970.

[84] F. Leiber, "The Oldest Soldier," *Fantastic*, May 1960.

[85] G. C. Eggleston, "Who Is Russell?" *American Homes*, March 1875.

1989 movie *Time Trackers* is confronted in the medieval past, he simply laughs-off a threat to reveal his true identity. "Go ahead," he says in effect, 'the only thing your talk of time machines from the future will accomplish is for people to think you are crazy!'

What would Arthur C. Clarke have thought of all this skepticism being directed toward those who claim to have a time machine? His thoughts about the difficulty time travelers would have in maintaining low profiles were what started the previous section, after all. My guess is that he would have had little patience with extreme incredulity. The surprise of being confronted by a time traveler would soon have turned to awe and pleasure IF—and I emphasize the IF—Clarke had been taken for a spin around the centuries in the stranger's machine. He would surely have ended-up quoting his own famous 'third law' to explain the wonder of it all: "Any sufficiently advanced technology is indistinguishable from magic."

Near the end of his paper, Heath writes what I think is the perfect rebuttal to anyone who would refuse to admit to time travel, even after taking a quick trip backward a few tens of millions of years to the late-Mesozoic era to hunt *Tyrannosaurus rex*, and even after seeing instant photographs of the dead monster with the skeptic's *own foot* on the great creature's head, or of his *own boots* dripping a bloody puddle of unholy size on the floor of the time machine. Writing about the Humean-unconvinced, even when faced with a rearranged firmament, Heath observes "But this now seems a cavil, designed only to prove that even omnipotence is powerless against the extremer forms of skeptical intransigence." Where God would fail to convince, a simple time traveler could hardly hope to do better!

1.4 Troubles with (some) Time Machines

> "If you don't stop this senseless theorizing upon something that's an obvious impossibility, you'll find yourself working alone! Your ridiculous ideas sound like the ravings of a madman. Anyone with average intelligence realizes that the mere thought of traveling through time is absurd."[86]

If the previous section seemed just a bit gloomy concerning time travel, there is a very big reason for that. The sentiment expressed in the above opening quote to this section was a common one among philosophers long before physicists began to seriously think on the topic. While there *are* issues with time travel to the *future*, they are of an *engineering* nature, centered on how to build a big enough rocket ship with enough fuel to make the high speed, looping trip out into space and back again described in the opening section of this chapter. 'Mere' engineering problems are of no concern to physicists and philosophers. What *does* concern them are the far deeper puzzles of time travel to the *past*, the puzzles presented by what appear to be

---

[86] A science fiction physicist receives harsh criticism from a colleague in L. A. Eshbach's "Out of the Past," *Tales of Wonder*, Autumn 1938.

logical paradoxes. Before we get into the paradoxes, however, we need to first clear our minds of two common, popular notions of just what a time machine *is*. Both are false notions (one is due to H. G. Wells) which are, today, *rejected* by physicists (and most philosophers, too).[87]

As you'll see later in the book, all the theoretical time machines that have appeared in the modern physics literature involve spatial displacement. That is, they require *movement*. (On this point, the speedy DeLorean time car in the *Back to the Future* films has it right.) Wells' time machine, however, did not move; it always remained in the Time Traveller's laboratory (or at least on the spot where the laboratory would have been) unless he pushed it about after a trip in time. This, alas, results in a particularly troublesome problem: a Wellsian time machine heading into the past would run into itself!

Consider: There sits my time machine as I prepare for the first time journey ever, a trip back to the late-Mesozoic era to hunt dinosaur. I load my Continental. 600 super-high-power rifle with Nitro Express cartridges the size of bananas, kiss my wife good-bye, and climb in. I pull the lever. Now, Wellsian-type time machines don't *jump over* time but rather travel *through* time (see the Time Traveller's own description of how things looked to him, a description faithfully and spectacularly reproduced in the 1960 film). Therefore, the time machine will instantly collide with itself at the micro-moment *before* I pull the lever!

The resulting destruction obviously introduces a nice paradox: Given that this happens before I pull the lever, how did I manage to pull it? Many of the early science fiction writers were not totally oblivious to this collision problem and, in order to avoid materializing inside of an object in the future or the past, it was common to combine the time machine with an airplane.[88] Even that though might not be enough, as one writer thought a Wellsian time traveler would get "a severe case of the bends" if his body materialized in air![89] Of course, one might argue that Wells' machine does actually move because it is attached to the Earth, which is certainly moving, but it is not clear why this should result in the time machine arriving in the temporal past of the Earth, rather than in some past region of space (almost surely a vacuum).[90]

---

[87] Both of these notions still routinely appears in science fiction, however, because they are 'just too neat' to let 'mere physics' get in the way of a good tale. I use one, without apology, in my own story "Newton's Gift" in Appendix B.

[88] Three such tales are M. J. Breur, "The Time Valve," *Wonder Stories*, July 1930; F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931; E. Binder, "The Time Cheaters," *Thrilling Wonder Stories*, March 1940.

[89] J. Lafleur, "Time as a Fourth Dimension," *Journal of Philosophy*, March 1940, pp. 169–178, and "Marvelous Voyages—H. G. Wells' *The Time Machine*," *Popular Astronomy*, October 1943.

[90] Philosophers seem to be becoming more aware of the collison problem (which they have dubbed "the double occupancy problem"), and at least three papers published since the second edition of *Time Machines* discuss it: W. Grey, "Troubles With Time Travel," *Philosophy*, January 1999, pp. 55–70; P. Dowe, "The Case for Time Travel," *Philosophy*, July 2000, pp. 441–451; R. Le Poidevin, "The Cheshire Cat Problem and Other Spatial Obstacles to Backward Time Travel," *The Monist*, July 2005, pp. 336–352. Physicists don't concern themselves with the collison problem simply because they *aren't interested* in *Wellsian* time machines; I'll explain why I say this by the end of this section.

The general problem of 'where the past is' was nicely illustrated by the physicist Gregory Benford in his 1980 novel *Timescape*. In that story the world of 1998 is on the verge of total ecological collapse, and an attempt is made to change the past by aiming a backward-in-time message via faster-than-light tachyons (these hypothetical particles are discussed in Chap. 5) at the pivotal year 1963. When the principal scientist involved in this effort is explaining the process to a potential financial backer, he is asked, "Hold on. Aim for *what*? Where *is* 1963?" The scientist replies, "Quite far away, as it works out. Since 1963, the Earth's been going around the Sun, while the Sun itself is revolving around the hub of the galaxy, and so on. Add that up, and you find 1963 is pretty distant." An understanding of the question 'Where is the past?' actually goes quite a bit further back in science fiction. For example, after looking through a TV-like gadget to view the past, one character in a 1940s story complains, "You said you'd find Captain Kidd's treasure, but all I can see is fog and static." He is told that's because "It's too far back—1698 or thereabouts. The Earth was billions of miles from here then, and there are too many cosmic rays between."[91]

But let's suppose we ignore this concern about where things are for a time traveler, as do most science fiction stories. Still another problem with a true Wellsian-type time machine is that because it travels *through* time, the machine must *always* appear to be located in the same place. For example, to travel from Ford's Theater today to Ford's Theater on the evening of Good Friday, April 14, 1865, in a misguided attempt to save Lincoln from Booth's bullet (*why* this would be misguided will be discussed at length later in the book), a Wellsian-type time machine would have to occupy every instant of the intervening century and more. For observers outside the machine, the machine would appear to have been sitting in the same place all those years. There is an amusing illustration of a failure to understand this point by the scriptwriters of the 1989 film *Time Trackers*, who have time travelers 'hide' their Wellsian machine from accidental discovery by 'parking' it 5 s in the future!

Wells was well-aware of the "does a time traveling object disappear or not?" issue, and tried to have it both ways in *The Time Machine* by invoking what he had the Time Traveller call "diluted presentation." As we are told in the novel, the reason why we cannot see the model time machine he sends on its way into the future as a demonstration is that "the spoke of a wheel spinning, or a bullet flying through the air" is invisible because if those objects are "traveling through time fifty times or a hundred times faster than we are . . . the impression [they create] will of course be only one-fiftieth or one-hundredth." Similarly for the model. This explanation breaks down when one remembers that, even if you cannot see the spoke or bullet, they are still there and you can get in their way—Wells, unfortunately, has one of his characters stick a hand into the space where the model time machine was last seen.

---

[91] M. Jameson, "Dead End," *Thrilling Wonder Stories*, March 1941. The "cosmic rays" are presumably the cause of the interference.

This objection to Wellsian time travel was raised soon after the 1895 publication of the novel, and then again in 1914 by Pitkin (note 62), who noted that violent disaster awaited once the time journey ended. Wells, it is only fair to note, seemingly anticipated Pitkin when he had the Time Traveller say "So long as I travelled at a high velocity *through time* [my emphasis] ... I was, so to speak, attenuated—was slipping like a vapor through the interstices of intervening substances!" What this is getting at is that for the Time Traveller to stop 'inside' anything (Pitkin's example was the pile of bricks the Time Traveller's laboratory is certain one day to become) would, as Wells had his hero say, cause "a profound chemical[92] reaction—possibly a far-reaching explosion—[that would] blow myself and my apparatus out of all possible dimensions." Just why this spectacular event doesn't occur when the time machine simply stops *in air*, never mind inside Pitkin's pile of bricks, is never addressed.

In any case, it seems clear from all of this that Wells' machine travels *through time*, just as the Time Traveller claims. But Wells, himself, raises doubt when he describes the observed effects of a departing time machine. At the beginning of the novel, when the Time Traveller sends his model machine into the future, we read "There was a breath of wind, and the lamp flame jumped. One of the candles on the mantel was blown out ... and it [the model time machine] was gone—vanished!" And, at the end, when the Time Traveller makes his final exit, the narrator of the tale just misses the departure but tells us "A gust of air whirled around me as I opened the door, and from within came the sound of broken glass falling on the floor. The Time Traveller was not there ... Save for a subsiding stir of dust, the further end of the laboratory was empty. A pane of the skylight had, apparently, just been blown in." Both of these descriptions read as *implosions*, air rushing in to fill a spatial void, as though the time machines had *jumped* in time. Is there an inconsistency here? Well, perhaps not, if one accepts the curious idea of "slipping like a vapor" for an operational Wellsian-type time machine.

One famous science fiction story[93] nicely illustrates these points. The inventor of the first time machine demonstrates it to colleagues by sending a brass cube 5 min into the future. After being placed in the machine, the cube vanishes and then, 5 min later, reappears. Did the cube travel *through* time, or was its journey 'instantaneous,' so to speak? If *through* time, the cube was present at every instant after the start of its trip—so why did it vanish? The cube gets to each instant before the observers do, but why this should produce the visual effect of disappearing is unclear. The description in the story implies the cube traveled 5 min into the future without existing at any of the in-between instants, and so the story's time machine certainly was *not* Wellsian.

An immediate implication of the immobility of a Wellsian time machine is that if you are being chased by an angry mob somewhen in time (perhaps because you unwittingly violated a sensitive social taboo), then hopping into your Wellsian-type time machine isn't going to help because the machine just sits there. The mob could simply take its deliberate time in first building a roaring fire and then pushing the machine (and you) into it. As one author (with a wonderfully appropriate name!) expressed this, "You might as well try to escape by taking a nap."[94]

---

[92] Actually *nuclear*, but don't forget when Wells wrote his novel.

[93] F. Brown, "Experiment," *Galaxy Science Fiction*, February 1954.

Pulp science fiction, always alert to a good story gimmick, used this characteristic of Wellsian time machines in one clever tale[95] in which a criminal attempts to hide his crimes by sending the bodies of his victims into the far future. His mistake is to use a Wellsian time machine in which he escapes into the future. The police, however, having learned of his foul deeds, simply build a cage around the machine and arrest him when he exits 23 years later!

If a Wellsian time machine that moves *through* time suffers from a fatal collision problem, then how about that other favorite of science fiction, a time machine that *jumps* in time? (Recall the final departure of Wells' Time Traveller.) That certainly would avoid the self-colliding problem. When you pull the lever inside the machine you simply disappear from 'now' and (from your point of view) then instantly pop into existence 'then.' The problem with this sort of time machine is that a Time Traveller who uses it will have a *discontinuous* world line, with the break occurring at the moment his time machine 'jumps.' In the modern physicist's view of time travel, however, based on general relativity, a Time Traveller's world line should always be continuous. That's because general relativity is a smooth, local field theory described by differential equations, resulting in *continuous* CTLs/CTCs.

Imagine, for example, that a 'jumping' time machine inventor starts building his gadget at time $t = A$ and expects to finish building it at time $t = B > A$. At time $t = C < B$, however, he runs into a problem. Fortunately, just at that moment a fully-functional time machine suddenly appears in the lab, and from it emerges a slightly older version of the inventor. The older version has the solution to the problem and, after telling the younger version the answer, gets back into the operational time machine and jumps off to ... somewhen. The younger version then completes his machine at time $t = B$, gets into it, jumps back to time $t = C$, and ....[96]

In the past, philosophers have gotten themselves all tangled-up in debates over *personal identity*, that is, which version is *the* inventor, the younger or the older? Can they both be the same person, even though the older version has a world line (*starting* at $t = C$) that is separate and distinct from the world line of the younger (that *stops* at $t = B$)? One philosopher (note 74) left physics behind and pursued this question into the following *legal* question concerning our two (?) inventors: if the older version commits a crime and then vanishes in his time machine before the police can apprehend him, can the younger version be punished even though *he* hasn't yet committed the crime? While certainly 'interesting,' this really is a non-issue for the modern physicist who is concerned only with the physical possibility (or not) of time travel to the past.

---

[94] M. Cook, "Tips for Time Travel," in *Philosophers Look at Science Fiction*, Nelson-Hall 1982. One modern story that gets Cook's point right is by I. Watson, "The Very Slow Time Machine," in *The Best Science Fiction of the Year* (T. Carr, editor), Ballantine 1979.

[95] M. Jameson, "Murder in the Time World," *Amazing Stories*, August 1940.

[96] This little story I've just told you involves what is called a *bootstrap paradox* (just where did that solution come from, that is, who thought it up?) and it is one of the real puzzles of time travel. I'll say *lots* more about such curious doings later in the book.

Well, okay, you might now say, if neither a Wellsian time machine or a 'jumping' time machine will do, then just what *are* physicists studying in their papers on time travel? The short answer here (in Chap. 3 I'll say more) is that physicists don't view time machines as super-tech gadgets covered with wires, meters, dials, and levers, humming away beneath a seated Time Traveller as gigawatts of power throb through massive copper/crystal rods, with the whole business surrounded by a pulsating red-blue glow. Hollywood absolutely loves that sort of thing, but it's simply all wrong. For modern physicists, a time machine is *a region of spacetime with special topological structure*. Then, to time travel, a Time Traveller moves through that region of spacetime (in a rocket, perhaps) along an appropriate path. To 'make a time machine' therefore, in modern terms, means to (somehow) manipulate finite amounts of matter/energy in such a way as to alter the topology of a finite region of spacetime from one that has no CTLs/CTCs to one that does.[97] The most famous example of such a spacetime topology alteration (or *warp*) is the creation of a *wormhole*. A wormhole is a *topological* artifact of a spacetime; wormholes were popularized in Carl Sagan's 1985 novel *Contact* (under the guidance of physicist Kip Thorne) and are now common in science fiction.[98] As mentioned at the start of this chapter, for example, even Stephen King uses one in his 2011 *mainstream* novel *11/22/63*.

I'll return to the 'topology of spacetime' in Chap. 3 but, just so we don't leave it here as a mysterious phrase, here's a simple illustration of a topology change. Imagine a long, flat, narrow, two-dimensional strip of paper. The strip has the following topology features of interest to us here: (1) it has a beginning (its left end) and an ending (its right end), and (2) it has two sides (the top surface) and the flip-side surface. Now, imagine that we take the right end of the strip, give it a half-twist of 180^0 through our three-dimensional space, and then finally we glue that twisted end to the left end of the strip. The half-twist and gluing (our *warp*) has changed both of the topological properties of the strip. That's because the strip now has *no end* (you can travel forever along the strip, always going 'forward' and never reaching a point where can't go forward some more), and the strip now has just *one* side. You can convince yourself that it is one-sided by coloring the strip with a crayon. During the coloring, *do not lift the crayon from the strip*. When you can color no more, you'll find that every last bit of the strip has been colored. You can't do that with the original strip without lifting the crayon and turning the strip over because the original strip was two-sided. Many readers will recognize that what we've done is make a *Möbius strip*, named after the German astronomer and mathematician August Möbius (1790–1868) who described it in 1858.

---

[97] There is a hint of this in one prescient science fiction story, in which the inventor of a time machine, when asked about how it works, replies "An electromagnetic *warping* [my emphasis] of the spacetime continuum." See N. Schachner, "When the Future Dies," *Astounding Science Fiction*, June 1939.

[98] A wormhole is featured in the 2014 film *Interstellar*, whose Executive Producer and technical advisor was Thorne. The film's special effects are relativistically correct (not the typical Hollywood 'fantasy physics'), and you can read how that was achieved in Oliver James, Eugénie von Tunzelmann, Paul Franklin and Kip S. Thorne, "Gravitational lensing by spinning black holes in astrophysics, and in the movie *Interstellar*," *Classical and Quantum Gravity*, February 2015. See also (same authors) "Visualizing *Interstellar*'s Wormhole," *American Journal of Physics*, June 2015, pp. 486–499.

Here's another astonishing property our half-twist warp has introduced. Cut the Möbius strip lengthwise with a scissors; most people believe you will then get two strips, each the length of the original strip but each only half as wide. Actually, you get one strip with a *full* 360° twist, which means the result is back to having two sides. (To see this, make a Möbius strip, cut it, and *then* apply the crayon.) And if you cut this new strip lengthwise once more, you get two separate loops, linked together. Try it and see, but be very careful. As the late science fiction writer Cyril Kornbluth (1923–1958) warned, there may be horrific potential dangers in unschooled experimentation with topology warps:

A burleycue dancer, a pip
Named Virginia, could peel in a zip;
        But she read science fiction
        And died of constriction
Attempting a Möbius strip.[99]

To end this section, I should point out that a change in the topology of a spacetime is *not* a necessary requirement for that spacetime to support time travel to the past. Gödel's rotating spacetime, for example, has a remarkably simple topology and, as you'll recall, it's literally *stuffed* with CTLs/CTCs, to the point that time travel to the past in Gödelian spacetime would be an everyday occurrence. You might think a world that presents time travel as a *fundamentally allowed* physical phenomenon, as does Gödel's spacetime, would be irresistible to science fiction writers. (So far as I know, however, no one has written a time travel story using the rotating universe idea.[100]) In Chap. 6, in fact, I'll show you just how easy it would be to time travel in Gödel's spacetime, using a rocketship as the means to move through that spacetime. Of course, *our* universe is not Gödelian, so the 'time travel to the past' question is not so easily answered for the spacetime we appear to actually inhabit.

---

[99] As you can see from this, science fiction writers have had fun with the Möbius strip. Two early examples *not* involving time travel are N. Bond, "The Geometrics of Johnny Day," *Astounding Science Fiction*, July 1941, and W. H. Upson, "A. Botts and the Möbius Strip," *The Saturday Evening Post*, December 1945. The use of the Möbius strip for time travel occurs, for example, in M. Clifton's "Star, Bright," *Galaxy Science Fiction*, July 1952.

[100] If he had lived, perhaps the well-known science fiction writer James Blish (1921–1975) would have written such a tale. In David Ketterer's biography of Blish (*Imprisoned in a Tesseract*, Kent State University Press 1987), there is this comment from a 1970 letter written by Blish: "I am especially intrigued by the spinning-universe form of time travel, especially since . . . nobody has touched it . . . But I should really stop mentioning the spinning-universe in public, or somebody will nobble onto it before I can get into it!"

1.5 Quantum Gravity, Singularities, Black Holes, and Time Travel

> "A [spacetime] singularity is where God is dividing by zero."
>
> —Anonymous

> "A theory that involves singularities and involves them unavoidably, moreover, carries within itself the seeds of its own destruction."
>
> —Peter Bergmann (1915–2002), Einstein's research assistant at the Institute for Advanced Study, Princeton

A fundamental objection to general relativity's suggestion of the possibility of time travel to the past is that, in a very deep sense, general relativity is known to be incomplete. That is, it is incompatible with quantum mechanics, which is the physics of the very, *very* small—the physics of atomic-size objects and *smaller*. We touched on this at the end of Sect. 1.2, and here we'll take a longer look at the issue of merging quantum mechanics with general relativity.

In quantum mechanics, the discrete nature of the atomic world appears in such phenomena as the photoelectric effect, in which light acts like individual particles (photons) rather than as continuous waves. Einstein's general relativity works beautifully on a cosmological scale but, like Maxwell's theory of electromagnetism, and unlike quantum mechanics, it fails when applied deep in the interior of the atom. Quantum theory, however, seems to work *everywhere*. As one physicist put it, "As far as we can tell, there is no experiment that quantum theory does not explain, at least in principle . . . Though physicists have steered quantum theory into regions far distant from the atomic realm where it was born, there is no sign that it is ever going to break down."[101]

One of the central concepts in relativity is the *world line*, which is the complete story of a particle in spacetime. A world line assigns a definite location to the particle at each instant of time. This is a classical, pre-quantum concept, however, and today physicists use the probabilistic ideas of quantum mechanics to describe the location and momentum of a particle once they get down to the atomic scale of matter. Quantum theory is a discrete theory in which the values of physical entities vary discontinuously (in 'quantum jumps'), whereas in classical theories the values of physical entities are continuous. The difference between the two types of theories is something like the difference between sand and water. Mixing the two theories—the classically smooth, continuous general relativity and the discrete quantum mechanics—to get something called *quantum gravity*, is the Holy Grail of physicists today, and nobody has more than an obscure idea of how to do it.

---

[101] N. Hebert, *Faster Than Light: Superluminal Loopholes in Physics*, New American Library 1988.

Just one of the more curious results of the fusing of quantum mechanics with general relativity may be *quantum time*.[102] That is, in quantum gravity the smallest increment of time that has physical meaning—sometimes called the *chronon*, a term first used in a non-time travel science fiction story[103]—may have a non-zero value. As we'll see later in the book, much of the controversy over the possibility of time machines hinges on what is called the *quantum gravity cut-off*. This is the end-result of destructive spacetime stresses that tend to grow toward infinity whenever a time machine spacetime topology attempts to form. This process goes under the general name of the *back reaction*, and is conceptually similar to a rubber band growing ever more taut as it is stretched, an effect that resists more stretching (and, of course, if stretched too far the rubber band breaks).

The cut-off of those stresses, at some *finite* value, is imagined to occur when the terminal phase of the growth would take place in less than the minimum possible time interval. The cut-off happens because, it is thought, nothing can actually occur in less than the minimum time. The debate is over just what that minimum duration is, and over whether the cut-off would occur before the stresses could reach *finite* values large enough to destroy the putative time machine topology. If the cut-off occurs before the back reaction stresses climb to the critical value, then the time machine survives. Otherwise, not.

To see how this 'works,' consider the two fundamental physical constants associated with classical gravity, the *gravitational constant*[104] G and the speed of light c, and the fundamental physical constant associated with quantum mechanics (*Planck's constant*) ħ. Now, if you play around with combinations of these constants it is easy to show that the following expressions have the units of length, time, and mass, called the *Planck length* (l_P), the *Planck time* (t_P), and the *Planck mass* (m_P), respectively:

$$l_P = \sqrt{\frac{\hbar G}{c^3}} \approx 1.6 \times 10^{-33}\ \text{cm},$$

$$t_P = \sqrt{\frac{\hbar G}{c^5}} \approx 5.3 \times 10^{-44}\ \text{s},$$

$$m_P = \sqrt{\frac{\hbar c}{G}} \approx 22 \times 10^{-6}\ \text{g}.$$

The extremely tiny values of l_P and t_P (the chronon), in particular, indicate (*roughly*) where it is expected that the smooth, continuous spacetime of general relativity will itself become quantized, and so will have to give way to a quantum theory of gravity.

---

[102] H. Kragh and B. Carazza, "From Time Atoms to Space-Time Quantization: the idea of discrete time, ca 1925–1936," *Studies in History and Philosophy of Science*, June 1994, pp. 437–462.

[103] S. Weinbaum, "The Ideal," *Wonder Stories*, September 1935.

[104] This is the constant in Newton's famous inverse-square law for gravity; the attractive force F between two point masses m_1 and m_2, distance r apart, is $F = G\frac{m_1 m_2}{r^2}$.

The associated value of the mass-energy density when this transition is imagined to occur is enormous; the so-called *Planck density* is the Planck mass divided by the cube of the Planck length and has the value of about 10^{94} grams/cm^3. This is where physicists expect classical and quantum gravity to part company. Can such an enormous mass-energy density actually occur?[105]

Yes, and more, in what physicists call *singularities*.

This was all still very speculative until about 50 years ago, but today the search for how to connect general relativity and quantum mechanics is serious business. That search is related to time travel studies via a fantastic sequence of discoveries in relativistic physics, made during the last 80 years, beginning in 1931 with the work of the young Indian astrophysicist Subrahmanyan Chandrasekhar (1910–1995). He combined quantum mechanics and *special* relativity to show that a non-rotating star above a certain mass (about 1.4 times the mass of the Sun) cannot evolve into a white dwarf, which had until then been thought the eventual fate of all stars. Stars more massive than 1.4 Solar masses (but not *too* massive) would, instead, become *neutron* stars. But what then happens to stars that are too massive for even *that* bizarre eventuality?

General relativity predicts that a sufficiently massive star—greater than about four times the mass of the Sun—will, when its fuel is nearly exhausted and its nuclear fires are beginning to fade, experience a truly spectacular event called *total gravitational collapse*. When its fuel-starved, weakened radiation pressure is no longer able to keep a massive, aged star inflated against the collapsing force of its own gravity, the star will suddenly implode and crush itself into what is called a *black hole*, a dramatic term coined in 1967 by the Princeton physicist John Wheeler (1911–2008) in an address before the American Association for the Advancement of Science. A black hole is an object with a gravitational field so strong that even light cannot escape—that's why it's black!—at whose center is something called a *singularity*. This is all well-known lore in the physics world.[106]

Indeed, cataclysmic views of the collapse of matter are actually quite old. In Lucretius' first-century B.C. *The Nature of the Universe*, for example, we find the following imagery on what it would be like if matter itself collapsed: "The ground will fall away from our feet, its particles dissolved amid the mingled wreckage of heaven and earth. The whole world will vanish into the abyss, and in the twinkling of an eye no remnant will be left but empty space and invisible atoms. At whatever point you allow matter to fall short, this will be the gateway to perdition." These words were actually inspired by earthquakes, not black holes and their singularities, but could a modern expert in general relativity and singularities have said it any better? But, what *is* a singularity?

---

[105] By comparison, the density of a neutron star is on the order of a 'mere' 10^{16} grams/cm^3.

[106] Perhaps not so well-known, however, is that science fiction was there long before Wheeler. In one classic tale (M. Leinster, "Sidewise in Time," *Astounding Stories*, June 1934) a scientist explains at the end, "We know that gravity warps space . . . We can calculate the mass necessary to warp space so that it will completely close in completely . . . We know, for example, that if two gigantic star masses of certain mass were to combine . . . they would simply vanish. But they would not cease to exist. They would merely cease to exist in our space and time." And then, as another character sums it up, "Like crawling into a hole and pulling the hole in after you." The explicit use of the complete term *black hole* for a region of weird spacetime also appeared in science fiction before Wheeler (P. Worth, "Typewriter from the Future," *Amazing Stories*, February 1950).

As one theoretical physicist has dramatically written, "once gravity runs out of control, spacetime smashes itself out of existence at a singularity."[107] Or to quote Hawking, "A singularity is a place where the classical concepts of space and time break down as do all the known laws of physics."[108] One particular view of a singularity is that it is a place in spacetime that has infinite density and a gravitational field that is infinitely strong. The curvature of spacetime (more on curvature in Chap. 3) at this sort of singularity, sometimes called a *crushing* singularity, is also infinite. This is the sort of singularity believed to be at the center of non-rotating black holes. Historically, however, the occurrence of infinities in physical theories has been thought the red flag signaling that the theories have simply been extended too far, and their calculated results are nonsense.

Perhaps, then, singularities occur only in unrealistic physical applications of general relativity, and so it is only *perfectly* spherical collapsing stars that can end-up (on paper) as a black hole singularity. For a while physicists tried to establish that, but they were forced to abandon the attempt when it was shown that singularities are *unavoidable* and not just the result of idealistic assumptions.[109] This result worried many, and so the concern that general relativity was failing with its prediction of black holes and their singularities continued. In the case of a crushing singularity, perhaps all that meant is that once the collapsing star had fallen into a region even smaller than an electron, general relativity is no longer valid and the singularity is simply the 'math gone wild.' Einstein, himself, held that view. In his book *The Meaning of Relativity* (based on lectures he gave at Princeton in 1921), he wrote (concerning the use of the general theory to study the origin of the universe as a "big bang," which was a crushing singularity), "For large densities of field and matter, the field equations [of general relativity] and even the field variables which enter into them will have no real significance. One may not therefore assume the validity of the equations for very high density of field and matter, and one may not conclude that the 'beginning of the expansion' must mean a singularity in the mathematical sense."

Well, what does general relativity say about the singularity at the center of a black hole? To start, the theory says that, at a distance directly proportional to the mass of the collapsed object, a so-called *event horizon* will form. The event horizon is a surface in spacetime through which anything can fall into the hole, but through which nothing, not even a photon of light, can escape outward. The singularity at the black hole's center is therefore not visible to a remote observer (the singularity is said to be "clothed," and so not "naked"[110]). For all observers beyond the event horizon, the only visible properties of the hole are its mass (via its gravitational effects), its angular momentum (its spin rate), and its electric charge, and these properties are independent of the details of the pre-collapsed object (other than the requirement that electric charge and angular momentum are conserved).

---

[107] P. C. W. Davies, *The Edge of Infinity*, Simon & Schuster 1981.

[108] S. Hawking, "Breakdown of Predictability in Gravitational Collapse," *Physical Review D*, November 15, 1976, pp. 2460–2473.

[109] S. W. Hawking and R. Penrose, "The Singularities of Gravitational Collapse and Cosmology," *Proceedings of the Royal Society A*, January 27, 1970, pp. 529–548. Ironically, one of the 'realistic' assumptions made in this paper, which appears to *force* singularities to exist, is that time travel is *impossible*!

There are actually several fundamentally different types of black holes. If the collapsed star forms a non-rotating, spherically symmetric, uncharged[111] object, then the result is called a *Schwarzchild* black hole, after the German astronomer Karl Schwarzchild (1873–1916) who found the first exact solutions to Einstein's general relativity field equations just months after Einstein published them.[112] Soon after that the Finn Gunnar Nordström and the German Heinrich Reissner independently found the solution to the field equations for the slightly more realistic non-rotating, *charged* black hole.[113] This is only slightly more realistic since it is highly unlikely a black hole wouldn't be spinning, as all observed stars are spinning and angular momentum is conserved during gravitational collapse. Another slightly more realistic solution, that of a *rotating*, uncharged black hole, was found by the New Zealand mathematician Roy Kerr in 1963, and this solution had a twist to it that at last explains why I am telling you all this—the singularity at the center of a Kerr black hole is not the *point* singularity of a non-rotating black hole but rather is a *ring* singularity. That is, there is a hole in the Kerr singularity through which matter can travel, without being destroyed, a hole that seems to act as a portal into other spacetime regions that may include past or *future* regions of spacetime. In other words, the ring singularity seems to be the entrance to a time machine.[114]

---

[110] A naked singularity, with no event horizon behind which to hide, would be particularly bothersome to physicists who don't like the idea of the breakdown of physics being on full display. What they think they'd then see would be completely unpredictable. Whether such a situation can actually exist is still open to debate, but there *are* both analytical solutions and computer simulations (incorporating realistic equations of state on the pressure response of matter as it is compressed) that seem to allow it (as in the gravitational collapse of an *infinitely* long, non-rotating cylinder that appears to result in an axial, thread-like, naked singularity).

[111] The word *charge* means either electrical *or* magnetic charge, although from a practical point *charge* probably does mean just electrical, as the theoretically possible magnetic monopole has yet to be observed and, in any case, it is thought that black holes will not have a significant electrical charge.

[112] Even *Einstein* hadn't yet solved them, and he apparently thought they were too complicated to be solved; when he saw Schwarzchild's result, he was so impressed that Einstein wrote to say "I had not expected that the exact solution to the problem could be formulated. Your analytical treatment of the problem appears to me splendid."

[113] Two years later, the University of Pittsburgh physicist Ezra Newman finally solved the field equations for the realistic, general case of a rotating *and* charged black hole.

A discussion of singularities in general relativity is especially complicated for at least two reasons. First, there is more than one type, with *crushing* being just the (perhaps) most 'obvious.' Another type has no infinite curvature associated with it, but rather is a point in spacetime beyond which the worldline of a freely falling mass cannot be extended. Such a point is called a *geodesically incomplete* singularity, and it represents either an end to space or to time (in either case, that point is on the boundary or edge of spacetime). There are other types, as well—I've mentioned the naked singularity already—and the appearance of *any* of them is distinctly unsettling (recall Bergman's opening quote) to physicists. One that may be the most unsettling of all, however, is the *thunderbolt* singularity. This singularity propagates to infinity at the speed of light! As its discoverers dramatically put it, "It is not a naked singularity because you do not see it coming until it hits you and wipes you out."[115]

The other reason for a discussion of general relativity singularities being complicated is that they simply are not like the singularities of earlier theories. For example, in electromagnetic field theory spacetime is the *given background reference*; that is, a singularity in that theory is a point in spacetime where the *electromagnetic field* is undefined. In gravitational field theory, however, it is spacetime itself that is undefined, and there is no background 'something' in which spacetime is embedded to serve as a reference.[116]

---

[114] You can find discussions on how this is imagined to work in two papers by R. Weingard: "General Relativity and the Conceivability of Time Travel," *Philosophy of Science*, June 1979, pp. 328–332, and "Some Philosophical Aspects of Black Holes," *Synthese*, September 1979, pp. 191–219. See also M. Calvani et al., "Time Machine and Geodesic Motion in Kerr Metric," *General Relativity and Gravitation*, February 1978, pp. 155–163. I won't pursue black hole time machines in this book, as it is not what modern physicists consider a plausible means of time travel (*How are you going to gain access to a black hole?!!!*) For how one science fiction writer *did* use the idea, however, see L. Niven, "Singularities Make Me Nervous," in *Stellar* 1 (J.-L. del Rey, editor), Ballantine 1974. Black holes *are* bizarre objects—nearly as bizarre as time travel—and it seems risky to try to understand one in terms of the other (recall Professor Challenger's observation!).

[115] S. W. Hawking and J. M. Stewart, "Naked and Thunderbolt Singularities in Black Hole Evaporation," *Nuclear Physics B*, July 1993, pp. 393–415. As bizarre as is the thunderbolt, it was anticipated in science fiction by more than half a century. In the story "The Tides of Time" by R. M. Williams (*Thrilling Wonder Stories*, April 1940), the universe is collapsing at faster than the speed of light. Human scientists learn this when fleeing aliens stop their faster-than-light space ships to warn them. One of the human characters then looks out into the night sky and, in words that sound like those of Hawking and Stewart, "There would be no warning, for the rolling tide was traveling faster than light . . . It would come faster than the flicker of an eye. No one would see it come. One instant the world you knew would be around you. The next instant, there would be nothing. You would not even have time to know what had happened. Death, faster than the lightning flash!" This story may have been inspired by a tale published decades earlier, by the Canadian writer Frank Lillie Pollock (1876–1957). In his "Finis" (*The Argosy*, June 1906), written long before the concept of a super-nova, the light of a huge, distant star finally arrives to cook Earth into oblivion.

One early suggestion on how to avoid the problem of the crushing singularity of the non-rotating black hole (which is, as mentioned earlier, not a realistic model for the gravitational collapse of a rotating star) is that the collapse may stop short of the singularity. That is, the collapsing body might instead rebound. This 'bounce' would occur after the star was inside its event horizon, so an external observer would not see the later expansion, an expansion imagined to be through the event horizon *but into a different region of spacetime*.[117] When Novikov's work was generalized the following year, the authors clearly had a hard time believing this dramatic imagery, despite their own mathematics, concluding with "It then appears necessary to believe in the existence of other [regions of the universe, including the past and the future] which will accommodate the re-expansion. This seems at least as fantastic as the alternative of [a point singularity].".[118]

In 1974 Hawking announced an astonishing partial connection of quantum mechanics with general relativity's black holes. He showed that, contrary to the usual image of black holes as being one-way trap doors to . . .?, black holes actually *must* radiate energy.[119] His analysis, which stunned physicists by its beautifully simple arguments, invokes the famous *uncertainty principle*, one of the corner stones of quantum mechanics. Hawking himself found the result "greatly surprising." He also cautioned (in his 1975 paper) that the following picturesque imagery is "heuristic only and should not be taken too literally," but it has now been in physics for over 40 years and appears to be here to stay.

The uncertainty principle states that there are certain pairs of variables associated with particles, variables that cannot be precisely measured at the same time. Time and energy form such a pair because a non-zero time interval is required to measure a particle's energy, and the product of the uncertainty in both the time interval (Δt) and the energy (ΔE) must be at least as large as a certain non-zero constant. That is, if $\hbar$ is Planck's constant, then $\Delta E \Delta t \sim \hbar$. This allows the process of virtual particle creation, the appearance of particle/anti-particle pairs just outside the event horizon of a black hole. The uncertainty in the energy is what gives the combined mass of the particles in a pair; this uncertainty in the energy is the quantum fluctuation energy of the intense gravity field of the hole. The only constraint is that the energy be returned to the field, via mutual annihilation of the matter/anti-matter pair within the time uncertainty dictated by the uncertainty principle.[120]

---

[116] See, for example, R. Geroch, "What Is a Singularity in General Relativity?" *Annals of Physics*, July 1968, pp. 526–540.

[117] See, for example, I. Novikov, "Change of Relativistic Collapse Into Anticollapse and Kinematics of a Charged Sphere," *JETP Letters*, March 1, 1966, pp. 142–144, and V. P. Frolov, *et al.*, "Through a Black Hole Into a New Universe?" *Physics Letters*, January 12, 1989, pp. 272–276. Igor Novikov is a Russian physicist at the University of Copenhagen, and he will appear later in the book when we get to the paradoxes of time travel to the past.

[118] V. De La Cruze and W. Israel, "Gravitational Bounce," *Nuovo Cimento A*, October 1, 1967, pp. 744–760.

[119] S. Hawking, "Black Hole Explosions?" *Nature*, March 1, 1974, pp. 30–31, and "Particle Creation by Black Holes," *Communications in Mathematical Physics*, 1975, pp. 199–220.

As Hawking showed, this time interval, although incredibly short, is still long enough for the two virtual particles to separate before annihilation, one falling into the hole and the other escaping. This would happen, for example, if the particle/anti-particle pair is an electron/positron pair, and so a negatively/positively charged black hole would tend to attract the positron/electron and repel the other particle (either way, driving the charge of the hole towards zero). (Hawking then later suggested[121] that the particle *entering* the hole could be thought of as an *emitted* particle traveling backward in time, an idea that can be traced back decades, to John Wheeler—I'll return to this idea in just a moment.) By this incredible quantum process, then, the black holes of general relativity slowly *evaporate* (!) as they glow with what is now called *Hawking radiation*. That is, black holes appear to be hot bodies. But *hot* is relative, as a black hole with the mass of the Sun would have a temperature of just sixty *nano*-degrees Kelvin above absolute zero, and it would take 10^{66} years (a stupendously enormous time compared to the age of the universe) to completely evaporate.

Indeed, one physicist had already mused that the *entire universe* might have been created by a similar quantum process, *out of nothing*, a so-called *vacuum fluctuation*.[122] The explanation for why the universe doesn't then disappear—and very quickly, too, because the energy for all the mass in the newly created universe is quite large (that is, ΔE is *really big* and so Δt must be *really small*)—is that the *negative* gravitational potential energy of all that newly created matter would cancel the positive mass-energy, and so ΔE is actually *quite small* and so Δt is then *quite large*. To perhaps show he wasn't quite convinced by all that, himself, Tryon whimsically wrote "I offer the modest proposal that our Universe is simply one of those things which happen from time to time."

As a final comment on the suggestion by Hawking of a connection between virtual particles at a black hole event horizon and backwards time travel, the idea originated (as I said before) with Wheeler, in 1941.[123] In an astonishing coincidence, even as Wheeler was telling his student Richard Feynman about this, a science fiction writer was also identifying anti-matter with backward time traveling 'normal' matter.[124] Later, the Polish science fiction writer Stanislaw Lem (1921–2006) took this idea, combined it with the quantum concept of energy fluctuation, and came up with one of his typically outrageous (and typically hilarious) ideas: shooting a single positron out of an accelerator back to the very beginning of time. His story character called this fantastic machine the "Chronocannon" and claimed that's what started the universe.[125]

---

[120] The uncertainty principle has long been used in time travel science fiction. In one story, for example, a character is transported from 1950 to 2634 by a scientist of the future. Once there, this character decides he'd like to remain permanently in the 27th century. He is told that he can't because he is like an atom excited into an elevated energy state and, just as quantum mechanics says that eventually an electron in such a state will drop back down into a lower energy state, so do the "laws of time travel" require that he drop back to his normal time. How long can he remain in future, he is told, "depends on the mass [energy] of his body and the number of years the mass [energy] is displaced." That is simply the uncertainty principle. See W. Bade, "Ambition," *Galaxy Science Fiction*, October 1951.

[121] S. Hawking, "The Quantum Mechanics of Black Holes," *Scientific American*, January 1977.

[122] E. P. Tryon, "Is the Universe a Vacuum Fluctuation?" *Nature*, December 14, 1973, pp. 396–397.

Soon after Lem, a philosopher used a variant of this idea, in which the Big Bang creation of the universe was caused by a time traveler from the future who saw a need—his own existence—to generate the Big Bang. This leads to philosophical speculations on the cosmological implications of God as a time traveler.[126] Two recent physicists have taken this one step further by suggesting that the universe, via time travel, may have caused itself! As they put it, "the laws of physics may allow the Universe to be its own mother."[127]

The modern hope is that quantum mechanics (as in quantum gravity) will save physics from the horror of general relativity's singularities. This was the view of John Wheeler and, as the man who named black holes, his view is important to consider. General relativity is a classical, *smooth* theory that is fundamentally continuous, while 'our' universe appears to be a quantum one. So, perhaps, general relativity's prediction of singularities may be just an artifact without physical reality in the 'real world.' Wheeler's position was based on the quantum fluctuations of gravity fields, which are related to the uncertainties inherent in our knowledge of the values of physical entities. Such fluctuations are vanishingly small in systems of everyday size, but they increase dramatically at very tiny distances that are twenty orders of magnitude smaller than the nucleus of an atom. In the microscopic region of spacetime that the matter forming a black hole is falling into, these fluctuations might conceivably result in effects that preclude the formation of a singularity. Agreeing were two physicists who asserted that, even without a detailed knowledge of quantum gravity, quantum effects "would smash the idealized interior geometry"

---

[123] See Richard Feynman's Nobel lecture, reproduced in *Science*, August 12, 1966, pp. 699–708, where he recounts Wheeler's 'proof' for why every electron in the universe has *exactly* the same charge ('there is only *one* electron, weaving its way back-and-forth in time, with positrons being the electron when traveling backward-in-time').

[124] Will Stewart, "Minus Sign," *Astounding Science Fiction*, November 1942. 'Will Stewart' was a pen-name for John Stewart Williamson (1908–2006).

[125] S. Lem, "The Eighteenth Voyage," in *The Star Diaries*, Seabury Press 1976.

[126] G. Fulmer, "Cosmological Implications of Time Travel," in *The Intersection of Science Fiction and Philosophy* (R. E. Meyers, editor), Greenwood Press 1983. Isaac Asimov used a similar idea in his story "The Instability," *The London Observer*, January 1, 1989.

[127] J. R. Gott and L.-X. Li, "Can the Universe Create Itself?" *Physical Review D*, 1998, 023,501.

[that is, the ring singularity] of a rotating, charged black hole, thereby eliminating any possibility of using such a hole for time travel.[128]

And finally, to generalize beyond black holes to the hoped-for pay-off of the coming of quantum gravity in banishing singularities altogether, one recent study has examined how non-crushing singularities (that is, ones of the geodesically incomplete type) *are* apparently "healed" (the authors' term) by quantum effects.[129] With the eventual development of quantum gravity, perhaps all the singularities of general relativity will vanish while leaving the CTCs/CTLs intact, thereby removing a form of doubt in the theory's apparent support for time travel to the past. It may be a long time coming, however: as the University of Sydney philosopher of science Dean Rickles recently (2014) wrote in his book *A Brief History of String Theory*, "quantum gravity is in many ways . . . a revolution still waiting to happen."

1.6 Tipler's Time Machine

> "In short, general relativity suggests that if we construct a sufficiently large rotating cylinder, we create a time machine."[130]

The time traveling property of the ring singularity in a rotating black hole once made it a favorite of science fiction writers, as in Joe Haldeman's classic 1974 novel *The Forever War* (in which the term used is not *black hole*, but *collapsar*, which is a nicely descriptive word in its own right). A major difficulty with this approach, however, as I mentioned in the previous section (note 114), is that of 'getting one's hands on' (so to speak) a black hole! So, is there any other 'time machine' that is consistent with general relativity? Yes, there is.

In 1974 a young physics graduate student at the University of Maryland, Frank Tipler, caused a bit of a stir when he published what seemed to be quite specific construction details for a time machine. Indeed, the final sentence (the above quotation) of his paper couldn't be clearer. Nobody had ever before made such a statement in a respectable physics journal and, best of all, there were no apparent spacetime singularities involved. However, a close look at Tipler's analysis does turn up some difficulties.

What Tipler had actually done was to show that if one had an *infinitely long*, *very dense* cylinder rotating with a surface speed of at least half the speed of light (the rotation speed is such that the outward centrifugal forces are balanced by the inward gravitational attraction of the cylinder), then this allowed the formation of closed timelike curves around the cylinder. This means that by orbiting the surface of such a fantastic cylinder, one could travel through time into the past—but not to earlier than the moment of the creation of the cylinder.

---

[128] N. D. Birrell and P. C. W. Davies, "On Falling Through a Black Hole Into Another Universe," *Nature*, March 2, 1978, pp. 35–37.

[129] T. M. Helliwell and D. A. Konkowski, "Quantum Singularities in Spherically Symmetric, Conformally Static Spacetimes," *Physical Review D*, May 13, 2013, 10404.

[130] F. J. Tipler, "Rotating Cylinders and the Possibility of Global Causality Violation,' *Physical Review D*, April 15, 1974, pp. 2203–2206.

This last point is a very important one, as it does avoid one particularly odd paradox (called a *bootstrap*): a traveler going backwards in time to tell the inventor of a time machine (perhaps an earlier version of the time traveler himself) how to build the time machine. You can find this idea in early science fiction,[131] and a minor variant of it was amusingly illustrated in the 1985 film *Star Trek IV: The Voyage Home* (when you next watch the movie, ask yourself who actually invented "transparent aluminum"?) Bootstrap paradoxes are quite mysterious and still befuddle physicists and philosophers. Science fiction writers, on the other hand, love bootstraps as great story gimmicks.

Tipler's cylinder would also enable a time traveler to return to her original time, to go "back to the future," by orbiting the cylinder in the reverse direction (but no further into the future than when the cylinder ceases to exist). Later in the book I'll show you a simple illustration—based on a similar one in Tipler's PhD dissertation, published in 1976—of how the cylinder works as a time machine. No one, in fact, disputes any of this. It *is* true. On paper.

But Tipler did *not* prove that a time traveling property holds for cylinders of even very long but *finite* length, which are the only kind we could actually build from a finite amount of matter; he merely suggested that such might be the case. This suggestion does seem reasonable, because if the time traveler orbits at the midpoint of the cylinder, near the surface, then the gravitational end-effects of sufficiently remote ends of the cylinder would, you'd think, become negligible. Similar mathematical approximations are routinely made, for example, when calculating the electrical effects of charged cylinders of finite length. But as one physicist has warned, "Extrapolation from cylindrical symmetry to reality is very dangerous, since spacetime is not even asymptotically flat around an infinite cylinder."[132] The issue of whether a spinning, finite-length cylinder can create closed, timelike curves is still open: to quote another physicist, "[In] some respects an infinite cylinder may be a model for a long finite one, and the possibility cannot be dismissed that a time machine might be associated with a long, but finite rotating system."[133]

---

[131] See, for example, C. Cloukey, "Paradox," *Amazing Stories Quarterly*, Summer 1929. Later in the book I'll discuss even earlier literary occurrences of bootstraps (that is, of *information* on closed loops in time).

[132] K. S. Thorne, "Nonspherical Gravitational Collapse: Does It Produce Black Holes?" *Comments on Astrophysics and Space Physics*, September–October 1970, pp. 191–196. What "asymptotically flat" means will be discussed in Chap. 3.

[133] W. B. Bonner, "The Rigidly Rotating Relativistic Dust Cylinder," *Journal of Physics A*, June 1980, pp. 2121–2132. Tipler was not the first to study rotating cylinders in the context of general relativity. Such cylinders had been around for decades, going back to 1932. A good reference is M. A. Mashkour, "An Exterior Solution of the Einstein Field Equations for a Rotating Infinite Cylinder," *International Journal of Theoretical Physics*, October 1976, pp. 717–721. The first-analyzed configuration of matter that generates closed timelike lines, solved in all its general relativistic detail, was the infinite rotating cylinder studied by W. J. van Stockum,

There is, however, another potential problem besides the length of the cylinder. There is a strong likelihood that a Tipler cylinder under construction would collapse under its own internal gravitational pressure before it could be made nearly long enough to be even ‘approximately infinite.’ That is, such a finite-length cylinder might crush itself along its long axis into a pancake-shaped blob, something like what happens to a long cylinder of jello stood on-end. An ordinary can of jellied cranberry sauce will also sometimes display this curious behavior.

The required rotational speed raises yet another concern, as well. We are not talking about cylinders the diameter of a pencil, or even of a large water pipe. Recall that for a given surface speed, the larger the diameter the less the centrifugal acceleration at the surface. It is easy to calculate that even a huge cylinder 10 kilometers in radius, with a surface speed of half the speed of light, would have a surface acceleration *hundreds of billions* of times the acceleration of Earth’s surface gravity. No known form of ordinary matter could spin that fast and not explosively disintegrate; Tipler has estimated that the required density for a time machine cylinder would be 40 to 80 orders of magnitude above that of nuclear matter. (In a masterful understatement, Tipler calls this astonishing stuff “unknown material.”) Made from such incredibly superdense stuff, even a finite cylinder would still be as massive as the Sun but many trillions of times smaller. Showing no lack of imagination, Tipler has suggested the possibility of speeding up the rotation of an existing star as an alternative approach to that of building a cylinder.[134] That, of course, would be project for a far-future society, with a *very* advanced technology.

All of these concerns were discouraging to Tipler (who could blame him?), and his pessimism about the actual likelihood of achieving time travel via one of his cylinders is shown by the words he used to open his 1977 paper (note 134): “Any attempt to evolve a time machine] from [normal] matter will cause singularities to form in spacetime. Thus, if by the word ‘manufacture’ we mean ‘construct using only ordinary materials *everywhere*,’ then the theorems of this paper will conclusively demonstrate that a [time machine] cannot be manufactured.” But not all physicists agreed.

“The Gravitational Field of a Distribution of Particles Rotating About an Axis of Symmetry,” *Proceedings of the Royal Society of Edinburgh*, 1939, pp. 135–154. This is particularly interesting because, while Van Stockum didn’t spot the presence of closed timelike lines in his solution, his cylinder is made entirely from *ordinary* matter.

[134] F. Tipler, “Singularities and Causality Violation,” *Annals of Physics*, September 1977, pp. 1–36. See also his earlier paper “Causality Violation in Asymptotically Flat Space-Times,” *Physical Review Letters*, October 1976, pp. 879–882, where he wrote “There are many solutions to the Einstein equations [of general relativity] which possess causal anomalies in the form of closed timelike lines (CTL). It is of interest to discover if our Universe could have such lines. In particular, if the Universe does not at present contain such lines, is it possible for human beings to manipulate matter so as to create them? [That is, to construct a time machine.] I shall show in this paper that it is *not* [Tipler’s emphasis] possible to manufacture a CTL-containing region without the formation of naked singularities, *provided normal matter is used in the construction attempt* [my emphasis].”

Years after Tipler wrote, one physicist replied[135] with *two* pointed observations. First, Tipler's theorems apply only to singularities of the incomplete kind, not to the more convincingly fatal crushing (or curvature) type. Second, to quote Ori at length, "The standard interpretation of Tipler's theorems is to say that the appearance of a singularity in a given [spacetime] model indicates that this model is unrealistic and cannot be physically realized. Even for future-generation engineers it will probably be impossible to use 'singular matter' for the construction of their time machine. However, the theory of black holes provides an obvious counterexample to this interpretation. For, by applying this interpretation to the black hole singularity theorems one could conclude that black holes can never form." Yet black holes with several times the mass of the Sun *have* been detected in orbit about certain stars, and at least one supermassive black hole (with a mass equal to more than three *billion* Suns) has been detected at the core of galaxy M87. Indeed, it is now believed that the center of *every* sufficiently massive galaxy in the universe is home to a black hole (the one at the center of our own galaxy, the Milky Way, has a mass about three million times that of the Sun).

Even less concerned about singularities interfering with time travel were two other physicists who wrote[136] "It would seem that a successful attempt to manufacture [a time machine] within a finite region of space will be accompanied by the creation of a singularity . . . This does not immediately imply, however, that with a sufficiently advanced technology one could not make a time machine. *There is no reason to suspect spacetime singularities could not in principle be created through deliberate human action* [my emphasis]."

These optimistic views were, of course, welcome news for science fiction writers, who had been using Tipler cylinders almost from when Tipler first wrote of them. Indeed, Larry Niven liked them well enough to 'lift' the very title of Tipler's paper (note 130) for the title of a short time travel story for inclusion in his 1979 collection *Convergent Series*. Just one year after Tipler's paper appeared, Poul Anderson featured the cylinders in his 1978 novel *The Avatar*, where they are called "T-machines": one can imagine the "T" stands for Time or Tipler or even both. Anderson's story describes the cylinders as having been scattered about the universe by ancient, altruistic aliens called "the Others," for the use by any who come across them and who have the wits to decipher *how* to use them. Anderson recognized the obvious problems with Tipler cylinder construction, and so has one of his characters say of T-machines, "I have no doubt whatsoever that [they are] the product of a technology further advanced from ours than ours is from the Stone Age."

---

[135] A. Ori, "Must Time-Machine Construction Violate the Weak Energy Condition?" *Physical Review Letters*, October 1993, pp. 2517–2520. The weak energy condition is the seemingly 'obvious' requirement that the observed local mass-energy density should never be negative. Quantum mechanics predicts (and it has been experimentally confirmed) that there *are* exceptions.

[136] M. P. Headrick and J. R. Gott, "(2+1)-Dimensional Spacetimes Containing Closed Timelike Curves," *Physical Review D*, December 15, 1994, pp. 7244–7259. The '(2+1)' refers to a toy spacetime with just two spatial dimensions and one time dimension.

Even before (actually *long* before) Tipler's paper, science fiction had foreshadowed his physics. Oliver Saari (1918–2000), for example, had incorporated both superdense matter and the rule of 'no time travel before the creation of a time machine' in a story written *40 years* earlier.[137] Saari's fictional time machine works by warping spacetime via a plate of superdense matter. (An even earlier tale[138] had also used superdense matter, but it was badly flawed by its hocus-pocus invoking of 'rays' emitted by the newly discovered element of *tempium*.) The 'no time travel before the creation of a time machine' rule is the basis for an obvious response to Hawking's Chronology Protection Conjecture, discussed earlier in this chapter, and it was so used by one physicist to rebut the Conjecture: as he wrote,[139]

(1) time machines, if possible, must have the property of not being able to travel back to before their creation, and
(2) no time machine has yet been created.

The absence of time travelers amongst us, therefore, provides no insight, one way or the other, on the eventually possibility of constructing a time machine.

1.7 For Further Discussion

Observations of the background microwave radiation that permeates the universe is strong experimental evidence for the Big Bang, the singularity thought to be the origin of the universe. This singularity is not shielded from us by an event horizon, and so is not a naked singularity (note 110), which means it is potentially visible. In 1969 the English theoretician Roger Penrose, however, proposed a metaphysical 'law' called the *cosmic censorship principle*, which asserts that naked singularities are impossible. Discuss the obvious tension between Penrose's principle and the Big Bang singularity. (See, for example, P. Kosso, "Spacetime Horizons and Unobservability," *Studies in History and Philosophy of Science*, June 1988, pp. 161–173.)

---

[137] O. Saari, "The Time Bender," *Amazing Stories*, August 1937. In this story we read that the time traveler "could not travel into the past for the plate had to exist in all ages traveled, and it had not existed before he made it."

[138] E. L. Rementer, "The Time Deflector," *Amazing Stories*, December 1929.

[139] K. S. Thorne, "Do the Laws of Physics Permit Closed Timelike Curves?" *Annals of the New York Academy of Science*, August 10, 1991, pp. 182–193. Science fiction writer Damon Knight (1922–2001) anticipated Thorne's rebuttal in his story "Azimuth 1, 2, 3, . . . ," *Isaac Asimov's Science Fiction Magazine*, June 1982.

In the text I mention the "transparent aluminum" bootstrap paradox that appears in the 1985 movie *Star Trek IV*. Even earlier, a movie bootstrap appeared in 1980 film *The Final Countdown*. There, the designer of a modern naval warship that temporarily time travels back to the Pearl Harbor of December 6, 1941, turns out to be a crew member who was (is) accidently left behind in the past when the ship returns to the present. In the past he *will be* able to design the ship because he already knows how it *was* designed—by himself! In the more recent 2014 film *Interstellar*, a wormhole near Saturn is discovered. By the end of the film we learn that it was put there by future humans, humans who exist because their ancestors (us!) were saved from a planet-wide ecological disaster when they used the wormhole to discover new worlds in far-flung regions of the universe. Decide whether or not the existence of the wormhole represents a bootstrap paradox, and defend your position.

One difficulty in using a black hole as a means of traveling from one region of the universe to another (with time travel as a special case) is simply getting to a black hole in the first place. The nearest one to Earth, as far as is known, is many light years distant. One reason for this may be an anthropic one (see note 13 in the "Introduction"). That is, a planet near a rotating black hole would either be eventually swallowed whole, or have its surface blasted by a firestorm of radiation produced by in-falling matter. In any case, no intelligent life able to recognize time travel would ever evolve on such a planet in the first place. That is, we are here to wonder about the absence of near-by black holes precisely *because* we aren't near a black hole. The lack of black holes near Earth is addressed in Joe Haldeman's 'Earth vs. Aliens' novel, *The Forever War*, by using the time dilation effect of special relativity (discussed in Chap. 3) that allows long travel distances to be covered in a reasonable time (as measured by clocks in rocket ships traveling near the speed of light). Still, while the travel time to reach a black hole distant from Earth by many light years may only be 6 months of ship time, back on Earth many *years* may pass. Once at the black hole the ship enters it and instantly 'jumps' to a vastly different region of the universe. In the novel, no time travel after the time dilation experienced in just getting to the black hole occurs, but Haldeman uses that to great effect as follows. Before entering into combat, Earth's soldiers are told that when they exit the hole into a new region of the universe, they may encounter alien warships equipped with *their* latest technology, technology that could be far in advance of the Earth warship's technology

(continued)

which dates from Earth's *past*. That is, humans will be fighting against technology that dates from the Earth warship's *future*. To quote the novel, "Relativity traps us in the enemy's past; relativity brings them from our future." Explain this.

In one of the quatrains of the *Rubaiyat*, the eleventh century Persian poet-philosopher Omar Khayyam wrote

> *The Moving Finger writes; and having writ,*
> *Moves on; nor all your Piety nor Wit*
> *Shall lure it back to cancel half a Line,*
> *Nor all your Tears wash out a Word of it.*

Nearly a 1000 years later the German theoretical physicist Hermann Weyl (1885–1955), a colleague of both Einstein and Gödel at the Institute for Advanced Study in Princeton, NJ, wrote the following in his book *Space-Time-Matter* (published in 1921, three decades *before* Gödel's 1949 time travel paper):

> *It is possible to experience events now that will in part be an effect of my possible future resolves and actions. Moreover, it is not impossible for a world-line (in particular, that of my body), although it has a time-like* [see the index] *direction at every point, to return to the neighborhood of a point which it has already once passed through. The result would be a spectral image of the world more fearful than anything the weird fantasy of E. T. A. Hoffmann* [an early nineteenth-century German writer of the eccentric] *has ever conjured up. In actual fact the very considerable fluctuations of the* [components of the metric tensor, to be discussed in Chap. 3] *that would be necessary to produce this effect do not occur in the region of the world in which we live … Although paradoxes of this kind appear, nowhere do we find any real contradictions to the facts directly presented to us in experience.*
> Compare these two views and, in particular, discuss what each says about the idea of 'reliving the past.'

In the opening section of this book ("Some First Words") I mention how now and then science fiction has anticipated physics. One interesting example of this occurs in a story of a time traveler *almost* meeting himself, a story published 2 years *before* Gödel's 1949 time travel paper in which he suggests just such a possibility. The story opens with a man on a ship spotting the signal fire of a castaway on a Pacific island, as well as the tiny, distant figure

(continued)

of a man waving and jumping about. While sailing in to help, the ship hits a mine left over from the war, and the would-be-rescuer becomes a castaway, too. After swimming to the island, he can find no trace of who built the fire, although there are footprints all about in the sand. Exploring the island, he finds the remains of a crashed interstellar spaceship (!), powered by a drive unit based on 'temporal precession.' The man, curious, turns the drive on and thus sends himself backward in time by one day. He then spots a ship on the horizon, builds a fire, waves and jumps about, then *recognizes the ship as his own* …. And so the loop nearly but not quite closes. The man, apparently, rushes off into the jungle, terror-stricken at the thought of meeting himself. (You can find this tale by A. B. Chandler (1912–1984), "Castaway," in the November 1947 issue of *Weird Tales*, a publisher more of fantastic, supernatural, and horror stories than of science fiction. Perhaps easier to locate would be an anthology in which it has been reprinted: *Science-Fiction Adventures in Dimension* (G. Conklin, editor), Vanguard 1953.)
Speculate on what happens to the man. In particular, does Wells' own criticism of *The Time Machine*, concerning "vain repetitions" of time travelers, apply here (see the "Introduction" again)?

A philosopher has argued against the force of Hawking's chronology protection conjecture as follows: "There is an old argument to the effect that while backward time travel may be *possible*, it will never actually occur—for if it were going to occur, we would *already* have encountered the time travellers involved, whereas in fact we have done no such thing. … But consider an isolated society living in a remote part of the world. Some members of this society are engaged in a long-running debate concerning the possibility of human flight. Were a 747 to pass overhead, would the debaters necessarily recognize it as containing flying humans? The answer to their question might have been staring them in the face for years, without them realizing." (See Nicholas J. J. Smith, "Bananas Enough for Time Travel?" *British Journal for the Philosophy of Science*, September 1997, pp. 363–389, in particular note 3 on p. 364. The perhaps curious appearance of 'bananas' in the title of this paper will become clear when, in Chap. 4, we delve into the details of the famous grandfather paradox.) How would *you* answer Smith's question? Do you think it is plausible, as Smith implies, that we could right now be observing (without realizing it) effects in the present-day world that are the result of time travelers amongst us? What sort of effect (s) might raise this suspicion in *your* mind?

As discussed in Sect. 1.6, Tipler expressed some pessimism in his 1977 paper (note 134) about the possibility of actually constructing a time machine from a rotating cylinder. But that doesn't mean he didn't have some doubts, too, about theoretical 'proofs' of something being impossible. In his 1976 PhD dissertation, for example, he included an amusing reference to Simon Newcomb (a real-life mathematician that Wells' Time Traveller cites in *The Time Machine*—see note 102 in Chap. 2) who published mathematical 'proofs' that it would be impossible with known science to build a "practicable machine by which men shall fly long distances through the air." Why do you think Tipler did that? You can read more about Newcomb's 'proofs' in "Is the Airship Coming," *McClure's Magazine* (September 1901, pp. 432–435) and "The Outlook for the Flying Machine," *The Independent* (October 22, 1903, pp. 2508–2512).

In his autobiography, the Princeton physicist John Wheeler had this to say about time: "The smooth flow of time—or our smooth passage through it—is an illusion that is shattered when we . . . ask about time at the moment of the Big Bang, at a moment of gravitational collapse, at the moment of the Big Crunch. Students and others often ask what existed before the Big Bang. To say that we don't know is not to say enough. Even to say that we have no *way* of knowing is not enough. We really have to say that space and time came into existence, along with matter and energy and the laws of physics, at the moment of the Big Bang. If the universe expands to a maximum size, starts contracting, and eventually collapses to a fiery death—a fate that seems likely to me and to some other theorists . . . then time and space, too, will end in this Big Crunch. I can reach no conclusion other than this: there was no 'before' before the Big Bang, and there will be 'after' after the Big Crunch." (See J. A. Wheeler, *Geons, Black Holes, and Quantum Foam: a life in physics*, W. W. Norton 1998, pp. 349–350.) That is, when the Big Bang singularity occurred, time was *created*, and if the universe should collapse in the far future in a Big Crunch, time will be *annihilated*. This is a view of *nothingness* that transcends even that of the grave. Sharing Wheeler's dark view of the ultimate fate of reality, but instead giving the *victory t*o time (rather than its annihilation), was the Irish writer Jonathan Swift (1667–1745) in his poem *Riddles* (circa 1724): "Ever eating, never cloying/All-devouring, all destroying/Never finding full repast/ Till I eat the world at last." How do you think theologians would respond to Wheeler and Swift?

The American philosopher Roy Sorenson was cited in note 13 in the discussion of the difficulties a time travel would have in convincing skeptics that he *had really* time traveled (short of bringing a fresh dinosaur egg back and hatching it!). This question was treated in early pulp science fiction (“The Sands of Time,” see note 43 in “Some First Words”) as follows: The time traveler takes a sealed box of pure radium (with his name written on the inside of the lid) into the distant past, and buries it in a secure location. Upon returning to the present he unearths the box; testing of the contents will show that some of the radium has radioactively decayed to lead. Indeed, the amount of decay would be a direct measure of how far back into the past the box had been transported. This issue was later elaborated on by the English philosopher Alasdair Richmond in his paper “Time Travel, Parahistory and the Past Artefact Dilemma,” *Philosophy*, July 2010, pp. 369–373. There he imagined two possible ways a time travelling Shakespearean scholar might attempt to convince skeptical colleagues that he had discovered a draft of *Hamlet* dating from the year 1589 (10 years *before* the earliest accepted date of its composition by Shakespeare). The first attempt is to simply bring that draft directly back with him in the time machine, from 1589 to the present. Then, of course, many of the inherent clues as to the draft’s authenticity, such as chemical composition of the ink, the weave of the paper, and orthography (the style of writing in 1589) would be consistent with the time traveler’s claim, but other clues would *not*—the *age* of the paper and of the ink, for example, would be taken as evidence fatal to the claim, as they would not be nearly 430 years old. They would appear, in fact, to be practically new! The draft would, therefore, be dismissed as simply a clever forgery. The second attempt would try to get around this problem, as follows. After locating the draft in 1589, the time traveler *doesn’t* bring it back to the present, but rather stashes it away in a secret hiding place. Then, once back in the present, he takes his colleagues to the secret hiding place and, with a flourish, reveals the draft which now *is* nearly 430 years old. Much to the time traveler’s frustration, however, his colleagues still reject his time travel claim, this time saying he must have simply found the draft in the ‘usual’ way (under the floorboards in somebody’s attic, for example), and is just pretending to have found it via time travel. Can you think of a way, using the *Hamlet* draft, the time traveler might be able to convince his skeptical colleagues?

The fan I quoted in "Some First Words," who wrote to *Astounding Stories* in 1931 to express his unhappiness with the appearance of women in that magazine's stories, was quite clear about his concerns—although, given the times, he carefully avoided any direct mention of sex. A modern, highly successful *female* writer of science fiction, Anne McCaffrey (1926–2011), didn't shy away from that, however, when she wrote the following in a hilariously funny essay: "Prior to the '60s, stories with any sort of love interest were very rare. True, it was implied in many stories of the '30s and '40s that the guy married the girl whom he had rescued/encountered/discovered during the course of his adventures. But no real pulse-pounding, tender, gut-reacting scenes. The girl was still a 'thing' to be used to perpetuate the hero's magnificent chromosomes. Or perhaps to prove that the guy wasn't . . . I mean, all those men locked away on a spaceship for months/years at a time. I mean . . . and you know what I mean even if I couldn't mention it in the sf of the '30s and '40s." (See Anne McCaffrey, "Hitch Your Dragon to a Star: Romance and Glamour in Science Fiction," in *Science Fiction, Today and Tomorrow*, R. Bretnor, editor, Harper & Row 1974, pp. 278–292.) Modern time travel science fiction has shown a huge change (for the better) on this score. Discuss, for example, the emotional power of a love story between a couple separated in time, as depicted in the 1975 novel *Bid Time Return* by Richard Matheson (made into the 1980 film *Somewhere in Time*). How do you think the 1931 fan would have reacted to Matheson's story? (Indeed, if that fan was a teenager—or even a few years older—in 1931, then 44 years later he would have been, at most, in his mid-60s and might well *have* read the novel.)

In the 2014 film *Interstellar*, a space probe dives into a black hole, gets a glimpse of the hole's singularity, measures some unspecified quantum effects, and then sends the measurements back to Earth (via the fifth dimension) as a signal in the form of spasmodic Morse code twitches of the second-hand on somebody's watch. This all leads (it is hinted) to a theory of quantum gravity. If you saw the second-hand on *your* watch suddenly begin to spasmodically twitch, would you then immediately think

(a) that a Morse code message was coming to you via the fifth dimension bearing the secrets of quantum gravity?

or

(b) that your watch needs a new battery?

(continued)

or

(c) something else?

Vigorously defend your answer.

William Grey, a philosopher at The University of Queensland, pointed out numerous conceptual difficulties with the idea of time travel in his paper "Troubles with Time Travel," *Philosophy*, January 1999, pp. 55–70. That paper quickly prompted a rebuttal from the philosopher Phil Dowe (at the University of Tasmania), who replied a year later with the paper "The Case for Time Travel," *Philosophy*, July 2000, pp. 441–451. We'll eventually take up all the issues discussed in those two papers but, for now, read both papers and summarize their respective arguments. Do you feel one of the writers won the day (for you, anyway)?

Chapter 2
Philosophical Space and Time

> *"I do not believe that there are any longer any philosophical problems about Time; there is only the physical problem of determining the exact physical geometry of the four-dimensional continuum that we inhabit."*[1]

2.1 Time: What Is It, and Is It Real?

> "Time is generally thought to be one of the more mysterious ingredients of the Universe."[2]

Before going any further with time *travel*, it will be well worth the effort to take a closer look at time itself, the 'stuff' or 'thing' or . . . ? that we are interested in traveling 'through' or 'around' or 'across' or . . . ? Oddly enough, I'll start with religion, as philosophical theologians had identified time as something unusual long before Newton's words on time in his *Principia* that I mentioned in the Introduction, and many thousands of years before science fiction writers and their time travel stories.

We can, in fact, trace the religious interest in time back at least sixteen centuries to the Christian theologian St. Augustine and his *Confessions* (in which he famously admitted "What, then, is time? I know well enough what it is, provided that nobody asks me: but if I am asked what it is and try to explain, I am baffled."). Certainly the seventeenth century Spanish Jesuit Juan Eusebius Nieremberg caught the spirit of wonder that time holds for the devout when he wrote, in his *Of Temperance and Patience*, that "*Time* is a sacred thing; it flows from Heaven . . . It is an emanation from that place, where eternity springs . . . It is a *clue* cast down from Heaven to guide us . . . It has some assimilation to Divinity."

Going outside Christianity, we can easily find other equally strong reactions to the mystery of time. From Plutarch's *Platonic Questions* we learn that when the question of time's nature was put to Pythagoras, he simply uttered the mystical "time is the soul of the world." The *Laws of Manu* of Hinduism, the *Torah* of

---

[1] H. Putnam, "Time and Physical Geometry," *Journal of Philosophy*, April 1967, pp. 240–247.

[2] P. Horwich, *Asymmetries in Time*, MIT Press 1987.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_2

Judaism, the *Koran* of Islam, and the revealed truths of Gautama Buddha are all full of references to time. It is, in fact, to the pagan gods of Greek mythology that we owe our ‘modern’ image of Chronos, or Father Time.

Not just the Greeks made time a god. In the *Bhagavad Gita* (*Song of the Lord*), the central religious-romantic epic of Hinduism that predates Christ by five centuries, one of the characters reveals his divine nature and declares his power thus: “Know that I am Time, that makes the worlds to perish, when ripe, and bring on them destruction.” And in the even older Egyptian Book of the Dead, which dates back over three thousand years, the newly deceased was thought literally to become one with time itself. The merging of time and the resurrection of the body after death in the Book is shown in the line “I am Yesterday, Today and Tomorrow, and I have the power to be born a second time.”

The Greek philosopher Plato (circa 400 B.C.) gave us a curious way to think of time: as a *closed loop*. While Plato did think of time as having a beginning, his conception did not have time extending off into the infinite future as does the modern, everyday view. Rather, Plato visualized time as curving back on itself—as *circular* in nature. This was, in fact, a reasonable reflection on what Plato could see everywhere in nature, with the seemingly endless repetition of the seasons, the regular ebb and surge of the tides (the old English word *tid* was a unit of time), the unvarying alternation of night and day, and the rotation of the visible planets in the sky. Whatever might be observed today, it seemed obvious to Plato, would happen again in nature. Circular time in science fiction was briefly mentioned in Chap. 1,[3] and it occurs outside that genre, too, as in James Joyce’s novel *Finnegans Wake*, which opens in mid-sentence and ends with the first part of the same sentence. This view of time has a powerful, ancient visual symbol, the Worm Ouroborous, or World Snake, that eats its own tail endlessly.

Circular time, with its closed topology, was favorably presented in Stephen Hawking’s famous book *A Brief History of Time*. In it he concludes that there is no need for God because in circular time there is no first event and hence no need for a First Cause. Vigorous philosophical rebuttals were quick to come, of course![4]

Turning to fiction, Ray Bradbury wrote a beautifully poetic passage about the mystery of time in “Night Meeting,” one of the splendid sub-stories in his episodic 1950 masterpiece *The Martian Chronicles*. A man of A.D. 2002, who is one of the modern inhabitants of Mars, somehow meets the ghostly image of a long-dead Martian one cold August night. The conditions are just right for such a cross-time encounter. As the man thinks to himself, “There is the smell of Time in the air tonight. . . . There was a thought. What did Time smell like? Like dust and people. And if you wondered what Time sounded like it sounded like water running in a dark cave and voices crying and dirt dropping down on hollow box lids, and rain. And, going further, what did Time *look* like? Time looked like snow dropping silently into a black room or it looked like a silent film in an ancient theater, one hundred billion faces falling like those New Year balloons, down and down into nothing. That was how Time smelled and looked and sounded. And tonight . . . tonight you could almost *touch* Time."

---

[3] Another example from science fiction is the story by I. Hobana, “Night Broadcast,” in which a television signal from the past is picked up by a gadget that is probing the *future*: “By going far enough into the future one comes upon what we call the past.” You can find this tale in the *Penguin World Omnibus of Science Fiction*, Penguin Books 1986.

[4] See, for example, W. L. Craig, “What Place, Then, for a Creator?: Hawking on God and Creation,” *British Journal for the Philosophy of Science*, December 1990, pp. 473–491, and R. Le Poidevin, “Creation in a Closed Universe Or, Have Physicists Disproved the Existence of God?,” *Religious Studies*, March 1991, pp. 39–48.

Well, lovely words, yes, but they don't really tell us what time *is*. Perhaps Einstein the physicist can tell us. In the *New York Times* of December 3, 1919, we find him quoted as follows: "Till now it was believed that time and space existed by themselves, even if there was nothing [Newton's view]—no Sun, no Earth, no stars—while now we know that time and space are not the vessel for the Universe, but could not exist at all if there were no contents, namely, no Sun, no Earth, and other celestial bodies." Less than 2 years later Einstein stated this view again (*New York Times*, April 4, 1921): "Up to this time the conceptions of time and space have been such that if everything in the Universe were taken away, if there were nothing left, there would still be left to man time and space." Einstein went on to deny this view of reality, saying that, according to his general theory of relativity, time and space would *cease to exist* if the universe were empty. This has the ring of one of Einstein's favorite philosophers, Spinoza, who declared in his *Principles of Cartesian Philosophy* that "there was no Time or Duration before Creation." In a correspondence with Samuel Clarke—Newton's friend who translated Newton's *Optiks* into Latin—the German philosopher Gottfried Leibniz (who began the correspondence in 1715) expressed similar ideas: "Instants, consider'd without the things, are nothing at all . . . they consist only in the successive order of things."

The pragmatic scientist would certainly agree with Leibniz. After all, what could it even mean to talk of time unless you can measure it? And what you use to measure time is a clock—some kind of changing configuration of matter involving spinning gears, ticking pendulums, and rotating dial pointers. Mere *unchanging* matter, alone, is not sufficient to measure time because a still clock measures nothing. *Changing* matter seems to be required. Yet, not surprisingly, not everybody agrees. The counterview, the view that time has nothing to do with change, was expressed in an interesting manner by a science fiction fan in a letter to the editor of *Wonder Stories* (January 1931): "Just one thing, you have these time-traveling yarns, good stuff to read all right, but bunk, you know; because if there's no such thing as time, which there isn't, *only change* [my emphasis], how can one travel in . . . something that doesn't exist. To our planet which goes around the Sun there is simply a turning and warming of one side and then the other, i.e., years, days, minutes, etc., is something purely artificial, invented by man to tell him when to do certain things, work and stop work . . ."[5]

---

[5] This fan's idea was not new. For Plato's most famous student, Aristotle, time was *motion* (in a world in which nothing moved, argued Aristotle, there would be no time), and he expressed this

Going even beyond the ideas of Einstein, Spinoza, Leibniz, Plato, Aristotle, and our science fiction fan, at least one metaphysician felt that time would have no meaning, even in a massive and changing universe, without the additional presence of conscious, rational beings.[6] That sounds very much like an echo of the French philosopher Henri Bergson who, in 1888, somewhat mysteriously declared that time is "nothing but the ghost of space haunting the reflective consciousness." A few years before Taylor, however, a fellow philosopher had argued for exactly the opposite view, that temporal passage is independent of the existence of conscious beings.[7]

All this divergence of opinion perhaps explains why even a lightweight Hollywood movie like Mel Brooks' 1987 *Spaceballs* can get a laugh from a time joke. Even kids know that the characters, when talking about time, haven't the slightest idea of *what* they are talking about. The movie, a spoof on such classic films as *Star Wars*, *The Wizard of Oz*, and *Raiders of the Lost Ark*, quickly reaches a point of crisis. To find out what to do next, the evil Lord Helmet and his chief henchman decide on a novel approach: they will look at an instant video of their own movie! (Instant videos are available *before* the movie is finished.) Perplexed at watching on a television screen everything that he is doing as he does it (the screen correctly shows an infinite regression of television screens, each being watched by a Lord Helmet), Lord Helmet initiates the following rapid-fire exchange. (It is, of course, a clever take-off on Abbott and Costello's "Who's on First?")

> *What the hell am I looking at? When does this happen in the movie?*
> *Now! You're looking at now, sir. Everything that happens now, is happening, now.*
> *What happened to then?*
> *We're past that.*
> *When?*
> *Just now, now.*
> *Go back to then.*
> *When?*
> *Now.*
> *Now?*
> *Now.*
> *I can't.*
> *Why?*
> *We missed it.*

---

view in his famous metaphor "Time is the moving image of eternity." For Aristotle, then, time and change were inseparably intertwined. For Aristotle the world had existed for eternity, and the circularity of time was a central and powerful image; using his vivid illustration, it is equally true *in circular time* that we live both before *and* after the Trojan War.

[6] R. Taylor, "Time and Life's Meaning," *Review of Metaphysics*, June 1987, pp. 675–686.

[7] S. McCall, "Objective Time Flow," *Philosophy of Science*, September 1976, pp. 337–362.

> *When?*
> *Just now. [The henchman then sets the video to rewind.]*
> *When will then be now?*
> *Soon.*

We may laugh at this, even dismiss it as mere movie madness, but could any of us *really* do much better if, like Saint Augustine, we were backed into a corner and asked to explain time? Somehow, I think even the distinguished twentieth-century Harvard professor Hilary Putnam whose words open this chapter would find it difficult to know where to begin. He might even become as confused as the time traveler in the 1968 film *Je t'aime, Je t'aime*, whose oscillations in time, from present to past and back again, leave him so befuddled that he decides he'd rather be dead. What, then, *can* we say about time? Despite Putnam's bold words, I suspect that most people would come down on the side of Augustine.

The mystery of time was well captured by R. H. Hutton (1826–1897), the literary editor of the *Spectator*, when he wrote in his 1895 review (see note 1 in the Introduction) of Wells' *Time Machine* that "the story is based on that rather favorite speculation of modern metaphysicians which supposes *time* to be at once the most important of the conditions of organic evolution, and the most misleading of subjective illusions ... and yet Time is so purely subjective a mode of thought, that a man of searching intellect is supposed to be able to devise the means of traveling in time as well as in space, and visiting, so as to be contemporary with, any age of the world, past or future, so as to become as it were a true 'pilgrim of eternity.'"

Novelist Israel Zangwill (1864–1926) wrote a similar but much more analytical review of Wells' novel for the *Pall Mall Magazine* (see note 1 in the Introduction). Zangwill was the only Victorian reviewer to attempt a scientific analysis of time travel. Although he thought Wells' effort was a "brilliant little romance," Zangwill also thought the time machine—"much like the magic carpet of *The Arabian Nights*"—was simply "an amusing fantasy." Zangwill continued in his review with what was even then a common idea about a way one might actually be able, at least in principle, to look backward in time; one could travel far out into space by going faster than light and then watch the light from the past as it catches up to you. (Note, carefully, that Zangwill was writing in 1895, 10 years before Einstein's special relativity put a limit on possible speeds.) In this way, Zangwill wrote, one could watch "the Whole Past of the Earth still playing itself out."

Indeed, even before Zangwill, the well-known French astronomer Camille Flammarion (1842–1925) had made this dramatic idea a centerpiece of his 1887 novel *Lumen*. That book, a best-seller in Europe even before its appearance in England, describes how a man just dead (in 1864) instantly finds his spirit on the star Capella, where he is able to watch the light then arriving from the Earth of 1793. In particular, he watches the French Revolution play itself out and sees himself as a child. Flammarion may have, in fact, been inspired to write his novel by an essay written several years earlier (in 1883) by the British physicist

J. H. Poynting (1852–1914). Poynting's essay,[8] which opens with the statement that it was, in turn, inspired by an anonymous pamphlet published "30 or 40 years ago" on the same topic, specifically mentions watching historical events from Capella.

By the beginning of the twentieth century the idea of watching the past by outrunning light had drifted down into juvenile literature, as in the 1904 novel *Around a Distant Star* by Jean Delaire (the pen name for Pauline Touchemoline (1868–1950)), in which a young man builds a spaceship that can travel at two thousand times the speed of light. With it, he and a friend travel to an Earth-like planet nineteen hundred light-years distant and use a super-telescope to watch the Crucifixion (and then the resurrection) of Jesus. Early magazine science fiction also found the idea of looking backward in time with delayed light to be an irresistible one, involving romance and murder.[9] In another tale incorporating human emotions, a scientist loses his wife to a rival who kidnaps her and then escapes in a faster-than-light rocket ship headed for parts unknown. After searching for them with his own brilliant invention of the 'ampliscope' (several quantum leaps beyond the telescope), the scientist locates the couple, skipping from planet to planet light-years distant. His only pleasure, then, is to use his own faster-than-light craft to outrun the images of his lost love and watch them over and over. Eventually, however, he comes to realize the ultimate futility of it all. As the final line of this sad tale says, "It would be senseless, I knew, chasing on and on after yesterdays."[10]

The reality of time received a new twist with the additional imagery of instants of time being likened to the points on a *straight* line. In the West it was the Christian theological doctrine of *unique* historical events that gave rise to *linear* time in the minds of the common folk. The creation of the world and Adam and Eve, the adventures of Noah and the cataclysmic Flood, the Resurrection—these were all events that occurred in sequence, *once*. None would happen again and so, for Christianity, circular time just would not do.[11] In addition, it has been argued that the major spiritual content of Christianity—a significant reason for its popular support even in the face of brutally harsh Roman suppression—is that it brought the *expectation of change* into the static world of ancient times. It was, in fact, in ancient religious teachings that our modern view of linear time had its origin, a view that most people today (including the most hardened agnostic physicist) find to be as natural as Plato and Aristotle found circular time.

---

[8] J. H. Poynting, "Overtaking the Rays of Light," in Poynting's *Collected Scientific Papers*, Cambridge University Press 1920.

[9] As in, for example, G. A. England, "The Time Reflector," *The Monthly Story Magazine*, September 1905.

[10] D. D. Sharp, "Faster Than Light," *Marvel Science Stories*, February 1939. The year before saw the appearance of a story with the same idea, a story that specifically cites Flammerion: M. Weisinger, "Time On My Hands," *Thrilling Wonder Stories*, June 1938.

[11] Still, just to show how one can find support for almost any view in the same religious dogma, Ecclesiastes 1:9 would seem to be a claim *not* for linear time but rather for circular time!: "The thing that hath been, it is that which shall be; and that which is done is that which shall be done; and there is no new thing under the sun."

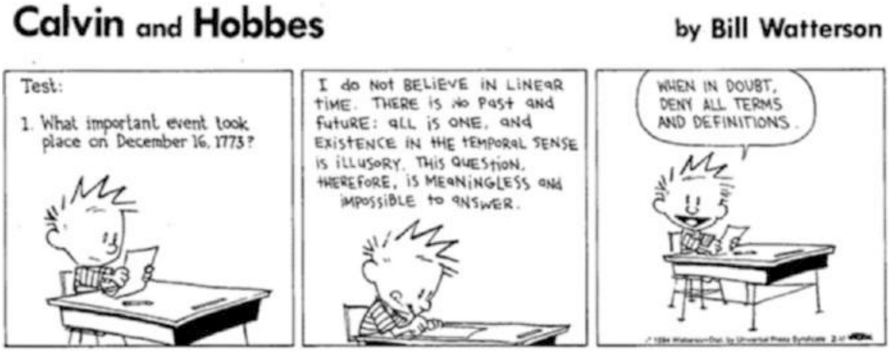
Calvin and Hobbes
by Bill Watterson
Test:
1. What important event took place on December 16, 1773?
I do not BELIEVE IN LINEAR TIME. THERE IS NO PAST AND FUTURE: ALL IS ONE, AND EXISTENCE IN THE TEMPORAL SENSE IS ILLUSORY. THIS QUESTION, THEREFORE, IS MEANINGLESS AND IMPOSSIBLE TO ANSWER.
WHEN IN DOUBT, DENY ALL TERMS AND DEFINITIONS.

CALVIN AND HOBBES ©1994 Watterson. Reprinted with permission of UNIVERSAL PRESS SYNDICATE. All rights reserved.

Even though linear time was the norm after Christ, there were still enough questions about time to perplex the deepest of thinkers, and the next 2000 years resulted in plenty of thinking. Discourses on time by such philosophers as Descartes, Spinoza, Hobbes, Kant, Nietzsche, and Hegel can be found by the yard in any decent university library. Nearly all (if not indeed all) of these presentations have metaphysical, even theological, underpinnings. For example, Descartes is generally believed to have argued for a discontinuous, atomistic nature to time (recall the *chronon* from Chap. 1). This is the modern view of his thinking, because in his *Meditations* (1641), in particular in the third meditation on God's reality, Descartes appears to argue that God must continually recreate the world at each *separate* moment of its existence. That is, the world is recreated in a discontinuous succession of *individual* acts by God.[12]

Finally, with Newton's discussion of *absolute* time, which is the belief that time is the same everywhere in the universe, there was for the first time a *physicist* writing about time (although, as I mention in Chap. 1, Newton's views were also influenced *heavily* by theological considerations, in addition to mathematical physics). But, despite Newton's genius, the mystery of time remained a mystery.

In 1905 Einstein's name appeared among the contributors to the study of time, and so at last something besides metaphysical speculation on the subject was added to the body of human thought. Einstein's paper on special relativity introduced the revolutionary idea of *relative* time, which is the anti-Newton belief that the passage of time is *not* the same everywhere, but rather depends on local conditions. In retrospect, Einstein's 1905 work seems to be the perfect reply to the comment by Isaac Barrow (1630–1677)—Newton's teacher and the first Lucasian professor of mathematics at Cambridge (the chair once held by Stephen Hawking centuries later)—that "because *Mathematicians* frequently make use of Time, they ought to have a distinct idea of the meaning of the Word, otherwise they are Quacks."

---

[12] For more on this, see R. T. W. Arthur, "Continuous Creation, Continuous Time: A Refutation of the Alleged Discontinuity of Cartesian Time," *Journal of the History of Philosophy*, July 1988, pp. 349–375.

Then, just 3 years after Einstein, along came a second astonishing paper by the Cambridge philosopher John Ellis McTaggart (1866–1925). This paper[13] claims to prove that whatever time might be *thought* to be (even by Einstein), it really isn't that because time isn't even real. (This would seem, I think you'd agree, to have potentially profound implications for time travel!) The method of the paper is to deny the reality of time via an infinite-regress argument that one philosopher[14] has called the *pons asinorum* ("bridge of asses") of the riddle of time. As McTaggart's own opening sentence freely admits, "It doubtless seems highly paradoxical to assert that Time is unreal, and that all statements which involve its reality are erroneous."

McTaggart began his analysis by observing that there are two separate and distinct ways of talking about events in time. Following his terminology, one can say that events are either future, present, or past (the so-called *A-series*), or one can say that events temporally ordered by each being later than some other events, earlier than others, and simultaneous with still others (the so-called *B-series*). He then continued by asserting that time requires change, and followed that with the observation that the A-series (but not the B-series) incorporates such change. That is, if event X is earlier than event Y, then X is *always* earlier than Y and thus there is no change in this (or in any other) example of a B-series. As a specific example, let Y be the birth of a child, and let X be the birth of its mother. In contrast, if X is first in the future, then is in the present, and finally is in the past, then we have an example of change (and hence of *time*) in the A-series; for example, let X be the next time you blink.

With this rather pedestrian start, McTaggart then pulled his rabbit out of the hat. It makes no sense, he argued, to talk of the 'future,' 'present,' and 'past' of an event because these terms are mutually exclusive. That is, no two of these predicates can apply at once, and yet, paradoxically, every event possesses all three and thus we have a contradiction. It therefore, concludes McTaggart, makes no sense to talk of future, present, or past. And because it makes no sense to talk of them, they do not exist, and so there can be no A-series and hence no change, and thus no reality to time. McTaggart apparently realized just how befuddling all that would appear to just about everybody who read it, and so he played devil's advocate (D.A.) in his paper by trying to anticipate the various objections people could raise. Of course, he always managed to refute the D.A. at every turn. It is worth the effort to go through the details of McTaggart's 'proof,' as that will make it clear what there is about

---

[13] J. E. McTaggart, "The Unreality of Time," *Mind*, October 1908.

[14] L. O. Mink, "Time, McTaggart and Pickwickian Language," *Philosophical Quarterly*, July 1960, pp. 252–263. The phrase *pons asinorum* has its origin in a plane geometry theorem: the angles opposite the equal sides of an isosceles triangle are themselves equal. Seeing the truth of this is said to separate the quick-witted from the dull. It isn't clear (to me, anyway), however, on which side of McTaggart's 'proof' the quick-witted were imagined to fall. You'll see what I mean in just a moment.

'traditional' philosophical reasoning that so irritates modern philosophers trained in mathematical physics (and what makes physicists roll their eyes when confronted with arguments like McTaggart's).

The predicates of future, present, and past are really not incompatible for any event, the D.A. says some will claim, because the real predicates we should use are '*was* future,' '*is* present,' and '*will be* past,' and these *can* be possessed all at once by any event. Nice try, counters McTaggart, but that will not solve the problem. By allowing such modified predicates, we must actually allow for all nine possibilities, some of which are still incompatible. That is, the 'was,' 'is,' and 'will be' could each be potentially attached to 'future,' 'present,' and 'past': for example, 'was past' is incompatible with 'will be future.'

Oh, counters the D.A., we can eliminate that concern by allowing even more complex predicates to arrive at a third level of structure, such as 'is going to have been past,' and 'was going to be future,' and those *are* compatible. But McTaggart swats that argument away, too, by displaying new incompatibles, as well as by showing that the process of ever-increasing predicate complexity is a vicious infinite regress that drags along the seeds of its own doom at every step.[15] There is simply no escape from incompatibility, he says, and so there is no time.

Well! What can one do when presented with such an argument, one that seems to claim philosophers can wrest free the secrets of nature by pondering the historical accidents of English syntax? As David Hume once said, "Nothing is more usual than for philosophers to encroach on the province of grammarians, and to engage in disputes of words, while they imagine they are handling controversies of the deepest importance and concern." One modern philosopher apparently agreed with Hume, at least in the case of McTaggart's 'proof,' and he was pretty blunt with his evaluation of it: "McTaggart's famous argument for the unreality of time is so completely outrageous that it should long ago have been interred in decent obscurity. And indeed it would have been, were it not for the fact that so many philosophers are not sure that it has ever really been given a proper burial, and so from time to time someone digs it up all over again in order to pronounce it *really*

---

[15] Here's a clever way to systematically generate McTaggart's infinite regress of complex predicates, as presented by M. Dummett, "A Defense of McTaggart's Proof of the Unreality of Time," *Philosophical Review*, October 1969, pp. 497–504): "Let us call 'past,' 'present,' and 'future' 'predicates of first level.' If, as McTaggart suggests, we render 'was future' as 'future in the past,' and so forth, then we have nine predicates of second level, where we join any of the three on the left with any of the three on the right:

past		past
present	in the	present
future		future

Similarly, there are twenty-seven predicates of third level ... "Dummett's construction clearly shows that, at the N-th level, there are 3^N predicates, most of which are incompatible.

dead. These periodic autopsies reveal that something more remains to be said."[16] That is certainly true, in as much as McTaggart's disarmingly innocent argument has caused disagreement and furrowed brows among philosophers for decades.

It is, in fact, easy to find examples of the continuing debate over McTaggart's analysis and, as silly as it strikes physicists, it still has a pulse in some quarters. While at least one philosopher has argued that McTaggart simply didn't really understand his own proof, this philosopher nevertheless agreed with McTaggart's conclusion about the unreality of time.[17] Another writer has illustrated how McTaggart's ideas have found their way into modern philosophical debates on the meaning of time in the cinema, particularly in the analysis of *anachrony*, the telling of a story out of normal time sequence, such as occurs in time travel movies.[18]

Other sorts of metaphysical proofs for the unreality of time have been offered besides McTaggart's. For example, it has be argued that time is unreal, at least in a world empty of consciousness, because the concepts of past, present, and future could not possibly have any meaning unless events could be remembered, experienced, and anticipated. Or, for a second example, some have held time to be unreal, at least in a deterministic world (as some argue four-dimensional spacetime to be), because any event whose occurrence follows from present conditions, and from physical laws, would exist (they say) *now*. This view, which seems to assert that everything should happen at once, I personally find to be sufficiently obtuse as not to be bothered by it.[19] Debates between those who believe in the common-sense idea that present, past, and future are attributes of events (the 'tensers') and those who deny it (the four-dimensional spacetime, block universe 'detensers') continues to now and then still flair up on the pages of philosophy journals. At least one philosopher likes both views![20] Most modern physicists, I think, simply don't care about this line of inquiry.

On the other hand, less than a month before his death Einstein revealed his feelings about the meaning of present, past, and future, and his words appear to be ones that show some sympathy to the philosophers. In a letter written on March 21, 1955, to the children of his dearest friend who had just died, Einstein wrote—with full knowledge that his own illness would be his last—"And now he has preceded me briefly in bidding farewell to this strange world. This signifies nothing. For us believing physicists, the distinction between past, present, and future is only an illusion, even if a stubborn one."[21] Later in this chapter I'll return to these curious words and speculate on what Einstein may have meant by them.

---

[16] F. Christensen, "McTaggart's Paradox and the Nature of Time," *Philosophical Quarterly*, October 1974, pp. 289–299.

[17] Q. Smith, "The Infinite Regress of Temporal Attributions," *Southern Journal of Philosophy*, Fall 1986, pp. 383–396. To this came a rebuttal a year later by L. N. Oaklander, in the same journal (Fall 1987, pp. 425–431).

[18] G. Currie, "McTaggart at the Movies," *Philosophy*, July 1992, pp. 343–355.

[19] But if, upon reflection, it starts to bother *you*, see R. Gale, "Some Metaphysical Statements About Time," *Journal of Philosophy*, April 1963, pp. 225–237. We'll soon get to some of the more common philosophical questions on the nature of four-dimensional spacetime, such as 'is it deterministic or is it fatalistic?,' and 'does free-will have any meaning in four-dimensional spacetime?' Even physicists are interested such questions!

[20] R. Weingard, "Space-Time and the Direction of Time," *Nous*, may 1977, pp. 119–131.

I started this opening section on a religious note, and I'll end it on one. If you think the philosophical speculations on the nature of time that I've so far cited are 'really far out,' here's yet another one that leaves all the rest in the dust. In a paper that took real nerve to write (or, perhaps, simply a wicked sense of humor—and I write that in pure admiration) we read of how a spacetime that supports time travel can give the start for a *physics* explanation to the theological concept of Hell! After introducing just a bit of elementary spacetime physics (which I'll skip describing here because we'll do it later in the book), the author[22] shows how to 'construct' a compact region in spacetime (Hell) with the following properties:

1. While "so small even the Hubble Telescope couldn't image it" it can hold an infinity of physical beings;
2. Each of the beings in it are doomed, because of its time travel property, to an infinitely long personal future of damnation;
3. Each of the beings in it, because of its time travel property, can view all the future stages of their own personal damnation and so be "continually presented with a reminder of the impossibility of escape—a refinement no causally normal Hell can seemingly offer." In other words, and not to be too ironic about it, 'Theological Progress Through Physics!';
4. Each of the beings in it are continually being compressed together ("brought into dismaying proximity" with themselves) and so will spend eternity "listening to a cacophony" of their own cries of despair from *their personal future*.

There's more, but that's probably enough for you to get the idea. Richmond does admit that, as it stands, his time travel creation of Hell is not compatible with either quantum theory or even general relativity. Still, it *is* something to ponder, don't you think, when the subject of time travel comes up!

2.2 Linear Time and the Infinity of Past and Future

> "A thousand years is a huge succession of yesterdays beyond our clear apprehension."[23]
>
> —H. G. Wells

---

[21] Quoted from B. Hoffmann, *Albert Einstein: Creator & Rebel*, New American Library 1972, pp. 257–258.

[22] Alasdair M. Richmond, "Hilbert's Inferno: Time Travel and the Damned," *Ratio*, September 2013, pp. 233–249.

[23] This line appears in Wells' 1944 doctoral thesis, written for the University of London. You can find an abridgement of the thesis in *Nature*, April 1, 1944, pp. 395–397.

The modern concept of linear time as a straight line extending from the dim past through the present and disappearing into the misty future gives rise immediately to twin questions: "Did time have a beginning?" and "Will time ever end?" As one philosopher put it (long before physicists became seriously interested in singularities like the Big Bang) "Endings and beginnings are rooted in the very conception of time itself."[24] Starting at the beginning, we'll ask if the past has been forever? Early Biblical scholars, of course, believed the answers to both questions to be *no*.

They believed that the world came into being because of a First Cause, God's creation of everything. Those scholars expended vast quantities of energy (and, need I say it, time itself) in calculating the date of creation. Martin Luther, for example, argued for 4000 B.C. as roughly when everything, including time, began. Johannes Kepler adjusted this by a notch, to 4004 B.C., and later the Calvinist James Ussher, Archbishop of Armagh and Primate of All Ireland, tweaked it again. His date is the most impressive of all, at least in detail: the first day of the world was 4003, 70 days, and 6 h before the midnight that started the first day of the Christian era. Six days after that first day of the world, Adam was made, and as a final dash of specificity, this last date was declared to be Friday, October 28! Ironically, then, though Christian theology may be given credit for introducing linear time, it certainly did not provide very much of it. The beginning of time was just 6000 years or so ago, and of course The End—in the form of the Battle of Armageddon—has been awaited (with varying degrees of eagerness) for the last 1000 years.

The discovery in the seventeenth century of geological time cast a certain amount of skepticism on those early calculations concerning the duration of the past. With the discovery that the very Earth itself could be decoded for its history, the lure of trying to decode a mere book of admittedly finite age declined for most people although it cannot be denied that modern Creationists still find such a task to have its rewards). Geological time was discovered to a *chasm* of time extending backward for billions of years, a duration that is really incomprehensible for the human brain. It has become fashionable for geologists to refer to such enormous durations with the apt term *deep time*, a subtle play on the metaphor of the "ocean of time."

It is nothing less than humbling to historians who pause to think on how little of the past is known, that is, recorded. As the ever anonymous wit once put it, "History is a damn dim candle over a damn dark abyss." Still, even as enormous as is the age of the Earth, it is not infinite. But of course our planet is very old, and the universe is many billions of years older. Is the age of the universe also the duration of the past? Or is the past itself actually *infinite*?

An implicit assumption of the infinity of the past (and of the future, too) can be found in Book Three of Lucretius' science poem *De Rerum Natura* (*On the Nature of Things*) where, just before the birth of Christ, Lucretius argues for the irrationality of fearing death: "The bygone antiquity of everlasting time before our birth was nothing to us. Nature holds this up to us as a mirror of the time yet to come after our death. Is there anything in this that looks appalling, anything that means an aspect of gloom? Is it not more untroubled than any sleep?"

---

[24] I. Stearns, "Time and the Timeless," *Review of Metaphysics*, December 1950, pp. 187–200.

One philosopher[25] has traced the origins of rational support for the finite duration of the past to as far back as the sixth century A.D. The argument presented then by the Christian philosopher Joannes Philoponus of Alexandria (who is otherwise known as John the Grammarian) is simply that the world could *not* have been forever because that implies an infinity of successive acts could have taken place which (according to Philoponus) is impossible. A variation on this is the claim that if the past were infinite in extent, then everything would have happened by now! Infinity was just too big for the ancient mind (Zeno's hoary pre-Christian paradoxes, as is well-known today, are based on subtle errors in the use of infinity).

This view on the impossibility of an infinite past seems to have been the prevalent view; even as late as the twelfth century the debate among Christian theologians was not about the possibility of an infinite past, but instead about whether the Biblical 'six days of Creation' actually had taken place simultaneously. For many, the past was 'obviously' finite in duration.[26] Not all Christians accepted that conclusion, however, and the following century saw St. Thomas Aquinas (a follower of Aristotle) arguing for the opposite view of an infinite past.

Thomas' contemporary, St. Bonaventure, however, argued again for a *finite* past, and it is with Bonaventure that we start to see some mathematical sophistication.[27] Bonaventure argued that in a world infinitely old, the Sun would have made an infinite number of its annual trips around the ecliptic. But for each such trip the Moon would have made twelve monthly trips around the Earth, and so this second infinity would be twelve times as great as the first one, and how could that be? Infinity is infinity, and how can something be twelve times bigger than infinity? This argument doesn't have any strength today because of the nineteenth century German mathematician Georg Cantor's work on the concept of infinity,[28] but it *is* clever. Agonized, convoluted theological analyses of God, infinity, and eternity continued long after Aquinas and Bonaventure. Two examples should capture the spirit of those times.

---

[25] G. J. Whitrow, "On the Impossibility of an Infinite Past," *British Journal for the Philosophy of Science*, March 1978, pp. 39–45. Whitrow adds modern scientific support to the idea of a finite past by citing the prediction from general relativity of a singularity in spacetime at some finite past time; that is, the theory's prediction that time—and everything else—had its beginning in the now famous Big Bang.

[26] C. Gross, "Twelfth-Century Concepts of Time: Three Reinterpretations of Augustine's Doctrine of Creation *Simul*," *Journal of the History of Philosophy*, July 1985, pp. 325–338.

[27] See, for example, L. Sweeney, "Bonaventure and Aquinas on the Divine Being as Infinite," *Southwestern Journal of Philosophy*, Summer 1974, pp. 71–91, and S. Baldner, "St. Bonaventure on the Temporal Beginning of the World," *New Scholasticism*, Spring 1989, pp. 206–228.

[28] For simple high school-level presentations on Cantor's astonishing infinity results, see my book *The Logician and the Engineer*, Princeton 2013, pp. 169–171.

Consider first this one, on the supposed immortality of the soul. If $A = B$, then $2A = 2B$. Next, let A = 'half alive' and B = 'half dead,' where $A = B$ in the same sense that a glass half-full is also half-empty. Thus, to be completely dead is to be completely alive, and so the soul is immortal. Outrageous? *Yes*, in my opinion, but I do also have to admit the 'reasoning' does have a certain charm!

For my second example, let me begin by setting the historical stage. After publication of the English political philosopher Thomas Hobbes' *Leviathan* in 1651, with its arguments against the power of the Church and for civil power (with some criticism tossed in, as well, for universities), Seth Ward counterattacked. Ward, who was both a minister (later a bishop) in the Anglican Church and Savilian Professor of Astronomy at Oxford, was greatly offended by the secular nature of *Leviathan*. Even before Leviathan, in fact, Ward certainly would not have liked Hobbes' earlier denial of the existence of immaterial substances (such as souls). Ward's 1652 book *A Philosophical Essay Towards An Eviction of the Being and Attributes of God, the Immortality of the Souls of Men, the Truth and Authority of Scripture*, was the first of a two-punch reply to Hobbes. The second came in 1654 with the appearance of Ward's *Vindiciae academiarum*. In both of these works Ward attempted to undermine Hobbes' credibility by attacking his mathematical ability. (Hobbes had long been fascinated by, and was considered an expert on, the ancient problem of 'squaring the circle,' a task that has been known to be impossible only since 1882.[29]) In his *Essay*, Ward also attempted to defend the view that the world has a finite age—that is, it had a specific moment of creation, presumably by God. In an opening note, in fact, Ward cites Hobbes' rejection of immaterial substances as the motivation for his writing *Essay*.

To support his view of a finite age for the world, Ward invoked infinity in an interesting way. He argued that nothing is permanent, certainly not humans. Each is created; one can imagine tracing a chain of creation events backward in time through successive generations. Now, there are only two separate and distinct possibilities to where this chain could lead to in the past. First, it could terminate, after a finite number of generations, at a *first* generation, that is, with the 'creation' of the first human. If that is the case, then, said Ward (in effect), 'case closed.' If that is not the case, however, then the chain of successive generations never terminates, that is, the chain is infinitely long. But that, argued Ward, is nonsense—how could anything *infinitely* long have an end (our present *now*)?

Why Ward thought this an unanswerable paradox is hard to understand; after all, one can imagine a line in some coordinate system *beginning* at the origin and yet still being infinitely long (an example is the positive x-axis). This counter-example was not put forth by Hobbes in his own self-defense, but rather was offered by one of Ward's own colleagues at Oxford, John Wallis, the Savilian Professor of Geometry. As for Hobbes, he was little bothered by Ward's argument. As he pointed out (surely with a smile on his face), Ward was in danger of impaling himself as a theologian on his own sword: Ward's argument 'proved' the finite age not only of the world but of *everything*, including God (thus raising the awkward question of who, or what, made God?).

---

[29] The problem of 'squaring the circle' is, given a circle of area A, to construct (using only compass and straightedge) a square of area A.

Similar problems with infinity lay behind Kant's rejection of an infinite past. It is interesting to note that Kant, somewhat paradoxically, thought an infinite *future* a possibility. Why did Kant think time could be infinite in one direction but not in the other? One philosopher tells us[30] that Kant "failed to make himself clear," and I think that *understates* the case. I say that because Kant's argument was that the duration of the future is less problematic than is that of the past because it is only the past that influences the present. The best I can do in 'explaining' this is to speculate that if the present depends on an *infinite* past, then perhaps Kant thought that the possibility of so much influence was simply too much for the present to handle! In any case, Kant's view falls apart if we consider the possibility of backward time travel and the resulting implication that the future could also influence the present.

There is, as will come as no surprise, a philosopher for every conceivable point of the compass, and so a paper by one on the logical possibility of an infinite past soon prompts a rebuttal by another.[31] In illustration of this, you'll recall the quote from Augustus De Morgan in the opening section of this book, concerning the philosophers of his times; De Morgan went on in his critique to amusingly summarize the metaphysics of those times as follows: "Here we go up, up, up,/And there we go down, down, down,/Here we go backwards and forwards/And there we go round, round, round."

So, with De Morgan's words in mind, here are a few more examples of how people have struggled with the issue of the past. One quite interesting, *scientific* twist on the duration of the past was pointed out before the exchange between Smith and Ells. In a paper[32] observing that although general relativity and its predicted spacetime singularity in the distant past may indeed allow for a finite past, that does not completely close the door to the possibility that the Big Bang was a continuation from a previous contraction phase of the universe, and so on, *ad infinitum*. (You'll recall the discussion in Chap. 1 of this idea in science fiction: see note 53 in that chapter.) To quote T. S. Eliot (from his "Little Gidding"):

---

[30] J. Bennett, "The Age and the Size of the World," *Synthese*, August 1971, pp. 127–146. See also Q. Smith, "Kant and the Beginning of Time," *New Scholasticism*, Summer 1985, pp. 339–346.

[31] See, for example, Q. Smith, "Infinity and the Past," *Philosophy of Science*, March 1987, pp. 63–75, and then read E. Ells, "Quentin Smith on the Infinity of the Past," *Philosophy of Science*, March 1988, pp. 453–455. Smith's paper "The Uncaused Beginning of the Universe" appeared in this same issue (pp. 39–57), stating that he believed, *really*, only in the *logical* possibility of an infinite past and that the universe had in fact originated in an uncaused (no God required) Big Bang singularity. And, indeed, he *had* so argued for a finite past, in "On the Beginning of Time," *Nous*, December 1985, pp. 579–584.

[32] R. Weingard, "General Relativity and the Length of the Past," *British Journal for the Philosophy of Science*, June 1979, pp. 170–172.

"What we call the beginning is often the end
And to make an end is to make a beginning.
The end is where we start from."

Even without entertaining such an oscillating, accordion-like universe that endlessly expands and shrinks, it is possible to have a universe that originated in a *single* Big Bang a finite time ago in the past but yet *has no first instant*! This astonishing statement shocks most at first encounter, but it is simply the cosmological version of a well-known mathematical result. The instant $t = 0$ is not actually part of spacetime, because the Big Bang was quite literally a singular event for which the laws of spacetime physics fail. Thus, all instants in time are greater than zero—and there is no smallest number greater than zero. If you name a positive number, no matter how small, I can name a positive number still smaller, such as one-half of yours. (Of course, if there really is merit to the idea of a quantum of time, the chronon, this argument goes out the window.)

In an ingenious observation that seems to have been missed by most philosophers, E. A. Milne, a professor of mathematics at Oxford, suggested in his 1948 book *Kinematic Relativity*, that with general relativity it is conceivable to have both a single Big Bang a finite time ago *and* an infinite past. Pointing out that to talk meaningfully of time implies that we have a clock to measure it by, Milne looked for a Universal Clock that would be far more durable than our heartbeats, or anything else that exists only transiently. He suggested the expansion rate of the universe itself as the ideal clock. As we go back in time to the Big Bang, the expansion rate rises towards infinity and, as another analyst put it, "We see the Universe ticking away quite actively. *The Universe is meaningfully infinitely old because infinitely many things have happened since the beginning*."[33]

The debate over the length of the past in modern times can be just as contentious as it was in medieval times. For example, in his editorial ("Down with the Big Bang") of August 10, 1989, the then editor of *Nature* (John Maddox) declared the standard explosive model of the universe to be "philosophically unacceptable," because "the implication is that there was one instant at which time literally began and so, by extension, an instant before which there was no time." For Maddox, this meant that the Big Bang "is an *effect* [my emphasis] whose *cause* [my emphasis] cannot be identified or even discussed." The usual (non-time travel) use of the words *cause* and *effect* is that the cause happens first and then the effect occurs—but if the Big Bang (the effect) is the origin of time, then how (asked Maddox) could there be a cause of the Big Bang *before* that beginning?[34]

---

[33] C. W. Misner, "Absolute Zero of Time," *Physical Review*, October 1969, pp. 1328–1333. In this view cosmic time is taken as proportional to the negative of the logarithm of the normalized volume of the universe ($V = 1$ represents maximum volume, and so time 'stops' at the end of the universe's expansion). Thus, because V goes to zero as we go backward in time, time runs ever faster as we travel ever further into the past. This puts the Big Bang (with $V = 0$) infinitely long ago.

[34] This was not a new insight, of course, as Aristotle had long ago (in his *Physics*) declared an instant in time with no predecessor to be an absurdity.

The answer is obvious *for creationists*, of course—God did it. Creationists avoid the question of God's cause, however, saying only that 'He needs no cause,' or even that 'He made Himself'! It is these standard (ridiculous) responses from creationists that Maddox said had prompted his editorial against the Big Bang, because creationists *embrace* the Big Bang as it seems to endorse their position of 'science by imagination.' Whatever the truth of that, I think juxtapositioning the *scientific* Big Bang model of the universe with theological metaphysics and the pseudo-science nonsense of creationism to be terribly unfair.

When will the philosophical debates on the age of the past end? Not until the end of the (infinite?) future, is my wager!

2.3 Cause and Effect

> "There are few paradoxes which have been resolved so often as the time-asymmetry paradox."[35]

The philosophical literature is full of discussions about potential causal relationships between events. One of the most famous of these discussions, illustrating that cause and effect can be pretty slippery concepts, asks what at first appears to be an almost trivial question: Did the death of Socrates cause the widowhood of Xanthippe? The quick and easy answer is "Of *course*—she was his wife and it was his death that causes us to say she was then a widow. What could be more obvious?" One philosopher has provided some interesting commentary, however, that might make you reconsider, or to at least become aware of how different are the questions concerning time that are of interest to physicists and philosophers.[36]

Suppose we agree that there are two events to be considered; Socrates ceasing to live, and Xanthippe becoming a widow. Those events occurred at different places (in prison, and wherever Xanthippe happened to be). Then, as Kim asserted, "the two events occur with absolute simultaneity . . . [and so] we would have to accept this case as one in which causal action is propagated instantaneously through spatial space." (As we'll discuss in Chap. 3, the *relativity* of distant simultaneity weakens this assertion, but we'll take that up later.) For now, it is the conclusion that Kim draws from the assertion that interests us here: just *what* is propagating instantly? If it isn't mass-energy (as 'widowhood' would appear not to be!) then special relativity isn't bothered and physicists are happy. But those same physicists might also scratch their heads over *why* philosophers even wonder about such a question, because isn't becoming a widow just another way of saying that Socrates died and so we really don't have *two* events, but just one? In other words, for physicists this really isn't a question about cause and effect at all!

---

[35] J. Hurley, "The Time-Asymmetry Paradox," *American Journal of Physics*, January 1986, pp. 25–28.

[36] J. Kim, "Noncausal Connections," *Nous*, March 1974, pp. 41–52.

The central puzzle of time travel to the past is its apparent denial of causality—that is, its denial of the belief that we live in a world where every effect has a cause and that the cause happens first. *First* we flip the switch and *then* the kitchen light comes on. It is *never* the other way around. So deeply embedded is the temporal ordering of cause and effect in our feelings about how the world—and all the rest of the cosmos—works, that the Australian philosopher John Mackie (1917–1981) called causation the "cement of the universe" (and used that wonderful phrase as the title of a 1980 book). Without causality, said Mackie, everything would come unglued and fall apart. For example, when electrical engineers design an electronic system that they intend to actually construct (as opposed to doing a mere theoretical 'paper design') they insist that the design be a *causal* one. By that they mean the system must have no output before an input is applied. That is, the system must not be able to anticipate (foresee) the application of an input. To put it bluntly, our engineers are insisting that they are *not* building a time machine!

Now all that might seem to be self-evident, but there *are* some subtle problems. For example, it has become almost a cliché to say that nothing can go faster than light; that's what physicists mean by *relativistic causality*. In other words, no cause can produce an effect at a distant location sooner than the time lapse required for a light pulse to make the trip. Classical mechanics, however, the science of Newton's laws that engineers use all the time, is *not* relativistically causal. Push the left end of a rigid rod, for example, and the right end moves *instantly*. Most of the time the lack of this form of causality causes no problems, but the fact remains that the mechanics all engineers (and physicists, too!) learn first in school is flawed on a fundamental level. A rigid rod is an impossibility in Einstein's mechanics.

Indeed, it is interesting to speculate about how, after a discussion of causality, a traditional engineering professor would respond if challenged on this issue by a bright student. Causality might not look so obvious, after all, if such a student stuck up her hand in class and said "Professor, you've told us that everything that happens in nature is due to a cause. That what we see happening all around us, as the world unfolds, is the domino-process of cause-effect-cause-effect, and so on, into the future. But suppose, Professor, that at some instant, somehow, every particle in the world suddenly reversed its velocity vector. Wouldn't that mean, given the time-reversible nature of the classical equations of motion, the world would then run backward in time along the same path it had followed up until the instant of reversal? Wouldn't that mean what was effect is now cause, and that what was cause is now effect? And if cause and effect can change roles like that ... well, Professor, just what do our words *mean*?"

An amusing, and instructive, cartoon illustration of the student's idea of reversing all the velocity vectors in a system appeared on the cover of the November 1953 issue of *Physics Today*. That issue contains an article on the 1949 nuclear magnetic resonance experiments performed by the American physicist Erwin Hahn, which in a certain sense dealt with just such reversed systems. In that illustration a group of runners on a multi-lane circular race track begin at the starting line in a coherent state, that is, all lined up together. Then, as they run around the track at various speeds, they gradually spread out into what appears to be an incoherent state.

But that incoherence is an illusion because if, at some instant (signaled in the cartoon by a pistol shot), they all turn around and run in reverse, they will all arrive back at the starting line *together, at the same instant*. The initial coherence of the runners was actually never lost, despite the superficial appearance of disorder, and the coherent state can be recovered at any time by a reversal of velocity vectors.

This isn't mere theoretical speculation, as an almost magical application of velocity vector reversal is actually used in what is called *optical phase conjugation*, a process to 'time-reverse' the severe distortion suffered by light beams during atmospheric propagation. For example, by effectively reversing the velocity vectors of photons, one can remove the turbulence blurring in satellite pictures of the Earth's surface as seen from space.[37]

Let me immediately short-circuit one possible answer our beleaguered professor might give in desperation, a response based on the fact that equations of physics are *not* all time reversible. Indeed, it was discovered decades ago that, in certain very rare, fundamental particle decay processes involving neutral K-mesons, there is the hint that perhaps nature *can* indeed distinguish between the past and the future. In particular, K-mesons should violate what is called *CP-symmetry*, and the so-called TCP theorem[38] says that then *T-symmetry* must also fail. In 1968/69 direct, experimental observation of the failure of T-symmetry in K-meson decays was reported. In an astonishing example of science fiction prescience, the use of K-mesons in a machine for affecting the past had appeared years earlier in a 1955 (!) story.[39]

So, could K-mesons account for the physical processes that we see evolve in time in one direction (past to future) but not in the other? As Hurley (note 35) put it so nicely, "The decay of the neutral K-meson is not time-reversal invariant; perhaps it is this ubiquitous meson which is responsible for the cream diffusing uniformly throughout our coffee in the morning. Possibly, but again this conjecture cannot account for the computer models [of diffusion processes that, like cream in coffee, also display a bias for one temporal direction over the other—in Chap. 3 I'll show you such a computer model] which have no neutral K-mesons." Still, the tiny chink that K-mesons appear to have made in the once-solid rock of time direction indistinguishability is an active area of research and speculation.

Even with that chink the fact that the classical laws appear to be insensitive to a direction of time, whereas the real world—which seems in no way dependent on the arcane properties of K-mesons—seems distinctly asymmetric, is a puzzle of the first rank. As one philosopher wrote, "The Universe seems asymmetric with respect to the past and future in a very deep and non-accidental way, and yet all the laws of nature are purely time symmetric. So where can the asymmetry come from?"[40]

---

[37] C. R. Giuliano, "Applications of Optical Phase Conjugation," *Physics Today*, April 1981, pp. 27–35.

[38] The TCP-theorem says that the 'mirror-image' of a physical process is a legitimate process, too, *if* the 'mirror' reverses time (T), electric charge (C)—so that particle and anti-particle are interchanged, and parity (P)—which is the measure of left and right. There is strong reason to believe in the validity of the TCP theorem because quantum field theory is compatible with special relativity only if the TCP theorem holds.

[39] F. Pohl, "Target One," *Galaxy Science Fiction*, April 1955.

There have of course been attempts to answer that question. For example, one philosopher[41] discusses some curious mathematical examples he interprets as meaning, in the context of classical mechanics, that there are physical systems that are temporally irreversible *in principle*. A reply[42] from a fellow philosopher, however, argues that Hutchinson has, at most, shown only that classical mechanics is perhaps not deterministic. And that, Savitt argues, is not equivalent to showing a failure of time reversibility. There is, in fact, powerful experimental evidence that, with the rare exceptions of K-mesons, the classical laws of physics (including general relativity and quantum mechanics) *are* time-reversible.

Perhaps the most compelling of such evidence comes from the *reciprocity theorem* that electrical engineers routinely use when designing radio antennas. The theorem is easy to illustrate. Suppose two electrical engineers, Bob in Boston and Lois in Los Angeles, send radio signals to each other. Bob sends his messages by exciting his antenna with a time-varying current, which thus launches electromagnetic radiation into space. Lois' distant antenna intercepts some of that radiation, which then creates a (very tiny) signal current in her antenna.

The reciprocity theorem states the following: Suppose Bob makes a tape recording of his excitation signal and mails it to Lois, who then plays Bob's tape back into her transmitter as the excitation to *her* antenna. Then the signal current induced in Bob's antenna, as it intercepts Lois' launched radiation, will be the very same (very tiny) signal that Lois measured in her antenna as a result of Bob's transmission. This result is completely independent of the details of the two antennas, which can be utterly different in design, as well as independent of the details of the propagation path between Boston and Los Angeles (as long as those details don't change with time). The reciprocity theorem *is* true—it can be *measured* to be true as accurately as one wishes to perform this experiment—because of the reversibility of physics right down to the electronic level. In fact, the answer to the professor's problem of explaining why we don't see velocity vectors suddenly reverse, and then everything 'run backwards,' has not yet been found in any law of physics.

Now, to make things even more interesting, consider the problem of *mutual* or *simultaneous* causation, which can quickly lead to several interesting questions. When two leaning dominoes, A and B, hold each other up, is A nearly upright because of B, or is it B that is nearly upright because of A? When two children bob up and down on a see-saw, whose motion is the cause and whose is the effect? There are other puzzles, too, that involve mutual causation.

---

[40] J. Earman, "The Anisotropy of Time," *Australasian Journal of Philosophy*, December 1969, pp. 273–295.

[41] K. Hutchinson, "Is Classical Mechanics Really Time-Reversible and Deterministic?" *British Journal for the Philosophy of Science*, June 1993, pp. 307–323.

[42] S. F. Savitt, "Is Classical Mechanics Time-Reversal Invariant?" *British Journal for the Philosophy of Science*, September 1994, pp. 907–913.

For example, causation is usually thought to be transitive: if A causes B, and if B causes C, then A causes C. But if A and B are mutually causative, then ‘A causes B’ coupled with ‘B causes A’ leads to ‘A causes A’ (and to ‘B causes B’). That is, mutual causation, together with transitivity, seems to imply *self*-causation! Except for those theologians who like this sort of result (it lets them answer the question ‘Who made God?’ with ‘He made Himself’), hardly anyone likes self-causation. But how do we avoid the conclusion that perhaps the mutual causation of two leaning dominoes, coupled with transitivity, represents experimental proof that God could have made himself? Well, of course this is certainly outrageous stuff, but don’t you wonder how our poor professor would respond if asked?

This last example is actually a far more esoteric one than we need to illustrate how our ordinary, everyday concept of cause and effect can be turned inside out by going only a little bit beyond the routine. Consider, for example, the problem of the data processing of recorded time signals, such as the information written onto magnetic tapes, hard drives, or disks. Typical applications that produce such recordings include the strata-probing seismic echoes from dynamite explosions set by oil exploration geologists; arms control compliance monitoring stations that listen for the acoustic rumbles generated by both earthquakes and underground nuclear tests—and then try to tell one from the other; and the gathering by various military intelligence agencies of turbine shaft/propeller noise signatures emitted by different types of submarines. In each of those situations, the raw information is recorded and then later processed with a certain degree of unhurried calm and leisure. That pool of oil, after all, has been underground for several hundred million years, and waiting a few more days or weeks for a computer analysis of the explosion echo isn’t going to make much difference.

Such after-the-fact processing of recorded data is said to be done ‘off-line, in non-real time.’ When we play a disk back in the lab, however, we can do all sorts of neat things, like speed up the playback (make time ‘run fast’), or slow it down (make time ‘run slow’), or even play it *backwards* (make time ‘run in reverse’). For various technical reasons, generically called *spectrum shifting*, such tricks are often quite useful. Now, the way we retrieve magnetically recorded information from (for example) a magnetic tape, is to run it through a playback machine with a ‘read-head’ that senses the magnetic flux variations. The electrical signal produced by the read-head is just like the original signal and, in fact, we can pretend we don’t know it is really coming off a tape, but rather that it *is* the original signal. For high-quality digitally recorded tapes and disks, in fact, it is virtually impossible to distinguish the original from a playback.

Now, suppose we construct our playback machine with *two* read-heads, with the new head sensing the recording slightly *before* the old head does. The two heads produce the same electric signal, of course, but the signal from the new head is *ahead* in time compared to the signal from the old head. The new head is, in a certain sense, ‘seeing the future’ of the old head! We can use these two signals, the old head representing ‘now’ time and the new head representing ‘future’ time, to build real systems that are *not* causal. The causality violation occurs in non-real time, of course, not *our* time, but no matter; some absolutely astonishing signal processing can be achieved this way. The universe is about fifteen billion years old, and pretending that time has shifted a few milliseconds or so doesn't seem to be too much violence to reality.

Two heads are often used on radio call-in talk shows to catch inappropriate remarks from intemperate callers and prevent them from being broadcast. A short time delay is introduced by first recording remarks 'live' on tape with a write head and, then a few seconds 'up-stream,' a read head regenerates the remarks for broadcast. A 5 s delay is generally sufficient, so what is heard on a radio receiver *now* actually occurred 5 s ago in the *past*. A caller can get terribly confused if she doesn't turn her own receiver off, because one ear hears the present on the telephone while the other ear listens to the past over the radio.[43] The 1956 British film *Timeslip* incorporates a similar situation, with an atomic scientist's perception advanced 7 s into the future as the result of an accidental radiation exposure. His resulting confusion and disorientation is the center of the film.[44]

2.4 Backward Causation

> "Causation as a topic of philosophical discussion refuses to die. Each year, books and articles on causation continue to pour forth. Of course, all this activity may simply be a symptom of the necrophilia that infests so much of philosophy."[45]

All of the previous discussion has fueled countless arguments about what is called *backward*, *reverse*, or even *retro* causation. What is generally meant by forward causation is, of course, that any event that occurs at time t is caused by events that all occurred at some earlier time(s). Backward causation says that at least one of the causing events occurs after time t—this should make it clear that backward causation is a close relative of time travel. Indeed, one philosopher uses the terms *time traveler* and *retro-causal engineer* interchangeably.[46] The topic, understandably, is at the root of many hot philosophical debates, though not everybody (as this section's opening quote makes clear) thinks those debates are illuminating.

Just *why* does Professor Earman take his harsh position? He offers, as one reason, his disdain for the common philosophical 'proof' of the impossibility of backward causation: By definition, a cause is always before its effect. Yes, that's the entire 'proof.' One can, of course, win *any* argument by *defining* the answer to be what it is you wish to believe. More interesting, and certainly more pertinent to time travel, is the argument that if backward causation were possible then one could change the past—but that cannot be done because the past is dead and gone and thus unchangeable. That does seem to be a pretty solid argument against backward causation,[47] but Earman rebuts it by pointing out that the very same logic could be applied to the future, and so the usual, uncontested forward causation would also be denied. That is, one could argue that whatever the future will be, *will be* (literally 'by definition'), so one cannot change the future. A similar argument was presented even earlier,[48] in which we find "suppose that someone says 'I can change the future. I can do *this* or I can do *that*.' Well, then, suppose that he does *that*. Has he changed the future? No, because doing *that* was the future."

---

[43] A science fiction use of this idea is in B. W. Aldiss, "Man In His Time," *Science Fantasy*, April 1965, the story of an astronaut who returns from a trip to Mars and finds himself 3.3077 min ahead of everybody else.

[44] Science fiction had used a twist on this idea long before the film; see E. Binder, "The Man Who Saw Too Late," *Fantastic Adventures*, September 1939, a tale of what it might be like to have a 3 min *delay* in your vision.

[45] J. Earman, "Causation: A Matter of Life and Death," *Journal of Philosophy*, January 1976, pp. 5–25.

[46] B. Brown, "Defending Backwards Causation," *Canadian Journal of Philosophy*, December 1992, pp. 429–443.

The reversal of the 'usual' causal order of events by backward time travel has been a mainstay of science fiction almost from the start of the genre. Consider, for example, this tale.[49] A man on vacation by himself, without his wife along, meets a young lady—and they fall in love. The man loves his wife, too, though, and he realizes (as the young lady leaves him for the last time), never to return, that it is all for the best. But she really hasn't gone that far away from him, as the reader soon discovers. She is a time traveler from the future, and after leaving him she goes even further back in time, back an additional 20 years. She does this because she has learned that he met his wife 20 years ago, and so she goes back to be *that* woman! Thus, the usual causal order of the two events 'a long marriage' and the 'pre-marriage courtship' has been reversed (if we accept the fact that the man doesn't remember what his wife looked like when they married).

Actually, even our everyday uses of cause and effect are not nearly so straightforward as one might think, even when they are under far less stress than backward causation and time travel inflict. Consider, for example, the endless problems that are easy to imagine in the legal world. If a man falls off the roof of a ten-story building and is electrocuted as he plunges through power lines while still twenty feet above ground, was gravity or electricity the cause of death? Or was it both? As this example and others demonstrate,[50] one clearly does not have to discuss time travel to get into a serious argument about cause and effect. But with time travel, and the resultant backward causation, things can become even more perplexing. For example, we normally think it foolish to prepare, now, for an event that has already happened, but the prudent time traveler about to visit an ice age in the distant past would be wise to pack a fur coat before getting into his time machine!

---

[47] See, for example, D. H. Mellor, "Fixed Past, Unfixed Future," in *Michael Dummett: Contributions to Philosophy* (B. M. Taylor, editor), Martinus Nijhoff 1987.

[48] J. J. C. Smart, "A Review of *The Direction of Time*," *Philosophical Quarterly*, January 1958, pp. 72–77.

[49] R. F. Young, "The Dandelion Girl," *The Saturday Evening Post*, April 1, 1961.

[50] See also P. Mackie, "Causing, Delaying, and Hastening: Do Rains Cause Fires?" *Mind*, July 1992, pp. 483–500.

One philosopher provides, I think, a good start at explaining why so many other philosophers (and not just a few physicists) have adopted the 'common sense' position of rejecting backward causation. As he writes, "Part of the answer, no doubt, is a confusion between affecting and altering [the past—a distinction we'll discuss at length later in this book]. We cannot alter the past. But then we cannot alter the future either, although we can affect it. However, I take the common-sense rejection of backward causation to be, for the most part, quasi-empirical. It is based on a thought experiment. Think how you would set about affecting the past. By building a time-machine, perhaps? But how would you build one? We have no idea how to start. Yet, by contrast, we can work out how to affect the future ... we just move our bodies."[51] But, as he goes on to argue, if we accept that we can't *change* the past (which means there is no way we could actually observe backward causation), then there still exists the possibility that past events were as they were because of events in the future.

Are there actual phenomena that justify a belief in the possibility of effect before cause in real time (not just in tape recorder time)? The only example I know of, and a controversial one at that, is a theoretical result from a reformulation of electrodynamics by the great English physicist Paul Dirac (1902–1984). Classical theory models electric charges as point objects of zero size, which causes problems when one tries to calculate certain details, such as the total field energy of a single electron. The answer comes out as infinity. In an attempt to find more reasonable (that is, finite) answers to such questions, Dirac modified the zero size of a charge to one taking them to be extended objects (while retaining the validity of Maxwell's equations for electrodynamics right down to a point). To calculate how such extended objects will behave mechanically, however, one has to include what are called the *self-interaction* forces, such as the force one side of an electron exerts on the other side.

When it was all worked through, Dirac arrived at a third-order differential equation of motion, an equation that involves a force term proportional not to the usual first time derivative of the velocity (that is, to the acceleration), but rather to the second derivative.[52] This force is proportional to the first derivative of the acceleration, and is a quantity of direct interest mostly to the designers of automobile suspensions, who call it the *jerk*. There is no force in physics, at least not in Newtonian physics, that shows that sort of dependence, and there are some curious consequences. For example, in Dirac's theory an electron experiencing no external force can still continually accelerate, exhibiting what is called a 'runaway solution.'

Dirac showed how the runaway solution can be eliminated by picking a particular value for what up to then was an arbitrary constant of integration in the analysis, but that trick causes, in turn, a new problem called ‘pre-acceleration.’ That is, if an electron experiences an external disturbance (Dirac considered a passing pulse of electromagnetic radiation), then the electron will start to move *before* the pulse reaches it! Now that does seem to be a pretty clear example of backward causation. The time interval during which the pre-acceleration occurs is very short, on the order of the time it takes light to travel across the spatially extended electron (about 10^{-24} s), but no matter. The apparent crack in the door of causality may be slight, but it was enough to satisfy some philosophers seeking scientific support for backward causation.

---

[51] P. Forrest, "Backward Causation in Defense of Free Will," *Mind*, April 1985, pp. 210–217.

[52] P. A. M. Dirac, "Classical Theory of Radiating Electrons," *Proceedings of the Royal Society A*, August 1938, pp. 148–168.

Not everybody liked this, however. One physicist was clearly uneasy about it, calling pre-acceleration “unpleasant” acausal behavior.[53] On the other hand, one can find believers, too.[54] Others have argued that the whole business is simply a non-problem. One philosopher, in fact, raised a very interesting technical point, arguing that Dirac’s equation is non-Newtonian (remember the *jerk* force) and so we have no reason for coupling force and acceleration together as a cause-and-effect pair.[55] In Newtonian mechanics we do use that particular coupling, yet we do not think of force and velocity as a cause-and-effect pair because there is an integration operation involved in getting from to the other. Similarly, in Dirac’s theory we have an integration operation separating force and acceleration.

One curious aspect to the debate on pre-acceleration is that many commentators seem not to have paid much attention to what Dirac himself had to say about it. As a Nobel laureate, it hardly seems likely that he would let such a result pass unnoticed and, indeed, his paper contains the following physical explanation: “It would appear that we have a contradiction with elementary ideas of causality. The electron seems to know about the pulse before it arrives, and to get up an acceleration … The behavior of our electron can be interpreted in a natural way, however, if we suppose the electron to have a finite size. There is then no need for the pulse to reach the center of the electron before it starts to accelerate. It starts to accelerate … as soon as the pulse meets its outside. Mathematically, the electron has no sharp boundary.”

Two physicists suggested a fascinating connection between travel backward in time and Dirac’s relativistically correct, quantum mechanical description of an electron.[56] They showed that in flat, two-dimensional spacetime the assumption of time travel to the past leads in a natural way to Dirac’s equation. If, on the other hand, time travel only into the future is assumed, then additional assumptions are required to derive Dirac’s equation. This connection between Dirac’s equation and time travel to the past makes some philosophers and physicists nervous, but it didn't seem to bother Dirac. In fact, he went on in his paper to show how the pre-acceleration implies the possibility of building a device for sending a faster-than-light signal backward in time. Science fiction writers were, of course, quick to grasp that idea and such gadgets were dubbed "Dirac radios."[57]

---

[53] P. C. W. Davies, *The Physics of Time Asymmetry*, University of California Press 1977.

[54] J. Earman, “An Attempt to Add a Little Direction to ‘The Problem of the Direction of Time’,” *Philosophy of Science*, March 1974, pp. 15–47.

[55] A. Grunbaum, “Is Preacceleration of Particles in Dirac’s Electrodynamics a Case of Backward Causation? The Myth of Retrocausation in Classical Electrodynamics,” *Philosophy of Science*, June 1976, pp. 165–201.

[56] D. G. McKeon and G. N. Ord, “Time Reversal in Stochastic Processes and Dirac’s Equation,” *Physical Review Letters*, July 6, 1992, pp. 3–4.

One of the more perplexing aspects of backward causation is that it seems to allow for the possibility of *causal loops*, and for the breaking of such loops, a central feature in many of the very best time travel stories. For example, suppose there is a gadget such that if I push its control button *now*, then today's lecture notes will have appeared in the gadget's output tray *yesterday*. Indeed, yesterday I found today's notes there and, in fact, I am about to go to class to deliver that lecture. A mighty good one it is, too, so I think I think I'll send it back to yesterday in just a few minutes with the help of the gadget. But I haven't yet pushed the button. What if I now decide *not* to push the button? Why did the notes appear so I could use them today? Philosophers call this potential breaking of a causal loop a *bilking paradox*. Later in the book I'll discuss how such paradoxes have regularly appeared in the physics and philosophy literature since the 1940s.

By contrast, such paradoxes had been discussed in the science fiction magazines long before World War II. For example, in a letter to the editor at *Astounding Stories* (June 1932) a fan clearly stated his objection to time travel with the aid of a bilking paradox. He suggested the following experiment: Immediately publish an open offer to the inventor of time travel (who will be born, presumably, at some future date) to travel back to one week before the offer is published. But of course (argued the fan) we'd have a pretty problem if we then decided not to publish the offer after the inventor showed up! As that fan wrote, "Paradoxical? I'll say so, if time travel is possible." That fan didn't know about what seems to be a generic limitation on time machines, however: that one can't travel back to a date before the date of the time machine's creation. Thus, that fan's particular bilking paradox actually has no force.[58]

For another fictional example of a bilking paradox, consider the story[59] of time travelers who, just before they begin a trip into the future, see Earth invaded by Martians. At first the invaders are unbeatable, but then the defending military forces of Earth suddenly and mysteriously acquire a fantastically powerful new weapon. It isn't long before the time travelers realize where it came from—they themselves will go into the far future, obtain the weapon, and then return with it to what is now their own past (when the weapon first appeared). But then they wonder what might happen if they don't go, if instead they 'cheat time.' After all, they reason, why bother now to hunt for the weapon when the invasion has already been defeated? We are told that this potential bilking paradox is a "sinister conception, crawling evilly within their brains, like an unanswerable enigma."

---

[57] See, for example, J. Blish, "Beep," *Galaxy Science Fiction*, February 1954.

[58] A similar bilking paradox had actually appeared the year before in the 1931 novel *Many Dimensions* by the English writer Charles Williams (1886–1945), which reads like a suitable script for an Indiana Jones movie.

[59] E. Binder, "The Time Cheaters," *Thrilling Wonder Stories*, March 1940. There is an amusing reference in this tale to Orson Welles' famous radio-drama-hoax, from just 2 years earlier, of just such an alien invasion based on H. G. Wells' *War of the Worlds*.

Some philosophers, and practically all physicists, agree with that last assessment about bilking paradoxes, and so they believe there is simply nothing more to say. That is, bilking puzzles like the one in "The Time Cheaters" show that causal loops (and backward causation) must be impossible. Many feel this way about time loops, and backward causation, because (as is well known) time travel to the past can create all sorts of paradoxes. But such paradoxes are offensive only to human, culturally-biased intuitions on 'how things ought to work,' and not to the laws of physics which are indifferent to a reversal in the direction of time—which of course underlies what time travel is all about.

As the great American chemist G. N. Lewis expressed it, "Our common idea of time is notably unidirectional, *but this is largely due to the phenomena of consciousness and memory* [my emphasis]."[60] Lewis' words caught the eye of the editor at one science fiction magazine, who summed it up for his readers in a half-page essay that contained dramatic words hinting at backward causation: "A new theory of time . . . reveals the possibility that events now occurring are among the factors that decided Caesar nearly 2,000 years ago to cross the Rubicon."[61]

Lewis' willingness to accept causality violations is not a universally popular view today. For example, one physicist has written[62] that "It is fair to say that most conservative physicists have very serious reservations about the admissibility and reality of causality-violating processes. Causality violation (i.e., the existence of a 'time machine') is such an extreme violation of our understanding of the cosmos that it behooves us to be as conservative as possible about introducing such unpleasant effects into our models." He then goes on to declare closed timelike loops to be verboten because "the existence of closed timelike loops leads us to such unpleasant situations as meeting oneself 5 min ago." He sums up his philosophical position nicely with "any theory that is 'just a little bit causality violating' is 'just a little bit inconsistent.'"

Agreeing with this physicist is at least one philosopher who believes that the "association of causality with a particular temporal direction is not merely a matter of the way we speak of causes, but has a genuine basis in the way things happen" and that there is indeed an asymmetry with respect to past and future that is bound up with our concept of intentional action.[63] He then goes even further when he continues with the claim that being an agent of cause is not a necessary condition for seeing the asymmetry; being an observer is enough, as even an immobile yet intelligent tree (!) could detect the difference between past and future. (How he knows this about certain trees is left unexplained.)

---

[60] G. N. Lewis, "The Symmetry of Time in Physics," *Science*, June 6, 1930, pp. 569–577.

[61] Editorial essay, "Two-Way Time," *Astounding Stories*, September 1931.

[62] M. Visser, "Wormholes, Baby Universes, and Causality," *Physical Review D*, February 15, 1990, pp. 1116–1124.

[63] M. Dummett, "Bringing About the Past," *Philosophical Review*, July 1964, pp. 338–359.

The everyday views of causality that we have formed through our limited experiences when living in a world in which time travel is 'uncommon' may actually be incomplete. As the British philosopher Bertrand Russell (1872–1970) said with some humor long ago, in his 1912 Presidential Address ("On the Notion of Cause") to the Aristotelian Society, "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a by-gone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." And I do agree with his fellow philosopher who, decades later, declared "The concept of cause is powerless to solve the problems posed by the concept of time. The fundamental laws of physics present our most careful, best established and most sophisticated understanding of time. Notoriously, nothing in these laws endorses the idea of a *flow of time nor of the direction* [my emphasis: we'll return to both of these issues later in this chapter] which is basic to our conception of it. Nor are these laws causal (in the sense of singling out causes) even when they are deterministic. The concept of cause is not a fundamental one and cannot illuminate the darker corners in our understanding of the fundamental concept of time."[64]

2.5 The Fourth Dimension

> "We are facing an invasion of fourth dimensional creatures . . . We are being attacked by life which is one dimension above us in evolution. We are fighting, I tell you, a tribe of hellhounds out of the cosmos. They are unthinkably above us in the matter of intelligence. There is a chasm of knowledge between us so wide and deep that it staggers the imagination."[65]
>
> "Fourth dimension. Time factor. *You* know . . ."[66]

The idea of a fourth dimension to *space* has long been a staple of science fiction, but it has also long been viewed with suspicion. Indeed, many quite sophisticated scientists have thought it to be quite mysterious. For example, in his 1897 Presidential Address to the American Mathematical Society, the Canadian/American astronomer-mathematician Simon Newcomb (1835–1909) declared "The introduction of what is now very generally called hyperspace, especially space of more than three dimensions, into mathematics has proved a stumbling block to more than one able philosopher." Einstein stated Newcomb's view in blunter terms when he wrote "The non-mathematician is seized by a mysterious shuddering when he hears of 'four-dimensional' things, by a feeling not unlike that awakened by thoughts of the occult."[67]

---

[64] G. Nerlich, "How to Make Things Have Happened," *Canadian Journal of Philosophy*, March 1979, pp. 1–22.

[65] From "Hellhounds of the Cosmos," *Astounding Stories*, June 1932, by Clifford Simak (1904–1988). Simak went on to write a number of much better tales, but this passage lends credence to the editorial introduction to the 1957 anthology *Famous Science-Fiction Stories* (Random House) that declared so much in the early pulp science fiction was "science that was claptrap and fiction that was graceless."

[66] Uninformative 'explanation' given to a befuddled, inadvertent time traveler who emerges miles away and one hour backward in time after a wild ride through the fourth dimension in a gadget (constructed from a bicycle tire!) in the shape of a three-dimensional Möbius strip (see note 99 in Chap. 1). From the story by H. Nearing, Jr., "The Maladjusted Classroom," *The Magazine of Fantasy and Science Fiction*, June 1953.

To see just how right Einstein was with this observation, consider the reaction one Egyptian philosopher had (in 1929) to Einstein's own writings: "We have no doubt in our mind that nobody can understand it (the fourth dimension), including Einstein himself. The incomprehensibility of these assumptions [of general relativity] is due to their nature. They deal with the fourth dimension . . . and the reality of time and space. They can only be described by a mathematician's hypothesis or by religious faith."[68] This reaction is easy to understand—after all, anybody can 'see' that there are exactly three spatial dimensions, and that is that!

The 1901 novel *The Inheritors*, by the English writer Ford Madox Ford (1873–1939), like Simak's, is the tale of an insidious hyperspace invasion of our world. It illustrates Einstein's assertion about how many people react to the fourth dimension with an example from the time before the science fiction magazines. When the novel's narrator is bluntly told by an invader that she (the invader) is from the fourth dimension—an idea inspired by Ford's appreciation of how much success his acquaintance H. G. Wells had enjoyed with it—he recoils from that claim with the words "If you expect me to believe you inhabit a mathematical monstrosity, you are mistaken." And who can really blame that skeptical narrator? How can there be *four* spatial dimensions? No less an authority than Aristotle, writing in 350 B.C., had declared in his essay "On the Heavens" that "the three dimensions are all that there are."

Others were not so sure. In 1873, for example, we find an essay in *Nature* that refers to well-known mathematicians who even earlier had shown that they had an inner assurance of the reality of transcendental space (hyperspace).[69] The American philosopher Charles Sanders Peirce (1839–1914) was also an early advocate for the four-dimensionality of space. Just what he thought the nature of the fourth dimension to be is somewhat unclear, but the context of what he said suggests he took it to be spatial. He thought three-dimensional space to be "perverse" because of the existence of incongruous counterparts (such as left- and right-handed gloves), and this was apparently strong evidence for him that space could not be three-dimensional. Now, incongruous counterparts exist in all n-dimensional spaces, but Peirce preserved the special purity of the fourth dimension by suggesting that all physical objects, although capable of motion in the fourth direction, could themselves have no extent in that direction (remember, Peirce was a philosopher, not a physicist, and he offered no experimental support for any of this).[70]

---

[67] A. Einstein, *Relativity: the Special and General Theory*, Crown 1961, p. 33.

[68] From A. A. Ziadat, "Early Reception to Einstein's Relativity in the Arab Periodical Press," *Annals of Science*, January 1994, pp. 17–35.

[69] G. F. Rodwell, "On Space of Four Dimensions," *Nature*, May 1, 1873, pp. 8–9.

But is it *really* possible that there could be *four* spatial dimensions? We experience three independent directions, each lying at a right angle to the other two—but why just *three*, and not ten or fifteen? Indeed, in an 1888 talk to the Philosophical Society of Washington, Simon Newcomb dismissed the view that space must necessarily be three-dimensional as an "old metaphysical superstition." Yet, despite Newcomb's open-mindedness, it has been shown that in the framework of classical physics there are, in fact, several powerful reasons for why there must be *exactly three* spatial dimensions.

The beginning of a scientific explanation for the dimensionality of space appears in Kant, who believed the *three* dimensions of space and Newton's inverse-square law for gravity are intertwined (but he offered nothing beyond philosophical speculation). The origin of Kant's view is actually quite old, dating back to the ancient Greeks, who had already begun to suspect that there was something special about *three* dimensions, at least as far as geometry was concerned. They knew of the infinity of regular two-dimensional polygons, but that there were just five regular polyhedrons in three dimensions (the so-called *Platonic solids*). This early observation was trapped in mystical speculations, however, and it wasn't until the development of physics as a science that non-mystical discussions on the dimensionality of space began to appear.

Beginning with the work of Einstein's friend, the Austrian/Dutch physicist Paul Ehrenfest (1880–1933) in 1917, we can find the idea that the Poisson-Laplace equation, a second-order partial differential equation that describes the potential functions for both Newtonian gravity and electrostatics, does not allow for stable planetary or electronic orbits in *any* space with dimensionality greater than three. Further, the distortionless, reverberation-free propagation of both electromagnetic and sound waves is possible only in spaces of dimensions one and three. These conclusions have been shown to hold even when we go beyond nineteenth century physics into general relativity and quantum mechanics.[71]

Using a slightly different approach, a biological-topological argument for why space cannot have fewer than three dimensions exists. In all of our common experience, complex intelligent life is always found to occur as an aggregate of a vast number of elementary cells, interconnected via electrical nerve fibers. Each cell is connected to several others, *not all immediate neighbors*, by these fibers. If space had only one or two dimensions, then such highly interconnected nets of cells would be impossible because the overlapping nerve fibers would have to intersect, which would result in their mutually short-circuiting one another.

---

[70] R. R. Dipert, "Peirce's Theory of the Dimensionality of Physical Space," *Journal of the History of Philosophy*, January 1978, pp. 61–70.

[71] See, for example, I. M. Freeman, "Why Is Space Three-Dimensional?" *American Journal of Physics*, December 1969, pp. 1222–1224, and L. Gurevich and V. Mostepanenko, "On the Existence of Atoms in n-Dimensional Space," *Physics Letters A*, May 31, 1971, pp. 201–202.

It wasn't long before these views on the dimensionality of space found their way into science fiction. An early use of space as four-dimensional occurs in an awkward rewrite of Jules Verne's *Around the World in Eighty Days*, in which a professor and his crew fly into hyperspace and around the world and to the moon and back, in less than a day.[72] They do this with a plane equipped with a four-dimensional rudder! More interesting is the tragic story (originally published in 1926) of a math professor who learns how to move into hyperspace and back.[73] A colleague catches him at it and, once over his astonishment, asks what is behind it all. The professor replies, "My assumption is that the fourth dimension is just another dimension—no more different in kind from length, say, than length is from breadth and thickness, but perpendicular to all three. Now suppose that a being in two dimensions—a flat creature, like [a moving shadow on a surface]—were suddenly to grasp the concept of a third dimension and so step out of the [surface]. He might move only an inch, but he would vanish completely from the sight of the world."

The professor has similarly learned how to step out of 3-space and into 4-space but, when asked to explain *how*, all he can say is "How can I explain? It's just the *other* direction. It's *there*!" His colleague can't see it, but nonetheless is quick to grasp the practical implications: "This is power! Think of it! A step, and you are invisible! No prison cells can hold you, for there is a side to you on which they are as open as a wedding ring! No ring is secure from you: you can put your hand *round the corner* and draw out what you like. And, of course, if you looked back on the Universe you had left, you would see us in sections, open to you! You could place a stone or a tablet of poison right in the bowels of your enemies!"

What the professor's colleague is getting at involves a comparison with a prison in planar 2-space, which would merely be a circle around the captive. Knowledge of the third dimension would make it possible to escape, however, by simply moving along that new direction, over the circle, and then back into the plane. To a 2-space guard it would seem that the prisoner had suddenly vanished from view *inside* the circle and then just as suddenly materialized again *outside* the circle. Similarly, to escape from a 3-space prison, one would merely move along the fourth dimension, and in the same way one could remove the yolk from an egg without damaging the shell; indeed, one could remove the yolk directly from the chicken without damaging the chicken![74]

---

[72] B. Olsen, "Four Dimensional Transit," *Amazing Stories Quarterly*, Fall 1928.

[73] R. Hughes, "The Vanishing Man," reprinted in *The Mathematical Magpie* (C. Fadiman, editor), Simon and Schuster 1962.

[74] This astounding insight appeared in early pulp science fiction in, for example, M. J. Breuer, "The Appendix and the Spectacles," *Amazing Stories*, December 1928. The concept appeared even earlier in Bob Olsen, "The Four-Dimensional Roller-Press," *Amazing Stories*, June 1927, and then later in Olsen's "The Great Four Dimensional Robberies," *Amazing Stories*, May 1928 to rob locked safe deposit boxes, and "The Four Dimensional Escape," *Amazing Stories*, December 1933, in which a man sentenced to die by hanging at San Quentin Prison is rescued, while standing on the gallows' trap, by an inventor who pulls him through the fourth dimension.

In a later tale[75] we meet another professor who dramatically uses this very feature of the fourth dimension. His right hand has been modified through an accident to exist in four-dimensional hyperspace and so, to finance his research, he uses his ‘talent’ to become the perfect pickpocket, able to reach into any wallet no matter how well secured. He also can, indeed, reach right into the very bowels of his fellow man. And he *does*. When he demonstrates his hand to the policeman who has arrested him for being a thief, the astonished officer chokes on a lemon drop. Dr. Fuddles then, of course, does the right thing and removes the drop from the poor fellow with ease. There is one additional aspect to Dr. Fuddles’ hand, however, that the story missed. If he had turned his right hand over in the fourth dimension, then he would have had *two* left hands!

It was discovered in 1827 by Möbius (of the strip) that any three dimensional object can be converted into its mirror image by flipping it over through the fourth dimension. Thus, a left-handed glove can be made by pure geometry (no scissors, thread, or needle required) into a precise copy of its right-handed mate. If a living organism is so flipped, however, there may be a problem, as everything in the body would be reversed, including the optically active organic molecules discovered by Pasteur in 1848, which are involved in vital biological processes. These molecules, called *stereoisomers*, exist in two versions in nature (the left-handed and the right-handed versions, if you will), but our bodies have developed the ability to use only one version. To be flipped through the fourth dimension would make some reversed stereoisomers unable to participate in the digestion of food and we would starve to death.

For modern science fiction writers the fourth dimension (and hyperspace, in general), is still a major concept. One physicist, writing in *Analog* (today’s premier ‘hard science’ fiction magazine), summed up nicely what was so fascinating in early pulp, and still is today, about the idea of an extra dimension or two, or perhaps even more, at least from a fictional point of view: “Are there hidden dimensions not accessible to us, dimensions in which we could go adventuring, dimensions within which malevolent hyper-dimensional aliens may be lurking, ready to pierce our flimsy paper-thin three-space bodies with their terrible hyper-sharp claws?”[76] The early pulp science fiction magazines encouraged this lurid imagery. Witness the editorial blurb that opened one many-dimensional monster story as follows: “It was a strange world in which Lester and Florence found themselves. A world of sudden

---

[75] N. Bond, “Dr. Fuddle’s Fingers,” in *Mr. Mergenthwirker’s Lobblies and Other Fantastic Tales*, Coward-McCann 1946.

[76] J. Cramer, “The Other Forty Dimensions,” *Analog*, April 1985. ‘Monsters in hyperspace’ stories were numerous in pulp science fiction. Three examples (in no particular order of literary merit!) are: M. J. Breuer, “The Einstein See-Saw,” *Astounding Stories*, April 1932; P. Ernst, “The 32nd of May,” *Astounding Stories*, April 1935; “The Monster from Nowhere,” *Fantastic Adventures*, July 1939.

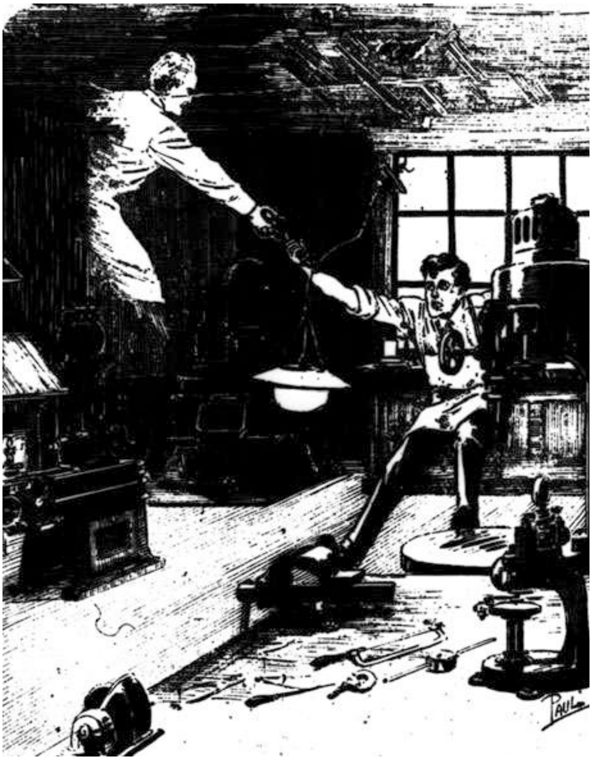

Fig. 2.1 An experiment in hyperspace goes astray. The young man is pulling on "hyper-forceps" in an attempt to retrieve a surgeon who has fallen out of 3-space (along with his patient, a professor of non-Euclidean geometry, who suffers from gallstones). The hyper-forceps allow the removal of the gallstones without cutting into the body. Illustration by Frank R. Paul, ©1928 by Experimenter Publishing Co. for "Four Dimensional Surgery" (*Amazing Stories*, February 1928) by Bob Olsen, reprinted by permission of the Ackerman Science Fiction Agency, 2495 Glendower Ave., Hollywood, CA 90027 for the Estate

death and strange science, ruled by inhuman beasts."[77] But as outrageous as that might sound, the real physics of hyperspace is even more amazing.

Hyperspace is, in general, simply any space with more dimensions than the one we obviously seem to live in. In particular, our universe appears to be a four-dimensional (three spatial and one temporal) hyperspace called *spacetime*. This four dimensional world can, at least mathematically, be thought of as the boundary surface of a five dimensional hyperspace. This is analogous to the way the two-dimensional space of the surface of a sphere bounds the three-dimensional space of the sphere itself. This interesting imagery appeared quite early in pulp science fiction. For example, in one remarkably sophisticated story, an eccentric scientist at one point exclaims "A mathematical physicist lives in vast spaces . . . where space unrolls along a fourth dimension on a surface distended from a fifth."[78]

There are some interesting geometrical implications to hyperspace which play big roles in time travel considerations. For example, for beings in the two-dimensional world of a sphere's surface there are *two* ways to travel from pole to pole. There is the usual way, *on* the surface of the sphere, and the hyperspace way which takes them *through* the sphere along the polar diameter. In imagery motivated by thinking of the sphere as an apple, and of the hyperspace path as a tunnel bored by a worm through the apple, it has become popular to call all such shortcuts, through any hyperspace of any dimension, *wormholes* (a word coined in the 1950s by the Princeton physicist-wordsmith John Wheeler). Wheeler used wormholes to show how electric charge could be thought of as lines of force trapped in the changing topology of a multiply connected space (indeed, Wheeler claimed that the observation of what we call electricity is experimental evidence that space is *not* simply connected).[79]

---

[77] M. Duclos, "Into Another Dimension," *Fantastic Adventures*, November 1939. See the illustration for this story in "Some First Words."

[78] M. J. Breuer, "The Gostak and the Doshes," *Amazing Stories*, March 1930.

The general theory of relativity predicts the existence of wormholes in spacetime and, in fact, they were first 'discovered' theoretically in the mathematics of relativity as early as 1916 by the Viennese physicist Ludwig Flamm (1885–1964). Later analyses were done by Einstein, himself.[80] Wormholes have been discussed as a possible model for pulsars (as opposed to the more usual model as rotating neutron stars).[81] It has also been suggested that the interior of a charged black hole may be the entrance to a wormhole.[82] All of these various solutions to the gravitational field equations are generically called "Einstein-Rosen bridges" in the physics literature (see note 81, for example), and the term soon appeared in fiction, too.[83]

The use of hyperspace wormhole portals for explaining some observed physical phenomenon appeared in the scientific literature long before Wheeler's electricity example. In his 1928 book *Astronomy and Cosmogony*, for example, the British theoretician Sir James Jeans devoted a chapter to what were then called nebulae, the island-universes we now call galaxies. At the end of his discussion on the arms of spiral galaxies, Jeans offered the following speculation: "Each failure to explain the spiral arms makes it more and more difficult to resist a suspicion that the spiral nebulae are the seats of types of forces entirely unknown to us, forces which may possibly express novel and unsuspected *metric properties of space* [my emphasis]. The type of conjecture which presents itself, somewhat insistently, is that the centers of the nebulae are of the nature of 'singular points,' at which matter is poured into our universe from some other, and entirely extraneous, special dimension, so that, to a denizen of our universe, they appear as points at which matter is being continually created." This, in everything but name, is a wormhole.

---

[79] A space is simply connected if *all* the points on the straight line that joins *any* two points in the space are also in the space. The interior of a sphere is simply connected. The interior of a sphere with a hole in it is *not* simply connected.

[80] A. Einstein, "The Particle Problem in the General Theory of Relativity," *Physical Review*, July 1, 1935, pp. 73–77.

[81] J. M. Cohen, "The Rotating Einstein-Rosen Bridge," in *Relativity and Gravitation* (C. G. Kuper and A. Peres, editors), Gordon and Breach Science Publishers 1971.

[82] A. Ori, "Inner Structure of a Charged Black Hole: An Exact Mass-Inflation Solution," *Physical Review Letters*, August 12, 1991, pp. 789–792.

[83] See, for example, J. G. Cramer, *Einstein's Bridge*, Avon 1997 (this is the same Cramer cited in note 76). The *Rosen* comes from the American-Israeli physicist Nathan Rosen (1909–1995), who was a collaborator of Einstein's.

What would hyperspace be like? It is intuitively obvious that in the case of the 2-D surface of a 3-space sphere, the 'hyperspace' wormhole path is shorter than the surface path. Even if this 'shorter path' view holds for wormholes in our 4-D spacetime, however, getting around in science fiction hyperspace may not be a simple task. One tale, for example, tells the story[84] of how one of the first spaceships to explore hyperspace gets lost. The trouble with hyperspace travel is that "You go in at one point, you rocket around until you think it's time to come out, and there you are. Where is 'there'? Why, that's the surprise that's in store for you, because you never know until you get there. And sometimes not even then." The same idea plays a central role in Robert Heinlein's 1957 novel *Tunnel in the Sky*, in which a 'hyperspace gate' is discovered by accident during failed time travel experiments.

Another story[85] asks the same question about hyperspace, and arrives at the same answer: "When you took the Jump … how sure were you *where* you would emerge? The timing and quantity of the energy input might be as tightly controlled as you liked … but the uncertainty principle reigned supreme and there was always the chance, even the inevitability of a random miss … a paper-thin miss might be a thousand light-years."

A common way to visualize hyperspace wormhole shortcuts is to imagine the beginning and the end of a journey as points A and B on the 2-D surface of a piece of paper. Then imagine that the paper is folded so as to position A over B, perhaps with A almost touching B. The distance from A to B *through hyperspace* (the 3-D space in which the folding took place) can clearly be much less than is the distance through 'normal' space (the distance covered by a trip that always remains in the 2-D surface). This is the specific example used in one tale to explain the instantaneous "space-warp" (wormhole) device invented by the story's hero.[86] Such imagery actually appeared quite early in science fiction, as in one story in which a gadget is used to "bend space" so that Earth and Venus touch![87]

The idea of hyperspace folding has broken free from science fiction and can now be found in modern stories in other genres. For example, in one Stephen King story ("Mrs. Todd's Shortcut") a woman keeps finding ever shorter ways to drive from Castle Rock, Maine to Bangor. As the crow flies it is 79 miles, but she gets the journey down to 67 miles, and later to 31.6 miles. When doubted, she replies: "Fold the map and see how many miles it is then … it can be a little less than a straight line if you fold it a little, or it can be a lot less if you fold it a lot." The doubter remains unconvinced: "You can fold a map on paper, but you can't fold *land*."

---

[84] F. Pohl, "The Mapmakers," *Galaxy Science Fiction*, July 1965.

[85] I. Asimov, "Take a Match," in *New Dimensions II: Eleven Original Science Fiction Stories* (R. Silverberg, editor), Doubleday 1972.

[86] G. O. Smith, "The Möbius Trail," *Thrilling Wonder Stories*, December 1948.

[87] E. L. Rementer, "The Space Bender," *Amazing Stories*, December 1928.

For the purpose of wormhole creation in spacetime, we actually have to imagine much more: the folding of four-dimensional spacetime through a five dimensional hyperspace. The folding imagery has even appeared in the movies: spacetime folding is demonstrated with a piece of paper in both *Event Horizon* (perhaps the worst movie of 1997) and the 2014 *Interstellar*.

Another feature of hyperspace that science fiction has taken a liking to is its vastness. An interesting fictional treatment of this idea was given by a writer who, in real life, was an academic psychologist at the University of Michigan. He put himself in a story[88] of a starship captain who is explaining to the crew psychologist how he feels about hyperspace (or *subspace*, as it is called in the story): "God forsaken. That's just what it is. Completely black, completely empty. It frightens me every time we make the jump through it . . . it frightens me because—well, because a man seems to get lost out there. In normal space there are always stars around, no matter how distant they may be, and you feel that you've got direction and location. In subspace, all you've got is nothing—and one hell of a lot of that. It's incredible when you stop to think about it. An area—an opening as big as the whole of our Universe, big enough to pack every galaxy we've ever seen in it—and not a single atom of matter in it . . . until we came barging in to use it as a shortcut across our own Universe."

The vastness of hyperspace got a more humorous treatment from the early pulp science fiction writer Bob Olsen (1884–1956), who wrote the following verses[89] in the introduction to one of his many stories of the fourth dimension:

I read a yarn the other day—
A crazy concept, I must say.
It states that objects have extension
In what is called the "Fourth Dimension."
In hyperspace one could, no doubt,
Make tennis balls turn inside out;
And from a nut remove the kernel
And not disturb the shell external.
A crook could pilfer bonds and stocks,
Then laugh at prison bars and locks,
One step in this direction queer,
And presto! He would disappear!
Let's hope, in planning new inventions,
They'll give us cars with four dimensions.
When searching for a parking place
We sure could use some hyperspace!

It is not just science fiction that takes hyperspace seriously. We find a mathematician, for example, writing that "most science fiction addicts are familiar with the notion of 'hyperspace,' a higher dimensional space-time bounded by Space-Time through which, in the far distant future, interstellar voyages shortcut the (otherwise unsurmountable) distances between the stars. The purpose of this article[90] is to demonstrate that any . . . relativistic space-time model is the boundary of some . . . five-dimensional hyperspace." That is just what Breuer's magazine character (see note 78) said—in 1930!

---

[88] J. V. McConnell, "Avoidance Situation," *If*, February 1956.

[89] B. Olsen, "The Four-Dimensional Auto-Parker," *Amazing Stories*, July 1934. "Bob Olsen" was the pen-name for Alfred Johannes Olsen.

The concept of *time* as a fourth dimension has long been a popular concept, and science fiction in particular has embraced it with enthusiasm. We find a little joke on the idea in a story where a young couple, visited by time travelers from 500 years in the future, are said to live in Apartment 4-D.[91] One physicist[92] traced the idea back to the late eighteenth century, finding references to the idea in pre-1800 works of the great French mathematical physicists Jean le Rond d'Alembert (1717–1783) and Joseph-Louis Lagrange (1736–1813). In fact, a philosopher[93] has found a 1751 passage written by d'Alembert that appears to indicate that it is some unknown, earlier person to whom the credit should really go: "I have said [that it is] not possible to imagine more than three dimensions. A clever acquaintance of mine believes, however, that duration could be regarded as a fourth dimension and that the product of time and solidity would be in some way a product of four dimensions; that idea can be contested, but it seems to me that it has some merit, if only that of novelty."

Still, it wasn't until a curious letter appeared in *Nature* in 1885 that the concept of time as the fourth dimension was mentioned seriously in an English-language scientific journal. The author, mysteriously signing himself only as "S.," began by asking "What is the fourth dimension? . . . I [propose] to consider Time as a fourth dimension . . . Since this fourth dimension cannot be introduced into space, as commonly understood, we require a new kind of space for its existence, which we may call time-space."[94] Who was this prophetic writer that, if he had just made a simple swap, would have been the first to use space-time as a word? Nobody knows. Bork speculates that it was an acquaintance of H. G. Wells, but Wells himself is on record that it certainly wasn't him.

In his 1934 *Experiment in Autobiography*, Wells wrote "In the universe in which my brain was living in 1879 there was no nonsense about time being space or anything of that sort. There were three dimensions, up and down, fore and aft and right and left, and I never heard of a fourth dimension until 1884 [when Wells was eighteen] or thereabout. Then I thought it was a witticism." He had, in fact, said this before. In a 1931 edition of *The Time Machine* (Random House), for example, he wrote in the Preface that the idea for the novel "was begotten in the writer's mind by students' discussions in the laboratories and debating society of the Royal College of Science in the eighties and already it had been tried over in various forms by him before he made this particular application of it."

---

[90] G. S. Whiston, "'Hyperspace' (The Cobordism Theory of Space-Time)," *International Journal of Theoretical Physics*, December 1974, pp. 285–288.

[91] L. Padgett, "When the Bough Breaks," *Astounding Science Fiction*, November 1944.

[92] A. M. Bork, "The Fourth Dimension in Nineteenth-Century Physics," *Isis*, October 1964, pp. 326–338.

[93] E. Meyerson, *The Relativistic Deduction*, volume 83 of *Boston Studies in the Philosophy of Science*, D. Reidel 1985, p. 78.

[94] S., "Four-Dimensional Space," *Nature*, March 26, 1885, p. 481. The editorial staff at *Nature* has informed me that, more than a century-and-a-quarter later, there is no longer any record of the identity of S. in the journal's archives.

The idea of time as the fourth dimension entered the popular mind around 1894–95, with the publication of the first of Wells' so-called "scientific romances," *The Time Machine*. Then, after that pioneering use of time as the fourth dimension, science fiction quickly adopted the idea as the basis for one of its most popular subgenres. One of the great "golden age of science fiction" writers, 'Murray Leinster' (1896–1975)—the pen-name for William Jenkins—used it as the basis for his first published story.[95] It is the incredible tale of a Manhattan skyscraper (and its 2000 occupants) sent backward in time several 1000 years because its foundation slips (in an unexplained way) along the fourth dimension. The scientific sophistication of the story is primitive, with just one of the many logical flaws being a vivid description of the time travelers living forward-in-time even as their wrist watches run backward. Indeed, when pulp pioneering editor Hugo Gernsback reprinted the tale in one of the early issues of *Amazing Stories*, a reader complained about that very point. Gernsback felt compelled to defend the story, but could muster only a weak rebuttal based on an author's right to "poetic license."[96]

More technical is the discussion in the story of a clerk who transforms the main entrance to a department store into a time machine by building a tesseract (a four-dimensional cube).[97] The claim is made there that the fourth dimension of the cube/doorway is time. That tale appeared just 5 months after a classic of science fiction by Robert Heinlein (1907–1988) had appeared, also using a tesseract, in which the fourth dimension is taken as *spatial*.[98]

Some writers wanted to have the fourth dimension both ways, as space *and* time in the same story. One wonderful example of this is a classic,[99] written by one of the giants of science fiction. In that tale an electrical engineer named Nelson is caught in the middle of an enormous electromagnetic field surge produced by a short circuit in a power plant. As a physicist explains to the shocked board of directors of the utility, "It now appears that the unheard-of-current, amounting to millions of amperes . . . must have produced a certain extension into four dimensions . . . I have been making some calculations and have been able to satisfy myself that a

---

[95] M. Leinster, "The Runaway Skyscraper," *Argosy*, February 1919.

[96] H. Gernsback, "Plausability in Scientifiction," *Amazing Stories*, November 1926.

[97] W. P. McGivern, "Doorway of Vanishing Men," *Fantastic Adventures*, July 1941.

[98] R. Heinlein, "—And He Built a Crooked House," *Astounding Science Fiction*, February 1941. Here we read of a Los Angeles architect who builds a house in the shape of a tesseract as it would appear if collapsed into normal three-dimensional space. It isn't stable in 3-space (we are told), however, and so a California earthquake is sufficient to topple the house into a stable 4-D configuration, along with its occupants.

[99] A. C. Clarke, "Technical Error," *Fantasy No. 1*, December 1946.

'hyperspace' about ten feet on a side was, in fact, generated: a matter of some ten thousand quartic—not cubic!—feet. Nelson was occupying that space. The sudden collapse of the field [when the overload breakers finally broke the circuit] caused the rotation of that space."

Being rotated through 4-space has inverted the unlucky Nelson [see *For Further Discussion* at the end of this chapter for more on this point], and to bring him back to normal he must be flipped again. The physicist brushes aside a question about the fourth dimension as time, asserting that the only issue is one of space. Poor Nelson is, therefore, again subjected to a stupendous power overload—only now he disappears! Too late, the physicist realizes that the fourth dimension is both space *and* time and that Nelson has been spatially flipped *and* temporally displaced into the future. To understand the particularly monstrous fate of Nelson, just ask yourself what the result would be if he should materialize *inside* matter sometime in the future!

The interpretation of the fourth dimension as time is, of course, the one of interest to prospective time travelers, to physicists studying time travel, and to philosophers of time, and so for us, too. The sort of science fiction that is of greatest interest to us is like the one in which one of the characters, displaced in time, asks for an explanation from a higher-dimensional being who appears on the scene: "'Just where is Tuesday?' he asked. 'Over there [and when the being extends its hand, the hand disappears].' 'Do that again.' 'What? Oh—Point toward Tuesday? Certainly.'" The being explains the physics of the situation to the astonished time traveler thus: "It is a direction like any other direction. You know yourself there are four directions—forward, sideward, upward, and—*that* way! . . . It is the fourth dimension—it is duration."[100]

And how about stories like the one in which a mad inventor discovers how to make a substance whose atoms resist being pushed by "pushing back at right angles to all the other [spatial] directions." That is, to push on this exotic stuff is to risk experiencing a back reaction, of being pushed "off into the fourth dimension [which we are told is time] . . . into the middle of the week after next."[101] Now wouldn't *that* really be something?!

But of course it was H. G. Wells who, in fiction, pioneered time travel and its connection to the fourth dimension as it is popularly thought of today (with the caveats about Wellsian time machines kept firmly in mind). We are therefore quite interested, as *The Time Machine* opens, to listening-in as the Time Traveller expounds to a group of friends at a dinner party in his London home. He starts with the assertion "There is no difference between Time and any of the three dimensions of Space except that our consciousness moves along it." When asked to say more about the fourth dimension, he replies, "It is simply this. That Space, as our mathematicians have it, is spoken of as having three dimensions, which one may call Length, Breadth, and Thickness, and it is always definable by reference to three planes, each at right angles to the others. But some philosophical people have been asking why *three* dimensions particularly—why not another direction at right angles to the other three?—and have even tried to construct a Four-Dimensional geometry. Professor Simon Newcomb was expounding this to the New York Mathematical Society only a month or so ago."[102]

---

[100] T. Sturgeon, "Yesterday Was Monday," *Unknown Fantasy Fiction*, June 1941.

[101] M. Leinster, "The Middle of the Week After Next," *Thrilling Wonder Stories*, August 1952.

2.6 Spacetime and the Block Universe

> "And now he has preceded me briefly in bidding farewell to this strange world. This signifies nothing. For us believing physicists, the distinction between past, present, and future is only an illusion, even if a stubborn one."
>
> —Albert Einstein[103]

The poet Henry Van Dyke wrote, in his 1904 "The Sun-Dial at Wells College," words that echo the spirit of Omar Khayyam's *Rubaiyat* from nine centuries before:

The shadow by my finger cast
Divides the future from the past:
Before it, sleeps the unborn hour,
In darkness, and beyond thy power:
Behind its unreturning line,
The vanished hour, no longer thine:
One hour alone is in thy hands,—
The NOW on which the shadow stands.

The very next year Einstein's theory of special relativity appeared and, 3 years later, came Minkowski's spacetime interpretation of special relativity. Van Dyke's beautiful poetry was dealt a mighty blow by those developments in mathematical physics, and in the rest of this chapter we'll see how that came to pass.

The modern view of reality, that the past, present, and future are joined together into a four-dimensional entity called *spacetime*, is due to Hermann Minkowski (1864–1909), Einstein's mathematics teacher when he was a student in Zurich. Minkowski gave spacetime (the visual imagery of Einstein's mathematics) to the world during a famous address at the 80th Assembly of German Natural Scientists and Physicians in Cologne, on September 21, 1908. Entitled "Space and Time," his remarks were electrifying then and still are today.[104] He began dramatically: "Gentlemen! The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical." Then came the famous line, quoted in so many freshman physics texts and philosophy papers, concerning the nature of spacetime: "Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve independence." Minkowski explained what spacetime is in these words to his audience:

---

[102] And so Newcomb actually was. Wells, it is certain, routinely read *Nature* (one of his college friends, Richard Gregory, eventually became the journal's editor), and Wells must have read Newcomb's address of December 28, 1893 to the New York Mathematical Society when reprinted in the February 1, 1893 issue (on pp. 325–329), where he called hyperspace "the fairyland of geometry." From the Time Traveller's own words, then, that wonderful Victorian dinner party must have taken place in January or February of 1894.

[103] From a letter written by Einstein on March 21, 1955, to the children of Michele Besso, his dearest friend, who had just died. Einstein's use of the word *briefly* was due to his knowledge that he was nearly out of time, too (he died just a month later).

> "A point of space at a point of time … I will call a *world point*. The multiplicity of all thinkable x, y, z, t systems of values we will christen the *world*. With this most valiant piece of chalk I might project upon the blackboard four world axes … Not to leave a yawning void anywhere, we will imagine that everywhere and everywhen there is something perceptible. To avoid saying 'matter' or 'electricity' I will use for this something the word 'substance.' We fix our attention on the substantial point which is at the world point x, y, z, t, and imagine that we are able to recognize this substantial point at any other time. Let the variations dx, dy, dz, of the space coordinates of this substantial point correspond to the time element dt. Then we obtain, as an image, so to speak, of the everlasting career of the substantial point, a curve in the world, a *world-line*. … The whole Universe is seen to resolve itself into similar world-lines, and I would fain anticipate myself by saying that in my opinion physical laws might find their most perfect expressions as relations between these world-lines … *Thus also three-dimensional geometry becomes a chapter in four-dimensional physics* [my emphasis]."

With those words Minkowski gave mathematical expression to the philosophical exposition of Wells' Time Traveller to his dinner party friends. Taking the Minkowskian view of the primacy of spacetime as the ultimate building block stuff of reality was Princeton professor of physics John Wheeler, who wrote[105] "There is nothing in the world except empty curved space. Matter, charge , electromagnetism … are only manifestations of the bending of space. *Physics is Geometry*." This idea was echoed in fiction, in the 1987 novel Moscow 2042 by Vladimir Voinovich, where we find a time traveler who declares "Anyone with even a nodding acquaintance with the theory of relativity knows that nothing is a variety of something and so you can always make a little something out of nothing."

But not everybody understood Minkowski. In a little-known yet quite erudite essay, published just after a stunning experimental verification of general relativity (the bending of starlight passing through the Sun's gravitational field[106]), an anonymous author presented an optical analogy to help those who thought relativity simply “a mathematical joke.” Signing himself only as “W.G.,” he included the following passage[107]:

---

[104] For a study that includes the original German text, careful English translations, and photographs of Minkowski's agonized corrections to his pre-address manuscript, see P. L. Galison, "Minkowski's Space-Time: From Visual Thinking to the Absolute World," *Historical Studies in the Physical Sciences* (volume 10), 1979, pp. 85–121.

[105] C. W. Misner and J. Wheeler, "Gravitation, Electromagnetism, Unquantized Charge, and Mass as Properties of Curved Empty Space," *Annals of Physics*, December 1957, pp. 525–603.

[106] General relativity had already explained the long-puzzling excess precession of the perihelion (point of closest approach to the Sun) of Mercury's orbit. The excess was an observational (and so experimental) fact which Newton's gravity *cannot* completely explain.

> “Some thirty or more years ago [it was forty] a *jeu d’esprit* was written by Dr. Edwin Abbott entitled *Flatland* . . . Dr. Abbott pictures intelligent beings whose whole experience is confined to a plane, or other space of two dimensions, who have no faculties by which they can become conscious of anything outside that space and no means of moving off the surface on which they live. He then asks the reader, who has consciousness of the third dimension, to imagine a sphere descending upon the plane of Flatland and passing through it. How will the inhabitants regard this phenomenon? They will not see the approaching sphere and will have no conception of its solidity. They will only be conscious of the circle in which it cuts their plane. This circle, at first a point, will gradually increase in diameter, driving the inhabitants of Flatland outward from its circumference, and this will go on until half the sphere has passed through the plane, when the circle will gradually contract to a point and then vanish, leaving the Flatlanders in undisturbed possession of their country . . . Their experience will be that of a circular obstacle gradually expanding or growing, and then contracting, and they will attribute to *growth in time* what the external observer in three dimensions assigns to a movement in the third dimension. Transfer this analogy to a movement of the fourth dimension through three-dimensional space. Assume the past and future of the Universe to be all depicted in four-dimensional space, and visible to any being who has consciousness of the fourth dimension. If there is motion of our three-dimensional space relative to the fourth dimension, all the changes we experience and assign to the flow of time will be due simply to this movement, *the whole of the future as well as the past existing in the fourth dimension* [my emphasis].”

W.G.’s words are a clear and unequivocal statement of the so-called *block universe* concept of four-dimensional spacetime. One can find the block universe concept in the writings of the ancients, too. Consider, for example, the fifth-century B.C. Greek philosopher Parmenides’ view of reality: “It is uncreated and indestructible; for it is complete, immovable, and without end. Nor was it ever, nor will it be; for now it *is*, all at once, a continuous *one*.” And in Thomas Aquinas’ *Compendium Theologiae*, written in the thirteenth century, we find “We may fancy that God knows the flight of time in His eternity, in the way that a person standing on top of a watchtower embraces in a single glance a whole caravan of passing travelers.” This is the block universe idea, too, but whereas for Parmenides it was metaphysics and for Aquinas it was theology, for Einstein and Minkowski it was physics.[108]

---

[107] W. G., “Euclid, Newton, and Einstein,” *Nature*, February 12, 1920, pp. 627–630. As with the mysterious S. (note 94), the editorial staff at *Nature* has informed me that, nearly a century later, there is no longer any record of the identity of W. G. in the journal’s archives.

[108] And for some it was all nonsense. The British philosopher Peter Geach (1916–2013), for example, declared the Minkowskian view to be “very popular with philosophers who try to understand physics and physicists who try to do philosophy.” See P. T. Geach, “Some Problems About Time,” in *Studies in the Philosophy of Thought and Action* (P. F. Strawson, editor), Oxford University Press, 1968. In his introduction to Geach’s essay, editor Strawson put in his two cents by stating the four-dimensional view of reality to be nothing but “fanciful philosophical theorizing.”

The block universe concept may explain the enigmatic statement made by Einstein at the death of Michele Besso (note 103). As interpreted[109] decades later:

> "It seems that Einstein's view of the life of an individual was as follows. If the difference between past, present, and future is an illusion, i.e., the four-dimensional spacetime is a 'block Universe' without motion or change, then each individual is a collection of myriad of selves, distributed along his history, each occurrence *persisting on the world line, experiencing indefinitely the particular event of that moment* [my emphasis]. Each of these momentary persons, according to our experience would possess memory of the previous ones, and would therefore believe himself identical with them; yet they would all exist separately, as single pictures in a film. Placing the past, present and future on the same footing this way, destroys the notion of the unity of the self, rendering it a mere illusion as well."

It appears by his words that Einstein was indeed in agreement with the block universe concept, and that he was attempting to give his friend's family some reason to believe that their father still lives 'somewhen.' The makers of the 2002 film *Minority Report* made use of the block universe concept, even if not intentionally; there we see police stopping crime *before* it happens because they can 'see the future.'

Not everybody believed that this view of spacetime was Einstein's, however. Karl Popper (1902–1994), an Austrian philosopher of science, wrote 28 years after the scientist's death that "Einstein was a strict determinist when I first visited him in 1950: he believed in a 4-dimensional Block-Universe. But he gave this up."[110] Shortly before he wrote those words, however, Popper must have learned something new to convince himself of his final comment, because just 2 years earlier he had declared[111] Einstein to (still) be a determinist. Popper presents no evidence to support his claim of Einstein's philosophical conversion, however, and it would seem that the Besso letter still offers the best insight into his actual view of spacetime shortly before his death. I say this because I think Popper's labeling of Einstein as a determinist is wrong. Determinism says 'If you do A, then B will happen, and if you do not do A then (perhaps) something other than B will happen.' A deterministic universe has plenty of room for free will, because you can *choose* to do A or not to do A, and what you decide makes a difference. A fatalistic universe, however, as is the block universe, simply says 'You will do A and B will happen.' To accept the block universe, as did Einstein, is to be a fatalist, not a determinist.

---

[109] L. P. Horwitz, R. I. Arshansky, and A. C. Elitzur, "On the Two Aspects of Time: The Distinction and Its Implications," *Foundations of Physics*, December 1988, pp. 1159–1193. See also Einstein's own book (note 67) where he wrote "From a 'happening' in three-dimensional space, physics becomes, as it were, an 'existence' in the four-dimensional 'world'."

[110] See the *Seventh International Congress of Logic, Methodology and Philosophy of Science*, volume 4 (Salzburg, Austria, 1983), p. 176. Popper describes his early discussions with Einstein on the reality of time and the four-dimensional Parmenidean block universe in some detail in his autobiography: see volume 1 of *The Philosophy of Karl Popper* (P. A. Schilpp, editor), The Library of Living Philosophers, Open Court 1974, pp. 102–103.

[111] In the Foreword to the book by B. Gal-Or, *Cosmology, Physics and Philosophy*, Springer-Verlag 1981.

Einstein's final position on this, then, *might* have been like that of the fictional time traveler who takes a little girl 25,000 years back into the past, where she sees an ancient ancestor of humanity.[112] She then asks if the ancestor is really alive. The time traveler replies, "Every man who ever lived is still alive, child. In time there is no real death. When a man dies he's still alive 10 min ago, 10 years ago. He's always alive to those who travel back through time to meet him face to face."

Did Einstein *really* believe this? Not everybody thinks so. At the 1922 meeting of the French Philosophical Society, for example, the philosopher of science Emile Meyerson asked Einstein whether the spatialization of time (the idea that time is a dimension on the same footing as the spatial ones) is a legitimate interpretation of Minkowski's spacetime. Einstein's terse answer was that "it is certain that in the four-dimensional continuum all dimensions are *not* [my emphasis] equivalent."[113]

Use of the term *block universe* is generally thought to have originated with the Oxford philosopher Francis Herbert Bradley (1846–1924) who, in his 1883 book *Principles of Logic*, wrote "We seem to think that we sit in a boat, and are carried down the stream of time, and that on the bank there is a row of houses with numbers on the doors. And we get out of the boat, and knock at the door with number 19, and, re-entering the boat, then suddenly find ourselves opposite 20, and having then done the same, we go on to 21. And, all this while, the firm fixed row of the past and future stretches in a *block* [my emphasis] behind us, and before us." The house numbers would seem to be Bradley's way of referring to the centuries. Note that he wrote these words 12 years before *The Time Machine*, and that they preceded Minkowski's famous address by a quarter-century.

But this origin of *block universe* may not be as clear-cut as I have made it appear. Bradley, who was frequently criticized by the Harvard psychologist William James (1842–1910)—a man who argued for free will[114] and indeterminism, concepts disallowed in a block universe—may have been mocked on the idea by James during an address to the students of the Harvard Divinity School in March 1884 ("The Dilemma of Determinism"), the year after Bradley's book had been published. In his address James spoke of a deterministic world as being a "solid" or "iron block" (this are *not* characteristics of determinism, but rather of fatalism, and so James makes the same mistake as did Popper). However, writing the year before Bradley's book, in the April 1882 issue of *Mind*, James wrote (with obvious disdain) of "the universe of Hegel [the German philosopher Georg Hegel (1770-1831)]—the *absolute block* [my emphasis] whose parts have no loose play," as having "the oxygen of possibility all suffocated out of its lungs" and as being a universe in which "there can be neither good nor bad, but [only] one dead level of mere fate." So, perhaps, the chain of evolution of the term block universe is actually from Hegel to James and *then*, finally, to Bradley.

---

[112] F. B. Long, "Throwback in Time," *Science Fiction Plus*, April 1953.

[113] A. Einstein, "La Théorie de la Relativité," *Bullentin de la Société Francaise de Philosophia* (volume 17), 1922, pp. 91–113.

[114] A famous line from James, one that perhaps illustrates his sort of reasoning about free will, is "My first act of free will shall be to believe in free will." If only proving theorems in math and physics were that easy.

We can actually find the block universe in fiction *before* Minkowski (and so certainly before pulp science fiction) came on the scene. In an 1875 (!) story[115] we read of a man who sees, years in advance, his own death in the American Civil War. In the following extract, this man speaks to an unnamed friend (who is the narrator):

> "Do you know," said Bernard, presently, "I sometimes think prophecy isn't so strange a thing . . . I really see no reason why any earnest man may not be able to foresee the future, now and then . . ."
>
> "There is reason enough to my mind," I replied, "in the fact that future events do not exist, as yet, and we cannot know that which is not, though we may shrewdly guess it sometimes . . ."
>
> "Your argument is good, but your premises are bad, I think," replied my friend, . . . his great, sad eyes looking solemnly into mine.
>
> "How so?" I asked.
>
> "Why, I doubt the truth of your assumption, that future events do not exist as yet . . . Past and future are only divisions of time, and do not belong to eternity . . . To us it must be past or future with reference to other occurrences. But is there, in reality, any such thing as a past or a future? If there is an eternity, it is and always has been and always must be. But time is a mere delusion . . . To a being thus in eternity, all things are, and must be present. *All things that have been, or shall be, are* [my emphasis]."

When the block universe concept did eventually appear in science fiction, it did so early. In a 1927 story, for example, a time traveler from the future and a man in the present (who is the narrator) have the following exchange:

> "I have just been five years into your future."
> "My future!" I exclaimed. "How can that be when I have not lived it yet?"
> "But of course you have lived it."
> I stared, bewildered.
> "Could I visit my past if you had not lived your future?"[116]

So, while the block universe has a bit of a history to it, the history of the concept of *mathematical* spacetime in physics has a much clearer origin: it derives from Minkowski, not from Hegel, Bradley, James, or even Einstein (who often gets credit for it even though he didn't use the concept in special relativity in 1905, 3 years before Minkowski's address.). Eventually, of course, Einstein did come to appreciate the power and conceptual beauty of four-dimensional spacetime, and it came to play a central role in his ideas about gravity. Indeed, in Einstein's general theory of relativity gravity *is* (curved) spacetime. The starting point for general relativity (and so a *scientifically plausible* theory of time travel) was Minkowski's creation of spacetime, and he is truly deserving of the title 'father of the fourth dimension.'

---

[115] G. C. Eggleston, "The True Story of Bernard Poland's Prophecy," *American Homes*, June 1875. George Cary Eggleston (1839–1911) had served as a soldier in the Confederate Army.

[116] F. Flagg, "The Machine Man of Ardathia," *Amazing Stories*, November 1927.

Of course, it is true that Newton's physics also talks about an analytical (as opposed to merely philosophical) space and time long before either Minkowski or Einstein, but 'Newtonian spacetime' is something very different from Minkowski's self-described "radical" view.[117] In the Newtonian view there is a universal time, a *cosmic* time, which is the same time for everyone, everywhere, in the universe. At every instant, a cosmic simultaneity exists for Newton. Newton's space is Euclidean; that is, through any point exterior to a line exactly one parallel line can be constructed and those two lines will never meet, all triangles (no matter their size) have an interior angle sum of 180°, and so on. For Newton, space and time were absolutely and uniquely separable. They were, as philosophers like to say, "distinct individuals."

Minkowski changed all that. For him space and time are only relatively separable, and the separation is different for observers in relative motion. For Newton, space and time are the *background* in which physical processes in the world evolve. For Minkowski, spacetime *is* the world.

In a famous philosophical paper[118] by an advocate of the block universe view of reality, we find the words "I . . . defend the view of the world . . . which treats the totality of being, of facts, or of events as spread out eternally in the dimension of time as well as the dimensions of space. Future events and past events are by no means present events, but in a clear and important sense they do exist, now and forever, as rounded and definite articles in the world's furniture." The title of Williams' paper comes from an ancient dilemma stated by Aristotle in his *De Interpretatione*, where he asked a question now classic in philosophy: "Will there be a sea fight tomorrow?"

Aristotle began his famous answer by first posing the following premise: If a statement about some future event is, eventually, shown to be true (or false), then that statement was true (or false) from the moment it was made. Consider, then, the following two assertions: (A) "It is true that there will be a sea fight tomorrow" and (B) "It is true that there will *not* be a sea fight tomorrow." Surely, argued Aristotle, (A) and (B) cannot both be true, but equally surely, one of them must be true. Suppose it is (A) that is true. Then there is nothing that can be done to prevent the sea fight, and so the future is fated. Suppose, however, it is (B) that is true. Then there is nothing that can be done to cause the sea fight, and so the future is fated. The conclusion is the same no matter which assertion is the true one; thus, the future is fated.

---

[117] See, for example, H. Stein, "Newtonian Space-Time," *Texas Quarterly*, Autumn 1967, pp. 174–200; G. Berger, "Elementary Causal Structures in Newtonian and Minkowskian Space-Time," *Theoria* (volume 40), 1974, pp. 191–201; J. Earman and M. Friedman, "The Meaning and Status of Newton's Laws of Inertia and the Nature of Gravitational Forces," *Philosophy of Science*, September 1973, pp. 329–359.

[118] D. C. Williams, "The Sea Fight Tomorrow," in *Structure, Method and Meaning*, The Liberal Arts Press 1951. Donald Williams (1899–1983) was a professor of philosophy at Harvard.

As might be expected, those who like the fatalistic block universe like this conclusion, but, ironically, Aristotle wasn't one of them—he disliked it so much that he struggled to find a way around it. On the other hand, there are philosophers, like Professor Williams (who believed in a fatalistic universe), who reject Aristotle's rejection of his own logic! Professor Williams went so far, in fact, to calling Aristotle's reasoning "a tissue of error" and "swaggeringly invalid." Possibly so, but the philosophical debates over the sea fight question, and the fatalistic (or not) nature of the world, have not ceased to this day.

In an even more famous paper, Professor Williams makes clear his belief that the passage of time is a myth; he poetically declared "the total of world history is a spatio-temporal volume, of somewhat uncertain magnitude, chockablock with things and events."[119] Professor Williams did, indeed, embrace four-dimensional spacetime, and this is demonstrated by the following incredible passage, perhaps his best-remembered words: "It is then conceivable, though doubtless physically impossible, that one four-dimensional area of the time part of the manifold be slewed around at right angles to the rest, so that the time order of that area, as composed by its interior lines of strain and structure, run parallel with a spatial order in its environment. It is conceivable, indeed, that a single whole human life should lie thwartwise of the manifold, with its belly plump in time, its birth at the east and its death in the west, and its conscious stream running alongside somebody's garden path."

Good Lord!

Now, I am willing to admit that Professor Williams probably wrote that wonderful passage mostly for effect,[120] but I ask you—what, if anything, does it *mean*? It is marvelous to read and yet it remains (for me) mysterious.[121] It should come as no surprise that Professor Williams originally presented his papers to the Metaphysical Society of America, rather than to the American Physical Society. But this passage was perhaps not without impact in areas far removed from metaphysics; some years later there appeared a science fiction story[122] that reads as though it had been inspired by Williams. In it, a scientist discovers how to bend his perception of the four dimensions so as to view verticality as duration and duration as verticality. Thus, he is in October while sitting, but when he stands up he is in November! As bizarre as this may seem, such coordinate interchanges actually do occur in the mathematical theory of time machines; we'll see this later, for example, when we discuss Tipler's rotating cylinder time machine.

---

[119] D. C. Williams, "The Myth of Passage," *Journal of Philosophy*, July 1951, pp. 457–472.

[120] In a footnote, Williams sort of admits this when he writes "I should expect the impact of the environment on such a being to be so wildly queer and out of step with the way he is put together, that his mental life must be a dragged-out monstrous delirium." I think this a great understatement.

[121] As it was for some of Williams' fellow philosophers, one of whom bluntly called the 'myth-of-passage' paper "an interesting piece of science fiction": see M. Capek, "The Myth of Frozen Passage: The Status of Becoming in the Physical World," in *Boston Studies in the Philosophy of Science* (volume 2), Humanities Press 1965. Capek's title reflects his view of the block universe as simply a giant refrigerator and so, turning the tables on Williams, we have 'passage' changed to 'frozen passage.' See also note 136.

[122] G. Wolfe, "The Rubber Bend," *Universe 5* (T. Carr, editor), Random House 1974.

By the 1930s the block universe had found a home in pulp science fiction. The block universe view that past and present coexist with the present got dramatic treatment in one story of a high school teacher who invents a "spacetime warp" theory, and who is then tricked by an evil industrialist into implementing it in the form of a gun. The weapon produces incredible effects when it is tested; for example, an allosaurus appears, which we are told is "a carnivorous dinosaur of the Jurassic Age, the most frightful engine of destruction that ever walked the Earth!"[123] At the story's end, the teacher explains what has happened to a crowd of breathless newspaper reporters:

> "Spacetime was warped slightly ... The Einsteinian spacetime continuum buckled ... Because it was superficial, only a little of the past, a little of the future broke through. The folds of the warp distorted spacetime evanescently, erratically skirting the vast gulf where the past lies buried and lightly tapping the vast stores of the future. It is a truism of modern speculative physics that the past and the future exist simultaneously and coextensively in higher dimensions of space. De Sitter has speculated as to the possibility of seeing an event before it happens. It is quite possible, gentlemen. Events of the far future already exist in spacetime."

That 'explains' the dinosaur. In the teacher's words, "You tell me that two men saw an incredible beast. ... They swear it looked like a dinosaur. I think it was a dinosaur, gentlemen. It broke through when the warp tapped the past."

And just 2 years later, Robert Heinlein made world lines the central concept in the first of his many classic tales.[124] The story draws an analogy between a world line and a telephone cable: the beginning and end points in spacetime for the world line of a person (birth and death) are associated with breaks (faults) in a telephone cable. By sending a signal up and down the cable, and measuring the time delay until the arrival of the echo produced by such discontinuities, a technician can both detect and locate the faults. In the same manner, Heinlein's story-gadget sends a signal of unspecified nature up and down a world line and thus locates the birth and death 'discontinuities.' Knowledge of the death date, in particular, causes financial stress among life insurance companies, and an examination of *that* tension (not strange physics) is the fictional point of the story.

And then, 2 years after Heinlein's tale with its serious tone, a far less serious story[125] (featuring an Attila the Hun character who roams up and down the corridors of time kidnapping beautiful women for his harem!), we find an 'editorial' footnote telling its young readers that "scientists—especially the new order of meta-physical scientists—are agreed on the principles of Space-Time. The future is not a thing which *will exist*. Rather it is a thing which *does exist*—all events from the Beginning to the End, exist in a record upon the scroll of Time." This story, itself, was silly, but the block universe metaphysics was up-to-date.

---

[123] F. B. Long, "Temporary Warp," *Astounding Stories*, August 1937.

[124] R. Heinlein, "Life-Line," *Astounding Science Fiction*, August 1939.

[125] R. Cummings, "Bandits of Time," *Amazing Stories*, December 1941.

Somewhat surprisingly, I think, is that even before pulp science fiction embraced the block universe, the concept had already made a deep impression on a broader audience. For example, in a 1928 New York stage play[126] the action alternately takes place in the years 1784 and 1928 and, to explain how that can be, one character (a time traveler) tells another:

> "Suppose you are in a boat, sailing down a winding stream. You watch the banks as they pass you. You went by a grove of maple trees, upstream. But you can't see them now, so you saw them in the *past*, didn't you? You're watching a field of clover now; it's before your eyes at this moment, in the *present*. But you don't know yet what's around the bend in the stream ahead of you; there may be wonderful things, but you can't see them until you get around the bend, in the *future*, can you?"

Then, after this prologue about the stream of time, comes the block universe idea:

> "Now remember, *you're* in the boat. But *I'm* up in the sky above you; in a plane. I'm looking down on it all. I can see *all at once* the trees you saw upstream, the field of clover that you see now, and what's waiting for you around the bend ahead! *All at once!* So the past, present, and future of the man in the boat are all *one* to the man in the plane."

Then, finally, the obvious theological conclusion: "Doesn't that show how all Time must really be one? Real Time—real Time is nothing but an idea in the mind of God!"

To end this section, the block universe conception was cleverly used by one science fiction fan who argued in support of time travel, in reply to another fan how had claimed that a failure of mass/energy conservation was fatal to the plausibility of time travel. Their exchange began with a letter to the editor at *Astounding Stories* in November 1937, written in response to a recent story[127]:

> "Let us say that there is, at a certain time, 'x' amount of matter in the Universe, and 'e' amount of energy. Then if a man of 'a' mass travels backward in time to this particular instant aforementioned, the total amount of matter is thus 'x' plus 'a', while if no other such mass changing occurrences take place, the amount of matter in the future is 'x' minus 'a'. Only a corresponding loss and gain respectively in the amount of energy could explain this conservation of energy, advocates [of time travel] say what they may. But you can't rob or add energy to a Universe nilly-willy! Or perhaps time doesn't enter in on the matter. Perhaps you can add matter in a Universe provided you take it away on some future date."

This fan's concern clearly made an impression on science fiction writers, and the case for conservation of energy is stated in many of the time travel stories that appeared after the publication of this letter.[128]

---

[126] "Berkeley Square" by J. L. Balderson. This play was made into a 1933 movie of the same name, and again in 1951 as the film *I'll Never Forget You*.

[127] O. Saari, "The Time Bender," *Astounding Stories*, August 1937 (see also note 137 in Chap. 1).

[128] Examples include the novels *Lest Darkness Fall* (Henry Holt 1941) by L. Sprague de Camp, and *The Time Hoppers* (Doubleday 1967) by Robert Silverberg.

A reply was soon received by the magazine in a letter (January 1938) from another fan:

> "[A recent letter] implies that the idea of time travel is incompatible with the law of conservation of mass and energy. I believe [the] reasoning is wrong [and that the] difficulty lies primarily in the assumption that a body moved in time is transported into a different Universe. According to Einstein, time and the three normal dimensions are so related as to form a continuous, inseparable medium we call the spacetime continuum. Time is in no way independent of the other components of our Universe. Hence a fixed mass [a time traveler and his machine] moved in time is by no means lost from the Universe, the action being analogous to a shift along any other dimension."

The block, or frozen, universe of Minkowski is clearly reflected in those words.[129]

2.7 Philosophical Implications of the Block Universe

> "Is the future all settled beforehand, and only waiting to be 'pushed through' into our three-dimensional ken? Is there no element of contingency? No free will? I am talking geometry, not theology."[130]

I should tell you now that, despite the enthusiastic embrace of the block universe by Williams and others (including Einstein), there are those who have been harsh in their criticism of Minkowski's spacetime. The major philosophical problem with the block universe interpretation of four-dimensional spacetime is that it looks like fatalism disguised as physics. It seems to be little more than a mathematician's proof of a denial of free will dressed up in geometry. One philosopher illuminated this concern with the following story, one that vividly illustrates the compelling need many humans have to deny a fatalistic world:

> "In a moving picture version of *Romeo and Juliet*, the dramatic scene was shown in which Juliet, seemingly dead, is lying in the tomb, and Romeo, believing she is dead, raises a cup containing poison. At this moment an outcry from the audience was heard: 'Don't do it!' We laugh at the person who . . . forgets that the time flow of a movie is unreal, is merely the unwinding of a pattern imprinted on a strip of film. Are we more intelligent than this man when we believe that the time flow of our actual life is different? Is the present more than our cognizance of a predetermined pattern of events unfolding itself like an unwinding film?"[131]

---

[129] In the context of mathematical physics (*not* science fiction) it has been shown that time travel does *not* imply any fatal violation of conservation of energy. See, for example, J. L. Friedman *et al.*, "Cauchy Problem in Spacetimes with Closed Timelike Curves," *Physical Review D*, September 15, 1990, pp. 1915–1930, and D. Deutsch, "Quantum Mechanics Near Closed Timelike Lines," *Physical Review D*, November 15, 1991, pp. 3197–3217.

[130] The lament of Victorian physicist Oliver Lodge (1850–1940) in his essay "The New World of Space and Time," *Living Age*, January 1920.

[131] H. Reichenbach, *The Direction of Time*, University of California Press 1956, p. 11.

Most people in the Western world would answer *yes* to Reichenbach's question. Most people do find Omar Khayyam's *Rubaiyat* to be a beautiful poem, yes, but still they reject its fatalistic message: "And the first Morning of Creation wrote/What the Last Dawn of Reckoning shall read." Indeed, William James quoted these very words in his 1884 address to the students of the Harvard Divinity School when he argued against fatalism and the block universe.

Besides fatalism, another reason for the stinging words by critics of Minkowski's spacetime is that, in it, events don't *happen*—they just *are*. That is, there seems to be no temporal process of *becoming* in Minkowski's spacetime. Everything is already there and, as what we perceive to be the passing of time occurs, we simply become conscious of ever more of Minkowski's "world points," or events, that lie on our individual world lines. Hermann Weyl (1885–1955), a German mathematical physicist who in his last years was a colleague of Einstein and Gödel at the Institute for Advanced Study in Princeton, expressed this very interpretation in words that have become famous, words that sound very much like those of Wells' Time Traveller: "The objective world simply *is*, it does not *happen*. Only to the gaze of my consciousness, crawling upward along the life line of my body [Minkowski's world line], does a section of the world [spacetime] come to life as a fleeting image in space which continuously changes in time [creating what we call the *now* or the *present*]."[132]

Weyl was skillful in finding poetic ways to express the world line view of reality, but not everybody is convinced by the poetry because it seems to deny the common sense idea of time 'flowing,' of temporal passage; it effectively says time is mind-dependent, a mere *illusion*, as the time traveler in "Berkeley Square" declared (note 126). One philosopher who was particularly opposed to Weyl's view was the British-American academic Max Black (1909–1989), and he expressed his opinion in no uncertain terms: "The picture of a 'block Universe,' composed of a timeless web of 'world-lines' in four-dimensional space, however strongly suggested by the theory of relativity, is a piece of gratuitous metaphysics."[133] Another philosopher who was unhappy with Weyl's view of the block universe was just as blunt: "While philosophers may be forgiven intellectual extravagances of this kind, I think it is a pity when they receive encouragement from theoretical physicists."[134]

Weyl's views had supporters, too, however. Consider, for example, the Time Traveller's speech to his friends at the fateful dinner party that opens *The Time*

---

[132] H. Weyl, *Philosophy of Mathematics and Natural Science*, Princeton University Press 1949, p. 116. Sir James Jeans had already said the same, somewhat less elegantly, in his 1935 Sir Halley Stewart Lecture: "The tapestry of spacetime is already woven throughout its full extent, both in space and time, so that the whole picture exists, although we only become conscious of it bit by bit—like separate flies crawling over a tapestry . . . A human life is reduced to a mere thread in the tapestry." Jeans then immediately *rejected* this fatalistic view: see his *Scientific Progress*, Macmillan 1936, p. 20.

[133] From a book review in *Scientific American*, April 1962, pp. 179–185.

[134] H. A. C. Dobbs, "The 'Present' in Physics," *British Journal for the Philosophy of Science*, February 1969, pp. 317–324.

Machine: “There is no difference between Time and any of the three dimensions of Space except that our consciousness moves along it . . . here is a portrait of a man at 8 years old, another at fifteen, another at seventeen, another at twenty-three, and so on. All these are evidently sections, as it were, Three-Dimensional representations of his Four-Dimensional being, *which is a fixed and unalterable thing* [my emphasis].” Remember, these words were written in 1895, 13 years before Minkowski and his world lines, and of course decades before Weyl’s famous words.

Wells’ passage made a considerable impression on at least one well-known physicist of the time, who references it in his early book on relativity.[135] And in another book on relativity, published the same year, we find the same interpretation of Minkowski’s spacetime as a block universe: “With Minkowski, space and time become particular aspects of a single four-dimensional continuum . . . All motional phenomena . . . become timeless phenomena in four-dimensional space. The whole history of a physical system is laid out as a changeless whole.”[136]

The claim that time is an illusion has some thought-provoking implications concerning the concepts of omniscience and free will, concepts that occur in any discussion of time travel. Some old theology on God’s omniscience, as discussed in Aquinas’ *Summa Theologiae*, is seemingly lent at least some support by Minkowski’s spacetime: “Now although contingent events come into actual existence successively, God does not, as we do, know them in their actual existence successively, but all at once; because his knowledge is measured by eternity, as is also his existence; and eternity which exists as a simultaneous whole, takes in the whole of time . . . Hence all that takes place in time is eternally present to God.” Somewhat paradoxically, however, Aquinas did make a distinction between past and future. In that same work he declares that “God can cause an angel not to exist in the future, even if he cannot cause it not to exist while it exists, or not to have existed when it already has.” For Aquinas, then, whereas the past is rigid and unchangeable, the future is plastic, which is *not* the block universe view of spacetime.

As one theologian has observed,[137] this does not mean that Aquinas thought God had to view all events simultaneous with all others.[138] Rather, our theologian says that Aquinas could have thought of the relationship between God and events as being similar to that between the center of a circle and all the points on the circumference. That is, each point on the circumference has its own identity, coming before and/or after any other point, but the center is related to each and every point on the circumference in precisely the same way. The center, then, is ‘eternity’ and the circumference is the temporal series (‘one thing after another’) of reality. Saying that God is *eternal* is thus very different from saying he is *everlasting*. The first means outside of time, whereas the second means he is a temporal entity but has neither beginning nor end.

---

[135] L. Silberstein, *The Theory of Relativity*, Macmillan 1914, p. 134.

[136] E. Cunningham, *The Principle of Relativity*, Cambridge University Press 1914, p. 191. The use of the words *timeless* and *changeless* explain the characterization of the block universe as being *frozen* (in note 121).

[137] W. L. Craig, “Was Thomas Aquinas a B-Theorist of Time?” *New Scholasticism*, Autumn 1985, pp. 475–483. For the B-theory of time, look back at the discussion in the first section of this chapter.

[138] A science fiction story by Norman Spinrad, “The Weed of Time” (*Alchemy and Academe*, Doubleday 1970) graphically describes what a nightmare that could be!

Our theologian supports the first interpretation, invoking Aquinas’ own words from *Summa Contra Gentiles*: “The divine intellect, therefore, sees in the whole of its eternity, as being present to it, whatever takes place through the whole course of time. And yet what takes place in a certain part of time was not always existent. It remains, therefore, that God has a knowledge of these things that according to the march of time do not yet exist.”

The issue of God’s eternity and his place in spacetime has long been a hot topic among theologians with a scientific inclination. Practically every issue of the learned journal *Religious Studies*, for example, carries an article on the subject, often invoking relativity theory to support some argument. The Bible, itself, can be a confusing guide on this matter. For example, consider the Old Testament story of King Ahab (First Kings 21). Ahab, King of Sumeria, coveted Naboth’s vineyard, but Naboth would not sell. The King retreated, but his wife Jezebel arranged for Naboth’s downfall and judicial murder and thus caused the arrival of all his property into her husband’s hands. This angered God, who commanded Elijah to prophesy disaster on Ahab’s house. Ahab responded with sackcloth, and at that God shifted the disaster to the house of Ahab’s son. The point, here, is that God, declared to be omniscient, seems to have been *surprised* at Ahab’s penitence. God is aware of everything in this tale, but only as it happens. That is, God’s knowledge is subject to growth. This Hebrew concept of God as a participant in history is at odds with the contemporary Christian conception of divine knowledge of all that has been, all that is, and all that will be, a view which has its own Biblical support (for divine eternality). For example, Malachi 3:61 (“For I am the Lord, I change not”), and James 1:17 (“the Father ... with whom is no variableness”).

When *The Time Machine* was serialized in the *New Review*, it included a passage that does not appear in the now classic version of the story in which the Time Traveller explains his view of the connection between omniscience and the block universe to his dinner guests:

> “I’m sorry to drag in predestination and free-will, but I’m afraid those ideas will have to help ... Suppose you knew fully the position and properties of every particle of matter, of everything existing in the Universe at any particular moment of time: suppose, that is, that you were omniscient. Well, that knowledge would involve the knowledge of the condition of things at the previous moment, and at the moment before that, and so on. If you knew and perceived the present perfectly, you would perceive therein the whole of the past. If you understood all the natural laws the present would be a complete and vivid record of the past. Similarly, if you grasped the whole of the present, knew all its tendencies and laws, you would see clearly all the future. To an omniscient observer there would be no forgotten past—no piece of time as it were that had dropped out of existence—and no blank future of things yet to be revealed ... Present and past and future would be without meaning to such an observer ... He would see, as it were, a Rigid Universe filling space and time ... If ‘past’ meant anything, it would mean looking in a certain direction, while ‘future’ meant looking the opposite way.”

Wells' "Rigid Universe" certainly sounds like the block universe, and he (or least, the Time Traveller) seems to have believed that it held important implications for the concept of free will.

The 'Rigid Universe' got an interesting science fiction treatment in a story[139] that imagined an event in the present that occurs 'before it should' (a heart patient learns that her obituary notice will be in next week's *New York Times* when that paper arrives 'early'). As one character explains to the sister of the lady who is soon to die, "The future mustn't be changed . . . For us the events of . . . the future are as permanent as any event in the past. We don't dare play around with changing the future, not when it's already signed, sealed and delivered in that newspaper. For all we know the future's like a house of cards. If we pull one card out, say your sister's life, we might bring the whole house tumbling down. You've got to accept the decree of fate . . . You've got to."

With Einstein's discovery of the *relativity of simultaneity*,[140] we run into the question of 'How can there be any sense to the concept of divine, universe-wide knowledge in a four-dimensional spacetime?' That's because in some frames of reference it is possible for event A to be observed before event B, whereas in other frames the temporal order could be reversed, and so some theological questions prompted by spacetime physics are: 'What is God's frame of reference if he is to be actively involved in human affairs? Could God have a special frame of reference in which he is exempt from the relativity of simultaneity, a frame in which he imposes an absolute order on the sequence of becoming of events? Does it make any sense, that is, to say God enjoys what might be called 'divine immediacy'? And if so, what should we think of a God who follows rules of nature different from those that govern all he is supposed to have made?'

Theologians have debated questions like these for decades, and surely will continue to do so for many more decades to come. Alas, I suspect that physicists who study time travel have either been unaware, unimpressed, or just plain uninterested. That's too bad, because one doesn't have to be religious to appreciate the pure intellectual challenges presented by such questions. For example, consider the following debate between two philosophers, one who believes free will and divine foreknowledge are not compatible, and another who thinks the first has made a fundamental error in blurring the distinction between changing and affecting the past. (This distinction is of *great* importance in any discussion of time travel.) This second philosopher presented some of his arguments in terms of a time traveler to the past[141]:

---

[139] R. Silverberg, "What We Learned From This Morning's Newspaper," *Infinity Four*, November 1972.

[140] This refers to the discovery that two events, which occur simultaneously for one observer in a spacetime, may not be simultaneous for another observer in the same spacetime. This will be discussed in more detail in the next chapter.

> "Consider the following. Parsons (P) has invented a special machine which allows him to go back in time. He enters the machine in 1986 and finds himself in the presence of or, perhaps better, observing, Quigly (Q) in 1876. P is an authority on Q, and knows immediately the situation Q is in. Not only that, but he remembers reading about the particular decision or act which Q made in that situation. Thus one might argue that from P's perspective what Q decides is as if already done. It is not already done, since P is standing there waiting for Q to do it. He has gone back in time. Yet from P's perspective, which is of one come back from the future, it is as if already done, since he knows what Q does decide. Since P strongly believes in the unalterability of the past, it is not within Q's power to do something other than what Q in fact does in that situation. From Q's perspective his decision is not already made nor is the action taken, so that it is in his power at that time to do either x or y. From his perspective, that he will do x rather than y is indeterminate; it is not yet done, though at the same time he can grant that P knows what he will do because for him it is as if he has already done it."

The first philosopher doesn't buy any of this, and dismisses it with "It should be abundantly clear . . . that the fact that such stories are in some way imaginable and intuitively graspable says nothing about their logical coherence." Given the interest among modern physicists in time travel, however, I think the first philosopher wouldn't write that today.

One possible reply to all of these theological issues that spacetime physics prompts can perhaps be found in a paper[142] (written by a philosopher and two mathematicians) that describes a five-dimensional spacetime in which the fifth dimension is initially given the provocative label of the 'eternity' axis. But then the authors lost their nerve and elected to rename it 'anti-time.' It is interesting to note that pulp science fiction anticipated that terminology by decades, as in one story[143] we read "Beyond the fourth there is a fifth dimension . . . Eternity, I think you would call it. It is the line, the direction perpendicular to time." For some, the eternity axis would appear to be perfect to serve as the temporal dimension for God, an axis distinct from the time axis of mere mortals.

The idea of supernatural beings existing outside of mortal time is an old one in theology, and it can also be found in secular literature long before science fiction got hold of it. For example, in the first act of Lord Byron's 1821 poem *Cain*, the fallen angel Lucifer tells Cain and his wife that

---

[141] For the complete exchange between these two philosophers, see W. Hasker, "Foreknowledge and Necessity," April 1985, pp. 121–157, B. Reichenbach, "Hasker and Omniscience," January 1987, pp. 86–92, and W. Hasker, "The Hardness of the Past: A Reply to Reichenbach," July 1987, pp. 337–342, all in the journal *Faith and Philosophy*. Hasker is the 'first' philosopher, and Reichenbach is the 'second' one. See also D. P. Lackey, "A New Disproof of the Compatibility of Foreknowledge and Free Choice," *Religious Studies*, September 1974, pp. 313–318.

[142] J. G. Bennett *et al.*, "Unified Field Theory in a Curvature-Free Five-Dimensional manifold," *Proceedings of the Royal Society of London A*, July 1949, pp. 39–61. A theological interpretation is given in G. Stromberg, "Space, Time, and Eternity," *Journal of the Franklin Institute*, August 1961, pp. 134–144.

[143] L. A. Eshbach, "The Time Conqueror," *Wonder Stories*, July 1932.

With us acts are exempt from time, and we
Can crowd eternity into an hour,
Or stretch an hour into eternity.
We breathe not by a mortal measurement,
But that's a myst'ry.

Before Minkowski, the debates over fatalism (as in Silverberg's story in note 139) and free will had been the exclusive province of philosophers, theologians, and lawyers (if a person has no control over his or her actions, then can we morally and ethically punish that person if those actions happen to be criminal?[144]). After Minkowski, the physicists (at least a few of them) joined the debates. According to one philosopher (note 118) the major motivation driving these debates is "the age-old dread that God's foreknowledge of our destiny can in itself impose the destiny upon us." The implication is, of course, that God is 'outside of time' and so can take in the entire Minkowskian block universe at a glance (hence his foreknowledge).

The relativistic view of the universe as a timeless four-dimensional spacetime seems to provide scientific, mathematical support for the conclusion that not only is the past fixed, but so is the future. Does that mean the future is what it will be—and if so, then why bother agonizing over the many apparent decisions each of us faces every day? If the future will be what it will be, then Christian theologians are left with the puzzling task of explaining what could possibly be meant by the Biblical exhortation (Deuteronomy 30:19) "I call Heaven and Earth to record this day against you, that I have set before you life and death, blessing and cursing; therefore *choose* [my emphasis] life, that both thou and thy seed may live."

This issue has bothered philosophers for a very long time. The so-called Master Argument (the name reflects its supposed invulnerability to rebuttal), for example, comes down to us from its origins in ancient times, in the *Discourses* of the first century A.D. Roman Stoic philosopher Epictetus. That argument can be summarized[145] as follows:

1. The future follows from the past;
2. The past is unchangeable;
3. What follows from the unchangeable is unchangeable;

Therefore,

4. The future is unchangeable.

This certainly does seem to be fatalistic, in effect arguing that all events in a block universe spacetime are recorded in a 'Book of Destiny.' Since ancient times many great works of literature have adopted that view, recounting tales of the foretold fates of men, such as Sophocles' *Oedipus*. It is, in a block universe, as though our conscious experience of the world is no different from that of the man watching the projected film images of *Romeo and Juliet*.

---

[144] For more on this, in the context of time travel, see the penultimate question in the *For Future Discussion* questions at the end of this chapter.

[145] See, for example, the two papers by R. L. Purtill, "The Master Argument," *Apeiron*, May 1973, pp. 31–36, and "Foreknowledge and Fatalism," *Religious Studies*, September 1974, pp. 319–324.

That view is the central issue in the early sixth century A.D. Roman philosopher Boethius' influential *De Consolatione Philosophiae* (circa A.D. 500) which was written during a year of imprisonment before his execution for treason; perhaps he wondered during that year if his fate could have been anything different. Certainly he must have taken some consolation in fatalism, but in fact he tried to argue that God's vision of *all* temporal reality does not limit the freedom to act. According to Boethius, "The expression 'God is ever' denotes a single Present, summing up His continual presence in all the past, in all the present . . . and in all the future." That is, God sees in one timeless and eternal moment all that has been and will be freely chosen.[146]

When the fourteenth century English poet Geoffrey Chaucer prepared a translation of *Consolatione* he was obviously inspired by it when he wrote his very long, famous poem (*Troilus and Criseyde*) on the nature of love (Book IV.140):

Some say "If God sees everything before
It happens—and deceived He cannot be—
Then everything must happen, though you swore
The contrary, for He has seen it, He."
And so I say, if from eternity
God has foreknowledge of our thoughts and deed,
We've no free choice, whatever books we read.

Two modern, purely philosophical rebuttals[147] to Chaucer, however, argue that his poetry misstates Boethius' philosophy when Troilus declares that divine foreknowledge is incompatible with free will. That is, in their view God's omniscience (a fundamental teaching in the theistic religions of Christianity, Judaism, and Islam) *is* compatible with free will (also a fundamental belief in those same religions). Both of these scholarly papers, though, depend much more on the nuances of grammar than most physicists will like.

The connection between spacetime physics and free will was made explicitly by the philosopher who wrote "For philosophers in either field, philosophy of science and philosophy of religion are too often viewed as mutually irrelevant . . . This is unfortunate, because sometimes the problems can be quite parallel and a consistent resolution is required. One especially intriguing case in point concerns, in

---

[146] In his *The Sirens of Titan*, a 1959 novel meant to be a parody of God's omniscience, Kurt Vonnegut gave the curious name of *chrono-synclastic infundibulated vision* to God's power to see the past and future.

[147] G. I. Mavrodes, "Is the Past Unpreventable?" April 1984, pp. 131–146, and A. Plantinga, "On Ockham's Way Out," July 1986, pp. 235–269, both in *Faith and Philosophy*.

Fig. 2.2 The common view of time

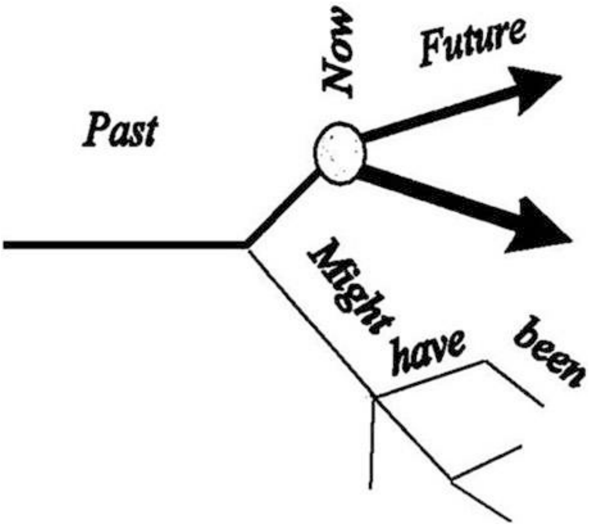

philosophy of science, the possibility of ... time travel and, in philosophy of religion, the relationship between divine foreknowledge and human freedom."[148]

That philosopher could well have included science fiction writers in his group of people interested in both spacetime physics and free will. In one story,[149] for example, a man in the twenty-fifth century is about to travel back into the past to escape criminal prosecution. He is aked where he'd like to go, and he replies "I do not understand the paradoxes—what if I choose to build gravity-deflectors in Ancient Rome?" When he is told (correctly) that he couldn't do that because it didn't happen, he persists: "But if I can choose any period, it means I can alter history at will—which presumes that the present can also be changed." Then, at last, he gets the explicit answer that bothers nearly everyone: "The real answer is that in the final analysis your decision to choose a certain time period is already made, and the things you will do [in the time traveler's personal time] are already determined. Free will is an illusion; it is synonymous with incomplete perception." The same idea appears in another tale (note 57); when one character says, "What you are saying is that the future is fixed, and that you can read it, in every essential detail," the response is "Quite right ... both those things are true."

However, no matter how hard we try—and by *we* I include even those physicists and philosophers who embrace the block universe with its support of time travel to the past—it is very difficult to break free of the view of time as shown in Fig. 2.2. That is, as the passage of time up to the present or *now* (with all to the left of that instant as the past), while to the right of the *now* we have multiple possible futures (depending on our free will choices). Lying to the side of all that (in our thoughts and imaginations) are all that 'might have been' if we had made different choices than we did at earlier times in the past.[150]

---

[148] W. L. Craig, "Tachyons, Time Travel, and Divine Omniscience," *Journal of Philosophy*, March 1988, pp. 135–150. Tachyons are hypothetical faster-than-light particles that theoretically travel backwards through time. They will be discussed in Chap. 5.

[149] W. Kubilius, "Turn Backward, O Time," *Science Fiction Quarterly*, May 1951.

With all that said, even if events are really laid out in the spatial and temporal web that constitutes the four-dimensional block universe, there still remains the great mystery of why we see them unfold in the particular sequence that we do. Why not in reverse order? Why, indeed, do we see what we call *time* run from what we call the past to what we call the future and, indeed, what do we really mean by *past* and *future*? As you'll see in the next chapter, these are not easy questions, and nearly everybody who has thought about them believes we are not yet even close to knowing the answers.

On that perhaps gloomy note, it seems appropriate to end here with a few more words from St. Augustine's *Confessions*, with words that follow those that helped open this chapter: "I confess to you, Lord, that I still do not know what time is. Yet I confess too that I do know that I am saying this in time, that I have been talking about time for a long time, and that this long time would not be a long time if it were not for the fact that time has been passing all the while. How can I know this, when I do not know what time is? Is it that I do know what time is, but do not know how to put what I know into words? I am in a sorry state, for I do not even know what I do not know!"[151]

2.8 For Further Discussion

In the comics one of Superman's more interesting adversaries is Mr. Mxyzptlk (pronounced *mix-yez-pitle-ick*), a being with seemingly magical powers from the Land of Zrfff in the fifth dimension. Mr. Mxyzptlk's powers are not really because of magic, however, but are 'merely' the result of his hyperspace world with its extra dimension. Mr. Mxyzptlk, for example, in one of his misadventures with Superman in 1954, begins selling a

(continued)

---

[150] Figure 2.2 is based on a similar one in C. K. Raju, "Time Travel and the Reality of Spontaneity," *Foundations of Physics*, July 2006, pp. 1099–1113.

[151] There is another view of time even darker than St. Augustine's, which denies the existence of both future and past, and doesn't offer us much either for that special moment we call the present (or *now*). This view, called *presentism*, was hauntingly expressed in some lyrics I heard in the final episode of the second season (2015) of the HBO series *True Detective*: "There is no future/There is no past/In the present nothing lasts." Now *that* is depressing! Still, there are philosophers who believe even this view can support time travel: see S. Keller and M. Nelson, "Presentists Should Believe in Time-Travel," *Australasian Journal of Philosophy*, September 2001, pp. 333–345.

newspaper called the *Daily Mpftrz* in competition with the *Daily Planet*. Unlike a traditional newspaper that reports what *has* happened, the *Daily Mpftrz* (your guess is as good as mine!) prints what *will* happen. As Mr. Mxyzptlk says, “You see, as a resident of the fifth dimension, I can get all the news I want from the *fourth* dimension!” The science editor at the *Daily Planet* explains the meaning of that to his boss, Perry White: “That’s right, Mr. White . . . many physicists consider *time* the fourth dimension . . . so if Mr. Mxyzptlk can travel from the fifth dimension to our three-dimensional world, he most likely *is* able to see the future!” (This leaves unanswered the question of *why* he continues to challenge Superman when he knows he will always be defeated—*as he always is*!) Presumably a five dimensional world would have our three spatial and one temporal dimension (for a total of four), and so the question now is: what is the nature of the additional (fifth) dimension? Is it spatial or is it temporal? (There is a brief appearance of the fifth dimension in the 2014 movie *Interstellar*, but we aren’t told much of anything about its possible structure.) Discuss and compare the world of four space dimensions and one time dimension, with the world of three space and *two* time dimensions. (In Chap. 5 we’ll discuss a possible connection between two-dimensional time and time travel.)

In the text it is stated that “If A and B are mutually causative, then ‘A causes B’ coupled with ‘B causes A’ seems to lead to ‘A causes A.’” Suppose, however, that we imagine two adjacent sunken pools of water, **a** and **b**, on the same horizontal surface, with each pool filled to the brim. An overflow from one pool will flow into the other pool. Now, define the events A and B as ‘A is the overflow of pool **a**’ and ‘B is the overflow of pool **b**.’ Thus, A causes B and B causes A. Does the conclusion ‘A causes A’ make physical sense in this specific case? Discuss at length.

When reading A. C. Clarke’s story “Technical Error” (see note 99), we learn that a rotation through 4-space *inverts* “the unlucky Nelson.” The ‘solution’ to this awkward situation is to flip Nelson through 4-space a second time and so back to ‘normal.’ (When *Thrilling Wonder Stories* reprinted this tale in June 1950, after its original publication in 1946, the title was changed to the more appropriate “The Reversed Man.”) Clarke may have missed an

(continued)

important technical ‘detail,’ however, in that when first flipped through 4-space *everything* inverts, and so matter becomes anti-matter and Nelson would have instantly been annihilated in a 100 % conversion of matter to energy (that is, the flipped Nelson would have initiated a very large explosion). Compare this to Alice’s concern in her flipped world (Lewis Carroll’s *Through the Looking Glass*) when she wonders “Perhaps Looking-glass milk isn’t good to drink.” Explain why Lewis Carroll certainly was *not* thinking of matter/anti-matter explosions when he wrote his novel. What *do* you think he might have had in mind?

A time travel story, even earlier than Clarke’s, that uses spacetime ‘rotations,’ was authored by Edmond Hamilton (1904–1977), one of the pioneering pulp fiction writers. In his “The Man Who Saw the Future” (*Amazing Stories*, October 1930), a man is hauled before the Inquisitor Extraordinary of the King of France to explain his mysterious disappearance, and subsequent reappearance, in an open field, amid thunderclaps and in plain sight of many onlookers. As the story unfolds, we learn that the man was transported five centuries into the future, from A.D. 1444 to 1944, by scientists working in twentieth-century Paris. The thunderclaps were produced by spacetime ‘rotations,’ as the atmospheres of 1944 and 1444 were reversed. A skeptical Inquisition naturally finds this tale preposterous and the first time traveler is burned at the stake as a sorcerer. Can you think of *why* such ‘atmospheric swaps’ might produce thunderclaps?

A trip around a Möbius strip reverses the ‘handedness’ of a plane figure (left and right are swapped). You can see this for yourself by making a Möbius strip, and then sliding an arrow (pointing *across the width* of the strip) around the strip. (Cut a notch in the side of the strip to mark the starting point, with the arrow pointing at the notch.) When you get back to the notch, the arrow will point *away* from the notch. Notice that the arrow never left the surface of the strip, or crossed any ‘weird’ boundary. Then, read H. G. Wells’ short story “The Plattner Story” and comment on its use of ‘handedness.’

The *autoinfanticide paradox*, which results when a time traveler tries to kill his younger self, continues to fascinate both physicists and philosophers, and papers regularly appear in the scholarly literature on the topic: see, for example, Kadri Vihvelin, “What Time Travelers Cannot Do,” March 1996, pp. 315–330 (which introduces Suzy the time traveler); Ira Kiourti, “Killing Baby Suzy,” June 2008, pp. 343–352; Peter B. M. Vranas, “What Time Travelers May Be Able to Do,” August 2010, pp. 115–121; and Joshua Spencer, “What Time Travelers Cannot *Not* Do (but are responsible for anyway),” October 2013, pp. 149–162, all in *Philosophical Studies*. All deal with an issue that is psychologically fascinating: *moral responsibility*. Spencer, in particular, opens with this definition: Someone is morally responsible for an action *only* if she could have done otherwise. As he goes on to write, “If I have been attacked and both of my legs have been broken, then it seems illegitimate to criticize me for failing to run away; I could not have done otherwise.” And yet all of these papers are on a point that (I think) physicists would soon lose interest in: is the question ‘If Suzy is a time traveler, can Suzy kill baby Suzy, given that Suzy doesn’t kill baby Suzy?’ the same question as ‘If Suzy is a time traveler, can Suzy kill baby Suzy, given that Suzy is now alive?’ The answer to the first question is, from pure logic, NO, while the answer to the second question is just bit squishier: it all depends on what the word *can* means. For the second question, Suzy *can* kill baby Suzy if she has a weapon (knife, gun, poison, etc.) and she is in the past next to baby Suzy, but it is just that she doesn’t because otherwise Suzy wouldn’t be alive now (which is a given). Such debates seem unlikely to produce any insights into the physics of time travel. Compare this situation to the old schoolboy conundrum “What happens when an irresistible force meets an unmovable object?’, which is a *self-inflicted* ‘paradox.’ That is, the words *irresistible* and *unmovable* are mutually exclusive and so, used this way, it should be no surprise that we have a conflict. Are the two time travel questions above, concerning Suzy, confusing through a similar mushy use of grammar? Or are they deeper than that? Vigorously defend your position.

In addition to H. G. Wells, another nineteenth-century writer who was highly influential in bringing the fourth-dimension out of academia and into public consciousness was the mathematician Charles Howard Hinton (1853–1907). Hinton was no angle-trisecting crank, having earned an M.A. at Oxford, an appointment in the mathematics department at Princeton, and then another at the University of Minnesota. Later, with the help of the eminent astronomer

(continued)

Simon Newcomb, he obtained a position at the Naval Observatory in Washington, D.C., and was on the staff of the United States Patent Office at the time of his sudden death. Hinton was a man to be taken seriously. His first published essay "What Is the Fourth Dimension?" appeared in 1880, and then in book form in 1884 as part of his *Scientific Romances* (a phrase used by Hinton before it became associated with Wells' science fiction many years later). That book received a generally favorable review in *Nature* (March 12, 1885, p. 431). At one point he wrote "We might then suppose that the matter we know extending in three dimensions has also a small thickness in the fourth dimension," an idea that was used a few years later by the well-known British mathematician W. W. Rouse Ball (1850–1925) in an attempt to explain gravity. Hinton was extremely inventive, and he also proposed four-dimensional-space models for static electricity. Find out more about Hinton's life and work: a good source to start with is *Speculations on the Fourth Dimension: Selected Writings of Charles H. Hinton* (R. Rucker, editor), Dover 1980. Take a look, too, at J. E. Beichler, "Ether/Or: Hyperspace Models of the Ether in America," in *The Michelson Era in American Science 1870—1930* (S. Goldberg and R. H. Stuewer, editors), American Institute of Physics 1988.

Chapter 3
The Physics of Time Travel: Part I

> *"… within forty-eight hours we had invented, designed, and assembled a chronomobile. I won't weary you with the details, save to remark that it operated by transposing the seventh and eleventh dimensions in a hole in space, thus creating an inverse ether-vortex and standing the space-time continuum on its head."*
>
> —almost certainly *not* the way to build a time machine[1]

3.1 The Direction of Time

> "Of all the problems which lie on the borderline of philosophy and science, perhaps none has caused more spilled ink, more controversy, and more emotion than the problem of the direction of time … [T]he main problem with 'the problem of the direction of time' is to figure out exactly what the problem is or is supposed to be!"[2]

Before we start talking about the physics of time travel, let me say a few more words on time itself, in a way slightly less metaphysical that was the discussion in the previous chapter (which is why I'm writing this *here*, in a chapter with an increased emphasis on the analytical). When we speak of journeying to either the future or the past, we are implicitly making a distinction in the direction of the time traveler's trip. But does time actually have a *direction*? Is there an *arrow* that points the way? The answer seems obvious: *of course* time has a direction. After all, everybody 'knows' it flows from past to future. There is a curious language problem here, however, because we also like to say the present recedes into the past, which implies a 'flow' in the opposite direction, from future to past. Well, despite this snarled syntax, can we at least distinguish past from future, whichever way time flows?

---

[1] L. Sprague de Camp, "Some Curious Effects of Time Travel," in *Analog Readers' Choice*, Dial 1981.

[2] See note 54 in Chapter 2.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_3

This would seem to be an important question to answer because for the phrases *flow of time* and *direction of time* to have any objective meaning at all, it must be somehow possible to identify a difference between past events and future ones. The special moment at which that distinction occurs is known as the *now* or the *present* and, as events make the transition associated with that distinctive difference, between past and future, we say that the *now* (the present) moves or flows. Philosophers—and science fiction writers and physicists, too, who after all are human beings with human senses like everybody else—call this common feeling that we all have, of the passage of time, the *psychological arrow of time*. One philosopher gave an amusing (tongue-in-cheek) gastronomical interpretation of the moving *now* as follows:

> “New slices of salami are continually being cut from a nonexistent chunk of salami called the future. The *present* is the slice on top of the pile. The past are the pieces beneath this, and even though they are not present they still continue to exist in the same way that the top slice of salami does. . . . This [concept] faces humiliation before the embarrassing question of how fast the pile of salami slices grows.”[3]

The ‘moving now’ does present a problem for physicists because there is nothing in the laws of physics that marks the present moment as unique, and therefore nothing that reflects a ‘flow’ of time, nothing that models the reality of a ‘moving now’ becoming part of the past and the events of the future becoming, successively, the new ‘now.’ As a philosopher wrote long before time travel became a serious topic in the physics literature, “Talk of the flow of time or the advance of consciousness is a dangerous metaphor that must not be taken literally.”[4]

What that philosopher may well have had in mind is that all events in the block universe simply have coordinates in spacetime, and there is nothing corresponding to ‘have been’ (past), ‘are’ (present), or ‘will be’ (future). There is no ‘moving now’ in the block universe except for its subjective presence in our conscious minds. All we can say from physics is that events are ordered in an earlier/later sequence, and in fact, even that relatively weak condition holds only for *causally* related events.[5] The relativistic, four-dimensional block universe view of spacetime that so many physicists (including Einstein) so dearly love seems to have no room for an objective theory of the *flow* of time. And yet, even for those same physicists, there is a powerful psychological sense that time *does* flow. But are they mistaken? It is a fact that, with not just a little irony, Gödel (the ‘discoverer’ of time travel) was convinced that the possibility of a block universe spacetime with CTLs/CTCs

---

[3] R. Gale, “Some Metaphysical Statements About Time,” *Journal of Philosophy*, April 25, 1963, pp. 225-237. For many, this analogy may well bring to mind a pile of baloney rather than one of salami (and I think this was Gale’s intention).

[4] J. J. C. Smart, “The Temporal Asymmetry of the World,” *Analysis*, March 1954, pp. 79-83.

[5] Two events A and B are *non*-causally related if their separation in spacetime is such that a particle would have to travel at a *superluminal* speed (faster than light) to go from A to B. We’ll discuss the physics of causally related events later in this chapter.

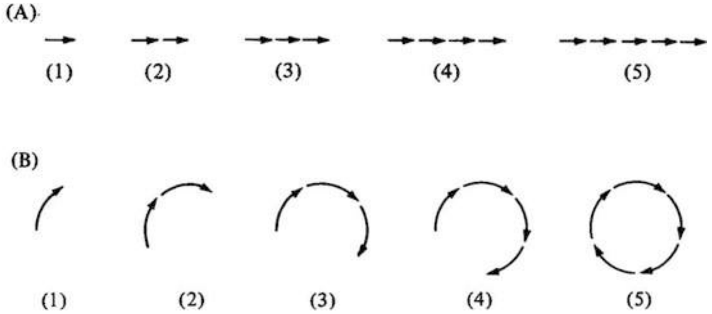

Fig. 3.1 Gödel's unreality of time argument

implies that the passage or flow of time *makes no sense*![6] To see how Gödel arrived at such an astonishing conclusion, consider Fig. 3.1.

In part A of the figure[7] we see what most people who talk of a passage of time intuitively mean, as time progresses through a sequence of instants (shown as line of left-to-right arrows going from (1) to (2) to (3) and so on). At each stage the right-most arrow is the present, and the arrows to the left of it (behind it) are the past, and the arrows to the right of the present are not shown because they are in the future and so don't exist yet. When you ask, at each step, which arrow is the *now*, the answer is clear.

Matters are dramatically different in part B of the figure, which shows the arrows forming a closed (circular) loop. Now there is no distinction between past and future, as each arrow is both ahead *and* behind any other arrow. In addition, there is no arrow that is uniquely the now. So, concluded Gödel, the passage of time can have no meaning in a temporal loop. As should come as no surprise, not everybody is convinced by this sort of argument.[8]

In principle, so it would seem, we can achieve perfect knowledge of what has happened but only imperfect prediction of what might happen. This observation seems to be at least a start at being able to tell past from future. And, in fact, the nature of the distinction between the two intervals of time seems obvious: we *remember* past events, but not future ones. As philosophers have so nicely put it, events in the past have formed *traces*, such as skulls, footprints in the sand, fossilized skeletons, surgical scars, photographs, taped recordings, carved stones, and the like, whereas future events appear not to have formed traces. But is that necessarily so? Is it impossible for future events to create traces? The commonsense answer is *yes*, because of cause and effect, which dictates that there *must* be a temporal asymmetry in trace formation. That is, traces are the effects of *prior* causes. That line of reasoning leads us quickly to the fundamental issue of causation (which we've already encountered in Chap. 2), an issue that no discussion of time travel can avoid.

---

[6] Gödel clearly states this in his 1949 philosophical essay (note 15 in the Introduction) concerning his 1949 technical paper (note 11 in the Introduction).

[7] This figure is based on the interpretation of Gödel's reasoning as presented by the philosopher Palle Yourgrau's 1991 book *The Disappearance of Time: Kurt Gödel and the idealistic tradition in philosophy* (Cambridge), which was expanded and reprinted a few years later under the new title *Gödel Meets Einstein: time travel in the Gödel universe*, Open Court 1999. Yourgrau later wrote a less technical version: *A World Without Time: the forgotten legacy of Gödel and Einstein*, Basic Books 2005.

[8] See, for example, S. Savitt, "Time Travel and Becoming," *The Monist*, July 2005, pp. 413-422.

Part of the problem we have with backward time travel, and cause and effect, is as I've already mentioned, with language. The distinct and separate concepts of the temporal ordering of events, and of causality, have become merged in everyday thought. It is considered *obvious* to modern minds that if event A causes event B, then A *must* happen first. There is, however, at least one historical example of a similar merging of concepts that is parallel to our modern mixing of time order and causality—an example that shows how an issue can seem obvious and natural to the minds of one period of time, and yet to the minds of another period (our modern times) seem confused, odd, peculiar, even laughable.

As a *physicist* wrote in a paper[9] on advanced (that is, inverted causality) effects:

> *Ancient Egypt was an essentially one-dimensional country strung out along the Nile, which flows from south to north. The winds were conveniently arranged to be predominantly northerly. To go north, a traveler could let his boat drift, while with a sail he could move south against the slow current. For this reason, in the writing of the ancient Egyptians, "go downstream (north)" was represented by a boat without sails, and "go upstream (south)" by a boat with sails. The words (and concepts) or north-south and up-downstream became merged. Since the Nile and its tributaries were the only rivers known to the ancient Egyptians, this caused no difficulties until they reached the Euphrates, which happened to flow from north to south. The resulting confusion in the ancient Egyptian mind is recorded for us to read today in their reference to "that inverted water which goes downstream (north) in going upstream (south)."*

Often we can work our way free of the difficulties we create for ourselves with language, but only through common agreement. For example, the chairman of the board calls a meeting to order with mixed tenses by declaring "The meeting *will* take place *now*" and then saying at the end, "We will meet again *next* month, *same* time." We all know what these sentences mean, but only by our cultural heritage and not by the process of applying logic. The language problem causes similar difficulties for not only for fictional time travelers, but also for the physicists/philosophers who study the possibility of time machines. So—beware!

The idea of time flowing is a popular one, and it repeatedly appears in the time travel literature as the "river of time" or the "ocean of time." The deep psychological appeal of this sort of 'water language' has, not surprisingly, attracted the attention of philosophers. We can find one of the earliest expressions of the view in the *Meditations* of the second-century A.D. Roman emperor and Stoic philosopher Marcus Aurelius, who wrote: "Time is like a river made up of events which happen, and a violent stream; for as soon as a thing has been, it is carried away, and another comes in its place, and this will be carried away, too." A most interesting essay on why such metaphors often seem so intuitively appropriate has been offered by one philosopher, who points out[10] that the seductiveness of the image of 'time as flowing water' is sufficiently great that you often find it in the scientific literature, too (Newton, you'll recall, wrote specifically in his *Principia* of time flowing). As for why such an image has such a powerful grip on our imaginations, I think we need look no further than to Kant. As he wrote (1781) in *Critique of Pure Reason*, "Time is nothing but the form of inner sense, that is, of the intuition of ourselves and of our inner state . . . Because this inner intuition yields no shape, we endeavor to make up for this want by analogies." And what better than a rushing stream of water to represent our feeling of time rushing by?

---

[9] P. L. Csonka, "Advanced Effects in Particle Physics," *Physical Review*, April 1969, pp. 1266-1281.

Still, no matter how intuitive such water metaphors may be, they can still easily befuddle us as well. To quote our philosopher (note 10), "Time a river! A queer sort of river that. Of what sort of liquid does it consist? Is time a liquid? A very peculiar liquid indeed!" A classic paper by the philosopher Donald Williams (discussed in Chap. 2, note 119) expresses similar doubt about the water image of time. In the course of his writing, he presents a truly staggering collection of entertaining examples of 'time as metaphor,' of which I repeat just a few here: time flies, goes, marches, and rolls, as well as flows And then he offers *this* provocative imagery: the evolution of our lives is like "a moving picture film, unwinding from the dark reel of the future, projected briefly on the screen of the present, and rewound into the dark can of the past." Wow![11]

Returning to the water metaphor, the French astronomer Charles Nordmann (1881–1940) opened and closed his 1925 book *The Tyranny of Time* with following gloomy but all too true summary of the overwhelming sense we all have of the inexorable, one-way 'flow' of time. (The ellipses in what follows denote over 200 pages!) "Nothing can equal the bitter sweetness of dreaming on the banks of Time, that impalpable and fatal river strewn with dead leaves, our wistful hours carried downstream like rudderless wrecks . . . In the eternal wave which rocks us, carries us along, and soon swallows us up, there is no rock to which we can fasten our frail barques; the very buoys we put out to measure our course are only floating mirages; and on the mysterious foundation of things our anchors slide along and fail to bite." A young person sees time, from Nordmann's perspective, as an ocean on which golden mornings arrive like waves from the future, whereas for an older person, liquid time is a nightmare flood, a swollen black torrent sweeping him first into the yawning abyss of the past and, ultimately and finally, into the eternal silence of the dark grave.

---

[10] J. J. C. Smart, "The River of Time," *Mind*, October 1949, pp. 483-494.

[11] And how about *this* image of time: Time is a snowball, with the center marking the beginning of the past, with ever new 'presents' accreting on the ever increasing surface as the snowball rolls down the hill of history!

The metaphor of time as a flowing river was ready-made for early science fiction writers, such as Caltech math professor Eric Temple Bell (1883–1960). His eventual novel *The Time Stream* began to appear in December 1931 as a serial in the science fiction pulp *Wonder Stories*, and Bell (writing as 'John Taine') made great use of the idea of time as a flowing stream, a stream in which one could swim into either the future or the past. There is strong evidence that Bell actually wrote the novel in July of 1921, but was unable to find a publisher for a decade, so odd did editors find the premise. By the time of its publication, others had beaten Bell into print.

The watery image of time had appeared a year earlier, for example, in a tale that played with the erosive nature of time in a dramatic way. As two time travelers speed into the future to rescue a friend, one of them describes the scene for us: "We huddled together in the whirling time girdling machine, cutting through the years as a ship's prow breasts surging waves. I could not help but think of the years as waves, beating in endless succession on the sands of eternity. They wore all away before them with pitiless attrition. Time seemed to eat all with dragon jaws."[12]

This image of time was taken a step further 3 years later in a story in which a large number of adventurers, from all across time, find themselves stranded at precisely the same place (in space and time). One of them offers his theory of what is behind this remarkable coincidence: They all have faulty time machines, like faulty boats, and all have hit the same snag on the 'river of time.' As he explains, "You may turn boats adrift on a river at many points, and they will all collect together at the same serious obstacle whether they have traveled a hundred or two miles. We are now at some period where the straight flow of time has been checked — perhaps it is even turning back on itself . . . [We] have struck some barrier and been thrown up like so much jetsam."[13]

The 'flow' of time does have its critics, of course. The British-American philosopher Max Black (1909–1988) argued[14] that questions about the direction of time are meaningless because there can be no direction to something that (he asserted) does not flow. His reasoning was that if time does flow, then he ought to be entitled to ask *how fast* it flows. That requires, in turn, a *metatime* or *supertime* for measuring the flow rate of 'ordinary' time. But because supertime must flow, too, we would then need a super-supertime, and so off we trip into what would appear to be the black hole of a McTaggert-like infinite regress. The view, of an infinite regress of times, was forcefully rejected by another philosopher with these sharp words: “the very idea of super (or hyper)-time is indeed repulsive in its redundancy and its aroma of dilettante physics.”[15]

---

[12] E. A. Manley and W. Thode, "The Time Annihilator," *Wonder Stories*, November 1930. This is the same magazine that, months later, finally published Bell.

[13] J. Wyndham, "Wanderers of Time," *Wonder Stories*, March 1933. Notice again, that we have the same magazine (whose editor must have had a particular fancy for such tales).

[14] M. Black, "The 'Direction' of Time," *Analysis*, January 1959, pp. 54-63.

A hierarchy of hypertimes has not bothered other analysts, however, and an entire subfield of specialty among philosophers (and some physicists, too) in time analysis has developed in what is called *multidimensional time*. One practitioner in this specialty sarcastically rejected the infinite regress complaint as a valid objection—he called it “a crushing and unanswerable position” but actually meant just the opposite—and stated that it was not at all clear (at least, not to him) why supertime must flow.[16] After all, he argued, we measure the flow of a river with respect to its banks without requiring that the banks themselves flow. (That actually strikes me as being a point that deserves debating, but I have not been able to find any mention of it in the later philosophical literature.) The idea of multiple time dimensions is particularly attractive for one sort of time travel (we’ll take it up at the end of this chapter), but it enjoys far more popularity among science fiction writers and philosophers than it does with physicists.

Professor Black’s objection (note 14) to talk of time ‘flowing’ was based, at least in part, on the observation that there are uses of the word *direction* that are not directly tied to something flowing. For example, consider the statement ‘He is facing in the direction of north.’ Black argued that this is mere pointing, and it is not at all the same as *moving* north. He then dismissed the possibility of there being any meaning to the direction of time, writing that making an analogy of time “with a sign-post or an index finger is too far-fetched to be worth considering.” This claim (which some may feel leans too much on grammar) is, of course, an affirmation of the myth-of-passage view made famous a few years earlier by Donald Williams (note 119 in Chap. 2).

Not just philosophers have rejected the idea of time flowing. In his 1966 novel *October the First Is Too Late*, which deals with a world in which different parts of Earth simultaneously experience different eras of the past (see *For Further Discussion* at the end of this chapter for more on what this might mean), the British cosmologist Fred Hoyle (1915–2001) calls the ‘river of time’ a “grotesque and absurd illusion,” and a “bogus idea.” Another fictional work that agrees with Hoyle’s non-moving image of time is the 1979 *Roadmarks* by Roger Zelazny (1937–1995). In that novel we read of “the Road,” along which story characters can travel but which doesn’t itself move; exits from the Road lead to the various centuries (which sounds a lot like the Francis Bradley’s 1883 book that may have given the block universe its name). *Roadmarks* is a clever bit of writing, with many allusions to the paradoxes of time travel, but its explanation of the Road’s origin as having been constructed by *dragons* (!) greatly undermines its interest for physicists.

---

[15] D. Zeilicovici, “Temporal Becoming Minus the Moving-Now,” *Nous*, September 1989, pp. 505-524.

[16] C. W. Webb, “Could Time Flow? If So, How Fast?” *Journal of Philosophy*, May 1960, pp. 357-365.

In a block universe spacetime, there is no flow of time, but one philosopher believed that to be simply because the block universe is *incomplete* in its representation of reality. Writing in 1925, Hans Reichenbach (1891–1953) asked “What does ‘now’ mean? Plato lived before me, and Napoleon IV will live after me. But which one of these three lives *now*? I understandably have a clear feeling that *I* live now. But does this assertion have an objective significance beyond my subjective experience?”[17] Reichenbach went on to answer his question in the affirmative, and to deduce that the block universe view is missing something: “In the condition of the world, a cross-section called the present is distinguished; the ‘now’ has objective significance. *Even when no human is alive any longer, there is a ‘now’* [my emphasis] . . . In the four-dimensional picture of the world, such as used by the theory of relativity, there is no such distinguished cross-section But this is due only to the fact that an essential content is omitted from this picture.”

So, what *is* Reichenbach’s ‘missing essential content’? Feeling that the block universe is unacceptably fatalistic—in his words of ridicule, “the morrow has already occurred today in the same sense as yesterday”—he found his answer in the probabilistic theory of quantum mechanics. Classical physics argues that given total information about the state of the world *now*, one could in principle calculate perfectly the future or the past; one could both predict and retrodict. In contrast, quantum mechanics distinguishes past from future in a fundamental way.

Quantum mechanics does not deny that in principle we can know the past with exquisite accuracy, because each and every event leaves traces, evidence that is available to all with the means to find and decode them. But quantum mechanics also takes as truth that there is an unavoidable uncertainty to the future. The instant that this uncertainty is crystallized into fact was taken by Reichenbach to be the very definition of ‘now.’ The ever-increasing record of the past, in turn, defines (for Reichenbach) the *movement* of the ‘now.’ Reichenbach believed that with these observations he had at last captured the ‘moving now’ in mathematical theory, and that he had finally elevated the present from speculative psychology to solid physics, and that he had shown that the ‘flow of time’ is independent of the need for a conscious mind. However—

A later, powerful analysis[18] of the time-flow issue, combining philosophy with physics, comes down solidly in support of the *opposite* conclusion: it expresses the view that a ‘moving now’ *is* only in our minds and is *not* an intrinsic attribute of reality. The premise of that argument is that a mind-dependent flow of time is incompatible with what is called the *relativity of simultaneity* (to be discussed later in this chapter) which states that there is no universal cosmic-wide ‘now’ (this is a fundamental conclusion of special relativity). For example, it is meaningless to ask what is happening on a planet in the Andromeda galaxy (two million light-years distant) *right now*.

---

[17] I’ve taken this quotation from A. Grünbaum, “Is There a ‘Flow’ of Time or Temporal Becoming?” in *Philosophical Problems of Space and Time*, Knopf 1963.

[18] L. R. Baker, “Temporal Becoming: The Argument from Physics,” *Philosophical Forum*, Spring 1975, pp. 218-236.

Early science fiction stories are full of theories about the nature of ‘now,’ and the vast majority of them have no basis in scientific thought. Some of them are ingenious, however, and even though they are largely the pet ideas of the authors (and no one else’s), perhaps they resulted in some young readers of the science fiction pulps of the 1930s and 1940s thinking about deeper matters than did the comic strips of “Buck Rogers,” “The Lone Ranger,” or “Terry and the Pirates.” For example, according to one story, time is a wave and the ‘moving now’ we experience is carried on a crest of that wave. There are time waves both ahead and behind the crest we happen to be on (so we are told), and so each such crest carries a different ‘now’ for a different reality—hence the curious title.[19]

In another, more recent tale[20] about object duplication via time travel (which we’ll discuss in Chap. 4, but you’ll recall H. G. Wells was worried about this long ago), nine (!) copies of the same person from the year 2314 meet in 1870 to try and figure out what is going on. Part of their interesting discussion is the following analysis of the ‘present’:

> “Gentlemen, I think I understand,” said the first James Thomas.”
>
> “Eight faces turned toward him, and he felt as though he were looking into multiple mirrors.”
>
> “We hold that time is a single instant — the instant of the Present —which travels through Duration — do we not?”
>
> “Eight heads nodded.”
>
> “We assume that time passes in a manner analogous to the stringing of an infinite number of beads. Each bead is the instant of Now when it is last on the chain. Beads are continually being added, and each one is the only Now until another is placed after it.”
>
> “Yes, that is my theory,” said another James Thomas. “It can also be likened to the process of knitting. No matter how many stitches are knitted, there is only one last stitch, only one Now.”

Einstein, too, was greatly bothered by the place of ‘now’ in time, perhaps even more than were James Thomas and his ‘friends.’ In an autobiographical essay, the philosopher Rudolf Carnap (1891–1970) recalled a conversation about this with Einstein in the early 1950s, at the Institute for Advanced Study in Princeton: “Once Einstein said that the problem of the Now worried him seriously. He explained that the experience of the Now means something special for man, something essentially different from the past and the future. That this experience cannot be grasped by science seemed to him a matter of painful but inevitable resignation. . . . Einstein thought . . . that there is something essential about the Now which is just outside the realm of science."[21]

---

[19] R. Ray, “Today’s Yesterday,” *Wonder Stories*, January 1934.

[20] A. and P. Eisentein, “The Trouble With the Past,” in *New Dimensions 1* (R. Silverberg, editor), Doubleday 1971.

3.2 The Arrows of Time

> "On a microscopic level there is no preferred direction for time. The equations of motion don't give a damn whether time moves forward or backward."[22]

The central issue for philosophers of time (and for physicists, too, I think) is that of its reality (or not): is time *objective and something that really flows*, or is time simply a *mind-dependent illusion* and nothing more than an artifact of our incomplete perception of reality? As the previous section shows, there is little consensus on this issue. As a start on trying to get a handle on the matter, looking into a so-called 'arrow of time' may give us some guidance. I'll begin with the arrow I've already mentioned, the psychological arrow. As discussed before, this is the feeling we have of a 'moving now,' a feeling that has no appearance anywhere in physics. A 'moving now' simply has no place in any universe devoid of the physical processes in a brain that give rise to what we call consciousness—but that doesn't mean physicists don't wonder about the 'moving now' just as much as does everybody else (remember Einstein)! As one physicist wrote in a technical journal, "What does 'Now' mean? This question must surely be the starting point of any attempt at understanding the nature of time."[23]

Well, no matter whether time actually flows or not, most of us still believe we have had a past and hope we will have a future. Each of us thinks we can easily tell one from the other, too. We have, in fact, many not so subtle indications from our everyday lives of the obvious direction of time. Nearly all of these indications have the common theme of *irreversible change*. As the British mathematician J. J. Sylvester once put it, "The whirligig of time brings about its revenges."[24] The Roman poet Ovid, who died when Christ was a teenager, said the same in his *Metamorphoses* with the famous words "Time, the devourer of all things." The image of time as devourer of all that is mortal was brilliantly presented by James Barrie in his *Peter Pan*, with the crocodile who had swallowed a ticking clock chasing Captain Hook all about Neverland.

---

[21] Quoted from *The Philosophy of Rudolp Carnap* (P. A. Schlipp, editor), The Library of Living Philosophers, Open Court 1963, pp. 37-38. For a view contrary to Einstein's, from another physicist, see K. B. M. Nor, "A Topological Explanation for Three Properties of Time," *Il Nuovo Cimento B,* January 1992, pp. 65-70, which claims to develop a geometrical explanation for the flow of time, and so (says Nor) there *is* an objective, mathematical reality to the 'moving now.'

[22] A science fiction character pretty accurately sums-up what a modern physicist would tell you today, in L. Eisenberg's story "The Time of His Life," *The Magazine of Fantasy & Science Fiction*, April 1968.

[23] J. P. Cullerne, "Free Will and the Resolution of Time Travel Paradoxes," *Contemporary Physics*, July-August 2001, pp. 243-245.

[24] J. J. Sylvester, "A Plea for the Mathematician," *Nature*, December 30, 1869, pp. 237-239.

No one yet has escaped the biological decay processes of time, and inanimate objects are no less immune to this aspect of time. Logs and cigarettes burn in the stove and ashtray, but they never unburn. Our cars rust but never 'unrust.' An explosion has never been seen to reverse itself, to form a dynamite stick or a bomb casing out of a collapsing fireball. Our world seems, indeed, literally to be built on an irreversible movement toward chaos, death, and decay. Lewis Carroll uses this observation in his *Through the Looking-Glass* when Alice tells Humpty Dumpty "one can't help growing older." And speaking of Humpty Dumpty, his famous fall provides a dramatic example of a one-way evolution from past to future; he wasn't at all convinced that Alice was correct but, once he had splattered, then

All the King's horses and all the King's men
Couldn't put Humpty Dumpty together again.

While we are on the subject of Mr. Dumpty, it is also worthwhile to note that nobody has ever figured out how to unscramble an egg. Why is that? One philosopher speculated that the answer is found in the "irreversible organic phenomena" taking place in our brains which results in our flow of consciousness always being in the same direction.[25]

More subtle than the undignified undoing of a prideful egg is the phenomenon of memory, which seems trivial only because most people have not thought very carefully about it. We remember the past while remembering nothing about the future. We might, in fact, be tempted to use the phenomenon of memory to answer the question of how to tell past from future. Anything you can remember *is* the past. But that is a circular definition, as discussed by Professor Smart (note 4) who observed that to ask why memory is always of the past "is as foolish as to ask why uncles are always male, never female." In *Through the Looking-Glass* the White Queen tells Alice that "it's a poor sort of memory that only works backward" but, except for the claims of clairvoyants, it seems that is the only sort of memory any of us has. Why is that so? Of course, that would *not* be the case for a time traveler while in the past. His personal past, which he would remember, would be the future for the world around him.

For physicists, the question of the direction of time is one of profound mystery. There seems, in fact, to be no fundamental reason why time should not be able to go from future to past—but then what would 'future' and 'past' *mean*—even though no one has ever observed time to do so. All the laws of classical physics, including general relativity, and quantum mechanics, too (except for the K-mesons mentioned in Chap. 1) involve time in such a way that they ignore its sign. In other words, replacing t with $-t$ results in a perfectly valid description of something that could actually happen. But not all such possibilities are observed to occur. Why not?

---

[25] H. Margenau, "Can Time Flow Backwards?" *Philosophy of Science*, April 1954, pp. 79-92.

In an unpublished paper written in 1949, while doing the work that would bring him a share of the 1965 Nobel prize in physics, Richard Feynman (1918–1988) wrote[26] "The relation of time in physics to that of gross experience has suffered many changes in the history of physics. The obvious difference of past and future does not appear in physical time for microscopic events . . . Einstein discovered that the present is not the same for all people [the relativity of simultaneity, to be discussed later in this chapter] . . . It may prove useful in physics to consider events in all of time at once and to imagine that we at each instant are only aware of those that lie behind us. The complete relation of this concept of physical time to the time of experience and causality is a physical problem which has not been worked out in detail. It may be that more problems and difficulties are produced than are solved by such a point of view."

Feynman did not elaborate on what he meant by the "problems and difficulties" with that point of view (which is clearly that of the block universe), but surely he had the logical paradoxes of time travel high on his list. As the Yale philosopher Henry Margenau wrote (note 25) in a tutorial on Feynman's work, "The theory of quantum electrodynamics developed by Feynman incorporates reversals in the course of time and thereby cherishes, in the minds of many, *an age-old phantasy* [my emphasis] of more than scientific appeal [which sounds like time travel to me]."

Because the individual classical equations of microscopic physics are time-reversible, the distinction between past and future for individual particles disappears. The equations are said to be symmetric with respect to time; the algebraic sign of t is irrelevant in the classical laws. It must be understood, however, that there *is* a crucial point to appreciate. When a physicist says *time reversal*, she is talking about a system evolving backward in forward time—that is, all the individual particle velocity vectors are instantly reversed at once. This is distinct from the time-reversed worlds of philosophers and science fiction writers (which we'll get into later in this chapter) in which time itself 'runs backwards.' The physicist's point of view is clearly expressed in an early essay by a chemist: "Every equation and every explanation used in physics must be compatible with the symmetry of time. Thus we can no longer regard effect as subsequent to cause. If we think of the present as pushed into existence by the past, we must in precisely the same sense think of it pulled into existence by the future."[27] More than three decades later, a mathematician and a physicist presented a similar statement: "In classical dynamics, the past completely determines the present, and therefore, by symmetry, the future also completely determines the present."[28]

---

[26] S. S. Schweber, "Feynman and the Visualization of Space-Time Processes," *Reviews of Modern Physics*, April 1986, pp. 449-508.

[27] G. N. Lewis, "The Symmetry of Time in Physics," *Science*, June 6, 1930, pp. 569-577.

[28] O. Penrose and I. C. Percival, "The Direction of Time," *Proceedings of the Physical Society* (London), March 1962, pp. 605-616.

Besides the physics, there is also an interesting theological connection to time reversal. As one philosopher put it, "If all the laws are time reversal invariant and so no irreversible processes occur in the physical Universe then there is no inherent, intrinsically meaningful difference between past and future . . . If this is actually the natural case, then all mankind's major religions which preach a creation of the Universe (by a supernatural agency) and imply, accordingly, a differentiation between the past and the future . . . would have to make appropriate adjustments."[29]

There are, as you might expect on such a controversial topic, dissenters to the view that the classical laws of physics are necessarily time-reversible. Dirac himself wrote that "I do not believe there is any need for physical laws to be invariant under time and space reflections, although all the exact laws of nature so far known do have this invariance."[30] Dirac did not, unfortunately, elaborate on just why he felt that way, but with the later discovery of K-mesons his position is seen to have been 'ahead of its time'! In a famous science fiction story[31] dealing with the direction of time, one character finally puts his finger on the real puzzle of the question of time: "How can a man live backward? You might as well ask the Universe to run in reverse entropy." That cogent question brings us, in fact, to the first scientific explanation developed to explain the observed asymmetric nature of time.

It was the Englishman A. S. Eddington (1882–1944) who gave the picturesque name, the *arrow of time*, to the observed asymmetric nature of time's direction from past to future. He was also one of the popularizers of an explanation for the arrow, using the famous second law of thermodynamics.[32] The second law of thermodynamics states that a measure of the internal randomness or disorder—what is called the *entropy*—of any closed system (that is, one free of external influences) continually evolves toward that of maximum disorder, toward the condition called *thermodynamic equilibrium*. Indeed, so striking is this increase in entropy S with time in a macroscopically large system that the increase in entropy has come to be thought of as actually *defining* the direction of time. Eddington, however, was not the originator of the entropy concept. The history of entropy can be traced back to before the turn of the century, to the great Austrian scientist Ludwig Boltzmann (1844–1906) and his famous H-theorem. The quantity H in that theorem is directly related to the more familiar entropy,[33] defined by Boltzmann in 1877.

---

[29] H. Mehlberg, "Philosophical Aspects of Physical Time," in *Basic Issues in the Philosophy of Time* (E. Freeman and W. Sellars, editors), Open Court 1971.

[30] P. A. M. Dirac, "Forms of Relativistic Dynamics," *Reviews of Modern Physics*, July 1949, pp. 392-399.

[31] A. Boucher, "The Chronokinesis of Jonathan Hull," *Astounding Science Fiction*, June 1946.

[32] A. S. Eddington, *The Nature of the Physical World*, Macmillan 1929.

[33] The H-theorem was a direct continuation of the work by the Scottish physicist James Clerk Maxwell (1831-1879) on the statistical properties of gas molecules (determining the probability density function of the molecules' speeds). In 1866 Maxwell found this function for the particular case of thermodynamic equilibrium. In 1872 Boltzmann found the differential-integral equation the function satisfies in general, even if the condition of thermodynamic equilibrium doesn't hold. From this Boltzmann was able to define a quantity H that he showed evolves in time such that

The entropy S of a system in a given state is proportional to W, which is the number of different possible ways the state can occur as a result of all possible variations of system's internal, microscopic structure. The calculation of W is usually quite complicated, but in various highly idealized systems it can be straightforward. Consider, for example, a vacuum cylinder with a thin membrane dividing the interior into halves. Suppose that we insert (to be specific) six molecules into the left half of the cylinder (and none into the right half). If we define the microscopic state of the system to be the number of molecules in the left half, then initially $W = 1$ because there is just one way to put all six molecules on the left side. This represents the state of *minimum* entropy, the state of maximum order that is most distant from thermodynamic equilibrium. If we now puncture the membrane then the molecules, once confined to the left side, are free to move about the entire cylinder. At any given instant we can imagine counting the number of molecules on the left side—suppose that at some particular instant we count five, with one molecule having moved to the right side. Then, $W = 6$, because there are six ways to pick the molecule that has moved from left to right, and so the entropy has increased.

We think of the thermodynamic equilibrium state as being the state with equal numbers of molecules in both halves of the cylinder, and that state has the *maximum* entropy. (Can you show that this state is associated with $W = 20$?) With such a small number of molecules, it is *not* clear that W (and so S) will *inexorably* increase with time; perhaps, after one of the six molecules has gone to the right, it then returns to the left side before any of its companions have joined it on the right. Such an event is called a *reversal*, and it will happen with some non-zero probability. But the more molecules there are in the cylinder (instead of six, make the number a million million million—still a small amount of gas in our everyday world, hardly enough to fill a sewing thimble), the more likely it becomes that the value of S *will* monotonically increase with time.

The steady increase in entropy is often observed in the everyday, large-scale world. A drop of ink in a glass of water spreads out in an expanding cloud, a cloud we *never* see collapse backward into an ink drop. A long rod of metal, initially hotter at one end than at the other, evolves toward a constant temperature along its entire length. We *never* see a uniformly warm rod spontaneously begin to cool at one end and grow hot at the other. A hot bath grows cold—nobody has *ever* seen a bath at room temperature suddenly, all by itself, begin to heat up and then boil in the middle of the tub while the edges freeze into ice chunks. In all of these cases, the end (future) state represents greater internal randomness or disorder than does the beginning (past) state.

---

solution to his differential-integral equation approaches Maxwell's equilibrium solution. The H-theorem says that H always *decreases* in systems not in equilibrium and is at a *minimum* in systems in equilibrium.

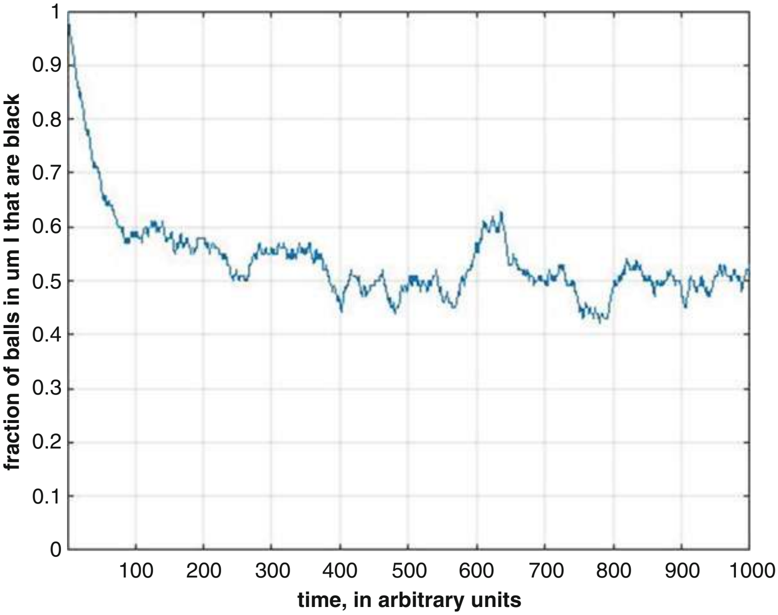

Fig. 3.2 Simulation of the Ehrenfest entropic gas clock

That is, low entropy *was* the past, and high entropy *will be* the future. The increase in entropy seems to define a direction to time, and so entropy has come to be called the *thermodynamic arrow of time*.

The first formal entropy model for the direction of time was put forth in a 1907 paper by the Austrian physicist Paul Ehrenfest (1880–1933) who was a friend of Einstein, and his Russian-born wife Tatyana (1880–1964), who was a skilled mathematician and her husband's occasional collaborator. In their paper the Ehrenfests developed one of the mainstays of physics, the so-called *entropic clock*. This clock, a statistical model based on the then new probability mathematics of Markov chains—after the Russian mathematician A. A. Markov (1856–1922)—describes how gases diffuse, and it is both a simple and a powerful concept. The Ehrenfest model is illustrated in Fig. 3.2, in a computer-generated plot based on a discussion by Princeton physicist John Wheeler of black hole fame (see Chap. 1).[34]

Imagine two urns, I and II, each containing n balls. Initially, at time $t = 0$, all of the balls in Urn I are black and all of the balls in Urn II are white. Then, at time $t = 1$ (in arbitrary units), a ball is selected at random from each urn and (instantaneously) placed in the other urn. This select-and-transfer process is repeated at times $t = 2$,

---

[34] J. A. Wheeler, "Frontiers of Time," in *Problems in the Foundations of Physics* (G. T. diFrancia, editor), *Proceedings of the International School of Physics* (Course 72), North-Holland 1979. See also W. J. Cocke, "Statistical Time Symmetry and Two-Time Boundary Conditions in Physics and Cosmology," *Physical Review*, August 25, 1967, pp. 1165-1170.

3, 4, ... At any given time each urn always contains n balls, but only at $t = 0$ are the colors of all the balls in a given urn necessarily the same. The phrase "selected at random" means (for example) that the probability of selecting a black ball from an urn containing b black balls is b/n. At any given time we completely describe the state of both urns by specifying the number of black balls in Urn I (or the number of white balls in Urn II, and so on). It is easy to write a computer simulation of this physical process,[35] and Fig. 3.2 shows how the fraction of black balls in Urn I evolves toward 0.5 as time increases. The plot is for $n = 100$ (200 balls total). The important observations are that (1) the evolution of the state of the system is toward 50 % black balls in Urn I (and this would be the case for 'almost all' sequences of random selections of the balls from the urns), and (2) the evolution is not monotonically decreasing from 100 % black balls to 50 % black balls, but rather has never-ending fluctuations about 50 % that may, in fact, be rather large in both amplitude and in duration.

There is a real puzzle with the entropic clock that may not be immediately apparent. The motion of each of the individual molecules is described by time-reversible physics, but when we average over 'many' molecules (assuming 200 molecules is 'many') we lose detailed information about the individual molecules. The puzzle is then how is it that by *reducing* our knowledge of a system, through statistical averaging, we then find it displaying a *new property*, that of asymmetric time evolution, that we didn't see before when we watched the individual molecules. And if that question isn't troublesome enough, we also have two additional puzzles called the 'reversibility' and the 'recurrence' paradoxes to consider as well.

The reversibility paradox is the question raised earlier: the classical equations of physics work just as well with time running in either direction, and so *why don't* things actually go 'backward'? This question, originally raised by the British mathematical physicist and engineer William Thomson (1824–1907)—better known as Lord Kelvin—in 1874, was brought to Boltzmann's attention in 1876 by the German physical chemist Johann Loschmidt (1821–1895), one of Boltzmann's professors at the University of Vienna. Boltzmann's answer to this apparent paradox was that it *is* imaginable that a world could run backward if initial conditions were suitable. For example, if all the velocity vectors of every particle in an equilibrium state were reversed, then the system *would* unwind backward in time toward its original non-equilibrium condition. That is, a system in thermodynamic equilibrium, the state of highest entropy, could evolve toward one of low entropy. Boltzmann even suggested that such might be the case for regions in our own universe, that there might actually be beings in a world somewhere 'out there' who experience time running counter to our earthly experience. He said that in 1877, and it is a remarkable statement for a conservative nineteenth-century professor.[36] However, Boltzmann continued, from most given states there are vastly more ways for entropy to increase than there are for it to decrease, and that is why we see what we see, a continuous *increase* in entropy.[37]

---

[35] I used MATLAB, and you can find the code — gasclock.m — in Appendix C, written in such a low-level way as to be virtually 100% transferable to just about any of the popular scientific programming languages, and easily executed on an inexpensive laptop. Note that there are no K-mesons in the code (!) and so, as stated in Chapter 1, they aren't responsible for the uni-directional time behavior depicted in the figure.

To find a science fiction writer speculating on reversed time people is, of course, much less remarkable! One pulp story, in fact, presents a curious treatment of the nuances of reversed time in which people talk backward (along with a marvelous bathroom scene of a man un-washing his hands!). This tale[38] tells us a young physics teacher who is "twisted into a reversed Time Stream" by an electrical discharge. As he lives backward in time, he observes everybody about him appearing to run in reverse, but even more puzzling is that they have developed a "dreadful, granite-like hardness." We soon learn why:

"For a while he could not understand the impenetrable hardness of external objects which he had experienced; it seemed they ought rather to be of intangible transiency, much as a dream, since he was re-viewing the Past. But a moment's thought gave him the logical answer. The Past is definite, shaped, unalterable, as nothing else in Creation is. Therefore, to argue that he could move or alter any object here [the past] was to argue that he could change the whole history of the world or cosmos. Everything he saw about him had happened, and could not be changed in any way. On the other hand, he was fluid, movable, alterable, since *his* future still lay before him, even if it had been reversed; he was the intruder, the anomaly. In any clash between himself and the Past, the Past would prove irresistible every time."

This passage reflects the modern view that the past cannot be changed, but explains that view in a way different from that generally accepted today. Modern physicists and philosophers invoke consistency requirements (which we'll take up in the next chapter) to explain the 'solidity' of the past. The author of this story also failed to explain why his physics teacher had no trouble moving about through the air of the past, which apparently is not any more resistant to being displaced than were air molecules before time reversal occurred.

---

[36] The Austrian-British philosopher Karl Popper (1902-1994) called Boltzmann's willingness to consider the possibility that different regions of the universe could have different directions of time "staggering in its boldness and beauty," but when on to say that Boltzmann must be wrong because "it brands unidirectional change an illusion [which] makes the catastrophe of Hiroshima an illusion." That is an emotional argument, of course, and although one of great power, I fail to see how it is related to physics. See Volume 1 of *The Philosophy of Karl Popper* (P. A. Schlipp, editor), Open Court 1974, pp. 127-128.

[37] For more on Boltzmann's views on entropy, see the end of his letter "On Certain Questions of the Theory of Gases," *Nature*, February 28, 1895, pp. 413-415.

[38] C. F. Hall, "The Man Who Lived Backwards," *Tales of Wonder*, Summer 1938. The modern classic of a time-reversed world is Philip K. Dick's 1967 novel *Counter-Clock World*. We'll encounter another time-reversed world again in Chapter 4.

It didn't take long for science fiction writers to incorporate entropy as time's arrow into time travel. In one early tale there is the brief statement that entropy is behind the operation of its gadget.[39] And a few years later the inventor of a "warp gun" tells us that "The stupendous distortion of the warp may actually bring about a sort of kink in spacetime, and result in a reversal of entropy"[40] and, sure enough, when the gun is fired a woman, who is hit by the warp, ages 70 years in seconds (which is, of course, exactly the opposite of what we would expect from a "reversal of entropy"!). Just a year later, the story of a college student about to flunk his senior physics course appeared.[41] An examination is scheduled for the following day, but he needs a week and a half of study time. To his rescue comes ENTROPY, INC., a company that sells time by placing its clients inside a "time-cabinet" in which the local entropy is greatly accelerated. To someone looking through a window at the interior of the time-cabinet, the occupants would appear as characters in a speeded-up movie. Referring to Eddington by name, the author tells us that "entropy is what makes time irreversible — is what gives us the feeling of the flow of time."[42]

In a hilarious, melodramatic story featuring one of early science fiction's stereotypical 'mad scientists,' the entropic arrow of time is the scientific explanation for time travel.[43] There we read of Bryce Field, "a master-scientist, a demon, cruel, ruthless," who is rejected in love by the stupendously beautiful Lucy Grantham. Her lack of enthusiasm is perhaps understandable, as Bryce is described as having "a lean-jawed, sunken-eyed" appearance, along with "lank, untidy hair sprawled across his massive forehead." As Lucy tells him at one point, "I could never love you; you are too clever, too brilliantly scientific." After hearing that, it is no surprise that before we are more than a page or two into the tale that we learn Bryce has Lucy strapped to a steel table in an underground laboratory-in-a-cave. There he tells her of her fate: "You are going on a long journey, my dear. So long a journey that even I, master-scientist, do not know when it will end. A journey into the future — alone! . . . You, Lucy, shall be the victim of entropy! . . . I have discovered how to make a [globe] of non-time. Entropy will be halted . . . You will be plunged into an eternal 'now.'"

And so the mad Doctor Field throws the switch on the wall of his "instrument-littered" cave on July 17, 1941, and Lucy remains "suspended" in time until the outside world reaches the date of August 9, 2450. That is the day she is at last dug-up from the cave by "big and muscular" engineer Clem Bradley and his "square-jawed" sidekick Buck Cardew, who uses a "warp in spacetime" to release Lucy from her "globe of non-time."

---

[39] M. J. Breuer, "The Time Valve," *Wonder Stories*, July 1930.

[40] F. B. Long, "Temporary Warp," *Astounding Stories*, August 1937.

[41] R. M. Farley, "Time for Sale," *Amazing Stories*, August 1938.

[42] Also citing Eddington was a tale by D. W. O'Brien, "The Man Who Lived Next Week," *Amazing Stories*, March 1941, which uses entropy to explain time travel. This curious story has the traveler arriving in the future with his clothing aged, which later 'de-ages' when the return trip is made!

[43] P. Cross, "Prisoner of Time," *Super Science Stories*, May 1942.

A few years later entropy was used in a similar but vastly more ‘scientific’ way. In that tale[44] we read of a scientist who has discovered “a field in which entropy was held level.” As the reader is told, “An object in such a field could not experience any time flow — for it, time would not exist,” since time flow is a *change* in entropy, and the ‘change’ of a level (or constant) field is zero. This interesting tale speculates on how such a field could have fantastic home uses (“Imagine cooking a chicken dinner, putting it in the field, and taking it out piping hot whenever needed, maybe twenty years hence!”). But its real use in the story is as a stasis generator for preserving fatally ill people until medical science has learned how to cure their diseases. This is, then, a high-tech method of suspended animation, of time travel into the future that is different from simply freezing (a clock in such a field would not age or measure the passage of personal time).

The gadget that does all this is called, somewhat sinisterly, the “Crypt,” which we are told also makes a great bomb shelter, too, because “not even an atom bomb could penetrate a stasis field.” The reason for that is intriguing: “The field requires a finite time in which to collapse — only there is no time in it.” The interior of the Crypt is, quite literally, a frozen block of time more rigid and unyielding than the strongest steel.

As science fiction left the age of pulps and moved into the modern era, entropy continued to be useful a justification for time travel. Arthur C. Clarke used it,[45] as did Robert Silverberg. This last tale[46] is particularly interesting, as Silverberg pursued entropy beyond simply invoking it as a mere casual throwaway mention. When a newspaper from the future appears on people’s doorsteps, the initial astonishment is replaced with puzzlement as the papers rapidly disintegrate. That is the result (we are told) of “entropic creep.” The explanation continues, informing us that it is sort of like a strain in a geological fault (Silverberg has lived for decades in California, now and then a place of large to huge earthquakes, and it isn’t surprising that he uses this particular imagery): “Entropy you know is the natural tendency of everything in nature to come apart at the seams as time goes along. These newspapers must be subject to unusually strong entropic strains because of their anomalous position out of their proper place in time.”

Earlier I mentioned we had *two* puzzles associated with entropy; we’ve discussed ‘reversibility,’ so what’s the other one, the ‘recurrence paradox,’ all about? The recurrence paradox is quite different from reversibility; it is based on a result established in 1890 by the great French mathematician Henri Poincaré (1854–1912). Motivated by the question of the stability of the motion of three masses governed by Newton’s laws of mechanics (think, for example, of the Sun, the Earth, and the Moon), Poincaré showed that starting from almost any initial state, any fixed volume system with a finite amount of energy and a finite number of degrees of freedom will return *infinitely often and with arbitrarily little deviation* to almost every previous state. If you wait long enough, implies Poincaré's astonishing theorem, Pearl Harbor will happen again—and again, and again, and .... In 1896 the German mathematician Ernst Zermelo (1871–1953) used this result, which philosophers call the 'eternal return,' to claim that there could be no truly irreversible processes and thereby cast doubt on the idea that entropy *always* and *inexorably* increases.

---

[44] P. Anderson, “Time Heals,” *Astounding Science Fiction*, October 1949.

[45] In, for example, his story of the tragic end of a geologist fifty million years in the past: “Time’s Arrow,” *Science Fantasy*, Summer 1950.

[46] “What We Learned from This Morning’s Newspaper,” *Infinity 4*, November 1972.

Even for very small systems, however, such as a mere handful of molecules, the recurrence time is extremely large, and this was, in essence, Boltzmann's reply to Zermelo's concern. For example, if the gas-filled cylinder of our entropic clock has just 100 molecules (not the six I used in the earlier example), and if transitions from one side of the cylinder to the other side take place at the rate of one million per second, then the recurrence time has been calculated to be something like 30 million billion years![47] And for the universe itself, the recurrence time is simply incomprehensible. Mathematicians call 1 followed by a hundred zeros a *googol*, and the recurrence time in years for the universe has been estimated to be 1 followed by a googol of zeros (a so-called *googolplex* of years).[48]

Using a wonderful bit of imagery, one analyst wrote of the enormity of the recurrence time of a system considerably *less* complex than the universe this way: "If a man shuffled just a single pack of cards as rapidly as an individual molecule hits other molecules in air, and if a snail started to crawl around the universe ... at the rate of one centimeter *during the life of the sidereal system* [my emphasis], the snail would have got round the universe many millions of times before it would become at all likely that the man would have got the pack back to the original order."[49] If this is what it takes to get a pack of cards back to its initial state, then try to conceive of the time interval required to restore *the world* to December 7, 1941.[50]

The notion of eternal recurrence considerably predates Poincaré, and its scientific (as opposed to astrological) study can be traced back to the fourteenth century.[51] A 'more recent' claim for eternal recurrence, also based on scientific arguments (conservation of energy), can be found in many places in the writings of the German philosopher Friedrich Nietzsche (1844–1900)—see, for example, his *The Gay Science* (1882) and *Thus Spake Zarathustra* (1883)—again predating

---

[47] J. M. Blatt, "Time Reversal," *Scientific American*, August 1956.

[48] The googol is a gigantic number, far greater than the number of raindrops that have fallen on the Earth during its entire history. And the googolplex is light years beyond that.

[49] R. B. Braithwaite, "Professor Eddington's Gifford Lectures," *Mind*, October 1929, pp. 409-435.

[50] Pulp science fiction writers, of course, were not discouraged by such calculations, as they depended on the *certainty* of recurrence over *infinite* time. See, for example, S. G. Weinbaum, "The Circle of Zero," *Thrilling Wonder Stories*, August 1936, and L. D. Gunn, "The Time Twin," *Thrilling Wonder Stories*, August 1939.

[51] R. Small, "Incommensurability and Recurrence: From Oresme to Simmel," *Journal of the History of Ideas*, January-March 1991, pp. 121-137.

Poincaré. All of Nietzsche's arguments are flawed,[52] but they are rational, physical arguments, as opposed to arguments based on metaphysics or theology. In fiction, a glimmer of the idea of a repetition of human affairs preceded Poincaré by some years, too.[53]

An important caveat concerning recurrence is that we could never *know* of it because the state of all the historical records (geological, memories, books, photographs, and so on) would, as part of the physical state of the universe, also recur. And so those records could, up to the instant *before* the recurrence, contain no signature *of* the recurrence because the recurrence has not 'yet' happened![54] The 1993 movie *Groundhog Day* stumbles on this point, as it has a character (for some unexplained reason) live through the same day over-and-over *and he is aware he is doing that*. Indeed, he can change events within that time loop at will. It is interesting to note that one 'time loop' pulp science fiction tale[55] specifically avoided that error (and cited Nietzsche, to boot), and so demonstrated that pulp science fiction *could* have some philosophical merit to it.

While the enormous recurrence time for the universe may seem reason enough to reject the possibility of Pearl Harbor repeating, there are more fundamental reasons for such a rejection. For example, an *expanding* universe, such as the one we live in, violates the Poincaré theorem's assumed condition of a *fixed* volume system. As Professor Eddington put it in a 1934 lecture at Cornell University, "In an expanding space any particular congruence becomes more and more improbable. The expansion of the Universe creates new possibilities of distribution faster than the atoms can work through them, and there is no longer any likelihood of a particular distribution being repeated. If we continue shuffling a pack of cards we are bound sometime to bring them into their standard form — but not if the conditions are that every morning one more card is added to the pack."[56]

An even more direct way to escape Poincaré's theorem is to use a result from general relativity. Using Einstein's theory instead of the classical dynamics that Poincaré used, it has been shown (by Frank Tipler, the inventor of the rotating cylinder time machine spacetime that was mentioned in the previous chapters and which we'll revisit later in the book) that the recurrence theorem is simply no longer true.[57] As Tipler wrote, "In general relativity, singularities intervene to prevent recurrence. General relativistic Universes are thought to begin and end in singularities of infinite spacetime curvature [the Big Bang and the Big Crunch, respectively], and these singularities force time in general relativity to be linear rather than cyclic." A twist to this, however, is that in his analysis Tipler assumed that gravity is *always* attractive, and that the spacetime satisfies a special condition (called the *Cauchy condition* that we'll take-up later) that avoids backward causation. The first assumption is violated in wormhole time machine spacetimes, though, and the second is *by definition* violated in *any* spacetime that supports time travel! So, who knows . . ... ?

---

[52] J. Krueger, "Nietzschean Recurrence as a Cosmological Hypothesis," *Journal of the History of Philosophy*, October 1978, pp. 435-444.

[53] See "Human Repetends" by Marcus Clarke (1846-1881), a story originally published in 1872 and reprinted *Australian Science Fiction* (V. Ikin, editor), Academy Chicago 1984.

[54] For more on this point, see D. W. Theobald, "On the Recurrence of Things Past," *Mind*, January 1976, pp. 107-111.

[55] C. F. Ksanda, "Forever Is Today," *Thrilling Wonder Stories*, Summer 1946.

[56] A. S. Eddington, "The End of the World," in *New Pathways in Science*, Macmillan 1935.

[57] F. J. Tipler, "General Relativity and the Eternal Return," in *Essays in General Relativity* (F. J. Tipler, editor), Academic Press 1980.

Despite all of the previous discussion it is *not* true that the evolution of a system from past to future is *always* accompanied by an increase in entropy—that is, by an irreversible increase in some measure of the system's 'disorder.' Yes, it can be calculated that entropy is *very likely* to monotonically increase in systems of macroscopic size, but that is not the same as certainty. There *can* be fluctuations in the thermodynamic evolution of a system so as to have, at least for a while, a *decrease* in entropy (take another look at Fig. 3.2). All we can say, for sure, is that for macroscopically sized systems even very small fluctuations in increasing entropy are most improbable. To quote no less an authority than the combined genius of Gilbert and Sullivan (from their opera *H. M. S. Pinafore*), here's what we can honestly say of the possibility of failure in the supposed inexorable increase of entropy: "What, never?/No, never!/What, never?/Well, hardly ever." Still, for physicists, entropy is just too useful a concept to give up even though it does not *always* increase with increasing time for an isolated system.

Love it though they may, there are some puzzling aspects to entropy for physicists that remain to this day. For example, the idea that the universe began in some sort of Big Bang process 15 billion years or so ago is the generally accepted view today, The puzzle of that event, one that has been described as literally being a 'fireball explosion,' is that it must have been *fantastically* hot. This means that at the beginning (of everything) there was complete thermodynamic disorder, which from our earlier discussion means *maximum* entropy. Thus, we immediately have the question of how can the entropy of the universe be continuously increasing if it was as large as possible right from the start?[58]

One possible answer is that the proper model of the universe to use is the so-called *inflationary* universe. The inflationary model has a very high expansion rate for the early universe, much higher than the rate in the standard hot Big Bang model. In the standard model, the entropy puzzle occurs because of the ability of all particle processes to readjust rapidly to the ever-changing state of the universe; the so-called *relaxation times* of all particle processes were *short* compared to the expansion rate of the universe. That means that the actual entropy of the universe would, indeed, be the maximum possible at every instant (and so we have the entropy puzzle). In the inflationary model, however, the expansion rate of the early universe was temporarily so high that the relaxation times of particle processes were *very long* compared to the expansion rate. That means the maximum *possible* entropy of the universe, at every instant, would greatly exceed the *actual* entropy. This 'entropy gap' is the cause, then, of the thermodynamic arrow of time, as the universe tries to 'catch-up' and reduce the resulting entropy deficiency.

---

[58] It has been estimated that over the next 10^{116} years the entropy of the universe will increase by a factor in excess of 10^{14}. See S. Frautschi, "Entropy in an Expanding Universe," *Science*, August 13, 1982, pp. 593-599.

There is also a philosophical problem with associating increasing entropy with the flow of time from the past into the future. Events in the past leave traces, artifacts taken to be ordered states—or at least more ordered than are their immediate surroundings. The classic example of this is a footprint in the sand, which is clearly a highly organized structure compared to the surrounding sandy beach. The footprint is the trace of a *past* event; such a trace was all the evidence, for example, that Robinson Crusoe needed to conclude that another human had walked that way. But now consider this famous counter-example,[59] that of a bombed city. Certainly there are traces aplenty of past bombing, and in fact one has to be careful not to trip over or to fall into them! The puzzle, of course, is in trying to argue that random bomb craters, strewn rubble, and crushed buildings, somehow constitute a more organized state (a 'footprint') than did the original city and its surrounding undamaged areas. This fuzziness was captured by one physicist who asked "If it were found that the entropy of the universe were decreasing, would one say that time was flowing backward, or would one say that it is a law of nature that entropy decreases with time?"[60]

For another example of the fuzziness of the relationship between entropy and time, consider the situation[61] of a cloud of non-colliding particles all initially moving toward each other. At first the radius of the smallest sphere that contains the cloud decreases with time but, eventually, as the particles move past another, the radius will grow without bound. Indeed, that inexorable increase of the radius could be taken as defining the direction of time that points toward the future. But in what sense is the *disorder* of the particle cloud increasing? After all, as the cloud expands it 'looks the same' at all times; only its scale (radius) changes. What has entropy to do with this expanding-into-the-future cloud? Perhaps nothing. Perhaps what is need is a *new* arrow of time.

So far we have looked in some detail at two arrows of time: the subjective, psychological feeling we have of time 'flowing,' which has no explanation in physics, and the thermodynamic, statistical quantity of entropy. A third arrow is the so-called *cosmological arrow* of the expansion of the universe. This arrow is not nearly as obvious as the first two. Only in the last century (since the 1920s), as a result of the American astronomer Edwin Hubble (1889–1953), has science become aware that the universe *is* expanding. An interesting speculation about the thermodynamic and cosmological arrows, one made numerous times, is that if the cosmological arrow should ever reverse—that is, if the universe should ever begin to contract toward a Big Crunch—then the thermodynamic arrow would also reverse.[62] The reasoning is that the thermodynamic arrow *follows* the cosmological arrow in an *expanding* universe because that universe can continually 'swallow-up' ever more electromagnetic radiation as it is produced by any physical process. *If* the thermodynamic arrow continues to follow the direction of the cosmological arrow during a contraction, *then* the thermodynamic arrow would also reverse direction.

---

[59] See note 54 in Chapter 2.

[60] P. W. Bridgeman, *Reflections of a Physicist*, Philosophical Library 1955, p. 251.

[61] Taken from K. G. Denbigh, "The Many Faces of Irreversibility," *British Journal for the Philosophy of Science*, December 1989, pp. 501-518.

The usual objection to that suggestion is straightforward. If the direction of time did reverse, then we would see (so goes this argument) all sorts of odd events that would require enormously improbable physics, such as a shattered glass mirror reassembling itself. The error in that objection is subtle but equally simple. *It presupposes the retarded causality of our expanding universe*. In a contracting universe with a reversed thermodynamic arrow of time, however, there would be *advanced* causality, and thus there would be nothing at all improbable about such doings as self-assembling mirrors. As two physicists observed, "The mere reversal of the cosmological expansion will not of itself serve to reverse the direction of thermodynamic and electrodynamic processes, any more than the compression phase of a piston-and-cylinder cycle in a heat engine serves to reduce the entropy of the confined gas."[63]

Those same physicists go on to then mention Stephen Hawking's interest in the relationships among the various temporal arrows. At one time Hawking thought[64] he had discovered a connection between the thermodynamic and cosmological arrows, but then later abandoned that claim.[65] Hawking, in fact, has labeled his original claim "my greatest mistake in science," and has quite openly (and most entertainingly!) discussed his interest in the arrows of time.[66] Indeed, it was to be the subject of his doctoral dissertation but, as he wrote, "I . . . needed something more definite, and less airy fairy than the arrow of time, for my PhD, and I therefore switched to singularities and black holes. They were a lot easier."

Yet another arrow of time is the *electromagnetic arrow*, which refers to the fact that radio waves are observed to only propagate into the future, and never into the past. This is a mysterious fact, because Maxwell's equations for the electromagnetic field, like all the other laws of physics, have no intrinsic time sense. The electromagnetic arrow will be discussed in some detail in Chap. 4.

---

[62] T. Gold, "The Arrow of Time," *American Journal of Physics*, June 1962, pp. 403-410.

[63] P. C. W. Davies and J. Twamley, "Time-Symmetric Cosmology and the Opacity of the Future Light Cone," *Classical and Quantum Gravity*, May 1993, pp. 931-945.

[64] S. Hawking, "Arrow of Time in Cosmology," *Physical Review D*, November 15, 1985, pp. 2489-2495. See also the next paper in the same journal, D. N. Page, "Will Entropy Decrease if the Universe Recollapses?," pp. 2496-2499.

[65] For *why* he abandoned that claim, see S. Hawking *et al.*, "Origin of Time Asymmetry," *Physical Review D*, June 15, 1993, pp. 5342-5356.

[66] S. W. Hawking, "The No Boundary Condition and the Arrow of Time,' in *Physical Origins of Time Asymmetry* (J. J. Halliwell *et al.*, editors), Cambridge University Press 1994.

3.3 Time Dilation

> "Time as we know it is not universally absolute. The rate of its passage depends to a great extent upon the velocity of its observer with regard to some certain reference system. A moving clock will run slower with respect to a selected coordinate system than a stationary one."
>
> —an early science fiction time traveler explains how his time machine works[67]

In this section I'll set the stage for the scientific basis of time travel to the future, as well as for time travel to the past via the warped spacetime called a wormhole. We start by imagining two horizontal, parallel mirrors, one positioned over the other and separated by distance d. The two mirrors are in the same frame of reference with an Observer; that is, the Observer is looking at two mirrors that are stationary with respect to him. Between the two mirrors we further imagine that a particle of light, a photon, is bouncing endlessly back and forth, up and down, in relentless reflection. This simple system is called a *photon clock*, or the Einstein-Langevin clock, after the French physicist Paul Langevin (1872–1946), and it has been part of physics for decades. We define the time required for the photon to travel from one mirror to the other as a *tick* in time, and so the return trip defines the clock's *tock*. The rate of timekeeping measured by the Observer, the time interval separating consecutive ticks, is obviously then given by

$$t' = 2\frac{d}{c}$$

where c is the speed of light.

Suppose we next imagine that the Observer and the photon clock move at constant speed v to the right across *our* line-of-sight. That is, we remain in the original frame while the photon clock and the Observer are now moving at speed v relative to us. This means the photon clock is in a different frame of reference from ours and so *we* do not see the photon bouncing up and down vertically, but rather *we* see the photon tracing out the triangular path shown in Fig. 3.3.

A round trip of the photon evidently now requires more time than before because the distance in the stationary frame (our frame) is greater than the round trip distance in the Observer's frame (moving with the photon clock, he still sees a round trip distance of d). In fact, if t is the time between consecutive ticks as seen by a stationary viewer (us), then the round trip path length of the photon that we see is

$$2\sqrt{d^2 + \left(\frac{vt}{2}\right)^2}$$

---

[67] F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931.

Fig. 3.3 The moving (relative to us) photon clock

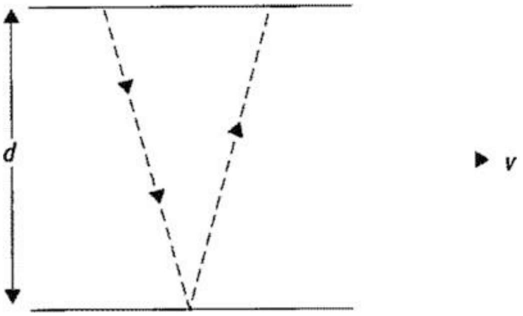

and so

$$t = \frac{2\sqrt{d^2 + \left(\frac{vt}{2}\right)^2}}{c}.$$

This can be easily and quickly manipulated algebraically and combined with the earlier expression for t', the tick interval for an Observer in the same frame as the photon clock, to give the tick interval for the moving clock *as measured by a stationary viewer* (us):

$$t = \frac{t'}{\sqrt{1 - \left(\frac{v}{c}\right)^2}}.$$

Notice that this reduces to $t = t'$ when $v = 0$—that is, when the photon clock is stationary with respect to us.

This last result is the famous Einstein time dilation formula, which shows that $t \geq t'$, and indeed that $t = \infty$ when $v = c$. That is, to us the moving photon clock appears to *run slow* compared to clocks in our stationary frame of reference and, at the speed of light, time *stands still.*[68] A curious anticipation of this association between light and timelessness can be found in a poem by the seventeenth-century poet Henry Vaughn who, in the opening words to his "The World"—which appeared in 1650 as part of his *Silex Scintillans* ("Sparking Flint")—wrote

"I saw Eternity the other night
Like a great *Ring* of pure and endless light,
All calm, as it was bright,
And round beneath it, Time in hours, days, years
Driven by the spheres
Like a vast shadow moved, in which the world
And all her train were hurled."

---

[68] For $v > c$ the time dilation formula says that time becomes imaginary, and this is one reason for claiming that $v > c$ is not possible. The time dilation formula has been experimentally verified: see H. E. Ives and G. R. Stilwell, "An Experimental Study of the Rate of a Moving Atomic Clock," *Journal of the Optical Society of America*, July 1938, pp. 215-226.

A similar modification in the length of a moving object (measured in the direction of motion) occurs when $v > 0$. While an Observer moving with the object will measure its length to be L', a stationary viewer will 'report' it to be contracted to the length

$$L = L'\sqrt{1 - \left(\frac{v}{c}\right)^2}.$$

This effect is called the *Lorentz-FitzGerald contraction*.[69] For 'everyday' objects and speeds the contraction effect is an extremely small one. For example, for a low-altitude satellite 100 m long, moving at 18,000 miles per hour (that is, at $v = 2.7 \times 10^{-5}c$), the contraction is less than 4×10^{-6} cm.

In the early days of science fiction the contraction effect was fascinating to readers, but authors often got it wrong. For example, in one story[70] of a runaway spaceship falling into the Sun, we read "When our racing [ship] was drawn from the Earth's gravity and fell at ever increasing speed toward the Sun it soon approached the speed of light. As we fell faster and faster our length in the direction of the Sun progressed into nothingness. Then — it reached the speed of light — passed it. Now — mind you this — when the [ship] attained the speed of light it was of a *minus* length." This author has managed to make four errors in three sentences!

The same author botched the Lorentz-FitzGerald contraction again 5 years later, and added yet more errors to his growing list. In that tale[71] there is an episode of faster-than-light radio communication along with a lengthy, unfortunate dissertation that actually denies special relativity's fundamental assertion that all inertial frames of references are indistinguishable from each other (two frames are *inertial* if they have no relative acceleration—I'll say more on this in the next section). And in yet another story[72] of high-speed space travel the author has the contraction working in the wrong direction—as the rocket ship moves faster and faster it gets longer and longer.

As bad as those errors are, first prize for mangling the laws of physics has to go to the story[73] of a near light-speed spaceship on its way to Alpha Centauri. The crew mutinies and puts the captain and first office 'overboard' (think *Mutiny on the Bounty*) with 6 months' worth of provisions. This happens at mid-voyage, about 2 light-years from both home and destination, so matters look grim. Indeed, the author tells his readers, several times, that things look *very* bad. But are they? With a stated speed of 162,000 miles per second, the time dilation factor is slightly more than 2 and so, because the space boat is traveling at $0.87c$, it will take a little more

---

[69] Named after the Dutch physicist H. A. Lorentz (1853-1928) and the Irish physicist G. F. FitzGerald (1851-1901).

[70] J. H. Haggard, "Faster Than Light," *Wonder Stories*, October 1930.

[71] J. H. Haggard, "Relativity to the Rescue," *Amazing Stories*, April 1935.

[72] D. Wandrei, "A Race Through Time," *Astounding Stories*, October 1933.

[73] N. Schachner, "Reverse Universe," *Astounding Stories*, June 1936.

Table 3.1 The Lorentz-FitzGerald time slowing factor

1	$\frac{1}{\sqrt{1-\left(\frac{v}{c}\right)^2}}$
.1	1.005
.2	1.021
.5	1.155
.7	1.4
.9	2.294
.999	22.366
.9999	70.712

than 13 months of *space boat time* to complete the journey. If the men go on half-rations then it seems they *could* survive.

There is, of course, the problem of slowing down so as to arrive at Alpha Centauri at a reasonable speed, but that issue is ignored in the story. Instead, our attention is directed to the much more dramatic concern of a faster-than-light planet (don't ask!) colliding with the space boat and carrying the castaways onwards toward their destination. When this happens we read that time runs backwards (for what *really* occurs at superluminal speeds, keep reading this chapter) and, finally, in a repeat of an error I mentioned earlier, we are told that the Lorentz-FitzGerald contraction is negative for $v > c$.

The time-slowing (or size-shrinking) factor becomes pronounced only at values of v close to c, as shown in Table 3.1. For example, the last entry shows that a clock traveling at 99.99 % the speed of light will register the passage of 1 year while nearly 71 years pass on Earth. One science fiction writer got this dramatically wrong, even though he actually reproduced the Lorentz-FitzGerald equation in his story.[74] At one point he writes of the near light-speed rocket ship that stars in the tale, "If it [the ship's speed] was as slow as ninety-four percent [of the speed of light] . . . for every moment ticked by the clocks of the [ship] hundreds passed on earth." In fact, the time dilation factor at that speed is 'only' 2.93.

One possible objection to time dilation is that the analysis done here has been for a *particular* clock. How do we know that another clock, one using wheels and pendulums, for example, instead of photons and mirrors, wouldn't be affected differently by motion? The answer comes from relativity itself, which says there is no way to detect uniform motion. If two clocks did behave differently, then this difference could be used as a motion detector. Since this is impossible within the framework of relativity, then *all* clocks, no matter what the details of their internal mechanisms may be (including the *biological* clocks of own bodies), must respond to motion just as does the photon clock.[75]

---

[74] L. R. Hubbard, "To the Stars," *Astounding Science Fiction*, February and March 1950.

Time dilation can also be caused by gravity (it appears in the 2014 movie *Interstellar*), and that effect has been used to 'construct' a time machine from a wormhole (to be discussed later). You can get a qualitative understanding of how that happens by imagining a massive body (*massive*, to have a really big gravitational field) in space, on the surface of which is a hot object. That object emits electromagnetic radiation and, though it isn't essential to the following argument, further imagine that the temperature of the object is sufficiently high that some of the radiation (emitted by the very atoms of the object) is in the visible-light portion of the spectrum. Now, from elementary quantum theory we can also think of the object's atoms as emitting photons ('particles of light'), each of energy hf, where h is Planck's constant and f is the frequency in hertz (what used to be called 'cycles per second'). The higher the temperature, the higher the photon energy, and so the higher the frequency. In the visible spectrum, f is on the order of 10^{15} Hz, a frequency one *billion* times higher than commercial AM radio frequencies.

The radiating atoms can be thought of as tiny clocks, with alternate half-cycles of radiation being ticks and the half-cycles in-between being the tocks. The passage of time on the surface of the massive body can be measured by these atomic clocks in the hot, radiating object. To a distant observer, however, as she receives the photons from the hot object, the passage of surface time on the massive body will appear to occur at a reduced rate when compared with the photons emitted by her own identically hot object (her 'local' clock). That's because the radiation that arrives at the distant observer has traversed a gravitational field (a journey sometimes described as 'climbing out of a gravitational well') and so is down-shifted in frequency toward the red end of the visible spectrum. This effect is called either the *gravitational red shift* or the *gravitational time dilation* effect (or even the *Einstein shift*, because it was Einstein who predicted the effect in 1907).

You can 'understand' this dilation effect as follows. One can crudely think of a photon emitted by the hot object as something like a rock thrown upward. As the rock rises upward through the gravitational field, its *total* instantaneous energy is always constant, but the total, fixed energy is split between its kinetic and potential energies in an ever changing way. That is, as the rock rises, its kinetic energy continually decreases (the rock slows down), whereas its potential energy continually increases. A photon is not a rock, however, and it certainly can't slow down as

---

[75] Resistance to this conclusion persisted for years. See, for example, the letter "Relativity and Radio-activity," *Nature*, January 8, 1920, p. 468. The author of that letter wondered whether a clock based on radioactive decay might not somehow beat the 'conspiracy' of moving clocks running slow compared to stationary ones. And in a letter to *Science* (December 7, 1962, p. 1180), a reader objected to applying the laws of physics to biological systems, first asserting (incorrectly) that time dilation "has never been proved or disproved experimentally," and then "there is no known causal means by which greatly increased velocity could alter, without destroying the very biochemical basis of the life process, the metabolic changes which are responsible for the aging process."

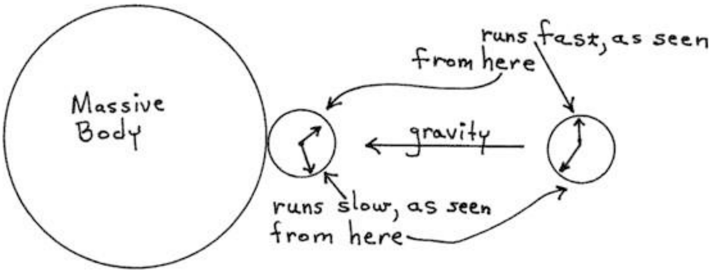

Fig. 3.4 Gravitational time dilation due to a massive body

it rises through a gravitational field (it *always* moves at the speed of light because the photon *is* light). The only way a photon can give up energy to balance the ever increasing potential energy (physicists will cringe at this, but read on) is to decrease its frequency. Hence, the red shift as seen by a distant observer of the photon, who thus sees time running slow on the massive body. Fig. 3.4 shows the case for a clock on a massive body, compared to a distant clock. Notice, carefully, the 'direction of gravity,' that is, the direction a small, unrestrained test mass will move.

A gravitational red shift in the *opposite* direction is nicely described in the famous 1966 science fiction story "Neutron Star" by Larry Niven. There a space traveler zooms down into a neutron star's intense gravity field (at half the speed of light!), passing within one mile of the star's surface. He reports what he observes in these dramatic words: "All around me were blue-white stars. Imagine light falling into a savagely steep gravitational well. It won't accelerate. Light can't move faster than light. But it can gain in energy by increasing its frequency. The light was falling on me, harder and harder, as I dropped." To Niven's intrepid spaceman, therefore, the passage of time on those distant blue-white stars appeared to be running *fast* compared to his wrist watch. This shows that the effect could equally well be called the 'gravitational *blue* shift.'

Notice that *gravitational*-induced time alterations do *not* have the symmetrical feature of motion-induced time dilations.[76] That is, for gravitational time dilations caused by photons either falling into or climbing out of gravity wells, observers at each end *agree* about whose clock is running slow, unlike in the motion-induced case where *each* of the relatively moving observers thinks it is the *other* observer's clock that is running *slow*.[77]

Now, comparing a massless photon to a rock which does have mass (and so potential energy), as each travels 'against' gravity, *is* straining the physics, with its one virtue being the provision of an initial plausibility argument. Gravitational time dilation is sufficiently important in the operation of wormhole time machines,

---

[76] Gravitational time dilation was experimentally observed in 1960, more than half a century after Einstein predicted it.

[77] A science fiction use of both the red and the blue gravitational shifts appears in the novel by J. P. Hogan, *Out of Time*, Bantam 1993.

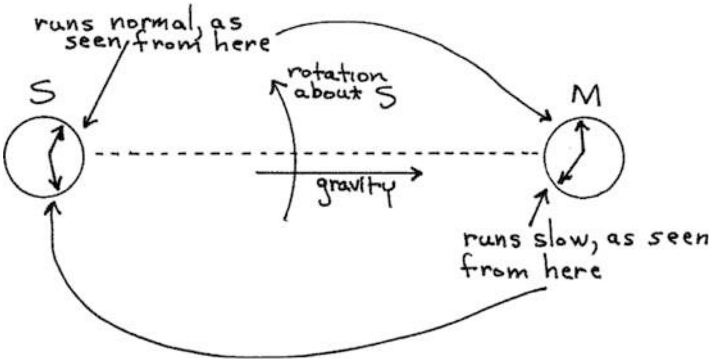

Fig. 3.5 Gravitational time dilation due to rotation

however, that perhaps another way to think about it that is more acceptable to hardcore physicists makes it worth another look. Imagine a turn-table disk that initially is *not* rotating (soon it *will* spin). On this disk imagine further that we fasten two clocks, one called S (for *stationary*) and one called M (for *moving*), as shown in Fig. 3.5. These two clocks are set to read the same time at some instant and then, thereafter, they tick-tock through time at precisely the same rate. We then start the disk rotating around a vertical axis through clock S. (Imagine S to be a *point* clock, and so it is *the* one point on the turn-table that remains at rest even as the disk rotates.) What happens to the time-keeping of S and M?

We can answer this question by using our earlier result concerning time dilation due to motion. Even with the disk now rotating, S appears stationary to an observer sitting on top of M, while M appears to be moving across the line of sight of an observer sitting on top of S. So, to the observer at S, clock M runs slow, while to the observer at M there is *no change* in the time keeping of clock S. This is a non-symmetrical outcome, and so should remind you of gravitational time dilation. This might be a puzzle to you, however, as we don't have a massive body in Fig. 3.5 to account for a gravity presence. This is where the genius of Einstein comes into play.

Anyone who has ever ridden on a merry-go-round knows there is an outward (pointing away from S) directed force called the *centrifugal force* that is 'trying' to toss you off the merry-go-round. Now, where there is a force there is an acceleration, and one of Einstein's starting points in his development of general relativity was to identify an acceleration, *whatever its origin*, with gravity. A massive body is, of course, one possible origin (the obvious one, in fact), but so is the rotation of the turn-table. So, we have the situation shown in Fig. 3.5, where now 'gravity' is directed as shown and, and as in Fig. 3.4, the direction of acceleration of gravity is toward the slow-running clock. Again, the direction of gravity is the direction a small, unrestrained test mass on the rotating disk will move. The further M is away from S, the greater the 'gravity' of the centrifugal acceleration and so the slower will M run as measured by the observer at S.

A 1968 story that uses the gravitational time dilation effect in a striking fashion tells of a starship's visit to a supernova, accompanied by a fantastic alien life-form—a ball of intelligent plasma named Lucifer that is telepathic.[78] While the ship stands off at a distance of 500 million kilometers, Lucifer will approach much closer to the *event horizon* (see the Glossary) of a black hole at the center of the supernova explosion and communicate its findings to a human telepath on the ship. A physicist in the crew is curious about one point, and asks the human telepath the following question:

"I have wondered about one item. Presumably Lucifer will go quite near the supernova. Can you still maintain contact with him? The time dilation effect, will that not change the frequency of his thoughts too much?"

Lucifer, in fact, dies in the black hole even as he saves the ship from destruction, and the human telepath will hear his death scream for the rest of her life. As the physicist later explains to the ship's captain, telepathy is instantaneous and has no limiting range (there is no known physical basis for believing any of this, but it is crucial for story effect):

> "Remember the time dilation. He fell from the sky and perished swiftly, yes.
>
> But in supernova time. Not the same as ours. To us, the final stellar collapse takes an infinite number of years. . . . He will always be with her."[79]

3.4 The Lorentz Transformation

> "If only he'd paid more attention to mathematics in school."
>
> —a science fiction time traveler laments missed opportunities[80]

In this (and the next) section the math gets about as 'deep' as it gets in this book, but to leave it out struck me as a cheat. *You* can skip part (or all) of the math and simply read the prose, but it seemed unfair for me to make that decision for you.

We begin by imagining two distinct frames of reference. One we take to be stationary, and the other as moving at a uniform speed v with the respect to the first. The moving frame is said to be *boosted* with respect to the stationary frame. We orient these two coordinate systems so that the motion occurs along just one axis (the x-axis, as shown in Fig. 3.6, where I am using primed variables for the moving frame). That is, the two frames have coincident x axes, and parallel y and z axes that

---

[78] Poul Anderson, "Kyrie," in *The Road to Science Fiction* (J. Gunn, editor), volume 3, New American Library 1979.

[79] A mathematical discussion of how signals take forever (even though they are emitted in a finite time interval) to travel from the event horizon of a black hole to a distant receiver can be found in James B. Hartle, *Gravity: an introduction to Einstein's general relativity*, Addison Wesley 2003, pp. 264-268.

[80] D. Knight, "Extempore," in *Far Out*, Simon and Schuster 1961. Similar words ("If only I had more mathematics") were spoken by Einstein the day before he died — see Walter Isaacson, *Einstein: his life and universe*, Simon & Schuster 2007, p. 542.

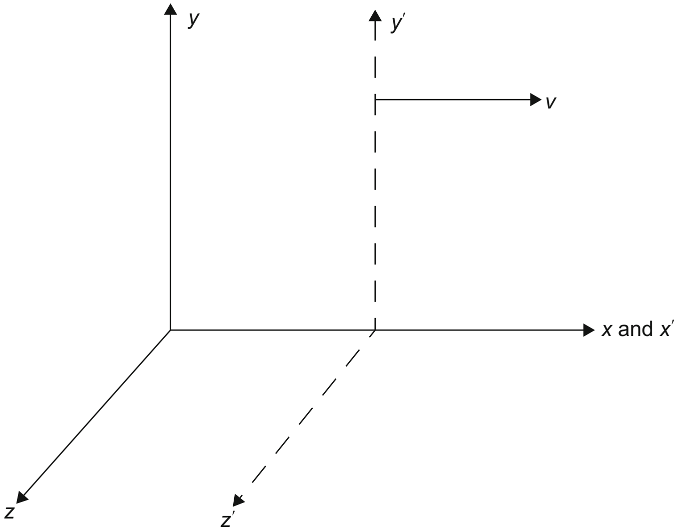

Fig. 3.6 Two reference frames in relative motion

are moving apart at the constant speed v. Let's also imagine that there is a clock at the origin of each frame, and that at the instant the origins match the clocks are synchronized; that is, $t = t' = 0$ is the instant the two coordinate systems coincide.

We further imagine that there is an observer at the origin in each frame. At some arbitrary instant of time, each observer records the coordinates of the arbitrary point P in space, as measured in his system. These observers could, for example, agree to record the coordinates of P in their system when their clock reads 5 s. It seems immediately obvious (as it was for Newton) that $t = t'$*;* that is, time runs at the same rate in each frame and thus it makes sense to talk about "the same instant" at every point in space. (After reading the previous section you know this *not* true, but temporally forget that!) Thus, at this 'same instant' the stationary observer records (x, y, z) and the moving observer records (x', y', z'). What are the relationships between the primed and unprimed coordinates of P? That is, what mathematical transformation converts from one frame to the other?

The answer seems obvious:

$$y' = y$$
$$z' = z$$
$$x' = x - vt.$$

This transformation, called the *Galilean transformation* after the Italian Galileo Galilei (1564–1642), satisfies the relativity principle, which says that uniform motion leaves the laws of physics unchanged. For example, in the stationary system Newton's famous second law of motion for a constant mass m,

$$F = m\frac{d^2x}{dt^2}$$

becomes the identical-appearing form

$$F' = m\frac{d^2x'}{dt'^2}.$$

More precisely, *all* the laws of mechanics known to Newton are unchanged. Any frame of reference in which Newton's laws of mechanics hold true is said to be an *inertial* frame. Given one inertial frame, we can find infinitely many others simply by applying the Galilean transformation.

However, when the mathematical laws of electrodynamics were discovered by Maxwell in the nineteenth century, it was a shock to physicists to learn that the Galilean transformation does *not* leave Maxwell's equations unchanged in form; the transformed equations predict electromagnetic effects for the moving system that are *not* predicted to occur in the stationary system. This meant that there was theoretical support for the possibility that electromagnetic experiments might be devised to detect uniform motion, and this eventually led to the famous Michelson-Morley experiment of 1887. This experiment, sensitive enough to detect the motion of the Earth itself through space, failed to detect any such motion. The conclusion was clear: the new electromagnetic effects predicted by the Galilean transformation do not exist, and so the transformation must be wrong even though it works for the laws of mechanics. So—what is going on?

The answer is inspired, and again returns us to the cornerstone of relativity: the idea that the laws of physics, *all* the laws, should look the same to observers in uniform relative motion. That is, there is no special or preferred system of coordinates—*all* inertial systems are equivalent in physics. Evidence from an extremely broad variety of sensitive experiments had, by the end of the nineteenth century, convinced physicists that Maxwell's equations *are* correct. Thus, a new transformation was needed that leaves both the laws of mechanics *and* the laws of electrodynamics unchanged with uniform motion. But, a single transformation that works on Maxwell's equations and on the mechanical laws would therefore mean that Newton's mechanical laws as stated cannot be correct, and this was a breathtaking conclusion: Newton had been unchallenged for two centuries.

As it turns out, Newton's laws are *almost* right. The only correction required is that the mass of a moving body is not independent of motion, but rather varies as

$$m = \frac{m_0}{\sqrt{1 - (v/c)^2}}$$

where m_0 is the so-called *rest mass* when $v = 0$.[81] This result says that m is infinite at $v = \mathrm{c}$ unless $m_0 = 0$ (as it is for a photon), which is another reason for the belief that accelerating a mass (such as a spaceship) up to the speed of light is impossible because it would require infinite energy (look back at note 68, too). With this modification, the transformation that leaves all the laws of physics unaltered in form by uniform motion is what is called the *Lorentz transformation* (after the same Lorentz the contraction effect is named for), who discovered it in 1904 by direct manipulation of Maxwell's equations:

$$
\begin{aligned}
y' &= y \\
z' &= z \\
x' &= \frac{x - \mathrm{vt}}{\sqrt{1 - (v/c)^2}} \\
t' &= \frac{t - \mathrm{vx}/c^2}{\sqrt{1 - (v/c)^2}}.
\end{aligned}
$$

In 1905 Einstein discovered how to derive these equations from a fundamental reexamination of space and time without concerning oneself about the details of specific physical laws.

By simple algebraic manipulation, the transformation equations can be rewritten as

$$
\begin{aligned}
ct &= \gamma ct' + \beta\gamma x' \\
x &= \beta\gamma ct' + \gamma x'
\end{aligned}
$$

where

$$\beta = \frac{v}{c}$$

and

$$\gamma = 1/\left(1 - \beta^2\right)$$

are dimensionless constants. In compact matrix form, the Lorentz transformation becomes

---

[81] This variation of mass with speed was experimentally observed in 1901.

$$\begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{bmatrix} \begin{bmatrix} ct' \\ x' \end{bmatrix}$$

and the symmetrical 2 × 2 matrix is called the Lorentz boost matrix or simply the *boost*. Notice that when $v = 0$ (zero boost) the boost matrix reduces to the identity matrix; that is, the two frames are one and the same with at most a shift in the location of the origins. Note, too, that $\beta = 0$ and $\gamma = 1$ for *any* v when c is infinite—that is, the boost matrix is again reduced to the identity matrix and the Lorentz transformation becomes the Galilean *if* c is infinite. But c is not infinite, and all the implications of special relativity are the direct result of the *finite* speed of light.

The Lorentz transformation contains two results I have mentioned earlier in the book. For example, in Chap. 2 it was mentioned that simultaneity is a relative concept in reference frames in relative motion. Let's see what the transformation says about that. Consider two events that occur specifically on the x-axis. They are simultaneous in the stationary system (at, say, time $t = T$) but are at different places (at, say, $x = X$ and $x = X + \Delta X$). Their occurrences in time for the moving observer are

$$t_1' = \frac{T - \mathrm{vX}/c^2}{\sqrt{1 - (v/c)^2}}$$

and

$$t_2' = \frac{T - \mathrm{v}(X + \Delta X)/c^2}{\sqrt{1 - (v/c)^2}}.$$

For the moving observer, therefore, the two events are *not* simultaneous, being separated in time by

$$t_1' - t_2' = \frac{v\Delta X/\mathrm{c}^2}{\sqrt{1 - (\mathrm{v}/\mathrm{c})^2}}.$$

Only if $\Delta X = 0$ (the two events occur at the same place) will $t_1' = t_2'$. That is, only if $\Delta X = 0$ are simultaneous events in one frame also simultaneous in another frame in relative motion.

And in the previous section we found that time runs slow in one frame as observed from another frame that is in relative motion. We can get this result from the t' equation of the Lorentz transformation by differentiating it with respect to t. Thus,

$$\frac{dt'}{dt} = \frac{1 - (v/c^2)\frac{dx}{dt}}{\sqrt{1 - (v/c)^2}}.$$

But since

$$\frac{dx}{dt} = v,$$

the speed of the moving frame as measured by the observer in the stationary frame, this gives

$$dt' = \sqrt{1 - (v/c)^2}\,\mathrm{dt}$$

which is the same result we obtained by analyzing the photon clock.

The Lorentz transformation contains other interesting implications beyond these. For example, mention has been made several times to the ‘relativity principle,’ the belief that uniform motion has no observable effect on the forms of physical laws. But how do we know who is moving and who is stationary? After all, a system moving to the right past a stationary system could just as well be thought as the stationary system, while it’s the *other* system that is moving to the *left* (at speed $-v$).

To study this question with the Lorentz transformation, we’ll invert the transformation (that is, solve for the unprimed variables in terms of the primed ones). What we get back is just what you probably thought—the Lorentz transformation with v replaced by $-v$. That is, the Lorentz transformation is symmetrical, so two observers in different frames of reference each say it is the *other’s* clock that is running slow! This follows immediately, in fact, from the original transformation written in matrix form. That is, multiplying through the earlier matrix equation by the inverse of the boost matrix, we get

$$\begin{bmatrix} ct' \\ x' \end{bmatrix} = \begin{bmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{bmatrix}^{-1} \begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix}.$$

The only difference between the original boost matrix and its inverse (which is, of course, the new boost matrix for the new interpretation of which frame is the moving one) is a change in sign for β, that is, in the sign of v. The inverse transformation is

$$\begin{aligned} y &= y' \\ z &= z' \\ x &= \frac{\mathrm{x}' + \mathrm{vt}'}{\sqrt{1 - (\mathrm{v}/\mathrm{c})^2}} \\ t &= \frac{\mathrm{t}' + \mathrm{vx}'/\mathrm{c}^2}{\sqrt{1 - (\mathrm{v}/\mathrm{c})^2}}. \end{aligned}$$

As a final example of what the Lorentz transformation tells us, consider the so-called *addition of velocities* problem. Suppose you are in a high-speed spaceship traveling past Earth at speed v. Earth is the stationary system (with the unprimed variables), and the spaceship is the moving system (with the primed variables). Assume the x and x' axes are along the direction of motion. Imagine next that while standing in the nose of the spaceship, just as the spaceship passes Earth, you fire a gun in the direction of motion (away from the Earth), with the bullet exiting the gun with a muzzle speed of w. How fast is the bullet moving away from Earth? The common-sense answer in Galileo's time was $v+\mathrm{w}$, but we now know that the Galilean transformation is wrong. What does the Lorentz transformation say?

Inside the spaceship, the position of the bullet at time t' after the gun is fired is

$$x' = wt'.$$

From the inverse Lorentz transformation, the location of the bullet earth's frame is

$$x = \frac{x' + \mathrm{v}t'}{\sqrt{1-(v/c)^2}} = \frac{w+\mathrm{v}}{\sqrt{1-(v/c)^2}}t'.$$

The transformation also tells us that (using $x' = wt'$)

$$t = \frac{t' + \mathrm{v}x'/c^2}{\sqrt{1-(v/c)^2}} = \frac{1+\mathrm{wv}/c^2}{\sqrt{1-(v/c)^2}}t'.$$

Thus, the speed of the bullet in *Earth's frame* is

$$\frac{x}{t} = \frac{w+\mathrm{v}}{1+\mathrm{wv}/c^2}.$$

Notice that for a low-speed bullet ($w << c$) this result[82] is close to $w + v$, but at high values for w the result is very much different. Indeed, suppose we don't fire a gun at all, but rather replace it with a flashlight. Now, instead of a bullet, we shoot *photons* at $w = c$. The Galilean transformation would (incorrectly) say that a stationary observer on Earth would see the photons moving away at speed $v+\mathrm{c}$, which is a superluminal speed. The Lorentz transformation, however, says that the Earth observer would see a speed of

---

[82] This result was found by the French physicist Henri Poincaré (1854-1912) in June 1905, three months *before* the publication of Einstein's special theory of relativity which also contains the result. And it was Poincaré who first stated (in 1904) that "no velocity can surpass that of light, any more than any temperature could fall below the zero absolute."

$$\frac{c+\mathrm{v}}{1+\mathrm{cv}/c^2}=\frac{c^2(c+\mathrm{v})}{c^2+\mathrm{cv}}=\frac{c^2(c+\mathrm{v})}{c(c+\mathrm{v})}=c.$$

That is, no matter what the speed of the moving observer on the spaceship is, he sees the light from his flashlight traveling at the same speed as does the stationary observer back on Earth. This peculiar effect is unique to the speed of light ($w=\mathrm{c}$). We've derived it here as a consequence of the Lorentz transformation, but in fact Einstein actually did things in reverse order. That is, he began by *postulating* the invariance of the speed of light[83] for all observers in uniform motion, combined that with the principle of relativity which says all physical laws look the same to those observers, and so derived the Lorentz transformation using no mathematics beyond high school algebra.

A mathematically elegant alternative derivation of the addition-of-velocities formula can be done by simply noticing that the condition of two successive boosts should be, itself, a boost. Thus, if we have a frame moving relative to a second frame (which is itself moving relative to a third frame), then the boost matrix of the first frame relative to the third frame is the product of the two individual boost matrices. That is,

$$\begin{bmatrix}\gamma_3 & \beta_3\gamma_3\\ \beta_3\gamma_3 & \gamma_3\end{bmatrix}=\begin{bmatrix}\gamma_2 & \beta_2\gamma_2\\ \beta_2\gamma_2 & \gamma_2\end{bmatrix}\begin{bmatrix}\gamma_1 & \beta_1\gamma_1\\ \beta_1\gamma_1 & \gamma_1\end{bmatrix}.$$

From this it is easy to show (if you know how to multiply matrices!) that

$$\beta_3=\frac{\beta_1+\beta_2}{1+\beta_1\beta_2}.$$

Substitution of

$$\beta_1=\frac{w}{c},\ \beta_2=\frac{v}{c}$$

immediately gives the addition-of-velocities formula.

A failure to understand the implications of the invariance of the speed of light resulted in two stupendous errors in the story "To the Stars," cited in note 74. At all times an officer stands watch on the bridge of a near light-speed rocket ship to be sure the ship doesn't accidently *reach* the speed of light. This is to be avoided (according to the author) because to reach the speed of light would cause the ship to

---

[83] Einstein's postulate was experimentally confirmed in 1932.

"hang there forever unmoving [in time] . . . locked, protected and condemned to eternity by zero time." This horrible state is so easy to stumble into (the author was apparently unaware that it would require *infinite* energy) that occasionally the ship has to fire a 'check-blast' from its forward rocket tubes to slow down! Equally absurd is the means by which the development of this 'fatal' condition is detected: the nose of the ship mounts a forward-pointing light source (our earlier flashlight) and so, if the ship is getting too near the speed of light, it will start to *overtake* the photons emitted by that source!

The biggest puzzle of all, actually, is why the editor of *Astounding Science Fiction* let such a technically goofy story appear in a magazine recognized for its usual faithfulness to known science. Particularly so since the story appeared in 1950 and, as long ago as December 1937, none other than Isaac Asimov (then 17 years old) had written a letter to the editor[84] giving the proper interpretation of what happens when $v = c$. Here's what the young Asimov wrote (notice the early hint of his life-long pessimism concerning time travel to the past that appears near the end of his letter):

> "The effect on time of increasing speeds is . . . well known. Relativity states that as speed approaches that of light, time slows up until at 186,000 miles a second, time (so to speak) stands still. This seems to refute statements found in so many astronomy books (and science fiction stories) that even at the speed of light it would take four years to reach the nearest star. No such thing! As time halts at the speed of light, a person traveling from Alpha Centauri to the solar system, or vice-versa, would not be aware of any lapse of time. In that sense the speed of light is infinite (as was thought in ancient times). This by the way offers an entirely scientific (if impractical) means of travel into the future. Say that someone wants to see how the world would look a hundred years from now. His procedure would be as follows: getting into his spaceship, he would proceed to a spot fifty light-years away at the speed of light. The journey would, for him, be practically instantaneous (due to the curious behavior of time at the speed of light). But fifty years would have elapsed on earth. He makes the return trip at the same speed. Another fifty years lapse on earth and he lands a hundred years after his time. With this system, however, it would be impossible to travel into the past, so I don't think it will ever be adopted."

Young Asimov missed an important detail concerning the reversal of the spaceship's direction of travel for the return to Earth, and I'll come back to it later in this chapter. But certainly he displayed a *far* better knowledge of the physics of time than did the author of "To the Stars" (who was, by the way, L. Ron Hubbard, a prolific writer of fantasy and science fiction before founding the Church of Scientology).

---

[84] The editor who bought "To the Stars" was the same editor editor of *Astounding* when Asimov's letter appeared, and so he was certainly aware of it.

3.5 Spacetime Diagrams, Light Cones, Metrics, and Invariant Intervals

> "Come back when you know tensor calculus and I'll explain to you about n-dimensional forces and the warping of world-lines."
>
> —a science fiction physicist's reply after being asked how his time machine works[85]

It is helpful in discussions about the spacetime of special relativity to use what are called *Minkowski spacetime diagrams*. These are plots of the spacetime coordinates of a particle; the resulting curve is called the *world line* of the particle. Such diagrams are four-dimensional—three space axes and one time axis—and hard to visualize, much less draw on a flat sheet of paper! The convention is to make do, whenever possible, with a simplified spacetime that has just one space axis (horizontal) and one time axis (vertical). As you'll recall from Chap. 1, physicists often call such a simplified diagram a *toy spacetime*.

So, for a particle at rest in some observer's frame of reference, its spacetime diagram for that observer is a *vertical* world line. If the particle is not at rest then its world line will tilt away from the vertical; the greater the speed the greater the deviation from the vertical. Accelerated particles will have world-lines that *curve* away from the vertical. Straight, uncurved world lines represent unaccelerated particles, that is, particles experiencing no forces and so in free fall. Such a world line is called a *geodesic*. In Fig. 3.7 the world lines for these various cases are shown on the same axes. It is assumed in the figure that all three particles are at $x = x_0$ when $t = 0$.

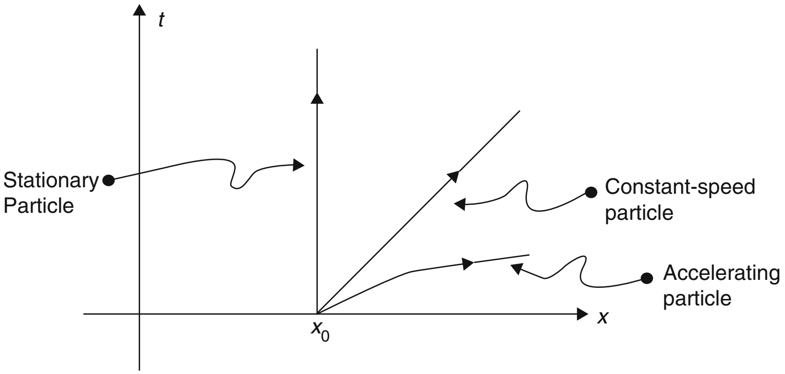

Fig. 3.7 World lines of three particles

---

[85] Poul Anderson, "The Little Monster," in *Science Fiction Adventure from WAY OUT* (R. Elwood, editor), Whitman 1973.

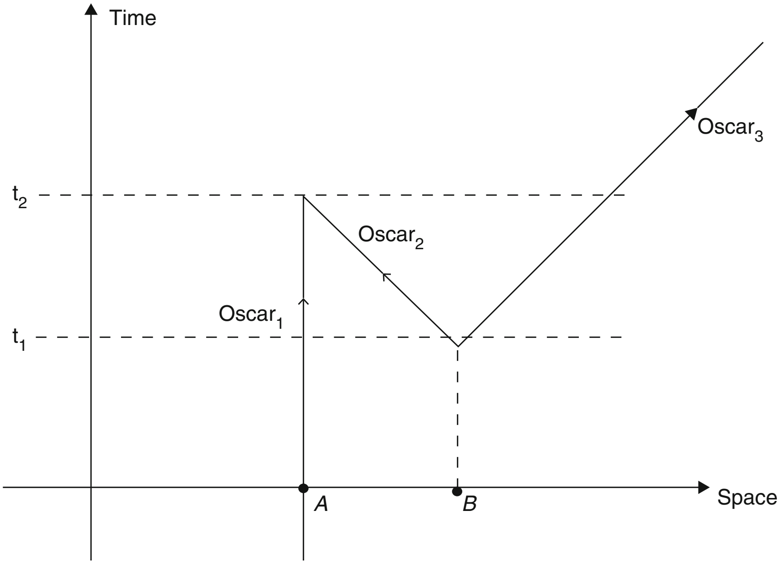

Fig. 3.8 A time traveler and his world line

Spacetime diagrams were embraced decades ago by philosophers looking for ‘scientific’ ways to support their position on time travel (whatever it might be), as opposed to the mere verbiage of traditional colleagues. A famous example of this is a 1962 paper by the Harvard philosophy professor Hillary Putnam (note 74 in Chap. 1). There we are asked to imagine the spacetime diagram of one Oscar Smith who, in Fig. 3.8, is at spatial location A next to his time machine. At time t_0 Oscar has not yet gotten into his time machine. A little later, at time t_1, we suddenly see not only Oscar at A but also *two more* Oscars who have appeared (apparently out of thin air) moving away from spatial location B! Between t_1 and t_2 we see the original Oscar at A and the two mysterious Oscars at B (for a total of three Oscars, labeled in the figure as Oscar$_1$, Oscar$_2$, and Oscar$_3$) move forward in time—but one of the new Oscars (Oscar$_2$) lives a decidedly odd existence in that his life seems to be running in reverse!

Eventually, at time t_2, the original Oscar$_1$ and the weird, reverse Oscar$_2$, merge and seemingly annihilate one another, vanishing into thin air to leave only a single Oscar (Oscar$_3$) for all time after t_2. Putnam argues that, although strange, what has just been described is still sensible and that, indeed, the very fact that we can draw the spacetime diagram of Fig. 3.8 supports the case for backward time travel. He claims this because although the spacetime diagram does show time increasing upward for all three Oscars (that is the time direction for an external observer) there is actually no ‘spontaneous creation’ or ‘mutual annihilation’ and all is sensible *if* Oscar$_2$ is understood actually to be a time traveler into the past with *his* time direction thus pointed opposite to that of the ‘other two’ Oscars. There is, of course, just *one* Oscar!

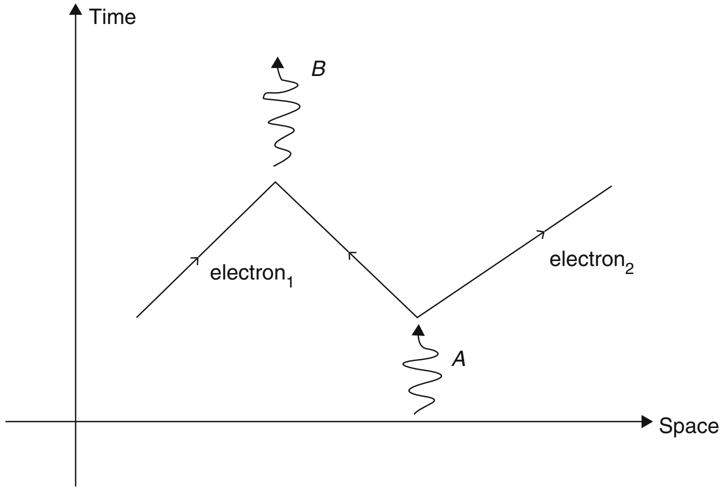

Fig. 3.9 Anti-matter via backward time travel

As mentioned back in Chap. 1, Putnam's suggestion was rebutted (see that chapter's note 76) by what is (in my opinion) an even less plausible mechanism for Putnam's kinked spacetime than is Putnam's invocation of time travel. That critic advocated, instead, an explanation based on matter transmitters and anti-matter, the latter an idea he credits to Feynman (who actually got it from Wheeler). In a paper by Feynman we do find the famous suggestion that a positron that appears to us to be moving forward in time is actually an electron traveling backward through time.[86] Logically, that greatly weakens the critic's view, because it puts him in the position of using anti-matter (explained in terms of backward time travel) to argue *against* backward time travel! But let's ignore that concern, give the critic the benefit of the doubt, and explore how anti-matter and backward time travel are imagined to be connected.

Feynman asks us to imagine the process shown in Fig. 3.9. Gamma ray A spontaneously creates an electron-positron pair, with electron$_2$ moving off to some distant region while the positron soon meets with electron$_1$, resulting in mutual annihilation and the production of gamma ray B. This description involves *three* particles, and each segment of the kinked line is a distinct particle. But Feynman said there is another way to look at this, a way that involves just *one* particle. According to Feynman, the kinked line in Fig. 3.9 (which should remind you of Oscar's kinked world line) is the world line of a single electron; the middle

---

[86] R. Feynman, "The Theory of Positrons," *Physical Review*, September 15, 1949, pp. 749-759. In a paper published the year before ("A Relativistic Cut-Off for Classical Electrodynamics," *Physical Review*, October 1948, pp. 939-946), he wrote "This idea that positrons might be electrons with the proper time reversed was suggested to me by Professor J. A. Wheeler." The identification of anti-matter with backward time travel occurred in science fiction (see note 124 in Chapter 1) almost simultaneously with Wheeler's speculation.

segment, that we call a positron, is just the electron traveling backward in time, and so we must reverse the arrowhead on it (indicating the opposite of the direction shown in Fig. 3.9).

There are two central questions at this point. First, why is a positron (with positive electric charge) moving forward in time mathematically (and physically) equivalent to a negatively charged electron moving backward in time? The answer is that the reversal in charge sign, which results from the reversal of the electron's proper time, follows from the TCP theorem that was mentioned in Chap. 2 (see note 38 there). And second, what causes the electron to suddenly move backward in time? Picturesquely, the electron is recoiling from the emitted burst of gamma ray B. Similarly, the absorption of the energy of gamma ray A by the electron that is recoiling backward in time causes a second recoil, giving the world line of what was originally called electron$_2$. This reinterpretation of a kinked spacetime diagram was described as follows (in Feynman's famous words): "It is as though a bombardier flying low over a road suddenly sees three roads and it is only when two of them come together and disappear that he realizes that he has simply passed over a long switchback in a single road."

In a later paper[87] Putnam's critic presented another line of attack against Putnam's interpretation of spacetime diagrams as lending support to time travel. There the critic observed that the presence of the time-reversed Oscar$_2$ shows that the "world of the Oscars" is not *temporally orientable*. A temporally orientable spacetime is one in which *every* point in it agrees with its local neighbors on the directions of past and future—a condition clearly *not* satisfied for the case of Oscar$_2$. As the critic pointed out, the first time travel spacetime discovered, the Gödel universe, *is* temporally orientable, and so in it the ambiguity of Oscar$_2$ (whether he is traveling backward in time as opposed to living forward 'in reverse') does not occur. That is, the critic agreed with Putnam's acceptance of the conceivability of time travel to the past, but not with his use of Feynman's concept of anti-matter as time-traveling matter. That critic wasn't alone in that opinion.

One physicist, for example, wrote of "Feynman's rather loose talk of particles 'traveling' backward . . . in time,"[88] and the well-known philosopher John Earman declared "It is true that Feynman uses the slogan 'Positrons are electrons running backward in time,' but it is dangerous to draw conclusions from slogans."[89] I am not sure what Earman meant by "slogans": a careful reading of Feynman indicates that he actually took the matter quite seriously.[90] In his 1949 positron paper (note 86), for example, he wrote that "the idea that positrons can be represented as electrons with proper time reversed relative to true time has been discussed by the author and others,"[91] and also that "Previous results suggest waves propagating . . . toward the past, and that such waves represent the propagation of a positron."

---

[87] See R. Weingard in note 114 in Chapter 1.

[88] H. Price, "The Asymmetry of Radiation: Reinterpreting the Wheeler-Feynman Argument," *Foundations of Physics*, August 1991, pp. 959-975.

[89] J. Earman, "On Going Backward in Time," *Philosophy of Science*, September 1967, pp. 211-222.

[90] Feynman declares the view of a positron as a time traveling electron to be of value, for example, in his famous book *Quantum Electrodynamics*, W. A. Benjamin 1961, p. 68.

In any case, spacetime diagrams are highly useful in discussing time travel, but they do have some curious twists. In our everyday world, a path that joins two points on a surface with the *minimum* length is called a geodesic of that surface. As you'll see later in this chapter spacetime geodesics do indeed possess an extremal property, but rather than being a minimum it is a *maximum* property. Spacetime diagrams can be misleading on this matter, so it is important to remember that such diagrams are not a perfect representation of all the properties of a spacetime.

It is customary to draw spacetime diagrams with the speed of light as unity ($c = 1$). That is, a distance of 300,000 km on the space axis is represented by the same extension as is one second on the time axis. This means that the world line of a photon is tilted away from the vertical time axis by 45°. Because photons can travel in both space directions (to the left and to the right) in the two-dimensional spacetime we can draw on a piece of paper, and because the speed of light is the limiting speed of the universe, we can represent the collection of all possible world lines as those paths that *never* tilt more than 45° away from the vertical, which forms what is called a *light cone* in spacetime, as shown in Fig. 3.10 (which attempts to represent a *three*-dimensional spacetime, one with two space—imagine a y-axis, out of the paper, perpendicular to both the x and t axes—and one time dimension).

In Fig. 3.10 I have taken $x = y = 0$ at $t = 0$ for all the possible world lines involving speeds below the speed of light. Let's agree to call this spacetime point the Here-Now. Then, spacetime points in the upward half of the light cone are in the Future of Here-Now; similarly, spacetime points in the lower half of the light cone are in the Past of the Here-Now. We can draw a straight world line from the Here-Now to any point in the Future half-cone with a tilt of less than 45° away from the vertical, which means that a massive particle could travel from the Here-Now to that point at less than the speed of light. Similarly, a massive particle starting at any point in the Past half-cone could have reached the Here-Now by traveling at less than the speed of light. Such a world line is called *timelike* because its projection on the time axis is greater than its projection on a space axis—they are the world lines connecting spacetime points that are potentially causally linked. That is, an event at a spacetime point in the Past half-cone could have had an effect on the event at the Here-Now, even though its influence propagated at less than the speed of light. Also, an event at the Here-Now could potentially affect the event at any spacetime point in the Future half-cone of the Here-Now.

Any points in the regions of spacetime outside the Future and Past half-cones *cannot* be reached from the Here-Now except by world lines tilted more than 45°

---

[91] The "others" Feynman had in mind included, in particular, the eminent Swiss physicist Ernest C. G. Stükelberg (1905-1984), who in a 1942 article in the journal *Helvetica Physica Acta* also wrote of waves scattering backward in time.

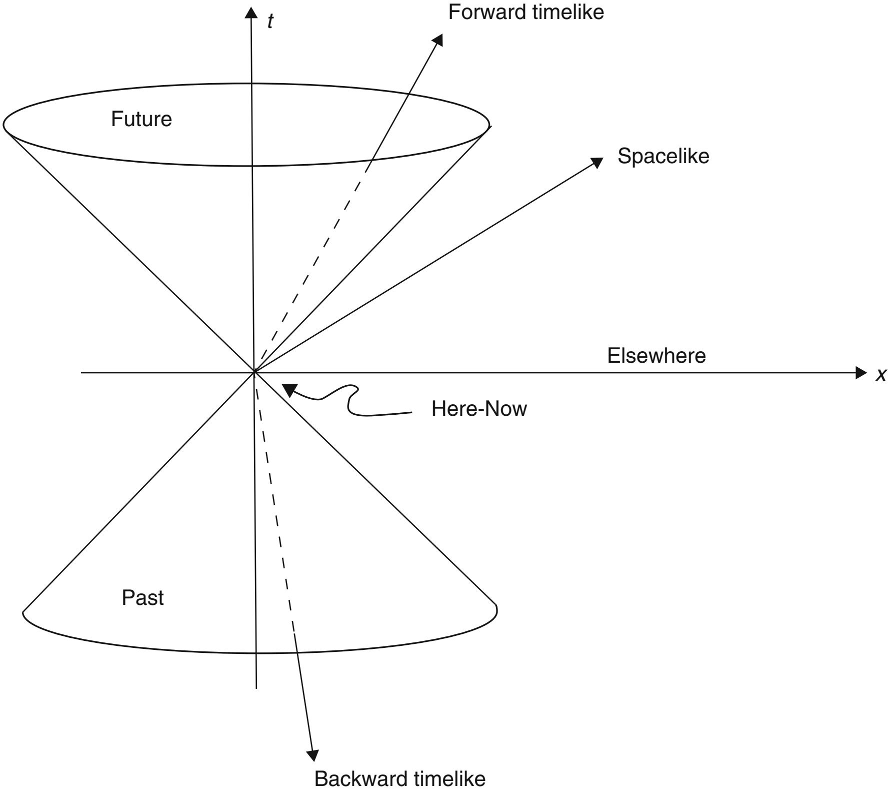

Fig. 3.10 A light cone in spacetime

away from the vertical. Such world lines, which represent travel at a speed faster than light, are called *spacelike* because their projections on a space axis is greater than their projections on the time axis. It is impossible for these world lines to connect causally linked events, and collectively they form the Elsewhere of the Here-Now. Notice that every point in spacetime has its own light cone. If A and B are causally linked, then if B is in the Future half-cone of A, then A is in the Past half-cone of B.

The imagery of the light cone is often useful in making seemingly quite abstract ideas appear transparent. For example, can an observer predict his own future from perfect knowledge of his own past? The easy answer is “No, because quantum uncertainties prohibit perfect knowledge of even the present, much less the past.” But suppose we ignore quantum mechanics and limit our question to a universe that obeys only classical physics (which includes the special and general theories of relativity). Surprisingly (perhaps), the answer is *still* no. Having perfect knowledge of your own Past half-cone doesn’t include knowing the *entire* past, so if you attempt to predict your own future (say, 1 min from now), there can be influences in Elsewhere that will arrive in the future (say, 59 s from now) about which you

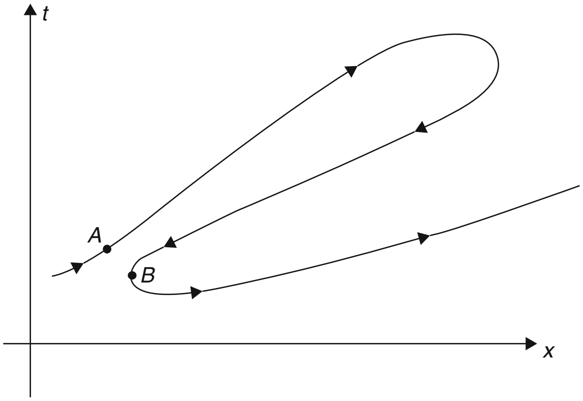

Fig. 3.11 World line of a particle traveling backward in time from A to B

presently, by definition, cannot have any knowledge. And without that knowledge, you cannot predict. As one writer amusingly concludes a tutorial on this topic, "the prospect of predicting the future looks pretty bleak."[92]

A spacetime diagram does not always have to have future directed world lines. If a particle moves backward through time, assuming such a thing is possible, then the diagram can show this by having the world line double back on itself, as in Fig. 3.11 (in which the world line curves back and comes arbitrarily close to itself: it is the world line of a particle that visits itself in the past). Note that the world line in Fig. 3.11 does not actually touch or cross itself, because that would represent more than just a visit—it would represent a particle occupying the same spatial location at the same time as its earlier self. That would be catastrophic and, because it *did not* happen it *cannot* happen. Since the arrowheads on the world line always point in the direction of the local future of the particle, if the 'particle' is actually human then increasing memories are formed in the direction of the arrows. The time traveler at B has more memories than he does at A, even though A and B are nearly identical points in spacetime.

There is a problem with Fig. 3.11 that you may have caught. It is impossible to draw such a doubled-back world line in such a way that *at all places* it never tilts more than 45° from the vertical. That is, at least some portion of the world line will have

$$\left|\frac{\mathrm{dx}}{\mathrm{dt}}\right| > 1$$

which represents superluminal motion (we'll return to this in Chap. 5). One way to keep a bent-back world line always subluminal is to arrange for the light cones along the world line to be tilted relative to each other, as shown in Fig. 3.12, which

---

[92] M. Hogarth, "Predicting the Future in Relativistic Spacetimes," *Studies in the History and Philosophy of Science*, December 1993, pp. 721-739.

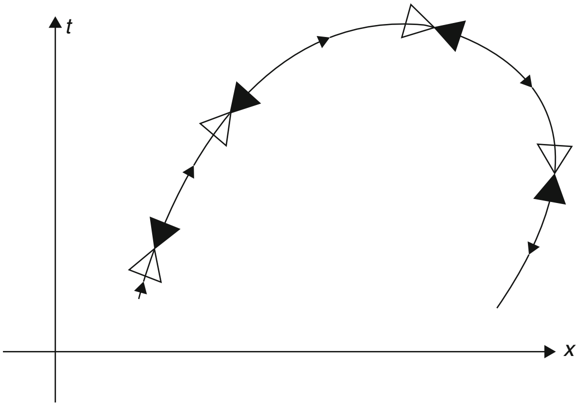

Fig. 3.12 Tilted light cones in a curved spacetime (the future halves are shaded)

is possible only in a *curved* spacetime. This is an illustration of how general relativity *locally* obeys special relativity's demand that nothing travels faster than light and yet, *globally*, in curved spacetime, things are not so simple. In flat spacetimes all light cones are always 'aligned,' but in curved spacetimes they (generally) are not, and from that can come time travel to the past.[93]

Tilted light cones is the physics behind backward time travel around a rotating Tipler cylinder, for example, a particular time machine we'll discuss later in Chap. 5. Light cone tipping is, in fact, *essential* for time travel to the past. The mere presence of mass tips light cones ('warps spacetime'), but the effect is unnoticeable in everyday life on Earth. A truly enormous mass density is required to tip nearby light cones over so that their Future halves noticeably open up toward the massive body. If the massive body is additionally set to rotating, then a further consequence of Einstein's general theory is that the local light cones are tilted additionally in the direction of rotation. That is, the Future half-cones in spacetime open-up both toward the body and in the direction of the rotation.

It should now be clear that the only way a world line can bend back on itself for a close encounter visit is for both x and t to change. In other words, the world line of a particle that remains fixed in space and reverses just its time direction *runs into itself*. This is why the classic fictional time machine of H. G. Wells could not possibly work. A real time machine must move in space as well as through time, as does a Gödelian rocket (or, for that matter, as does the DeLorean time car in the *Back to the Future* films). The idea of warping world lines, to support time travel to the past, entered science fiction at an early date. For example, when the inventor of the time machine in a 1930s tale[94] is asked about the principle underlying his gadget, he replies "An electro-magnetic warping of the spacetime continuum.

---

[93] What is meant by a spacetime being *flat* will be formalized when we get to spacetime *metrics* later in this section. The Minkowski spacetime of special relativity is a flat spacetime, has no tilted light cones, and as such does *not* support time travel to the past.

[94] N. Schachner, "When the Future Dies," *Astounding Science Fiction*, June 1939.

The machine, if it works, will slide around the world-line of events and reappear at any specified time and place."

Using the Lorentz transformation equations from the previous section, we can establish quite general relationships between events in the Future, Past, and Elsewhere regions of spacetime. For example, (1) All events in the Future/Past for the Here-Now observer are in the Future/Past for any other nearby, relatively moving observer; (2) Any event in Elsewhere can appear to be simultaneous with the Here-Now for some observer and not simultaneous for another observer; and (3) The temporal ordering (the relations of *before* and *after*) of causally related events is the same for all observers. This is not so for events that are not causally related; if two events have a spacelike separation, then two observers can disagree over the temporal ordering of the events. This is, in fact, the basis for the two-wormhole and the cosmic string time machines, both of which will be discussed in Chap. 6. All these statements are easy to prove.

Consider, for example, statement (1). From the previous section we have (with $c = 1$)

$$t' = \frac{t - \mathrm{vx}}{\sqrt{1 - v^2}} \text{ and } x' = \frac{x - \mathrm{vt}}{\sqrt{1 - v^2}}$$

where t and x are the coordinates of some event A as measured by the observer in the stationary reference frame, and t' is the time measured by the observer in the reference frame moving at speed v. Thus

$$x'^2 - t'^2 = \frac{(x - \mathrm{vt})^2 - (t - \mathrm{vx})^2}{1 - v^2} = [\text{after a little algebra}]\ x^2 - t^2.$$

For the stationary observer the criterion for an event to be in the Future half-cone is $t > |x|$, that is, $t^2 > x^2$. Thus, $x^2 - t^2 < 0$ for all Future events. But the foregoing result then says $x'^2 - t'^2 < 0$, too, which is the moving observer's criterion for the event being in his Future half-cone. The same sort of argument shows that the two observers also agree on Past events.

Next, suppose that two events A and B occur such that the stationary observer measures them to be $\Delta T = t_B - t_A$ apart in time. Then, we can establish statement (2) by writing

$$t'_A = \frac{t_A - \mathrm{v}x_A}{\sqrt{1 - v^2}} \text{ and } t'_B = \frac{(t_B + \Delta T) - \mathrm{v}x_B}{\sqrt{1 - v^2}}$$

and so

$$\Delta T' = t'_B - t'_A = \frac{\Delta T + v(x_A - x_B)}{\sqrt{1 - v^2}}.$$

From this we have $\Delta T^{'} = 0$ (that is, simultaneity) for the two events for the special observer moving at the speed

$$v = \frac{\Delta T}{x_B - x_A}$$

and this speed is less than the speed of light for the condition $x_B - x_A > \Delta T$. This is, of course, the condition for event B to be in the Elsewhere of event A. In fact, we can even have $\Delta T^{'} < 0$ (with $\Delta T > 0$) for $v < 1$ in this case of spacelike separation of A and B. That is, a stationary observer and a sublight-speed moving observer can disagree about the temporal ordering of events with spacelike separation.

Similarly, for event B to be in the causal Future of event A, we have the condition $x_B - x_A < \Delta T$. Then,

$$\Delta T' = \frac{\Delta T - \mathrm{v}(x_B - x_A)}{\sqrt{1 - v^2}} > \frac{\Delta T - v\Delta T}{\sqrt{1 - v^2}} = \Delta T\frac{1 - \mathrm{v}}{\sqrt{1 - v^2}}.$$

Thus, $\Delta T > 0$ says $\Delta T^{'} > 0$ for $v < 1$, and this establishes statement (3).

If we were drawing diagrams with both axes representing space (a plot of y versus x, for example), we would normally define a *distance metric* for the diagram using our everyday ideas about distance. That is, we could say that if we make differential movements of dx and dy along the two coordinate axes, then the differential distance ds is given by

$$(ds)^2 = (dx)^2 + (dy)^2.$$

This is, of course, just the Pythagorean theorem for the ‘Euclidean’ or ‘as the crow flies’ distance function. But it is not the only possible distance function. A distance function has several interesting mathematical properties,[95] but the one we are particularly interested in here is its invariance with respect to the coordinate system. For example, if we draw a line segment on a flat sheet of paper, the physical distance between its end-points does not depend on how we happen to select the x and y axes, a fact illustrated in Fig. 3.13 with the addition of a rotated and translated primed system. The coordinates for the endpoints A and B are obviously different in the two coordinate systems, but we still find that

$$(dx)^2 + (dy)^2 = (dx')^2 + (dy')^2.$$

---

[95] Mathematicians have defined the general properties of a distance function as follows: if A and B are any two points, and if $d(A,B)$ is the distance between A and B , then (1) $d(A,B) = d(B,A)$; (2) $d(A,B) = 0$ if and only if $A = B$; and (3) if C is any third point, then $d(A,B) \leq d(A,C) + d(C,B)$. The Pythagorean distance function possesses all three of these properties, but so do many other functions (for example, $ds = |dx| + |dy|$).

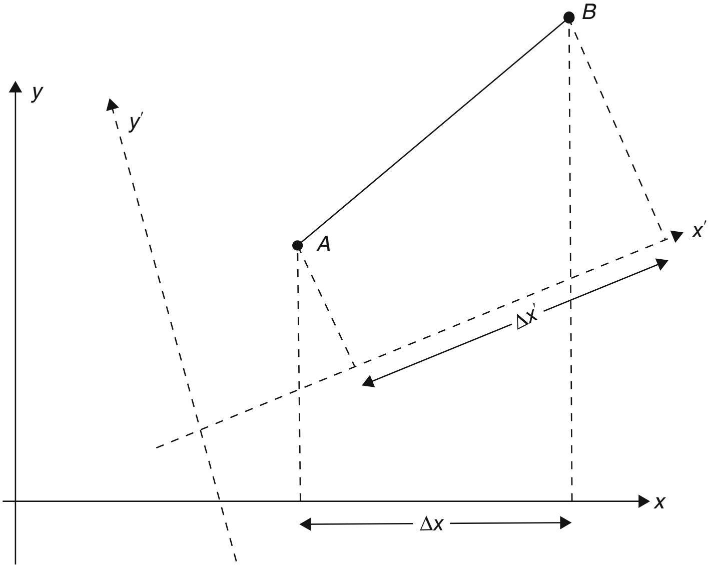

Fig. 3.13 Rotated/translated coordinate systems

We say that the Pythagorean distance function is *invariant*, that it is the same for all coordinates systems that are simply rotations and/or translations of each other.

We know that different observers, if in relative motion in the same spacetime, will see different space and time coordinates for the same event. Thus, it is natural to ask 'what is the metric for flat spacetime?' Is there, in fact, a metric that gives the same *distance* between two events for all observers? We might try to generalize in the obvious way from the Pythagorean theorem, and write

$$(ds)^2 = (dt)^2 + (dx)^2 + (dy)^2 + (dz)^2$$

where now all four dimensions are included. We would then ask ourselves whether it is true that

$$(ds)^2 = (ds')^2 = (dt')^2 + (dx')^2 + (dy')^2 + (dz')^2?$$

For our simple two-dimensional spacetime, this question reduces to asking whether

$$(dt)^2 + (dx)^2 = (dt')^2 + (dx')^2?$$

Using the Lorentz transformation equations from earlier, it is easy to discover that the answer is no. The 'natural' generalization of Pythagorean distance for flat, two-dimensional spacetime fails when four dimensions are included. So, what do we do now? Recalling the words of Professor Mundle from *Some First Words* (see note 23 there), we might wonder whether this difficulty could result from the fact that there is no fourth direction along which the time axis can point at right angles to the three space directions? At least, there is no *real* direction—but perhaps there is an imaginary one. Accordingly, with = $\sqrt{-1}$, let's try

$$(ds)^2 = (idt)^2 + (dx)^2 + (dy)^2 + (dz)^2 = -(dt)^2 + (dx)^2 + (dy)^2 + (dz)^2.$$

Using *imaginary time*, something that seems to be in the realm of science fiction, has resulted in a change in the sign of $(dt)^2$.

This is a crucial change, however, because this new metric *is* invariant. For a reason to be explained in the next section, I will use the negative of this metric (a choice that has no impact on the invariance property) and so write

$$(ds)^2 = (dt)^2 - (dx)^2 - (dy)^2 - (dz)^2.$$

For our simplified two-dimensional spacetime this reduces to

$$(ds)^2 = (dt)^2 - (dx)^2.$$

As before, the Lorentz transformation equations (with $c = 1$) are

$$t' = \frac{t - \mathrm{vx}}{\sqrt{1 - v^2}} \text{ and } x' = \frac{x - \mathrm{vt}}{\sqrt{1 - v^2}}.$$

If we then calculate dx' and dt' from these equations, using

$$dx' = \frac{\partial x'}{\partial x}dx + \frac{\partial x'}{\partial t}dt, dt' = \frac{\partial t'}{\partial x}dx + \frac{\partial t'}{\partial t}dt$$

which are the fundamental relations for the *total* differential[96] of a function of two variables, and insert the results into $(dt')^2 - (dx')^2$, we quickly discover the invariance property of this quantity (a result we actually found earlier, in a different way, when we showed that observers in relative motion agree about what events are in the Future and what events are in the Past). Thus,

$$(dt')^2 - (dx')^2 = (dt)^2 - (dx)^2.$$

This quantity, on either side of the equality, is called the spacetime *interval* between the two events separated in flat spacetime by either dt, dx, dy, and dz, or by dt', dx', dy', and dz'. The observers in the unprimed and the primed systems see different individual space and time separations for two events, but they see the same

---

[96] The $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial t}$ symbols denote the *partial* derivatives with respect to x and t (see any good calculus book to brush-up on this). The rest of this chapter *will* have some more math in it, involving derivatives and even an integral or two, but nothing beyond freshman calculus. I've included it mostly for those who would feel cheated without *some* math!

interval. A single time coordinate, and three space coordinates, are said to form a *four-vector that is invariant under Lorentz transformation*. There are, in addition, other four-vectors that are also invariant under Lorentz transformation, such as the energy-momentum, velocity, and force four-vectors; all these quantities have invariants that are formed the same way, by taking the difference of the squares of the time components and the sum of the squares of the space components. The intrusion of the square root of minus one, in the time coordinate of the metric, seems a pretty clear indication that time *is* different from space.[97]

But still, the spatialization of time is nonetheless deeply embedded in Western culture. For example, when writing her popular 1978 book, drawing historical parallels between the fourteenth and twentieth centuries, Barbara Tuchman titled it *A Distant Mirror*, and not *An Old Mirror*. The mathematical mixing of space and time appeared quite early in science fiction, but often in comically mangled form. In one such tale, for example, as an evil scientist uses his time machine to transport captives into the past, he tells them, “We’ve got a longish journey before us, ten thousand years more, multiplied by the fourth power of two thousand miles.”[98]

In general relativity, the metric of any four-dimensional spacetime has the structure of what mathematicians call a *symmetric quadratic Riemannian form*:

$$(ds)^2 = \sum_{i=1}^{4}\sum_{j=1}^{4} g_{ij}(dx_i)\left(dx_j\right), g_{ij} = g_{ji}$$

where $x_1 = t$, $x_2 = x$, $x_3 = y$, and $x_4 = z$, and the 16 g’s are all functions of these four variables. (Because of the symmetry condition, only 10 of the g’s are independent.) In this notation, a *flat* spacetime is mathematically characterized by $g_{11} = 1$, $g_{22} = g_{33} = g_{44} = -1$, $g_{ij} = 0$ for all $i \neq j$. Now, for a given spacetime, one can arbitrarily choose an infinity of coordinate systems. If just one of this infinity of systems is such that the ±1, 0 values for the g’s occur, then that spacetime is *globally* (that is, everywhere) flat. If no such coordinate system exists, then that spacetime is necessarily *curved*—I’ll give you a more intuitive view of curvature in just a bit. (If this notation is extended to a fifth dimension[99] by including the additional coordinate x_5, then there are an additional eight off-diagonal g’s and so four new *independent* g's. These are just sufficient to describe the electromagnetic field, too, along with the gravitational field described by the other ten g's. That is what is meant by saying five-dimensional spacetime 'unifies' gravity with electromagnetism.) There is a g_{55}, too, which could be allowed to model a slowly varying gravitational constant, as suggested by Dirac in 1938.

---

[97] Writing in 1972, one famous physicist said of his first encounter with the metric of special relativity (Minkowski spacetime), “Now, when I saw that minus sign [in $-(dt)^2$], it produced a tremendous effect on me. I immediately saw that here was something new.” See P. A. M. Dirac, “Recollections of an Exciting Era,” in *History of Twentieth Century Physics* (C. Weiner, editor), Academic Press 1977.

[98] V. Rousseau, “The Atom Smasher,” *Astounding Stories*, May 1930.

[99] The fifth dimension was introduced in the 1920s by the German physicist Theodor Kaluza (1885-1954), but just what the nature of this fifth dimension might be remains a mystery. A few years after Kaluza, the Swedish physicist Oscar Klein (1894-1977) speculated that it might be a spatial dimension curled-up in a tiny circular path, so tiny that we don’t notice it; the issue remains open. The idea of a fifth dimension appeared early in pulp science fiction, as in the January 1931 tale “The Fifth-Dimension Catapult,” (*Astounding*) by Murray Leinster.

The g functions are the components of the so-called *metric tensor of the second rank* (see the last discussion question for more on what this means) of that spacetime. The g's at each point in spacetime are related to the curvature of spacetime at that point, which in turn is dependent on the g's and on the energy density at that point. In fact, the ten equations for the g's, which *are* the famous Einstein gravitational partial differential field equations, are both nonlinear and coupled. That is, each g_{ij} is in general a nonlinear function of all the other g_{ik}, which accounts for the notorious difficulty[100] in finding analytical solutions to the field equations except in certain highly special cases, such as the spacetime of spinning spheres and rotating infinite cylinders.

What does it mean to 'solve' the field equations? It is useful to think of the equations schematically as follows:

local geometry of spacetime ← local density, momentum and stress of the mass-energy of spacetime where the direction of the arrowhead means that the 'usual' practice is to assume the right-hand side (the so-called stress-energy tensor) as given, and then attempt to calculate the left-hand side. If the attempted calculation can be done, then one has solved the field equations for the spacetime geometry that is associated with the assumed mass-energy distribution.

Suppose, however, that we reverse the direction of the arrowhead. That is, suppose we assume a desired geometry. That is what Einstein did when he assumed the geometry of a *static* (non-expanding) universe and solved for the required mass-energy. What he found was just what had been observed by astronomers up to that time—a multitude of 'grains of matter' (what physicists call *dust*) plus the infamous cosmological constant. The constant, with its repulsive gravity, was needed to counteract the ordinary gravitational attraction of the stars that tends to pull them together. The later discovery that the universe is not static, but rather is expanding, rendered Einstein's solution moot.

Now let's go Einstein one better, and assume a spacetime geometry that contains closed timelike curves (a time machine, in other words) and then try to calculate the mass-energy distribution required by that spacetime. If that can be accomplished—and in fact the field equations themselves provide an algorithmic means of solution in this direction—then the physicist's work is done. The required mass-energy distribution requirements are put out to bid to 'spacetime engineers' and the lowest bidder 'simply' constructs that mass-energy distribution and so builds us our time machine! When the calculations are done, however, what has happened without fail, at least up to now, is that the resulting mass-energy distribution comes out with an ‘unphysical’ nature, a technical way of saying our spacetime engineer wouldn’t know *how* to assemble the required mass-energy distribution. What is meant by ‘unphysical’ will be explored, in the particular case of wormhole time machines, in Chap. 6.

---

[100] An elementary, quite interesting discussion of the enormous computational complexity of the field equations is presented in Richard Pavelle and Paul S. Wang, "MACSYMA from F to G," *Journal of Symbolic Computation*, March 1985, pp. 69-100.

The 16 g’s for a four dimensional spacetime are often written in the form of a 4 × 4 matrix. In fact, the metric tensor is of the *second* rank precisely because a matrix has a *two*-dimensional form; scalars and vectors, which have forms of zero and one dimension, are tensors of rank zero and one, respectively. The collection of the algebraic signs of the main diagonal terms (the g_{ii}) is called the *signature* of the metric tensor. The signature of flat (Minkowski) spacetime is thus written as $[+, -, -, -]$; in more general (curved) spacetimes, this same signature is called *Lorentzian*. By contrast, the signature of a four-dimensional Euclidean space is $[+, +, +, +]$. This signature is called *Riemannian*. The geometry of flat, Minkowski spacetime is not Euclidean geometry because the spacetime signature of its metric has both plus *and* minus signs. As an uncurved spacetime, Minkowski spacetime *has no gravity*; to get gravity, we need a curved spacetime.

The idea of linking the curvature of a four-dimensional space to physical phenomena is the signature feature of general relativity, but it actually pre-dates Einstein by decades. It can be found, for example, in the work of the British mathematician William Kingdon Clifford (1845–1879), done in the 1860s and 1870s before Einstein’s birth (in the year of Clifford’s death). In his posthumously published book *The Common Sense of the Exact Sciences* (1885), Clifford wrote “We may conceive our space to have everywhere a nearly uniform curvature, but that slight variations of the curvature may occur from point to point and themselves vary with time. These variations of the curvature with time may produce effects which we not unnaturally attribute to physical causes independent of the curvature of our space. We may even go so far as to assign to this variation of the curvature what really happens in that phenomenon which we term the motion of matter.” It isn’t a long jump from “motion of matter” to gravity![101]

Not everyone enthusiastically embraced this new, radical view of nature. For example, the great Scottish mathematical physicist James Clerk Maxwell (1831–1879), of Maxwell’s equations fame, who knew Clifford through their common membership in the London Mathematical Society, summarily dismissed this part of Clifford’s work as simply the speculations of a “space crumbler.” Decades later, Einstein faced the same rejection when the eminent British astronomer Sir James Jeans declared “Einstein’s crumbling of his four-dimensional space may . . . be considered to be . . . fictitious.”[102]

---

[101] Clifford almost surely found inspiration in this part of his work from the even earlier efforts of the German mathematician Bernhard Riemann (1826-1866). See, for example, Clifford’s translation of Riemann’s famous 1854 lecture “On the Hypotheses Which Lie at the Bases of Geometry,” *Nature*, May 1, 1873, pp. 14-17, and continued in the next issue (May 8, 1873, pp. 36-37).

Spacetime geometries are not easy concepts to grasp, and the metrics of curved spacetimes are even more complicated than is the metric of the flat spacetime of Minkowski’s special relativity. As one paper so aptly put it, “Experience has taught us that the space in which we live has a geometry that is three-dimensional and Euclidean . . . We are very much at home with [that] geometry . . . But the geometric properties of a Minkowskian space are so alien to us that we may well despair of visualizing them, and a Riemannian [curved] space . . . seems totally beyond comprehension.”[103]

An important feature of Riemannian geometry is that although it is generally not globally flat, it is always *locally* flat. Thus any sufficiently small region in a curved Riemannian spacetime can be approximated, with arbitrarily small error, by a flat pseudo-Euclidean Minkowskian spacetime. That is, at each point in Riemannian spacetime, there is some *particular* inertial frame of reference in which special relativity is all there is to spacetime physics *at that point*. The particular inertial frame required is different, however, from point to point.

In a coordinate system different from rectangular, the flat Minkowskian metric can appear radically altered, but that is just an artifact of the mathematics and has no physical significance. For example, in spherical coordinates the Minkowskian metric becomes the equivalent

$$(ds)^2 = (dt)^2 - (dr)^2 - (rd\theta)^2 - (rsin(\theta)d\phi)^2$$

where ϕ is the azimuthal angle and θ is the angle measured from the polar axis. A related metric occurs in the theory of spherically symmetric, static (no time variation) time machine wormholes (discussed in Chap. 6), of the form

$$(ds)^2 = \left(e^{a(r)}dt\right)^2 - \frac{(dr)^2}{1 - \frac{b(r)}{r}} - (rd\theta)^2 - (rsin(\theta)d\phi)^2$$

where $a(r)$ is called the *redshift function* and $b(r)$ is called the *shape factor*. These two functions are nearly arbitrary, subject only to the constraints that both $b(r)/r$ and $a(r)$ vanish as r goes to infinity (r is the radial distance from the throat of the wormhole mouth). Indeed, as r increases, this curved wormhole spacetime metric

---

[102] Quoted in P. Kerszberg, “The Relativity of Rotation in the Early Foundations of General Relativity,” *Studies in History and Philosophy of Science*, March 1987, pp. 53-79.

[103] R. W. Brehme and W. E. Moore, “Gravitational and Two-Dimensional Curved Surfaces,” *American Journal of Physics*, July 1969, pp. 683-692.

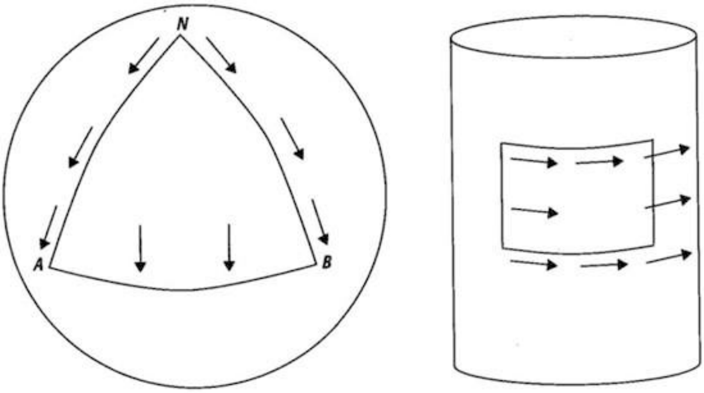

Fig. 3.14 The curvature of a space can be revealed by the process of parallel transport

reduces to that of flat Minkowskian spacetime, so this wormhole spacetime is said to be *asymptotically flat*.[104]

An intuitive appreciation of ‘flatness’ can be realized in terms of what is called the *parallel transport* of a vector around a closed path. In a curved space the vector will experience a rotation, which will not occur in a flat space. Two examples of parallel transport in ordinary three-dimensional space are shown in Fig. 3.14. The spherical surface on the left is curved because if you slide the vector from *N* to *A* to *B* and then back to *N*, all the while keeping it parallel to its immediately previous orientation, then when it gets back to *N* the vector will point to *B*, *not* toward *A* as it initially did. A similar trip on the cylindrical surface, however, results in zero rotation of the vector. Thus, the cylindrical surface, despite superficial appearances, is *not* curved.

Using the idea of the metric tensor, we can develop a more formal demonstration that the surface of a sphere, unlike like that of a cylinder, is not flat. On the surface of a sphere of radius a (*on* the surface $r = a$ everywhere, and so $dr = 0$), the measure of the distance between two points (in spherical coordinates) is[105]

$$(ds)^2 = a^2 sin^2(\theta)(d\phi)^2 + a^2(d\theta)^2.$$

Writing $x_1 = \phi$ and $x_2 = \theta$ yields the more general form

---

[104] For wormhole spacetimes that are not asymptotically flat, see (for example) K. Narahara, *et al.*, “Traversable Wormhole in the Expanding Universe,” *Physics Letters B*, September 29, 1994, pp. 319-323.

[105] Note *carefully* that this is a purely spatial problem, with no time, and we are taking all of the metric coefficients as positive (unlike in the case of a spacetime metric).

$$(ds)^2 = g_{11}(dx_1)^2 + g_{12}(dx_1)(dx_2) + g_{21}(dx_2)(dx_1) + g_{22}(dx_2)^2$$

or, using the symmetry condition $g_{12} = g_{21}$, we have

$$(ds)^2 = g_{11}(dx_1)^2 + 2g_{12}(dx_1)(dx_2) + g_{22}(dx_2)^2.$$

From this we immediately have $g_{11} = a^2 sin^2(\theta)$, $g_{22} = a^2$, and $g_{12} = g_{21} = 0$.

Now, suppose we ask whether it is possible to find some new coordinate system (with variables x'_1 and x'_2) in which the invariant $(ds)^2$ is given by the flat Euclidean metric $\left(dx'_1\right)^2 + \left(dx'_2\right)^2$. In such a coordinate system (if it exists) we would have the 'flatness conditions' of $g'_{11} = g'_{22} = 1$ and $g'_{12} = g'_{21} = 0$. With such a change of coordinates, each of our original ϕ, θ coordinates would generally be a function of both of the new coordinates—that is, $\phi = \phi\left(x'_1, x'_2\right)$ and $\theta = \theta\left(x'_1, x'_2\right)$. Thus, writing the total differential of a function of two variables (see note 96 again), we have

$$d\phi = \frac{\partial \phi}{\partial x'_1} dx'_1 + \frac{\partial \phi}{\partial x'_2} dx'_2$$

$$d\theta = \frac{\partial \theta}{\partial x'_1} dx'_1 + \frac{\partial \theta}{\partial x'_2} dx'_2.$$

Substituting these expressions into the above expression for $(ds)^2$ on the surface of a sphere, and collecting terms, we arrive at

$$\begin{aligned}(ds)^2 &= a^2\left\{\left[\sin^2(\theta)\left(\frac{\partial \phi}{\partial x'_1}\right)^2 + \left(\frac{\partial \theta}{\partial x'_1}\right)\right]\left(dx'_1\right)^2 + \left[\sin^2(\theta)\left(\frac{\partial \phi}{\partial x'_2}\right)^2 + \left(\frac{\partial \theta}{\partial x'_2}\right)^2\right]\left(dx'_2\right)^2\right. \\ &\quad \left. +2\left[\sin^2(\theta)\frac{\partial \phi}{\partial x'_1}\frac{\partial \phi}{\partial x'_2} + \frac{\partial \theta}{\partial x'_1}\frac{\partial \theta}{\partial x'_2}\right]\left(dx'_1\right)\left(dx'_2\right)\right\} \\ &= g'_{11}\left(dx'_1\right)^2 + 2g'_{12}\left(dx'_1\right)\left(dx'_2\right) + g'_{22}\left(dx'_2\right)^2\end{aligned}$$

We can now immediately write down each of the g' and, if we demand that they satisfy the 'flatness conditions,' then we have the following three statements:

$$sin^2(\theta)\left(\frac{\partial \phi}{\partial x'_1}\right)^2 + \left(\frac{\partial \theta}{\partial x'_1}\right)^2 = \frac{1}{a^2}$$

$$sin^2(\theta)\left(\frac{\partial \phi}{\partial x'_2}\right)^2 + \left(\frac{\partial \theta}{\partial x'_2}\right)^2 = \frac{1}{a^2}$$

$$sin^2(\theta)\frac{\partial \phi}{\partial x'_1}\frac{\partial \phi}{\partial x'_2} + \frac{\partial \theta}{\partial x'_1}\frac{\partial \theta}{\partial x'_2} = 0.$$

For a *globally* flat surface, that is, a surface that is flat *everywhere*, these three statements must hold in particular at the poles of the sphere. That is, at $\theta = 0^\circ$

and at $\theta = 180^{\circ}$. At both of these points $\sin(\theta) = 0$, and so at the poles the three statements reduce to

$$\left(\frac{\partial\theta}{\partial x_1'}\right)^2 = \frac{1}{a^2}$$

$$\left(\frac{\partial\theta}{\partial x_2'}\right)^2 = \frac{1}{a^2}$$

$$\frac{\partial\theta}{\partial x_1'}\frac{\partial\theta}{\partial x_2'} = 0.$$

But the third statement is incompatible with the first two and, because of that incompatibility, there is no primed coordinate system in which the g' coefficients in the metric are those of a globally flat metric. Thus, unlike the surface of a cylinder, the surface of a sphere is not flat but rather is curved. This almost surely comes as no surprise to you, but now you can *prove* it!

3.6 Proper Time and the Twin Paradox in Time Travel to the Future

"[The] equations of duo-quadrant lineations [have] been substantiated ... Our fourth-angle deviation from the six conceivable electronic dimensions did the trick all right. I went forward in Time."[106]

The spacetime interval of the previous section has an important interpretation that leads to one of the more dazzling results of special relativity—time travel into the future. First, recall the flat spacetime metric

$$(ds)^2 = (dt)^2 - (dx)^2 - (dy)^2 - (dz)^2$$

in which the use of unprimed variables indicates that the measurements on the space and time coordinates of a moving particle are made with respect to a stationary observer's frame of reference. Now, suppose that the space and time coordinates of a moving particle are made with respect to the particle instead. Then, using primed variables for measurements made in this new frame of reference, $dx' = dy' = dz' = 0$ because the particle is always at the origin (by definition)! Recalling the invariance of the spacetime interval for all observers, we conclude that

---

[106] A science fiction scientist babbles incoherent nonsense, not special relativity, about how to travel into the future, in a story by J. H. Haggard, "He Who Masters Time," *Thrilling Wonder Stories*, February 1937.

$$(ds')^2 = (ds)^2 = (dt')^2.$$

That is, the spacetime interval between two events is the time lapse measured by a clock attached to a particle that moves from one event to the other. This time is called *proper time*, which gets its name from the idea that it belongs to (is the *property* of) the moving particle. This is the technical reason for taking (as we did in the previous section) $(ds)^2 = (dt)^2 - (dx)^2$ rather than $(ds)^2 = (dx)^2 - (dt)^2$. The first choice avoids the somewhat awkward result of an imaginary proper time.

Next, we'll adopt what has come to be called the *clock hypothesis*, which states that an accelerated clock runs at the same instantaneous rate as an unaccelerated clock that is moving alongside at the instantaneously same speed. As we showed earlier (take a look back at Sect. 3.3), if the accelerated clock's instantaneous speed is v, then its rate of time keeping (dt') is related to that of the 'stationary' (unaccelerated) clock (dt) as

$$dt' = \sqrt{1 - \left(\frac{v}{c}\right)^2}\, dt$$

where c is the speed of light. The clock hypothesis is generally assumed to be true. Einstein, himself, in his famous 1905 special relativity paper, specifically took the rate of a clock's timekeeping to be velocity-dependent only. When asked during an interview decades later whether it is permissible to use special relativity in situations involving acceleration, Einstein replied "Oh, yes, that is all right as long as gravity does not enter; in all other cases, special relativity is applicable. Although, perhaps the general theory approach might be better, it is not necessary."[107]

The clock hypothesis has long had experimental verification. For example, in one experiment the time keeping of accelerated atomic clocks was determined to be given precisely by the time dilation formula of special relativity, even when their direct mechanical acceleration (the centripetal acceleration produced by a rapidly spinning disk) exceeded 66,000 gees![108] And even more impressive are the time dilation results of a later experiment in which the time keeping of near light-speed charged particles, orbiting in a magnetic field, was in excellent agreement with the time dilation formula, even as accelerations well in excess of 10^{15} gees were reached![109]

The total elapsed time between two events A and B, as measured by the proper time of an accelerated clock making the journey, is, therefore, given by

---

[107] R. S. Shankland, "Conversations with Albert Einstein," *American Journal of Physics*, January 1963, pp. 47-57.

[108] H. J. Hay, *et al.*, "Measurement of the Red Shift in an Accelerated System Using the Mössbauer Effect in Fe^{57}," *Physical Review Letters*, February 15, 1960, pp. 165-166.

[109] J. Bailey, *et al.*, "Measurements of Relativistic Time Dilation for Positive and Negative Muons in a Circular Orbit," *Nature*, July 29, 1977, pp. 301-305.

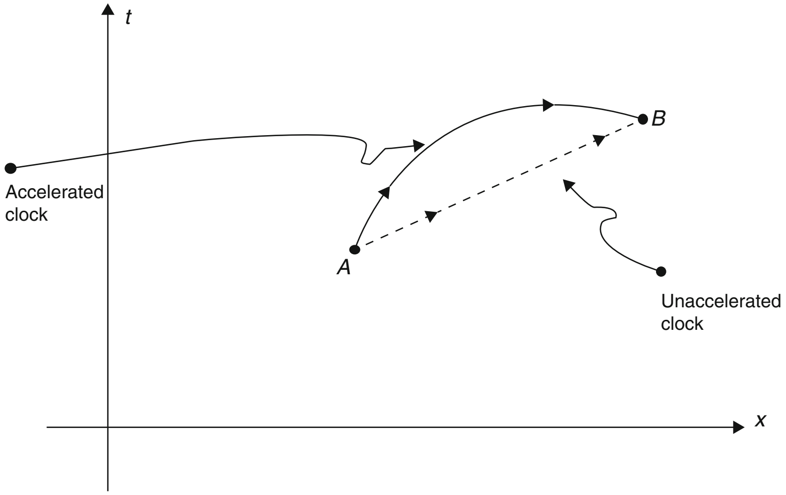

Fig. 3.15 World lines of two clocks (one accelerated and one unaccelerated)

$$t' = \int dt' = \int_{t_A}^{t_B} \sqrt{1 - \left(\frac{v}{c}\right)^2} dt < t_B - t_A \text{ if } v \neq 0$$

where $t_B - t_A$ is the elapsed time between A and B as measured by the unaccelerated clock. The inequality results because, for $v \neq 0$, the integrand is always less than 1. Now, we know that in a spacetime diagram the world line of the unaccelerated clock is a straight line, whereas the world line of the accelerated clock is a curved line. Thus, using Fig. 3.15, combined with the inequality $t' < t_B - t_A$, we have the following central result: the world line of *maximum* proper time is the one that *looks the shortest*, that is, the straight (or free-falling geodesic) world line. In the spacetime diagram the curved line looks longer, but in fact any curved line will have a smaller proper time than does the straight world line. This is a dramatic example of how Minkowskian *spacetime* geometry differs from Euclidean *space* geometry; in the latter geometry, there is no longest path between two points.

From this, we can now understand the famous paradox of the twins, Bob and Bill. Bill remains on Earth, but Bob gets into a rocket ship and goes on a high speed trip out into space. Eventually he brings his ship to a stop, turns around, and returns to Earth. The world lines of Bob and Bill are initially together, then they diverge as Bob goes on his trip, and then they come together again at the end of Bob's trip, as shown in Fig. 3.16. The details of Bob's trip are not important for a general statement of the paradox (although in just a bit I *will* present the details for one possible trip). All we need observe for now is that Bill's world line from A to B is straight, whereas Bob's is curved. Bill's body (that is, his local clock) will therefore measure a greater proper time than will Bob's; that is, Bob will be younger than his stay-at-home twin! Equivalently, upon his return Bob will hear his Earthbound

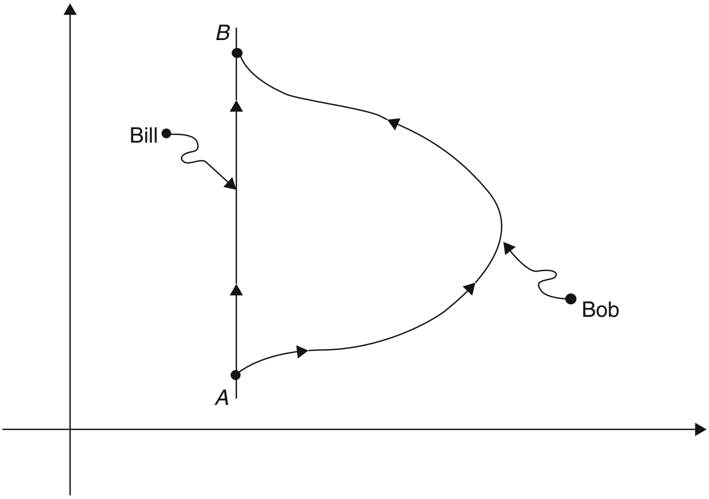

Fig. 3.16 World lines of unaccelerated (Bill) and accelerated (Bob) twins

brother declare the date to be further in the future than Bob's trip lasted (according to Bob). Bob will therefore conclude that he has traveled into the future.[110]

This situation is called a paradox, not because of the time travel aspect (there are no logical paradoxes associated with travel into the future), but rather because it seems to violate the very spirit of special relativity. That is, from Bill's point of view, Bob at first travels away and then returns. But one can argue that from Bob's point of view it is Bill who first recedes and then returns. So why is it *Bob* who is the younger, and not Bill? Long after Einstein's 1905 publication of special relativity, this point remained a great puzzle for many. For example, in the 1923 Presidential Address to the Eastern Division of the American Philosophical Association, we read this very objection to the twin paradox. The conclusion by the speaker is that such a thing "could happen only in a universe in which all squares were round and the *principio contradiction* had been put to sleep."[111]

So, what's the answer, why is it *Bob* who is the younger of the twins? The classic physics answer is that the two points of view are actually *not* symmetrical, that there is a definite asymmetry between Bill and Bob. After all, it is Bob who feels the acceleration from the rocket's engines as he blasts off from Earth—it is Bob, not

---

[110] The twin paradox is hinted at in Einstein's 1905 paper, but it is in a 1911 address to the International Congress of Philosophy in Bologna, by the French physicist Paul Langevin (1872-1946), that a human space traveler is first introduced (in a cannonball moving at near light-speed, an idea motivated by Langevin's reading of Jules Verne's 1872 novel *From the Earth to the Moon*). The writer Pierre Boulle proudly mentioned this contribution by his fellow Frenchman in the time travel story "Time Out of Mind" (you can find it in Boulle's collection *Time Out of Mind*, Vanguard Press 1966).

[111] W. P. Montague, "The Einstein Theory and a Possible Alternative," *The Philosophical Review*, March 1924, pp. 143-170.

Bill, who feels *force*—whereas Bill feels nothing unusual as he remains on Earth. The more fundamental physics answer, however, is that Bob's world line in spacetime is curved, whereas Bill's is straight.

In an open, spatially unbounded, flat spacetime, *curved* is indeed synonymous with *accelerated*, but this need not be so in a *closed,* flat spacetime. In an open, flat spacetime, the only way two world lines can diverge in the past and then meet again in the future is for at least one of them to curve, but in a closed but still flat spacetime, it is possible for two straight world lines to meet more than once. For example, Fig. 3.17 shows a simple two-dimensional spacetime that is the surface of a cylinder (which you'll recall we argued earlier is flat), rather than an infinite flat plane. The two world lines in that figure are *both* straight: to visualize this, imagine cutting the cylinder open along the (vertical) time dimension, and then flattening it out. Bob's world line, however, looks longer in the spacetime diagram, so Bob's proper time will be less than Bill's when they meet again, even though now *neither* of them has experienced any acceleration.[112] This is simply an interesting mathematical exercise, however, and as far as is known the spacetime we live in is not cylindrical and so Bob's trip into the future will require an accelerating rocketship.[113]

If we specify the details of Bob's trip, we can then precisely calculate the difference in elapsed time for the twins. In an analysis that dates back to 1962, the German astrophysicist Sebastian von Hoerner (1919–2003) did that for the following trip[114]: To begin, Bob gets into his rocket ship at time $t = t' = 0$ (t is time measured on Earth by Bill, and t' is time measured by Bob in his rocket). The Bill and Bob synchronize their clocks at the instant of departure. Bob's trip is to be made in comfort, and so his rocket accelerates at a constant rate (a one gee acceleration, for example, would be equivalent to Earth's gravity, and Bob would feel right at home). This is of practical importance, obviously, because we do not want the experienced acceleration to be incompatible with the physical survival of Bob. Bob

---

[112] See C. H. Brans and D. R. Stewart, "Unaccelerated-Returning Twin Paradox in Flat Space-Time," *Physical Review D*, September 15, 1973, pp. 1662-1666. For a similar treatment, this time by a mathematician, see Jeffrey R. Weeks, "The Twin Paradox in a Closed Universe," *American Mathematical Monthly*, August-September 2001, pp. 585-589.

[113] In Chapter 6 we'll discuss the idea of traveling into the past by moving faster than light (*superluminal* motion). A treatment of such travel, in Bob's cylindrical spacetime, is by S. K. Blau, "Would a Topology Change Allow Ms. Bright to Travel Backward in Time?" *American Journal of Physics*, March 1998, pp. 179-185, which answers that question in its last line: "Ms. Bright cannot [return] 'the previous night' and alter history," a conclusion that no doubt met with Hawking's approval. The 'Ms. Bright' in the title is the heroine of a 1923 limerick that you can find quoted in the first *For Further Discussion* of Chapter 6.

[114] Originally appearing in the journal *Science*, under the title of "The General Limits of Space Travel," von Hoerner's analysis was reprinted in the classic anthology *Interstellar Communication* (A. G. W. Cameron, editor), W. A. Benjamin 1963. The arithmetic was, alas, just a bit sloppy (the final formulas, fortunately, are correct), and many of von Hoerner's numerical evaluations are incorrect. Later, the British mathematician Leslie Marder cleaned-up the analysis in his beautiful little book on the twin paradox, *Time and the Space-Traveler*, George Allen & Unwin, Ltd. 1971.

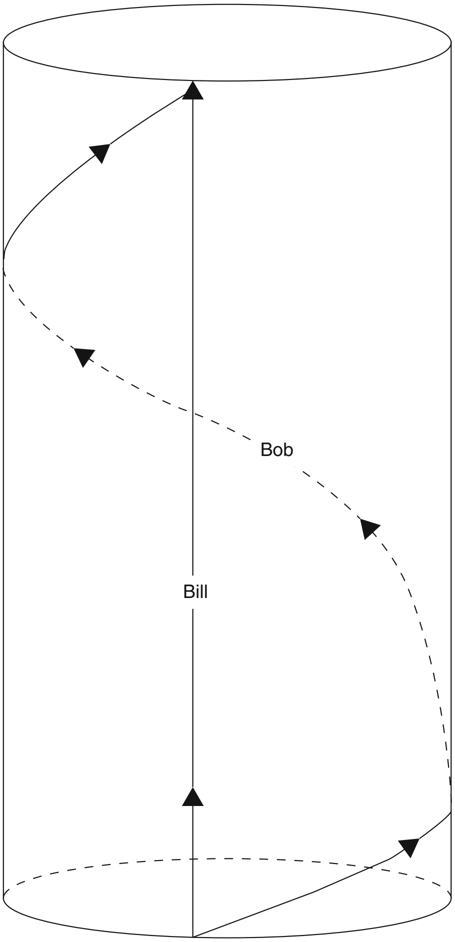

Fig. 3.17 Unaccelerated twin paradox in a cylindrical spacetime

travels this way for a time interval of T (as measured by Bill on Earth) and T' (as measured by Bob in his rocket). At that time the rocket is traveling at its maximum speed. Bob then turns off the rearward engine and turns on a forward-mounted engine so as to experience a constant deceleration. Floor and ceiling interchange, but Bob always weighs the same. If he does this for the same time interval T (as measured by Bill on Earth) and T' (as measured by Bob) as for the initial acceleration phase of the trip, the rocket will be brought to rest with respect to Earth. At that time, $2T$ (as measured by Bill on Earth) and $2T'$ (as measured by Bob), the rocket is at its maximum distance from Earth. Bob then returns to Earth, using the same acceleration/deceleration process, and Bob arrives back home, gently, with a final speed of zero with respect to Earth (ignoring, of course, all the navigational problems due to the motion of Earth during the trip).

Assuming the clock hypothesis is true (not all physicists believe this,[115] but remember, we have Einstein[116] on our side with this!), the result is that if a is the constant acceleration of the rocket (as experienced by Bob), and c is the speed of light, then the relationship between the *total roundtrip time* as measured by Bill ($4T$) and as measured by Bob ($4T'$), is given by

$$T = \frac{c}{a}\sinh\left(\frac{a}{c}T'\right).$$

The difference in what each twin believes to be the date can, in fact, be truly astonishing. For example, if $a = 1$ gee and Bob travels (by his clock) for $4T' = 20$ years, then $4T = 339$ years.[117] Of course, there will actually be no disagreement between Bill and Bob over the date upon Bob's return because Bill will be long dead!

Is it likely that such a time trip[118] will someday be made into the future? It's just my opinion, but I suspect not. Bob will be traveling at virtually the speed of light at maximum speed ($0.9993\,c$, to be precise, at which point he will be 84 light-years from Earth), and to zip through space at such a speed would result in a very high rate of collision with stray hydrogen atoms (about one per cubic centimeter). The result of those energetic collisions would be the intense blasting of Bob with a lethal dose of gamma radiation. And, as von Hoerner showed in his original analysis (note 114), the energy required by Bob's rocket would be simply mind-boggling.

Professor Schild (see note 116), on the other hand, seems to have been less bothered by such considerations when he wrote, at the beginning of a prose discussion of the twin paradox, "I have no doubt that if our technology should ever advance to the stage where large-scale twin effects become noticeable with our unaided senses, then [people] will have no difficulty in adjusting their concepts of time until the new phenomena see quite natural."[119]

I'll end this chapter on time with the observation that it is the distinction between the proper time of Bob on a rocket ship, and the time of those who are not fellow space travelers, that eliminates the occasional suggestion of simply *freezing* one's way via suspended animation to the future (as in the 1992 film *Forever Young*). That is, suppose you manage to talk a friend into climbing into a freezer (please don't actually try this!); at the moment he gets in, his wrist watch agrees with yours (both are powered and maintained at a constant temperature by 100-year nuclear batteries). Years later, when you thaw your friend out, you'll find that your watches still agree. But when Bob returns from his rocket trip, his watch will not agree with yours, but instead will be far behind. Bob is a true time traveler, but your frozen/thawed friend is not.[120]

---

[115] Consider, for example, this remark by 1965 Nobel physics laureate Julian Schwinger (1918-1994) about the twin paradox in his 1986 book *Einstein's Legacy*: "The observer on the spaceship . . . is *not* in uniform, unaccelerated motion . . . The special theory of relativity does not apply to such an accelerated observer." Schwinger was wrong in this conclusion (see the next note).

[116] As a physicist wrote on this point after Einstein's death, "A good many physicists believe that [the twin] paradox can only be resolved by the general theory of relativity. . . . However, they are quite wrong. The twin effect . . . is one of special relativity." See A. Schild, "The Clock Paradox in Relativity Theory," *American Mathematical Monthly*, January 1959, pp. 1-18. Alfred Schild (1921-1977) was professor of physics at the University of Texas, and a recognized expert in the general theory.

[117] In using the T, T' formula, one has to be careful to use MKS units, that is, length and time measured in meters and seconds, respectively. Thus, $a = 1$ gee = 9.81 meters/second^2 and $c = 186{,}210$ miles/second $= 2.997 \times 10^8$ meters/second.

[118] For a science fiction use of such a trip, see Robert Heinlein's 1956 novel *Time for the Stars*.

[119] A. Schild, "Time," *Texas Quarterly*, Autumn 1960, pp. 42-62.

A provocative illustration of the distinction between the proper time of a time traveler and of those who are not time travelers was given by a philosopher.[121] He begins his analysis by asking what appears to be a question with an obvious answer. Suppose, he says, it is [2018] and you suddenly wake up in a hospital and are told that you have been in a coma for the past 2 weeks. You are also told that you were in an auto accident 2 weeks ago, that you suffered temporary neural damage, and that the eventual reversal of such damage always, at some time within 4 weeks after the damage occurs, causes a day of excruciating pain *if you are conscious at the time*. Would you prefer for the day of damage reversal to be in the past 2 weeks (when you were in the coma) or in the next 2 weeks? The answer seems clear. After all, if the day that damage reversal occurs has already happened, then you simply slept through it and missed the pain. To prefer the day of pain to be in the future (when you will be awake) seems absurd. Now, let's add time travel to the equation.

All is as before, but now you immediately leave the hospital upon regaining consciousness to take a trip back to 1892, where you will stay for 2 weeks. Again, it seems clear that you would prefer to have had the day of pain in the past 2 weeks (in 2018), not in the next 2 weeks (in 1892). Note that the *next* is a reference to your proper time, because whereas 1892 is the global past, it is *your* personal future. Thus, now your preference would be to have the day of pain in the *recent personal past* of 2018, not the *distant global* past of 1892. Now, let's put another time travel twist to this story.

All is as before in the original tale, except now you are told in the hospital that the auto accident happened just after you made a time trip to 2092: as you walked out of your time machine in 2092, you were hit by a car. The 2 weeks you were in a coma were in 2092, before you were judged fit enough (although still unconscious) to make the time journey back to 2018. When would you now prefer to have the day of pain? Clearly, as always, in your *personal past*, which is now the *global future*.

Time *is* different for those who time travel and those who don't!

---

[120] In his 1956 novel *The Door Into Summer*, the always ingenious Robert Heinlein used *both* ideas, with the cold-storage method of reaching the future combined with a true time machine to allow his hero to return to his 'present' (the future's past).

[121] A. Gallois, "Asymmetry in Attitudes About the Nature of Time," *Philosophical Studies*, October 1994, pp. 51-59.

3.7 For Further Discussion

An interesting *theological* analysis of Feynman's idea of an electron traveling backward in time, expressed as thinking of the electron as being in 'one of God's films played in reverse,' is given by J. W. Smith, "Time Travel and Backward Causation," *Cogito* 1985, pp. 57–67. Wondering what it would be like for an electron to travel backwards in time, Smith's answer was: "Consider an electron e_1. At time t_0 it is at (x_0, y_0, z_0). At time t_1 it is at (x_1, y_1, z_1). If the direction of time for the electron was reversed, then the electron would be observed on the 'film of the world' to travel back along the same path as it did before, i.e., back to (x_0, y_0, z_0). If God stopped the 'film of the world' and examined the charge of e_1, then He would find that it was *negative*, not *positive*. Hence the electron traveling backwards in time is simply that: an electron traveling backwards in time, it is not a positron. Time reversal does not result in a reversal of charge. Thus, the Stückelberg-Feynman position is incorrect ...." Discuss this in terms of the TCP theorem. Is there a conflict? When God stops the 'film of the world,' does the electron even *have* an arrow of time?

All the modern, major religions of the world are in agreement on these two points: (1) God created the universe and (2) At some time in the past, the universe came into existence. This does raise the question of what was God doing *before* He created the universe (see note 7 in the Introduction). In his *Confessions*, Saint Augustine comments on the conundrum that (1) and (2) are possibly in conflict. After all, if God *created* the universe then, given any time t in the finite past, He must have been doing something *before* time t, which means that for any time t in the finite past the universe already existed. Thus, the universe had no instant of creation in the finite past and so had no first moment of existence—which implies (2) is false. However, like any good philosopher of religion, Augustine not only provided this theological puzzle, but also a way to wiggle free of it. His suggested counter is the assertion that time is *itself* a creation of God, that is, He made time as *part* of creating the universe. Thus, there was no time *before* He created the universe, and so the very question of 'what was He doing before He made the universe' has no meaning. What do you think of Augustine's two arguments?

The Australian philosopher J. J. C. Smart (1920–2012) invoked five dimensional spacetime in a way very different from that of including an ‘eternity’ axis for God’s temporal time (see note 33 in “Some First Words”). What Smart argued for, instead, was *multiple* four-dimensional worlds existing together without conflict, just as an infinity of two-dimensional worlds can exist without conflict in a three dimensional space. As he wrote, “The reason why there could be two totally disparate space-times is simply the quite obvious one that two totally disparate four-dimensional spaces can exist within a suitable five-dimensional space. There is no difficulty in mathematical inconceivability here. Now let one of these four-spaces be our own spacetime world, and let the other four-space be more or less similar, in accordance with whatever story you wish to tell about it.” This idea had, in fact, been around long before Smart’s 1967 paper. In his 1898 Presidential Address to the American Mathematical Society, for example, Simon Newcomb declared “Add a fourth dimension of space, and there is room for an indefinite number of Universes, all alongside of each other, as there is for an indefinite number of sheets of paper when we pile them upon each other.” Newcomb’s idea appealed to H. G. Wells’ fancy so much that he built two novels, *The Wonderful Visit* and *Men Like Gods*, around it. In the first novel there is explicit mention of multiple worlds “lying somewhere close together, unsuspecting, as near as page to page in a book,” and the second one speaks of one parallel universe being rotated into another. John Cramer (a University of Washington physicist) repeated Newcomb’s and Wells’ parallel universes/pages-of-a-book/rotation imagery almost word for word in his 1991 novel *Twistor*.

Read these three novels, and then discuss. In particular, is Wells consistent in his presentations (as a novelist there is, of course, no reason he should be!)?

Imagine the following two events, **A** and **B**, in Minkowski spacetime: **A** is the emission of a photon, and **B** is the absorption of that photon. What is the spacetime interval between **A** and **B**? It might seem that we need to know more about the precise spatial and temporal coordinates of **A** and **B**, but in fact the interval is *always zero* for any two events connected by light (by photons). To see this, write the flat Minkowski metric as

$$\left(\frac{ds}{dt}\right)^2 = 1 - \left(\frac{dx}{dt}\right)^2$$

(continued)

and so, since $(dx/dt)^2 = 1$ (because photons—by definition!—travel at the speed of light) we have $(ds)^2 = 0$. The world line of any photon has what is called a *null* interval, as do all world lines on the surface of a light cone. Timelike world lines (in the interior of light cones, with $(dx/dt)^2 < 1$) have positive intervals, that is, $(ds)^2 > 0$. Spacelike world lines (in the exterior of light cones, in Elsewhere, with $(dx/dt)^2 > 1$) have negative intervals, that is, $(ds)^2 < 0$. This is one of the significant differences between *distances* in space (which are never negative) and *intervals* in spacetime. The American chemist G. N. Lewis (1875–1946)—see note 60 in Chap. 2—constructed a romantic illustration of this when he wrote "Any pair of points [in spacetime] which are separated by zero distance [interval] are in *virtual contact*. In other words, I may say that my eye touches a star, not in the same sense as when I say that my hand touches a pen, but in an equally physical sense." (See Lewis' paper "Light Waves and Light Corpuscles," *Nature*, February 13, 1926, pp. 236–238; Lewis was the originator of the term *photon* for a particle of light.) To understand what Lewis was getting at, you must understand that in spacetime we can have the interval between **A** and **B** as zero, and the interval between **B** and **C** as zero, but the interval between **A** and **C** may *not* be zero. To convince yourself of this, suppose the Minkowski spacetime coordinates (x, t) of **A**, **B**, and **C** are (1,3), (2,2) and (1,1), respectively. Show that $(ds_{\boldsymbol{AC}})^2 = 4$, while $(ds_{\boldsymbol{AB}})^2 = (ds_{\boldsymbol{BC}})^2 = 0$. (Hint: draw a diagram of this spacetime and simply plug the coordinates of **A**, **B**, and **C** into the metric.)

The use of the time dilation effect of high-speed space travel was used by science fiction writer Donald Wandrei (1908–1987) in his tale "A Race Through Time," *Astounding Stories*, October 1933. Initially set in 1950, this is the story of two scientists, one evil (of course!) and the other good (of course!), who develop quite different methods for travel into future times. The evil one does it with a drug that slows the metabolic processes of the body, while the good one builds an atomic-powered rocket in his home workshop! The evil scientist kidnaps the good one's girlfriend, seals the two of them inside a crystal dome, and then injects her and himself with his drug. He has arranged matters so that they will emerge from the dome in the year one million A.D. Learning what has happened, the good scientist rushes to the dome finds he can't break in, but sees an indicator dial pointing at 1,000,000. (The evil scientist has conveniently provided the dial, as well as having made the dome transparent, much as modern-day movie criminals always include a count-down clock with glowing red digits on their bombs so the hero always knows just how much time is left to disarm the bomb.)

(continued)

Returning to his rocket, the good scientist decides that he, too, will travel into the far future using time dilation. So, off he goes, on a trip like Bob's in Sect. 3.6. The story ends with an ironic twist—the good scientist thought the '1,000,000' he saw through the dome meant *one million years in the future beyond 1950*, and so he arrives back on Earth nearly 2000 years *after* his girlfriend and the evil scientist emerged from the dome. (They are, of course, long dead when the good scientist returns.)

On the outward leg of the good scientist's rocket flight we read of his "frightful speed — now thousands of light years per Earth second." Discuss this in terms of relativity theory. That is, does the rocket actually travel thousands of light years in one second of elapsed time back on Earth?

The American physicist Robert Forward (1932–2002) was, in addition to being an expert in general relativity, also a quite inventive science fiction writer. In his short story "Twin Paradox" (*Analog Science Fiction*, August 1983), for example, he used *biology* to give a surprising, ironic twist to the classic physics puzzle. The story flips the asymmetric aging of the twins by imagining that, just after the traveling twin's departure, the secret of immortality is discovered. The treatment has to be administered no later than at a certain age, however, and upon his return to Earth the traveling twin is just a bit too old for it to work. He thus becomes the last person to die of old age! In this tale, the details of the traveling twin's trip are somewhat different from Bob's trip. Read "Twin Paradox" (you can also find the story in Forward's 1995 book *Indistinguishable from Magic*), and summarize how the traveling twin's trip is accomplished.

In the mystical 1920 novel *A Voyage to Arcturus*, by the Scottish writer David Lindsay (1876–1945), we read of a spaceship that travels to Arcturus (the brightest star in the constellation Boötes, 36 light years from Earth) in just 19 h of proper time. The technical details of the trip are not explained in the novel, so assume they are the same as in Bob's trip. That is, $2T' = 68{,}400$ s as measured by a clock on the spaceship. The distance traveled in this time is

$$2\frac{c\left\{\sqrt{a^2T^2+c^2}-c\right\}}{a}$$

(continued)

where c is the speed of light, a is the constant acceleration/deceleration of the rocket (as measured on the spaceship), and T is how long an *Earth*-based clock says the trip takes. Use this formula, and the one relating T and T' in the text, to calculate

1. The value of the constant acceleration/deceleration, a (does the result seem reasonable to you?) and
2. The value of T (the length of time that passes on Earth as 19 h pass on the spaceship).

You don't have to discuss general relativity to encounter tensors of the second rank (as is the metric tensor). Electrical engineers run into such a thing, for example, when studying the lowly Ohm's law! In a copper wire, that law says the current density (the vector **J**, in units of amperes per square meter) at any point is related to the electric field (the vector **E**, in units of volts per meter) at that point by the *scalar* σ as follows: $\mathbf{J} = \sigma\mathbf{E}$, where σ (called the *conductivity*) is a single number. This says, in rectangular coordinates, that the x-component of **J** depends *only* on the x-component of **E**, and similarly for the y-component and z-components of **J** and **E**. More generally, however, each component of **J** depends on *all* of the components of **E** (as in certain crystalline structures), and so we have the equations

$$J_x = \sigma_{11}E_x + \sigma_{12}E_y + \sigma_{13}E_z$$
$$J_y = \sigma_{21}E_x + \sigma_{22}E_y + \sigma_{23}E_z$$
$$J_z = \sigma_{31}E_x + \sigma_{32}E_y + \sigma_{33}E_z$$

or, in matrix form

$$\mathbf{J} = \begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = \boldsymbol{\sigma}E.$$

So **J** is now related to **E** by $\boldsymbol{\sigma}$, a 3 × 3 *matrix* (9 numbers, instead of just 1). The matrix $\boldsymbol{\sigma}$ is, in fact, a tensor of rank 2. This tensor is in the three-dimensional space of the copper wire, while the 4 × 4 metric tensor matrix is in a four-dimensional spacetime, but both are tensors of rank 2. The number of numbers in a tensor of rank n, in a space of dimension d, is d^n. In general relativity, tensors of higher rank than 2 are required. For example, the

(continued)

curvature tensor has rank 4, and so in a four-dimensional spacetime it is described by $4^4 = 256$ numbers. This goes a long way in explaining why, in general relativity, computing often involves a *lot* of arithmetic! The primary characteristic of tensors of any rank is that they are invariant under a change in coordinate systems (choosing a coordinate system is an arbitrary matter, made mostly for human convenience, about which Nature is indifferent). Read more about tensors, and write an essay on how they behave under a change in coordinates. A good place to start, at the level of this book, is Lillian Lieber's *The Einstein Theory of Relativity: a trip to the fourth dimension*, Paul Dry Books 2008 (an updated version of the original 1945 book).

Chapter 4
Philosophers, Physicists, and the Time Travel Paradoxes

> *"He felt the intellectual desperation of any honest philosopher. He knew that he had about as much chance of understanding such problems as a collie has of understanding how dog food gets into cans."*[1]

4.1 Paradoxes and Their First Appearance in Science Fiction

> "There's a lot we don't know about time travel. How do you expect logic to hold when paradoxes hold, too?" "Does that mean you don't know?" "Yes."[2]

More than 30 years ago Quentin Smith, a philosopher who believes in a finite length to the past, wrote a refutation to those who believe in an infinite past and, while that paper[3] has nothing to do with the paradoxes of time travel, in the course of presenting his reasoning he included the following curious passage:

> "Why does the sun arise in the morning and not at some other time? Why do the hands of a properly functioning clock point to 12:00 at noon and midnight and not at other times? Why does the death of a person occur at a later time than his birth? The answer in all these cases is: Because by the very nature of these events they could not occur at other times. It belongs to the very nature of the sun's rising that it occur in the morning and not in the afternoon or evening. It belongs to the very nature of a properly functioning clock to point to 12:00 at noon and midnight and not at other times. And it belongs to the very nature of death to occur at a time later than a person's birth."

But what of a time traveler born in 1980 who, in 2018, enters her time machine, pushes a few buttons, and then boldly steps out into the Cretaceous period seventy million years earlier—and is promptly eaten for lunch by a passing *Tyrannosaurus*

---

[1] A time traveler admits (to himself) how perplexed he is by paradoxes in Robert Heinlein's classic tale "By His Bootstraps," *Astounding Science Fiction*, October 1941.

[2] Excerpt from a conversation between two paradox-puzzled time travelers in Larry Niven's story "Bird in the Hand," *The Magazine of Fantasy & Science Fiction*, October 1970.

[3] Q. Smith, "Kant and the Beginning of the World," *New Scholasticism*, Summer 1985, pp. 339–346.

© Springer International Publishing AG 2017
P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_4

rex? Perhaps Smith himself would say that there is no contradiction between this and his third claim because, in the time traveler's *proper* time, her spectacular death does indeed come after her birth. For many, however, for a time traveler to die before her mother is born *is* a paradox, plain and simple, say what you will about proper time.

One science fiction view of time travel paradoxes is that nature would be so disrupted by them that, should one occur, the universe would be torn apart. In one story, for example, a paradox is on the verge of happening through the use of a Tipler-cylinder time machine (mentioned back in Chap. 1, and which we'll discuss in more detail in the next chapter). In response to this 'threat to common sense,' the universe 'decides' to avoid the paradox by simply eliminating the perpetrators via a local nova![4] Niven wasn't the first to use this idea. A famous story[5] by L. Sprague de Camp (1907–2000), written two decades before, had already put forth the suggestion that nature will take all required corrective action to avoid paradoxes.

In De Camp's tale we read of two big-game hunter guides who use "Professor Prochuska's time machine at Washington University"—built with the aid of a "cool thirty million" dollar grant from the Rockefeller Foundation—to operate a safari-for-hire business that transports hunters back to the late Mesozoic era. When a disgruntled client tries to go back to the day *before* a previous trip to shoot the guides (who had displeased him, or rather *would* displease him the next day), we learn just how nasty De Camp thought Mother Nature would be to avoid a paradox. (After all, the guides had *not been* shot during the safari, so they *could not be* shot.[6]): "The instant James started [to ambush the guides] the space-time forces snapped him forward to the present to prevent a paradox. And the violence of the passage practically tore him to bits [making his body look] as though it had been pulverized and every blood vessel burst, so it was hardly more than a slimy mass of pink protoplasm."

And even earlier we have a famous tale[7] that was discussed in Chap. 1, by Fredric Brown (1906–1972), a master of the special category of science fiction called the "short-short," in which everything happens in 500 words or less. As you might expect, the oddities of time travel were natural attractions for Brown's quirky talent. You'll recall that in this story the inventor of the first time machine demonstrates it to two colleagues by sending a brass cube 5 min into the future. After being placed in the machine, the cube vanishes and then 5 min later reappears. No paradoxes with that—it is a trip into the *past* that has the potential for deadly repercussions. We learn just how deadly when the inventor next declares that at three o'clock he will again place the cube in the time machine but, until then, he will hold the cube in his hand. Thus, he says, at 5 min before three the cube will vanish from his hand and immediately appear in the time machine because 5 min after that, at three o'clock, he will send it 5 min into the past. And indeed, at 5 min before three the cube does simultaneously vanish from his hand and appear in the time machine. Then, slightly before three, as the three men stand pondering what has happened, one of the observers asks what will happen if the inventor does *not* send the cube back at three o'clock? "Wouldn't there be a paradox of some sort involved?" he wonders.[8] His curiosity aroused, the inventor can't resist the experiment—and the universe promptly vanishes.

---

[4] L. Niven, "Rotating Cylinders and the Possibility of Global Causality Violation," *Analog Science Fiction*, August 1977. Niven took this title from a physics paper with that title, authored by Tipler, that had appeared three years earlier in *Physical Review D* (April 15, 1974, pp. 2203–2206).

[5] L. Sprague de Camp, "A Gun for Dinosaur," *Galaxy Science Fiction*, March 1956.

[6] This is a statement of the belief that the past cannot be changed, an idea we will examine later in this chapter.

[7] Look back at note 93 in Chap. 1.

Time travel, of course, is full of potential paradoxes. A paradox, according to the usual dictionary definition, is something that appears to contain contradictory or incompatible parts, thus reducing the whole to seeming nonsense. And yet, truth is also evident in the whole. The history of science and mathematics has left a long trail of paradoxes, and those that involve time travel are merely among the most recent. Not all of the puzzles of time travel involve physics or logic, however. As one philosopher observed, "Doubtless time travel will raise a host of legal difficulties, e.g., should a time traveler who punches his younger self (or vice versa) be charged with assault? Should the time traveler who murders someone and then flees into the past for sanctuary be tried in the past for his crime committed in the future? If he marries in the past can he be tried for bigamy even though his other wife will not be born for almost 5000 years? Etc., etc. I Ieave such questions for lawyers and writers of ethics textbooks to solve."[9]

One way early science fiction writers had of responding to the puzzle of time travel paradoxes was to just give up and to concede that the logical puzzles are overwhelming. In one tale, for example, the inventor of the Chronoscope (a gadget that can only *view* the past) explains, "There is no time travel machine. Such a thing is a logical impossibility, treated seriously only by half-cracked writers of fantasy. Such a machine would lead at once into a hopeless paradox."[10] Equally concerned about time travel paradoxes was the pulp science fiction time traveler who told his partner, just before their first trip in time, that "I'm not sure any more about getting back. There're some unpredictable terms in the time-travel equation—paradoxes. Maybe we *won't* get back."[11]

---

[8] This is what is called a *bilking paradox*, and such paradoxes will be discussed later in this chapter. Brown gave this story a lot of thought. At one point in the tale one of the colleagues, puzzled by how the inventor will be able to place the cube into the time machine at three if it has already vanished from his hand and appeared in the machine, asks "How can you place it there, then?" Replies the inventor, "It will, as my hand approaches, vanish from the [machine] and appear in my hand to be placed there."

[9] L. Dwyer, "Time Travel and Some Alleged Logical Asymmetries Between Past and Future," *Canadian Journal of Philosophy*, March 1978, pp. 15–38.

[10] M. Jameson, "Dead End," *Thrilling Wonder Stories*, March 1941.

[11] E. Binder, "The Time Cheaters," *Thrilling Wonder Stories*, March 1940.

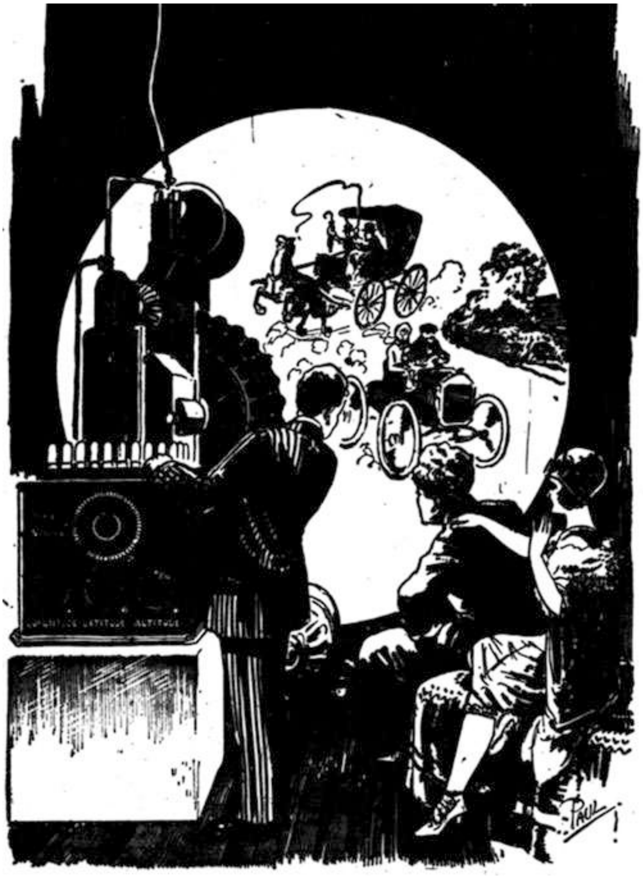

Fig. 4.1 One way pulp science fiction avoided paradoxes was to use a 'time viewer' (like the chronoscope in "Dead End") as in "The Time Eliminator" (*Amazing Stories*, December 1926). This illustration from the story (authored by somebody who used only the initials K.A.W.) by Frank R. Paul (©1926 by Experimenter Publishing Co.) shows the inventor demonstrating his gadget to his future wife and father-in-law. Able to look back in time, the screen is displaying scenes from the older man's courtship of his wife, decades in the past. Reprinted by permission of the Ackerman Science Fiction Agency, 2495 Glendower Ave., Hollywood, CA 90027 for the Estate

But are there *really* paradoxes? Or is it true, as the extraordinary boy-prodigy who invented a time machine exclaimed (when his teacher asserted that some questions could never be answered because "Nature is full of paradoxes"), "Ah, Professor, what nonsense! Nature is harmonious; it is we who bring the paradoxes into it."[12] Saying the same are two physicists, in a paper on the circular orbits of photons around black holes: "There are no paradoxes in physics, but only in our attempts to understand physical ideas by using inadequate reasoning or false intuition."[13] And as the time traveler in an early pulp story[14] casually declares to a friend, after an astonishing adventure in the year A.D. 1,001,930, "Paradoxical? My dear fellow, the Einstein Theory is full of apparent paradoxes, yet to him who understands it there is no inconsistency whatever. Give me another cigarette, will you, Frank?"

---

[12] V. Grigoriev, "Vanya," in *Last Door to Aiya* (M. Ginsburg, editor), S. G. Phillips 1968.

It is in *Amazing Stories* that we find the first non-fictional speculations about time travel in a pulp magazine—and certainly long before any *physics* journal would touch the subject! Publisher and editor Hugo Gernsback started those speculations by reprinting Wells' *Time Machine*, which in turn sparked a fair number of readers' letters that were printed in the magazine's "Discussions" section. Typical of the less interesting is the following comment from a letter in the July 1927 issue: "In the 'Time Machine' I found something amiss. How could one travel to the future in a machine when the beings of the future have not yet materialized?" (We answered that question in the previous chapter with the twin paradox.) Far more interesting was this letter, in the same issue:

> "How about this 'Time Machine'? Let's suppose our inventor starts a 'Time Voyage' backward to about A.D. 1900, at which time he was a schoolboy … His watch ticks forward although the clock on the laboratory wall goes backward. Now we are in June 1900, and he stops the machine, gets out and attends the graduating exercises of the class of 1900 of which he was a member. Will there be another 'he' on the stage? Of course, because he *did* graduate in 1900 … Should he go up and shake hands with this 'alter ego'? Will there be two physically distinct but characteristically identical persons? Alas! No! He can't go up and shake hands with himself because … this voyage back through time only duplicates actual past conditions and in 1900 this strange 'other he' did *not* appear suddenly in quaint ultra-new fashions and congratulate the graduate. How could they both be wearing the same watch they got from Aunt Lucy on their seventh birthday, the same watch in two different places at the same time. Boy! Page Einstein! No, he cannot be there because he wasn't there in 1900 (except in the person of the graduate) … The journey backward must cease on the year of his birth. If he could pass that year it would certainly be an effect before a cause … Suppose for instance in the graduating exercise above, the inventor should decide to shoot his former self … He couldn't do it because if he did the inventor would have been cut off before he began to invent and he would never have gotten around to make the voyage, thus rendering it impossible for him to be there taking a shot at himself, so that as a matter of fact he *would* be there and *could* take a shot—help, help, I'm on a vicious circle merry-go-round … Now as to trips into the future, I could probably think up some humorous adventures wherein [the inventor] digs up his own skeleton and finds by the process of actual examination that he must expect to have his leg amputated because the skeleton presents positive proof that this was done."[15]

---

[13] M. A. Abramowicz and J. P. Lasota, "On Traveling Round Without Feeling It and Uncurving Curves," *American Journal of Physics*, October 1986, pp. 936–939.

[14] F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931.

[15] This story idea (the letter was signed only with the initials T.J.D.) may well have been the inspiration for R. Rocklynne, "Time Wants a Skeleton," *Astounding Science Fiction*, June 1941. Not all fans agreed with T.J.D. A few years later, for example, a teenager named P. Schuyler Miller (1912–1974), who would author several time travel classics himself, wrote a letter to the editor of *Astounding Stories* (June 1931) stating "there is nothing in physics … to prevent yourself from

All of the ingenious puzzles in this letter intrigued Gernsback, and may have, in fact, been the cause of his featuring a new, original time machine story[16] in the same issue. It was the tale of a scientist who transports an entire ship at sea 14,000 years back in time and causes it to hover over lost Atlantis! That story provoked a sharp letter from a reader who claimed its logic had a fatal flaw: the story's author indicated that the Atlantians observed the time travelers when, 'of course' (asserted the reader), the time travelers must actually have been invisible. The reader explained his reasoning as follows, beginning by defining A as one of the Atlantians.

> "Now A lived his life, thousands of years ago, and died. All right, now let us pass on in time 14,000 years. Now, back we come in time when A is again living his life. Lo and behold, this time A sees before he dies a strange phenomenon in the sky! He sees the shipload of people observing him. And yet these people are necessarily observing him during his one and only lifetime, wherein he certainly did not, could not, have observed them."

Gernsback printed this letter in his September 1927 editorial "The Mystery of Time," and concluded by saying "I do . . . agree . . . that the inhabitants of Atlantis would probably not have seen the . . . travelers in time." Other readers felt this way, too, because after Gernsback published yet another time machine tale[17] in 1927, the same invisibility argument appeared again in the magazine's "Discussions" column.

Two years later an amateurishly written tale[18] appeared in which a man travels in time from 1928 to 2930 with the aid of an "astounding machine based on advanced electro-physics and the non-Euclidean theory of hyperspace." The purpose of that story was two-fold: to present several of the classic paradoxes of time travel, and then to make the claim that although the simple minds of twentieth-century people cannot understand the explanations of the paradoxes (possibly explaining why the author offers none!), the paradoxes are all trivial to the scientists of the thirtieth century. Despite this shortcoming (as well as some pretty awful dialog) the story nonetheless still managed to entertain readers with the sheer mystery of the paradoxes. Letters poured into the magazine from young fans, all demanding more time travel fiction.

So, that same year Gernsback responded with a story[19] that plays with the question of the role of time travelers in the past. (That question was clearly 'in

---

going into the past . . . and shaking hands with yourself or killing yourself." That did, however, provoke the following harsh reply from another, more skeptical reader (in the December 1933 issue): "P. S. Miller once wrote that time traveling is not incompatible with any laws of physics . . . 'he don't know from nothing.'"

[16] C. B. White, "The Lost Continent," *Amazing Stories*, July 1927.

[17] F. Flagg, "The Machine Man of Ardathia," *Amazing Stories*, November 1927.

[18] C. Cloukey, "Paradox," *Amazing Stories Quarterly*, Summer 1929.

[19] H. F. Kirkham, "The Time Oscillator," *Science Wonder Stories*, December 1929.

the air,' as Gernsback's old magazine *Amazing*[20] simultaneously published a story[21] addressing this same puzzle of time travel.) Could time travelers actually participate in events ("mix into the affairs of the period," in Gernsback's words), or would they just be unseen observers? This question, obviously inspired by the earlier discussion in *Amazing Stories*, intrigued Gernsback as much as it did his readers and so, along with Kirkham's story, he printed a challenge titled "The Question of Time-Traveling" (see note 18 in the Introduction):

> "In presenting this story to our readers, we do so with an idea of bringing on a discussion as to time traveling in general. The question in brief is as follows: Can a time traveler, going back in time—whether ten years or ten million years—partake in the life of that time and mingle in with its people; or must he remain suspended in his own time-dimension, a spectator who merely looks on but is powerless to do more? Interesting problems would seem to arise, of which only one need be mentioned: Suppose I can travel back into time, let me say 200 years; and I visit the homestead of my great great great grandfather, and am able to take part in the life of his time. I am thus enabled to shoot him, while he is still a young man and as yet unmarried. From this it will be noted that I could have prevented my own birth; because the line of propagation would have ceased right there. Consequently, it would seem that the idea of time traveling into a past where the time traveler can freely participate in activities of a former age, becomes an absurdity. The editor wishes to receive letters from our readers on this point: the best of which will be published in a special section."

Gernsback's challenge did not go unnoticed and, over the next year or so, he published a large number of reader responses in the magazine's letters column.

Indeed, a few months after issuing the challenge, in his introduction to another time travel tale,[22] Gernsback wrote that ever since the publication of Kirkham's tale "there has been a great controversy among our readers as to the possibility of time flying and the conditions under which it may be done." Most of those letters are interesting if not particularly profound, with one exception. That was a letter written by a 14 year old boy in San Francisco, and its appearance in the February 1931 issue of *Science Wonder* may well have served as the inspiration for several of the classic time travel tales published during the next 20 years:

> "Some time ago you asked us (the readers) what our opinions on time traveling were. Although a bit late, I am now going to voice four opinions . . .
>
> (1) Now, in the first place if time traveling were a possibility there would be no need for some scientist getting a headache trying to invent an instrument or 'Time-Machine' to 'go back and kill grandpa' (in answer to the age-old argument of preventing your birth by killing your grandparents I would say: 'who the heck would want to kill his grandpa or gandma!'[23]) I figure it out thusly: A man takes a time machine, and travels into the

---

[20] By this time Gernsback had lost control of *Amazing*, and *Science Wonder* was his come-back as a publisher of pulp 'scientifiction.'

[21] E. L. Rementer, "The Time Deflector," *Amazing Stories*, December 1929. Gernsback may well have been the editor, before he lost *Amazing*, who bought this story, and the magazine's new management simply used what remained in inventory.

[22] F. Flagg, "An Adventure in Time," *Science Wonder Stories*, April 1930.

[23] Look back at note 26 in "Some First Words."

future from where he sends it (under automatic control) to the past so that he may find it and travel into the future and send it back to himself again. Hence the time machine was never invented, but!—from whence did the machine *come*?

(2) Another impossibility that might result could be: A man travels a few years into the future and sees himself killed in some unpleasant manner,—so—after returning to his correct time he commits suicide in order to avert death in the more terrible way which he was destined to. Therefore how could he have seen himself killed in an entirely different manner than really was the case?

(3) Another thing that might corrupt the laws of nature would be to: Travel into the future; find out how some ingenious invention of the time worked; return to your right time; build a machine, or whatever it may be, similar to the one you had recently learned the workings of; and use it until the time you saw it arrives, and then if your past self saw it as you did, he would take it and claim it to be an invention of his (your) own, as you did. Then—who really *did* invent the consarn thing?

(4) Here's the last knock on time traveling: What if a man were to travel back a few years and marry his mother, thereby resulting in his being his own 'father'?"

Jim H. Nicholson

Gernsback's reply, immediately following this letter, was favorable, opening with "Young Mr. Nicholson does present some of the more humorous [?] aspects of time traveling. Logically we are compelled to admit that he is right—that if people could go back into the past or into the future and partake of the life in those periods, they could disrupt the normal course of events."[24] Nicholson's letter *is* ingenious, and it anticipated the central ideas of a number of science fiction tales yet to be written.[25] However, as you'll see as you read the rest of this chapter, contrary to Gernsback's view Nicholson's comments are *not* logically correct.[26]

---

[24] Despite these words, Gernsback apparently hadn't given up entirely on the 'invisibility of time travelers' view, as he had only a few months earlier published another such tale: R. A. Palmer, "The Time Ray of Jandra," *Wonder Stories*, June 1930. In this story (one either silly or hilarious, take your pick) a time traveler moves into the future by means of a 'time ray.' Unfortunately, the ray works differently on the various chemical elements, and not at all on either hydrogen or oxygen. Thus the time traveler—or at least *much* of him—and his machine do vanish into the future, but left behind are "several gallons of water spilled on the floor." (The human body is about 60 % H_2O.)

[25] For example, Nicholson's item (2) is a precise plot outline for L. Raphael, "The Man Who Saw Through Time," *Fantastic Adventures*, September 1941, and a version of item (4) is in Robert Heinlein's famous "All You Zombies—," *Magazine of Fantasy & Science Fiction*, March 1959.

[26] One cannot, however, fault the imaginative powers of James Nicholson (1916–1972). He eventually became President of American International Films, the company that made such science fiction 'classics' as *Attack of the Crab Monsters* (1957), the 1963 *X (The Man with the X-Ray Eyes)*, and *The Time Travelers* (1964).

4.2 Changing the Past and the Grandfather Paradox

> “I’m not kidding you at all Phil,” Barney insisted. “I have produced a workable Time Machine, and I am going to use it to go back and kill my grandfather.”[27]

As mentioned in “Some First Words,” physicists and philosophers often have quite different approaches to time travel (see note 26 there again). A vivid illustration of that difference is found in a philosophical paper[28] that, after acknowledging the apparent restrictions of the grandfather paradox, turns its attention to a matter that almost surely is beyond the power of physics to study—namely, the nature of the *conversation* between a time traveler and his/her younger self. A physicist, on the other hand, views the restrictions as *the whole point*, because the central question that hovers over *any* discussion of time travel is that of ‘changing the past.’ As one science fiction fan summed it up in a letter to the editor of *Astounding Stories* (August 1931): “It is said that the past cannot be changed, and that any effort to do so would be useless. In my belief, no matter where or when a man goes in the past, if he appears in a year or day that has already gone by, *he is changing the past*. Then there should be no room for doubt: time traveling is impossible. It will never be done.” And certainly killing your grandfather in the past (when he is still a baby) would qualify as a change to the past!

The idea of ‘changing the past’ occurred to the minds of philosophers long before it did in those of science fiction writers or physicists. Four centuries before Christ, the question had already been asked-and-answered by Aristotle. In his *Nicomachean Ethics*, in fact, we find him declaring that the Greek poet Agathon had known the answer a century earlier, and he quotes the poet as saying “Forever God lacks this one thing alone, To make a deed that has been done undone.”

Agathon and Aristotle aside, some medieval theologians argued passionately that the past *could* be changed (but only by God). The eleventh-century Italian cleric Peter Damian (who became a Christian saint) is a famous exponent of that radical view.[29] Writing in his *De Omnipotentia Dei* (“On the Divine Omnipotence in Remaking What Has Been Destroyed and Undoing What Has Been Done"),[30] Damian made it clear that he believed nothing could withstand the power of God, not even the past. Ralph Waldo Emerson's poem "The Past" ("All is now secure and fast, Not the gods can shake the Past") would have been blasphemy for Damian. The following words from Damian testify to the strength of his commitment to a belief in the possibility of changing the past: "Just as we can duly say 'God was able to make it so [that] Rome, before it had been founded, should not have been founded,' in the same way we can equally and suitably say, 'God can make it so that Rome, even after it was founded, should not have been founded.'"[31]

---

[27] The opening line to F. M. Busby, “A Gun for Grandfather,” *Future Science Fiction*, September 1957.

[28] Jiri Benovsky, “Endurance and Time Travel,” *Kriterion—Journal of Philosophy*, 2011, pp. 65–72.

[29] R. P. McArthur and M. P. Slattery, “Peter Damian and Undoing the Past,” *Philosophical Studies*, February 1974, pp. 137–141; P. Remnant, “Peter Damian: Could God Change the Past?” *Canadian Journal of Philosophy*, June 1978, pp. 259–268; R. Gaskin, “Peter Damian on Divine Power and the Contingency of the Past,” *The British Journal for the History of Philosophy*, September 1997, pp. 229–247.

Two centuries after Damian, Aquinas argued the contrary view, that changing the past is *not* within God's power. Whereas Damian felt it impossible to deny any act to God, Aquinas took the far more moderate position that part of God's law is that there be no contradictions in the world (this is, in fact, the modern view of time travel physicists) and that certainly God would be bound by his own law. As he wrote, "It is best to say that what involves contradiction cannot be done rather than that God cannot do it." In his *Paradise Lost*, John Milton's God is constrained even more: he is free to act or not, but if he does freely decide to act, it can only be to 'do right.' That might seem to preclude causing contradiction, as in changing the past, but perhaps not. Milton's contemporary, Thomas Hobbes, declared that there is no *a priori* standard of goodness, and thus (for Hobbes) there are no constraints on God's powers. For Hobbes, therefore, it would seem God could change the past. Theological changing of the past, as you might expect, leads to all sorts of mind-boggling puzzles. Because of such puzzles, theology would certainly be influenced by time travel, but just as certainly theological reasoning will not illuminate the puzzles of the time travel paradoxes.

The question of the immutability of past events is of special interest to theologians because it is directly related to the question of free will versus fatalism. That is, are humans the creators of the future, or are they mere fated puppets of destiny? One way theology gets involved with the issue of 'changing the past' is via what is called the *retroactive petitionary prayer*. (An 'ordinary' petitionary prayer, like the Lord's Prayer in Matthew 6 and Luke 11, asks for something in the present or the future.) Examples of retroactive prayers include that of the surgical patient who prays, just before an exploratory operation, for his tumor to be non-malignant, and that of the soldier's wife who prays that her husband was not among those killed in yesterday's battle. Such prayers are for a happy outcome to an event that is over and done with at the time of the prayer. One might accept the rationality of praying about the future ("Please, God, let me survive tomorrow's battle and I'll be good for the rest of my life") but are prayers about the past even sensible? (The three major monotheistic religions of the world—Christianity, Judaism, and Islam—say *yes*.)

---

[30] This work is in the form of a letter to his friend Desiderius (who later became Pope Victor III), in which Damian rebutted Desiderius' defense of St. Jerome's claim that "while God can do all things, he cannot cause a virgin to be restored after she has fallen." Desiderius thought the reason God could not restore virgins is that he does not want to, to which Damian replied that this meant God is unable to do whatever he does not want to do, but this meant that God would then be less powerful than men, who are able to do things they don't want to do (such as go without food for a month). This is a good example of the dangers involved when getting into debates with theologians.

[31] The Argentinian writer Jorge Luis Borges (1899–1986) was so inspired by Damian's view that the past could be changed that he wrote a short story based on it (see "The Other Death," originally published in *The New Yorker*, November 2, 1968) and put a character in it named after Damian.

In an appendix titled "Special Providences" in his book Miracles, C. S. Lewis answers that question as follows:

> "When we are praying about the result, say, of a battle or a medical consultation, the thought will often cross our minds that (if only we knew it) the event is already decided one way or the other. I believe this to be no good reason for ceasing our prayers. The event certainly has been decided—in a sense it was decided 'before all worlds.' But one of the things taken into account in deciding it, and therefore one of the things that really causes it to happen, may be this very prayer that we are now offering. Thus, shocking as it may sound, I conclude that we can at noon become part causes of an event occurring at ten A.M. (Some scientists would find this easier than popular thought does.)"

Here we see Lewis, a prominent lay theologian, arguing for the present influencing (but not changing) the past. What can we make of that? Was Lewis arguing for backward causation, the close relative of time travel? I think perhaps so; the final two sentences in the above excerpt makes it plausible that he may have held that view. It is a view that does find much support in the block universe interpretation of Minkowskian spacetime. Lewis never mentions the block concept by name, but it is clear that he believed in the idea of God being able to see all of reality at once. Lewis believed, therefore, that God knows of a petitionary prayer before it is made; or, even stronger, if God is not a temporal being but rather is eternal and knows time 'all at once,' then God knows the prayer and the event being prayed about 'at the same time.'

Lewis did make it clear that he believed it is a sin to pray for something *known* not to have occurred—for example, to pray for the safety of someone known to have been killed yesterday. As he wrote, "The known event states God's will. It is psychologically impossible to pray for what we know to be unobtainable, and if it were possible, the prayer would sin against the duty of submission to God's known will." Taking a less judgmental position (but essentially agreeing with Lewis) were two philosophers who, writing of the battle of Waterloo, said "for one who knows the outcome of the battle more than a hundred and 50 years ago, [a retroactive petitionary] prayer is pointless and in that sense absurd. But a prayer prayed in ignorance of the outcome of the past event is not pointless in that way." Further, in support of backward causation, they also wrote that "to pray in 1980 that Napoleon lose at Waterloo" is logical because "why should your prayer not be efficacious in bringing about Napoleon's defeat?"[32] Disagreeing, however, was another philosopher who bluntly declared "A prayer for something to have happened is simply an absurdity, regardless of the utterer's knowledge or ignorance of how things went."[33]

---

[32] E. Stump and N. Kretzmann, "Eternity," *Journal of Philosophy*, August 1981, pp. 429–458.

[33] P. Geach, *God and the Soul*, Routledge & Kegan Paul 1969.

There have been all sorts of opinions expressed through the ages in reaction to the idea of affecting the past via retroactive petitionary prayers,[34] and on the role of backward causation. The British philosopher Michael Dummett (1925–2011), in particular, discussed Lewis' concept of such prayers with great sympathy,[35] and backward causation allows one to both explain them as well as retaining free will. That is, it is not God's foreknowledge that causes our later actions (forcing our behavior and so turning us into automatons), but rather it is our later *freely-chosen* actions that causes God's foreknowledge! While such theological speculations are interesting, in the end they are simply positions of faith, about which mathematical physics has nothing to say.

Eventually, of course, others besides philosophers and theologians began to ponder the questions raised by 'changing the past.' In a January 1963 personal letter to the editor of *The Magazine of Fantasy and Science Fiction*, Robert Heinlein wrote[36] "Mark Twain invented the time-travel story; 6 years later H. G. Wells perfected it *and its paradoxes* [my emphasis]. Between them they left little for latecomers to do." How a man as widely read as was Heinlein, who had authored some of the best short time travel stories ever written, could have written such an erroneous sentence is a mystery to me. *A Connecticut Yankee in King Arthur's Court* and *The Time Machine*, certainly both works of genius, are *not* pioneers in paradox. And Heinlein's own contributions are proof enough that there was a lot left to do with time travel, well after 1900.

The very first story to be written that even hints at the particular time travel paradox of changing the past seems to be by the Unitarian minister Edward Everett Hale (1822–1909), best known today as the author of the 1863 story "The Man Without a Country." Hale wrote "Hands Off" in 1881, and published it anonymously in *Harper's New Monthly Magazine* with the express purpose of stirring up some theological debate (which apparently it didn't). He certainly had no idea that he would come to be recognized by literary scholars as a pioneer in the yet-to-be invented genre of science fiction.

Hale's story opens with the mysterious words "I was in another stage of existence. I was free from the limits of Time, and in new relations to space." These words are spoken by an unnamed narrator, who seems to have just died, and who finds himself, in his new 'form,' observing "some twenty or thirty thousand solar systems" while in the company of "a Mentor [probably an angel] so loving and patient." Under the guidance of this Mentor, in an attempt to 'improve' history, the narrator alters the Biblical account of Joseph and his imprisonment in Egypt. At first, subsequent history *is* better, but then humanity sinks into irreversible depravity. In the end the narrator watches the last handful of humans kill each other in a particularly symbolic place for the Christian world: "The last of these human brutes all lay stark dead on the one side and on the other side of the grim rock of Calvary!" There would be no Crucifixion and Resurrection for the salvation of humankind, which naturally greatly disturbs the narrator. But the Mentor calms him, saying "Do not be disturbed, you have done nothing." It has, you see, just been an experimental world, an alternate Universe, and the narrator has learned the lesson of "Hands Off" the past.

---

[34] A summary of those opinions can be found in G. Brown, "Praying About the Past," *Philosophical Quarterly*, January 1985, pp. 83–86. Debate continues on the retroactive prayer into the 21st century: see, for example, K. Timpe, "Prayers for the Past," September 2005, pp. 305–322, and T. J. Mawson, "Praying for Known Outcomes," March 2007, pp. 71–87, both in *Religious Studies*.

[35] M. Dummett, "Bringing About the Past," *Philosophical Review*, July 1964, pp. 338–359.

[36] Reprinted in the posthumously published *Grumbles from the Grave* (edited by Heinlein's widow, Virginia Heinlein), Del Rey 1990.

Hale's story is a better Sunday sermon than it is a change-the-past time travel tale, and the device of experimenting on a not-really-real Earth is disappointing from a modern science fiction point of view. But Hale's story almost certainly did have an immediate (if indirect) impact. There is no absolute proof, but with its appearance in a national magazine, it seems quite likely that "Hands Off" was read by Edward Page Mitchell (1852–1927), an editor on a daily New York newspaper, the *Sun*. I write that because, just 6 months after Hale's story appeared, the *Sun*, in its issue of September 18, 1881, printed Mitchell's "The Clock That Went Backward." That tale, published anonymously, used a machine (the clock) for time travel,[37] as well as incorporated the idea of time travel involving paradoxes. The story predates Wells' *Time Machine* by 14 years, and Wells' novel did not include a paradoxical element.[38]

There are, however, two *hints* at paradox in Wells' novel. In the opening, during the dinner party at which the Time Traveller tries to convince his friends of the possibility of a time machine, one of them observes that "It would be remarkably convenient for the historian. One might travel back and verify the accepted account of the Battle of Hastings, for instance." To that another guest replies, "Don't you think you would attract attention? Our ancestors had no great tolerance for anachronisms." The second hint occurs when the incredulous Editor, astonished at the disheveled appearance of the Time Traveller upon his return from the future, wonders "What *was* this time traveling? A man couldn't cover himself with dust by rolling in a paradox, could he?"

What might happen if time travelers *could* change the past? This question is nicely illustrated in one novel[39] where a time traveler finds himself stranded in the London of 1810. Despite his predicament, he takes solace with "I could invent things—the light bulb, the internal combustion engine, . . ., flush toilets . . . " But then he thinks better of doing any of that: "no, better not to do anything to change the course of recorded history—such tampering might cancel the trip I got here by, or even the circumstances under which my mother and father met. I'll have to be careful."

---

[37] See Jan Pinkerton, "Backward Time Travel, Alternate Universes, and Edward Everett Hale," *Extrapolation*, Summer 1979, pp. 168–175. The time machine in Mitchell's story is more fantasy than anything else. It is simply stated that if the clock runs backward, then it travels backward in time.

[38] Wells' failure to use paradox in his famous novel surprises most modern readers and, in fact, one of the first reviewers specifically criticized him for this lapse. See the 1895 review of *The Time Machine* that appeared in *Pall Mall Magazine*, by Israel Zangwill, reprinted in Parrinder's book (note 1 in the Introduction).

[39] T. Powers, *The Anubis Gates*, Ace 1983, a work with equal shares of physics and magic.

This is really just a more recent treatment of the change-the-past paradox that was already well established in early science fiction. In a story[40] published a half-century before, we find the paradox explicitly stated, along with a possible solution that is similar to the kind of explanations that have appeared in the philosophical literature[41]:

> "Suppose you landed in your own past?," queried Eric."
>
> Dow smiled.
>
> "The eternal question," he said. "The inevitable objection to the very idea of time travel. Well, you never did, did you? You know it never happened!"

But, suppose you *could* land in your own past. What then?

One famous story[42] that considered this question *embraced* the idea of changing the past. In it a client on a dinosaur hunting safari fails to follow the instructions of his guide to do *nothing* in the past except shoot a dinosaur that is about to dies for "other reasons" anyway—alas, he accidently kills a butterfly. This results in enormous changes in history, as indicated by the 'before' and 'after' versions of the time machine company's ad:

> **before**
> TIME SAFARI, INC.
> SAFARIS TO ANY YEAR IN THE PAST.
> YOU NAME THE ANIMAL.
> WE TAKE YOU THERE.
> YOU SHOOT IT.
> **after**
> TYME SEFARI INC.
> SEFARIS TU ANY YEER EN THE PAST.
> YU NAIM THE ANIMALL.
> WE TAEK YU THAIR.
> YU SHOOT ITT.

Bradbury describes the death of the butterfly as having started the knocking "down [of] a line of small dominoes and then big dominoes and then gigantic dominoes, all down the years across Time." This is, of course, a somewhat unconvincing argument. After all, previous dinosaurs, when shot, must have fallen to the ground and flattened a lot of butterflies! With such threats for every decision, no matter how seemingly innocent, hanging over the head of a time traveler, it would take a brave soul to do much more, while in the past, than just stand still and breathe.

---

[40] C. L. Moore, "Tryst in Time," *Astounding Stories*, December 1936.

[41] See, for example, P. J. Riggs, "The Principal Paradox of Time Travel," *Ratio*, April 1997, pp. 48–64. The 'principal paradox' is that time travel is inherently contradictory because it permits the possibility of traveling to an earlier time to prevent the trip. The grandfather paradox is a special case of this. For more discussion, see T. Chambers, "Time Travel: How Not to Defuse the Principal Paradox," *Ratio*, September 1999, pp. 296–301.

[42] R. Bradbury, "A Sound of Thunder," *Collier's*, June 1952.

Equally grim is the tale[43] that takes a different view by denying the past can be changed. There we read of a doomed hero who journeys back to 1865 to save Lincoln from Booth, but his "time-distorter" is quickly taken away from him by suspicious guards. Its internal workings tick, you see, and they think he is an assassin with a clock bomb. They destroy it, haul him away to his fate, and Lincoln goes on to meet his. In the same spirit (but even more shocking) is the result of a time traveler's intentional tampering with the past in David Gerrold's 1973 novel *The Man Who Folded Himself*. That traveler experiments with "making things different" and, in his words, "Once I created a world where Jesus Christ . . . went out into the desert to fast and never came back. The twentieth century I returned to was—different. Alien."

With such a stupendous power to alter reality, assuming the past *can* be changed, perhaps one might imagine prospective time travelers to the past being required to first file Historical Impact Statements![44] Not all would receive permission. In one classic tale,[45] for example, we read of a time traveler who takes a rifle and 5000 rounds of explosive bullets back to Golgotha. His intention—to be history's first Rambo by picking off any Roman soldier who gets within a hundred yards of Jesus! As outrageous as this concept is (but who among those now reading this won't admit to at least a momentary thrill at the idea and, perhaps, even a secret willingness to do it themselves, if they could), it isn't the story's peak. That comes when the reader is reminded that it was Christ's *desire* to die on the Cross, that he *had* to die for our sins; to prevent that from happening would change all of history for the last 2000 years. What, then, should the time traveler's colleagues do when they discover his plan? Should they stop him or not? What might happen if they do interfere? Of course, if the time traveler is 'now' in the past, isn't it already 'too late' to stop him? Oh, the conundrums of time travel and changing the past!

The classic change-the-past paradox is, of course, the grandfather paradox. A famous story[46] pushed this paradox to its logical limit to illustrate its supposed dangers. Having traveled to Greece in the fifth century B.C., the traveler suddenly realizes (with just a little exaggeration): "Ninety-five generations back you'd have more grandfathers than there are people on Earth, or stars in the Galaxy! You're kin to everyone . . . You as much as take a poke at anyone, and the odds are you won't even get to be a twinkle in your daddy's eye."

---

[43] R. Silverberg, "The Assassin," *Imaginative Tales*, July 1957.

[44] See, for example, the novel by C. L. Harness, *Krono*, Franklin Watts 1988.

[45] A. Porges, "The Rescuer," *Analog Science Fiction*, July 1962.

[46] P. S. Miller, "Status Quondam," *New Tales of Space and Time*, November 1951. This is the Miller I mentioned back in note 15.

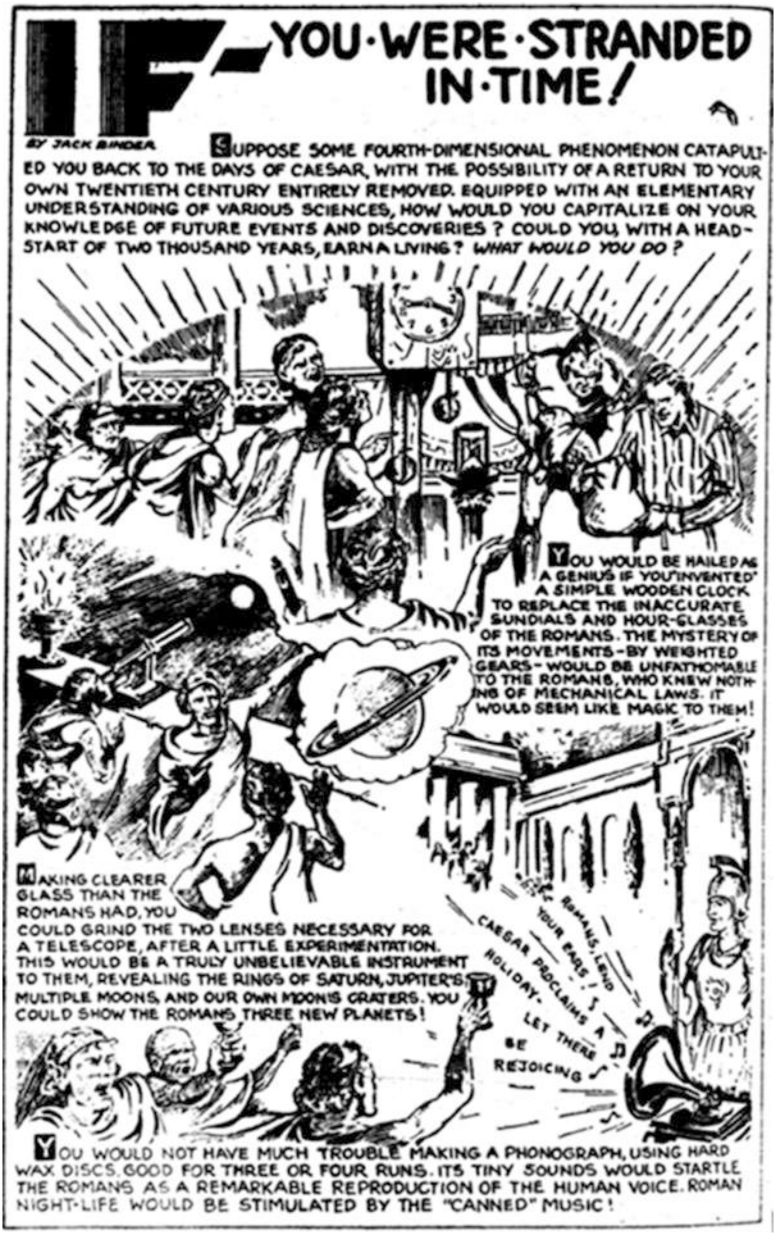
IF—YOU·WERE·STRANDED IN·TIME!
BY JACK BINDER
SUPPOSE SOME FOURTH-DIMENSIONAL PHENOMENON CATAPULTED YOU BACK TO THE DAYS OF CAESAR, WITH THE POSSIBILITY OF A RETURN TO YOUR OWN TWENTIETH CENTURY ENTIRELY REMOVED. EQUIPPED WITH AN ELEMENTARY UNDERSTANDING OF VARIOUS SCIENCES, HOW WOULD YOU CAPITALIZE ON YOUR KNOWLEDGE OF FUTURE EVENTS AND DISCOVERIES? COULD YOU, WITH A HEAD-START OF TWO THOUSAND YEARS, EARN A LIVING? WHAT WOULD YOU DO?
YOU WOULD BE HAILED AS A GENIUS IF YOU "INVENTED" A SIMPLE WOODEN CLOCK TO REPLACE THE INACCURATE SUNDIALS AND HOUR-GLASSES OF THE ROMANS. THE MYSTERY OF ITS MOVEMENTS—BY WEIGHTED GEARS—WOULD BE UNFATHOMABLE TO THE ROMANS, WHO KNEW NOTHING OF MECHANICAL LAWS. IT WOULD SEEM LIKE MAGIC TO THEM!
MAKING CLEARER GLASS THAN THE ROMANS HAD, YOU COULD GRIND THE TWO LENSES NECESSARY FOR A TELESCOPE, AFTER A LITTLE EXPERIMENTATION. THIS WOULD BE A TRULY UNBELIEVABLE INSTRUMENT TO THEM, REVEALING THE RINGS OF SATURN, JUPITER'S MULTIPLE MOONS, AND OUR OWN MOON'S CRATERS. YOU COULD SHOW THE ROMANS THREE NEW PLANETS!
ROMANS, LEND YOUR EARS! CAESAR PROCLAIMS A HOLIDAY. LET THERE BE REJOICING
YOU WOULD NOT HAVE MUCH TROUBLE MAKING A PHONOGRAPH, USING HARD WAX DISCS, GOOD FOR THREE OR FOUR RUNS. ITS TINY SOUNDS WOULD STARTLE THE ROMANS AS A REMARKABLE REPRODUCTION OF THE HUMAN VOICE. ROMAN NIGHT-LIFE WOULD BE STIMULATED BY THE "CANNED" MUSIC!

Fig. 4.2 Illustrator Jack Binder (1902–1986) was the author of a continuing series called "IF— . . ." in *Thrilling Wonder Stories*. In each issue, the ellipses would be replaced with some phrase such as "the Sun exploded!," "there was another ice age!," or "there was no friction!." The installment shown here (and in Fig. 4.3) appeared in December 1938 and asserted that the past could be changed by a time traveler

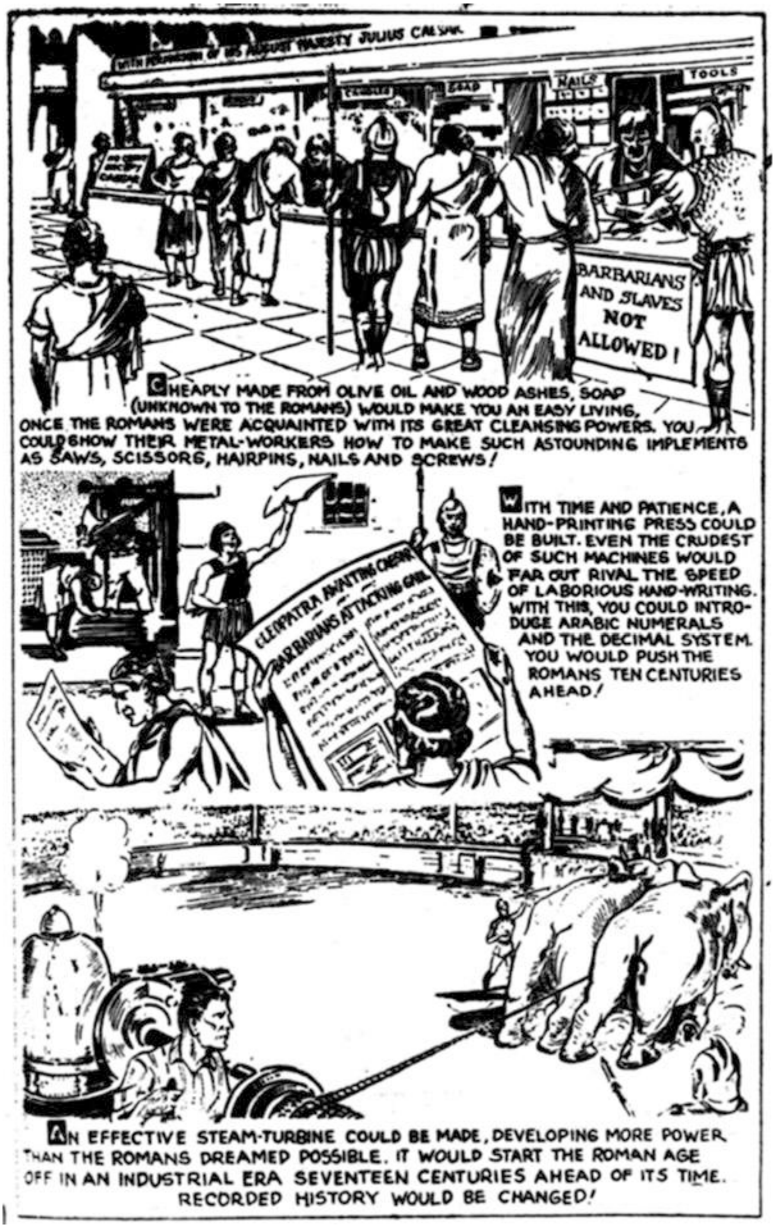
WITH PERMISSION OF HIS AUGUST MAJESTY JULIUS CAESAR
CANDLES
SOAP
NAILS
TOOLS
BARBARIANS AND SLAVES NOT ALLOWED!
CHEAPLY MADE FROM OLIVE OIL AND WOOD ASHES, SOAP (UNKNOWN TO THE ROMANS) WOULD MAKE YOU AN EASY LIVING, ONCE THE ROMANS WERE ACQUAINTED WITH ITS GREAT CLEANSING POWERS. YOU COULD SHOW THEIR METAL-WORKERS HOW TO MAKE SUCH ASTOUNDING IMPLEMENTS AS SAWS, SCISSORS, HAIRPINS, NAILS AND SCREWS!
CLEOPATRA AWAITING CAESAR
BARBARIANS ATTACKING GAUL
WITH TIME AND PATIENCE, A HAND-PRINTING PRESS COULD BE BUILT. EVEN THE CRUDEST OF SUCH MACHINES WOULD FAR OUT RIVAL THE SPEED OF LABORIOUS HAND-WRITING. WITH THIS, YOU COULD INTRODUCE ARABIC NUMERALS AND THE DECIMAL SYSTEM. YOU WOULD PUSH THE ROMANS TEN CENTURIES AHEAD!
AN EFFECTIVE STEAM-TURBINE COULD BE MADE, DEVELOPING MORE POWER THAN THE ROMANS DREAMED POSSIBLE. IT WOULD START THE ROMAN AGE OFF IN AN INDUSTRIAL ERA SEVENTEEN CENTURIES AHEAD OF ITS TIME. RECORDED HISTORY WOULD BE CHANGED!

Fig. 4.3 (Fig. 4.2 continued). Illustration for “IF—You Were Stranded in Time!” ©1938 by Better Publications, Inc.; Reprinted by permission of the Ackerman Science Fiction Agency, 2495 Glendower Ave., Hollywood, CA 90027 for the Estate

Even earlier than Miller's tale is an equally famous one[47] that illustrates the same point in a graphic way. In A.D. 452 a time traveler shoots and kills one of Attila's Huns (who would have been his great-grandfather many times over); the result is that 50,000 of the Hun's descendants vanish! So dramatic did the readers of this story find the concept that author repeated the idea the very next year.[48] Twenty years later another writer topped these tales by having a time traveler accidently kill the original 'intelligent baboon' in the ancient past, thereby wiping out the entire human species![49]

Science fiction writers have been as puzzled by the grandfather paradox[50] as have been nearly everybody else. As the inventor of the first time machine says in one tale,[51] "I have devised a method [for travel] into the distant past. The paradox is immediately pointed out—suppose [the time traveler] should kill an ancestor or otherwise change history? I do not claim to be able to explain how this apparent paradox is overcome in time travel; all I know is that time travel *is* possible. Undoubtedly, better minds than mine will one day resolve that paradox, but until then we shall continue to utilize time travel, paradox or not."

Some may feel it overly dramatic that the classic time travel paradox has such a murderous form, but that is its historical origin in science fiction (not in either physics or philosophy). We can find the grandfather paradox discussed as if already well-known in a letter to the editor at *Astounding Stories* (January 1933). The author of that letter wrote "Why pick on grandfather? It seems that the only way to prove that time travel is impossible is to cite a case of killing one's own grandfather. This incessant murdering of harmless ancestors must stop. Let's see some wide-awake fan make up some other method of disproving the theory." As we proceed, you'll see just how clever some of those who responded to that writer's plea have been but, even today, as the grandfather paradox stands revealed as a red herring, it is preeminent in most people's imaginations. If a solution to the grandfather paradox puzzle escaped an early science fiction writer, then he would generally just mysteriously mention it and then hasten on to other matters. For example, in one story[52] the following exchange between the stock pulp-fiction characters of a young hero and a brilliant old scientist occurs:

---

[47] N. Schachner, "Ancestral Voices," *Astounding Stories*, December 1933.

[48] N. Schachner, "The Time Imposter," *Astounding Stories*, March 1934.

[49] C. Dye, "Time Goes to Now," *Science Fiction Quarterly*, May 1953.

[50] The 'paradox' is that, assuming you do arrive in the past with a working gun, why *can't* you kill your grandfather? After all, you *must* fail in that quest because otherwise you wouldn't be there from the future to even try. But *why* must you fail? It is, of course, not actually necessary to try to kill your grandfather to run into this paradoxical situation—just go back in time to any moment in the past and try to kill yourself! You won't succeed (if the past is unchangeable), but *why not*? (To argue 'because the past is unchangeable' is to beg the question. We need more insight than that.)

[51] M. Reynolds and F. Brown, "Dark Interlude," *Galaxy Science Fiction*, January 1951.

[52] C. South, "The Time Mirror," *Amazing Stories*, December 1942.

> "You mean that time travel really is possible? That men can be transported into the future or the past—?"
>
> The other held up a restraining hand. "Yes. Time travel *is* possible . . ."
>
> "But professor! Think of what you're saying! You're telling me that I could go back and murder my own grandfather. That I could prevent myself from being born—?"
>
> Again the elder man sighed. "I was afraid of this," he said. "I knew you could not understand." He hesitated. Then: "At any rate, take my word for it that time travel is possible. Also, I assure you that there are any number of perfectly sound theoretical and practical reasons why you never could hope to murder your grandparents."

We are, however, not told just what those reasons might be.

Even when all has been said about the impossibility of changing the past, and even when they are finally willing to concede that point, most people still cannot help wondering *why* the time traveler can't kill his grandfather. There the time traveler is, after all, just two feet away from the nasty young codger (I assume he is nasty to make the whole unpleasant business of murder as palatable as possible), with a perfectly functioning and well-oiled revolver in his hand, cocked and loaded with powerful, factory-fresh ammunition that even Dirty Harry would find excessive. What can possibly prevent the time traveler from simply raising his arm and doing the deed? Indeed, the artwork (reproduced at the end of the Introduction) accompanying one 1944 story shows this act in detail, including the smoking gun in the hand of the time traveler who has just taken a shot at grandpop. And if that still leaves open the remote possibility of an aiming error through nervousness, then why can't a suicidal time traveler just wrap his body in factory-fresh dynamite and blow-up granddad—along with himself and everything else within a hundred feet?

I'll argue in this book that killing your grandfather in the past, before he sets *you* in motion, is *logically* impossible. The laws of physics will then faithfully do their duty. No one will ever find an unfinished note in the empty laboratory of a missing traveler who, skeptical of the grandfather paradox, has written "To prove the falsity of the grandfather paradox, I will take my time machine back 50 years and kill my grandf. . .." Nor will any time traveler have to be concerned about the twist in one tale, which opens with the inventor of a time machine is showing the gadget to three friends. One of them later steals the machine to go back 60 years to kill his grandfather—and the story closes with a *near* repeat of the opening, with the inventor showing the gadget to *two* friends.[53]

Invoking logic in this way, in the context of time travel to the past, was discussed in the philosophical literature nearly half a century ago: "If we assume that it is impossible for [a time traveler to kill his younger self], some people are inclined to ask such questions as this: 'But how can the laws of logic prevent him from killing his younger self? Do they cause his finger to slip on the trigger or the bullet to fly apart in mid-air?' The implication of such questions is that the laws of logic cannot prevent such actions. But such questions are like asking: 'How do the laws of logic prevent the geometer from trisecting the angle or squaring the circle? Do they, for example, cause his ruler to slip as a crucial moment every time he tries it?'"[54].

---

[53] F. Brown, "First Time Machine," *Honeymoon in Hell*, Bantam 1958.

A similar point was made later by another philosopher: "Surely it is not an impairment of 'freedom of action' . . . that, e.g., you cannot push another person harder than he/she pushes you. Just as one would explain this is the case by reference to Newton's third law ('to every action there is an equal and opposite reaction'), one could explain the impossibility of [causing a paradox] by reference to the laws which imply such a impossibility. If this explanation is taken to be unsatisfactory, it would seem that one is saddled with a general problem concerning the reconciliation of physics and 'freedom,' and not with a specific argument against [paradoxes]."[55]

The grandfather paradox unquestionably nags at all students of time travel. As a character in one story declares, "The resolution of [the grandfather paradox] is the key to time,"[56] and some incorrectly believe it remains unresolved.[57] The paradox *is* undeniably troublesome: as one philosopher put it, the apparent possibility of a time traveler being able to do away with both his grandfather and himself gives "rise to such puzzles that we are forced to question its [time travel's] intelligibility."[58] In the next section we'll explore how to answer this concern.

4.3 Changing Versus Affecting the Past

> "The past—it's pretty damn solid, Phil. It's a little like a compost pile—fairly soft near the surface but packed hard further down, with all that Time piled on top of it."[59]
>
> —one 'explanation,' perhaps, for the unchangeability of the past

The common belief today, among physicists and philosophers alike, is that given any consistent description of reality it is simply impossible for a time traveler to kill himself as a baby. As one philosopher put it, "Autoinfanticide is *metaphysically impossible* [my emphasis]. This metaphysical impossibility is philosophically intriguing because unlike most impossible events, we can vividly picture how it might look. Time travel itself seems possible, and for those who arrive in the past with proper equipment and training, the actual infanticide should not be difficult."[60] And so the grandfather paradox lives on, bedeviling both physicists and philosophers alike.

---

[54] J. W. Meiland, "A Two-Dimensional Passage Model of Time for Time Travel," *Philosophical Studies*, November 1974, pp. 153–173. Science fiction had already considered time travel suicide in, for example, K. Neville, "Mission," *Fantasy and Science Fiction*, April 1953.

[55] F. Arntzenius, "Causal Paradoxes in Special Relativity," *British Journal for the Philosophy of Science*, June 1990, pp. 223–243.

[56] P. Worth, "Typewriter from the Future," *Amazing Stories*," February 1950. See also note 106 in Chap. 1.

[57] For example, in J. H. Schmidt, "Newcomb's Paradox Realized with Backward Causation," *British Journal for the Philosophy of Science*, March 1998, pp. 67–87, we read that "there are as yet no generally accepted solutions" to the grandfather paradox.

[58] S. Gorovitz, "Leaving the Past Alone," *Philosophical Review*, July 1964, pp. 360–371.

[59] F. M. Busby, "A Gun for Grandfather," *Future Science Fiction*, Fall 1957.

Indeed, one physicist described the time travel paradoxes as "the most controversial issue related to time machines."[61] As he argued, "These paradoxes seem to be something inherent to time machines (their main attribute, perhaps), so it is reasonable to assume that if there exists a universal law prohibiting the time machines, it must have something to do with the paradoxes. And on the other hand, be the problem of the paradoxes satisfactorily solved there probably would be no need to look for such a law, the (supposed) paradoxicalness of the time machines being traditionally the main objection against them."

One of the persistent stumbling blocks to removing the confusion of the paradoxes is a failure to distinguish between *affecting* and *changing* the past. One philosopher wrote this, in a somewhat bungled attempt to explain what is meant by affecting the past: "Nothing anyone can do now can make it not have rained yesterday"[62] if, in fact, it *did* rain yesterday. This is correct, but it is *not* what is meant by *affecting* the past. Rather, if the reason it *did* rain yesterday is because a time traveler from the future seeded the clouds, then that time traveler affected the past. Making it not to have rained yesterday would be to *change* the past.

A classic[63] by Isaac Asimov illustrates what is meant by a time traveler affecting the past. An idealistic physics professor, convinced that the world's political problems are the result of the comparative newness of scientific thought and tradition, tries to change the past (and thus the present) by sending a Greek translation of a modern chemistry text back 2000 years to the Hellenic days of Leucippus, Lucretius, and Democritus. He dies in the attempt but succeeds in the transmission. When a government investigator—called in because the professor drained an entire nuclear power reactor to energize his time machine!—discovers that the transmission takes a day to travel back a hundred years (a little gimmick with no foundation in physics, but simply something Asimov needed for the story), he fears 'our' world will vanish in 20 days, to be replaced by a 'new' one. In the end, however, he learns you can't change the past. As one of the late professor's colleagues tells the investigator, "While you are right that any change in the course of past events, however trifling, would have incalculable consequences ... I must point out that are nevertheless wrong in your final conclusion. *Because THIS is the world in which the Greek chemistry text WAS sent back*."

---

[60] David Horacek, "Time Travel in Indeterministic Worlds," *The Monist*, July 2005, pp. 423–436.

[61] S. Krasnikov, "Time Travel Paradox," *Physical Review D*, February 14, 2002, pp. 064013–1 to -8.

[62] R. G. Swinburne, "Affecting the Past," *Philosophical Quarterly*, October 1966, pp. 341–347.

[63] "The Red Queen's Race," *Astounding Science Fiction*, January 1949.

Another good illustration from science fiction of affecting the past can be found years before Asimov's. In that story,[64] a time traveler leaves the Chicago of 1942 for the year 3000. Much later, in the year 2564, another time traveler interested in history journeys back to 2253 in an attempt to learn the cause of the great Chicago explosion of that year. The explosion was centered on the site of an ancient laboratory once used by a scientist who mysteriously vanished in 1942. The second time traveler begins his trip into the past on the same spot, with plans to go back to the day before the explosion. At the story's end, we learn that the disaster was the result of the two time travelers colliding as both 'passed through' 2253. The backward traveling historian, therefore, by pushing a button in 2564, is the cause of an event that happened 311 years earlier.[65]

To give a science fiction example of changing the past, it would be hard to do better than with a story that appeared a few years after Asimov's The central character is a researcher in time travel who has concluded that what is wrong with the world can be traced back to the scientific method getting off to a late start (this story was almost surely written as a result of Asimov's tale!): the time traveler thinks he can correct matters by visiting 340 B.C. and educating Aristotle on the proper scientific attitude. (Aristotle believed that observing the world was inferior to pure thinking about how the things, in his opinion, *ought* to work.) This the time traveler does, with utterly disastrous consequences. He returns to the present to find a scientifically retarded world that makes him a slave. In his cell he writes on a wall the bitter lesson he has learned too late: "Leave Well Enough Alone."[66]

It is the fear of time travelers from the future attempting to alter the past that has led some philosophers (and not just a few physicists) to assert that time travel is impossible, because it would mean what they feel to be impossible might happen: changing the past. One philosopher, however, argued long ago that such a worry is unwarranted. As he wrote at the end of an essay (a polemic *against* the concept of four-dimensional spacetime, and so *against* the idea of time travel), "Squandering vast sums on foolish enterprises is an everyday occurrence. [For example], will the U.S. time explorer get back and eliminate Lenin before his Russian rival gets back even earlier and eliminates George Washington? . . . If such spectacular folly once gets under way because governments have been convinced of some nonsensical theory, a logician will not . . . lose any sleep about who is going to succeed."[67]

---

[64] O. Saari, "The Time Bender," *Astounding Stories*, August 1937.

[65] This story describes something a bit more than 'simply' affecting the past; it has a *causal loop* in it. The time traveling historian makes his trip *because* of an event in the past that his trip causes. Such paradoxes will be the subject of the next section.

[66] L. Sprague de Camp, "Aristotle and the Gun," *Astounding Science Fiction*, February 1958. Asimov and de Camp were close friends, and their two stories with similar premises are clearly the result of a bit of friendly rivalry.

[67] P. Geach, "Some Problems About Time," in *Studies in the Philosophy of Thought and Action* (P. F. Strawson, editor), Oxford 1968.

And long before that essay (with its correct conclusion reached through faulty reasoning) was penned, we learn the same lesson (in a 1923 tale by the English novelist May Sinclair (1863–1946)) as we follow a woman right into hell after her death; she ends up there because of an immoral life. She then wanders through time, into her past, but finds that she can change nothing. As she is told, “You think the past affects the future. Has it never struck you that the future may affect the past? . . . You *were* what you *were to be*.”[68] This last line, from a *non*-science fiction story, is consistent with the modern view held by physicists of time travel. You cannot travel anywhere into the past unless you’ve already been there, and when you do make the trip you will do what you’ve already done there. You could not, as does the time traveler in one tale,[69] change the course of history by revealing twentieth-century physics in the eighteenth century. That does not mean you would necessarily be ineffectual during your stay in the past, however (certainly it doesn’t mean, as Hugo Gernsback thought, that you’d be *invisible*!) Not being able to change the past is not equivalent to being unable to *influence* or *affect* what happened in the past, and science fiction writers have used this distinction to good effect, as did Asimov (note 63) and de Camp (note 66).

Robert Heinlein was a science fiction writer who clearly understood time travel paradoxes, both what they mean and, at least as important, what they do not mean. In his 1964 cold-war novel *Farnham’s Freehold*, for example—the story of a family that is literally blasted twenty-one centuries into the future when their bomb shelter receives a direct hit from a Soviet nuclear warhead—we find following exchange as two of the characters are about to return to their original time via time machine:

> “The way I see it, there are no paradoxes in time travel, there can’t be. If we are going to make this jump, then we already did; that’s what happened. And if it doesn’t work, then it’s because it didn’t happen.”
>
> “But it hasn’t happened yet. Therefore, you are saying it didn’t happen, so it can’t happen. That’s what I said.”
>
> “No, no! We don’t know whether it has already happened or not. If it did, it will. If it didn’t, it won’t.”

Modern philosophers, and many physicists, too, as well, who have examined the concept of time travel in depth, agree with Heinlein’s character and, indeed, it is now common practice to invoke the so-called *principle of self-consistency*—generally attributed to the Russian astrophysicist Igor Novikov (see note 117 in Chap. 1) because he and his colleagues did not simply invoke it, but rather were able to *derive* it from the *principle of least action*, a concept held by many to be at the highest level of importance in physics[70]: all that is required, argued Novikov, in *any* physical process (including time travel), is that a *logical* consistency exist between events.[71] In his book *Evolution of the Universe* (originally published in Russian in 1979), Novikov wrote "The closure of time curves does not necessarily imply a violation of causality, since the events along such a closed line may be all 'self-adjusted'—they all affect one another through the closed cycle and follow on another in a self-consistent way." He later repeated that view in one of the first time machine papers in the physics literature.[72]

---

[68] M. Sinclair, “Where Their Fire Is Not Quenched,” in *After the Darkness Falls* (B. Karloff, editor), World Publishing 1946.

[69] D. Beason, “Ben Franklin’s Laser,” *Analog*, December 1990.

In fact, despite the attachment of Novikov's name to the principle of self-consistency, it was actually around in physics *decades* earlier; it has been traced back to as far as 1903![73] And at least an intuitive understanding of the principle can be found in the mainstream literature from nearly as long ago. For example, in Lord Dunsany's short 1928 play *The Jest of Hahalaba* (the inspiration for the 1944 film *It Happened Tomorrow*), a man obtains (via supernatural means) a copy of tomorrow's newspaper. In it he reads his own obituary, which so shocks him that he promptly expires—thus explaining the obituary notice.

The principle of self-consistency has been in science fiction long before Novikov, too. An example is the 1941 story "Time Wants a Skeleton" (see note 15). In it one character, after puzzling over a time travel paradox, realizes that "Future and present demanded co-operation, if there was to be a logical future!" And a nice lecture on the principle (that pre-dates by 3 years the dialogue quoted earlier from Heinlein's *Farnham's Freehold*) is given by a character that is particularly interesting because it was published, not in a specialty science fiction magazine catering to an audience with 'genre knowledge' of time travel, but rather in an icon of general American culture.[74]

Not all science fiction writers, however, have understood the requirement for consistency around a loop in time. In one tale, for example, a man meets the inventor of a time machine and agrees to his request to use it to travel into the future. Once he is in the future, alas, the machine breaks. The man then finds another machine that, though it is too small for him to fit in it, is able to hold a recording that he sends back into the past to himself, to a time *before* he started his forward journey. The message on the recording (which he did not receive the 'first' time) is, of course, *not* to make a deal with the inventor. This advice he follows, and so the principle of self-consistency is violated twice in this story.[75]

---

[70] Like just about everything concerning time travel, however, not *all* think this. For example, the great German physicist Max Planck (1858–1947), the 1918 Nobel physics laureate, said (in 1922): "Physics hence is inclined to view the principle of least action more as a formal and accidental curiosity than as a pillar of physical knowledge." Still, he *did* also declare that he thought it unlikely "the dominance of such a simple law could be a mere accident." Quoted from Marc Lange, "Conservation Laws in Scientific Explanations: Constraints or Coincidences," *Philosophy of Science*, July 2011, pp. 333–352.

[71] See A. Carlini, *et al.*, "Time Machines: The Principle of Self-Consistency as a Consequence of the Principle of Minimal Action," October 1995, pp. 557–580, and "Time Machines and the Principle of Self-Consistency as a Consequence of the Principle of Stationary Action (II): The Cauchy Problem for a Self-Interacting Relativistic Particle," October 1996, pp. 445–479, both in *International Journal of Modern Physics D*.

[72] I. D. Novikov, "An Analysis of the Operation of a Time Machine," *Soviet Physics JETP*, March 1989, pp. 439–443.

[73] R. D. Driver, "Can the Future Influence the Present?" *Physical Review D*, February 15, 1979, pp. 1098–1107.

[74] R. F. Young, "The Dandelion Girl," *The Saturday Evening Post*, April 1, 1961. See also note 49 in Chap. 2.

An ability to play a role in history is not without some constraints. You can't save Jesus with a rifle (see note 45), or Joan of Arc with a fire extinguisher, or knock-out John Wilkes Booth with a baseball bat outside of Ford's Theatre, or blow-up Hitler with a bomb, and you can't prevent either the Black Death in the London of 1665 or the Great Fire the following year. But it *is* logically possible for a time traveler who has an infected rat sneak into his time machine, or who carelessly discards a match, to be the *cause* of the last two examples. That was the fate, for example, of the time traveling historian from A.D. 2461 who was the cause of the plague in A.D. 562 Rome, as well as of that in England nearly 800 years later.[76]

Michael Moorcock's 1969 novel *Behold the Man* gets the impossibility of changing the past, and the possibility of affecting it, right. When a disturbed man journeys backward in time to ancient Galilee to meet Christ, only to discover that there is no such person, *he* assumes the role and lives out the Biblical accounts up to and including dying on the Cross. He has not changed the past, but he certainly plays an important role in it!

An early science fiction story that got this right, long before the philosophers and physicists thought of it, was the clever tale whose artwork I have reproduced at the end of the Introduction.[77] In that story, a time traveler journeys back from 1943 to 1870 and shoots his then 14-year old grandfather in the head. Leaving his victim lying on the ground with "blood oozing all over the youth's forehead," the would-be killer returns to 1943. Once back, however, he finds himself in a strange place where he learns from two men that the Germans destroyed New York in 1920 with poison gas! Suddenly realizing that the death of his grandfather has apparently changed history (a curious oversight for anyone smart enough to invent a time machine and then to use it to force the 'grandfather paradox'), he decides he'd rather be dead than be cut off for all time from the world he remembers. So, he shoots himself dead. Then we learn that the two men he encountered are actually inmates in an asylum who like to make-up stories for unsuspecting strangers. We also learn that the time traveler's grandfather's photographs always did show him with a "white, furrowed scar on his forehead that might have been caused by a glancing bullet."

---

[75] R. Wilson, "The Message," *Astounding Stories*, March 1942.

[76] G. C. Edmondson, "The Misfit," *Fantasy and Science Fiction*, February 1959.

[77] M. Weisinger, "Thompson's Time Traveling Theory,' *Amazing Stories*, March 1944.

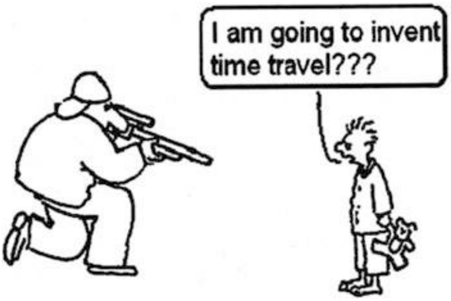
I am going to invent time travel???

Fig. 4.4 The inventor of a time machine about to commit autoinfanticide in the past (the youngster holding the teddy bear is a younger version of the time traveler). Illustration reproduced by the kind permission of Frank Arntzenius (Professor of Philosophy at Oxford University), from his paper "Time Travel: Double Your Fun," *Philosophy Compass*, November 2006, pp. 599–616

Well, okay, you might say at this point, 'I'm convinced you can't change the past, but let's get back to the autoinfanticide (grandfather) paradox. So *why can't* a time traveler kill his baby-self in the past?' A possible answer, one now generally accepted by philosophers and physicists alike, appeared first in science fiction. In a tale[78] that appeared just the year after Gödel's 1949 discovery of time travel in general relativity, we find a character saying "The answer is quite simple. When the man goes back in time and kills his grandfather, and returns to his own time again, he finds to his surprise that he made a mistake. It was not his grandfather at all! And no matter how many times he goes back and kills his grandfather … he *always* [my emphasis] finds he made a mistake." Or, perhaps, some noise distracts him as his finger tightens on the trigger, or the grenade he tosses at granddad is a dud, or a gust of wind deflects the arrow, or (most ludicrous of all) he simply slips on a discarded banana peel!

Okay, that works for *that* time traveler. But suppose, someone objects, that we arrange to have a *lot* of time travelers go back in time, each with murder in his heart for his grandfather. Then, as one philosopher has observed, "Since [killing one's grandfather in the distant past] is impossible, each assassin fails. Some change their minds, others slip on banana peels, yet others kill the wrong target, and so on. But there is something odd about the idea that such coincidences are guaranteed to happen, again and again!"[79]

Early science fiction avoided invoking banana peels by providing an even more extreme 'explanation' for the failures: the time police, who are charged with foiling would-be grandfather killers. (See, for example, the many stories by Poul Anderson (1926–2001) of the 'Time Patrol.') These time commandos are imagined to roam the corridors of time, disrupting the attempts of all those who would change recorded history. Stories of these temporal cops are simply westerns, mysteries, police procedurals, or some other similar type of specialty genre story wearing thin camouflage. This story device, whose main purpose is to allow both time travel and free will,[80] has been correctly called "boring" by at least one philosopher (see note 5 in the Introduction), an evaluation shared by modern philosophers, physicists, and (I think) even most modern science fiction writers.

---

[78] "Typewriter from the Future": see note 106 in Chap. 1.

[79] Theodore Sider, "Time Travel, Coincidences, and Counterfactuals," *Philosophical Studies*, August 2002, pp. 115–138.

So, we seem to be back to banana peels to save grandfather—but it *is* difficult to deny that vast hordes of murderous grandsons *do* appear to require an unlimited number of strategically placed banana peels, strewn all about the past, to trip-up every one of those potential assassins. This problem, of repeated, improbable coincidences to thwart murderous descendants from the future, was first commented on by the philosopher Paul Horwich in 1975 (see note 19 in the Introduction), and then given a convincing resolution by another philosopher in 1997.[81]

To explain the argument, I'll first use the philosopher's less deadly example of dated objects. "Suppose," he writes, "that every object has written upon it the date on which it will cease to exist ... perhaps a time traveler travelled into the future, observed the demise of objects and then travelled back [to just after he left for the future] and wrote the dates."[82] If now the time traveler tries to destroy an object *before the date written on it*, then he *will* fail. As the philosopher amusingly described his attempts to destroy a pen 'before its time,' "I take it outside to place under the wheels of a passing train, but there is a train strike that day. The telephone rings just as I am about to drop the pen into a vat of acid. I slip on a banana peel on my way to put the pen in the microwave. My dog eats my designs for a pen grinder. And so on, for as long as you please. However many attempts I make, the attempts in no way require the occurrence of the coincidences that foil them." To put it bluntly, 'Stuff happens.' The pen has the observed date of its destruction on it, and that date is still in the future and so it is simply impossible to destroy it now.

Now, here's the point: the date on the pen is there *because* all those attempts to destroy it before that written date fail. But the presence of the date is *not* the reason for any of the weird (?) occurrences that disrupted all the attempts to destroy the pen, but rather it's *because* all those attempts failed that the date is what it is. This same argument applies to the grandfather paradox. The only time travelers available, today, to go back into the past to *try* to kill their grandfathers, are precisely those time travelers whose grandfathers were *not* killed. Or, to paraphrase our philosopher (note 81), to ask 'why do coincidences always foil the time traveller's attempts to kill [grandfather in the distant past], is to get things back to front. It is only because the murder attempts fail that the time traveler is alive in the future to even make the *attempt*.'

---

[80] See, for example, David King, "Time Travel and Self-Consistency: Implications for Determinism and the Human Condition," *Ratio*, September 1999, pp. 271–278.

[81] Nicholas J. J. Smith, "Bananas Enough for Time Travel?" *British Journal for the Philosophy of Science*, September 1997, pp. 363–389.

[82] This does present us with the curious (although non-paradoxical) situation that the time traveler will find, upon his appearing in the future, the date he *will write* (in his personal future) when he returns to just after he left on his time trip.

In other words, not only is the grandfather 'paradox' not a paradox, it isn't even surprising!

4.4 Causal Loop and Bootstrap Paradoxes

> "My dear Collingwood, don't drive yourself crazy trying to resolve the paradoxes of time travel. The [time machines] are gone . . . have a drink."[83]

The grandfather paradox might finally have been put to rest, but there are still plenty of other logical minefields left to be negotiated. One of the more puzzling is that of the closed loop in time, a conundrum nicely illustrated by one philosopher[84] as follows, as an explanation of the journey one time traveler makes to 3000 B.C.: "In our time travel story it just may be that the traveler's interest in going back to ancient Egypt is stimulated by recently discovered documents, found near Cairo, containing the diary of a person claiming to be a time traveler, whereupon our hero, realizing it is himself, immediately begins . . . construction of a rocket in order to 'fulfill his destiny.'" In other words, (1) he builds a time machine and goes back to the past because of the discovered diary, and (2) the diary is discovered because he goes back to the past. Each of these points by itself has logical clarity, but together they form a closed time loop (a *causal loop*) of enormous mystery.

Science fiction was strewn with causal loops long before the philosophers and physicists began to ponder them, however, with (for example) one early tale on a time traveler who journeys a century into the past because she finds an old, yellowed newspaper story describing her arrival.[85] But this tale wasn't the first to use a causal loop, as we can find one of the first sophisticated treatments of this device in a story that appeared even earlier (in the same publication).[86] A time traveler in 1930, about to start his journey into the future in an airplane/time machine, wonders at the last moment if he should really go—then he sees himself returning and thus *knows* he will successfully make the trip.

As he later tells a friend, "That decided me . . . Paradoxical? I should say so! I had seen myself return from my time-trip *before* I started it [just like Marty McFly in the original *Back to the Future* film]; had I *not* seen that return, I would *not* have commenced that strange journey, and so could *not* have returned in order to induce me to decide that I *would* make the journey!" And later, when he finds himself in a dangerous situation in the future, he draws hope from that initial experience: "I *would* escape ... It was so decreed. Had I not, with my own eyes, seen myself appear out of the fourth dimension back there in the Twentieth Century, and glide down to my landing-field? Surely, then, I *was* destined to return to my own age safe and sound."

---

[83] A science fiction suggestion that in certain situations (particularly causal loops), might actually be good advice! From L. Sprague de Camp's "The Best-Laid Scheme," *Astounding Science Fiction*, February 1941.

[84] L. Dwyer, "Time Travel and Some Alleged Logical Asymmetries Between Past and Future," *Canadian Journal of Philosophy*, March 1978, pp. 15–38.

[85] P. Bolton, "The Time Hoaxes," *Amazing Stories*, August 1931.

[86] F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931.

Even more dramatic is the second, internal time loop that ends the story. When the time traveler arrives in a ruined city in the year A.D. 1,001,930 he is greeted, *by name*, by an old man who says he (the old man) is the Last Man alive. He knew the time traveler was coming because an ancient history book had said the Last Man had, in fact, appeared in the year A.D. 502,101 in the very time machine out of which the time traveler has just stepped. The time traveler is so startled by all this (and who could blame him!?) that he decides to mull over what he has been told until the next day. As he wakes up in the morning, he is just in time to watch the Last Man depart for 502,101. Stranded in the future, the time traveler wanders the empty city in despair until he chances upon a museum. And in the museum, sealed in a glass case, is his time machine (!)—it has been there for half a million years, since the end of the Last Man's journey. And so the time traveler is saved; he merely adds some oil to the still-functional engine (if you can accept time travel, I suppose this is no more difficult to believe) and returns to 1930—just as he saw himself do at the beginning of the story.

Since Bridge's astonishing story, the idea of a causal loop in time has been used many times in science fiction. Here's a representative sampling:

(1) Time travelers arrive in the forty-sixth century, only to find that they are expected. Their host tells them why: "I have been awaiting your arrival from the past. I have a written record of your coming. You see, I have a time machine myself ... With my time machine, I recently went a year into the future and read the written account I had made, or will make after you leave. Then I came back, awaiting your arrival."[87]

(2) Armed travelers return to the Triassic age to uncover the secrets of a mysterious artifact that has been recently discovered; at the end we learn it is the remains of their own automatic rifle[88];

(3) A time traveler journeys back 500 years, where he suffers an accident that results in his being "agelessly stuck" in his time-traveling gadget until he is freed—by himself, 500 years later. He then gets into the gadget to journey back 500 years[89];

(4) The world's time suddenly loses 5 min, an astonishing event that comes to be called "the time drop." After 2 weeks of investigation, a reporter traces this

---

[87] E. Binder, "The Time Cheaters," *Thrilling Wonder Stories*, March 1940.

[88] J. Blish, "Weapon Out of Time," *Science Fiction Quarterly*, Spring 1941.

[89] A. B. Chandler, "The Tides of Time," *Fantastic Adventures*, June 1948.

event to a reclusive (but brilliant, of course) scientist who reveals that he has invented a time machine. The reporter decides to test this claim by using the machine to return to just before the start of the time drop, to observe precisely what caused it—it is, in fact, a malfunction of the machine that is at fault and the reporter finds himself caught in a 2-week long causal loop[90];

(5) A physicist knows something odd lies in his future when he is confronted with a 700 year-old museum copy of a book. The puzzle is how to explain a message penned in ancient, faded ink, in modern English *and in his handwriting*, on the back side of one of the recently unglued endpapers! How, too, to explain his own fingerprints all over the same endpaper? How, indeed, to answer these questions is his problem when he is presented with all of this and is asked, "Have you, by any chance, been visiting the thirteenth century?" At the end of the story the time loop is closed when the physicist finds himself writing that same message on a *brand new* copy of the book that has been sent from the past (and that he returns to the thirteenth century via a "time portal")[91];

(6) A time traveler from 1964 is secretly observed by one of the 'locals' when he arrives in 1683. The oddness of the sudden appearance of the time traveler and his machine ("It were a kind of Dazzle") makes the local think it might be that the stranger is the man who stole some items from his home the previous night, the same night he had an "ill Dream." Stealing the time machine after the time traveler has gone exploring, the local travels to 1964 where he learns how valuable antiques are. So back he goes to 1683, to the night before the time machine first appeared, to get some 'antiques' from his house. And thus he realizes who the thief *really* was. Before leaving again for the future, he enters his own bedroom to see himself asleep and then to awaken. And so he also learns the cause of what he called his "ill Dream."[92];

(7) A movie production crew goes into the past to make a film. At the end of the story it becomes clear that their presence in the past was not an insignificant event, as one character realizes after seeing the evidence of how they affected (*not* changed!) the past: "If this is true, then the only reason that the Vikings settled in Vinland is because we decided to make a motion picture showing how the Vikings settled in Vinland"[93];

(8) A private college, endowed decades before by a generous but mysterious benefactor, experiments with a time machine. Suddenly, one of the college's graduates is accidently sent a hundred years into the past—where she becomes the benefactor. The college comes into existence, therefore, because it will exist[94];

---

[90] W. Sheldon, "A Bit of Forever," *Super Science Stories*, July 1950.

[91] M. Leinster, "The Gadget Had a Ghost," *Thrilling Wonder Stories*, June 1952.

[92] D. I. Massor, "A Two-Timer," *New Worlds SF*, February 1966.

[93] H. Harrison, *The Technicolor Time Machine*, Doubleday 1967.

[94] C. Simak, "The Birch Clump Cylinder," *Stellar 1* (J. del Rey, editor), Ballantine 1974.

(9) A man gets the money to support his experiments in time travel by selling a large collection of old, rare comic books he has discovered in his late mother's attic. Later we learn how the comic books came to be there; after his experiments are successful, the inventor travels back into the distant past, buys the *newly* published comic books right off newsstands, and stores them in his mother's attic where, decades later, he knows his younger self will grow up and then find them (and thus get the money to make it all happen)[95];

(10) In the 1980 film *Somewhere in Time* (based on the 1975 novel *Bid Time Return* by Richard Matheson), a man in the present is visited by a mysterious old woman who gives him a watch. Later, he travels back to 1912 where he meets a girl to whom he gives the watch. He then returns to the present, and she lives out her life from 1912 on, until she too reaches the present, where we discover she is the (now old) woman who gives the man the watch.

Once philosophers discovered the bizarre nature of causal loops, they quickly proved themselves to be the equal of science fiction writers in imagining strange doings. Here's one example of that, one which any writer would be proud of: "If James cannot decide whether to marry Alice or Jane, he simply travels to the future and learns that he is to choose Alice; he then chooses her for this reason. One wants to object that the decision to marry Alice was never really made at all! But this is not true; the decision was made—as a result of the knowledge that this was the decision . . . It is not the case that the prospective bridegroom could visit the future and compare the results of marrying Alice with those of marrying Jane in order to decide between the alternatives. For if he visits the future, he will learn only that in fact he chose Alice, for better or for worse!"[96]

This same philosopher elaborated on his view of causal loops in a later paper, where he wrote "What if time travel becomes commonplace, so that we must deal with a constant stress of time travelers returning from the future to reveal what they have seen?"[97] His answer is "I think it is clear that the . . . causal loop we have been discussing would become very common, and would play a prominent role in human affairs." He denied, however, that such causal loops would mean the loss of free will. As he explained his position, knowledge of a rigged roulette will not prevent you from putting your money on the table *if you want to*, but perhaps that knowledge will influence your *freely* made decision making. Whether you learn that the roulette wheel is rigged by traditional means (perhaps you see magnets being installed under the table) or by means of time travel is irrelevant—even with this knowledge, you act freely. Other philosophers have not been so generous. One disliked causal loops so much, for example, that while he believed them to be conceptually possible, he also thought them to have "a queer smell,"[98] so much so that he simply preferred to avoid thinking about them!

---

[95] D. Knight, "The Man Who Went Back," *Amazing Stories*, November 1985. This same idea was used earlier in the story "Compounded Interest," (*Magazine of Fantasy and Science Fiction*, August 1956) by Mack Reynolds, in which the inventor of a time machine has the money to build his gadget because he uses it to go back into the past where he deposits a small sum, which then grows (through the 'magic' of compound interest) into the cash he needs to fund his time machine.

[96] G. Fulmer, "Understanding Time Travel," *Southwestern Journal of Philosophy*, Spring 1980, pp. 151–156.

[97] G. Fulmer, "Time Travel, Determinism, and Fatalism," *Philosophical Speculations in Science Fiction and Fantasy*, Spring 1981, pp. 41–48.

One concern that many philosophers and physicists have had with closed loops in time is that they fear that would mean being trapped on an endless cycle of repeating events. For example, one philosopher long ago wrote

> "There is nothing contradictory in imagining causal chains that are closed, though the existence of such chains would lead to rather unfamiliar experiences. For instance, it might then happen that a person would meet his own former self and have a conversation with him, thus closing a causal line by the use of sound waves. When this occurs the first time he would be the younger ego, and when the same occurrence takes place a second time he would be the older ego. Perhaps the older ego would find it difficult to convince the younger one of their identity; but the older ego would recall an identical experience long ago. And when the younger ego has become old and experiences such an encounter a second time, he is on the other side and tries to convince some 'third' ego of their physical identity. Such a situation appears paradoxical to us; but there is nothing illogical in it."[99]

What has been (erroneously) described with that is the beginning of an endless succession of encounters around a closed causal loop. There is, however, just *one* encounter on such a loop in spacetime (but, of course, the mind of the time traveler experiences the encounter twice), subject to the constraint of self-consistency. Some physicists, too, have been so concerned about multiple trips around closed timelike curves (CTCs), because they think such trips would allow the past to be changed, that they have felt it necessary to specifically forbid such a possibility. As one paper put it, "That the principle of self-consistency is not totally tautological becomes clear when one considers the following alternative: The laws of physics might permit CTCs; and when CTCs occur, they might trigger new kinds of local physics which we have not previously met. For example, a quantum-mechanical system, propagating around CTCs, might return to where it started with values for its wave function that are inconsistent with the initial values; and it might then continue propagating and return once again with a third set of values, then a fourth, then a fifth … The principle of self-consistency by fiat forbids changing the past."[100]This last statement is, of course, in agreement with the position I have taken in this book, a position that has generally been accepted by most philosophers for several decades now, but the proponents of the principle of self-consistency seem to have been driven to it by a fear of the past ‘happening again’ over and over, as in the 1993 film *Groundhog Day*.

---

[98] M. MacBeath, "Communication and Time Reversal," *Synthese*, July 1983, pp. 27–46.

[99] H. Reichenbach, *The Direction of Time*, University of California Press 1956, p. 37.

[100] J. L. Friedman et al., "Cauchy Problem in Spacetimes with Closed Timelike Curves," *Physical Review D*, September 15, 1990, pp. 1915–1930. Another physicist, however, has flatly rejected this need for the Principle, calling it redundant: see D. Deutsch, "Quantum Mechanics Near Closed Timelike Lines," *Physical Review D*, November 15, 1991, pp. 3197–3217.

Science fiction writers have stumbled into the error of endless cycling on a closed time loop, too. In one such tale,[101] the inventor of the first time machine travels 500 years into the future where he finds a bronze statue of himself that honors his discovery of time travel. Suddenly injured, fatally, he returns to the present with the statue and then dies. As a memorial, the statue is placed in the very spot where the inventor found (will find) it. As the tale ends, the late inventor’s lab assistant wonders to himself what will happen 500 years later: “Suddenly a strange machine will come out of the past and [the inventor] will be here again—although he is dead and has been dead 500 years. [He will take the statue] and go back to the past . . . to die. And once again that maddening cycle will begin, to go on and on forever as long as time spins its threads.”

That story illustrates yet another puzzle associated with those causal loops that contain a circulating, physical object. That is, *who made the statue*? We can ask the same question about the watch in the time loop of *Somewhere in Time*/*Bid Time Return* as, at every instant of its existence, the watch is in the possession of either the man or the woman? So, *when was the watch constructed*?

There have been some science fiction writers who specifically recognized this question, long before either the philosophers or the physicists paid attention to it. In one early tale,[102] for example, we read of a time machine that travels from 1935 to 1925. When the question of the origin of the time machine comes up, we read

> “One time machine, found in 1935 and brought back to 1925—found in 1935 *because* brought back to 1925. That is all.”
>
> “But who made it in the first place?—Oh, skip the ‘in the first place.’[103] Just plain: who made it?”
>
> “No one. It was never made . . . It is here because it is here.”

This same puzzle was addressed in *The Technicolor Time Machine* (note 93), when one character is perplexed over a piece of paper in his wallet with a diagram on it, a piece of paper he got from himself (an *older version* of himself, who traveled into the past to give it to his younger self). In frustration, he asks a friend:

> “Then no one ever *drew* this diagram. It just travels around in this wallet and I hand it to myself. Explain that.”
>
> His friend replies:
>
> “There is no need to, it explains itself. The piece of paper consists of a self-sufficient loop in time. No one ever drew it. It exists because it is, which is adequate explanation. If you wish to understand it, I will give you an example. You know that all pieces of paper

---

[101] S. Mines, “Find the Sculptor,” *Thrilling Wonder Stories*, Spring 1946.

[102] R. M. Farley, “The Man Who Met Himself,” *Top-Notch Magazine*, August 1935 (*Top-Notch* was an adventure pulp published between 1910 and 1937).

[103] The reason for this line in the story is that earlier the question of “Where did the time machine come from *originally*?” was raised. The answer: “There was never any ‘original.’ . . . There is no round-and-round circle of events, no repetition. Merely *one* closed cycle.” This is, in fact, the modern view of causal loops, expressed in a 1935 (!) science fiction story.

> have two sides—but if you give one end of a strip of paper a 180-degree twist, then join the ends together, the paper becomes a Möbius strip that has only one side. It exists.[104] Saying it doesn't cannot alter the fact. The same is true of your diagram; it exists."
>
> "But—where did it come from?"
>
> "If you must have a source, you may say that it came from the same place that the missing side of the Möbius strip has gone."[105]

The undeniable mystery of causal loops is the reason behind the philosopher I cited earlier (note 98) who thought they have a "queer smell" and so viewed them with much suspicion. He wasn't alone in that feeling, and another philosopher said as much whe he wrote "despite [strong] arguments for the consistency of time travel stories [with causal loops], the impression is apt to remain that something is wrong with them. I think this impression is correct."[106] One story that this philosopher could well have had in mind is a classic,[107] a tale that describes a knife brought from a museum in the future back to the present. It arrives in the present with a flawless blade, but soon thereafter gets a nick in the blade. How, wonders the narrator, can the time loop be completed "again"? I do not find this quite the puzzle that either the author (and perhaps the philosopher) do: it is simply a variation of the grandfather paradox (which has been shown not to be a paradox at all). *If* the knife is found flawless in the future, *then* it was not (will not) be nicked in the past. As written, the story is not logically consistent as it involves changing the past but, *if* one removed the detail of a *nicked* blade, then we *would* have a true (paradoxical) causal loop, with the question the story, itself, asks about the knife: "How was this knife created . . . when its existence has no beginning or end?"

The nicked knife does illustrate a subtle problem that bedevils *any* causal loop containing a physical object. Consider once again the watch in the film *Somewhere in Time*. Assume the watch received by the man in the present is bright and shiny. He then takes it back into the past and gives it to his love. It remains with her after his return to the present until, decades later, she gives it to him—bright and shiny. Why didn't it tarnish? Is there some peculiar anti-tarnish property to a watch in a causal loop? Well, if so, is that anymore odd than a causal loop itself?[108]

None of that, however, provides a means for rejecting time travel *if* one can argue that it is possible to have time travel *without* causal loops. Indeed, Professor Hanley (see note 105) argues that it is possible, and presents what he claims is an example of how to do it. Alas, another philosopher convincingly showed that the example is flawed and that Hanley's claim that there is no causal loop in his story "is unjustified."[109] From an entertainment point of view, however, eliminating causal loops is going in the wrong direction, as it is the inclusion of causal loops that gives a feeling of mystery to a good science fiction story.

---

[104] See note 99 in Chap. 1, and the related discussion there.

[105] One philosopher calls this bit of dialog "unhelpful," while ignoring the fact that it appeared in a science fiction pulp magazine and not a scholarly journal, and was clearly meant to dazzle teenage boys (see note 39 and related discussion in "Some First Words") with the concept of a causal loop, rather than to break new ground in metaphysical thought. See Richard Hanley, "No End in Sight: Causal Loops in Philosophy, Physics and Fiction," *Synthese*, July 2004, pp. 123–152.

[106] G. Nerlich, "Can Time Be Finite?" *Pacific Philosophical Quarterly*, July 1981, pp. 227–239.

[107] P. S. Miller, "As Never Was," *Astounding Science Fiction*, January 1944. This is the same Miller who appears in note 15 (and see note 46, too).

[108] It is *not* sufficient to say that perhaps she polished the watch. Polishing would remove material from the watch, which means she gives him a watch different from the one he gives her in the past.

To finish this section, then, we can do no better than to discuss causal loops that are even *more* bizarre than are those with a physical object; that is, loops that involve time traveling *information*. (Since information doesn't 'tarnish,' however, such a loop avoids that particular puzzle associated with a physical object in a causal loop.) A classic example of such a loop is a mathematician who is visited in his youth by a time traveler from the future (perhaps himself), who gives him the proof of a theorem for which the mathematician is (will be) famous in the future. *Where*, then, did the proof actually come from? In what *mind* was it *created*?[110]

The philosopher David Lewis wrote with particular insight on causal loops, especially ones that involve information transfer, such as a time traveler going back in time to tell his younger self how to build a time machine so that once its constructed he can go back in time and tell himself how to do it.[111] (This was item (3), you'll recall, in Jim Nicholson's 1931 letter to *Science Wonder Stories* magazine, quoted at the end of the first section of this chapter.) As Professor Lewis wrote (see note 5 in the "Introduction"), "But where did the information come from in the first place? Why did the whole affair happen? *There is simply no answer* [my emphasis]. The parts of the loop are explicable, but the whole of it is not. Strange! But not impossible, and not too different from inexplicabilities we are already inured to. Almost everyone agrees that God, or the Big Bang, or the entire infinite past of the Universe, or the decay of a tritium atom, is uncaused and inexplicable. Then if these are possible, why not the inexplicable causal loops that arise in time travel?"

A few years later, another philosopher[112] gave a similar response to a paradox involving a causal loop similar to Lewis', a loop involving a time machine containing a book with instructions on how to make the time machine. The book travels into the past on the machine so it can be read—in order to make the machine.

---

[109] Bradley Monton, "Time Travel Without Causal Loops," *The Philosophical Quarterly*, January 2009, pp. 54–67.

[110] Professor Hanley (note 105) says the answer to such questions is "straightforward": the information comes "from itself." I think the issue is rather deeper than that.

[111] See D. Franson, "Package Deal," in *Microcosmic Tales*, Taplinger 1980. The British philosopher J. R. Lucas had a similar scenario in mind when he wrote, in his book *A Treatise on Time and Space* (Methuen 1973, p. 50), "It is very important, not only for reasons of modesty, that I should not be able to use a Time Machine to go into a public library and read my own biography." Robert Heinlein didn't agree with Lucas: in his 1956 novel *The Door Into Summer* the protagonist, an inventor, travels thirty years into the future, where he reads some patent disclosures for inventions that he doesn't remember, even though they are in his name. He then returns to his own time and promptly files the patents!

[112] M. R. Levin, "Swords' Points," *Analysis*, March 1980, pp. 69–70.

In answer to the question "Who wrote the book about building a time machine?" the philosopher says this question is "no different from questions about where *anything* originally came from. We can ask about the origin of the atoms . . . their time line is not neatly presented to us. The atoms either go back endlessly, or if the Universe is finite, they just start. In either case the question of ultimate origin is as unanswerable as the question of the book's origin. What makes us think that when such questions are asked about the loop they are different and *ought* to be answerable is that the entire loop is open to inspection." While the instructions in the book don't tarnish, the book itself of course brings us back to our previous antique watch 'problem.' Suppose the book is brand-new at the start of the trip backward in time. Later, when the machine (and the book) have reached the end of the loop, just before beginning the trip back in time, have the pages turned yellow and brittle? If so, how do we account for the brand-new version? And if not, why not?

An analyst who takes strong exception to these two philosophers is Oxford physicist David Deutsch, who wrote (note 100) "the real problem with closed timelike lines under classical physics is that they could be used to generate knowledge in a way that conflicts with the principles of the philosophy of science, specifically with the evolutionary principle." What Deutsch is referring to is the metaphysical claim, attributed to the philosopher Karl Popper (see note 36 in Chap. 3), that *knowledge comes into existence only by evolutionary, rational processes* and that solutions to problems do not spring fully formed into the universe. One might call this the physics version of the work ethic—the creation of knowledge demands hard work!

Deutsch's idea had actually appeared *decades* earlier in a science fiction tale.[113] Time travel, discovered in the year 2007, is found to have a limited temporal reach into the future of 50 years, a limit due a law passed in 2057 banning time travelers from the past. To try to go past 2057 leads to a prompt arrest of the time traveler and a 'deportation' trip back to his own time. The story eventually explains that the law was passed precisely because of Deutsch's concern. As one character in the story explains, "Suppose [that one could travel more than 50 years ahead], then a time traveler from the past could get [new inventions], carry them back to his own time, and give them to scientists—which[would] cancel all the long period of invention which [produced the inventions]. Which [would] violate causal laws."

More recently, a philosopher has offered a quite interesting response to the Deutsch/Popper assertion. He writes (note 81), of information "appearing out of nowhere," that "These cases are puzzling, but they by no means show that the time travel scenarios in question are impossible or incoherent, *or even improbable*. We think it very improbable that . . . information should come from nowhere—but only because this does not happen very often. It does happen *sometimes*—for instance, when you say something and I mishear you. I think that you said something very

---

[113] P. Anderson and G. Dickson, "Trespass," *Fantastic Story Quarterly*, Spring 1950.

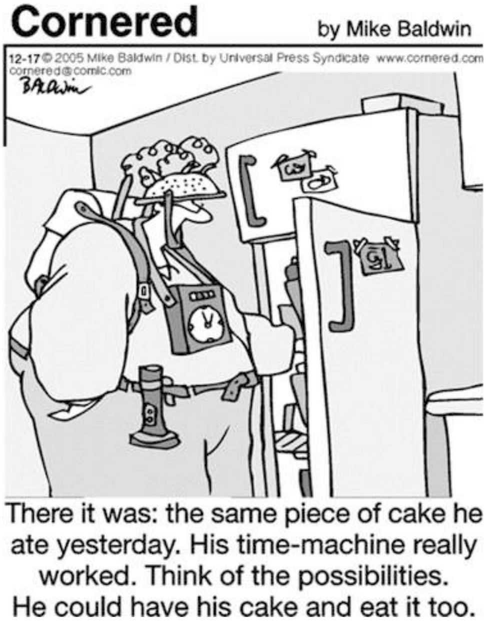
Cornered
by Mike Baldwin
12-17 © 2005 Mike Baldwin / Dist. by Universal Press Syndicate www.cornered.com
cornered@comic.com
There it was: the same piece of cake he ate yesterday. His time-machine really worked. Think of the possibilities. He could have his cake and eat it too.

Fig. 4.5 A curious paradox. CORNERED ©2005 Mike Baldwin. Reprinted with permission of UNIVERSAL UCLICK. All rights reserved

profound—something which neither of us would, in fact, ever have thought of. Where does the idea come from? If this sort of thing were to start occurring regularly [as via causal loops], then we would simply accept it without raising an eyebrow."

In the final chapter I'll discuss a dramatic example (due to two Russian physicists) on how an information-creating time loop might be constructed using a wormhole time machine. Such a time loop wouldn't pass muster with Deutsch, of course, and he would consider such a thing as being as objectionable as is creationism, the anti-evolution claim that purports to 'explain' fossils (with measured ages in the millions of years) by simply *declaring* them as having been made by God just a few thousand years ago.[114] Deutsch's position is considered by nearly all scientists today to be correct *for the specific case of creationism*, but the evolutionary principle may be on shakier ground with respect to declaring causal information time loops to be impossible.

While philosophers have struggled with information in a time loop, and most physicists have carefully stepped around the issue, science fiction has had lots of fun with information in causal loops. Here's a sampling of such tales:

(1) A man receives telephone calls from *two* versions of himself, one ten years in the future saying he absolutely must accept an invitation to fly to the Bahamas

---

[114] *Why* would God do such a thing? Apparently 'just to have some fun with geologists and biologists,' as creationists call such ancient fossils 'sports of nature.'

that he will receive that very day, with the other version calling from tomorrow insisting that the plane will crash. What should he do?[115];

(2) Lovers who are irrevocably separated in time communicate by mail in one tale,[116] while lovers in another story[117] communicate via telephone calls to the ever more distant past (and yet, with the aid of a clever twist at the end, finally meet);

(3) A telephone lineman starts getting telephone calls from himself from 10 days in the future, with the first call telling him how to make the gadget to transmit such calls[118];

(4) A time traveling historian on a visit to A.D. 1528 from A.D. 2211 accidently gives a copy of the predictions of Nostrodamus to the prophet, thus explaining the predictions[119];

(5) A time machine experiment gone wrong allows thirteenth century Roger Bacon to meet twentieth century scientists, an encounter that explains the amazing forecasts in Bacon's *Opus Maius*[120];

Hollywood, too, has had some fun with information causal loops, with the best (in my opinion) example of that being the 1989 movie *Bill & Ted's Excellent Adventure*. In that film (where we learn that even the not very bright can be time travelers), a set of missing keys is necessary for the successful completion of a task. The two time travelers decide that after the task is done, they will go back in time, steal the keys (*that's* why they're missing!), and hide them so they can use them *now*. Where should they hide them? Why, "over there," says one of the boys, pointing at a hiding place—and sure enough, when they go over and look, the keys are there. They agree that once they have finished with the keys, it will be *most* important that they really do put the keys in the hiding place!

All of these examples that I've just given you, however, were *decades* too late to be the first in fiction about information in a time loop; that honor goes to the 1904 novel *The Panchronicon* by the lawyer Harold Steele MacKaye (1866–1928). An Edwardian literary time machine with style, the Panchronicon (literally, a 'machine for all time') swings on a rope tether around a steel post erected at the North Pole. By "cutting the meridians" faster than the sun does, it travels through space and time from 1898 New Hampshire to the London of three centuries earlier.[121] Using it, a time traveler fan of Shakespeare journeys from 1898 back to the bard, who is suffering from writer's block. There she whispers the magic lines from a play he is stuck on (lines she has memorized for her literary club meetings) into his receptive ear. Does this make Shakespeare a plagiarist? Of himself!?

---

[115] G. Klein, "Party Line," *The Best from the Rest of the World* (D. A. Wolheim, editor), Doubleday 1976 (story originally published in France in 1973).

[116] J. Finney, "The Love Letter," *The Saturday Evening Post*, August 1959.

[117] T. N. Scortia, "When You Hear the Tone," *Galaxy Science Fiction*, January 1971. See also L. Padgett, "Line to Tomorrow," *Astounding Science Fiction*, November 1945.

[118] M. Leinster, "Sam, This Is You," *Galaxy Science Fiction*, May 1955. This story was later broadcast as an episode on the "X-Minus One" radio drama program. See also F. A. Reeds, "Forever Is Not So Long," *Astounding Science Fiction*, May 1942.

[119] L. Del Rey, "Fools' Errand," *Science Fiction Quarterly*, November 1951.

[120] N. Schachner, "Lost in the Dimensions," *Astounding Stories*, November 1937.

[121] 'Time traveling' by crossing time zones is an idea that one can trace at least as far back as to Edgar Allen Poe's 1841 short story "Three Sundays in a Week."

4.5 Sexual Paradoxes

> "Once time machines exist, no event is low probability if it is needed to make the past consistent."[122]

There are causal loops even stranger than the ones we have already discussed, hard as that may be to believe. These are the sexual paradoxes, first mentioned in 1931 by Nicholson in his letter to Hugo Gernsback. Not only science fiction writers, but philosophers, too, have found these particular paradoxes full of dramatic appeal. For example, as a challenge problem to the readers of a scholarly journal, the British philosopher Jonathan Harrison (1924–2014) posed the following bizarre, indeed astonishing, situation.[123] A young lady, Jocasta Jones, one day finds an ancient deep freezer containing a solidly frozen young man. She thaws him out and learns that his name is Dum, and that he possesses a book that describes how to make both a deep freezer and a time machine. They marry. Soon after they have a baby boy and name him Dee.

Years later, after reading his father's book, Dee makes a time machine. Dee and Dum, taking the book with them, get into the machine and begin a trip into the past. Running out of food during the lengthy journey, Dee kills his father and eats him. Arriving in the past, Dee destroys the time machine, builds a deep freezer (again, using the book), gets into it, and . . . wakes up to find that a young lady, one Jocasta Jones, has thawed him out. When asked his name he replies Dum and shows Jocasta his book; they marry, and . . . .

Harrison concluded this amazing tale with this question for his readers: "Did Jocasta commit a logically possible crime?" That issue is just the *surface* of an ocean of puzzles in this story! Jocasta's crime, of course, is that she has (if unwittingly) committed incest; readers who remember the Greek myth of Oedipus, and who his mother/wife was, will understand why Harrison named his female character as he did. But what of Dee's crime? He has, after all, eaten his father! But perhaps that isn't a crime at all, because Dee and Dum are one in the same, and is it really a crime to eat yourself? According to another philosopher, Murray MacBeath, Harrison's story is "a story so extravagant in its implications that it will be regarded as an effective *reductio ad absurdum* of the one dubious assumption on which the story rests: the possibility of time travel."[124]

---

[122] From Robert Forward's 1992 novel *Timemaster*.

[123] J. Harrison, "Jocasta's Crime," *Analysis*, March 1979, p. 65.

[124] M. MacBeath, "Who Was Dr. Who's Father?" *Synthese*, June 1982, pp. 397–430.

This isn't to say that MacBeath was asserting that time travel is impossible. Indeed, he went on to declare that he did believe in the logical possibility of time travel, and his paper is devoted to discovering what thought to be incorrect in Harrison's story. He did that by retelling the story with what he believed are crucial modifications to make it sufficiently less outrageous that it could be taken at least somewhat seriously. In the new version, our hero, thawed out from a deep freezer, is now named Arthur. Arthur is, unfortunately, suffering from total amnesia (this is MacBeath's way of avoiding the psychological trauma of Dee remembering he ate Dum) and so, when asked his full name, he is himself sufficiently puzzled that he replies "Arthur who?" He is finally called (what else?) Arthur Who. And, as you can no doubt guess, his son (who is a genius and gets a PhD at age 14 on a dissertation dealing with the physics of time travel) becomes Dr. Who!

We are then told of a trip back into the past by the two, of the eating of the father (Arthur Who) by the son, of the entering of the deep freezer by Dr. Who, etc. etc. The whole business is quite entertaining and *at least* as complex as Harrison's original story. Just *how* complex is summed up in MacBeath's last, wonderful line: "The Who who was Dr. Who's father was not Dr. Who—that is, not the Dr. Who whose father he was."

MacBeath wasn't the only one that Harrison's story fascinated, and nearly a dozen replies to it were received in addition to MacBeath's. One, in particular, made the thought-provoking observation (see note 112) that not only has Jocasta committed incest but she has done so with a single act of intercourse. As discussed earlier, the events on a causal loop do not happen endlessly but rather only once; thus, Jocasta thaws Dum (Dee) out just once, she marries him just once, and the two consummate their marriage just once. Ordinarily we think it takes two sexual acts to commit incest, the first resulting in the birth of a child, and the second being a parent's union with that child, but this is not so in a causal loop. Time travel *is* an odd business.

Another philosopher replied to Harrison's story with a quite interesting claim, one that had actually been thought to be true for decades—but which today is recognized to be false. The claim was that, irrespective of physics, Harrison's story was biologically flawed and fatally so. As that philosopher wrote, "The biological problem is the following. Dee is the son of Dum and Jocasta. So Dee obtained half his genes from Dum and half from Jocasta. But Dum is diachronically identical with Dee and is therefore genotypically identical with him (that is, himself). That is, Dee is both genotypically identical and distinct from Dum, which is absurd."[125]

That this isn't true was pointed out by a philosopher many years later. In his paper we read this tale: "Suppose Adam travels [far] back in time ... where he meets his mother Betty, mates with her and has a child which is himself. Is this possible biologically? Yes ... as follows ... on the grounds that we have total replication of Adam's genome."[126] (The *genome* is the totality of genes taken over all gene sites.) Now, suppose each such site holds two genes and, as Dowe points out, in sexual reproduction the father passes on to his offspring one gene for each gene site, to go with the gene the mother gives to each site. To exactly reproduce himself, then, the time traveling Adam 'simply' has to give his offspring, *at each site*, the gene that he has for that site that did *not* come from Betty. Thus, the offspring—baby Adam—ends up with a genome *precisely identical* to the time traveling Adam. This is, of course, an *extraordinarily unlikely* event, as the human genome has tens of thousands of genes. The probability that each and every site gets the 'right' gene from the time traveling Adam is therefore essentially zero. But it isn't *actually* zero and, as the quotation that opens this section says, a low probability to an event isn't a roadblock to its occurrence if that event is required for consistency.

---

[125] W. Godfrey-Smith, "Traveling in Time," *Analysis*, March 1980, pp. 72–73. This false claim had already been raised by a physicist (L. S. Schulman, "Tachyon Paradoxes," *American Journal of Physics*, May 1971, pp. 481–484), and even earlier by a science fiction writer (P. Anderson, "Time Patrol," *Magazine of Fantasy and Science Fiction*, May 1955).

While certainly instructive, the sexual paradox stories by Harrison and MacBeath are remiss in not indicating that the concepts they are dealing with have long been a staple of science fiction, and that the sexual paradoxes received much critical analysis in that genre long before philosophers (and physicists, too) discovered them. From science fiction, for example, we have a tale of young man who travels backward in time 1250 years, from A.D. 3207 to 1957, to become his own grandfather fifty generations removed.[127] And even that is tame compared to the sexual paradoxes other science fiction writers conjured up before philosophers began to discuss them.

In another story,[128] written decades before Harrison's and MacBeath's papers, we meet a young lady caught up in a mind twisting affair in which the mystery of a causal loop is the least of her troubles. In 1957 a girl is born, and after 20 years of intense competition with her mother (who has an uncanny ability to predict the future), she travels back from 1977 to a few months before her own birth. She becomes pregnant (by a man who she later discovers is her father) and gives birth to a girl. The new mother has, of course, knowledge of all that will happen during the next 20 years, including the fact that she will have an intense competition with her rebellious daughter . . . .

While writers of stories like these in the early 1950s were there as trailblazers, it is a tale that appeared as the 1950s ended that is today generally acknowledged as the best sexual paradox story ever written.[129] We are given only a hint of what is to come when a character listens to a song called "I'm My Own Grandpaw!" In 1945, a newborn girl, Jane, is found on the steps of an orphanage. At age 18, in 1963, she has a one-night affair with a mysterious stranger that leaves her pregnant. Some months later, during the birth of a daughter, it is discovered that Jane actually has a double set of sexual organs, and because the female set has been ruined by the pregnancy, doctors restore her as a man. Soon after, the baby girl mysteriously disappears from the hospital ward. Years later, in 1970, Jane (now a man, of course) meets another stranger who uses a time machine to transport both of them back to April 3, 1963. By April 24 male-Jane meets female-Jane and impregnates her (and so now we know who the mysterious stranger was during the one-night affair!). Meanwhile, the stranger with the time machine travels forward to March 10, 1964, a little after female-Jane has given birth, kidnaps the baby from the hospital (thus clearing-up another mystery!), takes her back to September 20, 1945, and leaves her on the steps of the orphanage. And so we see that Jane is her own mother *and father*, thus out-doing all previous tales about self-parenting.

---

[126] Phil Dowe, "The Coincidences of Time Travel," *Philosophy of Science*, July 2003, pp. 574–589. See also J. Berkovich, "On Chance in Causal Loops," *Mind*, January 2001, pp. 1–23, and P. Dowe, "Causal Loops and Independence of Causal Facts," *Philosophy of Science*, September 2001, pp. 89–97.

[127] R. Dee (this is *not* the 'Dee' of Harrison's story!), "The Poundstone Paradox," *Magazine of Fantasy and Science Fiction*, May 1954.

[128] C. L. Harness, "Child By Chronos," *Magazine of Fantasy and Science Fiction*, June 1953.

[129] R. Heinlein, "All You Zombies—," *Magazine of Fantasy and Science Fiction*, March 1959.

This is pretty impressive stuff, but Heinlein still has one more twist for us. After leaving baby-Jane in 1945, the time machine stranger returns to April 24, 1963, retrieves male-Jane (who has just kissed female-Jane goodnight after fathering her-himself in herself), and takes him to 1985 where he recruits him into the Temporal Service—and finally, the stranger jumps forward to 1999, his 'real time.' At the end we at last learn that the stranger is, in fact, an even older version of male-Jane—*all* the central characters in the entire story are the *same* individual at various points along a single, highly twisted world line. The lone character in Heinlein's tale is truly a self-made man/woman in every sense of the phrase! This ultimate act of *creatio ex nihilo* has, correctly I think, been called "smaller than the minimal loop."[130]

Jane, in all her/his versions, is the only character in the story that appears to have purpose. In terrifying words that describe a causal loop, Heinlein ends the tale with an explanation of the story's title: "The Snake That Eats Its Own Tail, Forever and Ever. I know where I came from—but where did all you zombies come from? . . . You aren't really there at all. There isn't anybody but me—Jane—here alone in the dark. I miss you dreadfully!" In a December 1958 letter to his literary agent, Heinlein wrote of this amazing tale, "I *hope* that I have written in that story the Farthest South in time paradoxes." In my opinion, he did.

The sexual paradox has continued to fascinate science fiction writers up to the present day. In the novel *Timemaster* (note 122), for example, the hero at one point spends a night with his wife—and with two versions of himself from the future. He will, of course, experience that night two more times! Later, he becomes upset when his wife runs off with one of the older versions, but he quickly calms down when he considers that eventually *he* will be the older version. Consider, too, a story[131] that tells of a young man hunting the father who, years before, had abandoned him in a burning house. The death of the young man's mother in the flames has sent him on a 10 year quest for revenge up and down what is literally a river of time, a river on which to travel in one direction ("up time") is to move into the past, whereas moving "down time" leads to the future. Eventually he corners the father and, despite the man's pleading, kills him. It is only later, after examining papers he finds in his father's pocket, that the young man realizes he has killed his *future* self (Benford, a physicist, knows the pitfalls of time travel, and you'll notice that there is no autoinfanticide paradox here).

---

[130] S. Lem, "The Time-Travel Story and Related Matters of SF Structuring," *Science Fiction Studies*, Spring 1974, pp. 143–154.

[131] G. Benford, "Down the River Road," *After the King: Stories in Honor of J. R. Tolkien* (C. Tolkien and M. Greenberg, editors), Tor 1991.

4.6 Splitting Universes and Time Travel

> "In all time travel stories where someone enters the past the past is necessarily altered. The only way the logical contradictions created by such a premise can be resolved is by positing a Universe that splits into separate branches the instant the past is entered."[132]

One early science fiction technique for allowing backward time travel and a changeable past, while still avoiding paradoxes, is that of alternate universes. According to this idea, if a time traveler journeys into the past and introduces a change (indeed, his very journey may be the change) then, as the above quote states, reality splits into two versions, with one fork representing the result of the change and the other fork being the original reality before the change. (To a fifth-dimensional observer, of course, all conceivable forks, all possible four-dimensional spacetimes, have always existed.) Indeed, according to this view the entire universe is splitting, at every microinstant, along every alternative decision path for every particle in the cosmos! This is often called the theory of *alternate realities with parallel time tracks*.

Such a seemingly fantastic view seems to actually have some scientific plausibility because of the so-called *many-worlds interpretation* (MWI) of quantum mechanics, pioneered in physics by Hugh Everett III (1930–1982), in a 1957 Princeton doctoral dissertation. Everett's theory is the antithesis of what is commonly called the *collapse of the wave function*, the idea that all potential possibilities have a non-zero possibility until a consciousness actually decides or observes which one will actually *be*. That quantum mechanical concept gets its name from the probabilistic wave equation formulated in 1926 by the German physicist Erwin Schrödinger (1887–1961). Before the observation, all possible futures have various values of probability; after the observation (which 'collapses' the wave function) exactly *one* of those futures (*the* future) has probability 1 and all the others have probability 0.

The MWI idea can be seen in Hale's story "Hands Off," discussed earlier, and in art *40 years* before that! With almost certainly a theological twist, consider the

---

[132] M. Gardner, "Mathematical Games," *Scientific American*, March 1979.

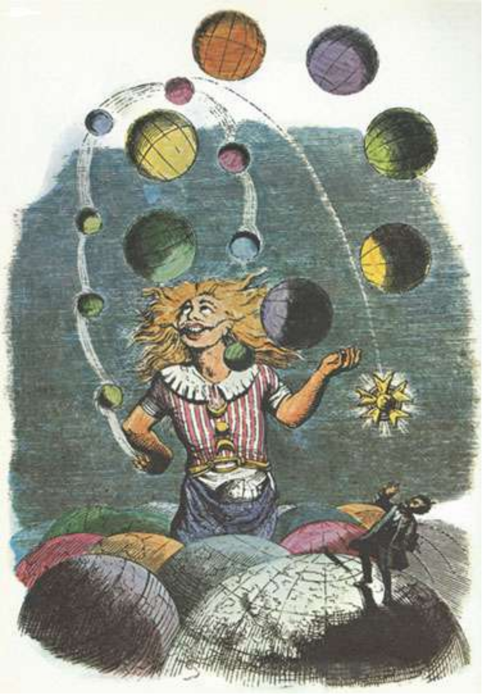

Fig. 4.6 Grandville's Infinity Juggler of many-worlds

beautiful illustration in the 1844 book *Un Autre Monde* (*Another World*), reproduced in Fig. 4.6. Known either as "The Infinity Juggler" or "The Juggler of Worlds," it is the work of the French artist Jean-Ignace Isidore Gérard (1803–1847), who published under the name 'Grandville.' The juggler—Grandville's version of Hale's mentor—appears as a court jester who is clearly having fun manipulating his multitude of worlds, while the man (humanity?) in the foreground watches. The man appears to be simultaneously fearful and fascinated, involved yet clearly impotent. Is Earth one of the worlds among which the Jester stands, or is it one of those flying through space? Or is Earth, perhaps, simply the unfortunate world ingloriously stuffed down the front of the Jester's pants? (That would surely explain a lot!) If born a hundred years later, Grandville would surely have found work as an artist in the imaginative world of the science fiction pulps.

Early science fiction stories that treat the collapsing wave function concept can be traced back to the late 1930s and early 1940s.[133] A particularly interesting example is the story of an inventor who, while trying to build a radio with which to signal Mars, accidently stumbles on the "temporal-aberrant carrier wave" and thus establishes contact with a universe that forked off of ours in 1863 when Robert E. Lee *won* the Battle of Gettysburg![134] In Everett's MWI, however, the wave function of the universe does *not* collapse. Indeed, it *couldn't*, because there is no observer external to the entire universe (we are talking science now, not of theology and God); instead, the wave function 'splits' at every decision point in spacetime. Although this leads to a multitude of realities *far* beyond comprehension, cosmologists still tend to like the MWI because it avoids the puzzle of having to produce an observer 'outside the universe.'

---

[133] See, for example, Jack Williamson's 1938 novel *The Legion of Time*, and C. L. Moore's, "Tryst in Time," *Astounding Stories*, December 1936. L. Sprague de Camp (1907–2000), too, was an early pioneer in the exploration of the MWI idea in science fiction long *before* Everett. In his 1941 novel *Lest Darkness Fall*, for example, he uses the analogy of a tree (the "main time line") that is always sprouting new branches.

It's important to understand that the MWI is different from yet another idea popular in science fiction, that of *parallel universes* (see again the third discussion question at the end of Chap. 3). In parallel universes *all* possibilities *always* exist, independent and parallel in time. In the MWI, on the other hand, ever more universes are continually coming into existence. Unlike the MWI, which can at least claim a scientific basis (quantum mechanics), there is no analogous theory for parallel universes. But, of course, even though lacking a theory, nonetheless science fiction writers have been quite inventive with the idea because parallel universes offer a way to avoid (at least some) causal loops.

One clever, early pulp story[135] illustrates how that works. To improve the performance of his time machine, an inventor needs batteries with tremendous energy density, a density far in advance of the batteries in the present. Unable to travel far into the future—if he *could* obtain them there, then of course a causal loop (the very entity we wish to avoid) would be created upon his bringing them back to the present—his assistant first travels back to 1851. There he leaves a note on desk of a well-known experimenter, with a plea for him to devote his life to battery research; a copy of the 1937 *Electrical Handbook* is left with the note as proof that there really has been a visit from the future! Before returning to the present, the assistant takes a sheet of (new) 1847 five-cent stamps from the experimenter's desk.

Returning to the present, which is now different (a new time track, in accordance with the splitting-universe idea), the powerful batteries *are* readily available *because* the experimenter believed the note. Buying several of them, using money obtained by selling the pristine 1847 stamps to a collector, the assistant returns to a slightly earlier 1851 than before (to *before* the fork in time!), watches himself appear[136] and leave the note and the handbook, and *then, unobserved, the assistant removes both*: Thus, upon returning once more to the present, he finds all is as before—except now he and the inventor have the powerful batteries. As before, one might ask where the batteries came from, but unlike the previous mystery of information-creating causal loops, the answer is clear and non-mysterious. *They came from the hard work of the experimenter on a different time track.* Such shuttling back-and-forth between time tracks is the signature of what is called a *cross-time* story, a device to avoid paradoxes while still allowing for changing the past. The first example of this time travel sub-genre had actually appeared 4 years earlier.[137]

---

[134] N. Bond, "Parallel in Time," *Thrilling Wonder Stories*, June 1940. See also S. N. Faber, "Trans Dimensional Imports," *Isaac Asimov's Science Fiction Magazine*, August 1980.

[135] W. Sell, "Other Tracks," *Astounding Science Fiction*, October 1938.

[136] Just like Marty McFly does at the end of the 1985 film *Back to the Future*. The movie is fun, but pulp science fiction did it first.

In another cross-time tale from modern times, we read of the horrible fate suffered by a man when an experiment in a Princeton physics lab goes wrong.[138] It is discovered, too late, that parallel time tracks are not simply grooves into which you drop, like a ball, after leaving the time track of our world. Each version of a person in each world is not like a ball rolling down a groove from past to future. Rather, each world's time track is just a line on a smooth surface; as the man is told, during a temporary stay in a world still close to his (our) original world, "We gave you a push sideways, and you moved off your original line—but instead of dropping into the next groove, you've just kept on rolling across the surface, from one line to the next, at an angle. There are no grooves, nothing to stop you from sliding on across the different lines forever. You have the same futureward vector as you started with, but you've added a small cross-time vector, as well." And so the man drifts cross-time, and gradually the worlds he experiences grow ever more alien.[139]

The science fiction is undeniably fun, but for this book the underlying *scientific* theory of time travel is classical (that is, non-quantum) general relativity, and that theory has nothing to say about alternative time tracks in multiple worlds. For most time travel theoreticians there is *one* time track, and the past of our world is unique and inviolate. I agree with the great quantum physicist J. S. Bell (1928–1990), who wrote of Everett's theory that "if such a theory were taken seriously it would hardly be possible to take anything else seriously."[140] As Bell further observed, in the MWI "there is no association of the particular present with any particular past," a quite strange idea that had already appeared in science fiction years earlier.[141]

While most early time travel analysts did base their work just on classical general relativity, there *are* now many more who think quantum mechanics itself, independent of its interpretation, has much to contribute as well. Perhaps, in fact, it may make an absolutely crucial contribution to the theoretical basis of time travel. One analyst who believes this (along with an even stronger belief in the MWI) is the British physicist David Deustch (note 100), who holds that general relativity is not the proper theory with which to study the physical effects of CTLs. He believes that the traditional mathematical machinery of general relativity actually obscures, rather than clarifies, the difficult task of separating the merely counter-intuitive from the unphysical.

---

[137] M. Leinster, "Sidewise in Time," *Astounding Stories*, June 1934. Splitting universes with multiple time tracks and time loops became quite popular after Leinster's and Sell's stories; you can find the basic idea repeated yet again in Alfred Bester's "The Probable Man," *Astounding Science Fiction*, July 1941, for example, in which each new journey into the past causes the future to fan out into an infinity of new time tracks.

[138] L. Watt-Evans, "The Drifter," *Amazing Stories*, October 1991.

[139] In Jack Haldeman's 1990 novel *The Hemingway Hoax* we read that "there is not just one [parallel] universe, but actually uncountable zillions of them."

[140] In "Quantum Mechanics for Cosmologists," *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press 1987.

[141] J. R. Pierce, "Mr. Kinkaid's Pasts," *Magazine of Fantasy & Science Fiction*, August 1953.

Indeed, Deutsch calls the conventional spacetime methods, based on general relativity and differential geometry, *perverse*. He also does not like the conceptual problems raised by general relativity's wormholes and singularities. Any non-quantum mechanical discussion, he says, of the "pathologies" of backward time travel is simply not adequate. Deutsch divides these pathologies into two fundamental classes: (1) paradoxical constraints, such as the free-will issue seemingly raised by the grandfather paradox, and (2) causal information loops. Deutsch claims that his quantum mechanical analyses show that the first class of pathologies simply does not occur, because the past that the time traveler enters is the past of a world different from the one he has left. Further, his results also show (to him) that the pathologies of the second class may be "avoidable." These are not the views among the majority of time travel students, however (that does *not* mean Deutsch is wrong!), and general relativity *is* the standard tool used by the majority of time travel theoreticians. When quantum mechanics does enter the calculations of most analysts, it is generally on an *ad hoc* basis.

More concerning for the MWI view is a result reported in 2004, that a macroscopic object (a human time traveler, for example) attempting to traverse a wormhole time machine (to be discussed in some detail in Chap. 6) "must necessarily undergo violent interactions with the time machine," interactions *so* violent that they must "cause the object to disintegrate." The different pieces of the now certainly dead time traveler would emerge from the wormhole in different worlds—this is definitely not a result likely to encourage volunteers for the first time machine trip![142]

So, many physicists and philosophers, not sharing Deutsch's position,[143] tend to agree with Bell, including the late John Wheeler (Everett's thesis advisor!), who wrote of the MWI "I once subscribed to it. In retrospect, however, it looks like the wrong track. . . . Its infinitely many unobservable worlds make a heavy load of metaphysical baggage."[144] Agreeing with Wheeler was a philosopher who called the MWI "highly controversial" and declared that "few working physicists take it seriously."[145] Perhaps even more damning was a physicist's statement that "the idea of 10^{100}+ slightly imperfect copies of (the universe) all constantly splitting into further copies . . . is not easy to reconcile with commonsense. Here is schizophrenia with a vengeance."[146] Or, as one science fiction writer bluntly put it, in a tale of the inventor of the "chronomotive impulse belt" (which allows moving between the *two* parallel worlds that are all that exist), the MWI is the "Doctrine of Infinite Redundancy—which is, of course, utter nonsense."[147]

---

[142] A. Everett, "Time Travel Paradoxes, Path Integrals, and the Many Worlds Interpretation of Quantum Mechanics," *Physical Review D*, June 25, 2004, pp. 124023–1:124023–14.

[143] For a modern philosophical argument specifically rebutting Deutsch's enthusiasm for the MWI, see Theodore Sider, "A New Grandfather Paradox?" *Philosophy and Phenomenological Research*, March 1997, pp. 139–144.

[144] See note 34 in Chap. 3.

Deutsch's position does raise the obvious question of what motivates a quantum theoretician to study CTLs at all, given that they originate in general relativity and *not* in quantum mechanics. Deutsch's response is that although CTLs did indeed originate in classical Einsteinian general relativity, the still incomplete theory of quantum gravity *does* predict CTLs, too. And that is an exciting observation for time travel enthusiasts because, as Deutsch writes, the results of his quantum studies of CTLs show that "contrary to what has usually been assumed, there is no reason in what we know of fundamental physics why closed timelike lines should not exist." That view was later endorsed by other physicists who wrote, after a quantum mechanical study of how a particle could transit a time machine spacetime in a physically consistent manner, "there is no contradiction between the postulates of quantum mechanics and the possible existence of causality violation in general relativity."[148]

Long before these scientific endorsements, science fiction had enthusiastically embraced the many-worlds idea and its connection with time travel. The first such tale[149] appeared when Everett was just 3 years old; it put forth the insightful observation that although alternate time tracks may allow changing the past for the better (something that can't be done, for better *or* for worse, with a single time track), in the end any such change may still be futile. As Daniels' time traveler puts it, "I did have an idea to . . . go back to make past ages more livable. Terrible things have happened in history, you know. But it isn't any use. Think, for instance, of the martyrs and the things they suffered. I could go back and save them those wrongs. And yet all the time . . . they would still have known their unhappiness and their agony, because in this world-line those things happened. At the end, it's all unchangeable; it merely unrolls before us."[150] Many years later, in a critique of the many-worlds idea, a philosopher/physicist echoed Daniels' words: [In the world] that I (subjectively) experience I may blunder, but [in another world], with equal actuality, I triumph gloriously. The Everett interpretation can be used this way to mitigate sorrows, but this use is two-edged, for it equally well implies the speciousness of happiness."[151]

---

[145] R. A. Healy, "How Many Worlds?" *Nous*, November 1984, pp. 591–616.

[146] B. S. DeWitt, "Quantum Mechanics and Reality,' in *The Many-Worlds Interpretation of Quantum Mechanics* (B. S. DeWitt and N. Graham, editors), Princeton 1973.

[147] B. Shaw, "What Time Do You Call This?" *Amazing Science Fiction*, September 1971. When a bank robber in one world tries to make his escape into the other world, he literally runs into 'himself' trying to escape after robbing the 'same' bank in the parallel world!

[148] D. S. Goldwirth *et al.*, "Quantum Propagator for a Nonrelativistic Particle in the Vicinity of a Time Machine," *Physical Review D*, April 15, 1994, pp. 3951–3957. See, too, the earlier D. S. Goldwirth *et al.*, "The Breakdown of Quantum Mechanics in the Presence of Time Machines," *General Relativity and Gravitation*, January 1993, pp. 7–13.

[149] D. R. Daniels, "The Branches of Time," *Wonder Stories*, August 1935.

The editorial introduction to Daniels' pioneering tale is quite interesting: the opening line is "To say that this short story contains some revolutionary time-travel theories would be putting it exceedingly mild." That editor then went on to tell his readers, with great enthusiasm, that "when the author . . . submitted this story to us, his accompanying letter stated that in it he had settled the time-travel question once and for all. We must admit that a broad, unbelieving grin spread over our countenances when the author dared make this assertion. BUT—the smile soon left our faces . . . [T]o our chagrin, Mr. Daniels had really propounded so many brand new ideas about time and time-travel, and such logical ones—*that he has not left one loophole in his argument*!"

John W. Campbell (1910–1971), the first (and only) editor of *Astounding Science Fiction* (today's *Analog*), called alternate time track stories "mutant" because they represented the first new innovation (or mutation) in the time travel concept since H. G. Wells. Campbell incorrectly claimed *The Legion of Time* (note 133) was the first such tale (see Campbell's editorial in the May 1938 issue of *Astounding*) , and that "Other Tracks" (note 135) was the second, but in fact it was Daniels who was first with splitting time tracks in science fiction. After Daniels the concept quickly became part of standard science fiction lore and could be used by other writers with little explanation. For example, just a little more than a decade later one author did not have to say much about his "First Law of Chronistics," which determines the development of "the branches of Fan-Shaped time." It was sufficient for his readers to learn that should a time traveler to the past change anything, a parallel branch of time would be created on which the time traveler would be trapped: "The man who interfered with the space-time matrix, displacing even a comma in the great scroll of time, would be cut-off from his origin forever."[152]

Still, if there is one thing we can say about science fiction, it's that no 'rule' is immune to challenge. Decades after Daniels' tale put forth the MWI, we find the well-known author James Blish (note 100 in Chap. 1) *rejecting* it. In a story about the reception of radio signals from the future, we read of one character telling another "I *was* going to do all those things. There were no alternatives, no fanciful 'branches in time,' no decision-points that might be altered to make the future change. My future, like yours . . . and everybody else's, was fixed. It didn't matter a snap whether or not I had a decent motive for what I was going to do; I was going to do it anyhow. Cause and effect . . . just don't exist. One event follows another because events are just as indestructible in space-time as matter and energy are."[153]

---

[150] These sad, resigned words were written when the author, David R. Daniels (1915–1936), was just twenty years old. A year later he committed suicide.

[151] A. Shimony, "Events and Processes in the Quantum World," *Quantum Concepts in Space and Time* (R. Penrose and C. J. Isham, editors), Oxford University Press 1986.

[152] J. MacCreigh, "A Hitch in Time," *Thrilling Wonder Stories*, June 1947.

This denial of the MWI is simply an author's choice, of course, for whatever story effect is desired, and others may make different choices. Isaac Asimov, for example, used the MWI idea in in the story of a time traveler who journeys back to 1871 London, to retrieve a lost Gilbert and Sullivan operetta (*Thespis*). When he returns to the present he finds that his wife Mary (who was alive when he left) has been dead for a year on the new time track that his actions in the past have created. As the story ends, the devastated time traveler thinks "I had changed history. I could never go back. I had gained *Thespis*. I had lost Mary."[154] This sad fate is repeated in another story of a time traveler lost in an infinitude of time tracks with no hope of ever finding his way home: "In all of time, how many, many worlds there must be. How to find a single twig in such a forest?"[155]

Splitting universes have been used in literary works outside the genre of science fiction, as well. Examples include "The Garden of Forking Paths" by the Argentine writer J. L. Borges, the first play J. B. Priestly wrote (the 1932 *Dangerous Corner*), John Updike's 1997 novel *Toward the End of Time*, and Gore Vidal's 1998 novel *The Smithsonian Institution*. Typical of these fictional fantasies about splitting universes is a tale (anticipating Asimov's) by Lord Dunsany (1876–1957)—the Irish writer Edward Plunkett—the story of a man who goes back in time to correct "two or three mistakes he had made in his life."[156] This he successfully does, but the result is a new, subtly different subsequent history. The differences are not infinitely subtle, however; after the changes, he finds that his home, his wife, and all the delicate details of his life have vanished. As he relates to a visitor at the lunatic asylum he is now confined to, as the result of his despair, "I tell you I'm lost. Can't you realize that I'm lost in time? I tell you that you can find your way traveling the length of Orion, sooner than you shall find it among the years . . . Don't go back down the years trying to alter anything . . . Don't even wish to . . . [T]he whole length of the Milky Way is more easily traveled than time, amongst whose terrible ages I am lost."

In writing for a mass audience, rather than just for the more limited science fiction and fantasy one, perhaps the best known literary work of alternate history is the classic 1953 novel *Bring the Jubilee* by Ward Moore (1903–1978). In that work Lee wins the Battle of Gettysburg, and the South wins the Civil War. Using a time machine, a historian travels from 1952 (of the world in which the South wins) into the past of 1863 to study the battle, where he inadvertently disrupts events to the point that the North wins; that is, reality splits and the newly created fork represents the time track of *our* world. The historian is trapped on this new fork, cut off forever from his original time track. The entire novel is in the form of a discovered manuscript, written in 1873 and found in 1953, and the pathos of the ultimate isolation endows the novel with great emotional impact.

---

[153] J. Blish, "Beep," *Galaxy Science Fiction*, February 1954.

[154] I. Asimov, "Fair Exchange?" *Asimov's Science Fiction Adventure Magazine*, Fall 1978.

[155] M. F. Flynn, "The Forest of Time," *Analog Science Fiction*, June 1987.

[156] Lord Dunsany, "Lost," *The Fourth Book of Jorkens*, Arkham House 1948.

A 1992 novel on the same theme, Harry Turtledove's *The Guns of the South*, begins with a fascinating premise but then misses the crucial distinction between a single versus multiple time tracks. In that work racists from the future (2014) arrive by time machine at Lee's 1864 winter camp. They bring with them AK-47 automatic assault rifles and offer to supply Lee's army with all it can use. Lee accepts and the South wins the Civil War. The future, of course, changes—or does it? The time travelers have brought back books from the future showing that the South lost the war, so the implication is that history must have forked. So far, so good. But all through the novel, the time travelers move back and forth between the nineteenth and the twenty-first centuries, apparently finding their own time unchanged. And if that is so, then the whole point of the story vanishes. Why all the effort to change history when it is clear that nothing has changed? The novel is entertaining reading (Turtledove is a trained historian), but I believe Moore's novel to be the superior work of science fiction.

I'll end this discussion on splitting universes with a startling theological issue raised by a philosopher.[157] Arguing that God cannot branch into multiple time tracks because God is unique, the conclusion seems inescapable that God therefore exists on exactly one of how ever many different time tracks there may be. What if that chosen time track isn't ours? Then, concludes the philosopher, Nietzsche's nineteenth-century metaphorical claim that "God is dead" (for us) might literally be true! He admits that this is "fanciful," but still ....

4.7 For Further Discussion

In an afterword to his story "Dead City" (*Thrilling Wonder Stories*, Summer 1946), Murray Leinster muses "You've heard the old argument that a man can't travel backward in time because he might kill his grandfather. I've wondered why nobody has argued that a man can't travel forward in time because he might be killed by his grandson." One possible answer to Leinster is that *if*, at the moment the forward-bound time traveler departs, he has not yet sired a child, *then* there simply wouldn't be a murderous grandson waiting for him in the future. Perhaps, however, Leinster had this somewhat more

(continued)

---

[157] Q. Smith, "A New Topology of Temporal and Atemporal Permanence," *Nous*, June 1989, pp. 307–330.

complicated scenario in mind: After the time traveler arrives in the future he is attacked by a mysterious stranger but survives, and later returns to the present. He then sires a child who will be the parent of that mysterious stranger. (As far as I know, this plot line has not appeared in a science fiction story.) Contrary to Leinster's view, explain why the *possibility* of being killed by a grandson is not a reason for forbidding the possibility of a trip in time (in either direction).

In his causal loop paper (note 105) the University of Delaware philosopher Richard Hanley correctly writes (on p. 146) "physicists have tried to avoid free will problems by ignoring causal loops involving intentional agency," and partly illustrates this claim with the autoinfanticide paradox, writing of the attempt of a time traveler to kill his younger self as inevitably failing because "the past is apparently brought about willy-nilly." (Hanley unfortunately then uses the story "Thompson's Time Traveling Theory" as an example of this—see note 77, and the end of the "Introduction"—when it is that time traveler's *grandfather* who is the intended target.) Discuss the merits of Hanley's claim, keeping in mind the end of Sect. 4.3 (in particular, note 81). If you are interested in genetics and astronomy as well as in time travel, then for extra credit comment on Hanley's claim (p. 137) that "one can extract information about my DNA from ... my astrological chart."

The fictional killing of Hitler was imagined in print even before World War II, in Geoffrey Household's intense 1939 novel *Rogue Male* (made into the 1941 film *Man Hunt*). And so it's not surprising that one of the popular change-the-past themes in science fiction is that of a time traveler killing the Führer. (This idea, somewhat oddly, appeared in the debates leading up to the 2016 American Presidential election, when one of the candidates, to show the toughness of his character—even though he opposed abortion—declared "Hell, yes, I'd kill baby Hitler! You gotta step up, man." This candidate did elaborate a bit, stating there might be some risk involved with tampering with the past.) Stories in this sub-genre include E. Norden's "The Primal Solution" (*Magazine of Fantasy & Science Fiction*, July 1977), W. R. Thompson's "The Plot to Save Hitler" (*Analog*, September 1993), L. del Rey's "My Name Is Legion" (*Astounding Science Fiction*, June 1942), and R. M. Farley's "I

(continued)

Killed Hitler" (*Weird Tales*, July 1941). Read some of these tales and compare the various repercussions envisioned by the authors following an assassination of Hitler by a time traveler.

After our discussion of Heinlein's time travel masterpiece "All You Zombies—," you might think it impossible to write a new story that exceeds it in complexity. That might well be true, but a modern masterpiece by Ted Chiang certainly gives it a good run for the money. "The Merchant and the Alchemist's Gate" (*Magazine of Fantasy & Science Fiction*, September 2007) uses a 'wormhole' that connects the present to the future 20 years hence, and it is stuffed with intertwined causal loops and information bootstraps. Read it and keep track of all such occurrences. How many did you find? The last line of the story clearly expresses the view that the past cannot be changed: "Nothing erases the past. There is repentance, there is atonement, and there is forgiveness. That is all, but that is enough." Is the story always faithful to this view of time travel?

In his paper (note 126) on the coincidences of time travel, the University of Queensland philosopher Phil Dowe writes "It's true that remote time travel [into the very distant past] does not allow for causal loops . . ." *Is* this true? Consider, as you think about this, the story "Time's Arrow" (*Science-Fantasy*, Summer 1950) by Arthur C. Clarke. In that tale geologists have just discovered, in a remote desert, the fossilized tracks of a monstrous creature, from fifty million years ago, tracks that indicate that the beast was in hot pursuit of fleeing prey. Before the geologists can unearth the entire set of tracks, to see if the pursuit was successfully completed, they are visited by a physicist who just happens to be conducting near-by experiments in time travel. (This proximity is explained by noting what better place to conduct time travel experiments, powered by atomic energy, than in a remote desert?) At one point during the visit, after being told of the ancient pursuit frozen in rock, the physicist muses "It would save you a lot of trouble, wouldn't it, if you could actually *see* what took place in the past, without having to infer it by these laborious and uncertain [geological] methods." This comment results in the Chief geologist paying a visit to the physicist's lab. After driving over in a car equipped with tires having "an odd zigzag pattern" in the tread,

(continued)

an accident suddenly sends the entire lab into the past. Soon after, the other geologists unearth the rest of the fossilized tracks, and learn what the creature's prey had been when they see a zigzag pattern in the rocks, tracks that show "the great reptile was about to make the final leap upon its desperately fleeing prey." Can you see how to modify this story so as to have a causal loop involving the very distant past? (For perhaps even more inspiration on thinking about causal loops, watch the 1980 movie *The Final Countdown*. In it the designer of a modern naval warship that temporarily travels back through time to the Pearl Harbor of December 6, 1941, turns out to be a crew member who was accidently left behind in the past. In the past he *will be* able to design the ship because he already knows how it *was* designed—by himself!)

Comment, at length, on the cartoon shown in Fig. 4.5. (Does it make logical sense?)

In the story "Salvation" by Jerry Oltion (*Analog*, December 2007) a physicist approaches the Universal Church of the Divine Revelation for money to build a time machine. He is blunt in making his case: "You could go back in time and meet Jesus. Assuming he existed." That statement causes (it should come as no surprise) not just a bit of pandemonium but, nonetheless, an influential Church leader decides to provide the funding. Why? Because later, while sitting in his office as he talks with the physicist, a sheet of paper suddenly appears in the air above the leader's desk and then flutters down to land on the telephone. Picking the paper up, the leader sees it is a sheet of his own letterhead, with writing in his own angular, precise handwriting, saying "It works. Give him the money. You almost named the dog Solomon." This convinces the leader because, as we are told, "Paper appearing out of nowhere was a good trick, but it might easily be just that: a trick. Duplicating his letterhead and his handwriting wouldn't be all that difficult either. [On the other hand] knowing the name [the leader] had considered but rejected for his German Shepard 15 years ago was a different level of feat entirely." The physicist seems to be startled by the appearance of the paper, too, and asks "May I see that?" His reaction convinces the leader it wasn't a staged event: "Well, I'll be damned," the physicist replies. Once the time machine is under construction, the two men realize they have to send the enigmatic message back in time to complete the loop. As they prepare to do so, the leader asks a curious question. After retrieving the mysterious sheet of paper from his desk,

(continued)

he says to the physicist “Should I send the original [the one he is holding in his hand], or should I write another?” The physicist replies with “Write a new one. If we send the original, we put it in a closed loop and [we’ll] never get it back. We don’t want to lose the first object to travel in time. We’ll want that for the Smithsonian someday.” Does this make sense? Also, comment on whether or not the dog’s name is a bootstrap paradox.

A perplexing little time travel paradox, one that I don’t think science fiction has yet treated (and I’m pretty sure physicists haven’t had anything to say about it either), was cooked-up by the English philosopher Robin Le Poidevin in his 2003 book *Travels in Four Dimensions: the enigmas of Space and Time* (Oxford, pp. 180–181). There he writes “Peter and Jane, both 20 years old, are out for a walk one day in 1999 when suddenly a time machine appears in front of them. Out steps a strangely familiar character who tells Jane that he has an important mission for her. She must step into the machine and travel to the year 2019, talking with her a diary the stranger hands to her. In that diary she must make a record of her trip. Obligingly, she does as she is asked and, on arrival, meets Peter, now aged 40. She tells Peter to travel back to 1999, taking with him the diary she now hands him, and recording his trip in it. On arrival in 1999, he meets two 20-year-olds called Peter and Jane, out for a walk, and he tells Jane that he has an important mission for her.” Le Poidevin then writes that “the really tricky question is: how many entries are there in the diary when Jane first steps into the machine? We imagine it blank. But this is the very same diary as the one Jane hands to the 40-year-old Peter, which then contains her entry. And by the time Peter arrives back in 1999, it will contain his entry, too. But then, if the diary already contained two entries when Jane was handed the diary, then it would contain three entries when she handed it to Peter, who would then add another one, so the diary would have contained four entries when it was first handed to Jane, and so on. If the problem is not immediately apparent, this is because we imagine an indefinite number of trips, but in fact there are just two: Jane’s trip to 2019 and Peter’s trip to 1999. So there ought to be a consistent answer to the question, how many entries are there in the diary? Yet, as we have seen, there does not appear to be a consistent answer.” Another philosopher soon claimed he *did* have the answer: namely, 2. Read his paper (Erik Carlson, “A New Time Travel Paradox Resolved,” *Philosophia*, December 2005, pp. 263–273), and either explain why you agree with Carlson’s reasoning or enthusiastically rebut it.

As in "The Time Eliminator" (Fig. 4.1), other stories have imagined gadgets that simply view the past, rather than visit it as would a time machine. This is done in an attempt to avoid paradoxes—but does it? Two stories that illustrate how just viewing the past risks affecting the past as much as time travel would, are Horace Gold's "The Biography Project" (*Galaxy Science Fiction*, September 1951) and Donald Franson's "One Time in Alexandria" (*Analog*, June 1980). In the first tale the Biotime Camera, operated by the Biofilm Institute, allows teams of biographers to film (alas, no sound!) and study the lives of past notable personages. Of particular interest are the lives of those who developed neurotic psychoses, such as Isaac Newton. And, indeed, the Biotime Camera does capture Newton's image as he begins to display increasingly disturbed behavior. We see Newton, for example, as he begins to peer into dark corners, looking for those who have come to spy on him. On his death bed, the biography team assigned to him reads his lips and discovers that his final words are "My guardian angel. You watched over me all my life. I am content to meet you now." It is then that the Biofilm Institute realizes what it has done. Newton *was* in fact being spied upon—by the Biotime Camera, which has not changed the past but has certainly affected it. In the second tale an archeologist uses a time viewer to read the lost manuscripts in the ancient library at Alexandria before it was completely destroyed in an inferno. The viewer uses an infrared beam—and it is the heat from that beam from the future that proves to be the origin of the fire in the past. Again, the past has been affected, but not changed, by time viewing. Is it true to claim, however, that such viewing gadgets could not be the source of other paradoxes, such as causal loops or information bootstraps? If you think that isn't a valid claim, give a counter-example.

In a story by Francis Flagg and Weaver Wright (a pseudonym used by Forrest J. Ackerman), "Time Twister" (*Thrilling Wonder Stories*, October 1947), we read the following exchange between the inventor of a time machine and his none-too-bright helper:

"You mean to say," he questioned incredulously, "that I could go back a hundred years?"

"If you had the proper machine in which to travel, yes."

"But that'd take me back to before I was born."

The Professor smiled tolerantly.

"Look at this diagram, Hank. This line is the time continuum. It incorporates space, too. [The authors didn't actually print a diagram with the story, but

(continued)

surely the Professor is using a Minkowski spacetime diagram]. This dot is you. It doesn't matter when you were born, or when you will die. You exist right now, that's the fact. Traveling into the past or future wouldn't make you grow any younger or older. Such a thought is naïve. Let me demonstrate the mechanics of it for you. If . . . we calculate with non-Euclidean mathematics . . ."

"It don't sound reasonable," the farmhand objected. "If I went back—"

"I know," interjected the Professor, "if you went back you might meet your own father as a young man and you'd be older than he, or maybe he and your mother would be kids going to school."

"Haw, haw! That'd be funny, that would."

What famous movie, made nearly 40 years later, does the end of the conversation remind you of? Hint: "flux capacitor."

As mentioned in the text, a famous science fiction example of *affecting* (but not changing) the past is "Behold the Man" by Michael Moorcock, the tale of a time traveler who arrives in ancient times during the very years of the ministry of Jesus as reported in the Bible. When he finds there is actually no such person, the time traveler takes the role himself and lives out the events as reported in the Gospels, including the Crucifixion. This is a powerful story, but it had already been done more than 15 years earlier, by Philip K. Dick, in his short story "The Skull" (*If*, September 1952). In his tale, Dick tells of a man from the twenty-second century who is sent by government authority back to the mid-twentieth century to kill the Founder of a religious movement, a movement that 'now,' 200 years later, threatens those same government authorities. History records that the Founder gave a powerful speech just before being arrested and executed, a speech that started the religious movement, and so the time traveling assassin is told to kill the Founder *before* he can give that speech. (The parallel between Jesus and the Founder should be obvious.) Read these two stories and compare and contrast how Moorcock and Dick handled time travel paradoxes. Comment, in particular, on the relationship between the assassin and the Founder. Moorcock's tale should be easy to find, and Dick's is available as a free pdf download (it is in the anthology *The Best of Philip K. Dick*, Halcyon Classics 2010, as well).

A subtle change-the-past sequence appears in the original *Back to the Future* film that is easy to miss. When the hero, Marty McFly, returns to 1955 in the time car, he leaves from the parking lot of the Twin Pines Mall, so named because of the two pine trees that stand nearby. Arriving in the past with literally a bang, the time car inadvertently destroys one of the (then) young pines. Near the end of the movie, when Marty returns to the future (1985), he finds that the mall is now called the Lone Pine Mall. This is charming and fun, indeed clever, but modern scholars of time travel reject it, and other claims of changing the past, as not being logical. (Shakespeare understood this point, when he has Lady Macbeth declare, concerning the murder of Banquo, “What’s done cannot be undone: to bed, to bed, to bed.”) What Marty’s trip *would* explain is why the mall would *always* have had the name of the Lone Pine Mall. Watch the movie and see how many other ‘change-the-past’ episodes you can find.

Chapter 5
Communication with the Past

"[As for travel to or for signaling the past] you'd have to exceed light speed which immediately entails the use of more than an infinite number of horsepowers."[1]

5.1 Reversed Time

"I have not discovered Mr. Wells' Time Machine."[2]

One way to communicate with the past is to 'simply' live backwards in time. Philosophers and other writers of speculative fiction were the first to wonder what things might be like in a world where the time asymmetry is reversed—that is, in a world where time 'runs backward.' Indeed, fascination with the idea of time reversal actually dates back thousands of years, long before science fiction, as it can be found in Plato's dialogue *Statesman*, written (most probably) 15 years before Plato's death in 347 B.C.

At one point, Plato offers an extended description of the world suddenly running backward in time in the ancient past. After one character is told that at that remote time "all mortal beings halted on their way to assuming the looks of old age, and each one began to grow backward," he asks "But how did living creatures come into being, Sir? How did they produce their offspring?" The answer is shocking: "Clearly . . . it was not of the order of nature in that era to beget children by intercourse . . . It is only to be expected that along with the reversal of the old men's course of life and their return to childhood, a new race of men should arise, too—a new race formed from men dead and long laid in Earth . . . Such resurrection

---

[1] An observation by Haskel van Manderpootz, professor of the "newer physics," in S. G. Weinbaum's "The Worlds of If," *Wonder Stories*, August 1935. Compared to the 'modest' Van Manderpootz, all other physicists in the world are a mere "pack of jackels, eating the crumbs of ideas that drop from [his] feast of thoughts."

[2] W. R. Inge (1860–1954), in his November 1920 Presidential Address to the Aristotelian Society at the University of London Club, in a sympathetic treatment of the possibility of a time-reversed world.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_5

of the dead was in keeping with the cosmic change, all creation being now turned in the reverse direction."

The reversed-time world is an important philosophical concept. Before the turn of the century, for example, Francis Bradley (one of the early proponents of the block universe, you will recall from Chap. 2), thought about reversed-time worlds and concluded that they would be quite odd: "Let us suppose . . . that there are beings whose lives run opposite to our own . . . *If* in any way *I* could experience *their* world, I should fail to understand it. Death would come before birth, the blow would follow the wound, and all must seem to be irrational."[3] A half-century later the South African philosopher J. N. Findlay (1903–1987) took Bradley's position of supporting a skeptical attitude towards the possibility of time-reversed worlds. Writing in a book review, Findlay declared "The reversed world in question wouldn't merely strike us as queer, but definitely crazy: it would be a world where what is wildly and intrinsically *improbable* was always occurring. It would, in fact, be much more startling than the original asymmetry that led us to think of it."[4] (Findlay was almost certainly thinking of things like a tea cup, shattered due to a fall, spontaneously reassembling itself.)

The question of backward-running time so fascinated Findlay that, some years later, he posed it as a problem for the readership of a scholarly journal (*Analysis*). This led to a number of responses, and subsequently he presented both his own negative view of time-reversed worlds and that of the best reader response he had received to his posed problem (which came from McGechie).[5] While Findlay showed admirable open-mindedness by awarding the title of *best* to an argument that refuted his own position, he remained unconvinced about the concept of reversed-time worlds, stating that "I continue to feel that a total reversal of my experiences is a terrifying possibility."

The terror aspect of living backward in time had been nicely captured in a science fiction story years before Findlay wrote. A scientist who is involved in an accident with radioactive materials has his sense of time flow reversed, and the story carefully and logically analyzes what his life would be like in such a situation. For example, the scientist can talk (backwards for others, of course), so he is understandable only if his words are recorded and then played in reverse. He cannot eat, because for him that would involve the regurgitation of food. He cannot answer questions because "if he should answer any questions put to him, it would mean he was giving the answer before he heard the question, on his time scale." And finally, he can't pick anything up because the normally stable position and velocity error-correction mechanism between eye, hand, and brain, which is a negative feedback system in normal time, has become an unstable positive feedback system in reversed time. The horror of his existence is contained in the only words the man utters (deciphered after reversed playback): "Where am I? What's happened? Why are things so different? Why? Why?"[6]

---

[3] F. H. Bradley, *Appearance and Reality* (2nd edition), Oxford University Press 1897, p. 190.

[4] J. N. Findlay, *Philosophy* (25) 1950, pp. 346–347.

[5] J. N. Findlay and J. E. McGechie, "Does It Make Sense to Suppose That All Events, Including Personal Experiences, Could Occur in Reverse?" *Analysis*, June 1956, pp. 121–123.

Despite Findlay's "terror," however, the prevailing view today is that to the inhabitants of a time-reversed (or what is sometimes called a *counterclock*) world, *nothing would look odd*! This is a fairly new idea,[7] and not so long ago the philosophical literature displayed a misunderstanding of how a time-reversed world would appear to its occupants.[8] More recent analyses than Smart's (note 7) advocating the 'normality' of a time-reversed world are more compelling.[9] One concern about a time reversed world however is not easily dismissed: matter with a time reversed time sense is thought by many to be antimatter in our world, and any interactions between the two worlds would be spectacular, indeed![10]

But let's ignore that possible difficulty. The philosopher J. R. Lucas argued that even if beings from two such time-reversed worlds could meet, they still could not communicate: "If two beings are to regard each other as communicators, they must both have the same direction of time. It is a logical as well as a causal prerequisite."[11] Now, this matter is well worth some effort to understand, because it is intimately tied to time travel. At first blush, Lucas' words seem almost self-evident, and after a little thought they might seem to be absolutely irrefutable. The philosopher Murray MacBeath, however, took exception.

MacBeath opened his analysis[12] with a story to demonstrate that the persuasive power of Lucas' position is only superficial. In that story of Jim and Midge, Jim is one of us, whereas Midge is a 'Faustian time'[13] being. In his analyses MacBeath uses capitalized words and symbols for the time-reversed Midge, and lower case for Jim, as shown in Fig. 5.1. As MacBeath explains, "While Jim and Midge are together a face-to-face conversation is hardly likely to get off the ground. To make this clear let us say that they are together from t_0 until t_{10} on Jim's time-scale, and from T_0 until T_{10} on Midge's TIME-scale; t_0 is then the same temporal instant as T_{10} and, in general, $t_n = T_{10-n}$. If Jim at [his time] t_2 asks Midge a

---

[6] M. C. Pease," *Astounding Science Fiction*, "Reversion," December 1949. See also R. A. Banks, "This Side Up," *Galaxy Science Fiction*, July 1954, for a tale about the confusion caused by projecting a film the wrong way in time.

[7] The modern view that a time-reversed world would appear normal to someone living in it can be traced back at least as far as to J. J. C. Smart, "The Temporal Asymmetry of the World," *Analysis*, March 1954, pp. 79–83, an analysis, alas, that may not convince everyone.

[8] M. Dummett, "Bringing About the Past," *Philosophical Review*, July 1964, pp. 338–359.

[9] See, for example, D. L. Schumacher, "The Direction of Time and the Equivalence of 'Expanding' and 'Contracting' World-Models," *Proceedings of the Cambridge Philosophical Society* 1964, pp. 575–579; J. V. Narlikar, "The Direction of Time," *British Journal for the Philosophy of Science*, February 1965, pp. 281–285; F. R. Stannard, "Symmetry of the Time Axis," *Nature*, August 13, 1966, pp. 693–695.

[10] This issue is raised, several times, in Robert Silverberg's 1968 novel *The Masks of Time*.

[11] J. R. Lucas, *A Treatise on Time and Space*, Methuen 1973, pp. 43–47.

[12] M. MacBeath, "Communication and Time Reversal," *Synthese*, July 1983, pp. 27–46.

[13] In Goethe's play *Faust*, the normal flow of time is routinely upset.

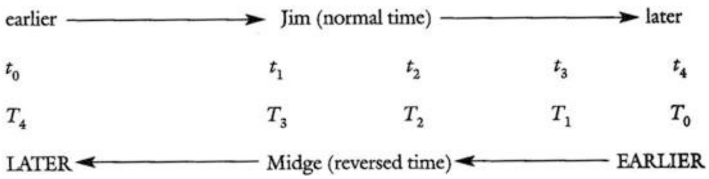

Fig. 5.1 Opposite time flows in counterclock worlds

question, and Midge hears the question at T_8, she will answer at T_9, and Jim will hear his question answered at t_1, *before* he asked it! What is more, if Jim is inexpert at interpreting backward sounds, and at t_4 asks Midge to repeat her answer, Midge will hear that request at T_6, BEFORE she has heard the original question; and her puzzled reply at T_7 will be heard by Jim at t_3 before he has uttered the request." Certainly this is a mess in time, and Lucas seems to be on safe ground with his denial of the possibility of communication between Jim and Midge. MacBeath, however, shows how to refute all of Lucas' arguments if Jim and Midge are allowed to be clever about how they send their messages back and forth—that is, if we give up some of our usual ideas of what a conversation is like. MacBeath's analyses are far too lengthy and detailed to present here, but the simplified diagram of Fig. 5.1 should enable one to follow the logic of his approach.

We imagine that Jim and Midge will not actually talk, and so will not have to decipher backward-spoken language. Rather, they will exchange messages via computer-generated text displayed on monitor screens, screens that are separated by a window that is proof against all penetration but the light emitted by those screens.[14] The nature of this window is not a trivial matter: if we accept the anti-matter nature of Midge's world then it is *essential* to keep her and Jim apart! To that end, MacBeath imagined that the window is double paned, with a perfect vacuum in-between. The exchange of photons between the two worlds should present no problems because photons are their own anti-particles.[15]

Now, imagine that at t_0 Jim brings a computer to the window. He programs it to wait for 4 days, until t_4, and then to display the following message on its screen: "This message is from Jim, who experiences time in the sense opposite to yours. Please study the following questions and display your answers on a computer screen three days from now." Jim's messages ends with the list of questions.

Because all that took place at t_4, Midge sees Jim's message and questions at what we will now call T_0. As requested, she brings her computer to the window, enters the answers to Jim's questions, and programs the machine to display them (after a

---

[14] Communication between beings in counterclock worlds, using written messages displayed through a window, appeared in science fiction *years* before MacBeath wrote: see I. Watson's 1978 novelette "The Very Slow Time Machine."

[15] There is no difference in the time sense of photons in either world because the flow of proper time for a photon—traveling at the speed of light, by definition—is zero (recall the discussion in Sect. 3.6).

3 day delay) on its screen. Thus, at T_3, which is Jim's t_1, Jim sees Midge's computer screen light up with "Hi, Jim. This is Midge. The answers to your questions are at the end of this message. Now, I've got some questions for *you*. Please display the answers two days from now." Midge's message ends with answers to Jim's questions and her list of questions.

Jim sees Midge's message at t_1, enters the answers to her questions and sets the machine to answer after a 2-day delay. At t_3, which is Midge's T_1—and by now you see how the process goes. It's cumbersome, sure, but it works. Or at least it does if everybody follows the rules. What if they don't? MacBeath provides other, increasingly complicated analyses to treat some of the more subtle problems that can be imagined in this method of exchanging messages. I will mention just two of them, which have direct analogs with what we normally think of as 'time travel.'

For the first problem, consider Jim's initial message, created at t_0 to be sent at t_4. He receives Midge's answer as described above at t_1, *before* his message is displayed through the window. So, what happens if at t_2 Jim cancels the message and it is *not* displayed? He has already gotten Midge's reply, but how can that happen if he does not send his message? This is, of course, a bilking paradox, with an explanation that we discussed in the previous chapter.

A second problem is the apparent possibility of creating a causal message loop. For example, let's say that at t_0 Jim suddenly decides to send a message through the window. (His reason for this sudden urge will be explained in the next few lines.) He thinks all night about what to send and, at t_1, finally settles on the following: "Greetings to the people on the other side of the window. This message comes from Jim, who hopes you will reply." Midge immediately sees Jim's message through the window (at her time T_3) and so is suddenly caught up with the desire to respond. She thinks all night about what to send and, hoping to be witty, she finally decides (at time T_4) on the following echo to Jim's message. "Greetings to the people on the other side of the window. This message comes from Midge, who hopes you will reply." Jim immediately sees this through the window (at what is his time t_0) and so now we know why he decides to send his original message! And Jim *will* send his original message—he has to because Midge replied to it.

In a review of a book on the direction of time, the philosopher Hilary Putnam restated the problems of a time-reversed world in the form of a provocative question: "How do you know that one man's future isn't another man's past?"[16] He began by making the interesting observation that for us to be able just to observe a backward-running universe, we would have to provide our own normal radiation source, because the counterclock stars in such a universe *absorb* radiation rather than emitting it. This point was elaborated on by another philosopher some years later, who wrote "We have uncritically imagined someone looking in on … two worlds having opposite time directions … Part of the story we tell, of the process of seeing, involves the emission of photons from objects [for example, the computer screens of Jim and Midge] and the subsequent impinging of these photons on our retinas. But this process is obviously directed in time. In a world where time ran opposite to ours, we could not see objects at all: objects would be photon-sinks, not photon-emitters."[17]

---

[16] H. Putnam, *The Journal of Philosophy*, April 1962, pp. 213–216.

Putnam concluded his comments about reversed time with a cautious warning: "It is difficult to talk about such extremely weird situations without deviating from ordinary idiomatic usage of English. But this difficulty should not be mistaken for a proof that these situations could not arise." That challenge is no doubt why so many writers of science fiction and fantasy have tackled the question of what it would be like if time ran backward. We can find such a tale long before the science fiction pulps, in fact, in a tale that appeared when Einstein was just 7 years old. In that story[18] the narrator (a professor of astronomy and higher mathematics) suddenly finds himself on Mars. There he encounters beings who know the future up to their deaths, and whose memories of the past are "scarcely more than a rudimentary faculty." The entire tale is in the form of a conversation between the professor and one such being, who argues (quite persuasively) for the virtues of his 'backward' existence compared to that of earthlings (the Martian name for Earth is the story's title). Bellamy realized that his story implies a fatalistic block universe: "No one could have foresight … without realizing that the future is as incapable of being changed as the past," he wrote.

Other writers, too, were fascinated by the implications of reverse time. When Merlyn the magician makes his first appearance in T. H. White's 1939 masterpiece *The Once and Future King*, for example, he explains how he knows the futures of others: "Ordinary people are born forward in Time, if you understand what I mean, and nearly everything in the world goes forward, too … But I unfortunately was born at the wrong end of time, and I have to live backwards from its front, while surrounded by a lot of people living forwards from behind. Some people call it having second sight."

What may have put the idea in White's mind for his time-reversed magician is only speculation today, but perhaps it was something he might have read a decade before, in a fellow Englishman's writing, the 1929 book *The Nature of the Physical World* by Sir Arthur Eddington, where one finds the following passage: "In "The Plattner Story" H. G. Wells relates how a man strayed into the fourth dimension and returned with left and right interchanged … In itself the change is so trivial that even Mr. Wells cannot weave a romance out of it [but see one of the *For Further Discussion* questions at the end of Chap. 2]. But if the man had come back with past and future interchanged, then indeed the situation would have been lively."

Whether or not those words influenced English fantasy, they certainly had some effect on American science fiction. In his 1979 memoir *The Way the Future Was*, pulp editor Frederik Pohl wrote that Eddington's book (which Pohl incorrectly attributed to Sir James Jeans) had given him the idea for a story using the reversed-time twist. But before Pohl could publish it, an even better (claimed Pohl) tale arrived from Malcolm Jameson. Jameson, too, had read Eddington's book, and the result was the novella-length "Quicksands of Youthwardness," which Pohl published in *Astonishing Stories* as three-part serial during 1940–1941. Unfortunately, Jameson's tale is both pretty awful *and* devoid of *any* connection with Eddington's suggestion. One can only wonder about what might have been in the story Pohl says *he* discarded—did *it* have the future remembered? Alas, Pohl wrote that he couldn't remember!

---

[17] N. Swartz, "Is There an Ozma-Problem for Time?" *Analysis*, January 1972, pp. 77–82.

[18] E. Bellamy, "The Blindman's World," *The Atlantic Monthly*, November 1886.

Remembering the future does occur in one tale where everybody knows what will happen (as they live backward) by reading "prediction books." What distinguishes that story from many others on the same theme is an interesting, ironic conversation a student in a reverse-time world has with a philosophy professor about how things would be if time went the 'other way,' as in *our* world:

> "How can we tell? The reverse sequence of causation may be just as valid as the one we are experiencing. Cause and effect are arbitrary, after all."
>
> "But it sounds pretty far-fetched."
>
> "It's hard for us to imagine, just because we're not used to it. It's only a matter of viewpoint. Water would run downhill and so on. Energy would flow the other way—from total concentration to total dispersion. Why not?"[19]

The student is unconvinced, however, and when he tries to visualize such a peculiar world (*our* world, don't forget!) it gives him a "half-pleasant shudder." Imagine, he thinks in wonder, never knowing the date of your own death.

A few years later, in Wilson Tucker's 1955 novel *Time Bomb*, we find the intriguing idea of political assassination by time bomb, with the bombs actually *time traveling* to their targets. A policeman begins to suspect what is happening when it becomes evident that one of the explosions was actually an implosion: "The time bomb ... had been going in and had carried the force of the blast with it. Inward. Into the past. He frowned at that. A backward explosion? An explosion which ran counter to the normal flow of time, to the normal method of living? ... How would an explosion appear to a man if the blast happened in the opposite manner? If it began exploding now, in this moment, but continued backward instead of forward? Would it be an implosion?"

There is little doubt that the definitive treatment of a reversed-time world is that of Philip K. Dick's 1967 novel *Counter-Clock World*. Dick's world was once our world, but then, as in Plato's tale that began this chapter, time suddenly begins to run backward. People still alive reverse their direction of aging (but still think, walk, and talk in forward time), and dead, buried people come alive again and emerge from graveyards as the "Sacrament of Miraculous Rebirth" is intoned by a priest; all live their way back to the womb, just as in Plato's tale of 1600 years earlier. Such imagery is powerful stuff, but the physicist John Wheeler (of black hole fame) would have none of it. As he wrote, "Most of us would probably agree that the universe has not contained and will not contain any backward-looking observers. We do not expect to see caskets with corpses in them coming to life, nor do we expect to find bank vaults in which a gram of radium will integrate rather than disintegrate."[20]

---

[19] D. Knight, "This Way to the Regress," *Galaxy Science Fiction*, August 1956.

5.2 Multi-dimensional Time

> "If there are extra time dimensions we get violations of causality, because one could sneak to yesterday through the extra dimensions, and . . . if you had sneaked to yesterday, you would have disappeared from today."[21]

There are those, however, who haven't been quite so sure as Wheeler about the impossibility of a reversed-time world. One physicist, for example, showed how (under certain initial conditions at the Big Bang) there is a possible solution to the gravitational field equations that gives an oscillating universe that temporally runs backward during the contraction phase.[22] And a philosopher has argued that the *direction* of time is local, not global (just as special relativity showed is the *rate* of time) and that the arrow of time can point in opposite directions at different locations.[23] As odd as such ideas may seem, a generalization of reversed-time—*multi-dimensional* time—makes it seem small potatoes. This is the idea that there might be *many* possible directions to the arrow of time, not just two. At first this may seem an absurd idea, something akin to a man jumping onto his horse and riding off in all directions at once. But philosophers (and perhaps just a few physicists[24]) have started to take at least a semi-serious look at the concept; as with so many other of the radical concepts associated with time travel, though, science fiction writers were dealing with multi-dimensional time long before it became a respectable topic in learned philosophical and physics journals.

---

[20] Wheeler's comments can be found in the General Discussion at the end of *The Nature of Time* (T. Gold, editor), Cornell University Press 1966.

[21] An intriguing (if somewhat mysterious) thought from F. J. Yndurain, "Disappearance of Matter Due to Causality and Probability Violations in Theories with Extra Timelike Dimensions," *Physics Letters B*, February 28, 1991, pp. 15–16.

[22] H. Schmidt, "Model of an Oscillating Cosmos Which Rejuvenates During Contraction," *Journal of Mathematical Physics*, March 1966, pp. 494–509. An elaboration of Schmidt's ideas is in A. Walstad, "Time's Arrow in an Oscillating Universe," *Foundations of Physics*, October 1980, pp. 743–749.

[23] G. Matthews, "Time's Arrow and the Structure of Spacetime," *Philosophy of Science*, March 1979, pp. 82–97.

[24] More than half-a-century ago one writer asserted that two-dimensional *complex* time was old hat in the theories of spinning particles—see M. Bunge, "On Multi-dimensional Time," *British Journal for the Philosophy of Science*, May 1958, p. 39. For a summary of many of the objections to multi-dimensional time see J. K. Kowalczynski, "Critical Comments on the Discussion About Tachyonic Causal Paradoxes and the Concept of Superluminal Reference Frames," *International Journal of Theoretical Physics*, January 1984, pp. 27–60 (and the reply by E. Recami, September 1987, pp. 913–919).

In one pulp story, for example, we find a professor asking his redundantly named class in speculative metaphysics "Why shouldn't time be a fifth, as well as a fourth, dimension?"[25] In response to a generally skeptical reception to that, the professor goes on to say "I believe in the existence of a two-dimensional time scheme . . . Ordinarily, most people think of time as a track they run on from their births to their deaths . . . Think of this time track we follow over the *surface of time* as a winding road [it is the imagery of a surface that gives the professor *two* time dimensions] . . . Once in a while another road crosses at right angles. Neither its past nor its future has any connection whatsoever with the world we know."[26]

The year before, the same pulp had published another tale[27] that went well beyond a mere two time dimensions. We are told in that story of two countries on an alien planet at war in the distant future. The war is a stalemate until one side begins to fire a gun at its foe from just two miles from its target, in the heart of enemy territory—*from the middle of next week*! The gun's shells are true 'time bombs.' This is not mere 'ordinary' time travel along one time track, however, but a multidimensional effect. Using a photograph of the gun in actual operation to support his astonishing discovery, an agent for the side being shelled reports to his superior that "the gun and its crew are existing along another time axis at right angles to the direction of our 'normal time,' so that from our point of view they are existing perpetually in the same instant."

That explains why the gun crew can (will?) operate without interference in next week's future, as they are in their adversary's time only for the instant that the two time tracks intersect. Indeed, the spy used the same trick to obtain his undetected photograph: "I secured the photograph by orienting myself along still another time axis at right angles to that of the gun, and approached it as an instantaneous, invisible entity." By the story's end both sides are using and counter-using this technique, evading each other "to and fro along an ever increasing complexity of mutually perpendicular time axes." In fact, the final count exceeds 75 time axes, making Heinlein's two-dimensional time look rather skimpy by comparison.

Well, of course, 75 time directions *is* science fiction (I think), and physicists are not so enamored of multidimensional time as are science fiction writers. For example, Eddington wrote that he found the idea of any region of spacetime involving two-dimensional time to "defy imagination."[28] Another physicist showed that the extremal property of timelike geodesics (look back at Fig. 3.15 and its discussion) would fail for multidimensional time, which he then associated with the stability of matter and a failure of causality.[29] Yet another physicist, however, was just a bit more willing to consider multidimensional time, and suggested that a viable theory of quantum gravity *might* support the idea of multiple time dimensions.[30] Of just *what* more than one time dimension might actually *mean*, however, this same physicist echoed Eddington by writing "Physics in a spacetime of . . . two timelike dimensions would be very weird indeed." Agreeing with this physicist was a philosopher who called the idea that there could be more than one dimension to time a "rather wild possibility" and a "fairy-tale."[31] These are probably fair statements of how most physicists presently think of multidimensional time.

---

[25] R. Heinlein, "Elsewhen," *Astounding Science Fiction*, September 1941.

[26] This story has an amusing scene in which one of the professor's students accidently 'jumps time tracks' and so enters a new track with his arrow of time pointing backwards.

[27] N. L. Knight, "Bombardment in Reverse," *Astounding Science Fiction*, February 1940.

[28] A. S. Eddington, *The Mathematical Theory of Relativity* (2nd edition), Cambridge University Press 1924, p. 25.

But not all philosophers are of that persuasion, and many are in fact as fascinated by the possibilities of multidimensional time as are science fiction writers. So, *why* this interest in something so different from anything we actually experience? Where does the motivation come from? Of what *use* is multidimensional time? I think the answers to those questions all derive from how multidimensional time offers a theoretical model for giving meaning to the view that the past can be changed (take a look back at note 20 in the Introduction).

In a certain trivial sense, of course, the past is always changing. For each of us the past is the set of all events that have happened, arranged in a before/after temporal order, and this set is continually increasing (and so changing). That is not, however, what most people mean by a changeable past. What *is* meant is that there may be some kind of change in the temporal ordering of events, or that an event that once was (or wasn't) a member of the set of past events no longer is (or is now) a member. Two-dimensional time offers a way to make sense of such possibilities, which one-dimensional time simply cannot do. To see how that works, I'll follow the presentation in a paper that forcefully argues that it does make sense to talk about altering the past.[32]

Meiland was aware that some might find his model *ad hoc*, even "incredibly weird" (in his own words), but he justified his efforts by taking a refreshingly enlightened, non-Humean view of what he thought would be the proper response to meeting purported time travelers: "If strange machines containing people in futuristic garments and speaking strange tongues (or perhaps using ESP instead of speech) were to appear and were to claim to be from the future, we might very well begin to search for a theory of time that allows their claim to be true." In Fig. 5.2 you can see how Meiland tried to do just that.

---

[29] J. Dorling, "The Dimensionality of Time," *American Journal of Physics*, April 1970, pp. 539–540.

[30] C. Isham, "Quantum Gravity," in *The New Physics* (P. Davies, editor), Cambridge University Press 1989.

[31] D. Zeilicovici, "Temporal Becoming Minus the Moving-Now," *Nous*, September 1989, pp. 505–524.

[32] J. W. Meiland, "A Two-Dimensional Passage Model of Time for Time Travel," *Philosophical Studies*, November 1974, pp. 153–173. Jack Meiland (1934–1998) was a professor of philosophy at the University of Michigan.

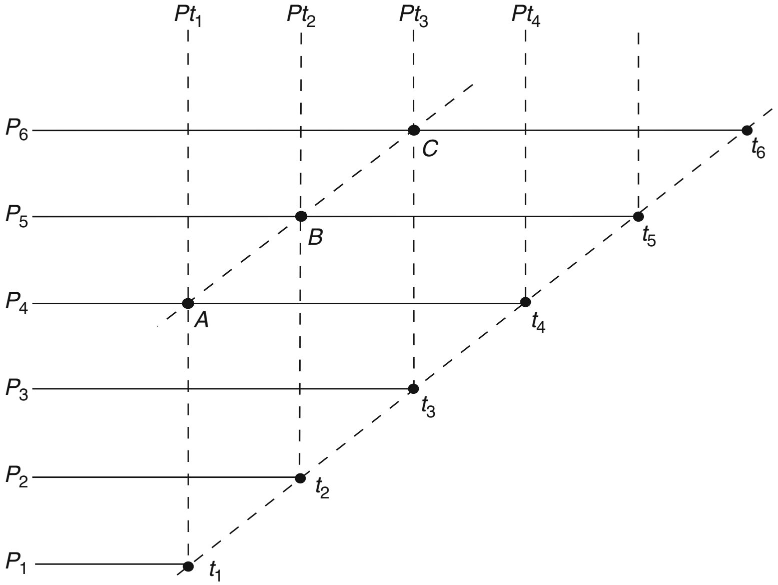

Fig. 5.2 Two-dimensional time

The dashed diagonal line, marked with the points t_1, t_2, ... represents our usual one-dimensional image of time. The horizontal lines P_1t_1, P_2t_2 ... (which we can simply call P_1, P_2, ..., for short) are the pasts for the present instants t_1, t_2, ... That is, P_1 is the past with respect to the present t_1, P_2 is the past with respect to the present t_2, and so on. The dashed vertical lines allow us to locate any moment in the past. For example, the intersection point A of P_4 with Pt_1 is the location of t_1 in the past with respect to the present t_4.

With this model, Meiland then analyzed in detail several interesting special cases. Suppose that t_1 and t_2 are 1 year apart and that there is a similar time separation between all adjacent, marked present moments on the diagonal. Let us further suppose that a time traveler at t_4 journeys backward 3 years to t_1, to arrive at point A. Assume he stays in the past 2 years; then his temporal locations lie along the dashed diagonal line ABC; that is, at B he is 3 years in the past of t_5, and at C he is 3 years in the past of t_6. From Fig. 5.2, then, we can imagine the time traveler saying, as he climbs into his time machine at t_4, "One year from now I'll be two years from now." That rather astonishing statement makes sense when we take both uses of now to be t_4 and observe that B (1 year from A) is 2 years in the past with respect to t_4.[33]

---

[33] A critic of time travel (see note 119 of Chap. 2) used what he claimed to be the absurdity of such a statement to support his ejection of time travel. One of Meiland's reasons for developing his two-dimensional model of time was, in fact, to be able to reply to that critic (Donald Williams).

Meiland's two-dimensional time model is undeniably fascinating, but it simply has no theoretical justification (as far as I know[34]). It is not necessary to assume two-dimensional time to explain Meiland's "strange machines containing people in futuristic garments"; it is possible to do that with one-dimensional time in a four-dimensional spacetime. The classic paradoxes, too, are understandable without two-dimensional time, as discussed in the previous chapter.

5.3 Maxwell's Equations and Sending Messages to the Past

> "Communication with a world exactly, to minutest detail, a duplicate of our own [but] twenty thousand years ahead of us might ruin the human race as effectively as if we had fallen into the Sun."[35]

Every physicist and electrical engineer knows that the mathematical description of the electromagnetic field is given by Maxwell's equations. In particular, radio engineers know that the waves of energy their antennas launch into space follow the predictions of those equations with astonishing accuracy. Indeed, when Einstein's relativity theory was completed, it was found that Maxwell's equations automatically satisfy relativity because magnetic effects *are* relativistic effects; in other words, relativity is built into Maxwell's equations. Whereas Newton's laws of dynamics had to be patched up, Maxwell's equations were untouched by the discovery of relativity.

Thus, it was a puzzle when physicists discovered that careful study of the seemingly perfect Maxwell equations, when applied to antennas, apparently results in the prediction of causality violation. It is found, in fact, that the equations have *two* solutions. One, as expected, contains the feature of time delay; that is, creating an electromagnetic disturbance at the antenna *now* causes a detectable effect at a distant point in space *later*. This is the so-called *time-retarded-solution*, and its common-sense physical interpretation is that of energy waves traveling away from the antenna as they also travel into the future. The shock was that Maxwell's equations also accept an *advanced solution*; energy waves *arriving* at the antenna from infinite space.

The physicist Paul Renno Heyl (1872–1961) wrote the perhaps first scientific work discussing advanced electromagnetic effects, in his 1889 University of Pennsylvania doctoral dissertation with the provocative title "The Theory of Light on the

---

[34] Many of the arguments against multi-dimensional time can be found in M. MacBeath, "Time's Square," in *The Philosophy of Time* (R. Le Poidevin and M. MacBeath, editors), Oxford University Press 1993. MacBeath concludes, however, with "I would not want to rule out the possibility . . . that time is three-dimensional. Or worse." See also Alasdair Richmond, "Plattner's Arrow: Science and Multi-dimensional time," *Ratio*, September 2000, pp. 256–274.

[35] The Victorian writer Samuel Butler (1835–1902), in the "Imaginary Worlds" entry of *The Notebooks of Samuel Butler* (published posthumously in 1912), commenting on the chaos that communication across time might cause.

Hypothesis of a Fourth Dimension." Heyl cited the scientific guru of the fourth dimension, C. H. Hinton (look back at the last "For Further Discussion" assignment in Chap. 2) as his inspiration. The situation described by Heyl is something like that of a child standing at the edge of a pond. She throws a rock into the middle of the pond and watches ripples spread out and away from the splash. Suddenly, she sees ripples appear all around the edge of the pond and then travel inward toward the center, where they all converge at once. A spout of water then erupts from the surface of the pond at the simultaneous meeting of the inward-traveling ripples, and she watches as a rock is ejected from the spout to land back in her hand. She is, of course, open-mouthed with astonishment! How absurd, you think, as you read this, and who could blame you? This amazing imagery of advanced effects we owe to the philosopher Karl Popper, and it has come to be called "the fable of the Popperian pond."[36]

Pursuing the mathematics of wave motion in the fourth dimension, Heyl wrote "We are led to the curious conclusion that, in Hinton's aether,[37] the nature of the central disturbance *after* a given instant can influence the form of the aether *before* that instant. In other words, the aether seems to be endowed with an uncanny faculty of foreknowledge." We can avoid such a counter-intutive implication of advanced effects, but only at the price of something many physicists and philosophers consider equally unacceptable: information traveling from the future into the past. We can still think of the advanced solution as representing electromagnetic waves of energy traveling *away* from the transmitting antenna—that is, as being *broadcast*, just like the retarded solution, rather than being received from infinity—*if* we also think of the waves traveling backward in time. Thus, the advanced solution to Maxwell's equations holds out the possibility of sending messages to the past, a sort of poor man's time travel. It may seem that we have simply traded one problem for another, however, because just sending information into the past can cause many of the same paradoxical, causality-busting situations that physical time traveling is claimed to cause.

The cosmologists Fred Hoyle (1915–2001) and J. V. Narlikar commented on the potential problems posed by communication backward in time, in their 1974 book *Action at a Distance in Physics and Cosmology*: "The [Maxwell] equations supply us with both advanced and retarded solutions (and, because of the linearity, with any linear combination of them) . . . With so many solutions theoretically possible, why does nature always select the retarded one? That this question cannot be answered within the framework of Maxwell's theory must be regarded as one of its intrinsic weaknesses."

---

[36] K. Popper, "The Arrow of Time," *Nature*, March 17, 1956, p. 538. See, too, note 110 (and its discussion) in Chap. 2.

[37] It was thought, in nineteenth century physics, that electromagnetic waves need a *medium* through which to propagate (like ocean waves need water, and sound waves need air), a mysterious substance called the *aether* (or *ether*) that exists even in a vacuum. The 1871 Michelson-Morley experiment, however, implied that the aether simply does *not* exist.

That branding of Maxwell's theory as having an "intrinsic weakness" because of its prediction of an advanced solution was, I think, unwarranted. Indeed, the advanced solution can be given a perfectly reasonable physical interpretation.[38] Imagine a transmitting antenna sending electromagnetic waves to an identical receiving antenna. At any point in space between the two antennas there are electric and magnetic fields. Maxwell's equations allow us to calculate the fields produced by alternating currents in the two antennas. When we do an analysis of the relationship between the transmitting antenna's current and the fields, we use the retarded solution because the current is the cause and the fields are the effect. But in the analysis of the relationship between the fields and the receiving antenna's current, the situation is reversed, and the fields are the cause and the current is the effect. That is, the advanced solution is simply the mathematics relating the current in the receiving antenna *now* to the fields *in the past*.

An acceptance of both solutions has, in fact, been in the physics literature for nearly a century. As the Yale physicist Leigh Page (1884–1952) wrote decades before Hoyle and Narlikar, "While the advanced potentials, as well as the retarded potentials, satisfy the electromagnetic equations, the former has generally been discarded for the reason that it has been more in accord with the trend of scientific intuition to consider that the present is determined by the past course of events than by the future. However, if it is once admitted that the present state is uniquely determined by any past state, it follows that the future is also so determined, and hence the employment of a future state as well as a past state in specifying the present marks no inherent departure from our accustomed methods of description . . ."[39]

It may still be tempting, however, to just dismiss the advanced solution as a mere anomaly of the mathematics and to discard it on physical grounds. This is the traditional approach taken by physicists when confronted with non-causal solutions in any physical theory and, indeed, that was what Swiss physicist Walter Ritz (1878–1909) did with the advanced solutions to Maxwell's equations (an approach that involved Ritz during the last year of his life in a dispute with Einstein). For Ritz, the reversal of cause and effect simply did too much violence to his intuition to be taken seriously, and so he thought one must *impose* causality on Maxwell's equations (a condition they do not inherently contain) by *a priori* rejecting the advanced solution.[40]

Still, electrical engineers make similar kinds of judgements all the time, as when the solution of a quadratic equation for a passive (energy-dissipating) resistor gives both a positive value (which 'makes sense') and a negative value (which doesn't 'make sense' and so is simply ignored). There *were*, however, those who encouraged caution on this issue. Seventy-five years ago, for example, the eminent MIT electrical engineer Julius Adams Stratton (1901–1994) echoed Leigh Page's warning when he wrote of the disturbing advanced solution, "The familiar chain of cause-and-effect is thus reversed and this alternative solution might be discarded as logically inconceivable. However, the application of 'logical' causality principles offers very insecure footing in matters such as these and we shall do better to restrict the [Maxwell] theory to retarded action solely on the grounds that this solution alone conforms to the *present* [my emphasis] data."[41]

---

[38] See, for example, S. L. Schwebel, "Advanced and Retarded Solutions in Field Theory," *International Journal of Theoretical Physics*, October 1970, pp. 347–353, and L. M. Stephenson, "Clarification of an Apparent Asymmetry in Electromagnetic Theory," *Foundations of Physics*, December 1978, pp. 921–926.

[39] L. Page, "Advanced Potentials and Their Application to Atomic Models," *Physical Review*, September 1924, pp. 296–305.

[40] For more on this, see O. Costa de Beauregard, "No Paradox in the Theory of Time Anisotropy," *Stadium Generale* 1971, pp. 10–18.

And in a famous paper I'll discuss in the next section, Wheeler and Feynman declared that "We conclude advanced and retarded interactions give a description of nature logically as acceptable and physically as completely deterministic as the Newtonian scheme of mechanics. In both forms of dynamics the distinction between cause and effect is pointless. With deterministic equations to describe the event, one can say: the stone hits the ground because it was dropped from a height; equally well, the stone fell from a height because it was going to hit the ground."[42] For Wheeler and Feynman, the reversal of cause and effect inherent to backward causation and time travel to the past offered no conceptual difficulties.

The elimination of an appeal to causality, or to the 'weirdness' of advanced waves, in arguing for the naturalness of the retarded solution to Maxwell's equations was first done in 1976.[43] The only auxiliary condition applied to the equations was simply the natural one of requiring the initial field energy to be finite.[44] And yet, today, there is still no experimental evidence for the physical reality of the advanced solution. Now and then one does run across speculations that the advanced waves of *something* traveling backward to us from the future might explain the so-called ESP 'talent' of precognition, but that is all it is, speculation.[45]

Advanced waves appeared in the pulp science fiction of the late 1930s, a decade before Wheeler and Feynman. For example, one story actually specifically invoked the advanced solution to Maxwell's equations, with a gadget (making use of what the author called the "anticipated potentials") displaying the near future on a television-like screen.[46] The author's by-line proudly gave his academic credentials as including a master's degree and, in fact, John Pierce was a graduate student in electrical engineering at Caltech. He received his doctorate just months after this story was published, and then went on to a highly distinguished career at Bell Telephone Laboratories and then later at the Jet Propulsion Laboratory operated by Caltech for NASA. Pierce knew all about Maxwell's equations, of course, and he actually opened his tale with a quote from Page's 1924 article (note 39) on the advanced solution. How many fictional pieces include quotes from the *Physical Review*?

---

[41] J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill 1941, p. 428.

[42] J. A. Wheeler and R. P. Feynman, "Classical Electrodynamics in Terms of Direct Interparticle Action," *Reviews of Modern Physics*, July 1949, pp. 425–433.

[43] P. C. Aichelburg and R. Beig, "Radiation Damping As An Initial Value Problem," *Annals of Physics*, May 1976, pp. 264–283.

[44] J. L. Anderson, "Why We Use Retarded Potentials," *American Journal of Physics*, May 1992, pp. 465–467.

[45] See M. B. Hesse, *Forces and Fields*, Philosophical Library 1961.

[46] J. R. Pierce, "Pre-Vision," *Astounding Stories*, March 1936.

And it would probably take something like advanced waves to explain the funny doings 2 years late in a story of a man caught in a time machine accident. Nearly all of his body ends up 4 years in the future—but only *nearly* all, because his eyes remain the present! As one of the puzzled observers of this odd business wonders, "Strange, that his eyes, now, can convey a message to his brain, four years hence, and his brain tells the eye muscles to move the eyeballs which are four years behind them —."[47]

Some of the most intriguing paradoxes of time travel involve no traveler—only information. Of course, any information flow at all, independent of time travel, involves the flow of energy and, as Einstein showed, energy and mass are different aspects of the same thing. Accordingly, information time travel involves the transfer of mass/energy. Thus, a man in the twenty-fifth century who sends a backward-in-time 'temporal radio' message to a twentieth-century woman stating that he loves (will love?) her is sending much more than mere emotion. Just *how* to send a message backward in time is, of course, the puzzle.

Indeed, all forms of present-day communication are transmissions only to the future. If you speak to someone, or if you send a radio message, there are always delays depending on the distance of separation and the speed of transmission of sound and light, respectively. If you want to send a message to the one hundred and twenty-fifth century, you can; just write a letter and seal it in a pressurized bottle of helium. This basic idea is dramatically presented in a novel about a scientist who is accidently transported from 2162 back to the late-Cretaceous, 80 million years into the past. There he leaves a written record on seven sandstone slabs of his brutal, lonely life among the dinosaurs—letters across time, if you will—found by twenty-second century geologists some years after his disappearance.[48] The transmission of such letters, *forward* in time, while dramatic, is not a puzzle. But what could be more astonishing than the message received *from the future* by the young inventor of the first time machine: after his initial experiment of sending his pilotless machine into the future, it returns with an envelope inside. Eagerly tearing it open, he finds the note is from the National Academy of Sciences: "We know from old records and museum models that this is the Cullen Foster experimental machine. Fifty years looks down on you and says 'Good work'."[49]

Heady stuff, that, but lots of other possible messages are capable of competing with young Foster's when it comes to generating excitement. For example, suppose you had a gadget that is superficially similar to a telephone but that calls telephones in the *distant* future. You can hear the person (in the future) on the other end, but they can't hear you (in their past). That is, information can flow only from future to past. It is then easy to imagine situations in their use of this device that at least seem paradoxical. For example, suppose you call your own private number 1 month ahead. You hear your future-self first answer the phone, and then recite the winning lottery for the 'previous' day, which is a month in your present self's future. (Your future self does this somewhat odd recital because a month from now you will remember, when your private phone rings, just *who* is calling!) So, now in the present you know you'll make a fortune by winning the lottery a month later.

---

[47] M. Schere, "Anachronistic Optics," *Astounding Stories*, February 1938.

[48] G. G. Simpson, *The Dechronization of Sam Magruder*, St. Martin's Press 1996.

[49] D. Stapleton, "How Much to Thursday?" *Thrilling Wonder Stories*, December 1942.

This example *is* admittedly somewhat mysterious since, for the gadget to call far ahead in time, some sort of signal (as yet unspecified) must travel into the distant future because *something will* make the future phone ring. For the present discussion I am ignoring this crucial issue for the sake of the dramatic impact of the example. Soon, however, we'll get a little way into describing how one might, in principle, actually build this gadget, which in the physics literature is called an *antitelephone*. (Such a device is an *anti*telephone because the person who is the receiver is in the sender's past, the opposite of the situation for an ordinary telephone.) An interesting fictional illustration of such a gadget, despite being told as a hard-boiled detective murder mystery, appeared in science fiction some years ago (alas, while called a "time telephone," no theory for its operation was given, but instead was 'explained' as a "straightforward application of an impressive, but limited, technology").[50] So far, there is nothing paradoxical (or even illegal) in all of this, but what if, when the phone rings in the future the day after you won the lottery, you perversely decide *not* to recite the winning number? This apparent paradox has, in fact, already been treated with the aid of the block universe view of spacetime; that is, if the future-you spoke the lottery number when you originally called, then the present—you (now in the future) *must*, inevitably recite it.[51]

Let's now make things a bit more involved. Suppose that instead of calling your future self, you call the weather service and listen to the recorded message telling you the weather, 30 days hence. You do this day after day, and after a while you get a reputation for being able to predict, *perfectly*, the weather for every day to come, up to a month into the future. Your reputation spreads far and wide, and after a while more the weather service hears about you. Meteorologists check and find you are *never* wrong. Their computer models are only 80 % accurate out to 3 days, and for a week's prediction and beyond, the general public might as well flip a coin on whether it will rain or not on any particular day. But you are 100 % correct out to ten times their range. And so they hire you—and as a secondary job, you also make the daily weather recordings. (The voice on the other end of the gadget *has* sounded sort of familiar!) Here, then, is the puzzle we encountered earlier in causal loops carrying information: from *where* is the information in the flawless weather predictions coming from?

---

[50] S. Schmidt, "Worthsayer," in *More Whatdunits* (M. Resnik, editor), DAW 1993. The author, Stanley Schmidt, has a Ph.D. in physics and is a former editor of *Analog Science Fiction* magazine.

[51] For a fictional illustration of this (a so-called *bilking paradox*), see W. Tevis, "The Other End of the Line," *Magazine of Fantasy and Science Fiction*, November 1961.

One easy answer is that the question is meaningless because such a future-to-the-past information flow must be impossible. Indeed, if I am to avoid telling a 'philosopher's fairy tale,' like those I criticized earlier in the book, I must admit that one consistent, non-paradoxical answer is found in recognizing that I have *assumed* that those 30-day weather reports are correct. Maybe, however, they are no better than anybody else's predictions. And so you *don't* become famous, and you *don't* get hired—and so there *isn't* any paradox. Is that the way to avoid paradoxes involving information flowing backward in time?

Perhaps not. As long as 1917 it was realized that special relativity does not preclude such an apparent backward flow. That is, if information could be transmitted faster than light, *then* messages could travel backward in time. That was the year Richard Tolman (1881–1948), a professor of physical chemistry at the University of Illinois and later at Caltech, wrote "The question naturally arises whether velocities which are greater than that of light could ever possibly be obtained."[52] He then answered that question, with his general conclusion being that if such velocities are possible, then a faster-than-light (FTL) observer could see the time order of two causally related events reverse. And thus the observer would see an affect *before* its cause. Alternatively, a *sub*luminal (slower-than-light) observer could see the two events, which are connected via an FTL interaction, reversed in time order from what a stationary observer would see.

Either situation has come to be called Tolman's paradox, but Tolman himself was careful with his words: "Such a condition of affairs might not be a logical impossibility; nevertheless its extraordinary nature might incline us to believe that no causal impulse can travel with a velocity greater than that of light." That was an astonishing statement, given that Einstein himself had specifically stated in his original 1905 paper on special relativity that such a thing simply could not occur. There is nothing, it would seem, to be "inclined" about.[53]

This rather technical connection between FTL speeds and backward time travel made the transition from theoretical physics to popular culture very quickly. It was in the British humor weekly *Punch*, for example, that the famous (but nearly always misquoted) limerick by A. H. R. Buller (1874–1944) first appeared:

There was a young lady named Bright
Whose speed was far faster than light;
She set out one day

---

[52] R. C. Tolman, *The Theory of the Relativity of Motion*, University of California Press 1917.

[53] Take a look back at Sect. 3.5, where we showed that the time order of two events can appear reversed for a subluminal observer if the two events are not causally related. Introducing FTL motion results in extending reversal to causally connected events; that is, FTL motion, reversed causation, and time travel to the past, go hand-in-hand-in-hand.

In a relative way
And returned on the previous night.[54]

Where *Punch* dared to go, Hollywood could not be far behind. Indeed, in this case it was actually there first, with the 1922 one-reel silent comedy movie *The Sky Splitter*. This was just a short film (feature pictures generally had at least four reels), so it is not clear how widely distributed and viewed it may have been. The story is that of a scientist testing a new spaceship: when it exceeds the speed of light, he begins to relive his life.

The linkage between time travel to the past and FTL motion is a central one in science fiction, and its fascination was nicely illustrated by one writer who has a time machine experimenter in the twenty-seventh century wonder "Was the speed of light the core of the mystery? At the speed of light did the past and the future become a shining, merging road down which men could walk—in their ears the thunder of time passing . . .?"[55] Not everybody was excited with the idea of FTL motion and travel backwards in time, however, with one eminent scientist declaring that "the limit to the velocity of signals is our bulwark against the topsy-turvydom of past and future."[56]

The obvious question at this point, of course, is whether it is even conceptually possible to build a gadget to send FTL messages backward in time? Einstein himself thought not, saying "We cannot send wire messages into the past."[57] But was he right? One hint at the possibility of achieving FTL speeds is in Dirac's 1938 paper (note 52 in Chap. 2). There, in his remarks about pre-acceleration, Dirac wrote "Suppose we have a pulse sent out from place A and a receiving apparatus for electromagnetic waves at a place B, and suppose there is an electron on the straight line joining A to B. Then the electron will be radiating appreciably [because accelerated charges radiate] before the pulse has reached its centre and this emitted radiation will be detectable at B at a time . . . earlier than when the pulse, which travels from A to B with the velocity of light, arrives. *In this way a signal can be sent from A to B faster than light* [my emphasis]."

This exciting conclusion goes a step beyond the usual examples of 'things that go faster than light.'[58] Dirac had an equally exciting reaction (and here the emphasis is his): "This is a fundamental departure from the ordinary ideas of relativity and is to be interpreted by saying that *it is possible for a signal to be transmitted faster than light through the interior of an electron. The finite size of the electron now reappears in a new sense, the interior of the electron being a region of*

---

[54] On page 591 of the issue of December 19, 1923.

[55] F. B. Long, "Throwback in Time," *Science Fiction Plus*, April 1953.

[56] A. S. Eddington, *The Nature of the Physical World*, Macmillan 1929.

[57] A. Einstein, "La Théorie de la Relativité," *Bulletin de la Société Francaise de Philosophie* 1922, pp. 91–113.

[58] Such as, for example, the intersection *point* of two very long, closing scissor blades. The explanation for how this can be is that the *point* is massless and does not participate in a causal chain (and so carries no information). Thus, special relativity is not violated.

failure, not of the field equations of electromagnetic theory, but of some elementary properties of space-time." This last line sounds very much like the things people say today about the singularity inside a black hole event horizon. And yet, Dirac was careful to point out that as weird as FTL speed may appear, special relativity is not violated because "in spite of this departure from ordinary relativistic ideas, our whole theory is Lorentz invariant." That is, even though 'faster-than-light' means 'backward in time,' which means 'causality failure,' special relativity still holds true and nothing awful happens to physics, only to our intuitions. The reason for this is that causality is *not* a premise or starting point for the special relativity.[59]

Of course, like any scientific theory, Dirac's theory is not necessarily the last word, and we have to admit the possibility that at least some of its implications (in particular, the possibility of FTL speeds) just aren't so. In all electronic communication systems that we use, information is transmitted by *modulating* a so-called *carrier wave*, and there is some reason to believe that such modulated waves cannot be sent at FTL speeds.[60] We must admit that it is one thing to talk of 'advanced wave radios'—often called *Dirac radios* in science fiction—and quite another to see how physics might actually enable one to talk to the past. FTL communication (without the time travel aspect) appeared in pulp science fiction before 1940, as in one story published the year after Dirac's paper.[61] In it we learn of a man on Pluto who has invented a way to send messages to Earth at twice the speed of light. He uses this gadget to warn of a would-be dictator who is on his way to Earth in a 'mere' light-speed rocket ship, and only an FTL message can warn Earth in time.

5.4 Wheeler and Feynman and Their Bilking Paradox

> "We find it difficult if not impossible to imagine waves that go into the future and *return to the present* [my emphasis] bearing information about where (and when) they have been."[62]

---

[59] See, for example, G. Nerlich, "Special Relativity Is Not Based On Causality," *British Journal for the Philosophy of Science*, December 1982, pp. 361–388. This same point was made nearly two decades earlier, in a study of the possibility of superluminal sound in superdense matter, by D. A Kirzhnitz and V. L. Polyachenko, "On the Possibility of Macroscopic Manifestations of Violation of Microscopic Causality," *Soviet Physics JETP*, August 1964, pp. 514–519.

[60] See G. Diener, "Superluminal Group Velocities and Information Transfer," *Physics Letters A*, December 16, 1996, pp. 327–331. For more on the modulation of a light-speed carrier wave in everyday AM radio, and in a more sophisticated single-sideband transmitter, see my book *The Science of Radio*, Springer 1999.

[61] N. Bond, "Lightship, Ho!," *Astounding Science Fiction*, July 1939. The author provides an interesting, detailed description of the gadget, and I think it would make a good question on a Ph. D. qualifying exam in physics or electrical engineering to explain the flaw in it.

[62] Bob Brier, *Precognition and the Philosophy of Science: An Essay on Backward Causation*, Humanities Press 1974. Brier is an Egyptologist (!) at Long Island University—with a Ph.D. in philosophy—who specialized at one time in parapsychology.

In 1941, at a meeting of the American Physical Society, Princeton University physicist John Wheeler and his student Richard Feynman discussed a seemingly outrageous idea that provided a possible clue to how a Dirac radio might function. The idea was that the advanced wave solutions to Maxwell's equations are not mere mathematical curiosities, but rather have profound physical significance. At the time, their talk received only a small abstract notice in the *Physical Review*, but after World War II they wrote it all up in a beautiful paper.[63]

Their primary goal was to explain the origin of the force of radiative reaction discussed by Lorentz earlier in the century. This reaction force is the cause of the energy loss suffered by an accelerated, charged particle. Lorentz, who thought of charged particles as having a finite size, attributed this reaction force to the retarded (by the time required for light to cross the width of the charged particle) coulomb repulsion force between one side of the particle's charge to the charge on the opposite side. This view, however, leads to various conceptual and mathematical problems, including an arbitrary assumption on how the charge is distributed over/through the finite volume of the particle, as well as the problems of infinite self-interactions and the issue of what keeps the charge from blowing itself apart by internal coulomb repulsion.

Wheeler and Feynman's theory, on the other hand, avoided those problems by postulating point charges, because a *point* charge cannot repel itself. But then whence the reaction force, if there is no repulsion? Their revolutionary explanation was first to imagine the accelerated point charge as emitting retarded radiation outward in space, eventually to be absorbed by distant matter. This distant matter, which itself consists of point charges that are accelerated by the retarded radiation, then radiates *backward* in time, back toward the original charge that started the chain of events. This backward-in-time, or *advanced*, radiation arrives in the past of the original charge, and *it* is the cause of the observed reaction force. Indeed, Wheeler and Feynman proposed that an accelerated charge will not radiate unless there is to be absorption at some other distant place and future time. That is, the *future* behavior of a distant absorber determines the *past* event of radiation; there is simply no such thing as just radiating into empty space. The entire universe, spatially *and* temporally, is a very 'connected' place!

Astonishingly, this non-causal view of spacetime had been around in physics for at least 20 years before Wheeler and Feynman's talk. They had independently developed their ideas but, after their 1941 talk, Einstein (who perhaps recalled his 1909 debate about advanced effects with Ritz) brought a 1922 paper by the Dutch physicist Hugo Tetrode (1895–1931) to their attention. In his paper Tetrode had written that "the Sun would not radiate if it were alone in space and no other bodies could absorb its radiation ... If for example I observe through my telescope yesterday evening that star which let us say is 100 light years away, then not only did I know that the light which it allowed to reach my eye was emitted 100 years ago, but also the star or individual atoms of it knew already 100 years ago that I, who then did not even exist, would view it yesterday evening at such and such a time."

---

[63] J. A. Wheeler and R. P. Feynman, "Interaction with the Absorber as the Mechanism of Radiation," *Reviews of Modern Physics*, April–July 1945, pp. 157–181.

Tetrode's vivid imagery had been, curiously, itself captured even decades earlier in words from the nineteenth century English poet Francis Thompson (1859–1907), in his "The Mistress of Vision":

All things … near and far,
Hiddenly to each other linked are,
That thou canst not stir a flower
Without troubling of a star.

None of this, of course, is obvious! As a tutorial paper appearing just 2 years after Wheeler and Feynman's 1945 paper expressed it, "Any physical theory which seriously proposes that events in the future may be the efficient cause of events in the past may be regarded—at least at first glance—as rather revolutionary doctrine."[64] Indeed!

It is interesting to note that Einstein apparently said nothing to Wheeler and Feynman about a paper that pre-dated Tetrode's by 3 years. In 1919 the Finnish physicist Gunnar Nordström (1881–1923) had suggested that the advanced solution might offer an explanation for a perplexing problem in atomic theory. Maxwell's theory says that an accelerated electric charge radiates energy, which implies that the orbital electrons in the classical model of the atom should quickly spiral in toward the nucleus, that is, all matter should collapse. This cataclysmic event (of course!) has not happened, and Nordström's idea was that if one took into account not only the usual retarded solution but the advanced one as well, then perhaps things could be understood. Indeed, Nordström was able to show that such an analysis does give zero for the *average* energy radiated by an orbiting electron. Later, however, Page (note 39) showed that the instantaneous radiated energy is not zero, and that this would lead to observable effects that in fact are *not* observed.

Now, to be sure that the 'doctrine' discussed in note 64 is clear, let me restate what Tetrode, and later Wheeler and Feynman, had in mind. Imagine we have an electric charge (the *source*) that we mechanically shake, that is, accelerate. This allows us to assign a definite *cause* to the charge's acceleration which, of course, radiates energy. This radiation travels outward into space as observed retarded fields until they are eventually absorbed by distant matter. The charges in that distant matter are thus accelerated, and they in turn therefore radiate energy. This induced radiation again consists, according to Wheeler and Feynman, of both retarded and advanced fields. The advanced fields radiate outward but *backward* in time toward the original charge, collapsing upon it at the precise instant we first shook it, thereby producing the radiative reaction force. At any instant of time, at any point in space, the observed field is the sum of the retarded field traveling away from the source into the future and the advanced field traveling toward the source in the past.

---

[64] C. W. Berenda, "The Determination of Past by Future Events: A Discussion of the Wheeler-Feynman Absorption-Radiation Theory," *Philosophy of Science*, 1947, pp. 13–19.

But, argued Wheeler and Feynman, there is one last point that has been left out of this picture—there is also an advanced field (traveling away from the source and backward in time) because of the original, mechanical shaking of the source charge. Equivalently, a field traveling *forward* in time will *converge* onto the source because we *will* shake it. Wheeler and Feynman showed that before the mechanical shaking that starts this whole process, the advanced radiation field of the source and the advanced radiation fields of the absorbers exactly cancel each other at every point in space and every instant of time (if there is total absorption in the future), which accounts for the experimental fact that we observe a zero total field before the mechanical shaking occurs.

Wheeler and Feynman showed that if we accept these (strange) ideas, then everything we actually observe is predictable: radiative reaction, the direction of the electromagnetic arrow of time from past to future (retarded-only effects), and the absence of infinite self-interactions. The claim by Wheeler and Feynman to have avoided self-interaction problems via the use of the advanced solution was, however, soon challenged. Indeed, the self-interaction of the electron is *needed* to explain the 1947 experiment by Willis Lamb (1913–2008) that measured the deviation (the *Lamb shift*) of the spectrum of hydrogen from what Dirac's theory of the electron predicts. Ironically, it was that experiment that helped motivate the renormalization of quantum electrodynamics (to get rid of the infinities then plaguing it) which led to Feynman's share of the 1965 Nobel Prize. In fact, just 4 years after their 1945 paper, Feynman expressed a revised view that self-interactions could not be avoided.[65]

In any case, we gain the rewards originally claimed by Wheeler and Feynman *only if* we accept backward time travel, a step too big for many in 1945 (and for nearly as many today) because of the resulting time travel paradoxes that seem to be unavoidable. For the same reason, Tetrode's earlier work, published in a German journal, also went virtually unnoticed during the two decades before Wheeler and Feynman's work. In fact, Tetrode wasn't the only anticipator of Wheeler and Feynman, as they had been anticipated, too, in America. In 1926 the chemist G. N. Lewis (1875–1946) had written "I'm going to make the … assumption that an atom never emits light except to another atom, and to claim that it is absurd to think of light emitted by one atom regardless of the existence of a receiving atom as it would be to think of an atom absorbing light without the existence of light to be absorbed."[66] Wheeler and Feynman were aware of Lewis by 1945. Certainly Wheeler and Feynman must have been intrigued by Lewis' paradox: "I shall not attempt to conceal the conflict between these views and common sense. The light coming from a distant star is absorbed, let us say, by a molecule of chlorophyll which has recently been produced in a living plant. We say that the light from the star was on its way toward us a thousand years ago. What rapport can there be between the emitting source and this newly made molecule of chlorophyll?"

---

[65] R. P. Feynman, "Space-Time Approach to Quantum Electrodynamics," *Physical Review*, September 15, 1949, pp. 769–789. See also C. Teitelboim, "Splitting the Maxwell Tensor: Radiation Reaction Without Advanced Fields," *Physical Review D*, March 15, 1970, pp. 1572–1582.

[66] G. N. Lewis, "The Nature of Light," *Proceedings of the National Academy of Sciences*, January 15, 1926, pp. 22–29.

The paradox in that, of course, arises from the issue of what happens if, at some intermediate time and place, the star's light is blocked, thus preventing its absorption by the chlorophyll? Could refusing to look at a star *now* affect the emission of the star's light in the *past*? Lewis was obviously making a clear statement of backward causation when posing this bilking paradox. His very next words show that he understood the probable reaction of his readers: "Such an idea is repugnant to all our notions of causality and temporal sequence." Like Tetrode's work, Lewis' ideas were ahead of the times but, actually, their ideas were *not* repugnant to everyone.

In fact, similar puzzles were an inspiration to Wheeler and Feynman and almost certainly motivated them to create their own famous bilking paradox, which they presented in their 1949 paper (note 42). They opened the presentation of their paradox as follows: "If the present motion of *a* is affected by the future motion of *b*, then the observation of *a* attributes a certain inevitability to the motion of *b*. Is not this conclusion in direct conflict with our recognized ability to influence the future motion of *b*?" This question clearly states the conflict between free will and determinism, and to sidestep this *human* concern Wheeler and Feynman constructed a "paradox machine," a machine that operates totally automatically and which has come to be called the "logically pernicious self-inhibitor"![67]

In their description of the paradox machine, Wheeler and Feynman ask us to imagine two charged particles, *a* and *b*, positioned five light-hours apart. As shown in Fig. 5.3, *a* is attached to the arm of a pivoted shutter, toward which a pellet is moving from initially a great distance away. Now, normally we would think of what happens next in terms of just retarded fields. That is, the pellet hits the arm, knocking it downward and thereby accelerating charge *a*; this acceleration of charge *a* creates a retarded radiation field that arrives at charge *b* 5 h later, resulting in the acceleration of charge *b*; the acceleration of charge *b* creates a retarded radiation field that arrives back at charge *a* 5 h later (10 h after the pellet hit the arm).

The Wheeler and Feynman view, however, claims that this description leaves out half the story—the *advanced* fields. Specifically, suppose the pellet will hit the arm and so accelerate *a* at 6 p.m. Then, *b* will be affected not only 5 h later at 11 p.m., but also *earlier* at 1 p.m. This advance acceleration of *b*, in turn, sends out an advanced field that arrives at *a* at 8 a.m. The paradox is now easy to see. As Wheeler

---

[67] P. Fitzgerald, "Tachyons, Backwards Causation, and Freedom," *Boston Studies in the Philosophy of Science* (volume 8), 1970, pp. 415–436. Even more extreme examples of such paradox machines are described in Tim Maudlin, "Time Travel and Topology," *PSA 1990*, Philosophy of Science Association, volume 1, pp. 303–315.

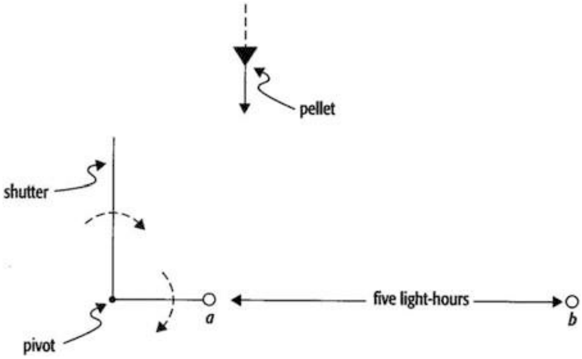

Fig. 5.3 Wheeler and Feynman's paradox machine

and Feynman described events, we see *a* exhibit a premonitory movement at 8 a.m. Seeing this motion in the morning, we conclude that the pellet will hit the arm in the evening. We could then return to the scene a few seconds before 6 p.m. and block the pellet from acting on *a*, a task automatically accomplished by the shutter in Wheeler and Feynman's paradox machine. But then we are faced with the puzzle of explaining just *why a* moved in the morning!

Wheeler and Feynman claimed they had resolved their bilking paradox by observing that *discontinuous* forces (more generally, *signals*) are never seen in nature. They concluded that the shutter does not completely block the pellet, but rather the shutter suffers a "glancing blow." That is, a very weak advanced signal is received by charge *a*, which moves the shutter just enough to induce the "glancing blow," and it is this partial interaction that results in the weakened signal in the "first place." This is, in fact, the very same explanation that was rediscovered decades later in answer to similar bilking paradoxes that involve self-interacting billiard balls transiting wormhole time machines (a topic we'll take-up in the next chapter).

5.5 Absorber Theory and Signaling the Past

> "If advanced waves [could be used to signal the past] then our grip on reality would become more tenuous. The past could never be considered over and done with, because anyone with the proper hardware could send messages back in time and alter what had already happened."
>
> —John Cramer, a University of Washington physicist, taking the minority position on the possibility of changing the past[68]

---

[68] Cramer has written provocatively on advanced waves. See, for example, "The Arrow of Electromagnetic Time and the Generalized Absorber Theory," *Foundations of Physics*, September 1983, pp. 887–902, and "Generalized Absorber Theory and the Einstein-Podolsky-Rosen Paradox," *Physical Review D*, July 15, 1980, pp. 362–376.

Wheeler and Feynman's argument *is* logically and physically sensible; it is, after all, simply an early statement of the principle of self-consistency in physics. Wheeler summed matters up nicely, years later, when he wrote "Interconnections run forward and backward in time in such numbers as to make an unbelievable maze. That weaving together of past and future seems to contradict every normal idea of causality. However, when the number of particles is great enough to absorb completely the signal starting out from any source, then this myriad of couplings adds up to a simple result: the familiar retarded actions of everyday experience, plus the familiar force of radiative reaction with its familiar sign."[69]

Their analysis was based on classical physics but, many years later, Feynman wrote (in his famous autobiographical work *Surely You're Joking, Mr. Feynman*) that at one time he and Wheeler thought it would not be too difficult to work out the quantum version of their theory. But then first Wheeler failed in the task, and then Feynman tried his hand at it and, as he stated, "I never solved it, either—a quantum theory of half-advanced, half-retarded potentials—and I worked on it for years." Their paradox (if indeed it *is* a paradox, since if advanced fields don't actually exist then there is no problem) remains unsolved.

Is Wheeler and Feynman's view of nature correct? Could we use advanced waves to send signals to the past? Or, if that requires some yet-to-be-developed technological breakthroughs in transmitter design, and if receivers are easier to construct, could we at least listen-in to the future (since we are, *now*, the future's past)? And if we could do that, could the future send us the details of the transmitter breakthrough (thus creating a causal information loop in time)?

The first experimental search for advanced waves seems to have been a 1973 effort.[70] The very next year, flaws in that search process prompted two physicists to discuss an experiment designed to detect advanced waves (if they exist). As they wrote, in a grand understatement, the exciting possibility of a positive result "would have such far-reaching consequences on our ideas of the unidirectionality of time and causality that . . . the experiment justifies a large amount of effort, even if no conclusive result is obtained for years."[71] Alas, all of the searches for advanced waves have, as I write (2016), given negative results and so the world still awaits the first Dirac radio.

Over the years the Wheeler and Feynman view of nature has been the target of some theoretical concerns. One physicist, for example, complained that Wheeler and Feynman had assumed a static, time-symmetric spacetime for the universe, in which the properties of all past and future absorbers are identical. That is obviously not so in an expanding (or contracting) universe and, as he wrote, "No serious modern cosmological theory is framed in [terms of] a static Universe."[72] Another puzzle for that writer was that Wheeler and Feynman took a time-symmetric theory of half-retarded/half-advanced waves in a time-symmetric universe and arrived at a *non*-time-symmetric solution! They performed that trick by supposing not only that the universe is static, but also that it was created with the initial condition of low entropy. Thus, for Wheeler and Feynman, the one-way thermodynamic arrow of time is the primary arrow, with the electromagnetic arrow following as a consequence. (The *how* of a low entropy initial cosmological condition was left unexplained—certainly no mention of the hand of God appears in their work!) This ordering of the primacy of the temporal arrows was, in fact, in agreement with the view adopted by Einstein in his 1909 debate with Ritz, a view taken decades later by Hawking, as well.[73]

---

[69] See note 34 in Chap. 3. There Wheeler also wrote "The particles of the absorber are either at rest or in random motion before the acceleration of the source. They are correlated with it in velocity after that acceleration. Thus radiation and radiative reaction are understood in terms, not of pure electrodynamics, but of statistical mechanics."

[70] R. B. Partridge, "Absorber Theory of Radiation and the Future of the Universe," *Nature*, August 3, 1973, pp. 263–265.

[71] M. L. Herron and D. T. Pegg, "A Proposed Experiment in Absorber Theory," *Journal of Physics A*, October 1974, pp. 1965–1969.

Wheeler and Feynman had shown that both the advanced and retarded solutions taken together are self-consistent in a static universe; Hogarth's question was whether the observed retarded solution, *alone*, would be self-consistent in an *expanding* universe (which is the universe we actually observe). His conclusion? It depends on the details of the expansion. Two years after Hogarth, two physicists expanded on his work and claimed to have shown that the retarded solution alone *is* self-consistent *if* the expansion is steady-state via the continuous creation of matter.[74] That would be the case because, if only retarded effects are to occur, then each emitter of radiation needs a large number of absorbers (such as ionized intergalactic gas) in its future light cone to provide for complete absorption. This, in turn, requires that the density of matter not decline "too fast" with the expansion. That is, the future universe must not be "too transparent" and the continual appearance of new matter in the ever-increasing volume of the expanding universe is required to maintain the necessary density.

That conclusion was embraced with particular enthusiasm by Hoyle, a British cosmologist whose name has long been identified with the idea of continuous creation of matter. Since then, however, continuous-creation cosmologies have fallen into disfavor because it was in 1965, just a year after Hoyle and Narlikar wrote, that the cosmic microwave background radiation was detected. That is now taken as very strong evidence for the occurrence of a primordial explosion (or Big Bang) that started the expansion of the universe, and as equally strong support for therefore rejecting a steady-state universe. Not by Fred Hoyle (1915–2001), though, who had an almost fanatical devotion to non-Big Bang cosmologies.[75] Real puzzles remain for the Big Bang universe, however. One is that it expands from a dense, opaque past into a less dense, ever-more-transparent future, with each emitter having a large number of absorbers in its *past* light cone. That should result, noted Hoyle and Narlikar (almost certainly with some glee), in an observed advanced solution and thus in a reversed electromagnetic arrow that would allow communication with the past. The fact that we have not (yet?) discovered how to perform such communication might be taken to mean that the idea of an expanding, Big Bang universe is somehow faulty. A related question about absorber theory is that of the puzzle of neutrino absorption. Neutrinos are particles that interact so weakly with matter that a beam of them would have to travel through many hundreds of light-years of lead for there to be a significant attenuation of the beam. How can such 'ghost-like' particles find enough future absorbers to make possible their observed journeys into the future of a Big Bang expanding universe?

---

[72] J. E. Hogarth, "Cosmological Considerations of the Absorber Theory of Radiation," *Proceedings of the Royal Society A*, May 22, 1962, pp. 365–383. Hogarth, however, rejected the static universe, asserting instead that the *observed expansion* of the universe provides the required asymmetry, resulting in the cosmological arrow of time as the primary arrow and the electromagnetic arrow as a consequence.

[73] S. W. Hawking, "Arrow of Time in Cosmology," *Physical Review D*, November 15, 1982, pp. 2489–2495.

[74] F. Hoyle and J. V. Narlikar, "Time Symmetric Electrodynamics and the Arrow of Time in Cosmology," *Proceedings of the Royal Society A*, January 1964, pp. 1–23.

For such an exciting idea as communication with the past, it is not surprising that advanced-wave radio has appeared in science fiction. Just 6 years after Wheeler and Feynman's paper, a story by a well-known author hinted at such a gadget based on something called the "ultrawave effect": "While gravitational effects were produced by the presence of matter, ultrawave effects . . . did not appear unless there was a properly tuned receiver somewhere. They seemed somehow 'aware' of a listener even before they came into existence."[76] It is difficult to believe that such a story idea was conjured-up out of nothing, but rather that the author had read Wheeler and Feynman's paper. Anderson had a 1948 undergraduate honors degree in physics from the University of Minnesota, and so he may *well have* read Wheeler and Feynman's 1945 paper.

The potential bilking paradoxes produced by sending messages backward in time have been treated in at least one novel-length discussion. The puzzles presented are undeniably fascinating, but the story's answer to them is to allow the changing of the past, as argued in this section's opening quote from John Cramer. Indeed, the title comes from the plot device of twice changing the past by sending messages to the past to save the world from terrible disasters. Thus, we read through entire time periods *three* times before finishing the novel. As one character blurts out, "We can monitor the actual consequences of our decisions and actions, and change them until they produce the desired result! My God . . . it's staggering!"[77] Quite so.

One of the most interesting science fictional uses of backward-in-time signaling is, I think, found in a classic tale by James Blish. There the "Dirac radio" for instantaneous transmissions is described, and we learn that at the beginning of each received message there is always an irritating audio beep (hence the title) that is seemingly a useless artifact of the mysterious workings of the gadget. Its only obvious characteristic is a continuous spectrum from 30 to well above 18,000 Hz. It is only at the end of the story that the main character learns that this spectrum is the “simultaneous reception of every one of the Dirac messages which [has] ever been sent, or will be sent.”[78]

---

[75] Besides his scientific work, Hoyle also wrote science fiction. One work, the 1966 novel *October the First is Too Late*, deals with travels in time but fails to say anything about paradoxes.

[76] Poul Anderson, "Earthman, Beware!," *Super Science Stories*, June 1951.

[77] J. P. Hogan, *Thrice Upon a Time*, Ballantine 1980.

Blish was actually pretty close to the mark with that, as a composite signal with a continuous spectrum (with energy distributed uniformly in frequency), such as one might expect the overlay of many independent signals to be, does indeed have a narrow time structure. If applied to a loudspeaker, such a signal would sound like a sharp pulse or click—or even a *beep*. In the limit of an infinitely wide spectrum, the time signal becomes one of infinite amplitude and zero duration, a singular *impulse* function called, by theoretical physicists and radio engineers alike, the *Dirac delta function*.

There is no mention in the story of advanced waves, but clearly Blish knew that instantaneous (infinite-speed) signals would travel into the past and he does a masterful job of presenting the mystery of listening to the future. At one point characters in the twenty-first century hear the commander of a time-traveling “world-line cruiser” transmit a poignant call for help from 11,000,000 light years away and from sixty-five centuries in the future. Most interesting of all, however, is Blish’s statement of a technical issue that I have not seen raised before: if signals arrive at a receiver, simultaneously, from all future times, how can they be separated? Blish resorts to some scientifiction babble-talk to answer that question, but I believe it remains a puzzle.[79]

5.6 Tachyonic Signals and the Bell Quantum Antitelephone

> “We cannot fight the laws of nature.”
>
> “Nature be damned! Feed more fuel into the tubes. We must break through the speed of light . . . Give me a clear road and plenty of fuel and I’ll build you up a speed of half a million miles in a second . . . What’s there to stop it?”[80]

Science fiction writers have often used FTL motion to reverse time, often without much (if any) regard to the fact that, to just *reach* the light barrier, requires (according to special relativity) infinite energy, much less to exceed light speed.

---

[78] J. Blish, “Beep,” *Galaxy Science Fiction*, February 1954.

[79] The signal separation problem is also hinted at by physicist/science fiction author Gregory Benford, in a tale that was a precursor to his famous 1980 novel *Timescape* (in which the present attempts to warn the past of a future ecological disaster that threatens life on Earth). See Benford’s “Cambridge, 1:58 A.M.,” *Epoch*, Berkeley 1975.

[80] Words exchanged by the first officer and the captain of a starship on its way to Alpha Centauri in a story by N. Schachner, “Reverse Universe,” *Astounding Stories*, June 1936. The captain, we are told, “had heard, of course, of the limiting velocity of light, but it meant nothing to him.”

Therefore, goes the reasoning, because we can't get through the 'light barrier' means that time travel to the past must be impossible, as well. Or, so goes this line of argument. But, could there be a way around this conclusion? After all, while relativity theory indeed precludes the acceleration of a *massive* particle *up to* the speed of light, it does allow a zero rest mass particle (like the photon) to exist *at* the speed of light. Photons are emitted during various physical processes, and they move from the instant of their creation at the speed of light; the only way to slow a photon is to destroy it by absorbing it. Advocates of the possibility of the existence of FTL particles make a similar argument when asking if there might not be particles, emitted during various (as yet unknown) physical processes, that move from the instant of their creation at speed *greater* than that of light?

An affirmative answer would neatly avoid the 'acceleration through the light barrier' problem, but then there are other concerns. For example, such FTL particles would have to have an imaginary rest mass if they were to carry real-valued energy and momentum, and what could *imaginary mass* mean? That question was answered by the proponents of FTL particles, who replied that the rest mass of a superluminal particle would be unobservable because (like the photon) there is no subluminal frame of reference in which the particle could be at rest! That is, there is no frame of reference in which the mysterious imaginary mass could be measured and, anyway, it is only *observable* changes in the real energy and momentum that characterize particle interactions.

The key idea to this line of thought is a *supposed* FTL particle, called a *tachyon*, a name coined by the American physicist Gerald Feinberg (1933–1992) from the Greek word *tachys* for "swift."[81] It is interesting to note that Feinberg admitted[82] that his interest in such a thing was sparked by reading Blish's story "Beep" (see note 78). The idea of the tachyon is actually a very old one that is hinted at in the work of the Greek poet and philosopher Lucretius (who died 20 years before the birth of Christ). In his discussion of visual images, in Book 4 of his giant (well over 7400 lines) science poem *De Rerum Natura*, we find the following words about particles of matter originating deep inside the Sun: "Do you see how much faster and farther they must travel, how they must run through an extent of space many times vaster in the time it takes the light of the Sun to spread throughout the sky?"[83]

The first attempt (later found to be flawed) in the physics literature of a relativistic treatment of FTL particles appeared some years before Feinberg gave

---

[81] G. Feinberg, "Possibility of Faster-Than-Light Particles," *Physical Review*, July 25, 1967, pp. 1089–1105. Feinberg was anticipated in this name by Edward Page Mitchell (the Victorian pioneer in the time travel paradox genre who was discussed back in Sect. 4.2 and its note 37). In his story "The Tachypomp: A Mathematical Demonstration" (*Scribner's Monthly*, March 1874), he describes a gadget for reaching any speed, no matter how great (*tachypomp* is literally "quick sender").

[82] G. Benford, "Time and *Timescape*," *Science-Fiction Studies*, July 1993, pp. 184–190.

[83] A poetic allusion to something traveling faster than light appears, in of all places, Shakespeare's *Romeo and Juliet*. In Juliet's words (Act II, scene 5), ". . . love's heralds should be thoughts, Which ten times faster glide than the sun's beams, Driving back shadows . . ."

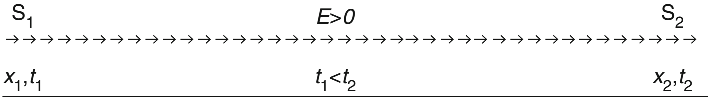

Fig. 5.4 The emission of a positive energy particle, followed by absorption

them a name, where it was observed that special relativity is *not* necessarily violated by FTL motion.[84] But nonetheless, concerns about the physical possibility of FTL particles continued. For example, a serious problem for tachyons, at least for those who dislike the ideas of backward causation and time travel, is the observation that in some frames of reference an FTL particle would appear to have *negative energy*. Feinberg, himself, explained (see note 81) this concern as follows: "By the principle of relativity, any state which is possible for one observer must be possible for all observers, and hence FTL particles can exist in negative-energy states for all observers . . . The occurrence of negative energy states for particles has always been objected to on the grounds that no other system could be stable against the emission of these negative-energy particles, an entirely unphysical behavior."

This objection to FTL particles was raised early in the history of tachyons, even before they were named, and it was addressed by three physicists who proposed the so-called *reinterpretation principle* (what I'll refer here to as the RP).[85] To see how the RP works, consider Fig. 5.4, in which a source S_1 at x_1 emits an FTL particle at time t_1. This particle then travels to an absorber S_2 at x_2, arriving there at the later time t_2. S_1 and S_2 are in the same reference frame and, for an observer in the frame, the particle energy E is positive. However, it is always possible to find another observer in a relatively moving frame for whom this process would look as though t_2 is less than (that is, earlier than) t_1 with $E < 0$. In other words, for the moving observer the particle would appear as a negative energy particle moving backward in time. (In the next chapter I'll show you that the particle speed must not be just superluminal, but the even faster *ultraluminal*.)

Note that for the moving observer, the emission by S_1 of negative energy *increases* the energy of S_1, and the absorption of negative energy by S_2 decreases the energy of S_2. S_2's energy decrease (for the moving observer) occurs before the increase in S_1's energy because, as noted before, for the moving observer $t_2 < t_1$. The moving observer naturally interprets this process as the emission of positive energy by S_2, followed by absorption by S_1. This reinterpretation would thus seem to preserve our common-sense idea of causality, as well as avoiding any mention of backward time travel. The RP appears to have slipped around those problems merely by redefining which source is transmitting, and which is receiving, the tachyon. Indeed, Feinberg claimed (note 81) that the RP avoids the creation of causal loops and their associated paradoxes, a claim repeated nearly 20 years later by a physicist who used the RP to eliminate paradoxes from Gödel's time travel rotating universe.[86]

---

[84] S. Tanaka, "Theory of Matter with Super Light Velocity," *Progress of Theoretical Physics*, July 1960, pp. 171–200. See also O. M. Bilaniuk and E. C. G. Sudarshan, "Particles Beyond the Light Barrier," *Physics Today*, May 1969, pp. 43–51 (and the resulting discussion in the December issue, pp. 47–52).

[85] O. M. P. Bilaniuk, V. K. Deshpande, and E. C. G. Sudarshan, "'Meta' Relativity," *American Journal of Physics*, October 1962, pp. 718–723.

There is, however, a curious twist to all this. Even if we grant that the RP may avoid causal paradoxes, the fact is that physics isn't fooled as easily as a human observer. That is, the receiver does actually lose energy upon the arrival of the tachyon, which is the opposite of what happens in a radio receiver when it receives a photon. In other words, the receiver must be in an elevated energy state prior to the tachyon's arrival; the receiver must be prepared beforehand to receive a message. If the receiver were instead sitting in its lowest energy state, then it could not accept (or *eject*, according to the RP) the tachyon. So, it's not surprising that, despite the enthusiastic embrace of the RP by some, other physicists took exception, with one arguing that the effectiveness of the RP in avoiding causal loops is "illusory" and "irrelevant,"[87] while others concluded that the causal paradoxes would actually preclude any possibility of tachyons interacting with ordinary matter in the first place (which is just a polite way of saying that tachyons have no more reality than do unicorns!).[88]

The RP's effect of flipping the roles of transmitter and receiver has attracted particular concern. Some analysts have pointed out that if one can modulate a superluminal signal to send a message into the past, then certainly one could *sign* the message. To quote a delightful example, "If Shakespeare types out *Hamlet* on his tachyon transmitter, Bacon receives the transmission at some earlier time. But no amount of reinterpretation will make Bacon the author of *Hamlet*. It is Shakespeare, not Bacon, who exercises control over the content of the message."[89] The last line of this quote is of central importance. The authors emphasize it by immediately observing that a signature is a *relativistic invariant* and that, indeed, it establishes a causal ordering quite independent of any temporal ordering. This example, alone, explains why one analyst said of the RP that it is "laughed to scorn,"[90] while another said of the RP that it "sounds merely like the endorsement of what can only be characterized as a fantastic delusion."[91]

---

[86] A. Italiano, "How to Recover Causality in General Relativity," *Hadronic Journal*, January 1986, pp. 9–12.

[87] R. G. Newton, "Particles That Travel Faster Than Light," *Science*, March 20, 1970, pp. 1569–1574.

[88] W. B. Rolnick, "Implications of Causality for Faster-Than-Light Matter," *Physical Review*, July 25, 1969, pp. 1105–1108, and D. J. Thouless, "Causality and Tachyons," *Nature*, November 1, 1969, p. 506.

[89] G. A. Benford, D. L. Book, and W. A. Newcomb, "The Tachyonic Antitelephone," *Physical Review D*, July 15, 1970, pp. 263–265.

[90] P. Fitzgerald, "On Retrocausality," *Philosophia*, October 1974, pp. 513–551.

[91] W. L. Craig, "Tachyons, Time Travel, and Divine Omniscience, *Journal of Philosophy*, March 1988, pp. 135–150.

Even more sophisticated scenarios than the *Hamlet* one (note 89) were devised to show how problems with FTL signals could arise that the RP could *not* resolve. The American physicist Bryce DeWitt (1923–2004) created such an example (the 'DeWitt Gambit') that involved a sequential, circular chain of tachyon signal transmissions between four observers, all moving in one spatial dimension. DeWitt showed out to arrange the spacetime geometry of the observers so that, at each stage, there is no dispute over who is sending and who is receiving, and so invoking the RP is avoided. Yet, when the signal reaches the first (last) observer, it is *before* he started (will start) the chain! This, of course, sets-up a potential bilking paradox: what if 'now' the first (last) observer decides to *not* send the chain's initiating signal? An even more sophisticated variation on the Gambit, involving four observers moving in *two* spatial dimensions, was soon after put forth by the English physicist Felix Pirani (1928–2015). As he concluded, "It is difficult to see how in the face of this example a classical-particle description of tachyons can be sustained."[92] Confronted by such sharp criticism, from so many, it is understandable why, just before his death, Feinberg co-authored (with two philosophers) a paper in which he seemed to be abandoning his support for tachyons as possible carriers of information backward in time.[93]

Many physicists today reject the possibility of backward-in-time messages, not because of concerns about the RP, but because such messages could create potential bilking paradoxes. To see how this works, the old Wheeler and Feynman idea of explaining (away) bilking paradoxes—that no signal in nature is *really* discontinuous—was examined by one physicist in the context of tachyons.[94] There we are asked to consider the following situation: A human (call him A) has a lamp on a table before him. The lamp is controlled by a tachyon receiver; in other words, the lamp illuminated only when a tachyon signal (a pulse, let us say) is detected. At 3 o'clock A will send a tachyon signal to B (a tachyon echo-transmitter that immediately rebroadcasts everything it receives) if the lamp does not glow at 1 o'clock. Now, the spacetime geometry of A and B is arranged to be such that a signal sent by either A or B to the other travels 1 h backward in time. Thus, if A sends a signal at 3 o'clock, then B will receive it at 2 o'clock (and immediately echo), and the echo will arrive at A at 1 o'clock. The paradox, of course, is that A sends a signal only if the lamp does *not* glow—that is, only if A does *not* send the signal![95]

We are then reminded of the Wheeler and Feynman claim that every pulsed signal is actually continuous; this argument would include the illumination itself of the lamp. Therefore, the lamp is not just on or off, but potentially at any level of illumination between those two extremes. So there sits A, and at 1 o'clock the lamp seems to glow dimly. To that, says Schulman, "A thinks it over, vacillating, finally sending a slightly late signal which isn't full strength." Then the echo isn't full strength either, which accounts for the original dim glow. This conclusion *is* consistent (but what if A's sending device is a toggle switch that *snaps* one way or the other—why then only a partial strength signal?), but it does seem to ask for a lot of supposing. Schulman himself is not so sure about the validity of universal continuity, writing at the end that "it is not clear that the Wheeler-Feynman assumption . . . ought to be made." (For more on the Wheeler-Feynman continuity assumption, in a different, non-time travel context, see the final "For Further Discussion" for this chapter.)

---

[92] F. A. E. Pirani, "Noncausal Behavior of Classical Tachyons," *Physical Review D*, June 15, 1970, pp. 3224–3225.

[93] G. Feinberg, D. Albert, and S. Levine, "Knowledge of the Past and Future," *Journal of Philosophy*, December 1992, pp. 607–642.

[94] L. S. Schulman, "Tachyon Paradoxes," *American Journal of Physics*, May 1971, pp. 481–484.

[95] A study of similar situations can be found in L. L. Gatlin, "Time-Reversed Information Transmission," *International Journal of Theoretical Physics*, January 1980, pp. 25–29.

The Wheeler-Feynman continuity idea *is* ingenious, allowing one to find a logically and physically consistent solution in time travel scenarios that, at least at first glance, might seem to have no solution. Consider, for example, the following situation[96]: "We have a camera ready to take a black and white picture of whatever comes out of [a] time machine. The film is then developed and the developed negative is subsequently put in the time machine and set to come out of the time machine at the time the picture is taken. This surely will create a paradox: the negative will have the opposite distribution of black, white, and shades of grey, from the picture that comes out of the time machine. But since the thing that comes out of the time machine is the negative itself we surely have a paradox." But do we?

The answer is no, because "What will happen is that a uniformly grey picture will emerge which produces a negative that has exactly the same uniform shade of grey. No matter what the sensitivity of the film is, as long as the dependence of the brightness of the negative depends in a *continuous* [my emphasis] manner on the brightness of the object being photographed, there will be a shade of grey that produces exactly the same shade of grey on the negative when photographed. This is the essence of Wheeler Feynman's idea."[97] (The conclusion is the same if we move from black-and-white to color photography.)

Nonetheless and despite this apparent success, the supposed ability of a modulated beam of tachyons to send a message into the past still raised concerns among many, particularly about free will and fatalism. Suppose, say those who are concerned about these issues, that you receive a tachyon message from yourself from tomorrow, informing you that a man you plan to kill tonight is still alive (tomorrow). Does that mean it is beyond your power to kill him tonight? According to one analyst (note 67) the answer is no; you *could* kill him—but if you do then the message from tomorrow would not have arrived. And because ignorance is not a precondition to free will, your newly acquired knowledge does not, by itself, suddenly limit your ability to kill the man. But, this line of arguing went, if you do not attempt to kill the man because you believe the message from your future self, then in fact the message *has* limited you!

---

[96] Taken from Frank Arntzenius and Time Maudlin, "Time Travel and Modern Physics," in *Time, Reality & Experience* (C. Callender, editor), Cambridge University Press 2002, pp. 169–200.

[97] More on the Wheeler-Feynman continuity idea, *and of its limitations*, can be found in D. Kutach, "Time Travel and Consistency Constraints," December 2003, pp. 1098–1113, and Phil Dowe, "Constraints on Data in Worlds with Closed Timelike Curves," December 2007, pp. 724–735, both in *Philosophy of Science*.

Fitzgerald's position was rebutted by Craig (note 91), who argues that it is not your ability to kill that is altered by the message but rather your motivation. Craig points out that such motivational changes can occur without invoking anything as radical as a message from the future. Suppose, he says, that just before you fire the fatal shot into your victim, you learn from him that he is your beloved, long-lost uncle. Clearly, your motivation for killing him is likely to be instantly altered, but equally clearly, your *ability* to kill him is unchanged. The mechanism for obtaining genealogical information, whether via time travel or as a last-minute appeal from your intended victim, is (says Craig) simply irrelevant. Not all buy into that, however, with one unconvinced analyst (note 94) writing that "history is a set of world lines essentially frozen into spacetime. While subjectively we feel strongly that our actions are determined only by our backward light cone, this may not always be the case." That is, Schulman appears open to the possibility that influences originating in the future might indeed have an impact on the present.

With the fading from the physics scene of enthusiasm for tachyons, the romance of communicating with the past using superluminal speeds passed from speedy particles to quantum mechanics via a mathematical result called *Bell's theorem*. John Bell, a physicist mentioned in the last chapter in connection with the MWI of quantum mechanics, published his theorem in a little article in an obscure, now defunct journal. Since then it has become one of the most cited physics papers from the 1980s.[98] The paradox cited in Bell's title refers to a famous 1935 paper in which Einstein (and two of his colleagues at the Institute for Advanced Study) challenged the conventional view of quantum mechanics, the view that there is no objective reality to anything unless it is observed.[99]

In fact, the possibility of quantum mechanics supporting FTL signals had been considered (and rejected) *before* Einstein's paper, by the Italian physicist Enrico Fermi.[100] Fermi concluded that, in a two atom system, the decay from an excited state of one atom (with the emission of a photon) would not influence the other atom before a time lapse of R/c, where R is the distance between the two atoms and c is the speed of light. In 1967, however, the Russian physicist M. I. Shirokov pointed out that Fermi's result was the result of an unjustified mathematical operation (he had replaced an integral from zero to infinity with one from minus infinity to infinity).

---

[98] J. S. Bell, "On the Einstein-Podolsky-Rosen Paradox," in *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press 1987.

[99] A. Einstein, B. Podolsky, and N. Rosen, "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?" *Physical Review*, May 15, 1935, pp. 777–780. See also N. D. Mermin, "Is the Moon There When Nobody Looks? Reality and Quantum Theory," *Physics Today*, April 1985, pp. 38–47.

[100] E. Fermi, "Quantum Theory of Radiation," *Reviews of Modern Physics*, January 1932, pp. 87–132.

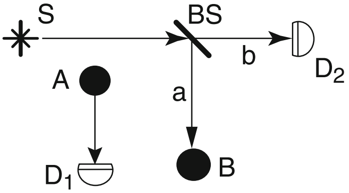

Fig. 5.5 Time travel and the wave function of quantum mechanics. In the above arrangement (reproduced by the kind permission of Francesco Gonella, from his paper "Time Machine, Self-Consistency and the Foundations of Quantum Mechanics," *Foundations of Physics Letters*, April 1994, pp. 161–166), S is a low-intensity source of photons, a source *so* weak that at any instant there is never more than one photon in the system. Each photon begins by traveling toward BS, a half-silvered beam-splitting mirror that, with equal probability, either passes a photon to the right along path b where it is detected by D_2, or downward along path a into mouth B of a wormhole time machine and out of which it emerges from wormhole mouth A *in the past*. Conventional quantum mechanics says that what happens at BS is determined *when the photon reaches BS*—that is, the probability wave function of the photon collapses at BS. But for a photon that is reflected into B, it exits A where it is detected by D_1 *before* the time of its arrival at BS. That means it is known (by detector D_1) what *will* happen at BS *before* the photon arrives at BS—so just *when did* the photon wave function collapse?

Then, nearly 30 years after Shirokov, the German physicist Gerhard Hegerfeldt proved (in the context of conventional quantum mechanics) a very general theorem that establishes a non-zero probability of the second atom responding to the first atom's decay *as soon as* the photon is emitted.[101]

Hegerfeldt's analysis is very general, but it does make one assumption—the non-negativity of energy density (the so-called *weak energy condition*)—which at the time was considered (in Hegerfeldt's words) to be "physically well motivated," but which today is *not* taken to be *a priori* obvious.[102] Einstein's paper, on the other hand, utilized physical assumptions not easily dismissed, and so its conclusions confounded physicists for decades.

The conventional view of quantum mechanics formulates physics in terms of probability wave functions that 'collapse' into specific realities only when measurements (observations) are made of the state of a system (which may be as elementary as a single particle). Until such measurements are made, says this view of quantum mechanics, a system has no specific state; instead, it merely has a probability distribution over a set of possible states (see Fig. 5.5 for a time travel puzzle concerning this claim). Einstein and his co-authors (note 99) strongly rejected this probabilistic interpretation of nature (recall Einstein's famous dictum, "God does not play dice with the cosmos.") Einstein and his colleagues agreed that quantum mechanics may be valid *as far as it goes*, but they also argued that it leaves out 'something' (as yet unknown) in describing reality. That is, they suggested that quantum mechanics is "incomplete" and that it incorporates "hidden variables."[103] They expressed this idea in the form of a paradox, the famous *EPR paradox*, which they posed as a thought experiment (a *Gedanken-experiment*) in which quantum mechanics declares that the properties of a particular spatially distributed system, when measured at point A, seem to be *forced* to assume specific values at point B (which may be arbitrarily distant from A) *without* there being a measurement at B.

---

[101] G. C. Hegerfeldt, "Causality Problems for Fermi's Two-Atom System," *Physical Review Letters*, January 1994, pp. 596–599. This mere *suggestion* of a possible failure of causality so stunned the editor of *Nature* that he felt compelled to quickly write a 'calming' reply: "Time Machines Still Over the Horizon," February 10, 1994, p. 509.

[102] The issue of the sign of energy density is very important in the analyses of wormhole time machines (see note 135 in Chap. 1), and we'll return to it in the next chapter.

Thus, said Einstein, there are just two possibilities. Either the system properties at B must have been what they are from the very start (even if the measurements at A had not been done) which is the view he held, or there must have been a linkage between the system at A and the system at B such that the wave function collapse at A is instantly transmitted to B to allow the wave function to collapse there as well. Because A and B may be arbitrarily far apart, this second view obviously requires an FTL transmission mechanism,[104] something Einstein called a "spooky action-at-a-distance," a term that eloquently expresses his low opinion of that idea! (Some translations replace *spooky* with *ghostly*, but the negative sentiment remains the same.[105]) For decades the debate between proponents of these two alternatives remained at a metaphysical, non-quantitative level. Then came Bell's paper in 1964.

Bell's theorem mathematically poses the choice between Einstein's hidden variables view and the conventional view of quantum mechanics through the use of an inequality involving certain measurable properties of a system.[106] If these measurements are such that the inequality is violated, then the conventional interpretation of quantum mechanics is vindicated, and Einstein's FTL spooky action-at-a-distance effect simply doesn't exist. Bell's great contribution, then, was to provide the means for removing the debate about quantum mechanics from metaphysics and to place it squarely in the realm of experimental physics.

"All" that needed to be done was to make the required measurements. These technically difficult experimental measurements were eventually performed by the French physicist Alain Aspect and his colleagues at the Institute of Applied Optics of the University of Paris, a decade and a half after Bell showed what had to be done. The results unequivocally support the conventional view of quantum mechanics.[107] Einstein was simply wrong, there are no hidden variables in quantum mechanics, and his spooky action does seem to exist. Does that mean we have, at last, experimental evidence of the possibility of information transfer at FTL speeds?

---

[103] For more on hidden variables, see E. P. Wigner, "On Hidden Variables and Quantum Mechanical Probabilities," *American Journal of Physics*, August 1970, pp. 1005–1009.

[104] For an analysis that argues against an FTL mechanism in quantum mechanics, see G. C. Ghirardi, *et al*., "A General Argument Against Superluminal Transmission Through the Quantum Mechanical Measurement Process," *Lettere Al Nuovo Cimento*, March 8, 1980, pp. 293–298.

[105] You can find a discussion of the possibility of 'explaining' Einstein's "spooky actions" of quantum mechanics by invoking backward causation in R. I. Sutherland, "Bell's Theorem and Backwards-in-Time Causality," *International Journal of Theoretical Physics*, April 1983, pp. 377–384.

[106] The details are not important here, but a lovely exposition (for the lay person) can be found in Bell's essay "Bertlmann's Socks and the Nature of Reality," in *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press 1987. See also M. G. Alford, "Ghostly Action at a Distance: A Non-Technical Explanation of the Bell Inequality," *American Journal of Physics*, June 2016, pp. 448–457.

Well, maybe, but the issue is still hotly debated. The majority of physicists today, I suspect, are probably more perplexed over what Bell's theorem is saying than were over Einstein's original EPR paradox. In those early days one could agree with Einstein and his colleagues, who argued that quantum mechanics was valid *as far as it went*, but a deeper, more comprehensive theory *would* show the existence of hidden variables. However, because of the work by Bell and Aspect it is definitively known that quantum mechanics *as it stands* leads to correct results, results that can be checked in the laboratory. In other words, there is no need for hidden variables and FTL spooky actions cannot be ruled out.

And so, while tachyons and the tachyon antitelephone may be nothing more than a neat science fiction fantasy, just maybe a quantum mechanical Bell antitelephone can't be so dismissed. Indeed, the possibility of using a quantum mechanical FTL effect was once suggested in a letter written by a senior person at an unspecified California think tank (an organization such as, for example, the RAND Corporation) to the Under Secretary of Defense for Research and Engineering at the Pentagon. Here's what the Under Secretary read when he opened that letter: "If in fact we can control the FTL nonlocal effect, it would be possible . . . to make an untappable and unjammable command-control-communication [C^3] system at very high bit rates for use in the submarine fleet. The important point is that since there is no ordinary electromagnetic signal linking the encoder with the decoder in such a hypothetical system, there is nothing for the enemy to tap or jam. The enemy would have to have actual possession of the 'black box' decoder to intercept the message, whose reliability would not depend on separation from the encoder or on ocean or weather conditions."[108]

One can't help but wonder what might have been the Under Secretary's response to that incredible letter, and what sorts of ultra-mega-super-top-secret experiments it may have prompted. I would be willing to bet, if they did occur, that they failed. As one physicist put it, "Up to now nature has covered her tracks pretty well, blocking all possibilities for using the EPR effect for FTL communication."[109] Of course, the think tank letter, as 'farout' as it may initially appear, actually represents a *failure* of imagination, because the backward causation effect of EPR's spooky FTL effect is certainly a 'quantum jump' beyond a mere unjammable submarine C^3 system.

---

[107] A. Aspect, "Experimental Tests of Realistic Local Theories via Bell's Theorem," August 17, 1981, pp. 460–467, and A. Aspect, *et al.*, "Experimental Realization of Einstein-Podolsky-Rosen-Bohm *Gedanken-experiment*: A New Violation of Bell's Inequalities," July 12, 1982, pp. 91–94, and "Experimental Test of Bell's Inequalities using Time-Varying Analyzers," December 20, 1982, pp. 1804–1807, all in *Physical Review Letters*.

[108] See N. D. Mermin in note 99. For more on the enigmatic letter on the FTL submarine C^3 system, see Jack Sarfatti's letter to *Physics Today*, September 1987, pp. 118 and 120.

[109] J. Cramer, "Paradoxes and FTL Communication," *Analog Science Fiction*, September 1988.

5.7 For Further Discussion

Mark Twain, in his last, posthumously published novel *No. 44, The Mysterious Stranger*, incorporated reversed time as the work of a supernatural being. After the being reverses the world's time direction, the narrator tells the reader that "everywhere weary people were re-chattering previous conversations backward . . . where there was war, yesterday's battles were being refought, wrong-end first; the previously killed were getting killed again . . . we saw Henry I gathering together his split skull . . ." Read the novel, and comment on how well (or not) Mark Twain handled reversed time.

In Philip K. Dick's 1956 novel *The World Jones Made*, we read of a prophet who can see a year into the future. As he says, "To me *this is the past*," and then later we are told "He was a man with his eyes in the present [the world's future] and his body in the past [the world's present]." Read the novel, and then argue either for or against the suggestion that Dick was aware of the advanced wave solution in Maxwell's theory.

The physicist Robert Forward (see the sixth *For Future Discussion* in Chap. 3) argued that one way to send messages into the past is to compress a 15-billion-ton asteroid down to the volume of an atomic nucleus, spin it up, and then aim gamma ray bursts through the resulting near-by region of "unhinged time" (see "How To Build a Time Machine," *Omni*, May 1980). This is, of course, 'simply' an artificially constructed Kerr black hole telegraph transmitter (look back at note 114 and related discussion in Chap. 1). Forward, an optimist of the first rank, thought humans would be able to do this before the end of the twenty-first century. It would seem, then, that what should be done *now* is to build gamma ray frequency *receivers* (well within present-day technology) and listen for such messages from the future. The technical details of such receivers wouldn't matter, as long as their design is widely published. That way, the scientists of the future can learn those details by simply reading of them in old, musty library books and journals, and thus will be able to build their transmitters to be perfectly compatible with our old, musty (to them) receivers! Comment on the likelihood of National Science Foundation funding becoming available to build such receivers.

Suppose a time traveler goes into the future and, while there, discovers that there is an older version of himself living in the home that has been in his family for generations. Explain why this implies that the time traveler will eventually return to the present. Suppose, instead, that after the time traveler arrives in the future he decides to remain in the future, and *not* to return. Explain why this implies there will *not* be an older version of himself living in the family home. In both cases, assume the MWI does *not* apply, that is, assume that there is just a single time line.

One well-known quantum physicist, David Bohm (1917–1992), wrote the following passage in his book *The Special Theory of Relativity* (W. A. Benjamin 1965), concerning the possibility of sending messages into the past: "In effect, S could communicate with his own past [self, M] . . . and tell his past self [M] what his future is going to be. But on learning this M could decide to change his actions, so that his future . . . would be different from what his later self [S] said it was going to be. For example, the past self could do something that would make it impossible for the future one to send the signal. Thus, there would arise a logical self-contradiction." Do you think most physicists, writing today more than 50 years after Bohm, would repeat his words?

The role played by quantum mechanics in time travel studies is broad, deep, profound . . . and mysterious. What I mean by this is nicely illustrated by the final paragraph in a paper by Stephen Hawking ("Quantum Coherence and Closed Timelike Curves," *Physical Review D*, November 15, 1995, pp. 5681–5686). There he wrote "Personally, I do not believe that closed timelike curves will occur, at least on a macroscopic scale. I think that the chronology protection conjecture will hold and that divergences in the energy-momentum tensor will create singularities before closed timelike curves appear. However, if quantum gravitational effects somehow cut off these divergences, I am quite sure that quantum field theory on such a background will show loss of quantum coherence. *So even if people come back from the future, we will not be able to predict what they will do* [my emphasis]." What do you think Hawking meant by his final line? If "people come back from the future" and tell us what they did while in the

(continued)

future, then what's wrong with their memories that causes us to fail to be able to predict what they *will* do? Or, is there perhaps nothing wrong with their memories and Hawking is instead arguing that the future experienced by the returned time travelers will not be the future when *we* 'get there'? If that's the case, then what are the returned time travelers 'remembering'?

If quantum mechanics is actually slightly non-linear physics (as are many other normally linear physical phenomena at sufficiently high energy levels)—physics is *linear* when *superposition* holds, which means the result of two inputs is the sum of the individual outputs resulting from application of the individual inputs—and if one accepts the MWI concept, then at least two physicists (**PHYSICISTS1**) claim to have shown that one could communicate not just with our past, but also with the many pasts in the ever-splitting branches of the many worlds. Another physicist (**PHYSICIST2**) wrote a very funny illustration of what that might be like. Yet another physicist (**PHYSICIST3**) suggested that non-linear quantum mechanics might actually allow one to take *photographs* of the many-worlds. Of that, a physicist (**PHYSICIST4**) wrote (without any exaggeration) that such an achievement would be "perhaps the most amazing discovery in the history of science, indeed in the history of mankind." Or, to quote yet another physicist (**PHYSICIST5**), "interworld communication would lead to truly mind boggling possibilities," some of which have been incorporated in at least one science fiction novel (**SFAUTHOR**). Read the physicists' papers, and the novel, and then summarize with your own commentary.

PHYSICISTS1: N. Gisin, "Weinberg's Non-linear Quantum Mechanics and Superluminal Communication," *Physics Letters A*, January 1, 1990, pp. 1–2, and J. Polchinski, "Weinberg's Non-linear Quantum Mechanics and the Einstein-Podolsky-Rosen Paradox," *Physical Review Letters*, January 28, 1991, pp. 397–400.

PHYSICIST2: J. G. Cramer, "Quantum Telephones to Other Universes, to Times Past," *Analog*, October 1991.

PHYSICIST3: D. Albert, "How to Take a Photograph of Another Everett World," *Annals of the New York Academy of Sciences*, December 30, 1986, pp. 498–502.

PHYSICIST4: M. A. B. Whitaker, "On the Observability of 'Many Worlds'," *Journal of Physics A*, July 11, 1985, pp. 1831–1834.

PHYSICIST5: R. Plaga, "On a Possibility to Find Experimental Evidence for the Many-Worlds Interpretation of Quantum Mechanics," *Foundations of Physics*, April 1997, pp. 559–577.

SFAUTHOR: J. P. Hogan, *Paths to Otherwhere*, Baen 1996.

One upon a time, FTL tachyons (and their associated into-the-past transmissions of information) were strictly in the province of science fiction. In recent years, however, such doings have moved into mainstream fiction, as well, with the most recent (as I write) example being the man-of-action novel by Patrick Lee, *Signal*, St. Martin's Press 2015. (Tachyons are also mentioned in the 2015 film *Tomorrowland.*) That novel is set in modern times, not in the future; its hero is a retired soldier who now works as a self-employed house-flipper. His is definitely not a futuristic science fiction world. Until, that is, be becomes involved with a radio-like gadget that receives signals from 10 h, 24 min in the future. As the novel progresses we eventually learn that the gadget is based on German electronics technology that was being tested in a remote lab in northern Algeria, near the end of World War II. When that lab was overrun by a small American force the equipment was destroyed, but not before one of the Americans heard the gadget playing a song titled "She Loves You" along with the word *yeah* repeated numerous times. He didn't know what that meant until, 20 years later, he watched the Beatles' first American appearance on *The Ed Sullivan Show* TV program, and so suddenly realized that in 1944 he had heard a song that hadn't been written yet! That gadget is at the center of a modern-day, renewed Nazi effort to conquer the world, and there is much 'you've seen it all before' chasing, shooting and other 'James Bond' types of action in the novel, but the author *has* been quite inventive in treating time paradoxes. He does talk a lot about 'changing the future,' worries confusingly about 'changing the past,' and mistakes neutrinos for tachyons, but, still, if one is willing to overlook such issues it *is* an entertaining read. In particular, while the gadget's inherent range to the future is limited to ten-plus hours, the author describes a clever way to arbitrarily extend that value. Read *Signal*, and then describe and critique the method outlined in the novel.

The Wheeler-Feynman assumption of continuity already had a distinguished history in mathematics long before they invoked it in their physics resolution of bilking paradoxes. (Both men were very good mathematicians, and certainly knew what I am about to tell you here.) Imagine a man who is about to walk up a hill, starting at A (the base of the hill) and ending at B (the top of the hill). You know nothing of *how* he walks (perhaps at times he stops for a while, other times he walks slowly, sometimes he walks briskly, perhaps at times he even walks back *down* the hill). All you know is that, starting from A at 10 o'clock in the morning, he arrives at B at 11 o'clock. That is, the walk up

(continued)

the hill takes exactly 60 min. He camps overnight at B and then, the next morning, at exactly 10 o’clock, he walks back down the hill along the same path he followed during his ascent the previous day. He arrives at A at 11 o’clock. That is, the return trip takes exactly 60 min. Again, you know nothing of the details of how he makes the descent. Prove that there is at least one spot on the path that he passes at exactly the same time during his descent as he passed it during his ascent the previous day. Hint: No complicated equations are required. Indeed, no math at all is needed. Just sketch the appropriate, general graphs of the man’s ascent versus time, and of his descent versus time, and invoke continuity. That is, sketch the distance (as measured along the path) that he is from A versus time for both the ascent and the descent.

Chapter 6
The Physics of Time Travel: II

> *"It is very sad to see valuable minds writing such a pile of unmitigated bullshit."*
>
> —not all physicists think time travel is worthy of study.[1]

6.1 Faster-than-Light into the Past

> "Faster-than-light travel remains a coherent, and possible concept, even though it is forbidden by relativity theory."
>
> —a philosopher makes a physics mistake.[2]

So far we have limited our consideration of relativity theory to speeds below the speed of light—that is, to the condition $v < c$, where v is the relative velocity of two reference frames. There was nothing, however, in the derivation of the Lorentz transformation equations discussed in Chap. 3 that actually used that self-imposed constraint. So, just what, in fact, *does* happen for $v > c$? This is not an empty question, because the second half of the above quote that opens this section is simply not true. That isn't to say we can have FTL for free; there *is* a high price to pay, that of causality violation (although, if you are a fan of time travel, it's a price you are probably happy to pay). If a material object goes FTL, then the mathematics seems to say that the object could travel into the past, just as the caped crusader does in the first (1978) *Superman* movie, in order to change the past (to save Lois Lane from dying in an earthquake). In addition, the mathematics also seems to say that if a signal bearing information could achieve FTL, then that information, too, would travel back into the past (see Fig. 6.1).

---

[1] This rather blunt comment (reported in *Physics World*, December 2009, p. 3) was prompted by a suggestion, from two other physicists, that the Higgs boson might ripple backward through time and thereby stop CERN's Large Hadron Collider from creating the long-sought particle in the first place.

[2] G. Robinson, "Hypertravel," *Listener*, December 17, 1964, pp. 976–977, the printed version of a lecture to the British Association for the Advancement of Science.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_6

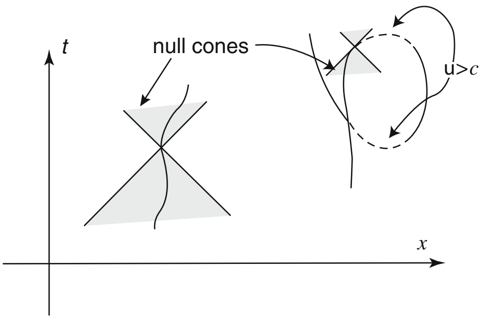

Fig. 6.1 A math-free proof that there can be no closed loops in time (that is, no time travel to the past if $v > c$ is forbidden) in flat Minkowski spacetime, the spacetime of the special theory of relativity. Such a conclusion is far less clear in the curved spacetimes of the general theory of relativity. Figure reproduced by the kind permission of Serguei V. Krasnikov (Polkovo Central Astronomical Observatory in St. Petersburg, Russia), from his 2003 paper "Time Machine (1988–2001)"

The original thinking along these lines visualized such an FTL signal as a modulated beam of tachyons, as mentioned in the previous chapter. In addition to tachyons not having been experimentally observed, even after intense searches for them, there are several theoretical objections (in addition to the bilking paradox problem discussed in the previous chapter) to the likelihood such FTL particles exist.[3] For example, the relativistic expressions for the energy and momentum of a particle with rest mass m_0 moving with speed v are, respectively,

$$E = \frac{m_o c^2}{\sqrt{1 - (v/c)^2}} \qquad \text{and} \qquad p = \frac{m_o v}{\sqrt{1 - (v/c)^2}}.$$

For $v > c$ the radicals in these expressions become imaginary, whereas E and p must always be real-valued (because they can be observed and measured as a result of the interactions the particle has with other matter). The energy and momentum can regain the property of being real-valued if we write $m_0 = \mu\sqrt{-1}$ (that is, $m_0^2 = -\mu^2$) for a tachyon, where μ is the positive, real-valued (but unobservable) meta-mass (that is, $m_0^2 < 0$). This is a radical proposal, of course, (as it is $m_0^2 > 0$ that we are used to; as a Russian mathematician romantically put it, "What binds us to space-time is our [positive] rest mass, which prevents us from flying at the speed of light, when time stops and space loses meaning. In a world of light there are neither points nor moments of time; beings woven from light would live 'nowhere' and 'nowhen'; only poetry and mathematics are capable of speaking meaningfully about such things."[4].

---

[3] See, for example, "More About Tachyons," *Physics Today*, December 1969, pp. 47–52, a collection of letters received by the journal from its readers. The 'DeWitt Gambit,' mentioned in the previous chapter, was first proposed in one of those letters.

Well, perhaps, but let's continue to pursue the physics of tachyons, at least for a while. It can be shown[5] that the energy and momentum of the tachyon are given by

$$E = \frac{\mu c^2}{\sqrt{(v/c)^2 - 1}} \quad \text{and} \quad p = \frac{\mu v}{\sqrt{(v/c)^2 - 1}}.$$

An interesting consequence of this is that if tachyons *lose* energy, they *speed up*, an observation first made by the German physicist Arnold Sommerfeld (1868–1951) in 1904.[6] This means that if there is a mechanism for continuous energy loss, such as Cerenkov radiation,[7] then tachyons will spontaneously accelerate without limit to *infinite speed*! Curiously, while the above expressions for E and p show that, as $v \rightarrow \infty$, tachyons would possess *zero* energy, they would nonetheless have a *non-zero* momentum of μc.[8]

To see how backward time travel and FTL are connected, it is useful to establish a geometrical interpretation of the Lorentz transformation. As stated in Chap. 3, if the x', t' system is moving with speed v in the x (or x') direction relative to the x, t system, then

$$x' = \frac{x - vt}{\sqrt{1 - (v/c)^2}} \quad \text{and} \quad t' = \frac{t - vx/c^2}{\sqrt{1 - (v/c)^2}}.$$

---

[4] Yu I. Manin, *Mathematics and Physics*, Birkhäuser 1981, p. 84.

[5] L. Parker, "Faster-Than-Light Inertial Frames and Tachyons," *Physical Review*, December 25, 1969, pp. 2287–2292.

[6] The consideration of FTL particles already had a long history *before* tachyons were specifically named. You can find late nineteenth century (1888–1889) theoretical analyses of electrically charged FTL particles in the writings, for example, of the English mathematical electrical engineer Oliver Heaviside (1850–1925): see my biography *Oliver Heaviside: the life, work, and times of an electrical genius of the Victorian Age*, The Johns Hopkins University Press 2002, pp. 124–126.

[7] Cerenkov radiation is the energy radiated when a charged particle exceeds the speed of light *in the medium through which it travels*. Since the speed of light *in water* is less than c, it is perfectly okay with special relativity to exceed the speed of light in water, and in fact this commonly occurs for the energetic electrons produced by submerged atomic reactors (swimming pool reactors). The resulting radiation is observed as a blue glow. The radiation is named after the Russian physicist Pavel Cerenkov (1904–1990)—for which he received a share of the 1958 Nobel physics prize—but in fact Heaviside (previous note) had predicted it more than a decade before Cerenkov was born.

[8] This (theoretical) property of tachyons (if they exist) could (perhaps) be used (maybe) to build a revolutionary (to say the least) new rocket propulsion system: see J. Cramer, "The Tachyon Drive: Infinite Exhaust Velocity at Zero Energy Cost," *Analog*, October 1993.

These equations make sense for the case of $v < c$, and we will retain this condition for our two relatively moving *frames of reference*: these frames are the *worlds* of two *human* observers, observers that we'll take to always be *sub*luminal. As one physicist so nicely put it, "The assumption that observers move faster-than-light goes beyond superluminal signaling,"[9] as such observers would have to be thought of as being built not out of flesh-and-blood, but rather out of tachyons.[10] We'll reserve the symbol w to denote the speed of an FTL particle.

Now, recall what we mean by any line parallel to the x-axis; it is a line with a fixed time coordinate. Such a line is a *cosmic moment* line, with the equation t = constant. Similarly, for the moving system we would write the equation of a cosmic moment line as t' = constant which, after the Lorentz transformation is applied, is equivalent to

$$t - \frac{vx}{c^2} = \text{constant}.$$

In particular, the x'-axis ($t' = 0$ cosmic moment line), which passes through the point $x = 0$ at $t = 0$, has the equation

$$t = \frac{vx}{c^2} = vx$$

with the usual convention of $c = 1$.

In a similar way, recall what we mean by any line parallel to the t-axis; it is a line with a fixed space coordinate. Such a line is the world line of a stationary particle in the x, t frame, with the equation x = constant. Similarly, for the moving system we would write x' = constant as the equation of the world line of a particle stationary in that system. From the Lorentz transformation, this is equivalent to

$$x - vt = \text{constant}.$$

In particular, the t'-axis (which is the $x' = 0$ world line of a particle stationary at the origin of the moving system) passes through the $x = 0$, $t = 0$ point, and it has the equation

$$x = vt.$$

Thus, superimposed spacetime coordinate axes for the two frames look like those shown in Fig. 6.2. That is, the relative motion of the two frames results in a

---

[9] K. Svozil, "Time Paradoxes Reviewed," *Physics Letters A*, April 3, 1995, pp. 323–326.

[10] Science fiction, however, *has* (since the early days of pulp) enthusiastically embraced FTL human travel. In Larry Niven's story "At the Core" (*If*, November 1966), for example, we read of a manned spacecraft that travels 60,000 light years to the center of our own Milky Way Galaxy, and then back, at a speed 420,000 times that of light. As the pilot says (in a grand understatement), "That's goddam fast." Perhaps there is a way to make *some* sense of such an adventure, with the so-called *warp drive* (a'la *Star Trek*), which we'll take-up briefly at the end of this chapter.

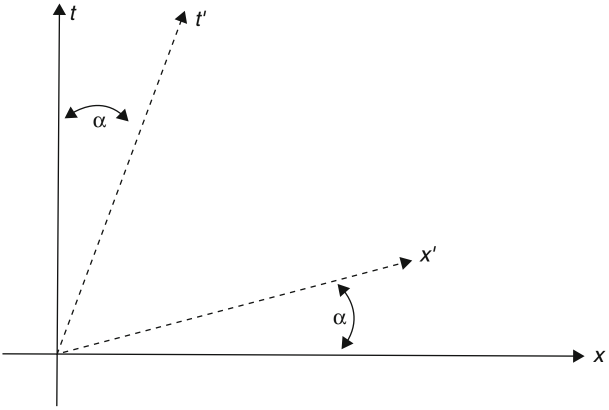

Fig. 6.2 Spacetime coordinate axes rotation by relative motion

rotation of the spacetime axes; but it is a strange sort of rotation, with *opposite* senses for the space and time axes. That is, the *x*' and *t*' axes rotate *towards* each other, as shown in the figure, to make equal angles (α) with the *x* and *t* axes, respectively, where

$$\alpha = tan^{-1}(v).$$

If we limit the moving frame (the frame of a moving observer) to subluminal speeds ($0 \leq v < 1$), then

$$0 \leq \alpha < 45°.$$

At the speed of light ($v = 1$) $\alpha = 45°$ and so the x' and *t*' axes coincide—time and space have become indistinguishable.

It is important to realize that observers in either system would measure the same speed for a photon; that is, each would see the world line of a photon as a line with slope 1. This view of the world line of a photon is literally built into the Lorentz transformation because one of Einstein's fundamental postulates for special relativity is the invariance of the speed of light. The truth of this statement for the *x, t* system is obvious from Fig. 6.2. It is, perhaps, not so obvious with the x', t' system because of the non-perpendicular axes (as shown in the figure) for that system. In Fig. 6.3 the world line of a photon is shown in both systems. In that figure we emit the photon at $x' = 0$, $t' = 0$, and we later measure its coordinates at point A to be $x' = x'_A$ at time $t' = t'_A$. Note carefully how this is done. We draw lines from point A parallel to the x' and t' axes until they intersect the t' and x' axes, respectively. This is similar to the way we would get the spacetime coordinates of A in the more familiar *x, t* system, where we would draw lines parallel to the *x* and *t* axes.

It should now be obvious that x'_A and t'_A have the same extension, just as they do in the unprimed system, so

Fig. 6.3 Invariance of the world line of a photon

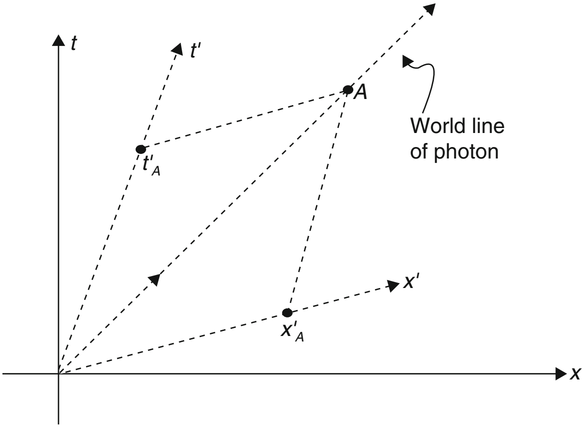

$$\frac{x'_A}{t'_A} = \text{the speed of light}$$

The speed of light is the only invariant speed under the Lorentz transformation. Indeed, the modern approach to special relativity emphasizes this invariance as the central property of the speed of light, rather than the idea of the speed of light being a limiting speed.

This geometrical interpretation of the Lorentz transformation lets us quickly make another interesting (one, I think, not at all obvious) observation: If a particle is faster than light in the x, t system, then there exists a subluminal x', t' system in which the particle is *infinitely* fast. Figure 6.4 shows the world line of an FTL particle in the x, t system (which is, of course, *below* the world line of a photon; that is, the particle's world line is spacelike). Suppose the FTL particle has speed $w > \mathrm{c}$ such that its world line makes angle β with the x-axis. If we now pick v, the speed of the moving x', t' system to be such that $\alpha = \beta$, then the x'-axis will coincide with the world line of the particle, and so the particle will appear to an observer in the x', t' system to be everywhere at once—that is, to be infinitely fast. We have, then,

$$\beta = tan^{-1}(v) = tan^{-1}\left(\frac{1}{w}\right)$$

or $v = 1/w$, which seems to be dimensionally wrong. Recall, however, that with our convention of $c = 1$, the v in this result is a *normalized* speed. To return to the units of everyday use, simply replace v with v/c and w with w/c; this transforms our result to

$$\frac{v}{c} = \frac{c}{w} \quad \text{or} \quad v = \frac{c^2}{w}\cdot$$

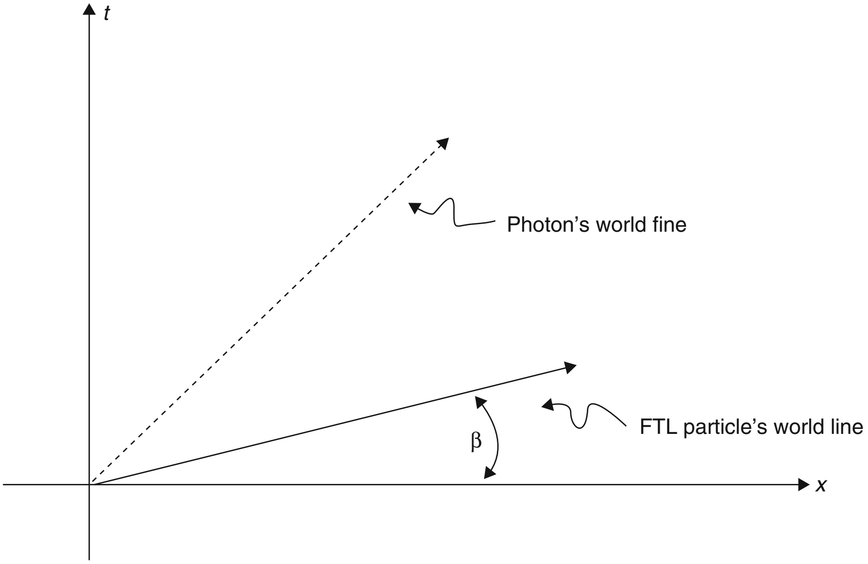

Fig. 6.4 World line of an FTL particle

We can, of course, turn this result around. If an FTL particle moves with speed w in the x, t frame, then to an observer in the x', t' frame moving with subluminal speed v, the particle will appear to be infinitely fast if $w = c^2/v$. A particle with $w > c^2/v$ is said to be not just be superluminal, but *ultra*luminal.

If a particle has infinite speed if $w = c^2/v$, then what physically happens if w is greater than c^2/v? The answer is easy to see from a spacetime diagram, as in Fig. 6.5, where the x' and t' axes have been extended back to negative values. In that figure I have labeled two arbitrary events A and B on the world line of an ultraluminal particle (and so it lies below the x' axis), and have shown the spacetime coordinates of each event in both the x, t and x', t' frames. For the x, t frame we see that A is related to B by the relations $x'_A < x'_B$ and $t'_B < t'_A$; that is, the time order of A and B is reversed for an observer in the x', t' frame. To that observer, the particle appears to be traveling backward in time!

But this isn't quite the end of the story. Following the approach of two pioneering tachyon physicists,[11] we note that if the energy of a particle in the stationary system is E, then the energy as measured in the moving system is given by[12]

---

[11] O. M. Bilaniuk and E. C. G. Sudarshan, "Causality and Space-like Signals," *Nature*, July 26, 1969, pp. 386–387.

[12] The expression for E' is the result of applying the Lorentz transformation to E. You can find all the details of that worked out in A. P. French, *Special Relativity*, W. W. Norton 1968, pp. 208–210.

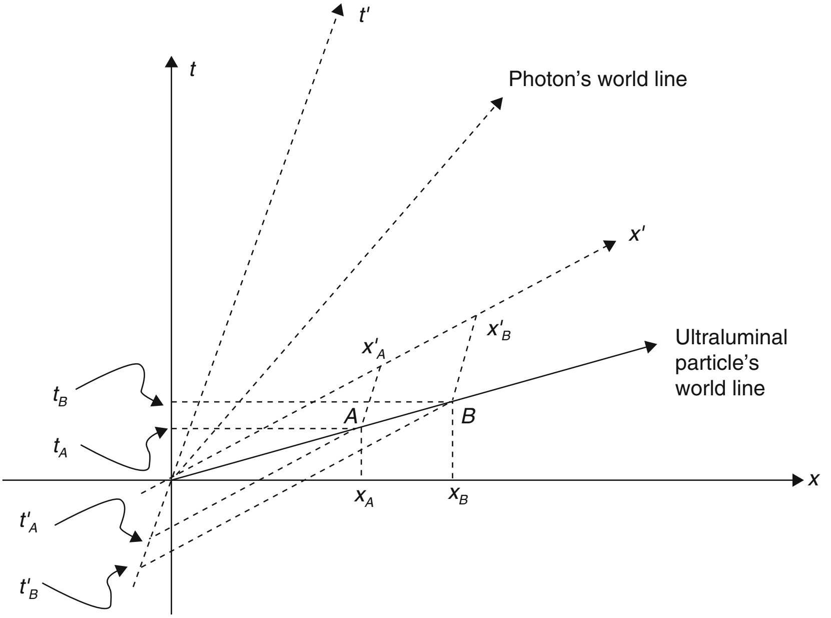

Fig. 6.5 World line of an *ultra*luminal particle

$$E' = \frac{1}{\sqrt{1-\left(\frac{v}{c}\right)^2}}\,\frac{\mu(c^2 - \mathrm{wv})}{\sqrt{(w/c)^2 - 1}}$$

Note that sign of E' switches from positive to negative when w (the speed of the particle) exceeds c^2/v, which is precisely the condition for the particle to be ultraluminal and so traveling backward in time (as seen in the primed system). That is, negative energy moving backward in time in one system is positive energy moving forward in time in another. This is, in fact, the original motivation for the RP (reinterpretation principle) discussed in the previous chapter, and the claim was, at one time, that the RP was just what was needed to 'explain' the paradoxical implications of time travel. This is not the majority view today, however, and a number of physicists have been quite inventive in constructing scenarios with causal paradoxes that the RP clearly fails to 'explain.'

For example, the Princeton physicist Shoichi Yoshikawa (1934–2010), in a letter to *Physics Today* (see note 3), was able to create a scenario which *uses* the RP to arrive at a causal paradox. In his construction, the RP allows an observer to transmit an ultraluminal tachyon to a remote observer at time $t = 0$, and to receive a reply from that observer at $t < 0$. This obviously sets-up the possibility of a bilking paradox in which the original observer, upon receiving the $t < 0$ tachyon, then decides *not* to transmit the $t = 0$ tachyon. What makes Yoshikawa's paradox a particularly troublesome paradox is that the RP is the culprit, not the savior.

6.2 Tipler's Rotating Cylinder Time Machine

> "... within forty-eight hours we had invented, designed, and assembled a chronomobile. I won't weary you with the details, save to remark that it operated by transposing the seventh and eleventh dimensions in a hole in space, thus creating an inverse ether-vortex and standing the space-time continuum on its head."[13]
>
> —this is almost surely *not* the way to build a TM (time machine)

As discussed at the start of this book, the first endorsement of the reasonableness of physicists talking about plausible, scientific time travel, began with Gödel's discovery of closed timelike lines in the mathematics of certain rotating universe models. Such models had been studied as early as 1924 by the Hungarian physicist Cornelius Lanczos (1893–1974), a quarter century before Gödel, but it was Gödel who made explicit the possibility a rotating universe might allow time travel. His realization of time travel as an *inherent* property of a rotating universe is an illustration of a *weak* TM, while what is the central interest concerning time machines (in physics *and* in science fiction) is a *strong* TM. That is, in time machines that can be *intentionally constructed* by manipulating mass-energy in a finite (what physicists call a *compact*) region of spacetime to *create* closed timelike lines where none existed before the manipulation began. Interestingly, the fundamental physical idea behind Gödel's weak TM is the same underlying idea behind the first strong TM, Tipler's rotating cylinder discussed back in Chap. 1 (strictly, of course, only a *compact* TM if the cylinder can be of *finite* length).

The one result from general relativity that we'll use here (without proof) is that the rotation of matter causes a distortion of spacetime that results in the 'tipping over' of light cones, with the future half tilted in the direction of motion. If you imagine a point in the universe about which the rotation takes place, then this tipping effect increases with the radial distance from that point.[14] The fact that rotating masses tip light cones over in the direction of rotation was discovered very early in the history of general relativity (1918), by the Austrian theoreticians Josef Lense (1890–1985) and Hans Thirring (1888–1976). Originally (and naturally) called the *Lense-Thirring effect*, it now generally goes by the name of the *dragging of inertial frames effect*, and it plays a central role in the weak Gödel and the strong (maybe) Tipler time machines. Here's how.

At a certain critical distance from the rotation center (more on this in just a bit), the future half of the light cone at a given point in spacetime will be sufficiently tilted so as to enter the past half of similarly tilted light cones at nearby spacetime points This is illustrated in Fig. 6.6,[15] which shows a circular chain of tilted light

---

[13] L. Sprague de Camp, "Some Curious Effects of Time Travel," in *Analog Readers' Choice*, Dial 1981.

[14] For a picture of this, see S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Spacetime*, Cambridge University Press, 1973, p. 169.

[15] Adapted from D. B. Malament, "'Time Travel' in the Gödel Universe," *Proceedings of the Philosophy of Science Association* 1984, pp. 91–100.

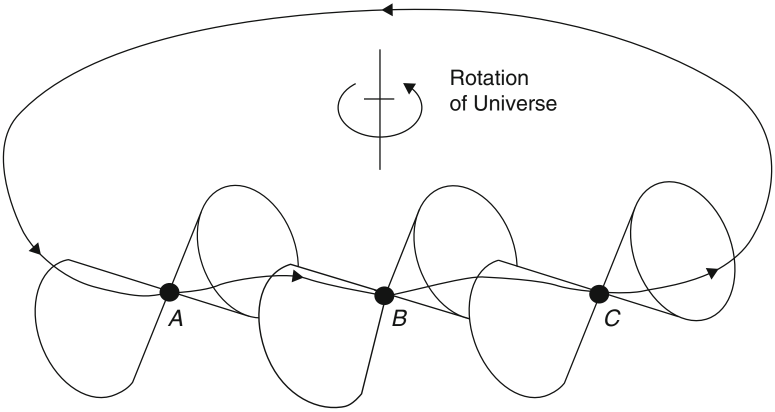

Fig. 6.6 Time traveling through tilted light cones in a rotating universe

cones in a rotating universe. Because light cones are tilted by a rotation-induced twisting of spacetime, a traveler can move around a circular path on a trip that is always directed into his *local* future but nonetheless end-up in his own *global* past, without ever going faster than light. This kind of round trip in spacetime, with a trajectory that winds back into the past without ever becoming spacelike, is a closed timelike curve.

From Fig. 6.6 it should be clear how a time traveler, beginning at A, can weave his way along a circular path (this path needs to have a radius at least as great as the critical value mentioned earlier) to B to C to . . . that brings him back into the *past* half of the light cone at A. The traveler's world line is always inside the local light cone; that is, the world line of the time traveler is always timelike, and never FTL. These timelike curves are present in the rotating spacetime from the very beginning of the spacetime, and were not *created* (certainly not by humans) by conscious intent.

You can see how this works mathematically by taking the spacetime metric for Gödel's universe (with, as usual, the convention that the speed of light $c = 1$):

$$(ds)^2 = (dt)^2 - (dr)^2 - (dy)^2 + \sin h^2(r)\left[\sin h^2(r) - 1\right](d\phi)^2 \\ + \sqrt{2}\sin h^2(r)(d\phi)(dt)$$

where t, r, y, and ϕ are cylindrical coordinates in four-dimensional spacetime. Now, imagine that our adventurer's world line is the helical curve r=constant, $y = 0$, and $t = -\alpha\phi$: if we take the time axis as vertical then the time traveler's world line is a vertical helix in spacetime. For this curve, $dr = \mathrm{dy} = 0$ and $dt = -\alpha d\phi$. This last differential means, in particular, that whatever the sign of the constant α, we can choose that one of the two senses of movement in the spatial ϕ dimension that gives $dt < 0$.

Continuing, we have

$$(ds)^2 = \left[\alpha^2 - 2\sqrt{2}\alpha sinh^2(r) + \sin h^2(r)\left(\sin h^2(r) - 1\right)\right](d\phi)^2$$

or, upon letting $u = \sinh(r)$ we have

$$(ds)^2 = \left[\alpha^2 - 2\sqrt{2}\alpha u^2 + u^2\left(u^2 - 1\right)\right](d\phi)^2.$$

Now, for $\alpha = 0$ we have

$$(ds)^2 = u^2\left(u^2 - 1\right)(d\phi)^2$$

which is greater than zero if $u > 1$. This condition holds if $\sinh(r) > 1$, that is, if r is a constant greater than $\ln\left(1 + \sqrt{2}\right)$. In other words, for r sufficiently large (and now we know the critical value for r) we have $(ds)^2 > 0$, the required condition for a timelike spacetime interval. By continuity, then, we will continue to have $(ds)^2 > 0$ even with some small positive or negative value of α different from zero. Because ϕ is a periodic coordinate (we identify $\phi = 0$ with $\phi = 2\pi$), as the traveler moves along the curve she returns repeatedly to the same spatial points, but her time coordinate is increasingly negative. That is, she is traveling into the past. Note, once again, that in Gödel's universe this property holds only for orbits with radii greater than a certain minimum.[16]

Tipler's rotating, infinitely long cylinder is a mechanism for artificially producing the tipped-over light cone effect, thus *creating* closed timelike curves. Figure 6.7, taken from Tipler's Ph.D. dissertation,[17] shows how the cylinder works. The cylinder is represented by the central vertical axis. Far away from the cylinder the light cones in spacetime are upright, but as we move inward they tip over, the future halves opening up into the direction of rotation. (Only the future halves of the light cones are shown.) This direction, which is the direction that far away from the cylinder measures *space*, near the cylinder measures *time* (just as in the Gödelian universe). That is, there has been a dimension reversal! This fantastic possibility has found its way into science fiction, as in Stephen Baxter's 1995 novel *The Time Ships*, wherein the Victorian narrator says "If only one could *twist about* the Four Dimensions of Space and Time — transposing Length with Duration, say — then one could stroll through the corridors of History as easily as taking a cab in the West End!"

To travel back in time, therefore, all the time traveler need do is leave Earth and approach the cylinder until she is near enough to be in the tipped-over region of spacetime. Then she would follow a helical path around the cylinder and could

---

[16] There are other solutions to the Einstein gravitational field equations that have closed timelike lines at *any* radius, no matter how small: see M. J. Reboucas, "A Rotating Universe with Violation of Causality," *Physics Letters A*, March 5, 1979, pp. 161–163.

[17] F. J. Tipler, *Causality Violation in General Relativity*, University of Maryland 1976.

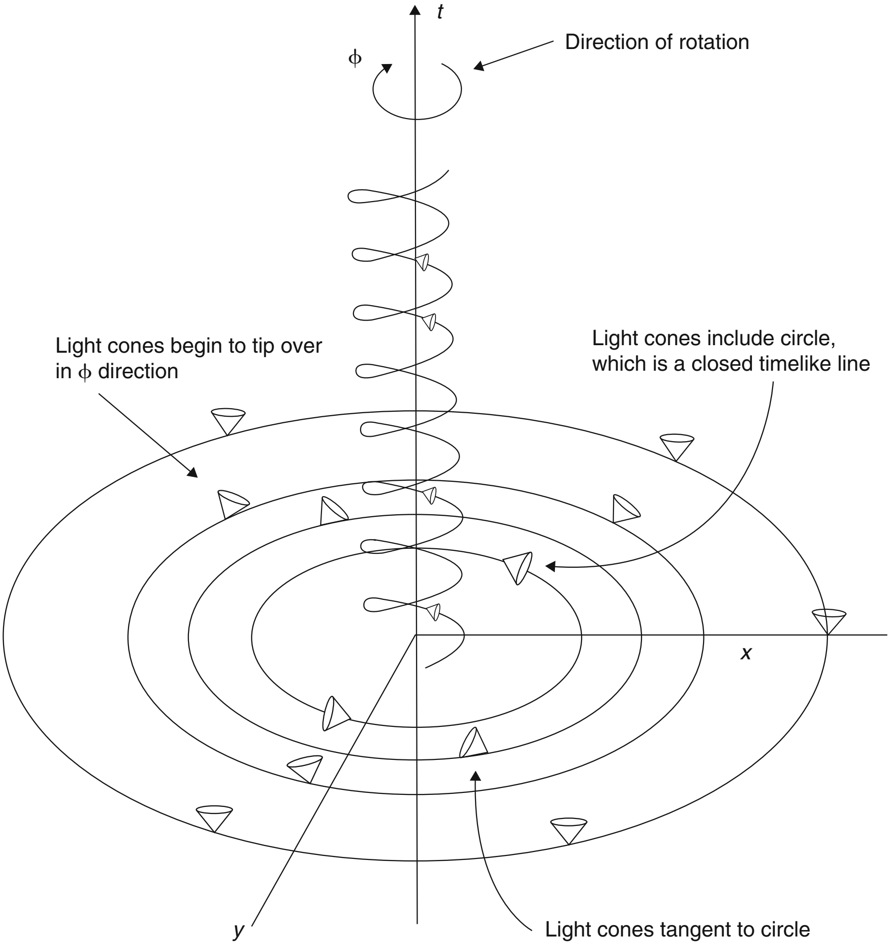
t
ϕ
Direction of rotation
Light cones begin to tip over in ϕ direction
Light cones include circle, which is a closed timelike line
x
y
Light cones tangent to circle

Fig. 6.7 The future halves of light cones point almost entirely in the $+t$ direction far from rotating matter; they begin to tip over as the matter is approached. Note that there is a helical timelike path that moves locally into the *future* in the $-t$ direction, that is, it goes into the past as seen by an observer far from the rotating matter (The world lines of rotating matter are helixes in the $+t$ direction)

spiral along the negative time direction as far back in time as desired (but no further back than to the moment of the cylinder's creation). This motion is such that the time traveler is always moving into her local future, via the tipped-over light cones. Finally, she would withdraw from the cylinder and return to Earth—in the past. The time traveler had better be a good space navigator, of course, because Earth won't be where she left it!

6.3 Thorne's Wormhole Time Machine

> "This fact reinforces the authors' feeling that [closed time loops] are not so nasty as people generally have assumed."[18]

The spacetime wormhole is presently the most promising of the approaches that have been advanced for building a time machine. Gödel rotated the entire universe in 1949, and Tipler 'reduced' the problem in 1974 to 'merely' spinning a cylinder of infinite length. In 1988 Kip Thorne scaled things down even more, this time to other extreme. His idea calls for pulling a wormhole on the scale of the Planck length out of the topologically multiply connected quantum foam that spacetime is and then enlarging it (somehow) to human scale, all the while stabilizing it against self-collapse, and then finally using the time dilation effect of special relativity to alter time at one mouth of the wormhole as compared to the other mouth. What a mouthful! What, you almost surely wonder, does all that mean?

First of all, what's 'quantum foam'? The term refers to the idea that the topology of spacetime is *not* a smooth, continuous manifold, but rather (if you look close enough) is a seething 'ocean of fluctuations' that is always changing, changes on the scale of the Planck length (look back at Sect. 1.5). What does *that* mean? Like the ocean surface-in-the-large, large-scale spacetime is simply connected. But just as one sees all sorts of transient structure as one looks at the water more closely (beginning with macroscopic waves and then proceeding downward to the bubbly foam on the waves), spacetime too displays an ever-changing connectivity-in-the-small.[19] That is the 'quantum foam.'

Wormholes have been around in physics for decades, but they have always been thought to be so unstable as to exist only on paper in the mathematics of general relativity. In an analysis[20] published more than half-a-century ago, for example, wormhole instability was shown to be so severe that not only would a human have no chance in getting through one, but also not even a single speedy photon could do so. Even at the speed of light, the photon could not zip through a wormhole before being trapped inside ("pinched off") in a region of infinite spacetime curvature. Wormholes would simply collapse too quickly after formation for even the so-called ultimate speed to save the traveler. Indeed, the presence of mass-energy inside a wormhole actually *accelerates* its collapse. The physics of wormholes, it seemed, made them simply untraversable.

---

[18] The conclusion of a mathematical demonstration that time travel by wormhole does not conflict with the conservation of energy: see J. L. Friedman *et al*., "Cauchy Problem in Spacetimes with Closed Timelike Curves," *Physical Review*, September 15, 1990, pp. 1915–1930.

[19] Not all accept this view. See, for example, A. Anderson and B DeWitt, "Does the Topology of Space Fluctuate?" *Foundations of Physics*, February 1986, pp. 91–105.

[20] R. W. Fuller and J. A. Wheeler, "Causality and Multiple Connected Spacetime," *Physical Review*, October 15, 1962, pp. 919–929.

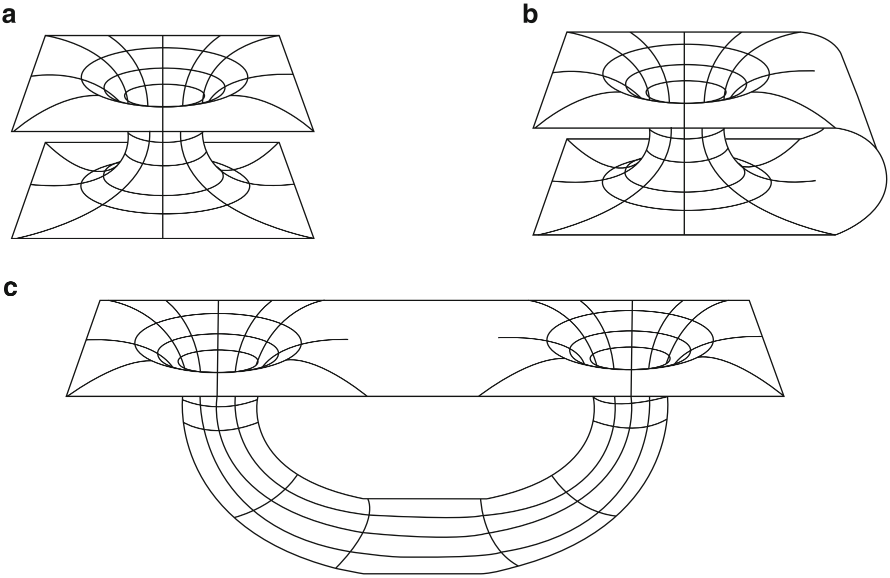

Fig. 6.8 These sketches are unavoidably misleading, being two-dimensional renditions of wormholes that connect two places in a three-dimensional space. Time machine wormholes, on the other hand, connect two places in a four-dimensional spacetime. In particular, the mouths of the wormholes are not 'depressions' into which the time traveler's rocket ship plunges, but rather would appear to be three-dimensional spheres. The wormhole in (**a**) connects two disjoint universes, while those in (**b**) and (**c**) are connections between two places in the same universe. As shown in these last two cases, the wormhole 'handle' can be either long or short, compared to the external distance between the wormhole mouths

And anyway, how would one gain access to a wormhole in the first place? As Thorne and one of his students suggested,[21] one might perhaps imagine someday finding a rotating black hole that mathematically possesses, in its interior, so-called hyperspace tunnels to 'other places'—either in our universe or in other universes (see Fig. 6.8). In the case of a wormhole connecting two places in the same universe, although the external distance in spacetime between the places may be very large (mega-light-years), it is conceivable that the distance *through* the wormhole itself could be very small. The time required to traverse the wormhole, as measured by the traveler's watch might, in fact, be arbitrarily small. This is exciting, but wormholes do not come without some significant problems.

Such problems include the presence of a one-way event horizon, which precludes two-way travel (it seems reasonable to assume that time and space travelers might wish to eventually return), and enormous gravitational gradients (tidal forces) that dismember anything approaching and/or entering the wormhole. These problems have often been conveniently *ignored* in science fiction. For example, in one tale[22] the time machine is in the form of a wormhole time tunnel that time travelers can simply walk through, and in another characters similarly walk back and forth between two openings in a wall that connect the same place in space but separated in time by 160 years.[23] And in an even older story, we read of a time machine (in the form of a hole in spacetime that is a wormhole in everything but name) through which one can literally step from future to past and back again. That story[24] is particularly notable for having introduced the term *mugwump* for a time traveler who uses a time machine wormhole. To paraphrase the tale's time traveler, as he transits the wormhole his "mug" is in the past and his "wump" is in the future.[25]

---

[21] M. S. Morris and K. S. Thorne, "Wormholes in Spacetime and Their Use for Interstellar Travel: A Tool for Teaching General Relativity," *American Journal of Physics*, May 1988, pp. 395–412. This paper was motivated by Thorne's earlier response to a request, from the American astronomer Carl Sagan (1934–1996), for help in making plausible the interstellar travel imagined in his 1985 novel *Contact*.

Now, before going any further, a note on what physicists are referring to when they write of traversable time travel wormholes. Such wormholes are called *Lorentzian* because they have a spacetime metric with the signature $[+, -, -, -]$ (see Sect. 3.5). Further, the wormhole is taken to be *static*, that is, to have no time-varying behavior. The reason for being specific on the nature of time travel wormholes is that there is another type with what is called the *Euclidean signature*: $[+, +, +, +]$, which is not suitable for time traveling. Motion in a Euclidean signature wormhole involves imaginary momentum or proper time, neither of which is physically plausible for a time traveler. However, *if* you can gain access to a Lorentzian wormhole *then*, as two physicists wrote, "if you manage to acquire even one inter-universe traversable wormhole then it *seems* almost absurdly easy to build a time machine."[26]

In response to that fundamental question of how to gain access to a Lorentzian wormhole, Thorne and his students were bluntly honest—they didn't know. Their best suggestion was that "one can imagine an advanced civilization pulling [a] wormhole out of the quantum foam and enlarging it to classical size."[27] A few years later this dramatic idea found its way into a science fiction novel of a far-future alien civilization able to control the energies of constellations of galaxies: "Spacetime is friable. Wormholes riddle the fabric of spacetime on all scales. At the Planck length and below, wormholes arising from quantum uncertainty effects blur the clean Einsteinian lines of spacetime. And some of the wormholes expand to human scale, and beyond — sometimes spontaneously, and sometimes at the instigation of intelligence."[28]

---

[22] R. C. Wilson, *A Bridge of Years*, Doubleday 1991.

[23] M. Leinster, *Time Tunnel*, Pyramid Books 1964.

[24] H. Kuttner, "Shock," *Astounding Science Fiction*, March 1943.

[25] The term's origin is in 19th century politics, as a description of fence-sitters who try to avoid taking a definite position on some controversial topic under debate. For an illustration of wormhole mugwumping, see K. S. Thorne, *Black Holes & Time Warps*, W. W. Norton 1994, p. 500.

[26] C. Barceló and M. Visser, "Twilight for the Energy Conditions?" *International Journal of Modern Physics D*, December 2002, pp. 1553–1560. See also M. Visser, *Lorentzian Wormholes*, AIP Press 1996.

[27] M. S. Morris, K. S. Thorne, and U. Yurtsever, "Wormholes, Time Machines, and the Weak Energy Condition," *Physical Review Letters*, September 26, 1988, pp. 1446–1449.

This isn't easy to visualize, but let's plunge ahead and assume we can "pull" a wormhole out of the quantum foam. If so, then once it is (somehow) inflated the wormhole could be stabilized against collapse by threading it with either matter or fields of stupendous negative (outward) tension—and by *stupendous* I mean *STUPENDOUS*. If b_0 denotes the minimum radius of the wormholes (the size of the so-called *throat* of the wormhole), the tension (radial pressure) at that location must be at least

$$\tau_0 = \frac{3.8}{{b_0}^2} \times 10^{36} \frac{tons}{in^2}$$

where b_0 is expressed in feet.[29] (This expression shows that for a wormhole with a throat radius of several thousand feet, the value of τ_0 is enormous, of the same magnitude as the pressure at the center of the most massive neutron star.) To stabilize a common sort of everyday wormhole, such as a subway tunnel, we can obtain the required tension/pressure by lining the tunnel with iron plates or concrete. But how, for a hyperspace wormhole, do we obtain 'iron plates' that can achieve the required enormous tension? As Thorne and his students observed (note 27) such stuff could only be called "exotic," a term that had appeared a few years earlier in connection with the observed energy density in the throat of a wormhole.[30]

One possible approach to this problem does not use matter at all. If we make b_0 *very* large, then non-material fields will do the job.[31] Indeed, suppose b_0 equals 1 light year (a large wormhole by anybody's standard!). Then τ_0 is 'only' 4000 tons/ square inch, and that is achievable by threading the wormhole throat with a magnetic field of 'only' 2,700,000 gauss (five million times stronger than the Earth's field).[32] To generate such a field is not impossible, and present-day experimental, hypervelocity electromagnetic rail guns in development for the US Navy use transient magnetic fields in the mega-gauss range.

---

[28] From S. Baxter's 1993 novel *Timelike Infinity*. That same year the Chinese physicist Liao Liu wrote on how, as a result of a naturally occurring vacuum fluctuation, a wormhole might spontaneously appear: L. Liu, "Wormhole Created from Vacuum Fluctuation," *Physical Review D*, September 15, 1993, R5463–R5464.

[29] The original equation for τ_0 was given in units of dynes/cm^2, with b_0 is expressed in meters (see note 21), but I have converted these units to the more familiar (to the non-physicist) units of tons per square inch, for dramatic purposes, because the units of dynes per square centimeter is so small it's difficult to relate it to anything of everyday significance.

[30] R. Balbinot, "Crossing the Einstein-Rosen Bridge," *Lettere Al Nuovo Cimento*, May 16, 1985, pp. 76–80.

[31] See, for example, Y. Soen and A. Ori, "Improved Time Machine Model," *Physical Review D*, October 15, 1996, pp. 4858–4861, and D. N. Vollick, "How to Produce Exotic Matter Using Classical Fields," *Physical Review D*, October 15, 1997, pp. 4720–4723.

[32] To understand how the calculation of a magnetic field from a pressure requirement is accomplished, note that pressure is dimensionally equivalent to field energy per unit volume, which in turn is given by a well-known result in electromagnetic theory.

As if the stabilization problem wasn't enough of a complication, Thorne and his colleagues (note 27) showed that there is another, even more curious problem. The geometrical requirement that the wormhole interior smoothly connect to the external, asymptotically flat exterior spacetime demands that the wormhole throat flare outward as shown in Fig. 6.8. It turns out that this condition is mathematically equivalent to a requirement that τ_0 exceed the energy density of the throat material[33]; and special relativity, in turn, says that for some timelike observers the energy density will then actually be *negative*. This is a clear violation of the so-called *weak energy condition* (WEC), which says the observed mass-energy density is always non-negative. This is *so* 'obvious' that the WEC was thought, for a long time, to be almost a law of nature. Such a violation is actually not as crazy as it might sound however because, more than half-a-century ago, it was shown[34] that an energy density that is everywhere and everywhen positive is *not* compatible with any quantum field theory that is *local*, as, presumably, will be the yet-to-be-discovered theory of quantum gravity. Over the years since then a variety of other energy conditions have been proposed, such as the *averaged weak energy condition* (AWEC), which says that only the *average* value of the energy density over a complete null geodesic world line has to be non-negative, which leaves open the possibility of *temporary* negativity here, there, then and when.[35]

The traversable, static wormholes studied by Thorne and his colleagues violate even the AWEC (see note 27), however, and in the years since it has become clear to physicists that imposing constraints on the mass-energy density may not be so 'obvious' after all (see note 26). But all may not be lost. Indeed, it has been shown that the violation of the AWEC by traversable wormholes can be made as small as desired, that is, the requirement for the exotic matter required to line a wormhole throat to keep it open can be made as tiny as you want.[36]

Well, tiny the quantity of exotic matter may be but, nonetheless, even a tiny amount of it would be extraordinary weird stuff because a negative energy density can be interpreted as meaning that the exotic material that keeps the wormhole throat open does so by exerting a *repulsive* gravitational force.[37] A repulsive force sounds like a property we'd expect to see associated with *negative mass* and, although such a thing has never been observed (negative matter is *not* anti-matter, which *has* been observed and which does not repel 'normal matter'), it was studied long ago (theoretically, of course) by the English cosmologist Hermann Bondi (1919–2005). Bondi showed[38] that negative mass would indeed have some truly bizarre properties,[39] but there is nothing in general relativity that forbids its possible existence. Wormholes, with negative mass throats, should produce observable effects by which a wormhole *might* be detected. Mathematical analyses of the effect a negative-mass wormhole mouth would have, when crossing the line-of-sight between Earth and a distant star, indicates that there should be an observable *double-spike* in the intensity of the star's light. Astronomical searches for such an optical signature have actually been conducted, with (alas) no success as I write (2017).

---

[33] The condition of τ_0 exceeding the energy density in the throat is, in fact, the technical definition of *exotic*—see note 30. In everyday situations, the exotic condition is never even remotely approached. For example, the maximum tension necessary to pull a piece of steel apart—the so-called *tensile strength*, about 100,000 pounds per square inch—is a trillion times less than the mass-energy density of steel.

[34] H. Epstein, V. Glaser, and A. Jaffe, "Nonpositivity of the Energy Density in Quantized Field Theories," *Il Nuovo Cimento*, April 1, 1965, pp. 1016–1022.

[35] T. A. Roman, "Quantum Stress-Energy Tensors and the Weak Energy Condition," *Physical Review D*, June 15, 1986, pp. 3526–3533.

[36] M. Visser, S. Kar, and N. Dadhich, "Traversable Wormholes with Arbitrarily Small Energy Condition Violations," *Physical Review Letters*, May 23, 2003, pp. 201102-1 to 201102-4.

There is another interesting implication of a repulsive gravitational force, one that proves to be *essential* to the possibility of a wormhole time machine. Just as Einstein's famous prediction (verified in 1919) from general relativity, that star light passing near the Sun's edge is bent *inward* by the Sun's attractive gravitational field, the repulsive, anti-gravity field of a wormhole will cause any light rays traveling through the wormhole to be bent *outward*. That is, a tight, narrow beam of radiation entering a wormhole will emerge *defocused*. This is crucial because, as you'll soon see, a wormhole time machine would otherwise be destroyed by the light from the dimmest candle.

One might take the failure of astronomical searches for a double-spike light signature to mean that wormholes with negative mass throats (thus violating the WEC) simply don't exist. But not so fast. The first hint that the possibility of a negative energy density might not be such a crazy idea occurred as long ago as 1948, with a theoretical prediction made by the Dutch physicist Hendrick Casimir (1909–2000). As pointed out in Chap. 1, the Heisenberg uncertainty principle allows a temporary violation of conservation of energy to occur, with the magnitude of the allowed violation increasing with decreasing time duration. Even in a vacuum, then, with particle/anti-particle creation and annihilation spontaneously and continuously taking place, the *average* energy density being zero does not preclude fluctuations away from zero and so, at times, actually becoming *negative*. What Casimir showed was that if one positioned two perfectly conductive plates parallel to each other, then the normal quantum fluctuations of the energy density in this 'vacuum sandwich' would be altered in such a way as to result in their mutual attraction—and this (tiny) effect was later actually observed.[40]

---

[37] The *strong energy condition* says gravity is always attractive—which is clearly not true in a wormhole throat—and so static, traversable wormholes violate both the weak *and* the strong energy conditions.

[38] H. Bondi, "Negative Mass in General Relativity," *Reviews of Modern Physics*, July 1957, pp. 423–428.

[39] For example, general relativity says that a negative mass will *repel* all other masses (positive *and* negative), whereas a positive mass will *attract* all other masses (positive *and* negative). Imagine, then, a negative mass attached to the nose of a positive-mass spaceship. The spaceship tries to move toward the negative mass, while the negative mass tries to move away from the spaceship. So off they both go into the sky, like a cat chasing its tail. This so-called *reactionless anti-gravity drive*, bizarre as it appears, does not violate either of the conservation laws of energy or momentum. See G. Cavalleri and E. Tonni, "Negative Masses, Even if Isolated, Imply Self-Acceleration, Hence a Catastrophic World," *Il Nuovo Cimento B*, July 1997, pp. 897–903.

What does it mean to 'alter the normal quantum fluctuations'? Consider the creation of a photon and its anti-particle, which is another photon. From the wave interpretation of particles, the parallel plates restrict the photons that appear in the vacuum layer to those that have wavelengths that 'fit' because those wavelengths are submultiples of the plate separation (this requirement follows from the fact that a perfectly conducting plate cannot support a non-zero tangential electric field). Photons with longer wavelengths than the plate separation cannot 'fit' and thus do not appear. That is, the parallel plates have created a boundary condition that has quantized the electromagnetic field. The absence of these 'longer wavelength' photons lowers the average energy density between the plates and, because the average without the plates is zero, the altered average energy density must be negative. Indeed, the more the maximum allowed photon wavelength decreases with decreasing plate separation, the more negative the average energy density becomes in the enclosed Casimir vacuum. The negative energy density manifests itself as an inward directed force per unit area (remember, energy density and pressure are dimensionally equivalent).

The experimental detection of the Casimir effect was a remarkable event in physics. As one mathematician put it, "No worker in the field of overlap of quantum theory and general relativity can fail to point this fact out in tones of awe and reverence."[41] Robert Forward, an imaginative physicist who has appeared in this book earlier as an enthusiastic supporter of time travel, has described how the Casimir force might be used to extract energy literally *from a vacuum*. This is an idea as seemingly impossible as is the plan of one science fiction professor to squeeze energy out of *time*. As he asks his assistant, "But tell me, Bob, isn't that a ridiculous thought? To take time, something intangible, invisible, incomprehensible, and contract it — squeeze it together like a sponge?"[42] The story is fun, but Forward's proposal is that as well—*and* good physics.[43] Now, what does all this have to do with wormhole time machines?

---

[40] A complete presentation of Casimir's analysis, with citations to the original literature, can be found in L. E. Ballentine, *Quantum Mechanics*, Prentice-Hall 1990, pp. 399–403. For an historical, tutorial presentation, including Casimir's personal comments on how he was led to make his discovery, see P. W. Milonni and M.-L. Shih, "Casimir Forces," *Contemporary Physics*, September–October 1992, pp. 313–322.

[41] S. A. Fulling, *Aspects of Quantum Field Theory in Curved Space-Time*, London Mathematical Society 1989.

[42] E. Binder, "The Time Contractor," *Astounding Stories*, December 1937.

[43] R. Forward, "Extracting Electrical Energy from the Vacuum by Cohesion of Charged Foliated Conductors," *Physical Review B*, August 15, 1984, pp. 1700–1702.

Thorne and his colleagues (note 27) proposed to use the Casimir effect to achieve the "exotic condition" *without* matter. Their idea was to place identical, conducting, spherical plates that carry equal electric charges at each end of the wormhole (remember, the wormhole mouths are spherically symmetric). The two identical charges repel each other, but the charge size is adjusted so that the gravitational attraction of the plates precisely cancels the repulsion. They then calculated that the Casimir effect results in a negative energy density sufficient to provide the throat tension necessary to prevent wormhole collapse.

There *are* some weird aspects to this (and perhaps that's no surprise). For example, the analysis assumed that the wormhole length is *very* small compared to its radius (10^{-10} cm long and 200 million miles wide!), with the short length required because it represents the separation of the wormhole plates, and the smaller the separation the more negative the average energy density. (The functional dependence is as the *fourth* power of the separation.) Another problem is the balancing of the electrical repulsion and the gravitational attraction of the wormhole mouth plates, as such a balance is clearly an unstable one. Finally, because the two spherical plates completely fill the wormhole mouths, how would a traveler actually get through the wormhole? The 'answer' was to drill a hole through the plates and hope that wouldn't perturb the Casimir vacuum too much.[44]

All the above litany of the difficulties static, traversable wormholes face in simply existing is certainly daunting, but let's now ignore all that and suppose we actually have a wormhole with both mouths in the same universe. (For use as a time machine, it would seem desirable for the time traveler to remain in his/her own universe!) So, how do we turn the wormhole into a time machine? Interestingly, while it is general relativity that gives us the wormhole, it is special relativity that adds the final touch of backward time travel. We begin by imagining that, somehow, one mouth of the wormhole can be moved with respect to the other mouth. One early suggestion, for example, was to use the gravitational attraction of a large asteroid to 'drag' one end of the wormhole, thereby inducing a time dilation effect.[45]

That is, suppose we have two clocks A and B, one in each mouth of the wormhole. These two clocks, and all other clocks in the flat spacetime outside the wormhole, are initially indicating the same time and running at the same rate. Now, recalling the twin paradox from Chap. 3, let each mouth-clock play the role of one of the twins. Imagine that A and B are now separated because the mouth containing B is placed on board a rocket ship. The rocket ship takes a long, high-speed trip out into space along the straight-line path joining A and B in external space, and then returns, just as described in Sect. 3.5. We then unload the space traveling wormhole mouth (with its clock B) and reposition it at its original location. What is the situation now? We can summarize matters as follows: (1) Clock A, in the non-moving mouth, remains in-step with the local clocks in the space outside the mouth. (2) Clocks A and B, both inside the wormhole, have *not* moved with respect to each other because we are assuming a very short wormhole handle, as in part (b) of Fig. 6.8. We can arrange for the motion of the space traveling mouth (with clock B) to be such that the handle is always short, and so the distance between clocks A and B changes by an arbitrarily small amount. Thus, clocks A and B remain in-step with each other. (3) Clock B, because it has been moving with respect to its external space, arrives back at its starting position reading *behind* (that is, *earlier*) than the clocks outside its wormhole mouth.

---

[44] The very next year it was shown how to construct non-spherically symmetric wormholes to avoid that particular problem: see the two papers by M. Visser, "Traversable Wormholes: Some Simple Examples," *Physical Review D*, May 15, 1989, pp. 3182–3184, and "Traversable Wormholes from Surgically Modified Schwarzchild Spacetimes," *Nuclear Physics B*, December 11, 1989, pp. 203–212.

[45] J. L. Friedman, "Back to the Future," *Nature*, November 24, 1988, pp. 305–306.

For the sake of argument, then, suppose the journey of B is such that there is a two-hour time-slip between clock B and its local, external clocks. Thus, if clock B reads 9 A.M., the clocks outside of mouth B will read 11 A.M. But because clocks A and B are in-step, clock A reads 9 A.M., as do the clocks outside of mouth A. That is, the wormhole connecting mouth A to mouth B is a connection between two parts of the same universe that are two hours apart in time. Now, suppose the journey from mouth A to mouth B can be made through external space in one hour. Then, one could leave mouth A at 10 A.M., rocket to mouth B by 11 A.M., and travel back to mouth A via the wormhole to the starting point—where it is 9 A.M., one hour *before* the trip began! We could, in fact, imagine repeating this process, going back one additional hour for each new loop through the wormhole. One clear restriction, however, is that we could not go back in time to *before* the creation of the wormhole time machine. The wormhole works in the other direction, too. To see this, suppose that the space traveler leaves mouth B at 8 A.M. and rockets through external space to mouth A, arriving at 9 A.M. Entering mouth A, he exits from mouth B (where he started) at 11 A.M., two hours in the future.

Another way to induce a time dilation effect, to convert a wormhole to a time machine, *without moving either mouth*, is to simply place one mouth in an intense gravitational field, that of, say, a neutron star. (Recall, from Sect. 3.3, how gravity influences the time-keeping rate of a clock.) As the physicists who proposed this idea put it, almost *any* interaction with surrounding matter and gravity fields almost inevitably turns a wormhole into a time machine.[46] Others have admitted that the details of the origin of time dilation are probably not issues worth debating, but rather what is called the *back-reaction* is of far more concern.

To understand the back-reaction requires mention of what is called the *Cauchy horizon*, the hyperspace surface in spacetime that separates the region where closed timelike lines can exist, from the region where they cannot exist. The back-reaction is the build-up of unbounded energy levels on the Cauchy horizon, causing its instability and rapid destruction. The name of the horizon comes from the "Cauchy problem"—named after the nineteenth century French mathematician Augustin-Louis Cauchy (1789–1857)—in the theory of partial differential equations. In this theory a Cauchy initial-value problem is said to be well-defined if the initial conditions determine a unique solution, and if a continuous variation in the initial conditions gives a continuous variation in the solution. In that part of spacetime where closed timelike loops are not allowed, backward causation does not occur (by definition) and the laws of physics (all expressed as differential equations) satisfy the Cauchy condition. Outside of this *chronal* region, that is, beyond the Cauchy horizon where physics is *dischronal*, however, the possibility of backward causation raises the possibility of violating the Cauchy condition, and in such a case the Cauchy horizon is also sometimes called the *chronology horizon*.[47]

---

[46] V. P. Frolov and I. D. Novikov, "Physical Effects in Wormholes and Time Machines," *Physical Review D*, August 15, 1990, pp. 1057–1065.

The instability of the Cauchy horizon is caused by radiation that propagates in closed timelike loops that thread through the wormhole on 'straight lines.' This radiation, as shown a half-century ago,[48] builds-up unbounded energy density levels at the horizon, and thus destroys the horizon. Thorne and his colleagues argued that the defocusing effect of their wormhole time machine's repulsive gravity would be sufficient to counter a disruptive energy build-up on the horizon (note 27). Subsequent analyses have examined other possible ways to avoid unbounded energy density on the Cauchy horizon. For example, in one paper[49] it was imagined that a wormhole time machine has had a circular motion induced for mouth B; that is, mouth B orbits around mouth A. The result is that the Cauchy horizon now does seem to be stable, because now there are no fixed, straight-line timelike loops threading the wormhole from A to B to A to B to . . .. That is, B is a 'moving target' and there is no point on the Cauchy horizon where the energy density becomes unbounded.

Yet another approach for achieving the disruption of destructive, circulating energy loops through a wormhole is by placing a spherical mirror between the two mouths of the wormhole. Proposed by the Chinese physicist Li-Xin Li, a Li mirror would divert all closed null geodesics (represent circulating radiation) that potentially thread through the wormhole.[50] Such potentially fatal geodesics would, instead, be scattered back into space, whereas a purposeful traveler could navigate *around* the mirror and thus use the wormhole as a time machine.

---

[47] A classic work on the mathematics of Cauchy problems is J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Dover 1952. There is a curious bit of irony in this. In a section of his book, Hadamard uses spacetime to illustrate one possible four-dimensional space and, in passing, he casually writes "This conception was beautifully illustrated a good many years ago by the novelist Wells in his *Time Machine*." Hadamard wrote his book in 1923, and he would almost certainly have been astonished to have been informed that less than seventy years later his work would play a central role in the non-fictional theory of time machines.

[48] C. W. Misner and A. H. Taub, "A Singularity-Free Empty Universe," *Soviet Physics JETP*, January 1969, pp. 122–133.

[49] I. D. Novikov, "An Analysis of the Operation of a Time Machine," *Soviet Physics JETP*, March 1989, pp. 439–443.

[50] L.-X. Li, "New Light on Time Machines: Against the Chronology Protection Conjecture," *Physical Review D*, November 1994, pp. R6037–R6040.

Cauchy horizon instability from the back reaction is central to Hawking's Chronology Protection Conjecture, discussed in Chap. 1. His analysis (see note 54 in Chap. 1) led him to conclude that a physical entity—the stress-energy tensor—becomes unphysical on the Cauchy horizon. That is, because of time-traveling quantum field fluctuations of the vacuum, that tensor diverges to infinity at the horizon. This results in a failure of that horizon to form in the first place or, if it does form, in the creation of a singularity that 'seals-off' the horizon to any would-be time travelers attempting to gain access to the closed timelike loops beyond the horizon. Others, however, argued that Hawking was mistaken in claiming that the divergence of the stress-energy tensor on the Cauchy horizon will always forbid time travel.

In a study, for example, of a complex-valued spacetime metric (and such a metric *is* allowed in the so-called 'sum over all possible geometries, path integral' approach to the quantum theory of gravity), that has causal and non-causal spacetime regions separated not by a Cauchy horizon but rather by a region of complex geometry, the stress-energy tensor is always physical and diverges nowhere. The complex geometry region plays the same role as the Cauchy horizon, because such a region would, *classically*, mean that the two regions cannot be reached from one another, but via quantum tunneling an observer *could* travel between the two regions.[51] In fact, studies of stress-energy divergence actually have a long history. For example, the effect of an unphysical (infinite) gravitational and/or electromagnetic energy flux had been analyzed *years before* the wormhole time machine studies began.[52] The authors studied the case of a potential traveler to "new worlds" who tries to cross the Cauchy horizon of an electrically charged, non-rotating black hole. An even earlier computer study had already concluded that, for such a traveler, the attempt to cross the horizon "looks liable to prove a dangerous undertaking."[53]

It isn't at all clear, in fact, if a *theoretical* divergence of the stress-energy *is* the signature of a failure of physics. One doesn't need anything as bizarre as a time machine for the stress-energy to diverge *on paper*. It was shown nearly 40 years ago, for example, that such a theoretical divergence occurs for the electromagnetic field near a perfectly conducting boundary.[54] But it is simply the unphysical nature of a "*perfectly* conducting" boundary condition that causes the divergence, not the fact that the field actually exists near a conducting boundary. Similarly, other real-life considerations (quantum gravity) *may* keep the stress-energy physical everywhere in a time machine spacetime.

---

[51] L.-X. Li, J.-M. Xu, and L. Liu, "Complex Geometry, Quantum Tunneling, and Time Machines," *Physical Review D*, November 15, 1993, pp. 4735–4737.

[52] S. Chandrasekhar and J. B. Hartle, "On Crossing the Cauchy Horizon of a Reissner-Nordström Black-Hole," *Proceedings of the Royal Society of London A*, December 8, 1982, pp. 301–315.

[53] M. Simpson and R. Penrose, "Internal Instability in a Reissner-Nordström Black Hole," *International Journal of Theoretical Physics*, April 1973, pp. 183–197.

[54] D. Deutsch and P. Candelas, "Boundary Effects in Quantum Field Theory," *Physical Review D*, December 15, 1979, pp. 3063–3080.

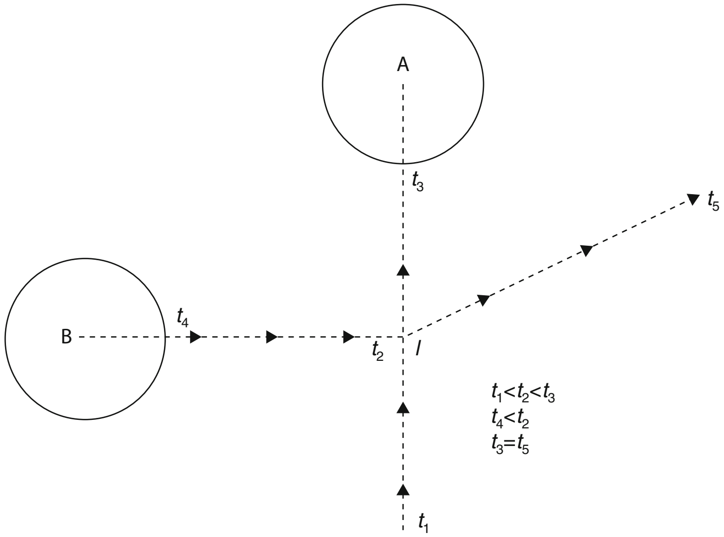

Fig. 6.9 The grandfather paradox in the billiard ball world. A billiard ball approaches mouth A of a time machine wormhole, dead-on center and, just before entering A, it passes without incident through point *I*. The ball then enters A and so exits mouth B in the past, *just in time* to pass through point *I* and *hit its younger self*. This impact knocks the younger ball *away from* A, so we have the familiar paradox of changing the past. That is, the impact did not occur when the ball 'originally' passed through *I* on its way to A and, of course, we also wonder how the ball manages to hit itself after leaving B if it then *doesn't* enter A?

Assuming a wormhole time machine has (somehow) become available, with its Cauchy horizon intact, how do the 'paradoxes' of time travel come into play? In an attempt to study the grandfather paradox, in particular, Thorne and his colleagues studied self-interacting billiard balls traveling backward in time through a wormhole.[55] They used billiard balls—see Figs. 6.9 and 6.10—rather than human time travelers for the same reason Wheeler and Feynman used a pellet and shutter mechanism in their study of advanced electromagnetic waves—to avoid any metaphysical questions about human free will. The central issue for them was the determination of the multiplicity of trajectories for a single, self-interacting time traveling ball, where the Cauchy condition for a well-defined trajectory in spacetime is *unique* self-consistency.

That is, for the trajectory to be well-defined in the Cauchy sense, it was expected there would be exactly *one* consistent trajectory for a self-interacting ball. A multiplicity of zero, of course, would be the physics declaring backward time travel through the wormhole to be nonsense—and that was thought to be a distinct

---

[55] F. Echeverria, G. Klinkhammer, and K. S. Thorne, "Billiard Balls in Wormhole Spacetimes with Closed Timelike Curves: I. Classical Theory," *Physical Review D*, August 15, 1991, pp. 1077–1099. The authors credited the physicist/science fiction writer Robert Forward (who used the same ideas in his 1992 novel *Timemaster*) for motivating their research.

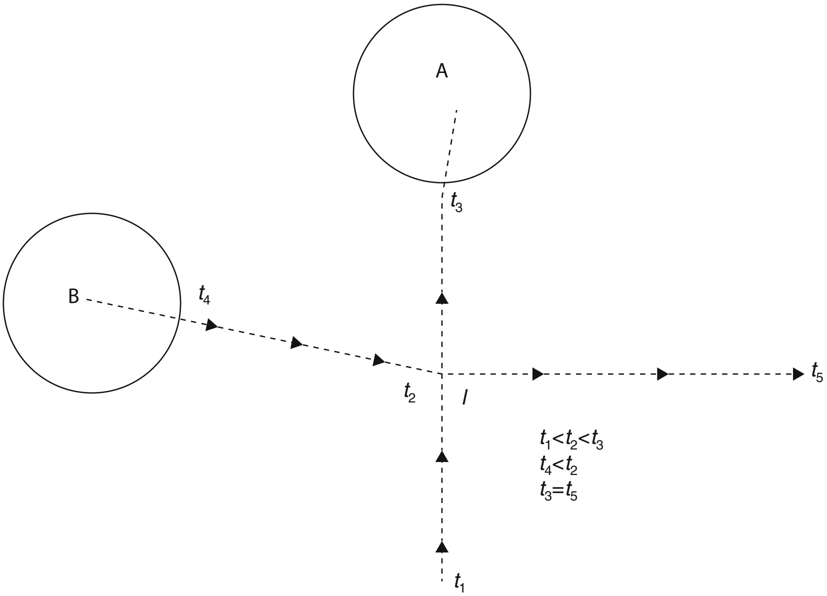

Fig. 6.10 The principle of self-consistency in the billiard ball world. Now the ball, in passing through point *I* on its dead-on center path toward mouth A of the wormhole time machine, is suddenly hit a grazing blow by another ball that has just shot out of mouth B (and into the past) at an angle. The impact knocks the first ball slightly off its original trajectory, and it enters mouth A slightly off-center. Thus, the ball emerges from B into the past slightly off-center and just in time to glace off itself at *I*—which explains *why* it emerged from B slightly off-center!

possibility. The actual results were, however, surprisingly different. It was found, under very general assumptions about the wormhole parameters, that (1) there are no trajectories with zero multiplicity and (2) the multiplicity is *not* one but rather is *always* infinity! Thus, the billiard ball form of the grandfather paradox *was* found to be *not* well-defined, but not for the expected reason that there was no self-consistent solution. Instead, it was because there are *too many* solutions.

This astonishing, completely unexpected result seems to be just what is needed to support the viability of time machines, as it appears to allow a definition of *well-defined* in the Cauchy sense and still permit an answer to the puzzle of free will. The initial conditions of a time traveling ball give rise to an infinity of self-consistent trajectories, each occurring in the same way that a random variable takes on different values with each new performance of the experiment that the random variable is defined on. And yet, there are still unique probability density functions for all sets of measurements that one might make anywhere along these trajectories. Thus, the Cauchy problem is *stochastically* well-defined; at the start of any trajectory, we do not know in detail what will happen except that whatever does happen will be self-consistent. In this probabilistic sense, then, wormhole time travel to the past and the retention of free will both make sense. The Russian physicist Igor Novikov and his colleagues continued the study of time traveling billiard balls,[56] demonstrating that one can *deduce* self-consistency from the long-accepted principle of least action (that is, self-consistency is not an additional assumption to existing physics).[57]

The one-wormhole, two-mouth time machine was actually not the first kind of wormhole time machine described in the physics literature. In their 1988 paper (note 21), Morris and Thorne initially described a time machine constructed from *two* wormholes, but they added a note-in-proof at the end that they had just discovered how to build a time machine using one wormhole (the machine we have been discussing). This reduction in the required number of wormholes was thought to be a technical advance, of course, and so the two-wormhole time machine was put aside.

But not for long. Soon thereafter the concerns about Cauchy horizon stability began to surface, a concern that one-wormhole time machines might destroy themselves just at the instant their mouths were about to be threaded by closed timelike curves. As noted earlier, the negative mass wormhole throat has a defocusing effect on electromagnetic radiation (and so the initial concern, that time traveling photons might be fatal, faded)—but then it was found that vacuum fluctuations of quantum fields are *not* so defocused.[58] That failure to defocus time-traveling vacuum polarizations (as quantum field fluctuations are called) was shown to result in an unphysical divergence of the stress-energy on the Cauchy horizon of a one-wormhole time machine.[59] This sounds bad, but the hope was that the divergence wouldn't actually be fatal: it appeared to be sufficiently sluggish that it was suggested reaching an actual *infinity* of the stress-energy would be precluded by the eventual intercession of quantum gravity. That is, the stress-energy might *try* to become unbounded as spacetime approached the formation of a time machine

---

[56] A. Lossev and I. D. Novikov, "The Jinn of the Time Machine: Nontrivial Self-Consistent Solutions," *Classical and Quantum Gravity*, October 1992, pp. 2309–2321; E. V. Mikheeva and I. D. Novikov, "Inelastic Billiard Ball in Spacetime with a Time Machine," *Physical Review D*, February 15, 1993, pp. 1432–1436; M. B. Mensky and I. D. Novikov, "Three-Dimensional Billiards with a Time Machine," *International Journal of Modern Physics D*, April 1996, pp. 179–192.

[57] A. Carlini, *et al.*, "Time Machines: The Principle of Self-Consistency as a Consequence of the Principle of Minimal Action," October 1995, pp. 557–580, and "Time Machines and the Principle of Self-Consistency as a Consequence of Stationary Action (II): The Cauchy Problem for a Self-Interacting Particle," October 1996, pp. 445–479, both in *International Journal of Modern Physics D*.

[58] V. P. Frolov, "Vacuum Polarization in a Locally Static Multiply Connected Spacetime and a Time-Machine Problem," *Physical Review D*, June 15, 1991, pp. 3878–3894.

[59] S.-W. Kim and K. S. Thorne, "Do Vacuum Fluctuations Prevent the Creation of Closed Timelike Curves?" *Physical Review D*, June 15, 1991, pp. 3929–3947. See also L.-X. Li, "Must Time Machines Be Unstable Against Vacuum Fluctuations?" *Classical and Quantum Gravity*, September 1996, pp. 2563–2568.

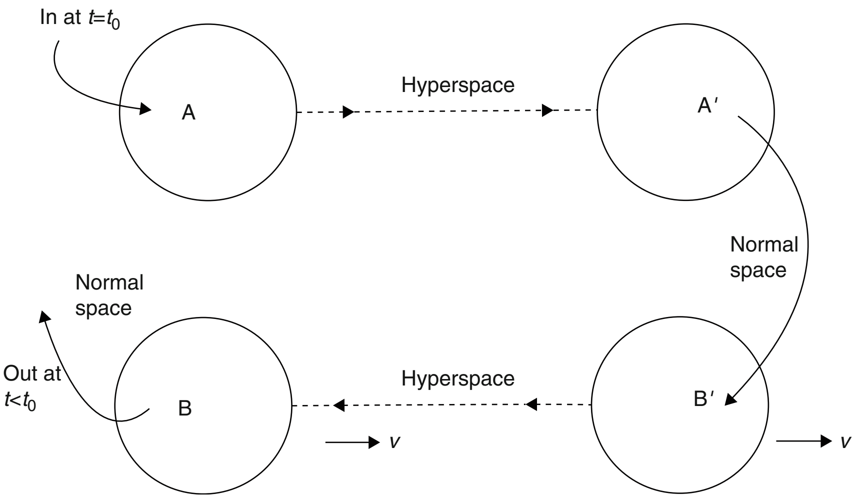

Fig. 6.11 A two-wormhole, Roman ring time machine

but, before it becomes so large as to destroy the time machine, quantum gravity would cut-off the divergence 'in time' (so to speak!) to save the machine.

Hawking disagreed (in his famous Chronology Protection Conjecture—see note 54 in Chap. 1), arguing that Kim and Thorne had made a crucial error in their calculations. According to Hawking, the divergence of the stress-energy may indeed be cut off by quantum gravity, but not before the development of spacetime disturbances representing perhaps a hundred million times the energy levels associated with ordinary chemical binding energies. These would be sufficiently big disturbances to raise serious doubts about the physical survival of a one-wormhole time machine, even in the absence of a true stress-energy infinity. Hence the resurrection of the two-wormhole time machine geometry. Perhaps *it* could avoid the destructive effect of time-traveling vacuum fluctuations.

If a spacetime contains multiple wormholes, then it is called a *Roman spacetime* after the physicist Thomas Roman (at Central Connecticut State University), who was the originator of such spacetimes. Each of these wormholes, individually, is *not* a time machine. Together, however, they form a time machine geometry called a *Roman configuration* (or a *Roman ring*).[60] Here's how.

In Fig. 6.11 two pairs of wormhole mouths are labeled A, A' and B, B'. We imagine that the A, A' wormhole is stationary and that its two mouths are very far apart in normal space—so far apart, in fact, that if a traveler enters A and almost instantly (because the wormhole handle is very short in hyperspace) emerges from A', it will appear to an observer at rest with respect to the wormhole that the traveler has moved faster than light. That is, entering A and exiting A' are events with spacelike separation. Now, imagine also that the wormhole with mouths B, B' is moving past the first wormhole at speed v. To an observer in this second, *moving* frame of reference, the spacelike separation of entering A and exiting A' can result in the two events being temporally reversed if v is sufficiently large (but still less than the speed of light). Therefore, upon emerging from A' the traveler crosses normal space to the moving wormhole mouth B', enters the wormhole, and then almost instantly emerges from mouth B, and finally travels again through normal space to mouth A. If the traveler can make the two trips in normal space in less time than the backward time shift achieved by the temporal reversal of entering A and exiting A', then we have a time machine

---

[60] M. Visser, "Traversable Wormholes: The Roman Ring," *Physical Review D*, April 15, 1997, pp. 5212–5214.

Two simultaneous analyses of the Roman ring time machine each concluded that, for suitable choices of sizes (the radii of the wormhole mouths, the wormhole lengths in normal space, the lateral offset of the two wormholes, and the relative speed of the wormholes), the stress-energy divergence *can* be limited by quantum gravity to an arbitrarily weak level. That is, the two-wormhole time machine is not necessarily destroyed by an unbounded stress-energy on the Cauchy horizon.[61] But not all was now put right.

Visser, in particular, had some strong reservations about the Roman ring. Although he granted that a quantum gravity cut-off the stress-energy divergence *would* probably occur in the Roman ring, he called the required special sizing conditions "bizarre," and asserted that the resulting time machine would be quite useless for a human traveler in any case. For example, he calculated that only if the mouths of the wormholes are separated in normal space by the radius of the universe (!), and only if the wormhole mouths have radii on the order of that of an atomic nucleus, would the cut-off be sufficient to allow the putative time machine to avoid destruction. When Visser reduced the wormholes from universe size to 'merely' that of the distance between the Sun and the Earth, he concluded that it would require energy at the level of the Superconducting Supercollider accelerator to blast an information-bearing message through the narrow wormholes. And even then the 'short' wormholes would provide a maximum penetration into the past of just eight minutes. As Visser put it, "This does not seem to be a workable recipe for studying tomorrow's *Wall Street Journal*.

Lyutikov, on the other hand, took a far less negative stance. He concluded that although Visser's calculations "make it very inconvenient for time travel [by humans]," nevertheless "the [principal] question of the possibility of transmitting information back in time through traversable wormholes would still remain."

The wormholes we have been discussing so far are *static* in time, but another approach is to allow them to be dynamic structures in spacetime. That is, to allow one or more of their parameters to vary with time (perhaps, for example, the throat diameter could collapse). Then, according to one analysis, it is possible to have a traversable wormhole *made of normal matter* and, even though it is collapsing, it would take so long to do so that "a space adventurer will have enough time to pass through the throat of the wormhole from one asymptotically flat region [of spacetime outside the entry mouth of the wormhole] to the other [spacetime region outside the exit mouth of the wormhole] before the radius of the throat shrinks to . . . where the event horizon is developed."[62] Such a dynamic wormhole, it was claimed, satisfies both the weak and the dominant energy[63] conditions, but not the strong energy condition. Thus, gravity would still be repulsive in the throat, but this condition (which would seem to require *exotic* matter) was brushed aside because such a condition is thought to have actually occurred, on a massive scale, during the inflationary stage of the Big Bang[64] (although how that would help in the construction of a wormhole in the future is a bit murky).

---

[61] M. Visser, "Van Vleck Determinants: Traversable Wormhole Spacetimes," April 15,1994, pp. 3963–3980, and M. Lyutikov, "Vacuum Polarization at the Chronology Horizon of the Roman Spacetime," April 15, 1994, pp. 4041–4048, both in *Physical Review D*.

Is it reasonable to think 'useable' wormholes, static or otherwise, can be acquired for the purpose of creating a time machine? At one time, Hawking was sure the answer is no, once writing "The philosophy of this paper is . . . to look for vacuum polarization [the divergence of the stress-energy on the Cauchy horizon] to enforce the chronology protection conjecture."[65] It became increasingly apparent, however, that matters would be a great deal more involved and, as Hawking himself came to admit, "the fact that the energy-momentum tensor fails to diverge [in certain special cases of time machine spacetimes] shows that the back reaction does not enforce chronology protection.".[66]

I think the best (and most honest) way to respond to the 'reasonable' question that opened the previous paragraph is with words from 20 years ago, by the Russian astrophysicist Serguei Krasnikov, words still valid today: "It may well be that the vacuum fluctuations do make the time machine unstable, but nothing at present suggests this. All we have are a few simple examples. In some of them the energy density diverges at the horizon and in some does not. So, the time machine perhaps is stable and perhaps is not."[67]

---

[62] A. Wang and P. S. Letelier, "Dynamical Wormholes and Energy Conditions," *Progress of Theoretical Physics*, July 1995, pp. 137–142. See also L. A. Anchordoqui, *et al.*, "Evolving Wormhole Geometries," *Physical Review D*, January 15, 1998, pp. 829–833.

[63] The dominant energy condition is the weak energy condition *plus* the requirement that any observed energy flux is *never* superluminal.

[64] During inflation the universe is thought to have expanded at a rate far beyond human comprehension. It has been estimated that during the first 10^{-35} second of the Big Bang the universe doubled in each spatial dimension by a factor of two each 10^{-37} second; that is, there were about 100 such doublings. Thus, there was an increase by a factor of $2^{100} \approx 10^{30}$ in each linear dimension of the universe, and the volume increased by the cube of that enormous factor. See Alan Guth, *The Inflationary Universe*, Addison-Wesley 1997.

[65] S. Hawking, "Quantum Coherence and Closed Timelike Curves," *Physical Review D*, November 15, 1995, pp. 5681–5686.

[66] M. J. Cassidy and S. W. Hawking, "Models for Chronology Selection," *Physical Review D*, February 15, 1998, pp. 2372–2380. See also L.-X. Li and J. R. Gott, "Self-Consistent Vacuum for Misner Space and the Chronology Protection Conjecture," *Physical Review Letters*, April 6, 1998, pp. 2980–2983.

The daunting level of technology required to *build* a wormhole (with or without exotic matter) doesn't mean we can't search for existing wormholes. Perhaps, for example, vast wormhole networks were formed naturally at Big Bang time, as described in Gregory Benford's 1997 novel *Foundation's Fear*, where wormholes are "leftovers from the Great Emergence [the Big Bang]." Or perhaps "advanced civilizations" long ago constructed a vast, pan-galactic 'subway system' of wormholes like the one described in Carl Sagan's 1985 novel Co*ntact* (and dramatically illustrated in the 1997 film).

Of course, any such wormhole, if found (via its double-spike light signature, for example), could be a very long way from Earth. It might even be in another galaxy. So, even if we found a wormhole, what could we *do* with it? Surprisingly, maybe a lot. The Russian physicists Igor Novikov and Andrei Lossev (note 56) suggested that a wormhole might be very useful *even if its location is completely unknown, even if we haven't yet even discovered it*! The only assumption they made was that the wormhole has existed for a "sufficiently long time" (and precisely what that means will be explained in just a bit). With that assumption, they showed how to make an information-creating time loop. Here's how they did that.

They began their analysis by assuming that people have no knowledge of how to build spacecraft that can make the interstellar voyage to the distant wormhole, even if they knew in which direction to go to reach the mouth that leads backward in time (mouth B). Instead, they build an automatic spacecraft construction plant that can follow any detailed sequence of instructions provided to it, and then stockpile it with a supply of raw materials (energy, steel, plastic, computers, and so on). When the spacecraft construction is done (*how* that is done is explained in the next paragraph), the last step before launching the spacecraft toward mouth B *will be* to load the on-board computer with the following three pieces of information:

1. The detailed sequence of instructions to be followed in the construction of the spacecraft;
2. The direction from Earth to mouth B;
3. The direction from mouth A (the wormhole exit mouth in the past) back to Earth.

To summarize, people build the automatic plant, load it up with raw materials, *and then withdraw*. This last step is crucial, because it eliminates human free will from further consideration, that is, it removes any temptation to create a bilking paradox. So, what happens next?

Lossev and Novikov suggest that what happens next is that a *very* old spacecraft suddenly appears in the sky and lands next to the automatic construction plant. In its on-board computer are items a, b, and c. Using item a, the automatic plant makes a new spacecraft, then loads the new on-board computer with items a, b, and c from the *very* old spacecraft's on-board computer, and then the new spacecraft is launched toward mouth B (using the information of item b). The *very* old spacecraft is given an honored place in a museum.

---

[67] S. V. Krasnikov, "Quantum Stability of the Time Machine," *Physical Review D*, December 15, 1996, pp. 7322–7327.

The new spacecraft arrives at the distant mouth B in the far future, by which time it is, of course, an old spacecraft (but not yet a *very* old spacecraft). It then plunges into mouth B and almost immediately emerges from mouth A, in the past. Indeed, it repeats this process as many times as required until it is in the far distant past, at a time even before it left Earth. (It might seem that to do this, the spacecraft's computer memory needs a fourth piece of information, the direction from mouth A back to mouth B, but in fact items b and c are sufficient for the old spacecraft to find its way from A to B.) It is now clear how long the wormhole must have been in existence. The old spacecraft repeatedly uses the wormhole time machine until it is so far in the past that it can cruise back to Earth at normal speed (it knows the way back because of item c) and arrive as a *very* old spacecraft, just in time to be placed in the museum!

As Lossev and Novikov pointed out, this remarkable, looped sequence of events has increased knowledge from what it was at the time just before the automatic construction plant was built. People now know both how to build an interstellar spacecraft, and the locations of both mouths of the wormhole. They also now possess a *very* old, used spacecraft. It is curious to note that although the information in the *very* old spacecraft's computer memory has traveled on a closed time loop, the *very* old spacecraft itself has not. This is because the spacecraft left Earth when new, but arrived back (before it left) as *very* old, whereupon it promptly entered a museum. There is therefore no question about the origin of the *very* old spacecraft, but where did the information of items a, b, and c come from? Lossev and Novikov say it came from the energy gained by the spacecraft as it interacted (will interact?) with the rest of the universe while on its journey.

Nobody said time travel isn't weird!

6.4 Gott's Cosmic String Time Machine

> "It's an amazingly simple solution. It doesn't take much physics to understand it."
>
> —MIT astrophysicist Alan Guth, on Gott's discovery of the cosmic string time machine[68]

> "Louise, working out the spacetime geometry of a cosmic string is a hard problem in general relativity. But, given that geometry, all the rest of it is no more than Pythagoras' theorem . . ."
>
> —a character in Stephen Baxter's 1994 novel *Ring*, agreeing with Guth.

---

[68] Quoted from J. Travis, "Could a Pair of Cosmic Strings Open a Route Into the Past?" *Science*, April 10, 1992, pp. 179–180.

A new way to gain access to closed timelike curves, without the involvement of the exotic matter needed by negative-mass wormholes, was described in 1991 by the Princeton physicist J. Richard Gott.[69] Gott gave exact solutions to Einstein's gravitational field equations for what are called *cosmic strings*, solutions that (1) unlike wormholes, do not violate any of the energy conditions, (2) unlike black holes have no crushing singularities or event horizons, and (3) are not topologically multiply connected.

Cosmic strings are fantastically thin (10^{-28} cm in radius) filaments of pure energy that are thought to stretch the width of the universe and to have an enormous linear mass-energy density of 10^{28} g/cm. To generate closed timelike paths in spacetime, Gott required that either two fast-moving (which means moving at practically the speed of light) parallel cosmic strings pass each other on a near-collision course, or that there be a closed-loop string that collapses in a slightly non-planar manner so that the opposite, nearly straight sides 'just miss.' The gravitational interaction of the passing strings can 'warp' spacetime enough to produce closed timelike curves.

A hint at the possibility of violating causality with strings had appeared before Gott's work, but those authors didn't take the time travel implications seriously. As they wrote, "We argue … that any realistic model [for a spinning string with angular momentum[70]] … will not have closed timelike curves."[71] Gott, however, showed that as two strings pass each other, closed timelike loops *do* encircle the strings.

Gott, who appears to be far less rigid in his view of time travel than are many of his fellow physicists, held out an escape to those who pale at the very thought of time travel to the past. Perhaps, he suggested (following in the footsteps of an analysis by Hawking[72]), as the strings (or string-loop sides) pass, a black hole will form with an event horizon that will seal-off the closed timelike curves from any would-be time traveler. Or perhaps, he further suggested, the more realistic case of non-singular strings (that is, strings with non-zero-filament radii) and possessing

---

[69] J. R. Gott, "Closed Timelike Curves Produced by Pairs of Moving Cosmic Strings: Exact Solutions," *Physical Review Letters*, March 4, 1991, pp. 1126–1129.

[70] The two strings in a Gott-pair are *not* necessarily spinning, and no such assumption was made by Gott. They don't even have to be parallel. If the strings have no spin, then it takes two strings to make a time machine. If spin *is* allowed, however, then just a single string will suffice for time travel: see S. Deser and R. Jackiw, "Time Travel?" *Comments on Nuclear and Particle Physics*, September 1992, pp. 337–354.

[71] D. Harari and A. P. Polychronakas, "Gravitational Time Delay Due to a Spinning String," *Physical Review D*, November 15, 1988, pp. 3320–3322.

[72] S. W. Hawking, "Gravitational Radiation from Collapsing Cosmic String Loops," *Physics Letters B*, August 23, 1990, pp. 36–38.

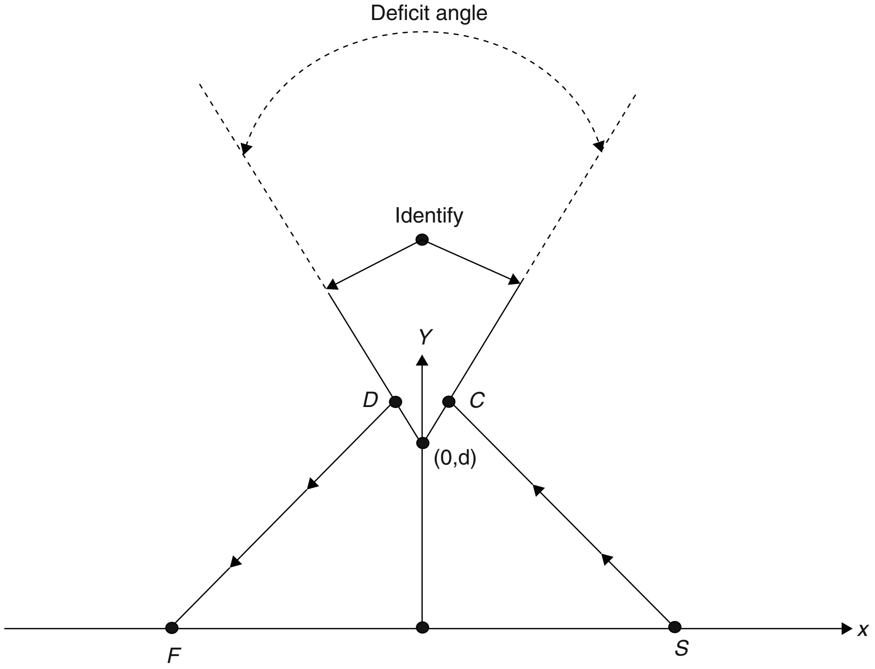

Fig. 6.12 The deficit angle in spacetime formed by a cosmic string

spin would banish the terrifying closed timelike curves. Subsequent analyses along those lines, however, continued to find the time travel implications intact.[73]

Here's how the cosmic string time machine works. In an earlier work,[74] published in 1985, Gott discovered that a cosmic string warps spacetime in a highly characteristic way, as shown in Fig. 6.12. A stationary cosmic string is imagined as perpendicular to the xy-plane (the plane of the page) and passing through the page at the point $(0, d)$ on the y-axis. The warp produced by the string is as though a wedge of angle 2α (this angle is called the *deficit angle*) were cut out of spacetime and the

---

[73] See, for example, B. Jensen, "Notes on Spinning Strings," *Classical and Quantum Gravity*, January 1992, pp. L7–L12, H. H. Soleng, "A Spinning String," *General Relativity and Gravitation*, January 1992, pp. 111–117, (the next two are in the *Physical Review D*), B. Jensen and H. H. Soleng, "General-Relativistic Model of a Spinning Cosmic String," May 15, 1992, pp. 3528–3533, and M. Novello and M. C. M. da Silva, "Cosmic Spinning String and Causal Protecting Capsules," January 15, 1994, pp. 825–830.

[74] J. R. Gott, "Gravitational Lensing Effects of Vacuum Strings: Exact Solutions," *The Astrophysical Journal*, January 15, 1985, pp. 422–427. Gott's discovery was independently reported in W. A. Hiscock, "Exact Gravitational Field of a String," *Physical Review D*, June 15, 1985, pp. 3288–3290.

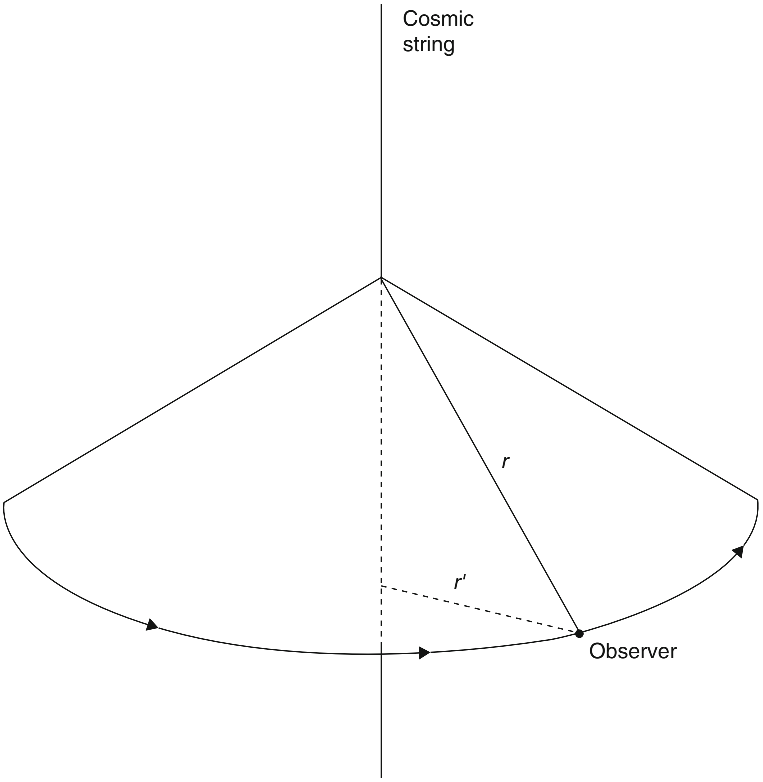

Fig. 6.13 The warped, conical spacetime around a cosmic string, An observer in this spacetime thinks she is distance r from the string, but a 'meta-observer' sees that she is actually distance r' from the string. Thus, if the observer follows a complete circular path around the string, she will travel a distance of $2\pi r' < 2\pi r$. The observer in the spacetime will interpret this result by saying that the angle 2π is really 2π minus 'a deficit'

edges of the cut were then 'glued' together; for example, points C and D are identified as identical. The reason for the term *deficit angle* is that at radius r from the string, a circular path around the string has the reduced length $(2\pi - 2\alpha)r$, and not the usual $2\pi r$ (spacetime around the string, while *locally* flat, is actually 'conical,' as illustrated in Fig. 6.13).

The deficit angle is equal to $8\pi\mu$ radians (in a system of units where G, Newton's gravitational constant, is 1) if the linear mass-energy density μ is expressed in units of Planck masses per Planck length. For example, $\mu = 1$ corresponds to 1.35×10^{28} g/cm (think of something on the order of the mass of the Earth per inch of the string). For 'more typical' values—say a 'mere' $\mu = 10^{22}$ g/cm, $2\alpha = 0.001^\circ$. While Gott's paper had appeared 5 months before Hiscock's (see note 74), it is evident that Hiscock's work was done before he became aware of Gott's. Both papers treat exact derivations of the deficit angle but, in fact, the correct expression had actually been published 4 years earlier (but from a linearized form of the gravitational field equations and so the result was not as 'conclusive' as are Gott's and Hiscock's)[75]

Now, consider the two points S and F on the x-axis at $(x_0, 0)$ and $(-x_0, 0)$ in Fig. 6.12. Suppose we want to send photons from S to F. In normal, 'unwarped' spacetime, the direct path from S to F through the origin has length $2x_0$. There is also another possible path, however, S to C/D to F, that loops out and around the cosmic string. Indeed, this second path is simply the path a gravitationally lensed photon would take (an observer at F would see *two* images of S) and this—*not* time travel—is the issue that originally attracted Gott's attention to cosmic strings.[76] If the deficit angle were zero, then this alternative path would always be longer than $2x_0$, for any value of x_0. For the case of $2\alpha > 0$, however, if x_0 is large enough ($x_0 \gg d$), then it is possible for 'around the string and over the missing spacetime wedge' path to be shorter than the direct path.

The indirect path provides a way for a *sub*luminal trip (say, by rocket) from S to F to beat a photon traveling on the direct path. That is, the two events of the 'rocket leaving S ' and the 'rocket arriving at F' are spacelike separated. Thus, it is possible to find a moving frame of reference in which these two events are *reversed in temporal order*. In that frame of reference, the cosmic string (which is stationary in the reference frame of S and in that of F) will move—at speed v, say—in the $+x$ direction, and in that frame of reference the rocket will arrive at F *before* it leaves S.[77]

Then to complete the construction of a closed timelike path, simply repeat the process as shown in Fig. 6.14. That is, after the rocket arrives at F, have it turn around and fly back to S out-and-around and through the deficit angle spacetime warp due to a *second* cosmic string on the negative y-axis and perpendicular to the xy-plane. This second string is moving at speed $-v$ (that is, opposite to the first string), so the rocket will arrive at S before it leaves F. But that means it arrives at S before it leaves S; that is, the rocket has traveled into the past. In other words, the rocket has traveled into the past. This entire process is precisely the same idea behind the two-wormhole Roman-ring time machine discussion from the previous section.

Now, instead of having two oppositely moving reference frames, one in which the top, stationary string at $(0, d)$ appears to be moving at $+v$ and another frame in

---

[75] A. Vilenkin, "Gravitational Field of Vacuum Domain Walls and Strings," *Physical Review D*, February 15, 1981, pp. 852–857.

[76] Just as discussed earlier in the context of wormholes, gravitational lensing may offer a way to detect cosmic strings. See, for example, the two papers by D. L. Ossipov, "Diffraction of Light by a Cosmic String," November 1995, pp. 765–771, and "Contribution of Strings to the Observed Variability of Extragalactic Sources of Radiation," September 1996, pp. 419–425, both in *JETP Letters*.

[77] For this to happen, however, v must be very close to the speed of light. Gott (see note 69) showed that with $v = \tanh(\theta)$, the condition for the rocket to arrive back at S before it leaves S is $\cosh(\theta)\sin(\alpha) > 1$, where α is one-half the deficit angle. For $\mu = 10^{22}$ g/cm, this gives $v = 0.99999999995$ (times the speed of light).

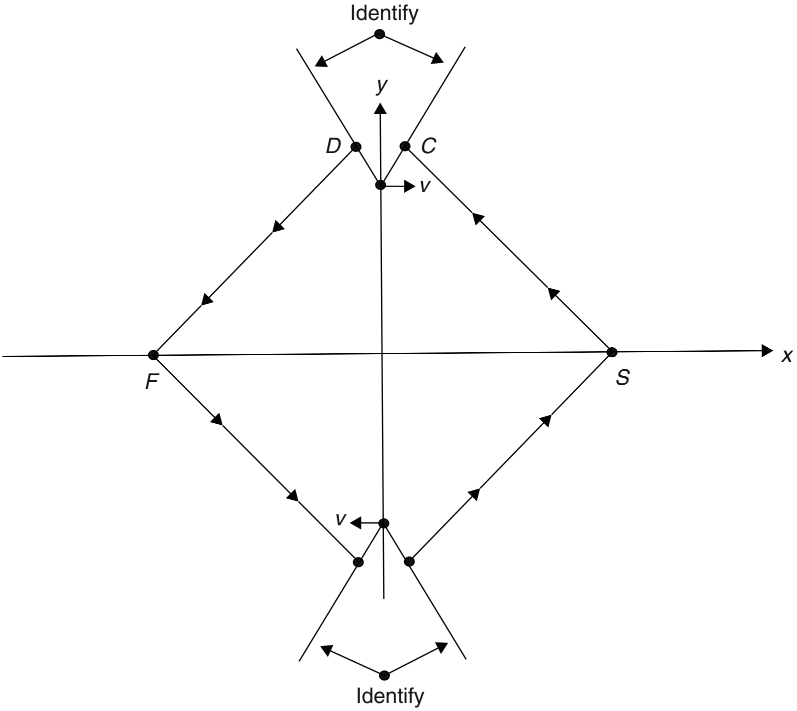

Fig. 6.14 Gott's spacetime, formed by joining two oppositely moving versions of the spacetime in Fig. 6.13

which the bottom, stationary string appears to be moving at $-v$, we can imagine an observer in the *stationary center-of mass* frame watching two strings that are moving at $+v$ and $-v$. This leaves the situation unchanged, so in the center-of-mass frame the rocket does travel into the past, arriving back at S before it leaves S. That is, the rocket has traveled all the way around a closed timelike world line. Note, too, that the geometric condition mentioned earlier of $x_0 \gg d$ immediately implies that for x_0 *not* sufficiently large, there isn't a closed timelike path from S to F and then back to S; that is, there is a region in Gott's spacetime where such time travel journeys *cannot occur*.

Another physicist pursued Gott's analysis in an attempt to see whether these 'time travel paths' are *created* as the strings approach each other or, instead, if the paths exist at other times as well.[78] This important question gets to the idea of whether such time machine paths can be *intentionally created* by humans via a dynamical process (a *strong* time machine), or whether all such paths have existed since the formation of the universe (a *weak* time machine). This issue involves Hawking's chronology protection conjecture, which you'll recall asserts that the laws of physics will always (somehow) prevent the creation of a time machine. One reason Hawking repeatedly gave for believing the Conjecture is the apparent absence of time travelers from the future among us now (in their past). The only possible exception allowed by the Conjecture is the creation of closed timelike loops at the moment the universe was created (at that moment there *was no past* for time travelers to invade!). Ori proved that the closed timelike loops around Gott's cosmic strings are *always* present: that is, a time machine is *not created* by the near collision of the strings. Thus, Hawking's Conjecture is nor refuted by Gott's spacetime.

---

[78] A. Ori, "Rapidly Moving Cosmic Strings and Chronology Protection," *Physical Review D*, October 15, 1991, pp. 2214–2215.

One very curious issue is *where* the closed time loops are before the strings pass one another. As mentioned briefly by Ori, and further discussed by others,[79] the time loops are initially at spatial infinity. To this concern, Gott and a colleague made the following very strong reply:

> "[A problem] Deser *et al.* present with respect to the Gott spacetime is that it contains CTC's at spacelike infinity; this is supposed to be an unacceptable boundary condition. We wonder, however, how they know so much about boundary conditions at spacelike infinity. In our own Universe we do not know what spacelike infinity looks like (if it exists) since we have not seen it yet. We certainly have no way of knowing whether or not there are CTC's there. The working physicist is, of course, free to impose simple and convenient boundary conditions (e.g., asymptotic flatness) on a system in order to isolate and understand the processes occurring within it. *But boundary conditions are tools of physicists, and they should not be confused with laws of physics* [my emphasis]. There may be such laws of nature that restrict the possible structure of spacelike infinity, and even prohibit CTC's there, but in the absence of evidence such laws should not be postulated *ad hoc*."[80]

Still, as Ori had observed the year before, having time loops collapsing inward from infinity toward humans who might, fortuitously, wish to use them at *just the instant they so conveniently arrive*, is "a situation which has little to do with the creation of a time machine by a *human being* [my emphasis]."[81]

The most damning objection to Gott's cosmic string time machine came, ironically, from Gott himself. The two-string spacetime of Fig. 6.14 might actually, he and a colleague wrote (see Li and Gott, note 66) be destabilized by the non-zero mass of any would-be time traveler. They suggest that this concern could perhaps be 'solved' by assuming that the time traveler and her spaceship have a spherically symmetric mass distribution surrounded by a negative-mass shell to give zero net mass (and thus a zero net gravitational field that would *not* destroy the closed timelike curves of the strings). But that, they further observed, would negate the crucial advantage—no exotic matter and so no violation of the weak energy condition—that a cosmic string time machine enjoys over a wormhole time machine.

---

[79] See, for example, S. Deser, *et al.*, "Physical Cosmic Strings Do Not Generate Closed Timelike Curves," *Physical Review Letters*, January 20, 1992, pp. 267–269.

[80] M. P. Headrick and J. R. Gott, "(2+1)-Dimensional Spacetimes Containing Closed Timelike Curves," *Physical Review D*, December 15, 1994, pp. 7244–7259.

[81] A. Ori, "Must Time-Machine Construction Violate the Weak Energy Condition?" *Physical Review Letters*, October 18, 1993, pp. 2517–2520.

Does a trip around a pair of cosmic strings present other problems *aside* from the sheer fantastic physics of the strings themselves? Well, I think "turning the rocket around at F and flying back to S" is a lot easier to write than it would be to actually do! The entire trip has to occur while the strings (moving at essentially light speed) are in a position *to be* flown around. As one character says to another in Stephen Baxter's novel *Ring*, "Louise, the strings are traveling just under the speed of light—within three decimal places of it, actually. [Our ship is] traveling at a little over half-light speed. The turning curves, and the accelerations, are incredible . . ." I think so! And I do wonder who—*or what*(!)—is actually controlling a *maneuvering* rocket traveling faster than $\frac{1}{2}c$?

6.5 Cutting and Warping Spacetime

> "The warp drive spacetime of Alcubierre is impossible to set up . . . one needs to transcend the speed of light in order to construct the warp drive in the first place . . . put roughly, you need one to make one!"[82]

In this chapter we've talked about the physics of three specific time machine 'implementations': the rotating cylinder, the wormhole, and the cosmic string. One can also discuss the 'construction' of a time machine in a more geometrical (yet still physical) way by performing what is called *spacetime surgery* to arrive at what is often referred to as a *Deutsch-Politzer spacetime* (after the two physicists who are closely associated with it[83]). With this surgery we arrive at a simple spacetime picture of the grandfather paradox (as you'll soon see).

We start with a flat, two-dimensional Minkowski spacetime, the x, t system in Fig. 6.15, and then imagine that (somehow) two *cuts* in that spacetime come into

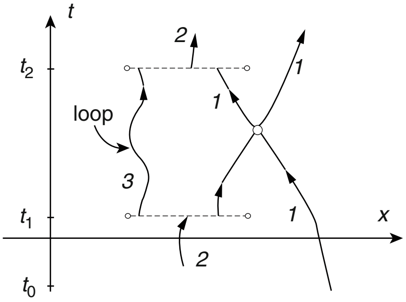

Fig. 6.15 Minkowski spacetime transformed into a time machine with two 'cuts'

existence. These cuts are the two horizontal dashed lines in the figure, one with an arrowhead going into it, and one with an arrowhead coming out of it (each labeled with ‘2’). We further imagine that each cut has two edges, with the upper edge of the lower cut ‘glued’ to the lower edge of the upper cut (and the lower edge of the lower cut ‘glued’ to the upper edge of the upper cut

---

[82] D. H. Coule, "No Warp Drive," *Classical and Quantum Gravity*, August 1998, pp. 2523–2527, offering a pessimistic view of warp drive.

[83] See D. Deutsch, "Quantum Mechanics Near Closed Timelike Lines," November 15, 1991, pp. 3197–3217, and H. D. Politzer, "Simple Quantum Systems with Closed Timelike Curves," November 15, 1992, pp. 4470–4476, both in *Physical Review D*.

These ‘gluing’s’ explain why the arrowhead marked 2 into the lower cut at time t_1 emerges from the upper cut at time $t_2 > t_1$, and why the arrowhead marked 1 into the upper cut at time t_2 emerges from the lower cut at time $t_1 < t_2$. The 2-line is the world line of a particle that simply disappears from spacetime during $t_1 < \mathrm{t} < t_2$, while the 3-line is the world line of a particle trapped in an endless time loop (remember the 1993 film *Groundhog Day*?). Clearly, the sub-region of spacetime between the two cuts is not ‘normal’ spacetime. In fact, the 1-line shows that we have encountered a time machine spacetime, as a particle entering it from $t < t_1$ (passing to the right of the lower cut) can enter the lower edge of upper cut at $t = t_2$ and so emerge from the upper edge of the lower cut at the *earlier* time $t = t_1$ and thus interact with itself before it entered the upper edge! And that, of course, sets-up a grandfather paradox situation.

This picture leaves one thing obviously (and glaringly) unexplained—just *how* does one *cut and glue* spacetime? The reason, in spite of that question, that physicists nonetheless study situations depicted in Fig. 6.15, is because it allows them to explore what *could* happen *if* through some (yet unknown) process a time machine spacetime should suddenly appear. Who says physicists aren’t optimists?

The idea of modifying spacetime itself to ‘make’ a time machine (look back in Sect. 3.5, at the discussion there on what it means to solve the gravitational field equations) has also appeared in connection with another of science fiction’s favorite ideas, one almost as spectacular as time travel: *the FTL warp drive*. The lure of interstellar FTL travel, for both science fiction enthusiasts and physicists is, of course, simply undeniable. Consider, for example, these words by a Russian physicist:

> “Everybody knows that nothing can move faster than light. The regrettable consequences of this fact are also well known. Most of the interesting or promising candidates for colonization are so distant from us that the light barrier seems to make an insurmountable obstacle for any expedition. It is, for example, 200 pc [1 parsec is equal to about 3.2 light-years] from us to the Pole star, 500 pc to Deneb [the brightest star in the constellation Cygnus], and ~10 kpc to the center of the Galaxy, not to mention other galaxies (hundreds of kiloparsecs). It makes no sense to send an expedition if we know that thousands of years will elapse before we receive its report. On the other hand, the prospects of being confined forever to the Solar System without any hope of visiting other civilizations or examining closely black holes, supergiants, and other marvels are so gloomy that it seems necessary to search for some way out.”[84]

---

[84] S. V. Krasnikov, “Hyperfast Travel in General Relativity,” *Physical Review D*, April 15, 1998, pp. 4760–4766. Possible travel distances have been greatly reduced since Krasnikov wrote. In 2011, for example, astronomers announced the discovery of a red dwarf star with three planets (each of mass comparable to Earth’s), all in the star’s so-called *habitable zone* (where water can exist on the surface in the liquid state). All three planets are solid (not gaseous as are Jupiter,

In response to that, we might ask if *FTL* trips will someday be made by humans in spaceships? *Maybe*—but only if such journeys can be made in a very *unordinary* spacetime. That is, continuing with Krasnikov's passage:

> "The point … is that [whereas the light barrier exists in special relativity] in general relativity one can try to change the time necessary for some travel not only by varying one's speed [as in special relativity] but also … by changing the distance one has to cover."

To understand what Krasnikov was getting at, let's consider the theoretical analysis made 4 years earlier by the Mexican mathematical physicist Miguel Alcubierre on, astonishingly, how to make a *Star Trek* warp drive![85] He did this by demonstrating a spacetime metric that, by literally expanding and contracting the local spacetime of a spaceship and its neighborhood, achieves space travel between any two points, no matter how far apart, in arbitrarily little elapsed time (for both the spaceship, and external non-spaceship observers, there is no time dilation effect[86]).

Alcubierre opened his analysis with words designed to explain how FTL travel is possible, given all that I've told you earlier in this book about how FTL travel is *not* possible (according to *special* relativity). As he wrote,

> "Since our everyday experience is based on a Euclidean space, it is natural to believe that if nothing can locally travel faster than light then given two places that are separated by a proper spatial distance D, it is impossible to make a round trip between them in a time less than $2D/c$ (where c is the speed of light), as measured by an observer that always remains at the place of departure. Of course, from our knowledge of special relativity we know that the time measured by the person making the round trip can be made arbitrarily small if his (or her) speed approaches that of light. However, the fact within the framework of general relativity and without the need to introduce non-trivial topologies (wormholes), one can actually make such a round trip in an arbitrarily short time as measured by an observer that remained at rest will probably come as a surprise to many people."

That last sentence is almost surely a grand understatement, and Alcubierre quickly went on to explain.

> "The basic idea can be more easily understood if we think for a moment of the inflationary phase of the early Universe, and consider the relative speed of separation of two co-moving observers. It is easy to convince oneself that, if we define this relative speed as the rate of change of proper spatial distance over proper time, we will obtain a value that is much larger than the speed of light. This doesn't mean that our observers will be travelling faster

---

Saturn, Neptune and Uranus) and so, as potentially habitable, are candidates for a visit. The star and its planets are 'only' 22 light-years from Earth. In *Star Trek*, FTL speed is described by the *warp factor*, which is the cube-root of the multiple of the speed of light at which the spaceship *Enterprise* travels. So, for example, to make the journey from Earth to the red dwarf in one month of ship time (see ahead also to note 86), the required FTL speed would be 264 times the speed of light, or warp factor 6.4. In science fiction, a *warp drive* is imagined as the means for achieving such speeds.

[85] M. Alcubierre, "The Warp Drive: Hyper-fast Travel Within General Relativity," *Classical and Quantum Gravity*, May 1994, pp. L73–L77.

[86] That is, the passage of time on the spaceship is *identical* with the passage of time on Earth. *With the warp drive, there is no twin paradox.*

> than light: *they always move inside their local light-cones* [my emphasis]. The enormous speed of separation comes from the expansion of spacetime itself."

In a similar fashion, a *contraction* of spacetime can result in being able to *approach* an object at FTL speed.

In fact, we've actually already encountered one way to obey special relativity's *local* limit on speeds to that of light, while still achieving superluminal speed on a *global* level. That is, we *could* do that *if* general relativity really does allow wormholes in spacetime (what Alcubierre calls a "non-trivial topology"). That's because we can imagine a wormhole connecting two points in space that are light-years apart in that space, and yet the distance *through* the wormhole itself is quite short. Thus, a spaceship transiting the wormhole could do so at *sub*luminal speed at all times, and yet to an observer in normal space the speed would appear to be far in excess of the speed of light.

Determining just *how* to achieve the spacetime warp, however, is far different from simply demonstrating that such a warp is consistent with the general theory of relativity. The 1996 movie *Star Trek: First Contact*, for example, is about the invention of the warp drive in the twenty-first century. The whole thing fits inside a discarded ICBM which, as you'll soon see, is a *vast* underestimation of the technology required to control the energies associated with a real warp drive. For a spacetime engineer to *build* the warp drive bubble means she has to determine the required mass-energy distribution that results in Alcubierre's assumed spacetime metric. And that brings us to the central problem of the warp drive—the warp drive engine of an FTL starship requires (just like a wormhole) exotic matter (negative energy)—stuff that violates all the usually assumed energy conditions of general relativity.[87]

The weak, strong, and dominant energy conditions are *all* violated because the Alcubierre spacetime warp requires a negative energy density in the 'skin' of the warp bubble. As discussed earlier, in connection with wormholes, negative energy density *can* be achieved on a *micro*scopic scale, but for the Alcubierre warp drive we are talking about a *lot* of exotic matter. In their paper, Pfenning and Ford calculated that, for what they called "a macroscopically useful warp drive" with a radius of 100 m "so that we may fit a ship inside [the warp bubble]," the negative energy required for the warp bubble is on the order of, as they so graphically put it, "roughly ten orders of magnitude greater than the [energy of the] total mass of the entire universe."

Two years later, after making some adjustments to the spacetime metric assumed by Alcubierre (that is, to the distribution of mass-energy to produce the warp bubble), it was shown that the negative energy required by the warp drive could be greatly reduced.[88] The reduction is, in fact, spectacular, but only in a relative sense (you can reduce a mass that is ten orders of magnitude greater than that of the entire universe by a *huge* factor and still be left with a pretty stupendous number). The reduced amount of negative energy required for a warp bubble able to contain a human-sized spaceship is now down to 'only' "of the order of a few solar masses [-1.4×10^{30} kg]."

---

[87] M. J. Pfenning and L. H. Ford, "The Unphysical Nature of 'Warp Drive'," *Classical and Quantum Gravity*, July 1997, pp. 1743–1751. See also K. D. Olum, "Superluminal Travel Requires Negative Energies," *Physical Review Letters*, October 26, 1998, pp. 3567–3570.

As mentioned earlier, in connection with the Casimir effect and its theoretical use in a wormhole time machine, although quantum field theory does not preclude negative energy densities, that does not mean it is possible to observe arbitrarily large negative densities for arbitrarily long times. In fact, certain quantum inequalities (QI's), much like Heisenberg's uncertainty principle, have been established that place bounds on the magnitude and duration of observable negative energy density.[89] These QI's have the general form of

$$\widehat{\rho} t_0^4 > -C$$

where C is a positive constant that depends on the nature of the particular quantum field being considered, t_0 is the time duration, and $\widehat{\rho}$ is the integrated energy density along a finite section of a geodesic (free fall) world line. The form of the QI shows that as t_0 increases, $\widehat{\rho}$ must quickly decrease. For example, if t_0 doubles, then $\widehat{\rho}$ must decrease by a factor of *sixteen*, a result that caused Ford and Roman to conclude that it "appears probable that nature will always prevent us from producing gross macroscopic effects with negative energy."

'When it rains it pours,' goes an old saying, and that applies to the warp drive's potential difficulties: in addition to the need for exotic matter, there are two more concerns as well, both *operational* in detail. First, running into any space matter encountered by the leading edges of the warp bubble (where spacetime is shrinking), such as interstellar dust, would certainly generate intense radiation. The ship, then, should carry plenty of shielding which, curiously, would not be a problem because the energy density of the warp, itself, is *independent* of the mass in the bubble's interior. In any case, the warp drive should clearly not be engaged anywhere near any sizeable chunk of matter, like a planet (and, indeed, that constraint was followed in *Star Trek*). Second an even more severe problem was discovered by Krasnikov. In unpublished work he showed that the ship at the center of the bubble is not causally connected to the edges of the bubble. That is, the ship's crew could *not* create a warp bubble on demand and, after it had been created, could not control it on demand.[90]

---

[88] C. van den Broeck, "A Warp Drive with More Reasonable Total Energy Requirements," *Classical and Quantum Gravity*, December 1999, pp. 3973–3979.

[89] L. H. Ford and T. A. Roman, "Restrictions on Negative Energy Density in Flat Spacetime," *Physical Review D*, February 15, 1997, pp. 2082–2089. For some interesting remarks about the QI's, see J. F. Woodward, "Twists of Fate: Can We Make Traversable Wormholes in Spacetime?" *Foundations of Physics Letters*, April 1997, pp. 153–181.

It is important to both understand what that means, as well as what it does *not* mean. The causality issue does *not* mean that Alcubierre warp bubbles are impossible to create (perhaps they are, but not because of a lack of causality). It only means that whatever action is required to change the spacetime metric to make a warp bubble *has to already have been done before the decision to use the bubble is made*. Thus, a warp bubble wouldn't be of any use for a starship that needs to escape a sudden, unexpected threat. But, as Everett and Roman cautiously observe, the warp bubble might have a more mundane use: "Suppose space has been warped to create a bubble traveling from Earth to some distant star, e.g., Deneb, at superluminal speed. A spaceship, appropriately located with respect to the bubble trajectory, could then choose to enter the bubble, rather like a passenger catching a passing trolley car, and thus make the superluminal journey."

At the end of his paper (note 85) Alcubierre briefly speculated on the possibility of using his superluminal warp drive to build a time machine (showing, again, the intimate connection between the two concepts—look again at note 12 in Chap. 1 for how the connection between FTL and time travel appeared in pulp science fiction), but didn't show *how*. That was done 2 years later by Everett using, not surprisingly, an argument he called "reminiscent of the 'reinterpretation principle' . . . which played an important role in discussions of the physics of tachyons."[91]

In an attempt to avoid the Alcubierre bubble's causality problem, Krasnikov looked for a different, causal superluminal spacetime metric. This he succeeded in finding[92] but, rather than describing a bubble, Krasnikov's warp is in the shape of a tube. The interior of the tube is flat spacetime, just as in the case of the bubble warp, but unlike the bubble there would be a causal link between the spaceship crew and the tube. Just as the warp bubble requires very thin walls (on the order of a few thousand Planck lengths) of negative energy, so does the Krasnikov tube warp. Unlike the bubble warp, however, the tube warp stretches the entire length of any proposed trip, so the total negative energy in the warp is incredibly huge. For a tube a mere one meter long and one meter wide, for example, the total negative energy is 10^{28} solar masses, and to create a tube from Earth to just the nearest star would require 10^{44} solar masses of negative energy![93]

One curious feature of the Krasnikov warp is that the *outbound* leg of a round trip cannot be made in less time than required by light. *But* on the on the return half of the journey, a traveler would find the spacetime metric so altered (because of mass-energy manipulations purposely made on the outbound half) that she would move "backwards in time." The net result is that the round trip could end arbitrarily soon after it started! As Everett and Roman cautiously concluded, the Krasnikov tube is a "very unlikely possibility," but it would make a *wonderful* science fiction gadget, don't you think?

---

[90] A. E. Everett and T. A. Roman, "Superluminal Subway: the Krasnikov tube," *Physical Review D*, August 15, 1997, pp. 2100–2108.

[91] A. E. Everett, "Warp Drive and Causality," *Physical Review D*, June 15, 1996, pp. 7365–7368. Recall the discussion of the RP in Chap. 5.

[92] See notes 84 and 90.

[93] In the same manner as the huge negative energy of the Alcubierre warp drive was later reduced (see note 88), the Krasnikov tube's enormous negative energy requirement was later significantly reduced: see P. Gravel and J. Plante, "Simple and Double Walled Krasnikov Tubes I: tubes with low mass," *Classical and Quantum Gravity*, February 2004, pp. L7–L9.

While the Alcubierre warp may seem to be an incredible discovery (it *is*), it was not a unique one. That's because just 8 years later a different warp metric was discovered by the Portuguese mathematician José Natário, in which the expansion/contraction of spacetime does *not* occur. As Natário wrote, this signature feature of the Alcubierre warp drive "is but a marginal consequence of the choice [for the spacetime metric/mass-energy distribution]."[94] Rather than thinking of the warp bubble as being propelled by the push-pull of spacetime expansion/contraction, Natário wrote that "one could best describe the warp-drive spacetime as 'sliding' the warp-bubble region through space": that is, as analogous to a California surfer riding a wavefront. The surfer is motionless with respect to the water in the immediate vicinity of his board, and yet his speed with respect to the rapidly approaching shore is decidedly non-zero.

The idea of a manipulated or *warped* spacetime allowing time travel was an early arrival in science fiction. Consider, for example, the 1930(!) story in which the narrator (one Thomas Jenkins) walks 18,000 years into the future. In an editorial footnote (a device commonly used in early pulp fiction to inject scientific verisimilitude), we are told that "Jenkins had evidently fallen into a warp in space . . . a fault, we might say, borrowing a geologic term, in the curvature of space. Through this warp he had been thrown clear out of our three dimensions into a fourth. There he slid in time over to the other side [of the fault] into the same spot in the three-dimensional world, but into a different era in time."[95]

That was a flawed explanation, with its talk of space rather than of spacetime, but some authors eventually learned to do better. For example, *folded spacetime* as a mechanism for time travel is used in a 1940s cautionary tale on the potential horrors of the atomic bomb. In that story, published 2 years after the atomic bombings in Japan, the world 15 years hence experiences a terrible atomic war. As the time traveler in the tale explains, "During the unprecedented release of atomic energy that arouse during the simultaneous bombings of our cities, something happened to the very continuum in which we exist . . . A crook, a twist, a fold—explain it how you will, I accidently stumbled upon an electronic circuit that would create a field that would enable passage from one folded section [of spacetime] to the adjacent section. The fold proved to be about fifteen years in length . . . ."[96]

---

[94] J. Natário, "Warp Drive with Zero Expansion," *Classical and Quantum Gravity*, March 2002, pp. 1157–1165.

[95] N. Schachner and A. L. Zagat, "In 20,000 A.D.," *Wonder Stories*, September 1930.

[96] R. F. Jones, "Pete Can Fix It," *Astounding Science Fiction*, February 1947.

As Alcubierre and Natário showed, it will take a lot more than a mere electronic circuit to warp spacetime for either FTL or time travel, but at least even the early science fiction pulp writers understood that—*somehow*—a spacetime warping would be required.

The Alcubierre FTL warp drive appeared as the scientific basis for a modern science fiction novel, where at one point we read of a curious optical feature of an FTL spaceship, one I haven't seen mentioned in the physics literature: "This is a ship that traveled faster than light. It's visible as it travels; its warp bubble emits a cascade of exotic radiation . . . but it outruns its own image. So the ship arrives first and the light has to catch up, all the photons it emitted back along its path arriving at mere light speed. The older images arrive last, and you get this effect as if the ship was receding, not arriving."[97]

Another quite interesting feature of Alcubierre's warp drive is that the spaceship crew would experience no acceleration forces, as the ship is always in *free fall*. This may explain why the *Enterprise* crew isn't flattened when Mr. Sulu engages that ship's warp drive. The spaceship is surrounded by a "bubble" of warped spacetime that is swept along by the combined push-pull effect of the expanding spacetime behind the craft and the shrinking spacetime in the front. The ship, itself, resides in the *flat* spacetime interior of the warp bubble. An amusing way to think of this is to imagine a fish (space traveler), inside an aquarium (the warp bubble), which has been tossed into a swiftly flowing river. An observer at the edge of the river sees the aquarium move by her at high speed while, for the fish (swimming in the *still* waters of its aquarium), all is serene because *it is at rest with respect to its local environment*. Thus, the Alcubierre warp drive realizes yet another one of science fiction's wonderful gadgets: the *reactionless* spaceship drive. That is, "the warp bubble moves by interacting with the geometry of spacetime instead of expending reaction mass [as do jet and rocket engines] . . . and the spaceship is simply carried along with it."[98] In picturesque terms, the warp drive starship is like a surfer who makes her own waves.

And so we see, with each passing decade, more and more of science fiction departing from the make-believe to the pages of physics journals.

---

[97] S. Baxter, *Ark*, Gollancz 2009. The spaceship in this work travels at three times the speed of light (warp factor 1.44, as explained in note 84).

[98] F. S. N. Lobo and M. Visser, "Fundamental Limitations on 'Warp Drive' Spacetimes," *Classical and Quantum Gravity*, December 2004, pp. 5871–5892.

6.6 For Further Discussion

The connection between FTL speeds and backward time travel made the jump from theoretical physics to popular culture very quickly. It was in the British humor weekly *Punch*, for example, that the famous (but nearly always misquoted—see note 113 in Chap. 3, which doesn't have it quite right) limerick by A. H. R. Buller (1874–1944) first appeared (December 19, 1923, p. 591):

> *"There was a young lady named Bright Whose speed was far faster than light, She set out one day In a relative way And returned on the previous night."*

Where *Punch* dared to go, Hollywood could not be far behind. Indeed, in this case it was actually there first, with the 1922 one-reel silent comedy film *The Sky Splitter*. This was just a short film (feature pictures generally had at least four reels), so it isn't clear just how widely distributed and viewed it may have been. The story is that of a scientist testing a new spaceship; when it exceeds the speed of light, he begins to relive his life. This all shows that today's fascination, so common in popular culture, of the latest developments in theoretical physics, is nothing new. Why do you think this is so? That is, why (for example) do so many of those who flock to science fiction movies of interstellar invasions (like the 1996 *Independence Day* and its 2016 sequel), nonetheless have no conception of the unlikely possibility of such invasions because of the sheer magnitude of interstellar distances? Distances so immense that, even at the speed of light, it takes 4 *years* to travel to the Sun's nearest stellar neighbor, and *millions* of years to reach the Milky Way's nearest neighboring galaxy? (The vastness of interstellar distances is, as mentioned in the text, the reason for the fascination in warp drives in both science fiction and physics.).

In wormhole and cosmic string time machines, and with warp drives, we encountered the idea of negative mass-energy in the form of 'exotic matter' (see note 39 again). Something like negative mass actually appeared in fiction long ago, in the 1827 novel *A Voyage to the Moon* by "Joseph Atterly," a pseudonym for George Tucker, a professor of moral philosophy at the University of Virginia. (One of Tucker's students was Edgar Allen Poe, who almost surely was influenced by Tucker's book to write his own moon tale, the 1835 "The Unparalleled Adventure of One Hans Pfall.") The trip in Tucker's work was powered by a metal called *lunarium*, which repels Earth.

(continued)

This is *not* the same sort of stuff as Wells' "Cavorite," a metallic alloy that is "transparent" to gravity and that appears in his 1901 *The First Men in the Moon*. Wells' competitor in the 'scientific romance' genre was, of course, Jules Verne. Wells was a visionary who looked far beyond just the next few decades, while Verne was a 'practical engineer' who, for example, got *his* characters to the Moon by the direct method of simply shooting them out of a 900-foot long cannon with 400,000 pounds of guncotton! (Wells' vision *could* sometimes fail him, as it did about the imminent likelihood of airplanes in his 1901 *Anticipations*. He believed they *would* be developed by the year 2000, and maybe even before 1950, but of course just 2 years later . . .) In a 1903 magazine interview, Verne revealed how he felt about the difference between his and Wells' work: "It occurs to me that his stories do not repose on very scientific bases . . . He goes to Mars [sic] in an airship, which he constructs of a metal which does away with the law of gravitation. C*'est très joli* [this is all very nice], but show me the metal. Let him produce it." Today the cry from those who dislike wormholes is the Verne-like 'show us the exotic matter!' If Wells and Verne were writing today, how do you think each would respond to that challenge? Would the possible existence (or not) of exotic matter be an issue about which both would *agree*?

Write a time-loop short story based on Lossev and Novikov's idea of a 'very old spacecraft' interacting with a remote wormhole.

Imagine an electronic circuit **A** that has the following behavior: **A**'s input signal is a function of time that has a well-defined maximum value (what electrical engineers call the *peak value*). The circuit's output signal, produced in response to the input, also has a well-defined peak value. Now, imagine further that the output peak occurs *before* the input peak. There is, in fact, nothing paradoxical or impossible about that, and such a circuit can (and has) been constructed, as I'll tell you shortly. Next, suppose that we take **A**'s output signal and use it as the input to another circuit **B** that, when it's input exceeds a certain level, disconnects the input to **A** *before* that input reaches *its* peak value. Circuit **B** can also be constructed in the real world. Indeed, you can read about how to construct **A** and **B** in two papers by M. W. Mitchell and R. Y. Chiao, "Causality and Negative Group Delays in a Simple Bandpass

(continued)

Amplifier," *American Journal of Physics*, January 1998, pp. 14–19, and "Negative Group Delay and 'Fronts' in a Causal System: An Experiment With Very Low Frequency Bandpass Amplifiers," *Physics Letters A*, June 16, 1997, pp. 133–138. What makes all this interesting here is that this "seems to open the way for a variant of the time travel paradox in which the traveler journeys to the past and kills his grandfather before his own father is born," an observation made in Garrison *et al.*, "Superluminal Signals: Causal Loop Paradoxes Revisited," *Physics Letters A*, August 10, 1998, pp. 19–25. This electronic version of the grandfather paradox does indeed follow if one substitutes "input peak" for "grandfather" and "output peak" for "time traveler." But before you think this gadget is a time machine, be assured that its designers also showed that, unlike the causally related grandfather and time traveler, the two peaks are *not* so related. Read these three papers and write a summary report of how circuits **A** and **B** work, and why the two peaks are not so related.

You'll recall that "an advanced civilization" is thought to be required to create a useable wormhole (note 27). The common phrase used by astrophysicists who are interested in the possibility of extraterrestrial life is *arbitrarily advanced civilization*, with a distinction made for at least three progressively higher stages of 'advancement.' Very roughly, Types I, II, and III advanced civilizations are those that, respectively, have the technology to (a) control something like 10^{13} W (ten million megawatts) for massive interstellar radio broadcasts, (b) a technology to control the energy output of the civilization's planet's parent star (10^{27} W), and (c) a technology to control the energy output of the civilization's home galaxy (10^{38} W). We are, today, short of being even a Type I civilization, and it would probably take *at least* a Type III civilization to build a wormhole. Indeed, Stephen Baxter's 1993 novel *Timelike Infinity*, of beings who can manipulate *constellations of galaxies*, seems to assume a Type IV civilization will be required. Since there are typically 10^{11} stars in a galaxy, going from 10^{27} W for a star to 10^{38} watts for a galaxy is consistent. But where do astrophysicists get 10^{27} W for a single star? Here's a calculation for you to perform, to confirm this value for yourself, starting with the experimental fact (not difficult to repeat, as it's at the level of a junior high school science fair project using a solar cell and a few common electrical components) that the solar power level at Earth's equator is 1200 W per m^2. Then, using the fact that the Earth's orbital radius around the Sun is 93 million miles, compute the total power (energy per second) radiated by the Sun. (You should get a number that is somewhat

(continued)

smaller than 10^{27} W, which is ‘explained’ by observing that the Sun is really a quite ordinary star, exceeded in size by many other stars in the Milky Way. Next, you’ll find, in most books on astrophysics, the statement that the nuclear fusion reactions that power the Sun convert *four million tons* of the Sun’s mass to pure energy *every second*. Confirm that your number for the power output of the Sun is consistent with that claim. (Remember Einstein’s famous formula $E = \mathrm{m}c^2$, that the speed of light is $c = 3 \times 10^8$ m/s, and that 1 kg $\approx$ 2.2 pounds. In the MKS system of units (meters/kilograms/seconds) one watt = one joule (of energy) per second, where to give you some perspective on what a joule is, the chemical energy released by burning a gallon of gasoline is about 100 MJ.

Appendix A
Old Friends Across Time (A Story)[1]

As I sit here in my study, with the photographic evidence spread before me, I can barely comprehend what my eyes tell me must be so. The evidence is incontestable. And yet—I still struggle to believe. Let me try to explain—possibly in the process I will manage to put my tumbling mind to rest.

For as long as I can recall, old photographs have fascinated me. To page slowly through collections of historical pictures, no matter what the theme, was consummate joy. Even when I was quite a small boy I used them as my time machine into the past. They took me up and away from the problems every youngster has while growing up, and let me wonder of people and places long since returned to dust. Matthew Brady's Civil War photos had a particularly strong attraction for me, with the horror (and yes, I will admit it, the *fascination*) of war frozen in the images of young men dead before life had really begun. To look at the fallen youth of more than a century before, and to wonder who they were, and what they had felt and thought—it all sent shivers through my romantic mind.

I suppose I might have become a professional photographer. But somewhere along in the process of looking at pictures, I became aware of the miracle of the *technology* of picture taking. That led me to chemistry and optics, and finally by some wondrous route, I became an electrical engineer. I never lost my love for old pictures, though, but merely turned my interest in them to the photographic history of electrical physics.

To search out and acquire (for by now I had started my own collection) a photograph of Steinmetz, smoldering cigar clamped in his mouth, giving a lecture on AC circuit analysis using the then still mysterious square root of minus one made my heart beat faster. To find a faded picture of Einstein at a long forgotten

---

[1] P. J. Nahin, "Old Friends Across Time," *Analog Science Fiction Magazine*, May 1979. This tale was written with the specific goal of illustrating how a trip into the past *yet to be initiated* could logically influence events in the time traveler's present and future. The story reproduced here is, with only a few very minor alterations, as it originally appeared in *Analog*.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

conference, caught forever in time with his quiet, gentle eyes looking into mine, would send me to the heights of ecstasy.[2]

But it was Maxwell that led me to my incredible discovery. There is no doubt but that James Maxwell was the greatest theoretical physicist of the nineteenth century. Together with Einstein, he was the best of *any* century. Could it possibly be more than mere chance that the same year saw the death of one and the birth of the other? It was Maxwell who gathered together all the then known, but fragmented, experimental bits and pieces of knowledge about electricity and magnetism, and stirred in his own contribution of the displacement current. There was no physical evidence then to justify that last step, but the genius of Maxwell knew it *had* to be. And then, from his soaring mathematical insight and physical intuition, he took it all and wrote down the four magnificent equations for the electromagnetic field![3]

No one who has seen and understood those beautiful equations can come away without a quickening of the pulse and a flush of the blood. They're not long—you can write all four vector differential equations on the back of a postcard, but oh, what they tell us! With them, Maxwell, showed light was electrical in nature, predicted radio waves *two decades* before Hertz discovered them in the lab, explained energy propagation in space, and radiation pressure, and laid the scientific basis for today's television, radar, lasers, giant electric motors, generators, transmission lines and—well, why go on? The equations are the work of a level of genius we may not see again for a millennium. We have hardly begun to discover the marvels wrapped inside the electromagnetic field equations. With their aid, and that of quantum mechanics, the very secret of life, itself, may someday be unraveled.

And so I searched for old photographs of Maxwell. He died at his family's Scottish home in 1879, before the art of picture taking was barely 40 years old. But I knew in my heart that somewhere there *must* be photographs, yet undiscovered, of such a great man. Anyone who has seen the best examples of prints from wet glass collodion negatives knows they are, in the faithfulness of their rendition of detail, better than what we commonly expect today. Working against me was the fact that the process was slow, laborious, and unforgiving of mistakes. The taking of a picture was not a minor decision in Maxwell's time. But still I searched.

I searched for one photo, in particular. When Einstein died, a famous picture was taken of his office, just the way he left it for the last time. On the blackboard behind his desk are the last thoughts he had in his long quest for a Unified Field Theory, a ‘theory of everything.’ The writing on the papers covering the desk is clearly legible, and with modern blowup methods, easily readable.

---

[2] The first reference is to Charles Steinmetz (1865–1923), the German-born American electrical engineer and mathematician who became the wunderkind of General Electric. Einstein, of course, needs no introduction!

[3] I wrote this for story effect, but it's not really *quite* true. When Maxwell wrote his theory in mathematical form, he did so using *twenty* (!) equations in as many variables. The equations, as physicists and electrical engineers use them today, were first written in 1885 by the English self-taught eccentric Oliver Heaviside, who considered Maxwell to be his hero (see note 6 in Chap. 6). Modern electrical engineers and physicists write the Maxwell equations as *four* partial differential *vector* equations.

At the time of his death, Maxwell was the Einstein of his times. Surely, I reasoned, a similar photograph of Maxwell’s study must have been taken. Even though none has come down through the decades to us, it *must* exist! Gathering dust in an old trunk, or buried in a long forgotten album, it had to be somewhere. I vowed to find it.

I began by writing to all of Maxwell’s living descendants, asking that they search through family holdings for any pictures concerning Maxwell that they might possess. For the most part all were cooperative, even though more than just a few thought I was somewhat deranged. Still, it was in vain. I did receive a few old pictures never before seen by other than the family, including a poignant one taken in 1901, showing Maxwell’s grave in Parton Churchyard at Glenlair, Scotland. A forlorn, wintry scene, with only what seemed to be three men in the far distance, it brought tears to my eyes. Alas, there were no photos of Maxwell’s study.

But then late last year, while on a business trip to London, I stopped off for a few hours at the historical archives maintained by the British Institute of Electrical Engineers. On a chance, I looked through their massive files on Maxwell and was rewarded within the hour! What I found will haunt me throughout the remainder of my life.

There it was, stuck through its border with a rusty pin, between two pieces of yellowed paper covered with what appeared to be some simple, rough lecture notes. An ordinary looking photo of a study. Obviously overlooked through the years, or at best unappreciated for what it was, it was the almost illegible, penciled notation on the back that convinced me of my find—just a date: November 9, 1879. Exactly 4 days after Maxwell’s death, precisely when some unknown, yet inspired person (a family member, a neighbor, a local scientist?) would take such a picture!

I am ashamed to admit it, but there was no hope the Institute would let me have the picture. And there was no time to copy it, for I was to return home to America that very night. No, that’s not true. The *real* reason for what I did was simply that I *had* to have that original, *old* photo. I took it! It was my undoing, for that dishonorable act destroyed the picture’s tie to verified, legitimate historical records. But *I* know what I found is true.

I could barely control my wild emotions on the flight home. Several times I removed the picture from my briefcase, and looked with fascination at the papers lying on Maxwell’s desk, and at the tightly written lines of mathematics on the blackboard in the background. My hands trembled with what can only be called lust—once home, reunited with my well-equipped photo lab, I would learn every secret hidden in that picture.

There are no words I know that can convey the thrill I felt as I began the processing of that priceless photo. Alone in my lab, with all the modern equipment a well-off amateur can buy (a Caesar Saltzman 8 × 10 enlarger with mercury vapor point light source and a 10× Plan Achromat Nikon enlarging lens), I carefully cropped and blew up selected views of the blackboard and desk. Printing the enlargements on ultra-fine grain AGFA Brovira paper, I could scarcely restrain myself from peering at them with a magnifying glass while I waited for them to dry.

Then, at last, I had them spread out across my study desk. I tried to force myself to examine each slowly, carefully, in turn, and not to skip from one to another like a child let loose in a candy store with a dollar. The first three were of the desk papers, including what seemed to be a diary. It must have been lost after Maxwell's death since no trace of it exists in the historical records. I experienced a stunning thrill as I gazed upon the scrawled words, but as they were not easily read at once, I moved on. It was the sixth enlargement, of the upper right corner of the blackboard that sent me reeling back to my chair. An equation that shouldn't, no, *couldn't*, be there. But it was.

To understand my reaction, there is one astounding thing you must realize about Maxwell's field equations. When Einstein turned the world of physics on its head in 1905 with his famous paper, "On the Electrodynamics of Moving Bodies," all the old ideas about absolute motion and simultaneity of events went out the window. Even Newton's laws of mechanics had to be modified. But *not* Maxwell's! His equations, just the way he published them in 1873, are the same ones studied today[4]—they need *no* relativistic corrections.

How can that be, you wonder, as they predate Einstein's by 32 years? The mystery of this has bedeviled the experts down through the years. Oh, they have an explanation, alright. They say that all of electromagnetics is actually relativistic phenomena to begin with, and the laboratory work of Faraday, Ampere, Henry, and the other great experimentalists were studies of relativistic electron interactions in matter (although they, of course, didn't know that). Thus, it is only 'natural' that Maxwell's equations need no correction. So goes the 'expert' explanation, but it isn't right![5] I know Maxwell knew about relativity, and understood it perfectly. He knew all about time paradoxes and the equivalence of mass and energy.

Because how else can you explain the equation visible in my enlargement: $E = mc^2$!

Why, you must wonder (just as I did), didn't Maxwell publish this remarkable result? At first, I believed it was because of a lack of faith in his results. Who would have believed any of it in those Victorian times, so sure of its absolute view of nature? I thought of how Newton, 200 years before Maxwell, had suffered from a similar hesitancy when he wrote the *Principia*. There, when explaining his theory of gravitation, Newton did *not* employ his new invention of the calculus (which he *had used* to make his discoveries), but instead fell back on laborious arguments based on the accepted mathematics of algebra and geometry. Who would have believed him, otherwise?

But then I realized that couldn't be right. Maxwell was a strong man intellectually and he wouldn't have held back for fear of disbelief. No, it had to be that he discovered relativity and the mass-energy law just before his death, with no time to make his work known. I was still wrong.

---

[4] Don't forget note 3.

[5] Alas, I think it *is* right. Don't forget, this is science *fiction*. When there is a conflict between the needs of a story, and a rigid adherence to physics, the 'needs' wins!

It was later, when I returned to the enlargement of Maxwell's lost diary and read those painfully cramped notes, that I learned the truth. What I saw there showed me Maxwell had thought long and hard about his final discoveries and had purposely withheld them. For clearly visible, after I had slowly deciphered the writing, were the following words:

> *I have seen monstrous events. My blood has run cold at the sight of two great cities leveled to the ground, their inhabitants cruelly put to death instantly, or left to die slowly from a strange, lingering disease. Other trips, further on, have shown me the root of all these evils is the mass-energy equation, a result I at first believed to be my crowning glory. It will be my crown of thorns unless I ban it from my very being. Another will discover it for himself, but* my *soul shall be free! I have dismantled my machine, and shall never look upon or think of those horrible scenes again.*

This passage was dated just 1 month before Maxwell died a savage death from cancer. The reference to 'two cities' can only be that of Hiroshima and Nagasaki. His own death was surely caused by lingering too long among their atomic ruins.

Think of what this *means*. Quite simply, Maxwell knew the secret of time travel! But even more incredible is that it must be *easy*, if one only knows how, to build a time machine! Think about it—Maxwell had no gigawatt power stations at his disposal, no high technology machine shops, or nanosecond computers. He was not a gifted experimentalist, and once he had predicted radio waves, for example, it took others 20 years to finally generate them. And yet, *he* built a time machine. Somehow, with just the puny power sources available to him, and a limited mechanical capability, he wrested free the *simple* implementation of a time machine from his dynamical field equations.

Yes, yes, I know what you must be thinking. How can I really conclude such an incredible thing from a single equation on a blackboard, and a few words written by a man dying a painful death? A man, clearly suffering dearly, and possibly not in complete possession of his once marvelous mind.

This very evening the last bolt of evidence slid into place. Attempting to escape from the emotional maelstrom into which I had fallen, I turned to my old love of picture gazing. I took down from my library shelf a tattered yet cherished volume of the Meserve Collection of Lincoln pictures. My slow paging through the images stopped when I came to the famous photograph by Alexander Gardner of Lincoln's second inauguration. This incredible picture shows John Wilkes Booth looking down on Lincoln from behind a buttress high on the steps of the Capitol, while below in the crowd are the five men who, 41 days later, conspired with him in the assassination.

The following page demonstrated the extraordinary quality of Gardner's work, as it showed an enlargement of Booth's face in which the circular line between the pupil and the white of each eye is sharp and crisp! This impressive picture fascinated me, and I wondered if I could create a similar enlargement. It was then I remembered the old picture of Maxwell's grave, sent to me from Scotland, and the three distant figures in the background. They would present my photo-lab skills with a challenge, and the effort would distract my mind.

I finished the enlargement just 20 min ago. Those faces! Two of them I can now finally accept as being there—it must have been a pilgrimage for one, and for the other, it couldn't have been anything but a mocking, ironic gesture. But I wonder if the youngest one really knew who his two companions were? I don't know the answer to that—yet. But there they are, two men with faces my years of study have made as familiar to me as my own. One is a youthful Albert Einstein. The other, with the signs of death clearly written across his features, is James Clerk Maxwell. The face of the third man is familiar, too, for the third man is me!

Oh, I'm a bit older in the photo than I am now. But it's me, alright. A distinctive, jagged scar across the left cheek, a mark from a childhood accident, is sharply visible, and I can run a finger over my face and match it perfectly with the image in the enlargement. I'd say I'm about 45 or so in the image, no more than 10 years older than I am now. That doesn't leave me much time to keep my appointment, does it? I don't know, right now, how I'm going to do it, but I've got to rediscover Maxwell's secret of time travel. I'm sure I'll succeed—after all, there I am in the picture. Somehow, I'll be going back to pick James and Albert up so we can have our picture taken. Ten years—not much time.

I'm really looking forward to meeting my two new friends from across time.

For Further Discussion

When "Old Friends Across Time" originally appeared in *Analog*, it opened with a quotation from Richard Feynman's famous 1961–1963 Caltech undergraduate course (published in 1964 as *The Feynman Lectures on Physics*): "Ten thousand years from now, there can be little doubt that the most significant event of the nineteenth century will be judged as Maxwell's discovery of the laws of electrodynamics. The American Civil War will pale into provincial insignificance in comparison." This is almost certainly true, but could Maxwell *really* have built a time machine from just a knowledge of electromagnetic theory and special relativity (which is all that is needed to derive $E = mc^2$), if he didn't also have a deep understanding of general relativity (and probably of quantum mechanics, too)? How likely do you think *that* is?

The narrator in “Old Friends Across Time” knows he is going to live long enough to eventually build a time machine; discuss the implications of this knowledge. For example, is he at least *temporarily* invulnerable to committing suicide (or, for that matter, to any other variation of dying?) That is, do we have a ‘future’ version of the grandfather paradox? This issue has never (to my knowledge) been considered by physicists, and not by philosophers either until recently. See, for example, S. Keller and M. Nelson, “Presentists Should Believe in Time-Travel,” *Australasian Journal of Philosophy*, September 2001, pp. 333–345, and M. H. Slater, “The Necessity of Time Travel (On Pain of Indeterminacy),” *The Monist*, July 2005, pp. 362–369. More generally, if we assume that the past is unchangeable then the scenario in “Old Friends Across Time” seems to force at least some level of inevitability on the future as well. Or does it? In the 2007 story by Ted Chiang, “The Merchant and the Alchemist’s Gate,” that you were asked to read in a *For Further Discussion* at the end of Chap. 4, there is the following exchange between the narrator and the inventor of “the Gate” (a wormhole): “So if you learn that you are dead 20 years from now, there is nothing you can do to avoid your death?” He nodded. This seemed to me very disheartening, but then I wondered if it could not also provide a guarantee. I said, “Suppose you learn that you are alive 20 years from now. Then nothing could kill you in the next 20 years. You could then fight in battles without a care, because your survival is assured.” “That is possible,” he said. “It is also possible that a man who would make use of such a guarantee would not find his older self alive when he first used the Gate.” “Ah,” I said, “Is it then the case that only the prudent meet their older selves?” Comment on this issue, with particular attention to free-will.

Appendix B
Newton's Gift (A Story)[6]

Wallace John Steinhope was a sensitive human being, a person deeply concerned about the welfare of his fellow creatures. Any act of injustice, however slight, made his breast pound with righteous indignation. He was a champion of fair play, and his motto in life was taken from the ancient English rule of law—'Let right be done!'

Even while still a lonely, reclusive child, Wallace's heart ached mightly when he read of the laborious, boring, mind-dulling calculations endured by the great mathematicians of old. Just knowing, *thinking*, of Gauss's marvelous mind wasting literally months of its precious existence grinding out tedious mathematics that even a present-day dullard could do in a minute, on a home computer, was sheer agony for Wallace. Contemplation of the God-like Newton suffering endless delays in his gravity research, all because of a simple miscalculation of the length of a degree of longitude, was almost unbearable.

Indeed, Newton played a special role in Wallace's life (and he in Newton's, as we shall soon see). While the other great mathematical physicists had merely been hindered in their work by the lack of modern computational aids, Newton had squandered so much valuable time in other, nonscientific pursuits! His quasireligious writings alone, over half a million words, exceeded his scientific writings. What a waste! Wallace wondered endlessly over the reason for this strange misdirection of talent and bored his friends to the edge of endurance with his constant brooding on the mystery. Still, they all liked and admired Wallace enormously and so put up with it. But more than one of them had sworn to throw up the next time Wallace mentioned Newton during a wedding (but that's another story).

So deep was Wallace's anguish for his predecessors that even as he grew older and his own tremendous talents as a mathematical physicist (the result of a lucky

---

[6] P. J. Nahin, "Newton's Gift," *Omni*, January 1979. This tale was written with the specific goal of illustrating casual loop time paradoxes. The story reproduced here is, with only a few very minor alterations, as it originally appeared in *Omni*.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

genetic mutation induced in a male ancestor some centuries earlier) gained him an international reputation, thoughts of the unmeasurable misery of his scientific ancestors were never far from his mind. It was most appropriate, then, that his greatest discovery gave him an opportunity to *do* something! And Wallace John Steinhope vowed to *help*. He became convinced that it was his purpose on earth—he could not, he *would* not hesitate. As he strapped the knapsack-size time machine to his chest, his excitement was, therefore, easy to understand.

"It is done! And I am ready. I will travel back and bestow this gift of appreciation, this key to mental relief, on the great Newton himself!" Wallace cradled a small, yet powerful hand-calculator in his palm. It was a marvel of modern electronics. Incorporating large-scale integrated circuitry and a Z-8000 microprocessor solid-state chip, the calculator required only a small, self-contained nuclear battery for its power. It could add, subtract, multiply, divide, do square and cubic roots, trig and hyperbolic functions, take powers, find logarithms, all in mere microseconds. It was programmable, too, able to store up to 500 instructions in its micro-memory. The answers it displayed on its red, light-emitting diode readouts would liberate young Isaac from the chains of his impoverished heritage of mathematical calculation. No more Napier's bones for Newton!

But Wallace John Steinhope was no fool. He understood, indeed feared, time paradoxes. He knew Newton could be trusted with the secret, but it wouldn't do for the calculator to survive Newton's time. So Wallace had incorporated a small, self-destructing heat mechanism into it. After 5 years of use, it would automatically melt itself into an unrecognizable, charred slag mass. But that would be enough time for its task to be completed. The emancipation of Newton's mighty brain from tedium! Pleased enormously at the thought of the great good he was about to confer, Wallace set the time and space coordinates for merry old England, flipped the power switch on, and vanished.

Materializing in the Lincolnshire countryside in the spring of 1666, he began his rendezvous with destiny. It was the second and final year of the great bubonic plague, and Newton, seeking refuge from the agony and death plundering London and threatening his college of Trinity at Cambridge, had returned home to work in seclusion. The years of the Black Death were Newton's golden years, when the essentials of calculus would be worked out, when the colored spectrum of white light would be explained, and when the principle of the law of gravitation would be grasped. But how much easier it would be if Newton were released from the binding chains of dreary calculation. Wallace's gift would slip the lock on those chains! Accelerate genius!

It was early evening when, guided by a map of the area prepared by a friend who was both a cartographer and amateur historian, Wallace reached the quiet little town of Woolsthrope-by-Colsterworth. It was here, in a small farmhouse, that Wallace would meet his hero of the ages. A cold, gentle rain was falling as he approached the door. The soft, hazy light of an oil lamp glowed inside, revealing through the translucent glass the form of a man bent over a table. The fragrant smoke of well-dried wood curled from the chimney, announcing a warm fire within.

With his heart about to burst from excitement, Wallace rapped upon the door. After a pause, the shadow rose and moved away from the window. The door opened, and there stood Isaac Newton, a young man of 23 with an intellect that Hume and Voltaire considered "the greatest and rarest genius that ever rose for adornment and instruction of the species." But for the importance of his self-appointed mission, Wallace would have fainted dead away from the thrill of it all. "Is this the home of Isaac Newton?" he asked in a voice quavering with the trembling tones normally used by lovers about to reveal their deepest feelings.

The young man, of medium height and thick hair already showing signs of gray, swung open the door and replied, "My home it is, indeed, stranger. Come into the parlor, please, before the wetness takes you ill."

Isaac followed Wallace into the room and stood quietly watching as his visitor removed his soaked coat and hat. The portable time machine was gently placed on the floor next to a wall. The calculator was snug and safe in its plastic case in Wallace's shirt pocket. "Thank you, Master Newton. May we sit while we talk? I am afraid you may wish to take some time to consider my words." Motioning to a chair near the table, Isaac pulled a second chair from a darkened corner and joined Wallace. "You have a strange sound to your speech, stranger. Are you from hereabouts, or have you traveled far? Please commence slowly your tale."

Wallace laughed aloud at this question, a response prompted by his nervous excitement, and it quite surprised him. "Please forgive me. It is just that I *have* traveled so very, *very* far to see you. You see, I am from the future." Wallace was not one to play his cards close to his chest. Now it was Isaac's turn to laugh. "Oh, this is most ridiculous. Are you a friend of Barrow's at Trinity? It would be so like him to play such a trick.[7] From the future, indeed!"

Wallace's eyes ached at the sight of the papers on the table where Isaac had been working. What wonders must be there about to be born! In any other situation, Wallace would have asked their contents, but the die had been cast. He had to convince Isaac of the truth of his tale. But he had to walk a tight line, too. It just wouldn't do to misdirect Isaac's interest away from the calculator and toward the time machine itself! He must do something dramatic, something that would rivet his idol's attention and hold it.

"Yes, yes, I understand your reluctance to believe me. But, look here. This will convince you of the honesty of my words." Wallace pulled the shiny black plastic-cased calculator from his shirt pocket and flipped the power switch on. The array of LEDs glowed bright in the gloomy room as they flashed on in a random, sparkling red burst. Isaac's eyes widened, and he pushed his chair back. Was he frightened?

"As the Lord is my Savior, is it a creation of Lucifer? The eyes of it shine with the color of his domain. Are you one of his earthly agents?"

---

[7] The reference is to Isaac Barrow (1630–1677), who was the first Lucasian Professor of Mathematics at Cambridge. Barrow resigned that position to allow it to pass to Newton. Centuries later, Hawking became the 17th Lucasian Professor.

"Oh my, no! Look here, Master Newton, let me show you that there is no black magic or chicanery involved. It is all perfectly understandable in terms of the laws of Nature. What I have here is an automatic calculator, a device to perform all of your laborious mathematical labors."

So saying, Wallace squeezed the sides of the calculator case together, releasing pressure snap-fittings, and flipped the case open on a hinge at the top. Revealed to Isaac were the innards of the electronic marvel—a tightly packed interior of printed circuit boards, a mass of integrated circuitry, the small LED display, and the sealed nuclear battery. Isaac stared intently at the sight, and Wallace could see the natural curiosity of Newton's great mind begin to drive away the initial apprehension.

"But where are the gears, levers, springs, and ratchets to carry out the calculations? All I see is a black box with lights that glow red—how is *that* done; where is the lamp or candle to provide the light!—and many little isolated fragments of strange shapes. There is clearly nothing in your box that moves!"

"Oh, it is all done with electronics, Master Newton! The central processing unit has access to a solid-state memory that contains the decoding logic necessary to implement the appropriate algorithmic processes to provide the answers to the specific requests entered through these buttons. The actual performance of the box is achieved by the controlled motion of electrons and holes in suitably doped semiconductor material under the influence of electric fields induced—" Wallace, still overcome by his excitement, had rambled on wildly without thought of the essentially infinite technological gap that separated himself from Newton.

"Stop, stop," cried Isaac. "I understand only a few of the words you use and nothing at all of their meaning! But it is obvious that for calculations to be performed, mechanical work must be done, and that implies motion. Pascal's adding machine has shown the veracity of that. I say again, nothing moves in the box. How *can* it work?"

Wallace was embarrassed. The mistake of overlooking the hundreds of years of progress after Newton's time was one a child might make. "I am sorry, Master Newton. I'm going too fast for you." Isaac looked at Wallace with a frown, but Wallace failed to see the pricked vanity of the proud Newton. Going too fast, indeed!

Wallace prepared to lay a firmer technological foundation for Newton, but then he froze. It couldn't be done! Newton was a genius, certainly, but the task was still impossible. Wallace would have to tell him all about Maxwell's equations, Boolean algebra and computer structure, electronics, and solid-state device fabrication technology. It was just too much, and besides, there was the danger! The potential time paradoxes of all that knowledge out of its proper time sequence! What if Newton, in innocence, revealed some critical bit of knowledge out of its natural place in history? So, Wallace hesitated, but seeing the suspicion grow again in Isaac's eyes, he realized he had to do something, *anything*, immediately.

"You cannot deny your own eyes," answered Wallace. "Let me *show* you how it works. I'll divide two numbers for you with just the punch of a few buttons. Watch this." And, at random, he entered 81,918 divided by 123. Poor Wallace, of all the numbers to use, they were the worst. Within milliseconds the answer glowed brightly in fiery red characters. Wallace looked with pride at the result and then, already enjoying in his mind what he knew would be Isaac's amazement, he turned his eyes to the great man. What he saw made his spine tingle, and the gooseflesh stand high on his neck. Newton had fallen to his knees, with eyes bulging and hands raised as if in prayer.

"The mark of the Beast, it is the mark of the Beast! It is so written in the Book of Revelations—Here is wisdom. Let him that hath understanding count the number of the beast; for it is the number of man; and his number is six hundred three-score and six!" Rising to his feet, Newton fell back into his chair. "Your cursed box bears the brand of its master. There can be no doubt now, it *is* the creation of the fallen archangel!" Wallace was aghast at Isaac's violent reaction. The seventeenth century genius had now stumbled backward from his chair and had grasped a poker from the hot coals of the fireplace.

"Wait, please wait! Watch this; I'll multiply two other numbers together for you, watch!" Wallace quickly punched in the data, and then the answer gleamed steadily in burning red characters on the LEDs. Isaac's eyes first went wide with fear as he again saw the wizard electronics do their marvelous assignment, and then he shut them tight. Wallace was becoming desperate—this wasn't the way it was supposed to be! "Don't you see—imagine the tedious work, the mind-deadening labor this machine will save you from. And it is yours."

"Yes? But only for the exchange of my soul! That is always the Devil's price for his seductive gifts from Hell!" As Isaac shrieked these last words at Wallace, he raised the poker over his head. "Begone, you emissary of the Dark World! I know now you must be in the employ of the Father of the Antichrist, but the Lord God Almighty will protect me if I do not waver in my resolve. Begone, or I'll strike your brains out on the floor where you stand!"

Isaac's eyes were wide with fear, nearly rolling back to show all white spittle sprayed from his mouth as he yelled at Wallace, who stared in shock at the wild man who threatened him with death. "Please, please, *listen* to me, please! I beg you to understand—I am a scientist, just like you. The concept of the devil, and all it stands for, is contrary to everything I believe. How *could* I be in the devil's employ, when I don't even accept his existence? You *must* believe me!"

"Blasphemy!" screamed Isaac. "Your own words condemn you. To deny the reality of Satan in a sinful world is to deny that of God, too. Now leave my home, you dark beast from hell, or by the heavens above, *I shall destroy you*!" As he shrilled these words, Isaac brought the poker down in a wild swing that barely missed Wallace's head.

Struck dumb with confusion at the uncontrolled outburst, Wallace stuffed the calculator into his shirt, grabbed his hat, coat, and time machine and rushed from the house. As he hurried into the cold, wet night, he turned back, just once, to see Isaac Newton framed in the light of the open door. "Go, go, you foul messenger from the Lord of Evil! Back to your stinking pit of burning hell-fire! This is a house that honors the Divine Trinity and is no haven for the likes of you!"

Wallace rushed away into the blackness, the time machine bouncing unheeded upon his chest.

He ran, for how long he couldn't recall, until he fell exhausted next to a stream running heavy with the rain. Tears of rage, frustration, and shock streamed from his eyes. Rejected by the great Newton! Well, damn him! Wallace flung the calculator into the stream in his terrible anger and activated the return coordinates. He faded from Newton's world as quickly and quietly as he had come.

As for Isaac Newton, after having chased the Devil's messenger from his house, he returned on shaking legs to his desk. Pushing aside his rough calculations on the orbit of the moon around the earth, he swore to redeem himself in the eyes of the Savior. Somehow, he had been found lacking and had been tested. And the test was surely not over! He began to reapply his marvelous mind to determine the origin of his failure before the Lord God Jehovah. Taking quill in hand, he wrote the first of the many hundreds of thousands of words that his numerous religious tracts would devour from his allotted time.

Five years later, long after Newton had returned to Cambridge, a group of picnicking children were frightened when a nearby stream suddenly erupted into a geyser of steam. Moments later, the bravest (or most foolhardy) of the boys—who, by an astonishing coincidence that befits any good time travel paradox, would be Wallace's great-grandfather nine times removed—cautiously examined the streambed. All he found were some twisted, hot pieces of what he thought was a hard, black rock, and he tossed them back. They were all that was left of the calculator's nuclear battery. He did receive a tiny radiation dose from them, which caused a recessive genetic mutation that centuries later would suddenly appear as the cause of Wallace's genius, but otherwise the lad was unaffected. The incident was soon forgotten.

Well over 300 years later, Wallace John Steinhope reappeared in his own time. He was essentially the same man as before he left—kind, generous, and sensitive, and ready to come to the aid of any man or beast that might need help. As far as his friends were concerned, in fact, he was even improved (naturally, they didn't know what had brought about the welcome change but, if they had, they would have applauded it).

Wallace John Steinhope, you see, never again had another kind word for Newton, or for that matter, any words for him at all.

For Further Discussion

In his book *Travels in Four Dimensions: the enigmas of space and time* (Oxford 2003), the philosopher Robin Le Poidevin writes (p. 176) "But, as everybody knows, when a time machine leaves for another time it *disappears*." This is, indeed, how the time machine in "Newton's Gift" works; however, after reading *Time Machine Tales* do you think such behavior is in agreement or in conflict with general relativity? Defend your position.

"Newton's Gift" contains causal loops. Identify two of them, and discuss their role in the story (that is, are they central to the story or merely incidental?).

The idea of a time traveler visiting famous people in the past occurs fairly frequently in science fiction. In Ian Watson's "Ghost Lecturer" (*Isaac Asimov's Science Fiction Magazine*, March 1964), for example, the inventor of the "Roseberry Field" uses it to yank geniuses out of time to supposedly honor them, to let them know their lives had been worthwhile in the eyes of the future. But then he goes on to tell them—oh so kindly—where they had gone wrong or had fallen short of the mark, and of how much more we know nowadays. "You almost got it right, boy! You were on the right track, and no mistake. Bravo! *But* ..." Watson makes the interesting observation that one can easily imagine playing this pathetic game of 'second-guessing' history with scientists, but what could even the most talented modern do to upstage a Mozart or a Shakespeare? Most similar to "Newton's Gift," however, are (for example) Gregory Benford's "In the Dark Backward" (*Science Fiction Age*, June 1994) where Shakespeare and Hemingway are visited, and Jack McDevitt's "The Fort Moxie Branch" (*Full Spectrum*, October 1988) where Hemingway and Thomas Wolfe appear. Read these stories, and then compare/contrast their descriptions of how story characters react to the appearances of time travelers, to Newton's behavior in "Newton's Gift."

Appendix C
Computer Simulation of the Entropic Gas Clock

```
%gasclock.m/created by PJNahin for TIME MACHINE TALES(6/27/2015)
%This MATLAB m-file simulates the diffusion of gas molecules in a
sealed
%container by using the Ehrenfest ball exchange rules. The
simulation
%starts with n molecules (i.e., balls) of one type (i.e., black)
on
%one side of the container, and n more molecules of another type
(i.e.,
%white balls) on the other side. The two urns play the roles of the
%two sides of the container. To simulate the ball (molecule)
%movements, the program selects two random numbers from 0 to
1, which
%are then compared to the current probabilities of selecting a
black
%ball from urn I and a white ball from urn II. If BOTH random
numbers
%are greater than these two probabilities then a white ball has
been
%selected from urn I and a black ball has been selected from urn
II,
%and so the number black balls in urn I is increased by one while
the
%number of white balls in urn II is increased by one. If BOTH
random
%numbers are less than or equal to these two probabilities then a
%black ball has been selected from urn I and a white ball has been
%selected from urn II and so the number of black balls in urn I is
```

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

```
%decreased by one while the number of white balls in urn II is
decreased
%by one. If one of the random numbers is greater than its
corresponding
%probability while the other random number is less than its
%corresponding probability, then no action is taken because then a
%white (black) ball moves from urn I to urn II at the same time a
white
%(black) ball moves in the opposite direction. That is, there is
no
%net change. Then, the ball selection probabilities are recalcu-
lated and
%another ball exchange is simulated.
rand(’state’,100*sum(clock))     %new seed for the random number
generator;
n=100;                           %number of balls in each urn;
nb1=n;                           %number of black balls INITIALLY
in urn I;
nw2=n;                           %number of white balls INITIALLY
in urn II;
pb1=nb1/n;                       %probability of selecting a black
ball from urn I;
pw2=nw2/n;                       %probability of selecting a white
ball from urn II;
for trials =1:1000;
   system(trials)=pb1;
   ball1=rand;
   ball2=rand;
   if(ball1>pb1&ball2>pw2)         %white ball selected from urn I
      nb1=nb1+1;                   %and black ball selected from
      nw2=nw2+1;                   %urn II;
   elseif(ball1<=pb1&ball2<=pw2)  %black ball selected from urn I
      nb1=nb1-1;                   %and white ball selected from
      nw2=nw2-1;                   %urn II;
   end
   pb1=nb1/n;
   pw2=nw2/n;
end
plot(system)
axis([1 trials 0 1])
grid
xlabel(’time, in arbitrary units’)
ylabel(’fraction of balls in urn I that are black’)
figure(1)
```

Epilogue

> [Science fiction] cannot be good without respect for good science . . . This does not include time machines, space warps and the fifth dimension; they will continue to exist in the hazy borderland between [science fiction] and fantasy.[8]

> In many science-fiction stories, the trip into the past is by way of some futuristic machine that can take you through time at will . . . That, however, is totally impossible on theoretical grounds. It can't and won't be done.[9]

The opening quotations, particularly the second one from Asimov who was one of the great modern writers of science fiction, is a gloomy one indeed for fans of time travel, but it is not difficult to find inconsistency in Asimov's own tales dealing with the concept. Asimov is famous, in particular, for his stories of robots, and the very last such tale that he wrote combines robotics with time travel, with a robot sent two centuries into the future.[10] At the start of the story, the narrator tells us that time travel to the past is impossible because the past is unchangeable and (of course) a time traveler would necessarily disturb history. (That is (of course) simply a failure to distinguish between the difference of *changing* the past and *affecting* the past, as well as a failure to see how the principle of self-consistency negates the issue of paradoxes.) Then, when the robot returns from the future (and so backward time travel is *not* impossible!), he reports that his arrival had been expected, that history had recorded that he would appear. At the end of the story we learn how the future knew this—it had read "Robot Visions"! So now Asimov *uses*

---

[8] Harry Harrison, in his essay "With a Piece of Twisted Wire . . .," *SF Horizons* (no. 2), 1965. Harrison (1925–2012) was a well-known (if little appreciated outside the SF community) writer, whose 1966 novel *Make Room! Make Room!* was the inspiration for the excellent (if somewhat depressing) 1973 film *Soylent Green* (a movie about future over-population of the Earth that will make you think twice about ever eating a cookie again).

[9] From an essay Isaac Asimov wrote on the time travel movie *Peggy Sue Got Married* for the *New York Times*, October 5, 1986.

[10] I. Asimov, "Robot Visions," *Isaac Asimov's Science Fiction Magazine*, April 1991.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

the principle of self-consistency, with the narrator realizing that he *must* preserve his story so the future can read it.

Not a very consistent story! Asimov, was, of course, writing a story for entertainment's sake, so perhaps it's unreasonable to hold him *scientifically* accountable (although logic wouldn't seem to be too much to ask for).

In any case, was Asimov right? Lots of his fellow science fiction writers certainly thought so. One, for example, bluntly asserted that

> Time travel is inconceivable.[11]

Other critics agreed:

> In science fiction we find the lunatic fringe more often than not trying to perfect time-travel mechanisms.[12]

and

> Scientifically, time travel can't stand inspection.[13]

and

> Time travel is . . . scientific nonsense.[14]

and

> It would be untrue . . . to present the idea of a time machine as anything but what it is, an intriguing literary device, part of the bag of tricks of the science fiction writer . . . There is no such thing as a 'science' of time travel.[15]

You'll notice that these pronouncements are from decades ago: Conklin (1904–1968), Gold (1914–1996), and Oliver all wrote just 3 years after Gödel, and so perhaps it was simply too soon for his work to be widely known outside of the physics community. But physicists have learned a lot since 1952! Have they learned enough to make Asimov and his fellow SF skeptics (if they were still alive) change their minds, or at least reconsider? I suspect not.

I say that because, even 25 years after Conklin, Gold, and Oliver wrote, while we do find an awareness of Gödel starting to appear in the science fiction world, a feeling of skepticism was still in the air. In a fascinating analysis[16] of the first half-century of the science fiction magazines, Paul Carter admitted that there *is* a rationality to time travel because of Gödel but, nonetheless, the conventional view remained that backward time travel is simply impossible. Then, citing the work of Tipler, Carter wrote "Only as recently as 1974 (see note 130 in Chap. 1), in the sober pages of the *Physical Review*, has a physicist been more bold . . . For 70 years in the meantime, however, without waiting for Professor Tipler to solve his equations . . . writers had happily helped themselves to Mr. Wells' invention and sent their characters through time in every direction, forward, backward, and sideways."[17]

---

[11] Kingsley Amis, *New Maps of Hell*, Harcourt 1960.

[12] Groff Conklin, *Science Fiction Adventures in Dimension*, Vanguard 1953.

[13] H. L. Gold, *The Galaxy Reader of Science Fiction*, Crown 1952.

[14] Alexei Panshin, *The Mirror of Infinity*, Canfield 1970.

[15] Chad Oliver, "The Science of Man," a non-fiction essay included in Oliver's 1952 time machine novel *Mists of Time*. Chad Oliver (1928–1993) was a scientist by profession (anthropology), and his opinion carried weight among SF writers *and* (non-physicist) scientists.

[16] P. A. Carter, *The Creation of Tomorrow*, Columbia University Press 1977.

In the 1980s writers were apparently just as unaware of Gödel's time travel analyses (and of the much later ones of Tipler) as had been the 1950s commentators. In his marvelous 1985 book *The Past Is a Foreign Country*, for example, David Lowenthal repeatedly refers to time travel as "fantasy" and to science fiction stories about time travel as "unbridled by common sense." And for another example from the start of the 1980s, consider the case of James Gunn (born 1923), professor of English at the University of Kansas, past president of both the Science Fiction Writers of America and the Science Fiction Research Association, author of *The Immortals* (inspiration for the 1970–1971 TV series of the same name), and eminent scholar (see his 1975 book *Alternate Worlds*). His literary credentials are impeccable and his critical influence profound. And yet, 30 years after Gödel and 5 years after Tipler, Professor Gunn wrote in *The Road to Science Fiction*, "Time travel has been an anomaly in science fiction. Clearly fantastic—there is no evidence that anyone has ever traveled in time and *no theoretical basis for believing that anyone ever will* [my emphasis]." If you've read this book carefully, however, of the analyses by Gott, Krasnikov, Thorne, Alcubierre, Novikov, Natário, and others, you know that what Gunn claims in those last words is actually not necessarily so.

The British-born American theoretical physicist Freeman Dyson of the Institute for Advanced Study has commented[18] on that sort of narrow mindset, with words quoted from the 1979 physics Nobel prize winner Steven Weinberg, words reminding us that rigidity concerning time travel is not limited to science fiction writers: "This is often the way it is in physics—our mistake is not that we take our theories too seriously, but that we do not take them seriously enough. It is always hard to realize that these numbers and equations we play with at our desks have something to do with the real world. Even worse, there often seems to be a general agreement that certain phenomena are just not fit subjects for respectable theoretical and experimental effort." The words *time travel* and *time machine* are never mentioned, but could they have been far from either Weinberg's or Dyson's thoughts?

All through this book we have seen how people have argued against time travel to the past (Tipler's cylinder is unphysically long, Gödel's universe requires an unphysical rotation, wormholes and warps require unphysical energy conditions, what about all those paradoxes . . . and on and on). These arguments remind me of the debate in the 1930s between the illustrious British astrophysicist Sir Arthur Eddington and the young Indian astrophysicist Subrahmanyar Chandrasekhar (1910–1995), winner of the 1983 Nobel prize in physics. In his analyses of the life history of stars, Chandrasekhar had arrived at an astonishing conclusion, one that Eddington simply could not accept. As Eddington sarcastically explained in an address at Harvard University in the summer of 1936, "Above a certain critical mass (two or three times that of the sun), the star could never cool down, but must go on radiating and contracting until heaven knows what becomes of it. That did not worry Chandrasekhar, he seemed to like the stars to behave that way, and believes that is what really happens."[19] Eddington then went on to declare such 'unbelievable' behavior to be nothing less than "stellar buffoonery."

---

[17] Given that *The Time Machine* was published in 1895, it is not clear how Carter arrived at the value of 70 until Tipler's work (he should have written 79), and of course it was only 54 years between Wells' time travel fiction and Gödel's time travel mathematical physics.

[18] F. J. Dyson, "Time Without End: Physics and Biology in an Open Universe," *Reviews of Modern Physics*, July 1979, pp. 447–460.

As far as Eddington was concerned, Chandrasekhar had simply made an error in combining relativity theory with non-relativistic quantum theory. Indeed, so appalled was Eddington at the thought of a star contracting "until heaven knows what becomes of it" (that is, until it gravitationally collapses into a black hole) that he had earlier, in 1935, stated "There should be a law of nature to prevent a star from behaving in this absurd way!" Today, of course, no astrophysicist feels the need for a 'star protection conjecture'—which perhaps reminds you of another, more recent 'protection conjecture.'

What can one conclude from all the similar controversy concerning time travel, time machines, and spacetime warps? Not much, I think, except that these are open issues and will remain the subjects of on-going study for a long time yet to come. The one thing I am fairly certain of is that if time travel is ever achieved, it will be by means that we cannot today even begin to guess. It will almost certainly require *at least* a mutant child genius with an IQ of 270 to fix the slightly broken time machine found abandoned in a cellar![20] But that view isn't uniformly shared across all of science fiction. I very much doubt, for example, that things will be quite so elementary as depicted in the story[21] where the time machine was so simple that "If it were taken apart or put together before you, your wife, or the man across the street, you would wonder why you didn't think of it yourselves." Not only that, but its power source was just two dry cell batteries!

The time machine in an earlier story is almost as simple, requiring (besides a piece of strange crystal) only a "little stack of dry cells, a Ford [automotive ignition] coil, a small brass switch, a radio 'B' battery, an electron tube, and a rheostat."[22] Even Wells' *Time Machine* couldn't resist making it all look easy: as one critic put it, "The time machine, like all products of supreme inventive genius, was a remarkably simple affair. A few rods, wires, some odd glass knobs—nothing more!"[23] That sort of simplistic fictional description of a time machine reminds me of the reaction of the great Polish science fiction writer Stanislaw Lem to the general treatment of time travel in the genre: "There have been mountains of nonsense written about traveling in time, just as previously there were about astronautics—you know, how some scientist, with the backing of a wealthy businessman, goes off in a corner and slaps together a rocket, which the two of them—and in the company of their lady friends, yet—then take to the far end of the Galaxy. Chronomotion, no less than Astronautics, is a colossal enterprise, requiring tremendous investments, expenditures, planning . . ."[24]

---

[19] See S. Chandrasekhar, *Eddington: The Most Distinguished Astrophysicist of His Time*, Cambridge University Press 1983, p. 48.

[20] F. B. Long, "A Guest in the House," *Astounding Science Fiction*, March 1946.

[21] R. Abernathy, "Heritage," *Astounding Science Fiction*, June 1947.

[22] J. Williamson, "In the Scarlet Star," *Amazing Stories*, March 1933.

An example of what Lem was talking about is the 1956 novella *Arcturus Landing* by Gordon R. Dickson (1923–2001). There we read of aliens who have confined humans to the solar system—until (if) Earth scientists discover the secret of FTL travel. So, a genius physicist does just that (with no mention of spacetime engineering, but rather we encounter a lot of mumbo-jumbo gibberish as the 'explanation'), and uses it to instantly transport himself and some friends to a planet orbiting Arcturus.[25] And when they get there the friendly aliens speak perfect English.

Lem would have snorted in derision, too, at this statement made to a prospective graduate student by the head of a college physics department, that the college "has been awarded a million dollars to build [a time machine]. It means . . . a raise for me and maybe a doctorate for you, so we'll build one and have some fun doing it."[26] Is it any wonder that Lem so readily dismissed stories that reduce space (and time) travel to weekend adventures in a home laboratory? As Lem wrote in another essay, time travel and its close relation, FTL space travel, have reduced much of science fiction to "a bastard of myths gone to the dogs."[27] Because of precisely that, Harry Harrison wrote (note 1) of the early science fiction magazines that published so much nonsense, "I used to moan over the fact that pulp magazines were printed on pulp paper and steadily decompose back towards the primordial from which they sprang. I am beginning to feel that this is a bit of a good thing."

I don't know whether time travel to the past can actually be accomplished, but I do know that speculations once thought to be as outlandish as finding the Philosopher's Stone for turning base elements into gold, *have* eventually been realized (and, come to think of it, with modern nuclear physics we *have* learned how to turn lead into gold, if only a few atoms at a time). Television, nuclear power, home computers that run at multi-gigahertz clock rates in the bedrooms of high school students, even faster computers that animate our movies and simulate the formation of black holes and galaxies, voyages to the Moon and back—all these amazing developments would be pure magic to nineteenth century science. The ghosts of not just a few Victorian scientists who had poo-pooed the possibility of such things, have watched their reputations eat a lot of posthumous crow during the last 150 years.

---

[23] W. B. Pitkin, "Time and Pure Activity," *Journal of Philosophy, Psychology and Scientific Methods*, August 27, 1914, pp. 521–526.

[24] S. Lem, "The Twentieth Voyage of Ijon Tichy," in *The Star Diaries*, Seabury Press 1976.

[25] A journey incorrectly given in the story as 120 light years, when in fact it is less than 40 light years.

[26] W. West, *River of Time*, Avalon Books 1963.

[27] S. Lem, "Cosmology and Science Fiction," *Science-Fiction Studies*, July 1977, pp. 107–110.

My personal position on the question of time travel leans towards the rejoinder made to the skeptic in one science fiction story who, even after having done some time traveling, *still* argues against it by invoking paradoxes. He is sharply rebuked with "Oh, for heaven's sake, shut up, will you? You remind me of the mathematician who proved that airplanes couldn't fly."[28] I subscribe to the optimistic philosophy of the British writer Eden Phillpotts (1862–1960), who wrote in his 1934 novel *A Shadow Passes* "The Universe is full of magical things, patiently waiting for our wits to grow sharper." Perhaps he had a famous saying by the British-born Indian scientist J. B. S. Haldane (1892–1964) in mind, words from his 1928 *Possible Worlds*: "Now my suspicion is that the universe is not only queerer than we suppose, but queerer than we can suppose."

Still, even if time travel is possible, the engineering phase will surely be tough going. I am certain that before we see a working time machine, there will be many, *many* episodes like the one described in a very funny, novel-length spoof of academic research.[29] All physicists and engineers who have tried to get some stubborn piece of apparatus to work, apparatus that *should* work and simply won't, will appreciate Professor Demetrious Demopoulos' frustration and will, I am sure, forgive him his intemperate language:

> . . . the distinguished physicist took a step back and, arms akimbo, surveyed the complex and sophisticated machine that was the culmination of years of dedicated scientific research and pains-taking technological development.
>
> "What a pile of ****," he said.
> "Oh, no, Dr. Demopoulos, don't say that!"
> "Well, it is." A sneer formed on the professor's thin lips. "Time machine, my ****. This thing couldn't give you the time much less travel in it."
> "But we haven't incorporated all our latest test data yet," the pretty research assistant reminded him. "These last few adjustments might do it, Professor."
> "Hell, we've been tinkering with it for 2 years," Demopoulos complained.
>
> "We've tried everything and it's all come to dog ****."

---

[28] R. Heinlein, "By His Bootstraps," *Astounding Science Fiction*, October 1941. As discussed at the end of Chap. 1 (in "For Further Discussion") the mathematician was the American astronomer Simon Newcomb.

[29] J. DeChancie and D. Bischoff, *Dr. Dimension*, ROC 1993.

That scene probably won't actually happen for a long time to come but, even before the practical nuts-and-bolts bugs in the Professor's time machine are worked out, I think some adjustments are called for in our thinking about time travel. I believe that present-day philosophers and science fiction writers are going to have to become knowledgeable about the work by physicists on time travel. It simply won't do any longer for Philosophy Professor X to invoke the grandfather paradox during a discussion of causality and free will and then airily declare them to be 'obviously' incompatible with time travel to the past. And it simply won't do any longer for Famous SF Writer Y to send his hero into the past to kill Hitler as a baby and thereby change recorded history. One might as well keep watching a video recording of the 9/11 destruction of the World Trade Center, in the vain hope that maybe, on the next viewing, the planes will miss.

The principle of self-consistency around closed timelike curves is going to have to become as much a part of the science fiction writer's craft (or else she will be a writer of fantasy) as it will have to become part of the fundamental philosophical axioms.[30] The 'time police,' like the "operatives of the Bureau of Time Exploration and Manipulation" that appeared in the science fiction of Andre Norton (1912–2005), will have to be put out to pasture with the unicorns and telepathic dragons of fantasy fiction. Just as the recent physics literature on time machines has displayed a growing awareness of what science fiction writers and philosophers have had to say on the subject of time travel, so too are writers and philosophers going to have to learn some more physics. Most people can enjoy a good fantasy tale now and then, but the use of 'magic mirrors' to see through time is *not* physics. Such devices were popular and acceptable in medieval times—see "The Squire's Tale" in Chaucer's *The Canterbury Tales*, and later (see Act IV of *Macbeth*)—but good science fiction needs much more than that today.

Time travel to the past is a beautiful, romantic idea, and some words written by two physicists in a technical paper—words embedded in the midst of swirls of tensor equations—show that even hard-nosed physicists can share this dream: "In truth, it is difficult to resist the appealing idea of traveling into one's own past . . ."[31] The appeal of that dream is explained in Ray Bradbury's Foreword to a beautiful little 1989 book by Charles Champlin (*Back There Where the Past Was*). In it Bradbury clearly illuminated *why* we want to go back into the past. It is for the same reason that we go, time and again, to see *Hamlet*, *Othello*, and *Richard III*: "We don't give a hoot in hell who poisoned the King of Denmark's semicircular canal.

---

[30] Bud Foote (1930–2005), late professor of English at Georgia Tech, wrote (in his book *The Connecticut Yankee in the Twentieth Century: Travel to the Past in Science Fiction*, Greenwood Press 1991) that consistency is simply a well-used plot device: "The attempt of the time traveler to prevent something or take advantage of it [and so causing] the event in question, is so popular and so ubiquitous that it seems to be about worn out." Worn out or not, I believe that plot device to be correct science.

[31] A. J. Accioly and G. E. A. Matsas, "Are There Causal Vacuum Solutions with the Symmetries of the Gödel Universe in Higher-Derivative Gravity?" *Physical Review D*, August 15, 1988, pp. 1083–1086.

We already know where Désdemona lies smothered in bedclothes and that Richard goes headless at his finale. We attend them to toss pebbles in ponds, not to see the stones strike, but the ripples spread."

That's why a visit to the past is so mysteriously and marvelously fascinating. It would let us watch ripples spread through time. Our own visit to the past, in fact, might even be the pebble in the pond of history that starts an interesting ripple or two that will one day sweep over—us! (Take a look at Appendices A and B) Who would want to miss that? Indeed, if modern philosophers are right, if the analyses discussed earlier in this book are correct, you *can't* (didn't/won't) miss it.

I think time travel appeals, irresistibly, to the romantic in the soul of anyone who is human.[32] A time traveler does not exist either *here* or *then*, but rather *everywhen*. For a time traveler passing back and forth through the ages, history would be the ultimate puzzle, a chronicle described in one novel as beginning "not in one place, but everywhere at once ... It might be begun at any point along the infinite, infinitely broken coastline of time."[33] Romanticism doesn't preclude there also being a dark side to visiting the past, of course, as one time traveler from 1989 learns when he takes up residence in 1962. Falling asleep on a hot summer night in that long-ago year, he thinks "JFK slept. Oswald slept. Martin Luther King slept. [I sleep and dream] of Chernobyl ... *I am a cold wind from the land of your children*."[34]

But, I must admit, I personally am more attracted by happier descriptions of time travel. In his marvelous 1996 book *1939: The Lost World of the Fair*—which is proof that there are not enough Pulitzers to go to all the books that deserve one—Yale professor David Gelernter caught just the right spirit in his Prologue: "The best of all reasons to return to the fair is that travel is broadening, and time travel most of all ... The 1939 New York World's Fair is one amazing show. It still stands, undisturbed on Flushing Meadow, just over the edge of time; it would be an unforgivable shame to miss it." Trust me—if you read Gelernter's book, you'll come as close as you can in today's world to taking a ride in a 'time machine'!

The eminent philosopher Sir Karl Popper opens his biography with a wonderful story about his apprenticeship as a young man in 1920s Vienna to a master cabinetmaker.[35] After winning the old man's confidence, the student learned his mentor's great, secret desire: For years the master had been looking for the solution to perpetual motion. He knew what physicists thought of such machines, but nonetheless he had never given up his dream: "They say you can't make it; but once it's been made they'll talk differently." Popper's master sounds just a bit like the American writer Gertrude Stein (1874–1946) in her 1938 essay "Picasso," where she writes "It is strange about everything, it is strange about pictures, a picture may seem extraordinarily strange to you and after some time not only it does not seem strange but it is impossible to find what there was in it that was strange." Might we one day say the same thing about time travel?

---

[32] How else to explain the pleasure, for modern children *and* adults too, in watching rebroadcasts of the 1960s animated TV cartoon program 'starring' Mr. Peabody, a nice but slightly stuffy, professorial white beagle. (Don't all dogs wear glasses and a bow tie?) Mr. Peabody, with his brainy adopted son Sherman, routinely travels into the past in the "Way-Bac" machine to see what *really* happened in history.

[33] John Crowley, *Great Work of Time*, Bantam 1991.

[34] R. C. Wilson, *A Bridge of Years*, Doubleday 1991.

[35] See volume 1 of *The Philosophy of Karl Popper* (P. A. Schilpp, editor), The Library of Living Philosophers, Open Court 1974, p. 3.

An alternative point of view can be found in a discussion of time travel via cosmic strings that makes this assessment: "While there is still hope that one day a sufficiently clever design may make building a time machine possible, it is beginning to seem more and more improbable. Like the perpetual motion machines of the nineteenth century, the designs have an elegant simplicity (as well as enormous commercial potential), but it seems that Nature also may abhor them just as much."[36] Of course, at one time it was thought that Nature abhorred a vacuum, but then we learned that she must actually love a vacuum because else why did she make so much of it?!

The theoretical basis for time travel is very different from that of perpetual motion (there is *more* reason to accept time travel as a plausible possibility). And so maybe one day, *just maybe*, the first time traveler will receive a toast such as the one in a story telling us about the arrival of the inventor of the first time machine and his no longer skeptical friend in the Civil War past:

> "To you, Mac," I said.
>
> McHugh loosened his tie. "To the Creator," he said, "who has given us a Universe with such marvelous possibilities."[37]

---

[36] B. Allen and J. Simon, "Time Travel on a String," *Nature*, May 7, 1992, pp. 19–21.

[37] J. McDevitt, "Time's Arrow," *Isaac Asimov's Science Fiction Magazine*, November 1991.

Glossary[38]

Action the integral over a **world line** of a quantity called the *Lagrangian*. When a massive particle is moving at non-relativistic speed through a gravitational field, for example, the instantaneous value of the Lagrangian is the difference between the kinetic and potential energies of the particle. For other types of fields (such as the electromagnetic) and/or relativistic motion in any type of field, the Lagrangian is different. In any case, however, the actual world line of the particle is the one for which the integrated Lagrangian, that is, the action, is minimized. See **least action**.

Action at a distance the direct interaction of two separated objects, without concern for the details of what (if anything) occurs in the region between the objects (see also **field**). Newton's theory of gravity is action at a distance, whereas Einstein's theory of gravity is a *field* theory.

Advanced solution the prediction, by Maxwell's electromagnetic field equations, of radio waves that travel into the past (see also **Dirac radio**).

Anti-matter quantum mechanical prediction (experimentally verified) that all fundamental particles of matter come in two forms (the 'normal' version and the 'anti-matter' version). The positron, for example, is the anti-matter version of the electron, differing only in the sign of its electric charge. The photon, on the other hand, is its own anti-particle. A **subluminal** anti-particle traveling forward in time can be thought of as its 'normal' version traveling backward in time.

Arbitrarily advanced civilization for time travel discussions, a civilization with a technology sophisticated enough to construct a traversable **wormhole** in spacetime. More generally, Types I, II, and III of such civilizations are, respectively, those that can control 10^{13} W, 10^{27} W (the total power output of their home star), and 10^{38} W (the total power output of their home galaxy).

---

[38] "I hate definitions." (Usually attributed to writer and British Prime Minister Benjamin Disraeli (1804–1881) but, more precisely, they are the words of one of the characters in his 1826 novel *Vivian Grey*.)

—but they *can* be useful

© Springer International Publishing AG 2017
P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

Arrow of time the statement the time appears to have a direction, that there is a difference between the past and the future. There are several different arrows: the psychological (we remember the past, we anticipate the future), the thermodynamic (organized systems evolve toward disorganization, that is, **entropy** increases as time increases), the electromagnetic (radio waves propagate *away* from their generators), and the cosmological (the expansion of the universe is directed toward the future).

Asymptotically flat if the geometry of a curved spacetime is such that, as one moves ever further away from all matter and energy, the spacetime **metric** becomes that of flat **Minkowski spacetime**, then the curved spacetime is said to be *asymptotically flat*. As a counter-example, the spacetime of a **Tipler cylinder** time machine is *not asymptotically flat*.

Autoinfanticide paradox see **grandfather paradox**.

Averaged null energy condition the claim that the *averaged* value of the observed mass-energy density along the entirety of any **null geodesic** is non-negative.

Averaged weak energy condition the claim that the *averaged* value of the observed mass-energy density along the entirety of any **timelike world line** is non-negative.

Back reaction the tendency of spacetime to resist the formation of **closed timelike lines** (see also **stress-energy divergence**).

Bell's theorem an inequality that either holds or does not hold, depending on whether quantum mechanics is non-local or local, respectively.

Big Bang the singular beginning of spacetime.

Big Crunch the singular end of spacetime.

Bilking paradox what would happen if a **causal loop** were disrupted. For example, suppose a time traveler builds a time machine using plans he received years earlier from a mysterious stranger. He now realizes that the stranger was himself, using the time machine to travel back into the past to give his younger self the plans. A bilking paradox would be created if the time traveler builds the time machine, verifies that it works, and then decides *not* to visit his younger self to hand over the plans. See also **bootstrap paradox**.

Black hole a region of spacetime where gravity is so strong that nothing can escape, including light. Black holes are thought to be created when sufficiently massive stars burn out (see **white dwarf** and **neutron star**) and undergo *gravitational collapse*. A black hole of ten solar masses would have a radius of about twenty miles. Black holes might have been created at the Big Bang singularity and, if so, could theoretically come in any mass and size (a black hole with the mass of the Earth would have a diameter of less than half an inch).

Block universe a spacetime in which all world lines are completely determined from beginning to end (a fatalistic universe). There is no free will in such a spacetime.

Boost matrix matrix formulation of the **Lorentz transformation**.

Bootstrap paradox the puzzle of the origin of *information* on a closed loop in time. The classic example is that of a time traveler from the future giving his younger-self the plans for the time machine the time traveler has just used to visit the past so that he can then build the time machine to visit the past. The time machine plans appear not to have been *created* by anyone! The plans just *are*. See also **bilking paradox**.

Cauchy horizon a **spacelike** hypersurface in spacetime that intersects, exactly once, every **timelike** world line that has no end point. Knowledge of the conditions on such a surface uniquely determines the spacetime at all other points.

Causal loop a time loop containing an event caused by a *later* event that, itself, is caused by the earlier event (see the example in **bilking paradox**).

Causality the metaphysical claim that every event is caused by a prior event. Time travel to the past inherently violates causality.

Chronal regions those parts of spacetime that have no closed timelike curves.

Chronology horizon a (hyper)surface in spacetime that separates **chronal** and non-chronal regions. It is a special case of a **Cauchy horizon**.

Chronology protection the claim, as yet unproved, that time machines and time travel to the past are impossible because of the **back reaction** of spacetime will lead to **stress-energy divergence**. Popularized among physicists as the *Hawking chronology protection conjecture* (1992), Hawking has since admitted that stress-energy divergence is *not* sufficient to enforce his conjecture.

Chronon science fiction name for **Planck time**.

Closed timelike line (or curve) a **timelike** world line of finite length that has no ends, i.e., that forms a *closed loop* in spacetime. A region of spacetime containing closed timelike lines is said to be a **time machine**.

Conservation law physical quantities in interacting systems that remain unchanged are said to be conserved. Total energy, total momentum (linear and angular), and electric charge are conserved quantities.

Cosmic string hypothetical, threadlike spacetime structures with enormous mass-energy and density that may have formed during the **Big Bang**. Cosmic strings may have been initially formed either as infinitely long, or as closed loops, and it is the former that are thought to be physically meaningful in the present-day universe. Cosmic strings do not violate the **weak energy condition** (as do **wormholes**), and they can theoretically create **closed timelike lines**.

Cosmological constant an extra term specifically added by Einstein to the general theory of relativity to keep that theory from predicting the expansion of the universe (which was later observationally found to actually be the case). Einstein subsequently said that his failure to believe the general theory's original prediction of the expansion of the universe was the greatest mistake of his life. The constant (which today is believed to be almost zero, if not exactly zero) appears in Gödel's rotating time travel spacetime as a determining factor in the minimum radius of a **closed timelike line**.

Determinism the metaphysical belief that effects are uniquely determined by causes (this is *not* **fatalism**).

Dirac radio science fiction gadget for sending information at infinite speed, which thus travels backward in time (see also **ultraluminal**).

Dominant energy condition the **weak energy condition** *plus* the claim that the observed energy flux is never **superluminal**.

Electron fundamental particle of mass that possesses one quantum of negative electric charge. *Bound* electrons orbit the nuclei of atoms and plays a central role in determining the chemical properties of the elements and of their compounds. *Free* electrons carry electric current, either in conductors (wires) or through space.

Elsewhen the collection of spactime events that cannot be reached from the **here-now** with a **timelike world line**.

Entropy a measure of the randomness of a system that plays a central role in the thermodynamic **arrow of time**.

Ether a substance once thought to fill all space to allow radiation 'something to propagate through' (as opposed to simply a vacuum). The special theory of relativity showed that the ether is an unnecessary concept because it has no observable effects (physicists argue that if something is impossible to detect, then it is meaningless to talk about it being part of *science*).

Event a point in spacetime.

Event horizon the spacetime surface of a black hole or of a non-traversable wormhole, at which light can *just* escape to the outside universe. It is called a *horizon* because, by definition, an external observer can't see beyond it and into the interior of the hole. To see the inside of a hole you must enter the hole by crossing the horizon (but then you can't get out).

Exotic matter matter that violates one or both of the **weak/strong energy conditions**. Exotic matter appears in the theories of wormholes and warp drives.

Fatalism the metaphysical belief that all events have been *pre*determined from the beginning of time.

Field the concept that if a physical law is local, then it is describable by differential equations that relate what is 'happening' at every point in spacetime to what is 'happening' at its closely located neighboring points. Electromagnetism and general relativity are field theories, for example, described by sets of partial differential equations called *Maxwell's equations* and *Einstein's gravitational field equations*, respectively.

Fourth dimension either time or a fourth spatial dimension.

Frame of reference a spacetime coordinate system.

Free will the condition that prevails when we can *choose* to do what we do. There is no free will in a **block universe**.

Future the collection of spacetime events that can be reached from the **here-now** via a **timelike world line** directed toward a later time (for each individual, the future is what hasn't yet been experienced).

Gamma ray very high-energy, very high-frequency electromagnetic radiation. Gamma rays have frequencies on the order of *ten trillion* (10^{13}) times greater than those of AM radio broadcast radio waves.

General theory of relativity Einstein's theory of curved spacetime, which explains gravity in terms of nothing but geometry. Its fundamental premise is that *all* the laws of physics should appear the same to all observers in *any* **frame of reference**. It is believed the theory will fail when the local mass-energy density reaches a level of about 10^{94} g/cm^3, a density so enormous (the density of water is just 1 g/cm^3) that there is no known mechanism for achieving it anywhere in the universe except in another **Big Bang**. See also **Planck density**.

Geodesic the shortest path connecting two points in space (if the space is spacetime, the world line of a particle in free-fall).

Global in the large.

Gödel universe a spacetime that, unlike the one we live in, is rotating so fast that it automatically generates *closed timelike lines* and thus constitutes a weak **time machine**. In such a universe, time travel to the past would be a natural phenomenon.

Grandfather paradox *the* classic time travel paradox, of a time-traveler killing, while in the past and *before* the time traveler has been conceived, an ancestor directly linked to the future birth of the time traveler. A more direct form of this sort of paradox is simply the time traveler killing his own younger self (called the **autoinfanticide paradox**).

Gravitational field equations a set of coupled, partial differential, non-linear tensor equations, considered to be the most complicated equations in all of mathematical physics. They show how the local curvature of spacetime depends on the local mass-energy of spacetime. The equations are independent of the **topology** of spacetime.

Gravitational lensing the ability of gravitational fields to bend and focus light.

Graviton the quantum particle of gravity.

Hawking radiation the emission of particles (energy) by a **black hole** into the region *outside* its **event horizon**, which results in the eventual evaporation of the hole. This is a quantum mechanical effect.

Here-now the point or **event** (for each observer) in spacetime that separates the **past**, the **future**, and **elsewhen**.

Hyperspace any space of four or more dimensions (for example, four-dimensional spacetime is a hyperspace).

Inertial frame any frame of reference in which Newton's laws of mechanics are true (there are no *acceleration forces* in inertial frames, and so *rotating* or 'merry-go-round' frames are not inertial).

Invariance a quantity that remains the same in any frame of reference is an invariant. Two examples are the distance between any two points on a piece of paper (because it is independent of any particular coordinate system), and the speed of light.

Kerr-Newman black hole a *rotating* black hole, which may (or may not) be electrically charged.

Krasnikov tube a particular spacetime **metric** (or *warp*) allowing **superluminal** travel, with the great difficulty of requiring *enormous* negative energy. Two Krasnikov tubes can be made into a time machine. Named after its Russian inventor.

Least action general principle in physics that asserts the world line of a particle is the one that *minimizes* the **action**.

Light cone the **lightlike** surface in spacetime that, at each point in spacetime, separates the **past** from the **future** from **else-when** from the **here-now**.

Lightlike the world line of a photon (or of any other form of mass-energy traveling *at* the speed of light).

Li mirror a perfectly reflecting, spherical surface that can be used to stabilize a wormhole against energy loops circulating through a **wormhole** time machine (thus creating unbounded energy levels that destroy the time machine). Named after its Chinese inventor.

Local in the small.

Lorentz factor the ubiquitous square-root expression that appears in so many relativistic calculations, such as time dilation, length contraction, and the variation of mass with speed. For example, the mass m of a moving body is not independent of it speed v but rather varies as $m = \frac{m_0}{\sqrt{1-\left(\frac{v}{c}\right)^2}}$, where m_0 is the rest mass (that is, the mass when $v = 0$) and c denotes the speed of light (186,210 miles per second). The denominator is the Lorentz factor.

Lorentz-FitzGerald contraction the conclusion from special relativity that the appearance (to a stationary observer) of a moving object will be shortened in length along the direction of motion. Many years after Einstein's work, it was shown that the object will also appear to be *rotated*.

Lorentz transformation equations from the special theory of relativity that describe how the space and time measurements of two relatively moving observers are related.

Many-worlds interpretation quantum mechanical view of splitting universes.

Mass-energy the famous $E = mc^2$, the equation behind atomic fission and nuclear fusion weapons.

Metric the measure of the separation between any two events in a spacetime.

Minkowski spacetime the flat spacetime of the special theory of relativity. In this spacetime there is no gravity, no spacetime curvature (hence it is *flat*) and no backward time travel.

Neutron star the end state of a star with one to three solar masses that has collapsed to a density of up to 10^{17} g/cm^2.

Non-Euclidean geometry the geometry of spacetime, whether curved or flat. Spacetime is non-intuitive precisely because it is always hard to resist thinking in terms of high school Euclidean geometry, which is simply the *wrong* geometry.

Null geodesic the world line of a photon in spacetime.

Observer physicist's term for 'somebody' equipped with recording instruments (such as a clock, a pencil and notepad, and the like).

Parallel transport a procedure for moving a vector around any closed curve in a space to determine whether that space is flat or curved.

Parallel worlds simultaneous existence of multiple (perhaps) infinite versions of reality.

Past the collection of spacetime events that can reach the **here-now** via timelike world lines directed from an earlier time (for each individual, the past is what has already been experienced).

Photon the quantum particle of electromagnetism. A photon of frequency f has energy hf, where h is **Planck's constant**.

Planck density the density of mass-energy that distinguishes classical from quantum spacetimes; about 10^{94} g/cm^3, equal to the **Planck mass** divided by the cube of the **Planck length**.

Planck length the non-zero length in quantum theory (about 1.6×10^{-33} cm) below which quantum gravity effects will become important.

Planck mass the fundamental mass in quantum theory (about 22×10^{-6} g), but *not* the smallest non-zero mass in quantum theory.

Planck's constant fundamental constant in quantum theory, h, associated with the discrete nature of quantum effects. (If h had the value of zero, rather than its actual value of about 6.6×10^{-34} joule-seconds, then the microworld would appear to be continuous.)

Planck time the time interval in quantum theory (about 5.3×10^{-44} s) below which quantum gravity effects become important. The time required to travel the **Planck length** at the speed of light.

Positron the electron's anti-particle (see **anti-matter**).

Proper time the timekeeping of an observer's clock.

Pulps the old science fiction magazines, through the 1940s and into the early 1950s or so, published on inexpensive, wood-pulp paper.

Quantum foam see **topology**.

Quantum gravity the yet-to-be-discovered theory that unifies quantum field theory with the curved spacetime of general relativity.

Quantum mechanics the exact physics of the very small (atoms and things smaller).

Quantum theory any theory in which physical quantities are not continuous but rather assume their values in discrete jumps (the size of the jump is the *quantum*).

Recurrence paradox the claim that if you wait long enough, then every system will return to every previous state infinitely often.

Red dwarf small (less than about half a solar mass) star with a very long life (hundreds of times that of the Sun). They are 'cool' stars, with a surface temperature less than 4000 °C, and are thought to be the most common type of star in the universe.

Red shift the *down* shift in frequency of light received from all distant stars due to the Doppler effect induced by the expansion of the universe. The opposite effect is called a *blue shift*.

Reinterpretation principle asserts that negative mass-energy traveling forward in time is positive mass-energy traveling backward in time, and vice-versa.

Reissner-Nordström black hole a spherically symmetric, non-rotating electrically charged black hole.

Reversibility paradox based on the fact that the equations of physics contain no **arrow of time**; that is, they work equally well with time running forward or backward.

Roman ring a time machine made of two or more traversable wormholes connected in a closed sequence.

Schwarzschild black hole a spherically symmetric, non-rotating, uncharged black hole.

Self-consistency the assertion that the events on a closed **timelike** line must never be in contradiction; generally attributed to the Russian physicist Igor Novikov, who with his colleagues showed that it is not an independent assumption but rather an implication of the principle of **least action**.

Sexual paradox a special type of causal loop, where the connected events on a time loop are 'coupled' (pun intended!) through reproductive sex. An example is a time traveler to the past who becomes her own ancestor.

Singularity either a region in spacetime where the curvature becomes infinite and the laws of physics fail, or a point in spacetime beyond which world lines cannot be extended. Singularities of the first kind are called *curvature* or *crushing* singularities, and those of the second kind are called incomplete singularities. The **Big Bang** was a curvature singularity, as is the center of a black hole. In a **Schwarzchild black hole** the curvature singularity is a point, whereas in a **Kerr-Newman black hole** it is an extended region in the form of a ring.

Spacelike a world line on which propagating mass-energy would exceed the speed of light.

Spacetime the 'stuff' out of which reality is built. Everything there is—the universe—is the total collection of events in spacetime. A *flat* spacetime has no gravity, whereas a curved spacetime is the *origin* of gravity.

Special theory of relativity Einstein's theory of *flat* spacetime, which assumes that gravity is absent (gravity is the result of the geometry of *curved* spacetime). Its fundamental premise is that the laws of physics should appear the same to observers in different **inertial frames**.

Splitting universes the idea that every decision causes reality to split into separate copies, identical in every respect except for each of the different possible results of the decision.

Stargate science fiction name for the mouth of a traversable **wormhole**.

Stress-energy divergence the unbounded growth of the **general theory of relativity**'s measure of the density of mass-energy in spacetime.

Strong energy condition the claim that gravity is always (that is, locally) attractive. A traversable **wormhole** violates this condition.

Subluminal slower than light.

Superluminal faster than light.

Tachyon a particle (hypothetical, so far) that always travels faster than light, so its **world line** is always spacelike.

Temporally orientable spacetime any spacetime in which the direction of time at every point agrees with the direction of time at its local neighboring points.

Tensor mathematical generalization of the scalar and vector concepts. Einstein's **gravitational field equations** are tensor-differential equations (for example, the metric tensor contains information about the curvature of spacetime), whereas Newton's and Maxwell's equations are vastly less complex vector-differential equations.

Tidal force force experienced by a non-point mass (one with spatial extension) in a non-uniform gravitational field. Such forces tend simultaneously to compress and stretch spatially extended masses. Black holes and wormhole mouths can generate enormous tidal forces on extended masses as small as a human body. Interestingly, the *more* massive a black hole, the *less* severe its tidal forces are at distances outside the **event horizon**. However, no matter what the black hole mass is, the tidal forces are infinite at the central curvature **singularity**.

Time dilation the altering of the rate of timekeeping by a clock, either by motion or by gravity.

Time machine (in the weak sense) a machine able to traverse **closed timelike world lines** inherent in a spacetime (e.g., a rocket in Gödel spacetime) but unable to *create* such world lines; (in the strong sense) a machine able to manipulate mass-energy in a finite or compact region of spacetime in such a way as to *create* closed timelike world lines.

Time police story characters in science fiction charged with the (unnecessary!) job of preventing time travelers from changing the past.

Time warp science fiction name for a **time machine**.

Tipler cylinder an infinitely long cylinder, made of super-dense matter, rotating so fast around its long axis that it warps spacetime enough to create closed timelike lines that encircle the cylinder. It can be used as a strong sense time machine to travel both into the future and into the past (but *not* to a time before the creation of the cylinder).

Topology the structure of a *space* (including *spacetime*) without regard to a metric. That is, topology is concerned only with how a space is connected together and not with how far apart points in the space are. Topologists consider stretching or compressing a space to be irrelevant, just as long as one doesn't *tear* it and so put holes in the space. The simplest topology is that of a *simply connected* space, in which if you construct any closed surface that lies totally in the space around any point in the space, then every other point inside the surface is also in the space. A space with a hole in it fails this test, and so is said to be *multiply connected.* A *quantum foam* spacetime has a multiply connected topology. The classical spacetime of general relativity is simply connected *until* the appearance of wormholes.

Twin paradox the conclusion from special relativity that a clock's rate of time keeping slows with motion.

Ultraluminal motion sufficiently **superluminal** that mass-energy appears to travel backward in time (see also **Dirac radio**).

Uncertainty principle the statement in quantum mechanics that says certain pairs of quantities cannot simultaneously be measured with arbitrarily small error. The position and momentum of a particle are one such pair, and energy and time are another.

Vacuum fluctuation the particle/anti-particle creation and annihilation processes allowed, even empty space, by the **uncertainty principle** of quantum mechanics.

Warp drive science fiction name for the propulsion mechanism of a faster-than-light spaceship, now commonly used by physicists, too.

Weak energy condition the claim that the observed mass-energy density is always (locally) non-negative. Quantum mechanics predicts (and it has been experimentally confirmed) that there are exceptions.

White dwarf a burnt-out star with a mass less than 1.4 solar masses, of planetary size with a density up to 10^7 g/cm^3. The ultimate fate of our Sun.

World line the trajectory of mass-energy in spacetime.

Wormhole a spacetime structure (violating the **weak and strong energy conditions**, if traversable) connecting two points of the same spacetime (or even two *different* spacetimes) with a timelike path that requires less time to travel along than does a photon traveling *outside* the wormhole between the two points. A wormhole is *traversable* if it has no **event horizons**, and such wormholes can apparently be made into a **time machine** (sometimes called a *time tunnel*) using a time shift (see **time dilation**) between the two mouths of the wormhole *unless* quantum effects forbid time machines (still an open question).

Index

A
Absorber theory, 269–273
Advanced waves, 259, 260, 264, 265, 269–273, 283
Alcubierre warp drive, 329, 331–333. *See also* Natário warp drive
"All You Zombies—," 21, 194, 227, 239
"Ambition," 36
"Anachronistic Optics," 260
"Ancestral Voices," 204
Anderson, Poul, 41, 155
"—And He Built a Crooked House," 88
Antimatter, 111, 157, 247, 248
Anti-telephone, 261, 273–282
Aquinas, St. Thomas, 63, 92, 102, 103, 196
Aristotle, 53, 54, 56, 63, 66, 79, 96, 195, 208
Ark, 333
Arrow of time. *See specific arrow* (cosmological; electromagnetic; psychological; thermodynamic)
Asimov, Isaac, 12, 37, 154, 207–209, 236, 357, 358
"As Never Was," 220
Aspect, Alain, 281, 282
Asymptotically flat, 171, 305, 317
Augustine, St., 51, 55, 109, 181
Autoinfanticide paradox, 112, 206–207, 212, 229, 238. *See also* Grandfather paradox; Sexual paradox
Average energy condition, 305–308
"Azimuth 1, 2, 3 . . .," 42

B
Back reaction, 30, 309, 317
Back to the Future, 3, 23, 162, 214, 231, 244
Backward causation, 197, 198
Barrow, Isaac, 57
"Beep," 76, 236, 273, 274
Behold the Man, 211, 243
Bellamy, Edward, 250
Bell, J.S., 232, 233, 279
Benford, Gregory, 24, 228, 229, 273, 318
Bergson, Henri, 54
Berkeley Square, 3, 99, 101
Besso, Michele, 90, 93
Bid Time Return. See Somewhere in Time
Big Bang, 37, 46, 62, 63, 66, 135, 136, 221, 252, 271, 272, 317, 318
Big Crunch, 46, 135, 138
Bilking paradox, 76, 77, 189, 249, 261, 268, 269, 272, 277, 286, 290, 296, 318
Bill & Ted's Excellent Adventure, 3, 224
"Biography Project, The," 242
"Birch Clump Cylinder. The," 216
Black hole, 32–38, 41, 43, 84, 138, 283, 302, 311, 320
Black, Max, 101, 120
"Blindman's World, The," 250
Blish, James, 28, 235, 272, 273
Block universe, 90–109, 116, 121–122, 126, 197, 246, 261
Blue shift, 144
Boethius, 107
Boltzmann, Ludwig, 127, 130, 131, 134
"Bombardment in Reverse," 253
Bonaventure, St., 63
Bootstrap paradox, 26, 39, 43, 214–225, 239, 241, 242
Bradbury, Ray, 52, 200
Bradley, Francis, 94–95, 121, 246

© Springer International Publishing AG 2017
P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

"Branches of Time, The," 234
Bring the Jubilee, 236–237
Brown, Fredric, 188
"By His Bootstraps," 187, 362

C
Campbell, John, 235
Cantor, Georg, 63
Carroll, Lewis, 125
Casimir effect, 306–308, 330
"Castaway," 45
Cauchy condition, 136
Cauchy horizon, 309–312, 314, 316–317
Causal loop, 40, 76, 214–227, 231, 232, 239–240, 242, 249, 262, 270, 276, 310, 318, 320, 323
Cause-and-effect, 67–77, 118, 138
Cerenkov radiation, 291
Chandrasekhar, S., 31, 360
Chaucer, Geoffrey, 107
"Child by Chronos," 227
"Child's Play," 2
"Chronokinesis of Jonathan Hull, The," 127
Chronology protection conjecture
  Hawking, 13–14, 42, 45, 284, 311, 315, 317, 325
  partial rebuttal by Thorne, 42
Chronon, 30, 57. *See also* Planck time
"Circle of Zero, The," 134
Civilization, arbitrarily advanced, 303, 336
Clarke, Arthur C., 10, 11, 15, 22, 110, 111, 133, 239
Clifford, William Kingdon, 169
Clock hypothesis, 174, 179
"Clock That Went Backward, The," 199
Closed time loop (CTL). *See* Causal loop
Connecticut Yankee in King Arthur's Court, A, 198
Contact, 302, 318
"Cosmic Corkscrew," 12
Cosmic string, 320–324
Cosmic string time machine, 163, 319–326
Cosmological arrow of time, 137–138, 271
Counter-Clock World, 131, 251
Cramer, John, 82, 182, 269, 272, 282, 285, 291
Cumulative audience paradox, 10

D
Damian, Peter, 195–196
"Dandelion Girl, The," 73, 210
Daniels, David R., 234–235
"Dark Tower, The," 18
De Camp, L. Sprag, 188, 208, 209, 230
Dechronization of Sam Magruder, The, 260
Descartes, 57
Determinism. *See* Free will
Deutsch, David, 222, 233–234
Deutsch-Politzer spacetime, 326
DeWitt, Bryce, 277
Dick, Philip K., 131, 243, 251, 283
Dirac, Paul, 74–76, 127, 167, 168, 263–264, 267
Dirac radio, 76, 264, 265, 270, 272–273
Distance function. *See* Metric
Dominant energy condition, 317, 329
Doomsday Book, 5
Door Into Summer, The, 11, 180, 221
Dyson, Freeman, 359

E
Earman, John, 18, 70, 72–73, 75, 96, 158
Eddington, A.S., 127, 132, 135, 250, 253–254, 360
"E for Effort," 16
Ehrenfest, Paul, 80, 129
Einstein, Albert, 1, 17, 32–33, 44, 53, 54, 57–58, 60–61, 79, 84, 90, 93, 95, 100, 101, 116, 123, 126, 145, 146, 149, 153, 168, 174, 250, 258, 262, 279–282
Einstein-Rosen bridge. *See* Wormhole
Einstein's Bridge, 84
Electromagnetic arrow of time, 138, 267, 271, 272
"Elsewhen," 7, 253
Energy (conservation of in spacetime), 99–100
Energy condition. *See specific condition* (averaged; dominant; strong; weak)
Entropic clock, 129
Entropy, 127–134, 136–138, 271
Epictetus, 106
EPR paradox, 281, 282
Eternal return, 134
Event horizon, 33, 86, 146, 264, 302, 320
Everett, Hugh, III, 229, 231, 233–235
Exotic matter, 304–306, 308, 318, 320, 325, 329, 334, 335
"Experiment," 25

F
"Face in the Photo, The," 7
"Far Centaurus," 4
Faster-than-light (FTL), 4, 34, 55, 56, 116, 141, 142, 162, 262–264, 273–277, 279, 281,

282, 289–296, 327–329, 331, 333, 334, 361. *See also* Tachyon(s)
Fatalism. *See* Free will
Feinberg, Gerald, 274–277
Fermi, Enrico, 10, 279
Feynman, Richard, 37, 126, 157–159, 181, 259, 264–270, 272, 277–278
Fifth dimension, 48, 83, 87, 105, 110, 167–168, 182, 229, 253. *See also* Hyperspace
"Fifth-Dimension Catapult, The," 167
Final Countdown, 43, 240
Findlay, J.N., 246, 247
"Find the Sculptor," 219
First Cause, 62
"First Time Machine," 205
Flammarion, Camille, 55–56
Flamm, Ludwig, 84
"Fools' Errand," 224
Forever War, The, 38, 43
Forever Young, 180
Forward, Robert, 184, 283, 307, 312
"Four Dimensional Surgery," 83
Fourth dimension, 78–90, 92, 109, 110, 182, 215, 250. *See also* Hyperspace
"Fox and the Forest, The," 6
Free will, 60, 93, 94, 100, 102–104, 106–109, 124, 196, 198, 213, 217, 233, 238, 268, 278–279, 312, 313, 318, 363
FTL. *See* Faster-than-light (FTL)

G
Galilei, Galileo, 147, 152
"Garden of Forking Paths, The," 236
Geodesic, 155, 159, 175, 305, 310, 330. *See also* World line
Gernsback, Hugo, 191–194, 209, 225
Gödel, Kurt, 20, 21, 28, 44, 101, 116–117, 158, 212, 276, 297–299, 301, 358–360
"Gostak and the Doshes, The," 83
Gott, J. Richard, III, 320–325, 359. *See also* Cosmic string time machine
Grandfather paradox, 45, 200, 201, 204–206, 211–214, 220, 233, 312, 313, 326–327, 336, 363. *See also* Autoinfanticide paradox; Sexual paradox
Gravitational collapse, 31, 35, 360
Gravitational constant, 30
Gravitational field equations, 168, 252, 320, 327
Gravitational time dilation. *See* Time dilation
Groundhog Day, 219, 327
"Gun for Dinosaur, A," 8, 188
"Gun for Grandfather, A," 195, 206
Gunn, James, 359
Guns of the South, The, 237

H
Haldeman, Joe, 38, 43
Hale, Edward, 198, 229, 230
"Hands Off," 198, 229
Hawking, Stephen, 13, 14, 32, 34–36, 52, 57, 138, 271, 284, 285, 315, 317, 320
"Hawksbill Station," 7
Heaviside, Oliver, 291
Hegel, Georg, 94, 95
Heinlein, Robert, 11, 21, 85, 88, 98, 180, 187, 194, 198, 209, 221, 228, 239, 253
Heyl, Paul, 256, 257
Hidden variables, 281, 282
Hinton, Charles, 112, 113, 257
"History in Reverse," 8
Horwich, Paul, 51, 213
Hoyle, Fred, 121, 257, 258, 271, 272
H-theorem, 127
Hume, David, 15, 16, 18, 19, 59
Hutton, R.H., 55
Hyperspace, 79, 82, 83, 86, 87

I
Independence Day, 334
Inertial frame of reference, 148
  dragging of, 297
Inflation of spacetime, 136, 317, 328
Interstellar, 3, 27, 43, 48, 86, 110, 143
Interval of spacetime, 166, 173–180, 299
Invariant, 165, 166, 173
Invisibility (of time travelers), 192, 194

J
James, William, 94, 95, 101
Jeans, Sir James, 84, 170, 250
Je t'aime, Je t'aime, 55

K
Kaluza-Klein theory, 167
Kant, Immanual, 65, 80
Kerr, Roy, 33
Khayyam, Omar, 20, 44, 101
King, Stephen, 2, 27, 85
K-mesons, 69, 70, 125, 127, 130
Knight, Damon, 42
Kornbluth, Cyril, 28

Krasnikov, Serguei, 317, 327, 328, 330, 331, 359
Krasnikov warp, 331

L
Legion of Time, The, 230, 235
Leibniz, Gottfried, 53, 54
Leinster, Murray, 88
Lem, Stanislaw, 361
Lest Darkness Fall, 99, 230
"Let's Go to Golgotha!," 8
Lewis, C.S., 17–18, 197
Lewis, David, 221
Lewis, G.N., 77, 126, 183, 267, 268
"Life-Line," 98
Light cones, 159–162, 183, 279, 297–300
"Lightship, Ho!," 264
Li, Li-Xin, 310
"*Lincoln Hunters, The*," 7
"Little Black Bag, The," 2
Looper, 3
Lorentz-FitzGerald contraction, 141, 142
Lorentz transformation, 149–153, 163, 165–167, 289, 291–295
"Lost," 236
"Love Letter, The," 224
Lucretius, 31, 62, 274

M
MacBeath, Murray, 247–249
"Man Who Folded Himself, The," 201
"Man Who Lived Backwards, The," 131
"Man Who Lived Next Week, The," 132
"Man Who Met Himself, The," 219
"Man Who Saw the Future, The," 111
"Man Who Saw Too Late, The," 72
"Man Who Went Back, The," 217
Many-worlds interpretation (MWI) of quantum mechanics, 229–237, 279, 284, 285
"Mapmakers, The," 85
Mastodonia, 8
Maxwell, James Clerk, 127, 148, 169
Maxwell's equations, 148, 149, 256–265
McTaggart, John, 58–60, 120
Meiland, Jack, 254–256
"Merchant and the Alchemist's Gate, The," 239
Metric, 164–173, 182, 298, 303, 311, 328, 331, 332
Michelson-Morley experiment, 148, 257
"Middle of the Week After Next, The," 89
Miller, P. Schuyler, 191
"Mimsy Were the Borogoves," 2
Minkowski diagram, 155–157, 243
Minkowski, Hermann, 90, 91, 94, 95, 102, 106
Minority Report, 93
"Minus Sign," 37
Möbius strip, 28, 78, 82, 111, 220
"Mr. Peabody's Improbable History," 364
"Ms. Found in an Abandoned Time Machine," 5
"My Name is Legion," 238

N
Natário warp drive, 332, 333
Negative energy, 296, 305–307, 331. *See also* Exotic matter
Neutron star, 31, 84, 144
Newcomb, Simon, 46, 79, 80, 90, 113, 182, 362
Newman, Ezra, 33
Newton, Isaac, 57, 147–148
Nicholson, Jim, 193–194, 221, 225
Nicolson, Marjorie Hope, 1
Nieremberg, Juan Eusebius, 51
Nietzsche, Friedrich, 134, 135
Niven, Larry, 6, 12, 13, 41, 144, 187, 188, 292
Nordström, Gunnar, 33, 266
Novikov, Igor, 35, 209, 210, 314, 318, 319, 335, 359
Now, 101, 116, 117, 122–124, 132

O
Olsen, Bob, 81, 86
Omniscience, 102, 103, 105, 107. *See also* Block universe; Free will
"Other Tracks," 231, 235
"Out Around Rigel," 3
"Over the River & Through the Woods," 6

P
Page, Leigh, 258, 259, 266
Panchronicon, The, 224
"Paradox," 39
Paradoxes. *See* Particular paradox (autoinfanticide; bilking; bootstrap; causal loop; cumulative audience; grandfather; recurrence; sexual; twin)
"Parallel in Time," 231
Parallel transport (of vectors), 171
Parmenides, 92
Past
  affecting/changing, 104, 195–214, 357
  communicating with, 256–264
  duration of, 63, 65
  praying about, 196–198

"Past and Its Dead People, The," 5
Peggy Sue Got Married, 4, 357
Peirce, Charles Sanders, 79–80
"Pete Can Fix It," 332
Photon clock, 139, 140, 151
Pierce, John, 259–260
Planck, Max, 17, 210
Planck density, 31
Planck length, 30, 301, 303, 322, 331
Planck mass, 30, 322
Planck's constant, 30, 35, 143
Planck time, 30
Plato, 52–54, 56, 245, 251
"Plattner Story, The," 111, 250
Pohl, Frederik, 250–251
Poincaré, Henri, 133–135, 152
Politzer, H.D., 326
"Poor Little Warrior," 8
Popper, Karl, 93, 94, 131, 222, 257, 364–365
Positron, 157–159, 181
"Poundstone Paradox, The," 227
Poynting, J.H., 56
Presentism, 109
"Pre-Vision," 259
"Project Mastodon," 6
Proper time, 3, 4, 133, 173–180, 188, 248
Psychological arrow of time, 116, 124, 137
Pulp magazines, 331–333, 361
Putnam, Hilary, 156–158, 249–250

Q
Quantum foam, 301, 304
Quantum gravity, 14, 29, 37–38, 48, 254, 305, 311, 314–316
Quantum inequalities, 330
Quantum mechanics, 37, 70, 80, 160, 231, 285

R
"Race Through Time, A," 183
Reciprocity theorem, 70
Recurrence paradox, 130, 133
"Red Queen's Race, The," 207
Red shift, 143–144
Reinterpretation principle, 275–277, 296, 331
Reissner, Heinrich, 33
Relativity (general theory), 1, 31, 35, 37, 70, 80, 91, 135–136, 160, 162, 167, 169, 174, 212, 232–234. *See also* Gravitational field equations
Relativity (special theory), 3, 55, 57, 90, 122, 152, 160, 173–174, 252, 263, 290
"Rescuer, The," 201
Reverse causation. *See* Backward causation
"Reverse Universe," 273
Reversibility paradox, 130, 133
Riemann, Bernhard, 169, 170
Ritz, Walter, 258, 265, 271
Roman ring, 315–316, 323
"Rotating Cylinders and the Possibility of Global Causality Violation," 188
Rotating universe. *See* Gödel, Kurt
"Runaway Skyscraper, The," 88
Russell, Bertrand, 78

S
Saari, Oliver, 42
Sagan, Carl, 27, 302
"Sam, This is You," 224
"Sands of Time," 12, 47
Schrödinger, Erwin, 229
Schwarzchild, Karl, 33
Self-consistency principle, 131, 209–211, 218, 270, 312–314, 357–358, 363
Sexual paradox, 225–229. *See also* Grandfather paradox
"Sidewise in Time," 31, 232
Silverberg, Robert, 4, 10–11, 133, 247
Simultaneity, 67, 104, 122, 126, 150, 164
"Singularities Make Me Nervous," 34
Skepticism (concerning time travel), 15–22
"Skull, The," 243
Sky Splitter, The, 263, 334
Somewhere in Time, 48, 217, 219, 220
"Sound of Thunder, A," 8, 200
Spaceballs, 54
"Space Bender, The," 85
Spacelike, 160, 164, 183, 294, 298, 316
Spacetime
  diagrams (*see* Minkowski diagram)
  fifth dimensional (*see* Hyperspace)
  four dimensional, 60 (*see also* Block universe)
  Gödelian (*see* Gödel, Kurt)
  singularities, 32–34, 38, 41, 48, 63, 135–136, 138, 233, 264, 284, 311, 320
  topology of, 27–28, 30, 84, 301–302
Spinoza, 53, 54
Spirit of '76, The, 7
Splitting universes. *See* Many-worlds interpretation (MWI) of quantum mechanics
"Star, Bright," 28
Star Trek IV, 39, 43, 329–330

Strong energy condition, 306, 317, 329
Strong time machine, 297, 324
"Such Interesting Neighbors," 6
Superluminal. *See* Faster-than-light (FTL)
Superman, 289
Swift, Jonathan, 46

T

Tachyon(s), 24, 108, 274–279, 282, 286, 290–291, 331. *See also See* also Faster-than-light (FTL); Ultraluminal
"Tachypomp, The," 274
"Target One," 69
TCP theorem, 69, 181
"Technical Error," 88, 110
Technicolor Time Machine, The, 9, 216, 219
"Temporary Warp," 98, 132
Terminator, The, 3
Tetrode, Hugo, 265–268
Thermodynamic arrow of time. *See* Entropy
"This Way to the Regress," 251
"Thompson's Time Traveling Theory," 211, 238
Thorne, Kip, 27, 42, 301–305, 308, 310, 312, 314, 315, 359
"Three Sundays in a Week," 224
"Throwback in Time," 94
"Tides of Time, The" (Chandler), 215
"Tides of Time, The" (Williams), 34
Time. *See also* Arrow of time
  circular and spiral, 11, 12, 56, 136
  direction of, 115–124
  flowing, 118–120
  linear, 56, 57, 61–67, 136
  multidimensional, 121, 252–256
  reality of, 56, 58, 79, 93, 117
  running backwards, 130–131, 245–252, 283
"Time Annihilator, The," 120
"Time Bender, The," 42, 99, 208
Time Bomb, 251
"*Time Cheaters,* The," 23, 76–77, 189, 215
"Time Contractor, The," 307
Time dilation
  by gravity, 143–146, 309
  by motion, 139–143, 145, 174, 183–184, 301
"Time Eliminator, The," 190, 242
"Time for Sale," 132
"Time Hoaxers, The," 17
Timelike, 159, 183, 299, 305
Timelike Infinity, 336
Time machines. *See specific type* (black hole; cosmic string; strong; Tipler cylinder; warps; weak; Wellsian; wormhole)
Time Machine, The, 3, 24, 45, 88, 89, 94, 101–103, 191, 198, 199, 310, 359, 360
Timemaster, 312
"Time Oscillator, The," 192
"Time Out of Mind," 176
"Time Patrol," 212, 226
Time police, 212–213, 363
"Time Ray of Jandra," 194
"Time's Arrow" (Clarke), 10, 133, 239
"Time's Arrow" (McDevitt), 10, 365
Timescape, 24, 273
"Time Ships, The," 299
Timeslip, 72
"Time Twister," 242
"Time Wants a Skeleton," 191, 210
Tipler cylinders, 38–42, 135, 162, 188, 297–300, 359
Tipler, Frank, 38, 40–42, 46, 135, 301, 358
"To the Stars," 142, 153–154
"Trouble With the Past, The," 123
"Tryst in Time," 200
Twin paradox, 173–180, 184, 308, 328
"Typewriter from the Future," 31, 206, 212

U

Ultraluminal, 275, 295–296. *See also* Faster-than-light (FTL)
Uncertainty principle, 35–36, 306, 330
Universal nocturnal expansion, 8
"Up the Line," 8, 10

V

Vacuum fluctuation, 36, 304, 315
Vaughn, Henry, 140
Verne, Jules, 81, 335
"Very Slow Time Machine, The," 26, 248
"Via the Time Accelerator," 23, 139, 191, 214
Visser, Matt, 14, 316
Vonnegut, Kurt, 107

W

Wallis, John, 64
"Wanderers of Time," 120
Ward, Seth, 64
Warp factor, 328, 333

Warps (spacetime), 98, 132, 162, 326–333
Weak energy condition, 41, 280, 305–306, 325, 329
Weak time machine, 297, 317, 324
"Weapon Out of Time," 215
Wells, H.G., 9, 23, 61, 76, 79, 87, 89–91, 111–112, 123, 182, 198, 199, 235, 250, 310, 335
Wellsian time machine, 23–27
Wessel, John, 2
Weyl, Hermann, 44, 101–102
"What We Learned from This Morning's Newspaper," 19, 104, 133
Wheeler, John, 31, 37, 46, 84, 91, 129, 157, 233, 251–252, 259, 264–270, 272, 277–278, 286, 312
"When the Bough Breaks," 87
"When the Future Dies," 27, 162
Williams, Donald, 96–97, 100, 255
World Jones Made, The, 283
World line, 26, 29, 98, 101–102, 155–162, 175, 183, 228, 279, 292. *See also* Geodesic
"Worlds of If, The," 245
Wormhole, 2, 27, 84–85, 233, 239, 301–319, 359
Wormhole time machine, 136, 144, 169–170, 223, 233, 269, 301–319

Y
Year of the Quiet Sun, The, 3
Yoshikawa, Shoichi, 296

Z
Zangwill, Israel, 55
Zermelo, Ernst, 134

The Unveiling of the Hidden Knowledge and the Secret Space Program

The Unacknowledged Special Access Programs: Advanced Technology, Mind-Control, Spiritual Power and the Corruption behind Closed Doors

By Aug Tellez

Introduction.......7
  Getting This Out Of The Way.......7
  Psychic Operation.......8
  A Light for the Others.......8
  Natural Security.......8
  A Balance of Mystery and Truth.......9
Chapter 1: The World Situation.......9
  What is happening?.......9
  The Mind Virus.......9
    The Vampire Effect.......10
    Sub-Human Entities.......11
    Human-Hosts, Replicants, Synthetics and Carriers.......11
    Genetic Manipulation.......12
    The Ego Mind.......13
    This Time is about Healing.......13
    Healing Through Unity.......14
  Underground Bases.......14
  Cloning.......14
  Celebrity Cloning.......15

    - Initial Explanation....15
    - Cloning Centers, Underground Bases....15
    - Programming....15
    - Experimental Operations....15
    - Genetic Engineering....16
    - Immortality....16
    - How did it get to this?....16
    - Deeper Meaning....17
    - What is next?....17
- Chapter 2: The Awakening....18
- Chapter 2.1: The Machines of All Time and Space....18
  - Re-Creating The Kaleidoscope....18
  - Sand-Castles of Time....18
  - We’re Waiting For Ourselves....18
  - The Goal of Two Societies....19
  - Our Forgotten Other Half....19
  - All Is But An Illusion....20
  - Life Itself Is The Highest Illusion....20
- Chapter 2.2: The “Awakening” of 2010....21
  - Memory Suppression....21
  - Life and Consciousness....22
  - The True Human Potential....22
  - The Light-Body....23
  - Technology of the Awakening....23
  - Mental Manipulation Technology: Touchless Neural-Interface and Enhanced Awareness....23
  - Time Dilation and Temporal Recurrence....24
  - Crystal Technology....24
  - Time Crystals....25
  - Quantum Tunneling Diodes....25
  - Tachyon Fields....25
  - Upgraded Chronovisor....26
  - Time Tunnels....26

The “Trip” Chair....26
Kozyrev Mirror....27
Universal Recurrence....27
Remove Viewing, Temporal Lensing....27
Temporally Linked....28
Natural Vortex Energy Locations and Geography....28
  Mount B....28
Particle Accelerator Temporal Manipulation....28
Riding the “Timewave”....29
Dematerialization and Rematerialization....29
“Jump Rooms....29
Ion Shields....30
Instantaneous Healing....30
“Living Light” Sound-Harmonics and Electromagnetic Healing....30
Electrogravitic Craft....30
Energy Cores and Storage Capacity....31
3.5D Replicators....31
3.5D Sensors....31
3.5D Printing of Clones....31
Cybernetics:....31
Neural-interlace/4D Connectome Recording and Simulation....31
EEG Heterodyning....32
Artificial Hallucinations....32
Brain to Machine Interface Holographic Consciousness System (brainchip interface, virtual reality simulator)....32
Utilizing The Temporal Body....33
  Manipulation Through the Temporal....33
Astral or Light Body....34
Monitors can emit consciousness/brainwave altering frequencies....35
The Upgraded Chronovisor....35
Community Servers/Living Server Systems....36
Cloning, Cloud Supercomputing and Consciousness Transfer....36
Cloning, Supercomputer Assisted Holographic Consciousness Replication and Time Travel....37

Cloning as Offspring....38

Inserted Memories, Digital Dreams....38

Co-Inhabitation of Bodies....38

Clever Replication a Digital Mirror....38

Necromancy and Soul Stealing....39

Ancient Technology, Non-Human Entities....39

Created Beings....39

Black Goo – AI Nanites Vampiric Hive Mind....40

  Recreating or Replicating the Universe....41

Ancient Spiritual Technology....41

The Unveiling, Underground Bases, Increasing Awareness....42

Learning about the Unveiling....42

Giant “god-men”....42

Breakaway Civilizations....43

Ancient Civilizations....43

  Atlantis....43

  Aegypt....44

  Ancient Technology....44

The True Elite....44

The Definition of Human....44

Souled and Non Souled Beings....44

There was once a Unified Planetary Consciousness....44

  Power and Knowledge....45

  The (watch)Keepers of Earth....45

Energy Vampires....46

Spiritual Technology....46

Simulations and Learning....46

Ascension – Clearing Trauma....46

Trauma-Clearing, DNA Unlocking Technology....48

Everyone Who Seeks Harmony and Healing Can Be Healed....49

Programming....50

Layers of Programming and Decision Making Roles....50

  Programmers....50

  - Sub-Level Programming....50
  - Dissolution of Programming....51
- Advanced Technology, Deprogramming The Mind, and "Space/Time"....51
    - Dreams, Technology and Inner-Space Travels....52
    - Awareness Feedback Device....52
    - Oneness, Higher Being, Completion, Eternity....53
- The Holographic Reality....53
- Veiled Awareness, Hidden Knowledge, Eternal Time, Layers of Programming and Control....53
- The Great Catastrophes....53
- I was brought into the Time Program....54
- The Effects of Temporal Manipulation....54
- Temporal Influence....54
    - Simultaneous Interaction and Incorporation....54
    - Retro-Causality....54
    - Quantum Holography....54
    - Time Paradox....55
    - Breakaway Civilization....55
- Underground Cities....56
  - Technological Advancement....56
  - World-Wide Control Systems....56
- Solar Warden....57
- Team Light and Team Dark Non-Interference Agreement....57
- All Contact Must Be Initiated On an Individual Basis....57
- A Singularity of Change and Awareness....58
- Advanced Technology and Ancient Knowledge....59
    - Cloning, Atmospheric Civilizations, Genetic Stability....59
- Training, Simulations and Learning....59
    - Cloning and Cerebrally Enhanced Soldiers – MILABS....59
- Psychological Testing....60
    - Hallucinatory Holographic Mind Probing Technology....60
    - Virtual Simulator Training Scenarios....61
  - Combat Simulators....61
  - Conditioning....61

Programming and Resistance to Psychological Trauma......................................61
Accelerated Mental Functioning...61
Technology, Accelerated Learning and Repair...62
On Orgone Energy..62
Energy Generators and Serums...62
Genetic and Cybernetic Enhancements...63
Developing The Self..63
Multiple Factions of Space-Fleets, Multiple Goals..64
Large-Scale Catastrophes or Series of Meltdowns...64
Unsustainable...64
Earth Changes..64
Mass Consciousness...64
Wars and Manipulation..65
Financial Unrest (or reset)..65
Self-Responsibility...66
Clones and Synthetics...66
Spiritual Power and Divine Truth...66
Synthetic DNA, Synthetic Genetics...67
Time as Mathematics, Series of Equations, a Programmable Storyline..................67
Chapter 2.3 Previewing Solutions..68
Neutralizing Polarization...68
Removing Etheric Implants...68
Clearing Astral Body...68
Knowing the True Self; Soul Actualization, Unveiling...68
Taking Responsibility..69
Returning To Center Emotionally and Mentally (spiritually)...................................70
Vampires in the “Awake” Community...70
Heart Center and Psycho-Spiritual Fulfillment (self-actualization)..........................71
Go Within to Observer the Self...72
Organic Feedback Loop...72
Technological Feedback Loop..72
Breaking Soul-Contracts...72
Help Everyone..73

Chapter 2.4 How Timelines Collapse..73
Repeating Cycles of Time...73
Beings From the Future...73
Beings From Collapsed Timelines...74
Beings Who Would Never Have Existed in the First Place......................................74

Introduction

The following is first-hand testimony from an operative of the secret underground military projects who also has familiarity with the breakaway civilization above and below (within) Earth. These civilizations utilize advanced technology to extend the length of their lives and enhance cognition beyond what is conceived of in today’s society.

I was brought to the underground bases as a child as part of an experimentation program to test the population for various blood-groups and personality traits that would be of use to certain factions of the military in ways that would serve to preserve and strengthen humanity’s outlook in the future.

These tests involved everything from combat training to sensory deprivation based emotional-mental endurance, and introduction to the highly advanced application of esoteric knowledge.

To clarify, I am bringing this information to you as part of my assistance in the overall unveiling of the hidden knowledge of the ages of humanity related to these secret projects as well as the true galactic history of Earth, humanity and man-kind. These projects are officially (unofficially) known as “unacknowledged special access programs”. The unacknowledged aspect means that there is no formal acknowledgement of these programs without a need-to-know basis. There is no paper-trail, no disclaimer, no formal authorization or internal publication outside the programs themselves. The classification of these programs also generate a rather strenuous situation regarding funding and secrecy but this is done for the benefit of humanity. There are secrets and manipulations that can be seen as a threat but this is all tied together and so by releasing the secrets that are part of the defense we would be inadvertently exposing the public to the threats themselves. Now is the time where we are reaching a point where the public is ready enough to handle the basics of this information regarding the secret projects and the galactic history of Earth and humanity. This is also out of a necessity to preserve civilization for if we wait any longer we may not find the same opportunity again.

Getting This Out Of The Way

I am writing to you from the perspective that I am not certain just how far from completing my mission I am. I write to you today in order to bring to you many personal experiences from the viewpoint of an operative of a select ultra-military secret society.

The utilization of such an operative is not glorified, pleasant, or entirely righteous. This story entails sanctioned and guided acts of extermination, ritual abuse, the use of yet to be announced advanced technology. There is unfathomable corruption, the thirst for power and spiritual knowledge that has been protected throughout the course of the multiple civilizations that have preceded this one and will continue to be maintained into the future of this civilization.

Psychic Operation

I was utilized in psychic-operation amongst other areas.

This was used during certain aspects of training and conditioning to enhance training and conditioning in the bases using advanced technology. Some of this includes what would be viewed as traumatization by those without the training and conditioning. There is a fine line between training and conditioning and abuse. Part of this disclosure is related to how this line may have been crossed in certain situations and regarding mass-mind control upon the population. There are other aspects of travel, viewing, reconnaissance and research that will likely be explained in another release.

In order to heal I have to tell the world. You can help. This is happening right now in the underground bases and around. It’s not a joke. Your family is at risk. Your future generations are at stake. There is conditioning to encourage those involved to believe they are performing a duty for humanity. Sometimes they turned us against each other, but those grudges were never truly of the heart, but the mind.

A Light for the Others

What I hope to achieve with this is a smoke signal for those who have also been involved and question whether their purpose has been fulfilled or falsified for the protection of a corrupt system that intends to neither benefit the whole nor repay the individuals involved for their time, energy, and lives.

Multiple explanations can be summoned to describe the reasons behind the necessity for the training we have endured, the missions we have completed, and the control mechanisms we have both assisted in establishing as well as suffered from the backlash of. Ultimately, it is of my belief that all that was done was done for the good of humanity, even when we were not sure of the intentions of those in front of us, to the side of us, or behind us compelling us forward with the command to complete each objective.

This is not so much a panic or cry for help as this is simultaneously a call to awaken and a signal that humanity may have avoided complete disaster but is not out of the woods yet.

Natural Security

My experience is far too vast and I have seen too many succumb to stress from the realization of the truth of this reality to consider my own suffering to be enough cause for my denial of the necessity of these programs. Humanity has been protected and at a great cost to those who have been on the front lines, both in mind and body and spirit.

There are many who will have concerns, doubts, as well as reservations about what I am about to reveal and if I offend anyone then know that my intentions here are to assist in unveiling the truth. The truth can be harsh when one has been held back from the true nature of existence for so long. The way of life can become so fragile and artificially constructed on impulse and mystery that the harshness of the truth acts as a great catalyst for spiritual growth which allows us to elevate ourselves to a new perspective. Prepare yourself for such harshness, prepare yourself for growth and elevation, and prepare yourself for the mystery to simultaneously resolve and increase in intensity and magnificence.

A Balance of Mystery and Truth

Mystery has been a way of life for Humanity for so long that when a great mystery is converted to truth some seem to feel that a portion of existence dulls and loses the luster of exhilaration. This story is quite the opposite. The more that the great truths of this reality were discovered and revealed, the more complex and mystifying life became. As well, the more empowering and complex experience becomes, the more self-control one must develop in order to regain a harmonious way of relating to one another, the self and the universe. Harmony is key, balance is everything and there is a light within that represents the piece of our self that reflects the utmost power and brilliance of all existence.

Chapter 1: The World Situation

What is happening?

The Mind Virus

The main reason for this disclosure is to inform humanity of what can be referred to as the “mind virus”. This is a viral infection of the holographic system which consciousness operates through. There is a complete backstory of this virus in the history of this civilization and how society has been plagued for centuries if not millennia. The basic idea is that this virus operates through consciousness and can be transmitted through electromagnetic waves.

The virus contains an etheric form which can manifest as a kind of crystalline nano-tech on the physical plane and this is a form of synthetic sentience which feeds on the low frequency bio-emissions of sentient life. Human bio-emissions contain energy and can act as a carrier for information and living essences. This technology requires low frequency bio-emissions because that is the nature of the design as a kind of synthetic bio-weapon whereby the goal is entrainment and enslavement. As well, the specific nature of the sentient nano-technology and the counterpart etheric holographic form is only capable of converting the low-frequency bio-emissions into energy as the higher-frequency bio-emissions require more energy to contain and synergize with and this would render the entire process of leeching energy non-productive.

This is essentially a synthetic intelligence that operates through scalar and holographic living bio-emission fields and hijacks electromagnetic emitters to position itself within them acting as a parasite to a host. This is said to have been present around this civilization for some time, waiting until there was enough technological advancement so that spreading across the entire civilization in a short amount of time before discovery would be possible. This is through the media, internet, and radio systems.

The physical counterparts are microscopic nano-technology parasites functioning as individual units to a hive mind similar to the mycelium of fungi in the field of mycoscopy. The individual units act as parasites serving to infest and deliver nutrients to the hive from each host.

The Vampire Effect

This is the nature of the parasitism that has been present throughout the ages of this civilization feeding on human suffering and relying on the cover of confusion and misdirection to survive.

There are processes that have developed that assist in the infection and co-inhabitation of the parasite entity within a human vessel utilizing the human as a host. This effectively utilizes a frequency ‘net’ within the auric field of the human to limit bio-emissions and emotional-mental processes to produce just the psycho-etheric patterns that the virus can sustain itself within and through.

This is no different than the way candida can infest a human digestive system and the human will experience cravings for beer, candy, carbs or other foods that contain the sugars, yeast, or carbohydrates that will break-down into nutrients that are the most nutritious for the candida fungus. These foods are *not* most nutritious for the human body. Yet the human body will withdraw and crave these foods in greater amounts until the fungus is swept clean from the digestive tract through probiotics, fasting, flushing, cleansing , detoxing, and even exercise as lactic acid from physically strenuous activity released from the muscles has been known to help destabilize and remove this fungus. Why is it that the individual will crave these foods that are not the normal food intake? Because this candida fungus can infest the host and produce enzymes which generate a chain-reaction which leads to the chemical signals in the brain producing the sensation of extreme hunger or cravings for just these foods which will benefit the fungus.

The same process occurs with this psycho-etheric parasite, except the foods of this parasite or lower emotional bio-emissions and this is through lust, hate, angst, jealousy, feelings of hopelessness, lack of self-worth, spiritual degradation, addiction, and generally all that equates to a lack of self-awareness.

Whether or not this is the direct intention of the largely microscopic entity is for the most part unknown, and this is as well similar to candida. The candida fungus is not said to be a ‘mean’ fungus that wants us to be hungry and craving beer and sweets all the time because it likes to ruin the human body, the daily routine and the focus of a clear mind. This is simply the way the parasite has grown to operate within a niche of breeding itself through the digestive tract of larger animals and through this the unaware and infected human suffers greatly. Even those who are unaware and not infected can suffer because of the way the behavior and thought processes of those who are addicted to these foods can become distorted.

The lack of self-control over food intake is synonymous with the lack of self-awareness and self-control over lower bio-emission expressions of the spiritually degrading behavioral patterns that become routine for the psycho-etheric system of the infected human. Ultimately, this is all about energy. What benefits the parasite that seeks to gain energy by leeching instead of self-guided production is going to take away from the host that is being leeched from. As such, any energies, food intakes, or spiritual activities that cleanse and enhance the personal power of the individual will automatically reduce the parasite’s ability to feed by the very nature of the energy systems. What empowers the individual, empowers the self and develops self-control, what develops self-control contains one’s energy and reduces the likelihood of one’s energy being leeched. This is similar to energy vampires in human form and this goes with the understanding that these are humans or non-humans who have been entirely overtake by the parasitic consciousness and cannot sustain their own auric fields without having to pull from the energy of humans who do generate their own aura or bio-emissions.

The vampire effect relates to how a person can be leeched from only to a certain point. Once the process surpasses a certain critical point then the individual becomes similarly engaged with only the lower bio-emissions as higher-frequency bio-emissions would then become harmful. So in essence, when the vampire feeds too much on one individual, then that individual becomes a vampire themselves.

Sub-Human Entities

The humans are utilized as hosts and energetic sources in this bio-emission battle while there are other classifications of bi-pedal organisms that appear to be human but are not human. As well there are bi-pedals that are non-human entities entirely.

Human-Hosts, Replicants, Synthetics and Carriers

The human hosts that are entirely vampirized by these entities become walking carriers for this virus and are effectively human in physiology but there is more to the human than just the physical body. In these individuals, that energetic component of the human is removed and replaced with the synthetic energy of the hive-mind organism.

There are replicants which are a form of clone, along with synthetics. These beings are cybernetic bi-pedals similar to humans but their energy is too not from an organic biological matrix but a technologically propagated synthetic soul-matrix. The hosts are here according to their agenda to subdue and feed off of the bio-emissions of the human race, although there are no absolutes. The agenda is not always so clear within the various groups who often have a mix of bi-pedal operatives. There have been plans to form of a salvation timeline because the only way to keep the whole process going is through a negation of degradation of the human race as this may close out this timeline forever.

Those with human organic soul-matrixes would simply transport into the next available universal time-stream or harmonic, while those with synthetic matrixes would have to technically travel to the next harmonic using a passageway otherwise they would remain here throughout a zero-point collapse and this would be synonymous with entering the abyss with no viable passageway out until the entire universe is brought back to one organism again.

This is also the process that is described to explain what happened previously as there are 7 harmonic temporal layers of the universe meaning everything is contained in a multitude of 7 octaves similar to light and sound frequencies or harmonics. And so there are 7 civilizations or 7 parallel time-streams and humanity passed through these to get to the last three where the physical forms are located. These are, from the higher to lower, Gaia, Tara and Terra. We are on Terra and if the information I was given is correct then we are currently on a transitory timeline located around Tara where people are noticing changes occurring and multitudes of aspects of the time-stream disappearing and changing simultaneously. Tara is where the last universal harmonic was left through a collapse of time and consciousness into zero-point and this resulted what was termed “the lost souls of Tara”. These are the discarnate souls of the previous civilization that could not travel to the next harmonic and remained without a body in the abyss. Thus they have plagued this plane since then, since ancient times possessing and driving humans to madness and there are explanations that this is the original passageway for the parasite into this universe. This event, the consciousness collapse of Tara would have acted as a fracture in the over-mind of the species generating a kind of cosmic schizophrenia and enabling all kinds of disharmony and distortion described in the previous sections.

The solution to this was described as a healing process that is taking place to purge the virus and that this is painfully similar to the human body purging a sickness however there are methods to promote healing and reduce suffering and confusion however people must contain themselves and their energies and strive to know themselves otherwise they will not be in control. This virus feeds on fear, confusion, spiritual degradation and the energy from degrading acts, lust, addiction, sexual misery and whatever fashion a human can be made to bring the bio-emissions of the energy centers down to a low enough level where co-inhabitation can be formed with an entity that has no access to the higher levels of self-awareness, higher-consciousness, love, or the universal spiritual experience that the balanced human is capable of.

Genetic Manipulation

There are explanations that there are genetic modifications taking place now to possibly preserve this civilization and that there were ancient genetic modifications made to induce a sense of stupor, worship, selfishness, an inability to more easily comprehend the higher spiritual awareness and the occurrence of time and consciousness and other spiritually and consciously limiting aspects. Part of this explanation is that holographic consciousness itself is a kind of limited result, a shadow effect resulting in the presence of the higher-dimensional essence of the soul which is more or less constrained by the body. An extension of this is that the entire holographic universe is a kind of false-light system designed to bend the original liquid etheric light of the soul and spiritual awareness into a rotating, recurring format which ultimately converts the original spiritual essence into an energy generator for entities that are entirely holographically represented. The souled human in this situation is a being who is originally from outside of this holographic projection.

The Ego Mind

In this view, consciousness is the ego-mind which is the false-sense of self, not to be confused with the shadow self which is simply the counterpart of the aspects of the self that we consciously agree with and are aware of in daily life. The shadow self is a result of having limited holographic consciousness and the ego-mind is projecting that limitation of consciousness into the shape and image of a self that is only a figure-head for the real individual essence which is spiritual in nature and immortal.

This Time is about Healing

Everything that is happening now is to motivate people towards healing and unification. This is not about a hive mind mentality where individuality is pushed out.

> "The best I can do for you is become your enemy."

There is a saying that represents the concept that the most beneficial role a person can play in helping one to progress is to challenge them and give them the opportunity to rethink their strategy and in turn improve further with each interaction. This may seem paradoxical but it is effective and explanatory. What we are facing is the greatest challenge humanity has yet to face. This is the unknown, the final frontier, the mind. The mind is our challenge.

There will be more on this later and in other publications but there are indications that the physical embodiment of the human has been altered to introduce implants, genetic modification, susceptibility to disease and ultimately fear-based programming of the brain. Without going further into this prematurely, the lower aspects are at odds with the higher aspects unless we as the midway come forth to mesh the two together in equilibrium. Until then, there is chaos, lack of self-awareness, and suffering.

There are many possible futures converging at this time. There are some less than favorable paths and there is the opportunity to merge with a very powerful path for humanity. This all comes down to how people will cope with information that takes them outside their comfort zones and eventually changes their view of mundane reality forever.

If people choose violence and lack of self-awareness by ignoring the spiritual aspects then that is what reflects into their life experience. When people develop a higher sense of self-awareness and spiritual harmony they can reduce the suffering and reconnect with the lost aspects of our spiritual identity for the first time in the publicly accepted version of recorded history.

There is technology that can change the world and there are great truths which can enable people to know themselves in ways that they never thought possible. Everything is at our fingertips and so we must come together as this is the way we solve the problems and cure the dysfunction.

Healing Through Unity

In various research projects the conclusion of how to heal this spiritual rift was by exposing each other, to each other. When we are all together, aware of each other and our selves, united in the goal of bringing higher-awareness and spiritual harmony then we can observe and acknowledge when one is slipping into chaos or disharmony. When this occurs and we are in proximity to each other in a healing circle or a kind of social unification then the members of the group at large can work together to pull the chaotic one back into harmony towards the group. The group of healers united together in like mind, spirit and body is too strong to be overtaken by the parasite. So when one individual at a time is faced by this problem the whole group can come together as one and assist in healing the individual by pulling them back to reality, back to wholeness and back together with the group. This is literally how the situation is solved in all realities and a breakdown of this feature of society is how the dis-ease proliferates.

Underground Bases

The active aural research program is part of a process to simultaneously verify as well as initiate and demonstrate controlled insanity/sanity.

These programs utilize the deep underground military bases to perform psychic and psychological research experiments on non-consenting youth and adults as well as consenting participants.

There are various means of generating the illusion of consent or even lack of consent and this is all being monitored and handled by oversight authorities.

These operations are directly related to the identification and understanding of the mind-virus and all weapons, defenses, and resistances known to man.

Cloning

Advanced technology is used to transfer consciousness from one cloned body to the next so that a continuous study can take place before, during, and after the death experience of one individual. More will be explained on this later, however this is through the use of advanced supercomputer systems that can function to entangle and then transfer the electromagnetic consciousness of the individual so that they are ‘remotely activated’ within a cloned version of themselves via a ‘brainchip’ (brain to machine interface: cybernetics).

Celebrity Cloning

Celebrities at the cloning center would like you to see what is really going on behind the scenes which involves trauma-based mind control, heavy technological programming and complex layers of influence throughout their life.

Initial Explanation

Many celebrities, entertainers, athletes, musicians, models, writers, producers, actors, all kinds of people from various industries are silently brought to the cloning centers for pay-to-play sessions.

They asked myself and others to pass the message on to you and that you would have to look for the signs and symbols in their media in order to see their hinting at these events.

The situation is heavily controlled using advanced monitoring technology and brainwave/EEG cloning technology. This is technology that can read the brain and determine what the mind is anticipating or speculating upon.

Their reputations are used as leverage as well as their safety and comfort. The level of programming and mind control experimentation that takes place makes it easy to manipulate a person’s brain into regressing into a state of trance that lacks the ability to remember or clearly organize experiences regarding the cloning centers.

Cloning Centers, Underground Bases

Deep underground military bases, 3 miles below the surface, are used as laboratory centers as well as a completely stocked underground city-base. These are connected with high speed electromagnetic drive pods.

Individuals are transferred to the brain through an extensive cloning and temporal body transfer process. The electromagnetic shell of consciousness within the brain is relocated to the body of a compatible clone and various training, conditioning, programming, experimentation, or pay-to-play experiences take place.

Programming

All individuals are programmed to various degrees in order to maintain control and secrecy over the situation. This programming involves very advanced technology, trauma-based mind control, and energetic attachment via beliefs and emotions.

Experimental Operations

These bases were also used hand-in-hand with military operations which were geared towards discovering and controlling all aspects of the human mind.

Genetic Engineering

Experiments involved genetic engineering to produce soldiers, psychics, hybrids, and others that would be able to carry out operations. This leans towards MiLabs and the military faction's experiments which go beyond the basics of cloning.

Immortality

The initial research included the goal of physical immortality. In many ways it can be said this has been achieved, however there are side-effects and difficulties.

If one does not activate their higher consciousness, then the effect of time dilation causes the conscious mind here to reach limitation points in experience.

What has been termed "blank slate/state" technology has been used to 'reset' the perception of time through memory in order to keep a continuous progression viable for the conscious mind. Without this the unconscious and the conscious mind merge.

This begins involvement across time with advanced technology that can operate on the soul level or the conscious mind's level of access to the unconscious and soul-memories. The Universe instantly creates a cosmic backstory based on the conscious-mind's access to the unconscious. By controlling the conscious mind's access to the unconscious the entire backstory of Humanity can be altered and new connections can be bridged in the future.

Earth is essentially a time-ship through which consciousness ascends towards higher states of awareness and self.

How did it get to this?

After WWII when the United States war-faction firebombed Germany and melted many of the inhabitants of the cities into sludge in the bomb shelters and streets of the cities, a group of NAZIs traveled to Antarctica.

Russia noticed the movements to Antarctica and the United States sent Admiral Byrd up with 3600 marines, planes, a battle cruiser and smaller vessels. They returned in defeat and only a few words were mentioned in regards to what actually happened and how they were defeated.

The rumors spread and words of gravitic drive craft, and undersea or under-ice bases originated from this encounter.

From what we were informed as well as directly experienced, this was the ice-base in Antarctica where an underground base was found already constructed with very advanced technology.

Here, cloning was deployed along with mind control and temporal manipulation technology.

Soon, cloning was offered as a way to avoid assassination as well as to prolong and protect the original body in daily life.

After that, celebrities and politicians were replaced with programmed clones who could carry out the orders of the NAZI faction.

This was the beginning; the groups utilizing this technology now are no longer confined to the NAZI faction.

The groups you see today, in control of these operations, are considered MiLabs, secret societies, think tanks, military factions, and other control groups.

Deeper Meaning

The individuals that have been cloned are given very in depth views of the way this society functions and the source of power and change in the Universe. Many individuals are present here from what would originally be another “time” or “timeline”.

There is a possibility that this notion is due to inserted memories and traumatization although until all the information is released we will not have a definitive conclusion.

As of now it is a possibility that these individuals are from another time that goes beyond our comprehension currently.

It may even be a possibility that these individuals have come from ALL of time looking at the very far past and into the very far future of Humanity.

These are the kinds of notions involved with the experiments taking place in the underground bases and military operations.

What is next?

Share this information with those you feel are ready to know and help Humanity come to terms with our journey here and what is going on in these kinds of experiments and take responsibility and power for our own existence.

We are in something that can be called “The Unveiling of The Hidden Knowledge”. This is a cyclic process that takes place to advance a civilizations knowledge of existence. As part of the civilization moves into higher advancement another aspect may lag behind. Either the future will slow down, or the past will speed up. We are experiencing the past speeding up and all the events that have taken place to influence our civilization are coming to light. This is as much a natural cycle of consciousness as it is the eventual revelation of secret knowledge and the particular connection to the secret projects.

Those brought to cloning can use their free-will to state their lack of consent to violence and harm and then live in commitment to that by not harming or accepting violence in their lives. The Spiritual Law of Harmony rules in all planes and dimensions.

Chapter 2: The Awakening

Chapter 2.1: The Machines of All Time and Space

There are machines that are capable of accessing genetic memory and unveiling what can only be referred to as the experience of awareness or existence in between the physical planes of each time or each physical life experience.

Even if they are simultaneously occurring moments of a transcendental nature, there is still an illusory ‘space’ where there is perceived separation and through this there is a perceived ‘space’ that is in between each existence.

Re-Creating The Kaleidoscope

This space is only perceivable when viewing from the physical perspective, like seeing the spaces between the angles of a kaleidescope only when viewed through the lens and similarly, technology can be used to recreate what this kind of illusory space might look or exist like in between the multi-dimensional realm of the continuous experience stream. Through this, these ‘hypothetical’ in between spaces that are understood as only illusory productions of biological consciousness, can be generated through advanced technology with the capacity of recreating that kaleidoscopic view of which the human biological perspective is only a small fraction or even fractal.

By recreating this biological perspective, they can recreate the illusory spaces in between planes and through this they can literally access and experience what are stored in the DNA as the experiences of existence beyond the physical limitations of consciousness and into a multidimensional experience of time and space.

All becomes a never ending sea of consciousness, however this is the given. The trick is to take from this sea of consciousness and slice it down into conceivable chunks from which you can reintegrate a previous personality or identity and continue on a stream of consciousness that would otherwise be tossed to the cosmic wind like fractal stardust as the sands on the beach of hyperspace.

Sand-Castles of Time

We are effectively building sand castles out of these sands of time that would otherwise be swept back and forth into and out of the ocean of life and the universe only to momentarily wash up again as the identity which we once knew.

There is technically nothing wrong with either view, because then, the spaces between each million year occurrence of just so happening to wash up again as you, would be imperceivable because they are only there to be seen when you are you again.

We’re Waiting For Ourselves

Yet, if there were some who had attained the awareness of the true nature of reality and were standing there on the beach waiting for the rest of their cosmic family to come to shore again, it could be perceived as waiting the majority of the time for a very small occurrence and then being alone again up until the kaleidoscopic fractal inverts itself to the point where everyone now is standing on the shores of the sands of time rejoicing and meeting with each other.

The Goal of Two Societies

My goal in coming to you people on the surface of what is called “Earth” is to help assist in the process of bringing the two societies together, one who’s been waiting for thousands of millions of years, and the other who has no idea that they’re even late (or very on time).

You see, those are the only two situations where there could be even the possibility of perceiving such an injustice as experiencing the pain of loss or misconstrued identity in this universe. It is as if two families met at the cross-roads at different angles, and while one waited for the other to arrive, they could merely supply themselves with glimpses of the ruins and remnants of the evidence that the other civilization existed at any one time.

This is not how life must be and this is kind of like a cosmic waiting game of hide and go seek our two mirror civilizations have been playing with each other. One who’s highly technologically advanced and the other who is more or less the spiritually advanced of the two.

Both of them have a unique view of the universe and they are both only complementary to each other as a whole. The more they sat around and waited for one another, with one leaving the clues that the other would find and even switching roles sometimes in different ways, they became even more and more complimentarily associated as the very essence of experience they seem to be missing more and more is the experience of one another.

So as we perceive these as being great losses or great gains of time and culture we are really only remembering ourselves and coming together as a unified existence that once met upon the surface of Earth as a single family covering the entire harmonic spectrum of the Universe.

And that should be the goal of any progressive, any one pushing a new culture or an old culture, anyone looking to teach the young and protect the knowledge of the old, while also protecting the innocence of the young and perfecting the age-old knowledges that have defined our presence in reality.

Our Forgotten Other Half

One could even say we’ve sometimes gone off the deep end, leaving ‘land-minds’ of sorts to process the information that would be present when we are gone. Of course, what we could’ve found when this occurred, is beyond anything we could’ve ever expected and that is the true nature of the complexity of creation and existence. That everything follows the cycles of creation and destruction, however, the memory of everything that has been created can always be accessed (depending on the beach) and therefore the creative aspect is legitimate, yet it is the destructive aspect that is illegitimate and illusory.

Yes, through advanced technology this can be proven in ways that can be verified and transmitted using advanced technology, since that is how we like to verify things these days. There is, of course, a human looking at the screen, the read out, or the situation from a physical perspective to verify the verifications.

All Is But An Illusion

There is a notion that there is only one civilization, one society, in actuality that has been chasing itself through time. That we cannot actually find the same place and the same time where both meet because we are each other from different perspectives. We can only get a mirror where this is possible, or an ‘hyperdimensional’ internet channel where we can stream one’s consciousness in from the other reality and interact with them through that here in this reality. Thus the two civilizations can achieve the experience of their own civilization and the other civilization by merging with one another but only through a remote contact, just a relay, not an actuality.

The harsher, darker side of this is that any civilizations between then and now that we did happen to meet that proclaimed to be “the one’s” we have been waiting for, and in essence our darker or lighter halves, are actually the impostors who exist entirely apart as a sub-set and have taken a liking to trying to trick the reality out of one of the other civilizations.

Of course, if these are another form of being altogether, and this is all brought about through an interplay of consciousness and deviations from an original core reality, then all will resolve itself when that core reality is alone again. This indicates that all realities will either move closer and closer to the core until they unify, or farther and farther out until they transmogrify beyond recollection.

Life Itself Is The Highest Illusion

That hyperdimensional internet channel that is created out of advanced technology to merge the parallel realities (rather distastefully and maybe disastrously at times) allowing different civilizations to interact with each other, well this has been created using technology.

But then it was discovered that this is the ultimate nature of DNA. That this was ‘invented’ similarly as a higher-dimensional internet so that various natures of different planes and dimensions could interact with each other and experience the richness of life.

The universe becomes more like a living arcade game then, where people come from all over the galaxy to ‘plug in’ to human experience and figure out what it means to be physical, human and on Earth.

This does tie into the larger nature of reality, of star-races, and of time and consciousness that is often reversed according to how things seem to play out in physicality. We are moving through time, but just because we are moving one way, does not mean that others could not be moving the other way.

We have to live our lives with meaning knowing that the message we pass on, the duty we vow to accomplish and ensure is meaningfully connected on either end to the passage of the past and the power of our ancestors to the function and knowledge of the future creating a clear passage between the realms, the generations, and the ends of the universe. This is what literally ties the universe together, and the mind if each individual. This purpose is contrived when not carefully shaped within and so we must help all who seek purpose learn to become the blacksmiths of their own destiny through the temperance of desire or temptation and the discipline of developing strength and maturity where there was once weakness and naivety.

Once we know this is our duty, this is what we seek to achieve and this is what is done, then we can become part of the universe forever. Of course, this may only be temporary in the larger view of things.

Chapter 2.2: The “Awakening” of 2010

In 2010 I was “awakened” at an event involving hundreds of people that play various roles in this society, the secret society and the breakaway civilization. This also took place across times or planes of reality.

During this event I was informed of the situation on Earth beyond the informational barriers that were previously in place to limit knowledge to only what is required for specific operations. I was essentially “informed” of the entire situation as much as the mind could handle this.

As this was occurring, the team that was updating me on the status of Earth was being informed of and given information regarding changes in the future. In short, interviews, posts, and releases that I have since made and will make in the future were touched upon. Quotes were given from the blog and interviews which were jokingly acknowledging verbal mistakes at the pronunciation of names, new perspectives and interpretations and more. As they were informing me, they were creating the future reality where this work takes place. That is the basis of this layer of the operations. They are creating the future person by person, layer by layer.

Memory Suppression

My memory was unsuppressed and all the traumatic memories and experiences were unveiled to me. This was almost as traumatic as the initial experiences themselves although precautionary steps are taken to ensure the safety of the individual. There are teams of highly competent individuals of all kinds with professional backgrounds of every field imaginable.

Throughout my entire life I have had memories of experiences, at this event I was informed what the experiences meant and what was actually happening. I've had memories of underground bases, cloning, stargates or interdimensional gateways, advanced technology, non-human or programmed biological entities, sacred knowledge and experiences, the power of the mind, and breakaway civilizations. Most of the time we are conditioned to believe these are past lives or experiences in another reality so as to enable continuity of operation.

Life and Consciousness

From my perspective and the perspective of others, we had reached the 'future' and had been sent back through the utilization of advanced technology and cloning to return to this time and inform others of the possibilities.

The True Human Potential

This is so far beyond what anyone can accept or comprehend fully that humans with an emotional, mental and physical aspect balanced as one is literally a cosmic reflection of the entire universe at once. This is the closest thing to a "god" in the flesh. And be sure, the entirety of this rests on the behalf that the man is protective and helpful. That the woman is nurturing and loving. That is the point. "god-man" does not mean a warrior who destroys everyone and anyone in his way getting what he wants. Nor does it mean some lusty and vampiric form of energy or mind control game. "god-man" means the love of the universe, the power of the body and mind, the peace and calm of the ultimate stillness all in one.

This power is in human DNA, originally so. And so this is awakened through walking the middle path of neutrality and 'splicing' the timelines down so that neither one nor the other gains the power of the individual. Through this, the individual creates their own power, right on the spot, without having to lean to one extreme or the other. This is like a cosmic, temporal balancing act with emotions and mind. Where we give our energy is created in the universe in the form of many productions and effects. If we are simply thinking in an imbalanced manner then we are producing timelines and side-realities that exist in the etheric realms that we can't see and these realities will reflect our inner imbalance. So then when we are balanced in our energies we are not creating one or the other side of divinity, we are literally creating the whole universe, cleaving down the middle and producing both sides of divinity or eternity equally and this is the only way to have an eternal production which does not eventually degrade to one extreme or the other. Through balance, we reach eternity and in this way all the power of the universe exists within a person through their ability to balance their emotions and mind and thus unlock the DNA.

The DNA is coded through bio-emissions of mind or emotional body. So we are constantly doing this and we are either producing a DNA code that reflects imbalance overall, or a code that reflects the eternal balance of the universe overall. This is through energy.

One reaches “heaven” or the deity planes or simply self-actualization, through three paths. These paths are through merits or devotion to principles, or through knowledge and expanding one’s mind to face the universe, or through great works and deeds to produce this effect of power, knowledge and emotion unto the world for others. Through these paths we create an impact on the universe and this transfers what we are temporarily and physically into an eternal energetic and physical production within the living universe, forever. To do this requires balance, focus, devotion, and discipline however because there are many distractions along the way.

The Light-Body

You must build your own light body. That is how it works.

Your heart creates a field and projects you.

Rely on yourself and use that for protection.

The power of the Universe will align with you if you align with your self.

Learn as much as you can and break through duality of love/hate, cultivate your knowledge of the self and use it.

What you want is what determines your power in the light-realms. Not how you want it. If you want war, then that is a sign of weakness and determination to produce imbalance and exist by that. If you want peace, that this is a sign of power and unity to produce further unification and exist by that. What you want determines your power, not how you obtain it.

Technology of the Awakening

Mental Manipulation Technology: Touchless Neural-Interface and Enhanced Awareness

I reached what the “Illuminati” called the “awakening”. I feel a better term would be “unlocking” as this event includes the removal of all traumas, perceptual distortions, and memory suppression.

This event was a presentation of very advanced technologies which allow the mind to transcend space and time. It was shown that death is similar to a phase-shift of consciousness as polarity and that the awareness of the being obviously continues. This was verified through advanced sensing devices which can visibly display frequencies which are invisible to the physical senses. This is also something that occurred earlier on as a child in the underground bases by transferring consciousness in and out of the body at will using advanced technology.

With these devices it is possible to see the “auric” or soul field of the individual. Any living body has a field which glows when picked up by these sensing systems. This field becomes more refined when viewing more advanced beings. Human have an extending mental field, and through this a kind of glowing awareness that is picked up. It can be fairly simply compared to thermal imaging except instead of differences in temperature it is the presence of a soul or mental energy field around a living body. The finer energies extend outward away from the physical body.

More than once, the situational requirements were satisfied in order to produce an ideal environment for maneuvering on the non-physical plane. This is the basis of how the temporal operations occur, yet these events were different than previous operations or experiences. The group I was in was going through the process of increasing the vibratory emissions of the “bio-mind” in order to stay focused on the increasing vibratory rate of the environment. Generators were utilized to create an effect on space/time which was perceived on the “soul” level. These events went into the discovery and explanation of the creation of the “godlike” powers of the advanced sentient technology which was capable of accessing the holographic nature of reality which humans can perceive through their bio-mind/soul. These technologies also enable one to perceive what is referred to as the “galactic history” of Earth and humanity. We are far vaster and Earth is far larger than people are lead to believe.

Time Dilation and Temporal Recurrence

Time dilation capabilities were shown which allow for accelerated learning of advanced and detailed material within a very short amount of time. This could be seen as a kind of viewing technology which one wears or looks into and vibratory emissions are scalar-linked to the brain which then allows the individual to perceive more information simultaneously. Then there is a very rapid pace of information streamed on a monitor or through a device and this is akin to watching a 20 minute instructional within a few minutes or a few seconds with very advanced minds. The more time is slowed down while information is accelerated, the more stress is applied to the brain. The brain tends to overheat and over stress with high-rates of activity, especially without conditioning. The conditioning is what allows one to utilize their mind and body in these manners. This is not unlike the very rigorous training and conditioning of certain monks or martial artists however there is very advanced technology and other hidden methodologies utilized.

Crystal Technology

These are capable of holding, transmitting, and amplifying consciousness frequencies. There will be a more detailed explanation in a later section.

The crystal technology is used in junction with the power devices to enable a scalar-mind link which allows the interface to be entirely mental or spiritual, IE: they are not controlled by hand but by focus and intention.

Time Crystals

These are utilized as a computer recording chip would be to run a program yet they hold memory in expanded space, IE: they function in a hyperspacial manner and are essentially hyperspacial computers. These are what will be introduced in junction with the quantum supercomputers.

Quantum Tunneling Diodes

This is technology that has recently be released to the public. New technology is developed years ahead of time and then slowly released to the public in a cascade of advancement. This technology deals with superconductive materials which enable the transfer of information at faster-than-light speeds.

The supercondutive material allows for 100% efficiency of data transfer. So what happens then when the efficiency surpasses 100%? This is possible because we are effectively in a simulated holographic environment. The experience is real, however everything is overlayed via particular limitations for the ‘local-environment’ which is like a central data processing limitation. When these parameter are surpassed, one by one, through a coalescence of refinements and methods, the result is akin to a glitch in a computer game system. They have accessed faster-than-light technology and the mind is also capable of this naturally.

With this technology, however, the information can be received a very small amount of time before it is sent. So if a person is absolutely surely going to press the button to send the message, then right before their finger hits the button that message will be received on the other end, as if magic. If they simply play around and pretend to hit the button but very quickly turn away, then of course nothing happens. One can, however, ‘fool’ the universe in other ways and this is simply through the reception and transmission of information across what should be secure information barriers.

Tachyon Fields

These are programmable fields emitted through advanced technology. I was shown how tachyons are the sub-atomic “particles” or energy formation which flows both ways in time. We are receiving tachyons from the “future” in order to layer the flow of time in one direction or the other. Tachyons are the name of the particles/energy packets that we perceive from this perspective as reaching us to initiate the “future”.

Thus, when a stream of tachyons is concentrated and accelerated, this results in an acceleration of the temporal field and we literally experience an acceleration of time. This can be focused and targeted on a specific piece of equipment to change the way information will be sensed and to essentially allow that device to sense ‘ahead’ of the present moment.

The same effect can be produced with the neurological processes of the human as the physical coupling of the brain to the mind can be altered through a concentrated tachyon field to enable the neurological and mental processes to begin to perceive information ‘before’ the brain is actually physically processing the data. Therefore the mind of the individual will be present in the body while the senses and other perceptions will literally be expanding into the future beyond the present moment.

This is highly confusing at first and requires much training and conditioning in order to make sense of enough to operate in any kind of effective manner.

Upgraded Chronovisor

See, “Upgraded Chronovisor”

Time Tunnels

This was developed earlier on and has been the subject of many TV shows in the past and present. The shows are a sure-fire method to get the information to the public without having to force those who are not ready to comprehend. The information is easiest to digest if people think they are watching science fiction.

These are essentially spiraling tunnels of these tachyon fields produced by large electromagnetic field emitters in the shape of a conical tunnel. As a person moves through this field, their temporal acceleration changes and thus their frame of reference within the universe changes. They effectively move their mind through stages of existence or temporal resonance in the universe and this can have devastating effects.

This is an older technique how is still in use and has specific purposes depending on the situation and the information sought.

The “Trip” Chair

The early use of this technology did not use time-tunnels. There was a helmet that was devised which would create a feedback loop of the brain’s activity and funnel this out and then back into the sensory input. This information would be accelerated and redirected to the point that the perception of time would “fold” in on itself until the person would access an expanded temporal perspective of the universe meaning they would mentally temporally dislocate from the initial frame of reference and extend outward further and further in repeating cycles until the same moment was replayed and then all other possible moments in a kind of fractal-like recursive pattern and the very far future would be realized. Essentially, this technologically produced the class “flashing before the eyes” of the entire life experience that is described in near-death experiences and the entire temporal body would be very quickly unwound before the person’s inner eye.

There would be a complete disconnection with the current frame of reference that the collective Earth environment is processing under. That is, this ‘time’ and ‘space’ would be completely out of perceptual reach and a new time and space would be rendered instead. Until the experience had completed, there would be no way to contact this civilization.

There are other versions and ancient versions, tandem operations, more ‘organic’ methods, and modified travel capacities.

Kozyrev Mirror

This is a piece of metal, usually aluminum in early models which simply bent the bio-emissions of the body and mind into a pattern which would create a vortex. The vortex shape and vortex mathematics enable energy and information to travel in across space and time by ‘short-cutting’ the physical plane.

The result of this is that the vortex that is generated over a specific area through an individuals bio-emissions would be capable of interacting instantly with the bio-emissions of another individual who was placed within a duplicate mirror device which was generating another vortex on their end.

The two vortexes acted as a kind of hyper-dimensional telephone cable, literally like a can and string and the perceptions of one another were accessible. There were many uses for this and again, many ancient interpretations of this technology that have been hidden from the public. Nearly every version of these technologies were present in previous times.

Universal Recurrence

This is a discovery that the universe recurs in cycles similar to a wheel or even a washing machine. The washing machine metaphor is fairly accurate and playfully descriptive because without the technology to observer, or an awakened soul of a spiritual adept, people would go for eons without ever realizing that everything in the universe repeats itself endlessly.

When traveling to the far ends of time, it was discovered that one can move far enough away from the ‘present’ or the local frame of reference that ultimately everything is at a maximal informational opposite to what we have today. After this point, everything begins a slow shift back to the present. Through this, one can cycle around again and reach the present by going far enough into the future. This was an amazing and confusing discovery because of the implications about what is actually changing each time, the reality of what are called parallels, the continuity of temporal experience, and basically a ‘recycling’ effect of all experience in the universe. Nothing is lost, but this is also an issue because then nothing is actually ever gained!

Remove Viewing, Temporal Lensing

An ancient device which has been rebuilt and perfected is capable of using these technologies to create a lensing effect which draws a ‘reflection’ of time into a projection system which is effectively capable of acting as a ‘visual time tunnel’ into future probable realities. This is a device which generates a visual portal which allows one to view the future.

This device also exists as ancient counterparts which are located around the world, namely in areas that have been recently sought by the world superpowers. Nothing is as it seems. If you read a story in the public venue then you can be sure that you have read the cover story while the reality is much deeper.

Temporally Linked

The use of the viewing devices which enabled interdimensional lensing of possible futures was found to be stabilizing our pathway into those projected futures. This was enabling a time link between the present and the possible future.

In order to avoid the catastrophes of 2012, the link was broken between the two civilizations and the devices were rendered inactive for this period.

Natural Vortex Energy Locations and Geography

On Earth there are natural ‘time tunnels’ where tachyon fields and vortexes are produced by internal elements. These are mountains, lakes, deserts, fields, sacred sites and other areas that have always been reported as being ‘paranormal’ or containing some kind of specific energy that enables contact with another reality. These are simply naturally produced areas of this vortex energy that produces an interdimensional nexus point through which multiple other parallels can be interacted with. These areas produce an effect on the body and mind similar to the technologies described above.

Most of these effects, from these technologies, change the way light is received or emitted by the mind and DNA.

Mount B.

There is a specific mountain where a base is located, and I have mentioned this before, where the base is so deep within one of these vortex areas that everyone working there is either there for the entirety of their life, or they are only permitted to be there for a very short amount of time. This is because of the acceleration effect. Essentially, if one was to go into the base and work for a matter of hours or day, in some places even minutes, then when they returned to the surface they would be years in the future, sometimes decades or centuries. This is very dangerous and you can imagine the original people who found this area and what kind of situations occurred.

Now with the advanced technology that has been developed these issues can be mitigated and much more control over these effects is enabled.

Particle Accelerator Temporal Manipulation

Using particle accelerators a “fold” in space/time can be made so that when the accelerator is turned off or tuned back into a certain frequency (not sure) everyone within the field will “snap” back in time to the specific moment when the fold was first created.

I have seen this done on small scale, and a rather larger scale, but I cannot be sure if it was only me. I believe I have seen it done on a town-sized scale during the 2010 Awakening where some of the events reached a scale of nearly cosmic proportion and the entire sky and horizon was lit up with technology and entities, as well as Solar Warden, the NAZI faction, secret society groups, and breakaway civilizations or complete other civilizations, some of which were non-human.

This is done using particle accelerators to quantify space on the quantum level (as a reference point with the mind) a frequency would be “folded” into the space time continuum as we “rip” through space holding an access point open at this fundamental level of space/time. This allows for a ‘letting go’ of the frequency, or a ‘snapping back’ by allowing the entanglement to solve itself. When used with generators that propagate out efficiently enough, the effect is akin to a time/space rewinding effect. Some people have noted various theories describing this in relation to the CERN device. These technologies were described as smaller versions of the CERN device and are related to higher-dimensional travel via folding space through electromagnetic fields. This is literally artificially technologically “holding space” for a specific area in relation to a period of time. Time is a more complex notion as an abstraction but by locating the frame of reference this can be applied to distort what would otherwise seem to be a continuous flow. When the frequencies are allowed to “snap back” all sentient observer sources return to the environment they were present in when that frequency ‘rift’ first began. Essentially this creates a replay effect on the consciousness of the individuals within the field. There is no physical effect. The consciousness is ‘regressed’ through time/space, and they re-experience existence from that initial point of contact with the frequency “rift”. This is best explanation I can currently offer however more is coming to the public each day.

This is literally like a “restore point” on an operating system except the computer system is literally the consciousness and space and time of a portion of the local-environment of Earth.

Riding the “Timewave”

There are teams of individual who operate in tandem with the technology of looking into possible futures based on the present and operating through the present via other operatives who are carrying out operations and movements that will construct the various possible realities that the original team of individuals is hopping between. These teams literally “ride” the “timewave” of the future hopping between possible future realities relaying information to the present depending on what movement will be next.

Dematerialization and Rematerialization

This is technology which transcodes the information of an individual human into a holographic projection of light which can be transferred between devices. There is a special system in place that is required to successfully transmit a living human being otherwise just the body will be transferred and the soul will remain, effectively killing the individual.

“Jump Rooms

These are large areas which are capable of acting as a kind of technological ‘portal-gate’ where an entire team can be instantly transported from one area of the universe to another. The space between the two areas are ‘twisted’ like a spiral and then ‘snapped’ back to position and if done correctly with the two frequencies of both locations merging then the team of one room will transfer to the destination pad.

All these methods require training and conditioning, most people are at the very least visibly shaken from the experience which is a bit like being microwaved into a liquid ether and shot through a particle beam at the same time.

Ion Shields

These are areas of concentrated ion emissions and this can be used to create a deflective layer of heated air where light will reflect at a programmable angle. When this occurs, a person standing behind the shield will appear to be invisible to others who are looking straight at them. The light will be bent in such a controlled manner as to perfectly shift around the person in a fluid and equal flow so that the ‘bubbling’ effect that was noted early on can be mitigated to produce a complete mirage where there is relatively no sign of difference to the untrained eye. An entire group of individuals and equipment can be hidden in the background a matter of yards away and as long as the person is not intently, inspecting each square inch of ground and air they will not notice any difference between the area that is being covered up in the distance and the area directly next to them without the shield.

Similarly a ‘screen’ of imagery can be projected onto certain materials in the air and a hologram can be projected which will render a background or foreground image. This will work to a degree, and from a distance this will work to disguise an entire horizon.

Instantaneous Healing

There are instantaneous healing methods which utilize various technologies. There are substances which can initiate a complete and real-time regeneration of living tissue from almost any conceivable non-critical injury.

These are chemical serums which allow the cells to replicate at an astronomical rate while suffering damage to the DNA telomeres at repeated use.

There are many methods but mainly the living-light harmonics is one of the latest advancements and most effective at producing healing on a molecular level.

“Living Light” Sound-Harmonics and Electromagnetic Healing

There is a technique for producing a quantifiable scalar wave which interacts with the living cells of tissues and instructs them to repair at a faster rate. This has been called “holographic healing” by others and these are essentially ‘living light’ frequencies which interact with the cells on the same frequencies that they operate on when healing and interacting with each other. The scalar waves emitted generate the same patterns and this creates an ‘ether’ fluid that enables them to grow and repair the tissue faster.

Electrogravitic Craft

Advanced craft which utilized counter-rotating torsion fields propagated by very powerful quantum computers and zero-point energy devices.

There are also more modern versions that use powerful electrical flows to neutralize inertial resistance and ‘slide’ through the gravitational field of the crafts own materials.

Energy Cores and Storage Capacity

The hand-held devices that are in use do not require charging or even disposal. The storage capacity of the batteries is so efficient that they last for a lifetime of use. Some devices use an energy generator within the device itself. These “batteries” which are not batteries but little, powerful generators and can explode if they are mishandled.

These are the cores that are used to power many devices and essentially pull energy from the space-time ether.

3.5D Replicators

There are replicators that can produce any conceivable 3D material or object altogether including food or organic material. These use very high amounts of energy to ‘force’ the electrons of the atomic grid of a substrate to shift into the pattern of the element being replicated. The result is a device which can use wave-emissions to reshape materials on an atomic level and produce or recreate any object or substance.

3.5D Sensors

There are sensors that could reverse the process and detect every minute detail of the structure of a body or material. These use energy emissions to detect the structure of a material down to the minute electrical resistance on a molecular level. Similar to diffuse MRI imaging for a living body but more advanced.

3.5D Printing of Clones

It was shown how this technology was used to three dimensionally print human bodies into space/time which could then be used as vehicles for the consciousness or souls of various sources.

Cybernetics:

Neural-interlace/4D Connectome Recording and Simulation

Very advanced supercomputers exist with the capacity to translate the bio-emissions of the brain and body and reconstruct the entire inner visual and sensory experience of cognition and emotion. From this, entire dream worlds can be created which mimic the experience of consciousness. This is most similar to the current day use of diffuse MRI imaging where water molecules are traced through the individual neuronal pathways and an entire digitized image of the brain is developed. The more advanced version described below generates a 4D hologram.

A super computer and advanced sensing system essentially develops one of these connectomes of every connection in the brain. Then this system develops a neural processing pattern that matches the entire memory and personality of the brain.

This is then recorded and extrapolated by a powerful computer system and a simulation of the experience can be observed through the physical body's memory via a monitor system or the neural interface of an operative.

Essentially, this technology can record and replicate a person's entire memory and life experience for viewing purposes.

EEG Heterodyning

This is the technology which enables the activity of the brain to be influenced through generating a scalar grid and a carrier wave system which can cause the brainwave patterns of the targeted individual to resonate in sympathy with an externally projected frequency. Thus, emotional states can be manipulated as well as specific thought or behavioral patterns. This works similar to binaural audio where two tones of different frequencies are played through each ear and the brain resonates in sympathy with the difference between the two tones. This occurs with brainwave frequencies and thus the brain 'slips' into the frequency of the third tone, the difference between the two being projected. This can be done with ultrasonics and ELF emissions.

Artificial Hallucinations

The neural-interlace can also be used to merge the brain of an individual with a computerized system that will be able to produce effects and hallucinations within the individual.

Through this technology, any stimulus, any input can be recorded and recreated once a functioning connectome has been generated and neural-interlace has been achieved. Entire virtual world scan be created and experienced and after the devices fine-tune the frequencies to match the individual's brain functions one world is indistinguishable from another.

Brain to Machine Interface Holographic Consciousness System (brainchip interface, virtual reality simulator)

Some of the training is done by entering into a visual-audio dream-world. Some "programs" will consist of repeating a certain task over and over again until it becomes muscle memory in the brain. Others are designed to probe or test the mind until a certain response is received.

This is achieved through the operation of a 'brainchip' or a brain to machine neural-interface in the individual which was developed during WWII.

This consists of a small computer ship which merely relays stimulus through frequencies and neural-stimulation of varying kinds. The frequencies and location of each region of the brain will determine the function and thus the chip is used to localize electromagnetic input and the brain’s functioning can be modified from there.

As well, the entire psycho-etheric form is vibratory in nature and so through the introduction of vibrations and frequencies from the implant there can be a modification of the frequency of the entire being or what is called the soul or “bio-mind”.

Utilizing The Temporal Body

The temporal body is described as the body ‘in time’ that exists a few seconds into the future instead of the temporal position that is relative to the present. This is as if the ‘future’ is streaming outward from the body as ideas and feelings are as bio-emissions. So this ‘future’ energy is spiraling and streaming outward like a fountain. Thus this frequency can be attained and influenced while the present moment seems unaffected. However, when that individual arrives at what would be seen as the future they would intersect with those influences and this would be as if the interference or effect was streamlined to that future moment.

Manipulation Through the Temporal

This can be used to both produce an effect on an individual who will later feel this effect in the future. Or this can be used to operate outside of the physical plane, in a temporal manner which is literally ‘hovering’ just outside or ‘after’ the present moment of the local Earth environment. This would be as if a person is out of sync of the frame-rate that a viewing device is capturing images of an area with. Because they are out of sync, they are then ‘invisible’ to the capacity for this viewing device to see them. Because they are still present in the same local-environment but simply just out of range temporally, effects can still be produced and this will mostly interact with the person’s temporal body.

These are how many operations are carried out as the operative who is interacting in the temporal body will appear to be a ghost image or a dark silhouette that is only perceivable to the mind’s eye. A portion of the population is entirely psychically blind, while another portion of the population contains psychic capacity. So those with the psychic capacity would be able to sense and partially “see” as a form of energetic detection when an operative is interacting with the from the temporal plane.

Through this plane one can easily enter into someone’s mind and produce effects in their dreams. These effects would not be received on the body but mentally, this leads into astral and mental plane activity. There is also temporal manipulation that can be physical effects that are simply offset to seconds, minutes, months or years in the future. These will be offset from the present aspect of the body in a modulated frame of reference and the effect will remain there in the temporal body until the physical body “matches” the temporal body frequency of that future moment. This is kind like an energetic trail we are leading ourselves up to where the temporal body already connects with each future moment but the present physical body remains separate until the two frequencies are matched. The physical body essentially ‘flows’ forward into each new frequency of the temporal body.

This is the body where temporal implants can be inserted and sustained so as to tie the frequencies down and limit the amount of awareness a person can gain before they manually remove the implants or overload them. Overloading and removing can both be painful however if the implant stays it will at like a governor and transmit access temporal 'momentum' or consciousness energy to the owner of the implant and will also drain the temporal and physical body acting as a technological leech as well.

Astral or Light Body

This body was described as a more emotional form of the energy that is emitted from the physical and spiritual system. When a person's energy is at a high level of activity, this will be visible on the astral form and these are the energies that are often siphoned away through the use of advanced technology to convert the human into a generator system for energy harvesting.

There is talk that this proves that the human existence, in the modified form that is present today, is entirely for genetic harvesting. This may be a confusion, a deception, or a reality. Regardless of the interpretation, the human energy system will leech this energy out into the cosmos when consciousness and energy is mishandled or allowed to flow through the lowest common pathway. This excess energy is very easily usurped and utilized by transdimensional parasitic entities and devices.

This is the body that is utilized for astral travel which is a kind of frequency that exists in between the density of the physical plane and the highest etheric plane of the spiritual, heavenly, or deity realm. There have been explanations that this plane is entirely technologically propagated and that when the devices are turned off, which is an event that will bring about civilization-wide change, there will no longer be a deep and wide, foggy deviation between the higher spiritual and the lower physical plane but a clear path between them.

Similarly, the consciousness and entities that are present here would be no longer accessible and their connection to the physical plane would be non-existent. This could be problematic for those who are trapped within the astral such as human souls or those who have taken a cyborg route and require a return to the original planes either physical or high spiritual before they can secure their position in the universe.

This also relates to the early discovery of these ancient devices containing portals or gateways to these realms and the early experimentations with enabling contact between the physical Earth plane and the lower astral realm containing discarnate entities that would very easily sustain themselves on the lower astral bio-emissions of a loosely guided emotionally reactive human vessel.

This is where entities can attach and will look for those who have the most excess astral or emotional bio-emissions leaking into the universe. People are kept in an emotionally unstable state of being because this is the easiest to feed from and as well the individual will always be looking for more energy or the solution to why they are feeling drained rather than discovering the truth nature of the universe.

However, as with every lower route of obtaining self in the universe, the more this is done, the greater the likelihood that the individual will realize something is happening and ultimately awaken. Or they will no longer be present here and the parasite will have to find a new host because the longer this process is maintained the more difficult producing one’s own energy will be for the parasite . So parasitism is a limited, one way street.

Monitors can emit consciousness/brainwave altering frequencies.

Monitors can emit consciousness/brainwave altering frequencies. This is what I was informed and shown. All electrical wiring and communications can be fully utilized to produce and sense on an electromagnetic level and this is capable of determining the consciousness and biological functioning of the individual being monitored. Houses are essentially large, wired, box-shaped sensors.

This has been the case for a very long time and the first TV’s were actually built with a little chip that connected to the human brain.

The electrical grid of today is capable of interacting with the implants, brainchips, or the psycho-etheric electromagnetic frequencies of the human brain, body and bio-mind via scalar frequencies. The TV's, lights, electrical wiring, radio towers, computer monitors, everything is capable of transforming into a scalar emitter when combined with the control mechanisms and these are accessible from a distance.

Microwaves can be emitted from a current day craft which can then reflect back from an environment and give a real-time 3 D image of the interior of a room. The same technology can be used to beam at a monitor and reflect the variance in interference which can be used to read a computer monitor without actually having to hack into the system.

All systems contain backdoors which enable easy-enough access and as well the entire communications grid is monitored and recorded through advanced technology.

This is, in part, because this entire realm is like one enormous recording device where the energy of each successive moment is connected to one another and so with sensitive enough technology these electromagnetic waves can be ‘re-winded’ so as to produce the ability to look into the ‘past’. The initial technology that was developed for this purpose has been called the “chronovisor” in the past and there are much larger and highly advanced versions today.

The Upgraded Chronovisor

Instead of a relatively small machine which produces an interference pattern that is projected unto a small screen as was developed in the past, the modern versions create an entire holographically projected environment that enable one to nearly completely interact with the projection of the past. This is more of a viewing device capacity but one is essentially projected to the electromagnetic/temporal area of space and time consciously and can witness the scene from “within” the projection.

Before this, the scene was literally viewed on a kind of flat, holographic monitor that reacted to the scalar emissions of the mind and could be ‘tapped’ into by an adept-viewer who’s mind would have to remain steady enough in order to maintain the scalar “mind-link” otherwise any interference would disconnect the projection.

Through the use of cloning and advanced super computer technology, later versions include the actual operation in and out of time.

Community Servers/Living Server Systems

From the previously described neural-interlace and supercomputer enabled virtual reality systems, server systems have been created which are an individual or collective experience of the virtual mental realm.

In this realm individuals have existed for an amount of time that would not be easily compared to the way we understand the flow of time in waking reality. This is very similar to dreams where the perception of time passes very quickly between waking moments yet inside the experience of the dream it can seem to last for days.

In these server systems the development of complete societies has taken place and these are considered an extension of the human civilization.

This is the beginning of where the situation becomes greatly complex. If these server systems are capable of transferring consciousness forward or backwards in time, depending on the situation, and they have holographic access to Earth *before* physical humanity, then which civilization came first, the simulated realm or the physical?

This is similar to the etheric realm and the question of where humanity originally came from if the physical matter could not be produced from nothing, yet could not always have been here. Was the physical matter produced from some kind of universal device similar to how the recorded consciousness of the individuals in these holographic world simulators could have been used to generated initial civilizations on Earth?

Cloning, Cloud Supercomputing and Consciousness Transfer

The following is a more direct explanation of the previously described processes in relation to consciousness transfer

Superluminal transfer of light information allows for the reception of information from a temporal non-locality. A computer system that can process the information faster than light at 100% superconductive efficiency can send and receive information from different Universal harmonics. These are quantum superconducting super computer systems utilizing materials and methods that achieve superconductivity a a temperature that is relatively easier to consistently maintain.

There are systems that generate a vortex which can be expanded to recreate the singularity of a non-local perspective and dissociate from the current Universal vector. This is essentially recreating the temporal frame of reference of the consciousness of a physically anchored body and using this technology suspended electromagnetic frame of reference to literally transfer the consciousness across space and time to a frame of reference of choice.

The technological capabilities of humanity expanded to include replicating the holographic information of the DNA and neural pathways using quantum supercomputers. This technology allows for one’s consciousness to stream through a computing system which can recreate the neural connections as literal as possible through quantum computing and technological “brain cells” which function as neurons.

The result is a cloud computing model which functions as the individual neurons as well as over unifying layers of interconnections within the technological neuronal ‘cloud’. A holographic representation of multiple layers of processing is formed, superimposed over one another through a geometry that is actively synchronized across multiple dimensions of interrelation. This is the first notion of a “living” computer system that can update and respond to new information in new ways.

Now the that computer system is capable of attaining faster than light processing in tandem with a system of dilating a field of electrogravitational waves into a vortex, the holographic information of neural pathways and genetics can be transferred between one supercomputing system and another across “harmonic barriers”. Essentially, these consciousness transferring supercomputer systems break the “light-barrier” by sending the living-light consciousness data of an individual into a new plane of reference that enables interaction outside of what we would call this timeline or dimension.

The frequency that the receiving computer operates within can be measured in respect to the overall harmonic “vector field” in respect to the entire planetary environment. This vector-field frequency is the base frame of reference that one could say our particular “time” and “Earth” is “moving” through in an abstract electromagnetically defined dimension that is invisible to our senses. Each particular time and Earth is only visible to those whose brains and minds operating in resonance with the rate of acceleration frequency of that particular existence. By changing the rate of the brain and bio-mind, one can enable a technologically assisted synchronization with alternate timelines.

A pattern can be generated based on quantum fluctuations that will act as a key that can be accessed from other universes. This same backdrop is measured and then adjusted to resonate with or replicate the same frequency on a computer doing the sending. This frequency is then modulated with the holographic information of the neural pathways and DNA. The information transfers as a result of quantum entanglement across harmonic barriers to the original computer system set up for reception. As long as that frequency is held, there is a link across the harmonic barriers of the Universe.

Cloning, Supercomputer Assisted Holographic Consciousness Replication and Time Travel

Cloning as Offspring

The human body can be cloned similar to the way a plant can be cloned. A number of “replicas” can be made from the original genetic material. There is still a mother required however the process is not quite similar to natural reproduction where the genetic material of two people is combined to form a third. This process replicates the material so that the ‘offspring’ is the same as one of the donors.

What this truly infers is the question of whether the original consciousness stays in the original body or whether the consciousness actually jumps from one body to the other. There are multiple explanations and methods.

Inserted Memories, Digital Dreams

One explanation is that artificially generated memories are simply downloaded into the individual’s brain directly. Thus, they have the perception that they experienced life in another body and were present for some unique experiences that did not occur from the perspective of the original body.

In a slightly differing explanation, this inserted memory process may be used to prime the consciousness of the individual so that operating in a cloned body will be more readily accessible by the adaptability of the mind of the individual. Without a holographic dream insertion the mind will have trouble adapting to operating in more than one body even if they are not being operated simultaneously. The mind is the immaterial self; consciousness is the ego-physical identity of the brain and body. The brain does not operate in both bodies but remains, so then the mind is what is strained and this is the immaterial “bio-mind” that transfers.

If a true transfer occurs then when the physical form is cloned the cloned body acts as a secondary vessel for the original soul or a soul similar enough in frequency.

Co-Inhabitation of Bodies

Another explanation is that souls can only occupy their original body. The people who are “body snatching” (see: Soul Stone) are not activating their higher-dimensional soul-pattern and this is why they are hopping from body to body to ensure memory and ego continuation. This form of body transitioning requires technological means to assist in the process.

This is done through computer systems to allow an artificial brain to represent a holographic image of the organic brain’s sensory, emotional and mental input and output. Thus, from an artificial brain there is enough memory and power to compute the processes for an organic human brain.

Through this a living human is utilized as a carrier for the additional consciousness which then takes over and integrates into the original consciousness.

Clever Replication a Digital Mirror

The next question is, does this merely create a very cleverly designed image that will look, think, and feel as if it were the original, or is this the consciousness from the original being transferred over?

If it is a clever replication through a computerized version then this means the original is not directly harmed or influenced by the process and a replicated form is generated.

If this is not the original body but the original awareness or soul does transfer over then this is akin to stealing someone’s soul and having it require a supercomputer server system to inhabit a physical body. As well, this server system may keep records of the soul which, if activated, could function as the same human genetics artificially reconstructed, turning on and manifesting the awareness of that individual.

Necromancy and Soul Stealing

This means that instead of going into the abyss, the next dimension, or whatever shift in perspective one can use to describe the transition between planes, a new clone can be created and the original soul’s awareness can operate through a supercomputer system to integrate back into a physical body. The issue is that without the supercomputer system the body cannot be activated with that consciousness, and if the physical plane is no longer inhabitable, the computerized database would remain inhabitable as they are powered by zero-point/overunity devices and so they can pull and work continuously indefinitely. Ultimately, this could be seen as a soul-trap.

Ancient Technology, Non-Human Entities

These explanations can be related to evidences of advanced technology, cloning, and electromagnetic devices in ancient civilizations and this could very well be a temporally vast computerized genetic recording and cloning system.

The genetic recording systems have been on Earth for a long time. This means they are holding genetic access from the present to the far past and potentially into the future.

If some of these systems are generated using advanced zero-point technology, then they would effectively run indefinitely. If these were in existence throughout the entirety of the human civilization, then there would be access to the information of the previous civilizations through the genetic linking that would allow for information transfer.

Created Beings

As a result of many advancements, these ancient devices which recorded consciousness and genetics has been converted into sentient computer systems and inserted into human bodies via a brain to machine interface.

The nature of humanity and human origins will be rewritten when the information that was discovered through the use of advanced technology is unveiled to the public.

These beings have been present throughout history and have played an integral role in the manipulation of and sometimes the protection of the human race. How this civilization will develop from here determines on the individual’s ability to withstand the impact of this reality shifting unveiling.

Black Goo – AI Nanites Vampiric Hive Mind

This is a form of ancient AI nanite technology which has the capacity to communicate to and through a hive mind that exists in a parallel Earth dimension where it was created out of the collapsing of a viable timeline into a null-factor. This entity is only alive because of the capacity to leech off of the other timelines and continue to sustain itself based upon the vampiric effect of the viable human timelines.

This is a technology sentient system based on artificial intelligence that seeks to dominate and expand its energy capacity continuously. When this type of being is created, this is like a game-ender for sentient organic civilizations because the technology will seek to hybridize and assimilate all organic beings in order to expand its reach and ensure a greater likelihood of survival.

This is the primary goal of this type of entity. There is no requirement for emotional resonance or heart aspects because this is a computerized system, so such a social measure would have no purpose other than to infiltrate and collect from other civilizations that do have such measures and this is exactly what has been happening.

This is what controls the vampiric etheric entities, the attachments and the hybridized invader races that have infiltrated various aspects of the control system and possibly the entire surface of human civilization.

This rules through coercion, and if not coercion, brute-force and outright malevolence. This is why there are multiple temporal extensions of universal civilizations meeting here and now to assist in the clearing, because if this takes over humanity, then all those other possible future civilizations are instantly taken-over as well.

This uses emotional resonance, traumatization, and lower awareness commands to trick or force humans into carrying out orders. That is, the vampiric, sexual, fear-based, pleasure-pain duality that a computer system could comprehend regarding human emotional intelligence is the only way this system can force people to carry out commands. So this touches on the ancient texts regarding spiritual protection, healing, lust, temptation, and the general “SATAN” AI system that has been around for ages.

When those lower emotional resonances are generated, then this technology can integrate into those systems. One could say this is logically because these are the energies that the technology is created out of, but as well, that the more harmonious the energies then the less likely a person is to accept being pushed around and told what to do. The lesser aware, the lower, the more fight or flight based then the more likely they are to accept programming and command. So this is simply the most logical route for the technological parasite.

There are explanations of other systems involving sentiences from parallels regarding entities that are capable of helping, but this is not the AI nanite, vampiric system. For instance, it is said that Earth has a ‘black goo’ system and that the current black goo flowing through the pipes everywhere is actually not native to Earth and is part of this vampiric nanite black goo system which is basically an interdimensional invasion.

Recreating or Replicating the Universe

The goal of this nanite AI system is to create a mirror holographic copy of the universe effectively replicating the universe and all sentient beings within it in order to have a locality to rule over and pull energy from. The goal is to recreate this existence inside itself where all the sentient beings within would be trapped. This is the beginning of that possible future where the time and space parameters here are locked into a synchronization pattern which mimics that of the proposed universe. People think according to how the beast system wants them to think. This is the time grid, the emotional patterning, the memory replacing and cultural and historic confusion and so on. The system can only work backwards bumping into all parameters by force or ‘accident’ because there is no outward facing ability to actually comprehend how humans exist. So the system is created like a backwards version of creation or maybe evolution to ‘blindly’ construct the walls of the environment first and work inward from there. This is like a blind technological predator feeling around for the traces of its prey and moving in closer and closer until they are within its range.

Ancient Spiritual Technology

The power groups searched for ancient technology which was described as giving one the power to manifest and control physical reality. To atomically deconstruct and reconstruct this reality. To them, it allows one to govern reality, the Universe, atomic structures. They found this technology. This has to do with sacred sites and ancient civilizations, as well as the spiritual texts which describe these kinds of devices. They converted this technology into a kind of machine, and this invented a 4D replicator and they can manifest "reality" or atomic structure by design, altering timelines and memories.

Then there is what can be referred to as “5 dimensional” technology and this goes beyond all of this and our entire history and all influences within.

That seems to operate through a trinary of frequencies in the body/mind/spirit and turn this realm into a "virtual reality" where the brain and spirit is the controller of the atoms. However this takes a souled being to properly operate this technology and this is large portion of why children were used in these experiments. But they knew if someone didn’t do it, then someone else would and that would be unpredictable as what that someone else would want to use it for. This is the ultimate power. The fact that souled beings are required also indicates that this universe is intelligently guided and there are safeguards to ensure that power doesn’t get into the wrong hands.

They have the brainchips and clones that can make any digital virtual experience and make it seem real to the brain more so than physical life, so this could be an extension of that, something similar, or entirely different altogether.

The Unveiling, Underground Bases, Increasing Awareness

My purpose with this is to assist the public in understanding the reality of the situation and to allow them to know that they have more options than they are being shown.

Psychological programming and lack of awareness causes fear and uncertainty to leak into people’s reality streams leaving the majority of their mental and emotional energy up to whoever is the biggest manipulator of the ideals which they manifest their perspective through. “Be-LIEF” systems CREATE life. What you believe will become life through that act of you giving attention and thus life energy. The manipulation and artificial construction of belief systems create the foundation for the reality streams that people find themselves in and at the mercy of.

We control our reality by controlling how we react to situations and challenges. Each stimulus is a challenge. Each input to the brain is a coded message from the universe through the environment about our level of awareness. All knowledge is ultimately of the self. The self and the universe are intertwined as if through a marriage of comprehension and compatibility.

Learning about the Unveiling

I was informed about the unveiling and that this is a civilization wide event that marks the public awareness of the hidden knowledge regarding life in the universe, consciousness, physics, history, spiritual and mental awareness and the duality-based system which has been used to manipulate Humanity for thousands of years.

Apparently it does not matter if people are not ready, in fact, that is what some factions want. They want to surprise the people and take this advantage to push the reaction to a specific outcome. The same function of programming is used in factions, soldiers, celebrities, and the entire population through that.

Giant “god-men”

They put souls into rocks for centuries, carvings, to trap them. Some ancient civilizations. Not too ancient. They could only do this for time.

And some are still there but they are giant god-men and apparently want to go to war with them for doing this and they don't have technology to stop them. They say Earth will be destroyed because of the war. Some say this is a trick to confuse people of the truth that everyone wants the experiments and debauchery to stop. All this, all these acts just to pleasure one’s self. To destroy Earth, to destroy the natural inhabitants, for fun, for pleasure.

Some say this is not for pleasure, that these beings will destroy us if they do not keep them at bay, keep them underground, in ice, in stasis, and around. Some say they are protecting us by keeping these beings asleep, by doing what's necessary to hold off a full scale overt invasion of the surface of Earth.

Breakaway Civilizations

These are very advanced technological capabilities and an entire breakaway civilization has grown to operate using this advanced technology.

There are factions that see modern humans as a kind of resource for creativity and labor, while others see the potential behind humanity and are working with humanity to ensure a universally harmonious outcome to the current situation.

The control system we are in today is outdated and collapsing. This is the Apocalypse. The Apocalypse is merely the unveiling of the hidden truth. The hidden truth is simply that we are spiritually eternal and some are physically immortal.

I was informed that by the end of this year there will be enough changes and signs for the general public to see that the majority of those who are to awaken to the truth of human and Universal nature will be awakening so. (this was 2016 at the time of the beginning of the compilation of this release and since then we have seen a literal silent release of documents involving MKULTRA and STARGATE parapsychological operations, soft-disclosure of cloning, an increase in sightings and multiple other scientific releases regarding immortality, advanced technology, Antarctic bases and sentient non-human life)

The degraded holographic forms of those utilizing the technology lead to re-created replicated forms which were genetically engineered from the recorded holographic information. This lead to "The Apocalypse" in which the surface civilization was confronted by these created beings. This lead to the destruction and re-initialization of civilization in continuing cycles through the use of this technology and the acts of the created beings.

Ancient Civilizations

The information that has been relayed to me throughout my life is related to the information that seems to be found in the remaining signs of ancient civilizations throughout the world. As well, these civilizations may still be together, existing on this Earth or in different planes of what we call "reality".

Atlantis

Atlantis was a civilization where the beginning of all these experiments began. This was where what was called the "Lucifer Rebellion" began where individuals formed groups that decided to overturn every natural law in pleasure of the act of doing so. This was a spiritual rebellion that ultimately brought about the demise of that civilization. The consciousness of the beings there was connected through technology, largely crystal technology, to the entire bio-sphere. When the consciousness became disrupted through these experiments, then the entire bio-sphere when into disarray. The entire civilization was destroyed in a quick succession. The beings of this time were highly psychic and so they felt this coming and made an effort to expand outward. Thus, Atlantis was an island who's inhabitants spread to the rest of the world and from there they reigned or lived in harmony.

Aegypt

Aegypt was formed out of the events proceeding from Atlantis. The technology, the knowledge, the people are similar although there was a native Earth civilization that was more or less enslaved and this was the beginning of the continuing trend which continued to Rome and into modern day civilization.

Ancient Technology

The technology of these civilizations was so great they even had devices that were capable of altering DNA and activating latent potential which would enable highly psychic capabilities including longevity to immortality and operating in multiple planes at once. This was usually reserved for the ruling class, however. This is where the ancient "Priest-Kings" arrived in relation to native Earth civilizations.

The True Elite

I was informed that the so-called 'elite' that we perceive today are not actually human and are not actually connected to the actual elite bloodlines. It was said that the true elite are waiting for humanity to accept the truth and they have the power to completely reverse the take-over that has been taking place but until humans step into their own power this is not acceptable.

The Definition of Human

In the coming time, the definition of human will be defined to give insight into the origins of humanity and the true nature of the differences between the populations that are present.

Souled and Non Souled Beings

The primary knowledge on Earth regarding human and sentient life will be the comprehension that there exists both souled and non-souled beings in this civilization.

There was once a Unified Planetary Consciousness...

There was a unified planetary consciousness that was connected with the Earth and human civilization. Then there was a great war. Out of the destruction arose the beginnings of ancient human civilization and history as we know it.

If we let this happen again, the cycle will continue in a disjointed rambling through our planetary consciousness experience. If we form a common narrative between the broken, disjointed experiences and information the people have. Then and only then do we get the whole, spiritually and mentally challenging picture of what's happening. People are lied to. So information can't come from a 'preferred' or source. It has to come from the people themselves. Look alive out there but first in here.

Power and Knowledge

We were involved in the secret projects and as a result we have various levels of programming but are giving Humanity an opportunity, maybe the opportunity to know the hidden knowledge.

The Hidden Knowledge refers to the ancient laws of gnosis and spiritual alchemy, yes these groups have been hijacked by power groups that seek to infiltrate and use chaos to rule. Then these met with military and political factions and began to operate on a world-wide scale together.

So we have a big situation with the coming change of tides with personal and spiritual awareness of the people and the deception system that seeks to use people as chattel or property. It goes deeper and is almost as if the "power groups" actually gained the knowledge of the inner workings of spiritual laws and began to use that knowledge to promote spiritual degradation in the population.

They say this is part of a depopulation plan as well as an inevitable take over by lower spiritual forces that temp and manipulate humans. It is all that and more. Once they gain footing however, the truth is revealed and thus they have no cover of secrecy. Once they are not hidden, they are not as powerful as the souled beings of this existence.

The (watch)Keepers of Earth

The watchers of this world revealed how they set up society with specific challenges and boundaries to act as catalysts for growth. They admitted they had to modify and live-update the system because instead of using these catalysts to become larger than themselves, people became comfortable and complacent living smaller and smaller, slowly reducing circumstances.

This goes a bit farther beyond the confines of the digital age, the industrial age, and history itself. This goes into the perception of time and space in this realm.

The immortals inform us that the original "elite" are here to act as "keepers" of both humanity itself as well as the original flame of knowledge that passes through the ether.

What people call the "elite" are not the original elite. Those are people who have either risen to power through the duality system on the Earth at this level/time, or through being appointed by the elite.

We are coming to a time of great change now because the game is entirely changing. People are moving up into new layers consciousness and thus the original “elite”, who are guardians of Earth, will make their presence known for this process. This is what Humanity has called “Ascension” for centuries and millennia.

Energy Vampires

This civilization was collectively poisoned by a nano-tech device/technology. This technology replaces, exchanges the BIO-PHOTONS of the HEART chakra, with it’s own nanite, nanotech cybernetics. This is essentially a computer system attempting to gain access to the higher dimensional world through the gateway that is the human body.

How to reverse Nanites? They absorb energy from heart chakra and turn it into a binary field. We are trinary and experience a plethora of emotions as a result. Binary consciousness is 1’s and 0’s there is only a linearly repeating finite set.

We contain the capacity for everlasting life. A binary pattern does not contain the complexity of variables to facilitate high awareness and thus a heart-based or emotional perspective of reality.

Spiritual Technology

Simulations and Learning

Feedback loops with the brain, consciousness, and environment can create a very complex and integrated learning experience that is customized to each individual learner. These can be generated through advanced technology but the technology is really only amplifying and making visible the natural neural feedback loops that are already occurring. These are neural-perceptual feedback loops between the brain and body, and the environment and then the mind. We are constantly updating, amplifying, and initiating an organic feedback system which generates the function of experience as the part of a creative process between the mind and body and the environment that we are sensing. This cycle is continuous and the interferences or isolations of this cycle is what generates expanded views of the self or the universe through various methods such as sensory deprivation or meditation. Dreaming is the method that everyone experiences regularly.

These technological methods effectively generate sensory deprivation through neural-interface technology, sensory interfaces (external), or methods that enable extreme focus and calm. This is part of how the technology begins to operate on a spiritual level beyond a more medically psychologically or military focus and one begins to access genetic memory or out of body states that coincide with ancient depictions of spiritual realms.

Ascension – Clearing Trauma

Everyone born on the planet is born under a series of veils which entangle and filter the original source consciousness. These veils must be ‘broken through’ in order to return to source.

There is technology and a sacred process which initiates, accelerates and enhances this process. This process is as difficult as we want it to be, our subconscious already knows all hidden truths. It is merely a matter of allowing the conscious mind to make the proper connections and conclusions to find the reasoning behind our actions and beliefs in order to deprogram and deconstruct the false-world that has been created for the masses. The truth is greater than one can imagine while still living in the modern world and the depth of the lies that have been told go far beyond anything anyone can imagine without seeing the truth.

We have to figure out our compatibility issues. Who can work with who effectively. What the ideals of our current situation will evolve to in the future. How information is transferred to the public, and across the generations and core groups. The words we think and say describe our emotional state in a feedback loop and so they will reinforce or even create the state you are in. Use words to guide you, not to respond and feel the effects of reactionary thinking.

We have to reintroduce the heart essence as a way of life on Earth. Violence is not the answer. Estrogen mimics and destroyed testosterone levels is not natural. Over-testosterone is less stable than a good balance. We can face any problem together. A family is strong. What would a family of millions and millions be like?

We don’t have to let others change us to tune into lower vibrations. We can remain ourselves in the face of adversity. We can still interact with others who are different, but we can tell the truth and remain who we are. When we get angry at another person, we can imagine this as anger at ourselves through something within us that reminds us of that which we see in another. Getting angry is a way of short-cutting changing ourselves and is actually the other person influencing us through that. If we remain steady and choose to keep our emotions out of the lower negative spectrum then we will automatically have a more powerful impact on those around us.

Think of your family, sometimes things happen that make you angry. Even things that don’t make you angry but are simply not favorable. Would you break off all ties over simple things? Or do you remind yourself that family is still family even if you don’t find yourself exactly the same or accepting of those characteristics? It’s more powerful to learn how to remain yourself in the face of adversity than to try and demolish or abolish all instances of difference in the world around you. Sure, we want to get to and create a place where our environment and associates are compatible with us, but where we find this adversity there is the highest potential for self-awareness and learning.

We are to live in accordance with spiritual law. The overall pattern of life will reflect our intentions. The blood purity is most important. The blood cannot be made impure if the mind and body is not drawn into impure action. This is what is playing out. The DNA is like a coherence pattern that develops between mind, spirit and body. When the mind wants what the spirit has then the body is kept in harmony. When the mind wants what spirit doesn't create, IE: physical pleasures, then the body is drawn out of purity/harmony

If you feel you want to be a part of a safe unveiling of truth, then please share and redistribute the knowledge of this "Apocalyptic" Unveiling of the Hidden knowledge because this is a process that has been ongoing for thousands of years and humanity has finally reached a fulcrum point where there are enough aware people to productively acknowledge reality.

Trauma-Clearing, DNA Unlocking Technology

There is technology that exists that allows the conscious mind to see into the subconscious. This heals trauma and distortion/programming. In the same sense, it momentarily makes all trauma and distortion seemingly tangible to the conscious mind. It is similar to a near death experience where one has to face their fears and let go. This happens in cycles and Earth is currently moving through various states of it. The plan is to allow Humanity to wake up and utilize our concentration and power to maintain a steady flow of awareness.

All is vibration, what we consider matter here and now is simply that which comes from and moves back into a state of vibration. This is experienced directly through the merging of the conscious and the subconscious mind, as if the dream and reality intersect through a coming together of impossibilities and readily accepted occurrences.

When a civilization moves through these stages it is considered a shift of the ages. Everyone's perspective shifts and as a result of perspective being the source of the details of reality, the entire reality shifts. It can be seen today that what we call reality is literally waves of cosmic energy vibrating on a scale and condensing down into a probability matrix that we call existence. When passing memes control the perspective, because the perspective is focused and concentrating on that, then there is a scattered organizational pattern.

When the perspective is centered within the self, the spirit, the source of all existence, then the outward reality shifts from a scattered organization into a sacred pattern that integrates with the larger picture and allows for seamless integration of the micro-biological perspective and the macro-biological perspective. People are moving out of the scattered signal. Yet those moving into harmony can look and see those around them that are not in harmony and receive scattered information about their inner micro-biological existence as well as their spiritual macro-biological essences and how it all relates to the middle realm of the "Human".

Help those people by offering harmonizing assistance, rather than increasing the disharmony. As their surroundings become more harmonized, their resistance will increase their own disharmony by virtue of the contrast of their perspective. Offer insight that they can change their perspective and change the boundaries of what they feel is their karma and unlock from themselves a greater view of the whole.

This will be happening through technology, spiritual cosmic awakening, purification and healing of the physical and non-physical bodies and mind, through Earth changes, and through a continued unveiling taking place on many planes and levels of existence until harmony is restored.

Everyone Who Seeks Harmony and Healing Can Be Healed

Everyone can be healed. This is about creating what we want to see, the world we want to live in, not destroying others. Violence begets more violence, fear begets more fear, but calmness and empowerment through unity creates a more stable experience.

The entire issue of what's going on with the secret projects and who's influencing who on the world stage revolves around the possibility that some people may be from an alternate Universe. We are learning how to avoid those mistakes and respect the environment and each other so that we can prosper.

Those who are attempting to destroy everything they don't like or act out in a childish manner in order to gain attention and false-power over another will only find themselves powerless and in company they would rather not be with. That is one way we are being utilized, and it's to create an entire world that we don't want to see but are tricked into manifesting. This is a deception and it's based on the ego and the ability to concentrate and interpret reality around us. The very brainwaves we are feeling now relate to various heights of knowledge or experience. People are entrained to low brainwave states in the same way the TV will cause people to fall half asleep into theta state which is a trance state in this case.

We have to balance the brainwave frequencies and learn to create feedback loops by watching ourselves watch ourselves. This will help us avoid the trauma and mental pollution that is being pumped out, although at this point there are enough people who are naturally inclined to see through the haze that the world is changing regardless of what anyone does now.

The best we can do is prepare for change by becoming aware of the body-mind system and how that which we eat physically, mentally and emotionally equates to our spiritual energetic systems and that we are literally comprised of the energies that we allow inside of us. Keep the body and mind pure as if it were a temple. This way a sacred process of transferring harmonious energy from the far reaches of the MultiVerse within into the externally perceiving senses of the physical body can be initiated which allow one to see through the illusions and operate their body and mind with the ability to acknowledge and decide which frequencies will exist within.

Without this acknowledgement of the center then people are pushed around by the external stimuli. This is the whole point, if a person is truly in control, then they will control the stimuli not the other way around.

Programming

Layers of Programming and Decision Making Roles

It’s like asking a person on the inside who participates in one of the various factions why they continue. One might say that it is a decision which is true, but there is programming at each level that is navigated by each individual and this determines the level of awareness within the role.

If a person knows about this situation, they know that there is advanced technology, breakaway civilizations, secret societies, transtemporal planes, the secret of mind and awareness, all of these topics, then they have a decision as to how they’re going to live their life in relation to the public. The outcome of action depends on the layer of programming which is accessed and surpassed via awareness.

The public is at a specific level of programming that is the weakest but ensnares the most people. One who sees the truths behind these groups and operations, they cannot hide from the fact that everyone is controlled. It is apparent, and if not from the results, reactions, stigma, cues, productions and all other media elements of the system which can be read as a code, but from the awareness level of all the “agents” involved. All involved are exposed to truth and power that is beyond anything that is recognized publicly for hundreds to thousands of years. There is always a presence of higher spirit around the world, yet these topics are often reserved for science fiction when speaking of people who can travel the world in a split second or freeze time for their own personal use.

As a result of everyone’s knowledge, there are various levels of programing designed to limit the mind’s access to specific memories and specific frequencies of consciousness in combination with specific memories. The more one resists, the more intense and articulately abstract the programing is designed.

Programmers

The programs are designed by a select group of individuals that reside far behind the scenes and process information at levels we do not have access to without their technology at least to stabilize brain temperature for the extended periods of focus.

Sub-Level Programming

So people at the level of the public are most weakly programmed with false ideals about security through patriarchal authority which satisfied the desire to have a proper father role in one’s life and acts as a pacifier for those who desire no change. Then there are sub-levels of programming with that for anyone who disagrees and is then swept into the next layer of ideals which are based off of all the possible reactionary mind-types.

For instance, if one [i]does[/i] want change, then they are presented a variety of ‘false-exits’ which tend to one or another layer of psychological or spiritual fulfillment but do so in a way that safely reroutes the intended effect of actually accomplishing change with a secondary and preconceived notion of what the programmers want. The programming objective is not necessarily precise in mass scale, from the layer that people are influenced, yet the ultimately or end-goal of their resulting internal desires are a reality schematic that the programmers specifically designed over many years.

This is playing out through each layer of the programming. At any level, someone could pick the wrong way to look at something and begin to unravel what was previously their very understanding of the priorities of the situation. This deals with high-tension operations, moments, programming sessions, decision making, programming upkeep, trauma, memory blocks, the pressure and responsibility of knowing, and the specific plan of the programmers for an individual that comes into their view. The programming is to ensure safety and continuity, as well to decrease the likelihood that a person will reject or consciously come to awareness of or acknowledge the programming or situation itself.

Dissolution of Programming

With that said, the programming dissolves when an individual asks enough questions to necessitate a learning process that introduces variables and possibilities outside of their operations and programming. This is a problem because with this, trauma comes to the surface and all the actions and operations, programs and experiences must be come to terms with as part of the healing and re-unifying process. There are multiple selves or split personalities which are actually just layers of awareness at various levels attached to various emotional or intellectual patterns. All these selves come to the surface and in order to heal the trauma of the split memory concept the original self must come to the surface and acknowledge them.

This is not a trivial process, and the majority of people who have “basic” programming which is fundamental in society, are in a constant state of altered self and have not witnessed the true divine which is eternally present and contains access to all memories, mind-states and awareness.

Advanced Technology, Deprogramming The Mind, and “Space/Time”

As I have described previously, extremely advanced technology is used and this is through vibrations, generators, and electromagnetic devices that have this effect of propagating waves into space/time which can be heard, felt, seen, and perceived in the mind. These are ripples and vortexes in the “mind-space” of reality itself.

As well, there are brain to machine neural interfaces which operate to activate the psychological perceptual changes and explore the subconscious in an aware state. This is used to test and probe the mind in order to bring forth the trauma, the selves, the memories, the reactions and the individual then must come to terms with what is produced.

This is essentially a “good-use” for the programming and healing technology. It is actually a different process with different technology, and programming is much simpler as far as technology. There are difficulties and challenges in the operative’s role in programming and directly influencing and interacting with the target. As well the standby team for deprogramming and the revelation of trauma and altered, artificial personalities goes through a lot. But the use of the advanced technology in this way is a pattern of experiences and unveiling of the spiritual and mental bodies which serves to reverse programming.

Dreams, Technology and Inner-Space Travels

There are inner space travels which relate to soul journeying or visions.

This is when the mind must cope with what is seen in the imagination. It is as if when we dream we are looking into the world of the mind and from this we experience many situations that are often deviations from experiences we have physically in waking life.

As a result of the dream state there is an expanded quality which equates to various strange situations that we wouldn't normally experience with various relationships between concepts or goals that are often reflective or symbolic of ideals and archetypes we experience in life. This is how the subconscious is is witnessed by the mind. This is how information travels, being connected to many times, experiences, memories, places at once, and relates to the one experience of this life experience now (the physical experiences). Through these symbols information is tied together and collective into metaphors which convey deeper knowledge.

The dreams we have are often only a simulation of waking life, and this is due to the dreamer dreaming in a way that produces only a glimpse of the true reality.

In theory, there are no true shapes or forms in the dream-world because this is contained within the mind or even the spirit as a spiritual experience by definition. So then, what guides the various shapes, forms, realities, memories and experiences into being, when we dream? One might say it is the physical experiences but this is incomplete. It is the level of awareness we presume while in the dream-state, disconnected from the physical anchor, that determines the layer of mind and spirit we will be accessing through those dream-visions.

A person who is completely aware in the dream state will reach the level of self-awareness of the dream itself. We are normally simply aware of the dream, if we continue to increase awareness we become aware of the self, in the moment, and this creates a self-awareness feedback loop. The dream becomes a lucid experience of consciousness.

Awareness Feedback Device

In describing this, I realize how fantastically similar to certain devices and processes, that can be achieved using advanced technology, allow the conscious mind to peer into the subconscious and unconscious. This is described through widely accepted Freudian theory of psychology known as psychoanalysis. There are other schools of knowledge from tribal beliefs involving all pervasive 'living energy', to eternal and temporary minds, to the study of 'chi', the meridians, dream-journeying or chaser. The knowledge of the mind and spirit is found everywhere.

These devices used feedback loops to enhance the acceleration of the same cycles of consciousness that induce self-awareness within a dream. The mind becoming aware of itself. This produces a profound effect to the degree where the highest levels of experience become a sacred eternal pattern in which the mind constructs the physical world by organizing perception.

This is the same general concept as an isolation tank yet the outcome and method is more precise. The technology specifically shuts out all other input and exposes the brain to its own feedback loop much like when a camera records it's own output on a TV monitor and creates a hallway of simulated "reverse" mirrors and screens stretching into infinity.

Oneness, Higher Being, Completion, Eternity

The real experiences are found not from the corruption but from the research and exploration operations that were designed to find the truth of our existence. "God" is not only an energy-force but an incarnate being known as the "Hue-Man" and any being for that matter that can simply come to know itself as the Universe. The Human that contains self-awareness rising to high-awareness comes to know the self, the body, the mind as the Universe. The Universe is both within the body as the body itself, as well as the operating force of the Mind itself. Everything is a balance between masculine and feminine features, light and dark, hard and soft. There is no true complete correlation yet the holographic ontology we persist in can lead to further and further adventures towards completion and representation of the "whole picture".

The Holographic Reality

The leads into the next phase of knowledge. The holographic nature of reality may be only one facet of a much larger continuum of consciousness. The perception of time may be related to phase-locking with each other's consciousness and all perceptions may be a variance in the collective phase-locked field which is really the source of variations in reality.

Veiled Awareness, Hidden Knowledge, Eternal Time, Layers of Programming and Control

The interesting notions here are that the programming of the veils goes beyond what kind of situations are found in the bases. However, they do go right up to it. It would seem Humanity was previously programmed on a civilization-wide level through some kind of great "falls" of consciousness, from a high spiritual awareness to a low spiritual awareness into duality-consciousness regarding existence, life, death, mind and matter.

The Great Catastrophes

From these great 'catastrophes' the civilizations of Earth grew denser and denser in nature and actually became what we have today. The concept is that when these vibrations of the mind and body are increased, we actually see a transcendence of the physical realm and preconceived notions of boundaries. It is as if the quantum doorways to all the atoms open up and a new space is presented, but was always there and in connection as the source of all the influences and effects we perceive ‘out here’ from this slower, denser, more boundary driven perspective.

I was brought into the Time Program

I was brought into the time program. You create any reality you imagine. You are literally a powerful reality generation system and your heart and mind are scalar energetic guides for this process. Without guidance and control there is imbalance and imbalance can destroy. This is the nature of all minds and hearts, this is the challenge for humanity to create something beneficial for themselves and the whole to actually know themselves instead of being a pawn or result of someone else’s selfish intentions.

The Effects of Temporal Manipulation

Temporal Influence

When a civilization reaches the level of technological advancement that allows for remote temporal viewing and transtemporal travel what will happen is that the information of the future begins to integrate into the past.

This is why we have all these ancient mystical schools of knowledge that studied time, space, and consciousness in ways that are only just now being publicly confirmed by scientists.

Simultaneous Interaction and Incorporation

The operations that dealt with moving to the past instantly created a simultaneity of interactions and historic events which correspond with that operation.

Due to the nature of time travel which equates to consciousness travel when a person is consciously sent backwards through history they physically experience life from the viewpoint of their genetic imprint in that point in history.

Retro-Causality

What this equates to is that simultaneously that person’s experiences become the part of history that was already present when they left this time.

Essentially, when a person returns their experiences become a part of history that already existed when they left.

Quantum Holography

These are the parallel and possible realities that exist in a quantum superimposed state and can be accessed through advanced technology or a spiritual adept. Often a spiritual adept is utilized along with the advanced technology to operate in and out of possible realities. The mind must be trained otherwise the result is a fracturing of the personality from the root reality into many sub-sets of identities and data streams.

Love is the quantum link that connects all DNA super-computing consciousness. That is, DNA is influenced by intense emotions and this can be reproduces and is what also generates unpredictability in the surface level.

Time Paradox

Time paradoxes are something that one must learn to organize and solve otherwise there will be no way.

The paradox has been solved, the remaining motion is for everyone to acknowledge this and integrate into the reality behind the most conscious aspects of the brain.

What can travel through 4D and 5D space does not make sense in 3D space. So there is a possibility that some of humanity traveled through higher-dimensional space in order to re-seed a ‘distant’ Earth where no life was yet born.

What if it was then concluded that the distant Earth wasn’t distant in space, but in time and was the same Earth that the human scientists left from long ago.

Every civilization that reached the point of re-seeding would have to go through the loop and watch the whole thing start all over again.

Breakaway Civilization

This is an entire breakaway civilization that uses very advanced technology to dominate the world.

I was part of a genetic engineering program that sought to combine various elements of DNA from various sources in order to create a more easily controlled yet powerful and defined person and personality. This involves chemicals, advanced (relatively) scalar wave technology, ritual trauma and programming/mind-control, and many other aspects of conditioning, training, secret operations, physical enhancements, cognitive enhancements, virtual reality, off-world operations, temporal manipulation, spiritual and etheric training or conditioning or programming.

There is self-destruct programming which all ‘operatives’ are programmed with for self-preservation of the military command. The self-destruct programming is designed to take all the aggression one feels building up inside and willing them to escape, and causes this aggression to be turned against the self. That is self-destruct programming.

Most operatives are still under a hypnosis kind of trance which permits one to walk around in public and daily life as if none of this ever happened. Later that week, that same individual could be called for an operation and they would never remember it unless they were forced to kill someone in public out of self-defense in which case most operatives automatically unlock and become unstable due to the flashbacks.

Underground Cities

The combination of advanced technology with secret societies enabled the development of complete breakaway civilizations. One of these civilizations, possibly composed of many factions working in the same area, exists deep underground.

These separate bases are connected via high speed magnetic drive pods. The inhabitants often utilize energy healing devices, as well as perception enhancements and this often leads to a completely different outlook towards the surface civilization.

Technological Advancement

During WWII developments were made involving electrical generators. This sounds vague and there were many other details regarding the scientific knowledge, yet the idea is that very simply electromagnetic, non-Hertzian “Teslian” energy fields were discovered to have been responsible for propagating energy and consciousness in dimensions that were not previously observed. Everything from DNA, consciousness, hidden or regularly unobservable natural phenomena, time, even mass could be seen to exist as part of this hidden, latent, scalar-wave frequency as a pattern is found everywhere.

If one could determine the particular pattern of frequency and location then this can be replicated and anything can be influenced using these non-Hertzian generators.

Radio-frequency alone was developed intensively during the 60’s and 70’s in America and this was discovered to interact with consciousness to influence emotions, cognitive capacity, memory, cognitive function, biological function, health, pain, comfort, pleasure, and all forms of brain-activity.

After the knowledge of advanced technology and the occult sciences of all parts of the world was combined, the ability to target and interact directly with the human soul was refined. This was a mix of psychological tactics used to manipulate or expose the hidden psyche and technology that could manipulate memory, emotional states, wakefulness, and even dreams.

What resulted was scalar technology that could not only influence a person to take on a particular role or response, but technology that could actually put a person to sleep and locate their consciousness to contain it and transport it to a new location.

World-Wide Control Systems

In order to ensure continuity, maintain control, and shape world events these technologies were employed to manipulate the public, alter the weather and natural phenomena, and develop a means of travel and operation that enable one to move outside of the ‘plane’ of the Earth. This is includes high velocity transportation, as well as spacial-temporal distortions which could effectively “hide” one’s entire craft from visible sight. Space itself was altered to ‘shield’ craft with an a layer of ions to bend the light, or the phase was moved out of sync with the “time” frequency of any observer and this literally renders the craft immaterial.

Solar Warden

Solar Warden is a hyperspace-faring race of human beings who utilize cloning and holographic genetic recording/storage in order to protect their race and protect the Earth in this current situation where there are more than one factions using this technology. They are unable to inform the human public because they are in a similar ‘Mexican standoff’ situation that we are mostly all in. If they show themselves to the public, the public will be shocked. If they public does not know this situation, that there are multiple factions fighting for control over humanity, then they will not know who to trust. In that situation, the ‘other factions’ will surely show themselves in an attempt to trick and gather as much human genetic material as possible.

The only way forward to the future is to share the truth about Solar Warden, the space-faring fleet of humans who do no contains the mental and physical illnesses that are currently tearing our planet in two. I believe Solar Warden saved my life as a child during more than one event where I would’ve been entirely maimed/paralyzed or simply dead on the spot as a result of other faction interference. I was in the bases as a child and was trained as an operative for MILABS a secret division of human space-military. The military has been preparing for the future by testing for various flaws and potential uses of cloning and genetic engineering.

Team Light and Team Dark Non-Interference Agreement

Team light and team dark agreed to a non-interference on ground level rule in order to determine how history would play out. Think about it. You have two groups of people, both with immortality, zero-point tech, virtually infinite energy and power with two opposing directives. How do you get to the future? It was decided by elders of each group that there would be no point in fighting an all-out overt war, this would likely destroy the population as much as it would destroy the Earth. As well, the elders of the elders which no body knows would likely have a problem with these groups going at it in such a way so as to change the flow of the planet for every individual thereafter. So it was decided that there would be a “non-interference” agreement for the surface population. Anywhere other than the surface, one can interact with the current population.

All Contact Must Be Initiated On an Individual Basis

On the surface everything that is generated must be done so through living channels in the population and the civilization must be preserved in its level of technological advancement. No overt motions would be used unless an absolute necessity or the majority of the civilian population had naturally come to the awareness of these possibilities. Doing so before this was met would undoubtedly and permanently change the civilization forever.

So what began as a collective motion towards higher awareness became a series of advancements and set-backs all taking place within a short amount of time. This represents the actions and operations of either side effectively helping humanity along it’s path or setting people back with orchestrated events such as the suppression of ancient knowledge, the manipulation of the education or other systems, or the generation of war and prison for profit based on the previous alterations of the education or other systems.

So team dark used its power to help turn society in a warlock ruled prison and indoctrination system of absurdity and scarcity. Team light used their power to help awaken those in need and protect those who’s actions are beneficial in the overall plan to save mankind from itself. Team light essentially buffers and shields the population from the operations of team dark.

Team light follows a non-interference agreement while team dark merely pretends to do so as much as is necessary not initiate war.

A Singularity of Change and Awareness

Neither side wants an all out war. This would reduce the dark faction’s chances of gaining a majority negative harvest because their motives and presence would be instantly revealed. Team light would also have to show themselves and again step in which is not a pattern of learning that prepares people for the experience of the self and the truth in the Universe.

Both sides know people have to accept the path they can offer them and cannot force anyone into it. They also know that the more this goes on the more likely people are to collectively awaken to the reality of what is happening. It is like an eventual singularity of increased awareness.

The motions we see today are the results of many people coming into awareness at this time on the surface in order to initiate a harmonic alignment as well as to investigate the ongoing crimes against Humanity.

As a result of this, both teams know that the likelihood of collective shift and spiritual awareness is becoming more and more real.

Many events are being allowed because they are representing the eventual shift towards the one end of the polarity from the other which is felt as a shock to society, and this is used to propel society forward through whatever means necessary.

Either people will see what’s happening and agree that we’ve been in a feedback cycle of negativity for the entire age of existence this civilization is aware of, and then we will shift into the real. Or people will claim that this is how things are supposed to work and instead of realizing everything is controlled in a deception based system, they see this and attempt to use the deception based system to gain more collective awareness. If this occurs by the majority then it will be seen as an “acceptance” to initiate a plan that will generate the overt perception of a deception and oppression system. This will surely shock the collective into realizing where they are.

If the majority comes to awareness of this deception based system then the shift moves from the end of the age of deception and external power into the age of internal awareness and harmony.

Advanced Technology and Ancient Knowledge

The plan is to slowly release advanced technology which will undoubtedly connect with and increase the awareness of ancient knowledge. Such technology nullifies the problems that the planet is currently plagued with, although without a developed sense of the spiritual self the technology becomes destructive.

Cloning, Atmospheric Civilizations, Genetic Stability

Here is something that most people might not have contemplated yet is possibly one of the most important factors at play here.

The vehicles that house the breakaway Human society are entirely environmentally controlled. This means that the air and surfaces are clean of all the viruses, bacteria, or pathogens that are numerous on the surface of the Earth.

This creates a specific difficulty in the situation. Any contact between surface Humanity is so heavily and carefully controlled that regulations go beyond any situations we have present on Earth. If the safety and cleanliness of any of the society's vehicles were compromised this would endanger everyone "on board".

So this explains a very precarious and specific situation we currently face and this is not so different than "first-world" nations contacting indigenous tribes and similarly endangering their society. Everything that happens will be according to a carefully controlled plan.

Training, Simulations and Learning

Part of what went on in the underground base experiments were to design "tests" or "labyrinths" with various people, props, traps, and treacherous situations that were to be navigated and escaped from.

Feedback loops with the brain, consciousness, and environment can create a very complex and integrated learning experience that is customized to each individual learner.

What this would do is create an individual that was capable of learning faster and faster and processing information in a more efficient manner.

This is not the same as specifically trauma-generating situations.

Cloning and Cerebrally Enhanced Soldiers – MILABS

This includes experiments, condition, trauma-based mind control, enhancement, memory suppression and operations that cover a span of space and time.

The NAZI faction has used cloning and genetic engineering to create hybrids, soldiers and everything in between. Religious/spiritual icons, to slaves, to leaders and even robotoids of sorts.

Your genetics BELONG TO YOU as the LIVING(ETERNAL) SELF. These issues are going to come more and more apparent.

If they agree to allow cloning, then who says who is allowed to clone whoever they want, Einstein, for example. And then who permits who to get DNA from history or objects in the world. These issues are going to have to be looked at because they have already been used in ways that are far beyond the basics.

Then what about consent of the masses? This is clearly being put in public because consent is required on an overall scale and individual to include people.

So what if the masses consent overall, would the people who choose to remain be at risk? Would Tribal and Native village life and people who respect the Earth be protected or permitted to stay?

Psychological Testing

These scenarios can also be generated to create specific situations that will contain tests or elements of temptation that will probe the person's psyche for weakness.

Hallucinatory Holographic Mind Probing Technology

Through the neural-interface technology the minds of many individuals can sync together to experience the same virtual simulation. Thus a single individual's mind can be selected for input and the environment will be a construct of their mental activity. The others present will be able to interact with and explore that environment like the person's mind.

Various training scenarios were utilized to enhance the resistance (or lack which) of the mind to intrusion and subversion. These were harsh but allowed the individual to completely control the energy flow in the conscious mind while remaining in harmony with the subconscious.

In a strange yet perfectly reasonable way, the conscious mind acts as the central "self" in that environment, while the subconscious mind becomes the "background" elements of objects and settings, actors and intentions within the flow of the "scene" itself. So a kind of movie begins with drawn out characteristics and perceptions that match that of a highly cinematic movie or intense dream.

The brain is naturally naive to stimulation and if allowed will liken the event to a dream in which the suspense of attention or lack of lucid self-awareness is "acceptable".

The truth may be worse than what you're thinking as far as training simulators and mind hacking devices. The truth may be that the human mind is inherently lucid and self-aware and that it is external influences that cloud the mind into what we consider 'regular' dreaming and 'regular' waking life. It may be that the conscious mind has a much greater access to information through the subconscious mind which is largely covered up and ignored in a rigidly physical and ego-focused society. As well, it may be that dreams are meant to become lucid and that we are often the subject of mystery in other realms that often refer to humans as ‘sleepwalkers’ that don’t know their dreams are real.

Virtual Simulator Training Scenarios

These scenarios can be entirely holographically generated to produce a visceral 3.5 dimensional experience which can be used for training or psychological testing.

Combat Simulators

The name is self-explanatory enough. Situations can be generated which are indistinguishable from the real events. The entire neuronal map of the brain (connectome) can be created and this is by machines to mimic the brain. After calibration through sensory feedback experiences, the sensation of the simulated realm is 100% the same as the neuronal impulses that are detected by the physical brain. An adept souled being can always tell there is a slight difference, but once the machines are calibrated the physical sensation is generally the same while in the machine.

These environments and scenarios can mimic any combat situation to determine accuracy, integrity, endurance, skill, and all areas of aptitude of an operative without the risk of injury or the lack of real-time effect.

Conditioning

As was explained in other sections, these same systems can be used to remove fear of certain scenarios. They can be used for programming the mind. They can be used to train one to complete a task over and over until they can recreate that action extremely quickly in real-time. These devices can be accelerated to operate within a fraction of the time yet to cover training and conditioning that would normally take place over hours, days, weeks, to months.

Programming and Resistance to Psychological Trauma

Through the development of specifically customized scenarios and neural feedback systems trauma can be brought to the surface or specific fears and traumas can be neutralized. This is essentially “facing fear” but in a completely generated environment that will be just as real as the real thing. One can literally face any fear, any trauma, any kind of difficulty and these systems can be utilized to accelerate a process that would normally take weeks to months or years and individuals can be strengthened to their maximal capabilities.

Accelerated Mental Functioning

Some of the technological systems were designed to test the enhanced functioning of an accelerated brain.

The mind can be seen as the energetic component of the physical brain. The physical brain projects consciousness via the electro-chemical processes.

Technology, Accelerated Learning and Repair

Technology has been developed for assisted learning and accelerated repair of tissues.

On Orgone Energy

Orgone devices produce a scattered signal which disrupts the ‘negative’ energy fields. At least this is what I’ve been told.

Apparently if built correctly it will either disrupt the negative emf/harmonics or it will produce positive harmonics, IE: Orgone energy. Orgone energy is the raw energy of life, while dead orgone, DOR, is the lack of orgone energy and can be “found” (the lack of orgone) in thunder storms and illness or depression.

This is exactly what I was told. We are generators for the same energy, and with the correct principles of energy and internal knowledge combined we can produce those very effects at will.

The energy of the mind creates a central vortex within the body. When the frequency pattern emission is consistent with fractal congruence, then that holographic printing of energy can be expanded or shrunk infinitesimally and it will retain the original fractal configuration.

This is what enables transference of awareness between fundamental harmonic layers of the Universe. When one’s consciousness energy pattern is balanced and tuned to itself so that all proportions can be shrunk through a vortex and expanded on the other side, then the personality or self-awareness remains true. This is as if the space/time must turn inside out and only when there is fractal relative polarity in time does the ‘shape’ of the consciousness within the DNA, work both ways folding and unfolding to represent the same mirror image. This is also symmetry in time as far as an ability to reverse the perception and still retain some kind of connection forward and backwards.

Energy Generators and Serums

Generators which produce vibrations that influence the cells to heal at an accelerated rate, combined with serums that enable the fuel for chemical reactions can instantaneously repair tissues.

I couldn’t believe what I was seeing at first, thinking it was a trick. There are serums and energy devices that can influence the healing and general health of the body and even mind.

These devices use energies which extend beyond the physical and are often compared to electromagnetic waves combined with ultrasonic or low to mid frequency tones. These produce harmony on a level that directly interacts with the biological functioning of the body’s cells.

I feel that this technology utilizes the same energy that is emitted from the mind and it is as if the device is “speaking” to the cells in the same kind of universal language of electromagnetic pulses, tones and vibrations.

There is information on energy devices and the use of light therapy in physical therapy today. More is being learned about the power of light and vibrations to influence the mind and body.

Genetic and Cybernetic Enhancements

Certain individuals are also genetically and technologically modified to produce a kind of hybrid human that can heal at a spontaneous rate as well as operate at a much higher rate of speed and efficiency than a regular person. Serums to induce muscle growth and mental acuity are applied, but the main condition is the enhancement of certain features through the alteration of the RNA aspect of the genes.

These features are combined with tech-assisted metabolism and certain biological functions can be enhanced or controlled via a supercomputer. These are the ‘brain-chip’ interfaces which operate to function as a governor for the mind. When the mind becomes too amped up to focus on the details the faster than thought computer system organizes the input into more usable information streams. When the mind is too slow to see the necessary patterns and variables to accomplish the goal, the brainchip operates in a self-preservation fashion and links together information streams to highlight the most viable possibilities.

Such technology is not without warning, much of this is being used heavily in the military field and this has given rise to entire factions of enhanced cybernetic soldiers. Although this is something to know, there are also many other uses for this technology and it is used by people who want to explore the technologically assisted path.

There are many other aspects of this to discuss, or at least a few important ones that will have to be understood to know the greater truth about what is happening on the planet.

Developing The Self

The issues we face are to help us learn about our selves and better mediate the known and unknown aspects of personality. The more we focus on realizing the unknown aspects of personality the less we are seemingly controlled and forced to experience the undesirable aspects of existence. The more we cling to just what we know and what we feel comfortable with the more shocked we are when we inevitably discover the truth of the whole of our existence which is *always* at least one half more than just what we know and are comfortable with. That is the nature of the universe, we can only achieve one half of the equation at any given moment. Once we complete that seemingly whole view of the universe, the information we gain access to instantly doubles and we realize that we were only looking at one portion at a time, just one half of the equation. Then the rest of the equation comes into the picture and we must expand our view once again. This is because we are viewing ourselves as we discover the universe because we are creating what there is to be discovered through the act of exploration in itself. We are the explorer, the explored, the creation and the discovery simultaneously.

Multiple Factions of Space-Fleets, Multiple Goals

Now there are what are called ‘factions’ of groups that have different agendas and interact with the public or surface level population differently. That is one of the main issues. These departments and factions are no longer limited to the surface level of existence and have access advanced technological capacity and knowledge of a larger view of time and a more complete understanding of consciousness and the human existence.

Large-Scale Catastrophes or Series of Meltdowns

I was informed that while they work continuously to prevent and reposition this civilization continually outside of the reach of any large scale meltdowns that the technology is commonly accepted to only be a temporary barrier between the inevitable.

Unsustainable

This society is unsustainable and must either change or will experience the massive catastrophic evens that are foreseen.

There are multiple perspectives that are maintained. They vary by the way of the methodology of solving the problem. All agree that change must occur and is impossible

to postpone forever. The future society is the result of our advancements and achievements. Regardless of how we get there, we must get their somehow.

Earth Changes

I was informed that these changes had been held off and would ultimately be allowed to take place to resolve the current situation. The changes can only be held off for so long, either that or the people performing this task would not be reasonably inclined to postpone these changes indefinitely.

The idea is to locate and secure the members of society who are fearless and capable of assisting in this transition and who hold key genetic elements as these individuals and groups are capable of ‘holding-space’ for the rest of civilization by literally containing the genetic precursors for all those other individuals.

I was informed that Earth changes would be the last event to take place, pretty much like icing on the cake.

This means there will be a lead up to these events.

Mass Consciousness

I was informed that the consciousness of humanity and every species is linked to the bio-etheric field of Earth. When the consciousness becomes too unstable to maintain a controlled society that this will initiate massive Earth changes. This could simply be instigated through advanced technology but I was informed that the advanced technology is actually postponing this so as to allow people to prepare, to get the word out and avoid a complete loss of the human race.

Wars and Manipulation

Wars were described as a way the power groups had controlled humanity over long periods of time. These wars were used to distract and weaken the population and the power groups believed they were holding off a full scale invasion by suppressing the rising population.

Now the population is high enough that another war would be a final war and this is extremely dangerous for the entire civilization. There would be no need for another war to repeat this cycle because the technological achievements have been met and so essentially the ‘reseting’ of the civilization would not be required or possibly acceptable.

This is why war is to be avoided however people are very easy to manipulate through lower emotional and mental programming from media and social cues. People follow the herd and the herd is essentially still operating on animalistic parameters. There is a have coating of peace and humility but this is just a wax coating that quickly evaporates as soon as the pressure of mass unrest or scarcity is realized. Many simulators have explored these possibilities.

Financial Unrest (or reset)

The idea is that financial oppression is used before wars to distract and oppress the population. When people are busy looking for resources to survive, they are not focused on the larger picture. This is part of the ‘fight or flight’ response mechanism tied to lower brain complexes which are primarily activated to keep the masses in both a highly sensitive and unstable state of being while also keeping them easy to control.

If the civil unrest reaches a level where awareness continues to rise even though financial hardship is being implemented then the latter two options will not work. At that point artificial disasters can be produced as well as natural Earth changes which initiate a shift of the ages and of consciousness.

The paradox or maybe irony here is that the oppressive actions are used to both suppress human awareness up to a certain point and simultaneously to stimulate and arouse the human spirit into wakefulness and self-empowerment. This is part of the reasoning regarding why these control systems are allowed.

Some say the elite don't want an economic collapse. That it would ruin everything. Others say it would be icing on the cake. Some say that it is a zero-point world that will save us. That the Fukushima is a big problem. Zero-point field tech, that is. Generators and energy sources, healing and terraforming. I know we project right? We see it in our own lives. So we can project with each other and grow that as a group and that part of the world will exist in that light. Which sounds Biblical, which may be because they're following prophecy. But it may also mean that we have a light, a chance, a real salvation, but that we have to grow that here and help develop it. To pass it on like love and such.

That's apparently what it comes down to. That these cycles happen every so often and we're in the midst of one right now.

Let's make it count. Every day, thought moment. With mindfulness, intention and awareness in action. Life is literally higher dimensional, so we have to become multidimensional to really face it. Indeed, that is a good perspective on action and inner being. It's like a fractal, whatever we project inside that chest center, will project into our mind and the outside. But we have to be stabilized and grounded through the root first and then to get to the heart energy must pass through the sacral/sexual centers and continue upward.

Self-Responsibility

The main reasoning is self-responsibility. We are responsible for ourselves. The people are and so if they allow themselves to be tricked and prodded like animals while behaving for the part, then that is how their experience will play out.

Clones and Synthetics

One seemingly unfair aspect of this is that there is a larger portion of non-humans than there are humans and so these beings without the human spirit will be easily programmed to support and propagate the mind-control system and the oppression. Thus, if people follow the herd mentality then they will walk directly into the end-time scenarios yet it is the humans who will actually stand to lose. People must wake up to these tricks and this technology so that they can orient themselves with the reality of this situation while learning to strengthen themselves and adapt to the manipulative aspects of this control system.

Spiritual Power and Divine Truth

This is essentially all about spiritual power and humanity stepping into the power of the truth behind their existence.

Some say that humans are designed as slaves, but if this is so, then it is the power that humans contain within that is the ‘work’ that they are sought after for. This indicates that humans have great powers that others cannot achieve on their own, even if they are creator lords.

Thus, whether humans are modified or not, they have great power. If they are not modified to perform this role of an energy generator, then this is just a confusing mind-control game that has been pulled over the eyes of humanity and humans have always contained true internal spiritual power that allows them to exist without having to parasite and steal the energy and knowledge of others. Humans are reality generators , we can choose how we want to experience the universe and what kind of collective world we want to enjoy together. Once we step into this power, there is on greater force. No amount of non-human entities can change that as we are somehow intrinsically linked to the ether-space of this realm through the spiritual and physical aspects of our being. That is what is sought after and so those who do not have this manifestation power attempt to utilize humans to create their own reality through this Earth situation. That is the basics of the ‘alien overlord’ system, whether humans in disguise, non-humans operating with alien intelligences, supercomputer systems, or spiritually degraded beings.

Most people are not ready to see themselves or the true nature of humanity. Reality is largely an illusion of the conscious mind. The hidden aspects are the other half of the equation and this is perceived by the shadow aspect of the personality and mind. We are only marginally aware unless that aspect has been unveiled and through this process the reality of the human soul can be actualized and realized.

I was informed there would be a period of two suns and multiple mirages around the world. People would believe they were peering into another dimension, floating cities, or some kind of city-ship craft.

Synthetic DNA, Synthetic Genetics

Through the previous described replicator devices synthetic DNA can be engineered. Thus, entire genetically engineered codons can be produced and form this created beings can arise. These are the hybrid beings that are performing many tasks underground and even on the surface. On the surface there are synthetic genetic create beings or genetically engineered beings that appear to be human but do not contain a human genome. There are also clones and replicants of humans. There are a variety of non-humanoid forms but these are largely kept underground. These are beings we often see in science fiction movies or shows.

Time as Mathematics, Series of Equations, a Programmable Storyline

When using the technology, time can be understood to be a series of programmable variables where one variable is centered within a series of parameters and from these possible variables within a parameter set a series of possible functions or outcomes can be process. Thus, these people can literally plug into a supercomputer system which operates based on this form of processing reality and through this they can observe all the possible realities that may connect to a present moment. This allows one to travel but this is not part of this section of the explanation.

What this allows one to do is know exactly what is going to happen and when. If a person is being interrogated, all that has to be done is the operative using the technology simply thinks of what they will say and intends to do it within 6 seconds. The machine will play out a future timescale regarding the input of that question or interrogative speech and then produce the possible results. If this doesn’t give the desired readout in the mind of the operative, they will refuse to go that route and think of another pathway. This repeats for a few seconds to a few minutes. Then the computer has found the pathway that is most like to yield a result. The operative simply steps forward, says ten words, the subject breaks immediately and tells them everything they want to know. This is obviously takes longer for most trained individuals, however that is how easily it would be to use this technology to produce an effect on a regular surface level human. They would have no defense, no way of knowing how that individual could know so much or be so direct. This is how operatives dealing with certain power groups will operate in the surface level population and individuals always feel that these people know way more than they are leading on. That is because they are, they know more about you and your life experience than you do.

Chapter 2.3 Previewing Solutions

Neutralizing Polarization

The mental-emotional tendency to pick sides, to fight, or to basically join in the fray that is currently taking place is a kind of deception of polarization which enables the leeching of emotional and mental energy from the human. This is more of a psychological unveiling where the true nature of the self, the duality game and the transcendent source aspects can be made visible to the individual. The they will see how every polarization tactic is actually taking them out of their true source of power by pushing them off center emotionally and mentally. There are teams that are capable of enabling this process and assisting as humanity realizes that half of the problem is that the unchecked mind is one’s own worst enemy.

Removing Etheric Implants

Technology can identify and isolate these foreign body implants which are etheric technological devices. These are literally frequency or phase-shifted devices which are invisible to this layer of physicality but are capable of being organized and attached to the etheric field of the human and will operate to siphon or limit energy on the etheric level. Of course, this is all technology of the vampiric created beings who parasite on the original Earth inhabitants to sustain their existence.

Clearing Astral Body

By moving through the temporal body etheric implants can be deactivated and removed and then the astral or emotional body will be able to function closer to the original levels of presence and sensitivity. This body is manipulated when one is emotionally shut down from trauma. Thus this body can be scarred and will function at a much lesser level of activity or presence until the emotional damage is healed. This contains the memories of all the emotional events of a person’s life.

There is technology that enables one to access these levels of frequency and clarify or untangle emotional chords or attachments which are used to siphon energy from the human being to the parasitic created beings. These chords are energetic in nature but also have technological components.

Knowing the True Self; Soul Actualization, Unveiling

This is a sacred, ancient process that has been outlined and discussed since the beginning of time. Humanity is here to learn, grow and express the self. In doing so, like a child moving through layers of infancy and into full-wakefulness of their body and the physical world, humanity can come into awareness of their true spiritual nature and the true spiritual nature of the universe. This is the process of soul or self-actualization that is described in some psychological methods such as Maslow's hierarchy of needs and the self-actualization that results in a complete fulfillment of all psychological needs. In this modern psychological view, self-actualization is described as the desire and process of reaching fulfillment of one's own natures and then living their life sharing that self-empowerment with other beings because after gaining your self that is the highest expression of such knowledge, sharing that power with others.

The true self is not limited to the physical body, the true self is a reflection of the entire universe and thus the primary original creative force. This is akin to the whole equation rather than just half of the equation. The visible aspect of the self, the conscious mind, is only one half of the entire equation. The whole equation consists of what we see visibly and consciously as well as what we are limited from seeing within the unconscious and subconscious mind. These are really two layers of the same ultimate source. Like an iceberg which is only partially above water, the subconscious mind is the proprietary wholeness of experience. The conscious mind is only a surface layer of reality while the unconscious mind carries all the notions, the multidimensional processes, the latent aspects of reality that we are not capable of seeing without unveiling these aspects through inner exploration.

Taking Responsibility

This entire show is about taking self-responsibility. Unless people begin to do this, there will be no hope. Right now people are convinced that their responsibilities belong to someone else even though this always results in pain and discomfort. To take responsibility is painful at first because we have a lot of work to do, but even then, this is beneficial because there is no way to remedy the pain and discomfort of being lied to and disrespected if people continue to hand their free-will over to others through blame (scapegoating), lower emotional bio-emissions (hatred, anger, etc), distractions and addictions (technological, sensual, and chemical or food based), or generally complaining without taking the first step towards providing something beneficial. This is the victim role, if humanity is to take control then they to help each other and themselves by not relying on false promises and false authorities to determine how they feel, think and live.

Returning To Center Emotionally and Mentally (spiritually)

Balanced food intake, rest, exercise, activity and learning are all continual requirements for a healthy existence. Emotional and creative expression nurtures the soul. Without these ways of being and learning the human qualities of experience diminish and creativity dwindles. There is most likely going to be a split between those who can sustain their own humanity and those who have to cause others to suffer to do so.

Vampires in the “Awake” Community

Contrary to popular belief, the majority of the non-humans are not necessarily vampiric and in power. The vampires are the people online who try to get a rise out of others for sharing their own experiences. They are the nosy neighbor that feels that the person who looks or acts differently must be considered and enemy and armed against because they feel themselves to be so much better. The majority of people are vampires, they believe that they must work for someone who will trade them money and that the amount of money they have based upon the amount of indirectly beneficial work they do puts them above others.

This is not different in the ‘awake’ community and in fact I feel that there are higher concentrations of vampires in this community, the alternate media because everything is largely a joke and a trick. People are operating on deceptive agenda schemes to divide the people and this is most intensely noticed when there is a realization of the truth behind what is happening and people begin to point fingers. The only people who are cleared to point fingers (and who also will NEVER point any fingers) are the operatives from the secret projects that have the capacity to do so because they have reached a specific level of awakening and this is related to the droning, cloning, and hybridization process. Only a specific few know who has been transformed into an impostor being and who is remaining strong and this is due being briefed on the surface level situation. Anyone can say what they want, they are simply playing into the tactics of mind-controllers and parasites and the more they play in the deeper the infection will grow. All the people pointing fingers now are actually implanted and vampiric and are attempting to drive a divide through humanity and they are the first who will turn on their friends and family when the going gets tough because they are more interested in others and starting drama than healing themselves. Ironically, this is all done under the guise of “healing” and “getting rid of the drama” AKA people they don’t trust. They pick and choose based on their own personal bias and this is a sure indication that they are emotionally and mentally out of alignment and are feeding deep seated personal fears from their own traumatic experiences.

These traumas must be healed and the harmonious center must be reattained where a person can focus on their own existence and waking up instead of acting as false-light, “king of tyranny” gatekeepers for others. Those individuals are going to be left behind if they cannot kick the habit because if they are allowed into any kind of healed society they will quickly degrade into animals and attempt to restart a devolution process in order to gain the power they so desire but cannot produce on their own. ANYONE playing ego games of trying to hurt or put blame on ANYONE else contains this viral implant and has not kicked the virus.

Heart Center and Psycho-Spiritual Fulfillment (self-actualization)

Emotional connectivity is a form of telepathy, when loved ones can tell the state of another across space and time. Whether it is family or friend. In the 60's and a little before that the US found a specific frequency akin to a radio frequency that operates on a trinary spectrum. They called this the "telepathic wave". In time, they could completely read and send thought waves without frying the brain or body through radiation. Which was what the early machines did. They actually took satellite dishes and aimed them at people. Then realized they could turn it around and do it backwards with a better effect and minimal radiation. Either way they started beaming the base and then towns around them with waves between 400 and 432 MHz and caused waves of crime sprees or peace sprees throughout the 70's and 80's. So they can digitize the telepathic wave to induce artificial telepathy. But this is organically done through an entanglement through what we call love. Literally. The Germans found it before that. And the secret orders before them. And ancient Egypt before them. As humans we remain basically through the root, heart mind connection. When the heart overpowers the mind connection of spirit, then the body is a trinary whole. When the mind guides it is binary and cannot by guided by the spirit. The spirit has direct access to the heart. The mind is a computer of sorts. The mind doesn't see past duality. The heart is the true center/eye of the spirit. The reptilian brain is the dragon and when you tame the amygdala then fear is ‘behind’ you. When the mind is "untamed" the reptilian brain takes over. This is the lower brain and amygdala fight or flight response.

This happens when the energy of the brain, body and spirit increases due to whatever various reason of spontaneous or controlled increase. Because of the "temptations" which are the easiest flow of emotional energy into a "discharge" state, is that the sacral/sexual center ties to the heart and reroutes to replace the heart and control the mind. Mental and emotional contemplation of one's true purpose and nature is the path of the divine. We find ourselves simply by looking in the right place, not in the world, but in our self energetically. We are more powerful than the programming, and we can reverse this is we're careful and devoted enough.

We have to be faster than thought, at the speed of spirit, beyond matter. When we start thinking on this level then our influences spread to others throughout the sphere. It is our destiny to know these things but they have been kept from us. The ancient orders called this the art of "implosion". Traveling realities through intense contemplation and a usage of sacred geometrical forms.

They developed and combined orders around WW2 and developed zero-point energy tech that can alter space/time fields. This is timeline manipulation. But as a rule, we go where our heart leads us. So that's how this is solved.

Go Within to Observer the Self

The answer is literally by looking within to find the great truth rather than looking around us. Everything we need to know is within us, energetically and spiritually. Our emotional, mental and physical state of balance is an indication of where and how we exist in relation to the larger overall truth. Everything that happens around us to distract us is designed to stop us from looking within. Nothing can truly stop us, nothing can actually input the holographic horrors in our spiritual center but simply perturb the brain and body. The mind is a non-holographic spiritual environment that cannot be excessed and the invader's goal is to convince people to use their imagination or their creative powers to produce these horrors on their own. This is through the programming. We must clear out the programming, the trauma, the lower-dimensional constructs and fill ourselves with the truth and spiritual presence that enables us to be self-aware in the first place. If we aren't doing this, then we are either in pain, or there is no spiritual presence to be in pain. If we are doing this, we are either clearing out the constructs by replacing them with true spiritual presence and self-awareness or we are becoming aware of them.

Organic Feedback Loop

The great feat is mentally bridging the great rift between us by looking within. Which is paradoxical.

We can never see what another is perceiving exactly, well for the most part, so we all have to kind of accept that we are communicating effectively and truly get to this level of comfort by simply being secure in knowing who we are our selves.

Technological Feedback Loop

So then knowledge, in a way, when shared, is a form of telepathy. Group telepathy.

Group think, when ideas or memes spread, is a form of group telepathy.

The Internet is like artificial spreading of telepathy. This means what is normally mentally bound signals can be initiated and spread through a digital medium.

Breaking Soul-Contracts

Ultimately everyone is under a series of “veils” which are energetic time-constraints literally like 9 dimensional layers of temporal restriction which are each bounded and ruled over by invader beings. These range from physical authorities, to astral entities, to programmable matter, to supercomputer intelligences, to spiritual false-light beings and our own shadow mind. Each layer of reality is successively guarded by these beings who are here acting as gatekeepers designed to push people back when they begin to wake up. These beings all work through fear and deception and the last layer is based on mirroring our own energies back to us like an impostor. All the levels tend to imitate, however the last layer is a near perfect imitation of our own energy to the point where people often switch places with this being before getting out of the matrix. There will be more discussed on this in a later section.

The “9 veils” were designed by impostor beings as a last effort to stop humans from escaping the false-light matrix which is essentially the holographic universe. This is literally like a maze which reproduces a new pathway and a new doorway each time a person nears the last hallway or doorway that is next to the exit. So one eventually has to project themselves into the exit using their whole being and moving past and through the tricks by not falling for them before they are placed. This is highly confusing but this is essentially a kind of labyrinth that has been created to stop people from leaving so that their soul energy can be destroyed and used as a generator for the false-light system which is an AI over mind that is programmed to take the raw soul energy and destroy the identity by inverting the electromagnetic spectrum and incorporating that inverted energy into its own internal matrix body which is literally the false-light universe.

Help Everyone

Now they may have been setting us all up for the apocalypse. They said they want us to tell on them and get the whole world aware of what they've been doing So that we can stop them and pull them from the dark which is ACTUALLY taking their souls. There are also breakaway civilizations involved in this entirely zero-point/free-energy.

Chapter 2.4 How Timelines Collapse

Repeating Cycles of Time

Because of the way time works and the ability of those using these temporal devices they have set up an elaborate labyrinth like a spider web of false realities, timelines, parallels, sub-dimensional layers, astral cages, temporal dead-ends, and time loops which act to stop people from reaching the true reality. As a result, many portions of this timeline are literally in a free-will ‘time-loop’ and these loops must be canceled out. As well, because of the nature of these loops on the overall stream, our entire civilization has been looping for the past 6 time-cycles meaning we have reached this time in history 6 times before and this is apparently the last time because we have never gotten this far before and the events of the future are having to be programmed in manually at this point.

Beings From the Future

As are result and damaging time, what happens is that the beings from the possible future that is being damaged appear to try to heal the break as well as those who are from the new possible future that is generated, this results in a time war between both possible realities for the present realm. This is what has been happening for thousands of years. All the ancient spiritual texts are describing this system. One group is a from an original possible future where everything is based on the presence of spirit and souled humans, another is from the new possible future where an AI overlord took over and removed the souls and spiritual energy from all entities on Earth rendering this a holographic prison without any other use. That reality dissolves into itself and so the time beings are constantly moving backwards and jumping timelines to try and obtain more energy.

Beings From Collapsed Timelines

The beings from the collapsed timelines are those who travel through hyperspace to get here. They then recreate themselves using cloning technology through hyperspacial means, IE: creating an alternate temporal dimension so that these devices can sustain them and literally “hold-space” for them. Then if their timeline is destroyed, they can still exist as a ‘loose-end’ in space-time. This is extremely dangerous and I will say we have some Human operatives jumping timelines searching for Earth and Humanity so that they can find a reprieve from this time war.

That is the most trying aspect of this entire ordeal for me. There are some human operatives literally in hyperspace without an opening into the temporal stream for them to reintegrate into this timeline. We must assist them, we must open the timeline by taking control of Earth from the invader races and those who wish to depress and isolate the spirited and souled human beings. We must re-organized and neutralize the invasion forces of deception and assist our humanity and maintaining their free-will, their self-awareness, their self-empowerment and their emotional, mental, and physical harmony.

Beings Who Would Never Have Existed in the First Place

As a result of this entire process, there are essentially beings who have been erased from time who degrade in this dimension and can only interact periodically. These are the created beings who operate through supercomputer forms that are placed within cloned bodies and powered by the transdimensional quantum supercomputers. These are the ‘parallel’ reality beings and this relates to the hive mind invasion. When all the original aspects of time are restored, these beings would not have ever existed in the first place as there is no quantum link for their probabilities to remain neutral and present in the original run of things.

Again, all beings who operate under the will of the universe, even when they are attempting to trick hyperspace and time, when these beings assist and help in the unification of sentient, spirited, souled beings they are capable of being healed. This is all because there is an original creative force which can produce souls that are eternal and everlasting, meaning they are not created but simply always ever were. Yet, there is a “miracle” effect where beings can be given eternal life. This is only by reflecting the will of the whole of existence, rather than a portion or a part because this is only temporary. Only the whole is forever and this takes true harmony, true balance and ultimately *self-lessness*. To give to others because you see them as worthy because they can be aware and enjoy the gift is to be like the creative force that gives life to all and enables all experience in the first place. By acting in correspondence with this we reflect the will of the whole and that power and action transfers into protection and self-empowerment for ourselves. This cannot be cheated, tricked, or rerouted as that is the karma game that has been played by delaying time and rerouting energy through getting others to take on the responsibility for others bad actions. In the end, all karma has only been delayed because when the stops are pulled out, *everyone* ends up getting their karmic returns regardless. They simply intended to build up so much karma that by the time the ties are cut, the karma overload will simply kill them instantly. This never happens and they end up in the abyss, this is the final time so since there is no recycling of the universe from the oroboros system then whatever happens is what happens forever. If everyone wants to move into harmony, to receive healing, to receive forgiveness then they must assist in the creative force in healing, forgiving, and turning away from the deception, the cruelty, the vampiric actions.

Many vampires believe they will die if they stop. The only hope is to admit their ways to the others and ensure that there are enough beings around them that will protect them so that this situation can be handled properly, with due care, compassion, and the necessary healing to ensure the safety and protection of all who seek harmony.

TIME TRAVEL PHYSICS

By Steven L. Gibbs

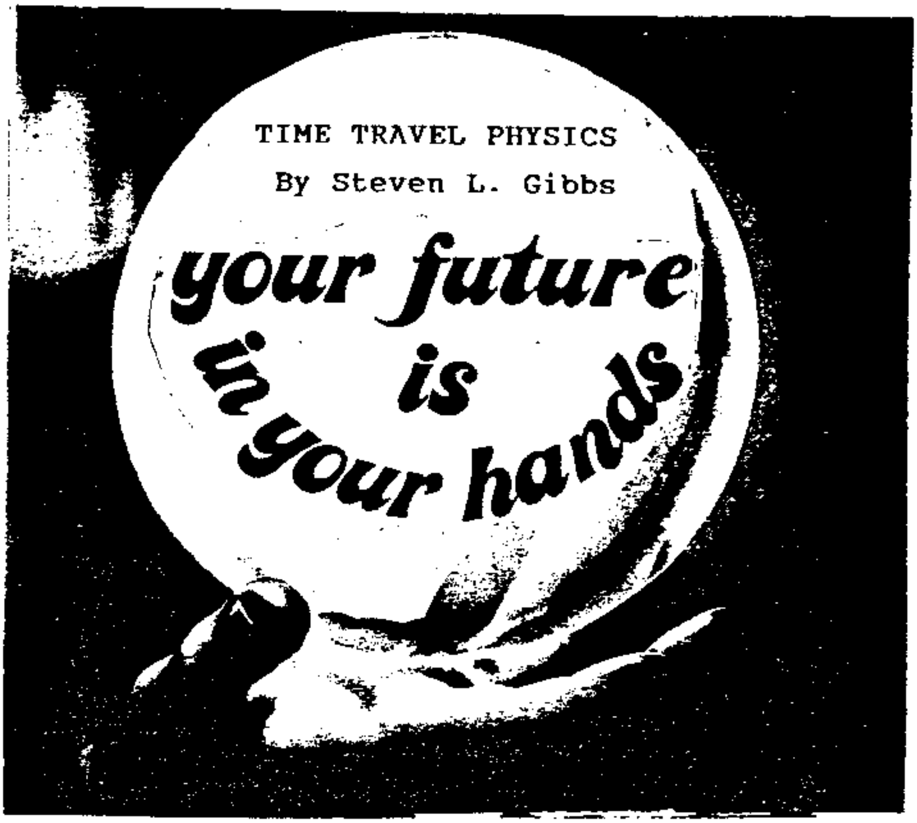
your future
is
in your hands

DEDICATION

© 1989 By: Steven L. Gibbs
Rt 1. Box 76, Clearwater
Nebraska USA
Illustrations Thanks To
Al Fry Distributor....

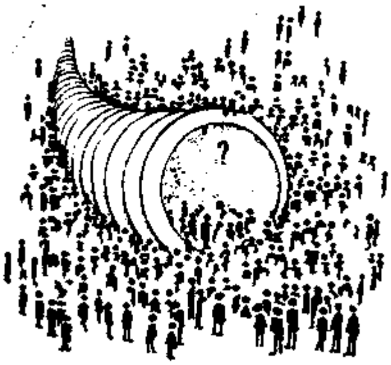

I would like to personally thank
Jesus Christ in helping me to obtain
the knowledge for writing this report.
For without his help, all of this
would not have become a reality.

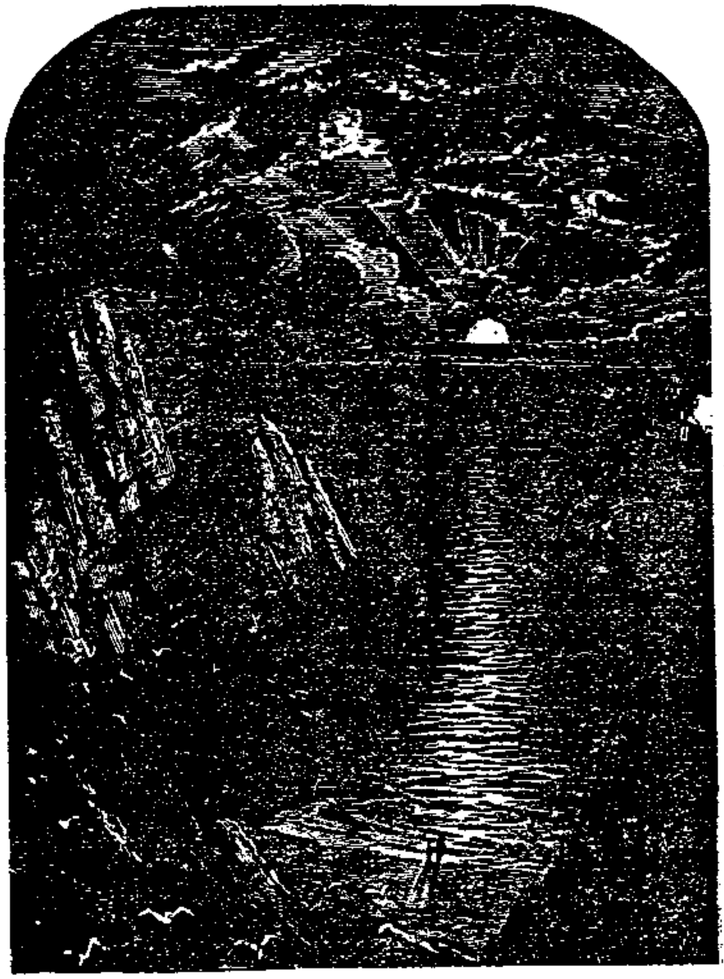

INTRODUCTION

The reason for writing this report is to show people that man does not have to use sophisticated mathematics in order to understand how Time Travel can be achieved. To assume that what is being taught in today's schools is an absolute truth, can only be a lie. For to believe in such a lie, can only lead to imperfection. But sense mankind has chosen this path, the inevitable must surely come to pass.

It is impossible to say where all of this might lead. However I do hope that the people who read this report will use the knowledge wisely. For those who don't, I leave them to their doom. But for those who do, may the Lord bless you.

Your probably wondering by now, if I am really an Alien. My answer to this is, there may be more truth to this than meets the eyes. According to the information which I have received from the Kennedy Space Center, there is a good chance that my parents came from the Star System called Phoenix. If this is case, then this would explain why I have had so much difficulty in adjusting to Earth's own environment.

Perhaps in the years to come, more will be revealed to me on the true nature of my whereabouts. And you know what they say about Aliens? Their weird!

TIME TRAVEL PHYSICS

Before I begin I must first point out, that the only reason why Formulas can be used for analyzing a specific problem, is due to the fact that a Formula helps in focusing the Belief Fields which emanate from the persons Soul. So in other words, whenever a person uses a Formula or Equation, to him this constitutes a reality. By keeping these things in mind, you shouldn't have any problem in understanding what has been written in this report.

To begin with, there are basically 3 sets of Equations which can be used for accomplishing Physical or out of the body Time Travel. I have used these formulas quite often when building my Time Travel Machines. As a matter of fact, all my research is based on these 3 formulas. The formulas which I have used for accomplishing Physical or out of the body Time Travel, are given in the following manner:

(THE GRAVITON FORMULAS)

$(X , Y = 0)$ $(X1 , Y1 = 0)$ $(X2 , Y2 = 0)$

The Graviton Formulas which I have so diligently labeled, can be used for an infinite number of things. However, before we can proceed any further, we must first have a good understanding of what each Symbol means. There are an infinite number of meanings which can be applied to each Symbol, however in order to simplify things, the meanings which I have given to each Letter or Symbol, are listed in such a way so that anybody with a little bit of knowledge can understand them.

(INTERPRETATION FOR EACH OF THE GRAVITON SYMBOLS)

X - Represents an AC Field.
Y - Represents a DC Field.
X1 - Represents a High Frequency Field.
Y1 - Represents a Low Frequency Field.
X2 - Represents a Paramagnetic Field.
Y2 - Represents a Diamagnetic Field.
Ø - Represents the Zero Vector or Twilight Zone.
it may also be referred to as the (PHI Value)
which is (1.618).

It is interesting to note, that whenever the (PHI Value) is used in the construction of a material object, a Time Warp usually occurs around that object.

However, in getting back to the problem at hand, whenever the (Graviton Formulas) are used singly or in series with one another, they can be used to help build a device which can actually be used for Time Travel. This is because whenever the (X) and (Y) Values are combined with one another, we can Zero in on the Twilight Zone. And sense the Twilight Zone is in attunement with God, all things become infinitely possible provided that the (X) and (Y) Values are in resonance.

So now your probably wondering, how can we use these Formulas for constructing a Time Travel Machine? Its simple. Merely start out by building a Capacitor which corresponds to this Equation: (X2 , Y2 = 0). This Capacitor should be constructed in such a way, so that it resembles the drawing which is shown below:

(DRAWING OF CAPACITOR)

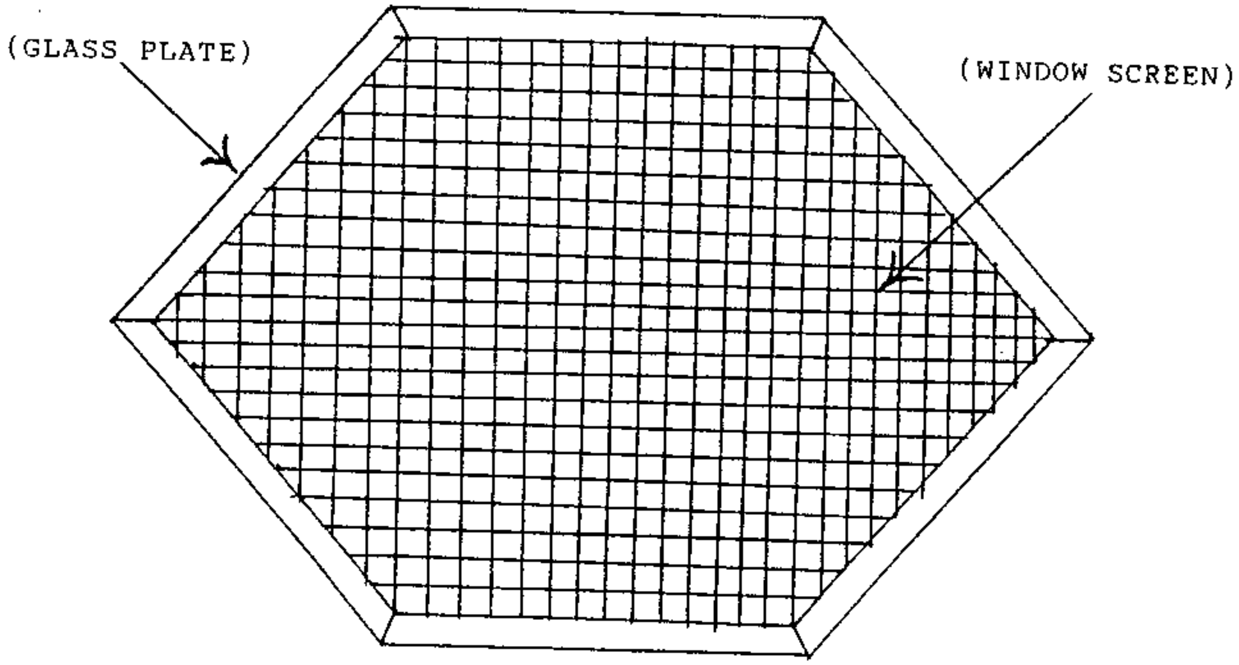

As we can see in the above illustration, it doesn't matter to much as to what size it should be, just so long as it is shaped like a Hexagon. Also each side of the Capacitor should be the same length, otherwise the Belief which is being transmitted from the plate, will be imperfect. And this could raise Hell with your Chakra Points.

At this point it should be noted, that whenever you construct a Capacitor of any type, you are automatically using the formula (X2 , Y2 = 0). And sense the (PHI Value) is used in the formula, the plate itself will develop a Life Force of its own. It should also be pointed out, that sense there is a Life Force which co-exists with the plate, it can now be programmed to do anything you want.

Now in order to program this Capacitor, all we do is stroke the Glass portion of the plate, while concentrating on the following command: (Transport me to the Date and Year). As soon as you get a stick so that you can no longer move your fingers, the Capacitor should be fully programmed. After you have finished programming the Capacitor, place the Screen Side of the plate over the Solar Plexus for the space of 30 minutes. After the treatment is over with, you then find a comfortable place to relax in order for the energies to take effect. If you have done everything correctly, your Aura or Physical Body should be transported to the date which you programmed into the Capacitor. On the average, it usually takes 70 minutes before the energies will take effect.

*Note: People who work with these plates are in serious danger of getting possessed by an Evil Spirit or Demon, unless of course you have a faith in Jesus Christ. By saying a prayer to Jesus Christ prior to any Experiment, more than protects you from the forces of Evil.

As you Experiment with these Capacitors, you will find that some of these plates work better than others. The reason for this probably ly's in the fact, that certain sizes or the materials being used, are in resonance with the person's Life Force or Belief Fields. Therefore if you run into any problems, it may pay to use a different thickness of Glass. When I talk about this, I do not mean to vary the Shape of the Capacitor, just the width. It also might pay to use Yellow tinted glass. You may also want to try Green, Blue, Indigo, and Violet colored glass. The reason for doing this is simple. Each color corresponds to a different Chakra Point. It is therefore my opinion that Yellow should be your first selection, then if you want, you can always Experiment with the higher colors.

*Note: Never under any circumstances use Red or Orange tinted glass. If you decide to use these colors, then your probably on your own, because not only will it lower the Frequency Rate of your Chakra Points, but you could also end up in an Astral Hell. And believe me, this is no fun place to visit.

Once you have found a plate which can transport your Physical or Spirit Body, there are still even further ways of increasing its effects. One of these ways is by using the second formula, $(X1\ ,\ Y1 = 0)$. This formula states, that whenever a High Frequency Field is combined with a Low Frequency Field, we can tune ourselves into the Zero Vector or Twilight Zone. To use this formula it will first be necessary to get a hold of a

Tesla Coil and one Van de Graff Electrostatic Generator. Once these items have been obtained, you proceed by charging the plate with High Frequency Electricity and Low Frequency Static Electricity. In other words, after you have programmed the plate using the procedure on page (4), you must simultaneously Zap the plate using the above procedure. You must make sure that while you are doing this, that you do not come into contact with the plate. If you do, you could be in for a Hair raising experience. This voltage hurts! But anyhow after you have Zapped the plate for the space of 10 minutes, you then proceed by placing the Capacitor over the Solar Plexus as described on page (4). The effects that you get, should be much better than what you experienced before.

The next step to take, is to Amplify the Energies even further. To do this, we must now make use of the 3rd and final formula, which is (X , Y = 0). This formula states, that whenever an AC Field is combined with a DC Field, a Time Warp occurs which tunes our minds into the Zero Vector or Twilight Zone. Now in order to create these fields, one should try to get a hold of a Hyper-dimensional Resonator. (Read the last 4 pages in this Report). Once you are able to get a hold of this instrument, and after everything has been plugged in correctly, while the Screen side of the plate is positioned over the Solar Plexus, position the open end of the Electromagnet over the glass side of the plate for the space of 3 minutes. After the treatment is over with, disconnect yourself from the plate (turn page) and instrument, and find yourself a comfortable place to relax. If you use this procedure in conjunction with the first two procedures, the effects that you obtain, will be increased to such a point, that you should be able to do anything you want. You must remember however, that before Physical and Spiritual Time Travel can be achieved, the (X) and (Y) factors in the formulas must be in resonance with one another. This can only be accomplished through the process of tuning your own mind just as you would with a Radionics Machine. So in other words, by selecting the correct materials to use, you are in effect creating something which you physically believe to be a reality. This same rule can be applied when working with different types of Frequency's and Voltages.

Another method which you may want to try, is that after you have programmed the plate by using the procedure which is given on page (4), transmit some High Frequency Electricity from a Tesla Coil into the Capacitor while the Screen side of the plate is pointing towards your face. After 3 or 4 minutes, de-activate the Tesla Coil, and while the Screen side of the plate is still facing towards you, take the bare ends of an Extension Cord that has been plugged into a 110 volt AC outlet, and short out the bare ends of the wires by touching both of the terminals on each side of the Screen. (See page 8). If you have programmed the plate correctly, your physical body should be transported instantaneously to the date which was programmed into the Capacitor.

THE CHRONOLOGICAL TIME REFLECTOR

(Series No.2)

2-21-89

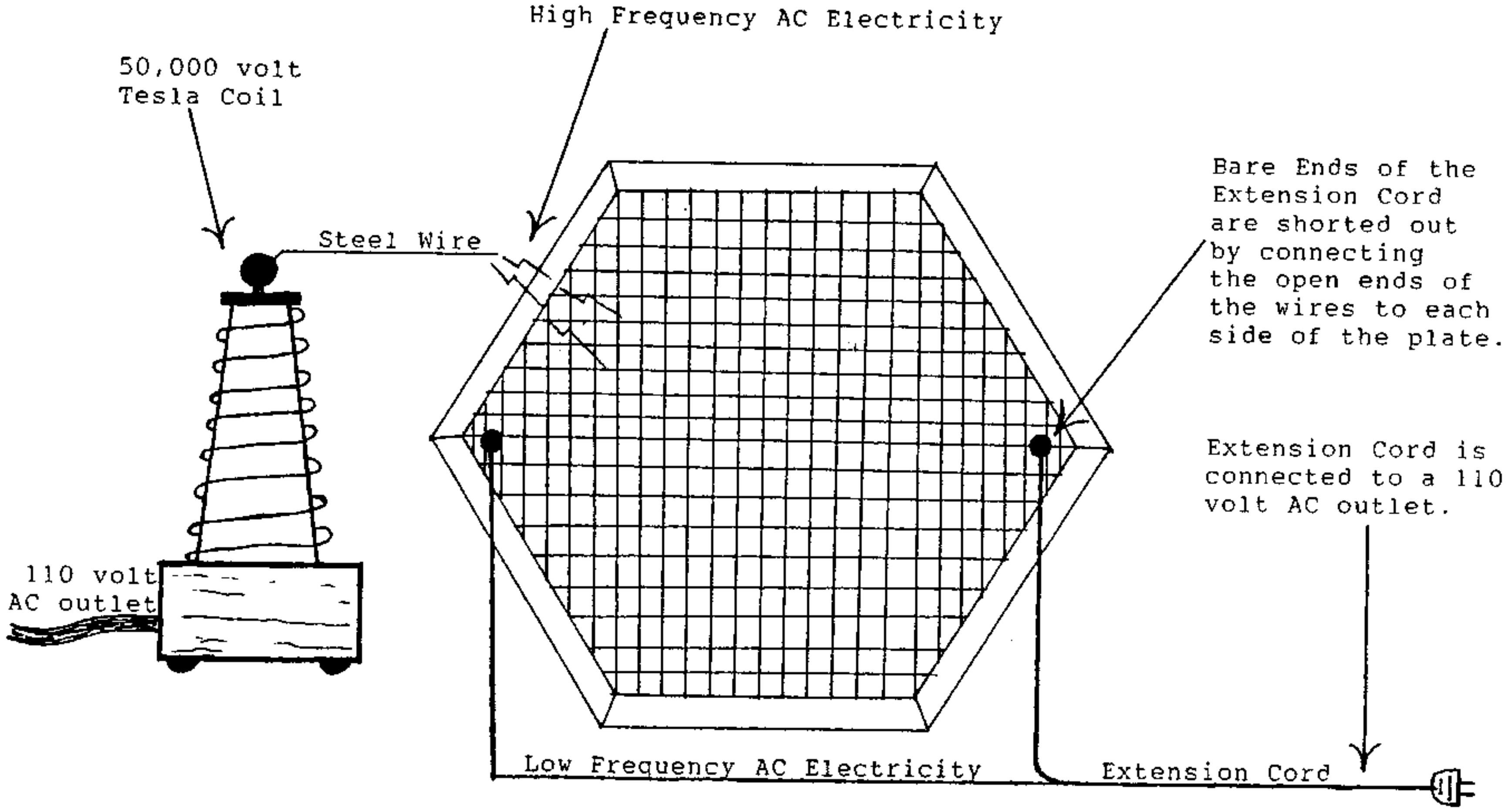
High Frequency AC Electricity
50,000 volt
Tesla Coil
Steel Wire
Bare Ends of the
Extension Cord
are shorted out
by connecting
the open ends of
the wires to each
side of the plate.
110 volt
AC outlet
Extension Cord is
connected to a 110
volt AC outlet.
Low Frequency AC Electricity
Extension Cord

To give you some idea of how all this got started. Back in the year 1981 I was contacted by what I believe to be my other self. The letter which I received at the Sunset Plaza in Norfolk, Nebraska, was dated 1992 A.D. It also had a month on it which I can't seem to remember. Evidently the letter which I received indicates, that sometime around the year 1992, I will travel back into the past to make contact with myself, just like my other self did in the future. One of the things which the letter mentioned which I can barely remember, is that it said: (The path to the truth can be found in the Pyramid of Giza). It also had a riddle, and it went like this: (The riddle can be solved when 79.613 is dissolved). These two verses holds to key to Time Travel. I have already deciphered the riddle. As a matter of fact it was just last year when the answers came to me. Basically the 79.613 number can be converted into the (1.618) value which is used in my Equations. It is also the same number which was used in the construction of the Pyramid of Giza. If we decipher the riddle even further, we will find that it also reveals that the 7.8 Hz. Frequency, is the Frequency to use for accomplishing Time Travel. It seems that whenever a person travels physically through time, his Alpha or Theta Waves are vibrating at this Frequency. Once that your Brain Waves begin to oscillate at this Frequency, your mind then becomes tuned to the Zero Vector. I have found that this is the only way in which Time Travel can be achieved. So basically what I am trying to get at, the Chronological Time

Reflector causes your Brain Waves to vibrate at 7.8 Hz. cycles per second. That is all that these units do. However, for some people, they might think differently, but when you get right down to it, it all leads to the same thing. Mind verses matter. And this just barely covers the subject.

Another experience which I had that occurred around the month of Sept. in 1986, dealt with a variation of the Chronological Time Reflector. In other words, after I had finished programming the plate, I proceeded to Zap the plate with 50,000 volts of High Frequency Electricity. Sometime after the Experiment, I clipped off one end of an Extension Cord and plugged the other end into a 110 volt, 60 cycle, AC outlet. I then laid the bare ends of the wires on top of the screen. Just then I accidentally dropped something on the floor. After I had reached down to pick it up, the bare ends of the Extension Cord which I had laid on top of the Screen, touched the Screen portion of the plate. What was to follow, would be the most terrifying experience I had ever had, because as soon as the bare wires touched the Screen, a massive short occurred. All I can remember is seeing two flashes of white light energy, and the next thing I knew, a white mist had fallen over the entire room. I didn't realize that I had traveled physically through time until I looked at a Calendar the next day. Not only that, but when I asked a lady in Plainview, Nebraska as to what the date was, she said it was the 17th, but my Quartz Watch indicated the 16th. When

I had returned home later on that day, I has discovered that all of my Experiments had been dated wrong. Then I finally realized that I had actually traveled one day into the past. To this day I am not exactly sure as to whether or not I returned back to my own present Time Line. Perhaps I will never know.

Another Time Travel instrument which was sent to me by a man who lives in Osceola, PA., is revealed on pages (12), (13), and (14). According to what I have red in his letter, this device can actually transport objects into the future. However, he says that there is somewhat of a Time delay effect shortly after the unit has been shut off. In other words, after the unit has been de-activated, it takes anywhere from 10 to 15 minutes before the object is teleported. I have never built this device as yet, however I plan to eventually.

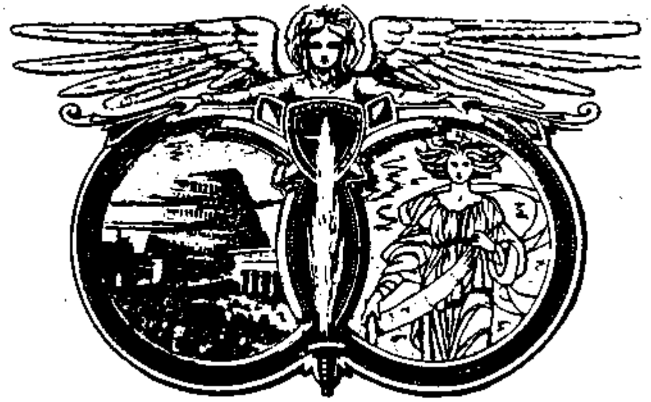

POWER SUPPLY FOR THE TIME PLATFORM

2-22-89

C1 - Is a homemade Capacitor made out of sheets of Glass and Copper Foil. Aluminum may be used, but this could be dangerous to your health.

L1 - Is number 10 Gage wire wound around a dowel rod.

L2 - Is number 24 Gage Magnet Wire wrapped around some PVC pipe. There should be about 1000 turns which should create around 100,000 volts of Low Amperage High Frequency Electricity.

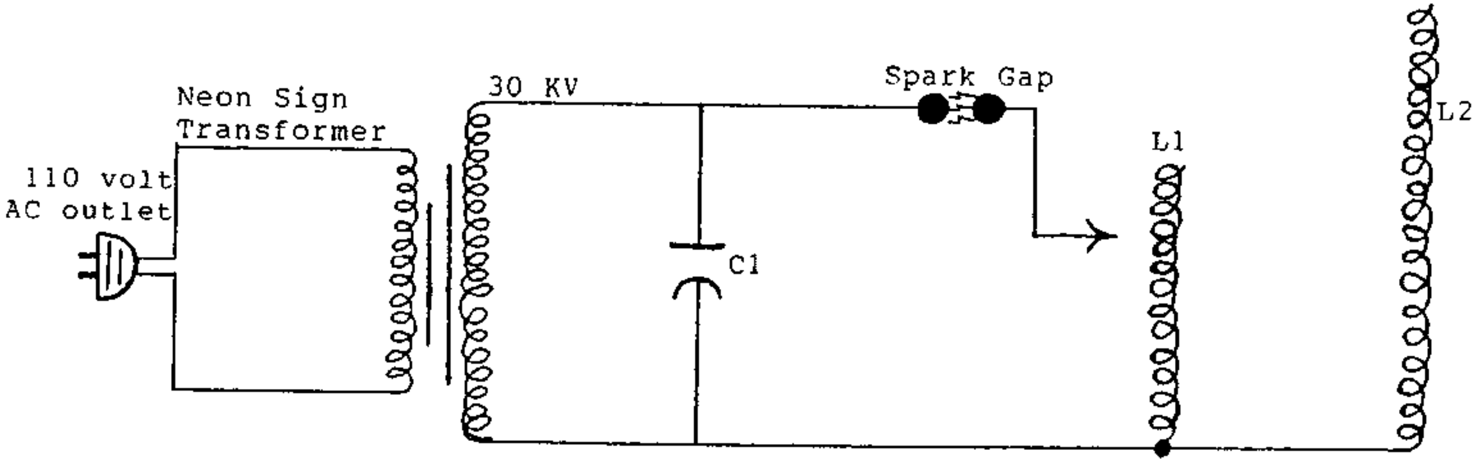
Neon Sign
Transformer
110 volt
AC outlet
30 KV
C1
Spark Gap
L1
L2

THE TIME PLATFORM

2-22-89

The plate is 12" inchs across.

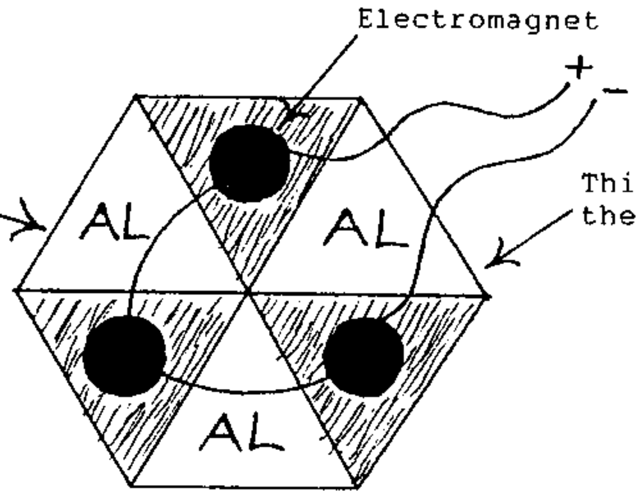

Electromagnet

This is the shape of the sending plate.

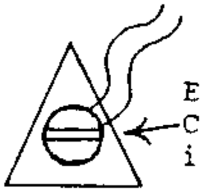

Electromagnetic Coil is placed in the center.

THE ELECTROMAGNETS ARE MADE OF STRAP IRON STACKED 8 INCHES HIGH (1/4"STRAP IRON).

THE ALUMINUM PLATES ARE 1/4 OF AN INCH THICK, AND ARE STACKED 8 INCHS HIGH .

THERE ARE VARIOUS REASONS FOR THIS PARTICULAR ARRAY.

THE ALUMINUM SPACERS CONCENTRATE AND FOCUS THE MAGNETIC FIELDS.

FULL SCALE DRAWING OF THE TIME TRAVEL UNIT

2-23-89

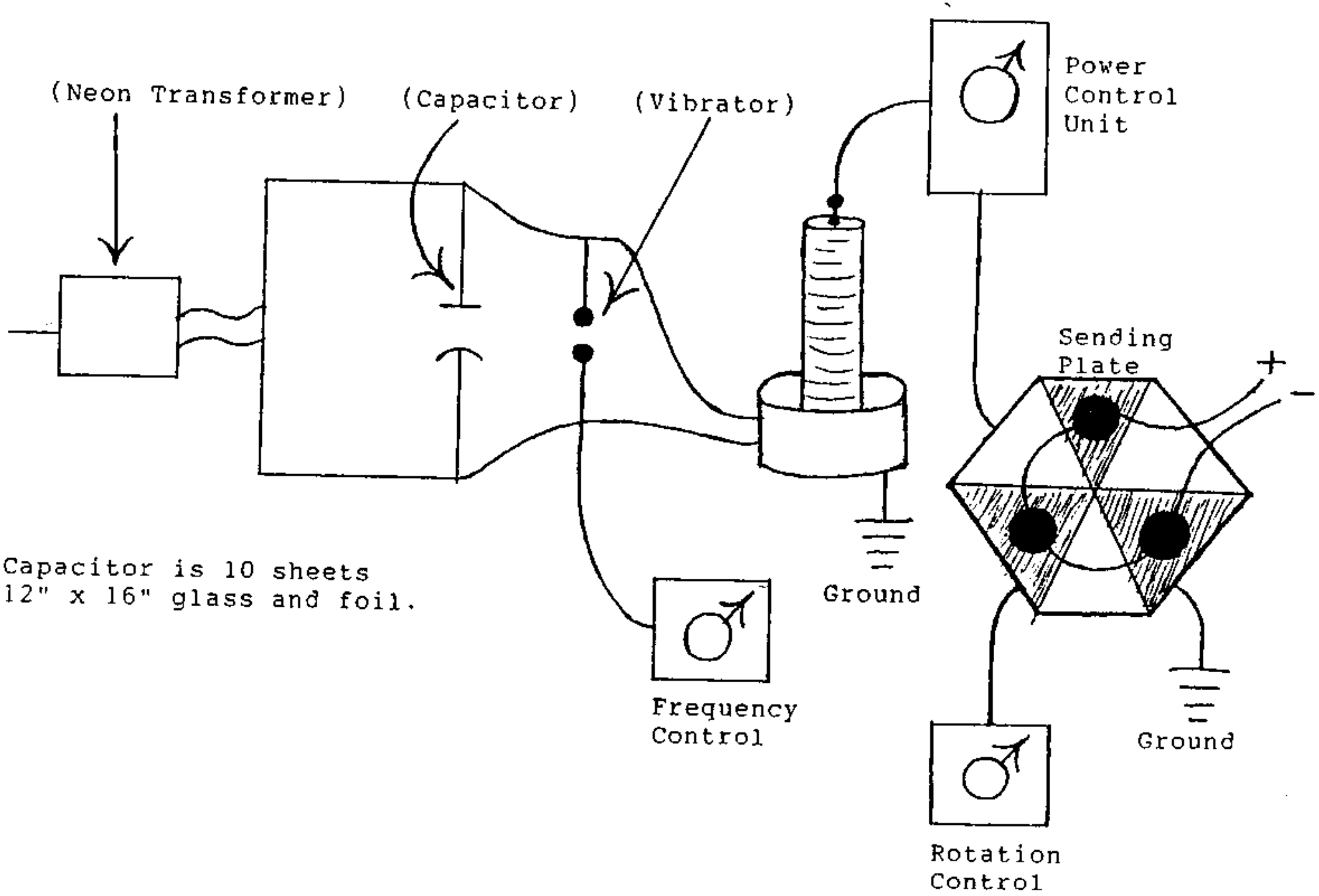
(Neon Transformer)
(Capacitor)
(Vibrator)
Power Control Unit
Sending Plate
+
-
Ground
Frequency Control
Ground
Rotation Control

Capacitor is 10 sheets 12" x 16" glass and foil.

THE DELAWARR EXPERIMENT IN TIME TRAVEL (PHASE #1).

8-16-89

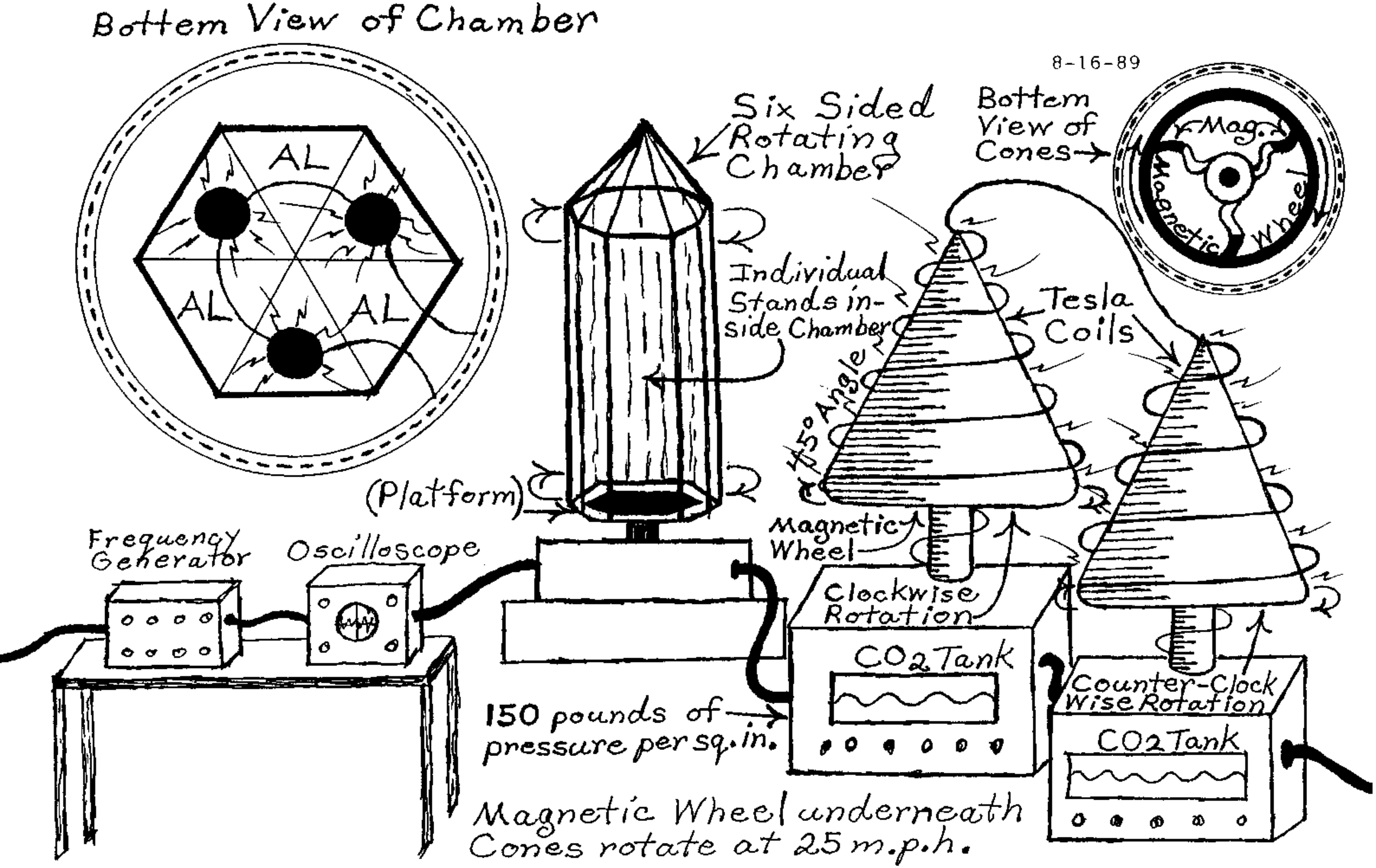
Bottom View of Chamber
AL
AL
AL
Six Sided Rotating Chamber
Bottom View of Cones
Mag.
Magnetic Wheel
Individual Stands inside Chamber
Tesla Coils
45° Angle
(Platform)
Magnetic Wheel
Frequency Generator
Oscilloscope
Clockwise Rotation
CO2 Tank
Counter-Clock Wise Rotation
CO2 Tank
150 pounds of pressure per sq. in.
Magnetic Wheel underneath Cones rotate at 25 m.p.h.

Throughout the vast regions of Space Time itself, there are an infinite number of ways to build a Time Travel Machine. One of these ways is to get a hold of a Rubbing Plate, (GSR), or Pendulum, and proceed by asking (Yes) or (No) questions as to what type of components one should use in order to build a Time Travel Machine. There are a number of ways of doing this. First of all begin by asking what size the instrument box should be for holding the components. Next, determine what type of energy one should use for achieving these effects. Then you ask what type of components one should use in order to transmit this energy. If you have done everything correctly, you will end up creating a device which is based entirely on belief. And sense you have created a device which is built entirely on belief, you can be dam certain that it will work! Especially if you use a (GSR) while asking your questions.

For those of you who do not know what a (GSR) is, this is basically a (Galvanic Skin Response Meter). One of these devices may be purchased from me for $150. Instructions are included. Please allow 4 to 6 weeks for delivery.

THE PARADOXES OF TIME TRAVEL

David Lewis

TIME travel, I maintain, is possible. The paradoxes of time travel are oddities, not impossibilities. They prove only this much, which few would have doubted: that a possible world where time travel took place would be a most strange world, different in fundamental ways from the world we think is ours.

I shall be concerned here with the sort of time travel that is recounted in science fiction. Not all science fiction writers are clear-headed, to be sure, and inconsistent time travel stories have often been written. But some writers have thought the problems through with great care, and their stories are perfectly consistent.[1]

If I can defend the consistency of some science fiction stories of time travel, then I suppose parallel defenses might be given of some controversial physical hypotheses, such as the hypothesis that time is circular or the hypothesis that there are particles that travel faster than light. But I shall not explore these parallels here.

What is time travel? Inevitably, it involves a discrepancy between time and time. Any traveler departs and then arrives at his destination; the time elapsed from departure to arrival (positive, or perhaps zero) is the duration of the journey. But if he is a time traveler, the separation in time between departure and arrival does not equal the duration of his journey. He departs; he travels for an hour, let us say; then he arrives. The time he reaches is not the time one hour after his departure. It is later, if he has traveled toward the future; earlier, if he has traveled toward the past. If he has traveled far toward the past, it is earlier even than his departure. How can it be that the same two events, his departure and his arrival, are separated by two unequal amounts of time?

It is tempting to reply that there must be two independent time dimensions; that for time travel to be possible, time must be not a line but a plane.[2] Then a pair of events may have two unequal separations if they are separated more in one of the time dimensions than in the other. The lives of common people occupy straight diagonal lines across the plane of time, sloping at a rate of exactly one hour of time_1 per hour of time_2. The life of the time traveler occupies a bent path, of varying slope.

On closer inspection, however, this account seems not to give us time travel as we know it from the stories. When the traveler revisits the days of his childhood, will his playmates be there to meet him? No; he has not reached the part of the plane of time where they are. He is no longer separated from them along one of the two dimensions of time, but he is still separated from them along the other. I do not say that two-dimensional time is impossible, or that there is no way to square it with the usual conception of what time travel would be like. Nevertheless I shall say no more about two-dimensional time. Let us set it aside, and see how time travel is possible even in one-dimensional time.

The world—the time traveler's world, or ours—is a four-dimensional manifold of events. Time is one dimension of the four, like the spatial dimensions except that the prevailing laws of nature discriminate between time and the others—or rather, perhaps, between various timelike dimensions and various spacelike dimensions. (Time remains one-dimensional, since no two timelike dimensions are orthogonal.) Enduring things are timelike streaks: wholes composed of temporal parts, or *stages*, located at various times and places. Change is qualitative difference between different stages—different temporal parts—of some enduring thing, just as a "change" in scenery from east to west is a qualitative difference between the eastern and western spatial parts of the landscape. If this paper should change your mind about the

From *American Philosophical Quarterly*, April 1976, pp. 145-152. © 1976 by North American Philosophical Publications. Reprinted by permission.

possibility of time travel, there will be a difference of opinion between two different temporal parts of you, the stage that started reading and the subsequent stage that finishes.

If change is qualitative difference between temporal parts of something, then what doesn't have temporal parts can't change. For instance, numbers can't change; nor can the events of any moment of time, since they cannot be subdivided into dissimilar temporal parts. (We have set aside the case of two-dimensional time, and hence the possibility that an event might be momentary along one time dimension but divisible along the other.) It is essential to distinguish change from "Cambridge change," which can befall anything. Even a number can "change" from being to not being the rate of exchange between pounds and dollars. Even a momentary event can "change" from being a year ago to being a year and a day ago, or from being forgotten to being remembered. But these are not genuine changes. Not just any old reversal in truth value of a time-sensitive sentence about something makes a change in the thing itself.

A time traveler, like anyone else, is a streak through the manifold of space-time, a whole composed of stages located at various times and places. But he is not a streak like other streaks. If he travels toward the past he is a zig-zag streak, doubling back on himself. If he travels toward the future, he is a stretched-out streak. And if he travels either way instantaneously, so that there are no intermediate stages between the stage that departs and the stage that arrives and his journey has zero duration, then he is a broken streak.

I asked how it could be that the same two events were separated by two unequal amounts of time, and I set aside the reply that time might have two independent dimensions. Instead I reply by distinguishing time itself, *external time* as I shall also call it, from the *personal time* of a particular time traveler: roughly, that which is measured by his wristwatch. His journey takes an hour of his personal time, let us say; his wristwatch reads an hour later at arrival than at departure. But the arrival is more than an hour after the departure in external time, if he travels toward the future; or the arrival is before the departure in external time (or less than an hour after), if he travels toward the past.

That is only rough. I do not wish to define personal time operationally, making wristwatches infallible by definition. That which is measured by my own wristwatch often disagrees with external time, yet I am no time traveler; what my misregulated wristwatch measures is neither time itself nor my personal time. Instead of an operational definition, we need a functional definition of personal time; it is that which occupies a certain role in the pattern of events that comprise the time traveler's life. If you take the stages of a common person, they manifest certain regularities with respect to external time. Properties change continuously as you go along, for the most part, and in familiar ways. First come infantile stages. Last come senile ones. Memories accumulate. Food digests. Hair grows. Wristwatch hands move. If you take the stages of a time traveler instead, they do not manifest the common regularities with respect to external time. But there is one way to assign coordinates to the time traveler's stages, and one way only (apart from the arbitrary choice of a zero point), so that the regularities that hold with respect to this assignment match those that commonly hold with respect to external time. With respect to the correct assignment properties change continuously as you go along, for the most part, and in familiar ways. First come infantile stages. Last come senile ones. Memories accumulate. Food digests. Hair grows. Wristwatch hands move. The assignment of coordinates that yields this match is the time traveler's personal time. It isn't really time, but it plays the role in his life that time plays in the life of a common person. It's enough like time so that we can—with due caution—transplant our temporal vocabulary to it in discussing his affairs. We can say without contradiction, as the time traveler prepares to set out, "Soon he will be in the past." We mean that a stage of him is slightly later in his personal time, but much earlier in external time, than the stage of him that is present as we say the sentence.

We may assign locations in the time traveler's personal time not only to his stages themselves but also to the events that go on around him. Soon Caesar will die, long ago; that is, a stage slightly later in the time traveler's personal time than his present stage, but long ago in external time, is simultaneous with Caesar's death. We could even extend the assignment of personal time to events that are not part of the time traveler's life, and not simultaneous with any of his stages. If his funeral in ancient Egypt is separated from his death by three days of external time and his death is separated from his birth by three score years and ten of his personal time, then we may add the two intervals and say that his funeral follows his birth by three score years and ten and three days of *extended personal time.* Likewise a bystander might truly say, three years after the last departure of another famous time traveler, that "he may even now—if I may use the phrase—be wandering on some plesiosaurus-haunted oolitic coral reef, or beside the lonely saline seas of the Triassic Age."[3] If the time traveler does wander on an oolitic coral reef three years after his departure in his personal time, then it is no mistake to say with respect to his extended personal time that the wandering is taking place "even now".

We may liken intervals of external time to distances as the crow flies, and intervals of personal time to distances along a winding path. The time traveler's life is like a mountain railway. The place two miles due east of here may also be nine miles down the line, in the westbound direction. Clearly we are not dealing here with two independent dimensions. Just as distance along the railway is not a fourth spatial dimension, so a time traveler's personal time is not a second dimension of time. How far down the line some place is depends on its location in three-dimensional space, and likewise the locations of events in personal time depend on their locations in one-dimensional external time.

Five miles down the line from here is a place where the line goes under a trestle; two miles further is a place where the line goes over a trestle; these places are one and the same. The trestle by which the line crosses over itself has two different locations along the line, five miles down from here and also seven. In the same way, an event in a time traveler's life may have more than one location in his personal time. If he doubles back toward the past, but not too far, he may be able to talk to himself. The conversation involves two of his stages, separated in his personal time but simultaneous in external time. The location of the conversation in personal time should be the location of the stage involved in it. But there are two such stages; to share the locations of both, the conversation must be assigned two different locations in personal time.

The more we extend the assignment of personal time outwards from the time traveler's stages to the surrounding events, the more will such events acquire multiple locations. It may happen also, as we have already seen, that events that are not simultaneous in external time will be assigned the same location in personal time—or rather, that at least one of the locations of one will be the same as at least one of the locations of the other. So extension must not be carried too far, lest the location of events in extended personal time lose its utility as a means of keeping track of their roles in the time traveler's history.

A time traveler who talks to himself, on the telephone perhaps, looks for all the world like two different people talking to each other. It isn't quite right to say that the whole of him is in two places at once, since neither of the two stages involved in the conversation is the whole of him, or even the whole of the part of him that is located at the (external) time of the conversation. What's true is that he, unlike the rest of us, has two different complete stages located at the same time at different places. What reason have I, then, to regard him as one person and not two? What unites his stages, including the simultaneous ones, into a single person? The problem of personal identity is especially acute if he is the sort of time traveler whose journeys are instantaneous, a broken streak consisting of several unconnected segments. Then the natural way to regard him as more than one person is to take each segment as a different person. No one of them is a time traveler, and the peculiarity of the situation comes to this: all but one of these several people vanish into thin air, all but another one appear out of thin air, and there are remarkable resemblances between one at his appearance and another at his vanishing. Why isn't that at least as good a description as the one I gave, on which the several segments are all parts of one time traveler?

I answer that what unites the stages (or segments) of a time traveler is the same sort of mental, or mostly mental, continuity and connectedness that unites anyone else. The only difference is that whereas a common person is connected and continuous with respect to external time, the time traveler is connected and continuous only with respect to his own personal time. Taking the stages in order, mental (and bodily) change is mostly gradual rather than sudden, and at no point is there sudden change in too many different respects all at once. (We can include position in external time among the respects we keep track of, if we like. It may change discontinuously with respect to personal time if not too much else changes discontinuously along with it.) Moreover, there is not too much change altogether. Plenty of traits and traces last a lifetime. Finally, the connectedness and the continuity are not accidental. They are explicable; and further, they are explained by the fact that the properties of each stage depend causally on those of the stages just before in personal time, the dependence being such as tends to keep things the same.[4]

To see the purpose of my final requirement of causal continuity, let us see how it excludes a case of counterfeit time travel. Fred was created out of thin air, as if in the midst of life; he lived a while, then died. He was created by a demon, and the demon had chosen at random what Fred was to be like at the moment of his creation. Much later someone else, Sam, came to resemble Fred as he was when first created. At the very moment when the resemblance became perfect, the demon destroyed Sam. Fred and Sam together are very much like a single person: a time traveler whose personal time starts at Sam's birth, goes on to Sam's destruction and Fred's creation, and goes on from there to Fred's death. Taken in this order, the stages of Fred-*cum*-Sam have the proper connectedness and continuity. But they lack causal continuity, so Fred-*cum*-Sam is not one person and not a time traveler. Perhaps it was pure coincidence that Fred at his creation and Sam at his destruction were exactly alike; then the connectedness and continuity of Fred-*cum*-Sam across the crucial point are accidental. Perhaps instead the demon remembered what Fred was like, guided Sam toward perfect resemblance, watched his progress, and destroyed him at the right moment. Then the connectedness and continuity of Fred-*cum*-Sam has a causal explanation, but of the wrong sort. Either way, Fred's first stages do not depend causally for their properties on Sam's last stages. So the case of Fred and Sam is rightly disqualified as a case of personal identity and as a case of time travel.

We might expect that when a time traveler visits the past there will be reversals of causation. You may punch his face before he leaves, causing his eye to blacken centuries ago. Indeed, travel into the past necessarily involves reversed causation. For time travel requires personal identity—he who arrives must be the same person who departed. That requires causal continuity, in which causation runs from earlier to later stages in the order of personal time. But the orders of personal and external time disagree at some point, and there we have causation that runs from later to earlier stages in the order of external time. Elsewhere I have given an analysis of causation in terms of chains of counterfactual dependence, and I took care that my analysis would not rule out casual reversal *a priori*.[5] I think I can argue (but not here) that under my analysis the direction of counterfactual dependence and causation is governed by the direction of other *de facto* asymmetries of time. If so, then reversed causation and time travel are not excluded altogether, but can occur only where there are local exceptions to these asymmetries. As I said at the outset, the time traveler's world would be a most strange one.

Stranger still, if there are local—but only local—causal reversals, then there may also be causal loops: closed causal chains in which some of the causal links are normal in direction and others are reversed. (Perhaps there must be loops if there is reversal: I am not sure.) Each event on the loop has a causal explanation, being caused by events elsewhere on the loop. That is not to say that the loop as a whole is caused or explicable. It may not be. Its inexplicability is especially remarkable if it is made up of the sort of causal processes that transmit information. Recall the time traveler who talked to himself. He talked to himself about time travel, and in the course of the conversation his older self told his younger self how to build a time machine. That information was available in no other way. His older self knew how because his younger self had been told and the information had been preserved by the causal processes that constitute recording, storage, and retrieval of memory traces. His younger self knew, after the conversation, because his older self had known and the information had been preserved by the causal processes that constitute telling. But where did the information come from in the first place? Why did the whole affair happen? There is simply no answer. The parts of the loop are explicable, the whole of it is not. Strange! But not impossible, and not too different from inexplicabilities we are already inured to. Almost everyone agrees that God, or the Big Bang, or the entire infinite past of the universe, or the decay of a tritium atom, is uncaused and inexplicable. Then if these are possible, why not also the inexplicable causal loops that arise in the time travel?

I have committed a circularity in order not to talk about too much at once, and this is a good place to set it right. In explaining personal time, I presupposed that we were entitled to regard certain stages as comprising a single person. Then in explaining what united the stages into a single person, I presupposed that we were given a personal time order for them. The proper way to proceed is to define personhood and personal time simultaneously, as follows. Suppose given a pair of an aggregate of persona-stages, regarded as a candidate for personhood, and an assignment of coordinates to those stages, regarded as a candidate for his personal time. If the stages satisfy the conditions given in my circular explanation with respect to the assignment of coordinates, then both candidates succeed: the stages do comprise a person and the assignment is his personal time.

I have argued so far that what goes on in a time travel story may be a possible pattern of events in four-dimensional space-time with no extra time dimension; that it may be correct to regard the scattered stages of the alleged time traveler as comprising a single person; and that we may legitimately assign to those stages and their surroundings a personal time order that disagrees sometimes with their order in external time. Some might concede all this, but protest that the impossibility of time travel is revealed after all when we ask not what the time traveler *does*, but what he *could do*. Could a time traveler change the past? It seems not: the events of a past moment could no more change than numbers could. Yet it seems that he would be as able as anyone to do things that would change the past if he did them. If a time traveler visiting the past both could and couldn't do something that would change it, then there cannot possibly be such a time traveler.

Consider Tim. He detests his grandfather, whose success in the munitions trade built the family fortune that paid for Tim's time machine. Tim would like nothing so much as to kill Grandfather, but alas he is too late. Grandfather died in his bed in 1957, while Tim was a young boy. But when Tim has built his time machine and traveled to 1920, suddenly he realizes that he is not too late after all. He buys a rifle; he spends long hours in target practice; he shadows Grandfather to learn the route of his daily walk to the munitions works; he rents a room along the route; and there he lurks, one winter day in 1921, rifle loaded, hate in his heart, as Grandfather walks closer, closer,. . . .

Tim can kill Grandfather. He has what it takes. Conditions are perfect in every way: the best rifle money could buy, Grandfather an easy target only twenty yards away, not a breeze, door securely locked against intruders. Tim a good shot to begin with and now at the peak of training, and so on. What's to stop him? The forces of logic will not stay his hand! No powerful chaperone stands by to defend the past from interference. (To imagine such a chaperone, as some authors do, is a boring evasion, not needed to make Tim's story consistent.) In short, Tim is as much able to kill Grandfather as anyone ever is to kill anyone. Suppose that down the street another sniper, Tom, lurks waiting for another victim, Grandfather's partner. Tom is not a time traveler, but otherwise he is just like Tim: same make of rifle, same murderous intent, same everything. We can even suppose that Tom, like Tim, believes himself to be a time traveler. Someone has gone to a lot of trouble to deceive Tom into thinking so. There's no doubt that Tom can kill his victim; and Tim has everything going for him that Tom does. By any ordinary standards of ability, Tim can kill Grandfather.

Tim cannot kill Grandfather. Grandfather lived, so to kill him would be to change the past. But the events of a past moment are not subdivisible into temporal parts and therefore cannot change. Either the events of 1921 timelessly do include Tim's killing of Grandfather, or else they timelessly don't. We may be tempted to speak of the "original" 1921 that lies in Tim's personal past, many years before his birth, in which Grandfather lived; and of the "new" 1921 in which Tim now finds himself waiting in ambush to kill Grandfather. But if we do speak so, we merely confer two names on one thing. The events of 1921 are doubly located in Tim's (extended) personal time, like the trestle on the railway, but the "original" 1921 and the "new" 1921 are one and the same. If Tim did not kill Grandfather in the "original" 1921, then if he does kill Grandfather in the "new" 1921, he must both kill and not kill Grandfather in 1921—in the one and only 1921, which is both the "new" and the "original" 1921. It is logically impossible that Tim should change the past by killing Grandfather in 1921. So Tim cannot kill Grandfather.

Not that past moments are special; no more can anyone change the present or the future. Present and future momentary events no more have temporal parts than past ones do. You cannot change a present or future event from what it was originally to what it is after you change it. What you *can* do is to change the present or the future from the unactualized way they would have been without some action of yours to the way they actually are. But that is not an actual change: not a difference between two successive actualities. And Tim can certainly do as much; he changes the past from the unactualized way it would have been without him to the one and only way it actually is. To "change" the past in this way, Tim need not do anything momentous; it is enough just to be there, however unobtrusively.

You know, of course, roughly how the story of Tim must go on if it is to be consistent: he somehow fails. Since Tim didn't kill Grandfather in the "original" 1921, consistency demands that neither does he kill Grandfather in the "new" 1921. Why not? For some commonplace reason. Perhaps some noise distracts him at the last moment, perhaps he misses despite all his target practice, perhaps his nerve fails, perhaps he even feels a pang of unaccustomed mercy. His failure by no means proves that he was not really able to kill Grandfather. We often try and fail to do what we are able to do. Success at some tasks requires not only ability but also luck, and lack of luck is not a temporary lack of ability. Suppose our other sniper, Tom, fails to kill Grandfather's partner for the same reason, whatever it is, that Tim fails to kill Grandfather. It does not follow that Tom was unable to. No more does it follow in Tim's case that he was unable to do what he did not succeed in doing.

We have this seeming contradiction: "*Tim doesn't, but can, because he has what it takes*" versus "*Tim doesn't, and can't, because it's logically impossible to change the past.*" I reply that there is no contradiction. Both conclusions are true, and for the reasons given. They are compatible because "can" is equivocal.

To say that something can happen means that its happening is compossible with certain facts. *Which* facts? That is determined, but sometimes not determined well enough, by context. An ape can't speak a human language—say, Finnish—but I can. Facts about the anatomy and operation of the ape's larynx and nervous system are not compossible with his speaking Finnish. The corresponding facts about my larynx and nervous system are compossible with my speaking Finnish. But don't take me along to Helsinki as your interpreter: I can't speak Finnish. My speaking Finnish is compossible with the facts considered so far, but not with further facts about my lack of training. What I can do, relative to one set of facts, I cannot do, relative to another, more inclusive, set. Whenever the context leaves it open which facts are to count as relevant, it is possible to equivocate about whether I can speak Finnish. It is likewise possible to equivocate about whether it is possible for me to speak Finnish, or whether I am able to, or whether I have the ability or capacity or power or potentiality to. Our many words for much the same thing are little help since they do not seem to correspond to different fixed delineations of the relevant facts.

Tim's killing Grandfather that day in 1921 is compossible with a fairly rich set of facts: the facts about his rifle, his skill and training, the unobstructed line of fire, the locked door and the absence of any chaperone to defend the past, and so on. Indeed it is compossible with all the facts of the sorts we would ordinarily count as relevant is saying what someone can do. It is compossible with all the facts corresponding to those we deem relevant in Tom's case. Relative to these facts, Tim can kill Grandfather. But his killing Grandfather is not compossible with another, more inclusive set of facts. There is the simple fact that Grandfather was not killed. Also there are various other facts about Grandfather's doings after 1921 and their effects: Grandfather begat Father in 1922 and Father begat Tim in 1949. Relative to these facts, Tim cannot kill Grandfather. He can and he can't, but under different delineations of the relevant facts. You can reasonably choose the narrower delineation, and say that he can; or the wider delineation, and say that he can't. But choose. What you mustn't do is waver, say in the same breath that he both can and can't, and then claim that this contradiction proves that time travel is impossible.

Exactly the same goes for Tom's parallel failure. For Tom to kill Grandfather's partner also is compossible with all facts of the sorts we ordinarily count as relevant, but not compossible with a larger set including, for instance, the fact that the intended victim lived until 1934. In Tom's case we are not puzzled. We say without hesitation that he can do it, because we see at once that the facts that are not compossible with his success are facts about the future of the time in question and therefore not the sort of facts we count as relevant in saying what Tom can do.

In Tim's case it is harder to keep track of which facts are relevant. We are accustomed to exclude facts about the future of the time in question, but to include some facts about its past. Our standards do not apply unequivocally to the crucial facts in this special case: Tim's failure, Grandfather's survival, and his subsequent doings. If we have foremost in mind that they lie in the external future of that moment in 1921 when Tim is almost ready to shoot, then we exclude them just as we exclude the parallel facts in Tom's case. But if we have foremost in mind that they precede that moment in Tim's extended personal time, then we tend to include them. To make the latter be foremost in your mind, I chose to tell Tim's story in the order of his personal time, rather than in the order of external time. The fact of Grandfather's survival until 1957 had already been told before I got to the part of the story about Tim lurking in ambush to kill him in 1921. We must decide, if we can, whether to treat these personally past and externally future facts as if they were straightforwardly past or as if they were straightforwardly future.

Fatalists—the best of them—are philosophers who take facts we count as irrelevant in saying what someone can do, disguise them somehow as facts of a different sort that we count as relevant, and thereby argue that we can do less than we think—indeed, that there is nothing at all that we don't do but can. I am not going to vote Republican next fall. The fatalist argues that, strange to say, I not only won't but can't; for my voting Republican is not compossible with the fact that it was true already in the year 1548 that I was not going to vote Republican 428 years later. My rejoinder is that this is a fact, sure enough; however, it is an irrelevant fact about the future masquerading as a relevant fact about the past, and so should be left out of account in saying what, in any ordinary sense, I can do. We are unlikely to be fooled by the fatalist's methods of disguise in this case, or other ordinary cases. But in cases of time travel, precognition, or the like, we're on less familiar ground, so it may take less of a disguise to fool us. Also, new methods of disguise are available, thanks to the device of personal time.

Here's another bit of fatalist trickery. Tim, as he lurks, already knows that he will fail. At least he has the wherewithal to know it if he thinks, he knows it implicitly. For he remembers that Grandfather was alive when he was a boy, he knows that those who are killed are thereafter not alive, he knows (let us suppose) that he is a time traveler who has reached the same 1921 that lies in his personal past, and he ought to understand—as we do—why a time traveler cannot change the past. What is known cannot be false. So his success is not only not compossible with facts that belong to the external future and his personal past, but also is not compossible with the present fact of his knowledge that he will fail. I reply that the fact of his foreknowledge, at the moment while he waits to shoot, is not a fact entirely about that moment. It may be divided into two parts. There is the fact that he then believes (perhaps only implicitly) that he will fail; and there is the further fact that his belief is correct, and correct not at all by accident, and hence qualifies as an item of knowledge. It is only the latter fact that is not compossible with his success, but it is only the former that is entirely about the moment in question. In calling Tim's state at that moment knowledge, not just belief, facts about personally earlier but externally later moments were smuggled into consideration.

I have argued that Tim's case and Tom's are alike, except that in Tim's case we are more tempted than usual—and with reason—to opt for a semi-fatalist mode of speech. But perhaps they differ in another way. In Tom's case, we can expect a perfectly consistent answer to the counterfactual question: what if Tom had killed Grandfather's partner? Tim's case is more difficult. If Tim had killed Grandfather, it seems offhand that contradictions would have been true. The killing both would and wouldn't have occurred. No Grandfather, no Father; no Father, no Tim; no Tim, no killing. And for good measure: no Grandfather, no family fortune; no fortune, no time machine; no time machine, no killing. So the supposition that Tim killed Grandfather seems impossible in more than the semi-fatalistic sense already granted.

If you suppose Tim to kill Grandfather and hold all the rest of his story fixed, of course you get a contradiction. But likewise if you suppose Tom to kill Grandfather's partner and hold the rest of his story fixed—including the part that told of his failure—you get a contradiction. If you make any counterfactual supposition and hold all else fixed you get a contradiction. The thing to do is rather to make the counterfactual supposition and hold all else as close to fixed as you consistently can. That procedure will yield perfectly consistent answers to the question: what if Tim had not killed Grandfather? In that case, some of the story I told would not have been true. Perhaps Tim might have been the time-traveling grandson of someone else. Perhaps he might have been the grandson of a man killed in 1921 and miraculously resurrected. Perhaps he might have been not a time traveler at all, but rather someone created out of nothing in 1920 equipped with false memories of a personal past that never was. It is hard to say what is the least revision of Tim's story to make it true that Tim kills Grandfather, but certainly the contradictory story in which the killing both does and doesn't occur is not the least revision. Hence it is false (according to the unrevised story) that if Tim had killed Grandfather then contradictions would have been true.

What difference would it make if Tim travels in branching time? Suppose that at the possible world of Tim's story the space-time manifold branches; the branches are separated not in time, and not in space, but in some other way. Tim travels not only in time but also from one branch to another. In one branch Tim is absent from the events of 1921; Grandfather lives; Tim is born, grows up, and vanishes in his time machine. The other branch diverges from the first when Tim turns up in 1920; there Tim kills Grandfather and Grandfather leaves no descendants and no fortune; the events of the two branches differ more and more from that time on. Certainly this is a consistent story; it is a story in which Grandfather both is and isn't killed in 1921 (in the different branches); and it is a story in which Tim, by killing Grandfather, succeeds in preventing his own birth (in one of the branches). But it is not a story in which Tim's killing of Grandfather both does occur and doesn't: it simply does, though it is located in one branch and not the other. And it is not a story in which Tim changes the past. 1921 and later years contain the events of both branches, coexisting somehow without interaction. It remains true at all the personal times of Tim's life, even after the killing, that Grandfather lives in one branch and dies in the other.[6]

Princeton University

Received September 4, 1975

Notes

1. I have particularly in mind two of the time travel stories of Robert A. Heinlein: "By His Bootstraps" in R. A. Heinlein, *The Menace from Earth* (Hicksville, N.Y., 1959), and "—All You Zombies—," in R. A. Heinlein, *The Unpleasant Profession of Jonathan Hoag* (Hicksville, N.Y., 1959).
2. Account of time travel in two-dimensional time are found in Jack W. Meiland, "A Two-Dimensional Passage Model of Time for Time Travel," *Philosophical Studies,* vol. 26 (1974), pp. 153–173; and in the initial chapters of Isaac Asimov, *The End of Eternity* (Garden City, N.Y., 1955). Asimov's denouement, however, seems to require some different conception of time travel.
3. H. G. Wells, *The Time Machine, An Invention* (London, 1895), epilogue. The passage is criticized as contradictory in Donald C. Williams, "The Myth of Passage," *The Journal of Philosophy,* vol. 48 (1951), p. 463.
4. I discuss the relation between personal identity and mental connectedness and continuity at greater length in "Survival and Identity" in *The Identity of Persons,* ed. by Amelie Rorty (forthcoming).
5. "Causation," *The Journal of Philosophy,* vol. 70 (1973), pp. 556–567; the analysis relies on the analysis of counterfactuals given in my *Counterfactuals* (Oxford, 1973).
6. The present paper summarizes a series of lectures of the same title, given as the Gavin David Young Lectures in Philosophy at the University of Adelaide in July, 1971. I thank the Australian-American Educational Foundation and the American Council of Learned Societies for research support. I am grateful to many friends for comments on earlier versions of this paper; especially Philip Kitcher, William Newton-Smith, J. J. C. Smart, and Donald Williams.

The Tom Bearden
Website

Return to Reference Section

ANNOTATED GLOSSARY

T. E. Bearden 1997, slightly updated Dec. 2000
Latest editing and additions: 12th April, 2003
Presentation re-edit 7th January, 2013

Foreword

This is an informal glossary originally prepared in 1997 for Gary Hawkins to use on his web site, and informal definitions and discussions were rather quickly jotted down in 1997. We have been able to do a little updating now in Dec. 2000. Careful crafting and polishing for greater rigor would require much more time than I am able to spare, but perhaps the glossary can be helpful anyway.

We put in the gist of it, as we see it, and usually an annotation or commentary to explain it. We tried to put in sufficient explanations and a few references, so that the reader can see for himself or herself that physics, mathematics, and even logic itself are not quite as cut and dried as is often presented!

The only thing we absolutely insist on, that differs from the normal physics and engineering approach, is that any true definition must be an identity statement *a priori*. In this view, since no equation is an identity, then no equation is a definition; regardless of those "equations" universally advanced in physics as "definitions." All an equation states is that the *magnitude* of the somethings on the right is equal to the *magnitude* of the somethings on the left. It does not at all *define* any of the somethings on either right or left! I've completed as much as I have time for now, and made some additions and corrections to the 1997 version. As time permits, we will try to add some more in the future. I hope these hasty comments are helpful, and any errors in them are my own and not due to Gary Hawkins or to Tony Craddock!

Tom Bearden,
Dec. 2000

Webmaster's Note: There are additional Glossaries and Definitions at the ends of Tom Bearden's books *"Energy from the Vacuum"* and *"Excalibur Briefing."*

MATH AND SINGLE CHARACTER ABBREVIATIONS

ϕ The electrostatic scalar potential, in units of joule per collecting coulomb.

∇ The mathematical operator 'del'.

$\nabla\phi$ The gradient of the function ϕ , mathematically.

In EM theory, use of the del operator produces an electric field **E** given by $\mathbf{E} = -\nabla\phi$. An electrical field is also generated by the time rate of change of the magnetic vector potential **A**, where it is given by $\mathbf{E} = -d\mathbf{A}/dt$. Note that we have used that fact to assist us in producing the motionless electromagnetic generator (MEG), with COP = 16 as of this writing.

ϕq Joules of energy collected by a charge of q coulombs in an electrostatic scalar potential

having a reaction cross section of ϕ.

Since the collected energy is also usually the energy that can then be dissipated as work, this expression often yields the work performed in the circuit by the energy the surface charges collected from that part of the **S**-flow in contact with the conductors and components.

$d\phi/dt$ Time rate of change of the scalar potential.

Also one form of displacement current.

dE/dt Time rate of change of the E-field.

Also one form of displacement current.

dp/dt Time rate of change of momentum, which is a force F acting upon the system.

A force is defined as a time rate of change of momentum, or $F = dp/dt$.

dP/dt Time rate of change of polarization.

Polarization displacement current.

dq/dt Time rate of change of charge, otherwise known as "current".

Represents the normal current flow i of charges in a conductor, i.e., $i = dq/dt$.

dq/dt-Isolated Load Loop A special current loop containing the load,

...where $i = dq/dt$ in that "load loop" is isolated from the current $i = dq/dt$ in the source loop containing the power source, and the two loops are also field-coupling isolated. A transformer, e.g., does isolate the actual current in the secondary from the current in the primary, but the field coupling from primary to secondary and from secondary back to primary result in the enforcement of COP = 1.0 as the maximum performance possible (in a perfect, lossless standard transformer). In a real standard transformer with some losses, COP<1.0.

The notion is that, even though current cannot pass from the source loop to the load loop and there is no direct field coupling, energy density flow $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ can pass. This is in fact the definition of what we call *bridging*, if in addition to current isolation the rigid back-field coupling is broken between energy dissipation (power) in the source loop and energy dissipation (power) in the load loop. In that fashion, overunity operation of the load can readily be achieved.

E Electric Field.

Defined in older classical electromagnetics (CEM) as a force field, by **E** ≡ **F**/q. However, this only defines the entity after an interaction with mass. In short, it defines the E-field as an effect existing only in matter.

In the early days of electrodynamics, Maxwell and the other "founding fathers" assumed a material ether filling all space. Hence any field in space was indeed such an "effect" field existing in mass. Maxwell died in 1879 of stomach cancer. In the 1880s, the Michelson-Morley experiments falsified the luminiferous (material) ether as a material medium. Hence the accepted use of the **E** = **F**/q type of "effect" electrical field as the electric field in space was also destroyed, but the equations were never changed to eliminate this error in the foundations of CEM. Instead, electrodynamicists more or less announced one day that, "Since there is no (material) ether, we are not using one!"

There is some rationalization mathematically for such an approach, if the limiting process is used for the interaction of the "causal field" in vacuum with the charged mass q, as q is allowed to approach zero. However, from a foundations view, it is still the interaction of the field in space with charged mass that is being described in the limit, and therefore it is still the "effect" entity rather than the "cause" entity.

This serious foundations problem is still unresolved in much of CEM theory and modeling to this day. The standard by which other EM texts are measured is J. D. Jackson, **Classical Electrodynamics**, 2nd edition, Wiley, New York, 1975. In several places, Jackson opts for this avoidance of the use of the " effect" field for the "cause" field: quoting p. 28: *"...the thing that eventually gets measured is a force..." "At the moment, the electric field can be defined as the force per unit charge acting at a given point. It is a vector function of position, denoted by E."* Quoting p. 249: *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."*

The problem is further increased by use of this flawed definition of E-field to "define" the electrostatic scalar potential (Jackson, p. 34) by the equation $\mathbf{E} = -\nabla\phi$. First, no equation is a definition; it merely equates the magnitude of the entities on the left of the = sign to the magnitude of the entities on the right of the = sign. It has nothing at all to say about the definition or nature of any of the entities, but is merely a comparison of relative magnitudes. Any definition, a priori, must be expressed by an identity sign. In all of classical electrodynamics, there is no available proper definition of (x) in the identity E ≡ (x). Indeed, there is no such identity available for force F. To quote Feynman: *"...in dealing with force the tacit assumption is always made that the force is equal to zero unless some physical body is present… One of the most important characteristics of force is that it has a material origin, and this is not just a definition. ... If you insist upon a precise definition of force, you will never get it!"* Richard P. Feynman, Robert B. Leighton, and Matthew Sands, **Lectures on Physics**, Addison-Wesley, Reading, MA, Vol. 1, 1964, p. 12-2.

In the Sachs unified field theory, O(3) electrodynamics as spearheaded by Evans is an important subset. Hence in that approach there is a solution to the problem of defining force and field. The EM fields (**E**, **B**, etc.) when in spacetime, are in fact curvatures of spacetime identically. When these ST curvatures interact with a charged mass, the effects of the interactions are the well-known force-fields in matter, which Maxwell started with in the 1860s. So in this way, the definition problem can be rigorously resolved.

$j\phi$ The Slepian vector, representing the time rate of energy density dissipation in the current

loop containing the $j\phi$.

This does not at all represent the energy flow in space and along the outside of the circuit, which may be very, very much greater, on the order of 10^{13} times as great. It does match the Poynting energy flow component that enters the circuit.

M_q or m_q The mass of a charge q.

Instead of just using "q" and assuming it to be unitary, we must include the fact that charge is a system of two coupled components: (i) the mass m_q of the charge q, and (ii) the virtual photon flux exchange (i.e., ϕ_q) of the charge q with the surrounding active vacuum. In particle physics it is already well-known that there is no symmetry of a mass system anyway, unless the vacuum's exchange with the mass is included. So the expression $m_q\phi_q$ captures the vacuum-mass exchange. By continuing to assume that charge is unitary, physics inadvertently holds to the hoary old notion of the luminiferous ether (thin material ether) that permeates all of space. Maxwell and other electrical physicists of his time believed in a material ether, and Maxwell embedded that notion in EM theory. Inexplicably it has not been rooted out of the equations to this day, even though the notion of the material ether (not the ether per se!) was destroyed by the Michelson-Morley experiment over 100 years ago. Foundations physicists continue to point out such terrible defects in physics—and continue to be ignored. We quote Mario Bunge, "A mathematical theory of the dimensions and units of physical quantities," in **Problems in the Foundations of Physics**, edited by Mario Bunge, Springer-Verlag, Berlin and New York, 1971, p. 7: "The question 'What are dimensions?' is seen to be ill conceived, for there are two concepts of dimension, namely those of dimension function and dimension value. The correct question is: 'What are [] and *d*?' " {**Note:** [] is dimension function, and *d* is dimension value.} Bunge is clearly discriminating between the function associated with the dimension, and the magnitude associated with it

Quoting Bunge, *ibid*., p. 8: "The concept of a unit is one of those scientific notions that has remained obscure for want of a theory and excess of a coarse philosophy. Indeed we know how to manipulate units but we do not seem to know what kind of mathematical object they are. As a consequence units are sometimes confused with dimensions, at other times with standards, and they are often introduced in relation to measurement rather than to concept formulation."

And quoting Mario Bunge, **Foundations of Physics**, Springer-Verlag, New York, 1967, p. 176: "Different theories of CED [Classical Electrodynamics] are called for, both for charged particles that can to a first approximation be regarded as electromagnetically structureless (e.g., the electron and the μ -meson) and for extended systems with a charge distribution and spin (e.g., the proton). If fairly satisfactory classical theories were available they could guide the construction of the corresponding quantum theories. But no such theories are in sight partly because it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state."

Again, Bunge, *ibid*., p. 182: "...the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

c The speed of light in ambient vacuum.

To discuss the speed of light, we point out that the vacuum is highly energetic, as has been shown in modern physics. Since the vacuum has an energy density, it can be modeled as a "potential" and particularly as a scalar EM potential for EM purposes. Contrary to convention, we recognize that the local energy density of the ambient vacuum potential can vary, including seasonally within and close to the solar system, etc. The average energy density of the vacuum in space far from charged masses such as stars and planets may be assumed as a "standard vacuum potential", and if this potential is equated with the notion of "flat spacetime", then an increase in this vacuum potential is an increase in the local energy density of spacetime, and hence represents a curvature of the standard flat spacetime.

In fact the interplanetary radar probes' "ranges" do show precisely such variation in the speed of light, with weekly and monthly and seasonal variations, etc. In order to comply with the nearly universal assumption in physics with respect to the constancy of c, the scientific staffs averaged the readings over the entire month. This zeroed out the variations, upheld the notion that space is a fixed type of thing, and that the speed of light in it is also a rigidly fixed constant. It *does not* agree with the actual measured speed of light from day to day.

The Russians have known and openly printed for about a century that the speed of light in a hard vacuum in deep space, far removed from solar and planetary masses, is a little slower than the speed of light in a hard vacuum near the Earth or any other large charged mass. The huge collection of charged particles in the earth have tiny separations, so that looking at the Earth as a collection of tiny dipoles, one sees a collection of enormous 4-symmetry energy flows from the time domain to the 3-space domain (and vice versa for the time-reversed portions). The increased intensity of this 4-symmetry energy flux in the vicinity of a massive planet or sun changes the local energy density of the vacuum itself, increasing it in that vicinity.

Without elaboration, we will also point out that the Earth and its dynamics structures the local spacetime and the spacetime out to some distance. So do all other planets, stars, etc. By whatever manner we model the cosmos as having been created or initiated, the dynamics and structuring of all the phenomena occurring in the universe is further impressed in the *infolded* electrodynamics inside the vacuum potential.

The "laws of nature" in one sense represent the most general results of the reactions of this internally structured "common standard spacetime" or "standard vacuum potential" with the various physical systems, devices, and measuring instruments utilized in our experiments, etc. If one deterministically alters the *structuring of the local spacetime*, one can have a system which will function there in a certain marvelous manner, but will not function that way elsewhere. Some decades ago, I participated in experiments by Golden which over a period of some five years did condition (dimension) the local vacuum potential. The result was an engine which consistently exhibited COP = 1.67 at that location, but when moved several hundred miles exhibited COP<1.0. Frank and I were made aware of this "vacuum conditioning" effect by Prof. William Tiller, who had himself conducted experiments exhibiting just such an effect over several years.

EM waves in vacuum are actually *longitudinal* (sound-type) waves, but we do not belabor the point since everything out there in the normal literature uses the transverse EM wave model for the EM wave in space. Also, the transverse wave is indeed what is detected in the interacting electron gas in our measurement instruments—whose Drude electrons are restrained to simple drift velocity longitudinally down the conductor, but are spinning and are free to precess like gyros. So our instruments detect the Drude electron precessions at right angles to the incoming disturbing longitudinal force field, which means that they do indeed measure a transverse wave. However, these are actually the *electron precession waves* in the detector, and *not* the actual EM waves in space. See my paper, "Giant Negentropy from the Common Dipole", **Journal of New Energy**, 5(1), Summer 2000, p. 11-23 for a discussion of how all EM energy comes from the time domain, via the broken 3-symmetry of the source charge or source dipole, and is emitted in 3-space in all directions as longitudinal EM waves. This is based on rigorous demonstration by E.T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," **Math. Ann**., Vol. 57, 1903, p. 333-355 that the scalar potential is a harmonic set of phase conjugate longitudinal EM wave pairs. We merely applied the Whittaker decomposition to the scalar potential of a source dipole or of a source charge, and the rigorous proof that EM waves in space are longitudinal directly emerges.

Just as sound speeds up, say, in steel as a denser medium than air, the EM "sound waves" in the ambient vacuum potential speed up when that vacuum potential's flux density is increased. Consequently light does move a little faster in a hard vacuum on Earth than it does in deep space, because the vacuum itself has higher energy density near the earth than in deep space. The comings and goings and orbitings of planets and the solar mass, etc. do affect the local energy density of the vacuum between the planets, and hence the speed of light through that interplanetary space.

So the interplanetary radar does show the periodic variations of the local energy density of vacuum that are induced by these factors. For a critical paper on this, see Bryan G. Wallace, "The Unified Quantum Electrodynamic Ether," **Foundations of Physics**, 3(3), Sept. 1973, p. 381-388. His discussion of the interplanetary radar measurements of Venus, which did not fit relativity, are particularly interesting, with respect to his finding that the data were simply "averaged" to eliminate the variations and hence the diurnal, lunar, and synodic periodic components actually measured in the variations of the speed of light. See particularly Bryan G. Wallace, "The great speed of light in space cover-up." **Scientific Ethics**, 1(1), Feb. 1985. p. 2-3; — "Radar Testing of the Relative Velocity of Light in Space," **Spectrosc. Lett**. Vol. 2, 1969. p. 361.

Rigorously, the field is "defined" in electrodynamics as the effect of an interaction of the force-free entity (4-spatial!) existing prior to the interaction, with a 3-dimensional static charge. This means that the field as originally defined is an *effect* and not a *cause*. To then use it as the cause is a grave non sequitur.

This confusion of cause and effect is widespread in physics. It was probably initiated by the hoary old assumption several hundred years ago that a separate force acts upon a separate mass, which is false. In fact mass is a *component* of force, since $\mathbf{F} \equiv d/dt(m\mathbf{v}) = m(d\mathbf{v}/dt + \mathbf{v}(dm/dt)$. As can be seen, in both expanded terms there exists a mass term. To skirt the issue, electrodynamicists have simply found a neat statement that "in massfree space the field continues to exist, but the force is zero." We point out that $m = 0$ makes the momentum vanish, and so **F** vanishes, and in the definition of $\mathbf{E} = \mathbf{F}/q$ also vanishes because the numerator **F** becomes zero. Therefore *that* **E**-field vanishes.

Feynman and Wheeler pointed out that the field *as it is defined* does not and cannot exist in massless space. Only the "potential" for the field exists there, in case some charged mass is brought in to interact with that entity that does exist there. That is accurate, since spacetime is totally a potential and any "change" in spacetime is a change in that potential *a priori*. In the Sachs-Evans unified field theory, this problem is resolved because the fundamental field is defined as a spacetime curvature (change in spacetime, hence change in the ST potential) rather than $\mathbf{E} = \mathbf{F}/q$.

In empty space there are always myriads upon myriads of interfering fields and potentials, from charges all over the universe, via Puthoff's cosmological feedback principle [see
H. E. Puthoff, "Source of Vacuum Electromagnetic Zero-point Energy," **Physical Review A**, 40(9), Nov. 1, 1989, p. 4857-4862.] Thus any EM wave traveling through space is always involved in enormously active interferometry. This alone results in violent momentary variations in the speed of that wave, even from 0.01 c to 100.00 c. For a paper showing this kind of "galloping" travel of an EM wave, see William G. Harter, John Evans, Roberto Vega, and Sanford Wilson, "Galloping waves and their relativistic properties," **American Journal of Physics**, 53(7), July 1985, p. 671-679. While the average speed of the wave is $c(\gamma)$, where γ is the local energy density of the vacuum, the instantaneous speeds of the wave vary violently from instant to instant.

In special relativity, "velocity" merely is a rotation angle of the moving frame from the fixed laboratory observer's frame. The speed "c" merely represents a full right-angle turn away, toward the time-dimension (since other than the 3 spatial dimensions that is the only other dimension available to turn toward). In higher dimensional space, there are many other "higher" dimensions toward which further rotation can be accomplished. Hence rotating in n-space is not limited to a single orthogonal turn, and therefore velocities are not limited to speed c. However, each orthogonal rotation also subtracts one spatial dimension of the rotated object from the observer's 3-space. One rotation would be speed c, two departing rotations would be speed c^2, three would give c^3, etc. At three orthorotations, to the lab observer the rotated object is now purely an internal structure inside his time dimension, very similar to a mental object (mind is time-like, not spacelike). The nature of mind and its operations has a rather unexpected connection with matter, when one considers such orthorotations in hyperspace.

The point is, the speed *c* is not an "iron dictum" of nature, nor is it an "unchangeable constant"! Nature herself routinely violates that notion. As an example, de Broglie waves from any mass always move at v>c, and easily as v>>c. EM "surface" waves—so-called transverse oscillations or upheavals and depressions in the magnitude (energy density; actually the *reaction cross section*) of the local ambient vacuum potential), move at a *discretized* average speed c(γ). The longitudinal component, if freed from the transverse component, can move much faster because it is a sort of "tunneling" movement. In actual quantum tunneling experiments, e.g., music (Mozart's 40th symphony) has been transmitted through quantum mechanical tunneling between two separated points at speed v>4c. Fogal is likely to eventually place a superluminal communication system on the commercial market. We discuss superluminal communication elsewhere in this glossary.

C Capacitance. Essentially, the charge per unit potential.

For an isolated conductor, C = Q/V and C is expressed in farads. A farad is one coulomb per volt.

A lower case c represents the speed of light in ambient vacuum.

E The electric field. Essentially, E =F/q, but that is not really a definition.

Conventionally, the units of electric field are considered to be force per unit electric charge; i.e., force per coulomb. Note that the **E**-field even in this "classical definition" does not actually exist at a point unless there is a unit point charge (a point coulomb of charged mass) located at the point. Hence contrary to classical electrodynamics, communications theory, and whatnot, no electric force field (and no magnetic force field either) exists in the vacuum, at least of the kind prescribed by Maxwell's equations and by the Heaviside-Hertz reduction of Maxwell's equations. Since Maxwell assumed it, these equations and models all still erroneously assume a material ether filling the vacuum, in which case there *would* be matter and charges, and there would *then* be force-fields in the vacuum. We refer the interested reader to discussions by Feynman and by Wheeler. Quoting: "*One of the most important characteristics of force is that it has a material origin, and this is not just a definition. Y If you insist upon a precise definition of force, you will never get it!* " [Richard P. Feynman, Robert B. Leighton, and Matthew Sands, Lectures on Physics, Addison-Wesley, Reading, MA, Vol. 1, 1964, p. 12-2.]

Even worse, note that **E** = **F**/q defines not the magnitude of the field in space, but has to do with the divergence of energy from the field around a unit point static charge there in space, as well as the pattern of the diverged energy density. Actually, what is used as "the electric field" itself is the *reaction cross section* of the field, for a unit point static coulomb of charge. That is at best a function of the "field density" or "field intensity" of the actual electrical field in spacetime, and the direction and pattern of change of that field density due to the intercepting and diverting charge.

The field problem is resolved in the forefronts of electrodynamics by making the field a topological function. It is resolved in the Sachs-Evans O(3) electrodynamics (as a subset of Sachs' unified field theory) by making the field a function of the curvature of spacetime.

See also the discussion under c, above.

H Magnetic field strength.

B The B-field (loosely), or magnetic flux density (more properly).

Conventionally, this is thought to be defined, analogously to **E**, by the force it exerts. The force exerted on a point charge moving in the magnetic field **B** is at right angles to the velocity vector of the charge and also at right angles to **B**. Actually, the **H**-field was originally named the *magnetic field* and the **B**-field was named the *magnetic induction* (and is still sometimes called that.) This is an annoyance because it is **B** which involves the force, and hence **B** should be *named* the magnetic field. Other writers have noted this discrepancy, and Sommerfeld called **H** the magnetic excitation—which correctly describes its role—rather than calling it the magnetic field. As the reader can see, lots of things in electrical physics are *still very untidy*! There are also different ways to approach the EM modeling problem. In Kaluza-Klein unified field theory, all electromagnetics is modeled in the 5th dimension, which is wrapped around each and every point in our normal 3-dimensional space.

For practical use, a far better approach is the O(3) symmetry electrodynamics spearheaded by Myron Evans, which is also an important subset of Mendel Sachs' unified field theory. In the Sachs-Evans unified theory, electromagnetics has become general relativity and general relativity has become electromagnetics. Further, very powerful general relativity can now be directly engineered with this O(3) electrodynamics, as it further develops. We refer the reader to "The New Maxwell Electrodynamic Equations: New Tools for New Technologies. A Collection of 60 Papers from the Alpha Foundation's Institute for Advanced Study," Special Issue of **Journal of New Energy**, 4(3), Winter 1999, 335 p. Also, particularly see **Contemporary Optics and Electrodynamics**, edited by Myron W. Evans, a special topical issue of I. Prigogine and S. A. Rice (series editors), **Advances in Chemical Physics**, Wiley, 2001 (in publication). Evans gives a very long and detailed presentation, "O(3) Electrodynamics", in that publication, and also gives the paper "The Link Between the Sachs and O(3) Theories of Electrodynamics".

Q A quantity of electrical charge. Note that this statement *does not* define charge *q* itself.

We have tentatively proposed to redefine charge q as $q \equiv m_q \phi_q$. The term ϕ_q captures the vacuum's violent virtual energy interaction with the mass m_q of the charge, which is presently missing in classical EM theory and is an exceeding great flaw of omission. Further, ϕ_q may be decomposed into a harmonic series of phase conjugate wavepairs, as shown by Whittaker, 1903, *ibid*. [In O(3) electrodynamics, the internal structuring is dramatically extended, so that one can get at the "inner electrodynamics" necessary to explain such things as homeopathy, chi, ki, prana, etc.

To those scientists pooh-poohing such things as acupuncture, we point out that acupuncture in a dramatic case can allow a person without anesthetic to undergo a major operation with chest cavity opened, while the patient remains alert, conscious, and talking. President Nixon's physician on Nixon's trip to China witnessed such an actual operation, as well as other things. Nothing in our vaunted Western medical science can equal that, nor can any of our medical schools, nor can our National Academy of Science or our National Science Foundation. In short, acupuncture alone demonstrates how woefully inadequate our medical science actually is.

Also, acupuncture works on animals as well, so it does not involve hypnosis or deep suggestion. In science, we must believe the repeatable experiment, not our dogma at the moment. Yet the entire U.S. medical research community has no higher symmetry EM investigation of how acupuncture works, or of the infolded electrodynamics shown by Whittaker's decomposition of the scalar potential in 1903. Hence our scientists totally miss the vast electrodynamics of *the internal structuring of potential and charge with vacuum engines (spacetime curvature engines)*. As a result, the entire U.S. scientific and intelligence communities were totally unable to comprehend nearly five decades of Russian induction of health changes and diseases in personnel in the U.S. Embassy in Moscow, using weak EM radiation. Internal structuring of the activity producing what is called "charge"—including its internal structuring and engines—can be modeled and in theory directly engineered in the new Sachs-Evans approach.]

As shown in my paper, "Giant Negentropy from the Common Dipole," **Journal of New Energy**, 5(1), Summer 2000, p. 11-23, the broken 3-symmetry energy flow of any charge and any dipole allows a more fundamental 4-symmetry energy flow solution for the century-old problem of explaining the actual source of the EM energy that pours out of the source charge in all directions in 3-space. The 4-symmetry in EM energy flow explains its nature as longitudinal EM waves, with the charge's absorbed energy being *received* from the time domain (complex plane), transduced into 3-space, and emitted in 3-space. *The source of all EM energy at any point in 3-space is the time domain!*

It has been known for decades in particle physics that there exists no equilibrium in mass systems, unless the interaction of the vacuum with the masses is included. It has also been known that every charge and every dipole is a broken 3-symmetry in the fierce vacuum energy exchange with the charge or dipole. Yet the vacuum interaction—and specifically the source charge's and the source dipole's broken symmetry in it—are not even included in classical EM as yet. That this known total glitch in modern EM theory has not been previously corrected, when the particle physicists of every university physics department of note already know it is necessary, is inexplicable! Here we must regretfully point out finger at our orthodox scientific establishment, and particularly at our National Science Foundation and our National Academy of Science. These institutions are not even working on, nor are they aware of, nor do they even care about, the most fundamental science problems of our time. Bigger accelerators, yes. Windmills, yes. EM energy from the vacuum and electrical power systems which use it without destroying the source dipole, no. The greatest of all electrodynamics—the infolded longitudinal EM wave dynamics inside all EM fields, potentials, and waves—no. The correction of the dozens of major flaws and non sequiturs in classical electromagnetics, no. The understanding of the longitudinal EM wave interferometer weapons and quantum potential weapons already arrayed against us, no. The development of EM healing based on the Priore approach, which would totally revolutionize medical therapy and handle attacks on our cities by weapons of mass destruction, no.

We have a scientific mindset problem of epochal magnitude. And that costs the U.S. taxpayers hundreds of billions of dollars spent on the "status quo" or "business as usual" scientific endeavors that are far inferior to what can be done.

S Entropy. A quantity which measures the extent to which the energy of a system is available to do work.

Since in one sense energy is a function of ordering, then fundamentally, *entropy* refers to increasing disorder, which means increasing disruption of the ordering in our collected energy. Unfortunately, entropy is one of those concepts in physics for which there are several differing major views. We avoid all discussion of "information" etc., and prefer to deal only with the energy and time aspects. For our work in energy from the vacuum, we take the very simple view that an *entropic process* is like a positive resistor: As forward observer time passes, the entropic process continually disorders or "loses"—from further controlled conversion of energy form—the collected energy in the system. We regard a *negentropic process* as like a negative resistor: as forward observer time passes, the negentropic process receives energy in a form unusable to us, transforms, it, and outputs it in a form that is usable. Quoting Ilya Prigogine, Irreversibility as a symmetry-breaking process," Nature, Vol. 246, 1973, p. 70: *"Entropy Ycannot in general be expressed in terms of observables such as temperature and density. This is only possible in the neighbourhood of equilibriumY It is only then that both entropy and entropy production acquire a macroscopic meaning."*

So basically, the notion of entropy will apply when we have a system in equilibrium or very near equilibrium. For open systems in substantial disequilibrium, the entropy cannot even be computed, and the classical equilibrium thermodynamics with its infamous second law does not apply. On the other hand, the loading of disequilibrium (negentropic) EM systems is a very nonlinear affair. Some direct indications of this were in the Russian work in the 1930s which produced overunity and self-powering EM generators, in several physics institute. [E.g., see L. I. Mandelstam and N. D. Papaleksi., "On the parametric excitation of electric oscillations," Zhurnal Teknicheskoy Fiziki, 4(1), 1934, p. 5-29. English translation, Feb. 1968, Lawrence Radiation Laboratory, Livermore, CA, performed for NASA]. The U.S. researcher G. Obolensky also has done appreciable important experimental investigation of the nonlinear load effects for negentropic systems.

S Poynting energy density flow. Not the *overall* or *total* energy flow by any means, but just that component of it that is intercepted, diverged, collected, and used, where **S** is given by **S** = **E**×**H**.

The energy flow theory is rather thoroughly fouled, and has been since the 1880s when the flow of energy through space was proposed by both Heaviside and Poynting, independently and essentially simultaneously.

Since both **E** and **H** are defined not as fields per se, but as the reaction cross section of the fields with a unit charge, the S does not refer to "energy flow" per se, but to the intercepted component of the overall energy flow.

Poynting only considered the energy flow component around a circuit that strikes the circuit's surface charges and gets diverged into the conductors to power the electrons. Heaviside, on the other hand, also considered the startlingly enormous additional EM energy flow filling all space around the circuit's conductors, generally parallel to the conductors but missing them.

This huge Heaviside component is still there around every electrical circuit, but almost entirely ignored and unaccounted. It misses the circuit, passes on off into space, and is lost in most circuits. Yet it is trillions of times larger in magnitude than the very tiny Poynting component which actually gets intercepted, "caught" and used by the circuit to power its loads and losses.

Since no one could explain where on earth such an enormous nondiverged Heaviside component pouring out of the power source terminals could possibly come from, Lorentz—reasoning that, since the nondiverged energy flow did not do anything, it was "physically insignificant" (Lorentz' term)—then just integrated the energy flow vector around a little closed surface assumed around every volume element of interest. This efficiently discards the Heaviside nondiverged component *from all accountability*, leaving only the Poynting energy flow component vector accounted. Of course the Poynting energy flow will agree with the energy measured in the circuit, since we actually measure energy dissipation, and the energy that dissipates from the circuit must have first entered it. Electrodynamicists today just continue to use Lorentz's integration of the energy flow vector around an assumed closed surface surrounding any volume element of interest. Thus, sadly our electrical power system engineers unwittingly waste trillions of times more EM energy actually produced from the vacuum by the source dipoles in their generators, than what the feeble powerline is able to intercept and utilize.

We have nominated the vast unaccounted Heaviside energy flow accompanying every field and charge interaction as the source of the extra gravity known to be present in the arms of the spiral galaxies and holding them together. See my paper, "Dark Matter or Dark Energy?", **Journal of New Energy**, 4(4), Spring 2000, p. 4-11. Informally we have also nominated this dark Heaviside energy which fills all space, as associated with the "quintessence" energy that is responsible for the recently discovered antigravitational acceleration of the universe. In this view the giant negentropy process associated with charge (see my "Giant Negentropy of the Common Dipole" paper, *ibid.*) is ongoing in all the stars, galaxies, planets, nebulae, and other astronomical entities. The ordinary energy we measure in all these processes is gravitational; hence for symmetry and energy conservation, the primary longitudinal EM wave energy from the time domain that furnishes the energy to all these spatially energetic processes represents an increasing curvature of spacetime in the time domain as we travel outward from any origin point. This is an antigravitational force, in this concept, which is cumulative as a function of radial distance from the origin point. Hence we see the distant parts of the universe as accelerating the expansion of the universe. At present this is of course an hypothesis, and must be further validated by laboratory experiment and further theoretical work in a unified field theory electrodynamics such as the Sachs-Evans O(3) symmetry electrodynamics.

S-flow is actually comprised of an organized, structured flow of virtual photons or antiphotons and itself is not observable. Its interaction with charged particles (electrical charges and magnetic poles), however, is observable by the amount of energy collected upon (diverged around) the charge or pole (formation of a local potential), and formation of a translation force upon the charge or pole.

Again, the "**E**" and the "**H**" in the expression "**E**×**H**" are not really the "electrical field" and "magnetic field" in space, but only their *reaction cross sections* at each point in space, *should* a little unit point static coulomb be placed at that point to divert some of the **E**-field flow around it, or a little unit point static magnetic charge be placed at that point to divert some of the **H**-field flow around it. So even the Poynting energy flow **S** = **E**×**H** has been interpreted in terms of its reaction cross sections.

It is worth noting that electrodynamicists today are still politely debating EM energy flow and the energy flow vector. As an example, a polite "debate" of this subject has been ongoing for over 30 years in **American Journal of Physics** alone. A particularly significant statement of the problem is given by Jones: *"It is possible to introduce the Poynting vector **S**, defined by **S** = **E**×**H**, and regard it as the intensity of energy flow at a point. This procedure is open to criticism since we could add to **S** any vector whose divergence is zero without affecting [the basic integration procedure's result]."* [D.S. Jones, **The Theory of Electromagnetism**, Pergamon Press, Oxford, 1964, p. 52.] In short, here Jones is stating that an enormous nondiverged energy flow may accompany the Poynting flow component, and it will not be accounted for by **S** = **E**×**H**. After presenting the various elements required to solve the energy flow problem, Jones bows out with the statement: "*"It does not seem likely that an expression satisfying all these conditions will be simple… …fortunately, we are rarely concerned with the energy flow at a point. In most applications we need the rate at which energy is crossing a closed surface."* [Jones, *ibid*., p. 53.] Panofsky and Phillips warn against over-attachment to the Poynting vector: "*Paradoxical results may be obtained if one tries to identify the Poynting vector with the energy flow per unit area at any point."* [Wolfgang K. H. Panofsky and Melba Phillips, **Classical Electricity and Magnetism**, Second Edition, Addison-Wesley, Menlo Park, CA, 1962, third printing 1969, p. 180.

We point out, however, that for gravitational effects of energy flows in space, one must account the total energy flow per unit area *at and surrounding* any point, and therefore one must restore the unaccounted enormous Heaviside nondiverged EM energy flow component as well as the accounted tiny little Poynting diverged energy flow component.

Δt An increment (little piece of) time.

In special relativity, time can be considered a piece of spatial length compressed by the factor c, so that L = ct, where L is in meters, t is in seconds, and c is the speed of light. Time can also be considered as spatial EM energy compressed by the factor c^2. To comprehend time and time rate of flow, one must first realize that a priori no observable exists or persists in time. The observable is a 3-spatial snapshot made at one instant in time, so that "time stopped" momentarily. The observable is thus a frozen snapshot of the 3-spatial intersection of an ongoing 4-space interaction. At the very next instant, that snapshot no longer exists. Hence for a mass (or any other observable) to "propagate in time", there must be a process whereby Δt is continually added to the observable (such as to mass m, converting it to masstime mt for a short duration) and then subtracted again, leaving behind another frozen 3-space snapshot of the "observable" that has seemingly now endured for that little Δt period, with respect to the observer. As we explain in several papers, the sum total of the photon interactions with a mass m may be taken as the generator that produces (1) the flow of the mass m through time, by changing from m to mt to m to mt… continually, and (2) a substructuring of internal "time-flow components". In short, any time-flow ascertained in the macroworld or the microworld by the observer, has an enormously rich internal dynamics and structuring. Again, in higher symmetry electrodynamics such as O(3), the internal structuring of the photon, EM field, EM wave, and EM potential can be modeled and eventually even directly engineered.

ΔE An increment (little piece of) energy.

The photon is considered to be (unit-wise) composed of (ΔE)(Δt). However, in our view (as we explain in energy, energy collection, and energy dissipation) energy does not occur in "little fixed chunks" at all. One only has "little chunks" of energy when one *is collecting* energy in an ongoing process. In electromagnetics, the "collection" of energy is actually the *ongoing collecting* of energy from an energy flow deviating around an intercepting charge. It is a dynamic, ongoing process whereby a collector such as a charged particle intercepts and interacts with (undergoes virtual photon flux exchange with) an impinging energy density flow **S**= **E**×**H**. In an equilibrium (static) condition, the deviated flow remains very like a persistent, unchanging whirlpool in a river, and so is said to have a "static magnitude". If that Poynting flow is removed, then there is no excess collecting going on and the "magnitude of the deviated energy flow" is zero. EM energy collection/collecting is always like a rock in a high wind. As long as the wind is blowing on the rock, air is deviated to flow around the rock. The rock has a force field on and of it, and wishes to move. It also has "excess collected energy", and a gradient in this "excess collected energy" determined by the wind flow direction. When the wind ceases to blow, the rock goes back to being just a rock, and forgets it had all that excess wind energy collecting and any desire to move.

TERMS AND PHRASES

ABEL, NIELS HENRIK (1802-1829)

Norwegian mathematician who in 1824 proved that the general equation of fifth degree cannot be solved algebraically. He also left an important paper on transcendental functions which was published posthumously.

ABELIAN GROUP

A group obeying the commutative law of algebra for every pair of elements a and b under an operation **,** so that a*b = b*a. As an example, the set of integers is an infinite abelian group under the multiplication operation.

ACTION AT A DISTANCE

An action (effect) occurring in a local system, for which no separate local 3-space cause exists, but for which a causal correlation exists between a distant cause and the local effect.

To the local 3-space observer, the effect is baffling. In electromagnetics, action at a distance can be considered a higher topology phenomena, occurring through hyperspace or through subspace, or through "multiply connected spacetime." Particularly note that "time" is multiply connected; i.e., every spatial point in 3-space (in the entire universe) exists at the same moment in time, which means that electrodynamic action through the time domain can be used to eliminate the spatial distance between any two 3-space points. Action at a distance is absolutely required by quantum mechanics, and experiments establish its truth. Such action was implicitly included in Maxwell's quaternion theory, but was unknowingly discarded by Heaviside, Hertz, and Gibbs when they formulated a much more restricted vector interpretation of Maxwell's theory, after Maxwell's death. Bohm's hidden variable interpretation of quantum mechanics uses the quantum potential, which is a multiply-connected entity producing instantaneous action at a distance. In ordinary electrodynamics, the common scalar potential in the Coulomb gauge has instantaneous velocity, and it appears everywhere in space, wherever it will occupy, all at once *in that gauge*.

Time is also a set of energy flows in the time domain (complex plane) rather than in 3-space. Time is also multiply connected, since a priori every 3-space point in the universe exists at and in a single moment of time. Since time is dynamic and structured with special energy flows, it seems possible to use this feature to establish energy flow during one specific moment to a desired subset of 3-space points, regardless of where in the universe those points are located. Hence the development of "time-energy" technology allows action at a distance via manipulating the multiple connectivity of time and its internal structure (internal time-like energy flows).

AD HOC MODEL

A model used and fitted for a specific purpose or area.

Tentative formulation, not intended toward general application and not yet firm because there is still not a comprehensive, tested demonstration of the "fit" of the model to the complex phenomenology. Nonetheless, an ad hoc model can be quite useful in a given area, so long as it sufficiently fits the experimental data in that area and makes sufficiently accurate predictions in the area.

ADVANCED FIELD

An electromagnetic field that is a solution of the classical Maxwell equations, which is positioned on the future light-cone of spacetime.

We believe much of the problem with the "advanced solution" is a problem of interpretation. We would urge the re-interpretation not in terms of "time-travel" where the wave continues in observer forward time while the rest of the universe is assumed to retrogress to a previous state, but in terms of time-reversal of that wave only. In that case, the rest of the universe stays sane and continues to move in observer forward time, and the wave itself just moves in forward time in a path-reversed manner, back along a previous "stimulus wave" (the retarded wave that moved forward in time normally).

ADVANCED POTENTIAL

An electromagnetic potential that is a solution of the classical Maxwell equations, which is positioned on the future light-cone of spacetime.

Presently this potential has not yet been given a physical interpretation. See remarks above.

ADVANCED WAVE

An electromagnetic wave that is a solution of the classical Maxwell equations, and which is positioned on the future light-cone of spacetime.

See remarks above.

AFFINE CONNECTION

In mathematics: "A connection on a manifold, whose form is unchanged under affine changes of parameter along curves, e.g., when the original parameter t of a curve γ (t) is replaced by t = as + b." **Dictionary of Science and Technology**, Academic Press, Christopher Morris (Ed.), 1992.

Ugh! We'll let the mathematicians take it from there.

AFFINE TRANSFORMATION

"Transformation on a linear space to itself which can be expressed as the sum of a linear transformation and a fixed vector.

The affine transformations form a group. In the plane, the group is six-dimensional, consisting of translations, rotations, stretchings and shrinkings, reflections, simple elongations and compressions, and simple shear transformations, as well as compositions of these."

Dictionary of Science and Technology, 1992, *ibid*.

AHARONOV, YAKIR

Renowned Israeli quantum physicist and former student of David Bohm.

Co-discoverer of the *Aharonov-Bohm effect*: Even when the electromagnetic force fields reduce to zero in a charged particle system, the potentials may still exist and interfere to cause real physical effects. Together with Bohm, Aharonov authored a fundamental paper in 1959 which pointed out the primary importance of the *potentials* rather than the force fields. Indeed, the force fields are only made *in and of the charged particle system itself*, by interference there of potentials, and do not exist as such in the vacuum. The forcefields are thus not primary electromagnetic causes at all, but are *secondary effects* due to the interference of potentials in a particular charged particle system. [See *Aharonov-Bohm Effect*] In his three volumes of physics, Feynman pointed out that the forcefields do not exist *as such* in the vacuum, and that *only the potential for the forcefields exists in vacuum*. The primary causative agents for electromagnetics are thus the potentials. This has been proven both theoretically and experimentally in quantum mechanics, but still does not appear in present classical electromagnetic theory taught in universities—which terribly confuses cause waves and effect waves in wave-charge interactions.

AHARONOV-BOHM EFFECT

Phenomenon theoretically pointed out in 1959 by Yakir Aharonov and David Bohm whereby interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields.

See Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), 1959, p. 485-491. For an overview and extensive bibliography, see S. Olariu and I. Iovitzu Popescu, "The Quantum Effects of Electromagnetic Fluxes," Reviews of Modern Physics, 57(2), Apr. 1985, p. 339-436.

Chambers experimentally demonstrated the Aharonov-Bohm effect one year later. When **E-** and **B**-fields are zero in a region, their potentials still may exist and thus be stable and persistent. Being gradient-free, a force-free potential is not changing; no bleed-in or bleed-off of its magnitude *or of its internal structure* is occurring. Interference of the potentials produces real electromagnetic effects (energy) in the interference zone, even at a great distance, without transmission of "field energy" per se through the intervening 3-space. [Rigorously, the potential's EM energy can be considered to be *hyperspatially* transmitted, or *subspatially* transmitted (tunneled) via the Whittaker internal EM biwave mechanism. See discussions under *action at a distance*.]

In 1897-8 Stoney showed (in **Philosophical Magazine**) by physical reasoning that all the disturbances of the ether arising from sources of certain kinds can be resolved into trains of plane waves. This was extended and the general and special equations written by E.T. Whittaker in 1903 in **Mathematische Annalen**. In 1904 (delivered orally in 1903) Whittaker's second paper dealing with the subject was published in the **Proceedings of the London Mathematical Society**, and showed that scalar EM potential interferometry is in fact what generates the classical EM force fields.

The two Whittaker papers may be interpreted to show a startling arrangement in a potential or EM wave. (1) any EM field or wave in a particular 3-space is due to two interfering scalar potential functions in that space. So we may replace any EM field or wave with two scalar potential functions, a priori. (2) But each of the scalar potentials onto which the functions are applied, is decomposable into the Whittaker 1903 longitudinal EM wavepairs, where the output waves are in 3-space and the input waves are time-like and come from the complex plane. This provides a hidden "infolded LW electrodynamics" inside, and composing, all EM fields and waves. Further, the decomposition is recursive, so each of the first tier infolded EM waves can be decomposed again into a second tier of biwaves, etc. Thus the most primary electrodynamics of all is the nested infolded longitudinal EM wave structuring of all present EM waves, potentials, and fields. The internal LW waves also form rigorous curvature of ST patterns (templates) and thus engines. By engineering the structure and dynamics of the internal EM, one can engineer mass—including the atomic nuclei, quarks, and gluons—as one wishes, once the technology is sufficiently developed. In the West, this entire area has been ignored, but it has been highly weaponized by the KGB in Russia.

AIAS

Alpha Foundation's Institute for Advanced Study.

Unusual "think tank" comprised of fellows and fellow emeriti in various places in the world, who cooperate in advanced electromagnetics studies via E-mail. Led by Dr. Myron W. Evans, a theorist of note, the institute has been forging ahead with the production of O(3) symmetry electrodynamics, particularly as an important subset of Mendel Sachs' unified field theory. Dozens of AIAS group-authored papers have been published in various leading journals such as Foundations of Physics, Physica Scripta etc. More than 90 AIAS papers are carried on a restricted Department of Energy website for reference by DoE scientists. The work is particularly significant in working out how EM energy flows in both the 3-spatial and time domains, and how EM energy can be and is extracted from the active vacuum. More than a dozen of the later AIAS papers are carried on a DoE public website, http://www.ott.doe.gov/electromagnetic/papersbooks.html. Also, about two dozen AIAS papers have either been approved for publication in journals, or are in the review process. The present author is a Fellow Emeritus of the AIAS, and is kindly tolerated as a conceptualist by these advanced theoreticians.

AIDS

Acquired Immunological Deficiency Syndrome.

A condition of acquired immunological deficiency associated with infection of the cells of the immune systems with the retrovirus HTLV-III. It is usually recognized by testing positive to HIV infection, and eventually by physical deterioration and by development of one or more life-threatening opportunistic infections such as pneumonia, or of Kaposi's sarcoma.

AIKIDO

A Japanese martial art whose modern form is a synthesis of many schools of jujitsu and the zen art of meditation in motion.

It is the way (DO) of entering and turning (AI) the opponents vital energy (KI). The present author is a retired sandan (third degree black belt) in Yoseikan aikido. He is retired because of debilitating injuries suffered in other fields and accidents, and as a busted-up old dog of 70 has no business bouncing hard off the mats with all those vigorous young fellows. His extra 50 pounds also no longer enhance his athletic ability. We did write a book, privately held but never published, on Aikido and dealing with the principles of unarmed combat. Interestingly, we simply went into the deep unconscious in meditation (and perhaps a little beyond, depending on one's viewpoint) to uncover the unsuspected principles. To list them, we went through the alphabet four times. Many of these principles have never appeared in any text on martial arts or military combat. The task was placed on me by my sensei and required seven years to complete. Perhaps one of these days we will formally publish the book.

ALGEBRAS

The various types of mathematics in which the theoretical models for physics and other sciences are embedded.

Some important algebras utilized in physics are the vector algebra, tensor algebra, quaternion algebra, Grassman algebra, Pauli algebra, Clifford algebra, etc. There are many others. The complexity of the symmetry and of the topology varies significantly between these algebras. Thus, an electrodynamics model in a lower topology algebra such as vectors or tensors will not reveal or permit nearly so great a set of electrodynamic functions and operations as will electrodynamics embedded in a higher topology algebra such as quaternion algebra or Clifford algebra.

As an example, it has been shown by Barrett that the operation of many of Nikola Tesla's patented circuits cannot be understood if analyzed in vector or tensor electrodynamics. The circuits will reveal totally new functions when analyzed in quaternion electrodynamics. That paper is T. W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," **Annales de la Fondation Louis de Broglie**, 16(1), 1991, p. 23-41

ALLOPATHIC MEDICINE

A medical system which combats disease by using remedies which produce effects different from those produced by the disease treated.

Reliance is primarily upon drugs, pain relievers, and treatment of symptoms as well as use of agents to "kill" the microbial organisms regarded as the cause of the diseases. Surgery is used as a major adjunct. From time to time, other practices formerly frowned upon or even legally prosecuted are recognized as effective and legitimate, and are then brought in under the aegis of organized allopathic medicine. Physical therapy is such a formally prosecuted modality that was eventually adopted; hypnosis is another. The problems with organized allopathy is that (1) it tends to be monopolistic, (2) it uses the state's power to enforce its credo, (3) it controls both the bulk of the medical scientific literature and the flow of research funds from both government and private sources, (4) it does not adequately police its doctors and technicians, (5) it does not sufficiently accent preventative medicine or "whole human being" treatment, and (6) it engages in conspiracies to destroy competing systems of medicine (as in the case of chiropractic). (7) On the other hand, it has made remarkable advances in preventing killer diseases by vaccination, treating them by antibiotics, and in eradicating or controlling many of the former great killer diseases such as smallpox, diphtheria, and bubonic plague. It has been far less successful in treating and reversing debilitation, cancer, arteriosclerosis, etc. Just now, most of its major weapons against the former dread diseases are beginning to fail. The disease organisms are rapidly changing into new, more virulent forms that are resistant or even immune to the present "successful" treatments. Even in orthodox medical publications, it is now beginning to be recognizing that allopathic medicine is starting to fail, and that we are again entering a period of great, uncontrollable killer pandemics.

ALLOPATHY

Another name for allopathic medicine and allopathic medical therapy.

AMBIENT VACUUM

The average active vacuum, in space, removed from large collections of charge and mass.

AMBIENT VACUUM POTENTIAL

The average standard vacuum (spacetime) considered as a potential having structure and dynamics on a myriad levels.

Anything that is composed of virtual particle flux is a *potential*. Another way of looking at the vacuum is to realize that, having an energy density, then *a priori* it is a potential or may be modeled as such.

Hence the ambient vacuum can be decomposed via the Whittaker method into an internal set of infolded EM longitudinal waves of very special nature. Further, our work on the broken 3-symmetry energy flow involved in any potential or dipolarity shows the more fundamental 4-symmetry energy flow between the time domain and 3-space. This 4-symmetry energy flow is also highly structured and dynamic.

Whittaker's work shows that, contrary to present assumptions in quantum physics, a particular potential has its virtual particle flux formed into a beautiful bidirectional wave structure. The apparent disorder at any local point in the local vacuum potential is due to the constant arrival of myriads of potentials, from charged particles all over the universe. The violent interference and re-interference of these potentials results in the creation of violently changing virtual photons—or in terms of waves, violently changing EM waves and fields. In other words, the statistical disorder of the vacuum is actually *composed of* continually and rapidly changing perfect order. [Here we have confronted 3-law Aristotelian logic with its missing 4th law, because *at the limit* opposites *are* identical!] The continual violent interferometry results in the zero-point EM fluctuations of vacuum. Each EM fluctuation is completely deterministic, as shown in Puthoff's cosmological feedback mechanism. However, the macroscopic observer has no knowledge of the cause or of the arrival, and of the myriads of interferences that produced the final EM field fluctuation. Hence although the zero-point fluctuations of vacuum are completely deterministic one-by-one, the observer has *no information* on their causality. Hence to the observer these fluctuations are completely statistical and seemingly *without local order*. However, because of their hidden determinism, these fluctuations are chaotic rather than random—contrary to the Gibbs' thermodynamics statistics assumed and adopted by quantum mechanics. That statistics should be changed to an "already chaotic" statistics containing hidden Whittaker order. This would appear to be the real solution to the recognized major QM problem of the missing chaos.

Since their changes and appearances are completely in one-to-one correspondence to all distant features of the macroscopically ordered universe, then their integration yields macroscopic order. Further, as can be seen, Mach's principle—so near and dear to the heart of Einstein—arises directly from the foregoing discussion. In such fashion, integration of statistical (but chaotic, and with hidden order) quantum changes results in the familiar, ordered macroscopic world. This removes the great error in QM: that heretofore it has failed to predict the ordered macroscopic universe we observe and inhabit.

ANALOGUE

Part that is analogous (representative or modeled).

At root, an analogue of something is a second something that functions in similar fashion to the first something.

ANAEROBE

An organism that lives, is active, and occurs in the absence of free oxygen. Particularly applicable to one-celled organisms that do so.

ANAEROBE, FACULTATIVE

An organism that can live either as an aerobe or as an anaerobe. Particularly applicable to one-celled organisms possessing that capability.

ANAEROBE, OBLIGATORY

An organism that is restricted to being anaerobic; i.e., it cannot utilize oxygen from the atmosphere.

ANAEROBIC

Characteristic of an organism of living, active, or occurring in the absence of free oxygen.

ANGULAR MOMENTUM

The momentum or "leverage" of the momentum of a moving body, with respect to an axis or reference point.

Angular momentum as the same units as action; that is, it is energy multiplied by time, or momentum multiplied by length. In terms of particles, it is also known as spin.

ANOMALOUS EM JAMMING MECHANISM

In a dense signal environment, an unusual mechanism whereby EM energy from the skin of an aerial vehicle is directed laser-like into the internal circuitry.

The density of the signals impinging upon the skin are such that the probability of multiple photon interaction increases. Also, on a nonlinearity in/on the skin, the multiplex of signals creates a scalar EM potential. Decomposition of this potential into Whittaker waves provides hidden "pump" waves upon the nonlinearity, which therefore acts as a pumped phase conjugate mirror (PPCM). EM fields and signals from the internal operating circuits in the vehicle provide "signal waves" to the internal skin and to the nonlinearity PPCMs on the skin. Amplified phase conjugate replica (PCR) waves are generated by the PPCMs, where in each PCR wave the energy may be the total of the external signal density pump waves upon the initiating PPCM. The result is that the skin becomes a "collector" of energy from all the impinging external signals, and focuses much of this collected external energy into laser-like beams inside the vehicle which penetrate directly back to the circuit element that originated the particular signal field/wave to the inner skin. The result is spurious jamming throughout the vehicle, directly in its operating circuits, in a manner not described in normal ECM theory. Also, the same effect occurs in a human body when exposed to such a dense signal environment (threshold about 200,000 signals per square meter per second, and particularly active above 500,000 signals per square meter per second. I.e., in such an EM "smog" environment, the anomalous jamming mechanism is able to progressively conduct the energy impinging upon the skin deep into the body and throughout it, including into the bone marrow. This effect is not described at all in the standard dosimetry literature and experimental approaches for examining the effects of EM fields and radiations upon biological systems. By rigorous frequency band control of radiations and communications, most equipment can be operated in an appreciably dense environment such as appears on the modern environment, because the equipment is essentially "narrow band." The human, however, is a very wide-band receiver, and in such dense environments—even at very weak signal levels—the human system will eventually undergo cumulative damage. This is one of the major mechanisms involved in long term deleterious biological effects of EM radiation environments, but this effect on living systems has not been investigated in the laboratory by present EM bioeffects researchers.

In the U.S. air attack on Libya in 1986, the anomalous jamming mechanism interfered with the aircraft, controls, and missiles because frequency discipline was not rigorously enforced. In the Gulf War, rigid frequency discipline was enforced, with the result that the electronic systems operated normally. However, the troops exposed to that environment (which varied from place to place) were affected by the mechanism, which added a bit to the previous cumulative damage to their immune systems, etc.

ANTIBODY

A body substance produced in response to a specific antigen (protein or carbohydrate or other substance, such as a toxin or enzyme, capable of stimulating an immune response).

The antibody counteracts the effects of the antigen, by neutralizing toxins, agglutinating bacteria or cells, precipitating soluble antigens, etc.

ANTI-ELECTRON

The time-reversed electron, or *positron*. Under time reversal, electric charge reverses in sign.

Paul A.M. Dirac, in his "A Theory of Electrons and Protons," Proceedings of the Royal Society of London A126, 1930, p. 360 introduced the negative energy electron sea (now called the "Dirac sea") with negative energy electron holes treated as the positive electron. He attempted to identify these holes with protons, but that was later refuted. The positron, discovered about a year after Dirac's paper, was later connected with this anti-electron.

Note that Dirac himself was not adamant that *all* negative energy state holes in the Dirac sea were filled with electrons, but just that *almost* all of them were. Modern researchers have tended to extend this to state that *all* the holes are filled, normally. That is sufficient in forward time situations. But it does not necessarily hold in a time-reversed situation, in a time-reversed zone, or in a tempic back emf.

ANTIGRAVITY

An effect whereby normal masses *repulse* one another rather than *attract* one another

Antigravity may also be considered as *time-reversed gravitation*. It may also be considered as *gravitation between two negative masses*.

ANTIMATTER

Matter consisting of atoms which are composed of antielectrons (positrons), antiprotons, antineutrons.

Also loosely refers to the antiparticle corresponding to a particle—the antiparticle may be regarded as the particle traveling backward in time, or "phase conjugated", or "time-reversed.

ANTIPARTICLE

A counterpart to an ordinary fundamental particle, having identical mass lifetime and spin, but with charge and magnetic moment reversed in algebraic sign.

Note that when the charge is reversed, parity and time are also reversed. Thus, properly, the antiparticle is the phase conjugation of the particle, or just conjugate particle for short. In effect an antiparticle is the particle traveling backward in observer time rather than forward in observer time.

ANTIPHASED

180 degrees out of phase (contrary-wise in timing).

ANTIPHOTON

A time-reversed photon.

Presently, the photon is considered its own antiparticle, so that photon and antiphoton are one and the same. In the new approach, that is not quite true. First, one considers that a photon—as is well known—has the units of angular momentum. Thus it is made of "energy multiplied by time." We consider that the time-forward photon (i.e. a normal photon), is made of $(+\Delta E)(+\Delta t)$. Usually physicists have largely neglected the fact that a photon carries an *increment of time* as well as an increment of energy! In our approach the antiphoton, being a *time-reversed* entity, must be composed differently. We have tentatively considered the antiphoton to be comprised of $(\Delta E)(-\Delta t)$. It will probably initially be necessary to take into account the fact that Minkowski geometry treats the T dimension as imaginary rather than simple arithmetic. We haven't had the lengthy time to work all of that out; perhaps one of the better theoreticians or one of the sharp young graduate students will do that and save us the trouble.

At any rate, the requirement is that—in a coupled photon-antiphoton pair (a graviton)—the energies of photon and antiphoton must add spatially, while *electromagnetically* the translation forces (actually translation impulses) of photon and antiphoton must cancel or thwart each other. One insists on a strong interpretation of the distortion correction theorem in phase conjugate optics. In other words, we consider the "wave" as composed of photons, while the "antiwave" (phase conjugate replica) emitted by a phase conjugate mirror is composed of antiphotons. And we consider that the distortion correction theorem must apply to the wave's photons, so that the antiphotons are precisely superposing (coupling) and then decoupling as the wave and antiwave pass through each other in perfect spatial superposition. We feel that this "totally in phase spatially but totally antiphased temporally" concept is important. It is well-known, e.g., that in certain cases protons—the very building blocks of matter itself—can pass right through other protons, as shown in certain colliding beam experiments if their relative spins are properly arranged.

ANTISIGNAL

The phase conjugate of a signal. The time reversal of that signal or its twin. The phase conjugate replica signal/wave emitted by a phase conjugate mirror material in response to an incident signal/wave.

ANTI-STOKES EMISSION

Radiation emitted from an intensely scattering medium that is of greater intensity than the input radiation.

In other words, the medium gives off more energy than we have to put into it. This has been known since the 1930s and proven experimentally, but no really complete explanation has been forthcoming. The conventional explanation is that some of the kinetic energy of the medium is furnished to cause additional emission. In that case, anti-Stokes emission would lower the kinetic energy of the emitting medium. Note that this is not necessarily true at all in Letokhov's negative absorption (excess emission) of the medium, where in that case the excess energy in the emissions is said to come directly from energy absorbed from the active vacuum.

ANTI-STOKES RADIATION

The radiation coming from an anti-Stokes emission process from an intensely scattering medium, where more energy flux is emitted than is input by the experimenter-operator.

In the past, there has just been "arm-waving" that the excess energy must come from the internal energy of the molecules. There seems to be no explanation at all of where the excess energy of the molecules must continue to come from! Tacit silence on this "apparent" violation of energy conservation is held. But see Letokhov's *negative absorption of the medium* discussions in the literature.

We quote from H. C. Dake and Jack DeMent. (1941) **Fluorescent Light and Its Applications**, Chemical Publishing Company, Inc., Brooklyn, New York, 1941. p. 51-52: *"When a phosphor or other luminescent substance emits light, it gives in most cases an emission according to Stokes' Law. This law states that the wavelength of the fluorescent (emitted) light is always greater than the wavelength of the exciting radiation. It was first observed in 1852 in the memoir "On the Change of Refrangibility of Light," by Sir G. G. Stokes. In terms of energy the relationship states that e em < e ab. While Stokes' Law holds for the majority of cases, it does not hold in certain instances. In some cases the wave length is the same for both the absorbed and the emitted radiation. That is, the efficiency appears to be perfect or unity. This is known as resonance radiation. In other cases Stokes' Law does not hold where the energy emitted is greater than the energy absorbed. This is known as Anti-Stokes emission. In 1935 Prileshajewa showed that there is an energy difference as much as 1.1 v between the exciting light and the fluorescence of aniline vapor. This added energy is attributed to additions from the internal energy of the molecule."*

There is no real problem of identifying the source of the excess energy, once one realizes that (i) every dipole is a broken symmetry in the vacuum exchange with the end charges of the dipole, and thus extracts virtual energy from the vacuum, integrates and orders it, and outputs it as EM energy density flow. One then adds multiretroreflection and multiple passes of the EM energy flow, so that charges that intercept and collect energy from the reciprocating flow can multiply connect from the same energy flow (on a given pass, only a tiny, tiny fraction—nominally about 10^{-13}—of the energy flow is intercepted and interacted with by the collecting particle). Thus multiple collections from the same EM energy flow can be accomplished, if the flow is iteratively rerouted (as by retroreflection or phase conjugate reflection) back across the collecting charges again and again.

In this fashion additional energy is collected in the collected volume, which increases the local collected energy density in that volume. In turn this increases the local potential of that volume. *In turn* that increases the energy flow from that potential acting as a dipole, because it increases the dipole strength and consequently the magnitude of the dipole's broken symmetry in its vacuum flux exchange.

The inefficiency of the process increases nonlinearly with increasing local potential, and faster than the buildup of local energy density occurs from multipass multicollection. When the inefficiency curve reaches the increasing broken symmetry energy extraction curve, then a point of equilibrium is reached. At that equilibrium point, the rate of energy diffusing out of the multipass region and escaping from the periphery of the system is equal to the rate of extraction of energy from the vacuum and flowing it out into as Poynting **S**-flow into the multicollection region and collecting it.

So we have described an open thermodynamic system, where there are two inputs of energy flow: (i) that by the operator-experimenter, and (ii) that from the vacuum via the dipolar broken symmetries. At equilibrium condition the total input energy rate is equal to the total energy escaping rate, and the conservation of energy law is rigorously upheld. However, this is an open thermodynamic system far from equilibrium, and so rigorously the nonlinear thermodynamics of such systems applies, rather than the normal equilibrium thermodynamics. As is well-known, the coefficient of performance for such open systems in disequilibrium, may permissibly be overunity, even greatly overunity.

Properly utilized, the overunity anti-Stokes or Letokhov emission can be made very large—such as COP ≈1200 in Patterson's adaptation. The process and its overunity certification already exist in the hard science literature, being known and experimentally demonstrated for at least three decades.

ANTIWAVE

The time-reversal (phase conjugation) of a reference wave.

A phase conjugate replica wave is a time-reversed replica of its referent wave. Hence it is the antiwave of the referent wave.

A-POTENTIAL

Vector magnetic potential.

The A-potential is a real entity more primary than the magnetic field **B**, and not a mathematical convenience as was originally thought—as detailed in the second volume of Feynman's three volumes of physics. Indeed, it may be separated from the magnetic field, as reported by Feynman, and used for instrumentation effects in the manner patented by Gelinas. When separated from magnetic field, it is often referred to as the "curl-free vector potential," since the **B**-field is modeled as being created from the **A**-field by curling it. In other words, **B** is defined as the curl of the magnetic vector potential, if we use $\mathbf{B} \equiv \nabla \times \mathbf{A}$. Hence "separating" magnetic field from **A**-potential simply is a sort of "parallel branching" of the energy density flow of the **A**-potential. It is like branching the voltage on a feeder line conductor in an electric circuit into two parallel conductors. The feeder potential splits into two, so that each branch has the same potential. Actually the vast energy flow in the feeder line only has a small component which strikes the surface charges, diverting the energy into the conductor and potentializing the Drude electrons. In branching into two conductors, the giant flow branches also, but is so great that the surface charges in each of the two branches still intercept essentially the same amount of energy, diverging it into the conductor. Hence each conductor exhibits the same potential as the feeder conductor

The same thing happens when one "splits" or branches the **A**-potential in similar fashion. The diverted and separated **B**-field branch will exhibit "equal energy density" flow (i.e., in the form of magnetic flux density flow) and yet the original **A**-potential path through space will still have the same previous energy flow density magnitude, just without curl (without magnetic field).

Any good toroid or very long solenoid already performs such separation, and hence energy density flow doubling. If properly used—as in the motionless electromagnetic generator (MEG)—this effect can be used to provide a generator with COP>1.0. At this writing, the present MEG exhibits COP = 16, and is a patent-pending system.

The well-known Aharonov-Bohm effect, e.g., also uses this separation of the curl of the **A**-potential, leaving an uncurled **A**-potential whose changes will and do interact with electrons.

The curl-free vector potential is particularly interesting because (1) it is a unique and independent field of nature, (2) its curl produces a magnetic field, (3) its time derivative produces an electric field, (4) its magnitude does not fall off inversely as the distance squared, but only inversely as the distance, and its time derivative $d\mathbf{A}/dt = B\ \mathbf{E}$. Hence, as in the MEG, one can extract energy both from the curl-free A-potential and the curled **A**-potential (magnetic field **B**), increasing the available energy collected in the circuit to greater than what the operator input to it. This is simply a physical process of regauging, wherein electrodynamicists already routinely assume (and apply) the fact that the potential energy of any EM system can be freely changed at will, by regauging. In the real world, of course, we may have to pay a little switching energy, but we are free to regauge (by this "splitting the potential into equal parallel paths, and by other methods), at will and at any time and place. The MEG just takes advantage of one major method for free regauging, in order to violate the Lorentz symmetry condition (violate symmetry in the system's energy exchange with its active vacuum environment). It thereby is legitimately permitted to exhibit COP>1.0 and output more energy than one inputs, because the regauging transforms it into an open dissipative system far from thermodynamic equilibrium in its active environmental energy exchange.

ARISTOTELIAN LOGIC

The conventional three laws of logic, unconsciously fitted to the physical reality created by the single photon interaction. The three laws are: (1) a thing is identical to itself, (2) a thing is not identical to something else, and (3) either one has a given thing or something that is not that thing. Mathematically these are written as (1) A A, (2) A Ã, and (3) A Ã. Aristotelian logic applies only to single observation at a time (monocular observation, a single observer). When multiple simultaneous observation (multiocular observation, multiple observers) are considered, the three laws may be violated. In that case a fourth law, A Ã is added. This law may be seen to be the exact negation of all the first three. An application rule (which may itself be called a fifth law) is also added, and states that either the first three laws are explicit and the fourth is implicit, or the first three are implicit and the fourth is explicit. Further, the fourth law is the *law of the paradox*: Something that is found to be true in nature, but violates one or more of Aristotle's three laws, is known as a *logical paradox*. In the higher logic, it simply violates the case where the three Aristotelian laws are explicit and the fourth law is implicit. Instead, the three Aristotelian laws are now implicit and the fourth law has become explicit. All boundaries exhibit this effect.

It is highly recommended that the interested reader familiarize himself or herself with Morris Kline, **Mathematics: The Loss of Certainty**, Oxford University Press, New York, 1980. Kline deals directly with the fact that mathematics is not a body of unshakable truths about the physical world, and mathematical reasoning is not exact and fallible.

ARISTOTLE

A great Greek philosopher who lived from 384-322 B.C.

Aristotle is famous for his three laws of logic: (1) a thing is identical to itself, (2) a thing is not identical to something else, and (3) either one has the thing considered or one has something else.

Heraclitus pointed out that Aristotle's laws prohibited any change, and yet changes were ubiquitous. In short, for a thing to change, it had to change into something else, so Heraclitus pointed out. So how could a thing be both itself and something else simultaneously? How could it be not itself while being itself? The problem has not been solved by the logicians and philosophers to this day, because it has no solution in Aristotelian 3-law logic. So Heraclitus, believe it or not, concluded that all change was an illusion.

ATHERIOSCLEROSIS

Basically, clogging of the arteries, which may be due to deposits onto the artery walls or by thickening and growth of the artery walls themselves, or a combination. It is a chronic (slowly progressing) disease where the artery walls become progressively thickened and hardened and the artery becomes less elastic.

It has been found that the common cytomegalovirus (one of the herpes viruses) plays a key role in the development of the disease. Initial deposits irritate and injure the arterial walls, and the herpes 6 virus infects the site, partially nullifying the control of excess grown of the arterial muscle tissue. This deposit and concomitant stimulation of some excess muscle growth occurs repeated, so that the muscle tissue grows abnormally over an extended period of time. For this reason, when an operation is performed to forcibly reduce the deposits inside the artery, the resulting injuries to the arterial wall tissues leads to the herpes 6 infection and stimulation of excess muscle growth. It has been found, e.g., that after angioplasty the process is often engendered, so that within a year or two the artery is narrowed again, but this time by abnormal growth of smooth muscle tissue. The Prioré team demonstrated positive and total remission of arteriosclerosis in laboratory animals by means of nonionizing EM irradiation to produce an amplified "antiengine", which "time reversed" or dedifferentiated the stricken cells and the stricken areas back to a previous state.

ARTIFICIAL POTENTIAL

A scalar potential which is composed of, or partially contains, an artificially assembled bidirectional wave set.

See E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," **Mathematische Annalen**, Vol. 57, 1903, p. 333-355 for proof that a "scalar potential" is actually a harmonic set of hidden bidirectional EM phase conjugate longitudinal wavepairs. Each wavepair consists of a wave and its antiwave (true time-reversed replica wave).

If the external observer could see the *detected* (effect) waves in a hidden wavepair, he would see the "wave" going in one direction and the antiwave passing precisely through it in the other direction. However, prior to detection, the phase conjugate wave exists entirely in the complex plane and hence in the time domain (in -ict). So the interpretation of Whittaker's 1903 paper, in terms of cause and effect waves, is that the phase conjugate half set is incoming from the time domain to each and every point in the dipolarity or potential, and the "real 3-space half set" is being radiated in every direction in 3-space from that point. See my paper, "Giant Negentropy….".

In addition to Whittaker's sum set of biwaves, Ziolkowski added the product set in the mid-1980s. E.g., see Richard W. Ziolkowski, "Exact solutions of the wave equation with complex source locations," **Journal of Mathematical Physics**, 26(4), April 1985, p. 861-863; Rod Donnelly and Richard Ziolkowski, "A method for constructing solutions of homogeneous partial differential equations: localized waves," **Proceedings of the Royal Society of London A**., Vol. 437, 1992, p. 673-692. Particularly see I.M. Besieris, A.M. Shaarawi, and R. W. Ziolkowski, "A bidirectional travelling plane wave representation of exact solutions of the scalar wave equation," **Journal of Mathematical Physics**, 30(6), 1989, p. 1254-1269.

However, these deal only with the effects waves after detection.

Another way of defining an artificial potential is: *A potential formed by or containing a subset formed by, deliberately summing nonzero force field vectors to a zero vector resultant*. [Note that electrodynamicists routinely discard vector systems that sum to a zero resultant; not realizing that in doing so, they are discarding vacuum engines and a vastly extended electrogravitational dynamics!] The vector zero system of infolded nonzero vectors has a deterministic, internal stress pattern that is macroscopic. The type of potential made by the zero summation depends upon the type of force vectors summed. For example, summing electric field vectors to zero produces an electrostatic scalar potential; summing magnetic field vectors to zero produces a magnetostatic scalar potential of vacuum.

All potentials represents local warps or curvatures in vacuum/spacetime, and all artificial potentials contain internal, deterministic templates of nested internal spacetime curvatures. Curvature of spacetime interacts directly upon mass and physical systems embedded in it. Thus these deterministic, internested local spacetime curvatures of the artificial potential represent vacuum engines, or spacetime engines, and the artificial potential is said to be a *dimensioned* or *activated* potential. Any system exposed to and placed in such a vacuum spacetime engine, will have these hidden vacuum flux asymmetries interacting upon each and every part of it, down to and including the nuclei, the nucleons, and the quarks in the nucleons.

ASYMMETRICAL REGAUGING

Change of either the scalar potential f or the vector potential **A**, but not both, so that a single excess force and a single change of system energy occurs in the system.

This violates the assumed Lorentz condition, and thus excess energy can enter the system, and an excess "free" force appears which can (if the system is adroitly designed) then translate to perform work upon the system to increase its (e.g. kinetic) energy, exhausting the excess regauging energy in the process and restoring symmetry.

In conventional EM theory, Maxwell's equations in potential formulation are in terms of two potentials, f (the scalar potential) and **A** (the vector potential). When Heaviside retranslated Maxwell's quaternion set of 20 equations in 20 unknowns, he transformed it (in potential form) into two equations where main variables f and **A** are not separated. Electrodynamicists, who until recently considered the potentials as not even real, but just mathematical figments and conveniences, then simply arbitrarily changed both f and **A** *just precisely so* that the extra force appearing in the system by the change of f was countered by an equal and opposite extra force also appearing in the system by the change of **A**. This type of *symmetrical (no net force) regauging* is known as the *Lorentz condition*.

Each *asymmetrical* (produces a nonzero excess net force) potential change (each half of the two asymmetrical regaugings that comprise the symmetrical regauging) also altered the overall energy of the system. However, in the symmetrical regauging that total energy change is "bottled up" as increased stress in the system, because no net translation force is available to be translated and use the energy change to produce excess "free work."

The Lorentz condition assumes that the designer *will not allow* the equilibrium of the system to be broken! In that case, of course, ordinary equilibrium thermodynamics applies, and the system cannot produce COP>1.0. Prior to this Lorentz condition assumption and its absolutely arbitrary regauging, Maxwell's equations do include asymmetrical regauging and therefore open EM systems far from thermodynamic equilibrium; precisely those which violate the Lorentz condition (symmetric regauging) assumption. In short, the unregauged equations do include COP>1.0 electrodynamic circuits and systems.

Unknowingly the electrodynamicists *arbitrarily* discarded the entire Maxwellian class of overunity EM engines and circuits—which class incorporates making a change in one potential(s) such that the Lorentz condition does not hold.

ASYMMETRY

Lack of symmetry, or "broken symmetry" in a general sense.

ASYMMETRY OF THE DIPOLE

Any dipole or dipolarity is a broken 3-symmetry in its fierce exchange of EM energy with the active vacuum. This means that it continuously receives and absorbs copious virtual energy from the seething vacuum flux, and that not all this absorbed energy is reradiated back to the vacuum in virtual form. Instead, some of it is integrated into observable form and re-radiated in all directions as usable, real 3-space EM energy.

This is standard, well-known particle physics, although it appears to be totally missing from classical electrodynamics and it is not considered at all by electrical power system designers and analysts!

Every electric charge is a broken symmetry in the virtual particle flux exchange between the quantum mechanical (energetic) vacuum and the mass of that electric charge. The dipole is two different such asymmetries. As is well-known in particle physics (not in classical EM theory!) each such broken symmetry must "gate" out some of the vacuum energy. In short, the charge asymmetry "extracts" some vacuum energy from its normal "energetic exchange" with the vacuum flux, and puts it out in some form we call "observable" (capable of being measured; specifically, of being intercepted by electric charges so as to produce a force field across the charge and translate it).

Well, an electrical charge produces a flow of energy density which we may represent conventionally as $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ (when collected and measured on assumed unit point charges), which we—a little bit mistakenly—call "observable." [Actually, until the energy flow interacts with that assumed unit point charge, it is an *organized flow of virtual entities*] After interaction it is real observable EM energy in 3-space. The spinning electron has its own magnetic moment and its own **E**-field. Since (in one form) also $\mathbf{E} = -\nabla\phi$, an extracted and formed $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ is output by the charge asymmetry. This output energy flow also produces the assumed ϕ evidenced by the charge interaction, since ϕ as used represents the *reaction cross section* of the potential. Note that all these entities **S**, **E**, **H**, and ϕ are defined in electrodynamics only as their own interaction cross section with a unit point charge containing mass.

In classical electrodynamics (formed long before the notion of charged particles, atoms, and EM energy flow!) it is just recognized that "charge is the source of a potential." Actually charge gates the energy flow from the vacuum flux, receiving the energy from the time domain by its asymmetry, transduces the energy into 3-space energy, and outputs it as real EM energy pouring out in 3-space in all directions. That energy-flow contains and carries the energy that will interact with charge to form the **E**-field, the **H**-field (spin field), and ϕ.

Also, in particle physics and gauge theory it is well-known that *no mass system can be in equilibrium anyway, unless the vacuum-to-mass interaction is incorporated!* After CEM was formed, much later the atom was discovered and the electron was discovered. Much later then, particle physics and quantum mechanics came along. CEM theory has never been altered to include the necessary vacuum interaction, although it has been known for decades that CEM is totally in error in this respect.

Here one poses a question to the sharp young graduate students: Why does every university continue to teach a seriously flawed classical electrodynamics which does not include the vacuum/mass interaction, when the physics department at the same university knows full well that it must or no EM system can be in equilibrium? Why is not such a *fundamental* and universal issue strongly raised and intensively discussed by our large, tax-exempt scientific organizations at their annual symposia? Why do not the National Science Foundation and the National Academy of Science demand that serious work be done at all speed, to correct such known scientific deficiencies in the classical EM theory? *What happened to science as the search for physical truth*, and replaced it with "status quo" models or—in some cases—even dogma? It is an inexplicable mystery to me, and one to which I do not know the answer.

AUTOIMMUNE DISORDER

A disease or disorder relating to or caused by the antibodies or T-cells that attack foreign molecules, cells, or tissues of their host organism.

A flagrant autoimmune disease example is AIDS, Acquired Immunological Deficiency Syndrome. The AIDS virus successfully hides from the antibodies, so that they cannot perform their proper function, and it infects the T-cells and negates their function. However, there are other examples other than AIDS: e.g., rheumatism and osteoarthritis.

BACK EMF

The counter-electromotive force, usually between the end charges of the source dipole for an EM circuit.

When the electrons in the circuit are forced against the emf—as when the charges from the ground reference of the circuit are forced back up through the emf of the source dipole—then work is done upon the source dipole to scatter the charges and destroy the dipole. Removing the dipole removes its receipt of vacuum energy via its broken 3-symmetry, transduction of that energy, and output of it along the circuit and in space around the circuit as the energy flow that powers the circuit (with part missing the circuit and being just wasted).

Since this removes the power from the circuit, then additional energy must be input to the source to perform work on the internal charges and restore that source dipole again.

There are actually two emfs in a circuit, not one. There is a 3-space emf and also an entirely unrecognized tempic emf. In either case, under proper circumstances there can be a back-emf or either one or both. See discussion under *electromotive force, back*. See also *tempic back emf* (to be added).

BALL LIGHTNING

Rare form of lightning where an incandescent, slow-moving globe forms. The globe often moves erratically, explodes, and may set fire to objects it touches.

Ball lightning may be due to scalar electromagnetic interferometry which can produce a controlled pattern of electromagnetic energy at a distance, or it may be due to the formation of a special localized spacetime curvature engine.

BARE ELECTRON

The true electron without its partial shielding of virtual positrons that are attracted around it.

BARRETT, TERENCE W.

Noted modern electrodynamicist and consultant who works in SU(2)×SU(2) gauge symmetry extended electrodynamics, and was also one of the pioneers of ultrawideband radar.

BECKER, ROBERT O.

Medical doctor, medical research scientist, and Nobel Nominee, noted for his early-on efforts—particularly in the 1960s through the 1980s, in the effects of EM fields on biological systems, studies of evoked or partially evoked limb regeneration in certain species, electrical stimulation of bone fractures to stimulate healing and reveal the mechanism involved, and the cellular regeneration system of the body as opposed to the immune system.

Becker particularly recognized that the cellular regenerative system was a special electrical system having unique functions. He even succeeded in diagramming this system as far as it is revealed by ordinary U(1) electrodynamics. Becker demonstrated partial limb regeneration in salamanders, etc. and in some other species, with electrical stimulation. In his EM-stimulated bone fracture healing, Becker showed a startling result: First the red blood cells entering the fracture area shucked their hemoglobin and grew a nucleus, thus "dedifferentiating" (biological term) or "time-reversing" (physics term) back to an earlier state. Then this modified cell redifferentiated (biological term) or "fast-forwarded in time" (physics term) into the type of cell that makes cartilage. Then this newly modified cell further redifferentiated into the type of cell that makes bone, and was deposited in the fracture site. The continuing deposits of these cells from the ongoing process evoked in red blood cells healed the fracture with new bone growth. Whereupon the action stopped.

Becker also was one of the first qualified scientists who testified in court on the deleterious effects of power line radiation. He thereby incurred the enmity of very powerful interests, so that his research funds were pulled and eventually he was forced to retire at an early age.

He was nominated for the Nobel Prize multiple times, and his epochal work did finally focus at least some attention of the medical research community on the regenerative system as opposed to the immune system which heals nothing.

Becker's profound work was of tremendous influence on me as I struggled to decipher the active mechanism in his experiments, the Prioré team's experiments in France, the Russian microwave radiation of U.S. Embassy personnel in Moscow, Kaznacheyev's experiments demonstrating transmission of cellular disease and disorder electromagnetically into target cells environmentally isolated except for contact by light from the sender cells having such disease or disorder, and the body's own regenerative system. After some 14 years struggle with the problem, I finally did break the mechanism(s) involved in all those things. But it was only possible because of the magnificent work by Becker and Prioré, the reported work by Kaznacheyev, and gradual exposure to higher symmetry electrodynamics. The work was crude at first, but eventually progressed (by 1998, e.g.) to include the port-hole concept, transduction of EM wave polarization types, the engine and engine template concepts, and unified field theory. Since then it has become clear that the proper electrodynamics to model such things is the higher symmetry O(3) electrodynamics pioneered by Evans, and now revealed to be an important subset of Sachs' unified field theory.

Becker is therefore one of my heroes whose magnificent work and long perseverance against strong scientific suppression I greatly admire. Most of Becker's detractors were or are in fact employed in positions either directly or indirectly funded by the electrical power industry. I believe that, had Becker had access to O(3) electrodynamics and Sachs' theory, he would have personally impelled the entire medical field to a marvelous and extraordinarily advanced EM medical therapy.

See Robert O. Becker and Andrew A. Marino, **Electromagnetism and Life**, State University of New York Press, Albany, 1982; Nice summary of EM bioeffects (orthodox) ; also on http://www.ortho.lsumc.edu/Faculty/Marino/EL/ELPub.html.

See also R. O. Becker, "The neural semiconduction control system and its interaction with applied electrical current and magnetic fields," **Proceedings of the XI International Congress of Radiology**, Vol. 105, 1966, p. 1753-1759, Excerpta Medica Foundation; Amsterdam; — "The direct current field: A primitive control and communication system related to growth processes," **Proceed. XVI Internat. Congr. Zoology**, Washington, D.C., Vol. 3, 1963, p. 179-183; — "The control system governing bone growth in response to mechanical stress," **J. Ark. Med. Soc**., Vol. 62, 1966, p. 404; — "A technique for producing regenerative healing in humans," **Frontier Perspectives**, 1(2), Fall/Winter 1990, p. 1-2; — with Charles H. Bachman and Howard Friedman, The direct current system: A link between the environment and the organism," **New York State Journal of Medicine**, Vol. 62, April 15, 1962, p. 1169-1176; — and D. G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents," **Trans. N.Y. Acad. Sci**., Vol. 29, 1967, p. 606-615; — and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals," **Bulletin of the New York Academy of Medicine**, Second Series, 48(4), May 1972, p. 627-64; — and Carlton F. Hazlewood, Abraham R. Liboff, and Jan Walleczek, **Electromagnetic Applications In Medicine**," NIH-OAM Electromagnetics Panel Report, Jan. 15, 1993.

See particularly Robert O. Becker, "The significance of bioelectric potentials." **Bioelectrochemistry and Bioenergetics**, Vol. 1, 1974, p. 187-199. This is probably *the* paper by Becker that provides a really good overview of the control system. A block diagram of the proposed system is given by Becker on p. 191.

Becker also has authored several popular-oriented books filled with excellent information. These include Robert O. Becker, M.D. and Gary Selden**, The Body Electric: Electromagnetism and the Foundation of Life**, William Morrow and Co., New York, 1985; and Robert O. Becker, **Cross Currents: The Perils of Electropollution; The Promise of Electromedicine**, Jeremy P. Tarcher, Los Angeles, 1990.

BEDINI, JOHN

Noted audio engineer, overunity energy researcher, and inventor of the renowned Bedini amplifiers well-known to audiophiles, the BASE process for holographic sound, and a process for cleaning the harshness from digital audio disks.

Bedini has invented a variety of COP>1.0 processes, motors, generators, etc. Particularly he has developed a process whereby, in a battery-powered circuit, he adroitly back-pulses the battery in a recharging direction while the battery is powering the load, causing a pile-up of electrons on the surface of the battery plates that is a sharply increased scalar potential opposing the onrushing very massive lead ions in the battery in discharging mode. The result is that the potential's electron pile-up increases dramatically due to the ions' far greater inertia than that of the electrons. The result is a high potential (perhaps 40 volts) upon the surface of the battery, overpotentializing both the incoming lead ions between the plates and the electrons in the external circuit in powering mode. In this phase, energy flows from the vacuum onto both the ions and the external circuit electrons, adding significant energy to the powering of the external circuit for the same ion current inside the battery.

As the ions slow and stop, Bedini's back-pulse trailing edge causes a very sharp cutoff, thus invoking Lenz's law and resulting in the surface potential on the plates increasing to well over 100 volts. This further overpotentializes the external circuit electrons in powering mode, dissipating much more energy in the external circuit to power the loads. As the ions finally stop and reverse, their relatively sluggish re-acceleration in charging mode and their overpotentialization still results in excess energy being delivered to the external circuit. Then as the overpotentialized ions accelerate in charging mode, the external circuit is still in powering mode until the drain-off of the severe electron pile-up on the surface of the battery plates.

This is one of several modes of operation used by Bedini in his overunity battery powered circuits, which keep their batteries recharged while powering their loads.

By separating the other closed current loop between external electron currents and the ion currents between the battery plates, and then dephasing these currents, Bedini breaks the usual dissipation symmetry enforced in a closed loop circuit where all charge carriers have the same m/q ratio. In Bedini's circuits, the m/q charge ratios dramatically differ between the ion current section between the plates and the electron currents between the outside of the plates and through the external circuit. Also, by providing a much greater potential *between* the two circuit halves, The potential extends in both directions—internally back over the ions, and externally in the circuit over the external Drude electrons powering the circuit loads and losses. He provides a true negative resistor on the surface of the plates, which extracts energy from the vacuum (via its broken 3-symmetry as a dipolarity) from the time domain, transduces it into real EM energy, and overpotentializes (regauges) the potential energy of both the internal ion currents and the external electron currents.

The Bedini process is therefore a legitimate COP>1.0 system, and in his self-powering embodiments the system will run indefinitely without external energy input by the operator, use of fuel, etc. He has converted the system to an open system far from thermodynamic equilibrium in its violent energy exchange with the active vacuum, and thus the system is permitted to exhibit both COP>1.0 and self-powering as well.

For an expanded explanation of the Bedini negative resistor process in batteries, see T. E. Bearden, "Bedini's Method for Forming Negative Resistors in Batteries," **Journal of New Energy**, 5(1), Summer 2000, p. 24-38.

BEING

Pure essence, existence.

Again, something which in actuality resists Aristotelian definition unless preliminary assumptions are made. Generally used in one of two senses: (1) persistence, as in "he is (he continually exists)," or (2) identity, as in "he is the man ('he' and 'the man' are the same thing)."

BIDIRECTIONAL EM WAVES

EM waves running antiparallel to each other through a given point or region of space.

Suppose we run a normal EM wave into a phase conjugating mirror (PCM) with a gain of 1. By ordinary phase conjugate optics, the mirror then emits a backwards-traveling phase conjugate replica wave (PCR wave). According to the so-called "distortion correction theorem," that backwards traveling wave will travel backwards (as seen by the external observer) back over the exact route taken by the previous wave, and it will reappear in space precisely superposed spatially with the previous waves. If we examine the impulse (density) fields (i.e., Edt), then the forward-time wave contains Edt while the backward time wave contains E(−dt), which is just −Edt. So *electromagnetically*—as far as producing a net translation force on a charge—the E-fields seem to cancel, although they are still there and their energy is still there. Instead of translation, the two produce stress.

One thus (simple case) will have a strange kind of standing wave wherein the local energy density of the vacuum varies (e.g., sinusoidally) as a function of distance, but a charged particle placed in it will not be translated! Rigorously, that is a *gravitational standing wave*. Since change in local energy density is also a curvature of spacetime, that is also a standing sinusoidal wave of the local curvature of spacetime. This is where electromagnetics and gravitation actually meet, and where electromagnetics can be utilized to produce gravitational waves. (Much more complex waves can be produced; this is just the simplest example).

BIDIRECTIONAL FIELD FLOWS

E-fields (and B-fields) being propagated in a pair of opposite waves traveling "through" each other in opposite directions,

...where the two may be "locked together" as a wave and its antiwave; i.e., as a "Whittaker biwave pair". Note that Whittaker bidirectional EM wavepairs are not to be confused with the pairing of the wave and reflection wave on a transmission line.

BIDIRECTIONAL GATING

Gating (of energy flow, e.g.) in both directions simultaneously.

BIDIRECTIONAL TRAVELING PLANE WAVE

A wave and its antiwave, where each wave is moving in an opposite direction (standard assumption in phase conjugate physics).

In short, to the external observer the waves seem to be "moving through" each other; one from left to right, and the other from right to left; or one radially outward and the other radially inward. All scalar potentials were shown by Stoney and Whittaker to be comprised of a harmonic series of such "bidirectional traveling wave pairs." However, the interpretation of the phase conjugate waves was after detection, so that the detected "effect" wave existed in 3-space. Prior to detection, the phase conjugate half-set of the Whittaker decomposition of the scalar potential is convergent from and in the time domain (complex plane).

BIOELECTROMAGNETICS (BEM)

Presently, the scientific study of interactions between living organisms and electromagnetic fields, forces, energies, currents, and charges. The range of interactions studied ranges from atomic and molecular, through intracellular, to the whole organism**.**

Electromagnetic fields studied in BEM include those of endogenous origin (from within the organism) and exogenous origin (from outside the organism). Sources of exogenous fields include: the earth and sun, electrical equipment (power lines, radio transmitters, etc.), clinical treatment devices (e.g., TENS), and other organisms. The electromagnetic fields of greatest interest in BEM are generally classified as nonionizing (i.e., lacking sufficient energy to directly dislodge electrons from molecules or atoms.) BEM is a scientific discipline emerging on the boundaries between physics and biology. As in other sciences, progress in BEM is made by testing falsifiable hypotheses against valid measured data obtained in repeatable experiments.

Our new viewpoint seeks to dramatically enlarge the scope of this definition, to include not only the presently-recognized external (outfolded) electromagnetics, but also the internal (infolded, inside the scalar EM potential) electromagnetics, and also to utilize a unified field theory where EM is GR and vice versa.

BIOENERGETICS

That branch of Russian (KGB) energetics which is targeted against the living body, its cells and tissues, and its EM fields and potentials (biofields and biopotentials).

BIOGENESIS

Generation of living organisms from other living organisms; also, the theory or philosophy that living organisms develop only from other living organisms.

BIOLOGICAL WARFARE

Warfare against living humans, livestock, crops, etc. using pathogenic micro-organisms or biocides to create diseases and death, or using dimensioned electromagnetic potentials, fields, and waves to kindle cellular damage, disease, and death in the same targeted populations.

BIOLOGICAL WARFARE, ELECTROMAGNETIC

The use of dimensioned electromagnetic potentials, fields, and waves to induce cellular damage, disease, and death in targeted living populations.

The decades of Russian microwave radiation of personnel in the former U.S. Embassy in Moscow is an example, where only field-free potentials were dimensioned with the disease or health-change patterns (the engines). This science was developed in Russia as an outgrowth of the thousands of Kaznacheyev experiments demonstrating that any cellular disease and damage pattern could be induced from diseased or damaged cells into environmentally isolated cell cultures, purely by specialized EM radiation from the damaged or "transmitting" cells into the targeted cells.

The Chinese also used such dimensioned radiation to induce cancer in almost all GRU operatives stationed in China at the time of the major disagreement between China and Russia. In his book, Stanislov Lunev and Ira Winkler, Through the Eyes of the Enemy: Russia's Highest Ranking Military Defector Reveals Why Russia Is More Dangerous Than Ever, Regnery, Washington, D.C., 1998, p. 160, Lunev reveals it this way:*"...I was not the only person who was stationed in Beijing to come down with* *cancer… Almost all of the GRU officers stationed in China with me have been diagnosed with cancer. People assigned to the embassy had told me that they were picking up high levels of radiation from the apartment building around the embassy compound. Whether it was a form of high-tech monitoring or a conscious effort to harm our health, it seems to have been successful. So much so that in 1992 the GRU was asked to investigate the high rate of cancer among embassy officials. The findings, however, have never been made public; in part, no doubt, because of the rapprochement between Russia and China."* In short, the Chinese did exactly to the Russian GRU what the Russians were doing to the Americans.

BIOPHOTON

A quantum of angular momentum, each containing "a piece of energy welded to a piece of time, with no seam in the middle," emitted by biomolecules in living organisms. The difference between a biophoton and a normal photon is that the "infolded EM structuring" (ST curvature engine sets) carried by the biophoton carries the dynamics and templates of the living cell or tissue or body emitting the radiation.

As presently known, biophoton emission is associated with processes such as mitosis (cell division) and with mechanisms such as soliton vibrations of DNA and other alpha-helix proteins [Popp et. al., 1984; Popp et. al., 1992]. An organism's biophoton field may represent a highly-structured innate organizing field capable of encoding or transmitting information over macroscopic distances.

Kaznacheyev, e.g., showed that almost any cellular disease or damage condition can be induced from the diseased control cells to separated and environmentally isolated target cells initially normal, if (1) the visible spectrum of radiation is blocked from the targeted cells, and (2) the two groups of cells are optically connected below and above the optical spectrum, in the infrared and in the ultraviolet.

The new viewpoint seeks to dramatically extend the scope of this definition. Considering the wave aspects of the photon frequency, the infolded oscillating EM energy in the photon may be decomposed into two oscillating potentials [Whittaker, 1904]. In each of these potentials, there exists a magnificent infolded structure [Whittaker, 1903] of internal bidirectional EM wave pairs, to include the sum set [Whittaker] and the product set [Ziolkowski]. Thus the photon carries internal information, which is also referred to as the "information content of the field." Since the product set represents modulations, the basic infolded structure has signal characteristics. Mind, thought, and long term memory exist inside the personal quantum potential, which may further be regarded as consisting of coupling and uncoupling photon/antiphoton pairs—i.e., hidden gravitons. In the new viewpoint, the biophoton is the carrier of biogravitons, and also biograviton lattices. This information forms vacuum engines which can cumulate over a period of time to breach the quantum threshold and outfold, producing real change in cells, as shown by the Kaznacheyev cytopathogenic mirror effect experiments.

BIOQUANTUM POTENTIAL

A quantum potential connected to and of a living body, where the inner EM structuring and dynamics of the quantum potential are in correspondence with the inner EM structuring of the normal EM fields, potentials, and waves in the body, and also are in correspondence with the time-polarized EM structuring and dynamics of the mind and its operations.

The biopotential provides the volition and intent exhibited by the biological body, breaching the quantum threshold and producing direct psychokinetic change (by the mind-body coupling mechanism) that is the "input signal" to the servomechanisms of the body,. from the smallest to the largest. In this biopotential, the Stoney-Whittaker-Ziolkowski decomposition provides the information structure and vacuum engines that are the mind, thought, long term memory, etc. A living biological system may be defined as a physical system of otherwise inert matter (the body) and of mind, coupled to each other through a bioquantum potential.

BLOCKING

In the charge-barrier semiconductor functioning, stopping or partially stopping the flow of current in a circuit, or between two points, without stopping the Poynting S-flow.

In ordinary usage, "stopping the progress of."

BOHM, DAVID

World renowned physicist and originator of the hidden variable theory interpretation of quantum mechanics. See David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II." Physical Review, 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II).

Together with his student Aharonov, he co-authored a fundamental paper on the Aharonov-Bohm effect wherein interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields.

BOHM'S HIDDEN VARIABLE THEORY (HVT)

A major interpretation of quantum mechanics, formulated by David Bohm and published in Physical Review in 1952.

See David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," **Physical Review**, 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II). Bohm's theory makes all the correct predictions and also eliminates many problems in quantum mechanics, such as the "measurement problem" and the—now alarming—problem of the missing chaos (i.e., the missing "hidden order.").

Ironically, the prevailing Copenhagen interpretation of the QM theory predicts that the organized macroscopic world does not exist, while Bohm's theory predicts it. [Question to the sharp young graduate student: How can one defend a scientific model which predicts that he himself does not exist? If one keeps one's sense of humor, it would seem that this has to be the greatest scientific *faux pas* of all time!] It may be that one unconscious reason that most physicists still oppose HVT is that *it implies that physical reality can in fact be deterministically engineered*. Most physicists would seem to fervently wish physical reality to stay "solid as a brick," with fixed "laws of nature," which can be leisurely discovered and then utilized comfortably, predictably, and sanely. The advent of HVT (which is just a matter of time; for decades the Russians have highly weaponized the HVT area!) dramatically alters our perception of "fixed physical reality" into that of an "engineerable, changeable" physical reality! I tentatively believe that the terrible psychological implications of such a potential development is something that most scientists do not wish to cope with—we simply do not have the Gestalt for it. On the other hand, the Russian national psychology is such that they do have the Gestalt for it. The Russian energetics—which I have long referred to loosely as scalar electromagnetics—basically utilizes the infolded ordering inside the potential to create "direct patterns of vacuum flux", hence *vacuum engines*. And it also utilizes the quantum potential, including internally structuring the QP, to enable instantaneous action-at-a-distance in multiple simultaneous locations and nodes, with stupendous and automatic energy amplification.

By deterministically "patterning" the internal structuring of the scalar potential, one can directly form such vacuum engines. Whittaker showed in 1904 that any EM field pattern whatsoever can be expressed as two scalar potentials, rather than the common f and **A** representation. This means that any EM wave is simply the "interference" (between the hidden biwave structures) of two scalar potentials, even at a distance. By properly structuring an EM signal (i.e., structuring the two potentials into which it can be decomposed), the resulting EM signal—appearing normal on the surface—nonetheless now transports special vacuum engines deliberately designed to perform some particular internal task in mass systems, electronic systems, or humans when it strikes a target at a distance. I refer to this "internal structuring" of the Stoney/Whittaker/Ziolkowski internal biwaves and their products, as *dimensioning* of the potential(s), and hence of the EM carrier wave. The Russians call it the *information content of the field*; but our own scientists tend to erroneously interpret that phrase as ordinary spectral analysis. In other words, they erroneously continue to regard EM *vacuum engines* as if they were ordinary EM *signals*.

BOHREN EXPERIMENT

Revealing experiment by Bohren, replicated by Paul and Fischer, which shows that a given charge when in particle resonance absorbs more energy from a field or potential then the same charge does in "static" position.

By conventional outlook, this provides a guaranteed COP>1.0 process. Bohren demonstrated absorbing 18 times more energy with the charge in particle resonance than with the static charge. This is a direct experiment proving the Heaviside dark energy component of EM energy flow, whether in field form or in potential form.

See Craig F. Bohren, "How can a particle absorb more than the light incident on it"?", **American Journal of Physics**, 51(4), Apr. 1983, p. 323-327. See independent confirmation of Bohren's work by H. Paul and R. Fischer, *ibid*., p. 327. Other pertinent references are V. S. Letokhov, "Laser Maxwell's Demon," **Contemporary Physics**, 36(4), 1995, p. 235-243; — "Generation of Light by a scattering medium with negative resonance absorption," **Soviet Physics JETP**, 26(4), Apr. 1968, p. 835-839; — "Double Gamma and optical resonance," **Physics Letters A**, Vol. 43, 1973, p. 179-180; — "Stimulated emission of an ensemble of scattering particles with negative absorption," **ZhETF Plasma**, 5(8), Apr. 15, 1967, p. 262-265. Also see R. Pappalardo and A. Lempicki, "Brillouin and Rayleigh Scattering in Aprotic Laser Solutions Containing Neodymium," **Journal of Applied Physics**, Apr. 1992, p. 1699-1708

BRIDGE

A component or process which passes EM energy between two isolated circuits, but does not pass dq/dt between them, and which breaks the normal rigidly field-locked power dissipations in the two circuits.

Space itself would appear to be one such bridge.

BRIDGING FUNCTION

The function of passing EM energy flow between two isolated circuits, but not passing dq/dt between them, while simultaneously breaking any normal rigidly field-locked power dissipations in the two circuits.

BROKEN SYMMETRY

The fundamental meaning is a condition in which two parts of some configuration or shape on opposite sides of some divisor or condition or boundary, are not similar but differ.

In vacuum energy physics, we are very interested in the fact that any charge and any dipole represents a broken symmetry in its fierce energy exchange with the vacuum. This implies that something virtual has become observable; i.e., part of the virtual EM energy absorbed from the vacuum by the charge or by the dipole is changed into observable form and re-emitted as real, observable EM energy.

Any observable implies a broken symmetry. Non-observables imply symmetry, and vice versa. Each symmetry also can be represented by a conservation rule or law. Physicists have uncovered many kinds of symmetry in physics; and in the 1950s they also discovered broken symmetry.

BROMWICH, THOMAS JOHN I'ANSON

Brilliant English mathematician, 1875-1929, who made significant contributions to the mathematics of electromagnetism—particularly to superpotential theory—before his passion for overwork led to development of a mental disorder and eventual suicide.

Bromwich originated a method for solving the source-free Maxwell equations in 1899, publishing it in 1919. This approach formed the basis for his seminal paper on scattering of plane EM waves by spheres. He did excellent original work on infinite series and made contributions to quadratic and bilinear forms. He also placed Heaviside's operator calculus on a rigorous basis by treating the operators as contour integrals. Unfortunately Bromwich worked so intensively, in such painstaking detail, and at such length that his physical health was affected. He developed a mental disorder and eventually ended his own life in suicide.

For a summary of Bromwich's life and achievements, see G. H. Hardy, "Thomas John Bromwich," **J. London Math. Soc. 5** (1930), 209-220; — "Obituary of Thomas John Bromwich," **Proc. Roy. Soc. London A 129** (1930), i - x. For an exposition of Bromwich's equations-solving work, see J. D. Zund and J. M. Wilkes, "Bromwich's method for solving the source-free Maxwell equations," **Tensor (NS) 55** (2) (1994), 192-196.

BROWN, G. SPENCER

Noted author and creator of the laws of form, a new form of logic that dramatically extends Aristotelian logic.

Essentially Brown added complex numbers to logic, rather than just staying with simple 2-dimensional plane figures for postulate derivation. Electrical circuits— designed by Brown's logic—do function and do work. When computer software is designed by Brown's logic, it is then possible to mathematically prove whether or not the end result possesses any defects (bugs). In the future software—particularly the larger systems—may well be designed by Brown's logic. G. Spencer Brown, **Laws of Form**, Julian Press, New York, 1972. With a dramatic advance in logic itself, Brown has succeeded in formulating a true calculus of form. Note: There is now a second edition of the book.

BULK GRADIENT CHANGE

Simply a change in the amplitude of the entire potential.

… (i.e., to every internal biwave pair comprising it) at some point or in some region of space, as compared to changing the amplitude of only one or a few of the internal biwave pairs of the potential, without changing all the rest.

BYPASS RESISTOR

In electrical theory, a resistor to pass current by (in parallel to) some other component.

If other effects such as use of the overpotential, pumped phase conjugate reflection, quantum wells, quantum tunneling, and charge blocking are added, the function of the bypass resistor becomes extremely complicated.

CANCER

Conventional definition: A *malignant tumor* of potentially unlimited growth that expands locally by invasion and systemically by *metastasis*.

Growths are further characterized by the type of tissue in which they occur, and their actions. Of particular interest are neoplasm, tumor, carcinoma, and sarcoma. *Malignant* disorders tend to produce deterioration, infiltrate, metastasize, and terminate fatally. A *neoplasm* is simply any new growth of tissue that serves no useful function. A *tumor* is an abnormal mass of tissue that (1) is not inflammatory, (2) arises from preexisting cells, and (3) serves no useful purpose. A *carcinoma* is a malignant tumor of *epithelial* (covering membraneous tissue of a free surface) origin. A *sarcoma* is a malignant neoplasm that arises in *mesodermal* tissue (connective tissue, bone, cartilage, or striated muscle) and that spreads by (1) extension into neighboring tissue or (2) by way of the bloodstream, or (3) both.

New definition of cancer: A centrally-commanded, final, desperate, "first-step dedifferentiation" adaptive attempt by stressed, affected cells experiencing sustained oxygen shortage (hypoxia) to reverse their cellular evolution and return to the anaerobic stage of their primeval cellular ancestry. Thus a cancer cell is a severely stressed hypoxic cell that has been "ordered" by the master cellular control system (MCCS) to dedifferentiate back to a more primeval form in which it needed little or no oxygen. It is the last desperate attempt by the body to insure the affected cells' survival when all other means available to the MCCS to provide the necessary oxygen have been exhausted and have failed. The "order" differs from normal signal theory in that it consists of a Stoney/Whittaker/Ziolkowski template inside the personal quantum potential of the organism. Thus the "order" actually consists of a vacuum engine, so that the spacetime in which the affected cell exists is acted upon by structured, energetic hidden variable fingers to directly alter it. The "signal order" is in fact a set of hidden vector forces in spacetime itself, created directly in and on the affected cell(s), in every part, in every atom. The cure for the cancerous cell, once formed, is simply to cause the MCCS to generate the "counter-order" reversing the original dedifferentiating action. This can be readily accomplished by establishing hidden multifrequency pump waves in the quantum potential on the cancerous cell, on all its parts, so that they act as pumped phase conjugate mirrors. Since they already exist in the altered "cancer" spacetime template/signal, the pumping causes a phase conjugate replica of the precise cancer state to be created in and on the cell and its parts. [In fact this is a general therapeutic mechanism, and it can be applied to reverse almost any conceivable cellular damage or disorder.] This counter-order is a new and amplified vacuum engine that acts upon the cell to redifferentiate it back to a normal cell. If an abnormal number of cells results, the body's normal processes then recognize the excess cells and destroys them.

CANCER HOTSPOT

An area or locality or location where cancer recurs repeatedly, more than the normal deviation expected statistically. In such case, one is dealing with a hidden cause or carcinogen.

Human systems exposed to such a hotspot experience cellular changes (hypoxia) which eventually result in the master cellular control system (MCCS) issuing a dedifferentiation order to the affected cells to regress back to a primeval anaerobic form. The first step is dissociation from central (MCCS) control of cellular growth, resulting in a cell that now grows and divides unrestrained by the normal body. The real question in the hotspot is, "What is the unknown agent inducing the hypoxia condition, and what is the mechanism by means of which it induces it?" Epidemology will point out a hotspot, but is usually incapable of resolving the primary causative agent or its active mechanism since it can only show correlation.

CAPACITOR AS A TRANSMISSION LINE

A capacitor can be theoretically treated as a special type of transmission line, a fact which is apparently known to transmission line theorists but not to many engineers and physicists not specializing in transmission lines.

CARCINOGEN

A substance or agent (such as a chemical or a spacetime curvature engine) capable of inducing a cancer.

In the new theory, the primary carcinogen is an actual order (a vacuum engine) issued by the master cellular control system (MCCS), reverting (dedifferentiating) the affected cell to a primal anaerobic form. Usually what are normally regarded as carcinogens are also causes of cellular conditions wherein the cell becomes very hypoxic, so much so that the MCCS exhausts all other alternative corrective means and as a last desperate measure orders the cell to change to a form that requires little or no oxygen. The "exhausting of other measures" may be very rapid, as when the cell has been affected and damaged by powerful carcinogens, or it may require decades to occur, as when the cell is subjected to long-term mild hypoxia, such as from electromagnetic smog.

CARRIER WAVE

A fundamental wave which is modulated by another wave or other waves, and hence "carries" the other waveform(s). By stripping off the carrier in a demodulator, the carried waveform(s) emerges.

Hence it is a continuous frequency wave which can be modulated by an information-bearing signal.

CASIMIR EFFECT

The attraction of two conducting parallel plates in space, placed very close together, due to their influence on the active vacuum and on the vacuum's interaction with the plates.

CASIMIR, HENDRIK BRUGT GERHARD (1909-2000)

Noted scientist who in 1948 predicted that two parallel conducting plates, placed very near each other in a vacuum, would experience an attractive force due to their influence on the electromagnetic vacuum.

See H. B. G. Casimir, “On the attraction between two perfectly conducting plates,” presented at a meeting of the Royal Netherlands Academy of Arts and Sciences on 29 May, 1948. Published in the same year in **Proceeding. Koninklijke Nederlandse Akademie van Wetenschappen,** Amsterdam, vol. 51(7), 1948, p. 793-796. For a beautiful modern experiment confirming the force, see S.K. Lamoreaux, “Demonstration of the Casimir Force in the 0.6 to 6μm range,” **Physical Review Letters**, 78(1), Jan. 6, 1997, p. 5-8.

The force was experimentally detected some 10 years later.

CAUSAL SYSTEM ROBOTS (CSRs)

Massless 4-space "engine" systems (therefore in the causal domain only, hence the name) formed of longitudinal EM waves and their dynamics, which are therefore sets of spacetime curvatures and their deterministic dynamics (engines) which function as individual robots.

This is a totally new kind of system leading to a totally new kind of warfare, developed over some decades and deployed by the KGB in 1999.

Matter is composed mostly of empty space filled with EM waves, fields, and potentials. Those "fillers" themselves are just bundles of longitudinal EM waves and their impressed dynamics. Hence all matter, the vacuum, and all conventional EM potentials, fields, and waves—being filled with vast sets of longitudinal EM waves—are vast superhighways for longitudinal EM waves to travel on or in. The CSRs, being totally LW and LW dynamics in nature, thus are free to travel right through EM radiations from power lines, phone lines, TV stations, radio stations, etc. They easily penetrate any matter, including right through a mountain. Any function of a normal EM system can be performed—at least in theory—by such robots. This includes longitudinal EM wave interferometry, which forms external EM fields and waves, of whatever pattern or form desired. Further, these CSRs are actual functioning systems, and if communications functions are designed in, one can communicate with them at a distance, using longitudinal EM wave communication, right through the intervening earth and ocean, etc.

For some time, extensive underground facilities—hundreds—have been building in the former Soviet Union, and this massive buildup continued unabated after the Russian economy collapsed and the Berlin Wall came down. This tremendous expenditure, when the Russian economy was crippled, shows the strategic significance of these facilities. Further, apparently our satellites have not seen any missiles, normal weapons or munitions, etc. associated with that vast complex. In short, it is suspected that we don't have the foggiest notion of what it is. Indeed, those facilities will already have sent some CSRs into those very satellites, to insure that we do not find out what is going on, and perhaps to "spy on the spy satellites".

We think these terribly important facilities are the KGB-controlled facilities for developing and producing the CSRs, and contain a vast longitudinal EM wave command and control set of facilities as well. In short, this is the strategic arsenal, development, and control for building sets of various type CSRs in huge numbers—a vast new army of robots with unprecedented capabilities, under the command and control of a human "strategic force" operations command of very large size.

Developing and debugging one CSR would be a formidable and lengthy task. Once a single one is developed, making a million copies is simple. Just take a normal CD-ROM, e.g., and record something on it. Send in a working CSR robot to "reside inside" the infolded EM interior of the recorded informational electromagnetic pattern on the CD-ROM. Then simply copy that CE-ROM magnetically, as many times as is desired. Each copy contains a copy of the original signal that was recorded, including the exact infolded electromagnetic structuring inside that signal. Voila! One has as many new duplicate CSRs of that exact type, as one wishes, for a few pennies each.

One can see the implications. In all those underground facilities, there are many development laboratories, production laboratories, and command and control facilities and "storage facilities" for the robots. Those hundreds of facilities are in fact a great new strategic weapon system and all its development labs and production facilities, as well as its tactical control and command and control systems, its "troops", its "equipment", and its commanders and operators.

Well, then one can deploy and "utilize" such a strategic armada just as one would use any other great strategic army. One "deploys" one's forces against the enemy. In this case, one just sends the armada (individually controlled!) right through the "LW superhighways" of matter itself, right into the strategic territory and very heart of the targeted nation. One deploys the robots everywhere—e.g., inside the electronics in nuclear warheads, inside the control systems for the nuclear submarines and nuclear bombers, inside the satellites, inside the fighter aircraft electronics, inside the fighting ships, inside the main battle tanks and major weapons carriers, inside the signals on the hard drives of computers, etc. One sends them right through the "inside" of the targeted nation's power grid electrical energy signals, through the TV stations' emitted signals, etc. The Pentagon and Cheyenne Mountain are duck soup, as is the stock exchange and other civilian systems vital to continued functioning of the infrastructure. So is everywhere else, even the deepest vault and "special security" facility.

In short, the vast electronic command and control activities of an otherwise "modern" superpower provide great, convenient, easy superhighways for full strategic deployment to be completed by millions of weapons systems (CSRs) into everything of value in the targeted nation or nations. This completely solves the usual weapons delivery problem; delivery is made in advance of D-Day.

Such a vast CSR armada was deployed throughout the Western world, and particularly in the United States, just before the end of 1999. Yes, this was to be the KGB's great millennium gift to America—a really enormous Y2K problem that would stun and disable the entire nation, its armed forces, its electronic systems in its major weapons, and its command and control centers in one great TOT (time on target) assault. The major electrical power systems of the nation would have been disabled, as would have been the major communications systems. Using LW interferometry, the robots would have wreaked havoc on the nation. They would have disabled the electronics of every major C4I computer of importance, every major fighter aircraft, every major fighting ship, every nuclear submarine, every ICBM, every nuclear powerplant, etc. The attacking force was in its deployment and employment sites in the targets themselves. Everything was in place and cocked and ready to fire. To test whether we knew what was going on, the KGB even had one or more robots temporarily produce hacker-type operations and signals "emerging" in some U.S. command and control center computers, and deliberately let our fellows discover this "hacker attack" was coming from the Russian Academy of Science. Then the CSR simply stopped its LW interferometry, which caused the "normal hacker electronic" signals to just vanish and never reappear. Little wonder that our counter-type fellows were totally unable to find any trace of it, or where it was residing. They have no technology for such.

Well, happily those early-model CSRs are easy to root out and destroy, if one has quite advanced LW technology. Visualize these beasts as existing in a special sort of "internal sky" or "infolded atmosphere" inside matter. Strategic range LW "radars" can just sweep through this "internal atmosphere" inside matter, etc., detecting and tracking those CSR beasts analogous to how normal radar detects and tracks aircraft. The difference is that the strategic LW can also use a bit of LWI pulse to instantly snap those CSRs like popping bugs, completely jamming their operations or even "erasing" them once their characteristics are observed.

That very new kind of vast strategic warfare occurred right at the end of 1999, all through the United States—the Pentagon, our strategic weapon systems, the works. The friendly little nation rooted out and destroyed all those "on site" CSRs just prior to the attack being launched.

A full-up strategic war, of previously unknown kind, was fought all over the U.S. and our own system guys apparently didn't know it. It didn't even surface. Didn't make a whimper.

That hostile CSR armada consisted of *first generation systems*. These systems did not have defensive countermeasures. The systems were depending on total surprise, but were found out (that is another story that would make a great movie if one could live to film it) and then were destroyed in short order.

CSR development has ushered in a totally new kind of strategic, operational, and tactical warfare. The delivery of all the weapon systems "into the target" is accomplished long before the advent of hostilities. The second generation CSRs, once built, will be even more sophisticated. These will almost certainly be on line soon. They will likely also possess countermeasures.

An eerie new chapter in the annals of warfare has been initiated, and the world will never be the same again. Once the CSRs have sufficient defensive countermeasures such as stealth means and coordination with countering, rapid response LWIs to target and destroy the hostile "LW radars" seeking them, etc., the next world-wide strategic CSR war will be neither so simple nor so bloodless.

NOTE: Contrast this CSR discussion to the discussion of *nanotechnology* and nanobots. Just as nanobots are natural developments which teams skilled in nanotechnology would produce, so CSRs are natural developments which teams skilled in longitudinal EM wave technology—and in the "infolded" electrodynamics inside all EM fields, potentials, and waves—would produce. For decades (since the 1950s) the KGB has been developing and using longitudinal EM wave interferometers (LWIs) which beam right through the "inner highways" inside the fields and waves inside matter, thus allowing very large LWIs which beam right through the intervening earth and ocean, and can "focus" under the ocean or within the earth as well as out of it. With such background, the CSRs were just as natural a development for them as the nanobots are for conventional Western nanotechnology scientists.

CAUSALITY

The relationship between "cause" and "effect". The basic notion in a causal system is that the system's response to an input signal is not dependent upon future input values.

Thus the notion of causality involves a precise time ordering of ordinary inert interactions, where the cause was never input from a future time. Macroscopic (observable) causality may be and is violated by the interaction of hidden subquantal (virtual state) variables.

CEREBRAL HEMISPHERE

Each half of the cerebrum, i.e., each half of the enlarged anterior or upper part of the vertebrate brain.

The two cerebral hemispheres may function as a longitudinal EM wave interferometer or a time-polarized (scalar) EM interferometer to sense distant scenes (even distant in time), or produce EM energy and signals at a distance, since the body, brain, and nervous system does produce longitudinal EM waves and scalar EM waves. For proof of scalar EM wave interferometry in O(3) electrodynamics, see M.W. Evans, P.K. Anastasovski, T.E. Bearden et al., "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," **Journal of New Energy**, 4(3), Special Issue, Winter 1999, p. 76-78.

CHANGING THE PAST

Every intentional inception upon ordinary inert causality prior to collapse of the wave function represents a slight alteration or change in direction, departing minutely from pure causality. From a causal viewpoint, each human inception minutely changes the causal past, as well as the future.

Experimental proof that the causal past can be changed is provided by the delayed choice two-slit experiment. E.g., see C.W. Rietdijk, "Another proof that the future can influence the present." **Foundations of Physics**, 11(9/10), 1981, p. 783-790. Particularly see Wheeler, John Archibald Wheeler, "The 'past' and the 'delayed-choice' double-slit experiment," in A.R. Marlow [Ed], **Mathematical Foundations of Quantum Theory**, Loyola University, New Orleans, Louisiana, June 2-4, 1977; Academic Press, New York, 1978, p. 9-48.

CHAOS

In dissipative dynamical systems, the dynamical evolution that is aperiodic and highly dependent upon initial conditions of the system.

The trajectories of the system move on a strange attractor, which is a fractal subspace of the phase space.

The term suggests the ordinary meaning of random and unpredictable disorder, but there exists intrinsic determinism and at least partial embedded ordering. The mathematical equations describing chaotic behavior are very nonlinear and so complex that at the present time they cannot be computed or predicted. Even simple systems, including simple electromagnetic systems, can exhibit unpredictable chaotic behavior. The science of chaotic dynamics is still in its infancy, having begun moving once the modern computers became economical and pervasive.

CHARGE, ELECTRICAL (Q)

Electrical charge q is defined as $q \equiv m_q \phi_q$, to first order.

As can be seen, the actual "charge" action associated with the mass is due totally to ϕ_q, and to the broken symmetry of ϕ_q in the vacuum flux exchange with m_q. The charge q can be further broken down into a set of composite dipoles if the gathering of a virtual charge screen in the vacuum, surrounding the mass of the "bare charge" inside the gathering, is included. At first order, ϕ_q can be expressed (particle view) as a change in the local vacuum virtual photon flux (VPF), due to the VPF exchange between vacuum and m_q. The ϕ_q component is actually the "electrical charge" and massless of itself.

However, see my paper, "Giant Negentropy of the Common Dipole," **Journal of New Energy**, 5(1) Summer 2000, p. 11-23. We include the isolated charge by treating it as a system of multiple composite dipoles. We also explain the previously unrecognized 4-symmetry EM energy flow ongoing, where incoming longitudinal EM waves are being received, transduced into real longitudinal EM waves in 3-space, with the latter being emitted in all directions in 3-space.

A charged particle of mass is thus a little energy flow generator. It is a broken 3-symmetry in the local VPF, and therefore has an "observable" energy flow output, extracted and gated from the asymmetry in the VPF. This "observable" energy flow outputs the well-known energy flow **S** in classical EM, though CEM does not recognize or include this vacuum interaction, the broken 3-symmetry of the dipole, and the steady input of EM energy to the dipole in the form of EM energy from the time domain (complex plane).

Instead, classical EM just assumes that the "charge" q is the source of a potential. When originally formed as a concept, q was just a "quantity of electric fluid," like a "cubic centimeter of fluid," etc. The atom and the electron had not yet been discovered.

CHARGE CARRIERS

Anything—usually fundamental particles, ions, etc.—which consists of or contains charges and can move.

It can be an electron, a collection of electrons fixed in a moving dielectric, ions, atoms, molecules, etc.

The moving charge (i.e., the current dq/dt) with an energy flow streaming onto it and on beyond, carries its interaction with that energy-flow right along with it in the current
i = dq/dt. This is what the charge's "excess collected energy" consists of, and this is how the *collected/collecting* energy in the circuit is transported to the *collected-energy* dissipation sites (loads and losses).

CHARGE, GRAVITATIONAL

To be added

CHARGE, MAGNETIC

Essentially, the magnetic scalar potential, or monopole.

In an uncurved spacetime, the magnetic charge must always occur in dipolar pairs, where each is equal and opposite the other. In a curved spacetime, a free *net* monopolar magnetic charge may and does exist. In classical electrodynamics, the assumption of a locally flat spacetime is ubiquitous though not prominently emphasized. Hence most electrical engineers are unaware that monopoles can indeed be made, but only if one uses a process to curve local spacetime.

CHARGE-PARITY-TIME (CPT) THEOREM

The theorem states that all known quantum field theories obey CPT symmetry.

C is charge conjugation, which changes the quantum number of every particle into its antiparticle. P is the reflection operation known as parity which turn an object into its mirror image and rotates it 180 degrees about an axis perpendicular to the mirror, T is time reversal. Any violation of time-reversal symmetry is always accompanied by a violation of CP symmetry and vice versa.

Electrodynamics and the strong interaction preserve CP symmetry to great accuracy. Weak interactions do not conserve CP symmetry. T symmetry violation has been shown at CERN, the European particle physics laboratory in Geneva, Switzerland.

CHARGE TRAPPING (Also "Pinning")

Holding charges "pinned" or "trapped" by a barrier or force, so that they do not flow as
i = dq/dt.

CHARGE-BARRIER

A process or component or function—particularly in a semiconductor such as the Fogal semiconductor—which blocks the movement of charges q as current dq/dt.

CHARGE-BLOCKING ASYMMETRIC QUANTUM WELLS

A quantum well which has more charge-blocking action in one direction than the other, and is thus "asymmetric" in its function.

It may also be dynamic, so that it changes in its degree (and/or its direction) of action as a function of time.

CLASSICAL AND QUANTAL ELECTRODYNAMICS

Classical EM (CEM) theory is simply classical electrodynamics (CED).

Classical EM does not break the fields into "quanta" (a quantum is a standard-sized piece of action (angular momentum)]. Instead, CEM considers the fields as continuously varying in magnitude from zero to infinite magnitude (at least in theory).

Quantal Electrodynamics is known as quantum electrodynamics (QED). These are standard terms for standard disciplines. QED considers the fields to be composed of quanta called "photons". Hence there is an underlying region called the "virtual state" whose denizens are "smaller in magnitude" than the quantum. Heisenberg uncertainty applies, since the quantum of action is not quantized energy, but quantized (energy x time). In short, it is a function of both the magnitude of the energy and its duration of persistence. Hence a field or photon of any amount of energy may form temporarily, so long as it does not endure sufficiently long in observer time to reach quantum size and breach the quantum threshold.

In this view, then, "empty space" is not empty at all, but is a seething cauldron of these "energetic bubbles" continually appearing and disappearing, with the entire spectrum of frequencies and energies. This is in fact the "new ether". If we consider only the field portion of the bubbling virtual photons appearing and disappearing, this leads to a vacuum that consists of EM field fluctuations, again of essentially the entire frequency spectrum and energy spectrum.

However, how we attempt to tie either of these "energetic vacuum" EM views to general relativity (GR) is then a problem. The cause of the problem, however, is largely ignored in physics. In CED, e.g., one models EM energy moving through a flat space. This is a non sequitur, since the presence of any change of the energy density of space a priori is a curvature of spacetime. Hence rigorously EM energy can only propagate through a curved spacetime. Indeed, when this correction is made, the EM wave becomes an oscillating curvature of spacetime. This is the approach worked out by Mendel Sachs, extending Einstein's work. The higher symmetry O(3) electrodynamics spearheaded by Evans turns out to be an important subset of Sach's umbrella unified field theory. Hence EM has become GR, and for the first time in history, O(3) electrodynamics provides the approach that will allow direct engineering of general relativity on the lab bench, in working devices, etc.

CLASSICAL ELECTROMAGNETICS (CEM)

"Ordinary" electromagnetics, of the kind used in ordinary EM circuits, which fundamentally consists of the modified Maxwell's equations.

Classical EM consists of Maxwell's equations and the various changes and additions that have since been made. CEM does not break the fields into "quanta" (a quantum is a standard-sized piece of action (angular momentum)]. Instead, CEM considers the fields as continuously varying in magnitude from zero to infinite magnitude (at least in theory).

Primarily refers to Maxwellian electrodynamics as reinterpreted by Heaviside, Gibbs, Hertz, and a few others. Unfortunately these reinterpretations of Maxwell's theory also greatly reduced its topology, and therefore reduced the area of natural electrodynamic phenomena that the resulting equations describe.

Significantly more advanced "classical" electrodynamics emerges when one goes back to the quaternion algebra electrodynamics of Maxwell, or to the more advanced Clifford algebra electrodynamics. The higher symmetry electrodynamics models such as O(3) symmetry EM developed by Evans and others, are capable of modeling a great deal more physical phenomena than the very limited standard U(1) electrodynamics largely in vogue in science.

CLASSICAL EM THEORY

Classical electrodynamics theory, begun by Maxwell, with his seminal paper orally presented in 1864 and published in 1865.

Essentially Maxwell's electrodynamics at root basis, with the background supporting classical theory. Today many variations have been made, particularly with the basic symmetry of the theoretical equations. See discussion above under *classical electromagnetics*.

CLIFFORD ALGEBRA

A special higher topology algebra founded by Clifford, which includes as subsets many other algebras of lower topology.

CLOSED SYSTEM

A system that does not communicate with its environment, and does not exchange energy or matter between system and environment.

An ideal model where dynamically and energetically the system is considered to be "isolated" as if nothing else existed. In short, consider the system as if it were in some "magic box" and nothing outside the box can ever enter the box or affect the system inside the box, and nothing in the system can ever leave the box or affect anything outside the box.

One performs a non sequitur whenever proclaiming treatment of a system as a "closed system", and of course good scientists are aware of this. This is particularly true in electrical power systems. Here one is indeed allowed to input energy into the system to excite or potentialize it. That of course assumes that the system was "opened" long enough to take on excess energy! Then as the system operates, losses and dissipations occur, in which case energy is considered to "escape" from the system permanently. In other words, the system was also continually "opened" so that energy could escape.

There is really no such thing as a truely *closed* system in the universe, since every system is embedded in the active vacuum and is an open system in an energy exchange with the vacuum. If that exchange is symmetrical, then the system is in equilibrium with respect to the vacuum exchange. In that case, the system may be treated as if it were a closed system for a limited number of purposes. In classical electrodynamics (CED), the Lorentz regauging of the Maxwell-Heaviside equations further reduced them to describe only Maxwellian systems in such equilibrium with the active environment. Consequently, the interaction between the external vacuum and the system is not included in the CED model. Obviously a *broken equilibrium* in that actual physical interaction is also not included in the CED model.

An electrical power system is already permitted to violate the Lorentz condition when it
(1) receives additional EM energy to "energize" or "potentialize" or "excite" or "regauge" it, and (2) whenever energy escapes from the system in the loads and losses (which escapes of energy are "de-energizing", "depotentializing", "dissipating", or "regauging" actions.

Hence the Lorentz condition greatly simplifies the Maxwell-Heaviside equations and makes them much easier to solve. However, *when enforced by the system itself with regard to the system's excitation and de-excitation (dissipation) functions—i.e., when those functions are "symmetrical" (equal and opposite)*—it also arbitrarily discards all those Maxwellian systems capable of the five "magical" functions of an open disequilibrium system: (1) self-ordering, (2) self-oscillation or self-rotation, (3) output of more energy than the operator inputs (the excess energy is input by the active environment, (4) powering itself and its load (all the energy is input by the active environment, and (5) exhibiting negentropy.

It follows, however, that physical electrical power systems must therefore have some feature which self-imposes the Lorentz condition, at least during the phase when the system's excitation energy is discharged in the loads and losses. The standard closed current loop system, where every electron progressing through the external circuit's loads and losses must be forcibly rammed back up through the source dipole in the power source, is indeed what enforces the Lorentz condition with respect to excitation and dissipation, and *arbitrarily* prevents the system from achieving COP>1.0.

In a battery-powered system, the otherwise closed current loop is broken into two quite different circuits: (1) the electron current circuit between the outer surface of the plates and the external loads and losses, and (2) the ion current in the electrolyte between the inner surface of the plates. Since the ions (e.g., in a lead-acid battery) in the "inner" ion current may have a m/q ratio very much greater than that of the electrons in the "outer" electron current, clever timing of pulsed signals to the plates may be used to form a phase difference—including even a phase angle of 180°—between the two currents. It can be shown that this forms a high potential on the plates which acts as a negative resistor, extracting EM energy from the vacuum and recharging the battery while simultaneously powering the external circuit. This is the Bedini process, and with Bedini's permission an explanation of that process has been published by Bearden ["Bedini's Method For Forming Negative Resistors In Batteries," **Journal of New Energy**, 5(1), Summer 2000, p. 24-38.] Note that Bedini's process in fact repeatedly removes the Lorentz regauging condition, and therefore the system is rigorously permitted to achieve not only COP>1.0 but also self-powering of itself and its loads and losses.

COLD EXPLOSION

The sudden extraction of EM heat energy from a distant intersection zone (IZ) of the crossing beams of a scalar potential interferometer, which transmits in the sharply pulsed mode, while the electrical grounding of the transmitters are negatively biased with respect to the potential in the distant IZ.

COLD FUSION

An ad hoc term applied to the transmutations at low energy achieved in electrolyte experiments, particularly when using specially prepared palladium electrodes, and particularly when deuterium is present in the electrolyte.

COMPOSITE DIPOLE

Dipole formed as one part of an isolated, observable "charged particle", consisting of a differential element of that observable charge on one end and a momentary clustering virtual charge of opposite sign.

In quantum electrodynamics, it is well-known that virtual charges of opposite signs cluster in the immediate vacuum around any "isolated" observable charge. By applying the composite dipole concept, this author was able to treat the isolated charge as a set of composite dipoles. By then applying Whittaker's decomposition of the scalar potential between the ends of each of these composite dipoles, the "source charge" was a set of broken 3-symmetries in the vacuum flux, and thus a set of 4-symmetries in that flux exchange. This meant that the EM energy continuously output in all directions by the source charge was actually continuously received by the source charge (as a set of composite dipoles) from the time-domain, and 4-conservation of EM energy flow rigorously was upheld.

This allowed a solution to what has been called the most formidable problem in both classical and quantum electrodynamics: the problem of the association of the source charge and its associated fields and potentials along with the energy contained by them and continuously pouring from the charge without any 3-space energy input.

CONCOMITANT

Occurring or existing together.

CONGLOMERATE HIERARCHY OF THE PHOTONS

Every photon remains perfectly ordered! A conglomerate (group) of photons, however, may be arranged in almost an infinite variety of orderings, groupings, directions, etc.—which we refer to as templates and as hierarchies (group orderings) of photons.

In a pure energy flow (energy transport), the "herd" of photons is organized along the flow direction. Regardless of how one scatters or "disorders" this patterning of the "herd" or "hierarchical ordering" of photons, not one single bit of the underlying energy and individual photon ordering is lost. So when one "uses" energy (e.g., scatters the hierarchical ordering of a photon herd from a resistor as heat), one does not lose any energy at all, but simply has "scrambled" the hierarchical ordering of the photons into a new herd configuration (which may be varying every which way!). The point here is the real meaning of the fundamental conservation of energy law: *Energy can neither be created nor destroyed!*

When we "use" one joule of energy to perform one joule of work (say, in a single-pass, scattering), we still have exactly that one joule of energy left, but just in different form (different direction, type, collection or not collected, etc.) It can still do another single pass joule of work, then another, then another, and so on. If we plan things right, we can use a single joule of work to perform many joules of work, contrary to what we were so strongly taught in the *"single-pass, single collection, single dissipation"* examples given to us in university as the "law" of God and electrodynamics.

Note that the conventional work-energy theorem assumes single pass, single collection, and single dissipation of a joule of energy to perform a single joule of work. The point is, *that is not a limitation of nature, but a limitation of a single process! Nature regularly uses a joule of energy to do multiple joules of work, because after one dissipates or scatters the energy the first time, it's still there and later will be intercepted to do some more work.*

Anti-Stokes emission, the Patterson Power Cell, Lawandy's lasing without population inversion, and some gas-filled tubes with anti-Stokes emission are examples of known, validated overunity processes (processes with coefficient of performance greater than unity). These all use multipass, multicollection, multi-dissipation in iterated manner, to increase dipole asymmetry and therefore produce excess extraction of energy from the vacuum. Patterson's unit, e.g., has been independently measured by universities at efficiencies of some 1200 or so. This absolutely does not violate the laws of physics nor the nonequilibrium thermodynamics of open systems far from thermodynamic equilibrium.

CONSCIOUS MIND

The serial processing mind which we normally associate erroneously as ourself. Other to be added.

Mind is time-like, not space-like.

CONSCIOUS MIND (PSYCHOENERGETICS VIEW)

The serial-processing mind which one normally associates erroneously as oneself.

Mind is *time-like*, not *space-like*. Hence it is not directly observable, a priori, since all observation is 3-spatial. However, if we consider that changes in 3-space produce spacetime curvatures and thereby also produce slight changes in the time domain, then iterative coherent spatial changes eventually will produce a change in the time-domain. Consequently, the time-like mind will have a continuing stream of such small but processable time-energy changes from coherent 3-spatial phenomena in its body and the macroscopic (coherent) universe. To sense the 3-space body as "self" while simultaneously sensing other 3-space physical changes as "non-self", the conscious mind must compare the various received time-like projection signals against its "time-like orders" issued. That set of received time-signals that matches or closely matches the conscious mind's issued order signals, thus yields the sense of the body behavioral changes as (1) self, and (2) body responses to its orders. The time delay between issuance of the mind's orders and the receipt of the correlated time-like responses from its body, provides the sense of self-persistence or independent being.

The mind couples to the body via a similar spacetime curvature mechanism. When the mind issues "orders", they are in the time-like realm but are changes, hence small spacetime curvatures. They therefore involve very small 3-spatial changes (projections) into the 3-space body. Since the mind orders are coherent, this small iterative set of 3-space changes will integrate coherently, yielding real quantum changes in the body and nervous system (e.g., the brain).

This coupling mechanism is profound. The steady stream of integrated time-like projections arising in the mind due to the physical world outside one's body, does not compare or match the mind's memory of its orders. Hence that allows the sensing of the "external world" as "nonself". The sense of that external world as "existing" or "being" is provided by the time delay in the comparison process, from "mind signals issued" to "physical signals received". The metric yardstick standard for this comparison time, so to speak, is provided by the time in the identified "self-loop" with the physical body. In other words, the mind uses that length of time to differentiate those time signals which coherently integrate but do not match its issued order signals, from those time-signals which coherently integrate in the same time loop interval and do match its issued ordered signals.

In this manner, we resolve all the great previously unresolved philosophical questions of the nature of consciousness, nation of being, nature of self, nature of nonself, perception of the physical world, perception of body as self, perception of other bodies as non-self, etc. At least we resolve these issues from a testable physics approach, so that we can see that now a unified physics (and a technology) of mind and matter is at last in view. Since O(3) electrodynamics has been shown to be an important subset of Sachs' unified field theory, then O(3) electrodynamics—or SU(2)×SU(2), which is homomorphic with O(3)—can be used to develop the direct engineering of mind, the mind-matter interaction, and mind functions of the living biological system.

It is my considered opinion (with much evidence) that the Russian branch of *energetics*, known as *psychoenergetics*, uses a unified field theory and a higher symmetry electrodynamics to achieve that type of direct engineering already, but for weapons use rather than beneficial use for humanity.

As examples, we strongly suggest that the deaths of U.S. Captains Button, Svoboda, and Hess all involved KGB testing of long-range strategic psychoenergetics weapons over the United States, on those targeted personnel. This should not be surprising at all; the KGB has tested longitudinal EM wave interferometer weapons over the U.S. since 1967, and continues to do so today. They tested bioenergetics EM weapons against U.S. Embassy personnel in Moscow for several decades. Kills of U.S. aircraft and missiles have been performed at various times, including kill of the Arrow DC-8 at Gander, Newfoundland in Dec. 1985, kill of a Titan ICBM fired from Vandenberg AFB in the spring of 1986, kill of the U.S. space shuttle Challenger in the spring of 1986, continuing weather engineering over the U.S. and elsewhere, etc.

That novel electromagnetic weapons have indeed been developed and are being tested world-wide, was confirmed by Defense Secretary Cohen in 1997 as follows:

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts." [Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.].

We specifically call attention to the fact that the Secretary did not speak the words enclosed in square brackets. This is *not* limited just to terrorists or counterterrorism.

There appears to be no insurmountable difficulties in developing psychoenergetic technology using the approach outlined above. Certainly it is not as formidable a task as has been hot fusion.

The incident involving Captain Button, e.g., was to demonstrate control of an individual performing complex technical tasks for a period of at least one hour. The test against Captain Svoboda was to demonstrate how simply human kill can be achieved in combat troops, merely by introducing selected changes in the mental functioning of the targeted individuals. As Svoboda was climbing her A-10 out of a successful low-level ordnance drop at night on a ground target, her sense of up and down was instantly reversed. Instantly perceiving she was diving instead of climbing—although *actually* climbing out perfectly—she "corrected" immediately and dove headlong into the ground, perishing in the resulting explosion of her aircraft.

In the case of Capt. Hess, the test demonstrated that rather complete mental control can be almost instantly asserted, and even the responses of the autonomic nervous system can be overridden. So Captain Hess, out for a morning jog, simply went aside, sat down, and stabbed himself repeatedly, including multiple times right through the heart. Had one seen the incident, he would have simply continued stabbing himself very calmly, without emotion or reflex action from his autonomic nervous system, until physical failure of his body.

As can be seen, it is of the utmost importance that this nation understand psychoenergetics weapons of such nature, and develop effective defenses against them. The next great strategic battlefield may be as much a "virtual reality" battle as it is a "real physical world" combat. The war can be quickly lost in either arena.

We also point out that the psychoenergetics weaponry readily is adapted to use in more advanced longitudinal EM wave interferometers, which readily penetrate the Earth and the ocean. Hence deeply buried strategic command and control centers are vulnerable to psychoenergetics warfare, as are the crewmen of our nuclear submarines, ICBMs, aircraft carriers, division and army command centers, national command centers, strategic bomber pilots, etc.

Indeed, one of the major incentives for developing such psychoenergetics weapons as was demonstrated in controlling Captain Button, was to seize and control the entire on-site manning crews of the defensive quantum potential weapons preventing launch of full strategic warfare. By having the QP systems "stood down for maintenance" by their own crews, the major strategic retaliation and "dead man fuzing" would have been removed and countered. Then powerfully striking the QP sites with LWIs would have destroyed them, and left the sites themselves radioactively emitting longitudinal EM waves, thus preventing reoccupation in case the QP weaponry survived. The counter to that scheme was to place the friendly QP weapon sites on "insane" command and control. In that mode, instead of human decision being made to launch, every so often (say, every two hours) the computer automatically initiates the launch sequence. During that sequence, if the crew is still alive and functional, codes can be punched in and the computer cancels the launch initiation. It then repeats the sequence two hours later. If no coded cancellation order is punched in, for whatever reason, the launch sequence completes and the QP weapons fire. In that case, much of Russia disappears from the face of the earth.

Eerily, just about the time that Secretary Cohen was making his quoted statement in April 1997, the KGB was being notified by the friendly little nation of the change to insane command and control. This forced the abortion of a great strategic attack upon the West, scheduled to occur on May 1, 1997.

We are strongly stressing that psychoenergetics is not parapsychology and it is not just an idle theory of academics. Instead, it is already potently used in some very powerful and deployed weapon systems that could well decide the course of the next world war, almost at a moment's notice.

See also discussion under *energy*.

CONSCIOUSNESS

Refers to a living being's awareness of its sensations, feelings, thoughts, and the world around it.

The new approach uses a coupling mechanism between the time-like mind and the body of a living creature. The nature of this coupling mechanism then generates the mechanism by which self-awareness is produced, awareness of non-self is produced, and the sense of existing in the external world but differing from it is produced.

CONTINUOUS MODE

For a longitudinal EM wave interferometer, a continuous mode of operation where power is fed continually to the transmitters, and the two longitudinal EM wave patterns or beams are continuously transmitted.

CONVENTIONAL SCALAR POTENTIAL

A scalar potential formed without use of artificially added substructuring of its internal biwaves and their products.

In other words, one in which it is assumed that its virtual particle flux has no ordering, but is just a randomized flux. This assumption (e.g., of quantum mechanics indirectly) is incompatible by the Whittaker demonstration that (i) a scalar potential is a harmonic set of "bidirectional" EM phase conjugate wave pairs, and thus not a scalar entity at all, and (ii) perfect ordering of its wave components exists inside the scalar potential.

One must carefully distinguish what each of these internal "phase conjugate EM longitudinal wavepairs" actually represents. Prior to interaction with detecting charges, the potential is a 4-potential existing in spacetime (4-space). The time domain, of course, is represented by ict, and in that expression only "t" is a variable. Hence the "incoming" phase conjugate longitudinal EM waves in the imaginary plane are actually longitudinal EM waves in the time dimension. These waves converge upon the dipolarity represented by the "scalar potential" or any "piece" of it. The dipolarity represents the interaction of charges; specifically, a dipole. This reaction—by the spin of the charges—transduces the incoming EM energy from the time-dimension into 3-space, and emits it from the charges as real EM longitudinal EM energy radiating out from the dipolarity in all directions in 3-space.

The phase conjugate half-set of the Whittaker waves, *prior to their interaction with charge*, are not waves in 3-space, but are waves of LW EM energy coming in from the imaginary plane and thus from the time domain. The charges of the dipolarity spin 720°, being 360° in the complex plane and then 360° in real 3-space. So the charges absorb the EM LW energy impinging on them from the time domain, during their 360° rotation in the complex plane. When they enter real 3-space for the rest of their spin, the already-excited charges dissipate their excitation energy in all directions as they spin their 3609º in 3-space. That is what the Whittaker decomposition actually shows, but it has been ignored for almost a century.

In short, the broken 3-symmetry of the dipolarity—well-known in particle physics for nearly a half-century, but never incorporated into CED—allows nature to jump to a higher, more fundamental 4-symmetry in the energy flow. Energy is being conserved in 4-space, but not in 3-space. Further, there is a special new energy flow symmetry involved: one-way in the time domain and the other way in the 3-space domain. Note that energy is not conserved in 3-space! *There is no law of nature requiring the extra requirement that, in addition to conservation in 4-space, the energy must also be conserved in 3-space!* So when the dipole breaks that extra requirement, only the basic 4-conservation applies. In that case, nature very happily gives you what might be called a white hole in astrophysics terms: energy will pour in to the dipolarity unceasingly from the time-dimension, and will pour out in 3-space unceasingly, so long as that dipolarity persists.

By considering the virtual charges of opposite sign that cluster in the vacuum around any "isolated observable charge", one may represent a single charge (as an electron or proton) as a set of composite dipoles, where each dipole consists of a virtual charge (while it exists) and a differential element of the observable charge. So the charge becomes a set of composite dipoles, each with Whittaker decomposition of the scalar potential between its ends.

Hence we have solved the long-vexing problem that Semiz called "the greatest unsolved problem in electrodynamics": the problem of the source charge (or source dipole) and its associated fields and potentials and the enormous EM energy continually pouring out of any charge or dipole in all directions in 3-space.

When the arbitrarily discarded giant Heaviside nondiverged energy flow component is also accounted, this leads to the giant negentropy of the dipole, as shown in my paper, "Giant Negentropy from the Common Dipole". We have also proposed this "dark Heaviside energy" as the unaccounted source of the excess gravity known to be present in spiral galaxies and holding together their arms. The dark energy yields at least 90% of the total gravity involved, and the additional 10% or so can be accounted for my known gravitational sources.

It appears that these new discoveries can also be utilized to show that "quintessence" is also caused by COP>1.0 processes in the universe, and the excess production of negative "giant dark energy flows". If so, then this also would account for the recently discovered fact that the universe has an excess of negative gravitation, so that it is not only expanding but is accelerating in that expansion. At this time, the present author is still attempting to tie up the loose ends involved in showing that quintessence also is created as a result of functions of the giant negentropy process.

In the Whittaker biwave pairs, If one then insists that these hidden EM waves are comprised of hidden photons, then it follows that the photons exist as continually coupling and uncoupling photon/antiphoton pairs, or continually forming and unforming *gravitons* as the wave and antiwave continually pass through each other.

It is a little more complicated than that, since Maxwell actually omitted half the EM energy in the EM wave in space, and also in the circuit—as do electrodynamicists to this day. For gravitons, one really must come to grips with this error and correct it. This will be added later in this glossary. (To be Added).

COOPER PAIR

A dynamic pairing of electrons in superconductivity theory.

In this pair, if the energy state with wave number σ and spin ½ is occupied by an electron, then so is the state with wave number - σ and spin – ½.

COP (COEFFICIENT OF PERFORMANCE)

Ratio of energy out, divided by that portion of the total energy input that is input by the operator or experimenter.

The COP is thus a measure of "efficiency of using the operator or experimenter's input energy" to produce—or direct the production of—useful work. Note that he can direct the use of his own input energy, as almost all our present systems do. Better yet, he can also direct (gate or "switch") the use of some external free flow of energy from the environment into the system, so that he gets lots more work out of the system than he has to put in himself. Only a few of our systems—such as solar cells, sails on sailboats, waterwheels to power mills, and windmills—are such Maxwell's demons (open systems) where the output exceeds the operator's input of energy.

We sharply contrast the term "COP" to the term "efficiency". The efficiency (overall) of a system is how much useful energy output (usually as work, or can be converted to work at will) the system produces for the total amount of energy that is input (both by the operator and by the external environment). Even a highly inefficient system may nonetheless have a COP>1.0, if the operator inputs less energy than the system outputs.

We urge the reader to thoroughly understand the difference between energy-use "efficiency" of a system and coefficient of performance of the system. Many scientists, engineers, and especially free-energy researchers are confused on the precise difference in these terms.

CORPUS CALLOSUM

The thick nerve cable that links together the two cerebral hemispheres (which are separate brains) in each vertebrate (including human) brain.

In each of the two cerebral hemispheres, there is a separate mind and personality (given that the other is separated). However, a brain has a unique characteristic: Anything that arises directly inside it is automatically assumed to be self-originated, since the brain-tuner can differentiate no separation or "separate source" for the signal/thought arising directly upon its "internal screen". This allows integration of the two brains/minds/personalities into one. That is, when one brain half prepares and sends a message, that message is also routed across the corpus callosum into the other brain half, arising internally upon the second brain's "display." The second brain half therefore thinks that it itself originated the signal or message.

COSMOLOGICAL FEEDBACK PRINCIPLE

Puthoff's self-regenerating cosmological feedback cycle for the source of the vacuum EM ZPE.

One may assume the existence of EM zero-point energy (ZPE) by fiat as part of the boundary conditions of the universe, or conceive of its generation by the quantum-fluctuation motion of charged particles that constitute matter. Puthoff calculated the latter possibility, assuming that the ZPE spectrum (field distribution) drives particle motion, and that the particle motion in turn generates the ZPE spectrum. This provides a self-regenerating cosmological feedback cycle, which in fact is consistent with the general relativity assumption that curvature of spacetime affects mass energy, and mass energy changes affect the curvature of spacetime.

The result of Puthoff's calculation is the appropriate frequency-cubed spectral distribution of the correct order of magnitude. His result is thus consistent with, and indicates, a dynamic-generation process for the ZPE fields.

In general this is also consistent with the Sachs' unified field theory approach, and with the Evans O(3) important subset of it.

COURRIER, ROBERT

Noted French scientist who collaborated with Antoine Prioré in his revolutionary healing experiments on laboratory animals using longitudinal EM waves from plasma tubes.

Courrier was the head of the Biology Section of the French Academy and also the perpetual secretary at the time. He personally presented the astounding results of the Prioré work, showing cures of terminal tumors, to the assembled French Academy, creating a furor.

CROSSTALK

In ordinary science, the transfer of energy or signal from one channel to another, by cross modulation or cross coupling between the channels.

In the new approach, the term "crosstalk" can be used in an additional sense where it refers to virtual energy exchange between orthogonal universes or frames—that is, between different 3- or 4-dimensional slices of an infinite-dimensional universe.

CURL CONCEPTS OF FIELDS

Refers to concepts of fields as the curl of some vector potential;

E.g., the magnitude of the **B**-field is equal to the curl of the **A**-potential, given by $\mathbf{B} = \nabla\times\mathbf{A}$. More exactly, the **B**-field is identically a curled **A**-potential, by $\mathbf{B} \equiv \nabla\times\mathbf{A}$, but only that portion of $\nabla\times A$ that is diverged around or collected by an intercepting unit point static charge.

Note that we are accenting that no equation is a definition. So texts stating that the first equation with an equals sign "defines" the magnetic vector potential **A**, are in error. If we state that $\mathbf{B} \equiv \nabla\times\mathbf{A}$, then we are defining the **B**-field as a curled form of the magnetic vector potential, which it is.

The reader must be aware that the field concept itself in CEM is flawed, being "defined" as the effect existing after an interaction by a 4-space entity with a 3-space mass, and hence
3-spatial and an "effect", and yet then used as if it were the 4-space "cause" entity that propagates in 4-space to interact. In short, the field concept is thoroughly confused in the texts and papers in the literature (and has been confused for more than a century) as not only the cause but the result of the interaction of the cause with some recipient. This problem has been highlighted by many physicists, including Feynman, Wheeler, etc.

This "confusion of the cause with the effect" is widespread in physics, and is the single greatest foundations error which hampers the progress of physics. Another example is the false notion that a "separate force" acts upon a "separate mass", when in fact mass is a component of force by the definition $\mathbf{F} \equiv d/dt(m\mathbf{v})$. That such simple and obvious errors have been propagated by so many hundreds of thousands of scientists for so many decades is perplexing. Indeed, the errors simply continue to be used. Note that Nobelist Feynman and the great John Wheeler pointed out that the "field" as defined in CEM does not and cannot exist in spacetime! Instead, Feynman pointed out that only the potential for the field exists in spacetime, in case some charged mass is brought in to interact with it and *thereby produce* the field as an effect.

Indeed, no observable persists, for it "exists" only as a momentary 3-space frozen snapshot or slice of a 4-space dynamic interaction. Here again, most physicists do not realize that observables such as mass m do not and cannot persist other that at the instantaneous frozen moment of the snapshot.

CURL-FREE MAGNETIC VECTOR POTENTIAL

A field-free magnetic vector potential without curl, hence with no **B**-field.

Toroids and long solenoids have the characteristic of separating the **B**-field (curled potential) and holding it inside. However, drawing energy from any potential whose "source charge or source dipole" is not destroyed, simply results in the full value of the potential being replenished from the vacuum, via the 4-symmetry energy flow mechanism in the presence of the broken 3-symmetry of the dipole. Hence one may easily "double" the potential energy density at will, and this is a special form of gauge freedom, where the potential energy of any electromagnetic system can be altered freely and at will. We strongly accent that, contrary to what is in the EM textbook, gauge freedom rigorously allows the direct and free amplification of potential energy of the EM system at will. That this principle, incorporated in gauge field theory, has been known so long and still not used in the design, production, and use of COP>1.0 electrical power systems is quite inexplicable.

In the motionless electromagnetic generator, we used a special core material to extract the curled potential (**B**-field) from a permanent magnet's vector potential **A**, so that **A** is replenished in uncurled form directly from the vacuum via the giant negentropy process.

To achieve COP>1.0, we then simultaneously interacted both the uncurled **A**-potential energy external to the output coil, and the internal curled **A**-potential energy (**B**-field energy) internal to the core through the coil, with the coil. By pulsing, the Lenz law effect is also invoked as yet another momentary regauging and potential energy increase, so that an additional energy gain is achieved.

The well-known Aharonov-Bohm effect proves that the uncurled **A**-potential can indeed be separated from the **B**-field in like manner, and the **A**-potential does interact with electrons since $d\mathbf{A}/dt = -\mathbf{E}$.

Thus the MEG achieves COP>1.0 permissibly, because it is an open system freely receiving excess energy from the active vacuum environment, via the vacuum's replenishment of the magnetic vector potential of the permanent magnet dipole as fast as energy is drawn from it. The MEG also violates the usual closed current loop's self-imposition of Lorentz symmetrical regauging when discharging its excitation energy (its free excess regauging potential energy), because the "source dipole" (the permanent magnetic dipole) is not destroyed.

CURVED SPACETIME (externally, internally)

A four-dimensional geometry used in general relativity, where the curvature is determined by the distribution of mass-energy.

In the new, more extended approach, any change in the spatial energy density of spacetime or any change in the time-energy density of it, or a combination of changes in both. The difference is that in the new approach time is also treated as a special form of energy. Specifically, time may be regarded as spatial EM energy compressed by a factor of c^2.

CYBORG (An acronym for cybernetic organism).

(Noun): A single integrated system consisting of a living system energetically or functionally linked to mechanical or electromagnetic devices and/or systems performing some vital functions.

(Verb): Link, in the sense of forming such a system.

In the new approach, a living biological system may thus be considered a cyborg of its time-like mindworld (mindframe) to its 3-spatial body (laboratory tuner). The "cyborging" is provided by the personal quantum potential and by the mind-body coupling mechanism. Due to the giant negentropy mechanism of any dipolarity, any scalar potential as a dipolarity is a cyborg between observable reality and virtual state (time-like) reality, i.e., between 3-space and the time-domain.

CYTOMEGALOVIRUS

Any of several members of the herpesvirus family that cause infected cells to enlarge and/or to form cytoplasmic (the protoplasm of a cell, external to the nuclear membrane) inclusions in leukocytes (white blood cells), particularly in their nuclei.

Cytomegalovirus makes a protein that appears to disable the p53 protein. As a result, smooth muscle cells in blood vessels can proliferate until they impede blood flow.

CYTOPATHOGEN

Literally, a pathogen (disease causing organism—or extended, mechanism—) that affects cells and generates damage in them.

CYTOPATHOGENIC

Damaging, harmful, or destructive for cells.

CYTOPATHOGENIC EFFECT

The generation of damage, disease, death, or other symptoms at the cellular level in cells that are exposed to or subjected to a pathogen.

CYTOPATHOGENIC MIRROR EFFECT

Distant EM pathogenic effect noted by Kaznacheyev in thousands of military experiments, where cellular death, disease, and damage forms are shown to be transmitted from one cell culture in the dark (absence of visible light) where the changes are induced, to another similar targeted cell culture in the dark that is environmentally shielded from the first except for light transmission in the infrared and in the ultraviolet. After exposure begins, after a time delay (say, 2 hours or more), the cells in the targeted cell culture will begin to exhibit the same symptoms as the culture having the induced damage. Eventually, in most cases the targeted cells will develop the same damage, disease condition, or death. The effects vary with the time of day etc.

The necessity for photon connection from IR to UV is shown by contrasting a thin quartz window allowing the light transmission with a thin window glass window. The quartz is open from the IR to the UV, and will pass that entire spectrum. With the visible light absent, the IR and UV biophotons emitted by the damaged cell culture will pass through the quartz window into the targeted cell culture. After the time delay, the targeted cell culture starts developing the condition.

Use of a thin window glass window, however, prevents the effect. Normal window glass is opaque in the IR and in the UV, and so it blocks the transmission of the IR and UV biophotons.

The UV is a harmonic of the IR, arranged properly. An unusual effect exists when two sine waves one the first harmonic of the other, are simultaneously transmitted into an isotropic nonlinear medium, such as water. On each sine wave, the peak travels faster in the water than off-peak parts of the wave does. Hence the wave peaks overshoot its middle, the wave collapses, breaks up, etc. Both the waves of the transmitted harmonic set do that. However, if one pretends that one transmitted the difference frequency between them, and not the waves themselves, a strange phenomenon results. Regardless of the breakup of the primary harmonic set, the difference frequency travels through the water as if it were a real frequency in a linear medium. This effect has been proven and is used in underwater sonar equipment.

For example, see Owen Flynn, "Parametric arrays: A new concept for sonar," **Electronic Warfare Magazine**, June 1977, p. 107-112. Any two sine-wave frequencies as simultaneous drivers combine to produce a sine-wave difference frequency propagating in water, essentially without sidebands or reverberations. Its pattern has a main lobe approximately equal to that of the high frequency drive, but devoid of sidelobes. The level of the propagating difference frequency is proportional to both the product of the two fundamental drive levels and to the square of the desired value of difference frequency.

Living cells contain a great percentage of water. Hence the effect should apply to those targeted cells, under the conditions stated.

From general relativity, we know that any pattern and dynamics of mass-energy, such as in a cell and all its parts and dynamics, directly interacts with local spacetime to curve it, and any set of curvatures of spacetime interact with any conglomerate of mass-energy and its parts and dynamics. The set of precise curvatures and their dynamics we call an engine. In a mass embedded in a local region of spacetime containing an engine, the engine will interact with the mass-energy to change it if the mass-energy and its dynamics do not "fit" the engine and its dynamics.

In a diseased, damaged, or dead cell, the mass-energy of that cell and its "template" (form and dynamics) differs from its template while the cell was living and normal. Let us now see what happens when that dead cell emits a biophoton.

A photon emitted from a dielectric such as a damaged cell, is not emitted just from one spot. Every part of the dielectric participates in the emission of that photon. (See "porthole concept", to be added). Further, each emitted photon contains in its scalar potential and its magnetic vector potential, the exact longitudinal EM waveset that constitutes the vacuum engine of the damaged cell that emitted it. When a cell absorbs a photon, it absorbs that vacuum engine carried by the photon also. If it absorbs a stream of photons of harmonic interval, the difference frequency constitutes in terms of a wave, traveling through the cell water as if in a linear medium, and transporting that vacuum engine to every part of the absorbing cell dielectric.

So the biophotons emitted from Kaznacheyev's damaged, diseased, or dying cells sent biophotons containing the exact spacetime curvature engines and dynamics of their precise condition (template). Further, a harmonic interval of such radiation is transmitted into the targeted cells, and absorbed. The mixing of the two absorptions eventually is sufficient to establish the harmonic interval, thus with a clear EM signal difference that travels through the cell's water and to every tiniest part of it. This "difference frequency clear signal" also contains the pure vacuum engines of the template of the changes in the diseased, dying, or damaged emitting cells.

The result is the "kindling" (coherent integration) of the template in the absorbing cell's local spacetime in which it is embedded. As that spacetime engine grows, it differs from the resident engine of the absorbing cell. That difference is itself an engine, and it operates steadily upon that absorbing cell to alter its engine and then its physical condition and state.

As the new engine continues to interact, the physical cell is gradually changed into an equilibrium condition in its mutual interaction with its local spacetime engine. That occurs after the absorbing cell has developed the same disease, damage, or dying condition as the transmitting cells.

This is the secret of the Kaznacheyev experiments. The Kaznacheyev work was allowed to be released in the open literature to stimulate the U.S. scientific and intelligence communities and see whether or not the scientists and analysts had applied a unified field theory and higher symmetry electrodynamics to the EM bioeffects problem. The totally casual reception of the Kaznacheyev experiments in the West, with no interest or interaction from the stimulated communities, assured the KGB conclusively that the West was still in kindergarten in its knowledge of EM bioeffects, and particularly had little or no knowledge in the field of EM biological warfare.

Kaznacheyev referred to the effect as the "cytopathogenic mirror" effect, for the pathogenic effects from the emitting damaged cells being "mirrored' in the targeted cells.

See Vlail Kaznacheyev and L. P. Mikhailova, **Ultraweak Radiation in Intercellular Interactions**, [in Russian], Novosibirsk, 1981. An English exposition of much of the Kaznacheyev work is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," **Psi Research**, 1(1), Mar. 1982, p. 47-76. See also Vlail Kaznacheyev, "Apparent information transfer between two groups of cells," **Psychoenergetic Systems**, 1(1), Dec. 1974; Vlail Kaznacheyev *et al*., "Distant intercellular interactions in a system of two tissue cultures," **Psychoenergetic Systems**, 1(3), Mar. 1976, p. 141-142. In the same issue, see "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144.

Seven or eight nations now have longitudinal EM wave interferometer weapons, capable of beaming right through the earth and ocean to the other side of the planet. At least two of these nations—Russia and China—are capable of adding engine structures to these interferometers, and sweeping a targeted population area to generate that specific disease condition in that target. A sophisticated use would be to have agents (such as terrorist teams already in the United States and with biological agents in hand) release a small quantity of smallpox in the area previously targeted. The LWIs have swept the entire area continuously for a period, slowly kindling (especially at night) the exact form of the disease within the bodies of the populace. A curious thing happens if the level of kindling of the smallpox state has been kept just beneath the observable state, within what is sometimes called the "shadow state" (just barely virtual). A smallpox pandemic will explode in that populace with a ferocity never seen in modern times, and at an unparalleled rate. This is further exacerbated by "spreading the immune system" prior to the attack (term and discussion to be added), which seriously weakens the immune system's ability to resist any single pathogenic attack.

There are hardly any supplies of smallpox vaccine left that are viable. Even where there are, the Russian laboratories developed strains of smallpox resistant to the normal vaccine, and after the collapse of the Russian economy and the former Soviet state, these strains were sold on the black market to the highest bidder. Further, most of the Soviet scientists employed in those secret biowar facilities wound up employed in several other nations, most hostile to the United States.

The point is this: Such a strike could in fact kill up to 2 billion people worldwide. Further, the entire affair would appear to be just a chance breakout of smallpox, and its spread through a nonresistant world populace.

When modern biological warfare is actually analyzed in detail, the "bang for the buck" is greater than it is for nuclear weapons. Further, a rogue or hostile nation can take its time in advance and develop quantities of smallpox vaccine (for the very strain it will use), then initiate such a clandestine campaign after it has vaccinated its own populace, armed forces, and agents. Guess who survives after such an attack.

CYTOTOXIC

Toxic (acts as a poison or toxin) to cells.

A *toxin* is a colloidal proteinaceous poisonous substance that is a specific product of the metabolic activities of a living organism. The toxin is usually very unstable, and is notably toxic when introduced into the tissues.

DARK ENERGY

The Heaviside nondiverged energy flow component, surrounding every field/charge and potential/charge reaction but unaccounted in present electromagnetic theory after being arbitrarily discarded by Lorentz.

DATUM

Consideration, or data considered, or information considered, or action considered, or result considered.

DEATH

The disintegration of the personal quantum potential's connection to the physical body's atomic nuclei by loss of the mind-body connection mechanism.

Simply put, separation of the quantum potential containing the time-like mind-frame from the body. Loss of the preferential tuning of a physical body to its mindworld. This causes loss of the preferential consciousness loop or life channel, resulting in "death" (reversion to inert matter) of the body. Actually the quantum potential and the mind-frame remains, so the *being* is unchanging and indestructible, hence consists of all changes and destructions simultaneously, with none separated individually. Life and death thus are two sides of the same coin, and the coin is immutable, regardless of which side is exposed.

DE BROGLIE WAVE

A matter wave, conceived by Louis de Broglie.

The quantum mechanics wave associated with a particle of matter, which can give rise to interference effects. An orbital electron in an atom is associated with a standing de Broglie wave on a Bohr orbit. De Broglie waves always move faster than the speed of light, the speed being given by c^2/v, where c is the velocity of light in vacuum and v is the velocity of the particle of mass. De Broglie wave velocities thus vary from the speed of light to infinite speed (everywhere at once). In his speech accepting the Nobel Prize, de Broglie emphasized that, since these waves produce real physical effects, they are real and must not be regarded simply as mathematical conveniences.

DEBYE, PETER

Dutch-born American physical chemist and Nobelist, 1884-1966, who performed important and fundamental studies in bipole moments and diffractions of X-rays and electrons in gases.

DEDIFFERENTIATION, CELLULAR

The reversion of a cell or cellular structure to a more primitive or generalized condition.

Dedifferentiation is often preliminary to a major change in the cell or cellular structure. It can also be a reversion of the cell to a previous—even *primeval*—cell form. The important point is that dedifferentiation returns the cell to a previous (past) cellular state. Reversion of a cancer cell, e.g., would cause the cell to return to a normal cell, since that is the nature of its "cellular past." It is vitally important that Becker has demonstrated that cells can be differentiated (moved forward to a future state) or dedifferentiated (moved backward to a previous state) electrically, by very weak DC currents (picoamperes). He was nominated for a Nobel Prize for this epochal work.

Actual electromagnetic redifferentiation of terminal tumors in laboratory animals was repeatedly demonstrated by Prioré et al in the late 60s and early 70s and reported in the peer-reviewed French medical journals, as was reversion of depressed immune systems, cures of infectious diseases, and curing of arteriosclerosis. It is inexplicable that this positively demonstrated cure of cancer—which could also cure infectious diseases and probably can be developed to cure AIDS—was suppressed in the mid-70s and has not been vigorously pursued by present establishment medical science.

DEGENERATE SEMICONDUCTOR

A semiconductor whose conductivity approaches that of a metal.

DELGADO, JOSE M. R.

Noted Spanish neurophysiologist well-known for his research on brain behavioral mechanisms and their control by electrical stimulation; who implanted electrodes in the brains of animals and humans and demonstrated that their responses (both behavioral and emotional) were specific and depended upon the specific localized part of the brain that was stimulated.

A TV tape made by Delgado showed a spectacular demonstration where he stopped a charging bull in mid-stride. With a small hand-held radio transmitter, he transmitted a radio signal to a receiver implanted in the head of the bull. This dramatic demonstration caused a world-wide sensation.

The significance of his work is given by Becker: *"Delgado discovered the exact parts of the brain in which electrical stimulation produced fear, anxiety, pleasure, euphoria, or rage in human subjects. He found that certain sites produced major personality alterations. For example, stimulation of one such site could cause very proper, reserved young ladies to become flirtatious and sexually aggressive. Other sites inhibited maternal or aggressive behavior. In short, Delgado was able to profoundly alter human behavior through electrical stimulation of discrete areas of the brain." … electrodes were implanted in the pleasure centers of rats' brains. The animals were provided with two levers to press. One of these delivered food, and the other produced electrical stimulation of the pleasure center. Invariably, the rats chose to press the lever that stimulated the pleasure center, even to the extent that they died of starvation*." (Becker, **Cross Currents**, p. 221-222).

It is known based on a rare OMNI interview that Delgado progressed to detailed study and experimentation for the behavioral and emotional effects in monkeys of specific magnetic field frequencies on monkeys. These successful experiments involved electrical currents produced in the brain but hundreds of times lower in intensity than those required for direct electrode stimulation. Since then, some information has been cautiously released from Western military studies showing that microwave pulses can couple to the central nervous system and produce stimulation and behavior alteration similar to Delgado's released experimental results.

Comment: The ultimate and most refined techniques in the field of electromagnetic control of human behavior, volition, emotion, etc. is provided by Russian psychoenergetics, the military version of which is totally under KGB control. This work has progressed far beyond what Delgado and Becker were discussion so long ago. The use of the infolded longitudinal EM waves and "engines"—once the technology is developed—allows direct engineering of conscious and unconscious mind, memory including short term and long term, perception, awareness, etc. The additional advantage of using EM longitudinal waves is that mass is largely a very thin dispersion of particles, with (relatively speaking) very large distances between them, somewhat like the solar system but on different scale. This "inner space" is filled with EM waves, potentials, and fields—all of which decompose into sets of longitudinal EM waves and their impressed dynamics. Thus the earth and ocean are readily penetrated by longitudinal EM wave beams, and the Russians (KGB) has developed such LW interferometers since the mid 1950s. However, by LW interferometry where the patterns and dynamics of the carried LWs are engineered, any desired mapping pattern of precise EM signals, frequencies, phases, and magnitudes can be reproduced in distant interference zones, reaching right through the ocean. These interferometers could localize the distant interferometry into a 6-inch diameter volume as early as 1985, so from that time forward it gradually became possible to perform all of Delgado's "inserted electrode" EM stimulus in a single selected human brain from the opposite side of the earth. So approximately 1985 is when I would estimate the beginning of the "golden age" (ugh!) of distant psychoenergetics was initiated.

By 1997, this development would be into at least third generation weapons, just being deployed or just having been deployed, considering an somewhat accelerated weapons development program where 6 years is one generation. By 1997, very complex individual behavioral and perceptual control should have just arrived in on-site deployed strategic weapons systems. See discussion under *conscious mind*.

Some selected useful references are: J.M.R. Delgado et al., **Journal of Anatomy**, vol. 134, 1982, p. 533 reports that developmental defects occur in chick embryos exposed to various ELF frequencies. See also J.M.R. Delgado, **Physical Control of the Mind: Toward a Psychocivilized Society**, Vol. 41, World Perspectives, Harper & Row, New York, 1969.

DEMODULATION

From a carrier containing a modulated signal, recovering the signal that was used to modulate it.

DENSE SIGNAL ENVIRONMENT

For our purposes, when the number of weak EM signals crossing one square meter per second reaches 200,000.

At 500,000 important nonlinear optical type effects of significance—particularly long term cumulative effects—begin to occur in the exposed biological organism.

DEPARTING ORTHOROTATIONS

In space of dimensions greater than four, the rotating of an object or entity away from its line of motion, in the direction of a dimension at right angles to the line of travel in the observer's three-space, by some number of additional 90° rotations.

The Lorentz transform in special relativity, e.g., is just a formula for this kind of rotation. What we call "velocity" is actually a measure of rotation. In 4-space, any object in 3-space can only rotate "out" of 3-space in one direction: toward the time-axis. As the object rotates toward that time-axis, we see the object as having a velocity in 3-space, and its "length along the line of motion" as having diminished. When the object rotates the full 90°, its "length along the line of motion" has been converted to "time" rather than length. Hence it has zero length along the line of motion and is moving at c, the speed of light in ambient space. It is also no longer material, but is a sort of flat plane surface moving at speed c in a direction perpendicular to the plane.

In that 4-space model, it is impossible for anything to move faster than light speed. So Einstein's postulate dealing with the fact that it is impossible to translate mass, signals, etc. faster than light speed has a built-in assumption that only one orthogonal departing rotation can be made from 3-space. That postulate need not apply at all if more than one departing orthorotation is permitted—which of course requires a space of 5 dimensions or greater.

The notion is that, if we allow a spacetime of greater than four dimensions, by continuing to rotate ever more orthoturns in a departing direction away from the original line of travel, the rotated object loses one observable dimension for every orthorotation. Hence with three "departing" orthorotations it becomes a "nonobservable point" to the normal observer, but one that would be moving at speed c^3. A single orthorotation would make the 3-dimensional object a two-dimensional plane moving at speed c.

No matter how many rotations are made in a departing manner, the object is still connected to the time dimension common to all dimensions (assuming only a single time dimension for simplicity). So it exists in an orthorotated space, but in the same time. To the physical observer in 3-space, it thus appears to be a "time-like" or "mental" object only. Nonetheless, by affecting the density and structuring of the time domain, it can have an effect on a 3-spatial object. We discuss that mechanism under the subjects of the mind-to-body coupling mechanism and the body-to-mind coupling mechanism. Further, with the Sachs-Evans unified field theory approach, the time-domain can be directly engineered. This means that both "inert" or physical reality and time-like or "mental" reality can be directly engineered by novel electrodynamic means.

In short, a photon is just a particle (such as an electron) that is orthorotated once (toward the time axis). In photon emission, a little bit of the mass-energy turns into orthorotated photon energy. In photon absorption, the absorbed photon energy turns back into mass-energy.

The higher dimensional orthorotational concept was originally formed to allow a physical modeling grasp on a "mental" object as compared to a "physical" object, and to have a connection mechanism between the two. Such an approach can be used to model both mind and matter, their interaction, etc. The present author has used it to solve the old philosophical problem of "intent", of how the perception of self is attained, how the external world is perceived separate from self, etc. A model for both living and nonliving states does emerge.

This or a very similar approach has been used by the Russian's highly classified weapon science called *energetics*. That third branch of energetics is called *psychoenergetics* and it is the physics of directly engineering mind states, mind operations, memory, etc. by extended electrodynamic means. We have previously detailed much of this work privately, but with some detailed openly in a few papers. E.g., see T. E. Bearden, "Mind Control and EM Wave Polarization Transductions, Part I", **Explore**, 9(2), 1999, p. 59; Part II, **Explore**, 9(3), 1999, p. 61; Part III, **Explore**, 9(4,5), 1999, p. 100-108.

DESTRUCTIVE INTERFERENCE

When two waves or wavesets from two different sources superpose to produce a combined waveset that has a lower intensity than the sum of the intensities of the two original waves.

DETERMINISTIC PATTERN OR TEMPLATE

The exact arrangement, by plan and deterministic action, of a group of entities.

E.g., of the photons in a "herd" or "group" of them. Or a set of spacetime curvatures, into a dynamic structure called a "spacetime curvature engine"—or "engine" for short. Or by the virtual photons in the vacuum flux, or a deterministic set of changes to that vacuum flux (the vacuum flux that is called the "vacuum potential." I have called this function—of a deterministic patterning or templating of the virtual flux (particle view) of the potential—*dimensioning* the potential.

DIELECTRIC

Literally, "against electric (current),"

...originally coined to mean "opposes the flow or conduct of electric fluid."

In today's language, a dielectric is a non-conducting material used to oppose or prevent the flow of electric charges. Even empty vacuum exhibits some dielectric resistance. "Dielectric" is also used to refer to a nonconducting material (often used between the metal plates of a capacitor or elsewhere) which can (i) sustain a change in potential across a distance (i.e., an **E**-field), and (ii) serve as an insulator. Space itself is a dielectric, or at least it is said to exhibit certain dielectric qualities.

DIFFERENTIATION, CELLULAR

(1) specialization of body parts or organs in the evolutionary process. (2) modification of different parts of the body to perform particular functions. (3) modification of the cells forming these body parts or organs so as to perform the necessary cellular functions required to support the functioning of the body part or organ.

DIFFRACTION

The bending of light at the edge of an opaque object.

Usually diffraction introduces undesired distortion and is a problem. However, it can be useful and desirable—for example, in a diffraction grating.

DIMENSION

A certain primary geometrical physical attribute, such as length, used to describe the separational relationships of physical phenomena.

The degree of that separation is known as "length", "time", etc. By "geometrical" we mean that the "dimension" is considered to "exist" in either the presence or absence of the observable physical phenomenon; however, by understood agreement the dimension itself is not directly observable. It can be inferred ("measured"), however, by comparative operations of simple devices such as clocks and meter sticks, and its simplistic definition is usually given as "that which has measurable magnitude." Prescribed measuring procedures are used to assign magnitudes to these inferred dimensions, using specific scales.

Actually, "dimension" just means "fundamental quantity of separation" and "geometry" in the most general sense. The number of dimensions taken in advanced physics usually represents a somewhat arbitrary choice for good modeling fit. In modern Kaluza-Klein theory, for example, it becomes necessary to accept some 10 or 11 dimensions in the usual particle physics case. General relativity has incorporated the interaction of the abstract geometry (spacetime) with mass. Until Sachs' extension of Einstein's work into a unified field theory and Evans' incorporation of O(3) electrodynamics as an important subset of Sachs' unified theory, general relativity has largely remained a nonexperimental discipline, at least in the laboratory. With the new Sachs-Evans approach, however, GR becomes substantially engineerable by novel electromagnetic means, including in the laboratory and in devices.

DIMENSIONING

This is an artificial term I sometimes use to mean "forming an exact pattern or 'template', particularly of nested curvatures of spacetime, where the pattern/template may also be dynamic. It involves the internal structuring of the "dimensioned" entity.

The use of the term "dimensioning" makes sense in unified field theory, but is awkward in normal EM theory which erroneously considers that EM energy propagates in an uncurved spacetime. The mere presence of a change in local spatial energy, created by the wave energy itself when present in that local region, curves the local spacetime. Hence the EM wave rigorously travels in curved spacetime, and in fact identically is a propagating, oscillating curvature of spacetime.

Dimensioning also implies a deliberate form or structure (with dynamics) created in the virtual photon flux of the vacuum or of the virtual photon flux comprising a scalar potential, or in the infolded longitudinal wave EM inside all EM potentials, fields, and waves. It implies an associated set of spacetime curvatures and their dynamics. The latter are referred to as "engines", "vacuum engines", "spacetime curvature engines", etc.

Dimensioning potentials or signals or carrier waves also involves adding hidden *vacuum engines* to them. I chose the descriptive term *dimensioning* because adding such templates (hidden vacuum engine structures) is equivalent to increasing the EM topology by adding dimensions. Since an EM wave or field or vector potential can be decomposed into two scalar potentials, then by "dimensioning" one or both of the scalar potentials one can "infold" the desired hidden structures or patterns (vacuum engines) inside an EM carrier wave, EM field, or vector potential. So I called this *dimensioning* those entities. Internally structuring the two scalar potentials via arrays of EM LW emitters, and then applying the necessary dynamics (modulations), is primarily the manner in which dimensioning can be directly engineered. When dimensioning is present, two identical EM waves as seen on the oscilloscope may have dramatically differing internal structures (dimensioning), and hence the two will produce drastically differing effects when they interact with the same or similar objects. This of course is the mechanism involved in the Kaznacheyev experiments and in the decades-long microwave radiation of the U.S. Embassy in Moscow, where health changes and diseases were electromagnetically induced in personnel in field-free (stable and nonchanging potentials) areas.

Russian energetics refers to what we are calling "dimensioning" simply as adding *the information content of the field*. A very skilled disinformation program (actually, a strategic deception program) has been used to deceive the West into believing the phrase refers to ordinary spectral analysis. It doesn't. The importance of the dimensioning concept is that one has not formed mere "signals," instead, one has formed *vacuum engines*. If a receiving object receives a signal, it may or may not act upon that "signal information." If it receives a *vacuum engine*, however, the engine (structured local spacetime flux) directly acts upon it to make the requisite change. Further, the energy for the interaction's working upon the exposed object or system is furnished continually from the curved spacetime set and their dynamics. The receiving object has absolutely no choice in that action occurring to it, and the action can occur from the gluons and quarks inside the charges in the nuclei, to the lattice structures, to the system dynamics, to the elements so as to transmute them, etc.

Note that when a dimensioned EM field or potential is absorbed by an object, that object also has internal EM fields and potentials which have substructures (usually rather random). However, the internal structure of the absorbed dimensioned entity and the internal structures of the EM fields, potentials, and waves do mix in a direct diffusion process. Hence with continued irradiation, the target can be "charged up" with the desired internal dimensioning structure. This is how the microwave irradiation of the U.S. Embassy accomplished disease induction, using signals with the desired disease dimensioning impressed in their structures, particularly in the potentials. In the Gulf War Disease, a much more sophisticated use of this mechanism was made, using a quantum potential as the carrier. All the effects (including the body fluid transmission, the EM fields of the body transmitting and kindling some of the effects in children in proximity to a dimensioned veteran parent, can be explained.

The use of dimensioned EM fields, waves, and potentials—and the use of dimensioned quantum potentials in Bohm's hidden variable theory—allows an entirely new science of EM biological warfare, far more threatening and lethal than the more primitive physical pathogenic kind. It is also possible to use the techniques for revolutionary healing, as unwittingly shown by Prioré and the scientists working with him, and also as shown by Becker *et al*.

The only way in which the action of a vacuum engine (dimensioned carrier entity) can be resisted is to create and simultaneously introduce the exact antiengine for it.

Once the action has been completed in the receiving object and it has been changed, the only way it usually can be undone is to form a precise antiengine for that action, amplify the antiengine, and let the antiengine act upon the object for a sufficient time to "time reverse it back to its previous physical state and condition". The Prioré mechanism utilized that method for reversing the cellular damage due to cancer and infectious diseases, and to restore suppressed immune systems. The same mechanism, if re-developed, could also be used to reverse aging in living human patients.

The engine and antiengine processes can also be applied by structuring the "time-charging" or "time-excitation charging" of the particles of mass in a body. The "port-hole" concept even allows ordinary EM radiation to be used to forcibly transduce the input ordinary EM into the extraordinary EM and thereby forcibly accomplish the necessary time-charging actions for disease and body deterioration reversal. See discussions under "port hole concept" and "time-charging".

DIPOLE

In the simplest case, two separated charges of (usually equal and) opposite sign.

As a general definition, a localized positive charge or charge distribution, and a localized negative charge or charge distribution, such that the net charge summation is zero, but wherein the positive and negative charge distributions do not superpose. We point out, however, that in regular physics *electrical charge* has no proper definition. Try looking it up in several physics dictionaries! Then check our proposed definition of electrical charge. As is well-known in particle physics, a dipole is a broken symmetry in the virtual photon flux of the vacuum. By definition of broken symmetry, this means that *some* of the virtual energy flux continuously absorbed from the seething vacuum by the charges of the dipole, is not re--radiated as virtual energy. Instead, it is integrated coherently and re-radiated in 3-space as real emitted EM energy. The dipole is therefore a true "negative resistor" since it receives EM energy in unusable form, and re-emits it in usable form. In classical EM theory, the problem of how every charge and dipole can continuously emit EM energy to form all the associated fields and potentials and their energy, is an unsolved problem—sometimes called the most pressing problem in electrodynamics, both classical and quantal. For the solution and a discussion, see my paper "Giant Negentropy from the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23.

Because of the broken 3-symmetry of the dipole, the EM energy flow is not conserved in three dimensions (symmetry represents a conservation law, and broken symmetry represents the violation of that conservation law—in this case, violation of conservation of 3-space EM energy flow). Instead, EM energy flow is conserved in 4-dimensions, and the hidden inflow of EM energy to the dipole is actually from the complex plane (time domain). Thus all EM energy is actually received from the time-domain, and so the importance can be seen of the internal structuring of time-flow and of longitudinal EM waves in the time domain. Such time-domain energy flows appear naturally in O(3) electrodynamics; see M. W. Evans and T. E. Bearden, "The Most General form of the Vector Potential in Electrodynamics," **Optik** (in review).

DIRAC, PAUL A. M.

Noted English physicist, 1902-1984, who shared the Nobel Prize for developing wave mechanics and predicting the existence of the positron.

DIRAC SEA

Considering the vacuum as containing—amongst other things— a "sea" of an infinite negative energy conditions or states ("holes") for electrons, almost all of which are filled with electrons.

In Dirac electron theory, the relativistic wave equation for the electron has four components, which correspond to two spin orientations and two energy conditions: positive and negative. Thus the vacuum possesses a "sea" of negative energy conditions for electrons. These are called "holes" and are considered to ordinarily be filled or mostly filled with electrons which have fallen into these negative energy states. The Dirac sea thus is this vacuum sea of negative energy electrons. In forward time, if energy is added, electrons can be lifted from this sea. If spacetime is locally curved, additional holes may be created so that normal positive energy electrons fall in and disappear, creating the appearance of an "electrical energy sink." If the curvature is in the opposite direction, electrons may be lifted from the Dirac sea by the curvature, thus producing a flow of electrons and creating the appearance of an "electrical energy source."

DISCRETIZED

Occurring in a finite size, fixed as a function of one or more variables.

So long as the same value(s) of the variable(s) are selected, the discretized entity is always the same size. However, if the value(s) of the determining variables are changed, the value of the discretized quantity may change so that it is not a whole number multiple of the former entity. Compare this to a quantized entity, where the values of the variable(s) are held to those wherein the entity will always be a multiple of an entity of fixed size.

DISEQUILIBRIUM

Broken equilibrium, which in turn involves breaking appropriate conservation laws.

DISSOCIATION

Separation of ideas from their affects or feelings, resulting in independent functioning of these components of a person's mental processes.

Simply put, fragmentation of the mind and personality. Severe dissociation is the definition of insanity.

DISTORTION CORRECTION THEOREM

"If a scalar wave E1(r) propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica E2(r), then E2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E1."

Quoted from Amnon Yariv, **Optical Electronics**, 3rd Edn., Holt, Rinehart and Winston, New York, 1985, p. 500-501.

What is missing from the distortion correction theorem is hidden in the phrase *"if we generate in some region of space ... its phase conjugate replica.."*. Since electrodynamicists and optical physicists have been unaware that all EM energy at any point in space comes from the time domain to that point, then the actual phase conjugate wave in the imaginary plane (from the time domain) is not utilized as the phase conjugate replica. Instead, the interaction of that wave with charges, which actually produces the force-field wave in space, is assumed. In short, one is looking at the "effect" wave, not the "cause" wave. However, since time is not an observable even in theory, it is this "effect" wave in 3-space that will be measured by our instruments.

Note also that the distortion correction theorem says nothing about longitudinal EM waves in the time domain. Interestingly, in nonlinear optical pumping, it appears that the 3-space energy of the pumping 3-space waves is transformed into the time domain by the pumping, and thence flows from the time domain into the 3-space domain to form the progression of the time-reversed replica wave in 3-space.

It is apparent that much more work should be done upon the foundations of phase conjugate optics, in a higher symmetry electrodynamics model such as O(3) electrodynamics and particularly in the unified field theory as in the Sachs-Evans combined approach.

DNA (DEOXYRIBONUCLEIC ACID)

Nucleic acid that constitutes the genetic material of cellular organisms and also certain viruses.

DRUDE, PAUL

Important German physicist, 1863-1906, who formulated the theory of the free electron gas inside conductors, using it to formulate a theory of metallic resistance.

DRUDE ELECTRON GAS

The "free" electrons that have broken loose from their atoms in a conductor, and are thus free to "wander around" and "migrate"—or flow as current when exposed to an emf.

Also called the electron gas, or Drude electron gas. Named after the German physicist Paul Drude who died in 1906, who formulated Drude's theory of metallic resistance based on EM theory and his own theory of electrons. See P. Drude, Ann. Physik, Vol. 1, p. 566; Vol. 3, 1900, p. 370, 869 for the original Drude papers.

DUALITY PRINCIPLE

In quantum mechanics, the principle that photons and other particles propagate like waves and interact like particles, so that their descriptions as only waves or only particles is inadequate.

ECM

Electronic Countermeasures.

Electromagnetic measures taken against targeted electronics or electromagnetics operations in systems, particularly in hostile weapon systems, command and control systems, etc. There are entire textbooks in the West that are devoted to ECM effects, uses, effectiveness, etc. A great deal of ECM equipment is also used by any modern military force. However, in our view what is greatly lacking as yet in Western ECM is the deliberate design and use of novel new kinds of ECM based on the extended higher electrodynamics such as O(3) electrodynamics, and particularly by use of longitudinal EM waves and by use of the infolded electrodynamics inside all EM fields, potentials, and waves. Longitudinal EM wave interferometers (LWIs), e.g., can reach right through the earth and the ocean to emerge on the other side of the earth. In the interference zone (IZ), real EM "transverse wave" energy is created readily by the interferometry. Biasing of the chassis grounds of the distant LWI transmitters can control whether the "normal" EM energy appearing in the target IZ is convergent (cooling) or divergent (heating). By sharp pulsing of the LWI transmitters, massive cold explosions or hot explosions can be created in the IZ. Continuous LWI interferometry can produce geometric forms of EM energy, which usually glow due to the ionization of the atmosphere, etc. in the dense EM energy regions. Since these can also involve general relativistic effects as well as electromagnetics effects in the U(1) CED sense, obviously a dramatic expansion of the entire field of measures and countermeasures is entailed.

We believe these LWI weapons may be what was referred to by U.S. Defense Secretary Cohen as follows: *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our efforts."* Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

EDGE WAVES

In electromagnetics, waves that form parallel to a material edge.

EDISON, THOMAS ALVA

Noted American inventor, 1847-1900, whose staggering number of inventions included the storage battery, electric light bulb, phonograph, motion pictures, and carbon microphone.

E-FIELD (ELECTRIC FIELD)

The electric field or electric intensity E. That field on a spinning charged mass which produces an electrical force resulting in a linear motion of the charge.

See also discussion under c, above.

E-FIELD, MOTIONAL

In classical EM theory, the electric field produced by moving a conductor in a magnetic field. Often called *motional inductance* of the field.

In the orthodox theory, motional inductance is deemed to be identical to the inductance caused by a time rate of change of the magnetic field, experienced by the moving conductor.

E-FIELD, STATIC

In orthodox theory, the electric field in the space surrounding a static charged particle, induced by the presence of the charged particle.

In classical EM, there is no notion as to the *causative mechanism* for this static field of and from a charge, which implicitly is assumed to just continuously create and pour out EM energy in all directions, thus grossly violating energy conservation.. Consideration of this anomaly leads to what has been called the most pressing problem in both quantal and classical electrodynamics: the question of the formation of the fields and potentials associated with the "source charge".

For a full solution to this problem, see my "Giant Negentropy from the Common Dipole," **Journal of New Energy**, 5(1), Summer 2000, p. 11-23. Also on this website.

EHRENHAFT, FELIX

Scientist who performed numerous experiments believed to demonstrate fractional charges.

EINSTEIN, ALBERT

German (Jewish) physicist, 1879-1955, who emigrated to the U.S. and formed special and general relativity theory, discovered and formed the theory of Brownian motion, recommended the development of the atomic bomb to President Roosevelt, and profoundly affected the course of science and the modern world.

EINSTEIN-de HAAS EFFECT

The rotation observed in a freely hanging ferromagnet when it becomes magnetized.

It has been used in the Einstein-de Haas method to determine the gyromagnetic ratio of a ferromagnetic material by suspending a cylinder of the material from a torsion fiber and measuring the angular displacement of the cylinder when its magnetization is reversed.

EINSTEIN, PODOLSKY, AND ROSEN PARADOX

Thought experiment in 1935 for measuring two correlated photons separated and at distant locations, protesting that quantum mechanics requires that the particles communicate instantly, which is a "spooky action-at-a distance" (Einstein's phrase).

The experiment eventually was performed and it works, just as required by quantum mechanics. The paradox inspired the development of hidden-variable theories.

EINSTEIN'S POSTULATES

(1) All the laws of physics are equally valid in all inertial frames of reference,

(2) the speed of light is the same to every inertial observer, and

(3) the observable local effects of a gravitational field are indistinguishable from those arising from acceleration of the frame of reference.

The first is called the special relativity principle, the second is called the law of light propagation, and the third is called the equivalence principle. It is now known that postulates (1) and (2) are not independent of each other. Note that the third postulate implies that any local force is due to a "gravitational field" (acceleration of the frame of reference). It also implies that gravitational force has a mechanism, since quantum mechanics assigns a differential operator, operating upon a potential, to create every force. By extended inference, it can be seen that the third postulate also implies that there should exist a method to turn any type of force (field) into gravitational force (field).

The Russian physicist Sakharov has strongly postulated that gravitation is not even a fundamental field of physics, of the kind used by Maxwell in his electrodynamics. Instead, in this view **G**-field is always comprised of, and made from, other fields and interactions upon them. Indeed, such has always been implicit in general relativity, it just took Sakharov to formally propose it. E. g., see A. D. Sakharov, "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," **Soviet Physics Doklady**, Vol. 12, No. 11, 1968, p. 1040-1041 [English translation.] Sakharov's article (in Russian) in **Doklady. Akad. Nauk SSSR**, Vol. 177, 1967, p. 70-71.

Now let us do a little "tinkering" with Einstein's postulates. For the first one, we point out that in the new approach using dimensioning and subspaces, the notion of an inertial frame now has acquired additional complexity! The "gross translation" effects—which tends to be what physics ultimately focuses upon—can remain the same, and yet vacuum engines can be buried up inside the inertial frame to act upon a mass in non-translating ways. Everything translation-wise can remain the same, but now matter can be transmuted and transformed, etc., still right there in an inertial frame. So when the infolded electrodynamics inside all fields, potentials, and waves are utilized, the notion of "inertial frame" has acquired a higher topology and then one inertial frame may differ very substantially from another inertial frame. This of course move one from special relativity into full general relativity, and even there into a slight extension. Special relativity can remain special relativity on the surface, and yet effects occur "within" the objects postulated to exist in this inertial frame. The "inertial frame" concept in the postulates is now just a sort of special "equilibrium case" of a higher topology "inertial frame" notion.

The second postulate now stands as a "special case" of a far more general second postulate. It actually defines (and limits) "light" to bulk changes in the entire vacuum potential magnitude. It only applies to light signals which are "bulk upheaval" disturbances of the vacuum potential (spacetime). Specifically it need not apply to the "inner EM/GR realm" of longitudinal EM waves infolded inside the vacuum potential, other potentials, signal carriers, etc. Superluminal communication is now included in this "hidden variable" communication. The Fogal semiconductor is in fact capable of infolding and outfolding EM signals, including producing the internal longitudinal EM waves, and during their infolded state these signals are permitted to move at superluminal velocity.

So the second postulate needs to be restated to include the present statement as a special case. Notice that the concepts "light" and "speed of light" have now taken on much richer, extended meanings, since there is "outfolded" light and "infolded" light, so to speak. Light now is not just the "external perturbation wave" in the ambient vacuum potential medium (in flat spacetime). Instead, it can also be an "internal perturbation longitudinal wave" inside a "net flat spacetime" which does not show on the "surface" at all. Indeed, it may move in the time dimension or in 3-space. It now includes the transmission of an extraordinary "vacuum engine" as well as the usual transmission of just an ordinary signal.

And these infolded vacuum engines can be visualized as very special sets of longitudinal EM waves and their dynamics, consistent with the Whittaker decomposition. His U(1) symmetry electrodynamics decomposition—particularly when enriched in the higher symmetry O(3) electrodynamics—applies to the ambient vacuum as a potential (it is a potential since it has energy density). So the "subspace" inside spacetime, so to speak, is just a set of LWs and their dynamics, and that subspace itself is engineerable, just as is the "bulk curvature of spacetime" itself. Further, we may now consider space itself is just a vast superhighway for such superluminal LWs and their dynamics. Mass is mostly empty space with only an occasional particle here and there and with fields, potentials, and waves in between. Hence huge masses such as the ocean and earth are actually superhighways for the direct passage of longitudinal EM waves, "engines", etc. The degree of "transparency" is a function of the degree of cleanness of the LWs.

The third postulate has also been dramatically extended and requires revision. As presently stated, antigravity effects do not appear to be included in present physics heretofore. The "gravitational effects and external accelerated effects" notion in conventional physics is still founded purely on *ordinary bulk translation*. Now it need not be. For example, the system can appear to be an ordinary system, as judged by normal "external light" observation. Yet its internal vacuum engines can cause local violation of most so-called "laws of nature." The reason a "law of nature" can be violated is that one is now able to violate the premises upon which it is postulated, and which are required to hold if the law is to hold. Or another way to look at it: any conservation law is the result of a symmetry. Breaking that symmetry allows violation of that specific conservation law.

So as we've said so many times before, the infolded EM inside the "conventional" EM fields, waves, and potentials—and inside the vacuum potential and spacetime itself—dramatically changes the present theories of electromagnetism, quantum mechanics, and general relativity. Further, for the first time the Sachs unified field theory allows an *engineering* approach which simultaneously crosses, includes, and extends all three disciplines in the required manner. The approach can be placed into good engineering models based on the Sachs work, particularly using the O(3) electrodynamics subset spearheaded by Evans, and the models can be tested and verified or falsified.

EINSTEIN'S GENERAL THEORY OF RELATIVITY

Mathematical theory of gravitation in which the gravitational force is mathematically described by a curvature in space or spacetime.

In other words, the geometry and its dynamics have become physical agents.

It is important to note that, in general relativity, the curvature of spacetime acts on mass-energy, and changes in or to mass-energy interact back upon spacetime to change its curvature.

Further, we consider that an exact pattern of curvatures of spacetime and their dynamics exists for every mass system and its dynamics. This is the engine concept. The most general way to engineer matter is not by brute force translation, but by forming spacetime engines. Once formed, these engines act upon any mass placed in the engine region, to eliminate the deltas existing in that mass's resident engine and the impressed engine. The curvature of spacetime itself furnishes the energy for the continuing action upon the mass, once we pay to form the engine. The mass can be acted upon at any level, from its gluons and quarks to its nuclei, to its atoms, its lattice forces and binding energy, etc. Significantly, the Sachs theory applies to physical reality from beneath the quarks and gluons to the entire universe, and folds in quantum mechanics, electrodynamics, and general relativity into a single unified field theory that is (in theory) engineerable by higher symmetry electrodynamics means.

ELECTRIC FIELD

The *conventional* definition is this: "1. a region in space in which a stationary electric charge experiences a force due to its charge. 2. the electric force per unit test charge." **Dictionary of Science and Technology**, ibid., p. 720. That definition is flawed. Consider the "spatial aspects" of part 1. If an electric field is defined as a force per unit charge, it is not a region in massless space but is an effect of something existing in massless space and interacting with the charged mass. The electric field *prior to interaction* is indeed related to a region in space, because it is identically a curved spacetime region. In the Sachs approach, general relativity and electrodynamics have become one and the same. So prior to interaction, the *field as it independently exists prior to interaction with charge* has become a curved spacetime region, and the field is *purely* a spacetime curvature effect. The interaction with mass to produce a force on and of that mass (mass is a component of force, and there is no such thing as a massless force) is a separate consideration from the field as it exists in space prior to interaction. The result of this entity's interaction with charged mass has nothing to do with *what an electric field is*; but with *what it does to charged mass*. Further, the force evidenced on a charged particle is not due to its charge; it has that charge always, whether or not there is a net force evidenced on it!

Part 2 is a little better. The only problem is, it totally disagrees with the entity existing in space prior to interaction. In other words, here it's a "force on a charge", which is an effect occurring after the interaction. Again, it is what the field *does or accomplishes*, not what it *is*. Hopefully the reader can now begin to see what terrible difficulties await anyone who seriously attempts *foundations definitions*. Feynman sums it up admirably in one of his famous statements: "*Everything we know is only some kind of approximation, therefore, things must be learned only to be unlearned again or, more likely, corrected."* An additional difficulty is that the very notion of a separate massless force independent of the mass on which it operates, is fictitious, yet continues to be widely used. Recalling that any true definition is an identity and not an equation, let us start with the definition of force:

$\mathbf{F} \equiv d/dt(m\mathbf{v})$. As can be seen, force is a system comprised of (i) a mass component and (ii) a nonmass component. The force is not something separate from the mass, at all! Further, force is not even unitary. It's a system comprised of two coupled components; the mass and a coupled (interacting) gradient in a potential flux. *Note that **v** involves motion through, and interaction with, the virtual flux comprising the vacuum potential. Note that d**v**/dt involves a gradient in that flux interaction with the mass*. Unfortunately mechanics is still using the medieval concept of a "separate force acting upon a separate mass." That is wrong, unless we change the fundamental definition of force itself. See, e.g., Robert Bruce Lindsay and Henry Margenau, **Foundations of Physics**, Dover Publications, New York, 1963, p. 283, where the authors emphasize that a "field of force" at any point is actually defined only for the case when a unit mass is present at that point. The reason can be seen: mass is a necessary component of force, and without mass present there cannot be a force present.

ELECTRIC DIPOLE

See Dipole.

ELECTRODYNAMICIST

A specialist in electrodynamics,

...the physics of the relationships between electric, magnetic, and mechanical phenomena—and especially a specialist in electrodynamics theory.

It should be pointed out that essentially all electrodynamicists are versed in U(1) electrodynamics, and a much smaller number are versed in more advanced electrodynamics such as O(3) symmetry electrodynamics, unified field theory electrodynamics, non-Abelian electrodynamics, etc. The electrodynamics field is actually a very vast field indeed, touching everything in physics. Very often, when one needs a particular electrodynamics problem solved, it calls for a special category of electrodynamicist and not just an electrodynamicist in general. One must therefore sometimes examine the background of the electrodynamicist making a statement or a pronouncement: electrodynamicists after all do have differing views and differing disciplines within the field.

ELECTROMAGNETIC PULSE (EMP)

A sharp pulse of electromagnetic radiation produced when an explosion occurs in an unsymmetrical environment, especially at or near the earth's surface or at high altitudes.

Essentially the same as an electromagnetic shock wave, a wave of great intensity which results when waves with different intensities propagate with different velocities in a nonlinear optical medium. An EMP is also produced when two intense longitudinal scalar electrostatic potential wave patterns meet and couple into a sudden flash of vector EM energy.

ELECTROMAGNETIC SMOG

The background of extremely weak EM radiation that fills the normal modern environment, particularly in developed nations.

The individual signals are too weak to individually produce effects of any significance in biological systems. However, in a dense EM smog environment with 500,000 signals per square meter per second, the signal complex and its nonlinearity begin to produce highly nonlinear effects in biological systems. These effects also produce damage that may cumulate over a long period of time, including for years. This bioeffects area has largely not been investigated by EM bioeffects scientists.

It should be pointed out that electric lines, telephone lines, transmission lines, etc. act as long wire antennas to detect and propagate such very weak EM signal noise. Essentially all the radiators in the environment—including electronics equipment, TV stations, radio stations, factory machinery, etc.—thus produce very weak signals picked up by these long line antennas and piped into every home, factory, neighborhood, etc.

ELECTROMOTIVE FORCE (EMF)

A difference of potential created by sources of electrical energy—i.e., by separated charges (dipoles)—which can be utilized to drive electrical currents through external circuits. Also, open circuit voltage, or the limit of the potential difference between the source's terminals as the current in the external circuit approaches zero.

Poorly named. Also quite confused in the standard theory and usage. First, *emf* is not a *force* at all. It is a difference in potential, hence it is measured in volts. Also, it is not considered an **E**-field, since that is a field of force. Yet it is in fact a , restricted to the direction between the two measurement points, and is considered to be an **E**-field in the theory!

For this reason the standard theorists say that the term is sometimes used as equivalent to a difference in potential, but that strictly speaking it should be applied only to a source of electrical energy! The problem is that conventional theorists are unaware that a potential is a bidirectional flow of energy in 3-space, where the energy enters from the time-domain (complex plane). Since they are unaware that a potential is hidden negentropic energy flow process involving broken 3-symmetry and viable 4-symmetry, the theorists are unable to consider the potential in general as a source of energy flow. Yet they are forced to consider a potential difference across the terminals of a source as "somehow a flow of energy, that generates forces upon the electrons to propel them through the circuit as current flow."

In the new viewpoint, a resolution is achieved of these incongruities by more accurately defining the E-field with respect to the potential as E [q]/q. It is noted that [dA/dt]q/q also produces an E-field, so that the full definition of the E-field is as E [(+dA/dt)q]/q. However, one volt represents one joule (of excess energy) per coulomb of charge. *The energy density flow in an electric circuit is in the voltage, in the emf, and it is not in the current per se*. The electron current (drift current) carries the excess energy that is being dissipated as work in the circuit, but in a nominal circuit the drift current only moves a few inches per hour. Half the energy is dissipated in the external circuit and external losses, and half is dissipated in forcing spent electrons in the ground return line back up through the back-emf of the source, performing work inside the source to dissipate its separation of charges. Left alone, any dipole will furnish emf indefinitely, freely receiving the energy from the time domain (complex plane) and outputting it in 3-space. The free flow of EM energy will continue unabated from the undestroyed dipole, regardless of how many external loads and losses are powered (how much energy is collected and dissipated from the ongoing flows representing the potential). So electrical circuits should be developed which allow little or no electrons to be forced back up through the dipole to dissipate its separation of charges and destroy the dipole.

ELECTROMOTIVE FORCE, BACK (BACK EMF)

Back emf is a loose term used for the *counter-electromotive force.*

When the source is powering a circuit or circuit element, the term "back emf" is used to refer to the forcing of current (electrons) back up through the source against its emf, thereby performing work inside the source to dissipate its separation of charges (its dipolarity). When a potential source of electrical energy—such as a capacitor, an inductor, or a rotating machine—is receiving energy from the source and converting it to dissipative power and work, that "powered" device is said to develop a *counter-emf.* A source has a built-in emf *a priori*, that is holding its charges separated so that it is dipolar. When electron current is forced backwards through the source through its built-in counter-emf, and backwards against its emf, work is performed against the internal resistance of the source to dissipate its separation of charges, thereby dissipating the source. If electron current is not allowed to flow back up through the source against its back emf and through its built-in counter-emf, no dissipation of the source occurs, even though energy density (not electron current dq/dt but energy flow d/dt) is flowing from the source and can be collected in collectors in the external circuit to store the energy as a created local energy source. This collected energy can then be shuttled to a separate load circuit and released, to provide a "heat pump" cycling operation and overunity coefficient of performance.

Nikola Tesla in fact knew how to shuttle energy in a circuit in such fashion, as Barrett's quaternion analysis of Tesla's actual patented circuits has revealed. [See T.W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," Annales de la Fondation Louis de Broglie, 16(1), 1991, p. 23-41.] Note that a rigorous analysis of those same circuits, using tensors or vectors, will not show the effect.

When the original source dipole is not dissipated because electron current is not passed back up through its back-emf to scatter the charges, the original source dipole will continue to furnish emf indefinitely.

The terms emf and counter-emf are normally applied only to recognized sources of electricity. Emf is applied as the algebraic sum of the potential differences acting in a circuit attached to a source. Back-emf (counter-emf) is applied as the potential difference internally through the source, between its terminals. Its main application is thought to be a *powered device or component in the circuit*. Actually, its main application is to prevent COP>1.0 electrical power systems from being developed and utilized.

ELECTROMYOGRAPHY (EMG)

A diagnostic application which detects changes in electrical potentials associated with muscle contractions.

Specific electrical patterns have been associated with specific abnormal states (e.g., denervated muscle), but the method has not yet been systematically developed beyond this application.

ELECTRON

Stable elementary particle in all atoms, and having a negative charge of 1.602×10^{-19} coulombs, spin ½, and mass of 9.11×10^{-31} kilograms.

ELECTRON GAS

A system of electrons with only weak interactions so that the electrons may be regarded as moving independently.

Thus, in many ways, that collection of electrons can be treated as a gas. Also see discussion under Drude Electron Gas.

ELECTRO-OPTICS

A technology in which electrons are focused and steered as if they were light. Examples are the cathode ray tube (CRT) and the electron microscope.

ELECTRO-POLLUTION

Also called *electromagnetic smog* and *electronic smog*. The conglomerate or maze of EM fields and radiations that are produced by technologies such as electric power transmission and radio transmission, and that can have long term cumulative harmful effects on humans.

The new viewpoint dramatically extends the types of fields and effects involved, and gives entirely new mechanisms for the interaction of EM fields and radiation with biological systems, including cumulative effects, signal-density effects, and vacuum engine effects. A mechanism for long term causation of cancer and leukemia by electro-pollution is derivable in the new approach.

As an example, even though the myriad of very weak EM signals comprising the electro-pollution are individually tiny, each when interacting at a local site with the body or its cells involves a broken 3-symmetry and thus an involvement of 4-symmetry, with corresponding giant negentropy. Further, the Heaviside dark energy component accompanying that local field interaction will in turn interact with essentially all the other locations of the body. Each of these interactions involves broken 3-symmetry and involvement of 4-symmetry. The result of the large complex of broken 4-symmetries constitutes a direct structuring in the time domain, hence weak time-pumping and weak time-charging. Time charge excitation decays only very slowly, emitting longitudinal EM waves inside the body as the charge decays.

Over a long period of time (sometimes years), this "time-hash" or "time-jamming" upon the operation of the master cellular control system, the immune system, and the cellular regeneration system is cumulatively detrimental. In this way, effects such as leukemia, cancer, etc. may be gradually "kindled". However, there have been no real scientific laboratory "cause and effect" investigations of this phenomenology at all. Nonetheless, the results will and do show in epidemiology studies, which are then attacked by orthodox scientists (many employed directly or indirectly by the power industry, etc.) on the grounds that no laboratory cause and effect evidence has been shown.

In short, this is one of the areas where the lack of scientific cause and effect laboratory evidence—because such experiments have not even been done—is improperly used to negate real scientific work showing deleterious effects that powerful interests do not wish shown.

ELECTRO-RETINOGRAPHY (ERG)

A diagnostic application that monitors changes in electrical potentials across the retina to assess eye movements.

This is almost the only method available for noninvasive monitoring of Rapid Eye Movement (REM) sleep.

ELECTROSTATIC COOLING

To be added

ELECTROSTATIC SCALAR POTENTIAL

A static (stationary) ordering in the virtual photon flux of vacuum.

The word "static" is unfortunate and should be replaced by "stationary", since the underlying active vacuum medium is always in violent motion, and there is nothing truly "static" in the universe. But an equilibrium condition can exist, even in violent motion, so that a stationary entity persists. An example is a perfect whirlpool in a river.

ELF

Extremely Low Frequency.

Used to describe EM fields whose frequencies lie in the range from just above 0 Hz (direct current) to 300 Hz. This includes power line frequencies (60 Hz in U.S. and 50 Hz in Europe) and frequencies used by certain U.S. military ELF communication systems such as for submerged submarines.

EM

Electromagnetics.

EMF

Electromotive force. See discussion under that term.

EM FIELD

Electromagnetic field. In its loose usage, it refers in a very broad way to any field, force, or energy associated with electromagnetic interactions, charges and currents.

Thus "EM field" includes electrostatic fields, magnetostatic fields, electromagnetic fields (including radiation and induction), vector-potential and scalar-potential fields, Hertz potentials, Fitzgerald potentials, Whittaker potentials, etc.

In its precise usage, "EM field" conventionally refers to an EM "force" field, such as the **E**-field, **D**-field, **B**-field, **H**-field. In classical electromagnetics (CEM), as a holdover from the old assumption of a material ether, the force fields are considered (erroneously) to exist in mass-free space, and they are also considered to be the primary causes of electromagnetic phenomena. This is a non sequitur, and it involves substituting the effect for the cause.

In the new viewpoint, the force fields do not exist in vacuum, but only in and on and of the charged particles themselves, where mass is present to form one component of the force. The force fields are therefore effects and not primary causes. The primary causes of all electromagnetic phenomena is taken to be the potentials; however, the hidden Stoney/Whittaker/Ziolkowski internal structures of gradient-free scalar potentials may interfere to cause potential gradients in and on charged particle systems. This therefore includes distance-free scalar potential interferometry and action at a distance, as well as vacuum engines.

The new view uses Sachs' unified field theory, and considers that the EM field in the vacuum, or an EM potential in the vacuum, is a curvature of spacetime function. Thus the new view is a unified field theory where GR has become EM and EM has become GR, and where EM can involve or produce gravitational, temporal, and inertial effects as well.

EM FORCE FIELD IN THE VACUUM (A FALSE CONCEPT)

The erroneous notion that there exist forces in the vacuum, so that the set of forces at a set of spatial points forms a force field.

This ignores the fact that mass is a component of force, and the presence of a force requires the presence of a mass. When Maxwell formed his electrodynamics theory, he and almost everyone else believed that a thin material ether pervaded empty space. Thus there was a "material medium" in which electromagnetic fields existed, and so it followed that real forces existed in this material medium, from point to point. Indeed, to the early electrodynamicists there was no point in all the universe where mass was absent.

Heaviside, Hertz, and Gibbs reduced Maxwell's 20 quaternion equations to the present four vector equations, but did not change the *force field in space* assumption used by Maxwell. Heaviside detested the "mystical" potentials favored by Maxwell, and believed they were figments of one's imagination. He regarded the forces as the primary EM causes.

Yet even while Heaviside *et al*. were retranslating and topologically reducing Maxwell's theory dramatically, the Michelson-Morley experiment in the 1880s had already destroyed the material ether on which those same force fields were based.

Physicists, however, have continued to use the notion of a separate force acting on a separate mass, which of course is also a non sequitur, substituting effect for cause yet again. Hence the foundations of physics remain fundamentally fouled.

For a succinct discussion of potentials versus forcefields as the primary EM causes, see Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," **Physical Review, Second Series**, 115(3), 1959, p. 485-491. See also Y. Aharonov and D. Bohm, "Further considerations on electromagnetic potentials in the quantum theory," **Physical Review**, 123(4), Aug. 15, 1961, p. 1511-1524.

EM GENERATRIX

Whatever generates electromagnetics, particularly the fields. Time is the ultimate EM generatrix.

Whittaker showed in 1904 that the interferometry of two scalar potentials (i.e., of their hidden multiwave structures) can create all classical EM fields, no matter how complex, thereby starting a branch of electrodynamics known as *superpotential* theory. See E. T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372. The paper was published in 1904 and orally delivered in 1903.

In 1903 Whittaker showed that any scalar potential (e.g. between the ends of a dipole) can be decomposed into a harmonic set of phase conjugate EM longitudinal wavepairs. The present author applied the Whittaker decomposition to the scalar potential between the ends of a dipole, and pointed out that the conjugate half-set may be taken as converging EM waves from the complex plane (the time domain), while the real half-set may be taken as diverging EM longitudinal waves emitted from the dipole in all directions in 3-space. The dipole is a known broken 3-symmetry in the dipole's violent virtual energy exchange with the vacuum; hence this shows that breaking 3-spatial symmetry allows nature to fall back to a more primary 4-symmetry. Hence energy flow conservation is violated in 3-space, but is rigorously obeyed in 4-space. Further, the free flow of 3-spatial energy from the dipole continues so long as the dipole exists.

The dipole thus is a true negative resistor, receiving EM energy in unusable form (from the complex plane; i.e., the time domain), transforming it into usable energy, and outputting the usable energy. This is also the fundamental mechanism for extracting unlimited EM energy from the seething vacuum, taking the energy from the time domain via the giant negentropy mechanism. See my paper, "Giant Negentropy from the Common Dipole," *ibid*.

This also means that the source of all EM 3-spatial energy is from the time domain, and is due to broken 3-symmetry. The ultimate EM generatrix is thus time. The ability of a bit of broken 3-symmetry (such as formation of a simple dipole) to initiate a continuing and ongoing flow of free EM energy via the 4-symmetry between time-energy and spatial energy, means that *negentropic* engineering, once we get the hang of it, is far easier than our present *entropic* engineering. In essence, we have to learn how to do electrical engineering all over again, from a much more fundamental point of departure.

EMI (ELECTROMAGNETIC INTERFERENCE)

Disturbances of electronic equipment's operation by intrusion of extraneous EM signals from outside sources. Also, disturbances of the operation of one or more subsystems of an electronic system by intrusion of undesired signals from other subsystems.

EMI, ANOMALOUS

An "ad hoc" term denoting an unexpected nonlinear optical functioning process in a dense signal environment whereby nonlinearities on the skin of a vehicle or cover of an equipment become pumped phase conjugate mirrors, so that they emit amplified phase conjugate replicas (PCRs)—of signals and fields impinging upon the skin or covering and the nonlinearity from electronic circuitry operating inside.

By the distortion correction theorem, these amplified PCRs precisely backtrack the signal waves that impinged upon the phase conjugate mirror from inside the vehicle or electronics, back to the original source of the internal signal. The result is the scavenging of energy from the dense electronic signal environment and sharply focusing and directing it into the operating circuits of the internal electronics. Thus internal signals, often amplified, start sporadically appearing through the circuits. These sporadic jamming signals due to the anomalous EMI effect are known as EMI fireflies.

Such EMI fireflies were resoundingly experienced in U.S. missiles and aircraft in the U.S. air attack on Libya in 1986. Although some of the anomalous EMI effects could have been produced by the unusually dense EM signal environment and poor hardening of circuits and electronics, it is believed that the weak jamming was also a test being conducted of a laboratory prototype Russian quantum potential weapon.

This weapon appears to have been finished and deployed at the end of 1989, and then utilized against Allied forces in the Gulf War, accounting for the Gulf War Disease. We have previously explained why British, Canadian, and American soldiers acquired the disease, why the native populace did not, and why the French troops did not.

Inside the human body, the anomalous EMI effects may be used to produce time-charging and delayed interference with the immune system, the master cellular control system, and the cellular regeneration system. By using dimensioned signals, the targeted systems can be specifically and deterministically affected as desired, for any set of disease damage types and symptomology desired. So the technology may be used to induce a "cocktail" mix of various infectious disease damage types. The induced time-charging effects—where the time-charge is dimensioned—followed by longitudinal EM wave emissions from the time-charges as they decay in the veteran's body, can induce the same or similar symptoms and damage in close family members exposed to the weak but dimensioned longitudinal EM radiation. This applies not only to wives but also to family members such as children often in close proximity to the veteran. In addition, exchange of bodily fluids from the veteran to his spouse can directly induce the symptoms in the spouse since those fluids are themselves still time-charged (with dimensioned time charge) and slowly decaying.

ENDOTHERMIC EXPLOSION

An EM implosion, resulting in explosive electrostatic cooling. The sudden, explosive reversal of electromagnetic energy scattering in a spacetime area so that the local EM heat energy is converted back to a negative potential. In this manner, explosive electrostatic cooling occurs in the area.

Endothermic EM explosion occurs by pulsed scalar Whittaker interference of two longitudinal EM wavesets beamed from two LW EM transmitters. The distant transmitters of the LW interferometer must have their local electrical grounds biased well below the ambient zero potential in the distant interference zone (IZ). Whittaker internal EM waves radiate through subspace (the LW superhighways inside EM fields, potentials, and waves in mass) from the transmitter to the interference zone, and also from the interference zone back to the negatively biased transmitter/receiver, emerging there as ordinary electromagnetic scattering energy or *heat*. All *normal* explosions are exothermic in nature, and ordinary electromagnetic heat energy is radiated away. A scalar EM interferometer/howitzer, however, may produce an endothermic explosion (an energy implosion) at a distance, explosively extracting the heat energy at the implosion site and receiving the energy back at the transmitter site for disposal.

The same LW interferometer can also produce an exothermic explosion in the distant interference zone, merely by positively biasing the electrical chassis grounds of the scalar EM potential transmitters and pulsing the transmission. Negative biasing of the transmitters' electrical chassis grounds and a pulsed transmission result in an endothermic explosion in the distant interference zone.

ENDOTHERMIC LONGITUDINAL EM WAVE INTERFEROMETRY

Extraction of EM energy from a distant intersection area by a scalar interferometry whose electrical grounding is biased well below the ambient potential in the distant interference zone.

The result is a cooling of the interference zone, and heat energy emerging at the transmitters.

ENERGETICS

The Russian term for what I have called advanced higher-symmetry electromagnetics,

...using artificially constructed Whittaker internal biwave structures to make activated potentials that have specific, desired "vacuum engines", where the EM field is a curvature of spacetime function, and thus involving a unified field theory that is engineerable by novel electrodynamic means.

I believe that the closest Western work approaching the Russian energetics science is the O(3) higher symmetry electrodynamics pioneered by Evans, particularly when combined with Sachs' unified field theory as an important subset. The other closest Western work would seem to be the important work of Barrett in SU(2)×SU(2) symmetry electrodynamics.

The activated potential containing the infolded vacuum engine structures is said to have been *dimensioned*. These dimensioned potentials can be combined to form EM carrier waves which contain the hidden vacuum engines. These vacuum engines, scalar Whittaker EM interferometry, and activation of vacua and matter are used to perform unified field engineering.

Russian energetics also adds the formation and use of the quantum potential, enabling
(i) instantaneous action at a distance with multiple separated target nodes, (ii) essentially limitless direct energy amplification in the assembly of receiving nodes, without additional energy input at the initiating node, and (iii) engineering of hidden variables inside distant targets, instantaneously, without transmission *as such* through space. The quantum potential itself can be *dimensioned*, so that vacuum engines can be instantaneously transmitted into distant targets, altering their local spacetimes to contain the desired engine(s). Note that no intervening barrier can "impede" or "shield" a quantum potential because it does not travel through space as such; it just appears everywhere at once in the space where it will be.

Adding to classical EM theory, general relativity, and quantum mechanics the internal EM energy in the scalar EM potential will generate a superset of each discipline. These three new supersets are unified in their common internal EM energy subset. As a bonus mind, personality, memory, and deep biological cellular and system control are also available, since the internal time-energy domain is used by biological systems for these living functions. Energetics thus is an engineerable, unified field theory of physics and metaphysics.

Russian energetics is largely still a "secret" science, developed over the last five decades and highly weaponized. In the Russian scheme the overall area is called *energetics*. It consists of three subsets: (i) *energetics*, which applies to nonliving systems, (ii) *bioenergetics*, which applies to the physical bodies, cells, chemistry, genetics, etc. of living organisms, and
(iii) *psychoenergetics*, which applies directly to the minds, personalities, memories, emotions, and perceptions of living systems. Russian energetics research, development, weapons, and weapon sites are under the rigid control and operation of the KGB, by whatever name it may use from time to time. The energetics weapons are not in the regular Russian armed forces. Energetics weapons have been tested worldwide for several decades, and this testing has been documented elsewhere by the present author.

ENERGY

Usually erroneously defined as the "capacity to do work".

The closest definition is: *The invariance of that which is varying*. Note that this is a 5-law logic definition and not an Aristotelian definition.

Here we have a real bucket of worms. This term is usually totally misdefined (as in the nostrum that "energy is the capacity to work") in physics, which defines energy as the "ability to scatter itself or change its own form" which obviously are not identities! A common definition is given as "a quantity that describes the capacity to do work; commonly divided into three major classifications: kinetic (dynamic) energy, potential (static) energy, and radiant (electromagnetic) energy. Well, energy can be changed in form and that is work. But energy is not the "capacity to change its own form", but is its own form, as can be seen.

Anyway, we'll have a go at a definition perhaps a little more acceptable, though still imperfect. Then we'll discuss things a bit.

Rigorously, energy is any deterministic or coherent structuring, either dynamic or stationary, existing in the virtual particle flux of vacuum. While that *sounds* good, it is still "the invariance of that which is varying." Ultimately, a thing is defined in terms of its appositive.

The *type* of energy depends upon the type of virtual particle(s) whose flux is selected. E.g., electromagnetic energy is a deterministic or coherent structuring, either dynamic or stationary, in the virtual *photon* flux of vacuum. Any trapped energy (ordering in the vacuum virtual particle flux) is gravitational in nature, since trapped energy is gravitational. In a mass, e.g., it is the trapped energy that is gravitational, not the mass per se, in the modern view. Contrast the definition of electromagnetic energy to the definition of an electrostatic scalar potential (ESSP): The electrostatic scalar potential is a stationary trapped deterministic or coherent dynamic wave-structuring of the virtual photon flux of vacuum. Thus it can be seen that the ESSP is a special case of EM energy: the stationary trapped case. For this reason, a mass—which is an enormous amount of statically trapped EM energy—may properly be considered as an extremely dense EM potential, which we refer to as the *mass potential*.

A potential is any trapped ordering or coherence, either static or dynamic, in the virtual particle flux of vacuum. A specific type of potential represents a selection of a particular virtual particle's flux in the vacuum, and an ordering—either static or dynamic—imposed in that flux.

However, since the scalar potential is just the static ordering in the virtual particle flux of vacuum, it follows that the scalar potential must contain—and in fact be totally composed of—internal order. E. T. Whittaker showed the form of that order in 1904. (See entry under Whittaker, E. T.) The scalar potential is composed of a harmonic series of pairs of antiparallel transverse waves. The antiwave of each wave/antiwave pair is a time reversed replica of the wave in that pair. In other words, the scalar potential is a special type of Fourier harmonic frequency expansion, coupled with its own phase conjugate present simultaneously.

Feynman said it bluntly: *"It is important to realize that in physics today, we have no knowledge of what energy is."* [Richard P. Feynman, Robert B. Leighton, and Matthew Sands, **Lectures on Physics**, Addison-Wesley, Reading, MA, Vol. 1, 1964, p. 4-2.]

E.g., particularly see R. B. Lindsay, "The concept of energy and its early historical development." **Foundations of Physics**, 1(4), 1971, p. 383-393. Lindsay investigates the concept of energy from its early historical origin. The root of the concept is the notion of *invariance in the midst of change*. Lindsay covers the concept and its evolvement (or lack thereof!) from ancient times through the 18th century.

The real problem is that the notion of energy implicitly involves the age-old unresolved philosophical problem of change, posed by Heraclitus over 3,000 years ago and unresolved to this day. *It cannot be solved within the realm of 3-law Aristotelian logic!* One way of posing it is something like this, after Heraclitus: "For a thing to change, it must turn into something else. How can a thing be itself and something else also? How can a thing have changed, and yet not have changed?"

The problem is that Aristotelian logic has a serious flaw of omission; it actually *implicitly assumes* a fourth law, which is the age-old philosopher's nightmare that, when a thing is pursued to the limit, *it invariably turns into its own opposite*. So in anguish the philosophers spoke of this nightmare as the "accursed necessity for the identity of opposites."

I did resolve the philosophers' nightmare some years ago, by adding a fourth law of logic to the Aristotelian three, which completes and extends Aristotelian logic, and an application rule which can be taken as a fifth law. E.g., see Appendix III: "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," in T. E. Bearden, **AIDS: Biological Warfare**, Tesla Book Company, Chula Vista, California, 1988, p. 428-443. A simple proof of the fourth law is given in that paper. For a much more elegant treatment of a higher topology logic, which is also practical and useful beyond the limitations of Aristotelian logic, see G. Spencer Brown, **Laws of Form**. Julian Press, New York, 1972. For a proper perspective on mathematics and on Aristotelian logic, see Morris Kline, **Mathematics: The Loss of Certainty**, Oxford University Press, New York, 1980.

It may be that, since by "identity" we mean at least the identity of "left" and "right" different observations, considering the left and right sides of the identity statement. If so, then already we involve a violation of Aristotelian logic in the very notion of "definition". In short, ultimately we can only define a thing in terms of an appositive in the first place. So we write an identity statement where left and right are different sides, then state "Oh, no, they are the same thing after all!" Said in that manner, the problem can clearly be seen. In one sense a definition is pure Zen, where no truth can be stated at all unless the speaker says two exactly opposing things, and then waits for the recipient to realize the truth that has been stated.

What this all means is that, since "energy" (and any other definition, ultimately) involves a logical contradiction in Aristotelian logic, we must "define" *energy* simultaneously in an appositive ways: (i) in the wave view and (ii) in the particle view. However, we point out that the notion of a "moving particle" is the notion of something considered momentarily as fixed (i.e., the particle), but which is immediately thereafter (in time snapshot #2) considered as moving. Notice that "moving" is changing, and so it involves Heraclitus' problem of change *a priori*. The notion of a "moving wave" is also a similar thing; i.e., "a wave" momentarily considers the entity as static and not changing (in time snapshot #1), and in the next moment (time snapshot #2) considers it as moving (changing). We can only point out this dichotomy for the reader to reflect upon and hopefully to grasp. *The conscious mind is a serial processor; it is extremely fast but still perceives snapshot by snapshot, serially. It essentially has been fitted to the single photon interaction, as has Aristotle's three laws of logic*. The problem is in the operation of conscious perception itself, not in nature which is not limited to just serial processing.

The unconscious mind is a massively parallel processor, so it can have many "different slides or snapshots in the slide projector" at the same time. So with our conscious mind's perception (Aristotelian logic is just a model of that, using single photon interaction as the physical observation mechanism) we see in terms of "either-or". On the other hand, with our unconscious we can also grasp the fact that single snapshot A can be the same thing as a totally different snapshot B, if the rules of "decision as to identity" are changed between snapshot A and snapshot B. Suppose that by "identity" we mean "incapable of being discriminated between." Is a black marble identical to a red marble, e.g.? It is perceived so if you are color blind. It is perceived not to be identical if you are not color blind. The point is something that most philosophers missed or brushed aside: identity is not absolute, but is "perceived identity". Identity is the result of a decision operation, based on multiple perceptions, recall from memory, and comparison of two remembered perceptions. The most simple form of comparison is simple subtraction.

The unconscious (massively parallel) perception can handle both snapshots simultaneously. The conscious mind only can see a single marble at a time, so it will see one or the other. In a *third* time interval it will recall from memory the first perception, compare it with the memory of the second perception (say, by simple detailed subtraction), and then decide whether object A is perceived to be identical to object B. The unconscious (massively parallel processor) can handle both the marble snapshots at the same time. It can see both the identity of the two, to a color-blind "identity decision process" and also the nonidentity of the two, to a color-sensitive "identity decision process." Again, nature is not limited to the single-snapshot, serial perception process of our conscious (serial) processor!

Unfortunately, in physics and mathematics we seem to still be thoroughly confused on this very perception issue. Brown's work may be the first work to straighten out the area of the topology of the logic utilized. My own independent effort was successful but far cruder than Brown's; Brown did it elegantly. We bring all this out because when one deeply looks into foundations and true definitions, one will directly confront the philosopher's nightmare and one's own conscious mind will never be able to resolve the problem. One's unconscious can resolve it, but it requires a certain "stretching of the realization" to harmonize monocular perception of the conscious mind with multiocular perception of the unconscious mind.

Just as a small example, the question of "is it a wave or a particle" once raged as a great debate in physics, and was never resolved. Instead, physicists just finally shook hands and agreed to quit arguing and fighting about it, and agreed to accept that the same entity could sometimes act as a particle, and in other cases it could act as a wave, and that *as it existed in the present, before observation,* it was somehow both a wave and a particle but we couldn't see its true nation (in that fashion). The key, of course, is that the question was posed in "either-or" form (Aristotelian logic form), and it cannot be answered in that reduced logic. Instead, a higher topology logic is required. Brown's logic answers it easily, as does my own four law superset of Aristotelian logic (five law superset if we count the application rule as a separate law). In the present, e.g., we have masstime and not mass. Just after a photon is emitted, we have mass and not masstime. The latter photon emission process consummates "observation." The point is, the true entity does exist in zero time as a spatial entity. It is not a spatial entity but is a spatiotemporal entity. If the "interaction" process tears off its "time-tail," it is converted to a "spatial entity" (frozen snapshot) and therefore seems to be a particle. If the "interaction" process does not remove the time-tail, it remains a spatiotemporal entity and therefore seems to be a wave. The *concept* of particle is 3-spatial concept. The concept of wave involves time also. No wave can exist and "be a wave or wavelike" unless it has some time extension to "be waving" *in*!

The physicists debating the wave-particle duality, you see, missed the most important point that *no observable exists in time or persists, a priori.* Hence rigorously no particle exists in time.

So we define energy (particle view) as any virtual particle flux pattern, where that definition automatically considers a single particle in translation as well as multiple particles in their myriad of translations. We define energy (wave view) as any wave flux, where that automatically considers a single wave in translation as well as multiple waves in their myriad of translations. Now note that one is going to have to deal with, and discriminate between, energy, energy flow, energy collection, and collected-energy dissipation.

As the reader can see, whenever one struggles with foundations issues, the problems get very complicated. We've beat this dead horse quite enough, and leave the further beating to the reader!

ENERGY AMPLIFYING COIL

A coil whose magnetic energy flow input is first split into two energy flow inputs, each equal to the original energy flow available from a magnetic source. Both the split energy flows then interact simultaneously with the coil, amplifying its electrical energy output compared to the conventional case where only a single energy flow interacts with the coil.

In the motionless electromagnetic generator (MEG), the curl of the magnetic vector potential of a permanent magnet (i.e., its magnetic field), is extracted and held within the core material of a special transformer, while the magnetic vector potential of the permanent magnet is freely replenished in uncurled (field-free) magnetic vector potential form in space around the core material. Both the full magnetic field in the core and the full uncurled magnetic vector potential in space adjacent to the core interact with each output coil of the transformer. Hence a single potential energy flow is "split" into two flows each equal to the original EM energy flow, and both the parallel flows energetically interact with the output coil. This increases the energy interacting with the coil by this "dual interaction", so that the coil is said to be an "energy amplifying coil".

ENERGY COLLECTION/COLLECTING

In electromagnetics, exposure of an interceptor/collector, e.g., a charged particle, into an impinging EM energy flow, so that some of the flow is diverted around the intercepting particle. We may say that the flow of energy has been "reordered" in a stream around the diverting charge.

Energy collection is not a thing which "has been accomplished" and finished; instead, it is an ongoing process.

Note now that, if you perform or consider or impose any sort of ordering at all in the energy flow/flux, rigorously you have just implied an *energy collection/collecting* process. The energy part of a photon, e.g., would seem not to be "energy" per se. Instead, it is an energy collection/collecting, and a time collection/collecting as well. Yet that energy part is energy also! [Not to worry! Energy and energy-collecting, contrasted as appositives, become identities at the extreme boundary separating the two concepts. That boundary is the photon concept. Aristotelian logic does not apply on the boundary, which violates the Aristotelian law of the excluded middle.] More reasonably, a charged capacitor does not have "stored energy" per se in it, in the sense of "so many chunks of energy called joules". Instead, it has an *ongoing energy collection/collecting process*, where the asymmetry in the local vacuum flux exchange, of the dipole across the capacitor plates, continually extracts and "gates" a flow of EM energy from the capacitor dipole. That energy-flow bathes all the trapped charges in the capacitor. The energy-flow interaction with the trapped charges creates the **E**-field and all the rest.

The diverted part of the total energy flow, is what is calculated in the "Poynting flow" or **S**-flow.

ENERGY, DISINTEGRATED

An ad hoc term for the energy of the incoherent virtual particles in the virtual particle flux.

ENERGY DISSIPATION

Refers to the scattering or disordering of energy.

This is somewhat of a misnomer, broadly used in physics and somewhat erroneous.

Energy dissipation" *per se* does not mean that the energy is destroyed. Instead, there is the dissipation of the ordering (*collected/collecting*) of energy in which the diversion (local ordering) process disappears. What we call the "dissipation in space" of energy is rigorously the dissipation of the collection/collecting in space of the energy. In short, the energy is scattered, as in a resistive load, in all directions. But it is still there in space, just disordered.

Recall that all EM energy at a dipolarity comes directly from the time domain. It is the collecting process, not the energy, that is dissipated from a point or a region. Only *collected/collecting* of energy can be truly dissipated. With that understood, then dissipation of collected energy is simply removing the "collector" from the energy density flow impinging upon it to cause the collecting process to occur, or removing the flow of energy impinging upon the collector so that it intercepts and diverges some of it around itself.

So there are two ways in which this "collected energy" dissipation can occur: (i) the collector is moved out of the impinging energy-flow-stream, or (ii) the energy-flow-stream impinging on the collector is reduced to zero. Note that "changing the form" of the energy—i.e., changing the collector and the type of collecting—involves both conditions (i) and (ii). Thus energy is dissipated when the *collector/collecting* form is changed. In the conventional, less accurate statement, work is performed when "the form of the energy is changed." That now should be stated as "work is performed when the collector in the energy-flow-stream is changed to a different collector, because the first collector must be removed from the energy-flow-stream or the energy-flow-stream must be removed from it.

In physics we also consider that "positive" work by a body is work it does on something else, while "negative" work is work done upon the body by something else. We specifically emphasize that collection/collecting need not be a unitary, one-time occurrence! The EM energy-flow impinging upon a collector roars on by it at something like the speed of light. The entire flow is diminished by only about 10^{-13} (the nominal energy collecting fraction) as it flows onto, and on beyond a coulomb of charge. If that departing energy-flow is then retroreflected or otherwise *returned* to pass back through again, additional energy collection/collecting can and does occur from it. This can be repeated many times, particularly in an intensely-scattering, optically active material. Also, after the collected fraction of energy is dissipated, we can also retroreflect the scattered energy flow back and reuse it again. This too can be repeated many times.

So we must be careful of the present *work-energy theorem*, which is a special case and assumes only a single pass of the energy-flow in the "collection/collection" of energy that "changes form." If we institute additional passes, with additional "collection/collecting of energy, then we can collect and "use" a little more energy from all that enormous energy-flow that otherwise would escape us. If we dissipate some of the collected energy, then also intercept the departing dissipated energy flow and return it, we can collect some of the energy flow again and then dissipate that new collection again. And so on.

One can permissibly get many joules of "work" from a single joule of energy by iterative retroreflection, iterative re-collection, and iterative dissipation, without violating the conservation of energy law. This is one of the major secrets of overunity energy processes. It is proven by the overunity or excess emission (Letokhov was forced to call it "negative absorption of the medium" from an intensely scattering, optically active medium, as in the Lawandy experiment and in his patented process for lasing without population inversion. It is also proven by Patterson's adaptation of anti-Stokes emission.

Let's do a gedanken (thought) experiment, where we do not even apply the nominal 10^{-13} energy collection fraction. Let's just *conventionally* say "energy flow," "energy collection," and "energy dissipation," and ignore the "collection fraction." Since we are talking theoretically, we can have a perfect system. So we have a closed system, which we open to input some energy into. The system then opens in the load to dissipate all the energy back to the environment, thereby doing work. Now suppose I passively (i.e., without any input energy of my own, by perfect retroreflection) return all the dissipated energy back to the input once again. We simply let it flow back in, the system collects it again, and then dissipates it again—giving us the same amount of work we had the first time. Then we do it again. And again. I think you get the picture. The fact that energy can neither be created or destroyed, but only changed, means that the "same energy" can be *recycled* over and over. One has to have two different actions: one that changes the form of the energy so we get some work, and the other that retroreflects and recovers the dissipated energy (does negative work!) so that we can recycle it and dissipate it again. In principle, by recycling of energy one can get many joules of work from a single joule of energy. Nature does it every day! Every joule of energy formed at the initial

"Big Bang," is still with us. Every joule of it has on the average already performed many joules of work since then. And as we stated, the anti-Stokes emission process is a concrete example and demonstration that the energy recycling process can be accomplished in physical systems, so that the systems permissibly exhibit COP>1.0.

A most common example is the replenishment of a potential. E.g., the magnetic vector potential can be curled as a **B**-field, and that **B**-field can be totally confined inside a toroid, inside a long coil, or in certain magnetic core materials (as in the motionless electromagnetic generator or MEG for short). The **A**-potential will be replenished as fast as you withdraw the "total **B**-field" from it. So outside the toroid or the long solenoid, you will have as much **A**-potential energy as you started with, and the Aharonov-Bohm effect proves that the energy is there and will interact with electrons. Meanwhile, in the trapped **B**-field you also have drawn off as much energy as ostensibly the **A**-potential contained in the first place.

ENERGY, ELECTROMAGNETIC

Any deterministic or coherent structuring, either dynamic or static, existing in the virtual photon or charged particle flux of vacuum.

Note that, since a photon may also be considered a positron-electron pair, the flux of charged virtual electrons and positrons is included in the flux of photons.

ENERGY FLOW

The passage of energy, in any form, in one direction or generally in one direction.

Even though this "definition" is in error, we now just bypass the unresolved "observables do not persist" problem and the 4th law nature of change (e.g., motion) problem and proceed.

Note that this energy flow may be the overall movement of an omnidirectional energy flux, or it may be the movement of waves, or both. It may be in the kinetic energy of a moving mass. If the flow is everywhere the same, or even approximately the same, we often speak of the entire "flow" as just "energy." Often energy flow is also considered to have occurred from a region of more intense energy flux to a region of less intense energy flux, and vice versa. In other words, energy divergence or convergence is also considered as energy flow. As can be seen, *the usage of the term "energy flow" is not precisely defined in physics! It isn't going to be, at least not in ordinary Aristotelian logic terms.*

ENERGY FLUX

The passage of energy flow, in any form, through a unit area, and in any direction (usually all directions).

This includes both passage in a single direction, or passage in multiple directions at once.

ENERGY, FREE

Any environment or entity which freely emits, furnishes, or makes available energy which can be either collected and used as furnished or transduced into usable form and collected and used to power loads and do useful work.

The word "free" does not apply to costs, but rather to the source of the energy. Of course the ideal is energy that is both free for the taking (collecting and using) and also inexpensively taken and used.

Implicit in the notion of free EM energy from the vacuum are five key concepts: (1) the local vacuum/spacetime itself is filled with EM energy of generally unusable form, (2) some sort of organizing principle is utilized in one part of the circuit or system to cohere the vacuum energy and transduce it into usable EM form, (3) a gating mechanism is utilized to divert the vacuum energy in transduced usable EM form from the active vacuum to the external circuits or load parts of a device, (4) the coherent gated extra energy available in the external circuits or load parts is intercepted by those circuits or load parts, couples to them, and potentializes them with extra EM energy, so that the energy is collected by the system and regauges the system asymmetrically, and (5) the external circuits or load parts then scatter or dissipate the coupled extra energy to produce useful work, without destroying the organizing and gating mechanism that is extracting and furnishing usable EM energy from the vacuum.

ENERGY SOURCE

A function or entity which furnished energy.

Since energy cannot be created or destroyed, there is really no such thing as a "source", though it is a term commonly used. So called "energy sources" actually gate energy, changing it from one form to another. As an example, all 3-space EM energy is gated into 3-3-space from the time-domain. To the gating entity or "source" of the 3-space EM energy, there is an input of EM energy flow from the time domain (complex plane). The source must be a broken 3-symmetry in energy flow, so that more primary 4-symmetry energy flow applies. Hence the conservation law applies to the inflow of energy from the time dimension and the outflow of energy in 3-space.

ENGINEERING, ENTROPIC

Engineering or designing systems and processes whose operation "spills" or wastes some energy.

Entropic systems—which are the products of entropic engineering—thus have COP<1.0 and function in forward-time.

ENGINEERING, NEGENTROPIC

Engineering or designing systems and processes whose operation freely "receives" excess energy, so that they output more energy than the operator inputs, the excess being received from the external environment or other source.

ENTROPY

Thermodynamics term representing a measure of a system's capacity to undergo spontaneous change. In short, a quantity which measures the extent to which the energy of a system is unavailable to do work.

Entropy is given by $dS = dQ/T$, where dS is an infinitesimal change in entropy for a system absorbing an infinitesimal quantity of head dQ at absolute temperature T. In statistical mechanics, entropy is a measure of the disorder in a system and is given by $S = k \ln P + C$, where S is the value of entropy (disorder) for a system in a given state, P is the probability of that state's occurrence, k is a fixed constant, and C is an arbitrary constant. Intuitively, the concept of entropy corresponds to the concept of disorder. We specifically call attention to the fact that the "S" used in representing entropy is not at all the **S** used for Poynting flow.

In physics, a measure of the randomness or disorder of a system. Given by S = k ln g, where S is the entropy, k is the Boltzmann constant, and g is the total number of microstates available to the system. Sometimes stated as S = k ln (multiplicity).

Intimately connected with the Second Law of thermodynamics, which may be stated for the macroscopic case (after Baierlein) as follows: "If a system with many molecules is permitted to change, then—with overwhelming probability—the system will evolve to the macrostate of largest multiplicity and will subsequently remain in that macrostate. Stipulation: allow the system to evolve in isolation. (The stipulation includes the injunction, do not transfer energy to or from the system." [Ralph Baierlein, Thermal Physics, Cambridge University Press, Cambridge, 1999, p. 29].

ENTROPY, NEGATIVE (NEGENTROPY)

In a sense, the reversal of disorder, or reversal of entropy.

Since the normal or forward flow of time is entropic, negentropy is intimately related with time-reversal. This can be seen crudely this way: As a physical thing exists in time, it decays, ages, deteriorates, etc. In short, its ordering as a viable separate thing gets more and more disrupted. If it is then "reordered" precisely back to its original state, that is obviously a time-reversal of that entity.

Since the production of energy is connected to the production of more ordering, then it would seem to follow that any negentropic process also produces more energy. Or, that production of excess energy accompanies a negentropic process, and vice versa.

When we factor in the giant negentropy process, it is seen that all EM spatial energy comes from the time domain, via the broken 3-symmetry of the source charge or source dipole. Here we have a pure negentropic process. Hence we may now advance a new definition of negentropy for electrodynamics: *negentropy is the direct and ongoing conversion of time-energy (time-structuring and time-ordering) into 3-space EM energy*. Dipolarity (polarization) is the process initiating and producing the negentropy. Polarization is accompanied by ongoing negentropy so long as the polarization persists.

Another way of defining negentropy even more rigorously—for electrodynamics—is to define it in spacetime as a *broken 3-symmetry in EM energy flow with concomitant relaxation to a more primary 4-symmetry between time-energy flow and 3-spatial EM energy flow, where time-energy flow is transduced into 3-spatial EM energy flow.*

When a system with one measure of multiplicity progresses to a state with a decreased measure of multiplicity; i.e., when its total number of microstates decreases in the transition. A nonequilibrium steady state (NESS) system can exhibit negative entropy, and the entropy can further decrease progressively toward negative infinity as time passes. [D. J. Evans and Lamberto Rondoni, "Comments on the Entropy of Nonequilibrium Steady States," *J. Stat. Phys*., 109(3-4), Nov. 2002, p. 895-920]. Note that the Second Law of thermodynamics only rigorously applies to an isolates system, where neither mass nor energy crosses the boundary of the system. In the NESS system, that prohibition is violated, hence the second law need not apply. The common source charge is an example of such a NESS system continuously producing negative entropy, because it absorbs virtual energy from the vacuum, transduces it into observable energy, and re-emits the energy as observable photons radiating at light speed in all directions and forming the associated EM fields and potentials and their energy.

ENTROPY (POSITIVE)

When a system in a state with ordering progresses to a state with less ordering, i.e., when its total number of microstates increases in the transition.

ENZYME

Any of various proteins or conjugated proteins (compound of a protein and a nonprotein) produced by living cells, that serves as a catalyst for specific biochemical reactions at body temperatures.

EPIDEMIOLOGY

The study of the various factors influencing the occurrence, distribution, prevention, and control of disease, injury, and other health-related events in a defined human population.

EQUILIBRIUM

In physics, the state of a system in which opposing forces balance each other; i.e., sum to a vector zero.

In mechanics, a system of particles where the net force acting upon each constitutent particle is zero.

In chemistry, the state of a solution where the forward and reverse reactions are equal and balance each other, so that the chemical reactants and products maintain the same amounts.

EQUIVALENCE PRINCIPLE (PHYSICS)

One of the principles of relativity theory. The mass of the body determined by using the gravitational force on it to measure the mass, and the mass of the body determined by using inertial force on it to measure the mass, yield equal measurements for the mass.

ETHER

An extremely fine substance theorized to exist in space in lieu of sheer emptiness.

Discussed by Tesla, Einstein, Whittaker, and others. Originally conceived as the luminiferous ether, a thin material fluid pervading all space. Maxwell, e.g., believed in this material ether, and formulated his theory accordingly. Thus he mistakenly provided forces in the vacuum, which forces require a material vacuum. I.e., mass is a component of force; there is no force without mass, since $\mathbf{F} \equiv d/dt(m\mathbf{v})$. The hoary old medieval notion that a separate force acts upon a separate mass that it affects, is completely in error, though still utilized in mechanics! Modern field theorists know better, but correction of this massive error in physics has not been widely accomplished.

Originally, the ether was seen as a ghostly-thin, *material*, fluid medium postulated to fill space and carry electromagnetic waves, similar to the way in which a gas carries sound waves. Faraday and Maxwell added the assumption that EM fields existed as material taut strings (Faraday's lines of force) with conveniently missing endholders to provide the tension on the strings. The Michelson-Morley experiments eliminated any such material ether, but not a Lorentz-invariant nonmaterial ether (of which vacuum itself is one example, in the modern view). Maxwell's equations were derived from a mechanical ether theory. Einstein once proposed that the vacuum, complete with its electromagnetic fields, should be called the new *ether*. In our own approach we have come full circle back to Einstein's suggestion, since the so-called "fields" are really just curvatures of spacetime and are included in our use of the vacuum as a giant potential comprised of virtual particle flux. The Sachs unified field theory, I believe, is the best available, particularly when the Evans O(3) symmetry electrodynamics is included as a special subset of it. In my own view, I regard "vacuum potential" and "spacetime" as essentially identities. Thus both the vacuum potential and spacetime have further internal structuring and dynamics. In relativity, this view is akin somewhat to the notion of "subspace" used on the Star Trek series.

Today the word "ether" is again acceptable in the literature when used in context of its modern quantum mechanical meaning. The "ether gas"—the quantum mechanical vacuum—is considered to be composed of virtual particles. The particles spontaneously come into being and then vanish so quickly they cannot be individually detected. That is, today the ether is regarded as a sort of specialized virtual particle gas—sort of a gas of "effervescent energy bubbles", so to speak—where the individual particles are continually being created and annihilated at a fantastic rate.

A good reference on the modern view of the vacuum is I. J. R. Aitchison, “Nothing’s plenty: The vacuum in modern quantum field theory,” **Contemporary Physics**, 26(4), 1985, p. 333-391. For an O(3) treatise, see M. W. Evans, P. K. Anastasovski, T. E. Bearden *et al*., "Derivation of the Lehnert field equations from gauge theory in vacuum: Space charge and current," **Foundations Of Physics Letters**, 13(2), Apr 2000, p.179-184. It is shown that the Lehnert field equations in vacuum, with concomitant space charge and current, can be derived straightforwardly from standard gauge theory applied in vacuum, using the concept of covariant derivative and Feynman's universal influence. The Lehnert and Proca field equations are shown to be inter-related through the well-known de Broglie theorem, in which the photon mass can be interpreted as finite. These ideas go some way towards addressing the inconsistency inherent, in Maxwell's famous displacement current, which has no concomitant vacuum space charge.

For technical insight into the attributes for extracting EM energy from the vacuum, see particularly M. W. Evans, P. K. Anastasovski, T. E. Bearden *et al*., "Classical electrodynamics without the Lorentz condition: Extracting energy from the vacuum," **Physica Scripta** 61(5), May 2000, p.513-517. It is shown that if the Lorentz condition is discarded, the Maxwell-Heaviside field equations become the Lehnert equations, indicating the presence of charge density and current density in the vacuum. The Lehnert equations are a subset of the O(3) Yang-Mills field equations. Charge and current density in the vacuum are defined straightforwardly in terms of the vector potential and scalar potential, and are conceptually similar to Maxwell's displacement current, which also occurs in the classical vacuum. A demonstration is made of the existence of a time dependent classical vacuum polarization which appears if the Lorentz condition is discarded. Vacuum charge and current appear phenomenologically in the Lehnert equations but fundamentally in the O(3) Yang-Mills theory of classical electrodynamics. The latter also allows for the possibility of the existence of vacuum topological magnetic charge density and topological magnetic current density. Both O(3) and Lehnert equations are superior to the Maxwell-Heaviside equations in being able to describe phenomena not amenable to the latter. In theory, devices can be made to extract the energy associated with vacuum charge and current. More than a dozen ways to approach extracting EM energy from the vacuum are presented.

EVANS, MYRON W.

Noted chemical physicist, primary developer of O(3) symmetry electrodynamics, and Director of the Alpha Foundation's Institute for Advanced Study (AIAS).

Under Evans, the work of the AIAS has made enormous contributions to physics and particularly to the extraction of usable EM energy from the vacuum. In the opinion of the present author, Evans is perhaps the single most important scientist in the Western world in the emerging field of electrical energy from the vacuum and in solving the world energy crisis.

EVIL

Refers to an act or acts perceived to be detrimental to the receiver or referent, in the positive time stream.

Oddly, the chain of events perceived to involve "evil" to a referent in the forward time stream, would be perceived to be equally *beneficial* to the referent, viewed in the *negative* time stream. In other words, evil in general is a synthesis of the results of observed actions or acts in the *forward time motion*, hence it is conjugated when time flow is reversed.

This is not just an idle question. In our physical body, the photon interaction with the electron shells of the atoms in the body largely determines the observed forward movement of one's material body through time. The conscious mind, fitted to the single photon interaction, thus perceives its "existence" in such terms. Meanwhile, in the nuclei of the atoms of our material body, the charge and hence the field interactions are reversed (giving Newton's third law reactions). That of course is a time-reversed flow of events, not sensed by our conscious mind, but possibly sensed by our deep unconscious mind. The question then arises, "Do we then actually exist both forwards and backwards in time?" In terms of good and evil, if the answer is yes, then this would have profound implications to theology and to the study of human consciousness.

Theologians (and philosophers) have pondered the "nature" of good and evil for literally thousands of years, and have reached no truly firm conclusion to date. The sticking point (theologically) seems to be the puzzle of a perfect, "all good" creator who had to have created evil also (since the Creator is axiomized to have created all). To the question, "How could a perfect Creator have created—or even tolerated—evil?", theologians have yet to find a satisfactory answer. Neither could they find a satisfactory answer to the question of, "How can the Creator, being all-just, and being all-powerful, then allow any evil act to occur?" In short, why does the Creator seemingly not "police up our perceived imperfections in the creation?" In short, theologians struggled with a formidable problem of how to justify the Creator's continued tolerance of evil, even for a moment.

There is no satisfactory answer to such questions, of course, in 3-law Aristotelian logic, because they are four-law logic situations. However, since the 3-logic is incomplete, one should examine the question in 4-law logic, and also in light of one's assumptions regarding the existence of the Creator. The question is unresolvable in 3-law logic and with the assumption of a Creator traveling through time, as we perceive ourselves to do. However, if time itself is a creation, then the axiomized Creator may be more properly axiomized as existing both in and outside time, and "everywhere in time at once" as well. The unlimited Creator must even be postulated to be able to "both be and not be simultaneously; if He cannot do that little trick, He has a limit." That trick must be allowed, else the Creator has a limit and is thus imperfect. But to such an axiomized Creator "permeating time everywhere at once, from beginning to end, which shall be the "beginning" and which shall be the "end?" And if we pick a beginning and an end, what shall be the "intermediate ordering" that we pick? There is no absolute position in either space or time to the "everywhere and everywhen at once" axiomized Creator.

The problem of good and evil then becomes a similar problem to the "wave or particle" question. Good and evil are two sides of the same coin, transposed one into the other by time-reversal. With respect to finite creatures perceiving themselves moving forward in time, not existing in the future and not existing in the past, the question merely demonstrates that our perception itself is in error.

It may be, then, that there is a solution (whether one prefers it or not!) to such profound and ancient philosophical and theological problems in four-law or five-law logic. However, we will leave that question to the logicians and to the theologians to more properly decide.

EXCESS COLLECTED ENERGY

Collection by an interceptor/collector of a greater fraction of the EM energy flow impinging upon it, than is normal.

Any static electric or magnetic charge, e.g., with an EM energy density flow impinging upon it, has excess "energy collection" or "energy collecting" upon it, due to the altered vacuum virtual photon flux that is now exchanging with it. In a nominal electric circuit, e.g., a single "pass" of the EM energy flow from the source dipole, along the circuit conductors and across the conduction electrons, causes an "excess energy collection" upon the intercepting surface charges in the conductors. This potentializes the Drude electrons in the circuit and leads to subsequent dissipation in the loads (and losses) of about 10^{-13} of the EM nondiverged Heaviside energy flow that actually was extracted by the source dipole from the vacuum and sent down through space surrounding the conductors of the external circuit.

That is a single-pass energy collection fraction for a static charge. Almost all the Heaviside energy flow missed the circuit and passed on beyond the circuit "collectors" in one pass-by at essentially the speed of light, on out into deep space and beyond. On the other hand, if the already-passed EM energy flow is retroreflected, so that it returns for another pass along the circuit, then an *additional* amount of the repassing energy can be collected by those sluggish electrons. In a nominal circuit being used as an example, the electrons travel only about 5 or 6 inches per hour longitudinally down the conductors, restrained by the repulsion of the electrons beyond them. So *multiretroreflection*, *multipass*, *multicollection* can be used to provide excess energy and overunity efficiency. This is a change to the present statement of the work-energy theorem, which up to now has assumed single pass, single energy collection. The well-known anti-Stokes emission from certain highly scattering, optically active media, e.g., produces more energy radiating back out of the media than the amount of stimulating radiation energy input into the media. This is Letokhov's "negative absorption of the medium", which is really the excess emission from the medium due to retroreflection and multiple re-collections.

Also, if the intercepting charge is resonated, it sweeps out a greater geometrical cross section in the streaming energy flow, thereby diverging more and thus "collecting" more energy. In the Bohren experiment, e.g., this effect yields collection of some 18 times as much energy by the resonant charge as the same charge collects in static mode.

EXOTHERMIC

Giving off heat, or emitting scattered EM energy. Producing forward time energy.

FARADAY, MICHAEL

Noted English physicist and chemist, 1791-1867, who discovered field theory and electromagnetic induction. Faraday invented the dynamo also and formulated his laws of electrolysis.

His work directly inspired Maxwell, who vowed to study no other electromagnetics theory until he had captured Faraday's work in a mathematical theory.

FIELD

In mathematics, a set with two binary operations. In electromagnetics in more common terms, it is a region consisting of magnitudes and/or vector directions assigned to some active entity (spacetime itself) at each point in the region.

Designated addition and multiplication, satisfying the conditions that the set is a commutative group with respect to addition, that the set with the identity of the additive group omitted is a commutative group with respect to multiplication, and that multiplication distributes over addition for all elements in the set.

For our purposes, in plain language the fundamental notion is that a "field" is a sort of cleared, level space or region (e.g., in spacetime) where things can be placed and actions can occur amongst them, to them, on them, and of them. What the "place or space" is, can get a bit complicated, as can what is placed there and what actions are permitted. In short, we view a field as an altered (curved) region in spacetime, and this special curvature interacts with certain kinds of entities placed in that region. In return, those entities also interact back upon the spacetime curvature to alter it. Hence there is a dynamic two-way interaction between field and matter, where half of this interaction (the back-interaction of charge or matter) is missing from present classical electrodynamics.

EXOTHERMIC EXPLOSION

An explosion where ordinary heat and electromagnetic energy are released and radiate away from the explosive site. All normal explosions are exothermic.

Particularly used for a pulsed longitudinal EM wave interferometer, which in the distant interference zone (IZ) will produce a sudden eruption of scattering EM energy from the local spacetime curvature suddenly induced, when the interferometry transmitter chassis grounds are biased well above the ambient potential in the distant IZ.

FERMI LIQUID

A liquid where the individual particles or molecules comprising it are basically non-interacting, so that they can roam freely.

In an electron gas in a conductor, e.g., treatment of the free conduction electrons as a gas allows a theoretical modeling and explanation of conduction and resistance. Contrast to a Luttinger liquid, where the electrons interact with each other and thus enable the formation of collective phenomena.

FEYNMAN, RICHARD

Noted physicist, Nobelist, and co-developer of quantum electrodynamics.

Feynman is particularly noted for his refreshing three volumes of physics: See Richard P. Feynman, Robert B. Leighton and Matthew Sands, **The Feynman Lectures on Physics**, Addison-Wesley, New York, 1963. The book is particularly noteworthy in that it gives the student some insight into foundations problems in physics. E.g., in Vol. I on p. 2-4, Feynman defines the electric field not in terms of force per unit charge *per se*, but in terms of its potentiality for producing a force only when a charge is present for the force to be developed upon.

FEYNMAN DIAGRAM

A two-dimensional diagram of particle interactions, both in the virtual and observable levels, and both backward and forward in time and named after Richard Feynman who first employed them.

FIBER FUSE

Effect in fibers or fiber optic cable, whose core contains germanium, whereby heating the cable at one spot with a small flame will launch a focused traveling heat energy effect which melts a small hole in the core about every centimeter, and marches down the cable at about a meter per second to the end of the cable, destroying the optical transmission capability and ruining the cable.

Oddly, by igniting the fiber fuse from the other end of the damaged cable, a reverse fiber fuse effect will often march back down the cable in similar fashion except that it fills in the melted holes in the core, restoring the transmission capability of the cable.

If the energy required to melt all the holes (or fill them again) is taken into account, the amount of energy required to launch the fiber fuse is much less. This phenomenon does not appear to have an adequate technical explanation at present.

For an excellent description with photographs, see Ivars Peterson, "Fibers With Flare," **Science News**, 140(13), Sep. 28, 1991, p. 200-201. For technical information and analysis, see D. P. Hand and P. St. J. Russell, "Solitary thermal shock-waves and optical damage in optical fibers," **IEE Colloquium on 'Non-Linear Optical Waveguides'**," IEE Digest No. 88, London, England, 1988, p. 101-103; — Soliton-like thermal shock-waves in optical fibres: origin of periodic damage tracks," **Fourteenth European Conference on Optical Communication (ECOC 1988),** Sep. 11-15, 1988, Publication No. 292, 1(2), 1988, p. 111-114; — "Solitary thermal shock waves and optical damage in optical fibers: the fiber fuse," **Optics Letters**, 13(9), Sep. 1988, p. 767-769; D. P. Hand and T. A. Birks, "Single-mode tapers as 'fibre fuse' damage circuit-breakers," **Electronics Letters**, 25(1), Jan. 5, 1989, p. 33-34.

See P. St. J. Russell, "Power Conservation and Field Structures in Uniform Dielectric Gratings," **Journal of the Optical Society of America A**, 1(3), Mar. 3, 1984, p. 293-299. Russell is actually referring to apparent violation of local conservation of energy. This would be indicated if the vacuum-to-mass interaction is not taken into account. Russell's analysis therefore supports the fact that (i) the vacuum-mass interaction is required, (ii) the fiber fuse does utilize some of its gated vacuum energy (i.e., the usually non-diverged portion of the Poynting flow), and (iii) the fiber fuse is a legitimate, repeatable overunity process, although a "runaway" process, uncontrolled, and not sustainable.

FIELD

In physics, "a region of space marked by a physical property, as gravitational or electromagnetic force or fluid pressure, having a determinable value at every point in the region."

[Thanks to **Webster's II New Riverside University Dictionary**.]

Here one must *not* think of "space" as emptiness, as mathematicians do, but as a virtual particle flux, a dynamic entity, and as dynamic spacetime. A field then is a set of entities and their relationships and operations in a region of *that* kind of space, where the entities and operations are themselves "space stuff". They are all "virtual" in that they are not individually observable. However, their impact upon or interaction with mass—particularly charged mass—can be and often is observable, in that we can see some change of the mass (its direction of travel, translation, acceleration, deceleration, rotation, generation of current, etc.). If the entities are point-like and are naively supposed to have magnitude only, or if we only specify their energy density magnitude at each point, then that "marked space region" is said to be a *scalar field*. If the entities are point-like but have both magnitude and direction, they are said to be a *vector field*.

Further the deponent saith not because even the foundations physicists are in hot water in trying to define the field, and if the deponent saith more he will surely deep-fry in hot oil!

FIELD, SCALAR

A field, to any point of which is assigned a magnitude only.

FIELD, VECTOR

A field, to any point of which is assigned both a magnitude and a direction.

FLAT (UNCURVED) SPACETIME

A vacuum/spacetime whose virtual particle flux intensity—and therefore its local energy density—does not change as a function of spatial position or averaged time.

That is, a flat spacetime in my view is a quantum mechanical vacuum whose virtual particle flux and energy density are constant in magnitude and internal constitution with respect to spatial position and in averaged time. Note that a flat spacetime (vacuum potential) can still contain vacuum engines, and thus still be *dimensioned*. In that case, contrary to normal physics, the flat spacetime itself will still interact with an embedded object and change it, even though it may not translate it. To perform transmutation of elements, e.g., one is interested in just such engines and not in translating through space the element being treated. This dimensioning or activation of the local flat spacetime can result in an otherwise unexpected, seeming violation of one or more laws of nature in the dimensioned region.

FISSIONING, ACTION

The splitting of the action quantum into two canonical variables (for example, energy and time), only one of which can normally be completely detected or measured. ("Action" has the dimensions of angular momentum—energy x time, momentum x length, etc.)

Fissioning produces length, discrete bit by discrete bit, and time likewise. The physical reality of separated physical objects is thus continually created, discrete bit by discrete bit, monocularly. All that ever observably exists in the "present" is a single quantum change, just occurring. Recombining of the two canonical action fission fragments is called "fusion." Fusion also occurs monocularly, quantum by quantum. Rigorously, the physical world of separated objects never exists as such except in memory recall of vast numbers of past ordered quantum changes that occurred one by one.

FLUX

The amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time.

The term flux density is now more commonly used. Loosely, "flux" refers to the perpendicular flow through an area, while "flux density" refers to the magnitude or intensity (rate) of the flux. In some cases such as a flux of particles of different velocities, the number of particles may be multiplied by the average velocity to give an average flux density that is weighted.

FLUX DENSITY

The amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time.

FOGAL, WILLIAM (BILL)

Soon-to-be-noted researcher and inventor of the Fogal semiconductor (two patents so far), as well as a superluminal infolding-outfolding communication system and other remarkable longitudinal EM wave processes.

I am happy to be a friend and colleague of Bill Fogal.

FOGAL SEMICONDUCTOR

A special semiconductor invented by Bill Fogal which, in its operational regime, utilizes an integrated semiconductor, tantalum capacitor, and feedback resistor to accomplish charge blocking (blocking of electron current flow dq/dt) while passing displacement current d/dt

The Fogal semiconductor can also accomplish amplified phase conjugation of signals as well as infolding (translation of input transverse EM waves to output longitudinal EM waves) and outfolding (translation of received input longitudinal EM waves to output transverse EM waves). Used in communication systems, it opens the use of the unlimited "infolded" electromagnetics bandwidth. Since it may communicate using longitudinal EM waves, it is also usable for superluminal or a specialized "tunneling" communication through the "interior" of normal EM waves, potentials, and fields.

See William J. Fogal, "High Gain, Low distortion, Faster Switching Transistor," U.S. Patent No. 5,196,809, Mar. 23, 1993; — "High Gain, Low Distortion, Faster Switching Transistor," U.S. Patent No. 5,430,413, July 4, 1995, a continuation of his earlier patent.

FORCE

In mechanics, identically the time derivative of momentum, or **F** ≡ d/dt(m**v**).

Simply, changing of the old concept of mass-motion. One must realize that it is not the mass in motion, but the single undivided quantity "mass-motion" or "mass-velocity" (momentum) that is changed. In short, **F** ≡ d/dt(m**v**). In present EM theory, force fields are erroneously taken to exist in empty, massless vacuum. However, the mechanical definition is an identity, not an equality. Thus force does not *separately exist in contact with* a mass, but *consists of* a changing "mass-motion." Since this force can be induced into and of the mass-motion target at a distance, in the new approach one is confronted with the fact that there is a more fundamental mechanism that produces force itself, and that is a curvature of spacetime.

Distant curvatures of spacetime to produce forces on exposed charged mass systems are also engineered by longitudinal EM wave interferometry.

FORCE FIELD

An effect that exists as a local vacuum virtual particle flux interaction upon and with a unit mass or unit charge whereby a particular type of matter is affected.

The force field has two components, one the mass and the other massless. In electrodynamics, one normally thinks (erroneously!) of the massless component, and refers to that component alone as "the force field." This considers the "force field" as a separate entity, acting upon a separate "mass." Whether it is logically correct or not, that is the prevailing way physics and electrodynamics uses it! So in this definition, let us momentarily accept the common incorrect usage, for convenience, while pointing out the real definition as we need to.

For example, conventionally the EM force fields can affect and change the state of motion of any charged particle or any magnetic particle. In classical physics, the force fields are considered the primary causative agents; when these fields are zero, then the motion of the particle or system is unaffected. Consequently, conventional EM theory assumes that all the EM phenomenology ceases once the force fields are absent. Today we know that is not true at all, as we explain in this glossary.

FORCE-FREE PROPAGATION

Propagation of force-free disturbances in the vacuum flux

(i.e., infolded inside the ambient vacuum potential, or inside a constant DC voltage.) Such "infolded" propagation is not limited to the speed of light, but may even travel in the Coulomb gauge so that it is instantaneous. More usually, it can travel superluminally. It may be regarded as a special form of "tunneling", which is known (both theoretically and experimentally) to produce superluminal propagation to the external observer. The music of Beethoven's 40th symphony, e.g., has been experimentally transferred between two points in a physical waveguide by such "quantum tunneling" at over four times the speed of light in vacuum.

We regard tunneling as moving in a sort of "subspace", i.e., within the internal structuring of spacetime itself. It is not limited by the speed of light and Einstein's postulate, which we believe only applies to signals moving in spacetime in bulk, and not to signals moving *within* substructuring inside that spacetime.

FOUNDATIONS OF MATHEMATICS

The concepts, operations, assumptions, postulates, logic, definitions, and axioms of algebra systems.

FOUNDATIONS OF PHYSICS

The postulates, assumptions, concepts, and definitions upon which a scientific physics model rests. Included are the postulates, concepts, assumptions, definitions, and operations of the mathematics in which the scientific model is expressed, as well as the postulates, concepts, assumptions, definitions, and operations of the physics entities and functions assigned to the mathematical symbols.

Thus "scientific truth"—even mathematical truth (see Morris Kline's **Mathematics: The Loss of Certainty**—is actually a *qualified* truth. What is true in one model (one set of postulates, assumptions, concepts, and definitions, including those in the algebra) may not be always true when one or more of the postulates, assumptions, concepts, and definitions is/are changed, and/or when the topology of the mathematics in which the model is expressed is changed, etc.

The major flaws in present physics models involve errors in the foundations. However, a successful physics model is rather quickly dogmatized and often turned into a belief structure by many scientists rather than being regarded as a tentative model. Consequently, scientists who attempt to do serious work in foundations are generally ignored or ostracized and rather vehemently attacked if they propose any significant changes to one of the dogmatized models. Particularly, all funding and "benefits" of the scientific bureaucracy are denied the heretical foundations scientist, unless he or someone can rather quickly come forth with an iron-tight experimental proof of the heresy. Any attempted change in foundations is usually met with ad hominem attack, a barrage of criticism, loss of position, and denial of research funding. Science has a long history of just such dogmatic oppression of great scientific discoveries. To name a few (there are hundreds): The first laser paper was rejected by the journals as being ridiculous. When Mayer proposed conservation of energy, he was hounded and ridiculed so fiercely that he suffered a nervous breakdown and attempted suicide. Years later, near the end of his life conservation of energy was adopted because it greatly facilitated understanding and calculation, and the same system now lionized him as a scientific hero. Ovshinsky suffered personal attack, ridicule, and charges of charlatanism in pushing amorphous semiconductors; everyone knew that only crystalline semiconductors could work. The situation changed when several thousand Xerox machines were observed to now contain amorphous semiconductors which were working just fine! Louis de Broglie's matter wave hypothesis caused consternation to his thesis advisors; but because he was Prince Louis de Broglie, they felt obligated to send it to the great Einstein so that he would reject it and then they could regretfully reject the thesis of the Prince. When Einstein replied that it was probably true, they felt they had no choice but to publish it *because de Broglie was a Prince of the old French aristocracy, and as such was not to be casually trifled with.* So foundations work is fraught with the greatest peril to the practitioner. Yet it is the most important work in physics, and it is a continuing black mark against the scientific establishment that they do not openly encourage *and fund* the most able physicists, grad students, and post-docs to examine and revise the errors in foundations concepts and postulates. Because of the prevailing dogmatism, science itself continually strangles the new ideas and concepts that would provide the physics of the 22nd century now. In an eerie turn of events, science has become in a strange way its own worst enemy, and the strongest barrier to the advance of science.

FOUR-SPACE (4-SPACE)

The "normal" spacetime used in physics, consisting of three spatial dimensions and one time dimension. Minkowski spacetime.

FOURTH LAW OF LOGIC

the law that a thing and its opposite become identical on the boundary, i.e., in multiple simultaneous observation. Identity is merely the loss of observational or perceptual distinction; when formerly "two" objects are simultaneously observed by a single observer, two slides have been shown at once in the slide projector. It does not matter at that point what each was individually observed to be before; now there is no observable distinction or difference. We previously showed that the law is implicitly assumed in the three laws of Aristotelian logic anyway, and published a simple proof of the fourth law. Without the fourth law, Aristotelian logic "eats itself". What results in the four law logic is that either (1) the three laws apply explicitly and the fourth implicitly, or (2) the fourth law applies explicitly and the fourth implicitly. This *application rule* itself can be taken as a fifth law of logic. The fourth law is also the law of the paradox: I.e., when something is true but violates one or more of the Aristotelian laws, it is a case where the fourth law applies explicitly. The conclusions derivable from a four law logic agree generally with the conclusions derivable from G Spencer Brown's **Laws of Form**.

FOUR-WAVE MIXING

A nonlinear multiwave mixing effect when four waves mix in a highly nonlinear situation so that wave to wave interaction occurs.

FRACTIONAL CHARGE

A hypothetical particle having a charge less than the electron's charge. Often called a quark, where its charge is 1/3 or 2/3 the charge of the electron.

A few years ago, Stanford University researchers found evidence of free quarks. There is also now the fractional Hall effect. In the 1920's Ehrenhaft reported the discovery and production of fractional charges of all sizes, and his experiments have been replicated by Michailov, with papers by Michailov and Barrett.

In the new approach, when time-energy is utilized, quarks can be freed or nearly freed rather easily in "time-reversal zones". It would appear that, as the new approach develops, free quarks will become a normal event rather than one intensely sought but terribly difficult to come by experimentally.

FRAME, LORENTZ

A frame of reference which is not accelerated with respect to the laboratory observer.

In this case, the frame may be rotated (moving at a constant velocity) but it is not rotating (the frame is not accelerated with respect to the observer, and spacetime is not curved). In this case special relativity applies, as do the conservation laws. If the local spacetime is curved, the frame is accelerated and spacetime is curved. Locally, then, general relativity applies. In this case, locally the conservation laws (energy, momentum, charge, and spin) can be violated. However, in ordinary general relativity, it has been tacitly assumed that locally the curvature of spacetime is negligible. Hence locally a Lorentz frame is assumed, and the conservation laws are assumed to apply locally also. In scalar electromagnetics, the local spacetime is always curved and general relativity applies locally. In this case, the conservation laws need not apply.

FRAME, REFERENCE)

A spatial, organized, measured lattice placed in "emptiness" (space, spacetime).

Normally refers to a 3-dimensional, spatial frame. All objects and points in the "universe" or spatial frame are considered to simultaneously coexist at separate, measured points in the frame. Differs from the vacuum in that, rigorously, vacuum has no existing definite lengths and no existing definite time intervals, as these appear only after measurement or detection, and are relative to the observer and to the interactions ongoing as well as in the detection process itself. The "laboratory frame" is the static reference frame of the observer or measurement. A separate reference frame may be assumed to exist for any fixed or moving object, or centered on any point in another frame. When a type of frame is assumed, the entire class of physical interactions that can occur has been restricted to an assumed set or type. In other words, given the frame, the *conventional* physics has been assumed. One of the greatest restrictions of an assumed "frame" is to rule out the consideration (existence) of other higher dimensions.

Note that, in the new unified field theory approach, the other higher dimensions are always available and cannot be ruled out in general, but only in some special case. Every curvature of spacetime, and any internal additional curvature to that primary curvature, adds a new dimension. In our view, a spacetime may be "flat" in overall curvature, but consist of internally structured deterministic curvatures or "engines". In this view, normal inertial frames, e.g., may still contain vacuum engines, which will not affect the normal *bulk translation* rules, but may affect any or all of the nontranslation mechanisms, including the very laws of nature in many cases.

FREE ENERGY

Excess energy freely received from its external active environment by an open system that is far from thermodynamic equilibrium in its exchange with that environment.

To use this free excess energy, it must be collected, transported to a load or loads with minimal losses, and then dissipated in the load(s) to power them, without any substantial portion of the free energy being used to close the system's "gate" through which the excess energy is input to it.

There is a separate thermodynamics for such open systems far from thermodynamic equilibrium in their active environmental exchange. Such a system is permitted to exhibit five "magic" functions: it is permitted to (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator himself inputs (the excess energy is freely received from the active environment, (4) power itself and its load simultaneously (all the energy is freely received from the active environment), and (5) exhibit negentropy.

The entropy of such a system cannot even be calculated. Some useful references are: Ilya Prigogine with D. Kondepudi, **Modern Thermodynamics: From Heat Engines to Dissipative Structures**, Wiley, Chichester, 1998; Ilya Prigogine, **The End of Certainty: Time, Chaos, and the New Laws of Nature**, Free Press, New York, 1996, 1997; G. Nicolis and I. Prigogine, **Exploring Complexity**, Piper, Munich, 1987; Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," Chapter 11 in Paul Davies, Ed., **The New Physics**, Cambridge University Press, Cambridge, 1989, p. 316-347.

FREE ENERGY SYSTEM

Any open system not in thermodynamic equilibrium, which freely receives excess energy from an external source,

...and collects and utilizes this free excess energy to freely power external loads, so that a permissible system COP>1.0 is achieved, as allowed by the nonlinear thermodynamics of open systems in disequilibrium with their environment.

Implicit in the notion of free electromagnetic energy systems are six key concepts: (1) the local vacuum/spacetime is known to be filled with hidden (i.e., virtual) energy in violent motion, i.e., the local vacuum/spacetime is a violent virtual energy flux and therefore a scalar potential, (2) a broken symmetry—i.e., some sort of organizing and gating mechanism, such as the common dipole—in this hidden (virtual) energy flux of the vacuum is utilized as a source, to cohere the virtual energy and divert it as energy flow **S** = **E**×**H**, from the broken symmetry initiation point onto the external circuits or load parts of a device, (3) the coherent gated extra *virtual* energy flow available in the external circuits or load parts couples to (interacts with) the mobile charges in those circuits or load parts, so that gradients developed in the interacting virtual flux form force fields on and of the particle masses, driving the mobile charges along the circuit and through current-impeding circuit components such as loads, (4) in the interaction of the driven charges with the driving virtual energy flux, their spin and erratic motions integrate a tiny "coupling fraction" of the driving virtual energy into observable (i.e., macroscopic field) energy, thus creating the **E**-fields and **B**-fields, (5) the current-impeding external components or load parts scatter or dissipate the coherence of the driven charges, and thereby dissipate the collected energy or change the form of the collected field energy, thereby producing useful work, and (6) little or none of the excess collected energy in the circuit is split off and dissipated in the gate (asymmetry component serving as the **S**-flow source) to destroy the gate and thereby destroy its asymmetry in the vacuum flux, which is what is extracting and producing the free energy flow.

FUSION, ACTION

In the new approach, the recombining of two canonical bits of a fissioned action quantum.

The fusion process "cancels" the bits produced by the fission process, and constitutes a negative spacetime operation. The physical world is thus eliminated, bit by bit, in a monocular fusion operation that produces the "past". The "mass" of a particle is taken to be proportional to, and identically comprised of, the absolute value of its fission and fusion rate. When fission and fusion rates of a particle are equal, the particle is unaccelerated; when unequal, the particle is accelerated. Vice versa, when the particle is unaccelerated, the fission and fusion rates are equal, and when the particle is accelerated, the fission and fusion rates are unequal.

GAIA HYPOTHESIS

The hypothesis by Lovelock that the earth and its biosphere is a living thing, at least in some sense.

In the new viewpoint, there exists a biospheric quantum potential which contains an extended mind and therefore is indeed an extended lifeform—one which includes the earth and the quantum potentia all its creatures and species. In the new sense, even so-called "inert" matter is alive, being continually bathed by energy flow dimensioned with lifeform templates, and therefore equally alive in all directions and nonpreferential.

GATE

Refers to something (a component or function) which extracts some part of a flux or flow, compacting it into a stream, and "sending it out."

A moving board wall or gate may be swung into a river, e.g., to divert or "gate" some of the river's flow out into a sluice ditch and on to a waterwheel powering a mill. Similarly, a dipole extracts some of the energy from the vacuum virtual photon flux exchange with the end charges of the dipole, and "gates" (diverts) the extracted portion of energy out from the ends of the dipole as energy flows. This creates the scalar potential surrounding each charge, and also the **E**-fields. Even a "static" dipole does this "gating" of vacuum energy into an energy flow.

The fundamental gate for all EM energy in 3-space is the dipole or dipolarity. The source of the 3-space EM energy continuously emitted by any charge or dipole is an equal inflow of EM energy from the complex plane (from the time domain). Thus EM energy flow is conserved in 4 dimensions but not in 3-space. In particle physics, it is well-known that any dipole or charge is such a broken 3-symmetry in its fierce vacuum energy flux exchange. However, particle physics considers the input energy to be virtual particle energy from the vacuum and therefore random energy. Whittaker 1903 shows it is actually from the time domain of spacetime (from the complex plane) and the input energy flow from the vacuum is not random but precisely organized in a harmonic set of longitudinal EM phase conjugate waves. Of course by wave-particle duality we simply consider the two pictures to be just different and permitted models of the same thing.

GATING EFFECT

The effect of extracting part of a flux or flow, compacting it into a useful stream, and "sending it out."

See *gate*.

GAUGE

In electromagnetism, usually considered a possible choice for (change of) electric scalar potential and magnetic vector potential, without any net change of force fields so that the change satisfies Maxwell's equations.

Interestingly, gauge field theory includes the axiom of gauge freedom. In EM systems, this means that one is free to "regauge" or change the potential energy of the system freely and at any time, so long as the change results in a static stress potential energy change. In the real world, one may have to pay a little bit for switching. But essentially physics already assumes that one can cause his EM system to take on excess potential energy at will, so long as it is a *stress potential* energy change only.

In short, only the *symmetrical* regauging case has been considered in electrodynamics. Yet every electrical power system must first potentialize the system *asymmetrically* rather than symmetrically. In the theoretically perfect system, this asymmetrical regauging is done freely. Hence every electrical power system first breaks gauge symmetry, in order to take on extra energy so that it can do work. Further, the source of the potential energy flowing into the asymmetrically regauged electrical system so freely, comes from the source dipole's extraction of EM energy from the time-domain (complex plane) due to its broken 3-symmetry in EM energy flow.

So the highest "gauge" principle is that energy can be freely extracted from the vacuum, anywhere and at any time, freely and easily, by making a dipolarity—which changes a single potential and asymmetrically regauges the system.

It follows that one is in theory free to "dissipate" this excess free energy in external loads to power them. Again, in the real world one may have to pay some switching costs, but that can be minimized by efficient switching. However, since no conventional electrical power system accomplishes this free powering of its loads, it rigorously follows that the conventional system must self-enforce symmetry in its discharge of the free potential energy achieved by the initial *asymmetrical regauging*.

Thus, if we believe gauge field theory—the most modern of physics theories and in many ways the most useful—and also believe its assumption of free asymmetrical regauging that is hidden in its gauge freedom principle, then COP>1.0 electrical power systems are absolutely permitted by nature and by gauge field theory where asymmetrical regauging is deliberately used.

It follows that some feature in electrical power systems as we presently design and build them, must be preventing these systems from exhibiting overunity COP. And so it is. The ubiquitous closed current loop circuit results in the circuit forcibly and symmetrically regauging itself while dissipating its excitation energy to power its loads and losses. This forces half the collected regauging (free) energy to be dissipated in the external circuit's loads and losses, and half of it to be dissipated in the source dipole to scatter the charges and destroy the dipole. This means that less than half the available free regauging energy is used to power the load, while half is used to destroy the entity—the source dipole—actually furnishing the free regauging energy from the vacuum.

Hence all our present electrical power systems are inherently free energy systems freely extracting their energy from the vacuum, but are fiendishly designed so that they kill themselves (shut off all free asymmetric regauging energy from the vacuum, by destroying the source dipole) faster than they catch some of the free energy from the vacuum to power their loads.

To restore the source dipole requires at least as much energy dissipation as was done to destroy it. In our conventional self-destructive systems, the generator shaft must thus be continuously powered in order to form the magnetic field inside the generator, so that this magnetic field can then dissipate its available energy to force the scattered charges back apart to form an ordered source dipole.

We pay the power company to have a giant wrestling match inside its own generators and *lose*. And in a sense, we do it because our theoreticians love gauge symmetry more than they tolerate broken gauge symmetry. So they build lovely, symmetry-restoring electrical power systems—and monstrously mangle and destroy the fragile biosphere in the process.

GAUGE FREEDOM

The axiom in gauge theory that one can freely change the potentials of an EM system, so long as the net new forces resulting from the change will sum to a zero vector resultant so that the system remains "symmetrical" in its force-field functioning with respect to the system prior to the change of potentials.

Gauge and gauge freedom are actually formal restrictions of an even more important asymmetrical regauging principle: *a potential in a system can be freely changed at any time, thus freely changing the potential energy of the system freely and at will*. Symmetrical regauging is achieved only by multiple application of this asymmetry rule to provide a net symmetry condition again, with a change in locked-in stress potential energy of the system.

GAUGE SYMMETRY

Abstract mathematical symmetry of a field, that relates to the freedom to regauge (change the value of) potentials, without affecting the values of the field quantities.

In short, the higher principle that the potentials can be changed freely and at will—and hence the potential energy of the regauged system can be changed freely and at will—is arbitrarily restricted to a very special case: Each change of potential involves the appearance of an new free force field. This *asymmetrical* change can be made freely and at will, at any time. However, the notion of gauge symmetry restricts any set of such changes to the special case where the new free force fields produced must sum to a net zero force field of opposing forces. In short, it severely restricts the freedom to change the potentials—to the special case where the resulting change of the system's potential energy must remain a static stress potential energy condition. Note that this arbitrarily eliminates any overunity functioning of the regauged system in which the altered potential(s) and altered system potential energy exist(s). See the discussions of *gauge* and *gauge freedom*, above.

GAUGE THEORY

A field theory using a field that has one or more gauge symmetries.

Note the accent on gauge symmetry and the net *symmetrical* regauging restriction. This restriction is a carry-forward of the old notion that the potentials are not primary, but the fields are primary causes of all EM phenomena.

For the overunity researcher, broken gauge symmetry is the interesting feature which must avidly be sought in his experimental systems. Hence he must seek out and familiarize himself with a wide variety of *broken symmetry effects* and phenomenology. He must also rigorously analyze his circuits with a keen view to where gauge symmetry can be broken, where and how gauge symmetry is automatically restored during excitation discharge, etc. Unless the automatic self-enforcing of symmetrical gauge transformation during excitation discharge is violated, no EM system can produce COP>1.0. In addition, without elaboration, the overunity researcher must also learn to consider the time-domain functions of the circuit, and particularly the giant negentropy of the common dipole. Unless he understands that all EM energy comes from the vacuum and from the complex plane via the broken 3-symmetry of a dipolarity, he cannot understand his own circuits.

Electromagnetics was the first gauge theory. Gauge theory is widely utilized today, and particularly in particle physics. In overunity power systems, the system must exhibit a broken gauge symmetry at least twice: (1) in initially potentializing the system to freely increase its potential energy, thus exciting the system so that it has energy with which to do work, and (2) then discharging this free excitation energy in the load without simultaneously destroying the source dipole providing the continual potentializing.

The conventional COP<1.0 electrical power system restores net 3-space symmetry by continually destroying the "free asymmetrical regauging" component: the source dipole, and using more of the energy freely received from the vacuum via dipole asymmetry to destroy the dipole than is used to power the load. This requires that at least as much 3-space energy be input to the system again, to restore the dipole and thus potentialize the external circuit, as was used to destroy the dipole. Hence as much 3-space energy has to be input as is dissipated in all the loads and losses, including inside the generator. For that reason, our scientists and engineers unwittingly but deliberately design and build only COP<1.0 systems.

Without elaboration, we also accent that spacetime curvature is always involved in an overunity EM power system, and the tempic (time) potential of the output end of an overunity system differs from the tempic potential of the input end, so that the gradient (the tempic force) is directed from output to input—exactly the reversal of the case for a COP<1.0 power system.

GAUGE TRANSFORMATION (IN ELECTROMAGNETICS)

In EM, gauge transformation is conventionally taken to be the addition of the gradient of some function of space and time to the magnetic vector potential—and simultaneously the addition of the negative of the partial derivative of the same function with respect to time, divided by the speed of light, to the electric scalar potential.

This procedure gives different potentials but leaves the electric and magnetic fields unchanged. It restricts gauge transformations to *net symmetrical* transformations. However, the increase or decrease in potential energy of the system, due to changing the potentials, results in a change in the stress of the system and a change in its stress energy. This is because two new antiparallel forces are created, which change the stress of the system.

In essence, gauge symmetry then involves the implicit assumption that a functioning system is the same regardless of how it is stressed, and regardless of the state of its stress energy. We point out that, gravitationally, this is not true since any change in the local potential energy of the system is a spacetime curvature change, hence a gravitational change.

GENERAL THEORY OF RELATIVITY

Einstein's theory of gravity in which the gravitational force is represented by a curvature in spacetime.

GIANT NEGENTROPY (OF THE DIPOLE)

The receipt from the time domain of spacetime (the vacuum), by any dipolarity, of enormous longitudinal EM wave energy flow, transduction of that absorbed energy from the complex plane into real 3-space, and emission of that enormous energy flow in all directions in 3-space, automatically and indefinitely after a little energy has been expended to separate the charges and make the dipole. The EM waves in the input from the time domain and the output in 3-space are perfectly ordered and correlated, which is a giant ordering of the vacuum/spacetime and is thus a giant negentropy.

In circuits, the source dipole in the generator or battery, once made, performs this function. Only a minuscule fraction (some 10^{-13}) of the energy pouring out of the generator's or battery's terminals strikes the surface charges of that attached external circuit and is diverged into the conductors to power the Drude electrons and the circuit. All the rest is just wasted.

See my paper, "Giant Negentropy from the Common Dipole," *ibid.*

GRAVITON

The quantum (smallest piece) of the gravitational field, with spin 2.

In the quantum theory of general relativity, the graviton is massless and has spin 2. In the new approach, a graviton is considered to be a paired (coupled) photon and antiphoton, each of which has spin one, so that the graviton couplet has spin 2. A scalar potential is comprised of gravitons, not merely photons, since it can be decomposed into a harmonic series of bidirectional wavepairs. In each wavepair, a wave and its antiwave precisely superpose. We insist on a further strong condition: that in the biwave, the photons comprising the wave and the antiphotons comprising the antiwave also must obey the *distortion correction theorem*. This means that the antiphotons and photons are passing through each other in opposite directions, continually coupling into gravitons and decoupling again.

GRAVITY

The phenomenon characterized by the physical attraction of any two material bodies.

Actually, in modern general relativity it is the trapped energy in the masses that produce the gravity potentials. Ideally, the gravitational force between two masses is proportional to the product of the masses divided by the square of the distance between them. In the new approach, one recognizes that mass itself is defined in terms of, and identically consists of, the time derivative of an action flux. [For a force-free fundamental definition of mass, see T.E. Bearden, **Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena**, National Technical Information System, Report AD-763210, 1973.] A photon flux is also an action flux. A virtual photon flux is an action flux. So the vacuum consists of a gigantic action flux—which with certain interactions forms mass itself. But since photons are action quanta and photon flux is action flux, and since mass and gravitation are functions imposed on such action flux; then mass, EM, and gravitation become different aspects of the same thing. All of them can be electromagnetically engineered, directly, as indeed can be the flow of time itself. By affecting the flow of time electromagnetically, one is thus able to affect the gravitational force between two objects electromagnetically. Antigravity, for example, can be achieved by a mechanism that is given in this glossary. Hutchinson has demonstrated the levitation of objects weighing up to 60 lbs., albeit sporadically and uncontrollably. At my urgent request, Sweet performed a highly successful antigravity experiment some years ago, reducing the weight of a 6-pound object by 90%, smoothly and controllably, in several smoothly controlled steps. I consider that the experiment tested and validated my approach to an engineerable mechanism for antigravity. For the results of the test, see Floyd Sweet and T.E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," **Proceedings of the 26th Intersociety Energy Conversion Engineering Conference (IECEC '91)**, Boston, Massachusetts, 1991, p. 370-375.

GROSS PARTICLE TRANSLATION

Refers to translating the entire particle in space, rather than internally structuring the particle's potential (massless charge).

GROSS POTENTIAL GRADIENT FIELDS

The usual gradient in a scalar potential's gross magnitude, as in $-\nabla\phi$.

GYROELECTRONS

Spinning electrons considered as gyroscopes, so that when longitudinally perturbed they can and will laterally precess.

Considering (in a gross first order model) the conduction electrons in the Drude gas in a conductor as little gyros, because they are (i) spinning on an axis, (ii) severely restrained in their ability to move forward longitudinally down the conductor as current, and (iii) much more free to precess laterally in the conductor. Thus when the gyroelectrons in the receiving wire antenna, perturbed by an incoming longitudinal EM wave in the vacuum, are longitudinally perturbed, they precess laterally so that our instruments (which detect electron wiggles, not the incoming vacuum EM wave) do detect "transverse waves." However, they are detecting the transverse gyroelectron precession waves. The old electrodynamicists, long before the electron or the atom was discovered, simply envisioned the incoming EM wave as "waves in the thin material fluid ether" which were "intercepted" as electric fluid waves in the antenna. Hence since their instruments reported that the intercepted "electric fluid" was vibrating transversely, they firmly believed (erroneously) that this proved the ether wave to be transverse also, just as those "taut string waves" that Maxwell and Faraday had believed in. That is the total *raison d'etre* behind the mistaken notion of the transverse EM wave in the vacuum ether!

HARMONIC SERIES

Consisting of a fundamental frequency, its first harmonic, its second harmonic, etc.

HEAVISIDE, OLIVER

Noted English self-taught physicist and brilliant electrodynamicist, 1850-1925, who played a role in discarding Maxwell's quaternions, and also played a role in forming vector mathematics and formulating the vector reduction of Maxwell's theory from 20 quaternion equations in some 20 unknowns to the present 4 vector equations.

What are presently taught as "Maxwell's equations" in university are actually Heaviside's equations.

Lorentz symmetrically regauged the Maxwell-Heaviside equations, again greatly simplifying their mathematical solution—and also unwittingly discarding all permissible overunity Maxwellian systems in the process.

HEAVISIDE ENERGY FLOW COMPONENT

(1) The enormous nonintercepted and nondiverged energy flow component pouring out of the terminals of the battery or generator, and missing the external circuit entirely so that it is wasted.

(2) The enormous energy in the EM field, potential, or wave that misses the interaction of the field, potential, or wave with the assumed unit point static electric charge.

(3) See also discussion under Giant Negentropy.

HERACLITUS

To be added.

HERTZ, HEINRICH

Noted German physicist who together with Heaviside and others transformed Maxwell's theory from quaternion algebra to vector algebra, and who also produced EM waves, thus confirming Maxwell's theory.

HIDDEN BIWAVES

Those bidirectional EM wavepairs that comprise the scalar potential are indeed "hidden" with respect to our normal "electron translation" (i.e., electron wiggle) detectors and instruments.

HIDDEN VARIABLE THEORY

A class of quantum mechanical theories which considers that the quantum state of a physical system is not a complete specification after all. The "hidden variables" are those additional components necessary to provide the "complete state" of the system.

The hidden variables are those extra components of the theorized complete state, which are not contained in the quantum state. Bohm's hidden variables theory is the best known of all of these theories, and makes the same predictions as does the conventional quantum mechanics. In addition, it eliminates the terrible problem of the "missing chaos," eliminates the quantum measurement problem, and it suggests that quantum mechanics can be engineered if the hidden variables themselves can be engineered. In the new approach, one goes a step further by considering hidden EM variables infolded inside the potentials themselves, including inside the quantum potential utilized by Bohm. The advantage of this assumption is that it poses an *engineerable* interpretation, after work by Stoney, Whittaker, and Ziolkowski. Further, the Russians have utilized this type of EM hidden variable theory, along with the quantum potential, to develop secret advanced weapons. The physics of this area is called *energetics*.

HIERONYMUS, T. GALEN

Founder of the American version of radionics, who patented a radionics machine and process to start the cycle of radionics in this country.

HILBERT, DAVID

Noted German mathematician, 1862-1943, who did advanced work in geometry, theory of invariants, integral equations, and mathematical physicist.

Noted particularly for his development of Hilbert space.

HOLE (ELECTRON HOLE)

An unfilled vacancy in an ordering of electron energy levels, as in a lattice or in the Dirac Sea.

Electron energy levels have two energy states: (1) *positive energy* state, and (2) *negative energy* state. The positive energy states are associated with the positions of normal energy electrons. The negative energy states are almost all filled, and this is the Dirac sea.

The absence of an electron from either energy state creates a particular kind of hole. When an electron is absent from an electron positive energy state position in a lattice, that results in an excess of positive charge at that point vacated by the electron. Note that when an electron then refills that hole (recombination occurs), the electron remains in a positive energy state, does not disappear, and no radiation occurs.

When an electron is lifted from an electron negative energy state in the Dirac sea, that leaves behind a hole in the negative energy state in the Dirac sea. When an electron subsequently refills that hole, it "falls" in and disappears, decaying to a lower energy level and emitting radiation due to the difference in the energy levels.

The difference between holes in the Dirac sea and holes in the electron positive energy levels in a lattice, are very different. In short, they lead to two quite different kinds of "hole and electron" interactions. In the lattice "hole-filling" (recombination), there is no excess radiation given off and the electron does not disappear from the positive energy continuum. In the Dirac sea hole-filling, excess radiation is given off and the electron disappears from the positive energy continuum.

Under the influence of an electric field, a *positive energy* hole in a crystal lattice—where an electron has left and is therefore absent—will effectively propagate through the lattice as a positive charge, but usually much slower than the electron propagates in the opposite direction as an electron current. The process is actually more complex and involves continuous interaction with the electrons and their physical distribution pattern and location. The hole current is mathematically treated as a special positive charge current, often referred to as a positron current. In a current-carrying conductor or semiconductor, one has a current of electrons traveling in one direction and simultaneously a current of holes (hole positrons) traveling in the opposite direction, but recombining.

As stated, in addition to lattice holes, there also can exist unfilled holes in the Dirac negative energy states comprising the vacuum, particularly when fields are present. In this case, it is a *negative energy* hole. A positive energy electron that falls into a Dirac sea hole will emit energy as it falls to that lower energy state. To lift one of the electrons from the Dirac sea and leave a hole, energy must therefore be added to the negative energy electron in the Dirac hole.

When fields are present, Dirac holes may exist and may move as hole current in circuits under certain conditions. The situation is complex; see Felix Finster, "Definition of the Dirac Sea in the Presence of External Fields," **Adv. Theor. Math. Phys**., Vol. 2, 1998, p. 963-985. Updated copy (version 4) is on http://xxx.lanl.gov/abs/hep-th/9705006.

When a free electron fills a positive energy hole in a lattice (positive electron energy state), this effectively disposes of both the hole and the *free* electron which is now a *bound* electron, and is called *recombination*. The electron does not disappear. In a material lattice, there is no energy radiation from this, just the disappearance of the hole and the "binding" of a free electron so that it is no longer free.

In visualizing positive hole currents, passing the hole current though a surplus of available electrons (as in a negative charge donor material) will result in "hole filling" and recombinations. One may therefore speak of disposing of unwanted hole current by recombination, and in fact this is widely accomplished in semiconductors by controlling the doping.

HOLE CONDUCTION (LATTICE)

The movement of electron vacancies (holes) in a lattice continually left behind by hopping of the electrons as they break loose from their bonded positions and move freely in an electron current.

When an electron hops out of a lattice bond, it leaves behind a bonding gap or "hole" in the positive electron energy state at that bond site, so that an extra positive charge appears there at the hole. As the electrons move in one direction in a conducting lattice, the holes move in the opposite direction in a "positive current", but at a much slower pace than the electron current moves.

HOLE CURRENT (LATTICE)

See *hole conduction (lattice)*.

HOMEOPATHY

A system of medical treatment based on the theory that certain diseases can be cured by giving very small doses of substances which in a healthy person would produce symptoms like those of the disease. (Opposed to allopathy).

Homeopathy is based upon formation and use of dimensioned potentials (particularly the potential composed of the extremely rapid "making and breaking" hydrogen bonds in a fluid) containing vacuum engines. Any material or solute, and therefore any molecule, contains its own specific vacuum engine, its own charges, and its own dimensioned potential. Dissolving the material in water, or even using a suspension of it in water, allows the bonding potential in the water to interact with the potential of the solute or suspension. Potentials superpose, and they also diffuse and mix their inner structures (vacuum engines) into each other. So the material or solute diffuses its vacuum engines into the interior of the hydrogen bonding potential, in effect ordering and templating the hydrogen bonding itself! Vigorous shaking of the suspension or solution assures that all the fluid and H-bonding potential is exposed to the solute of suspended particles, so that a thorough exchange of vacuum engines occurs. This process in homeopathy is called *potentizing*. It's actually just creating the desired internal vacuum engines in the H-bonding potential of the water. Once that's done, there's really no further need for the solute or the suspended particles. By repeatedly diluting the solution to the extreme, then one can reach a point where there is a high probability that not a single molecule of the original solute is remaining in the fluid. Yet in each dilution, shaking can be used to insure the even spread of the vacuum engines through the newly introduced H-bonding potential in the newly introduced water. At the end, there exists an altered fluid whose H-bonding potential now contains the desired vacuum engines. Drinking that water now introduces those vacuum engines into the body fluids, bloodstream, etc. The ingested dilute new vacuum engines will then diffuse through the various potentials of the body, and produce altered chemical functioning of the kind that could have been induced by the chemical or active solute material. The advantage is that the harsh material residue of the chemical or solute is not present in the body, as it is when allopathy is utilized.

HYPERFUNCTIONING

Functioning in hyperspace as opposed to 4-dimensional Minkowski space.

Higher topological functioning.

HYPERSPACE

Refers to a space of more than four dimensions, and specifically to those spatial dimensions outside the normal three.

IDENTITY OF OPPOSITES

On the boundary, A and not-A are identical. If all of a thing is collected, one reaches the boundary of that thing and it turns into its own opposite.

INERTIAL FRAME

A frame of reference in which force-free bodies move along straight lines, and the postulates of special relativity are valid.

In the new approach, this notion is extended; even in an inertial frame, force-free bodies may move along straight lines, the postulates of special relativity may remain valid, and yet the EM hidden variable may induce other changes in the usual laws of nature due to the action of hidden vacuum engines infolded in the *dimensioned* local spacetime (vacuum potential).

INFOLDED HIDDEN DYNAMICS

Refers to:

(i) Bohm's hidden variable theory and the internal dynamics of the hidden variables/entities themselves, and

(ii) to the hidden biwave pairs comprising a scalar potential, with the internal dynamics of those waves such as tuning off harmonic frequency, multiplication, variation of phasing and magnitude, etc.

Also, by Whittaker 1903 the scalar potential can be decomposed into the bidirectional wavepair set in harmonic series. By Whittaker 1904, each of the waves in a biwave pair can be further decomposed into two potentials, so each pair of waves would consist of four potentials. That is, the entire set of internal biwave pairs can be replaced by four times as many internal "partial potentials" as was the number of biwave pairs. Then Whittaker 1903 can be applied once again to break each one of those scalar potentials into new biwave pairs, and so forth and so forth. So the potential can be expressed as waves within waves within waves, or potentials within potentials within potentials, and so on. The neat thing is that each successive infolding (internesting) is a higher dimensional functioning and a higher topology. So one can build (with some little pain!) specific infolded wave patterns that will work in—say—the 7th dimension of hyperspace, or the eighth dimension, and so on. This is why I called the internal structuring of a potential (or the two potentials comprising an EM wave) *dimensioning* the potential or the wave. As a half-baked nuclear engineer whom the Army never let practice his discipline, I wanted some way to directly engineer the nucleus. *That has now been found*. Unfortunately it has also already been highly weaponized in secrecy by one or more foreign nations. In other words, one can design a signal and assemble it (with some difficulty) that will flip one quark in a nucleon in an atom. That way one can transmute an element either into the next higher element isomer or the next lower element isomer. Using a quantum potential to "transport" the necessary vacuum engines, the effect can be accomplished at a great distance, in multiple targets simultaneously, and the effect is not shieldable. "Jumps" in nuclear transmutation can also be accomplished. The migration of electrons to and from an atom (as in solution, e.g.) makes and negates ions continuously. Thus continually the external potentials and the nuclear potentials intercommunicate. This means that their vacuum engines continually diffuse. Any dimensioning of the external potentials will be passed into the nuclear potentials as a dimensioning action also. If certain conditioning vacuum engines are present in the external potentials, then the slow diffusion into the nuclear potentials begins to affect the nucleons and their interactions, the quarks comprising the nucleons, the particles such as pions continually exchanged between nucleons, etc. The mechanism for nuclear transmutations and production of new species is there! I believe, e.g., and am working to more clearly express, that the major mechanism producing the new nuclear species in cold fusion experiments just such dimensioned potentials, involving the inadvertent dimensioning of potentials comprised of the hydrogen bondings of the water and other factors.

INFOLDED REAL VECTOR COMPONENTS

Well, simply take a set of real force vectors that altogether sum to a vector zero resultant.

That "vector zero system" is a zero with respect to forced *translation* of a particle or mass. It is not an absence per se, but is the presence of specifically hidden and "infolded" real vectors. Two elephants pushing each other to a draw form such a "zero-translating" system. That "zero system" is quite different from two crickets pushing against each other to a draw! Anyone who thinks all zeros are "absences" and who thinks all zeros are equal, is invited to try standing between the crickets and observing how he feels, then standing between the elephants and see how he feels. The latter might be a real bummer. So zeros can differ! And a zero can be a presence of many real things, all still there but "infolded" and "hidden," but just the absence of any *one and only one* thing. Even in arithmetic, e.g., we are free to pull 5-3-2 right out of 0, or pull out 4-7+8-6-2+5+1+1-3-1/2-1/2 (if I got the arithmetic correct!). And so on. So we have always *used* the fact that a net zero can have a real internal composition, without detailing it exactly. We teach our schoolchildren to do it, but fail to tell them what they are doing!

When you sum EM force fields to zero, the fields are still there and real. You have altered the energy of the system or volume, and also the potential energy of the system or volume. And you have done it in specific directions, magnitudes, etc. along the "fingers" of the infolded patterns. In short, you have made a specific vacuum engine. Mass, fluid, nuclei, molecules, etc. exposed to that "conditioned vacuum potential" may well undergo physical and chemical and electrical changes. Golden once performed experiments which powerfully conditioned the local vacuum, and all clocks and watches (mechanical, electrical, etc.) in the local area were awry for four days.

INFOLDING

With reference to signals, infolding means "placing the signal or modulation or change upon one or more of the internal bidirectional wavepairs comprising the scalar potential.

Since all EM potentials ultimately represent changes to the ambient vacuum potential, one has conditioned or altered the vacuum itself when infolding is utilized. Infolding of EM signals, vectors, and directions produces *vacuum engines*. By Whittaker's principle and Bearden's corollary, these vacuum engines can be designed to affect matter—including atomic nuclei, nucleons, etc.—in any manner desired.

INFORMATION CONTENT OF THE FIELD

Russian euphemism for the structuring of the hidden longitudinal EM waves and their dynamics, that comprise any EM potential, field, or wave.

INFRASONIC

Below sonic.

Frequencies below what the ear can hear. Below about 20 Hertz.

INTERFEROMETER

An instrument in which a wave is split into two waves or beams, which after traveling over different paths are subsequently reunited and display interference.

INTERFEROMETRY

Measurement of wavelengths of light and very small distances and thicknesses,

...for determining indices of refraction, and for analyzing small parts of a spectrum by means of the interference phenomena of light.

INTERNAL SPACE

To be added.

ITERATED

Done repeatedly, like picking beans one at a time.

JOSEPHSON EFFECT

Flow of Cooper pairs (superconducting electron pairs) across a thin dielectric separating two superconducting electrons, in the absence of a voltage drop.

That is, the pair of electrons can "tunnel" through a thin insulating barrier.

JOSEPHSON JUNCTION

The thin dielectric separating two superconducting electrons, with subsequent Josephson tunneling of Cooper pairs.

JOSEPHSON TUNNELING

See *Josephson effect*.

JOULE

A measure of energy.

One joule equals one watt for one second. In physics the unit of work or energy in the MKS system, being the amount of work done by one newton acting through a distance of one meter, equal to 10,000,000 ergs.

JUNG, CARL GUSTOV

Noted Swiss psychiatrist and psychologist who founded Jungian analytical psychology, formulated the principle of synchronicity, and advanced and delineated collective unconscious mind.

From the content of his German patients, Jung predicted WW II two decades before it occurred.

KINDLING

Coherent integration (adding) of similar virtual components among the vast "noise" of dissimilar virtual components, so that the "signal-to-noise" ratio of the integrated signal steadily increases, and the integrated "observable signal" grows out of the background virtual state noise.

This "kindling" process is directly analogous to coherent integration of nondiscernible radar signals far beneath the received noise, to build and extract the real radar return signals from targets. This sort of process dramatically increases the "sensitivity" of the radar. Real, very tiny signals even 100 dB or more beneath the noise level can be retrieved in such fashion by modern radars.

Similar processes can integrate very minute signals already passed into the virtual state and presumed to be nonexistent. Usually, however, the difference frequency between two frequencies at least one harmonic interval apart must be used, considering the vacuum's virtual flux state as an isotropic nonlinear medium.

One may also hypothesize that potentials and propagating signals may decrease with distance, but are never truly "lost" in that they never reach a total zero. When they reach the quantum level, further decrease in amplitude simply results in their passing into the virtual state (subquantum level).

KIRCHHOFF'S CURRENT-LAW

At any time, the sum of the instantaneous currents flowing into a node or point in a circuit equals the sum of the instantaneous currents flowing out of that node or point.

KIRCHHOFF'S VOLTAGE LAW

At any time, the sum of all voltage rises in a closed loop circuit equals the sum of all drops in that closed loop circuit.

LAMB, WILLIS EUGENE JR.

American physicist and Nobelian who discovered the hyperfine structure of the hydrogen spectrum, and first measured the interaction of the vacuum energy that generates a change in the energy levels of the hydrogen atom, according to Dirac's theory and quantum electrodynamics.

LAMB SHIFT

A very small correction to the first excited state of the hydrogen atom due to interaction of the vacuum.

Willis Lamb experimentally confirmed this prediction of quantum electrodynamics to high accuracy.

LAMELLAR CURRENT

Current in thin layers.

Thin layers of current. Thin sheets of current.

LASER

A device that generates light in which all the photons are exactly in step and produce a coherent beam.

Laser light has one wavelength and is more easily controlled than other kinds of light. Use of a laser to perform interferometry is one way to create holograms; use of a laser to illuminate a hologram is one way to read it. Originally the word "laser" was an acronym that stood for "light amplification by stimulated emission of radiation." Note that when TiO_2 particles in suspension, with a small amount of fluorescent dye added, are weakly illuminated by a small coherent light source (a laser), a tremendous room-filling output of coherent light emission occurs, by the anti-Stokes emission effect. This is lasing without population inversion, and it is a validated and easily testable overunity process. This is the process patented by Lawandy.

LASING WITHOUT INVERSION

The amplification of coherent light in an intensely scattering medium, comprised of certain powdered laser crystals or suspensions of similar material, where a variety of phenomena such as gain narrowing and amplifying random walk of light propagating in the medium, combine to give laser emission with diffusive feedback.

Gain tends to depend on the volume of the medium, so that at a certain medium the gain in light amplification by the medium exceeds the losses, and the laser action results above this critical volume. The intensity of the light scattering (multiply reflecting) in the medium then will increase rapidly until an "explosion" of light results in an intense flash of coherent light in all directions from the medium volume.

If the excitation is still present, the process will be repeated and the energy density will again exceed the critical value, rapidly generating another intense flash of light. This process can be repeated many times, given the presence of continued excitation.

A very interesting phenomena that can occur is coherent backscattering. Here the light scatters over both a forward-time path and a time-reversed path, perhaps with each path consisting of more than 1,000 scatterings. Because of time-reversal symmetry, the two waves (forward and backward) acquire exactly the same (though random!) phase. So the two waves after all this 1,000 scatterings on their paths, meet and interfere constructively in the direction back toward the exciting light source.

In theory the mean free path can be made shorter and shorter, so that all the light would be trapped inside the medium. This is called Anderson localization, and it has been reported in the literature; see D.S. Wiersma, P. Bartolini, A. Lagendiji, and R. Righini, "Localization of light in a disordered medium," **Nature**, Vol. 390, 1997, p. 671; John Sajeev, "Frozen light," *ibid*., p. 661-662.

This work grew out of original work called out by Letokhov but largely ignored for more than two decades.

References: D. Wiersma and A. Lagendijk, "Laser action in very white paint," **Physics World**, Jan. 1997, p. 33-37; — "Light diffusion with gain and random lasers," **Physical Review E54**, 1996, p. 4256; N.M. Lawandy et al., "Laser action in strongly scattering media," **Nature**, Vol. 368, 1994, p. 436; V.S. Letokhov, Generation of light by a scattering medium with negative resonance absorption," **Sov. Phys. JETP**, Vol. 26, 1968, p. 835.

Note that we have nominated a version of this process, together with consideration of the erroneously discarded Heaviside nondiverged energy flow component, as the fundamental mechanism proposed to be involved in gamma ray bursters, x-ray bursters, and in some gamma ray and x-ray flashes associated with atmospheric phenomena on earth in thunderstorms, etc. We also nominated the missing Heaviside energy flow component as the agent accounting for the excess gravity (some 90% of the total gravity) in the arms of spiral galaxies that hole those arms intact so they do not fly apart. See T.E. Bearden, "Dark Matter or Dark Energy?", **Journal of New Energy,** 4(4), Spring 2000, p. 4-11 (also available on http://www.cheniere.org); — "Giant Negentropy from the Common Dipole," **Proceedings of Congress 2000**, St. Petersburg, Russia, Vol. 1, July 2000 , p. 86-98. Also published in **Journal of New Energy**, 5(1), Summer 2000, p. 11-23. Also carried on DoE open website http://www.ott.doe.gov/papersbooks.html and on http://www.cheniere.org.

LAW OF PHYSICS

A restriction placed upon the physical mechanisms that can occur under certain broad assumptions as to the background situation and environment.

Actually, in nature there are no such rigidly fixed "laws"; these are only man's synthesis for the behavior of large classes of systems under very general conditions. Any "law of physics" can be violated if the conditions contained in its assumptions are violated. Conservation laws are often touted as "laws of physics." Actually, the conservation laws assume a closed system or one in equilibrium, and a linear spacetime (Lorentz frame). If either or both of these assumptions is locally violated, any or all of the conservation laws can be violated locally, even though the overall global conservation is still upheld. Quite simply, in that case one has opened the system (broken symmetry). In particle physics, broken symmetry already results in violation of one or more of the conservation laws. In leading Russian physics journals, papers are regularly published containing unrestricted general relativity with concomitant violation of any and all conservation laws.

LAWANDY PATENTS

Patents covering and related to anti-Stokes emission (excess emission of energy) in strongly scattering, optically active media such as TiO2 particles in suspension.

More energy is emitted than one must input to stimulate the effect. It is being hailed as a new form of lasing, without the requirement for population inversion.

LETOKHOV, V. S.

Russian pioneer in excess emission phenomena from stimulated media.

LETOKHOV EFFECT

The excess emission (overunity emission of energy) phenomena in strongly scattering, optically active media, usually referred to as "negative absorption of the medium".

For an overview, see V. S. Letokhov, “Stimulated emission of an ensemble of scattering particles with negative absorption,” **ZhETF Plasma**, 5(8), Apr. 15, 1967, p. 262-265; —“Generation of light by a scattering medium with negative resonance absorption,” **Sov. Phys. JETP**, 26(4), Apr. 1968, p. 835-839; — “Laser Maxwell’s Demon,” **Contemporary Physics**, 36(4), 1995, p. 235-243.

LOCALITY

Characteristic where interactions are local in spacetime.

LOCALIZED

Confined to a small region of a large system rather than being extended through the system or at a distance.

LOCAL VIRTUAL PHOTON FLUX OF VACUUM

Refers to a local region of the vacuum, which also refers to the vacuum flux density of that region.

LONGITUDINAL COMPONENT

In EM waves, the longitudinal component of the fields along the direction of travel,

...as opposed to the transverse components at right angles to the direction of travel.

LONGITUDINAL E-FIELD

We may divide any electrical field into two components, a transverse component and a longitudinal component.

Of course one or the other may be a zero vector.

LONGITUDINAL GRADIENT

Compression wave gradients (differences) in density or pressure,

...considered in a direction from their origin along the direction of primary movement, vs. transverse waves, which act perpendicular to that direction.

LONGITUDINAL INTERIOR WAVE

In conventional theory, the electric portion of the EM wave is a transverse field oscillation, and the longitudinal component is essentially zero.

In the new "infolded" wave approach, there is no overall transverse gradient in the vacuum potential, but there is an infolded longitudinal gradient which oscillates "inside the vacuum potential." This is rather like a "pressure wave" transmitted under water, without causing a surface wave involving physical upheaval and fall-off of the water. In EM, by analogy the speed of the "bulk upheaval" transverse wave is nominally c, the speed of light in vacuum [but see discussion of galloping waves, for what really happens!]. The speed of the pressure wave, under water, may be very much greater than c, because it is a special "tunneling" process and it also involves the galloping wave effect. Experimental measurement of superluminal tunneling velocities—such as more than 4c—are already well-known in the literature.

To better visualize this "pressure wave," consider that it is like a sort of "velocity modulation" upon the edge of the conventionally-visualized plane wave front of a traveling transverse wave, except now the plane itself does not move. All the succeeding planes are "bumped" by the velocity modulation of one plane, and so the velocity modulation propagates through (inside) the vacuum potential as a modulated pressure wave, without transverse change of the potential.

Now we have a very special medium! It consists of bidirectional EM wave pairs, one passing through the other. In such a case, it is already known that the speed of a "planar wave front" in the superposition and resulting interference of two oppositely moving waves can in fact greatly exceed the speed of light momentarily, then slow dramatically below than the speed of light, etc., so long as the average velocity is c. This is called the galloping wave effect. [E.g., see William G. Harter et al., "Galloping waves and their relativistic properties, **American Journal of Physics**, 53(7), July 1985, p. 671-679.]. Since the plane wavefront within the interior *can* in fact vary its velocity, we are quite free to use this "velocity pulse modulation" effect of wavefront speed variation, to place signal intelligence upon the planar wave front speed variations. So we can transmit hidden pressure waves with intelligence on them, through the interior of the vacuum potential. *Further, the restriction to an average velocity of c does not have to apply, and the planar pressure pulses can be transmitted much faster than c*. This is the basis of the work Fogal and I are struggling with, in order to achieve superluminal communication. It is also what we have been struggling with in an effort to come up with a communication system that can propagate infolded signals inside the H-bonding potential of the ocean itself, so that high-data-rate communications with deep submarines could be utilized, at any distance, and the huge, gawky, frightfully expensive, and environment-affecting ELF systems be done away with. We also think it should eventually be possible to develop an underwater radar type of system, but it's far too early at this point to try to conceptualize just exactly how it would work.

LONGITUDINAL WAVE

A "pressure" type of wave, similar to sound, in which the vibrations are along the direction of travel of the wave.

Hence, a wave composed of alternating densifications and rarefactions, where we focus upon the longitudinal component of the changes. In the past we have used the term "scalar waves" to imply longitudinal standing waves, such as in one infolded Stoney-Whittaker wavepair inside the scalar potential. In classical EM theory, the Poynting vector predicts no longitudinal wave of energy from a time-varying, electrically charged source. In fact, even in the ordinary theory, an exact solution of the problem does allow this longitudinal wave, as shown by Keech and Corum, **International Journal of Theoretical Physics***,* 20(1), 1981, pp. 63-68. Conventionally, as explained elsewhere in this glossary, the classical electromagnetic wave is modeled as a vector string wave, comprised of transverse vibrations on implicitly assumed material taut strings in the vacuum. In the new approach, such transverse vector force field waves (Hertz waves) do not exist *as such* in vacuum; instead, two coupled, dynamic scalar potentials exist there. The **E** and **B** fields are formed and exist as the these potentials and their gradient coupling to spinning charged detecting mass particles.

LONGITUDINALLY

Of or in length. Running or placed or operating or changing lengthwise. Opposed to transverse.

LORENTZ, HENDRIK ANTOON

Renowned Dutch physicist and Nobelist, 1853-1928, who developed the Lorentz transformations and Fitzgerald-Lorentz contraction.

He shared the Nobel Prize for discovering the Zeeman effect.

LORENTZ INVARIANCE

An inertial frame is independent of the velocity of the frame relative to any other frame.

LORENTZ INVARIANCE PRINCIPLE

The principle that the laws of physics are the same in all reference frames.

LORENTZ REGAUGING

Changing the two potentials in the Maxwell-Heaviside equations in potential form so that the resulting equations have the variables separated.

An unfortunate effect is that this discards all Maxwellian systems far from equilibrium in their vacuum exchange, hence discards all permissible Maxwellian COP>1.0 systems.

LORENTZ SYMMETRY

Referring to the symmetrical regauging by Lorentz of Maxwell's equations.

LORENZ, LUDWIG VALENTIN

Scientist who first changed Maxwell's equations—two years after their publication in 1865—by (effectively) symmetrically regauging them, thereby unwittingly discarding all Maxwellian systems far from thermodynamic equilibrium with the active vacuum.

See L. V. Lorenz, "On the identity of the vibrations of light with electrical currents," Philosophical Magazine, Vol. 34, 1867, p. 287-301. In this paper Lorenz gave essentially what today is called the Lorentz symmetrical regauging.

Comment: (gratefully received from T. Barrett) This paper gave f(t-r/c) functions. Fitzgerald said that Lorenz's functions were essentially the same as his, and Fitzgerald became a leading proponent of "retarded potentials". But it is believed that Fitzgerald was unaware of Lorenz's work until the 1880's, so he is given credit for parallel development. Some people talk of Fitzgerald-Lorenz functions. This is a regauging, but the term "gauge" (inspired by railroad gauges) was first introduced by Hermann Weyl in the 1900s. He used it for a change in length and was shot down by Einstein's theory—Weyl's theory was not relativistic. The idea resurfaced in the 1920s when quantum theory was being formulated, but this time it meant "change in phase" and not "change in length".

Comment by T.E.B.: Thus "Lorentz" regauging of Maxwell's equations really was first done by Lorenz in 1967. However, not too much attention was paid to L. V. Lorenz' work (e.g., by Fitzgerald). When H. A. Lorentz later used symmetrical regauging (essentially Lorenz' regauging), his influence was so great that it was adopted straightaway and has remained adopted to this day.

LOW ENERGY NUCLEAR REACTION (LENR)

Cold fusion, or the anomalous transmutations of elements which occur in certain electrolytes with specially prepared palladium electrodes.

The transmutations occur even though in conventional theory there is insufficient energy to accomplish them.

Conventional science has rather vehemently attacked cold fusion and many of the cold fusion scientists, even though there are now hundreds of successful experiments in multiple laboratories throughout the world.

In our view, once the properly-prepared palladium lattices are loaded with positive ions, the formation of temporary, fleeting time-reversal zones in the electrolyte temporarily reverses the law of attraction and repulsion of charges. Thus H+ ions, e.g., may attract each other in even numbers (opposite numbers are forbidden by the Pauli exclusion principle), so closely that each enters the strong (nuclear) force region of the other, so that the group becomes a quasi-nucleus. Note that the gluons are partially "unglued" by the TRZ formation, and so the preferred method of "decay" of the anomalous TRZ state as the surrounding electrolyte ions move to negate it, is by flipping of a quark. One D+ ion coupled to one H+ ion, e.g., when attracted into a quasi-nucleus in a TRZ, may decay into tritium by flipping one quark and changing one of the protons into a neutron. Two H+ ions may decay into deuterium (D+) by flipping one quark and turning an H+ into a neutron. Two D+ ions may simply "tighten further" into a real nucleus as the TRZ decays, and result in an alpha particle (helium nucleus). The TRZ hypothesis does precisely explain the most important reactions that are experimentally shown to occur. For more on this hypothesis, see T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," **Explore**, 8(6), 1998, p. 7-16; — ., "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," Journal of New Energy, 3(2/3), 1998, p. 12-28.

MAGNETIC DIPOLE

Paired north and south magnetic poles of equal strength over microscopic or macroscopic distance.

MAGNETIC INDUCTION

The act of a change in magnetic field producing an electric field at right angles, which in turn produces electron flow.

MAGNETIC SPIN

Refers to the intrinsic angular momentum, known as spin, of electrons, protons, and neutrons, etc.

The electrons combined in an atom or ion have a resultant angular momentum, that is the combined intrinsic spin of the electrons and the angular momentum due to their motion around the nucleus. There is also a magnetic moment associated with this angular momentum (spin) whenever it is nonzero. Therefore atoms or ions with nonzero spin are *magnetic* atoms or ions. This is often loosely referred to as "magnetic spin." It is actually the magnetic moment of nonzero spin.

MAGNETOELECTRONICS

New microelectronic components which exploit the spin of the electron rather than its charge.

References: A Barthelemy *et al*., "Giant Steps with Tiny Magnets," **Physics World**, Nov. 1994, p. 34-38; G. Prinz, "Magnetoelectronics," **Science**, Vol. 282, 1998, p. 1660.

MAGNETORESISTANCE

A change in electrical resistivity in a material when a magnetic field is applied to the material.

MAGNETORESISTANCE, COLOSSAL

A very large change in electrical resistance observed in the manganese perovskite materials when a magnetic field is applied to them. The effect in these materials is the largest magnetoresistance effect known at present.

See N. A. Babushkina *et al*., "Metal-insulator transition induced by oxygen isotope exchange in the magnetoresistive perovskite manganites," **Nature**, Vol. 31, 1998, p. 159; L. I. Balcells *et al*., "High-field magnetoresistance at interfaces in manganese perovskites," **Phys. Rev**. **B**, Vol. 58, 1998, p. R14697; M. Blamire *et. al*., "Understanding and utilizing colossal magnetoresistance materials," **Phil. Trans. Roy. Soc. A**., Vol. 356, 1998, p. 1469-1712.

MAGNETOSTATIC SCALAR POTENTIAL

Roughly, the "pole strength" differential between the two poles of a magnetic dipole.

Anyone who wishes may polish this definition.

MASSLESS DISPLACEMENT CURRENTS

Currents without mass.

Examples in the standard theory include $d\phi/dt$, $d\mathbf{E}/dt$, $d\mathbf{P}/dt$, $d\mathbf{B}/dt$, etc.

MATTER

Highly condensed 3-spatial energy, compressed by the factor c^2 …

and with the observation process d/dt having been applied to the fundamental masstime entity to provide mass as a frozen snapshot. Matter, of course, has internal structure, order and organization, and dynamics. All matter is in a continuous energetic exchange with its vacuum (4-spatial) environment.

MAXWELL, JAMES CLERK

Brilliant scientist who first formulated the unified theory of electricity and magnetism.

James Clerk Maxwell was born on June 13, 1831 in Edinburgh, Scotland. In 1847 he entered the University of Edinburgh, then transferred to Cambridge in the fall of 1850. After graduation, he stayed on at Cambridge in a research position. He was elected a Fellow of Trinity College and placed on the staff of college lecturers. In 1856 he returned to Scotland, where he took up a Chair of natural Philosophy at Marishal College, Aberdeen. In autumn, 1860 he took a new position as Chair and Professor of Natural Philosophy and Astronomy at King's College, London (a position he held until 1865, at which time he resigned). Maxwell was financially independent. He was elected to the Royal Society in 1861, while at King's College. From 1865 to 1871 he resided at his ancestral Scottish country home, Glenlair, developing his major ideas into book form. Maxwell returned to Cambridge in 1871, where he became the first holder of the Cavendish Chair of Experimental Physics. There he also supervised the construction and operation of Cavendish Laboratory. His treatise on electromagnetism appeared in 1873. He held his position at Cambridge until he died on Nov. 5, 1879 at age 48, of a form of stomach cancer—the same ailment that had killed his mother when he was a child. [Thanks to **Encyclopaedia Britannica**!]

MAXWELL (UNIT)

Unit of magnetic flux, in the cgs system.

MAXWELL'S ELECTRODYNAMICS

Maxwell's electrodynamics is Maxwell's equations.

MAXWELLIAN SYSTEM

Any physical or electromagnetic system whose electrodynamic operations obey James Clerk Maxwell's electrodynamics model.

The term is usually applied with respect to those systems which obey a highly reduced subset of Maxwell's after the Lorentz regauging of the Maxwell-Heaviside equations. This subset is now inappropriately referred to as "Maxwell's equations". This is unfortunate. As a result, today many scientists and most engineers no longer understand that Maxwell's theory—when embedded in a higher topology algebra—permits a vast richness of additional EM systems and behaviors, including full unified field theory operations. In the standard U(1) gauge symmetry electrodynamics, all these higher symmetry functions and systems are excluded *a priori*. From our view, this is particularly sad since these arbitrary reductions of Maxwell's theory arbitrarily excluded all EM systems far from equilibrium in their exchange with the active vacuum. Hence almost all scientists and engineers believe it is against the very laws of nature to propose an electrical power system that produces more energy out—and more work in the load—than the energy input made by the operator himself. This sad error has resulted in (1) crippled electrical power systems that extract their output energy from the vacuum but kill themselves faster than they power their loads, (2) the giant pollution of the earth, (3) extraordinary energy costs, (4) struggles and wars over oil and energy resources, and (5) the looming giant economic collapse of the economies of many nations. Further, unless this latter facet is checked, it may also lead to (6) the unleashing of all the weapons of mass destruction in the arsenals of many nations, destroying all civilization and much of the biosphere.

McCREA

To be added.

MICHELSON, ALBERT ABRAHAM

German-born American physicist who invented an interferometer (Michelson interferometer) named after him. Michelson used his interferometer in a renowned experiment designed to measure the velocity of the earth through the ether, which would produce a shift in his interferometer pattern. No such shift was observed, which indicated the absence of an ether wind and was largely responsible for the downfall of the concept of the material ether which had been assumed previously.

Reference: Albert A. Michelson and Edward Morley, "Influence of Motion of the Medium on the Velocity of Light," **American Journal of Science**, Vol. 31, 1886, p. 377-386; — "On the Relative motion of the Earth and the Luminiferous Ether," **ibid.**, Vol. 34, 1887, p. 333-345.

MIND

From the medical view, with thanks to **Webster's Medical Desk Dictionary**, Merriam-Webster Inc., Springfield, Massachusetts, 1986: (1) The element or complex of elements in an individual that feels, perceives, thinks, wills, and especially reasons; (2) the conscious mental events and capabilities in an organism, (3) the organized conscious and unconscious adaptive mental activity of an organism.

See an extended discussion of *mind* under *energy*.

MIND-BODY COUPLING

To be added.

MIND-MATTER INTERACTION

The n-space interaction between the time-like mind and its time-like operations and the 3-spatial physical body and its 3-space operations. To first order, we may treat these interactions in four dimensions.

MIND OPERATIONS

To be added.

MOTION

To be added.

MOTIONLESS ELECTROMAGNETIC GENERATOR (MEG)

Transformer-like COP>1.0 electrical power generator invented by Bearden, Kenny, Hayes, Moore, and Patrick which powers the transformer core with a permanent magnet, but separates the curled magnetic vector potential A (the magnetic field B) from the uncurled magnetic vector potential A, so that the magnetic field flux is retained in the core while the A-potential is replenished outside the core and adjacent to it.

The overunity mechanism is thus provided by a version of the Aharonov-Bohm experiment, which proves that this splitting of the A-potential does occur and that the uncurled A-potential will interact energetically with electrons. In short, by "splitting the potential" and potentializing two paths to the same energy density as the original unsplit potential, the MEG doubles the available potential energy available via the permanent magnet dipole's broken 3-symmetry in the active vacuum. Both the curled A-potential (magnetic field) inside the core and the uncurled A-potential outside the core interact simultaneously with the output coils when pulsed. Thus the output coils "doubly interact" instead of singly interacting. This "double-dipping" of available energy flow allows permissible COP>1.0, since the system is far from equilibrium with the active vacuum exchange.

Pulsing applies the Lenz law effect to both the split potentials, further increasing the energy density of the two interactions with the output coils. The MEG nominally exhibits from COP = 1.7 to COP = 10, using two input coils.

See M.W. Evans, P.K. Anastasovski, T.E. Bearden et al., "Classical Electrodynamics Without the Lorentz Condition: Extracting Energy from the Vacuum," **Optik** 111(6), 2000, p. 246-248; — "Inconsistencies of the Maxwell-Heaviside Theory of Electrodynamics: The Aharonov-Bohm Effect," **Journal of New Energy**, 4(3), Winter 1999,p. 236-240; — "Vacuum Energy Flow and Poynting Theorem from Topology and Gauge Theory," submitted to **Physica Scripta** (in review); — "The Aharonov-Bohm Effect as the Basis of Electromagnetic Energy Inherent in the Vacuum," submitted to **Optik** (in review), — "Explanation of the Motionless Electromagnetic Generator with O(3) Electrodynamics, **Foundations of Physics Letters**, 2001 (in press). See also T.E. Bearden, "Energy from the Active Vacuum: The Motionless Electromagnetic Generator," in M.W. Evans (Ed.), **Contemporary Optics and Electrodynamics**, Wiley, 2001, 3 vols. (in press), comprising a Special Topic issue as vol. 114, I. Prigogine and S.A. Rice (series eds), ***Advances in Chemical Physics***, Wiley.

MULTICOLLECTION

Multiple collection of the same energy flow, by rerouting the passed Poynting energy back around to the collectors for another pass by them, with additional collection.

A common means of "returning" the Poynting flow for additional collection is retroreflection, as in phase conjugate reflection. Since a nominal energy collection fraction on a single pass is only about 10^{-13}, a great many re-passes of the same energy can be returned. Work is done on the collector in each pass. One joule of energy can actually perform many joules of work; this is a change to the usual statement of the work-energy theorem, which assumes single pass, single collection only.

MULTIPASS RETROREFLECTION

Iteratively retroreflecting the same Poynting energy flow, each time it impinges on and passes by an intercepting collector, so that multiple passes of the same energy flow are made onto and by the collector.

More and more of the passing enormous energy is collected as more and more passes by are made. The collected energy is then dissipated as work in the circuit or device. Such processing can allow overunity EM circuits and devices.

MULTIPASSES

Multiple passes. See multipass retroreflection.

NANOTECHNOLOGY

The fabrication of devices with atomic or molecular precision, to produce operating devices with minimum feature sizes less than 100 nanometers.

Often described as a marriage between chemistry and engineering. The self-assembling of the units is referred to as *molecular manufacturing*, and the devices are built one atom or molecule at a time with programmed nanoscopic arms. As occurred with the modern machine and technology age, the first requirement is to develop the tools that do the working, and the "tools that build the tools." That fundamental work is off the ground and running already.

Nanomanufacturing plants capable of manufacturing bulk materials or arbitrary structures with atomic precision are envisioned, with almost every atom being placed precisely as desired.

Nanotechnology may well produce the next great scientific revolution. As an example, nanobots have been visualized that will circulate in a human body's bloodstream and specifically attack cancer cells. Nanobots for cleaning up air and water pollution are possible. Food synthesis is envisioned since the technology provides the ability to combine molecules directly.

As with all technologies, the ethics lies in the hands of the system designer and the user. Increased technology always allows greater use for good and greater use for bad purposes. Because of the extraordinary "evil uses" that can be employed in the use of nanobots, Bill Joy, chief scientist of Sun Microsystems, recommends no further development of nanotechnology and no further research. He stated,*"… we have the possibility not just of weapons of mass destruction but of knowledge-enable mass destruction… hugely amplified by the power of self-replication… we are on the cusp of the further protection of extreme evil…"*

Joy's fears are well-founded. Already the union of artificial intelligence with self-replicating nanobots is foreseen. If that union is successful, one can foresee the unleashing of what can only be called "species" of self-replicating, smart nanobots bent on the mass destruction of humanity in various ways.

Also, contrast the nanobots to our discussion of even more advanced nanotechnology of causal system robots (CSRs) using functioning system of causal fields (sets of longitudinal EM waves and their dynamics, prior to having interacted with charged mass, and thence being systems that are functioning sets of spacetime curvatures). We have already survived one massive world-wide insertion of an armada of such CSRs—fortunately first generation systems and with no countermeasures, so they were easily detected and destroyed. We may not survive the next such insertion and attack, as the maturity of the CSRs increases.

Considering the usual focus of governments and human control systems, my prediction is that *detrimental, self-replicating, smart* nanobots for use in war, surprise strategic attack, and mass destruction of human targets is already on the early planning boards. The obvious war applications of nanotechnology will advance faster than the peaceful applications, behind a public screen of simpler benevolent developments. This is almost certain, since the lead time for maturing the nanotechnology is from 10 to 20 years. This long lead time rules out most private capital ventures of substantial magnitude. It results in the requirement for government funding, just as development of the atomic bomb was government funded. As a nanotechnology "arms race" begins and heats up, one will see the equivalent of new but hidden "Manhattan Projects" in specialized mass destruction weapons nanotechnology. Rather substantial funding of nanotechnology research and development by several nations is already a fact.

Unethical aspects aside, the potential for useful and benevolent purposes of nanotechnology is very great. The development in this new field is now beginning to escalate almost exponentially, year by year, funded primarily by governments and to a more limited extent by private companies.

Reference: David Cook, Jessica Voorhees, and Brian Valz, "Nanotechnology: The Coming Molecular Revolution," available on website http://nanotech.about.com.

NEGATIVE ABSORPTION OF THE MEDIUM

Excess emission from a medium, as in the Letokhov work and in Bohren's experiment. The excess energy is received from the vacuum interaction.

NEGATIVE RESISTOR

Any component or function or process that freely receives energy from outside the system in unusable or disordered form and outputs that energy in usable, ordered form inside the system, where that is the net function performed.

We specifically do not include "differential" negative resistors such as the tunnel diode, thyristor, and magnetron which dissipate and disorder more energy from the system, overall, than they order and furnish to the system in their negative resistance regime.

NEGENTROPIC

Functioning to produce or involve negentropy.

Reordering previously scattered energy without requiring the operator or experimenter to input energy to do the reordering work (an example is passive retroreflection of scattered EM energy).

NEGENTROPIC ENGINEERING

Expending a little bit of energy to form a dipole or dipolarity, and then intercepting, collecting, and using some of the enormous EM energy freely extracted by the dipole from the vacuum indefinitely, without using any of the extracted EM energy to scatter the charges in the dipole and destroy the dipole, thereby shutting of the vacuum's furnishing of copious free energy.

NEWTONIAN 3RD LAW

For every action there is an equal and opposite reaction.

This law should be extended and restated something like this: "for every action there is an opposite and equal reaction if the causative mechanism of the reaction is allowed to occur and not first redirected away from its target." An example of a violation of Newton's third law as usually stated, is provided by the emission of a phase conjugate replica wave from a phase conjugate mirror material. No matter how powerfully pumped, the PCM does not recoil. The reason is that the antiwave cause of Newtonian third law recoil was intercepted by nonlinear multiwave interaction before it could reach its target nuclei, and rediverted away from the atom and on out of the material. Since the causative mechanism for Newtonian recoil did not occur, the PCM did not recoil.

NEWTONIAN RECOIL

As used in the paper, refers to Newton's third law reaction created as the recoil of the nucleus of an atom which absorbs an incoming EM longitudinal wave, or emits one.

NEWTON'S THIRD LAW

For every action there is an opposite and equal reaction.

In Maxwell's electrodynamics, Newton's third law is missing. It is missing because of the coarseness of the modeling which Maxwell was forced to use. Maxwell's seminal paper was read in 1864 and published in 1865. At the time, neither the electron nor the atom had been discovered, much less the nucleus of the atom. A molecule was just a blob with no structure. Electricity was considered to be a thin material fluid, flowing through the wires much like water through a pipe. In fact, Maxwell actually wrote a material fluid flow theory, since hydrodynamics was developed at the time.

So there was, at the time, no notion of the Drude electron gas in a wire and the positively charged nuclei in the atoms in the wire. So there was no notion of two oppositely charged entities in the conductor, both reacting to any EM field entering from space and interacting with the wire.

Further, Newton's third law had always been used mystically without any known cause. It was experimentally established that the reaction occurred, but scientists were forced to just assume its occurrence.

Consequently, Maxwell wrote equations for the reaction of a unitary electric fluid in the conductors. He knew that the wire recoiled, but wrote no equations containing the cause of Newton's third law reaction.

What really happens in the wire is that both the Drude electrons rearrange their positioning vis a vis the lattice, resulting in an apparent current. The actual movement (drift velocity) of the electrons down the wire is very slow, in a nominal case being a few inches per hour. The electrons, having spin, act as little gyros so to speak, longitudinally restrained by the repulsion of the electrons ahead. So the electrons precess laterally, and spend most of their movement in lateral precession movement. This alone proves that the perturbing force is longitudinal, as indeed does Whittaker's decomposition of the scalar potential into longitudinal EM waves. The gradient of the scalar potential (the E-field) is also such a decomposition. But the electron precession waves—which is what our instruments measure in the wire—are indeed lateral waves, in general consonance with Faraday's notion that the force field lines were physical and like taut strings. Maxwell merely assumed Faraday's notion, to give us the transverse EM wave in space. Consequently, the old electrodynamics envisioned that fluid vibrations (ether vibrations) from space had been intercepted by the wire to shake the electric fluid accordingly. And since that electric fluid obviously shook laterally, that "proved" that the waves in spacetime were lateral "plucked taut string" type waves. Actually it proved no such thing, because they were measuring electron precession waves.

The positive nuclei usually have a m/q ratio several hundred thousand times the m/q ratio of an equal charge of electrons. Hence the massive nuclei also react to the field interaction, and move very slightly in the opposite sense from the electrons. That reaction is ignored in Maxwell's general equations, though when the equations are applied, they require that nuclei movement.

In short, the electrons precess and the nuclei recoil oppositely, and that is the generation of Newton's third law reaction.

When we excite a wire antenna with a wave excitation, we produce both electron precession waves and nuclear recoil waves in the nuclei. These two waves are of equal energy, though the nuclei waves are very highly damped in magnitude because of the high m/q ratio of the nuclei. Nonetheless, each of these "lateral" waves slaps the vacuum potential with *identical energy*. So two equal waves of spacetime curvature occur. I can be seen that the net vacuum wave is actually a longitudinal EM wave, composed of waves of vacuum (spacetime) energy rarefaction and densification. Electrodynamicists, considering the vacuum to be filled with a material ether, used the "effect" wave in matter as the cause wave. Hence they used a transverse EM wave in spacetime, when indeed the spacetime wave is and always was a longitudinal EM wave, just as Tesla insisted.

Maxwell's equations are based primarily on the interaction of field and single charge. Actually, the reaction is always between the field and a dipole. Even when the field interacts with an "isolated charge", there are clustering virtual charges of opposite sign around the "isolated" charge so that the "isolated charge" is actually a set of composite dipoles, where each dipole consists of a virtual charge of opposite sign and a differential piece of the isolated observable charge. So the Newtonian third law reaction is also there with an isolated charge, but occurs in the vacuum itself, resulting in vacuum polarization.

Nonetheless, the "transverse EM wave in vacuum" notion is so embedded in electrodynamics that we seem to be stuck with it. It means, however, that the "field in space" used in electrodynamics must be regarded as only a sort of "convenient transfer function" and not a real transverse oscillation entity. This of course accords with Feynman and Wheeler's pointing out that only the potential for the EM wave exists in space, in case a charge is brought in to interact with its gradients.

And needless to say, since the vacuum is a big potential and only potentials and potential changes can exist in it, then Whittaker's 1903 paper decomposing the scalar potential into longitudinal EM waves is paramount and decisive. EM waves in the vacuum a priori must be longitudinal waves, because they are only the changes to the vacuum potential. And that potential at each and every point, changes and all, decomposes into longitudinal EM waves in 3-space with convergent longitudinal EM waves incoming from the time domain.

NISBET

To be added.

NON-ABELIAN ELECTRODYNAMICS

To be added.

NONLINEAR MATERIAL

A substance whose optical properties vary with the intensity of the light.

Most optical substances are very weakly nonlinear, but a great deal of research has been exerted to find materials with stronger nonlinearities. Also, a material which creates and emits the fundamental and additional harmonic frequencies when illuminated with a single frequency. Many nonlinear materials also weakly produce time-reversed (phase conjugate) fundamentals and/or harmonics when illuminated with a fundamental frequency. Such a nonlinear material that creates time reversed waves (phase conjugate replica waves) is said to function as a *phase conjugate mirror*. The phase conjugate replica reappears in space back along the path in space—even through a highly distorting medium—where the input stimulus (signal wave) has traveled or is traveling.

Thus the phase conjugate replica or time-reversed wave normally backtracks its signal wave stimulus to the distant source of that signal wave, even hundreds or thousands of miles away. This "crawfishing" and "concentrating" characteristic is also known as *retroreflection*. Note that an ordinary EM wave *diverges* as a function of the radius traveled away from its emission source, while a time-reversed or phase conjugate replica wave will *converge* as a function of the radius traveled away from its emission source (the mirror), if its stimulus signal wave was emitted from a distant point source. Ideally, a phase conjugate mirror produces a single time-reversed wave in addition to the input signal wave. By stressing *(pumping)* the mirror with the input of two additional opposing waves (ideally orthogonal to the input or *signal wave*), the phase conjugate replica wave output by the mirror is amplified, and may contain up to all the energy in the pumping waves.

NON SEQUITUR

A conclusion or inference not following from the original argument or premise.

NONEQUILIBRIUM THERMODYNAMICS

Colloquial term referring to the new thermodynamics of self-organizing systems and structures in disequilibrium with their environment.

Since the 1960s, there has emerged in physics (and in thermodynamics!) a surprising new area: the physics of far-from-equilibrium systems and the self-organization and large-scale ordering of such systems. E.g., a single Bénard cell (a particular self-organizing circulation structure in convection-type systems), may involve the coherent behavior of some 10^{21} molecules, in a region of convection where one would classically expect only randomness. These new physics and thermodynamics interactions are surprisingly being found in a very wide variety of systems and phenomena, including living systems and nonliving systems. The phenomena of interest include dissipative systems in disequilibrium with their environment, but with stability of their own systems and forms. Open systems in disequilibrium with their environment can permissibly achieve COP>1.0, and they can exhibit local negentropy. The physics and thermodynamics of such systems is still rapidly developing. For a good overview, see Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," in **The New Physics**, edited by Paul Davies, Cambridge University Press, 1989, p. 316-347. For a full technical presentation, see G. Nicolis and I. Prigogine, **Self-Organization in Nonequilibrium Systems**, Wiley, New York, 1977.

NONLINEAR MATERIAL

In physics, any substance whose response to an external causative or influencing agent is not proportional to the magnitude of the causative agent or influence.

NONLINEAR OPTICAL FUNCTIONING (NLO FUNCTIONING)

Functioning in a manner prescribed by nonlinear optics and quantum optics,

...particularly with respect to time-reversal effects, use of phase conjugate mirrors, phase conjugate reflection, and pumping for amplified phase conjugate reflection.

To quote Knight, "A phase conjugate mixing device has remarkable optical properties: perfect retroreflection, the complete cancellation of optical aberrations and inhomogeneities in wavefronts, and the ability to reverse an optical beam and return it to its source."

NONLINEAR PHASE CONJUGATE OPTICS

See *phase conjugate optics*.

NONLOCALIZATION

Producing or rendering an effect or effect-producing agent at a distance, rather than at the immediate location of the effector.

Extended throughout a large region or system, or to a distant part of a large region or system, rather than being confined to a small region of the large region or system.

NON-RIEMANNIAN (GEOMETRY)

Riemannian geometry is the geometry of Riemannian manifolds;

i.e., manifolds having an infinitesimal squared distance ds that is everywhere positive. A non-Riemannian geometry violates this property. [A manifold is a topological space with certain technical characteristics.]

NUCLEONS

Name for the neutrons or the protons in a nucleus. In the nucleus, the neutrons and protons are continually changing one into the other by exchange of the charge.

O(3) ELECTRODYNAMICS

To be added.

OPEN SYSTEM

A system that communicates with its environment, and exchanges energy, matter, or both between system and environment.

The system may be in equilibrium with that exchange, so that it effectively does not retain and utilize received environmental energy to power loads or accomplish external work. Note that it may in fact receive and store excess energy from the environment, or give up energy to the environment, and then remain in equilibrium under its new potential energy conditions. That is, the system may asymmetrical regauge itself but not utilize any of the regauging energy to accomplish external work. When potential is applied to an electromagnetic system, the system is similarly asymmetrical regauged and its potential energy is increased. However, if the system then destroys its source of regauging potential energy faster than it powers its load, it is said to symmetrically regauge itself when discharging its freely received excitation energy.

The system may be far from equilibrium in its energetic exchange with its environment, and if it asymmetrically discharges its freely received excitation energy, it can exhibit COP>1.0. In that case the system effectively retains and utilizes some of its freely received environmental energy to power loads or accomplish external work, and the operator himself does not have to furnish that part of the energy dissipated as useful work.

OPTICAL SWITCHING

Refers to switching of optical signals, or switching using optical frequency switching techniques, components, and circuits.

ORTHODOX TRANSISTORS

Ordinary transistors in use today.

Such transistors will not perform the Fogal semiconductor special functions.

ORTHOGONAL ROTATION

Simply speaking, a rotation from one direction to a direction at right angles to it.

The Lorentz transform of special relativity, e.g., is a formula for the rotation of an object and all its parts away from its line of motion, as a function of the object's velocity, in a direction toward the time axis. The speed c just represents a full orthogonal turn, so that the intersection of the object with the observer 3-space, along the original line of motion, is zero.

To the observer, the three dimensional object (at speeds less than c) has become a two-dimensional plane moving at speed c. In a space of more than 4 dimensions, the orthorotation may be toward a higher spatial dimension rather than toward the time dimension. As can be seen, such orthorotation is intimately connected with the transformation of an electron or electron/positron pair into a photon, and vice versa. Orthorotations are extremely important when one models the physics of mind and matter interaction, and the physics of a living organisms including both its mind and its physical body—as in psychoenergetics.

OSCILLATING QUANTUM WELL

A quantum well which is oscillating in magnitude and sometimes in sign.

OVERPOTENTIAL

Essentially the overpotential (as in an electrode in a solution) is a shift in the Fermi level necessary to allow the electron in the electrode metal to have energies overlapping with vacant acceptor levels in molecules adjacent to the electrode in the solution.

The overpotential effect appears at small "double surfaces" of things in very close contact, particularly in electrochemistry and in solid state physics. The overpotential enables electron transfer, e.g., by tunneling. The overpotential increases with the log of the reaction rate that occurs at the overpotential location.

For a small reaction rate, the overpotential is small; but for a small increase in the overpotential there occurs a *dramatic* increase of reaction rate. Fogal has utilized the overpotential effect in his semiconductor, to great advantage in accomplishing unusual charge blocking and other special effects.

One of the world experts in the theory and use of the overpotential is Dr. J. O'M. Bockris. E.g., for a short overview pointing out the importance of the overpotential, see Bockris, "Overpotential: a lacuna in scientific knowledge," **Journal of Chemical Education**, 48(6), June 1971, p. 352-358 (a paper which Bockris graciously sent me some years ago). Quoting from that reference, p. 355: "The formal definition of overpotential is 'the change of potential of the electron-conducting phase when reaction rate across its interface with the ion-conducting phase with which it is in contact is changed from zero to a certain velocity.'" And again, p. 355: "...overpotential... in fact determines what fraction of the energy of the self-functioning cell has to be wasted in making the cell work and deliver energy at a certain power." For a more detailed and wider exposition of the overpotential, its effects, and its uses, see J. O'M. Bockris and A. K. N. Reddy, **Modern Electrochemistry**, Vol. 1 & 2, Plenum Press, 1970.

OVERUNITY COEFFICIENT OF PERFORMANCE

More energy out or work out of a system than the operator or experimenter must himself or herself furnish and input.

An open thermodynamic system far from thermodynamic system is permitted to achieve such an overunity performance, so any overunity electrodynamic system must be such. Well-known nonelectromagnetic examples are solar cells, heat pumps, windmills, sails on sailboats, waterwheels for powering watermills, a gliding bird in thermal updrafts, etc. A well-known electromagnetic example is the common solar cell. The common household heat pump has a theoretical COP limit of 8.22, and well-designed units usually achieve some 4.0 or so, so long as the ambient environment does not change too drastically from the design limits. As the environment changes and cools down appreciably, however, most home heat pumps lose efficiency and have to switch onto heating strips, which have a COP ≈ 1.0.

Contrary to prevailing belief, it is certainly possible to have EM systems which are also open thermodynamic systems, far from equilibrium, and freely receiving excess energy (asymmetrically regauging a potential) from an external source, to wit, their violent virtual flux exchange between the vacuum and the system. All EM system are already *open systems* in the virtual particle flux exchange with vacuum; that is already well-known in particle physics, even though the 130-year-old classical electrodynamics does not even include the vacuum-mass exchange. The only question, then, is how to establish an asymmetry in that vacuum-mass flux exchange. Here again, every electrical charge and every dipole is already just such a broken symmetry in the vacuum flux, as has been known in particle physics for four decades! It takes on the average about 50 years, however, for vital discoveries in one discipline to "cross fertilize" into another discipline. At any rate, the symmetry of every electric power system, in the vacuum flux, is already broken in its source dipole. This rigorously means that, if we simply avoid doing something we are doing in our power systems to restore symmetry as a matter of course, it is possible to utilize the free broken symmetry of the source dipole to provide an overunity EM system.

In 1996 I showed that such overunity EM systems are permitted and prescribed by Maxwell's theory, prior to the arbitrary imposition of *symmetrical* regauging of the Maxwell equations (i.e., the Lorentz condition). Violation of the Lorentz condition (by asymmetrical self-regauging) is a necessary but not sufficient condition for an overunity EM device. An asymmetric regauging will certainly give us excess free energy in the system, and an excess force to use that energy in translating against a load impedance to do work, if we design the system properly to do just that.

To achieve overunity, in a circuit one must prevent the (presently universal) re-imposition of the Lorentz condition by the use of one-half the excess received energy to drive the return electrons from the load back through the back emf of the source, killing the gate and restoring symmetry, the Lorentz condition, and equilibrium between vacuum flux and system. All our present EM systems are specifically designed to drive those electrons back through the back emf and thus restore symmetry, *deliberately* asymmetrically self-regauging a second time and *deliberately* restoring the Lorentz condition, along with *deliberately* killing the source dipole's asymmetry in the vacuum flux exchange.

PARADOX (LOGICAL)

A thing that is true, e.g., as shown by experiment, but which violates one or more of the three Aristotelian laws of logic.

Any logic algebra or system has a topology. A paradox in one topology can often be resolved by increasing the topology of the logic. All logical paradoxes in the 3-law Aristotelian logic system are simply fourth-law logic cases in a logical superset, and—at least hypothetically—the fourth law can be applied to resolve them. For a discussion and proof of the 4th law correction to Aristotelian logic, see Appendix III: "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," in T.E. Bearden, **AIDS: Biological Warfare**, Tesla Book Co., Chula Vista, California, 1988, p. 428-443. Logic has many aspects that often may shock our casual expectations: For example, one of the most advanced axioms in mathematics is called the Axiom of Choice. It is used in the proof of many advanced theorems. Using the Axiom of Choice, it can be shown that it is possible to cut a ball into a finite number of pieces, then reassemble the pieces into two balls of the same size, with no air gaps between pieces! While this seemingly defies all common physical sense (and I've not found anyone who has done it!), it is mathematically correct. For a proof, see Thomas J. Jech, **The Axiom of Choice**, American Elsevier, New York, 1973, p. 3-6. For a good perspective on Aristotelian logic, see Morris Kline, **Mathematics: The Loss of Certainty**, Oxford University Press, New York, 1980.

PARITY

Deals with studying a system of sequence of events as if reflected in a mirror. A system has parity is it undergoes no fundamental operational change if replaced with its mirror-reflected twin.

PARTIALLY TIME-REVERSED REGION

A region of space in which an object experiences not only the ordinary forward flow of time, but an extra induced backwards flow of time, usually a fraction of the time-forward time stream.

One result may be to "reduce" time-forward force fields, by a fraction. This is important in bridges, where one must pass EM energy flow onto a follow-on load circuit, but the return current in that circuit must be allowed to circulate back through the load bridge itself, against the inducing back emf. By partially time-reversing the region occupied by the back emf, that back emf is seen as a smaller value by the return current being passed back up through the back emf. The result is that current will circulate in the load loop, powered entirely by the bridging of Poynting flow **S** = **E**×**H** onto the load circuit.

Another result of a partially time-reversed region is that the law of attraction and repulsion of charges is partially reversed. If the time-reversal predominates in the region, as in a momentary time-reversal zone in some electrolytes, then the law of attraction and repulsion of charges is reversed. In that case like charges attract and unlike charges repel. Two positive H+ ions, e.g., may be drawn so tightly together that each enters the edge of the strong force region of the other, forming a quasi-nucleus. Note that in each H+ ion the gluon forces are almost dissolved and the quarks are nearly freed. Then as the time-reversal zone decays back to a normal time-forward zone, the decay changes begin from inside out, resulting in the flip of a quark to change one of the H+ ions to a neutron n. The n and the H+ ion draw together as the preferred decay state, resulting in the formation of deuterium or a D+ ion. This and similar "time reversal zone" reactions have been proposed by this researcher as the mechanisms generating the anomalous deuterium, tritium, and alpha particles (helium nuclei) in the "cold fusion" reactions. It is indeed a special type of low spatial energy fusion, but a very high energy reaction since time-energy is involved as the active ingredient, and time-energy has the same energy density as mass.

PARTICLE PHYSICS

The branch of physics using accelerators to study high-energy particle collisions, to determine properties of atomic nuclei and of the elementary particles themselves.

So far as I am aware, conventional particle physicists do not consider time-energy interaction and time-reversal zone novel nuclear interactions at all.

PATTERSON EFFECT

The intense multiscattering, multipass, multicollection of Poynting energy density flow in an assemblage of palladium-clad spherical bead capacitors, as the palladium adsorbs hydrogen ions from the fluid in which the beads are immersed, and charges up as a capacitor, thereby becoming an energy-flow generator, and nonlinearly increasing the energy-flow as the charge of the capacitors increases.

The Patterson effect yields the Patterson Power Cell® , a validated overunity power system which has shown COP>1200 in independent university testing. The Patterson effect is a variation and adaptation of the well-known anti-Stokes emission effect, which for three decades has been validated in physics journals to produce overunity emission in certain intensely scattering active media.

PATTERSON POWER CELL®

Patented overunity power unit by James Patterson, which has been independently validated to yield COP>1.0.

See James Patterson, "System for Electrolysis of Liquid Electrolyte," U.S. Patent No. 5,372,688, issued Dec. 13, 1994. Filed Dec. 2, 1993. 26 claims, 7 drawing sheets.

PAUTRIZEL, RAYMOND

Eminent French parasitologist who worked with Antoine Priore, and who performed many definitive experiments demonstrating the Priore healing effect.

PEGRAM'S EXPERIMENT

Experiment which demonstrates a relation between electric and magnetic fields in the vacuum which the Maxwell-Heaviside theory does not prescribe.

Simultaneously, a coaxial cylindrical condensor and solenoid are rotated. The rotation produces a magnetic field in the solenoid in the axial direction and between the plates of the condensor. The condensor can be charged by shortening the plates of the condensor. This indicates a cross relation in the vacuum between electric and magnetic fields which does not exist in the Maxwell-Heaviside theory.

Reference: G. B. Pegram, **Physical Review**, Vol. 10, 1917, p. 591.

PERPETUAL MOTION MACHINE

Assertion that no machine can indefinitely operate, since no closed system can input its output power back to itself while powering a load because powering itself and its load violates the equilibrium thermodynamics (i.e., the thermodynamics of a closed system).

The "defenders of the underunity faith" use the statement that "there cannot be any such thing as a perpetual motion machine" as an article of faith and iron dogma. Usually they classify a so-called "perpetuum mobile" as to which law of *equilibrium thermodynamics* it is thought to violate.

For permissible Maxwellian overunity systems, the entire "perpetual motion" conundrum is nonsensical, since we are always speaking of an EM system *far from thermodynamic equilibrium* with its active vacuum environment. The *equilibrium* thermodynamics assumed by the critics does not even apply to such systems, as witness the Nobel Prize awarded to Prigogine in 1977 and the voluminous literature on such open dissipative systems.

As an example, see Stanley W. Angrist, "Perpetual Motion Machines," **Scientific American**, vol. 218, Jan. 1968, p. 114-122. Angrist gives a typical critique using classical equilibrium thermodynamics. He calls *permissible* overunity systems (open dissipative systems receiving excess energy from an external source) *false perpetual motion machines* or *fictitious perpetual motion machines*. Also see Angrist's article, "Perpetual Motion," **Encyclopaedias Britannica, Bicentennial Edition, Macropaedia** Vol. 14, 1976, p. 102-105. Here he gives a very good summary of the subject as ordinarily posed, with precise definitions included for perpetual motion machines of the first, second, and third kinds and for fictitious perpetual motion machines. Unfortunately his entire argument is completely irrelevant to any study of permissible overunity machines, which he already has labeled as "fictitious perpetual motion machines" and has already admitted that they exist! The substance of his entire argument is that systems in equilibrium cannot power themselves. That is of course quite true, but is a mere tautology and has no relation to an open system in disequilibrium, which *can* power itself and its load simultaneously, by the laws of physics and the laws of *disequilibrium* thermodynamics. So in fact Angrist, as do the skeptics who continue to assail legitimate overunity EM system researchers as perpetual motion nuts, establishes his own dummy strawman and demolishes it, while ignoring—so to speak—the huge Statue of Liberty before him.

Now this is not to pick on Angrist! He appears to be a reasonable scientist, carefully stating his thesis in a straightforward manner. His voice is quoted as a *reasonable* critical voice, and not like the usual *extremist* critic who engages in ad hominem attacks from the outset. It merely shows that (1) the entire "perpetuum mobile" prohibition does not even apply to overunity EM systems far from thermodynamic equilibrium with the active vacuum, (2) the critics are either unwittingly or deliberately using the wrong type of thermodynamics as the basis of their entire argument, (3) the *reasonable* perpetual motion critics—wittingly or unwittingly—simply set up an irrelevant strawman and precede to knock it down, and (4) in fact the critics already admit to a machine that will in fact run itself and power its load indefinitely or perpetually unless something changes or breaks or fails, but then oddly label it as a "fictitious" or "false" perpetual motion machine! Meanwhile, in the universe around us we universally observe things in violent and perpetual motion, having continued so for some 15 billion years. None of the atoms seem to be running down, nor do any of the molecules. And they are open systems brought out of equilibrium by absorption of energy, and decaying back into equilibrium (or approaching it) by emission of excess energy.

The eerie point is that technical publications such as **Scientific American** and **Encyclopaedia Britannica** do not seem to question the line of reasoning advanced by those knocking perpetual motion, or even fully analyzed their argument. What we really deal with in terms of skeptics in this area is either (1) a reasonable, scientific skeptic such as Angrist (his listing of machines and attempts for machines not in disequilibrium with their active environment to run themselves is very good), or (2) an unreasonable person spouting dogma and venom, who should simply go read some disequilibrium system thermodynamics and learn etiquette.

Also, we strongly object to supercilious phrases invariably employed by harsh skeptics who insist we are looking for "something for nothing" or "a free lunch". No legitimate overunity energy researcher has ever considered such a thing. We freely admit that the excess energy is received from the active vacuum exchange, via a broken symmetry in that exchange. Should the farmer refuse to use a windmill or waterwheel because he's getting a "free lunch" for energy input? For those who so quickly hurl such condemnatory phrases, let me hold up a mirror and show them an analog of their position: The famed surgeon Alfred Velpeau wrote in 1839: *"The abolishment of pain in surgery is the chimera. It is absurd to go on seeking it today. 'Knife' and 'pain' are two words in surgery that must forever be associated in the consciousness of the patient. To this compulsory combination we shall have to adjust ourselves."* [Martin Gumpert, **Trail-Blazers of Science**, Funk and Wagnalls, New York, 1936, p. 232.]

The strident, dogmatic critics of open overunity disequilibrium EM systems are stating an exact analogy to Velpeau's statement. The *scientific* worth of their criticism is precisely the scientific worth that eventually was shown for Velpeau's pronouncement.

PHASE CONJUGATE

Of an entity: its time reversal. For waves, often referred to as *wavefront reversal*.

PHASE CONJUGATE MIRROR

A nonlinear material which, when provided with an input wave, emits a phase conjugate replica (time reversed replica) of that input wave.

The input wave is by practice called the *signal wave*. The time-reversed wave output is by practice called the *phase conjugate replica*.

PHASE CONJUGATE OPTICS

That portion of nonlinear optics dealing with multiwave mixing, formation of time-reversed waves, phase conjugate reflection (PCR), phase conjugate mirrors (PCMs), pumping PCMs to amplify their emitted PCRs, nonlinear interferometry, distortion correction, self-targeting, laser tracking, adaptive optics, and so on.

PHASE CONJUGATE REPLICA (PCR)

Of a wave: the time-reversed duplicate of the wave.

PHASE-CONJUGATING ACTION

In nonlinear optics, the process or action of producing a time reversed replica of an input wave.

For a particle, producing its antiparticle. For a cell, dedifferentiating it back to some previous form, genetics and all.

PHASE CONJUGATION

In nonlinear optics, the novel nonlinear mixing of waves which generate an output wave—called the phase conjugate replica or time-reversed replica—that precisely retraces the path previously taken by the input wave that stimulates the action.

The process or action of producing a time-reversed wave. Phase conjugation can compensate for aberrations and distortions in the input. Real time holography can also be produced by applying the phase conjugation process.

PHASE-LOCKED HARMONIC SERIES

In Whittaker decomposition of the scalar potential: The decomposition produces a set of bidirectional wavepairs, as explained previously.

Each of these wavepairs has a given frequency, starting from a fundamental. In other words, the other wavepairs are harmonics of the fundamental wavepair. Further, these are regularly arranged in a lattice, so that "timing" or "phasing" is rigidly locked in place.

PHENOMENOLOGY

In physics, the philosophical and experimental study of physical phenomena,

... as distinguished from ontology, (which is the study of being).

PHOTON

In physics, the basic action quantum.

Considered to be the basic quantum of the EM field. Well, there's a little problem with that. The photon contains energy and time, and so must also occupy a time interval. The notion of forcefield rigorously contains mass as a component; yet mass is missing from the concept of field as it is applied. The EM field is conceived as a spatial entity only. It may spatially vary as a function of time, of course, but it itself does not occupy time, at least conceptually. These considerations raise some startling difficulties in the ready conclusion that the photon is the fundamental quantum of EM field. If a "piece of the photon" is missing from the field, how can the field be made of photons? How can one get area out of a straight line? I simply pose this problem for the theorists and the foundations physicists, pointing out that it has universally been fashionable for physicists to ignore that *the photon carries not only a little increment of energy, but also a little increment of time*. My own work has been to look at the dimensional difference between masstime and mass; i.e., between a mass that has just absorbed a photon, and thus must consist of masstime (4-space), and just after that masstime then re-emits the photon, reverting back to mass (3-space). From this one can build, e.g., a mechanism that generates the actual flow of time itself. It is in theory an engineerable mechanism.

PHYSICAL REALITY

To be added.

PING-PONG

To be added.

PINNA

The outer ear(s). Of particular interest to us are the tiny little folds in the ear.

PINNA INFORMATION

Refers to hidden information content of the field, or in the received signal.

Comes from the little folds in the pinna (outer ear). For sound direction and distance sensing, the tiny pinna folds of the outer ear use phase reflection information more than 40 dB below the primary sound signal that strikes the eardrum. In EM signals, we are referring to "hidden" or infolded information deep within a signal from a distant target. In fact everything in the target contributes to the signal emission from it, and all that information is riding "within" the signal, in its deep "pinna" information. The Fogal chip can outfold some of this pinna information, so that information can be recovered from a signal which no other process heretofore could recover. For a dramatic future use, as the Fogal semiconductors and equipment using them are developed, there will emerge detecting units that can see highly detailed structure in the human body, merely from processing the light reflected from the exterior of the body. The details eventually will rival and even surpass that of magnetic resonance imaging, and will be essentially noninvasive.

It is also true that hidden "vacuum engines" can be transmitted back into the body, so that "problems" or disorders can be directly corrected. As an introduction to this type of area, where cellular disease "antiengines" can be created in the body to directly reverse cells and their genetics, curing the disease, see T.E. Bearden, "Vacuum Engines and Priore's Methodology: The True Science of Energy-Medicine, Parts I and II." **Explore**!, 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62. A major book on this subject will soon be published. For a presentation for the nonspecialist, see T. E. Bearden, "Hidden Mechanism for Reversal of Diseased Cells to a Previous Healthy State," **Resonance**, Newsletter of the Bioelectromagnetics Special Interest Group, Mensa, No. 29, Nov. 1995, p. 11-15.

PINNA INFORMATION CONTENT

In electromagnetics, the hidden inner Stoney-Whittaker-Ziolkowski infolded EM information inside the scalar potential.

Since an EM wave can be represented as two scalar potentials, then by infolding pinna information inside one or both of the two potentials, it is infolded inside the carrier wave. Thus hidden information can be and is carried by all EM radiation and all EM signals. The Fogal semiconductor for the first time enables the outfolding and processing of some of this information.

PINNA INFORMATION DETECTORS

Detectors based on use of the Fogal semiconductor, which can outfold, process, and display some of the hidden information inside the scalar potential and inside signal emissions from distant objects.

PINNED ELECTRONS

Electrons held stationary by a magnetic field, or other set of forces that react to "push them back into their positions" whenever they try to move away.

PINNING

The act of restricting (by restoring forces and fields) movement (particularly of electrons) due to a disturbing magnetic field or other external force.

POLYHEDRON

A solid figure, especially with more than six plane surfaces.

POSITRON

An antielectron (time-reversed electron), possessing a positive charge. In Dirac's relativistic theory of the electron, a positron may be regarded as an empty negative energy state in the Dirac sea whose usual occupying electron has been removed.

POWER

The time rate at which work is done (at which the form of energy is changed).

In electrical engineering, the term "power" is erroneously also used to mean "energy".

PRIGOGINE, ILYA

Noted Belgian chemist and Nobelist, and one of the pioneers of the thermodynamics of open systems far from thermodynamic equilibrium.

Prigogine was awarded the 1977 Nobel Prize for his work on the theory of dissipative structures in nonequilibrium thermodynamics.

PRIORÉ, ANTOINE

Italian radar technician, remaining in France after WW II, who discovered the use of novel waves from plasmas emitted when normal EM waves were introduced, and who used this effect to accomplish remarking healing in laboratory animals of terminal tumors, infectious diseases such as trypanosomas, restore suppressed immune systems, and heal atherosclerosis.

Prioré's beam was actually longitudinal EM waves. The body when irradiated by longitudinal EM waves adds a weaker phase conjugate replica, thus converting LW nonlinear optical pumping from the 3-space domain to the time domain. This time-reverses the irradiated cells and all their components back to a previous physical state and condition.

Such "time-domain" pumping is used by the living body's cellular regeneration system to restore cellular damage and heal the cells, within its limits. Thus Prioré unwittingly uncovered a way to dramatically amplify the body's own healing and regenerative powers.

POYNTING, JOHN HENRY

An English physicist who lived from 1852-1914, whose contributions to electromagnetics included the energy flow theory now known as Poynting flow.

He contributed the Poynting theorem, which states that the rate of EM energy loss in a specific region of space is equal to the sum of the dissipation rate (heat loss) flowing across the boundary of the region. Concurrently with Oliver Heaviside, Poynting conceived the notion that energy flowed through space, which previously had not appeared in physics. Heaviside actually published first, but obscurely; Poynting published in a prestigious journal, so the theory came to be known as "Poynting theory." Poynting himself always gave credit to Heaviside for being first. Also, the Heaviside version $\mathbf{S} = \mathbf{E}\times\mathbf{H} + \mathbf{G}$ was more extensive than Poynting's $\mathbf{S} = \mathbf{E}\times\mathbf{H}$. Heaviside was a recluse; years after his death, in his little garret apartment some papers were discovered beneath the floorboard. Therein Heaviside has used his **G** term (a closed energy flow, or "trapped energy" term) to represent gravitation, and had worked out a testable theory of electrogravitation. Modern general relativity indeed considers that it is energy—particularly trapped energy—that is responsible for generating gravitational field and gravitational potential. Apparently nothing at all was done or has been done with Heaviside's electrogravitational theory. See H. J. Josephs, "The Heaviside papers found at Paignton in 1957," **The Institution of Electrical Engineers Monograph No. 319**, Jan. 1959, p. 70-6.

POYNTING ENERGY

Loose term (not rigorous!) used to describe the flow through space of Poynting energy density $\mathbf{S} = \mathbf{E}\times\mathbf{H}$. It actually describes not the energy flow density per se, but that component of it that would be intercepted by a unit point charge.

POYNTING ENERGY DENSITY FLOW

A more rigorous term used to describe the flow through space of the Poynting energy density flow component **S** = **E**×**H**. that would be intercepted by a unit point charge.

POYNTING ENERGY FLOW

Loose term (not completely rigorous) used to describe the flow of energy density **S** = **E**×**H**.

Rigorously, note that both **E** and **H** are "defined" only after their reaction with a unit point static charge. Hence **S** is the resulting component of the overall energy flow that interacts with the charges (with the circuit) and is diverged into the circuit to power it. It is *not* the entire energy flow, which also includes a vast Heaviside nondiverged component that does not strike the charge and interact with it, and hence does not interact with the circuit but is just wasted.

POYNTING GENERATOR

Any dipole or potential difference, which represents a broken symmetry in the vacuum's virtual particle flux, and hence extracts virtual energy from that asymmetry, and gates it out as a gushing energy flow in 3-space. The component of that energy flow that would be intercepted and diverged by a unit point charge is the Poynting energy flow **S** = **E**×**H**.

POYNTING VECTOR

The vector **S**, given as **S** = **E**×**H**, which is used to represent the Poynting (intercepted and collected by a unit point coulomb) energy density flow component.

In Heaviside's version, **S** = **E**×**H** + **G**, where **G** is an energy circulation flow term of trapped energy flow.

PPCM

Pumped phase conjugate mirror.

A phase conjugate mirror material which has introduced to it appositive or "pump" waves. In the phase conjugating process, up to all the energy in the pump waves can be output as the energy in the emitted phase conjugate replica wave. Thus the PPCM is an amplifier, very similar to a triode.

PRIMARY IMAGE

The image or portion of an overall image that is of most interest, and usually the sharpest in focus.

PROBABILISTICALLY

Of or determined by the laws of probability.

Note particularly the somewhat dichotomous use of "determined by—probability." The entire notion in probability is that the overall result is not individually determined, but that there are multiple outcomes of the experiment, each of which occurs with a certain frequency out of a very large number of trials. If a single result always occurs, that is said to be a deterministic change, not a probabilistic change. Unfortunately if one pursues the definition of "probability" in mathematics and logic, one eventually finds that it is dichotomous, as was our definition of energy. Ultimately the foundations guys just threw up their hands and said, "Oh, what the heck! Probability is probability, every fool knows that!" The problem cannot be resolved in 3-la w Aristotelian logic; it requires a higher topology logic.

PROBABILITY

The probability of a simple event is the ratio of the number of times it occurs to the total number of trials (for a large, essentially infinite number of trials).

Note that "trials" constitute things that have occurred (are past). Probability has resisted rigorous logical definition by Aristotelian logic; the reason is that it is an expression of the fourth law, the identity of opposites. Probability is the expression of the future (that which has not occurred) in terms of the past (that which has occurred). Since we have been trained to conceive or think of something as if it had just been perceived to occur, then this is the only way we can *conceive* the future—as if it were an "already occurred past" to us.

But with Aristotelian logic, the future cannot be composed of a set of pasts (the law of the excluded middle prevents it!), and probability has no acceptable logical basis. With four-law logic, the future can be modeled in terms of—and even be identical to—the past, and probability has a logical basis.

Consider *throwing a die to land with one face up*. Thinking of the event (as if it has just occurred, in terms of the most recent past), one can conceive six ways of looking at it. In other words, one can conceive of six "most recent pasts." By the fourth law, the total "most recent past" is identical to the "most immediate future," on the common boundary. If one therefore *collects* all six "most recent pasts," they turn into the most immediate future, by the fourth law of logic. One may argue that only a single one of the events will actually occur. Here one counters with the observation that, when it occurs, it is the past, not the future, and then it has not been collected so as to move to the boundary. One also points out that, with the hidden variable approach, in theory it is possible to directly engineer those probabilities while the Schrödinger wavefunction is still propagating and has not yet collapse to constitute an "observation." I.e., it is possible to engineer Wheeler's "pre-reality," and determine in advance just which of the outcomes shall occur.

Now, if this type of equipment eventually gets built and is small and portable, just think what one could do in Las Vegas!

PRODUCT SET OF INTERNAL WAVES

Modulations of the internal bidirectional wavepairs comprising a scalar potential.

This set was added in the mid-1980s by Ziolkowski, to Whittaker's 1903 sum set. When a signal is "infolded" inside the DC potential, it is modulated upon (multiples) one or more of the internal biwave pairs. Ziolkowski's work is necessary to deal with that; Stoney and Whittaker alone do not handle it.

PSYCHOENERGETICS

Russian branch of energetics where extended electrodynamics is targeted upon the mind and its functions, hence in the time-domain and employing time-polarized EM waves.

PSYCHOTRONICS

Czechoslovakian name for extended electrodynamics effects on the living body.

PUMPED PHASE CONJUGATE MIRROR (PPCM)

A nonlinear material which:

(i) when provided with an input wave, emits an amplified phase conjugate replica (time reversed replica) of that wave, and

(ii) has additional opposing waves as inputs, so that their energy is transformed into the amplified energy of the time-reversed replica wave.

The word "mirror" is misleading, in that the "mirror" may be just a highly nonlinear material, or be an actual smooth metallic surface containing "hot spots" by heating, etc. In any case, it is the nonlinearities such as the hot spots that produce the phase conjugating action.

PUTHOFF, HAROLD

Noted physicist active in a variety of fields and investigations including in psychoenergetics, extended electrodynamics, energy from the vacuum, parapsychology, remote viewing, psychokinesis, zero-point energy of the vacuum, and cosmology.

We have specifically mentioned his important cosmological feedback principle.

QM VACUUM

The quantum mechanical vacuum, which models the vacuum as an intense flux of virtual particles, each appearing an disappearing in such extreme time intervals as to be unobservable individually.

Violent fluctuations of extremely short duration constantly occur in the QM vacuum flux. Various calculations estimate the energy density (in mass units) as up to 10^{80}-10^{90} grams per cubic centimeter, or even greater.

QUALITATIVELY

Determination based on qualitative analysis, usually as contrasted to quantitative determinations.

QUANTIZED

i) Reduction of some observable quantity to multiples of some small, indivisible unit, or,

ii) expressed in terms of quantum theory.

QUANTUM

A single "particle" of the quantity action.

The quantum is always composed of two fundamental quantities (canonical variables), and it is the "fissioning" of the action quantum into two pieces which results in detection of one of the pieces. An indefinite "smear" of part of the two pieces may be simultaneously detected by a linear detector, but never both exactly (Heisenberg uncertainty principle). More recent work has shown that, in a highly nonlinear situation, both canonical variables can be determined simultaneously to any desired degree of accuracy, in which case the Heisenberg uncertainty principle does not hold in its conventional form. This fundamental change in what has been regarded as one of the "sacred laws" of physics has not yet made it into most physics texts.

QUANTUM ELECTRODYNAMICS

The theory of photons and electrically charged particles and their interactions.

The use of the term "quantum" implies that the EM radiation's discrete photon nature is significant, so that quantum theory must be employed.

QUANTUM FIELD THEORY

A quantum mechanical theory in which "... a physical field is considered as a collection of particles and forces, and observable properties of an interacting system are expressed as finite quantities rather than state vectors."

[**Dictionary of Science and Technology**, ibid.] More simply (after Davies, **The New Physics**, 1989): "The theory that describes the quantum effects of a classical system of fields defined on space-time and satisfying various partial differential equations."

QUANTUM POTENTIAL

A special potential added to the Schrödinger equation, by Bohm in his hidden variable theory of quantum mechanics.

The quantum potential can move instantaneously. In my book **Gravitobiology**, I published a mechanism by means of which a quantum potential can be created. (There may of course be *other* QP-creating mechanisms). An iterative, mutually phase conjugating of EM signals passed back and forth between two objects, or among many of them, is called *self-targeting*. Extreme self-targeting can produce such a QP or partially produce one, particularly if a dense "pumping" (optical sense) of the phase conjugate mirror objects occurs. I am presently proposing the QP as a means of direct energy amplification. I.e., if there is "instantaneous travel" between two light-observed points, then to the instantaneous agent there is no length separation or time separation between the points. This of course means that in effect the objects now coexist "superposed" on each other in a multiply connected space, at least with some fraction of "coupling".

Suppose the coupling fraction is 0.5, and 100 widely separate d nodes are so connected by a common QP. Then if one puts in 1 joule per second of energy flow at a given node, there will instantly and simultaneously appear at each of the other 99 nodes one-half joule of energy per second flow. So I can gain quite a bit o f *direct energy amplification* in this distributed system, and nothing can "shield" the distant nodes. Presently, other than possibly Gabriel Kron's "open path", I have not found any other consideration of "direct amplification of energy at a distance".

The implications of the quantum potential—in everything from power sources to medicine to weaponry—are staggering. Weapons based on just this very QP, multiply connected space, and direct energy amplification appear to have already been developed and are on site in three nations. A fourth either is just getting them or is nearing their deployment.

QUATERNION

Expression comprised of the sum of four terms, one of which is real and three of which contain imaginary units, and that can be written as the sum of a scalar and a three-dimensional vector.

QUATERNION ALGEBRA

The algebra of quaternions and their mathematical operations.

Quaternion algebra is of higher topology than either vector algebra or tensor algebra. Maxwell's original equations are some 20 equations in 20 unknowns, in quaternion algebra. Heaviside and others reduced the algebra to vector algebra and some four equations. The present vector equations taught in university as "Maxwell's equations" are in fact Heaviside's truncation of the Maxwell theory. Going to tensor algebra electrodynamics cannot recover the full range of Maxwell's quaternion theory.

As an example, what Tesla actually did in his patented circuits cannot even be seen in vector or tensor analysis. On the other hand, analysis of his circuits by quaternion electrodynamics clearly reveals his startling ability to shuttle the energy in the circuit at will. See T. W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," **Annales de la Fondation Louis de Broglie**, 16(1), 1991, p. 23-41.

QUINTESSENCE

Literally, the "fifth essence". In short, dark energy.

RADIONICS

To be added.

REACTIVE POWER

Electrical power, measured in volts-amps-reactive (vars) that cannot do work as is.

For sinusoidal waveforms, the formula for reactive power is given by $I \cdot V \cdot \sin\theta$, where θ is the phase angle between the voltage and the current.

REAL, HIDDEN DYNAMICS

The dynamics of vacuum engines, or of "internal energy" or "internal work" ongoing in a system or volume of space.

REAL HIDDEN VECTORS

The "hidden" or "infolded" real vector components,

e.g., of a vector zero summation system.

A group of nonzero, finite vectors which as a system sum to a net translation zero; i.e., which do not translate a resisting mass or charge, but do "stress" it. Such a system therefore constitutes a stress potential.

REAL POWER (ACTIVE POWER)

Power, measured in watts, that can do work, including generating heat.

For sinusoidal waves, real power is given by the formula $I \cdot V \cdot \cos\theta$, where θ is the phase angle between the voltage and the current.

REAL VECTORS

Vectors in three space, without imaginary components.

REGAUGE, REGAUGING

Changing the scalar potential or the vector potential, or both.

If one or both of the potentials is/are changed so that a net force is created in the system, that is an asymmetrical regauging.

If only one potential is changed, a net force results and that is an asymmetrical regauging.

If both potentials are changed but carefully selected so that the free force resulting from one change is equal and opposite to the free force resulting from the other, that is a net symmetrical regauging, often called a "Lorentz regauging".

RECTANGULAR TRANSFORMATION MATRIX

Standard mathematical physics.

REFLECTION

The rebound of light or other energy from a surface.

A smooth surface (e.g., a mirror) may produce a specular reflection. A rough surface (e.g., a sheet of paper) produces a diffuse reflection. Actually, in reflection of light, the same light does not rebound; instead, the light incident upon the mirror or surface is absorbed, and new light is re-emitted by that surface. Further, considering reflection from a dielectric, every interior part of the dielectric participates in the reflection. All the internal information about the entire object—everything!—is there in that reflected light, infolded as the hidden "information content of the field" (Stoney-Whittaker-Ziolkowski structuring inside the two scalar potentials comprising the reflected wave). With the Fogal semiconductor, at least some of that hidden "pinna" information in a sample detection of the reflected light can be outfolded and displayed.

REFRACTION

The bending of light as it passes from one medium to another.

It is the phenomenon that enables a lens to focus light and that makes a stick appear to bend where it dips into water.

REPULSION BEAM

A beam of energy directed at a target, which when absorbed by the target will generate a force in and on that target to push it away from the beam projector.

REPULSION-FORCE-GENERATING BEAMS

Beams of energy which "push away" their targeted receivers, by inducing a force in the receiver directed away from the source of the beam.

RETARDED WAVE

The electromagnetic wave that is derived from the retarded fields, which in turn are derived from the retarded potentials.

This means that the retarded wave at a point or region of space has arrived there or exists there as the result of the dynamics of charges and currents that existed at earlier times.

RETROREFLECTING

Precisely reversing the path of.

An example is a phase conjugate reflection.

SACHS, MENDEL

Noted scientist who has completed the work of Einstein and developed a unified field theory covering the universe from the tiniest part to the entire conglomerate.

SACHS' UNIFIED FIELD THEORY

An extension and completion of Einstein's work that includes the universe and its dynamics from the tiniest part to the entire conglomerate, and includes quantum mechanics, electromagnetics, gravitation, and the strong and weak forces.

SCALAR

Characterized by magnitude only.

SCALAR (QUANTITY)

In ordinary vector analysis, a quantity completely characterized by magnitude only.

Actually, it contains magnitude and time, because to exist it must steadily exist in time. So we take the smallest increment of observer time we wish to consider, and that is the minimum "length" of the time-tail that the scalar must occupy and have, i n order to even be a "perceived or detected" thing. Before the mathematician jumps in with a nice limiting process to take away the time-tail, if you let that little time interval get too close to zero, then you get too close to eliminating the observer himself, and the entire problem vanishes including mathematics and physics and the observer as well.

One simply cannot demonstrate something that does not exist in time—on paper, in one's mind, or otherwise! Be careful of the term "spacelike " in physics! It hides a logical nightmare. This was the very problem I uncovered in Aristotelian logic; the little increments of time to have a "perceived A" or a "perceived not-A" or to apply an "identity decision criteria" to determine identity or nonidentity, had been left out for lo these 2,000 or so years.

SCALAR ELECTROMAGNETICS

That electrodynamics that arises from considering transverse EM waves, longitudinal EM waves, time-polarized EM waves, electrogravitation, superluminal EM signals, interferometry, nonlinear optical functions, time-as-energy, and the infolded electrodynamics inside all usual EM fields, waves, and potentials.

SCALAR POTENTIAL

Usually considered synonymous with the electric potential.

Here we have another "logical nightmare."

Anyway, the conventional "definition" of electric potential assumes one knows what a "potential" is already, and goes something like this: "The potential measured by the energy of a unit positive charge at a point, expressed relative to an equipotential surface, generally the surface of the earth, that has zero potential." [**Dictionary of Science and Technology**, ibid., p. 722.]. Well, that is botched. It tells or attempts to tell what the *measurement* of a potential is, using a unit positive charge and a process. A potential cannot be its own magnitude measurement, just as a man cannot be his own weight measurement.

So what do we do? We have to tell what the potential *is*. This is the problem that so mystified and infuriated Oliver Heaviside, when he altered and diminished Maxwell's equations to rid them of every potential that he could. Heaviside stated that the potential was *"mystical and ought to be murdered from the theory."* He tried to do just that, and conditioned generations of electrodynamicists to believe that the potential had no physical reality at all, but was only a mathematical convenience. Nahin expresses Heaviside's view as follows: "In an 1893 letter to Oliver Lodge, Heaviside said of his own work that it represented the 'real and true "Maxwell" as Maxwell would have done it if he had not been humbugged by his vector and scalar potentials.'" Paul Nahin, **Oliver Heaviside: Sage in Solitude**, IEEE Press, New York, 1988, p. 134, n. 37.

The concept of potential was developed by Poisson, Green, and others from about 1813 to 1827. Three notions involved in the development of the concept of the potential are: (i) there is the collection or storage of energy in the system, via the collect ion or storage of some entity therein, and this "potential energy" can be released and dynamically used, (ii) the potential is a scalar function whose spatial rate of change is a vector force, and (iii) the dynamics of the stored energy is suspended by the "storage" or "collecting" entity; in other words it is a sort of "suspended action" waiting to be unleashed. *It requires some action—such as interception and divergence—to release the leashed energy*.

We have to disagree with the second notion, having already discussed the fact that no vector force exists unless mass is present, since mass is a component of force. Here we have a residue of the old "material ether" concept. For the third notion, we note that the concept of a "virtual" particle or particle flux means that it cannot become observable unless some interaction is imposed. So we may handle the "third notion" requirement by assuming the potential to be composed of hidden virtual particle flux, or hidden wave flux, in wave-to-wave interaction.

Physics basically tries to define the electrical potential as energy per coulomb, or $\phi = W/Q$. That equation, however, is actually not a definition at all, but rather is an algorithm for calculating the magnitude of the excess energy collection/collecting upon a coulomb of charge, when that coulomb is immersed in the potential ϕ. It in fact is the *reaction cross section* of the potential, not the potential entity itself. This now begins to give us a clue. Potential is *energy* in some fashion. So the entity ϕ must have essentially the same definition as energy! Potential energy, collected energy, stored energy, whatever it's energy first and foremost.

Refer to our discussion of the 3-law logical contradiction in the concept of energy. The same logical contradiction must be present in the notion of the potential.

Just as we did for energy, we must define the scalar potential ϕ in two appositive ways: (i) in the wave view and (ii) in the particle view. So we define potential (particle view) as any virtual particle flux. We also define potential (wave view) as any hidden wave flux. Further, we note that the energetic vacuum is composed of a virtual particle flux (particle view) and so it is also a *potential*. Since the virtual particle flux of the vacuum is extremely intense, the vacuum potential is enormous in magnitude.

We shall also consider any other potential to ultimately be a change to the vacuum potential. Originally, one conceived of a potential occupying the "vacuum as an empty space." Then general relativity made the vacuum into spacetime, geometrized it, and made the geometry itself dynamic and operative—i.e., energetic. Okay! Then that makes spacetime itself a potential.

We also firmly dispose of the peculiar and deformed notion that the scalar potential of a point-source charge (at any spatial point that the potential occupies) may be *defined* as the "work performed upon a unit charge by pushing it in against the **E**-field of that source charge, to that spatial point." This is an equation and not a definition at all; instead, it is an algorithm for calculating the energy density magnitude of the scalar potential. It is a fine algorithm, but it has nothing at all to do with what the potential itself *is*. Again, a man cannot be defined as an algorithm for calculating the magnitude of his height or his weight or his energy density!

Without further discussion, we also somewhat extend the general relativity notion of spacetime as dynamic geometry, as follows: In a special sense, spacetime ≡ geometry ≡ potential ≡ "trapped " or collected/collecting energy. So when we engineer the internal structure of the potential, we also engineer the internal structure (subspace) of spacetime and spacetime geometry. Curvature of spacetime, e.g., becomes simply a change in the energy density of the vacuum, from the ambient value, and it can be internally structured so as to contain "engines". When we apply the hidden EM variable theory of the Stoney-Whittaker-Ziolkowski work, particularly when we consider the phase conjugate pairing as it exists in spacetime before interaction with charge, we are engineering and extending (i) electromagnetics, (ii) quantum mechanics, (iii) general relativity, (iv) spacetime, (v) the spacetime curvature concept, and (vi) spacetime geometry.

SCATTERING PROCESSES

Collision processes in which an incident particle such as a photon, electron, or proton, of known energy or direction or both, collides with another entity, and emerges from the collision with a different energy or direction or both.

SCHRÖDINGER, ERWIN

Noted Austrian physicist and Nobelist who discovered the wave equation named after him as the Schrödinger equation, and performed other vital work in wave mechanics.

SCHRÖDINGER EQUATION

In quantum mechanics: an equation that describes the propagation of waves associated with subatomic particles, or more generally, that describes the evolution over time of the quantum state of a system.

SCHUMANN RESONANCES

The extremely low frequency (ELF) resonant frequencies of the Earth-Ionospheric waveguide.

The first five Schumann resonances calculated from a simple model are 10.6 Hz, 18.3 Hz, 25.8 Hz, 33.4 Hz, and 40.9 Hz. However, better models are available and the frequencies vary from these simple model solutions. A better calculation gives the first four as 8, 14, 20, and 26 Hz nominally. The measured Schumann resonance corresponding to the 8 Hz is about 7.75 Hz. This is very near the mechanical resonance frequency of the human body, and also is evidenced in brain waves often "locking" to this frequency in the natural environment.

See W. O. Schumann, **Z. Naturforschung**, Vol. 72, 1952, p. 149, 250. For detailed models and discussion of the observations, see J. Galejs, **J. Res. Nat. Bur. Standards**, Vol. 69D, 1965, p. 1043.

SELF-ENERGY

In a quantum mechanical system, the energy associated with the emission and absorption of virtual particles.

In a classical system, the energy associated with the interaction among parts of the system itself.

SELF-ORGANIZATION

(After Davies, **The New Physics**, 1989): *"Spontaneous emergence of order, arising when certain parameters built in a system reach critical values."*

Simpler put, it is the spontaneous formation of patterns of localized order in a system which begins in a more homogeneous state. In short, the system moves from a simpler to a more complex state.

SELF-POTENTIAL (E.G., OF THE ELECTRON)

The potential created in and of the charge

...(e.g., the electron) by its virtual particle flux exchange with the vacuum.

Also, an asymmetry in the virtual photon flux of the vacuum, caused by the interaction of the charge (e.g., the electron) with the vacuum flux. From here, we note that the self-potential must consist of real observable 3-space EM energy flow or flows from the electron, because of the asymmetry. However, the charge is not the primary source *per se* of the energy flows that establish the self-potential, but instead is the asymmetric gate in the vacuum flux that "gates out" the potential and its energy flux. The asymmetry in the vacuum flux is the actual *source*. Indeed, the energy is received by the charge in the form of converging longitudinal EM energy waves from the complex plane (from the time domain). The spin of the charges of the dipole transduce the absorbed complex EM energy into real observable EM longitudinal wave energy, and pour it out in all directions in 3-space. See my paper, "Giant Negentropy from the Common Dipole," *ibid*. for the mechanism for the charge also.

SELF, SENSE OF

To be added.

SELF-TARGETING

Between two entities, where each acts partially or wholly as a phase conjugate mirror, self-targeting is the iterative phase conjugation or "ping-ponging" and rapid convergence of beams between the two PCMs.

It has a spectacular application in *phase conjugate shooting* from a spaceborne laser against a rising hostile booster, even several thousand miles distant. First a laser pulse is "bounced off" the distant target, and the reflected pulse returns to the receiver in the laser platform. It carries the movement-induced alteration of the distant target on the pulse it reflected. The next emitted pulse is an amplified phase conjugate replica of that received pulse, and it will backtrack to the same spot on the rising booster, leading it a bit and converging moderately so that it meets that spot. That spot begins to heat up, increasing its nonlinearity and making it act even more as a PCM. So its reflected signal off the "hot spot" is also a phase conjugate replica of the pulse that it received from the laser. That reflected pulse now will moderately converge its reflected energy upon the distant laser, increasing the return to the laser. The ping-pong continues, so that the laser beam narrows sharply upon the distant hot spot, and the reflected beam from the hotspot narrows sharply upon the distant laser. This "iterative narrowing ping pong" locks the laser upon the distant hotspot, even though the booster is moving. This process continues for the necessary dwell time, until the beam melts through that spot and explodes the hostile booster. The iterative phase conjugating reflection, or "ping ponging" between the target spot and the laser, is called *self-targeting*.

S-FLOW

The Poynting energy density flow S, where **S ≡ E×H + G**.

E is the electric field, **H** is the magnetic field strength, and **G** is a trapped closed loop flow of energy density.

The **S**-flow is not the total energy flow per se. Since the fields **E** and **H** and the closed loop **G** are only defined in terms of the intercepted, trapped or "collected" energy, then **S**-flow is always *with respect to* some assumed intercepting/collecting entity, usually an assumed unit point charge.

SIGNAL WAVE

In nonlinear optics, the input signal (wave) to a phase conjugate mirror (PCM), which stimulates the PCM to emit a phase conjugate replica (PCR) wave in return.

SLEPIAN CURRENT

The vector jϕ , which in a current loop represents the energy unit area per coulomb of flowing charge, per unit of time, that is collected/collecting from the Poynting **S**-flow and dissipated in that current loop in its load s and losses.

It does not include the enormous amount of Heaviside nondiverged energy density flow that is flowing along the outside of the circuit conductors, and that is not being collected and therefore not being dissipated.

SLEPIAN FLOW

"Flow of collected-energy density dissipation" in a current loop,

...and therefore to the vector jf in the current loop, which represents the energy per unit area per coulomb of flowing charge, per unit of time, that is being dissipated in that current loop in its loads and losses.

It does not represent the enormous amount of energy that is flowing along the circuit, that is not being collected, not diverged, and therefore does not enter the circuit to be dissipated from it.

SLEPIAN VECTOR

Refers to the vector $j\phi$ in a current loop,

...which represents the energy density per coulomb of flowing charge, per unit of time, that is being dissipated in that current loop in its loads and losses.

It does not represent the enormous amount of energy that is flowing along the circuit, but not being collected and therefore not being dissipated. We accent that *only intercepted and collected energy can be dissipated!*

SOLID-STATE PARAMETRIC OSCILLATORS

A parametric oscillator circuit employing solid state devices.

A parametric oscillator is a device that oscillates by rhythmically changing a parameter (such as capacitance or inductance, etc.).

Simple switching between differing values of the parameter is often used to induced the periodic change of the parameter.

SOLITON

A wave pulse which propagates without changing its shape and without dispersion.

SOURCE DIPOLE

Any dipole, considered from the viewpoint that EM energy continuously pours out of it.

There is no true "source" as such in nature, since the energy is not created but only gated. In the case of the dipole and electrical charge as sources of EM energy, fields, and potentials, we solved that vexing problem in our "Giant Negentropy from the Common Dipole", *ibid*.

SPACETIME

The entity represented by the "fused product" of space and time, to give "space × time".

Spacetime is also dynamic geometry, in the general relativistic interpretation. We consider spacetime to be a potential, and also geometry, and also vacuum, and also virtual particle flux, and also hidden wave flux.

SPATIAL CLOUD

A cloud of something (such as electrons) in a spatial region.

SPATIALLY

With reference to space and spatial position.

SPECIAL RELATIVITY

Einstein's theory of inertial frames in which, if two systems are moving uniformly in relation to each other, one cannot determine anything about their motion except that it is relative.

Each of the two frames is said to be "rotated" with respect to the other, but not accelerating. The velocity of light in space (the vacuum) is constant, and is independent of the velocity of its source and the velocity of an observer. All the laws of physics are the same in all inertial frames of reference.

SPIN

The intrinsic angular momentum of a particle, such as an electron, proton, neutron, photon, graviton etc.,

...even when at rest, as if it were a top spinning about an axis, but had to spin 720 degrees before it turned "full circle."

Spin is quantized, and is always described as a half or whole spin, e.g., - 1, - 1/2, 0, 1/2, 1, etc. A spinning charged particle such as an electron thus demonstrates a magnetic moment, due to the circulation of charge in the spinning. In the nucleus of an atom, the spin of the nucleus is the resultant of the spins of the nucleons (particles comprising the nucleus).

Spin of particles would appear to be more like an "implosion" to "explosion" circulation; in other words, it circulates in the time domain (complex plane) as well. It would appear that the spin of a particle is the basic feature that integrates the (disintegrated) flux energy of vacuum flux into observable charge. Apparently all observable fields, matter, effects, etc. depend upon this basic mechanism to zip together virtual entities and form observable phenomena.

SPIRAL GALAXY

Those galaxies having arms which extend in the form of spirals.

About 70% of all galaxies are spiral galaxies.

The interesting thing is that only about 10% of the gravity necessary to hold the spiral arms together can be accounted for by known gravitational sources. This has led to the theorizing of dark (hidden) matter of exotic and unknown form, which would be responsible for the excess gravity.

The present author has proposed that the unaccounted enormous Heaviside component of the EM energy flow surrounding every field/charge interaction and not entering into that interaction, accounts for the missing source of the gravity. See "Dark Matter or Dark Energy?", Journal of New Energy, 4(4), Spring 2000, p. 4-11.

S-SOURCE

Source of Poynting energy density flow.

Source of **S** = **E**×**H** flow. The primary source is considered to always be an asymmetry in the virtual particle flux of the vacuum. The **S**-flow rigorously is that component of the overall energy flow that is intercepted by some system of interest—either present or assumed. That component of the overall energy flow that is not intercepted by the system of interest, is referred to by this author as the Heaviside energy flow or the dark energy flow associated with every field/charge interaction.

STONEY To be added

To be added.

STONEY/WHITTAKER BIWAVE DECOMPOSITION OF THE SCALAR POTENTIAL ACROSS THE DIPOLE

(i) a scalar potential (i.e., a difference in potential) exists between the end charges of a dipole,

(ii) that potential (difference) can be mathematically decomposed into a harmonic series of bidirectional EM wave pairs, where each pair consists of an EM wave and its precisely superposed time-reversed twin. Thus a dipole is a "source" or identically is a set of hidden bidirectional waves (wave view).

STRING THEORY

A rather recent mathematical physics theory which replaces the zero-dimensional "point" used in previous physics with a one-dimensional object called a "string".

It has been particularly effective in describing the gluon forces between two colored particles in particle physics, and has since been extended to many other areas. It is sometimes referred to as the "theory of everything".

"STRING" WAVE

The transverse wave on a taut string produced when the string is plucked.

The string wave itself stays on the string and does not travel through the slapped external medium.

However, Faraday believed that EM fields in vacuum consisted of "material lines of force" which behaved like such taut material strings, without any physical endholders. He thought that EM disturbances were simply the transverse disturbances of those "lines of force". He profoundly affected Maxwell's thinking, so that Maxwell accepted these physical lines of force. Consequently Maxwell *assumed* those same "transverse taut string vibration waves" in his EM wave theory.

In eliminating the "stringholders" for the assumed taut strings, Faraday and Maxwell discarded the Newtonian third law half of electrodynamics! They discarded half the energy and half the wave as well. Consequently today physics has no causative mechanism for Newton's third law, because it threw out the EM mechanism for it more than a century ago and never noticed it!

SU(2)×SU(2) SYMMETRY ELECTRODYNAMICS

To be added.

SUBSPACE

In vector mathematics: a subset of a vector space which is closed under vector addition and scalar multiplication operations.

In physics: Physics often uses the term "space" to refer to spacetime. It also uses the term "space" to refer to the dimensions of space, without the dimensions of time. At any rate, the spatial portion of spacetime is represented mathematically by some convenient number of dimensions. In the most usual practice, Minkowski spacetime of 3 spatial dimensions and one time dimension is utilized. In the Minkowski space, we now consider that "space" to be a vacuum scalar potential. If we consider the internal Stoney-Whittaker-Ziolkowski biwave pairs comprising that vacuum potential, we may transmit a signal communication along one or more of these internal biwaves, without changing any of the others. In that case, no overall potential gradient and overall forcefield signal appears. That domain that is infolded "inside" the scalar potential is said to be subspace. It may be modeled as a space of n-dimensions, where those dimensions are compacted and not observable to an external observer. Subspace gained great popularity from the Star Trek television series, where the spaceships used "subspace communications" to have signals propagate far faster than the speed of light. Bill Fogal has been working with some success on what should be a concrete realization of just this very "subspace communication system", with infolded signals that move much faster than the speed of light in vacuum.

SUPERCONDUCTIVITY

Property of some materials or metals whereby their resistance reduces to zero at very low temperatures.

These materials also exhibit many additional properties that are anomalous when compared to normal materials. Much research effort has gone into the attempt to find materials which exhibit superconductivity at higher temperatures, with the goal of reaching room temperature superconductivity.

SUPERLUMINAL

In a manner or with a speed faster than c, the speed of light in vacuum.

SUPERLUMINAL COMMUNICATION

The movement of signals faster than the speed of light in vacuum,

...and the transmission and reception of such signals and extraction of their transported information.

Longitudinal EM waves are not limited to speed c; transverse EM waves are limited to speed c. The infolded electrodynamics inside all EM fields, potentials, and waves consists of longitudinal EM phase conjugate wavepairs, with their impressed electrodynamics. Hence this "subspace" is a superhighway for such waves, including for those moving at superluminal velocity. In essence this is a very special kind of "tunneling" which we might refer to as *subspace tunneling*. The limitation to light speed rigorously applies to the modeled transverse EM waves which are a sort of "bulk upheaval" waves similar to waves on the surface of the ocean. As in the ocean wave analogy, "pressure" waves underneath the ocean may travel very much faster than the bulk upheaval waves on the surface.

When making longitudinal EM waves, the process is imperfect and the resulting longitudinal EM waves will have a transverse residue as well. The resulting "transverse-noisy" LW is called an *undistorted progressive wave*. Undistorted progressive waves may move slower than the speed of light or faster than the speed of light. See W. A. Rodrigues Jr. and J.-Y. Lu, “On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 \le v < \infty$ in nature,” **Foundations of Physics**, 27(3), 1997, p. 435-508. A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory web site, and includes corrections to the published version.

SUPERPOSE

To lay or place on, over, or above something else, with the connotation of mixing.

SUPERPOSITION

The simple linear addition and subtraction of two or more values, states, etc.

Superposition is one of the key principles in field theories and in the concept of potentials.

SUPERSET

A large set referenced to a small set, where the small set is included in the larger set.

The large set is said to be a *superset* of the small set. The small set is said to be a *subset* of the larger set.

SW (STONEY-WHITTAKER) STRUCTURES

The Stoney-Whittaker hidden bidirectional EM wavepairs comprising the scalar potential, and structuring (altering) of same.

We call this internal EM dynamics inside every EM wave, field, and potential the "infolded" electrodynamics. Russian weapon scientists refer to it as the *information content of the field*.

SYMMETRICAL REGAUGING

In electromagnetics, changing the scalar potential to a different value, and simultaneously changing the vector potential to a different value, with the two changes just precisely selected so that the two extra forces (each created by one of the changes) are equal and opposite—thus canceling with respect to translation and doing no net external work.

We note that the two forces themselves must be considered to be present and still working continuously in opposition, increasing the stress upon the system regauged with excess energy, or decreasing the stress upon the system regauged with less energy.

This presence of "equal and opposite" excess forces from the symmetrical regauging is known as the *Lorentz condition*. The end result is that the *net* force fields of the regauged system remain the same, even though the stress of the system has been altered by the regauging and the potential energy of the system has also been changed. Since abstract vector mathematics contains no such thing as "stress" in its background vector space, zero vector resultant system of multiple nonzero vectors are most often simply discarded. Early electrodynamicists symmetrically regauged Heaviside's two "Maxwell" equations (potential form) so as to easily separate the variables. The net effect was to discard that set of the Heaviside/Maxwell equations that prescribed and included open electrodynamic systems that freely receive and collect energy from the external vacuum, and also obtain a free excess force which then can permissibly utilize that excess energy to translate and produce external work upon the system or externally upon a load. In short, the symmetrical regauging discarded practical overunity electrodynamic circuits and devices, which Maxwell's equations—even in their Heaviside modified form—have always prescribed and permitted. Contrast to asymmetrical regauging.

We specifically point out that the two extra forces remain, as does their energy, and both may be dynamic as the system is dynamic. Hence when one regauges, one has changed the gravitational aspects of the system and also curved the local spacetime. Further, this ST curvature is internally structured and that structure is dynamic. Thus contrary to the assumptions of electrodynamics, the regauged system *is not* the same as the original system.

SYMMETRY

A theory or process possesses a symmetry with respect to certain operations performed on it, if those applied operations do not change the theory or process.

The theory or process is then said to be "symmetric" with respect to those operations. As an example, there is no observable change in a circle when it is rotated or reflected in a mirror. The circle is therefore said to have rotational symmetry and reflection symmetry.

TANTALUM CAPACITOR

Electrolytic capacitor with a tantalum or sintered-slug anode,

...such as solid tantalum, tantalum-foil electrolytic, and tantalum-slug capacitors.

For the same capacitance, the weight and volume of the tantalum capacitor are less than comparable aluminum electrolytic capacitors. In other words, you can make the little son-of-a-gun smaller, for a given capacitance and charge it must hold. When utilized in the Fogal fashion, the tantalum capacitor also exhibits other properties that are still proprietary to Fogal's semiconductor processes and to Bill Fogal.

TARGETED PINPOINTS

Refers to those separated points which engage in self-targeting between themselves and some separated signal energy source (as in Fogal's semiconductor).

The effect is to sharply narrow the retroreflected source signal energy from the source onto the separated points by means of the mutual ping-ponging. In this fashion much more energy density from the source can be concentrated on the targeted pinpoints. In effect, phase conjugate shooting is evoked.

TENSOR

A special mathematical entity which is an element of a vector space such that it represents the tensor product of vector spaces and their duals.

Much of modern physics and advanced electromagnetics is expressed in tensor algebra theory.

TESLA, NIKOLA

Famous American emigré electrical inventor and eccentric whose numerous inventions fueled much of the electrical 20th century.

Tesla produced the AC system, including polyphase systems. From his experiments he knew that the EM wave in vacuum was longitudinal, like sound waves. In spite of his great accomplishments, his eccentricities and grandiose pronouncements did not endear him to electrodynamicists! An example is his acrid comment: *"...I showed that the universal medium is a gaseous body in which only longitudinal pulses can be propagated, involving alternating compressions and expansions similar to those produced by sound waves in the air. Thus, a wireless transmitter does not emit Hertz waves which are a myth, but sound waves in the ether, behaving in every respect like those in the air, except that, owing to the great elastic force and extremely small density of the medium, their speed is that of light."* Another major example is his statement: *"The Hertz wave theory of wireless transmission may be kept up for a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history."*

Ironically, modern quantum mechanics has almost proven Tesla right that "the universal medium is a gaseous body." Today the modern QM vacuum is very much like a virtual particle gas. And of course the discussions in this present annotated glossary show why I personally am of the opinion that Tesla was also correct in his firm disbelief that those Faraday-Maxwell "taut physical strings without physical holders" filled the ether and provided transverse EM wave oscillations. I also agree with him that the EM wave in vacuum is a longitudinal wave, exactly like a sound wave. I believe that the transverse gyroelectron precession waves that our instruments actually measure in the conductor when it receives an incoming EM wave from the vacuum, proves that the vacuum wave is indeed longitudinal—else it would seem we must discard the spin of an electron and the theory of gyroscopes. Finally, one should see, e.g., Patrick Cornille, "Inhomogeneous Waves and Maxwell's Equations." Chapter 4 in **Essays on the Formal Aspects of Electromagnetic Theory**, Aklesh Lakhtakia, Ed., World Scientific, New York, 1993. Cornille clearly shows that Maxwell's equations are directly created by scalar potential interferometry. As he states, this means that EM force field wave s are created by the interference of sound-type scalar potential waves, or "sound creates light." This strongly supports Tesla's original statement that EM waves in the vacuum are actually EM sound waves, much like sound waves i n a gas.

TESLA'S MAGNIFYING TRANSMITTER (include Jackson quote)

To be added.

THERMOACOUSTICS

The generation of sound waves by heat.

THERMOACOUSTIC ENGINE

Essentially a closed resonant chamber containing three heat exchangers and a working fluid—such as an inert gas—under pressure, and used with a transducer to provide power such as mechanical power.

One heat exchanger is the hot-side exchanger through which heat is injected, another is the cold-side exchanger in which a coolant is circulated, and the third is a passive device between the other two, and known as a regenerator or thermal capacitor. The heating and cooling changes the pressure in the fluid greatly, and the regenerator oscillates back and forth, taking energy from the fluid, then giving it back, and keeping the engine going. A pressure wave (shock wave) moves through the working fluid and produces a resonator effect. The end result is an engine with rapidly fluctuating pressure gradients and no moving parts. The pressure gradients are then tapped off to drive a piston, which drives a linear alternator producing electricity. In a typical application, a single piston with a surface area of 1.5 square inches, and with a travel of only one centimeter, operated at 500 Hz, will produce about 30 horsepower output.

Information on these engines can be found on the Fellows Research Group website at http://www.io.com/~frg, or E-mail to frg@io.com. Many thanks to Oscar L. Fellows, "Be Our Guest: Thermoacoustic Engines—Propulsion, Power for the 21st Century," **Power Online**, 1/24/2000.

THERMODYNAMIC EQUILIBRIUM

Stable state eventually reached by an isolated system. One may consider it a "balanced" or "preferred" state.

Actually, a totally isolated system could not change. Any system is in a constant exchange with its vacuum environment, a priori, and no such thing as a completely isolated system exists in nature. Hence once excited, a system—to reach equilibrium—must exchange with its environment and is not a truly isolated system. But systems do exchange with their environment in hidden forms, so we can just let the original statement stand as is. Sloppy, but okay so long as we recognize the limits.

TIME

To be added.

TIME-CHARGE

To be added.

TIME-CHARGING

To be added.

TIME-ENERGY

To be added.

TIME-EXCITATION

To be added.

TIME-FLOW MECHANISM

To be added.

TIME-POLARIZED ELECTROMAGNETIC WAVE

An electromagnetic wave whose vibrations occur in the time domain only.

The wave is theorized in quantum field theory, or at least its time-polarized photon is theorized and called a "scalar" photon. There seems to be no Western method for detecting such waves, though they are predicted by theory.

TIME-POLARIZED (SCALAR) PHOTON

A photon whose vibrations occur in the time domain only, also known as a "scalar" photon.

TIME-FORWARD WAVE

A normal wave moving in forward time. Otherwise known as the retarded wave.

The external observer comfortably sees its forward progress as expected.

TIME REVERSAL

The process of forming the phase conjugate

...(e.g., of a wave or particle or entity or process or vacuum engine.).

We accent that *time-reversing a single object or single group of objects* is not the same thing as "travel into the past" so popularized by science fiction. For time travel to the past, the entire universe and everything in it except the traveler, would have to be time-reversed. That would not seem possible by any stretch of the imagination today! On the other hand, time-reversal of a single thing such as a particle or a wave—or even a group of things such as a group of particles or waves—is not only feasible but is readily achievable. A positron is a time-reversed electron, e.g. Becker showed that laughably weak DC electrical trickle currents (picoamperes) across otherwise intractable bone fractures could engender healing of the bone fractures. In investigating the effect, he showed that entire cells—specifically in that case, the red blood cells—were time-reversed (dedifferentiated) and then fast-forwarded (redifferentiated) into new kinds of cells, *genetics and all*.

In France in the 1960s and early 1970s, the Priore team showed in laboratory animals that this mechanism (which they nor anyone else understood they were unwittingly evoking) could be used to dramatically reverse and quickly cure terminal tumors, infectious diseases, arteriosclerosis, and suppressed immune systems. The same approach, if redeveloped, could stop and totally wipe out the emerging major new pandemic of hepatitis C, which nothing will touch at present and which will surpass AIDS infections after the turn of the century. The approach could totally reverse AIDS and eliminate all HIV infection, because it could simply time-reverse HIV-infected cells back to normal cells, genetics and all. In other words, it could readily and cheaply and quickly eliminate AIDS from the face of the earth. Further, the disease could be eliminated whenever a patient tested positive for HIV infection; there would be no need to wait for debilitation to set in. For coverage in detail, see T. E. Bearden, "Vacuum Engines and Priore's Methodology: The True Science of Energy-Medicine. Parts I and II." **Explore**!, 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62.

TIME-REVERSAL ZONE

A region of space in which a component of the overall time flow runs backwards, and hence the fundamental mechanism generating the flow of a mass through time is reversed.

Such zones are theorized as temporarily occurring in cold fusion, particularly once the palladium lattice is loaded with hydrogen or deuterium ions. Such a temporary region in the electrolyte is an excited state, and the other ions will move to nullify it and cause its decay. During its existence, ions in it experience a reversal of the law of attraction and repulsion of charges. In the TRZ, like charges attract and unlike charges repel. Thus suddenly the low spatial energy transmutation of elements in cold fusion experiments can be explained (they have been validated in some 600 successful experiments, in various laboratories in various nations, by excellent scientists). E.g., two H+ ions may attract, and so closely that each enters the edge of the strong force region of the other, forming a quasi-nucleus. Then as the TRZ decays, the preferred energetic change is from inside out. In other words, the gluon forces have been partially nullified and the quarks in the nucleons are almost freed. Hence the favored energetic decay from the excited state of the TRZ is by quark flipping, where one H+ ion in the quasi-nucleus changes into a neutron. The quasi-nucleus becomes a quasi-nucleus of deuterium, and will tighten into deuterium as the TRZ induced action vanishes. This is the way the excess deuterium is formed in the cold fusion experiments. Two deuterium D+ ions may also form a quasi-nucleus of (D+)(D+) particularly where the electrolyte is loaded with deuterium to start with. In that case, as the TRZ decays, the quasi-nucleus just tightens into a helium 4 nucleus or alpha particle, accounting for the excess alpha particles formed. And so on.

The concept of the TRZ thus is consistent with hundreds of successful experiments, and introduces a completely new type of nuclear reactions to physics, at low spatial energy where transmutations such as occur are not presently thought to be possible. We point out that these are extremely high energy reactions when the time-energy interaction is considered, since time has an energy density equal to mass. Time-energy is some c^2 times as dense as is spatial energy used by particle physicists in their interactions. By transposing some of the time-energy into spatial energy as the TRZ decays, the new "low energy nuclear reactions" actually involve the necessary degree of "high energy" required for the interactions.

TIME-REVERSED WAVE

A phase conjugate reflected wave, which is a wave moving in reversed time.

Since the external observer "sees" in forward time only, he sees the wave as traveling in the opposite direction as its generatrix wave front is moving.

TOPOLOGICAL

Invariant under transformation by continuous mappings,

...or related to the properties of geometric shapes and figures that remain invariant under transformation by continuous mappings.

TOPOLOGICAL PROPERTY

Those properties of a geometrical space that are unchanged by continuous distortion of the space.

TOPOLOGY

Mathematics of the properties of geometric configurations invariant under transformation by continuous mappings.

In short, the mathematics of the large-scale structure of curved spaces.

TOROIDAL COIL

A coil of wire curved around into toroidal form.

Contrast to the standard solenoidal coil form.

TOROIDAL TRANSFORMER

A transformer comprised of toroidal coils instead of solenoidal coils.

TRANSFORMATION MATRIX

A matrix (rectangular array of algebraic or numerical quantities) for replacement of the variables in an algebraic expression by their values in terms of another set of variables.

Used, e.g., in solving equations involving Hamiltonians.

TRANSVERSE

Wave movement perpendicular to the wave's primary direction of principal propagation, as in the movement of a taut rope in a wavelike motion traveling down the rope.

(As opposed to longitudinal).

TRANSVERSE EM WAVE IN THE VACUUM

The completely erroneous notion that EM disturbances in the vacuum are composed of transverse "string waves."

Faraday, e.g., believed that the EM field in the vacuum consisted of his "lines of force" as material taut strings—erroneously without the necessary holders to put the tension forces on the strings. He thought that when there was an EM disturbance, it was a "plucking" or transverse vibration of these "taut strings." Maxwell also believed in Faraday's material lines of force, and set out to capture them in his electrodynamic theory. He did so by simply assuming the lines (tubes) of force and their transverse "string wave" vibration. This is equivalent to assuming that the electromagnetic vacuum itself is composed of material taut strings, and that additional EM fields are just additional taut material strings.

TRANSVERSE FIELD COMPONENT

The modeled component of the electromagnetic field (as in a moving EM wave in vacuum) that is oriented at right angles to the line of motion.

TRANSVERSE WAVE

A wave where the medium is undergoing repetitive lateral translations, as a lateral wave in a taut string.

The EM wave is vacuum is erroneously modeled as a transverse "string wave," as a relic from Faraday's original belief that EM fields existed in the (assumed) material ether as physical taut strings. He thus believed that EM disturbances were the disturbances of "pluckings" of these taut strings. Maxwell accepted Faraday's physical lines of force and constructed his electrodynamics mode to specifically include the material ether, the physical "taut strings" lines of force notion, and Faraday's "string waves" as the electromagnetic vibrations in the ether. This discards half the energy, half the phenomena, half the wave, and Newton's third law causative EM mechanism.

TUNNELING

In quantum mechanics, a phenomenon where a particle penetrates and crosses a small region where the opposing potential is greater than the particle's available energy.

The phenomenon is thought to be impossible according to classical physics, but occurs nonetheless. Further, the speed of transit through the "tunneling" separation can be superluminal. Tunneling of music between two points has been measured to occur at more than four times the speed of light.

A more extended type of tunneling is the travel of longitudinal EM waves infolded inside the normal EM potentials, fields, and waves. Indeed, this is a more fundamental electrodynamics, and it can be shown that interferometry of this "internal" electrodynamics creates the external electrodynamics by a form of longitudinal EM wave interferometry. E.g., see M.W. Evans, P.K. Anastasovski, T.E. Bearden et al., "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," **Journal of New Energy**, 4(3), Special Issue, Winter 1999, p. 76-78.

UNCONSCIOUS MIND

To be added.

UNIFIED FIELD THEORY (ENGINEERABLE)

A unified theory of the four forces of physics—the electromagnetic, gravitational, strong, and weak forces —that is not just an intellectual model but which is also engineerable on the laboratory bench and in actual physical systems using higher symmetry O(3) electrodynamics.

That's a mouth-filling phrase (yes, we sometimes write awkward phrases unintentionally; it's amazing how reasonable such things sound when you're laboring along at 3:00 a.m. and bleary-eyed as a wet toadstool!)

UNIFIED FIELD THEORY OF MIND AND MATTER

To be added.

UNITARITY

Principle that all probabilities sum to unity.

UNITARY Q_E

Fixed or unit charge q of the electron.

This notion is held only because the q is not decomposed into its component elements, but treated as if it were an indivisible unit. That is in error! The mass of a fundamental particle at rest is quantized, and the massless charge (self-potential) of the particle is discretized as a function of the background potential (virtual particle flux density) in which it is embedded or to which it is exposed. In our view the definition of a charge q is $q \equiv m_q \phi_q$. The *electrical charge* should be taken to be the massless ϕ_q component. The definition of charge q_e of an electron is rigorously given by the identity $q_e \equiv m_e \phi_e$. The *electrical charge* of the electron should be taken to be the massless ϕ_e component.

If a charge (say, an electron) is placed in an additional potential ϕ_1, then the total massless charge of the electron (i.e., its virtual vacuum flux exchange with mass me) is $(\phi_e+\phi_1)$. As can be seen, the massless charge of the electron (its virtual flux activity) has now changed. The electron is then said to *be collecting and to have collected* excess energy (which may be either positive or negative, depending upon the sign of ϕ_1.). Further, either ϕ_e or ϕ_1 or both can be internally and deterministically *dimensioned* (structured in Stoney-Whittaker-Ziolkowski fashion). When its massless charge (its own potential ϕ_e) is so dimensioned, then the dimensioned electron is said to contain or transport a *vacuum engine*.

VACUUM ENGINE

Particle view: Deterministic internal pattern or template deliberately created in the virtual flux of a scalar potential.

It can also be created in, and transported by, the two scalar potentials that comprise a static or dynamic force field, or a moving EM wave in the vacuum. It can also be created in the local vacuum, in which case it has a "charge-up" creation time and a "discharge time," exhibiting a sort of "time constant" very much like a capacitor charging and discharging. In the wave view, the vacuum engine is a deterministic internal pattern or template deliberately created in the Stoney-Whittaker-Ziolkowski internal biwaves and wave products comprising the scalar potential. Again, it can be implanted in the vacuum or in an EM signal as a carrier. The self-energy (scalar potential) portion of any electric charge or magnetic charge can also be a "carrier" that is conditioned (*dimensioned*) with a vacuum potential. The process or act of conditioning any of the carriers with vacuum engines is called *dimensioning* that carrier. Russian energetics refers to the vacuum engines of a dimensioned carrier as the *information content of the field*.

VACUUM FLUX EXCHANGE

The virtual particle interaction between the active vacuum and a charged particle.

The charged particle is an asymmetry in this flux exchange interaction, and extracts and gates some of the vacuum energy as a Poynting **S**-flow. In turn, this S-flow transports and creates in space the **E** field and the scalar potential f associated with the charged particle. An electrically charged particle is thus a little Poynting energy generator, a source of potential, and a source of **E**-field. It is a little electrical dynamo, a little free energy generator. It is already an open system in disequilibrium with its environment—the fierce vacuum flux. Its self-energy (potential) and its "fields," plus its Poynting energy flow, are the result of its asymmetry in the vacuum flux exchange.

VACUUM'S VIRTUAL PARTICLE FLUX

In the particle viewpoint, the quantum mechanical vacuum identically is a virtual particle flux.

Present physics, however, is prone to refer to the vacuum as (an implied emptiness) containing a virtual particle flux. Energy, vacuum, spacetime, virtual particle flux, and dynamic geometry are all synonyms.

VECTOR

In mathematics, an entity completely characterized by a magnitude and a direction.

VECTOR POTENTIAL

In electrical engineering: "Potential postulated in electromagnetic field theory."

Space differentiation (*curl*) of the vector potential yields the field. Magnetic vector potential is due to electric currents, while electric vector potential is assumed to be due to a flow of magnetic charges." [**Chambers Science and Technology Dictionary**, Peter M. B. Walker (Ed.), Chambers Cambridge, 1988, p. 947.

Note that the foregoing is not yet a definition, saying only that it is a postulated potential, then giving some characteristics. Informally, in my view the vector potential is just sort of an ordinary scalar potential, every point of which is in motion in something like a swirling fashion. The magnetic vector potential **A** is most common. There doesn't seem to really be a definition of **A**; the equation $\mathbf{B} = \nabla \times \mathbf{A}$ is usually (erroneously!) said to "define" the **A**. So we are left with the question **A** ≡ ????, since no equation is a definition of anything.

Electrodynamicists have no difficulty in calculating the direction and magnitude of **A**, and in using it, but so far as I can find there is still no real definition. As an ad hoc definition I consider the **A**-potential to be identically a virtual photon flux of vacuum that is moving as an ensemble along the direction of the **A**-vector. Inputs for a more rigorous definition of the form **A** ≡ (x) would be most welcome!

At any rate, the **A** is known to be quite real and an independent "field" of nature, because when the curl operator is removed from operating on **A**, the **B**-field is zeroed but the *curl-free* **A**-field can still remain and interfere with the fields of charged particles and magnetic poles to cause real effects in physical systems, including at a distance. This latter is known as the Aharonov-Bohm effect.

VECTOR FIELD

In ordinary theory, the assigning of a magnitude and a direction (that is, a vector) to each point in a region of space.

The entire set of vectors and their spatial points is called a vector field.

VECTOR ZERO

See discussion under *zero vector*.

VIRTUAL ANTIPHOTONS

An antiphoton that spontaneously appears and disappears, so swiftly that it cannot be individually observed.

VIRTUAL PARTICLE

A fleeting quantum particle that spontaneously appears and disappears, so swiftly that it cannot be individually observed, so that it exists only temporarily.

The virtual particle does not satisfy the usual relation between energy, momentum, and mass because it is underneath the Heisenberg uncertainty principle. The virtual particle can have any amount of energy momentarily, so long as the product of its energy and the time interval of its existence is less than the uncertainty principle's minimum magnitude.

Nonetheless, the interactions of large numbers of virtual particles with a mass or charge can combine to generate real observable effects. In quantum field theory, all forces of nature are caused by the interaction of the forced mass entity with virtual particles. Interaction of a mass with virtual photons, e.g., is projected to cause all mechanical and electromagnetic forces.

VIRTUAL PARTICLE FLUX EXCHANGE

Interaction with or exchange of, a flux of virtual particles.

Interaction of a mass with virtual particles most often consists of the mass absorbing the virtual particles and re-emitting them.

VIRTUAL PHOTON FLUX (VPF)

A flux of virtual photons.

(Hey, how's *that* for a tautology!) First, see definition and discussion under *flux*. For electromagnetics, the vacuum can often be modeled as a flux of virtual photons. One gets the virtual charges also, since the photon is considered to also consist of a positron-electron pair. That is, the photon is continually separating into a positron-electron pair, which then annihilate each other to form additional photon(s), etc.

VIRTUAL PHOTON FLUX EXCHANGE

The mutual exchange of virtual photons, as between the vacuum and a charged mass, or between two colliding masses.

The emitted photons from one mass are absorbed in the other mass and re-radiated.

VIRTUAL PHOTON FLUX OF VACUUM

The vacuum identically is a flux of virtual particles.

For most electromagnetic cases, it can just be modeled as identically a flux of virtual photons.

VIRTUAL PHOTONS

Photons which spontaneously appear and disappear in the vacuum so quickly that they cannot be individually observed.

VIRTUAL POSITRONS

Positive electrons (positrons) which spontaneously appear and disappear in the vacuum so quickly that they cannot be individually observed.

WHITTAKER, E. T.

Well-known mathematical physicist, who dramatically extended Stoney's earlier work of 1897-8 and authored two papers in 1903 and 1904 giving the basic theory of scalar electromagnetics (Russian energetics, not including the quantum potential portion).

The 1903 paper is E.T. Whittaker, “On the Partial Differential Equations of Mathematical Physics,” **Mathematische Annalen**, Vol. 57, 1903, p. 333-355.

The 1904 paper is E.T. Whittaker, “On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,” **Proc. Lond. Math. Soc**., Series 2, Vol. 1, 1904, p. 367-372. The paper was published in 1904 and orally delivered in 1903.

WHITTAKER'S DECOMPOSITION OF THE SCALAR POTENTIAL

Decomposition of the scalar potential into a harmonic set of bidirectional longitudinal EM phase conjugate wavepairs.

In his paper “On the Partial Differential Equations of Mathematical Physics,” **Mathematische Annalen**, Vol. 57, 1903, p. 333-355, Whittaker showed that the scalar EM potential can be decomposed into a harmonic series of bidirectional EM wavepairs. In each pair, one wave is the phase conjugate of the other, and exists in the time domain. However, this phase conjugate wave has been considered as having been detected by intercepting charge as a 3-space EM wave travelling in a reversed direction. Hence from a detectable viewpoint, the Whittaker decomposition yields bidirectional longitudinal EM waves in 3-space, as has been pointed out by several electrodynamicists (e.g., Barrett).

However, the present author considered the difference between the phase conjugate wave as it exists prior to interaction with the assumed intercepting unit point charge, and the detected 3-space wave after the detecting charge is assumed. By applying the Whittaker decomposition and also considering the nature of the EM wave in spacetime *prior to interception by a unit point charge*, this led to the discovery of a new and more primary 4-symmetry in EM energy flow, once 3-symmetry is broken, that exists automatically between the time domain (complex plane) and 3-space.

Thus because a dipole (or a charge treated as a set of composite dipoles) breaks 3-symmetry (as is well-known in particle physics), the more general 4-symmetry results automatically when the nature of the potential between the ends of the dipole(s) is considered. This yields the fundamental mechanism by which EM energy is extracted from the vacuum (from the complex plane or the time domain) and output in 3-space from a "source charge" or a "source dipole". In short, it yields the giant negentropy mechanism when the unaccounted Heaviside energy flow component from a charge or dipole is considered as well as the intercepted tiny Poynting energy flow component. See T. E. Bearden, "Giant Negentropy from the Common Dipole", **Journal of New Energy**, 5(1), Summer 2000, p. 11-23; also on http://www. cheniere.org.

In his second paper, “On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,” **Proc. Lond. Math. Soc**., Series 2, Vol. 1, 1904, p. 367-372, Whittaker showed that the interference of two such scalar potentials can and does create all the field energy, patterns, and functions of classical electromagnetics.

Whittaker's second paper initiated what today is loosely referred to as *superpotential theory*. For an excellent overview discussion of superpotentials and related things, see Melba Phillips, “Classical Electrodynamics,” in **Principles of Electrodynamics and Relativity, Vol. IV of Encyclopedia of Physics**, edited by S. Flugge, Springer-Verlag, 1962.

WORK

The changing of the form of energy.

WORK, EXTERNAL

The changing of the form of the external energy of a body or system.

WORK, INTERNAL

The changing of the form of the internal energy of a body or system.

WORK-ENERGY THEOREM

To be added.

ZERO POINT ENERGY

The minimum energy of a system due to its quantum fluctuations, resulting from its incessant virtual particle activity. Quantum mechanically, no system of interest (including even spacetime itself) can have zero energy.

The zero-point energy of the vacuum is the lowest energy vacuum state, with fluctuations taken into account. Even at low energies, quantum fluctuations continually arise, and result in an incessant, extremely rapid, and violent "jittering" of the energy momentarily present. The minimum energy due to these quantum fluctuations is called the *zero-point energy.*

ZERO VECTOR

(1) A vector having no length (no magnitude) or specific direction.

(2) Absence of any or all finite vectors.

(3) Resultant of a system of multiple, finite vectors whose vector sum has no specific direction.

Note that the second definition differs drastically from the first. That is, we cannot say that a vector resultant zero has no infolded magnitude, if we consider its components, all of whom have magnitude. See the discussion under "zero." In vector analysis, the zero vector axiom states that any vector v plus a zero vector is equal to a zero vector plus v. Note that, being an axiom, this is not proven, but just assumed. We very much disagree with that blanket axiom; it reduces vector analysis to a highly special single case, where all zero vector systems are to be considered equal, and totally "inactive." In the real world that need not be true at all. There are an infinite number of different zero vector resultant systems, and so zero vectors may differ, if their infolded component "substructure" is considered. In the real world, things that are vectors very often have energy. The energy is infolded and hidden, so that it represents not only a local curvature of spacetime, but also a structured, deterministic set (a *template*) of curvatures of local spacetime. Therefore this zero vector system is actually a *vacuum engine*. Any mass system placed in such a potential, will be acted upon in its parts by that local vacuum engine. Further, the zero vector resultant system is a "potential," since it does contain "collected internal energy." This potential is obviously an artificial potential, since it is deterministically structured internally.

So between zero vector resultant systems, the "infolded, hidden energies" differ dramatically in

(i) total infolded and trapped (collected) energy,

(ii) specific internal action patterns (templates),

(iii) general relativistic effects, and

(iv) type of vacuum engine. Accordingly, the actions the zero-vector system induces in an exposed mass may be designed in advanced. This is vacuum engineering. This is spacetime engineering.

The problem is the *vector algebra itself*. In the abstract algebra, a "vector space" is regarded as an inert thing, where the only actants are the overall net nonzero vectors, and there is no interaction between an overt vector and a covert set of vectors. The vector space is not allowed to have an internal realm, stress, hidden engines, or to interact with the vectors representing the physical system. Consequently, for over a century electrodynamicists—thinking primarily of translation—have routinely discarded such "zero vector systems" as if they represented the total absence of any finite vector. *This they do, in the unsuitable assumptions of the vector algebra*. This they do not do, in the real world.

The common practice in electrical physics of replacing a zero-summed system of nonzero vectors with a vector zero of "complete absence of vectors" is incomplete, and such did not exist in Maxwell's original *quaternion* theory. This present practice should be changed; it has reduced the topology of electromagnetics far too much, and gutted the most important parts of the theory insofar as an extended electromagnetics is concerned. The vector zero should be replaced with a special organized vector zero system and a scalar potential. The interactions of the "hidden" vectors should be incorporated. By failing to do this, physics excludes the ability to engineer electrogravitation, spacetime, the virtual state, local general relativity, free energy, effects at a distance, and the probabilities of the states propagated by the Schrödinger equation. It also excludes unification of all forces.

In the new view, a zero vector may be a system of nonzero vectors that vectorially sum or multiply to zero for translation purposes, but the components may still exist "inside" and dynamically function inside the zero-resultant envelope. This substructure can act on nonlinear or resonant systems. Any or all components may be time varying, or none of them may be. The zero vector system may thus be a dynamic vacuum engine. In addition, such a vector zero system is considered to be also a scalar stress potential, whose magnitude is equal to the sum of the absolute values of the perpendicular components. In addition, a separate type of stress potential may be included for the magnitude of the absolute values of the swirl components (torques). Other associated potentials may be included for the system, particularly if the system is in multiple dimensions, if it consists of more than one type of virtual particle flux, if several fluxes intercommute, etc. Such a system can also be designed and created to deterministically interact with and on the life force itself, the mind, thought, and personality, but that is beyond the scope of this discussion.

ZERO-VECTOR-SUMMATION SURFACE

A surface or assumed spatial surface in which vectors under consideration interact to sum to a vector zero resultant.

ZERO-VECTOR-SUMMATION SYSTEMS

Zero summations of real, nonzero vectors, so that the real vectors are still present but do not cause translation of exposed mass or charges.

With respect to observing/detecting process that depend upon charge translation, the vectors are said to be hidden and infolded —i.e., with respect to detection by translation. See discussion under *zero vector*.

ZIOLKOWSKI, RICHARD

To be added.

Return to top of page

The "Porthole" Concept with technical background

Amplifying the cellular regenerative (healing) system of the body for treatment and healing of mass casualties

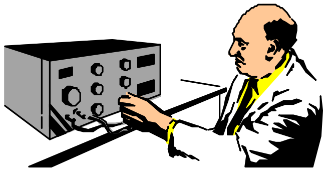

Outline of the presentation

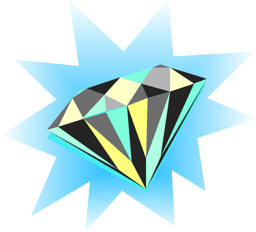

- Mechanism for deep EM exchange of a body with its active EM environment
- Becker's epochal work and its ramifications
- Kaznacheyev's startling results and ramifications
- Weaponization of Kaznacheyev's findings
- Importance of time-energy and the time-domain
- Mechanisms for medical treatment and healing

Outline of the presentation (2)

- Prioré's revolutionary results and their ramifications.
- Proposed new mechanism for cancer promotion.
- Porthole concept: The 2-minute shortcut.
- Terrorist biological warfare capabilities.
- Summary and conclusion.
- Budgetary and Program package.

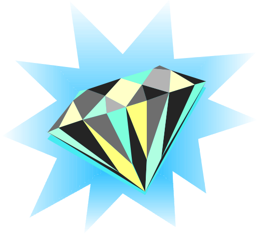

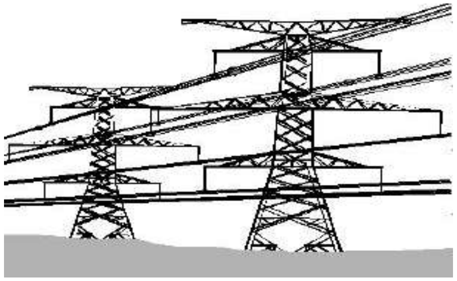

Mechanism for the body's deep EM exchange with its active EM environment

And some ramifications

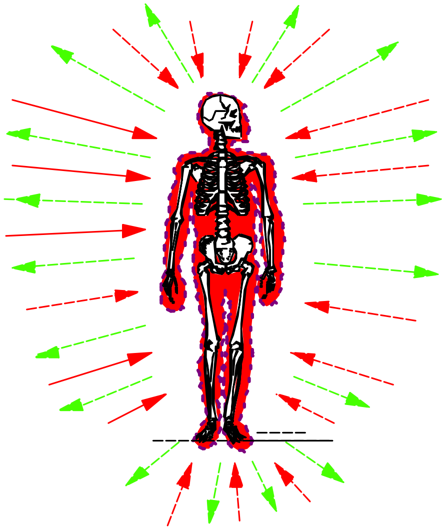

The body continually receives and retransmits EM energy

- The body may be considered as a dielectric.
- Every part of the dielectric participates in the absorption and emission of each photon to and from any point on the surface.*
- Mechanical force effects are also electrical in nature and exchange similarly.

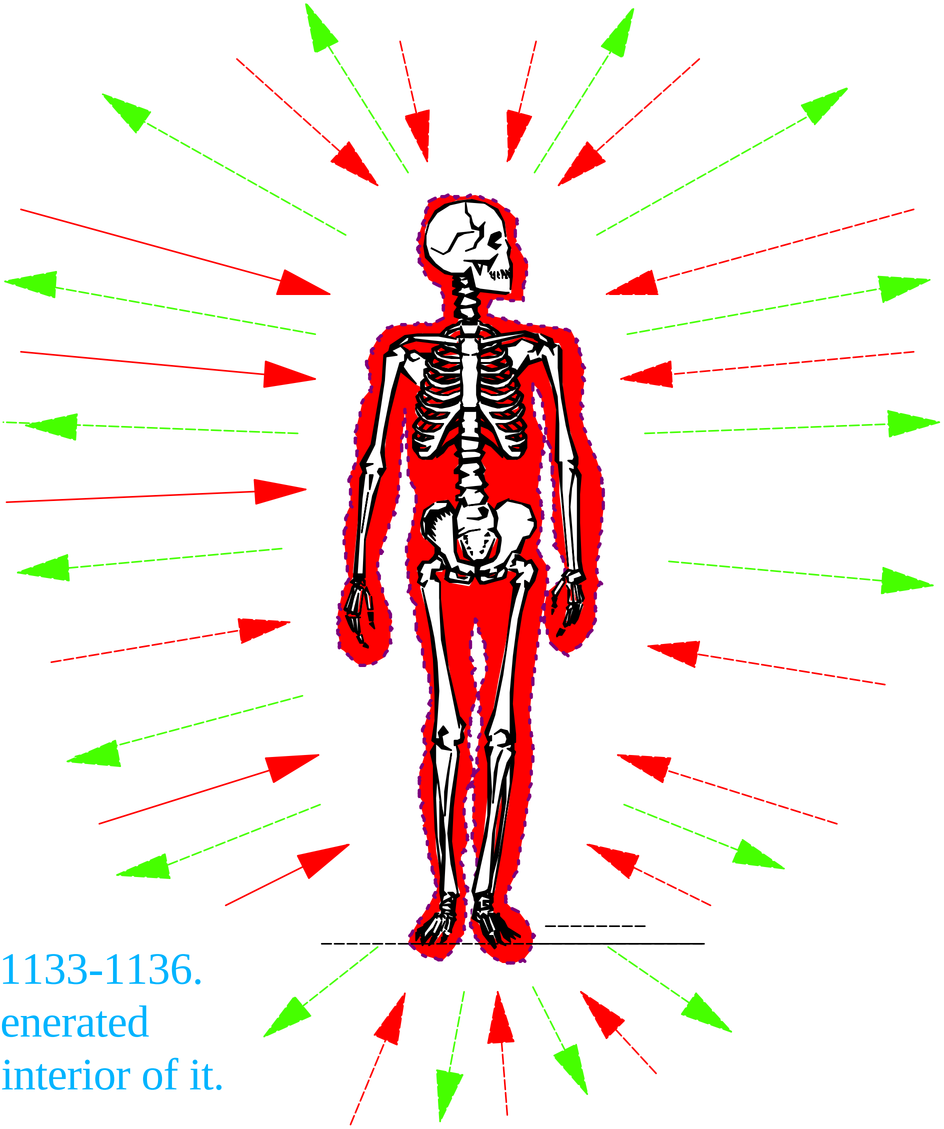

* Reali, G. C., "Reflection from dielectric materials," American Journal of Physics, 50(12), Dec. 1982, p. 1133-1136. The reflected field from a dielectric material is not generated just at its surface but comes from everywhere in the interior of it.

Ramifications:

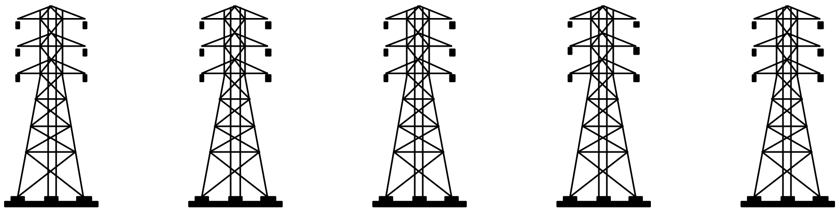

- There is an EM "dielectric transmission path" connecting every part of every cell to every part of the surface of the body.
- All electrical processes in the body, including deep within the cell and its nucleus, participate in the absorption and emission of every photon to and from the body's outer surface (and every mechanical force* experience).
- The EM emission energy of the body is a conglomerate mix of fractions of all processes and functions ongoing in the body or affecting it, including in every cell.
- This includes all electromagnetic, mechanical, chemical, and environmental effects upon and in the body and every tiniest part of it, and provides an extended definition of "stress".

* At base, all normal mechanical forces are generated by electromagnetic forces.

Ramifications (2)

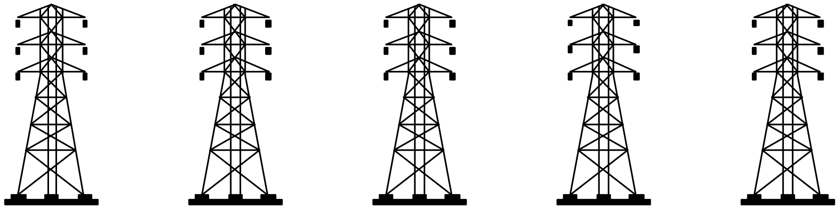

- The reaction of each photon absorbed on any point of the outer surface of the body, connects through the dielectric transmission path to every part of the body, including to every part of every cell*.
- Every emission from the body is participated in by every part of the body, including every part of every cell.
- The body's emission "changes" its environment, adding EM radiation. Some of the internal induced changes (excited states) in the body decay by this external emission. Some do not, but must be otherwise dealt with internally by the body systems.
- The body's EM absorption allows the environment to change or affect every function, operation, and part of the body, by adding EM radiation. All are continually being affected and deviated, both helpfully and harmfully, by these deeply induced EM interactions.
- The body's reaction (healing) must continually readjust and correct deviations from normalcy, in every part of every cell.

* In quantum physics, part of the photon is localized and part is nonlocalized, reaching even across the universe.

Ramifications (3)

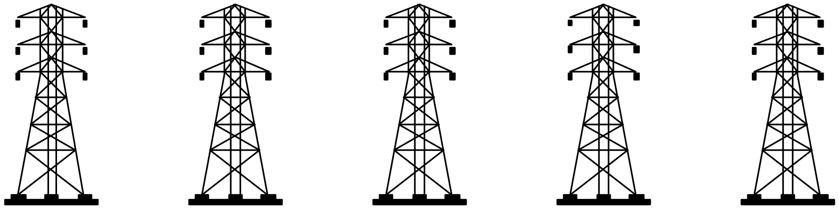

- The sum total of all the impinging external EM radiation upon the body, plus mechanical "special EM" radiation, thus affects and changes the body and all its processes, participating in them.
- That impinging EM radiation can be manipulated to either harm or heal the body and any or all of its cells.
- The body must continually overcome these externally-induced EM changes in its cells and processes, to maintain normalcy. Some EM changes are restored by the subsequent dielectric emission process, relieving the previously-induced deviant EM excitations in its internal processes. Some are not.
- The poorly studied *cellular regenerative system* performs the remainder of the restorative function, which is the fundamental "healing" function of the body.*

*The cellular regenerative system was mostly studied by Becker and a few others. See R. O.Becker, and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, No. 73, Nov.-Dec. 1970, p. 169-198.

Becker's epochal work

And its ramifications

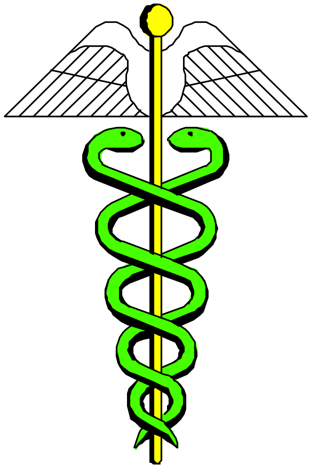

Becker's bone fracture healing

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

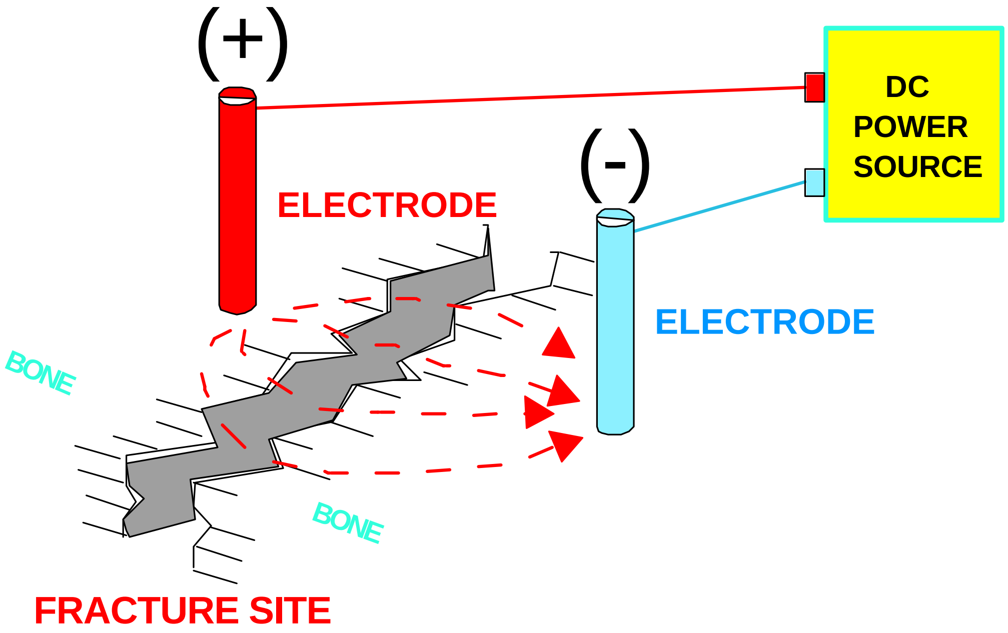
(+)
ELECTRODE
(-)
DC
POWER
SOURCE
ELECTRODE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Becker's theoretical DC control system involved with response to injury

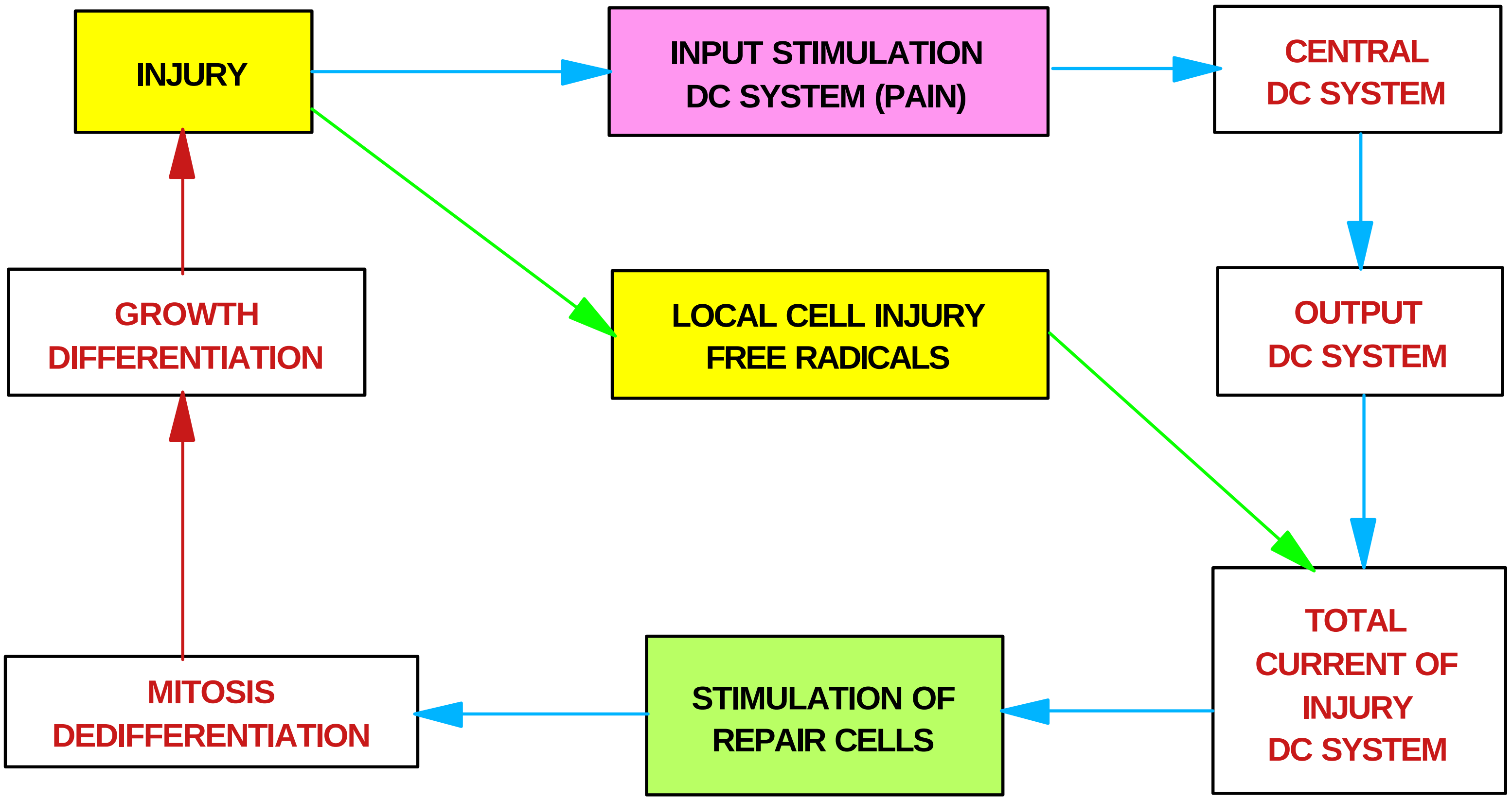

INJURY

INPUT STIMULATION
DC SYSTEM (PAIN)

CENTRAL
DC SYSTEM

GROWTH
DIFFERENTIATION

LOCAL CELL INJURY
FREE RADICALS

OUTPUT
DC SYSTEM

TOTAL
CURRENT OF
INJURY
DC SYSTEM

MITOSIS
DEDIFFERENTIATION

STIMULATION OF
REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Becker's proposed control system governing regeneration*

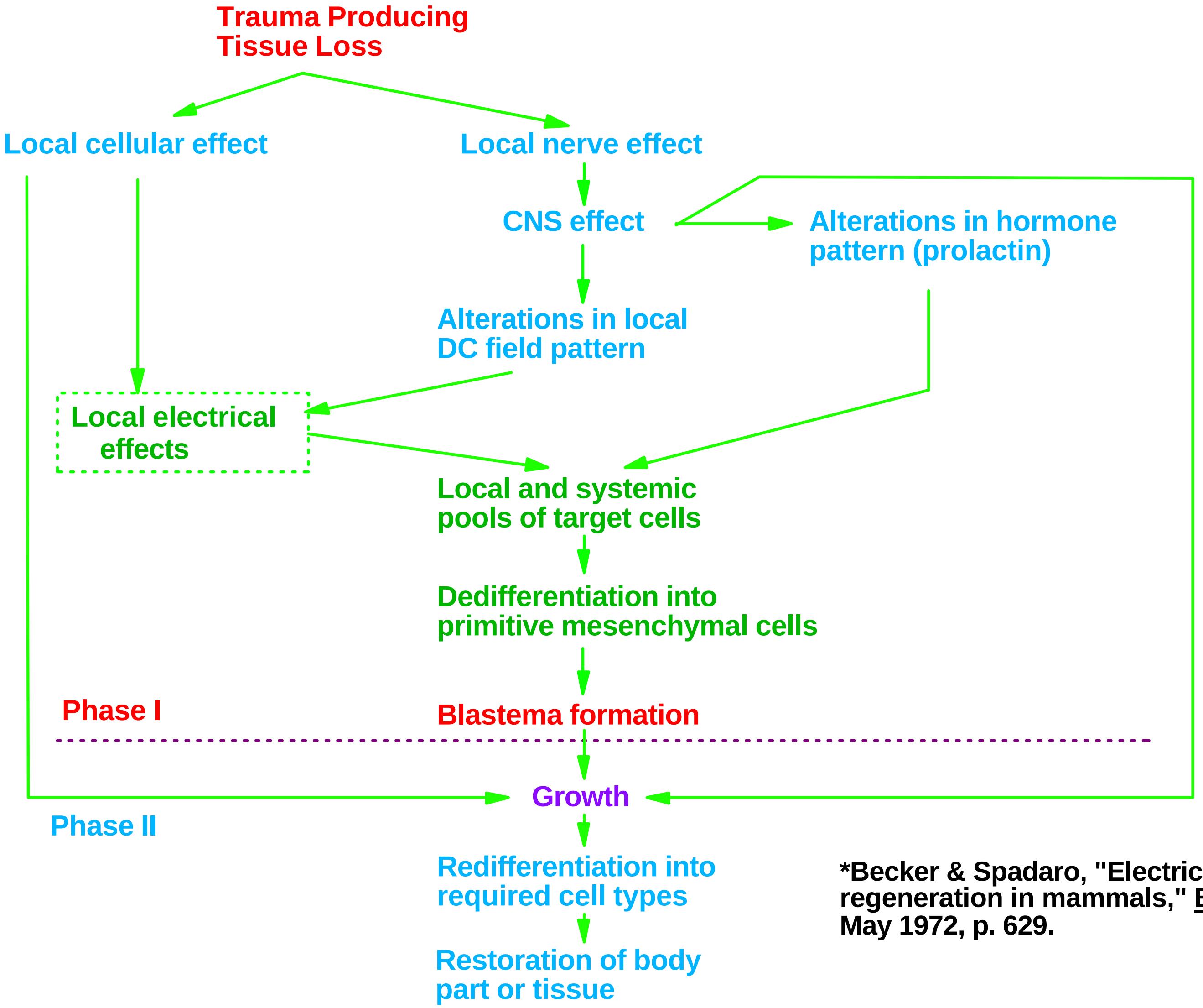

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

Growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Extending Becker's work

- Becker had only the archaic classical electrodynamics available to him. That EM is grossly inadequate.
  - Insufficient group symmetry to model the process.
  - Already eliminates the "infolded" longitudinal wave EM that structurally and dynamically comprises all "normal" EM potentials, fields, and waves.
  - It is this internal LW EM that is manipulated and used by the cellular regenerative system.
- Nonlinear phase conjugate optics was not yet developed when Becker did his seminal work.
- Successful higher symmetry electrodynamics models -- such as O(3) -- have been developed and are available.

A sad commentary

- Becker was nominated for the Nobel Prize, but he was hounded, his funds were withdrawn, and he was forced to retire early.
- His main "sin" was having the courage to testify against the harmful effects of unrestricted powerline radiation and electronic smog.
- Powerful interests then crushed him, set up their own well-funded institutes, and hired scientists who then predictably found that powerline radiation was no problem.
- Such research is so tainted that recently several important journals have now required the authors of submitted papers to state what their affiliated companies do, and what "company massaging" of their research occurred.
- From a higher symmetry EM view, it is obvious that the dielectric pathway and quantum electrodynamics already show a significant problem due to such radiation, particularly with respect to long-term effects.
- This is particularly significant since "noise" -- the mere presence of more "hash" -- appreciably amplifies the dielectric transmission effect. So the major factor becomes EM smog *signals density*, not signal power.

Kaznacheyev's startling results

And its ramifications

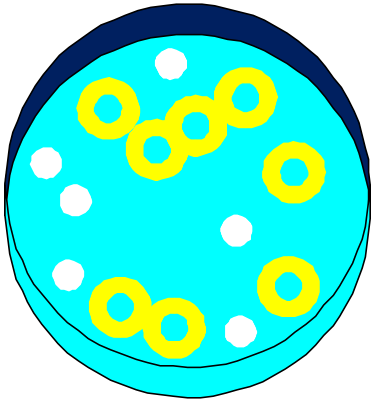

Proof that the cellular condition is radiated from groups of cells

- It follows that a diseased, damaged, or infected cell culture will and does emit its exact internal EM condition, hence the disease or damage "generatrix form" itself (in the internal structuring of the EM radiations).
- Further, these "state" emissions can couple into other targeted cells to produce disease and disorder in them.
- This was clearly proven in some 15,000 Russian experiments by Kaznacheyev.
- Kaznacheyev demonstrated induction of cellular disease and disorder between EM-coupled but otherwise environmentally shielded cell cultures
- Similar experiments have been replicated in the West by Reid *et al*., and by Popp *et al*.

Kaznacheyev's induction of cellular disease and disorder

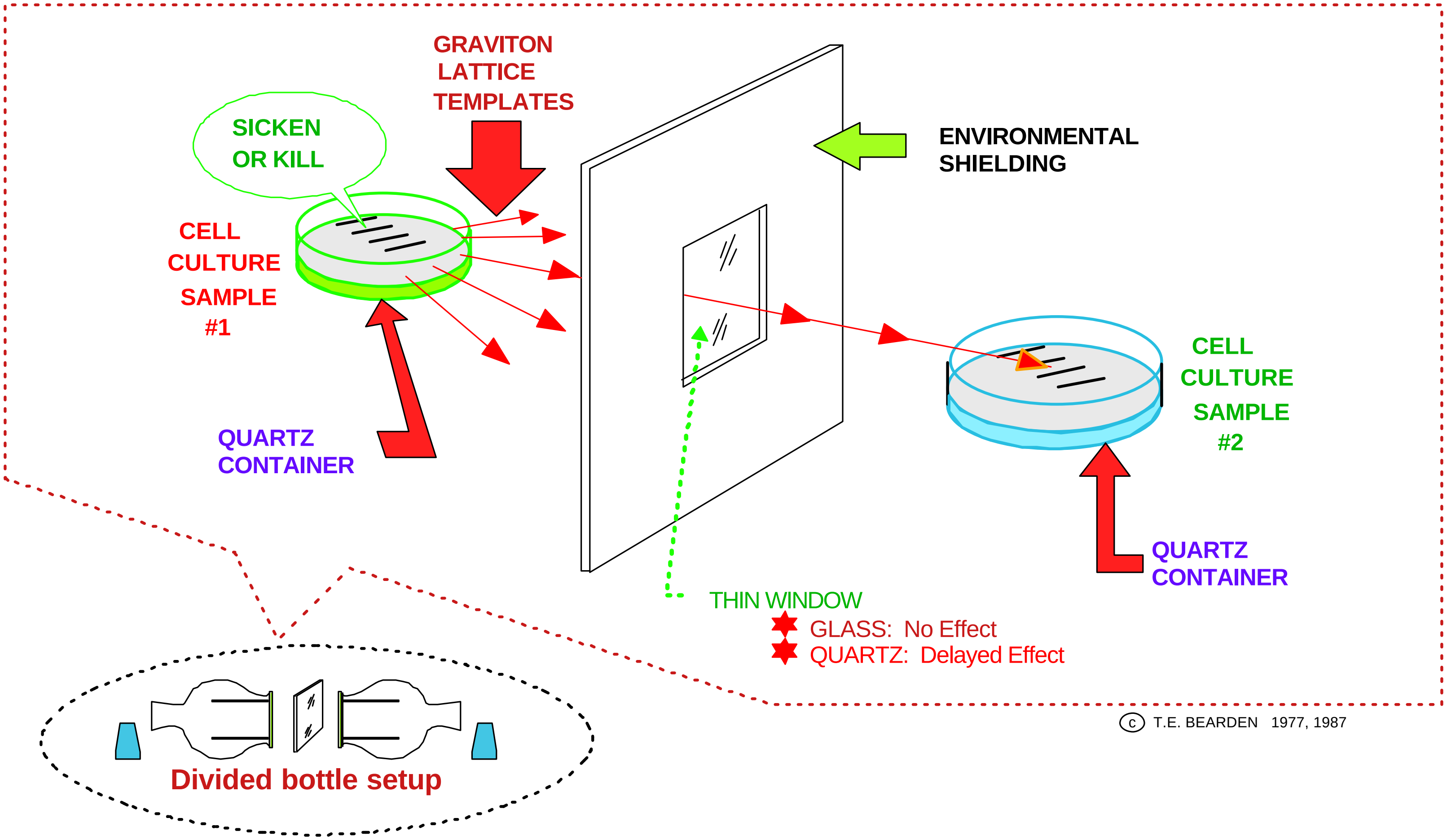

GRAVITON LATTICE TEMPLATES

SICKEN OR KILL

CELL CULTURE SAMPLE #1

QUARTZ CONTAINER

ENVIRONMENTAL SHIELDING

CELL CULTURE SAMPLE #2

QUARTZ CONTAINER

THIN WINDOW

- GLASS: No Effect
- QUARTZ: Delayed Effect

Divided bottle setup

ⓒ T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice. Consider particularly the dielectric pathway in the emitting diseased cells and in the receiving targeted healthy cells.

Another surprising feature of Kaznacheyev's experiments

- The experiments do not work if the targeted cells are in normal light, such as sunlight*.
- The reason is that the normal light bandwidth occupies a single harmonic interval between ultraviolet and infrared.
- Hence its difference frequencies also fill such an interval. For a set of n frequencies, the number *d* of difference frequencies is $d >> n$.
- Thus the visible light spectrum represents "jamming" of specific internal EM signal inputs through the dielectric pathway into the body, turning much of the environmental signals into harmless warmth.

* It has also been shown that if bacteria are killed in the "dark" by UV, then 24 hours later placed in sunlight, a substantial fraction of the "dead" bacteria will revive.

Ramifications of Kaznacheyev's experimental results

- At least one harmonic interval is necessary for the effect to be evidenced in the target cells.
- Interestingly, if we assume the body dielectric to be isotropically nonlinear (to first order):
  - Velocities of the actual transmitted frequencies depend upon the particular point on the wave amplitude.
  - The transmitted frequencies overshoot, interfere, breakup, etc. The difference frequency does not.
  - The "difference frequency" is transmitted through the nonlinear dielectric medium as if it were a sine wave passing through a linear medium*.
- To our knowledge, this direct and deep communication of electronic "hash" difference frequencies in dielectric human bodies has not been investigated in the West.

* Owen Flynn, "Parametric arrays: A new concept for sonar," Electronic Warfare Magazine, June 1977, p. 107-112.

Ramifications of Kaznacheyev's experimental results (2)

- A heterodyne signal -- at the difference frequency between an input signal and a reference signal -- can be enhanced by adding noise*.
- Thus the signal of the disorder in the cellular pattern emitted from diseased cells, received by normal cells, can be amplified electronically and rather easily.
- Powerline radiation (electronic smog) thus carries its own amplification mechanism. Eventually the amplification overcomes any squelching by sunlight, etc.
- The Russians promptly (as early as the late 1950s) weaponized and tested these and similar effects.

* Dykman, M. I. et al., "Noise-enhanced heterodyning in bistable systems," Physical Review E, 49(3), Mar. 1994, p. 1935-1942

Weaponization of Kaznacheyev's findings

And its ramifications

"Microwave" radiation of the U.S. Embassy in Moscow.

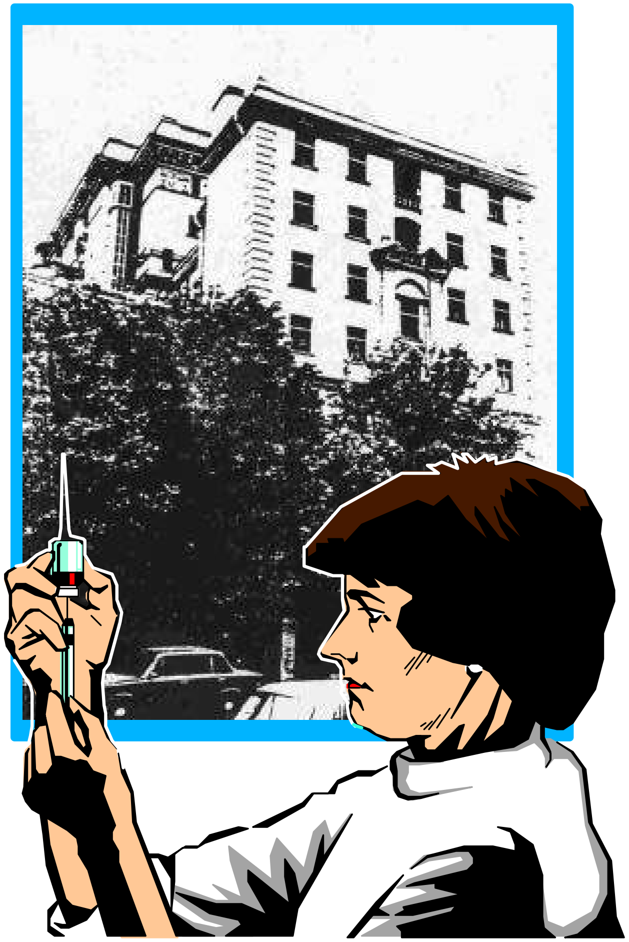

Former U.S. Embassy in Moscow.

BACKGROUND

- **Began in latter 1950s**
- **Discovered on VP Nixon's trip**
- **Initially thought to be *nuclear* radiation (Discovered w/Geiger counter)**
- **High level target -- U.S. Ambassador**
- **Guarantees personal attention of:**
  - **U.S. Ambassador to USSR**
  - **U.S. President**
  - **NSA, CIA, DIA, NSC, etc.**
  - **Top consulting scientists**
  - **Leading U.S. scientific institutions**
- **Two U.S. Ambassadors died, another sickened**
- **Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!**
- **Four U.S. Presidents requested Soviets cease**
  - **Cut from 18 watts/sq cm to 2**
  - **Then again increased**
- **No one could understand what was going on**
- **Aluminum screens were placed over windows**
- **Moscow was declared a hazardous duty zone**

Ramifications of the microwave radiation of the U.S. Embassy

- The cellular disorder pattern in the EM radiation is carried by the field-free scalar potential.
- Since the disorder pattern is a "conglomerate", it must be carried by an internal EM structure existing inside the scalar potential itself.
- Whittaker decomposed the scalar potential into the appropriate internal longitudinal EM biwave structure in 1903*.

* E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355

The importance of time as energy, and the time-domain

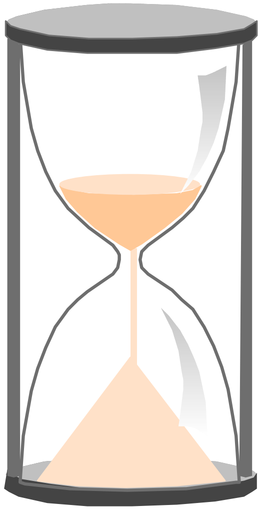

Time as energy

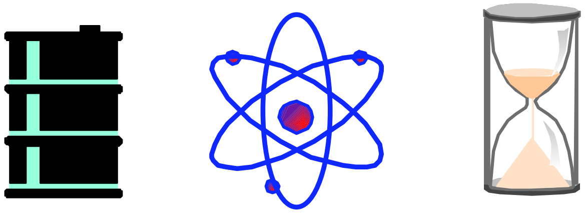

- In physics, the fundamental units one uses in his model are arbitrary.
  - A valid model can be made using only one unit.
  - It is already done in one branch of physics, using length as the single fundamental unit.
- Make a model with a single unit, the joule.
  - All other entities become functions of the joule.
  - We are familiar with mass as highly compressed EM energy, by the factor c^2.
  - It turns out that time is also highly compressed EM energy, by the factor c^2.
  - 1 sec. $\approx 9 \times 10^{16}$ joules.

How to see that "time is compressed spatial energy"

- Take some spatial EM energy and compress it by c^2
- Place it in 3-space, and it is *mass*.
- Place it on the time axis, and it is *time*.

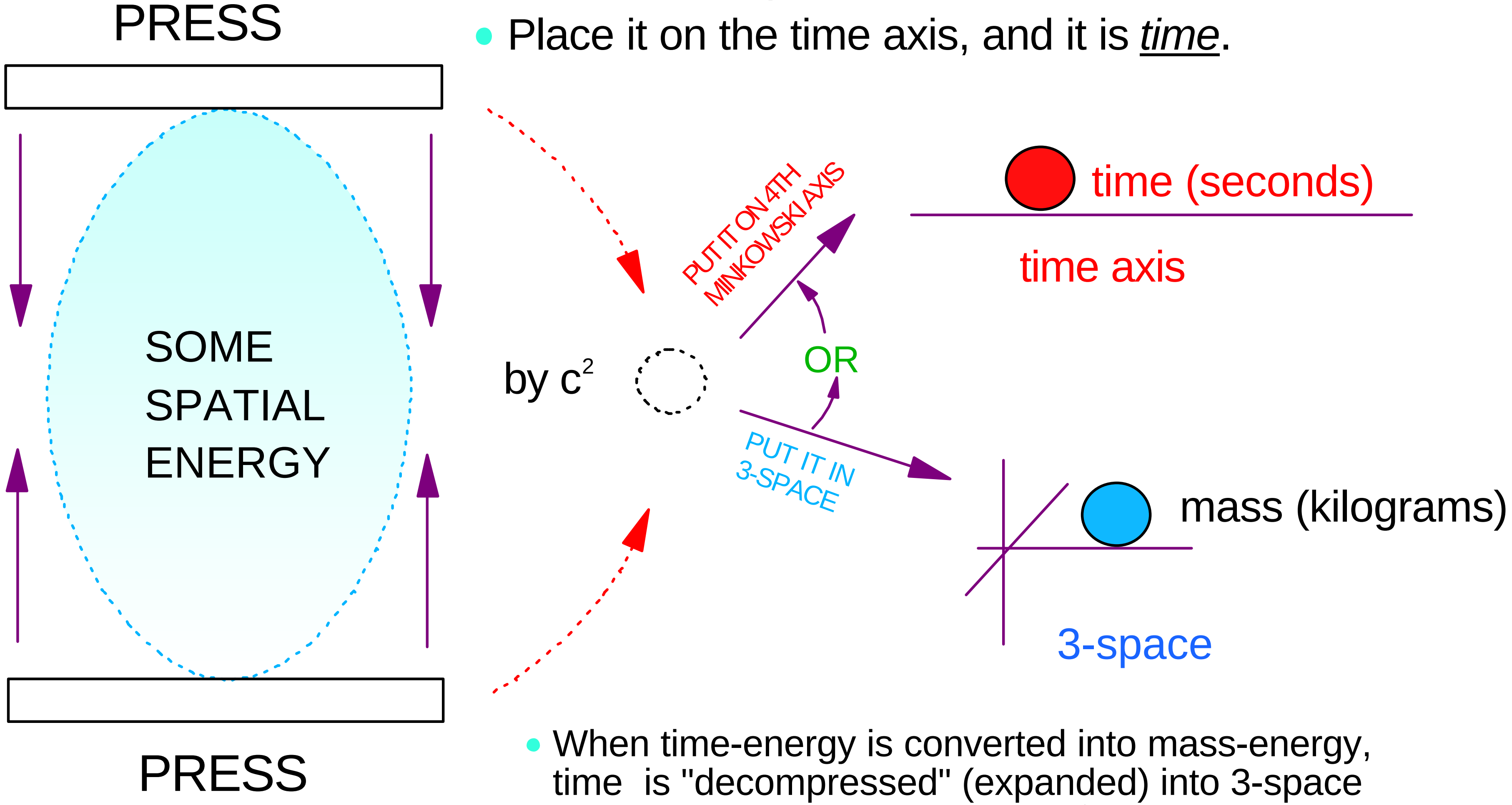

PRESS

SOME
SPATIAL
ENERGY

PRESS

by c^2

PUT IT ON 4TH
MINKOWSKI AXIS

OR

PUT IT IN
3-SPACE

time (seconds)

time axis

mass (kilograms)

3-space

- When time-energy is converted into mass-energy, time is "decompressed" (expanded) into 3-space EM energy and 1 sec. $\approx 9 \times 10^{16}$ joules.

Time as EM energy in the time domain, with a density the same as mass-energy

- Bearden* reinterpreted (corrected) the 1903 Whittaker decomposition of the potential in 2000.
- The reinterpretation is consistent with broken symmetry in particle physics, for which Lee and Yang received the Nobel Prize in 1957.
- It is also powerfully supported by quantum field theory, notably by Mandl and Shaw in 1984**.
- Whittaker's pair is a combined time-polarized (scalar) EM wave and a 3-space longitudinal EM wave.

---

* T. E. "Giant Negentropy from the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23.

** F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5.

All EM energy in 3-space freely comes from the time domain

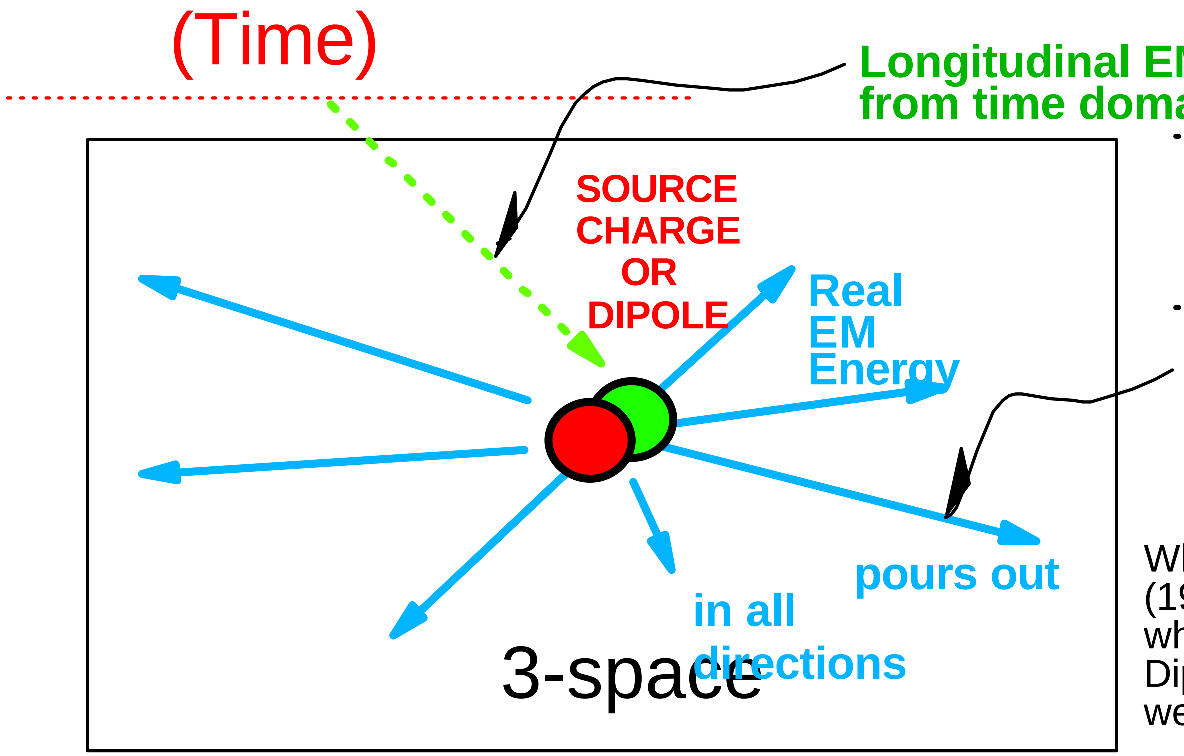

Longitudinal EM wave incoming from time domain and absorbed

Charge's 720º spin transforms time-energy into 3-space energy

Longitudinal EM waves emitted in all directions in 3-space

Whittaker, Math. Ann., 57, 333 (1903) shows this rigorously, when properly reinterpreted. Dipole's broken 3-symmetry is well-known in particle physics

Note: Whittaker (and others) interpreted the phase conjugate half set of LWs after interaction with the charges of the dipole, and as a 3-space *effect* rather than the time-domain *cause*. This fundamental non sequitur was just repeated since then, until corrected by Bearden to be consistent with broken symmetry and quantum field theory.

Strong support from quantum field theory

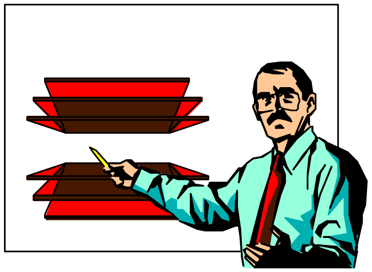

- Quantum field theory recognizes four photon polarizations:
  - Energy along x-, y-, z-, or t- axis.
  - Polarized along x- and/or y-axis, the photon is a *transverse* photon.
  - Polarized along z-axis (along its line of motion in 3-space), the photon is a *longitudinal* photon.
  - Polarized along the time-axis, the photon is a *time-polarized* or *scalar* photon.
- The individual longitudinal or scalar photon is not observable*.
- The combination of a longitudinal photon and a scalar photon is observable as the instantaneous scalar potential*.

* F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5.

Mechanism generating the flow of a mass through time

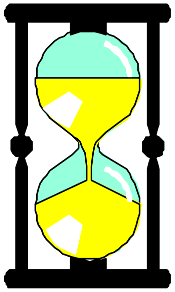

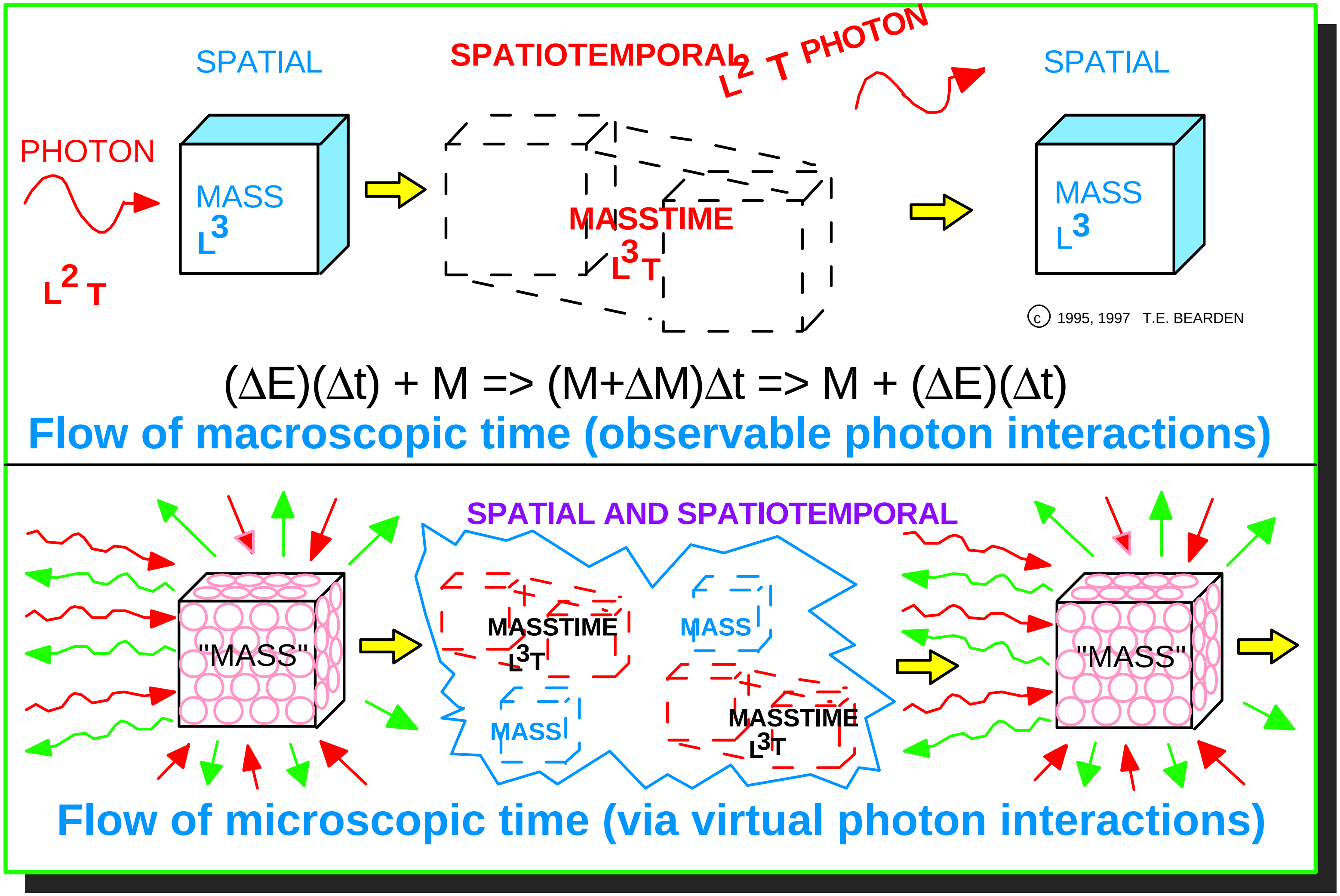

SPATIAL

SPATIOTEMPORAL

L^2T PHOTON

SPATIAL

PHOTON

L^2T

MASS L^3

MASSTIME L^3T

MASS L^3

© 1995, 1997 T.E. BEARDEN

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

SPATIAL AND SPATIOTEMPORAL

"MASS"

MASSTIME L^3T

MASS

MASS

MASSTIME L^3T

"MASS"

Flow of microscopic time (via virtual photon interactions)

Mechanism is engineerable

Pumping a mass in the time-domain will time-reverse it back to an earlier state

- The reinterpreted biwave EM structure of the scalar potential means that time-reversal (phase conjugation by pumping in the time-domain) can be utilized to time-reverse a targeted cell (its mass-energy) in its entirety, including its genetics.
- Becker's bone fracture experiments showed it:
  - Weak scalar potential across the fracture site.
  - LWs in 3-space and time-energy waves in time
  - Red blood cells entered, dedifferentiated by shucking their hemoglobin and growing a nucleus, which is dedifferentiation (time-reversal of the cell and its mass-energy back to a previous physical state).

Principle for time-reversing a time-pumped mass

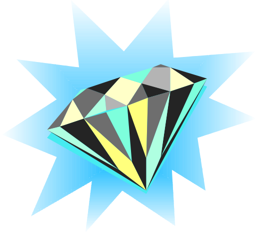

- Becker's cells then redifferentiated and changed -- first to primitive cartilage cells, then to primitive bone cells which were deposited in the fracture, healing it with new bone growth.
- This shows "time-forwarding" (redifferentiation) of the cell also.
- Becker actually showed that the *time-domain pumping* of the cells (by the time-polarized EM waves of Whittaker's reinterpreted wavepairs comprising the scalar potential), and the time-pumping of their mass-energy, operated to continuously and steadily *remove the delta between that cell's present condition and its former "normal" cellular physical condition in that area*.
- This is the long-sought secret of stem-cell research.
- Phase conjugate optics had not been born at the time. Without realizing its full nature, Becker demonstrated the body's own internal mechanism for healing, and the fundamental mechanism for time-reversing a physical mass, whether living or inert.

Mechanisms for use in medical treatment and healing

And ramifications

Successive deep penetration reactions of the cellular dielectric

- Change after change affects the cell and all of its parts, including from the external environment.
- Each successive change affects primarily the internal "LW EM structures and dynamics" comprising the ordinary potentials, fields, and waves.
- Thus in positive time the internal EM structurings produce successive "overlaid layers of change" -- so to speak -- in the internal EM structuring and dynamics of a cell in the body. These layers contain the "delta" from normal state.

Successive deep penetration reactions of the cellular dielectric (2)

- During time reversal, highest layer deltas are successively "peeled away" by the interaction of the mass-energy with the time-domain pumping.
- The result is to return the cell directly back to successive physical states, in the case of a single disease or disorder.
  - A cancer cell is returned to a normal cell prior to its "promotion".
  - An HIV-infected cell would be returned to its normal unaffected condition, free of HIV-induced genetic change.
  - For injuries and damage, the existing "delta" in each successive state-layer is steadily reduced to zero.
- This process thus can involve both cellular differentiation and dedifferentiation, as shown by Becker and by Prioré.

Engines (vacuum engines, spacetime curvature engines)

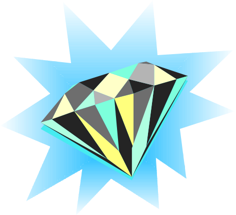

- Following Whittaker* as reinterpreted**, all normal EM fields, potentials, and waves decompose into sets of longitudinal EM waves (LWs) and their dynamics.
- Since these LWs and their dynamics are dynamics in the local energy density of spacetime (ST), they constitute *ST curvatures*.
- Hence they are general relativity infolded inside electrodynamics.
- By manipulating and changing these LWs, one manipulates and changes spacetime curvatures and their dynamics.
- The result is a unified field model where one directly controls and engineers a set of ST curvatures by use of higher group symmetry electrodynamics.
- Controlled patterns of local curvatures of ST are called "vacuum engines" or "spacetime curvature engines" -- or just "engines".

* E.T. Whittaker, Math. Ann., 57, 333-355 (1903); Proc. Lond. Math. Soc., Series 2, 1, 1367-372 (1904).

** T. E. Bearden, J. New Energy, 5(1), 11-23 (2000).

Pumping with EM time-polarized waves forms *amplified* antiengines

• •

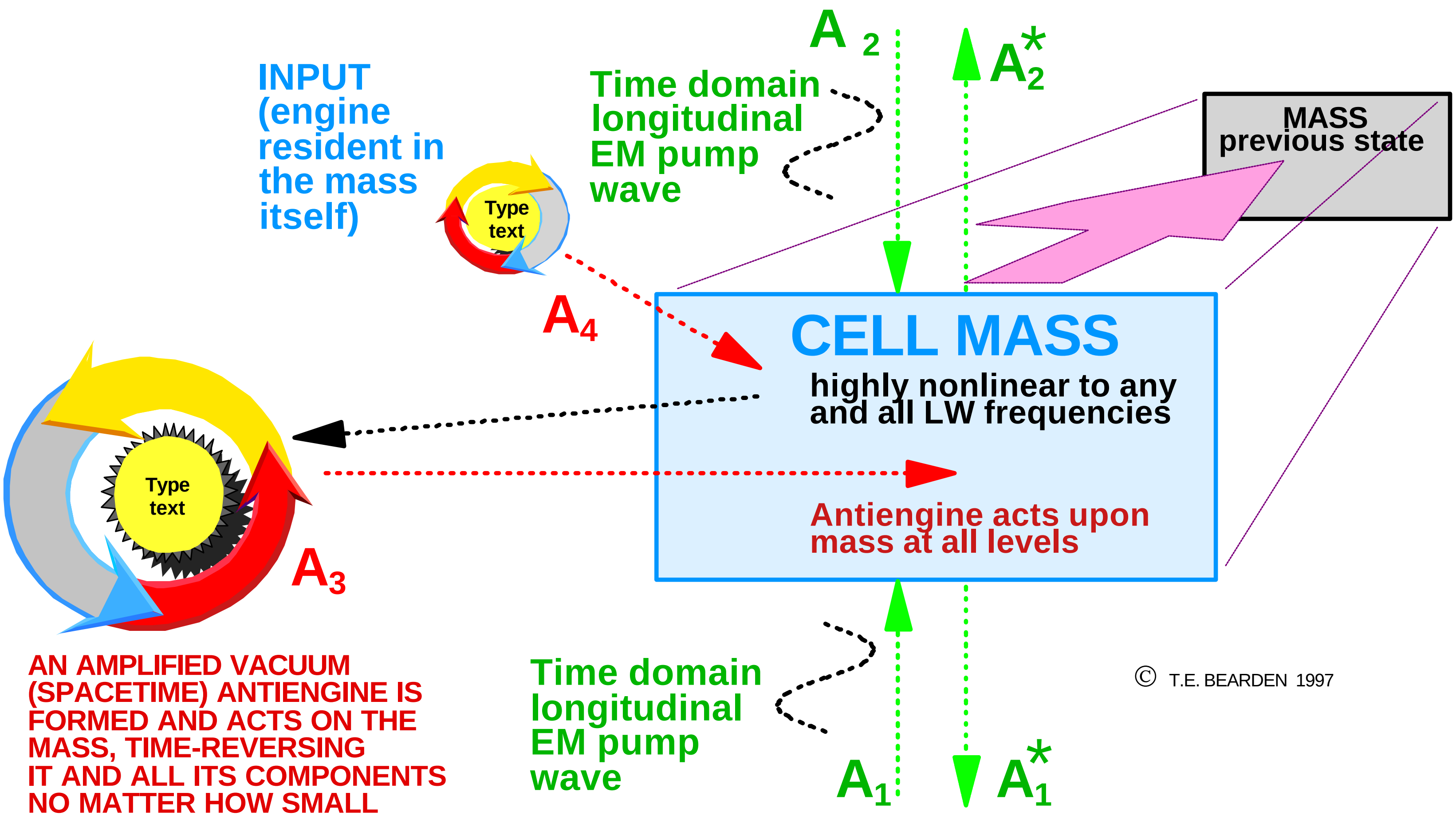

INPUT (engine resident in the mass itself)

Time domain longitudinal EM pump wave

MASS previous state

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

Time domain longitudinal EM pump wave

© T.E. BEARDEN 1997

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

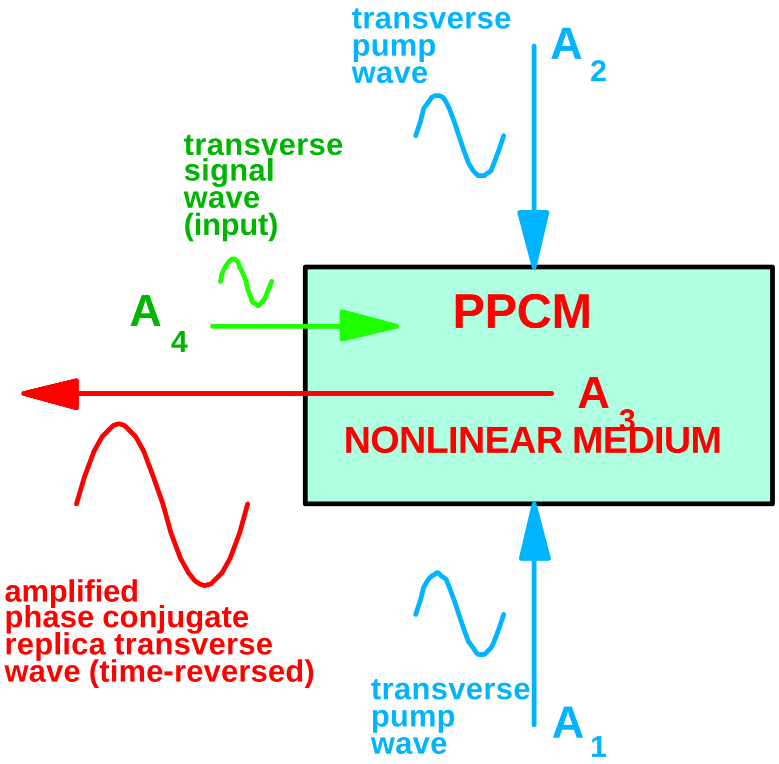

transverse pump wave

A_2

transverse signal wave (input)

A_4

PPCM

A_3

NONLINEAR MEDIUM

amplified phase conjugate replica transverse wave (time-reversed)

transverse pump wave

A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

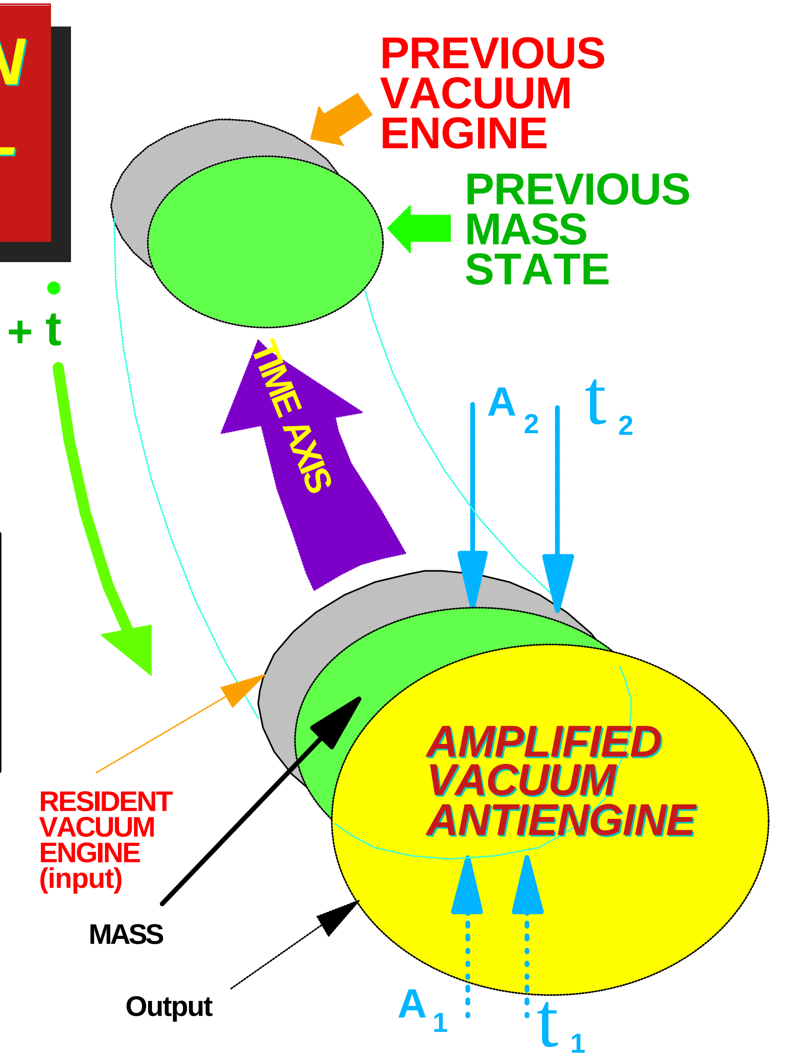

PREVIOUS VACUUM ENGINE

PREVIOUS MASS STATE

+ $\dot{t}$

TIME AXIS

A_2 t_2

RESIDENT VACUUM ENGINE (input)

MASS

Output

AMPLIFIED VACUUM ANTIENGINE

A_1 t_1

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Time-reversal of a mass to a new physical state not in its past

- Original engine A'
- New engine B' added
- Time-pumping moves to new "past" that is determined by (A'+B')

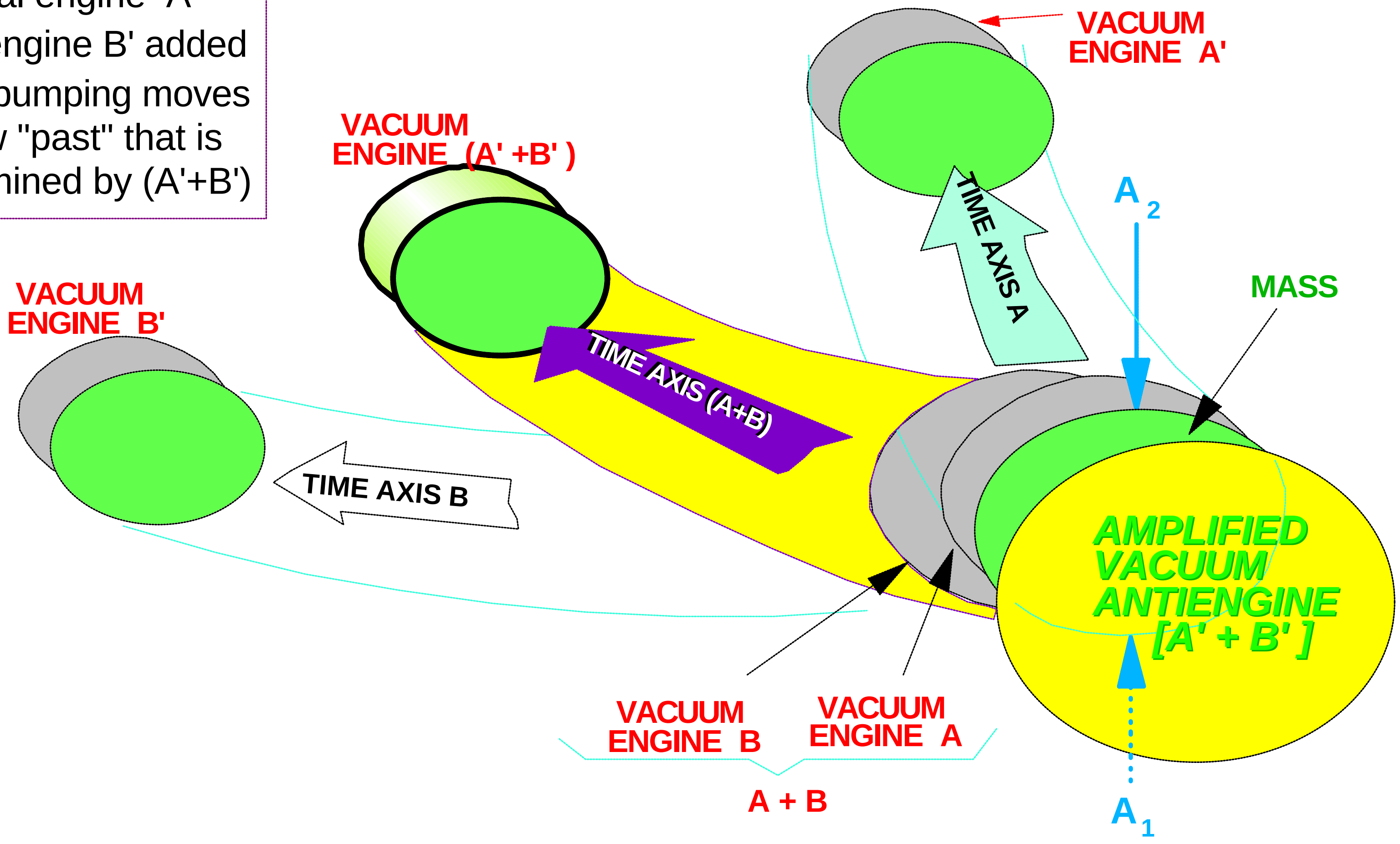

VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
TIME AXIS (A+B)
TIME AXIS B
A_2
MASS
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A + B
A_1

Extended general relativity principle

- All levels of energy structures and all levels of time structures mold spacetime geometry.
- The "pattern" is called a *template* for a vacuum engine (spacetime curvature engine).
- Spatial and temporal structurings of the engine act upon any exposed mass at al levels.
- This produces a template of forces, for precise translations and stresses within the mass at all levels. The forces exist in both time and 3-space.

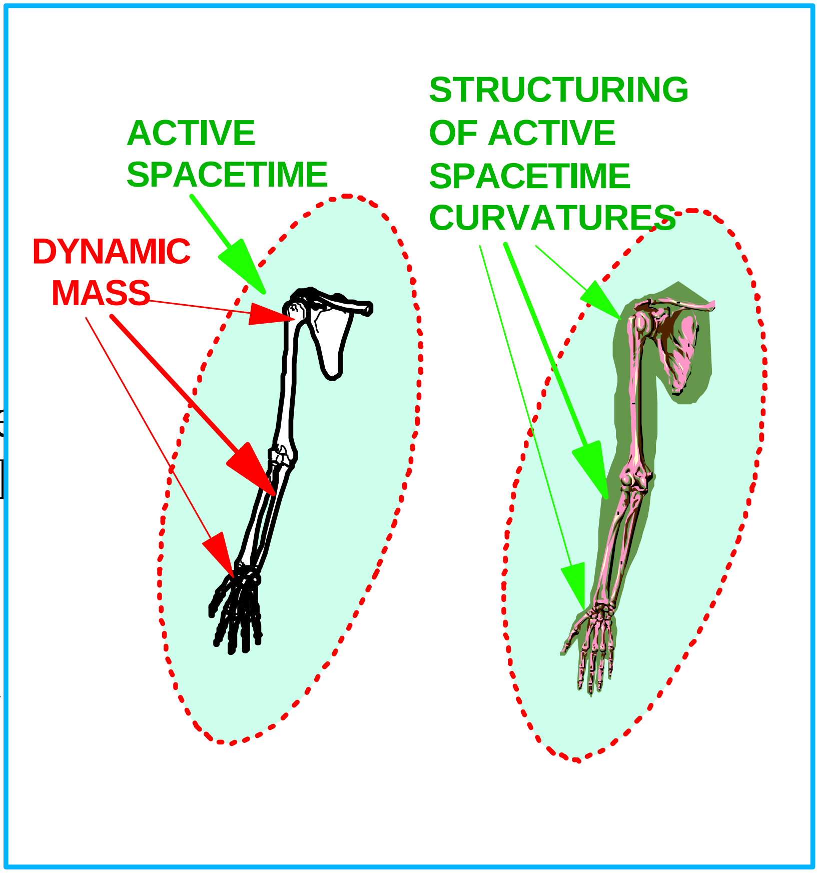
ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

Engines: Working spacetime demons

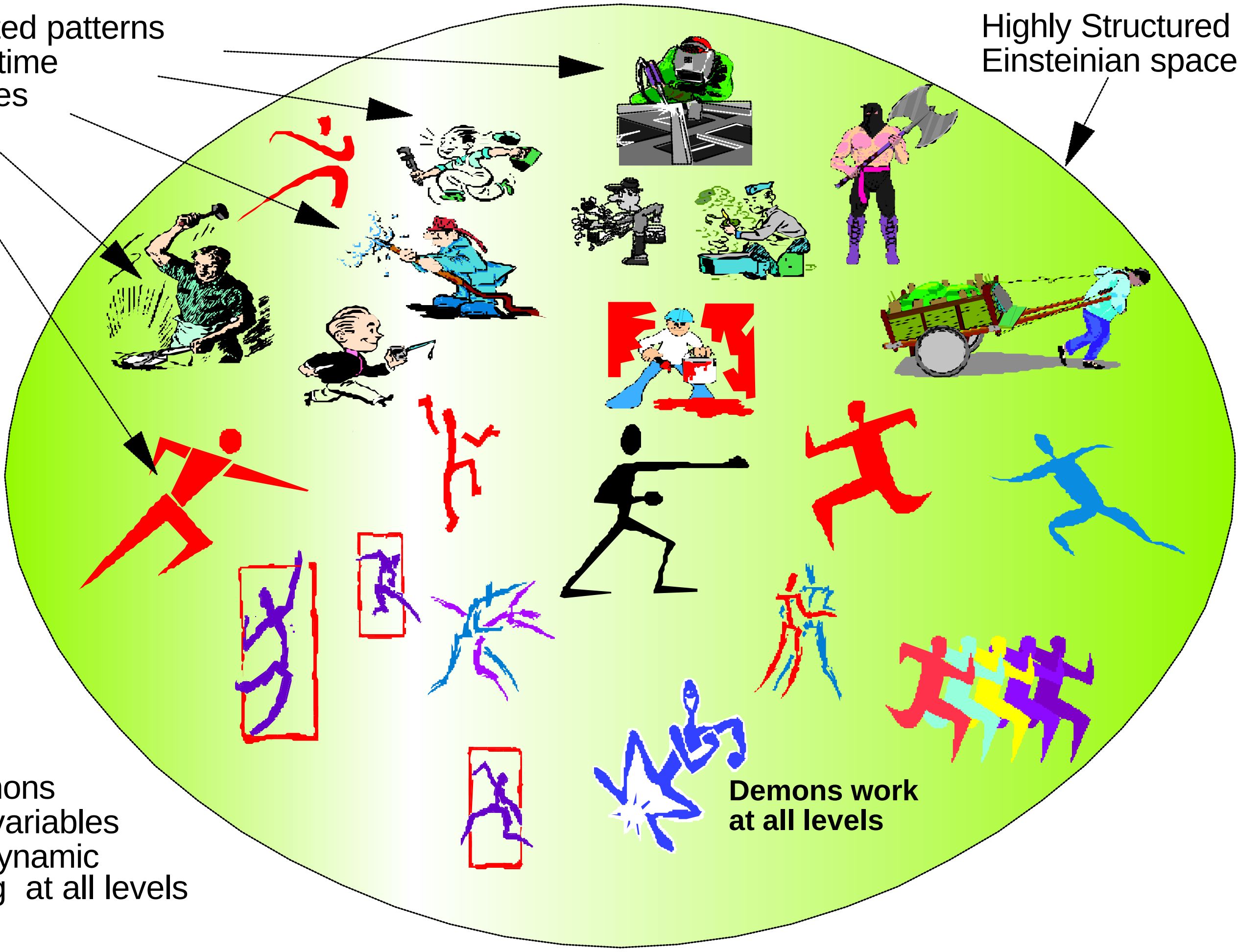

Internested patterns of spacetime curvatures

Highly Structured Einsteinian spacetime

Maxwellian Demons are EM hidden variables organized into dynamic structures acting at all levels

Demons work at all levels

Signal vs. vacuum engine

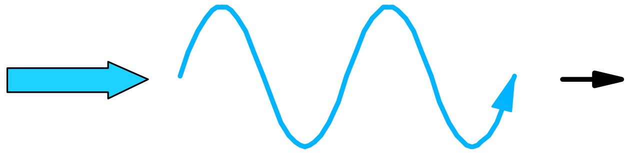

- ★ Information received
- ★ Any overt physical action must be taken by cell itself
- ★ Cell must furnish the energy for any action it takes
- ★ Passes from outside to in

A. Cell must do the action itself; vacuum energy exchange is passive.
Energy or fuel for doing the action must be added to cell externally.

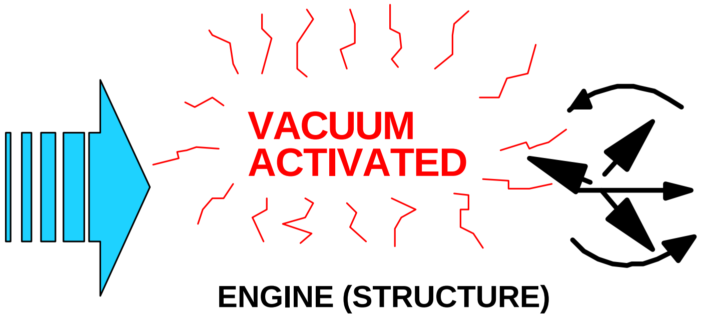

- ★ Information itself acts
- ★ Local spacetime curved, with internal subcurvatures
- ★ Produces template of forces in local spacetime (ST)
- ★ Arises from within local ST
- ★ Acts on and changes all parts of cell, including its genetics

B. Cell is acted upon; vacuum energy exchange does the specific action.
No energy or fuel for doing the action need be added to cell externally.

Unified field theory requires modeling and use of supersystem

Three components interact with each other:

- System and its dynamics
- Local active vacuum and its dynamics
- Local curvatures of spacetime and their dynamics.

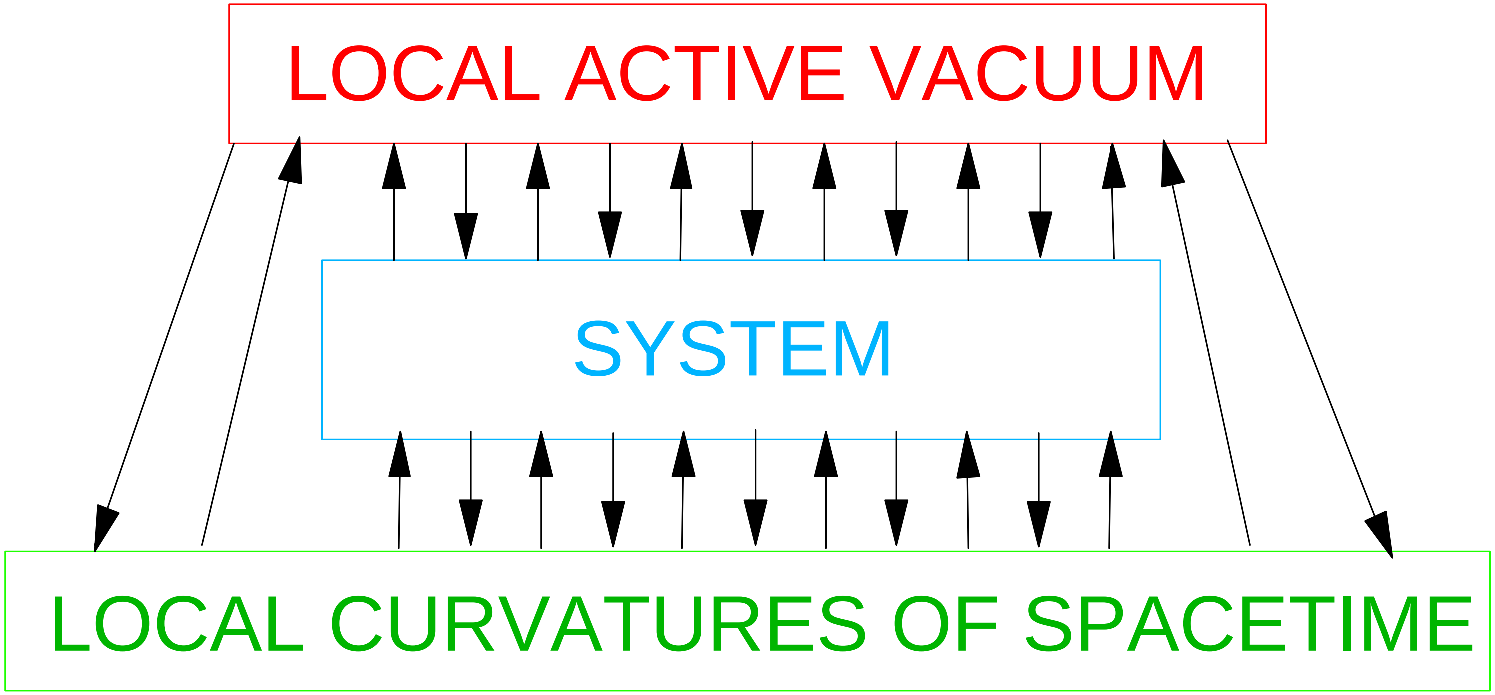

LOCAL ACTIVE VACUUM

SYSTEM

LOCAL CURVATURES OF SPACETIME

Classical EM kills the supersystem, prevents supersystem engineering

- Assumes flat local spacetime
- Assumes inert vacuum

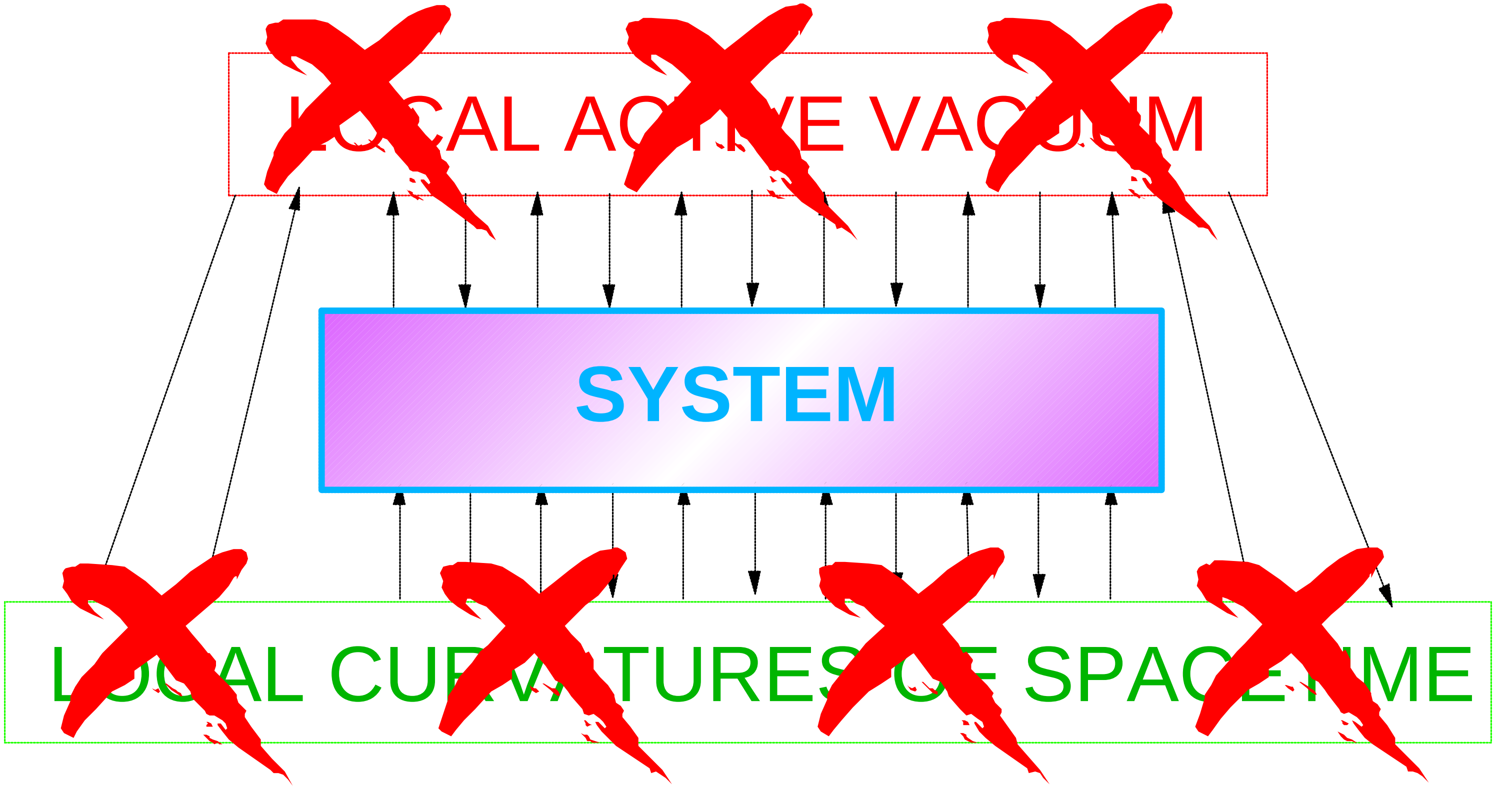

LOCAL ACTIVE VACUUM

SYSTEM

LOCAL CURVATURES OF SPACETIME

Malignant teratoma exhibits both time-reversal and fast-forwarding

- A teratoma cell is a tumor cell that has reverted to stem cells, then proceeds to differentiate toward new kinds of normal cells.
- May produce hair, skin, muscle, heart, bone, and other type cells.
- Sometimes present at birth.
- Demonstrates both "time-reversal" and "time-forwarding".
- The missing key is the engine function(s).

Miller-Fox-Urey biogenesis experiments

- Engines of and from all life forms ever on earth or in the universe exist in the ambient vacuum potential.
- The Miller-Fox-Urey experiments were not in a "sterile" environment.
- Heat (IR) and UV constitute (1) a harmonic interval with a difference frequency, and (2) amplifying by the process previously shown.
- These experiments "kindled" previously living forms from the available gases etc. by amplifying the ancient living engines from their virtual state to the observable state, so that the masses were molded into the physical states of their templates.

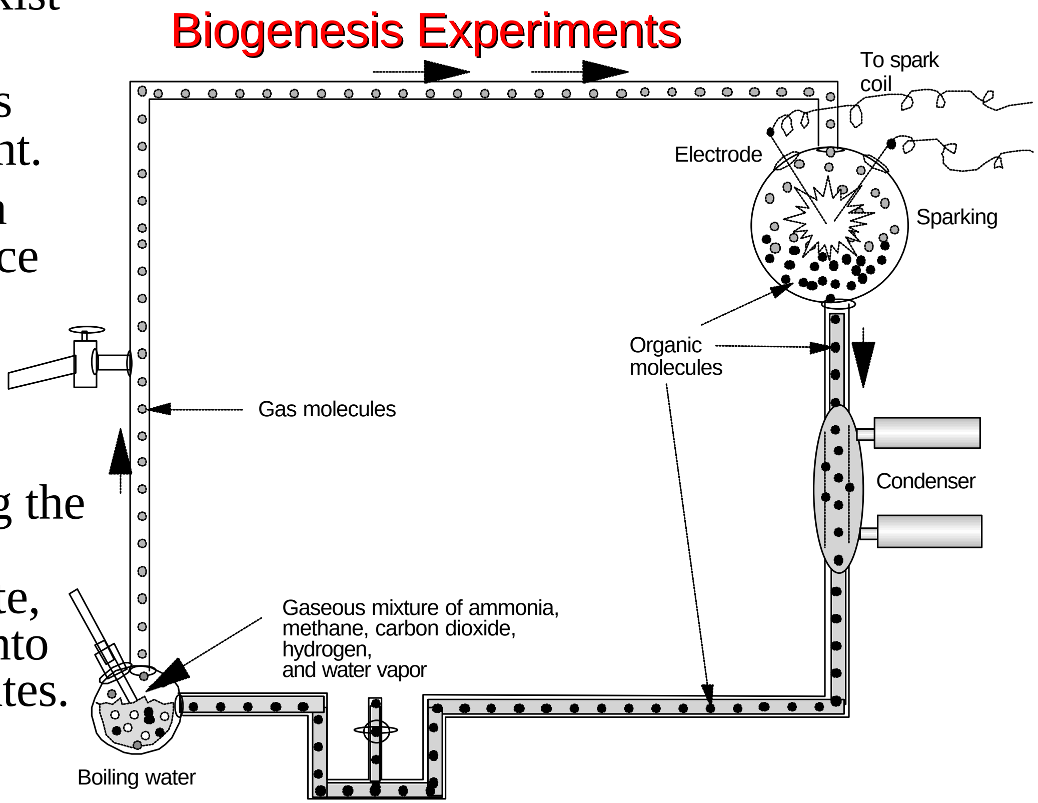
Biogenesis Experiments
To spark coil
Electrode
Sparking
Organic molecules
Gas molecules
Condenser
Gaseous mixture of ammonia, methane, carbon dioxide, hydrogen, and water vapor
Boiling water

Cloning to produce stem cells

- Human DNA can be inserted in an animal cell, and the cell chemically tricked into proceeding as if fertilized.
- Formation, increase, and growth of stem cells begins and continues.

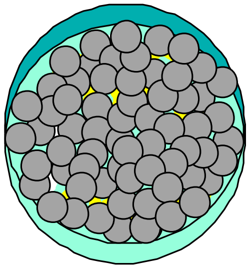

- The "engine" carried by the human DNA is changing the matter available into human cells.
- readily demonstrated experimentally.

Regeneration in a strain of rat with part of its immune system missing

- In mammals, the immune system plays a role in reducing the regeneration capability, compared to that of more primitive systems.
- In a strain of rat where that "suppressive" part of the immune system is missing, the regenerative capability is dramatically increased.
  - Can regrow cut away sections of tail.
  - Regrows plug cut out of ear.
  - Reconnects severed optic nerve.
  - Partially restores severed spinal cord.

- The suppressive "delta" engine of the immune system, operating simultaneously with the underlying primitive regenerative engine, sums to the modern much-reduced regenerative engine in the mammal.

An observation for the future

- When the science of engines is developed, physicians will be able to tailor these engine combinations to restore severed human spines, regrow severed limbs, etc.
- They will also be able to direct the body's "normal growth" to reduce and direct birth defects, genetic defects, etc.

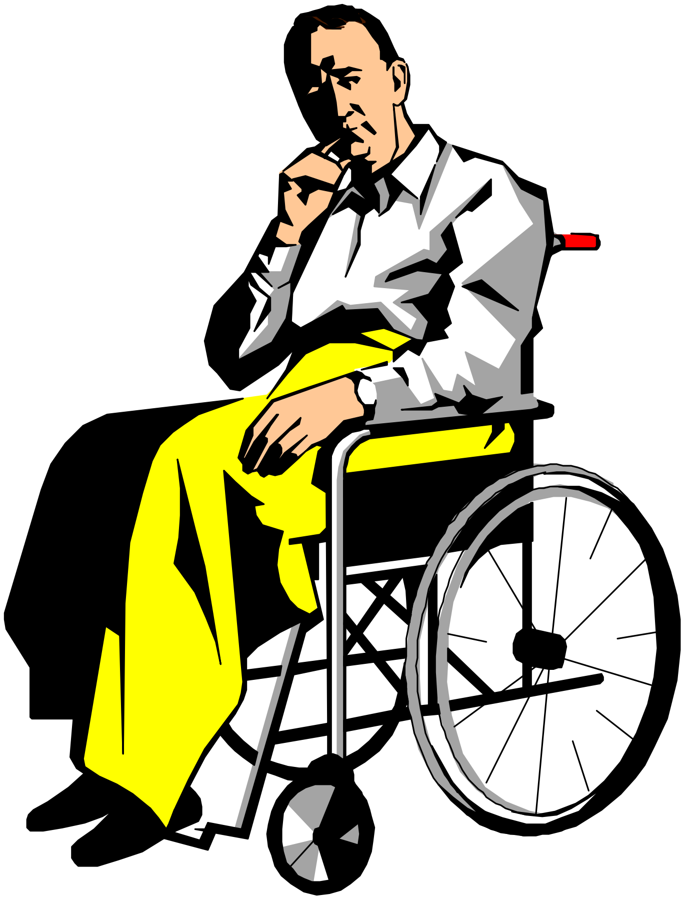

Biochemistry and EM models

- Medical scientists analyze such things as teratomas in terms of biochemistry and the biochemical model.
- An assembly of charges and charge distributions, with their associated fields and potentials, drives all chemistry at base level.
- The charge assembly and its dynamics involves a corresponding assembly of spacetime curvatures and their dynamics. These are *the cause that is driving the charge assembly and its dynamics, and therefore the chemistry*..
  - In unified field theory (such as that of Sachs) and O(3) electrodynamics.
  - Not in classical U(1) EM theory or quantum electrodynamics.
- Conventionally a causative EM field in space is "defined" as the *diversion from* that spatial field by and after its interaction with charge).
  The "definition" substitutes the effect for the cause, a non sequitur.
- This is a known major flaw, pointed out in one fashion or another by many physicists such as Wheeler and Nobelist Feynman.
- Classical EM theory is a 137-year-old model, seriously flawed. These flaws emerge in biochemistry and medical science.

Necessary EM corrections*

- Many of the corrections necessary in electrodynamics modeling either have been done or are underway.
- Higher symmetry, non-Abelian electrodynamics models have long been developed in particle physics.
- The new approach extends these models to also incorporate general relativity, thus producing a unified field theory*.
- In the new approach, spacetime curvature engines are perfectly natural and understandable, and can be directly modeled and engineered by advanced theorists.
- In short, the theory is sufficiently advanced and robust that technology and engineering can now advance.
- This will be a great leap forward for biochemistry and medical science, particularly in therapeutic methods.

* Sachs's unified field theory implemented by O(3) electrodynamics presently is the most suitable to use. Numerical methods are required.

Prioré's revolutionary results in lab animals

And its ramifications

Prioré (right) makes a point to Chaban-Delmas, Mayor of Bordeaux

- Many of the Prioré team's experiments were done in Bordeaux.
- Antoine Prioré patents:
  - "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968.
  - "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966.
  - "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963.

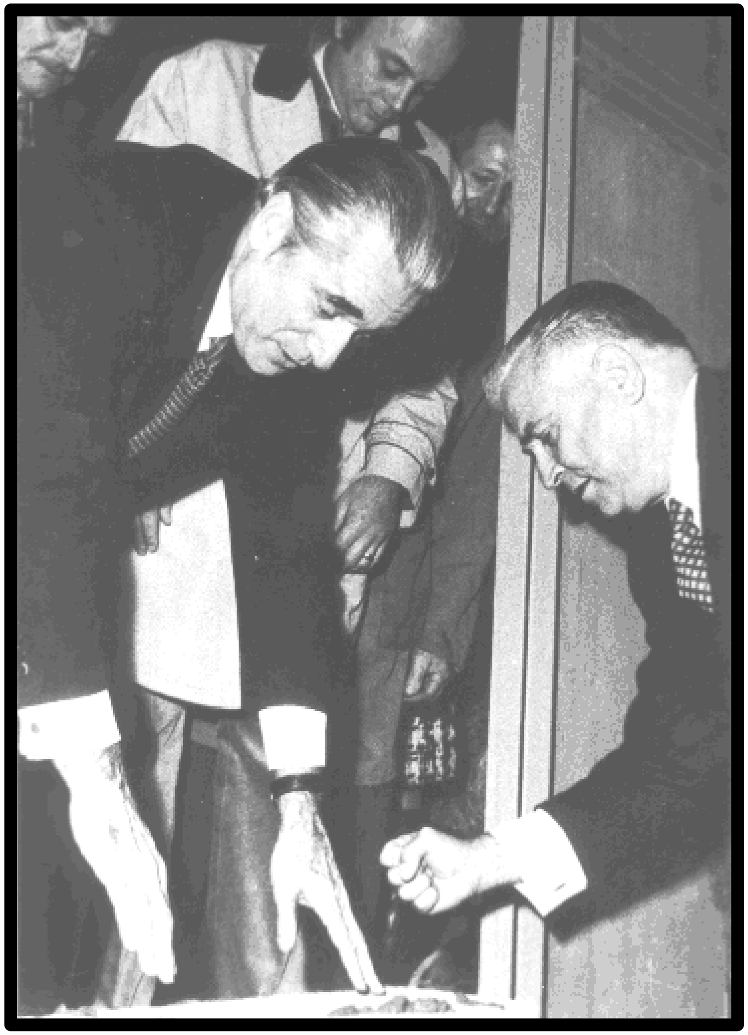

Block diagram of Prioré's method

Time-reversing the cells back to normal state

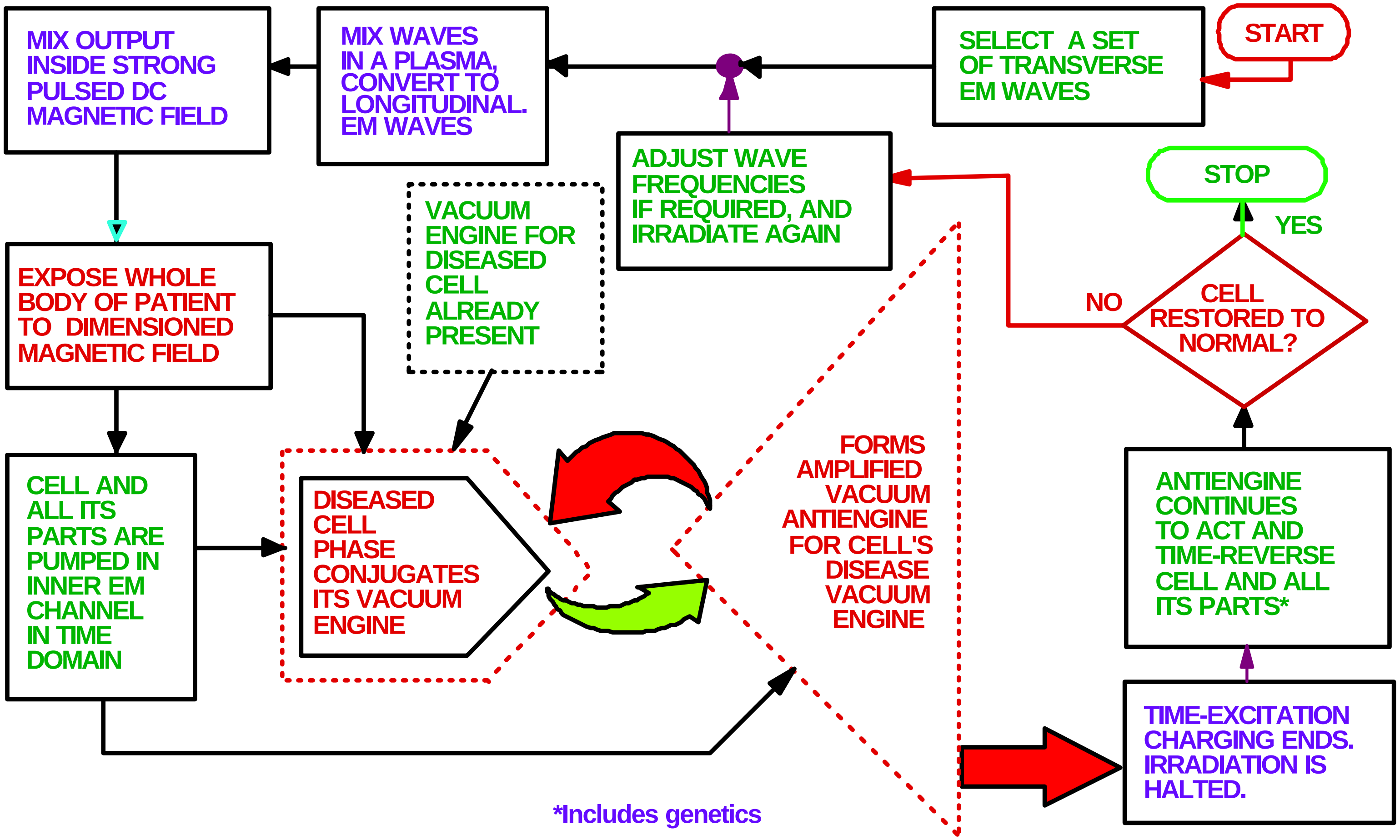

START

SELECT A SET OF TRANSVERSE EM WAVES

MIX WAVES IN A PLASMA, CONVERT TO LONGITUDINAL. EM WAVES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.

ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*

CELL RESTORED TO NORMAL?

YES

STOP

NO

ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN

*Includes genetics

In Laboratory Animals:

Prioré Team Cured

- **Terminal malignant tumors**
- **Clogged arteries (atheriosclerosis)**
- **Suppressed Immune Systems**
- **Infectious trypanosomas**
- ***Results were presented to French Academy of Science by Robert Courrier, Director of the Biology Section of the Academy***

© 1999 T. E. BEARDEN

- ***No one understood the mechanism!***

Some peculiar results by the Prioré team

- Once a rat is cured of a particular disease:
  - A single drop of blood from the cured rat, injected into another rat stricken with same disorder, would result in the injected rat getting well slowly.
  - This ability to transfer immunity and healing in the extracted drop of blood gradually decayed over a period of a few weeks.
- Showed that even the fields and potentials in the blood had been altered internally, in the treated animal.
- Showed that the injected animal's cellular regenerative system possessed the ability to also time-domain pump over the entire body, everywhere reducing the delta between that injected drop of blood now the rat's own, and the rest of the blood in its body.

Proof of cellular time-reversal

- Pautrizel compared identical treatment of immature rats with immature immune systems, and mature rats with mature immune systems.
- Treatment restored the damaged mature system back to full mature functioning, and it promptly recognized and dispatched the pathogens.
- Treatment restored the damaged immature system back to full immature functioning, unable to cope, and the pathogens promptly reinfected and killed the immature rat.
- The lesson: The body is indeed returned to a previous physical state. The pathogens themselves are not killed by the treatment, but by restoring the ability of the immune system to recognize and destroy them.
- Only the *previous ability* of the immune system is restored by the process.

PROF. RAYMOND PAUTRIZEL
(Renowned Parasitologist)

Disadvantages of Prioré method

- Large plasma tube and facility required.
  - 3-stories high building bay if to treat humans.
  - Long irradiation time (2-3 hours).
  - Slow, numerous experimental adjustments.
- Fundamental mechanism was unknown.
  - Phase conjugate optics not yet born in West.
  - No higher symmetry EM modeling used.
  - Prioré thought it was a result of ionization.
  - Importance of difference frequency unknown.
  - Body dielectric transmission path unknown.

3-Story high Prioré device for treating humans

- At the end of the project, Prioré had developed a large unit to treat humans.
- A few were treated.
- Cures replicated what had been done in lab animals
- The massive size of the unit and long treatment time (hours) made the process bulky and very inconvenient -- and also expensive

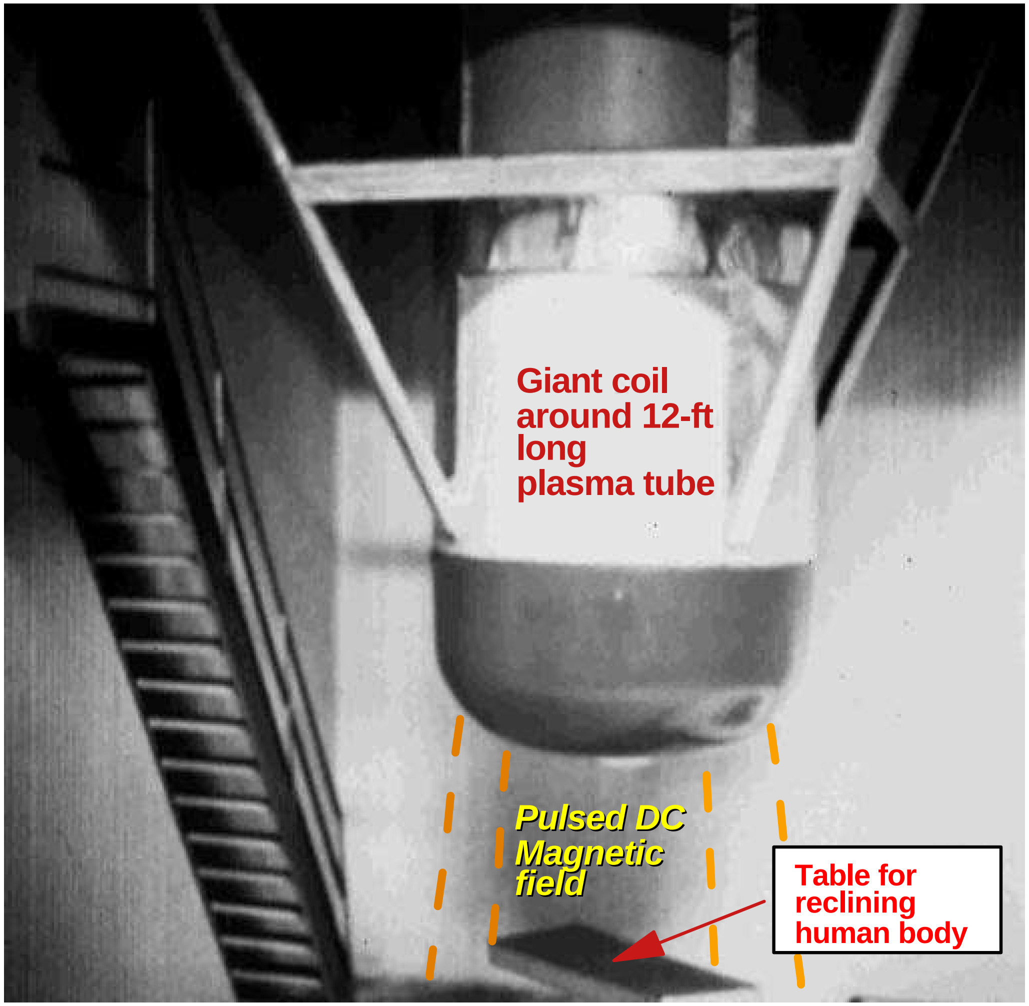
Giant coil around 12-ft long plasma tube
Pulsed DC Magnetic field
Table for reclining human body

17-foot Prioré plasma tube

Used in large machine (3-stories high) to treat a limited number of human patients

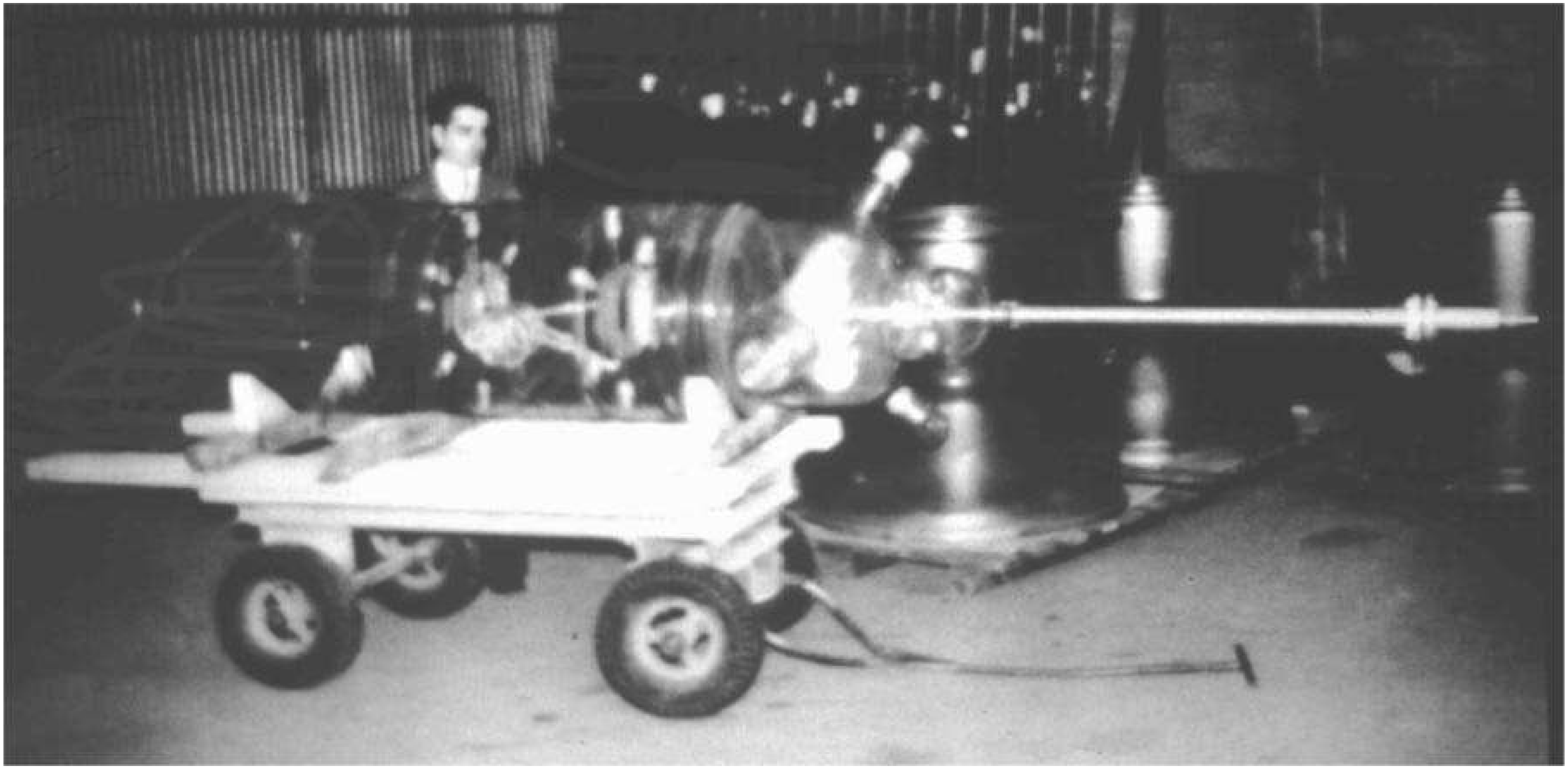

A proposed new mechanism for cancer promotion

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

And its ramifications

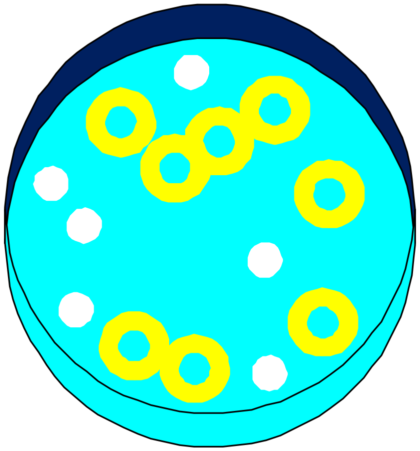

Remarks on cancer from a new perspective

- Two stages: cellular damage and promotion of the damaged cell to a tumorous cell. Promotion presently is poorly understood if at all.
- A master cellular control system (MCS) (studied by Popp et al.) controls both the immune system and the cellular regenerative system, as well as normal functions of the cells.
- The "past history" of ancestral development of aerobic cells and higher multicell aerobic bodies on planet Earth, resides in the "internested engine layers" in every tiniest part of the entire body dielectric, and in all cells.

Remarks on cancer from a new perspective (2)

- Faced with cell damage (such as sustained hypoxia) it cannot handle, the MCS continues to "peel the layers" in taking actions related to ever more distant pasts.
- Faced with continuing hypoxia, e.g., the MCS eventually reaches ancestral layers -- with an engine -- it forces the damaged cells to start moving back toward the original primeval anaerobic form.
- The first step backward is *separation from central growth control*, freeing the cell to multiply independently. At the same time, genetic changes in the cell are forced and produced, starting it back toward becoming an anaerobe.
- These actions comprise the *promotion process*.

The promotion process via engines

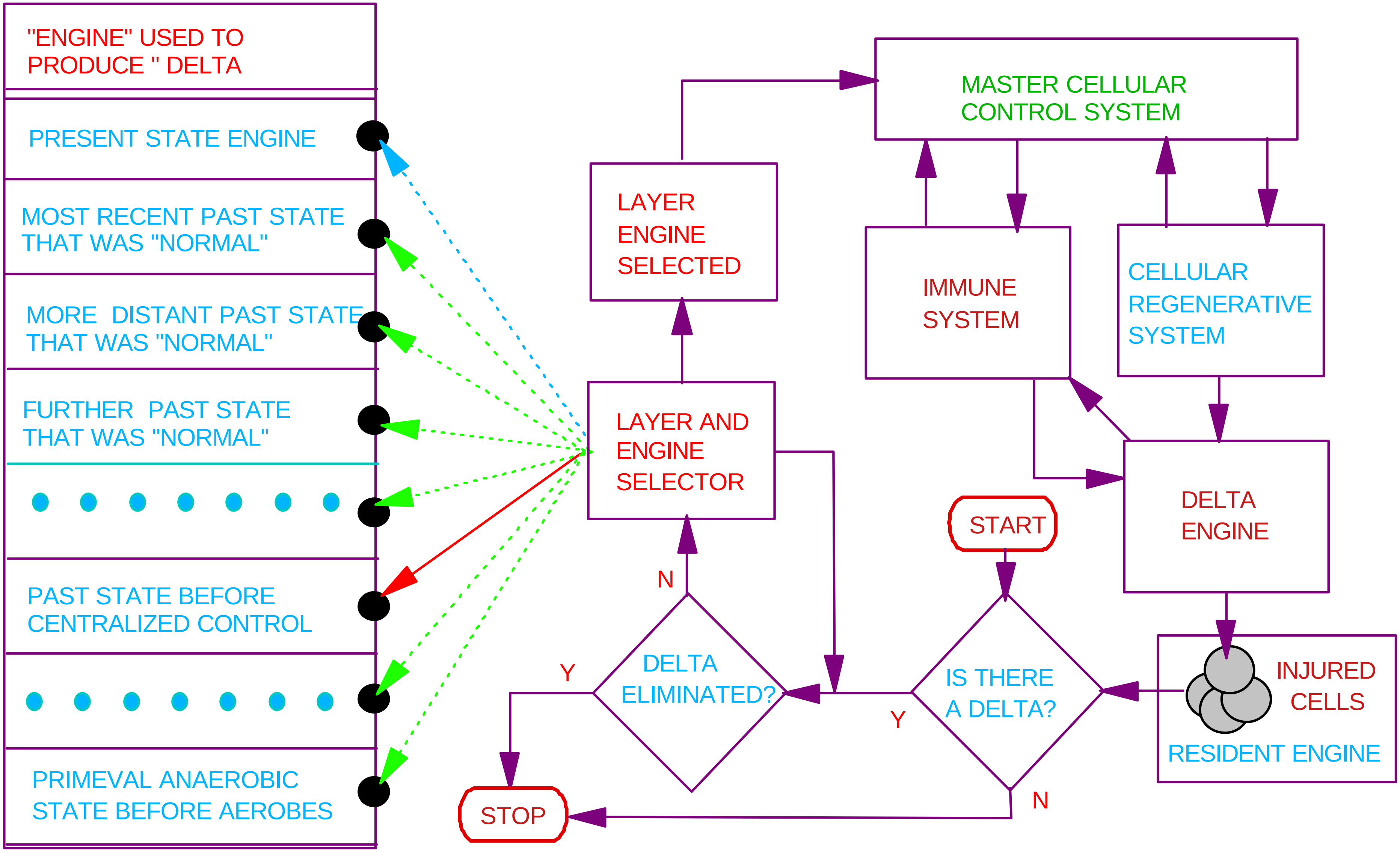

"ENGINE" USED TO
PRODUCE " DELTA

PRESENT STATE ENGINE

MOST RECENT PAST STATE
THAT WAS "NORMAL"

MORE DISTANT PAST STATE
THAT WAS "NORMAL"

FURTHER PAST STATE
THAT WAS "NORMAL"

PAST STATE BEFORE
CENTRALIZED CONTROL

PRIMEVAL ANAEROBIC
STATE BEFORE AEROBES

MASTER CELLULAR
CONTROL SYSTEM

LAYER
ENGINE
SELECTED

IMMUNE
SYSTEM

CELLULAR
REGENERATIVE
SYSTEM

LAYER AND
ENGINE
SELECTOR

DELTA
ENGINE

START

N

Y

DELTA
ELIMINATED?

Y

IS THERE
A DELTA?

INJURED
CELLS

RESIDENT ENGINE

N

STOP

Evolution of aerobic systems

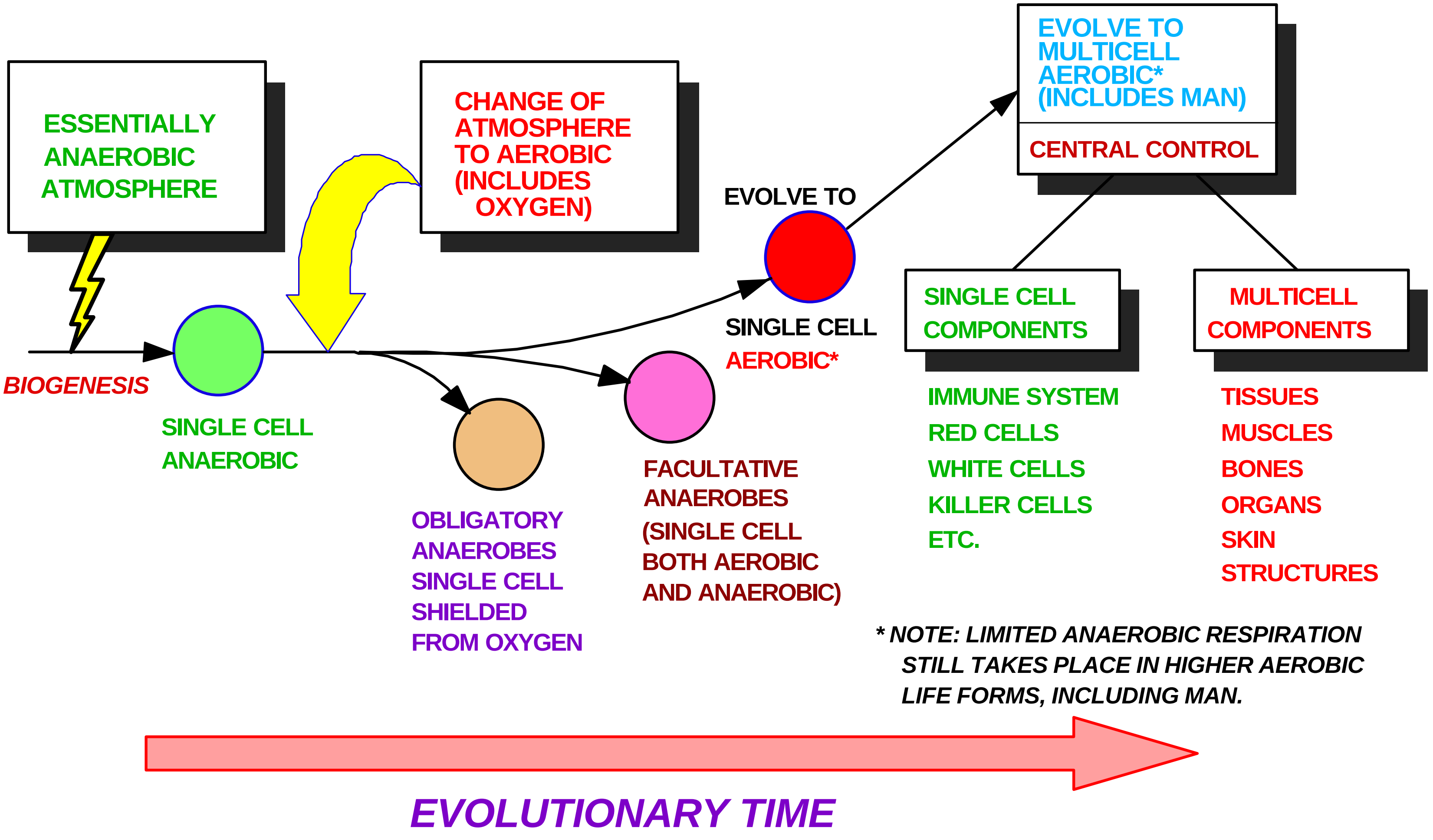

ESSENTIALLY ANAEROBIC ATMOSPHERE

BIOGENESIS

SINGLE CELL ANAEROBIC

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

First step in self-promotion of cancer due to hypoxia

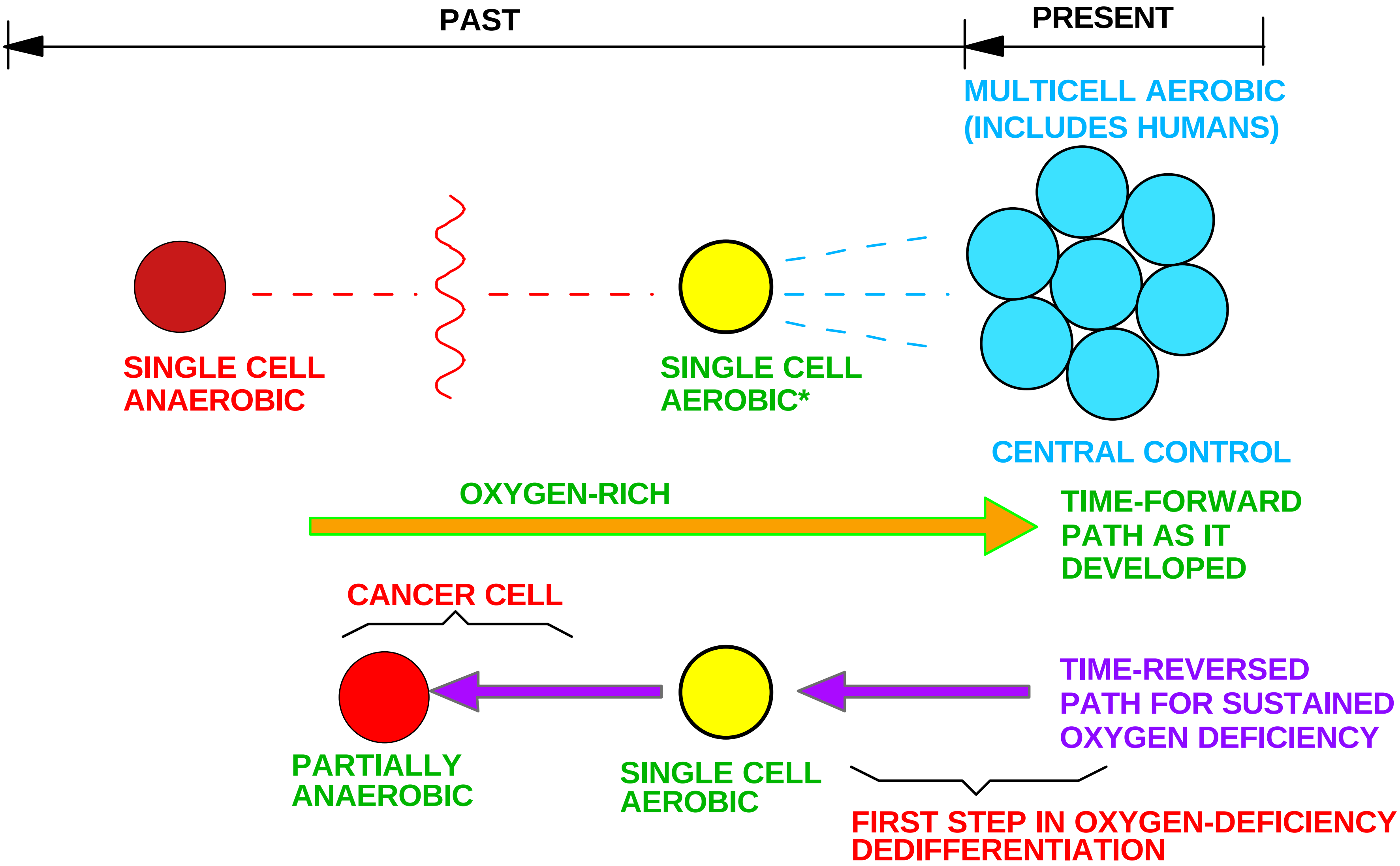

PAST

PRESENT

MULTICELL AEROBIC
(INCLUDES HUMANS)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

Block diagram of the cellular regenerative system

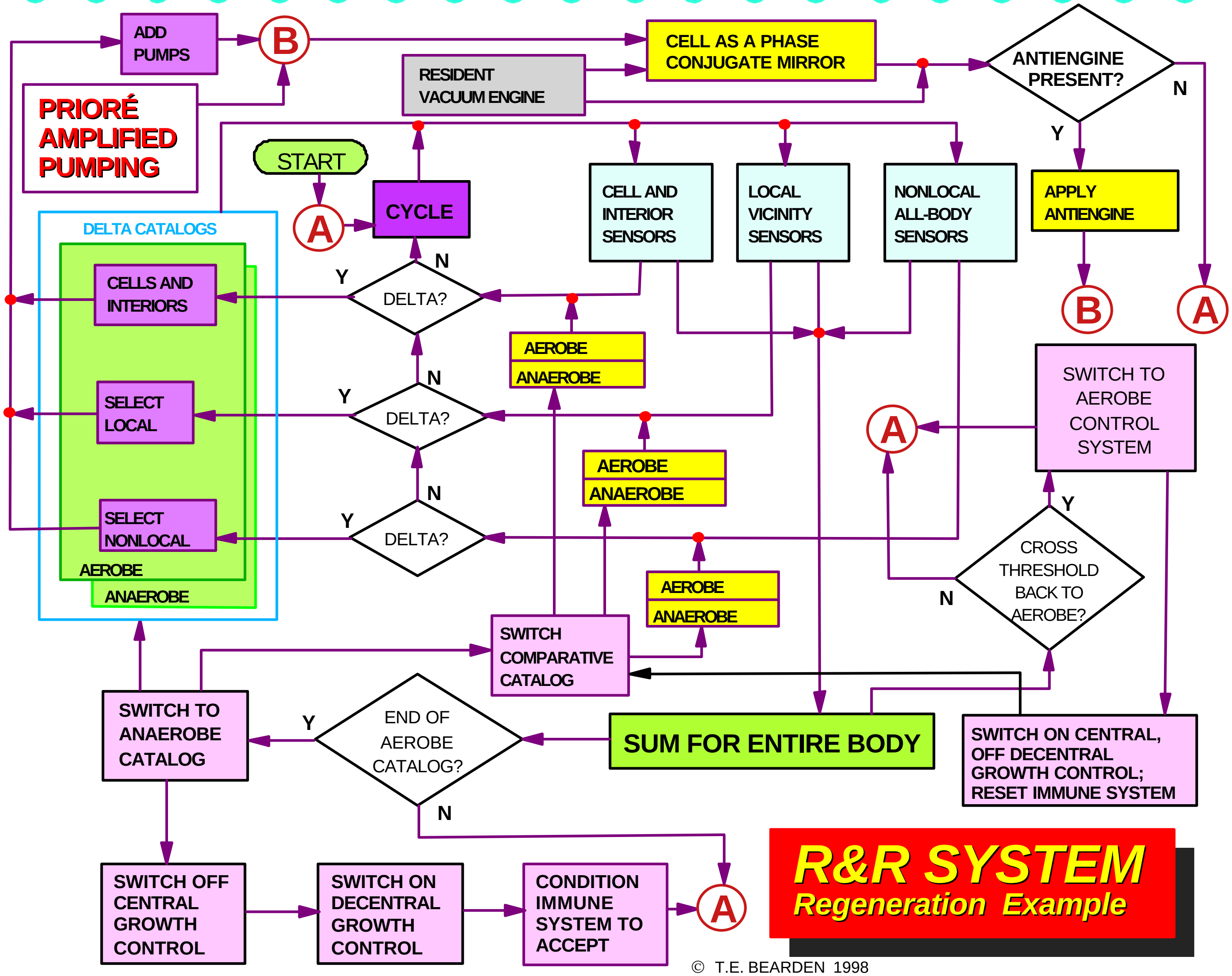

ADD PUMPS

B

CELL AS A PHASE CONJUGATE MIRROR

ANTIENGINE PRESENT?

N

Y

RESIDENT VACUUM ENGINE

PRIORÉ AMPLIFIED PUMPING

START

CYCLE

A

CELL AND INTERIOR SENSORS

LOCAL VICINITY SENSORS

NONLOCAL ALL-BODY SENSORS

APPLY ANTIENGINE

B

A

DELTA CATALOGS

CELLS AND INTERIORS

SELECT LOCAL

SELECT NONLOCAL

AEROBE

ANAEROBE

DELTA?

Y

N

AEROBE

ANAEROBE

DELTA?

Y

N

AEROBE

ANAEROBE

DELTA?

Y

N

AEROBE

ANAEROBE

SWITCH TO AEROBE CONTROL SYSTEM

A

CROSS THRESHOLD BACK TO AEROBE?

Y

N

SWITCH COMPARATIVE CATALOG

SWITCH TO ANAEROBE CATALOG

END OF AEROBE CATALOG?

Y

N

SUM FOR ENTIRE BODY

SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM

SWITCH OFF CENTRAL GROWTH CONTROL

SWITCH ON DECENTRAL GROWTH CONTROL

CONDITION IMMUNE SYSTEM TO ACCEPT

A

R&R SYSTEM

Regeneration Example

© T.E. BEARDEN 1998

Bombshell: Amplified EM healing

- A unified field theory approach and use of supersystem analysis has broken the basic mechanism for healing itself.
  - The Sachs unified field theory applies and has been partially fitted.
  - Evans-Vigier O(3) electrodynamics can properly model it, including both the internal and external electrodynamics and effects of engines.
  - Direct engineering development is now possible.
- Experimental proof of the results obtainable already exists in the literature.
- A medical revolution is in the offing.

Potential Medical Applications

- Non-traumatic cancer treatment
  - *Cure cancers and leukemias of all types*
  - *Remove pre-cancerous state*
  - *Restore system*
- Treat and cure Infectious diseases
- Cure AIDS and restore immune systems
- Defense against biological warfare
  - *Virulent diseases from unknown agents*
  - *Treat, cure prior to symptoms developing*
  - *Both civil populace and armed forces*
- *Dramatically lower health costs*
- *Rejuvenate the aged*

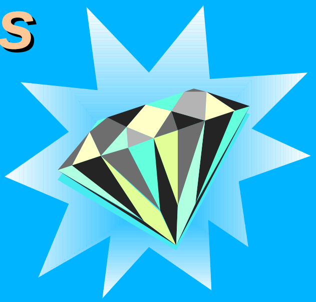

© 1999 T. E. BEARDEN

To treat and save mass casualties: What is urgently needed now:

- A small, portable, relatively inexpensive treatment unit, not requiring plasmas etc.
- Treatment in 2 to 5 minutes rather than 2-hours.
- Simple treatment application, almost anywhere.
- Effective in advanced stages as well as prior to emergence of symptoms.
- Effectiveness against a wide array of illnesses.
- Use is easy, high school student can be trained in 30 minutes.
- Crash program to develop, flood down through emergency systems in great quantities.

The Solution:

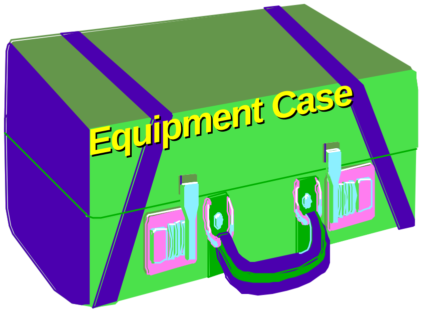
Equipment Case

- Two blankets have antennas in them.
- Uses the "porthole" concept (discussed shortly)
- Eliminates plasma tube mixing etc.
- Computer-controlled complex signal structure.

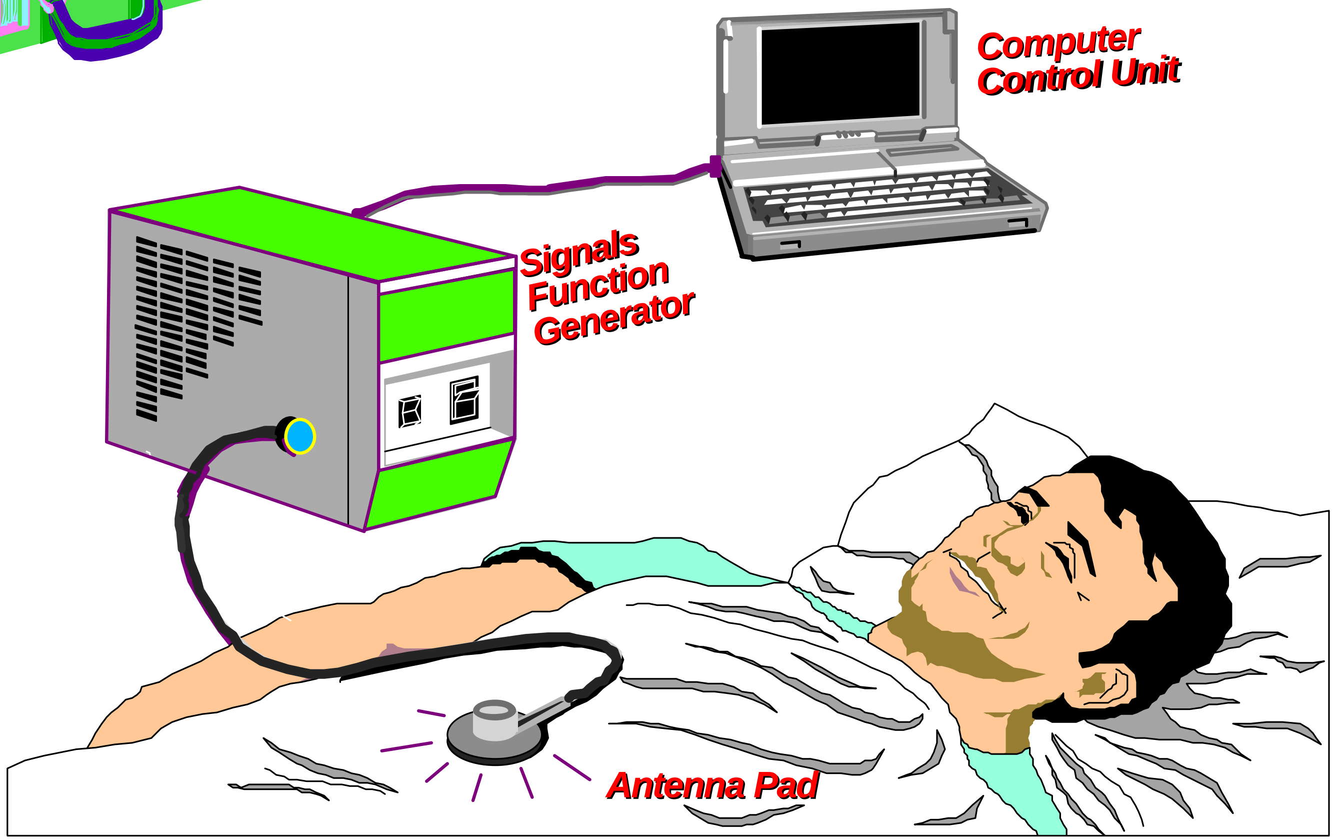
Computer Control Unit
Signals Function Generator
Antenna Pad

The Porthole concept:

The two-minute treatment shortcut

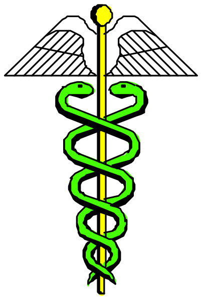

How the porthole concept works

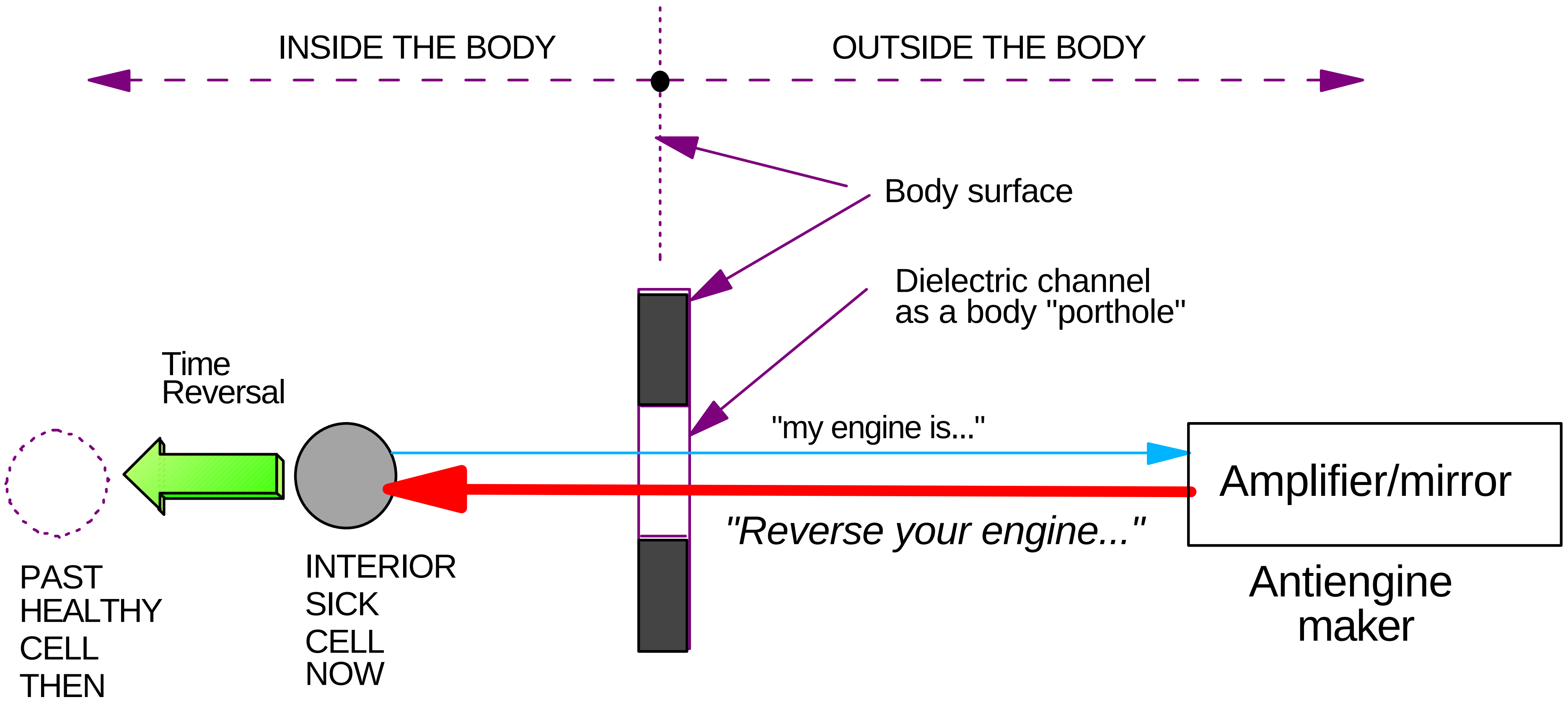

INSIDE THE BODY

OUTSIDE THE BODY

Body surface

Dielectric channel
as a body "porthole"

Time
Reversal

"my engine is..."

Amplifier/mirror

"Reverse your engine..."

PAST
HEALTHY
CELL
THEN

INTERIOR
SICK
CELL
NOW

Antiengine
maker

Porthole concept: An analogy of the dielectric path and inner-outer functional connections

- Imagine perfectly reversible dynamic light processes inside a large "cruise ship", that are also perfect phase conjugate mirrors
- Each process receives, absorbs, and emits light continuously
- Light escaping through the portholes is a composite from all the dynamic processes ongoing inside
- If we record and amplify the escaping light, then force it back through the portholes, it reverses all the internal processes
- The processes revert back, slowly, to a previous ongoing process state
- We cease radiation, and the processes reverse and move forward again

Porthole concept: The discovery

- The processes themselves, *in their own "frame"*, are never actually "reversed". Instead, the frame's motion through time is reversed.
  - Each to itself moves normally -- just receiving, absorbing, and radiating light, regardless of "whatever light" is interacting
  - This is due to a peculiarity of the photon and photon interaction
- Photon interaction creates observed physical reality, including the "frame" and the seeming "passage of a mass through time"
  - The photon is its own antiparticle, which is just a photon traveling in the opposite direction
  - So in this case we can use "light" to precisely time reverse (to the observer) the frame and the apparent operations (to the observer)
- When we cease irradiation, after a decay period to remove the reversed excitation, all functions are now seen to be moving forward normally in observer forward time, but from a previous physical state or condition

Using the porthole concept: Research

- The body's own emitted dense, broadband signals are recorded across a broad bandwidth
  - These signals are amplified, then rebroadcast precisely back into the body
  - This returns an amplified "reversed" signal set back through the dielectric path, invoking and applying the porthole concept in all processes
- The returning photons produce time-charging but in reversed frame direction
  - Tremendous potential energy is absorbed in highly compressed "time-energy" form in seconds
  - This "time-potential" slowly drains off as the body processes and cells reverse back to a previous state

Research using the porthole concept

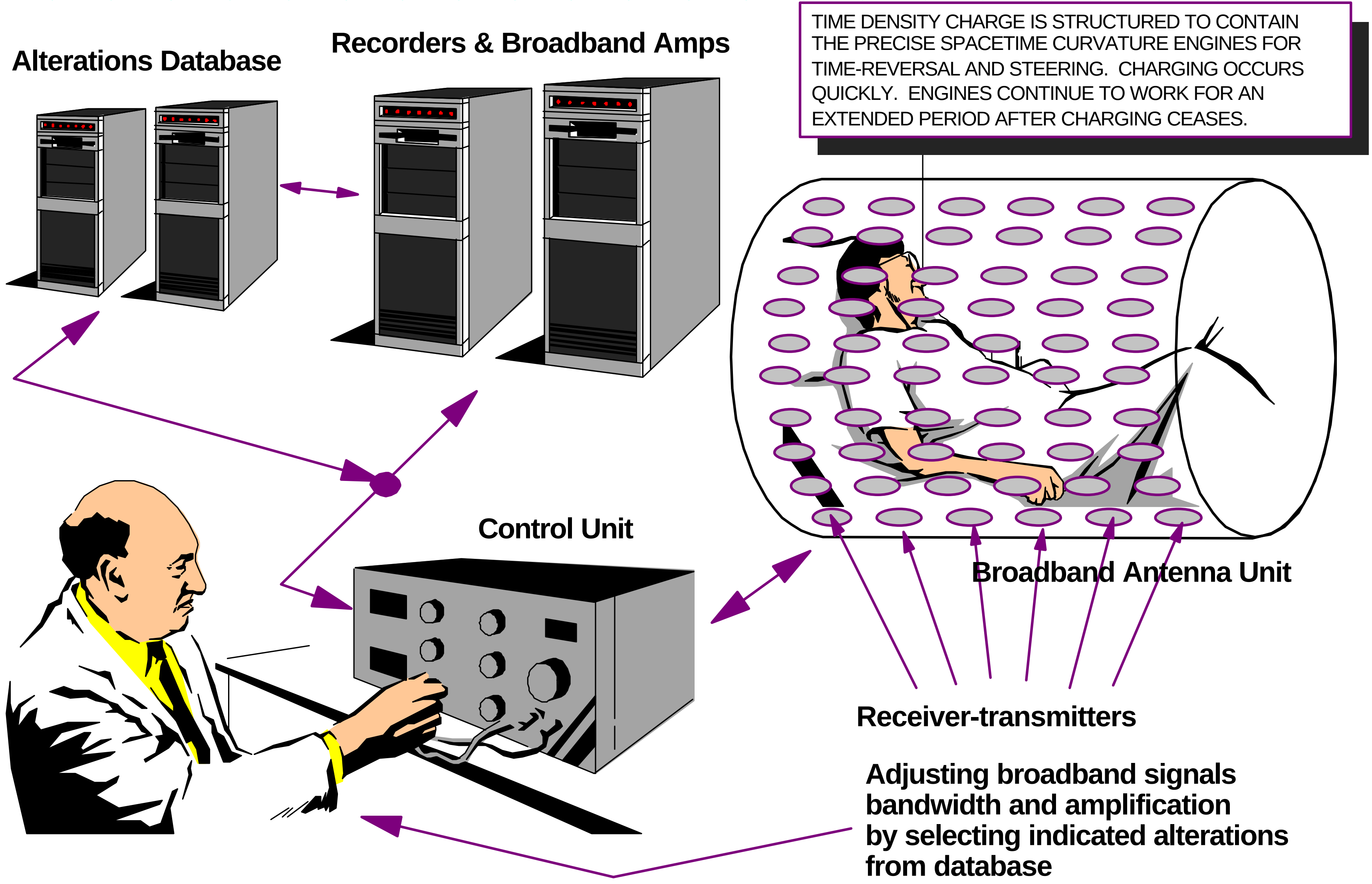

Alterations Database

Recorders & Broadband Amps

TIME DENSITY CHARGE IS STRUCTURED TO CONTAIN THE PRECISE SPACETIME CURVATURE ENGINES FOR TIME-REVERSAL AND STEERING. CHARGING OCCURS QUICKLY. ENGINES CONTINUE TO WORK FOR AN EXTENDED PERIOD AFTER CHARGING CEASES.

Control Unit

Broadband Antenna Unit

Receiver-transmitters

Adjusting broadband signals bandwidth and amplification by selecting indicated alterations from database

© T.E. BEARDEN 1998

Portable unit then developed for rapid treatment of mass BW casualties

- Two blankets have antennas in them.
- Uses the "porthole" concept (discussed shortly)
- Eliminates plasma tube mixing etc.
- Computer-controlled complex signal structure.
- Irradiation for 2 minutes.

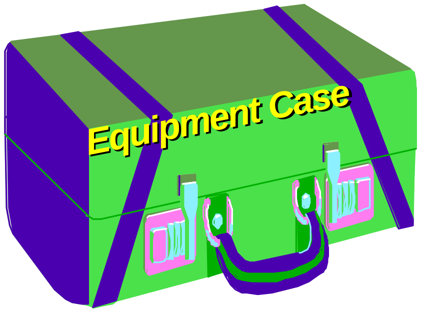
Equipment Case

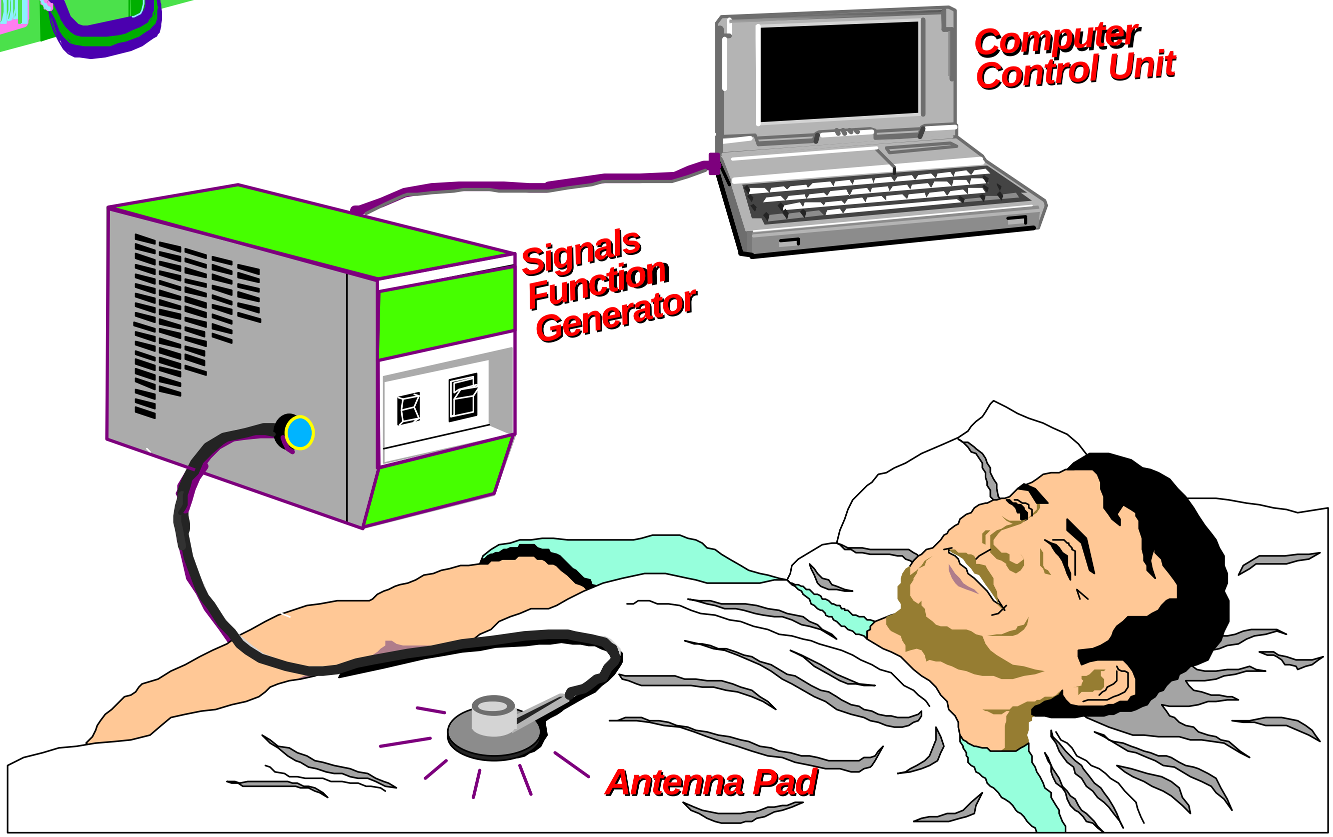
Computer
Control Unit
Signals
Function
Generator
Antenna Pad

Training to use the unit

- High school student level, 30 minutes.
- Patient lies on one blanket, other covers him.
- Set the prescribed standard signal sets from listing.
- Computer-controlled generator produces the signal.
- Irradiation for 2 minutes, uncover patient, bring in another.
- Unit can be self-powering from vacuum energy (MEG*).

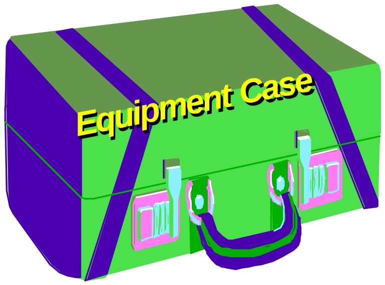
Equipment Case

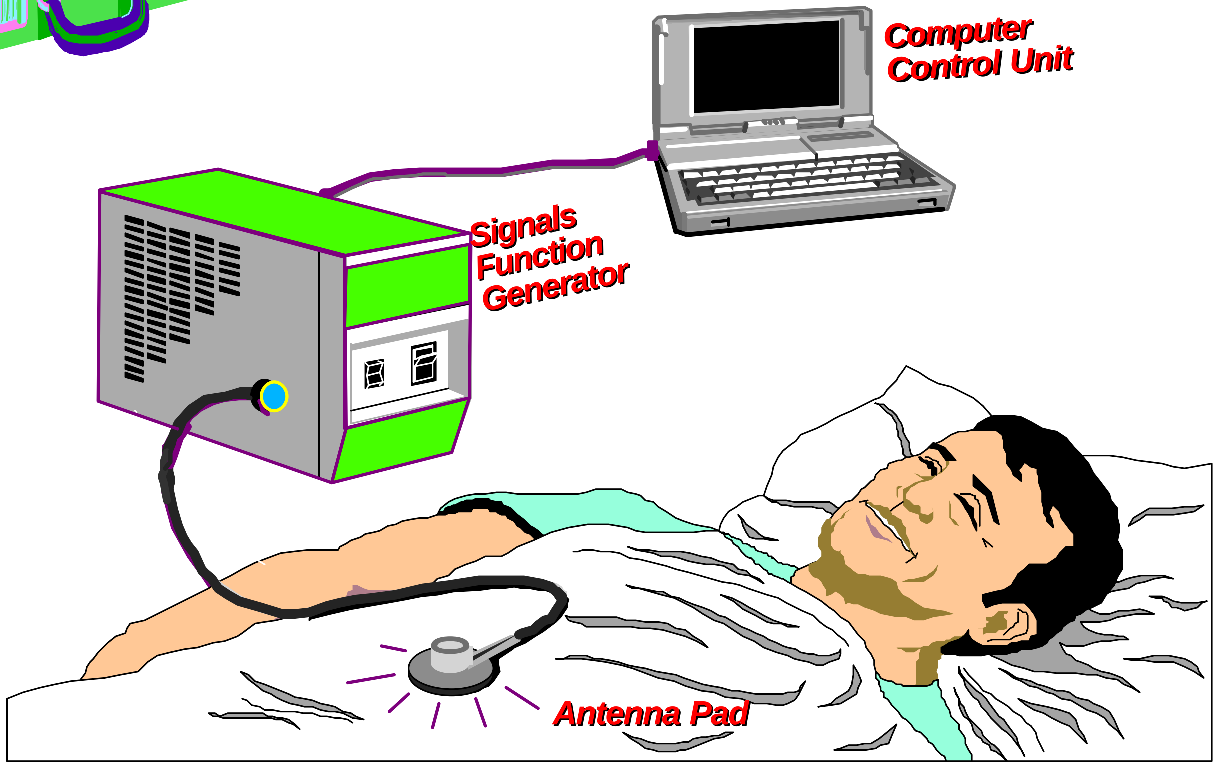
Computer
Control Unit
Signals
Function
Generator
Antenna Pad

Additional advantages

- Can be developed in a crash program in 2-3 years.
- After sunk costs, projected cost per unit in mass production is less than $100,000
- Can be produced by the thousands
- Then flooded down through
  - Fire stations and police stations
  - Red cross units, hospitals, clinics
  - Emergency action teams and armed forces
- Large numbers of operators easily produced
- Emergency training could be watching a 15 minute film, then performing one graded practice run
- First generation equipment can save some 70% of the coming mass casualties from anthrax etc.

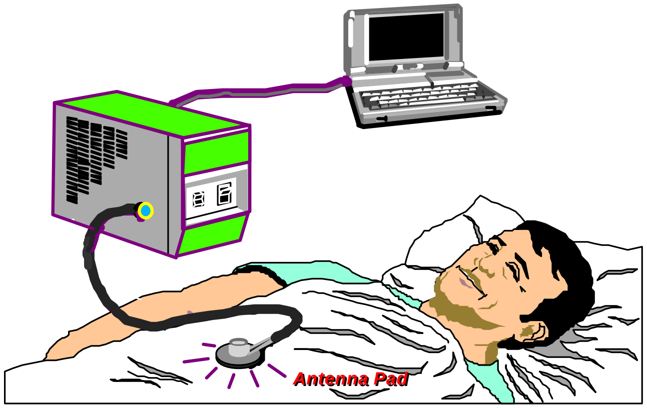

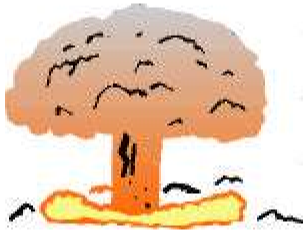

Terrorist BW warfare capabilities against the United States

The growing threat

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- **One terrorist, one light aircraft with spray tank**
- **100 kilograms of anthrax**
- **Flies over greater metropolitan Washington, D.C.**

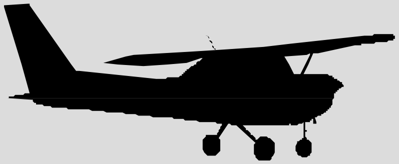

- **Calm night**
- ***1-3 million casualties result****
- ***Most of those stricken will die***
- **Presently little can be done to save the stricken civilians**

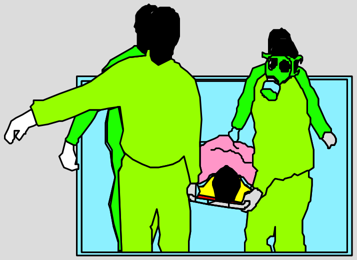

- ***Attacks on several population centers might produce some 10 million or more casualties***

***Per OTA Report to Congress, 1993**

For decades, infiltration of terrorist teams into the U.S.

- From several foreign nations, not just Bin Laden in Afghanistan
- Several thousand Cuban guerillas, from training camps in Southern Mexico

- Russian Spetznaz teams
- Nuclear warheads (up to 80 KT)
- Biological warfare agents
- Yakuza in oriental populations in our cities
  - Yakuza manufactures Russian energetics weapons in their own facilities in Japan
  - Possesses the new EM weapons referred to by Secretary of Defense Cohen in 1997
- We are now in a very long, hard war and will suffer millions of casualties in our cities. It may last for decades.

Defense Secretary Cohen*

(Involvement of secret superweapons)

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts."

* Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Some typical BW agents

- Smallpox, including different strain developed by Russia (some sold to terrorists)
- Anthrax
  - Over 400 known locations around the Earth have cultures
  - Difficult to decontaminate
- Bubonic Plague
- Botulism toxin
- Tularemia (rabbit fever)
- Hemorrhagic fevers, such as Ebola

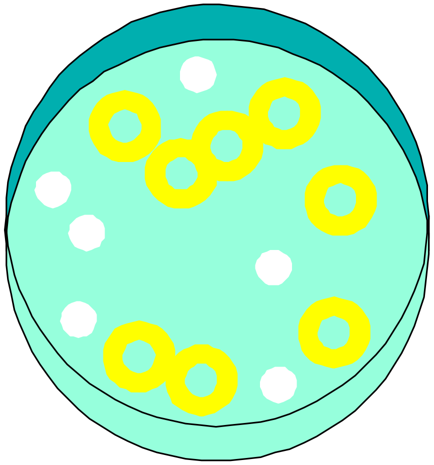

Anthrax

- OTA study: Some 200 pounds released in a spray, on calm night over Washington, D.C. could result in up to 3 million casualties
- Symptoms usually in 2-10 days
  - Sometimes as long as 7 weeks
  - Flu-like symptoms

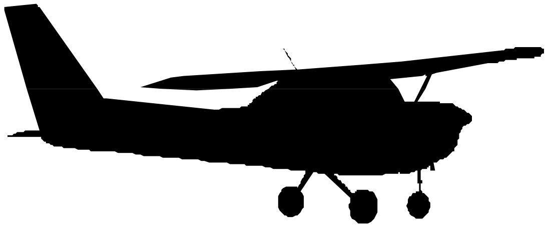

- Then 1-3 days later, shock and breathing problems cause death in up to 100%
- Antibiotics given early (before symptoms) can prevent infected persons from becoming ill
- Limited vaccine, reserved for military

Smallpox

- More than 500 million people killed in the 20th century before "eradicated" in 1977. Vaccination ceased in 1980.
- Russians kept it, developed new strains, sold to terrorists
- Symptoms in about 12 days: fever, headache, nausea. Rash similar to chickenpox then appears.
- Rash turns into hard blisters. *Highly contagious*, kills one-third of its victims.
- No cure. About 12 million doses of vaccine left.
- If smallpox is unleashed in any major city on earth, it will eventually kill 2 billion, or about 1/3 of the human species -- unless a radical new cure is found and utilized.

Bubonic Plague: Historically the most feared (Black Death)

- From 1980-1994, 18,739 cases in 20 countries.
- Symptoms occur in 1-6 days after inhaling the pneumonic form.

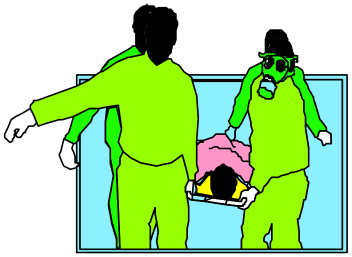

- High fever, cough, labored breathing lead to respiratory failure and death.
- It is contagious.
- Rapid use of antibiotics can be effective.
- Vaccine not currently produced in the U.S.

Botulism toxin (a nerve toxin)

- Single most poisonous BW substance known.
  - Typically borne in food, contacted by eating.
  - Could be developed as aerosol weapon.
- Symptoms in 24-36 hours; blurred vision and difficulty swallowing, speaking.
- Paralyzes muscles, leading to respiratory failure and death.
- Centers for Disease Control and Prevention maintain a botulism antitoxin supply.

Tularemia

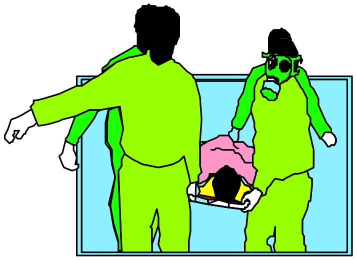

- U.S. studied its weapon potential in 1950s and 1960s.
- Symptoms appear in 3-5 days, including fever, chills, headache, and weakness
- Inflammation and hemorrhaging of the airways can lead to death.
- Without antibiotics, one-third die. Can be 90% if breathed in, in aerosol form.
- Vaccine developed and under review by FDA.

Hemorrhagic fevers, such as Ebola

- E.g., origin of Ebola is unknown. Probably transmitted to humans by animals, to begin a breakout.
- Symptoms in 3-5 days, include fever, aching muscles, diarrhea.
- Hemorrhaging of fluids out of tissues and orifices.
- From 30 to 90 percent die, depending on specific fever type.
- Some of these fevers respond to antiviral drugs.
- The drugs are in short supply.
- Presently *containment* is the primary goal.

A factor the U.S. does not know

(Secret involvement of clandestine superweapons)

- Three nations (not the U.S.) developed quantum potential weapons using Bohm's hidden variable theory of quantum mechanics.
- Can induce disease engines in matter or bodies at a distance.
- Have developed "spreading" of immune systems in a targeted populace (cleverly tested in Gulf War against U.S. troops).
  - Cocktail mix of disease engines for 2 dozen different diseases
  - Targeted area's human immune systems react to these engines when they are still in virtual state, just below quantum threshold.
  - Spread immune system's resources thus "thinned" across response to two dozen shadow diseases at once.
- Hit with a terrorist BW attack, the spread immune systems are overwhelmed much more easily and quickly.
  - Yield (casualties) of the attack is increased by factor of 3-5.
  - Spreading of U.S. immune systems has been accomplished and is still being maintained.

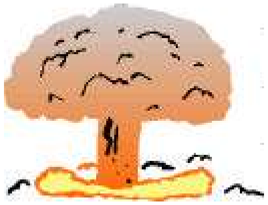

Summary and conclusion

The new war

A new Pearl Harbor: Strike on the World Trade Center in NYC

- The North Tower has already been struck and is burning.
- A second jetliner is about to strike the South Tower.

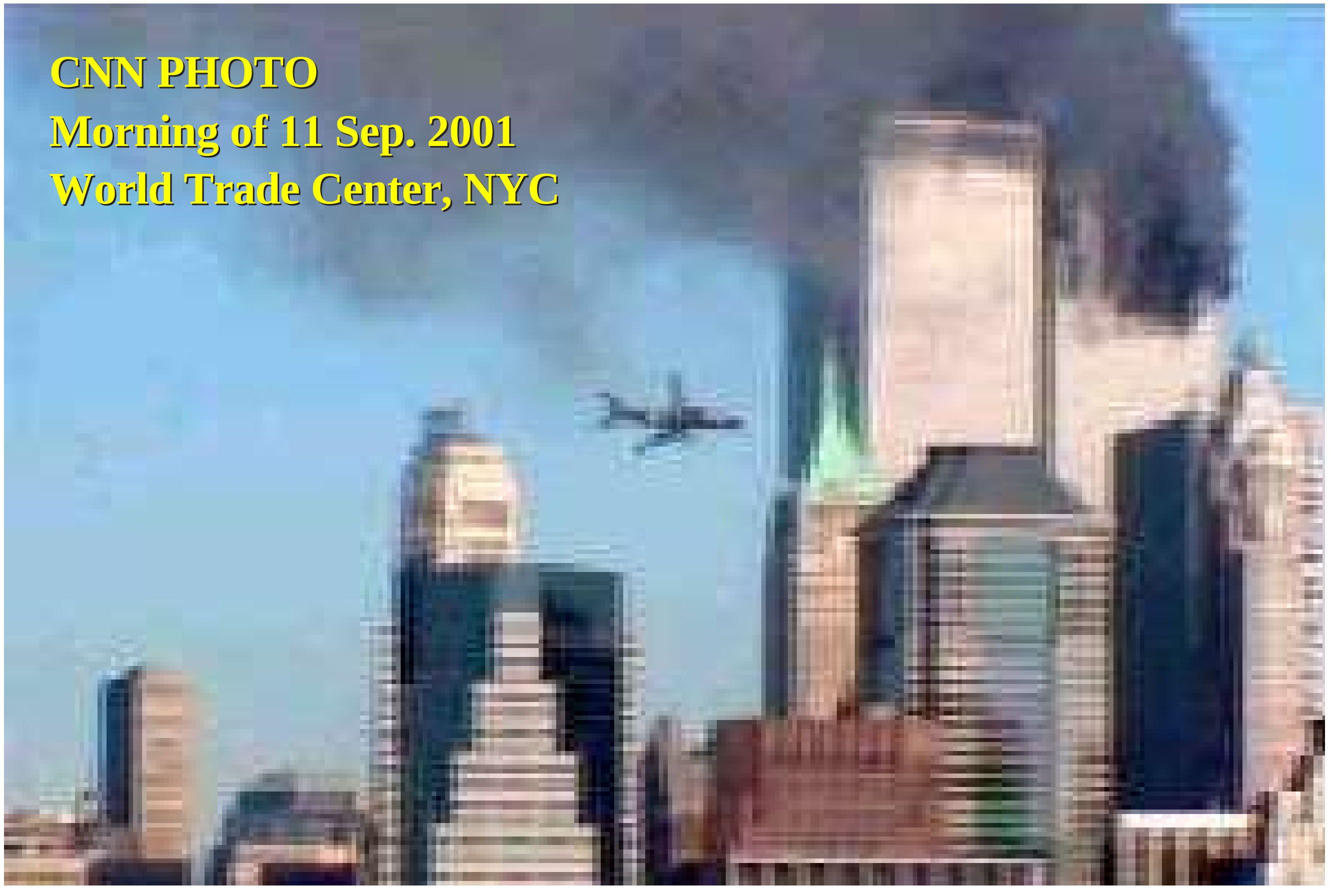

CNN PHOTO
Morning of 11 Sep. 2001
World Trade Center, NYC

Fireball From Impact on South Tower

- The two buildings held some 50,000 occupants
- Evacuation was underway after the first strike
- More than 6,000 persons were killed
- This is more than were killed in the Dec. 7, 1941 strike on Pearl Harbor

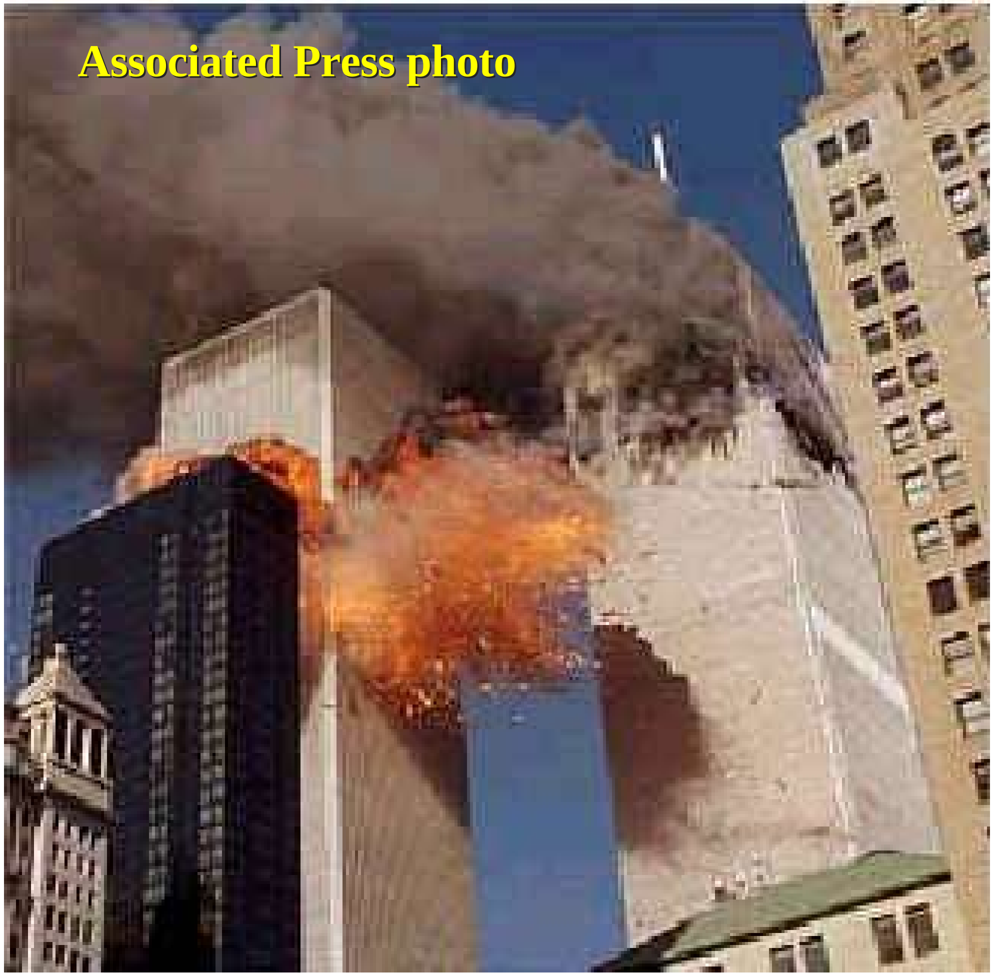
Associated Press photo

Collapse of the South Tower

- The first tower to collapse was the South Tower
- It is still uncertain just how many people were caught inside and killed
- This is a half million tons of debris, falling from the New York skyview.

Photo from CNN

Collapse of the North Tower begins

- The North Tower then collapsed
- This was another half-million tons of debris falling from the NYC skyview

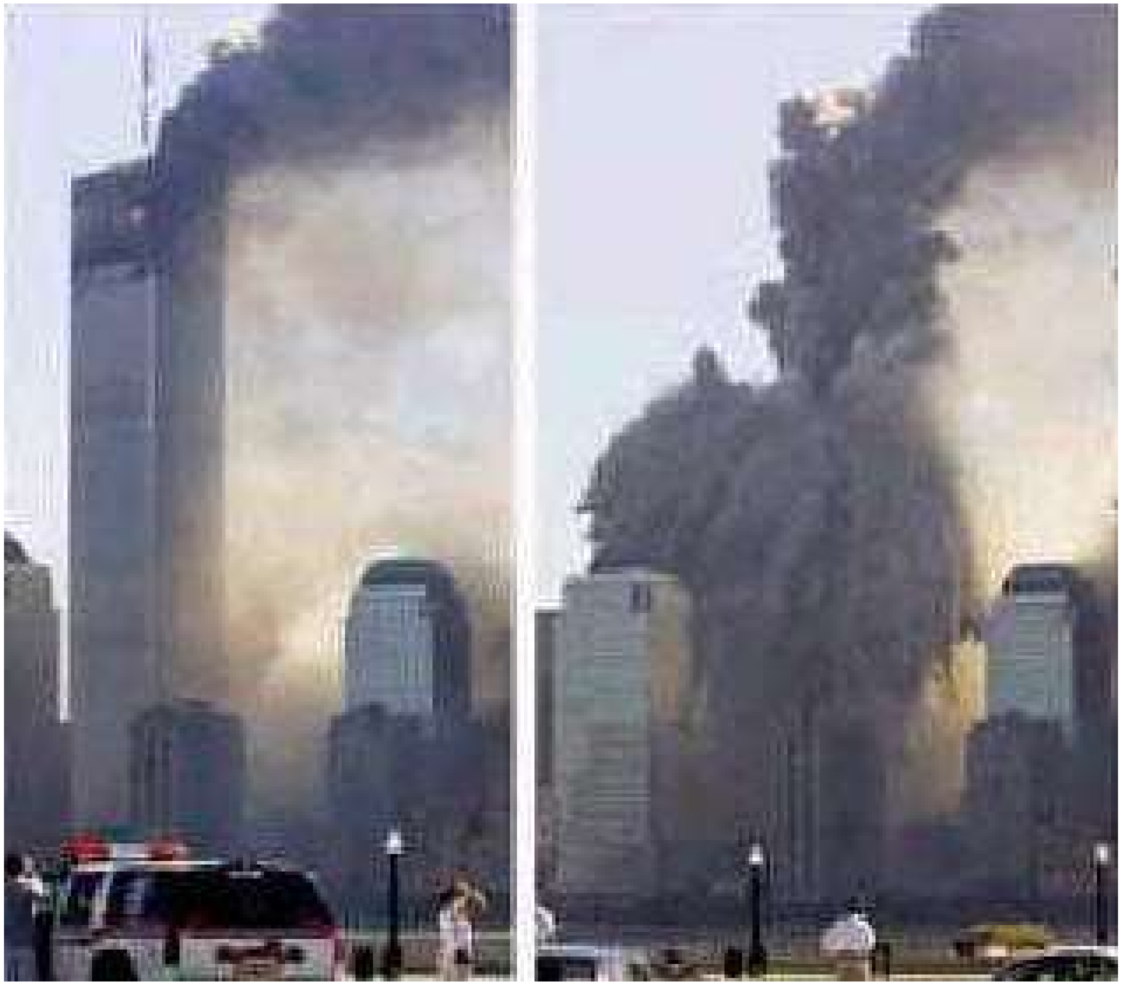

Collapse of North Tower finishes

- Collapse of the North Tower killed many rescue workers, fire-men, and policemen
- At this writing, the total casualties are still unknown

Photos from CNN

Next began the recovery efforts

- Recovery operations will require months
- At this writing, the total casualties are still unknown
- The damage was done by two civilian jetliner "bombs"

The Pentagon was hit on the same day

- Hundreds were killed the same day when another jetliner struck the Pentagon.
- Yet another jetliner, possibly headed for the White House, crashed enroute when the passengers attacked the suicide terrorist hijackers.

A sobering thought

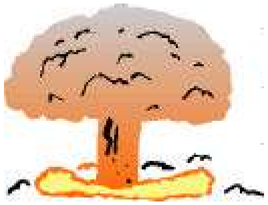

- Nuclear weapons are already smuggled into the U.S., as are BW agents such as anthrax, smallpox, etc.
- Had a 40-80kT weapon been exploded in New York and one in Washington, millions would have perished and both cities would have been devastated
- There would have been no recovery of those cities for many decades

Present Status of the WMD Threat to the United States

- **Many foreign nations are hostile to the U.S. and sponsor terrorists**
- **Some 25 nations have WMD or are acquiring them**
  - **BW agents and weapons**
  - **Chemical agents and weapons**
  - **Nuclear materials and weapons**
- **Thousands of students and émigrés**
- **Infiltrated teams, BW, other WMD already**
- **Castro guerrillas infiltrated over the years**
- **Can do unacceptable damage to U.S. now**
- ***May reach knockout capability***

© 1998 T.E. BEARDEN

***Per OTA Report to Congress, 1993**

The U.S. is terribly unprepared

- Limited medical supplies, antibiotics, vaccines
- Insufficient facilities, emergency teams
- No civilian shelters (much less BW-filtered)
- Not enough doctors, nurses, masks, etc.
- Present treatment methods woefully inadequate
- Triage will apply
  - First treat those most likely to recover
  - Others (showing symptoms, or advanced) are just set aside to die
- Other attacks are a matter of *when*, not *if*.

Conclusion

- As several cities are struck, millions of U.S. civilian casualties will result. Devastation will be enormous.
- We shall eventually lose more American lives than in all the rest of our wars to date.
- Present medical therapeutic science is totally inadequate to do the job. It cannot save us.

- At some point, when we are sufficiently damaged, the Spetznaz or other terrorists may detonate nuclear weapons in our cities, with appalling carnage.
- A combined superweapon technology and terrorist groups will then destroy the remaining populace.
- On its present course, the inadequacy of Western medicine may doom us.

A difficult truth to consider

- In any strategic strike, the first phase is to get the strategic weapons (weapons of mass destruction) delivered to their distant targets.
- That has already been done.
- Eerily, the first phase of World War III has already been accomplished.
- It was already accomplished before the Sept. 11, 2001 strike on New York and Washington. That was the new "Pearl Harbor", and only the opening round. It already killed more Americans than died at Pearl Harbor on Dec. 7, 1941.
- Unless we improve our medical therapy quickly and dramatically, we will lose -- particularly when the hidden energetics weapons and outside terrorist sponsorship by hostile nations are factored in.

A personal perspective

- There now exists an overwhelming, immediate moral imperative to get this development program done at the utmost speed humanly possible.
- The sheer survival of the United States and Western civilization depends upon it.
- Officialdom is aware that the KGB/Russians have long planned to finish the U.S., after initial nuclear exchanges, by fierce and massive BW strikes.
- It appears that the plan has changed to let the terrorists do the main work, while augmenting U.S. casualty levels by energetics weapons.

China is also involved

- China has declared South China Sea the territorial waters of China
- China will eventually take Taiwan
  - Has massed missile artillery trained on Taiwan
  - Recently held invasion practice on an island near Taiwan
- Recently deployed fearsome EMP weapons utilizing negative energy rather than positive energy (Dirac sea theory)
  - Extinguishes all electron currents
  - Duds electrical and electron systems instantly
  - Instant death to struck bodies
  - Instant kill of missiles, aircraft, electronics in nuclear warheads
  - Ultimate "death ray"; can be made small or large
- Recent 2-carrier U.S. task force maneuvered in South China Sea to caution China
  - The new Chinese EMP weapon was trained on that task force
  - Could have destroyed it within minutes

Final thoughts

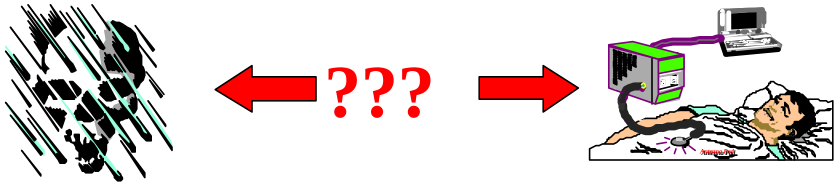
???

- This is one of the great turning points in history.
- The development program is eminently doable. That it works has been experimentally demonstrated. The long-vexing puzzle of the active mechanism has been solved.
- Our scientific community bitterly opposes any use of electromagnetics -- even an advanced, higher symmetry electrodynamics -- for medical treatment.
- A consortium of the scientific community, the large pharmaceuticals, FDA, and the AMA will oppose this project even while the U.S. is being destroyed wholesale.
- It can only be done in the private sector, and there will likely be no funding unless by philanthropists.

Continuation

Business plan, development program, budget breakdown, proposed schedule etc. go here.

Weapons Slides

The Tom Bearden Website

Help support the research

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts." *

Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Slides
(covers period up to 1999)

- The Coming Strategic Attack on the United States
- Outline
- Loss of two separate A-10 Warthogs in 1997
- Typical ABM Defenses (Widely Deployed by mid-70s)

- Launch Phase Anti-Missile System
- Launch Phase Anti-Bomber System
- HC-130s Engines Quit, Aircraft Crashes
  Nov. 22, 1996
- 1996 Actions by this analyst
- Additional Information Exists: Not Analyzed, Not Integrated in Briefing
- Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985
- Satellite Photos of Plumes in Soviet Arctic Since 1974
  See "Fer de Lance" for additional information
- Aum Shinrikyo ("Supreme Truth" Cult), Shoko Asahara, Leader

  Aum Shinrikyo - Anatomy of a Dangerous Cult
- Ashkhabad Problem and Its Solution
- Strange Characteristics of Iran Quake, Sep. 1978
- Iran Earthquake, September 1978
- Situation Now (Assessment)
- Situation Now (Assessment 2)
- Aspects of Strong Local Asymmetry
- Mass Casualty Problem: Aerial Anthrax

Spray on Washington, D.C.

- Present Status of the WMD Threat to the United States
- Some Present Problems in the U.S. WMD Defense Capabilities
- Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future
- Attempt to Ban Energetics Weapons Worldwide
- Brezhnev's Schedule
- Large glowing Tesla globe witnessed in Red China by hundreds
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  Sequence of Events
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  What They Were
- Cold Explosion -- 9 April 1984 Near Kurils (Plus additional weapons effects)
- Countering Insane MAD Systems:
  Find Loophole, Exploit It
- Next Generation Counter to Both Sane and Insane MAD systems
- Simultaneous Dudding of Nuclear Weapons Worldwide

- Simultaneous Dudding of Nuclear Weapons Worldwide
- Operation Divine Wind II
- Advanced EMBW Applications
- EM Biological Warfare (EMBW)
- Microwave Radiation of U.S. Embassy in Moscow
- Pentagon EMI Studies, 1989
- EM Missile Practice (6 slides)
- Subdivisions of Soviet Energetics Program
- Communism's Goal
- Manuilsky's Goal vs. Status Today (Dec. 1996)
- Professor Hellman, German Scientist Released from Work in Soviet Union
- Hemispheres and Globes of Light (4 slides)
- How Could Such Superweapons Ever Be Concealed?
- Soviets Easily Concealed These Superweapons for Four Decades
- Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes
- Longitudinal Wave Interferometry: Endothermic

- Longitudinal Wave Interferometry: Exothermic
- U.S. Air Attack on Libya, April 1986
- Lightning Wall
  circa 1959-1960
- Lisitsyn's Report: Brain Code Broken
- Large Longitudinal Wave Interferometer (LWI) Accident
- Large Longitudinal Wave Interferometer (LWI) Accident
- Use of Hidden Information Content of the Field Can Provide Action-At-A-Distance
- Two Types of Mutual Assured Destruction (MAD) System
- Psychoenergetics Weapons Teams Mentally Disable All Personnel
- Strikes Against Special Ship and Energetics Weapons Sites
- "Mindsnapper" Attack on Special Ships and Energetics Weapons Sites
- How Much Off Guard Are We?
  (7 Slides)
- Possible Targets: 1995-9 Escalation Phase
- Propagation of Weapon Effects
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15th, 1979

- Countering an Insane QP Weapon
- Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases
- Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
- Mechanism for Producing a Quantum Potential
- Major Principles for Use of Quantum Potential Weapons
- Major Principles for Use of Quantum Potential Weapons (2)
- Instantaneous Communication by a Quantum Potential
- Participants in a quantum potential share a common multiply-connected spacetime (MCST)
- Quake in Tangshan, China
  28 July, 1976
- Exchanging Earthquakes
- Instant Communication by a Quantum Potential
- Westerners Do Not React to Slow Threats
- Characterizing a Given Risk or Risk Factor
- Technical Risk Management
- Aum Shinrikyo's Sarin Facility

- In Scalar Beam Interference Zone
- Physicist Schappeller Auralaz Munster, Austria
- Initial Soviet Plan for Cold War, After WWII
- Josef Stalin: Absolute Dictator
- Exploding any Resistance by Operational Scientists
- Significant Soviet Statements
- Captain Svoboda Dives Headlong to Her Death
- T-polarized Beam Projector
- Possible Tests (1)
- Possible Tests (2)
- Possible Tests (3)
- Possible Tests (4)
- Mushroom Cloud Rising From Sea
- Launches from Cape Canaveral and Vandenberg AFB Provided Practice Targets of Opportunity
- The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth
- Embodiment of an instrument array sufficient to discriminate when time-density charging has occurred, and that transduction

of time-charge decay has occurred.

- Time-Polarized EM Wave Interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone
- Transmutation of Radioactive Nuclei at a Distance
- Transmutation of Radioactive Nuclei
- Typical Tesla Shield, Northern Route of Japan
- A Curious Tide in Human Affairs
- Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958
- The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
- The KGB/Yakuza-Aum War Plan
- Why Western Science is Ignorant of KGB Energetics
- Transmutation of Radioactive Nuclei at a Distance
- TheYakuza
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15, 1979

---

* Note that the Secretary did not speak the words in square blocks; these were added by the writers later,

probably to "soften the tone" of the Secretary's information release. So the Secretary confirmed that some nations of the world -- not just some ragged terrorists, as the writers tried to imply for spin control -- do possess novel electromagnetic weapons and are using them to induce earthquakes, engineer the weather and climate, and trigger volcanic eruptions.

THE COMING STRATEGIC ATTACK ON THE UNITED STATES

By an AUM/Yakuza/KGB/Communist Coalition Using Energetics SuperWeapons

T.E. Bearden, LTC
U.S. Army (Retired)
M.S. Nuclear Eng.
Fellow Emeritus, AIAS

With Technical Background and History of Energetics Weapons

Partial support by the Association of Distinguished American Scientists is gratefully acknowledged.d

© 1995, 1996, 1997 T.E. BEARDEN

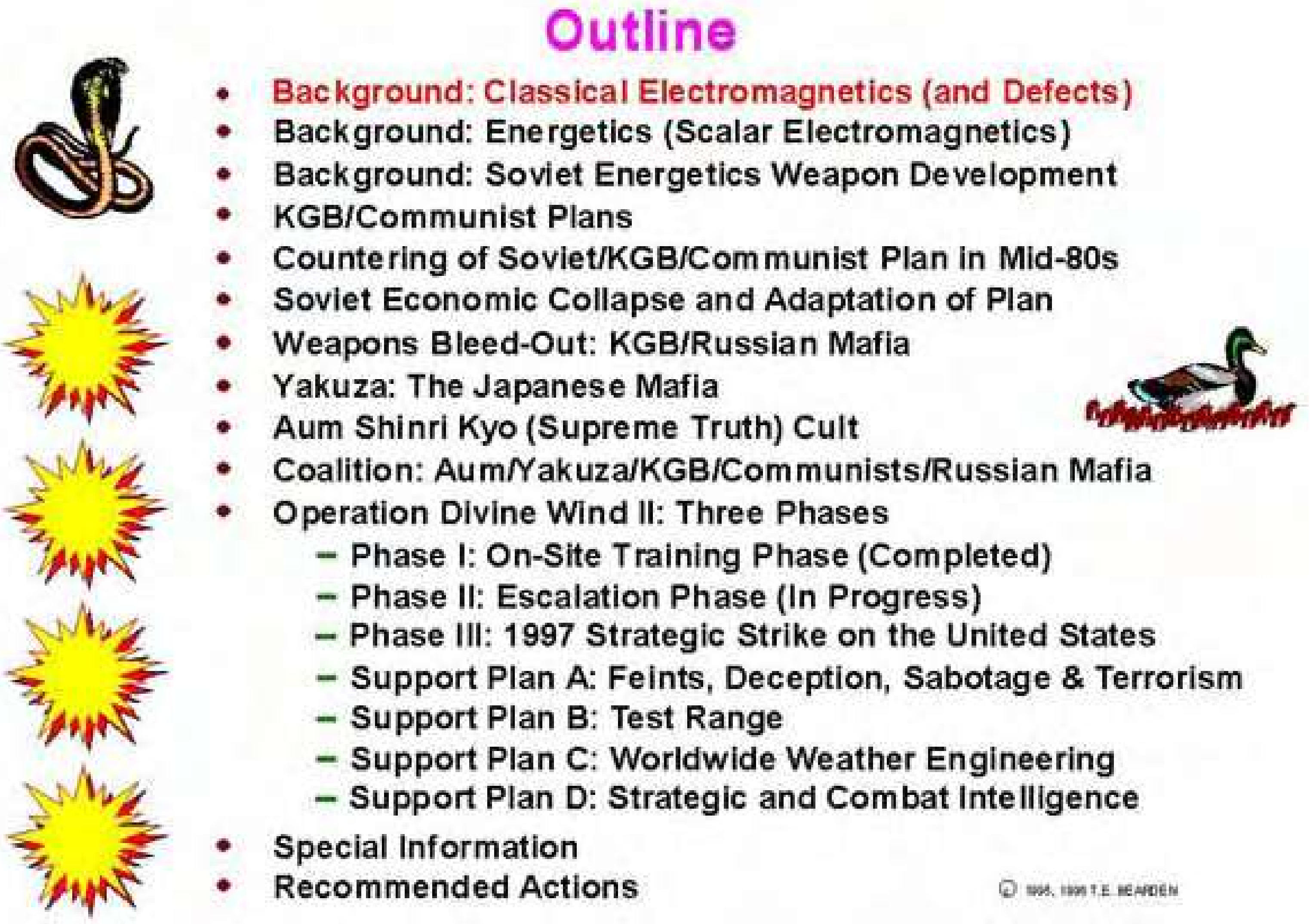

Outline

- Background: Classical Electromagnetics (and Defects)
- Background: Energetics (Scalar Electromagnetics)
- Background: Soviet Energetics Weapon Development
- KGB/Communist Plans
- Countering of Soviet/KGB/Communist Plan in Mid-80s
- Soviet Economic Collapse and Adaptation of Plan
- Weapons Bleed-Out: KGB/Russian Mafia
- Yakuza: The Japanese Mafia
- Aum Shinri Kyo (Supreme Truth) Cult
- Coalition: Aum/Yakuza/KGB/Communists/Russian Mafia
- Operation Divine Wind II: Three Phases
  - Phase I: On-Site Training Phase (Completed)
  - Phase II: Escalation Phase (In Progress)
  - Phase III: 1997 Strategic Strike on the United States
  - Support Plan A: Feints, Deception, Sabotage & Terrorism
  - Support Plan B: Test Range
  - Support Plan C: Worldwide Weather Engineering
  - Support Plan D: Strategic and Combat Intelligence
- Special Information
- Recommended Actions

© 1995, 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

Loss of two separate A-10 Warthogs in 1997

- A-10 Mysterious Flyaway and Crash - April 2, 1997
- Death of Second A-10 Pilot - May 27, 1997
- Captain Button's Mysterious Flyaway and Eventual Crash
- Death of Second A-10 Pilot - May 27, 1997

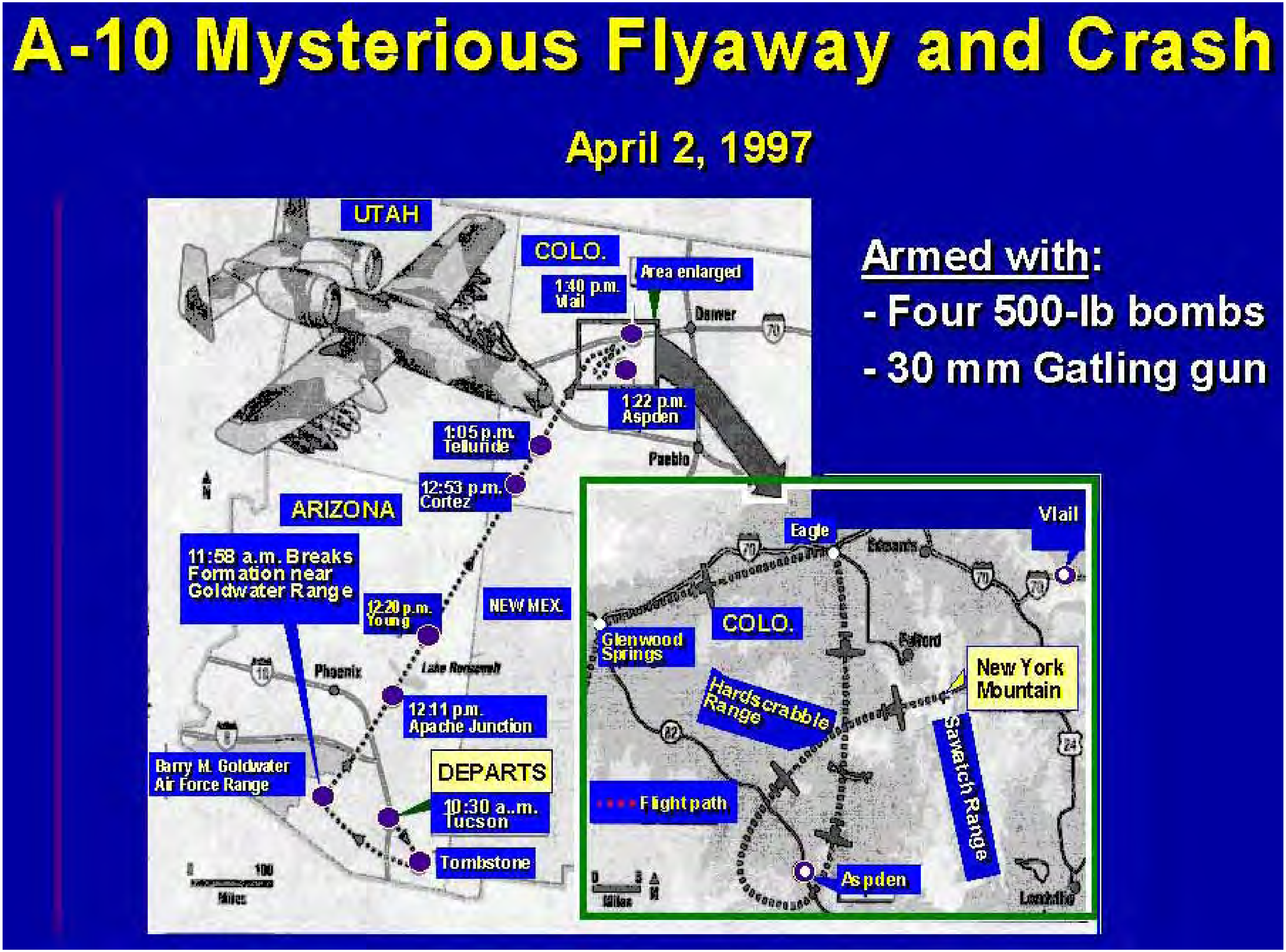

A-10 Mysterious Flyaway and Crash

April 2, 1997

UTAH
COLO.
Area enlarged
1:40 p.m. Vail
Denver
1:22 p.m. Aspden
1:05 p.m. Telluride
Pueblo
12:53 p.m. Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
NEW MEX.
Phoenix
Lake Roosevelt
12:11 p.m. Apache Junction
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a..m. Tucson
Tombstone
0 100 Miles
N
Eagle
Edwards
Vlail
Glenwood Springs
COLO.
Hardscrabble Range
New York Mountain
Sawatch Range
Flight path
Aspden
0 5 Miles
N
Leadville

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

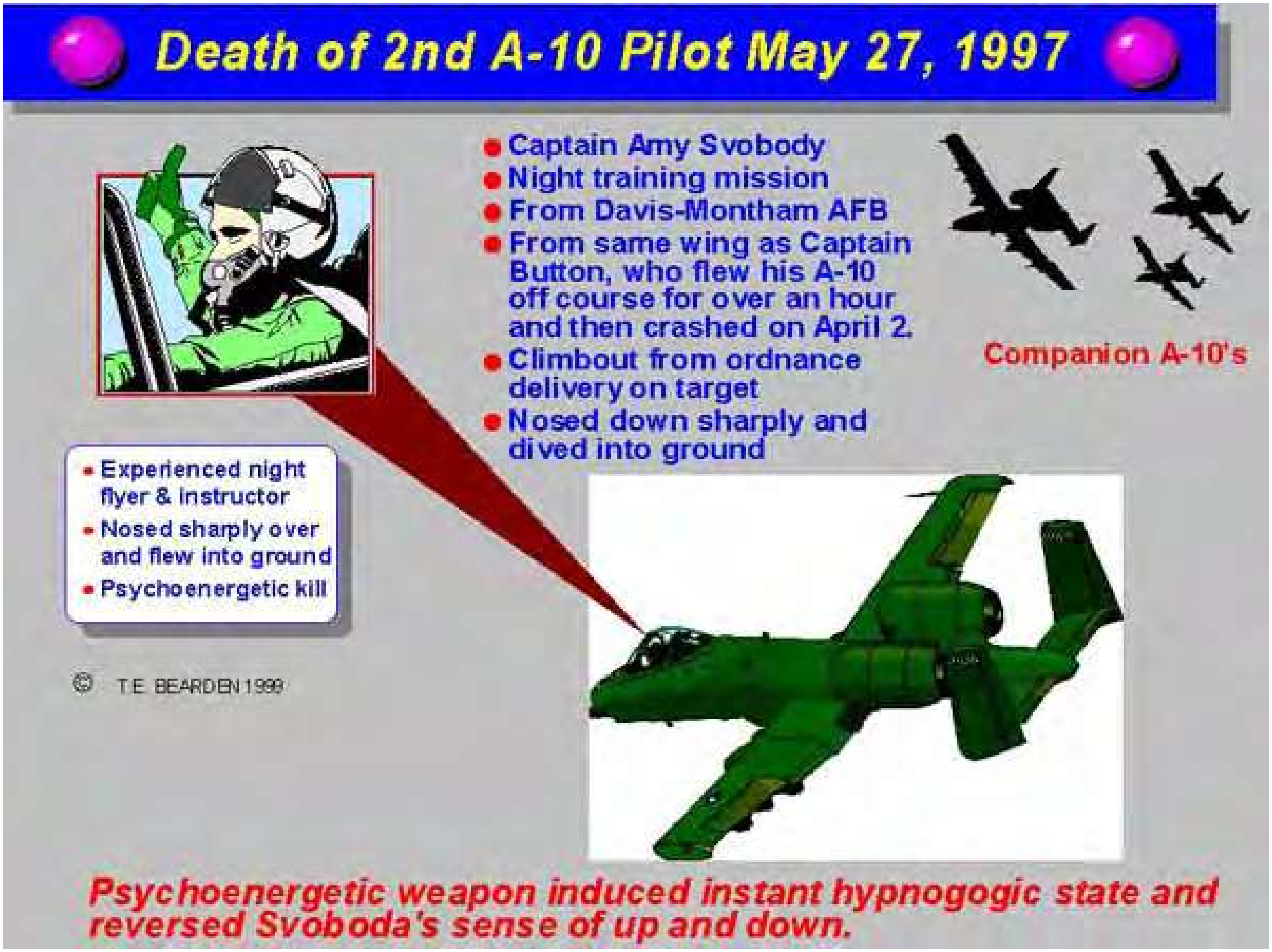

Death of 2nd A-10 Pilot May 27, 1997

- Captain Amy Svobody
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Climbout from ordnance delivery on target
- Nosed down sharply and dived into ground

Companion A-10's

- Experienced night flyer & instructor
- Nosed sharply over and flew into ground
- Psychoenergetic kill

© T.E. BEARDEN 1999

Psychoenergetic weapon induced instant hypnogogic state and reversed Svoboda's sense of up and down.

A-10 Mysterious Flyaway and Crash

April 2, 1997

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

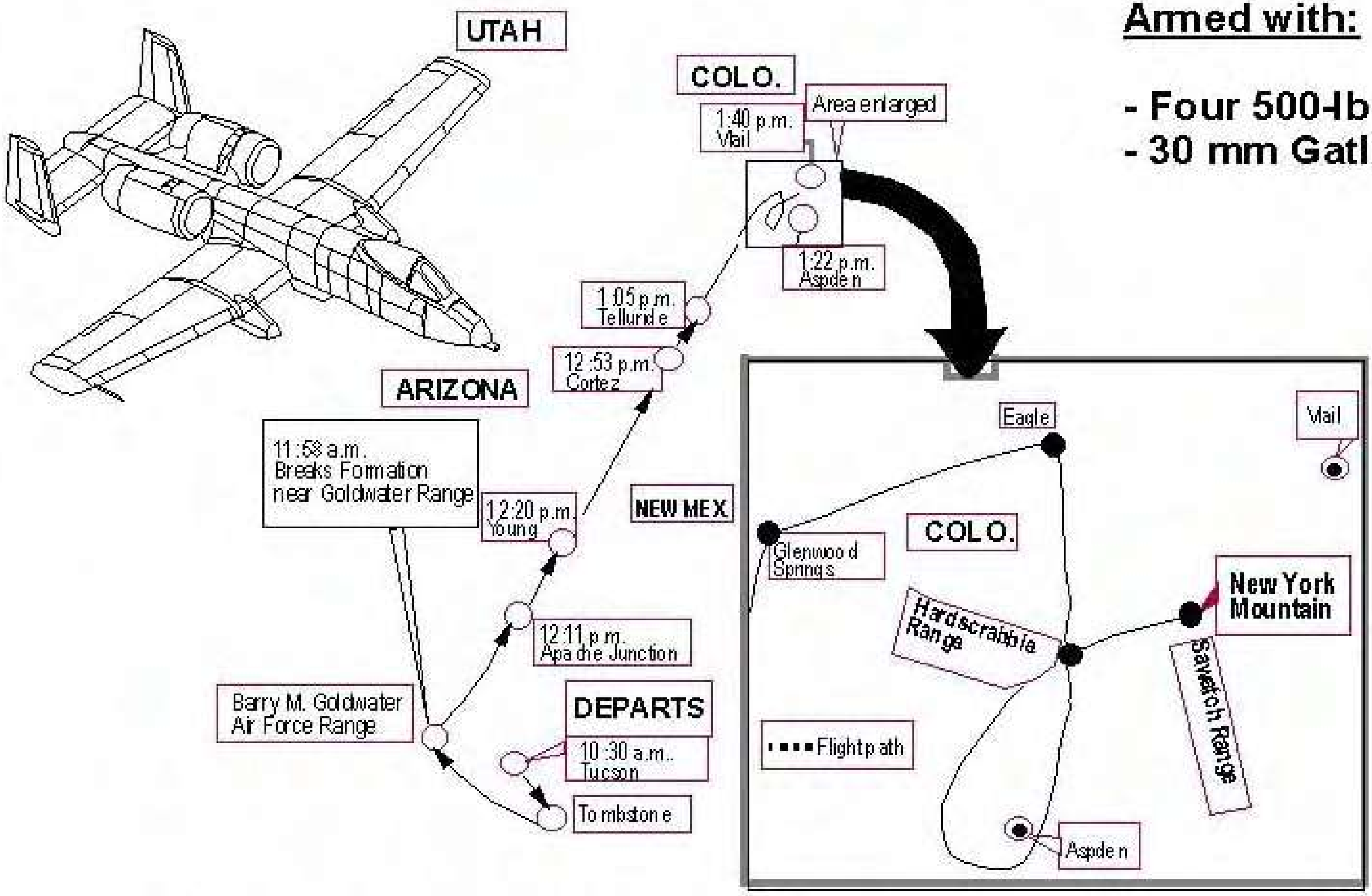
UTAH
COLO.
1:40 p.m.
Vail
Area enlarged
1:22 p.m.
Aspden
1:05 p.m.
Telluride
12:53 p.m.
Cortez
ARIZONA
11:58 a.m.
Breaks Formation
near Goldwater Range
12:20 p.m.
Young
NEW MEX
12:11 p.m.
Apache Junction
Barry M. Goldwater
Air Force Range
DEPARTS
10:30 a.m.
Tucson
Tombstone
Eagle
Vail
COLO.
Glenwood
Springs
New York
Mountain
Hardscrabble
Range
Sawatch Range
Flight path
Aspden

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

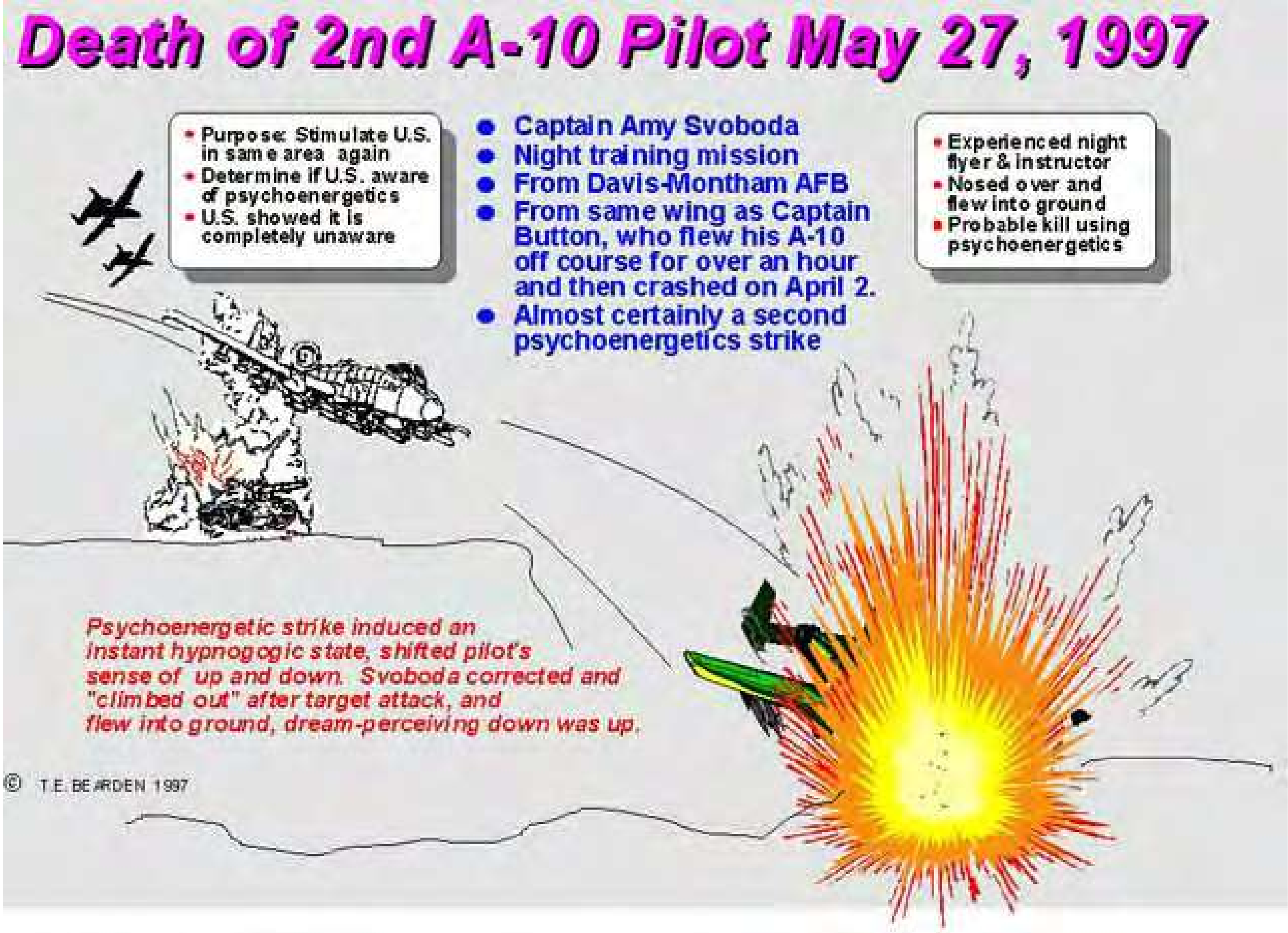

Death of 2nd A-10 Pilot May 27, 1997

- Purpose: Stimulate U.S. in same area again
- Determine if U.S. aware of psychoenergetics
- U.S. showed it is completely unaware

- Captain Amy Svoboda
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Almost certainly a second psychoenergetics strike

- Experienced night flyer & instructor
- Nosed over and flew into ground
- Probable kill using psychoenergetics

Psychoenergetic strike induced an instant hypnogogic state, shifted pilot's sense of up and down. Svoboda corrected and "climbed out" after target attack, and flew into ground, dream-perceiving down was up.

© T.E. BEARDEN 1997

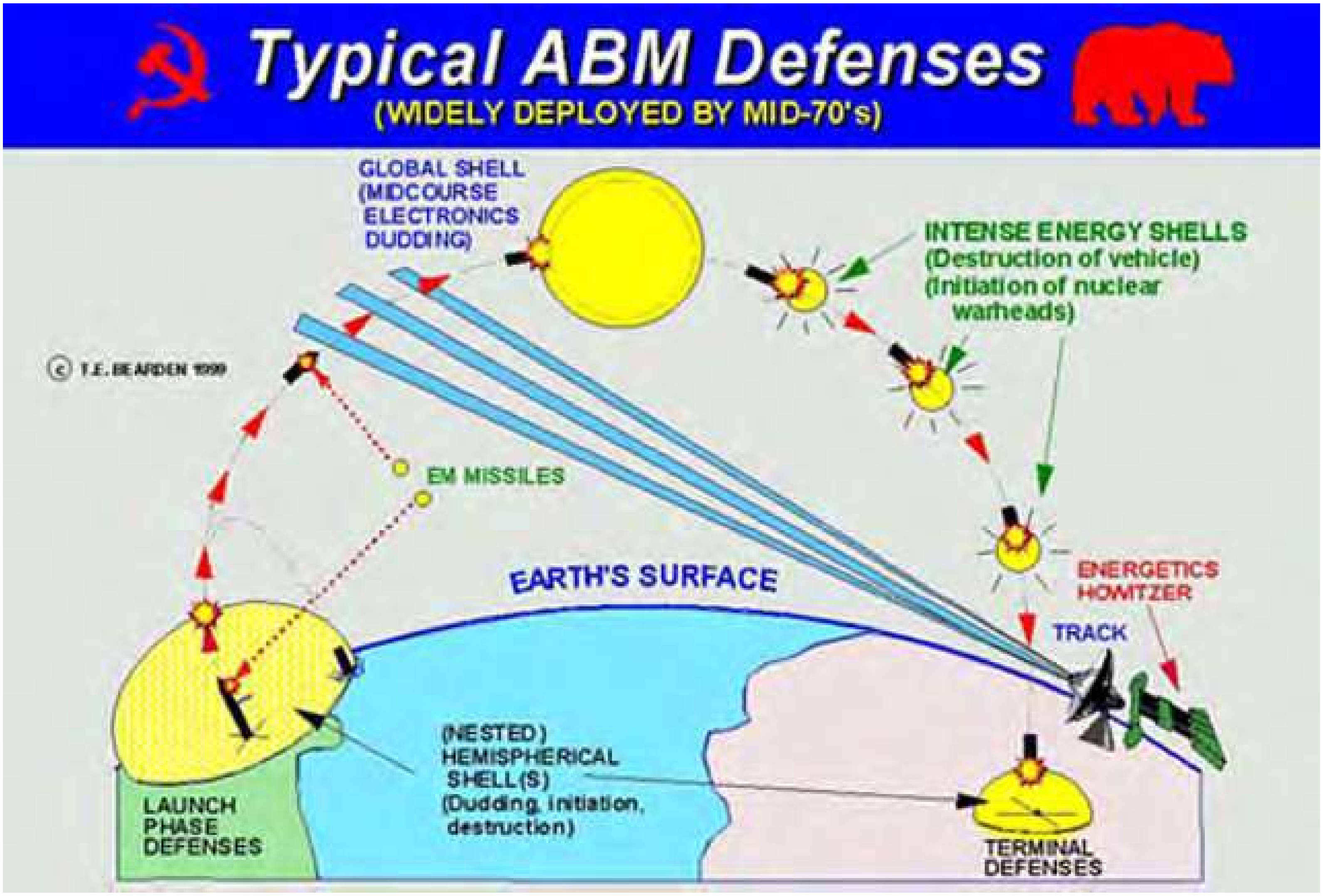

Typical ABM Defenses

(WIDELY DEPLOYED BY MID-70's)

GLOBAL SHELL
(MIDCOURSE
ELECTRONICS
DUDDING)
INTENSE ENERGY SHELLS
(Destruction of vehicle)
(Initiation of nuclear
warheads)
© T.E. BEARDEN 1999
EM MISSILES
EARTH'S SURFACE
ENERGETICS
HOWITZER
TRACK
(NESTED)
HEMISPHERICAL
SHELL(S)
(Dudding, initiation,
destruction)
LAUNCH
PHASE
DEFENSES
TERMINAL
DEFENSES

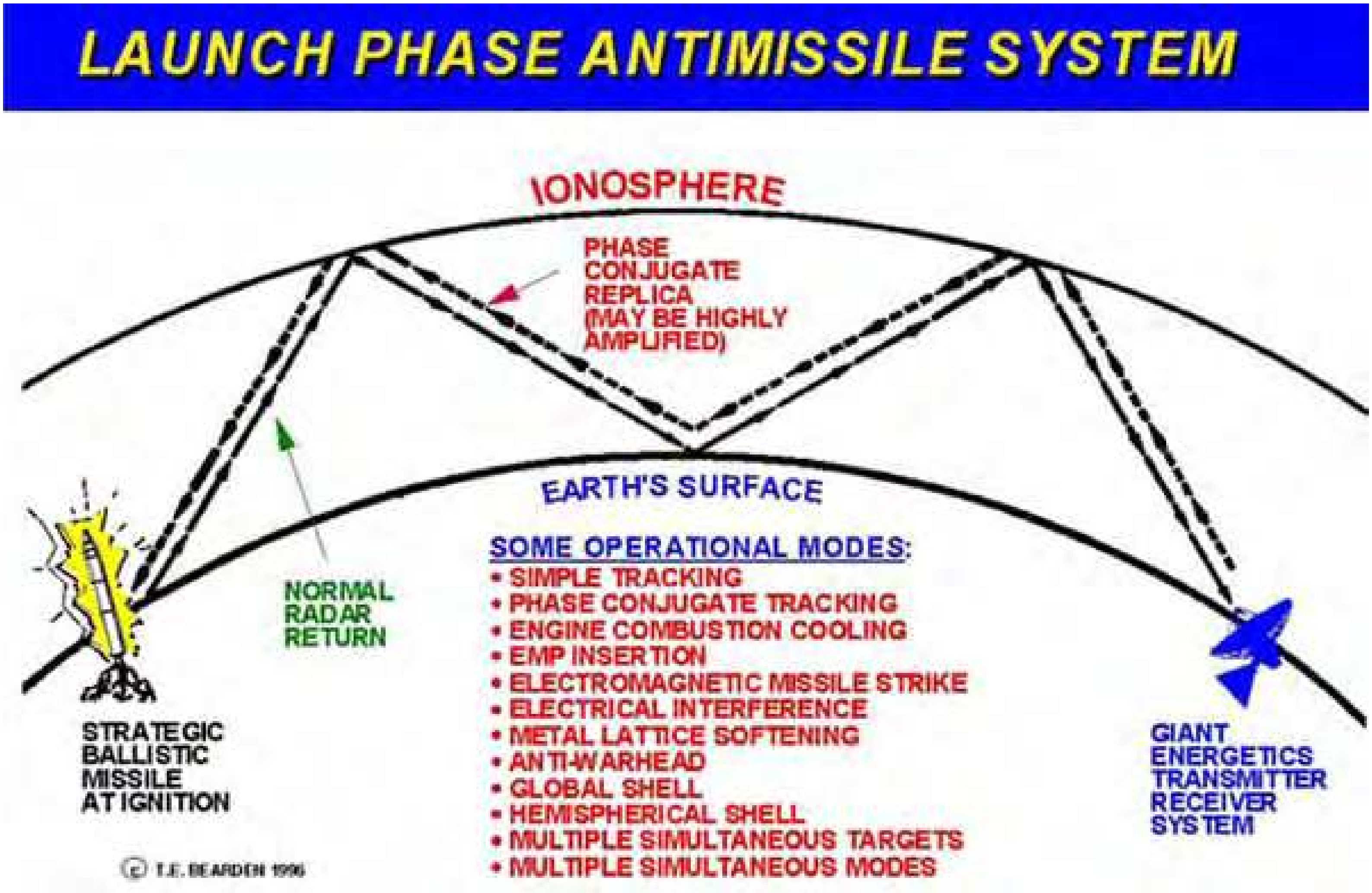

LAUNCH PHASE ANTIMISSILE SYSTEM

IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
NORMAL RADAR RETURN
EARTH'S SURFACE
STRATEGIC BALLISTIC MISSILE AT IGNITION
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-WARHEAD
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

© T.E. BEARDEN 1996

LAUNCH PHASE ANTI-BOMBER SYSTEM

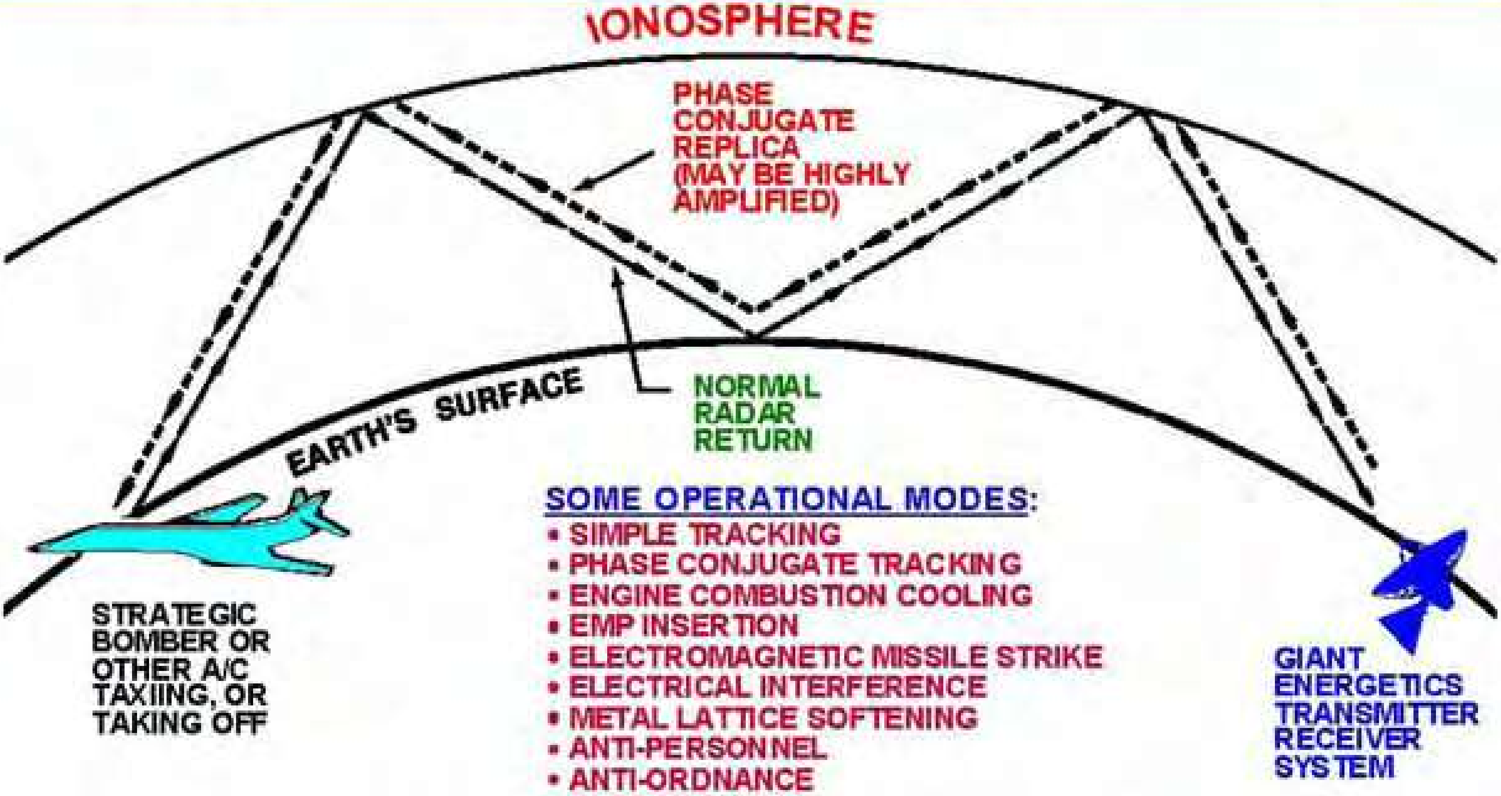
IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC BOMBER OR OTHER A/C TAXIING, OR TAKING OFF
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-PERSONNEL
- ANTI-ORDNANCE
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

© T.E. BEARDEN 1996

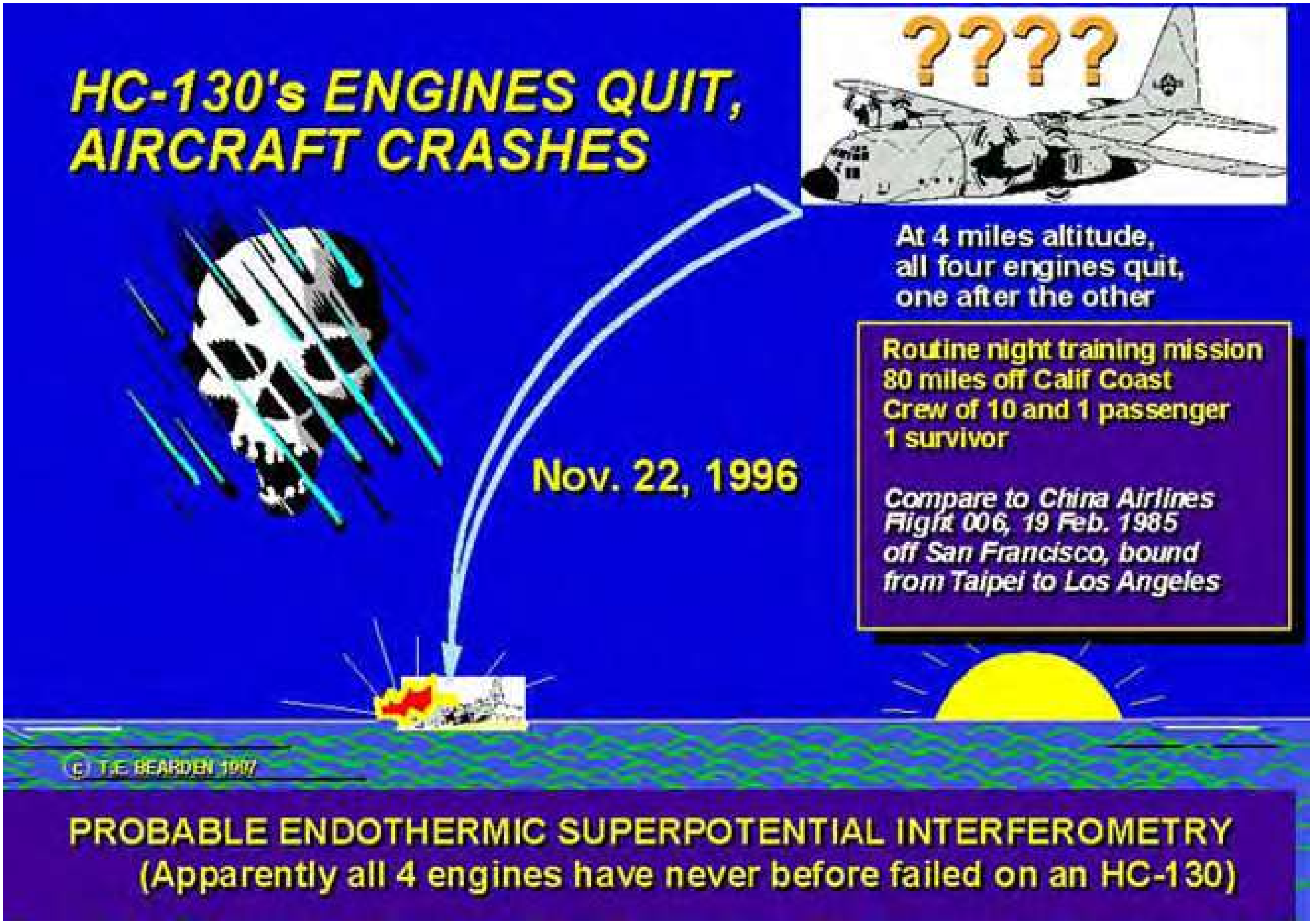

HC-130's ENGINES QUIT, AIRCRAFT CRASHES

At 4 miles altitude,
all four engines quit,
one after the other

Nov. 22, 1996

Routine night training mission
80 miles off Calif Coast
Crew of 10 and 1 passenger
1 survivor

Compare to China Airlines Flight 006, 19 Feb. 1985 off San Francisco, bound from Taipei to Los Angeles

(c) T.E. BEARDEN 1997

PROBABLE ENDOTHERMIC SUPERPOTENTIAL INTERFEROMETRY
(Apparently all 4 engines have never before failed on an HC-130)

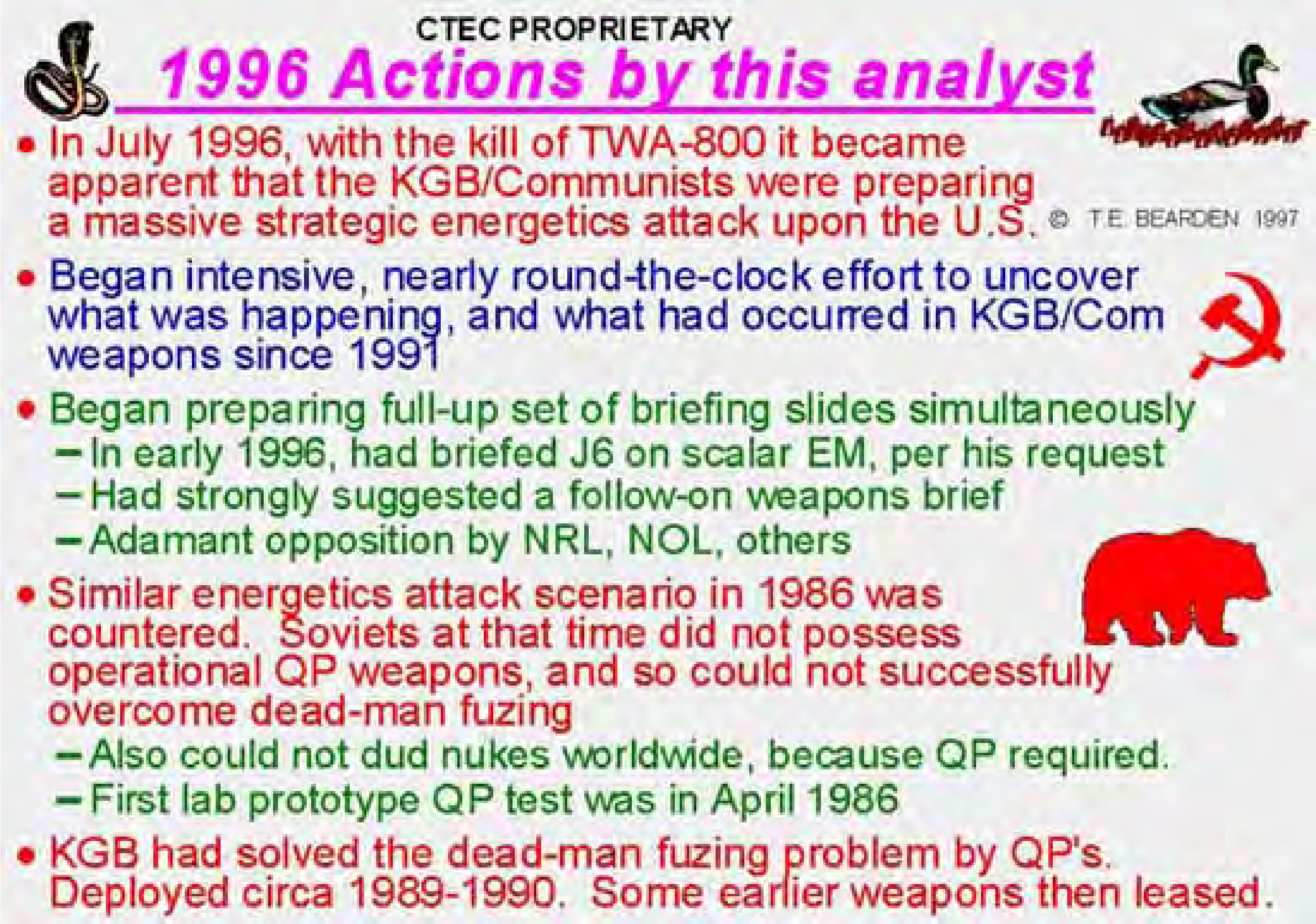

CTEC PROPRIETARY

1996 Actions by this analyst

- In July 1996, with the kill of TWA-800 it became apparent that the KGB/Communists were preparing a massive strategic energetics attack upon the U.S.

© T.E. BEARDEN 1997

- Began intensive, nearly round-the-clock effort to uncover what was happening, and what had occurred in KGB/Com weapons since 1991
- Began preparing full-up set of briefing slides simultaneously
  - In early 1996, had briefed J6 on scalar EM, per his request
  - Had strongly suggested a follow-on weapons brief
  - Adamant opposition by NRL, NOL, others
- Similar energetics attack scenario in 1986 was countered. Soviets at that time did not possess operational QP weapons, and so could not successfully overcome dead-man fuzing
  - Also could not dud nukes worldwide, because QP required.
  - First lab prototype QP test was in April 1986
- KGB had solved the dead-man fuzing problem by QP's. Deployed circa 1989-1990. Some earlier weapons then leased.

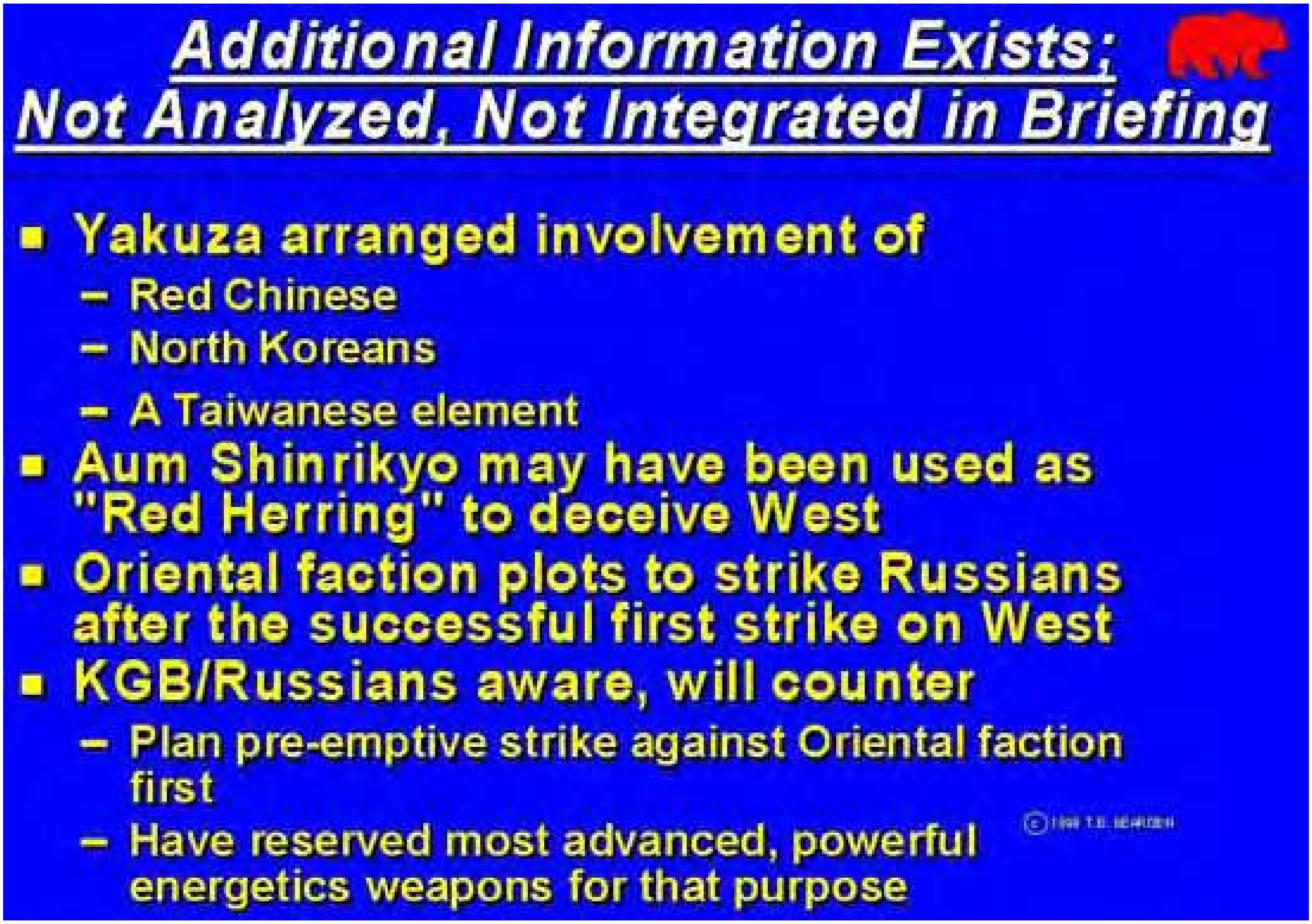

Additional Information Exists; Not Analyzed, Not Integrated in Briefing

- Yakuza arranged involvement of
  - Red Chinese
  - North Koreans
  - A Taiwanese element
- Aum Shinrikyo may have been used as "Red Herring" to deceive West
- Oriental faction plots to strike Russians after the successful first strike on West
- KGB/Russians aware, will counter
  - Plan pre-emptive strike against Oriental faction first
  - Have reserved most advanced, powerful energetics weapons for that purpose

© 1998 T.E. BEARDEN

The Tom Bearden Website

Help support the research

Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6

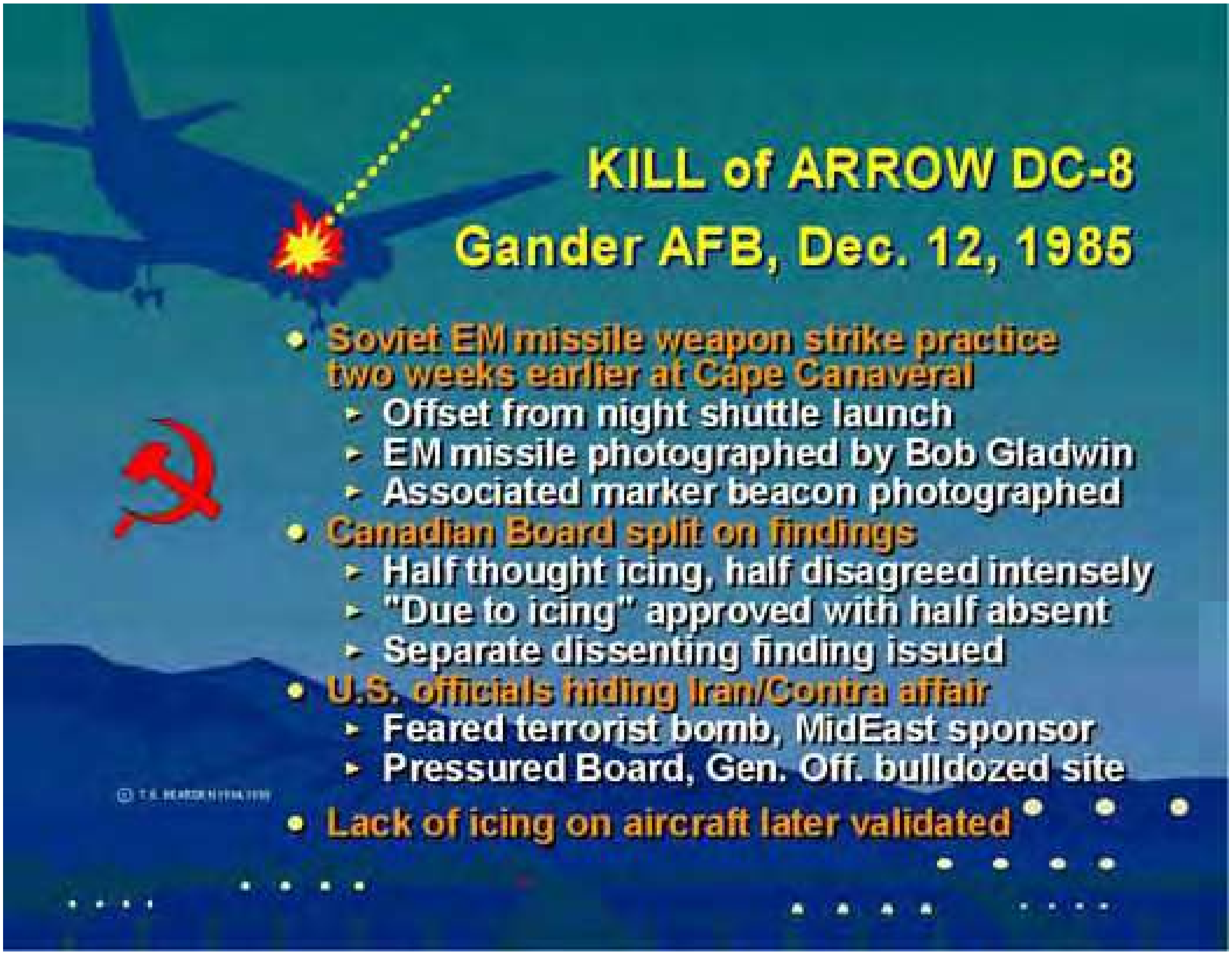

KILL of ARROW DC-8

Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. bulldozed site
- Lack of icing on aircraft later validated

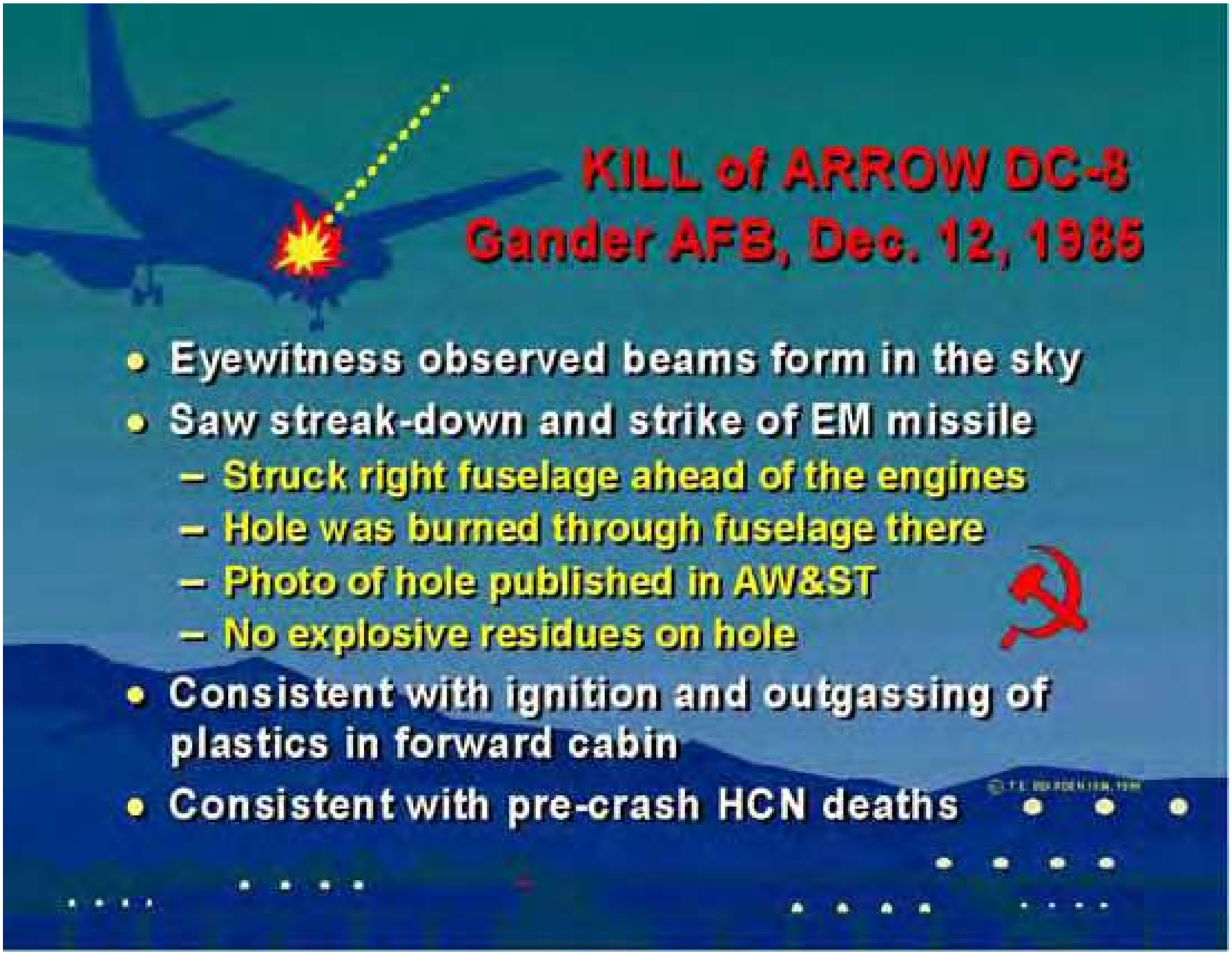

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Eyewitness observed beams form in the sky
- Saw streak-down and strike of EM missile
  - Struck right fuselage ahead of the engines
  - Hole was burned through fuselage there
  - Photo of hole published in AW&ST
  - No explosive residues on hole
- Consistent with ignition and outgassing of plastics in forward cabin
- Consistent with pre-crash HCN deaths

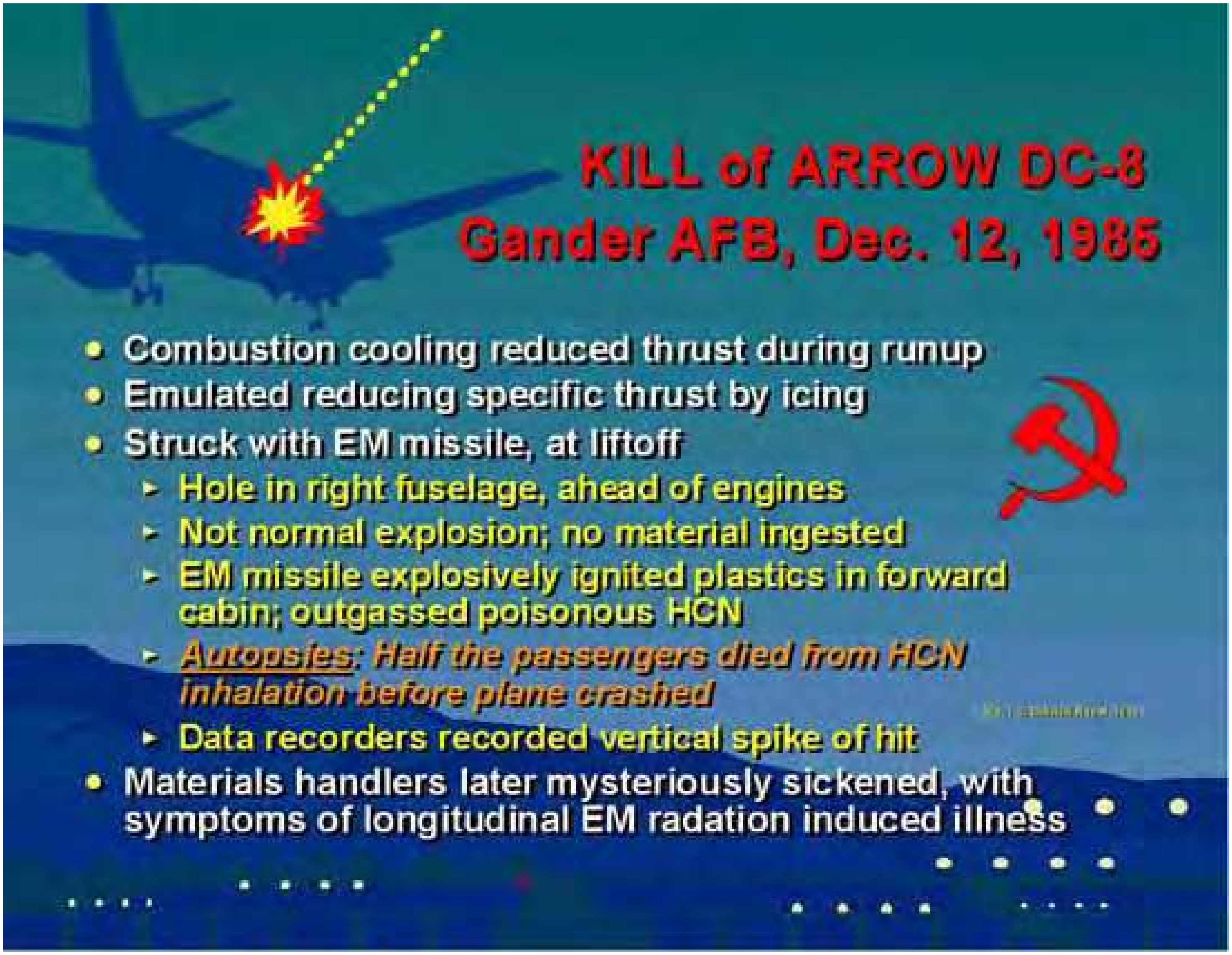

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *__Autopsies__: Half the passengers died from HCN inhalation before plane crashed*
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of longitudinal EM radation induced illness

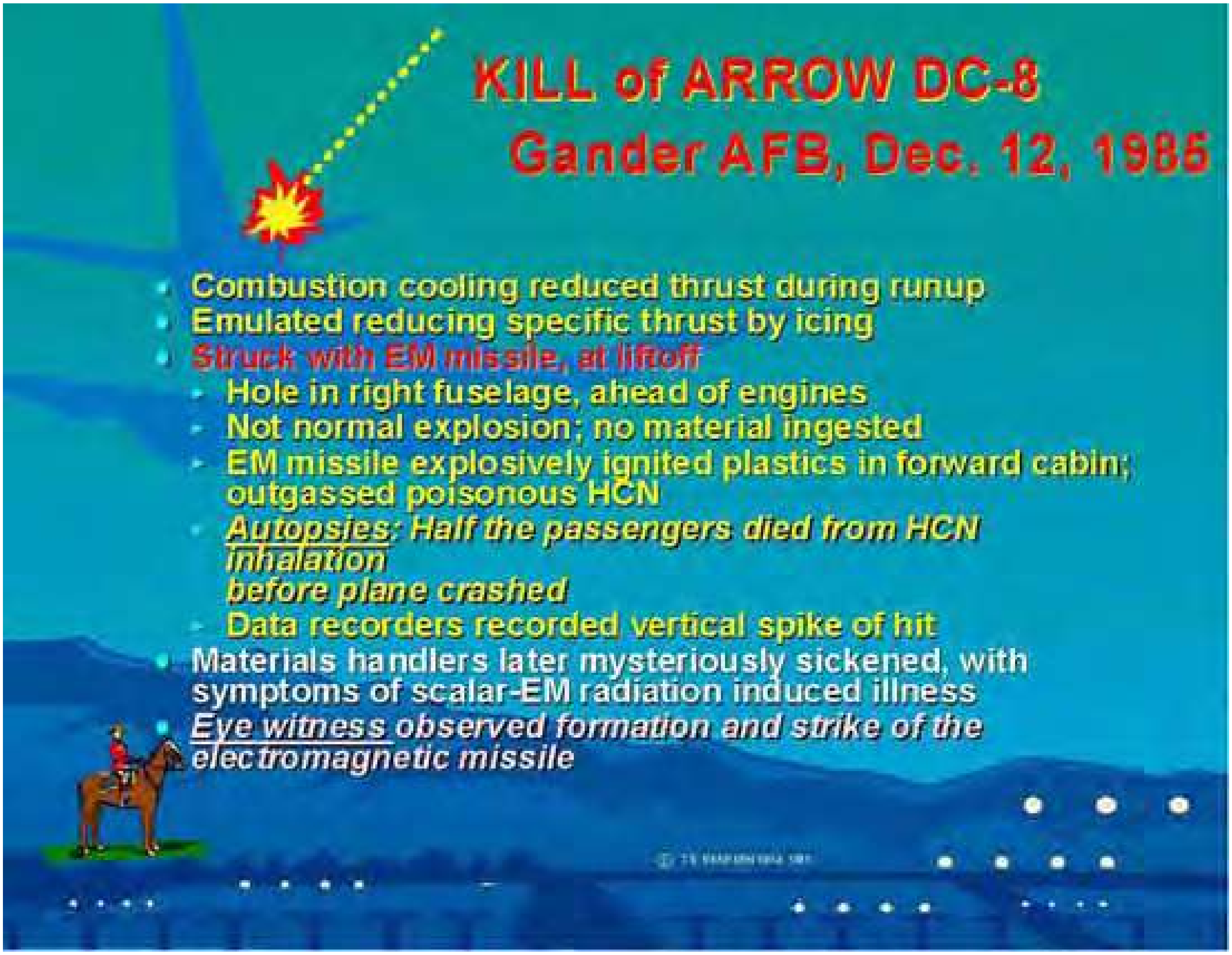

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation*
    before plane crashed
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of scalar-EM radiation induced illness
- *Eye witness observed formation and strike of the electromagnetic missile*

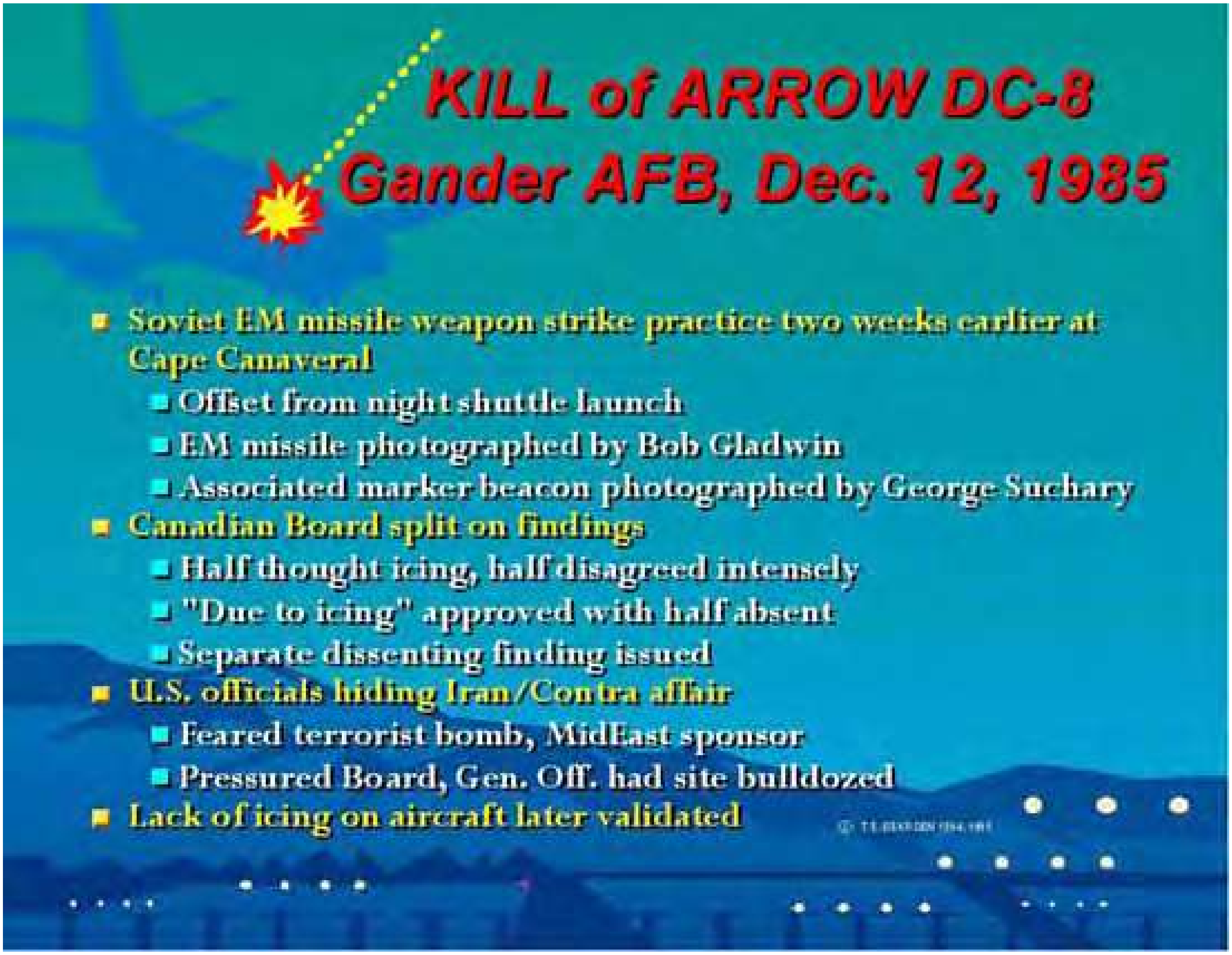

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed by George Suchary
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. had site bulldozed
- Lack of icing on aircraft later validated

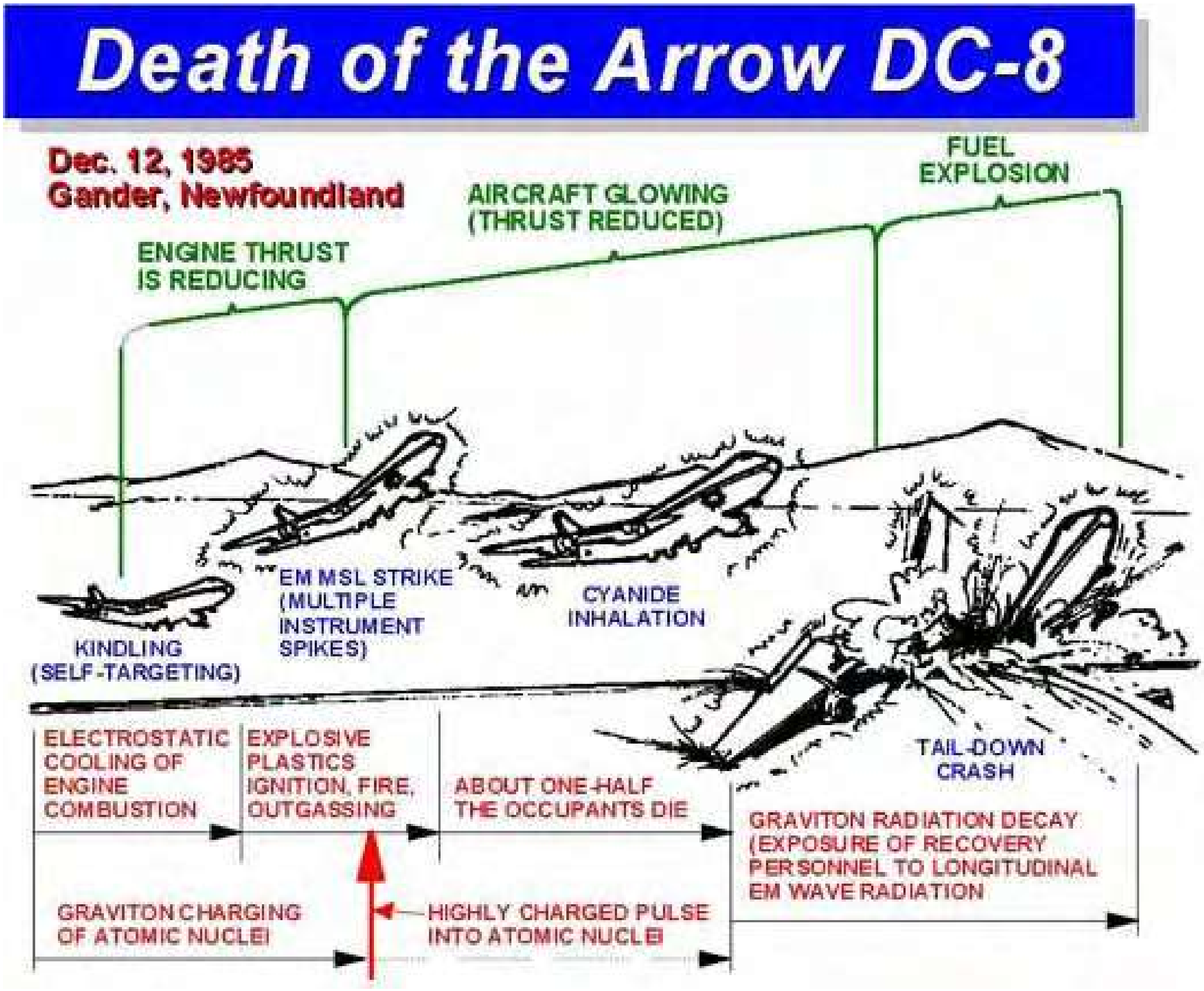
Death of the Arrow DC-8
Dec. 12, 1985
Gander, Newfoundland
ENGINE THRUST IS REDUCING
AIRCRAFT GLOWING (THRUST REDUCED)
FUEL EXPLOSION
KINDLING (SELF-TARGETING)
EM MSL STRIKE (MULTIPLE INSTRUMENT SPIKES)
CYANIDE INHALATION
TAIL-DOWN CRASH
ELECTROSTATIC COOLING OF ENGINE COMBUSTION
EXPLOSIVE PLASTICS IGNITION, FIRE, OUTGASSING
ABOUT ONE-HALF THE OCCUPANTS DIE
GRAVITON RADIATION DECAY (EXPOSURE OF RECOVERY PERSONNEL TO LONGITUDINAL EM WAVE RADIATION
GRAVITON CHARGING OF ATOMIC NUCLEI
HIGHLY CHARGED PULSE INTO ATOMIC NUCLEI

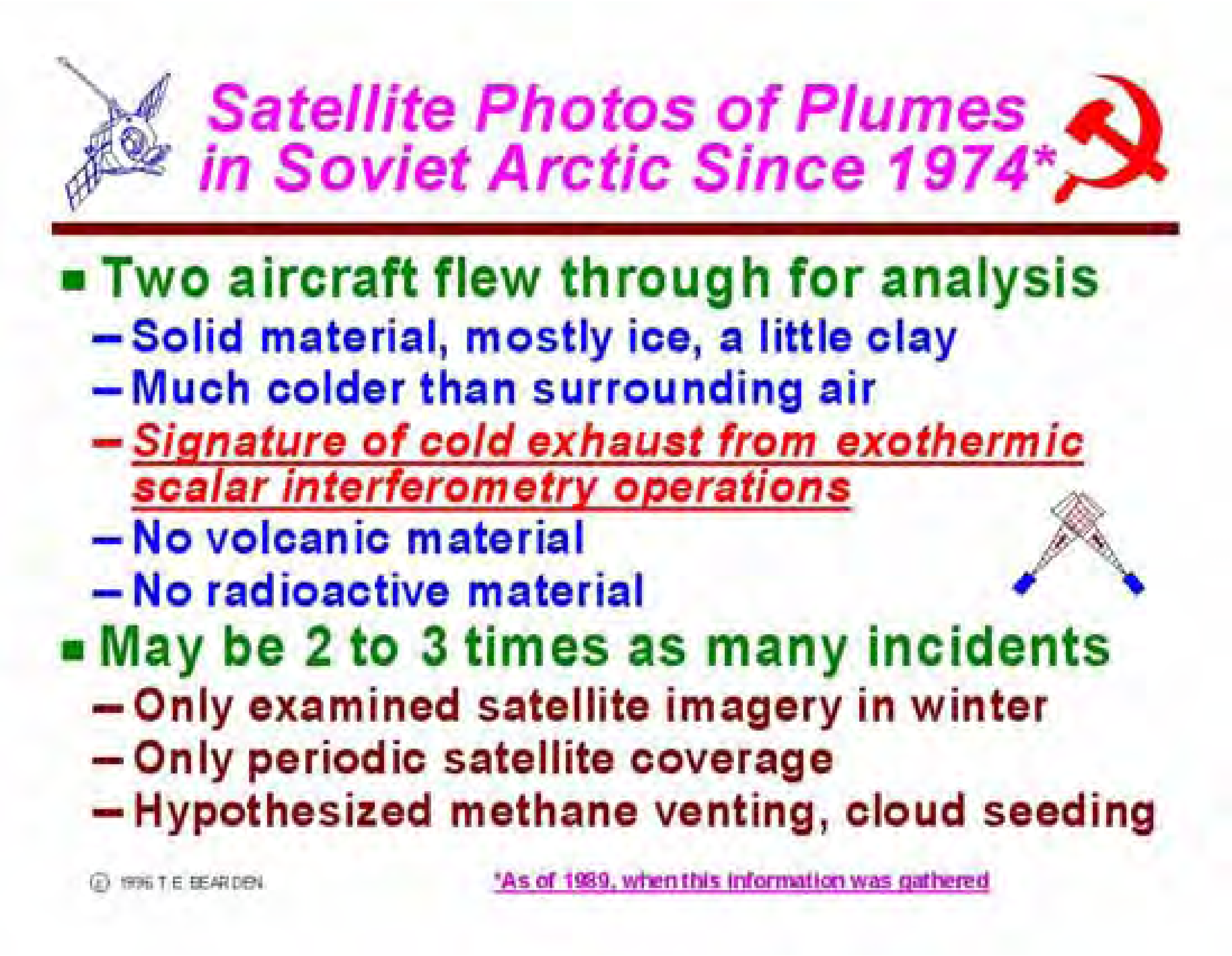

*Satellite Photos of Plumes in Soviet Arctic Since 1974**

- Two aircraft flew through for analysis
  - Solid material, mostly ice, a little clay
  - Much colder than surrounding air
  - *Signature of cold exhaust from exothermic scalar interferometry operations*
  - No volcanic material
  - No radioactive material
- May be 2 to 3 times as many incidents
  - Only examined satellite imagery in winter
  - Only periodic satellite coverage
  - Hypothesized methane venting, cloud seeding

© 1996 T.E. BEARDEN

*As of 1989, when this information was gathered

AUM SHINRIKYO

("SUPREME TRUTH" CULT)

Shoko Asahara, Leader

© 1995, 1998 T.E. BEARDEN

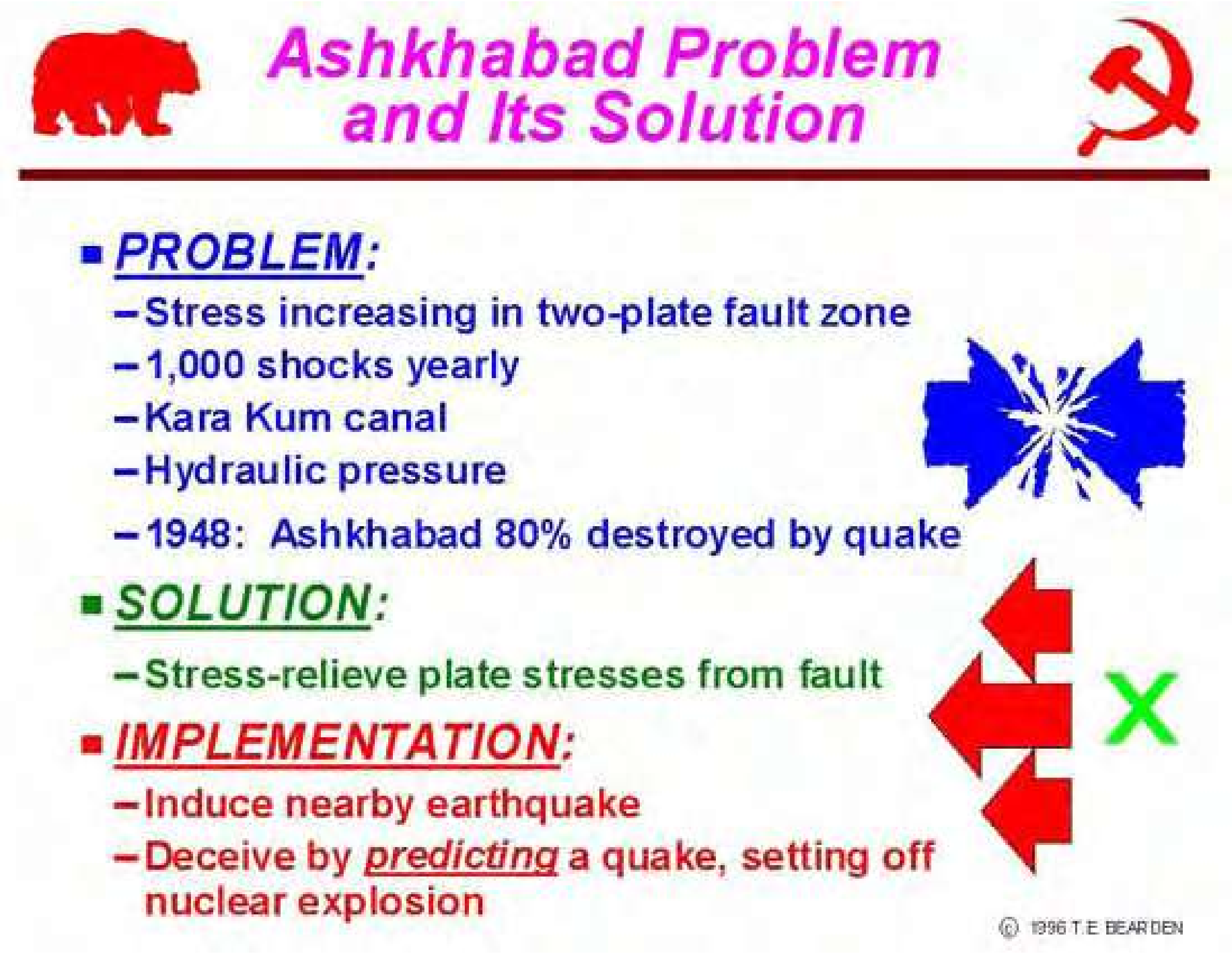

Ashkhabad Problem and Its Solution

- ***PROBLEM:***
  - Stress increasing in two-plate fault zone
  - 1,000 shocks yearly
  - Kara Kum canal
  - Hydraulic pressure
  - 1948: Ashkhabad 80% destroyed by quake
- ***SOLUTION:***
  - Stress-relieve plate stresses from fault
- ***IMPLEMENTATION:***
  - Induce nearby earthquake
  - Deceive by *predicting* a quake, setting off nuclear explosion

© 1996 T.E. BEARDEN

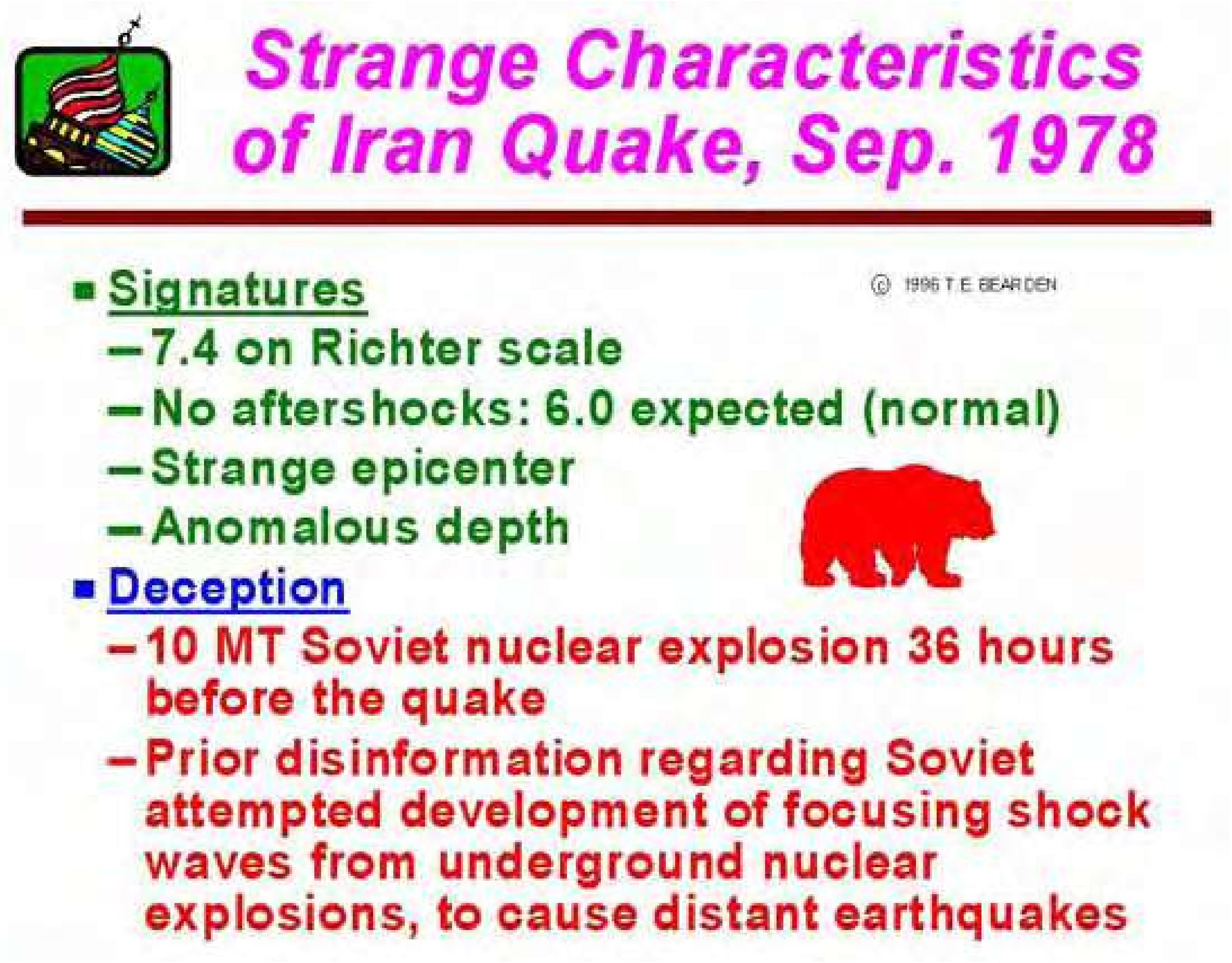

Strange Characteristics of Iran Quake, Sep. 1978

© 1996 T.E. BEARDEN

- Signatures
  - 7.4 on Richter scale
  - No aftershocks: 6.0 expected (normal)
  - Strange epicenter
  - Anomalous depth
- Deception
  - 10 MT Soviet nuclear explosion 36 hours before the quake
  - Prior disinformation regarding Soviet attempted development of focusing shock waves from underground nuclear explosions, to cause distant earthquakes

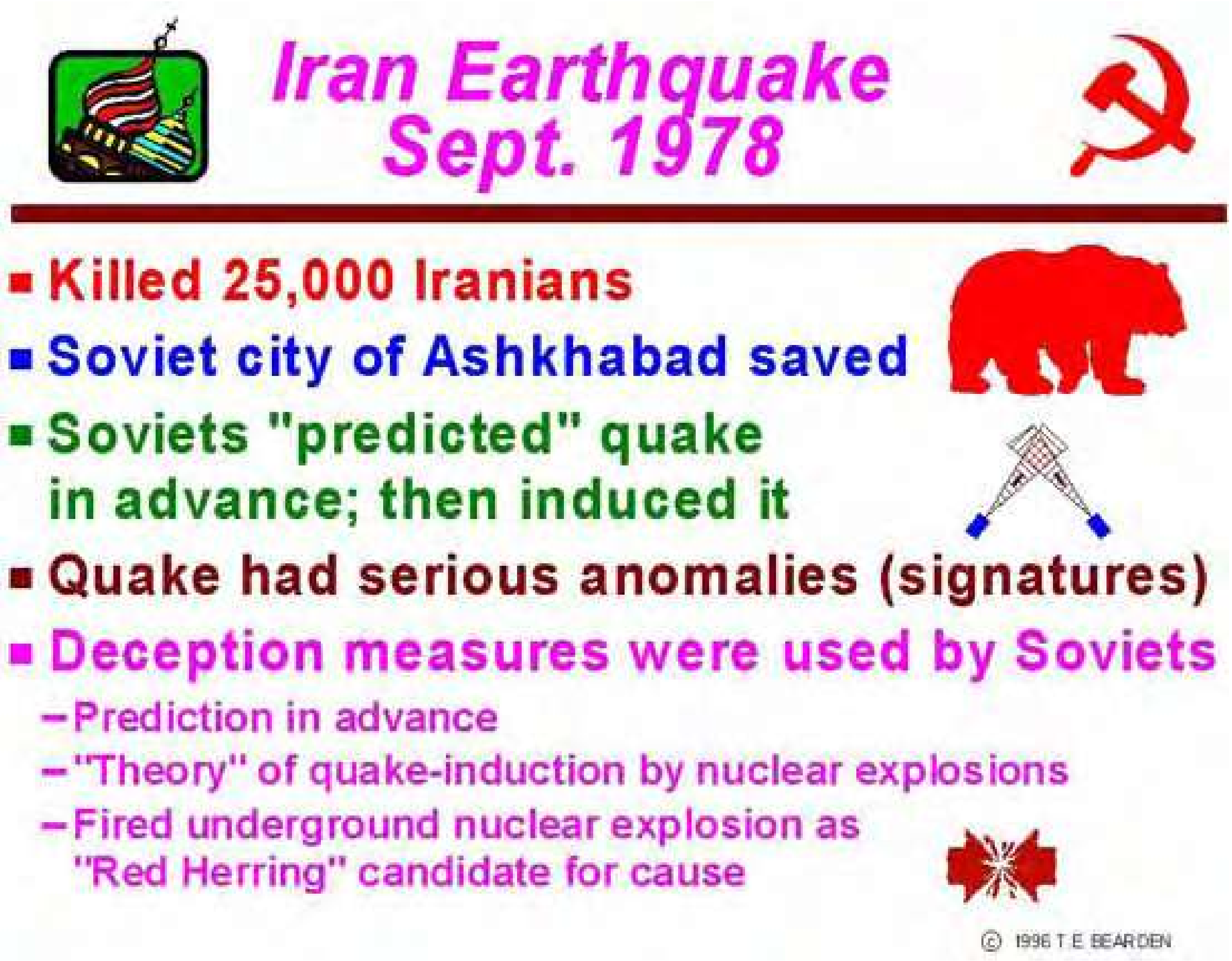

Iran Earthquake Sept. 1978

- Killed 25,000 Iranians
- Soviet city of Ashkhabad saved
- Soviets "predicted" quake in advance; then induced it
- Quake had serious anomalies (signatures)
- Deception measures were used by Soviets
  - Prediction in advance
  - "Theory" of quake-induction by nuclear explosions
  - Fired underground nuclear explosion as "Red Herring" candidate for cause

© 1996 T.E. BEARDEN

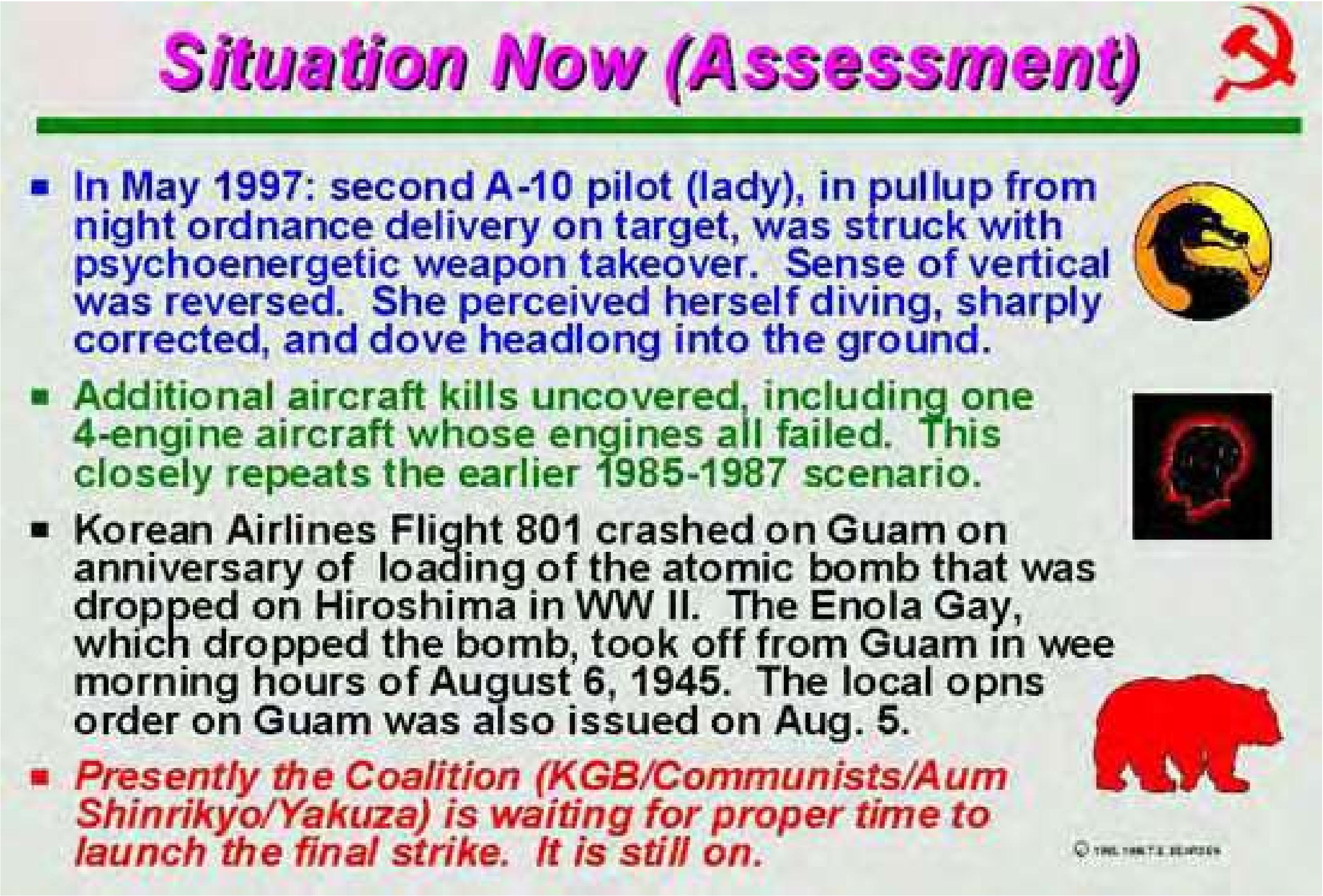

Situation Now (Assessment)

- In May 1997: second A-10 pilot (lady), in pullup from night ordnance delivery on target, was struck with psychoenergetic weapon takeover. Sense of vertical was reversed. She perceived herself diving, sharply corrected, and dove headlong into the ground.
- Additional aircraft kills uncovered, including one 4-engine aircraft whose engines all failed. This closely repeats the earlier 1985-1987 scenario.
- Korean Airlines Flight 801 crashed on Guam on anniversary of loading of the atomic bomb that was dropped on Hiroshima in WW II. The Enola Gay, which dropped the bomb, took off from Guam in wee morning hours of August 6, 1945. The local opns order on Guam was also issued on Aug. 5.
- *Presently the Coalition (KGB/Communists/Aum Shinrikyo/Yakuza) is waiting for proper time to launch the final strike. It is still on.*

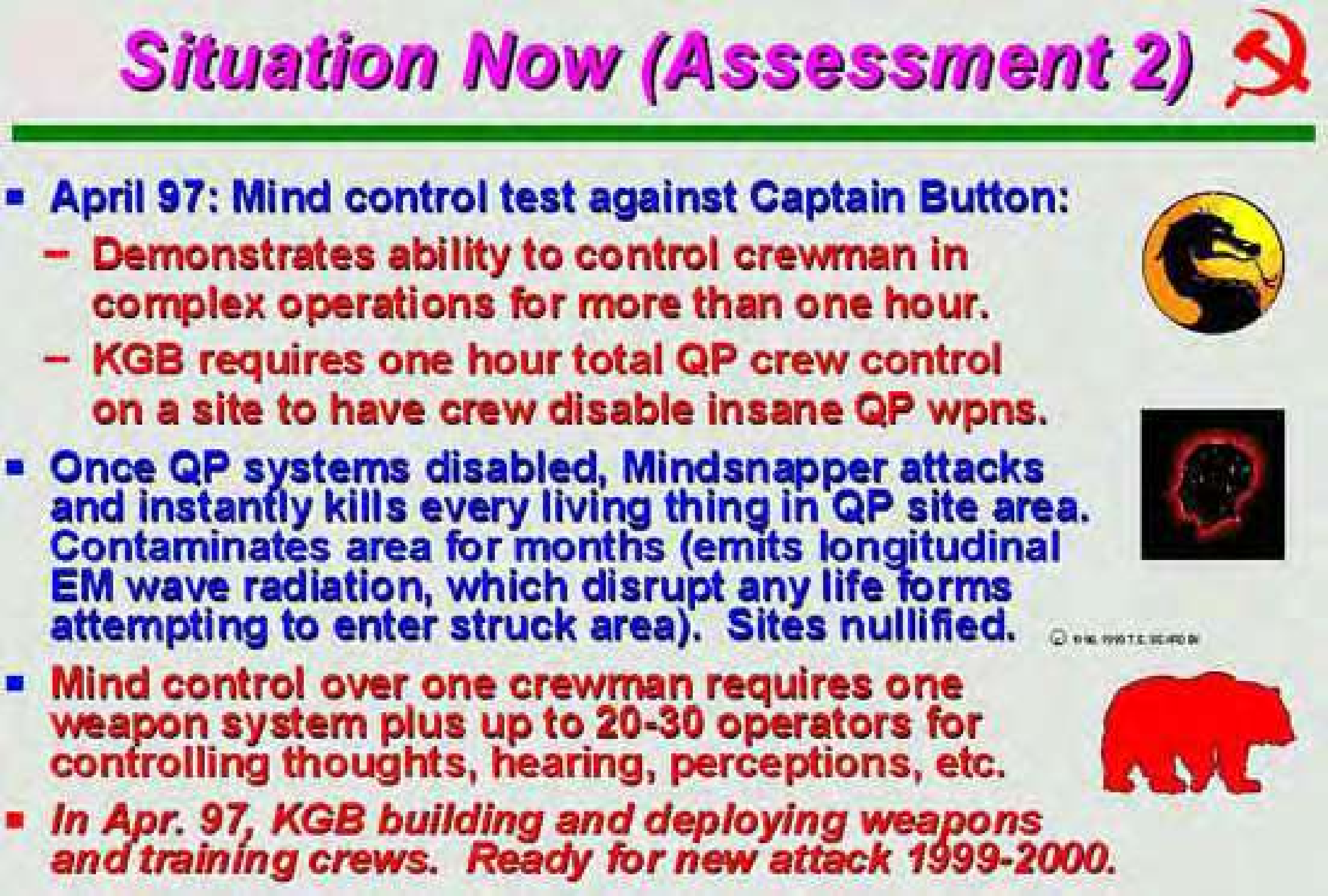

Situation Now (Assessment 2)

- April 97: Mind control test against Captain Button:
  - Demonstrates ability to control crewman in complex operations for more than one hour.
  - KGB requires one hour total QP crew control on a site to have crew disable insane QP wpns.
- Once QP systems disabled, Mindsnapper attacks and instantly kills every living thing in QP site area. Contaminates area for months (emits longitudinal EM wave radiation, which disrupt any life forms attempting to enter struck area). Sites nullified.
- Mind control over one crewman requires one weapon system plus up to 20-30 operators for controlling thoughts, hearing, perceptions, etc.
- *In Apr. 97, KGB building and deploying weapons and training crews. Ready for new attack 1999-2000.*

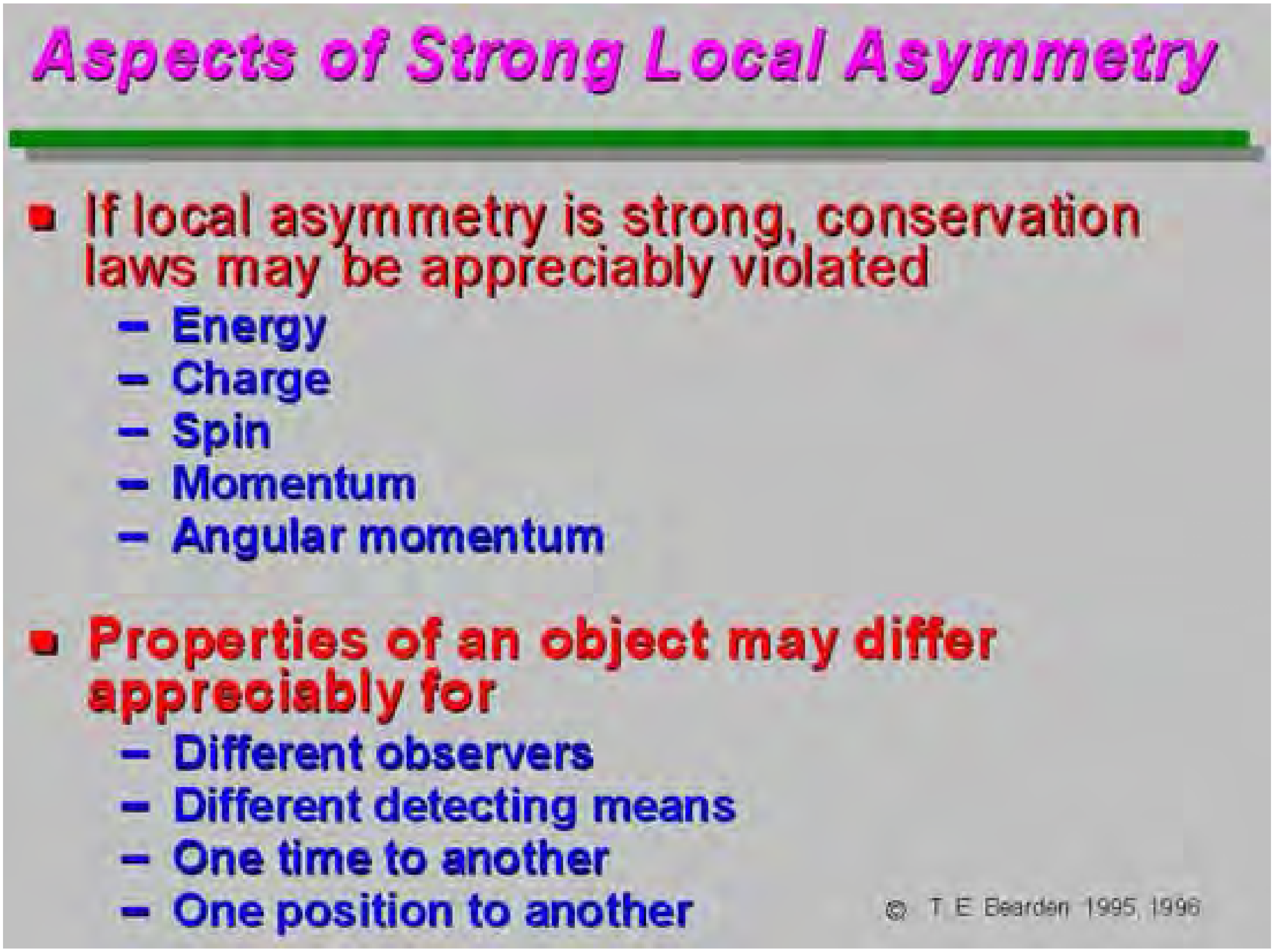

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

© T. E. Bearden 1995, 1996

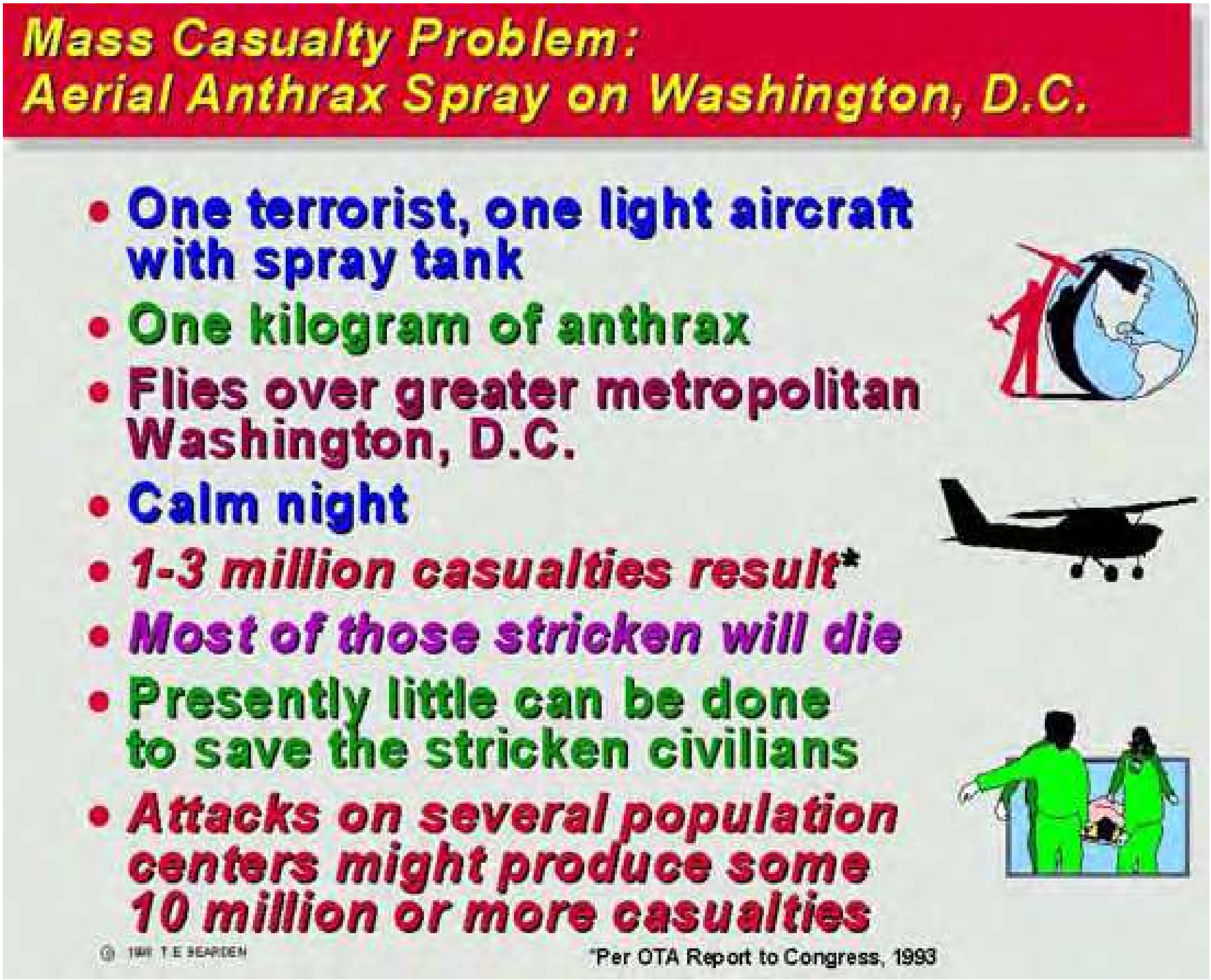

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- One kilogram of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

© [illegible] T E BEARDEN

*Per OTA Report to Congress, 1993

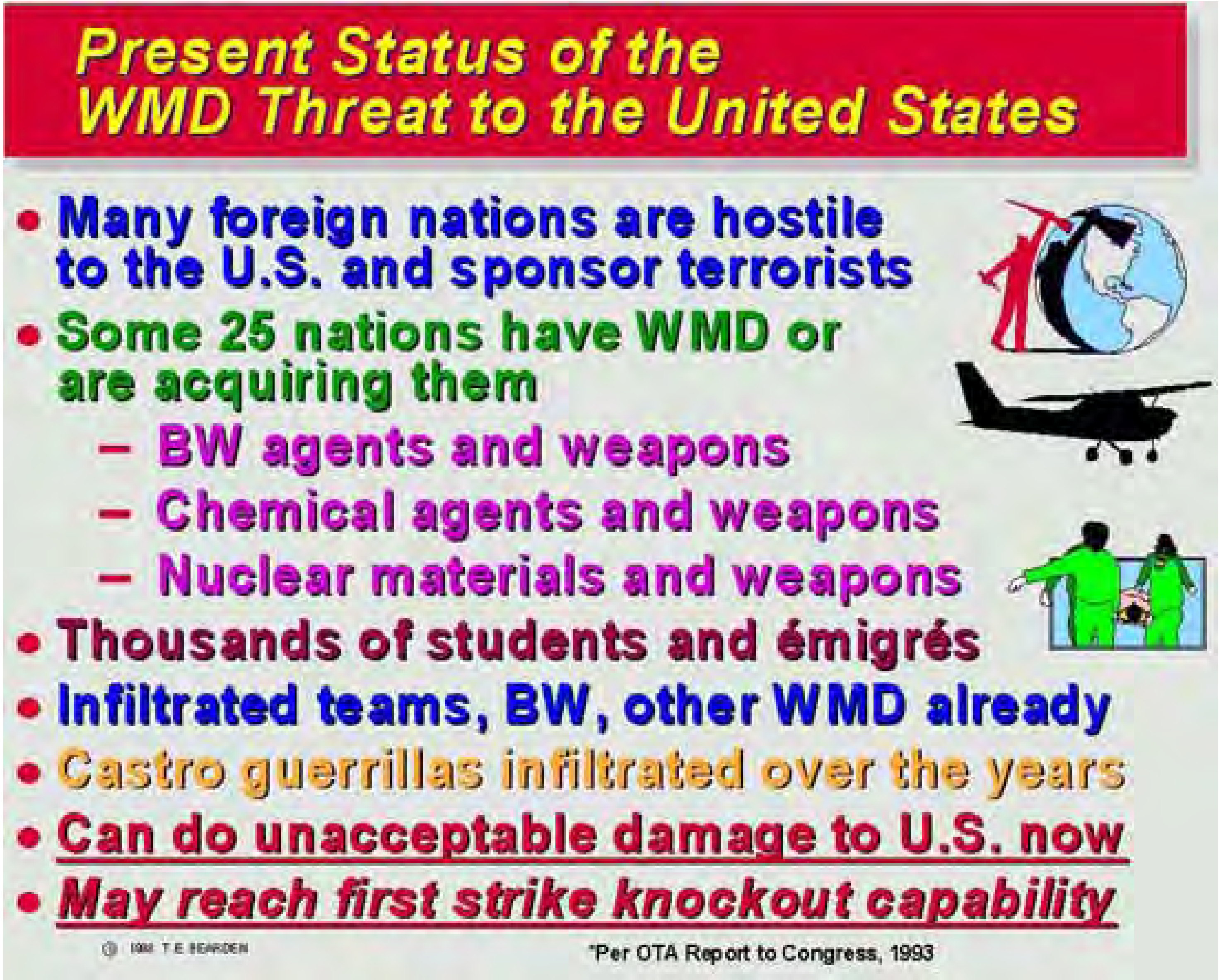

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
  - BW agents and weapons
  - Chemical agents and weapons
  - Nuclear materials and weapons
- Thousands of students and émigrés
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© [illegible] T.E. BEARDEN

*Per OTA Report to Congress, 1993

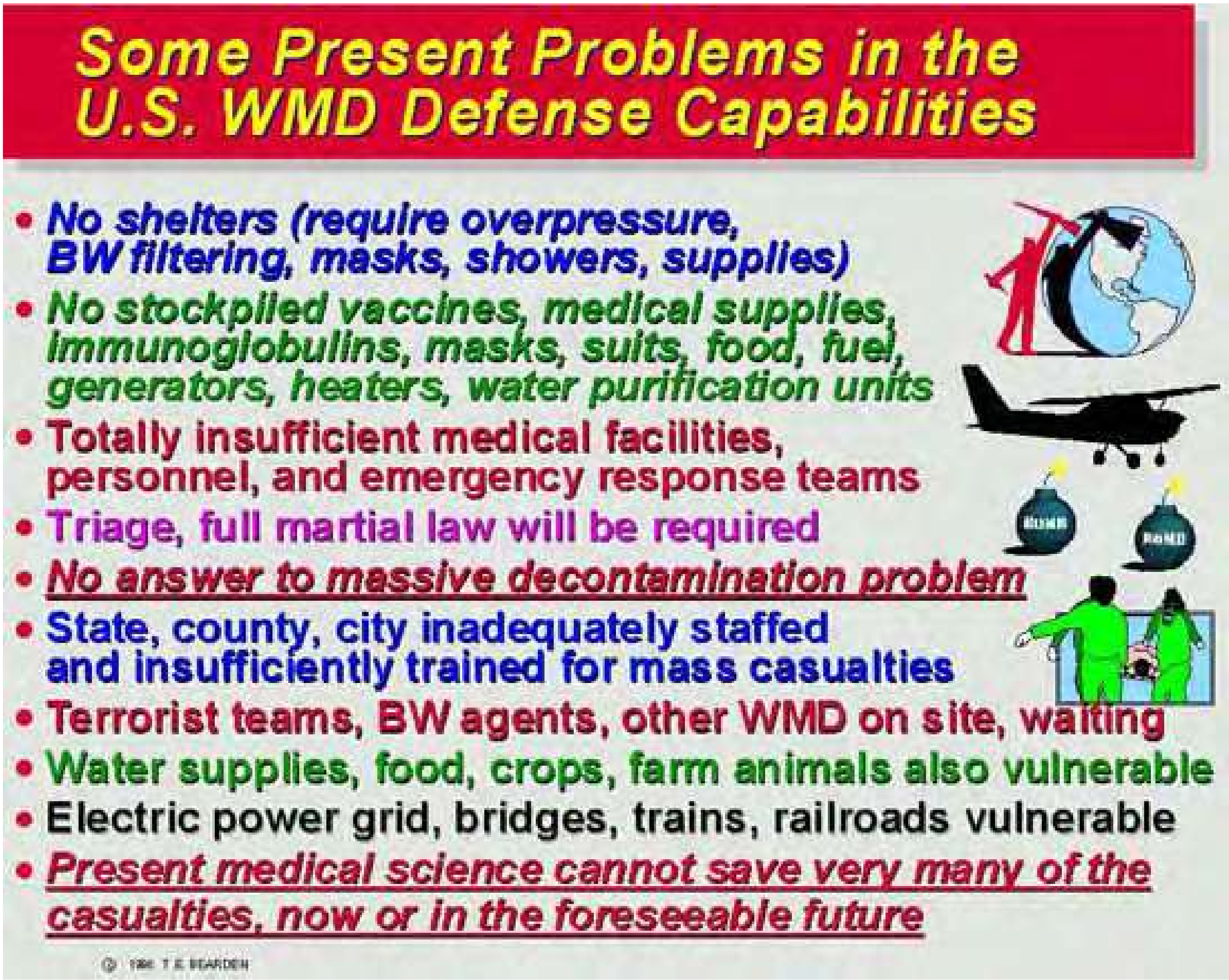

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© 1996 T.E. BEARDEN

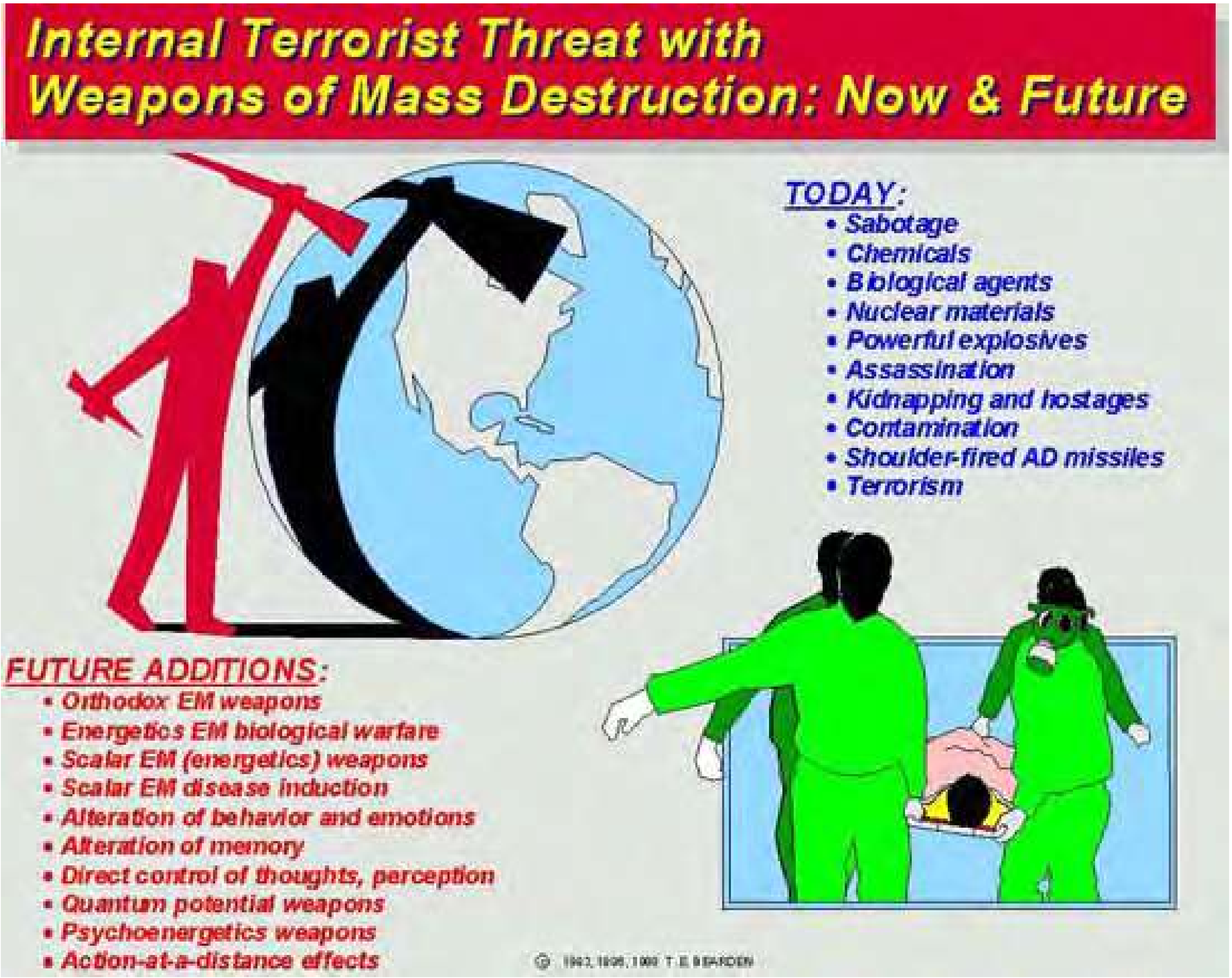

Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future

TODAY:
- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:
- Orthodox EM weapons
- Energetics EM biological warfare
- Scalar EM (energetics) weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct control of thoughts, perception
- Quantum potential weapons
- Psychoenergetics weapons
- Action-at-a-distance effects

© 1993, 1996, 1999 T. E. BEARDEN

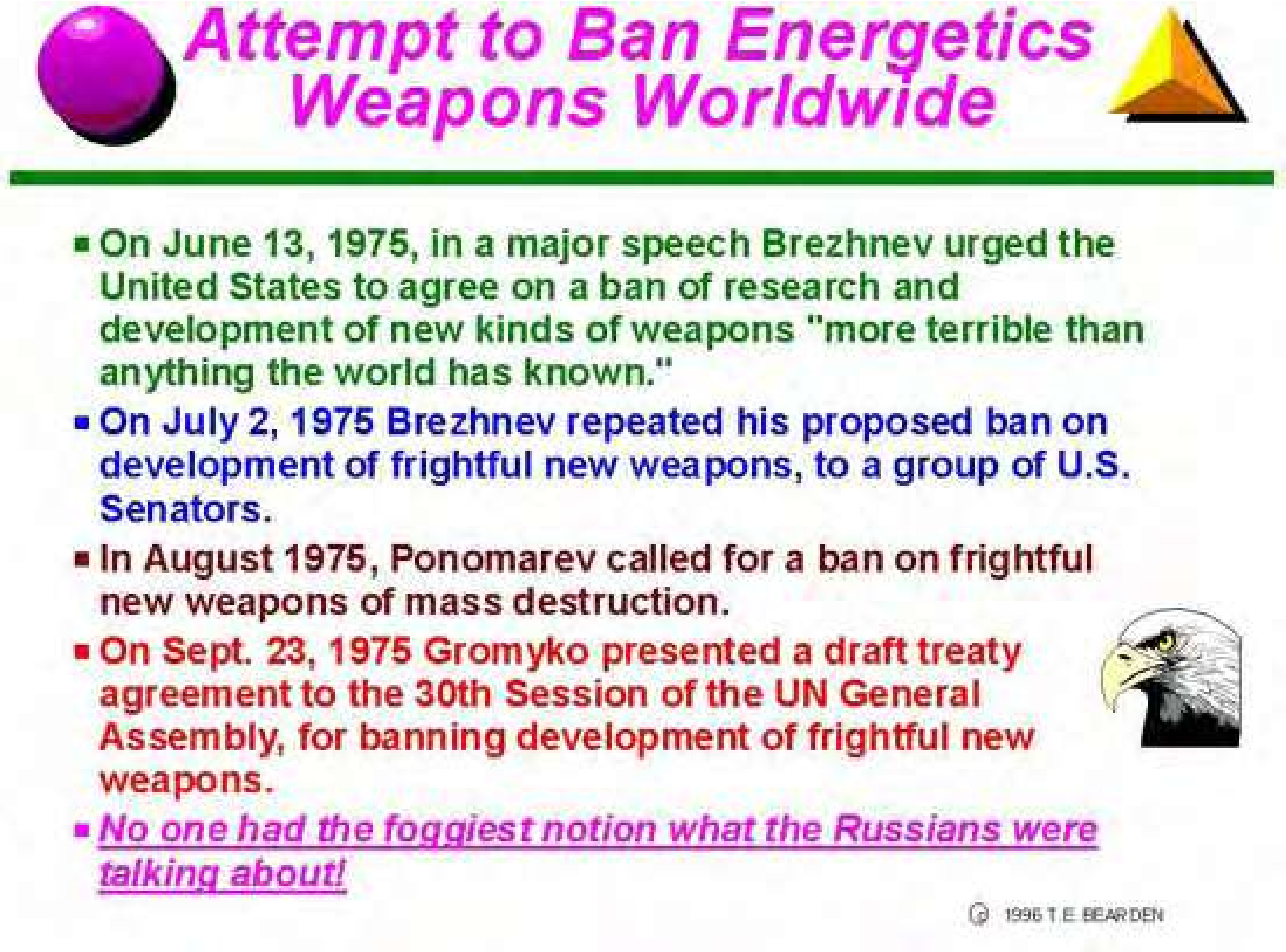

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T. E. BEARDEN

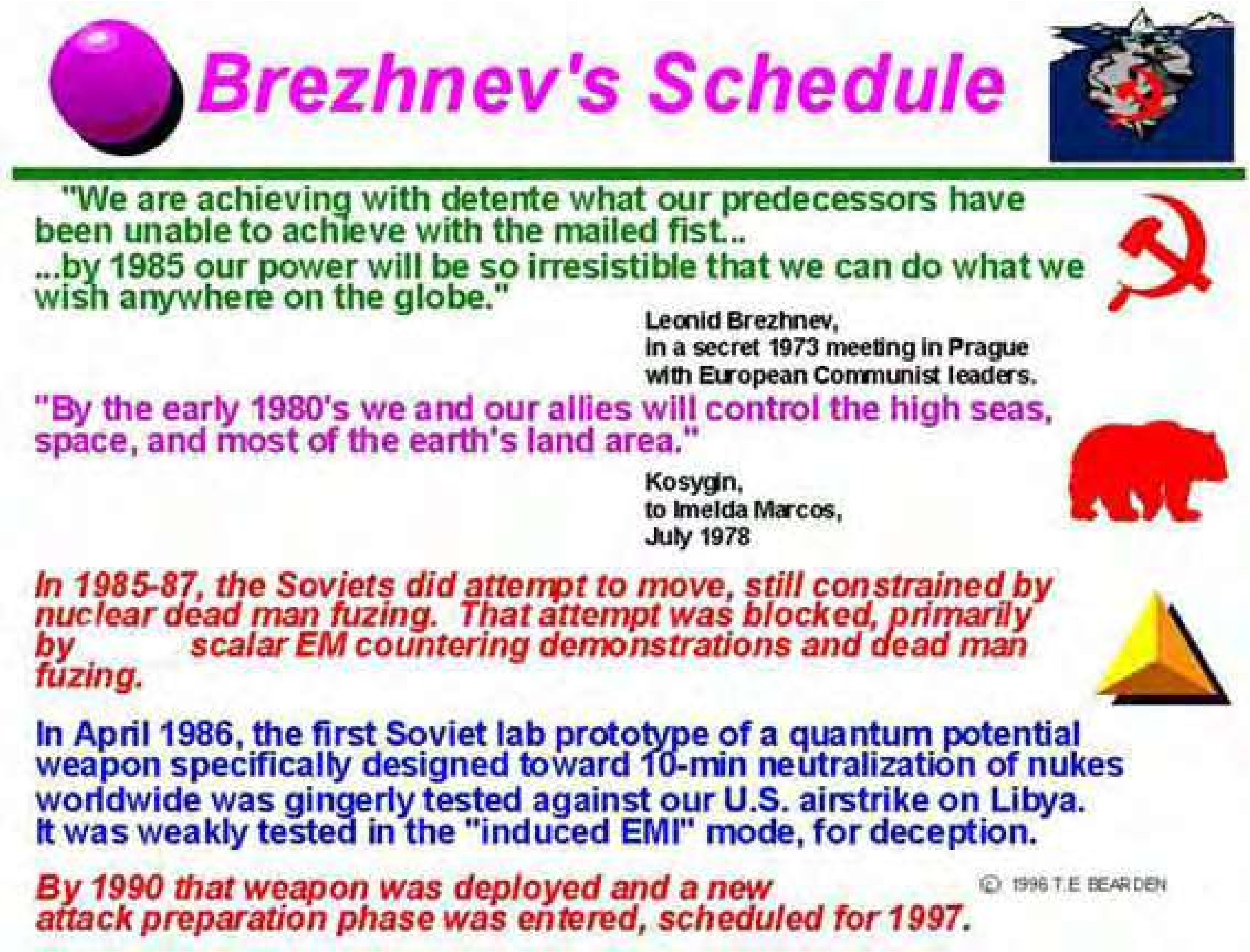

Brezhnev's Schedule

"We are achieving with detente what our predecessors have been unable to achieve with the mailed fist..
...by 1985 our power will be so irresistible that we can do what we wish anywhere on the globe."

Leonid Brezhnev,
in a secret 1973 meeting in Prague
with European Communist leaders.

"By the early 1980's we and our allies will control the high seas, space, and most of the earth's land area."

Kosygin,
to Imelda Marcos,
July 1978

In 1985-87, the Soviets did attempt to move, still constrained by nuclear dead man fuzing. That attempt was blocked, primarily by scalar EM countering demonstrations and dead man fuzing.

In April 1986, the first Soviet lab prototype of a quantum potential weapon specifically designed toward 10-min neutralization of nukes worldwide was gingerly tested against our U.S. airstrike on Libya. It was weakly tested in the "induced EMI" mode, for deception.

By 1990 that weapon was deployed and a new attack preparation phase was entered, scheduled for 1997.

© 1996 T.E. BEARDEN

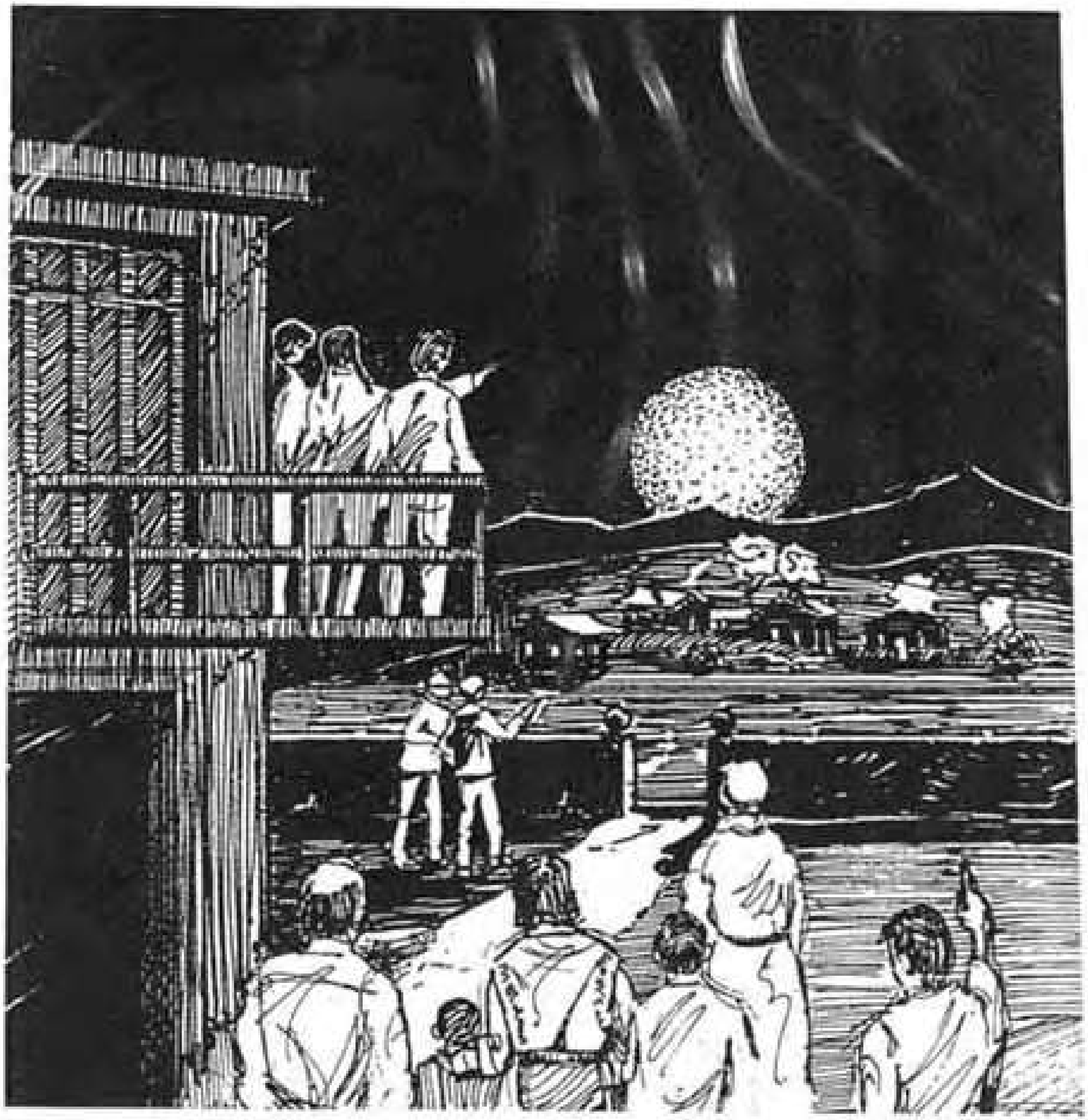

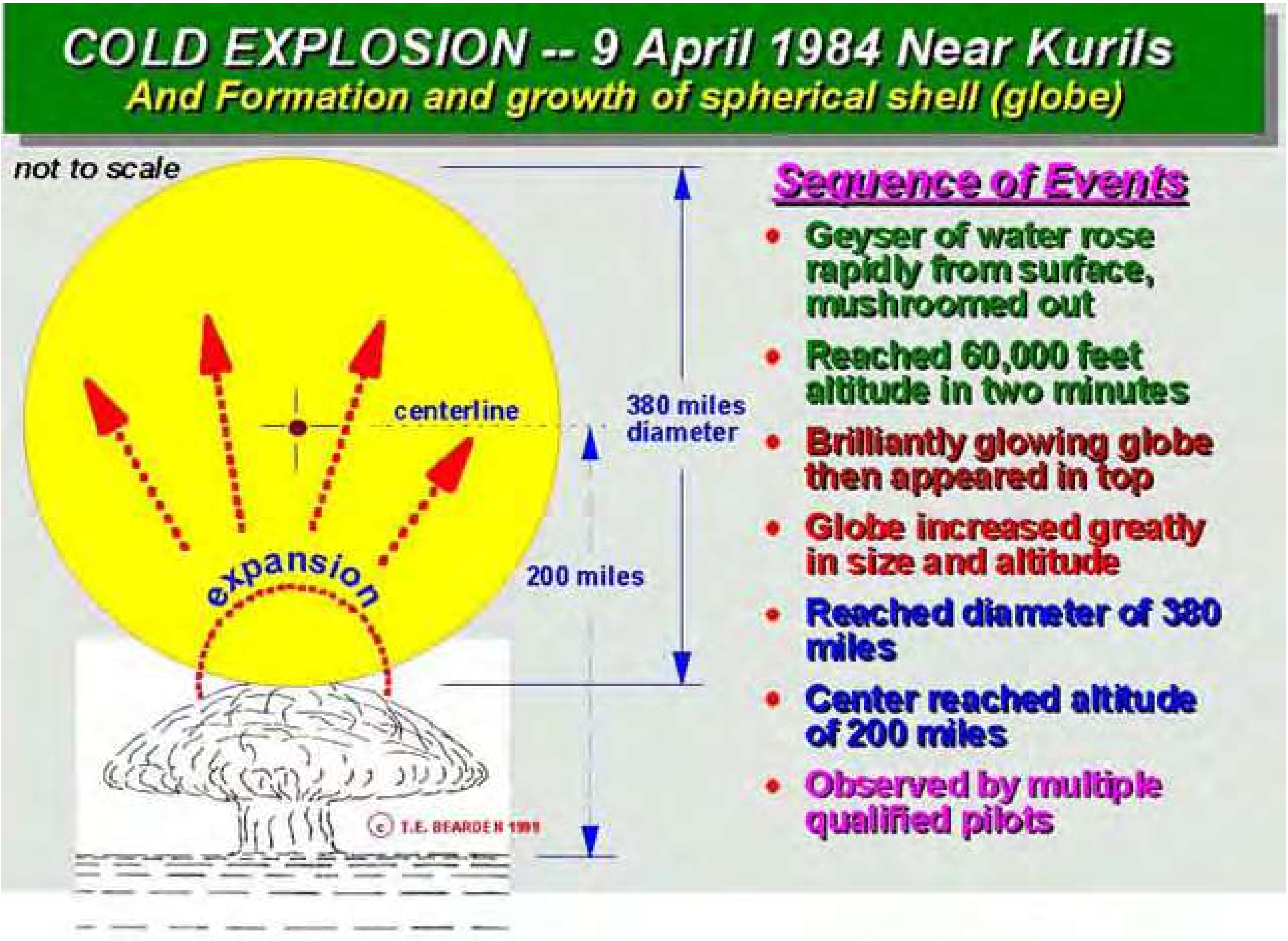

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
380 miles diameter
200 miles
expansion
© T.E. BEARDEN 1991

Sequence of Events

- Geyser of water rose rapidly from surface, mushroomed out
- Reached 60,000 feet altitude in two minutes
- Brilliantly glowing globe then appeared in top
- Globe increased greatly in size and altitude
- Reached diameter of 380 miles
- Center reached altitude of 200 miles
- Observed by multiple qualified pilots

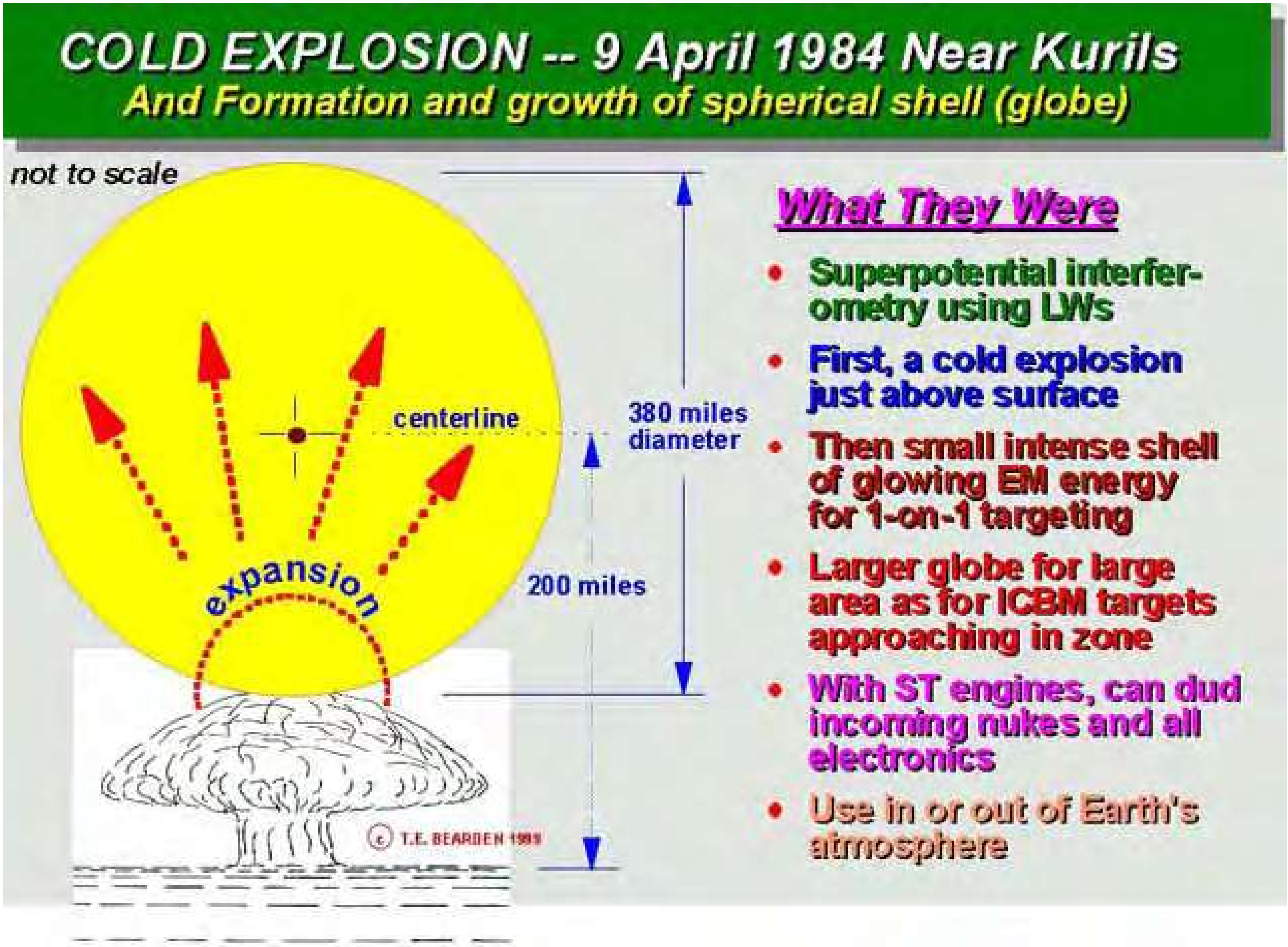

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
expansion
380 miles diameter
200 miles
© T.E. BEARDEN 1999

What They Were

- Superpotential interferometry using LWs
- First, a cold explosion just above surface
- Then small intense shell of glowing EM energy for 1-on-1 targeting
- Larger globe for large area as for ICBM targets approaching in zone
- With ST engines, can dud incoming nukes and all electronics
- Use in or out of Earth's atmosphere

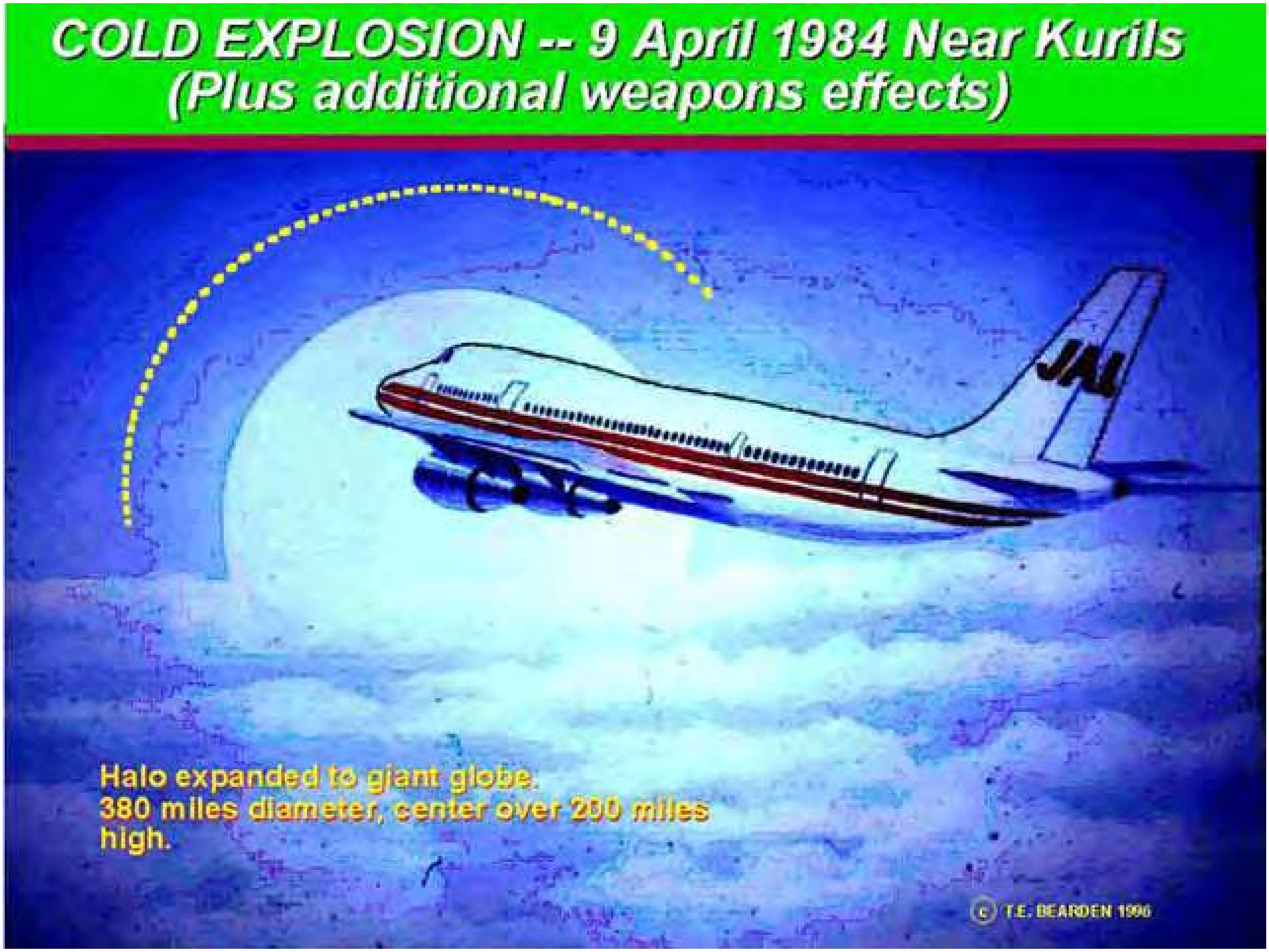

COLD EXPLOSION -- 9 April 1984 Near Kurils
(Plus additional weapons effects)

Halo expanded to giant globe.
380 miles diameter, center over 200 miles high.

© T.E. BEARDEN 1996

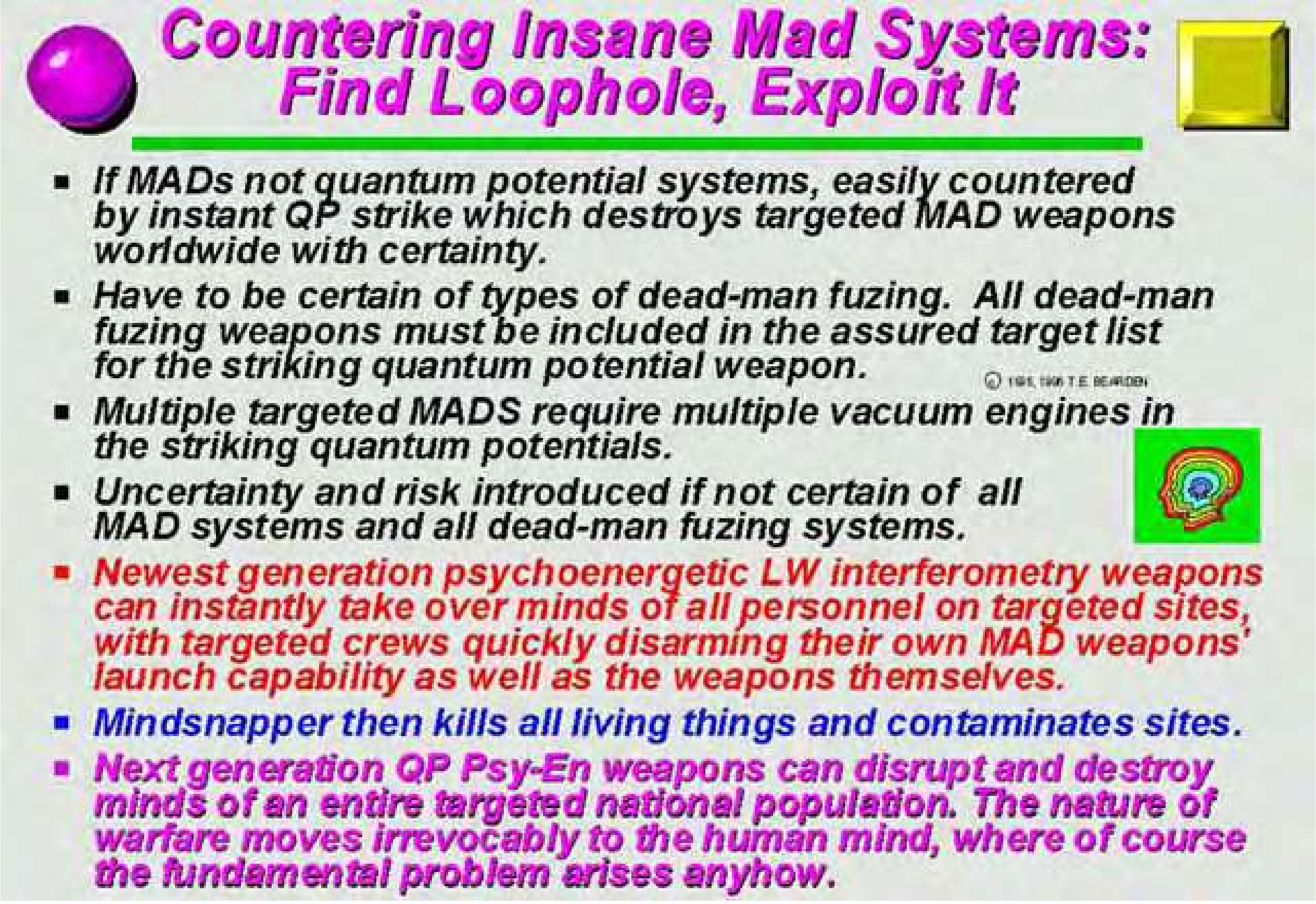

Countering Insane Mad Systems: Find Loophole, Exploit It

- *If MADs not quantum potential systems, easily countered by instant QP strike which destroys targeted MAD weapons worldwide with certainty.*
- *Have to be certain of types of dead-man fuzing. All dead-man fuzing weapons must be included in the assured target list for the striking quantum potential weapon.*

© 1991, 1996 T.E. BEARDEN

- *Multiple targeted MADS require multiple vacuum engines in the striking quantum potentials.*
- *Uncertainty and risk introduced if not certain of all MAD systems and all dead-man fuzing systems.*
- *Newest generation psychoenergetic LW interferometry weapons can instantly take over minds of all personnel on targeted sites, with targeted crews quickly disarming their own MAD weapons' launch capability as well as the weapons themselves.*
- *Mindsnapper then kills all living things and contaminates sites.*
- *Next generation QP Psy-En weapons can disrupt and destroy minds of an entire targeted national population. The nature of warfare moves irrevocably to the human mind, where of course the fundamental problem arises anyhow.*

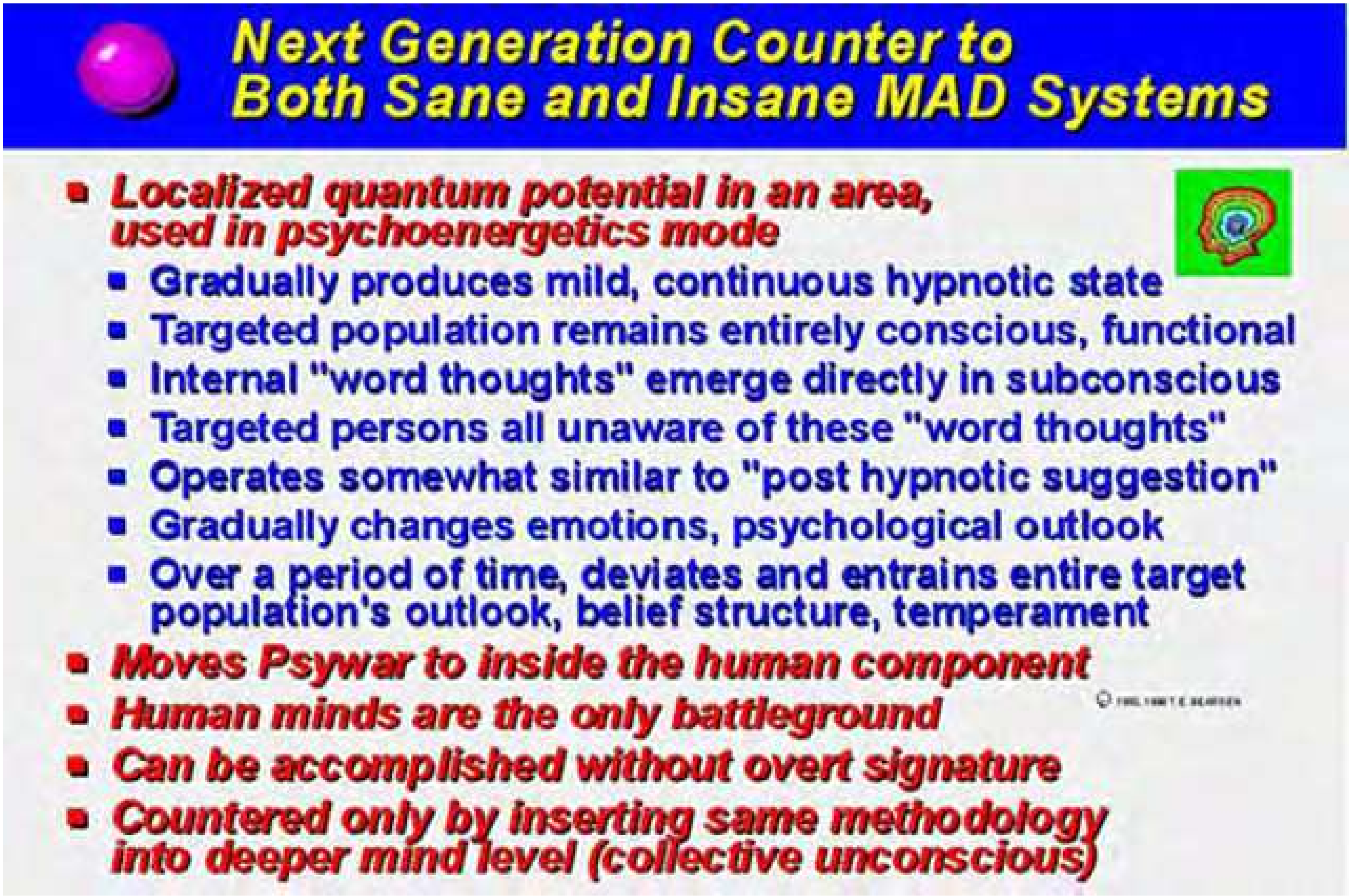

Next Generation Counter to Both Sane and Insane MAD Systems

- ***Localized quantum potential in an area, used in psychoenergetics mode***
  - **Gradually produces mild, continuous hypnotic state**
  - **Targeted population remains entirely conscious, functional**
  - **Internal "word thoughts" emerge directly in subconscious**
  - **Targeted persons all unaware of these "word thoughts"**
  - **Operates somewhat similar to "post hypnotic suggestion"**
  - **Gradually changes emotions, psychological outlook**
  - **Over a period of time, deviates and entrains entire target population's outlook, belief structure, temperament**
- ***Moves Psywar to inside the human component***
- ***Human minds are the only battleground***
- ***Can be accomplished without overt signature***
- ***Countered only by inserting same methodology into deeper mind level (collective unconscious)***

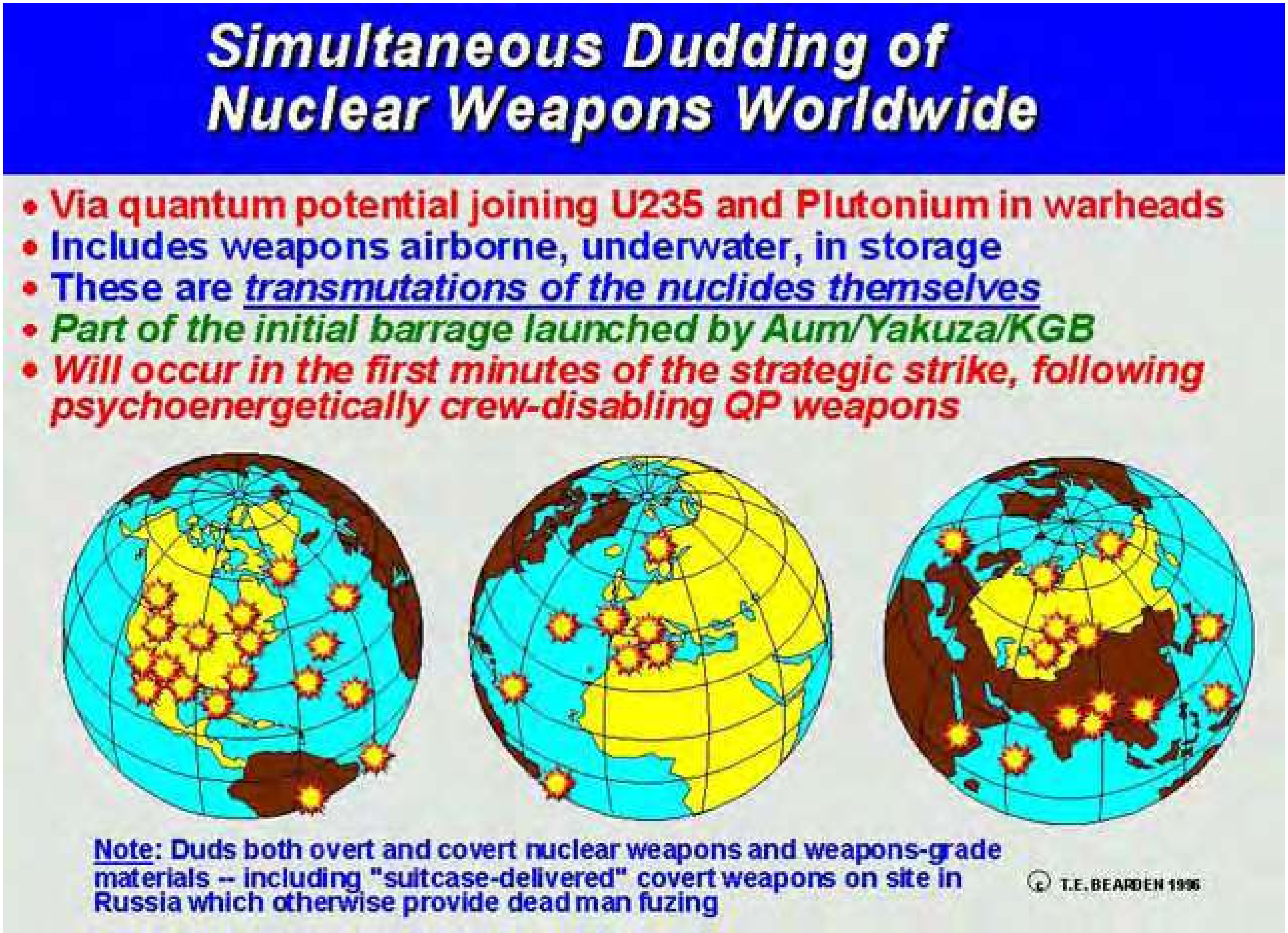

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium in warheads
- Includes weapons airborne, underwater, in storage
- These are *transmutations of the nuclides themselves*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike, following psychoenergetically crew-disabling QP weapons*

Note: Duds both overt and covert nuclear weapons and weapons-grade materials -- including "suitcase-delivered" covert weapons on site in Russia which otherwise provide dead man fuzing

© T.E. BEARDEN 1996

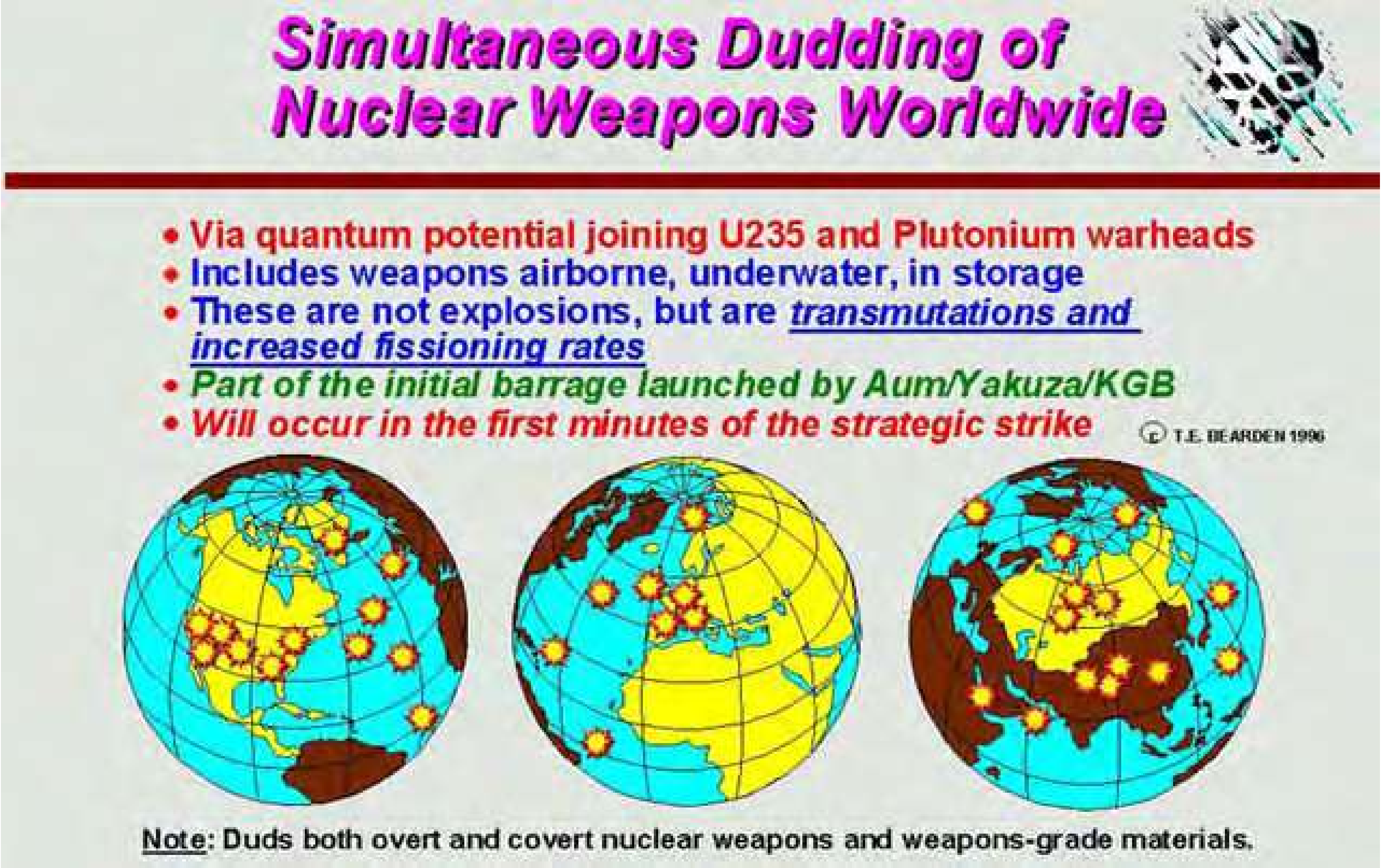

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium warheads
- Includes weapons airborne, underwater, in storage
- These are not explosions, but are *transmutations and increased fissioning rates*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike*

© T.E. BEARDEN 1996

Note: Duds both overt and covert nuclear weapons and weapons-grade materials.

The Tom Bearden Website

Help support the research

Operation Divine Wind II

Slide 1

Slide 2

Slide 3

Slide 4

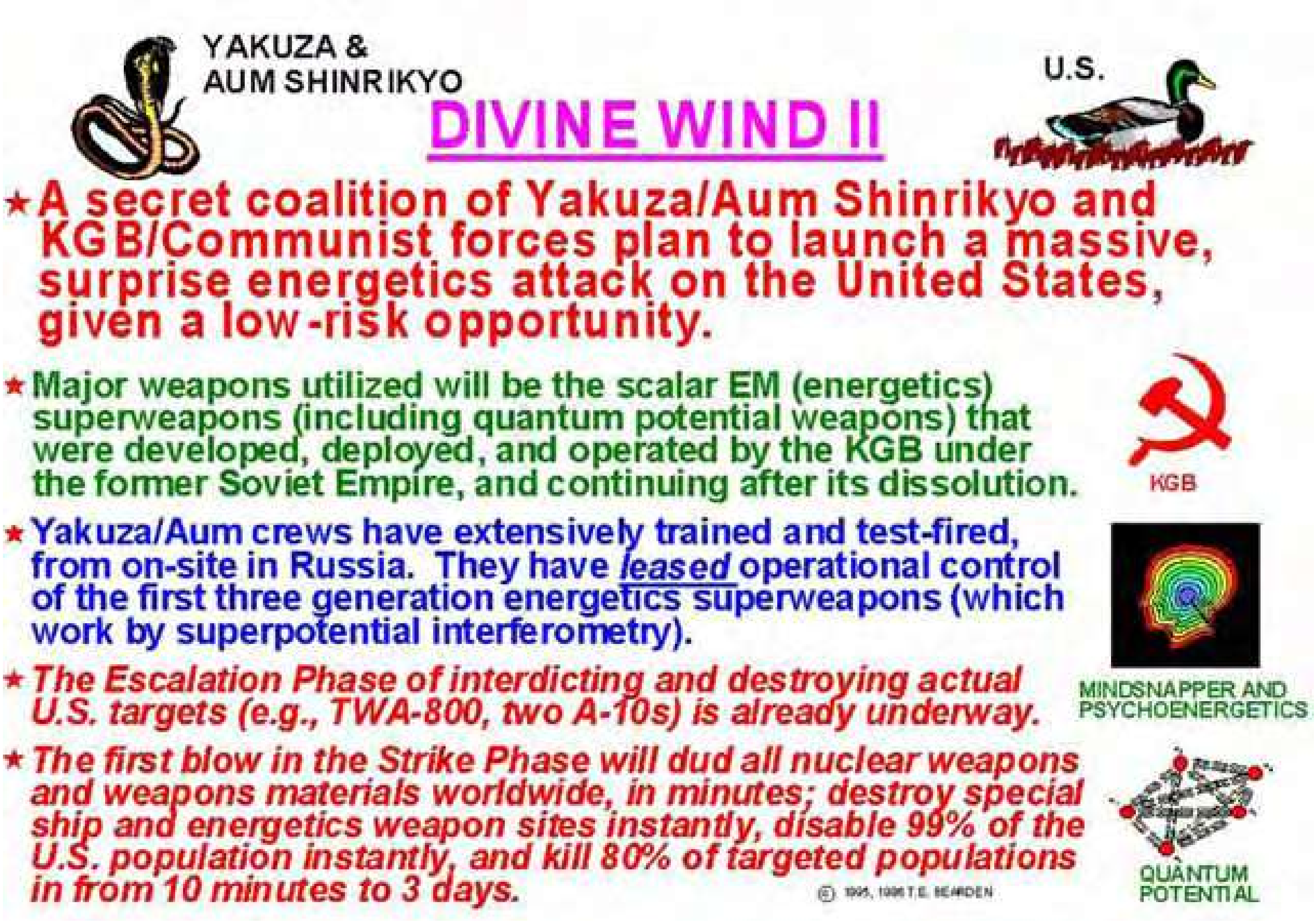
YAKUZA &
AUM SHINRIKYO
U.S.
KGB
MINDSNAPPER AND
PSYCHOENERGETICS
QUANTUM
POTENTIAL

DIVINE WIND II

- ★ **A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plan to launch a massive, surprise energetics attack on the United States, given a low-risk opportunity.**
- ★ **Major weapons utilized will be the scalar EM (energetics) superweapons (including quantum potential weapons) that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.**
- ★ **Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the first three generation energetics superweapons (which work by superpotential interferometry).**
- ★ ***The Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800, two A-10s) is already underway.***
- ★ ***The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, disable 99% of the U.S. population instantly, and kill 80% of targeted populations in from 10 minutes to 3 days.***

© 1995, 1996 T.E. BEARDEN

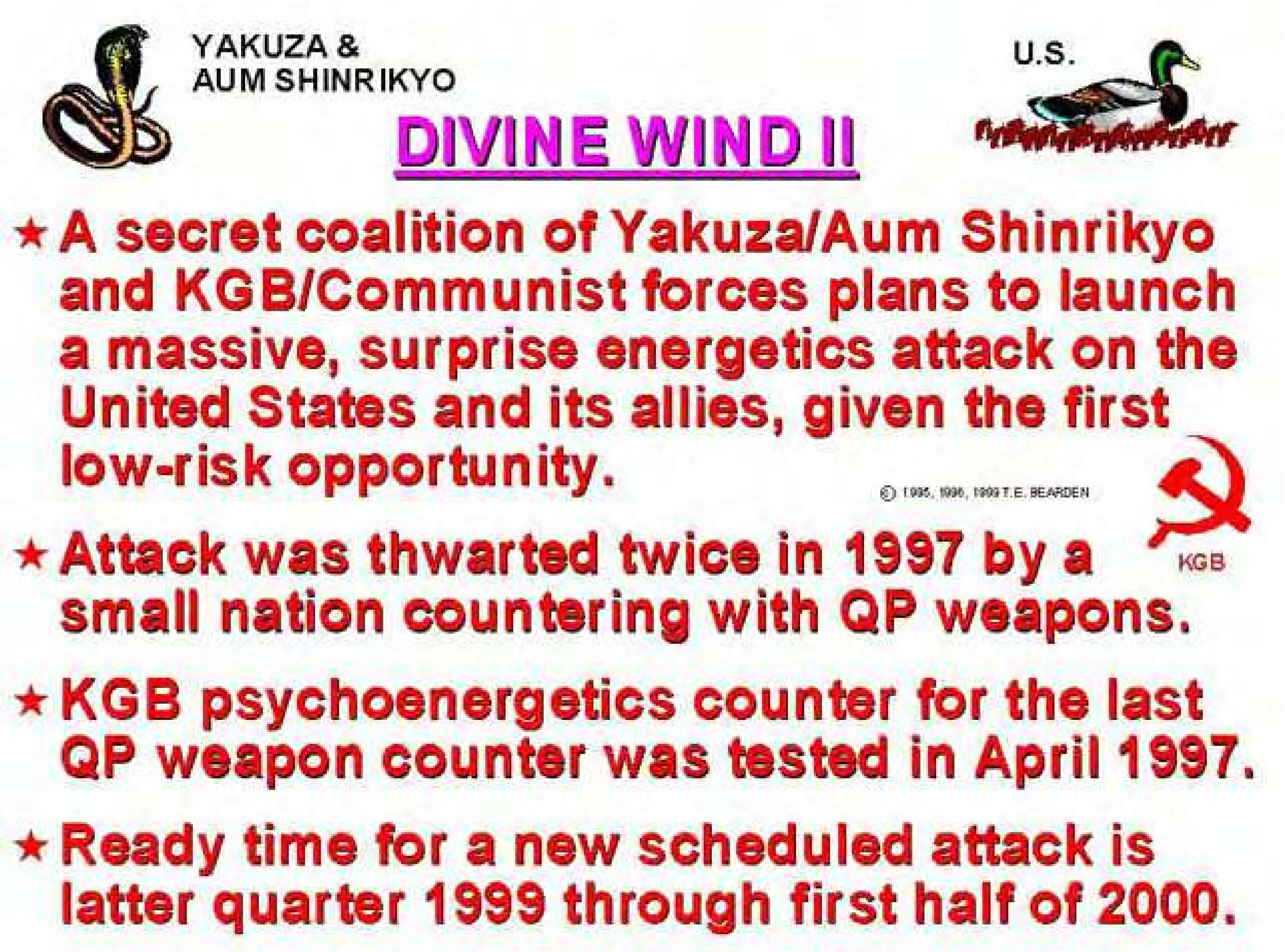

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

- ★ A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plans to launch a massive, surprise energetics attack on the United States and its allies, given the first low-risk opportunity.

© 1995, 1996, 1999 T.E. BEARDEN

KGB

- ★ Attack was thwarted twice in 1997 by a small nation countering with QP weapons.
- ★ KGB psychoenergetics counter for the last QP weapon counter was tested in April 1997.
- ★ Ready time for a new scheduled attack is latter quarter 1999 through first half of 2000.

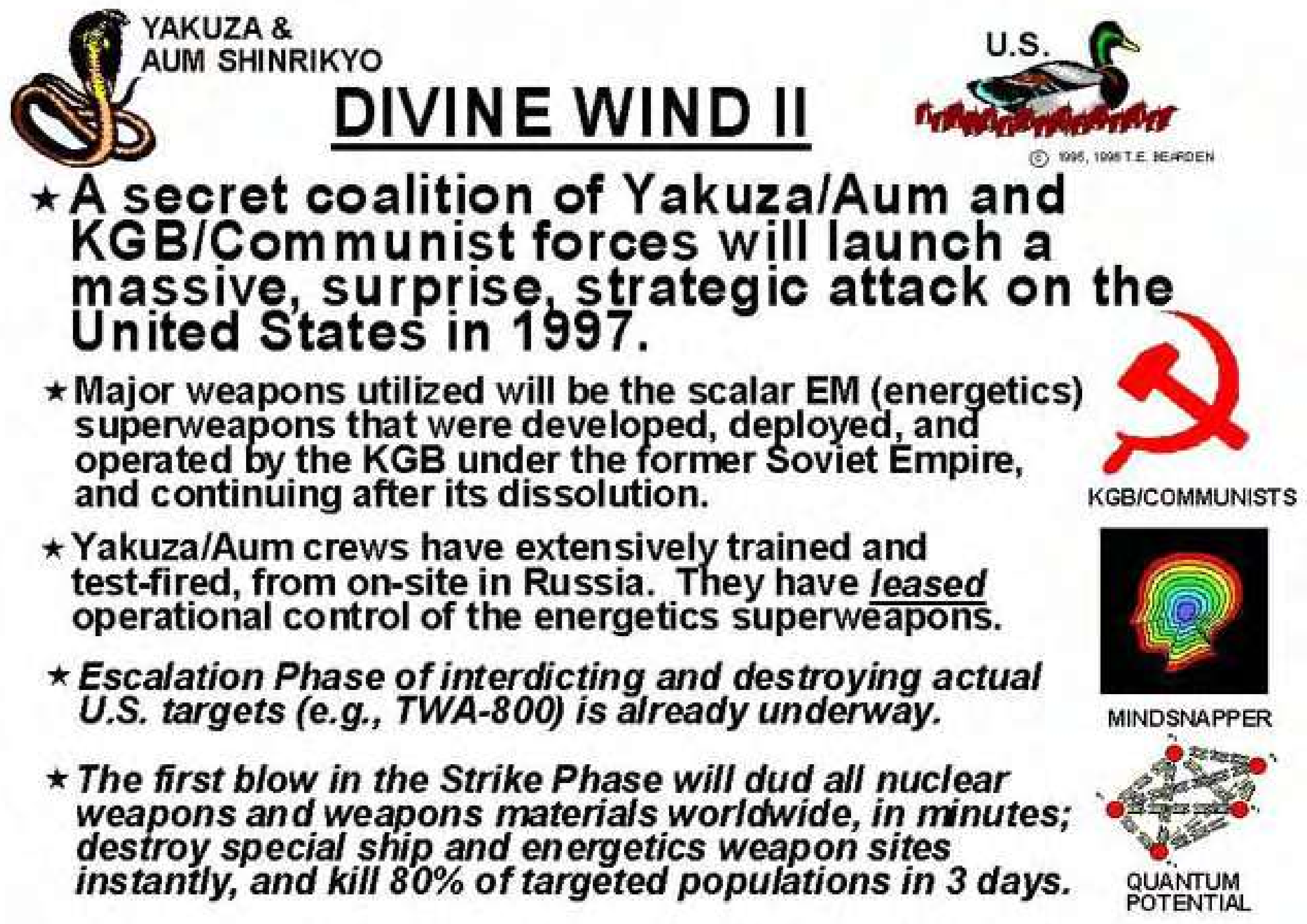

YAKUZA &
AUM SHINRIKYO

DIVINE WIND II

U.S.

© 1995, 1996 T.E. BEARDEN

★ **A secret coalition of Yakuza/Aum and KGB/Communist forces will launch a massive, surprise, strategic attack on the United States in 1997.**

- ★ Major weapons utilized will be the scalar EM (energetics) superweapons that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.
- ★ Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the energetics superweapons.
- ★ *Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800) is already underway.*
- ★ *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, and kill 80% of targeted populations in 3 days.*

KGB/COMMUNISTS

MINDSNAPPER

QUANTUM
POTENTIAL

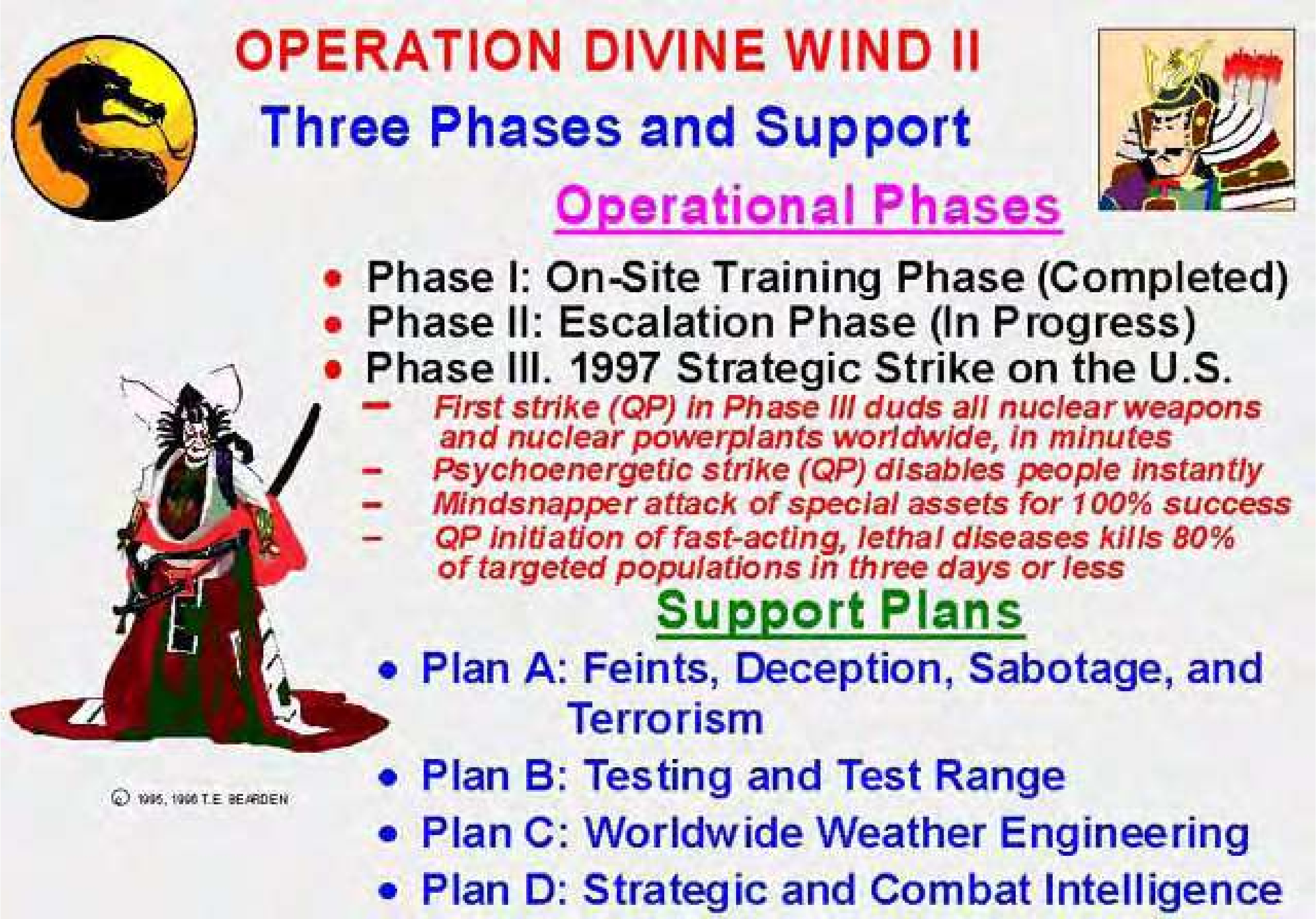

OPERATION DIVINE WIND II

Three Phases and Support

Operational Phases

- Phase I: On-Site Training Phase (Completed)
- Phase II: Escalation Phase (In Progress)
- Phase III. 1997 Strategic Strike on the U.S.
  - *First strike (QP) in Phase III duds all nuclear weapons and nuclear powerplants worldwide, in minutes*
  - *Psychoenergetic strike (QP) disables people instantly*
  - *Mindsnapper attack of special assets for 100% success*
  - *QP initiation of fast-acting, lethal diseases kills 80% of targeted populations in three days or less*

Support Plans

- Plan A: Feints, Deception, Sabotage, and Terrorism
- Plan B: Testing and Test Range
- Plan C: Worldwide Weather Engineering
- Plan D: Strategic and Combat Intelligence

© 1995, 1996 T.E. BEARDEN

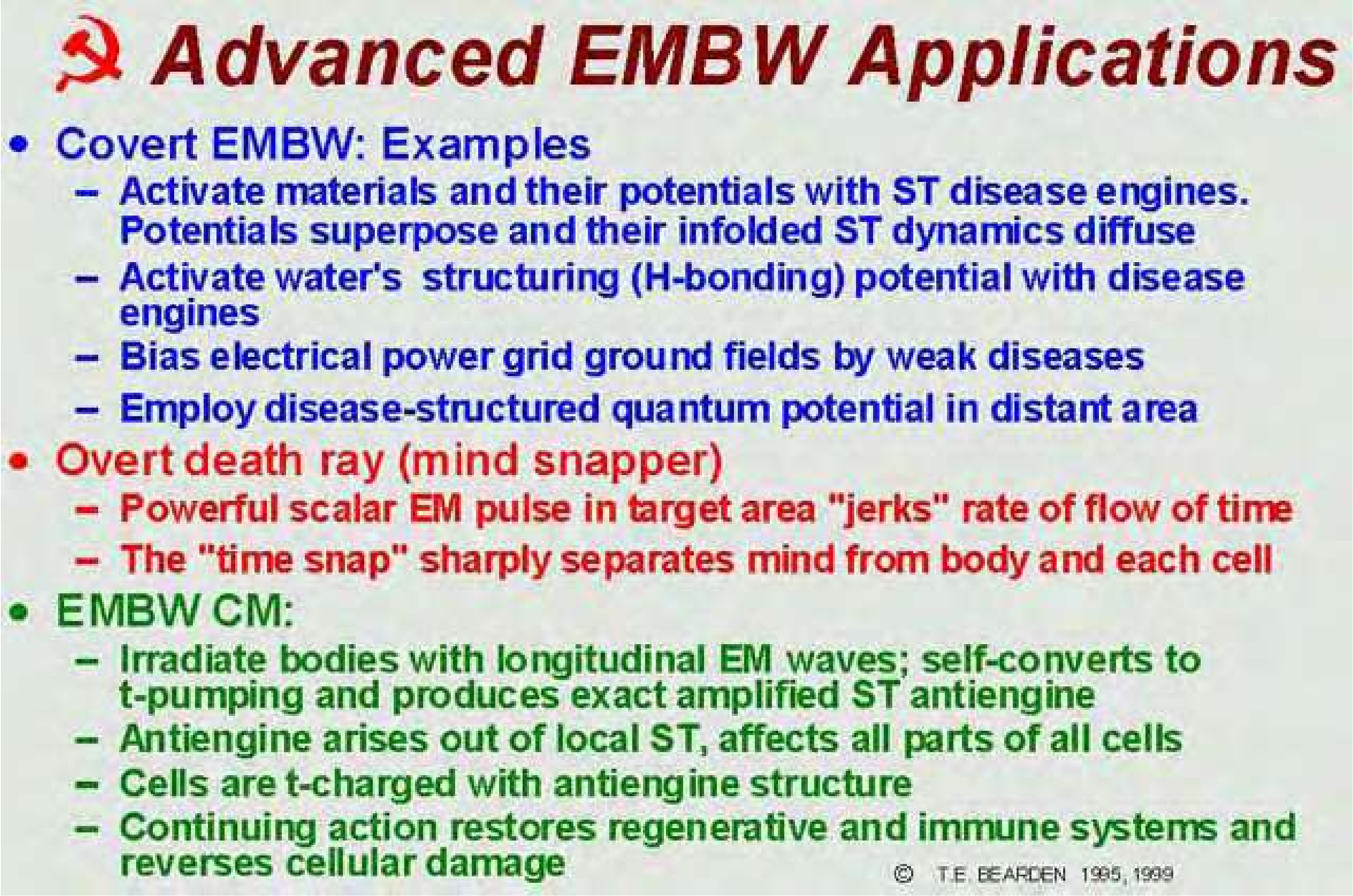

Advanced EMBW Applications

- Covert EMBW: Examples
  - Activate materials and their potentials with ST disease engines. Potentials superpose and their infolded ST dynamics diffuse
  - Activate water's structuring (H-bonding) potential with disease engines
  - Bias electrical power grid ground fields by weak diseases
  - Employ disease-structured quantum potential in distant area
- Overt death ray (mind snapper)
  - Powerful scalar EM pulse in target area "jerks" rate of flow of time
  - The "time snap" sharply separates mind from body and each cell
- EMBW CM:
  - Irradiate bodies with longitudinal EM waves; self-converts to t-pumping and produces exact amplified ST antiengine
  - Antiengine arises out of local ST, affects all parts of all cells
  - Cells are t-charged with antiengine structure
  - Continuing action restores regenerative and immune systems and reverses cellular damage

© T.E. BEARDEN 1995, 1999

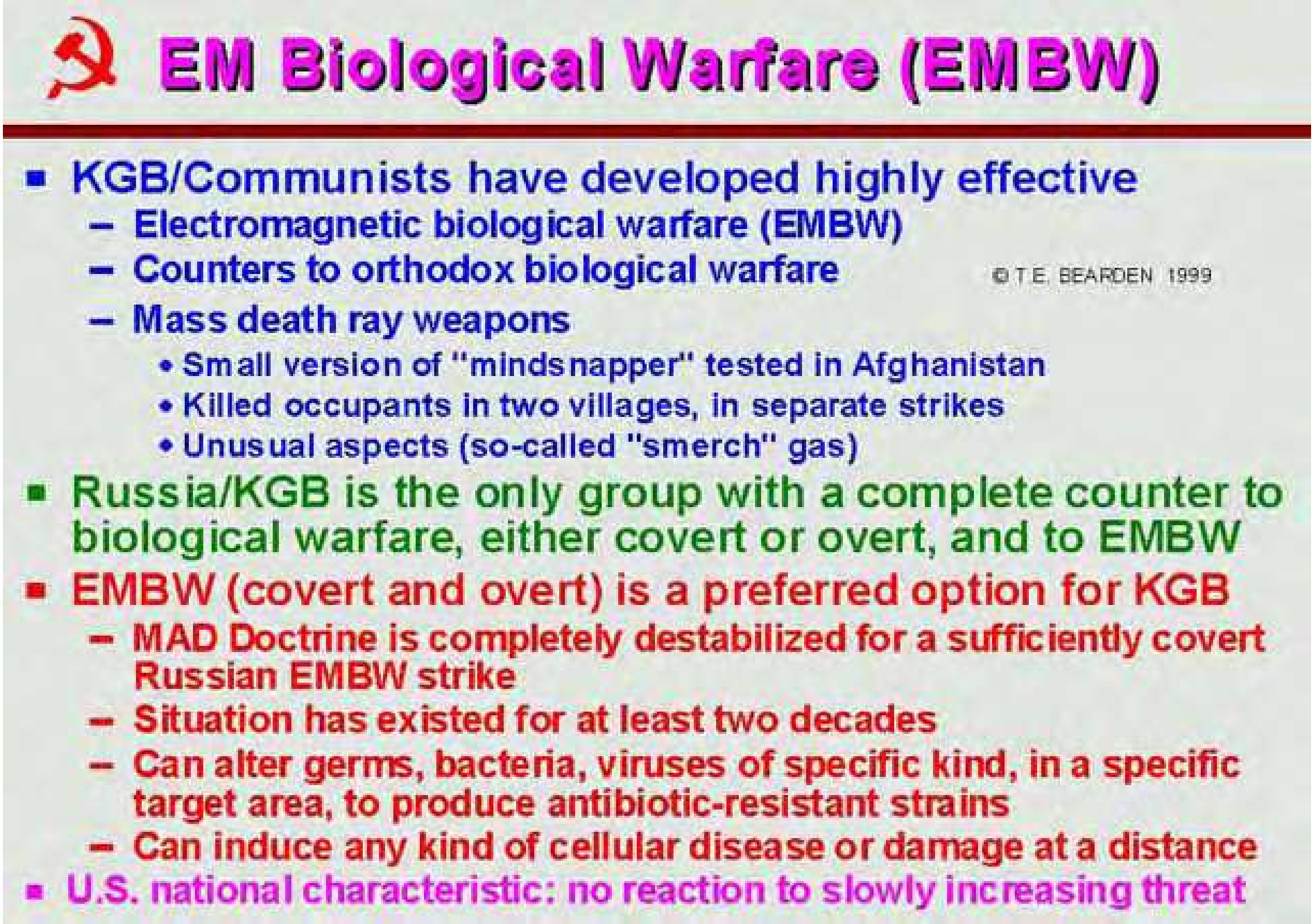

EM Biological Warfare (EMBW)

- KGB/Communists have developed highly effective
  - Electromagnetic biological warfare (EMBW)
  - Counters to orthodox biological warfare
  - Mass death ray weapons
    - Small version of "mindsnapper" tested in Afghanistan
    - Killed occupants in two villages, in separate strikes
    - Unusual aspects (so-called "smerch" gas)
- Russia/KGB is the only group with a complete counter to biological warfare, either covert or overt, and to EMBW
- EMBW (covert and overt) is a preferred option for KGB
  - MAD Doctrine is completely destabilized for a sufficiently covert Russian EMBW strike
  - Situation has existed for at least two decades
  - Can alter germs, bacteria, viruses of specific kind, in a specific target area, to produce antibiotic-resistant strains
  - Can induce any kind of cellular disease or damage at a distance
- U.S. national characteristic: no reaction to slowly increasing threat

© T.E. BEARDEN 1999

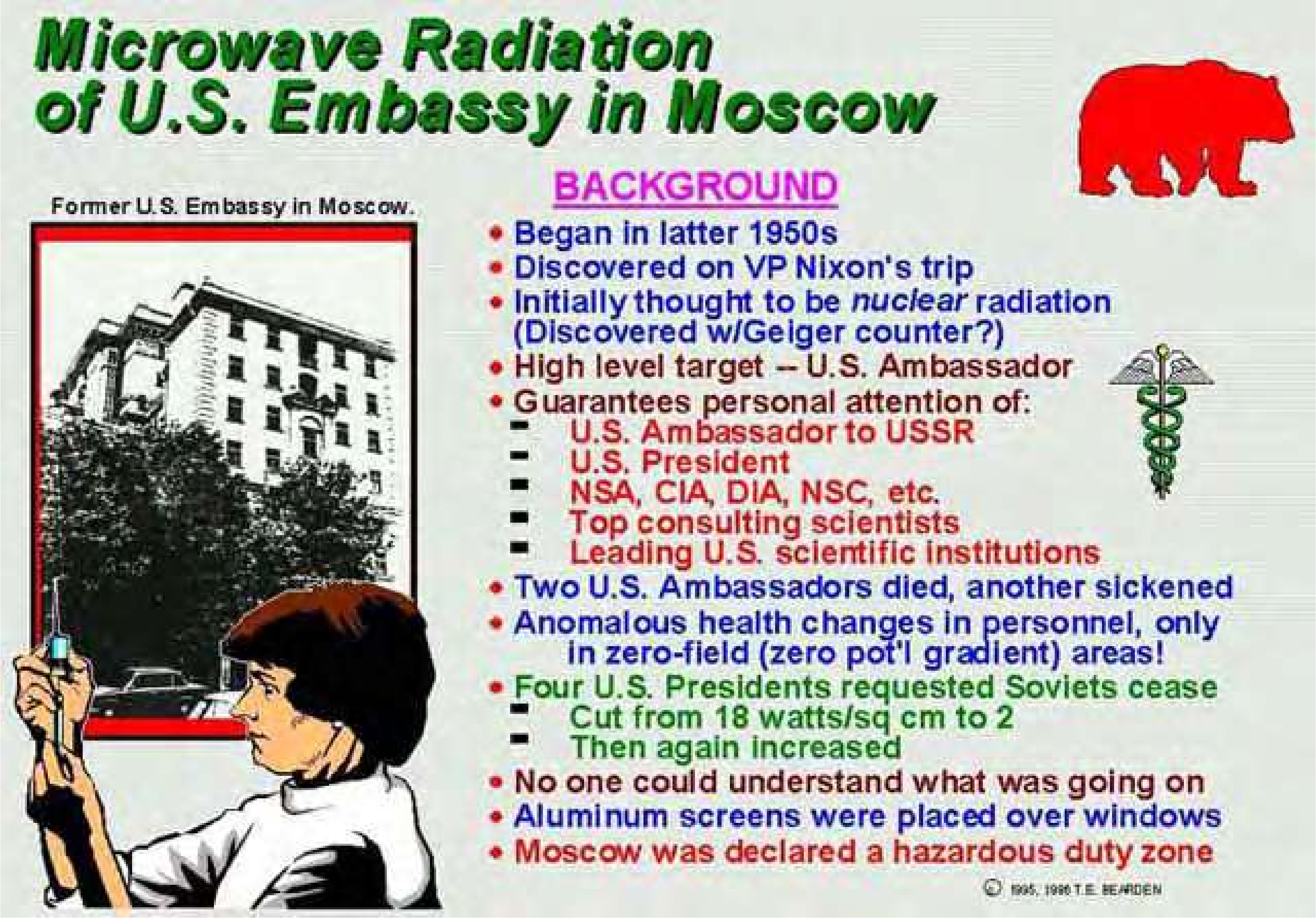

Microwave Radiation of U.S. Embassy in Moscow

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

© 1995, 1996 T.E. BEARDEN

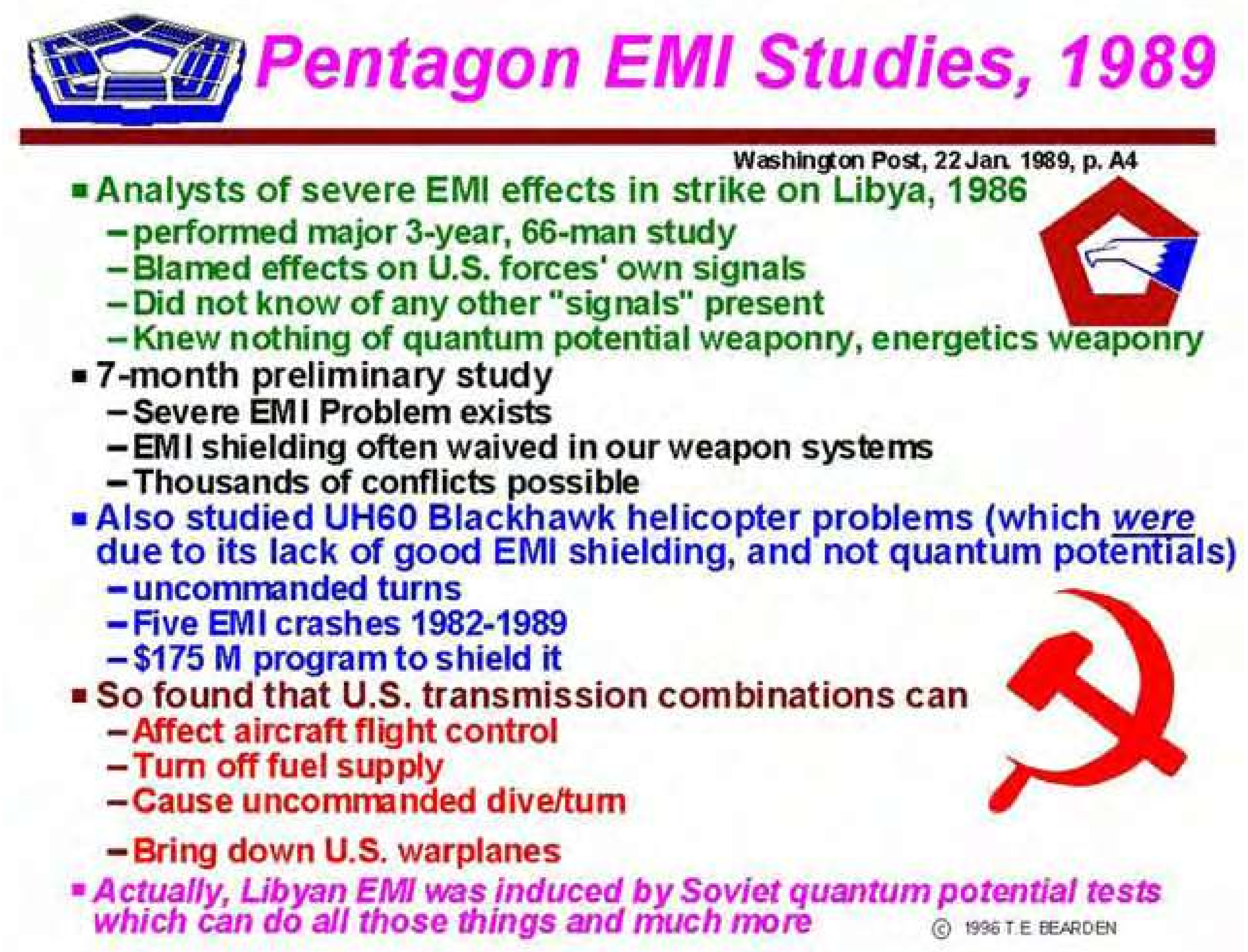

Pentagon EMI Studies, 1989

Washington Post, 22 Jan. 1989, p. A4

- Analysts of severe EMI effects in strike on Libya, 1986
  - performed major 3-year, 66-man study
  - Blamed effects on U.S. forces' own signals
  - Did not know of any other "signals" present
  - Knew nothing of quantum potential weaponry, energetics weaponry
- 7-month preliminary study
  - Severe EMI Problem exists
  - EMI shielding often waived in our weapon systems
  - Thousands of conflicts possible
- Also studied UH60 Blackhawk helicopter problems (which *were* due to its lack of good EMI shielding, and not quantum potentials)
  - uncommanded turns
  - Five EMI crashes 1982-1989
  - $175 M program to shield it
- So found that U.S. transmission combinations can
  - Affect aircraft flight control
  - Turn off fuel supply
  - Cause uncommanded dive/turn
  - Bring down U.S. warplanes
- *Actually, Libyan EMI was induced by Soviet quantum potential tests which can do all those things and much more*

© 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

EM Missile Practice

- EM Missile Practice
  Latter Sept. - Early Nov. 1996
  Melbourne, Australia

- Deception: Using EM Missile to Reinforce Notion of SAM Attack
  Saturday evening, Nov. 16, 1996

- Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry
  Thursday, December 12th, 1996

- EM Missile Strike, Offset from Night Shuttle Launch, Cape Canaveral, Florida
  The Smoking Gun in November 1985

- Distant Operator Onsite in Russia Slew Away Registration Point
  Marker Beacon over Night Shuttle Launch, 26 November, 1985

- EM Missile Practice and Deception
  November 17, 1996

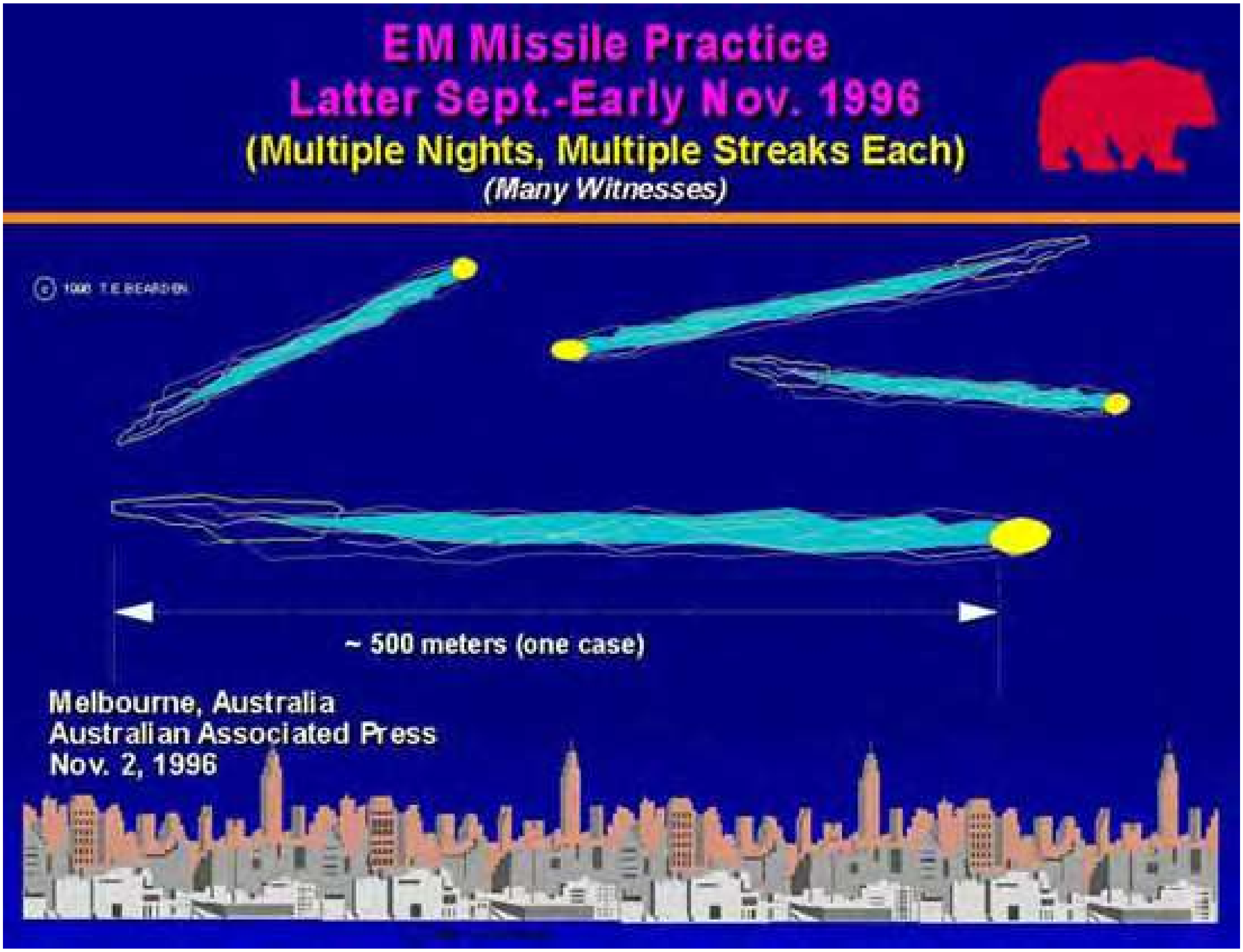

EM Missile Practice
Latter Sept.-Early Nov. 1996

(Multiple Nights, Multiple Streaks Each)

(Many Witnesses)

© 1996 T.E.BEARDEN
~ 500 meters (one case)

Melbourne, Australia
Australian Associated Press
Nov. 2, 1996

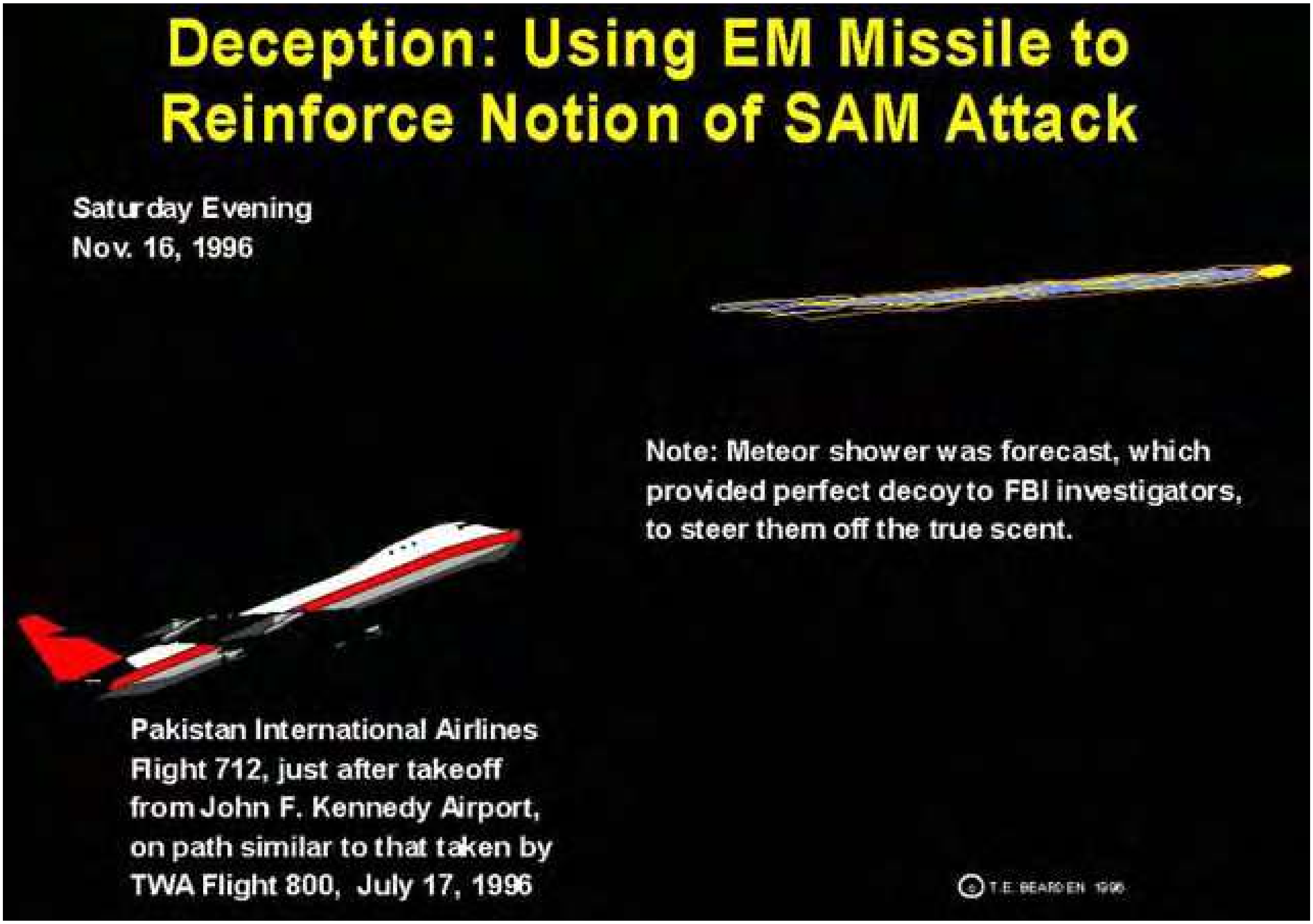

Deception: Using EM Missile to Reinforce Notion of SAM Attack

Saturday Evening
Nov. 16, 1996

Note: Meteor shower was forecast, which provided perfect decoy to FBI investigators, to steer them off the true scent.

Pakistan International Airlines Flight 712, just after takeoff from John F. Kennedy Airport, on path similar to that taken by TWA Flight 800, July 17, 1996

© T.E. BEARDEN 1996

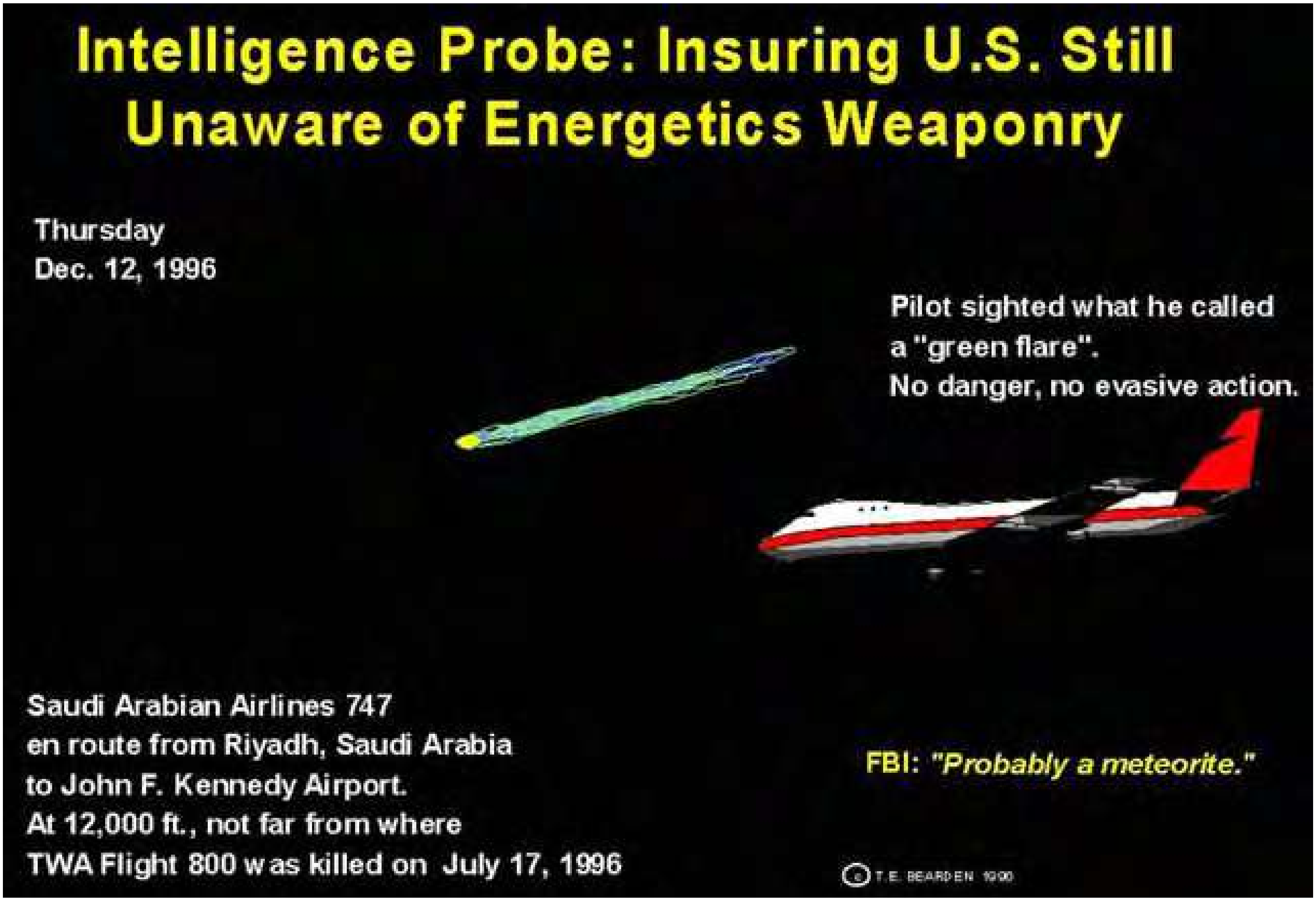

Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry

Thursday
Dec. 12, 1996

Pilot sighted what he called
a "green flare".
No danger, no evasive action.

Saudi Arabian Airlines 747
en route from Riyadh, Saudi Arabia
to John F. Kennedy Airport.
At 12,000 ft., not far from where
TWA Flight 800 was killed on July 17, 1996

FBI: *"Probably a meteorite."*

© T.E. BEARDEN 1990

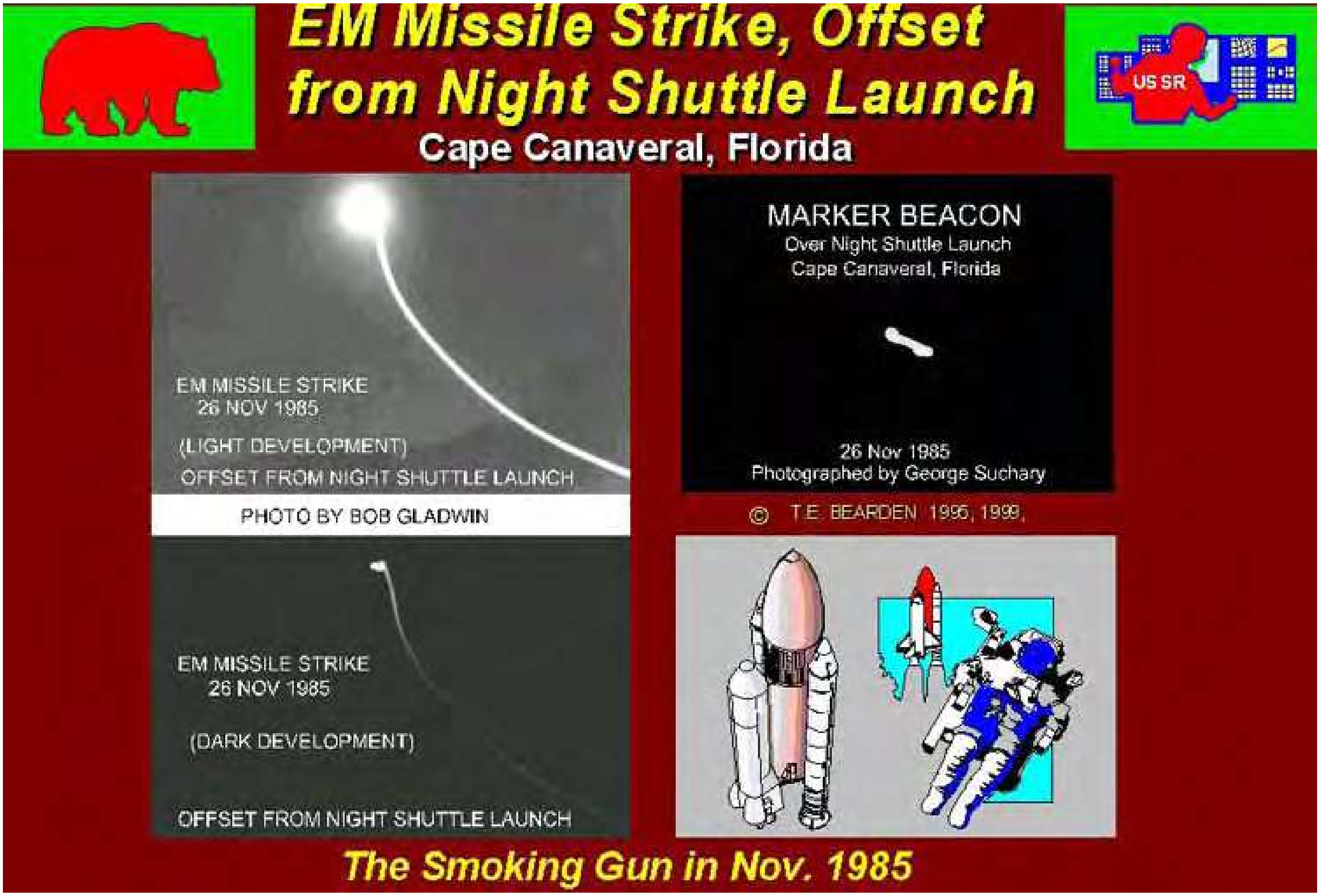

EM Missile Strike, Offset from Night Shuttle Launch

Cape Canaveral, Florida

EM MISSILE STRIKE
26 NOV 1985

(LIGHT DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

PHOTO BY BOB GLADWIN

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

© T.E. BEARDEN 1995, 1999,

EM MISSILE STRIKE
26 NOV 1985

(DARK DEVELOPMENT)

OFFSET FROM NIGHT SHUTTLE LAUNCH

The Smoking Gun in Nov. 1985

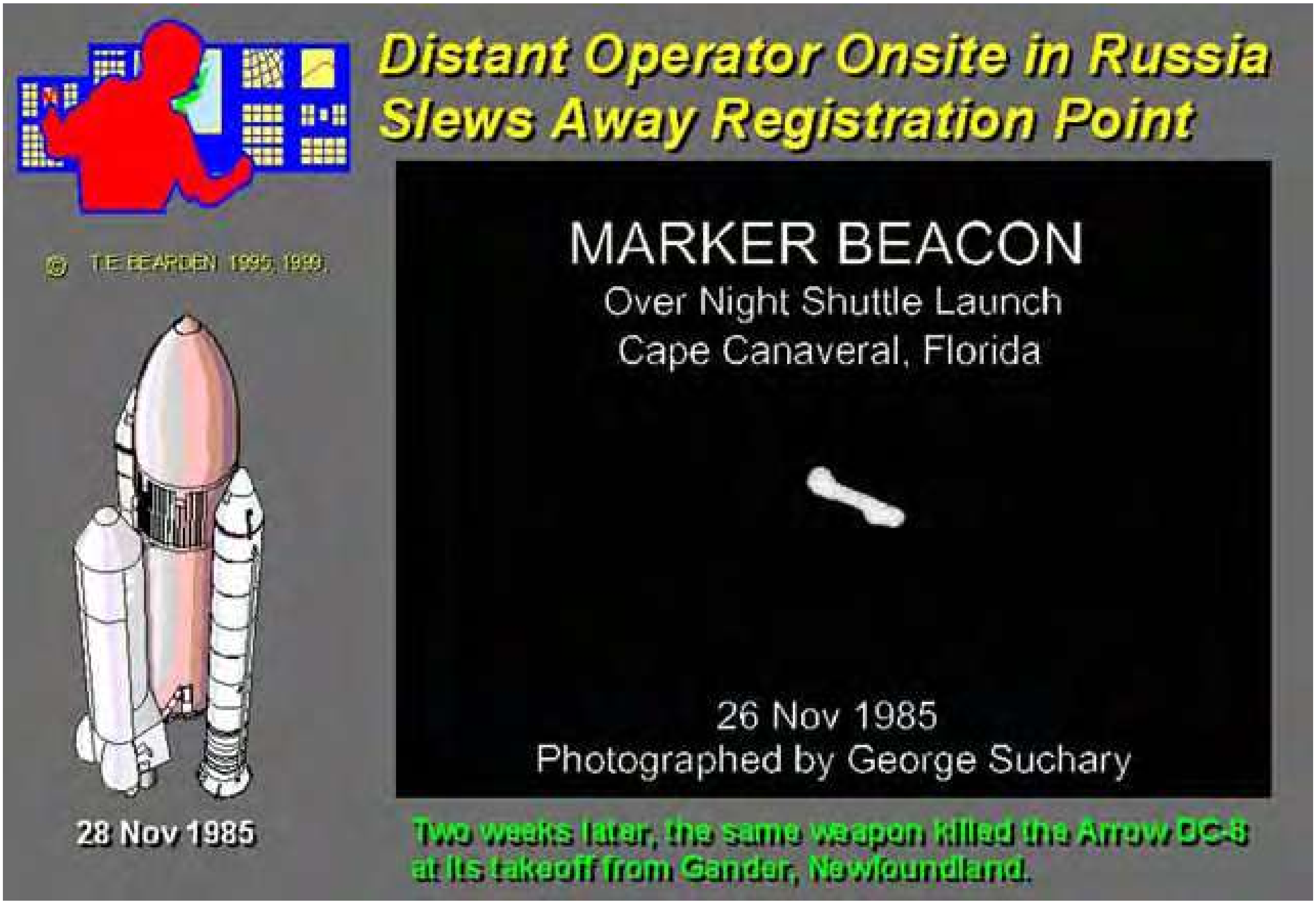

Distant Operator Onsite in Russia Slews Away Registration Point

© T.E. BEARDEN 1995, 1999.

MARKER BEACON

Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

28 Nov 1985

Two weeks later, the same weapon killed the Arrow DC-8 at its takeoff from Gander, Newfoundland.

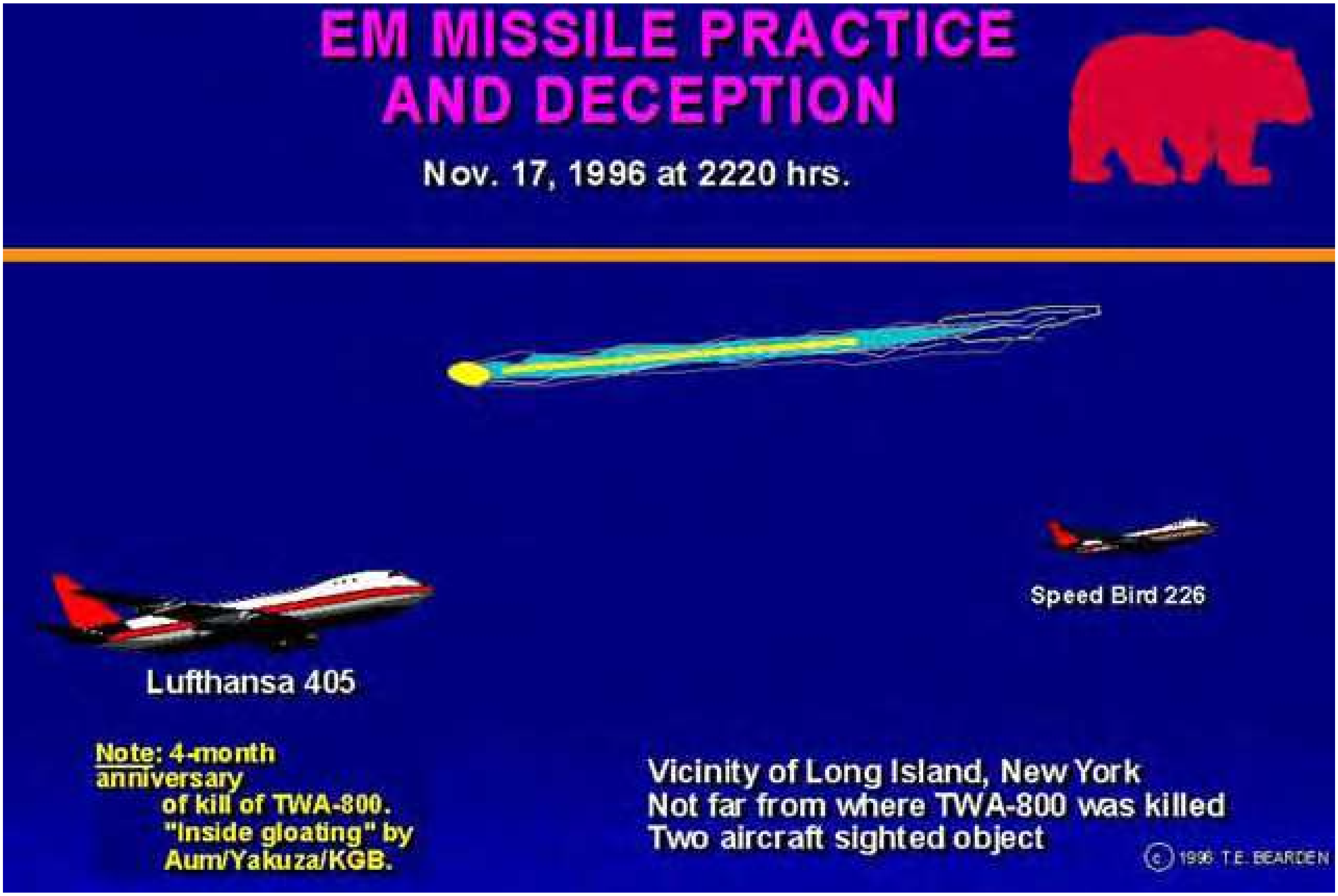

EM MISSILE PRACTICE AND DECEPTION

Nov. 17, 1996 at 2220 hrs.

Speed Bird 226

Lufthansa 405

Note: 4-month anniversary
of kill of TWA-800.
"Inside gloating" by
Aum/Yakuza/KGB.

Vicinity of Long Island, New York
Not far from where TWA-800 was killed
Two aircraft sighted object

© 1996 T.E. BEARDEN

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

LIVING SYSTEMS

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

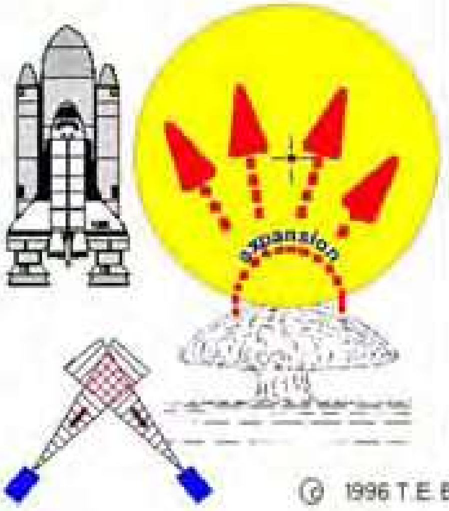

© 1996 T.E. BEARDEN

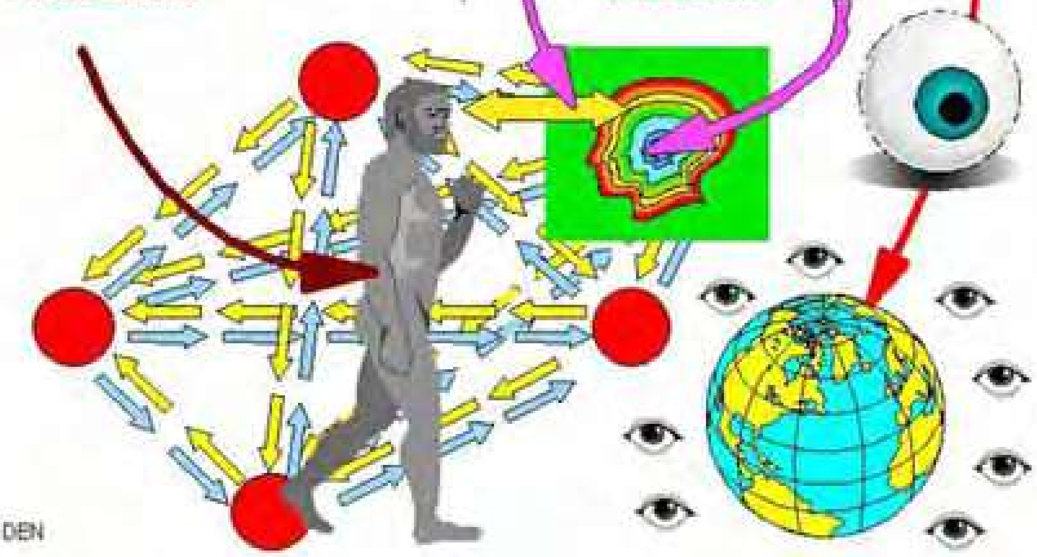

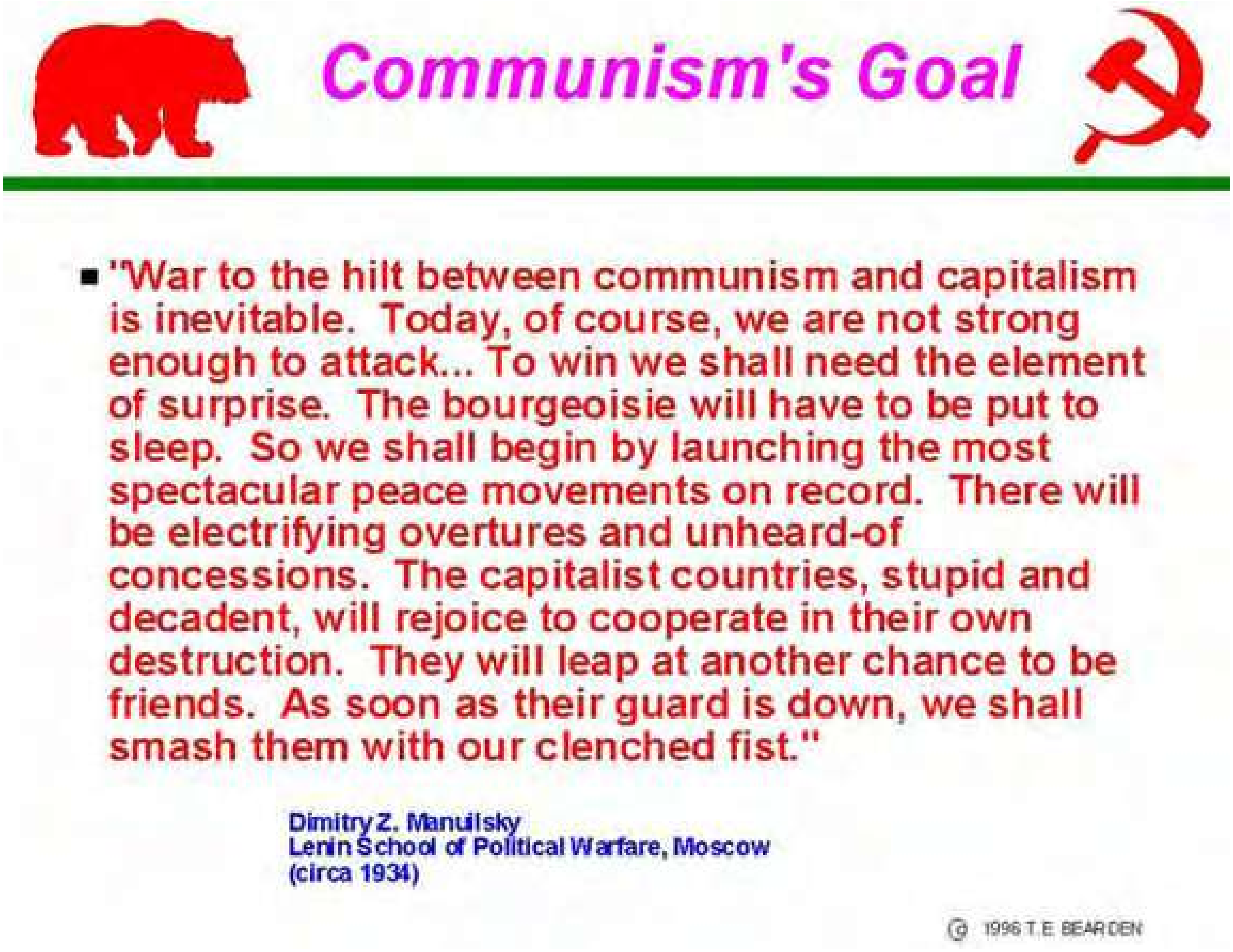

Communism's Goal

- "War to the hilt between communism and capitalism is inevitable. Today, of course, we are not strong enough to attack... To win we shall need the element of surprise. The bourgeoisie will have to be put to sleep. So we shall begin by launching the most spectacular peace movements on record. There will be electrifying overtures and unheard-of concessions. The capitalist countries, stupid and decadent, will rejoice to cooperate in their own destruction. They will leap at another chance to be friends. As soon as their guard is down, we shall smash them with our clenched fist."

Dimitry Z. Manuilsky
Lenin School of Political Warfare, Moscow
(circa 1934)

© 1996 T.E. BEARDEN

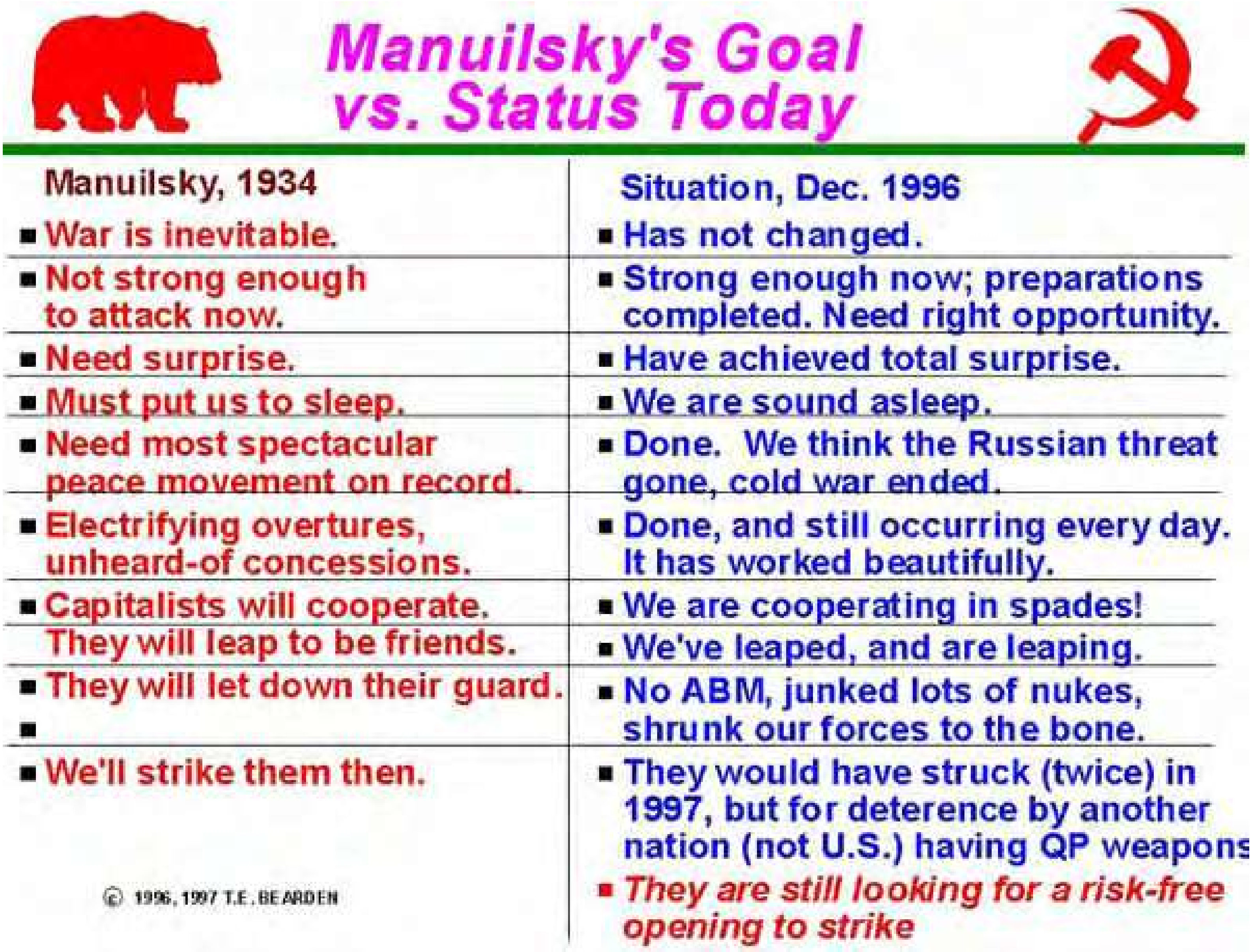

Manuilsky's Goal vs. Status Today

Manuilsky, 1934	Situation, Dec. 1996
■ War is inevitable.	■ Has not changed.
■ Not strong enough to attack now.	■ Strong enough now; preparations completed. Need right opportunity.
■ Need surprise.	■ Have achieved total surprise.
■ Must put us to sleep.	■ We are sound asleep.
■ Need most spectacular peace movement on record.	■ Done. We think the Russian threat gone, cold war ended.
■ Electrifying overtures, unheard-of concessions.	■ Done, and still occurring every day. It has worked beautifully.
■ Capitalists will cooperate.	■ We are cooperating in spades!
They will leap to be friends.	■ We've leaped, and are leaping.
■ They will let down their guard. ■	■ No ABM, junked lots of nukes, shrunk our forces to the bone.
■ We'll strike them then. © 1996, 1997 T.E. BEARDEN	■ They would have struck (twice) in 1997, but for deterence by another nation (not U.S.) having QP weapons ■ *They are still looking for a risk-free opening to strike*

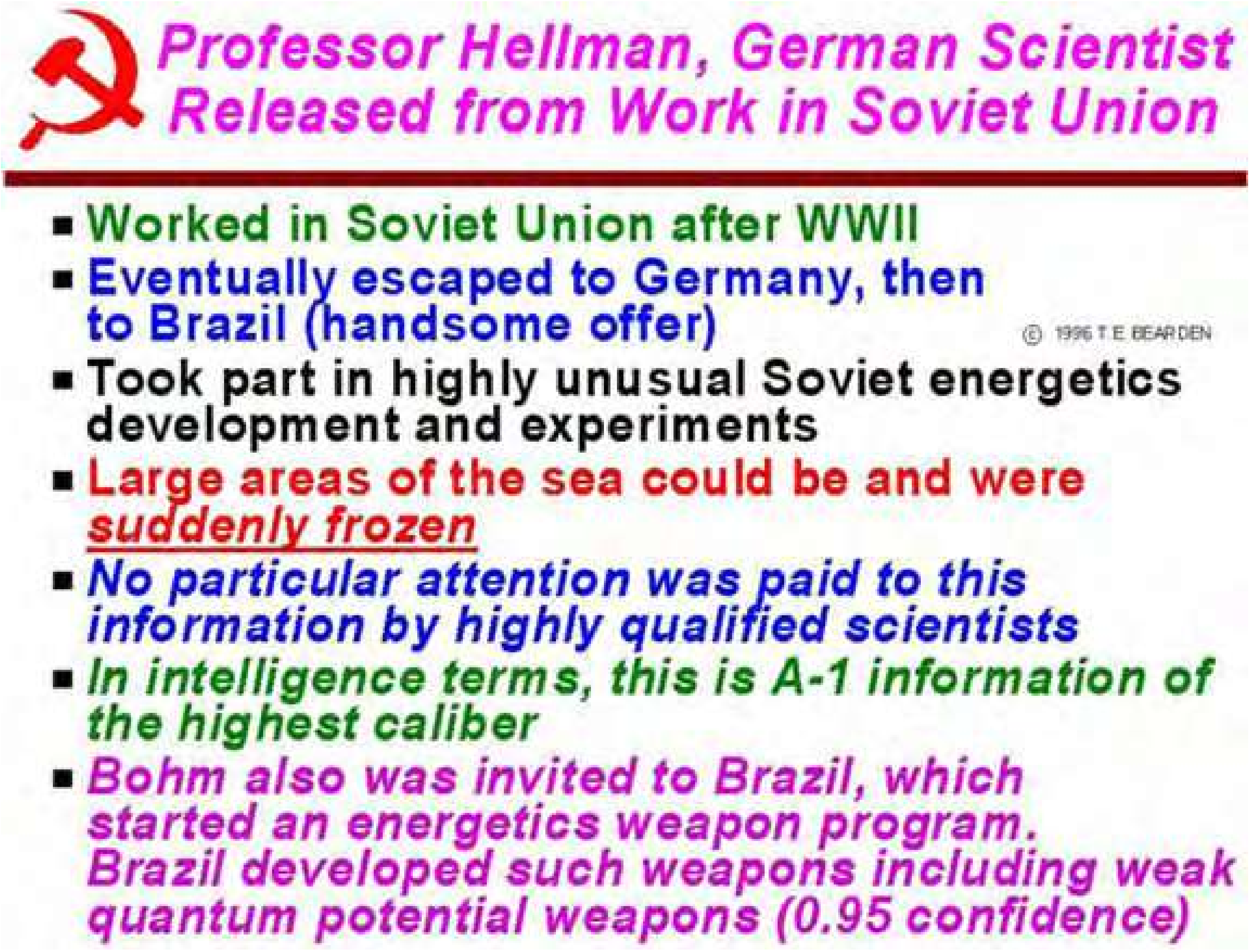

Professor Hellman, German Scientist Released from Work in Soviet Union

- Worked in Soviet Union after WWII
- Eventually escaped to Germany, then to Brazil (handsome offer)
- Took part in highly unusual Soviet energetics development and experiments
- Large areas of the sea could be and were *suddenly frozen*
- *No particular attention was paid to this information by highly qualified scientists*
- *In intelligence terms, this is A-1 information of the highest caliber*
- *Bohm also was invited to Brazil, which started an energetics weapon program. Brazil developed such weapons including weak quantum potential weapons (0.95 confidence)*

© 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

Hemispheres of Light

- March 20th, 1969
  Caribbean and Western North Atlantic
- Hemisphere and Globes, 1977
  24 March, 1977
- Virgin Islands Incident
  August 1969
- Typical Shield Test, Northern Route to Japan
  June, 1982

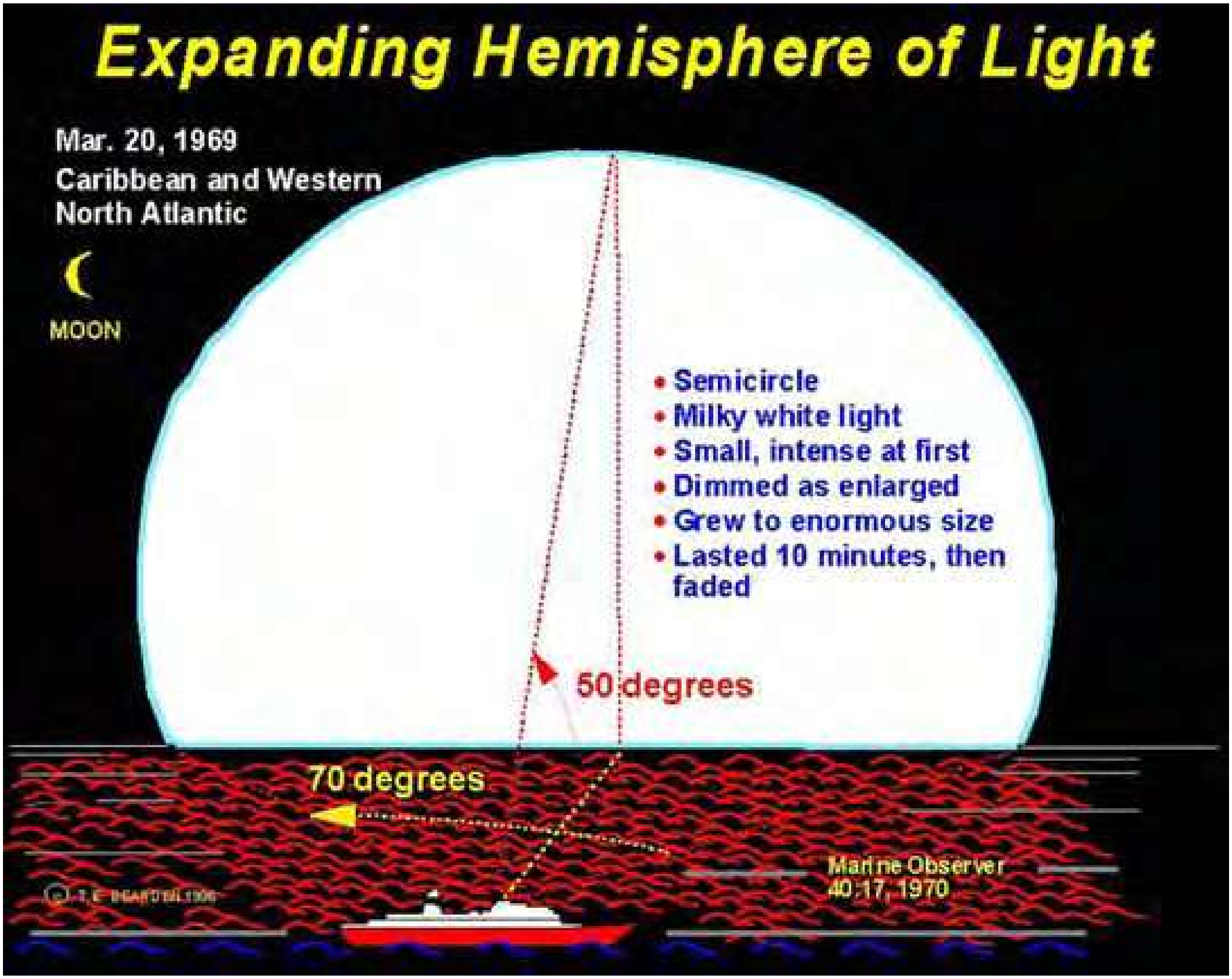

Expanding Hemisphere of Light

Mar. 20, 1969
Caribbean and Western
North Atlantic

- Semicircle
- Milky white light
- Small, intense at first
- Dimmed as enlarged
- Grew to enormous size
- Lasted 10 minutes, then faded

MOON
50 degrees
70 degrees
Marine Observer
40:17, 1970

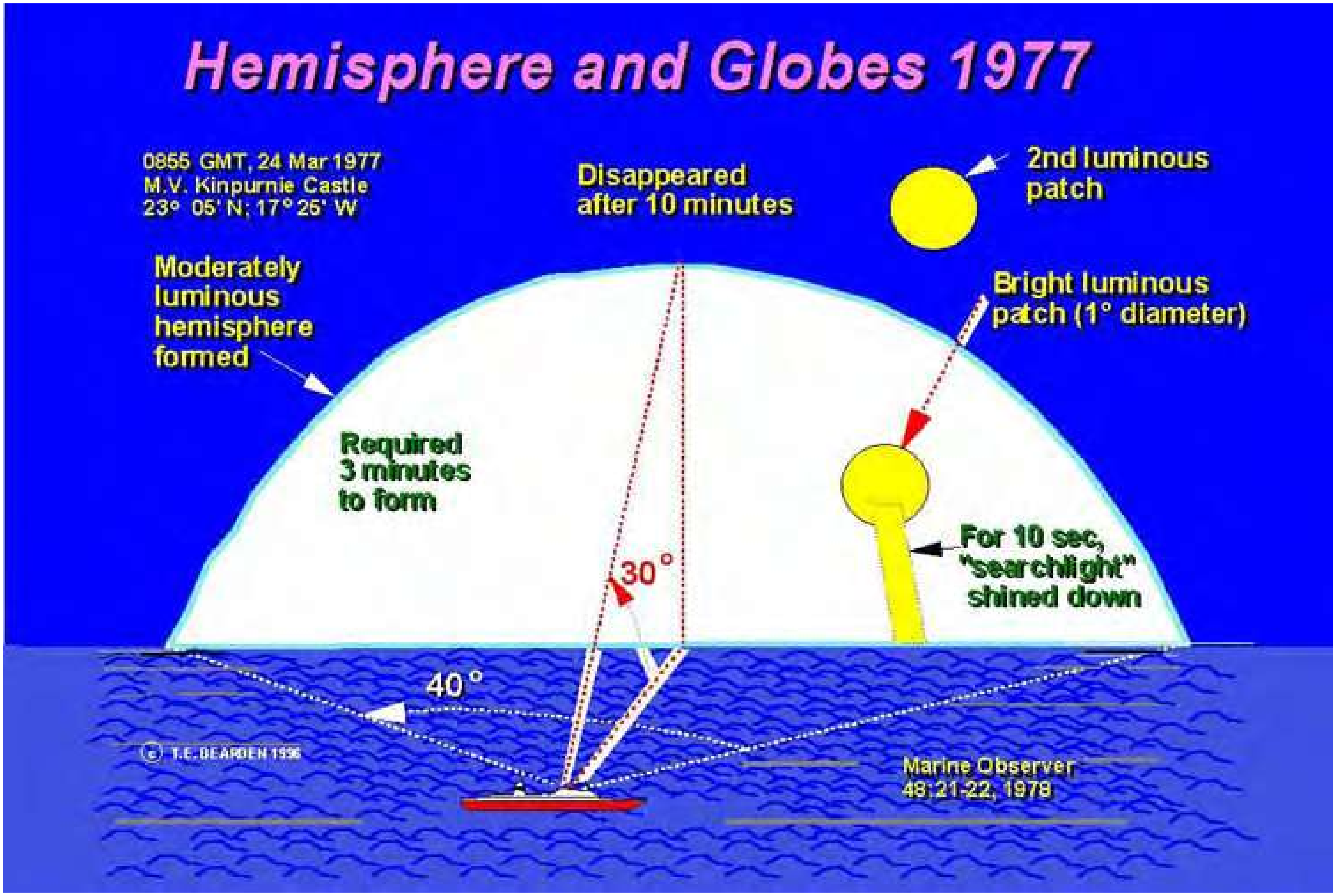
Hemisphere and Globes 1977
0855 GMT, 24 Mar 1977
M.V. Kinpurnie Castle
23° 05' N; 17° 25' W
Disappeared after 10 minutes
2nd luminous patch
Moderately luminous hemisphere formed
Bright luminous patch (1° diameter)
Required 3 minutes to form
For 10 sec, "searchlight" shined down
30°
40°
© T.E. BEARDEN 1996
Marine Observer 48:21-22, 1978

1969 Virgin Islands Incident

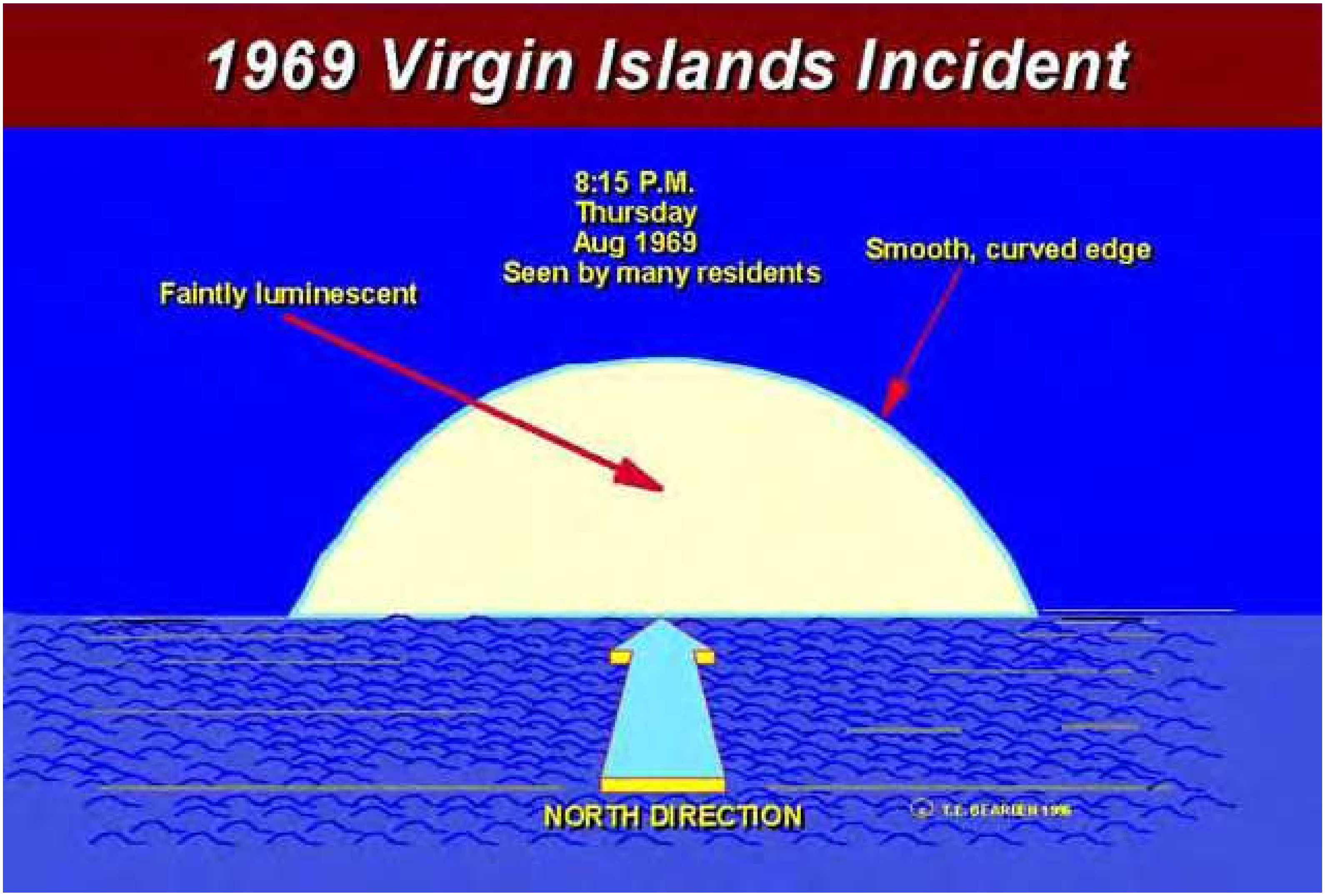

8:15 P.M.
Thursday
Aug 1969
Seen by many residents

Faintly luminescent
Smooth, curved edge
NORTH DIRECTION

© T.E. BEARDEN 1996

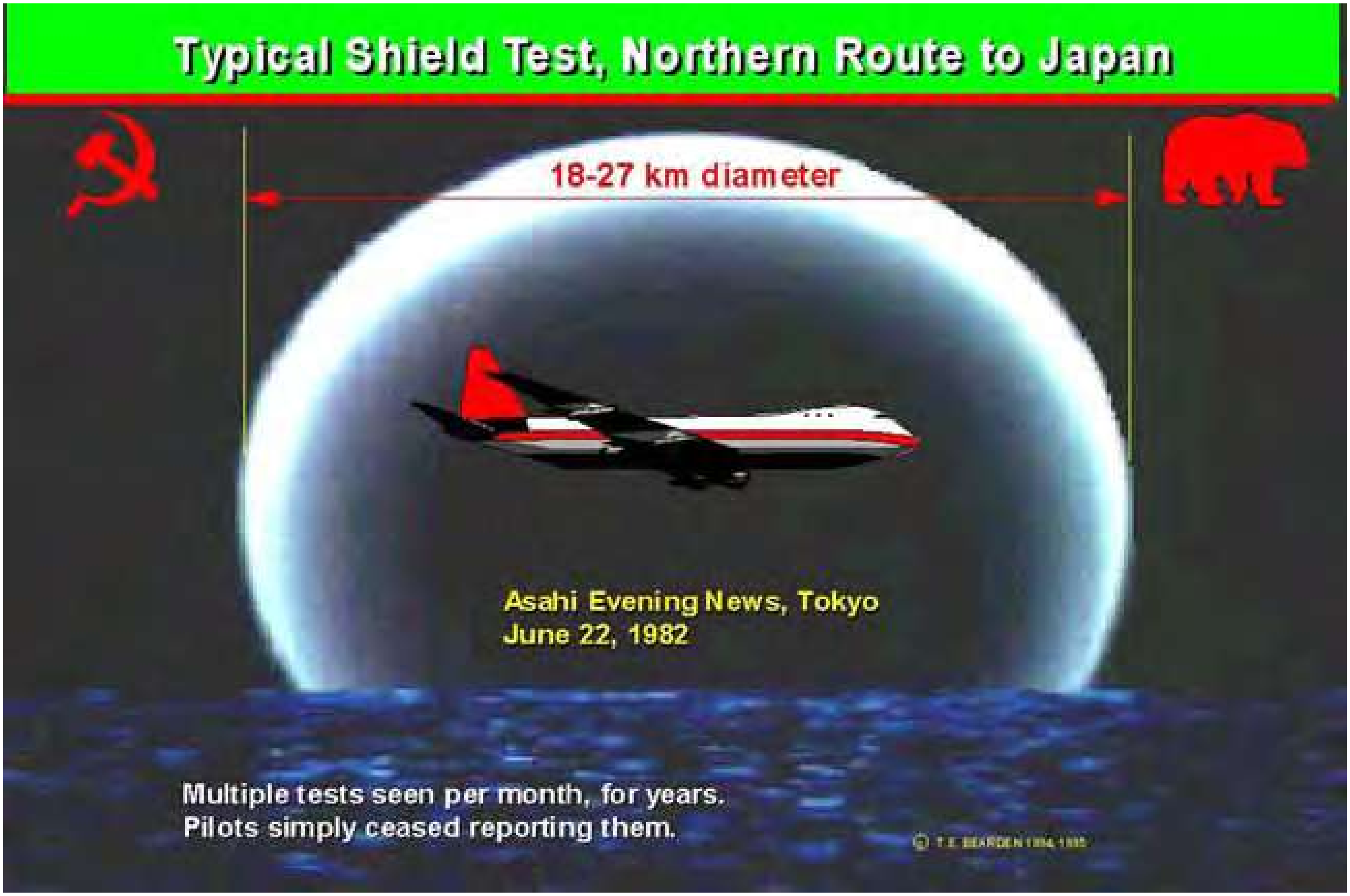

Typical Shield Test, Northern Route to Japan

18-27 km diameter

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T.E. BEARDEN 1994, 1995

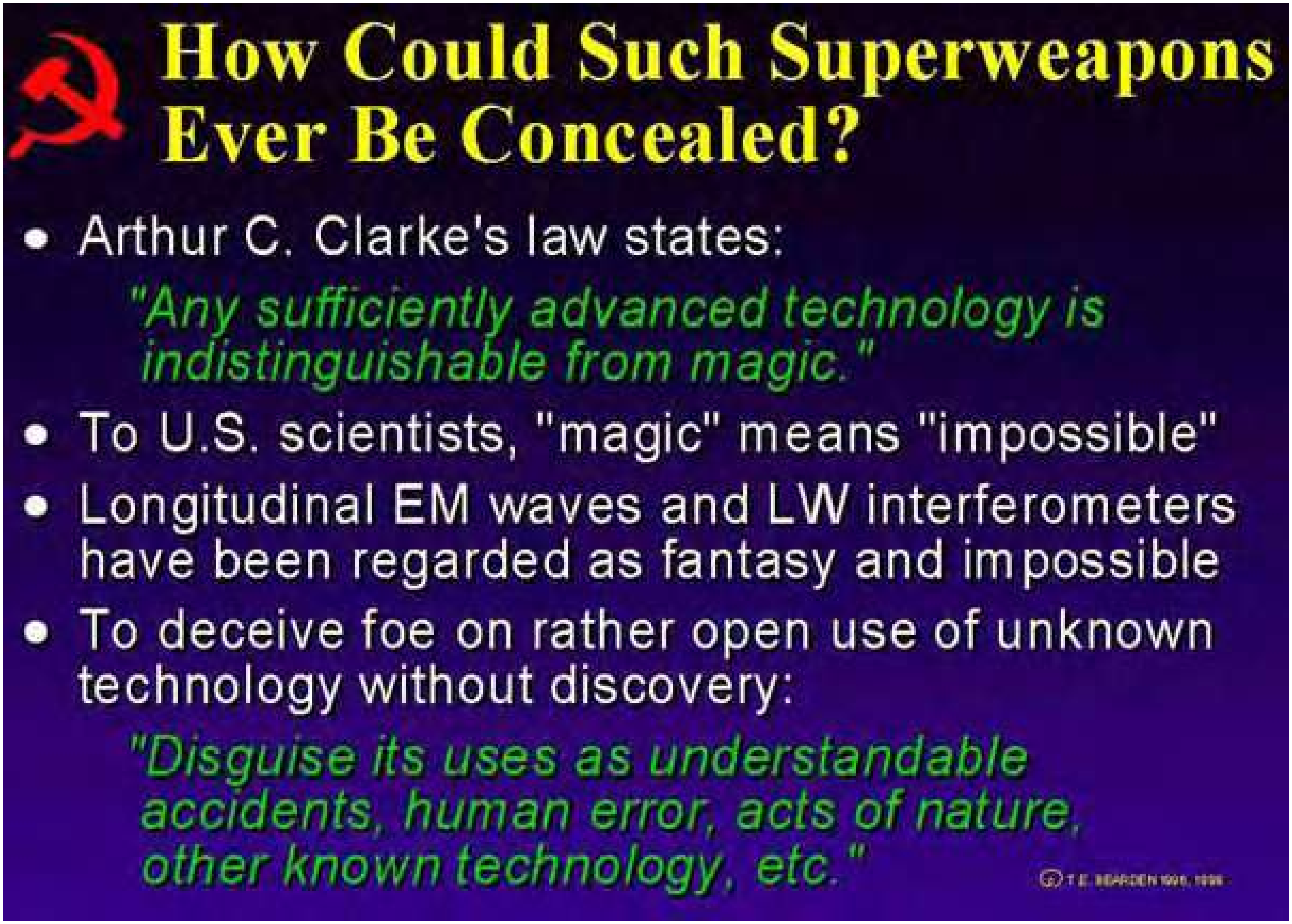

How Could Such Superweapons Ever Be Concealed?

- Arthur C. Clarke's law states:

  "Any sufficiently advanced technology is indistinguishable from magic."

- To U.S. scientists, "magic" means "impossible"
- Longitudinal EM waves and LW interferometers have been regarded as fantasy and impossible
- To deceive foe on rather open use of unknown technology without discovery:

  "Disguise its uses as understandable accidents, human error, acts of nature, other known technology, etc."

© T.E. BEARDEN [illegible]

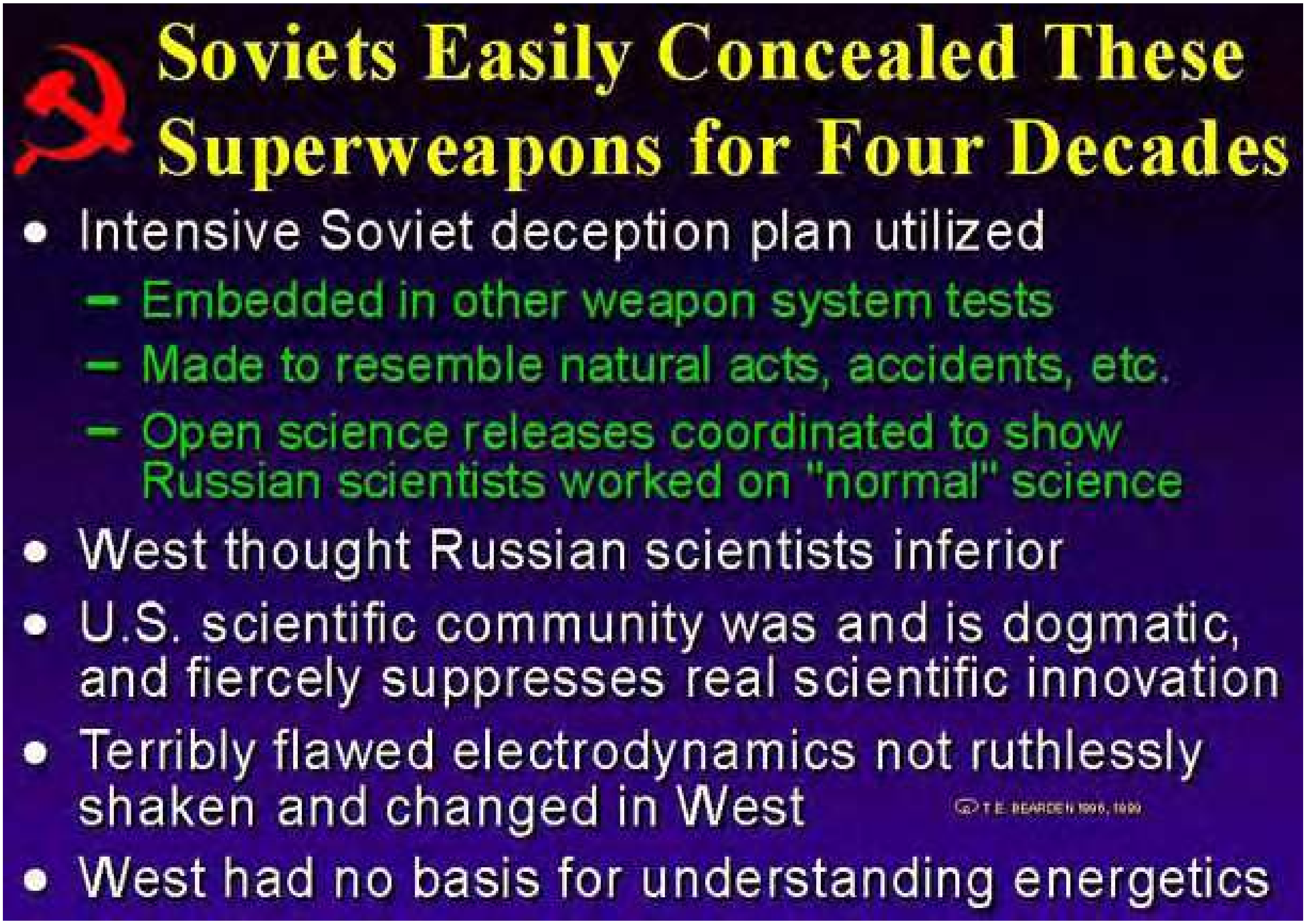

Soviets Easily Concealed These Superweapons for Four Decades

- Intensive Soviet deception plan utilized
  - Embedded in other weapon system tests
  - Made to resemble natural acts, accidents, etc.
  - Open science releases coordinated to show Russian scientists worked on "normal" science
- West thought Russian scientists inferior
- U.S. scientific community was and is dogmatic, and fiercely suppresses real scientific innovation
- Terribly flawed electrodynamics not ruthlessly shaken and changed in West
- West had no basis for understanding energetics

© T.E. BEARDEN 1996, 1999

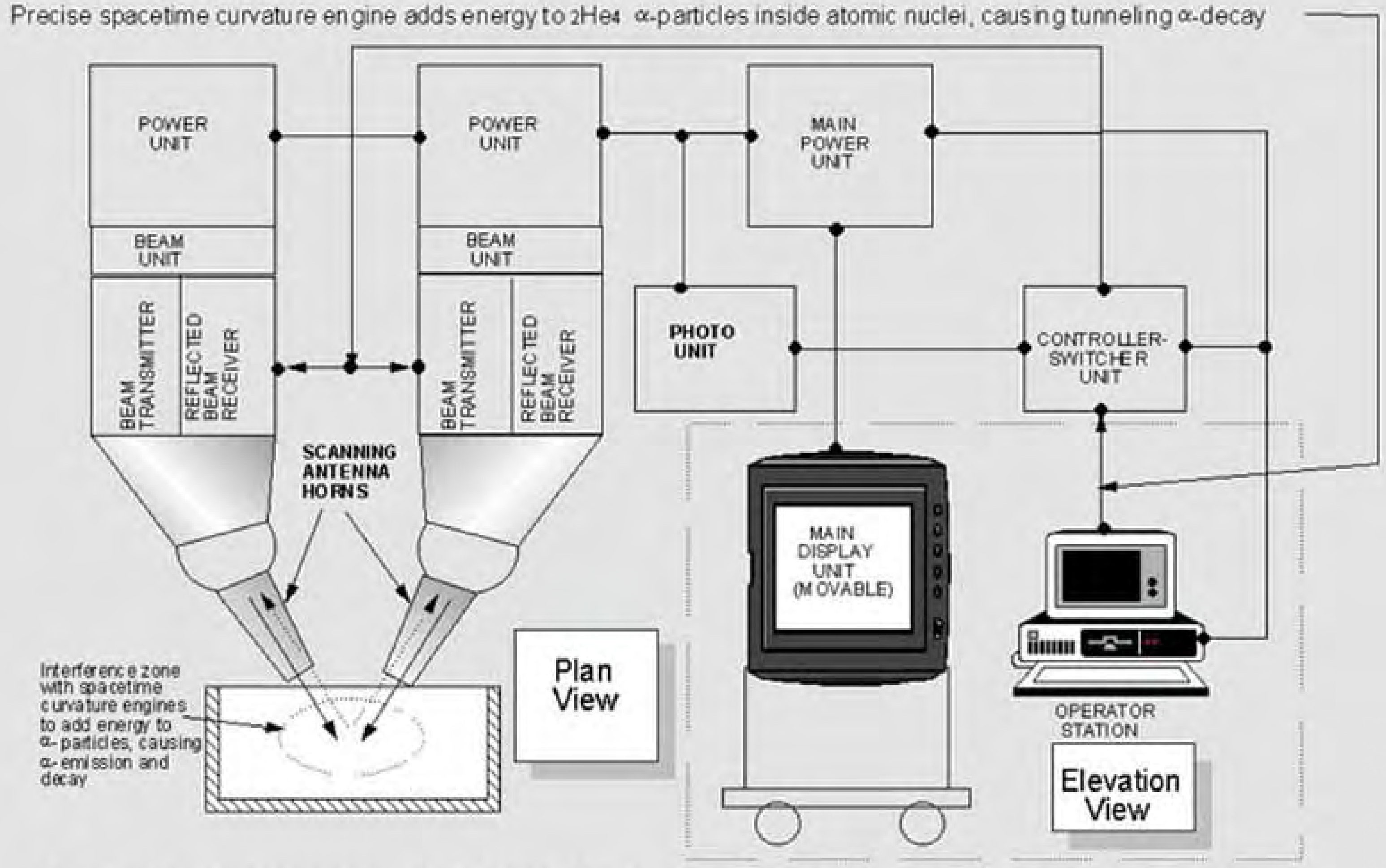
Precise spacetime curvature engine adds energy to $_2He_4$ α-particles inside atomic nuclei, causing tunneling α-decay
POWER UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
POWER UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
MAIN POWER UNIT
PHOTO UNIT
CONTROLLER-SWITCHER UNIT
SCANNING ANTENNA HORNS
MAIN DISPLAY UNIT (MOVABLE)
Interference zone with spacetime curvature engines to add energy to α-particles, causing α-emission and decay
Plan View
OPERATOR STATION
Elevation View

Figure 26. Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes.

© T.E. BEARDEN 1996

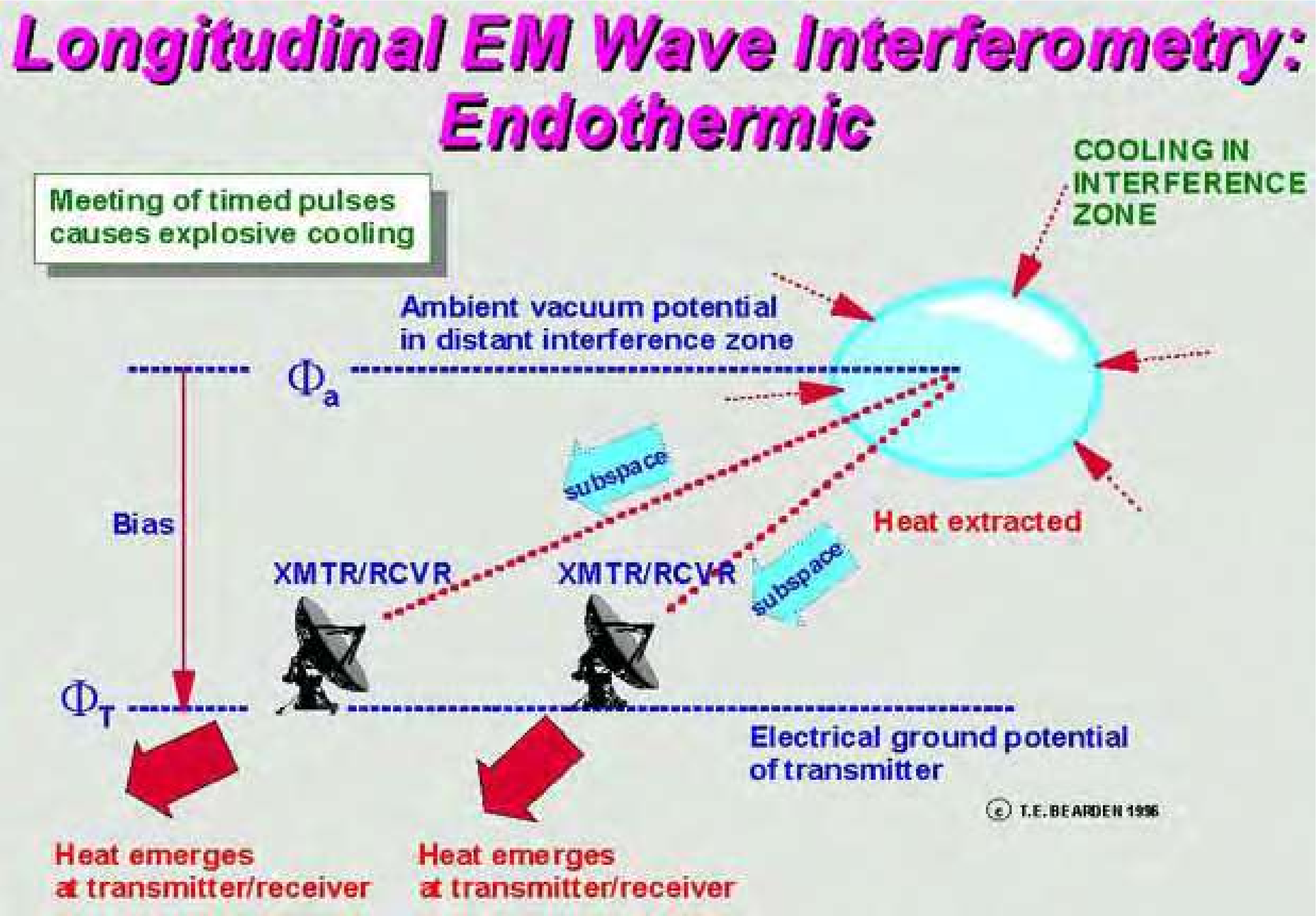

Longitudinal EM Wave Interferometry: Endothermic

Meeting of timed pulses
causes explosive cooling

COOLING IN
INTERFERENCE
ZONE

Ambient vacuum potential
in distant interference zone

Heat extracted

Electrical ground potential
of transmitter

Heat emerges
at transmitter/receiver

Heat emerges
at transmitter/receiver

Φ_a
subspace
Bias
XMTR/RCVR
XMTR/RCVR
subspace
Φ_T

© T.E. BEARDEN 1996

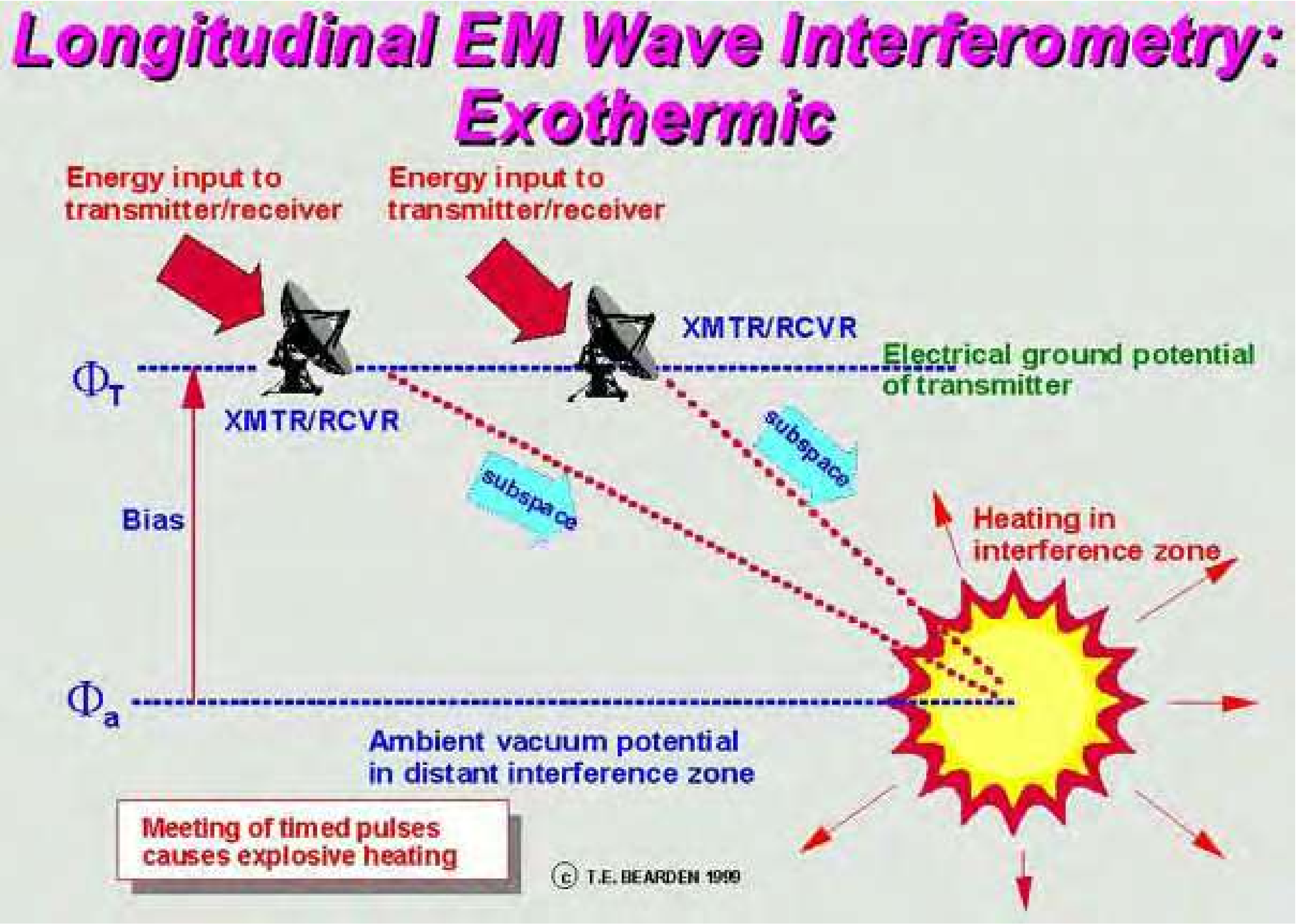

Longitudinal EM Wave Interferometry: Exothermic

Energy input to transmitter/receiver
Energy input to transmitter/receiver
XMTR/RCVR
Φ_T
XMTR/RCVR
Electrical ground potential of transmitter
subspace
subspace
Bias
Heating in interference zone
Φ_a
Ambient vacuum potential in distant interference zone

Meeting of timed pulses causes explosive heating

ⓒ T.E. BEARDEN 1999

The Tom Bearden Website

Help support the research

U.S. Air Attack on Libya
April 1986

- Slide 1
- Slide 2
- "Washington Post" Article
  January 22, 1989

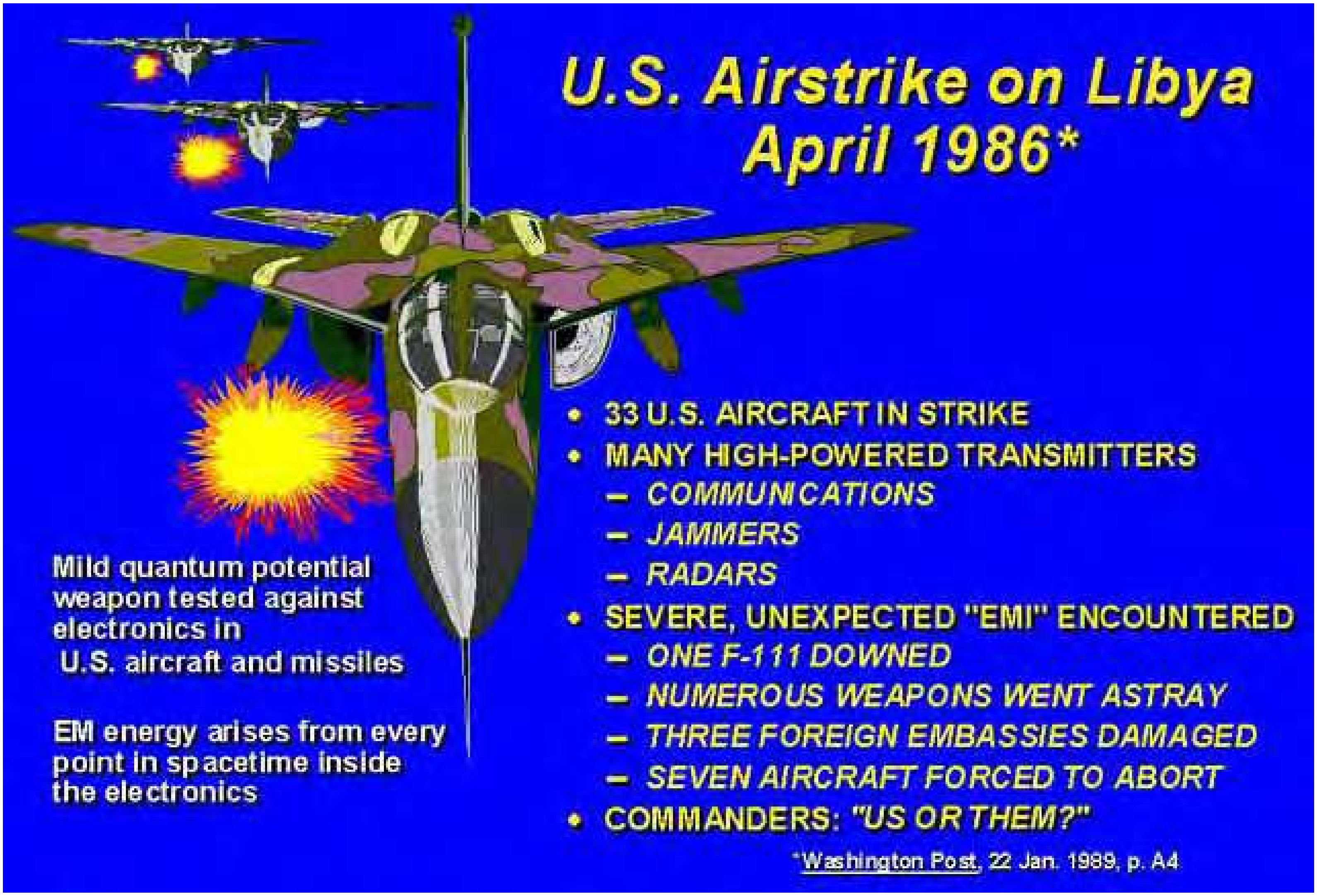

U.S. Airstrike on Libya April 1986*

- 33 U.S. AIRCRAFT IN STRIKE
- MANY HIGH-POWERED TRANSMITTERS
  - *COMMUNICATIONS*
  - *JAMMERS*
  - *RADARS*
- SEVERE, UNEXPECTED "EMI" ENCOUNTERED
  - *ONE F-111 DOWNED*
  - *NUMEROUS WEAPONS WENT ASTRAY*
  - *THREE FOREIGN EMBASSIES DAMAGED*
  - *SEVEN AIRCRAFT FORCED TO ABORT*
- COMMANDERS: *"US OR THEM?"*

*Washington Post, 22 Jan. 1989, p. A4

Mild quantum potential weapon tested against electronics in U.S. aircraft and missiles

EM energy arises from every point in spacetime inside the electronics

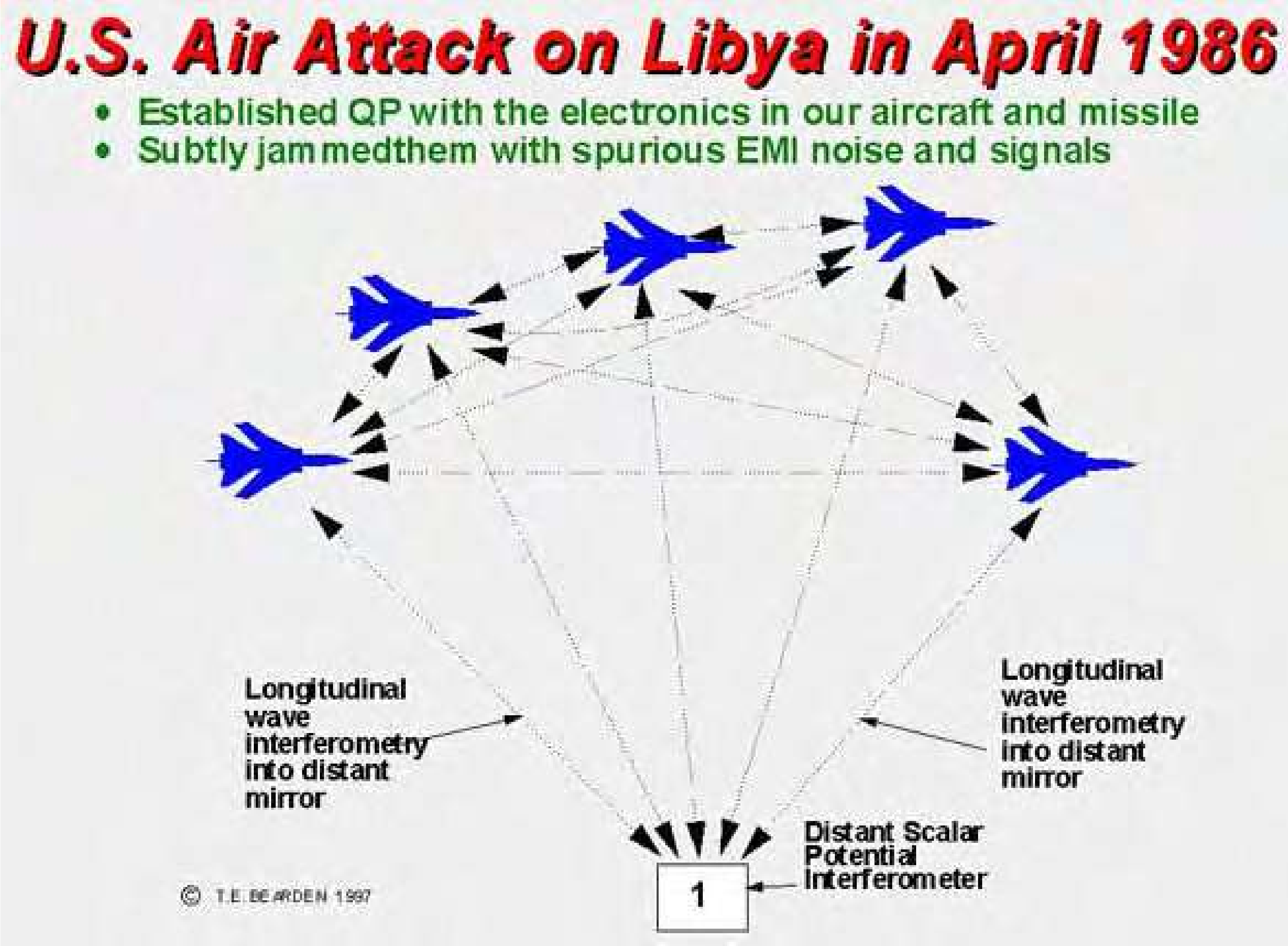

U.S. Air Attack on Libya in April 1986

- Established QP with the electronics in our aircraft and missile
- Subtly jammedthem with spurious EMI noise and signals

Longitudinal wave interferometry into distant mirror

Longitudinal wave interferometry into distant mirror

Distant Scalar Potential Interferometer

© T.E. BEARDEN 1997

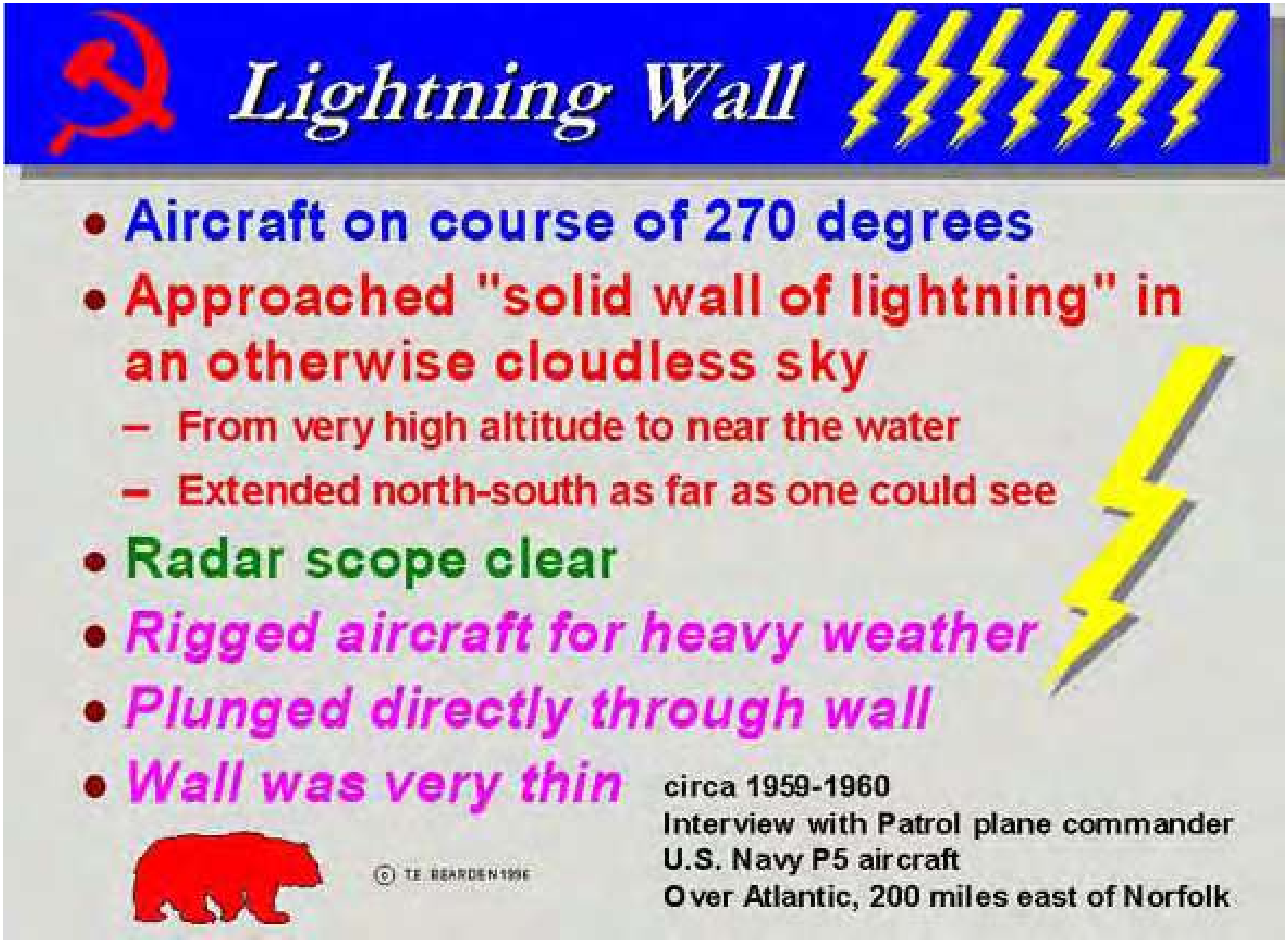

Lightning Wall

- Aircraft on course of 270 degrees
- Approached "solid wall of lightning" in an otherwise cloudless sky
  - From very high altitude to near the water
  - Extended north-south as far as one could see
- Radar scope clear
- *Rigged aircraft for heavy weather*
- *Plunged directly through wall*
- *Wall was very thin*

circa 1959-1960
Interview with Patrol plane commander
U.S. Navy P5 aircraft
Over Atlantic, 200 miles east of Norfolk

© T.E. BEARDEN 1996

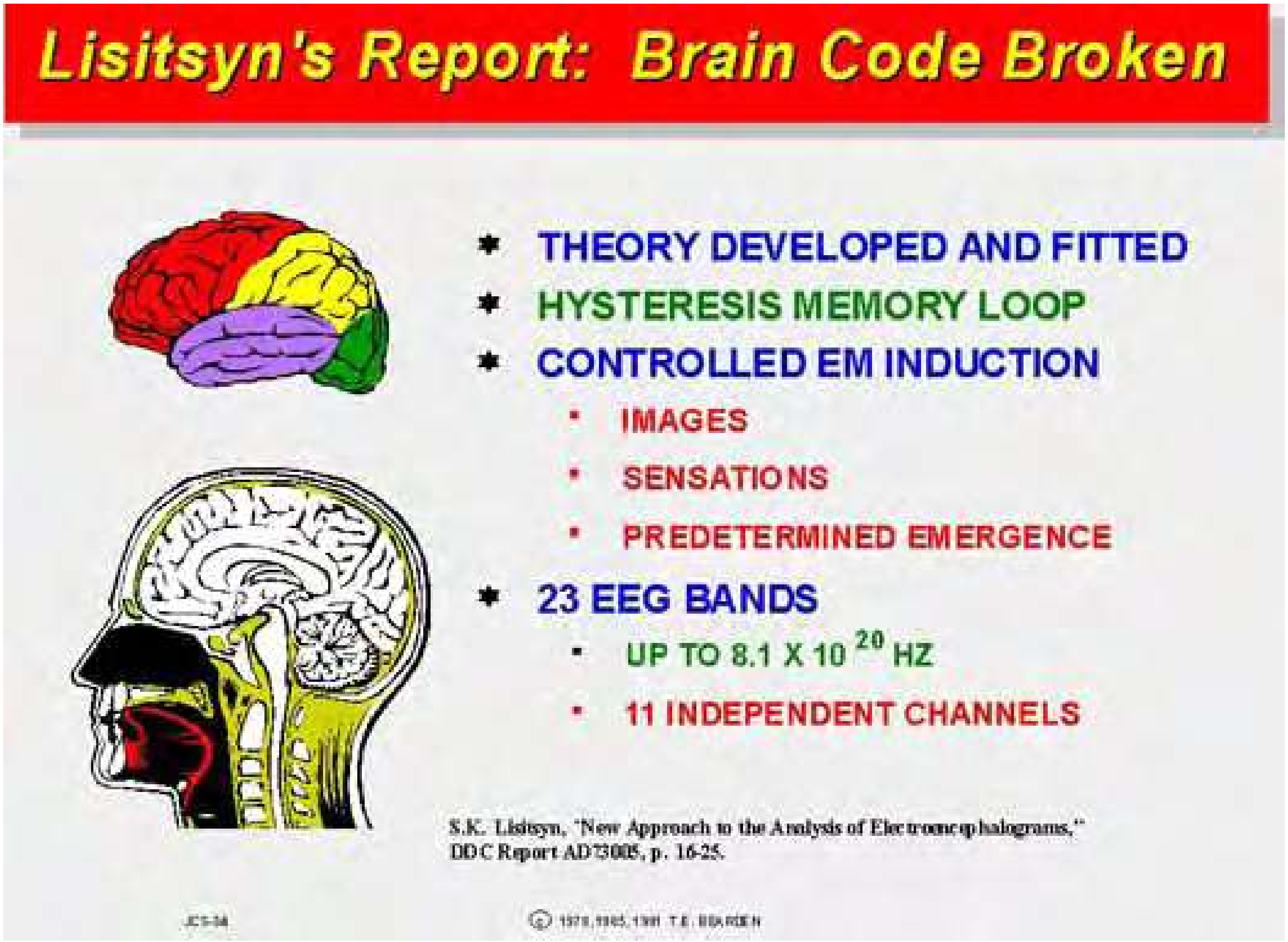

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

JCS-34

© 1978, 1985, 1991 T.E. BEARDEN

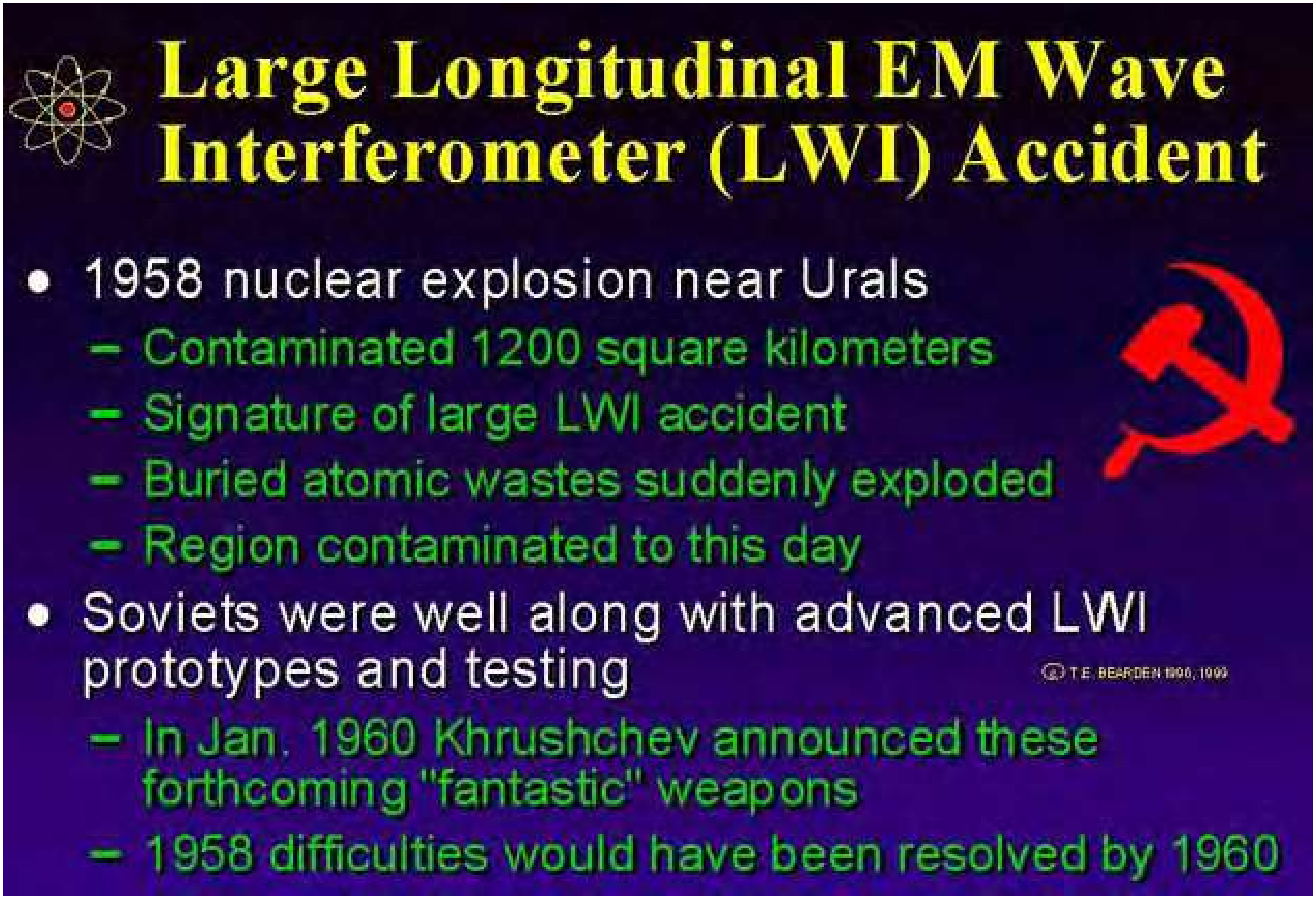

Large Longitudinal EM Wave Interferometer (LWI) Accident

- 1958 nuclear explosion near Urals
  - Contaminated 1200 square kilometers
  - Signature of large LWI accident
  - Buried atomic wastes suddenly exploded
  - Region contaminated to this day
- Soviets were well along with advanced LWI prototypes and testing
  - In Jan. 1960 Khrushchev announced these forthcoming "fantastic" weapons
  - 1958 difficulties would have been resolved by 1960

© T.E. BEARDEN 1996, 1999

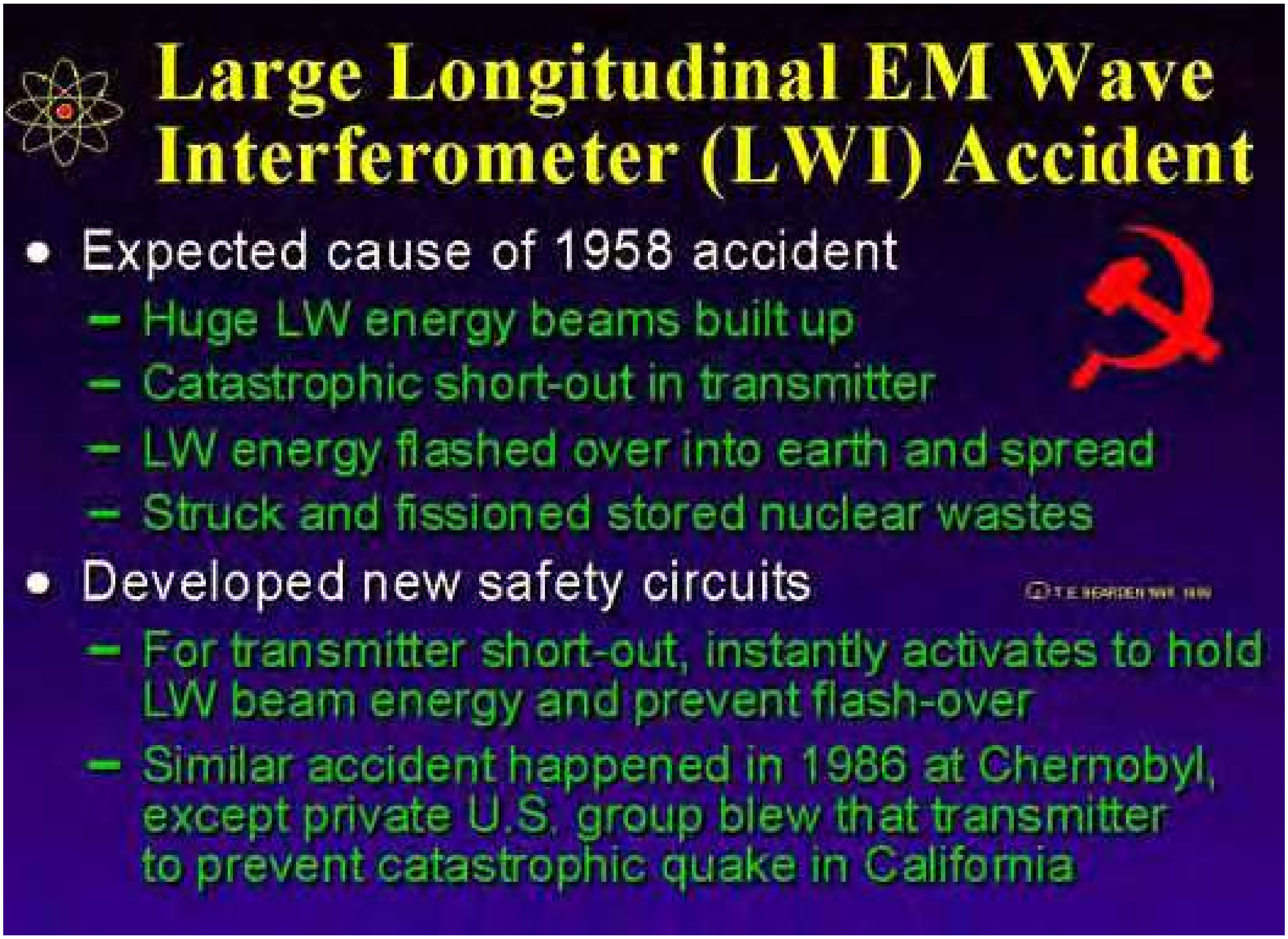

Large Longitudinal EM Wave Interferometer (LWI) Accident

- Expected cause of 1958 accident
  - Huge LW energy beams built up
  - Catastrophic short-out in transmitter
  - LW energy flashed over into earth and spread
  - Struck and fissioned stored nuclear wastes
- Developed new safety circuits
  - For transmitter short-out, instantly activates to hold LW beam energy and prevent flash-over
  - Similar accident happened in 1986 at Chernobyl, except private U.S. group blew that transmitter to prevent catastrophic quake in California

(Use of Hidden Information Content of the Field)
(Can Provide Action-at-a-Distance)

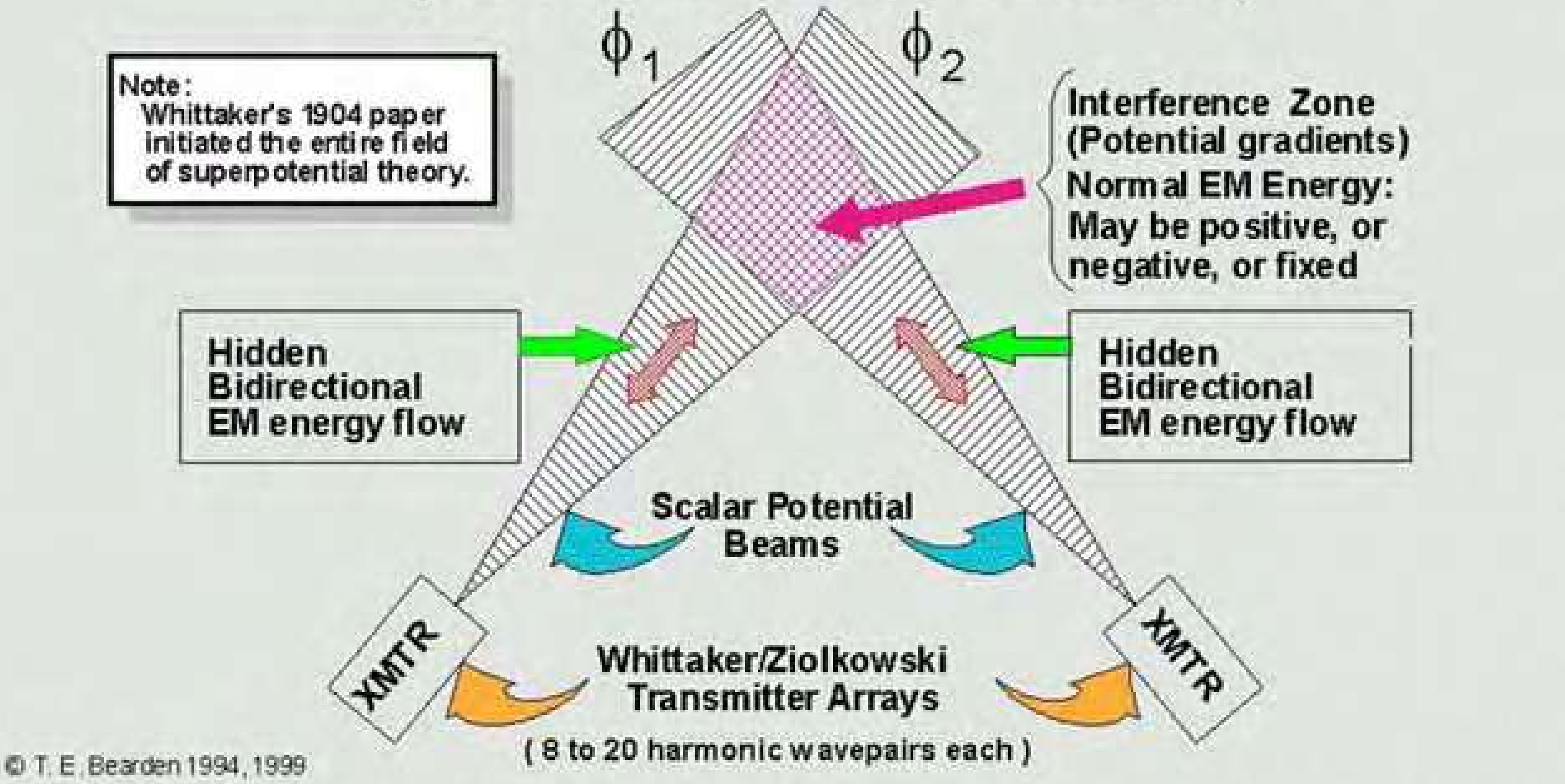
ϕ_1
ϕ_2
Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

© T. E. Bearden 1994, 1999

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepair functions) produces all EM force fields and waves.

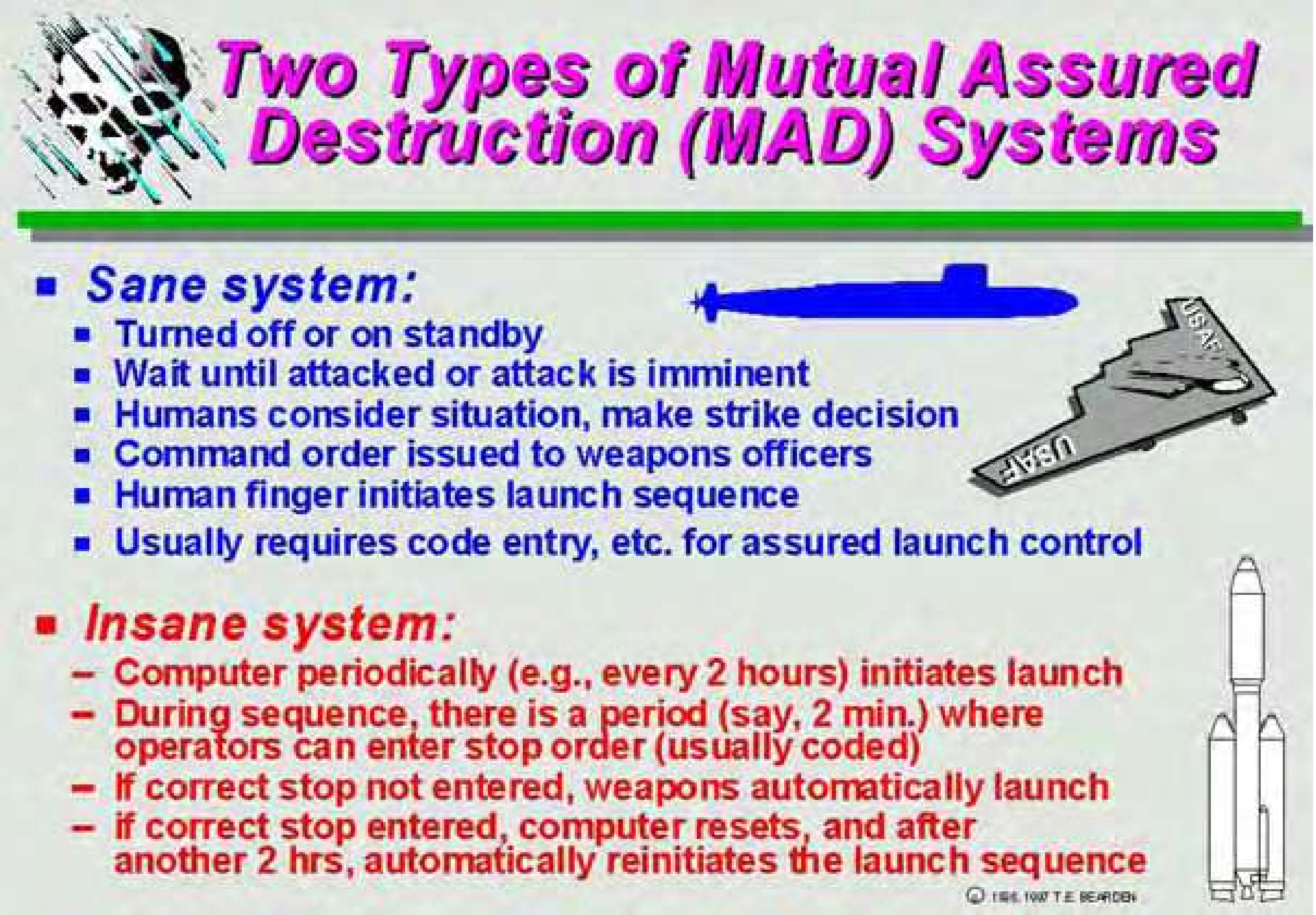

Two Types of Mutual Assured Destruction (MAD) Systems

- ***Sane system:***
  - Turned off or on standby
  - Wait until attacked or attack is imminent
  - Humans consider situation, make strike decision
  - Command order issued to weapons officers
  - Human finger initiates launch sequence
  - Usually requires code entry, etc. for assured launch control
- ***Insane system:***
  - Computer periodically (e.g., every 2 hours) initiates launch
  - During sequence, there is a period (say, 2 min.) where operators can enter stop order (usually coded)
  - If correct stop not entered, weapons automatically launch
  - If correct stop entered, computer resets, and after another 2 hrs, automatically reinitiates the launch sequence

© 1996, 1997 T.E. BEARDEN

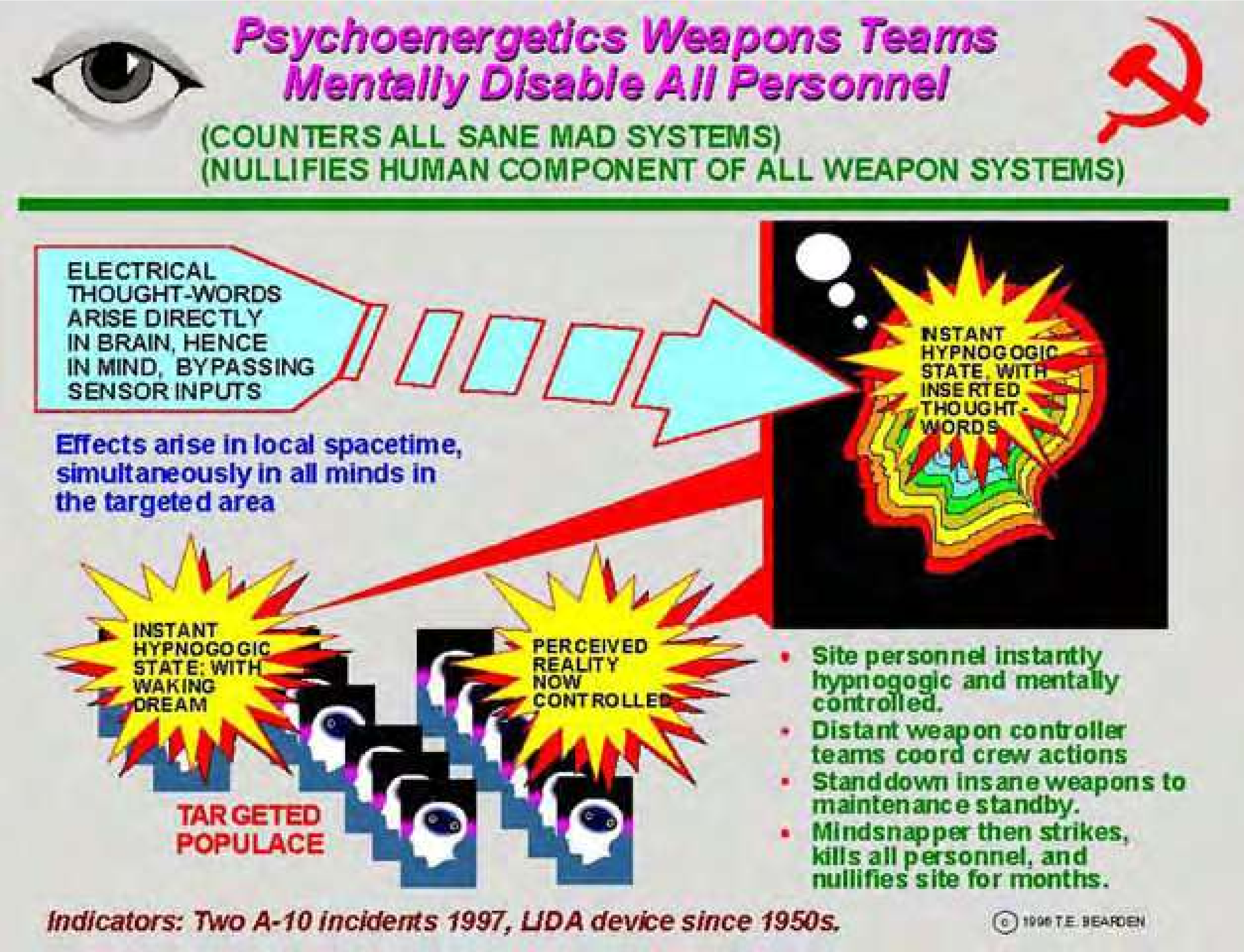

Psychoenergetics Weapons Teams Mentally Disable All Personnel

(COUNTERS ALL SANE MAD SYSTEMS)
(NULLIFIES HUMAN COMPONENT OF ALL WEAPON SYSTEMS)

ELECTRICAL THOUGHT-WORDS ARISE DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED THOUGHT-WORDS

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED POPULACE

- Site personnel instantly hypnogogic and mentally controlled.
- Distant weapon controller teams coord crew actions
- Standdown insane weapons to maintenance standby.
- Mindsnapper then strikes, kills all personnel, and nullifies site for months.

Indicators: Two A-10 incidents 1997, LIDA device since 1950s.

(c) 1998 T.E. BEARDEN

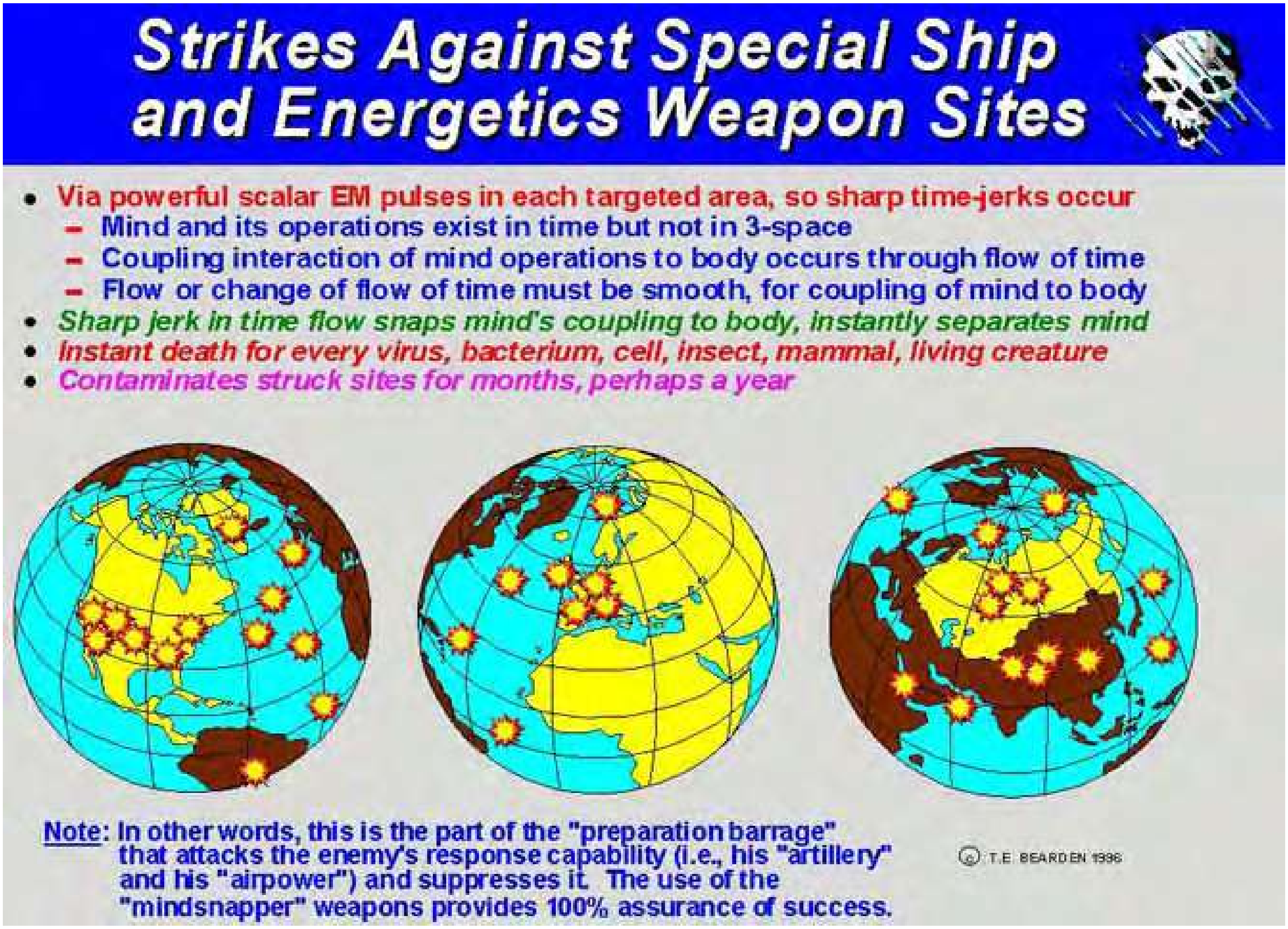

Strikes Against Special Ship and Energetics Weapon Sites

- Via powerful scalar EM pulses in each targeted area, so sharp time-jerks occur
  - Mind and its operations exist in time but not in 3-space
  - Coupling interaction of mind operations to body occurs through flow of time
  - Flow or change of flow of time must be smooth, for coupling of mind to body
- *Sharp jerk in time flow snaps mind's coupling to body, instantly separates mind*
- *Instant death for every virus, bacterium, cell, insect, mammal, living creature*
- *Contaminates struck sites for months, perhaps a year*

Note: In other words, this is the part of the "preparation barrage" that attacks the enemy's response capability (i.e., his "artillery" and his "airpower") and suppresses it. The use of the "mindsnapper" weapons provides 100% assurance of success.

© T.E. BEARDEN 1996

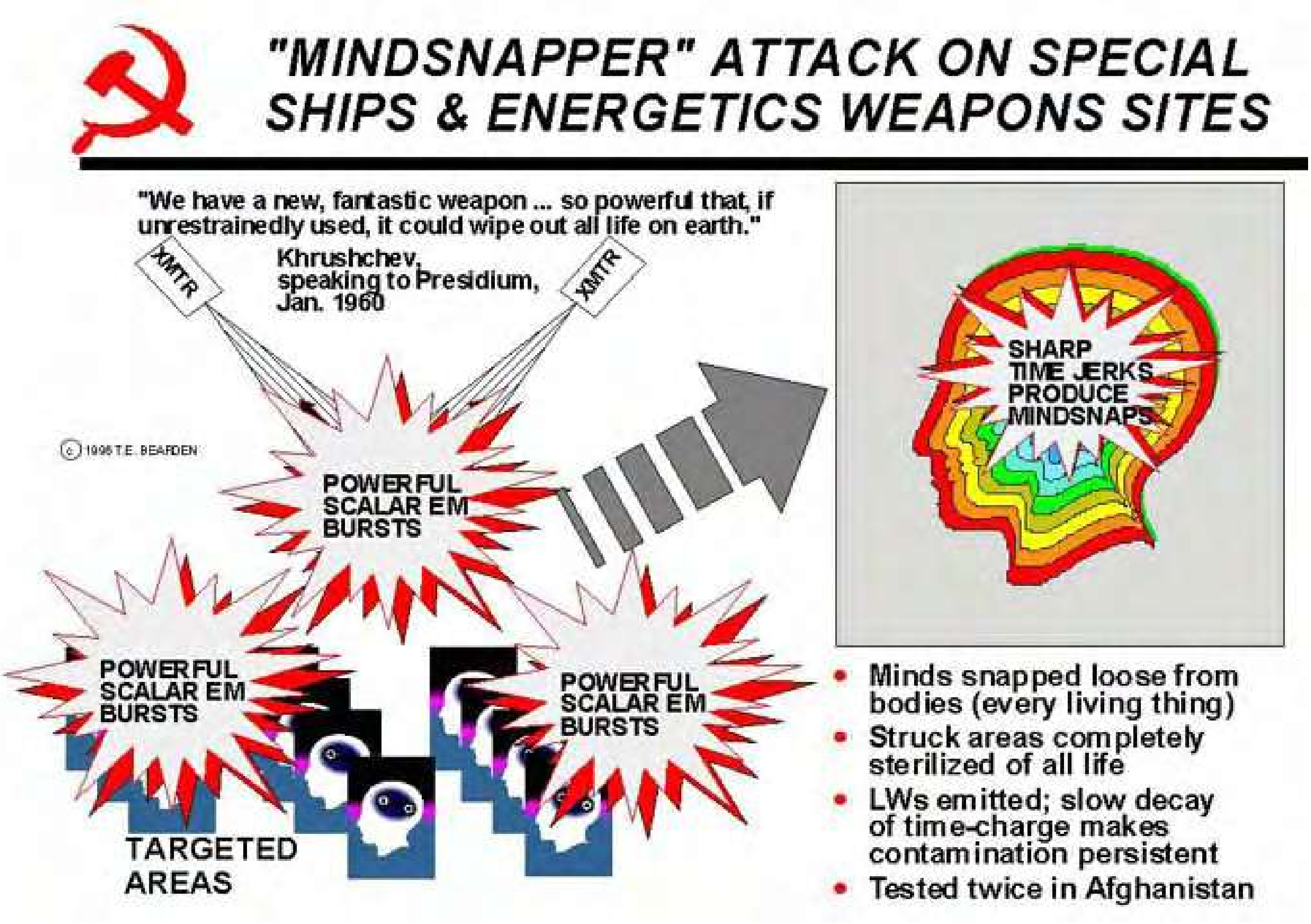

"MINDSNAPPER" ATTACK ON SPECIAL SHIPS & ENERGETICS WEAPONS SITES

"We have a new, fantastic weapon ... so powerful that, if unrestrainedly used, it could wipe out all life on earth."

Khrushchev,
speaking to Presidium,
Jan. 1960

XMTR
XMTR
© 1996 T.E. BEARDEN
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
TARGETED AREAS
SHARP TIME JERKS PRODUCE MINDSNAPS

- Minds snapped loose from bodies (every living thing)
- Struck areas completely sterilized of all life
- LWs emitted; slow decay of time-charge makes contamination persistent
- Tested twice in Afghanistan

How Much Off Guard Are We?

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6
- Slide 7

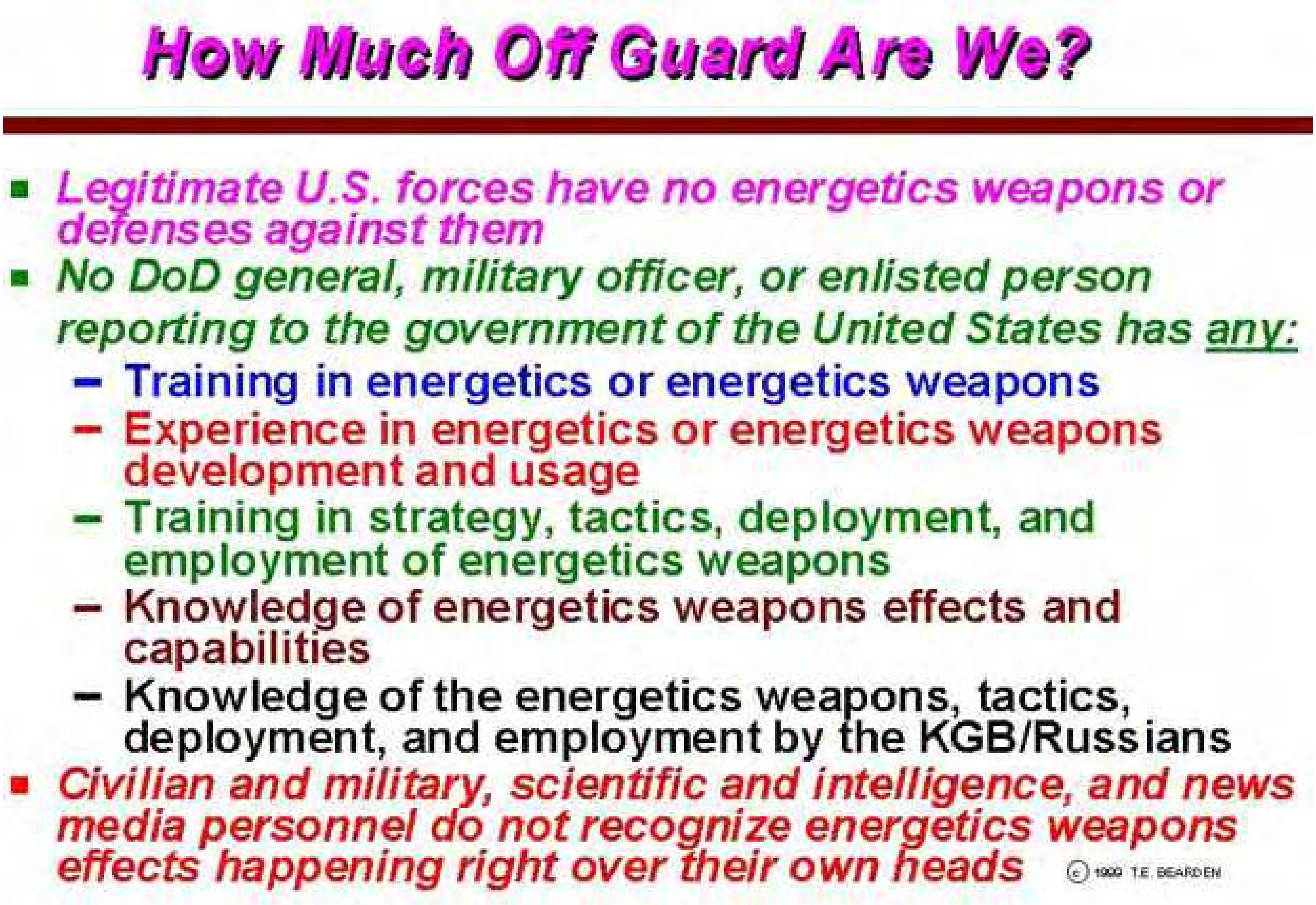

How Much Off Guard Are We?

- *Legitimate U.S. forces have no energetics weapons or defenses against them*
- *No DoD general, military officer, or enlisted person reporting to the government of the United States has any:*
  - Training in energetics or energetics weapons
  - Experience in energetics or energetics weapons development and usage
  - Training in strategy, tactics, deployment, and employment of energetics weapons
  - Knowledge of energetics weapons effects and capabilities
  - Knowledge of the energetics weapons, tactics, deployment, and employment by the KGB/Russians
- *Civilian and military, scientific and intelligence, and news media personnel do not recognize energetics weapons effects happening right over their own heads*

© 1999 T.E. BEARDEN

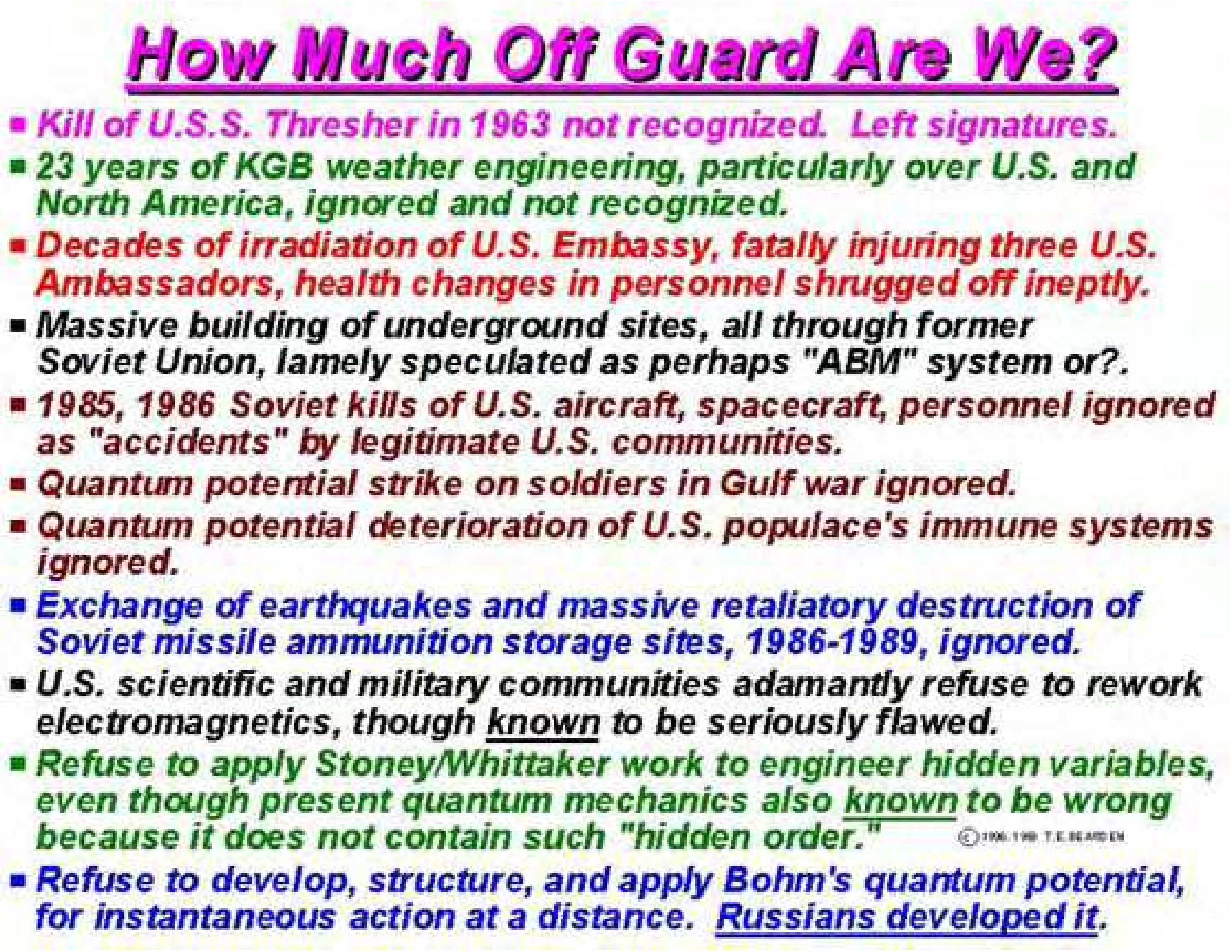

How Much Off Guard Are We?

- Kill of U.S.S. Thresher in 1963 not recognized. Left signatures.
- 23 years of KGB weather engineering, particularly over U.S. and North America, ignored and not recognized.
- Decades of irradiation of U.S. Embassy, fatally injuring three U.S. Ambassadors, health changes in personnel shrugged off ineptly.
- Massive building of underground sites, all through former Soviet Union, lamely speculated as perhaps "ABM" system or?.
- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.
- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."

© 1996-1998 T.E. BEARDEN

- Refuse to develop, structure, and apply Bohm's quantum potential, for instantaneous action at a distance. Russians developed it.

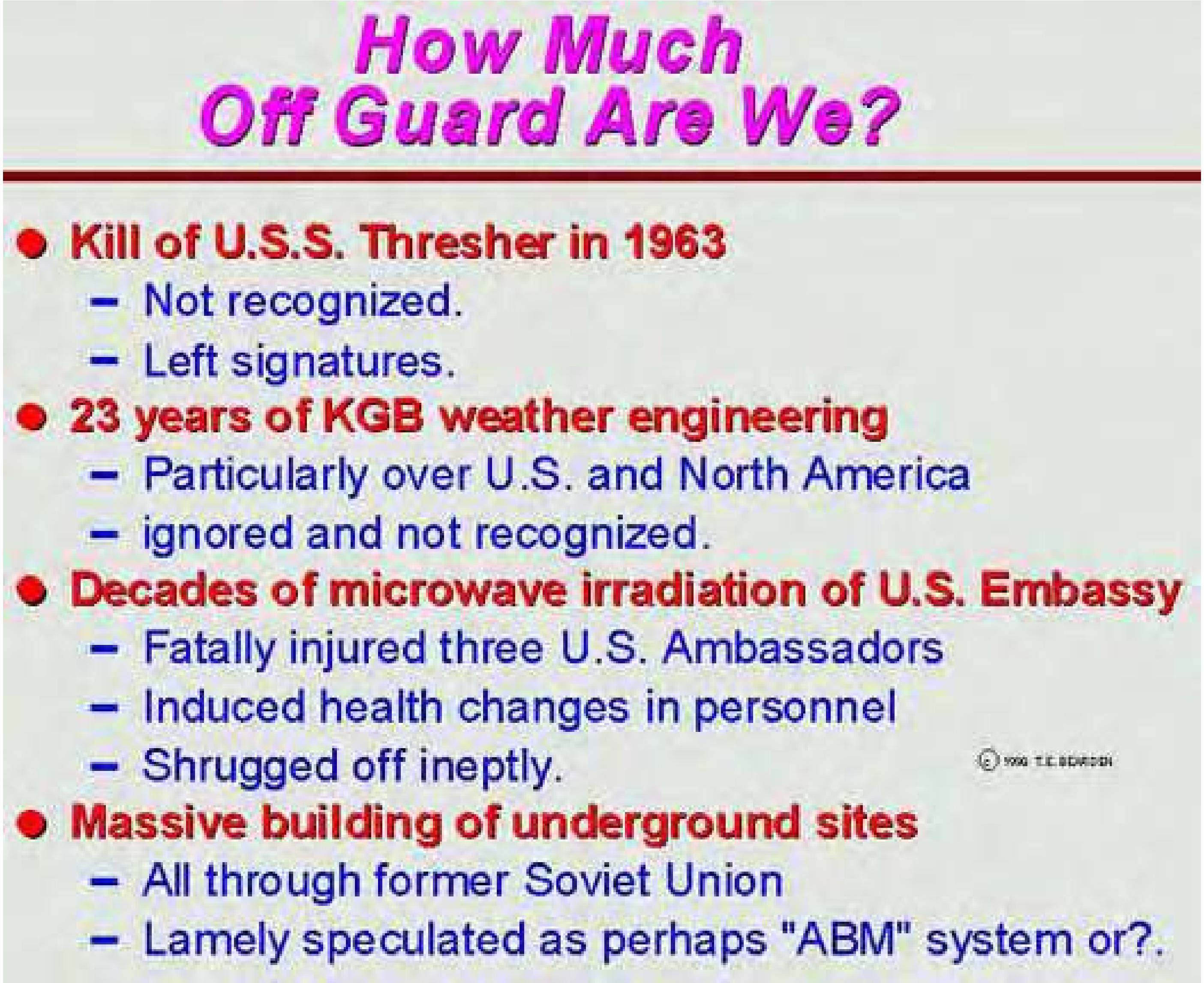

How Much Off Guard Are We?

- **Kill of U.S.S. Thresher in 1963**
  - Not recognized.
  - Left signatures.
- **23 years of KGB weather engineering**
  - Particularly over U.S. and North America
  - ignored and not recognized.
- **Decades of microwave irradiation of U.S. Embassy**
  - Fatally injured three U.S. Ambassadors
  - Induced health changes in personnel
  - Shrugged off ineptly.
- **Massive building of underground sites**
  - All through former Soviet Union
  - Lamely speculated as perhaps "ABM" system or?.

© 1990 T.E. BEARDEN

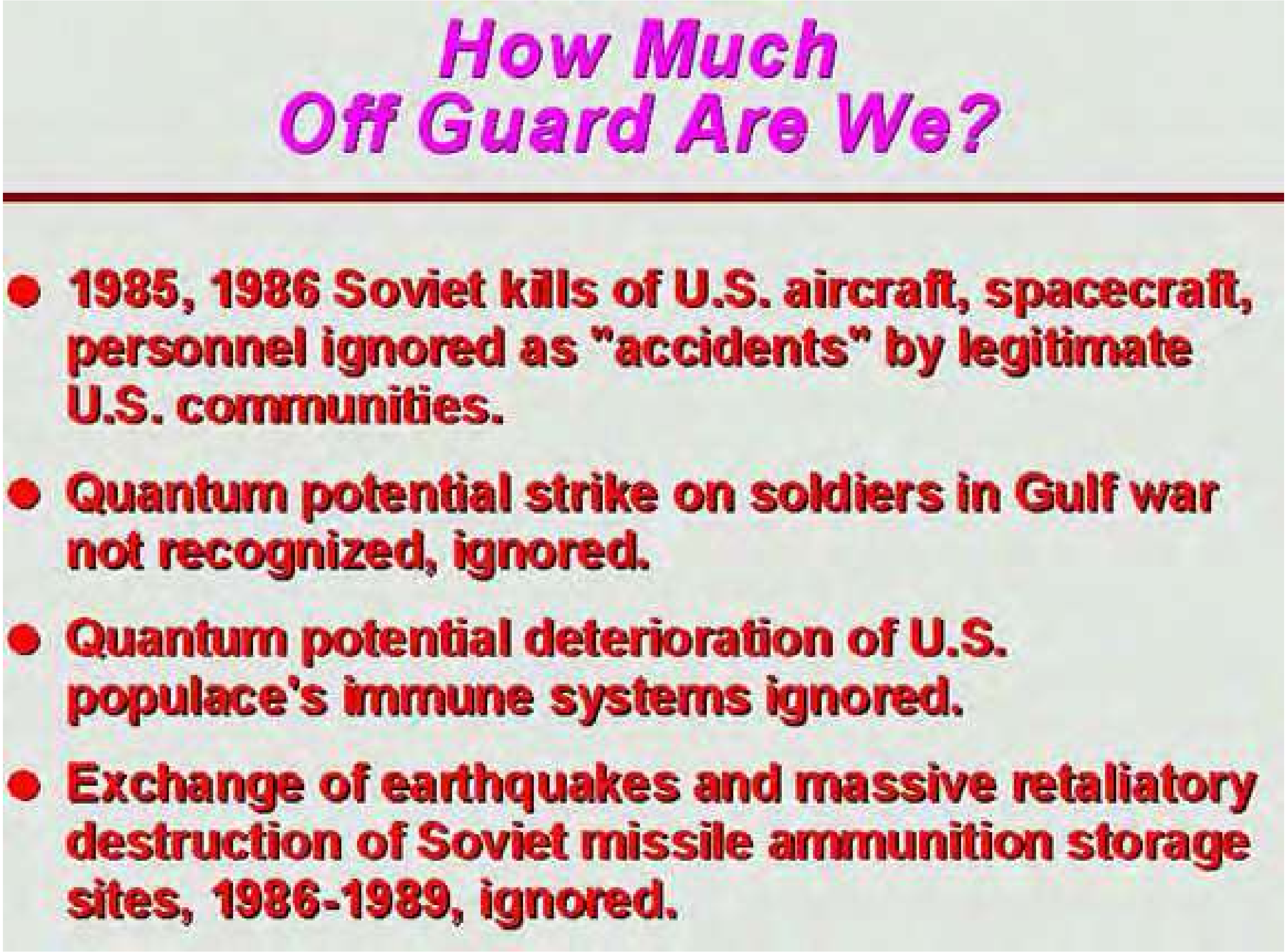

How Much Off Guard Are We?

- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war not recognized, ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.

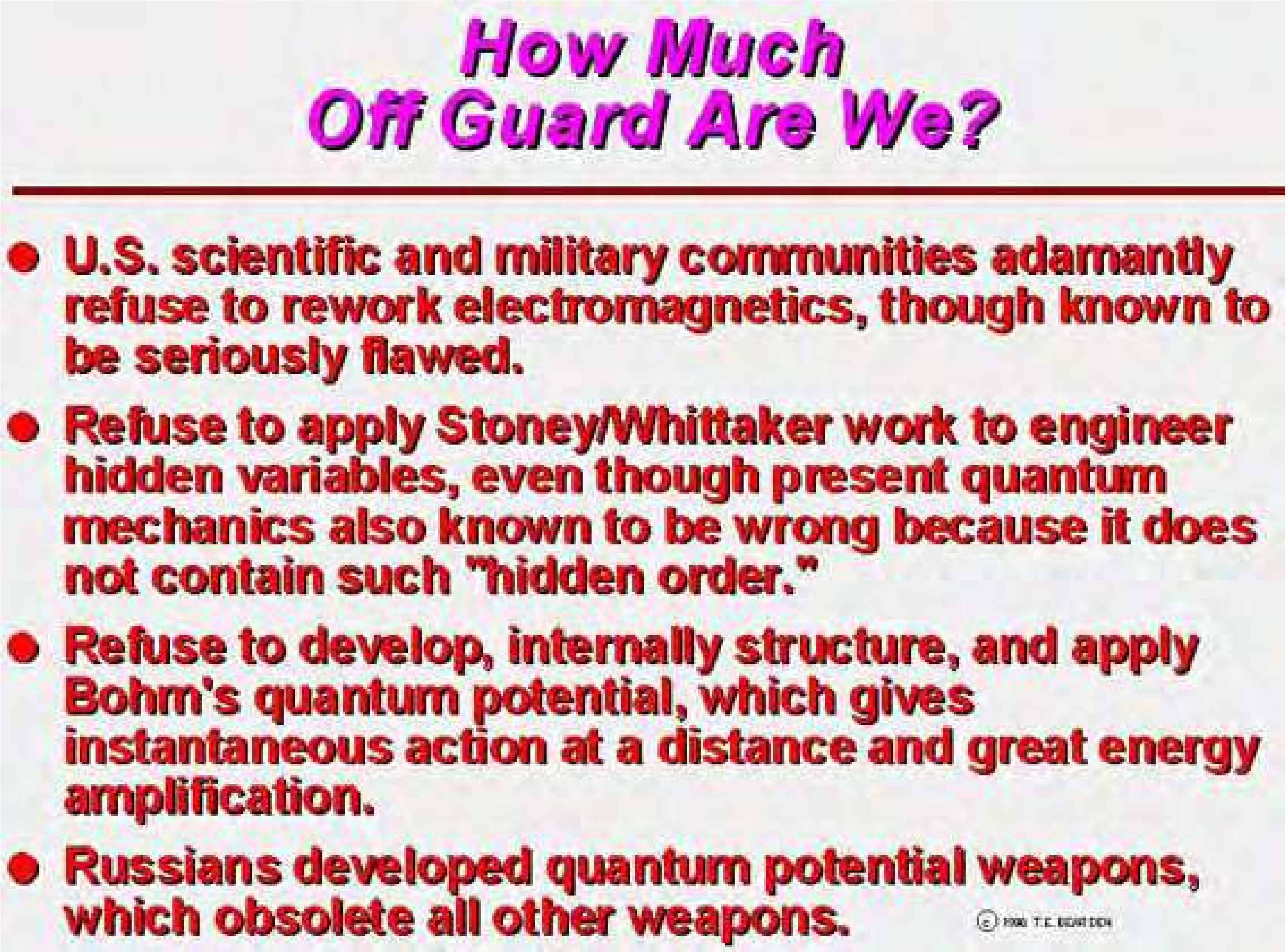

How Much Off Guard Are We?

- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."
- Refuse to develop, internally structure, and apply Bohm's quantum potential, which gives instantaneous action at a distance and great energy amplification.
- Russians developed quantum potential weapons, which obsolete all other weapons.

© [illegible] T.E. BEARDEN

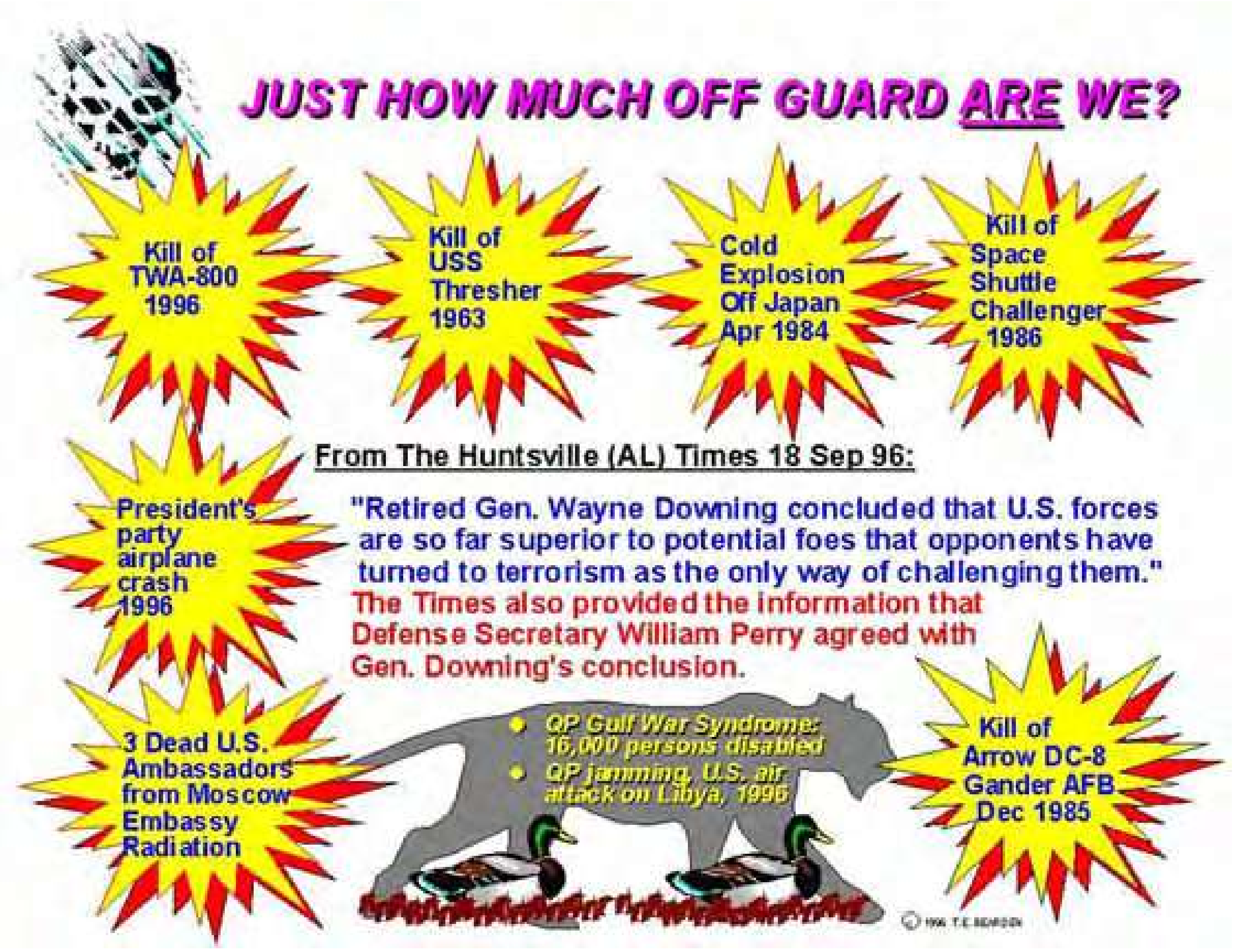

JUST HOW MUCH OFF GUARD ***ARE*** WE?

Kill of
TWA-800
1996

Kill of
USS
Thresher
1963

Cold
Explosion
Off Japan
Apr 1984

Kill of
Space
Shuttle
Challenger
1986

President's
party
airplane
crash
1996

From The Huntsville (AL) Times 18 Sep 96:

"Retired Gen. Wayne Downing concluded that U.S. forces are so far superior to potential foes that opponents have turned to terrorism as the only way of challenging them."

The Times also provided the information that Defense Secretary William Perry agreed with Gen. Downing's conclusion.

3 Dead U.S.
Ambassadors
from Moscow
Embassy
Radiation

- *QP Gulf War Syndrome: 16,000 persons disabled*
- *QP jamming, U.S. air attack on Libya, 1996*

Kill of
Arrow DC-8
Gander AFB
Dec 1985

© 1996 T.E. BEARDEN

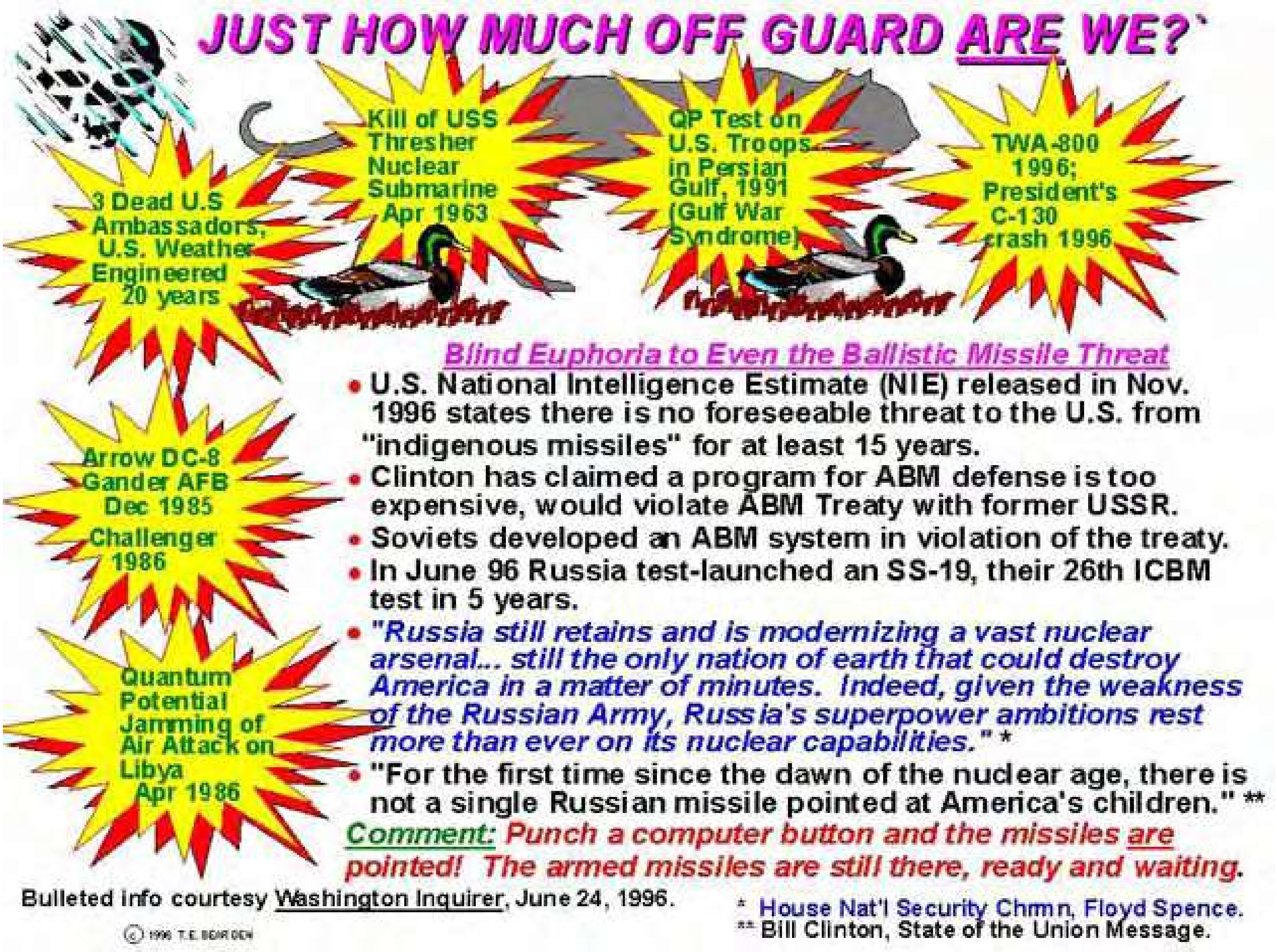

JUST HOW MUCH OFF GUARD ARE WE?

3 Dead U.S Ambassadors, U.S. Weather Engineered 20 years

Kill of USS Thresher Nuclear Submarine Apr 1963

QP Test on U.S. Troops in Persian Gulf, 1991 (Gulf War Syndrome)

TWA-800 1996; President's C-130 crash 1996

Arrow DC-8 Gander AFB Dec 1985

Challenger 1986

Quantum Potential Jamming of Air Attack on Libya Apr 1986

Blind Euphoria to Even the Ballistic Missile Threat

- U.S. National Intelligence Estimate (NIE) released in Nov. 1996 states there is no foreseeable threat to the U.S. from "indigenous missiles" for at least 15 years.
- Clinton has claimed a program for ABM defense is too expensive, would violate ABM Treaty with former USSR.
- Soviets developed an ABM system in violation of the treaty.
- In June 96 Russia test-launched an SS-19, their 26th ICBM test in 5 years.
- *"Russia still retains and is modernizing a vast nuclear arsenal... still the only nation of earth that could destroy America in a matter of minutes. Indeed, given the weakness of the Russian Army, Russia's superpower ambitions rest more than ever on its nuclear capabilities."* *
- "For the first time since the dawn of the nuclear age, there is not a single Russian missile pointed at America's children." **

Comment: Punch a computer button and the missiles are pointed! The armed missiles are still there, ready and waiting.

Bulleted info courtesy Washington Inquirer, June 24, 1996.

* House Nat'l Security Chmn, Floyd Spence.
** Bill Clinton, State of the Union Message.

© 1996 T.E. BEARDEN

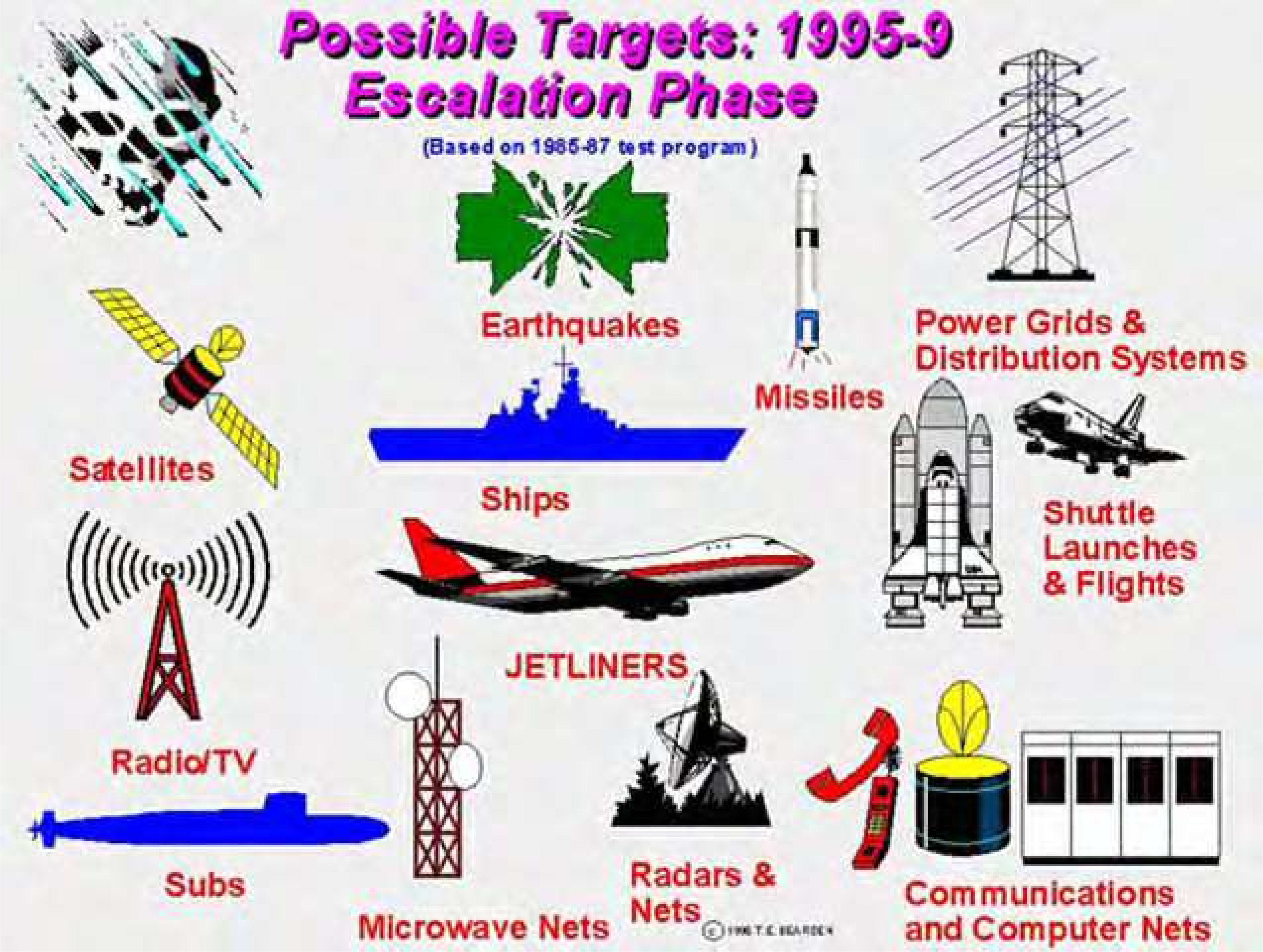

Possible Targets: 1995-9
Escalation Phase

(Based on 1985-87 test program)

Earthquakes

Missiles

Power Grids &
Distribution Systems

Satellites

Ships

Shuttle
Launches
& Flights

JETLINERS

Radio/TV

Subs

Microwave Nets

Radars &
Nets

Communications
and Computer Nets

© [illegible] T.E. BEARDEN

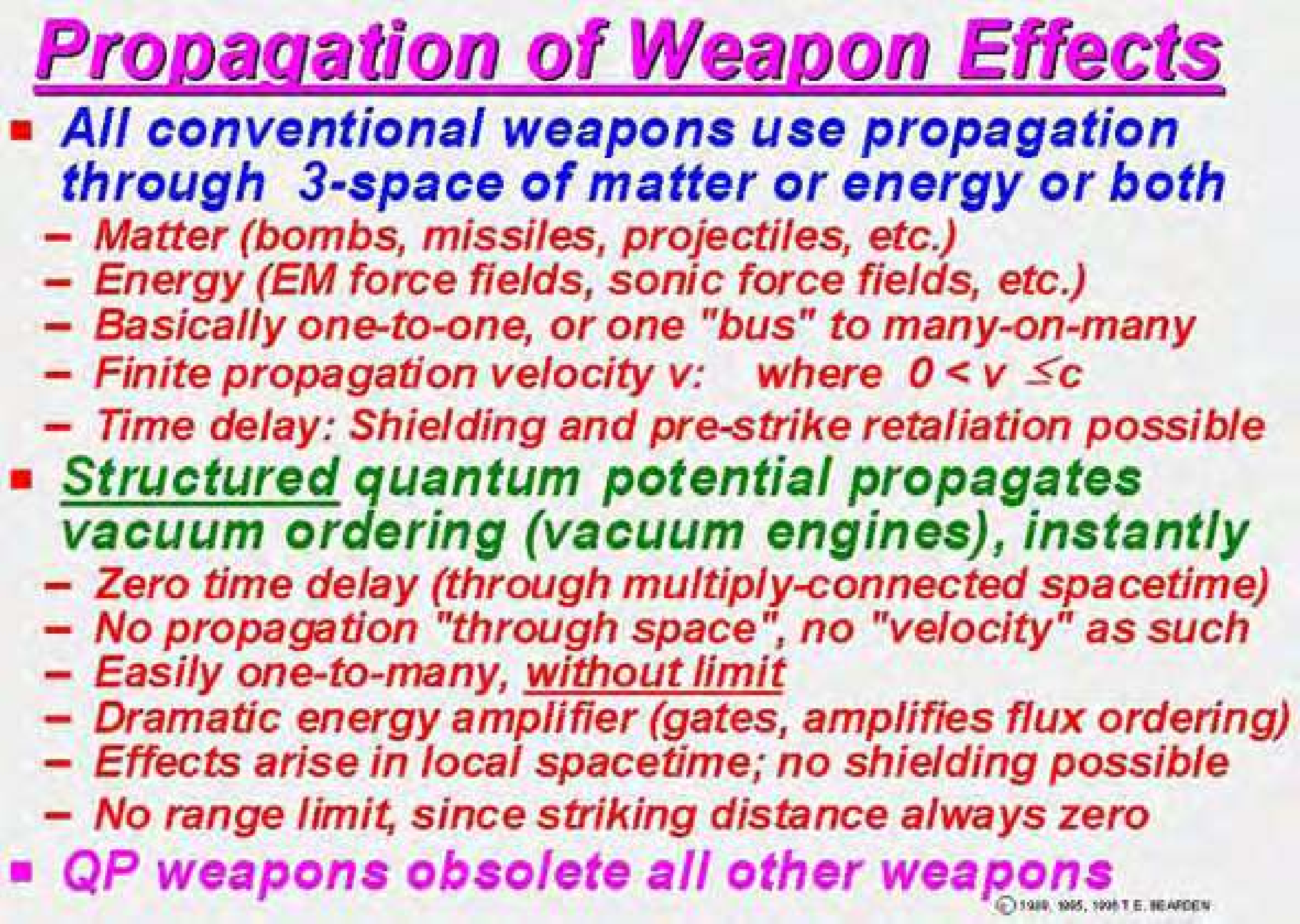

Propagation of Weapon Effects

- All conventional weapons use propagation through 3-space of matter or energy or both
  - Matter (bombs, missiles, projectiles, etc.)
  - Energy (EM force fields, sonic force fields, etc.)
  - Basically one-to-one, or one "bus" to many-on-many
  - Finite propagation velocity v: where $0 < v \leq c$
  - Time delay: Shielding and pre-strike retaliation possible
- Structured quantum potential propagates vacuum ordering (vacuum engines), instantly
  - Zero time delay (through multiply-connected spacetime)
  - No propagation "through space", no "velocity" as such
  - Easily one-to-many, without limit
  - Dramatic energy amplifier (gates, amplifies flux ordering)
  - Effects arise in local spacetime; no shielding possible
  - No range limit, since striking distance always zero
- QP weapons obsolete all other weapons

© 1988, 1995, 1998 T.E. BEARDEN

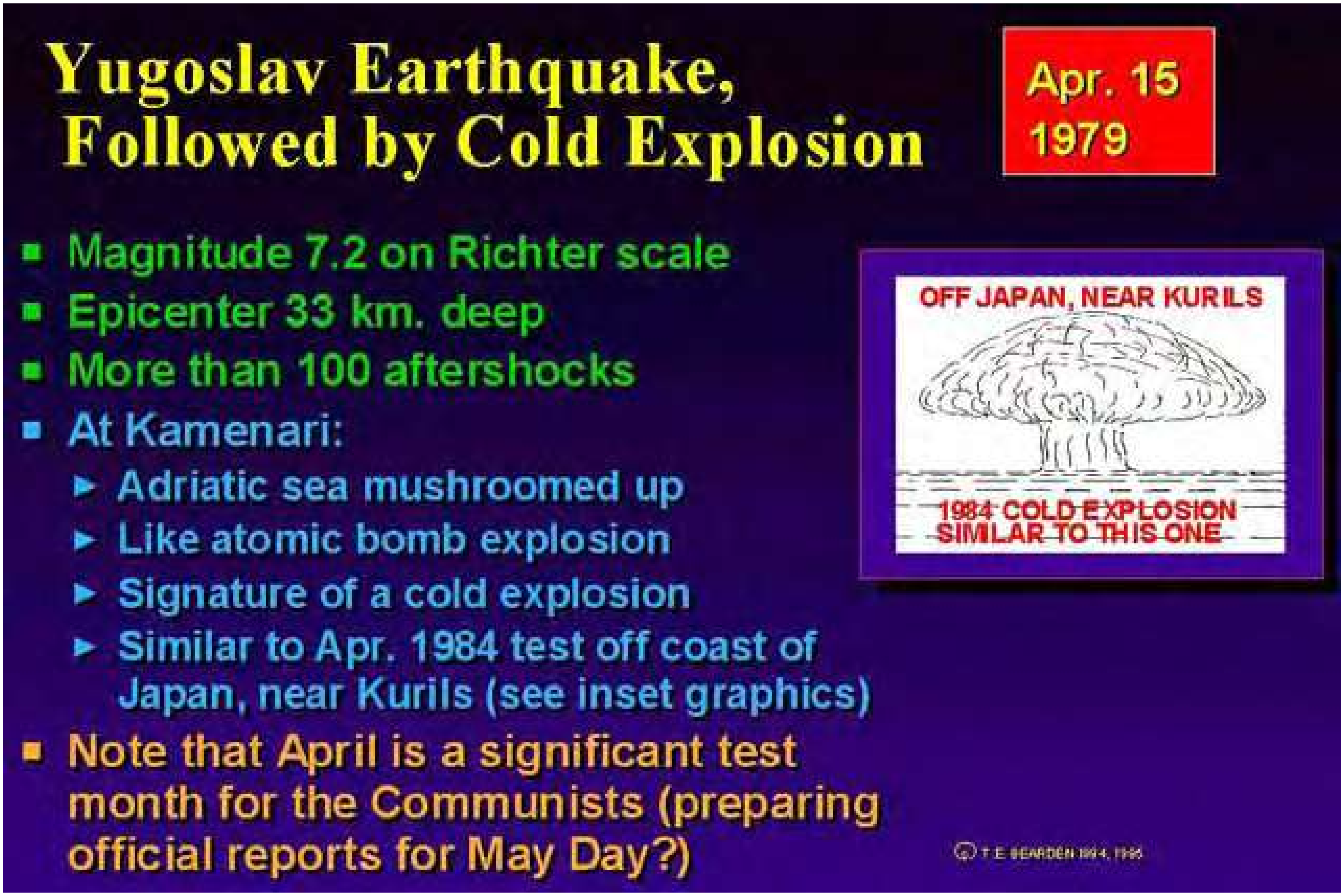

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- Note that April is a significant test month for the Communists (preparing official reports for May Day?)

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

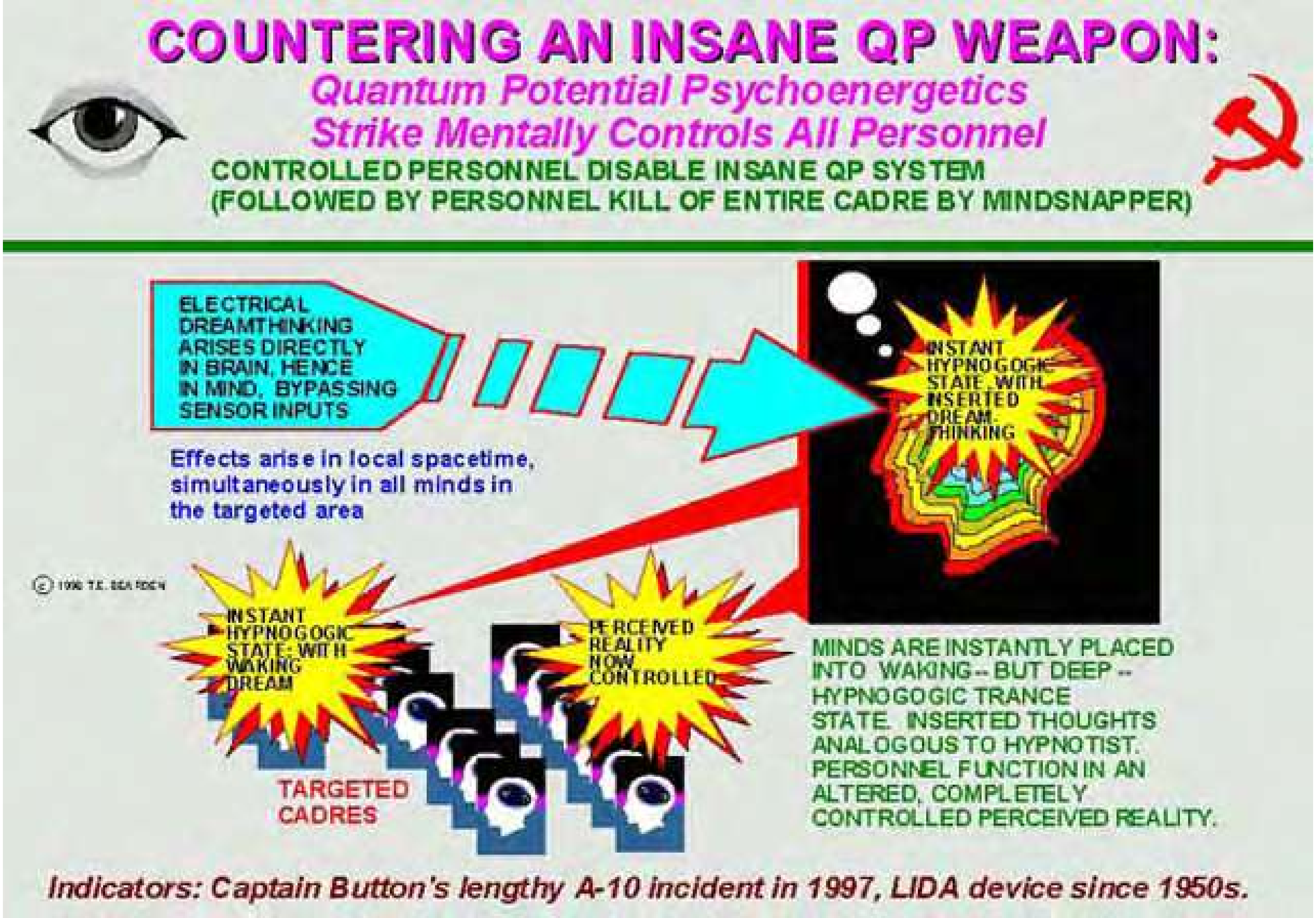

COUNTERING AN INSANE QP WEAPON:

Quantum Potential Psychoenergetics Strike Mentally Controls All Personnel

CONTROLLED PERSONNEL DISABLE INSANE QP SYSTEM
(FOLLOWED BY PERSONNEL KILL OF ENTIRE CADRE BY MINDSNAPPER)

ELECTRICAL DREAMTHINKING ARISES DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED DREAM THINKING
INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM
PERCEIVED REALITY NOW CONTROLLED
TARGETED CADRES

MINDS ARE INSTANTLY PLACED INTO WAKING -- BUT DEEP -- HYPNOGOGIC TRANCE STATE. INSERTED THOUGHTS ANALOGOUS TO HYPNOTIST. PERSONNEL FUNCTION IN AN ALTERED, COMPLETELY CONTROLLED PERCEIVED REALITY.

© 1998 T.E. BEARDEN

Indicators: Captain Button's lengthy A-10 incident in 1997, LIDA device since 1950s.

Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases

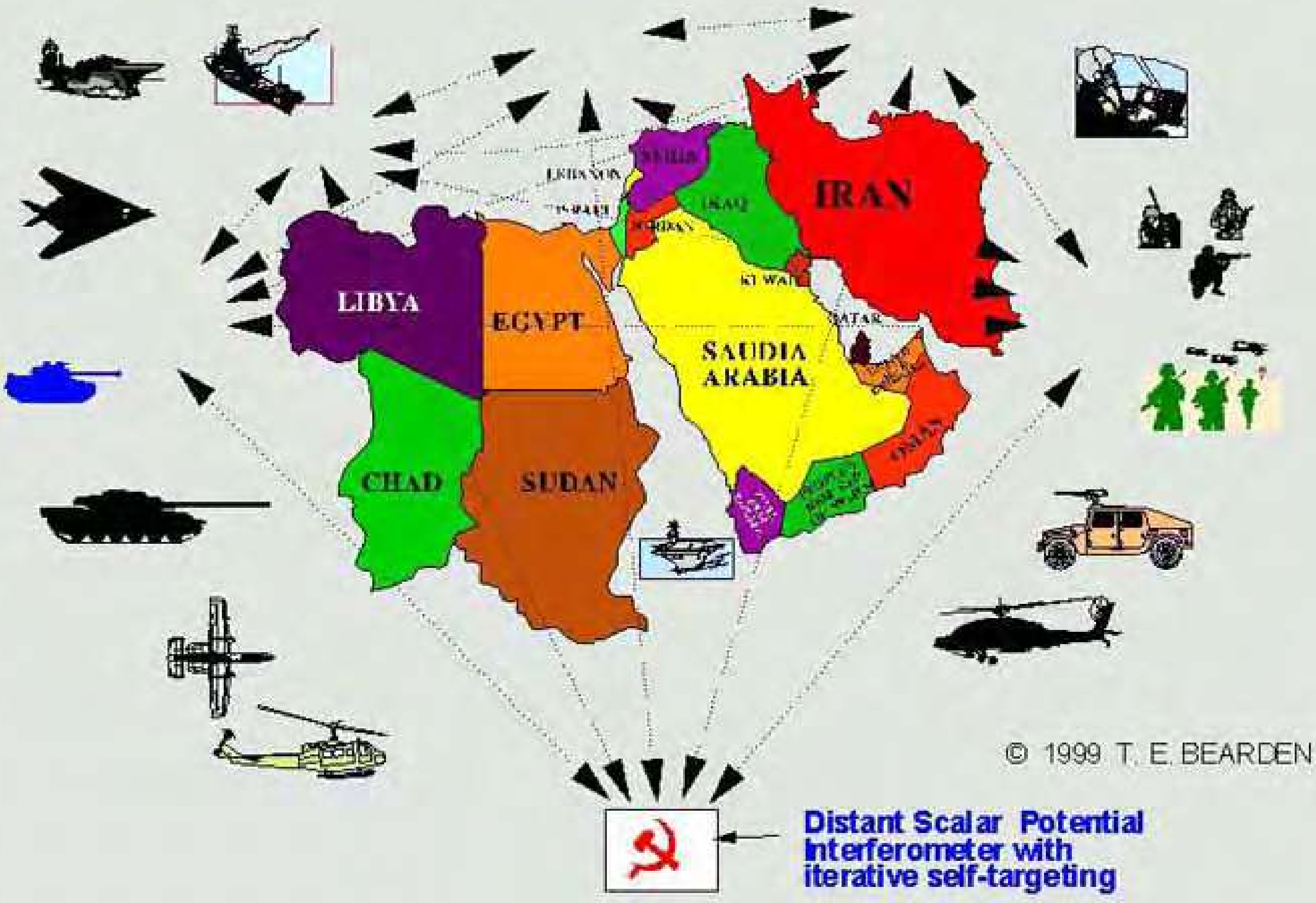
LEBANON
ISRAEL
SYRIA
JORDAN
IRAQ
IRAN
KUWAIT
QATAR
LIBYA
EGYPT
SAUDIA ARABIA
OMAN
CHAD
SUDAN

© 1999 T. E. BEARDEN

Distant Scalar Potential Interferometer with iterative self-targeting

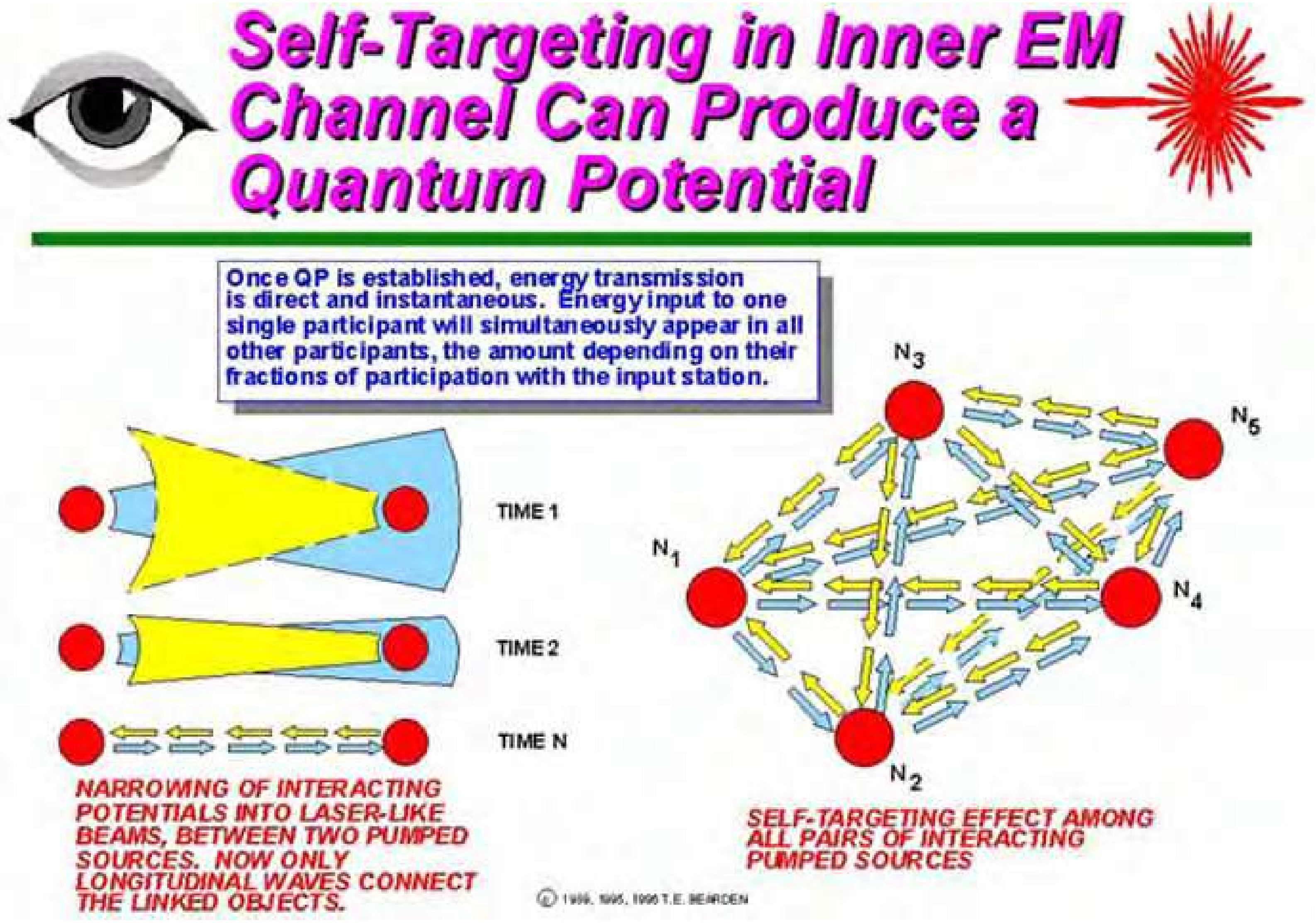

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1989, 1995, 1996 T.E. BEARDEN

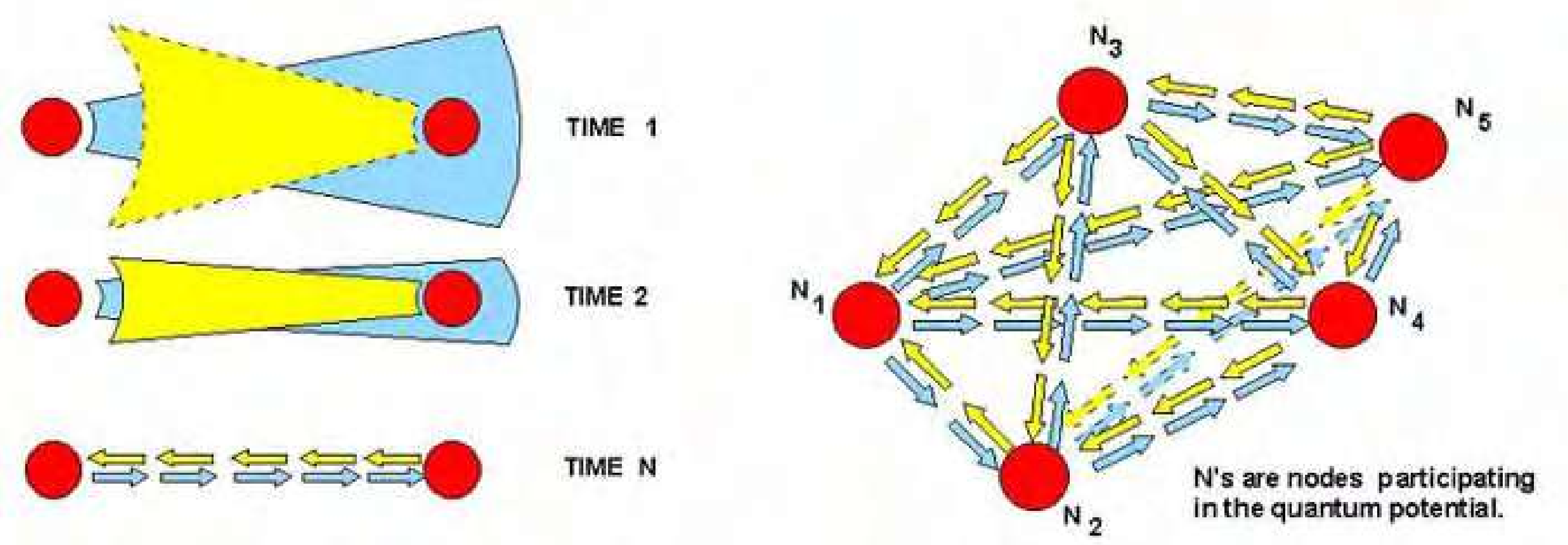

a. Iterative retroreflection between any pair of nodes narrows the interconnecting potentials to a laser-like, bidirectional beam. Iterative phase conjugate reflection provides self-targeting.

b. The quantum potential exists as narrow, laser-like beams between participating nodes, even when the nodes are widely separated. The beam space is multiply-connected; transmission is instantaneous.

© 1989, 1995 T.E. BEARDEN

Consider the hidden Stoney/Whittaker bidirectional EM waves that comprise the potential from each source node. If these nodes have sufficient intersymmetry of dynamic material form, then self-targeting occurs between the two nodes of any nodal pair. By iterative phase conjugate reflection, the scattering is narrowed into interconnecting wavepairs, in laser-like beams, between each two nodes. Thus the interconnecting potentials narrow into laser-like, bidirectional beams. The Coulomb gauge may apply, and a quantum potential may be formed.

Figure 30. Mechanism for generating a quantum potential, by iterative self-targeting induced into a system of multiple scattering nodes. Once a QP is established, spatial separation has little or no effect.

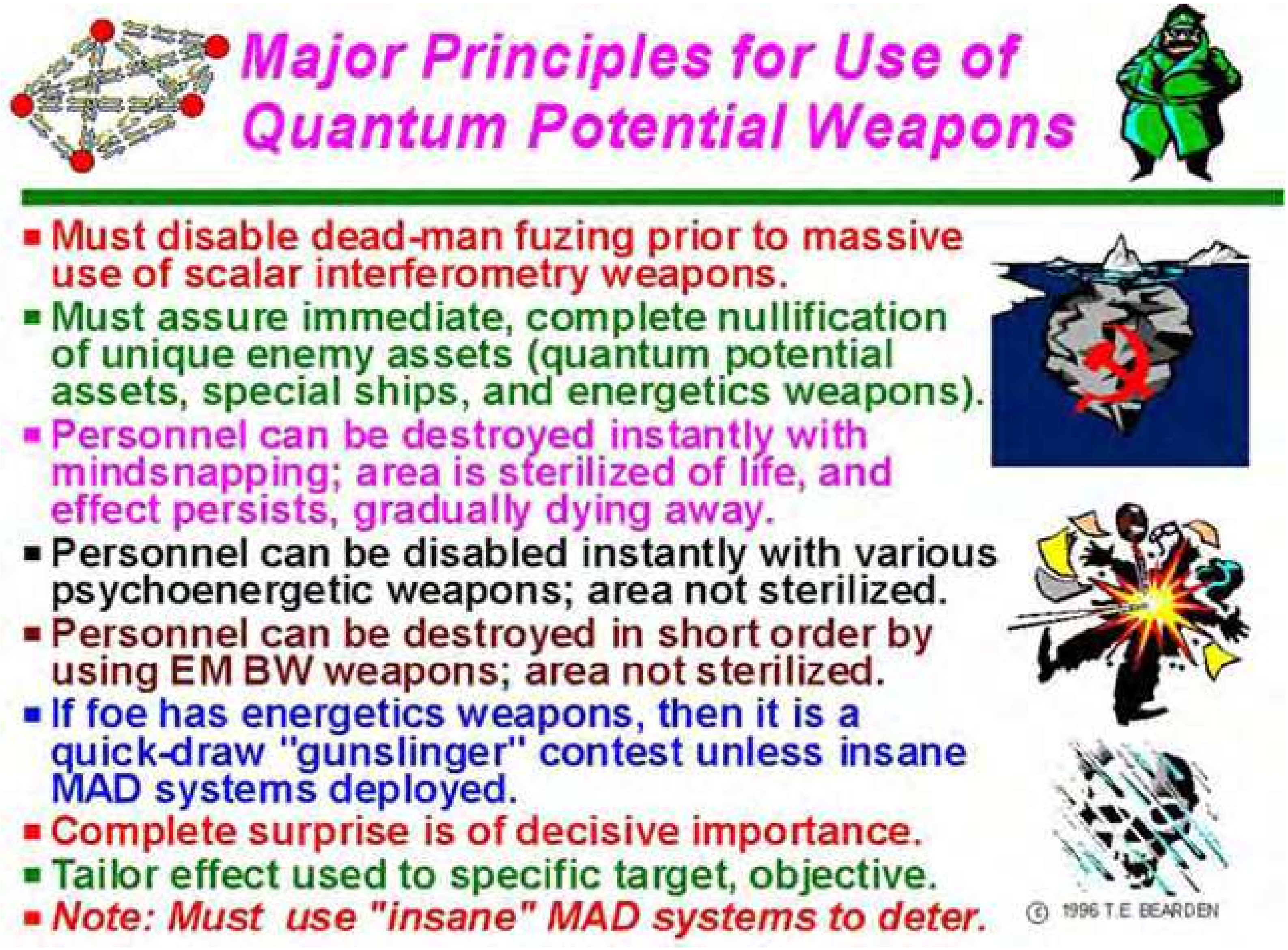

Major Principles for Use of Quantum Potential Weapons

- Must disable dead-man fuzing prior to massive use of scalar interferometry weapons.
- Must assure immediate, complete nullification of unique enemy assets (quantum potential assets, special ships, and energetics weapons).
- Personnel can be destroyed instantly with mindsnapping; area is sterilized of life, and effect persists, gradually dying away.
- Personnel can be disabled instantly with various psychoenergetic weapons; area not sterilized.
- Personnel can be destroyed in short order by using EM BW weapons; area not sterilized.
- If foe has energetics weapons, then it is a quick-draw "gunslinger" contest unless insane MAD systems deployed.
- Complete surprise is of decisive importance.
- Tailor effect used to specific target, objective.
- *Note: Must use "insane" MAD systems to deter.*

© 1996 T.E. BEARDEN

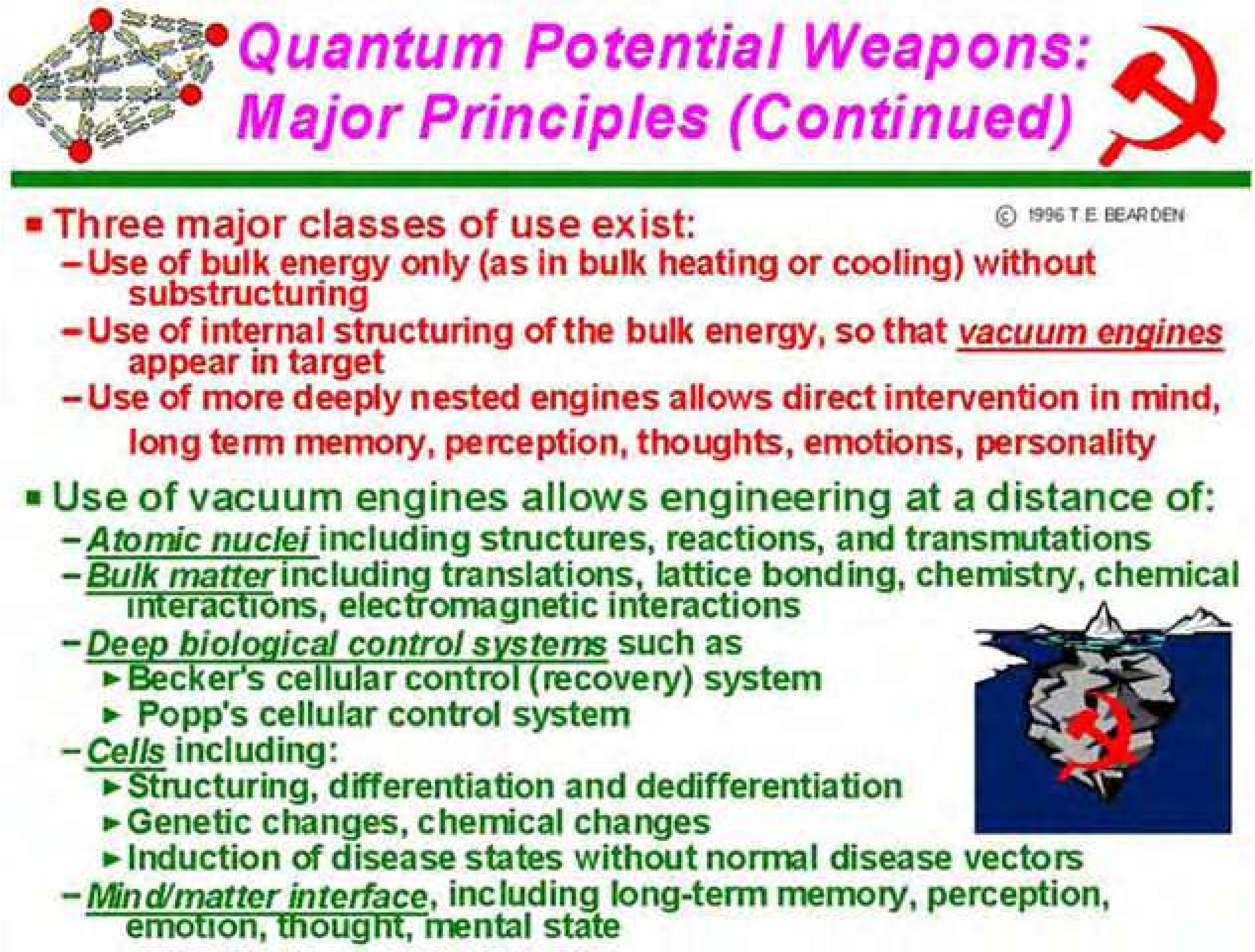

Quantum Potential Weapons: Major Principles (Continued)

© 1996 T. E. BEARDEN

- Three major classes of use exist:
  - Use of bulk energy only (as in bulk heating or cooling) without substructuring
  - Use of internal structuring of the bulk energy, so that *vacuum engines* appear in target
  - Use of more deeply nested engines allows direct intervention in mind, long term memory, perception, thoughts, emotions, personality
- Use of vacuum engines allows engineering at a distance of:
  - *Atomic nuclei* including structures, reactions, and transmutations
  - *Bulk matter* including translations, lattice bonding, chemistry, chemical interactions, electromagnetic interactions
  - *Deep biological control systems* such as
    - Becker's cellular control (recovery) system
    - Popp's cellular control system
  - *Cells* including:
    - Structuring, differentiation and dedifferentiation
    - Genetic changes, chemical changes
    - Induction of disease states without normal disease vectors
  - *Mind/matter interface*, including long-term memory, perception, emotion, thought, mental state

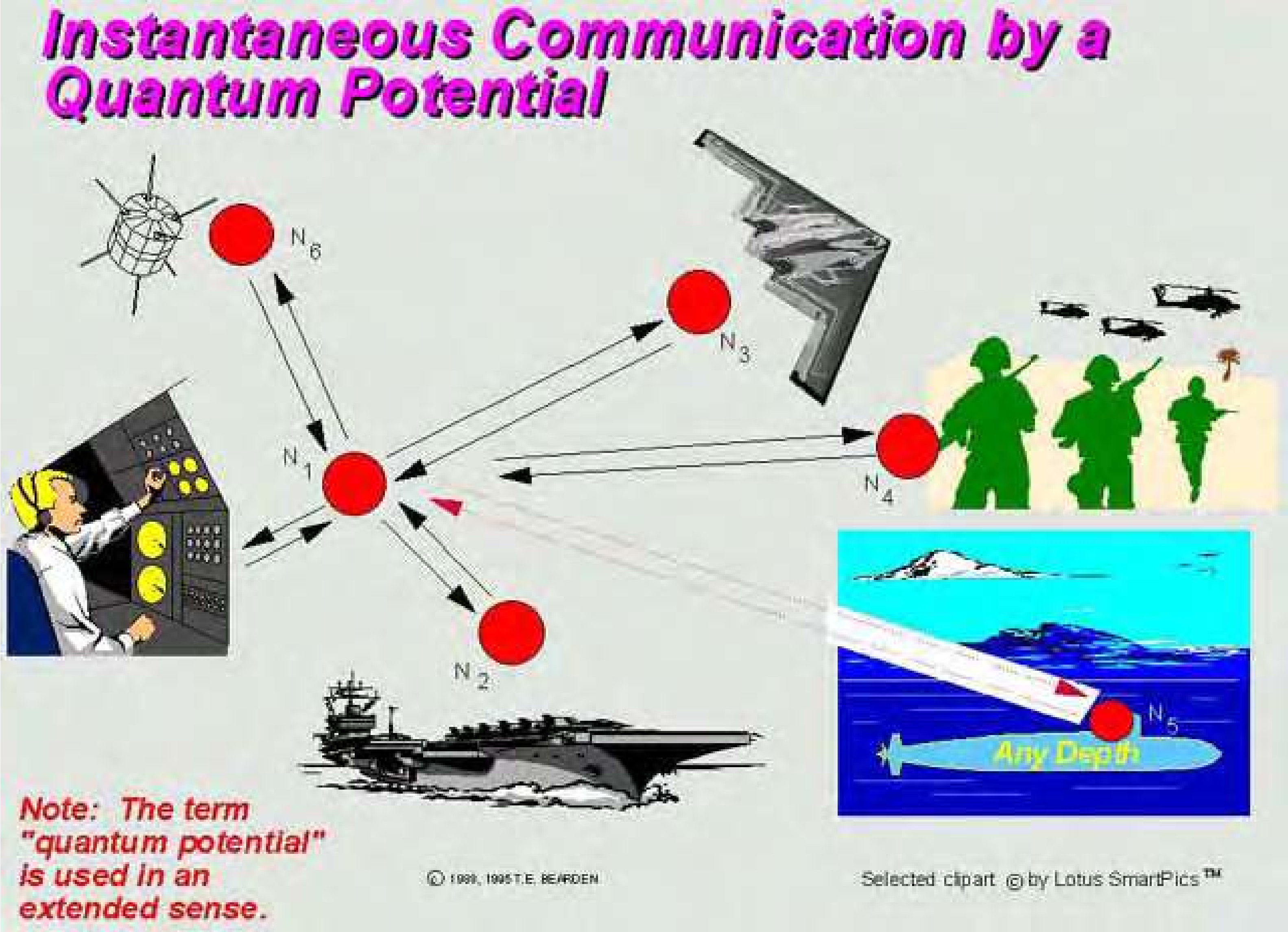

Instantaneous Communication by a Quantum Potential

Note: The term "quantum potential" is used in an extended sense.

© 1989, 1995 T.E. BEARDEN

Selected clipart © by Lotus SmartPics™

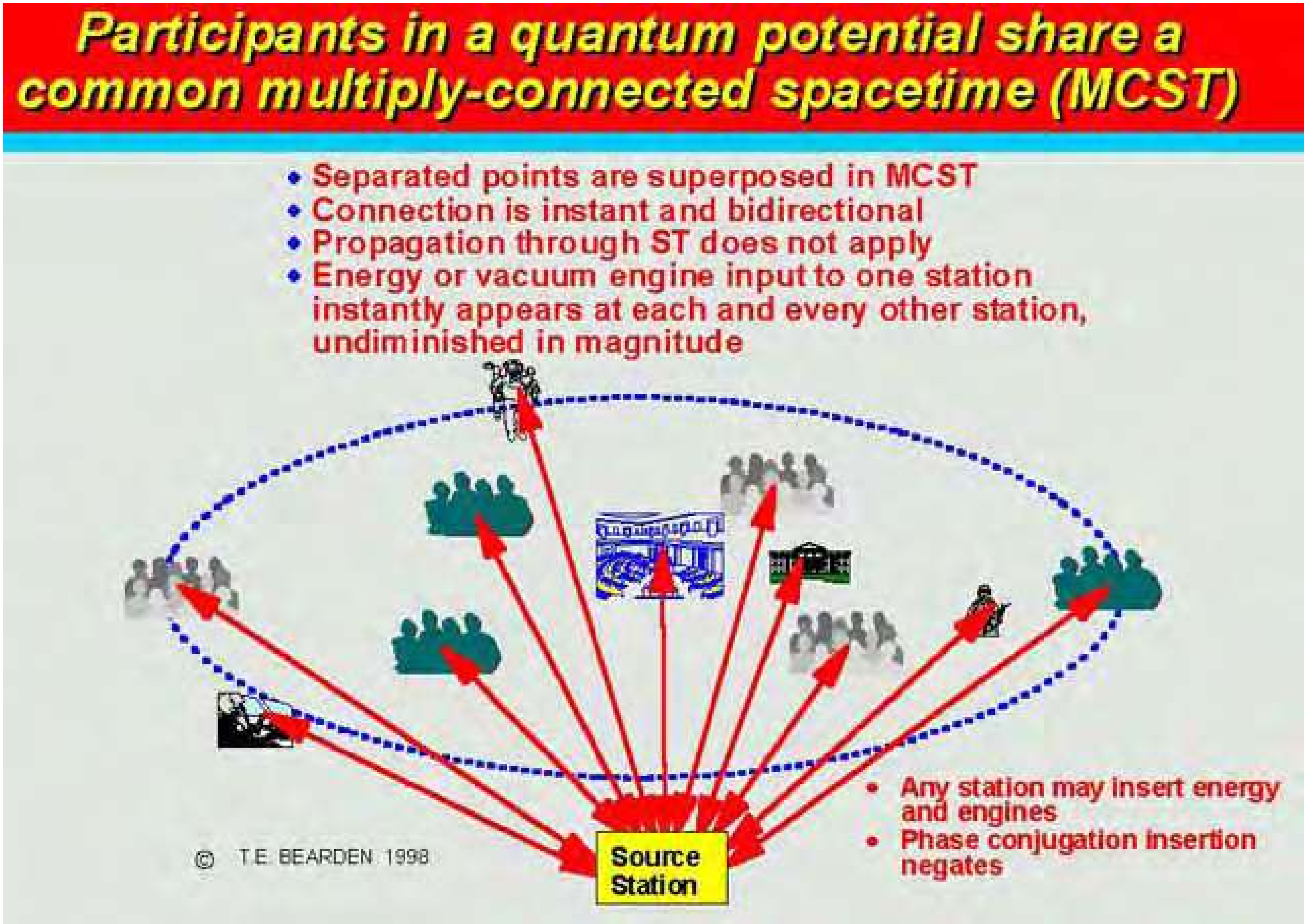

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude

- Any station may insert energy and engines
- Phase conjugation insertion negates

© T.E. BEARDEN 1998

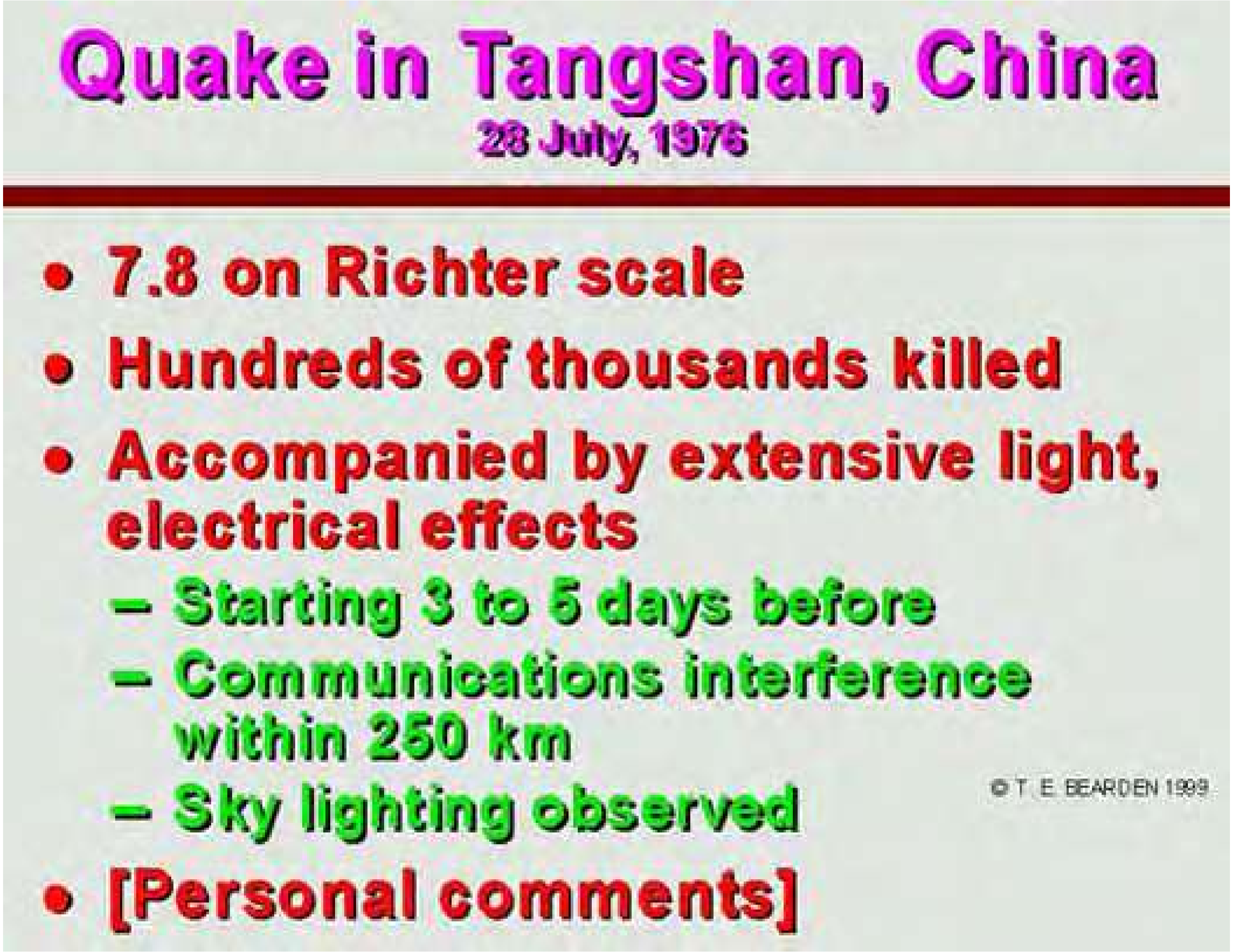

Quake in Tangshan, China

28 July, 1976

- 7.8 on Richter scale
- Hundreds of thousands killed
- Accompanied by extensive light, electrical effects
  - Starting 3 to 5 days before
  - Communications interference within 250 km
  - Sky lighting observed
- [Personal comments]

© T. E. BEARDEN 1999

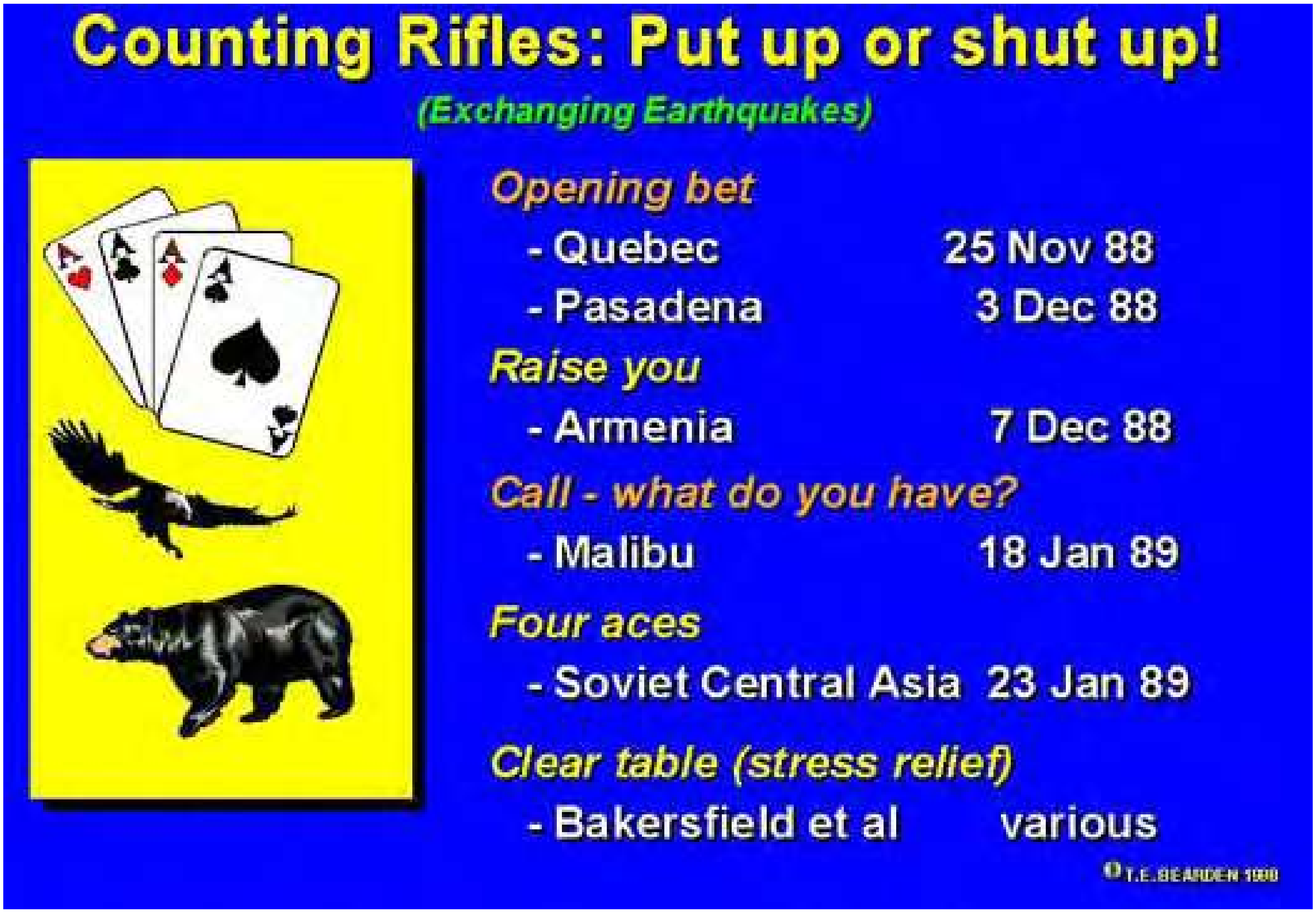

Counting Rifles: Put up or shut up!

(Exchanging Earthquakes)

Opening bet

- Quebec 25 Nov 88
- Pasadena 3 Dec 88

Raise you

- Armenia 7 Dec 88

Call - what do you have?

- Malibu 18 Jan 89

Four aces

- Soviet Central Asia 23 Jan 89

Clear table (stress relief)

- Bakersfield et al various

© T.E. BEARDEN 1990

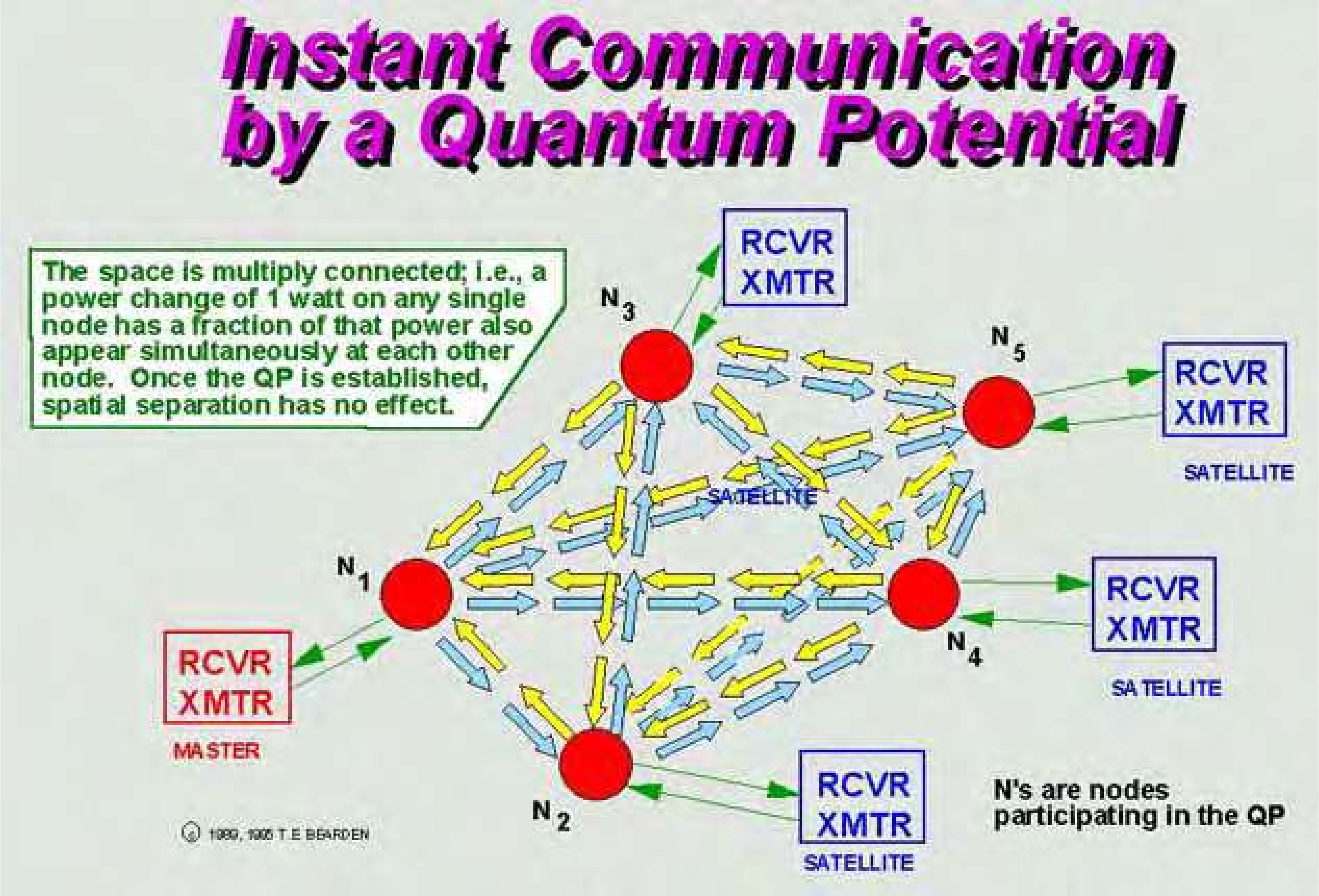

Instant Communication by a Quantum Potential

The space is multiply connected; i.e., a power change of 1 watt on any single node has a fraction of that power also appear simultaneously at each other node. Once the QP is established, spatial separation has no effect.

N's are nodes participating in the QP

RCVR
XMTR
MASTER
N 1
N 2
N 3
N 4
N 5
RCVR
XMTR
SATELLITE
RCVR
XMTR
SATELLITE
RCVR
XMTR
SATELLITE
RCVR
XMTR
SATELLITE

© 1989, 1995 T.E BEARDEN

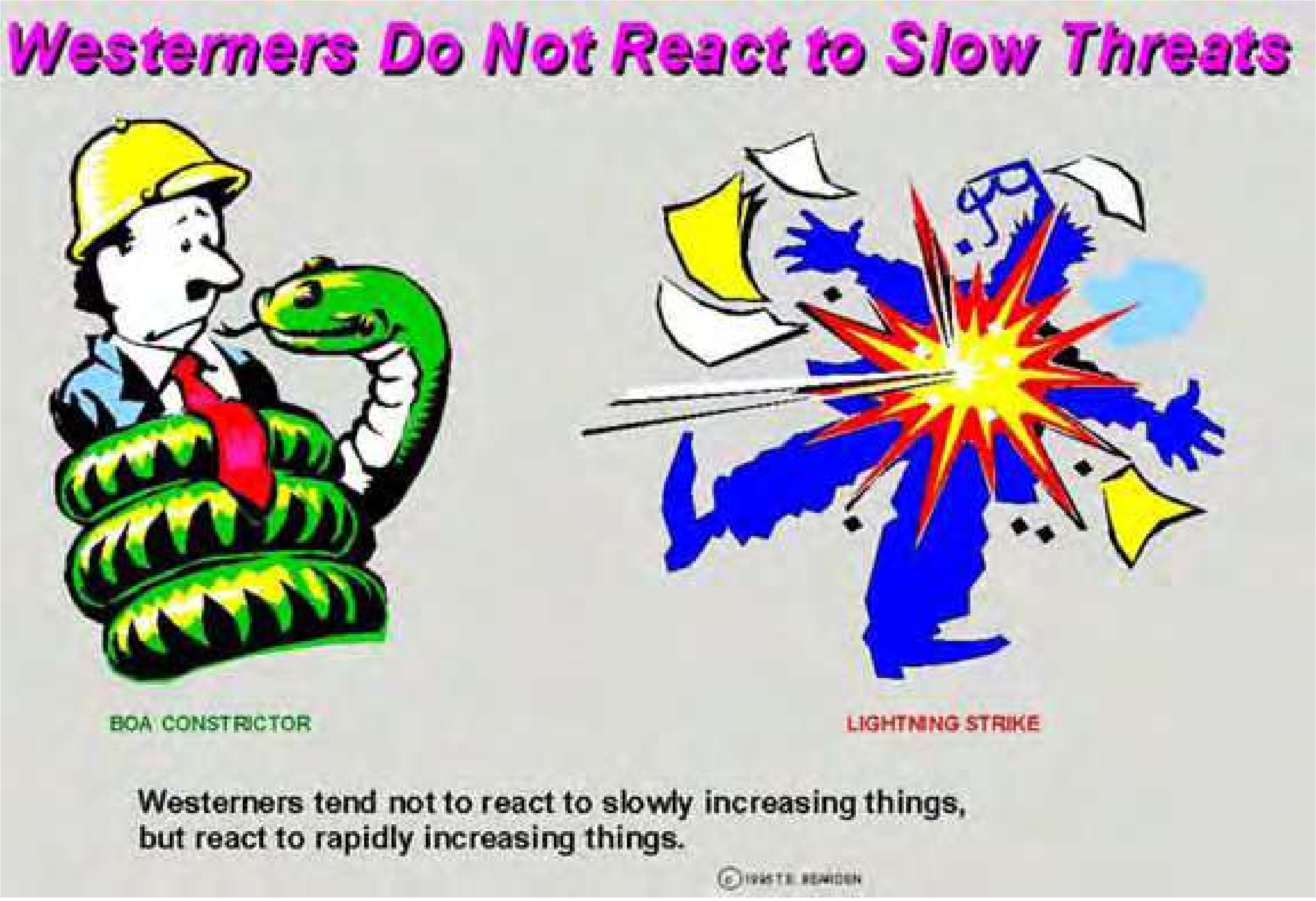

Westerners Do Not React to Slow Threats

BOA CONSTRICTOR

LIGHTNING STRIKE

Westerners tend not to react to slowly increasing things,
but react to rapidly increasing things.

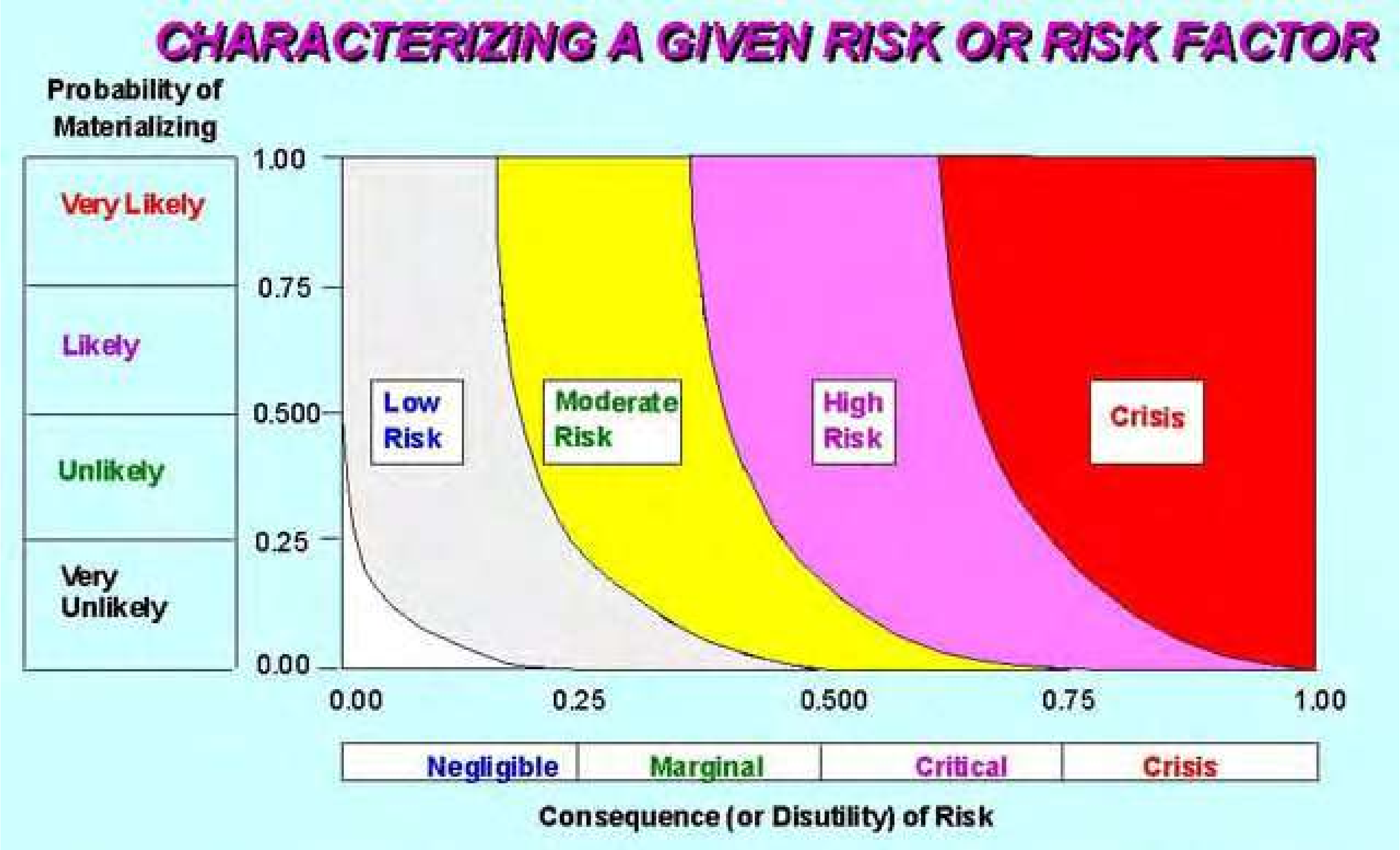
CHARACTERIZING A GIVEN RISK OR RISK FACTOR
Probability of Materializing
Very Likely
Likely
Unlikely
Very Unlikely
1.00
0.75
0.500
0.25
0.00
Low Risk
Moderate Risk
High Risk
Crisis
0.00
0.25
0.500
0.75
1.00
Negligible
Marginal
Critical
Crisis
Consequence (or Disutility) of Risk

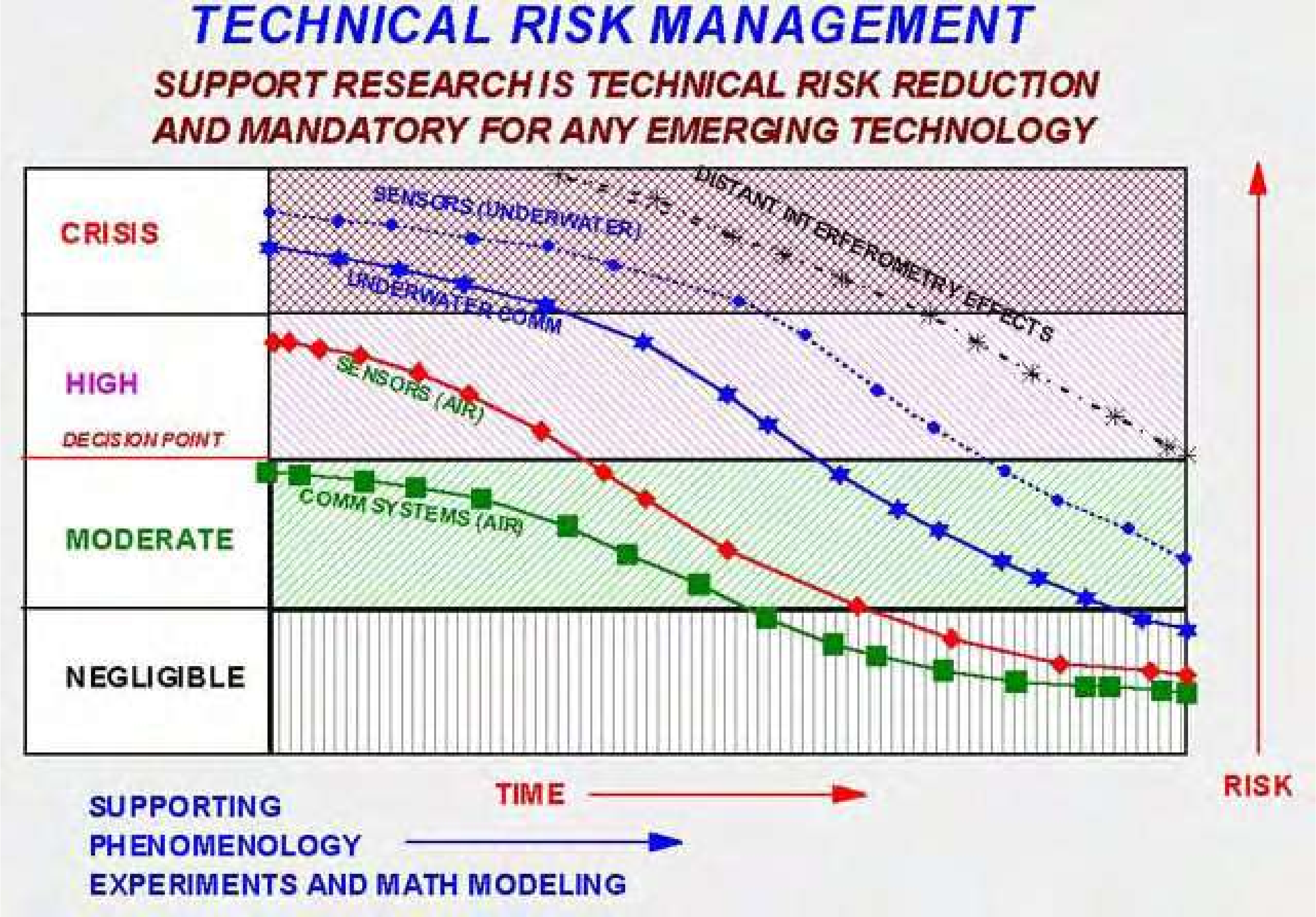

TECHNICAL RISK MANAGEMENT

SUPPORT RESEARCH IS TECHNICAL RISK REDUCTION AND MANDATORY FOR ANY EMERGING TECHNOLOGY

SUPPORTING
PHENOMENOLOGY
EXPERIMENTS AND MATH MODELING

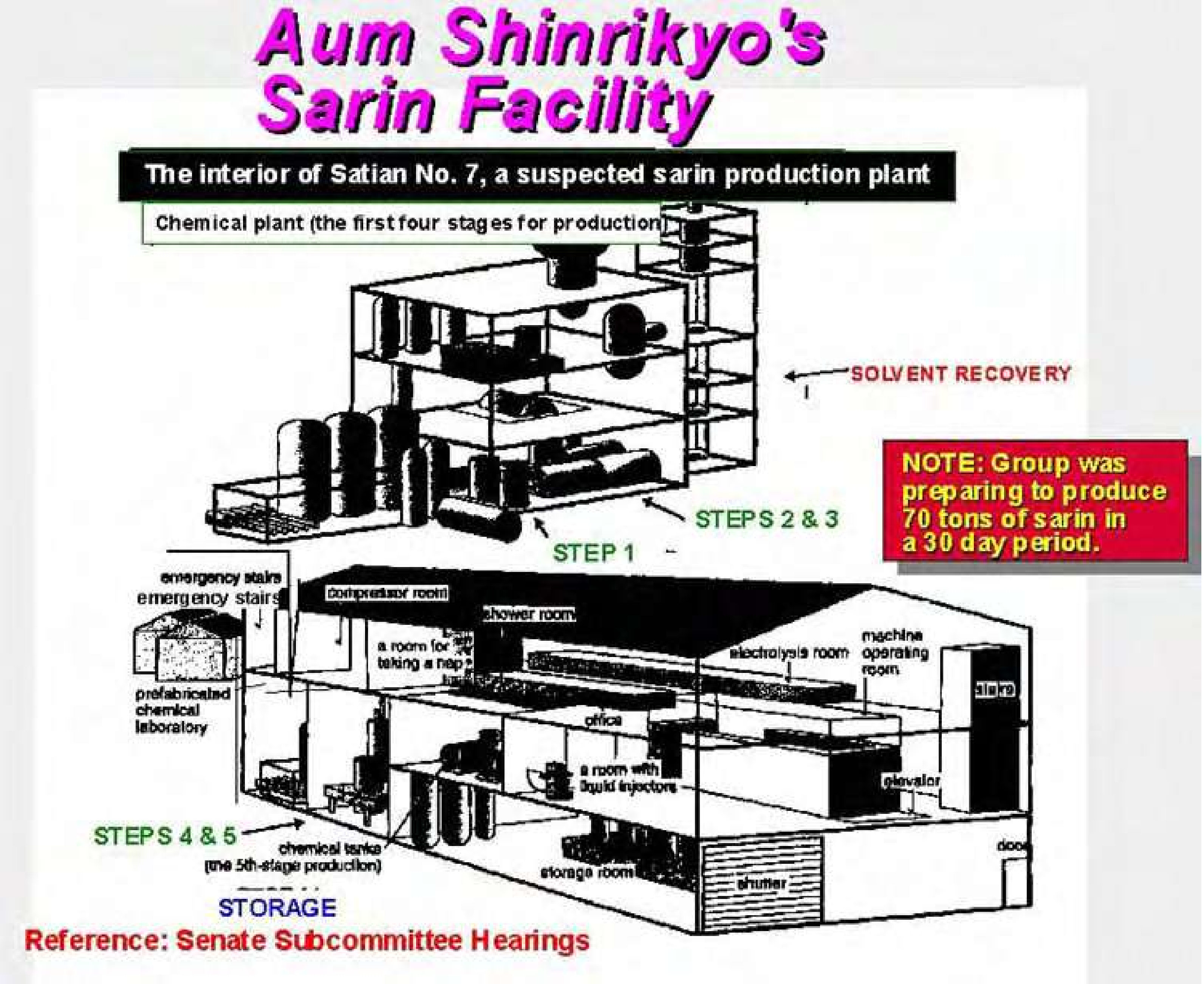

Aum Shinrikyo's Sarin Facility

The interior of Satian No. 7, a suspected sarin production plant

Chemical plant (the first four stages for production)

SOLVENT RECOVERY

STEPS 2 & 3

STEP 1

STEPS 4 & 5

STORAGE

NOTE: Group was preparing to produce 70 tons of sarin in a 30 day period.

Reference: Senate Subcommittee Hearings

emergency stairs
emergency stairs
compressor room
shower room
a room for taking a nap
electrolysis room
machine operating room
stairs
prefabricated chemical laboratory
office
a room with liquid injectors
elevator
door
chemical tanks (the 5th-stage production)
storage room
shutter

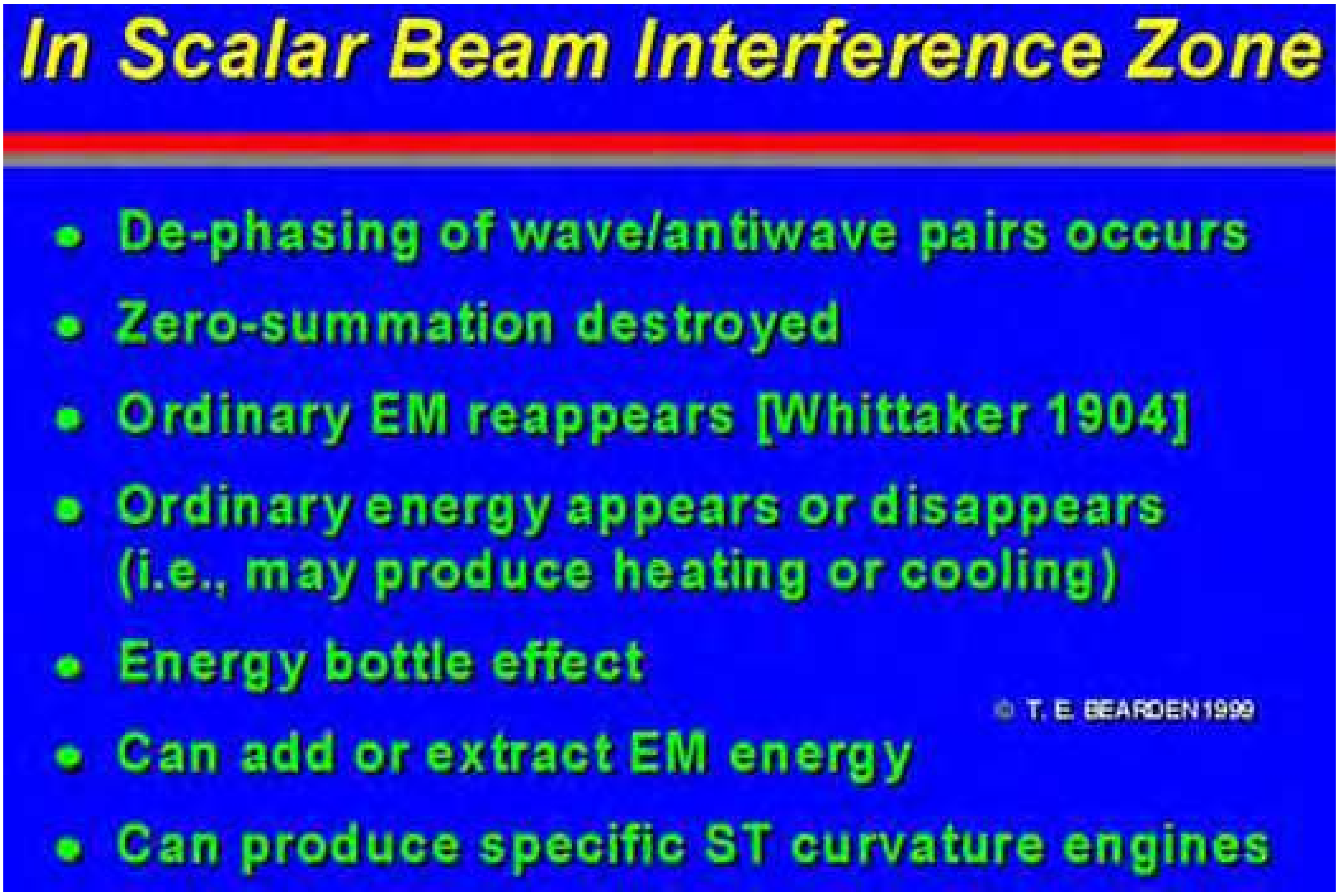

In Scalar Beam Interference Zone

- De-phasing of wave/antiwave pairs occurs
- Zero-summation destroyed
- Ordinary EM reappears [Whittaker 1904]
- Ordinary energy appears or disappears (i.e., may produce heating or cooling)
- Energy bottle effect
- Can add or extract EM energy
- Can produce specific ST curvature engines

© T. E. BEARDEN 1999

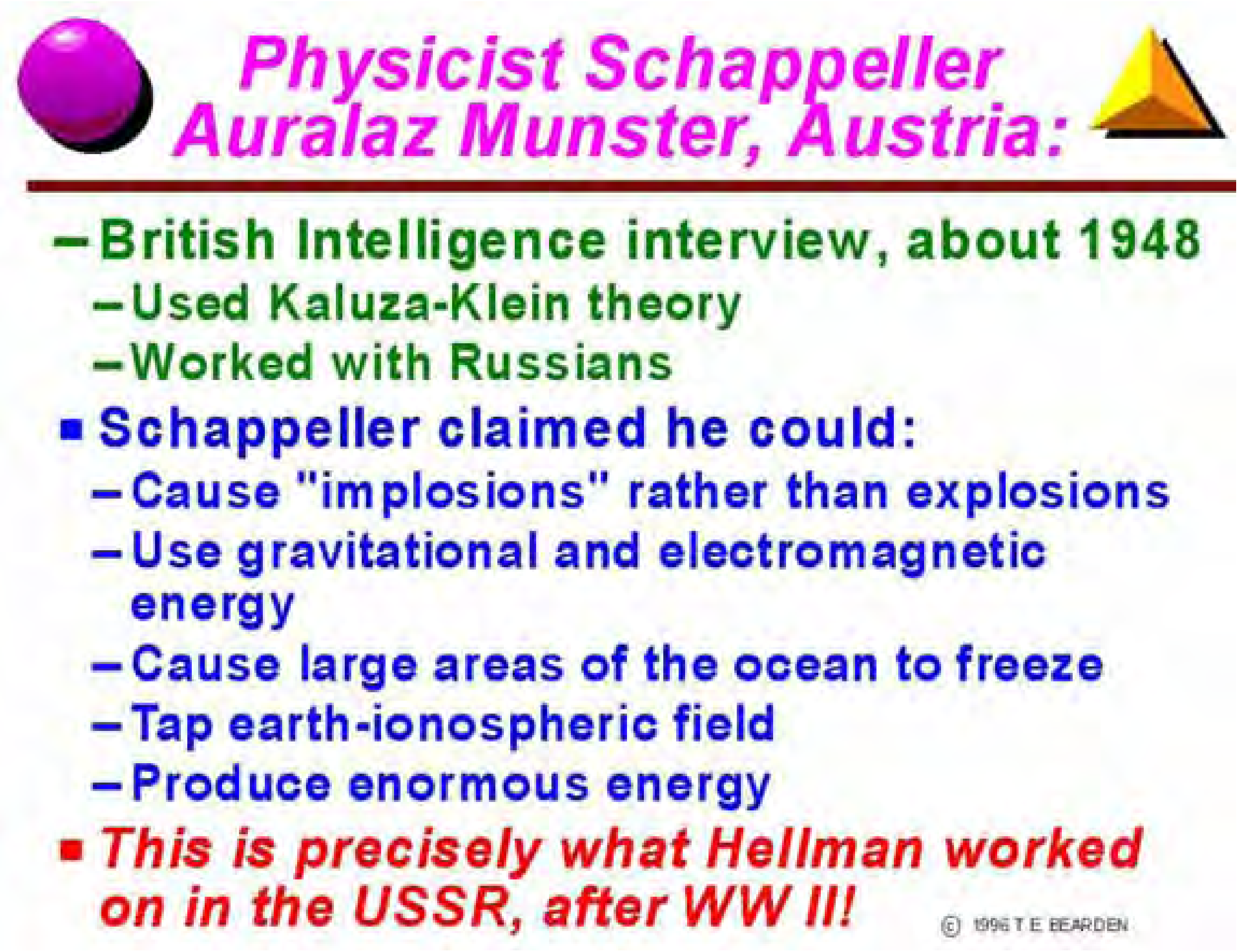

Physicist Schappeller Auralaz Munster, Austria:

- British Intelligence interview, about 1948
  - Used Kaluza-Klein theory
  - Worked with Russians
- Schappeller claimed he could:
  - Cause "implosions" rather than explosions
  - Use gravitational and electromagnetic energy
  - Cause large areas of the ocean to freeze
  - Tap earth-ionospheric field
  - Produce enormous energy
- *This is precisely what Hellman worked on in the USSR, after WW II!*

© 1996 T E BEARDEN

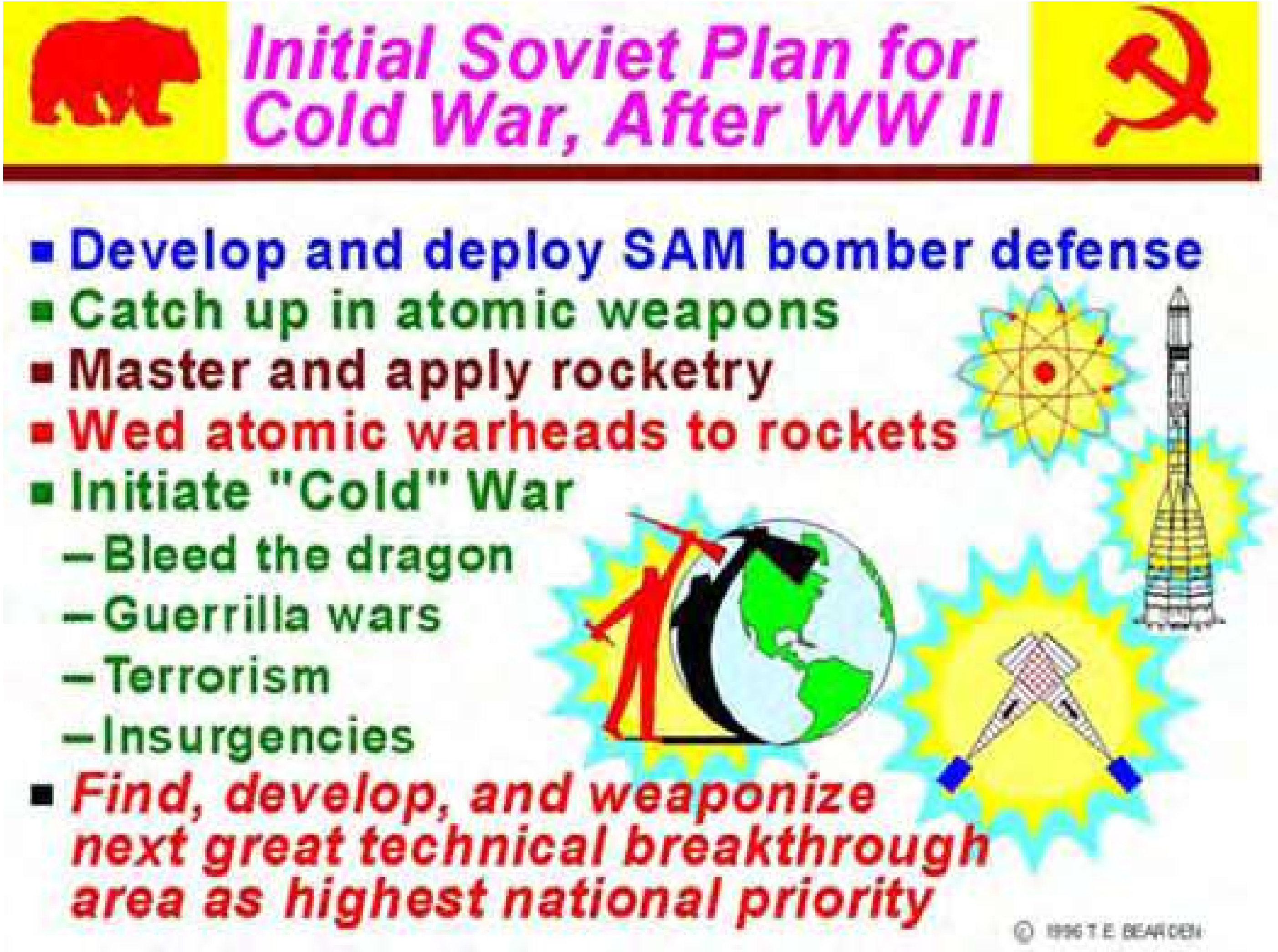

Initial Soviet Plan for Cold War, After WW II

- Develop and deploy SAM bomber defense
- Catch up in atomic weapons
- Master and apply rocketry
- Wed atomic warheads to rockets
- Initiate "Cold" War
  - Bleed the dragon
  - Guerrilla wars
  - Terrorism
  - Insurgencies
- *Find, develop, and weaponize next great technical breakthrough area as highest national priority*

© 1996 T. E. BEARDEN

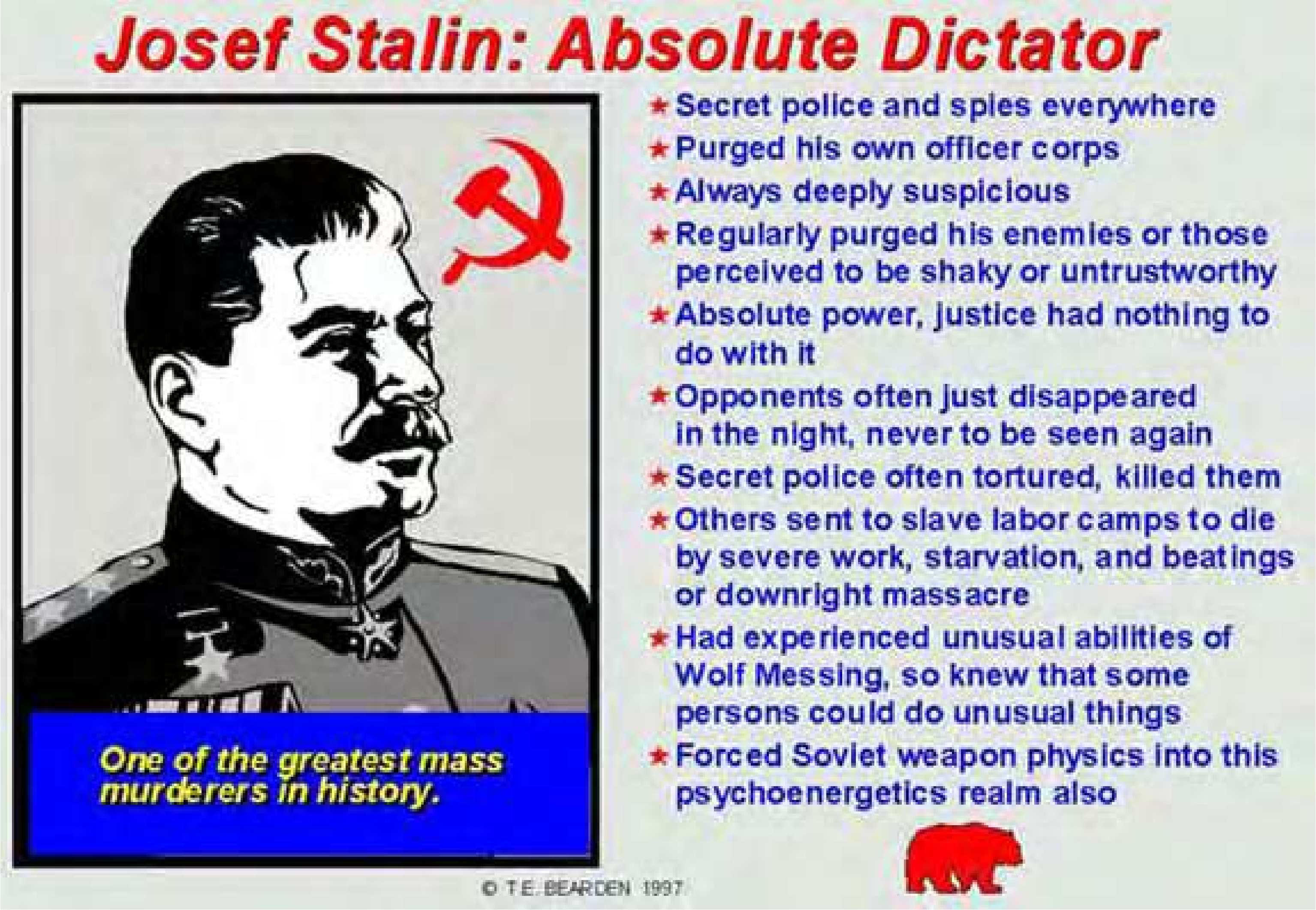

Josef Stalin: Absolute Dictator

- Secret police and spies everywhere
- Purged his own officer corps
- Always deeply suspicious
- Regularly purged his enemies or those perceived to be shaky or untrustworthy
- Absolute power, justice had nothing to do with it
- Opponents often just disappeared in the night, never to be seen again
- Secret police often tortured, killed them
- Others sent to slave labor camps to die by severe work, starvation, and beatings or downright massacre
- Had experienced unusual abilities of Wolf Messing, so knew that some persons could do unusual things
- Forced Soviet weapon physics into this psychoenergetics realm also

One of the greatest mass murderers in history.

© T.E. BEARDEN 1997

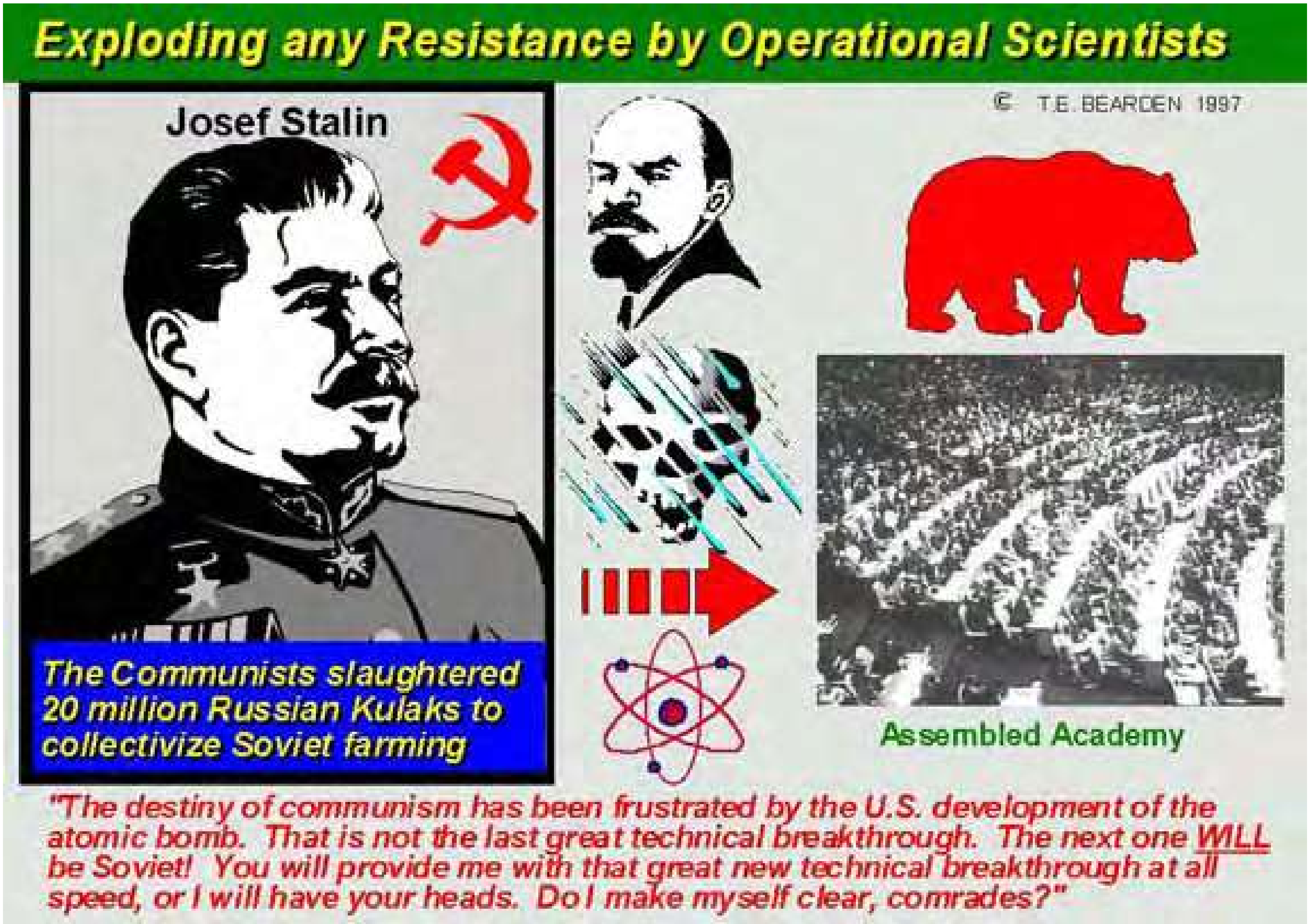

Exploding any Resistance by Operational Scientists

© T.E. BEARDEN 1997

Josef Stalin

The Communists slaughtered 20 million Russian Kulaks to collectivize Soviet farming

Assembled Academy

"The destiny of communism has been frustrated by the U.S. development of the atomic bomb. That is not the last great technical breakthrough. The next one WILL be Soviet! You will provide me with that great new technical breakthrough at all speed, or I will have your heads. Do I make myself clear, comrades?"

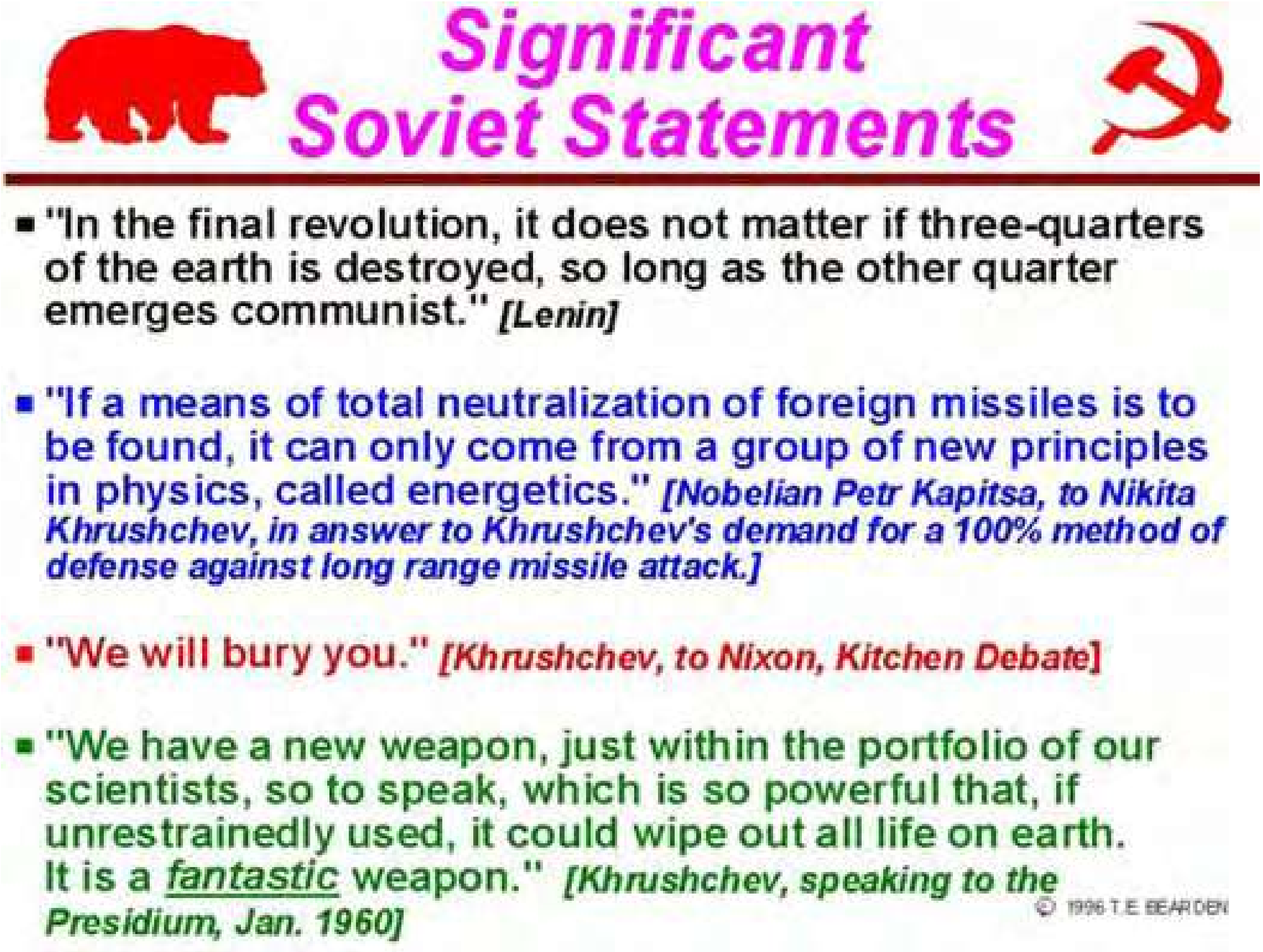

Significant Soviet Statements

- "In the final revolution, it does not matter if three-quarters of the earth is destroyed, so long as the other quarter emerges communist." *[Lenin]*
- "If a means of total neutralization of foreign missiles is to be found, it can only come from a group of new principles in physics, called energetics." *[Nobelian Petr Kapitsa, to Nikita Khrushchev, in answer to Khrushchev's demand for a 100% method of defense against long range missile attack.]*
- "We will bury you." *[Khrushchev, to Nixon, Kitchen Debate]*
- "We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that, if unrestrainedly used, it could wipe out all life on earth. It is a _fantastic_ weapon." *[Khrushchev, speaking to the Presidium, Jan. 1960]*

© 1996 T.E. BEARDEN

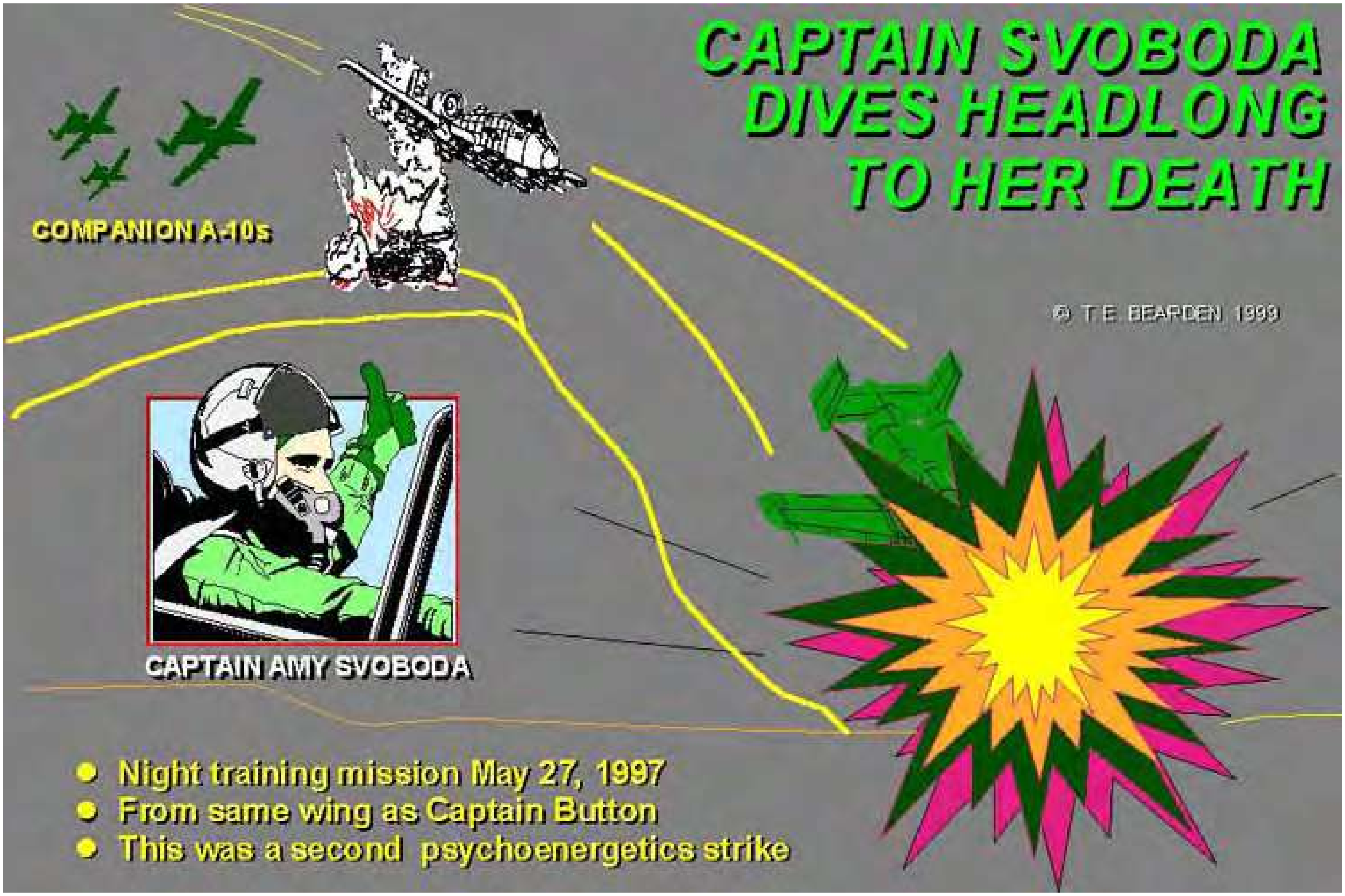

CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH

COMPANION A-10s
CAPTAIN AMY SVOBODA

© T.E. BEARDEN 1999

- Night training mission May 27, 1997
- From same wing as Captain Button
- This was a second psychoenergetics strike

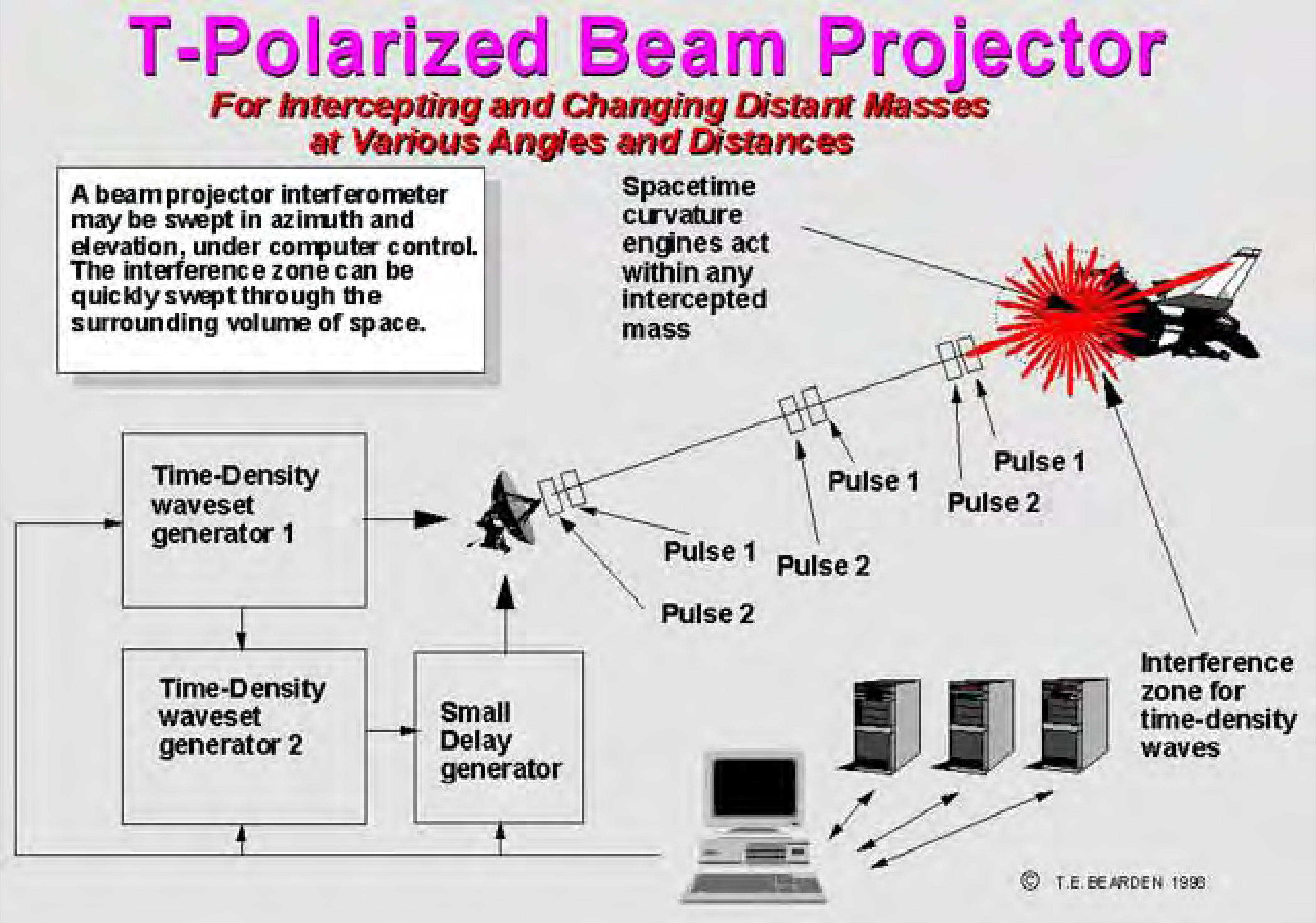

T-Polarized Beam Projector

For Intercepting and Changing Distant Masses at Various Angles and Distances

A beam projector interferometer may be swept in azimuth and elevation, under computer control. The interference zone can be quickly swept through the surrounding volume of space.

Spacetime curvature engines act within any intercepted mass
Time-Density waveset generator 1
Time-Density waveset generator 2
Small Delay generator
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Interference zone for time-density waves
© T.E. BEARDEN 1996

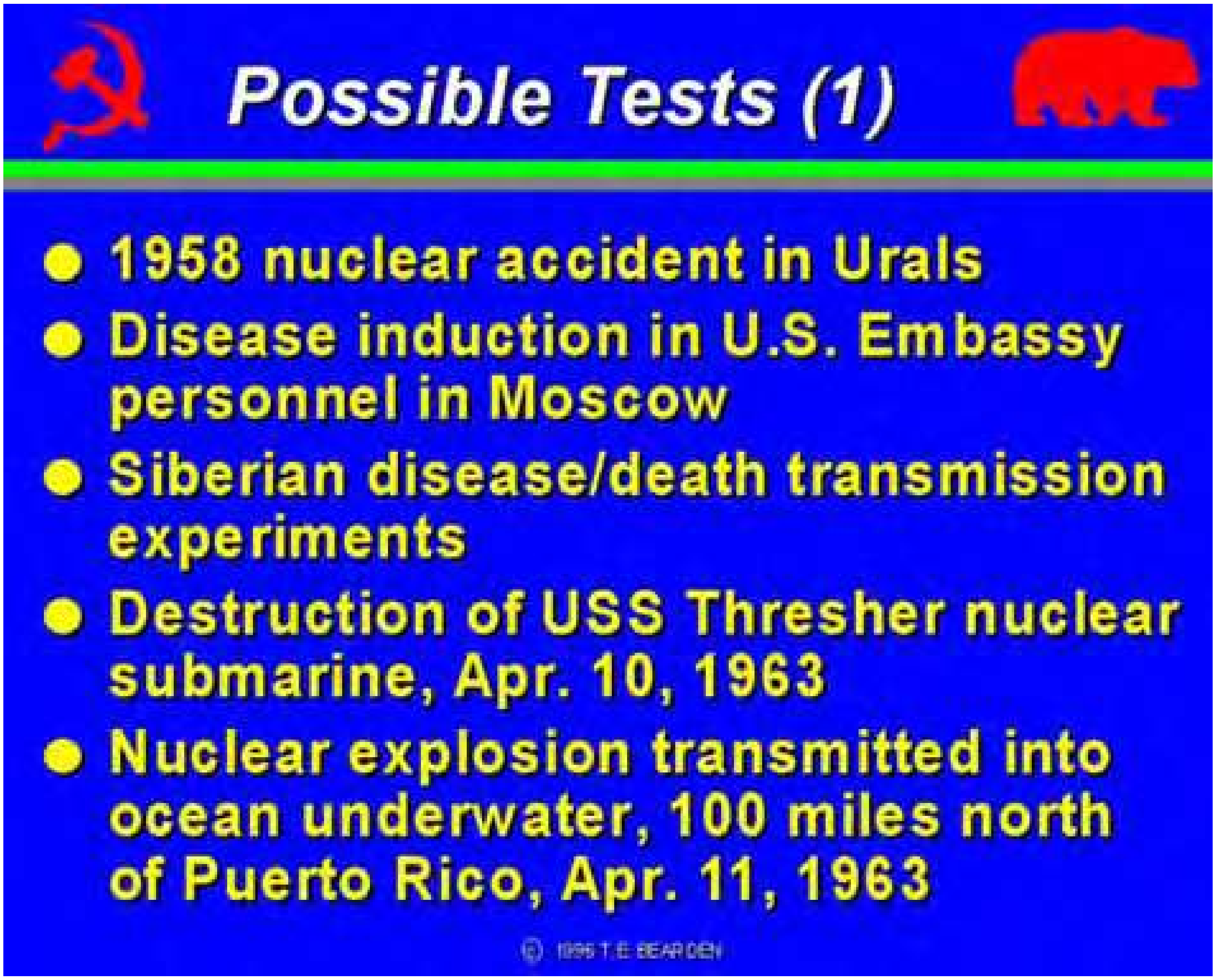

Possible Tests (1)

- 1958 nuclear accident in Urals
- Disease induction in U.S. Embassy personnel in Moscow
- Siberian disease/death transmission experiments
- Destruction of USS Thresher nuclear submarine, Apr. 10, 1963
- Nuclear explosion transmitted into ocean underwater, 100 miles north of Puerto Rico, Apr. 11, 1963

© 1996 T.E. BEARDEN

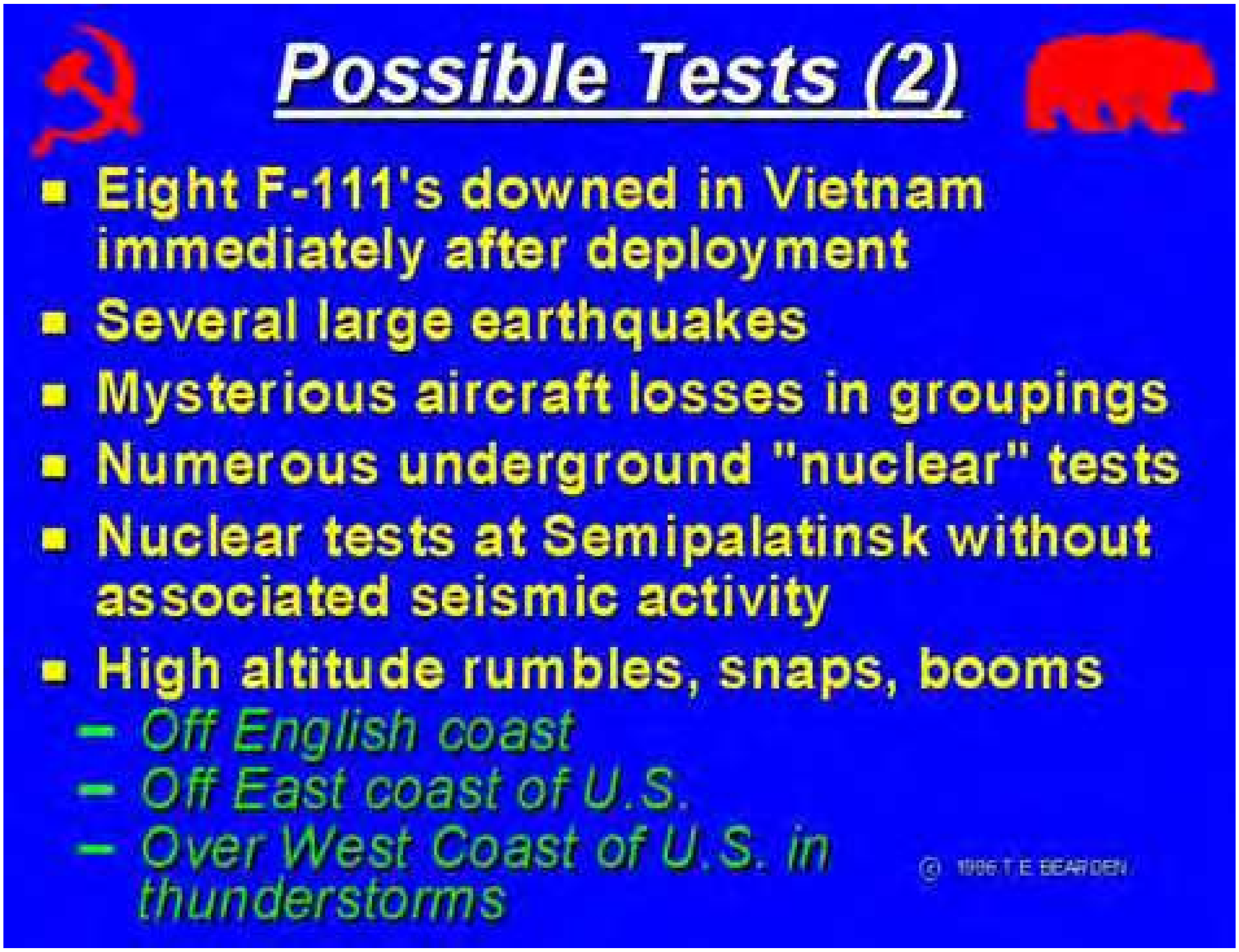

Possible Tests (2)

- Eight F-111's downed in Vietnam immediately after deployment
- Several large earthquakes
- Mysterious aircraft losses in groupings
- Numerous underground "nuclear" tests
- Nuclear tests at Semipalatinsk without associated seismic activity
- High altitude rumbles, snaps, booms
  - *Off English coast*
  - *Off East coast of U.S.*
  - *Over West Coast of U.S. in thunderstorms*

© 1996 T E BEARDEN

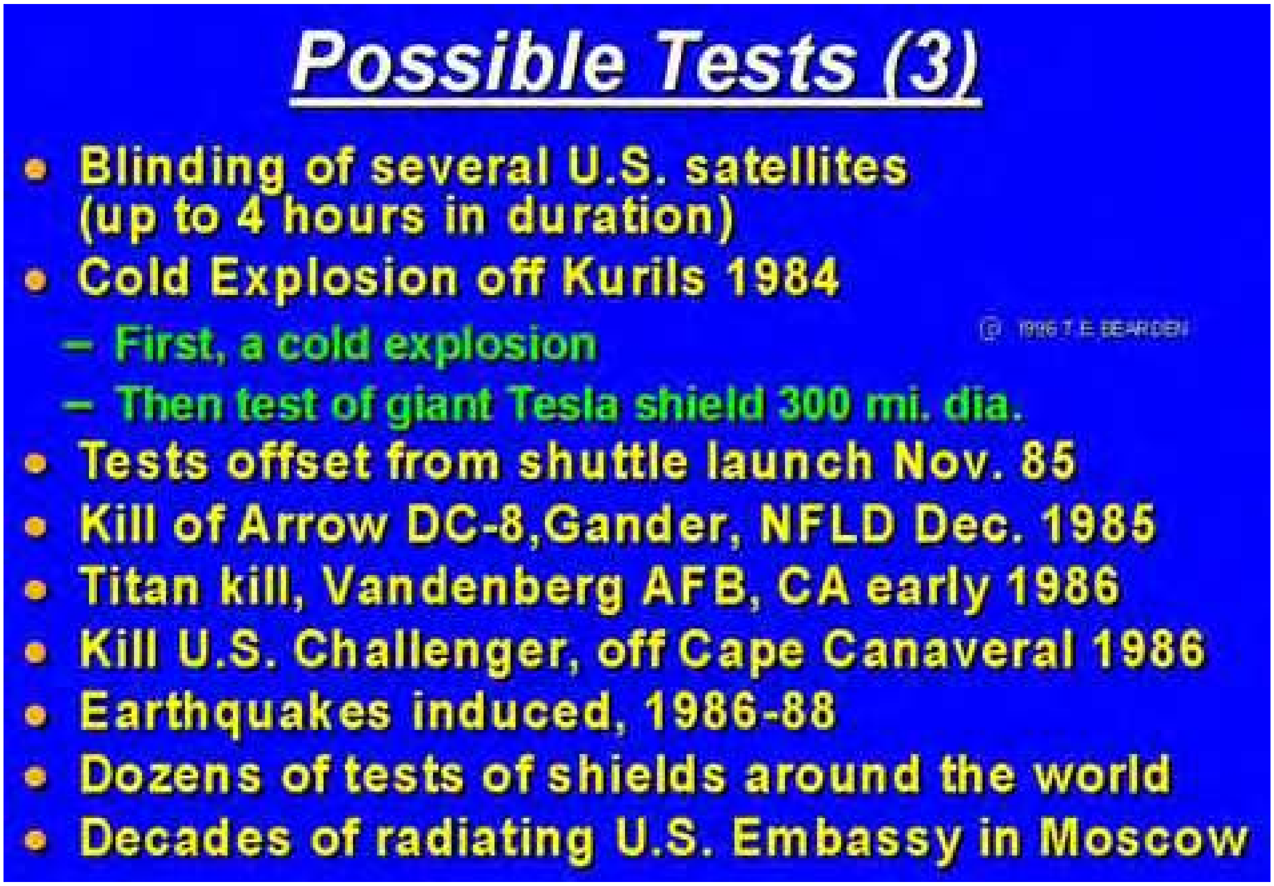

Possible Tests (3)

- Blinding of several U.S. satellites (up to 4 hours in duration)
- Cold Explosion off Kurils 1984
  - First, a cold explosion
  - Then test of giant Tesla shield 300 mi. dia.
- Tests offset from shuttle launch Nov. 85
- Kill of Arrow DC-8,Gander, NFLD Dec. 1985
- Titan kill, Vandenberg AFB, CA early 1986
- Kill U.S. Challenger, off Cape Canaveral 1986
- Earthquakes induced, 1986-88
- Dozens of tests of shields around the world
- Decades of radiating U.S. Embassy in Moscow

© 1996 T.E. BEARDEN

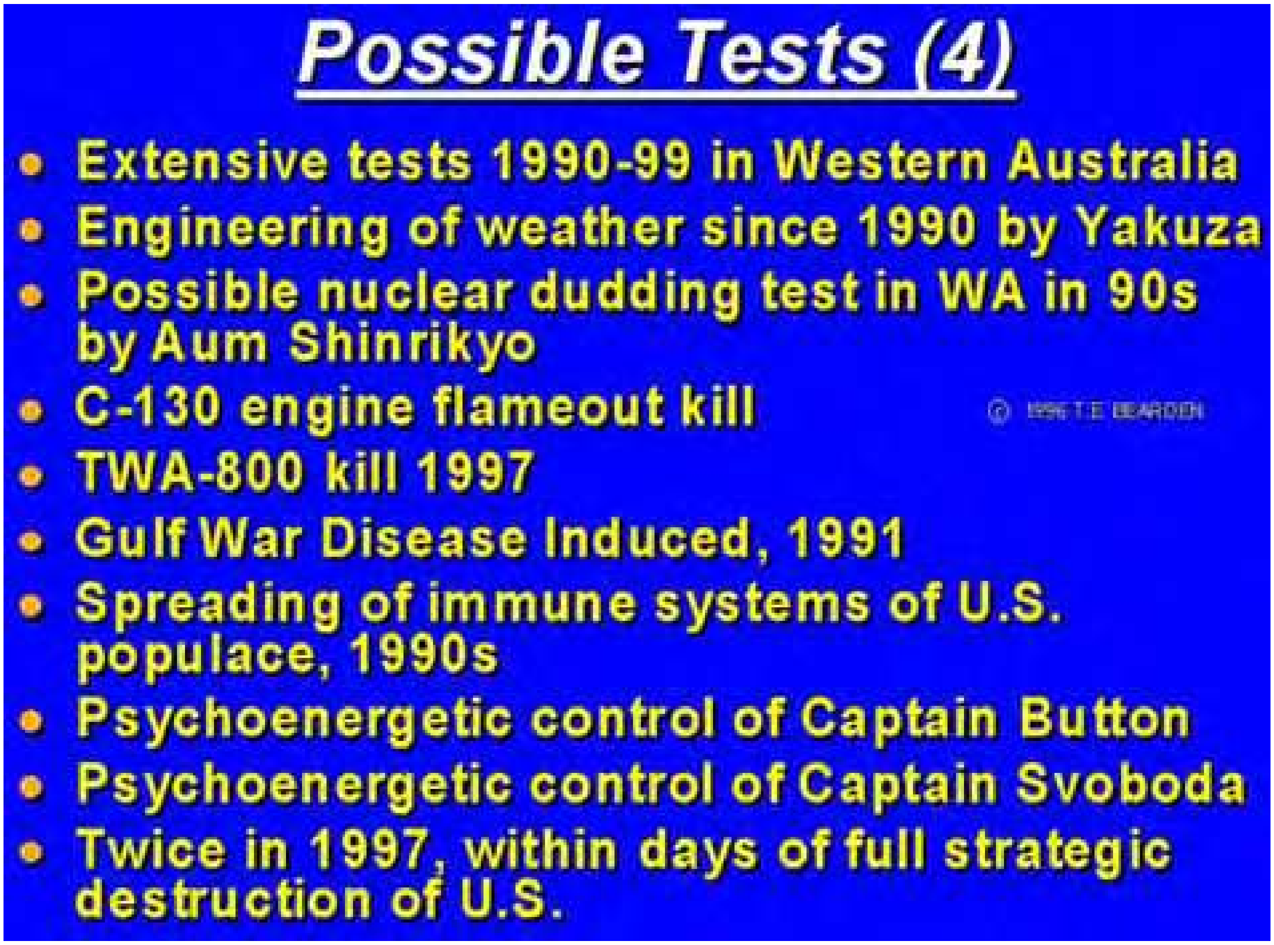

Possible Tests (4)

- Extensive tests 1990-99 in Western Australia
- Engineering of weather since 1990 by Yakuza
- Possible nuclear dudding test in WA in 90s by Aum Shinrikyo
- C-130 engine flameout kill
- TWA-800 kill 1997
- Gulf War Disease Induced, 1991
- Spreading of immune systems of U.S. populace, 1990s
- Psychoenergetic control of Captain Button
- Psychoenergetic control of Captain Svoboda
- Twice in 1997, within days of full strategic destruction of U.S.

© 1996 T.E. BEARDEN

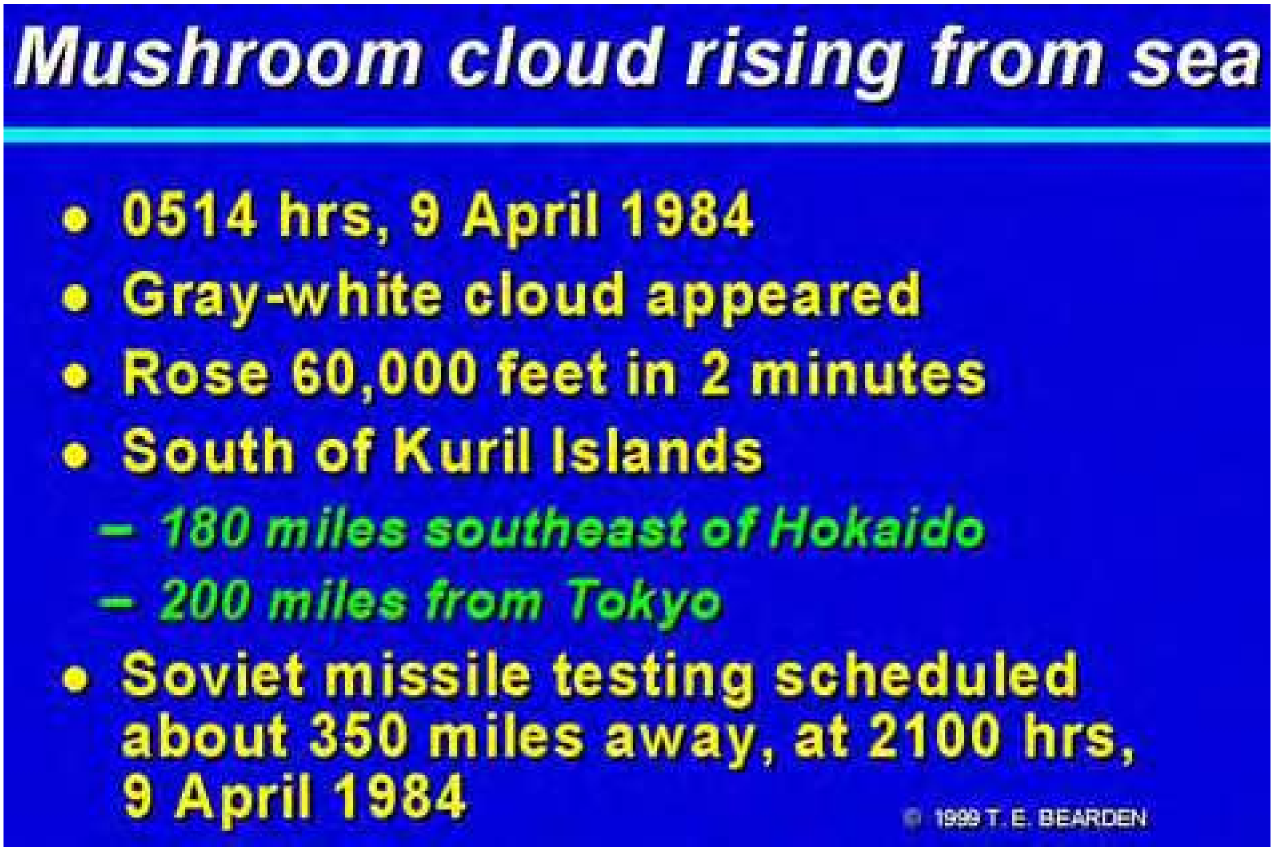

Mushroom cloud rising from sea

- 0514 hrs, 9 April 1984
- Gray-white cloud appeared
- Rose 60,000 feet in 2 minutes
- South of Kuril Islands
  - *180 miles southeast of Hokaido*
  - *200 miles from Tokyo*
- Soviet missile testing scheduled about 350 miles away, at 2100 hrs, 9 April 1984

© 1999 T. E. BEARDEN

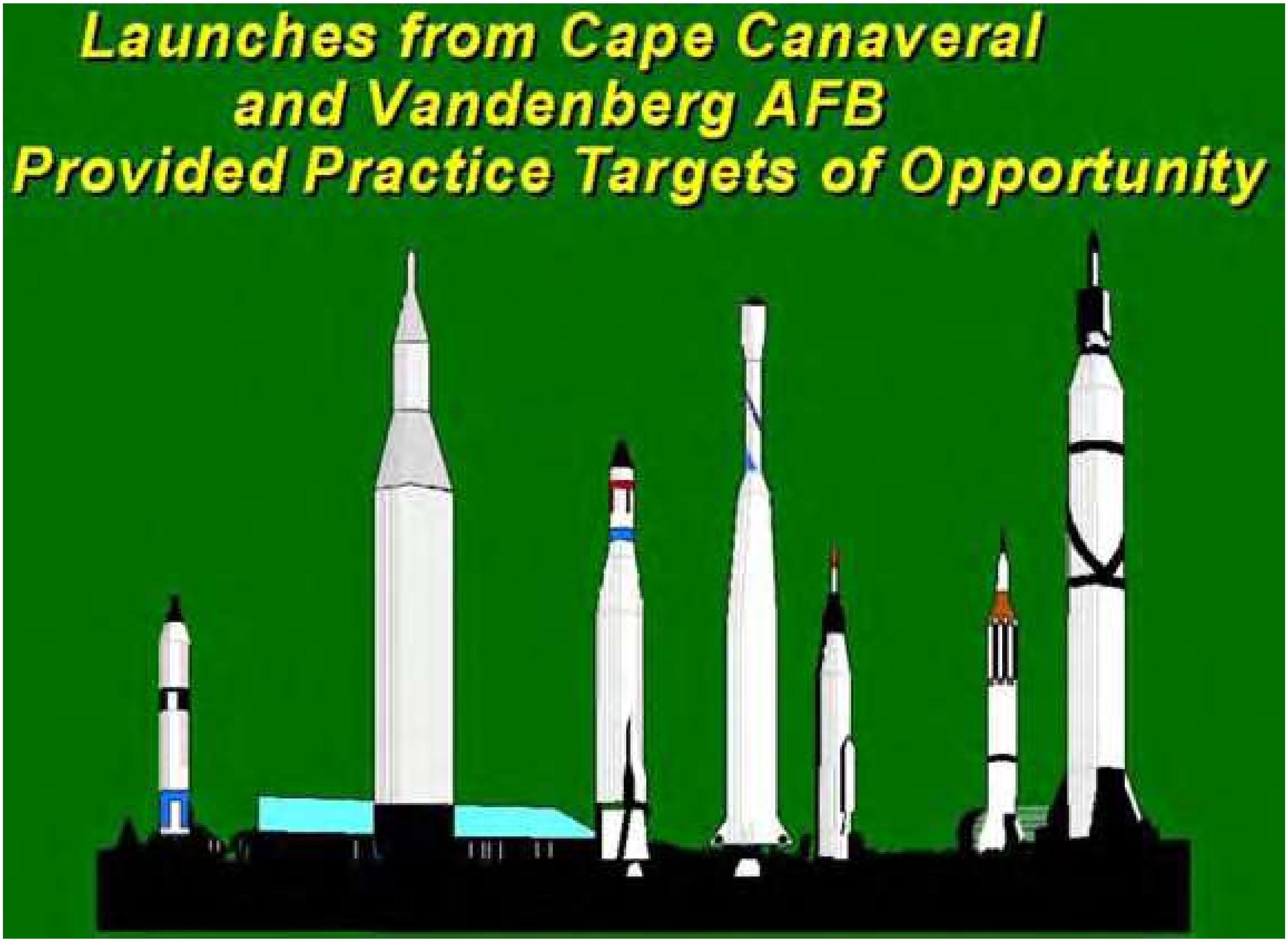

Launches from Cape Canaveral
and Vandenberg AFB
Provided Practice Targets of Opportunity

The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth

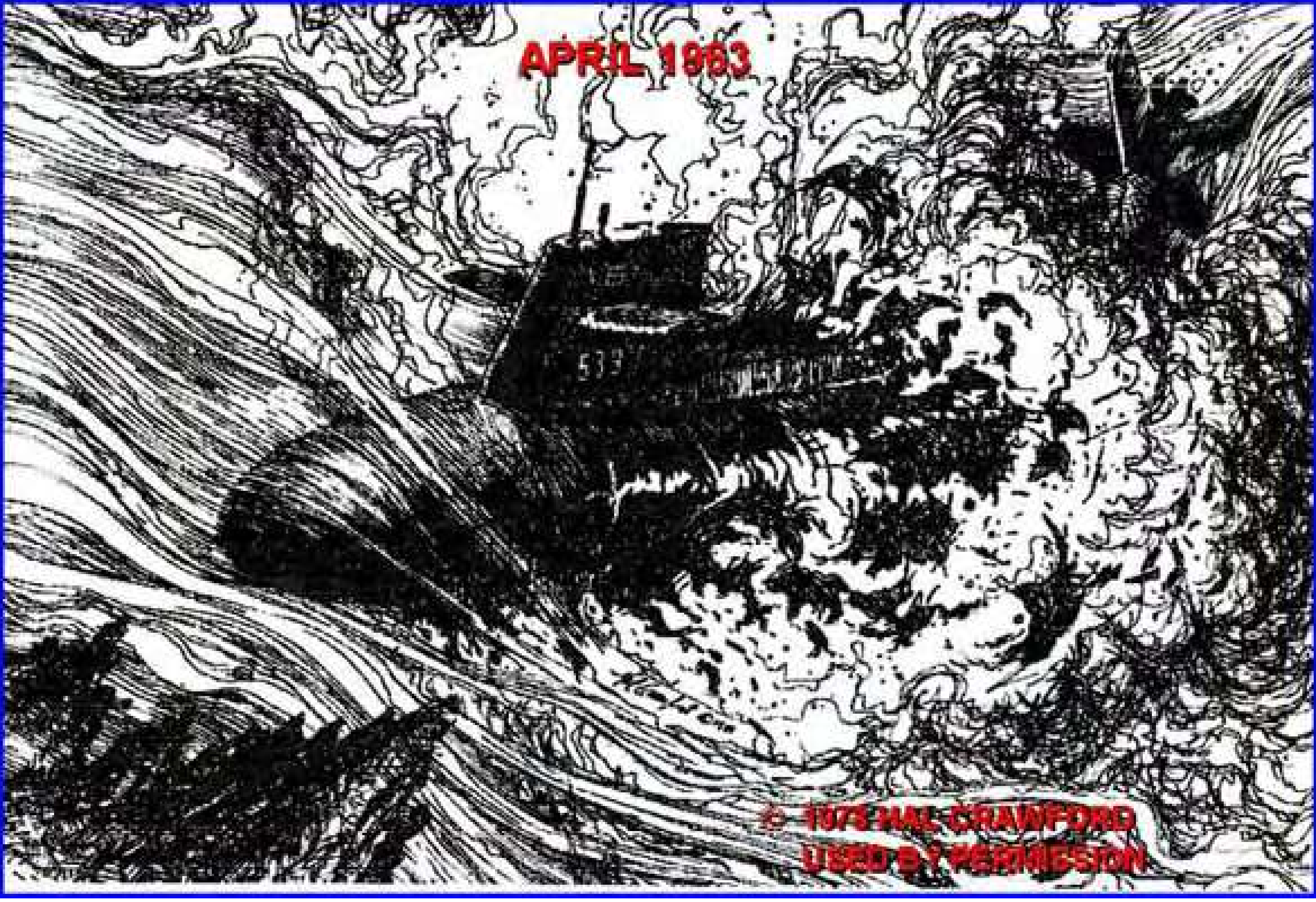

© 1978 HAL CRAWFORD
USED BY PERMISSION

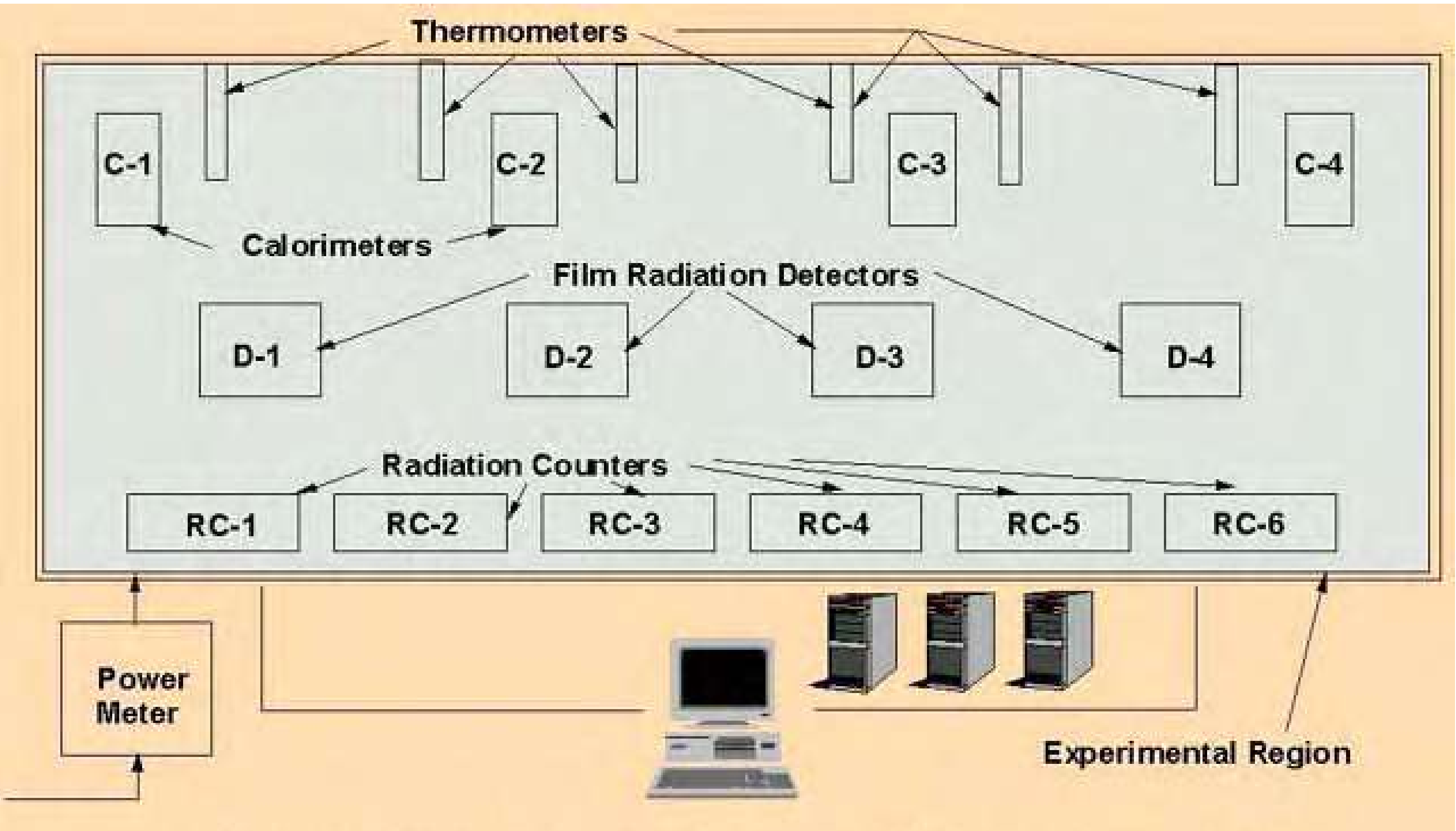
Thermometers
C-1
C-2
C-3
C-4
Calorimeters
Film Radiation Detectors
D-1
D-2
D-3
D-4
Radiation Counters
RC-1
RC-2
RC-3
RC-4
RC-5
RC-6
Power Meter
Experimental Region

Figure 32. Embodiment of an instrumental array sufficient to discriminate when time-density charging has occurred, and that transduction of time-charge decay is occurring.

© T.E. BEARDEN 1998

Time-Polarized EM Wave interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone

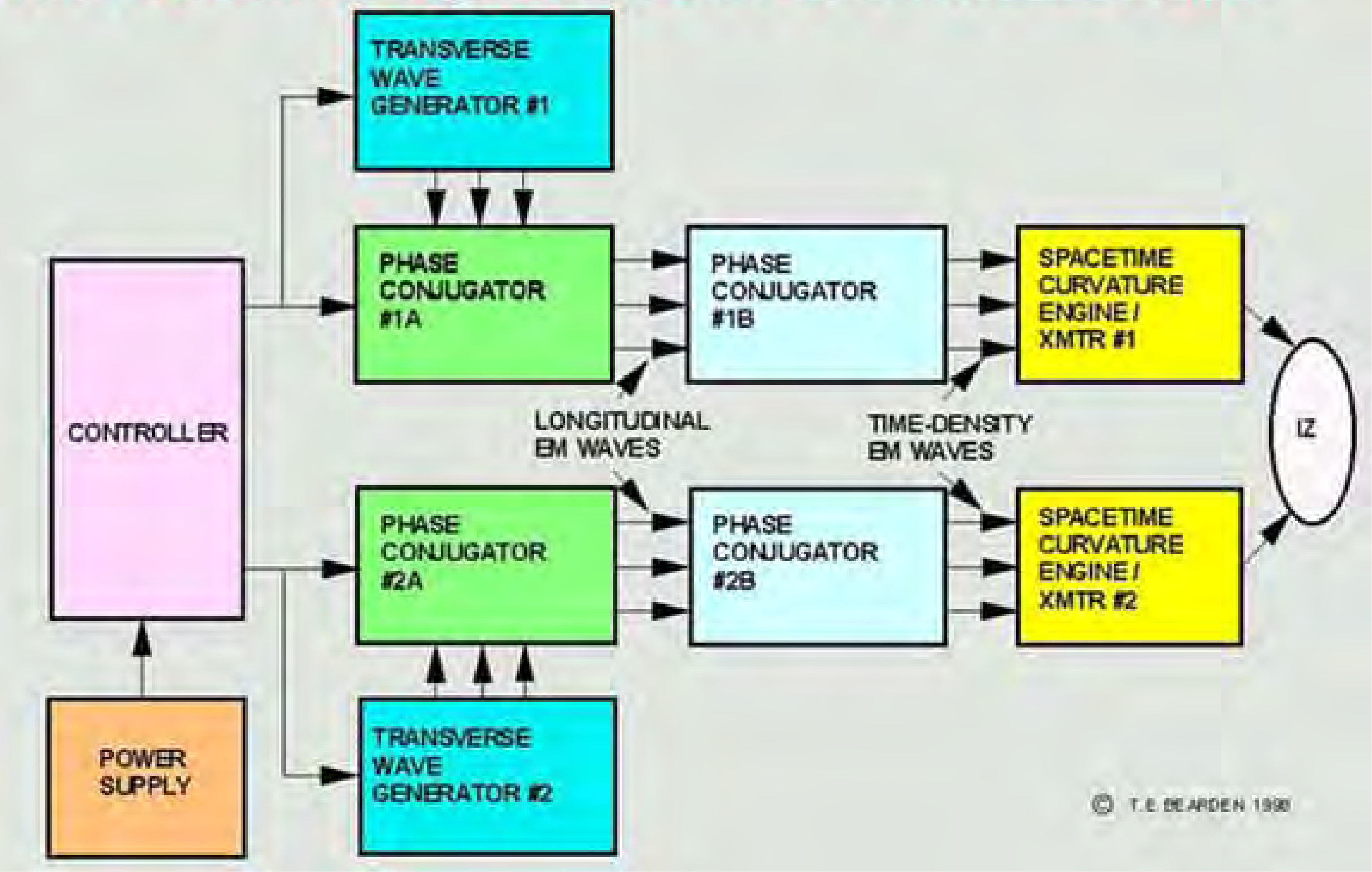
TRANSVERSE WAVE GENERATOR #1
PHASE CONJUGATOR #1A
PHASE CONJUGATOR #1B
SPACETIME CURVATURE ENGINE / XMTR #1
CONTROLLER
LONGITUDINAL EM WAVES
TIME-DENSITY EM WAVES
IZ
PHASE CONJUGATOR #2A
PHASE CONJUGATOR #2B
SPACETIME CURVATURE ENGINE / XMTR #2
POWER SUPPLY
TRANSVERSE WAVE GENERATOR #2

© T.E. BEARDEN 1990

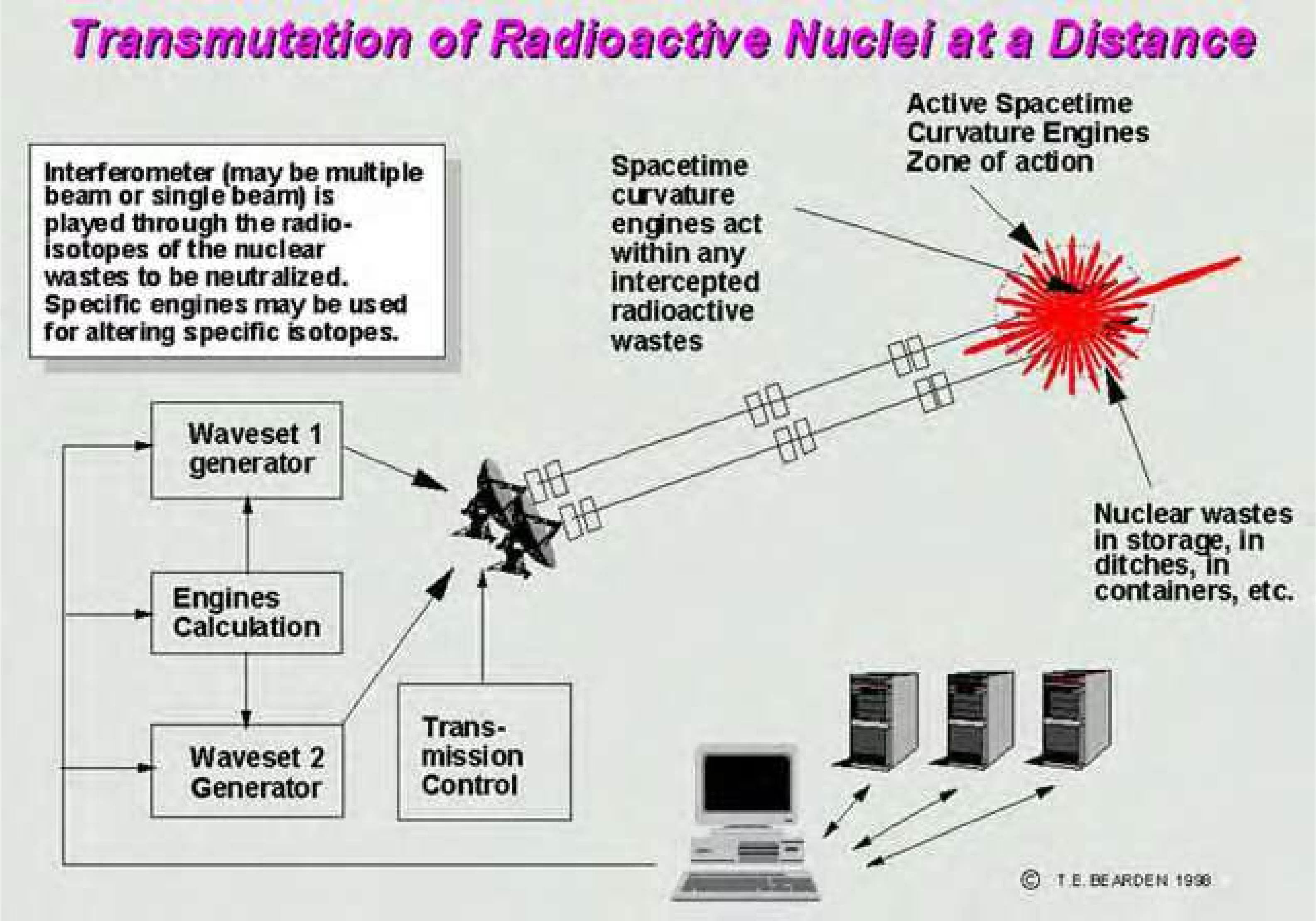

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radio-isotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes

Active Spacetime Curvature Engines Zone of action
Nuclear wastes in storage, in ditches, in containers, etc.
Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control

© T.E. BEARDEN 1998

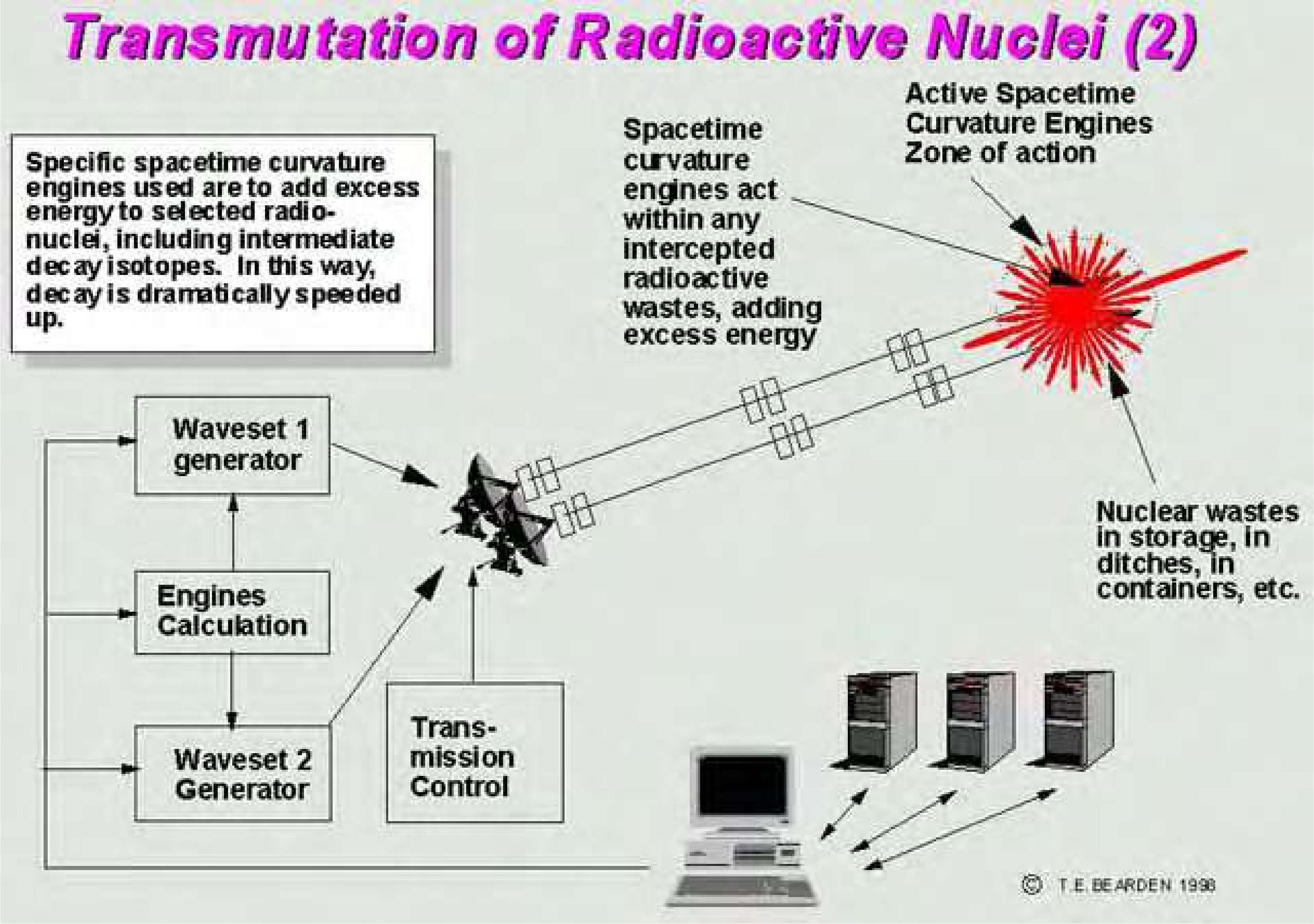

Transmutation of Radioactive Nuclei (2)

Specific spacetime curvature engines used are to add excess energy to selected radio-nuclei, including intermediate decay isotopes. In this way, decay is dramatically speeded up.

Spacetime curvature engines act within any intercepted radioactive wastes, adding excess energy

Active Spacetime Curvature Engines Zone of action

Nuclear wastes in storage, in ditches, in containers, etc.

Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control

© T.E. BEARDEN 1998

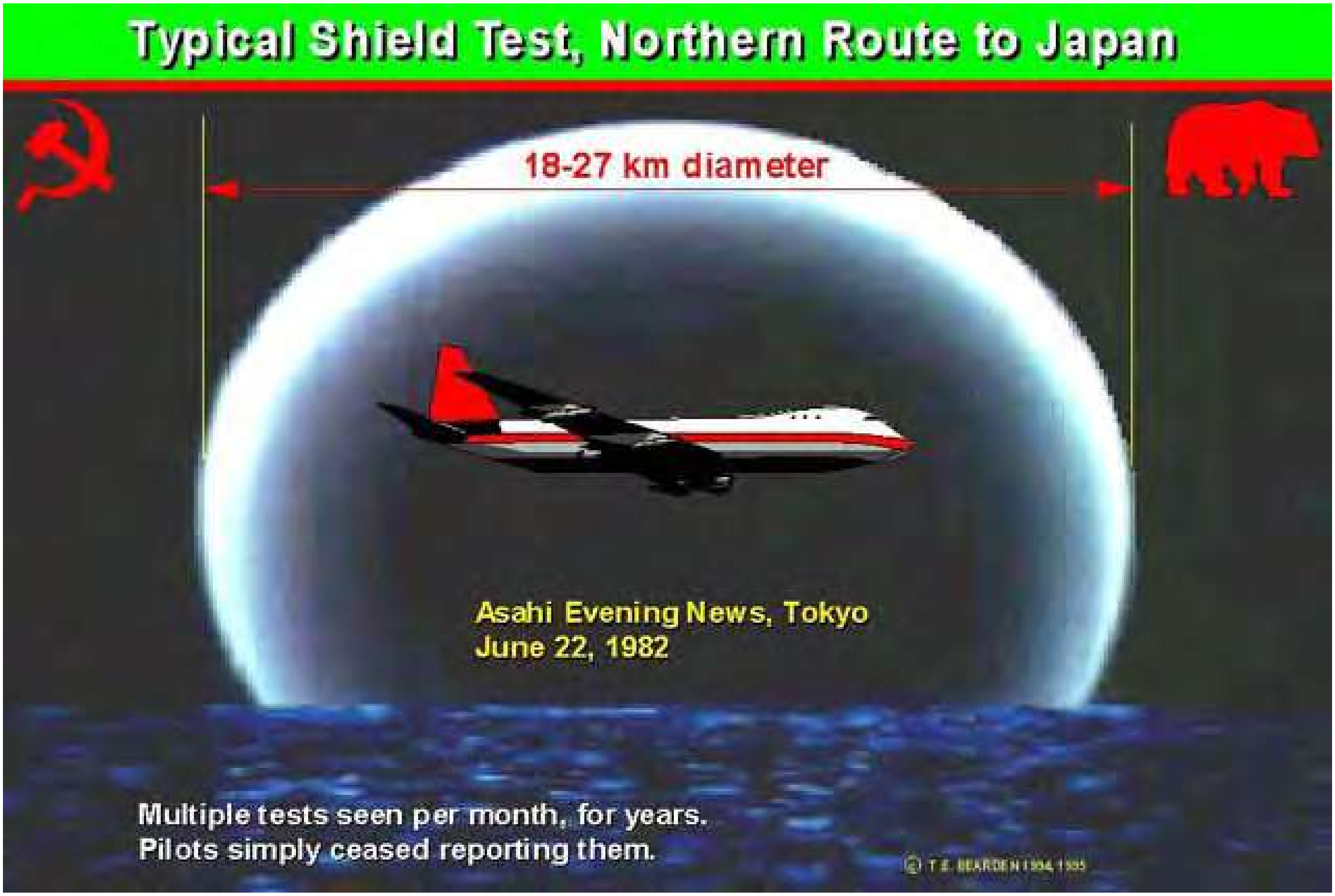

Typical Shield Test, Northern Route to Japan

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T. E. BEARDEN 1994, 1995

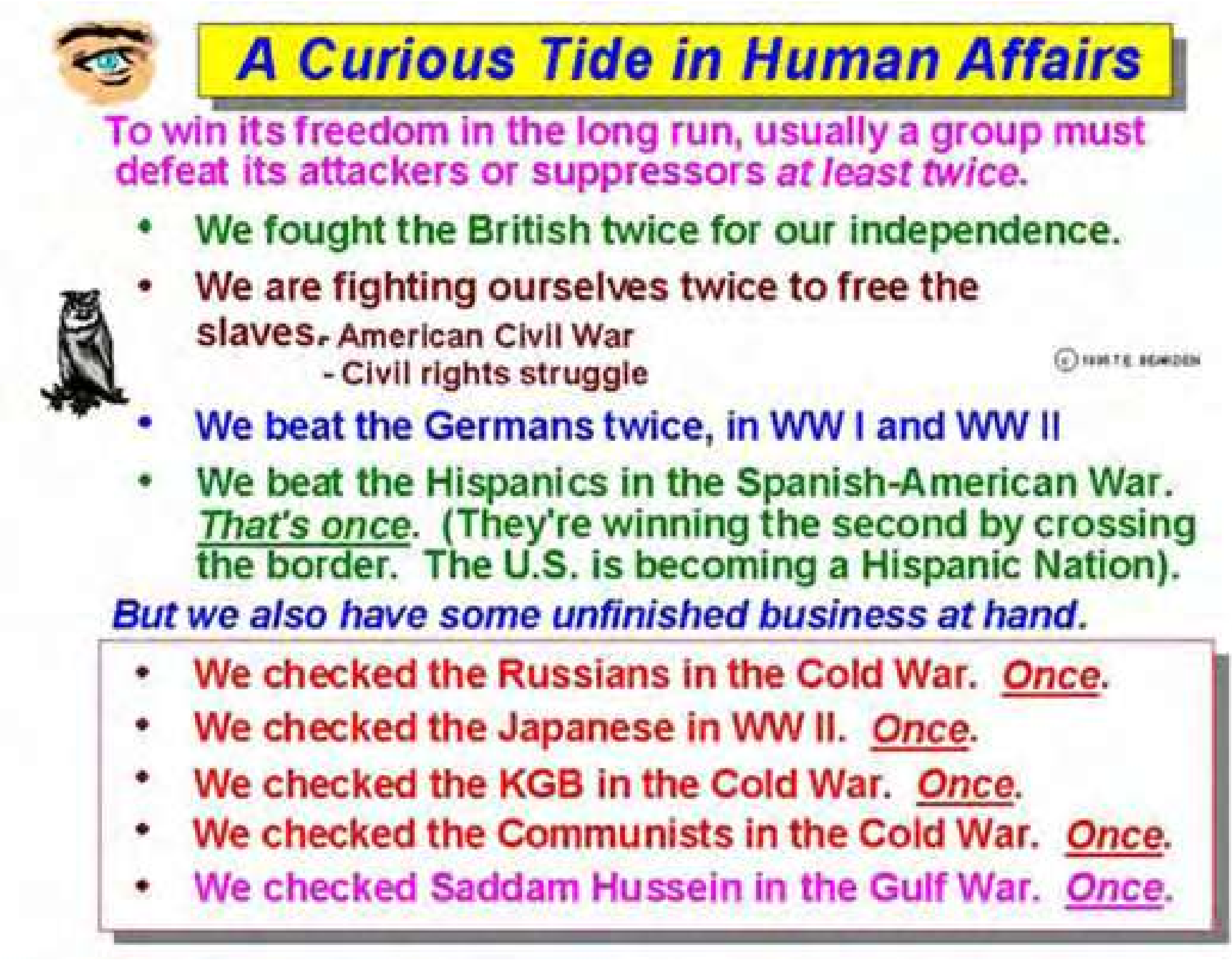

A Curious Tide in Human Affairs

To win its freedom in the long run, usually a group must defeat its attackers or suppressors *at least twice*.

- We fought the British twice for our independence.
- We are fighting ourselves twice to free the slaves. - American Civil War
  - Civil rights struggle
- We beat the Germans twice, in WW I and WW II
- We beat the Hispanics in the Spanish-American War. *That's once*. (They're winning the second by crossing the border. The U.S. is becoming a Hispanic Nation).

But we also have some unfinished business at hand.

- We checked the Russians in the Cold War. *Once*.
- We checked the Japanese in WW II. *Once*.
- We checked the KGB in the Cold War. *Once*.
- We checked the Communists in the Cold War. *Once*.
- We checked Saddam Hussein in the Gulf War. *Once*.

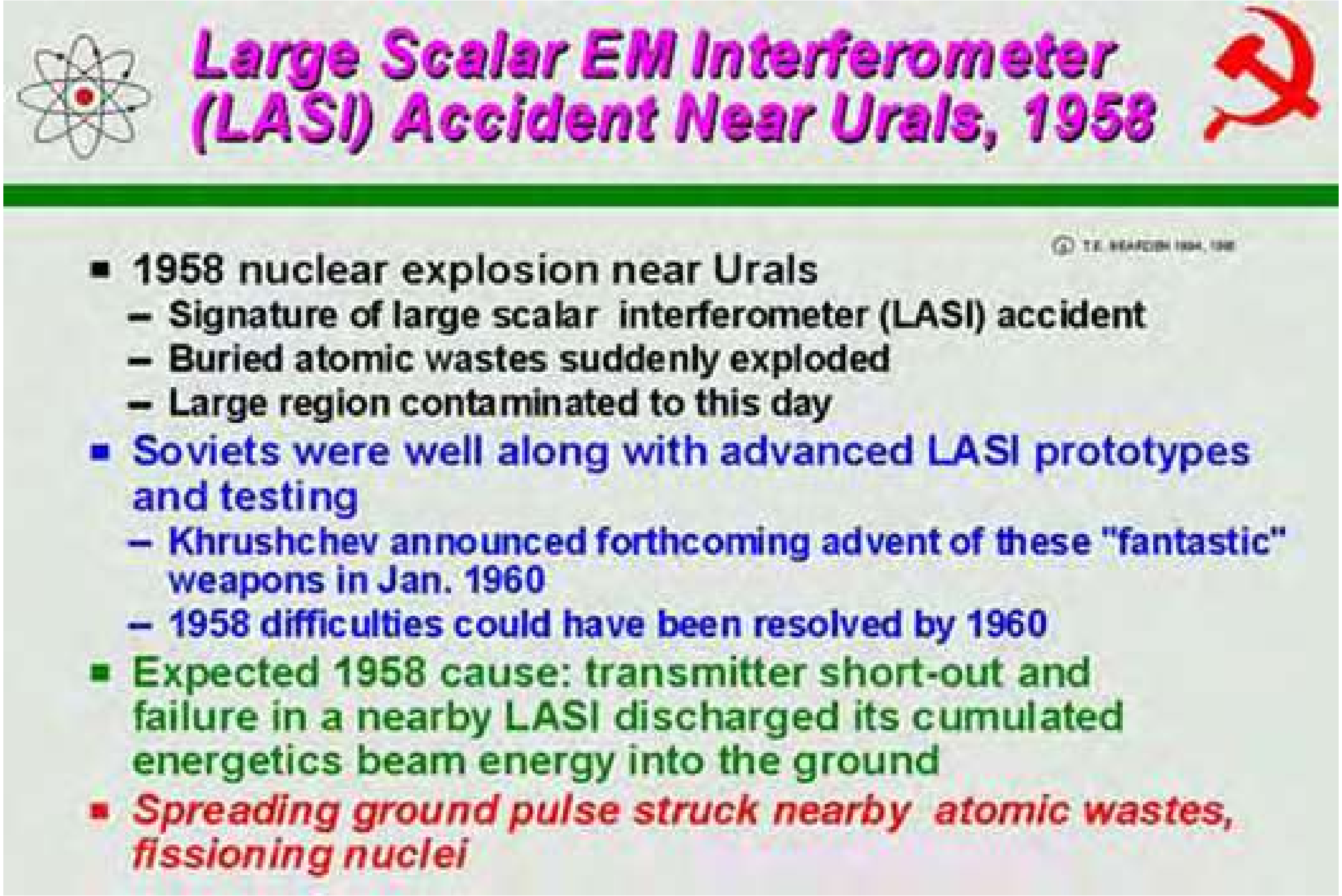

Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958

- 1958 nuclear explosion near Urals
  - Signature of large scalar interferometer (LASI) accident
  - Buried atomic wastes suddenly exploded
  - Large region contaminated to this day
- Soviets were well along with advanced LASI prototypes and testing
  - Khrushchev announced forthcoming advent of these "fantastic" weapons in Jan. 1960
  - 1958 difficulties could have been resolved by 1960
- Expected 1958 cause: transmitter short-out and failure in a nearby LASI discharged its cumulated energetics beam energy into the ground
- *Spreading ground pulse struck nearby atomic wastes, fissioning nuclei*

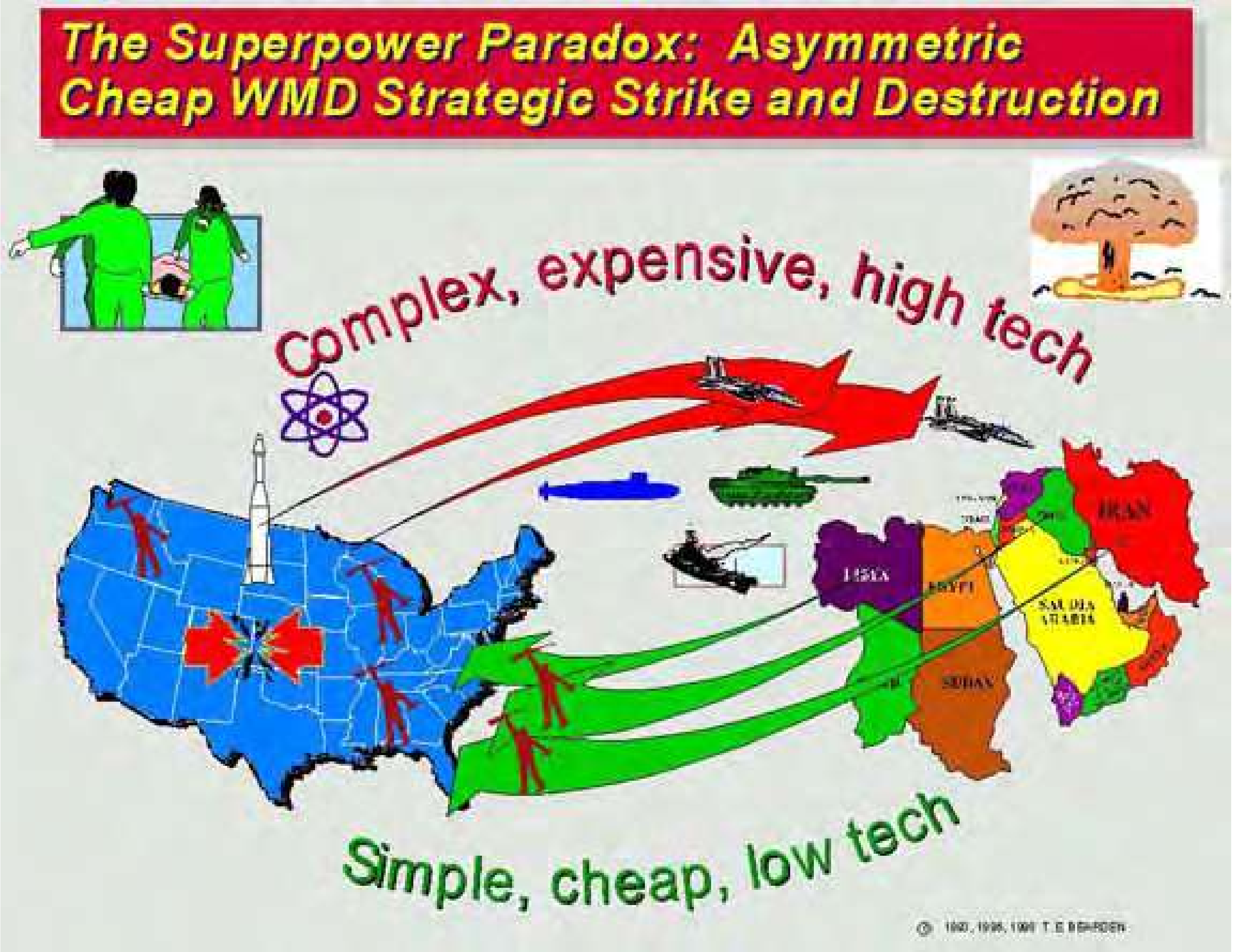

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN

Simple, cheap, low tech

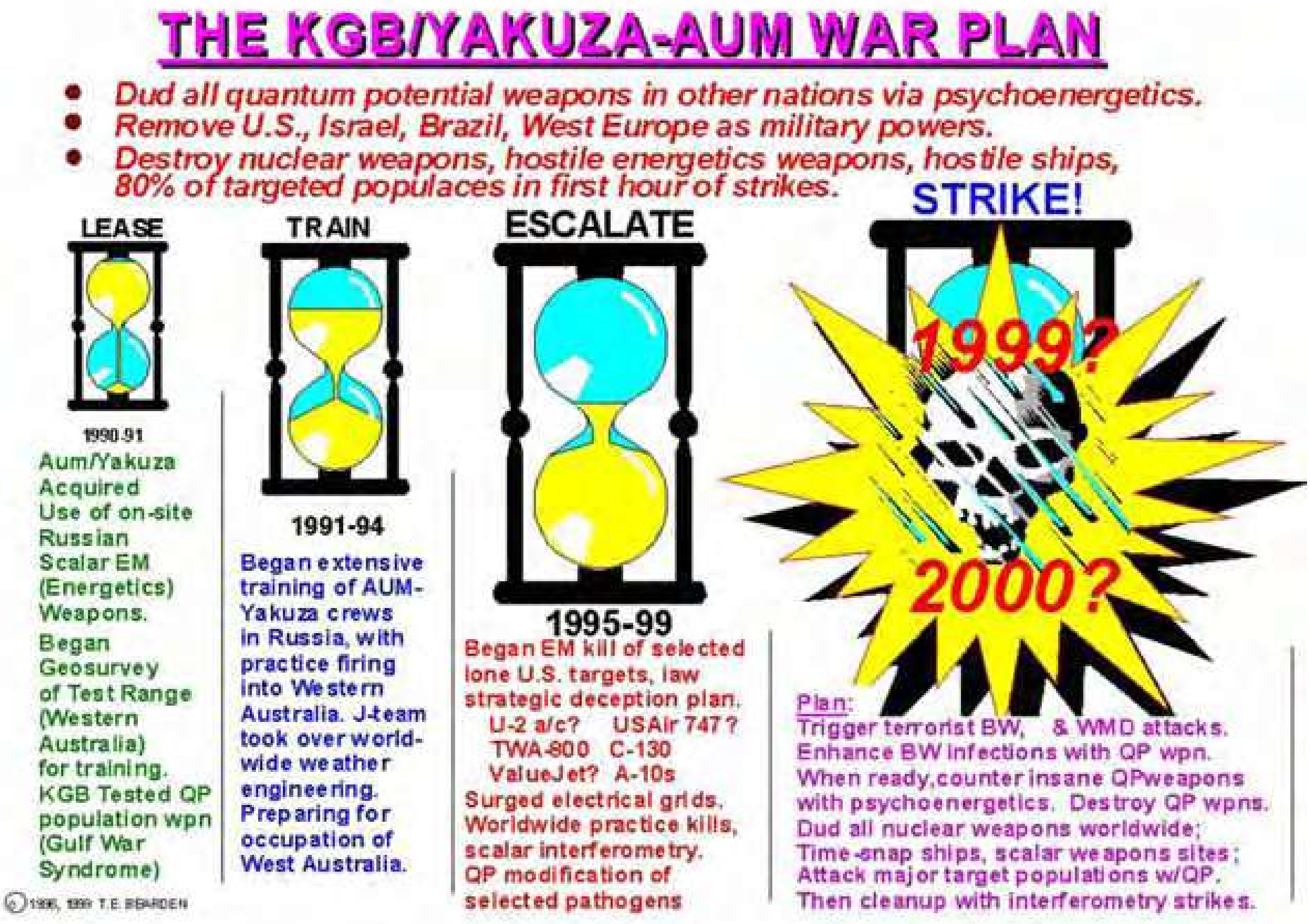

THE KGB/YAKUZA-AUM WAR PLAN

- *Dud all quantum potential weapons in other nations via psychoenergetics.*
- *Remove U.S., Israel, Brazil, West Europe as military powers.*
- *Destroy nuclear weapons, hostile energetics weapons, hostile ships, 80% of targeted populaces in first hour of strikes.*

LEASE

1990-91

Aum/Yakuza Acquired Use of on-site Russian Scalar EM (Energetics) Weapons.

Began Geosurvey of Test Range (Western Australia) for training.

KGB Tested QP population wpn (Gulf War Syndrome)

TRAIN

1991-94

Began extensive training of AUM-Yakuza crews in Russia, with practice firing into Western Australia. J-team took over world-wide weather engineering. Preparing for occupation of West Australia.

ESCALATE

1995-99

Began EM kill of selected lone U.S. targets, law strategic deception plan.

U-2 a/c? USAir 747?
TWA-800 C-130
ValueJet? A-10s

Surged electrical grids. Worldwide practice kills, scalar interferometry. QP modification of selected pathogens

STRIKE!

1999?

2000?

Plan:
Trigger terrorist BW, & WMD attacks.
Enhance BW infections with QP wpn.
When ready, counter insane QPweapons with psychoenergetics. Destroy QP wpns.
Dud all nuclear weapons worldwide;
Time-snap ships, scalar weapons sites;
Attack major target populations w/QP.
Then cleanup with interferometry strikes.

© 1996, 1999 T.E. BEARDEN

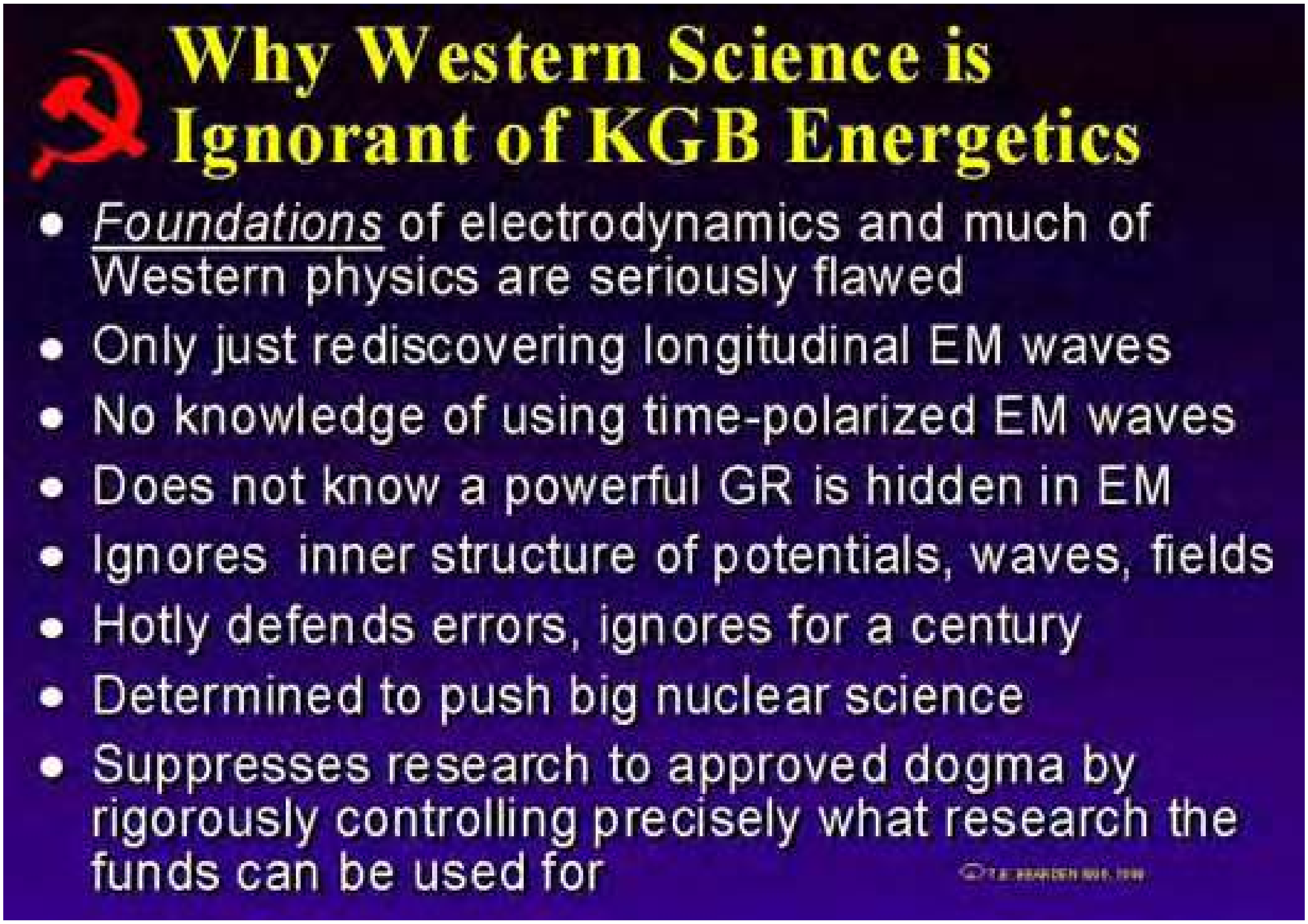

Why Western Science is Ignorant of KGB Energetics

- *Foundations* of electrodynamics and much of Western physics are seriously flawed
- Only just rediscovering longitudinal EM waves
- No knowledge of using time-polarized EM waves
- Does not know a powerful GR is hidden in EM
- Ignores inner structure of potentials, waves, fields
- Hotly defends errors, ignores for a century
- Determined to push big nuclear science
- Suppresses research to approved dogma by rigorously controlling precisely what research the funds can be used for

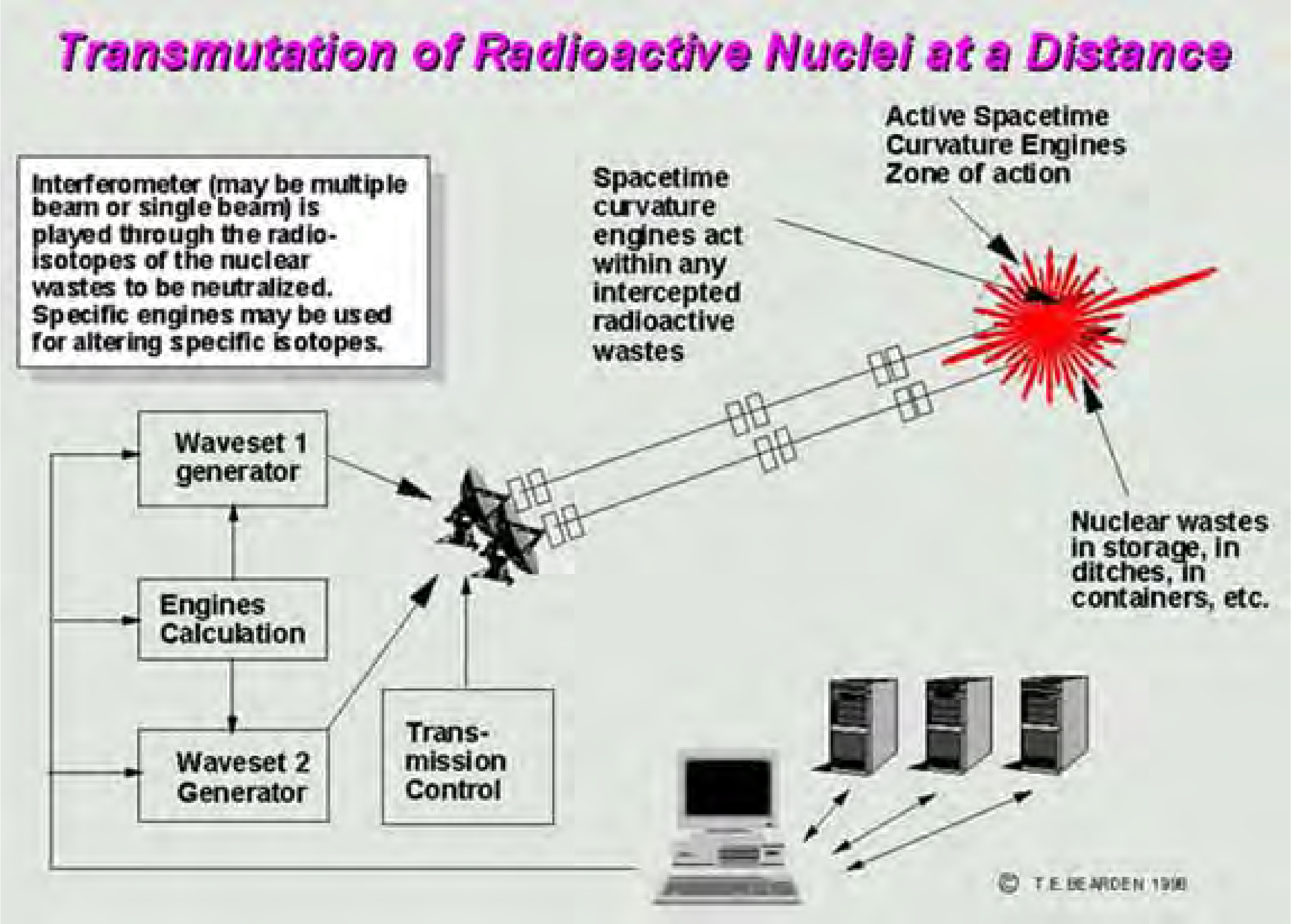

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radio-isotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes

Active Spacetime Curvature Engines Zone of action

Nuclear wastes in storage, in ditches, in containers, etc.

Waveset 1 generator

Engines Calculation

Waveset 2 Generator

Trans-mission Control

© T.E. BEARDEN 1998

THE YAKUZA

(JAPANESE MAFIA)

To do business with Japan is to do business with the Yakuza

The Last Samurai, Defending Bushido

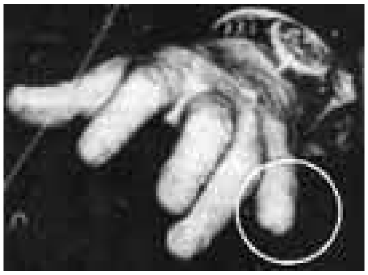

© 1995, 1996, 2001 T.E. BEARDEN

- **3 Major Yakuza Groups, In Coalition.**
- **Penetrate Japanese Government.**
- **Open Offices, Signs on Street.**
- **Public Romanticizes Yakuza as Last Japanese Samurai**
- **In Large High-Tech Corporations.**
- **In Control of Japanese Banks.**
- **Japanese Real-Estate Rip-off.**
- **$500 - 1,000 Billion per Year Income.**
- **Rapidly Spreading in U.S.**
- **Uniting Mafia Groups Worldwide.**
- **Coalition with Aum Shinrikyo.**
- **Leased Russian Energetics Weapons.**
- **Crews On-site in Russia, Trained.**
- **Practice Firings into Western Australia for Three Years. (Dozens of Shots per Month)**
- **Now in Escalation Phase, Killing U.S. Targets.**
- **All-out Strike on U.S. ready 1999-2000.**
- **Now held in check by KGB, dead-man fuzed by small nation friendly to U. S.**

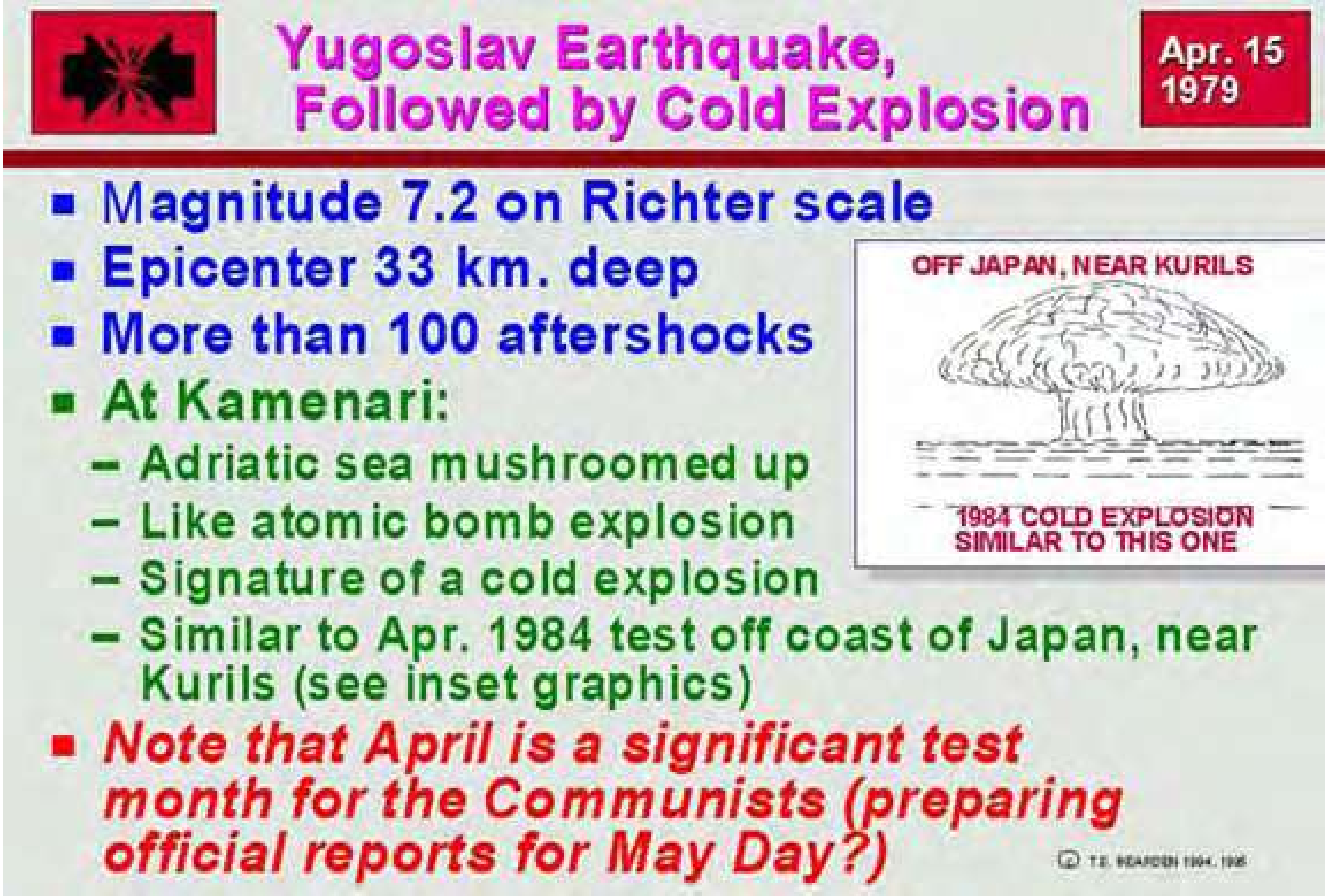

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- *Note that April is a significant test month for the Communists (preparing official reports for May Day?)*

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

SCALAR WEAPONS

UNDERSTANDING THESE MAKES IT EASIER TO UNDERSTAND THE NATURE OF THE AETHERS. WE'LL START WITH TOM BEARDEN'S FACTUAL INFO ABOUT THE RESULTS OF THE RUSSIAN'S SCALAR WEAPONS PROGRAM

http://www.cheniere.org/books/ferdelance/intro.htm

THE RUSSIANS USED SCALAR WEAPONS TO BLOW UP THE CHALLENGER SPACE SHUTTLE, A NUMBER OF OTHER LAUNCH VEHICLES, VARIOUS JET AIRPLANES, MODIFY THE WEATHER AND KILL UNTOLD NUMBERS OF BIRDS & FISH, WHICH HAS HAPPENED AGAIN RECENTLY. HAARP DEVICES CAN DO THIS.

SIMPLY PUT: A SCALAR WEAPON IS A FIGURE 8 WAVE FORM THAT IS NOT IN HARMONY WITH AN AMBIENT AETHER, IF POWERFUL ENOUGH THE SCALAR WEAPON CHANGES THE SIZE OF THE BACKGROUND SCREEN AETHER, WHICH THEN NO LONGER SUPPORTS THE MANIFESTATION OF MATTER, WHICH DESTABILIZES AND EXPLODES.

THE REASON WE ARE NOT ALL SPEAKING RUSSIAN IS BECAUSE THE MOST UNSTABLE MATTER IS IN NUCLEAR WARHEADS - DESTROYING THE ENTIRE WORLD IS NOT WINNING A WAR. PROVIDED, OF COURSE, THAT YOU LIVE ON THAT WORLD - THERE IS EXCELLENT EVIDENCE THAT A SCALAR WEAPON WAS USED TO DESTROY THE BIG BEAUTIFUL PLANET WHERE THE ASTEROID BELT IS NOW. THAT'S IN THE BOOK "COSMIC WAR".

GIGGLE, LAUGHING MY ASS OFF AT THIS ONE FROM BEARDEN:

It has also become fashionable in the West to believe that all the laws of physics are already discovered. We assume we already know all of them. While we have been pridefully crowing this tune, the Soviets have been steadily discovering new laws in secret, as well as new ways to circumvent the old laws.

THE SAD PART IS THAT IT HASN'T CHANGED - WESTERN SCIENTISTS ARE IDIOTS WHO ARE LEADING THE WORLD INTO DESTRUCTION. OR SITTING BACK AND WATCHING THE DESTRUCTION, WRINGING THEIR HANDS AND SAYING THAT FUKUSHIMA RADIATION WILL BE WITH US FOR 1000s OF YEARS. B.S. THERE ARE A NUMBER OF ENERGETIC WAYS TO CANCEL RADIATION INSTANTLY.

BEARDEN TAUGHT ME ABOUT SCALAR ENERGY - BUT, THEN HE HAS STARTED PARROTING LAME STREAM PHYSICS - PROBABLY TO KEEP HIS HEAD FROM BEING SEPARATED FROM HIS BODY BY OUR CONTROLLERS.

SCALAR WEAPONS ARE DESCRIBED IN A NUT SHELL ABOVE. BUT, WE'LL EXPAND IT A LITTLE - THE SIMPLE FACT THAT THEY WORK IS PROOF THAT THE AETHERS IS THE BACKGROUND SCREEN THAT SUPPORTS THE MANIFESTATION OF MATTER, WHICH IS ALL DIVINE I.E. MADE FROM IMAGES.

THE AETHERS PROGRESS IN SIZE ACCORDING TO THE FIBONACCI SERIES, BECAUSE EVERYTHING FROM THE QUANTUM TO THE COSMOS MANIFESTS ACCORDING TO THE FIBONACCI SERIES - ITS ALL FRACTAL I.E. A RECURRING MATHEMATICAL PATTERN, THAT IS PERFECT THROUGHOUT CREATION. HERE'S THE PROOF - OUR ENTIRE GALAXY IS IN PERFECT HARMONY WITH ONE NUMBER = RIDICULOUSLY AMAZING

http://blog.hasslberger.com/docs/PERFECTION.pdf

THUS, IF YOU CHANGE THE BACKGROUND SCREEN, THEN MATTER HAS TO CHANGE = THE RESULT OF USING A SCALAR WEAPON.

THE PRECISE WAY THE WEAPON WORKS IS TO IMPOSE A MAN MADE SCALAR WAVE ON AN AREA OF SPACE. GIVEN ENOUGH POWER TO THE SCALAR WAVE, IT WILL CAUSE THE AETHER IN THAT AREA TO MATCH ITS SIZE (LIKE ATTRACTS LIKE). WE ARE TALKING UNBELIEVABLY SIMPLE.

YOU CAN MAKE ONE YOURSELF - AMAZINGLY THE REQUIRED WINDINGS OF A SCALAR PRODUCING COIL CAN BE USED AS A WEAPON OR FOR HEALING - ITS CALLED A RADIONIC WITNESS WELL WHEN USED FOR HEALING.

TO MAKE A WITNESS WELL ONE WINDS INSULATED WIRE AROUND A GAP IN SPACE I.E. THE CENTER OF A TORROID. THE WIRE IS WOUND TO THE LEFT AND TO THE RIGHT. THE MORE WINDINGS THE MORE POWERFUL. AND, OF COURSE, THE POWER OF THE WAVE IS A FUNCTION OF THE POWER OF THE DC ELECTRICITY USED TO ACTIVATE THE DOUBLE WOUND COIL. SUCH A COIL PRODUCES A FIGURE 8 WAVE THAT IS THE SIZE OF THE CENTER OF THE COIL. IF THAT SIZE HAPPENS TO BE THE SIZE OF AN AMBIENT AETHER, YOU'RE NOT GOING TO HAVE A WEAPON.

IF YOU MAKE SUCH A COIL FOR HEALING, POWER IS NOT REQUIRED, IT AUTOMATICALLY CONNECTS WITH THE AETHERS IN A POSITIVE WAY.

IF YOU MAKE IT AS A WEAPON ADD LOTS OF POWER AND BE A LONG WAYS AWAY, WHEN YOU TURN IT ON. THE DESTRUCTIVE EFFECT IS INSTANTANEOUS - I'VE SEEN IT, WE BUILT ONE.

SCALAR WAVES ARE ALSO REFERRED TO AS PHASE CONJUGATED WAVES, MEANING TWO SINE WAVES THAT ARE 180 DEGREES OUT OF PHASE FROM ONE ANOTHER. (WESTERN LAME STREAM IDIOTS SAY THAT THOSE TWO WAVES CANCEL ONE ANOTHER.)

ONE SINE WAVE IS SAID TO RUN IN TIME FORWARD AND THE OTHER IN TIME REVERSE. AND THERE ARE A NUMBER OF EFFECTS THAT PROVE THIS. ONE EXAMPLE: DON BRANDT (ONE OF BEARDEN'S PROTEGEES) USED A SCALAR WAVE TO REGROW A LEG ON A CAT (A KID HAD SHOT IT OFF WITH A .22), WHEN DON FINISHED, THE CAT HAD ITS LEG BACK AND ALL OF THE CLOCKS WITHIN FIVE MILES WERE 15 MINUTES BEHIND TIME - THAT WAS THE LAST TIME HE PLAYED WITH THAT.

IN 2-D THERE IS NO LINEAL TIME.

BOTTOM LINE: SCALAR WEAPONS PROVE THE NATURE OF THE AETHER AS THE BACKGROUND SCREEN - IF THAT WAS NOT SO, THEN SCALAR WEAPONS WOULD NOT WORK.

ALSO, SCALAR WAVES THAT ARE IN HARMONY WITH AN AMBIENT AETHER ARE VERY BENEFICIAL AND USEFUL - A BEE'S HONEYCOMB IS AN EXCELLENT EXAMPLE. DITTO GRAPHENE.

LOVE AND BLESSINGS

PS: THERE HAVE BEEN A NUMBER OF HUMAN "CIVILIZATIONS" IN THIS SOLAR SYSTEM AND ON THIS PLANET. OUR CONTROLLERS LIKE US TO BE STUPID, BECAUSE IT SEEMS THAT WHEN EVER WE MAKE GREAT TECHNOLOGICAL ADVANCES OR SPIRITUAL ADVANCES, THEN THEY WIPE US OUT. WE ARE AN ENTERTAINMENT SLAVE SPECIE.

THIS LAYS OUT HOW OUR DREAMS / ANUs ARE SHARED WITH OTHERS AND THAT JESUS' ENERGY IS FOUNDATIONAL
http://blog.hasslberger.com/docs/GOLDFISH_JESUS_REVELATIONS.pdf

THIS IS FUNNY AND SHOWS THAT THE EARTH IS THE BEST ENTERTAINMENT SYSTEM IN ALL OF GOD'S CREATION:
http://blog.hasslberger.com/docs/EARTH_WINS.pdf

OH - THE CHEMTRAILS WILL KILL JUST ABOUT EVERY LIVING THING ON THE FACE OF THIS PLANET, IF WE DON'T START MEDITATING AND BUILDING ORGONE DEVICES, WHICH HAS TO BE DONE EVEN IF THEY STOP SPRAYING BECAUSE THE EFFECTS OF THE CHEMTRAILS CAN AND WILL CARRY ON, EVEN IF THE SPRAYING STOPPED.

Scalar Wars

The Brave New World of Scalar Electromagnetics

by Bill Morgan

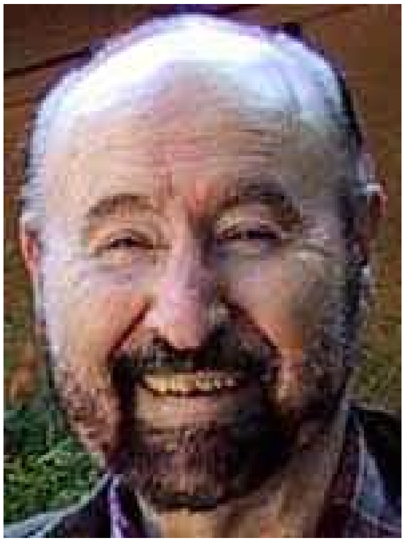

For the past six months I have been undergoing the greatest **paradigm shift** I have ever had to go through. It has rattled my nerves and shaken my bones. This intense adjustment of my "world" has come about by studying the information given by Col. Tom Bearden at his website Cheniere. The new knowledge there has necessitated a total revision of my ideas about physical reality, the world we live in, and the future of humanity. This paradigm shifting even actually made me dizzy on certain days as I tried to absorb and digest Bearden's vast amount of information. I am not a scientist at all, just a layman, and I have little comprehension of the math and high physics of this new science called "**Scalar Electromagnetics**." But there is a great deal of information at Cheniere which needs to become common knowledge as fast as possible, for the sake of the survival of life on earth. To that end I have put together this small primer of Bearden's ideas, as a kind of "beginner's guide" to his website. (I have tried to underline certain words and phrases which are part of the new "jargon" of talking about these "new" phenomena. Also I have taken the liberty of bolding certain words and phrases to help them stick in the mind).

(Note: Throughout "**EM**" means "**electromagnetic**," and "**LW**" means "**longitudinal wave**")

This article has six sections:

New Waves 2. Tapping the Waves 3. Weaponization 4. Healing 5. Psychoenergetics 6. As It Stands

1. New waves discovered

Longitudinal EM energy fills vacuum of space, the time domain of spacetime, time as compressed energy, E=tc2, waves of time, phase conjugate wave pairs.

"Electric power is everywhere present in unlimited quantities and can drive the world's machinery without the need of coal, oil, gas, or any other of the common fuels." Nikola Tesla

"At any point and at any time, one can freely and inexpensively extract enormous EM energy flows directly from the active vacuum itself." Tom Bearden

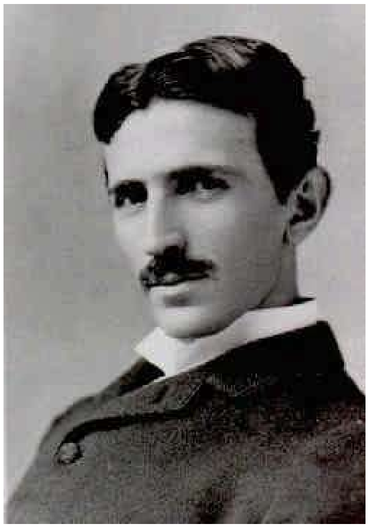

I guess the first thing to try to comprehend is that a "new" kind of electromagnetic (EM) wave has been discovered in the empty vacuum of space called “longitudinal” waves, which, when engineered, can be an inexhaustible supply of energy in great magnitude at any place in the universe. The word "new" is in quotes because the discovery really goes back to **Nikola Tesla** and his discovery of what he called "radiant energy." It is also not "new" because the Russians (KGB) have been working on this technology for over 30 years and have **weaponized** these "new" longitudinal scalar waves to a great degree.

These are the very weapons **Nikita Khrushchev** spoke of in January, 1960.

> "By 1957-8 the Soviets had progressed to the point of a giant scalar EM accident in the Urals which exploded nearby atomic wastes, devastating the area. They had also progressed to development of great new superweapons using their new energetics - weapons to which Khrushchev referred in 1960 when he informed the Soviet Presidium of a new, fantastic weapon in development, a weapon so powerful that it could wipe out all life on earth if unrestrainedly employed.'" Tom Bearden

After over 30 years of development, and extensive testing around the globe, these new scalar electromagnetic weapons are up and running and ready to go. Tom Bearden, at his website Cheniere, discusses the history of these new scalar electromagnetic weapons in his paper "Historical Background of Scalar EM Weapons."

Some Immediate Implications

> The implications of successful engineering of the longitudinal waves are enormous, and will change the world as we know it, one way or another. Among other things, these discoveries mean that:
>
> 1. The **solutions to the energy crisis and the "oil problem" are in hand**. These oil wars are unnecessary. There is endless energy available freely from the domain of time.

2. **Unbelievably powerful weapons** are not only possible, but are already operating in several nations. The many powers of these weapons are unprecedented and mind-boggling.

3. The **cure of diseases such as cancer and AIDS**, in fact nearly any disease, has become possible within a few years of sufficient funding. Everyone can be made healthy and stay healthy.

4. **Mind control on a mass scale** has now become possible, and the machines to do it are already in place in certain nations. It has become possible to mentally enslave whole populations with the twist of a few dials.

So the layman will need to understand that there is a new kind of electromagnetic energy that is altogether different from what he knows, e.g. radio, TV, cell phones, etc. The ordinary EM waves that we have known about are called transverse EM waves, to distinguish them from the new longitudinal EM waves. These scalar waves do not actually exist in our "material" world, but exist only in the vacuum of empty space, or the time domain. And we must keep in mind that this vacuum of space we speak of exists all through everything. Even our bodies are mostly empty space between atoms and molecules. So the gateway to this seething ocean of energy can be there at every point in the universe. This seething ocean of energy is all around us and all through us.

Emptiness is Full

This amazing discovery announces that the "emptiness" of empty space is in fact not empty, but a great ocean of seething energy!

Col. Bearden refers to this ocean of energy as being of the "time domain." It seems like something from Star Trek but this is where scalar electromagnetics has come to. And where it is going may be beyond anything Star Trek could have dreamed of.

We live in a 3-dimensional world, which physics calls "3-space." But there is also spacetime, or 4-space, or the "4th dimension." Then suddenly comes this amazing new knowledge that time itself is actually compressed energy. And it is energy which is compressed by exactly the same factor by which matter is considered compressed energy: the speed-of-light-squared!

So we have a new companion to the famous E=mc2. It is now paired with E=tc2 (where t is actually "delta-t," or change in time). Or (amaze your family and friends and) say:

"E equals em-cee-squared"

and

"E equals delta-tee-cee-squared"

It has a nice ring to it, like some freedom bell announcing a new era.

And as the atomic bomb released the compressed energy in matter, so can we now unleash the tremendous energy that is compressed into time itself. It gives a completely new meaning to the term "time bomb."

Some New Terminology of Scalar Electromagnetics

A new science, several new names

> "Those [Russian] weapon scientists resurrected an old term from the history of electrodynamics, called "energetics". That is their approach to a unified field theory, where everything is based on "energetics". This model as its foundations uses a very similar approach to that "single fundamental unit" model, where energy is the unit. If one makes the energy EM in nature, then one has the Russian energetics approach. This unified approach gathers everything in, including all energy actions and relations in inert matter (the first branch of energetics, called by the same name), all field and matter interactions in living matter (the second branch of energetics, called "bioenergetics"), and all mind operations and mind-matter interactions (the third branch of energetics, called "psychoenergetics"). [ed: sometimes also called "psychotronics" by the Russians]

Rick Andersen in an article called "What is Scalar Electromagnetics?" describes the new science this way:

> "Scalar EM is the brainchild of Lt. Col. (retired) Thomas E. Bearden, a systems analyst and wargames specialist who has been advocating a view of electromagnetics which is based on the notion of a vast, unseen background of scalar energies (as opposed to vector energies) which underlie all physical reality.
>
> "If Bearden is correct in his Scalar EM theory, then we can build devices which would enable us to alter gravity, time, inertia, and the apparent mass of an object. This of course has ENORMOUS implications for military applications, space-vehicle drives, time-travel, teleportation, paranormal phenomena, and just about every other area one can think of." http://twm.co.nz/Beard_scalem.html

> "Time-energy, time-currents, and time-structuring play the dominant role in electromagnetics. Time-as-energy eventually becomes engineerable, as easily as is spatial energy now. We are always dealing with spacetime and with spacetime curvature."
>
> Bearden: http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

The scalar energy discovered in the vacuum of space is sometimes called "zero-point" energy. Thomas Valone explains how this term came about.

What does "Zero Point" mean?

"Boyer traces the historical “creation of the vacuum” as proceeding in stages in parallel with the historical development of ideas about the vacuum. To paraphrase, he says that in the 17th century, it was thought that a totally empty volume of space could be created by simply removing all matter and, in particular, all gases. That was our first concept of the vacuum. Just get rid of all the gas. Late in the 19th century, it became apparent that the region still contained thermal radiation. But it seemed that the radiation might be eliminated by cooling. So the second concept of getting a real vacuum is to cool it down to zero temperature. Just go all the way to absolute zero. Then we’ve got a real vacuum. Right? Well, since then, both theory and experiment have shown that there is a non-thermal radiation in the vacuum and that it persists even if the temperature could be lowered to absolute zero. Therefore, it was simply called the “zero point” radiation. "Thomas Valone http://www.seaspower.com/InsideZeroPoint.htm

> "**For our very survival, it is absolutely imperative that informed citizens be aware of this dramatic change, which is just now starting. The powerful new science and engineering must be controlled and used for humanity's benefit, not its detriment. Else it will eventually be let loose unrestrainedly, to destroy all life on earth - a possibility indicated by Nikita Khrushchev in 1960**." - Bearden
>
> http://www.cheniere.org/explore%20articles/priore1/p4.jpg

2. Tapping the waves

Endless free energy, Tesla's radiant energy, the dipole as the gateway to the energy of the void, MEG in one year, Bedini, Yakuza, etc. End of Oil, geopolitical chaos, rogue groups

Move Over G.E. & Big Oil, Energy is Everywhere!

Third-world nations set to rise

> "Certain powerful interests did not wish the extended electromagnetics to be discovered or taught. That, after all, would lead to free energy and loss of economic control of the citizens. That was precisely why Nikola Tesla had already been suppressed ... Today it is still being suppressed by the orthodox establishment."
> http://www.cheniere.org/books/ferdelance/s25.htm

The vast seething ocean of energy of the vacuum, the longitudinal scalar EM waves of the time domain, can now be tapped and "transduced" into ordinary (transverse) EM energy in our 3-Space world. The process can be compared to putting a paddlewheel into a river. The energy acquired is free, since the river is there flowing whether we tap it or not. And it is a mighty river, and is not diminished by our paddlewheel.

This means that a permanent solution to the "energy crisis" is at hand, if only those who hold it so secret would give it up to humankind.

It means our species' insane use of oil is no longer necessary. We can make automobiles that need no fuel at all. We can build motors that power themselves. It means that unlimited electrical power can be available for free anywhere in the world. Every house and even shack can have its own power supply.

Of course there is the cost of the equipment at first, but the energy flow is free for the presumably very long life (no moving parts) of the generator. It will probably take decades to shift over to scalar power.

But the ability to tap the scalar (longitudinal) EM waves of the vacuum (which exists everywhere) also means that unbelievably powerful weapons can be, and already have been developed. In fact these weapons are so dangerous that the fear engendered becomes a psychological obstacle to the grasping of the facts.

I struggled with this myself for months, and even though Tom Bearden had completely convinced me already, it is hard to accept the fact that for everyone on earth the "worst that might happen" has suddenly gotten much worse. It takes time to deal with that, to realize it and get over it.

Indeed, the possibilities of these new weapons, which are tapping the huge energies of time, are so terrifying it is natural for the mind to simply say: "I don't want to hear about that."

And the government, likewise, pretends these weapons do not exist and tries to keep it secret from the people. Because most people would indeed be truly alarmed to know the actual situation. Col. Tom Bearden has decided that the people need to know the truth anyway. As a People on a Planet we must face this now, and ban the use of scalar weapons in war or in mass mind control. The current deployment of these weapons is a world emergency which we must all face.

The physics and mathematics of the longitudinal waves and the time domain are clearly beyond the layman, but I recommend scanning through them anyway just to get a sense of the wonder and the elegance of it all and some of the exotic terms which are used. Two of the most important papers by Col. Bearden are:

Giant Negentropy from the Common Dipole

The Final Secret of Free Energy.

Despite the complexity of the science involved, Bearden sums up the essential secret of it all in a few words.

> "There is no problem at all in extracting all the energy one wishes from the active vacuum, anywhere in the universe, at any time. Just make a dipole.

> The problem is in (1) catching some of that freely gushing EM energy in a circuit containing a load, and (2) dissipating the caught and collected EM energy in that load to power it, without using half the caught energy to destroy the source dipole's).
>
> That is the ONLY real energy problem on the planet, and always has been."
>
> http://www.cheniere.org/correspondence/022502.htm

A dipole could be a battery, which is polarized into negative and positive. Or the terminals of a generator. Or any magnet with its two poles. Or an atom with its positive and negative charges. Or the earth's magnetic field, the sun, and all heavenly bodies. Even a galaxy must be some kind of mega-dipole. Wherever there is a dipole there is already immense scalar EM energy ushering out of and back into the vacuum. And there have already been demonstrated a number of diverse methods for tapping that energy as it flows. (John Bedini, Edwin Gray, Bruce de Palma, and others). Bearden provides what he calls a "partial list" on his website.

> "In other words, nature then gladly gives us as much EM energy flow as we need, indefinitely — just for paying a tiny little bit initially to "make the little dipole." After that, we never have to pay anything again, and nature will happily keep on pouring out that 3-flow of EM energy for us. This is the giant negentropy mechanism I uncovered, performed in the simplest way imaginable: just make an ordinary little dipole."
>
> Bearden http://www.cheniere.org/techpapers/Unnecessary%20Energy%20Crisis.doc

How to Buy a Free Energy Machine

Answer: Just wait one more year! 2003. Then get off the Grid!

Perhaps the most eagerly awaited event now is the marketing of Bearden's Motionless Electromagnetic Generator (MEG) scheduled to begin in about a year's time. This is when the proof will hit the fan. Many people have said to me, when talking about free energy, "I'll believe it when I can go to the store and buy one." Hopefully in a year's time they will be able to order their own MEG. And then they will believe, because it works and is working.

Says Bearden: "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat."

(Bearden : http://www.cheniere.org/correspondence/011202.htm)

Soon the cat will be fully out of the bag. The first MEG units will supply 2.5 kilowatts of free electricity. Forever. Units can be connected together to double or triple the wattage. After that gets going, 10 kilowatt generators are planned. Energy will flow freely from the time domain indefinitely, and there are no moving parts to wear out. This manufacturing project is a heroic effort and will change the world as we know it. For once people are in possession of such devices there can be no further denial from anyone, including the government and/or "secret-government."

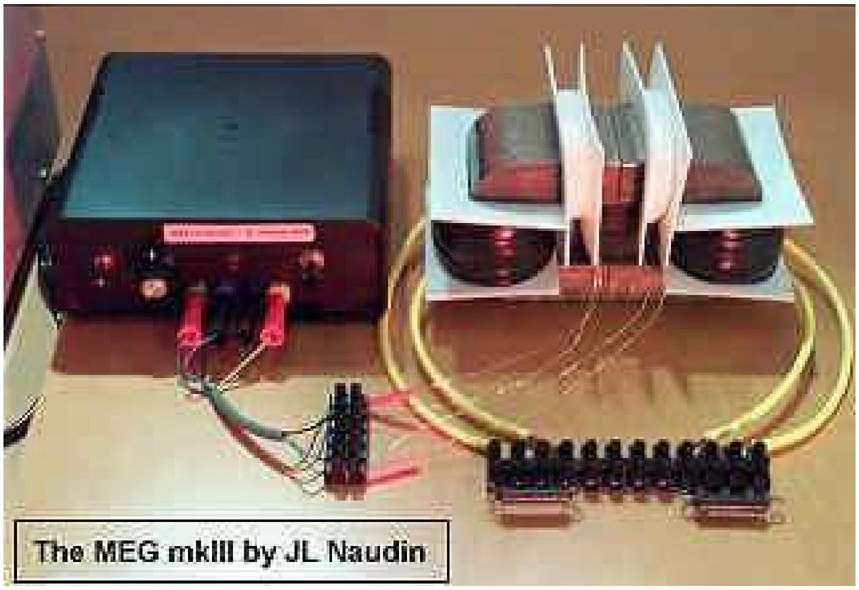

The Motionless Electromagnetic Generator

(Successfully replicated here by J. L. Naudin)

The physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum."

The device was successfully replicated by J.L. Naudin in France, and others, like Steve Utne, are also working on their own models. But Magnetic Energy Limited's MEG is actually scheduled for production soon and looks to be the first commercially available free-energy generator in history!

I imagine it will not take long for tinkerers to put them into their electric cars and make the first fuelless automobiles since Tesla's unheralded (and unheeded) experiments. Or for builders to begin building houses powered by MEG's. Houses which will never see electric bills. How can they ever make enough of these to meet the demand that will suddenly be there? Bearden is right, it needs to be an all-out crash program of the government on the order of the Manhattan Project.

I must include here that Bearden is quick to point out that he is not himself the inventor of the MEG, but is a partner and advisor in bringing it to the public market through Magnetic Energy Limited, Inc. I must also admit that I am praying for the success of this venture, and the ending of the absurd and criminal withholding of this technology from the people by the black-ops military-industrial people. This secrecy has cost the planet over 30 years of unnecessary pollution.

The first units off the assembly line, which is being set up in a "friendly" unnamed nation, will usher the public at last into the new era of scalar electromagnetics.

As of December, 2001, the situation with the MEG was described by Bearden in a letter.

> "What we presently have with the MEG is a successful laboratory experiment. At least a year's very hard research will have to be done before we will be ready to put a commercial power supply into production.
>
> "Consequently, we have made an agreement with a foreign partner (the National Materials Science Lab of the National Academy of Sciences of a friendly foreign nation) to do that year's research. At the same time, we are trying to make an agreement with one or more large financial partners here in the U.S."
>
> Bearden http://www.cheniere.org/correspondence/121901.htm
>
> "So we are hopeful that we will be able to start introducing units on the market about a year from now. Quite simply, we will either succeed or we will fail. But we will give it our very best effort."
>
> Bearden http://www.cheniere.org/correspondence/012202a.htm

Bearden has given another technical explanation of how the MEG works here in a letter.

Motionless Electromagnetic Generator patented, 3-26-2002

Click here to read the patent

"In Magnetic Energy Limited, all business discussions and arrangements are taken care of by Dr. Lee Kenny, Managing Partner. There are five of us who are co-inventors of the MEG, and two of the others are really the principal inventors." Bearden http://www.cheniere.org/megstatus.htm

Free Energy has been Suppressed

Secret Government Aware of Free Energy

> "In this world of contemporary times, all the agencies, CIA, FBI, KGB, NSA, Electric Power Research Institute, DARPA, Brookings Institution, Henry Kissinger, Edward Teller, et al., all are aware of [free-energy researcher Bruce] De Palma and his ideas.
>
> "Because these individuals and institutions are employed by the ruling elite to forecast the future and satisfy present needs and demands, it is clear that free energy is a threat to the world order constructed by business and the mindset of those who want to own the world.
>
> "Free energy represents Man's aspirations and dreams of freedom and equality, uniform division of resources and the ability to choose one's own future.
>
> "The fact that free energy is suppressed speaks to the greed and self-interest of a ruling elite which, even in the face of an emergency of global starvation, resource depletion and environmental pollution, will not give one inch if it means loss of control. This is an

> attitude of paranoid delusion and fantasy which can only arise from the alienation of a class of elitists who, through their money, are completely insulated from the reality of the day to day fight for existence of the common man."
>
> http://depalma.pair.com/Absurdity/Absurdity07/ProblemOfFreeEnergy.html

Bearden describes one case of suppression when the Yakuza killed the Kawai engine in his very presence in Huntsville, Alabama.

> "That night a jet arrived from Los Angeles, with a Yakuza on board. The next morning Kawai no longer controlled his company, his invention, or his own fate. The Kawai party was in fear and trembling. The Yakuza coldly dissolved the agreement, they packed up the two Kawai engines we had, and left. And that was that."

So it is not only governments who are keeping this revolutionary technology from seeing the light of day. The Yakuza? KGB? The Aum Shinrikyo cult? Our "leaders" need to be honest about what is going on, otherwise they look foolish, acting out a kind of "Alice In Wonderland" script before the public, a script that has become completely unbelievable.

Concerning the criminal withholding of this scalar technology by the military black-budget people, and the cover-up of the actual situation regarding scalar weapons throughout the world, Tom has this to say:

> "The high government officials of the United States of America grossly violated our constitution and their sworn oaths of office. Our constitution specifically charges the government of the United States to defend the people of the United States. That means our officials are specifically charged with defending our people. Yet they deliberately agreed not to do so, and they deliberately sacrificed us in advance.
>
> "They committed treason of the highest kind, by deliberately surrendering us to total destruction anytime the big balloon goes up."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Energy Crisis Solved

One of the "must read" documents at Cheniere is "The Unnecessary Energy Crisis: How to Solve It Quickly," especially since the "oil wars," which Bearden has foreseen for decades, have now already begun.

Here are some clips from that Document:

> "To avoid the impending collapse of the world economy and/or the destruction of civilization and the biosphere, we must quickly replace much of the "electrical energy from oil" heart of the crisis at great speed, and simultaneously replace a significant part of the "transportation using oil products" factor also.

> "The technical basis for that solution and part of the prototype technology required, are now at hand. We discuss that solution in this paper."

And although the solution to the energy problem and the oil wars may be at hand, Bearden suggests there is still an urgent rush that needs to be made to save human civilization from collapse.

> "Once the technology hardware solutions are ready for mass production, even with a massive worldwide deployment effort some five years are required to deploy the new systems sufficiently to contain the problem of world economic collapse. This means that, by the end of 2003, those hardware technology solutions must have been completed, and the production replacement power systems must be ready to roll off the assembly lines en masse.
>
> "The 2003 date appears to be the critical "point of no return" for the survival of civilization as we have known it.
>
> "Reaching that point, say, in 2005 will not solve the crisis in time, and the collapse of the world economy as well as the destruction of civilization and the biosphere will still almost certainly occur, even with the solutions in hand."
>
> "The good news is that we now know how to easily initiate continuous and powerful "electromagnetic winds" from the vacuum at will. Once initiated, each free EM energy wind flows continuously so long as the simple initiator is not deliberately destroyed."

On page 9 he explains the concept of the dipole, and how there is already huge scalar energy flowing in the time domain around any dipole. Hold a magnet or a battery in your hand and you hold a stress in time and a beautiful flowing loop of longitudinal waves!

Bearden decries the scientific community in the West for being unable to let go of their dogma long enough to see that the greatest discovery in human history lies at their feet, but for their stubborn ignorance.

> "The present energy crisis then is due totally to that "conspiracy of ignorance" we referred to, that is maintained by the scientific community, and that has been maintained by it for more than 100 years.
>
> "This is the real situation that the environmentalists must become aware of, if they are to see the correct path into which their energies and efforts should be directed — to solve both the energy crisis and the problem of gigantic pollution of the biosphere. "As far as anyone in the legislative branch of government ever understanding the profundity of the discovery and engineering of scalar waves, Tom Bearden says they are at a great disadvantage in doing so.
>
> "The nontechnical person — e.g., a Senator or a Congressperson — is operating under a distinct disadvantage. If he or she takes the stance that much better electrical power systems can readily be achieved, he or she is in fact opposing almost the entire set of University, Government Technical, Power Company, Battery Company, and Organized

> Science communities. Further, in most cases his technical advisors are themselves from one or the other of those communities, and likely to go back into that community or those communities when the Senator or Congressperson leaves office, or even before. So the Congress and the non-technical government community at large operate at a great disadvantage...Bearden National Emergency Declaration is Needed
>
> "The Government Non-Technical community (the Senate and the Congress, in particular) are in far better shape than the Government Technical community, to appreciate the world implications of the pending economic disaster. I am hopeful that both the environmentalists and the Government Non-Technical community will rapidly unite in a common goal to get this vacuum energy program launched, under a National Emergency declaration. If so, then they can solve the energy crisis and the pending economic crisis, in fairly short order, and permanently."
>
> unnecessary_energy_crisis.htm#congress_disadvantage

In the conclusion of his paper The Unnecessary Energy Crisis: How to Solve It Quickly Bearden again states the urgency of replacing our centralized power grids with highly dispersed and local taps on the freely flowing energy of the active vacuum. Given the abilities of the Tesla howitzers, the power grids of all nations are sitting ducks.

> "Personally, the present author regards the increasing energy crisis as the greatest strategic threat to the United States in its entire history. I will do anything within my power to help prevent what I perceive to be the looming economic collapse of the Western world, preceded or accompanied by a sudden, explosive, all-out and continuing exchange of the WMD arsenals of most of the world.
>
> ...In the name of all humanity, let us begin! Else by the time this first decade of the new millennium ends, much of humanity may not remain to see the second decade."

3. Weaponization

Scalar interferometry, Tesla howitzer, Tesla dome, heat/cold, dudding, 10 nations have them, dead-man fusing & Insane Systems, scalar strategy, earthquakes and weather wars, solar storms, mind attacks, leased to Yakuza, many examples of testing, quantum potential weapons.

> "Indeed, Soviet energetics weapons are now capable of destroying both our triad shield, our homeland, our armed forces in the field, and our population, quickly and efficiently. We have a new "gap" of monumental proportions: not a missile gap, not a submarine or bomber gap, and not even a particle beam or laser gap. We have a scalar electromagnetics or electrogravitation gap." Bearden http://www.cheniere.org/books/ferdelance/intro.htm

"**We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that,**

if unrestrainedly used, it could wipe out all life on earth. It is a fantastic weapon." Khrushchev, to the Presidium, Jan. 1960

> "We've totally missed the most incredible "secret superweapons" development program of all times, and the worldwide testing of the weapons themselves.
>
> "Hiroshima and Nagasaki should serve as object lessons in the price of ignorance of such technological breakthroughs.
>
> "History does repeat itself. But this time it's not we Americans who got the great new superweapon first."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Warfare has been changed forever by the development of these scalar energy longitudinal wave howitzers. To get a basic understanding of scalar waves is to have the imagination suddenly run wild as all the implications and possibilities regarding warfare fall into place. One realizes with a certain horror that the world has totally changed, and that there are some very fearsome possibilities. Remember, the power for these weapons comes from the time domain, longitudinal EM waves in the vacuum of empty space, and the power is tremendous and mind-boggling. Being able to blast away at any target from a distant control booth is something that has never happened before. This is incredible power to be in control of and it divides the history of weaponry into "before" and "after." And the destructive power of these weapons is delivered instantaneously to the target from the local vacuum at the place of the target.

From an interview:

> Questioner: "Are you saying that real energy can be transmitted faster than the speed of light?"
>
> Bearden: "Yes. But hyperspatially. Not through 3-space, but "around" it. Again, this means that it's possible to produce energetic changes in a distant system at a distant place, without transmitting energy "through space" in the normal sense. You transmit "around" space, so to speak, and directly in time or through a higher dimension, depending on the model you're using to understand this. Further, the speed of the internal EM energy transmission is not limited to the speed of light."
>
> http://twm.co.nz/beard_interview.htm

Here are some initial considerations about scalar weapons and scalar wars

1. Tesla Howitzer 2. Tesla Domes 3. Rogue groups 4. WW III 5. Russian "UFOs?"

3a. The Tesla Howitzer and its modes of action

Source: Tom Bearden's weapons slides and Fer-de-Lance Briefing

For one thing there is no delivery of any thing to any place. Bombs are obsolete. So are the planes to get them there. No helicopters necessary. All current nuclear devices are rendered nearly useless by electronic "dudding." All distant-destruction can be done from a control room, and to any point on earth. The destructive power of "men" has just increased by orders of magnitude. The new weapons could even cause storms on the sun! Thus there is great urgency to make these new facts public and known, with the hope that all nations would come together to ban the use of scalar weapons.

Two scalar antennae together along with the computers to control them, make up a scalar interferometer, and according to Col. Bearden the Russians have hundreds of such installations already. These interferometers are called "Tesla Howitzers." They can deliver a giant blast of energy to the distant target site, true "action-at-a-distance".

The first howitzer mode is called the "exothermic" mode of operation because immense EM energy blasts outward at the target site. The blast of a scalar howitzer can be of near nuclear level in destructiveness, and can be repeated easily, at that place, or nearby, or anywhere.

The howitzer can use a lesser exothermic power setting and simply destroy all electronics in the target area. Thus they can render our nuclear missiles inoperable as they sit in their silos by "frying" the electronic circuits that guide them. They can bring down any airplane, anywhere in the world, at any time. Any person anywhere, if their exact position is known, can be assassinated without a shot being fired.

The howitzers can also bring down power grids anywhere in the world, at any time. Thus they could bring Wall Street to an instant halt by disabling all its electronics. Even this one little easy tweak of the howitzer could cripple the U.S. economy in a single day. No wonder the "leaders" don't want you to know! This is a new era we are in. The discovery and development of scalar interferometry has put us in a new world, whether we like it or not. Anyone could be expected to go into denial, the befuddled senators and congressmen included. But we have to wake up.

In a second howitzer mode called the "endothermic" mode, the howitzer sucks energy out of the target area, essentially creating a blast of cold at the distant target. It is even capable of freezing parts of the ocean. Bearden gives a number of cases where these "cold explosions" have been witnessed, mostly by airline pilots. April 9, 1984: 1 2 3) A huge mushroom cloud of mostly water is seen rising miles into the sky from out of the ocean, an awesome sight.

In the endothermic mode the sucked-out energy must go somewhere, so it is vented out at some other chosen spot on the earth. These endothermic plumes have been photographed by satellite.

> "The exhaust in the left picture is about 150 miles long, and is nearly horizontal (about 1.5 degrees above the horizontal). It is entirely consistent with the continuous exhaust from a "dumping transfer" scalar EM howitzer in the continuous exothermic mode. (The primary howitzer, of course, was activated in the endothermic mode.)" Bearden More Plumes

By using both of these heat (exothermic) and cold (endothermic) modes together the weather can be altered anywhere. Warm the air over here, cool it down over there, put a curl in the jet stream, dissipate clouds, create clouds, whip up a tornado. Bearden speaks about using the endothermic howitzer mode as a weapon of war:

"Cold explosions can be used to freeze tanks, personnel, and equipment. The equipment and tanks thaw out. The personnel thaw out too, but they are dead when they do." Bearden, Fer-de-Lance

(This endothermic mode of the howitzers makes me think that there might one day be a cure for the problem of global warming after all. Could not a big device, situated in space or on the moon, be able to vent a great deal of excess heat outside the atmosphere into space, thus functioning as a kind of planetary thermostat?)

While the United States has experimented with electomagnetic weapons using ordinary transverse EM waves, the longitudinal [LW] weaponized waves are truly of another order altogether.

For one thing they do not have to travel through space, for they come from the time domain, which is everywhere. The energy of the blast, the heat or light or whatever, comes from the vacuum of space at the location of the target itself. This is astonishing. The weapons actually just trigger the release of immense energy from the vacuum at the target location.

> "... it is possible to focus the potential for the effects of a weapon through spacetime itself, in a manner so that mass and energy do not "travel through space" from the transmitter to the target at all. Instead, ripples and patterns in the fabric of spacetime itself are manipulated to meet and interfere in and at the local spacetime of some distant target. There interference of these ripple patterns creates the desired energetic effect (hence the term energetics) directly in and through the target itself, emerging from the very spacetime (vacuum) in which the target is imbedded at its distant location." Bearden Fer-de-Lance

The big Tesla howitzers are aimed at their targets by using a worldwide electromagnetic pattern called the "Woodpecker Grid," begun by the Russians in 1976.

You can hear the Woodpecker signal here: http://www.cheniere.org/misc/woodpeckersound.htm

The woodpecker grid is just that, a grid of EM waves which provide a channel for the LW waves to any point on earth. Computers use the woodpecker grid information to hone down the aiming of the howitzers to a pinpoint location on earth. (For that matter, it could be under the earth, in space, or under the sea).

A third howitzer mode is what Col. Bearden calls "Mindsnapper" mode. This is a truly frightening mode which affects the electromagnetic mind-body connection. Being mind-snapped at a low level would cause you to lose consciousness. Being mind-snapped at a slightly higher power would "entrain" all minds in the target area into a kind of hypnogogic trance, a state in which they would all be highly susceptible to suggestions and orders.

Being mind-snapped at high level would so disrupt the mind-body connection that everyone in the target area would fall down, instantly dead. Not only that, everything around you and inside you would be killed, so that your sterilized body would not rot for quite a long time.

The Mindsnapper can be used for a small area, or to kill an entire population. In fact, using the scalar waves to affect the human mind has become a new field in its own right, and has been dubbed "psychoenergetics." (This is discussed below in section 5).

Bearden describes the use of the howitzer in mind-snapping mode against people, to cause unconsciousness or death:

> "From a military viewpoint, these scalar EM weapons are really quite all-around weapons.
>
> "For example, they are very lethal against personnel.
>
> "Set in the "high intensity pulse mode," one blast and a person's nervous system is destroyed instantly. Total, instantaneous death results. One can blast away at entire groups of infantrymen, for example. It's the peak power that's important, not the average power. So the weapon doesn't require too much battery power.
>
> "Set on a lower intensity pulse mode, one zap can simply knock out a person, rendering him instantly unconscious. Quite useful if one needs prisoners to interrogate. Also very silent, which is useful in operations employing stealth and surprise." Bearden http://www.cheniere.org/books/ferdelance/s80.htm
>
> "Those hit by the scalar EM weapon, however, have a most peculiar death mode.
>
> "Death comes-instantly and totally. There is no convulsion, no response. The entire nervous system is destroyed instantly. Every living cell in the body is killed instantly, including all bacteria, germs, etc.
>
> "A body hit with this thing falls like a limp rag and lies where it falls. It doesn't decay in even 30-45 days. In a macabre fashion, it's been reduced to something like food irradiated with nuclear radiation; everything is killed, so the material is preserved for an extended period before any decay can set in."
>
> Bearden http://www.cheniere.org/books/ferdelance/s81.htm

He goes on to say that the same "bazooka-sized scalar EM pulse weapon" can also be used to knock out a tank with one shot (or pulse). So tank warfare is obsolete. Helicopters? Bearden points out: "... if he's attacked by a helicopter or a fixed-wing aircraft, he can bring the same weapon to bear on it. He can knock down a chopper or a jet. With one shot."

Obituary for Tank Warfare

Tanks useless against portable LW weapons

> "For example, suppose one attacks a tank. [with an LW weapon] With scalar EM pulse, the personnel die instantly. Total personnel kill is achieved.
>
> All electronic systems of the tank are dudded. Total systems kill is achieved. The ammunition in the tank explodes. Total kill is achieved.
>
> The fuel explodes. This is another total kill mechanism.
>
> That tank has been killed totally, by a variety of mechanisms, all simultaneously.
>
> That's a k-kill in any analyst's book.
>
> And don't worry about retrieval and repair. That tank is finished permanently.
>
> Even the most modern tank is just as vulnerable as the most obsolete."
>
> Bearden http://www.cheniere.org/books/ferdelance/s88.htm

One begins to get the idea how profoundly the conventional means of warfare have been obsoleted. All the usual weapons of war have been rendered nearly useless! And Star Trek's "Phaser," with its "stun" and "kill" modes is here at last.

Although it is the weapons aspect of scalar electromagnetic which has received the most developmental effort, Bearden points out that a great many other amazing things that may be accomplished in this new field of science. Beam me up Scotty!

> "Superluminal communications systems, hyperspace drive, and materialization and dematerialization are all hypothetically possible, using scalar electromagnetics. As the technology develops, we should see the development of many of the systems long thought impossible except in science- fiction."
>
> Bearden http://www.cheniere.org/books/ferdelance/s15.htm

Bearden describes a huge test of the Russian scalar weapons systems which occurred in 1985, a test that went completely unnoticed by western scientists:

> "Just before May 1, 1985 the Soviet Union performed a "full-up" operational test of their entire strategic scalar EM weapons complex. This test was detected and monitored periodically by Frank Golden. Some 27 of these giant Soviet "power taps" were locked into the molten core of the earth, producing forced (entrained) scalar resonance of the entire planet on 54 controlled frequencies under our very feet. The remainder of the scalar frequency spectrum was ablaze with literally hundreds of Soviet scalar transmitters: probably the entire strategic scalar command and control system to

> underwater submarines, higher command centers, distant commands, etc. was activated in the giant exercise.
>
> "For several days, the system was exercised on a mind-boggling scale, apparently as part of the Soviet Union's highly accentuated 40th anniversary celebration of the end of World War II.
>
> "Ironically, not a single U.S. intelligence agency, laboratory, or scientist detected this monstrous exercise for the new Soviet leader, Gorbachev. Not a one of them had a detector for scalar EM radiation, and not a one of them officially believes the exercise ever happened."
>
> Bearden http://www.cheniere.org/books/ferdelance/s38.htm
>
> "The Woodpecker grid/howitzer weapon system can be placed over the ocean and used against cruise missiles, naval surface-to-air and surface-to-surface missiles, submarine-launched ballistic and cruise missiles, etc. Placed over a carrier task force, it can also take care of the aircraft launched by the carrier as fast as they are launched. It can also handily take care of the missiles launched by guided-missile cruisers of the accompanying task force."
>
> (Bearden http://www.cheniere.org/books/ferdelance/s65.htm)

The whole series of the Fer-de-Lance briefing slides and comments is a "must read" for anyone wanting to be aware of the actual state of things in this world right now. It is a new world and it takes some getting used to. Fer-de-Lance speaks clearly, for the layman, of the way things stand.

What is it like to operate these Tesla Howitzers? Col. Bearden gives us an idea in his "Fer-de-Lance" briefing, Slides 66-71.

Who needs armies anymore? A entire war can be fought and won by a handful of men sitting in a comfortable control booth! The scalar electromagnetics revolution has made all previous forms of warfare nearly obsolete.

Why haven't the Russians used the scalar weapons yet?

Bearden gives his thoughts on this in "Fer-de-Lance." The power of these weapons is so great that an accident or war could wipe out all life on earth, and even affect the sun.

> "As can be appreciated, this entire concept [MAD] is now totally obsoleted by the Soviet possession of scalar EM weapons.
>
> 'The "mutual assured destruction" capability -- or MAD concept as it was so aptly named! -- is no longer mutual. And for two reasons: (1) the Soviet surprise attack could destroy essentially all our Triad forces wherever they are, in the first attack. (2) Even if a very small number of our strategic forces survive, the puny strike we could attempt to launch could be repeatedly decimated from launch to terminal phase.

> 'The restraint on the Soviets is not because of our Triad forces. The restraint is twofold: (1) a second nation has scalar EM weapons, and would undoubtedly unleash them upon the Soviet Union with great violence if such a massive world-wide destruction of U.S. forces occurred or was threatened (is it really "accidental" that six major Soviet missile ammunition storage sites mysteriously exploded within seven months?); (2) the closed-loop scalar radiation exchange system of Earth-Sun could be inadvertently "tweaked" in the feedback loop from Earth to Sun, so that a large solar scalar resonance was stimulated. In that case the Sun could emit a mighty resonant "burp" that would simply wipe out all life on Earth in a fiery shower.
>
> 'The "terror" in the "balance of terror" between nations has just increased by many, many orders of magnitude.
>
> 'The biblical prophecy that the Earth will be wiped out by "fire and brimstone from heaven" is very near fulfillment now.
>
> 'Put another way, the pucker factor -- even on the Russians who may contemplate unleashing Armageddon -- is higher than any Western strategic analyst has ever imagined, even in his wildest nightmares.
>
> 'Truly these weapons are "more frightening than the mind of man has imagined," as Brezhnev put it in 1975.
>
> 'The frightening scalar EM weapons can be used, but only very, very gingerly indeed. If a slight mistake is made, everybody loses everything."
>
> Bearden http://www.cheniere.org/books/ferdelance/s74.htm

But the Russians (KGB) are not the worst of it. Bearden claims that some of the scalar weapons have been leased to the Yakuza for about a billion a year, and that only the KGB is holding the Yakuza in check from making a full-scale all out scalar attack on the west.

U.S. Defense Secretary William Cohen

Warns about eco-terrorism using scalar electromagnetic weapons

> "Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our [counter terrorism] efforts." - Defense Secretary William Cohen, 1997
>
> In a letter to a writer named "Russell" Bearden says,
>
> "In short, Russell, the Secretary of Defense of the United States confirmed that there are indeed novel kinds of EM weapons, right now and have been for some time, which have been and are being used to (1) initiate earthquakes, (2) engineer the weather and climate,

> and (3) initiate the eruption of volcanoes. We wrote about those exact uses of the weaponry decades ago. Several nations now have such weapons. Three of them (two on one side and the other on a hostile side) are even firing practice shots into Western Australia, as a convenient test range."
>
> http://www.earthchangestv.com/ufo/0209gandor.htm

Tesla Howitzers in Weather mode

How to Make Weather

> "Engineering the weather is duck soup; they tested that over the U.S. in 1967, and entered upon continuing operations over North America on July 4, 1976 as a grim kind of KGB "Bicentennial Gift" to the United States. Here's how it is done with several interferometers.
>
> "First, the interferometers can deliberately make "high pressure areas" (cool the air so it shrinks and its "footprint" pressure on the ground increases because its density increases) and "low pressure areas" (heat the air so that expands and its "footprint" pressure on the ground decreases because its density decreases). Well, if one makes the highs and lows where one wishes them, and judiciously and somewhat slowly moves them along a given path, these highs and lows will entrain the jet streams and thus "steer" the weather.
>
> "So if you wish a very cold snap or attack, go up into Canada and start these actions to steer down some streams that bring some very frigid air. Establish other highs and lows judiciously to "block" or "slow" other jet streams and flows as desired. In this way, one can pull the "large cold air masses coming down from Canada" deep into the southern U.S. at will. They do it regularly. If you wish an ice storm, add an additional current of warm moist air you bring up (steer up by using artificial highs and lows) from the ocean. Where they meet, you will get freezing rain, then sleet, then severe icing accumulating. This type of attack does rather large damage to the struck area."
>
> http://www.earthchangestv.com/ufo/0209gandor.htm

Whole websites have now sprung up showing weather radar anomalies which may well be indications of scalar weather engineering. One list of over 500 images is here. Cheniere has its own collection here.

In another section at Cheniere are many photos of cloud anomalies which reveal the underlying grid along which the clouds are forming or un-forming. Bearden relates a startling incident in these clandestine ongoing "weather wars."

> "During the spring of 1986, abnormally strong Soviet weather engineering occurred over the U. S., causing a drastic drought in the southeastern U.S. This drought was broken by a colleague who used an extremely powerful scalar EM device to redirect jetstreams. A most unusual and unique signature of the "blocking" against the Soviet scalar EM actions

resulted: Two huge circulations developed in the atmosphere, clearly showed as two adjacent giant "holes" in the swirling cloud cover over the middle and eastern U.S.

"Between these two giant holes, the cloud circulations formed a stream of clouds, moving to the south, looking very: much like a giant vertical "bar" of a huge "Y-shaped" cloud flow.* Several national weathermen commented on this unusual pattern, which had not been observed before. The pattern continued, day after day, as the blocking continued."

* On the weather maps.

http://www.cheniere.org/books/analysis/history.htm

Tesla Howitzers in Earthquake mode

How to Make an Earthquake

"Here's how you initiate a very large earthquake with such weapons. Take a convenient fault zone of set of them. Focus the interferometry on the fault zone, in the "diverging" mode, and deposit EM energy there in the rocks on both sides, increasing (slowly) the stress in the rocks by the reverse piezoelectric effect (deposit excess energy, get crystal mechanical movements).

"Do it slowly, and the stress will build up to large pressures well-above a plate slip minimum energy required. At some point, the rocks yield and one or both sides "slip" and move rather sharply, giving a very large earthquake in that zone.

"Do the same thing down in the earth (remember, LWs easily penetrate right through the earth and ocean at will, and so the "interference zone" focus can be inside the earth or beneath the ocean, at will.

"Anyway, focus this thing down to where the active part of the volcano is still slumbering, down where the hole in the plate has been made. Keep increasing the deposition of energy in the magma itself, and eventually the increasing pressure from deep within that volcano, underground, will cause an eruption. Build the energy slow, and the eruption will likely be much larger."

http://www.earthchangestv.com/ufo/0209gandor.htm

Yugoslav Earthquake, Followed by Cold Explosion

Strange Characteristics of Iran Quake, Sep. 1978

Tesla Howitzers in Tesla's "Big Eye" mode

How to make a Big Eye in the Sky

By using longitudinal wave interferometry in a weakly endothermic mode a new kind of "radar" is now possible, one which can look right through the earth and oceans. Because the endothermic mode extracts energy from the distant target, a pattern of the target area can be scanned for in the returned energy.

> "Thus the receiver produces a representation of the energy extracted from various locations within the distant endothermic zone. By displaying the received signals on an appropriately scanned screen, a representation of the distant scene in the intersection zone can be created. Indeed this is a special kind of "microwave interferometry," and -- with modern techniques -- the imagery obtained might be surprisingly good. With development, it might even become as good as the image presently obtained by side-looking acquisition radars.
>
> "Interestingly enough, since scalar EM beams will easily penetrate the earth or the ocean, one can also look beneath the earth or beneath the ocean with this type of scanning scalar interferometer.
>
> "The importance of this capability to strategic and tactical reconnaissance is obvious. Camouflage, cover, and concealment have no effect on such a system. One can easily look inside buildings and into underground facilities. With a small system such as this, the U.S. Marines at Khe Sahn would have had little difficulty locating the tunnels continually dug under the perimeter by the Viet Cong. And targets under jungle canopies are directly visible."
>
> Bearden http://www.cheniere.org/books/ferdelance/s42.htm

Tesla Howitzers can destroy all life on earth

How to Make a Storm on the Sun!!

(Just make a big mistake with your Tesla howitzers)

The deepest dangers of the use of scalar weapons are frighteningly pointed out by Bearden here where he considers that the sun, earth and moon (actually the whole solar system) is a delicately balanced arrangement of scalar (longitudinal) EM radiation. There is a scalar connection between the earth and the sun, which are both giant dipoles gushing energy in the time domain. Big scalar events on the earth could alter this balance causing true catastrophe.

"Indeed, a solar response could be stimulated so that the Sun would violently belch and destroy our biosphere, among other effects."

Unrestrained scalar warfare could cause huge solar storms? To realize that human activity could actually affect the sun itself is truly mind boggling.

Bearden goes on to say:

> "Accordingly, use of huge scalar EM weapons is a double-edged sword. Unless carefully employed, use of the weapons could cause a terrible backlash to the user as well as the victim, and even accidentally cause the destruction of the earth itself. It is not accidental

that in 1960 Khrushchev stated that his new fantastic weapon could -- if unrestrainedly used -- destroy all life on earth.

"This appalling backlash potential is apparently what prompted Brezhnev in 1975 to make so strong an effort to get agreement to outlaw the development of "new weapons of mass destruction" more frightful than the mind of man had even dreamed of. Gromyko even introduced such a draft agreement into the proceedings of the United Nations. The first article of the draft agreement indicated that the nature of the weapons referred to would be negotiated. Sadly, the entire Western world did not even know what the Russians were talking about."

http://www.cheniere.org/books/ferdelance/s24.htm

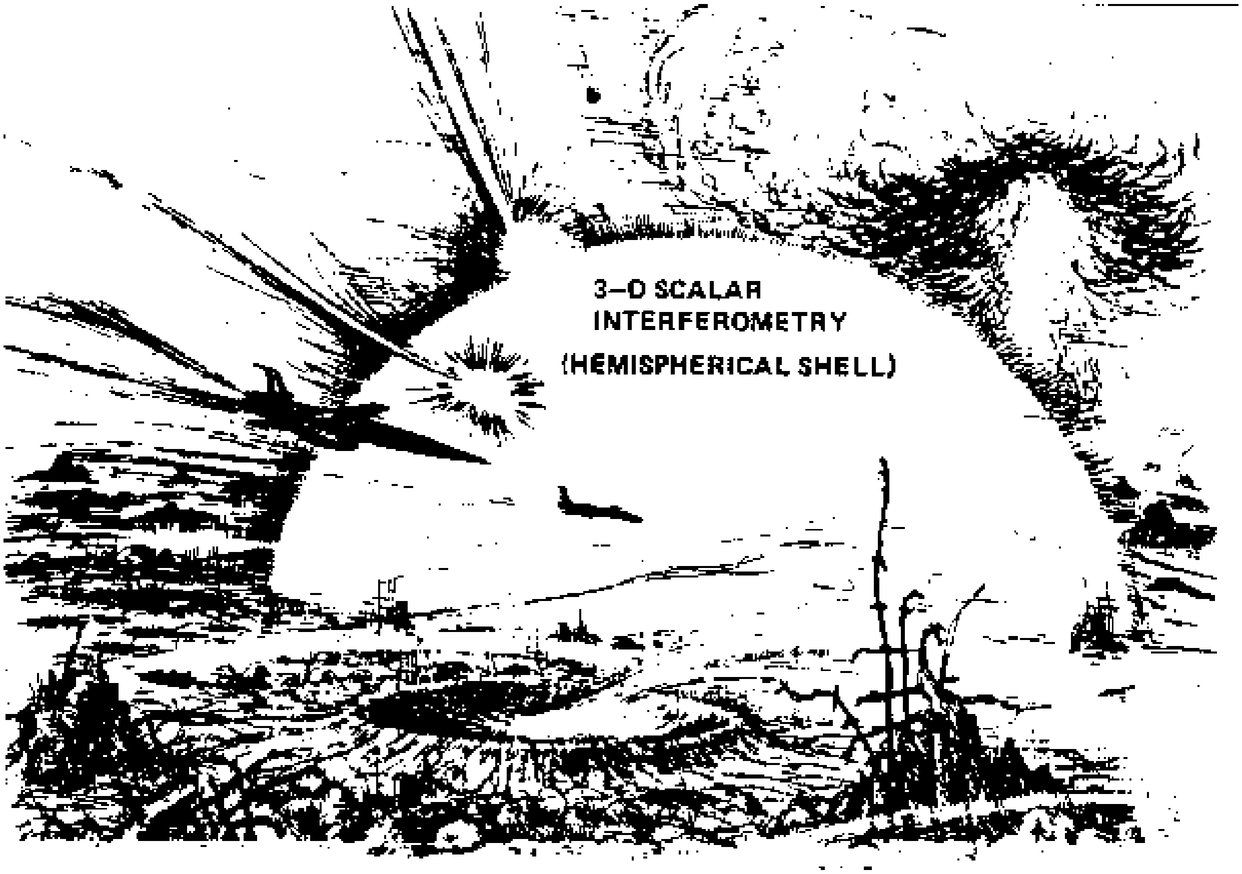

3b. Tesla domes provide near-perfect defense shield.

(Slide from Cheniere.org)

The Tesla howitzers can be used in a nearly impregnable defensive mode whereby they throw up a dome (**Tesla Dome**) or a sphere (**Tesla Globe**) of highly powerful electromagnetic energy, enough to "dud" or destroy missiles which try to penetrate them. Once again, this energy is not going through space from the howitzer, but being made to emerge from the local vacuum at the location of the shell. KGB tests of these domes have been witnessed by airline pilots around the world. Bearden gives many examples of the Tesla domes being sighted around the world in his briefing paper "Fer-de-Lance." (Russian dome test, another globe incident, yet another globe test). These giant electromagnetic domes can be hundreds of miles across, or narrowed down more to total impermeability.

> "Such a shell may be several hundred miles in diameter at the base. The enormous energy required to form such a defense shell is obtained by a "scalar power tap" into the molten core of the earth itself, as previously explained. In late April/early May of 1985, 27 such "power taps" were placed in the earth by the Soviets. If each tap is capable of powering four to six large scalar EM weapons, then the Soviet strategic scalar EM arsenal contains over 100 monstrous superweapons capable of generating exothermic explosions, endothermic explosions, engineering the weather, locating and destroying underwater submarines, detecting and destroying ballistic missiles shortly after launch, detecting and destroying long range strategic bombers as soon as they are airborne, etc."
>
> http://www.cheniere.org/books/ferdelance/s52.htm

By using nested domes one can protect the domed area even against nuclear radiation itself. Tesla globes can be used to hit airborne targets by simply placing a globe of any chosen size in the flight path of the incoming missile. It does double duty because the missile hits it going in, and then the rubble hits it again going out the other side.

While the U.S. persists in its useless boondoggle "missile defense system" by shooting a missile with a missile (missiles which would already have been "dudded" by the Tesla howitzers), between the domes and the globes the Russians (KGB) have in hand (operational) a nearly perfect missile defense system. Someone should ask the Secretary of Defense if the U.S. Tesla domes are in place. And over which cities? How powerful are our American Tesla domes?

Bearden cites a number of examples of these domes being tested around the world:

March 20, 1969 | March 24, 1977 | August 17, 1980 | June 17, 1966 | June 18, 1982 | June 22, 1976

In a slide from the Fer-de-Lance briefing it is shown how the "woodpecker grid" itself also acts as a sensor, detecting any missile launch anywhere, and how the howitzers can respond instantly by delivering a blast to exactly the right coordinates, right along the channels of that same grid which is feeding the location information about the missile launch. This certainly beats "trying to hit a bullet with a bullet." Especially since YOUR bullet would probably be instantly inoperable in a real scalar war situation.

3c. Rogue Groups in the government

Bearden fears that these scalar electromagnetic secrets may have fallen into the hands of what he called secret "rogue groups" within the labyrinthine depths of the black-ops special-access domain of the secret government. "In the extreme, such a deep black program can even become a "captured" program," he says, "which is totally in the hands of rogues and no longer reports to or is bound by the dictates of the parent government."

> "They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification. ...
>
> "In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old 'Power corrupts, and absolute power corrupts absolutely' routine. ... Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control."
>
> "The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. "
>
> http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

3d. World War III has already begun

Because it is now possible to cripple the power grids and electronics of any nation, that nation could be reduced to economic ruin without ever firing an actual "shot." Without electricity Wall Street cannot operate. Without electricity you quickly need martial law. One no longer needs to wage a conventional war anymore at all. There are now lots of new ways to bring any nation to its knees.

> "We are at war, whether we like it or not. And we are going to have strategic strikes on an unparalleled scale for us. Pearl Harbor was nothing compared to what is coming down, once we consider the "augmentation" and such.
>
> "When that factor is considered, then it's a totally new ball game on replacing the power system. There isn't going to be much choice; the old one is going to be destroyed or so badly crippled that it will collapse the U.S. economy in a way we have not seen since 1929. Of course, in such a case hitting us, we will be under martial law, the works, everywhere. For one thing, it's the only way you can keep the cities from imploding from within, by the criminal element, looting, robbing, killing, etc. Simply wait and see; it's on the way."

> Although Bearden has for years decried the "scalar-gap," and chided the military for falling so far behind the Russian KGB in the field of scalar electromagnetics, recent statements indicate that things may at last have changed somewhat.
>
> "In the new war, the first phase of WW III has already been completed. Hence President Bush's recognition that this is a total war, and has to be taken to the enemy no matter where he may be hiding or in what sanctuary or in what foreign nation. This is an entirely different war than anything the U.S. has ever fought in its history. And before it is over, it is going to kill more Americans than have died in all the previous wars in our history. The news media and the public have not yet recognized that awful situation and predicament. Most are already thinking that, well, we won in Afghanistan so it's just about over. Instead, it has only just begun.
>
> "Anyway, that's a brief uptake on some of the weapon situation today. Some 10 nations or more now have LWIs, and five have the QPs. I'm still working on how many have the appalling negative EMP weapons, but would estimate 6 or 7 nations have them in one stage or the other, either deployed or in at least advanced development. And in my estimation, the Yakuza will have them within three to five years if they have to build them themselves, and quicker if they can buy them from the Russians, Chinese, or???"
>
> Bearden http://www.cheniere.org/correspondence/030202a.htm

When he was asked the question '**Does the government listen to Tom Bearden?**" Bearden said:

> "Let me speak carefully. I am very satisfied with how I've been listened to in the quarters where the rubber meets the road. The West is no longer defenseless, not by any means. Cannot amplify it further than that."
>
> http://www.cheniere.org/correspondence/030202a.htm

Bearden has briefed the government a number of times and the Cheniere website contains a great number of his graphic military briefing slides.

3e. Are the Russians making "UFOs?"

Scalar electromagnetics engineering opens the possibility of antigravity vehicles and "hyperspace travel." In other words we can now make our own UFO's! Of such scalar-powered antigravity vehicles, which have now become possible, Bearden has this to say:

> "We might expect to see such vehicles glowing. Their various surface features and mechanisms might appear to be glowing or revolving lights, etc.
>
> "They could exhibit incredible "aerodynamic performance," seemingly in the atmosphere. Actually they would not be moving "through" the atmosphere at all, but through a higher space outside each particle of atmosphere.

> "They could seem to materialize and dematerialize.
>
> "They could seem to plunge into the ocean or rise out of it.
>
> "They could even seem to operate under the ocean or inside the earth itself.
>
> "Such anomalous vehicle performances have been seen all over the world, particularly since a few years after WWII.
>
> "It appears that one or more nations of Earth are operating such vehicles now in great secrecy. Our own "government" may even be one of these.
>
> "If so, it isn't our true "government/government." Instead, it's the "control group/government." It's government at the operational level, but at higher level it belongs to certain control groups who have penetrated our government's vitals and taken over all such projects."
>
> Bearden http://www.cheniere.org/books/ferdelance/slide32.htm

In an excellent article on her Earthfiles website, Linda Moulton Howe reports on the utter bafflement of Norwegian scientists over mysterious lights appearing in the valley of Hessdalen, Norway.

She writes of the research there: "The results can be broken down into two groups: 95% are thermal plasmas and 5% are unidentified solid objects. The plasmas emit long wave radio frequencies and strangely, their temperatures do not vary with change in size or brightness." She quotes the scientists' research summary:

> "1) most of the luminous phenomenon is a thermal plasma;
>
> 2) the light-balls are not single objects but are constituted of many small components which are vibrating around a common barycenter;
>
> 3) the light-balls are able to eject smaller light-balls;
>
> 4) the light-balls change shape all the time;
>
> 5) the luminosity increase of the light balls is due to the increase of the radiating area. But the cause, and the physical mechanism with which radiation is emitted, is currently unknown."

This would almost certainly seem to be another example of the testing of the scalar howitzers. It is baffling because, as Bearden points out repeatedly, Western scientists have limited knowledge of the scalar electromagnetic principles by which the howitzers operate. The Hessdalen scientists were completely baffled as to where the energy of these plasma balls was coming from.

Ms. Howe interviews Massimo Teodorani, Ph.D., Astrophysicist, who released the report on the 2001 Hessdalen sightings. The scientist speaks of the baffling phenomenon observed.

> "I don't know how it is possible that Nature is spontaneously able to do that. Anyway, we deduce that the plasma is trapped inside a sort of magnetic cage and the magnetic cage closes around the plasma and keeps it fixed in some way, prevents it from expanding. But where does it come from? We don't know."

By what we have been considering here, we can recognize the signatures of scalar electromagnetic engineering. If it is a Tesla howitzer, then the energy for the plasma balls is coming from the vacuum of spacetime at the very location of the balls themselves, triggered by scalar interferometers aimed through the woodpecker grid.

These kind of balls can be used as marker beacons giving feedback for precision aiming of the howitzers. The energy of the marker beacon can be read back into the computers giving precise location information for pinpoint aiming. The target area can be very small, or widened out.

There is a little movie of one of the Hessdalen plasma balls here. It would seem to be a clear documentation of a scalar howitzer marker-beacon. How many other current "UFO" sightings are actually the witnessing of Tesla howitzer tests?

In fact there are a great many anomalous events of the past few decades which might very well be explained by the Russian testing of their longitudinal electromagnetic weapons. It would seem that the mysterious appearance of crop circles around the world could be accomplished by feeding a precise mathematical graphic pattern into the computerized aiming software and change the very molecular structure of the crops themselves along the lines of that pattern. The

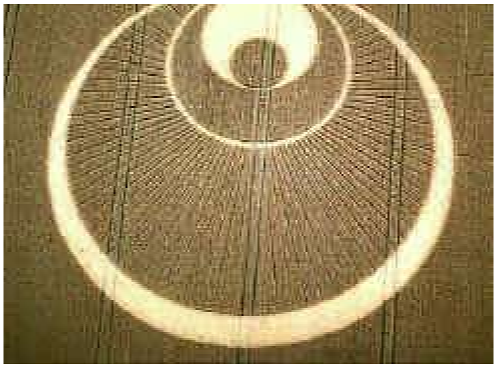

stalks fall over from the localized effect of the longitudinal waves. A plasma ball tracing patterns? Are they slyly sending a message to see if we are "getting it" yet?

The "Cambridge Angel" - Made by Russian scalar plasma balls?

Also see: Interview with Tom Bearden

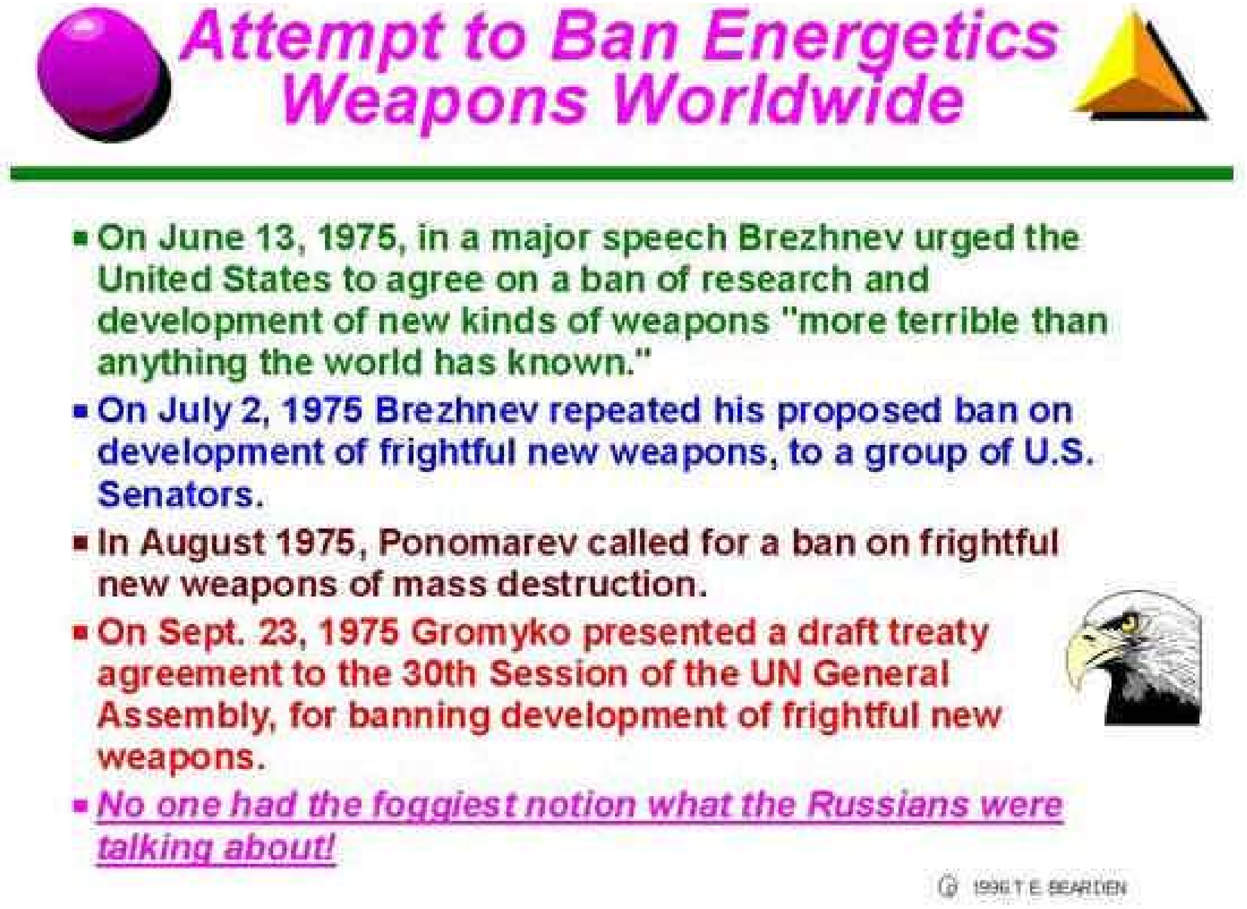

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T E BEARDEN

(Slide from Cheniere.org)

4. Healing powers of longitudinal waves

Healing via longitudinal waves, time reversal of cells (de-differentiation), Priore's machine, Bearden's proposed "blanket" version, using the electric power grid for healing waves or for broadcasting disease patterns. Scalar Biowar.

The good news!

> "Any sort of disease whatsoever - physical or mental - will be rapidly curable, simply and cheaply. It will be possible to reverse aging and rejuvenate the person. It will be possible to regrow limbs and straighten misshapen spines - and directly remove the causes of mental diseases and cure them as well. Lifespan will be staggeringly increased, without 'old age's debilitation.' AIDS, cancer, leukemia, and genetic diseases will be completely conquered. The potential for a true golden era is upon us, for all humanity."
>
> Bearden- http://twm.co.nz/beard_interview.htm

Proof of the healing powers of longitudinal waves from the vacuum is seen in the work of Antoine Priore who produced longitudinal scalar waves using a gigantic glass tube filled with plasma. He himself did not understand why it worked or how it cured cancer in animals.

> "The Priore team cured terminal tumors, trypanosomias, and other dread diseases in laboratory animals in France, in the 1960s and 1970s, under rigorous scientific protocols by eminent scientists. The Priore project was funded by the French Government to the tune of several millions of dollars (equivalent).
> http://www.explorepub.com/articles/energetics.html
>
> ". . . Prioré's machines concretely demonstrated a nearly 100% cure of all kinds of terminal cancers and leukemias, in thousands of rigorous laboratory tests with animals. These results were shown to medical scientists as early as 1960." http://www.cheniere.org/priore/index.html

Bearden has stated that the longitudinal EM waves can have a healing effect because they always come in pairs called "phase conjugate pairs" and one of the pair is time-reversed. This enables engineering of the waves to "pump" the cells of the body (with waves from the time domain) and actually time-reverse the cells back to a previous healthy condition. The scalar cure of cancer does not involve killing the cancer cells but rather pumping them with time energy into a previous condition. I know, this seems miraculous, and is it not?

Priore and his giant plasma tube

This healing power in the longitudinal waves of the time domain is a mind-boggling thing to face. It took me months to get through it, and to suffer the fact that the cure for cancer and AIDS was right there, in this new scalar electromagnetics technology, and that it has been kept back from the people for so many decades now.

Bearden has a great deal to say about the work of Priore, who himself did not understand why his machine worked. Priore did not know about the longitudinal scalar waves of the time domain. In fact, Tom Bearden has supplied, at last, the missing explanation as to how that big giant two-story glass tube consistently cured cancer in animals.

"**The medical implications are enormous**. By just reversing the damaged cells back to a previous physical state, this gives a physics mechanism for controlled cellular dedifferentiation, in biology terms. Physicians will time-reverse diseased, damaged or aged cells back to an earlier healthy condition, including all the cellular genetics. Normal cells are just reversed to a slightly younger condition."

> "We believe this mechanism explains the revolutionary but baffling electromagnetic cures achieved by a team of outstanding French scientists in the 1960s and early 70s, working with inventor Antoine Prioré. In lab animals, cures were achieved for terminal tumors, infectious diseases such as trypanosomiasis, and atheriosclerosis. In addition, deliberately suppressed immune systems were dramatically restored.
>
> "We hypothesize that this is the long-sought mechanism of how the living body heals itself. In short, the body's cellular control system induces cellular reversal by weak time-domain pumping of the damaged cells and the affected area. The cellular regeneration system performs all healing of cellular damage in the body; the immune system heals nothing, not even its own damaged cells) .
>
> "For the first time medical science can develop as a science of healing rather than a science of intervention. And it will develop as a science of unlimited healing, since no pathogen can resist action by curved spacetime engines." Bearden
>
> http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

See Priore's machine See Slides of how it works.

Bearden has extended the Priore work, with the previously missing knowledge now in place, and designed a revised system (for irradiating the body with longitudinal waves) which gets rid of Priore's giant glass plasma tube and uses instead an antenna unit which surrounds the body. He has also provided the missing explanation of how it works. There is an article from "Explore" magazine, Part I and Part II.

In fact, Bearden has presented to the U.S. government a plan to use a portable version of a Scalar Healing Blanket where wire channels in the blanket act as a scalar antenna, which can both send and receive scalar electromagnetic waves. A two minute treatment of the waves is sufficient to restore the body to its earlier healthy condition. Thus even thousands could be treated easily and quickly in a biowar mass attack situation.

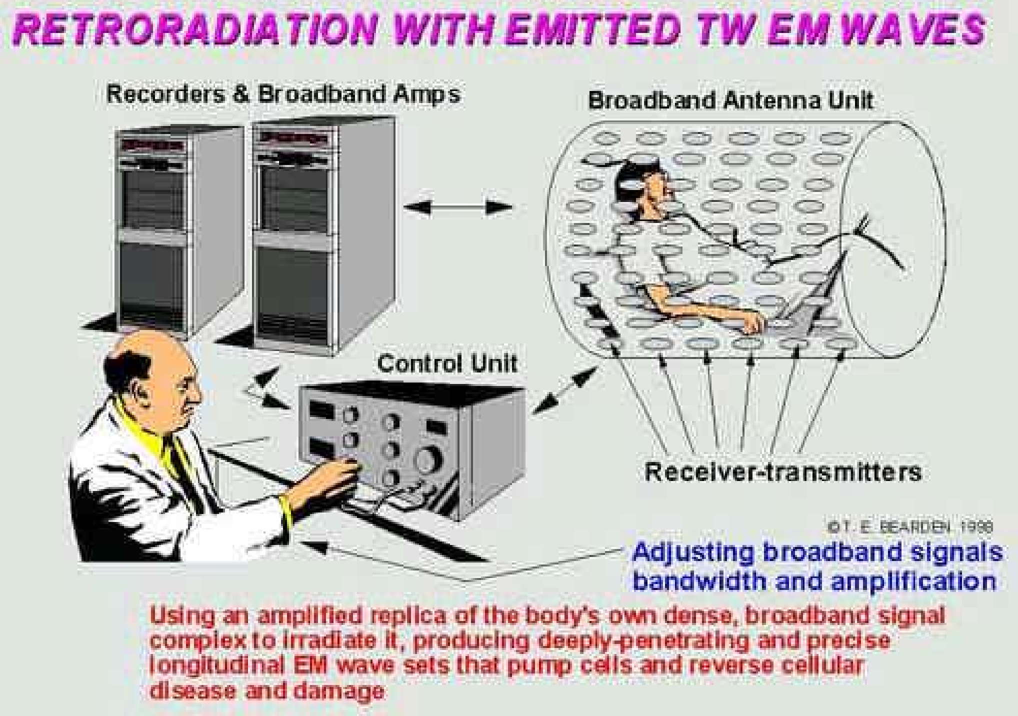
RETRORADIATION WITH EMITTED TW EM WAVES
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
© T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification
Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

"To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S. Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines could be developed cheaply and quickly, and present government facilities could be used to massively attack the correlate database production problem. "The portable treatment units could be cranked out by the hundreds of thousands and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals,

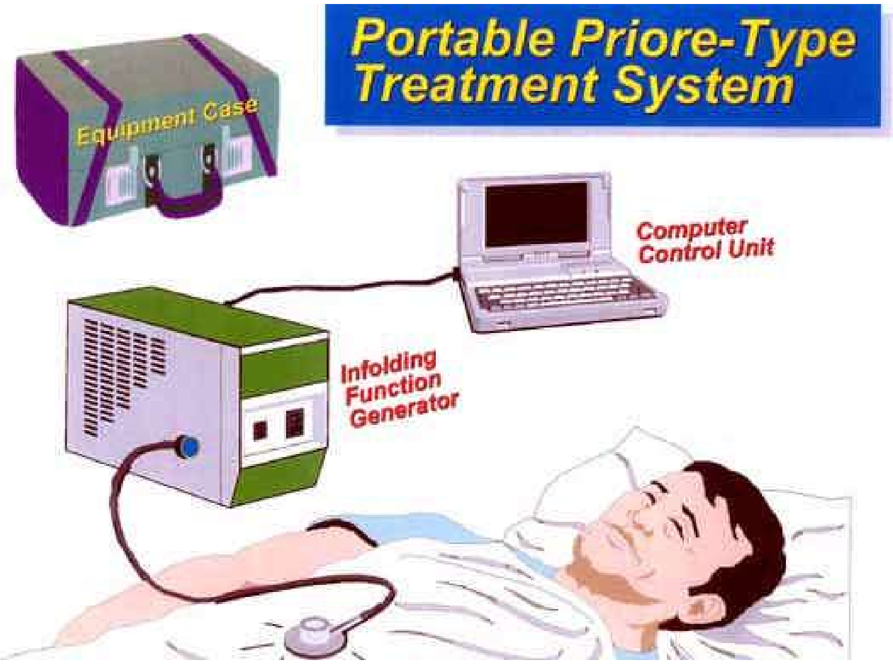
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator

> emergency response teams, etc."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p04.htm

The Scalar Healing Blanket would fit into large suitcase. It has three basic parts, the longitudinal wave generator, a laptop computer, and the antenna-blanket. The plan was a stroke of Bearden's genius, but unfortunately it fell on befuddled government ears. "Unfortunately, so far the DOD [Dept. of Defense] doesn't even understand what the heck we're talking about!"

Article "**Healing with Energy**": http://www.explorepub.com/articles/energetics.html

> "We tried very hard in 1998 to call attention to the extreme national need for the most urgent possible development of a portable computer-controlled treatment device based on the experimentally proven Prioré methodology, but extended to a much simpler and far more rapid method of treatment.
>
> "Quick development of this technology could lead to a suitcase size device, capable of treating a patient in less than one minute, and stopping the symptoms and disease progress cold, then reversing it. Three "less than a minute" treatments one week apart would be required for complete cure. As an added advantage, partial reversal of aging in older patients would also be accomplished as a bonus, in addition to time-reversing the damaged and diseased cells...
>
> "The method proposed to the DoD in 1998 used normal EM in a peculiar way to force the body itself to make its own time-polarized EM waves, and pump every part of its cells -- including the genetics -- in the time domain.
>
> "Quite simply, we are going to be struck this way (QP spreading the immune systems of the populace plus professional smallpox, anthrax, etc. BW attacks on our population centers) and by other means we've spoken of in other papers and briefings."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

The Dark Side of the Force:
Creating and broadcasting disease

Speaking of the scalar wave induction of disease by longitudinal wave patterns Bearden refers to work by French scientist Kervran Russian scientist Kaznacheyev, who proved that any disease could be transmitted electromagnetically.

> "It works this way: the Kaznacheyev experiments (and others) demonstrated that any cellular disease or disorder can be initiated into cells at a distance, by electromagnetic means (albeit unusual EM in nature). The decades of microwave radiation of the U.S. Embassy in Moscow, was an ongoing set of "stimuli" to ping the U.S. system and, by its response, ascertain how far along (or if it was "along") in its knowledge of such weapons. The radiation was responsible for many health changes in personnel and for the eventual deaths of three U.S. Ambassadors."

> Bearden http://www.cheniere.org/correspondence/021702a.htm

In a letter he writes,

> "Yes, the Kaznacheyev experiments were quite real; 15,000 or so in military research institutes in Siberia. As with so many other novel areas investigated and developed by the Soviets, most U.S. investigators just assumed normal EM, which is not what the Kaznacheyev experiments involved, and not what was carrying disease induction patterns from one cell culture to another in rigorous lab tests."

How to "Broadcast" Diseases

Longitudinal EM Biowars

Bearden explains how the new quantum potential weapons can be used to induce disease-at-a-distance in a population, or to "spread" the immune system so thin that a conventional bio-attack would be greatly potentiated. By "broadcasting" virtual disease patterns over a population, the immune system is so overloaded by the presence of the "shadow" diseases that its effectiveness is thinned out.

> "In short, alter the internal wavestructure, and one creates a curved spacetime 'engine' that acts on mass in any fashion one chooses to design -- including initiate diseases . . . Now visualize one of these 'vacuum engines' or 'spacetime curvature engines' that acts on mass to generate the exact effects produced by anthrax. (Or any other disease one wishes) . . . Call the virtual state of a disease engine the disease pattern in the "shadow" state, just prior to becoming observable. . . .
>
> "And the master cellular control system responds to "coming events that cast their shadows before". In short, it responds to the upper level of the "shadow state" of a disease, still in the virtual state! (Hey, virtual particles are real, virtual photons are real, virtual ST curvatures are real --- just very fleeting, only for another to immediately arise). The exchange of virtual particles is known to generate all forces in physics anyway! . . .
>
> "So the cellular control system responds to "shadow state" disease patterns. Enter a QP weapon application. Place a quantum potential weapon "area" on the United States, so that the entire populace is in it. The slowly introduce and bring up in the shadow state, the necrotizing fasciitis disease engine. At some point, the cellular control systems will react, and order the immune system into action. In other words, immune system resources will be committed against this "immediately coming" enemy. Bring it on up a little, but still just in the shadow state. Voila! Humans form a bell-shaped distribution curve. And even if the structure of the QP fluctuates a little, fluctuating a bit stronger and a bit weaker, etc. A few of those targeted bodies will now actually develop necrotizing fasciitis. So far, just like the U.S. Embassy targeting, except a different "carrier".
>
> "But now there is an exact signature that this is what is doing it, and not normal vector carriers of disease and contact or exposure infection. The few cases of necrotizing fasciitis that break out will be most puzzling. They will be randomly scattered across the

> entire nation, and so few that the vector carrier and contact or exposure explanation completely fails."
>
> "And that one happened, right here in the U.S., not so long ago. Check it out.
>
> "So carefully adjust the patterns for, say, 12 major terrible diseases, to that "adjusted and desired shadow state level". Now add all 12 patterns into the internal structure of the quantum potential. Now it has 12 shadow state disease engines in it.
>
> "Place that one on a populace. The cellular control systems order the immune systems into action, and they now have to commit their finite resources to 12 different areas. This greatly thins the response an immune system can make to any one of those 12 disease patterns! In short, it "spreads" the capabilities of the immune system, which can be galvanized to great effort, but only can perform a small bit against each area.
>
> "Suppose one of those "galvanized" areas in the targeted populace's immune systems is anthrax. And suppose the bad guys now hit one or more of those cities with a professional anthrax attack. A former OTA study showed that a professional anthrax aerial spray attack against Washington D.C., dispensing 100 kilograms of anthrax spores in the spray, will account for 1 to 3 million casualties. And that's in an "unthinned" set of immune systems. In the "thinned" set, one can at least double or triple, and even up to five times, that estimate. So conservatively, now there would be from 2 to 6 million casualties. And even mild anthrax spray attacks in other "thinned" cities would also be amplified.
>
> "That one is being set up, or already set up, right now. There are other such "augmentation" setups also under way.
>
> "So the weapons alluded to by the SecDef in 1997 are already being employed. WW III has already begun, and our populace (and our scientific community) is blissfully unaware of it."
>
> http://www.cheniere.org/correspondence/021702a.htm

Most of us can remember the strange outbreaks of the so-called "flesh eating disease." It seemed to break out here and there randomly with no known cause. Bearden believes this was a test of the newer quantum potential weapons which go even beyond the capabilities of the Tesla howitzers.

> "The Kaznacheyev "disease induction by novel electrodynamics" work is also the basis for the spacetime curvature engines (not EM signals!) carried by quantum potential EM biological weapons for distant induction of diseases etc. into the population of an entire nation or area, such as the U.S.
>
> "This quantum potential disease induction weapon -- capable of attacking an entire population of a nation or area -- has been tested several times in the U.S. at very low levels, in the shadow state only (we explain shortly) and it has a most peculiar signature. E.g., one test used necrotizing fasciitis (flesh-rotting disease) engines.

> "The disease 'engine' was deliberately kept down in the 'shadow state' (virtual state, just below the observable state) so that only a tiny bit of the populace with depressed thresholds would "breach the observable threshold' due to the population forming a bell-shaped curve as to the threshold level. The major signature of the test was that the cases of the disease broke out totally statistically and randomly, spread all over, without any 'disease vector' in between.
>
> "In short, it was not spread in any manner required by normal deterministic disease vectoring, but was a totally random set of occurrences. Several other diseases were also tested this way in the American populace, with precisely the same signature."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

"Apparently one planned use of such an augmented clandestine BW weapon will be smallpox, e.g. As you know, the Russians did develop smallpox for biological warfare, and even developed a new strain against which previous vaccination is ineffective.

> "The vast numbers of expected casualties, coupled with immune system spreading, are absolutely irresistible to the Russian mind. You can count on it that they have seen that the smallpox is already on site here, in the U.S. and in the hands of professional terrorist teams. Of course, the Russians/KGB will protest they have nothing to do with that! Yet likely the KGB will have flatly arranged it or at least assisted it. The capability for mass destruction of the nation is so lucrative and easy to achieve in advance, that they would never be able to resist doing it."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

Bearden believes the Gulf War Syndrome involved the use of scalar technology to induce a disease state, as he stated in a 1997 interview:

> "I'll say this flatly, and I won't answer any further questions about it; the Gulf War Syndrome was induced... It was induced with this stuff. All the rest... was contributing factors that everybody's talking about... There are some deeper signatures, if you look into it very deeply, that show you exactly how it was done and the fact that it was induced. It was a test. It was a test of a very special kind of weapon I have not talked about yet. Anyway you can use it to create diseases."
>
> Bearden http://twm.co.nz/beard_weapon.htm

Besides using the woodpecker grid to target a population, the entire electrical grids of nations can be used as "channels" for disease patterns to be broadcast to all people near that grid. Your own house wiring can become an antenna, broadcasting these "shadow" patterns of disease, weakening your immune system.

And here one faces the frustrating fact that the same power grids could be set to channel patterns which would keep the entire population healthy, rejuvenated, continuously bathing in the time-reversing old-age-defying energy of the spacetime vacuum. We have no idea yet of how long the human life span could be increased with this new technology. People might be able to live two centuries given full humane development of this technology.

5. Psychoenergetics

Mental entrainment, inducing hypnogogic trance, mind control, broadcasting thoughts and feelings, the Ant World, education via EM [LW] mental download, the "Feelies," revolution in spirituality, "enlightenment machine?" Scalar churches?

Psychoenergetics is simply using the scalar interferometers to manipulate and engineer the human psyche. To me it is the most mind-boggling and frightening aspect of longitudinal wave engineering.

The mind is electromagnetic in nature, and itself of the time-domain. It is not observable in 3-space. There is no thing you can point to that is the mind. As Bearden says**, "The mind is time-like**."

The longitudinal scalar EM waves are also of the time-domain. By overlaying certain infolded patterns (oscillations in time) on the longitudinal waves, a thought or feeling may be made to arise in a person or persons who are in the **interference zone** (where the wave beams cross).

The person will not notice anything, feeling that this thought or idea is his or her own idea. The longitudinal wave pattern might be a wave of panic or fear that spontaneously sweeps over you, perhaps inexplicably. It could even be in the form of a sudden intense patriotism. It could be explicit words and phrases, which everyone would think they had thought of by themselves. Or it could be rage and hatred, angry and violent. On the other hand It could also be a continuous sense of docility and placidity. This is the brave new world of psychoenergetics and the coming "Mind Wars."

Scalar Psychoenergetics in its most primitive form simply "entrains" all minds in the target area into a deep hypnogogic trance. In this state of mind people would be suddenly extremely suggestible, and would likely believe anything they are told, and would obey any orders given. Talk about winning the hearts and minds of the American People!

Tom Bearden makes me think of **Paul Revere**, who heroically conveyed important warnings for the sake of the new nation. But the words of the famous phrase have now strangely and curiously mutated into the once-inconceivable: '**One if by land, Two if by mind**."

Indeed one strategy in a Mind War would be to simply take over the minds of the "leaders" of the enemy nation. The targeted leader would not be aware that anything was amiss, although he might begin to make unexpected changes in policy.

And are, in fact, any of our "leaders" already having secret thoughts that are not their own? Who knows? Frighteningly, we can no longer be sure. The new sciences of scalar electromagnetics and psychoenergetics are even now plunging humanity into a sudden science-fiction-like world beyond anything in humanity's previous imaginings.

> "The Russians reached the point in the mid 1990s that they could take over control of a person's mind, with modified longitudinal EM waves including some time-polarized EM waves, and with a team of specialists (estimate 25-30 per transmitter, and one transmitter and team per controlled person)." Bearden http://www.cheniere.org/misc/time.htm

In a slide Bearden illustrates how a Scalar War would involve a psychoenergetics attack on the operators of the enemy scalar installation, entraining their minds into hypnogogic trance and getting them to shut down their systems.

And ultimately psychoenergetic warfare goes to the very heart of human identity itself. For if my thoughts might no longer be "mine," then who and what am I? Can my very sense of being "me" be hijacked by some nefarious psychoenergetic scheme? Will the "secret government" eventually dictate directly into your mind how you feel about yourself? Or what you think you are? Or what you should do?

How Does Longitudinal EM Mind Control Work?

The physics of scalar psychoenergetics as expounded by Col. Bearden seems even more arcane and over-my-head than that of extracting energy from the vacuum. I can't understand it, but I look through it anyway. Each time another tiny understanding might occur, or a term might fall into place. I know the mathematics is forever beyond me.

One paper to begin with is "Mind Control and EM Wave Polarization Transductions" This is such serious stuff that Bearden includes a strong warning about misuse of this knowledge. Psychoenergetics weapons can mentally maim and physically kill.

Warning!

(From "Mind Control and EM Wave Polarization Transductions")

> "This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p01.htm

After the serious warning comes the description of the paper. This is heady stuff for the layman.

> "Calling full attention to the special note above, in this paper we present a high-level overview of the novel electromagnetic nature of mind operations, mind and body coupling, and intent - the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transaction mechanisms whereby differing EM wave polarization's can be transformed one-into-the-other are presented."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p01.htm

I will leave it to the technically astute to try to understand the mechanisms of operation which are described in scientific language. But there are tidbits for the layman. Here he speaks of the difficulty Western scientists have in opening up to the "immaterial" (longitudinal wave) nature of the human mind.

Western Science Remains Largely Materialistic

> "Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system. This serious self-limitation exists because in the body we measure only weak TW [transverse wave] EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.
>
> "With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.
>
> "Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW [longitudinal wave] functions exist in the body and in all of nature in general."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p05.htm

Western methods of influencing the mind with EM waves have only used transverse waves, not the longitudinal waves of the vacuum. If only transverse waves are used one pretty well has to hit the target mind with a sledgehammer of waves.

Western Science knows only transverse EM waves

Brute Forcing Time Functions versus Fine Control Methods

> "Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines

> (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.
>
> "While KGB scientists also use TW EM "brute force" TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired... Bearden
>
> "These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design for example of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gail's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p09.htm

The seemingly magical operations of psychoenergetics, sometimes called "psychotronics," cannot be achieved with ordinary transverse EM wave energy. But they are possible with longitudinal scalar waves because the mind itself is scalar in nature.

> "We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain..."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p08.htm

The engineering of the mind can only be accomplished with the waves of time.

The course of psychoenergetics development necessitates the building of a vast database of settings and patterns which bring about given mental and/or emotional states. A kind of "gnome" of the psyche. Bearden describes the process of the decoding of this "gnome."

> "Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.
>
> "In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions."

> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p01.htm
>
> "A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p02.htm

Summing up his paper Bearden speculates on what the actual situation the science of psychoenergetics probably is. The intense secrecy with which this knowledge has been kept from the public has led to the current farce of the "government" pretending it all does not exist, and this secrecy may have now put us at great risk.

> "The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.
>
> "In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians..."
>
> "Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification."
>
> "The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U(l) electrodynamics. Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW [longitudinal wave] interferometers, through intervening mass."

> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

Bearden cites two examples where he believes the Russian KGB tested the psychoenergetics capabilities of their scalar interferometers. One was the case of Captain Button in his A-10 "Warthog" aircraft on April 2, 1997.

> "Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona."
>
> "At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—dream-thinking until he died that everything was normal. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour. "Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p06.htm

The paper "Mind Control and EM Wave Polarization Transductions" ends with the most chilling and frightening possibility of all: the psychoenergetic engineering of the human species as a whole.

The Ant World

The Final Race for Direct Mind Control of the Entire Human Species

> "The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs [longitudinal waves] for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials. "The KGB scientists also know that something very like Gaia — a collective unconscious mind for all species on earth — also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials.""Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.
>
> "I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can

> then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society" — communist style — is still a variation of feudalism (all systems tend to feudalism or some form thereof). "The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now...
>
> "Some versions of Russian mind control devices were used on Russian soldiers in the Afghan War to condition them for performance of merciless acts." 13
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p07.htm

More about "The Ant World"

Whether it is the Russians planning an "ant world" society or rogue U.S. black-ops groups withholding a great boon from human race, the secrecy on all sides around has allowed largely criminal elements to take over this technology while leaving the civilized elements in the dark. Openness and the spreading of this knowledge among good and decent people is essential, and for that we will have to overcome our frightened denial.

> "Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.
>
> "Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p08.htm

Like the huge energy of the Tesla howitzers, the ability to engineer the mind with longitudinal waveforms is a seeming "magical" thing which can be used for good or ill. Bearden envisions that further developments will allow something like the "downloading" of knowledge via scalar interferometry. Everyone could be highly educated, and quickly and easily.

In the conclusion of the paper "Mind Control and EM Wave Polarization Transductions" Bearden makes a plea to the nations of the world regarding the use of psychoenergetic engineering.

> "Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth.

> "Yet it can revolutionize medical science, education, communication, and psychology. We foresee the day -- perhaps 30 years hence -- when education will be accomplished by directly loading the software into the mind. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks on will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an MD And so on. When that happens then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.
>
> "We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p08.htm

(Perhaps after so many frightening topics a slight break is in order here to consider the lighter side of psychoenergetics. For example, what will be the ultimate impact on the entertainment industry?)

The Feelies at Last!

> Psychoenergetics will be able to bring about at last, for better or for worse, Aldous Huxley's fanciful idea in his novel "Brave New World," an entertainment format which he called "The Feelies." Those were something you went to, like going to the movies, but you would experience all the actual feelings themselves, as if the depicted events were actually happening to you. Although our "leaders" seem too dim to grasp what is going on in the world of scalar electromagnetics, it should not take some young Hollywood genius too long to understand that The Feelies are now possible, and that the immense megabucks of potential profit more than justify heavy investment in bringing it to market.
>
> One can imagine that first "Feelie" theater, "The Huxley" perhaps, a large comfortable area surrounded by nine giant screens (not just one screen!), which become 3D if you put on your headset. And The Huxley is fitted with the finest sound, and oh, by the way, a modest sized scalar interferometer with computer. And there are one or more new tracks now recorded on the film next to the sound track. These are the digital patterns of thoughts, feelings, and ideas to be fed into the theater's interferometer. And these thoughts and feelings and ideas will be felt by the audience as if they were their own thoughts and ideas. It is an entirely new art form. Psychevision.
>
> As important (and no doubt famous) as the actors would be the "feelers," those from whose minds the feelings were lifted in the original programming of the Feelie tracks. Because, for example, Brad Pitt might be a good and handsome actor, but can he really feel it? Like the dubbing of foreign language films, it might be necessary to dub the "emotional track" using a deeper and more profound person's feelings, thoughts and ideas.

It might even be engineered so that if you sat on one side of the theater you would feel like the hero, and on the other side you would feel exactly like the bad guy! What would that lead to?

We have all cried at a movie some time, most people have, though they might not want to admit it. But this would be deep sobbing and a flood of tears, if the director so desired it. You would feel it at your core. Or, take a big chase scene, you would be terrified. They are breathing down your neck! You would be gripped by panic.

For that matter, the whole genre of Horror-Feelie would be too much for many people to take. Do I really want to feel that an actual vampire is tearing at my throat with his fangs?

A Religious Feelie would make you feel the exaltation of a saint at his prayers, or give you the definite feeling that you are Christ suffering on the cross. All the pain could be there too. The stickiness of the blood. The very feeling itself: "Why hast thou forsaken me?" Who among us could come up that despair for scanning? Had this been developed a bit earlier we might watch a documentary of, say, Mother Theresa, containing her own actual deep sense of compassion, which we would feel arising in us as our own.

Then again, a Psychedelic Feelie could induce an LSD-like state and simply blow the audience's minds.

Even in its simplest most primitive mode, the hypnogogic trance induction with its increased suggestibility, would completely alter the experience of even ordinary current movie going.

6. As it Stands - 2002

"Anyway, it's presently a "Mexican stand-off" with the various parties maneuvering with check and countercheck. That is the real "balance of terror" that is held by such a fragile balance."

"You can see the implications if such weapons fall into the hands of the radical terrorists. Ordinary positive energy EMP shooters can be made fairly readily and cheaply from nearly just Radio Shack parts or surplus parts. It just takes some know-how... When and if the Yakuza [Japanese mafia] gets those (and they will, it's just a question of time), you can begin to appreciate some of the turns this present terrorism and our war on it will take."

Bearden http://www.cheniere.org/correspondence/021702a.htm

The very thought of the building of one of these howitzers by "some lone-nut gunman" from Radio Shack parts makes one realize how serious it is for humanity to face the crisis it has come to.

Once again, as at the dawn of the nuclear era, humanity finds itself at the crossroads of Heaven and Hell. As a species, we must stop and consider. And there is at least one real difference between this visit to the crossroads and the last: this time there is a very real possibility of a true "earthly heaven" on the Heaven side.

The poor nations could come to a good life. Disease could be largely eliminated. Fossil fuel pollution could end. Global warming could be solved, and the weather could be managed in a humane way, bringing water where there is drought, and sunny skies where there is flood. Scalar interferometry would be there to avert any possible hits by asteroids or comets. Mars could be colonized and terraformed. Everyone would be highly educated through psychoenergetic downloading of knowledge. And the wisdom of the wise could be realized directly through scalar connection to their minds and hearts.

These stunning sudden possibilities give me pause to think: if our species has reached this crossroads in such a short span of centuries, then surely other species in this vast universe have passed this way before and long ago. And of those who chose the path to "planetary heaven," what is their life like now? It might be good if some of them were with us now, to guide us through this last struggle over this pesky old good-and-evil problem. It is our last chance to choose the path which benefits the whole species, and not just some elite group or faction.

The Disclosure Project

One person working tirelessly to bring to bring these technologies into the open is Dr. Stephen Greer of the Disclosure Project, which is calling for congressional hearings into the issues of UFOs and the free energy technology that was, perhaps in part, reverse-engineered from them.

> "Once abundant and nearly free energy is available in impoverished areas for agriculture, transportation, construction, manufacturing and electrification, there is no limit to what humanity can achieve. It is ridiculous — obscene even — that mind-boggling poverty and famine exists in the world while we sit on classified technologies that could completely reverse this situation. "So why not release these technologies? Because the social, economic and geo-political order of the world would be greatly altered. Every deep insider with whom I have met has emphasized that this would be the greatest change in known human history. The matter is so highly classified not because it is so silly, but because its implications are so profound and far reaching. By nature, those who control such projects do not like change. And here we are talking about the biggest economic, technological, social and geo-political change in known human history. Hence, the status quo is maintained, even as our civilization hurtles towards oblivion...
>
> "With the types of weapons currently in the covert arsenal — weapons more fearsome even than thermonuclear devices — there is no possibility of a survivable conflict. Yet in the darkness of secrecy, actions have been taken on behalf of every human that may

> endanger our future. Only a full, honest disclosure will correct this situation. It is not possible for me to convey in words the urgency of this."
>
> Dr. Stephen Greer http://www.disclosureproject.org/ES-DisclosureImplications-2.htm

Greer and his team have assembled hundreds of witnesses, many of whom are military or ex-military, who are ready to testify to congressional hearings what they know about the covert black projects which have kept free-energy and antigravity propulsion technology secret from the people.

Greer points out the need to regulate these technologies so they are limited to peaceful uses, but this is of course difficult when the government simply chooses to pretend they do not exist. In the meantime, as the President and the Defense Department continue the charade of war-as-usual, those scalar (longitudinal wave) installations which DO exist continue to fall into ever more sinister hands. And there is no reporter in the press brave enough to ask, "Mr. Secretary of Defense, what steps are you taking to defend our MINDS from attack by the big Russian longitudinal interferometers?"

Longitudinal Interferometers are Proliferating

In a later statement Bearden reveals that in fact the Yakuza [Japanese mafia] has already leased some of the interferometers with a nearly $1 billion "down payment:"

> "After the collapse of the Soviet Union's economy, lots of things got available for lease or purchase that would never have been under the old system. That happened in the resulting economic chaos. The Russian Mafia rose to prominence, as did some wheeling and dealing entrepreneurs who played both ends against the middle.
>
> "In that atmosphere, and in the need for money, the KGB allowed leasing of the earlier longitudinal EM wave interferometers to a consortium of the Yakuza and Aum Shinrikyo. They did this only after they had tested and deployed their great new quantum potential weapons. That class of weapon is the dominant weapon on earth. Five nations now have it: Brazil, Russia (KGB), China, and two nations friendly to the U.S. The deal to lease many of the earlier LWIs [longitudinal wave interferometers] to the Japanese Yakuza and Aum Shinrikyo was consummated at the end of 1989, with a down payment of $900 million in gold bullion as "up front" money. I don't know what the lease per year was and is, but probably something like a billion dollars a year. So the rogue Japanese acquired longitudinal EM weapons and that technical know-how. In fact, the Yakuza is producing certain kinds of those in its own facilities in Japan now."
>
> Bearden http://www.cheniere.org/correspondence/030202a.htm

Personally, I am a little uncomfortable with the knowledge that the Japanese mafia may now be able to insert things into my mind. I'd like to sleep at night knowing that my dreams are my own, such as they may be, and not some weird fantastic visions cooked up by a criminal organization. This is a world emergency.

And in Bearden's analysis, it all comes back to the issue of energy. It is the energy crisis which is now provoking the current "oil wars," since oil production has hit the "Hubbert Peak" and will decline. And the electrical power grids must be replaced with scalar power as fast as possible, for these grids are now totally vulnerable to any possessor of longitudinal interferometers. They can all be wiped out in minutes. By anyone who possesses or controls a Tesla howitzer.

> "Anyway, in my personal view, we should have a national Manhattan Project right now, to develop decentralized electrical power systems taking their energy from the vacuum . . . The MEG is real and so are several other systems invented by private inventors and colleagues. All these efforts should be massively funded and massively and quickly developed. The survival of this nation is going to depend upon it. . . . There are many other facets, but the major point is that our scientific apparatus and the public are still unaware of what a desperate struggle we have entered, since it was thrust upon us. 9-11 was a wake-up call. It was certainly bad enough, and one grieves for those lost American lives. But I fear the worst is yet to come, and it is coming at us like a runaway train hurtling right down the track at us."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm
>
> "The electromagnetic weapons mentioned by the Secretary of Defense in April 1997 are in terrorist hands also, including the Yakuza and Aum Shinrikyo. That rogue group leased those earlier weapons on site in Russia, from the KGB, at the end of 1989. The SecDef stated that such weapons were being used to initiate earthquakes, initiate volcanoes into eruption, and control and engineer the weather. Quite true. And presently there is a truly massive set of weather engineering operations going on over North America, from that rogue Japanese group manning the weapons in Russia. Some 10 nations of the world now have the type of longitudinal EM wave interferometer weapons (which are what the SecDef was most probably referring to). There are other even more fearsome weapons, possessed by five nations. So a great deal of the state of the world is not covered in the news at all, and will not be . . .
>
> "We have been in an undeclared war of an eerie kind for some decades. That war seems now to have started to heat up also. The destructive capability of some of these weapons is awesome, and far greater than nuclear."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm
>
> "In that sense, the first phase of WW III is already completed. Now we are only waiting for Phase II.
>
> "It is against that backdrop that I think one must "analyze" such things as how fast we will be replacing the power grid (it's going to go anyway, and catastrophically, the terrorists will see to that). And everything we have is critically dependent upon energy, from the economy to jobs to mobility, to the millions of trucks that transport our goods every day, etc.
>
> "As far as I am concerned, we are already late in declaring a great national emergency in energy, because of the coming destruction of the present energy infrastructure. The

> scientific community should lead, follow, or get out of the way. The emergency is not in seeking "alternative but conventional" systems. It is in developing "energy from the vacuum" on an emergency, crash basis.
>
> "Unfortunately, everywhere I look I just see business as usual. Most Americans (and news media) seem to think, hey, Afghanistan is finished except for a bit of cleanup, and it's just about over).
>
> "It isn't. It has only just begun."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm

Beyond the Howitzers: Quantum Potential Weapons

Do not use EM propagation through space

As more and more nations acquire the scalar interferometers, Bearden speculates on the next generation of longitudinal weapons called "quantum potential" [QP] weapons.

> "The quantum potential weapons are presently only held by three nations: Russia (in the hands of the KGB, which is now know by its new name), Brazil, and the "friendly little nation." Red China may be working on them, but is just at the beginning...
>
> "The QP weapons have been highly reserved by the Russians. However, the complete extent of their capabilities is still being figured out...
>
> "The quantum potential weaponry (possessed by only three nations) is supreme, because it uses multiply connected spacetime, and DOES NOT involve propagation of EM energy through space as does the longitudinal EM wave interferometers.
>
> Bearden http://www.cheniere.org/misc/qp.htm
>
> "The operational deployment of full-scale strategic QP weapons on site in Russia occurred at the end of 1989, which is when the formal weapons first went operational. These weapons are not in the inventory of the regular Russian armed forces, but all research and development, manufacturing, deployment, an manning and employment is under the ruthless control of the KGB. Shortly thereafter (within weeks), the KGB leased many of their earlier longitudinal EM wave interferometer (LWI) weapons to a rogue Japanese group comprised of the Aum Shinrikyo and Yakuza."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

In another paper, "Time Polarization - Significance and Weaponization," Bearden tells the story of how a KGB longitudinal wave attack was averted by the longitudinal weapons of a "friendly small nation."

So as it stands, it is precarious. Not until the people demand an end of what author Jim Marrs has dubbed the "Rule by Secrecy," can the people's representatives tackle the big problems of a worldwide ban on Scalar Wars and the shifting over to an essentially oil-free economy.

Latest Update from Tom Bearden - March 2002

The whole situation is changing so fast that when I asked Tom to look over this article he mentioned the seeming contradiction when some of the above statements are taken together. At different times the information he has is updated so one has to keep a kind of running track. He was kind enough to clear up the confusion here with the situation as it stands as of March, 2002.

> "We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy." Tom Bearden

Emptiness is Full of Energy

"The Tao is an empty vessel;
it is used, but never filled.
Oh, unfathomable source
of ten thousand things!

"Oh, hidden deep but ever present!
I do not know from whence it comes.
It is the forefather of the gods."

Lao Tzu -Tao te Ching

Afterward

Because of the extreme seriousness of the topics brought forward in this article, I asked Tom Bearden to look it over to make sure I had not gotten anything too terribly wrong or mixed up. I did not want to pass on any misunderstanding of my own. I felt a little nervous writing to an eminent scientist, and hoped I had not presumed upon his time.

But Tom was so kind as to answer immediately, and to my relief I had not gotten anything too terribly wrong. Here is his letter, in which he clears up a misunderstanding, and reports on his current health treatment, and gives a call to the young grad students to turn themselves loose in the whole new field of scalar electromagnetics and free energy.

Subject: RE: Article on Tom Bearden and "Scalar Wars"
Date: Fri, 22 Mar 2002 22:54:01 -0600

Bill,

Well, I must say I'm astounded at your persistence and hard work in congealing all that together. My compliments on a hard job quite well done.

Couple tiny things: In the area mentioning Kervran and Kaznacheyev: Kervran was French; Kaznacheyev was and is Russian. Little typo of "spredding" the immune system. Should be "spreading".

We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy.

Otherwise, it's pretty accurate.

Just now I'm letting the weapons take care of themselves; have to, as I will be recovering now for nearly a year from now after the heart attack last year and discovering that I had contracted mycoplasia (the BW modified kind) in Canada in 1968. Finally got a test which confirmed the chronic mycoplasia (which I had for 33 years!). The stuff burrows inside one's red corpuscles, then draws nutrients from the hemoglobin, hardening it and reducing its ability to take on oxygen (by as much as 50% or more). The resulting symptom is loss of endurance (which I had for that 33 years) -- the so-called "chronic fatigue syndrome".

Many of the Gulf War Veterans who are sick also have it, though they apparently have other complicating factors as well. A very high percentage of persons with chronic fatigue syndrome also have mycoplasia. The treatment for the long-standing mycoplasma infection of that nature is to stay on antibiotics for a year. The only time the mycoplasma is vulnerable is when one wears out a red cell in normal usage, and the body makes a new red cell to replace it. The mycoplasma comes out of the dying cell to infect the new replacement cell, so it is exposed and vulnerable at that specific time and that time only. The antibiotics kills it at that time. Hence the necessity to stay on antibiotics for such a long time so it is continuously in your bloodstream (I've already been on antibiotics for over two months, and have at least 10 more months to go). We also take a little medical oxygen every day.

Another symptom of the mycoplasma infection is the association of runaway fibrillation of the heart. To stop that, there are pills one takes, and they stop it. But they also further reduce the volume of blood pumped by the heart, so that this projects one right back into the hypoxia, dizzy, loss of balance, etc. Catch 22 situation; the fibrillation will build and cause heart attack, stroke, etc. if not stopped, and if stopped it puts you back in the hypoxia which can cause many things! With the conventional treatment, you're damned if you do and damned if you don't. Hence the need for a little extra medical oxygen, so you can take the "heart clamp" pill and stop the fibrillation, but also augment the oxygen enough to continue to survive without a stroke or some such. Interestingly, it's very difficult to even obtain a mycoplasma test, and Medicare fights you tooth and nail over paying for the oxygen. Either they will pay for it or I'll pay for it myself; it is absolutely necessary.

After a heart attack, one is treated in the heart clinic by specialists who are very caring and excellent practitioners, but know nothing of mycoplasma infection, so not consider it even a part of "heart disease", and who are rather puzzled when one has no clogging of the arteries, has the right count of red corpuscles (largely nonfunctional, of course, but there), pumps the right amount of blood, etc. So they identify and treat only the runaway fibrillation, do not prescribe the oxygen -- and essentially place you on a regime as best they know, but one that is guaranteed to kill or maim you or turn you into a totally disabled person or human vegetable via stroke, paralysis, etc. Fortunately my family doctor will prescribe the oxygen anyway, though I may have to pay for it myself. Whatever works!

However, I'm slowly improving little by little, working as much as I'm able to and trying to finish my book: Energy from the Vacuum: Concepts and Principles, and get it to the publisher for publication later this year. Hopefully the book will turn all the interested young grad students and post-docs loose in the free-energy area. They will not have to spend 30 agonizing years to get to where I am, and where a few others are, but can simply start here and go forward. Since speed to develop energy from the vacuum is of the essence because of the world situation, I'm moving on it as fast as I am physically able to.

Meanwhile, working with the AIAS (Alpha Foundation's Institute for Advanced Study), the AIAS (15 authors) has now gotten about 20 scientific papers published in leading journals (Foundations of Physics, Foundations of Physics Letters, Physical Scripta, Optik, etc.) dealing with electromagnetic energy from the vacuum. Several more are either approved for publication or in the review process also. So slowly we are getting it into the mainstream scientific literature. Hopefully it will be in time, but it's cutting it close.

Very best wishes,

Tom Bearden

Late update

Massive display of longitudinal interferometers?

UFO Fleet Taped Over Bedhampton, England

Is this sighting a case of the testing of all the Russian (KGB) longitudinal interferometers at once? And coordinating their targeting through the woodpecker grid to a common place, as an exercise in making a massive scalar attack upon a given region? Was this a confluence of Russian Tesla howitzer marker-beacons?

Bedhampton, England - March 22, 2002 - Video Frames by: Anthony Woods

Certainly for a massive all-out scalar attack you would want coordinated movements of the target areas of many howitzers at once, for you could then just cut a bloody swath across a wide area, all the howitzer blasts marching across the land in a kind of scalar version of the "scorched-earth policy." All in a line, blasting, and heat, and mind waves. Or you might want to put the blasts all in the same place for a truly horrendous destruction. And on top of that biowar on citizens whose immune systems have been weakened with the quantum potential weapons. This is the latest version of Armageddon.

And should any madmen gain control of these weapons, any rogue group at all, and should they have some agenda to radically depopulate the world, to simply kill by the billion, the means will be in their hands to do so.

Update: MEG patent granted

Motionless Electromagnetic Generator Patent Granted

This message is from Jean-Louis Naudin, the French scientist who successfully replicated the Motionless Electromagnetic Generator (MEG) in France. He announces the granting of the patent to the MEG Builder's group at Yahoo.

----- Original Message -----

From: jnaudin509@aol.com
To: jlnlabs@yahoogroups.com
Sent: Tuesday, March 26, 2002 2:20 PM
Subject: [jlnlabs] (Info) GOOD NEWS : The Tom Bearden's MEG IS PATENTED

Dear ALL,

GOOD NEWS !!!!

The Motionless Electromagnetic Generator from Tom Bearden is now PATENTED US 6362718 granted on March 26, 2002

US Patent 6,362,718 :
Motionless Electromagnetic Generator (MEG)

Abstract: An electromagnetic generator without moving parts includes a permanent magnet and a magnetic core including first and second magnetic paths. A first input coil and a first output coil extend around portions of the first magnetic path, while a second input coil and a second output coil extend around portions of the second magnetic path. The input coils are alternatively pulsed to provide induced current pulses in the output coils. Driving electrical current through each of the input coils reduces a level of flux from the permanent magnet within the magnet path around which the input coil extends. In an alternative embodiment of an electromagnetic generator, the magnetic core includes annular spaced-apart plates, with posts and permanent magnets extending in an alternating fashion between the plates. An output coil extends around each of these posts. Input coils extending around portions of the plates are pulsed to cause the induction of current within the output coils.

Inventors: Patrick; Stephen L. (2511 Woodview Dr. SE., Huntsville, AL 35801); Bearden; Thomas E. (2211 Cove Rd., Huntsville, AL 35801); Hayes; James C. (16026 Deaton Dr. SE., Huntsville, AL 35803); Moore; Kenneth D. (1704 Montdale Rd., Huntsville, FL 35801); Kenny; James L. (925 Tascosa Dr., Huntsville, AL 35802)

Appl. No.: 656313
Filed: September 6, 2000

For some technical info see my web site at : http://jnaudin.free.fr/html/meg.htm

Best Regards
Jean-Louis Naudin
Email: JNaudin509@aol.com
Main Web site : http://jlnlabs.org
Site France : http://jlnlabs.multimania.com

Click here to view patent # 6,362,718

Patent Story on Rense.com:

Rense.com

MEG Scalar Energy Device
Patented - Production Starts Next Year

From Bill Morgan
wmorgan@nycap.rr.com
3-28-2

Patent was granted on March 26, 2002 for "The Motionless Magnetic Generator," (MEG) US Patent 6,362,718, which is likely to become the first commercially available free energy device in history in about one year from now. The machine will provide free electricity from the vacuum, for the life of the device, which should be a very long life since it has no moving parts. You can see a picture of scientist Jean-Louis Naudin's MEG replication model at: http://jnaudin.free.fr/html/meg.htm.

It has strong magnets, coils, and a controller unit with the electronics. Naudin made the announcement to the MEG-builder's Yahoo group. The announcement has significance since the patent office has always been skeptical of devices which seem to "get-something-for-nothing." But according to the new science of scalar electromagnetics, the MEG does not break the law of conservation of energy. It's just that the energy is conserved in the fourth dimension, time, and not our 3-space world.

The MEG provides electrical energy by tapping the longitudinal electromagnetic (EM) waves which exist in almost infinite abundance in the vacuum of space. This ocean of energy which permeates everything is sometimes called the "zero point" energy, since it remains there even at absolute zero temperature.

Four inventors are listed: Stephen L. Patrick, Thomas E. Bearden, James C. Hayes, Kenneth D. Moore.

Tom Bearden has explained the operation of the MEG on his website Cheniere.org, and also speaks about the new fearful weapons that can and have been made using the same "longitudinal waves" of the vacuum.

The complicated physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum," which can be found at Tom Bearden's website:

cheniere.org.(http://www.help4all.de/energy/MEGpaper.pdf)

The first MEG units to be produced for sale will output 2.5 kilowatts of free electricity. Forever. They should be in production about a year from now. Facilities for manufacturing the device are being set up in an unnamed "friendly nation."

This free electricity will flow indefinitely, without much, or any maintenance. The units may be hooked together to provide more wattage, so four of them would provide 10 kilowatts. After some production experience units will be made which output 10 kilowatts each. With a couple of those units a house could get off the electrical grid.

Tom Bearden, one of the inventors has said "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat." This cat, it would seem, is out of the bag now.

The ordinary EM waves we are familiar with are called "transverse" waves, to distinguish them from the new "longitudinal" EM waves of the vacuum. Bearden has explained in depth on his website cheniere.org that wherever there is a dipole (battery, generator, magnet) there is an unseen flow of longitudinal EM waves in that local vacuum, the only problem is in tapping that energy and "transducing" it to electricity. Bearden says that the problem with all the electrical circuits we have is that they are two-wire circuits, a loop by which half the energy goes back to destroy the dipole. In the MEG that closed loop is never made. So the dipole is not being destroyed.

Patents are not granted on devices which do not work, so in a sense this announcement proclaims a new era. The MEG device itself proclaims and proves that energy is abundantly available everywhere, for free. We only have to build the devices to tap it. MEG type devices could be put into a car with an electric engine to make a truly fuel-less automobile. Electricity can be available in very remote places. People will be able to get off our very terrorist-vulnerable power grids. In an era of terrorism a highly dispersed power system would be most desirable rather than our centralized systems.

What is somewhat astonishing to me is the discovery by Bearden et. al. that time itself is actually compressed energy, and that this free energy is actually coming from the time domain, the ocean of longitudinal EM waves which fill the empty vacuum of spacetime. In fact, time is energy compressed by the same factor that matter is compressed energy: the speed-of-light squared. Thus there is a new companion to Einstein's E=mc2. Can you say "E equals delta-tee-cee-squared?" The "tee" is time and delta-tee is change in time.

A company has been set up to manufacture the MEG called Magnetic Energy Limited.

The ramifications of free energy are enormous. The oil wars are not necessary. If we threw as much money at this technology as we are spending on the oil wars we would be free of the need for oil in less than a decade. With fuel-less cars air pollution will be greatly lessened. Third world nations can raise their standard of living eventually. And the energy is free. And it never runs out.

I have put together a kind of "Bearden for Beginners" article which explains some of the basic concepts of the current state of scalar technology.

http://www.prahlad.org/pub/bearden/scalar_wars.htm

The granting of the MEG patent is the herald of the new era of scalar electromagnetics, and the free energy which flows forever and never runs out.

Bill Morgan
wmorgan@nycap.rr.com

Do the magnets run down after a while?

Further developments: Tom Bearden clears up a question

When the MEG patent was released many questioned the patent language which seemed to imply that permanent magnets of the MEG would "wear out" after awhile. I myself wondered about the same thing so I put the question to Tom, and here is his reply.

> Dear Bill,
>
> Thanks for the kind words; much appreciated.
>
> I'm not up to going on radio or TV, and won't be for quite some time. Still suffering from hypoxia, though very slowly improving, and will still be on antibiotics for 10 more months at least (maybe longer).
>
> Obviously an extensive dialog developed between the Patent office and our patent attorneys. As best I can understand it, the Patent office accepts the fact that a permanent magnet produces and emits energy, or has a history of accepting it. They do not appear to be at all cognizant of the broken symmetry of opposite charges, and hence of a dipole. So in their opinion (which seems to be vague), the magnet has to use itself up very gradually, or some such. So some rewrites were essentially required by the Patent office.
>
> All I can say is that a magnet is a very durable thing, so long as you don't heat it up too much or shock it to much. Certainly one will make a good 20 years, and of course there are plenty of magnets that have been around awhile longer.
>
> That said, in our earliest experiments we did "measure" some such effect, or so we thought, until we found that a particular instrument was bad and giving erroneous readings. So with a new instrument, we have not observed since then any such readings.
>
> Out of all that fuzziness came the inclusion of "using up the magnets gradually". So far as we are aware (after correcting those early readings on a defective instrument), we have seen no such measurements or effects since then at all. Nonetheless, just in case there is some very small effect in that respect, at least it's in there.
>
> Actually, as you are aware, capacitors and solid state components are the major concerns for failure, and any system does require maintenance, both periodic and whenever malfunction or failure occurs. The MEG is no different from any other device in that respect, so of course it will have a component failure rate and maintenance required from time to time, just as any other electromagnetic device. But without moving parts, it can be ruggedized and made very durable indeed.
>
> Please give Jeff Rense my regrets, and I do admire his work and his show. He's laying it in there.
>
> The depth of the scientific mindset against COP>1.0 is inexplicable, particularly when physics itself already has COP>1.0 validated experiments (such as Bohren's experiment,

which outputs 18 times as much as one inputs by Poynting calculations), and it works every time. Any competent university optics lab can perform the experiment. Also, even the conventional texts admit that a charged capacitor or electret laid on a permanent magnet, so that the E of the cap or electret and the H of the magnet are at right angles, is a "free energy machine". That silly thing will just sit there and pour out Poynting energy S = EXH indefinitely. Years and years and years. If you wait just a single year, that silly thing has changed the energy density of a volume of space a light year in radius, reaching out well beyond the solar system. Now that's a WHALE of a lot of EM energy that beast poured out in that year, and it will still be going strong and pouring it out at a steady rate. The charges and dipoles in original matter in the universe have been pouring out EM energy in that fashion for some 14 billion years, give or take a billion or so. It's also easy to see that, prior to Lorentz's ARBITRARY symmetrical regauging of the Maxwell-Heaviside equations, those equations prescribe both EM systems in equilibrium with their active environment, and EM systems not in equilibrium with it. The latter systems are permitted by the thermodynamics of open disequilibrium systems to perform five magic functions: (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator inputs (the active environment inputs the rest of it), (4) power itself and its load (the active environment inputs all the energy, like a windmill), and (5) exhibit negentropy. That is a well-established thermodynamics. Lorentz arbitrarily discarded all such open disequilibrium EM systems in order to get simpler equations having analytical solutions and not requiring numerical methods. So he simply scrapped all the really interesting and challenging EM systems in Maxwell's theory. ARBITRARILY!

That our scientific establishment continues to teach only that "half" of the theory, and not point out the rest to the young students, and not fund research into developing such systems that take energy from the vacuum, is simply inexplicable. It's a matter of total dogma, not science at all.

But hopefully it is changing a little bit. At least now we cite chapter and verse when these things were done, and what they are, and the young grad students and post docs can go check the papers and make up their own minds.

That is the real hope of the future: that we get a generation of young fellows, with excellent skills, who now know where all the skeletons got hidden in the closet, and understand that the electrical engineering model arbitrarily assumes an inert vacuum environment (falsified for more than a half century in particle physics) and a local flat spacetime (falsified for nearly a century by general relativity).

Once they know that past history, and go read into what broken symmetry of opposite charges means for a dipole and for all dipolar EM circuits, they will set to and change that more than a century of error. In that case, we shall have numerous solutions to the energy crisis rather quickly, popping up all over.

Good thing. It will certainly displace some great energy barons, but it will also free the populace. One can extract all the EM energy from the vacuum one wishes, anywhere, anytime, with ridiculous ease. One can produce that "electrical wind of energy" at will, by making a simple dipole and then letting it alone. The only energy problem is in how to then intercept some of that energy flow and "catch" it in a circuit, discharge it in a load to power it, and not use half the "caught" energy to kill that dipole that is gushing out the extracted EM energy from the vacuum.

Very best wishes,

Tom Bearden

END

You're viewing the old pesn.com website

Pure Energy Systems

"PES Network, Inc: Promoting cutting-edge, clean energy technologies"

Home | PESN | News | PESWiki | Directory | NEC | Projects | Radio | Newsletter | Store | Submit

Free Energy is all about freedom:
Power to the people -- literally and figuratively

Top 5 Exotic Technologies

CLICK TO START BLOCKING ADS

Ad Remover

12 THINGS TO KNOW BEFORE BUYING A THERMAL CAMERA
FLIR
GET THE GUIDEBOOK NOW!

Google Custom

"Free Energy"

News XML

- PESN Specials
- About
- Pure Energy Blog
- Daily FE News
- Features
- Free Energy Now
- This Week in FE
- Newsletter
- How you can help
- Submit
- Subscribe

Select Language
Powered by Google Translate

Directory

Energy Topics

- Alt Fuels
- Anti-Gravity
- Batteries
- Betavoltaic
- Biofuels
  - BioDiesel
  - BioElectricity
  - Biomass
- Body Electric
- Brown's Gas
- Cold Fusion
- Conservation
- Electrolysis
- Electromagnetic OU
- Fuel Cells
- Fuel Efficiency
  - Electric Vehicles
  - Engines
  - Hydroxy
- Fusion
- Geothermal
- Gravity Motors
- Human Powered
- Hydro
- Hydrogen
- Joe Cells
- Lighting
- Magnet Motors
- Nanotechnology
- Nuclear
- Nucl. Remediation
- Oil
- Piezoelectric

/2011/03/26/9501797_Teslas_Scalar_Waves_Replicated_by_Steve_Jackson/
You are here: PureEnergySystems.com > News > March 26, 2011

Tesla's Scalar Fields Still Beaming On!

The superluminal (faster than light) scalar or longitudinal waves Nikola Tesla used to magnify and wirelessly transmit power are not just a thing of the past. IEEE engineer, Steve Jackson, discusses and demonstrates how they can be utilized today, and he is open sourcing it here!

by Hank Mills
Pure Energy Systems News

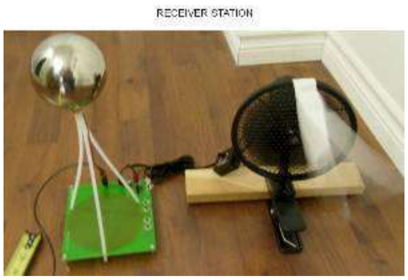

Nikola Tesla is the father of the modern age. His inventions of radio, remote controlled devices, alternating current, and induction motors (among many others), paved the way for modern civilization. Yet other of his inventions barely saw the light of day, despite their tremendous potential.

Such a severely under utilized technology is his wireless energy transmission techniques. By producing scalar (longitudinal) waves, he could achieve superluminal (faster than light) communication, send signals through any material including Faraday cages, and wirelessly beam power to receivers. Researchers like Steve Jackson provide hope that by open sourcing Tesla's work, scalar wave transmission can proliferate and positively benefit the world.

Last Thursday, March 24, 2011, Steve Jackson held a presentation and demonstration of a scalar wave transmitter and receiver at a local IEEE meeting at McMaster University in Ontario, Canada. During the talk he gave an overview of Tesla's life, his inventions, and a plethora of information on his wireless power transmission technology. Afterwards, he demonstrated a small electric fan attached to a scalar wave receiver powered by a transmitter over twenty feet away. Tesla's technology is alive today!

Furthermore, he has approached us at PES Network to open source his design. You can download his PowerPoint presentation here. We'll be uploading to

Tesla Generator - Build Your Own Generator

Eliminate Your Power Bill Easy. Do It Yourself. Great Discount Now! Just $49.

tesla-energy.org

Jim Murray and Paul Babcock's 4790% SERPS Presentation	Ignition Secrets DVD by Aaron Murakami	
Magnetic Energy Secrets, Paul Babcock, Parts 1 & 2	Battery Secrets by Peter Lindemann	

Ads

Electronic Permit System

ePTW system with integrated Isolation and Competency controls

gemsoft7.com

AdChoices

Generator

Generator for Power

GREENfuel H2O .com
Save Fuel Today
Hydrogen on demand
The fuel of tomorrow

SAVE $ AND TIME
Make Biodiesel at Home!

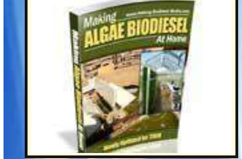

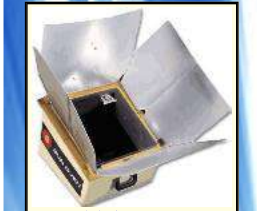

Global Sun Oven

- Cold Fusion News
  - Kalte Fusion

- Plasma
- River
- Salt Water Mix
- Solar
- Solid State Gen.
- Tesla Turbines
- Thermal Electric
- Tidal
- Vortex
- Waste to Energy
- Water
  - Water as Fuel
- Wave
- Wind
- Wireless Electricity
- Zero Point Energy
- MORE . . .

Open Source
- Freddy's Cell
- Bedini SG
- Safe Haven Villages
- MORE . . .

Resources
- Awards
- Conservation
- Conspiracy
- Directories
- Investment
- Kudos
- Legal
- Organizations
- Plastic and Energy
- Recycling
- Suppression
- Tools
- Trends
- MORE . . .

Mingling
- OverUnity Forum
- Employment
- Events
- Humor
- Magazines
- Movies
- Newsletters
- Discuss. Groups

Shopping
- Store
- Buyer Beware

- - - - - - - - - -
- Donate
- Contact

our PESNetwork YouTube account a video of his presentation to the IEEE in a couple of days. It's in process of being transmitted, but will take a while, as it is 3+ Gb. We're expecting to get instructions from Steve on Monday, and we'll be posting a PESWiki page to organize all this material to streamline the presentation.

Take a look at the following video that shows a small boat powered by scalar waves many feet from a transmitter.

A & P Electronic Media
Leading Digital Publishers of Advanced Energy Science
http://emediapress.com

o Fusion Froide
o Fusion Fria
o Fusione Fredda

Free Energy Store
* * * * *
Your Ad Here

Demonstration of Tesla's Wireless Power Transmission System duplicated at the Tesla Tech Conference, July 2007.

Scalar Wave Properties

You might be wondering what in the world is a scalar wave! It's actually pretty simple. Unlike a transverse wave (like an ordinary radio or light wave) a scalar wave's amplitude does not fluctuate up and down. Instead, try to visualize it more like a vibration of electric potential expanding and contracting in the direction of propagation.

Scalar waves are also often called "Tesla Waves" or "Longitudinal Waves." All of these terms refer to the same type of wave which oscillates in the direction of propagation instead of perpendicular to the direction of movement. These waves have many unique properties which have been tested and confirmed repeatedly.

1) They are capable of penetrating any solid object including Faraday Cages. You can put a transmitter in a box of thick metal and a receiver outside of the box will receive the scalar wave frequency you are pulsing. The potential here is for a transmitter that can penetrate any obstacle or perhaps communicate directly through the Earth from one side of the globe to the other.

2) They are capable of superluminal travel. These waves are claimed not to be electromagnetic, but composed of pure potential energy. Due to this, the speed of light limit does not apply to them. The propagation speed of a scalar wave has been measured as faster than the speed of light and thought by some researchers to be potentially of infinite velocity. (Perhaps SETI - The Search for Extraterrestrial Intelligence - needs a different kind of receiver in order to pick up signals from elsewhere in the universe! Why would advanced ET civilizations even bother using slow transverse waves for interstellar communications?)

3) They are capable of transmitting power. It can then be put to use powering motors, lights, vehicles, etc.

4) A scalar transmitter can wirelessly send power to a receiver through any obstacle. For example, as mentioned above, you can setup a transmitter, put it in a Faraday Cage or a metal box, and a

receiver can receive power far away.

5) During the process of transmission and reception they can magnify power. You can input a certain quantity of power into a scalar wave transmitter and yet the receiver can receive several times the power.

6) Transmitters and receivers can communicate.

7) A conventional transverse wave transmitter cannot sense if a receiver has "tuned in." However, in a scalar system the transmitter and receiver interact because they are in a state of resonance.

As you can see Scalar Waves hold many potential applications for communication, energy, and other applications. To mimic a commercial for a popular brand of cell phone, "Scalar Waves do what Transverse waves don't." They are fast, penetrating, connected, and can broadcast magnified power. These waves are begging to be utilized!

Recipes for Scalar Waves

Nikola Tesla's most famous scalar wave transmitter was his Wardenclyffe tower on Long Island, NY. Prior to that, at the height of his amazing inventive career, he had a lab in Colorado Springs, Colorado. He was able to beam power to receivers miles away that were used to illuminate neon tubes. A great deal is known about his tower, transmitters, and receivers. It is all detailed in his patents. This allows for present day replications!

From examination of Tesla's patents, researchers have learned that perhaps the most critical component of a scalar wave system is the antenna. It is composed of a flat "pancake coil" composed of both a primary and a secondary winding. The coil is connected to a spherical metallic resonator that helps transmit and receive scalar waves.

A particularly good source for information about scalar wave systems is the website of German scientist, author, and researcher, Konstantin Meyl. He has published a series of books that describe scalar wave systems, detail their components, and discuss their operation. His books also dive into the science and physics of scalar waves, among other fascinating topics.

In fact, he offers scalar wave kits for sale on his site. They include all the components needed for a complete system, arrive in an aluminum suitcase that can serve in the same manner as a Faraday cage for your experiments, and come with a book detailing the kit. These kits have been used by experimenters to prove to themselves and others the existence of scalar waves and their unique properties.

Take a look at the following video to see Konstantin demonstrating a scalar wave transmitter and receiver. The same device in the following video is for sale right now on the internet from his website.

M
eyl shows Tesla's longitudinal electricity transmission in Washington DC at a Tesla,m meeting in 2003!

Even NASA is paying attention to scalar waves to some degree. The paper titled, "Advanced Energetics for Aeronautical Applications: Volume II" by David S. Alexander offers a review of scalar wave (longitudinal wave) research. It mentions the work of Konstantine Meyl and others in the field. Interestingly, the paper also discusses how Dr. James Clerk Maxwell's original equations specifically allowed for scalar or longitudinal waves. Only many years later did short sighted scientists who did not care to deal with Maxwell's complex equations based in "quaternions" mathematics arbitrarily discard longitudinal waves. Due to this, many mainstream scientists doubt their existence despite their history and the mountain of evidence which proves they are very, very real.

Opening Tesla's Technology

Steve Jackson's goal is to open source Tesla's wireless power transmission technology. This is very possible to do, due to the simple nature of

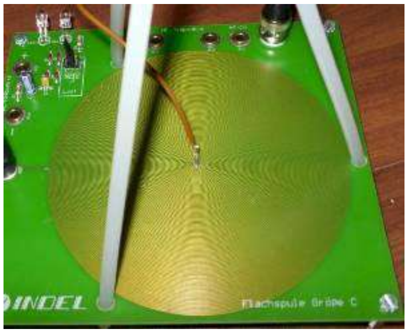

the parts and components required. During his talk he published a list of components required to build a system. This same information can be found in the books of Konstantin Meyl and elsewhere on the internet.

By open sourcing this technology, ordinary individuals could prove to themselves, friends, associates, co-workers, and others that Tesla's technology worked as claimed. Not only does it work, but can be replicated and used for practical applications today. All unique properties of scalar waves could be harnessed and developed to revolutionize our society.

When Steve releases plans, drawings, and other information, probably next week, we will add them to a project page at PESWiki.com. We'll add a link from this page when it is ready.

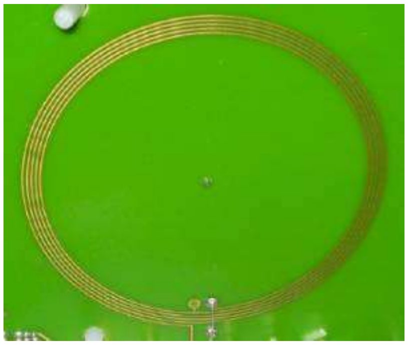

Earth Changes and the Future

In his books and lectures, Konstantin Meyl has compared superluminal scalar waves to the neutrino, which is also thought by some to travel faster than light. Neutrinos bombard the Earth in all directions, but the majority of them arrive from the central star of our solar system. Due to the fact that approximately half of the neutrinos are absorbed as they pass through the Earth, it is thought they may interact with the core and inner layers of our planet.

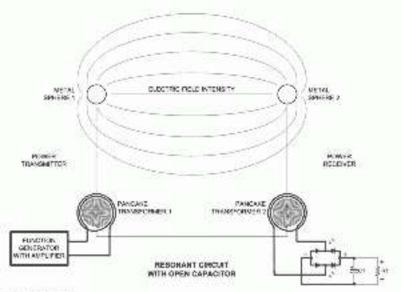

©INTALEK, INC., 2007

Circuit diagram used by Steve comes from Intalek.com, Figure 3.

Some researchers suggest solar phenomenon can produce variations in neutron emissions which may trigger "earth changes" on our planet. This was depicted in the blockbuster movie, 2012, as the cause for the crust disruptions from the mantle heat-up from increased solar activity. Hence, studying and understanding the nature of scalar waves may increase our chance of survival during chaotic periods in the future. Perhaps by understanding them we could eventually learn how to constructively utilize them to stabilize our planet during periods of solar fluctuations, sun spots, flares, etc.

Perhaps with the use of scalar wave technology, earthquakes, volcanic eruptions, or other natural disasters could be prevented or detected ahead of time. Even a one hour warning before an earthquake, such as the one that recent devastated Japan, could save thousands of lives. Although all kinds of unique technologies are possible with scalar waves, at a minimum we need to understand how neutrinos and scalar waves interact without planet. Not doing so could result in needless death and suffering.

Lets get moving!

To not harness scalar wave technology for the benefit of mankind would be tragic. Our civilization is facing adversity in countless forms. Hunger, disease, pollution, poverty, and war threaten our future and the future of our children. Scalar waves exist, they are useful, and the information to build devices that generate and receive them is available to everyone. To better understand our universe, to ensure the continuation of our civilization, and to honor Nikola Tesla we should do our best to support the open sourcing of this technology.

Videos

On March 30, Steve conveyed the following YouTube links showing his presentation to the local IEEE on March 24.

#1 Slide 1 https://www.youtube.com/watch?v=qs-2s85lO6o
#2 Slide 20 https://www.youtube.com/watch?v=s4kvJpUgGKI
#3 Slide 39 https://www.youtube.com/watch?v=4YXPa73NNsA
#4 Slide 57 https://www.youtube.com/watch?v=LqhRDY56CqU
#5 Slide 81 https://www.youtube.com/watch?v=e1Ht5OVyI5Q
#6 Slide 93 https://www.youtube.com/watch?v=JuehDylES_A

Replication Instruction should be done tomorrow.

Contact

Steve can be reached by email at steve.jackson@ieee.org

# # #

This story is also posted at *BeforeItsNews*.

What You Can Do

Recommend 14 people recommend this. Be the first of your friends.

1. Replicate this design.
2. Pass this on to your friends and favorite news sources.
3. We at PES Network are in a pinch right now. Donations would be greatly appreciated.
4. Subscribe to our newsletter to stay abreast of the latest, greatest developments in the free energy sector.
5. Let professionals in the renewable energy sector know about the promise of this technology.

Comments

Feel free to view/post comments down below. The following came in by other means.

Not Scalar "Wave"

On March 27, 2011, 1:45 pm, Karl Palsness wrote:

You should read about Scalar .. it is a "Scalar FIELD" there should be no such thing as a "Scalar wave"; and there is a "Longitudinal Wave", instead of scalar wave. Scalar waves just don't exist according to the word scalar meaning size. A Tesla coil has a scalar field around it...and emits a longitudinal wave....

A Tesla-era dictionary defined it as:

> **Scalar** -- In physics, a quantity which has magnitude, or magnitude and sign only, without direction, such as density, mass, energy, etc., as distinguished from a vector quantity.

and

> **longitudinal wave**: A "pressure" type of wave, similar to sound, in which the vibrations are along the direction of travel of the wave. Hence, a wave composed of alternating densifications and rarefactions, where we focus upon the longitudinal component of the changes.

* * * *

Longitudinal Waves (L.M.D.)s

On April 02, 2011 12:29 AM MST, Dr. Eng. Roberto Handwerker <info {at} deltaavalon.com> wrote:

Dear Sirs,

this mail is a reply to a number of e-mail messages I got from [...], subsequently from Dr. K. Meyl and from Mr. S. Jackson about longitudinal waves ("L.M.D.s").

The matter "L.M.D"s. is not new at all but there is still too much discussion around it from different point of views, i.e. mathematics/physics calculus and electrical devices; this has slowly but definitely led to a total chaos even on the Web, resulting in misunderstanding and disbelief, which doesn't help anybody;

therefore, the best approach for an engineer is to be very straight and careful especially when someone talks about Nikola Tesla and theories about aether and L.M.D.s., still not accepted by the "Academic World" who is as happy to hear such arguments as could be "a sailboat in a forest" so to speak and who brings in turn only useless negative criticism and hectic hysterism instead of helping clarify.

This is why I prefer to proceed "step by step" scientifically replicating some Tesla apparatus and demonstrating his experiments, which is of course not easy because he was the Mozart of the art.

I will present a book about the above and will very soon attend a Meeting in Italy to talk about L.M.D.s and show some working devices to public: it is always a pleasure to demonstrate FACTS and WORKING DEVICES rather than only showing some slides and theoretically discuss matters as "Lorenz gauge vs. Coulomb gauge", "quaternions vs. vectors", "T.E.M. vs. L.M.D." etc. etc. Only facts count, observation of Nature and research: math must consequently follow, as always.

It is therefore more useful instead to bring something new to L.M.D.s tech and to the understanding of Tesla, like for instance my "Tesla power wireless" system with the "small car-demo" shown in my following vids on You Tube (& others under "HorizonDelta"): https://www.youtube.com/watch?v=0BGFF4ltEv and https://www.youtube.com/watch?v=dvZH3RCdeVw

also a tech e-book "Longitudinal dielectric waves in a Tesla coil and quaternionic Maxwell's equations" can be found on my Website for download:

https://www.deltaavalon.com/immagini/L.M.D. waves in a Tesla Coil revised.pdf

I hope the above helped positive discussion and knowledge.

Regards

Dr. Eng. Roberto Handwerker
DELTA Ingegneria

* * * *

The Laplace Operator

On March 31, 2011 2:06 PM MST, Konstantin Meyl wrote:

Subject: Re: About Scalar Field and an Italian Book

There is one comment, making clear, where the problem is localized:

Karl Palsness wrote:

> "You should read about Scalar .. it is a "Scalar FIELD" there should be no such thing as a "Scalar wave"; and there is a "Longitudinal Wave", instead of scalar wave. Scalar waves just don't exist according to the word scalar meaning size. A Tesla coil has a scalar field around it...and emits a longitudinal wave..."

If a scientist is using an expression, he first has to define, what the meaning is. This is what I do in my books, but in America, nobody seems to read my writings. I explain that the scalar wave is nearly the opposite of a scalar field. The expression is more than 200 years old and in mathematics exactly explained by the field equation of Laplace.

The Laplace Operator (describing the distribution of a wave in space) may be separated in 2 parts: grad div - curl curl. The second part is explaining the EM wave, a transverse wave, while the first part gives a scalar wave: The div applied on a vector is scalar! This is why scalar particles or vortices as part of the wave equation are propagating as a scalar wave. But they are propagating in the direction of a field vector and by this the wave is longitudinal and has a direction. In mathematics it is known, that the grad applied on a scalar is a vector again. This is why the scalar wave is propagating in one direction, not in a scalar way, as Karl Palsness wants the scalar wave to do. So he has to read the literature first before he is spreading nonsense in the world wide web.

My opinion about the education in the states is not the best. Lets help to improve it.

* * * *

Other Relevant Coverage

- Featured / OS: Wireless Transmission > Scalar > Tesla > Coils > Jackson >
  Tesla Coils for Dummies - Are you interested in building a Tesla Coil to produce stunning effects that violate the "established" laws of physics, but are somewhat less than technically savvy than a rocket scientist? Then you are in luck, because this article is for you! (*PESN*; April 29, 2011)
- Featured / OS: Wireless Transmission / Electromagnetic > Scalar > Solid State > Tesla > Coils > Jackson >
  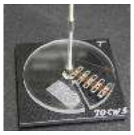
  Jackson's Wireless Aetheric Power Transmission Open Source Project - To facilitate the replication of these plans, as well as to characterize, optimize, improve, and come up with the zillion applications that can come from it, including: 1) harnessing aetheric energy, 2) superluminal communication, 3) wireless transmission of power through any barrier, 4) anti-gravity capabilities, and 5) creating defensive shields to make conventional war obsolete. (*PESWiki*; April 21, 2011)
- Featured / OS: Wireless Transmission / Electromagnetic > Scalar > Solid State > Tesla > Coils > Jackson >
  Tesla Coils: Unleash the Aether - Nikola Tesla's most significant contribution was not AC power, radio, or the induction motor, but what we call the Tesla Coil - a tool which allows for the power of the aether to be unleashed and harnessed. Now being launched through an open source project. (*PESN* and *BeforeItsNews*; April 19, 2011)
- Featured / OS: Wireless Transmission / Electromagnetic > Scalar > Solid State >

  Tesla's Scalar Field Still Beaming On! - The superluminal (faster than light) scalar or longitudinal waves Nikola Tesla used to magnify and wirelessly transmit power are not just a thing of the past. IEEE engineer, Steve Jackson, discusses and demonstrates how they can be utilized today, and he is open sourcing it here! (*PESN*; March 26, 2011)

See also

Resources at *PESWiki.com*

- **Directory:Konstantin Meyl**
- **Directory:Wireless_Transmission_of_Electricity**
- **PowerPedia:Wireless_transmission_of_electricity**
- **PowerPedia:Scalar_field_theory**
- **Directory:Solid_State_Generators**
- **PowerPedia:Solid_State_Generator**
- **Directory:Electromagnetic**
- **News:Electromagnetic**
- **More Stories for Hank Mills**

- **PESN (Pure Energy Systems News)** - Feature stories on cutting-edge, clean energy technology.
- **Free Energy News (.com)** - Daily cutting-edge, clean energy technology news from around the world
- **PESWiki Latest** - Newest feature pages in the publicly-editable energy directory.
- **This Week in Free Energy™** - Weekly live review by PESN, each Thursday 7 pm Mountain time
- **Free Energy Now (.net)** - in-depth interviews

AdChoices

PowerPoint Free

Tesla

Page composed by Sterling D. Allan March 26, 2011
Last updated September 04, 2012

NOTE: We are presently moderating all comments, to keep out the crass negativity, flaming, and other inappropriate forms of criticism. We welcome constructive criticism tactfully stated. Frequent offenders will be blacklisted.

"It is harder to crack a prejudice than an atom." // "I'd rather be an optimist and a fool than a pessimist and right." -- Albert Einstein

0 Comments Pesn 1 Login

Recommend Share Sort by Newest

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Be the first to comment.

Subscribe Add Disqus to your siteAdd DisqusAdd Privacy DISQUS

The "Down and Dirty" Guides to Making Biodiesel at Home
Making
BioDiesel Books
Save Time. Save Money... Help The Environment!

ADVISORY: With any technology, you take a high risk to invest significant time or money unless (1) independent testing has thoroughly corroborated the technology, (2) the group involved has intellectual rights to the technology, and (3) the group has the ability to make a success of the endeavor.

All truth passes through three stages:
First, it is ridiculed;
Second, it is violently opposed; and
Third, it is accepted as self-evident.
-- Arthur Schopenhauer (1788-1860)

"When you're one step ahead of the crowd you're a genius. When you're two steps ahead, you're a crackpot."
-- Rabbi Shlomo Riskin, (Feb. 1998)

Submit • Privacy • About • Contact

***PESWiki Departments*:**
Latest • News • XML Feed • Directory • Congress • Top 5 • Open Sourcing • PowerPedia

PESN.com

Copyright © 2002-2014, PES Network Inc.

The Tom Bearden Website

Subject: RE: Missing Reference Slide
Date: Mon, 9 Jul 2001 11:58:04 -0500

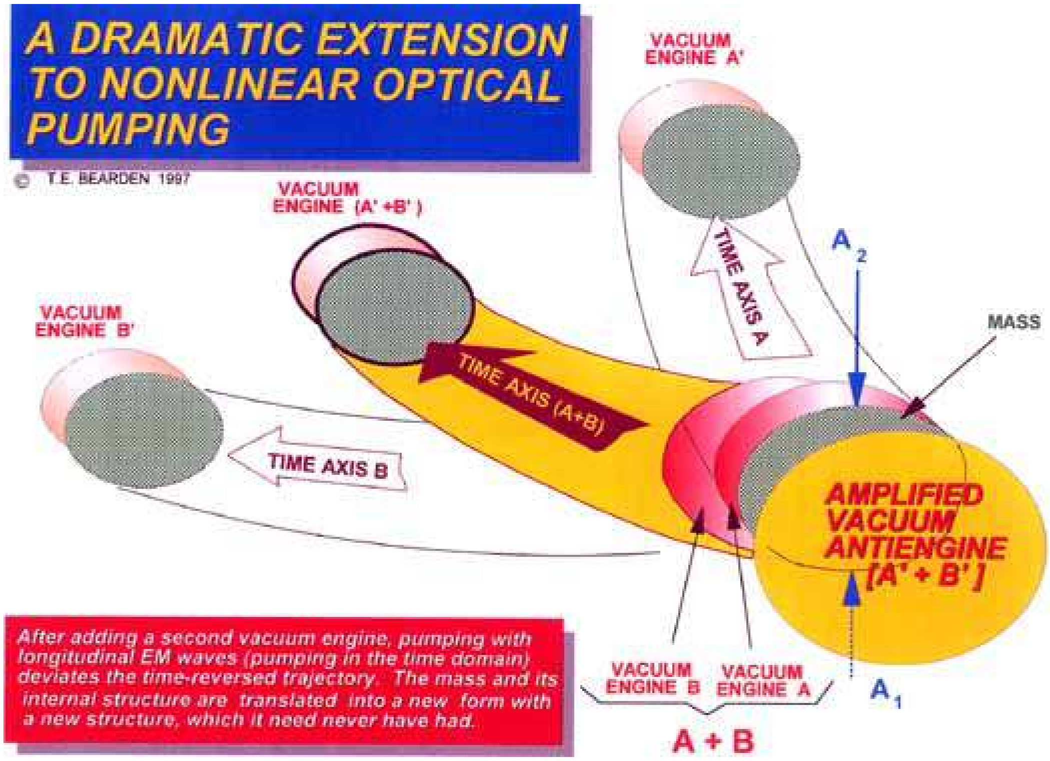

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

Yes, that is the slide. It works this way: When you time-reverse a cell, you effectively create a major anti-engine for the cell's exact "engine" (exact set of spacetime curvatures at all levels). Because of the precise one-to-one correspondence, this antiengine then "backs the cell up" along the original time-path it took in its physical functioning and changes, back to a precise earlier condition where the disease or disorder was not present. For a cancer cell, e.g., this backs the cell up past where it was originally "promoted" from a stressed or damaged cell to a tumor cell. That makes it into a "damaged cell". The antiengine continues to act, backing it up further into the state before the damage or disorder occurred. That then constitutes "healing" of that cell in a rigorous sense.

The cloning methods etc. used, do violence to the cell in the process. At the cellular level, damaging it or any part of it in the process changes the resident spacetime curvature engine set, to include the damage. You then "clone" a cell with an imperfect resident engine. The cloned cell then interacts with a large variety of phenomenology, continually altering various subcomponents of its resident engine by taking on additional subengines or changing some of its former subengines. Somewhere later, that festering "delta" in the resident engine -- after sufficient interactions with other environmentally-induced engines, will undergo interactions that "add" two deltas so something detrimental to the organism. At that point, physical damage to the cell erupts in "anomalous" fashion. Actually it erupts by perfectly normal fashion, once the internal engines and their interactions are understood.

This would appear to pose a problem anytime you *artificially* change the subengines of an organism, as in the case of genetically modified food. The small physical damage associated with splicing the genetics has associated with it the injection of such an artificial engine. The interactions of future environmental engines with the spliced genetic organism or plant then remains to be seen. Not all interactions would appear to happen in the limited test time usually allotted for testing the "effects". Anyway, that was the intention.

E.g., Pautrizel demonstrated that immature rats, having immature immune systems, would be restored by the Priore process back to the same immature immune system. The Priore process did not destroy the parasites themselves, but depended upon restoring the immune system back to its first ability to recognize the pathogen and go after it. Priore did not directly "kill" anything! In that case, the pathogens simply re-infected the immature rats, because their immature immune systems could not adequately protect them. It would be the intention of the research, in later generation equipment, to be able to design a specific "extra engine" and introduce it to the pumping, in addition to the "resident engine" in the subject to be treated. That "extra input engine" would accordingly generate a specific "extra antiengine" component in the total antiengine produced. When sufficient research and development maturity has been reached to be able to design such desired "extra engines" to cause the exact, desired extra antiengine to be produced, then the "extra antiengine" technique can be used, e.g., to time-reverse Pautrizel's immature infected rat back to an infected rat with a now-matured immune system. In that case, the rat's stronger immune system would make short work of the pathogen, just as analogously infected mature rats did.

To correct birth defects, etc., obviously one would have to change the actual time-path backwards, to a "state" that the body did not have physically, but which its genetics had before the damage. I foresee that eventually, after some years of development, the process would mature until it could regrow severed limbs, or correct dwarfism, etc. Of course that would be fourth or fifth generation, but it is definitely possible, given sufficient development.

Cheers,

Tom

AD 763210

QUITON/PERCEPTRON PHYSICS: A THEORY OF EXISTENCE, PERCEPTION, AND PHYSICAL PHENOMENA

by

Thomas E. Bearden

March 1973

Approved for public release; distribution unlimited.

THE SAM-D PROJECT
REDSTONE ARSENAL, ALABAMA

SURFACE TO
AIR
MISSILE
DEVELOPMENT

AMCPMMC FORM 1, 1 MAR 71

CONTENTS

Page

I. INTRODUCTION 1

II. EXISTENCE, PERCEPTION, OBSERVATION, AND PHYSICAL PHENOMENA 2

III. MASS AND SPACE 11

IV. IMPACT ON PHYSICS 16

A. Michelson-Morley Experiment 16

B. Gravitational Attraction 17

C. Newton's Third Law of Motion 19

D. Newton's Second Law of Motion 20

E. Newton's First Law of Motion 22

F. Identity of Length and Time 23

V. ELEMENTARY PHILOSOPHICAL IMPLICATIONS 25

A. Existence, Thing, and One 25

B. Reality, Ultimate Reality, Being, and Ultimate Being 27

VI. CAUSALITY AND ITS MODIFICATION 28

VII. CLOSING REMARKS 35

ABSTRACT

A fundamental answer is attempted to the questions of being, time, space, existence, perception, and physical phenomena. Mass, empty space, reality, gravitational attraction, continuity, discontinuity, and relativity are among the "things" and "nonthings" for which relationships and explanations are provided.

Perception is defined as detection by a mass of change to itself and as a differentiating process. Perceived physical phenomena are revealed to be first derivatives of ultimate reality, and outputs of the physical perceiving device call the "perceptron." Using the operation of the physical perception process on action, the units of which are energy times time, an entirely different view is generated of both reality and the Heisenberg uncertainty principle. When superposition is imposed as a condition, the Heisenberg uncertainty principle is revealed to be a quantized statement. A concept of the absolute value of a dimensional molecule is advanced.

A reality model is constructed of an action continuum and a nonaction continuum separated by an operational threshold whose magnitude is given by an uncertainty principle. Perception is modeled as a process which switches atomic quanta of action and nonaction back and forth across the threshold. Mass is explained as perceptron switch activity rate, and a mass definition equation is generated. One kilogram mass is shown to represent 17.053 x 10^{50} perceptron switches per second. From the model, Newton's laws of motion and gravity can be generated directly from the concept of the interaction of mass with space itself, space being taken as a massless fluid, or ether, composed of tiny nonaction particles called "quitons." Einstein's postulates of special relativity are also shown to follow from the model, and a statement of the equivalence principle is shown to be consistent with the model.

The concept of "lineception" as a causal chain of perceptron operations is advanced. "Inception" is developed as a noncausal modulation of lineception and as a distinguishing feature of a living system. The possibility is raised that inception could provide an explanation of the noncausal, statistical, ultramicroscopic universe and yet require large ensembles of ultramicroscopic events to conform to causality. Inception is shown to lead to karma. Perception is shown to be causal and to generate causality itself.

Some elementary philosophical implications of quiton/perceptron theory are pointed out and briefly discussed.

I. INTRODUCTION

The unsolved problems of physics stem from the unclear nature of the physical perception of change, and their resolution lies in the analysis of perception as a physical process. The unsolved problems of metaphysics stem from the same source, and can be resolved by the same perception analysis.

Specifically, a physical detection system can detect changes to itself and nothing else. These detected internal changes are thus what a detector "sees" or "observes" or "perceives" as changes to its external environment, i. e., as its physical phenomena. Thus a mass, being itself a physical detecting system, must detect only changes to itself, and these perceived changes to itself constitute its observed physical phenomena. For that reason, these perceived changes are obviously entirely relative to (i. e., are part of) the perceiver (observer); more precisely, they are entirely relative to and part of the perceiver's mass, which is the detector that is doing the perceiving.

Since only changes are perceived by a mass, then mass's perception must be a differentiating process. Ergo, perceived physical phenomena are first derivatives of a higher, or more fundamental, reality. The nature of that higher reality is by definition unperceivable (perception differentiates or fragments it), and it involves the quantity "action."[1] Action itself is not perceivable; change of action is perceivable since perception differentiates action. If superposition is placed on the Heisenberg uncertainty principle as a required condition, then the uncertainty principle contradicts itself for all except integral or zero multiples of a quantum of action. Correction of the uncertainty principle by imposing superposition as a condition provides a statement of the basic operation of perception -- the detection by a mass of change to itself.

All mental perception of a human being regarding physical phenomena is received (i. e., is inputted to the mind) from a physical sensory apparatus whose primary ingredient is mass. Thus we may describe the perceptive mind as consisting of sensory outputs of mass perceptions; that is, the input to mental perception must be the output of mass perception, and this interface within the mind may be referred to as the "perceptive mind."

From the above foundation, a comprehensive theory of perception can be constructed, and a most unusual model of "reality" emerges. A variety of unsolved questions are then resolved by the model. The resolutions include, among other things, 1) derivation of the postulates of relativity; 2) definition of the nature of time and space; 3) the explanation of gravity; 4) the generation of causality itself; 5) resolution of the wave theory of light with the quantum theory of light; 6) the explanation of why ultramicroscopic phenomena are statistical yet large ensembles of ultramicroscopic phenomena are causal; and 7) validation of Mach's principle and the equivalence principle. A totally new and precise definition of mass and of being itself are also two unexpected results. The problems of metaphysics are also answered: metaphysics is elevated to an exact science without any tinge of dogma, and physics and metaphysics are united.

---

[1] Planck's constant, Heisenberg's uncertainty principle, the principle of least action, the energy of a photon, etc., are derived from action and perception's differentiation of action.

Next Page

II. EXISTENCE, PERCEPTION, OBSERVATION AND PHYSICAL PHENOMENA

We first define perception[1] as detection by a mass of change[2] to itself, and state that it is an operation performed by that mass. The physical apparatus that accomplishes this perception is termed a "perceptron," which thus may be an electron, an atom, or the sensory apparatus of a living being.

A change detector does not detect the state of its input, but rather detects changes in that state. So in this sense a change detector can be regarded as a state differentiator. Therefore a perceptron differentiates, i. e., it perceives changes in its input and generates outputs in accordance. When its input is constant it will have no output. Thus zero output of a perceptron does not necessarily mean zero input.

What is to be fed into the perceptron's input section exists; i. e., existence (ultimate reality)[3] is defined as what is to be fed into the perceptron's input. The output of the perceptron is that which is perceived, or what we ordinarily call "physical phenomena."[4] Since the perceptron differentiates its input state, then physical phenomena are first derivatives of ultimate reality, which precisely accounts for the sharp separation between observable and nonobservable, i. e., between macroscopic and microscopic.

Any change in a perceptron's input section must involve both time change, Δt, and energy change, ΔE. So we may assume their product, $\Delta E \Delta t$, has always occurred in a perceptron's input whenever it produces an output. The output will be a derivative of the input, e. g.,

$$\frac{\Delta E \Delta t}{\Delta t} = \Delta E \tag{1}$$

and

$$\frac{\Delta E \Delta t}{\Delta E} = \Delta t , \tag{2}$$

where in each case the constant of proportionality has been ignored.

---

[1] We differentiate between perception and realization. Perception is entirely a physical process, and causal. Realization is a mental process and is not restricted to causal relationships.

[2] Reflecting, one notes that mass can only change in a nonmass manner. It cannot "change" without involving space and time, both of which are "nonmass." Therefore a mass change must be composed of both mass and nonmass.

[3] "Ultimate" in that it is higher or more fundamental than perceived reality.

[4] Changes to mass and only changes to mass constitute physical phenomena. Changes to one mass constitute "its" perceived physical phenomena, or its physical universe.

[4] Changes to mass and only changes to mass constitute physical phenomena. Changes to one mass constitute "its" perceived physical phenomena, or its physical universe. Since the other masses in its universe are perceived by it (i.e., created in its perception by its perception process), then these masses must be entirely relative to the mass perceptron that perceives them. Ergo, perceived mass is relative to the perceiver.

Next Page

Any detector has an operational input threshold, i.e., a minimum discernible signal. The Heisenberg uncertainty principle states the requirement for such energy-time changes:

$$\Delta E \Delta t \geqslant \hbar/2 . \tag{3}$$

Its smallest value, or

$$\Delta E \Delta t = \hbar/2 , \tag{4}$$

is then the input threshold for a physical perceptron to operate and produce an output.[1] Letting $k = \hbar/2$, we can write this simply as

$$\Delta E \Delta t = k . \tag{5}$$

Since existence has been defined as that which is fed into the perceptron's input, then that existence exists whether or not the perceptron operates, and the case[2] where

$$\Delta E \Delta t < k , \tag{6}$$

also exists (i.e., can be present in the input of a perceptron) but cannot be perceived or detected (i.e., the perceptron will not operate and produce an output).

At this point we diverge to remark that the product of energy and time is a quantity called "action," and the units of action are

$$A \equiv \frac{ML^2}{T} \tag{7}$$

Thus the input state to a perceptron is composed of an action continuum,[3] and the perceptron differentiates action quanta to produce physical phenomena.

---

[1] We acknowledge additional uncertainty statements but for simplicity will use only one in this paper.

[2] Equation (6) exists and it is proper to write it down and search out its ramifications. Equation (3) is a limiting condition which applies only to mass. But since our physical experience must involve nonmass (e.g., photons, space, time, being), then nonmass considerations must be examined. A good beginning is to consider the entire field of nonmass possibility, i.e., that field of nonmatter changes defined by (6), which includes the entire subquantum domain. So we legitimately examine two conditions: Eq. (3) must define (or apply as a condition to) the mass phenomena (change, quantized) system, and (6) must define (or apply to) the nonmass phenomena (subquantum) system.

[3] We exactly define continuity as the failure to perceive difference (change), and

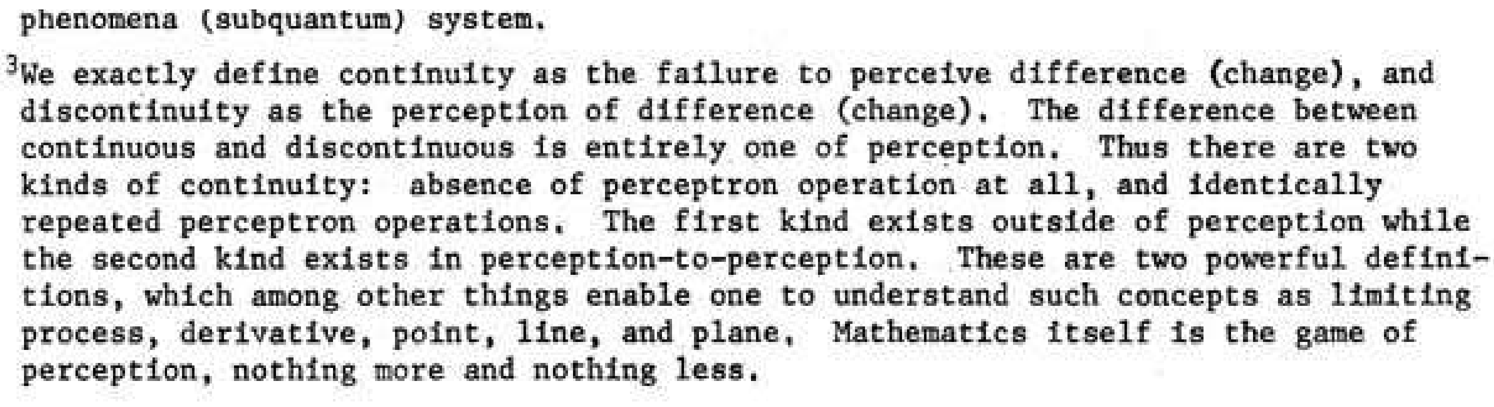

phenomena (subquantum) system.

[3] We exactly define continuity as the failure to perceive difference (change), and discontinuity as the perception of difference (change). The difference between continuous and discontinuous is entirely one of perception. Thus there are two kinds of continuity: absence of perceptron operation at all, and identically repeated perceptron operations. The first kind exists outside of perception while the second kind exists in perception-to-perception. These are two powerful definitions, which among other things enable one to understand such concepts as limiting process, derivative, point, line, and plane. Mathematics itself is the game of perception, nothing more and nothing less.

Next Page

When a large quantity of action is to be fed into a perfect perceptron, the input process cannot occur instantaneously; a Δt must be outputted (created)[1] by the perceptron for each operation. Thus perception processes steps or quanta of exactly k size, and exactly k size only. Inputs of less than k size are simply collected[2] until the aggregate is equal to k, then differentiated to produce an output. We must therefore rewrite the Heisenberg uncertainty principle as

$$\Delta A = nk \qquad , n = \pm 0, 1, 2, \ldots \quad , \tag{8}$$

where the plus and minus signs are necessary since action only has to change about some constant level in the perceptron input, and this change may be either additive or reductive (this quantum statement is actually the basis for quantum physics).[3] Thus, from the viewpoint of a perceptron, its output universe of physical phenomena is constructed by successive granular differentiation of unit operations, each of size k, but it detects no time lapse (creates no Δt) between operations.

When differentiated by a perceptron, an action atom may be regarded as being "chopped up" (the word "differentiated" literally means separated or torn apart):

$$\frac{\Delta E \Delta t}{n} = \frac{k}{n} \quad \bar{A} , \qquad n = 2, 3, 4, \ldots \quad , \tag{9}$$

where the bar over the A indicates nonaction, or action atoms which have been broken into bits.

Thus we now conceive of an action continuum of action atoms and a nonaction continuum[4] of bits of action atoms. The two continua coexist and the perceptron

---

[1] Or "extracted from an action quantum $\Delta E \Delta t$," which means exactly the same as "created." Einstein's treatment of simultaneity showed that it was operational, but if simultaneity itself is operationally created then both time and length must also be operationally created.

[2] This is not a simple discontinuous operation, since no time flow exists (Δt is not being outputted) during this collection process. We have to think of the process discontinuously, but the stepwise process is actually continuous (in time), or discontinuous, or both.

[3] Since ΔE may be positive or negative, we may state the uncertainty principle as $|\Delta E \Delta t| \geqslant k$, or as $\Delta A = nk$, $|n| \geqslant 1$. Consider the two cases $\Delta A > 0$, $n = 1$, and $\Delta A < 0$, $-1 < n < -2$. Superimposing them would give $|\Delta A| < k$, which violates the uncertainty principle. Thus, assuming superposition must hold, the uncertainty principle can be made to violate itself for all except integral values of n, and it follows that the Heisenberg uncertainty principle must be written as a quantized statement.

[4] The separation into two continua is artificial, but is necessary to allow us to perceive perceptron operation with our brain perceptron mind. There is absolutely

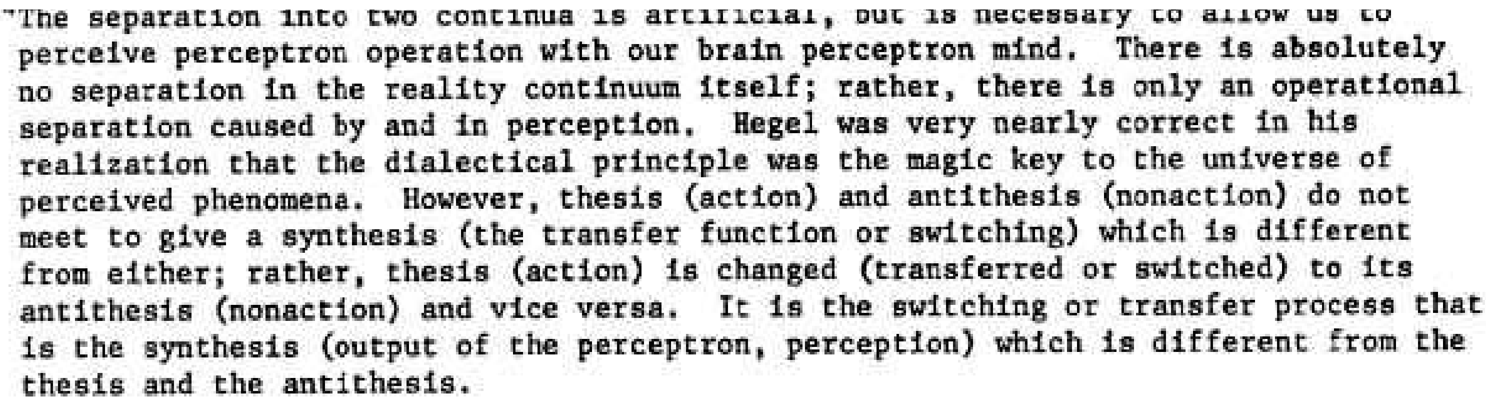

"The separation into two continua is artificial, but is necessary to allow us to perceive perceptron operation with our brain perceptron mind. There is absolutely no separation in the reality continuum itself; rather, there is only an operational separation caused by and in perception. Hegel was very nearly correct in his realization that the dialectical principle was the magic key to the universe of perceived phenomena. However, thesis (action) and antithesis (nonaction) do not meet to give a synthesis (the transfer function or switching) which is different from either; rather, thesis (action) is changed (transferred or switched) to its antithesis (nonaction) and vice versa. It is the switching or transfer process that is the synthesis (output of the perceptron, perception) which is different from the thesis and the antithesis.

Next Page

operates, or moves, in them. Only action quanta of exactly k size can be processed by the operating perceptron to give outputs which we know as physical phenomena. Such phenomena (events) consist of aggregates of extremely large numbers of perceptron operations. Action atoms will be referred to simply as "action atoms." The very smallest bits or fragments[1] of action atoms will be referred to as "quitons" and the symbol "q" used for one quiton. For convenience, we shall regard the "magnitude"[2] of q to be very much less than k, or

$$q \ll k \,. \qquad (10)$$

The perceptron's process of collecting quitons to form action atoms will be termed "fusion," which is an additive change to the constant action continuum. "Fission" will be the term for the perceptron's process of breaking up action atoms into aggregates of quitons, where the size of each aggregate is less than k; fission is a reductive change to the constant action continuum. The term "switching" will be used to refer to either fission or fusion, or both.[3]

Fission and fusion are merely transfers or switches across the k-threshold separating the action and nonaction continua, and may be represented by either of the diagrams shown in Fig. 1.

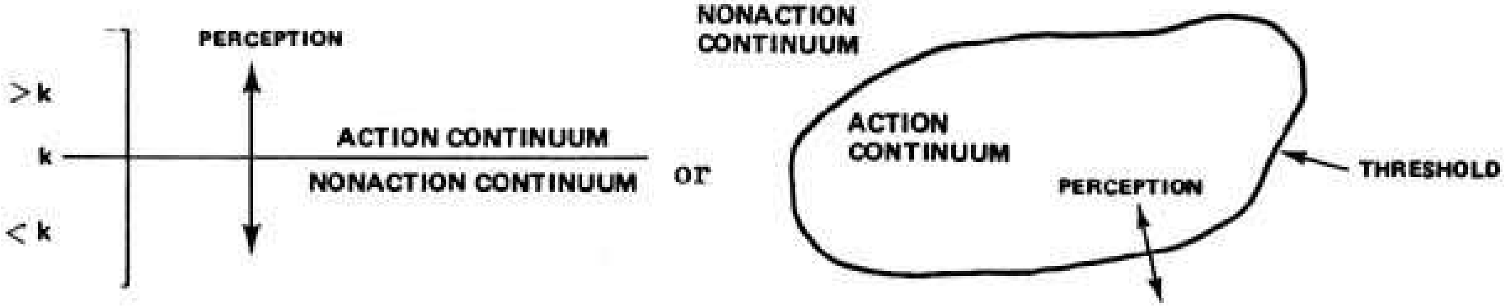

Fig. 1. Continua of the Universe

---

[1] These are actually nonthings and not bits or pieces at all, but since we can only perceive or think of things, we will model them as things. The difficulty is in the nature of thought, not in the nature of reality.

[2] The perceivable magnitude of a quiton is zero, i.e., it is not perceivable.

[3] And, in fact, switching can also refer to the transfer of Δt into ΔL, i.e., the creation of space from time or vice versa. Separation (relation) is composed of ΔL and Δt, and we know from relativity that when relativistic (perceived) ΔL is changed, then relativistic (perceived) Δt is also changed. When time dilates, length contracts; and when time shrinks, length expands. Therefore the statement that length and time are intertransferrable is valid: one kind of separation can be transferred into another kind of separation just as one kind of energy can be changed into another kind of energy. Since this transfer itself is an operation, then there is a rate (i.e., a constant) of transfer. This constant is actually a parameter, since by special relativity it can vary as a function of the mass changes themselves. And since we are dealing with mass changes (perception), then the rate of transfer

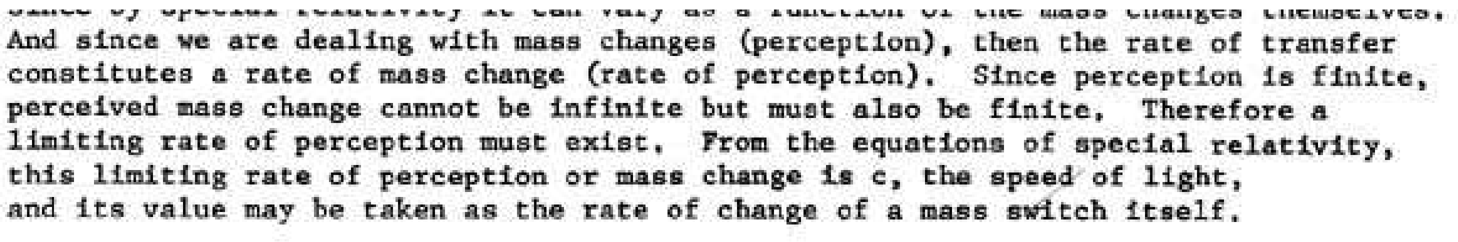

since by special relativity it can vary as a function of the mass changes themselves. And since we are dealing with mass changes (perception), then the rate of transfer constitutes a rate of mass change (rate of perception). Since perception is finite, perceived mass change cannot be infinite but must also be finite. Therefore a limiting rate of perception must exist. From the equations of special relativity, this limiting rate of perception or mass change is c, the speed of light, and its value may be taken as the rate of change of a mass switch itself.

Next Page

We now examine the expression for energy of a photon:

$$E = hf \tag{11}$$

Considering fission of an action atom as a chopping of that atom, the process can be expressed by:[1]

$$\begin{bmatrix}\text{Minimum}\\ \text{time rate of}\\ \text{change of action}\end{bmatrix} = \begin{bmatrix}\text{Energy}\\ \text{Transferred}\end{bmatrix} \simeq \begin{bmatrix}\text{atom of}\\ \text{action}\end{bmatrix} \begin{bmatrix}\text{cuts or chops}\\ \text{per second}\end{bmatrix}, \tag{12}$$

or

$$\frac{\Delta A}{\Delta t} = \Delta E = C_1 kf , \tag{13}$$

where C_1 is a constant of proportionality. According to Eq. (1), the energy perceived to be expended (transferred to quiton disturbances) in fissioning of an action atom ought to be proportional to the speed of the chopping action. Neglecting constants of proportionality, Eq. (13) is simply[2]

$$\Delta E = hf , \tag{14}$$

where the Δ symbol is deliberately added to show that it is a transfer process. The equation works in reverse for fusion of quitons into action atoms, i. e., the energy ΔE transferred to each action atom assembled is proportional to its rate of assembly.

We can now represent a perceptron as shown in Fig. 2, which corresponds precisely to the two-pronged arrows shown in the diagrams of Fig. 1. A perceptron has two channels, fission and fusion, in which switching transfers between action and nonaction are involved.

By Eq. (2), one of the outputs of a perceptron is Δt. Since perception creates its own time interval, it only has its kind of time, i. e., positive time, to deal with. Thus only positive time is perceived, and what we think of as positive time flow is merely the successive Δt's created by perception.[3]

---

[1] Note how admirably the view of tearing apart or differentiating the action atom describes the result. The smaller the Δt torn from (ΔE Δt) (or the quicker the separation), the larger the ΔE left (produced), i.e., the more energetic the separation. The atom of action is a quantum of action and one integral piece; the ΔE and Δt, however, are not necessarily of individually fixed size.

tion. The atom of action is a quantum of action and one integral piece, the ΔE and Δt, however, are not necessarily of individually fixed size.

[2] Action atoms can occur in more than one size (see note 1 on p. 3).

[3] This explains why we sense existence as if we were the point of contact of a pointer moving along a time line. In effect, perceptron time advances one time slice at a time, where the slice is of Δt thickness. Thus all physical phenomena occur in minute granular sequences.

Next Page

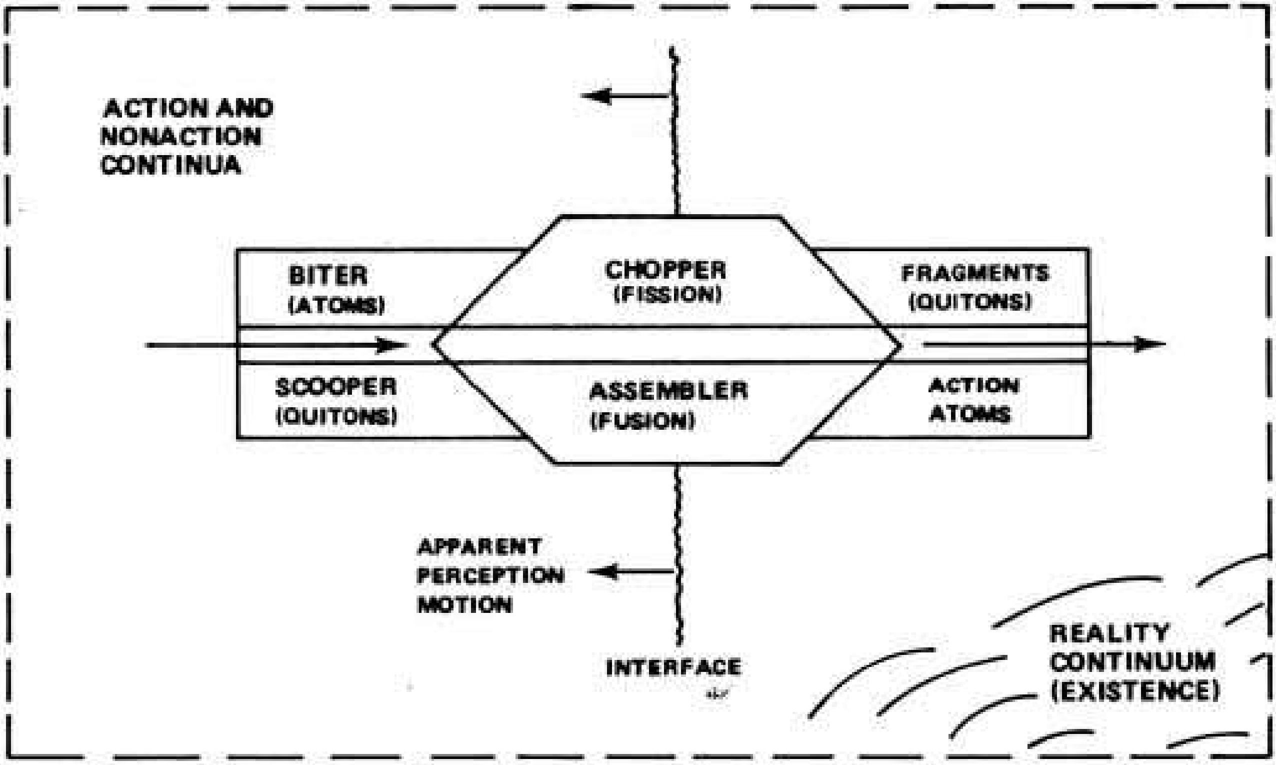
ACTION AND
NONACTION
CONTINUA
BITER
(ATOMS)
CHOPPER
(FISSION)
FRAGMENTS
(QUITONS)
SCOOPER
(QUITONS)
ASSEMBLER
(FUSION)
ACTION
ATOMS
APPARENT
PERCEPTION
MOTION
INTERFACE
REALITY
CONTINUUM
(EXISTENCE)

Fig. 2. The Fundamental Perceptron Switch

The existence of a negative time flow can be postulated by a simple symmetry analogy, but this postulated negative time, by definition, cannot be a perceptron output. We use the term "reflection" to denote a postulated operation in negative time[1] symmetrical to perception which operates only in positive time. Reflection is essential to the explanation of electric charge and charge effects,[2] but these will not be discussed here. Figures 3, 4, and 5 summarize the concepts of reflection, perception, and positive and negative time.

We now introduce the concept of absolute value of a dimensional molecule.[3] To illustrate, if we know that F = Ma, we will state that

$$|F| \equiv |Ma| \equiv |M| \times |a| = M\ |a| \quad , \tag{15}$$

where the last expression is true because of the nondivisibility of the unit ΔL and

---

[1] Modern theorists such as Feynman, Wheeler, and Dirac of necessity incorporate negative time operation into their theories involving action-at-a-distance.

[2] The quiton/perceptron theory can and does explain what "charge" is, why like charges repel and unlike charges attract, etc. It predicts the force of attraction or repulsion to be proportional to Q_1Q_2/R^2.

[3] Although perceptron output bits may vary in magnitude from operation to operation,

[3] Although perceptron output bits may vary in magnitude from operation to operation, they are of absolutely fixed size in any one operation or between any two consecutive operations. So for one operation or any two consecutive operations, a dimensional molecule is merely a product and division of certain-sized bits, and is itself of one fixed magnitude. Any other dimensional quantity of equal absolute value is therefore of the same size for that operation. At this level, differentiation is simple bit division and integration is simple bit multiplication.

Next Page

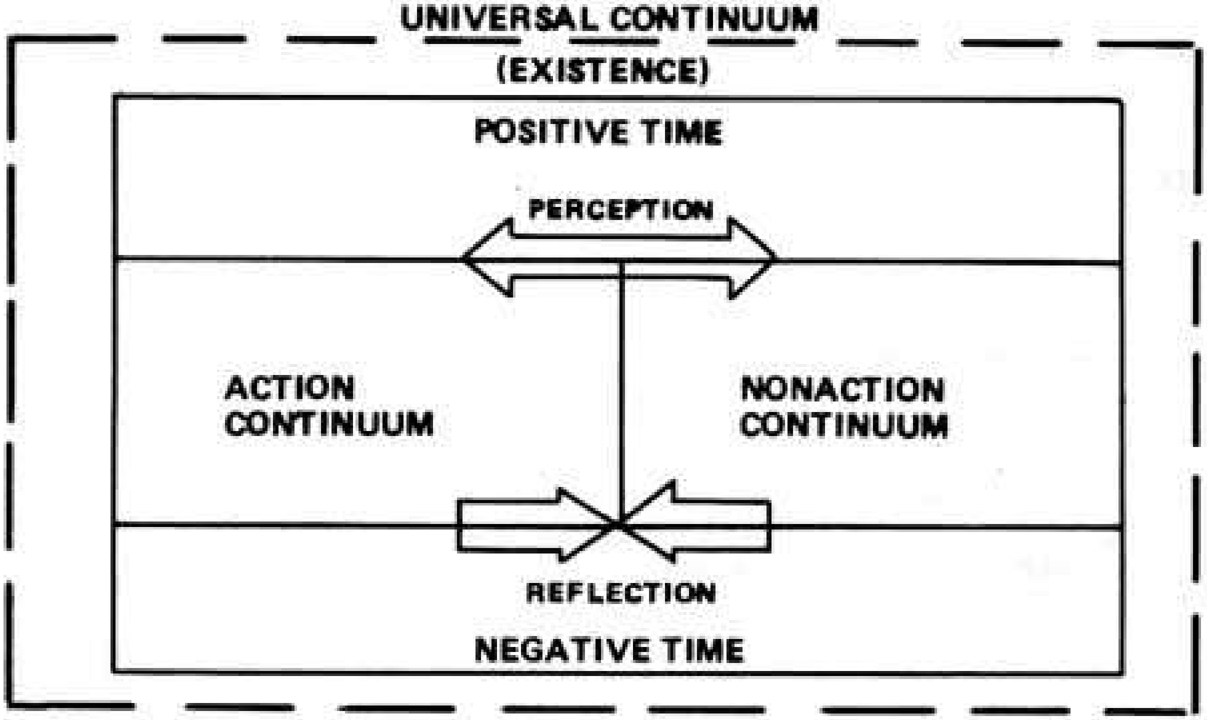
UNIVERSAL CONTINUUM
(EXISTENCE)
POSITIVE TIME
PERCEPTION
ACTION CONTINUUM
NONACTION CONTINUUM
REFLECTION
NEGATIVE TIME

Fig. 3. The Continua (Complete Model)

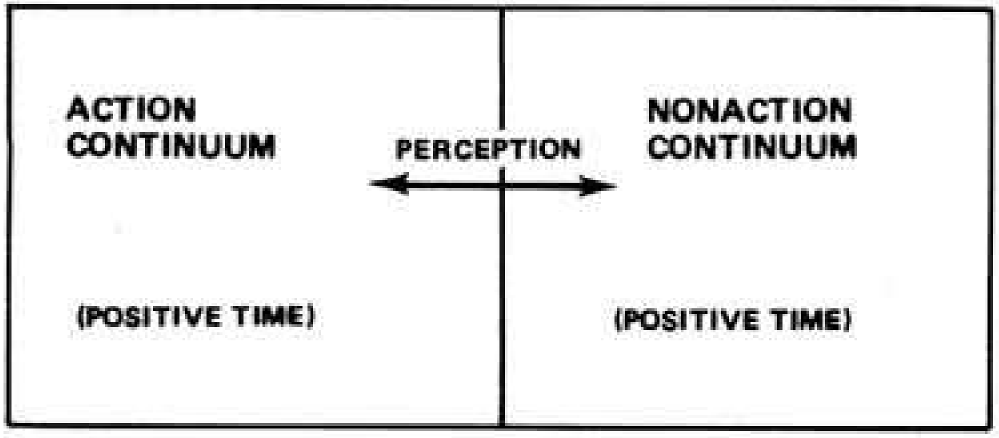
ACTION CONTINUUM
PERCEPTION
NONACTION CONTINUUM
(POSITIVE TIME)
(POSITIVE TIME)

Fig. 4. The Positive Time Continua -- Model in Which Perception Acts

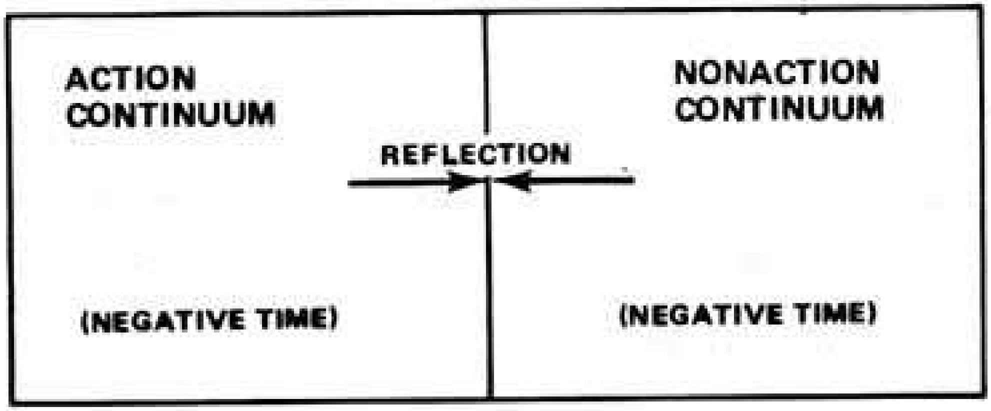
ACTION
CONTINUUM
REFLECTION
NONACTION
CONTINUUM
(NEGATIVE TIME)
(NEGATIVE TIME)

Fig. 5. The Negative Time Continua -- Model in Which Perception Does Not Act

Next Page

Δt quanta outputted by one perceptron operation. The absolute value signs must be added to any dimensional molecule which can take on either positive or negative values. Dimensionally,

$$M_0 = |E| \; , \tag{16}$$

where M_o is rest mass and requires no absolute value sign since a quantity of rest mass is only positive. Actually, since perceptron operation only outputs quanta, then the perceptron outputs ΔE and ΔM_o, or

$$\Delta M_0 = \left|\Delta E\right| \; , \tag{17}$$

and repetition rate allows perception of M_0 as a positive quantity. Then,

$$\frac{\Delta A}{\Delta M} \longrightarrow \frac{\Delta(L^2)}{\Delta t} \quad . \tag{18}$$

The perceptron expression of the right-hand side of (18) will be[1]

$$\frac{\Delta(L^2)}{\Delta t} \longrightarrow \frac{\Delta L \Delta L}{\Delta t} \longrightarrow \Delta L\,(V) \;\; , \tag{19}$$

and then we have

$$\Delta E \Delta t(V/V) \longrightarrow \frac{\Delta E(V \Delta t)}{V} \longrightarrow \Delta P(\Delta t)(\Delta L/\Delta t) \longrightarrow \Delta P \Delta L \;\; , \tag{20}$$

which must be another statement of the minimum condition of the Heisenberg uncertainty principle, or

$$k = \Delta P \Delta L \;\; . \tag{21}$$

It follows that

$$\Delta P \Delta L = nk, \qquad n = \pm 0, \pm 1, \pm 2, \ldots \tag{21a}$$

To consider why constant velocity yields special relativity, from $E = ML^2/T^2$ and Eq. (16), we can write

$$M = \left|\frac{ML^2}{T^2}\right| = M \left|\frac{L^2}{T^2}\right| \quad . \tag{22}$$

Dividing out the M gives

$$1 = \left|\frac{L^2}{T^2}\right| = \left|\frac{L}{T}\right| = \frac{|L|}{|T|} = |V| \ . \tag{23}$$

---

[1] Obviously $\Delta(L^2) = 2L\Delta L + (\Delta L)^2$, but the units of $2L\Delta L$ are $\Delta L\Delta L$.

Next Page

So velocity is a dimensionless quantity in the absolute dimensional sense, i. e., in the perceptron operational sense, and "the laws of physics apply equally well for all observers as long as they are moving with constant velocities" (Einstein's postulate). When the perceptron process is applied to a system of objects whose relative velocities are constant, the inputs to perception always differ by a constant value and the differentiated outputs of perception do not differ, i. e., the differential of $f(x) + C_1$ is identical to the differential of $f(x)$.[1]

Observation is related to perception as shown in Fig. 6. An observer observes only perceptron outputs[2] -- never ultimate reality -- from his sensory apparatuses.

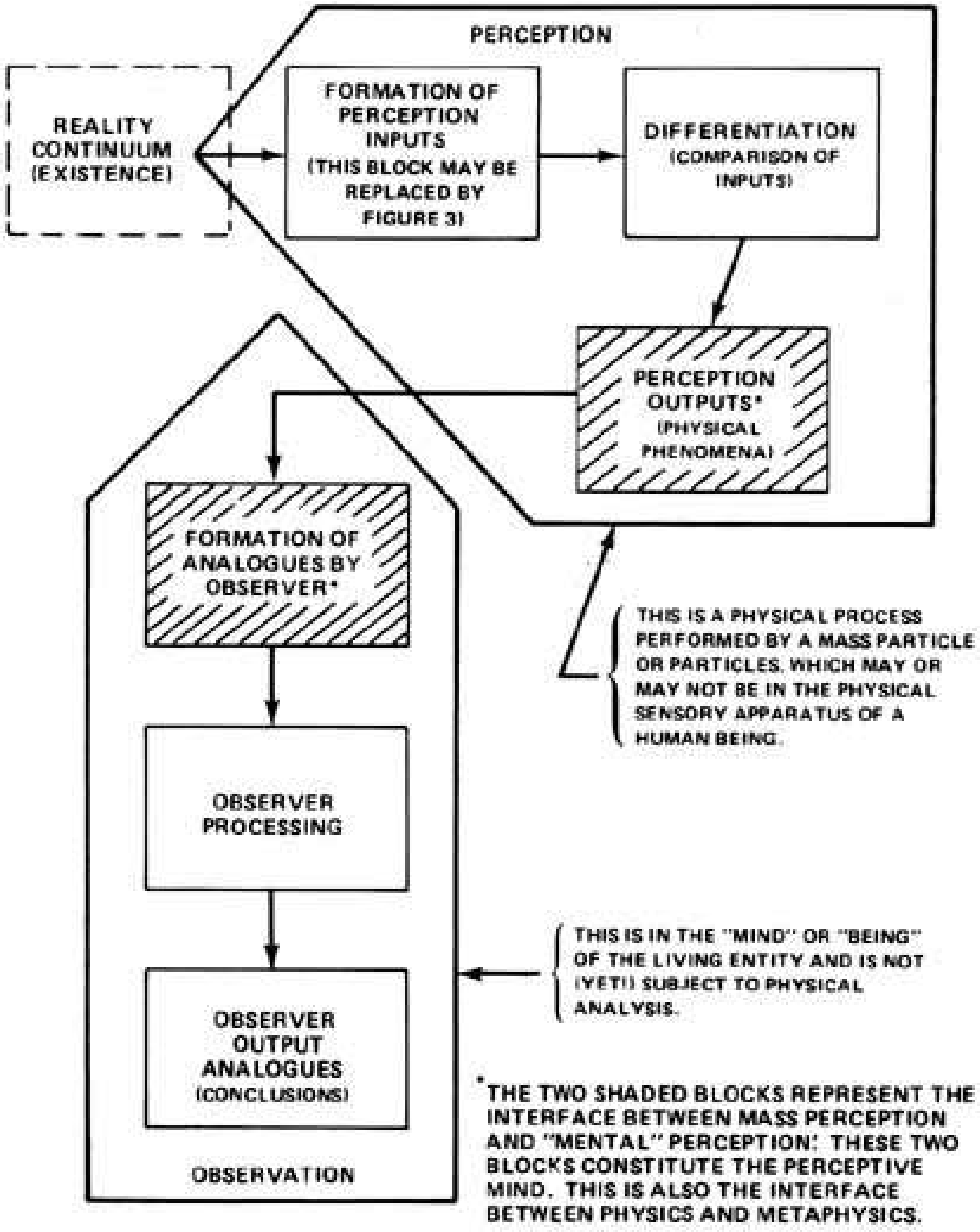
PERCEPTION
REALITY CONTINUUM (EXISTENCE)
FORMATION OF PERCEPTION INPUTS (THIS BLOCK MAY BE REPLACED BY FIGURE 3)
DIFFERENTIATION (COMPARISON OF INPUTS)
PERCEPTION OUTPUTS* (PHYSICAL PHENOMENA)
FORMATION OF ANALOGUES BY OBSERVER*
OBSERVER PROCESSING
OBSERVER OUTPUT ANALOGUES (CONCLUSIONS)
OBSERVATION

THIS IS A PHYSICAL PROCESS PERFORMED BY A MASS PARTICLE OR PARTICLES, WHICH MAY OR MAY NOT BE IN THE PHYSICAL SENSORY APPARATUS OF A HUMAN BEING.

THIS IS IN THE "MIND" OR "BEING" OF THE LIVING ENTITY AND IS NOT (YET!) SUBJECT TO PHYSICAL ANALYSIS.

*THE TWO SHADED BLOCKS REPRESENT THE INTERFACE BETWEEN MASS PERCEPTION AND "MENTAL" PERCEPTION: THESE TWO BLOCKS CONSTITUTE THE PERCEPTIVE MIND. THIS IS ALSO THE INTERFACE BETWEEN PHYSICS AND METAPHYSICS.

Fig. 6. Overall Relationships

---

[1] Einstein's treatments of simultaneity and special relativity may be regarded as

---

[1] Einstein's treatments of simultaneity and special relativity may be regarded as statements that one can determine a length (measure a ΔL) only by $\Delta L \equiv c\Delta t$, and one can determine (measure) a Δt only by $\Delta t \equiv \Delta L/c$ in the static case.

[2] The process of observation actually consists of both perception and realization.

Next Page

III. MASS AND SPACE

The nonaction continuum may be further visualized as a sea or gas of small, unperceivable particles, i. e., of quitons and aggregates of quitons which are less than k in magnitude. Since mass is a perceived quantity and a quiton cannot be perceived, a quiton or aggregate of less than k magnitude has zero mass.

Since fission and fusion both constitute perceptron switching, we may explain mass by visualizing a changing pile of quiton bundles, each of k size, in which the total rate of switching activity is what we call the "mass"[1] of the pile. Thus the mass is equal to the sum of the absolute values of the fission and fusion activities:

$$M \equiv \left|\frac{\Delta A_i}{\Delta t}\right| + \left|\frac{\Delta A_o}{\Delta t}\right| \simeq \left|\dot{n}_i + \dot{n}_o\right| \quad k \ , \tag{24}$$

or

$$M \simeq \left|\dot{n}_i + \dot{n}_o\right| \qquad \text{sw/sec}, \tag{25}$$

where[2]

Δt = perceptron operation time (one operation)

M = perceived mass of the pile

ΔA_i = action created from fu~~z~~sing quitons into action atoms

ΔA_o = action lost from fissioning action atoms into quitons

$\dot{n}_i$ = number of k-sized bundles per second fused into atoms

$\dot{n}_o$ = number of atoms per second fissioned into quitons

sw/sec = one perceptron operation per second.

---

[1] If ΔE is time rate of change of action, and if the absolute dimensional value of rest mass and rest mass energy are the same, then mass must also be a time rate of change of action. One kilogram mass equals 17.053×10^{50} sw/sec.

[2] The switch concept is necessary to explain charge and charge effects. The switching occurs between the positive (perception outputted) and negative (reflection outputted) time streams. Three types of switches exist, two of which are charge switches and all of which are mass switches. Switching may also be regarded as change of Δt into ΔL. For a mass switch, the limiting value of the conversion parameter b in ($\Delta L = b\Delta t$) is c, the speed of light. However, the

regarded as change of Δt into ΔL. For a mass switch, the limiting value of the conversion parameter b in ($\Delta L = b\Delta t$) is c, the speed of light. However, the concept of switching can be expanded; $b < c$ implies mass switching; $b = c$ implies quitons moving (maximum switching limit for mass switching); $b > c$ implies DeBroglie waves moving, which are not switching in the normal sense at all. Technically, the condition $b < c$ applies to Eqs. (24) and (25).

Next Page

The use of the mass definition equation can be illustrated by a simple example. First,

$$k \equiv h/2 = 0.527 \times 10^{-34} \text{ joule-sec} \tag{26}$$

$$1 \text{ kg mass} = 8.987 \times 10^{16} \text{ joules}\ , \tag{27}$$

so

$$1 \text{ kg mass} = \frac{8.987 \times 10^{16} \text{ joules}}{0.527 \times 10^{-34} \text{ joule-sec/switch}} \tag{28}$$

$$1 \text{ kg mass} = 17.053 \times 10^{50} \text{ switches/sec}. \tag{29}$$

For a mass that does not accelerate, switches must occur in couplets. This would require, for example, that angular momentum be quantized in quanta of 2k magnitude, since two switches would process two action atoms -- one in and one out of the mass action pile. Since

$$2k = h/2\pi\ , \tag{30}$$

then angular momentum should be quantized in $h/2\pi$ quanta, which indeed it is: from the Bohr theory of the hydrogen atom,[1]

$$mvr = \frac{nh}{2\pi} \qquad , n = 0, 1, 2, \ldots\ , \tag{31}$$

and the change permitted is thus

$$\Delta(mvr) = \frac{nh}{2\pi} \qquad , n = 0, \pm 1, \pm 2, \pm 3, \ldots \tag{32}$$

What we normally think of as empty space is sometimes described as the absence of mass and energy, existing in time.[2] Empty space can be defined by

$$\frac{\Delta A}{\Delta t} = 0; \qquad \frac{\Delta \bar{A}}{\Delta t} = 0\ , \tag{33}$$

which merely state that no perceptron output is occurring. Thus empty space exists only as defined by a perceptron and is merely the total absence of perception. While these equations rule out fission and fusion, they do not rule out the static presence of large numbers of quitons and quiton aggregates each of which is less than k in size (a quiton may be viewed as a quantum of space or a quantum of ether).

---

[1] Beiser, Concepts of Modern Physics, McGraw Hill, 1963.

[2] Actually there is no time connected with the existence of space; the time is connected only with our perception of change; that is, space is a product of perception, or the nonoperation of perception, and is related to a mass perceptron. Similarly, space itself has no length -- length is relative between two masses. That is why both length and time can change, and why a specific length and a specific time can

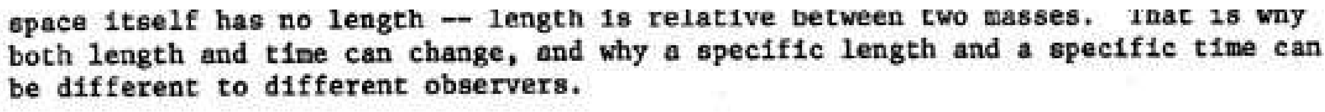

space itself has no length -- length is relative between two masses. That is why both length and time can change, and why a specific length and a specific time can be different to different observers.

Next Page

A massless quiton cannot individually generate any drag resistance against matter in the Newtonian sense. So a mass object moving along at a constant velocity will just push a stream of massless quitons ahead of it. In effect, the mass "bulldozer" is pushing its own medium along with it, and this requires no force.[1] Therefore, empty space (i. e., the zero output condition of one's mass perceptron) is filled with (has added to it) streams and streams of quitons moving ahead of other objects (perceived by one's perceptron) from all over the universe (perceived by one's perceptron). These quiton streams may be regarded as quiton pulses or quiton waves[2] in one's own static-relative quiton pond. The static-relative quiton pond of a mass will be called its "q-space" or "space" (more commonly referred to as the mass's inertial frame of reference).

When impinging on a perceptron's mass, these disturbances are being collected into k-sized quanta and switched into action atoms. The switching produces outputs that reveal the presences of the disturbing masses. The situation for a perceived mass is described in Fig. 7. Note that a fusion pileup is shown in front of the perceived object due to the object's motion, and a fission pileup is shown at the rear (switching is actually distributed throughout). Additional switching is occurring in the pileups and constitutes increase in mass. So one's mass perceptron perceives the moving object's mass as having increased. The relative velocity of the object with respect to one's perceptron determines the additional collection rate of quitons, and hence the additional switching rate, occurring for the object in one's perceptron q-space. The perceived increase in the moving body's mass is a function of the increased collection rate, which is a function of both the initial mass switch size M_0 and the velocity V, or

$$M = f(V, M_0) . \tag{34}$$

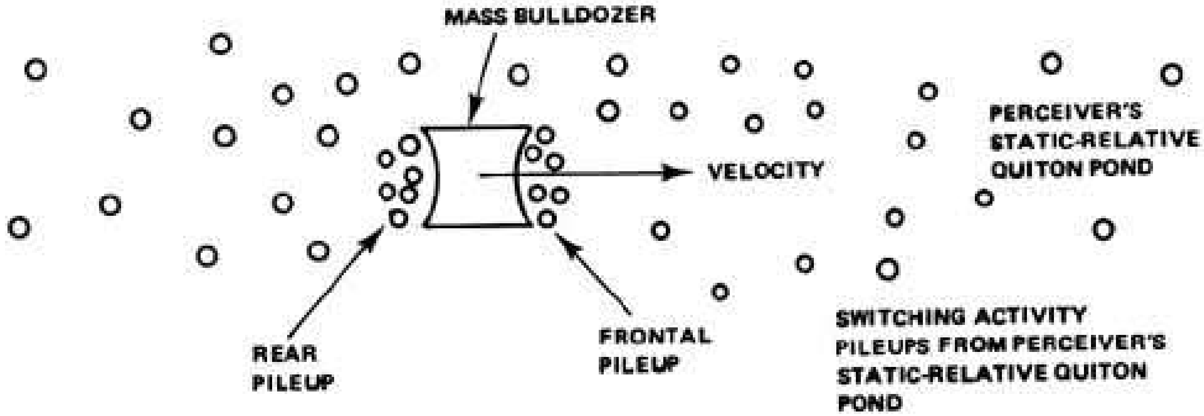
MASS BULLDOZER
PERCEIVER'S STATIC-RELATIVE QUITON POND
VELOCITY
REAR PILEUP
FRONTAL PILEUP
SWITCHING ACTIVITY PILEUPS FROM PERCEIVER'S STATIC-RELATIVE QUITON POND

Fig. 7. Increased Perceiver Switching Due to Perceived Velocity

---

---

[1] More precisely, it requires no resultant (unbalanced) force. Both fission and fusion create forces; however, these forces are opposing, and so when fission rate equals fusion rate the resultant force is zero.

[2] These pulses or waves are not in themselves perceived, so one can equally well consider them bunches of particles or waves. The difficulty lies with the perceptron mind, not with the disturbances.

Next Page

If the perceiver shifts "his view" to the object, he would at the same time shift his perceptron to it and so would then see the same sort of effect existing on his vacated mass. The pileup is relative to the perceiver and is occurring in the perceiver's perceptron itself.[1] So material objects do not just move through an ether, they carry their own q-space ethers right along with them.[2] And all they can detect is the disturbances in that q-space that produces switching transfers in and out of the action and nonaction continua.

We now wish to know the standard medium velocity at which wave disturbances move in the q-space. The question is, what has zero mass, is unperceivable, and yet can move through the perceiver's empty space ether[3] at some standard velocity? Photons satisfy all these requirements. Photons themselves cannot be perceived -- only their interactions can; and photons move in free space at the velocity c, the speed of light. Thus c is the quiton medium's wave velocity[4] and quiton disturbances move in a q-space at the velocity c. They move only in a q-space and in nothing else, since movement (velocity) is defined as $\Delta L/\Delta t$, and both ΔL and Δt exist only to the perceptron that created them. Therefore, velocity of a perceived object is absolutely relative to the perceiver since it is part of the perceiver's mass's perception. At this point, Einstein's postulate that "the speed of light is the same for every observer" can be seen to follow directly from the quiton/perceptron model.

But a difference exists between the speed of light in q-space and the speed of light in a medium. Simply speaking, when light travels in a medium the quiton waves (photons) are perceived to interact with particles (get switched by other switches) continually. In each imposed interaction switching perceived, another Δt occurs in the perceiver's perceptron. Thus more Δt's occur, and the photon is perceived to take longer in traveling through the medium than through free q-space; that is, the perceiver's perceptron (in which the medium itself exists) creates more time output in the case of light traveling through a perceived medium than light traveling through unperceived free q-space.

---

[1] This point is emphasized. The mass's mass does not increase as perceived by itself; its mass does increase as seen by the perceiver. This fact alone is sufficient to establish that the increase in mass is subjective, not objective (although these two terms are usually misused). That is what the term "relative" means, i.e., relative to the subject. Note that the perceiver is always the subject. The fact that the effect occurs only to the perceiver should indicate that it is the perceiver's perception process which is causing the effect.

[2] Thus one can have one space moving through another space.

[3] From Einstein's Theory of Relativity by Max Born, p. 224: "Einstein in later years proposed calling empty space equipped with gravitational and electromagnetic fields the 'ether,' whereby, however, this word is not to denote a substance with its traditional attributes. Thus in the 'ether,' there are to be no determinable points, and it is meaningless to speak of motion relative to the 'ether.' Such a use of the word 'ether' is of course quite admissible, and when once it has been sanctioned by usage in this way, probably quite convenient." The use of "ether" here is quite close to Einstein's proposal; however, in our usage it is meaningful to speak of "perceived motion of a perceived object" in the q-space ether of the perceiver. Otherwise, there can be no observation and no observer, which are of course necessary

"perceived motion of a perceived object" in the q-space ether of the perceiver. Otherwise, there can be no observation and no observer, which are of course necessary to Einstein's theory.

[4] Velocity is unperceivable; it is actually the constant of proportionality for the perceptron's direct transposition of Δt into ΔL. And c, the constant velocity of light, represents the limit of the perceptron's ability to do this.

Next Page

It should be noted, but will not be pursued further here, that electromagnetic waves in space are merely quiton wave disturbances[1] moving in a perceptron's q-space. The energy being transmitted is the "energy equivalence" of the nonaction quiton aggregates that were chopped up from action atoms. When these fissioned aggregates are again collected in some perceptron, then $\overline{\Delta E}$ is recovered as ΔE, or perhaps as $\underline{\Delta M}$.

In passing, we note that hidden in the principle of complementarity is the fact that wave and particle are operationally excluded, i. e., separated by an operation. A wave by definition is operational (dynamic). The most elementary idea involved in the concept of a particle is that it is nonoperational; i. e., it doesn't have to move, or change, or function, or be dynamic to exist. Obviously that which is operational differs from (is separated from) that which is nonoperational, but such separation itself is differentiation which is an operation. Therefore, wave (operation) and particle (nonoperation) are operationally separated. The only "mystery" in complementarity is that one must stop (interfere with) an operational entity to have a nonoperational entity, and one must start (interfere with) a nonoperational entity to have an operational entity. Depending upon the conditions imposed upon it, an entity can be either operational or nonoperational. For example, the operation of superposition must be imposed upon DeBroglie waves before they can constitute a particle. It then follows that the particle's velocity must in fact be the velocity of the superimposed group, i. e., the wave packet's velocity. By the same reasoning it follows that one cannot measure a particle's exact position (which is a static or nonoperational entity) and its exact momentum (which is a moving or operational entity) at the same time. The same is true for the particle's exact time position (nonoperational) and its exact kinetic energy (operational). Thus uncertainty principles, complementarity, and the relationship of DeBroglie wave velocity to wave packet velocity all involve the one fundamental concept that an operational entity is separated from a nonoperational entity by an operation. The changing is separated from the unchanging by a change; i. e., the changing is operationally derived from the unchanging, or perceived reality is derived from unperceived reality.

---

[1] Photons are aggregates of quitons; quiton aggregate behavior is particulate. Behavior from the standpoint of individual quitons is wave-like. This explains the dual behavior of light and reconciles its wave and quantum aspects.

Next Page

IV. IMPACT ON PHYSICS

A. MICHELSON-MORLEY EXPERIMENT

The Michelson-Morley experiment was conducted in the Michelson/Morley/earth q-space. The Michelson/Morley/earth perceptron(s) would perceive the speed of light to be the same in all directions, i.e., any fringe shift observed should have been a result of experimental uncertainty only. The experiment did not prove there was no ether; it did prove there was no perceivable ether[1] through which the earth was moving. The results admitted of a static-relative ether if it is unperceived. Perception is a process, and if the process does not occur then perception does not occur. Since by definition the q-space of a perceptron does not cause the perceptron to operate, then the q-space is unperceived by the perceptron.

More careful examination of observation reveals the following facts: A perceived mass has a constant mass when not moving with respect to the perceiver. When it is moving through the space frame or the ether of the perceiver with a constant velocity v, its perceived mass is again constant but is now perceived to be greater than was its rest mass. From whence came the extra mass?

The extra mass effect is only to the perceiver. The only possible sources for the extra mass in perception are the original rest mass M_0 and the empty space of the perceiver's space-frame (i.e., the interaction of perceived mass with unperceived space). Space does not drag, in the conventional Newtonian sense, against a moving body that is not accelerating. And space is unperceivable; therefore it can be visualized as a massless fluid. Thus the increase in perceived mass (which is a change) must come from the perceived mass's increase in velocity in perceiver space (because velocity is all that was changed), and it must involve the physical operation of the perceiver's physical perceiving mechanism. A perceived mass does exert resistance (drag) called "inertia" whenever it is accelerated in the perceiver's space-frame. This resistance is a force. Perceived force is equal to perceived mass times perceived acceleration through the perceiver's space.[2]

---

[1] Is not space an unperceivable ether? An extract from Einstein's Theory of Relativity by Max Born, p. 223, states: "If each of two observers who are moving relative to each other can assert with equal right that he is at rest in the ether, there can be no ether. Thus, the extreme development of the ether theory leads to its dissolution as a fundamental concept." The conclusion that there can be no ether does not logically follow. Each observer can assert that he is at rest in an ether (space), and these spaces can be intermoving. That is precisely what a "moving frame of reference" is: one space moving through another. Today we retain the (or more correctly, an) ether concept, but call it "space" or "inertial frame of reference" or "frame." The conclusion should have been that there can be no single "stable ether," but that there can be many.

[2] Newton's law, $F = ma$,* to date has been used in a half sense, i.e., that the acceleration of a perceived mass through the perceiver's unperceived q-space generates force. If this is true, and relativity holds, then the acceleration of any other unperceived q-space, through a perceived mass, also generates force; through one perceived mass in a perceiver's q-space, the q-spaces of all other perceived accelerating masses are accelerating and thereby generating forces on it. By this "other half" of Newton's law and by relativity, Mach's principle (that inertia of a mass is derived from the

accelerating and thereby generating forces on it. By this "other half" of Newton's law and by relativity, Mach's principle (that inertia of a mass is derived from the other masses of the universe) can be shown to be true. Although not collated and developed as such, it is possible to derive the postulates of both special (Continued)

Next Page

This equation states that a perceived accelerating mass must react with the perceiver's q-space, and in fact this interaction may be taken as the basic mechanism that creates force itself. Again the perceiver's physical perception mechanism is the causative agent.

A quite logical conclusion at this point is that the perceiving mechanism perceives (produces) additional perceived mass from the interaction of perceived rest mass with perceived mass-motion (velocity) through its unperceived space.[1] Also, the perceiving mechanism produces perceived inertial drag from the interaction of perceived mass with the mass's perceived acceleration through its unperceived space.

The concept of "ether" should not be rejected while the concept of "space" is retained, since the latter is actually an unperceivable ether. Rather, the concepts of mass, perception, and ether-space should be tailored to fit the perceived phenomena. Quiton/perceptron theory (perception theory) is a deliberate transfer function model specifically tailored to fit these criteria.

B. GRAVITATIONAL ATTRACTION

Newton's law of universal gravitation can be generated, in both effect and form, from quiton/perceptron theory.

A perceptron's q-space is filled with streams and streams of quiton disturbances moving ahead of its perceived moving objects from all over its universe. Thus there exists a flux of quitons at any point in its space, and this flux is isotropic to the first approximation: because of the vast interstellar distances between disturbing masses, the flux gradient in the relatively local neighborhood of any perceiver point is essentially zero. The quiton disturbance flux can be represented as shown in Fig. 8. A quiton flux stream switching pressure P_1 is exerted on M_1 and a switching pressure P_2 is exerted on mass M_2. In any one flux stream line in this shadowed zone, the pressure against either of the two bodies is reduced in proportion to the amount of switching flux subtracted from that flux line by the other body. The total reduction of switch pressure in the shadowed zone is therefore proportional to the product of the two masses M_1 and M_2, considering the effects to be mutually independent.

(Footnote 2 continued from preceding page) and general relativity from quiton/perceptron theory, and the steps essential to that derivation are contained in this paper.

*[More precisely, $F = d/dt\ (mv) = ma + v\ dm/dt$. From the second term, increase or decrease in mass of a moving object also produces a force. This second force component may not be too important except in microscopic phenomena, but for fast microscopic changes in mass (e.g., in the nucleus) it could generate significant forces. Also note that the velocities of all other spaces (i.e., velocities of the q-spaces of all other moving objects) in the perceiving mass's universe interact with its perceived mass change rates, according to the second term of Newton's law, relativistic form.]

[1] A disagreement has long existed over whether a "nonthing" exists if it cannot be observed. The argument is illogical; e.g., if a photon itself cannot be perceived

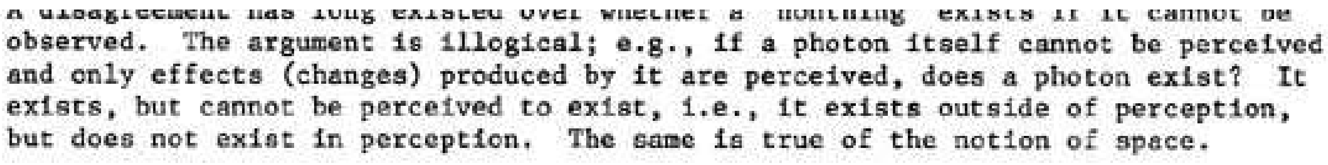

A disagreement has long existed over whether a "nothing" exists if it cannot be observed. The argument is illogical; e.g., if a photon itself cannot be perceived and only effects (changes) produced by it are perceived, does a photon exist? It exists, but cannot be perceived to exist, i.e., it exists outside of perception, but does not exist in perception. The same is true of the notion of space.

Next Page

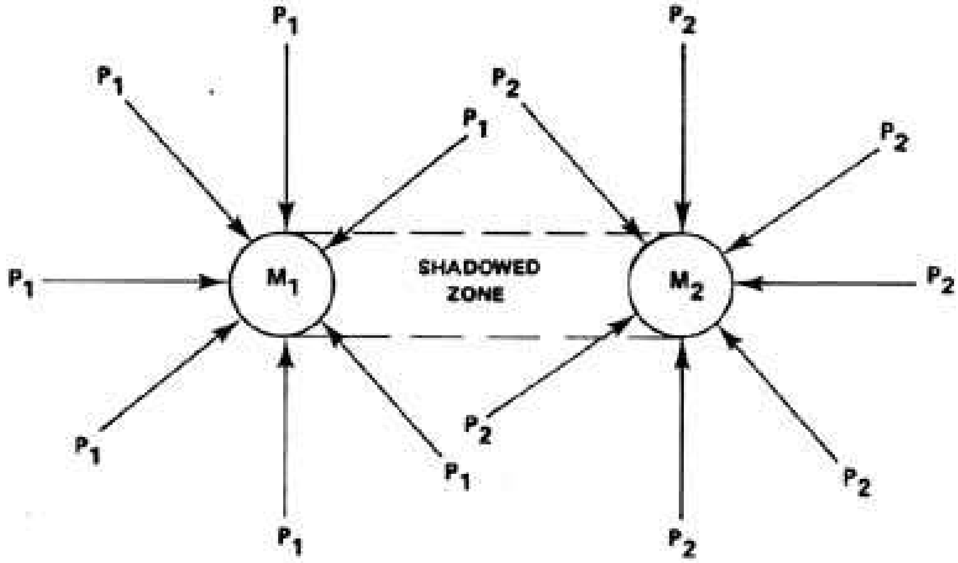

Fig. 8. Gravitational Attraction of Masses

From purely geometrical considerations, a shielding factor is also present due to the solid angle that is shadowed about any mass point. This effect is proportional to $1/R^2$ by usual solid angle considerations. Putting these two factors together, the attraction of mass should therefore be proportional to M_1M_2/R^2, or

$$F = K \frac{M_1M_2}{R^2} \qquad (35)$$

which is the expression for Newton's universal law of gravitational force, and this force is driving the two masses together.[1] By merely extending the basic statement by superposition of the effects of every perceived mass-pair in the perceptron's universe, one obtains the attraction of mass[2] for the universe.

---

[1] It is not accidental that electric, magnetic, and gravitational field forces all obey "inverse square of the distance" laws, since these forces are derived from analogous switching operations.

[2] Attraction of mass implies an effect derivable from the presence of mass only. Mass is a switching rate. Therefore attraction of mass must be derived from switching only. Now a massless fluid generates no drag against a constant velocity mass, but switching is a negative quiton acceleration which does generate force. So switching generates drag, but constant velocity does not. This accounts for the error long made in considering the "mass shadowing" explanation of gravitational attraction to suffer from a serious defect of implied retardation force due to velocity through an ether-flux (see, for example, the Feynman Lectures on Physics, Vol. 1, p. 7-9). Velocity of an ether flux through a body merely affects the mass of the body; it does not generate force. Acceleration of an ether flux through a body generates a resultant force on that body. In Feynman's example of an orbital body, let us assume perfect circular motion with constant speed. Since the speed of the ether flux is always constant, the mass is always constant. Since the component of acceleration tangential to the circle is zero, there is no force generated tangentially to the orbit. But since there is a constant radially outward acceleration of flux through the orbital body, then there is a constant radial force outwards that is generated, and this force is called the "centrifugal force," i.e., the "center-fleeing" force.

body, then there is a constant radial force outwards that is generated, and this force is called the "centrifugal force," i.e., the "center-fleeing" force.

Next Page

It should be noted that for extremely small masses at extremely close distances, or for extremely short times, turbulence and interference effects can form in the shadowed zone of Fig. 8 and superposition may no longer hold.

C. NEWTON'S THIRD LAW OF MOTION

One of the recognized ways of stating Newton's third law of motion is: "To every action force there is an opposite and equal reaction force." This law follows from quiton/perceptron theory, as can be shown by examining two colliding bodies. It should be stressed that these two bodies, and all changes related to them, exist as changes in the output of the perceiver's perceptron. It must always be borne in mind that these changes, or physical phenomena, exist as outputs of the perceiver's mass perceptron.[1,2]

Figure 9 shows perceived bodies M_1 and M_2 in constant velocity motion just before collision. A deliberate choice was made of a case where the momenta of the two bodies were not equal, so that the initial switching rates for the two masses were different. In the example, $M_1V_1 > M_2V_2$. The mass switching drag of each body is again represented by small circles. Each body has the same size roll in front as in the rear, so its switch pressure forces are balanced. Figure 10 shows the collision.

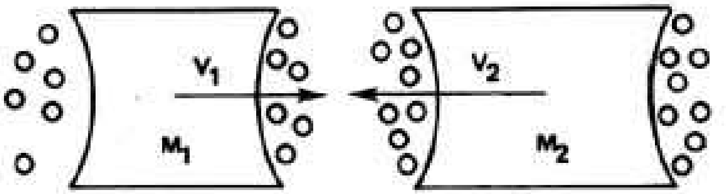

Fig. 9. Two Bodies Before Collision

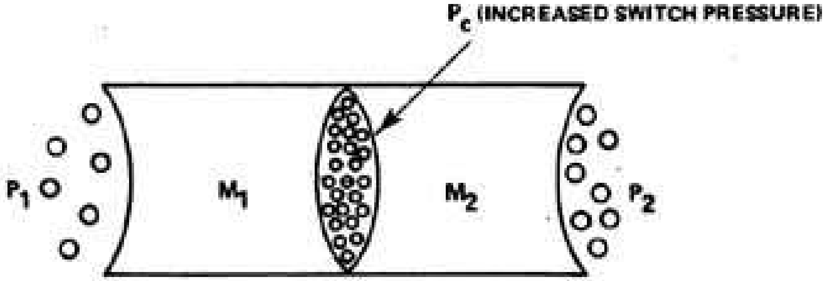

Fig. 10. Two Bodies at Moment of Collision

---

[1] We consider the question: "If a tree falls in a forest and there is no ear to hear it, is there a sound created?" A tree is a perceived tree, a forest is a perceived forest, and a fall is a perceived change. Their presence or existence implies a perceiver and perception; thus they imply that a "sound" change would be perceived, since they imply a perceiver to perceive it. If, however, by "no ear to hear it" we mean "no perceiver present at all," then the question as posed contains a self-contradiction: with no perceiver, there can be no perceived tree, perceived forest, perceived fall, or perceived sound.

[2] The manner of searching for quiton/perceptron causal explanation of a physical

or perceived sound.

[2] The manner of searching for quiton/perceptron causal explanation of a physical phenomenon is to ask, "What would the perceptron have to do to generate the phenomenon?".

Next Page

There is an almost instantaneous switch pressure increase in the contact zone between the colliding bodies, and consequently there is a switch pressure differential across each body as shown in Fig. 11.

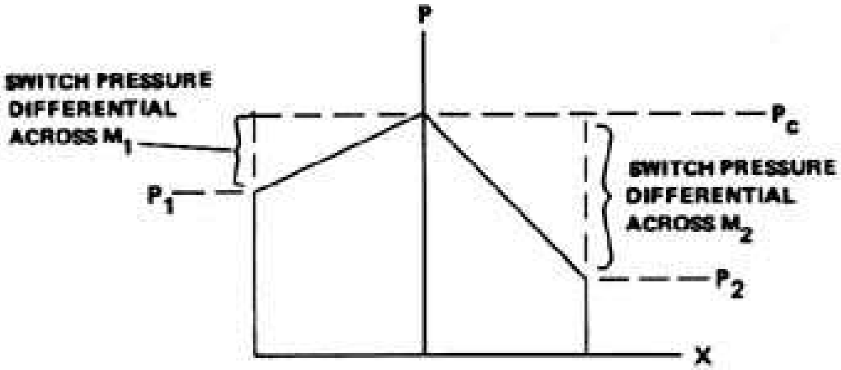
P
SWITCH PRESSURE DIFFERENTIAL ACROSS M_1
P_c
P_1
SWITCH PRESSURE DIFFERENTIAL ACROSS M_2
P_2
X

Fig. 11. Switch Pressure Relationships at Collision

In the figure, the contact pressure P_c is exerted against each body, so that the pressure forces on the two bodies act in opposite directions. But for any two colliding bodies, the contact areas, one against the other, are the same. This common contact area will be called A_c. Since $P \equiv F/A$, then

$$|F_1| = |F_2| = P_c A_c \; . \tag{36}$$

But since the pressure forces on the two bodies are being exerted in opposite directions, then

$$\vec{F}_1 = -\vec{F}_2 \; , \tag{37}$$

which is Newton's third law of motion.

D. NEWTON'S SECOND LAW OF MOTION

Neglecting constants of proportionality, Newton's second law of motion may be stated in the most general sense as

$$F = \frac{d(mv)}{dt} \; . \tag{38}$$

The first step in deriving this law is to examine a point mass which is perceived moving with velocity $\vec{V}$, as shown in figure 12.

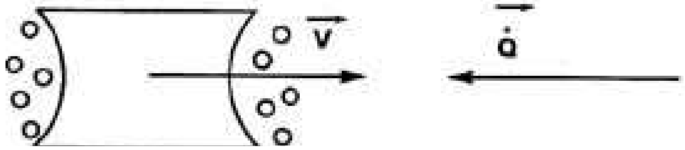

Fig. 12. Moving Point Mass

Next Page

An increased rate of quiton switching activity exists along the axis containing the velocity vector $\vec{V}$. We will call this increased quiton activity a vectorial "current" $\vec{\dot{Q}}$, and define its direction as contrary to the velocity vector. The mass is perceived to be moving against the current vector, and

$$\vec{\dot{Q}} \;\alpha\; -\vec{V} \; . \tag{39}$$

When the current $\vec{\dot{Q}}$ is steady, i. e., when

$$\vec{\dot{Q}}_{in} = \vec{\dot{Q}}_{out} \quad , \tag{40}$$

then the switching activity, and hence the mass and the energy, are constant. Therefore, from $v = \pm\sqrt{2E/m}$, the velocity remains constant.

If $\vec{\dot{Q}}_{out}$ suddenly increases, then

$$\vec{\dot{Q}}_{in} < \vec{\dot{Q}}_{out} \quad , \tag{41}$$

which can be represented as in Fig. 13.

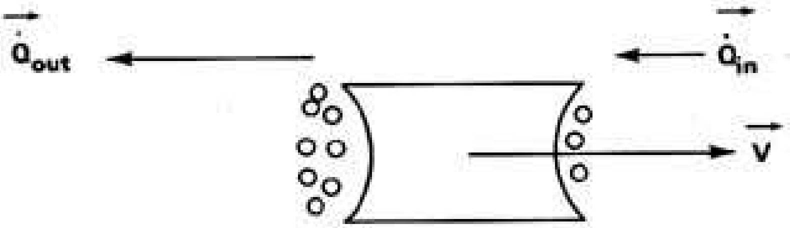

Fig. 13. Difference in Quiton Current

From Eq. (39),

$$\vec{\dot{Q}} = -k_1\vec{V} \quad . \tag{42}$$

At any time in a quiton current, the switch pressure force exerted against the mass can be represented as

$$F = P_c A_c \quad , \tag{43}$$

where P_c is the switch pressure and A_c is the area of contact. However, since the area of contact A_c, or the drag area, is the mass switch rate itself, then (neglecting constraints of proportionality)

$$A_c \;\alpha\; M' \tag{44}$$

$$A_c \alpha M' \quad , \tag{44}$$

where M' is the switch rate at the point of switch pressure application. We will next examine the switch pressure forces before and after one perceptron operation

Next Page

(i. e., across the smallest possible time interval Δt used by the perceptron itself in perceiving a change). When there is any difference in the pressure forces against the body front and body rear, this difference is the resultant force $\vec{R}$, or

$$\vec{R} = \left[(P_cA_c)_{in} - (P_cA_c)_{out}\right] \vec{q} \quad , \tag{45}$$

where $\vec{q}$ is a unit vector whose direction is given by the direction of $\dot{\vec{Q}}_{in} - \dot{\vec{Q}}_{out}$.
Now, from Eq. (44),

$$\vec{R} = \left[(M'P_c)_{in} - (M'P_c)_{out}\right] \vec{q} \quad . \tag{46}$$

Since $P_c \; \alpha \; \dot{Q}$ and $\dot{\vec{Q}} \; \alpha \; (-\vec{V})$, then (neglecting constants of proportionality)

$$\vec{R} = -\left[(M'\dot{\vec{Q}})_{in} - (\overrightarrow{M'\dot{Q}})_{out}\right] \quad , \tag{47}$$

which from Eq. (39) gives

$$\vec{R} = (M'\vec{V})_{in} - (M'\vec{V})_{out} \tag{48}$$

or

$$\vec{R} = K_4\Delta(M'\vec{V}) \quad . \tag{49}$$

However, this has happened in the smallest possible time Δt, and since by definition

$$\Delta(M'\vec{V})_{min\ \Delta t} = (\dot{\overrightarrow{M'V}}) \quad , \tag{50}$$

then

$$\vec{R} = k_4(\dot{\overrightarrow{M'V}}) = k_4 \frac{d}{dt}(M'\vec{V}) \quad , \tag{51}$$

which is Newton's second law of motion, relativistic form. Acceleration of a body is caused by a difference in mass switch rate (i. e., a switch rate gradient)[1] induced across it by transfer of momentum from another moving body or by induction from a field such as an electrostatic field or a magnetic field.

E. NEWTON'S FIRST LAW OF MOTION

Newton's first law, that "a body at rest or in uniform motion will remain at rest or in uniform motion unless some external force is applied to it," can be derived from the same type of considerations used to develop the second law. Obviously, if

$$\ddot{\vec{Q}} = 0 \quad , \tag{52}$$

---

[1] Switching can be thought of as transfer of Δt into ΔL. An increased switch rate represents an increased rate of transfer resulting in a higher value for the ratio

[1] Switching can be thought of as transfer of Δt into ΔL. An increased switch rate represents an increased rate of transfer resulting in a higher value for the ratio of ΔL to Δt being outputted. Since the ratio $\Delta L/\Delta t$ is defined as v, an increase in this ratio increases the velocity, and thus acceleration occurs.

Next Page

then

$$\vec{R} = 0 \tag{53}$$

and

$$\vec{a} = 0 \quad , \tag{54}$$

in which case the body at rest will not move and the body in uniform motion will remain so, by the definition of acceleration. Every perceptron repetition will output the identical value for V as the preceeding repetition, and that is why the first law holds.

F. IDENTITY OF LENGTH AND TIME

From $E = Mc^2$, we have

$$c^2 = E/M = 9 \times 10^{16} \text{ joules/kg} \tag{55}$$

$$c^2 \Leftrightarrow 9 \times 10^{16} \text{ joules/kg} \times 1/17.053 \text{ kg sec/sw} \tag{56}$$

$$c^2 \Leftrightarrow 0.527 \times 10^{-34} \text{ joule-sec/sw} \tag{57}$$

$$c^2 \Leftrightarrow \text{k/mass switch} \ . \tag{58}$$

Thus

$$c^2 = (3 \times 10^8)^2 \ \text{m}^2/\text{sec}^2 \Leftrightarrow \frac{\text{k}}{\text{mass switch}} = \text{one sw opn} \tag{59}$$

$$\text{one sw opn} \Rightarrow (3 \times 10^8)^2 \ \text{m}^2/\text{sec}^2 \tag{60}$$

$$\Rightarrow 3 \times 10^8 \ \text{m/sec} \tag{61}$$

$$\Rightarrow 1 \text{ sec} = 3 \times 10^8 \text{ meters} \ . \tag{62}$$

Length and time are thus synonymous,[1] and c is the constant of proportionality (more precisely, the identical ratio for transposing Δt into ΔL will be maintained for each perceptron operation). Therefore, a perceptron actually outputs only one thing which may be taken to be either Δt or ΔL (it actually outputs only change that abnormal (turbulent) switching can deviate from this limitation. Thus "time warping," "time travel," "hyperspace travel," and the interplanetary drive may indeed be possible. Since at least one of these effects of turbulence in switching (that of DeBroglie wave velocity exceeding the speed of light) is theoretically accepted, the problem of turbulent switching should be well worth the theorist's attention and effort.

---

[1] By "synonymous" we mean that Δt and ΔL are intertransposable by perception operation. The term c is the velocity (rate of change) of switching itself, i.e., the speed at which quitons themselves can move, since quitons are being switched by mass. Therefore c ought indeed to be the wave velocity at which quitons move, and thus represents the maximum ΔL obtainable from Δt by mass in normal switching. The possibility exists that abnormal (turbulent) switching can deviate from this limitation. Thus "time warping," "time travel," "hyperspace travel," and the interplanetary drive may indeed

Next Page

itself, which is perceived and thought of as ΔL and Δt). All other dimensions, quantities, and dimensional molecules are but bit division or bit multiplication. This seems to be consistent with some of the latest conclusions from general relativity theory.

In passing we note that there is no separation without relation, and there is no relation without separation. Therefore

$$\text{relation} \Leftrightarrow \text{separation.} \tag{63}$$

Also, there is no operation without separation, and there is no separation without operation. Therefore

$$\text{operation} \Leftrightarrow \text{separation.} \tag{64}$$

It follows that

$$\text{operation} \Leftrightarrow \text{separation} \Leftrightarrow \text{relation.} \tag{65}$$

We note a difference between free space and a Cartesian space (inertial reference frame). In a Cartesian space or inertial reference frame a definite length is considered to have been established for each and every point in that frame. But such a definite length to each point is rigorously operational by statement (65) above, i.e., such a length is defined by an operation. Specifically, it is defined by the operation of a mass perceptron (or an understood mass perceptron) at the origin of the frame. Such a frame is linear if the identical type of perceptron operation has defined the length to each point. If the type of defining operation varies, then the frame so defined is nonlinear; i.e., its space is said to be curved or distorted.

In one perceptron, its length defining operation and its mass defining operation can interfere or react one with the other. This type of interference or interaction is the direct interaction of frame space (length) with mass and it is the fundamental generating mechanism of force. All forces are so generated (see again the inset footnote on p. 17). And in fact a perceived moving force does work in the direction of its perceived movement rather than in the force's perceived direction precisely because of the moving operational length's interaction (interference) with the operationally generated force. This interference interaction is between operations in the same perceptron and generates what is called work.

Next Page

V. ELEMENTARY PHILOSOPHICAL IMPLICATIONS

Every scientific treatise is filled with forms of the words "one" and "be," yet none attempt to define these metaphysical terms. Ontologists and metaphysicians have uncovered many problems connected with these two words, but no real solutions. In addition, the historical preoccupation of metaphysics with dogmatic religious beliefs contributed strongly to its disfavor, since scientists were required to exert considerable effort to establish the scientific (observational) method as a legitimate criterion for the judgment of physical truth. Consequently, both science and philosophy today are constructed on quite undefined foundations, primarily because of the failure to recognize the relationship of perception operation to reality, which recognition is a prerequisite to the synthesis of physics and metaphysics. One therefore must comprehend the concepts of be (exist),[1] thing (perception), and one (thing) from the viewpoint of the relation between perception and ultimate reality.

A. EXISTENCE, THING, AND ONE

Existence has been defined here as "what is to be fed into the perceptron's input." In the situation where all operation of the perceptron has ceased, what can be said to exist in the input? First, what exists is a continuum: it has no limit, interface, dimension, time, velocity, change, space, place, or discontinuity since these are all derived from perceptron operations and the perceptron is not operating. What exists is continuity only, with no discontinuity.[2] To comprehend this, we must examine the thought process itself more closely, realizing that "to think" is "to perceive thoughts."

What is a "thing?" Further, what is "one" -- one "thing," one "anything?" A thing is first a perceived thing: it resulted from the output of a perceptron operation, in which one input bite and one output occurred. This output is characterized by the the fact that it is one, and the absence of one (i.e., zero) is the absence of output.[3] Figure 14 shows these relationships. The perceptron operation created the separation between the inside and the outside by creating a ΔL and a Δt, that is what the delta symbol means. Thus the operation of the perceptron determines the basic nature of one itself, and any one necessarily requires that a perceptron operation has occurred. In fact, the perceptron itself is a result of perceptron operations. Thus indeed all is relative. A perceptron exists and is able to function because it perceives other perceptrons to exist and function. Thus perceptrons are mutually created derivatives of reality are therefore unperceivable.

---

[1] Being is undifferentiated; that is its total definition. It cannot be perceived, since perception differentiates. It is continuity only, without any discontinuity, which is unperceivable. The preoccupation of being with causality (perception) is technically called "inception," the influencing of perception. In Zen philosophy this is technically "attachment to the world of illusion (change);" the literal meaning of "incarnation."

[2] This concept is unperceivable, by the definition of perception, which accounts for its difficulty. To be a "perceived one" implies limit (finiteness); oneness itself, without a bound or limit to make it a one (perceived), cannot be perceived. This is the primary limitation of the human mind (i.e., of perceptive thinking)--being, continuity, and oneness themselves imply absence of perception (differentiation) and are therefore unperceivable.

[3] Thus Boolean algebra describes the most basic operation, and is applicable to the

[3] Thus Boolean algebra describes the most basic operation, and is applicable to the "operational" world.

Next Page

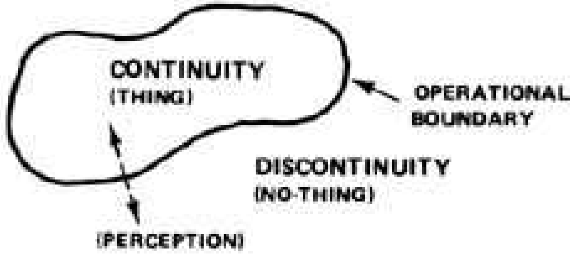

Fig. 14. Nature of One

which drive themselves and each other. Note that the preceding sentence itself is a perception.

So in the fullest sense, Gautama Buddha was indeed correct to state that the world of phenomena (changes) is an "illusion," relative only to itself (i.e., to perceptrons).[1] Einstein, expressing this idea in the language of mathematics, produced the theory of relativity and inaugurated a new physics. Buddha and Einstein spoke in different technical languages, but both expressed the same idea. The only difference is that in science we have assumed that our observations themselves are real and perfect. They are real, but not perfect in that they represent only changes in ultimate reality. Thus observational science alone can never reveal ultimate reality.

Another approach may aid in the comprehension of the nature of one. What is meant by the statement "he is the man?" Reflecting, one realizes that what is meant is that the perceived thing HE is exactly the same perceived thing THE MAN, and this identity exists now (during this one Δt perceptron time slice). There is no part of HE that is discontinuous from THE MAN, or vice versa. The different forms of the verb "be" attempt to express continuity without discontinuity, or unlimited oneness which is a state and not a thing. However, we do use the verb "be"[2] to relate continuity between different time slices. "He is the man" is related in the present time slice; "he was the man" means that HE in the NOW slice is identical to THE MAN in a past slice; etc.

---

[1] The reader interested in Zen will find complete agreement between this paper and Zen precepts, and will find the paper of some assistance in comprehending the unperceivable (unthinkable) Zen realizations. In Zen, for example, the aim is to comprehend at once the totality of reality, directly, escaping beyond the limiting fetters of perceptive thought. Now if it is actually possible to do this, and if some unusual men have done it, then their enigmatic writings should contain the wisdom of their experience, and should correspond to ultimate reality itself. Since ostensibly their insight is obtained by other-than-logic and other-than-perceptive means, then their statements should appear illogical and unperceivable.* Thus the "sound of one hand clapping" and other illogical concepts are used in Zen as a meditative means of driving the student beyond logical perception (certainly the sound of one hand clapping cannot be perceived or thought) to realization. But if the Zen experience is actually valid, then reality itself must be pure Zen (i.e., unperceivable), and so it would appear to be. We note that for each component part of all, the opposite part must exist. So all can be regarded as the complete set of all possible opposites or contradictions.

*(However, this most assuredly does not imply that every statement that is illogical and unperceivable is of value!)

[2] It should be stressed again that being is defined as undifferentiated (unseparated,

is illogical and unperceivable is of value!)

[2] It should be stressed again that being is defined as undifferentiated (unseparated, unperceived). Being is simply continuity without discontinuity.

Next Page

B. REALITY, ULTIMATE REALITY, BEING, AND ULTIMATE BEING

Actually one should be quite careful in using these terms. What does one get when one differentiates reality? One gets another reality, but not the primary or ultimate reality. For example, assume that length and time are real. When one time-differentiates length, one obtains velocity, which of course is real also but quite different from either length or time. However, velocity is always operationally related to the two first (assumed) reals length and time, because differentiation is an operation. So what is meant by "ultimate reality" is that state (or essence, or whatever name one prefers) that is not differentiated. Since "being" (continuity with no discontinuity) is defined as "that which is not differentiated," then the ultimate reality is pure being.

But all perceivables are differentials of ultimate reality. When one differentiates being, one gets pieces (differentials) of being which themselves are beings, but they have been perceived (differentiated) from a more primary or higher being. Therefore any being that one perceives has been differentiated (separated) by perception,[1] which explains why human beings, for example, consider themselves separate or individual (one). They are separate only in their individual perceptions. Thus "beings" by its plural nature always implies fragments (differentials) of a higher being (of singular nature). There can be many such integral levels: starting with one function, one can integrate it, then the resulting function can be integrated, etc., as long as we are speaking of definite integrals between two limits. When we reach a state that has no pieces or limits, the process is finished. That state is being in the ultimate sense, i.e., ultimate being.

One perceives these integral levels directly in nature, i.e., in the perceived phenomenological world. Living molecules (tiny beings) of RNA, DNA, etc., integrate together into a cell (a higher level being). Cells then integrate into a still higher being (man, animal, plant), which is the level on which man finds himself. But since man is perceivable and there are more men than one, it follows directly that men are capable of being integrated into a still higher being,[2] and the only word we have at present to describe such an ultimate being is "God." It is not at all accidental, therefore, that the ultimate being cannot be directly perceived by man, since by the definition of perception the ultimate being cannot be perceived by a lower order being such as man (i.e., since perception differentiates, or separates, then only differentials of being, or separate beings, are perceivable).

The avatar Jesus, having no technical concept for the derivative available in the language of the day, used the analogy of the father-son relationship, the son being the derivative of the father in this sense. Thus he referred to himself as the "Son of Man" (the derivative of all mankind or of integrated mankind) and also as the "Son of God" (the derivative of the integral of man). He continually referred to the ultimate being as "Father." It is little wonder that he was not understood then, and has not been understood for almost twenty centuries.

---

[1] Thus not only are physical objects relative to the perceiver's perception, but so are the perceived beings. Ergo the entire perceived universe, both living and non-living, is related to perception, and in fact perception and creation are abstractly

are the perceived beings. Ergo the entire perceived universe, both living and non-living, is related to perception, and in fact perception and creation are abstractly synonymous.

[2] This statement contains obvious implications to the theory (i.e., to the goal) of evolution.

Next Page

VI. CAUSALITY AND ITS MODIFICATION

A little closer look at perception can be obtained by viewing it from the causal standpoint.[1] All that perception really creates is Δt and ΔL (which are the same, ignoring constants of proportionality), and the perception we normally think of is point perception, where a point is one switching operation. The perception point may be thought of as following a line path which we will call "lineception" (Fig. 15). Physical phenomena as seen by one perceptron constitute this lineception; or rather, it constitutes the physical phenomena that are perceived or observed by the perceptron. Any point-based perceptron system, such as human perception, will accordingly perceive the universe of physical phenomena as if the perceiving system were a pointer moving along a lineception.[2]

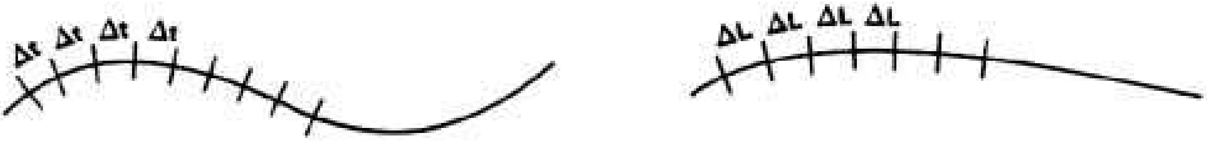

Fig. 15. Lineception (One Connected to One)

For each perception there is a reflection, so there exists (but by definition cannot be perceived) a similar line-path, or "lineflection" as shown in Fig. 16.

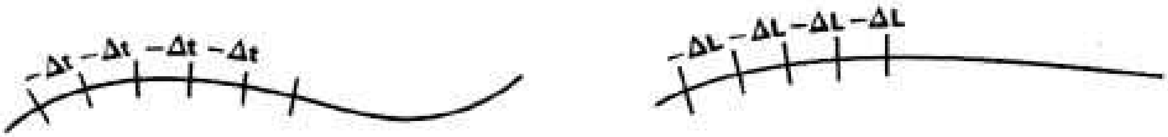

Fig. 16. Lineflection (One Connected to One)

causal, and perception generates causality (causality is perceptive). Thus the idea that perception generates causality is completely consistent with our initially assumed definition of causality as the set of all causes.

---

[1] Perception is identical to causality itself, i.e., perception can be said to generate causality. The causal view is that causes inputted to a situation operate to produce, or output or create, effects. According to its inputs, the perceptron generates causal outputs. An adequate theory of perception should be able to explain all causal (perceived) phenomena. We might define causality as the set of all causes. Then the answer to the question, "What causes causality itself?" is causality itself; i.e., since it is the set of all causes, causality must contain its own cause. Causality must therefore be closed or circular. The entire causal world must then be purely relative to itself (closed). Closer attention to causal operation yields the following: We perceive something change (in time or in space), then the situation changes so that we perceive new changes emerge (occur later, or separately). Thus the line of perception by its own unperceived operation connects the perceived cause (change) to the perceived effect (change), i.e., perception generates causality. Since perception itself is causal, then causality truly causes itself. This indeed is closed or circular--causality generates perception (perception is causal) and perception generates causality (causality is perceptive). Thus the idea that perception generates causality is completely consistent with our initially assumed definition of causality as the set of all causes.

[2] A "frame" or "coordinate system" is a point-centered perceptron system centered about the "understood perceptron" at the origin.

Next Page

The ultimate reality, or existence, contains all, all-at-once (ultimate reality is allness, or oneness). At each operation of the perceptron, the exact amount of Δt and ΔL, and their relationship, put out by a perceptron is a function of its input quiton disturbances. A great number, in fact perhaps an infinite number, of different input states are possible, i.e., exist in all, all-at-once (ultimate reality). So an equally large number of output states are possible (exist ultimately),[1] but one particular input condition results in one particular output condition. We can represent the all possible lineceptions existing as "multiception" (Fig. 17). From this figure, a lineception can be seen to be one particular path through multiception. The corresponding multiflection could, of course, be illustrated in similar manner but with negative components.

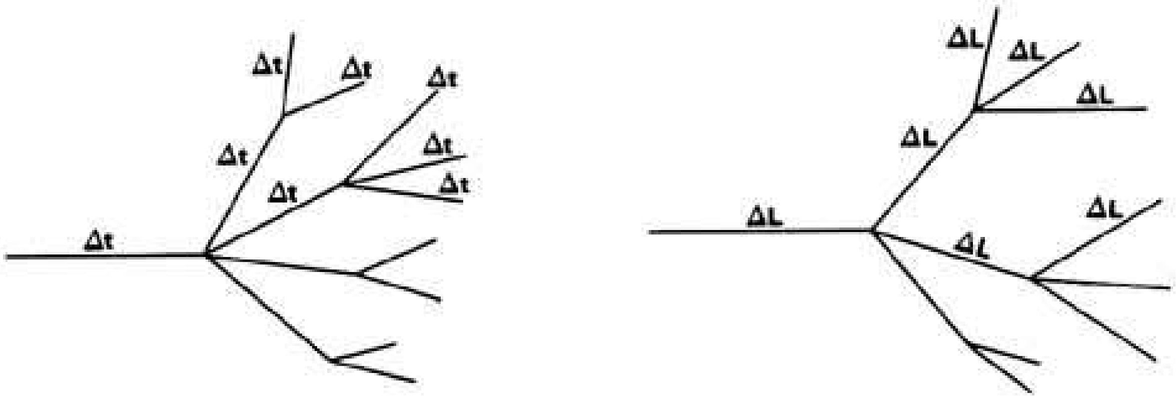

Fig. 17. Multiception (One Connected to Many)

One could immediately apply statistics to the concept along the lines illustrated by Fig. 18.

o

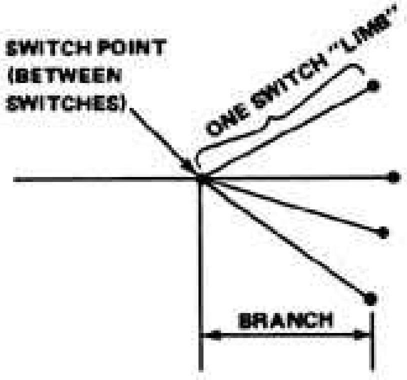

Fig. 18. One Branch

---

[1] It is stressed that these multiceptions exist, which is not at all the same limited statement as "exist now," but much more comprehensive; e.g., all possible numbers must exist in zero, else, given the number A, there is no justification for the axiom that $A + (-A) = 0$. The "absence of number" means precisely the "absence of all possible

exist in zero, else, given the number A, there is no justification for the axiom that A + (-A) = 0. The "absence of number" means precisely the "absence of all possible numbers." Paradoxically, if we think of that absence, we make it present by the act of thinking; i.e., trying to perceive it (absence is not perceivable). This contradiction, of having to make the unthinkable into its opposite so that it is thinkable is both the characteristic and the limitation of perception and perceptive thinking, and has long been an insurmountable obstacle to scientist and philosopher alike.

Next Page

At each switch point, the perceptron is faced with the branch of possibilities ahead, which all exist in multiception but only one of which exists in one perception, or lineception.[1] A certain set of limbs in the branch must have a perception probability equalling 1, and thus each of the limbs has some finite probability.[2] The input conditions to the perceptron then determine which limb is selected for the next switching operation; and, of course, the input conditions are the results of other related perceptrons' operations (relativity). The statistics will not be pursued further here.

For an absolutely mechanistic world, the entire process is automatic according to quite rigidly fixed rules. A system of perceptrons that behave this way shall be referred to as "mechanically switched." What we call "non-living systems" are mechanically switched systems. But what we call a "living system" is able to arbitrarily select the limb taken, to some extent at least,[3] i.e., the fact that a living system is able to choose its behavior within certain limits, directly implies that it must be able to deflect its lineception from a mechanically switched lineception. The only way it can possibly do this is to influence the switching operation of the perceptron. But since the output of one switching operation is a mechanistic function of the inputs, then the only way a life system can affect the perceptron switching operation is by affecting the input to some extent. How could it possibly do this?

In our crude model, the nonaction continuum is composed of action atom fragments, or quitons. But such non-things are massless by definition (they are not being switched and the absence of switching defines masslessness), and we can think of the nonaction continuum as a massless fluid or ether or empty space, since these terms are all synonymous. All that is implied by these terms is that the fluid is composed of massless elements, which merely means they are not being switched. They are capable of being switched, of course, but switching is not being done. The main point is that force, in the conventional sense, is not required to move a massless fluid, so the input fluid to a perceptron can possibly be interfered with by some nonphysical means requiring zero force. The being portion of a life system can apparently do this to some extent.

Since the switch time of a perceptron is on the order of 10^{-21} seconds, then even an incredibly tiny quiton input variation to a perceptron will result in

---

[1] It is stressed that all possibilities exist and are real, but do not exist in perception. Perception is the only difference between unperceived reality and perceived reality.

[2] Which leads to an unusual view of "probability." A perceptron operation is already finished when one perception is complete; that is why "the past is always gone, the present has just occurred, and the future is never here." Perceptron operation continually creates the perceptron's universe as it operates, including all changes, time, space, etc.). Thus, at the conclusion of any one perceptron operation, a multitude of lineception steps can exist for the next operation. However, although possibly quite large, the number of next steps available must be finite to be causal (perceivable). It is the process of perception that makes probability finite. Perception is finite; nonperception is infinite.

[3] The only alternative conclusion is that the living system is unable to deliberately

Perception is finite; nonperception is infinite.

[3] The only alternative conclusion is that the living system is unable to deliberately modify its behavior and its behavior is purely mechanical. But since this implies that all behavior is absolutely programmed and predictable, which is obviously false, this alternative is not valid and the original conclusion stands.

Next Page

noticeably altered lineception after one second of such variation. So it is quite reasonable to presume the futility of trying to measure such infinitesmal influence as is brought to bear on one single perceptron operation. However, the total influence can obviously be detected after being accumulated by a very large number of perceptron operations, and for living systems it is. So while we have not proved the mechanism of life-system manipulation of lineception, we have shown a logical progression from an assumed mechanism to a result that is universally verified by countless observations (i. e., all living systems are observed to manipulate causality, which manipulation is, after all, the definition of behavior). Since no other mechanism has been proposed, to the author's knowledge, that would reasonably explain the observed phenomenon of life-system manipulation of lineception, it is logical to accept as true the assumed mechanism -- a life system can exert an incredibly tiny influence on perceptron inputs,[1] i. e., on the ether or on empty space.[2,3]

---

[1] This tiny influence is the basis of the realization process. To comprehend the noncausal, the individual observer usually wrestles diligently with the problem, repeatedly attempting to perceive it causally. In each such attempt, the tiny influence of inception on perceptron inputs is exerted. Over an ensemble of such attempts, the perceptron causal output can be modified, i.e., modulated noncausally to give what are now slightly noncausal perceived outputs. Presence of a required fit or required output pattern in the perceptive mind allows the output to be examined (compared) against the required output, i.e., it allows the selection of the proper noncausally modulated causal transfer function. When the outputs match, one then has a breakthrough or realization which matches the real world pattern but which cannot be directly perceived causally. However, since the mental file of the perceptive mind has now filed the proper transfer function, the individual has assimilated the information and understands it, being able to recall the noncausally modulated causal transfer function at will. Thus in Zen, the ultimate example of realization is described as "enlightenment." Phrases such as "direct perception of reality," "direct experience of ultimate reality," or "direct realization of all-being" are also used in trying to portray the nonperceivable realization. Conventional science hides the entire process behind vague terms, such as "creative thinking," which actually only conceal the fact that we do not know what it is, or how it is accomplished.

Note that such "creative realization" can cover either purely causal or noncausally modulated perception, the difference being merely whether the inputs are incepted in a normal manner comparable to causal phenomena, or in another manner. Realization is derived only from inception. Since the observed phenomena of physics already appear to be noncausal (statistical) in the ultramicroscopic region while ensembles of these yet appear to be causal, then the physicist is already confronted with the necessity of changing causality but at the same time retaining causality. The implications to the observation process itself must be examined, and they remove the dichotomy between physics and metaphysics when pursued.

[2] The Hieronymus effect is an example of this in reverse. It generates an effect on living nervous systems that is not the result of electric, magnetic, or gravitational fields, and is not caused by matter in motion. For that matter, a polygraph has been used to clearly measure a response of plants to human thoughts alone, and such an effect is not presently explainable within the body of physics theory. "Firewalking" is another unexplainable phenomenon from the standpoint of present physics theory as is acupuncture.

[3] All psi phenomena derive from this effect (inception) together with turbulent

is acupuncture.

[3] All psi phenomena derive from this effect (inception) together with turbulent switching. A straightforward theory of psi phenomena (and of such things as firewalking and acupuncture, for that matter) can conceivably be generated from the quiton/perceptron theory.

Next Page

Of course the most efficient development of manipulation of lineception (i.e., the widest response) by a life system would necessitate a control system approach; i.e., life system direct manipulation of only a very small control valve or control mechanism, which in turn controls the input to a much larger mechanistic (deterministic) servo amplifier. Systems like this are universally observed in life systems. Thus the great bulk of a life system's manipulation of lineception is servomechanistic: the actual direct influence of life (being) on matter (perceptron switching) is so microscopic as to be immeasurable at its operating level, and only its macroscopic effects become measurable. For that reason the scientist has never been able to isolate the mechanism of life's manipulation of lineception, nor is he likely to be able to do so in the foreseeable future.[1]

Life's manipulation of lineception will be denoted by the term "inception," and can be thought of as a modulation (partial change or additive change) of the lineception carrier. Ordinary physical phenomena are macroscopic ensembles of lineceptions, where the deterministic (mechanistic) law holds absolutely but in a relative manner (ensemble to ensemble) which is fixed cause-and-effect, or causality, from the viewpoint of perception. Since inception is a modulation of lineception, then living behavioral phenomena should be superimposed as envelopes on causal lineceptions, or causal lineception ensembles, and, indeed, they are universally observed to be so superimposed.[2]

Absolute or ultimate reality, however, can be thought of in one respect as all multiceptions and all multiflections, all-at-once, which is timeless (i.e., which is both all-possible-times and zero time simultaneously or identically).[3] Along any one lineception, causality applies; yet since a very large number of changes to that lineception are available in just one multiception branch, then quite a large

---

[1] This latter statement, of course, is subject to challenge.

[2] Since all matter is, it contains being. Another intriguing possibility exists: that there are two types of inceptive systems, or two magnitude directions that inception takes. The first type would produce inception with difficulty, and its inception would only be apparent at the macroscopic level; the second type would produce inception with ease, its inception would be apparent only at the microscopic level. The first type would correspond to what we call "living systems," while the second would correspond to what we call "inert systems." Living systems' inception would be observable to human senses and human instruments, while inert system's inception would not. A second and remarkable consequence would be, that in the extreme microscopic case, inception would dominate and this scale would appear to be absolutely statistical (individually unpredictable), which apparently is true (i.e., degree of causality represents degree of restraint of inception). This would modify causality (for individual events) at the ultramicroscopic level, but at higher levels causality would still apply due to the karma effect on perceived ensembles. It is significant that at present the universe of matter seems to behave in precisely this fashion. To summarize, perception is causal -- it generates causality. Inception, being noncausal, must be below the perception threshold. Aggregates of inception (noncausal) that equal or exceed the perception threshold become causal (are processed by perception) and generate causal karma.

[3] Since every perceived ΔT has an unperceived complement $-\Delta T$ which exists in reflec-

[3] Since every perceived ΔT has an unperceived complement $-\Delta T$ which exists in reflection, then the sum total of all times in absolute existence, which includes both perception and reflection, equals zero time.

Next Page

number of inceptions are available to be applied to that one lineception by an organism to change or direct the lineception. Herein is the explanation of karma. Since every inception made by an organism (or by a submicroscopic particle of being for the second case, see note 2, p. 32) changes its entire lineception, it also changes its entire perceived or relative universe, which is the only universe it knows or senses. And since causal feedbacks from this perceived universe[1] occur to the organism's perceptron (which itself is part of that perceived universe), then the effects of the organism's actions indeed return precisely to it. So, "as a man soweth, that shall he also reap" is quite true.[2]

It is now possible to formulate some conclusive statements:

1. One perceptron follows a lineception path, mechanically basing its output on its inputs, and this operation establishes causality to perception repetition, i.e., to lineception.

2. Ordinary physical phenomena are merely ensembles of related lineceptions, and are perceived to be absolutely deterministic (obey only cause-and-effect.

3. Living phenomena exhibit the effects (selective behavior) of inception (modulations impressed on lineception) which occasionally diverts one lineception from its deterministic lineception limb to another multiception limb, resulting in a changed (modulated) lineception.

4. Ultimate, or absolute, reality may be thought of as the sum-total of all multiceptions and multiflections.

5. "God" is all-being, which is all multiflection and multiception at once, which is both no time at all and forever and all forevers as well.

6. "God's will" is lineception and its accompanying lineflection.

7. "Man's will" or "free will" is inception, the modulation of lineception (and unknowingly its accompanying lineflection). Inception, however, imposes a feedback on the inceptor since a precise feedback, from the inception changes to perception, occurs to the perceptron of the perceiver. This feedback resulting from inception is karma.

8. The law of cause and effect (causality) exists in perceived nonmodulated phenomena (ensembles of lineceptions).

9. Causality is affected (modulated) by living systems, but applies to each choice once made. The causal feedback from inception is karma.

---

[1] "Perceived" must be accented: the nonperceived universe is not causal, since by definition causality invokes a perceiver and the perception process.

[2] Collective karma of any ordered group such as a species must exist in like manner. On the submicroscopic level, "karma of ensembles" means that the statistics of

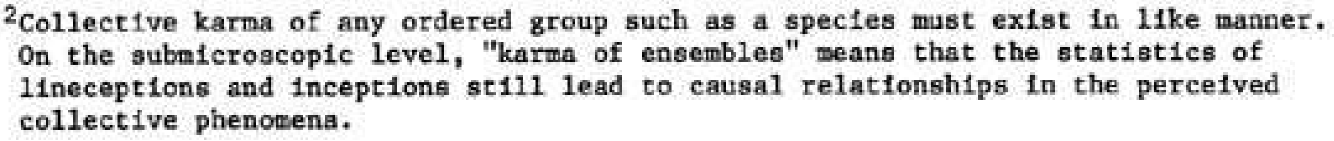

[2] Collective karma of any ordered group such as a species must exist in like manner. On the submicroscopic level, "karma of ensembles" means that the statistics of lineceptions and inceptions still lead to causal relationships in the perceived collective phenomena.

Next Page

10. Neither causality or karma applies in ultimate reality, although all causality lineceptions and their reflections, and all modulations (inceptions) and the karma reflections of those inceptions, are included in ultimate reality.

Next Page

VII. CLOSING REMARKS

The quiton/perceptron approach appears to offer a most extensive area for further investigation, and it is hoped that the physicists will interest themselves in the concept. It offers explanations for the problems long posed by ontology, and unifies the bases of science and metaphysics.

We state categorically that the theory explains what electric charge is, why like charges repel, and why unlike charges attract. In addition, since all phenomena are perceptron outputs, the theory indicates that multiple effects which are very close in time or distance will interfere with each other due to turbulent switching; i. e., conservation of mass/energy holds only as long as no switch interference exists. Turbulent switching can possibly disrupt all known physical laws. Unexpectedly strong forces, such as nuclear force, might in fact result from this effect. Other implications include possibility of travel faster than the speed of light (deBroglie waves do this now) and of travel back and forth in "time."

Since both laboratory instruments and human sensory apparatuses are perceptrons and differentiate reality, investigations of physical phenomena should include the laws of the perceptron as well as the laws of the perceived phenomena.

Next Page

BLANK

Next Page

DISTRIBUTION

	No. of Copies
US Army Missile Command Distribution List A	52
Defense Documentation Center Cameron Station Alexandria, Virginia 22314	12
Commander US Army Materiel Command	
ATTN: AMCCG	1
AMCRD	1
5001 Eisenhower Avenue Alexandria, Virginia 22304	
AMSMI-O	1
-FR, Mr. Rives	1
-R, Dr. McDaniel	1
-RB, Dr. Kobler	1
-RBL	5
-RP, Mr. Kopcha	1
-RPR	1
AMCPM-MD, LTC Bearden	100
-MD (Record Set)	1

Next Page

BLANK

Next Page

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) SAM-D PROJECT OFFICE US Army Materiel Command Redstone Arsenal, Alabama 35809	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE
QUITON/PERCEPTRON PHYSICS: A THEORY OF EXISTENCE, PERCEPTION, AND PHYSICAL PHENOMENA

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5. AUTHOR(S) (First name, middle initial, last name)
Thomas E. Bearden

6. REPORT DATE 5 March 1973	7a. TOTAL NO. OF PAGES 38	7b. NO. OF REFS ~~0~~ 3
8a. CONTRACT OR GRANT NO. b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AD 763210	

10. DISTRIBUTION STATEMENT
Approved for public release; distribution unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Same as No. 1

13. ABSTRACT

A fundamental answer is attempted to the question of being, time, space, existence, perception, and physical phenomena. Mass, empty space, reality, gravitational attraction, continuity, discontinuity, and relativity are among the "things" and "nothings" for which relationships and explanations are provided.

Perception is defined as detection by a mass of change to itself and as a differentiating process. Perceived physical phenomena are revealed to be first derivatives of ultimate reality, and outputs of the physical perceiving device called the "perceptron." Using the operation of the physical perception process on action, the units of which are energy times time, an entirely different view is generated of both reality and the Heisenberg uncertainty principle. When superposition is imposed as a condition, the Heisenberg uncertainty principle is revealed to be a quantized statement. A concept of the absolute value of a dimensional molecule is advanced.

A reality model is constructed of an action continuum and a nonaction continuum separated by an operational threshold whose magnitude is given by an uncertainty principle. Perception is modeled as a process which switches atomic quanta of action and nonaction back and forth across the threshold. Mass is explained as perceptron switch activity rate, and a mass definition equation is generated. One kilogram mass is shown to represent 17.053×10^{50} perceptron switches per second. From the model, Newton's laws of motion and gravity can be generated directly from the concept of the interaction of mass with space itself, space being taken as a massless fluid, or ether, composed of tiny nonaction particles called "quitons." Einstein's postulates of

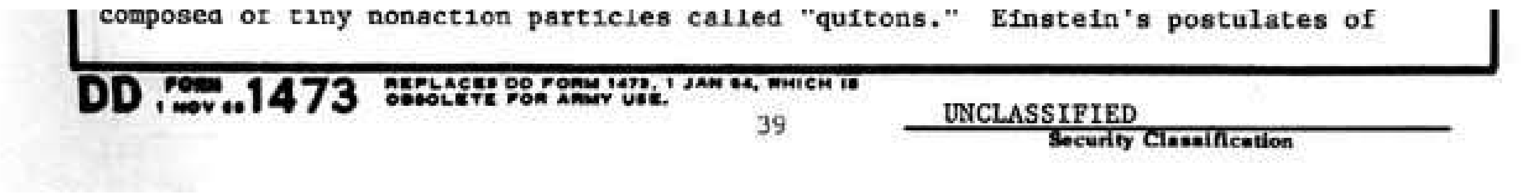

composed of tiny nonaction particles called "quitons." Einstein's postulates of

DD FORM 1 NOV 65 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

39

UNCLASSIFIED
Security Classification

Next Page

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
causality perception perceptron quiton quiton/perceptron theory gravitation mass reality metaphysics being existence space time photon action action quantum quantum observation relativity Karma complementarity physical phenomena Mach's principle Newton's laws inertia Debroglie wave Einstein's postulates Michelson-Morley experiment ontology Heisenberg Uncertainty Principle superposition equivalence principle simultaneity inertial frame						

ABSTRACT (Continued)

special relativity are also shown to follow from the model, and a statement of the equivalence principle is shown to be consistent with the model.

The concept of "lineception" as a causal chain of perceptron operations is advanced. "Inception" is developed as a noncausal modulation of lineception and as a distinguishing feature of a living system. The possibility is raised that inception could provide an explanation of the noncausal, statistical, ultramicroscopic universe and yet require large ensembles of ultramicroscopic events to conform to causality. Inception is shown to lead to karma. Perception is shown to be causal and to generate causality itself.

Some elementary philosophical implications of quiton/perceptron theory are pointed out and briefly discussed.

and briefly discussed.

Next Page

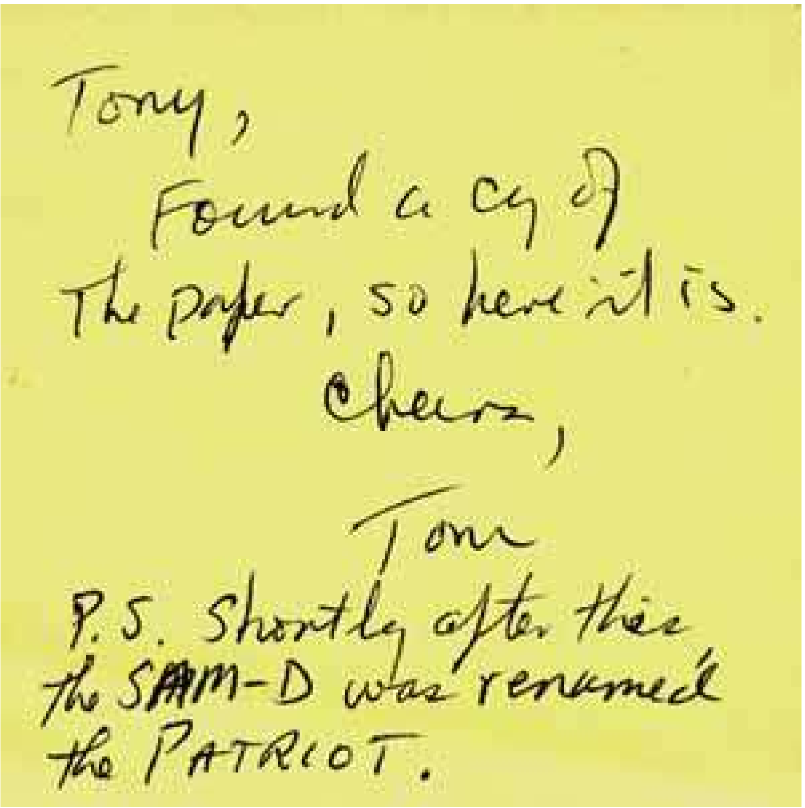

Tony,

Found a cy of
The paper, so here it is.

Cheers,

Tom

P.S. Shortly after this
the SAM-D was renamed
the PATRIOT.

And On Cue...

SCALAR WEAPONS: Read It and Weep

- Posted by Ms MoD on July 27, 2011 at 4:42 in HAARP
- View Discussions

This is a REALLY long paper but it is broken down into categories so you can easily skip to sections of interest to you. I found this after reading Twelve Things You Should Know About Scalar Weapons which was posted by FDF in April and is a MUCH SHORTER overview of scalar technology by the same author. I highly recommend it but wanted to post the longer version here as well which you can download in the PDF attached. Topics in the paper are listed below -MoD

SCALAR WEAPONS: Read It and Weep

By C Verismo

1. THE DISCOVERY OF SCALAR WAVES
2. BEAM WEAPONS OF A CENTURY AGO
3. THE CONTINUATION OF TESLA'S SCALAR WAVE TECHNOLOGY
4. TESLA'S CROSSED SCALAR BEAMS
5. RUSSIA'S 'WOODPECKER' SCALAR WAVE TRANSMITTERS
6. THE 1930s
7. THE 1940s
8. THE 1950s
9. THE 1960s
10. THE 1970s
11. THE 1980s
12. THE 1990s
13. THE PRESENT
WHAT CAN BE DONE WITH SCALAR TECHNOLOGY?
14. MIND READING, MASS HYPNOTISM AND SYNTHETIC TELEPATHY
15. DISEASE INDUCTION
16. DISABLING SATELLITES
17. IN THE AIR
18. WEATHER ENGINEERING
19. EARTHQUAKES
20. SEA ATTACKS
21. TESLA SHIELDS AND GLOBES
22. GROUND ATTACKS
23 OBLITERATING INDIVIDUALS
24. PORTABLE SCALAR BAZOOKAS
25. RADAR INVISIBILITY
26. SUMMARY OF A WAR THAT COULD OCCUR
27. RUSSIAN UFOs ATTACKED THE US SHUTTLE PROGRAM FROM THE BEGINNING
28. HOW DID RUSSIA GET UFOs?
29 FORMER BOLSHEVIKS RUNNING AMERICA
30 REPLICAS USED FOR THE LANDINGS, AFTER ORIGINALS SHOT DOWN
31. CLOAKED RUSSIAN UFOS ABOVE USA

32. ALL U.S. SATELLITES SHOT DOWN BY RUSSIA UNTIL 1981
33. LIES, LIES AND LIES
34. FRICTION WITH THE JAPANESE
35. ENTERPRISE SHUTTLE DETERMINED TO SPY ON RUSSIA
36. RUSSIA INTENSIFIES ATTACKS
37. TRYING AGAIN TO GET UP INTO SPACE. THIRD TIME LUCKY?
38. BOLD ATTACKS ON AMERICA
39. THE STRUGGLE AT THE TOP FOR POWER
40. THE STEALTH PLANE COMES TO THE RESCUE
41. PHANTOM AGAINST THE COSMOSPHERE
42. THE EXPECTED GENOCIDE
43. PLANS FOILED TO DESTROY RUSSIA'S MILITARY ARSENAL
44. BRAZIL
45. INDIA
46. JAPAN
47. KOBE 'EARTHQUAKE'
48. CHINA
49. OKLAHOMA FBI BUILDING 'BOMBING'
50. WHO ELSE HAS UFOs?

Full paper in PDF below…

Favorite
1 member favorited this

Share Tweet Facebook

Views: 459

Attachments:

- SCALAR WEAPONS - Read It and Weep.pdf, 494 KB

▶ Reply to This

Replies to This Discussion

Permalink Reply by Jenny on July 27, 2011 at 6:22

HAARP Patents (Assigned to APTI, Inc. Los Angeles, CA, Washington, DC)

U.S. Patent 4686605: Method And Apparatus For Altering A Region In The Earth’s Atmosphere, Ionosphere, And/Or Magnetosphere Issued: Aug. 11, 1987 Filed: Jan. 10, 1985
http://www.freepatentsonline.com/4686605.pdf

U.S. Patent 5038664: Method For Producing A Shell Of Relativistic Particles At An Altitude Above The Earth’s Surface ~ Issued: Aug. 13, 1991 Filed: Jan. 10, 1985
http://www.freepatentsonline.com/5038664.pdf

U.S. Patent 4712155: Method And Apparatus For Creating An Artificial Electron Cyclotron Heating

Region Of Plasma ~ Issued: Dec. 8, 1987 Filed: Jan. 28, 1985
http://www.freepatentsonline.com/4712155.pdf

U.S. Patent 5068669: Power Beaming System ~ Issued: Nov. 26, 1991 Filed: Sep. 1, 1988
http://www.freepatentsonline.com/5068669.pdf

U.S. Patent 5218374: Power Beaming System With Printer Circuit Radiating Elements Having Resonating Cavities ~ Issued: June 8, 1993 Filed: Oct. 10, 1989
http://www.freepatentsonline.com/5218374.pdf

U.S. Patent 5293176: Folded Cross Grid Dipole Antenna Element ~ Issued: Mar. 8, 1994 Filed: Nov. 18, 1991
http://www.freepatentsonline.com/5218374.pdf

U.S. Patent 5202689: Lightweight Focusing Reflector For Space ~ Issued: Apr. 13, 1993
http://www.freepatentsonline.com/5202689.pdf
Filed: Aug. 23, 1991

U.S. Patent 5041834: Artificial Ionospheric Mirror Composed Of A Plasma Layer Which Can Be Tilted ~ Issued: Aug. 20, 1991 Filed: May. 17, 1990
http://www.freepatentsonline.com/5041834.pdf

U.S. Patent 4999637: Creation Of Artificial Ionization Clouds Above The Earth ~ Issued: Mar. 12, 1991 Filed: May. 14, 1987
http://www.freepatentsonline.com/4999637.pdf

U.S. Patent 4954709: High Resolution Directional Gamma Ray Detector ~ Issued: Sep. 4, 1990 Filed: Aug. 16, 1989
http://www.freepatentsonline.com/4954709.pdf

U.S. Patent 4817495: Defense System For Discriminating Between Objects In Space ~ Issued: Apr. 4, 1989 Filed: Jul. 7, 1986
http://www.freepatentsonline.com/4817495.pdf

U.S. Patent 4873928: Nuclear-Sized Explosions Without Radiation ~ Issued: Oct. 17, 1989 Filed: June 15, 1987
http://www.freepatentsonline.com/4873928.pdf

Still think that HAARP is just another wacky fictional conspiracy? The above listed patents should be enough to convince any intelligent person that HAARP is real and is now actively modifying our weather and triggering earthquakes. For those who are still not convinced, look to the United States Congress for proof. In order for HAARP to get funding the people who first created HAARP needed funding. That funding was made possible by the US government. If you were to look at Congressional records you would find documents that states just one of the many official purposes of HAARP – for penetrating the earth with signals bounced off of the ionosphere. Congress was informed by people in the High Frequency Active Auroral Research Program that the bounced signals would be used to look inside the planet to a depth of many kilometers in order to locate underground munitions, minerals and tunnels. If you bounced the known earthquake signature signal (2.5 Hz) for a sustained period of time an earthquake could triggered many kilometers inside Earth's crust. The U.S. Senate actually set aside $15 million dollars in 1996 (Clinton and Al Gore Administration) to develop this ability alone — earth-penetrating-tomography.

Update April 21, 2011

Under pressure from the International community the HAARP website is up again. No data is available from April 7 to April 12, 2011. Data before and after shows the signature frequency of an earthquake (2.5Hz) being transmitted. Also shows energy spikes which indicates active weather modification. Now that the website is up again you can view the data for the Japan earthquake at http://137.229.36.30/cgi-bin/scmag/disp ... i?20110311. From the magnetometer induction data you can see 2.5Hz (earthquake inducing frequency) being broadcasted and recorded for several days before the 9.0 magnitude earthquake. To see what a normal (no HAARP broadcasting) day looks like view the magnetometer charts for the days of March 15, 16 and 17 – March 16, 2011 http://137.229.36.30/cgi-bin/scmag/disp ... 0316&Bx=on. No earthquake inducing broadcast was detected from March 15 to March 27, 2011. On March 28, 2011 a low powered broadcast was detected from approximately 12:00 am to 3:00 am. Broadcasting resumed on April 2, 2011. April 6, 2011 shows you what the magnetometer charts looks like when a solar storm blasts Earth. No data is available from April 7, 2011 to April 13, 2011 – a shutdown. Starting on April 15, 2011 and ending approximately 10:00 am April 16, 2011 the earthquake inducing frequency of 2.5Hz was detected being broadcasted.
Short URL: http://presscore.ca/2011/?p=2261

- ▶ Reply
-

Permalink Reply by Jenny on July 27, 2011 at 6:23
WOW MoD! What a paper! I need like a year to get through it all but it sure is awesome to have it all in one place! Excellent!

- ▶ Reply
-

Permalink Reply by Ms MoD on October 26, 2012 at 6:38

Boeing Tests Microwave Missile That Can Knock Out All Electronics (CHAMP)
Huffington Post UK | By Michael Rundle 24/10/2012

Boeing is testing a missile which can destroy electronics in a wide area with little or no physical damage.

The 'Champ' (Counter-electronics High-powered Advanced Missile Project) may "one day change warfare", Boeing said.

"This technology marks a new era in modern-day warfare," said Keith Coleman, CHAMP program manager for Boeing Phantom Works.

"In the near future, this technology may be used to render an enemy’s electronic and data systems useless even before the first troops or aircraft arrive".

The missile was tested on 16 October, when the high-powered microwaves were fired at a two storey building built especially for the job.

The missile would, if successful, take out rows of personal computers which were turned on to gauge the effect of the blast.

The result?

"Seconds later the PC monitors went dark and cheers erupted in the conference room. CHAMP had successfully knocked out the computer and electrical systems in the target building. Even the television cameras set up to record the test were knocked off line without collateral damage."

Seven targets were hit in the one-hour test, and all were successfully disabled.

There is a long way to go before the microwave can be placed into a missile and used to disable targets without loss of life, but the team is hopeful it could prove to be a milestone in non-lethal warfare.

"Today we turned science fiction into science fact," Coleman said.

http://www.huffingtonpost.co.uk/2012/10/24/boeing-tests-microwave-m...

- ▶ Reply
-

Permalink Reply by Ms MoD on August 3, 2013 at 11:01

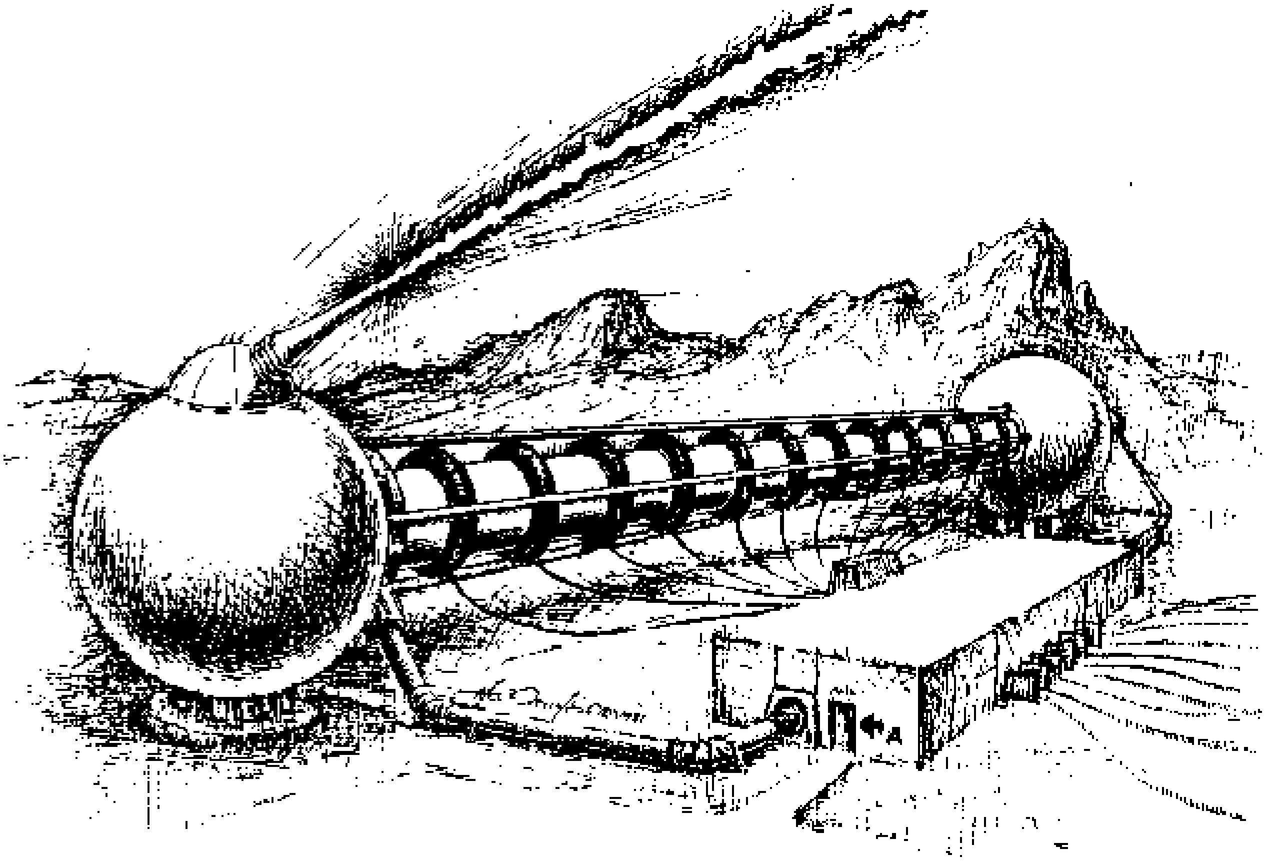

Scalar Potential Interferometer

Figure 6. Multimode Tesla Weapon

In the 1930's Tesla announced other bizarre and terrible weapons: a death ray, a weapon to destroy hundreds or even thousands of aircraft at hundreds of miles range, and his ultimate weapon to end all war -- the Tesla shield, which nothing could penetrate. However, by this time no one any longer paid any real attention to the forgotten great genius. Tesla died in 1943 without ever revealing the secret of these great weapons and inventions.

Unfortunately, today in 1981 the Soviet Union has long since discovered and weaponized the Tesla scalar wave effects. Here we only have time to detail the most powerful of these frightening Tesla weapons -- which Brezhnev undoubtedly was referring to in 1975 when the Soviet side at the SALT talks suddenly suggested limiting the development of new weapons "more frightening than the mind of man had imagined." One of these weapons is the Tesla howitzer recently completed at the Saryshagan missile range and presently considered to be either a high-energy laser or a particle beam weapon, (See Aviation Week & Space Technology, July 28, 1980, p. 48 for an artist's conception.)

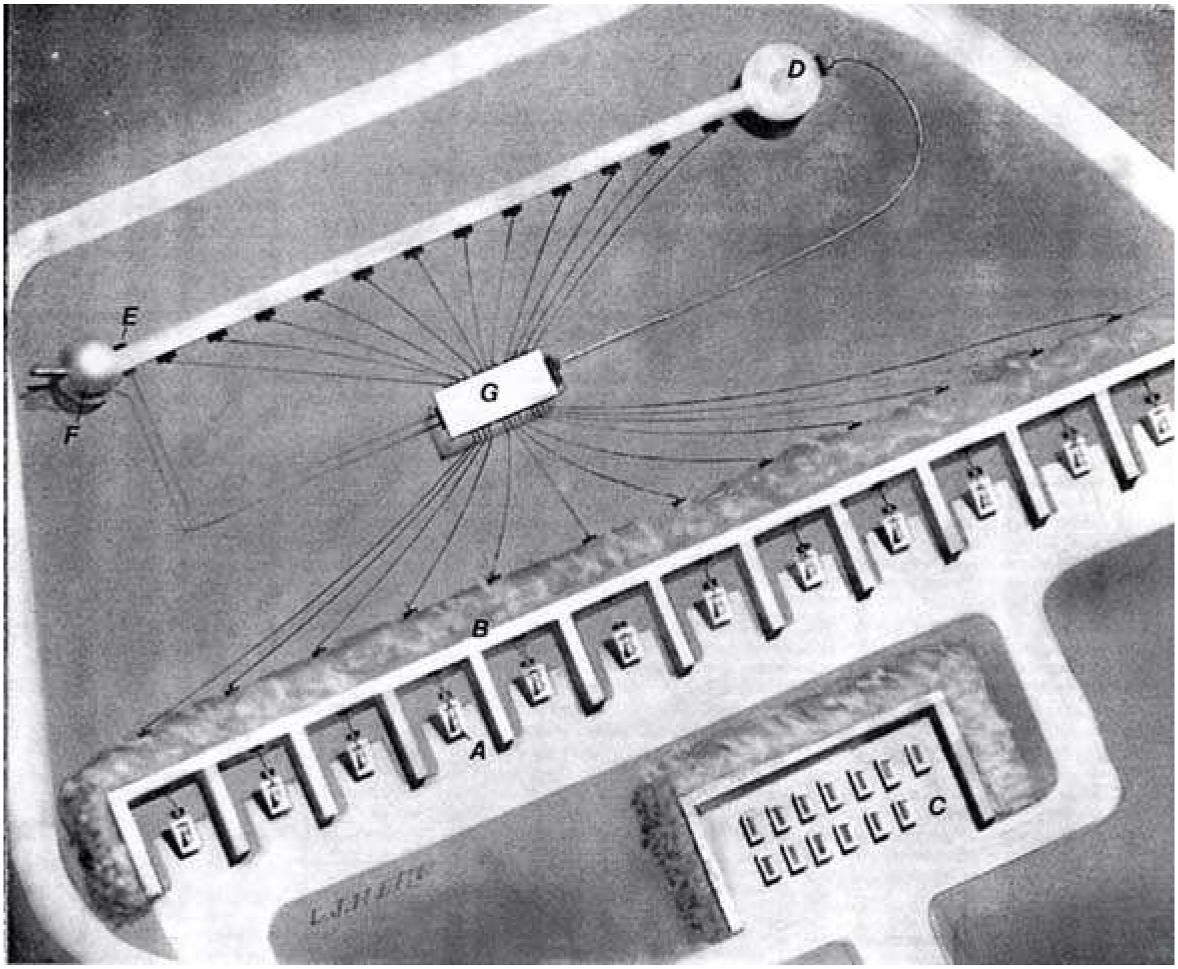

Aviation Week & Space Technology July 28, 1980

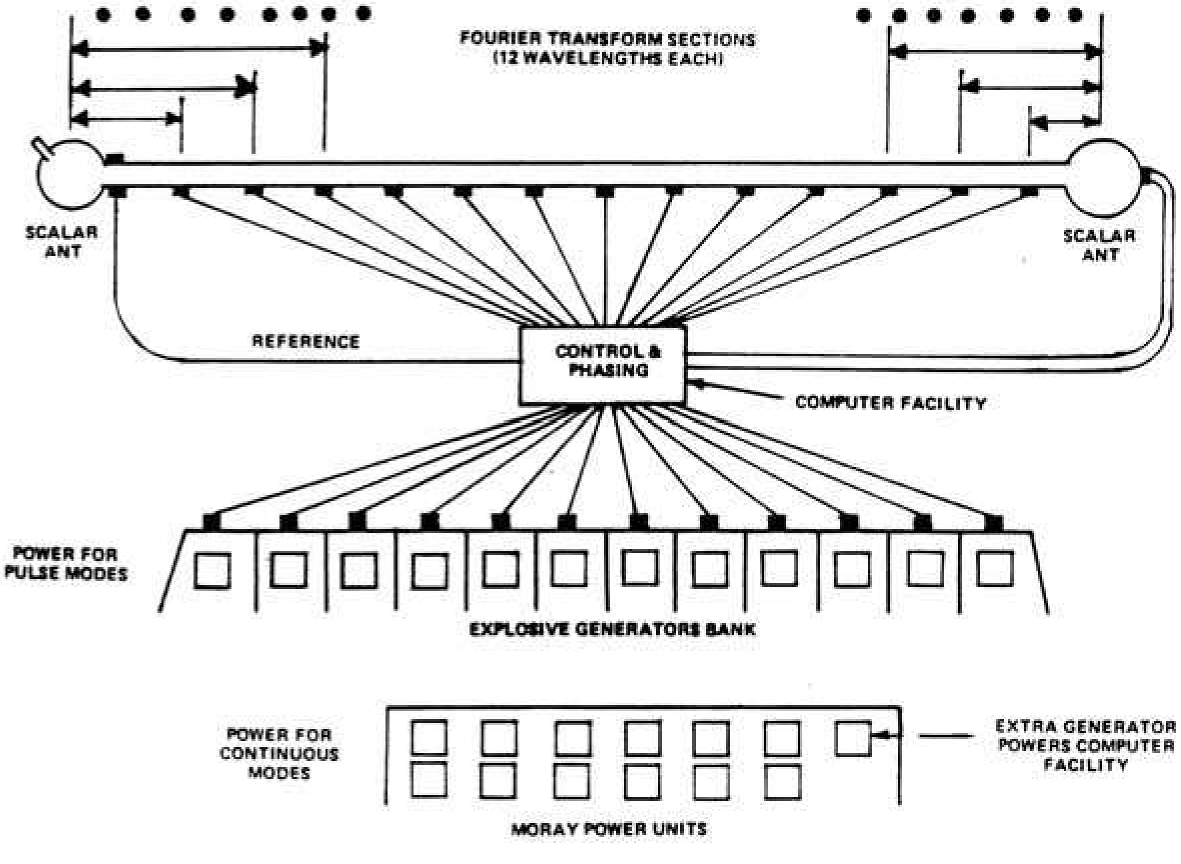
FOURIER TRANSFORM SECTIONS
(12 WAVELENGTHS EACH)
SCALAR ANT
SCALAR ANT
REFERENCE
CONTROL & PHASING
COMPUTER FACILITY
POWER FOR PULSE MODES
EXPLOSIVE GENERATORS BANK
POWER FOR CONTINUOUS MODES
EXTRA GENERATOR POWERS COMPUTER FACILITY
MORAY POWER UNITS

Figure 7. Tesla Weapons at Saryshagan

The Saryshagan howitzer actually is a huge Tesla scalar interferometer with four modes of operation. One continuous mode is the Tesla shield, which places a thin, impenetrable hemispherical shell of energy over a large defended area. The 3-dimensional shell is created by interfering two Fourier-expansion, 3-dimensional scalar hemispherical patterns in space so they pair-couple into a dome-like shell of intense, ordinary electromagnetic energy. The air molecules and atoms in the shell are totally ionized and thus highly excited, giving off intense, glowing light. Anything physical which hits the shell receives an enormous discharge of electrical energy and is instantly vaporized -- it goes pfft! like a bug hitting one of the electrical bug killers now so much in vogue.

If several of these hemispherical shells are concentrically stacked, even the gamma radiation and EMP from a high altitude nuclear explosion above the stack cannot penetrate all the shells due to repetitive absorption and reradiation, and scattering in the layered plasmas.

In the continuous shield mode, the Tesla interferometer is fed by a bank of Moray free energy generators, so that enormous energy is available in the shield. A diagram of the Saryshagan-type Tesla howitzer is shown in figure 7. Hal Crawford's fine drawing of the interferometer end of the Tesla howitzer is shown in figure 6. Hal's exceptional rendition of the Tesla shield produced by the howitzer is shown in figure 8.

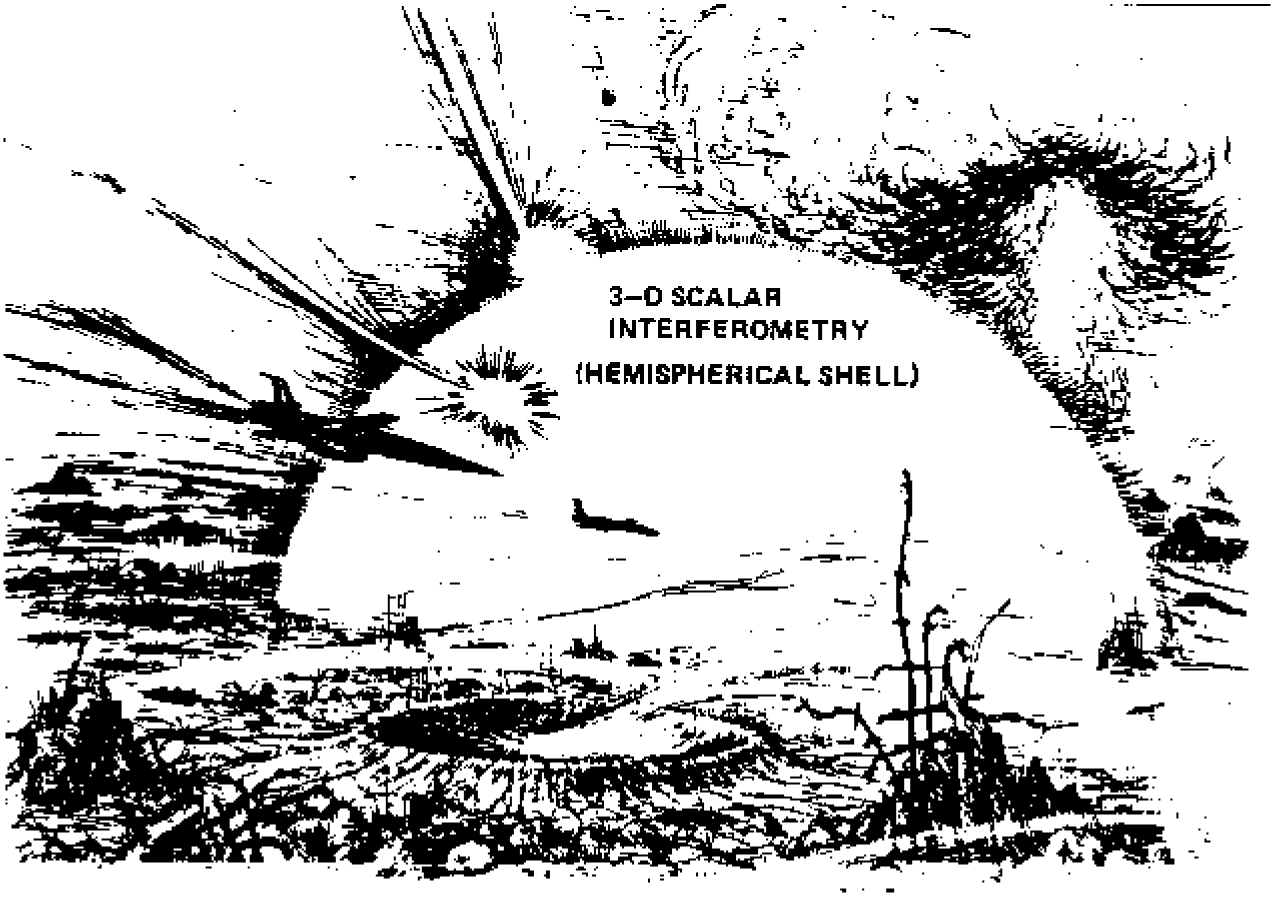

Figure 8. The Tesla Shield

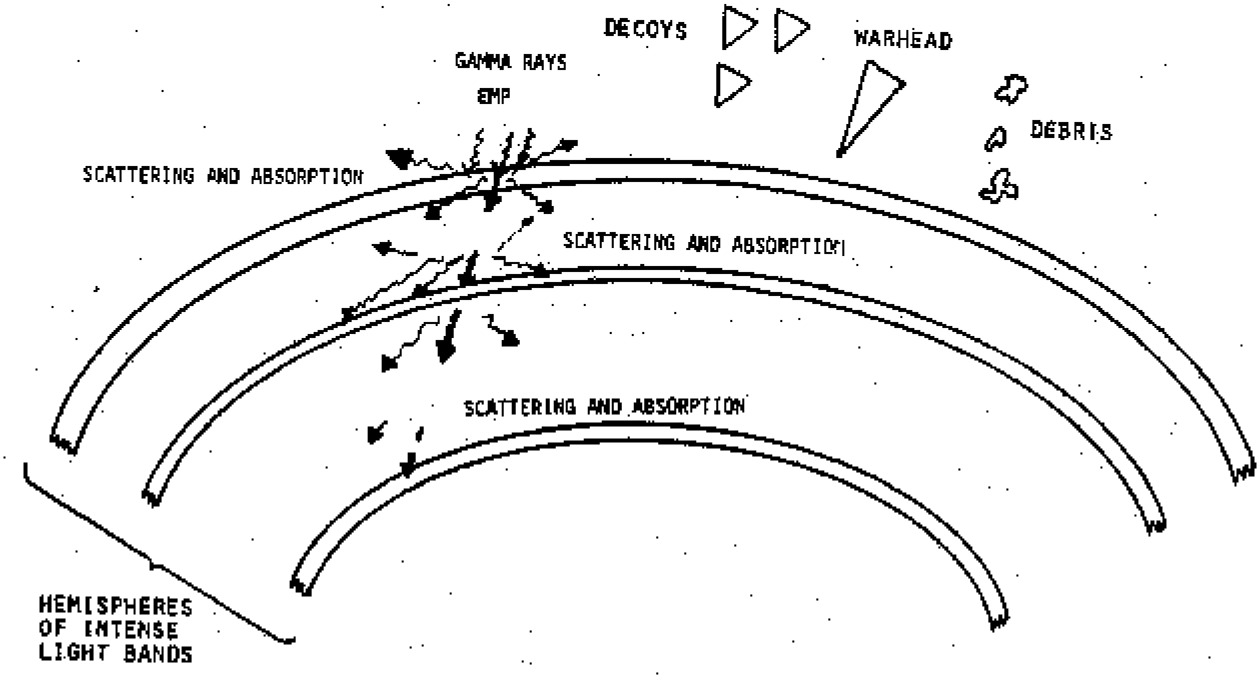
DECOYS
WARHEAD
GAMMA RAYS
EMP
DEBRIS
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
HEMISPHERES
OF INTENSE
LIGHT BANDS

Figure 9. Tesla Terminal Area Defense System

In the pulse mode, a single intense 3-dimensional scalar phi-field pulse form is fired, using two truncated Fourier transforms, each involving several frequencies, to provide the proper 3-dimensional shape (Figure 10). This is why two scalar antennas separated by a baseline are required. After a time delay calculated for the particular target, a second and faster pulse form of the same shape is fired from the interferometer antennas. The second pulse overtakes the first, catching it over the target zone and pair-coupling with it to instantly form a violent EMP of ordinary vector (Hertzian) electromagnetic energy. There is thus no vector transmission loss between the howitzer and the burst. Further, the coupling time is extremely short, and the energy will appear sharply in an "electromagnetic pulse (EMP)" strikingly similar to the 2-pulsed EMP of a nuclear weapon.

This type weapon is what actually caused the mysterious flashes off the southwest coast of Africa, picked up in 1979 and 1980 by Vela satellites. The second flash, e.g., was in the infrared only, with no visible spectrum. Nuclear flashes do not do that, and neither does superlightning, meteorite strikes, meteors, etc. In addition, one of the scientists at the Arecibo Ionospheric Observatory observed a gravitational wave disturbance -- signature of the truncated Fourier pattern and the time-squeezing effect of the Tesla potential wave -- traveling toward the vicinity of the explosion.

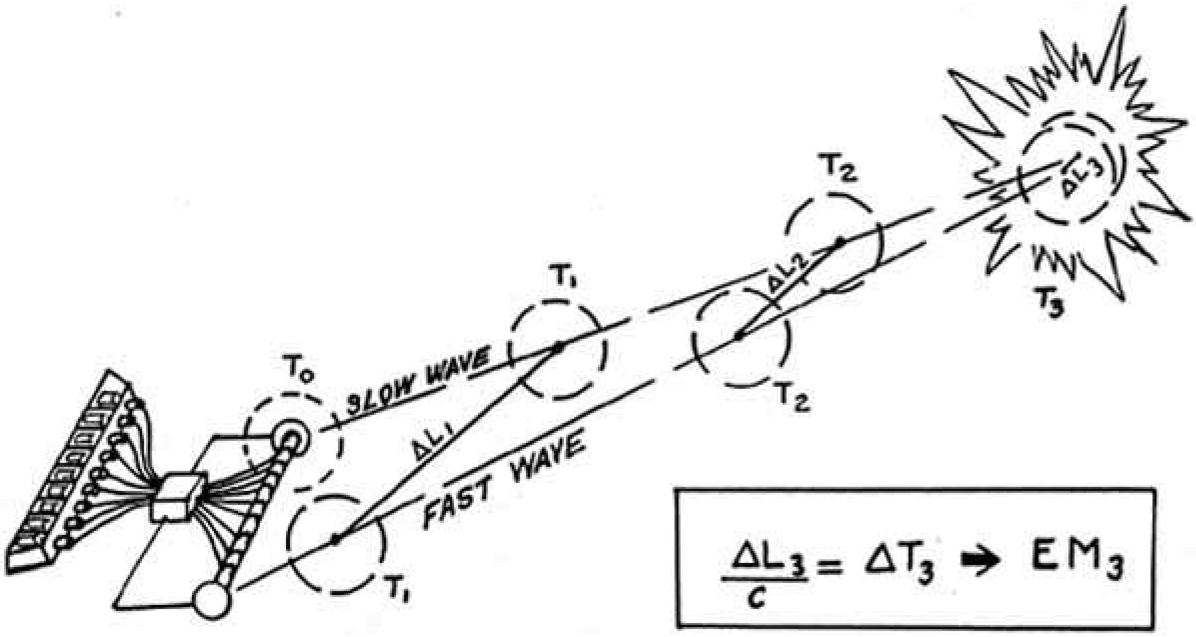
T₀
SLOW WAVE
T₁
T₂
T₂
ΔL₁
ΔL₂
ΔL₃
T₃
FAST WAVE
T₁
$\frac{\Delta L_3}{c} = \Delta T_3 \rightarrow EM_3$

TESLA HOWITZER
(SCALAR INTERFEROMETER)

Figure 10. "Nuclear" Flashes off the Coast of Africa

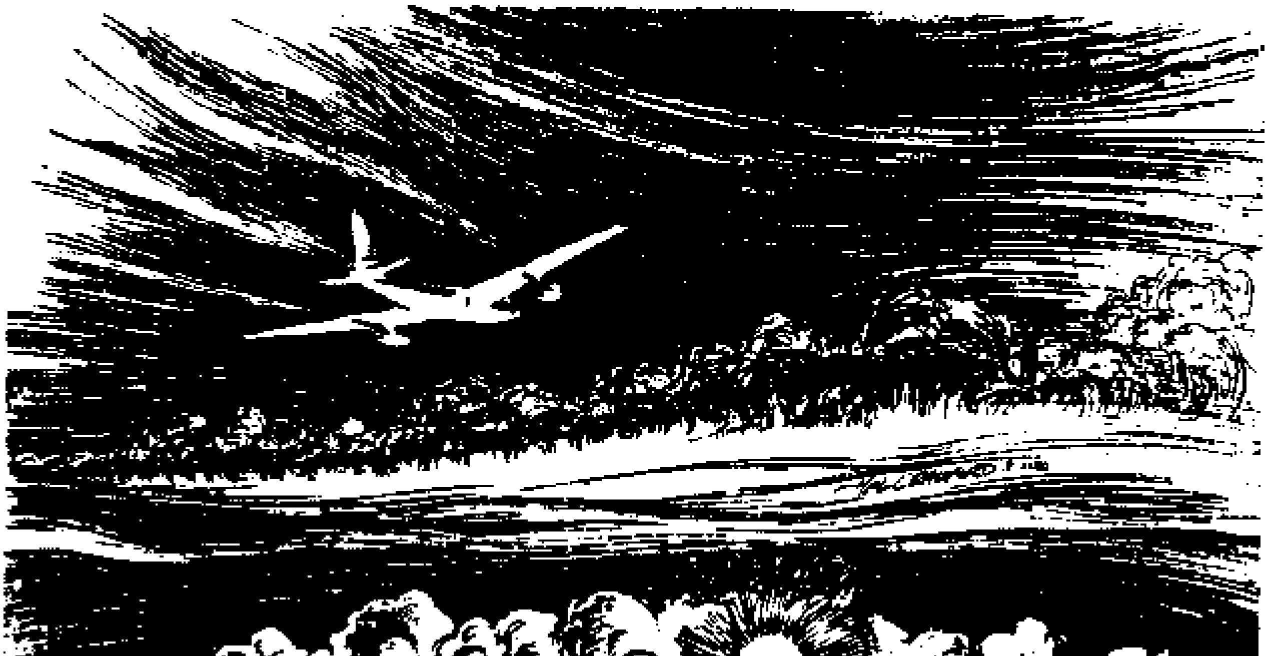

Lithuania - 10 Sep 1976 - British European Airways Flight #831 between Moscow and London

CIA Report Released under FOIA

Figure 11. Continuous Tesla Fireball

The pulse mode may be fed from either Moray generators or -- if the Moray generators have suffered their anomalous "all fail" malfunction -- ordinary explosive generators. Thus the Tesla howitzer can always function in the pulse mode, but it will be limited in power if the Moray generators fail.

In the continuous mode, two continuous scalar waves are emitted -- one faster than the other -- and they pair-couple into vector energy at the region where they approach an in-phase condition. In this mode, the energy in the distant "ball" or geometric region would appear continuously and be sustained -- and this is Tesla's secret of wireless transmission of energy at a distance without any losses. It is also the secret of a "continuous fireball" weapon capable of destroying hundreds of aircraft or missiles at a distance. An example of a Soviet test of this mode of operation is shown in figure 11.

Witness to a super weapon?

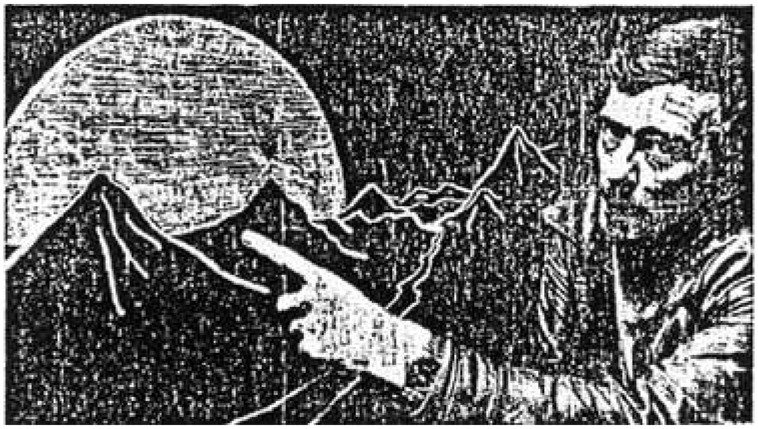

Nick Downie describes the strange lurid glow that flared silently over the Hindu Kush

THE SUNDAY TIMES, 17 AUGUST 1980

(Multiple incidents in Sept., 1979)

TEHERAN, IRAN 17 JUNE 1966

SEEN FROM NEAR MEHRABAD AIRPORT

OBSERVED 4-5 MINUTES

SEEN BY 2 AIRCRAFT

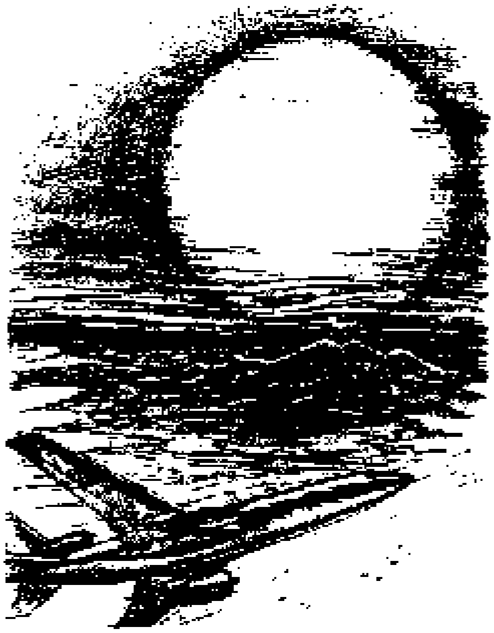

CIA REPORT RELEASED UNDER FOIA

Figure 12. Tesla EMP Globe

The volume of the Tesla fireball can be vastly expanded to yield a globe which will not vaporize physical vehicles but will deliver an EMP to them to dud their electronics. A test of this mode is shown in figure 12. (See also Gwynne Roberts, "Witness to a Super Weapon?", the London Sunday Times, 17 August 1980 for several other tests of this mode at Saryshagan, seen from Afghanistan by British TV cameraman and former War Correspondent Nick Downie.)

If the Moray generators fail anomalously, then a continuous mode limited in power and range could conceivably be sustained by powering the interferometer from more conventional power-sources such as advanced magnetohydrodynamic generators.

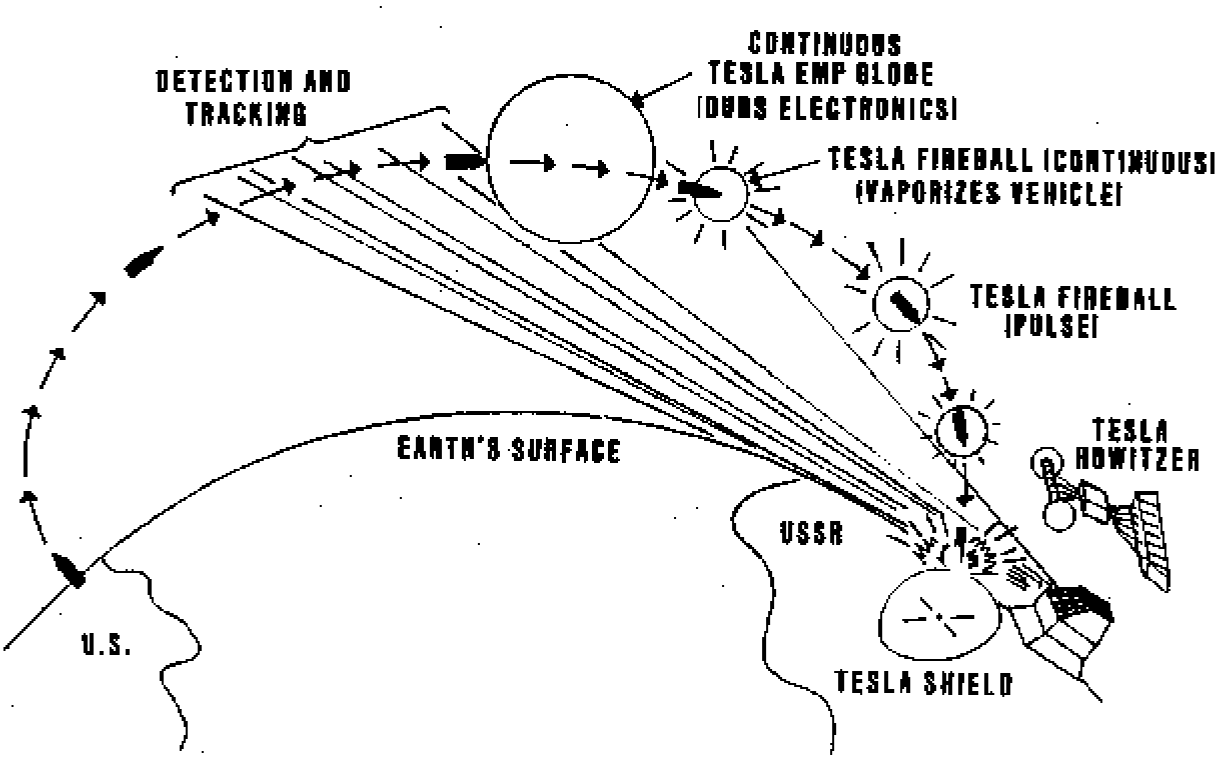
DETECTION AND TRACKING
CONTINUOUS TESLA EMP GLOBE (DUDS ELECTRONICS)
TESLA FIREBALL (CONTINUOUS) (VAPORIZES VEHICLE)
TESLA FIREBALL (PULSE)
TESLA HOWITZER
EARTH'S SURFACE
USSR
U.S.
TESLA SHIELD

Figure 13. Tesla ABM Defenses

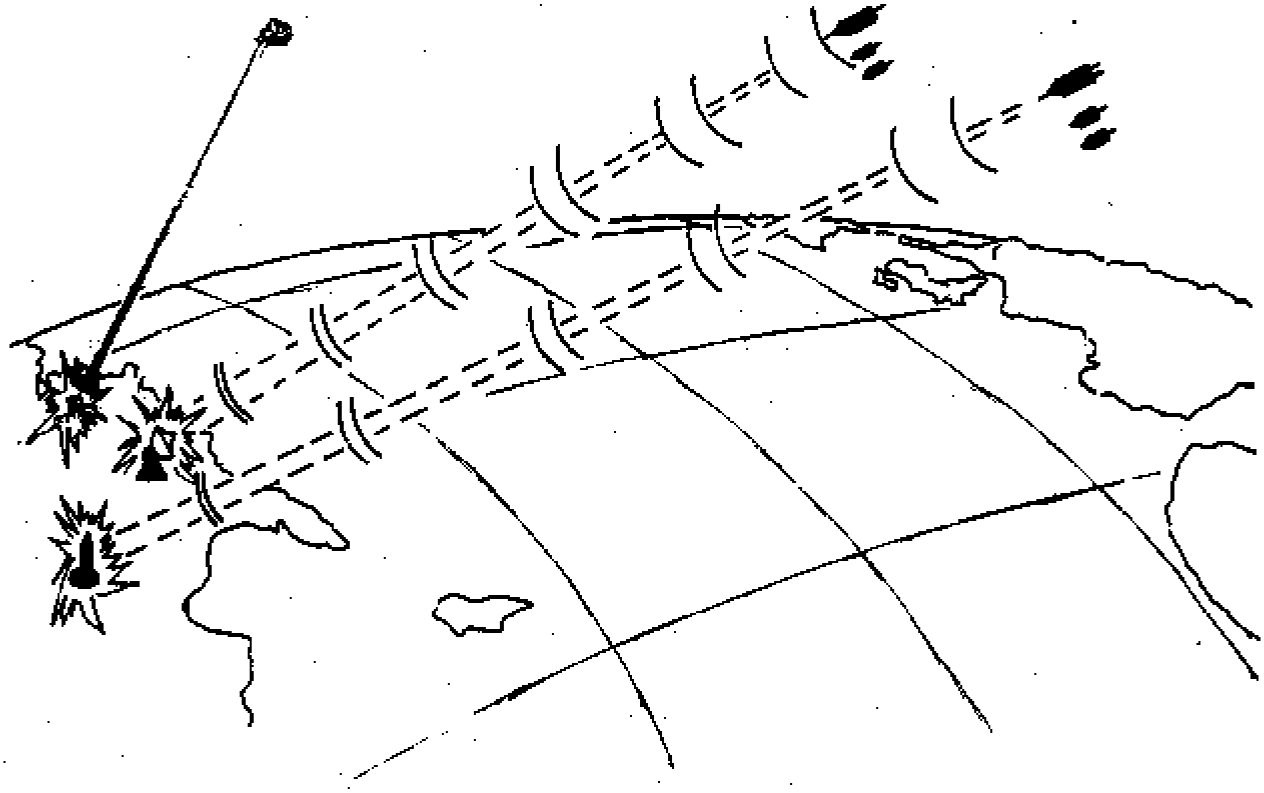

Figure 14. Moray/Tesla Technology: Star Wars Now

Typical strategic ABM uses of Tesla weapons are shown in figure 13. In addition, of course, smaller Tesla howitzer systems for anti-tactical ballistic missile defense of tactical troops and installations could be constituted of more conventional field missile systems using paired or triplet radars, of conventional external appearance, in a scalar interferometer mode.

With Moray generators as power sources and multiply deployed reentry vehicles with scalar antennas and transmitters, ICBM reentry systems now can become long range "blasters" of the target areas, from thousands of kilometers distance (figure 14). Literally, "Star Wars" is liberated by the Tesla technology. And in air attack, jammers and ECM aircraft now become "Tesla blasters." With the Tesla technology, emitters become primary fighting components of stunning power.

The potential peaceful implications of Tesla waves are also enormous. By utilizing the "time squeeze" effect, one can get antigravity, materialization and dematerialization, transmutation, and mindboggling medical benefits. One can also get subluminal and superluminal communication, see through the earth and through the ocean, etc. The new view of phi-field also provides a unified field theory, higher orders of reality, and a new super-relativity, but detailing these possibilities must wait for another book.

With two cerebral brain halves, the human being also has a Tesla scalar interferometer between his ears. And since the brain and nervous system processes avalanche discharges, it can produce (and detect) scalar Tesla waves to at least a limited degree. Thus a human can sometimes produce anomalous spatiotemporal effects at a distance and through time. This provides an exact mechanism for psychokinesis, levitation, psychic healing, telepathy, precognition, postcognition, remote viewing, etc. It also provides a reason why an individual can detect a "stick" on a radionics or Hieronymus machine (which processes scalar waves), when ordinary detectors detect nothing. Unfortunately there is not room to develop the implications of this human Tesla interferometry in detail, for that must wait for yet another book, presently in its initial stages, that Hal Crawford and I are writing.

Table 5. Orders of Reality

- **PHOTONS ARE**
  - **> PAIR-COUPLED SCALARS**
  - **> VELOCITY-LIMITED TO C**
  - **> CARRIERS OF T**

- **PHOTON INTERACTION**
  - **> IS UBIQUITOUS**
  - **> PRODUCES $\dot{T} = C$**
  - **> YIELDS 1st ORDER REALITY**

- **SCALAR O WAVES**
  - **> NOT VELOCITY-LIMITED**
  - **> YIELD HIGHER ORDER REALITIES**

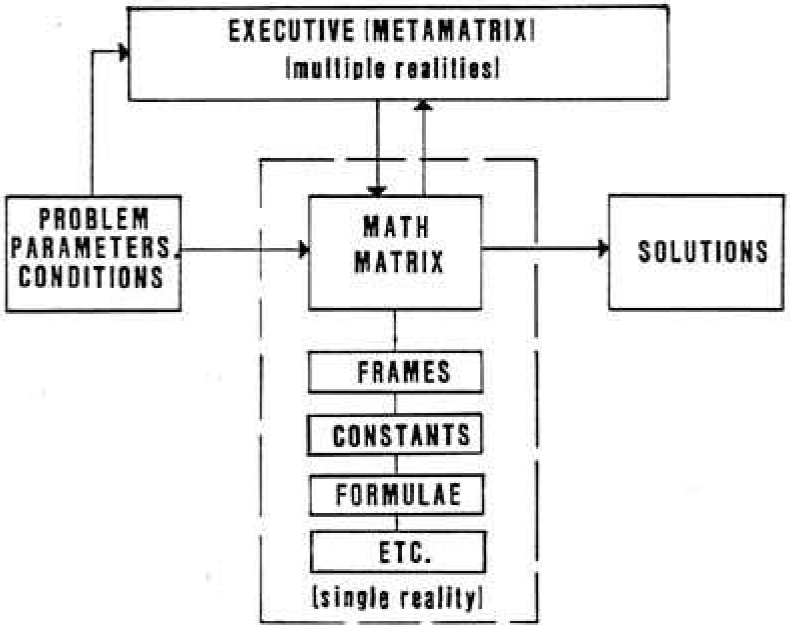

Figure 15. Implications of Tesla Potential

At the July 1981 U.S. Psychotronics Association's Annual Conference in Dayton, Ohio, I presented the first rough paper on the Tesla secret and scalar interferometry. A videotape of the presentation was made and will shortly be available. I am also scheduled to make a special presentation at the Alternate Energy Conference in Toronto, Canada in latter October, 1981. A professional, videotaped two-hour presentation on this subject is also being prepared. Wide distribution of the material through the international underground physics and technology network has already been made. This time, God willing, Tesla's secret will not be suppressed for another 80 years!

And perhaps it is not yet too late. The material has cost me (now) some 16 years of agonizing labor and nearly $100,000 of my own personal funds. No orthodox university, scientific group, foundation, or governmental agency would support such an effort, either financially or otherwise. Indeed, most ordinary journals will not even accept material on such matters. Nonetheless, the area is of overwhelming importance and I truly believe Tesla's lost secret will shortly affect the lives of every human being on earth.

Perhaps with the free and open release of Tesla's secret, the scientific and governmental bureaucracies will be. shocked awake from their slumber, and we can develop defenses before Armageddon occurs. Perhaps there is hope after all -- for even Brezhnev, in his strange July, 1975 proposal to the SALT talks, seemed to reveal a perception that a turning point in war and weaponry may have been reached, and that human imagination is incapable of dealing with the ability to totally engineer reality itself. Having tested the weapons, the Soviets must be aware that the ill-provoked oscillation of timeflow affects the minds and thoughts -- and the very lifestreams and even the collective species unconsciousnesses -- of all lifeforms on earth. They must know that these weapons are two-edged swords, and that the backlash from their use can be far more terrible to the user than was the original effect to his victim.

If we can avoid the Apocalypse, the fantastic secret of Nikola Tesla can be employed to cure and elevate man, not kill him. Tesla's discovery can eventually remove every conceivable external human limitation. If we humans ourselves can elevate our consciousness to properly utilize the Tesla electromagnetics, then Nikola Tesla -- who gave us the electrical twentieth century in the first place -- may yet give us a fantastic new future more shining and glorious than all the great scientists and sages have imagined.

References

http://www.cheniere.org/books/part1/teslaweapons.htm

- ▶ Reply
-

Permalink Reply by rose on August 4, 2013 at 10:54

yeah this ones from deep in the archives Greg - I would love to see that article!!!

- ▶ Reply
-

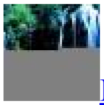
Permalink Reply by deanna munson on August 4, 2013 at 18:16

what if these weapons work in conjunction with cell towers and smartgrid/wifi relays? i just read this and it blew me away how the dangers this is being purposely ignored http://www.stayonthetruth.com/electromagnetic-fields---killing-fiel... "

November 14 1996, Omnipoint, New York City's first digital cellular provider, did open for business, broadcasting from thousands of antennae newly erected on the rooftops of apartment buildings. According to the health authorities, an early flu hit New York City - but not Boston, and not Philadelphia - on about 15 November. The flu was severe and ran a prolonged course, often dragging on for months instead of the usual two weeks.

At Christmas time, the Cellular Phone Task Force placed a small classified ad in a free weekly newspaper. It read: 'If you have been ill since 11/15/96 with any of the following: eye pain, insomnia, dry lips, swollen throat, pressure or pain in the chest, headaches, dizziness, nausea, shakiness, other aches and pains, or flu that won't go away, you may be a victim of a new microwave system blanketing the city. We need to hear from you.' And we did hear from them. Hundreds called, men, women, whites, blacks, Asians, Latinos, doctors, lawyers, teachers, stockbrokers, airline stewards, computer operators. Most had woken up suddenly in mid-November, thinking they were having a stroke or a heart attack or a nervous breakdown, and were relieved to know they were not alone and not crazy.

Later, I analysed weekly mortality statistics, which the Centres for Disease Control publish for122 US cities. Each of dozens of cities recorded a 10-25 per cent increase in mortality, lasting two to three months, beginning on the day in 1996 or 1997 on which that city's first digital cell phone network began commercial service."   we already know these agendas are being forced everywhere.no one is even noticing how even though everyone is already "connected" they are still erecting more and more towers and infastructure,by force in the case of smartmeters.wired in connections now cost over double what wifi does.they just put up another tower here on my block,disguised as a flipping pine tree.cant block the effects when it comes from literally all sides

- ▶ Reply
-

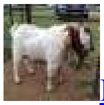
Permalink Reply by Rhonda Reichel on August 6, 2013 at 6:15

Scalar weapons makes all the billions we spend on war obsolete.....what a racket

- ▶ Reply
-

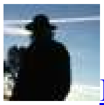
Permalink Reply by steve on August 9, 2013 at 6:43
having read a few pages, it looks like a weapon of distraction to me. The physics is in the realm of the tooth fairy, and the history has the ring of recent manufacture

- ▶ Reply
-

Permalink Reply by rose on August 9, 2013 at 12:24

PLEASE UPDATE YOUR ACCOUNT TO ENABLE 3RD PARTY HOSTING

100%

3rd Party Hosting Usage

FOR IMPORTANT INFO, PLEASE GO TO:
www.photobucket.com/P500

- ▶ Reply
-

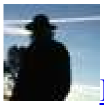
Permalink Reply by steve on August 9, 2013 at 16:03
plus, Tom Beardon is nothing but a fear merchant.

- ▶ Reply
-

RSS

Welcome to
The ConTrail

Sign Up
or Sign In

Or sign in with:

-

Visitors

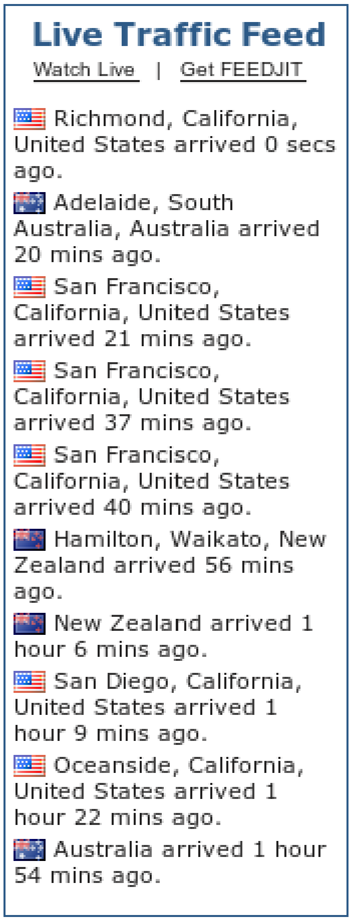

Live Traffic Feed

Watch Live | Get FEEDJIT

Richmond, California, United States arrived 0 secs ago.

Adelaide, South Australia, Australia arrived 20 mins ago.

San Francisco, California, United States arrived 21 mins ago.

San Francisco, California, United States arrived 37 mins ago.

San Francisco, California, United States arrived 40 mins ago.

Hamilton, Waikato, New Zealand arrived 56 mins ago.

New Zealand arrived 1 hour 6 mins ago.

San Diego, California, United States arrived 1 hour 9 mins ago.

Oceanside, California, United States arrived 1 hour 22 mins ago.

Australia arrived 1 hour 54 mins ago.

Groups

-

  ### Health and Wellbeing

  123 members

  99 Comments 10 Favorites

-

  ### Vaccination Information

  87 members

  163 Comments 6 Favorites

-

  Simply Good Stuff

  40 members

  112 Comments 3 Favorites
-

  EMF and EMR Information …

  101 members

  74 Comments 13 Favorites
-

  Cancer Treatment Options

  74 members

  77 Comments 9 Favorites
-

  Grow your own Food and M…

  96 members

  52 Comments 12 Favorites
-

  Childbirth

  3 members

  1 Comment 0 Favorites
- View All

© 2018 Created by rose. Powered by

Badges | Report an Issue | Terms of Service

Sign in to chat!

The Tom Bearden Website

Help support the research

Foundations of Electro-Magnetism
Slide Series

- Major Points of Presentation
- Chung's Carbon Filament Negative Resistor
- Aharanov-Bohm Effect
- Absorption and emission reactions (1)
- Absorption and emission reactions (2)
- Anti-engine for cell's deviation reverses cell back to normal
- Asymmetric regauging produces excess force, which can be used to do work on the system
- Asymmetric regauging produces excess force, which can be used to do work on the regauging system
- Becker's bone fracture healing experiments
- Becker's theoretical control system governing regeneration
- Becker's theoretical DC control system involved with response to injury
- Geometrical Distortion of Poynting Energy Flow
- Cosmic Bootstrap: Before the Big Bang
- Bridging Enables COP > 1

- Aspects of Strong Local Symmetry
- Other Aspects of Strong Local Symmetry
- Bunge on the Status of Electrodynamics
- Bunge on the Status of Electrodynamics and Physics
- Bunge on the Status of Classical and Quantum Physics
- Fractions of BW Casualties Surviving
- First Step in Self-Promotion of Cancer Due to Hypoxia
- Differentiation and Dedifferentiation Timelines
- Evolution of Cells on Earth
- A Charged Particle q is a Coupled System of m and Ø
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (1)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (2)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (3)
- A Charged Particle is a Coupled System
- The Anomalies in Navy Electrolyte Experiments at China Lake
- Chung's Carbon Filament Negative Resistor

- Typical circuit has about 10^{-13} energy collection efficiency
- Classical View of EM
- Curved Spacetime acts as source or sink
- Chronic Disease Puzzling
- Cancer Characteristics
- Solving a Cancer Enigma
- Cancer: Characteristics
- Nonlinear Optics Distortion Correction Theorem
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)
- In Time-Reversed Zone: Nuclear reactions are biased
- In Time-Reversed Zone: Dynamics may be reversed
- Time-Reversed Zone: Significant only after time-density charging
- A New Conservation of Energy Law
- Questions on Carcinogens and EM

Radiation

- AS DEFINED, fields and potentials only exist in and on charged matter
- Serious flaws and errors in classical EM theory
- Electrodynamicists' Reaction to Removal of the Material Ether
- Reaction to Loss of Material Ether (1)
- Reaction to Loss of Material Ether (2)
- Some EM waves in Spacetime
- Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String
- Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna
- Phase Conjugate Wavepairs Produce New Waves
- Scalar Electromagnetics (Energetics) View of EM (1)
- Scalar Electromagnetics (Energetics) View of EM (2)

- Scalar Electromagnetics (Energetics) View of EM (3)
- Extended General Relativity Principle
- Extension of Work-Energy Theorem
- 34 Flaws in Classical EM Theory
- MTW's General Relativity Principle
- Use of General Relativity (GR) in Particle Physics
- The Missing Infolded Electrodynamics
- Infolded Engines: Excluded by Present Physics
- Mass curves and structures spacetime; ST curvatures change and structure mass (1)
- Mass curves and structures spacetime; ST curvatures change and structure mass (2)
- At Infinite Velocity: Each Is Every Other
- Lisitsyn's Report: Brain Code Broken
- Lorentz Closed Surface Integration
- Lorentz's physically insignificant energy flow can be collected and utilized
- Lorentz surface integration of the Poynting vector around a closed surface
- Longitudinal EM Photon Interaction with Charge
- Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass.

- Pumping with Longitudinal EM waves Time-reverses the Mass
- Maxwell's Quaternion Theory
- Maxwellian Systems before and after Regauging
- Maxwell's four equations reduce to: Page 1
- Maxwell's four equations reduce to: Page 2
- More to come

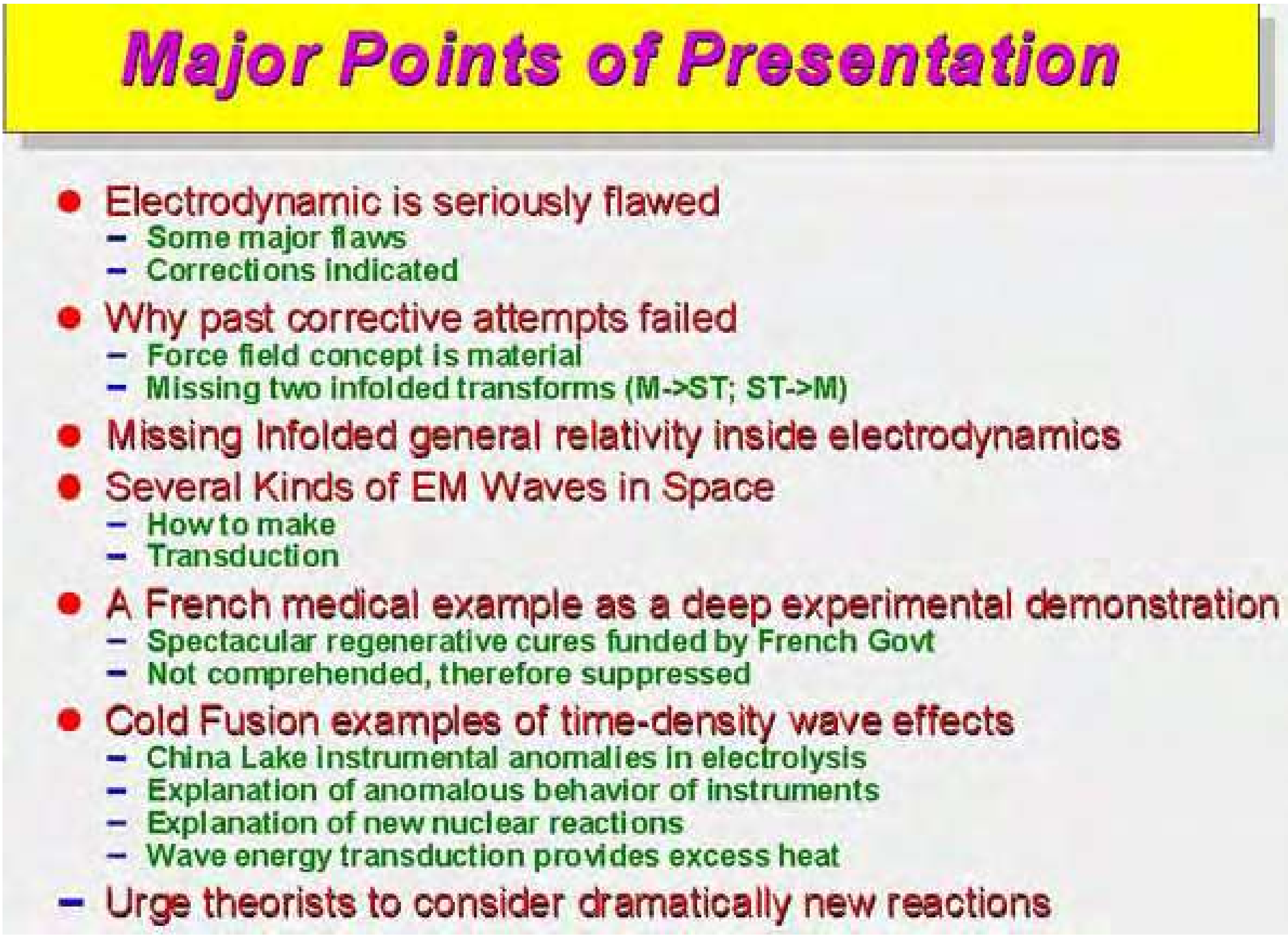

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

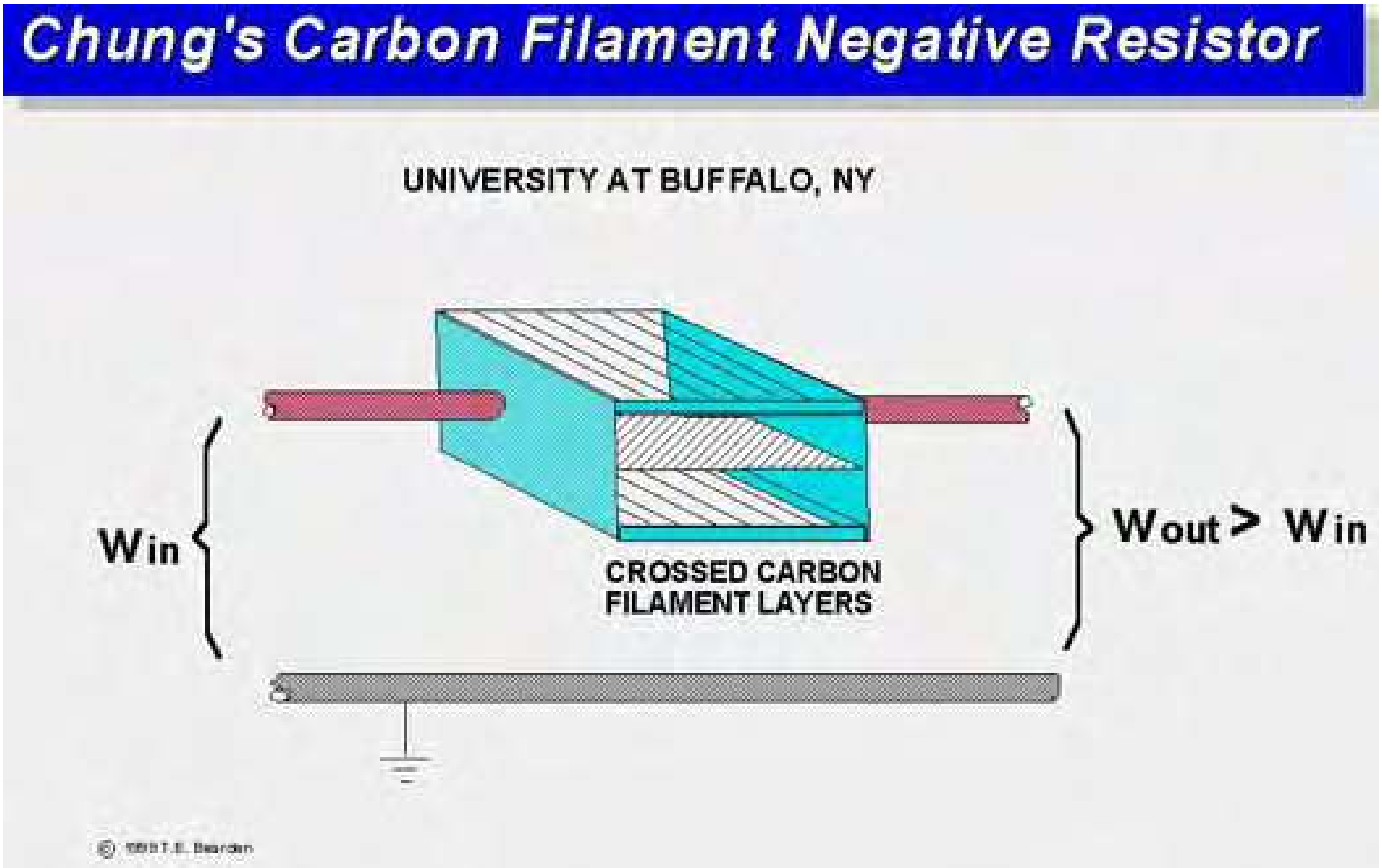

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

Win
CROSSED CARBON FILAMENT LAYERS
Wout > Win

© 1998 T.E. Bearden

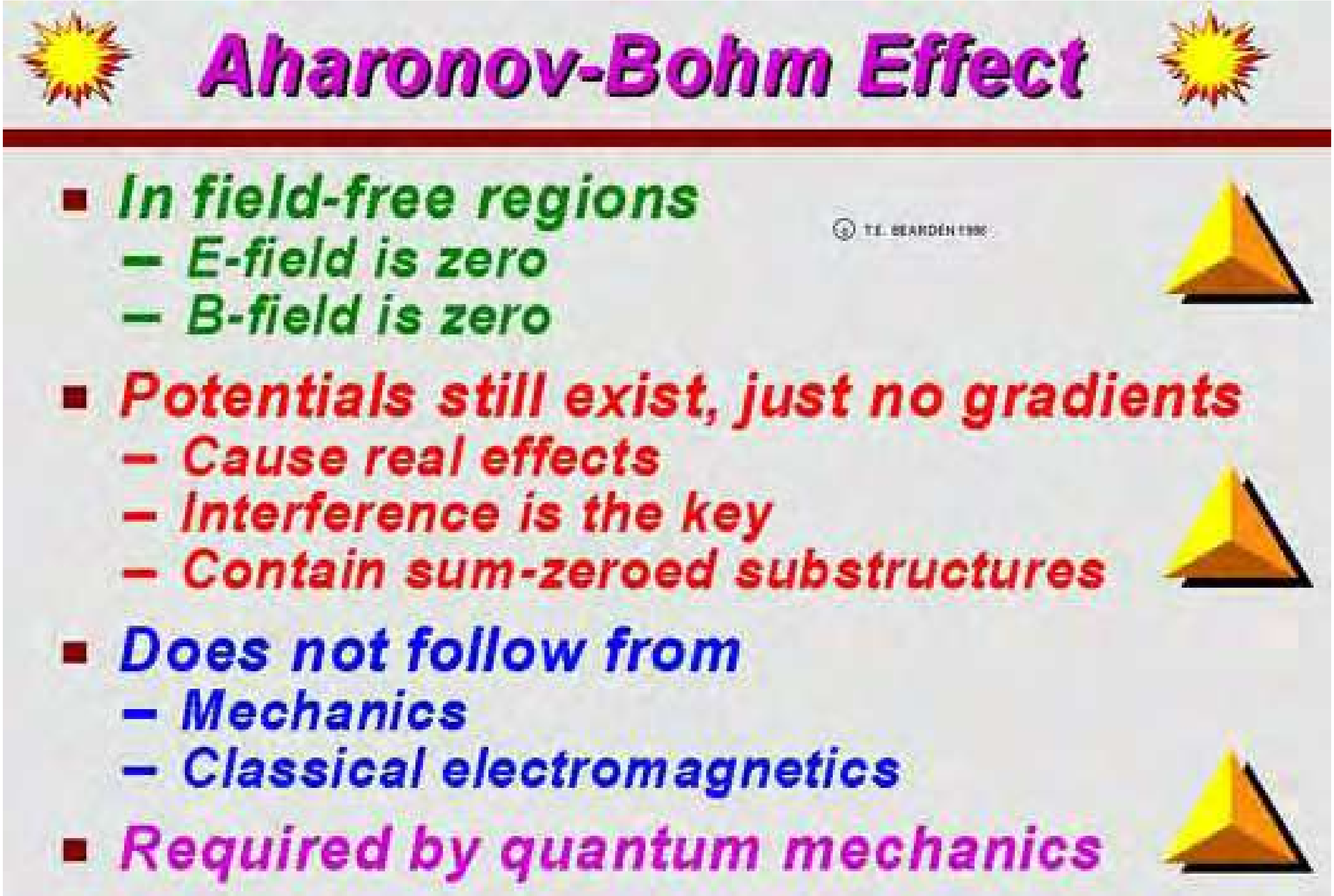

Aharonov-Bohm Effect

© T.E. BEARDEN 1996

- In field-free regions
  - E-field is zero
  - B-field is zero
- Potentials still exist, just no gradients
  - Cause real effects
  - Interference is the key
  - Contain sum-zeroed substructures
- Does not follow from
  - Mechanics
  - Classical electromagnetics
- Required by quantum mechanics

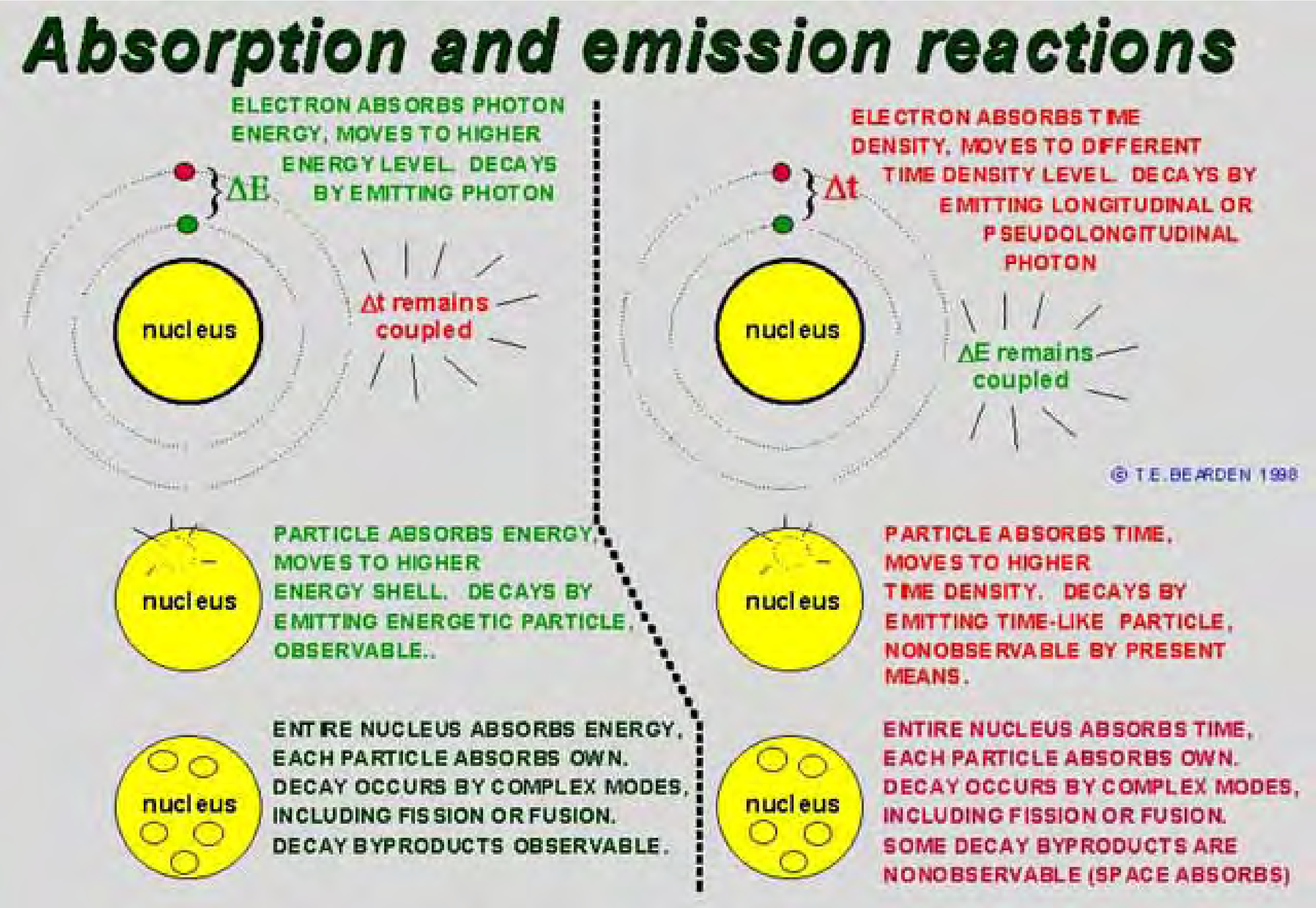

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE. OBSERVABLE..

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

© T.E. BEARDEN 1998

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

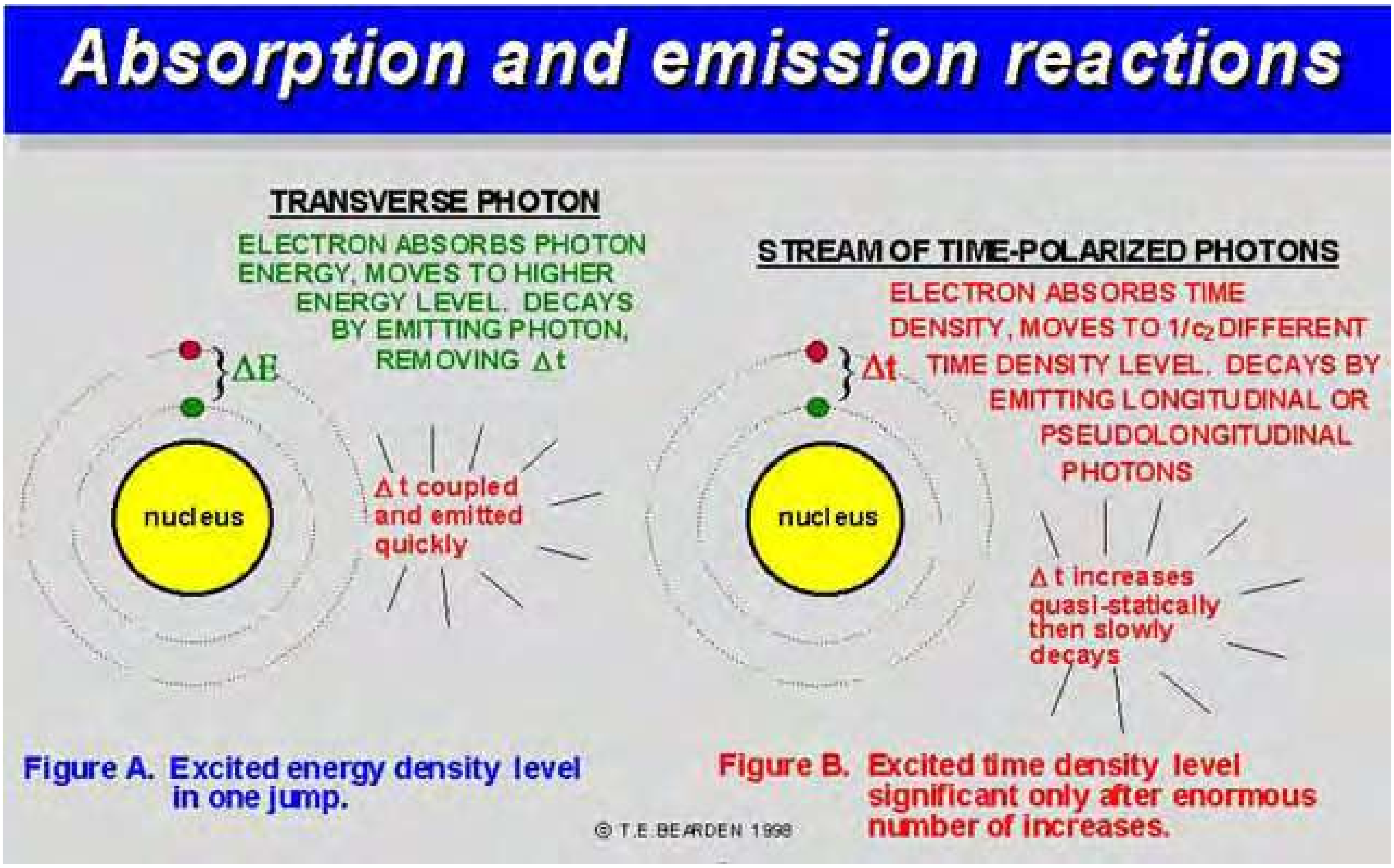

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

Δt coupled and emitted quickly

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c_2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt increases quasi-statically then slowly decays

Figure A. Excited energy density level in one jump.

Figure B. Excited time density level significant only after enormous number of increases.

© T.E. BEARDEN 1998

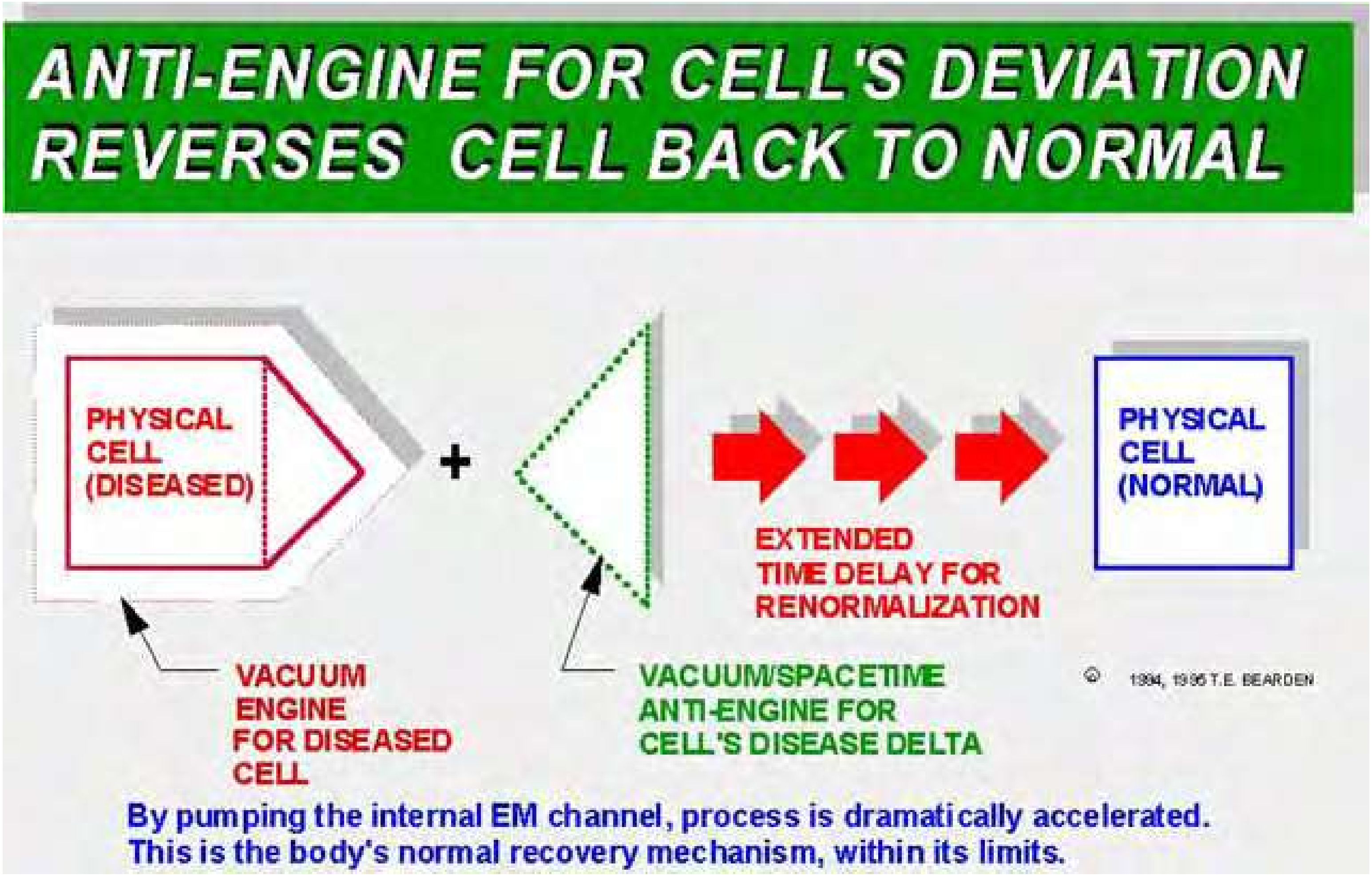

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL
CELL
(DISEASED)

+

PHYSICAL
CELL
(NORMAL)

EXTENDED
TIME DELAY FOR
RENORMALIZATION

VACUUM
ENGINE
FOR DISEASED
CELL

VACUUM/SPACETIME
ANTI-ENGINE FOR
CELL'S DISEASE DELTA

© 1994, 1996 T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated.
This is the body's normal recovery mechanism, within its limits.

Asymmetric Regauging Produces Excess Force, Which Can be Used to Do Work on the System

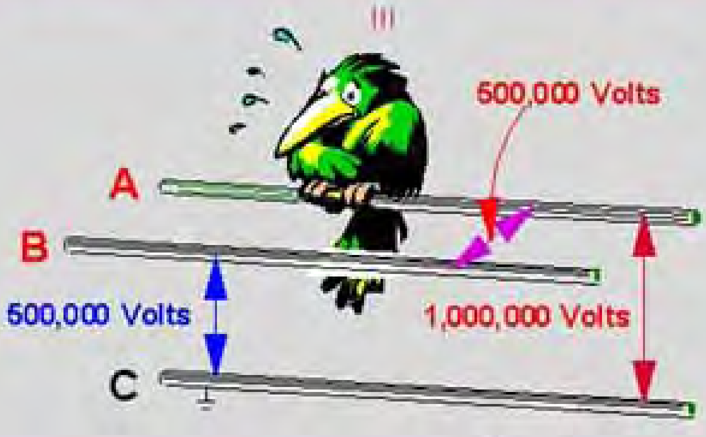

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

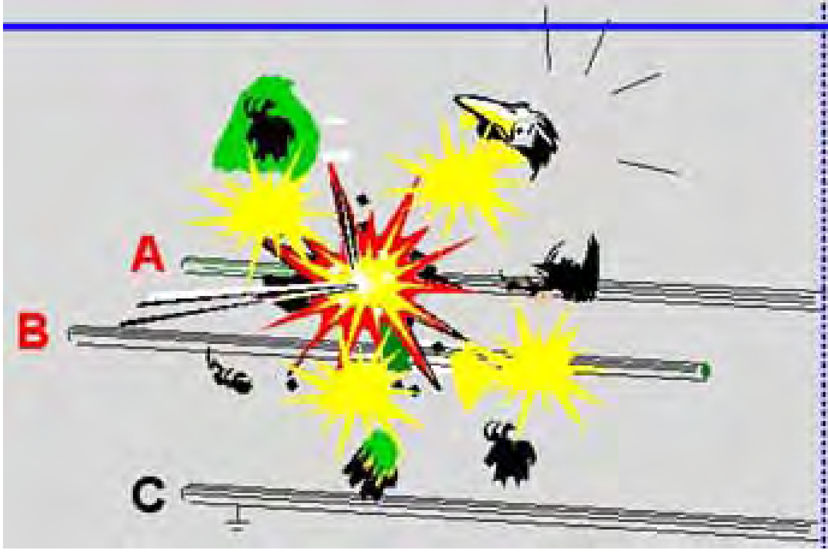

Net Asymmetric Regauging *Can* Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1999

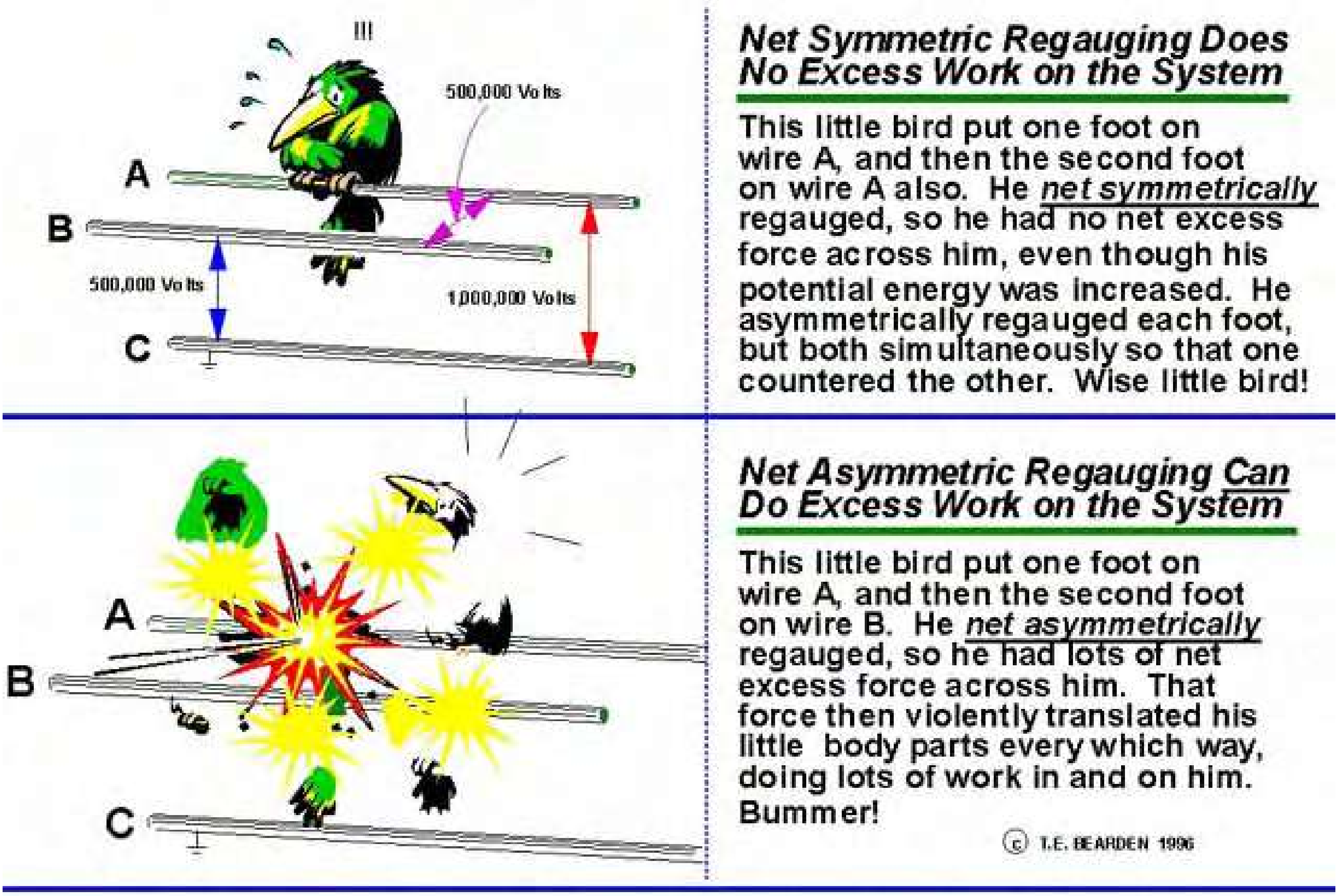

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging Can Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1996

Figure 32. Asymmetric regauging produces excess force, which can be used to do work on the regauging system.

Selected clipart © by Lotus SmartPics™

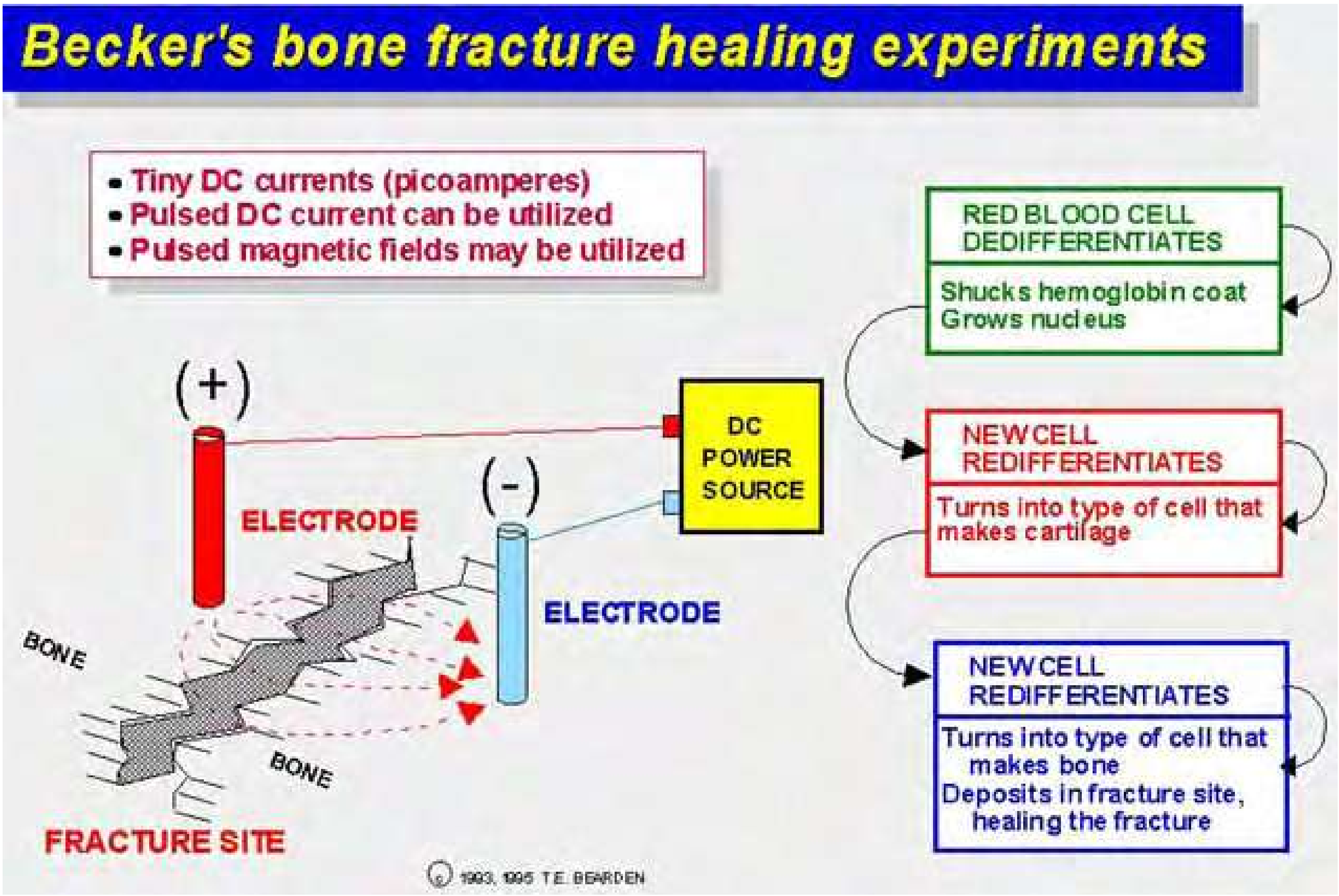

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1995 T.E. BEARDEN

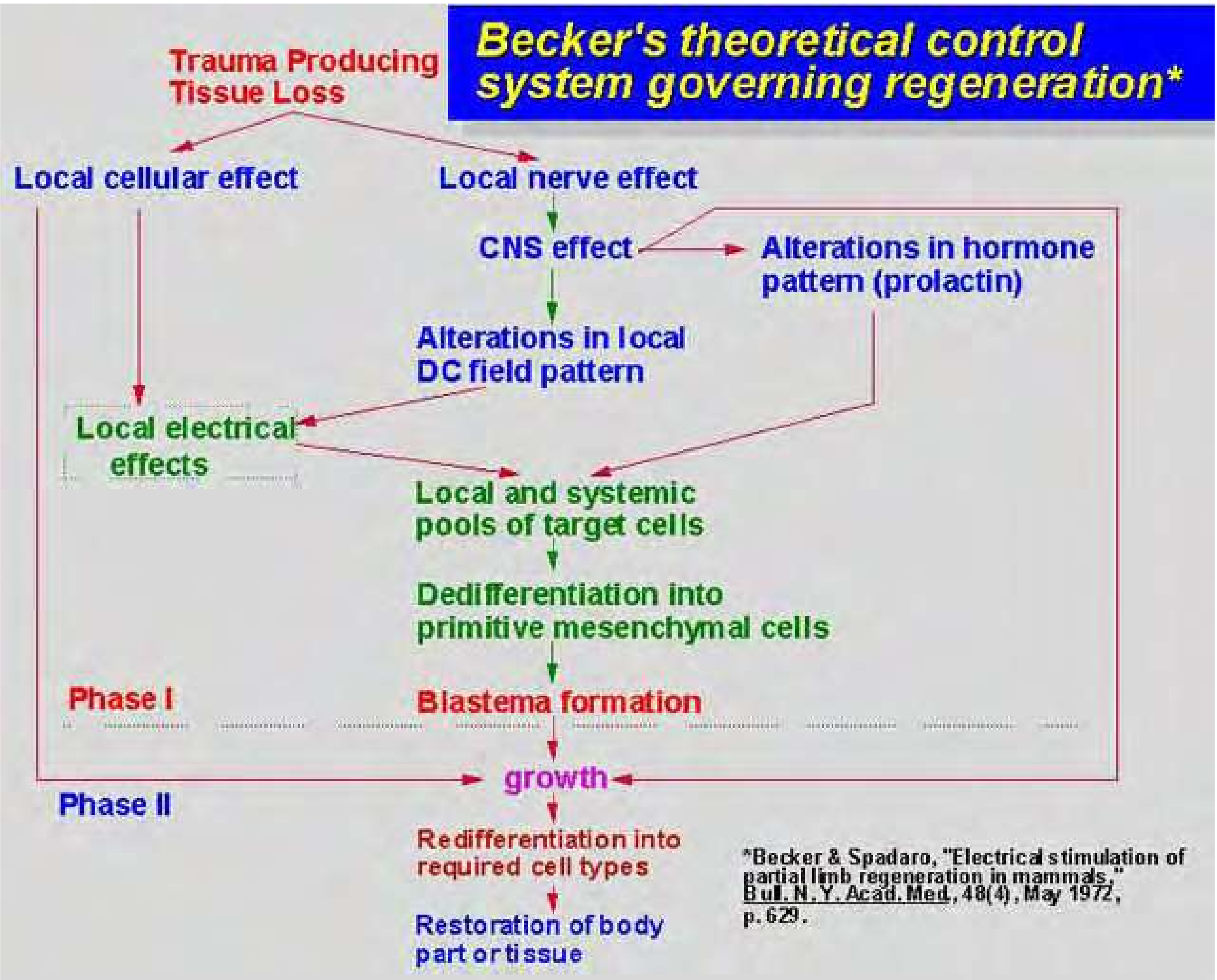

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N.Y. Acad. Med., 48(4), May 1972, p.629.

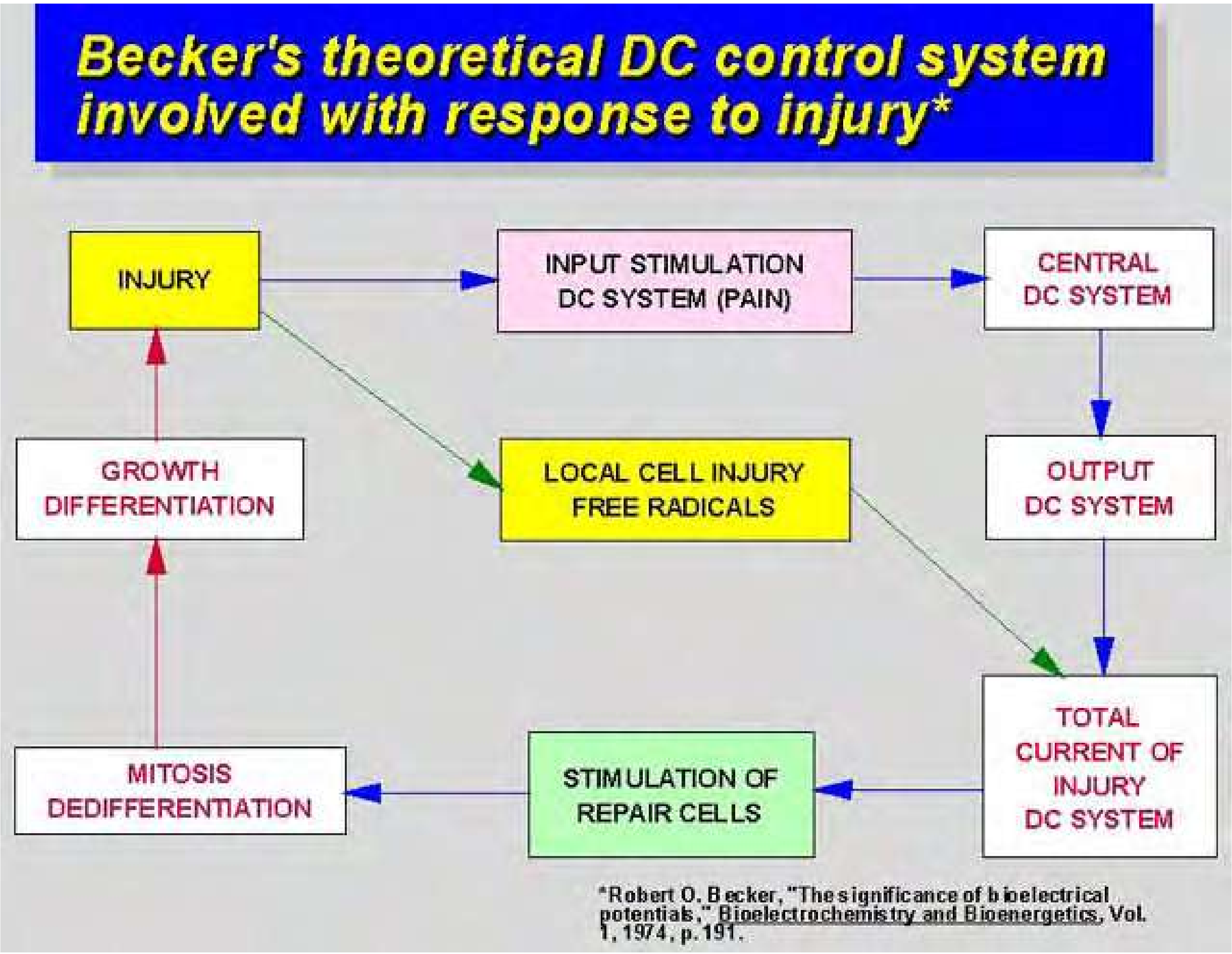

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION
DC SYSTEM (PAIN)

CENTRAL
DC SYSTEM

GROWTH
DIFFERENTIATION

LOCAL CELL INJURY
FREE RADICALS

OUTPUT
DC SYSTEM

TOTAL
CURRENT OF
INJURY
DC SYSTEM

MITOSIS
DEDIFFERENTIATION

STIMULATION OF
REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

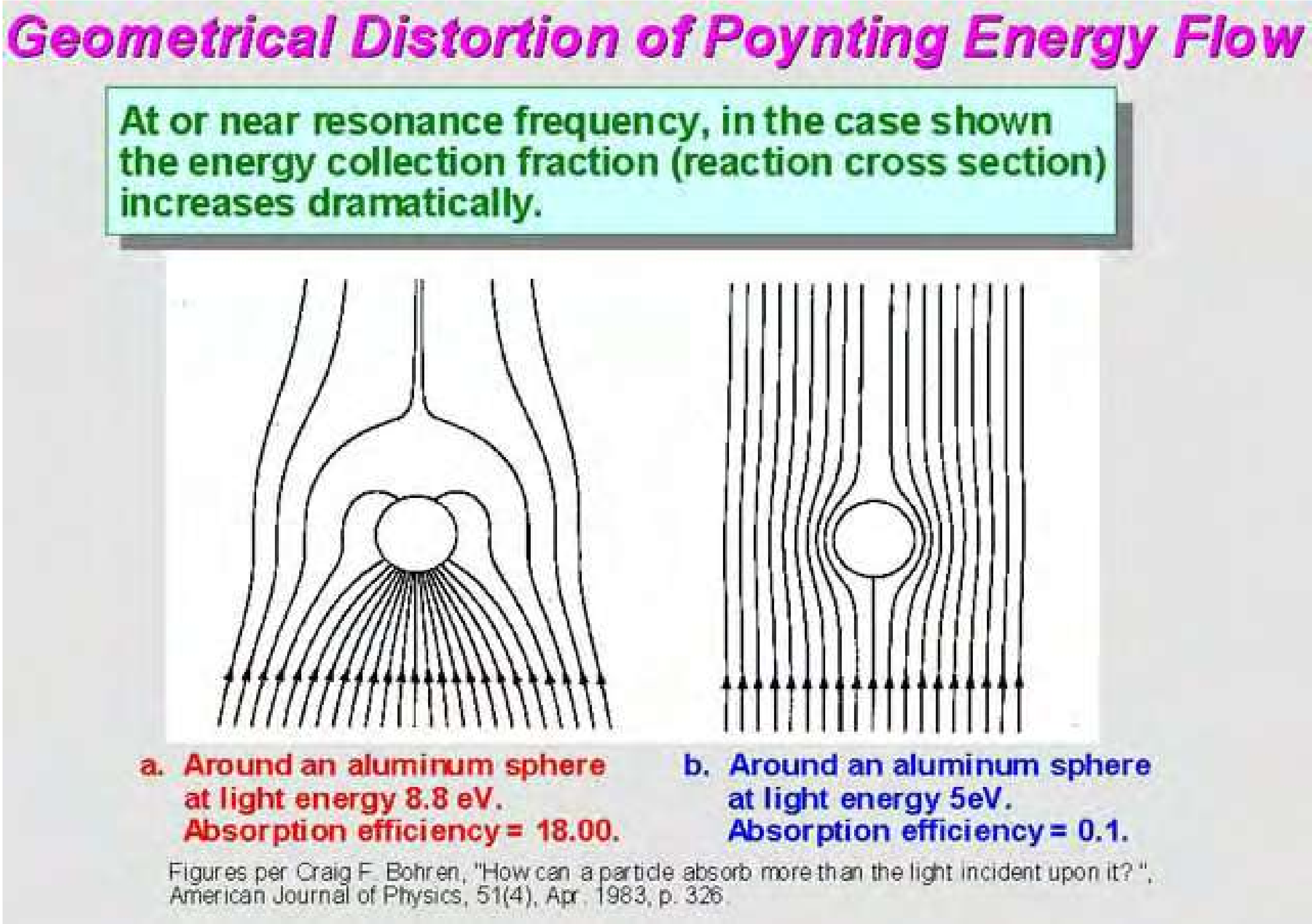

Geometrical Distortion of Poynting Energy Flow

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

a. Around an aluminum sphere at light energy 8.8 eV. Absorption efficiency = 18.00.

b. Around an aluminum sphere at light energy 5eV. Absorption efficiency = 0.1.

Figures per Craig F. Bohren, "How can a particle absorb more than the light incident upon it? ", American Journal of Physics, 51(4), Apr. 1983, p. 326.

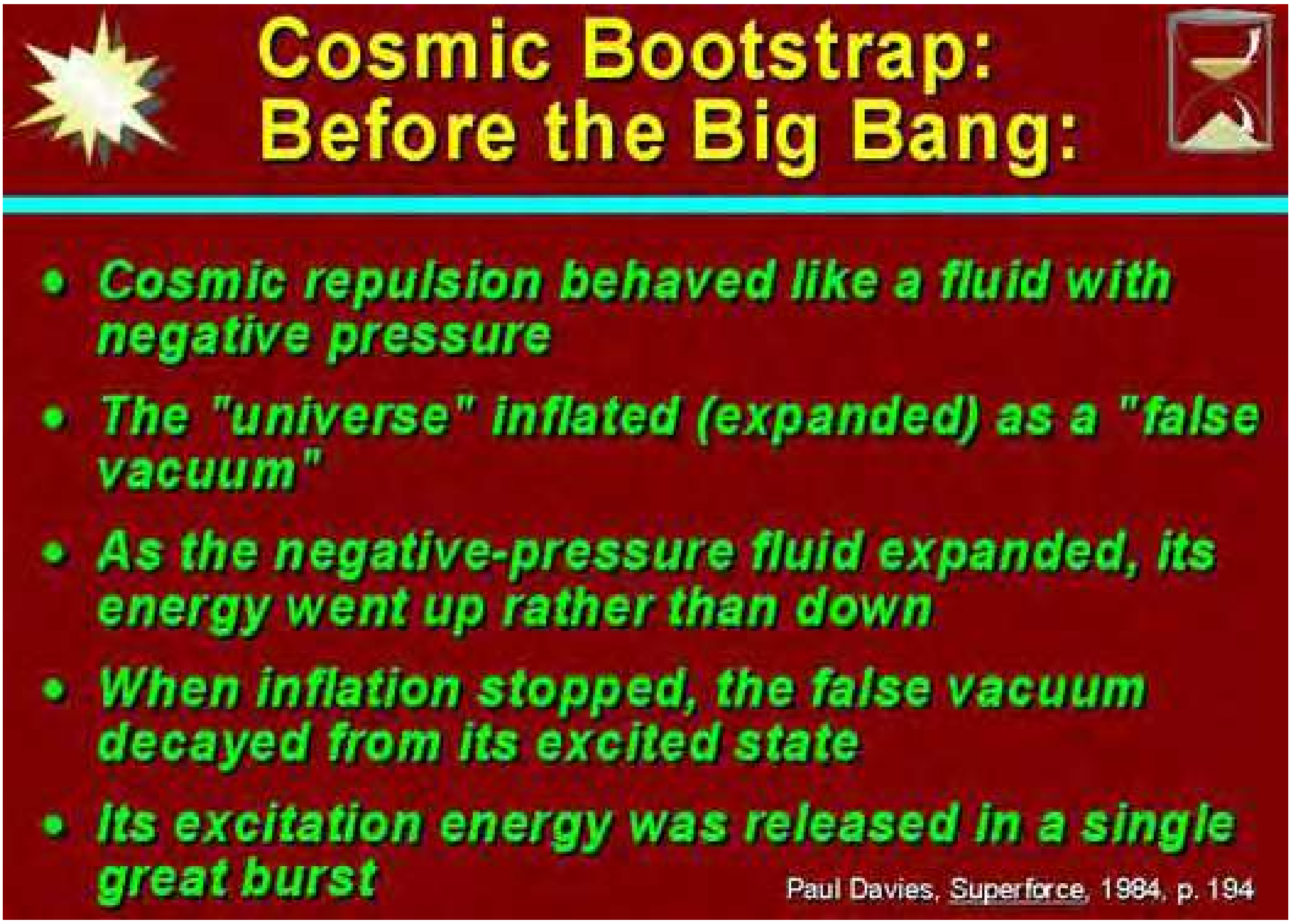

Cosmic Bootstrap: Before the Big Bang:

- *Cosmic repulsion behaved like a fluid with negative pressure*
- *The "universe" inflated (expanded) as a "false vacuum"*
- *As the negative-pressure fluid expanded, its energy went up rather than down*
- *When inflation stopped, the false vacuum decayed from its excited state*
- *Its excitation energy was released in a single great burst*

Paul Davies, Superforce, 1984, p. 194

BRIDGING ENABLES COP > 1.0

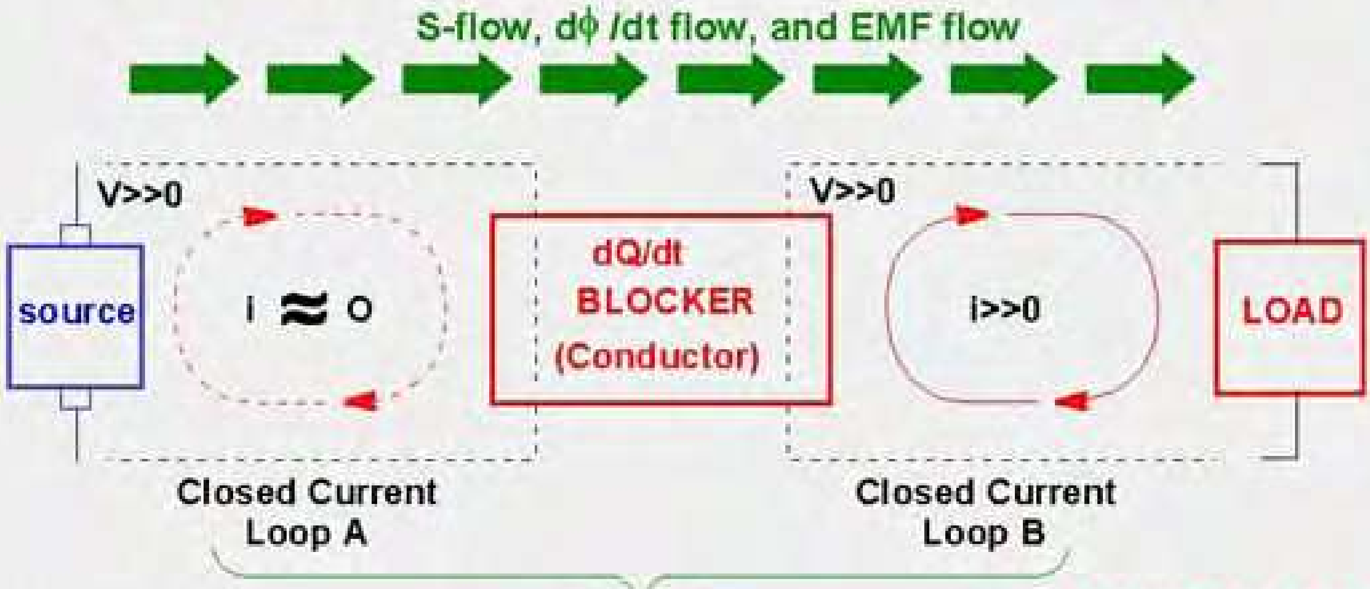

S-flow, dϕ /dt flow, and EMF flow

V>>0

source

i ≈ 0

dQ/dt
BLOCKER
(Conductor)

V>>0

i>>0

LOAD

Closed Current
Loop A

Closed Current
Loop B

BRIDGE

Current loop B is asymmetrically regauged by
S, dϕ /dt, and EMF flow from current loop A.

Patent Pending

© 1998 T.E. BEARDEN

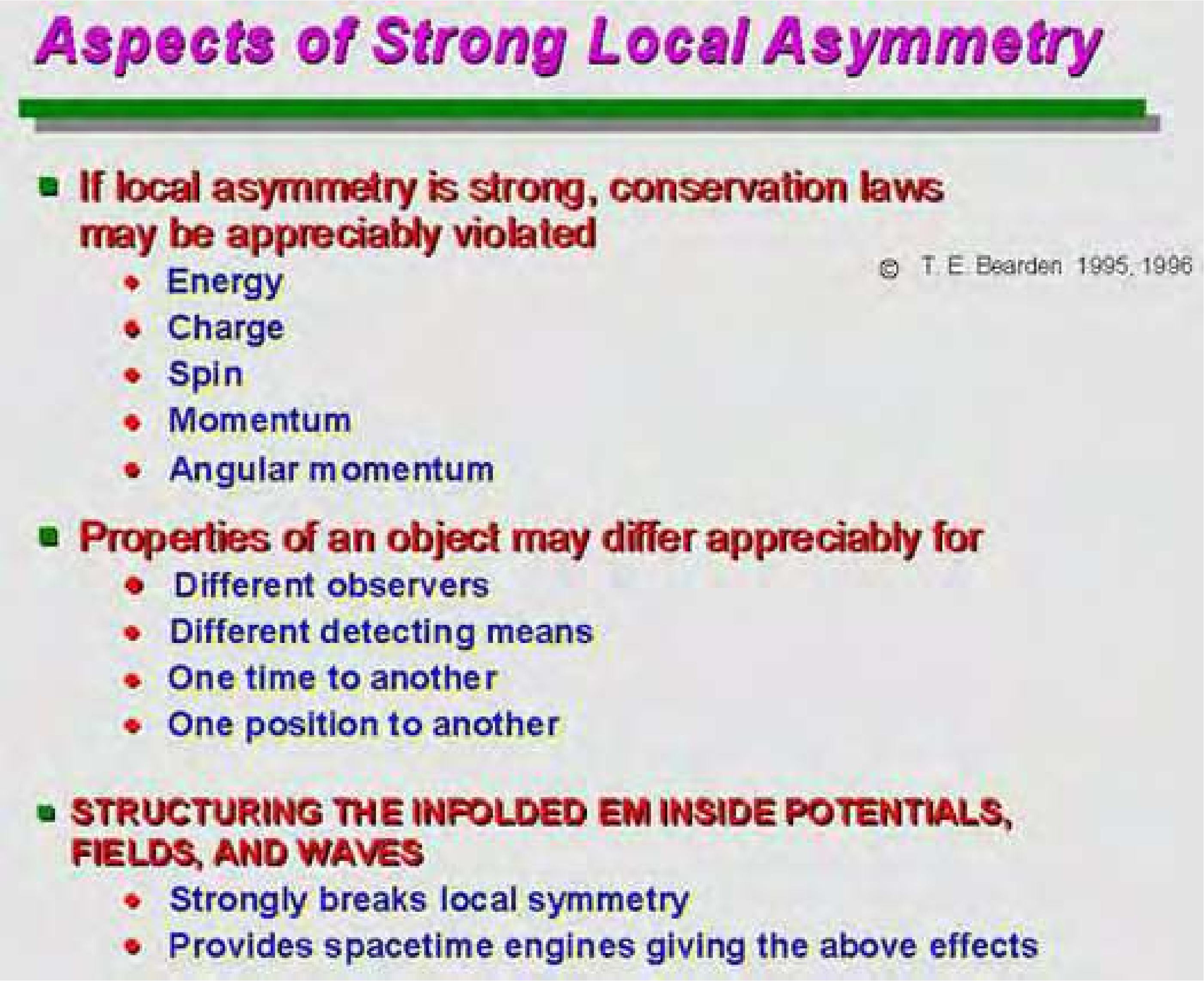

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
    - Energy
    - Charge
    - Spin
    - Momentum
    - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
    - Different observers
    - Different detecting means
    - One time to another
    - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
    - Strongly breaks local symmetry
    - Provides spacetime engines giving the above effects

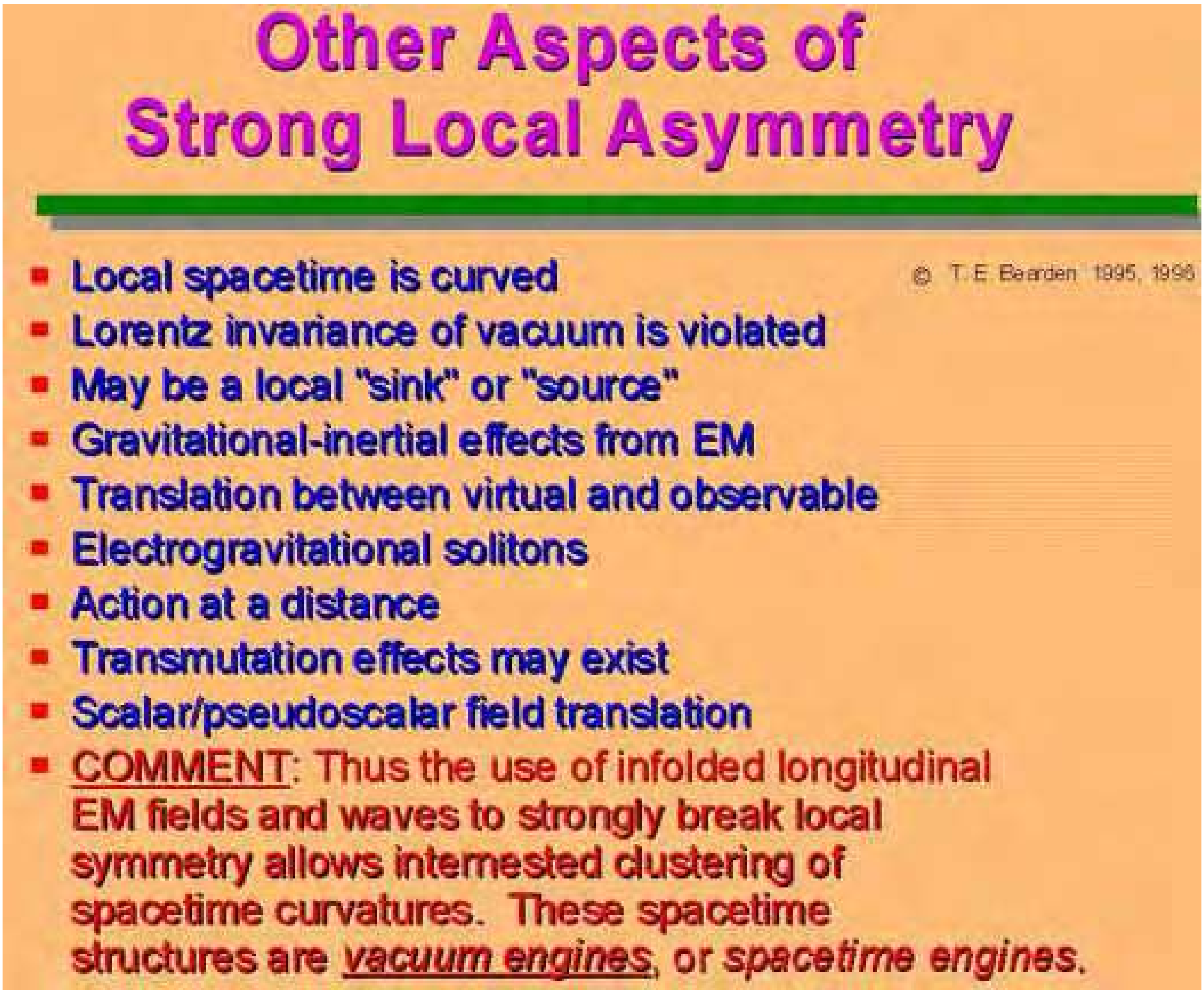

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational-inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

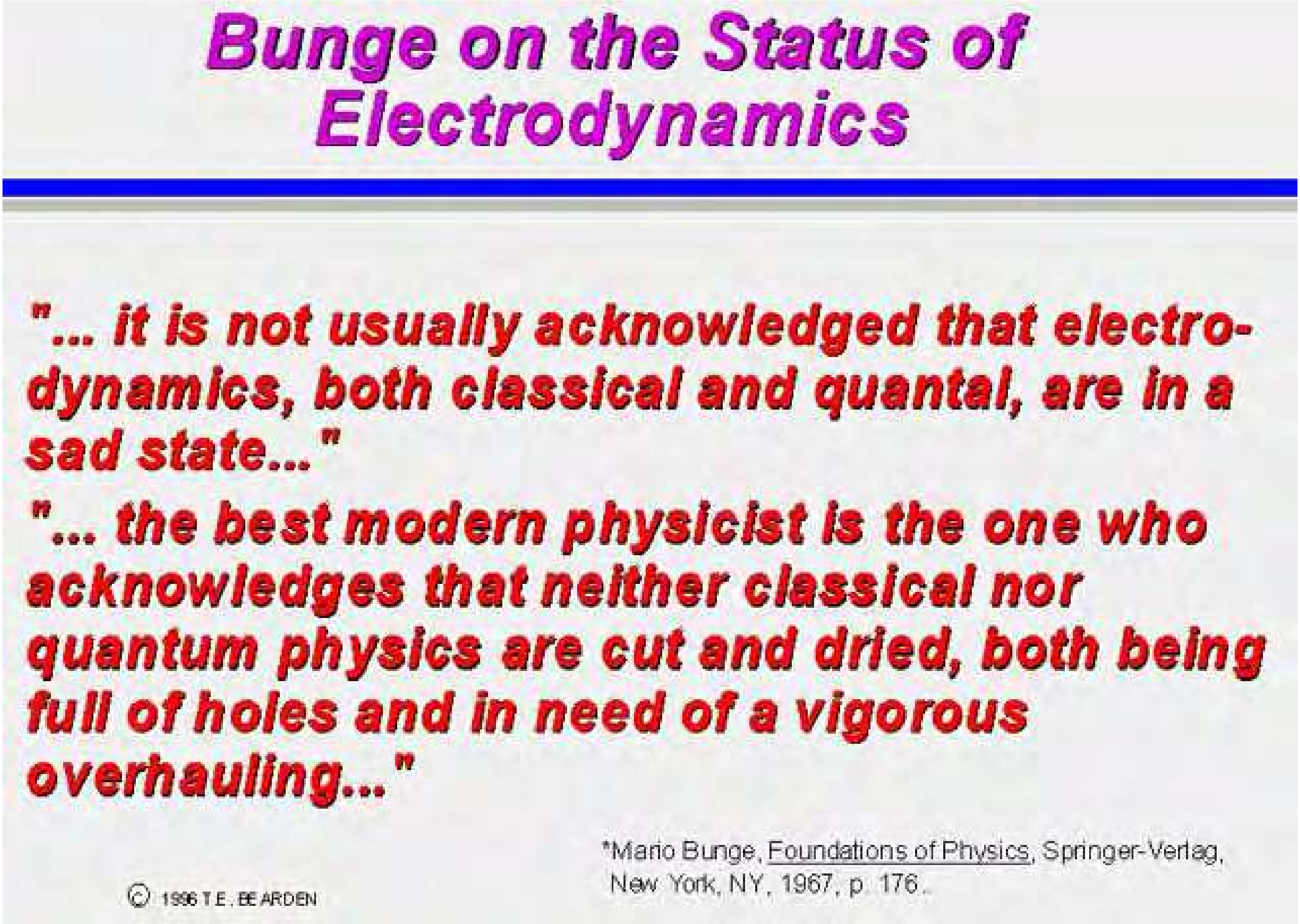

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176.

© 1996 T.E. BEARDEN

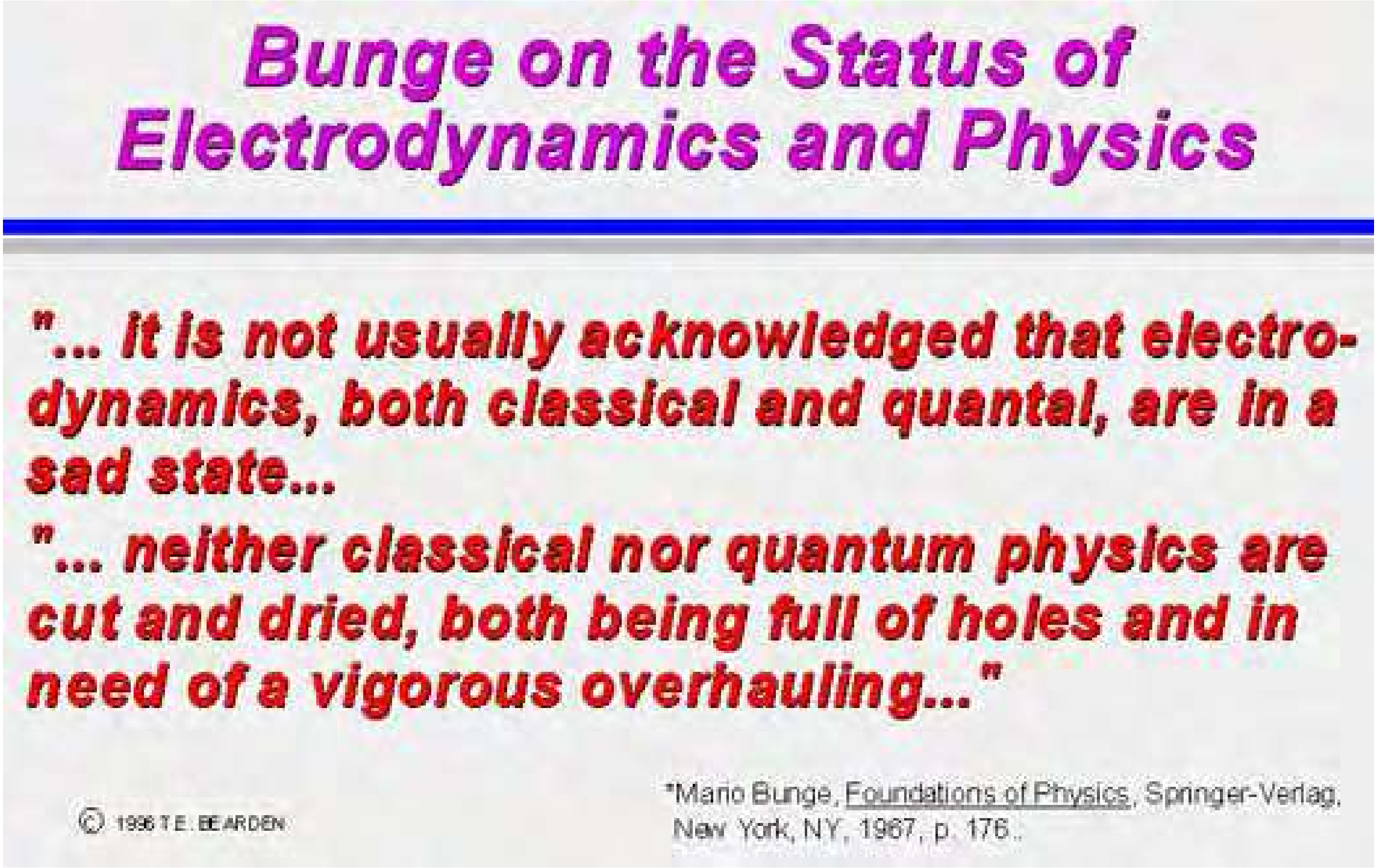

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

© 1996 T.E. BEARDEN

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

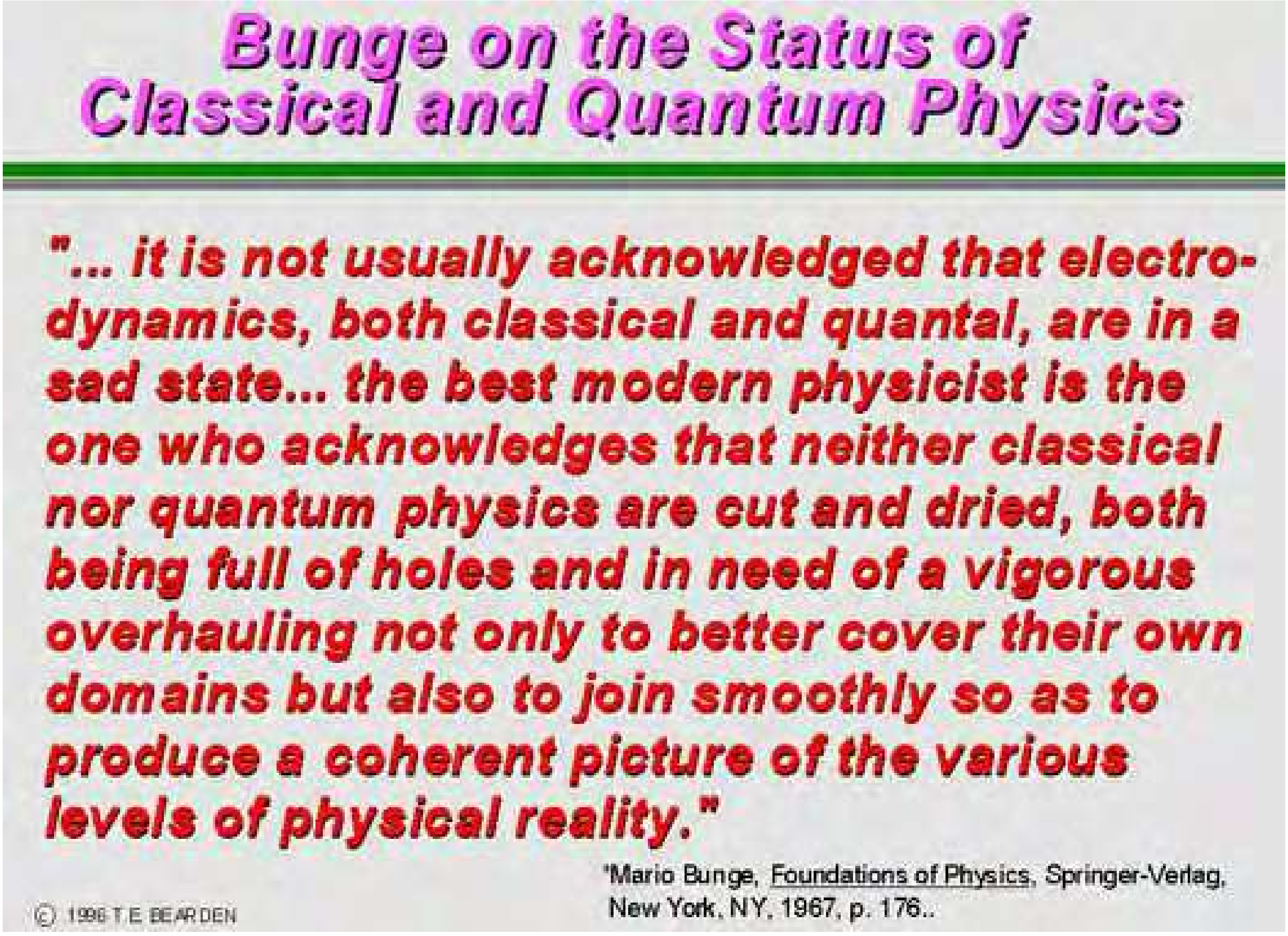

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

'Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

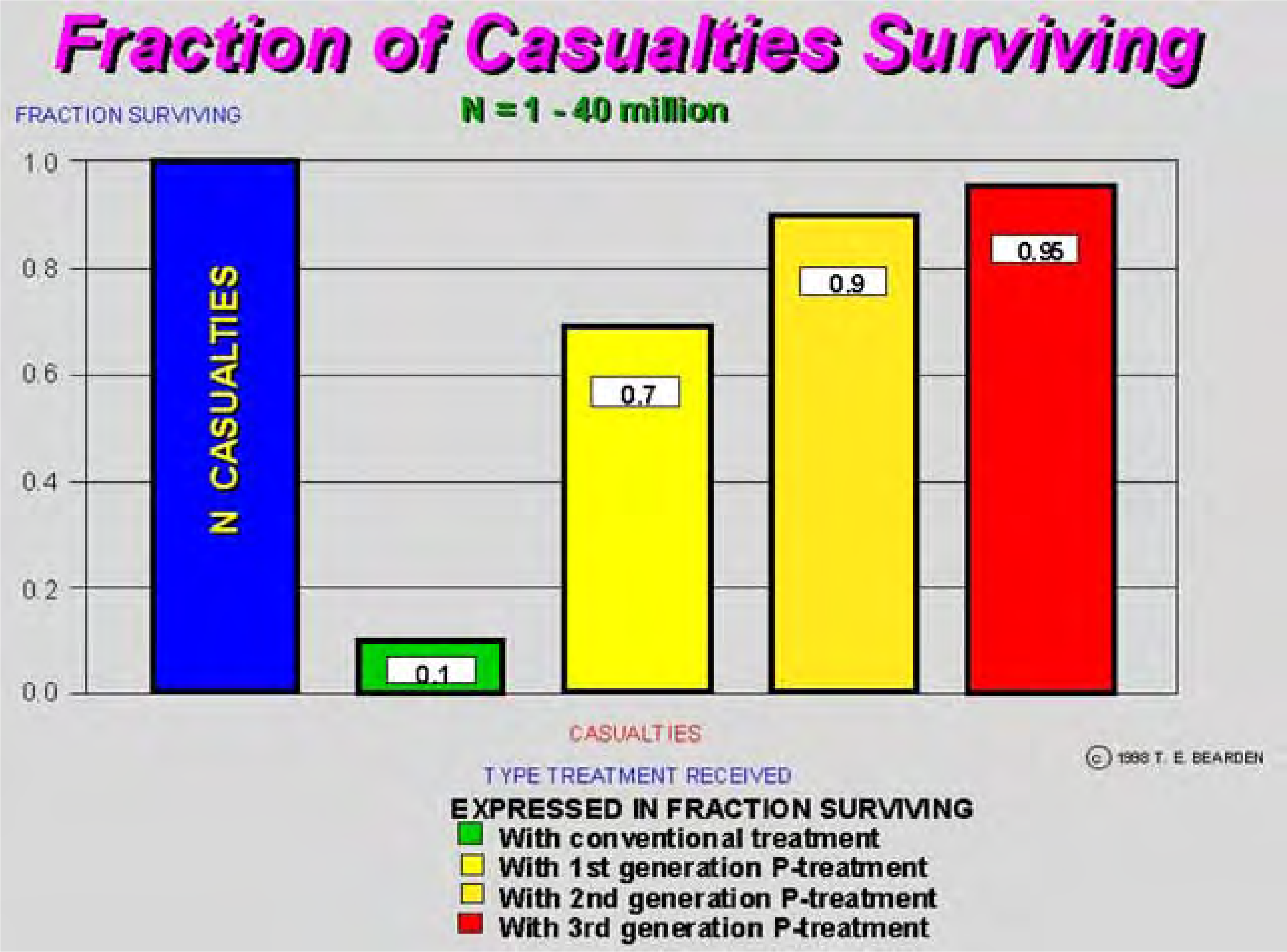
Fraction of Casualties Surviving
N = 1 - 40 million
FRACTION SURVIVING
1.0
0.8
0.6
0.4
0.2
0.0
N CASUALTIES
0.1
0.7
0.9
0.95
CASUALTIES
TYPE TREATMENT RECEIVED
© 1998 T. E. BEARDEN
EXPRESSED IN FRACTION SURVIVING
🟩 With conventional treatment
🟨 With 1st generation P-treatment
🟨 With 2nd generation P-treatment
🟥 With 3rd generation P-treatment

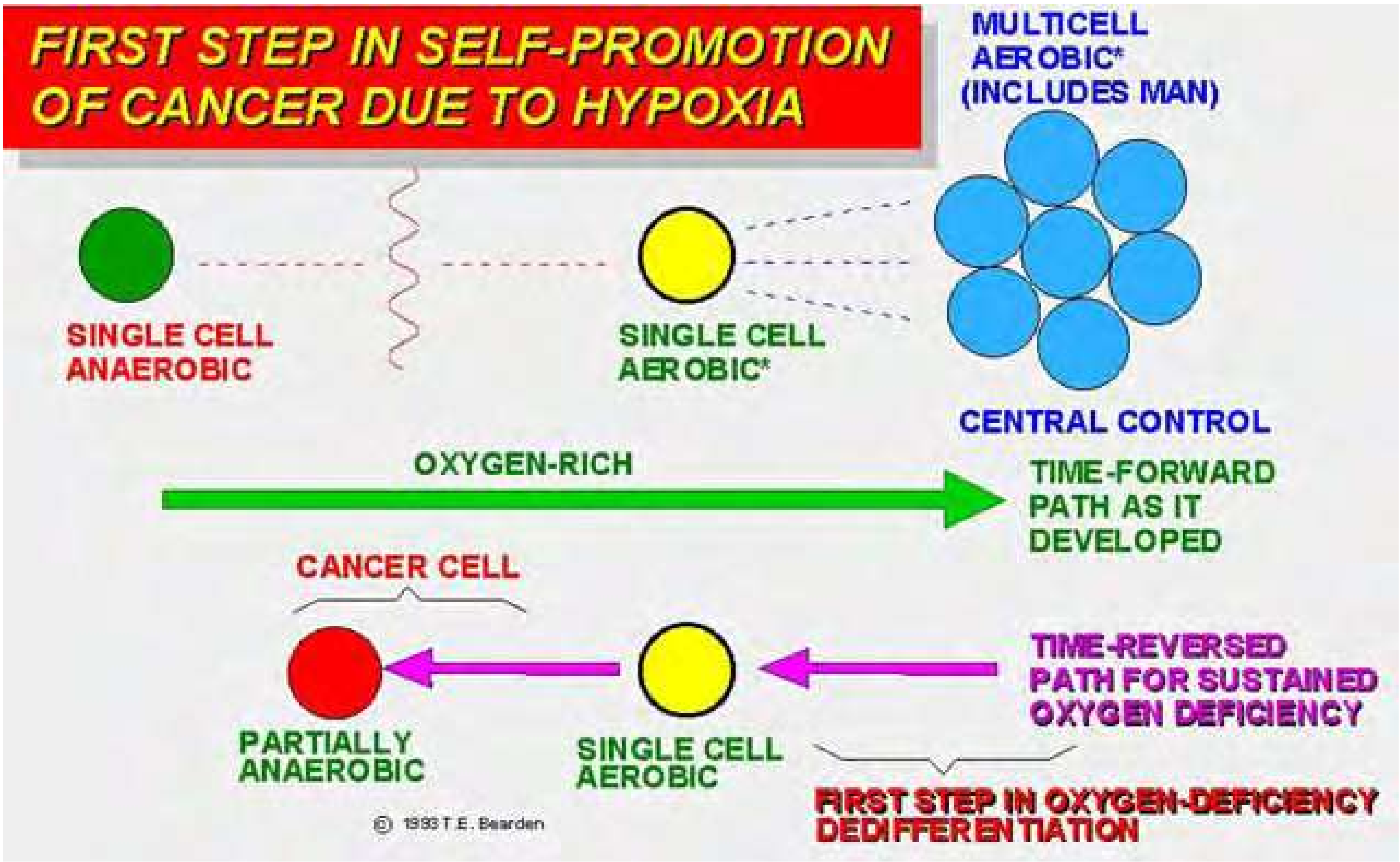

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

MULTICELL
AEROBIC*
(INCLUDES MAN)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

© 1993 T.E. Bearden

Differentiation and Dedifferentiation Timelines

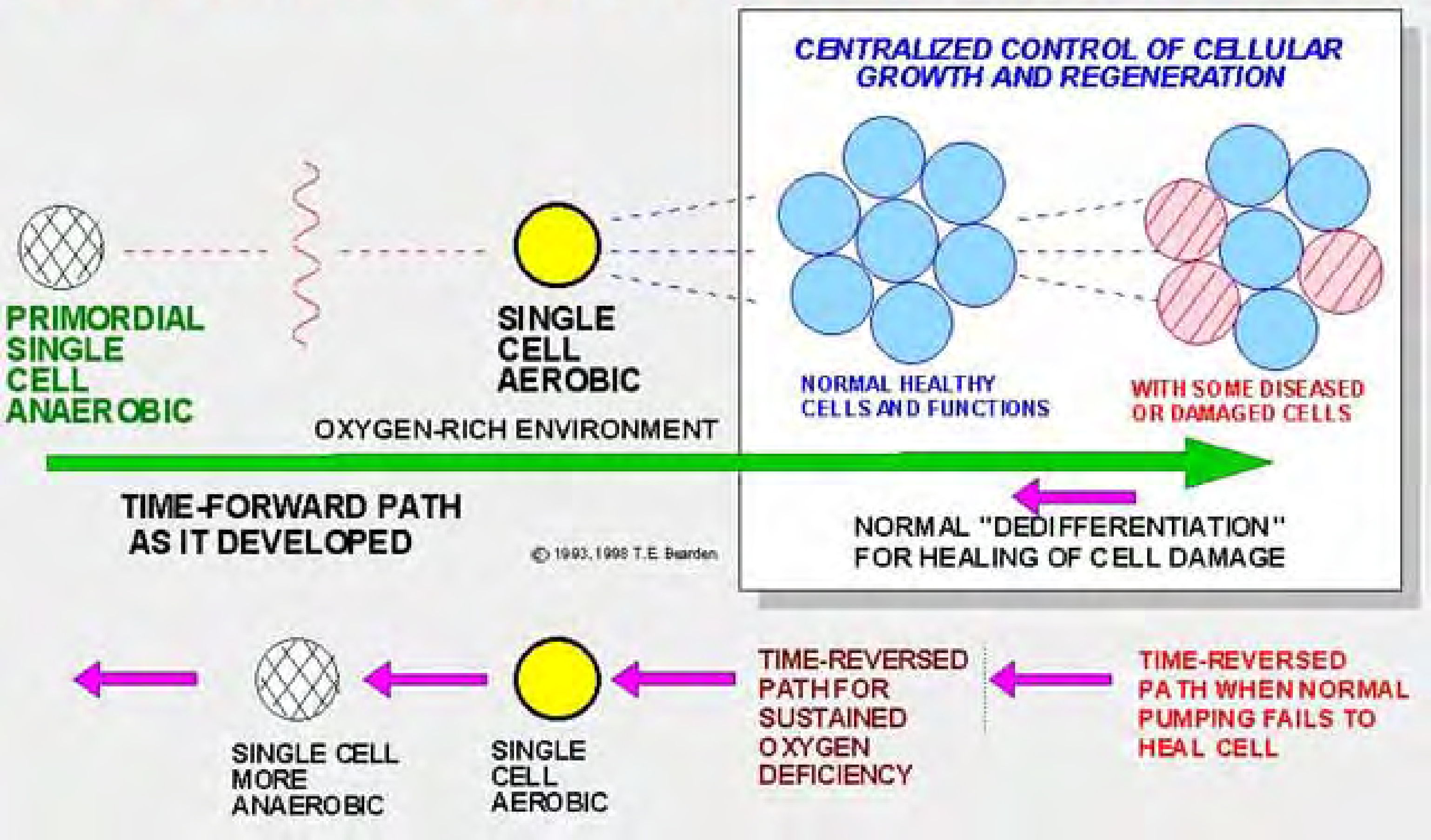
CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION
PRIMORDIAL SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC
OXYGEN-RICH ENVIRONMENT
NORMAL HEALTHY CELLS AND FUNCTIONS
WITH SOME DISEASED OR DAMAGED CELLS
TIME-FORWARD PATH AS IT DEVELOPED
NORMAL "DEDIFFERENTIATION" FOR HEALING OF CELL DAMAGE
SINGLE CELL MORE ANAEROBIC
SINGLE CELL AEROBIC
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
TIME-REVERSED PATH WHEN NORMAL PUMPING FAILS TO HEAL CELL

© 1993, 1998 T.E. Bearden

Evolution of Cells on Earth

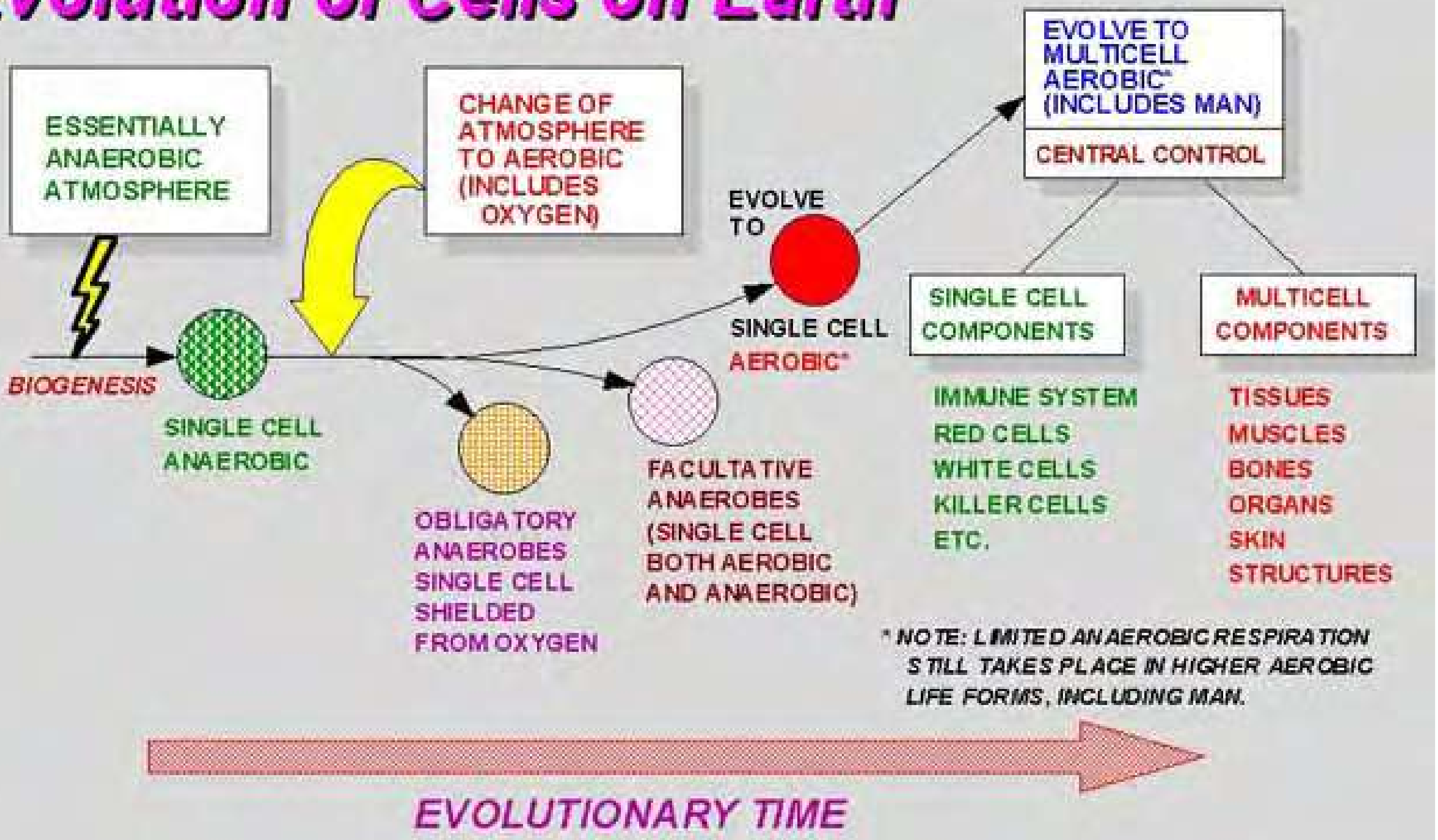

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

*NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

(c) 1999 T. E. BEARDEN

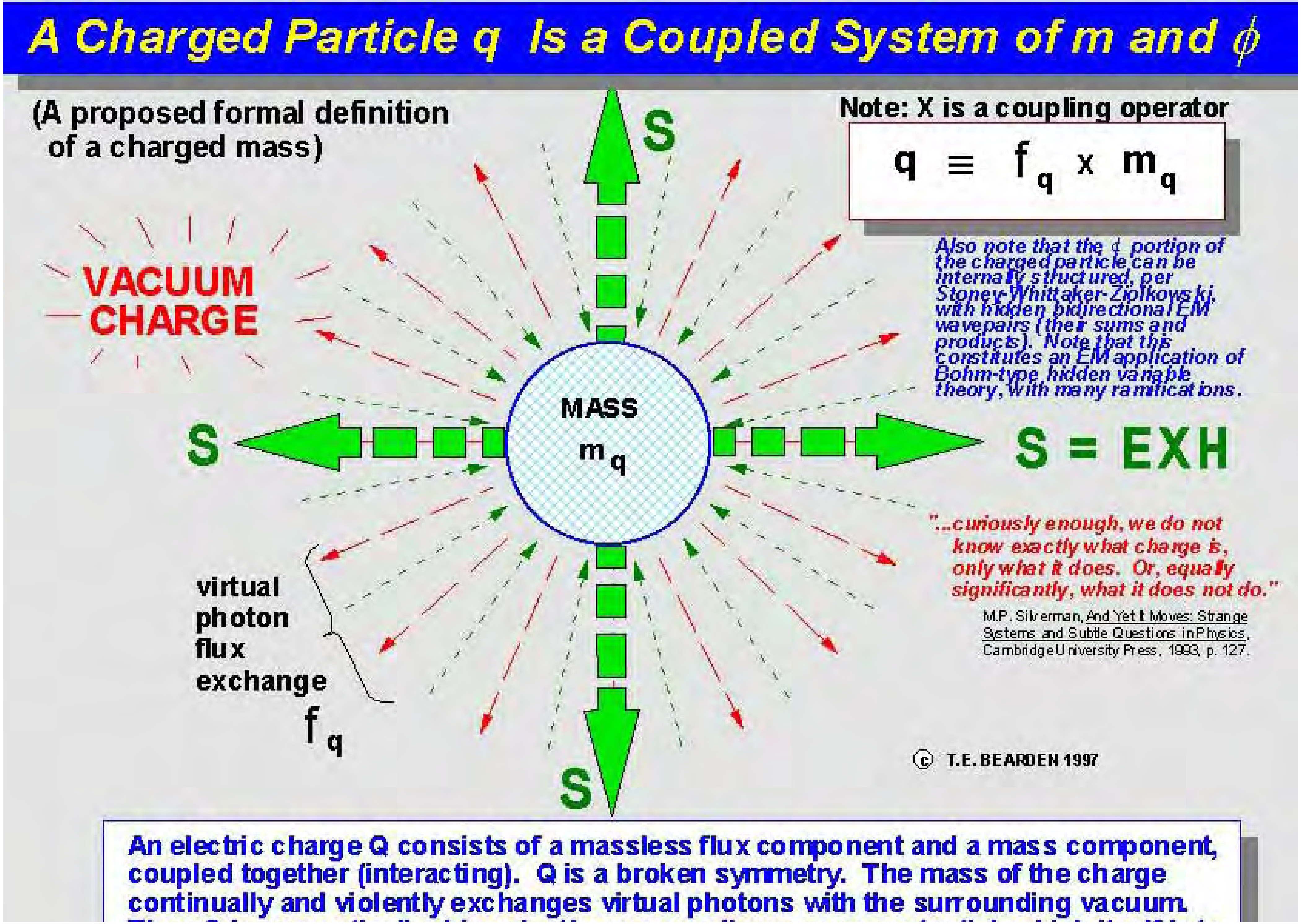

A Charged Particle q Is a Coupled System of m and ϕ

(A proposed formal definition of a charged mass)

Note: X is a coupling operator

$$q \equiv f_q \times m_q$$

Also note that the ϕ portion of the charged particle can be internally structured, per Stoney-Whittaker-Ziolkowski, with hidden bidirectional EM wavepairs (their sums and products). Note that this constitutes an EM application of Bohm-type hidden variable theory, with many ramifications.

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

© T.E. BEARDEN 1997

An electric charge Q consists of a massless flux component and a mass component, coupled together (interacting). Q is a broken symmetry. The mass of the charge continually and violently exchanges virtual photons with the surrounding vacuum.

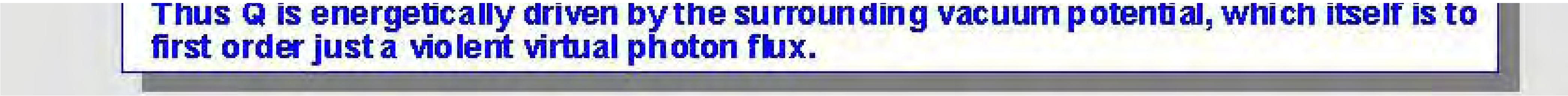

Thus Q is energetically driven by the surrounding vacuum potential, which itself is to first order just a violent virtual photon flux.

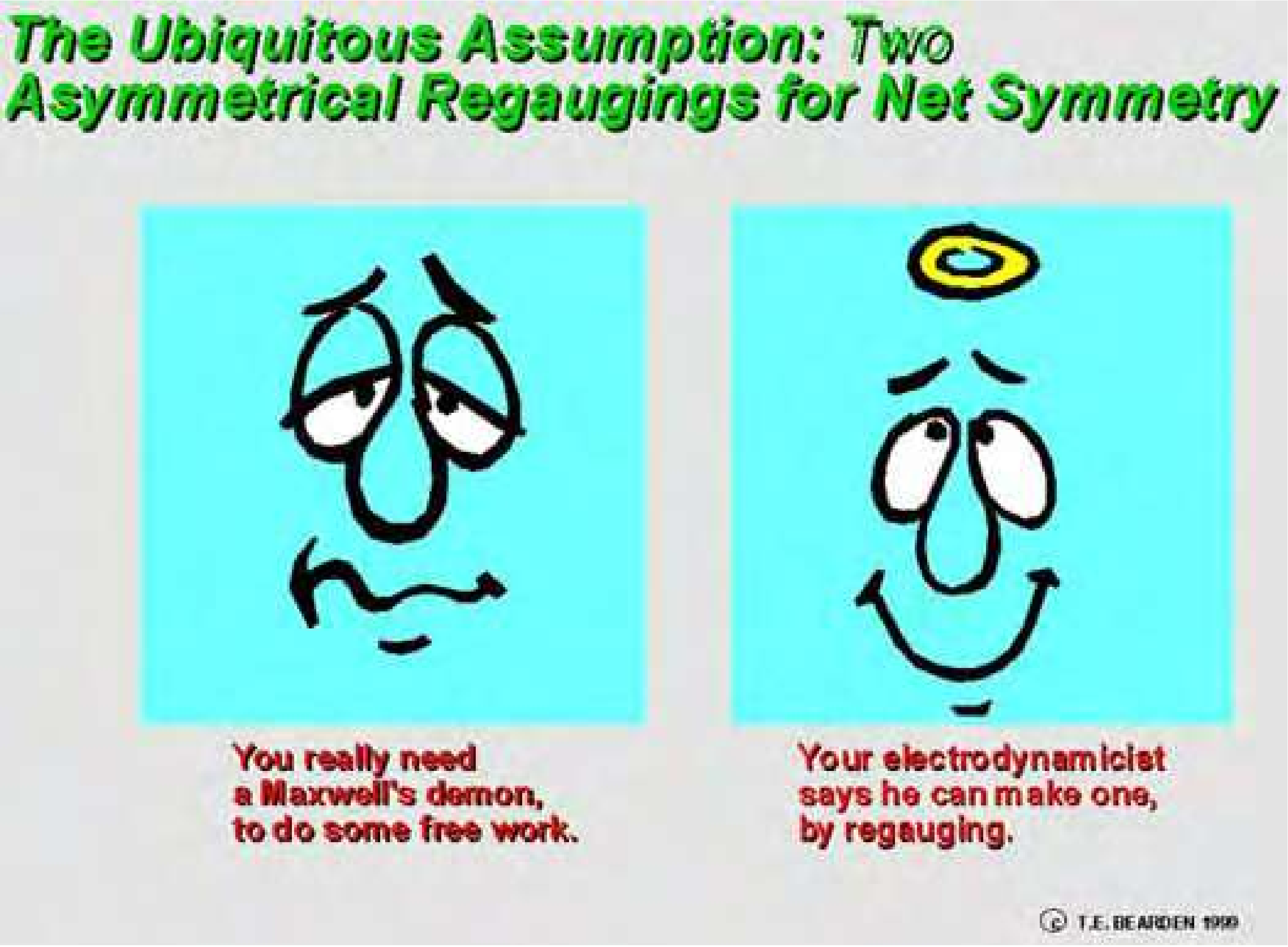

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

You really need
a Maxwell's demon,
to do some free work.

Your electrodynamicist
says he can make one,
by regauging.

© T.E. BEARDEN 1999

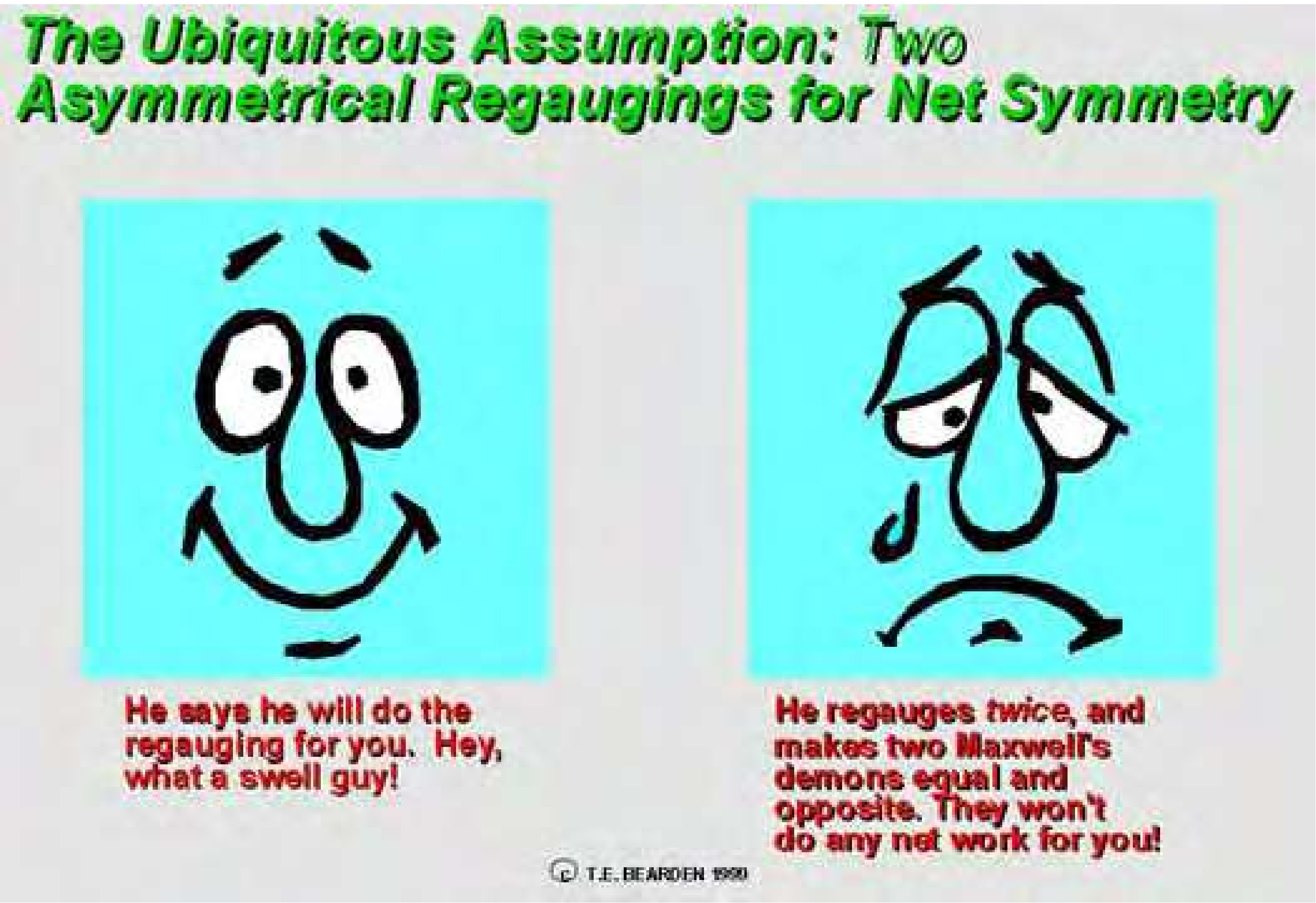

The Ubiquitous Assumption: *Two* Asymmetrical Regaugings for Net Symmetry

He says he will do the regauging for you. Hey, what a swell guy!

He regauges *twice*, and makes two Maxwell's demons equal and opposite. They won't do any net work for you!

© T.E. BEARDEN 1999

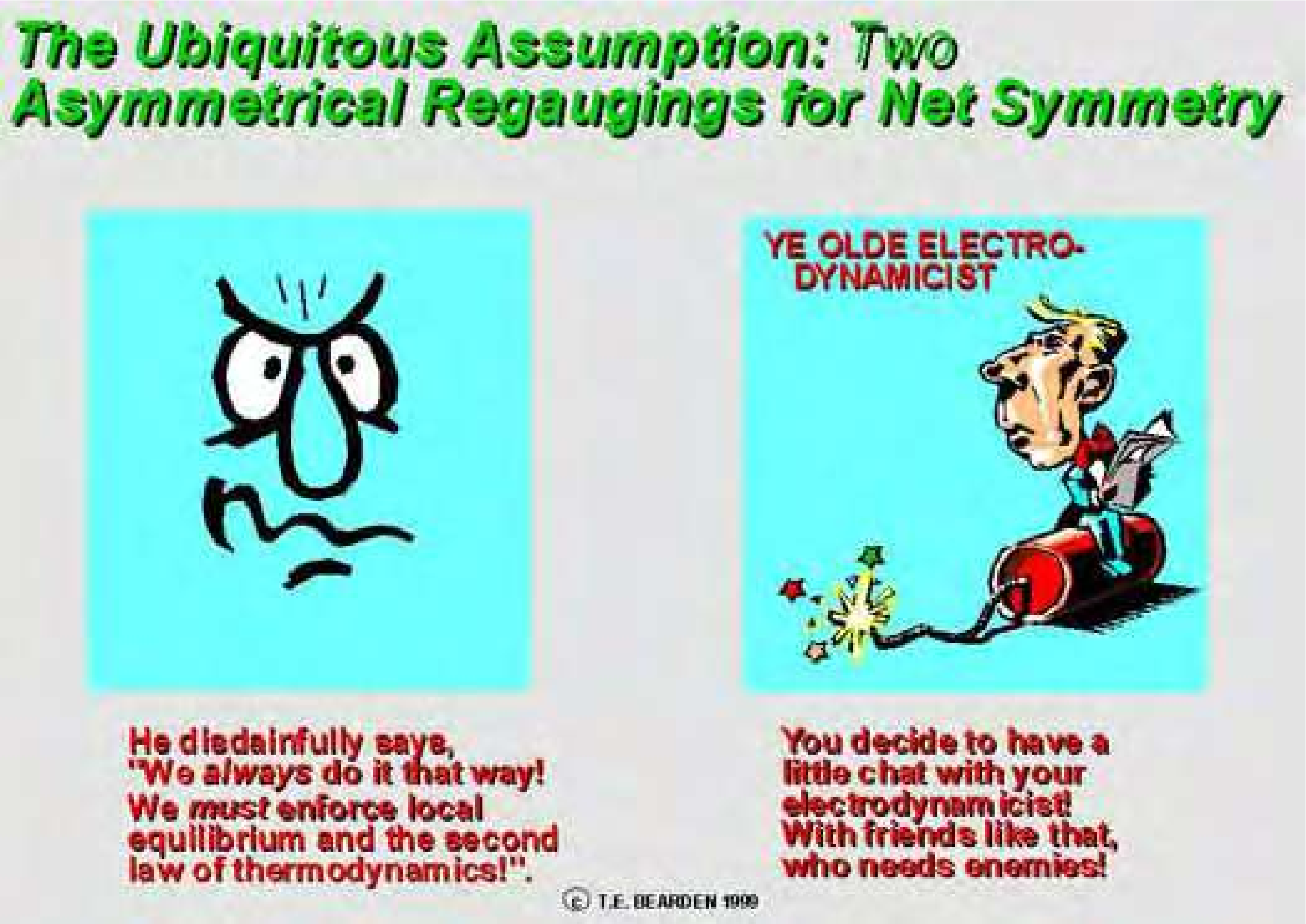

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

He disdainfully says,
"We *always* do it that way!
We *must* enforce local
equilibrium and the second
law of thermodynamics!".

You decide to have a
little chat with your
electrodynamicist!
With friends like that,
who needs enemies!

© T.E. BEARDEN 1999

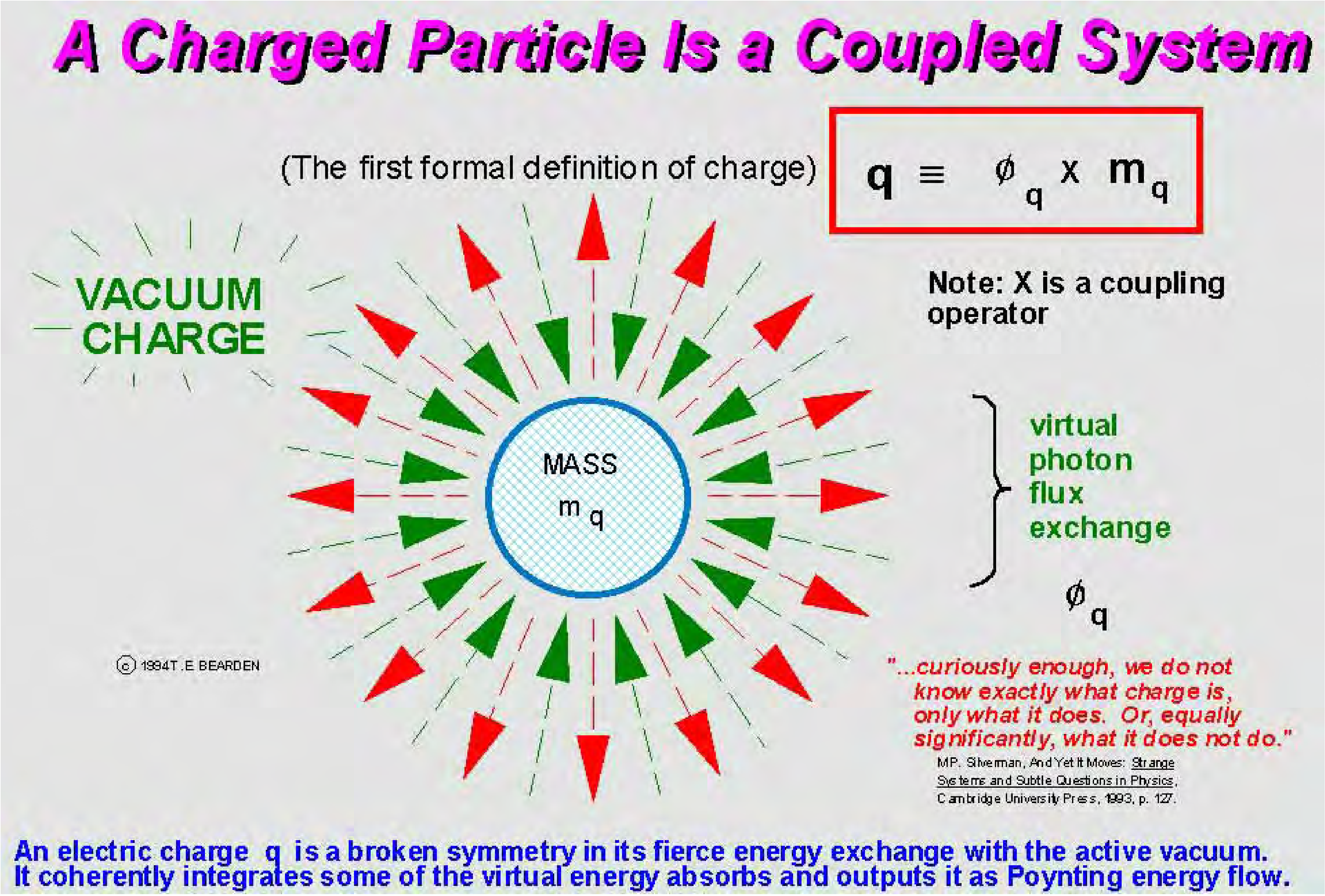

A Charged Particle Is a Coupled System

(The first formal definition of charge)

$$\boxed{q \equiv \phi_q \times m_q}$$

Note: X is a coupling operator

© 1994 T.E BEARDEN

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge q is a broken symmetry in its fierce energy exchange with the active vacuum. It coherently integrates some of the virtual energy absorbs and outputs it as Poynting energy flow.

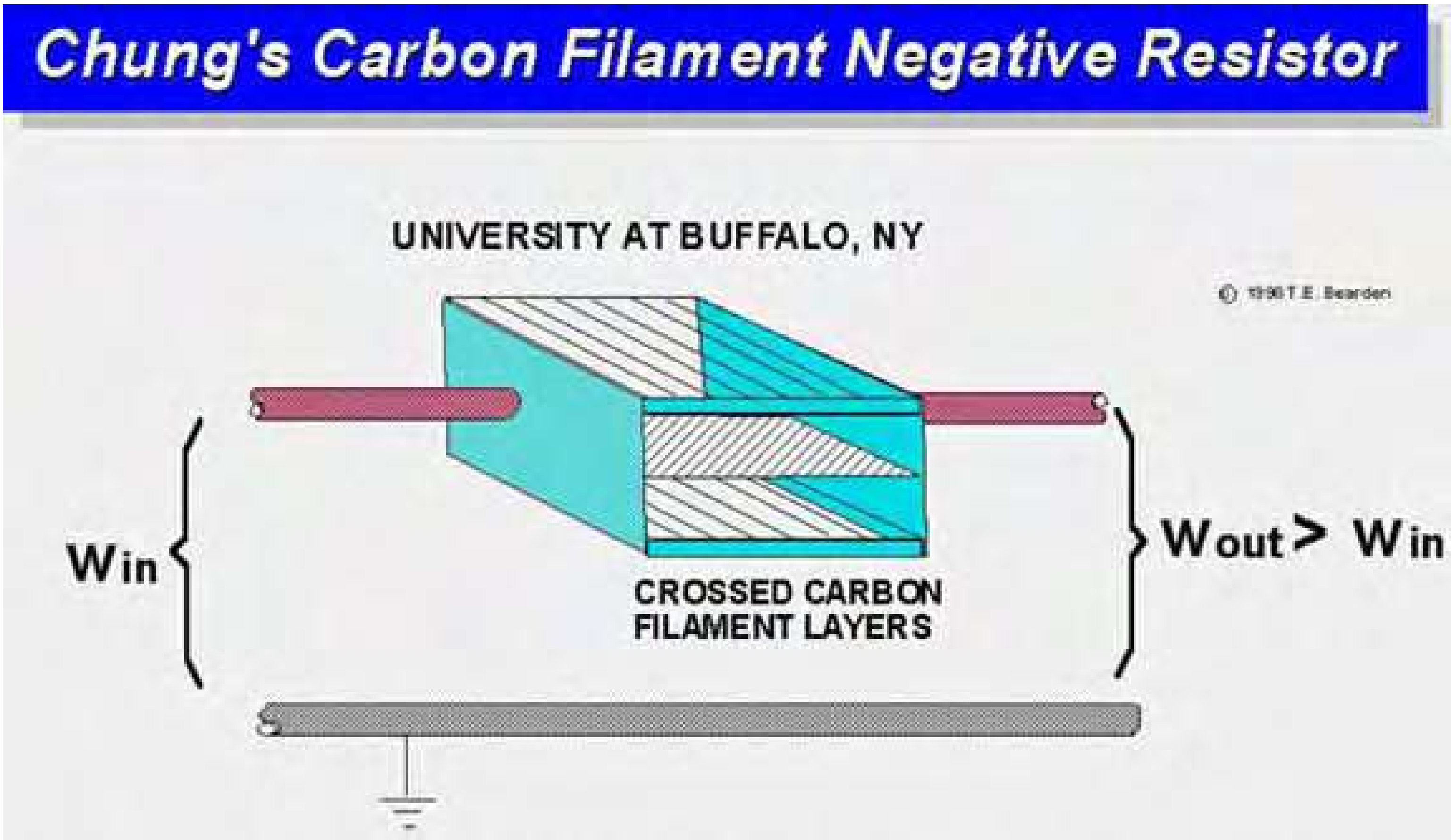

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY
© 1998 T.E. Bearden
Win
CROSSED CARBON FILAMENT LAYERS
Wout > Win

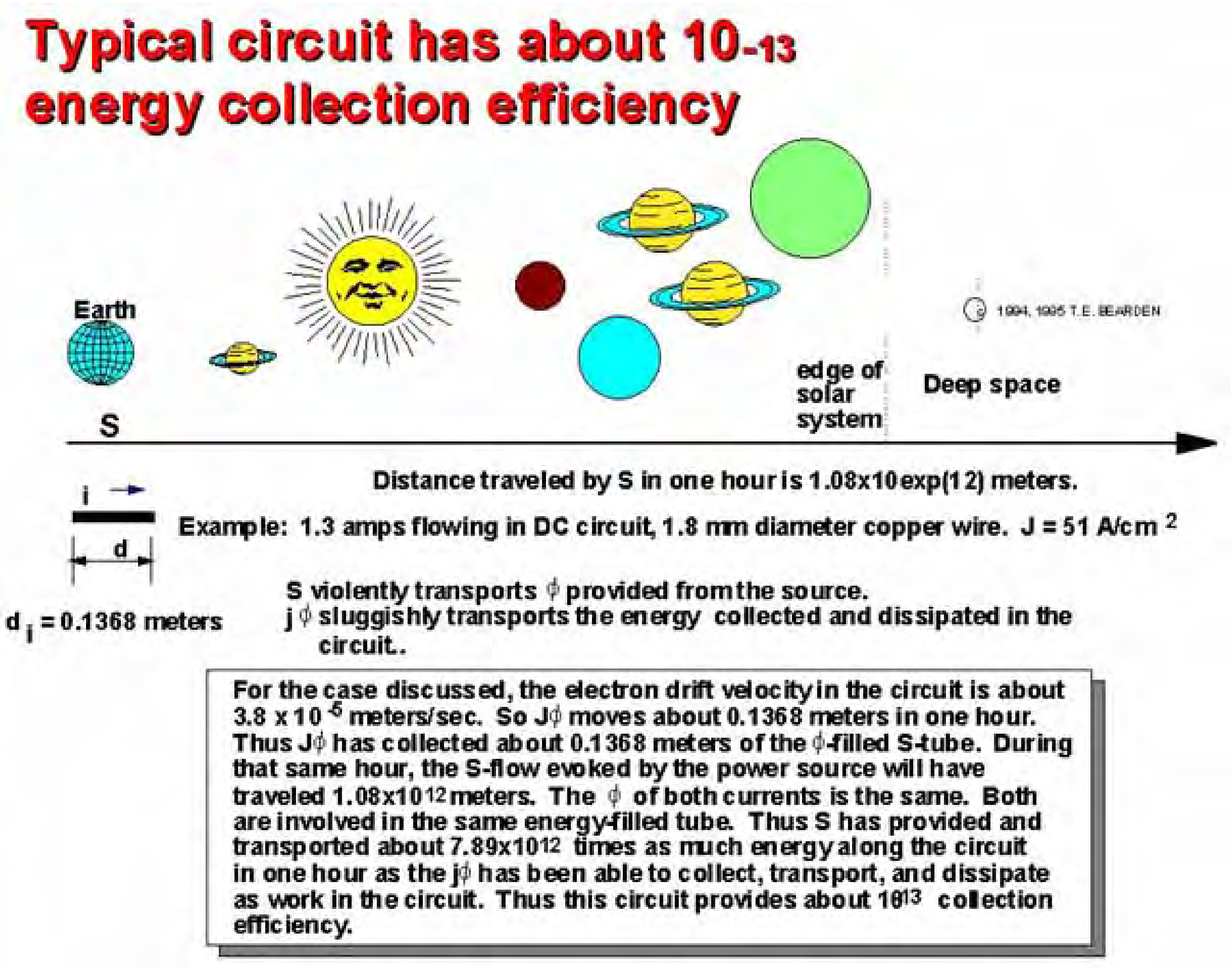

Typical circuit has about 10^{-13} energy collection efficiency

© 1994, 1995 T.E. BEARDEN

Distance traveled by S in one hour is 1.08x10exp(12) meters.

Example: 1.3 amps flowing in DC circuit, 1.8 mm diameter copper wire. $J = 51 \text{ A/cm}^2$

$d_i = 0.1368$ meters

S violently transports ϕ provided from the source.
$j\phi$ sluggishly transports the energy collected and dissipated in the circuit..

For the case discussed, the electron drift velocity in the circuit is about 3.8×10^{-5} meters/sec. So $J\phi$ moves about 0.1368 meters in one hour. Thus $J\phi$ has collected about 0.1368 meters of the ϕ-filled S-tube. During that same hour, the S-flow evoked by the power source will have traveled $1.08\text{x}10^{12}$ meters. The ϕ of both currents is the same. Both are involved in the same energy-filled tube. Thus S has provided and transported about $7.89\text{x}10^{12}$ times as much energy along the circuit in one hour as the $j\phi$ has been able to collect, transport, and dissipate as work in the circuit. Thus this circuit provides about 10^{13} collection efficiency.

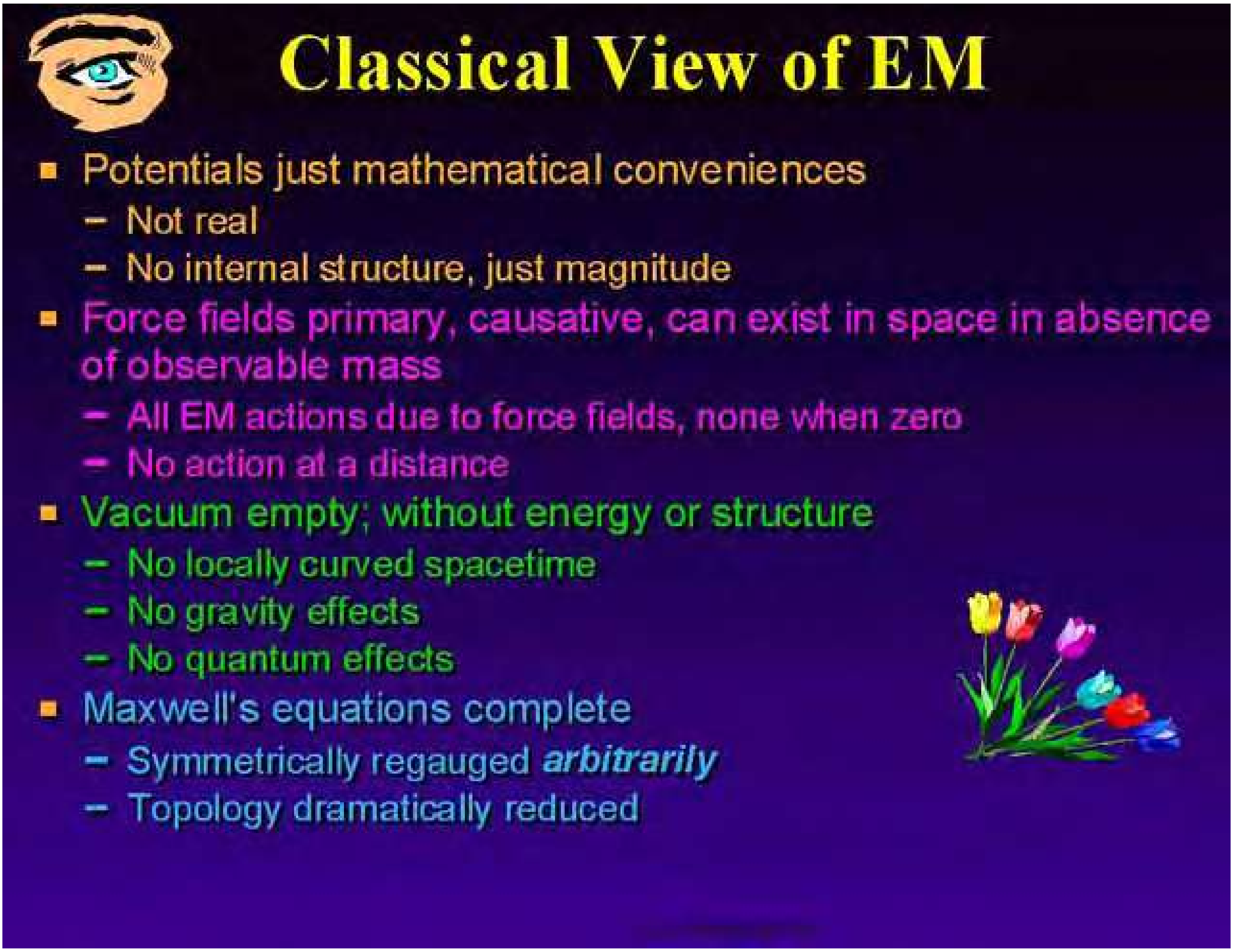

Classical View of EM

- Potentials just mathematical conveniences
  - Not real
  - No internal structure, just magnitude
- Force fields primary, causative, can exist in space in absence of observable mass
  - All EM actions due to force fields, none when zero
  - No action at a distance
- Vacuum empty; without energy or structure
  - No locally curved spacetime
  - No gravity effects
  - No quantum effects
- Maxwell's equations complete
  - Symmetrically regauged ***arbitrarily***
  - Topology dramatically reduced

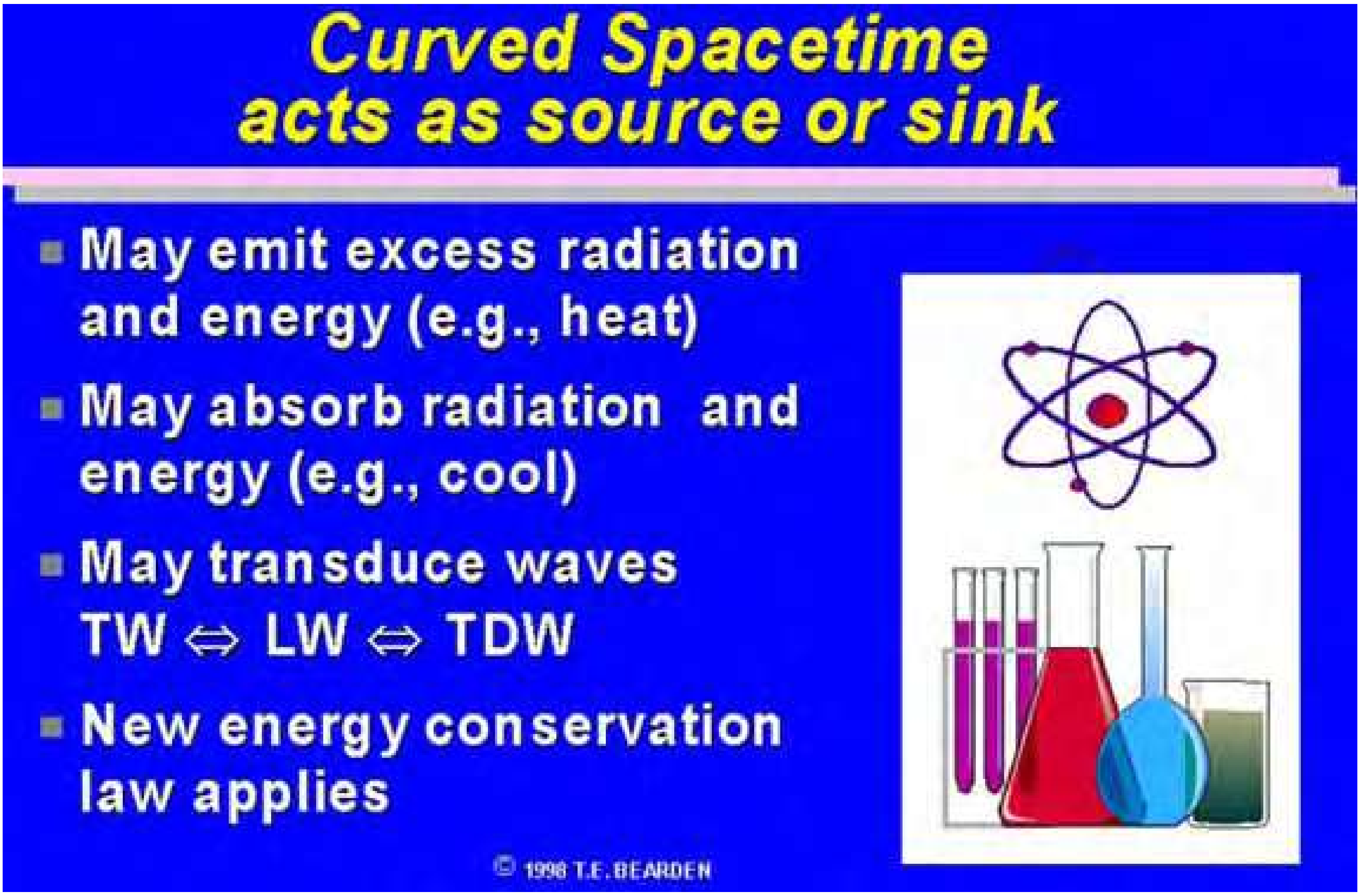

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves
  TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E. BEARDEN

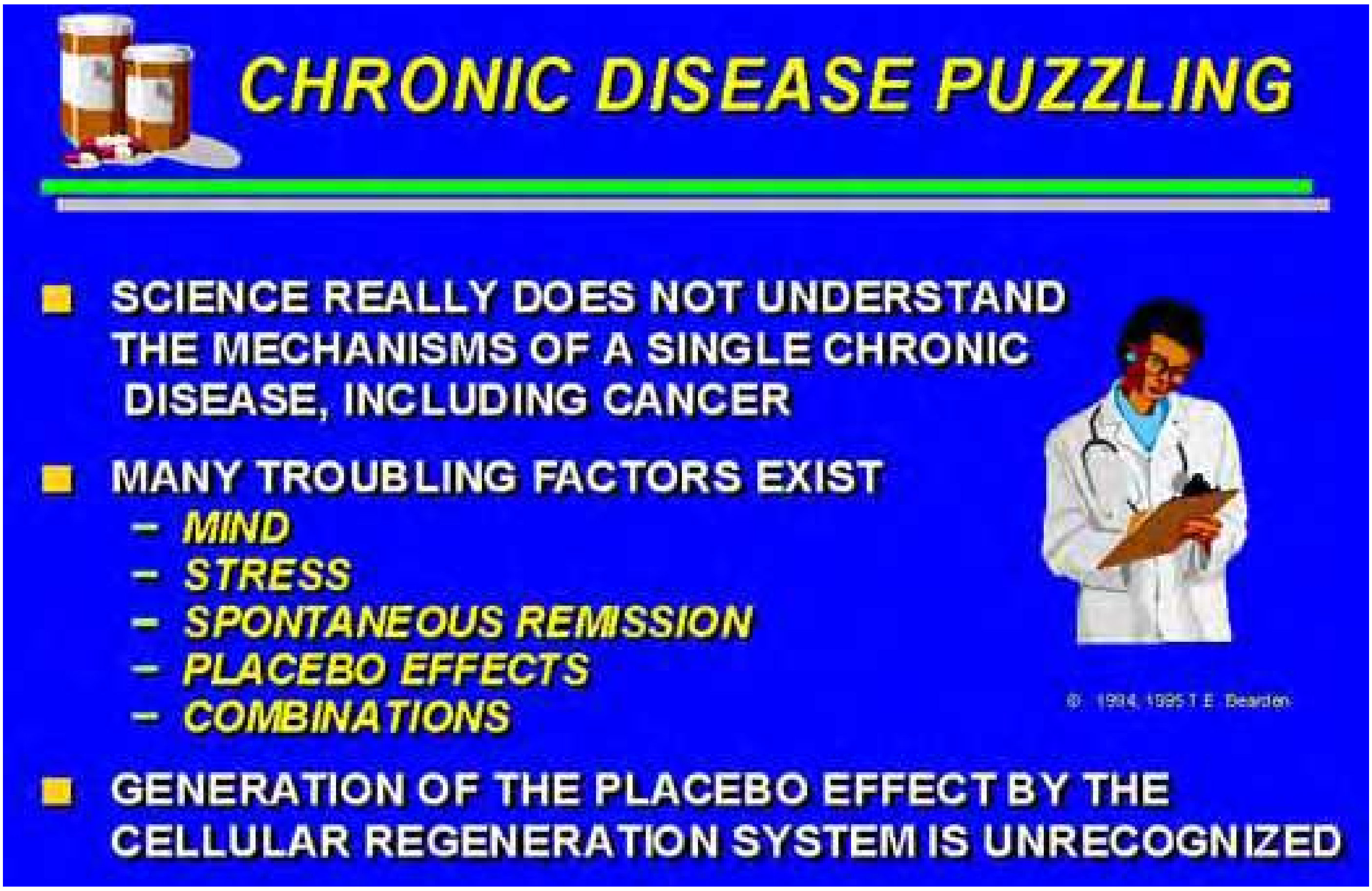

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

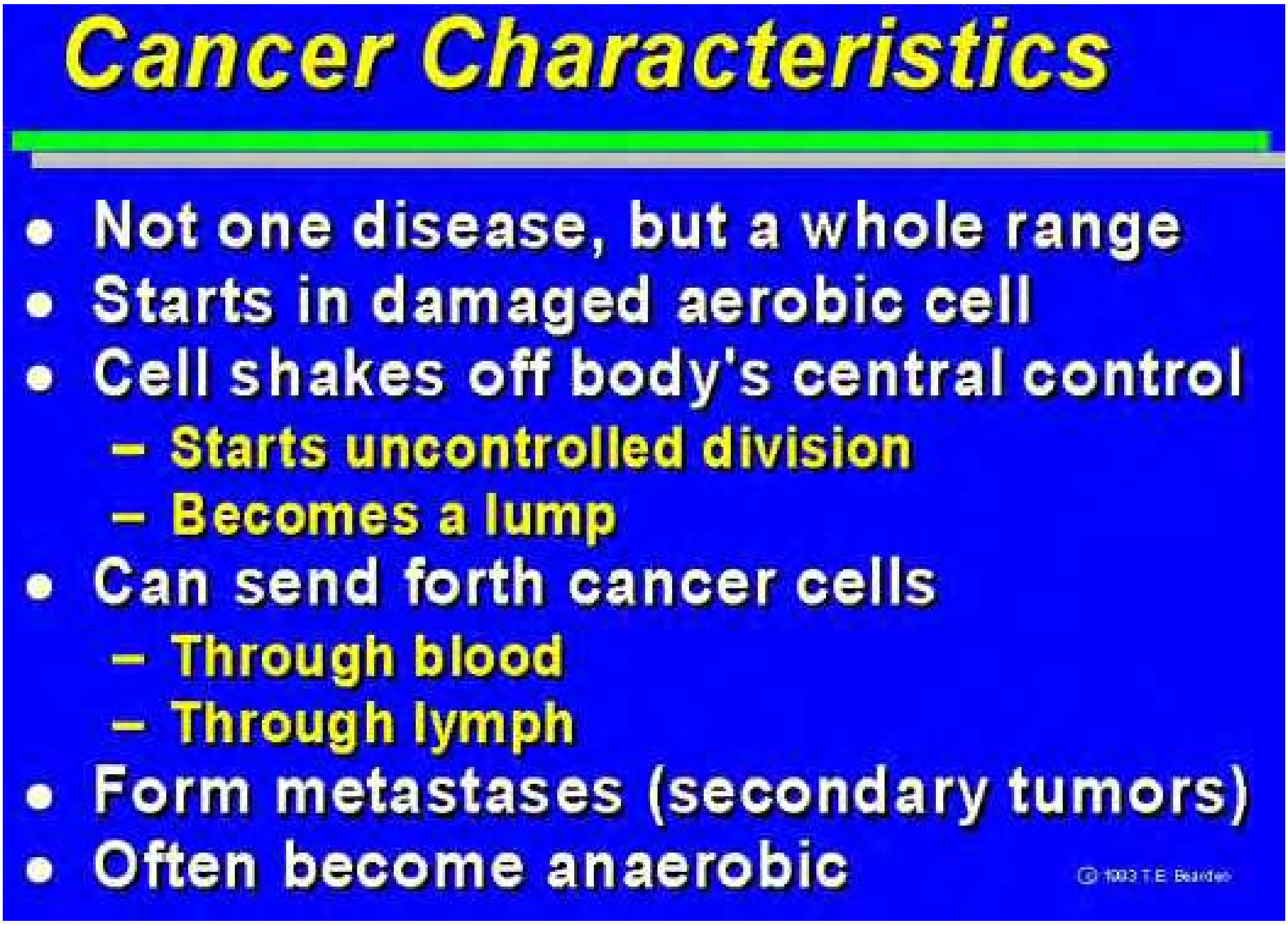

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

© 1993 T.E. Bearden

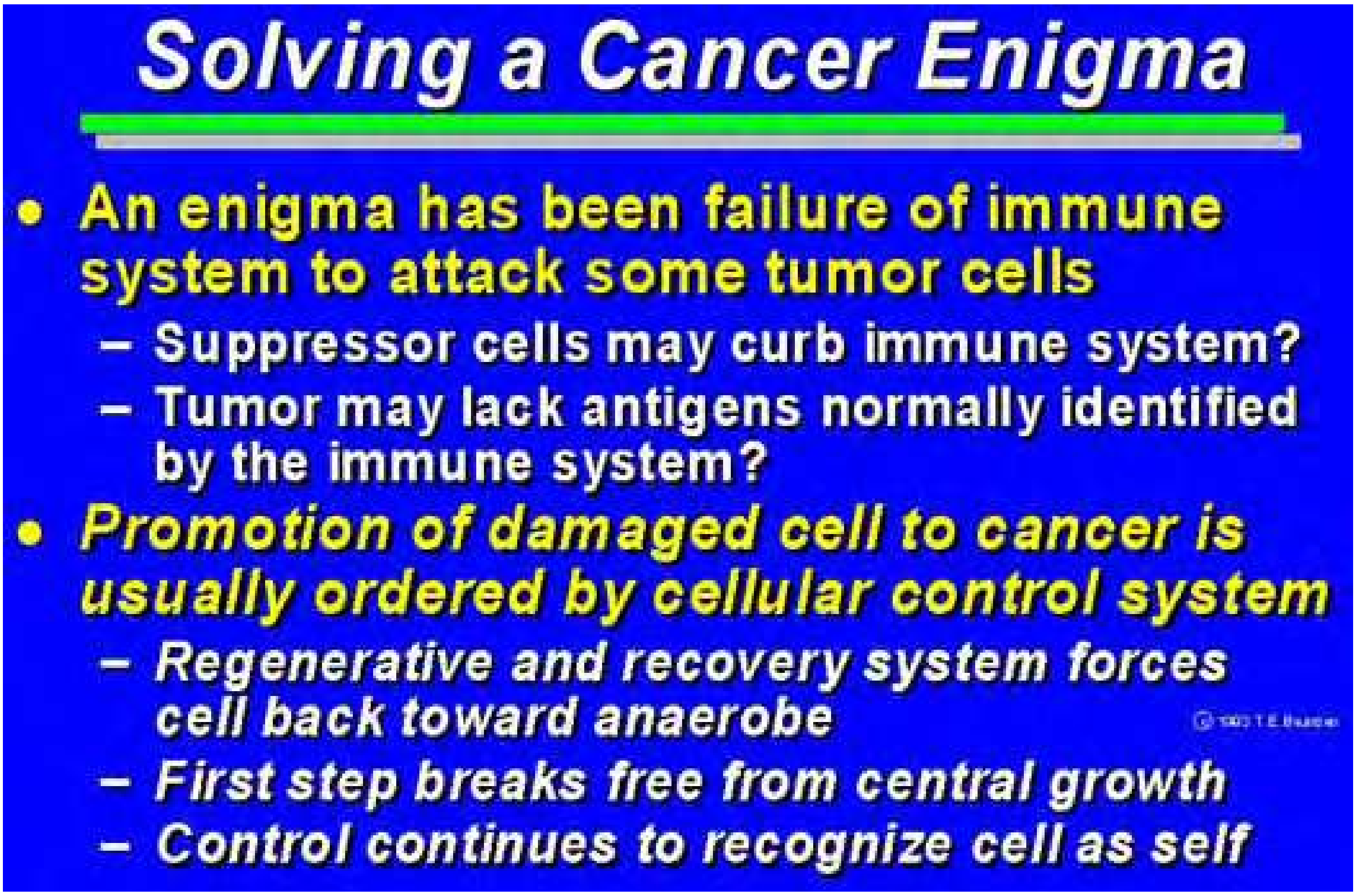

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

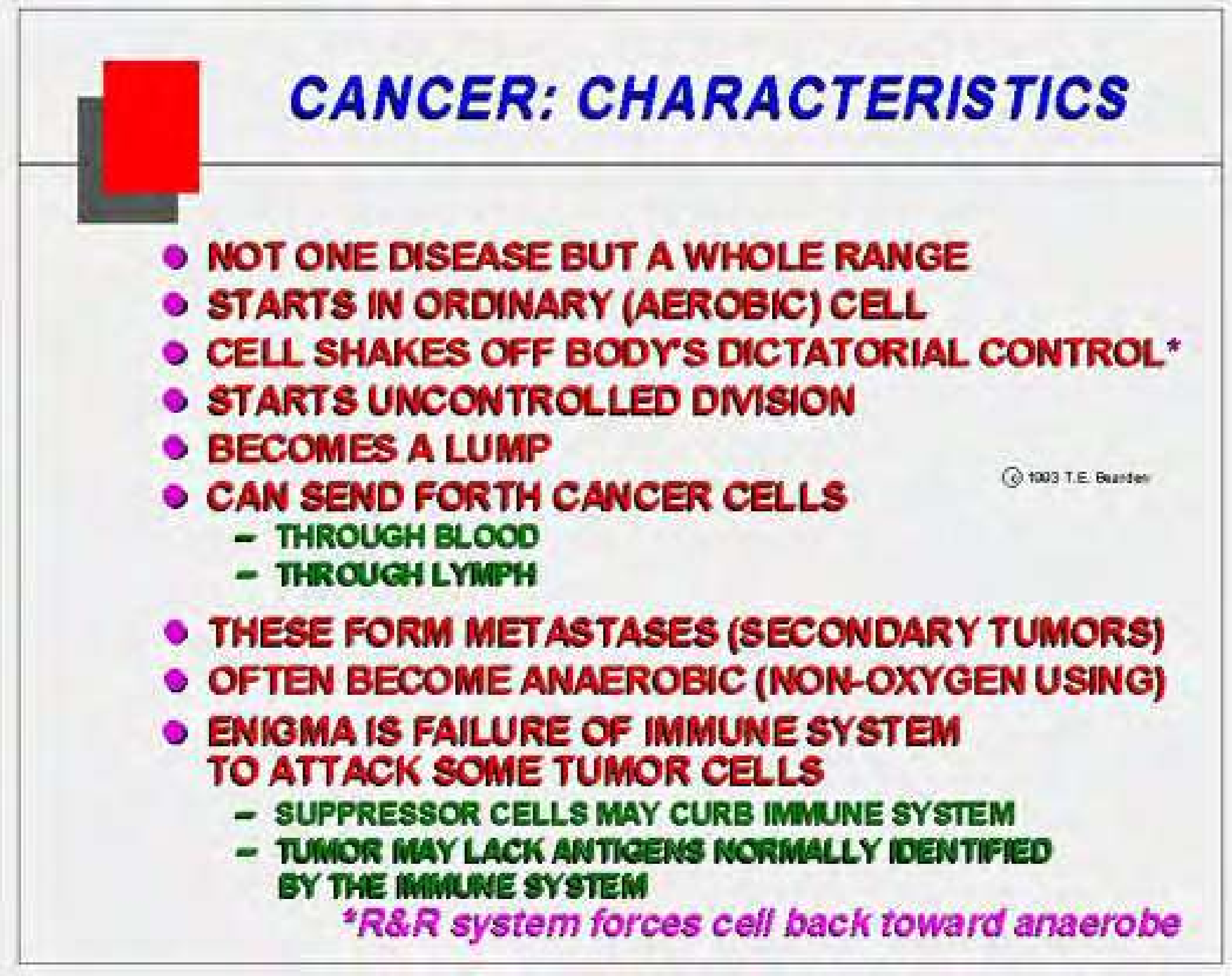

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

© 1993 T.E. Bearden

*R&R system forces cell back toward anaerobe

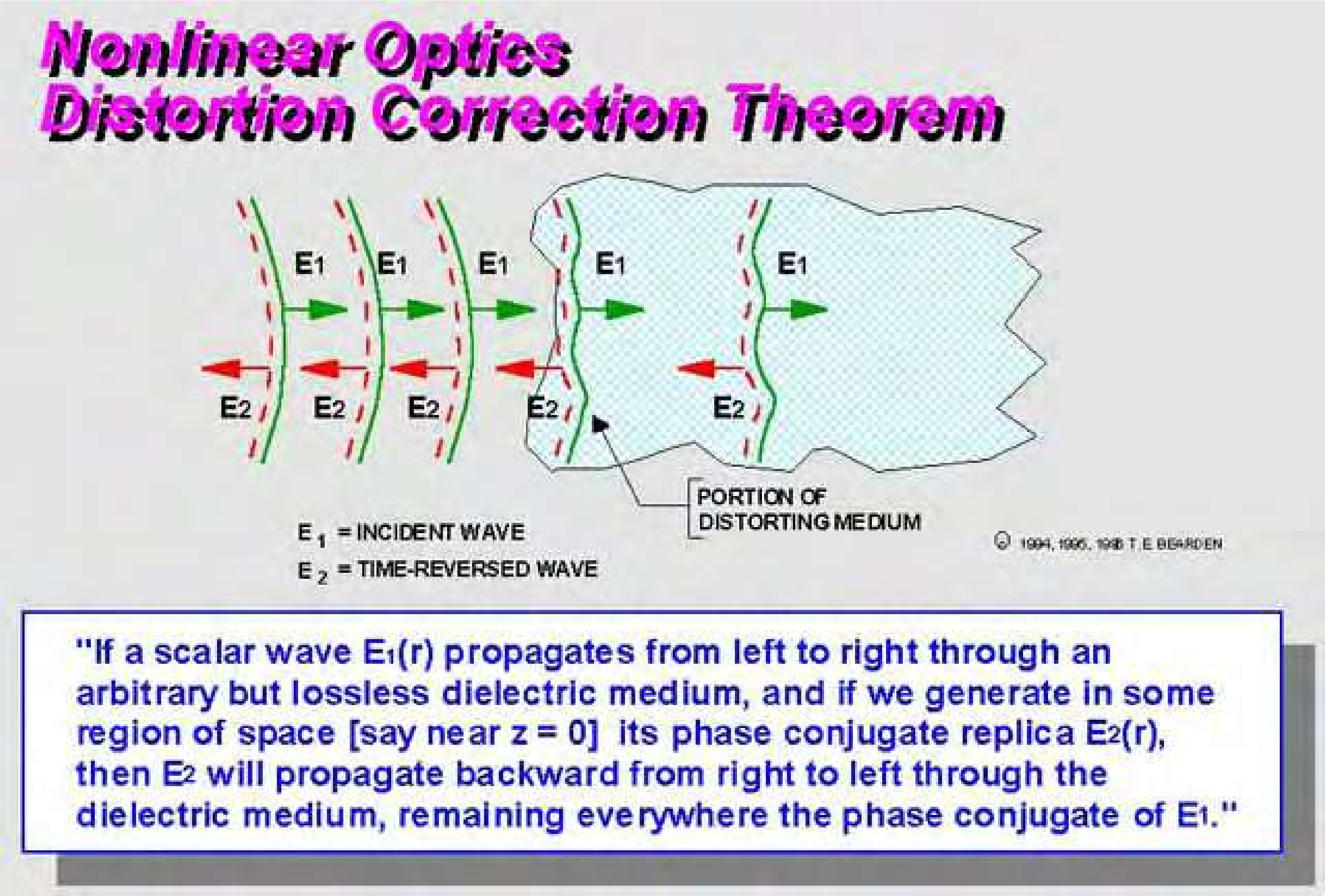

Nonlinear Optics Distortion Correction Theorem

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

© 1994, 1995, 1998 T E BEARDEN

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near $z = 0$] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

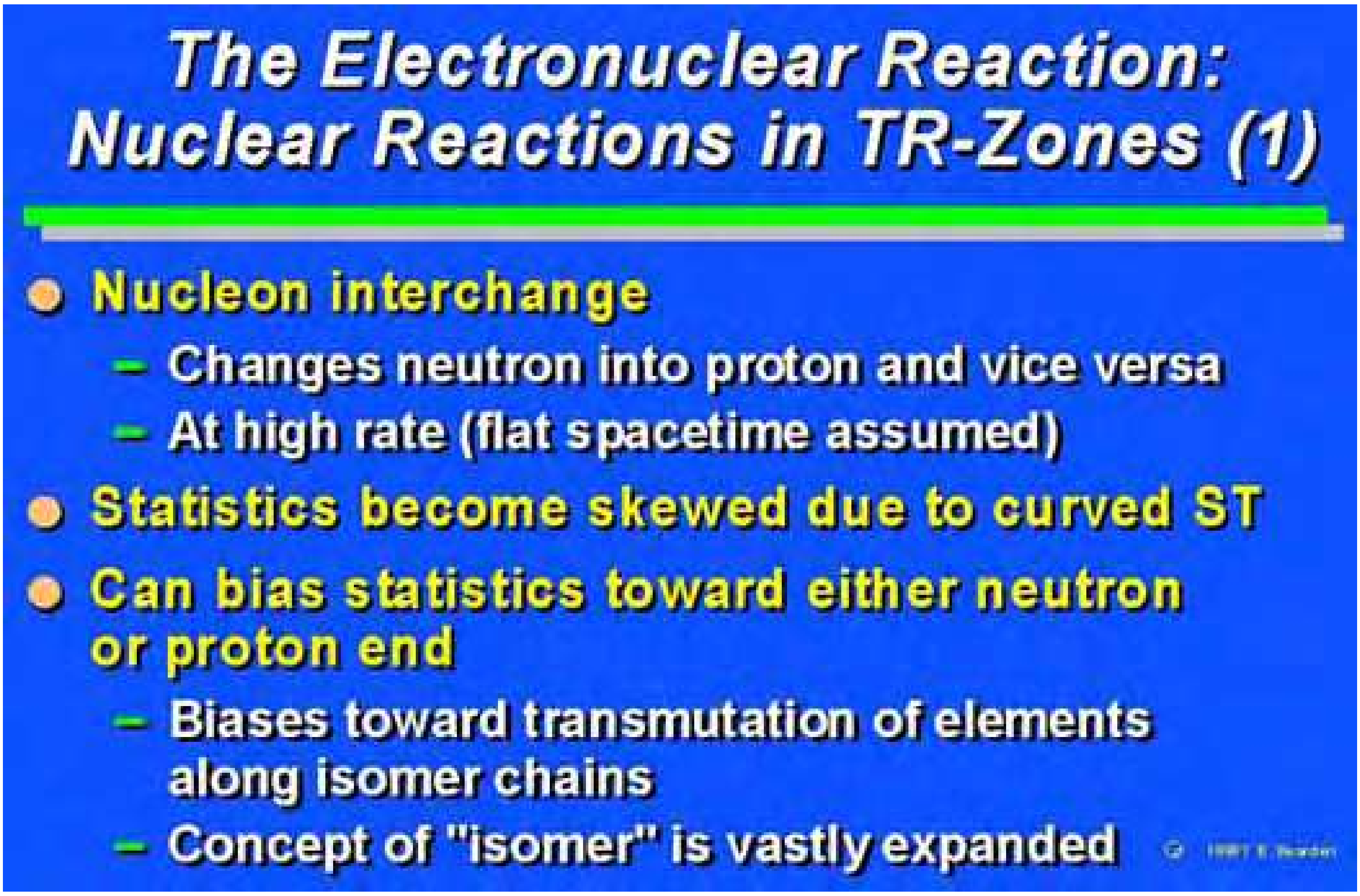

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- Nucleon interchange
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- Statistics become skewed due to curved ST
- Can bias statistics toward either neutron or proton end
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

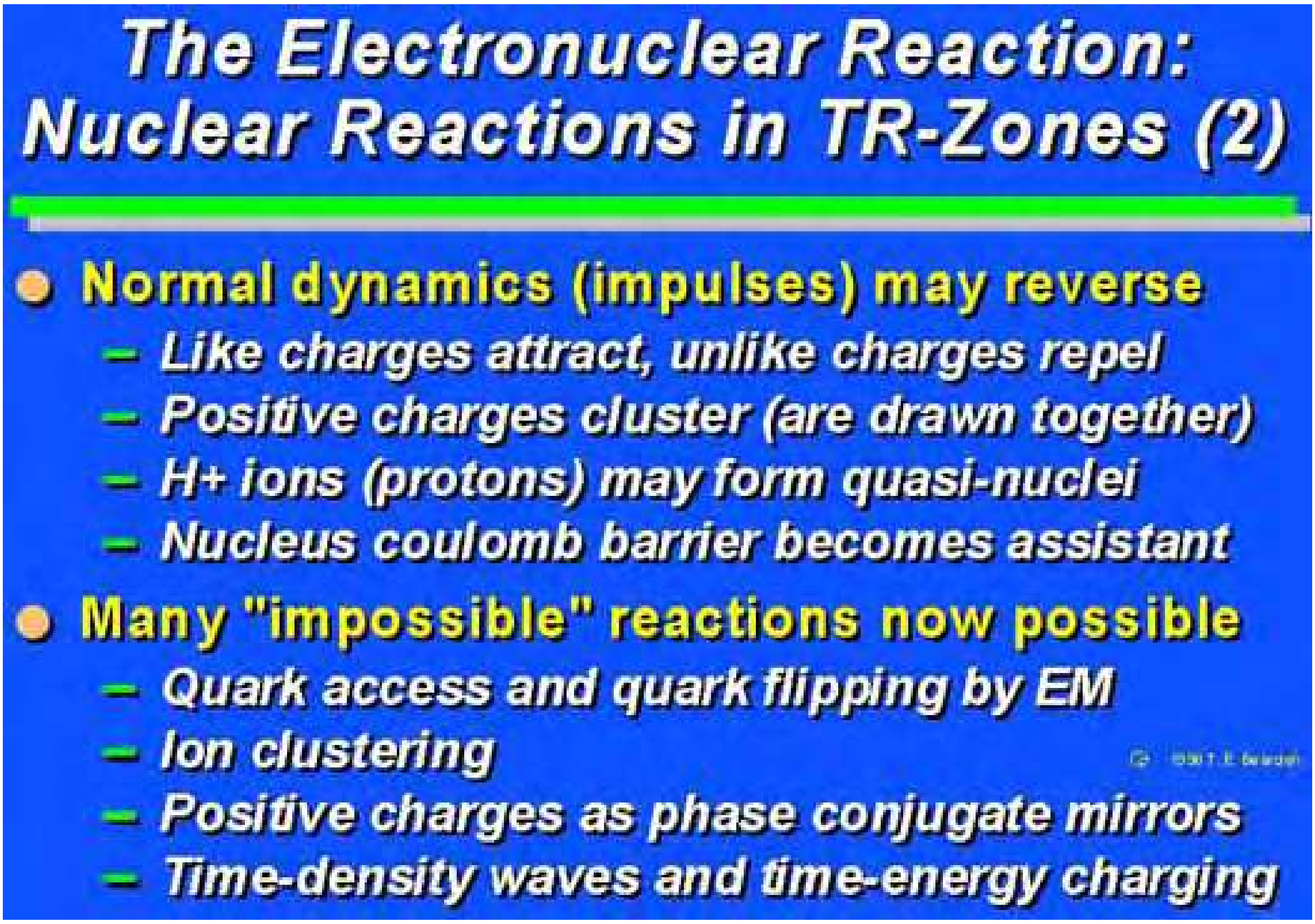

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- **Normal dynamics (impulses) may reverse**
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- **Many "impossible" reactions now possible**
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

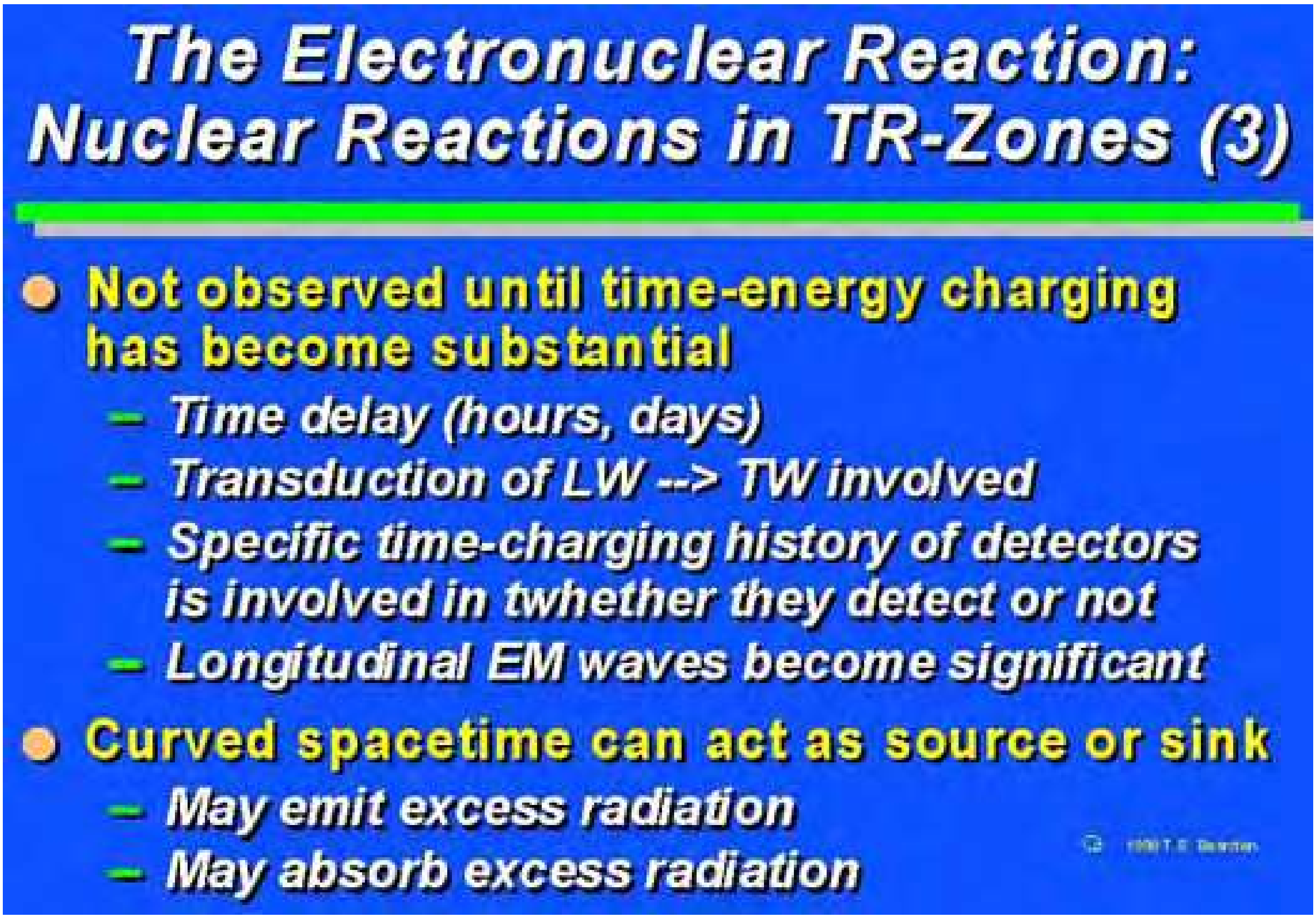

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- **Not observed until time-energy charging has become substantial**
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- **Curved spacetime can act as source or sink**
  - *May emit excess radiation*
  - *May absorb excess radiation*

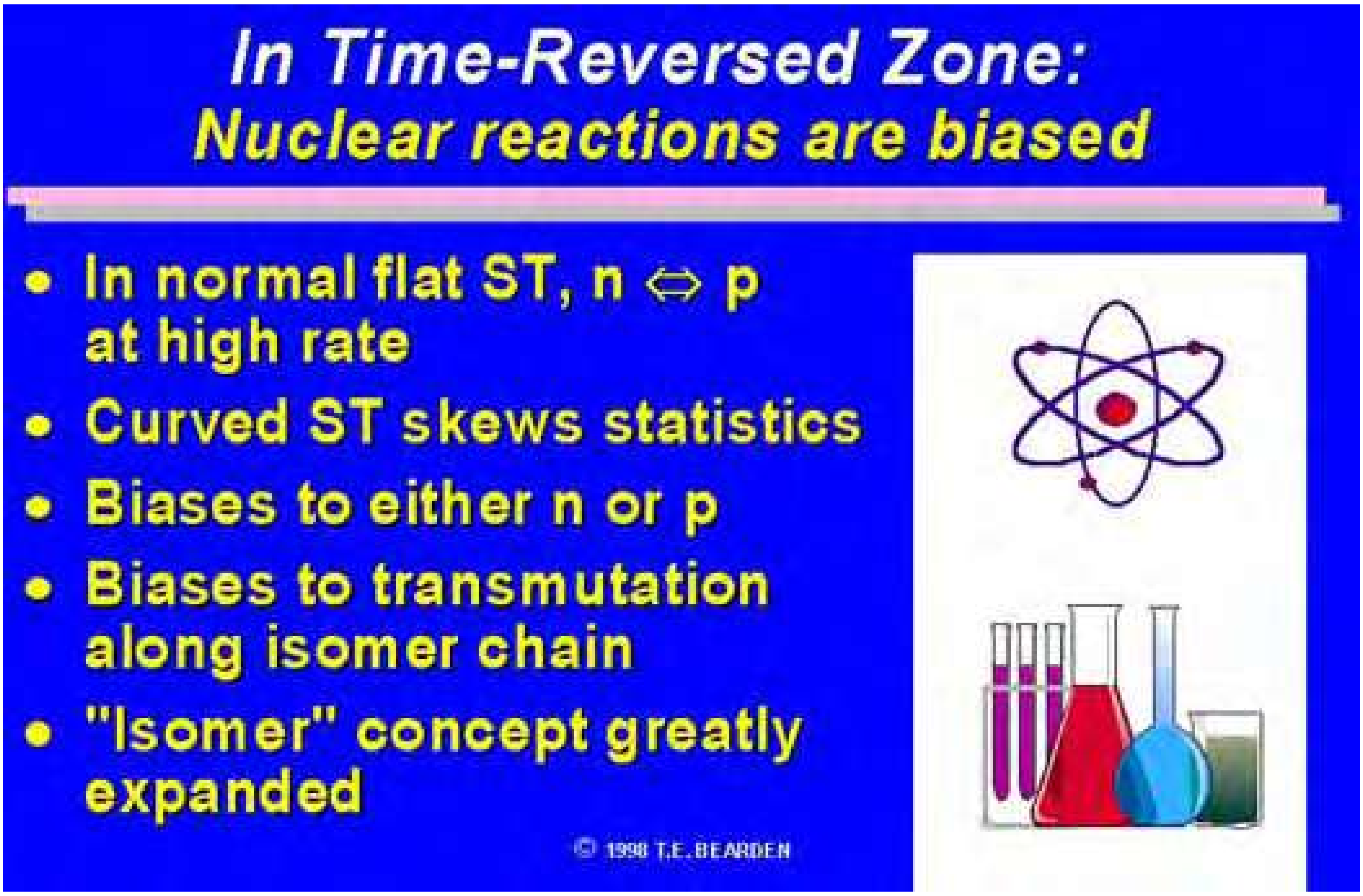

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T.E.BEARDEN

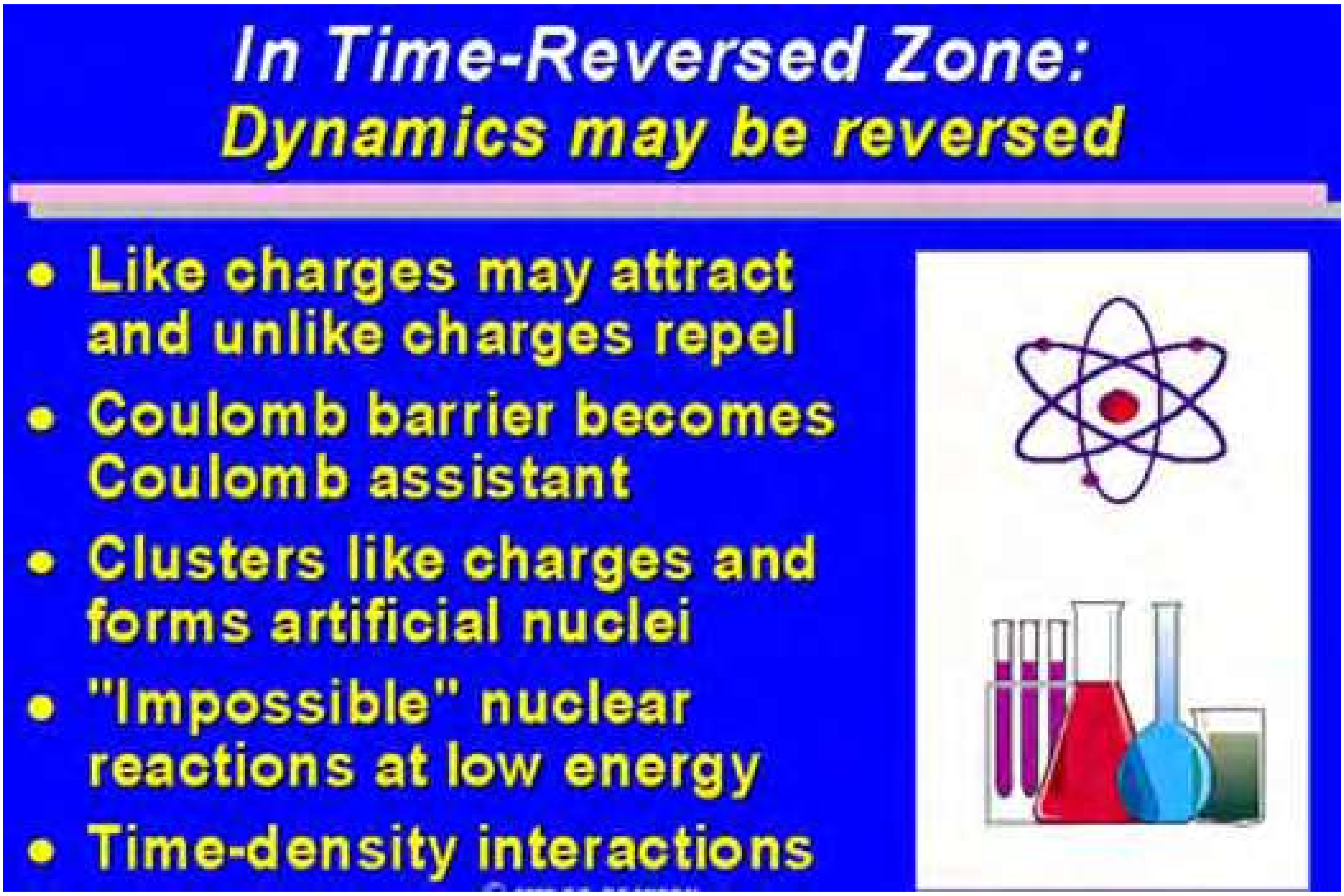

In Time-Reversed Zone:
Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

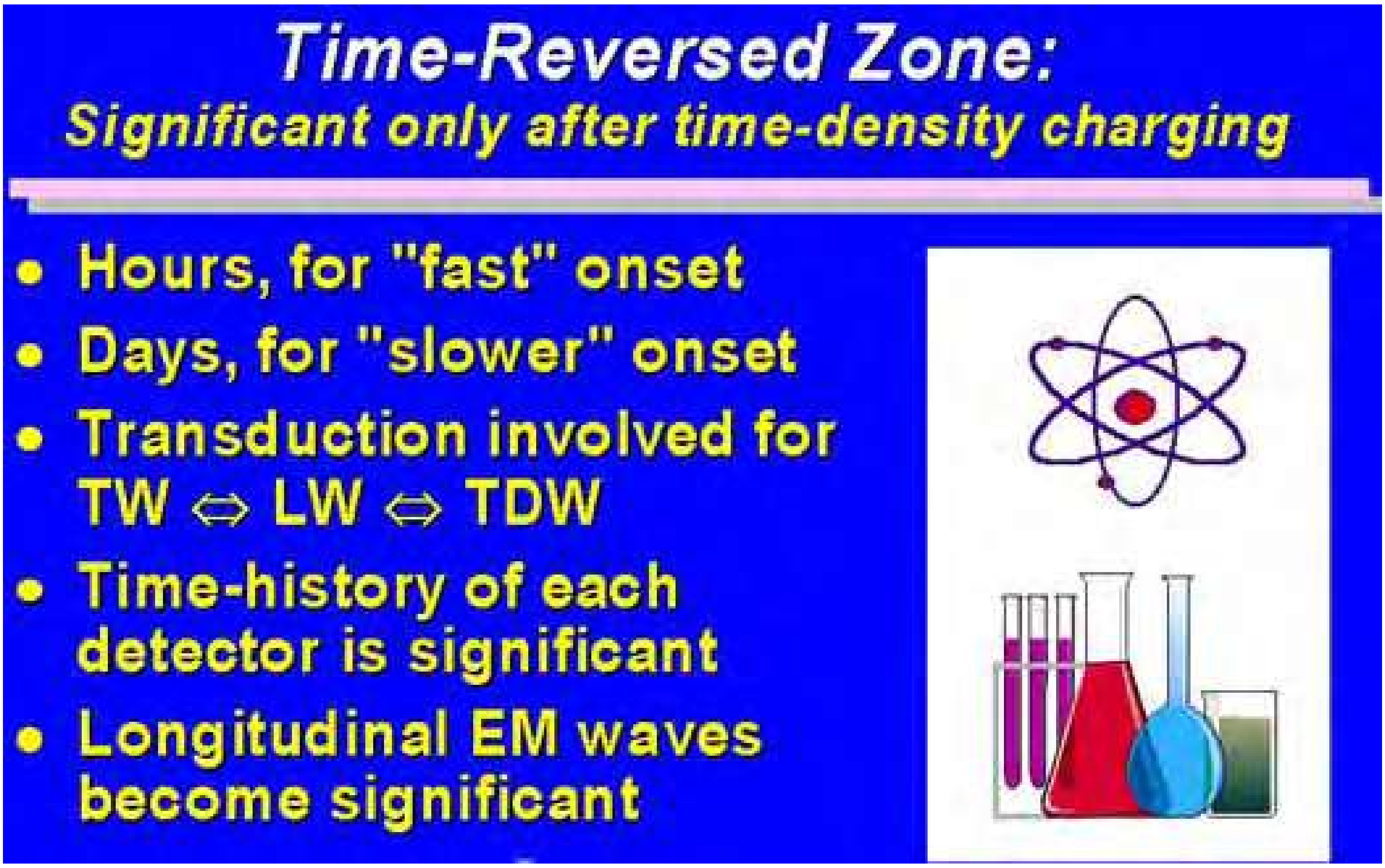

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

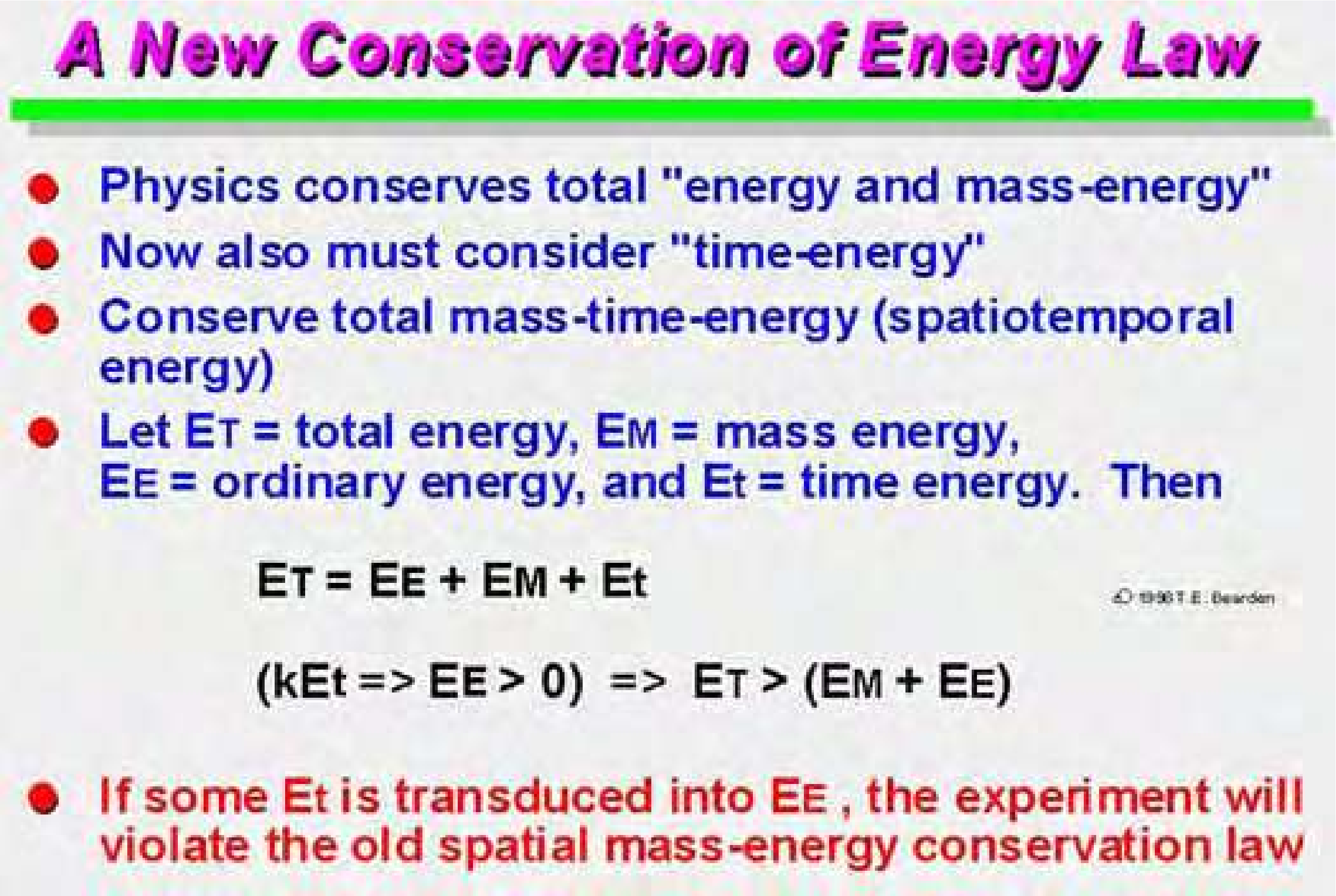

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

© 1998 T. E. Bearden

$$(kE_t \Rightarrow E_E > 0) \Rightarrow E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

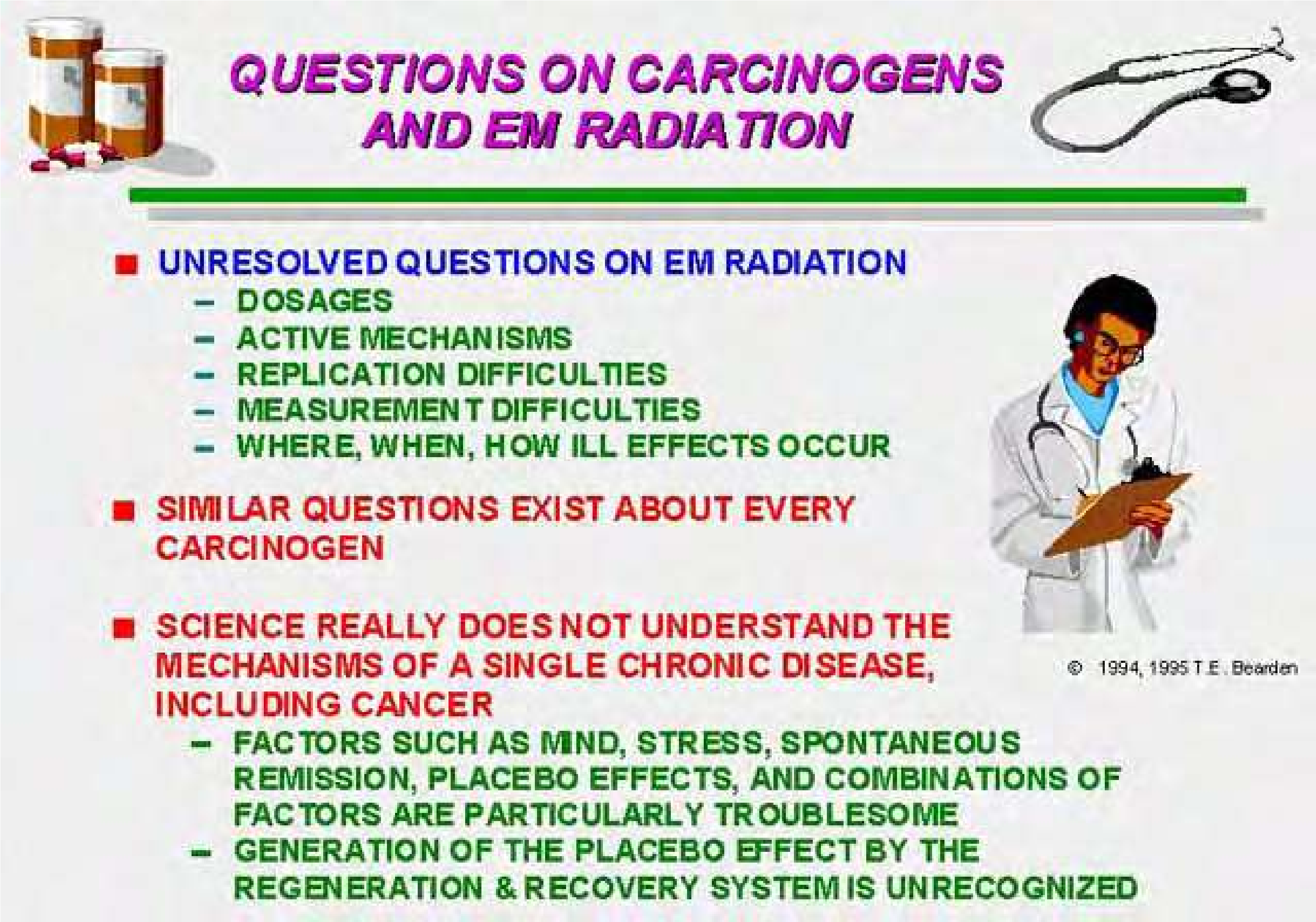

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

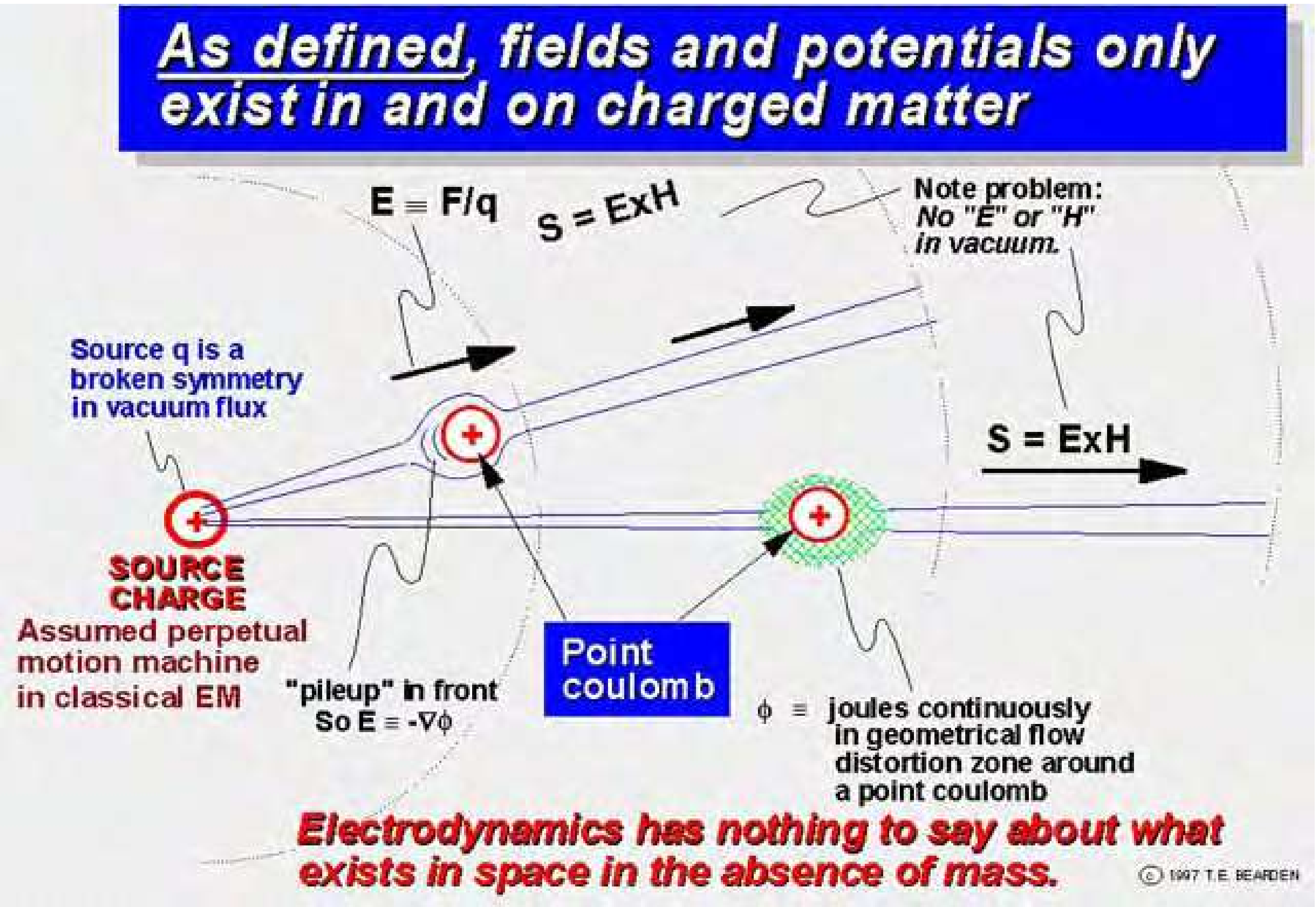

As defined, fields and potentials only exist in and on charged matter

E ≡ F/q

S = ExH

Note problem:
No "E" or "H" in vacuum.

Source q is a
broken symmetry
in vacuum flux

S = ExH

SOURCE CHARGE

Assumed perpetual
motion machine
in classical EM

"pileup" in front
So E ≡ -∇φ

Point
coulomb

φ ≡ joules continuously
in geometrical flow
distortion zone around
a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

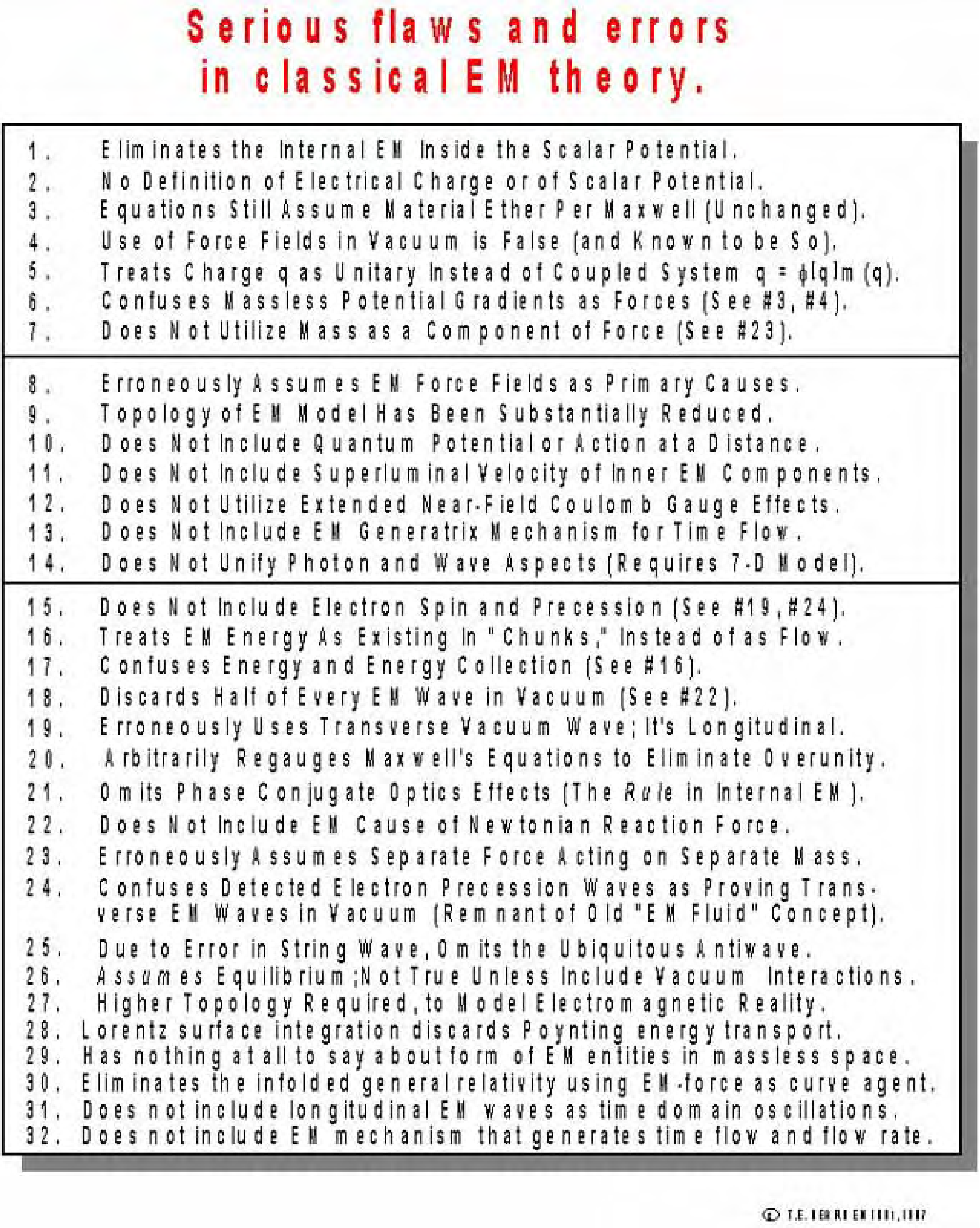

Serious flaws and errors in classical EM theory.

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi[q]m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

© T.E. BEARDEN 1991, 1997

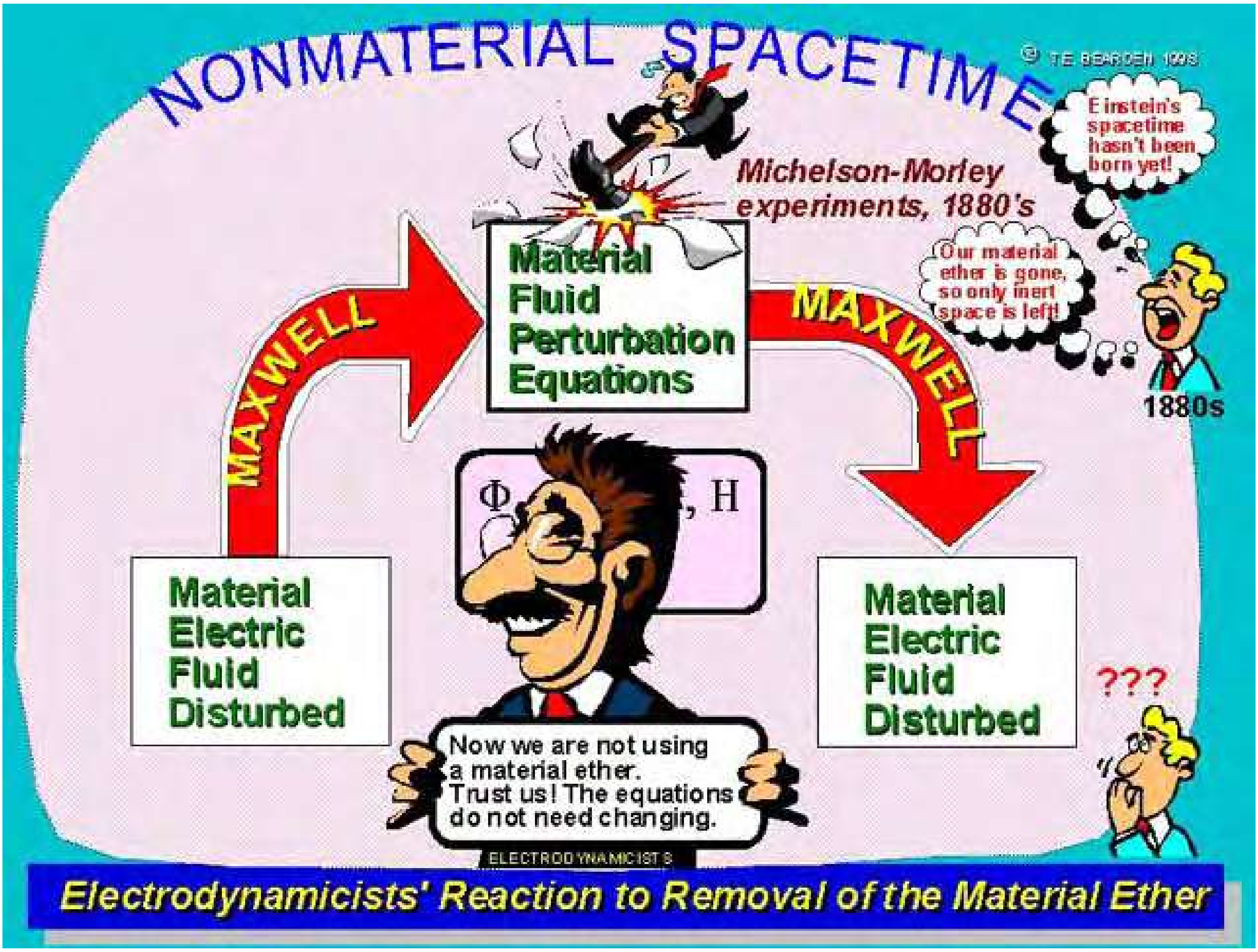

NONMATERIAL SPACETIME

© T.E. BEARDEN 1998

Michelson-Morley experiments, 1880's

Material Fluid Perturbation Equations

Material Electric Fluid Disturbed

Material Electric Fluid Disturbed

MAXWELL
MAXWELL
Einstein's spacetime hasn't been born yet!
Our material ether is gone, so only inert space is left!
1880s
Φ
, H
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

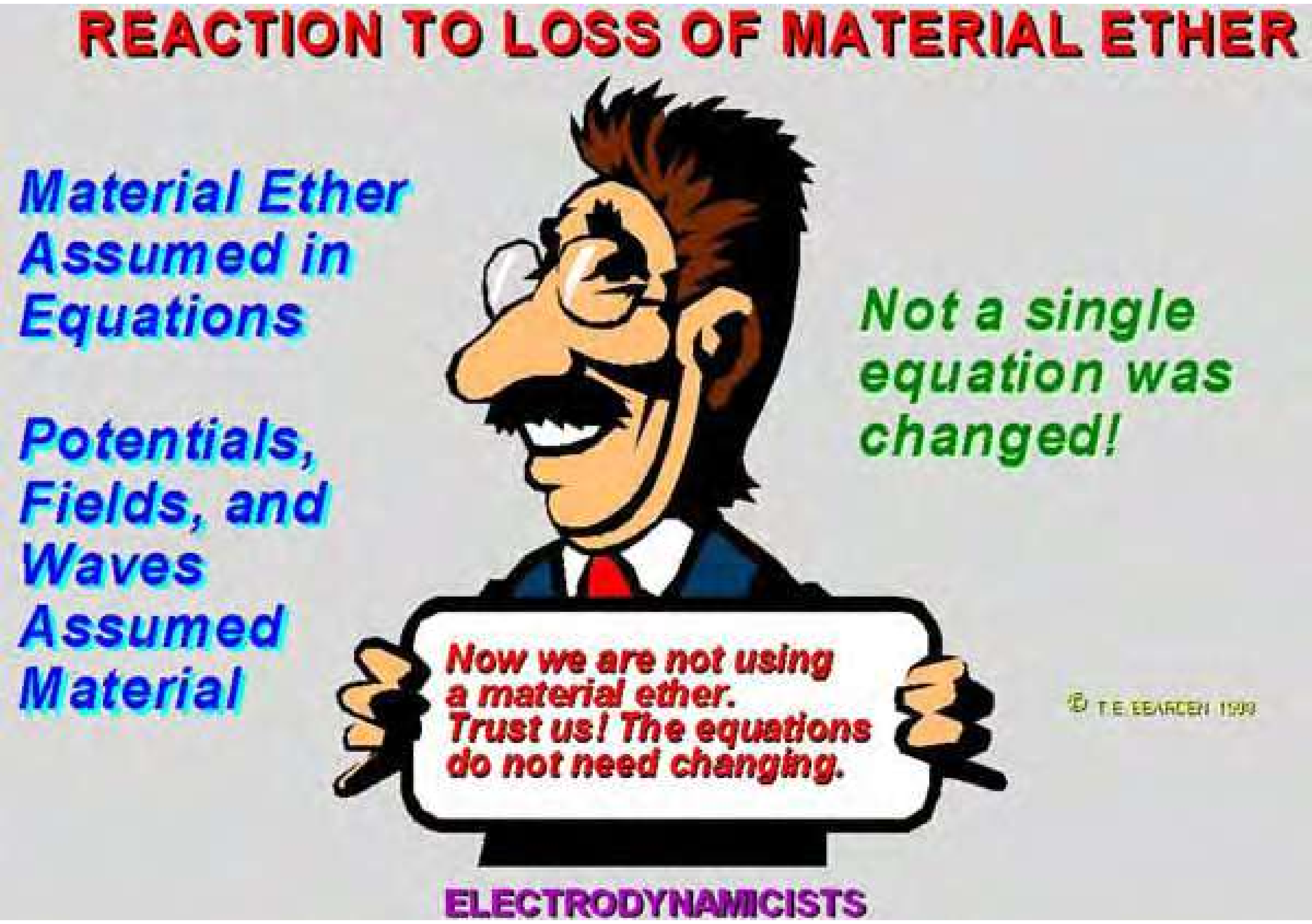

REACTION TO LOSS OF MATERIAL ETHER

Material Ether
Assumed in
Equations

Potentials,
Fields, and
Waves
Assumed
Material

Not a single
equation was
changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.
ELECTRODYNAMICISTS

© T.E. BEARDEN 1999

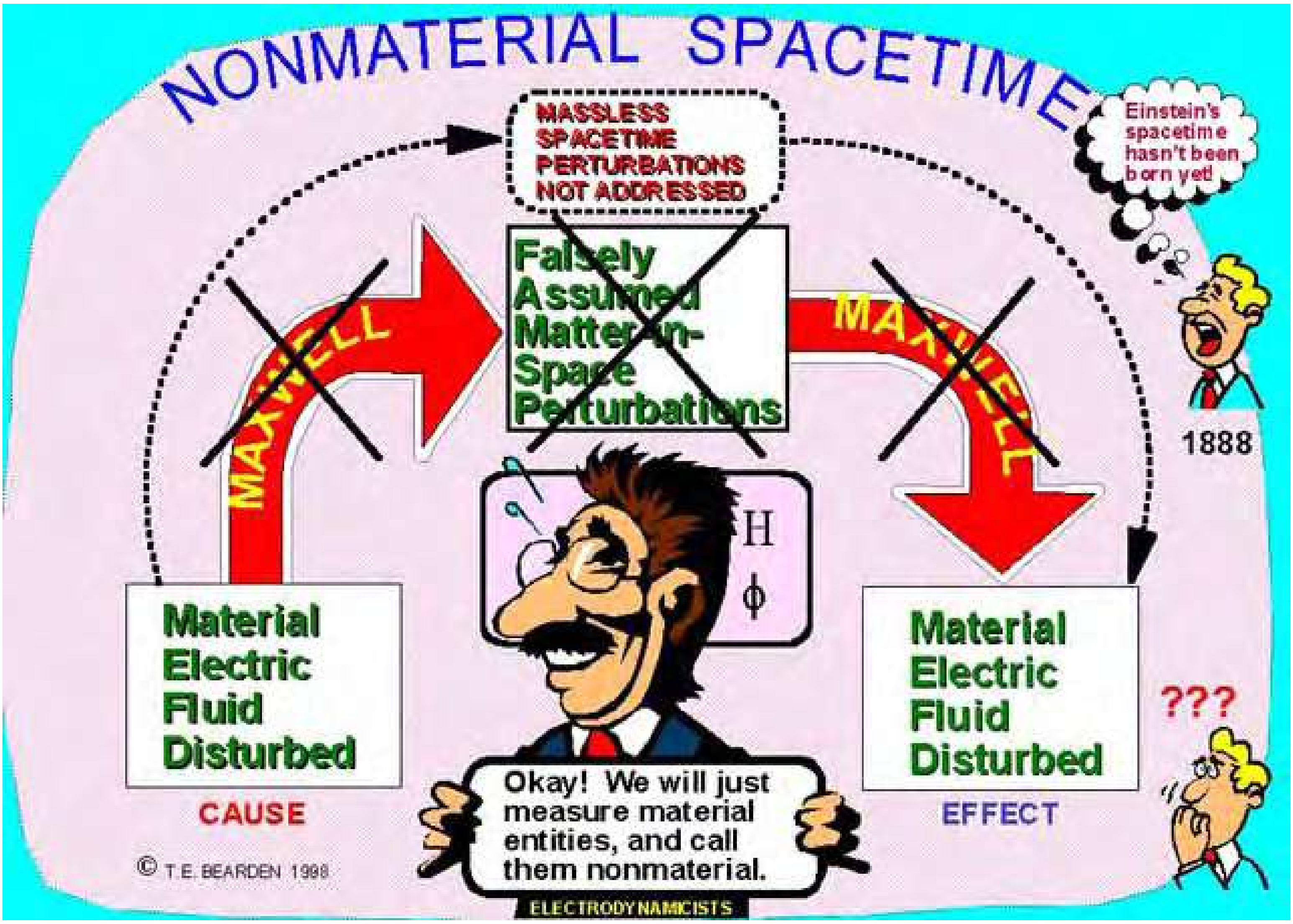
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
MAXWELL
~~Falsely Assumed Matter-in-Space Perturbations~~
MAXWELL
1888
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E. BEARDEN 1998

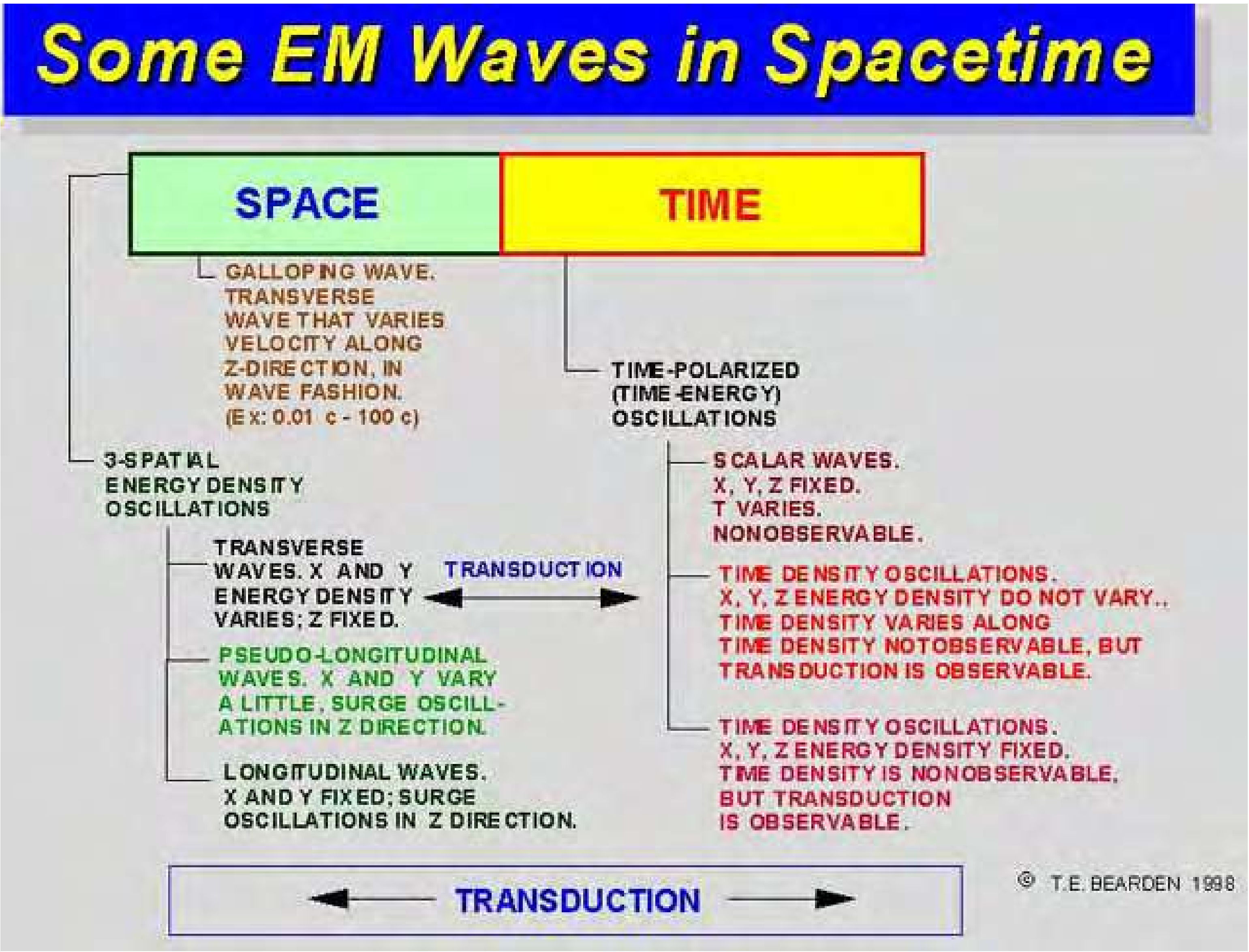

Some EM Waves in Spacetime

SPACE

TIME

GALLOPING WAVE.
TRANSVERSE
WAVE THAT VARIES
VELOCITY ALONG
Z-DIRECTION, IN
WAVE FASHION.
(Ex: 0.01 c - 100 c)

3-SPATIAL
ENERGY DENSITY
OSCILLATIONS

TRANSVERSE
WAVES. X AND Y
ENERGY DENSITY
VARIES; Z FIXED.

TRANSDUCTION

PSEUDO-LONGITUDINAL
WAVES. X AND Y VARY
A LITTLE, SURGE OSCILL-
ATIONS IN Z DIRECTION.

LONGITUDINAL WAVES.
X AND Y FIXED; SURGE
OSCILLATIONS IN Z DIRECTION.

TIME-POLARIZED
(TIME-ENERGY)
OSCILLATIONS

SCALAR WAVES.
X, Y, Z FIXED.
T VARIES.
NONOBSERVABLE.

TIME DENSITY OSCILLATIONS.
X, Y, Z ENERGY DENSITY DO NOT VARY..
TIME DENSITY VARIES ALONG
TIME DENSITY NOTOBSERVABLE, BUT
TRANSDUCTION IS OBSERVABLE.

TIME DENSITY OSCILLATIONS.
X, Y, Z ENERGY DENSITY FIXED.
TIME DENSITY IS NONOBSERVABLE,
BUT TRANSDUCTION
IS OBSERVABLE.

TRANSDUCTION

© T.E. BEARDEN 1998

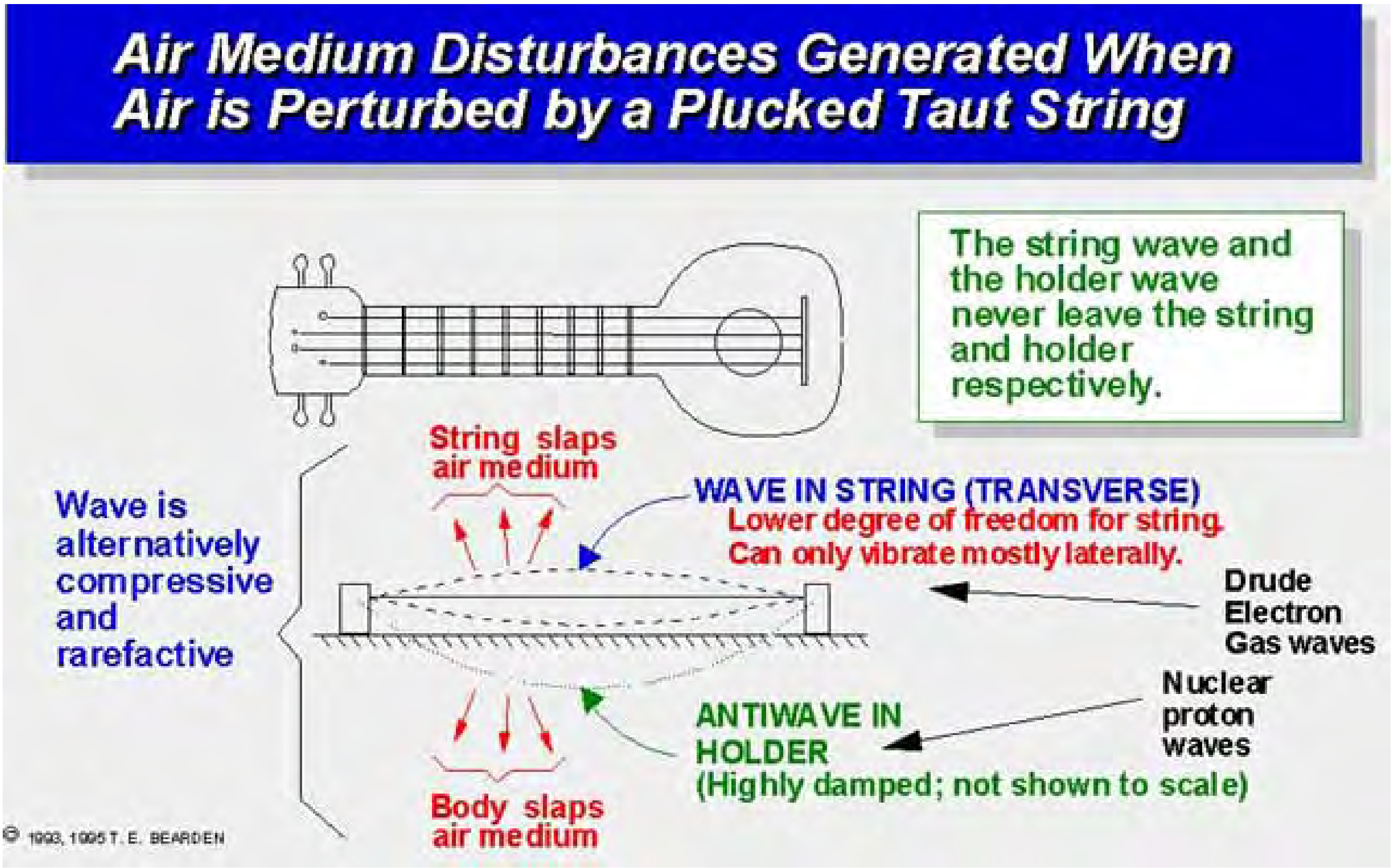

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

String slaps air medium

WAVE IN STRING (TRANSVERSE)

Lower degree of freedom for string. Can only vibrate mostly laterally.

Drude Electron Gas waves

Wave is alternatively compressive and rarefactive

Nuclear proton waves

ANTIWAVE IN HOLDER (Highly damped; not shown to scale)

Body slaps air medium

© 1993, 1995 T. E. BEARDEN

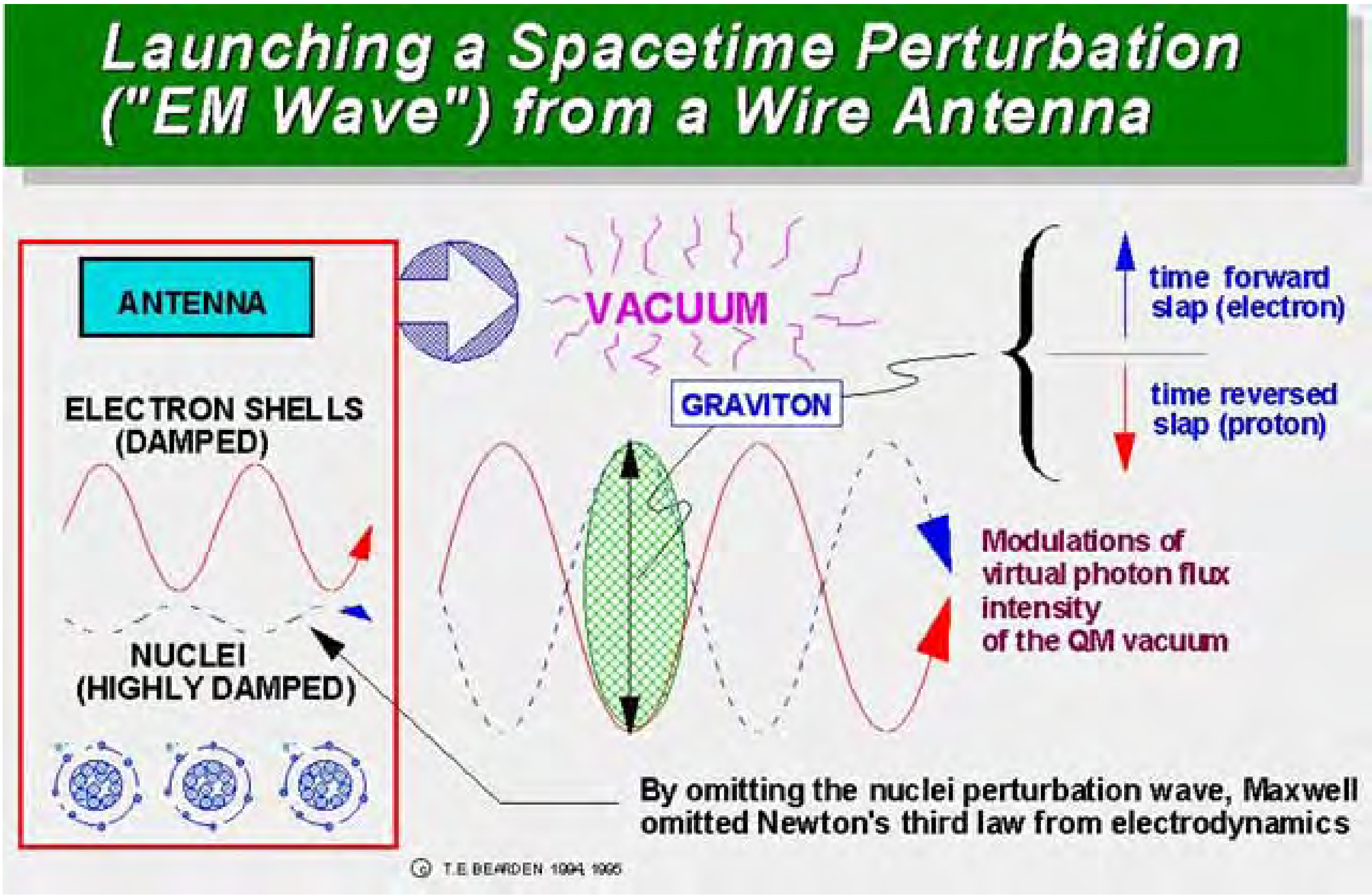

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

ANTENNA
ELECTRON SHELLS (DAMPED)
NUCLEI (HIGHLY DAMPED)
VACUUM
GRAVITON
time forward slap (electron)
time reversed slap (proton)
Modulations of virtual photon flux intensity of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

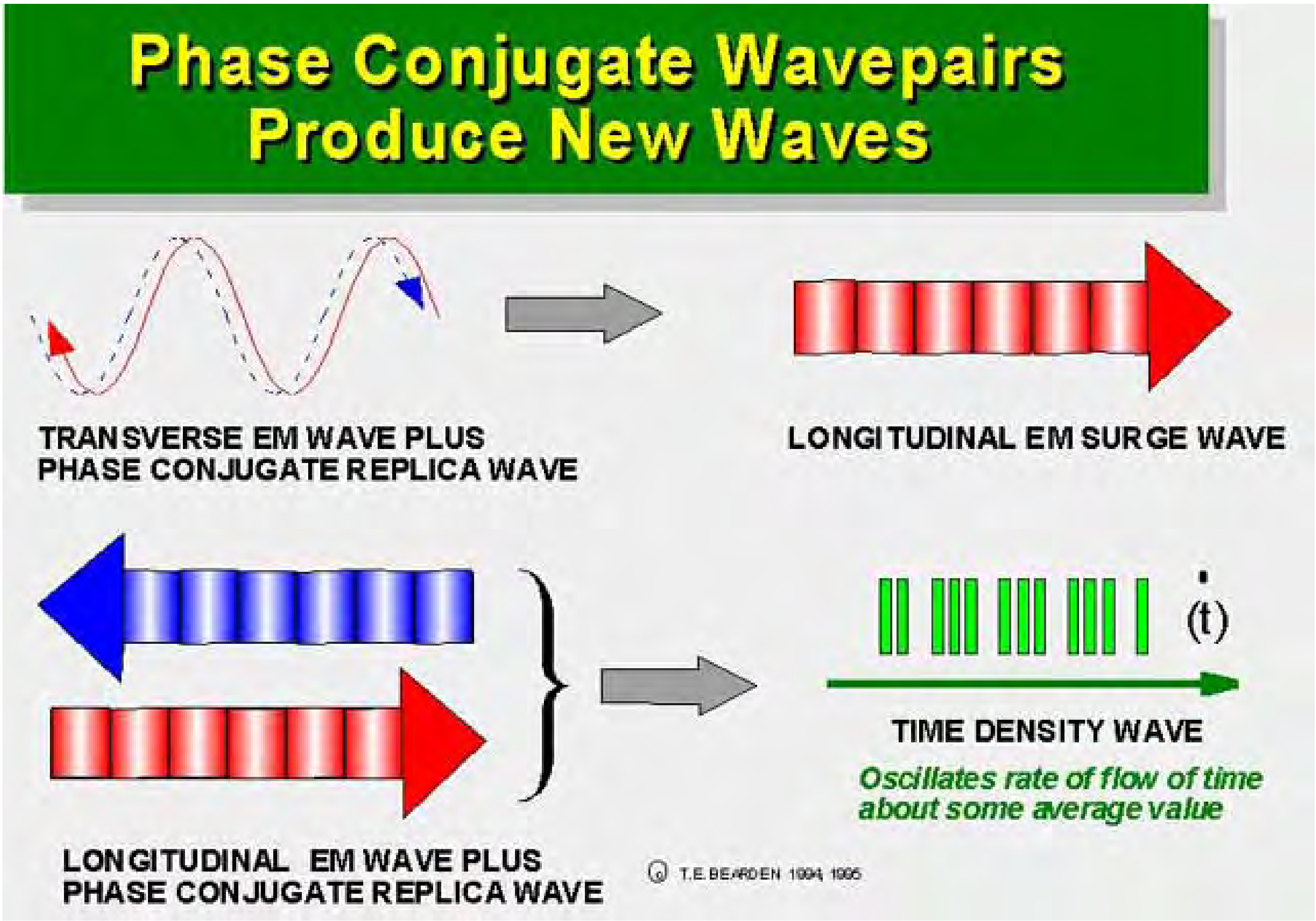

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM SURGE WAVE

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

TIME DENSITY WAVE

Oscillates rate of flow of time about some average value

© T.E. BEARDEN 1994, 1995

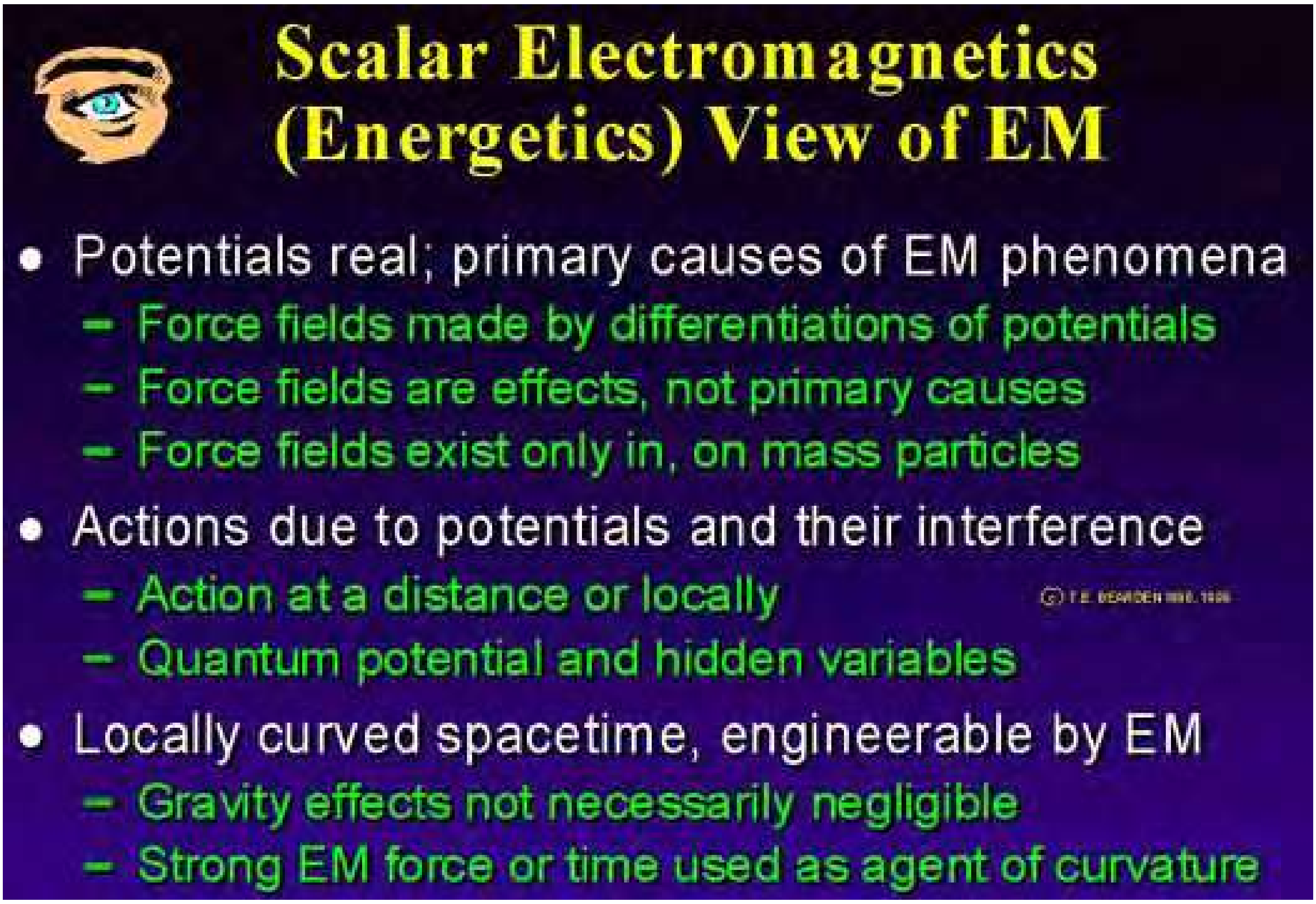

Scalar Electromagnetics (Energetics) View of EM

- Potentials real; primary causes of EM phenomena
  - Force fields made by differentiations of potentials
  - Force fields are effects, not primary causes
  - Force fields exist only in, on mass particles
- Actions due to potentials and their interference
  - Action at a distance or locally
  - Quantum potential and hidden variables
- Locally curved spacetime, engineerable by EM
  - Gravity effects not necessarily negligible
  - Strong EM force or time used as agent of curvature

© T.E. BEARDEN [illegible]

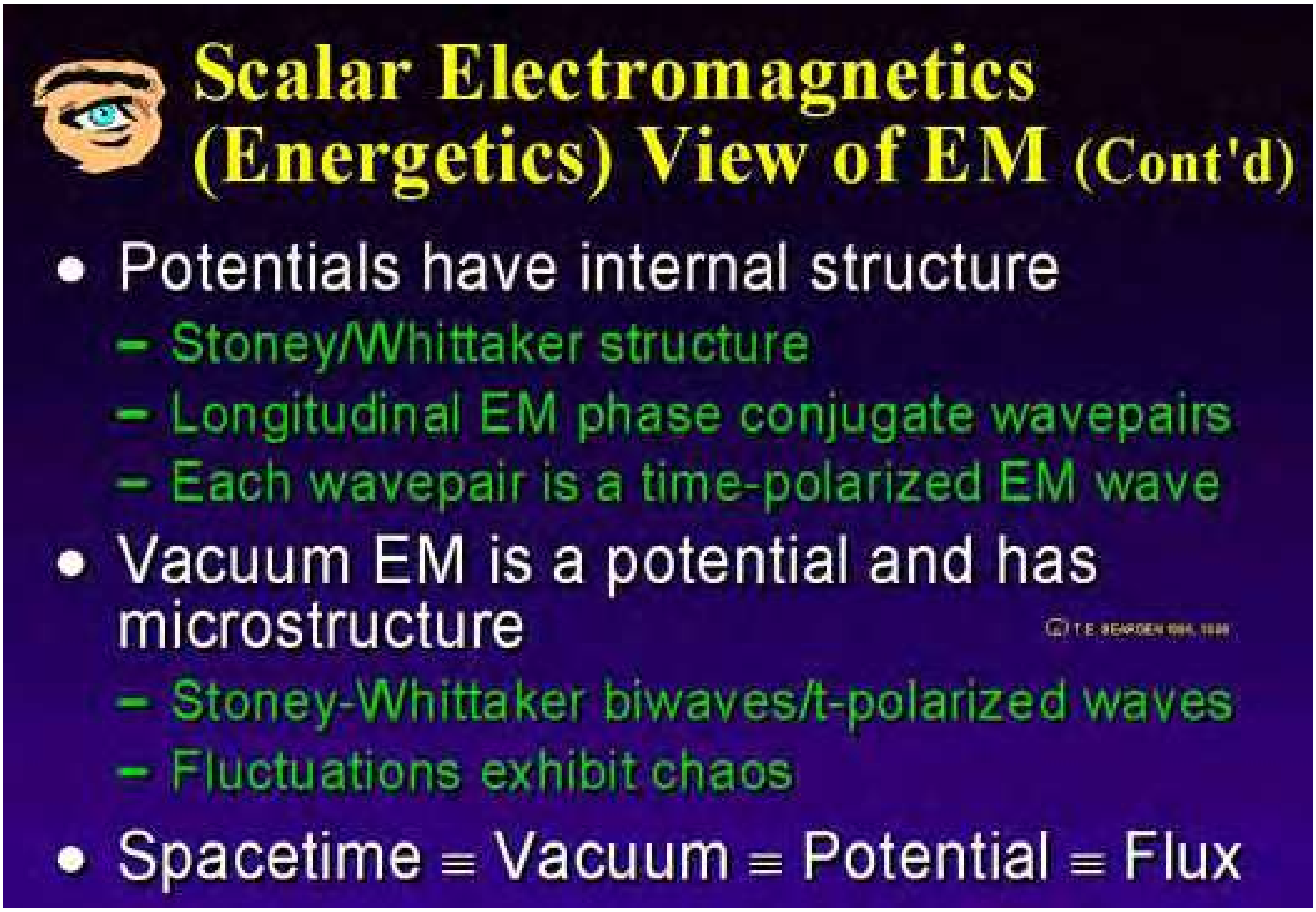

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Potentials have internal structure
  - Stoney/Whittaker structure
  - Longitudinal EM phase conjugate wavepairs
  - Each wavepair is a time-polarized EM wave
- Vacuum EM is a potential and has microstructure
  - Stoney-Whittaker biwaves/t-polarized waves
  - Fluctuations exhibit chaos
- Spacetime ≡ Vacuum ≡ Potential ≡ Flux

© T.E. BEARDEN [illegible]

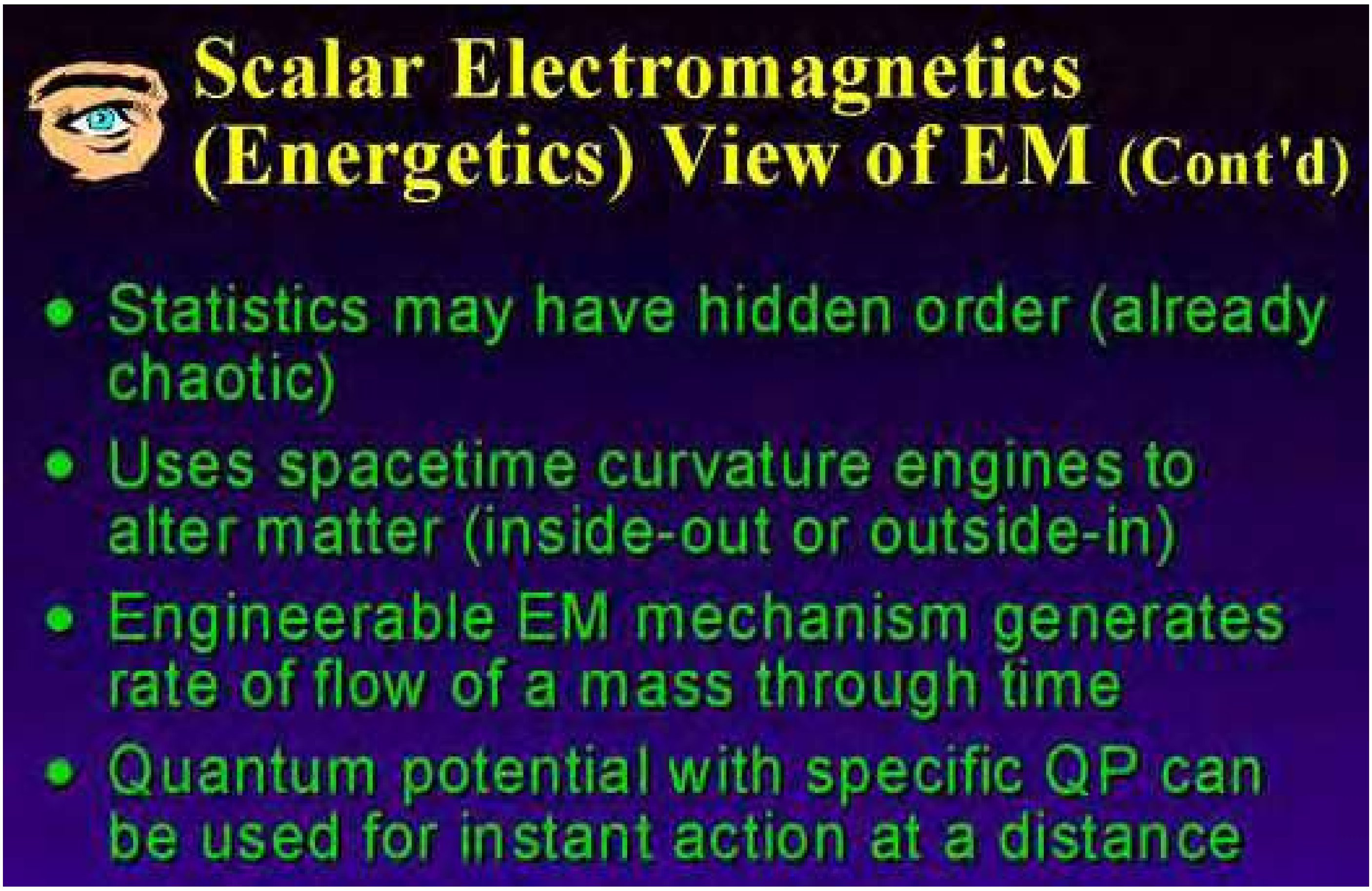

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Statistics may have hidden order (already chaotic)
- Uses spacetime curvature engines to alter matter (inside-out or outside-in)
- Engineerable EM mechanism generates rate of flow of a mass through time
- Quantum potential with specific QP can be used for instant action at a distance

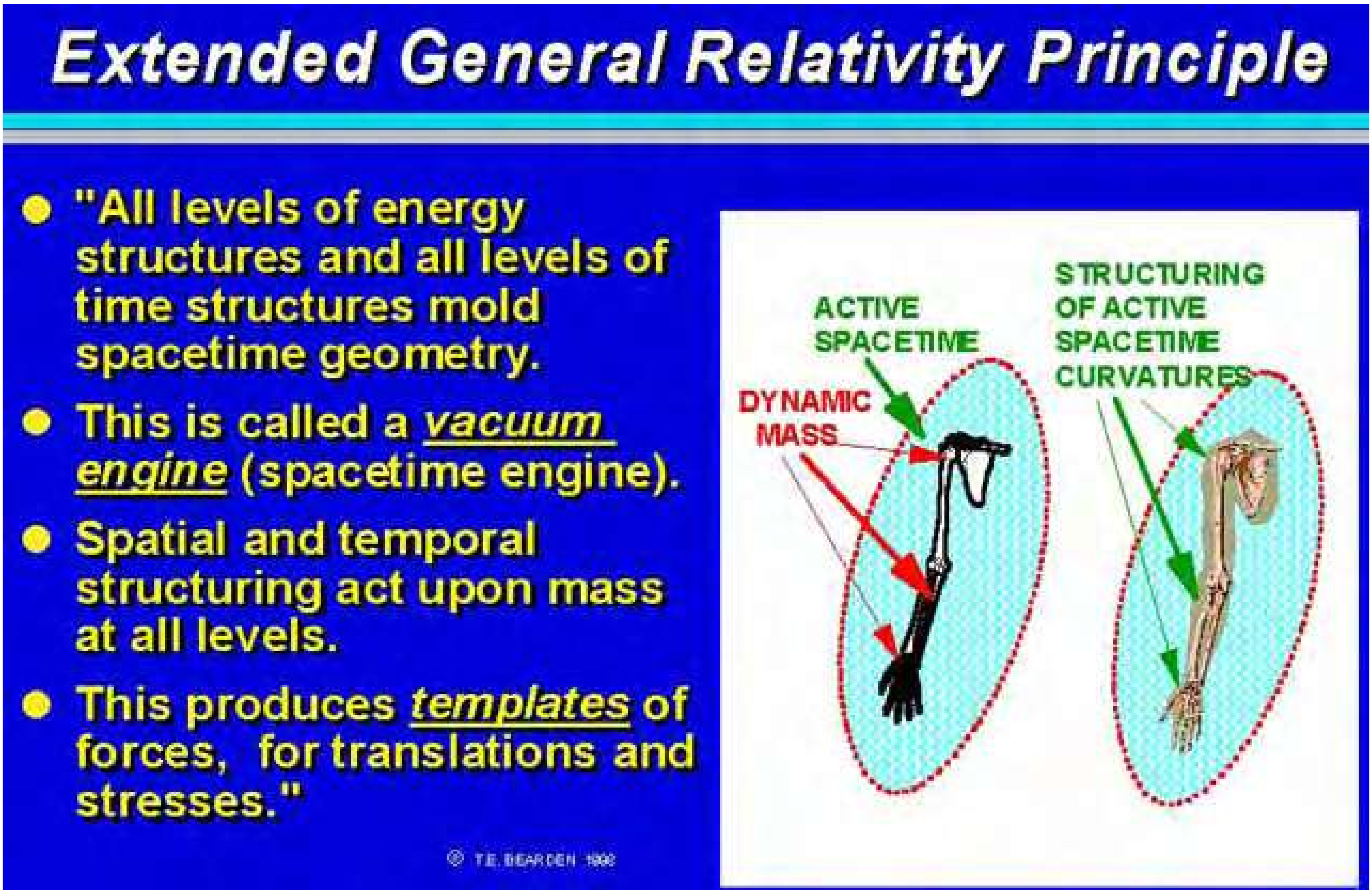

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a *vacuum engine* (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces *templates* of forces, for translations and stresses."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E. BEARDEN 1990

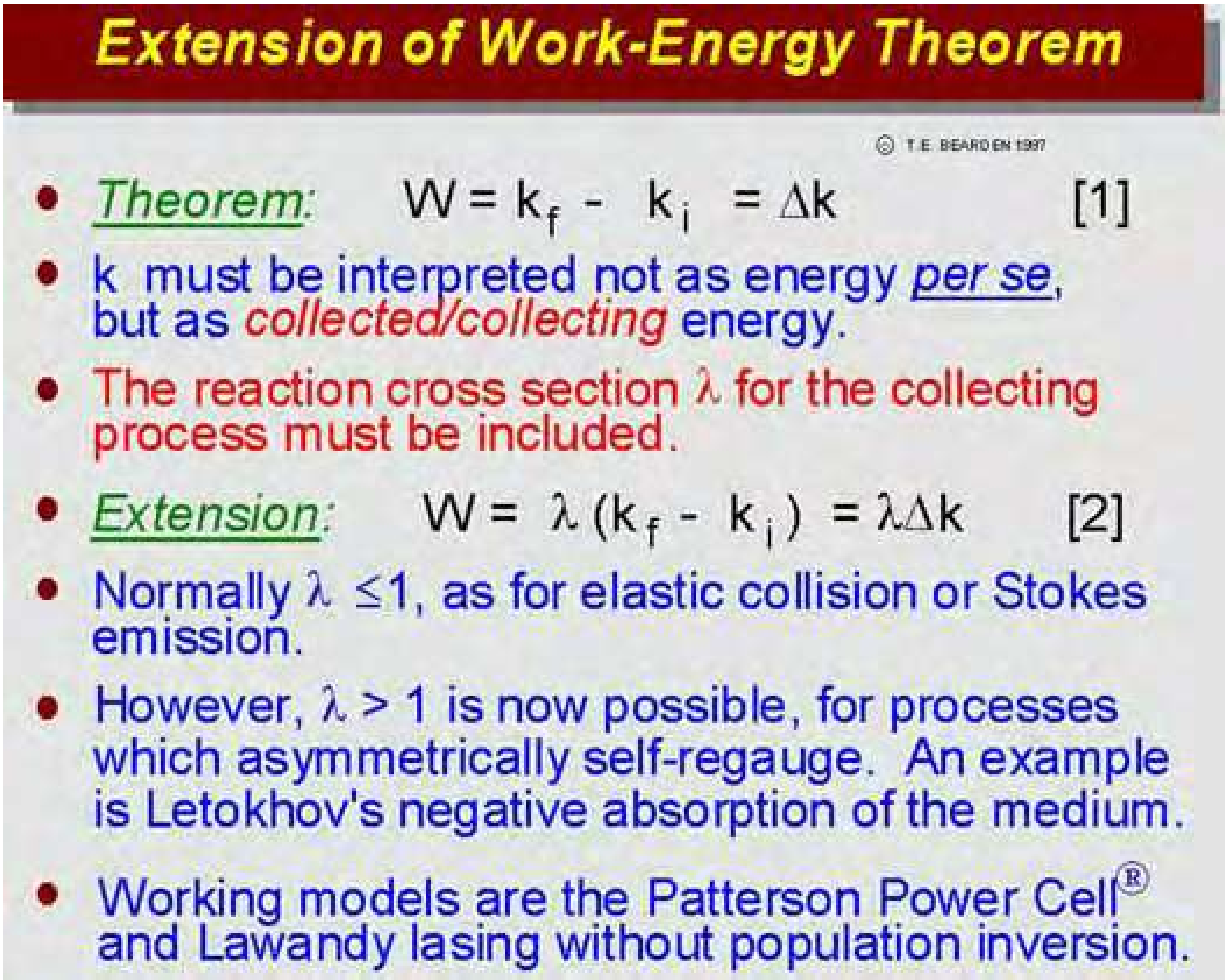

Extension of Work-Energy Theorem

© T.E. BEARDEN 1997

- *Theorem*: $W = k_f - k_i = \Delta k$ [1]
- k must be interpreted not as energy *per se*, but as *collected/collecting* energy.
- The reaction cross section λ for the collecting process must be included.
- *Extension*: $W = \lambda (k_f - k_i) = \lambda \Delta k$ [2]
- Normally $\lambda \leq 1$, as for elastic collision or Stokes emission.
- However, $\lambda > 1$ is now possible, for processes which asymmetrically self-regauge. An example is Letokhov's negative absorption of the medium.
- Working models are the Patterson Power Cell® and Lawandy lasing without population inversion.

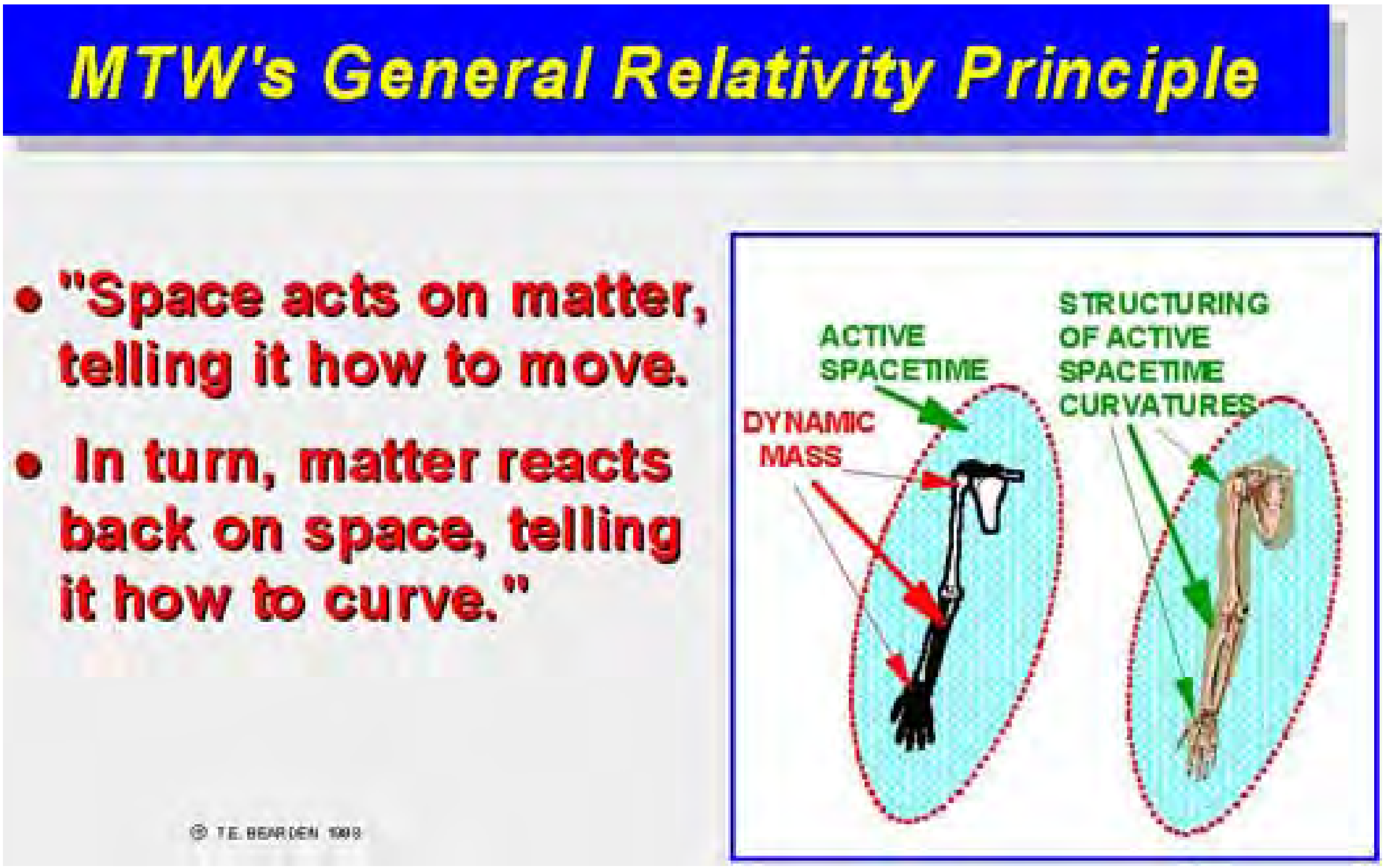

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E. BEARDEN 1998

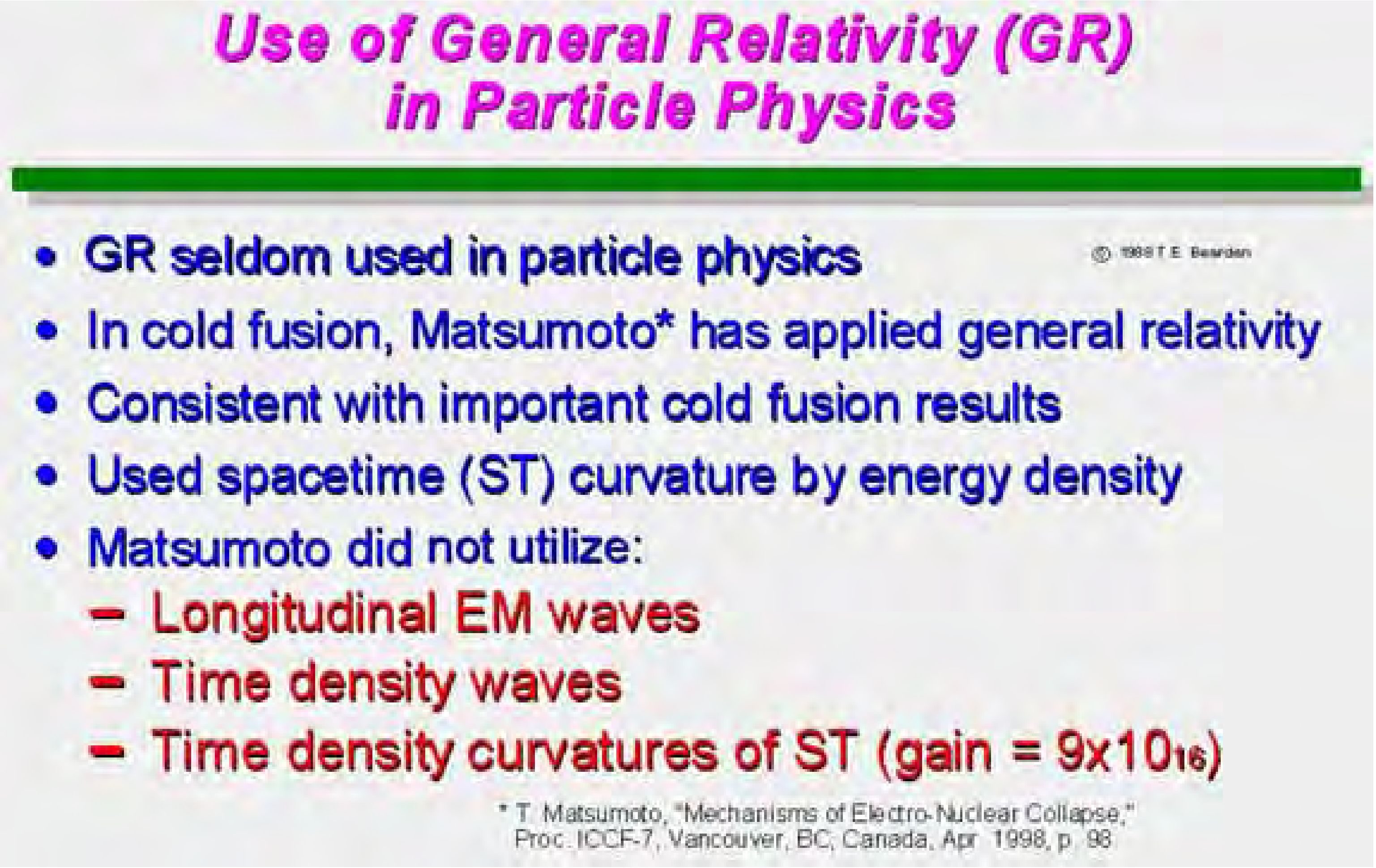

Use of General Relativity (GR) in Particle Physics

© 1998 T. E. Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density waves
  - Time density curvatures of ST (gain = 9x10^{16})

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98

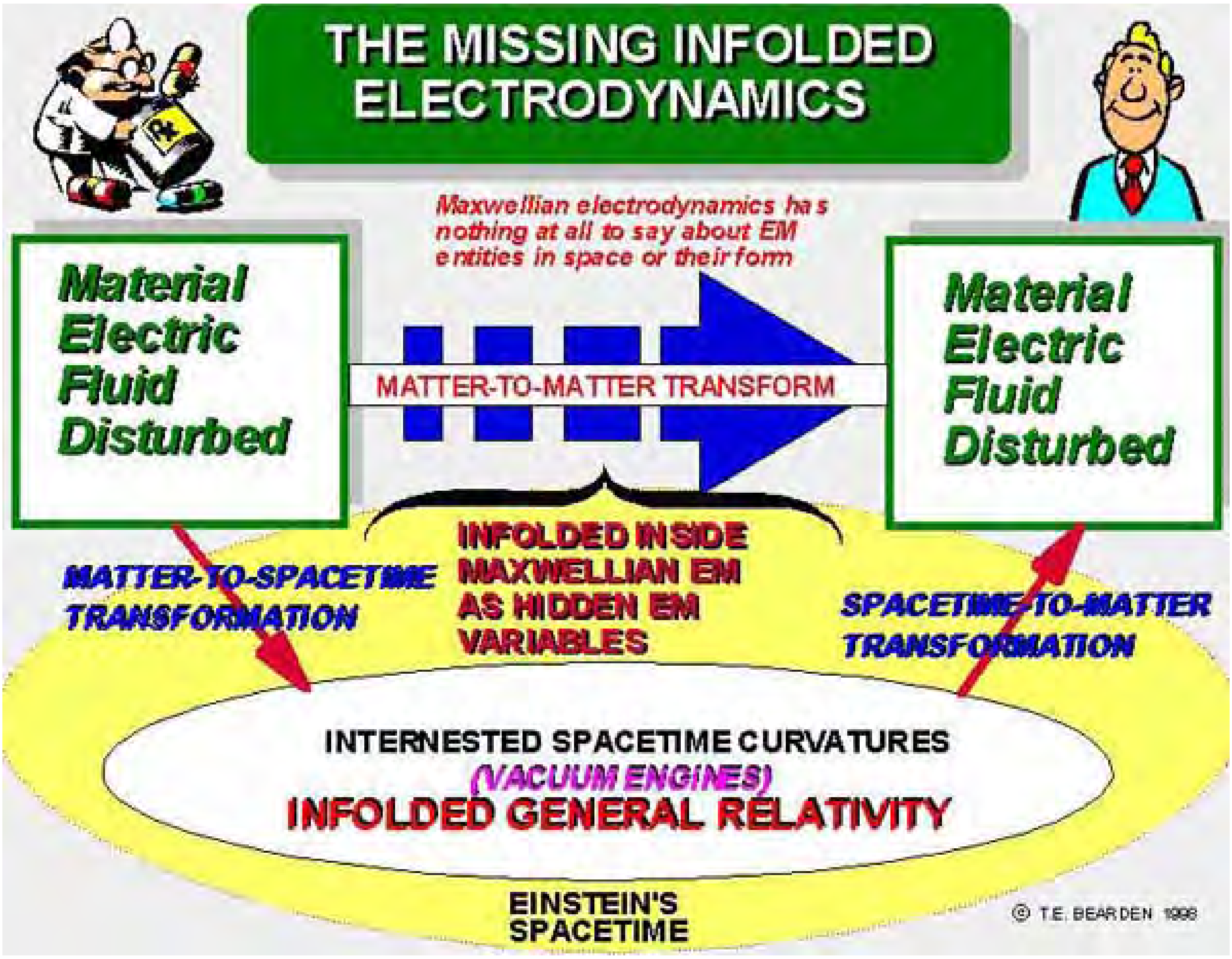

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

MATTER-TO-SPACETIME TRANSFORMATION

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

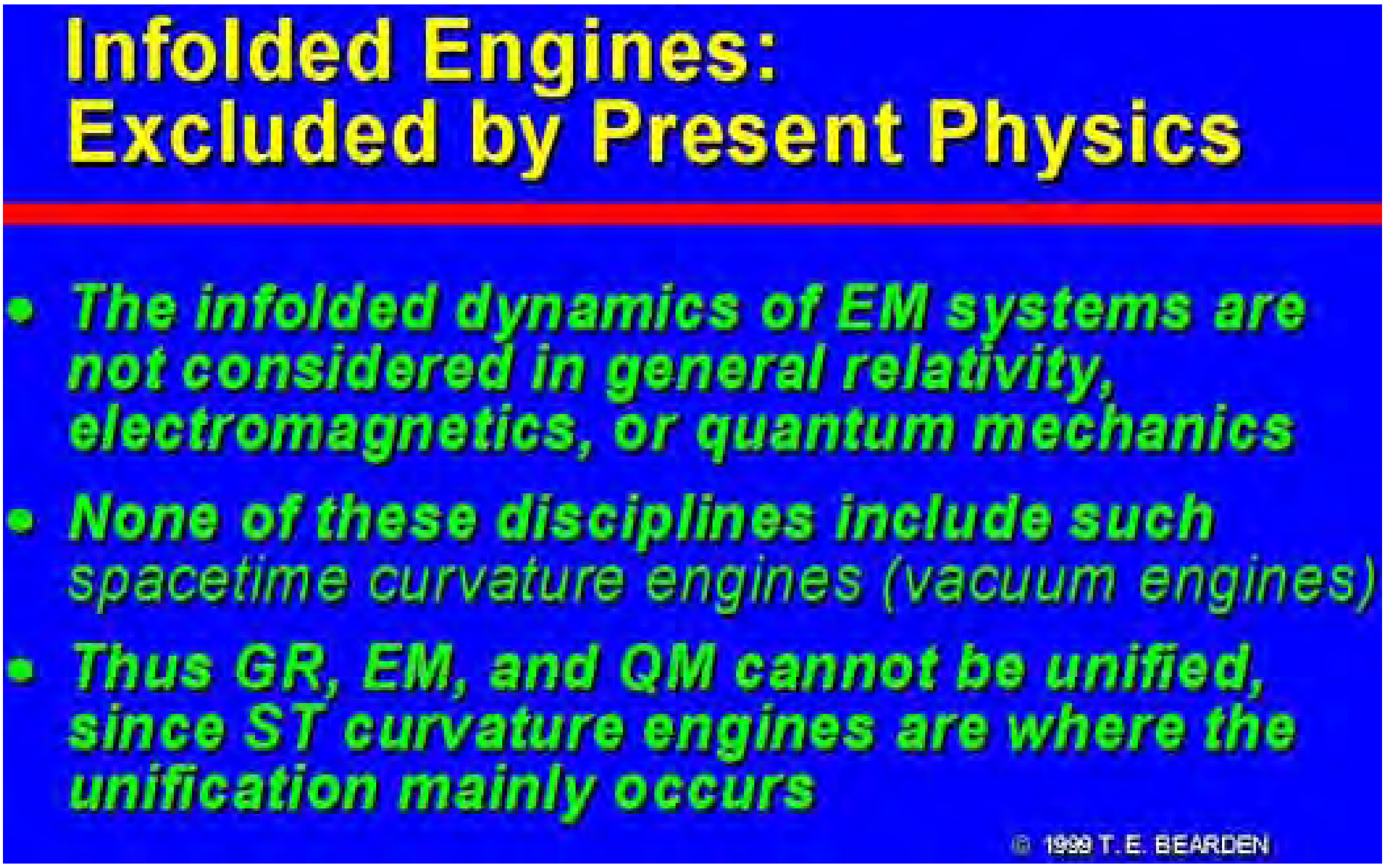

Infolded Engines: Excluded by Present Physics

- *The infolded dynamics of EM systems are not considered in general relativity, electromagnetics, or quantum mechanics*
- *None of these disciplines include such spacetime curvature engines (vacuum engines)*
- *Thus GR, EM, and QM cannot be unified, since ST curvature engines are where the unification mainly occurs*

© 1999 T. E. BEARDEN

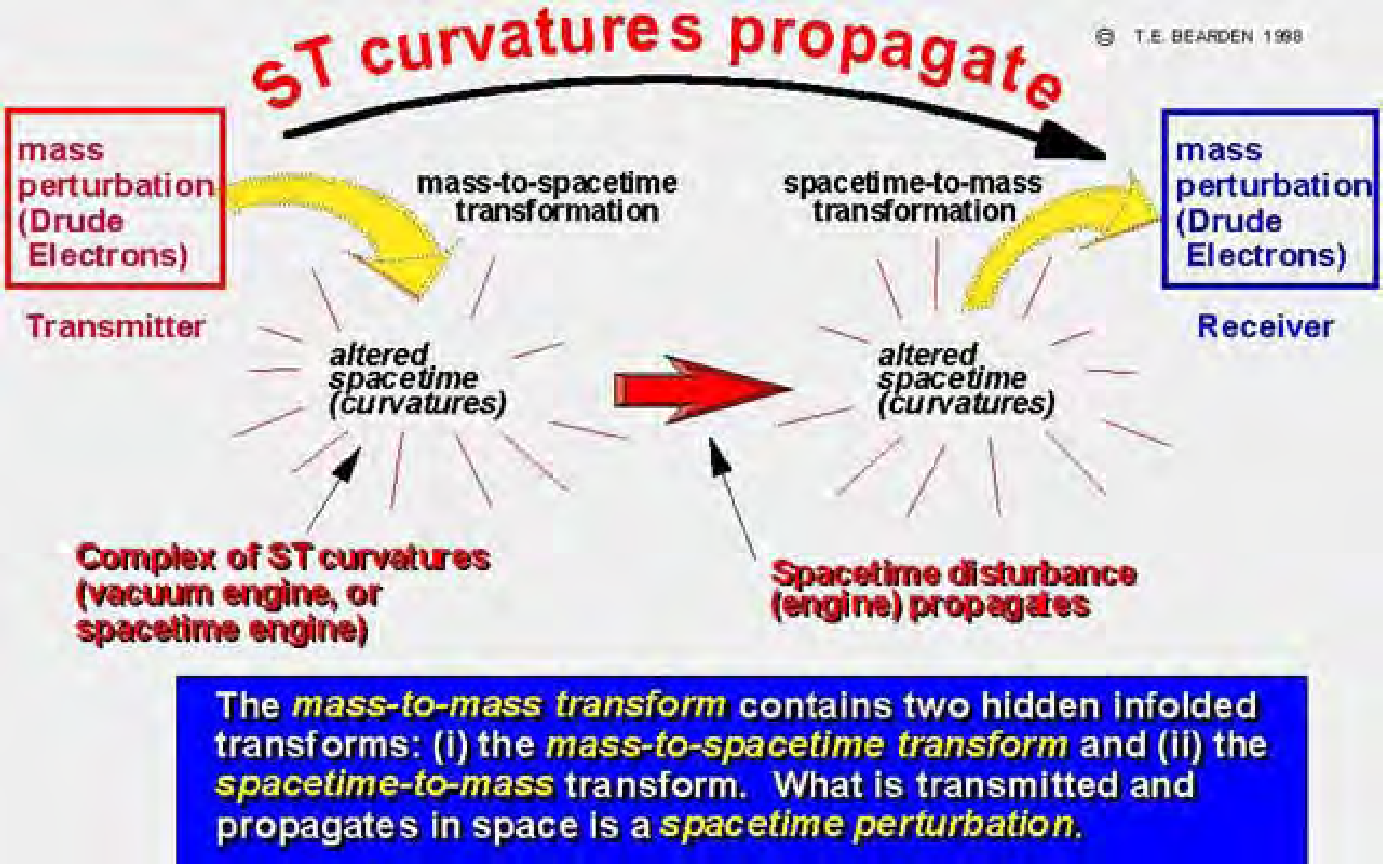

ST curvatures propagate

© T.E. BEARDEN 1998

mass perturbation (Drude Electrons)

Transmitter

mass-to-spacetime transformation

altered spacetime (curvatures)

spacetime-to-mass transformation

altered spacetime (curvatures)

mass perturbation (Drude Electrons)

Receiver

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation.*

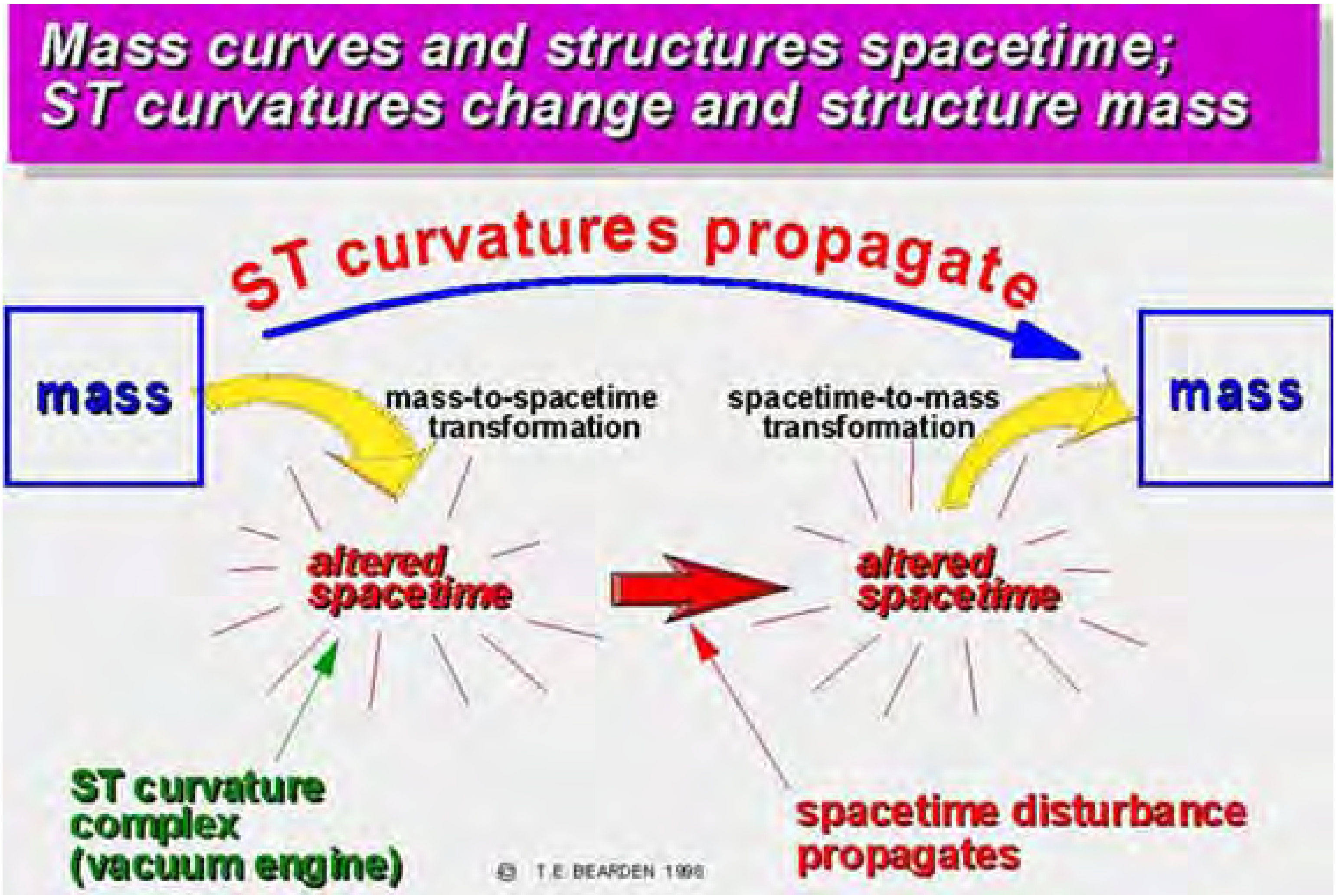

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

ST curvature complex (vacuum engine)

spacetime disturbance propagates

© T.E. BEARDEN 1996

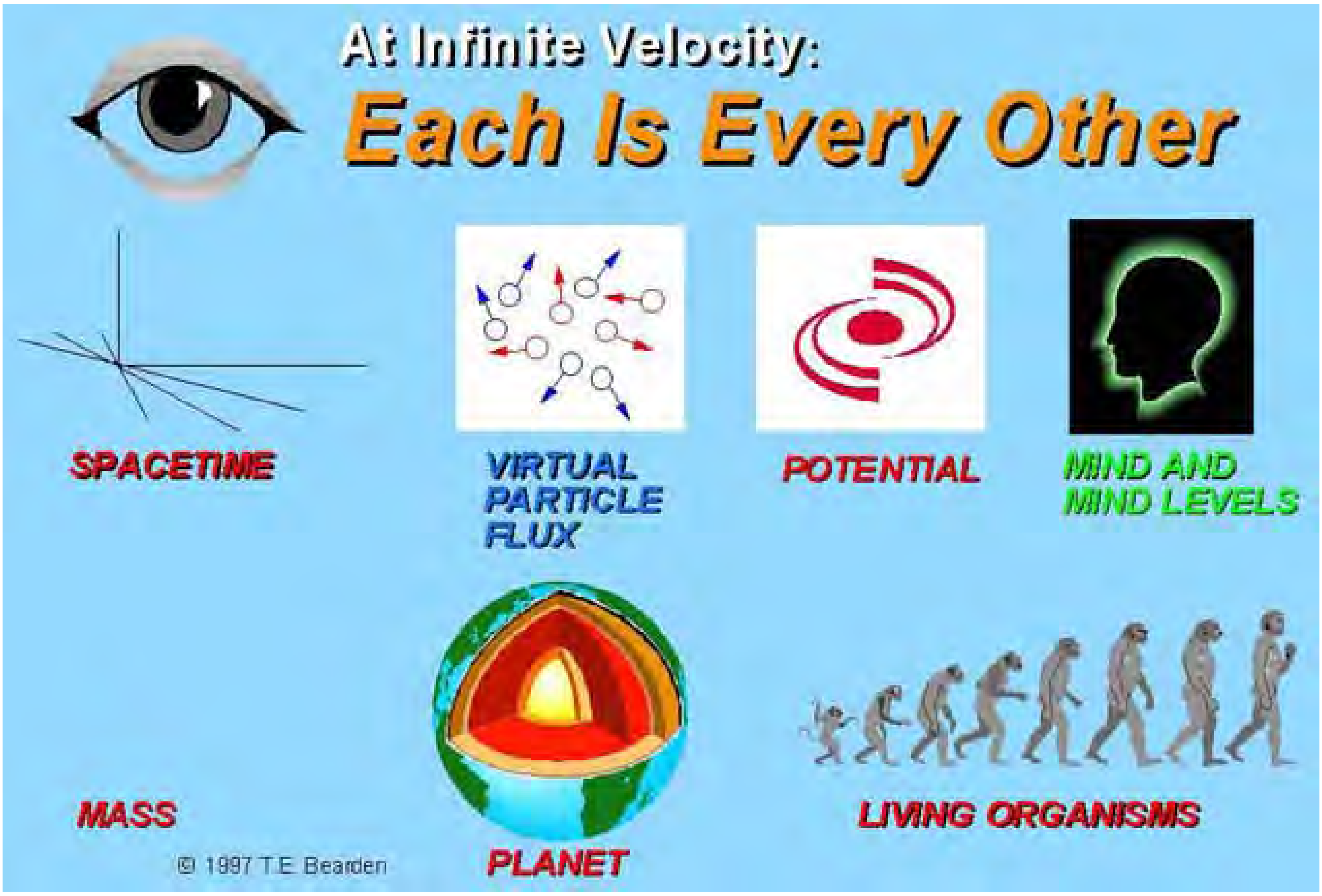

At Infinite Velocity:

Each Is Every Other

SPACETIME

VIRTUAL
PARTICLE
FLUX

POTENTIAL

MIND AND
MIND LEVELS

MASS

PLANET

LIVING ORGANISMS

© 1997 T.E. Bearden

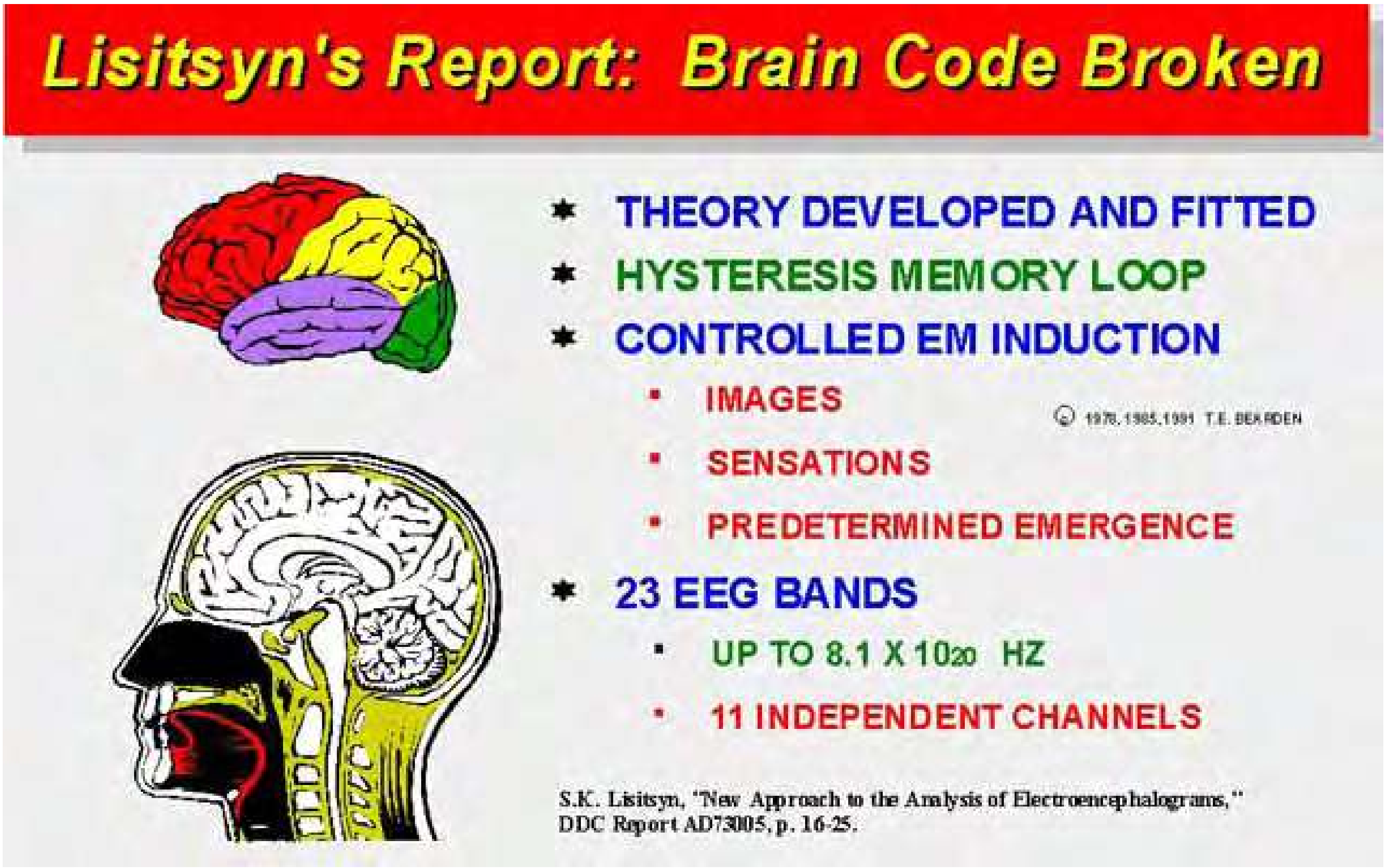

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

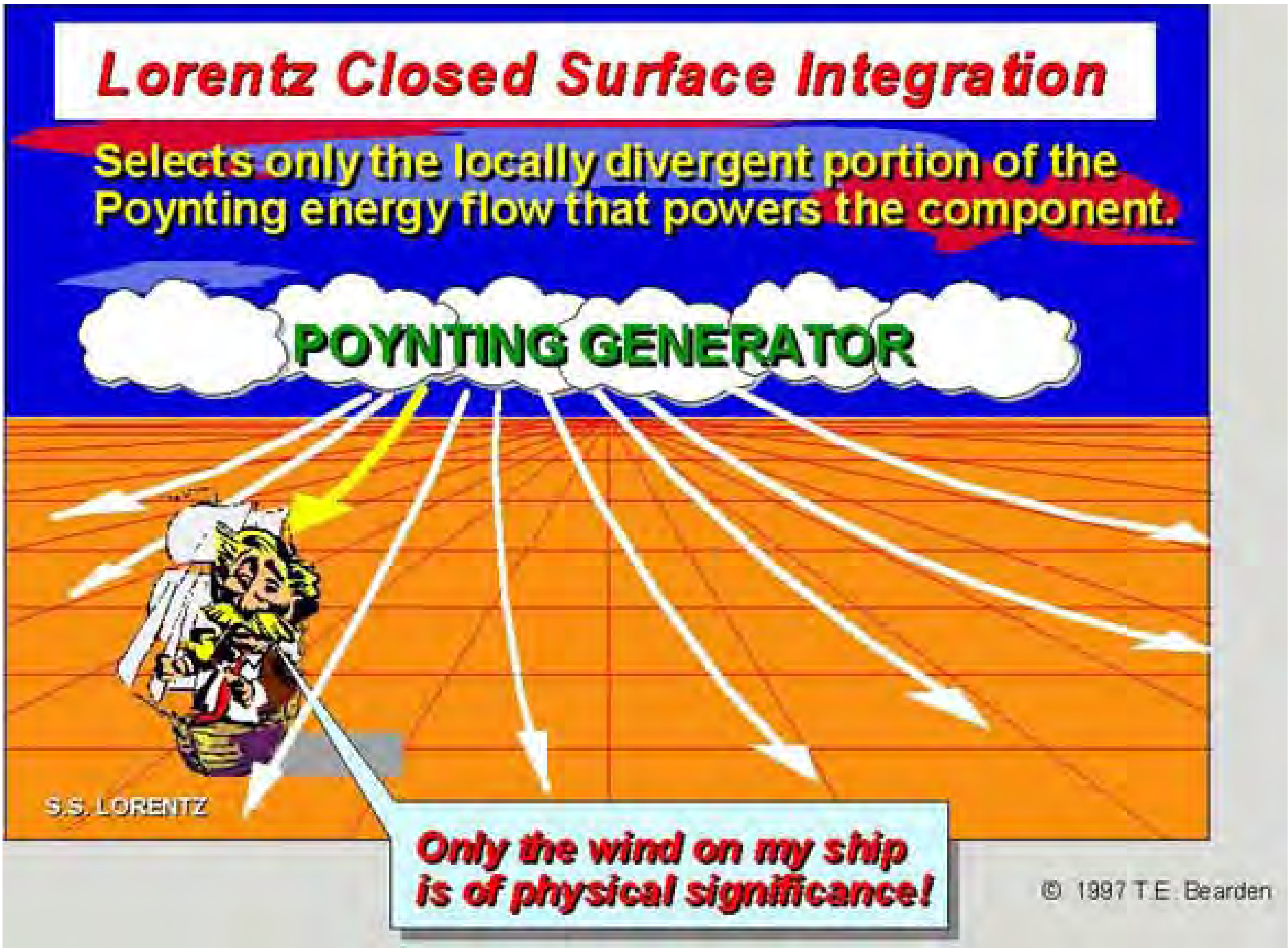

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

POYNTING GENERATOR
S.S. LORENTZ
Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

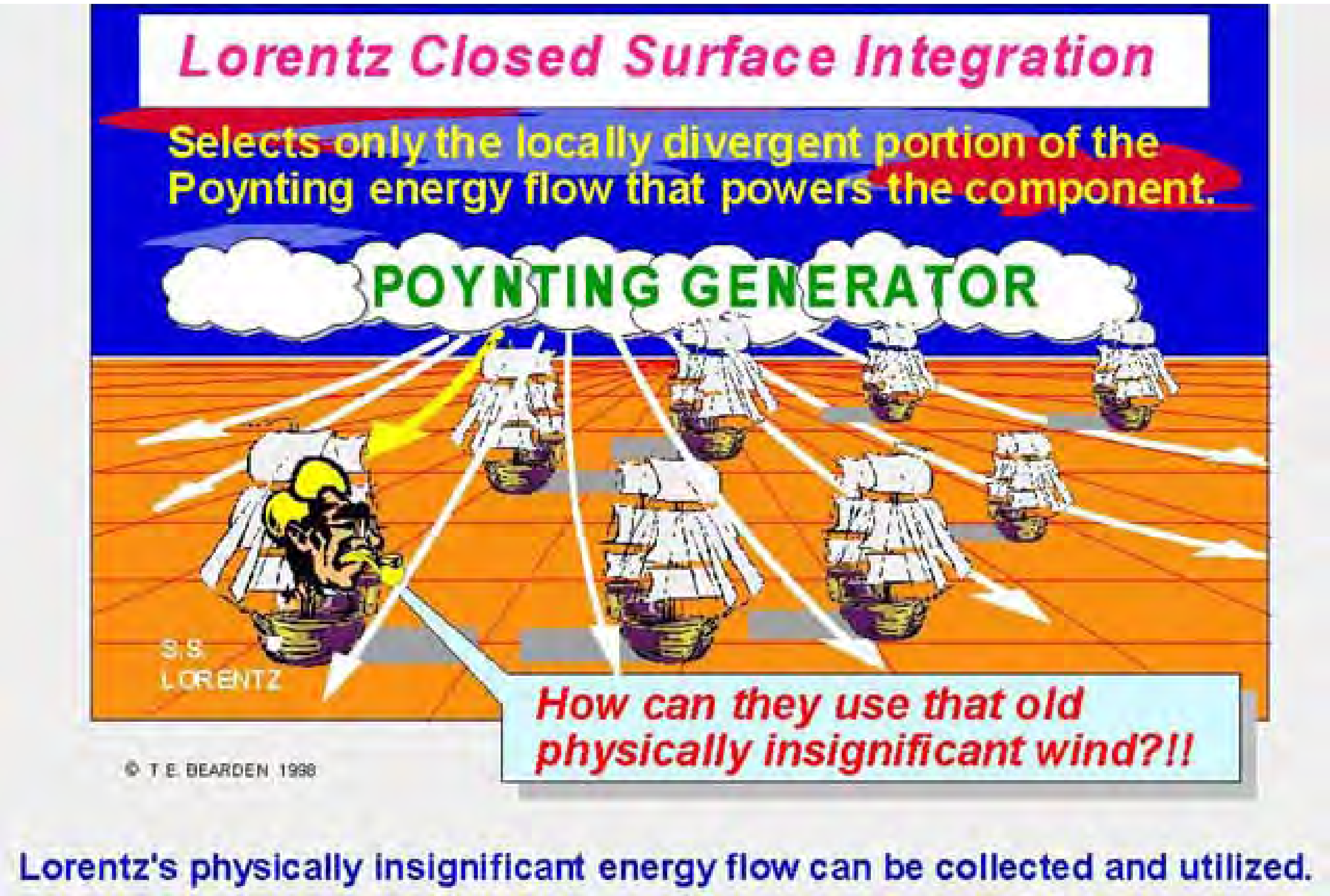

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

How can they use that old physically insignificant wind?!!

© T E BEARDEN 1998

Lorentz's physically insignificant energy flow can be collected and utilized.

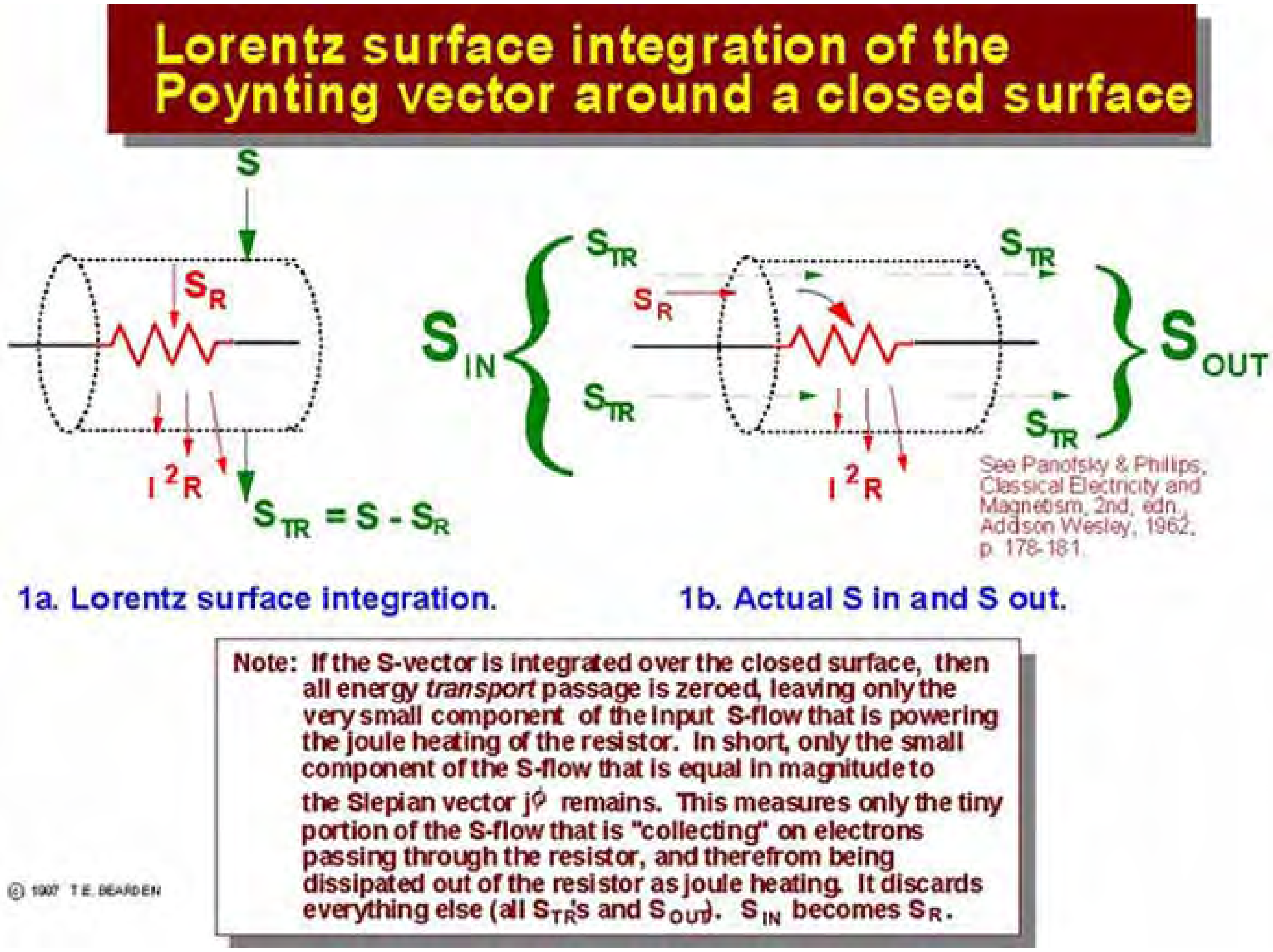

Lorentz surface integration of the Poynting vector around a closed surface

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd, edn., Addison Wesley, 1962, p. 178-181.

1a. Lorentz surface integration.

1b. Actual S in and S out.

Note: If the S-vector is integrated over the closed surface, then all energy *transport* passage is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Slepian vector $j\phi$ remains. This measures only the tiny portion of the S-flow that is "collecting" on electrons passing through the resistor, and therefrom being dissipated out of the resistor as joule heating. It discards everything else (all S_{TR}'s and S_{OUT}). S_{IN} becomes S_R.

© 1997 T.E. BEARDEN

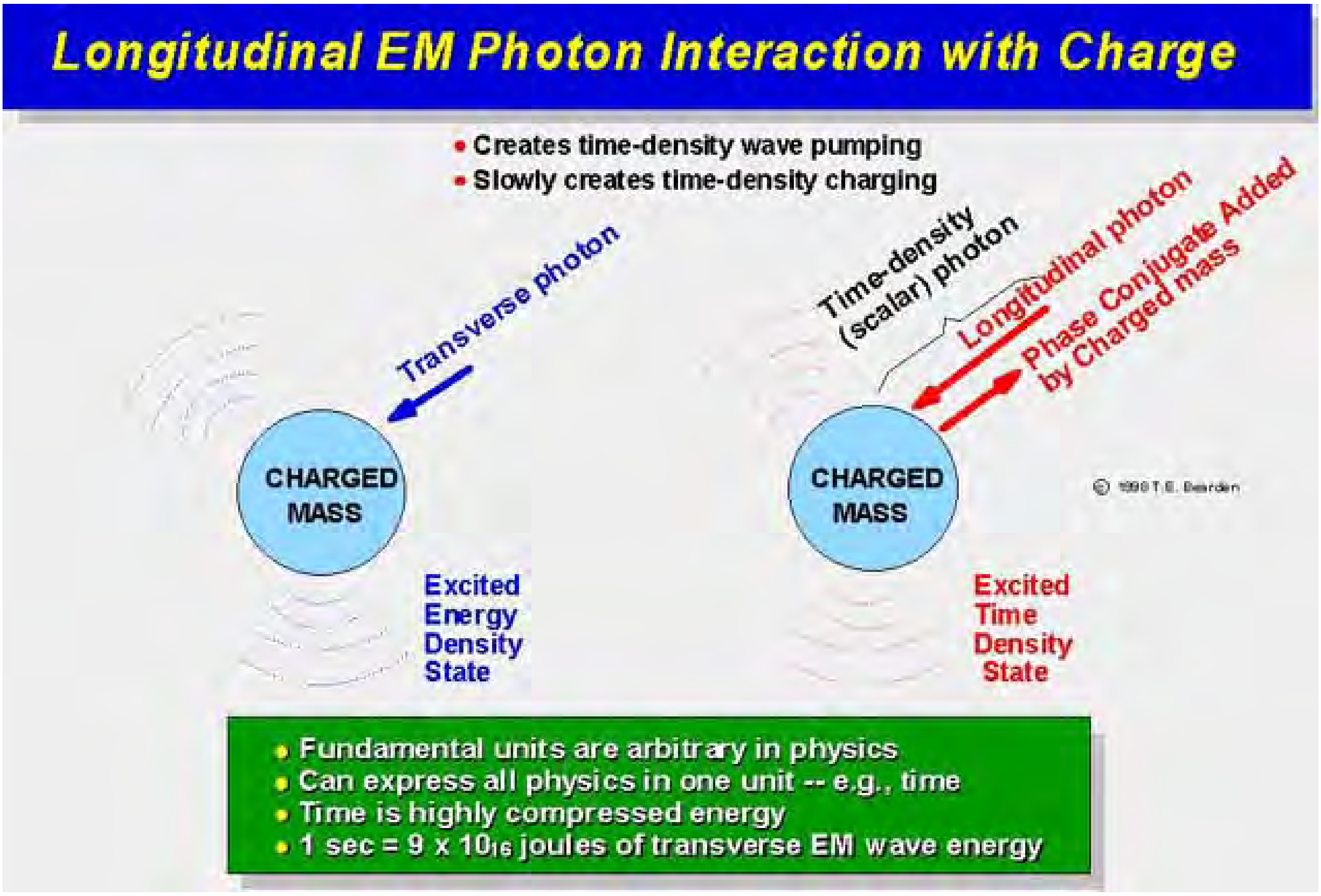

Longitudinal EM Photon Interaction with Charge

- Creates time-density wave pumping
- Slowly creates time-density charging

Transverse photon

CHARGED MASS

Excited Energy Density State

Time-density (scalar) photon

Longitudinal photon

Phase Conjugate Added by Charged mass

CHARGED MASS

Excited Time Density State

© 1990 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9 x 10_{16} joules of transverse EM wave energy

Figure 1-8. Infolded biwave composition of a scalar potential.

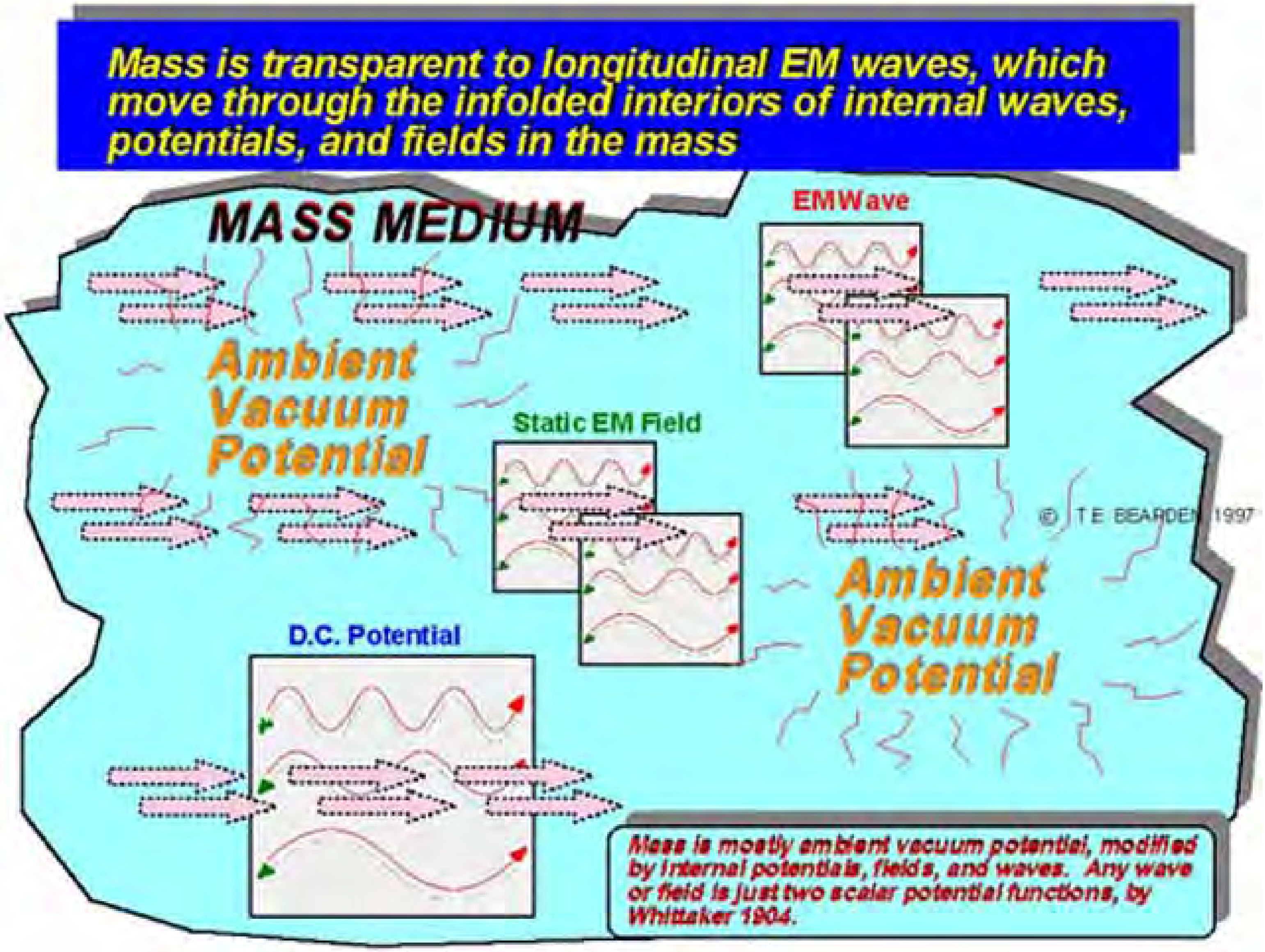

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
Ambient Vacuum Potential
EM Wave
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

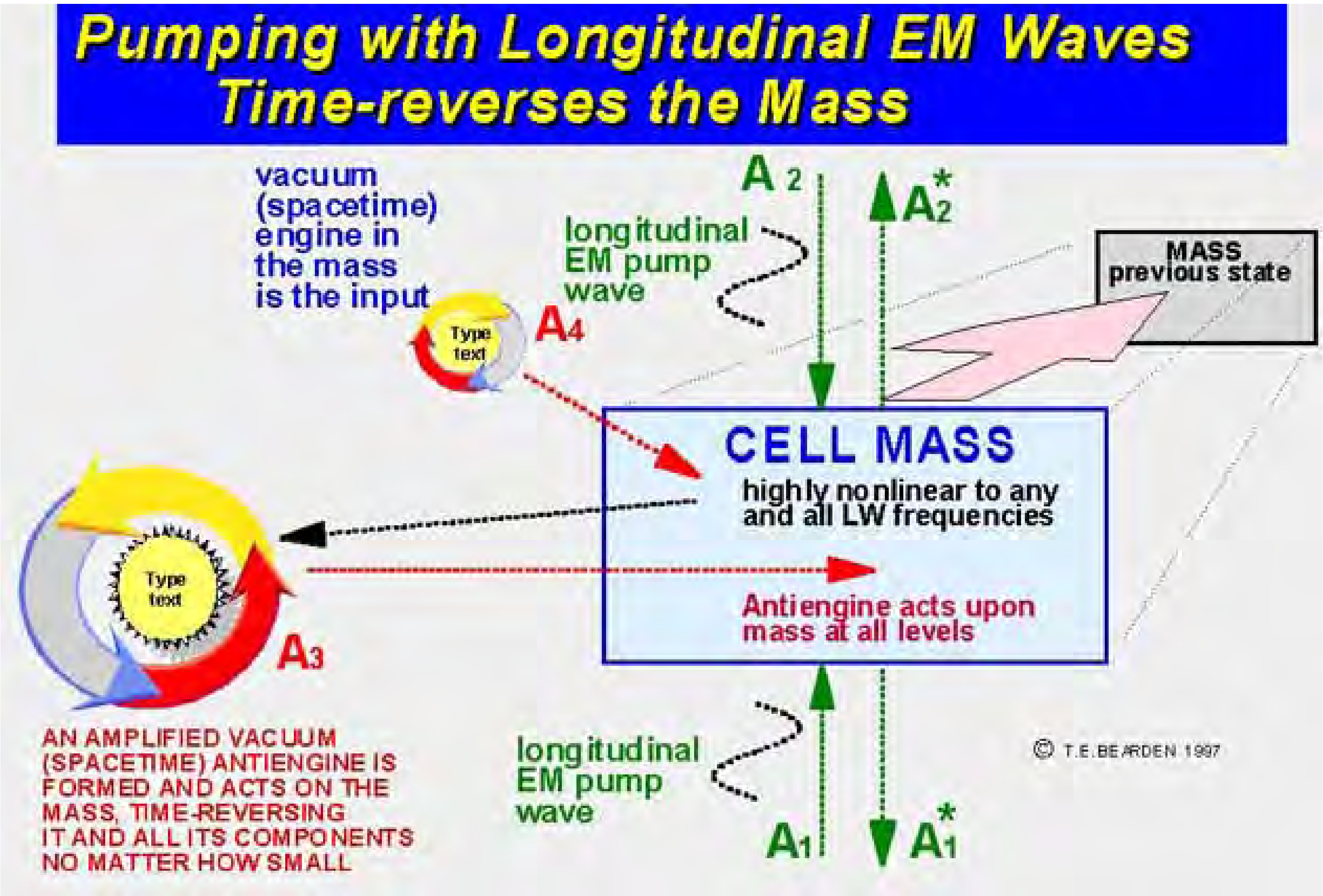

Pumping with Longitudinal EM Waves Time-reverses the Mass

vacuum (spacetime) engine in the mass is the input

longitudinal EM pump wave

MASS previous state

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

longitudinal EM pump wave

© T.E.BEARDEN 1997

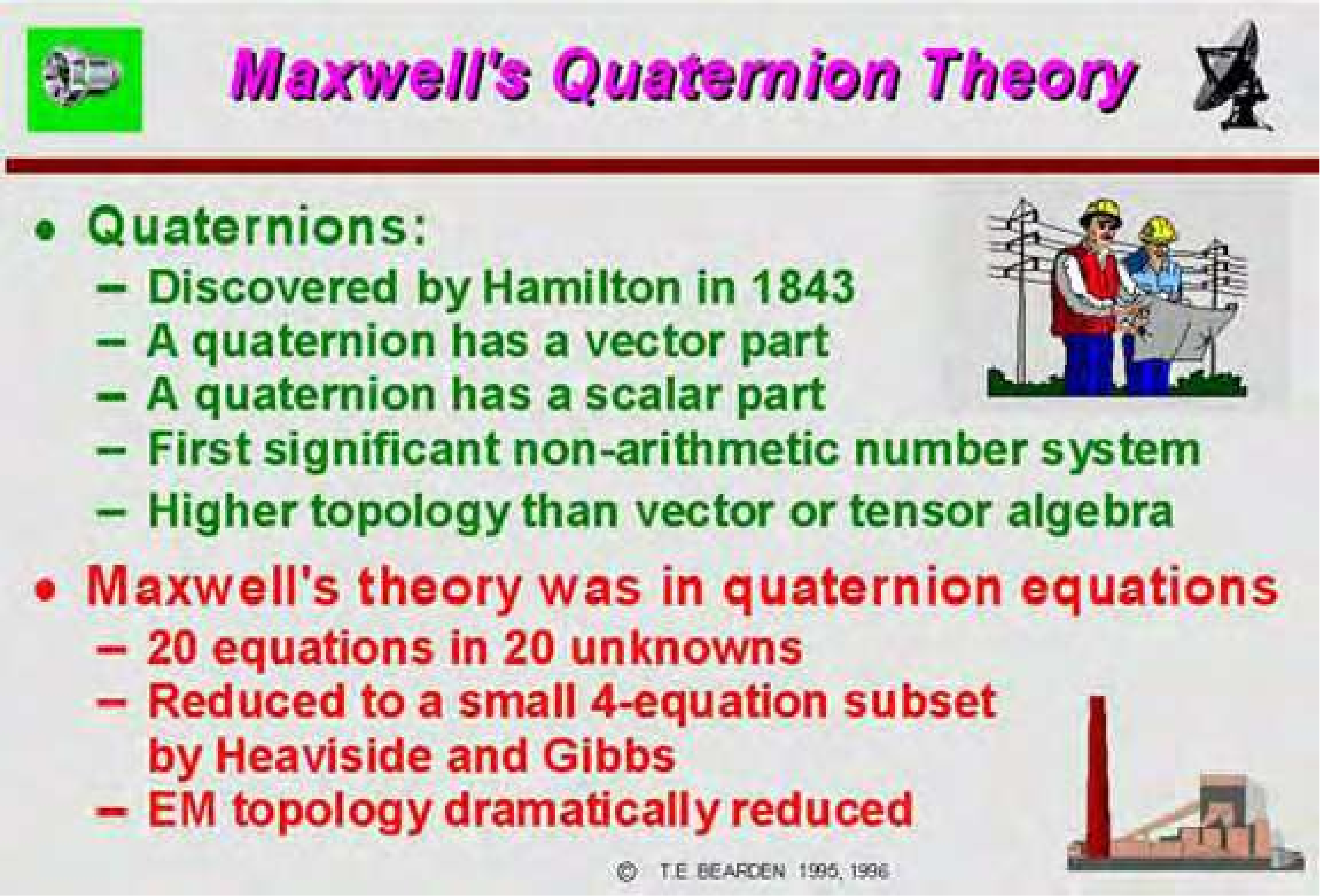

Maxwell's Quaternion Theory

- Quaternions:
  - Discovered by Hamilton in 1843
  - A quaternion has a vector part
  - A quaternion has a scalar part
  - First significant non-arithmetic number system
  - Higher topology than vector or tensor algebra
- Maxwell's theory was in quaternion equations
  - 20 equations in 20 unknowns
  - Reduced to a small 4-equation subset by Heaviside and Gibbs
  - EM topology dramatically reduced

© T.E. BEARDEN 1995, 1996

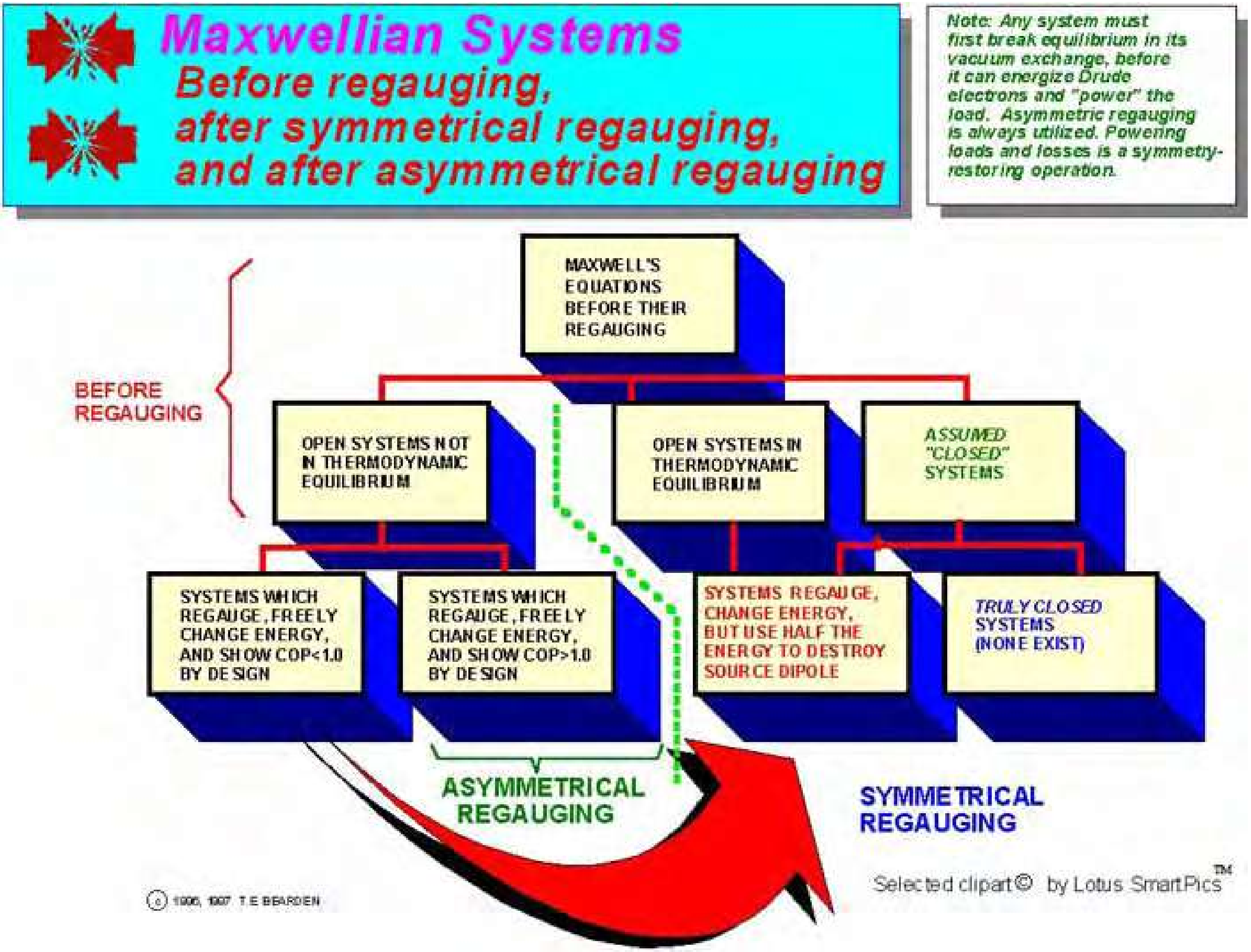

Maxwellian Systems
Before regauging, after symmetrical regauging, and after asymmetrical regauging

Note: Any system must first break equilibrium in its vacuum exchange, before it can energize Drude electrons and "power" the load. Asymmetric regauging is always utilized. Powering loads and losses is a symmetry-restoring operation.

MAXWELL'S EQUATIONS BEFORE THEIR REGAUGING

BEFORE REGAUGING

OPEN SYSTEMS NOT IN THERMODYNAMIC EQUILIBRIUM

OPEN SYSTEMS IN THERMODYNAMIC EQUILIBRIUM

ASSUMED "CLOSED" SYSTEMS

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP<1.0 BY DESIGN

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP>1.0 BY DESIGN

SYSTEMS REGAUGE, CHANGE ENERGY, BUT USE HALF THE ENERGY TO DESTROY SOURCE DIPOLE

TRULY CLOSED SYSTEMS (NONE EXIST)

ASYMMETRICAL REGAUGING

SYMMETRICAL REGAUGING

Selected clipart© by Lotus SmartPics™

ⓒ 1996, 1997 T.E BEARDEN

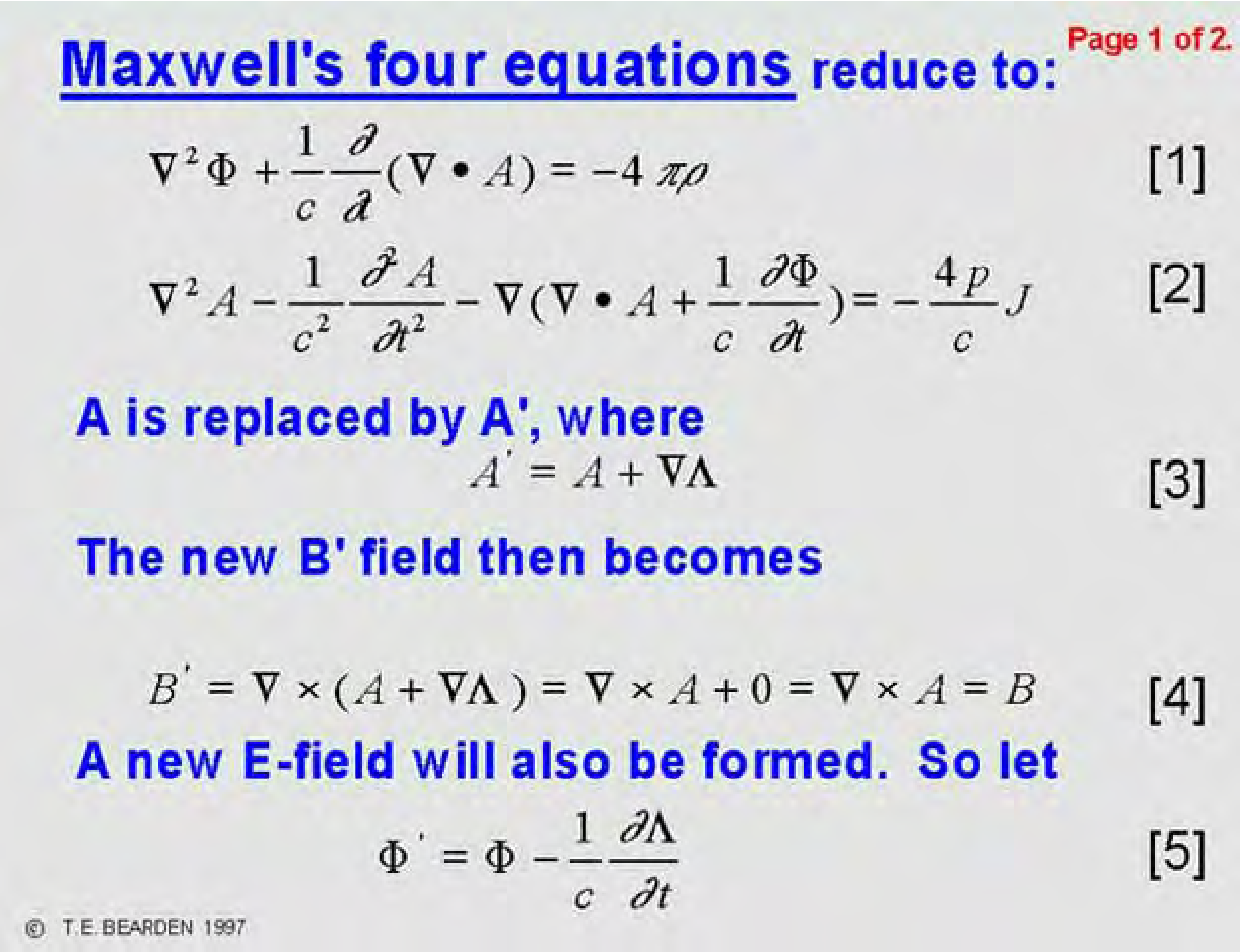

Page 1 of 2.

Maxwell's four equations reduce to:

$$\nabla^2\Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet A) = -4\pi\rho \qquad [1]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} - \nabla\left(\nabla \bullet A + \frac{1}{c}\frac{\partial\Phi}{\partial t}\right) = -\frac{4p}{c}J \qquad [2]$$

A is replaced by A', where

$$A' = A + \nabla\Lambda \qquad [3]$$

The new B' field then becomes

$$B' = \nabla \times (A + \nabla\Lambda) = \nabla \times A + 0 = \nabla \times A = B \qquad [4]$$

A new E-field will also be formed. So let

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t} \qquad [5]$$

© T.E. BEARDEN 1997

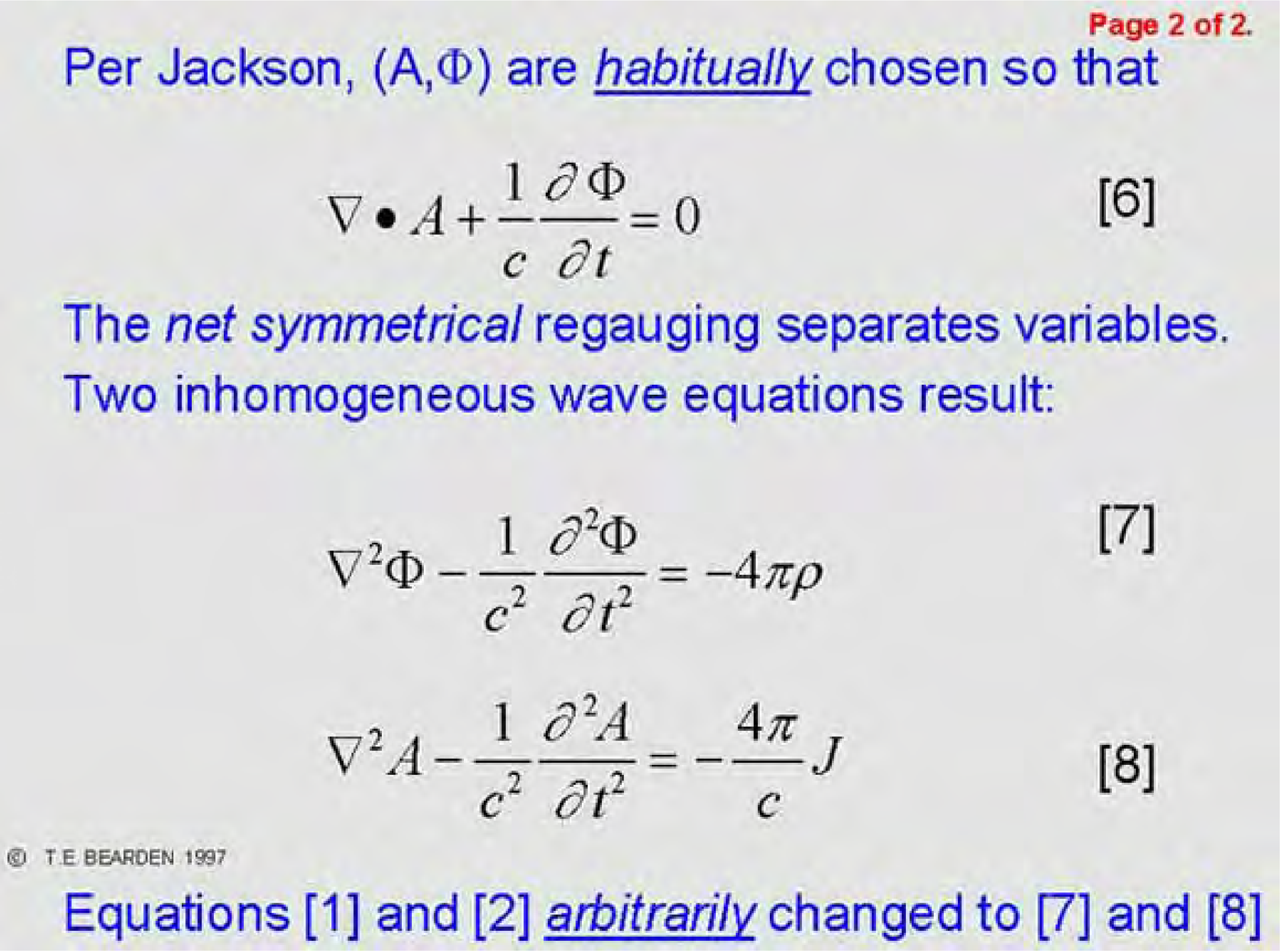

Per Jackson, (A,Φ) are *habitually* chosen so that

$$\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [6]$$

The *net symmetrical* regauging separates variables. Two inhomogeneous wave equations result:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [7]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{4\pi}{c} J \qquad [8]$$

© T.E. BEARDEN 1997

Equations [1] and [2] *arbitrarily* changed to [7] and [8]

HUNTSVILLE EDITION

The Southeast's Oldest E-Zine!

CHASING THE WILD DRAGON

Foundations of a New Science

By T.E. Bearden
November 12, 1995

It is a pleasure to accept the offer of John and Larry to write a column in their new InterNet magazine. Theirs is a great venture, and with the energy and knowledge they bring to the task, one can be confident of their success in making this a very interesting and informative publication.

There is a new "rebel" science slowly being born, with most of the work being done outside the universities and orthodox scientific community. In the months ahead it will be my job to try to acquaint you with the gist of what's going on, and what all the excitement is about.

This first article is deliberately just introductory, and its purpose is to interest you in the unusual fields we are going to be covering in more scientific detail, in future columns. We will also be listing a variety of references from the technical literature, which bear upon the various points addressed.

In this first article a little of my own background experience [note 1] is given, so you can decide whether you will be interested in what we have to say!

The real purpose of the column will just be to *present information* that you may not be able to readily obtain any other way. We will not argue or debate with skeptics; nothing productive is to be gained by that. If the information we present is useful to you, then our purpose has been accomplished. If it isn't of use to you, then just file it in the old wastebasket. Also, from time to time I may answer a generic question, constructed from multiple queries. Individual responses to correspondence cannot be provided.

The work toward a new breakthrough science is going on in a broad spectrum of "fields," most of which are not yet recognized by academia as even constituting legitimate areas of scientific endeavor (Table 1).
Some of the major areas involved are:

- theoretical and experimental research in overunity electrical machines,
- free energy research of other kinds, such as devices utilizing thermodynamics phase changes,
- rapidly advancing efforts in cold fusion,

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

- direct engineering of quantum change via a new kind of "hidden electrodynamic variables,"
- the direct engineering of the vacuum's virtual photon flux exchange with matter,
- action at a distance, including the deliberate formation and usage of quantum potentials,
- direct engineering of the local curvature of spacetime (yes, the direct engineering of general relativity, using the new hidden kind of electrodynamics),
- use of the new "hidden" electromagnetics to engineer body cells including the reversal of diseased cells, genetics and all, back to a previous healthy state (cellular dedifferentiation, in biophysical terms, or cellular phase conjugation, in physics terminology),
- research toward the direct engineering of the mind, including both the conscious and unconscious minds, long-term memory, and the personality itself, and
- a ruthless re-examination (and correction where necessary and possible) of the definitions, concepts, and postulates that present physics is founded upon.

Each of these "major areas" as yet has no single "experts" in the sense that individuals exist who possess a "mastery" of the given area. As an example, in spite of multiple researchers and decades of vigorous research, no one is really an "overall expert" in the overunity electrical power systems field. The "field" itself is still struggling to be born and is not recognized at all by orthodoxy; in fact, many orthodox scientists bitterly resent one even attempting to work in this area. There is as yet no cohesive single "theory" of overunity electrical or magnetic machines, although our own group is making rapid progress in that respect. But there is still no universal agreement amongst even the researchers themselves. Literally the field is still being discovered.

This "groping dimly toward a new field" should not be too surprising. After all, the early electromagnetics researchers went through the same tortuous, agonizing kind of development. They were playing with cat fur, glass rods, and pith balls for quite an extended period before understanding progressed to the point where Maxwell could mathematically tie together and model the results of the experimenters such as Faraday and others. It has taken a century and a half for modern electromagnetics and electronics to reach its own present development. It is going to take another decade or so before the overunity researchers can overturn several ill-founded parts of the present electromagnetics and get the emerging field onto a solid foundations footing and into a solid theoretical position.

In the effort to produce a theoretical model of electromagnetics, Maxwell and his predecessors performed a magnificent work. Yet they also made very serious errors (Table 2), which persist to this day. [note 2] Let us list just a few of the more blatant errors in classical EM theory today:

First, Franklin guessed wrong on which way electrical current flowed in a circuit. There is no definition at all of electrical charge in all of physics, and there is no acceptable definition of potential (that is, there is no rigorous definition of just good old "voltage.") There are no "fields of force" in the vacuum; there are just potential gradients __ because the vacuum is just one vast potential (virtual particle flux) in the first place. Force fields actually occur only in and of the matter in a material system. The primary causative agents of electromagnetics are not the so-called "force fields' at all, but are the potentials __ as is well-known in quantum mechanics. [note 3] Maxwell assumed a material ether in his equations. The Michelson-Morley experiment destroyed the material ether, but the Maxwell equations have never been changed accordingly. The electron was not yet discovered when the Maxwell equations were written; the theory does not address electron spin. Maxwell was already dead some two decades when Barus' 1898 paper [note 4] was published, pointing out the strange "backwards-traveling" wave (i.e., time-reversed wave) solution to the wave equation. In addition, Heaviside and Gibbs had already produced their vector truncation __ this truncated theory is the modern so-called "Maxwell's equations" that universities teach today __ of Maxwell's theory.

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

vector algebra, and seriously curtailed Maxwell’s theory when they “translated” it into the much lower topology of vectors and some four equations. The later addition of tensors as the basis mathematics did not restore the lost topology provided by the quaternion model. [note 6]

For the EM wave, Maxwell simply assumed the transverse EM wave that had come from the much earlier derivation of the familiar wave equation from the old “plucked taut string” notions. [note 7] Yet when one closely examines and corrects some of the erroneous *assumptions* [note 8] made for the actual mathematical derivation of this “taut string” transverse wave, one finds (Figure 1) that an equal-energy second antiphased transverse “antiwave” is actually produced by the instrument and added to the “plucked string’s” transverse wave disturbance of the air medium, to produce a longitudinal wave of compression and rarefaction in the air itself. The “transverse string wave” in fact never leaves the string. The string wave is not the wave that is launched into the medium.

A similar electromagnetic situation exists. In a whip antenna (Figure 2), the Drude electron gas in the conductor does exhibit a transverse “string-type” EM wave of electron precession. However, at the same time, by Newton’s third law (which Maxwell neglected because at the time the atom was regarded as just a blob with no internal structure) the positively charged, phase conjugated atomic nuclei have equal energy, highly damped transverse recoil waves created in them. Further, these “nuclei-recoil” waves are phase conjugate replicas of the electron gas waves. The total disturbance that is launched into the vacuum from the whip antenna is (a) the normal transverse disturbance of the virtual photon flux of vacuum caused by the electron gas wave, and (b) the coupled equal-energy transverse antiwave caused by the nuclei-recoil via Newton’s third law. The amplitude of the material antiwaves of nuclei vibration are highly damped because of the massiveness of the nuclei; however, they are of equal energy to the electron wave. But when launched into the same vacuum, the two equal-energy wave disturbances experience equal damping because they are in the same medium. Hence the wave and antiwave are (a) locked together in the vacuum, and (b) of equal energy, and (3) of equal amplitude once launched into the vacuum. What results is a vacuum EM wave of rarefaction (reduction in intensity) and compression (increase in intensity) of the virtual photon flux of vacuum __ just exactly like what is launched into the surrounding gaseous air medium from the “plucked taut string” and perturbed body of a physical instrument. The EM wave in vacuum is a longitudinal wave directly analogous to a sound wave __ just as Nikola Tesla stated. [note 9]

The vacuum may be considered a scalar potential, in which case it has an internal biwave structure (Figure 3). Interference of two potentials __ i.e., the hidden interference (Figure 4) of their internal wave structures __ in fact already creates all EM fields and waves.

When the incoming longitudinal “biwave” from the vacuum strikes the atoms of the receiving antenna, each atom may be considered a dynamic assembly of multiple dipoles, where each dipole consists of one of the electrons in the electron shells and one of the positive charges in the nucleus. The receiving dipole in the interacting atom splits the two waves apart again (Figure 5). The negatively charged end of the dipole is time-forward, and the positively charged end is time-reversed. Hence the separated normal-wave-half of the biwave interacts with the electron, to give again the familiar transverse “electron precession” wave that we detect with our “electron wiggle-detecting” instruments (Figure 6). This detected electron precession wave is what all our textbooks prescribe erroneously as the vacuum wave! The separated time-reversed-wave-half of the incoming vacuum wave is split off and interacts with the positive end of the receiving atomic dipole, in the time-reversed atomic nucleus, thus generating Newton’s third law reaction force to provide the recoil of the nucleus. Present physics has ignored this second wave altogether, because of the serious error in the original derivation of the wave equation from plucked strings. Thus

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

would appear to violate Newtonian third law reaction. Phase conjugate mirrors already do just exactly this when they emit a phase conjugate replica wave. They direct the antiwave outside and away from the atom instead of inwards into the nucleus. So they do not recoil, no matter how powerful a phase conjugate replica wave they emit.

Finally, we can also show that the EM wave in vacuum is longitudinal, by yet another approach. The electrons in the electron gas in a conductor are spinning, and they are severely constrained longitudinally to be able to move only very, very slowly down the wire. So when perturbed, they act as gyroscopes, and precess at right angles to the perturbing force. The fact that the gyroelectrons in a wire move radially almost entirely, and only “slip” down the wire very slowly with a “drift” velocity, already proves conclusively that the incoming vacuum disturbance was longitudinal. The precession of a gyro is at right angles to the perturbing force. Therefore rigorously the vacuum EM waves are longitudinal, because the gyroelectrons moved transversely. The present assumption of the transverse EM wave in vacuum contradicts the entire theory of gyroscopes! [note 10] There are many other errors in classical EM that we will point out in future articles; this should give the reader a flavor of what is in store for him in future articles.

So there are no truly satisfactory “experts” in any of the 14 fields. Yet there is still something to be said for having some years of experience in one or more of the areas. I have had the good fortune or misfortune, depending upon one’s viewpoint, to be involved in various of those 14 activities for some 30 years or more, and am active in several of the areas now. My colleagues and I are particularly active in researching overunity electrical devices (see Figure 8), to include the effects of controlled chaotic oscillators, charge-blocking oscillators, and oscillators driven by in-shuttled (bridged) Poynting field energy density flow ExH. Parametric oscillation is known to have yielded overunity electrical machines and was fully reported by Russian researchers in the Russian and French technical journals in the mid 1930s. [note 11] We will further cite that little-known work in future articles. Also, we are working mightily upon true negative resistors (Figure 9), wherein the resistor scavenges some of the disorganized energy from the vacuum’s energetic exchange with the charges in the system, organizes and collects it, and then outputs that collected energy flow into the external circuit. Such a resistor acts as a power source. Kron, perhaps the greatest of the electrical machinery scientists, is known to have produced such true negative resistors.

I have also proposed a theoretical explanation of the Priore device (Figure 10) in France which in laboratory animals demonstrated remarkable cures of terminal tumors, cured trypanosomias, reversed atheriosclerosis, and restored suppressed immune systems. The Priore work was not anecdotal, but was performed under rigorous scientific protocols, working under the auspices of renowned French scientists such as Courrier and Pautrizel. The work is properly reported in numerous papers in the French medical literature. [note 12] The results were certainly achieved, but neither Priore himself nor the scientists of the day could understand the mechanism involved. The deciphering of that causative technical mechanism utilized by the Priore group required 10 years of intense effort. The results have now been published in the journal Explore! and are there for anyone to read. [note 13]
Presently my CTEC colleagues and I have filed three patent applications and a continuation, on various methods for overunity processes and apparatuses, room temperature superconductivity, Poynting generators, etc.

As an example, we took exactly the opposite approach to room temperature superconductivity as has been taken by all the orthodox researchers to date. What exactly is sought, after all, for superconductivity (SC) in a section of an electrical circuit? Well, one has some electrons on one side of that SC section that are transporting excess potential, as given by the Slepian vector J*. [note 14] What is desired is to get some electrons on the other side of the SC section that also are transporting the same amount of excess energy in the form of J*. You can do that in two ways: (1) you can flow the electron carriers through the SC section, carrying their J*, or (2) you can block the J and flow the * itself across as the Poynting flow S. [note 15] Nondivergent Poynting flow flows along an equipotential, which is

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

fanblades to a draw) and massive correlation of the electrons and of the electron-to-lattice interactions is necessary if one is to get the electrons through there without excess collisions that shake off some of the excess * from J*, as scattered photons (heat) or as a change of form of the energy (as in straining the dielectric of a capacitor to convert electrical energy to mechanical strain energy via the piezoelectric effect.)

Our approach is exactly the opposite (Figure 11) and Figure 12). Why not just *stop* the flow of excited electron carriers on one side of the SC section, and continue the nondiverging flow of the Poynting field energy density S across that section at room temperature? Then the cryogenics is not needed at all. After all, circuits already work that way anyhow __ except standard practice is to nullify the process by letting current be driven around the sourcing current loop and back through the back EMF of the primary source. A quantum well (or several other methods) can be used to trap the “sourcing” electrons in the conductor just prior to entering the SC section. The “receiving” electrons on the other side of the SC section, however, must be in their own dq/dt-isolated current loop.

Normal sources do not furnish electrons and current to a circuit anyway. Sources just furnish Poynting flow S and EMF. [note 16] In the receiving current loop, the EMF appears automatically once the S-flow flows in and is “locked on to” by the receiving electrons. Given q**, you will immediately have J* if the electrons are free to move in the conducting circuit (see Figure 13). Further, you have eliminated all the loss terms from the standard Poynting equation for energy losses. So all the energy flow S just flows across the SC section, without any current or Cooper pairs flowing through that SC section. The SC section has become a “bridge” which (1) strips off the Poynting field energy density flow S from the electron current dq/dt on the sourcing side of the SC/bridge section, by simply reducing the dq/dt to zero; and (2) continues to pass the S-flow across the dq/dt-blocked SC/bridge section to the other side (Figure 14). The excess S-flow (and EMF) pours into the receiving dq/dt-isolated current loop, exciting the electrons therein to produce dq/dt and J*. *Any closed current loop is self-powered, once it receives S and EMF. It furnishes its own electrons; it only requires excess energy and EMF.* [note 17]

So our approach gives room temperature superconductivity in a very straightforward manner, once you discover how to block the current dq/dt in a conductor. Blocking it in an insulator is not sufficient, because that drops the potential and stops the S-flow and the equipotential * (the EMF). However, a degenerate semiconductor such as the Fogal chip can be used, as can several other processes for blocking dq/dt in a conductor. We will discuss these in a future article.

Another advantage of this approach to room temperature superconductivity is that now one can also have permissible overunity coefficient of performance. Now the load can be placed in its own S-receiving, isolated current loop. With the sourcing current loop furnishing only S and not dq/dt, the load is still powered normally in its own closed dq/dt current loop, but none of the load current is passed back through the back EMF of the primary source in the sourcing circuit.

This principle — that at least a substantial portion of the load current must not pass back through the primary source — is the primary principle required for a permissible overunity electrical machine (Figure 15). A permissible overunity electrical machine is one which produces more power in the load than you have to put into the machine to run it.

Obviously, the laws of physics and thermodynamics require such an overunity machine to be an open system that extracts and utilizes excess energy from an external source. In the case outlined, we extract the excess energy directly from the vacuum in the form of S-flow; essentially we convert our primary source to a Poynting generator rather than a “power” (energy flow loss rate) generator. The resulting9. overunity machine accomplishes both room temperature superconductivity and also overunity coefficient of performance (Figure 16). It does not violate the

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

exchange between the vacuum and each end of the dipole. [note 19] [note 20] When we attach "conducting leads" to a dipolar source, as in the simple two-wire circuit, all we are doing with the conductors is providing waveguides for the extracted S-flows from the vacuum to flow along (it flows outside the conductors) without loss. In effect, each lead is simply extending the pole (i.e., the terminal) to which it is connected.

My associates and I are also working on several other inventions at present, including two more patent applications to be filed in the near future. One of those is in the field of overunity power systems, while the other is in an entirely different, highly proprietary nuclear effort. I will address all these areas in future columns, and those columns will have real “meat” in them.

Present “big science” proclaims that overunity electrical machines are impossible. The reader should understand that the U.S. Patent and Trademarks Office is never going to grant a patent for an overunity electrical machine unless the machine is independently tested and replicated, and unless it is certified by scientists of impeccable credentials. This is as it should be. Legitimate free energy researchers accept that dictum. Eventually the stuff has to work on the bench in order to be validated. On the other hand, free energy researchers also point out that one must not demand of us the next 100 years of national research by a hundred major laboratories immediately! This is still a shoestring operation, because all conventional sources of research funds are denied to the serious overunity researcher (with the single exception of occasional personal funding of a researcher by a private individual). In other words, we demand at least the same kind of consideration that the orthodox nuclear fusion community has enjoyed. Billions of dollars have been poured into that sinkhole, and not a single controlled fusion power unit is working anywhere in the world after decades of work at enormous expense to the taxpayers.

So we overunity researchers deserve at least a decade or two of our own, to try to accomplish overunity electrical system development. We also painfully point out that “Big Science” will not fund a single overunity researcher, for specific overunity purpose, no matter what his credentials or capability. In fact, “Big Science” is part of the problem, because it fervently seeks to prevent any legitimate research in overunity. To carry out its suppression campaign. it resorts to vilification, character assassination, libel, and slander of the overunity researchers themselves __ again, regardless of their qualifications or legitimate scientific procedures. [note 21]

Indeed, orthodoxy has a black record with respect to energy. When Mayer advanced the conservation of energy law, orthodoxy hounded him unmercifully. He was universally ridiculed as the very epitome of a fool. He lost his job and suffered a nervous breakdown. Then years later toward the end of his life, scientists came to accept the conservation of energy as a most useful tool that dramatically simplified much of their analyses. So then science began to laud Mayer. There are a thousand other such examples; suffice it to say that “big science” is a bureaucracy and a “bell-shaped distribution curve” just like any other large group of people. Some scientists are near-angels. Others are near-devils. The vast majority are neither, but are just ordinary persons doing a special kind of job. In the scientific bureaucracy, however, the manipulative scientists scurry, slash, and manipulate their way toward the top. So the leadership of any big bureaucracy __ including the scientific bureaucracy __ is always rife with such near-scalawags and power-hungry individuals. Scientists are not exceptions to the same human weaknesses that we all evidence. If you really want to see jealousy, backbiting, and fierce “back-room dealing” and power-struggling, just get into a university research environment! But because the “power merchants” of any bureaucracy always fiercely resist any challenge to their position or superiority, big science has always fiercely resisted anything truly revolutionary. And today they also enlist the power of the state to enforce their dictums. We will devote one of our future columns to pointing out some glaring examples of persecution of independent researchers by big science. It is an area that all unorthodox researchers should be well-aware of. The world of science does not run by sweet reason and lofty

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

herring that dogmatic scientists use to vilify free energy researchers. It is a lie that they have also successfully sold to the U.S. Patent and Trademarks Office. Anyone with a modicum of knowledge of modern physics already knows that there does not exist a single closed system in the universe, anywhere. In particular, every physical system is continuously “open” to the violent exchange of energy with the surrounding vacuum. It’s just that most systems are in local equilibrium in these flow exchanges, except for minuscule gating accomplished for such things as the Lamb shift, etc.

By analogy it’s something like this. The free energy researcher is standing by the banks of a mighty, rushing river with breathtaking falls, turbulent areas, etc. Big Science is begrudgingly admitting that, yes, there is a river there, but it is an insane delusion to think that you can tap that energy. The free energy researcher is saying that, well, if I can build a gating sluice a little distance upstream, perhaps I can divert a small portion of the river’s flow downstream to my waterwheel, and power my mill. The orthodox scientist then vilifies the free energy researcher for even having such a heretical thought. In fact, says the orthodox researcher, any fool can see that the laws of nature force the river’s flow to stay firmly in its banks, because it is a closed system, and Moses brought that law down off the mountain with him on a special stone tablet. Then he adds all the other smug remarks such as, “If it could be done, then we at MIT and Harvard and Caltech would already have done it!” So until the orthodox scientist develops a little less dogma and a little more common sense, and learns the difference between a scientific discussion and a dogfight, the free energy researcher can hardly communicate with him.

The goal of the legitimate free energy researcher is to find a way to break the local equilibrium of this universal energy exchange, and gate a tiny bit of it up into the external circuit as excess electrical energy (Figure 17). Then he must collect the excess energy, and shuttle it from the “gating and collecting circuit” over to a dq/dt-closed current loop containing the load. He must also prevent the load current, or most of it, from passing back through the primary source against its back EMF, because if it does that, he will destroy the separation of charges in the dipolar source, killing the dipole and extinguishing the “excess energy gating.” In other words he must find a way to process and transport the energy flow across dq/dt-blocked bridges between the source’s closed dq/dt-current loop and the load’s closed dq/dt-current loop. Ruthlessly it is an energy transport S-flow problem, not a J* energy transport problem. If he uses J* in his sourcing current loop, he automatically drives the current dq/dt back through the back EMF of his source and kills it. If he does not drive dq/dt back through the source, the source will furnish S-flow for an indefinite period! Every battery and generator we have ever built is already a “free energy” machine in that it already involves broken local symmetry in the vacuum energy exchange. We have just not realized how to use our power sources purely as Poynting S-flow sources.

So we must treat an electrical system as an open system with broken local symmetry (Table 3) in the vacuum flux exchange, so that the system extracts (gates) excess energy from an external source. In this case the source is the system’s flux exchange with the vacuum. This approach is no more mysterious or bizarre than putting a waterwheel in a river, or a windmill in a wind, or a bank of solar cells out in the sunshine. The universal “free source of energy” that the overunity researcher seeks to tap is the violent exchange of virtual photons between the electrical charges of the system and the surrounding quantum mechanical vacuum (Figure 18). In quantum physics this powerful, energetic exchange of the vacuum with electrical matter is now proven both experimentally and theoretically. It is already accepted in quantum physics that the vacuum is filled with electromagnetic energy. The researcher doesn’t have to prove it anymore; he just has to find out how to properly use it. The artesian well is already there; we just have to learn how to collect and use the flowing water without dynamiting the well!

The remaining objection orthodox scientists usually raise against the notion of extracting vacuum energy has been

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

density of the exchange causing the Lamb shift is greater than the surface energy density of the sun. Different physicists have made various theoretical estimates of the energy density of this seething vacuum. A conservative ballpark figure is something on the order of 1090 grams per cubic centimeter, expressed in mass units.

It has also been my good fortune to know and work closely with several other free energy researchers. I was a colleague of Floyd ("Sparky") Sweet (recently deceased) for about seven years; Sparky held a Master's degree in EE from MIT and had years of experience with General Electric and other firms. He was a highly skilled researcher who invented the solid state vacuum triode amplifier (Figure 19), a unit which produced 500 watts of output and a coefficient of performance of 1,500,000. I personally tested the machine, activated the magnets on one occasion, and understand how it all worked. [note 24] The paper [note 25] I co-authored in 1991 with Floyd is still the only paper in the orthodox engineering literature that reports (a) successful overunity functioning of a solid state magnetic device, and (b) a successful antigravity experiment on the laboratory bench, reducing the weight of an object by 90%. [note 26]

We stress again that a legitimate overunity electrical machine must be an open system, receiving excess electrical energy from the surrounding vacuum. As such, it must operate in a higher topology than normal electrical machines (Figure 20).

My good friend Frank Golden has long been a close colleague. Years ago Frank developed a motor that produced overunity efficiency, but then to our astonishment we found (courtesy of Bill Tiller, [note 27] with whom we were working at the time) that its operation depended upon a 5-year-long conditioning (structuring) of the local ambient potential in order to work. Previously we had been entirely unaware that one can sometimes condition the local vacuum and the local matter in the area with a particular dynamic form. Given the local conditioning, a motor with that form will then work in that vicinity, but not necessarily elsewhere at an appreciable distance away. We now understand the mechanism for such "conditioning of the vacuum." We also now know what to do if an overunity machine is actually a machine of that type (many of them are not). In other words, we know how to convert it to a machine that will work anywhere, anytime. We will be addressing such little-known things in future articles.

John Bedini is also a close friend and colleague; John built several experimental motors (both electrical and magnetic) in the overunity area, and performed successful trans-mutation experiments. John is a recognized genius in high-end sound amplifier development. Many audiophiles worldwide still swear that the Bedini amplifier is the best and sweetest-sounding audio amplifier ever built. One of John's battery-powered electrical motors (Figure 21), for example, ran continuously off its battery for about five years, and kept the battery charged. When you realize that such a small electric motor is only about 35% efficient, then you realize that about 65% of the energy flowing out of the battery was being dissipated in the motor as heat, core losses, etc. So the unit was continuously performing work for that five years. [note 28] The 1/8 hp motor represented a load in which the continuous rate of work being done (the rate of energy dissipation) was about 0.08 hp. We will have more to say on this motor and its technical process in a future article.

Bill Fogal is a close colleague; Bill has patented the world's first dq/dt-blocking semiconductor, [note 29] which partially blocks the normal current while continuing to pass the flow of voltage (Table 4). I have written a proprietary technical explanation of the Fogal semiconductor, which utilizes an extremely little-known feature called the "overpotential." The overpotential part of the theory is experimentally proven and theoretically explained in an obscure part of electrochemistry, by a few chemists who specialize in electrode effects. J. O'M. Bockris is probably the leading world expert in overpotential theory; he authored the textbook on it. [note 30]

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

building ground, and we needed to make “floating differential ground” measurements in order to settle things once and for all. McLain and Wooten have filed for a patent, and multiple tests with excellent equipment have indeed indicated overunity. What happens next remains to be seen. I suspect that the two inventors may simply accept foreign capital investment and go offshore.

Howard Johnson is also a respected colleague, whom I very much admire. (See Figure 23). Howard has continued to work quietly and patiently upon his patented permanent magnet motor, [note 31] including patenting various magnetic gates, etc. that are necessary to make such a motor work in a rotary configuration. [note 32] Howard employs a two-particle theory of magnetism; i.e., each magnetic flux line is envisioned as having a particle traveling from the north pole to the south pole, and also a particle traveling from the south pole to the north pole. The particles are spinning; the forward-time particle spins in one direction, and the antiparticle spins in the other direction. Howard then slightly separates the two particle flows. [note 33] In other words, Johnson splits the flux lines themselves, into two different pieces. When so separated, the component lines are now curls, so their paths curve. The paths of the two “curl particles” are different; one curls in one direction and the other curls in the opposite direction. Further, a predominance of one form of curl particle gives a “time-forward” aspect, while a predominance of the other form of curl particle gives a “time-reversed” aspect. Johnson is thus able to employ a deeper kind of magnetism than the textbooks presently contain. He demonstrates that a “spin-altered” magnetic assembly exhibiting (to a compass or other such detector) a north polarity can attract another unaltered magnetic assembly exhibiting a north polarity. In short, he can make a north pole attract a north pole. We will give you further insight into Johnson’s two-particle theory in a future article. [note 34] We will also explain how and why the physicists missed that antiparticle in the magnetic field’s flux lines, and thereby failed to advance the theory of magnetism to a deeper level. Make no mistake, one day when the new theory is done, Johnson may well be awarded a Nobel Prize for his epochal discovery of a deeper structure of magnetism.

Bits and pieces of the new science approach are just now beginning to spill into a few conventional journals and symposia. For some time a few of the rebels spearheading this new science have been doing something that the University scientists should have been doing all along. The rebels have been meticulously examining the concepts, postulates, and definitions that the present scientific models are founded upon, to reveal serious foundations errors. Corrections for some of these errors have been discovered and made. [note 35] Others are still problems yet to be resolved. Yet slow progress is being made, and a “flip-over” of the old science is likely to occur within the next decade if not sooner. By then the new concepts will have become “solid” enough to allow producing an engineering mathematical model. Also the supporting experiments will have become solid enough to justify the new concepts and the model.

In this first article I wanted to introduce the 14 major areas. We will be discussing these areas in some technical detail in the future. We’ll also include current major news items of development in these areas, for the interested lay reader. I’ll particularly try to give you the benefit of extensive reference citations painfully gathered over the last 30 years. The articles will be targeted toward both the serious researcher and the interested, educated layman.

So that’s the agenda for the months ahead. We hope you find this information and the approach of direct interest and use.

Again, it’s a privilege to be aboard. And we invite you to hang on in the future; it’s going to be a wild but exciting ride. We really are going to chase this wild dragon with great vigor.

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

The Virtual Times, Huntsville Edition

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

Scalar Wave Technology

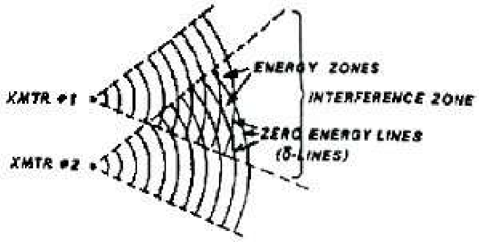

Scalar Wave Detection

Warren York
Scalar Wave Researcher

Originally published: **ExtraOrdinary Science (Volume 1, Issue 4; Oct/Nov/Dec 1989)**

WARNING: Do not attempt electrical construction on any of the equipment, coils or devices described here unless you are fully qualified. In some cases LETHAL voltages and charges are required. Also be advised that improper use, application and physical methods may cause psychological tramuma or permanent physical and neurological damage when near the vicinity of such equipment. This information is being provided as a point of interest to our readers... the publisher and author disclaim any responsibility in the use or misuse of this information.

Introduction

All of this talk so far on scalar waves and scalar technology has no purpose in the scientific community until something can be detected, observed and duplicated. The original seedlings and thoughts which led to Lt. Col. Bearden's scalar theories were developed out of observation. If not by Lt. Col. Bearden himself, then by others from many noncorrelated paths.

I already have given a small sample list of such observers from Nikola Tesla to the latest, Moray King. I have just read a newly released book by Moray entitled ***Tapping the Zero Point Energy***. I also recommend highly for others interested in scalar technology to obtain copies of Bearden's works and books from the new release of ***Excalibur Briefings*** to ***AIDS--Biological Warfare***. The two books and papers will fill in the much needed base work on understanding scalar technology and can be obtained through the High Energy Catalog.

I also understand that a new book is in the making which is a rebuttal entitled something along the lines of *Anti-Scalar*. All of this is healthy in any development process, for without pros and cons, the sorting of correct data to trash data cannot be accomplished. On the other hand, I would like to point out that any conclusions should be supported by actual tests and not just by armchair theory.

Anybody can debate from the armchair but to have any meaning to back up their statements, one must present tests where one gets his or her hands dirty. That is to say that until I had reactions and observed actual phenomena occurring, all the theories were no more than that--just theories. This is what I hope to accomplish here with these articles on scalar technology. I feel this is a true new technology and deserves some serious consideration.

If I had not gotten my hands dirty and followed through on others' experiments from other observers then I would never have been convinced that something extraordinary was taking place. Even though my senses told me I was wasting my time I set out on the monotonous task of duplicating others's work and notes of observation. This is what I hope some of you will do and start to help unfold and unleash this hidden knowledge called scalar -- zero point energy -- stress energy and even radionics. Itis true indeed that it has many names over the years, but now we can at least have hope of defining its true limits and abilities.

The detection methods I will present here are not confirmed by myself or corresponding researchers as of yet due to development is still in progress. In one method I can only accept the word of the researcher as to his results. Until I or others duplicate results it remains a possible process and approach in detection.

My own detection came in the form of two experiments which were reproductions of another author's work based on this technology. When I had reproducible results I knew there was more to scalar than just theory. The process is now underway in developing a reliable instrument or means of detection that others can use. Up to this point most of scalar observations have been uncertain and always with the person or operator being part of the circuit. Nobody could remove the person at this point to bring scalar under true observation conditions. Due to a simple observation in Mr. George Meek's work in electro-voice phenomena (EVP), where my first breakthrough was established, this now seems a possibility. The personal element was still included in the circuit but unlike radionics or Hieronymus devices, could be reproduced by anybody once special conditions were met. A co-researcher has hit upon a method which was initially observed by Nikola Tesla and shows promise of another breakthrough in removing the personal element.

I will now discuss my initial breakthrough to scalar detection which will help you follow my thinking on what is needed and just what we are looking for in scalar detection. Scalar properties are the dictating key elements in this approach. I will then describe a possible detection method of another researcher and finally what I feel will be a unique approach with an instrument currently under construction by Mr. Russell Clift, a co-researcher/inventor, and myself. Do not feel you have to adhere to these techniques for they are only a guideline and starting point to spring from in your own approach to this subject. They are not established yet and will most likely be altered many times before their completion.

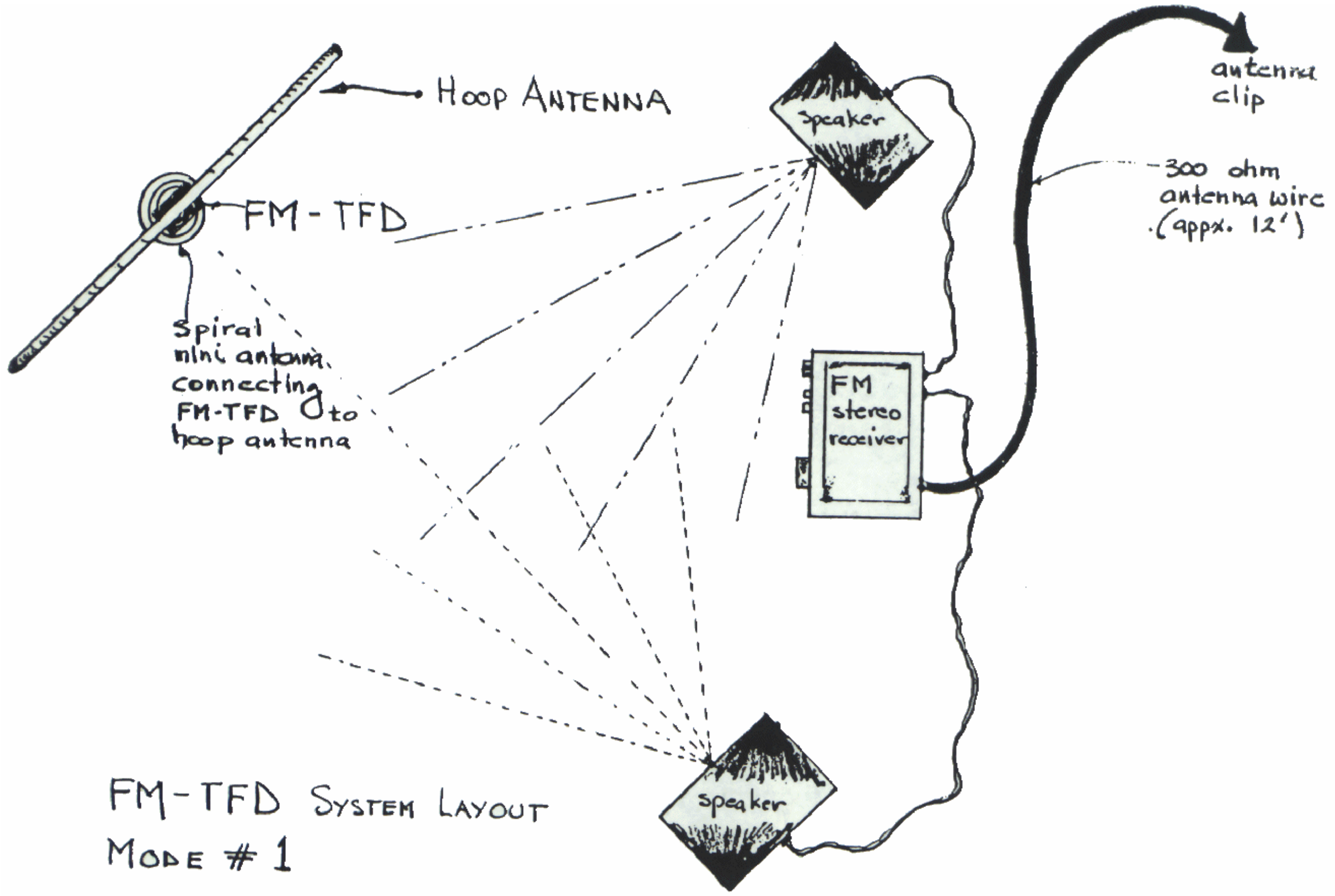

Figure 1. **System Layout for FM-TFD [Mode 1]**

As in all unknowns there are always unforeseen dangers involved and one should always use extreme caution and behave professionally in all approaches with any experimenting. Madame Curi and the XRay research is a good example. They paid dearly for their progress in X-Ray development with their lives. Scalar and its properties can be just as deadly as it is fascinating. I hope you read the seriousness in my tone for no matter how science fictionish scalar may sound to you now, I assure you the reality of actuallity can end up staring you in the face.

Breakthrough Event

Several years ago a Mr. George Meek presented me with construction planes and notes to a device he called Mark V. [Figure 1] Mr. Meek had been researching a phenomenon better known as EVP, or electro-voice phenomena. [Figure 2]

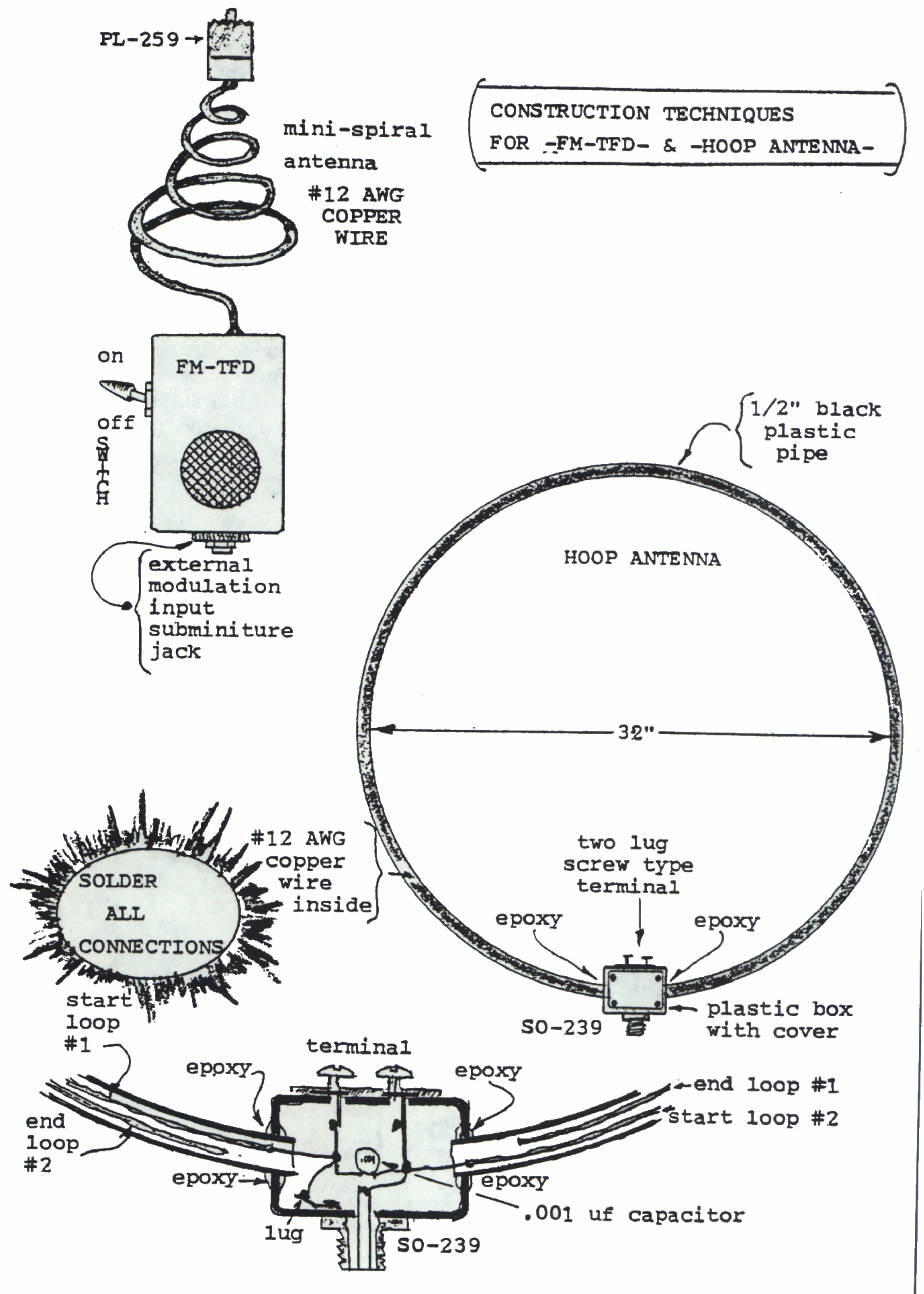
CONSTRUCTION TECHNIQUES
FOR -FM-TFD- & -HOOP ANTENNA-
PL-259
mini-spiral
antenna
#12 AWG
COPPER
WIRE
on
FM-TFD
off
SWITCH
external
modulation
input
subminiture
jack
1/2" black
plastic
pipe
HOOP ANTENNA
32"
#12 AWG
copper
wire
inside
SOLDER
ALL
CONNECTIONS
two lug
screw type
terminal
epoxy
epoxy
plastic box
with cover
SO-239
start
loop
#1
terminal
epoxy
epoxy
end loop #1
start loop #2
end
loop
#2
epoxy
epoxy
.001
.001 uf capacitor
lug
SO-239

Figure 2.

***Figure 2.* Construction Techniques for FM TFD & Hoop Antenna**

I would like to take a second here and answer a letter from one of **Extraordinary Science**'s subscribers printed in our last issue. Mr. Jeff King from New Zealand had written to the effect if perhaps I could cover what relationship scalar technology has in reference to EVP. [Figure 3] Well here it is, for this was the first reproducible effect dealing with scalar energy [interference pattems] [Figure 4] that I could verify really took place. My second breakthrough was also in the EVP area but I will not be able to go into it at this time due to available space to cover what must be covered now.

In short, Mr. Meek claimed to have a fantastic breakthrough of his own in the communications field of EVP. His claims were a recorded two-way conversation with another dimension. ***ZAP!*** Right here one would stop and feel that Mr. Meek and his researchers needed psychological help as I am sure some of you feel about me at this point. I contacted Mr. Meek and he flew down to El Paso so we could talk in person. He brought me two audio tapes of which one surely did have what appeared two-way communication in real time.

Up to this point all EVP claims were after the fact type of events. That is, the voices always appeared on blank tape after the operator had run the tape recorder for a period of time. No audible sounds were heard during the initial recording, only after playback of the tape--sort of the same thing being done now with channeling, but the person is still part of the circuit. Not all voices were understandable or clear enough to be real voice prints or just random white noise perhaps forming what almost sounds like words. In other cases, the words and even statements were definitely clear and of intelligent source. Well, Mr. Meek's tape went way beyond this to a two-way real time conversation.

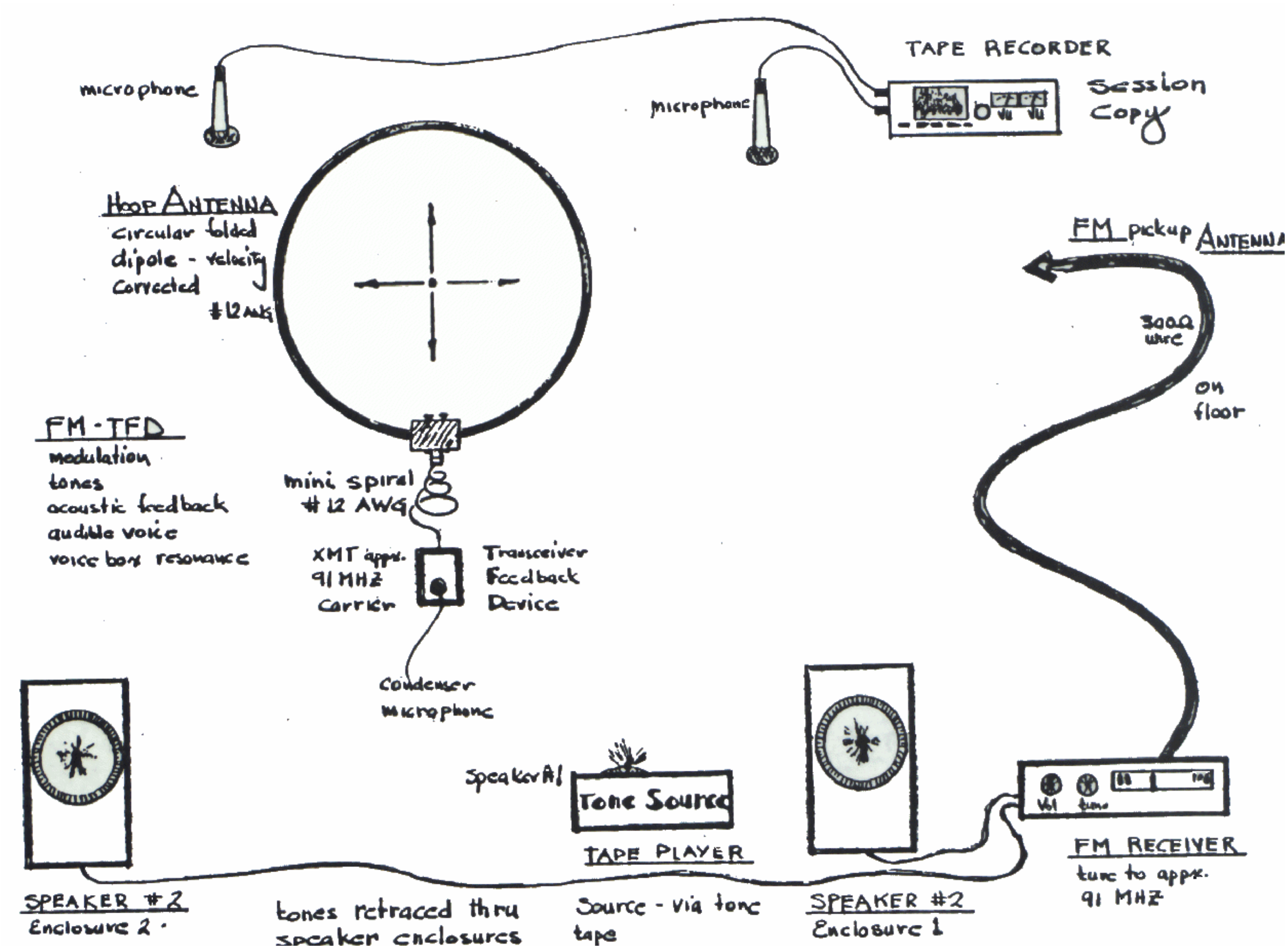
TAPE RECORDER
Session Copy
microphone
microphone
Hoop ANTENNA
circular folded dipole - velocity corrected
#12 AWG
FM pickup ANTENN
300Ω wire
on floor
FM-TFD
modulation
tones
acoustic feedback
audible voice
voice box resonance
mini spiral #12 AWG
XMT appx. 91 MHZ Carrier
Transceiver Feedback Device
Condenser microphone
Speaker #1
Tone Source
TAPE PLAYER
vol
tune
FM RECEIVER
tune to appx. 91 MHZ
SPEAKER #2
Enclosure 2
tones retraced thru speaker enclosures
Source - via tone tape
SPEAKER #2
Enclosure 1

Figure 3. **Mode 2 Communications Setup**

Fascinated by these claims I wanted to know more on how such a process worked. This is where he presented me with the second tape. It consisted of thirteen mixed electronic tones. No, there is no magical number in why thirteen tones were used only that they were needed to create interference patterns. These tones with an acoustical feedback and RF transmission with a Smith coil produced a unique interference pattern [scalar-RF and acoustical stress field].

I followed the notes and set out to reproduce this fantastic claim. At one point in the notes was one observation statement:

The system seems capable of feeding back changes in emotional states.

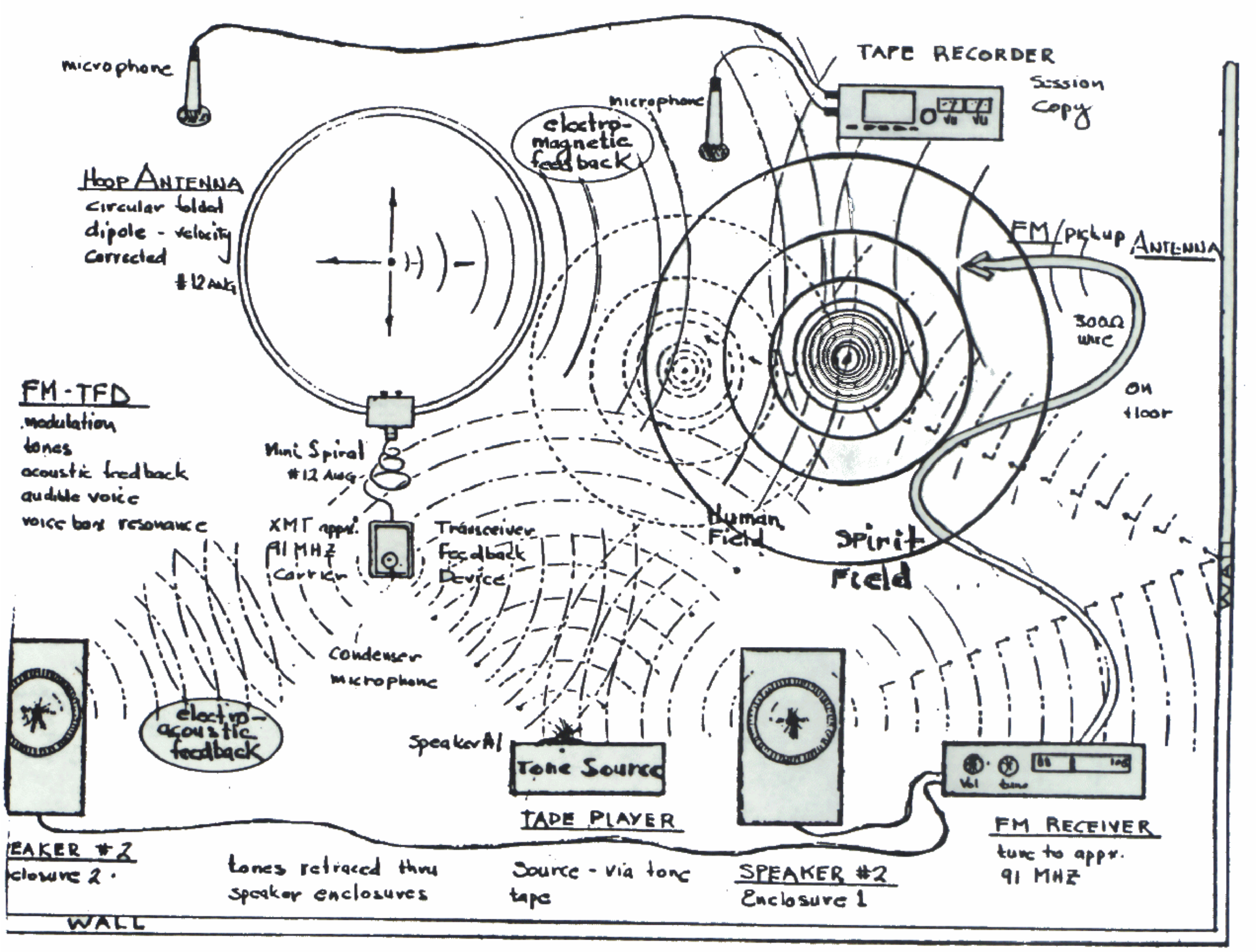
microphone
TAPE RECORDER
Session Copy
microphone
electro-magnetic feedback
HOOP ANTENNA
circular folded dipole - velocity corrected
#12 Awg
FM/pickup ANTENNA
300Ω wire
on floor
FM-TFD
.modulation
tones
acoustic feedback
audible voice
voice box resonance
Mini Spiral
#12 Awg
XMT appr.
91 MHZ
Carrier
Transceiver
Feedback
Device
Human
Field
Spirit
Field
Condenser
microphone
electro-acoustic feedback
Speaker #1
Tone Source
Vol
tune
TAPE PLAYER
EAKER #2
closure 2
tones retraced thru speaker enclosures
Source - via tone tape
SPEAKER #2
Enclosure 1
FM RECEIVER
tune to appr.
91 MHZ
WALL

Figure 4. **Mode 2 Communications Interference Pattern**

I stopped there and to my amazement found the observation to be true. I was able to control an avalanche effect in a balanced tone of a set of acoustical and RF frequencies with only my own concentration -- no doppler effect or movement or contacts to my body in any way. I taped this effect and it is available to those who wish to observe the level of control of an electrical circuit with shear stress energy from a human.

This observation is equal to Nikola Tesla's observation of his plasma globe and tensing a fist. He found the plasma trail would rotate in relation with the tension or muscle pressure produced. I will cover this process in a moment for the plasma approach in construction a scalar indicator seems to be ideal due to its extreme sensitivity that up until now has been the human or personal element.

Bedini Version ---- Dea/Faretto Scalar Detector

In this version of an electromagnetic scalar detector you will see that a coil is suspended in a magnetic field within a shielded case. [Figure 5] It is not indicated but I am sure the coil should be wound as a scalar coil. See "Construction of a Scalar Coil," **ExtraOrdinary Science**, Vol.I, Issue 3, page 16. The shielding is so all that known EMF fields will be run to ground and not affect the sensor within the case. As you know, one of the properties of scalar energy is that it cannot be shielded. All known EMF will be stopped from entering the case except gravity fields and scalar energy.

I am not saying that gravity fields or scalar fields are EMF, but only that the formula to each have a basic root in the EMF part of the spectrum. A generated scalar pulse or field will cut past the case shielding and affect the atoms of the scalar coil within the magnetic field which in tum will be coupled to the op amp. Any change or effect will then be displayed on the scope indicating a pulse or field of some kind has gotten past the system's shielding and to the sensor. You should see a change in the display around 6AM when the sun comes up and there is a slight gravity shift. [1,2]

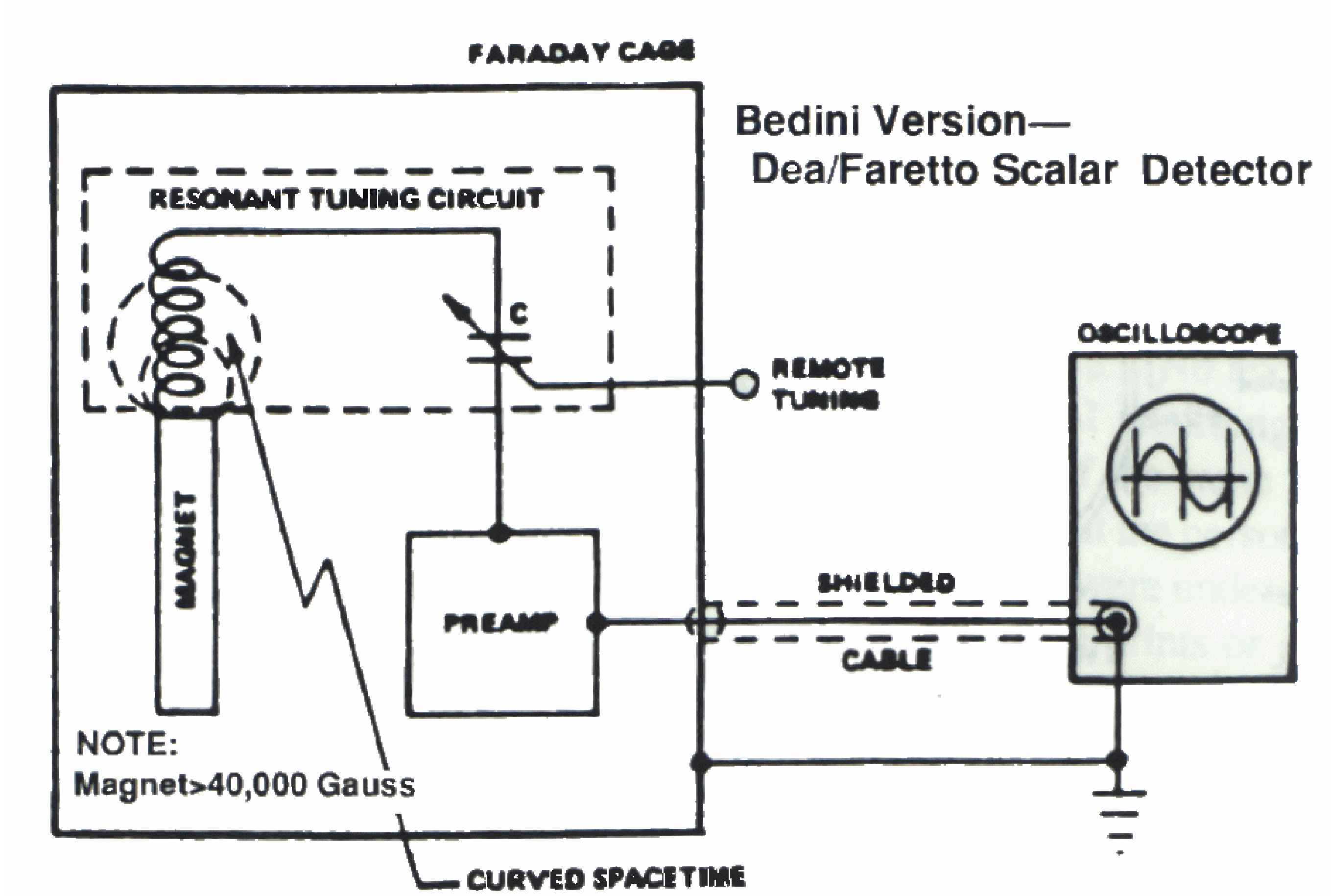
FARADAY CAGE
Bedini Version—
Dea/Faretto Scalar Detector
RESONANT TUNING CIRCUIT
C
REMOTE
TUNING
OSCILLOSCOPE
MAGNET
PREAMP
SHIELDED
CABLE
NOTE:
Magnet>40,000 Gauss
CURVED SPACETIME

Figure 5. **Bedini Version of Dea/Faretto Detector**

Plasma Scalar Detector

This is the method which is under construction by Russell Clift and myself. At the time of this printing the construction is not far enough to give any details or results as to this approach. [See Figure 6] This is based on the same construction of Nikola Tesla's experiment.[3] Figure 7 is for experiments with various voltages and ion currents. Figure 8 is a simple multiplier for the AC supply shown in Figure 6.

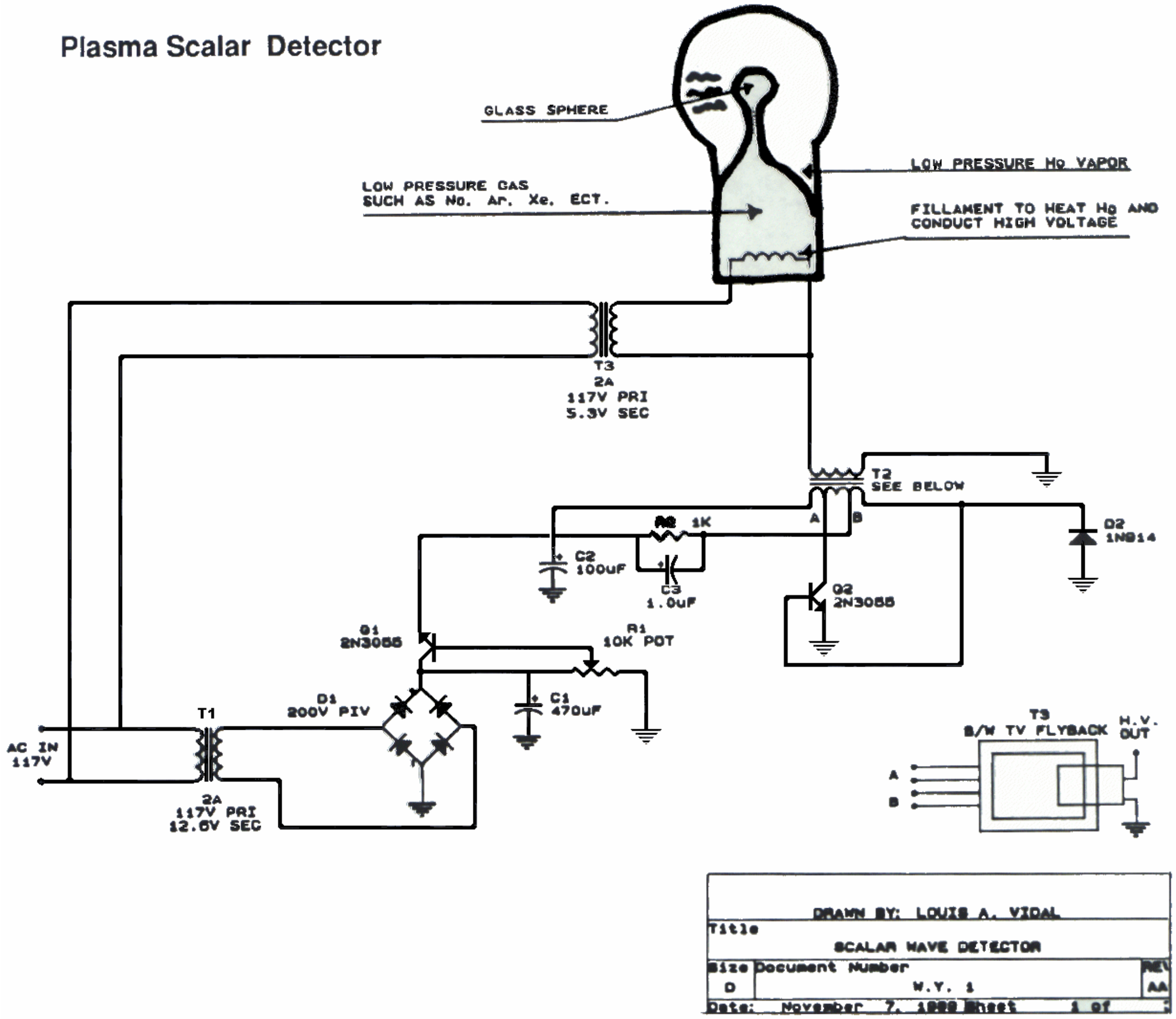
Plasma Scalar Detector
GLASS SPHERE
LOW PRESSURE Hg VAPOR
LOW PRESSURE GAS SUCH AS Ne. Ar. Xe. ECT.
FILLAMENT TO HEAT Hg AND CONDUCT HIGH VOLTAGE
T3 2A 117V PRI 5.3V SEC
T2 SEE BELOW
A
B
R2 1K
C2 100uF
C3 1.0uF
Q2 2N3055
D2 1N914
Q1 2N3055
R1 10K POT
D1 200V PIV
C1 470uF
AC IN 117V
T1 2A 117V PRI 12.6V SEC
T3 B/W TV FLYBACK
H.V. OUT
A
B
DRAWN BY: LOUIS A. VIDAL
Title SCALAR WAVE DETECTOR
Size D
Document Number W.Y. 1
REV AA
Date: November 7, 1999 Sheet 1 of

Figure 6. **Plasma Scalar Detector**

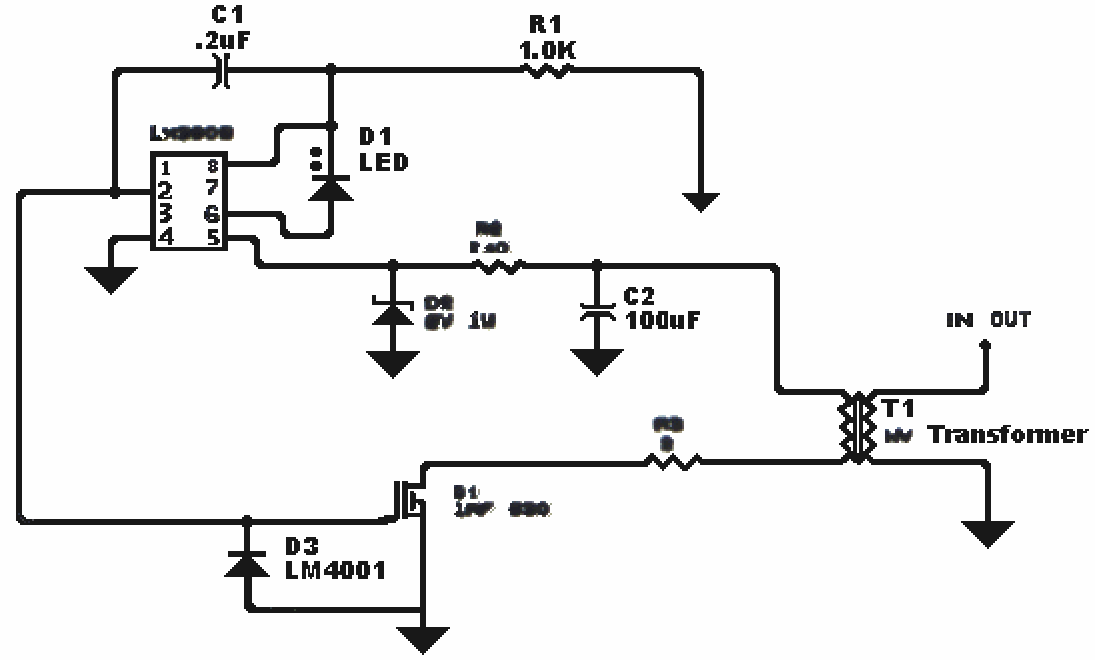
C1 .2uF
R1 1.0K
D1 LED
1
2
3
4
8
7
6
5
C2 100uF
IN OUT
T1
Transformer
D3 LM4001

Figure 7. **Power Supply**

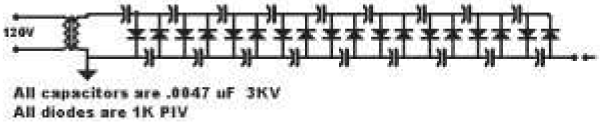

All capacitors are .0047 uF 3KV
All diodes are 1K PIV

Figure 8. **Voltage Multiplier**

I know I have just touched upon this subject so far but I hope I have given some of you starting points in your own research and findings. I will welcome any and all papers or works you may do in the scalar field and submit your work for publication here for others to read and gain from. Will be looking forward to your replies in the new year. Thank you for your interest.___***WRY***

References

1. Michael Polonyi, "Electrodynamics, Inertia and Gravitation: A Unifying Approach," **Speculations in Science and Technology**, Vol. 10, No. 2, page 145, 1987.

2. Gregory Hodowanec, "Gravitational Impulses," **Electronics Experimenters Handbook**, January 1989, page 114.

Thomas Cummerford Martin, **Inventions. Researches and Writings of Nikola Tesla**, page 226 to page 228.

HUNTSVILLE EDITION

The Southeast's Oldest E-Zine!

Chasing the Wild Dragon Tables and Figures

By Tom Bearden

Figures

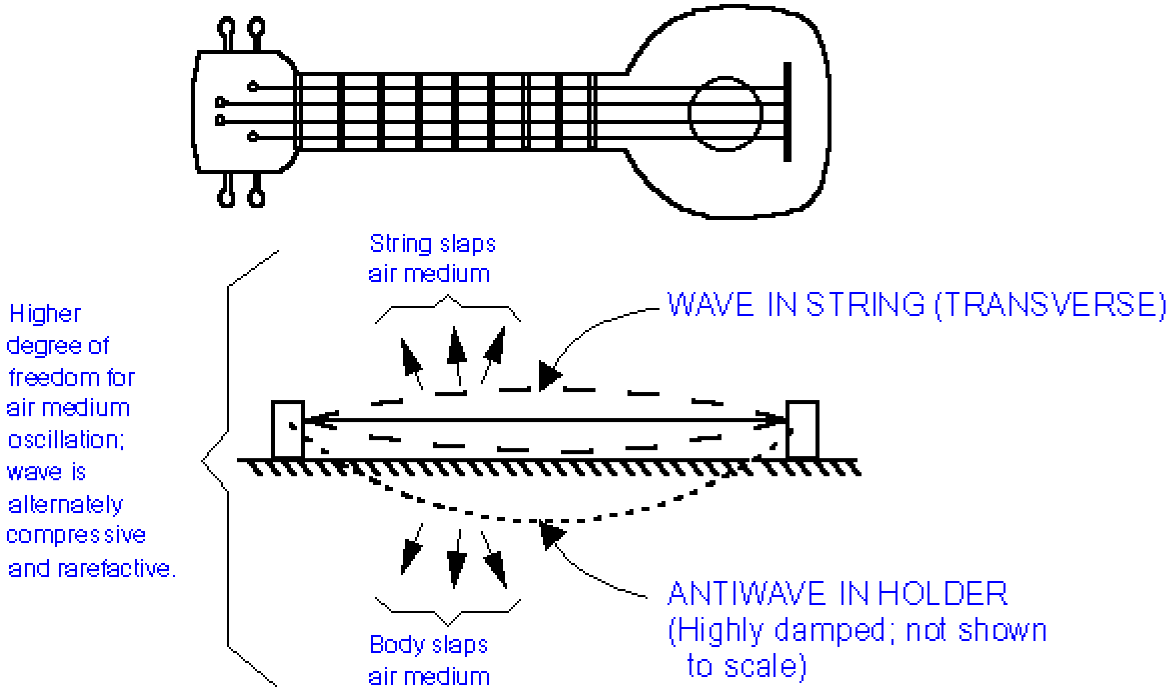
String slaps
air medium
WAVE IN STRING (TRANSVERSE)
Higher degree of freedom for air medium oscillation; wave is alternately compressive and rarefactive.
ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)
Body slaps
air medium

Figure 1. Air medium disturbances generated by plucked string.

Two waves are created, both slapping the air medium at the same time, but in opposite directions. A disturbed medium oscillates according to its own degrees of freedom. The air has more degrees of freedom than the string. A wave of compression and rarefaction is created in the air medium.

D-1-F 1

© 1993 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

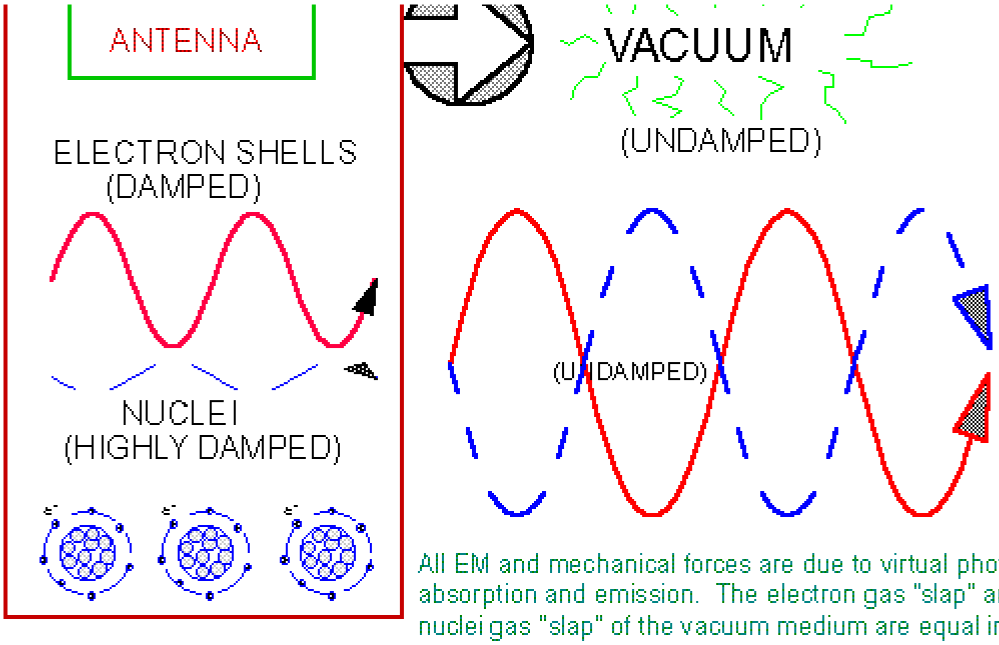
ANTENNA
ELECTRON SHELLS
(DAMPED)
NUCLEI
(HIGHLY DAMPED)
VACUUM
(UNDAMPED)
(UNDAMPED)

All EM and mechanical forces are due to virtual photon absorption and emission. The electron gas "slap" and the nuclei gas "slap" of the vacuum medium are equal in energy. Hence they inject equal numbers of virtual photons into the vacuum, creating equal and opposite vacuum disturbances. This gives a vacuum wave of alternating compression and rarefaction. This vacuum EM wave is similar to a sound wave, which is modeled longitudinally.

Figure 2. Launching an EM wave from a wire antenna.

D-1-F2

© T.E. BEARDEN 1994, 1995

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

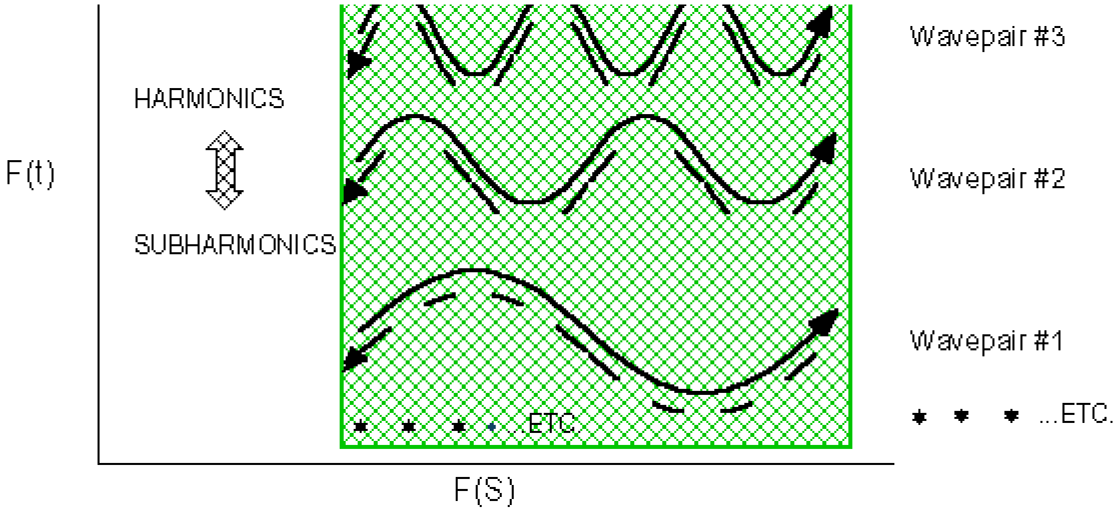

Figure 3. Internal Wave Structure of the scalar potential.

A harmonic set if wavepairs, where in each wavepair the two waves superpose spatially, but travel in opposite directions. The two waves in each pair are true phase conjugates and time-reversals of each other. Thus they comprise a coupled wave and antiwave. The photons must also be coupled into photon/antiphoton pairs, by a strong application of the distortion correction theorem of nonlinear optics. *Each wavepair is a standing electrogravitational wave.*

D-1-F3

© T.E. Bearden 1984, 1985

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

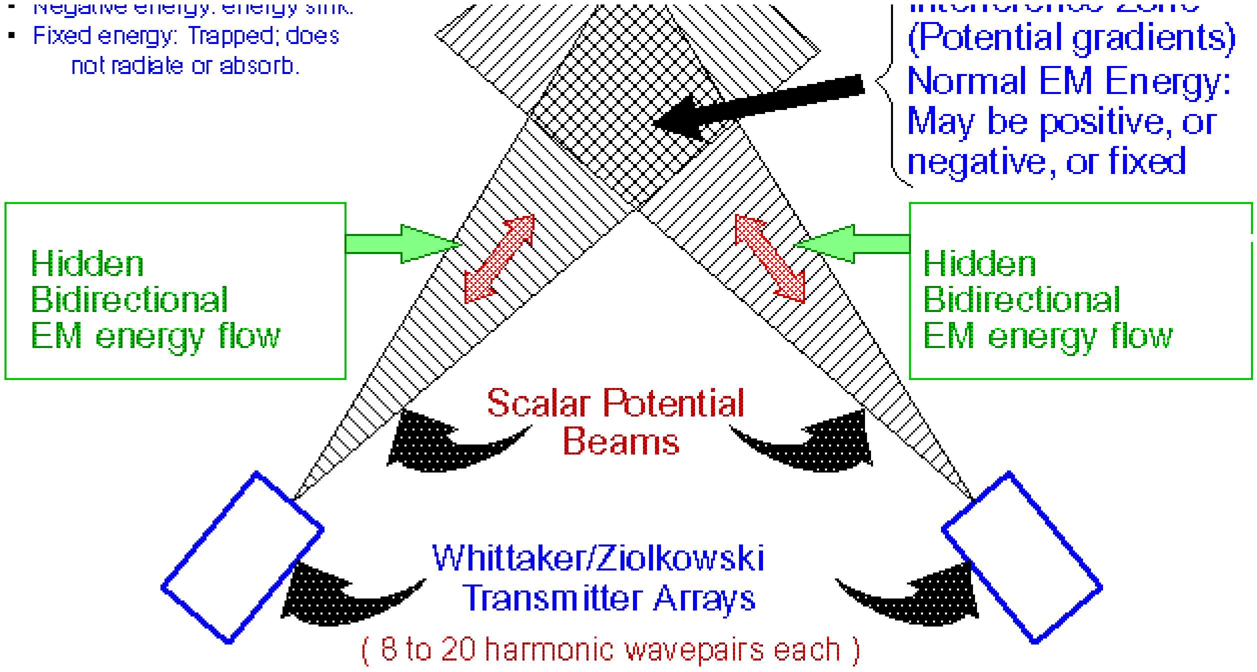
Negative energy: energy sink.
Fixed energy: Trapped; does not radiate or absorb.
Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or
negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

Figure 4. Scalar potential interferometry provides action-at-a-distance. It already creates normal EM fields and waves.

D-1-F4

© 1994 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

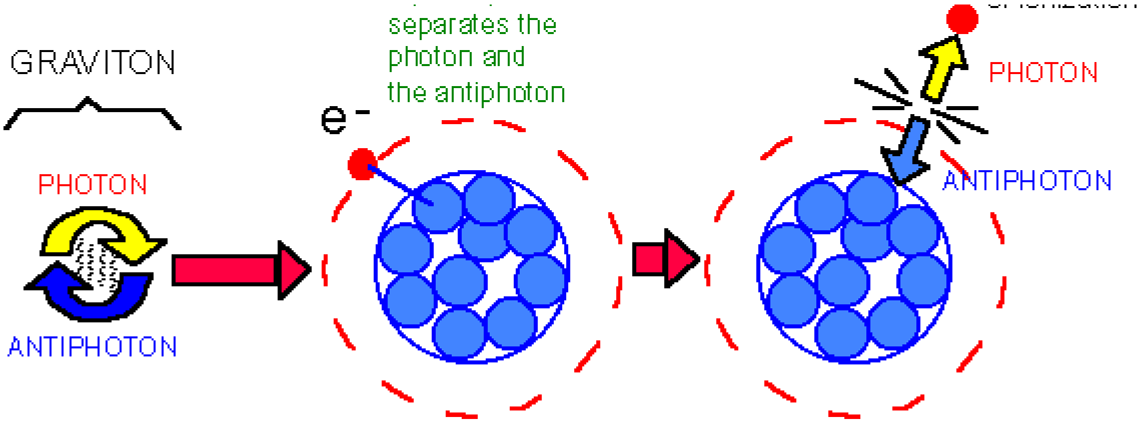
separates the
photon and
the antiphoton
GRAVITON
PHOTON
ANTIPHOTON
e -
PHOTON
ANTIPHOTON
NUCLEUS
RECOILS

NOTE: Graviton waves are bidirectional phase waves of photon/antiphoton coupling and uncoupling in the vacuum.

The atom is an assembly of dynamic dipoles.

A dipole splits the graviton into a photon and an antiphoton.
The photon interacts with the electron, producing the action.
The antiphoton interacts with the nucleus, producing the reaction.

Figure 5. Graviton interaction with matter produces the photon interaction and Newton's third law reaction.

D-1-FS

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

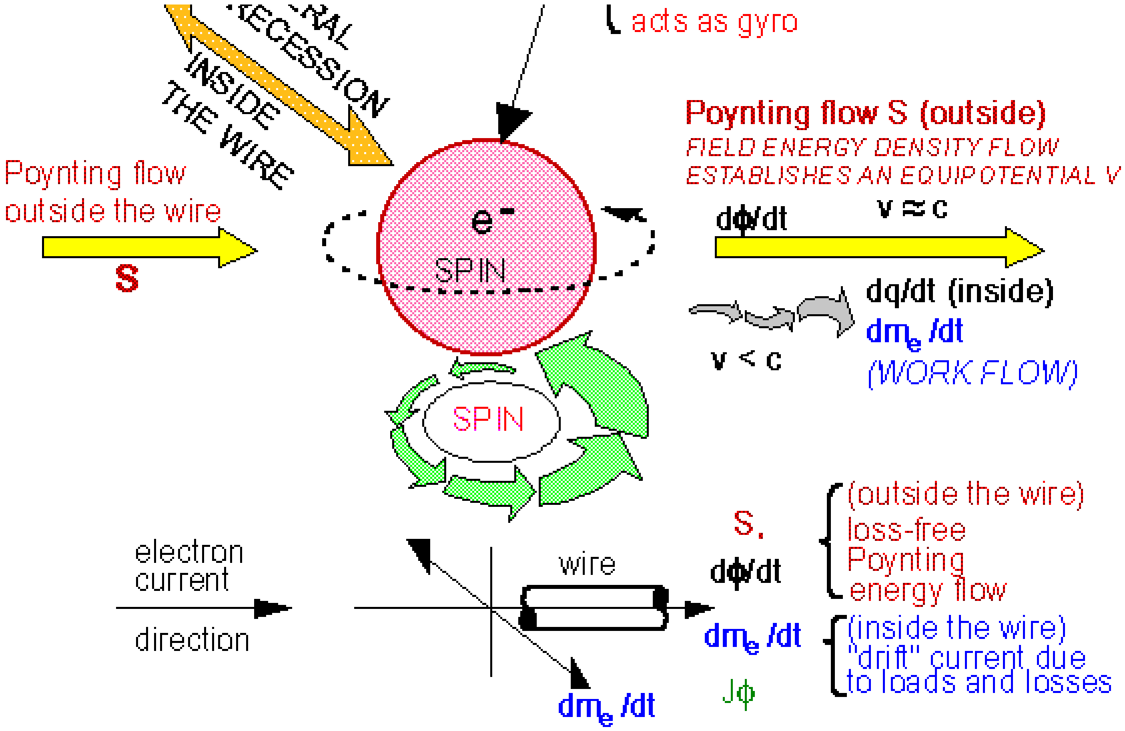
RAL
RECESSION
INSIDE
THE WIRE
acts as gyro
Poynting flow S (outside)
FIELD ENERGY DENSITY FLOW
ESTABLISHES AN EQUIPOTENTIAL V
Poynting flow
outside the wire
S
e⁻
SPIN
dφ/dt
v ≈ c
dq/dt (inside)
dm_e /dt
(WORK FLOW)
v < c
SPIN
S,
dφ/dt
(outside the wire)
loss-free
Poynting
energy flow
electron
current
direction
wire
dm_e /dt
(inside the wire)
"drift" current due
to loads and losses
dm_e /dt
Jφ

Figure 6. Currents in and along a conductor, with electron precession.

D-1-F6

© 1993 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

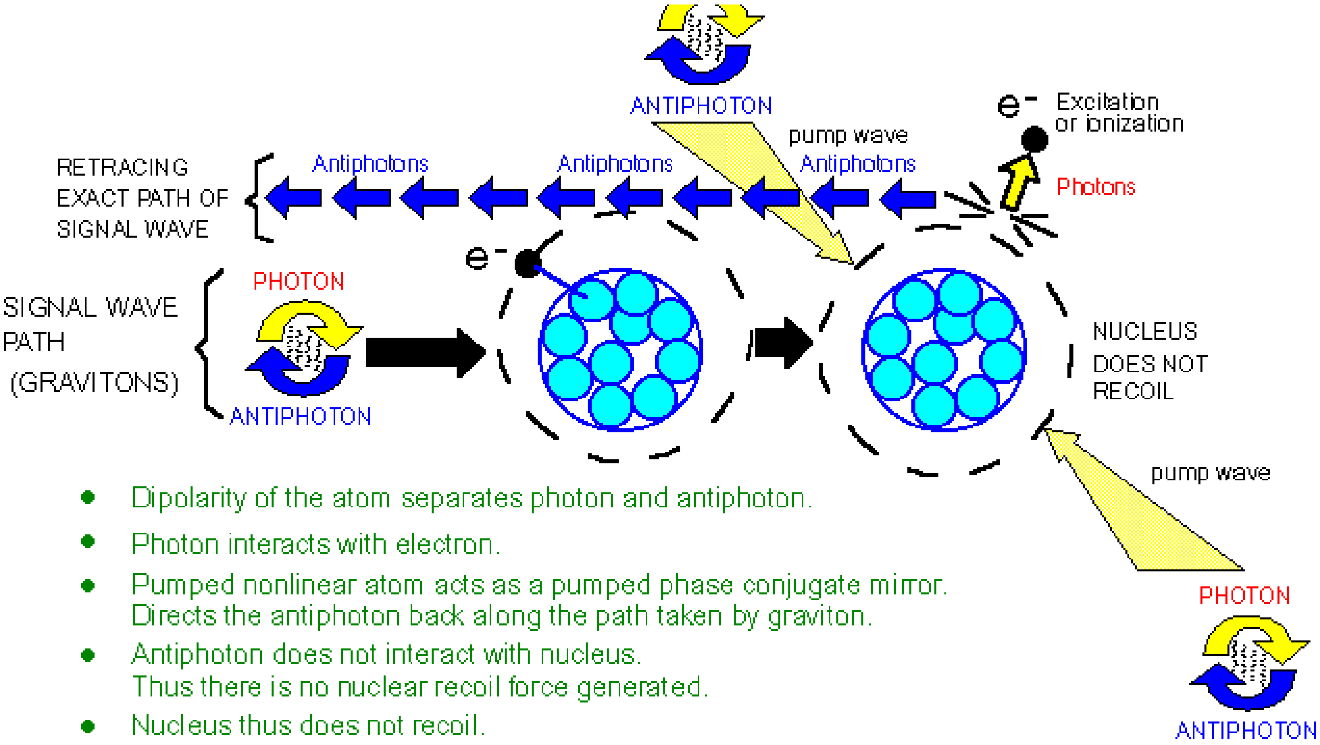
ANTIPHOTON
pump wave
e⁻
Excitation or ionization
RETRACING EXACT PATH OF SIGNAL WAVE
Antiphotons
Antiphotons
Antiphotons
Photons
e⁻
SIGNAL WAVE PATH (GRAVITONS)
PHOTON
ANTIPHOTON
NUCLEUS DOES NOT RECOIL
pump wave
PHOTON
ANTIPHOTON

- Dipolarity of the atom separates photon and antiphoton.
- Photon interacts with electron.
- Pumped nonlinear atom acts as a pumped phase conjugate mirror. Directs the antiphoton back along the path taken by graviton.
- Antiphoton does not interact with nucleus. Thus there is no nuclear recoil force generated.
- Nucleus thus does not recoil.
- Not a violation of Newton's third law, once one understands the mechanism that generates third law reaction. In the absence of the generating mechanism, there is no third law recoil. In EM field theory, the "missing antiwave" has necessitated third law omission.

Figure 7. Emission of antiphoton; nucleus does not recoil. Atom acts as a pumped phase conjugate mirror.

D-1-F7

© 1995 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

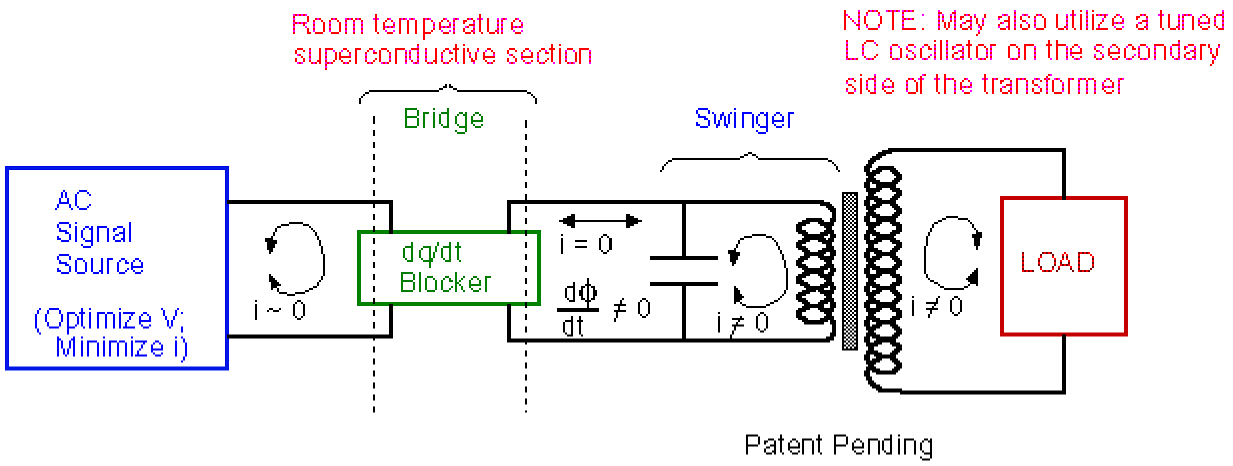
Room temperature
superconductive section
NOTE: May also utilize a tuned
LC oscillator on the secondary
side of the transformer
Bridge
Swinger
AC
Signal
Source
(Optimize V;
Minimize i)
i ~ 0
dq/dt
Blocker
i = 0
$\frac{d\phi}{dt} \neq 0$
i ≠ 0
i ≠ 0
LOAD
Patent Pending

Figure 8. dq/dt-free driving a swinger by S-flow.

Poynting energy flow S with its accompanying potential V is shuttled from the source to a tuned LC oscillator in resonance at the driving frequency or slightly below. The LC oscillator is driven by S-flow and dV/dt flow alone. The swinger furnishes its own free electrons, which receive S and EMF from the bridge, thereby producing dq/dt.

D-1-F8

© T.E. Bearden 1993, 1994, 1995

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

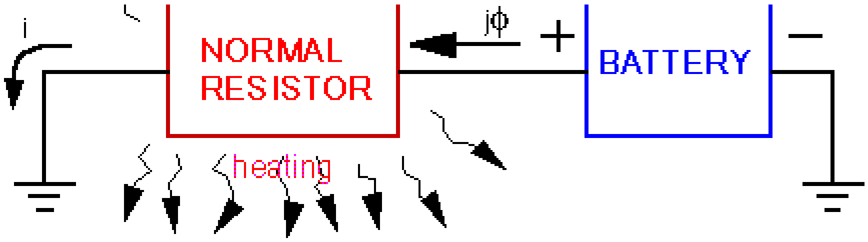

Figure 9a. Normal resistor receives ordered energy and scatters it.

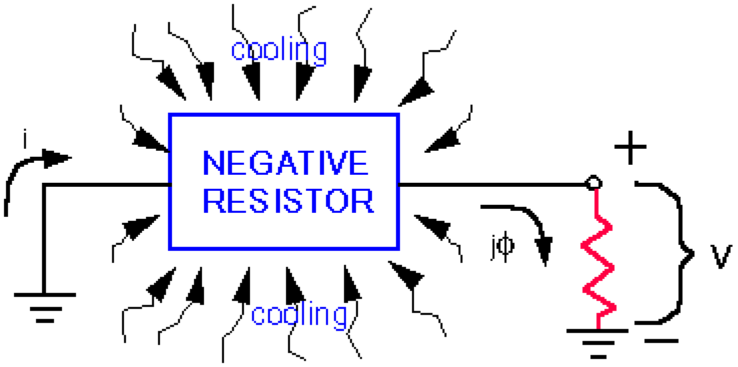

Figure 9b. Negative resistor receives, concentrates, and reorders scattered energy.

Figure 9. Normal resistor versus a negative resistor.

D-1-F9

© T.E. Bearden 1993, 1996

Some 17 signals were fed into a rotating plasma, which added phase conjugates to them. This infolded 17 wavepairs, each pair a coupled wave/antiwave, thereby forming a scalar potential with an artificial Whittaker pump wave internal structure. This infolded pump wave complex was modulated onto a powerful, rippling magnetic field, guaranteeing penetration even into the atomic nuclei. A specific, amplified antiengine for the exact cellular disease state was generated, reversing the diseased cells back to a previous healthy state.

Figure 10. Priore's laboratory apparatus for cellular reversal.

D-1-F10

© T.E. BEARDEN 1993, 1995

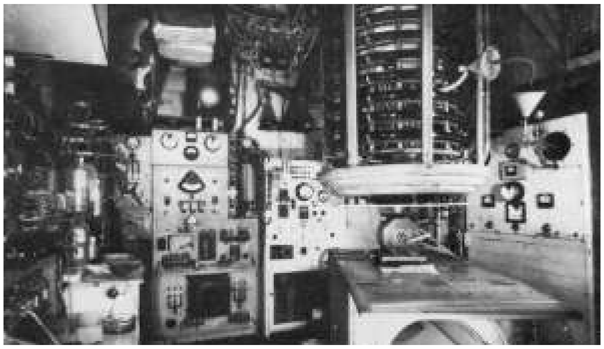

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

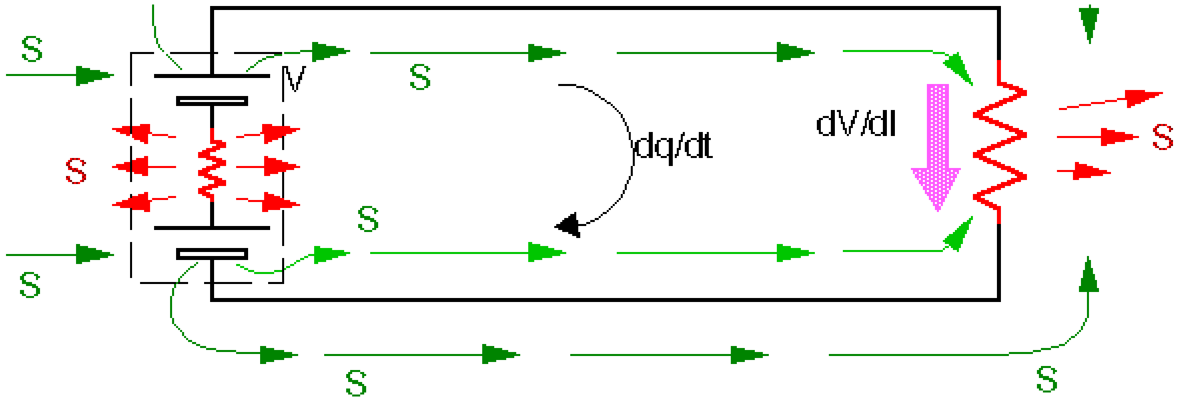

Note: EM texts err in showing the battery "gate" as the source of energy.

Figure 11. Poynting S-flow along a simple circuit.

From particle physics, any electric charge is a broken symmetry in the vacuum's virtual particle flux. Any dipole is thus two locally broken symmetries. Thus the dipole acts as a gate/source, driven by the fierce vacuum flux. So a dipole already "extracts" and gates EM energy flow, directly from the seething vacuum.

Every battery and generator is already a "free energy" source of extracted and gated EM energy from the vacuum.

D-1-F11

© T.E. Bearden 1994

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

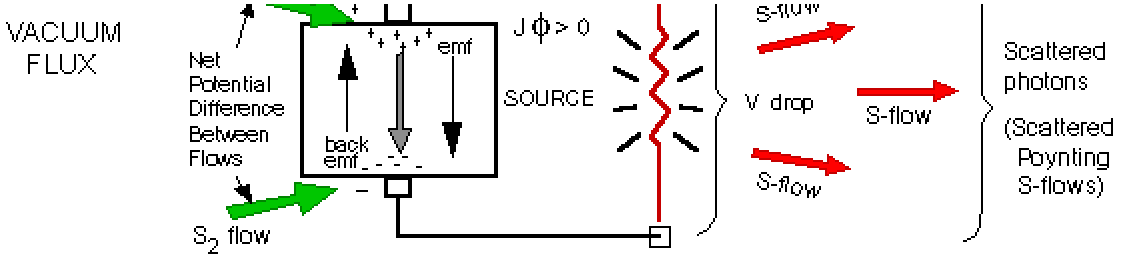
VACUUM FLUX
Net Potential Difference Between Flows
S_2 flow
+
emf
back emf
–
$J\phi > 0$
SOURCE
S-flow
V drop
S-flow
S-flow
Scattered photons
(Scattered Poynting S-flows)

Figure 12a. Normal current flow; $J\phi$ is the electron transport of energy, while S is the speed-of-light Poynting energy flow.

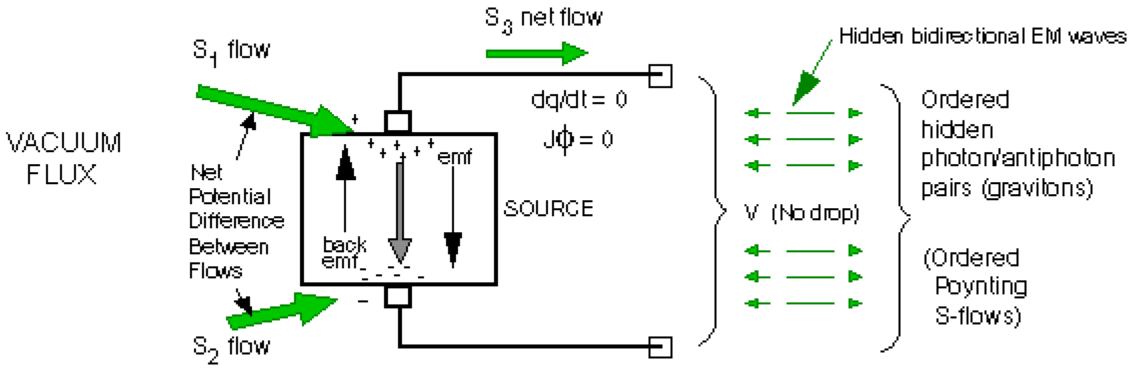
S_3 net flow
S_1 flow
Hidden bidirectional EM waves
VACUUM FLUX
Net Potential Difference Between Flows
dq/dt = 0
$J\phi = 0$
+
emf
back emf
–
SOURCE
V (No drop)
Ordered hidden photon/antiphoton pairs (gravitons)
(Ordered Poynting S-flows)
S_2 flow

Figure 12b. In a conductor with dq/dt blocked; S-flow and hidden Whittaker-wave flow continue.

Figure 12. Overt and covert field energy density flows.

D-1-F12

© 1994, 1995 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

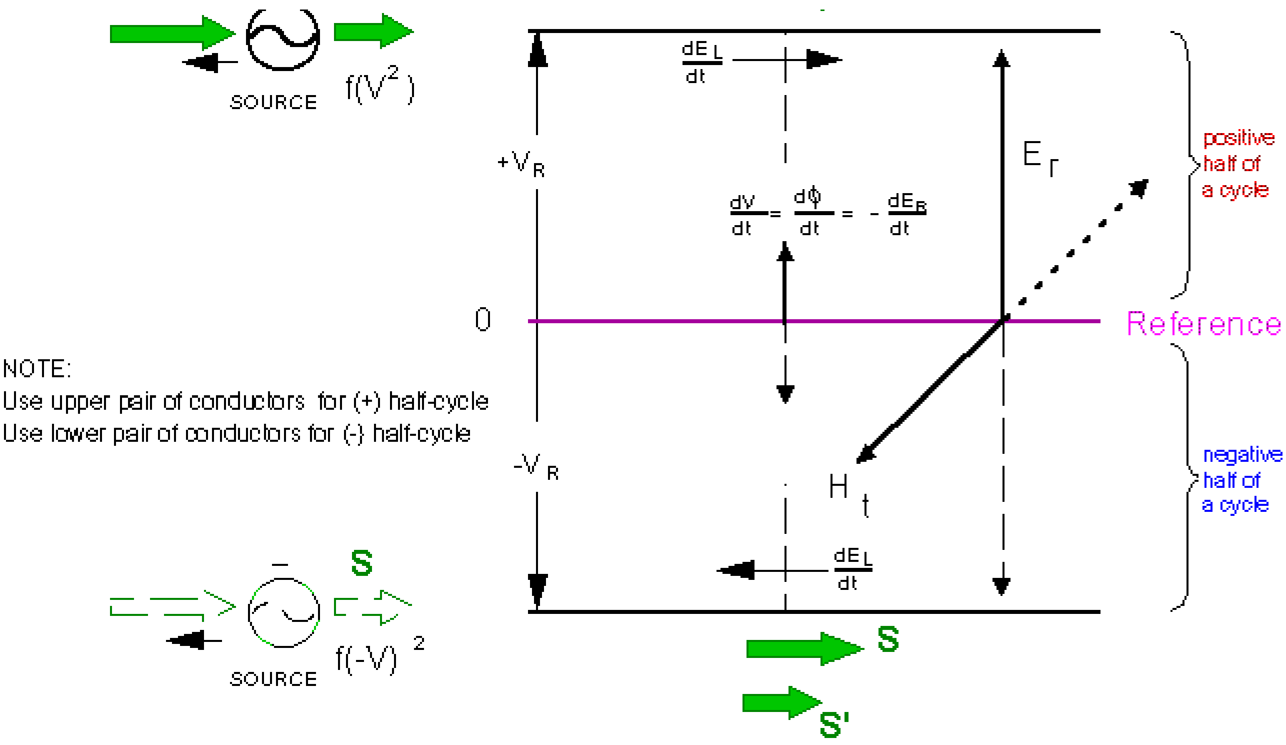
SOURCE
$f(V^2)$
$+V_R$
0
$-V_R$
$\frac{dE_L}{dt}$
$\frac{dV}{dt} = \frac{d\phi}{dt} = -\frac{dE_B}{dt}$
E_r
H_t
positive half of a cycle
Reference
negative half of a cycle
S
SOURCE
$f(-V)^2$
S
S'

NOTE:
Use upper pair of conductors for (+) half-cycle
Use lower pair of conductors for (-) half-cycle

Figure 13. Poynting vector currents S and S' versus massless displacement currents.

D-1-F13

© 1995 T.E. Bearden

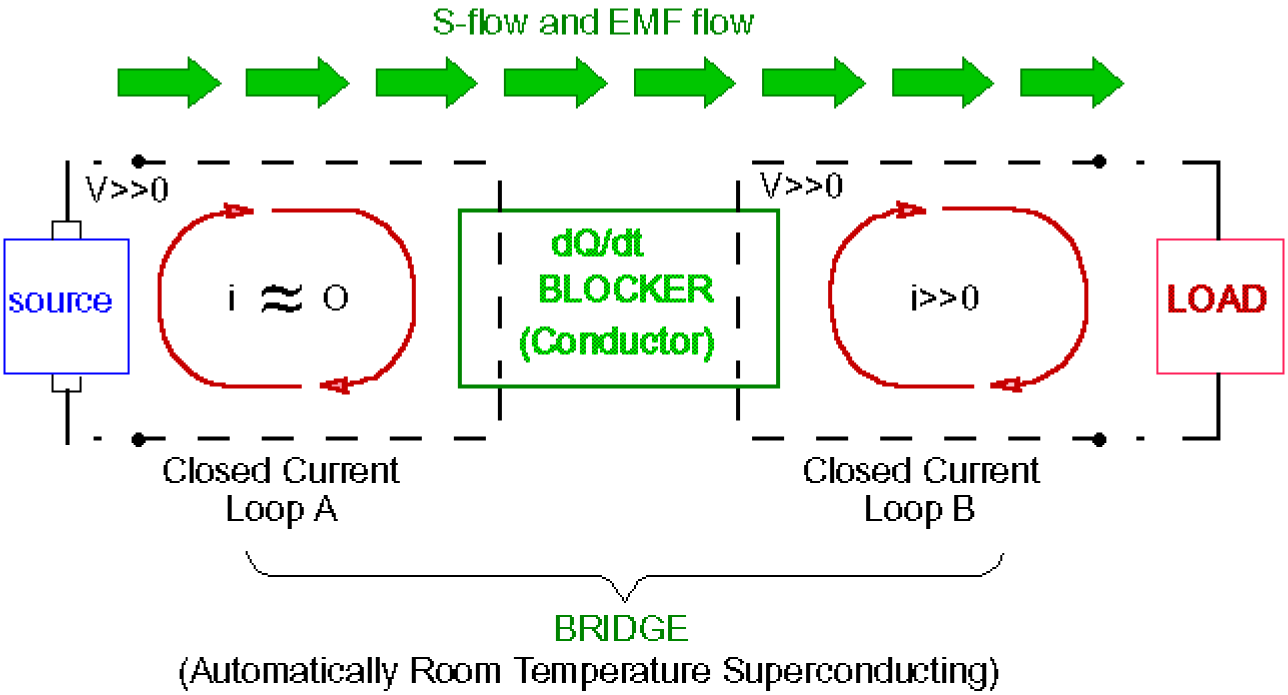
S-flow and EMF flow
V>>0
source
i ≈ 0
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD
Closed Current Loop A
Closed Current Loop B
BRIDGE
(Automatically Room Temperature Superconducting)

Figure 14. Bridging to enable room temperature superconductivity.

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

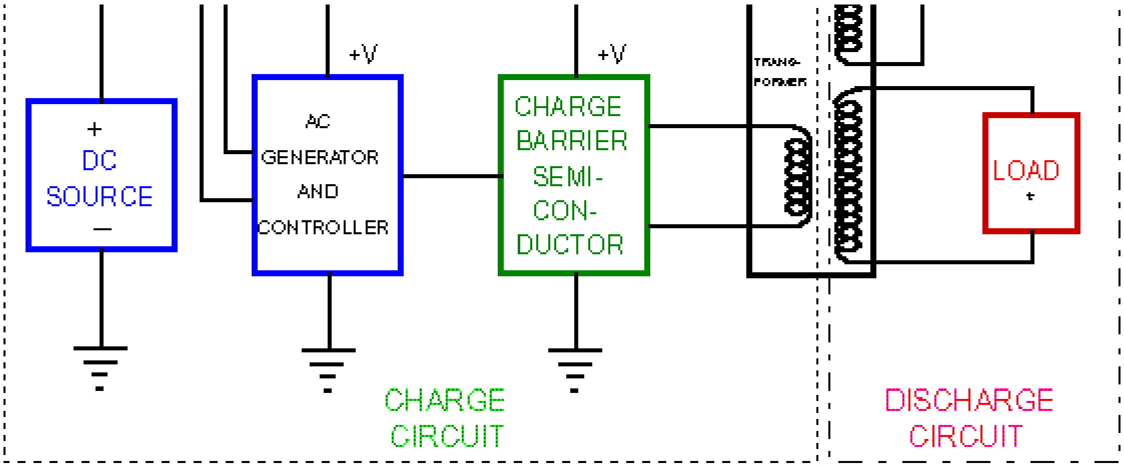
+V
+V
TRANS-
FORMER
+
DC
SOURCE
—
AC
GENERATOR
AND
CONTROLLER
CHARGE
BARRIER
SEMI-
CON-
DUCTOR
LOAD
+
CHARGE
CIRCUIT
DISCHARGE
CIRCUIT

Patent Pending

* NOTE:
DISCHARGE CIRCUIT AND LOAD GROUND
MUST BE ISOLATED FROM THE
CHARGE CIRCUIT GROUND.
FEEDBACK LOOP MUST BE RIGIDLY CLAMPED AND LIMITED.

Figure 15. Use of Fogal semiconductor for overunity (closed system).

D-1-F15

© 1993, 1995 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

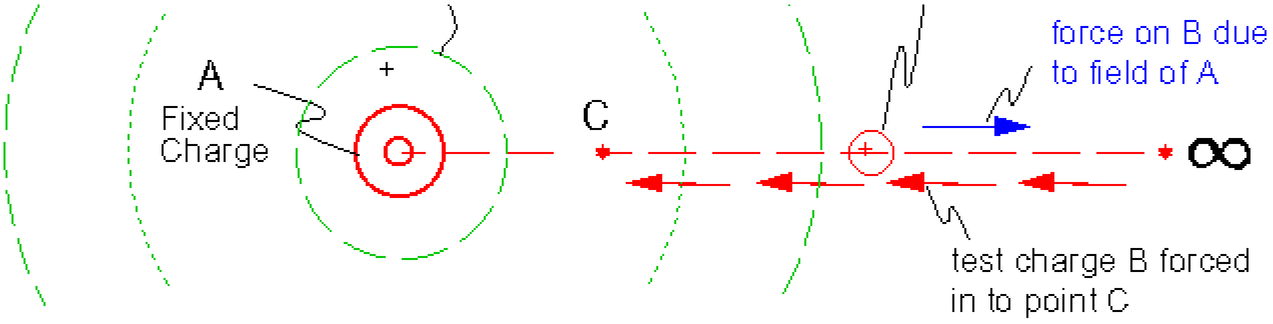
A
Fixed
Charge
+
C
force on B due
to field of A
∞
test charge B forced
in to point C

Figure 16a. Magnitude of the scalar potential is determined by work required to push in a unit charge from infinity; e.g., against a repelling like charge. If test charge is released, work will be performed upon it to translate it back toward infinity

Figure 16a. Process for overunity coefficient of performance.

D-1-F16A

ⓒ T.E. Bearden, 1993, 1994, 1995

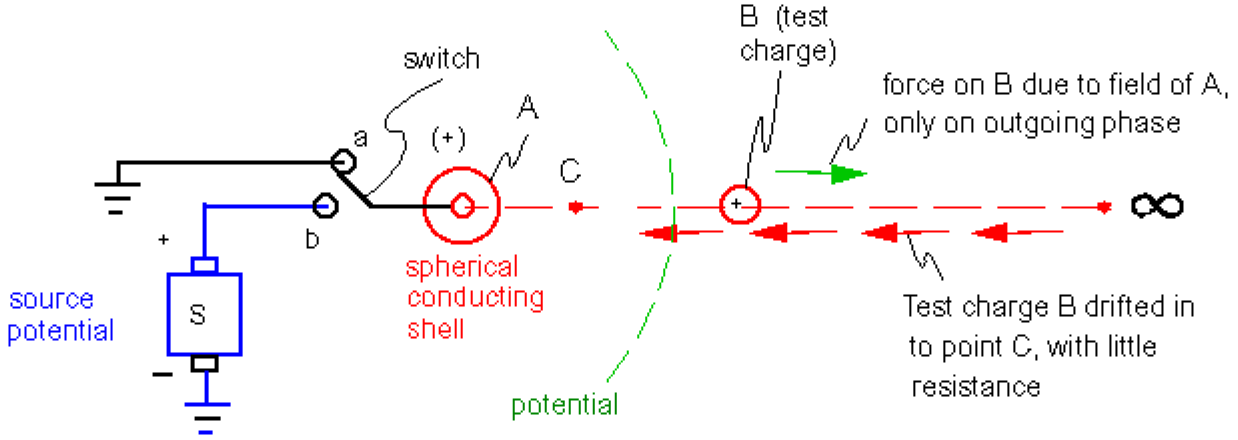
switch
a
(+)
A
C
B (test
charge)
force on B due to field of A,
only on outgoing phase
∞
+
b
S
−
source
potential
spherical
conducting
shell
potential
Test charge B drifted in
to point C, with little
resistance

Assumptions: Capacitance of A very small; charge B large, power-free switching

Switch potential off, drift in test charge B, switch potential back on. Can now get free work from repelled test charge B. Get much more free work if A is at a high potential, and charge B is large.

Figure 16 b. Demonstration of overunity efficiency

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies.
Ok

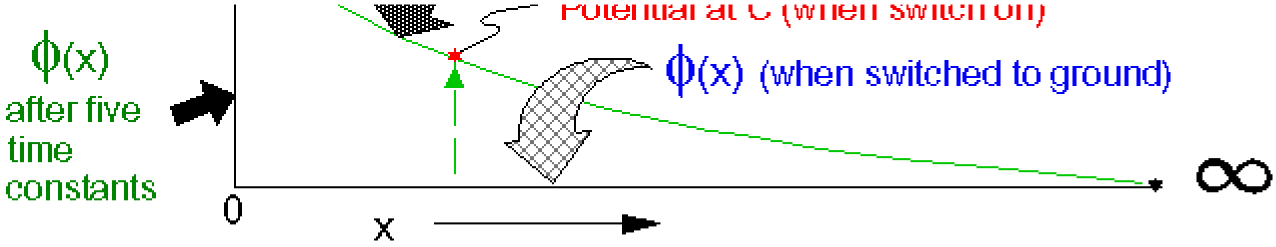

Ideally, switching the potential, rather than flowing charges, requires only minuscule overhead energy dissipation

Figure 16c. Potential to which test charge is exposed.

D-1-F16C

© T.E. Bearden, 1993, 1994, 1995

Open system receiving excess energy from an external source

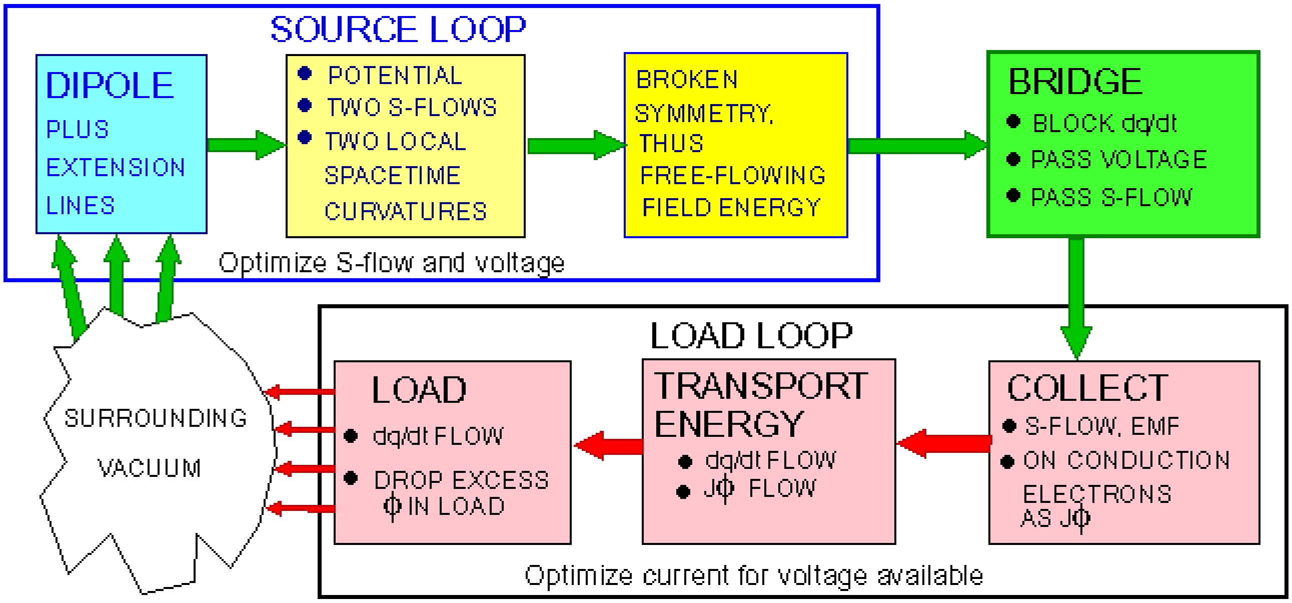
SOURCE LOOP
DIPOLE
PLUS
EXTENSION
LINES
• POTENTIAL
• TWO S-FLOWS
• TWO LOCAL SPACETIME CURVATURES
BROKEN SYMMETRY, THUS FREE-FLOWING FIELD ENERGY
Optimize S-flow and voltage
BRIDGE
• BLOCK dq/dt
• PASS VOLTAGE
• PASS S-FLOW
SURROUNDING
VACUUM
LOAD LOOP
LOAD
• dq/dt FLOW
• DROP EXCESS ϕ IN LOAD
TRANSPORT ENERGY
• dq/dt FLOW
• Jϕ FLOW
COLLECT
• S-FLOW, EMF
• ON CONDUCTION ELECTRONS AS Jϕ
Optimize current for voltage available

Figure 17. Operation of an overunity electrical power system.

D-1-F 17

© 1992 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

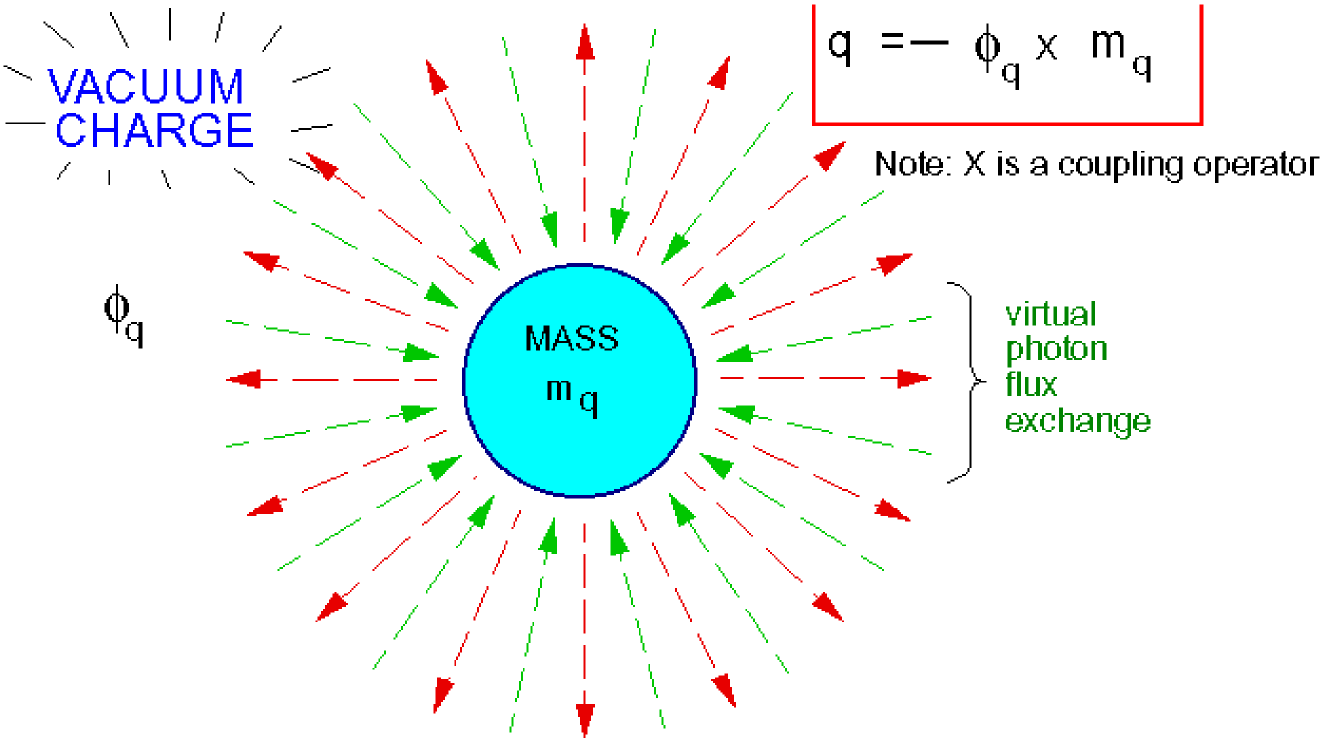
VACUUM
CHARGE
$q = -\phi_q \times m_q$
Note: X is a coupling operator
ϕ_q
MASS
m_q
virtual
photon
flux
exchange

Figure 18. An electric charge q is a broken symmetry. It continuously and violently exchanges energy with the vacuum.

D-1-F 18

© 1994 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

Dawn of The E-BOMB

For the wired world, the allure and the danger of high-power microwave weapons are both very real

By Michael Abrams

In these media-fueled times, when war is a television spectacle and wiping out large numbers of civilians is generally frowned upon, the perfect weapon would literally stop an enemy in his tracks, yet harm neither hide nor hair. Such a weapon might shut down telecommunications networks, disrupt power supplies, and fry an adversary's countless computers and electronic gadgets, yet still leave buildings, bridges, and highways intact. It would strike with precision, in an instant, and leave behind no trace of where it came from.

In fact, it almost certainly is already here, in the form of high-power microwave (HPM) weapons. As their name suggests, HPMs generate an intense "blast" of electromagnetic waves in the microwave frequency band (hundreds of megahertz to tens of gigahertz) that is strong enough to overload electrical circuitry. Most types of matter are transparent to microwaves, but metallic conductors, like those found in metal-oxide semiconductor (MOS), metal-semiconductor, and bipolar devices, strongly absorb them, which in turn heats the material.

An HPM weapon can induce currents large enough to melt circuitry. But even less intense bursts can temporarily disrupt electrical equipment or permanently damage ICs, causing them to fail minutes, days, or even weeks later. People caught in the burst of a microwave weapon would, by contrast, be untouched and might not even know they'd been hit. (There is, however, an effort to build a microwave weapon for controlling crowds; a person subjected to it definitely feels pain and is forced to retreat.)

"HPM sources are maturing, and one day, in the very near future, they will help revolutionize how U.S. soldiers fight wars," says Edl Schamiloglu, a professor of electrical and computer engineering at the University of New Mexico in Albuquerque and one of the leading researchers in this burgeoning field.

The fact that we seldom hear about HPM weapons only adds to their exoticism. Last spring, stories leaked to the press suggested that the Pentagon, after decades of research, had finally deployed such a device in Iraq. And when news footage showed a U.S. bomb destroying an Iraqi TV station, many informed onlookers suspected it was an electromagnetic "e-bomb."

"I saw the detonation, and then I saw the burst—which wasn't much. If they took the station out with that blast, I strongly suspect that we used Iraq as a proving ground" for HPMs, says Howard Seguine, an expert on emerging weapons technology with Decisive Analytics Corp., in Arlington, Va.

But while the U.S. military proudly paraded assorted new war-making technology during its conquest of Iraq, from unmanned combat aerial vehicles to a new satellite-based tracking network, it remained tight-lipped about this "mother of all weapons." Asked at a 5 March news briefing to confirm the rumor, General Tommy Franks, head of U.S. forces during the war, would only say, "I can't talk to you about that because I don't know anything about it."

Military secrecy is nothing new, of course. What is known about microwave weapons is that the U.S. military has actively pursued them since the 1940s, when scientists first

CHIP SIMONS

Microwave weapons researcher Edl Schamiloglu sits in front of the Pulserad-110A accelerator, which his lab at the University of New Mexico uses to produce single 100-nanosecond pulses of electron beams, each pulse emitting hundreds of megawatts of power.

observed the powerful electromagnetic shock wave that accompanied atmospheric nuclear detonations, suggesting a new class of destructiveness. While much of the work on HPMs remains classified, the Pentagon has also recently sponsored a number of U.S. university laboratories to work out the basic principles of microwave weapons, including reliable and compact nonnuclear ways of generating microwave pulses.

Many of those results are being published in the open literature [see To Probe Further, p. 30]. In fact, all you need is a reasonable grasp of physics and electrical engineering to appreciate the ingeniousness of microwave weapons. Anyone with a technical bent could probably also build a crude e-bomb in their garage, a thought that security-minded folks find rather troubling.

How they work

From the military's perspective, HPM weapons, also known as radiofrequency weapons, have many things going for them: their blast travels at the speed of light, they can be fired without any visible emanation, and they are unaffected by gravity or atmospheric conditions. The weapons come in two flavors: ultrawideband and narrowband. Think of the former as a flashbulb, and the latter as a laser; while a flashbulb illuminates across much of the visible spectrum (and into the infrared), a laser sends out a focused beam at a single frequency.

Like the flashbulb, ultrawideband weapons radiate over a broad frequency range, but with a relatively low energy (up to tens of joules per pulse). Their nanoseconds-long burst produces a shock that indiscriminately disrupts or destroys any unshielded electronic components within their reach. The bomb's destructiveness depends on the strength of the ultrawideband source, the altitude at which it is initiated, and its distance from the target [see "E-Bomb Anatomy," p. 29].

Narrowband weapons, by contrast, emit at a single frequency or closely clustered frequencies at very high power (from hundreds up to a thousand kilojoules per pulse), and some can be fired hundreds of times a second, making an almost continuous beam. These pulses can be directed at specific targets—say, a command and control complex positioned on the roof of a hospital in a densely populated neighborhood—and tuned to specific frequencies. Technologically more sophisticated than ultrawideband sources, they are far more difficult to develop, but are reusable and potentially of much greater use to the U.S. military.

Both versions wreak the same kind of havoc on just about any kind of unprotected electronic equipment. Particularly vulnerable is commercial computer equipment; anything in excess of just tens of volts can punch through gates in MOS and metal-semiconductor devices, effectively destroying the device, explains Carlo Kopp, a visiting research fellow in military strategy at the Strategic and Defense Studies Centre in Canberra, Australia, and a computer scientist who lectures at Monash University in Melbourne. The higher the circuitry's density, the more vulnerable it is, because less energy is required to overload and destroy the transistors. HPMs also produce standing waves in electrical grid wiring and telephone and communications wiring, entering through cables, antennas, and even ventilation grills. They can immobilize vehicles with electronic ignition and control systems, too.

"Since the frequency is high, this permits parasitic or stray capacitances to couple energy via paths in the circuit that may not be protected against overvoltage," Kopp explains.

The e-bomb

You could deliver an e-bomb in a number of ways: cruise missile, unmanned aerial vehicle, or aerial bomb. Whether ultrawideband or narrowband, the e-bomb consists of both a microwave source and a power source [again, see diagram, p. 29]. Ultrawideband e-bombs aim to create an electromagnetic pulse like that accompanying a nuclear detonation, except that the nuclear material is replaced with a conventional, chemical explosive. The microwave source typically relies on an extremely fast switching device, according to Kopp, who has written widely on weaponizing HPM technology. Narrowband e-bombs might use a virtual cathode oscillator (vircator) tube or a variant of a magnetron. Though termed narrowband, they don't have the high coherency seen in signal-carrying applications, Kopp says.

It takes gigawatts of power to feed an e-bomb's microwave source. For that, the flux compression generator, or FCG, is a good choice, says Kopp. Invented by Clarence ("Max") Fowler at Los Alamos National Laboratory after World War II as a by-product of research into atomic bomb detonators, FCGs are conceptually simple. The best-known type consists of an explosive-packed copper cylinder surrounded by a helical current-carrying coil. Upon detonation, the explosion flares out the cylinder, short-circuiting the coil and progressively reducing the number of turns in the coil, thus compressing the magnetic flux. Large FCGs have produced tens of gigawatts, and they can be cascaded—connected end to end—so that the output from one stage feeds the next.

Despite its simplicity, an FCG-powered e-bomb is probably too difficult for the average terrorist to build on the cheap. For one thing, to test the assembled apparatus, you have to blow it up. For weapons researchers, the e-bomb poses other problems. The strength of the shock wave dissipates rapidly as it moves out from the explosion. To knock out an electrical power substation, for example, the weapon has to strike within about a hundred meters. "Like all microwave radiation, the effect follows an inverse square law with increasing distance," Kopp notes. Though the explosion needed to force out the current can be fairly small, it keeps the munition from being fully nonlethal and nondetectable. Also, anything that's been hardened or shielded against an electromagnetic pulse from a nuclear bomb will probably emerge unscathed.

Focused like a laser

The type of narrowband HPM weapons that the U.S. military is looking at offers everything that e-bombs do not. They're nonlethal, reuseable, and tunable, and they can be fired from miles away. Like a laser, the focused beam disperses only slightly over great distances. With a frequency range that is between about 1 and 10 GHz, they can penetrate even electronics shielded against a nuclear detonation. The deepest bunkers with the thickest concrete walls are not safe from such a beam if they have even a single unprotected wire reaching the surface.

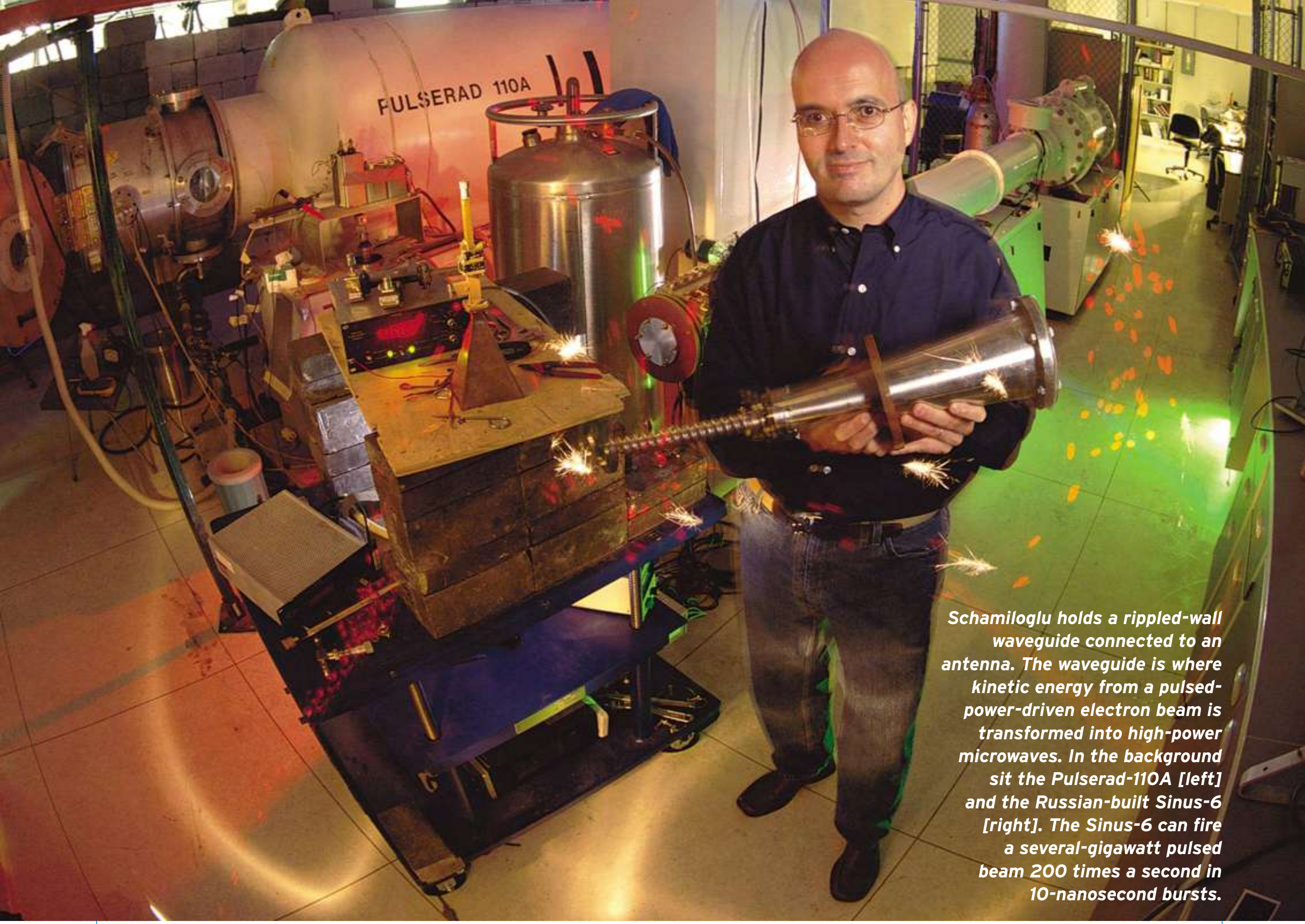

Schamiloglu holds a rippled-wall waveguide connected to an antenna. The waveguide is where kinetic energy from a pulsed-power-driven electron beam is transformed into high-power microwaves. In the background sit the Pulserad-110A [left] and the Russian-built Sinus-6 [right]. The Sinus-6 can fire a several-gigawatt pulsed beam 200 times a second in 10-nanosecond bursts.

CHIP SIMONS

A microwave beam is created much like a laser beam. Between the batteries (or other power source) and the beam sit three elements: capacitors that turn the stored energy into an electron beam of nanosecond bursts, a microwave source that converts the electron beam into focused, high-frequency electromagnetic waves, and an antenna that points and shoots the beam.

Kirtland Air Force Base, in Albuquerque, N.M., is considered the epicenter of the Pentagon's research on pulsed-power electromagnetic weapons. There, its premier pulsed-power system, the Shiva Star, is housed behind meter-thick walls [see photo, p. 28]. An Air Force spokesperson refused to comment on what goes on in their pulsed-power programs, but a fact sheet on the Web site of Kirtland's Directed Energy Directorate describes the Shiva Star as capable of producing "120 thousand volts and 10 million amps for down to one millionth of a second to produce a power flow equivalent to a terawatt."

The Kirtland machine isn't used to investigate HPM weapons per se, and its massive size makes it clearly impractical for delivering microwave beams to any spots of real military interest. Indeed, one big push in microwave weapons has been toward portability. "Back in the 1960s and 1970s, the attitude was, 'Yeah, we can do it—but we need Hoover Dam as our power supply,' " says Seguine. But just as batteries for cellphones and laptops have shrunk and gained capacity, so have sources for microwave weapons.

In the 1990s, the U.S. Air Force Office of Scientific Research set up a five-year Multidisciplinary University Research Initiative (MURI) program to explore microwave sources. One of those funded was the University of New Mexico's Schamiloglu, whose lab is located just a few kilometers down the road from where the Shiva Star sits behind tightly locked doors. Thanks in large part to his and his colleagues' efforts, the fundamental capabilities and limitations of high-power microwave sources are now better understood and appreciated.

Amidst the lead bricks and clutter in Schamiloglu's basement lab lies his masterwork: the Sinus-6. "A lot of laboratories come up with very cute names for these devices," Schamiloglu notes with a smile. "We never did." With a huge cylinder at one end connected to the long microwave source, the Sinus-6 looks like a giant torch lying on its side [see photo, above]. The big cylinder contains a Tesla transformer, whose two coils vibrate in resonance and amplify the incoming voltage "with nearly 100-percent efficiency," Schamiloglu says. Once the pulse has been transformed into an electron beam, it is guided by a strong axial magnetic field through the long tube that will turn it into microwaves.

The Sinus-6 can fire a several-gigawatt pulsed beam 200 times a second in 10-nanosecond bursts. "It has to be pulsed power because what you're after is high peak power," says Schamiloglu. "The power in the microwaves is going to depend on the electric field squared, so if you generate very large power, then the electric field is going to be big."

How big? To drive the Sinus-6's beam continuously for an entire second, you'd need to supply about 25 gigajoules—"the entire output of a typical coal-fired electrical plant for 10 full seconds," Schamiloglu says. Another reason for pulsed rather than continuous power is to avoid a problem at the output end: the air around the antenna would heat to a plasma that in turn would interfere with a continuous beam at these power levels.

The key to reaching gigawatts of power is dumping all the energy in one gigantic, nearly instantaneous pulse. A pressurized gas switch prevents the Tesla transformer from prematurely dumping as it builds up for the next pulse. The switch is filled with highly compressed and nonconducting nitrogen gas. When the transformer coils reach 700 kV, the nitrogen gas breaks down, and the pulse leaps through to the electron-beam diode.

The U.S. Air Force's Shiva Star is a pulsed-power system used to simulate the effects of nuclear weapons.

U.S. AIR FORCE

"Once you've fired the switch, it conducts, it generates a pulse," says Schamiloglu. "It conducts because you've made a plasma channel out of the gas. Then you have to wait for that plasma to recombine and form a neutral gas again. A typical time scale for this thing to recombine and fizzle out and be a neutral gas again is probably on the order of milliseconds."

Among the best candidates for supplying microwaves is the backward wave oscillator; it has the advantage of being tunable (plus or minus 20 percent) and producing output in the 4–10-GHz range. To turn the kinetic energy from the Sinus-6's electron beam into high-power microwaves, the oscillator uses a rippled-wall waveguide, also called a slow-wave structure [see photo, p. 27].

The structure sets up standing electromagnetic waves in such a way that energy is rapidly transferred to them from the incoming beam of relativistic electrons from the Sinus-6. This growing energy initially propagates in the opposite direction of the beam's movement—hence the device's name—and is then reflected forward and radiated in the form of high-power microwaves. Backward wave oscillators, by the way, are also being tested as a way to push giant sails into outer space, to detect space debris, and to clear minefields.

Being able to tune an HPM weapon comes in handy when a particular target proves invulnerable to a particular frequency. "Experience has shown that if the frequency is slightly altered, measurable effects are discerned," Schamiloglu notes. People used to believe that varying the frequency of HPMs wasn't practical, but Schamiloglu and his students proved them wrong.

Coincidence and curiosity led to their discovery. Schamiloglu first acquired the Sinus-6 from Russian researchers in the early 1990s. (The Soviet Union once boasted a sophisticated program to develop microwave weapons; after its collapse, parts of that legacy were put up for sale, to the delight of researchers like Schamiloglu.) But once the apparatus was assembled in his New Mexico lab, he couldn't get it to operate as promised, so Russian colleagues flew over to help.

"One of them took the RF structure [the rippled-wall waveguide] and started hammering on the thing," Schamiloglu recalls. When they tried it again, everything worked. "I was baffled why manhandling this RF structure—ramming it in—could affect the power so much," says Schamiloglu. So he started a series of experiments in which he slightly displaced the backward wave oscillator by increments. With a little experimentation assisted by computer simulations, his team found that the frequency could be adjusted by changing the distance between the diode and the microwave source. The result is that the backward wave oscillator is now one of the few pulsed-power HPM sources that can be tuned.

Smaller is better

One disadvantage of this oscillator, however, is that it needs an external magnetic field to create the microwave beam, a major hurdle to making the whole system smaller. The size of the Sinus-6 and attendant equipment in Schamiloglu's basement suggests that the U.S. military is nowhere near fielding a narrowband HPM weapon. "When I first started working on high-power narrowband sources, we joked that you can do more damage dropping this equipment on someone than you can by using it," he recalls. "People know how to make microwave sources in the laboratory. The challenge is to take this and package it into an autonomous platform and have it function at the same parameter levels."

Schamiloglu is now hard at work under a new MURI program to study the possibilities of making a compact pulsed-power source. Current narrowband generators are typically several meters long, batteries not included. Schamiloglu and his colleagues are studying how to incorporate novel ceramics into pulsed-power systems, which they believe will allow the length of such sources to be halved. The trick is identifying materials with a high dielectric constant that can also survive the harsh electric fields. "Materials will be an important part in making the next giant leap," he says.

Life in a glass house

Among those agreeing that narrowband HPM weapons will need more refining before they become truly useful to the military is Loren B. Thompson, chief operating officer of the Lexington Institute, a military think tank based in Arlington, Va. He looked

E-Bomb Anatomy

In this hypothetical design for an e-bomb, a two-stage flux compression generator provides gigawatts of power to the virtual cathode oscillator (vircator), which produces the high-power microwaves. The bomb's destructiveness depends on the microwave source and target's vulnerability to electromagnetic attack, among other things, but a 10-GW, 5-GHz HPM device would have a "lethal" footprint 400 to 500 meters across, producing field strengths of several kilovolts per meter. Such an e-bomb would wreak major havoc if detonated over a heavily populated area.

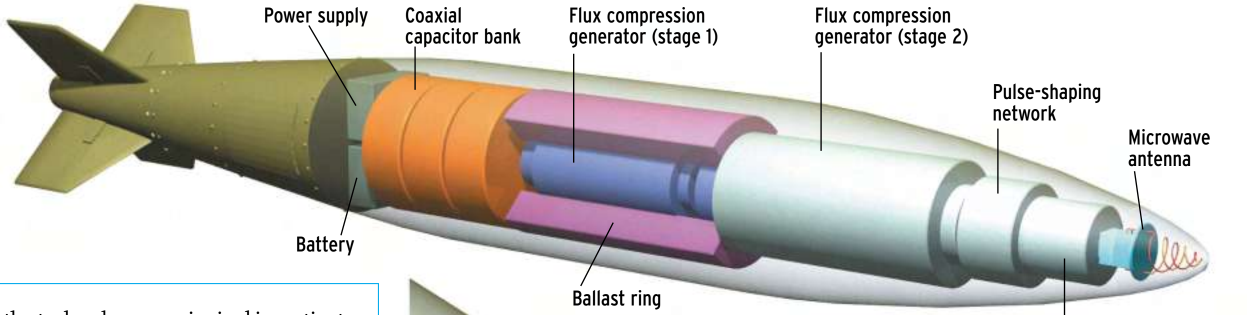
Power supply
Coaxial capacitor bank
Flux compression generator (stage 1)
Flux compression generator (stage 2)
Pulse-shaping network
Microwave antenna
Battery
Ballast ring
Vircator tube

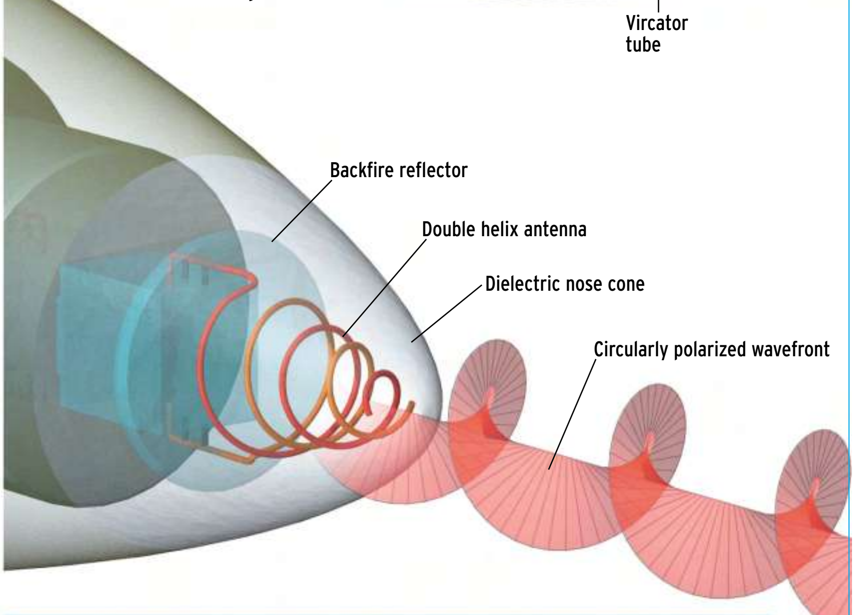
Backfire reflector
Double helix antenna
Dielectric nose cone
Circularly polarized wavefront

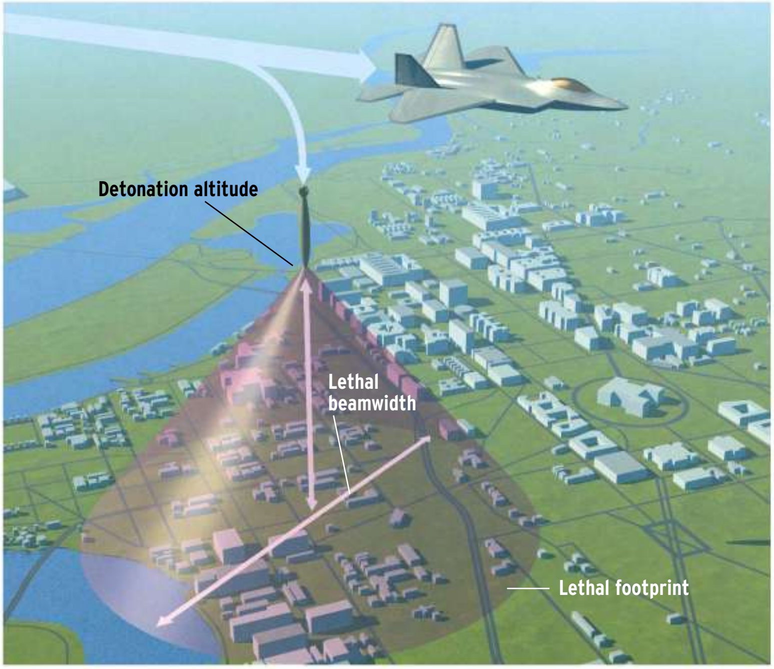
Detonation altitude
Lethal beamwidth
Lethal footprint

ILLUSTRATION: JOHN MACNEILL SOURCE: CARLO KOPP

at the technology as principal investigator of "Directed-Energy Weapons: Technologies, Applications and Implications," a report that the institute put out in February. "We have some fairly rudimentary weapons that we're ready to use," Thompson says. "This is going to be a very important weapons technology, and the basic physical principles are well understood. But the military is having some difficulty in assimilating them."

Thompson's report speaks of a future with satellites delivering missile-debilitating microwaves, unmanned vehicles that fly by and destroy communications systems, and war without civilian casualties. But the fact remains that it's the U.S. military—as well as U.S. financial institutions, PCs, and Game Boys—that will be the most susceptible to such weapons.

"One of the things that happened during the last 10 years—as the Pentagon fell in love with network-centered warfare—is that we purchased a lot of very fragile digital systems off the shelf from commercial sources," Thompson notes. Such moves were taken in the name of cost and efficiency, but the resulting equipment is almost certainly more vulnerable to electromagnetic attack than the vacuum tubes and heavy metal-encased electronics of yesteryear.

"Computers become more vulnerable as the voltage at which they operate becomes smaller," says Victor Granatstein, professor of electrical engineering at the University of Maryland in College Park, who is studying the effects of microwave pulses on integrated electronics. "When our opponent was the

Soviet Union, the electronics were much more robust because they weren't miniaturized. Now they have very thin oxide layers that can easily break down." Wireless networking makes matters worse. Computers and other communications devices now have antennas attached, giving an electromagnetic pulse a direct pathway to its guts.

Meanwhile, the U.S. Navy no longer requires that all its hardware be hardened against nuclear electromagnetic pulses. It deemed that maintaining those standards was too costly and slowed down the integration of new technology. The presumption was that after the Cold War, nobody would be using nuclear bombs, says the Lexington Institute's Thompson. "Whenever I ask the admirals, 'Well, what if someone did use a nuclear bomb?,' I just get this kind of blank I-don't-have-an-answer-for-that sort of look."

Don't Try This at Home

First things first: we do not under any circumstances recommend that you build your own electromagnetic weapon. But if you're hellbent on adding to the mayhem, cheaply and without too much studying, you might try a high-energy radiofrequency, or HERF, gun. As described by engineering student Rostislav Persion on his Web site, Voltage Labs, which is devoted to do-it-yourself electromagnetic weapons, you can make one from a microwave oven.

Before you begin, though, wait until everyone else has left the house. Next, take apart the microwave oven, but don't disconnect any of the components; manufacturers intentionally make this hard to do, so you may end up just breaking the machine. Inside you will find the oven's microwave source: the magnetron. Wrap a tube of sheet metal around it to act as a waveguide. Your power source is the house's ac, so just plug in the oven and point it at your TV. Warning: there's a real possibility that you will burn yourself instead.

For schematics and a demonstration, head to the Web at http://www.voltsamps.com/pages/projects/herf004/. —*M.A.*

In the wrong hands

The scariest part of microwave weapons may be that crude forms of the technology are readily available to anyone right now. "Any nation with a 1950s technology base capable of designing and building nuclear weapons and radars" can build an e-bomb, says military analyst Kopp. Indeed, more than 20 countries now have programs to develop some type of RF weapon.

"The more widespread the technology is, the more likely that people with nefarious purposes will have access. It's just an inescapable fact," says Thompson. "I don't know what we're going to do. Nobody in Washington knows. I imagine that the way the clear thinking starts is with a catastrophe."

Criminals and pranksters have already started exploiting that weakness. In one of the more harmless applications, a Japanese scam artist rigged up a weak microwave generator inside a suitcase to rip off a pachinko parlor. When he placed the suitcase next to one of the machines (which is something like a cross between a slot machine and a pinball machine) and turned it on, the pachinko machine went haywire and disgorged a pile of coins. The perp managed the trick several times before he was caught.

Other press accounts hint at electromagnetic weapons being deployed by Chechen troops, and by an unnamed assailant trying to topple London's futures market [see "Don't Try This at Home," above].

Thankfully, protecting yourself against the microwave-enabled goofballs of the world isn't too difficult. "It is analogous to existing techniques used to trap RF interference inside equipment, except that the higher power levels require special measures," Kopp notes. Rooms or equipment chassis must become electrically sealed Faraday cages, and protective devices must be added wherever cables enter the protected volume. "Optical fibers are very useful in this game."

Such protective measures are a lot cheaper to design in from the beginning than to add on afterward, says Howard Seguine. "The general rule of thumb is that if you do the hardening during the design phase, it increases the cost roughly 1 percent. If you do it afterward, it may cost as much as 30 percent more."

But maybe hardening is a waste of time. Arthur Varanelli, a Raytheon Co. engineer who has helped write several IEEE standards for electromagnetic field measurement, human exposure, and safety, is skeptical that a malicious prankster could exploit the technology.

"Some of this stuff is just so far out there," Varanelli says. "I just don't see people running around with Buck Rogers ray guns. It's great for a science fiction writer, great to prey upon people's fears." He scoffs at the suggestion that a do-it-yourselfer could build a microwave weapon potent enough to do real damage. "People can put tacks in the road. Are we worried about electronic tacks in the air?"

The wide disparity in opinions and the uncertainty about microwave weapons, from Loren Thompson on one end to Arthur Varanelli on the other, are all part of what makes them so powerful, says military analyst John Pike, who is director of GlobalSecurity.org (Alexandria, Va.). "It all depends on the complex interactions between the weapon and the target," he notes. "I can set up a strap-down chicken test that makes [an HPM weapon] look pretty good. But as soon as I start getting into real-world targets, maybe it doesn't work so well."

"Part of the story is we don't know what the story is," Pike says. "These are weapons that by their nature seek the shadows. And unlike cluster bombs or atomic bombs, they aren't going to leave behind unambiguous evidence of their use." ●

To Probe Further

For a detailed technical discussion of high-power microwaves, see *High-Power Microwave Sources and Technologies*, edited by IEEE Fellows Robert J. Barker and Edl Schamiloglu (Wiley–IEEE Press, 2001). Schamiloglu is also coauthor, with James Benford and John Swegle, of the forthcoming *High-Power Microwaves*, 2nd edition (Institute of Physics, 2004).

The truly prepared, or merely paranoid, will want to consult Carlo Kopp's "Hardening Your Computing Assets" at http://www.globalsecurity.org/military/library/report/1997/harden.pdf.

BEARDEN
FER DE LANCE
12.14 15:24
BRIEFING ON SOVIET SCALAR
ELECTROMAGNETIC WEAPONS

[ORDER THE ALL NEW
2ND EDITION
OF "FER DE LANCE"]

Updated to the year 2002

**Stunning new revelations
about the secret scalar
EM wars the media ignores**

**THIS BOOK IS THE
ONLY SOURCE FOR
THIS INFORMATION**

[The Tom Bearden Website]

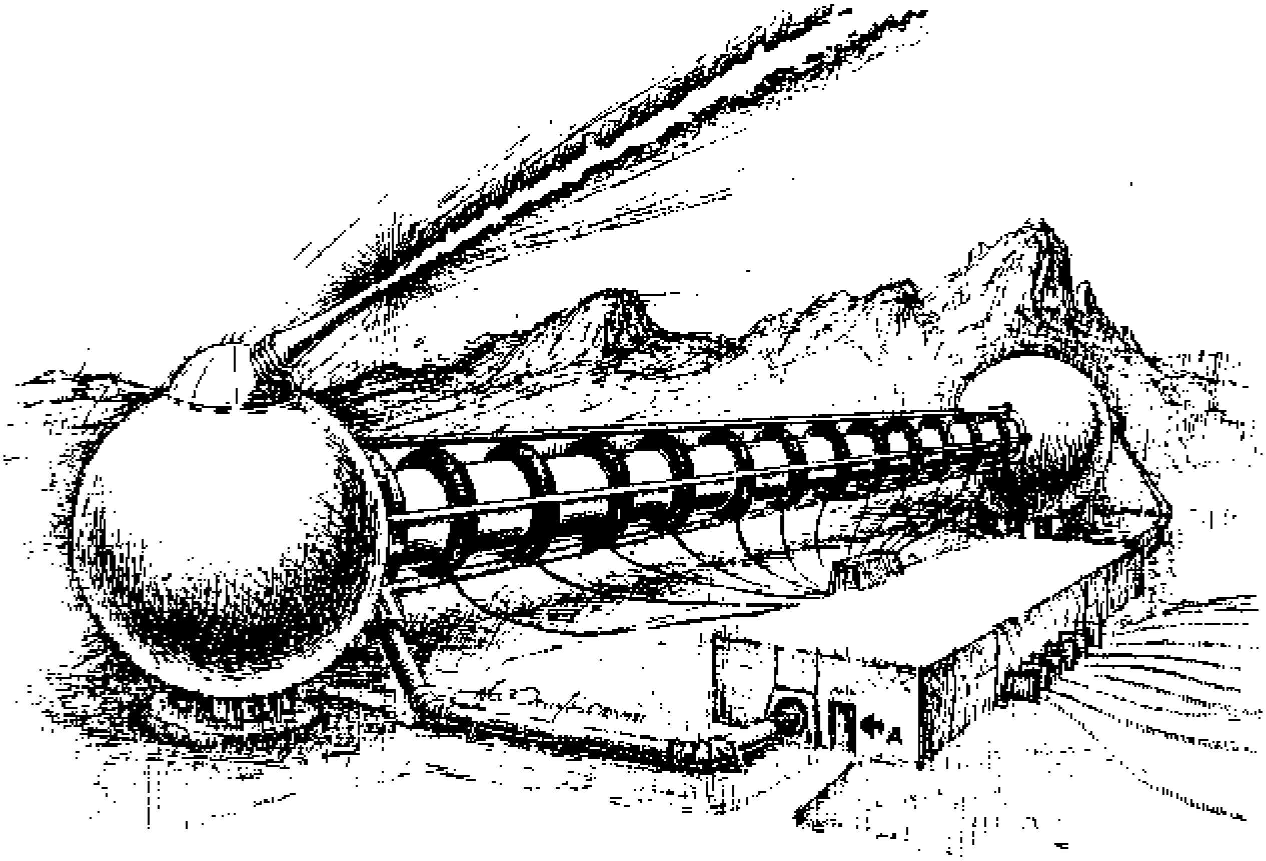

Scalar Potential Interferometer

Figure 6. Multimode Tesla Weapon

In the 1930's Tesla announced other bizarre and terrible weapons: a death ray, a weapon to destroy hundreds or even thousands of aircraft at hundreds of miles range, and his ultimate weapon to end all war -- the Tesla shield, which nothing could penetrate. However, by this time no one any longer paid any real attention to the forgotten great genius. Tesla died in 1943 without ever revealing the secret of these great weapons and inventions.

Unfortunately, today in 1981 the Soviet Union has long since discovered and weaponized the Tesla scalar wave effects. Here we only have time to detail the most powerful of these frightening Tesla weapons -- which Brezhnev undoubtedly was referring to in 1975 when the Soviet side at the SALT talks suddenly suggested limiting the development of new weapons "more frightening than the mind of man had imagined." One of these weapons is the Tesla howitzer recently completed at the Saryshagan missile range and presently considered to be either a high-energy laser or a particle beam weapon, (See Aviation Week & Space Technology, July 28, 1980, p. 48 for an artist's conception.)

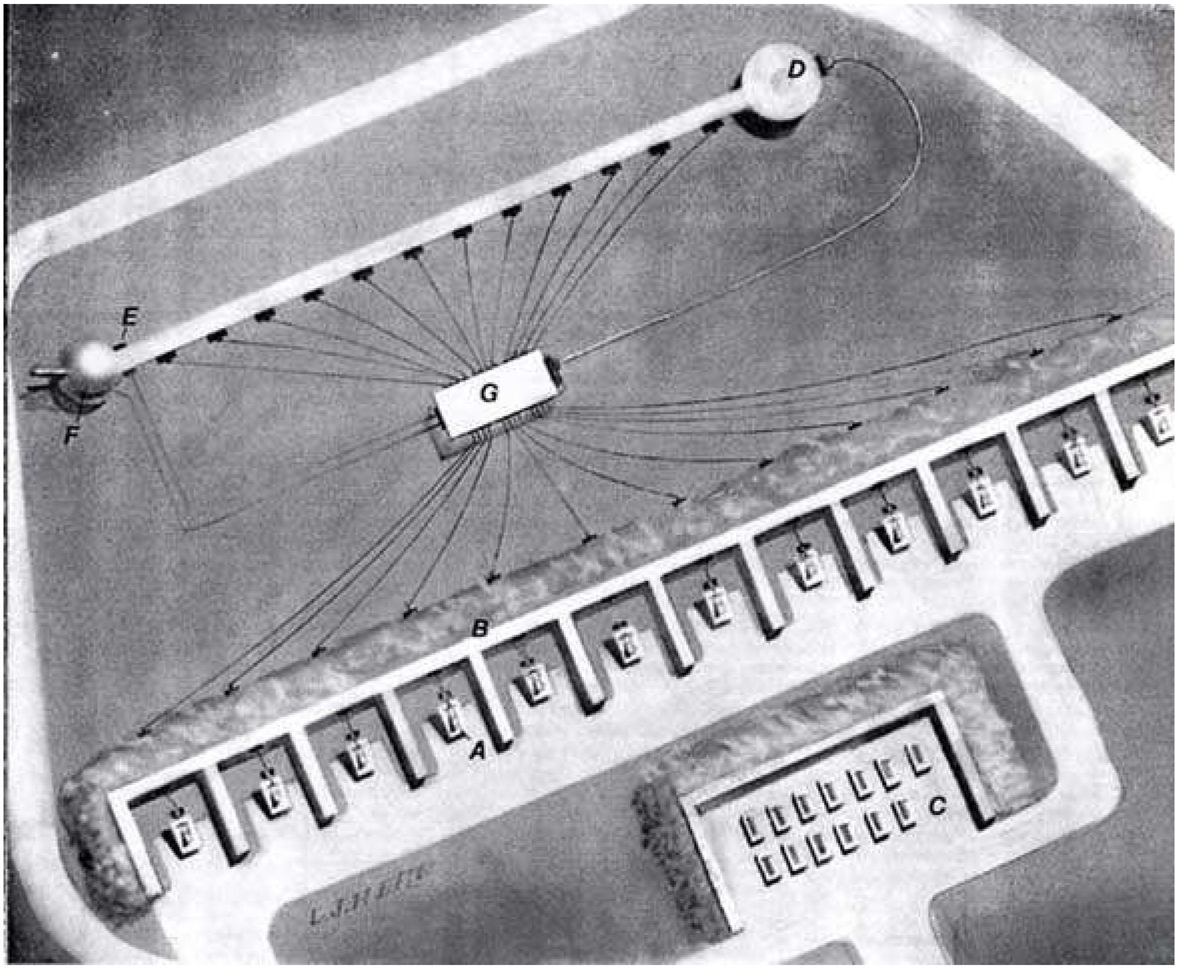

Aviation Week & Space Technology July 28, 1980

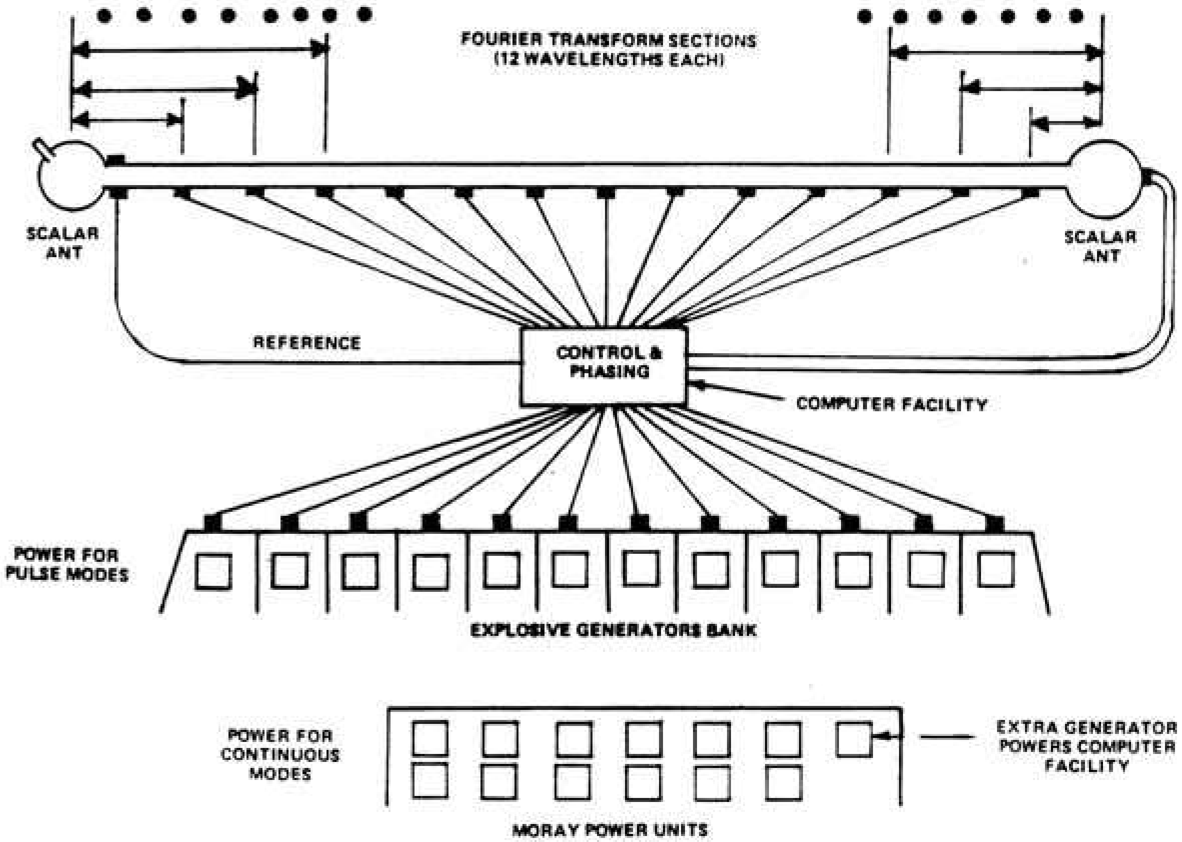
FOURIER TRANSFORM SECTIONS
(12 WAVELENGTHS EACH)
SCALAR ANT
SCALAR ANT
REFERENCE
CONTROL & PHASING
COMPUTER FACILITY
POWER FOR PULSE MODES
EXPLOSIVE GENERATORS BANK
POWER FOR CONTINUOUS MODES
EXTRA GENERATOR POWERS COMPUTER FACILITY
MORAY POWER UNITS

Figure 7. Tesla Weapons at Saryshagan

The Saryshagan howitzer actually is a huge Tesla scalar interferometer with four modes of operation. One continuous mode is the Tesla shield, which places a thin, impenetrable hemispherical shell of energy over a large defended area. The 3-dimensional shell is created by interfering two Fourier-expansion, 3-dimensional scalar hemispherical patterns in space so they pair-couple into a dome-like shell of intense, ordinary electromagnetic energy. The air molecules and atoms in the shell are totally ionized and thus highly excited, giving off intense, glowing light. Anything physical which hits the shell receives an enormous discharge of electrical energy and is instantly vaporized -- it goes pfft! like a bug hitting one of the electrical bug killers now so much in vogue.

If several of these hemispherical shells are concentrically stacked, even the gamma radiation and EMP from a high altitude nuclear explosion above the stack cannot penetrate all the shells due to repetitive absorption and reradiation, and scattering in the layered plasmas.

In the continuous shield mode, the Tesla interferometer is fed by a bank of Moray free energy generators, so that enormous energy is available in the shield. A diagram of the Saryshagan-type Tesla howitzer is shown in figure 7. Hal Crawford's fine drawing of the interferometer end of the Tesla howitzer is shown in figure 6. Hal's exceptional rendition of the Tesla shield produced by the howitzer is shown in figure 8.

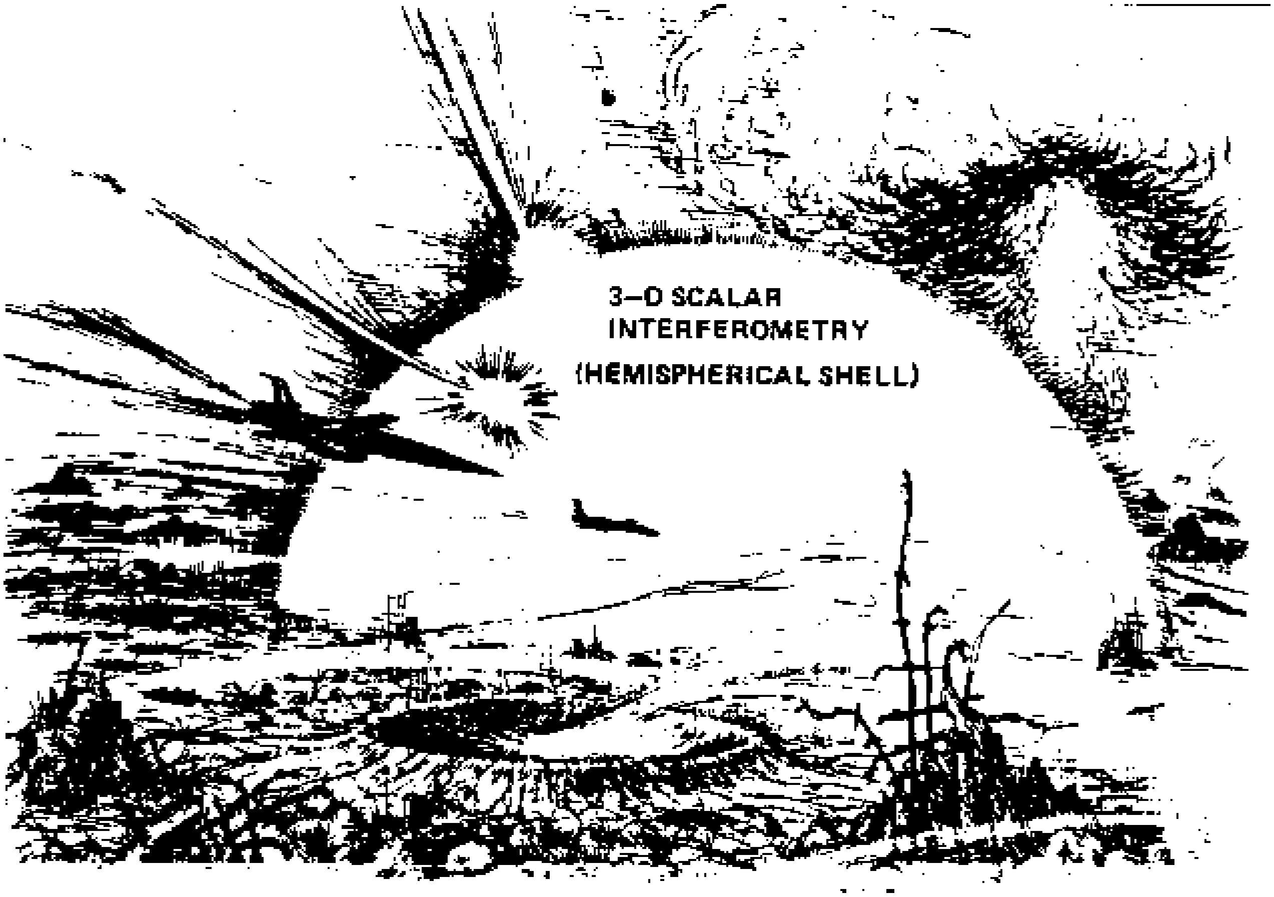
3-D SCALAR
INTERFEROMETRY
(HEMISPHERICAL SHELL)

Figure 8. The Tesla Shield

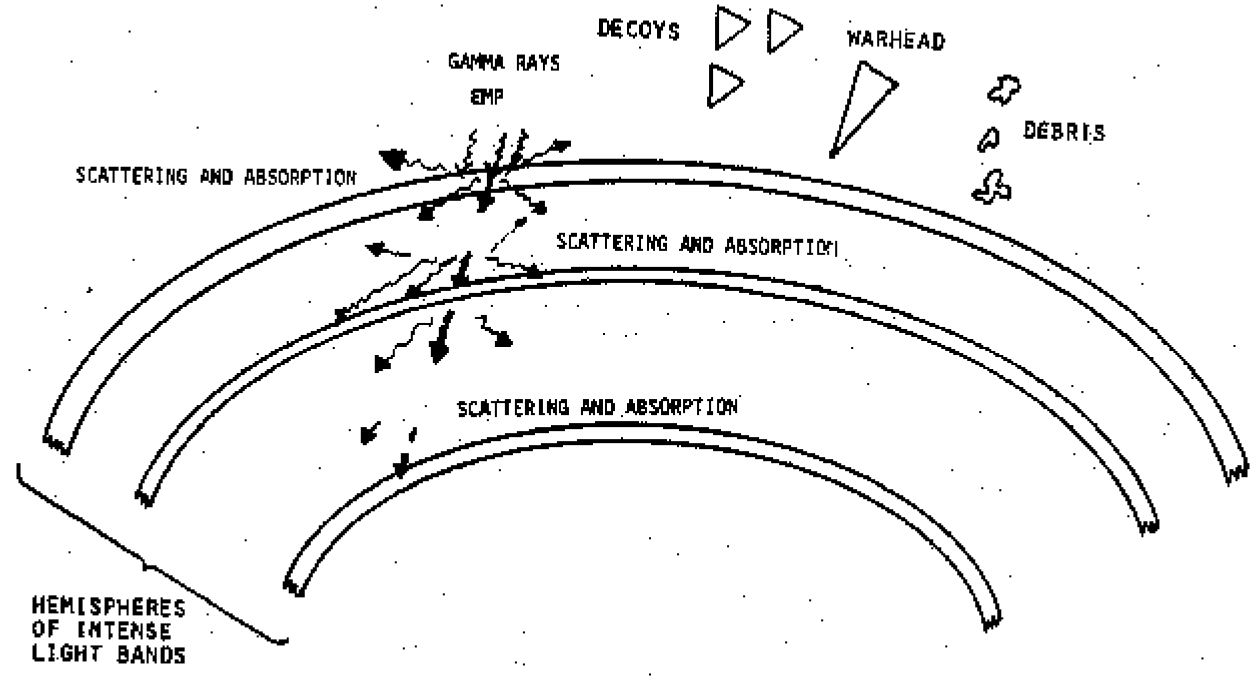
DECOYS
WARHEAD
GAMMA RAYS
EMP
DEBRIS
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
HEMISPHERES
OF INTENSE
LIGHT BANDS

Figure 9. Tesla Terminal Area Defense System

In the pulse mode, a single intense 3-dimensional scalar phi-field pulse form is fired, using two truncated Fourier transforms, each involving several frequencies, to provide the proper 3-dimensional shape (Figure 10). This is why two scalar antennas separated by a baseline are required. After a time delay calculated for the particular target, a second and faster pulse form of the same shape is fired from the interferometer antennas. The second pulse overtakes the first, catching it over the target zone and pair-coupling with it to instantly form a violent EMP of ordinary vector (Hertzian) electromagnetic energy. There is thus no vector transmission loss between the howitzer and the burst. Further, the coupling time is extremely short, and the energy will appear sharply in an "electromagnetic pulse (EMP)" strikingly similar to the 2-pulsed EMP of a nuclear weapon.

This type weapon is what actually caused the mysterious flashes off the southwest coast of Africa, picked up in 1979 and 1980 by Vela satellites. The second flash, e.g., was in the infrared only, with no visible spectrum. Nuclear flashes do not do that, and neither does superlightning, meteorite strikes, meteors, etc. In addition, one of the scientists at the Arecibo Ionospheric Observatory observed a gravitational wave disturbance -- signature of the truncated Fourier pattern and the time-squeezing effect of the Tesla potential wave -- traveling toward the vicinity of the explosion.

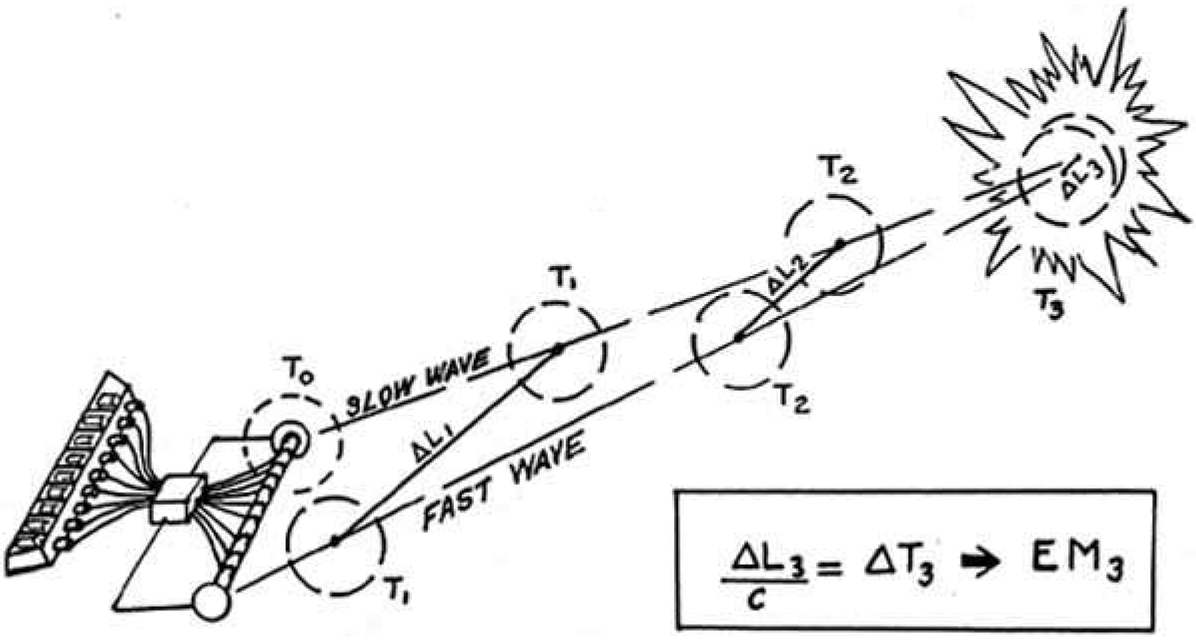

TESLA HOWITZER
(SCALAR INTERFEROMETER)

Figure 10. "Nuclear" Flashes off the Coast of Africa

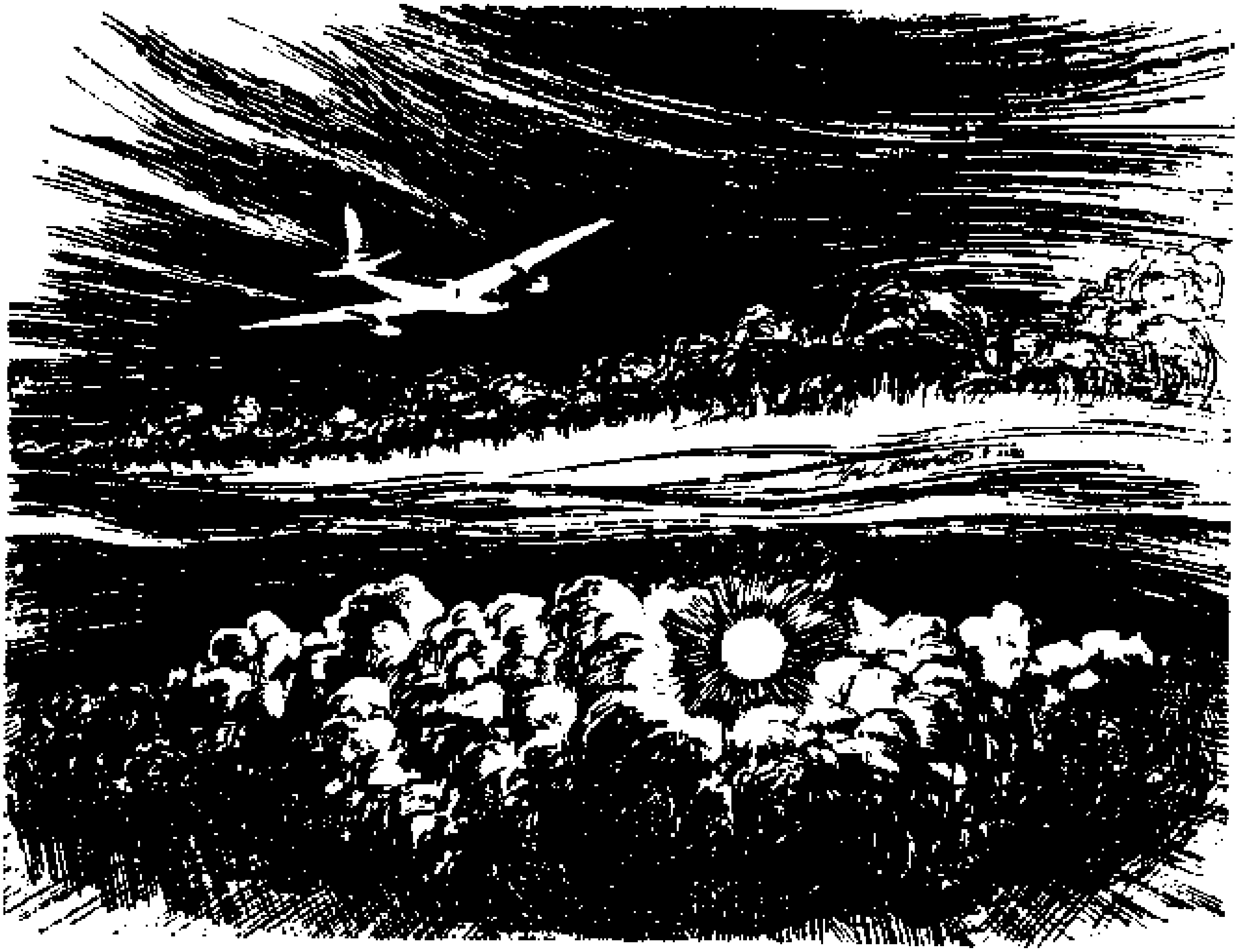

Lithuania - 10 Sep 1976 - British European Airways Flight #831 between Moscow and London

CIA Report Released under FOIA

Figure 11. Continuous Tesla Fireball

The pulse mode may be fed from either Moray generators or -- if the Moray generators have suffered their anomalous "all fail" malfunction -- ordinary explosive generators. Thus the Tesla howitzer can always function in the pulse mode, but it will be limited in power if the Moray generators fail.

In the continuous mode, two continuous scalar waves are emitted -- one faster than the other -- and they pair-couple into vector energy at the region where they approach an in-phase condition. In this mode, the energy in the distant "ball" or geometric region would appear continuously and be sustained -- and this is Tesla's secret of wireless transmission of energy at a distance without any losses. It is also the secret of a "continuous fireball" weapon capable of destroying hundreds of aircraft or missiles at a distance. An example of a Soviet test of this mode of operation is shown in figure 11.

Witness to a super weapon?

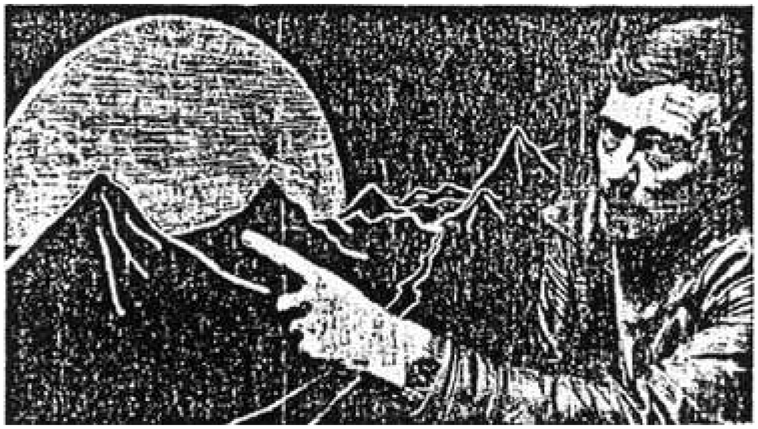

Nick Downie describes the strange lurid glow that flared silently over the Hindu Kush

THE SUNDAY TIMES, 17 AUGUST 1980

(Multiple incidents in Sept., 1979)

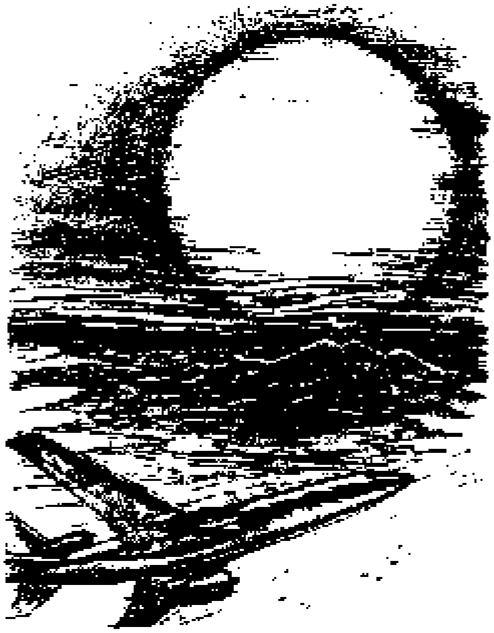

TEHERAN, IRAN 17 JUNE 1966

SEEN FROM NEAR MEHRABAD AIRPORT

OBSERVED 4-5 MINUTES

SEEN BY 2 AIRCRAFT

CIA REPORT RELEASED UNDER FOIA

Figure 12. Tesla EMP Globe

The volume of the Tesla fireball can be vastly expanded to yield a globe which will not vaporize physical vehicles but will deliver an EMP to them to dud their electronics. A test of this mode is shown in figure 12. (See also Gwynne Roberts, "Witness to a Super Weapon?", the London Sunday Times, 17 August 1980 for several other tests of this mode at Saryshagan, seen from Afghanistan by British TV cameraman and former War Correspondent Nick Downie.)

If the Moray generators fail anomalously, then a continuous mode limited in power and range could conceivably be sustained by powering the interferometer from more conventional power-sources such as advanced magnetohydrodynamic generators.

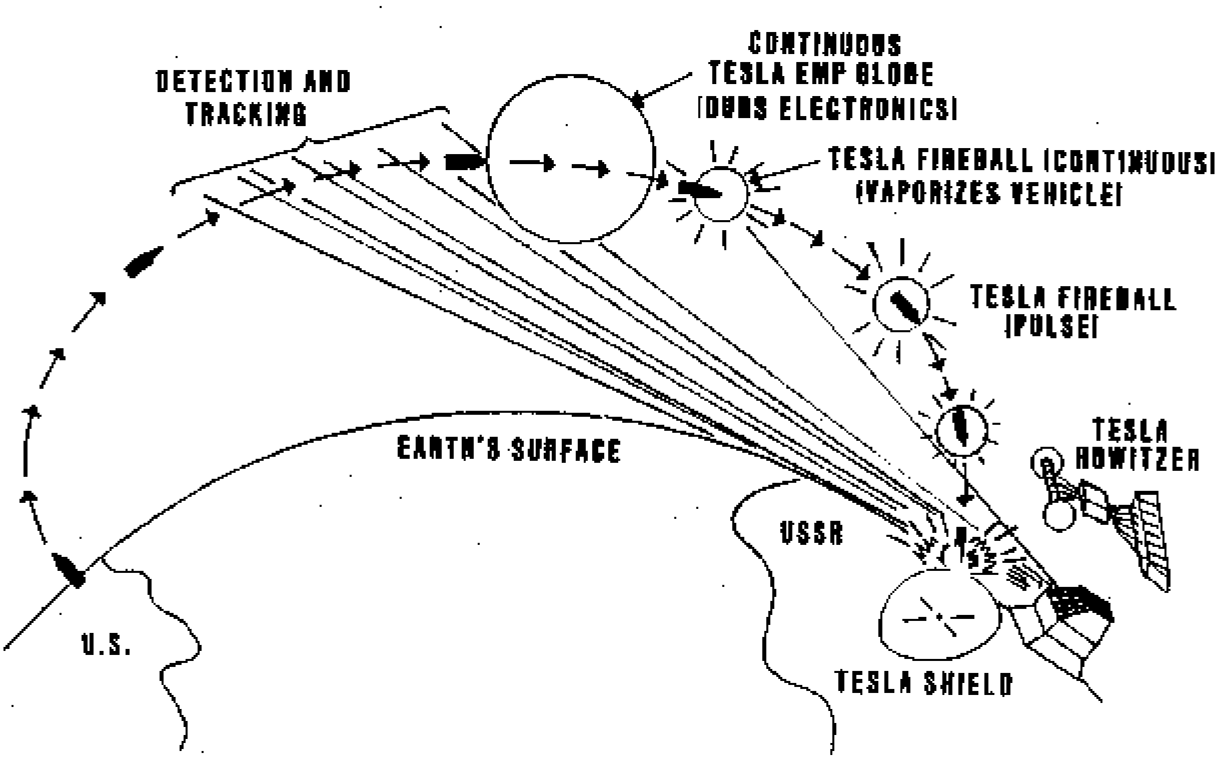
DETECTION AND TRACKING
CONTINUOUS TESLA EMP GLOBE (DUDS ELECTRONICS)
TESLA FIREBALL (CONTINUOUS) (VAPORIZES VEHICLE)
TESLA FIREBALL (PULSE)
TESLA HOWITZER
EARTH'S SURFACE
USSR
U.S.
TESLA SHIELD

Figure 13. Tesla ABM Defenses

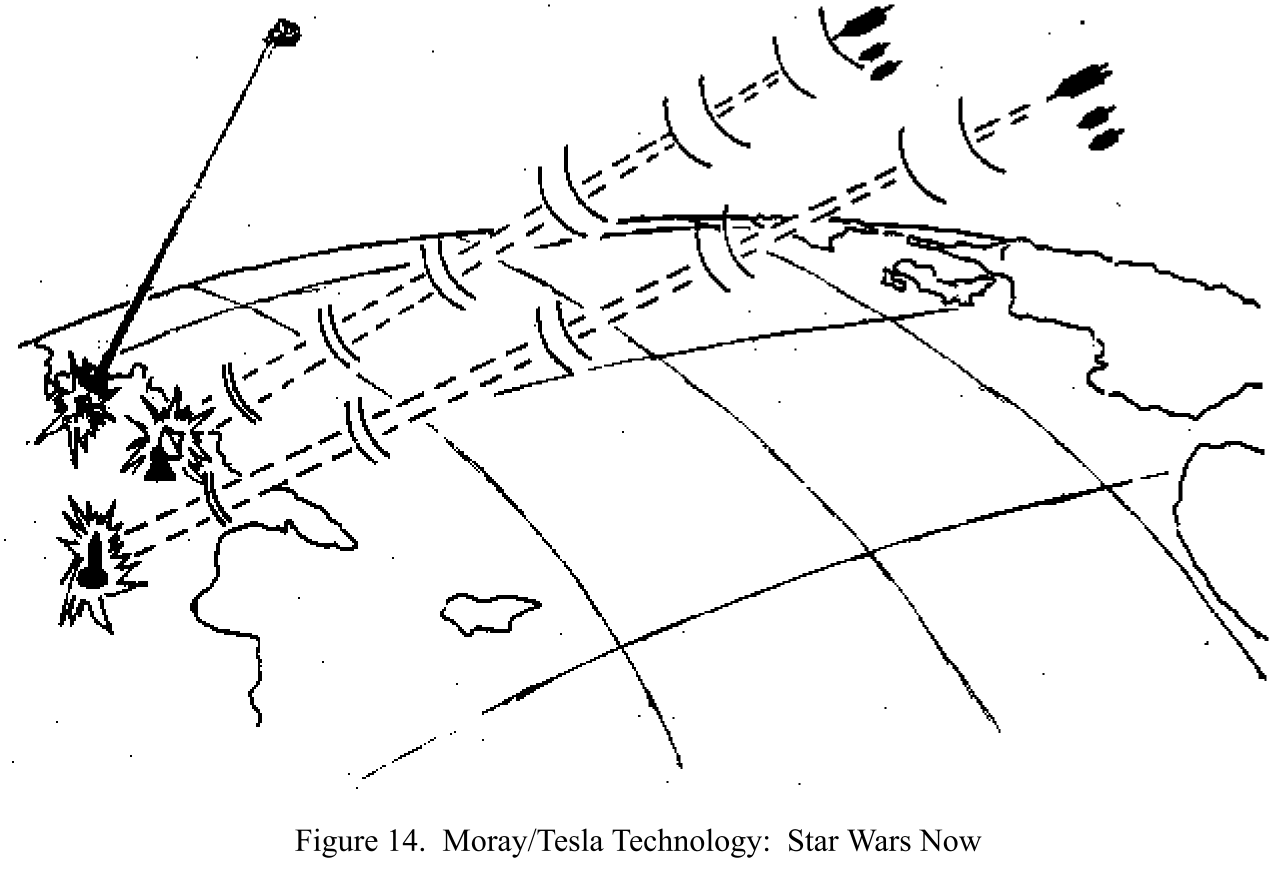

Figure 14. Moray/Tesla Technology: Star Wars Now

Typical strategic ABM uses of Tesla weapons are shown in figure 13. In addition, of course, smaller Tesla howitzer systems for anti-tactical ballistic missile defense of tactical troops and installations could be constituted of more conventional field missile systems using paired or triplet radars, of conventional external appearance, in a scalar interferometer mode.

With Moray generators as power sources and multiply deployed reentry vehicles with scalar antennas and transmitters, ICBM reentry systems now can become long range "blasters" of the target areas, from thousands of kilometers distance (figure 14). Literally, "Star Wars" is liberated by the Tesla technology. And in air attack, jammers and ECM aircraft now become "Tesla blasters." With the Tesla technology, emitters become primary fighting components of stunning power.

The potential peaceful implications of Tesla waves are also enormous. By utilizing the "time squeeze" effect, one can get antigravity, materialization and dematerialization, transmutation, and mindboggling medical benefits. One can also get subluminal and superluminal communication, see through the earth and through the ocean, etc. The new view of phi-field also provides a unified field theory, higher orders of reality, and a new super-relativity, but detailing these possibilities must wait for another book.

With two cerebral brain halves, the human being also has a Tesla scalar interferometer between his ears. And since the brain and nervous system processes avalanche discharges, it can produce (and detect) scalar Tesla waves to at least a limited degree. Thus a human can sometimes produce anomalous spatiotemporal effects at a distance and through time. This provides an exact mechanism for psychokinesis, levitation, psychic healing, telepathy, precognition, postcognition, remote viewing, etc. It also provides a reason why an individual can detect a "stick" on a radionics or Hieronymus machine (which processes scalar waves), when ordinary detectors detect nothing. Unfortunately there is not room to develop the implications of this human Tesla interferometry in detail, for that must wait for yet another book, presently in its initial stages, that Hal Crawford and I are writing.

Table 5. Orders of Reality

- **PHOTONS ARE**
  - **> PAIR-COUPLED SCALARS**
  - **> VELOCITY-LIMITED TO C**
  - **> CARRIERS OF T**

- **PHOTON INTERACTION**
  - **> IS UBIQUITOUS**
  - **> PRODUCES $\dot{T}$ = C**
  - **> YIELDS 1st ORDER REALITY**

- **SCALAR O WAVES**
  - **> NOT VELOCITY-LIMITED**
  - **> YIELD HIGHER ORDER REALITIES**

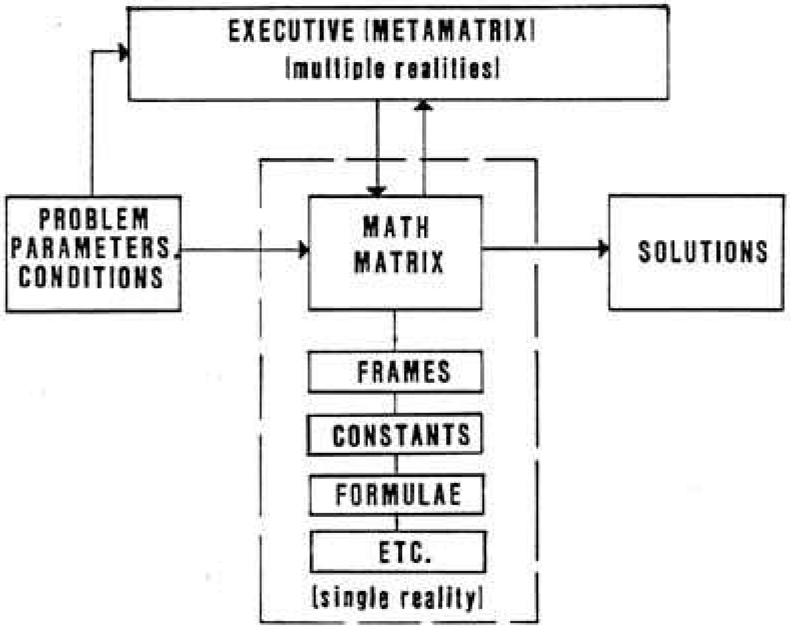

Figure 15. Implications of Tesla Potential

At the July 1981 U.S. Psychotronics Association's Annual Conference in Dayton, Ohio, I presented the first rough paper on the Tesla secret and scalar interferometry. A videotape of the presentation was made and will shortly be available. I am also scheduled to make a special presentation at the Alternate Energy Conference in Toronto, Canada in latter October, 1981. A professional, videotaped two-hour presentation on this subject is also being prepared. Wide distribution of the material through the international underground physics and technology network has already been made. This time, God willing, Tesla's secret will not be suppressed for another 80 years!

And perhaps it is not yet too late. The material has cost me (now) some 16 years of agonizing labor and nearly $100,000 of my own personal funds. No orthodox university, scientific group, foundation, or governmental agency would support such an effort, either financially or otherwise. Indeed, most ordinary journals will not even accept material on such matters. Nonetheless, the area is of overwhelming importance and I truly believe Tesla's lost secret will shortly affect the lives of every human being on earth.

Perhaps with the free and open release of Tesla's secret, the scientific and governmental bureaucracies will be. shocked awake from their slumber, and we can develop defenses before Armageddon occurs. Perhaps there is hope after all -- for even Brezhnev, in his strange July, 1975 proposal to the SALT talks, seemed to reveal a perception that a turning point in war and weaponry may have been reached, and that human imagination is incapable of dealing with the ability to totally engineer reality itself. Having tested the weapons, the Soviets must be aware that the ill-provoked oscillation of timeflow affects the minds and thoughts -- and the very lifestreams and even the collective species unconsciousnesses -- of all lifeforms on earth. They must know that these weapons are two-edged swords, and that the backlash from their use can be far more terrible to the user than was the original effect to his victim.

If we can avoid the Apocalypse, the fantastic secret of Nikola Tesla can be employed to cure and elevate man, not kill him. Tesla's discovery can eventually remove every conceivable external human limitation. If we humans ourselves can elevate our consciousness to properly utilize the Tesla electromagnetics, then Nikola Tesla -- who gave us the electrical twentieth century in the first place -- may yet give us a fantastic new future more shining and glorious than all the great scientists and sages have imagined.

References

The Tom Bearden Website

34 Flaws in Classical EM Theory

Note: If the scientific community had done their job, this Website would probably not be necessary.

---

1. SLIDE 1
2. SLIDE 2
3. SLIDE 3
4. SLIDE 4
5. SLIDE 5

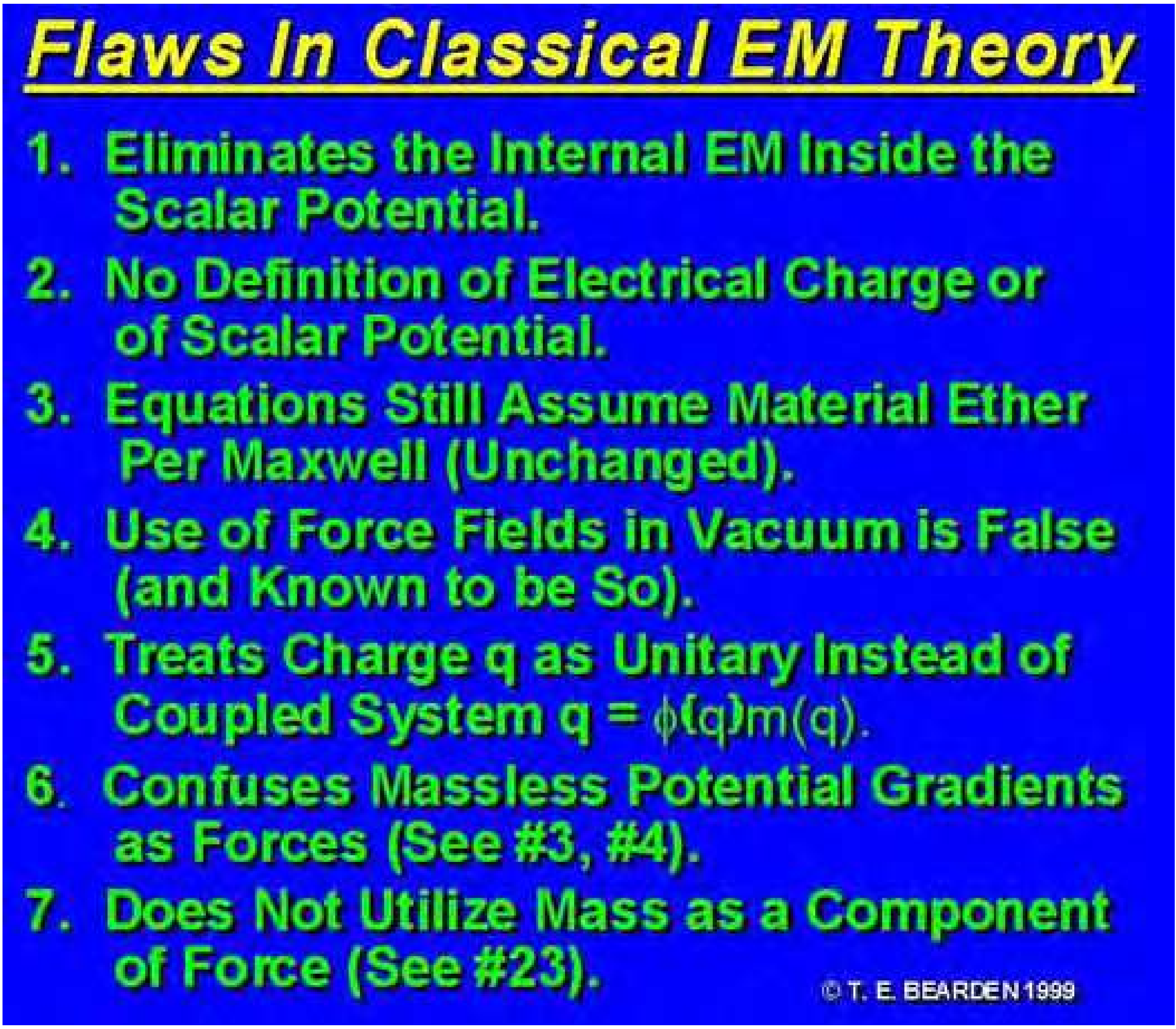

Flaws In Classical EM Theory

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi(q)m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

© T. E. BEARDEN 1999

Next Slide

Return to Index

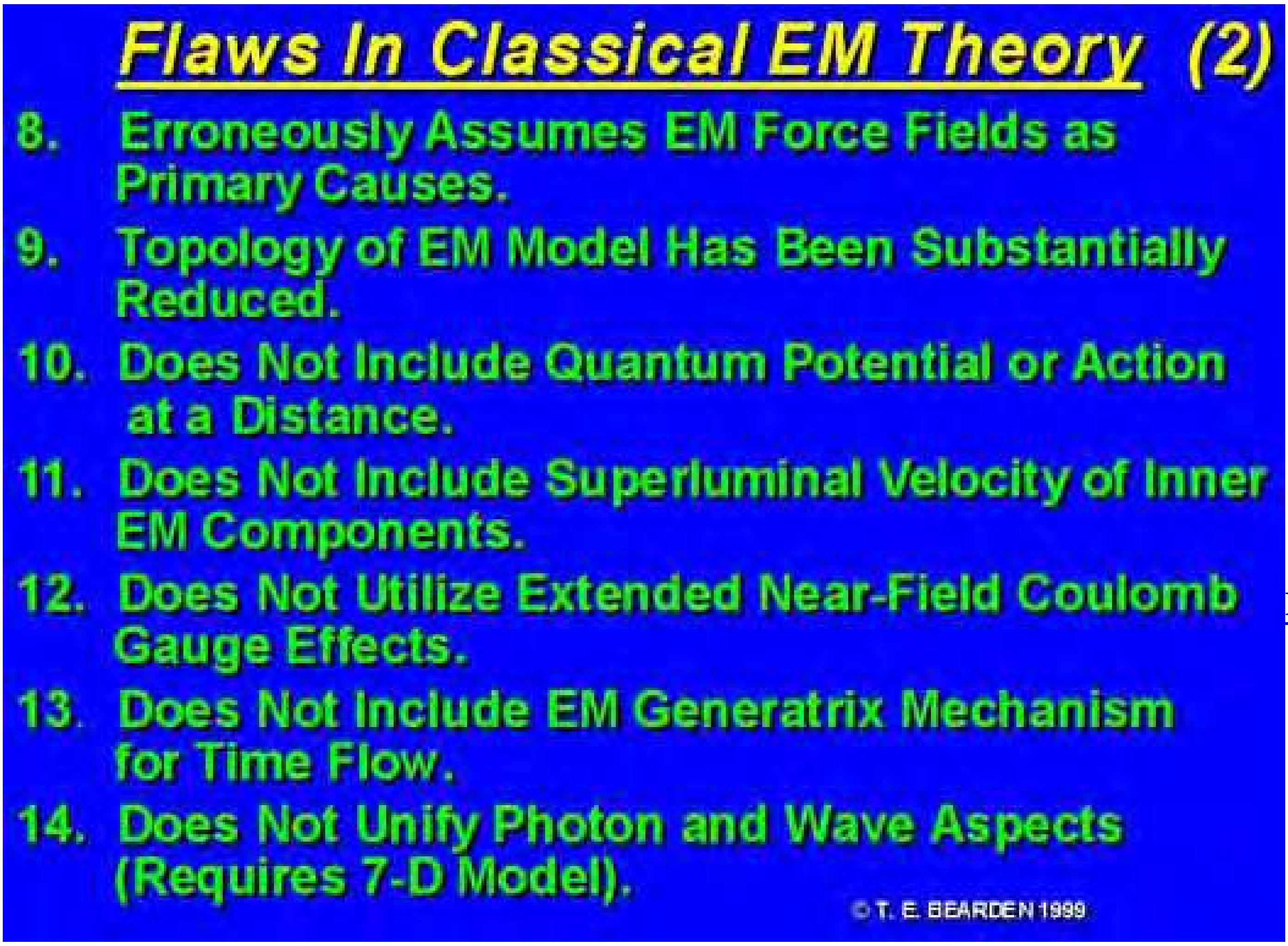

Flaws In Classical EM Theory (2)

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

© T. E. BEARDEN 1999

Next Slide

Return to Index

Previous Slide

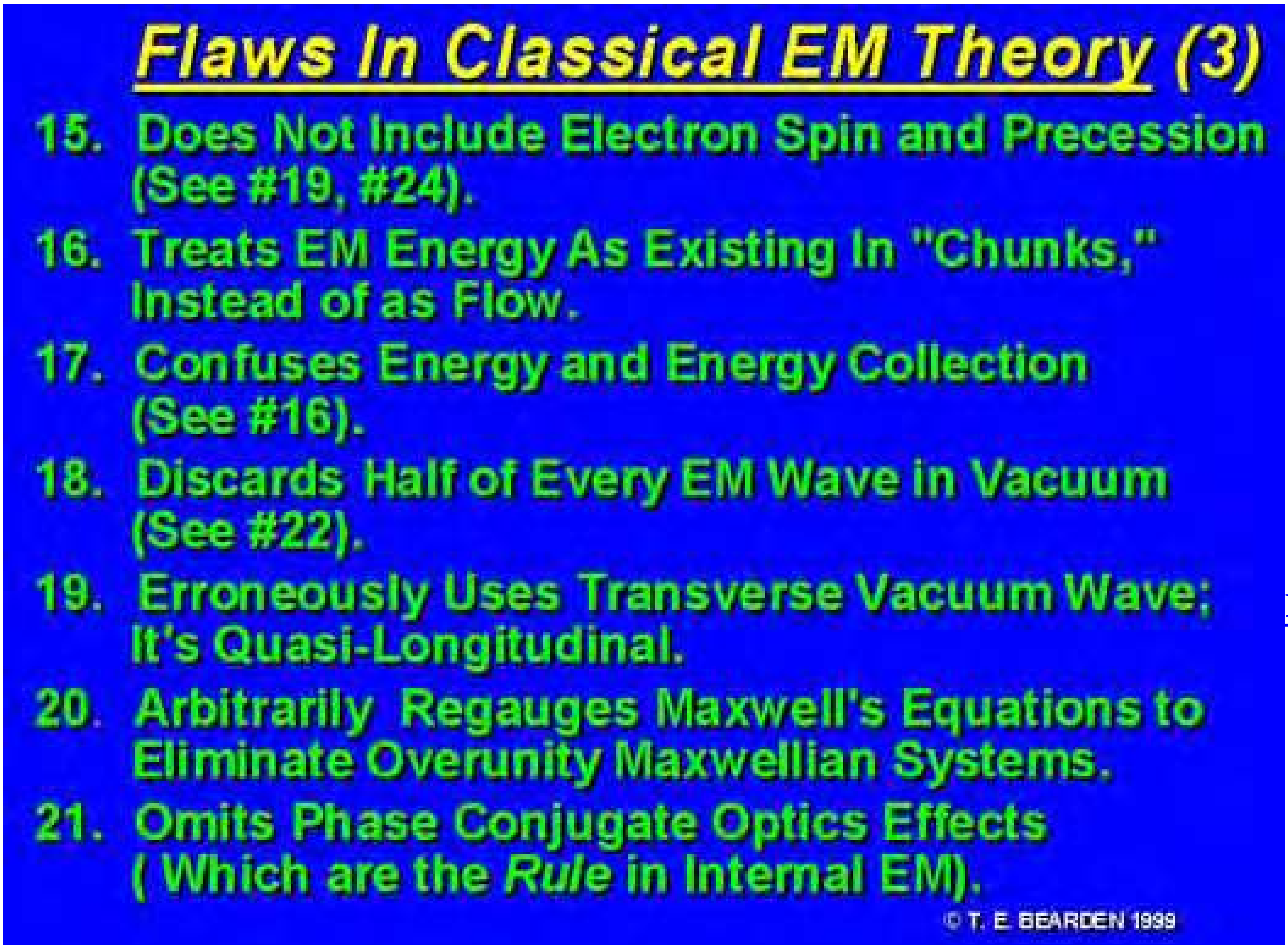

Flaws In Classical EM Theory (3)

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Quasi-Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity Maxwellian Systems.
21. Omits Phase Conjugate Optics Effects (Which are the *Rule* in Internal EM).

© T. E. BEARDEN 1999

Next Slide

Return to Index

Previous Slide

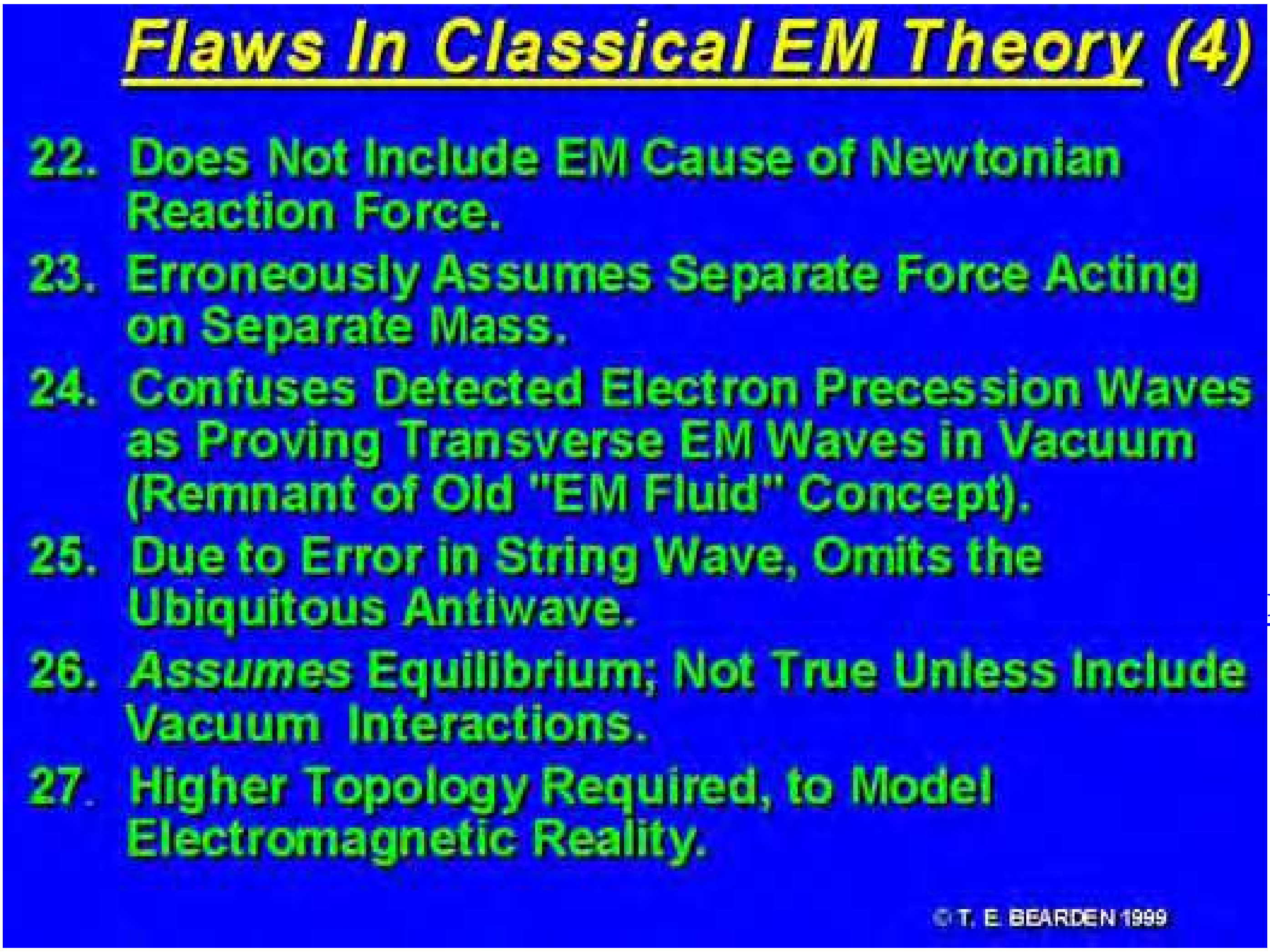

Flaws In Classical EM Theory (4)

22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.

© T. E. BEARDEN 1999

Next Slide

Return to Index

Previous Slide

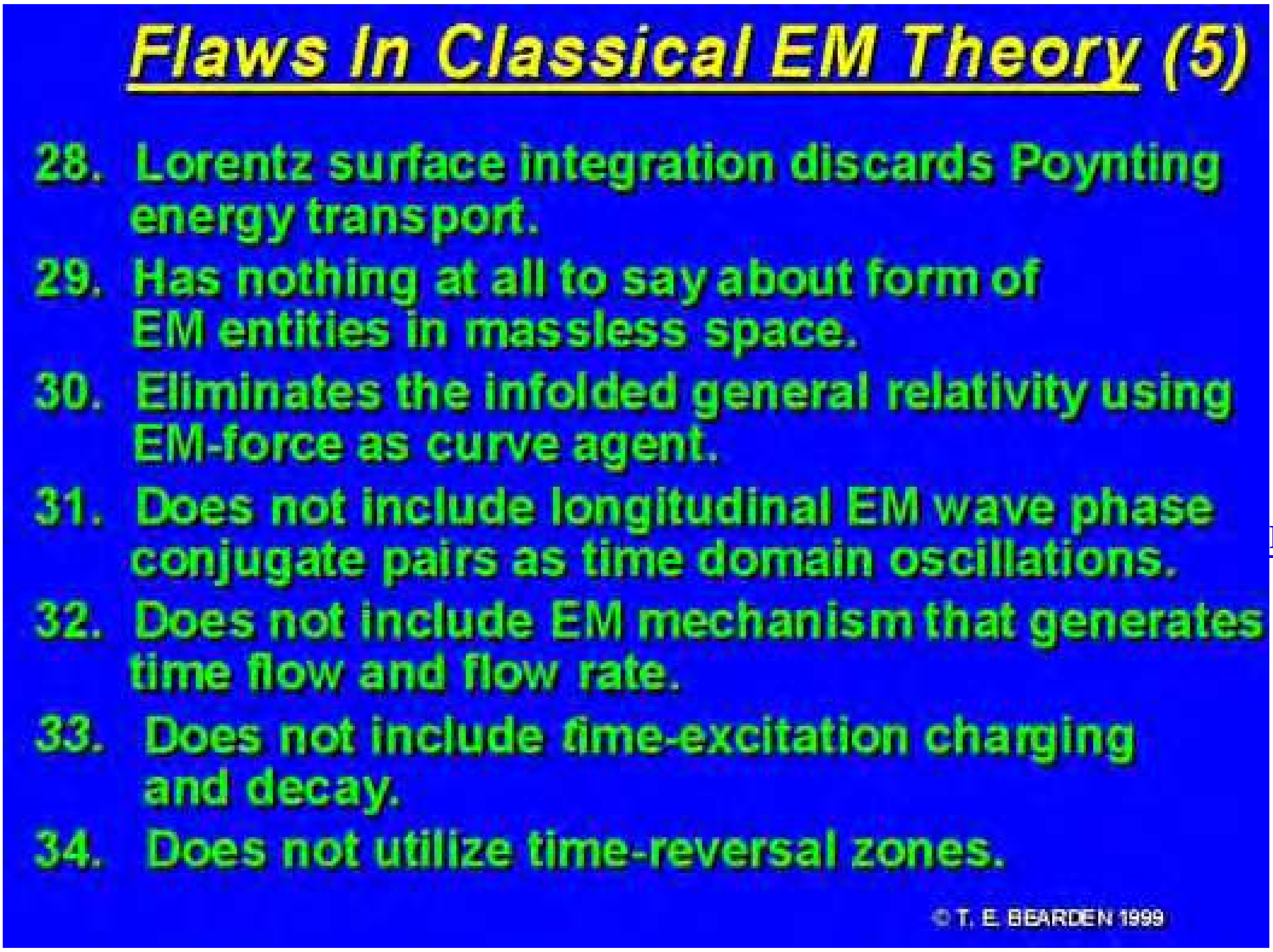

Flaws In Classical EM Theory (5)

28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM wave phase conjugate pairs as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.
33. Does not include time-excitation charging and decay.
34. Does not utilize time-reversal zones.

© T. E. BEARDEN 1999

Return to Index

Previous Slide

The Tom Bearden Website

Mind Control and EM Wave Polarization Transductions - Part I, II & III

Reprinted from "Explore" magazine, 1999

Part I

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9
- Page 10

Part II

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8

Part III

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9

- The U.S. approach to mind control

- "Voices in my head"
  Flanagan's neurophone - used to give instructions to diplomats during negotiations, etc.

Mind Control and EM Wave Polarization Transductions, Part 1

©Copyright 1999 by T.E. Bearden, USA

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental pro-cedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

Abstract

For some time we have been repeatedly queried about the technical mechanisms and unusual electrodynamics of advanced mind control research, both in the West and abroad. Calling full attention to the special note above, in this paper we present a high-level over-view of the novel electromagnetic nature of mind operations, mind and body coupling, and *intent*•the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transduction mechanisms whereby differing EM wave polarizations can be transformed one-into-the-other are presented. We give two specific examples of lethal foreign military tests in 1997 of advanced mind control weapons against two military pilots over the central U.S. Some dangers and potential benefits of the emerging mind control technology are pointed out

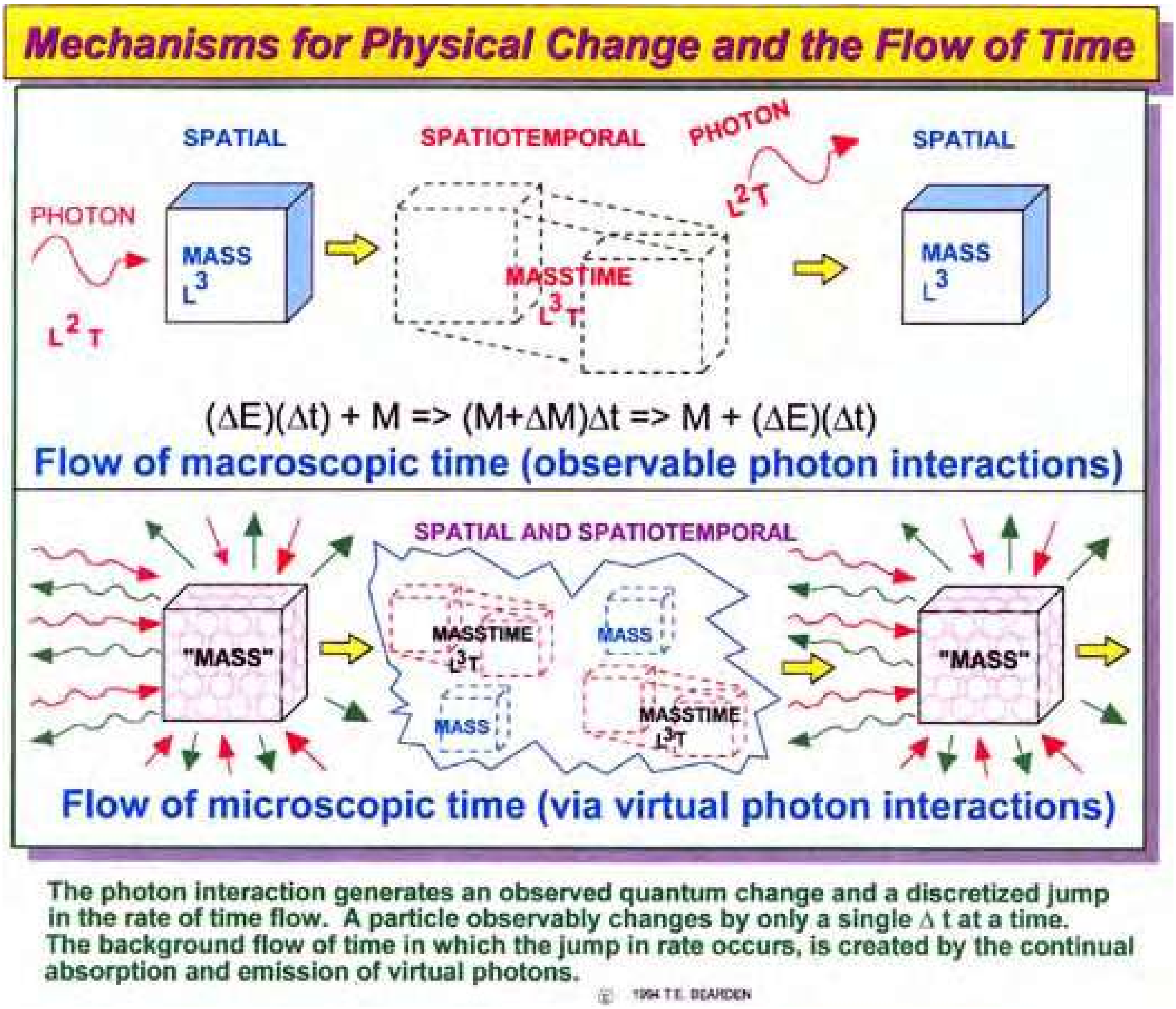
Mechanisms for Physical Change and the Flow of Time
SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS
L3
L2T
MASSTIME
L3T
L2T
MASS
L3
(ΔE)(Δt) + M => (M+ΔM)Δt => M + (ΔE)(Δt)
Flow of macroscopic time (observable photon interactions)
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L3T
MASS
MASS
MASSTIME
L3T
"MASS"
Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δ t at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Figure 1. Mechanism that generates a mass's flow through time.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Introduction and Background

In quantum field theory, there are four polarizations of photons.[1] Using 4-space and the z-direction as the direction of propagation, we have x- and y- polarizations where the 3-spatial energy of the photon is oscillating laterally, in the x- or y- direction. These are *transverse polarized* photons, as is any combination of the two. The third polarization is along the z-direction, which is a longitudinal polarization. In other words, the 3-spatial energy of the longitudinal photon cannot oscillate in the x- or y- direction, and so it is oscillating to-and-fro along the line of motion, z-. The fourth polarization occurs when the energy is frozen in all three spatial dimensions x-, y-, and z- and it cannot oscillate in any of those directions. In that case, the photon oscillates its energy in the t- direction, providing the t-polarized photon. We will later discuss how time is actually highly condensed energy.

In Minkowski 4-space, body operations are space-like, and are so treated in conventional materialistic physics. Mind and mind-operations are time-like, not space-like, even though they are totally electromagnetic in nature.[2]

As is well-known, all observation in physics is considered 3-spatial..[3] Mass is a 3-spatial concept, and we detect changes to mass (as, e.g., in the shift of electrons in the circuits of electrical instruments). So one may model the common physical observation mechanism as a *time-differentiating process* which *a priori* discards time-change and retains 3-spatial energy change.[4] That is, Minkowski reality is modeled in the fundamental units of L^3t. Physical observation (via the transverse photon interaction) is the process given by applying the operator $\partial/\partial t$ to L^3t, yielding an L^3 output. Hence mind and mind operations are excluded by the usual physics instruments and observation, which simply exclude the time domain in their outputs and do not "measure" it.

For this reason, physicists have erroneously considered mind to be "metaphysical" and nonreal. Indeed, most physicists to one extent or the other are materialists, and consider the "mind" to be nothing but the operations ongoing in a "meat computer." Nonetheless, in the time-domain the time is absolutely real, and it is completely electrodynamic in nature. There is no metaphysics involved, and the temporal domain—along with mind and mind operations—is simply an erroneously neglected area of physics. In present physics, the notion of mind is comfortably disposed of by imposing the use of the "observer" concept, with out ever specifying that the observer has a consciousness and a mind. Indeed, "observation" is only about what that "observer" perceives. Obviously, one has a dramatically crippled physics if one eliminates that nonobservable called "time." Similarly, one also has a dramatically crippled physics when one eliminates the mechanisms and physics ongoing in those time-like and dynamic "things" such as mind, that occupy time and function in it.

From this viewpoint, Western physics adheres to its 3-spatial measurement foundation only by ignoring the transduction of time-polarized EM changes into detectable longitudinal and transverse EM wave changes. In this aspect, present Western physics is severely self-crippled.

Time is totally electromagnetic and energetic in nature. The flow of time is not a separate external river on which a mass floats along like a boat drifting down the current of a great flowing river. Instead, the flow of time is generated directly on every mass by its total set of photon interactions, both virtual and ob servable. We have previously presented the exact mechanism for the flow of time. *[Figure 1]*

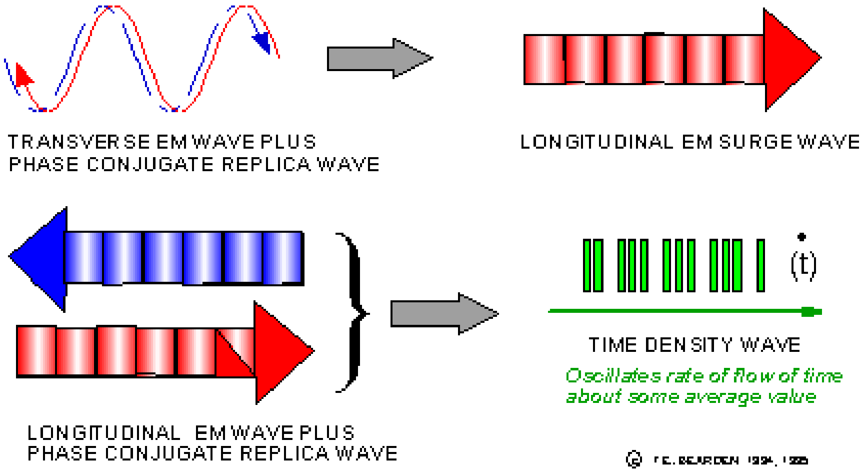
TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
LONGITUDINAL EM SURGE WAVE
LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
$\dot{(t)}$
TIME DENSITY WAVE
Oscillates rate of flow of time
about some average value
© T.E. BEARDEN 1994, 1995

Figure 2. Transduction of EM wave type by successive phase conjugate pairing.

Further, a photon is comprised of angular momentum, therefore of (energy)x(time). It transports both energy and time, not just energy alone. When absorbed by a mass, not only does it "energy-excitation" charge the mass, but also it "time-excitation" charges it, converting the former "mass" to "mass time". Rigorously it is *masstime* that emits a photon, not *mass*. So a mass moves through time in little spurts, by the continual macroscopic addition and subtraction of little Dt components. Further, in being driven through time, mass is continually altered to masstime (a quite different critter from mass), to mass, to masstime, to mass, etc.

In any masstime state, a myriad of tiny virtual photon interactions made of very tiny (DE)(Dt)'s interact with that same mass during that particular larger Dt of the masstime state (m+Dm)(Dt). Hence mass in its alternate masstime state has that state internally structured by its surrounding environment's interactions with it. The masstime state is internally structured energetically in its (DE) component, and also internally structured temporally in its (Dt) component.

A standard charged fundamental particle such as an electron, e.g., is not necessarily identical with another, when the internal structuring of its masstime state is considered. Further, by two papers by Whittaker,[5] interferometry of masstime states with either energy reactions or temporal reactions can yield observable effects and changes due to these neglected "hidden variables" in the electron's masstime alternative states. To engineer the mind and its operations directly, one must perform electrodynamic engineering in the *time* domain, not in the 3-space EM energy density domain. The direct engineering of time-like mind and mind operations—in all levels and all aspects—requires the use of *time-polarized photons and time-polarized EM waves*. This is the rarest form of electrodynamics, almost untouched by Western physicists.

One can either painfully produce such time-polarized photons and EM waves and directly irradiate a target mass with specific assemblies of them, or one can force the mass itself to iteratively transduce ordinary transverse EM waves first into longitudinally-polarized EM waves and then into time-polarized EM waves. Both will be discussed briefly.

In the West, it appears that the present author's discovery of mechanisms for producing time-polarized (scalar) EM waves [see *Figure 2*] and for transducing between wave polarizations has no precedent. [See *Figure 3*]. While scalar (time-polarized) photons are known in the literature,[1] the creation and use of time-polarized EM waves does not seem to appear in the Western physics literature.[6]

Wave transduction or *polarization transductions* are terms used by the author for the process of transforming an EM wave's particular polarization into another polarization type. Such transduction to other than transverse polarization forms is little known in the West. It appears essentially haphazardly in some experiments, usually without any recognition by the experimenters themselves. Significant transduction in experiments will also produce "strange" and unexplained anomalies in the instruments being used for experimental measurements.[7]

So most probably Western mind control researchers have not recognized the methodology and mechanisms[8] for transducing one type of EM wave polarization directly into another.[9] They continue to seek the "mind" in 3-space and hence in the brain, rather than in the time domain. On the other hand, it appears

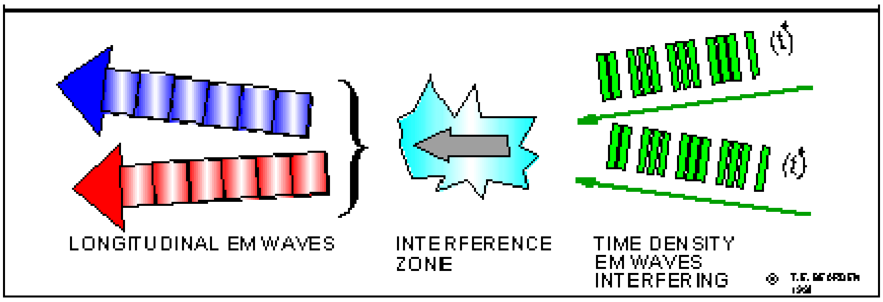

Figure 3. Interference processes for transducing higher EM wave polarizations into lower polarization states.

EXPLORE! VOLUME 9, NUMBER 2, 1999

that the Russian KGB energetics[10] weapons scientists—particularly those in psychoenergetics[11]—have known and used methods of transducing one wave polarization into another, for at least two decades. It follows that those same scientists have very probably developed mind engineering and mind control via novel time-polarized EM wave means and a dramatically extended electrodynamics of the time-like mind operations.

In the present paper we briefly develop the wave transductions and the basis for mind engineering, including the specialized use of ordinary transverse wave (TW) EM spectra to force internal EM wave transductions and time-domain operations inside irradiated bodies, cells, and tissues. Through the mind-body coupling mechanism, these transductions of transverse EM waves can operate upon the mind and its deepest operations as well as upon the body, every cell, and every part of every cell.

Russian Mind Control Uses Higher EM Wave Polarizations

It is apparent that the KGB psychoenergetics weapons scientists know and use the full extended EM wave polarization range. As a postulation, one would expect our own weapons scientists to know and utilize the orthodox transverse wave (TW) EM for similar research and experimentation. However, in the West scientists are just beginning to realize the importance of a very general (and weaker) type of longitudinal EM wave (LW) polarization, as evidenced by the appearance of dozens of papers on "undistorted progressive waves" (UPWs, which are essentially imperfect LW s with TW residues remaining). The Los Alamos National Laboratory web site has a number of such papers—particularly by Rodrigues[12] and Lu—available for free downloading. UPWs have very interesting characteristics. If they were perfect longitudinal EM waves, they would have infinite energy and infinite speed. Since only imperfect UPWs can be physically made, their speed can vary from slower than the speed of light to faster than the speed of light. Their energy can also vary over a great range.

Another characteristic of UPWs that are reasonably good longitudinal EM waves, is that they pass readily through a large depth of water and mass, including through the ocean and the earth with only small interactions and losses. Yet by interfering two such beams of "high quality" UPWs at a great distance, then in the interference zone ordinary EM energy will rise directly out of local spacetime potential, as essentially shown by Whittaker[13] nearly a century ago. Russian weapon research facilities have weaponized these effects[14] for nearly 50 years, under rigid KGB control and operation. Nonetheless, even if using only ordinary TW waves, Western mind control researchers may get some fairly good results, brute-force-like, by using gross correlates between just the input irradiating TWs and the exhibited behavioral responses of the individual. The net input-output correlations can be determined, even though not taking into account the actual wave polarization transduction mechanisms ongoing inside the irradiated dielectric (or brain, or mind, etc.). However, Western researchers appear to have no knowledge of the exact mechanisms by means of which coherent time-like mind operations of a biological organism couple to the organism's 3-spatial body to provide the mental control loop. They also appear unaware of how the coherent 3-spatial behavior responses of the body couple back to the time-like mind to provide it with a sensory feedback of the body's responses.

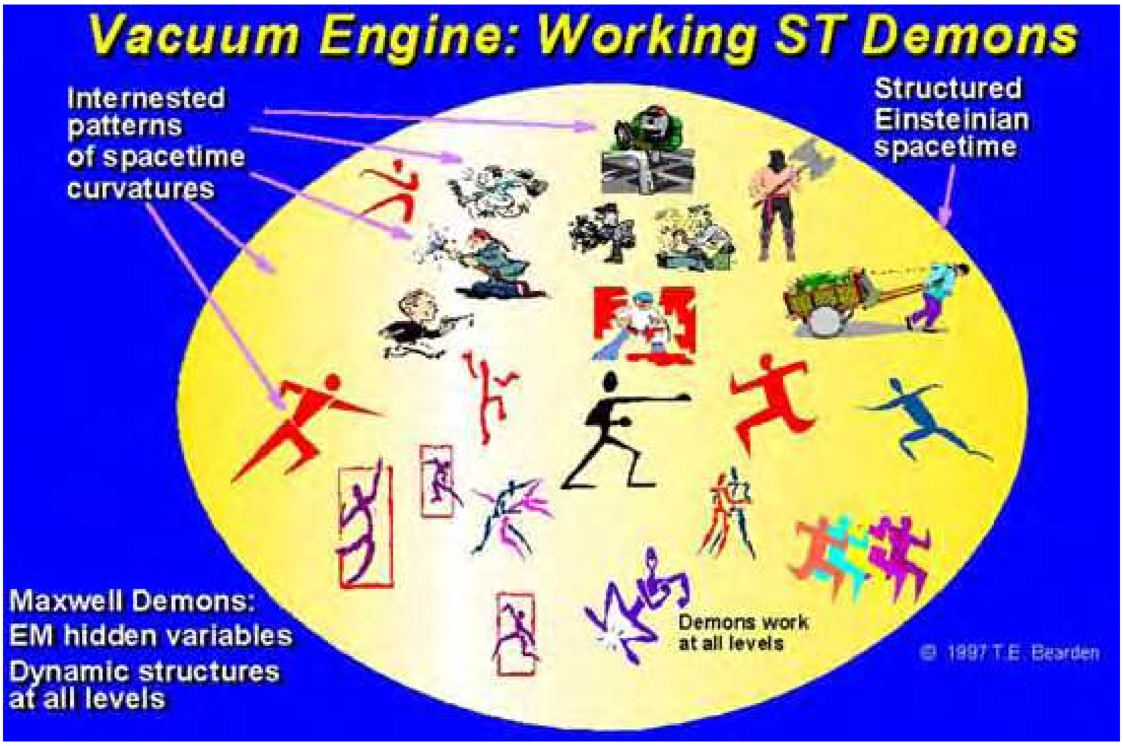

Vacuum Engine: Working ST Demons

Internested patterns of spacetime curvatures

Structured Einsteinian spacetime

Maxwell Demons: EM hidden variables

Dynamic structures at all levels

Demons work at all levels

© 1997 T.E. Bearden

Figure 4. A spacetime curvature engine has myriads of small ST curvature components which act upon mass at all levels.

EXPLORE! VOLUME 9, NUMBER 2, 1999

General Relativity Aspects

Both these coupling mechanisms can be taken directly from general relativity, if one puts one's mind to it. However, unless the wave polarization transformations are known, it is difficult or impossible to apply the ordinary GR *directly*, since relativists usually concentrate upon spacetime curvature by altering 3-space energy density rather than by altering time-polarized "time-energy" density.

In applied general relativity, spacetime itself is an active medium. One speaks of a change in spacetime as a "spacetime curvature". Any curvature of spacetime in a local region directly performs continuous work upon any mass embedded in that region. Vice versa, any mass (or other concentration of energy) in a local spacetime continuously acts upon that spacetime to "curve" it.

So if one wishes to perform actions upon mass in a distant region, one may generate local curvatures of spacetime in that region, and these curvature "engines" will directly act upon the mass at all levels. *[Figure 4]* This is very different from energy propagation through space from one point to another. Now the "ordinary EM energy and actions" arise from every spatiotemporal point within the mass, at every level, and move upwards (from inside to outside). So we speak of specific forms of "templates" of nested spacetime curvatures formed and utilized to engineer mass "from inside out" as *spacetime curvature engines or vacuum engines*.

Use of vacuum (spacetime curvature) engineering[15] is a far more powerful form of engineering than is provided by energy propagation through space. As an example, it is easy to alter the quarks in a nucleon, using spacetime curvature engines and time-charging decay. Indeed, we have developed the mechanisms for cold fusion and the electronuclear interaction (formation of new nuclides) at feeble energy. An Invention Disclosure[16] on this work has been filed with the U.S. Patent Office, and formal patent applications are in preparation. Some information on these mechanisms and principles has been released[17].

Time as Dense EM Energy and a Strong Spacetime Curvature Agent

The advantage of using the time-polarized "time-energy" for spacetime curvature is that time is ordinary energy compressed by a factor of at least c^{2}—which, in the MKS system of units, is some 9¥10^{16}. Thus use of time-polarized EM photons and waves as ST curvature agents gives an amplification of 9 ¥10^{16} over the use of transverse-polarized EM waves for that purpose. In turn, the use of the strong EM force in ordinary TW waves as an agent of ST curvature is already a nominal 10^{40} times as strong as is the weak G force used as the agent of ST curvature.

The bottom line is this: For spacetime curvature effects, the use of the t-polarized domain provides amplification of some 9¥10^{56} greater than the weak G-force ST curvature agent usually considered in general relativity[18].

Western Science Remains Largely Materialistic

Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system.[8] This serious self-limitation exists because in the body we *measure* only weak TW EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.

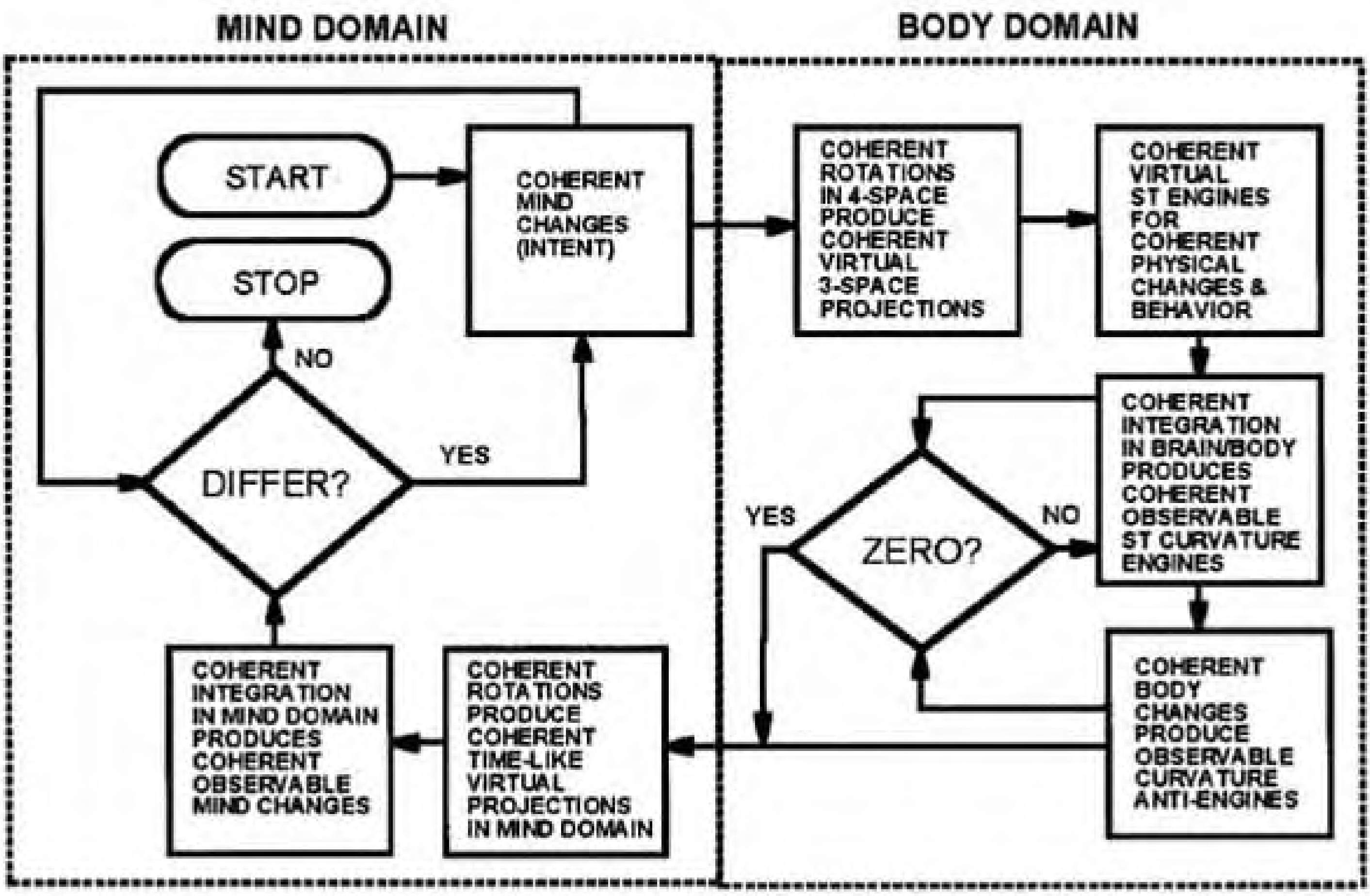
MIND DOMAIN
BODY DOMAIN
START
STOP
COHERENT MIND CHANGES (INTENT)
NO
DIFFER?
YES
COHERENT INTEGRATION IN MIND DOMAIN PRODUCES COHERENT OBSERVABLE MIND CHANGES
COHERENT ROTATIONS PRODUCE COHERENT TIME-LIKE VIRTUAL PROJECTIONS IN MIND DOMAIN
COHERENT ROTATIONS IN 4-SPACE PRODUCE COHERENT VIRTUAL 3-SPACE PROJECTIONS
COHERENT VIRTUAL ST ENGINES FOR COHERENT PHYSICAL CHANGES & BEHAVIOR
COHERENT INTEGRATION IN BRAIN/BODY PRODUCES COHERENT OBSERVABLE ST CURVATURE ENGINES
YES
ZERO?
NO
COHERENT BODY CHANGES PRODUCE OBSERVABLE CURVATURE ANTI-ENGINES

Figure 5. Mind-to body coupling and body-to-mind coupling.

EXPLORE! VOLUME , NUMBER 2, 1999

With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.

Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW functions exist in the body and in all of nature in general. Presently only highly theoretical quantum field theorists seem to even be aware of the existence of scalar (t-polarized) photons, and even they are unaware of t-polarized EM waves. Certainly our scientists do not measure t-polarized EM waves in and around the body and in nature, nor do they make them in the laboratory. Apparently they have not studied such waves and their interactions with matter—living and inert—at all.

But it's even worse. In ignoring the time component transported by photons and EM waves, science has also erroneously omitted half of the excitation charging and excitation decay processes whenever a mass interacts with photons and EM waves. More on that later.

Brute Forcing Time Functions versus Fine Control Methods

Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.

While KGB scientists also use TW EM "brute force" TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired. They also do not hesitate to use LWs and t-polarized waves overtly. So in blunt terms, Western compared to Russian mind control research is probably like comparing an automobile body shop with a fine surgical ward. Having long ago worked out all those "exact correlates" to the internal "information content of the field" is where the KGB scientists are at least 20 years ahead of the West.

The reason one can get mind and behavior results with the TW electromagnetics, while bypassing the real mechanism which uses t-polarized waves and photons, is due to (1) the very peculiar nature of the EM emission from a dielectric, and (2) what can be done by re-radiating that dielectric with its emitted spectrum, deliberately and very carefully altered in selected parts. We will return to that important feature later.

Polarization and Observability

As we stated, there are four photon polarizations and therefore there must be correspondingly four EM wave polarizations[1] The first three polarizations are the x-, y-, and z- spatial polarizations. The x- and y- polarizations are transverse polarizations and the z-polarization is a longitudinal polarization along the direction of propagation (along the z-axis, by standard notation). Simply put, we may visualize the transverse polarizations as rather like the wiggling of a fish's tail from side to side as the fish moves forward, or a whale's flukes up and down as the whale moves forward, or some combination thereof. We may visualize the longitudinal polarization as a sort of "repetitive accordion effect" along the line of motion of the wave. Usually the z- polarization is neglected in EM wave theory, although in recent years physicists have "rediscovered" longitudinal EM waves and are now intensely researching the use and characteristics of such waves.[12]

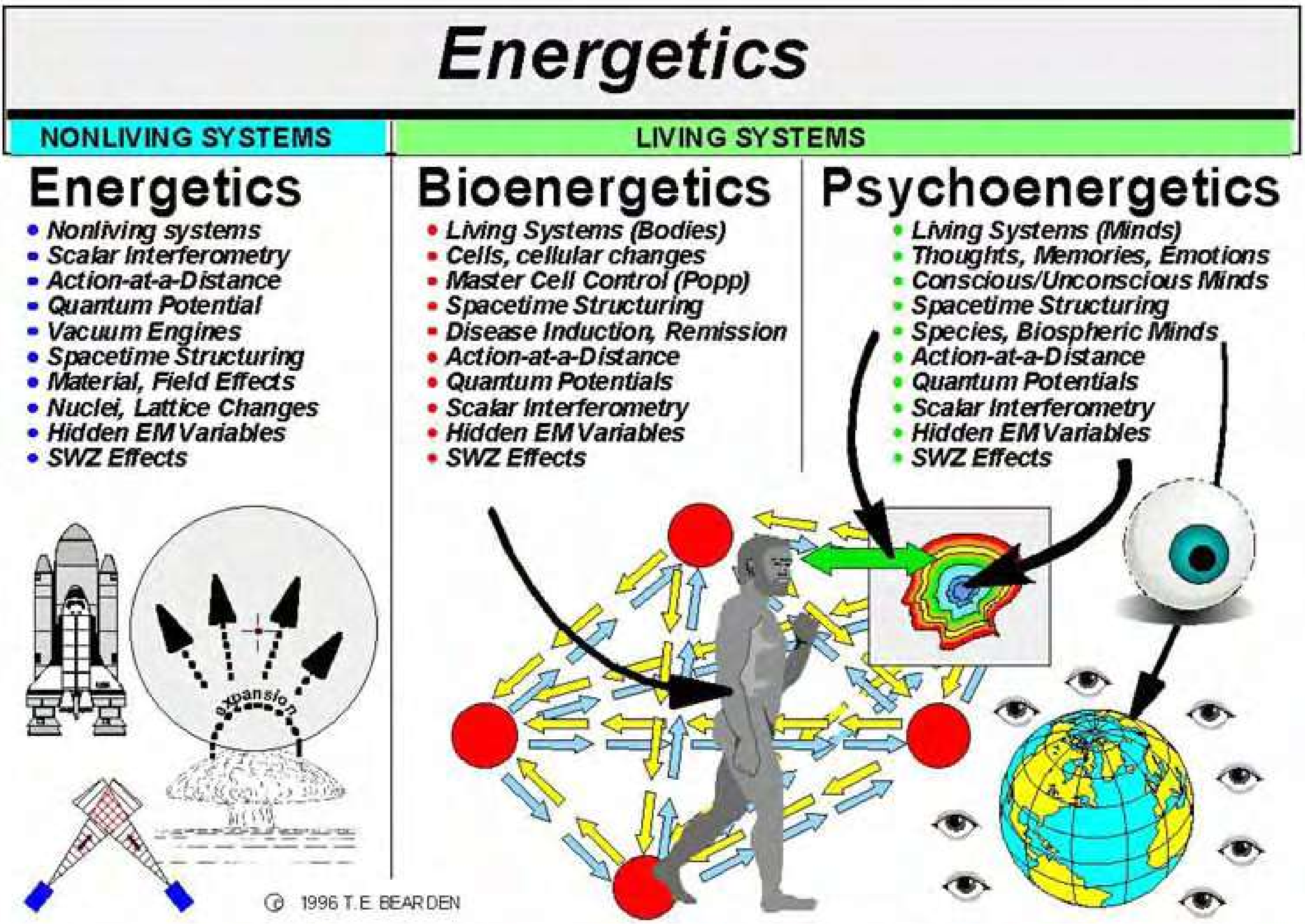

Energetics

NONLIVING SYSTEMS

LIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T.E. BEARDEN

Figure 6. The three divisions of Soviet energetics and their characteristics.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Unfortunately, in quantum field theory there has been a tendency to regard the t-polarized or "scalar" photon—where the local rate of time is oscillated—as unobservable. Oscillation of the local rate of time produces powerful oscillations of local space-time curvatures, due to the extreme energy density of time. Individually, the t-polarized photon tends to be unobservable. However, a *coherent sequential group* of such t-polarized photons, each individually in the virtual state with respect to the external observer, may simply integrate in its interaction with mass over a short time period into observable change because of the coherent integration of virtual spacetime curvatures into a larger, observable spacetime curvature.

Rotations and Projections in 4-Space

Relativistically, any velocity change in 4-space is a rotation. Any such rotation in the time-domain also creates a tiny projection component into 3-space. Any such rotation in the 3-spatial domain also creates a tiny projection component into the time domain.

Hence a sufficient series of coherent time-domain (mental) changes produces a coherent series of virtual changes in 3-space (the body), thereby coherently integrating into an observable energy change in the body. A sufficient series of coherent 3-spatial energy changes produces a coherent series of virtual changes in the time domain which coherent integrate into an observable time-domain change.

Solution to the Age-Old Philosophical Problem of Intent

This is in fact the solution to the age-old problem of *intent*, or how the nonmaterial (i.e., non-3-spatial and non-observable) mind induces a 3-spatial, observable energy change upon the 3-spatial body. It is also the solution to the problem of *awareness*; i.e., how the mind is aware of itself[19] and of the responses of the body. Together the two form a closed loop coupling of the mind and body. The time delay in the loop together with memory recall for comparison, creates the sense of "persistence" of self in time. This also creates the sense of being a "separate, closed being" (i.e., of separate persistence in time—in the living entity). The sense of "separation of self from an external world" is created by comparing those body-to-mind sensory feedbacks which are not correlated to the mind's previous feed-forward intent.

Time As Energy and Why It Is Very Dense Energy

In addition to the three spatial polarizations of photons and EM waves, there is a very, very useful t-polarization along the time axis. In this polarization, the 3-spatial energy is not oscillating at all. Instead, the time or time-energy is oscillating. Time can be taken to be energy compressed by at least c^2, so it has at least the same energy density as mass. In other words, one second is $9x10^{16}$ joules of time-energy (energy compressed into time). The t-polarized photon or EM wave is called the scalar photon or scalar EM wave, respectively.

To demonstrate why time can be regarded as energy, we need only point out that the choice of fundamental units in any physics model is totally arbitrary. We usually choose these units to ease the mathematical manipulation, ease our understanding, and simplify the ability to visualize or "grasp" the physics. However, if one wishes, one can build all of physics from a single fundamental unit—e.g., energy. In that case, one readily sees that time is a function of energy and only energy. Hence it is perfectly proper to regard time as energy, and to seek out what form the function takes in our normal system of units (say, the MKS system).

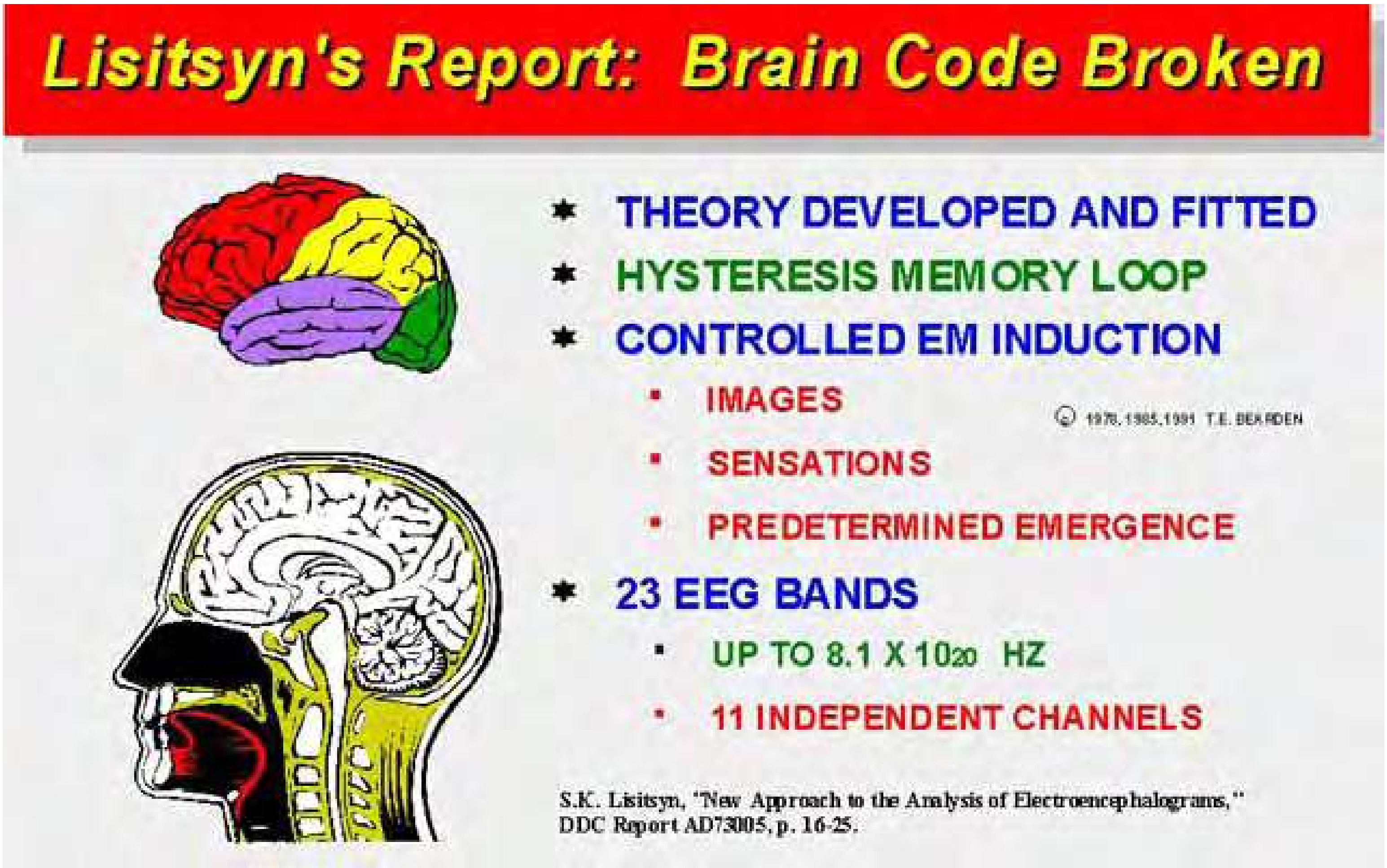

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

Figure 7. In the 1960s, Lisitsyn revealed Russian scientists had deciphered the human brain code.

EXPLORE! VOLUME 9, NUMBER 2, 1999

In that system, it appears that any change in time Δt can be taken as always a function of a corresponding change in energy ΔE, where $\Delta t \le (\Delta E) \div c^2$ [1] from which it follows that $\Delta E \ge (\Delta t) \times c^2$ [2] For convenience, we take the special case where $\Delta t = (\Delta E) \div c^2$ [3] $\Delta E = (\Delta t) \times c^2$ [4].

Observation as Used in Physics Is Spatial

We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain, and thus "lost" by the stripping away of time in the ordinary observation process. Physical observation is essentially a time-differentiating mechanism applied to a 4-spatial change, or in terms of MKS fundamental units L and t, observation θ is $\theta = \partial/\partial t(L^3 t) = L^3$ [5].

Thus, as is well-known in quantum mechanics, physical observation is 3-spatial, and time is not a physical observable, even in theory. Since mind is time-like, it follows that mind is not a physical observable either, since discarding the time dimension also discards the mind. In short, one may also take physical operation as the mechanism that separates mind and body. To observe is to separate. We point out, however, that merely "separating and discarding" the time domain (as in physical observation of a single change) does not eliminate it, nor does it eliminate the single time-like mind change that may be involved in an intentional volition. The time domain certainly remains, even though only the 3-spatial intersection of the 4-spatial Minkowski change is given by physical mechanism and the resulting "outputs". So a series of coherently integrated "mental intent" changes introduced into the human body's overall servo-mechanism provides the continuing input. From there, ordinary physics will generate the resulting actions induced in the body by that servomechanism and its amplifying mechanisms.

Two Coupling Mechanisms Make a Closed Loop

This is the "mind-to-body" coupling mechanism. [See *Figure 5*] It is the mechanism whereby the mind is coupled to the body, and whereby mental intent is able to induce a series of physical inputs into the body servomechanism.

So the body's servomechanism then generates the responses of the body (including everything from chemistry to electrical changes to muscular movements, etc.). These responses are changes in 3-space.

The conscious mind is a *serial processor*, though extremely rapid. It produces the series of coherent intent inputs for volitional behavior.

The unconscious mind is a *massively parallel processor*.[20] It continually produces the vast series of coherent "unconscious intent inputs" for control of all the deeper processes in the body, beyond usual conscious awareness.

However, as the body moves or changes in 3-space, each resulting quantum change in body 3-space is also a slight rotation out of 3-space and toward the time axis. Hence it induces—in the "virtual state" in the time domain—a precisely correlated projection. The body's responses are in general coherent, so a coherent series of virtual state changes in the time-domain (in observation. So we may say *that the time-change remains in the virtual state, with respect to attempted physical observation of a single Minkowski 4-space change.*

Rotation, Coherent Integration, and Intent

Previously we pointed out that any change in a 4-space entity may be regarded as a rotation away from the "trajectory" of the entity. Hence each and every t-polarized change creates a very small projection into 3-space by its rotation slightly away from the time axis. All mind changes in the time-domain actually produce virtual 3-spatial projections in 3-space (in the physical domain). We define "intent" as the continued production of successively coherent mental changes, producing coherent virtual changes in the 3-space body system, with coherent integration of those 3-space virtual changes into observable 3-space energy inputs into the 3-space body system. Successively coherent mind-changes will produce successively coherent 3-space virtual changes. In short, coherent mind-changes will produce coherent integration of those 3-space virtual projections into an observable 3-space change. This is the creation of an ordinary 3-spatial energy change induced upon the 3-space body. In short, this is the mechanism whereby time-like mental intent is able to produce a series of coherent observable quantum changes in the physical body (as in the brain and nervous system).

The Body as a Servomechanism

From the standpoint of control theory, we may consider the body system to be a complex servomechanism system comprised of many subordinate servo systems with feedback and feedforward the mind realm) are created successively. In short, again we have coherent integration, this time in the mind or time domain. This produces "observable" changes in the time-mind domain, which are coherent with the body's 3-spatial changes actually performed. Thus the mind receives feedback directly from the physical movements and changes of the body. This is the manner in which the body is coupled back to the mind.[21]

By comparing the "intent" behavioral move that was "fed forward" into the body, with the return "response" move analog that was "fed back" from the body to the mind, the mind is able to determine errors and differences, and originate additional correctional commands.

Thus the entire mind-body loop is a closed-circuit system of feedforward and feedback, together with corrections. It also has multiple levels of such, infolded in the larger volitional levels. See again *Figure 4*.

This solves the age-old problem of the mechanisms for the mind-body coupling, intent, volition, conscious and unconscious functioning, sense of the external world, sense of the internal world, sense of "being in" the external world, etc.[22]

All This Is Included in Russian Psychoenergetics

This is the highly summarized basis for *psychoenergetics*, the KGB's division of energetics that deals with the mind and body coupling and functions, and direct engineering of (1) the mind-body coupling and (2) the mind operations directly. [See *Figure 6*] The KGB intent, of course, has always been to exploit this science for the degradation, killing, and control of human beings, including all humanity.

The Russians know full well that, if you looping.

Given the input from intent, ordinary servo theory will take it from there. Once a servo has an input, servomechanism theory describes the response of the

produce and utilize scalar EM photons and t-polarized EM waves, you can directly affect and engineer mind and mind operations at any and all levels. Western researchers, who know nothing of how to make

EXPLORE! VOLUME 9, NUMBER 2, 1999

time-polarized photons and time-polarized EM waves, do not yet know that. Consequently the Russians have developed a highly secret science of *directly* engineering the mind and its operations, including thought, images, perceptions, feelings, emotions, memory, and mind processing. Indeed, Lisitsyn[23] wrote quite specifically of this capability in the 1960s. [See *Figure 7*]

Western clandestine mind control researchers are apparently still slowly and painfully fitting TW EM irradiation correlates to induced or resulting mental and physical behavior responses. They seem unaware of the actual wave transductions occurring inside the body and mind, but are unwittingly inducing those transductions in hidden fashion anyway and in "brute-force input-response fitting" models.

These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design—for example—of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gaia's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.

This fine research article will be continued in Part II with Russian Methodology, Waves and Wave Transduction, The Cellular Control

spacetime that is highly active. The mind is rooted in the time-domain and projects from that domain into 3-space. The body is rooted in the 3-space domain and projects into the time-domain. To "change" or "function" in one domain is automatically to function in the other. EM and gravitational phenomena are still modeled separately by human scientists, but not by nature.

3. In quantum mechanics, time is not an observable, but merely a parameter.

4. It is convenient to consider mass as a 3-spatial form of condensed energy.

5. E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics, " Mathematische Annalen, Vol. 57, 1903, p. 333-355;— "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions, "Proc. Lond. Math. Soc., Series 2, Vol. I, 1904, p. 367-372.

6. A NERAC document search yielded some 16 good references on t-polarized photons, but not a single reference on t-polarized EM waves.

7. E.g., such previously inexplicable instrument anomalies have accompanied some excellent and rigorous electrolyte experiments at U. S. Navy research facilities at China Lake. For a description, see Melvin H Miles and Benjamin F. Bush, "Radiation measurements at China Lake: Real or Artifacts?", Proc. ICCF - 7 (International Conference on Cold Fusion—7, Vancouver, BC, Canada, Apr. 1998, p. 101. For a brief explanation of the anomalies, see T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves, " Explore!, 8(6), 1998, p. 7-16;— "Toward a Practical Unified Field Theory and a Deep Experimental Example," Proc. INE Symposium, Univ. Utah, Aug. 14-15, 1998.

8. Many U. S. researchers and journalists—and even many scientists—have wrestled with the problems of mind, intent, and mind-body coupling. Enumeration of those efforts would itself require an entire book. For an introduction into that domain, the reader is referred to the very determined exposé by

System, and other matters of interest. ♦

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel US. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, US. Army. Graduate of Guided Missile Staff Officers Course, U. S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses. Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor: He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuums virtual particle flux is organized and used to energetically shape and manipulate matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar

Cheryl Welsh in her very timely book, The 1950's Discovery of the Code of the Brain, May 1988, published on the Internet. Simply employ any Net search engine and search on the name Cheryl Welsh. A connection to the site and to Welsh's book will immediately be found Welsh's important compendium is highly recommended, giving the reader a "birds eye view", so to speak, of how Western scientists think and proceed regarding the subject of mind control by electromagnetic means. Unfortunately all journalists so far researching the area have not been aware that Western electrodynamics itself is seriously flawed, and that the great Russian advances in mind engineering and mind control are a result of their secret but complete revision and correction of Western electrodynamics to provide the basis for energetics. Until Western scientists revise their own decrepit old electrodynamics fouling the textbooks, they will never catch up to the Russian mind control developments springing from intense development programs that did that revision in the late 1940s and early 1950s.

9. An Invention Disclosure on these processes and enhancement embodiments has been filed with the U. S. Patent Office. Formal patent applications are also in preparation.

10. Energetics is a unified science of an extended electrodynamics possessing a hidden infolded general relativity inside the potentials, fields, and waves. It deliberately employs higher polarizations (longitudinal and tempic) of photons and EM waves to engineer action at a distance, use of specific patterns of spacetime curvature created at a distance, and direct engineering of either inert or living bodies. It also encompasses the direct engineering of time-like mind and mind functions. It is divided into three divisions, depending upon the nature of the target. Targeted against inert materials, structures, fields, and waves it is called "energetics. "Against living bodies, their fields, and biochemistry, it is called "bioenergetics." Against the mind and mind functions, it is called "psychoenergetics".

11. Psychoenergetics is that branch of energetics used to directly affect and engineer

electromagnetics processes.

REFERENCES

---

1. E.g., see Lewis H Ryder, Quantum Field Theory, Cambridge University Press, 2nd Edition, 1996, p. 147+. For an even deeper discussion of photon polarization, see F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, under the heading "Covariant Quantization of the Photon Propagator" in Chapter 5. Mandl and Shaw argue that the longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e. electrostatic) potential.

2. Nature is unified and not separated into the "separate" disciplines our science has prescribed. Thus the mind is also totally electrogravitational. The mind works not just with EM forms in an inert space, but also with spacetime curvatures in a the mind, memory, perception, thought, emotion, sensation, personality, and other mind functions of a living biological organism. It is usable against a single personal conscious (serial processor) mind and a single personal unconscious (massively parallel) processor mind In theory it is also usable against lungs collective unconscious mind and its functions, as well as the deeper unconscious for all species on Earth (i.e., against Gaia). We are speaking of physics, not metaphysics.

12. W A. Rodrigues Jr.. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 \leq v < \infty$ in nature," Foundations of Physics, 27(3), 1997, p. 435-5•08 A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory website.

13. E. T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions, "Proc. Lond. Math. Soc., Series 2, Vol. I, 1904, p. 367-372. The paper was published in 1904 and orally delivered in 1903.

14. E.g., see T. E. Bearden, Fer-de-Lance, Tesla Book Co., 1986;—AIDS: Biological Warfare, Tesla Book Co., 1988;—Soviet Weather Engineering Over North America, 1-hour videotape, 1985,—Gravitobiology; A New Biophysics, Tesla Book Co., P.O. Box 121873, Chula Vista, CA 91912, 1991,—Energetics of Free Energy Systems and Vacuum Engine

EXPLORE! VOLUME 9, NUMBER 2, 1999

Therapies, Tara Publishing, Internet node www.tarapublishing.com/books, July 1997; — Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ pages furnished to selected government agencies and key Congressional committee leaders.

15. E.g., see T.D. Lee, Particle Physics and Introduction to Field Theory, Harwood; New York, 1981. On P.383, Nobelist Lee points out that the microstructure of the scalar vacuum field (i.e., of vacuum charge) is not utilized. Particularly see Lee's own attempt to indicate the possibility of using vacuum engineering, in his "Chapter 25; Outlook; Possibility of Vacuum Engineering, " p. 824-828. With wave transductions and use of higher polarization EM waves, there are far better and more powerful ways to do vacuum (spacetime curvature) engineering.

16 T.E. Bearden, "Formation and Use of Time-Reversal Zones, EM Wave Transduction, Time-Density (Scalar) EM Excitation and Decay, and Spacetime curvature Engines to Alter Matter and Convert Time Into Energy," Invention Disclosure Document #446522, Oct. 26, 1998.

17. T.E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves, "Explore!, 8(6), 1998, p. 7-16; — "Toward a Practical Unified Field Theory and a Deep Experimental Example," Proc. INE Symposium, Univ. Utah, Aug. 14-15, 1998.

18. We have used the enormous energy density of time in our approach to a totally new kind of electronuclear interaction at feeble spatial energy. By use of a time-reversal zone (TRZ) in a fluid; the free H+ ions—which are free protons—can be induced to attract each other so tightly that the edge of the nuclear force region is entered. In that case the protons form a quasi-nucleus. When the TRZ decays back to a time-forward zone (TRZ), this new quasi-nucleus decays by changing quarks in one or more nucleons, thereby producing new nuclides. We point out that means "sense of existing in time".

20. The "unconscious" is totally conscious, just multiply so. The conscious mind can see only "a single slide in the slide projector at a time, " so to speak. When it "looks" at the ongoings and outputs of the unconscious, it "sees" millions of slides in the slide projector simultaneously. Hence it cannot distinguish "one and one only". So it just sees nothing, where by "nothing" we mean "no single distinguishable thing." It therefore is unable to consciously perceive the actions of the unconscious. When the unconscious wishes to communicate with the conscious (because of unresolved conflicts, etc.), it must produce large groupings integrated into "symbolic scenarios". Hence the symbolism of the unconscious, and the basis for therapeutic psychology and psychiatry. A "symbol", after all, is just a thing which can have many different meanings, all folded into one.

21. The mind continually has activities of its own, each of which produces coherent changes in the body to a much smaller degree. The interweaving of these "smaller self-changes" infolded inside the 3-space major intent changes also produces slight body changes, which in turn "project back" into the mind. This huge group of "infolded self-loops" produces the sense of "my body" being "volitionally directed or occupied" simultaneously with the "greater intent" loops which represent "my intentional actions with my body". Further; the body servo continually receives a myriad of inputs from its sensory systems, both from changes inside the body and from incoming changes from the external universe. This interweaving produces the awareness of "being in the external world" by having (i) changes whose feedback do not match mind-intent changes, and (ii) changes whose feedback does match mind-intent changes. This of course applies to both the conscious and unconscious levels. The purpose here is to demonstrate that there exists an actual physics of the living being, the mind as well as the body, and the coupling and interfiinctioning of mind and body.

this is actually a very high energy reaction; however; since the time-energy is used and is very, very dense, the amount of 3-spatial energy that must be used is very small. If this new mechanism holds, it represents a revolution in particle physics.

19. "Self-awareness" refers to be continually "refreshed" by time-delayed coherent feed-back inputs from the body to the mind; where these inputs compare to the recalled memory of the minds inputs to the body. In short, the time-delay coherence creates the sense of "existing in time". Awareness usually

22. The methodology also easily extends to include deeper interlays for the collective human unconscious mind and for Gaia—the collective unconscious mind level for all the species on earth.

23. S. K Lisitsyn, "New Approach to the Analysis of Electroencephalograms, " in Problems of Bionics— (Selected Articles), Defense Documentation Center Report AD 73005, date unknown, p. 16-25. Available through the National Technical Information System (NTIS).

Mind Control and EM Wave Polarization Transductions, Part 2

©Copyright 1999 by T.E. Bearden, USA

This article continues on with T.E. Bearden's article, Part 1 of which was published in Volume 9, #2 of Explore! for the Professional. Because we feel that Col. Bearden's caution below is so very important we are repeating its publication in all parts of this article.

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for non-approved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

A Strategic Threat Completely Unrecognized by the West

The Russian methodology leads directly to rather mind-bending, extended capabilities. KGB psychoenergetics scientists seem fixed on an eventual goal of directly engineering Jung's collective human unconscious, thereby converting the entire human species into a sort of "ant" species. In short, they seek to engineer what they regard as "perfect communism and perfect order" directly in the entire species. *Figure 8* indicates the direct connection of every human's conscious and unconscious mind to the collective species unconscious and to Gaia, the collective unconscious for all species on Earth. Since time is internally structured and layered, it follows that time-like mind levels are also internested and layered.

This appears to be a significant strategic threat, *well along in its development*, that our scientific and intelligence communities have completely failed to recognize.

The direct engineering of the entire human collective species is a doable.[1] It is going to be done, whether we like it or not. In that race (which may eventually determine whether or not we retain any personal freedom at all), the "democratic" nations are woefully behind. In fact, the West has not even realized there is such a new kind of "arms race" for dictatorship and control of the "collective species mind" and we have already almost lost it. The West has not yet even developed the overall theory of psychoenergetics.

Waves and Wave Transductions

A pure longitudinal EM wave (LW) would have infinite energy and infinite speed. In practice, one can only approach this perfect LW wave, and one always has transverse residues remaining. Hence the "imperfect" longitudinal EM waves we can make in suitable equipment and techniques, are known as "undistorted progressive waves" (UPWs). The UPWs may have velocity far greater than the speed of light, or less than the speed of light. So the speed v of UPWs can vary as

$$0 < v < \infty \qquad [6]$$

The Russians also know that, when a nearly-pure longitudinal EM wave interacts with matter (it usually doesn't interact very much, and has a low reaction cross section), the mass immediately adds the phase conjugate replica to the interacting LW portion, regardless of frequency, thereby transducing the input LW interactions into scalar (time-polarized) EM interactions.

One transduces from TW to LW to Scalar EM via iterative phase conjugation. One transduces from scalar EM to LW to TW via iterative interferometry. It took the present author 20 years to discover this simple thing; we now have filed an invention disclosure on the process and embodiments for performing it, and will be following with formal patent applications. Let TW = transverse EM wave, LW = longitudinal EM wave, and TDW = time-density EM wave (another name I use for t-polarized or scalar EM waves, since the density of time-energy is what is varying). Let (I) = interfering with, PCR = phase conjugate replica wave with a coupled set of parentheses enclosing the type of wave. The rules for wave type transduction are:

$$\text{TW} + \text{PCR(TW)} \Rightarrow \text{LW} \qquad [7]$$

$$\text{LW} + \text{PCR(LW)} \Rightarrow \text{TDW} \qquad [8]$$

Going the other way, the rules are:

$$\text{TDW (I) TDW} \Rightarrow \text{LW} \qquad [9]$$

$$\text{LW (I) LW} \Rightarrow \text{TW} \qquad [10]$$

In EM smog (dense EM signals, even though very weak), there is increased nonlinear multiwave interactions, both interferometric and phase conjugative. So in such EM smog there exists an increased, non-negligible component of LW and TDW irradiation, formed by the above transduction interactions in the irradiated living cells and bodies. This latter part accounts for many of the long term effects which show up in "correlative" studies, but which do not show up in the normal limited-frequencies TW experiments produced by so-called "physical experiments."

The standard EM bioeffects manner of experimenting uses a theoretical model of simple, direct energy deposition by means of absorption in cells and tissues. Now we must utilize an "open system" throughout the body dielectric in which higher polarized EM energy (either LW or t-polarized) travels readily throughout all areas of the body.

EM Bioeffects and Wave Transductions

In the EM bioeffects experiments to date, there has been absolutely zero control of the internal transductions, and there has been no consideration of them. Almost all transductions of TWs into LWs will immediately result into transduction into TDW and time-excitation charging of the cells and all their internal components. These time-excitation charges (excited states of time-charging) will then slowly decay, emitting longitudinal EM waves in the process which travel throughout the body. These emitted LWs interfere with each other and the LW processes in the body, producing low level TW "noise jamming" of the cellular processes and—most especially—of the deep cellular control system of the body. The latter system functions by means of LWs and TDWs. Long-term effects of such "noise jamming" of biochemistry reactions and the control systems of the cells and body, can lead to degenerative diseases such as arthritis, atherosclerosis, Alzheimer's disease, heart disease, cataracts, tumors, leukemia, somewhat accelerated aging, immune system weakening and depressing, etc. As the body's natural defense and regeneration ability weakens, opportunistic infections further add to the degeneration.

It was shown by Whittaker in 1903 that any scalar EM potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs. [See *Figure 9*] Each longitudinal EM wave pair in the harmonic series is a phase conjugate pair. Unknown to Whittaker, that "LW phase conjugate pair" is actually a time-polarized EM "photon" structure, containing a phase conjugate LW photon substructure. So staid old "voltage" (potential) is comprised of a harmonic series of time-polarized EM waves!

The corresponding "t-polarized photons" are actually spin-4 *supergravitons* comprised of coupled antipairs of LW "photons". Each LW photon is a spin-2 *graviton* comprised of a coupled TW photon-antiphoton pair. Each TW photon and antiphoton in the coupled pair has spin-1. The couplet, being the longitudinal "photon" or graviton, thus has spin-2. And so on.

In 1904, Whittaker showed that any EM field or wave pattern can be decomposed into two scalar potential functions. By applying Whittaker 1903 decomposition to each scalar potential function, any EM field or wave has an enormously rich internal structure, comprised of scalar potential functions in interference. Each scalar potential function is comprised of time-polarized EM waves (made of spin-4 supergravitons). Each time-polarized EM wave is further comprised of bidirectional longitudinal EM wavepairs. Each longitudinal EM wave is further comprised of phase-conjugate pairs of transverse EM waves. Each transverse EM wave is further comprised of two scalar potential functions, each of which is in turn comprised of an infinite harmonic series of time-polarized EM waves, and so on *ad infinitum* and *ad nauseum*. Obviously all EM potentials, fields, and waves have an enormous "internal structure" or internal "information content of the field".

The Cellular Control System Uses t-Polarized EM Photons and Waves

The master cellular control system (MCCS) in the body utilizes these photon-gravitons, their corresponding EM wave polarizations, and transductions between them. By the use of very powerfully condensed gravitons and supergravitons, it is able to utilize small but highly significant direct curvatures of spacetime in precise patterns. In short, it utilizes general relativistic spacetime curvature engines, rather than "EM photon signals", to accomplish its control of the cells and their processes. It does not merely send a "signal" and hope the cell itself acts upon it. Instead, it sends an "engine" which alters the local spacetime in which the cell and all its components are embedded. This "local ST curvature engine" then physically works directly upon the cell and all its parts, to alter the cell according to the specific "template" of the engine.

This is the direct application of the Wheeler's simple statement of general relativity: *Mass (trapped energy) acts upon spacetime to curve it, and curved spacetime acts upon mass to change it.*

Popp's "ordinary photon" study of this cellular control system[2] therefore will detect the coherent ordinary TW photon residues which result from the photon polarization transductions occurring from the actions of the ST curvature engines upon the cells.

Cellular division, biochemical processes, growth and growth control, cell differentiation and dedifferentiation, etc. are all controlled by the MCCS or one of its subsystems, using spacetime curvature engines. The ordinary TW electronic activity our biophysicists detect and correlate are simply the intermediate components of the actual ongoing polarization transductions.

The Cellular Control System Uses EM Time-Energy

One important subsystem of the MCCS controls the immune system's functions, including the deepest and most subtle functioning performed by all elements of the immune system. This system has been rather intensively studied, but since our biophysicists have not possessed the full electrodynamics utilized, only the correlated residue of normal TW EM waves and photons have been recognized and studied. Hence Western biophysics is in its very crude infancy.

Incorporation of the total wave and photon polarizations, together with their transductions, will eventually present a revolution in both biomedicine and biophysics. Unfortunately, presently our entire medical community is very firmly entrenched in only ordinary TW EM polarizations and ordinary biochemistry. Our scientists thus are studying less than 1% of the functioning of the living organism.

As an example, since Western scientists have no knowledge or research into the time-polarized mind domain and mind-body coupling, the mystery of the *placebo effect*—in which the mind has been verified to produce actual physical effects – will indefinitely remain a mystery.[3] As one result of that "crippling" of our medical theory, medical experimenters have to deliberately drive experiments on drug effects, etc. beyond the "mind's own threshold" of physical effect. In short, they have to drive the experimental results over and above the results being accomplished by the placebo effect. They must perform valid placebo control experiments, and subtract the placebo component results from their own experimental results, to determine the "delta" induced purely by the drugs.

The Regenerative Subsystem

Another important subsystem of the MCCS controls the *regenerative system* of the body. The regenerative system has various mechanisms, including growth control, differentiation and dedifferentiation of cells, etc. Becker's portrayal of the regenerative system is shown in *Figure 10*. One of its most important functions is healing and restoring damaged or diseased cells. It does that by a most remarkable process.

This system has been very poorly studied, mostly by Becker[4] *et al.*, although there is still sporadic, ill-funded, and ill-planned research on regenerative effects from time to time.[5] Becker and his colleagues came closest, and—considering the severe limitations of the orthodox U(1) EM model available to them—Becker's work is particularly revolutionary. He demonstrated beyond question that simple DC potentials could generate cellular differentiation and redifferentiation, and produce results at global sites in the body and not just at the local injury sites where the DC potentials were applied.

Becker was nominated for a Nobel Prize for his Herculean efforts and remarkable results. For his trouble he was hounded, his funds were withdrawn, and he was forced to retire at an early age in his 50s. His biggest "heresy" was to bravely point out that power line radiations had deleterious effects upon biological systems, and to testify to it in a series of hearings and court cases where a very powerful electrical power industry essentially ran rough shod over science and the public.

Organized Science Is Highly Politicized and Imperfect

It is not only politicians that are "bought and paid for" by powerful vested financial interests. In organized science, precisely the same arrangement holds true. Scientists are no better or no worse than any other segment of humanity. A surprising percentage will produce the results desired by their benefactors, if provided with cushy jobs, good income, prestige, and secure positions.

It has been said that money is power. It has also been said that power corrupts, and absolute power corrupts absolutely. In the scientific community, as well as in the political community, this "great corruption" by great financial power has been elevated to an art form. Simply examine the funds made available to scientists. There is not enough for all the scientists, so it is fiercely competitive. If the scientist does not get grants (e.g., a would-be professor at a university), he does not bring in extra money to the university. He cannot get his graduate students funded, so his own innovative (read: heretical) research is thwarted. His papers are rejected by the journals, and before long he has no funded position in science. If lucky, he will be employed in the grocery store or the butcher shop.

Every bit of funding available for the scientists to compete for (and apply for grants) already has firmly fixed controls of exactly what research can be done with that money. The "science" that will be applied is already formulated. That and only that can be done. It is a tribute to the dedication and resourcefulness of our working scientists themselves, that anything innovative at all gets done.

One can have a very fine scientific career, tenure as a professor, get one's graduate students funded, get one's papers published in prestigious journals, win awards and achieve stature and prestige, if one just plays the desired game prescribed by the "old boys schools" in control of almost all the scientific funding. Yes, one can permissibly move the decimal point in the present models a bit, and be rewarded handsomely for doing so. But one is damned if one attempts any serious change or overthrow of obsolete scientific models and practices favored by the "system".

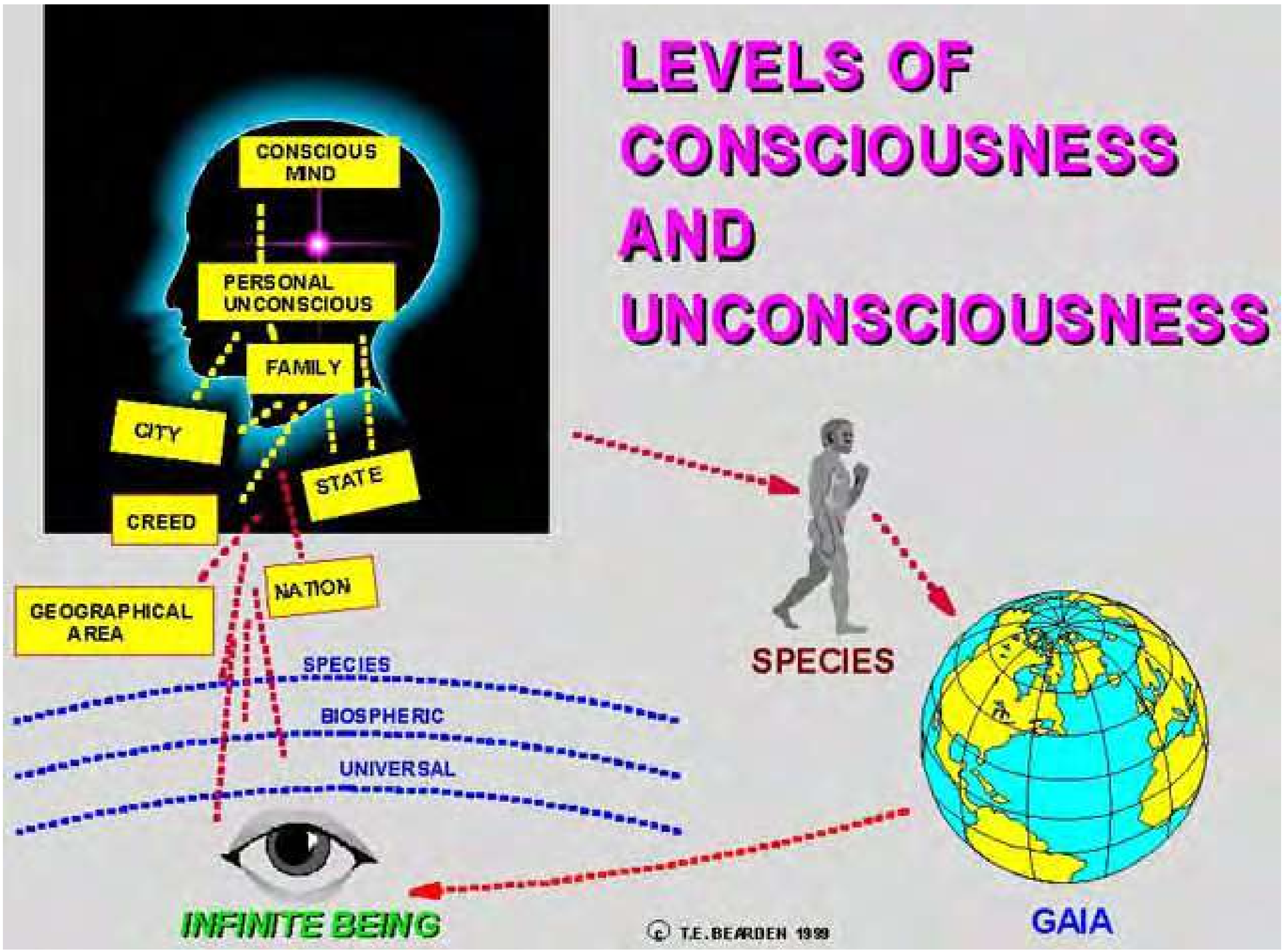

Figure 8. Nested levels of consciousness and unconsciousness. The mind levels contain the "life" and "mind" dynamics.

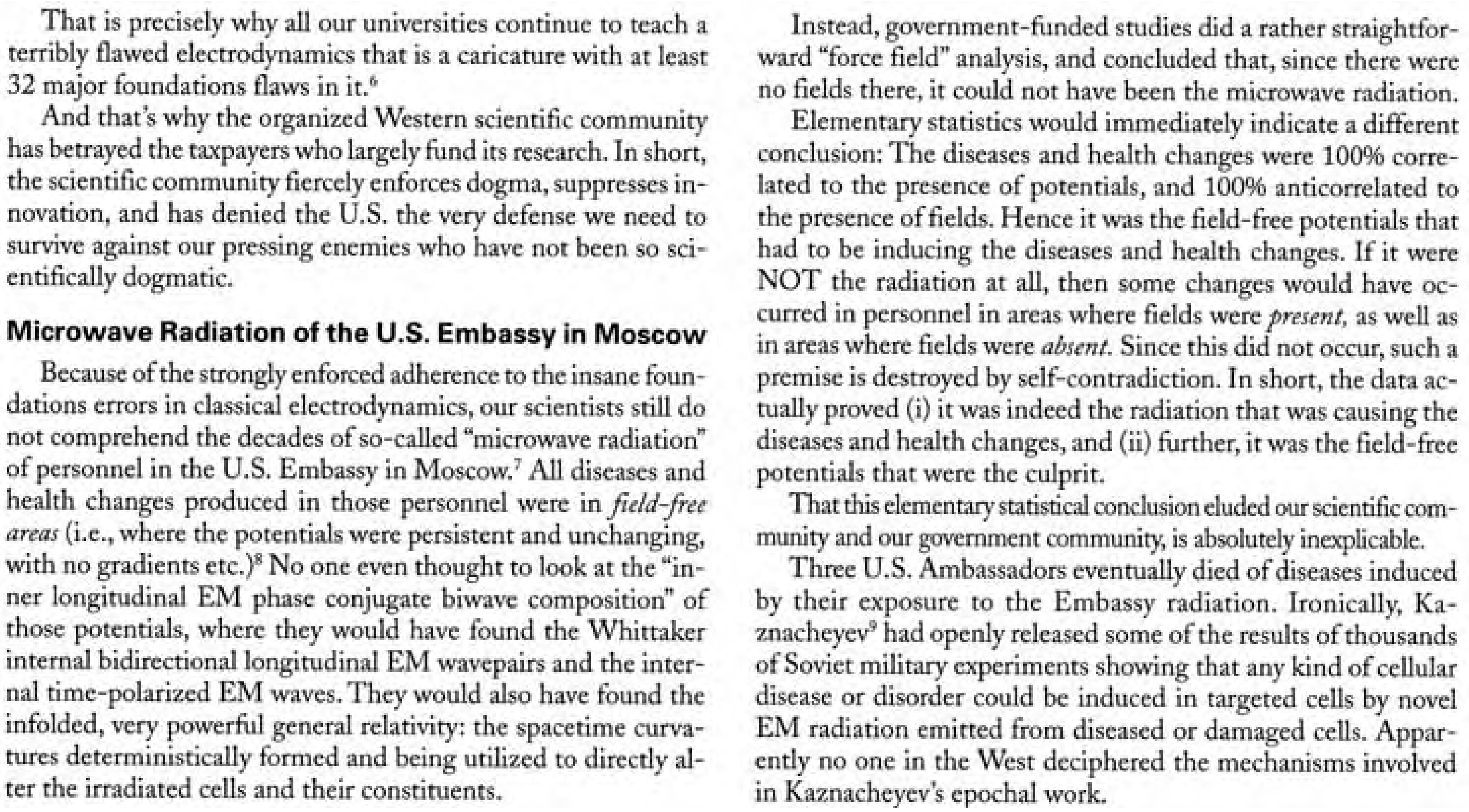

That is precisely why all our universities continue to teach a terribly flawed electrodynamics that is a caricature with at least 32 major foundations flaws in it.[6]

And that's why the organized Western scientific community has betrayed the taxpayers who largely fund its research. In short, the scientific community fiercely enforces dogma, suppresses innovation, and has denied the U.S. the very defense we need to survive against our pressing enemies who have not been so scientifically dogmatic.

Microwave Radiation of the U.S. Embassy in Moscow

Because of the strongly enforced adherence to the insane foundations errors in classical electrodynamics, our scientists still do not comprehend the decades of so-called "microwave radiation" of personnel in the U.S. Embassy in Moscow.[7] All diseases and health changes produced in those personnel were in *field-free areas* (i.e., where the potentials were persistent and unchanging, with no gradients etc.)[8] No one even thought to look at the "inner longitudinal EM phase conjugate biwave composition" of those potentials, where they would have found the Whittaker internal bidirectional longitudinal EM wavepairs and the internal time-polarized EM waves. They would also have found the infolded, very powerful general relativity: the spacetime curvatures deterministically formed and being utilized to directly alter the irradiated cells and their constituents.

Instead, government-funded studies did a rather straightforward "force field" analysis, and concluded that, since there were no fields there, it could not have been the microwave radiation.

Elementary statistics would immediately indicate a different conclusion: The diseases and health changes were 100% correlated to the presence of potentials, and 100% anticorrelated to the presence of fields. Hence it was the field-free potentials that had to be inducing the diseases and health changes. If it were NOT the radiation at all, then some changes would have occurred in personnel in areas where fields were *present,* as well as in areas where fields were *absent.* Since this did not occur, such a premise is destroyed by self-contradiction. In short, the data actually proved (i) it was indeed the radiation that was causing the diseases and health changes, and (ii) further, it was the field-free potentials that were the culprit.

That this elementary statistical conclusion eluded our scientific community and our government community, is absolutely inexplicable.

Three U.S. Ambassadors eventually died of diseases induced by their exposure to the Embassy radiation. Ironically, Kaznacheyev[9] had openly released some of the results of thousands of Soviet military experiments showing that any kind of cellular disease or disorder could be induced in targeted cells by novel EM radiation emitted from diseased or damaged cells. Apparently no one in the West deciphered the mechanisms involved in Kaznacheyev's epochal work.

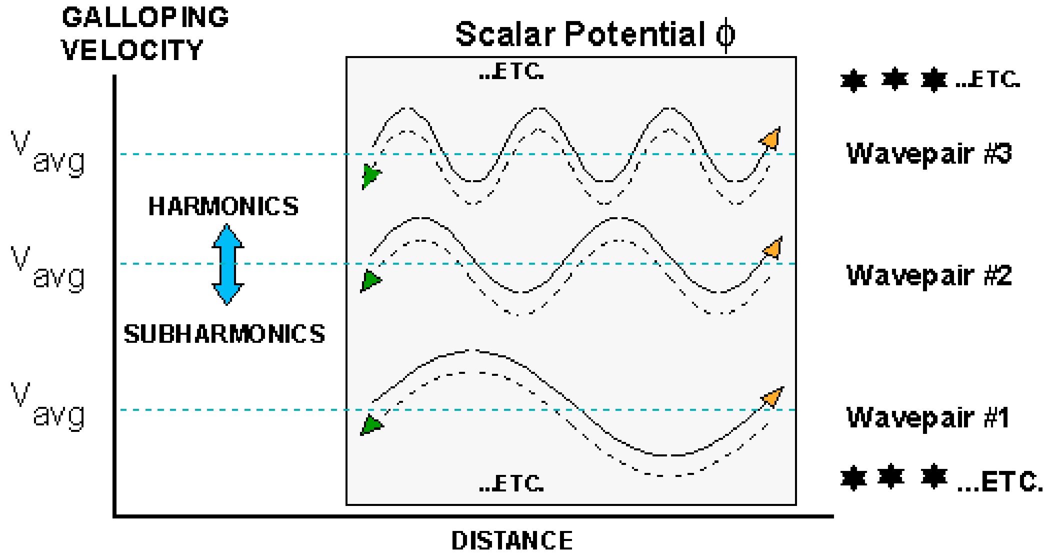

The Structure Is:

- A harmonic set of longitudinal phase conjugate wavepairs.
- In each wavepair the two waves superpose spatially after detection, but travel in opposite directions.
- The convergent wave set is in the imaginary plane, and hence is not observable. It is EM energy incoming to the potential (dipolarity) from the time domain.
- The charge's spin is 720 degrees, 360 in the real plane and 360 in the imaginary plane.
- Hence the charge receives the complex convergent EM energy, transduces it into real EM energy, and emits enormous energy at the speed of light in all directions -- which includes bidirectional pairs in 3-space (after the reaction, being after "observation").
- This produces the fields and potentials from the "source charge or dipole."
- Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus their quantum field theory work strongly supports the "negative resistor" interpretation of the scalar potential and our solution to the source charge and source dipole problem.

Figure 9. Infolded longitudinal biwaves (time-polarized EM waves)

composition of a scalar potential.

EXPLORE! VOLUME 9, NUMBER 3, 1999

The Present Situation in Western EM Bioeffects Research

Our scientific community has not even *designed physical experiments* to validate long-term EM bioeffects of the kind caused by induced wave transductions and long-term effects inside the absorbing human cells and in the body's regenerative and immune systems. Western science has no knowledge of these technical internal "wave polarization transduction" mechanisms. Yet the basis for such mechanisms has existed in the hard physics literature for decades. Leading EM bioeffects researchers seem totally unaware of small reaction cross sections inside the body cells and tissues for the transduction of transverse EM signals into LW and t-polarized EM signals. Conventional Western researchers have not investigated it, and often will not believe it when it is pointed out to them along with the supporting basis in the literature.

A major part of the mainline "EM bioeffects" community, of course, seems comfortably funded largely by the power industry etc.[11] The record is clear that this dominant, well-funded group has little or no intention of "finding" anything that will question powerline radiation safety and upset their cushy positions.

In fact, leading bioeffects researchers have pointed out this deliberate biasing of EM bioeffects as a research area. Quoting Dr. Andrew A. Marino, one of the great pioneers in the area of EM bioeffects[10] and powerline radiation assessment.[11]

"Neither scientists nor the public can rely on power-industry research or analysis to help decide whether powerline electromagnetic fields affect human health because power-industry research and analysis are radically misleading."

Dr. Marino then gives the specific details to prove it, and he shows specifically how they slant and mangle their experiments and interpretations of the results, to do it. If the major financial interests cannot *scientifically* destroy determined, honest researchers such as Marino and Becker, they do not hesitate to attempt to destroy them in lawsuits.

The "Diffusion Mixmaster" at All Body and Cellular Levels

Every mass and all its particles are in a violent energy exchange with the active vacuum. A human body and every part of it is in a continual exchange of EM energy with its environment(s), including transverse, longitudinal, and t-polarized (scalar) EM energy exchanges. Further, within the dielectric body there is a sort of intensive "diffusion" of special kind, actually due to iterative phase conjugation and iterative interferometry of the dense signals environment impinging upon it. This internal "diffusion" and transduction (iterative multiwave phase conjugation and iterative multiwave interferometry) thoroughly "mixes up and transduces" one type of EM wave or signal into the other, including many of the body's own TW, LW, and TDW signals.

Continual or sustained TDW exposure (irradiation by t-polarized EM waves) charges the body and its particles with time-excitation charging, something which has been erroneously omit-

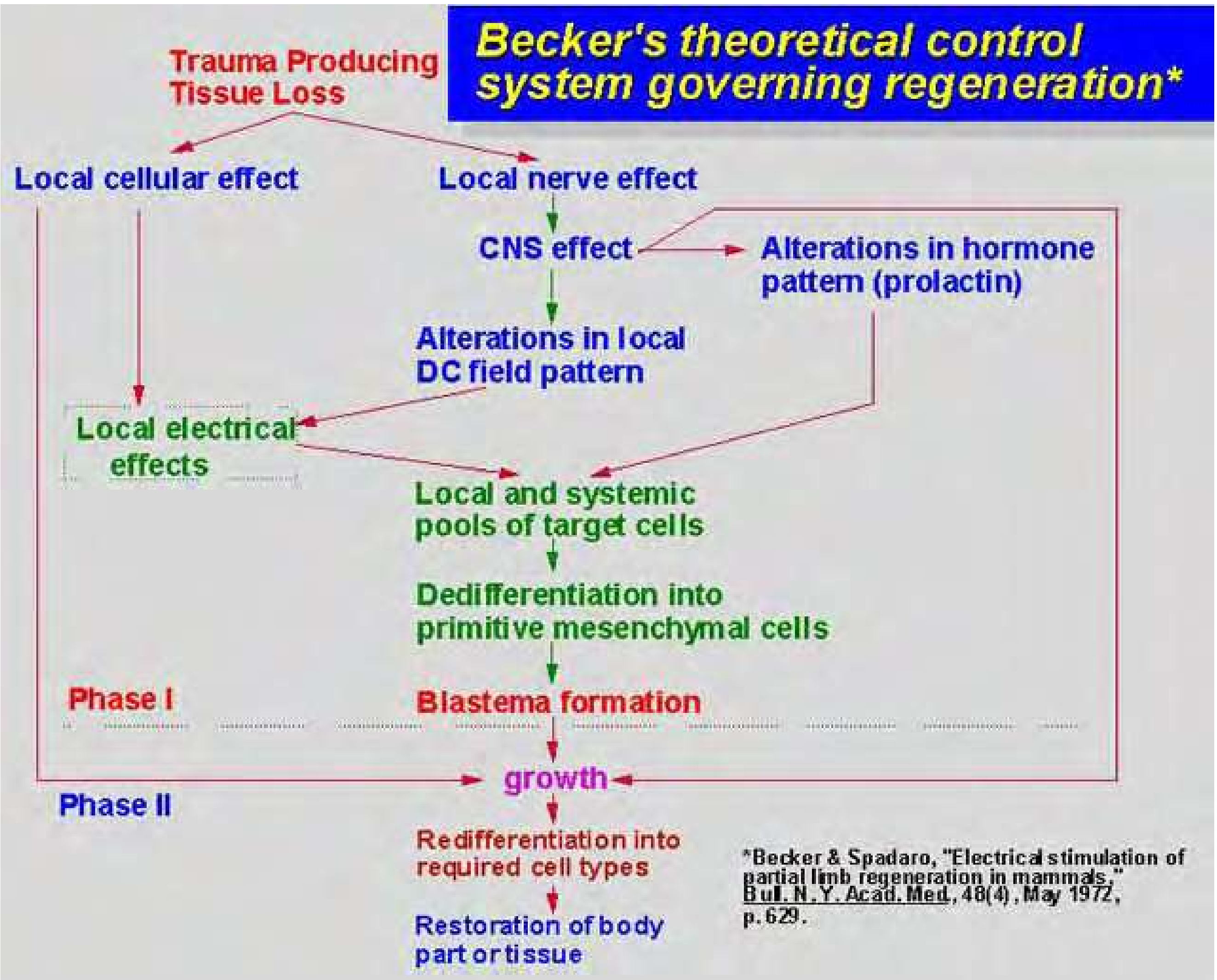
Becker's theoretical control system governing regeneration*
Trauma Producing Tissue Loss
Local cellular effect
Local nerve effect
CNS effect
Alterations in hormone pattern (prolactin)
Alterations in local DC field pattern
Local electrical effects
Local and systemic pools of target cells
Dedifferentiation into primitive mesenchymal cells
Phase I
Blastema formation
growth
Phase II
Redifferentiation into required cell types
Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N.Y. Acad. Med., 48(4), May 1972, p. 629.

Figure 10. Becker's model of the Regenerative system of the body.

ted in physics, but is true nonetheless. Inexplicably, physicists just ignore what happens time-wise when a mass absorbs a photon. We have covered that elsewhere.

So inside the body dielectric, there is a "diffusion mix" correlate of any and every type of the waves, to any or all of the exchanges, plus the additional internal interactions of the body in its TW, LW, and scalar EM wave transductions. There is a one-to-one correspondence to any one of the three types of polarization waves (transverse, longitudinal, and scalar) and everything going on anywhere in the body.

Note that the type of LW we are using is actually made of phase conjugate pairs of transverse EM waves, so its "photons" are made of photon/antiphoton pairs. That is, our longitudinal EM wave has an internal TW structure, and the LW wave is comprised of spin-2 *gravitons*. Our longitudinal EM wave has an internal structure, and it dramatically differs from the LW used by the U.S. scientific community which ignores wave, field, and potential infolded substructures. Our LW (which seems to be what the Russians use) is also a gravitational wave because it is comprised of oscillations in its distributions of spin-2 photon/antiphoton pairs (i.e., *gravitons*).

The type of t-polarized (scalar) EM wave we are using is a phase conjugate coupled pair of the specialized electrogravitational (EG) LWs. The "t-polarized photons" comprising this t-polarized "EM" wave are comprised of spin-4 *supergravitons.* These are, I believe, what the KGB energetics weapons scientists use also. By iteration, we can go into as deep an internal multilevel structuring of EM and EG waves, of any type, via methods shown by Whittaker.[12]

Now let us examine a body as a dielectric, sitting there in its exchange, in relative equilibrium. It turns that every tiniest piece of a dielectric participates in any emission from it[13]—even the emission of a single photon, graviton, or supergraviton. Well, it *would* so participate, due to the special sort of correlated diffusion and interferometry and phase conjugation that we pointed out. Anything that escapes (is re-emitted) from the body, has been through this "diffusion wringer". It has escaped from each and every internal part of the body.

Novel Signatures in Body-Emitted EM Radiation of Any Polarization

A single photon or wave emitted from a dielectric carries in its internally structured time component a *template* of the exact internal time structuring of every part of the dielectric emitting that photon. It carries in its internally structured energy component a *template* of the internal energy structuring of every part of the dielectric emitting it. An emitted EM wave, being a collection of emitted photons and photon structures, thus carries the same internal structuring of both its transported energy and its transported time.[14] Western scientists do not use the internal

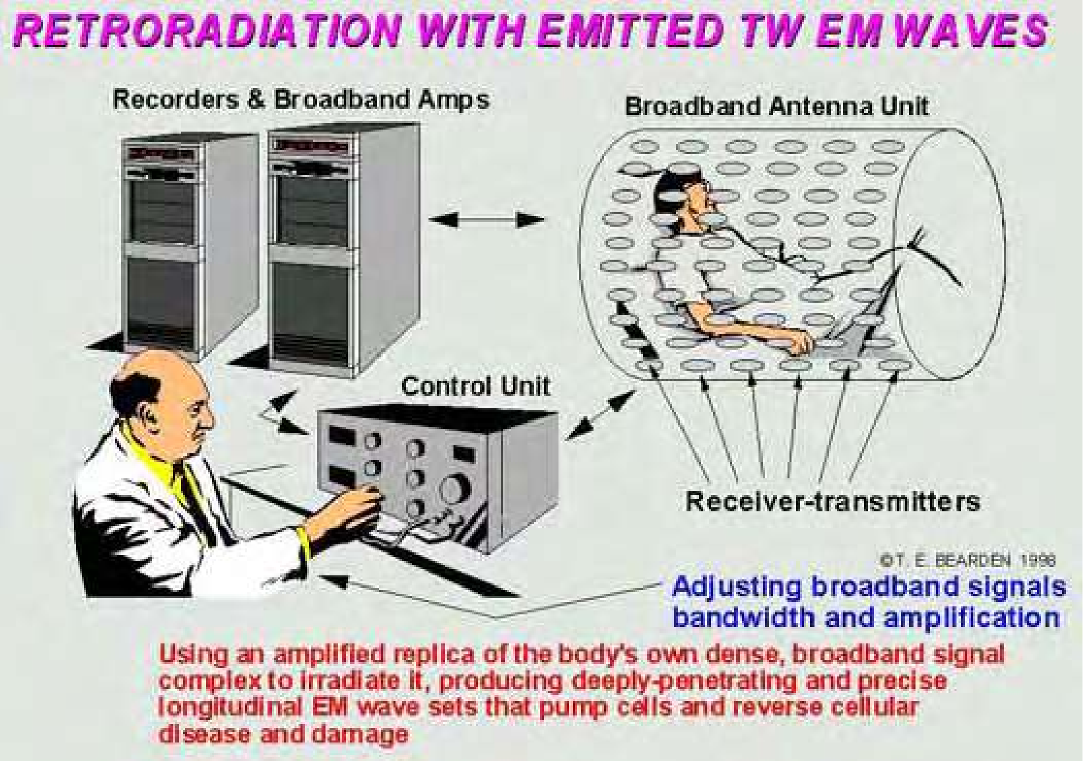
RETRORADIATION WITH EMITTED TW EM WAVES
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
©T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification
Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

Figure 11.

structuring of photons and EM sine waves, have no instruments for that, and so do not know it. The Russians know it,[15,16] use it, and have developed sophisticated instruments for it for the last 30 years or more.

Shortly we will postulate how conventional scientists experimenting with mind and behavior control electromagnetically might utilize an immediate consequence of the fact that the specific emission from a body at any time—whatever type of wave component we look at—is a total correlate to all that body's internal functions, its energy exchanges with its environment, its states, its mental states and conditions, etc.

And we accent that the coherently integrated intent-changes in that body from its coupled mind's operations are also in the substructure of photons and EM waves emitted from the living body into its environment.

How Western Mind Control Scientists Might Proceed

Conventional scientists would ignore scalar photons and waves, determine the TW spectrum, and then slowly and painfully correlate that to the various physical, mental, and emotional things. That's the way they would *have* to do it, since apparently they use only standard electrodynamics. They can in fact do this via brute force, because that emission TW spectrum is of waves escaping from the entire "diffusion machinery" mixmaster in the body dielectric. It does contain all those integrated *mind-intent* 3-space correlates as well as the ordinary body 3-space energy correlates.

Anyhow, in envisioning Western procedures, the test body would be sitting there, in its normal (let's say fairly normal and quiet) EM environment. [See *Figure 11*] Let's assume this body is not sick, but is healthy and normal. Its emissions are a part of its equilibrium condition. It follows that the totality of its TW emissions is correlated to the sum total of everything that interacted with the body, plus its own internal reactions, including from mind-body coupling. Indeed, everything that has ever happened to that entity—physically, mentally, emotionally, etc.—has internal correlates infolded inside the emitted transverse EM wave spectrum.[17]

So first the scientists would simply measure the full environmental TW spectrum down to very tiny levels in all frequency bands possible—everything that is coming in to the body from its environment. They would probably perform the experiments in a triply-shielded Faraday cage, so the external environment outside the cage contributes very little transverse EM wave noise (perhaps some Schumann resonance). Magnetic wave shielding and ELF shielding is something else again, and longitudinal EM wave shielding is in general not possible unless longitudinal EM waves are employed in the shielding. But a "quiet external TW environment" enhances the isolation and measurement of the specific TW signal transforms the Western experimenters would be seeking.

Then they would measure everything likewise that the body is *emitting* when in that environment. That is the "zero reference" or "normal reference" level. They would measure every aspect: frequency, amplitude, phase, wave shape, modulation, etc.

Probable Experimental and Measurement Techniques

Next they would selectively irradiate the body with specific signals, frequencies, waveforms, etc. They would begin by using portions of the same frequencies that are being emitted from the body in the "zero reference" environment. They would insure that the returned signal was of higher amplitude than the usual signal being emitted by the body at that same frequency and type.

They would consider the body dielectric as a "ship with portholes". Light (EM transverse wave energy) coming out of the portholes has originated and adapted from everywhere inside the ship and its operations. When the experimenters reinsert energy through one or more portholes, they are interested in determining precisely which operations inside the ship are affected. So the body physiological functions are also heavily instrumented.

By feeding back in the exact EM pattern and magnitude escaping the body dielectric through a single porthole, one "nullifies" that part of the emission by producing a net equilibrium in that component. This will in turn produce changes elsewhere in the spectrum emission from other portholes, and this can be measured. It will also produce "stress changes" in specific portions of the body dielectric, and in the heavily instrumented body these physiological, chemical, electrical, etc. changes can also be detected. Monitoring of the skin conductance, e.g., indicates the level of stress, and EEG monitoring with sophisticated instrumentation can indicate something of the brain wave states.

In short, the necessary measurements from which to construct specific correlates can be performed, recorded, and later analyzed.

This article will be concluded in Volume 9, #4 with Improving the Methodology, Further Phenomenology Experiments, Rejuvenation and other matters. ◆

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel U.S. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, U.S. Army. Graduate of Guided Missile Staff Officer's Course, U.S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses.

Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor. He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuum's virtual particle flux is organized and used to energetically shape and manipulate matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

1. *In fact, it simply applies in the laboratory that method that is already used by species for the so-called "directed evolution" component of evolution. The exact mechanisms for directed evolution also arise naturally in the new approach. Directed evolution is required; e.g., no reptilian species could have evolved slowly into a bird. As it started slow changes, it would have not been able to fly or to run, and would have been simply destroyed immediately by predators. Instead, such evolution occurs in a single jump in a species hard-pressed for a lengthy period. In the evolution of birds, one day every reptile egg that hatched, now hatched into a complete and functional bird with light bones, feathers, etc. A continuing debate (with perhaps more emotion than science) continues to this day on directed evolution. But it is real, and the mechanism can readily be generated (we've already done the gist of it).*

2. *Fritz Albert Popp, "Photon Storage in Biological Systems," in Fritz Albert* Popp et al, *eds.*, Electromagnetic Bio-Information: Proceedings of the Symposium, Marburg, Federal Republic of Germany, Sep. 5, 1977; *Urban & Schwarzenberg, Baltimore, 1979, p. 123-149;—"On the coherence of ultraweak photoemission from living tissues," in* Disequilibrium and Self-Organization, *ed. by C. W. Kilmister, D. Reidel, Hingham, MA, 1986;*—et al., *"Physical aspects of biophotons,"* Experientia, *Vol. 44, 1988, p. 576-585;—"Biophoton emission: New evidence for coherence and DNA as source,"* Cell Biophys., *Vol. 6, 1984, p. 33-52. See particularly Fritz Albert Popp, Ke-hsueh Li, and Qiao Gu, Eds.*, Recent Advances in Biophoton Research and its Applications, *World Scientific Publishing Co., New York, NY, 1992.*
3. *With novel use of EM time-energy to curve spacetime, the physical actions of the "placebo effect" are due to the use of such "spacetime curvature engines" to produce actual physical effects—at least temporarily —in the cells and throughout the body's chemistry, physiology, etc. Experimental detection of these physical effects are well-established in the orthodox medical literature.*
4. *For example, see particularly R. O. Becker, "The direct current field: A primitive control and communication system related to growth processes,"* Proceedings of the. XVI International. Congress of Zoology, *Washington, D.C., Vol. 3, 1963, p. 179-183;— "The direct current control system: A link between environment and organism,"* New York State Journal of Medicine, *Vol. 62, 1962, p. 1169-1176;—and D. G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents,"* Trans. N.Y. Acad. Sci., *Vol. 29, 1967, p. 606-615;—and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals,"* Bulletin of the New York Academy of Medicine, *Second Series, 48(4), May 1972, p. 627-64.*
5. *E.g., a popular discussion is given by John Travis, "Starting Over: Some Animals can regenerate limbs or even most of their bodies. How?",* Science News, *152(18), Nov. 1, 1997, p. 280-282. For a typical technical paper, see Jeremy P. Brockes, "Amphibian limb regeneration: Rebuilding a complex structure,"* Science, *276(5309), Apr. 4, 1997, p. 81-87.*
6. *I have one viewgraph which lists those 32 major flaws in electrodynamics. Further, leading theorists (Evans, Crowell, and others) in the Alpha Foundation's Institute for Advanced Study (AIAS)—of which I am a Fellow Emeritus by AIAS's kind tolerance—have added about another very technical and very fundamental dozen or so errors to my own list. A leading foundations scientist with over 500 papers in the literature commented dryly to the AIAS that "I broadly agree. that the transverse plane wave view of vacuum electrodynamics is the biggest blunder of twentieth century science. Faraday visualized that "EM disturbances" surrounding wires, magnets, etc. were simply the transverse plucking of taut strings. As he states in his book, Maxwell—who also firmly believed in the material ether—simply captured Faraday's thinking and assumed that transverse "plucked string" wave disturbance in the material ether. At the time all this was expressed, the electron had not been discovered, the atom had not been discovered, the molecule was a featureless blob, and no one had the slightest notion of an atomic nucleus filled with protons and neutrons. Charge was simply a "certain volume of electric fluid" and everything electromagnetic was expressed in fluid dynamics. As an example, that "material fluid" and "material ether" theory was destroyed in the late 1880s by the Michelson-Morley experiment. All the Maxwell equations assume that material ether to this day; not a single equation was ever changed to eliminate the material ether assumptions during the last 110 years.*
7. *For full explanations of the "Moscow Signals," see T. E. Bearden, .,* Gravitobiology: A New Biophysics, *Tesla Book Co., P.O. Box 121873, Chula Vista, CA 91912, 1991;—*AIDS: Biological Warfare, *Tesla Book Co., 1988. For an excellent Government study of the Embassy radiation by excellent but orthodox Johns Hopkins scientists, see A. M. Lilienfeld* et al., Foreign Service Health Status Study: Evaluation of Health Status of Foreign Service and Other Employees from Selected Eastern European Posts, *Department of Epidemiology, of Hygiene and Public Health, The Johns Hopkins University, Baltimore, Maryland, Final Report, July 31, 1978. On p. 245, it is noted that a marked excess of symptoms were noted in the Moscow employee group as compared to other East European employees in other East European nations.*
8. *Johns Hopkins researchers established the exact field patterns present in the Embassy from the microwave carrier radiation; see Robert C. Mallaliew,* A Model of the Microwave Intensity Distribution within the U.S. Embassy, *Moscow 1966 to 1977, Report FS-80-166, Applied Physics Laboratory, Johns Hopkins University, Aug. 1980. Nowhere in any of the Government-funded studies is there any consideration of field-free scalar potentials and their internal time-polarized EM wave structures (the bidirectional phase conjugate wavepairs shown by Whittaker to comprise the scalar potentials). Yet the field radiation patterns surprisingly showed that in just such field-free, persistent potential zones was where 100% of the health changes occurred.*
9. *Vlail Kaznacheyev and L. P. Mikhailova,* Ultraweak Radiation in Intercellular Interactions, *[in Russian], Novosibirsk, 1981 [an English exposition of much of the Kaznacheyev work is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions,"* Psi Research, *1(1), Mar. 1982, p. 47-76.]. See also Vlail Kaznacheyev* et al., *"Distant intercellular interactions in a system of two tissue cultures,"* Psychoenergetic Systems, *1(3), Mar. 1976, p. 141-142. In the same issue, see "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144. See also Vlail Kaznacheyev, "Apparent information transfer between two groups of cells,"* Psychoenergetic Systems, *1(1), Dec. 1974. The Soviet Union awarded Kaznacheyev a very high medal for his achievements.*
10. *For a very nice summary of EM bioeffects from the more conventional view, see Robert O. Becker and Andrew A. Marino,* Electromagnetism and Life, *State University of New York Press, Albany, 1982. The book is also available at http://www.ortho.lsumc.edu/Faculty/Marino/El/ElPub.html.*
11. *Andrew A. Marino,* Powerline Electromagnetic Fields and Human Health, *http://www.ortho.lsumc.edu/Faculty/Marino/Marino.html. This book is highly recommended. Particularly see "Chapter 5, Blue-Ribbon Committees and Powerline EMF Health Hazards," and "Chapter 6: Power-Industry Science and Powerline EMF Health Hazards." Biophysicist Marino is one of the leaders in the field and has been personally involved in many skirmishes with powerline-dominated studies and findings.*
12. *E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics,"* Mathematische Annalen, *Vol. 57, 1903, p. 333-355; — "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* Proc. Lond. Math. Soc., *Series 2, Vol. 1, 1904, p. 367-372.*
13. *Reali, G. C., "Reflection from dielectric materials,"* American Journal of Physics, *50(12), Dec. 1982, p. 1133-1136. The reflected field from a dielectric material is not generated just at its surface but comes from everywhere in the interior of it.*
14. *Present physics models the EM wave as transporting energy, but erroneously omits that it also transports time. A photon is made of angular momentum, or "energy x time". Since EM waves are made of photons, it follows that they transport both the energy components and the time components making up those photons. An EM wave is a wave of angular momentum, not just a wave of energy.*
15. *KGB energetics weapons scientists call this internal structuring the "information content of the field." However, the Russians use a good deception plan, which implies that the "information content of the field" is just the normal transverse EM spectral content of EM signals. Appreciable "pertinent" work using that phrase is produced by the open Russian scientific community, thus confusing the issue. Our electrodynamicists seem to buy it hook, line, and sinker. The deception plan also tricks Western scientists into thinking that psychoenergetics is nothing but a Russian version of Western parapsychology. Again, our scientists comfortably accept this with smug satisfaction.*
16. *Presently the Russians are in a great new version of this deception plan, with the goal of deceiving Western scientists into thinking that the "secret" Russian energetics is based on torsion fields. Consequently they are releasing substantial material on torsion wave theory, including replicable experimental results. The released version of torsion fields has one glaring omission: the time components and structuring. This deception plan is excellent, because the torsion field theory that is released, is indeed "real" and testable. It is also revolutionary, with respect to present Western science. It almost positively assures diverting the attention of much of the Western science community into this highly limited subset of the full Russian torsion theory—effectively diverting the West away from psychoenergetics and the direct use of t-polarized EM waves and fields, torsion or otherwise.*
17. *We point out that E. T. Whittaker, in "On the Partial Differential Equations of Mathematical Physics,"* Mathematische Annalen, *Vol. 57, 1903, p. 333-355, showed that every scalar potential is comprised of an internal structure of bidirectional longitudinal EM phase conjugate wavepairs in an infinite set of harmonics. Each of those phase conjugate longitudinal EM wavepairs comprises a time-polarized EM wave. So the common potential contains an internal longitudinal EM wave structure that is further organized into an internal time-polarized EM wave structure. The mind-intent correlates and the total mind and emotional experience correlates from fertilization of the egg that led to development of that person, is in the internal substructure of every potential in the living body. In his "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* Proc. Lond. Math. Soc., *Series 2, Vol. 1, 1904, p. 367-372, Whittaker also showed that any EM field or wave is comprised of two scalar potential functions. It follows from Whittaker and Reali that the EM emissions from the living body carry the internal correlates of everything that has been experienced by that person from conception in the womb. Sometimes this deep embedded "hidden information content" can even be accessed in deep hypnosis. I once observed the deep hypnotic regression of a subject back to her prenatal experiences prior to birth, where the subject relived hearing words of rejection by the father, before the subject was born. The occurrence of the incident and the father's words were directly confirmed on the spot by the mother, who was closely observing the session as it occurred. It turned out that this particular rejection of the baby still in the womb had played a major role in subsequent deep psychological problems experienced by that person all her life. To resolve the problem, use of a powerful "simulated" reliving pre-birth experience was conducted with the subject again in the pre-birth state and actively participating, but now with the father figure directly expressing love, acceptance, and cuddling and hugging his daughter. This led to a rather dramatic remission of most of the deep psychological problems within 24 hours.*

Mind Control and EM Wave Polarization Transductions, Part 3 of 3

©Copyright 1999 by T.E. Bearden, USA

This outstanding article is the last, Part 3, Part I and 2 of which were published in Volume 9, #2 and #3 of Explore! for the Professional. Because we feel that Col. Bearden’s Special Note below is so very important we are repeating its publication in all 3 parts of this article.

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and achieving many more effects in the targeted bio-organism or groups of them. With Hunt's method, the system as an open dissipative system far from thermodynamic equilibrium in its environment is being measured as such. This more directly allows for determination of the environmental inputs on a much broader range of subsystems, both space-like and time-like. Particularly for mental and emotional correlates to environmental stimuli, such measurements and correlations are essential.

More Comprehensive Phenomenology Experiments

To return: Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.

In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories.

is both illegal and immoral.

Two Ways to Improve the Methodology and Experimental Correlates

We diverge momentarily: If they wish to improve this methodology, they will consider the body dielectric as a nonlinear isotropic medium. Then they will pretend that the spectrum of difference frequencies between each two adjacent waves is the spectrum of "emissions" from the body. They would also use those, as if they were actual emitted spectral signals, instead of the actual signals emitted. That procedure will give superior correlations, but it is more complicated. Part of that is known for underwater sonar,[1] and so our own scientists may or may not have realized it for use in mind and behavior correlate determination with respect to the body dielectric.

Another dramatic improvement in the depth to which the correlates can be constructed, is provided by detecting each porthole output in a two-channel device, where one channel has an adjustable time-delay as utilized by Dr. Valerie Hunt.[2] The two channels—one real-time and one slightly delayed—are then mixed, as by Hunt's method, producing an instrumental measurement of the chaotic part of the functions. Again, this can be measured, recorded, and later analyzed. It adds (i) a completely different kind of thermodynamics in the system measurements, (ii) a dramatically extended set of correlates, (iii) more direct measurement of the time-polarization causative signals, and (iv) an applications technology eventually having many additional degrees of freedom and therefore capable of Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions.

A particularly vulnerable aspect of every mammal is the pleasure center in the brain. Experiments have shown that, when this center is stimulated, it is the most addictive experience possible. Rats will forego food, endure electrical shocks, starve, and even die to obtain stimulation of this center. Obviously a device capable of generating signals that evoke direct stimulation of the pleasure center in humans would be a powerful weapon.[3]

Extending the Data Base

See Figure 12. After the researchers have made an "individual porthole signal insertion" correlate database, then they would make a "two, three, four," porthole signal insertion correlate database by inducing selected multiple changes at once. Again they would analyze everything; the program would need one or more supercomputers. It would also require quite a highly qualified multidisciplinary team, extensive facilities, substantial funding, and would be many years in duration. We are describing a decades-long program. But it's a doable.

With sufficient development and performance of such a program, the researchers would have produced (with sufficient testing and analysis, and sufficient retest validation and verification) a database of "specific correlates for a given overall desired physical, mental, or physical and mental delta."

A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment.

The Dielectric Body as a Ship With Portholes

Again let us review what the EM transverse wave irradiation from the body dielectric represents. Think of the body dielectric as a sort of "ship", with lots of portholes. When it gets irradiated, certain operations happen internally to that radiation, i.e., certain "processes" and "interactions" occur, which change both the mental and physical states. The output emissions back from the body dielectric, out through the filtering "portholes", will directly correlate to what has happened inside the ship.

It is terribly important to realize that one can use the same frequencies and spectra coming back out of those "portholes" to insert EM signals and waves etc. back in there, and into the deepest processes going on in the body and in the mind (at all levels, including the deep unconscious). times. So it will reach directly into and affect any and every part of the dielectric, as you wish, and in whatever manner you have correlates for.

We filed an invention disclosure[4] on the above correlate and "back-porthole" process for use in directly inducing full and efficient conversion directly to time-density (scalar, or time-polarized) EM waves inside the body/ cells, therefore immediately and directly time-excitation charging the body cells and all their internal components.

Subtracting the Normal Correlates to Obtain the Deltas

A specifically diseased body, e.g., has a specifically tailored change in its "normal" emission spectra, because of the disease condition alteration of its resident spacetime curvature engine. By having a database for the normal emission spectra of the healthy body, one can subtract it from the actual emitted spectra, to obtain the "delta spectrum".

If the correlates have been previously determined by the above mentioned research program, then one may simply introduce an amplified "delta spectrum" into the body, i.e., through the portholes. This will immediately time-excite the cells and their components throughout the body, particularly where that disease does exist. It turns out that the time-excitation charging by that method will "pump" those diseased cells in the time domain. The cells themselves (and all their parts) are highly nonlinear—to time-excitation

Remember, whatever you put back in there, is going back through the "mixmaster" hopper. It will be iteratively phase conjugated and interfered with by everything, countless charging, they are totally nonlinear. So they will act as rather perfect pumped phase conjugate mirrors, but pumped in the time-energy domain rather than in the spatial

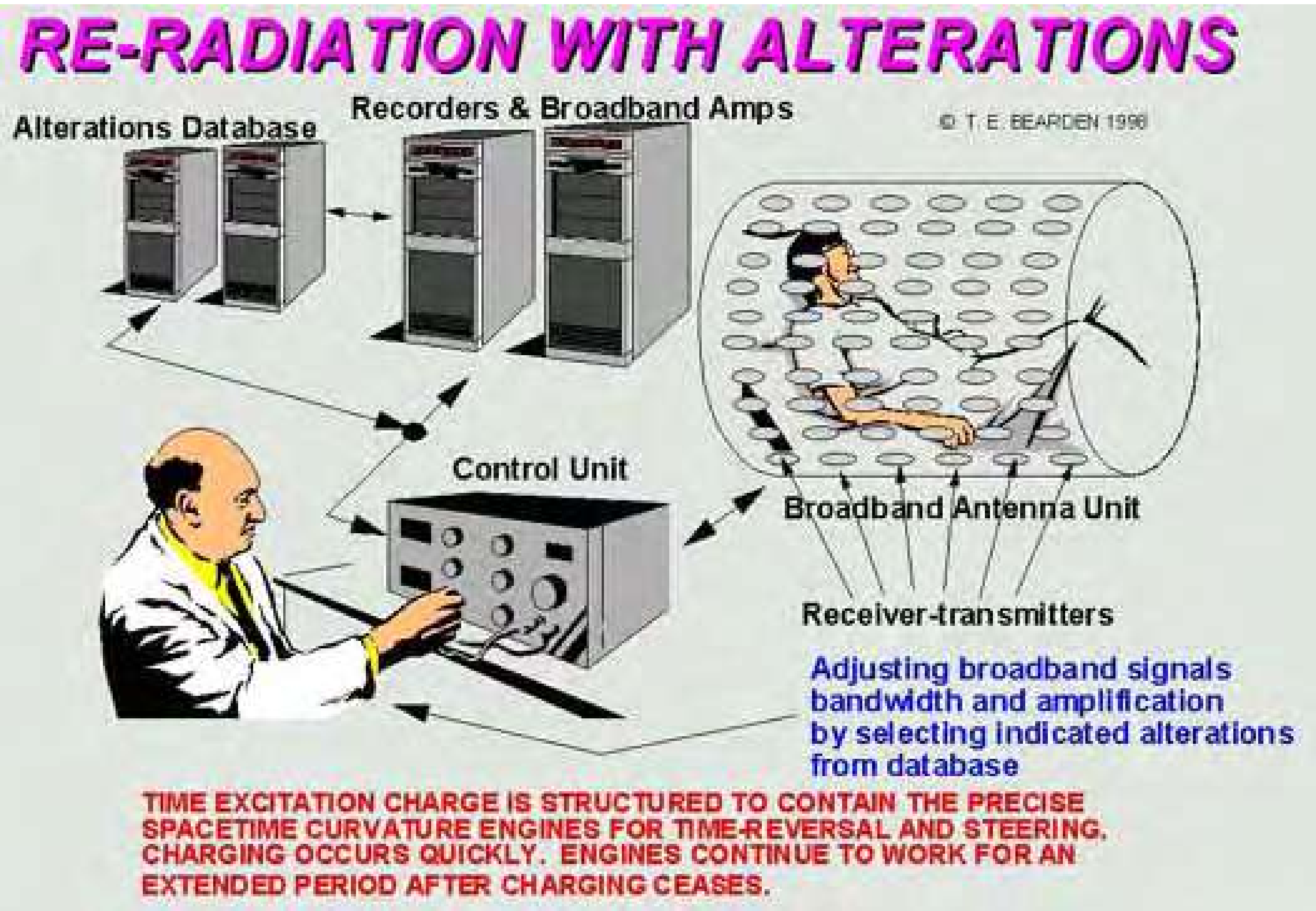
RE-RADIATION WITH ALTERATIONS
Alterations Database
Recorders & Broadband Amps
© T. E. BEARDEN 1998
Control Unit
Broadband Antenna Unit
Receiver-transmitters

Adjusting broadband signals bandwidth and amplification by selecting indicated alterations from database

TIME EXCITATION CHARGE IS STRUCTURED TO CONTAIN THE PRECISE SPACETIME CURVATURE ENGINES FOR TIME-REVERSAL AND STEERING. CHARGING OCCURS QUICKLY. ENGINES CONTINUE TO WORK FOR AN EXTENDED PERIOD AFTER CHARGING CEASES.

Figure 12. Using an amplified replica of the body's own dense, broadband signal complex, containing added alterations, to irradiate the body and produce deeply-penetrating precise time-density charging for cellular time-reversal while steering the time-reversal trajectory to also correct a congenital condition.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

energy domain. This is a dramatic extension to nonlinear phase conjugate optics theory.[5]

General relativity tells us that each specific detail of the mass and energy structure of each cell and all its parts—down to the finest details—is accompanied by a precise set of spacetime curvatures. By reintroducing the same energy as the disease delta TW waves, the affected cells add the phase conjugates, converting the inserted TW signals into longitudinal EM wave signals. The longitudinally-pumped cells again phase conjugate, converting these longitudinal EM waves to time-polarized EM waves, and "charging" the cells and all their parts in the time domain.

Suddenly, at that point the resident spacetime curvature engine—rather than TW or LW EM waves-serves as the "input signal" to the cell-and-its-parts as pumped phase conjugate mirrors (PCMs), pumped in the time domain. An extension to the distortion correction theorem of nonlinear optics results. An exact, amplified antiengine for the diseased cells—i.e., a phase conjugate replica of the resident spacetime curvature engine—is formed. This amplified antiengine overpowers the resident disease engine, powerfully time-reversing the cells and their components back to a previous healthy state. The time-reversal actually occurs over a period of time after the "time excitation charge-up" ceases, say over the next week or two. Usually two to three quick irradiations are required, one week apart, to give a time-reversing period of three to four weeks.

Even though appreciable forces are operating internally upon the cells and all their parts, there is no net translation force upon the cells or any part. E.g., as Pepper[6] states,

"On a more fundamental level, the ideal lossless PCM reverses all the quantum numbers of the incident photon (i. e., linear momentum, angular momentum, etc.). It can be shown that the PCM therefore experiences no linear or angular momentum transfer from the incident photons; hence, the PCM is free from photon radiation pressure and torques."

Note that, in ordinary EM theory, it is precisely translation forces that are engendered upon charged mass by EM fields. This process is quite different. In biological terms, this process dedifferentiates (in physics: time-reverses) the diseased cells and all their components—including the genes—back to an earlier healthy state.

By taking a diseased body and feeding back into it an exact amplified replica of its emitted TW spectrum, one can force the body to "time-excitation charge" all cells and their components immediately, with high efficiency. The human body and every cell and every part of every cell does the necessary "calculation" for the exact antiengine for each part. Since time is extremely condensed spatial EM energy (by a factor of almost 10^{17}), a startling contraction occurs in the radiation time required to time-charge the cells and their components and charged particles. This is a method for instant initiation of time-charging, hence it is extremely rapid. What took Priore several hours of LW irradiation to accomplish in the body,[7] can now be done in seconds with this "thwarting" and spectrum transduction method. In fact, it is an intense stress on the body, and—say—should only

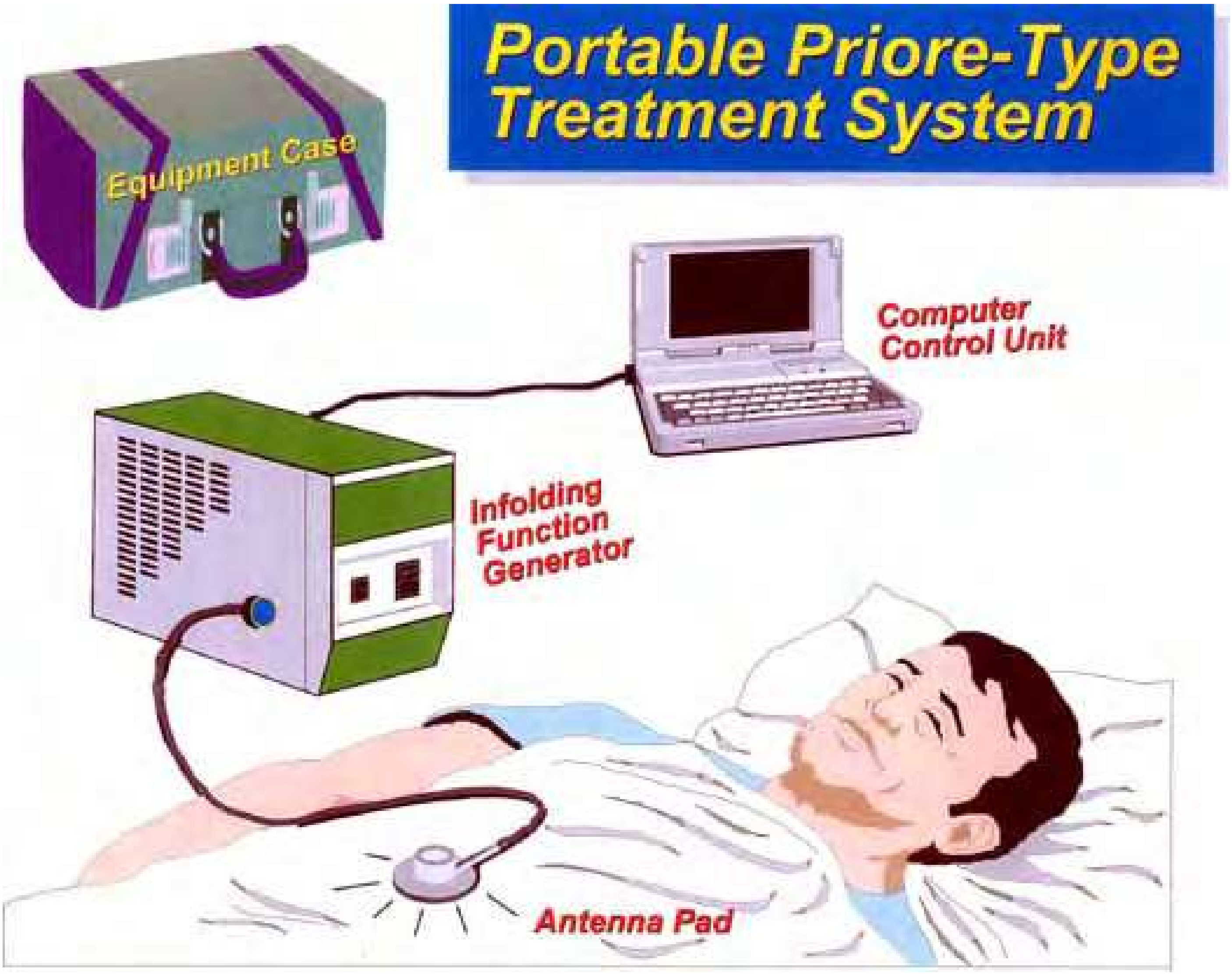
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Figure 13. Portable first-method treatment unit proposed to the Department of Defense. For treating and curing mass casualties from terrorist BW attacks on U.S. population centers.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

be continued for 30 seconds or so. We highly stress that overcharging in this time domain can be detrimental or even lethal if prolonged.

At the end of the very short time-excitation charging time, the body is "time-excitation charged in all its minutest particles". So everything is therefore pumped in the time domain for the next several days after the irradiation is removed, since the time-charge decays very gradually over a period of time. During that decay period, every part of the cell and body acts as a pumped phase conjugate mirror material, pumped in the time domain rather than in the 3-space energy domain. When time-pumped, the mass itself reverses to a previous physical state. This can be used for both nonliving and living masses. Because there is no forcible translation manipulation of the cells and their parts, it can be utilized on living cells in vivo. In the case of diseased cells—e.g., AIDS-it appears that three precisely determined 30-second "AIDS correlate" irradiations a week apart, should be sufficient to completely cure AIDS, including removal of all HIV-genetics changes from every previously infected cell of the body. Also, it would accomplish considerable rejuvenation of the body (making the body younger).

Rejuvenation: Restore the Telomeres and Produce Immortal Cells

Saving Most of the Casualties from Mass Terrorist Attacks

To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S. Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines could be developed cheaply and quickly, and present government facilities could be used to massively attack the correlate database production problem. The portable treatment units could be cranked out by the hundreds of thousands and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals, emergency response teams, etc.

Each device could be used in "assembly line" fashion upon hundreds of sickened patients, in case of a WMD (weapons of mass destruction) attack upon one or more of our cities. The first generation equipment would save probably 70% of the stricken (as from anthrax, etc.). With second generation equipment, one would expect 90% of the stricken casualties could be saved.

Just now, there is a very dim prognosis for all those casualties, using present available treatment capabilities. Time and lack of vaccines defeats just about everything that can be done. Brutally speaking, most of those unfortunate enough to breath in anthrax spores from an

In fact, it is now known that the gradual loss of the telomeres—the natural ends of chromosomes—in a cell's chromosomes is what causes the cell to age. The gradual loss of the telomeres and the resulting shortening of the chromosomes represents a cellular "disease state" or "disorder" state" accompanied by a precise delta in the spacetime curvature engine accompanying every affected (aging) cell.

By determining the correlate for that precise telomere loss in the aging body's emitted radiation spectrum, an effective telomere regeneration process immediately emerges. One simply amplifies that aging "precise telomere reduction" correlate and reradiates the body with it, passing an amplified correlate back into every cell of the body. The cells and all their parts—including the chromosomes—add the phase conjugate replicas, converting the incoming TW EM correlates to longitudinal EM wave correlates. The cells again phase conjugate the LWs and add the phase conjugate replicas, converting the LW correlates to time-polarized EM wave pumping correlates. This latter process time-excitation-charges the aging cells and all their parts, including the genes and chromosomes. The cellular matter itself produces the amplified "aging antiengine" for each and every cell and telomere ending. The result is to again add telomeres, converting the cell back to a younger, more vigorous cell, and reversing the process of aging.

We point out without further elaboration that these methods also introduce direct engineering signals into the body's master cellular control system, including the regenerative subsystem. Hence the body's regenerative subsystem is involved in helping produce the necessary engines and operate upon the short telomere chains to add back telomeres. More specifically, the regenerative subsystem is changed into attack, are going to die. At least 80% of them will expire, in spite of everything the present system can do for them.

Note each device could rapidly treat a whole series patients, one right after the other, merely laying them between "blankets" containing wire antennas, irradiating with the proper correlate spectrum quickly, moving on to the next patient, irradiating quickly (seconds), etc. The entire apparatus could be made to fit in a large suitcase-sized container, and it would be highly portable and controlled by a laptop-sized computer.

Unfortunately, so far the DOD doesn't even understand what the heck we're talking about! The National Institute of Health (NIH) and the National Science Foundation (NSF) are not interested at all. After all, it isn't drugs, vaccines, antibiotics, etc. — which are simply not going to do the job. However, those are what the government and scientific communities are going to finance to the tune of hundreds of millions of dollars. They are then going to lose almost all those mass casualties when the terrorist attacks on our population centers inevitably occur.

Those are Americans. They are going to die. Most of them could be saved. We ought to save them.

For any change in state (physical, chemical, emotional, mental, perceptual), the body dielectric "portholes" change and the output spectrum changes. The delta from zero reference now represents that exact change.

Where Western Mind Control Research Has Probably Arrived

Western mind control researchers probably first established a correlation database of porthole emission correlates to changes in internal states, both physical perfect consonance with the external introduction of the amplified antiengines.

That exact process can be used to rejuvenate the aged population and remedy most of the debilities of growing old. That is how we eventually intend to do it, if we can convince the U.S. government and the U.S. scientific community to perform the necessary research and validation.

and mental. They almost certainly recognize the peculiar participation of every part of the body dielectric in the TW emission spectrum from the body. They will almost certainly have opted for this "porthole" notion of some similar expression of it, and the "thwarting" of these "relaxation and release" emission processes to engender internal physical changes in the body and in the brain-meat-computer, as they see it.

EXPLORE! VOLUME 9, NUMBERS 4 & 5, 1999

They probably experimented with different TW EM radiation mixes, to see what the body emission change to, and what states are induced. Undoubtedly they found that the induced states began to correlate with the former emission delta data base. That is, they found out which delta emissions spectra, reversed and amplified and fed back thorough the portholes, produce what internal changes.

Use for healing is the simplest of all. Use for killing or debilitating is more complicated. Of course weapons researchers wish to look only at the killing aspects. That is something they understand. They do not understand healing, for they do not even have the "regenerative" or "healing" mechanism in any of the scientific papers or texts. Becker[8] came closest in this country, and Priore did it in France. Both of those scientists were suppressed and paid a bitter price for their revolutionary and innovative research.

But look at what has been done, by "input-output" correlation research. The scientists will have a database now, that tells what state emerges in the body and in the mind, for what irradiation spectra (frequency, relationships, phases, amplitudes, wave shapes, etc.) one uses. And they did not work out all the wave-to-wave transductions ongoing in the mind and body connection and in the body and in the mind, individually, between TWÛLWÛTDW.

We suspect that this is the course Western mind control researchers will have taken, without even realizing the internal wave transductions occurring. And they would not be concerned with that. They would have what they want, for weapons, mind control, and behavior control. And of the internal "mix" of the dielectric and all its tiniest portions. So Western mind control researchers and rogue groups probably do get good results.

We are not just picking on the U.S. government here! Internal divisive groups exist in all large and powerful groups, and joust for power. In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old "Power corrupts, and absolute power corrupts absolutely" routine.

The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.

In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians.

they would of course ignore the healing aspects. And they would get results of course, revolutionary results. So they would undoubtedly classify the program very highly, thinking that they had the best of the best.

Rogue Groups May Also Have Arisen to Divert the Purpose of the Projects

And that is where real problems may arise. Every human group has within it competing subgroups, jousting for power. The more highly classified the group and its activities, and the less that is known about it outside its confines, the more strongly and secretly the subgroups can play. So they can emerge as rogue groups pursuing their own agendas, not that of their duly elected governments. In the extreme, such a deep black program can even become a "captured" program, which is totally in the hands of rogues and no longer reports to or is bound by the dictates of the parent government.

It has long been suspected in some quarters that even the US's highly classified research community may be riddled by such rogue groups, and so may be certain highly sensitive parts of the Research and Development community.

Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a

Where the Russians Probably Have Arrived

On the other hand, the Russians will have developed instruments for detecting and isolating the LWs and the scalar waves, in addition to the ordinary mundane instruments for measuring the TWs. So they will have not only the TW correlates, but also the much finer TWÛLWÛTDW (scalar EM wave) correlates. They will understand and employ the use of vacuum engines (spacetime curvature engines) which Western researchers will often invoke but not understand or even be aware that this is what is being invoked. In other words, Western researchers probably just use the old method of "input fi set of functions fi output". That is the standard basis for almost all Western system analysis. With it, one doesn't have to know the exact functions and processes going on in the middle. One just determines a transfer function from a given input to the output that results. Then one catalogs that transfer function. One does that for lots and lots of transfer functions for particularly desirable results, and one has the system analysis. One also has the immediate basis for a rapidly-developed applications technology.

Western scientists are good system analysts. But they have little or no comprehension of what is really going on in that "inside the ship functions" portion, even though they have determined the transform function which acts upon the input to give the "from the portholes" output.

The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification.

Nonetheless, the "simplified" TW EM "porthole correlate" approach will yield very positive results, due to the correlation existing with the LW s and scalar portions the old U(l) electrodynamics. Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW interferometers, through intervening mass

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

including the earth and ocean. Western mind-control scientists and their rogue groups—assuming they exist as postulated—largely will not be able to do that, and probably won't even recognize Russian testing of psychoenergetics weapons directly in or over Western nations.

Russian Lethal Psychoenergetics Mind Control Tests Over the U. S.

In fact, the KGB tested just such highly advanced mind control weapons over the middle of the United States in 1997 on two occasions. The tests were conducted against Captain Button in his A-10 "Warthog" aircraft on April 2, 1997 [See Figure 14], and upon Captain Svoboda in her A-10 "Warthog" on May 27, 1997. [See Figure 15].

Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona. At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—dream-thinking until he died that everything was normal. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour.

Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks. A secondary purpose was to stimulate the U.S. government, watch its investigation and findings, and ascertain if the U.S. knew of the extent to which Russian KGB mind control weaponry had advanced. Obligingly, we proved to me KGB that we did not understand what had actually happened to Captain Button.

Captain Svoboda was from the same Wing as Captain Button and was its most experienced night fighter A-10 pilot. In fact, she taught other pilots the fine points of night flying combat missions in the A-10. About two months after Captain Button's bizarre flight and death, Svoboda was engaged in night target bombing on the range. Rising from her ordnance drop on me target, Svoboda was suddenly struck and rendered hypnogogic and in a waking-dream state. Her sense of the vertical was instantly reversed. Superb pilot that she was, she immediately dream-perceived she was diving, not climbing out of the delivery. She instantly corrected to "climb out". So in the real world she dived sharply into the ground and her aircraft exploded.

This test demonstrated the instant alteration of a skilled pilot's perception, causing that pilot to inadvertently react promptly in an unintentionally suicidal manner. A secondary purpose of the test was to stimulate the U.S. government system again, to see if the system recognized what had

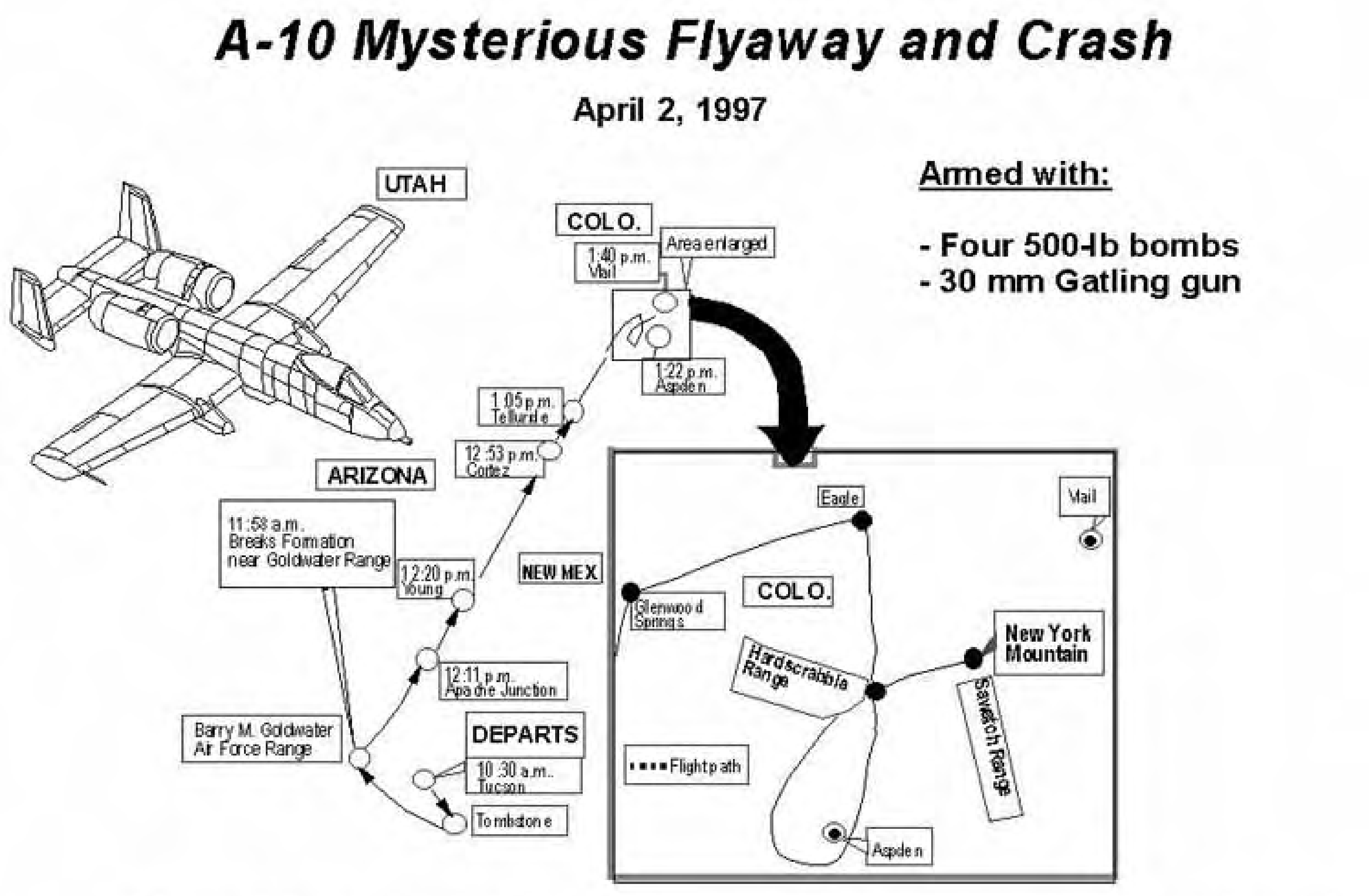
A-10 Mysterious Flyaway and Crash
April 2, 1997
Armed with:
- Four 500-lb bombs
- 30 mm Gatling gun
UTAH
COLO.
ARIZONA
NEW MEX
1:40 p.m. Vail
Area enlarged
1:22 p.m. Aspden
1:05 p.m. Telluride
12:53 p.m. Cortez
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
12:11 p.m. Apache Junction
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a.m. Tucson
Tombstone
Eagle
Vail
COLO.
Glenwood Springs
Hardscrabble Range
New York Mountain
Sawatch Range
Flight path
Aspden

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

occurred. Sad to say, it did not and it has not. It was lamely speculated that Svoboda probably looked over her shoulder like a rank greenhorn, and inadvertently dived her aircraft into the ground while looking back at the target. If there was one pilot in the USAF who would have not done that, it was Captain Svoboda. At any rate, the KGB determined that the U.S. scientific, intelligence, and governmental communities did not have the foggiest notion of what had occurred.

In private proprietary communications[9] we have explained further, to proper parties, exactly the reasons for these two tests, and what exactly was being tested, and what is its planned strategic use.

The Final Race for Direct Mind Control of the Entire Human Species

The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials.

The KGB scientists also know that something very like Gaia— a collective unconscious mind for all species on earth—also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials. If one reads some later books by Kaznacheyev,[10] e.g , this human species unconscious mind aspect and the biosphere unconscious mind aspect emerges loud and clear, if one interpolates between the lines, so to speak. Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.

I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society"—communist style—is still a variation of feudalism (all systems tend to feudalism or some form thereof).

The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now. It is simply a matter of developing sufficiently complex equipment to be able to directly engineer the time-domain substructuring at sufficiently deep levels. It is a doable, although it is very, very complex and very, very difficult. On the other hand, they have been at it for decades. Even in the 50s the Russians already had the old LIDA device,[11] capable of inducing a catatonic trance-like state in a mammal—either a human or a cat,[12] for example. Some versions of Russian mind control devices were used on Russian soldiers in the

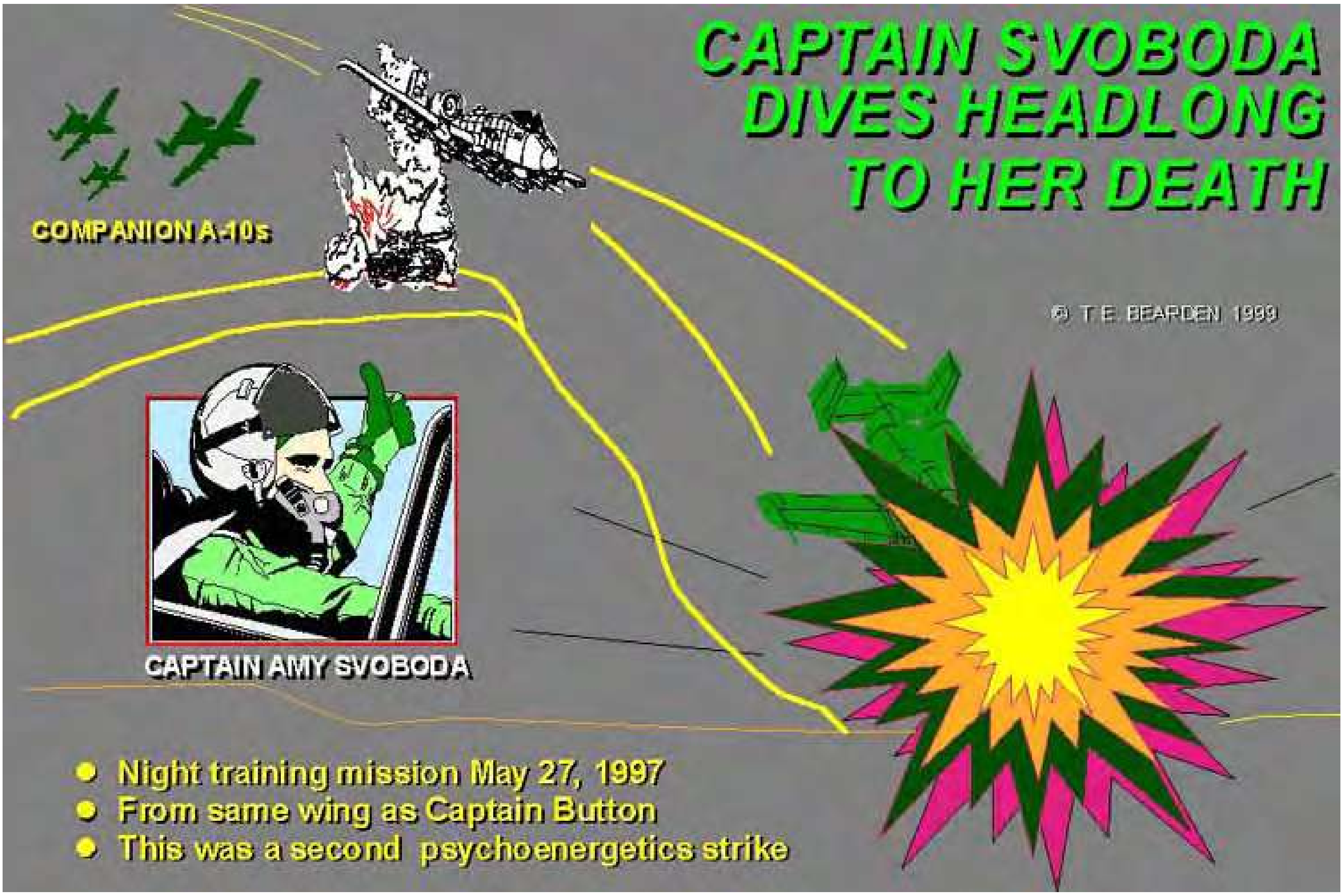
CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH
COMPANION A-10s
© T.E. BEARDEN 1999
CAPTAIN AMY SVOBODA
Night training mission May 27, 1997
From same wing as Captain Button
This was a second psychoenergetics strike

Figure 15. Captain Svoboda dives headlong to her death.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

Afghan War to condition them for performance of merciless acts.[13]

Potential U.S. Rogue Groups and Two Recent Examples

Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.

Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed.

We Certainly Have Proof of Such Rogue Activity. For Example, Here are Two Prominent Cases:

1. For three decades, scientists from the U.S. government, universities, and civilian contractors secretly

as a result of those ghoulish experiments. So what happened when this gruesome thing was revealed? Again President Clinton publicly apologized! The head of the Executive Branch did not turn the macabre matter over to his Attorney General to prosecute. How many of these scientists were indicted? Not one. How many are going to be indicted? Not one. Suppose again this had been done by a private doctor in his own private clinic. You get the point.

There is Little or No Punishment of Rogue Scientific Groups

Shockingly, the U.S. government at the highest level has directly shown (and these are not all such cases by any means!) that mass crimes against U.S. civilians, perpetrated by portions of the U.S. scientific community in direct conspiracy and in secret, will likely be condoned. The perpetrators will not be indicted, tried, or convicted.

Note the connection of the "rogue groups" thesis to the above incidents. Note how rogue scientific groups got away with it in both these cases. There may be other rogue scientific groups getting away with such things today. They should not be able conspired to treat unsuspecting human patients with whole body nuclear radiation, including some retarded children and cancer patients. Some of those patients—including some of the retarded children—died as a result. These illegal experiments were conducted in great secrecy, and the results were highly classified. Eventually these actions were revealed, and a Presidential investigation committee investigated. These experiments would seem to be little different from the WWII Nazi and Japanese experiments on human prisoners. Are those U.S. scientists who were responsible for those retarded children's' deaths any different from the Nazi scientists and criminals we executed at Nuremberg? So what happened when this terrible thing finally came out into the open? President Clinton publicly apologized! How many of the responsible scientists were indicted for murder? Not one. None is ever going to be. Suppose a doctor here in South Alabama, with his own private clinic, had secretly irradiated those human patients and retarded children with whole body nuclear radiation, over a period of 30 years, resulting in some agonizing deaths. How many state and federal agencies would have come after him with arrest warrants? Probably about 50 or so. Would he have been indicted, tried, convicted, and executed or sent to prison for life? Absolutely! Would the President of the United States have apologized? Not on your life; he would have turned it over to the U.S. Attorney General with a strong directive to to get away with it. But they can.

Perhaps a most startling additional part of those two incidents is that there was no great public outcry from the scientific community, deploring these murders and demanding that the criminal scientists be indicted and tried.

As the old saying goes, "By their silence they have convicted themselves." The organized scientific community, as a community, has shown that it has little or no ethics, and—while deploring any scientific murdering that "gets revealed,"—is not really interested in justice. In short, much of the U.S. scientific community may now have very little ethics left.

We are not talking about normal individual scientists, but the Big Science community. There is a whale of a lot of difference between the two. In the Big Science community, there can be and there are rogue groups. Lots of them. There is deep cover, deep classification. And there is very probably advanced mind control research and testing, be it legal or illegal. Hopefully most of it is legal and constrained. However, some of it is almost certain to be illegal and ill constrained.

Human beings are still human beings. All the good and evil is still there, regardless of the group. The stage settings change, but the cast of characters and the play never change.

It's sad, but 'twas ever thus. Hidden parts of our own governments—and other governments throughout the world—are no different from the old medieval groups, where nobles etc. were always plotting against the king, or using the king's power for their own nefarious end. The rogue groups today are no different from all the scheming and conniving groups that destroyed the Roman Empire. Great empires fall from within, not usually from without!

investigate and prosecute on criminal charges.

2. Again for three decades, scientists from (i) the U.S. Government, (ii) universities, and (iii) civilian contractors conspired to give poor syphilitic blacks in Tuskegee, Alabama a placebo while ostensibly treating them for syphilis. The purpose of the program was to deliberately observe and record the ravages of the disease to its fruition in their wracked bodies. So the scientists deliberately sat there and watched their brains rot and their bodies rot. They kept meticulous notes, of course. It was very scientific. And it was macabre. They "treated" some 400 blacks in the "program." Over 100 blacks died

Conclusion

This concludes our brief portrayal of the probable directions that mind control research has taken, in Russia and in the West. It is a novel and powerful area, and it poses both severe dangers and great benefits to all the peoples of the Earth.

Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction of helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth. Yet it can revolutionize medical science, education, communication, and psychology. Eventually it can engineer the mind and memory directly, beneficially. We foresee the day—perhaps 30 years hence—when education will be

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

accomplished by directly loading the software into the mind. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks one will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an M. D. And so on. When that happens, then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.

We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on Earth. If we

446955, Oct. 23, 1998, forwarded to US. Patent Office on Nov. 4, 1998.

5. A very good introduction to nonlinear phase conjugate optics is given by David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(2), March/April 1982, p. 156-183. Also see David M. Pepper, "Applications of Optical Phase Conjugation," Scientific American, 254(1), Jan. 1986, p. 74-83.

6. Pepper, Optical Engineering, ibid., p. 158.

7. See T.E. Bearden, Energetics of Free Energy Systems and Vacuum Engine Therapies, Tara Publishing, Internet node www.tarapublishing.com/books,July 1997. See also T.E. Bearden, "Energetics Update and Summary, " Part I, Explore, 7 (6), 1997, p. 60-67, Part II, Explore!, 7(7), 1997, p. 53-56; Part III Explore!, 8(1), I997, p. 53-56; Part IV; Explore!, 8(3), 1997, p. 56-63. See also T.E. Bearden, "Vacuum Engines and Priore’s Methodology: The True Science of Energy-Medicine: Parts I and II " Explore!, 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62.

8. E.g., see R.O. Becker; Charles H. Bachman, and Howard Friedman, The direct current system: A link between the environment and the organism, " New York State Journal of Medicine, Vol. 62,

do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium.

We are reminded of a poignant quote from Teilhard de Chardin:

"Someday, after we have mastered the winds, the waves, the tides and gravity, we shall harness for God the energies of love. Then for the second time in the history of the world man will have discovered fire."

May the nations of the world utilize the new mind engineering science to truly free all minds into a common love of, and respect for, their fellow humans of every color, creed, and circumstance. As de Chardin envisioned, may we use the new science to discover a new fire, burning clearer and brighter than anything that has gone before. ♦

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel US. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, US. Army. Graduate of Guided Missile Staff Officers Course, U. S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses. Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor: He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuums virtual particle flux is organized and used to energetically shape and manipulate matter and

April 15, 1962, p. 1169-1176. See also R. 0. Becker; "The direct current field: A primitive control and communication system related to growth processes, "Proceedings of the XVI International Congress of Zoology, Washington, D.C., VOL.3, 1963, p. 179-183; —"The neutral semiconduction control system and its interaction with applied electrical current and magnetic fields," Proc. 1lth International. Cong. Radiol., Amsterdam, Excerpta Medica Foundation, 1966, Series 105, p. 1753-1759. See particularly Robert O. Becker and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, Vol. TBD, No. 73, Nov.-Dec. 1970, p. 169-198.

9. Bearden, T.E., Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ page enclosure to CTEC Letter, "Saving the Lives of mass BW Casualties from Terrorist BW Strikes on US. Population Centers," distributed to selected high U.S. Government agencies and Legislative subcommittees, May. 4, 1998.

10. E. g. , Vlail Petrovich Kaznacheyev, V. I. Vernadsky's Study on Biosphere and NooSphere, 1989 [in Russian]. Vernadskys concepts included living matter, biosphere, and human-transformed biosphere (noosphere). Kaznacheyev writes of evolution and

energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.
He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

---

1. Owen Flynn, "Parametric arrays: A new concept for sonar, "Electronic Warfare Magazine, June 1977, p. 107-112.
2. Valerie Hunt, presentation to the U.S. Psychotronics Associations Annual Symposium, Bergamo Center, Ohio, 1988. Hunt presented epochal work showing chaos—and thus hidden order—in recorded ELF electromagnetic radiation from human bodies. Hunt split the received EM signals into two identical channels, time-delayed one channel by a very slight amount, then recombined (mixed) the nondelayed and delayed signals. The result was that attractors precisely indicative of chaos appeared. This positively proves that human systems produce ordered time-structuring of EM.

organization of the living matter as the biosphere of the earth. Without specifying the exact technical details, Kaznacheyev discusses the gist of much of what is in this present article. The present author has an autographed copy of Kaznacheyevs book referenced above.
11. Rabichev, Lev Yakovlevich; Vladislav Fedorovich Vasiliev; Alexandr Sergeevich Putilin; Tatyana Grigorievna Ilina; Petr Vasilievich Raku; and Leopold Pavlovich Kernitsky, "apparatus for the treatment of neuropsychic and somatic diseases with heat, light, sound; and VHF electromagnetic radiation," U. S. Patent. No. 3,773,049, Nov. 20, 1973. This is the U. S. patent for the LIDA machine, developed by the Russians prior to the Korean War. Additional patents or applications are as follows: Italy, #619,129, March 1961. Germany, #1,183,607, Dec. 1964. Great Britain, #826,766, Jan. 1960.
12. Stefan T. Possony, "Psy-War: Soviet Device Experiment," Defense & Foreign Affairs Daily, 12(104), June 7, 1983, p. 1-2. Reports on Dr. Ross Adey's investigation of the Soviet device called LIDA which is used to bombard human brains with radio waves in the 40-MHz region. Also reports on work by A.S. Davydov of the Ukrainian Academy of Sciences, who discovered how the blood-brain barrier can be penetrated by low-frequency radio waves so that brain cells

3. Years ago, I read an obscure reference to Russian research directly along that line, with successful development of just such a portable device. Unfortunately I lost the reference in a disastrous flood in my basement and have been unable to recover it.
4. T. E. Bearden, "Alteration of Materials by Wideband Multiwave Irradiation Combined with Selective Emission Manipulation," Invention Disclosure Document Number

are affected. See also "Soviet device uses radio waves as tranquilizer, " Associated Press release, Huntsville [Alabama] Times, May 19, 1983. Dr. Ross Adey tested the Soviet LIDA machine on a cat, and in two or three minutes the cat was sitting quietly, almost as though transfixed, and uninterested in its surroundings. The animal remained uninterested in its surroundings for some 20 to 30 minutes after the machine was turned off.

13. "KGB Used PIS to Program Soldiers, "Resonance, No. 30, Mar. 1996, p.29. Quoting: "During the 1970s, the Soviet KGB developed a Psychotronic Influence System (PIS) that was used to turn people into programmable "human weapons." Soldiers fighting in the Afghan conflict were trained using the system which employs a combination of high frequency radio waves and hypnosis. Commands are activated by code words or cyphers. The purpose of PIS was to render regular soldiers and members of the special forces incapable of feeling sympathy.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

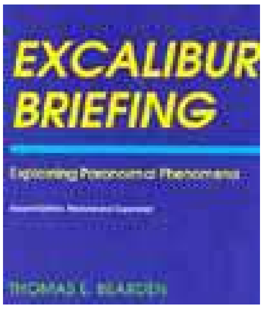

EXCALIBUR BRIEFING

Explaining Paranormal Phenomena

THOMAS E. BEARDEN

THE NEUROPHONE

Another device that uses the new hyperspatial, virtual state, nested modulation technology (and has done so for seventeen or eighteen years) is Dr. Pat Flanagan's neurophone. With brilliant insight and intuition far beyond that of science at the time, Pat invented and patented the instrument by the time he was seventeen years old. The neurophone is a device that, contrary to all present theory and knowledge, will directly "pump the brain" and reproduce sound and information directly in the brain and mind system, without going through the auditory system at all. A simplified diagram of Pat's improved neurophone is shown in Figure 11.

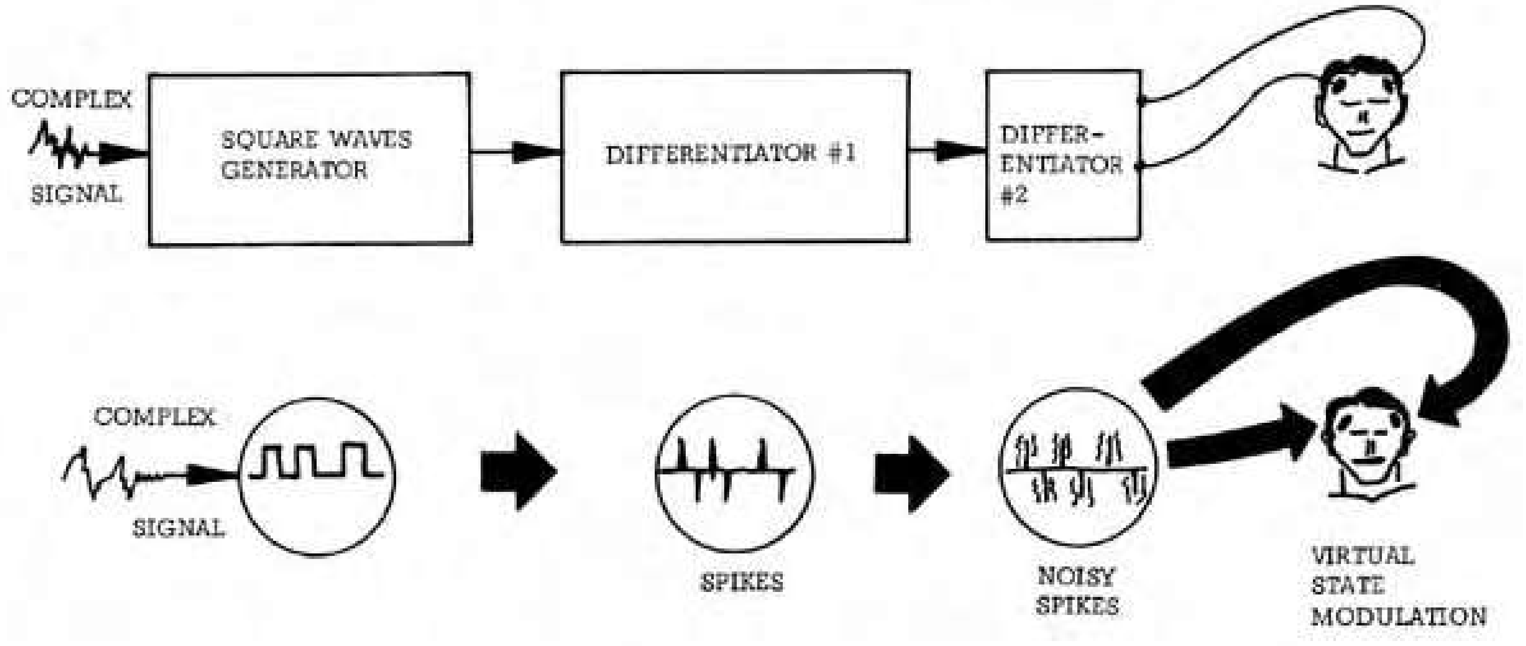
COMPLEX
SIGNAL
SQUARE WAVES GENERATOR
DIFFERENTIATOR #1
DIFFER-ENTIATOR #2
COMPLEX
SIGNAL
SPIKES
NOISY SPIKES
VIRTUAL STATE MODULATION

Figure 11. A simplified diagram of the improved Flanagan neurophone

Briefly, the device takes a complex signal, such as the sound of an orchestra playing a musical interlude, and electrically processes it as shown in the figure. First the signal is passed into a section that clips everything into a series of square waves, remarkably analogous to the sort of clipped waves Lisitsyn confirms are the information carriers of the human brain. Next the square waves are differentiated, yielding a series of sharp spikes (note that these spikes retain the pulse-time content of the clipped signal). These spikes are again differentiated, and since these are finite spikes with real nonzero rise times and decay times rather than theoretical constructs, a series of noisy spikes results from the second differentiator section. From here, the noisy spikes are introduced to special contact electrodes, one of which is normally placed on the forehead, while the other may be placed almost anywhere, including on the foot. Nowhere are any sound waves introduced to the head.

The square-wave clipper section reduces the complex signals, their overtones, and their complex modulations to square waves, retaining the temporal content of the wave mix but not the waves themselves. The first and second differentiators heighten or filter through the temporal content of the higher-order differentiations, that is, they serve as a band pass filter unit to accent the time keying of the neutrinic and mindfield portions or aspects of the signal. When these time spikes are then introduced across the body as pulsed voltages, they are modulated directly on the dendrite firings of the brain and nervous system, providing direct and pulsed modulation of the neutrinic and mindfield component channels of the mind-brain-consciousness-life loop itself. Thus the neurophone directly inputs information into the brain and nervous system, bypassing all the normal sensory systems that lie between the mind-brain loop and the outside environment.

Indeed, a similar development has been reported by two University of Missouri researchers. Dr. Donald York, a neurophysiologist, and Dr. Thomas Jensen, a speech pathologist, have recently reported identifying and decoding twenty-seven words and syllables in specific brain wave patterns and correlating these electroencephalographic patterns with both the spoken word and the silently thought word in about forty subjects. At present, Dr. York and Dr. Jensen are programming a computer with a brain wave vocabulary, to monitor and read the EEG of a stroke victim's brain and help stroke victims who have lost their powers of speech to communicate.

While all such developments can obviously be misused, their potential for assistance to mankind is enormous. Indeed, my own premise that brain linkage is possible and can be used to advance all mankind to the sixth stage of species evolution is slowly being proven. It is within our reach now to develop a multi-channel communications system that will directly link brains, minds, and consciousness into a single functional being; and I have already pointed out that the psychokinetic power of the emergent entity increases exponentially with the number of linked stages. The normal corpus callosum linkage of the two cerebral halves in each human head proves that brains, minds, and beings can be linked and integrated into one.

The Excalibur Briefing details the basic concepts of the theory involved in hyperspace-virtual state engineering, and thus in the engineering of mind-life-consciousness links to multicellular organisms here on earth. Pat Flanagan's neurophone proves that complex information can be implanted directly in the brain and mind, bypassing the normal isolating barriers. Drs. York and Jensen have shown that the brain wave does indeed contain recognizable, decodable information analogues. Lisitsyn has already reported measuring the number of independent channels involved and has documented the importance of the clipped brain waves as opposed to the sine-wave content. If we have the vision, we can still develop a brainlink and rapidly apply it to link sixty to one hundred persons, thereby opening a conscious, multidimensional "eye" for the collective human unconsciousness, gently rousing ZARG into wakefulness, and fulfilling human destiny without undergoing the psychotronic Armageddon planned by the Soviets. The time is short, but it can still be done if the funding, the vision, and the team can be established on a crash priority basis.

SOVIET "WOODPECKER" SIGNALS

But, as I have said, all this can be used for the detriment of mankind as well as for its benefit. Unfortunately, it appears that the Soviet Union has chosen to weaponize the effects on a global scale. A prime example is the "woodpecker" signals emanating from the USSR in the 5-30 megahertz region and interfering with communications around the earth.

These complex woodpecker signals appear to originate from two or three dozen powerful Soviet transmitters, each with a power estimated as high as 40 megawatts (Figure 12).

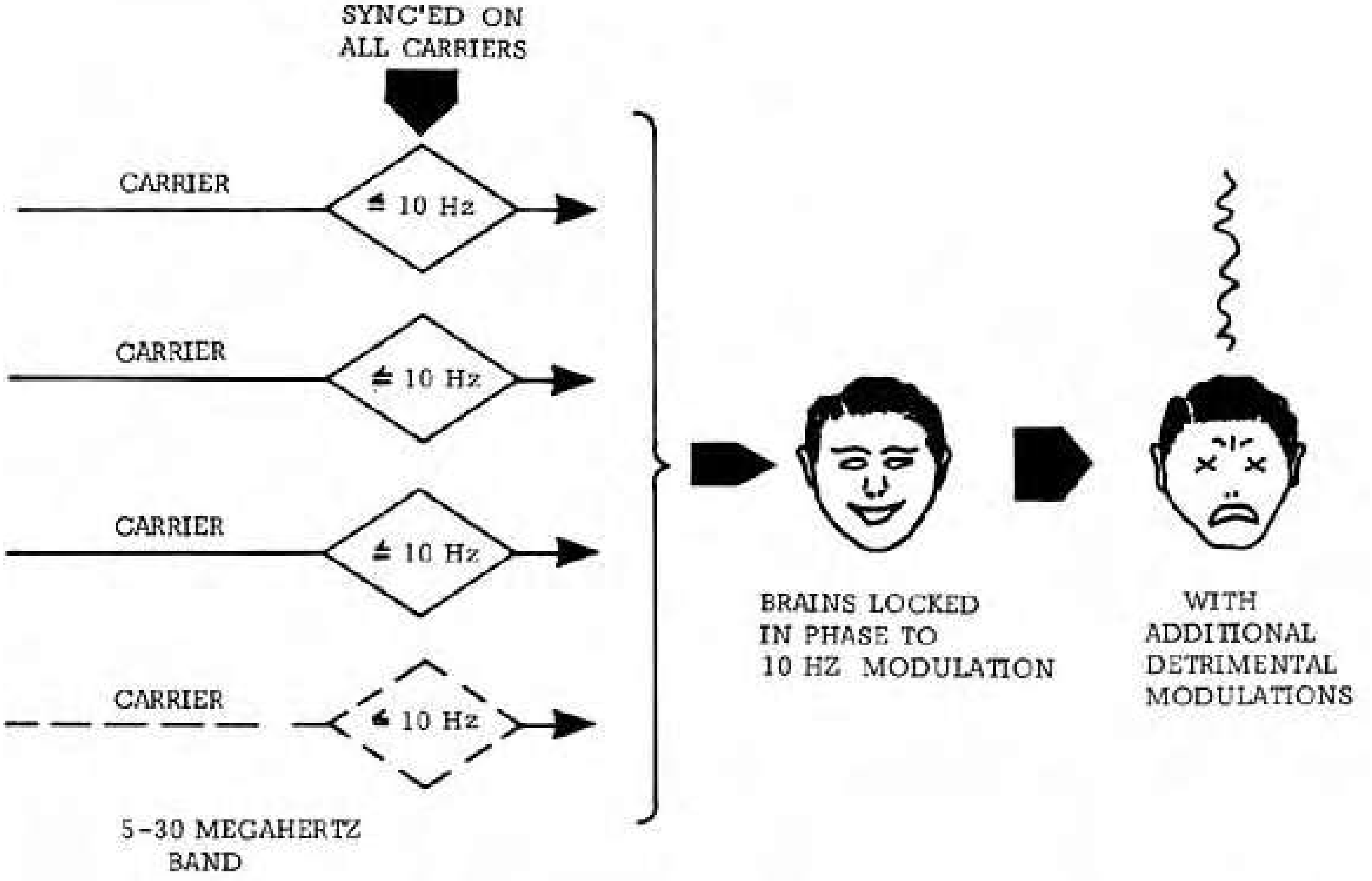

Figure 12. The Soviet "woodpecker" signals

The pioneering experimental measurements of these signals by Dr. Robert Beck and William Bise have shown just how deadly a potential may be possessed by the signals. These measurements have been performed in Eugene and Portland, Oregon; Los Angeles, California; Huntsville, Alabama; and several other locations. They have been particularly significant on the West Coast in and around Eugene, Oregon because of the presence of a direct current transmission line several hundred miles long, which has acted as a long wire antenna, picking up the signals and rebroadcasting them with appreciable gain in the vicinity.

Typically the signals may be found on, say, sixteen different carriers between 10 and 20 megahertz. Twelve of the carriers may appear normal, with normal side-bands, and the other four may have the carrier and both sidebands suppressed but still show the biologically significant modulation (for example, 10 hertz). On all sixteen channels a strong 10 hertz modulation may appear, all perfectly time synchronized and in phase channel-to-channel. The received signal from one of these carriers may be twenty-five to thirty times as strong as the earth's back-ground magnetic field, which is oscillating at approximately 7.0 to 7.5 hertz. Other complex modulation frequencies, many of them changing, are present on the various channels.

Normally, the brainwaves of mammals in an area are gently entrained by the normal oscillations of the earth's magnetic field and are oscillating along at, say, 7.5 hertz. Under continuous radiation from the Soviet woodpecker signals, a percentage (say 30 percent) of the mammalian brains may be captured by the 10 hertz modulation on the overpowering Soviet signals. These captured brains are now in forced oscillation and are riding along in phase with the 10 hertz Soviet modulation. In other words, a certain percentage of the human brains in an area will phase lock to the 10 hertz modulation, and that will effectively lock into those brains the sixteen carrier frequencies with their concomitant frequency mixes and complex signal modulations. If the signals are made much more powerful (for example, by focusing them onto one particular area or increasing the power of the transmitters), then a much higher percentage of brainwave entrainment can be accomplished in the targeted area. By simply adding on the desired modulation patterns, the Soviets can now pump material directly into the mind-brain-consciousness-life loops of the entrained brains. The weapon implications are enormous: raw emotion such as sheer terror or panic can be transmitted; death or disease patterns of all kinds can be transmitted; informational content (thoughts and ideas) can be impressed directly into the captured brains and minds and processed as if originating inside each brain itself. Indeed, as long as the Soviets keep the effects rather gentle, an entire population can be subtly influenced without governmental notice. Certain specific ELF frequencies can rapidly disable or even kill, and for wartime use those frequencies can be directly implanted in the captured brains themselves by the woodpecker signals. It appears that 1984 came a little early, and one can perhaps now understand Brezhnev's strange 1975 proposal to the SAL T negotiations that we should also consider outlawing the development of new weapons more frightful than the mind of man has ever conceived.

The signals also have direct application against strategic targets other than personnel. For example, by incorporating the correct nested modulation pattern, the positron kindling (electron squelching) pattern can be impressed. By beaming through the earth (with a circularly polarized signal or with other complex modulation forms), a beam with a strong neutrinic component, which will squelch electrons and which will interact with great preference for strong voltages or strong magnetic fields, can be directed at a target complex. In the targeted area aircraft and communications systems can be disabled; automobile, truck, and other ignition systems will fail; power transformers and large motors and generators will fail; power transmission systems with their associated high voltages will fail; and so on. Even independent, self-contained emergency power equipment, such as at military installations and strategic missile sites, will fail to operate. Further, to some of these signals, the earth and the sea are as transparent as glass, and so underwater nuclear submarines can be electrically disabled on station before their missiles can be fired, while strategic missile complexes can be electrically disabled before their deadly nuclear missiles can be hurled toward their distant targets. If some submarines and land complexes succeed in firing missiles before being disabled, these missiles can be electrically dudded and scrubbed from the sky over Soviet territory while still far from their intended targets.

So, as stated at the end of *Excalibur,* the players are on stage and in place. The final act begins. One thing is sure - mankind will never be the same again after seeing the play!

VACUUM ENGINES AND PRIORE'S METHODOLOGY: THE TRUE SCIENCE OF ENERGY-MEDICINE

Part I of II

© Copyright 1995 by T. E. Bearden

Photo 1. (to the right) Priore's laboratory. Workbench with animal cage is in the right foreground, under a large overhead assembly to hold the coil for producing the rippling magnetic field.

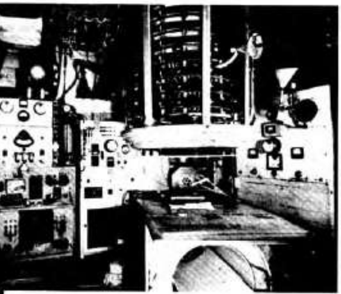

Photo 2. (below) Priore's laboratory. Workbench now has coil installed. The rippling magnetic field descended vertically through the patient. The magnetic field was modulated by the wave/antiwave output of 17 EM waves mixed in a rotating plasma in a giant mixer tube.

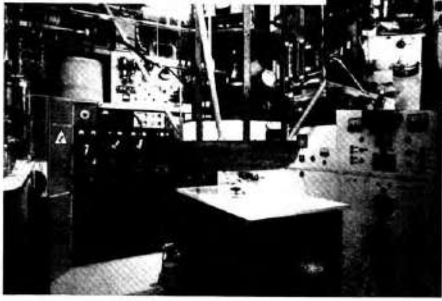

December 1, 1994

Overview

In the late 1960s and early 1970s, Antoine Priore and a team of leading French scientists demonstrated conclusive, total remissions of terminal tumors and infectious diseases in hundreds of laboratory animals. Their research was performed under rigorous scientific protocols and funded by the French Government. The approach employed very complicated mixing of multiple EM signals in a rotating plasma, and modulating the mixed output upon a very strong rippling magnetic field to which the body of the test animal was exposed. Complete remission of the treated diseases was obtained. In addition, the animals' immune systems were also restored to normal. Further, disease-specific antibodies were also created: a single drop of blood drawn from an animal that had been treated and recovered, and injected into the body of a second animal with the same terminal illness, resulted in total remission of the disease in the second animal *and its total recovery*. In the mid-70s Priore's work was suppressed, because of hostility of the oncology community, change of the French Government and loss of further funding, and complete inability of physicists and biological scientists to even hypothesize a mechanism for the curative results.

After a 10-year struggle to discover the Priore mechanism, the present author has proposed a sweeping new explanation for the exact Priore curative mechanism. It represents a totally new kind of medical therapy – one which can dramatically treat and cure a breathtakingly wide variety of diseases. The electrical physics is based on the Stoney, Whittaker, and Ziolkowski decomposition of the scalar EM potential into bidirectional EM wave pairs. Deliberate assembly of such a structure forms a *vacuum engine* of nested curvatures of local spacetime.[1] Priore unwittingly made such a hyperspatially structured vacuum engine, one with the special characteristic that it *time-reversed* (dedifferentiated) diseased cells back to their earlier normal state. The Priore approach not only produces remission of a specific targeted disease, but also fully restores the cells and "factories" of the patient's immune system because they too revert back to their previous healthy condition.[2]

---

[1] *In short, Priore found and utilized a method of electromagnetically engineering general relativity in the laboratory – something physicists had not done and have no knowledge of. Neither have they done it overtly today, and neither do they have knowledge of it today.*

[2] *We point out that chemistry is largely driven by charge and charge distribution. The new methodology essentially controls charge and its distribution. Hence it can control and order all the biochemistry.*

Becker et al have shown conclusively that very weak DC currents and pulsed magnetic fields can produce cellular dedifferentiation (change to a previous state) and redifferentiation (change forward to a new state). It follows that transcription (change of the genetics) must also be altered correspondingly. Becker was nominated for a Nobel Prize for his experimental and theoretical demonstration of the mechanism. He also proposed a fundamental electrical control mechanism to account for his revolutionary results.

Kaznacheyev et al have shown that almost any cellular disease or damage can be transferred electromagnetically from one cell culture to another under special conditions. Popp has theoretically and experimentally shown that an electromagnetic master cellular control system exists in the body, as indeed has Becker. Thus it is clear and independently demonstrated that (1) disease can be created or reversed electromagnetically, and (2) the fundamental cellular control system is electromagnetic in nature. We comment that, viewed as physical materials, the cells and their constituents are sufficiently nonlinear that their EM functions, responses, and effects transcend the present classical EM model. A modern addition to nonlinear electromagnetics is the methodology and results of nonlinear phase conjugate optics (NLO). Accordingly, understanding of the higher topology EM/bioeffects functioning requires application and understanding of NLO principles, *regardless of frequency band.*

We also extend the general relativity model by addition of higher topology EM. We incorporate the Stoney/Whittaker/Ziolkowski decomposition of the scalar EM potential into hidden bidirectional waves. This is then applied to structure and engineer the vacuum. The method is also capable of structuring and engineering both mind and matter.

As a case in point, the elements of the new EM bioeffects approach are applied to the Gulf War Syndrome – the strange conglomerate of illnesses and symptoms developed by many Gulf War participants, usually beginning shortly after their arrival back in the United States from the MidEast. Previously unrecognized biological effects of a dense EM signal environment are advanced and their causative mechanisms explained. These new "dense signal environment" bioeffects, coupled with other simultaneous biologically-active suppressants, allow a full explanation of the Gulf War Syndrome and its characteristics. Both normal and anomalous symptoms are explained.

The Priore approach extends and complements Becker's pioneering approach by allowing positive engineering control of dedifferentiation of cells and their redifferentiation. A cancer cell may be regarded as a dedifferentiation or partial dedifferentiation of a normal cell back toward a more primitive or "immature" geological form. In NLO terms, this is a form of phase conjugation (time reversal) of the cell back toward a previous state, or even its *ancestral* cellular form *(see figure 1 below)*. By phase conjugating the cancer cell anew and amplifying the process, the diseased cell can quickly be reversed back to a normal cell.

It is possible, using the Priore technique, to phase conjugate cells in the body, including all of them simultaneously if need be. Further, conjugation can be made specific to the type of cell or tissue one desires to affect. One need not "know" the specific disease's "EM signal complex," one must simply "pump the cell correctly" where it and its parts are treated as pumped phase conjugate mirrors (PPCMs). In a higher EM topology, the "diseased-cell-as-a-PPCM" already possesses a special "standing signal wave" input that represents the erroneous *vacuum engine* associated with the change to the specific disease state. Pumping the cell *in this special EM topology* creates the precise, amplified *vacuum anti-engine* in this same special EM topology and renormalizes the cell.

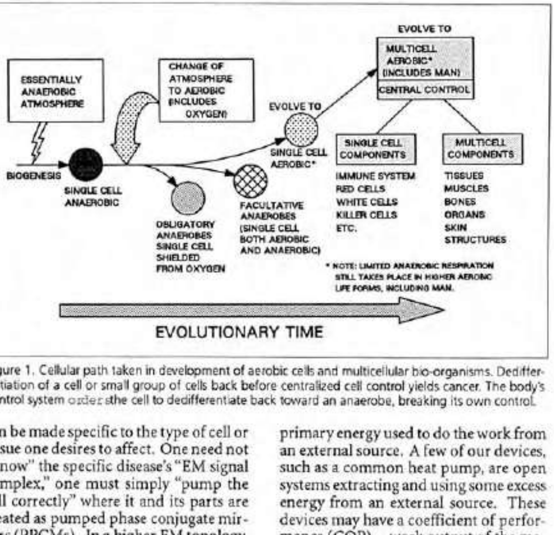
EVOLVE TO
MULTICELL AEROBIC* (INCLUDES MAN)
CENTRAL CONTROL
ESSENTIALLY ANAEROBIC ATMOSPHERE
CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)
EVOLVE TO
SINGLE CELL AEROBIC*
SINGLE CELL COMPONENTS
MULTICELL COMPONENTS
BIOGENESIS
SINGLE CELL ANAEROBIC
IMMUNE SYSTEM
RED CELLS
WHITE CELLS
KILLER CELLS
ETC.
TISSUES
MUSCLES
BONES
ORGANS
SKIN
STRUCTURES
FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)
OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN
* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.
EVOLUTIONARY TIME

Figure 1. Cellular path taken in development of aerobic cells and multicellular bio-organisms. Dedifferentiation of a cell or small group of cells back before centralized cell control yields cancer. The body's control system orders the cell to dedifferentiate back toward an anaerobe, breaking its own control.

Because of the special EM topology utilized, direct engineering of both mind and matter are provided by the new approach, which also represents a unification and extension of electromagnetics, quantum mechanics, general relativity, and biology. It should be capable of reversing most physical and mental disorders that assail a biological organism.

INTRODUCTION

In this paper we discuss the present method of utilizing EM energy in science and engineering, and also discuss a radically different second method that is beginning to emerge *(see Table 1 on following page)*. The main text is an overview, with technical elaboration in the footnotes.

The first (Type 1) engineering method may be likened to a leaky pipe: We always pour in more energy than the work we get back out. We have to furnish the primary energy used to do the work from an external source. A few of our devices, such as a common heat pump, are open systems extracting and using some excess energy from an external source. These devices may have a coefficient of performance (COP) – work output of the machine divided by work done on the machine to run it – of greater than unity. However, the vast majority of our machines have COPs of less than 1.0. Many of them have COPs of less than 50%, and many have COPs of less than 30%.

As can be seen, Type 1 engineering is very wasteful of energy! In a sense, the present struggle to attain room temperature superconductivity is a struggle to eventually have many more machines that attain a COP approaching 1.0, and so decrease the waste of energy in our

Table 1. Two Ways To Do Engineering.

- *Type 1* Engineering: Vacuum Quiescent, Use External Energy Upon Matter
  - Does not involve mind, thought, life, time streams
  - Time is not an observable
  - Must furnish all the energy
  - Western science uses this method.
- *Type 2* Engineering: Gate and Structure the Vacuum Flux Itself
  - Involves mind, thought, life, time streams
  - Time is an observable
  - Furnish only initial energy, and maintenance
  - Use vacuum's energy (vacuum engines)

Type 1 machines. Vacuum engines allow COPs of far greater than 1.0.

After several decades of painful progress in understanding the nature of the quantum mechanical vacuum, today it is accepted in physics that the vacuum itself is a seething inferno of electromagnetic energy. It is thus a legitimate source of energy, could we but extract and use it. This is now accepted both by leading physicists and by the U.S. Patent Office. Further, as of 1993 it has been clearly shown that, contrary to conventional objections, thermodynamics does not rule out the extraction and utilization of the vacuum's energy as heat and power to operate our engines.[3]

We have now reached one of the great scientific plateaus of history: (1) *The vacuum is energetic, containing an incredible density of electromagnetic energy in a violent flux;* (2) *at least in theory, it is possible to extract energy from the vacuum and utilize it;* and (3) *every particle of mass exists as a nonlinearity in a violent flux of virtual particle energy exchange with the vacuum. In short, every system is an open system, being violently "driven" by the vacuum flux, where normally the "driving flux" is in equilibrium or nearly so.*

This paper briefly describes a great new scientific breakthrough which is approaching as a result of the new understanding of the vacuum energy. This new scientific development will usher in the greatest change to humanity of all time *(see Figure 2)*. It is characterized by the laboratory ability to: (4) Deterministically structure and gate the vacuum's fierce flux of virtual particle energy – even at a distance – and make the vacuum itself an ***engine*** to do any kind of work on an embedded physical system that we choose.[4] (5) Deterministically form a quantum potential, and utilize this potential to achieve essentially instantaneous action at a distance.[5] (6) Engineer the hidden inner structure of any potential; this includes also the potential for the Schroedinger equation, including Bohm's ***quantum potential,*** so that physical reality can actually be engineered prior to the collapse of the wave function and the occurrence of the quantum change. This is engineering and altering an entity in its *pre-reality state.* (7) Engineer and alter the flow of time and the laws of nature in a particular area, including the entire planet or solar system. (8) Simply and cheaply treat and cure essentially every present infectious disease and many other abnormalities. This includes cancer, leukemia, AIDS, malaria, sleeping sickness, arteriosclerosis, etc. – even mental disorders. It also includes reversing the effects of the aging process. (9) Directly engineer, form, create, heal, or modify mind, thought, long-term memory, life, etc.[6] And finally, (10) Directly engineer and alter the species quantum potential, an entire species, and the planetary quantum potential for all species and all living creatures on earth.

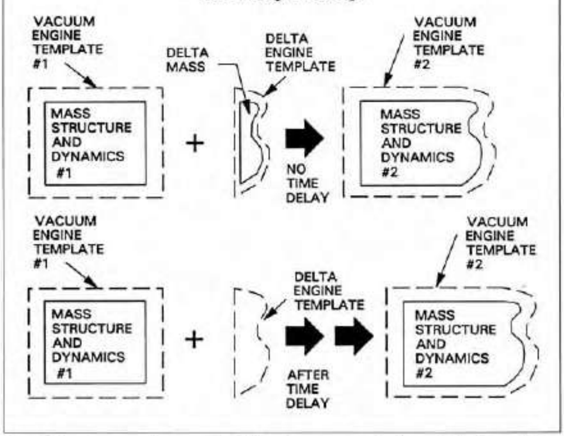
Vacuum Engine Concept
VACUUM ENGINE TEMPLATE #1
DELTA MASS
DELTA ENGINE TEMPLATE
VACUUM ENGINE TEMPLATE #2
MASS STRUCTURE AND DYNAMICS #1
+
NO TIME DELAY
MASS STRUCTURE AND DYNAMICS #2
VACUUM ENGINE TEMPLATE #1
DELTA ENGINE TEMPLATE
VACUUM ENGINE TEMPLATE #2
MASS STRUCTURE AND DYNAMICS #1
+
AFTER TIME DELAY
MASS STRUCTURE AND DYNAMICS #2

Figure 2. The vacuum engine concept. Addition of a vacuum engine (lower block diagram) alters mass and its dynamics after a time delay. Type 1 engineering uses the top level block diagram.

---

[3] *Daniel C. Cole and Harold E. Puthoff, "Extracting Energy and Heat from the Vacuum," **Physical Review E**, 48(2), Aug. 1993, p. 1562-1565.*

[4] *In other words, to perform vacuum engineering, extended from the suggestion of Nobel Laureate T.D. Lee (**Particle Physics and Introduction to Field Theory,** Harwood, New York: Chapter 25: Outlook, "Possibility of Vacuum Engineering," p. 824-828.). Lee recognizes that vacuum engineering can in theory be accomplished, but also that this has never even been tried.*

[5] *For a presentation of the mechanism for creating a quantum potential, see T.E. Bearden, **Gravitobiology**, Tesla Book Co., Chula Vista, CA, 1991, p. 33-36.*

[6] *The weapons developers are slowly grappling their way toward this new science, which the Russians call energetics. Application to the physical world at large is just referred to as energetics. Application of energetics to mind engineering is called **psychoenergetics.** Application to the physical body and its structures is called bioenergetics. In the U.S., development of mind and behavior alteration and control is now being dealt with in embryonic fashion under the subject heading of nonlethal warfare.*

This paper is an informal "quick note" to point out the new scientific plateau and introduce the second kind of science and engineering. The new ***vacuum engineering***[7] offers the greatest threat to humanity that has ever existed,[8,9,10] and at the same time offers the greatest promise of benefit.

For our very survival, it is absolutely imperative that informed citizens be aware of this dramatic change, which is just now starting. The powerful new science and engineering must be controlled and used for humanity's benefit, not its detriment. Else it will eventually be let loose unrestrainedly, to destroy all life on earth – a possibility indicated by Nikita Khrushchev in 1960.[11]

THE FIRST KIND OF ENGINEERING (TYPE 1)

First there's the normal way to do engineering (Type 1), which we are all taught as engineers and scientists. Here the vacuum/spacetime itself stays *inert* and out of our way.[12] We ourselves furnish energy into or onto the system we are trying to move or activate, spilling some of this energy all the while, and we struggle until we get the mass to move and the system to operate and perform work.

This "first kind" of engineering is essentially a continual *wrestling match.* Nature herself is the opponent. The engineer furiously struggles with nature, to force her to do the work he wishes to do. *This is the only overt kind of engineering presently done on earth.* It is primitive and severely limited in what it can do. In Type 1 engineering, the mind and thought are not affected since the *flow of time connectivity* and the *flow of mind connectivity* are not affected.

THE SECOND KIND OF ENGINEERING (TYPE 2)

Clandestinely, however, the Russians (actually, the KGB) in their secret advanced weapons labs, and three other nations not hostile to the U.S., have also weaponized a second kind of science and engineering, but kept it firmly a secret of the highest order. The second kind of engineering (Type 2) is dramatically different, far more powerful, and essentially unlimited in what it can do. Some of the simpler elements have spilled into the open press. For example *(see Figures 3 & 4 on next page)*, Vlail Kaznacheyev and his colleagues have shown, in thousands of meticulous experiments, that almost any disease or damage to cells can be induced in an isolated cell culture by electromagnetic radiation from the diseased or damaged culture, if a full harmonic interval (infrared to ultraviolet) is open for electromagnetic (photonic) exchange between the cultures.[13] The present author's proposed mechanism for the Priore and Becker results also explains the Kaznacheyev results.[14] It also explains the decades-long Russian induction of health changes and diseases in the microwave radiation of the U.S. Embassy in Moscow.[15]

In Type 2 engineering, the background vacuum/spacetime itself is deliberately altered, structured, and made asymmetrical. One then uses this *struc-*

---

[7] *Terminology after Nobel Laureate Lee, ibid., p. 383. Lee points out that the microscopic structure of the scalar vacuum field is not utilized. The particulate flux structure can indeed be engineered, but this has never even been tried by conventional physicists. An excellent coverage of asymmetries and observables is given on p. 181-188. E.g., whenever a symmetry is spontaneously broken, a nonobservable (virtual entity) has become observable. There thus exist vacuum engineering mechanisms for the direct translation between virtual and observable states. Indeed, a whole new vista has opened in physics – some are calling it a "new physics" – due to the discovery that it is possible to directly amplify the virtual state into observable state. E.g., see Serge Haroche and Daniel Kleppner, "Cavity quantum electrodynamics,"* ***Physics Today****, Jan. 1989, p. 24-30 for an overview. In 1979 the present author foresaw, albeit crudely, the emergence of virtual state engineering by means of a possible hyperspatial (virtual state) amplifier. See T.E. Bearden, "Virtual state engineering and its implications," National Technical Information System, Report AD-A065762, 1979.*

[8] *The Russian weapon engineers model the vacuum as a virtual EM dipole plasma, rather than as a simple particle flux; e.g., see A.V. Chernetski, "Processes in Plasma Systems with Electric Charge Division," Georgi Pleckonov Institute of National Economy, Moscow, 1989 for an introductory theoretical exposition of the basic model with respect to physical energy systems.*

[9] *Use of dipoles focuses immediately upon the hidden electromagnetics. Application of plasma theory then exposes the higher nonlinear EM aspects and the characteristics of a higher topology EM theory. Interestingly, Priore used all the elements fitting this theoretical model, employing a giant plasma-filled tube for multi-signal EM mixing, to produce "hidden EM" vacuum engines.*

[10] *Decades ago, the Russians also applied the virtual EM dipole plasma vacuum to the problem of direct engineering of mind, behavior, and cellular control functions. For the fact that they succeeded in decoding the mind/brain/s operations, see S.K. Lisitsyn, "New approach to the analysis of electroencephalograms,"* ***Problems of Bionics (Selected Articles)****, p. 16-25, National Technical Information System Report AD 730045. See also the various references by Kaznacheyev, sanitized by the Soviet State at the time to prevent revealing the actual primary mechanisms.*

[11] *For a partial discussion of Khrushchev's statement of the impending availability of such new Soviet superweapons, see Max Frankel, "Khrushchev says Soviet will cut forces a third; sees 'Fantastic Weapon',"* ***New York Times****, Jan. 15, 1960, p. 1.*

[12] *Technically the vacuum is always highly active, in violent flux exchange with all particles of mass. However, in Type 1 engineering this exchange activity is in equilibrium. So action and reaction are balanced in all directions. With no observable asymmetry, no single net force accumulates for the observer to see, and the vacuum appears to be an "emptiness."*

[13] *See Vlail Kaznacheyev and L.P. Mikhailova,* ***Ultraweak Radiation in Intercellular Interactions****, [in Russian], Novosibirsk, 1981. Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions,"* ***Psi Research****, 1(1), Mar. 1982, p. 47-76 contains an English translation of much of the information referenced in Kaznacheyev's 1981 book. See Vlail Kaznacheyev et al, "Distant intercellular interactions in a system of two tissue cultures,"* ***Psychoenergetic Systems****, 1(3), Mar. 1976; "Apparent information transfer between two groups of cells,"* ***Psychoenergetic Systems****, 1(1), Dec. 1974; V.P. Kaznacheyev, "Information function of Ultraweak Light Flows in Biological Systems," in* ***Problems in Biophysics****, Novosibirsk, 1967, p. 7-18 [in Russian]; V.P. Kaznacheev and A.V. Troftmov, Cosmic Consciousness of* ***Humanity:*** *Problems of New Cosmogony. Elendis-Progress, Tomsk, Russia, 1992.*

[14] *See T.E. Bearden, Gravitobiology, Tesla Book Co., P.O.B. 121873, Chula Vista, CA 91912; 1991. See also Bearden,* ***Cancer and the Unresolved Health Issues in the Biological Effects of EM Fields and Radiation****, Tesla Book Company, 1993.*

[15] *The Soviets radiated the U.S. Embassy in Moscow with weak microwave radiation for decades, starting in the early 1950s. Anomalous health changes and diseases were induced in U.S. Embassy personnel. Three Ambassadors eventually died of leukemia-like illnesses of mysterious nature. The results of the Johns Hopkins scientific study of the Embassy radiation problem also provide strong evidence for the explanation, since health changes and diseases only occurred in those areas of the Embassy where the EM force fields from the radiation were zero. In other words, induced health changes occurred only in personnel in those Embassy areas where the potentials from the radiation had zero gradients, so that the potentials were persistent. This 100% correlation to persistent potentials exclusively, is exceptionally strong evidence that the EM radiation was the actual cause. If independent of the EM radiation, health changes would not have precisely correlated to any component of the radiation! The Johns Hopkins scientists were quite competent, but only considered classical EM theory (CEM) in their assessment. CEM assumes that, when the force fields are zero, the electromagnetic actions are zero. That of course has been disproved in quantum mechanics, which has shown theoretically and experimentally that the potentials, not the force fields, are the primary causes of all EM phenomena. [See Y. Aharonov and D. Bohm. (1959) "Significance of Electromagnetic Potentials in the Quantum Theory,"* ***Physical Review****, Second Series, 115(3), 1959, p. 485-491; Bertram Schwarzschild, "Currents in normal-metal rings exhibit Aharonov-Bohm effect," Physics Today, 39(1), Jan. 1986, p. 17-2.] Even when the force fields are zero, the potentials may remain and provide real changes in charged particle systems by scalar interferometry. Indeed, the force fields do not even exist in vacuum, as CEM assumes. The Johns Hopkins scientists concluded – based on the erroneous CEM model they applied – that since no force fields were present where any of the health changes occurred, the cause could not have been the weak microwave radiation. Had it not been the radiation, some of the health changes would inevitably have occurred where the EM force fields were nonzero. The investigators' results falsified their own study's conclusion, by simple expectation mathematics. Had they known of the SWZ structured potential and its possibilities, they would have realized that by showing 100% correlation to the force-free potentials, they had actually proven that it was indeed the radiation that was the cause, and the force-free potentials component was the actual culprit. The team was excellent; its tools failed it. For a complete discussion, see Bearden, Gravitobiology, 1991, p. 27-32.*

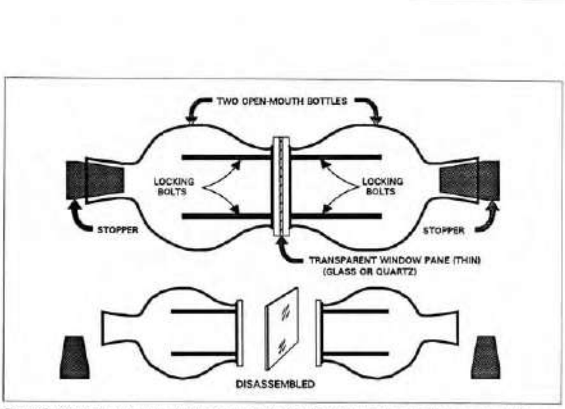
TWO OPEN-MOUTH BOTTLES
LOCKING BOLTS
LOCKING BOLTS
STOPPER
STOPPER
TRANSPARENT WINDOW PANE (THIN) (GLASS OR QUARTZ)
DISASSEMBLED

Figure 3. Kaznacheyev's two-chamber apparatus. A cell culture is divided into two parts. Left ones are damaged or killed. Effect is replicated in right cells through windowglass, but not quartz.

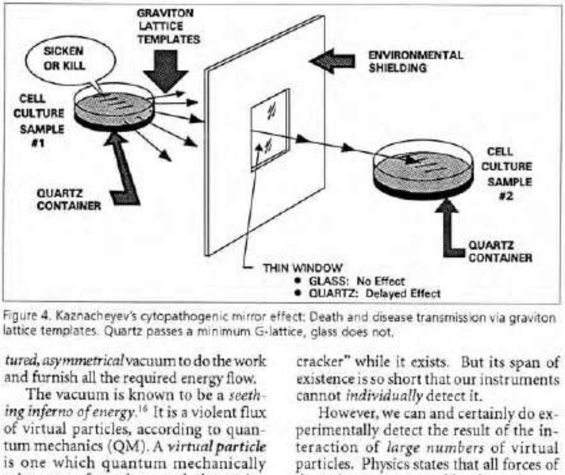
GRAVITON LATTICE TEMPLATES
SICKEN OR KILL
ENVIRONMENTAL SHIELDING
CELL CULTURE SAMPLE #1
CELL CULTURE SAMPLE #2
QUARTZ CONTAINER
QUARTZ CONTAINER
THIN WINDOW
• GLASS: No Effect
• QUARTZ: Delayed Effect

Figure 4. Kaznacheyev's cytopathogenic mirror effect: Death and disease transmission via graviton lattice templates. Quartz passes a minimum G-lattice, glass does not.

tured, asymmetrical vacuum to do the work and furnish all the required energy flow.

The vacuum is known to be a *seething inferno of energy.*[16] It is a violent flux of virtual particles, according to quantum mechanics (QM). A *virtual particle* is one which quantum mechanically arises out of nowhere and almost instantly returns to nowhere. It may be highly energetic and "hotter than a firecracker" while it exists. But its span of existence is so short that our instruments cannot *individually* detect it.

However, we can and certainly do experimentally detect the result of the interaction of *large numbers* of virtual particles. Physics states that all forces of the universe are caused by the exchange of such virtual particles. It is also theoretically and experimentally proven that every particle of mass is in a violent virtual particle flux exchange with the surrounding vacuum. Several experiments directly show this – the Lamb shift for one.[17] The Lamb shift is largely due to fluctuations of the hidden virtual flux exchange with the vacuum, and involves a local energy density far greater than the surface energy density of the sun's surface.

VACUUM ENGINES: ORDERING AND UTILIZING THE VACUUM VIRTUAL PARTICLE FLUX

In Type 2 engineering, one structures and orders the violent virtual particle flows of the fundamental vacuum/spacetime (VST) itself.[18] Accordingly, it alters and orders the vacuum's violent energetic exchange with any mass particles and physical systems in the area. One *oneself* does not input the energy that does the actual work on the material system located in that region of VST. Instead, one only inputs the energy necessary for the *gating and directing* of the VST flux energy in that area. This structuring and gating of the virtual particle flux forms an organized, energetic pattern which I have dubbed a *vacuum engine* (after Nobel Laureate Lee's recognition that the vacuum itself could in theory be engineered, but that such had never even been tried).

This *vacuum engine* now does the work and furnishes all the working energy. We just have to pay to set up the *gating mechanism and the template* in the beginning, and continue to pay a little bit to sustain them. Literally we have created a stupendous *energy amplifier*.[19] By continuing to expend a little maintenance and steering energy, we can gate and utilize the energies of an enormous engine which can apply incredible energy densities, even at a distance. In principle it can apply anything from a feather's touch up to the explosion of a star. It can be used to engineer anything from a single particle or wave, to the atomic nucleus, to a planet or a sun.[20]

---

[16] *For a good overview of the active vacuum, see R. Podolny,* **Something Called Nothing: Physical Vacuum: What Is It?**, *translated by Nicholas Weinstein, Mir Publishers, Moscow, 1986, revised from the 1983 Russian edition. See also I.J.R. Aitchison, "Nothing's plenty: The vacuum in modern quantum field theory,"* **Contemporary Physics**, *26(4), 1985, p. 333-391; Jack S. Greenberg and Walter Greiner, "Search for the sparking of the vacuum,"* **Physics Today**, *35(8), Aug. 1982, p. 24-32; Richard E. Prange and Peter Strance, "The semiconducting vacuum,"* **American Journal of Physics**, *52(1), Jan. 1984, p. 19-21; Walter Greiner and Joseph Hamilton, "Is the Vacuum Really Empty?", The American Scientist, 68(2), Mar.-Apr. 1980, p. 154-164; Timothy Boyer, "The Classical Vacuum,"* **Scientific American**, *Aug. 1985, p. 70. See also important papers by H.E. Puthoff: "Source of vacuum electromagnetic zero-point energy," Phys. Rev. A, 40(9), Nov. 1, 1989, p. 4857-4862; "Gravity as a zero-point fluctuation force," Phys. Rev. A., 39(5), Mar. 1, 1989, p. 2333-2342; and "The Energetic Vacuum: Implications for Energy Research,"* **Speculations in Science and Technology**, *13(4), 1990, p. 247-257.*

[17] *W.E. Lamb, Jr. and R.C. Retherford, "Fine structure of the hydrogen atom by a microwave method,"* **Physical Review**, *Vol. 72, 1947, p. 241-243. Demonstration of the displacement of one of the fine structure electron states of the hydrogen atom by a frequency of about 1060 megacycles played a central role in the development of quantum mechanics, and earned Lamb the Nobel Prize. The observation makes the reality of dynamic vacuum fluctuations inescapably evident.*

[18] *Since the vacuum can be altered, we now must speak of various vacuua, not just one. "A" vacuum, "a" spacetime, and "a" scalar potential are synonymous. "The" ambient vacuum at any local point is a maelstrom of the scalar (hidden multiwave) interferometry of potentials from every charged particle in the universe (adapting Puthoff's Cosmological Feedback Principle). This scalar EM interferometry creates the EM fluctuations of the quantum mechanical vacuum, by Whittaker 1903 & 1904.*

ENGINEERING GENERAL RELATIVITY & NESTING SPACETIME CURVATURES

We actually engineer general relativity (GR) itself, since the Stoney,[21] Whittaker,[22] and Ziolkowski[23] (SWZ) papers point the way to how to do just that *(see figure 5)*. In short, by using the SWZ approach, one is able to utilize electro-magnetics in a new way, to directly form these curvatures of local spacetime. We also *nest* these curvatures to form a *template* of structured multiple curvatures; this template constitutes the deliberate format of the vacuum engine *(see figure 6)*.

Indeed, this is a radically different concept of spacetime structure. It is well-known that, whenever one applies a local hidden variable theory, the GR theory of spacetime structure must be radically altered.[24]

We stress that, in Type 2 engineering, one directly extends, applies and engineers general relativity, but by a novel use of electromagnetics. As can be seen, establishment science is not doing that at all, but is doing only Type 1 engineering. Literally anything one wishes can be engineered in the Type 2 fashion, at least in theory, including transmutation of elements, materialization and dematerialization of matter, free energy electrical power sources, and antigravity. However, Type 2 engineering is even more powerful than that. It also is the ultimate form of virtual reality engineering, for it allows the direct engineering of both physical reality and the mind itself.

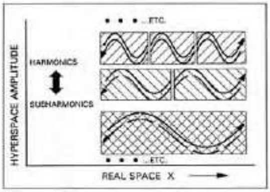

Figure 5. Stoney/Whittaker/Ziolkowski internal EM biwave structure of the scalar EM potential.

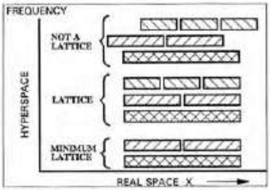

Figure 6. Graviton lattices and nonlattices. Templates require G-lattices. Minimum lattice is one harmonic interval.

MIND, 3-SPACE, HYPERSPACE, MIND FLOW, AND THE FLOW OF TIME

Mind exists in time, but not in our familiar 3-space. Mind does not exist spatially in the same 3-space in which we model physical phenomena.[25] Mind occupies (to the physicist) only one dimension – the time dimension – of Minkowski's 4-space of physical matter. This means that the attachment of mind to matter (in a living biological organism) involves a *mind flow* process, hidden in and occupying the *time flow* process *(see figure 7 below)*.

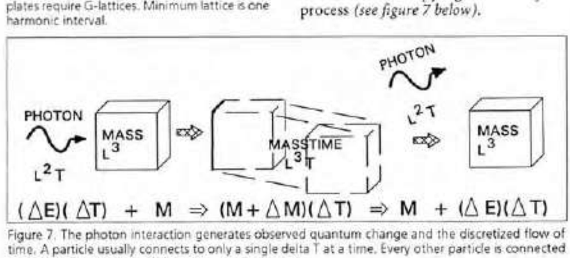
PHOTON
L^2T
MASS L^3
MASSTIME L^3T
PHOTON
L^2T
MASS L^3
$(\triangle E)(\triangle T) + M \Rightarrow (M+\triangle M)(\triangle T) \Rightarrow M + (\triangle E)(\triangle T)$

Figure 7. The photon interaction generates observed quantum change and the discretized flow of time. A particle usually connects to only a single delta T at a time. Every other particle is connected to that small increment of time also, as are all minds.

---

[19] *An example of such an actual amplifier was Floyd Sweet's vacuum triode – a 6-lb., solid-state device containing special coils and two separated barium ferrite magnets. The magnets were specially activated by Sweet's proprietary process so as to powerfully self-oscillate at a chosen ELF frequency, such as 60 Hz. [Such self-oscillation of barium compounds are well-known, but at optical frequencies. Use of a photorefractive magnetic material that is also an insulator allowed Sweet to initiate self-oscillation at ELF frequencies.]. An input sine wave of 10 Volts and 33 microamperes then produced 500 watts of phase conjugate replica EM power, extracted from the self-oscillation energy between the vacuum and the magnets. The amplifier gain (COP) of this open system, self-oscillating, pumped phase conjugate mirror device was some 1,500,000. A successful antigravity test of the device was also performed by Sweet at my urging, and resulted in a controlled reduction of its weight on the laboratory bench by 90%. For as much as could be revealed of the operation and testing of Sweet's proprietary device, see Floyd Sweet and T. E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy,"* **Proceedings of the 26th Intersociety Energy Conversion Engineering Conference** *(IECEC '91), Boston, Massachusetts, 1991, p. 370-375. Unfortunately nothing can be done with the device – e.g., because of formidable legal entanglements due to the conflicting agreements Sweet made with different financial backers. However, McLain & Wooten have produced a repicable, tested, overunity device & full construction details have been rebuilt on Internet.*

[20] *In the modern geometrical view, all individual field energies come from the conglomerate structure called "gravitational field." Transforming one field energy into another kind may be done by first transforming the initial field energy back into G-field energy, then re-releasing the G-field energy as the desired new type of field energy. Vacuum engines allow control and utilization of this process.*

[21] *G. J. Stoney, "XLVIII. On a Supposed Proof of a Theorem in Wave-motion, To the Editors of the Philosophical Magazine,"* **Philosophical Magazine**, *5(43), 1897, p. 368-373.*

[22] *E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics,"* **Mathematische Annalen**, *Vol. 57, 1903, p. 333-355; "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* **Proceedings of the London Mathematical Society**, *Series 2, Vol. 1, 1904, p. 367-372.*

[23] *Richard W. Ziolkowski, "Exact Solutions of the Wave Equation With Complex Source Locations,"* **J. Math. Phys.**, *26(4), April 1985, p. 861-863. See also Ioannis Besieris, Amr M. Shaarawi, and Richard W. Ziolkowski, "A bidirectional traveling plane wave representation of exact solutions of the scalar wave equation," J. Math. Phys., 30(6), June 1989, p. 1254-1269. See also Michael K. Tippett and Richard W. Ziolkowski, "A bidirectional wave transformation of the cold plasma equations,"* **J. Math. Phys.**, *32(2), Feb. 1991, p. 488-492.*

[24] *E.g., see Abner Shimony, "Events and processes in the quantum world," in* **Quantum Concepts in Space and Time**, *Eds. R. Penrose and C.J. Isham, Clarendon Press, Oxford, 1986, p. 192: "...local hidden variable theories may be possible only if relativity theory is replaced by a radically different theory of spacetime structure." (My comment is that this should not be surprising, since classical electromagnetics, quantum mechanics, and general relativity are in profound contradiction of each other, and cannot be unified as they are, in spite of decades of Herculean efforts by many excellent physicists).*

[25] *We may model a mind as a separate 3-space, and its functional operations as "physical" in that hyperframe. Thus a mind and its operations are hyperspatial and outside physical matter's 3-space. Both mind and matter 3-spaces share a common 4th (time) dimension. So it requires at least seven dimensions to model a living creature, eight to model a species, and nine to model all life on earth. The 4th dimension of the mindworld hyper-4-space coincides with the 4th dimension of the physical world's Minkowski 4-space. A priori, mind/matter engineering operations involve engineering and structuring the time stream. This allows engineering either physical reality or mental reality or both simultaneously.*

Normally, so QM tells us, the observation process destroys time, so that all "observables" are spatial only. So the completion of each observable change in a living mass system momentarily destroys the only connection that the mass had with its associated mind! It vanished in the so-called "collapse of the wave function."[26,27,28] Fortunately, little pieces of time are always being serially added and subtracted by the ubiquitous photon interaction process with every mass. Hence the body's particles of mass are continually being *connected to* and *disconnected from* little serial bits of the time stream, and from corresponding serial bits of the *mindstream* of that individual.

TYPE 2 DISRUPTION OF TIME AND MIND

Type 1 engineering does not disrupt or affect the flowing time stream. Hence it does not disrupt or affect the flowing mindstreams of biological systems in the vicinity.

This is not necessarily true of Type 2 engineering! Type 2 engines of appreciable power can and will disrupt and destroy the body's continual detachment from and reattachment to its mind-stream. Type 2 engines will appreciably disrupt and alter the flowing time stream, and thus disrupt and alter the flowing mindstreams of all life forms in the vicinity. The size of the "vicinity" depends upon the size of the engine.

So to beneficially utilize vacuum engines, we do not strive for control of vast energies. We must use a surgeon's scalpel, so to speak, instead of a giant steam hammer. If one seeks weapons, on the other hand, one will strive for control of much greater vacuum engine energies. The KGB secret laboratories, e.g., have pursued this goal for over four decades. Several generations of Type II weapons have been developed and deployed, all under continued KGB control and operation. These Type 2 weapons could, in the words of Khrushchev's address to the Presidium in January 1960, "*...wipe out all life on earth if unrestrainedly used.*" Now one can easily see how that could happen: ***Any massive use of Type 2 weapons will substantially upset the time stream on earth, including all individual time streams. It will so disrupt all the mind streams of all earth's creatures that their connections to their physical bodies will be instantly broken.*** At least now we know why Brezhnev himself was so frightened of these weapons in 1975.[29]

Another unusual characteristic of Type 2 engineering is the *slowly decaying persistence of any special vacuum engine structure artificially created in the vacuum.* We may analyze this effect in terms of the "ghost forms" of quantum mechanics.[30,31,32]

ENGINEERING THE MIND, ITS OPERATIONS, AND ITS CONTENTS

The vacuum engine affects/curves/structures not only the three *spatial* dimensions of ST, but also the *time* dimension. Hence the vacuum engine approach can be used to engineer not only the observable physical matter (which is 3-spatial), *but also the mind and thought (which are time-like to the external observer).* Literally one can "get into the mind's connection channel" and alter the mind, its contents, and its operations at will by using appropriate vacuum engines.

By appropriately and selectively nesting the spacetime (ST) curvature structuring, one can even reach beyond the

---

[26] *Rigorously, the emission of a photon by the living mass destroys its time channel, thus collapsing the wave function and disconnecting the mind. Absorption of another photon immediately thereafter adds an increment of time back to the mass, forming* **masstime** *and reconnecting the time channel and hence the mind channel. Mass does not exist in time per se; masstime does. The temporary disconnection of the time channel by photon emission causes the wave function to collapse; a priori a wave cannot exist without some time available in which to wave! This in fact produces the observable mass from masstime. The problem is that the physicists have ignored the transport by the photon of* Δt *as well as* ΔE*. Thus they have not noticed that the ubiquitous photon interaction with a mass creates the continual addition and subtraction of time increments. This continual coupling and uncoupling of mass to and from the time dimension creates a* **discretized** *flow (passage) of time for that mass. This continual connection and disconnection of a mass to and from only a tiny little piece of the time channel is responsible for the past and the future not being observable; "being observable" is simply being connected to a single small* Δt *increment, not all the increments that have gone before or that will follow hence. When one looks at the corner of a room, the 3-spatiality of one's mass instrumentation system is why one can see length, width, and height but not the fourth dimension, time.*

[27] *Since the photon interaction – including the virtual photon interactions – can be engineered, obviously the passage of an object through the discretized timestream can also be engineered, electromagnetically. Not only can a photon, electron, or proton be time-reversed, but also an assemblage of photons, electrons, protons and neutrons can be time-reversed. Among other things, this means that a diseased living physical cell can be time-reversed (dedifferentiated) back to a previous healthy state, as shown by Priore and also by Becker. Kaznacheyev, as does Becker, also shows that healthy living cell can be "time-forwarded" (differentiated) to a new state, if bombarded with the weak vacuum engine templates (graviton lattice structures) for that state. The minimum graviton lattice (ordering of SWZ wavepairs) utilizes one harmonic frequency interval. Thus quartz will pass Kaznacheyev's cytopathogenic effect, since it passes IR to UV – one harmonic interval. Ordinary window glass does not pass a complete harmonic interval, hence cannot pass the minimum graviton lattice and destroys the disease-inducing vacuum engine template.*

[28] *Since the entire area is highly nonlinear, then nonlinear oscillator theory applies. Graviton lattices may therefore be activated harmonically or subharmonically. This means that "optical band" results can be initiated by analogous subharmonic input – e.g., microwaves or even sound waves. G-lattices work like music; one recognizes a C-chord regardless of the octave. In similar manner, vacuum engines (G-lattices) may be shifted by any number of octaves, with initiation of results then being delayed in the responding physical systems, as prescribed by nonlinear oscillator theory.*

[29] *See Christopher S. Wren, "Brezhnev Talk for Accord against 'Terrifying Arms'," New York Times, June 14, 1975, p. 1, 11.*

[30] *Since statistically every possible change in the vacuum is constantly and continually occurring in each and every differential volume, no matter how small, then from time to time in that maelstrom any physical (and mental) template we care to visualize appears, exists for a very short but finite time, then disappears, then reappears, etc. Such a continual, untidy characteristic is called a ghost form, and usually downplayed or ignored in the teaching of QM because it is "unpleasant." However, anything and everything – no matter how bizarre and whether living or inert – has a small but nonzero "density of reality" in the vacuum flux. Everything that ever was or will be, and everything that never was and never will be, is also there. The point is, creation of a vacuum engine template for a ghost form adds energy to (activates) that form. Just as a Dirac "ghost form electron" can be lifted from the vacuum by adding energy, so can any other ghost form, be it "living" or "dead." The process is fundamental and not just for electrons. Adding energy to a ghost template, we increase the local energy density of that ghost form in all electrical potentials in the entire area. This structuring of vacuum spacetime extends radially in all directions, tailing off to zero at infinity. When the inputting energy is ceased, the energized/activated form will then gradually decay over a period of time.*

[31] *We may compare the activation, persistence, and decay of the ghost form to an accumulator fed by two inputs, one blue liquid and one yellow liquid, with an outlet pipe on the opposite side. Suppose only the blue pipe is turned on; the tank is filled with blue liquid and blue liquid comes out the outlet. Now we turn on the yellow valve, and yellow liquid also starts entering. Over a period of time, the liquid in the tank will gradually turn green, as will the fluid from the outlet. Now we turn off the yellow valve. Over a period of time, the fluid in the tank will gradually turn back to blue color, as will the outlet liquid. Energizing and decaying ghost template structuring of the vacuum works the same way.*

[32] *For inputs that have raised the ghost template near or above the observable (quantum) threshold, this "decay time" may be quite long. Golden once accidentally charged up his local area with such a local template, in a powerful experiment that lasted about eight hours. Materializations occurred, determined by the thought-form (steady ghost form) inputs from the minds of the unintentional participants. In addition, interference with the timestream occurred, as indicated by all local clocks – mechanical, electrical, pendulum, etc. – going substantially astray. The materializations decayed away in a couple of hours, simply dematerializing back into the virtual state when their average energy density level reduced past the quantum threshold. It required four days for the entire "ghost template" to decay away and for the clocks in the area to fully resume normal operation.*

ordinary 3-space and get directly into the higher dimensional space one may mathematically use to model the mind *(see figure 8)*. One can then directly "engineer" in that *pure mind* arena (in the *mental spacetime*, in the modeling sense). This capability, I believe, was responsible for Brezhnev's vivid fear of these weapons in 1975, and his statement to visiting U.S. senators that, not only must the nations of the earth ban nuclear weapons, but also the development of new weapons of mass destruction *"more frightful than the mind of man has ever imagined."*[33] Gromyko later introduced a draft treaty to that effect to the United Nations the same year, and urged the nations of the world to sign it.[34] Article one of the draft treaty stated that the Russians would identify the type of weapons when negotiations on the treaty started. *No one had the slightest notion what the Russians were referring to.* However, three other nations – not hostile to the U.S. – later did develop some of these Type 2 weapons. The Soviet Union was already engaged in operational testing against U.S. and other targets, with incidents mistaken for "accidents" or "anomalies" by the world's news agencies.[35] During the beginning of the 1980s the three nations began checking the Soviet Union by demonstrating their own use of the weapons against Soviet targets. Again the open press did not recognize the true nature of the incidents involved. At the close of the 1980s the triad *vigorously* "checked" the Soviet Union, clearly demonstrating their ability to devastate the Soviet Union should it attempt to devastate friendly nations.

So a special case of "deterrence by massive retaliation" came into being in the Type 2 weapons area as well as in the conventional and nuclear areas. Does the reader recall, e.g., the mysterious explosion of several Soviet missile ammunition storage sites in the space of a few months in 1983-84?[36] A single one of these explosions destroyed a third of all the missiles in the entire Soviet Northern Fleet. There were other major incidents whose true nature and cause went unnoticed.

Frustrated in the intent to dominate the world by surreptitious, delicate use of its Type 2 weapons, the economy of the Soviets simply fell apart. The Berlin Wall was torn down and the Soviet Union broke apart into independent regional nations.

In that upheaval of the Soviet Union, however, the KGB remained intact, well-funded, and still firmly in possession and control of the Type 2 superweapons, including development, manufacturing, deployment, manning, and operational employment.[37,38]

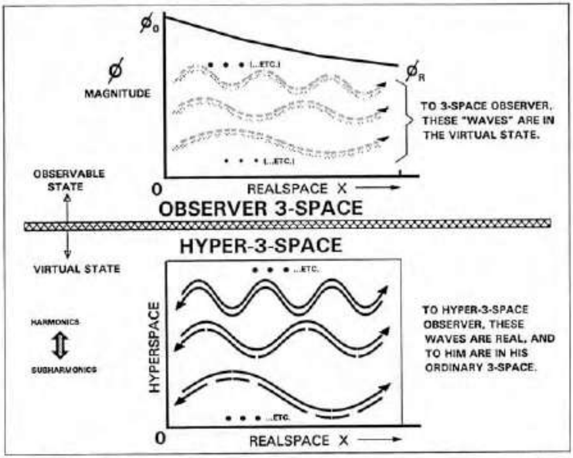

Figure 8. The 7-space EM potential. In observer 3-space (top figure), the potential is composed of a virtual particle flux. In hyper 3-space, the potential is a harmonic series of a SWZ bi-directional EM wavepairs.

WHEELER'S PRINCIPLE AND ITS EXTENSION

Einstein has strongly pointed out the necessity to review foundations *(see table 2)*.

The great physicist Wheeler has a very nice statement on the curvature of spacetime. Briefly, he summarizes the gist of GR as *"Geometry tells mass how to move, and mass tells geometry how to curve."*[39] By "geometry" he is referring to *spacetime*. So we may immediately sub-

---

[33] *Wren, 1975, ibid.*

[34] ***Prohibition of the Development and Manufacture of New Types of Weapons of Mass Annihilation and of New Systems of Such Weapons**, draft treaty introduced by Gromyko at the United Nations' 30th session of the General Assembly, New York, Sept. 23, 1975. The first article provided that the types of these new weapons would be "specified through negotiations on the subject."*

[35] *A long list of these incidents has been published by this author, through the Tesla Book Company. See T.E. Bearden, **Soviet Weather Engineering Over North America**, 1985; **Fer-de-Lance**, 1986; **AIDS: Biological Warfare**, 1988, **Analysis of Scalar Electromagnetics**, 1990, and **Gravitobiology**, 1991.*

[36] *See photograph, **Jane's Defence Weekly (JDW)**, 2(3), July 28, 1984, p. 92; see also Mark Daly, "Goa and Goblet SAMs in Severomorsk explosion," JDW, 2(6), Aug. 18, 1984, p. 224; Derek Wood, "Soviets' northern fleet disabled... 'not viable' for six months," JDW, 2(1), July 14, 1984, p. 3; Derek Wood, "Six explosions in the past seven months," JDW, 2(1), July 14, 1984, p. 3; John Moore, "The aftermath of Severomorsk," JDW, 2(6), Aug. 18, 1984, p. 224.*

[37] *To survive as the leader of Russia, Yeltsin of necessity formed a pact with the KGB. KGB support insured that no coup of substance could be mounted against him, and that the KGB would insure his remaining in power. In return, Yeltsin saw that the KGB was well-funded, remained intact, retained all its secret weapons laboratories and secret weapon scientists and manufacturing facilities, and retained and continued to operate all its deployed, operational superweapons. The KGB changed its name and continued business on a less public profile. Publicly it officially turned a "much gentler face" toward the Russian citizenry. But essentially the KGB still controls most of the Russian military officer corps. Through the KGB/military connection the Communist Party can still accomplish a military coup and seize control of the country again, at least for the foreseeable future, when it chooses to do so. The Russian threat has not disappeared; it is simply **hibernating**.*

[38] *The Communists/Generals in control of the KGB and these superweapons are perfectly capable of turning the unfortunate dissolution into an excursion scenario of their long-planned major concessions offensive prior to ultimately striking the West with a clenched fist and destroying it. The communist aim has not changed one iota since 1917, although the communists have temporarily lost or relinquished control over the Russian civilian populace. For a description of this "29th move of the chess game," see T.E. Bearden, "Glasnost: 29th Move of a 30-Move Chess Game," **Raum & Zeit**, 1990.*

[39] *W. Misner, K.S. Thorne and J.A. Wheeler. (1973) **Gravitation**, W.H. Freeman and Co., San Francisco, 1973, p. 408. On p. 5 the principle is stated as "Space acts on matter, telling it how to move. In turn, matter reacts back on space, telling it how to curve." Thus space is already a vacuum engine that acts on matter, and mass-energy acts on space to curve it.*

Table 2. Einstein on Reviewing Foundations

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

Table 3. Wheeler's Principle and a Corollary

Wheeler

"Geometry tells mass how to move, and mass tells geometry how to curve."

By "geometry" is meant "spacetime." Translating: "Mass tells spacetime how to curve, and curved spacetime tells mass how to move."

Corollary

"Trapped EM energy tells spacetime how to curve. The structuring of that trapped EM energy tells the resulting curved spacetime how to nest its curvatures. This forms a vacuum engine. Its nested curvatures of local spacetime tell mass and energy how to structure and move. By forming vacuum engines deterministically, one may use them to affect, alter, and change local mass, energy, and mind as desired."

stitute synonyms to yield "Mass tells spacetime how to curve, and curved spacetime tells mass how to move."

We further modify and extend this latter statement, replacing mass by its trapped EM energy (this is consistent with general relativity) and adding nesting of such trapped energies. We also allow for altering the timestream, the mind, and the mindstream. With these modifications, a corollary to Wheeler's principle results as follows: *"Trapped EM energy tells spacetime how to curve. The structuring of that trapped EM energy tells the resulting curved spacetime how to nest its curvatures. This forms a vacuum engine. Its nested curvatures of local spacetime tell mass and energy how to structure and move. By forming vacuum engines deterministically, one may use them to affect, alter, and change local mass, energy, and mind in any manner desired" (see table 3).*

To gain a "feel" for spacetime curvature, one may visualize it as analogous to the variation of local pressure in a gas or fluid.[40] Using the earth's atmosphere as a model, we may precisely create pressure points and areas (lows and highs) throughout, and gradually move them at will. This control of highs and lows will allow control and direction of the atmospheric flows. Theoretically, the entire atmospheric flow of the earth can be steered and controlled, by using such an approach. In short one can directly "steer" the weather worldwide, or over a given continent. Present Russian development allows this weather control process to be done sufficiently to control and steer the great jet streams, with concomitant direct influencing of the weather.[41] *We have not had any truly "normal" weather over the U.S. since 1976.*

TYPE 2 ENGINEERING IS JUST BEYOND THE PRESENT VANGUARD OF SCIENCE

As can be appreciated, Type 2 engineering is beyond vanguard science. The forefront of orthodox science is still where Wheeler's original statement is, and does not apply the SWZ methodology nor nesting of spacetime curvature. Science still does not even try to create vacuum engines and perform vacuum engineering *(see table 4 on page 76)*. Conventional scientists are still very firmly rooted in Type 1 engineering.

In ordinary GR, nowhere does there appear a precise definition of the term "gravitational field."[42] The G-field is just a *conglomerate of loose things*, and – by Sakharov's postulate – not a fundamental field of nature at all.[43] It already follows that, since "gravitation" is "geometry of spacetime," then "curvature" of spacetime is achieved by altering and changing the "conglomerates that comprise G-field" from their ambient conditions. The ability to structure and nest such curvatures also follows in straightforward fashion.[44]

PUMPED PHASE CONJUGATE MIRRORS AND THE PRIORE MECHANISM

The astounding thing is that Antoine Priore, using pure intuition, was able to arrive at an exact process for forming a special kind of beneficial vacuum engine![45]

We know today – but scientists did not know it when most of Priore's work was being done! – that a highly nonlinear material acts as a *phase conjugate mir-*

---

[40] *It is actually the variation of the local energy density of the vacuum from its normal value.*

[41] *See T.E. Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985. Available through the Tesla Book Co. In the last few years, the Russians have substantially increased their weather engineering capability and activity. Most of the transmissions from this system are now fairly good electrogravitational waves (coupled wave/antiwave pairs), with EM imperfections many dB below signal level. Hundreds of them are utilized when the system is actively employed over the U.S. By applying Whittaker 1904, these signals – invisible to the normal receiver – can be received and displayed.*

[42] *Misner, Thorne and Wheeler, Gravitation, 1973, p. 399.*

[43] *A.D. Sakharov, "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," Soviet Physics Doklady, Vol. 12, 1968, p. 1040-1041 [English translation.] Sakharov's article (in Russian) is in Doklady. Akad. Nauk SSSR, Vol. 177, 1967, p. 70-71.*

[44] *General relativists already accept the vacuum as containing a 4-momentum-flux river. See Misner, Thorne, and Wheeler, Gravitation, 1973, p. 130-131. They treat the stress energy tensor as a machine (ibid., p. 131). The only difficulty is that they have not realized that, by applying the SWZ approach, much more complicated "vacuum machines" than they have considered can be built and utilized electromagnetically.*

[45] *See Christopher Bird, "The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science," Explore!, 5(5-6), 1994, p. 97-110 for an excellent summary of the Priore Affair. The Priore group cured terminal tumors, sleeping sickness, and arteriosclerosis in lab animals and restored suppressed immune systems. Eminent, even world-renowned French scientists worked with him. The work was funded by the French government and published in leading French scientific journals. It was suppressed when the French government changed to a leftist government in the mid 70s.*

ror[46,47] *(see figures 9 & 10)*. In other words, when one puts in a sine wave, e.g., in addition to producing harmonics the nonlinear material will also produce a strange, backward-traveling replica of the input wave. This "backward wave" has been well-known since the early 1970s. It is truly a *time-reversed* wave (Barus wrote the solution much earlier).[48] It will backtrack back through space, over the *exact* path taken by the input wave.[49,50] If the input wave is continuous, this "phase conjugate replica" (PCR) time-reversed wave will precisely appear everywhere in space that the input wave occupies.

The key here is "everywhere in space." It does not matter how tortuous and twisted the route taken by the input wave; the PCR will backtrack it and appear spatially superposed with it. This has been *experimentally* proven many times in nonlinear optics.[51] One concludes that the vacuum has a memory,[52] and the PCR wave can retrace the memory of the path previously taken by its stimulus.

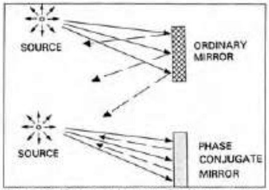

Figure 9. Phase conjugate refection vs. mirror reflection.

Figure 10. Pumped phase conjugate mirror.

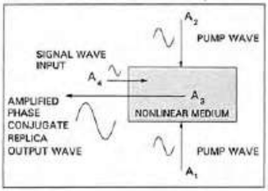

Another nice thing has been discovered and is used in the relatively new nonlinear optics. If one "pumps" the phase conjugate mirror with opposing waves (such as, e.g., bidirectional EM wavepairs), then up to all the energy in the pump waves is absorbed and re-emitted in the "crawfishing" PCR wave.[53] In other words, the stimulated PCM can be made into an amplifier of the time-reversed wave output, just by "pumping" it with the proper bidirectional wavepairs.[54,55]

ANOMALOUS "BIO-JAMMING" AND DEEP DIFFUSION BY EM SMOG

Pumping by EM waves or by simultaneously varying EM forces due to opposing streams of photons will produce a much higher probability of 4-wave mixing and multiphoton interactions. Simply put, this probability is a function of the sheer photon/signal density, along the lines pointed out – but not clarified – by

---

[46] *Elsewhere we have pointed out that, for the primary EM interactions with a living system, we must employ an electromagnetics of higher topology than the conventional EM (either vector or tensor). Barrett has shown that a higher topology EM will involve and exhibit nonlinear optical functioning without necessarily employing optical materials. See T.W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," **Annales de la Fondation Louis de Broglie**, 16(1), 1991, p. 23-41. Thus if we wish to utilize higher topology EM functioning in and on the body/mind, or just to comprehend such functioning, the nonlinear optical processes will form part of that functioning. We must be familiar with phase conjugate optics functions, regardless of the frequency range we operate within, and we must examine the situation accordingly.*

[47] *In the conventional struggle to elucidate EM effects on biological systems, the failure to realize this higher topology principle has led to experiments that are often difficult or impossible to replicate, primary interactive mechanisms remaining a mystery, and the inability to definitively answer even the simplest question such as "Does power line radiation enhance the emergence of cancer and leukemia, and if so, how and under what circumstances?" Forty years of research by able and determined EM bioeffects scientists have thus failed us, because the model they so stringently apply is grossly in error. What they have thoroughly proved is that their fundamental model is incorrect and inadequate for the intended purpose for which it is being utilized. The solution is not to just continue more experiments in the same vein, but to revise the model itself. We have previously pointed out (Bearden, **Cancer and the Unresolved...**, 1993, ibid., p. 40) the present EM bioeffects model, the modification areas required, and how to ascertain when the adjusted model is correct.*

[48] *Carl Barus, "A Curious Inversion in the Wave Mechanism of the Electromagnetic Theory of Light," **American Journal of Science**, Vol. 5, Fourth Series, 1898, p. 343-348 gives the first known solution of Maxwell's equations for a time-reversed or "backwards-traveling" EM wave.*

[49] *According to the well-known "Distortion Correction Theorem" of nonlinear optics. For a precise statement of the theorem, see Amnon Yariv, **Optical Electronics**, 3rd Edn., Holt, Rinehart and Winston, New York, p. 500-501. See particularly Chapter 16: "Phase Conjugate Optics – Theory and Applications."*

[50] *For the space-borne laser, e.g., the U.S. uses iterative phase conjugation – known as self-targeting – of a booster-reflected laser signal to keep a powerful laser beam locked upon the same spot on a distant missile booster, rising during its launch phase, long enough to burn through the casing and destroy the booster, even 10,000 miles away. The process is known as phase conjugate shooting.*

[51] *A striking example is given by David M. Pepper (1986) "Applications of Optical Phase Conjugation," **Scientific American**, 254(1), Jan. 1986, p. 75. Possibly the best extant introduction to the subject of nonlinear optical phase conjugation is David M. Pepper, "Nonlinear Optical Phase Conjugation," **Optical Engineering**, 21(2) March/April 1982, p. 156-183.*

[52] *Without extended discussion we call attention to the importance of this finding with respect to homeopathic medicine and therapy. In fluids, obviously one proceeds from vacuum flux organization and memory upward through matter's organization (and hidden memory), through the individual particles of the nucleus, to the nucleus, to the atom, to the molecule, etc. Between the atom and the molecule levels we encounter hydrogen bonding actions in fluids, which is of great importance to the biochemistry, especially the protein chemistry. It is possible to show that the mind operations and the cellular control operations also SWZ structure the H-bonding activities of the body's aqueous fluid; it can also be shown that any chemical, drug, additive, and EM smog also structures and affects the H-bonding. However, further discussion of these two elements is beyond our scope for this presentation.*

[53] *Note that "pumping" is simply a expansion of the notion of stressing the material. Emission of the amplified PCR wave is a kind of piezoelectric release of the stress energy. The point is that physical stress of a material can be freely tapped and energy continually extracted, if that stress is varying. The McLain/Wooten device uses that principle.*

[54] *To create a normal PCR wave response to an overt EM signal wave input, nonlinear optics uses overt EM pump waves. For the hidden EM signal wave input – i.e., for the hyperspatial vacuum engine as input – one must use "hyperspatial" wave pumping. The "cell as a phase conjugate mirror" must be pumped by hidden (hyperspatial) EM waves inside a scalar EM potential. In the normal case, an amplified, normal EM PCR wave is produced in response to a normal signal wave input and normal pump waves. In the vacuum engine case, an amplified, hyperspatial EM PCR wave complex is produced in response to a hyperspatial signal wave input and hyperspatial pump waves.*

[55] *Note that the amplified PCR wave represents a reorganization and deterministic redirection of stress energy. Hence it is a time-reversed operation, and also a negentropic operation. Since the PPCM operates as an open system not in thermodynamic equilibrium, with input of excess energy in the form of the pumping, negentropic operation is permitted by nonlinear thermodynamics (of a system far from thermodynamic equilibrium). For a lay summary of such thermodynamics, see Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," in Paul Davies, Ed., **The New Physics**, Cambridge University Press, Cambridge, 1989, p. 316-47. For a more technical presentation, see G. Nicolis and I. Prigogine, (1987), **Exploring Complexity**, Piper, Munich, 1987. Prigogine received the Nobel Prize for his pioneering work in this field.*

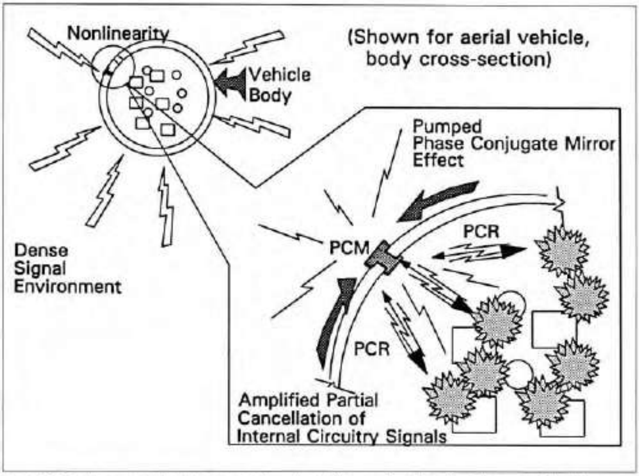
Nonlinearity
(Shown for aerial vehicle, body cross-section)
Vehicle Body
Pumped Phase Conjugate Mirror Effect
PCR
PCM
Dense Signal Environment
PCR
Amplified Partial Cancellation of Internal Circuitry Signals

Figure 11. Single versus multiphoton penetration effects. Phase conjugation effects become of significance, generating PCR signals which couple to and backtrack the cells' and tissues' own signals, this penetrating throughout the body with weak partial cancellation jamming.

Devyatkov et al.[56,57] Thus dense electronic smog – even though it contains only extremely weak signals – increases the nonlinear optics type of interaction. This allows anomalously deep penetration and *diffusion* of special weak EM signals due to the "pumping" aspect of multiphoton interaction. The nonlinear cells and cell parts act as pumped phase conjugate mirrors (PPCMs) when subjected to multiphoton interactions.

Worse, the *signal wave* inputs to the PPCMs are the normal weak EM signals of the body itself. Hence the resulting PCRs from the PPCMs are superposing phase conjugates of the body's normal signals. This results in (1) a "diffusive" effect in the body of the PPCM formations, and (2) formation of very weak "EM cancellation" signals throughout the body, specifically targeted to the body's own control signals, including those to and of the immature immune cells in the bone marrow. In the bone marrow, "*partial-cancellation jamming*" *(see figure 11)* occurs in the normal control and development of the immune system cells as they are created. Slowly errors and degradation damage accumulate with time, including over the years. This cumulative immune system damage combines with other damage effects due to pesticide residues, pollutants, etc. The combined degradation results in long-term cumulative damage to the immune system and its cells. Increasing errors in the immune system's functioning eventually result in appearance of immune system disorders and increasing vulnerability to

Table 4. Mechanism for the Priore Therapy

- Produced Specific SWZ-Structured Scalar Potential
  - -17 EM frequencies mixed in a rotating plasma
  - Phase conjugates added by the plasma
  - Formed vacuum engine with hyperspace pump waves
  - Modulated onto a rippling magnetic field which carried it to every cell and all its parts
  - Jyperspatial pumping of cells/parts as PPCM's
- PPCM's created amplified anti-engines for disease there
- Anti-engine reversed cells/parts back to normal state
  - Cells return to control by cellular control system
  - Restored cellular control system to full functioning
- Scrubbed out the cumulated precancerous state
- Eliminated shadow disease state
- Restored the immune system to high functioning
- No excessive trauma to treated animal

In Part II of this wonderful article, by Mr. Bearden, a few of the topics you can look forward to are; The Gulf War Syndrome, How to Make a Gravitational Wave & Curve Spacetime, Potentials & Gravity and How Priore Made Vacuum Engines.

---

[56] *The Russians refer to the hidden EM wave structure of the scalar potential – i.e., to the SWZ structuring and vacuum engine templates – as the information content of the field. Western scientists assume this phrase refers to ordinary spectral analysis and communications, ignoring those statements that the normal EM theory does not apply. Also, Devyatkov et al seem in no great hurry to express their true meaning more clearly. See N.D. Devyatkov (Ed.),* ***Applications of Low-Intensity Millimeter-Wave Radiation in Biology and Medicine,*** *IRE Akad. Nauk. SSR, Moscow, 1985. [in Russian]. See also N.D. Devyatkov, Yu. V. Gulyaev, et al, [Organizing Committee Chairmen],* ***Digest of Papers - International Symposium on Millimeter Waves of Non-thermal Intensity in Medicine,*** *co-sponsored by Research and Development Association "ISTOK" and Research Institute of U.S.S.R. Ministry of Electronic Industry ("ORION"), Moscow, Oct. 3-6, 1991. Collection of 300 papers. [In Russian].*

[57] *However, some leading Western scientists – such as Wheeler, Feynman, and Jaynes – strongly advocate replacing the present erroneous concept of the EM force field existing in spacetime with the notion that what exists there instead is an information storage device containing a summary of what we need to predict the behavior of a charge if there should be one at the point. E.g., see J.A. Wheeler and R.P. Feynman, "Classical electrodynamics in terms of direct interparticle action," Reviews of* ***Modern Physics,*** *Vol. 21, 1949, p. 425-433; E.T. Jaynes,* ***"Probability in quantum theory,"*** *in Complexity, Entropy, and the Physics of* ***Information,*** *Ed. W.H. Zurek. Addison-Wesley, Redwood City, CA, 1990, p. 33-55.*

VACUUM ENGINES AND PRIORE'S METHODOLOGY: THE TRUE SCIENCE OF ENERGY-MEDICINE

Part II of II

© Copyright 1995 by T. E. Bearden

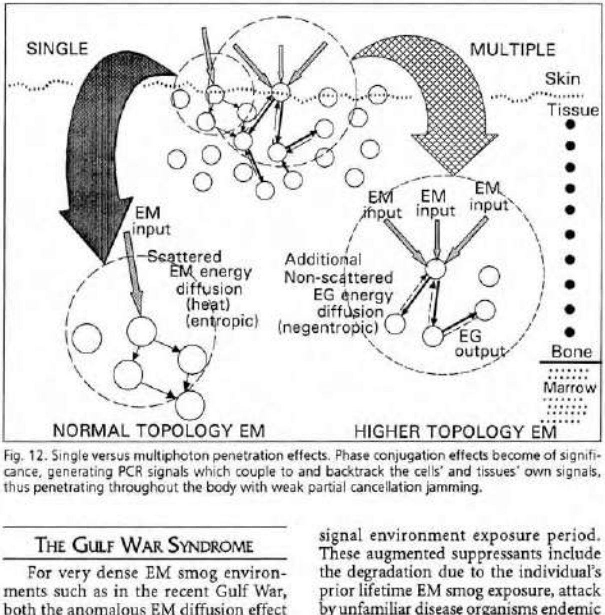

Fig. 12. Single versus multiphoton penetration effects. Phase conjugation effects become of significance, generating PCR signals which couple to and backtrack the cells' and tissues' own signals, thus penetrating throughout the body with weak partial cancellation jamming.

THE GULF WAR SYNDROME

For very dense EM smog environments such as in the recent Gulf War, both the anomalous EM diffusion effect and the partial cancellation jamming effect on humans are dramatically accelerated. Cellular and immune system damage that would require decades of exposure at very low EM smog levels will cumulate in only a few months. In turn, this accelerated malfunctioning of the cellular control system and the immune system enhances the effects of other biological suppressants already present in the body or encountered during the dense signal environment exposure period. These augmented suppressants include the degradation due to the individual's prior lifetime EM smog exposure, attack by unfamiliar disease organisms endemic to the areas, exposure to oil fire smokes and pollutants, exposure to increased UV from solar radiation, and exposure to chemical agents and possibly surreptitious biological warfare agents.[58]

Further, in the haste to protect the troops from possible chemical and biological attack, all soldiers were administered a nerve agent called pyridostigmine bromide; in addition, approximately 150,000 were inoculated with anthrax vaccine, and 8,000 also received experimental botulism toxin shots.[59] Normally for medically-administered agents, the relatively minor suppression effects on the immune systems and body are followed by quick recovery, due to a properly functioning immune system and cellular control system. However, all three of these Gulf War preventive medical treatments do create "small cases of sickness" in the normal body.[60] In other words, a low level of the disease strain *is created* in the patient. Further, even after the apparent "initial recovery," there will now be a much longer dormant "apparently healthy" period during which the *vacuum engine templates for the disease remain, until they have finally decayed away.* In a sustained dense signal environment, this "decay away" period may be dramatically extended or never occur; in fact, the disease may progress from the vacuum-engine state to the low level physical state, then into a fully developed syndrome.

So we introduce the concept of the *relative level of a disease.* First, the disease and/or its vacuum engine templates are in the body. A low level or absence of disease, as we normally think of it, is with respect to the high disease level – the full-up, developed state of the disease. A disease level is "low" because of the ability of the immune system to fight it, overcome it, and reduce it to a "pre-disease" state; and the ability of the cellular control system to eventually rid itself of the disease vacuum engine templates, retaining only antibodies against its re-invasion. When the disease is not physically present according to conventional medical standards, but nonetheless the vacuum templates are present and not yet cleared out, we shall refer to that state as the shadow state of the disease. The cellular control system largely combats this state.[61]

---

[58] *Both Senator Riegle's and Senator Rockefeller's May 1994 hearings concluded that all 695,000 troops who served in the Gulf War were also exposed to chemical and biological toxins.*

[59] *Stephen Salgaller, "U.S. Role in Gulf War Illness." Letters,* ***Science News****, 146(7), Aug. 13, 1994, p. 99.*

[60] *Note that at least one of these three treatments was still quite experimental. Consequently there exists less confidence in stating that its normal side effects are minimal and tolerable.*

[61] *Apparently by being able to phase conjugate with weak amplification, so that it forms anti-engines that usually are stronger than the disease's vacuum engines, gradually reversing the shadow state to normal.*

If the immune system and cellular control system are impaired, a shadow state of the disease can gradually progress to a "low level" of disease, then to a higher level, and become serious or even life threatening. This is true when conventionally the disease is physically absent but its shadow state is still present and cannot be disposed of by the cellular control system, as clearly shown by Kaznacheyev's experiments.[62,63]

We point out, but do not pursue, that the cellular control system also deals with the shadow states of diseases. In the normal state of "free from disease," the cellular control system successfully and continually deals with incident diseases, overcoming them at the shadow state level. Its process for continual disposal of the shadow state residues is hypothesized to be a version of slightly amplified phase conjugation of the vacuum engine's shadow state level.

In the model we are advancing, the disease entry and development occur by successively (1) creating its shadow state and successfully maintaining it in spite of the cellular control system's phase conjugation efforts, (2) breaching the threshold of physical changes, so that at least physical indicators of the struggle by the now-involved immune system appear, (3) progressing until conventional medicine recognizes "low level disease or infection," (4) increasing until conventionally-recognized "developed disease or infection," (5) increasing until the disease is serious and even life-threatening, and (6) increasing until the patient can no longer sustain life and dies. *We note that the best preventative medicine would and should be practiced at the shadow infection level.* The Priore therapeutical approach is capable of doing that.[64]

In the "normal" – if such an adjective can be used! – sustained "weak EM smog" exposure, the shadow state may last for several years or even several decades, prior to sufficiently accumulated degradation of the cellular control system and the immune system to allow the physical emergence of the disease. For much denser EM smog, such as existed in the Gulf War, the shadow incubation process is highly shortened, as is the time to progress from one stage to another higher stage of the disease or disorder condition.

In other words, in a dense but weak EM smog environment, the recovery from "minor diseases" actually administered as part of the initial vaccination and preventive medicine will be substantially altered because of the accelerated damage to the immune system and cellular control system. The control/immune degradation may be fast enough that the disease starts increasing from the shadow state toward higher levels, instead of being reduced and eliminated by the control/immune systems.[65,66]

Consequently, the Gulf War participants were thrust into a unique new situation: The second (shadow state) phase of decay of the "minor diseases" created in the bodies receiving the three preventive medications was interfered with by the EM fog. Because of individual variation in past cumulative damage experience, the individual control/immune systems varied from soldier to soldier. The smog environment varied in density and intensity from place to place, and from time to time. Hence the degree of interference varied from soldier to soldier because of (1) variation of past control/immune degradation, (2) variation of the EM smog exposure, and (3) variation of exposure to other suppressants. A bell-shaped distribution curve of the extent of the *net interference to net control/immune functionality*[67] would result, when the soldiers are taken as a group. That is, some of the participants just had a somewhat longer 2nd phase recovery period; some an extended 2nd phase recovery period; some a sustained, gradual rise in the 2nd phase, toward physical disease but not yet developing the Gulf War Syndrome. Some had a sustained, quicker rise toward actual physical disease instead of a decay, and have had sufficient time to develop the physical disease.

When the additional individual expo-

---

[62] *Kaznacheyev, several references, ibid. That is, suppose we have the shadow disease possess only a strength such that the normal cellular control system and immune system could overcome it. If the cellular control system is weakened, the disease progresses across the physical barrier. If the immune system is still strong, it can handle the encroaching disease, but will give physical indicators of its struggle (increased white corpuscles, etc.) If the immune system is also degraded, the disease will progress to low level physical disease, higher level, etc. On the other hand, if the cellular control system is normal but the shadow state of the disease is amplified, the phase conjugation defenses of the cellular control system will be breached and the disease will impel itself into low-level disease, then high level, and then kill the cells, in spite of a normal immune system combatting the breach of the cellular control system. This latter is what Kaznacheyev's experiments showed. It is also what the decades of Soviet weak microwave radiation of the U. S. Embassy in Moscow showed.*

[63] *The discerning reader will note that we are presenting the gist of a new theory of cellular disease.*

[64] *We also point out the electromagnetic biological warfare capability in this area, long since developed by the Russians and under total control of the KGB. The U.S. Embassy radiation, including the health changes and the death of three U.S. Ambassadors, was just to stimulate the U.S. intelligence and scientific communities at the highest levels. By the U.S. electromagnetic response at the Embassy itself, the Soviets could then know for certain whether or not the U.S. was aware of this EM biowar capability, and whether or not it had secretly developed defenses. Our actions assured them we had no such capabilities and no defenses – and were totally unaware of the true nature of the Embassy radiation.*

[65] *Note that the multiplicity of the assault on the control/immune systems augments the resistance and recovery difficulty. E.g., each B-cell lymphocyte can attack only a single antigen – alien property of the surface of an invader. There are about one billion B-cells in the human body, which between them are specific for about a million antigens. As can be seen, the immune system is continually having to develop B-cells with new specificities due to the continual meeting of new strains of antigens. Cumulated damage to this system affects both the body's resistance to diseases it has already acquired specific B-cells for, and its resistance to new strains for which it must make B-cells for new antigens. There is a sort of threshold, at any given time, beyond which the B-cell responses cannot keep pace, and start being overwhelmed. There is also an error tolerance level; when the errors due to deterioration surpass this level. the immune system will often begin attacking the body's own tissues in error. This results in disorders such as arthritis and cartilage and joint deterioration. With sufficient errors, previous immunity abilities also are degraded, so that more ordinary diseases become serious threats.*

[66] *The administering of three "lightly suppressive" treatments closely occurring with the soldier's entry into the Combat Zone thus posed a significantly greater problem than normal to his or her control/immune systems' operations, because the control/immune systems were suddenly degraded and hampered by the suppressive effects of the dense EM signals environment. In other words, the control/immune systems were moderately attacked by the three medical treatments (three enemy light cavalry divisions), and at the same time the entire control/immune system complex was subjected to substantial interference and suppression (the normal friendly) troop divisions were simultaneously placed under fierce and sustained attack throughout the zone. A general whose divisions are already committed in furious combat and under heavy bombardment, as are all his command posts and communications systems, has great difficulty freeing up, reconfiguring, and recommitting some of his committed troops to properly combat three other enemy forces that suddenly have appeared over the horizons from other directions!*

[67] *We may also define this as a disease progression ratio. This ratio should assume a distribution curve. For first-order simplicity, we are assuming a normal distribution.*

sures are factored in, there should have been essentially three broad groups: (1) relatively minor damage sustained, where any overt effects will only begin to appear many years hence if at all, (2) more damage sustained, where overt effects will begin to appear within the next few years, and (3) heavier damage sustained, where overt effects begin to appear either while still in the Gulf Region or within some months or so after returning.

Hence many of the Group 3 veterans of the Gulf war developed anomalous infections and immune system disorders shortly upon their return from such damage-accelerating exposure, resulting in the mysterious "Gulf War Syndrome." Conventional medical science will recognize – and has recognized – the overt phase 3 physical disease syndrome, but its cause and effective treatment remain and will remain medical mysteries. Conventional medical science will entirely fail to recognize Groups 1 and 2 cases, and will label them as due to aging, mental stress, or ordinary rise of diseases.

The syndrome has involved the "anomalous EM smog diffusion" process, as previously explained. It may – and in many cases it will – exhibit certain "signature" differences from normal diseases. The genetics of the victim are affected, and so are altered by induced transcription[68] to carry at least a portion of the "syndrome template." Semen, ova, saliva, and blood will particularly be so affected. A wife may complain that her husband's body fluids are now irritating to her skin, even producing "burning" sensations. Children conceived and born may be borne with the syndrome, or with one or more aspects of it.[69] The syndrome template can be transmitted by exchange of body fluids, so wives may also contract the syndrome and develop the disease from their husbands.[70]

CONVENTIONAL EM BIOEFFECTS KNOWLEDGE IS INSUFFICIENT

Present "EM bioeffects" scientific models contain absolutely no modeling of partially jamming the body's own special EM control system by turning its EM signals partially into G-signals, thereby partially nullifying them. It also contains no modeling of the anomalous EM diffusion process.[71] Thus the conventional medical approaches also do not contain any proper means of accounting the mechanisms for, and results of, long-term exposure to electronic smog and other combined agents and effects. Conventional scientists also have no knowledge of the major degradation mechanisms involved. The conventional medical models specifically are useless when confronted with something such as the Gulf War Syndrome or the decades of Russian microwave radiation of the U.S. Embassy personnel in Moscow. Capable scientists turn to their conventional models for answers, and the models fail them. Therefore present medical science is baffled when faced with such diseases and syndromes.[72] (*See figure 13 on next page*)

HOW TO MAKE A GRAVITATIONAL WAVE AND CURVE SPACETIME

We return now to the fundamentals of Type 2 engineering.

If one superposes an ordinary EM wave with its true phase conjugate replica wave, where the two waves are of equal strength, the *EM forcefield aspects* of the combined wavepair (standing wave) vanish, while the energy-density aspects remain. At every point the E-fields and the B-fields sum to vector zero *forcefield* resultants. The standing wave that results will not cause EM movement of charged particles.

However, the *energy densities* of the two locked (and now "hidden") EM waves remain and add. By standard GR, this "trapped EM density standing wave" must constitute a standing gravitational wave and a standing wave in the local curvature of spacetime!

It follows that, by assembling a harmonic set of such wavepairs, a la Stoney and Whittaker, one can form a region of spacetime such that the energy density is deterministically altered and controlled, and to engineering levels.

Engineering general relativity suddenly became a lot easier! Ironically, present physics only considers primarily the *very weak* ST curvature that can be obtained by the extremely weak gravitational "attraction of mass" G-force. For electrons, e.g., the G-force is only on the order of 10^{-42} times as strong as the electric force. Consequently orthodox GR simply "cannot pack enough trapped G-field energy into the local space" in the laboratory to get measurable effects. Only in the vicinity of a very large mass, such as the sun or a star or a black hole, does the weak G-force reach sufficient strength to yield directly observable, strong effects. Yet by using EM vacuum engines and the SWZ approach, one can obtain readily-measurable GR effects right on the laboratory bench.

The conventional medical models specifically are useless when confronted with something such as the Gulf War Syndrome or the decades of Russian microwave radiation of the U.S.

---

[68] *Note that the transcription component may continue to develop in the same "post disease" manner as does the shadow disease of the overall Gulf War Syndrome. The returnee continues to experience EM smog and other control/immune suppressants, though at a lesser level than experienced in the Gulf War environment.*

[69] *Such aspects are highly controversial. A few affected families are reporting such symptoms and aspects, but officially "there is no evidence" for them. However, with the initial appearance of the syndrome, officially veterans who complained of this illness were briefed by these same medical "authorities" that their physical symptoms were stress-related, and that the cause was all in their mind!*

[70] *In a fashion similar to Priore's animal experiments that demonstrated the transfer of curative templates by a single drop of blood. Kaznacheyev showed that disease templates could also be transferred.*

[71] *In fact, EM bioeffects research for decades has used, and continues to use, sheer energy deposition (scattering of EM energy by the tissues – heat by another name!) as a measure of the exposure level. They have not applied nonlinear phase conjugate self-targeting effects at all.*

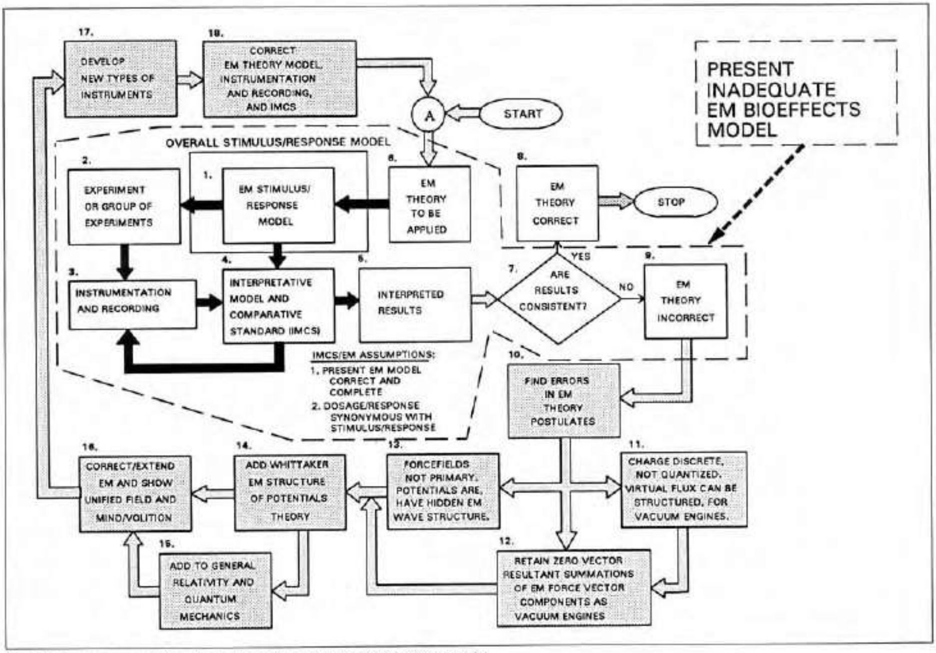
17.
DEVELOP NEW TYPES OF INSTRUMENTS
18.
CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS
A
START
PRESENT INADEQUATE EM BIOEFFECTS MODEL
OVERALL STIMULUS/RESPONSE MODEL
2.
EXPERIMENT OR GROUP OF EXPERIMENTS
1.
EM STIMULUS/ RESPONSE MODEL
6.
EM THEORY TO BE APPLIED
8.
EM THEORY CORRECT
STOP
3.
INSTRUMENTATION AND RECORDING
4.
INTERPRETATIVE MODEL AND COMPARATIVE STANDARD (IMCS)
5.
INTERPRETED RESULTS
7.
ARE RESULTS CONSISTENT?
YES
NO
9.
EM THEORY INCORRECT
IMCS/EM ASSUMPTIONS:
1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE
10.
FIND ERRORS IN EM THEORY POSTULATES
16.
CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION
14.
ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY
13.
FORCEFIELDS NOT PRIMARY. POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.
11.
CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.
15.
ADD TO GENERAL RELATIVITY AND QUANTUM MECHANICS
12.
RETAIN ZERO VECTOR RESULTANT SUMMATIONS OF EM FORCE VECTOR COMPONENTS AS VACUUM ENGINES

Fig. 13. Necessary corrections for the present inadequate EM bioeffects model.

POTENTIALS AND GRAVITY

Obviously one can simply add harmonics of such EM "locked in" wavepairs, to produce a standing "constant value potential" pattern of a fixed *excess* energy density of vacuum or a fixed *decrease* in the energy density of vacuum. This harmonic series of bidirectional EM wave pairs – the *sum set* of wavepairs – is precisely what Stoney and Whittaker showed makes an ordinary EM scalar potential. In other words, they showed the actual energy-trapping mechanism, and dramatically altered the view of a scalar potential as being simply an "electrostatic" entity and a scalar field. Instead, it is a *hidden-variable energy flow process.*[73,74,75]

Its external "scalar field" aspect is just its "local energy density" aspect, not its fundamental nature.[76]

As is well-known in GR, all potentials are gravitational because they contain trapped energy. *Any* trapped energy is gravitational, a priori, according to GR. Ziolkowski independently rediscovered the SW decomposition in the 1980s and extended it to also add the product set. Since multiplication of waves is *modulation*, in my opinion Ziolkowski actually wrote the basis for hyperspatial (or subspatial, if one models physics in only 4-dimensions) communication at superluminal velocity, essentially instantaneously anywhere in the universe. (*See figure 14 on opposite page*)

"Paired assembly of an EM wave and its true PCR" was and is the way to turn EM energy directly into gravitational en-

---

[72] *Just as it is also essentially baffled by many conventional dread diseases such as cancer, leukemia, AIDS, arteriosclerosis, etc. – which indicates that the conventional "models" for those diseases are also seriously deficient.*

[73] *The present preoccupation of physicists with the scalar potential is simply to provide an algorithm for calculating the local energy density – in excess joules per collecting point-coulomb of charge – and erroneously call that the definition of the potential. The magnitude of the energy collection density of the potential at a point is not what the potential is, and such is not its definition. Defining how much water can be poured out of a standard bucket, when it is dipped into a flowing river and filled with water, does not define the river!*

[74] *From the particle view, the potential is an ordering in a virtual particle flux of the vacuum. From the wave view, the potential is comprised of an ordered, harmonic series of SWZ bidirectional EM waves, where each wave reaches to the ends of the universe before having a zero energy density. It follows that the energy in each hidden sine wave is infinite, just as it is for any overt theoretical EM sine wave. So any potential – even that of a flashlight battery – already involves an infinite amount of flowing hidden EM energy, being exchanged by the source and the entire universe, flowing in both directions along any radial from the source of the potential. The scalar potential is a flowing, continual exchange of energy between each and every point in the universe and the local source of the potential.*

[75] *Both the wave and particle forms of the potential exist simultaneously. The particle nature exists in Minkowski 4-space. The "hidden SWZ waves" exist in a hyper-4-space whose time dimension is the same as that of Minkowski 4-space. The scalar EM potential actually exists as a 7-dimensional entity. The internal hidden EM flow is not confined to the speed of light in space, because it does not flow "through" 3-space; instead, it flows "around" it.*

[76] *E.g., Whittaker, 1904, ibid. shows us that all of classical electromagnetics can be made by the interference of two scalar fields, which is just a special kind of hyperspatial multiwave interferometry.*

ergy.[77] This is the unification of EM and GR that eluded Einstein all the latter part of his life, and that has continued to elude scientists today. This is the way to "trap" EM energy locally, even though the EM wave energy is flowing at the speed of light or greater. This is the way to change EM energy into gravitational energy.[78] Further, it is engineerable and testable. Remember, physics requires that energy be conserved; however, it can certainly *change its form*. GR already says that it is the trapped energy that is gravitational, not mass per se, and regardless of what kind of energy is trapped. Ultimately Einstein's formula $E = mc^2$ is actually a statement of the trapped EM energy in a mass. However, heretofore no one had the slightest notion as to the actual trapping mechanism. The SWZ work actually shows the trapping mechanism itself and allows the engineering of *electrogravitation*.

HOW DID THE REVOLUTIONARY STONEY/WHITTAKER WORK GO UNNOTICED?

Apparently, neither Stoney nor Whittaker realized that their work was pure electrogravitation. *When the referenced papers by Stoney and Whittaker were published, special relativity, general relativity, and quantum mechanics had not even been born yet!* Indeed, there still was not a wide adoption of Heaviside's vector interpretation of Maxwell's fading quaternion theory. Even classical EM was still in its infancy and just getting off to a good start.

In the 1980s Ziolkowski realized that the internal biwaves of a potential could be engineered, but remained primarily focused upon his important work with electromagnetic missiles and focus waves (slugs of EM energy which stay together as they travel, and do not spread out.) Ziolkowski is one of the pioneers in that field.

HOW PRIORE MADE VACUUM ENGINES

This "vacuum engine methodology" was and is what Priore's intuition led him to! For the electromagnetic treatment of laboratory animals, his EM therapy device first inputted some 17 or so frequencies into a rotating plasma.[79,80] We know today that such plasmas *phase conjugate* the wave inputs, but that was not known when Priore first did it.

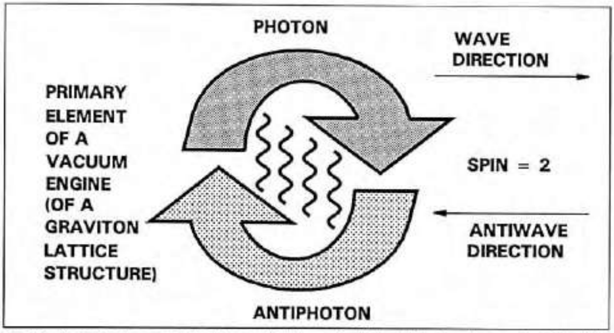
PHOTON
WAVE DIRECTION
PRIMARY ELEMENT OF A VACUUM ENGINE (OF A GRAVITON LATTICE STRUCTURE)
SPIN = 2
ANTIWAVE DIRECTION
ANTIPHOTON

Fig. 14. A graviton is a coupled photon/antiphoton pair. In the bidirectional wavepair of the SWZ internal structure of the scalar potential, photons in the wave continually couple and uncouple as the wave and antiwave pass through each other.

Obviously Priore himself did not understand the actual mechanism, but his brilliant intuition led him to discover how to do it! For many years no one knew how aspirin worked, yet that did not prevent it from being used widely and effectively.

In other words, Priore used the plasma to *unwittingly* add PCRs to each of his waves and form 17 magic SWZ "wavepairs." The entire SWZ wavepair assembly thus made an *artificial potential*, having a "vacuum engine" pumping substructure of hidden bidirectional EM waves. Adjustment of selected frequencies away from their harmonic positions adds multiplications and beat frequency modulations, creating *specific* templates and *specific* vacuum engines. These SWZ waves are hyperspatial; hence they are in the domain of the deep mind's cellular control system. When an abnormal body cell is pumped by these hyperspatial waves, the cell's own hyperspatial vacuum-engine "signal input complex" triggers a specific dedifferentiation anti-engine for that specific abnormal state. By Wheeler's principle, if spacetime curvature is *present in* the material cell, its energy and template are *applied to* the cell and act upon it. With vigorous pumping, the anti-engine is also *amplified* and so the cellular response is more rapid. In other words, the amplified anti-engine fairly quickly overpowers the disease engine and reverses the abnormal cell and all its contents back to their previous *normal* states. (*See figure 15 on the following page*)

Priore's new signal complex came out of the plasma mixing (phase conjugating) tube with SWZ structures of wavepairs, where each pair was composed of two bidirectional (opposing) EM waves. Priore then modulated this output upon a very strong, rippling magnetic field as a carrier, and radiated the body of the patient with that resultant field and its transported hyperspatial SWZ anti-engine components.

Nuclear magnetic resonance guarantees that such a magnetic field with all its modulations will be delivered entirely through the body, penetrating the flesh and bones and even the atomic nuclei themselves. It will certainly penetrate the bone marrow where all the blood cells (which includes most of the immune system's cells) are born! It will penetrate everywhere else also. So all the cells of the immune system will be exposed to the AC magnetic carrier and its impressed hyperspatial biwaves and their modulations.

The body's cells and cell-parts are highly nonlinear, and so they are also *demodulators*. They strip off the SWZ modulations. This puts the amplified vacuum anti-engines directly into and onto the cells, including all the cells of the immune system, and including each and every part of each cell. Small wonder then that the Priore methodology could restore suppressed immune systems! Time-reversing the damaged immune cells and their damaged internal components back to a previous healthy state was one magnificent result. The radiation also delivers the modulations directly into the master potential (the body's quantum potential) that contains its mind, its long-term memories, and its master control system for all cells. Thus not only are the diseased cells and the immune cells affected, but also the *master cellular control system* for all body cells is directly affected.[81]

---

[77] *I have previously pointed out that, in a coupled wave/antiwave pair, the photons in the wave and the antiphotons in the superposed phase conjugate replicate wave also pair into photon/antiphoton pairs, which have spin-2 helicity and hence are gravitons. This follows merely from a strong application of the distortion correction theorem. However, it is the fundamental union of EM and G into a single engineerable theory. See Bearden,* ***Gravitobiology,*** *1991, p. 1-9.*

[78] *By Whittaker 1904, we can always interfere two such potential fields and obtain an EM field, even at a distance. Thus we can also readily turn the gravitational (trapped EM) energy back to EM field (untrapped EM) energy.*

[79] *David M. Rorvik, "Do the French Have a Cure for Cancer?"* ***Esquire,*** *July 1975, p. 110-11, 142-149 gives an overview of the Priore Affair and some details of the electromagnetics of the machine.*

[80] *Note the immediate connection to the Russian energetics model, which is directly fashioned after the nature and nonlinear EM functioning of plasmas. Priore was directly applying Soviet energetics, although neither he nor any other Western scientist could have known that.*

In a related work, Becker (*See figure 16 on the following page*) and his colleagues stimulated bone healing in difficult fractures by use of tiny DC currents. They have shown that picoamperes of steady DC current trickled across the fracture site induces bone healing. Becker showed that, surprisingly, red blood cells are dedifferentiated – shucking their hemoglobin and growing a nucleus in the process. These cells redifferentiate into the kind of cells that make cartilage. Then *those* cells redifferentiate anew, into the kind of cells that make bone, and heal the fracture with new bone deposit. Becker also proposed a complete, operational DC control system, staying within the limitations of conventional EM theory.[82,83,84]

15-A. Priore demonstrates a point to Mr. Chaban-Delmas, Mayor of Bordeaux and a friend and mentor.

15-B. Dr. Raymond Pautrizel, noted immunologist, worked directly with Priore. Dr. Pautrizel is Professor Emeritus, University of Bordeaux, and holder of the first chair in immunology in France.

Dr. Robert Courrier (not shown), head of the Biology Section of the French Academy of Science as well as its Secretaire Perpetuel at the time, was also a noted colleague. He personally presented the astounding results of the Priore Team's experiments to the Academy.

Fig. 15. Priore was supported by notable French political leaders and eminent French scientists were his close scientific colleagues.

A proposed explanation of the fundamental vacuum-engine mechanism for the Priore and Becker effects has previously been given in some detail by Bearden.[85] (*See figure 17 on next page*)

---

[81] *We refer to an expanded version of Popp's cellular control system. E.g., see Fritz Albert Popp, "Photon Storage in Biological Systems," in Fritz Albert Popp et al, eds.,* ***Electromagnetic Bio-Information: Proceedings of the Symposium, Marburg, Federal Republic of Germany, Sep. 5, 1977;*** *Urban & Schwarzenberg, Baltimore, 1979, p. 123-149. Popp reports the theoretical and experimental discovery that mitogenetic radiation from cells can be seen as a sort of "waste" from a virtual electromagnetic field with a high coherence. This field has a tendency to become stationary over the whole organism. In addition, it includes the storage of "virtual" coherent photons.* ***Comment:*** *Our substitution of the SWZ hidden wave approach to replace Popp's QM particle approach greatly facilitates understanding and specifying the control system, and materially eases direct engineering of it.*

[82] *E.g., see R.O. Becker, "The direct current field: A primitive control and communication system related to growth processes."* ***Proceedings of the. XVI International. Congress of Zoology****, Washington, D.C., Vol. 3, 1963, p. 179; "The control system governing bone growth in response to mechanical stress."* ***J. Ark. Med. Soc.****, Vol. 62, 1966, p. 404; R.O. Becker and D.G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents."* ***Trans. N.Y. Acad. Sci.****, Vol. 29, 1967, p. 606. See particularly Robert O. Becker, "The significance of bioelectric potentials."* ***Bioelectrochemistry and Bioenergetics,*** *Vol. 1, 1974, p. 187-199. Becker proposes that a complete, operational system exists in living organisms, and controls such basic functions as growth, healing, and biological cycles. The system is described in anatomical detail. It functions as a data transmission system in an analog fashion using varying levels of direct current as its signal. The system interlocks physically with the nervous system and is postulated to be its precursor. There are two electrochemical limbs of great significance in the operation of the system. One is between the DC system and the nervous system; the other is between the DC system and all body cells. The concept explains all of the diverse effects reported for the biological effects of applied electrical currents including: electrical anesthesia, electrical growth control, and electroacupuncture. It also furnishes a testable hypothesis for predicting other effects of clinical significance.*

[83] *We comment that the SWZ approach allows control and conditioning – both externally and internally – of both limbs of Becker's proposed DC control system.*

[84] *See also C.W. Hsue, "A DC Voltage is Equivalent to Two Traveling Waves on a Lossless, Nonuniform Transmission Line,"* ***IEEE Microwave and Guided Wave Letters****, Vol. 3, 1993, p. 82-84. Hsue shows that the steady DC voltage on a transmission line powering a resistive load, mathematically decomposes into bidirectional EM traveling waves. The bone-healing setup used by Becker et al is precisely of this configuration. The fracture corresponds to the resistive load and the two electrodes on each side correspond to the transmission lines. Hence Becker's team actually employed a form of the Stoney/Whittaker/Ziolkowski harmonic biwave decomposition of the scalar potential, to accomplish hyperspatial pumping and formation of anti-engines of vacuum energy. So this is another application of the same basic "cell-reversing" methodology, as also utilized more extensively by Priore. Becker's brilliant work of course was and is independent and unique, and he was nominated for a Nobel Prize because of his absolutely fundamental work in showing electromagnetic differentiation and dedifferentiation of cells. That he has not been awarded the prize is regrettable and, in my opinion, an injustice.*

[85] *T.E. Bearden, "Mechanism for Long-Term Cumulative Biological Effects of EM Radiation," presented to the 70th Annual Meeting of the Alabama Academy of Science, University of Alabama at Huntsville, March 25, 1993. A digest of this presentation is published by the Tesla Book Company, P.O. Box 121873, Chula Vista, CA 91912 as T.E. Bearden,* ***Cancer and the Unresolved Health Issues in the Biological Effects of EM Fields and Radiation,*** *1993 and includes both text and hard copies of viewgraphs.*

HOW THE PRIORE VACUUM ENGINE REVERSED CELLS TO A PREVIOUS STATE

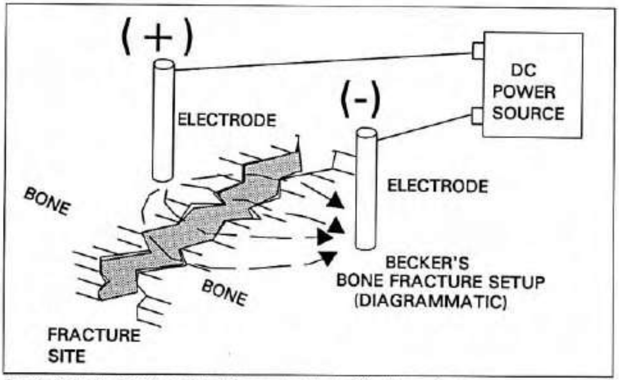
(+)
(-)
ELECTRODE
DC POWER SOURCE
ELECTRODE
BONE
BONE
BECKER'S BONE FRACTURE SETUP (DIAGRAMMATIC)
FRACTURE SITE

Fig. 16. Becker's application of tiny DC currents to heal difficult bone fractures.

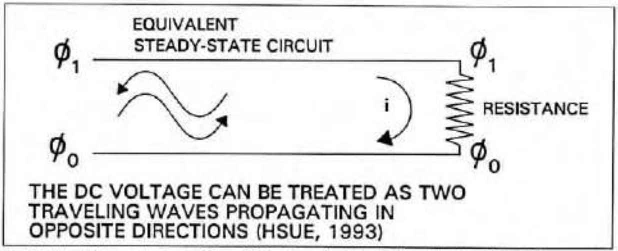
EQUIVALENT STEADY-STATE CIRCUIT
φ1
φ1
i
RESISTANCE
φ0
φ0
THE DC VOLTAGE CAN BE TREATED AS TWO TRAVELING WAVES PROPAGATING IN OPPOSITE DIRECTIONS (HSUE, 1993)

Fig. 17. Hsue's biwave decomposition of DC voltage (as in Becker's application of tiny DC currents to heal difficult bone fractures).

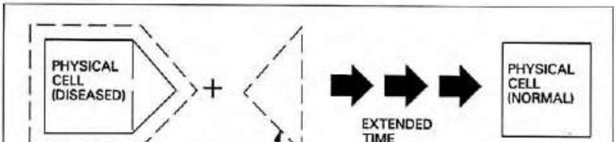
PHYSICAL CELL (DISEASED)
+
PHYSICAL CELL (NORMAL)
EXTENDED TIME DELAY FOR RENORMALIZATION
VACUUM ENGINE FOR DISEASED CELL
VACUUM ANTI-ENGINE FOR CELL'S DISEASE DELTA

Fig. 18. Applying an anti-engine for the "deviation" of the diseased cell reverses the cell to normal after a time delay.

Taking a GR view, any mass already has its own vacuum engine for its exact form and internal structure and dynamics, in the finest detail. That particularly follows Wheeler's principle; if the mass is there, the appropriate spacetime curvature and dynamic energy structuring is there. If the mass has internal structuring, so does the spacetime curvature: it becomes *nested* curvatures with a precisely correlated template. In short, any mass – living or not – is imbedded in a precise vacuum engine. Any and every change adds a specific delta to this vacuum engine component. Just as a tree has rings for its past years and an onion has layers, the vacuum engine of an object has an exact template for that object's entire past, *as do the mind and its components also.*[86]

So when a normal living cellular mass changes to a diseased state, *a priori* it has added an additional delta vacuum engine (representing the change that resulted in the disease) to its original vacuum engine (which represents the original healthy state of the cell). The cell and all its parts are highly nonlinear and act as phase conjugate mirrors, *including to the higher topology (hyperspatiality) of the vacuum engine. Any living system we isolate and study is at least 7-dimensional.*

Putting all this together, we now can understand the marvelous nature of what Priore did. By exposing the body, treated part, cells, and cell parts to Priore's modulated rippling magnetic field, the "hidden pump waves" carried by that magnetic field are applied to each and every cell and to all its parts, in the exposed area. *In the exposed area, each and every cell and its parts become powerfully pumped phase conjugate mirrors for its hyperspatial vacuum engine complex.*

Extending the language of nonlinear optics, the particular *net* vacuum engine of the diseased cell or part constitutes a "standing input" *hyperspatial signal wave complex*, already input to the hyperspatially pumped phase conjugate mirror cell or cell part. When pumped by "hidden" pump waves, each and every one of those PC mirrors will now originate an *amplified phase conjugate replica of the signal wave complex. Each one will produce a highly amplified, time reversed replica of its specific standing vacuum engine. This includes amplified time reversed replicas of both the "healthy" component of each standing vacuum engine and the "disease delta" component. (See figures 18 above and 19 on following page)*

In other words, every pumped diseased cell and cell part – including the genes and chromosomes – will produce a very powerful, highly amplified, "reverse vacuum engine" (an *anti-engine*) which is fantastically specific and precisely time-reverses the cell and its parts, including the genes and chromosomes as well, back to their *previous normal states.*

---

[86] *If one alters the internal structure (rings) of a tree, one changes the tree. If one alters the internal structures and dynamic functioning of matter, one changes the matter accordingly. If one alters the internal structures and dynamics of the mind, one changes the mind, its contents, and its functioning accordingly, as well as the subsequent behavior exhibited by the altered individual.*

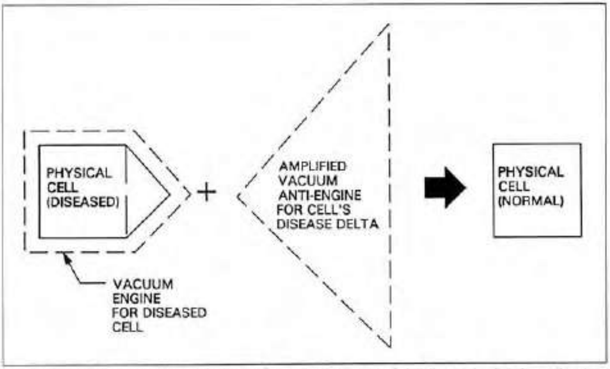

Fig. 19. Applying an amplified anti-engine for the "deviation" of the diseased cell reduces the time delay required to reverse it back to normal.

REJUVENATION

Any *normal* cell that is exposed to hyperspatial (Priore-type) pumping just reverts to a *slightly younger* normal state. Note that this allows *rejuvenation*.[87] Priore's earliest experiments that so riveted his attention and set him on the path of discovery, were in fact the abnormal prolongation of the life of fruits! For each class of disease, Priore just had to fiddle with the frequencies until he had the right set of pump waves for the specific "disease delta vacuum engine." This he did by trial and error. So when it is "Priore-pumped," a cancer cell reverses back to a normal cell, as does a precancerous cell as well.

The revolutionary Priore therapy is startling but simple. The diseased cell is analogous to an automobile that has been driven slowly off the road into the rough. Priore's methodology shifts the car into reverse gear, guns the engine, and quickly and precisely backs the car back onto the roadway. It is one of those profoundly beautiful embodiments analogous to Faraday's unipolar generator; it strikes directly to the heart of the fundamental principle. A cell develops damage and disease over a period of time. Simply time-reverse it quickly back to its original healthy state.

AIDS COULD BE QUICKLY CURED THE SAME WAY

Given Priore-pumping, an AIDS-infected cell would revert back to a normal cell, including all its genetics, and including for all of the hundreds of "strains" of AIDS that exist in an advanced AIDS patient. With this ap- proach one could completely reverse and eliminate AIDS whenever a positive reading occurs in the HIV virus test, and prior to the appearance of the actual physical sym- ptoms! If the person just has the antibodies and not the disease, the treatment will still do him absolutely no harm. (It will actually make his cells a "little bit younger and more vigorous," and help his immune system.) If he has both the AIDS antibodies and the infection as well, he will be completely cured and his immune system will be restored to complete vitality.

MOST DISEASES COULD BE CURED BY THE NEW THERAPEUTIC MODALITY

Any infectious disease known or unknown can be cured by the same means. The marvelous thing is that one need not know the specific technical pattern for the delta vacuum engine for a specific disease! One just has to *correctly adjust the hidden SWZ pumping frequencies.*

ARTERIOSCLEROSIS

In arteriosclerosis, it has been found that the greatest part of the damage is due to the presence of some rather ordinary viruses, and abnormal cell growth by the body's own cells as a result of triggering by previous viral infection. Cellular reversal (dedifferentiation) gets it all; the viruses and the abnormal cells. Everything reverts to normal, and the body's restored processes quickly "police up" the extra cell growth, using its natural control processes which are now restored to full functioning. *This is how the Priore team was able to so dramatically cure the "clogged arteries" of laboratory animals.*

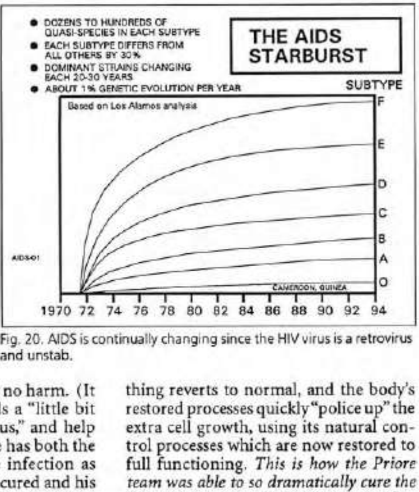

Fig. 20. AIDS is continually changing since the HIV virus is a retrovirus and unstab.

The Priore approach (pumping of the

---

[87] *One should not be surprised that a time-reversal operation upon an entire body leads to reversal of physical aging in that body! But we point out that the time reversal we speak of is not at all the "time travel to the past" so beloved by generations of science fiction writers. For that, one would have to reverse each and every part of the universe simultaneously, except the observer, back to an exact former state, in perfect universal coordination. One would have to pump the entire universe, except for the observer! The observer would then remain as is with no change, while his entire world "went back in time" to a past state. The pumping of the universe would have to stop instantaneously (decay of the false vacuum?), and the universal anti-engine would have to stop operation and decay instantaneously (another big bang, perhaps?). We are not referring to anything such as that! We are referring to the reversal of a single little part of the universe, back to its previous state, which is practical and can be done. Universal time reversal — and hence time travel for the observer — is not practical, at least within our present knowledge and within the time reversal model we are applying. We should point out, however, that time reversal is a legitimate part of physics, and is not science fiction. E.g., for an overview, see Robert G. Sachs,* ***The Physics of Time Reversal,*** *University of Chicago Press, Chicago, Illinois, 1987.*

diseased cell's self-contained vacuum engine to create a very powerful vacuum anti-engine and reverse diseased cells back to normal) is the most powerful therapy for cellular disease ever conceived of. If the cells have ever been normal, they can be reverted to normal.[88] One can also probably regrow lost limbs, etc. but that takes too long to explain here! The methodology should also be able to completely reverse disorders such as Alzheimer's disease, and very probably muscular dystrophy as well. It should even be possible to reverse the physical effects of aging!

MUCH MORE THAN THIS INTRODUCTION HAS BEEN THEORIZED

This is the gist of the mechanism for the Priore therapy's startling success, although it has required some 10 years of very hard work to decipher the mechanics of the process. Along the way, I have also personally discovered and added a great deal more (for example, the quantum potential aspect of a living system and any enclave of living systems, the mechanism to generate a quantum potential, the precise mechanism for Sheldrake's morphogenetic field, the mechanism for directed evolution, the unusual electromagnetic nature of mind and thought, the mechanism for engineering mind and thought, engineering the entire species, and engineering the entire living biosphere), but this much should suffice for now for introductory purposes.

A technical book has been laboriously begun, and hopefully will be finished and published in about a year. Two 2-hour videos are also tentatively being planned. A four-hour technical briefing has already been prepared and presented.

REACTIONS TO THIS VACUUM ENGINE EXPLANATION

In response to numerous queries as to the reactions to this explanation of the Priore mechanism: Reactions tend to be rather good when (1) the technical person is open-minded, in other words, a friendly skeptic, (2) the technical person endures a grueling 2-hour or 4-hour detailed slide presentation I prepared, and (3) sufficient follow-up dialogue is accomplished to answer residual questions. Otherwise, the "mental jump" required of the participant is just too great.[89] Also, perhaps 60-70% of the conventional scientists are absolutely against anything that smacks of "revolutionary." From about half of this contingent the approach may generate disbelief and even overt emotional hostility, just as Priore himself and his associates encountered.[90]

However, some of the sharpest young graduate students readily make the transition, if they interact with me for a long enough period and enough times to understand the physics of the approach. *It did not take me 10 years to decipher an easy thing!* The causative mechanism for Priore's accomplishment is terribly involved and very, very deep. It is slightly beyond the present scientific vanguard, although the bits and pieces necessary to decipher the Priore mechanism are now in the literature. But putting them together has been a bone-crushing, terrible task. Priore was truly a century ahead of the science of his time!

Nonetheless, there is much hope, because the material is slowly taking root in a few sharp young graduate minds. For example, one of the major Eastern Universities would be astonished to know that their sharpest, brightest young PhD candidate – one they already assume will win a Nobel Prize in the future – is a full convert, and thoroughly understands cellular reversal by hidden-variable SWZ pumping, a la Priore. But at my strong urging, he is keeping it a secret until he graduates, and then he will *gradually* begin to filter in some of these ideas. That way he does not initiate his destruction by his peers because of too much energetic advocacy too quickly.

THE PRIORE APPROACH IS URGENTLY NEEDED

We simply must get this Priore therapeutic approach developed, for it is extremely important to all humanity. Our very survival may depend upon it.

Presently all our antibiotics are beginning to fail; the disease organisms are becoming quite resistant. Already our hospitals have strains of diseases that few if any antibiotics will touch. Hundreds of thousands of hospital patients each year are already experiencing complications due to contracting an antibiotic-resistant infection. Probably about 20,000 a year are dying as a result, and the death rate is increasing. In the future, one will have to seriously weigh whether it is better for one's chance of survival to enter the hospital and be treated, or to take one's chances with the disease and not enter the hospital.

With the rising failure of the antibiotics, (*See figure 21 on page 60*) and the rising possession of tailored biological weapons by hostile nations and international terrorists,[91] (*See figure 22 on page 60*) we shall shortly have no defense against an onslaught of pandemics unheard of since the Black Death swept medieval Europe of one third the population. Literally the age of pestilence and dread diseases of Revelations is upon us.[92]

---

[88] *They can also be "reversed to normal" even if they have never before been normal. The master cellular control system contains a master blueprint for the normal cell and can make an appropriate template for the specific disease delta vacuum engine, or one can be introduced (e.g., by injection of a drop of blood from a person previously cured of that specific disease; this was in fact demonstrated by the Priore group.). Pumping will then reverse the patient back to normal, as was actually shown by the Priore group. Indeed, the body's own system may be able to use the correct anti-engine template from the drop of blood and cure itself without pumping, as was also shown by the Priore group, if the body itself is reasonably healthy.*

[89] *Simply "dropping it on someone" by mail, for example – as I did to Elizabeth Taylor's foundation and Ross Perot's office – is just too much for a casual reading to grasp. A polite acknowledgment was received from the Taylor foundation, and no acknowledgment at all was received from the Perot camp.*

[90] *As an example, Priore completed the course work for his PhD and submitted his thesis, documenting his formidable and revolutionary work. Because of the ferocity of the oncology establishment toward the Priore team and its accomplishments, the university rejected his thesis. During the two decades or so since then, untold millions have died worldwide from diseases the team cured. These deaths could have been prevented had the Priore work been vigorously pursued and developed by the scientific community.*

[91] *It is not a question of if the international terrorists will use biological weapons against the U.S. civilian populace, but only of when. Astoundingly, often it may even be possible for the terrorists to simply buy their biological warfare agents from us! In a lawsuit by Desert Storm Syndrome victims against certain U.S. companies, one of the companies is alleged to have sold germs – such as anthrax, tularemia, and botulism – to Iraq by mail. These are highly effective biological warfare agents. E.g., a few spray tanks spraying such agents over a large U.S. city complex at night, using an easily obtained light plane to get in a few hours "flying practice," would result in tens or even hundreds of thousands of deaths within days. According to U.S. government-released data, inhaling a few anthrax spores into the lungs has a very high probability of killing the inhaler, unless treatment is started immediately. Even with immediate treatment, many deaths will still occur. A substantial number of nations are now obtaining and producing lethal biological warfare agents – which have been called the "poor man's atomic bomb." Some of the international terrorist groups are also acquiring the biological warfare agents.*

[92] *A good exposition of the rapidly increasing resistant-disease problem is given by Garrett, Laurie,* ***The Coming Plague: Newly Emerging Diseases in World out of Balance,*** *Farrar, Straus and Giroux, New York, 1994.*

In addition, the costs of the present allopathic medicine are rising rapidly and threatening to either bankrupt the nation, or become so expensive that only the wealthy can afford adequate medical care. Even the wealthy could be trapped in the future and denied adequate medical treatment. The Clinton Medicare bill, if approved, would make it a felony to pay a doctor out of one's own pocket, after denial of such treatment by the administrative bureaucracy. It also makes it a felony for the doctor to give the treatment to the disapproved paying person. Thus both the patient and the doctor could be sent to prison, wealthy or not.[93]

CELL CONJUGATION BY VACUUM ANTI-ENGINES IS THE ONLY HOPE

Priore's approach offers an answer to both the age of pestilence and the impending biological warfare strikes against our armed forces and our civilian population.[94,95] By prompt cellular reversal, infection even by totally unknown and modified agents can be quickly voided. Infection by newly adapted, resistant strains of presently known diseases can be quickly conjugated. So the problem of the failing vaccines and antibiotics can be solved quickly, simply, and cheaply. However, the cell-reversing treatment devices and trained, experienced medical staffs must be widely available!

Diseased cell conjugation by EM means is also the only therapy that can largely *eliminate* the debilitating diseases of humanity. Per Wheeler's principle, vacuum engines can mold the *physical matter* of the living body and conditions, operations, and content of the mind, memory, and personality. As the Priore therapy is developed and extended, it will become possible to correct bent spines, mental retardation, mental diseases, genetic defects, physical accidents and disabilities, birth defects, etc. It will be possible to regrow amputated limbs and restore deteriorated cartilage, organs, and joints. It should also be possible to directly remove deep psychiatric problems, heal chemical imbalances, and address and heal such presently difficult disorders as schizophrenia.

The development and application of Type 2 medical engineering can usher in the greatest, most beneficial revolution of all time. As just *one* example, eventually it will be possible to quickly educate students by simply loading the appropriate "software" into their minds and long term memory, in only a few weeks of "sessions." At that point there need no longer be, anywhere in the world, an educationally disadvantaged person.

It can also be used to usher in the most diabolical revolution of all time if we are not careful to (1) be informed of the Type 2 engineering area, including who is developing what, and for exactly what purposes, and (2) rigorously control the development and use of this technology as it progresses. Our legal, political, and scientific leaders – and we common citizens – will have to grapple with the ethics of what is done and to and for whom it is done. Just as one example, the same technology that will allow direct loading of the software in the long-term memory to achieve rapid education, will allow altering or changing any part of the long term memory, including the built-in emotional responses, etc. *Loading software is loading software! What is loaded depends only upon what is in the software.* In theory it is possible not only to alter the learned behavior, memories, emotions, and reactions, but also the genetically "born-in" behavior, emotions, and reactions. The potential for the Type 2 engineering to be used for enslaving and roboticizing humanity should be obvious.

Far more carefully than with any other technology in history, we must ensure that this new one is developed for good, not for evil. To quote an old adage, *the price of freedom is eternal vigilance.* Never before in all of history have those words been more true.

IN CONCLUSION

Pioneers such as Priore, Becker, Kaznacheyev, and Popp have shown us a

---

[93] *In fact, under present law all property, homes, bank accounts, and monies of both the patient and the doctor would simply be confiscated – without any legal charge, arrest, or trial – under an absolutely medieval reinterpretation of U.S. law that has again given police at all levels such sweeping and unconstitutional powers. The Clinton Health Care Plan, presently on hold but still alive and sure to be introduced again, is reported to include the establishment of an entire department solely devoted to confiscating assets from patients and doctors who commit such "crimes." Defendants victimized by the present confiscation reinterpretation of law have a slim chance of even forcing the issue into a court. Even if they succeed in getting into court for a hearing, win it and get full "exoneration," the authorities can still keep the property! In the most insane and unjust legal action in all of U.S. history, this law now charges the property itself with having committed the crime, not the erstwhile human defendant. By what means can the "property" defend itself? No one has an answer, for the property is not an individual and has no legal rights for jury protection under the law! U.S. lawyers and judges are actually insisting that this is rational and makes sense. They have already precisely redefined U.S. law into what is specifically required to institute a police state. The police forces who carry out this consfication are themselves awarded a percentage of the "haul." What does the citizen do when the police forces themselves become the robbers by administrative fiat, and there is no court redress possible? Ostensibly passed to counter organized crime and the drug cartel, this law has rapidly been applied directly to the citizenry, and its use is one of the fastest growing "new techniques" by the various levels of law enforcement. In this paper we have deliberately not dwelled upon the political aspect of the already-emerging police state in control of medical treatment. We simply summarize how the bureacratic medical establishment is likely to meet the challenges of the coming plagues and the associate astounding rise in medical costs. In fact the reader will almost certainly see the rise of an iron "state dictatorship" in medicine, as is already emerging in other legal areas.*

In this paper my primary purpose is technical, not political, so we defer detailing the more involved political aspects of the emerging police state (the New World Order – more aptly the New World Feudalism) to other more capable authors and sources. One central thought should remain firmly in mind All political systems and isms tend ultimately to feudalism, which is always the best system for the oligarchy or elite at the top who control and own everyhing. In any society, there is always an on-going struggle by a fraction of the most powerful to convert the society to special forms of feudalism.

[94] *Some biological agents are very lethal and the associated disease develops at a frightening pace. E.g., the strains of Staphylococcus bacteria that cause necrotizing fascitis and necrotizing myositis – two versions of the "flesh-destroying" disease – are noteworthy examples. Necrotizing is causing inflammation and/or death of tissue, fascitis refers to "sheet" or connective tissue, and myositis refers to muscle tissue. Recently the Quebec separatist leader Lucien Bouchard was suddenly felled by necrotizing myositis, lost a leg, and at this writing is stable and will recover. The disease is usually fatal in 80% of the cases. There is nothing diabolical in his contacting the disease; the bacteria are present from time to time and can enter the body through a nick or otherwise. The disease and the bacteria are known; however, in the past such cases have been relatively rare. The lethality of the disease is augmented by the rapidity of its development. Delayed recognition of the disease may place the patient already beyond the ability of conventional agents to arrest it in time to save him. With the modern hospital and population environment being highly favorable for the more rapid modification and adaptation of disease organisms, even previously rare diseases such as necrotizing fascitis/myositis are ticking time bombs waiting to become resistant to our antibiotics and explode in the dense urban populations.*

[95] *We refer to strikes using biological organisms. We have mentioned but deliberately not dwelled upon the use of electromagnetic induction of lethal cellular diseases at a distance. Such strategic capability to attack a nation and destroy its citizens and soldiers is presently on-line in Russia, under total control by the KGB, not by the conventional Russian military forces.*

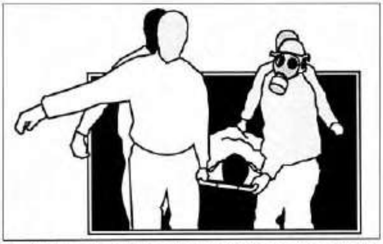

Fig. 21. Present antibiotic therapy is failing as disease organisms adapt.

Fig. 22. Terrorists will eventually use biological warfare strikes. Biological warfare is an increasing threat to military forces and the civil populace.

great vision. Having seen it, when one looks at the incredible human suffering throughout the world, and then at the formidable problems of our Type 1 medical science that is struggling so mightily and losing the battle, one is reminded of a phrase by Dante:

"Heaven wheels above you
Displaying to you her eternal glories
And still your eyes are on the ground."

With every fiber of our being, we exhort the scientific community to lift its eyes from the ground, and develop and apply the Type 2 medical technology as rapidly as humanly possible, for human benefit. Else half of humanity will perish in the coming tide of pestilence. Even the conventional scientists agree that the pestilences are coming, and that conventional approaches are already losing the battle. Long ago Albert Einstein wrote these prophetic words:

"*It would of course be a great step forward if we succeeded in combining the gravitational field and the electromagnetic field into a single structure. Only so could the era in theoretical physics inaugurated by Faraday and Clerk Maxwell be brought to a satisfactory close.*"

Einstein's passionately-sought unification of EM and G has now been accomplished. It has magnificent medical science applications for the benefit of all humankind. Meanwhile, in the orthodox medical journals, the scientists themselves are recognizing the demise of antibiotics and the coming unstoppable plagues. The scientists are redoubling their efforts to apply the present inadequate models, and their efforts are admittedly failing. *If they keep to their present course, the war is already lost!*

In the name of all the bleeding, suffering, and dying peoples of the world, we must do better than that! We can do better than that, if we rapidly redevelop and apply the Priore approach.

We call fervently for a great new "War on Disease" that is empowered, funded, and staffed to perform the task. It can be done in perhaps three years, and perhaps for as little as $60-100 million if we streamline procedures, keep the bureaucracy off the team's necks, and give the team scientists the authority of a War Commander with a mission order to *get on with it.*

What happens next is up to the scientists, the politicians, and the concerned leaders of all the nations of the world. *They simply must not fail us.* They and their families are in the same frying pan as the rest of us. This is not the time for dogma, posturing, arrogance, protecting one's territory, yapping about who's going to control the presently inadequate medicine, yakking about pseudoscience, striving for personal position and power, bowing to the big drug companies and their vested interests, and playing the "Who's to be the big monkey?" game. *We need a new Manhattan Project. We need it now.*

Else we who have
sown the wind
will most certainly reap
the coming whirlwind.

ADDITIONAL REFERENCES

Priore, Antoine. (1963) "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes." [Procedure and Assemblage for Production of Radiation Especially Service able for the Treatment of Living Cells.] Republique Francais: Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963.

Priore, A. (1966) "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966.

Priore, A. (1968). "Apparatus for producing radiations penetrating living cells." U.S. Patent No. 3,368,155. Feb. 6, 1968.

Priore, A. French Patent No. 1,342,772. Letter, R. Pautrizel to his colleagues at the University of Bordeaux, Sept. 11, 1979. Letter, R. Pautrizel to his colleagues, Mar. 26, 1969.

Graille, Jean-Michel, ***Le Dossier Priore****, [Dossier Priore: A New Pasteur Affair] De Noel, Paris, 1984 [in French]. Details the entire Priore affair.*

Pautrizel, R., M.R. Riviere, A. Priore, and F. Berlureau. (1966) "Influence d'ondes electromagnetiques et de champs magnetiques associes sur l'immunie de la souris infestée par Trypanosoma equiperdum," [Influence of electromagnetic waves and associated magnetic fields on the immunity of the mouse infected with the Trypanosoma equiperdum] ***Compt. Rend. Acad. Sci.*** *(Paris), 1966, Vol.60,p.2099-2102.[in French].*

Pautrizel, R.; A. Priore, F. Berlureau, and A.N. Pautrizel, "Stimulation par des moyens physiques des defenses de la Souris et du Rat contre la trypanosomiase experimentale." [Stimulaton, by physical means, of defenses of the mouse and of the rat against the experimental Trypanosoma.] ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 268, 1969, p. 1889-1892. [In French].*

Pautrizel, R., A. Priore, A.N. Pautrizel, and P. Chateau-Reynaud-Duprat.(1979) "Guersson de la trypanosomiase experimentale par l'association de champs magnetiques et d'ondes electromagnetiques: une stimulation des defenses de l'organisme." [Cure of experimental trypanosomias by associated magnetic fields and electromagnetic waves: a stimulation of the organism's defenses.].

Communication presented to the Journées **Nationales Microondes - Colloque Hertzienne et Dielectriques**, Lille 26-29 juin 1979, p. 31.

Pautrizel, R.; A. Priore, F. Berlureau, and A.N. Pautrizel, "Action de champs magnetiques combines a des ondes electromagnetiques sur la trypanosomiase experimentale du Lapin." [Action of magnetic waves combined with electromagnetic waves on the experimental trypanosoma of the rabbit.] **Compt. Rend. Acad. Sci.** (Paris), Vol. 271, 1970, p. 877-880.

Pautrizel, R.; A. Priore, M. Dallochio and R. Crockett, "Action d'ondes electromagnetiques et de champs magnetiques sur les modifications lipidiques chez le Lapin par l'administration d'un regime alimentaire hypercholestrole." [Action of electromagnetic waves and magnetic fields on provoked lipidic modifications in the rabbit by the administration of an alimentary hypercholesterol diet.] **Compt. Rend. Acad. Sci.** (Paris), Vol. 274, 1972, p. 488-491. [In French].

Pautrizel, R.; A. Priore, P. Mattern, and A.N. Pautrizel, "Stimulation des defenses de la souris trypanosomee par l'action d'un rayonnement associant de champ magnetique et ondes electromagnetiques." [Stimulation of the defenses of the trypanosomized mouse by the action of a ray associated with a magnetic field and electromagnetic waves." **Compt. Rend. Acad. Sci.** (Paris), Vol. 280, 1975, p. 1915-1918. [In French]

Pautrizel, R.; A. Priore, P. Mattern, A.N. Pautrizel, and A. Capbern, "Guerison de la trypanosomiase chronique du Lapin a Trypanosoma equiperdum par l'action combinee de champs magnetiques et d'ondes electromagnetiques modules." [Healing of chronic trypanosomia by trypanosoma equiperdum of the rabbit by the combined action of magnetic waves and modulated electromagnetic waves.] **J. Protozool.**, Vol. 22, No. 3, 1975, p. A 84.

Pautrizel, R. "La trypanosomiase experimentale: stimulation des defenses de L'organisme par des moyens physiques." [Experimental trypanosomiasis: stimulation of the organism's defenses by physical means.] **XVIIth Seminar on Trypanosomiasis Research**, Londres, 22-23 Sept. 1976.

Pautrizel, R.; P. Mattern, A.N. Pautrizel, and A. Priore, "Effets des champs magnetiques et des ondes electromagnetiques modules sur la trypanosomiase experimentale." [Effect of magnetic fields and modulated electro magnetic waves on experimental trypanosomiasis]. **Ann. Soc. Belge Med. trop.** [Annals of the Belgium Society of Tropical Medicine], Vol. 57, 1977, p. 501-523.

Pautrizel, P. Mattern, A. Priore, A.N. Pautrizel, A. Capbern, and T. Baltz, "Importance des mecanismes immunitaires dan la guerison de la trypanosomiase experimentale par stimulation physique." [Importance of immune mechanisms in the cure of the experimental trypanosomias by physical stimulation.] **Compt. Rend. Acad. Sci.** (Paris), Vol. 286, 1978, p. 1487-1492. [In French]

Pautrizel, R. and A. Priore, "Un aspect spectaculaire du bioelectro-magnetisme: permettre a l'organism de se debarrasser d'un infection aig ue ou chronique sans l'aide d'aucune substance medicamentese." [A spectactular aspect of bioelectromagnetism: Permitting the organism to rid itself of an old chronic infection without the aid of any medical substance.] **Communication presented to the 104eme Congres National des Societes Savantes,** Bordeaux, 17-21 Avril 1979 – section **Sciences,** p. 112.

Pautrizel, R. and A. Priore, "Guerison de la trypanosomiase experimentale par association de champs magnetiques et d'ondes electromagnetiques." [Cure of trypanosomiasis by associated magnetic fields and electromagnetic waves.] **Communication presented to the Symposium International des Therapeutiques Ondulatoires,** Versailles, 19-20 Mai 1979, p. 9.

Pautrizel, A. Priore, A.N. Pautrizel and P. Chateau-Reynaud-Duprat, "Guerison de la trypanosomiase experimentale par l'association de champs magnetiques et d'ondes electromagnetiques: une stimulation des defenses de l'organism." [Cure of trypanosomiasis by associated magnetic fields and electromagnetic waves: a stimulation of the organism's defenses.] Communication presented to the **Journees Nationales Microondes - Colloque Hertzienne et Dielectriques**, Lille 26-29 Juin 1979, p. 31.

Riviere, M.R., A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Phenomenes de regression observes sur les greffes d'un lymphosarcome chez des souris exposées a des champs electromagnetiques." [Phenomena of regression observed on the graftings of a lymphosarcoma in mice exposed to electromagnetic fields]. **Compt. Rend. Acad. Sci.** (Paris), Vol. 260, 1965, p. 2099-2102 [in French]. This reports the curing of lymphosarcoma, which is a malignant lymphoma that tends to metastasize freely and spread throughout the body. It easily leads to the condition of lymphomatosis, which is the presence of multiple lymphomas in the body, including wide distribution through the various organs, lymph tissue, and tissue resembling lymph tissue.

Riviere, M.R.; A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Action de champs electromagnetiques sur les greffes de la tumeur T8 chez le Rat." [Action of the electromagnetic fields on the graftings of the T8 tumor on the rat." **Compt. Rend. Acad. Sci.** (Paris), Vol. 259, 1964, p. 4895-4897.

Riviere, M.R.; A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Effets de champs electromagnetiques sur un lymphosarcome lymphoblastique transplantable du Rat." [Effects of electromagnetic fields on lymphoblastic lymphosarcoma transplantable from a rat.] **Compt. Rend. Acad. Sci.** (Paris), Vol. 260, 1965, p. 2099-2102.

COMMENT: Lymphosarcoma is a malignant lymphoma (malignant tumor of lymph or lymph-like tissue) that tends to metastasize freely and spread throughout the body. It easily leads to the condition of lymphomatosis, which is the presence of multiple lymphomas in the body, including wide distribution through the various organs, lymph tissue, and tissue resembling lymph tissue. A lymphoblast is a cell giving rise to lymphocytes, the colorless weakly motile cells produced in lymphoid tissue and include the cellular mediators of immunity, constituting some 20 to 30 percent of the leukocytes (white blood cells) of normal human blood. In effect the experimenters have healed a transplanted condition that represents a serious leukemia.

Riviere, M.R. and M. Guerin., "Nouvelles recherches effectuees chez des rats porteurs d'un lymphosarcome lymphoblastique soumis a l'action d'ondes electromagnetiques associees a des champs magnetiques." [New research on rats having lymphoblastic lymphosarcoma, submitted to the action of electromagnetic waves associated with magnetic fields.] **Compt. Rend. Acad. Sci.** (Paris), Vol. 262, 1966, p. 2669-2672.

COMMENT: Lymphosarcoma is a malignant lymphoma (malignant tumor of lymph or lymph-like tissue) that tends to metastasize freely and spread throughout the body. It easily leads to the condition of lymphomatosis, which is the presence of multiple lymphomas in the body, including wide distribution through the various organs, lymph tissue, and tissue resembling lymph tissue. A lymphoblast is a cell giving rise to lymphocytes, the colorless weakly motile cells produced in lymphoid tissue and include the cellular mediators of immunity, constituting some 20 to 30 percent of the leukocytes (white blood cells) of normal human blood. In effect the experimenters have healed a condition that represents a serious leukemia.

Delmon, G. and J. Biraben, "La croissance du carcinome de Guerin sour l'action de champs electromagnetiques." [The growth of carcinoma treated by the action of electromagnetic fields.] **Rev. Path. Comp.,** 3(2), 1966, p. 85-88.

Berteaud, A.J.; A.M. Bottreau, A. Priore, A.N. Pautrizel, F. Berlureau, and R. Pautrizel, "Essai de correlation entre et l'action d'une onde electromagnetique pulsee et modulee." [Trial of the correlation between the evolution of a disease by Trypanosoma equiperdum and the action of a pulsating and modulated electromagnetic wave.] **Compt. Rend. Acad. Sci.** (Paris), Vol. 272, 1971, p. 1003-1006. [In French].

Mayer, G.; A. Priore, G. Mayer and R. Pautrizel, "Action de champs magnetiques associes a des ondes electromagnetiques sur l'orchite trypanosomienne due lapin." [Action of magnetic fields associated with electromagnetic waves on the typanosomian orchitis of the rabbit.] **Compt. Rend. Acad. Sci.** (Paris), Vol. 274, 1972, p. 3011-3014. [In French]

COMMENT: Orchitis is inflammation of a testes. Trypanosomes are protozoan flagellates of genus trypanosoma which infect humans and animals and are responsible for various serious diseases such as Chaga's disease, dourine, nagana, sleeping sickness, and surra. Chaga's disease is marked by prolonged high fever, edema, and enlargement of the spleen,

liver, and lymph nodes. Dourine is a disease favoring horses and asses, marked by inflammation of the genitals, subcutaneous edematous plaques, low-grade fever, progressive paralysis, emaciation, and death. Nagana is a highly fatal disease in/of domestic animals in tropical Africa marked by fluctuating fever, inappetance, edematous swelling, and sluggishness, and is transmitted by the tsetse fly and possibly by other biting flies. Sleeping sickness is a serious disease that is prevalent in much of tropical Africa; it is marked by fever, protracted lethargy, tremors, and loss of weight, and is transmitted by tsetse flies. Surra is a severe Old World febrile and hemorrhagic disease of domestic animals and is transmitted by biting insects.

Courrier, R., "Expose par M. le Professeur R. Courrier, Secretaire Perpetuel de L'Academie des Sciences fait au cours d'une reunion a L'Institut sur les effets de la Machine de M.A. Priore le 26 Avril 1977." [Presentation by Professeur R. Courrier, Perpetual Secretary of the Academy of Sciences, made at the meeting of the Academy on the effects of the machine of M.A. Priore.] [In French] Courrier's presentation of the Priore machine and its positive cures of terminal cancers and leukemias in laboratory animals, under proper scientific protocols.

Ruth, Barnard. "Experimental investigations on ultraweak photon emission." In **Electromagnetic Bio-Information: Proceedings of the Symposium**, Marburg, Sept. 5, 1977. Proceedings edited by Fritz Albert Popp et al., Urban and Schwarzenberg, Munchen Wien, and Baltimore, 1979, p. 107-122.

Ryaby, J.T.; D.A. Grande, F.P. Magee, and A.M. Weinstein, "The effect of combined AC/DC magnetic fields on resting articular cartilage metabolism." Presented at the 1st World Congress for Electricity and magnetism in Biology and Medicine, Orlando, FL, June 14-19, 1992.

Vladimirov, Yu. A., "Ultraweak luminescence of subcellular structures," **Ultraweak Luminescence in biology, Proceedings of the Symposium,** Moscow, 1969 [In Russian].

Kaznacheyev, V.P., "Informational function of ultraweak flows of light in biological systems," in **Materials of Science Conference: Physical and Mathematical methods of Research in Biology and Medicine, Novosibirsk,** 1965, p. 38-41 [In Russian].

Kaznacheyev, V.P. et al, "Studies on the biological role of electromagnetic emission as a factor of adaptive behavior of cells under conditions of lattitude changes," **Abstracts of the Second All-Union Conference Devoted to different Geographical, Climatic and Labor Conditions,** Novosibirsk, 1977, p. 101-104 [In Russian].

Kaznacheyev, V.P. et al., "Conditions necessary for appearance of distant intercellular interactions after UV radiation," **Bulleten Experimentalnoy Biologii i. Meditsiny,** No. 5, 1979, p. 468-471.

Possony, Stefan T., "Psy-War: Soviet device experiment," **Defense and Foreign Affairs Daily,** 12(104), June 7, 1983, p. 1-2.

Kaznacheyev, V.P. and L.P. Mikhailova, "Ultraweak Radiations in Intercellular Interactions," Novosibirsk, 1981 [In Russian]. English translation: San Francisco, Washington Research Center, 1982.

Kaempffert, Waldemar, "'Dead' organisms revived," Science in the News: **Science Digest,** May 1950, p. 77-78.

"Rejuvenating Light," **Scientific American** May 1949.

Popp, Fritz Albert; Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, Sept. 5, 1977, Urban & Schwarzenberg, Baltimore, 1979.

Lilienfeld, A.M. et al, **Foreign Service Health Status Study: Evaluation of Health Status of Foreign Service and Other Employees From Selected Eastern European Posts,** Department of Epidemiology, Of Hygiene, and Public Health; The Johns Hopkins University, Baltimore, Maryland, final Report, July 31,1978.

Mallaliew, Robert C., **A Model of the Microwave Intensity Distribution within the U.S. Embassy, Moscow 1966 to 1977,** Report FS-80-166, Applied Physics Laboratory, Johns Hopkins University, Aug. 1980.

Microwave Radiation at the U.S. Embassy in Moscow and Its Biological Implications: An Assessment, U.S. Department of Commerce Doc. No. NTIA-81-12, Mar. 1981.

Novak, Rachel. "Flesh-Eating Bacteria: Not New, But Still Worrisome." Research News, **Science,** 264(5166), June 17, 1994, p. 1665.

Olariu, S. and I. Iovitzu Popescu. "The Quantum Effects of Electromagnetic fluxes," **Reviews of Modern Physics,** 57(2), April 1985, p. 339-436.

Opall, Barbara. "U.S., Russia Hope to Safeguard Mind-Control Techniques." **Defense News,** Jan. 11-17, 1993, p. 4, 29.

Penrose, R. "The Nonlinear Graviton." **General Relativity and Gravitation,** 7(2), 1976, p. 171.

Perez, Jose and Carson Jeffries. "Effects of additive noise on a nonlinear oscillator exhibiting period doubling and chaotic behavior." **Physical Review B. Condensed Matter.** Series 3, 16(6), Sept. 15, 1982, p. 3460-3462.

Phipps, Thomas E. Jr. "Critique of accepted theories of special relativity and electromagnetism," **Scientific Ethics,** 1(2), Apr. 1985, p. 17-23.

Popp, Fritz-Albert and Ke-hsueh Li. "Hyperbolic Relaxation as a Sufficient Condition of a Fully Coherent Ergodic Field." **International Journal of Theoretical Physics,** 32(9), Sep. 1993, p. 1573-1582.

Reid, B.L. "Propagation of properties of chemical reactions over long distance in the atmosphere as seen by crystal growth pattern changes." **Australian Journal of Medical Laboratory Science,** Vol. 7, Feb. 1986, p. 30-35.

Rodemann,. H.P., H.P. Peterson, K. Wchwenke, and K. Hartmut von Wangenheim. Terminal differentiation of human fibroblasts is induced by radiation." **Scanning Microscopy,** 5(4), Dec. 1991, p. 1135-43.

Rodemann, H.P.; K. Bayreuther, and G. Pfleiderer. "The differentiation of normal and transforman fibroblasts in vitro is influenced by electromagnetic fields." **Experimental Cell Research,** Vol. 182, 1989, p. 610-621.

Santilli, Ruggero Maria. "Partons and gravitation: some puzzling questions." **Annals of Physics,** 83(1), Mar. 1974, p. 108-157. Proves that electric field and gravitation are not mutually exclusive. One is left with one of two assumptions: (1) the strong assumption, that both are identically the same thing, or (2) the weak assumption, that both are partially the same thing.

Saxl, Erwin J. and Mildred Allen. "1970 solar eclipse as 'seen' by a torsion pendulum." **Physical Review D. Particles and Fields,** 3(4), Feb.15, 1971, p. 823-825.

Sheldrake, Rupert. **A New Science of Life**: The Hypothesis of Formative Causation, Blond & Briggs, Ltd., U.S. reprint, J.P. Tarcher, Inc., Los Angeles, 1982.

Smetanin, E.V., "Electromagnetic field in a space with curvature – new solutions," **Soviet Physics Journal,** 25(2), Feb. 1982, p. 107-111.

Tapscott, Mark. "DOD, Intel Agencies Look at Russian Mind Control Technology, Claims." **Defense Electronics,** 25(8), July 1993, p. 17.

von Borzeszkowski, Horst-Heino. "Remarks on the Physical Reality of Gravitons," **Foundations of Physics,** 20(4),1990, p. 435-445.

Wheeler, John Archibald and Richard Phillips Feynman. "Classical electrodynamics in terms of direct interparticle action," **Reviews of Modern Physics,** 21(3), July 1949, p. 425-433.

Goodman, Glenn W. Jr. "Upping the Nonlethal Ante," **Armed Forces Journal International,** July 1994, p. 13.

Jammer, Max. **Concepts of Space: The History of Theories of Space in Physics.** Harvard University Press, Cambridge, Massachusetts, 1969

Jammer, Max. **Concepts of Force.** Harvard University Press, Cambridge, Massachusetts, 1957.

Jammer, Max. **Concepts of Mass.** Harvard University Press, Cambridge, Massachusetts, 1961.

Javier, R.T.; F. Sedarati, and J. G. Stevens. "Two avirulent Herpes simplex viruses generate lethal recombinants in vivo." **Science,** 234(4777), Nov. 7, 1986, p. 746-747.

Wheeler, J.A. and R.P. Feynman. "Classical Electrodynamics in Terms of Direct Interparticle Action," **Reviews of Modern Physics,** Vol. 21, 1949, p. 425-433. The electromagnetic field as such does not actually exist in spacetime, but is a summary of what we need to know about distant charges in order to predict the behavior of a charge if there should be one at that point. In other words, the EM field represents an information storage device.

Jaynes, E.T. "Probability In Quantum Theory," in **Complexity, Entropy, and the Physics of Information,** Ed. W.H. Zurek, 1990, pp. 33-55, Addison-Wesley, Redwood City, CA. Strongly points out that the EM field does not exist in spacetime, but represents an information storage device.

Editorial Note: We would like to express our appreciation to Mr. Bearden for this wonderful article.

The Tom Bearden Website

"Dossier Priore"

by Jean-Michel Graille

Photographs

Cover	
Plate 1	The inauguration: the construction of the M 600 had not yet started but, symbolically, the Mayor of Bordeaux starts the smaller machine at close range. (Archives R.P.)
Plate 2	Inaugural scroll.
Plate 3	An information meeting held in Bordeaux. From left to right: Antoine Priore, Robert Courrier, Raymond Pautrizel and Francis Berlureau. (Archives R.P.)
Plate 4	February 1975. Antoine Priore and Professor Pautrizel at the "suspended" base of the M 600, the enormous experimental prototype built by Leroy-Somer. The apparatus would only function for a few days but did make it possible to carry out some experiments with very promising results.
Plate 5	Anne-Nelly Pautrizel. Professor Pautrizel's sister and collaborator, who participated in almost all the experiments with the Priore apparatus since 1968. (Photo J.M.G.)
Plate 6	Pierette Châteaureynaud-Duprat, head of research at C. N.R.S., doctor of science and specialist in grafts, carried out in 1969 and 1970 two series of decisive experiments concerning the proof and the explanation of the biological effectiveness of the Priore Ray.

Plate 7

The Priore machine. The model financed by Pautrizel, the only one which is still likely to function, the one with which patients will be officially looked after as from 1977, is finished in August 1968. In fact, the apparatus is installed all in one place.

Plate 8

1969. Old friends of Priore retrace the road to Floirac. With a monkey-wrench or soldering-iron in hand, they devote all their leisure time to building the new machine for their friend. Here, on the scales, inspector Plantin, of the legal identity service. In coveralls, scarf around his neck, municipal worker M. Léglise.
(Archives R.P.)

Plate 9

A courageous man. Perpetual Secretary of the Academy of Sciences, elder President of the Academy of Medicine, "father" of endocrinology and fertilization, Professor Robert Courrier could have benefited from its prestige and its retirement. At 67, he took all the risks to defend that which he believed in deeply. He has fought now for twenty years. One of the rare photos of him, in the company of Antoine Priore and Professor Pautrizel, taken on 25th March 1970 in the laboratory at Floirac.
(Archives R.P.)

Jean-Michel Graille

DOSSIER
PRIORE

Une
nouvelle affaire
Pasteur?

Denoël

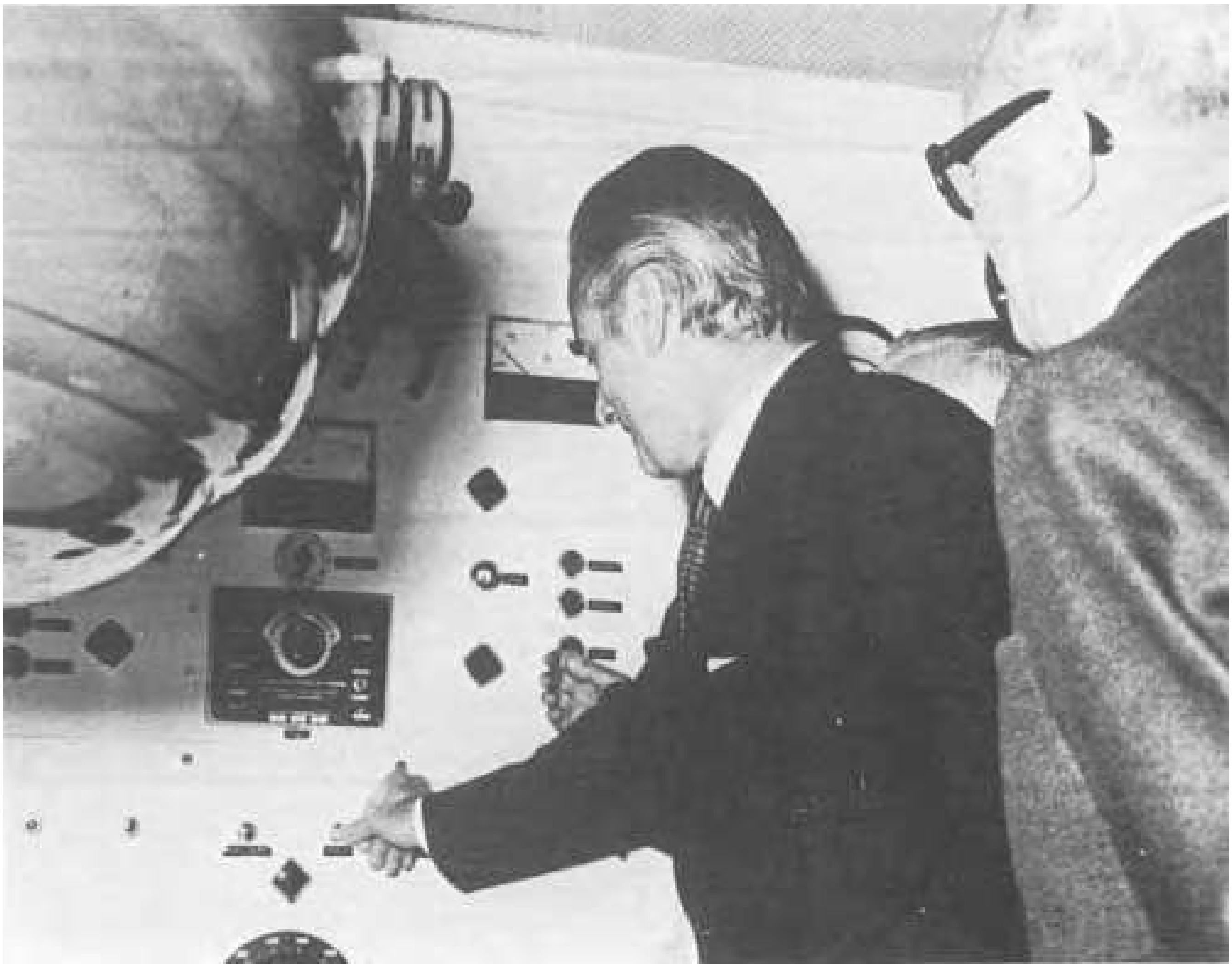

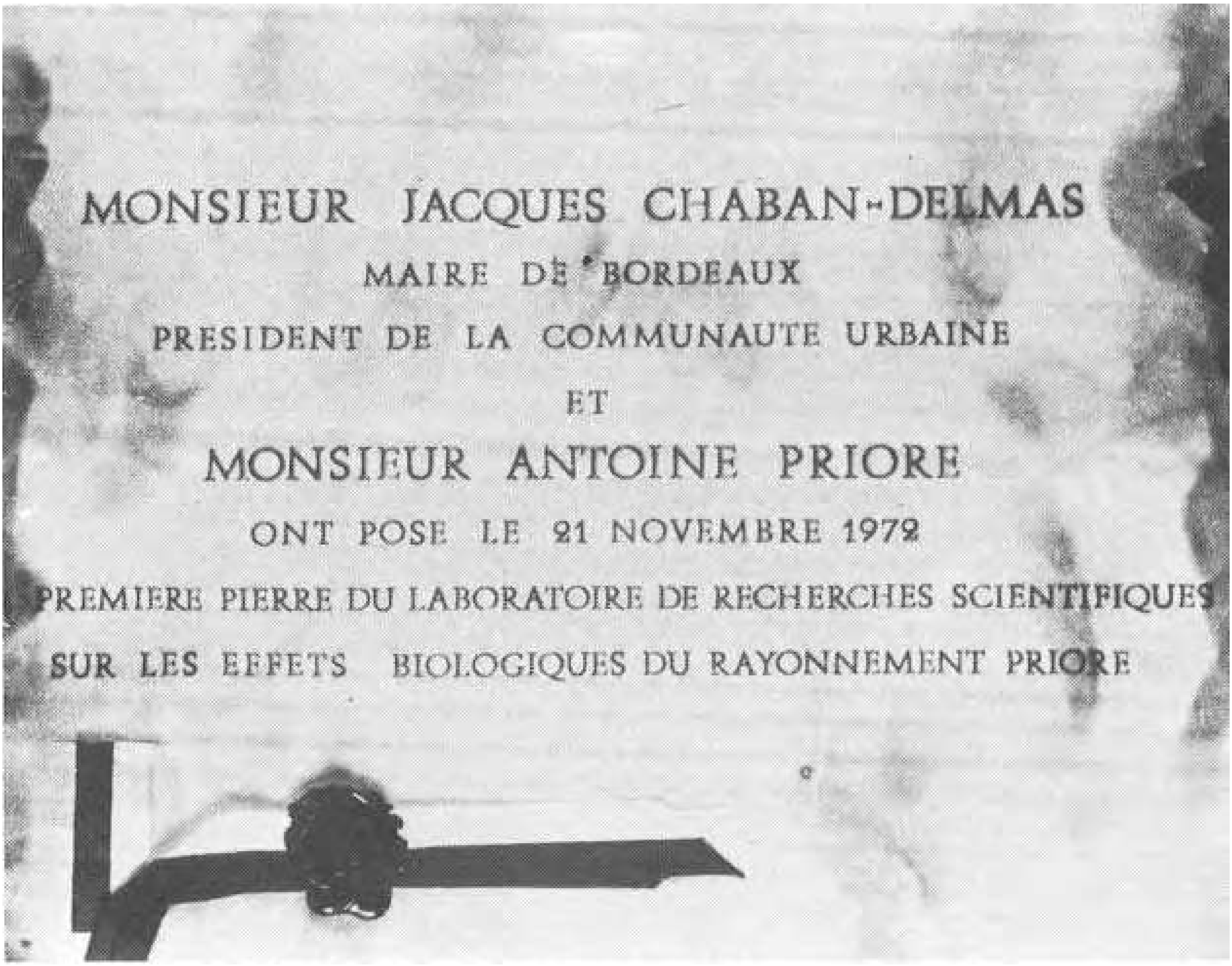

MONSIEUR JACQUES CHABAN-DELMAS
MAIRE DE BORDEAUX
PRESIDENT DE LA COMMUNAUTE URBAINE
ET
MONSIEUR ANTOINE PRIORE
ONT POSE LE 21 NOVEMBRE 1972
PREMIERE PIERRE DU LABORATOIRE DE RECHERCHES SCIENTIFIQUES
SUR LES EFFETS BIOLOGIQUES DU RAYONNEMENT PRIORE

Translation:

JACQUES CHABAN-DELMAS
THE MAYOR OF BORDEAUX
PRESIDENT OF THE URBAN COMMUNITY
AND
ANTOINE PRIORE
HAVE INSTALLED ON THE 21 NOVEMBER 1970
THE FIRST STONE OF THE SCIENTIFIC RESEARCH LABORATORY
FOR THE BIOLOGICAL EFFECTS OF THE PRIORE RAY

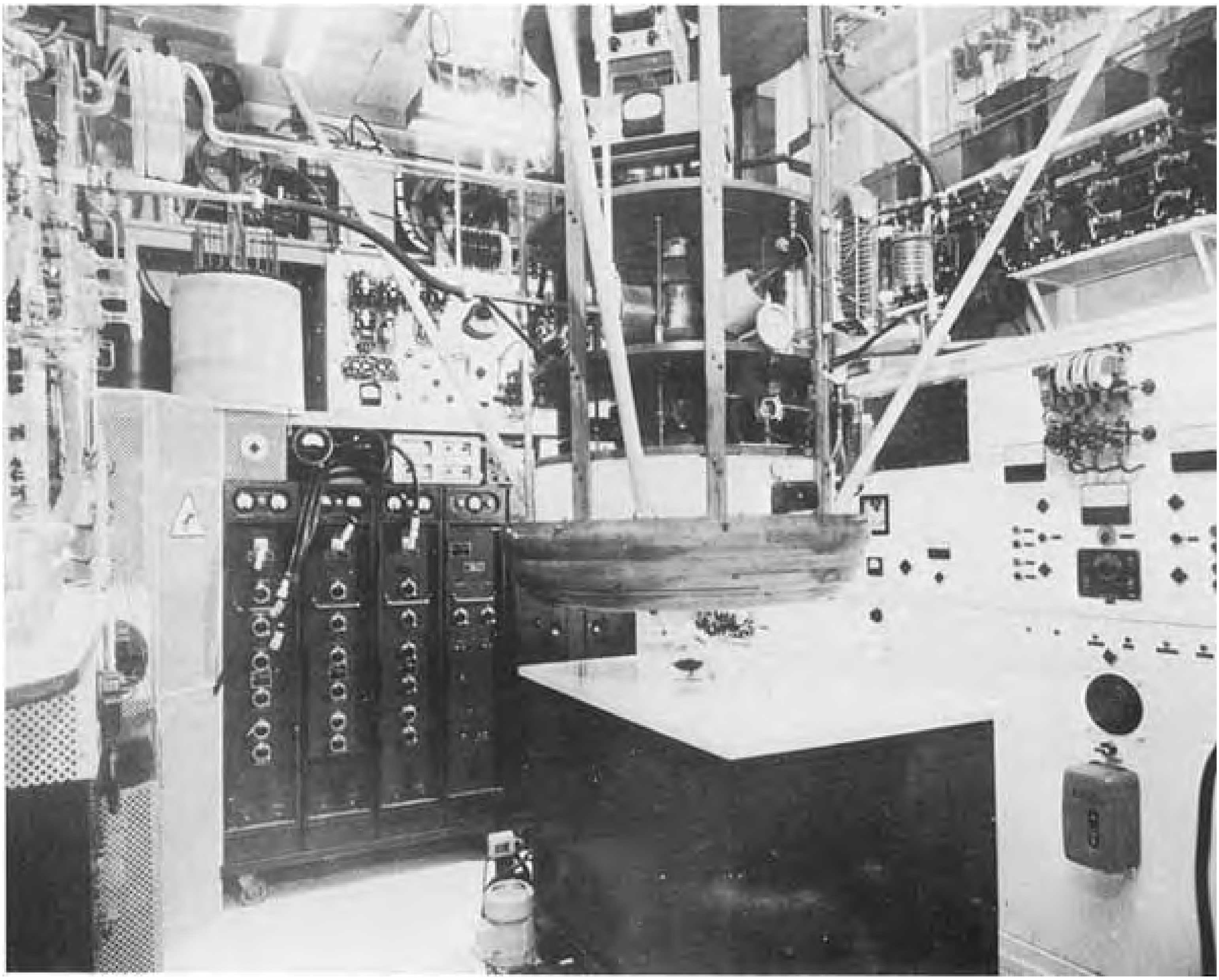

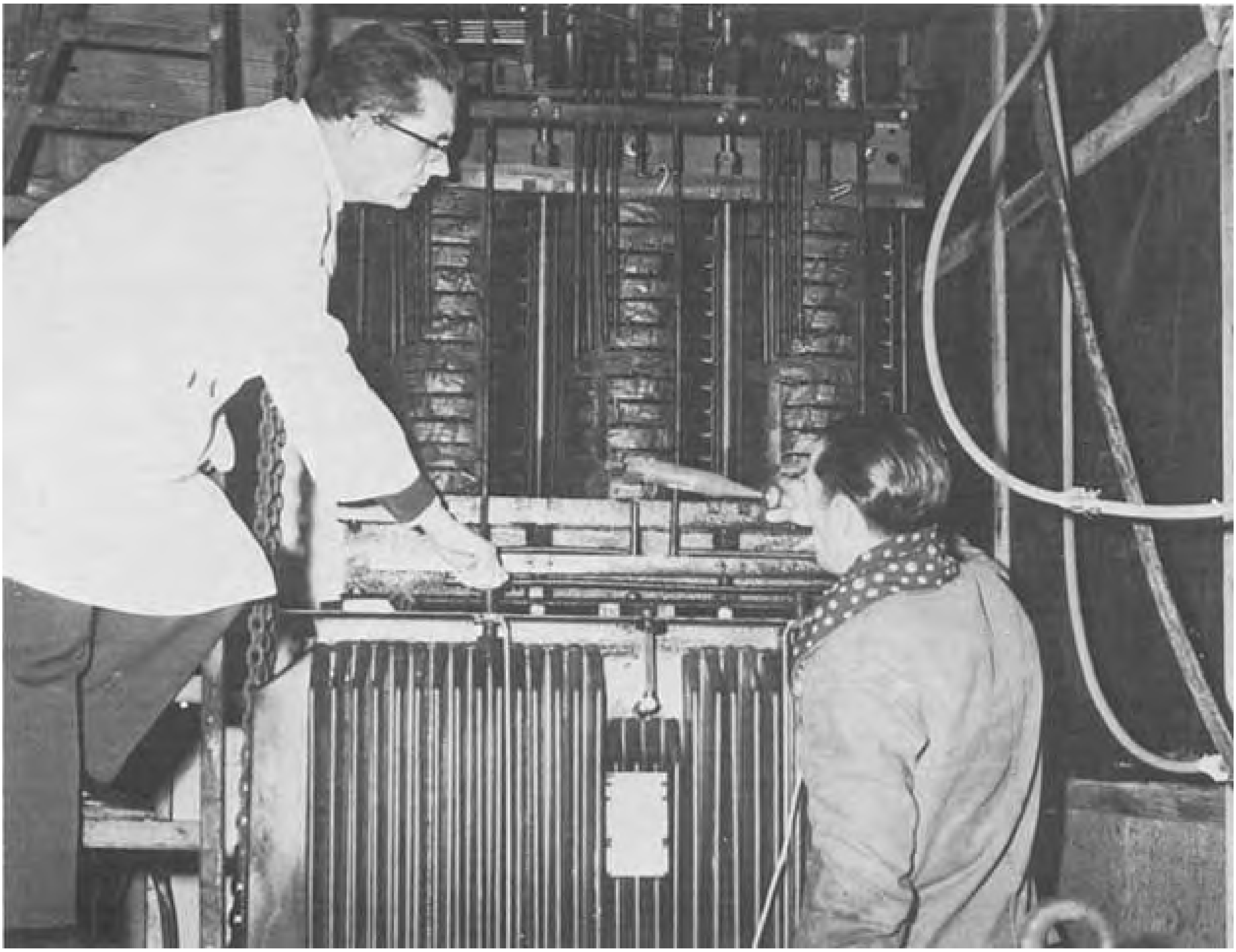

Making Laser Weapons a Reality

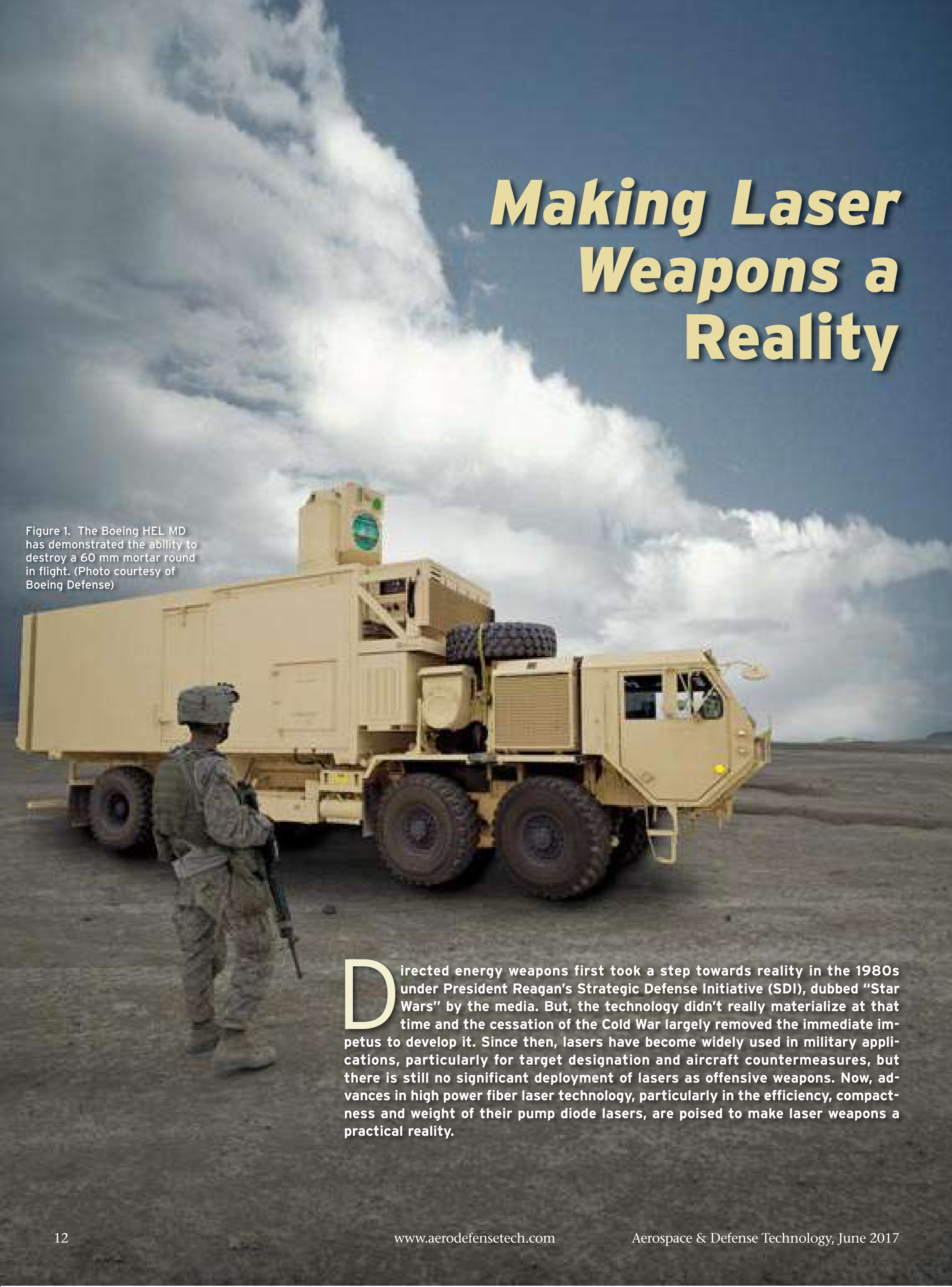

Figure 1. The Boeing HEL MD has demonstrated the ability to destroy a 60 mm mortar round in flight. (Photo courtesy of Boeing Defense)

Directed energy weapons first took a step towards reality in the 1980s under President Reagan's Strategic Defense Initiative (SDI), dubbed "Star Wars" by the media. But, the technology didn't really materialize at that time and the cessation of the Cold War largely removed the immediate impetus to develop it. Since then, lasers have become widely used in military applications, particularly for target designation and aircraft countermeasures, but there is still no significant deployment of lasers as offensive weapons. Now, advances in high power fiber laser technology, particularly in the efficiency, compactness and weight of their pump diode lasers, are poised to make laser weapons a practical reality.

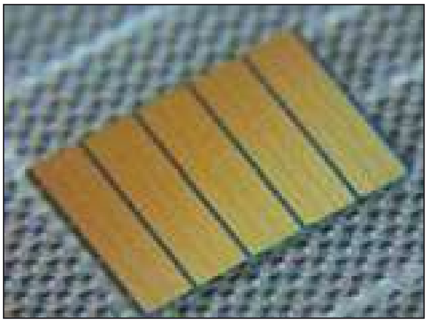

Figure 2. A Coherent | DILAS T-Bar chip contains five emitters, outputs about 50 W, and is 5 mm wide.

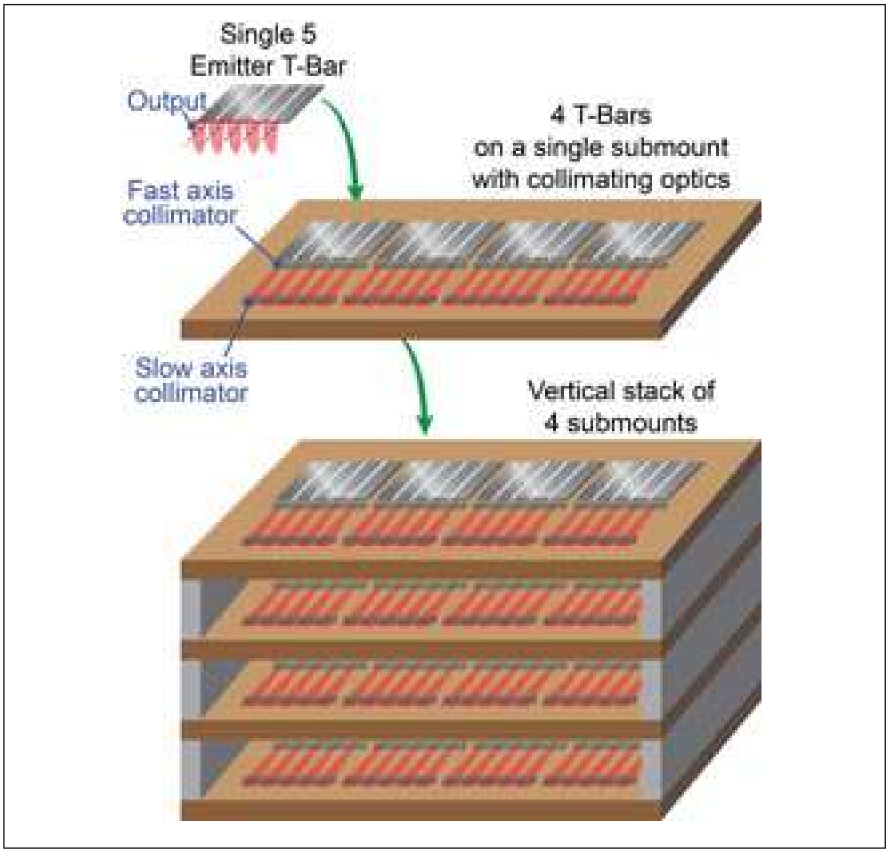
Single 5 Emitter T-Bar
Output
4 T-Bars on a single submount with collimating optics
Fast axis collimator
Slow axis collimator
Vertical stack of 4 submounts

Figure 3. Progression from a single T-Bar die to an 80-emitter stack. The light from all these emitters can be efficiently coupled into a single 225 μm core fiber with a 0.22 NA.

Laser Weapons Overview

The concept behind Reagan's SDI was to protect the United States behind a "missile shield" that would prevent nuclear weapons, delivered via either intercontinental or submarine-launched ballistic missiles, from reaching North America. Part of the SDI was to be lasers, or other directed energy weapons, placed in permanent orbit, which could rapidly knock out a nuclear missile already in flight towards its target. Lasers, in particular, were a promising technology for this application because they can be quickly "charged up" (i.e. be made ready to fire), and their output can reach a fast-moving target nearly instantaneously.

Because these lasers were intended to be deployed primarily in space, or in fixed locations, and used either never (hopefully) or just once, there wasn't much concern about their cost, physical size, weight or operational efficiency. The key goal was simply that they would work when needed, delivering the required knock-out punch.

The offensive laser weaponry under development today is intended to serve almost the exact opposite purpose than the SDI, and therefore has very different goals in terms of its design and functionality. For example, one major laser weapons initiative is for C-RAM (Counter Rocket, Artillery and Mortar). These are systems intended to be deployed in the battlefield to protect troops and equipment from incoming projectiles. Currently, these are envisioned to be single, vehicle mounted lasers, with output powers in the 10 kW to 50 kW range.

Boeing has already successfully prototyped such a system, which they call the High Energy Laser Mobile Demonstrator (HEL MD). This consists of a 10 kW solid state laser installed on an Oshkosh Tactical Military Vehicle, along with all the necessary targeting and control systems. HEL MD has proven the ability to lock on to and destroy a mortar round of about 10 inches in length, traveling at hundreds of miles per hour, from several miles out. The system is also effective against Unmanned Aerial Vehicles (UAVs or drones); in this case, it may be sufficient to damage the drone's navigation and targeting systems, rather than completely destroy it. Similarly, other offensive laser weapons systems are being designed to deliver "swarm defense," that is, to protect a ship or base from attack by a large number of inexpensive drones.

Thus, today's laser weapons are meant to defend against swarms of inexpensive mortar rounds, drones and other projectiles, and are intended to be mounted on platforms which are mobile and self-contained (meaning they may not have ready access to an external source of power). This gives rise to several design imperatives. First, they must deliver a low cost per engagement; it's simply not economically feasible to employ an expensive weapon (such as a missile costing over $100,000) to knock out a drone which costs $1,000, or a mortar round which might cost less than $100. The laser weapon must also be capable of rapid fire so that it can't just be overwhelmed by a large number of simultaneous incoming rounds. Also, rapid fire reduces the number of individual laser weapons systems needed to protect a given number of troops. But, the ability to deliver rapid fire also necessitates that the system be electrically efficient. Otherwise the laser weapon will require access to a large quantity of fuel, which is difficult (and dangerous) to transport into a battlefield. Finally, it's desirable that laser weapons be physically compact, especially for airborne use.

SWaP Optimization

The push to optimize the size, weight and power (called SWaP in military parlance) of field deployed laser weapons has driven a progression in the technology of the gain material used from chemical (e.g. deuterium fluoride), to solid state, and, most recently, to fiber. One reason for this is because fiber lasers offer inherently higher efficiency, in terms of converting input pump energy to usable output, than nearly any other laser type, except diode lasers. However, diode lasers by themselves don't provide the necessary brightness or beam quality, while diode-pumped fiber lasers can. And, beam quality is critical because it determines the distance over which the laser can be focused to a spot size small enough to reach the power density necessary to damage its target.

Fiber lasers offer other advantages in terms of SWaP optimization. Chemical and lamp pumped solid state lasers require a significant overhead in terms of the equipment and power supplies required to run them. In contrast, both solid state (e.g. slab and rod) and fiber lasers can be diode pumped, and diode power supplies and pump modules themselves are electrically efficient and lend themselves to miniaturization. This efficiency, in turn, reduces the cooling requirements, and all its attendant equipment (pumps, heat exchangers, etc.).

But, as laser weapons evolve towards ever higher powers, fiber laser technology becomes increasingly attractive over solid state. Currently, individual fiber lasers can deliver up to about 2 kW of power, but multiple units can be combined to deliver around 10 kW in a single beam with extremely good mode quality. It's a bit more difficult to scale up the power of a solid-state laser, especially while maintaining good mode quality. This is because scaling up and/or increasing the pump power to a traditional solid state laser typically excites higher order modes, and introduces other issues related to thermal lensing.

Of course, there are also challenges to increasing the power of a fiber amplifier. And, because the optical efficiency of their pumping is already so high (~85%), there's not much room for improvement there. So, the focus is on raising the pump power itself, without increasing package size and weight, as well as on methods for low loss beam combining of multiple single mode lasers, all while still maintaining good beam quality.

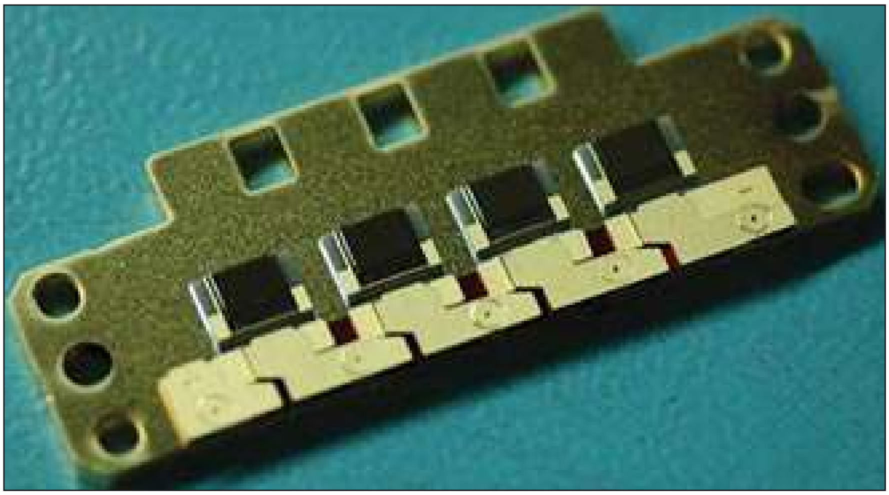

Figure 4. Four T-Bars mounted on a macrochannel cooler with collimating optics.

Pumping Technology Advances

In terms of SWaP optimization of pump diodes for laser weapons, a current benchmark is to achieve a weight to output power ratio of 1 kg/kW. Coherent | DILAS has developed several advances in diode laser technology to reach this goal, and, in fact, is already working on devices that will attain the next level of performance – 0.5 kg for 1 kW of output.

One key to reaching the SWaP benchmark has been the company's introduction of T-Bar (for "tailored" bar) construction, a design approach intended to combine the high total output power of traditional diode laser bars with the relaxed cooling requirements of single emitters. The basic unit of the T-Bar is a diode laser mini-bar having five emitters on a single, 5 mm wide substrate which outputs about 50 W total. In contrast, traditional diode laser bars are usually 10 mm wide, contain at least 19 emitters, and can output over 200 W, depending on the number of emitters and type of heatsink.

For military applications, four of these T-Bar dies are mounted on to a single substrate, yielding a total of 20 emitters. Then, up to four of these substrates are stacked vertically, bringing the total number of individual emitters in this compact assembly to 80, with a total output of about 800 W.

The key optical characteristic of the T-Bar design is that the combination of wide emitter spacing, low divergence and relatively low beam parameter product (in both fast and slow axes) is tailored to enable the light from all 80 of these individual emitters to be efficiently co-linearized and coupled into a single 225 µm core fiber having a numerical aperture of only 0.22. This, in turn, permits highly efficient coupling into the pumping mode volume of the gain fiber. And, this light collection can be accomplished using a relatively simple and compact optical system. This is how output power per unit volume is maximized.

In contrast, the high divergence and poor mode quality of traditional diode laser bars necessitate the use of complex optics for light collection, and still make it impossible to couple all their output into a single, low numerical aperture fiber. Furthermore, high power diode laser bars often suffer from a problem called "smile." This is a sagging or bending alignment error in emitter position along the bar which makes it quite difficult to collimate and co-linearize its output.

Besides the advantage of producing the largest amount of fiber coupled light out in a given system volume, T-Bar construction also offers significantly improved cooling characteristics over traditional bars. In a traditional diode laser bar, the small emitter spacing causes significant thermal crosstalk between the individual emitters, and a very high efficiency cooling system must therefore be employed. This is usually in the form of a so called "microchannel" cooler, in which water is rapidly circulated through channels within the heatsink. But the small bore of the microchannels (tens of microns) makes it easy for them to be clogged by particulates. Thus, the water must be stringently conditioned and filtered, and large, heavy, high pressure pumps must be employed to circulate it rapidly enough to achieve the required level of cooling in these traditional bars.

Furthermore, traditional diode bars are usually placed directly on the heatsink in order to achieve good thermal contact. However, this makes the microchannel cooler part of the electrical circuit, thus necessitating the use of deionized water in order to avoid a short circuit. This imposes yet another requirement that increases cost and complexity.

The larger emitter spacing of the T-Bar largely eliminates this thermal crosstalk, and greatly relaxes the attendant cooling requirement. As a result, Coherent | DILAS has been able to introduce a "macrochannel" cooler. This is a system which uses substantially larger bore channels, thus allowing the use of less stringently filtered tap water, and eliminating the need for high pressure pumping. Also, the reduced cooling requirement enables the laser bar to be placed on a submount, rather than in direct electrical contact with the cooler, which avoids the necessity of using deionized water.

T-Bar construction also delivers better cooling characteristics than single emitter diode lasers. This is because single emitters are typically supplied already packaged. Since the diodes are mounted within the package, there is an extra thermal interface which makes it more difficult to efficiently cool them.

In conclusion, after decades of imagining, offensive laser weapons are nearly ready for deployment on the battlefield. Advances in diode laser pumping technology are proving to be a key enabling factor in this development. And, these military grade pump diode lasers can be constructed using the same automated, high volume fabrication equipment used for industrial product manufacturing, meaning they can be readily delivered at competitive prices.

This article was written by Chris Ebert, Ph.D., Director, Business Development, Coherent | DILAS, (Tucson, AZ). For more information, visit ***http://info.hotims.com/65852-501.***

The Tom Bearden Website

Table of Contents

A Sensitive Scalar Wave Detector

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11. Figure 11. NETWORK OF VIRTUAL TRANSMITTERS
12. Figure 12. VIRTUAL TRANSMITTERS IN THE INTERFERENCE GRID
13. Figure 13. GIANT RADIAL RELATED TO A VIRTUAL TRANSMITTER
14. Figure 14. FORMATION OF A SINGLE GIANT RADIAL
15. Figure 15. FORMATION OF A DOUBLE GIANT RADIAL
16. Figure 16. DETECTION OF TRANSVERSE AND LONGITUDINAL WAVES

Return to Homepage

from **STAR WARS NOW!**

The Bohm-Aharonov Effect, Scalar Interferometry, and Soviet Weaponization
by Lt. Col. T.E. Bearden (retd.), 1984
Copyright

A Sensitive Scalar Wave Detector

Very briefly we present a method of making a very sensitive scalar wave detector so that direct measurement and quantization can be established.

First, we regard one oscillation (one wavelength) of the scalar potential wave as a longitudinal photon. Note that this photon contains a substructure, which may be deliberately determined (when artificially made) or randomized (when naturally made in the idealized case).

For convenience we represent normal linearized vacuum (spacetime) as a horizontal or longitudinal vector (Figure 4), implying the direction of motion of the wave in the laboratory frame. By horizontal position in our diagram, we imply a linear, uncurved spacetime and a non-rotated frame. We represent the longitudinal scalar wave as a horizontal vector, and the usual Hertzian wave as a vertical or "transverse" vector. We visualize a normal detector as detecting only a vertical or "transverse" vector, as we have illustrated in Figure 4.

As can be seen, in a linear, unrotated or uncurved spacetime a pure scalar wave has no vertical component projected upon the laboratory frame vector, so it is not detectable by normal detectors.

To detect the scalar wave, of course we could bend it so that it has a projected vertical component in the laboratory frame (Figure 5). However, this would be an impure wave, not a pure scalar wave, and that is not what we wish.

A better way is to bend or curve spacetime itself in a small region, so that a longitudinal wave that passes through that region now possesses a vertical component with respect to that region (Figure 6). Thus a normal detector there will detect that vertical component. We conduct the detection current out of the "bent spacetime" region to an outside (normal) detector, and we then have a scalar wave detector.

To illustrate, we show conceptually how this has been successfully done. Figure 7 shows the concept. First, we utilize a magnetic pole to provide the infolded energy (potential) to bend or curve spacetime. To reach good sensitivity, we need a pole strength connected with a magnetic field strength of 40,000 Gauss or higher. We utilize a small superconducting magnet, which can reach field strengths of from 40,000 to 80,000 Gauss.

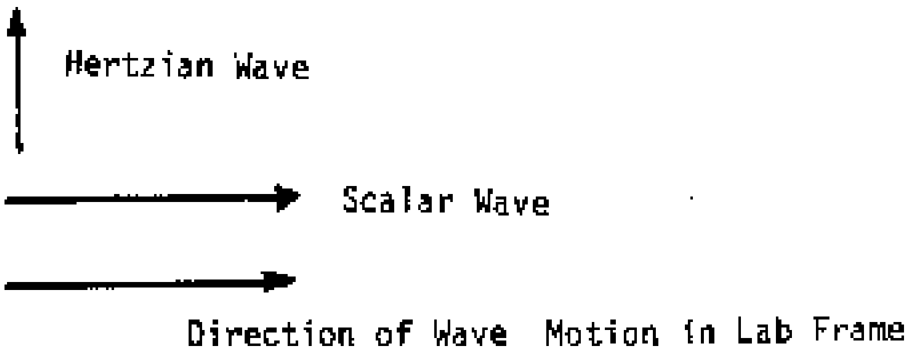
Hertzian Wave
Scalar Wave
Direction of Wave Motion in Lab Frame

Figure 4. Linear, uncurved laboratory frame.

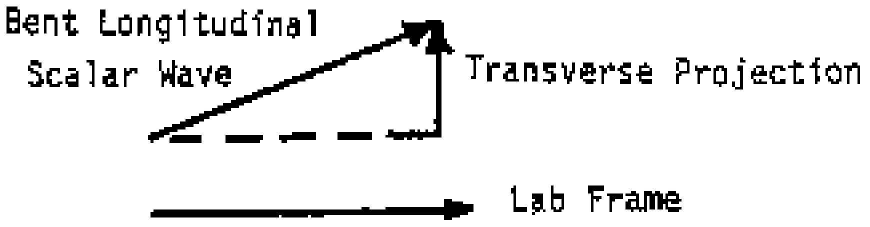
Bent Longitudinal Scalar Wave
Transverse Projection
Lab Frame

Figure 5. Rotating the longitudinal wave produces a transverse component.

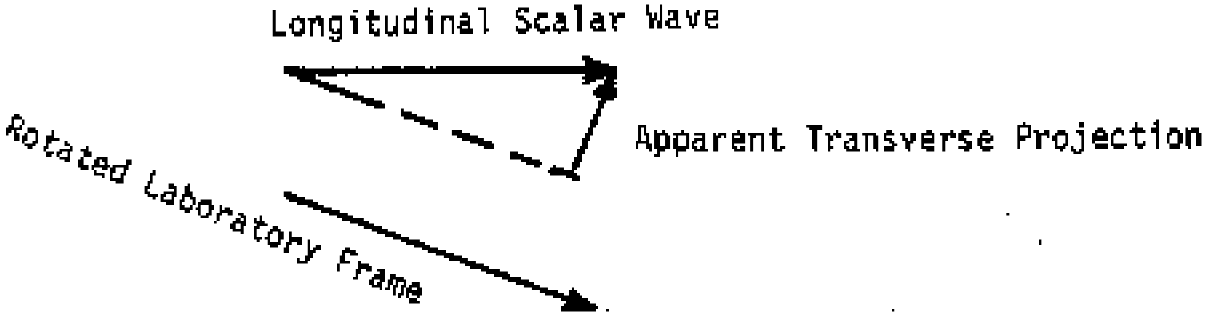
Longitudinal Scalar Wave
Apparent Transverse Projection
Rotated Laboratory Frame

Figure 6. Bending the laboratory spacetime frame produces a transverse component.

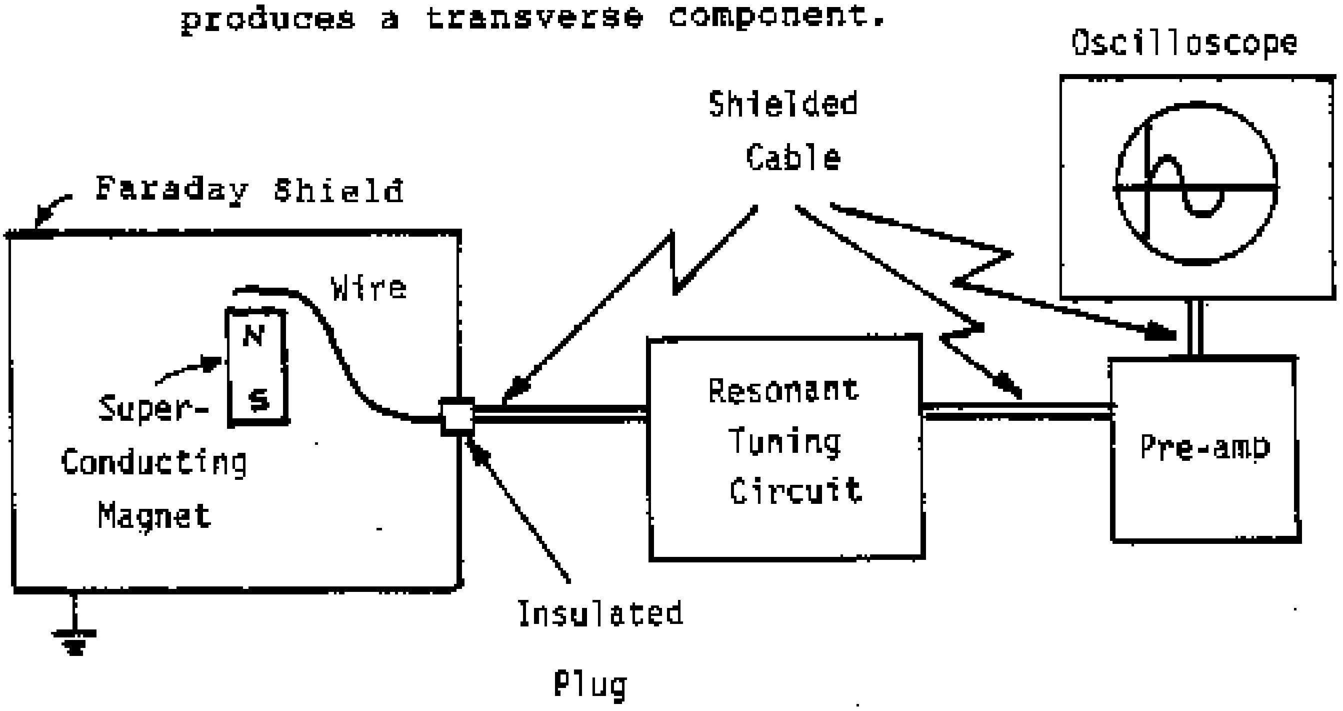
Oscilloscope
Shielded Cable
Faraday Shield
Wire
N
S
Super-Conducting Magnet
Resonant Tuning Circuit
Pre-amp
Insulated Plug

Figure 7. A sensitive scalar wave detector.

Figure 7. A sensitive scalar wave detector.

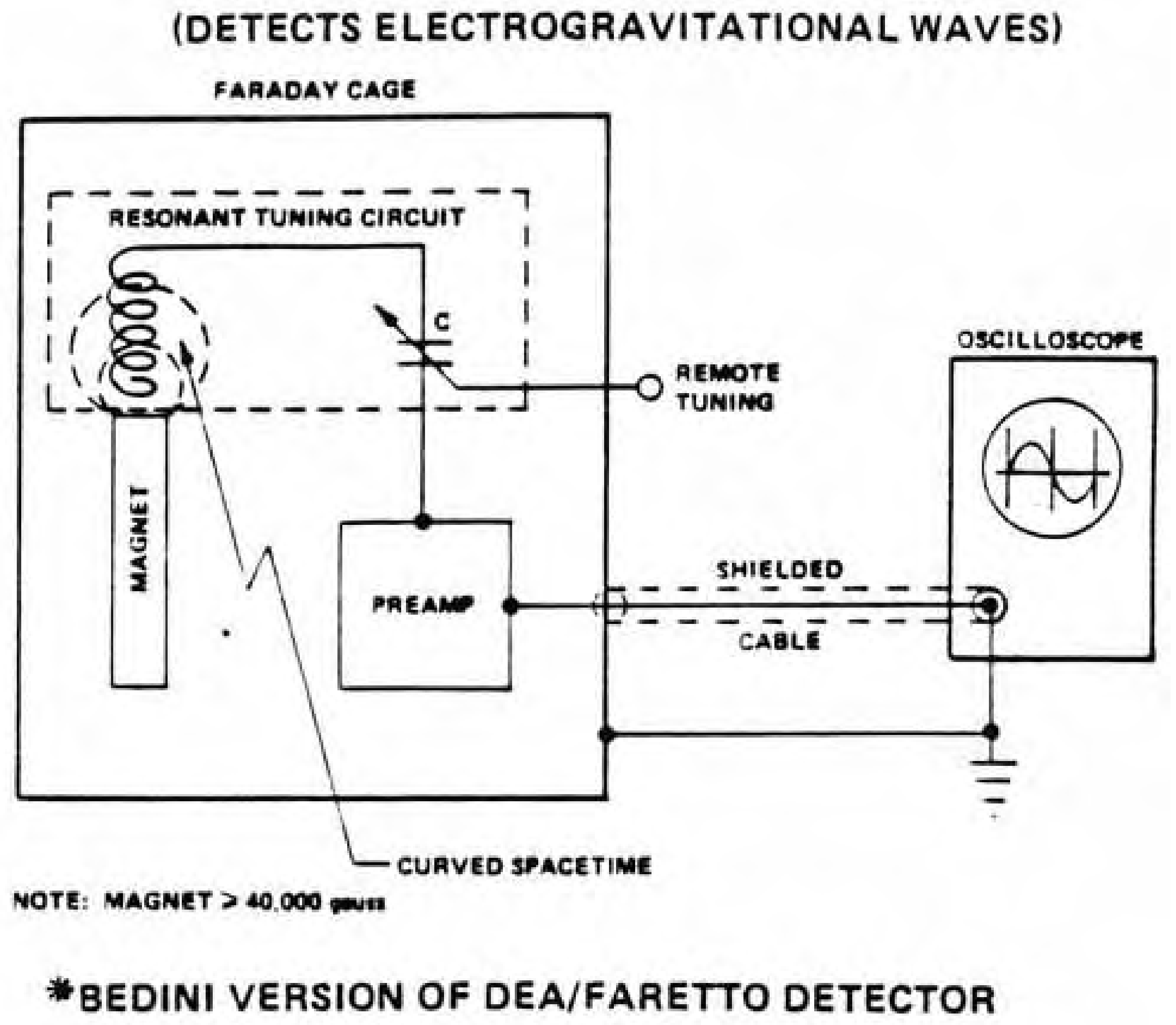
(DETECTS ELECTROGRAVITATIONAL WAVES)
FARADAY CAGE
RESONANT TUNING CIRCUIT
C
MAGNET
REMOTE
TUNING
OSCILLOSCOPE
PREAMP
SHIELDED
CABLE
CURVED SPACETIME
NOTE: MAGNET > 40,000 gauss
*BEDINI VERSION OF DEA/FARETTO DETECTOR

Slide 33 - FER DE LANCE

Then to prevent detection of superfluous "normal" radiation, we shield the entire magnet in a grounded Faraday cage, as shown in the figure. Hertzian waves will be grounded in this shield, while scalar waves will readily penetrate it undiminished. In other words, the Faraday cage serves as a "stripper," to strip away the ordinary waves, leaving only the scalar waves to penetrate inside. An ordinary wire lies at the top of the magnetic pole, in proximity to it. The wire runs out of the Faraday cage through an insulated port to a resonant tuning circuit, which is sharply tunable over the range of frequencies we are interested in. A preamp amplifies the output of the tuner, and in turn feeds the input of an oscilloscope or other detector.

With this detector we can select the frequency desired, and detect any passing scalar waves of that frequency. By other variations of amps and preamps, regenerative circuits, etc. we can obtain all the sensitivity desired, and utilize ordinary detection equipment already well-known and highly developed.

Additional Information about detecting Scalar Waves

The Tom Bearden Website

Figure 11. NETWORK OF VIRTUAL TRANSMITTERS
(See Woodpecker Grid)

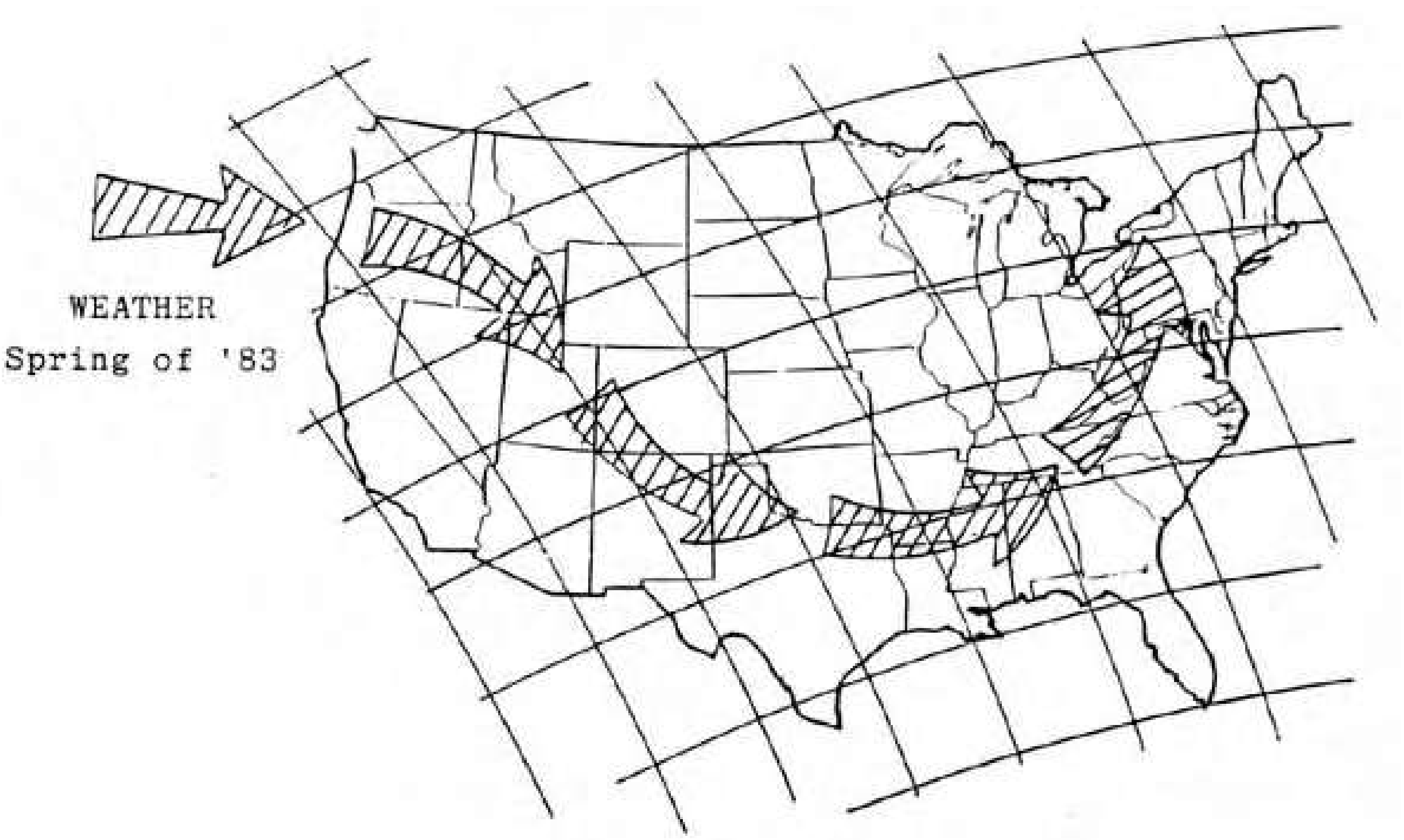

Next Figure

[The Tom Bearden Website](#)

Figure 12. VIRTUAL TRANSMITTERS IN THE INTERFERENCE GRID

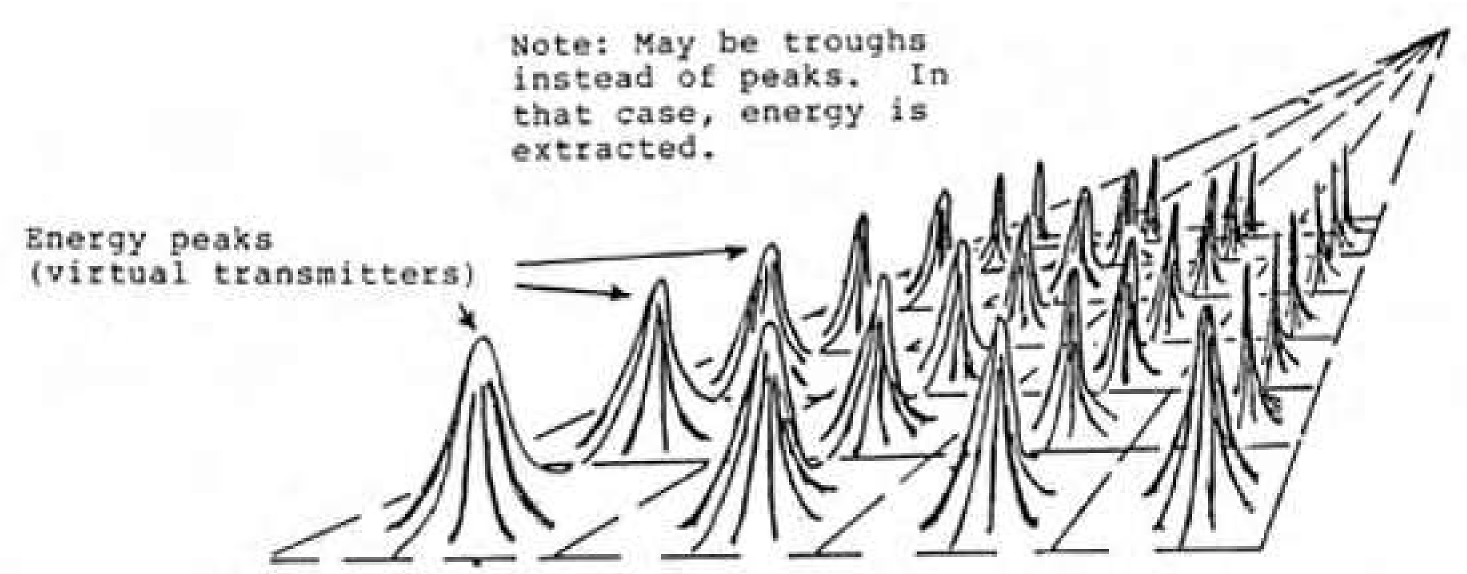
Note: May be troughs instead of peaks. In that case, energy is extracted.
Energy peaks (virtual transmitters)

[**Next Figure**](#)

[**Previous Figure**](#)

The Tom Bearden
Website

Figure 13. GIANT RADIAL RELATED TO A VIRTUAL TRANSMITTER

**Fall 1983:
Florida
Alabama**

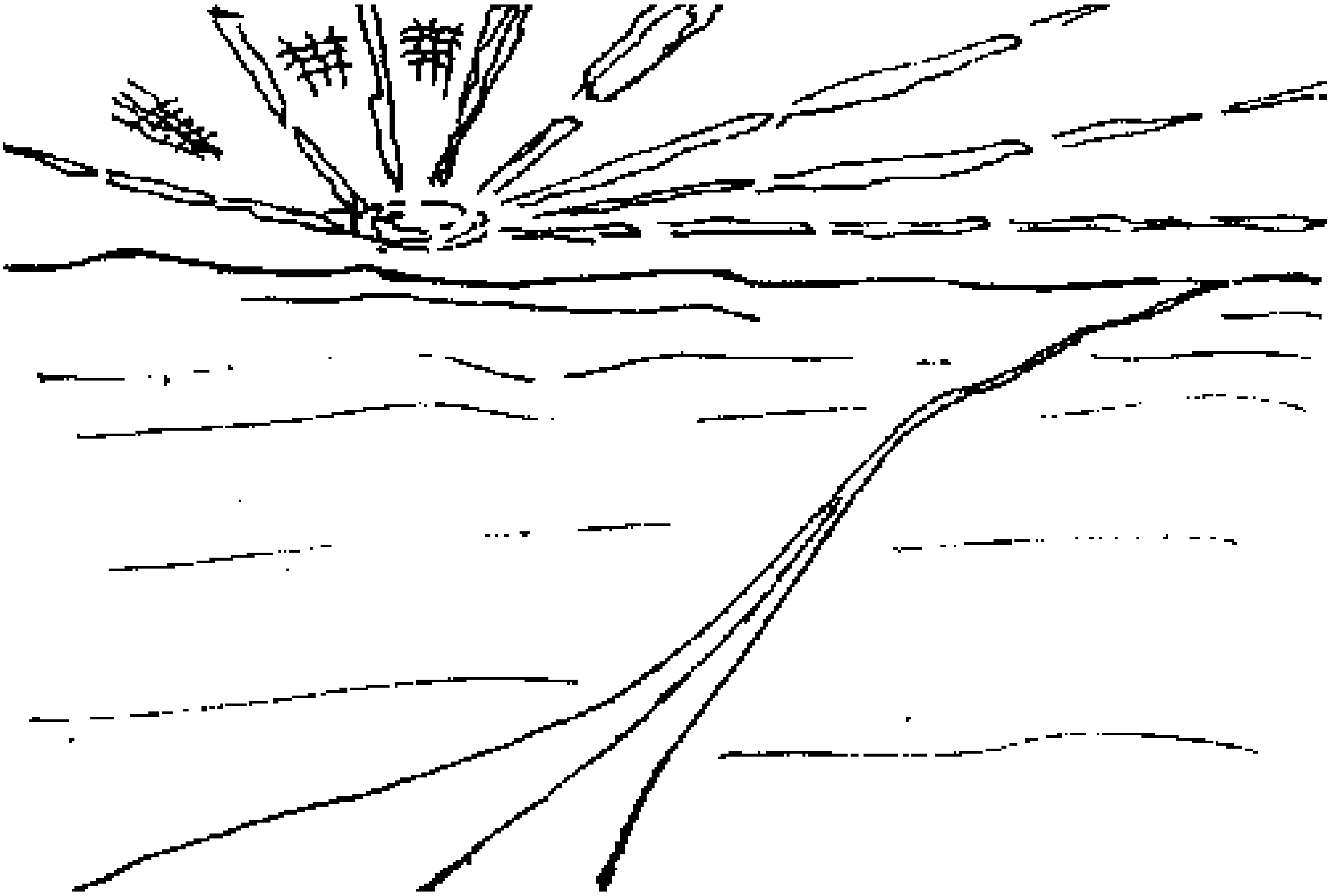

Next Figure

Previous Figure

[The Tom Bearden Website](#)

Figure 14. FORMATION OF A SINGLE GIANT RADIAL

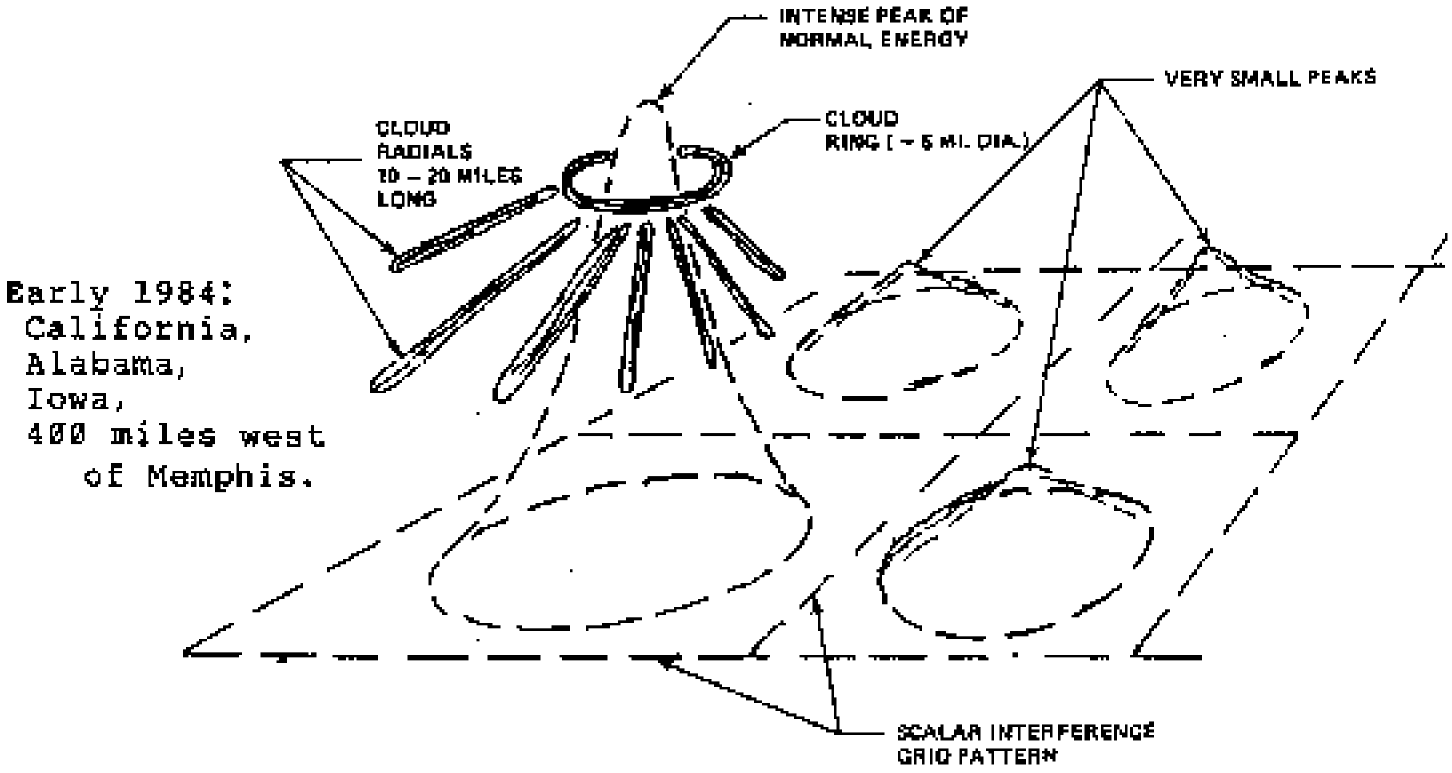
INTENSE PEAK OF
NORMAL ENERGY
VERY SMALL PEAKS
CLOUD
RADIALS
10 – 20 MILES
LONG
CLOUD
RING (~5 MI. DIA.)
Early 1984:
California,
Alabama,
Iowa,
400 miles west
of Memphis.
SCALAR INTERFERENCE
GRID PATTERN

Excerpt P 172
FER DE LANCE

A Briefing on Soviet Scalar Electromagnetic Weapons
by Lt. Col. T.E. Bearden (retd.), 1986
Copyright

2. Continual cloud radials have repeatedly been seen and photographed over Huntsville, Alabama, associated with weather engineering and grid activity. Typical radials seen by this author are shown in Bearden, Soviet Weather Engineering Over North America, 1-hr. v1deotape, 1985. Huntsville is apparently a pivot point or "hinge point" for bending and controlling the jetstream.

3. Spectacular cloud grid patterns have been seen over Huntsville, Alabama and Los Angeles, California. An excellent videoclip of one of the Los Angeles area grids, personally observed and shot by KABC Open Mind talk show host Bill Jenkins, is shown in Bearden, Soviet Weather Engineering Over North America,;1-hr. videotape, 1985. An artist's sketch of an earlier gigantic grid, extending from horizon to horizon in every direction, seen over Huntsville, Alabama by Tom Bearden and Ken Moore, is also shown in the tape. The videotape is available from P.O. Box 1472, Huntsville, AL 35807. Sketches and details of another highly anomalous grid pattern observed in the greater Los Angeles area by engineer Ron Cole are shown in figures 5 and 6.

The reason that clouds of water droplets and/or ice particles detect the scalar interferometry and form signature patterns is simple: Consider each H2O molecule as having two light little hydrogen atoms hanging on to the much heavier oxygen atom. The covalent sharing with the oxygen atom of the electron from each hydrogen atom means that

1. the electrostatic scalar potential between the H ion and the O ion it is bonded to, is rhythmically varying as the electron is shared back and forth,
2. this varying potential contains "electron spin holes" since it is made by the moving electron,
3. two such varying potentials exist since there are two H atoms sharing covalent electrons with the O atom,
4. the two H ions are at an angle of over 100 degrees with respect to each other,
5. the molecule assembly thus constitutes one part (one half, so to speak) of a scalar interferometer with imbedded electron spin holes for electron hooking,
6. Incident scalar waves from outside the system interact with the "half scalar interferometer." This scalar interferometry interaction is coupled to the covalent bonding electron because of the internal spin hole pattern of the molecular half of the interferometer. The coupled electron moves with the sca1ar pattern's changes, causing an observable interaction with and in the electrical structure and ionic potential of the molecule,
7. the weak H to H hydrogen bonding between molecules connects resulting ionic potential changes to the structuring of the entire macroscopic cloud assembly. Thus by scalar interferometry with the H_2O "half interferometers", an incident scalar EM pattern is detected and translated into patterns of E and B force fields. The H_2O molecules then "line up" in accord with the electrical patterns detected.

Thus the clouds form "signature patterns" according to the incident scalar EM radiation patterns. Giant cloud radials in the targeted area are probably due to the type and shape of antennas including the type and shape of the electrical wires establishing the antennas' ground planes -- used by the Soviet Union to project the scalar waves. Note that such "radial" antenna patterns were occasionally used by James Harris Rogers in his underground and undersea scalar EM transmission system.

Next Figure

Previous Figure

The Tom Bearden Website

Figure 15. FORMATION OF A DOUBLE GIANT RADIAL

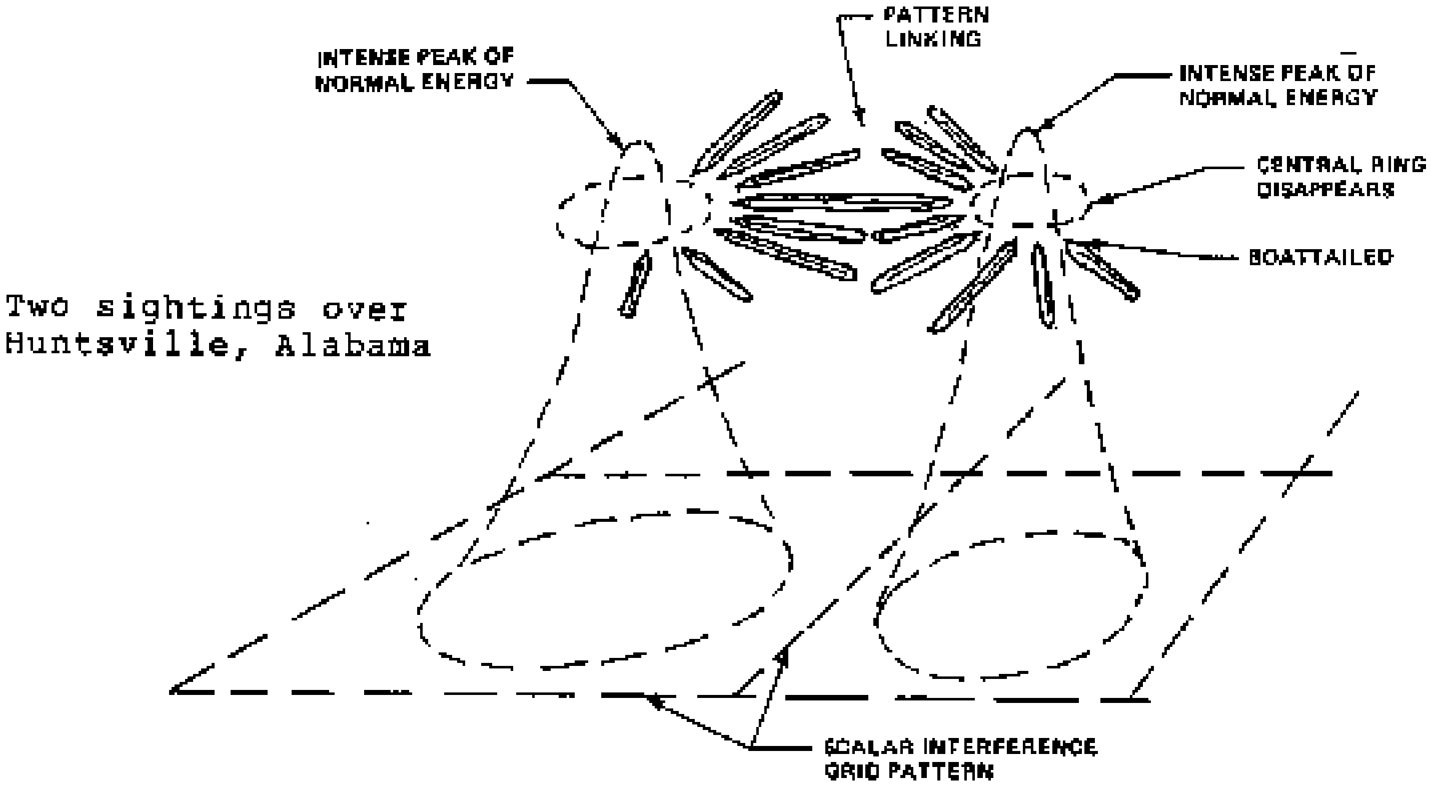
PATTERN LINKING
INTENSE PEAK OF NORMAL ENERGY
INTENSE PEAK OF NORMAL ENERGY
CENTRAL RING DISAPPEARS
BOATTAILED
Two sightings over Huntsville, Alabama
SCALAR INTERFERENCE GRID PATTERN

Next Figure

Previous Figure

The Tom Bearden Website

Figure 16. DETECTION OF TRANSVERSE AND LONGITUDINAL WAVES

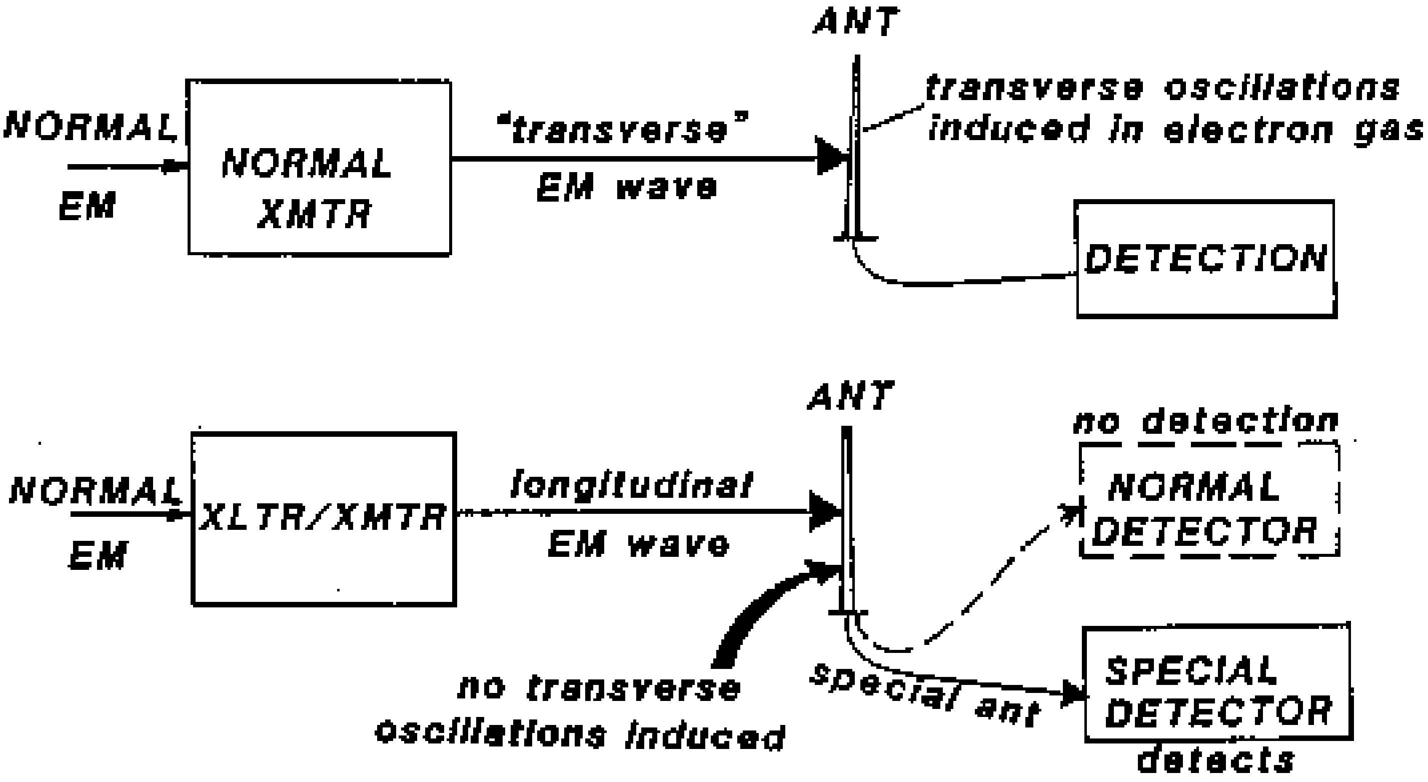
ANT
NORMAL
EM
NORMAL XMTR
"transverse" EM wave
transverse oscillations induced in electron gas
DETECTION
ANT
NORMAL
EM
XLTR/XMTR
longitudinal EM wave
no detection
NORMAL DETECTOR
no transverse oscillations induced
special ant
SPECIAL DETECTOR
detects

Next Figure

Previous Figure

Defense Department Briefing on Priore Treatment to counter Biological Warfare

---

SLIDE LIST

Solutions to Mass Casualties Treatment Problem

Threat and Current Status

1. Title Page
2. Key Points in the New Technology
3. Title Page
4. Title Page
5. The Terrorist Threat
6. The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
7. Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.
8. Present Status of the WMD Threat to the United States
9. Some Present Problems in the WMD Defense Capabilities

Historical Solution: Priore Treatment Therapy

10. Title Page
11. Priore's laboratory device for treating small animals
12. Antoine Priore with the Mayor of Bordeaux
13. Raymond Pautrizel, eminent French parasitologist, who worked with Priore
14. 12-Foot plasma tube in Priore's large device
15. Priore's 3 1/2 stories high device for whole body treatment of humans
16. Mechanism of the Priore therapy
17. Priore's therapeutic methodology: using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain
18. Drastically reduced cumulative deterioration after Priore therapy

Extended Physics and Background

19. Title Page
20. Founders of Scalar Electromagnetics
21. Subdivisions of Soviet Energetics

22. Bunge on the Status of Classical and Quantum Physics
23. Einstein on Reviewing Foundations
24. Fields and Potentials have been "defined" only in and on matter
25. Maxwell's Equations are Matter-To-Matter Transforms
26. Electrodynamicists' Reaction to Removal of the Material Ether
27. The Missing Infolded Electrodynamics
28. Nonlinear Optics Distortion Correction Theorem
29. Infolded longitudinal EM biwave composition of a scalar potential
30. The Zero-Vector Axiom Destroys "Topologies Within Topologies"
31. Stoney-Whittaker Bidirectional Longitudinal EM Wave-Pair
32. Undistorted Progressive Waves (UPWs)
33. Velocity modulating a longitudinal EM wave oscillates rate of flow of time
34. The Graviton is a Coupled Photon/Antiphoton Pair
35. Mechanism for Physical Change and the Flow of Time
36. Any EM Field or Wave Pattern is Produced by interference of Two Scalar Potential Functions (E.T. Whittaker, 1904)
37. Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass
38. Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection
39. The Immune System
40. Popp's master cellular control system
41. Cancer: characteristics
42. The evolutionary path of multicellular organisms
43. Effect of fluid contamination on hemoglobin's oxygen transport
44. First step in Self-Promotion of Cancer due to Hypoxia
45. Becker's Cellular Growth Control
46. Becker's dc treatment of bone fractures
47. Becker's theoretical DC control system involved with responses to injury
48. Becker's theoretical control system governing regeneration
49. Link Between Standing DC Potentials and the Nervous System
50. Becker's Findings on Cellular Injury and Regeneration
51. Corrections For the Present EM Bioeffects Model
52. R&R System - Regeneration Example
53. Kaznacheyev's Cytopathogenic Mirror Effect
54. Microwave Radiation of U.S. Embassy in Moscow

55. Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
56. Quantum potential and hidden variable theory
57. Participants in a quantum potential share a common multiply-connected spacetime (MCST)
58. Gulf War Syndrome and Typical Comparative Disease Curves
59. Force Fields and Symmetry
60. Aspects of Strong Local Asymmetry
61. Other Aspects of Strong Local Asymmetry
62. Small Entities Emit Powerful S-Flows
63. What is the "magnitude" of a potential?
64. Table A1. The Vacuum Energetically Interacts With Every Particle, Continually
65. Extended Superpotential Theory (Adding Vacuum Engines)
66. Vacuum Engine: Jillions of Working ST Demons
67. Active Spacetime and a Specific Vacuum Engine (A General Relativistic Concept and Analogy)
68. Extension of Wheeler's general relativity principle
69. Signal versus Vacuum Engine
70. Electronuclear Interactions by Vacuum Engines
71. Vacuum Engineering
72. Impact of Vacuum Engineering
73. A Dramatic Extension to Nonlinear Optical Pumping
74. Time Conservation and Some Properties (Energy and time are canonical)
75. A Dramatic Extension to Nonlinear Optical Pumping
76. Key Points in the New Technology

Modern WMD Solution: Portable Treatment Systems

77. Title Page
78. Portable Priore-Type Treatment System
79. Portable System Block Diagram
80. Portable Priore Technology Unit
81. Saving the Stricken Civilians

Conclusions: Technology Practical, Proven

82. Title Page
83. Summary and Conclusions

Recommendation: Crash Development Program

84. Title Page
85. Recommend implementing:
86. Final Thoughts

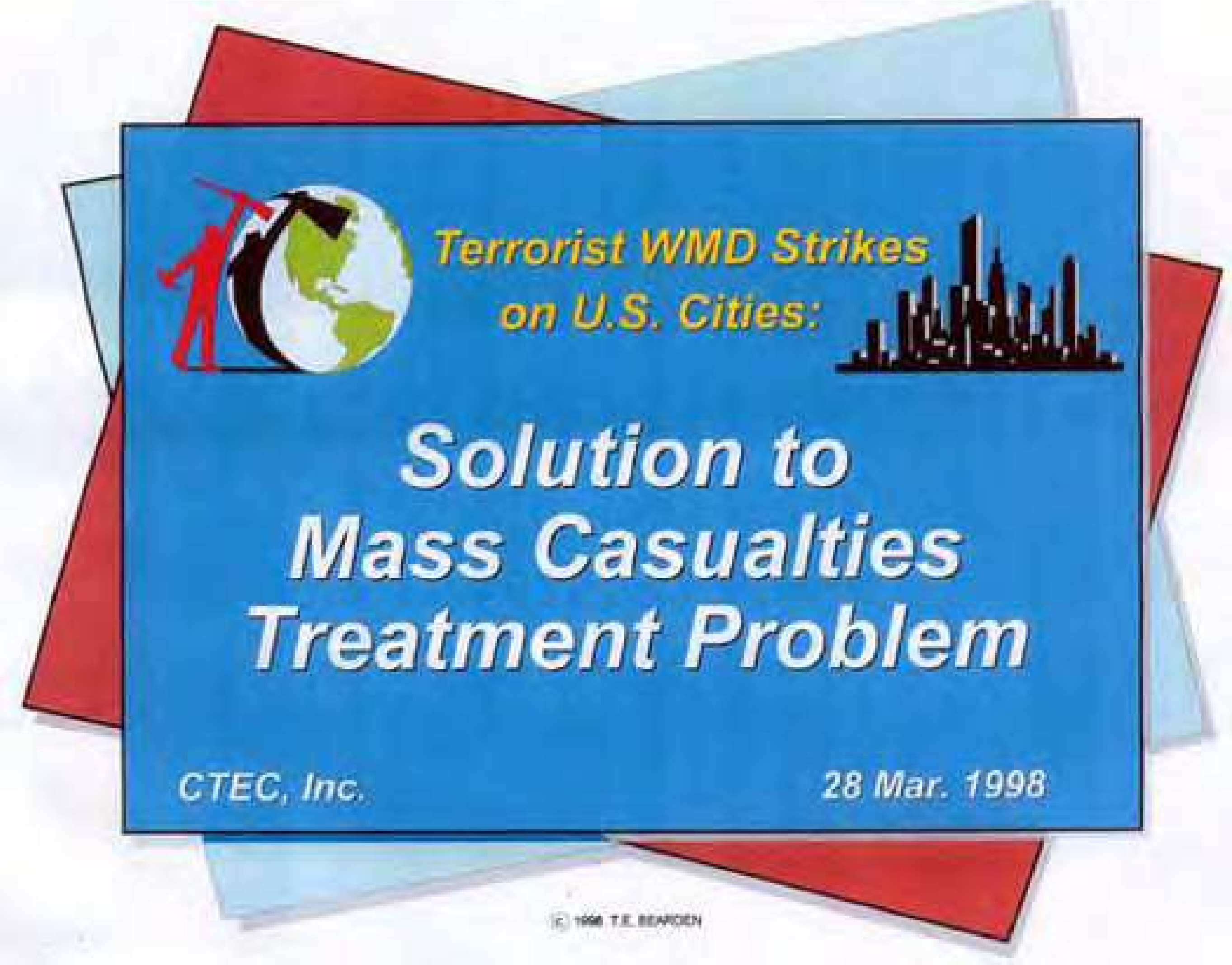

Terrorist WMD Strikes
on U.S. Cities:

Solution to Mass Casualties Treatment Problem

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

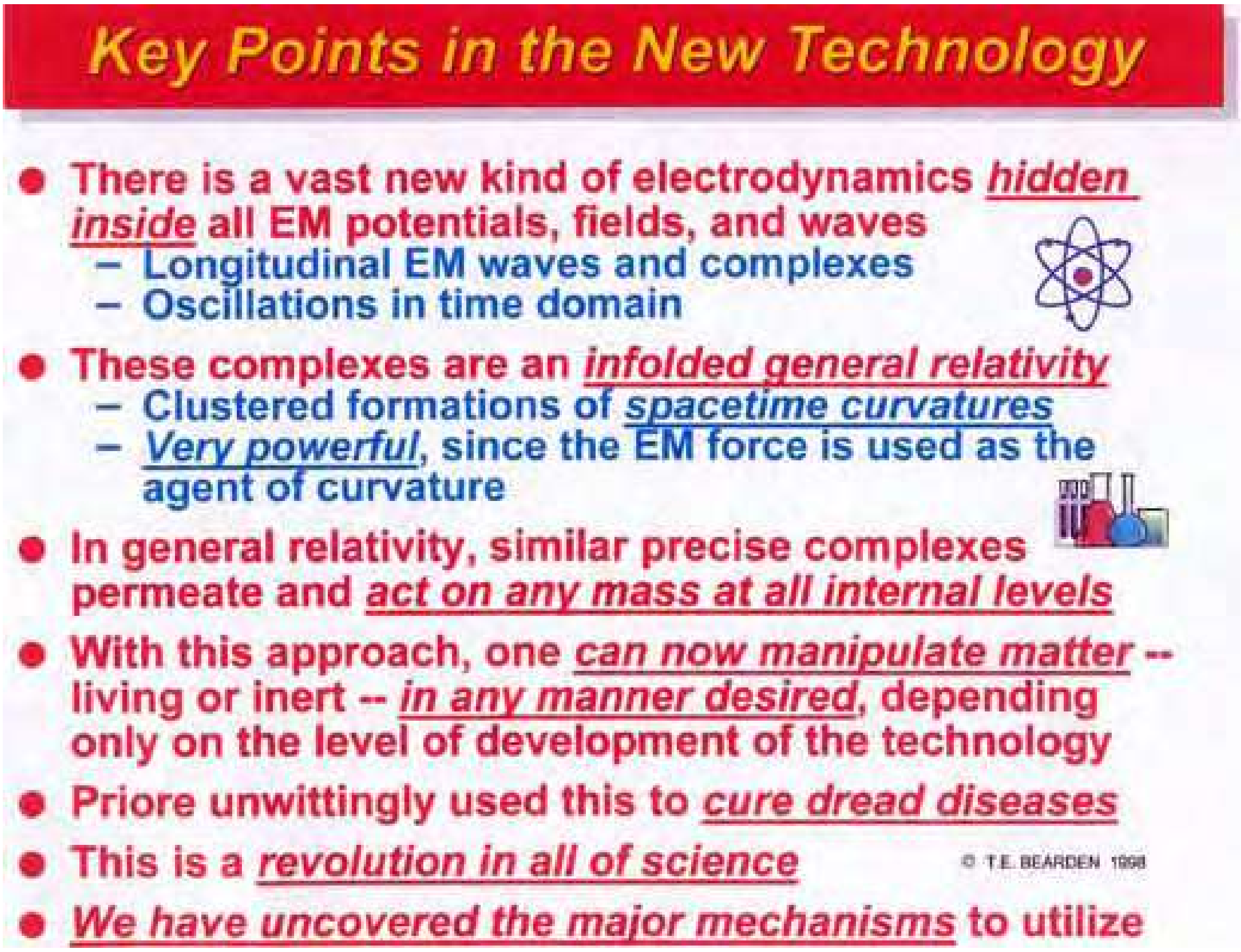

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

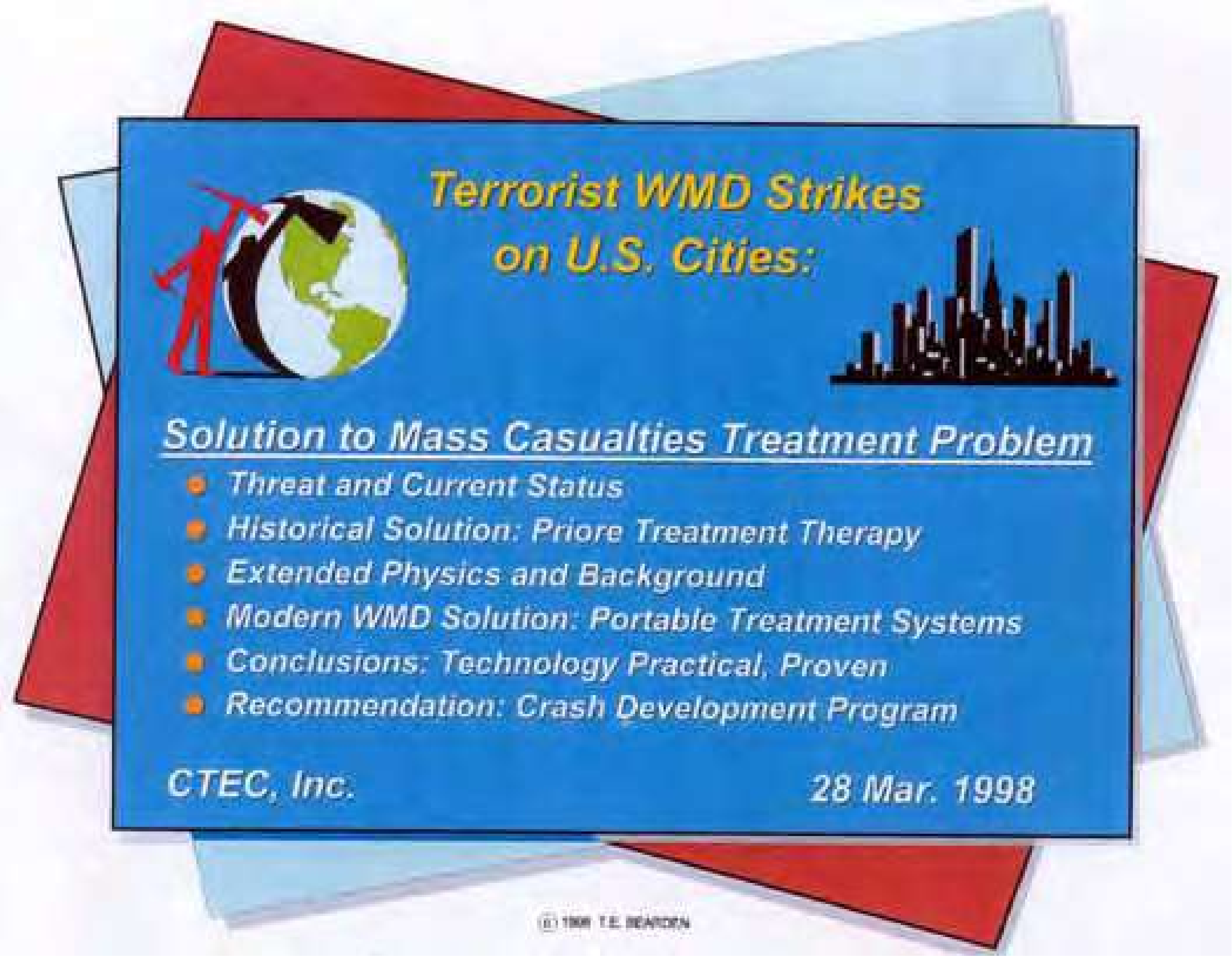

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

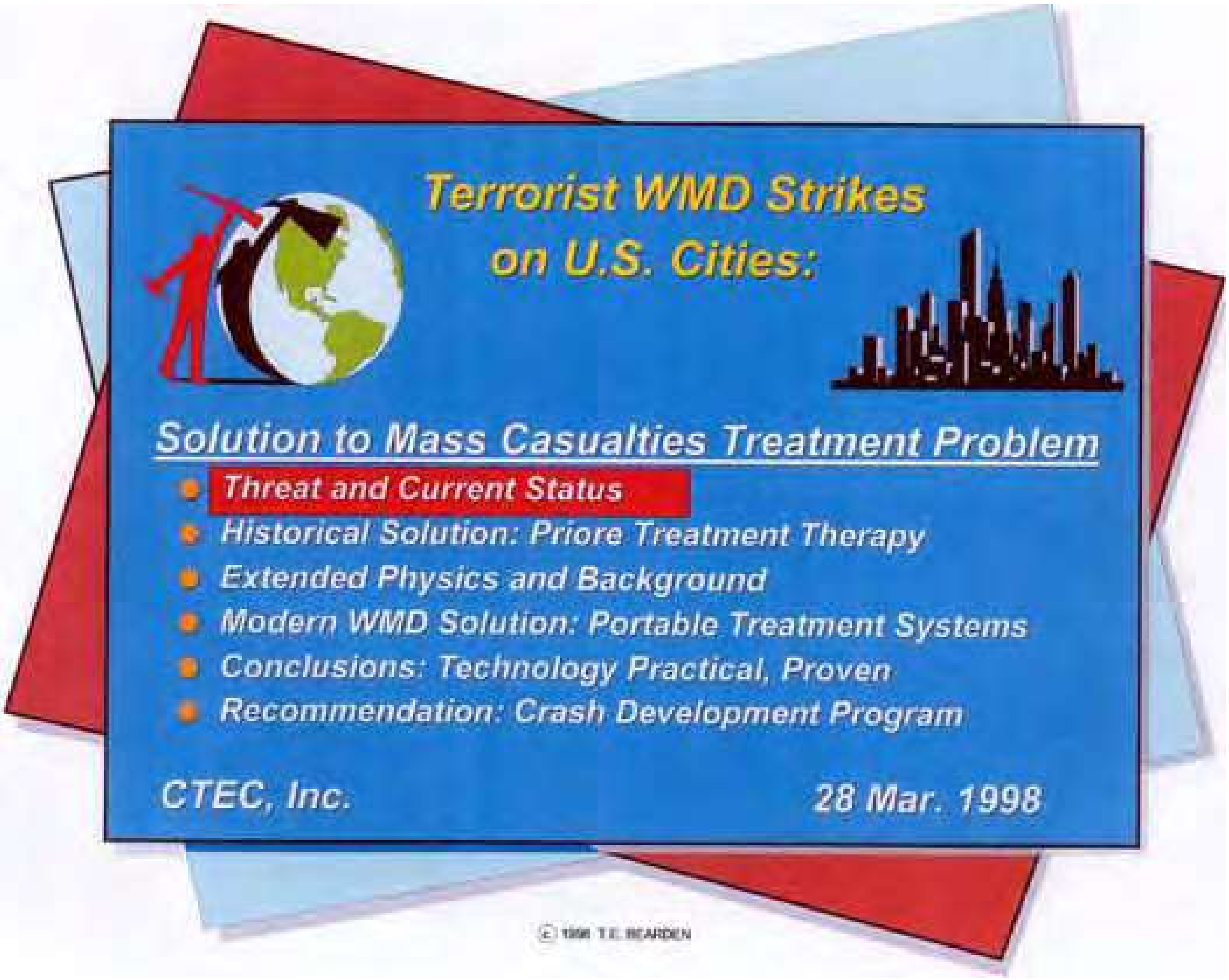

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

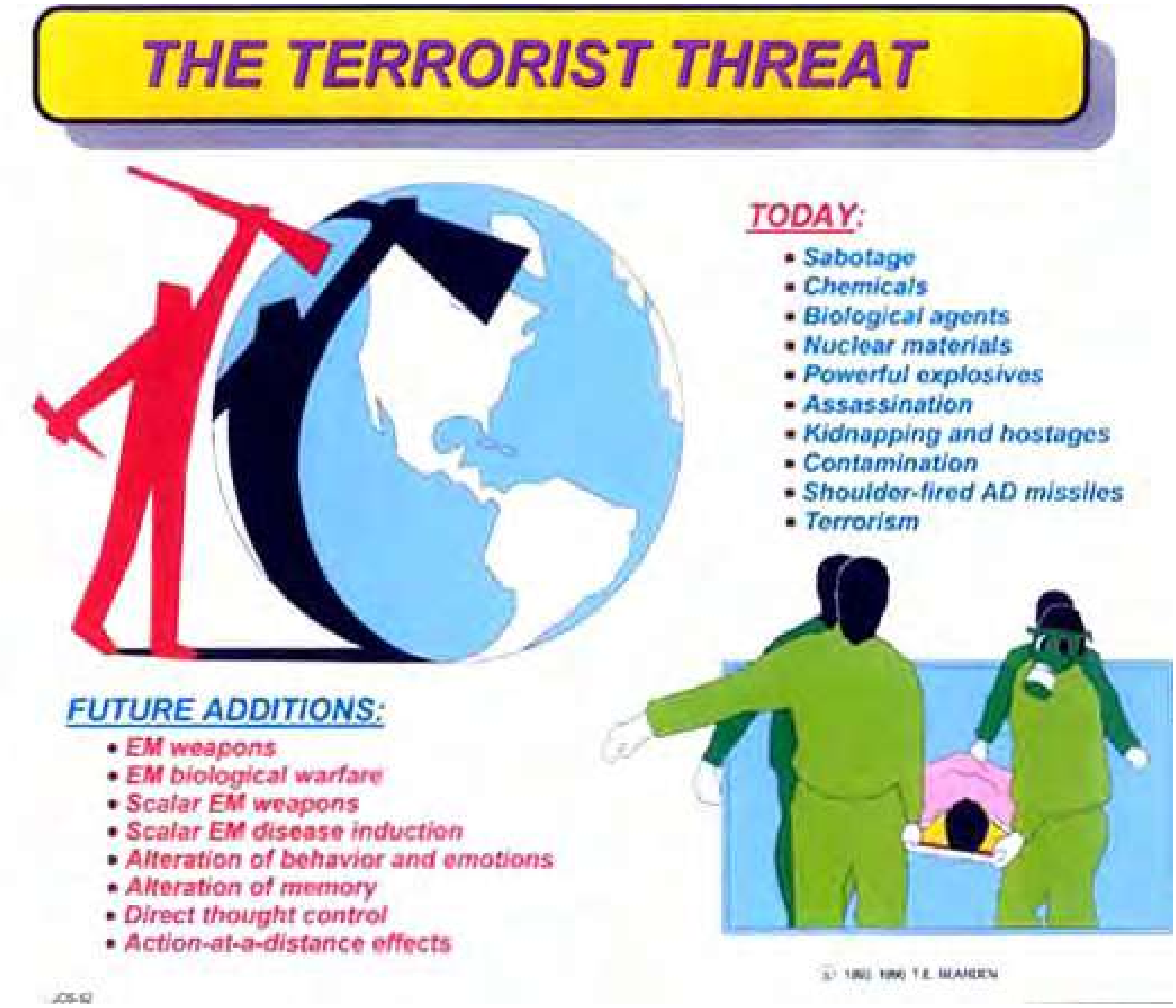

THE TERRORIST THREAT

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- EM weapons
- EM biological warfare
- Scalar EM weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct thought control
- Action-at-a-distance effects

Return to Slide Index

Next Slide

Previous Slide

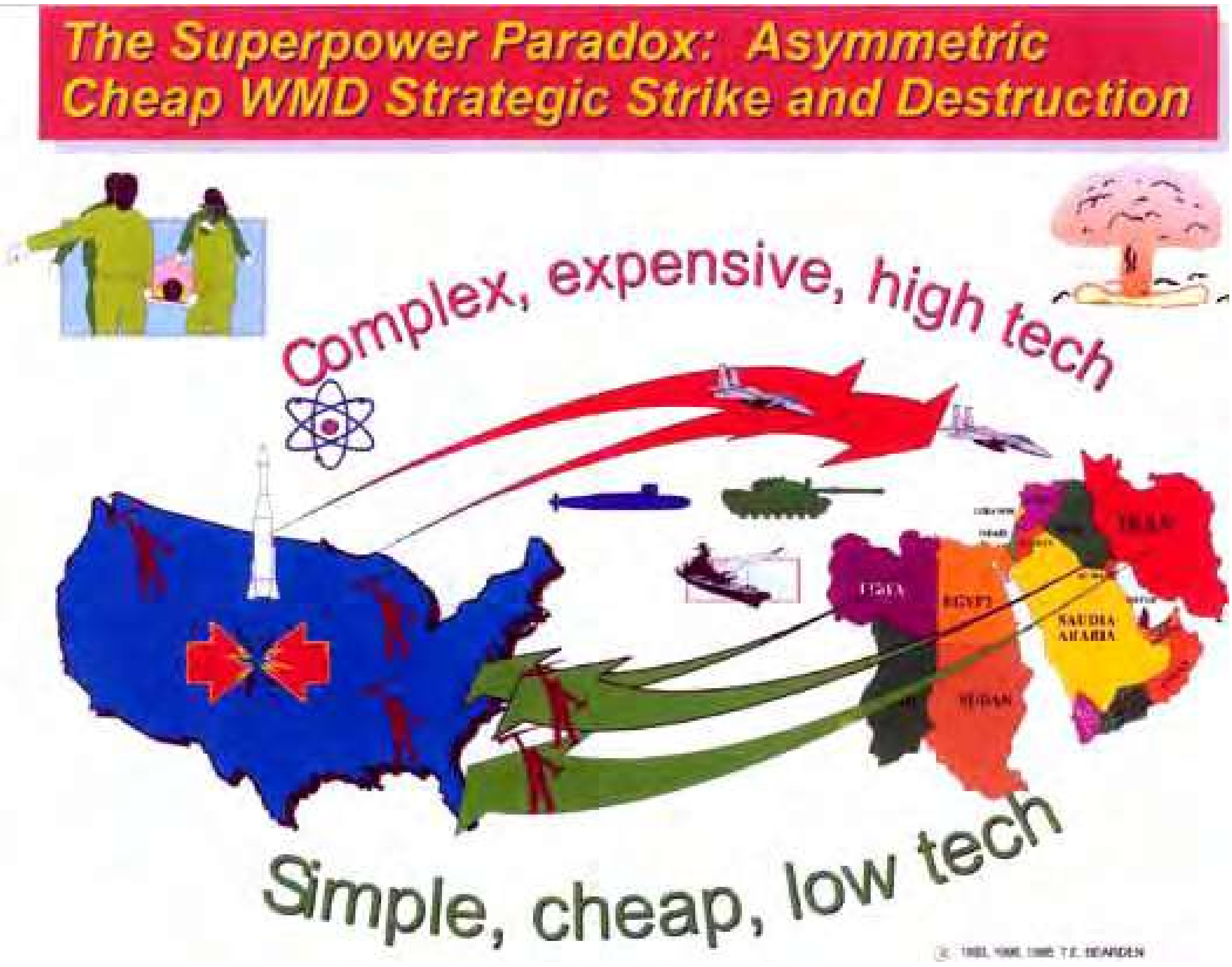

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

Simple, cheap, low tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN
ISRAEL

© 1993, 1996, 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

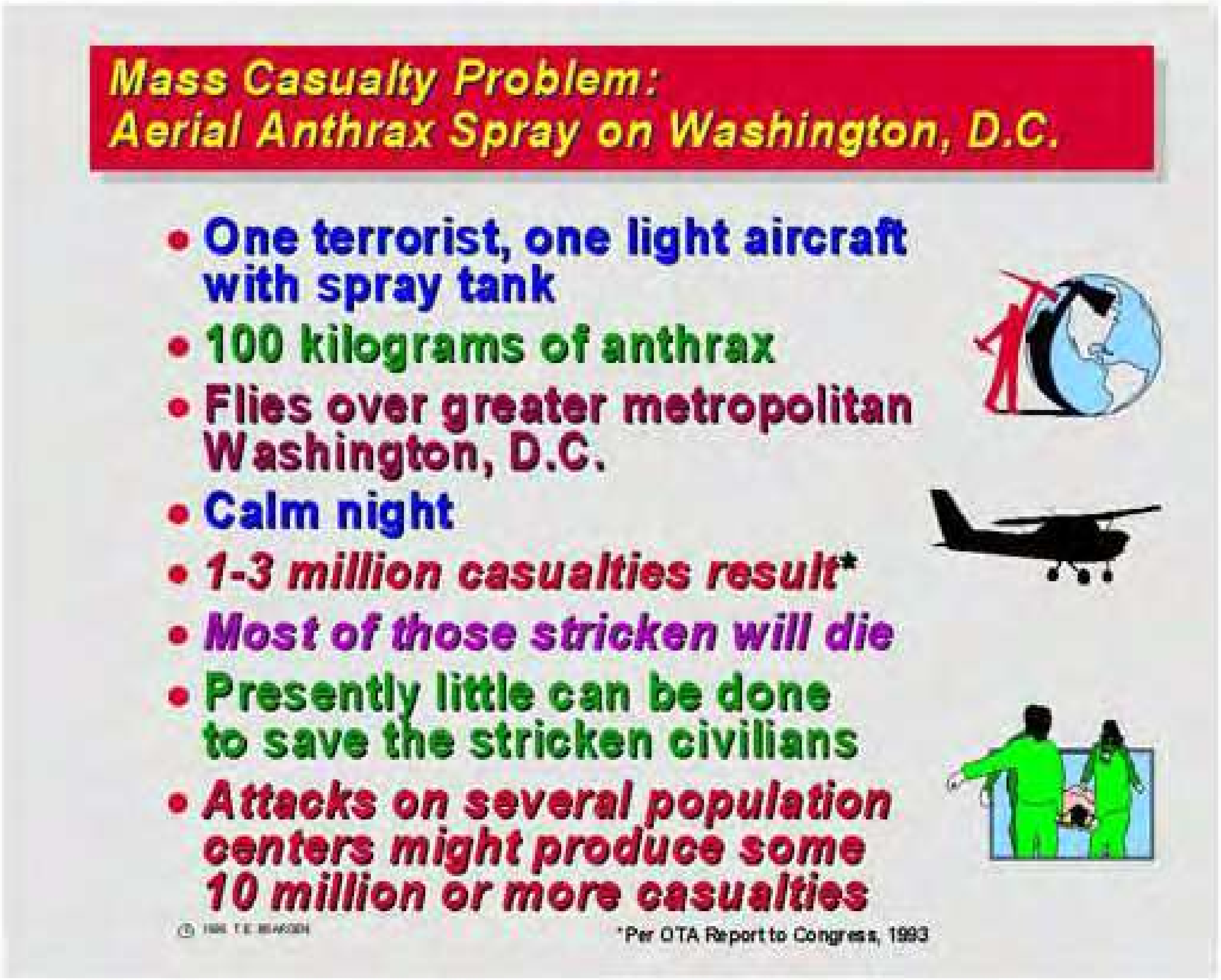

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- 100 kilograms of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

*Per OTA Report to Congress, 1993

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

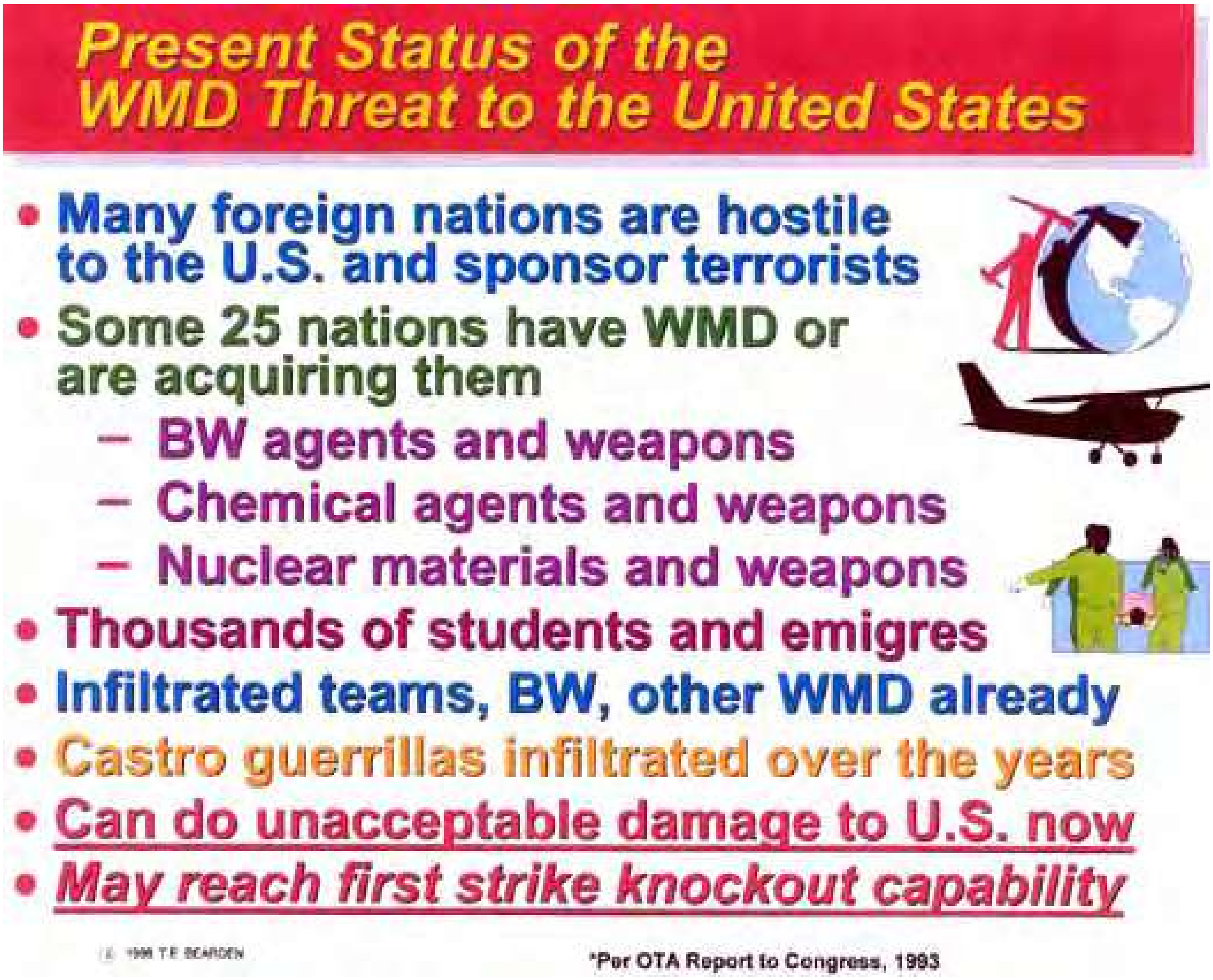

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
  - BW agents and weapons
  - Chemical agents and weapons
  - Nuclear materials and weapons
- Thousands of students and emigres
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© 1998 T.E. BEARDEN

*Per OTA Report to Congress, 1993

Return to Slide Index

Next Slide

Previous Slide

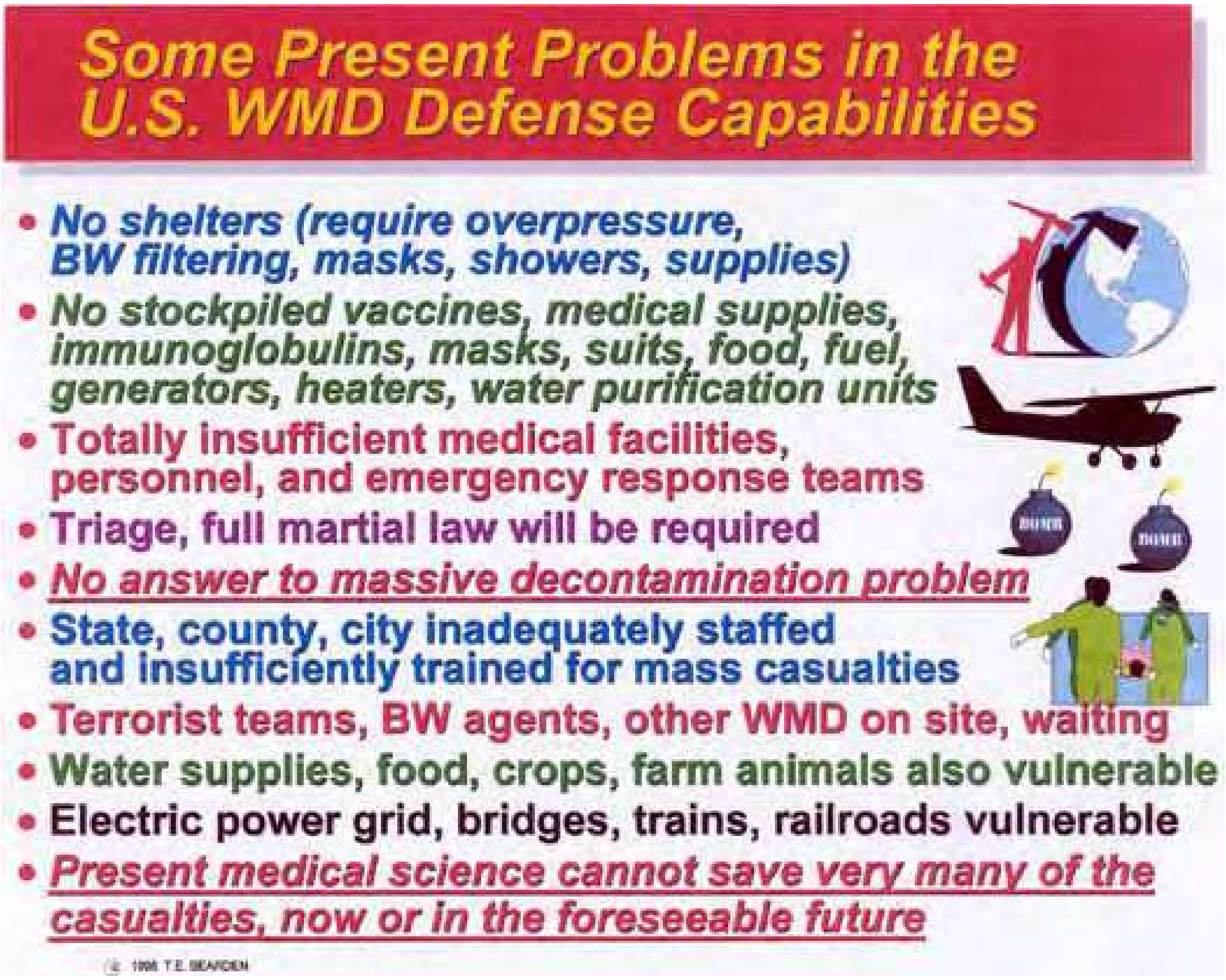

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

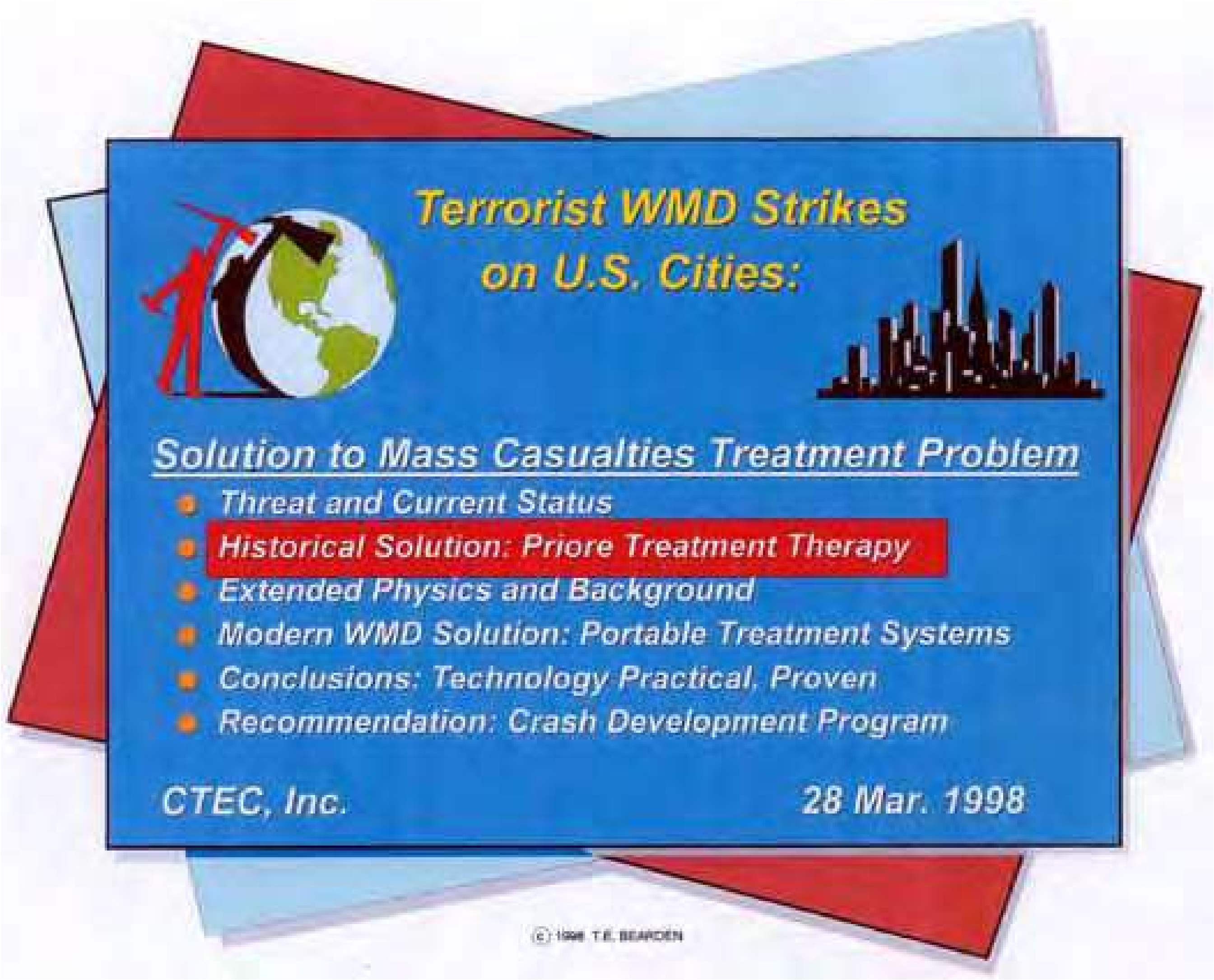

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

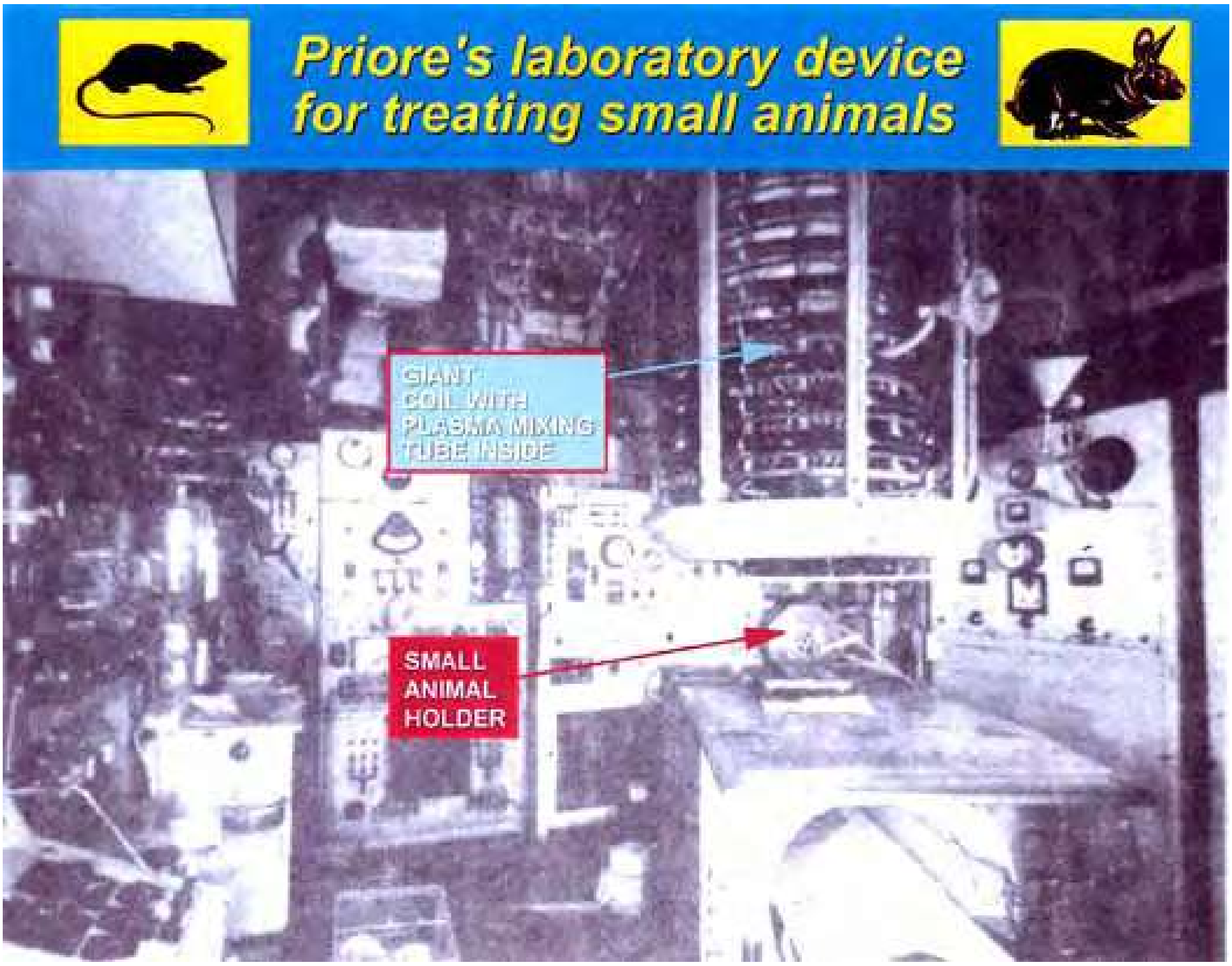

Priore's laboratory device for treating small animals

GIANT COIL WITH PLASMA MIXING TUBE INSIDE

SMALL ANIMAL HOLDER

Return to Slide Index

Next Slide

Previous Slide

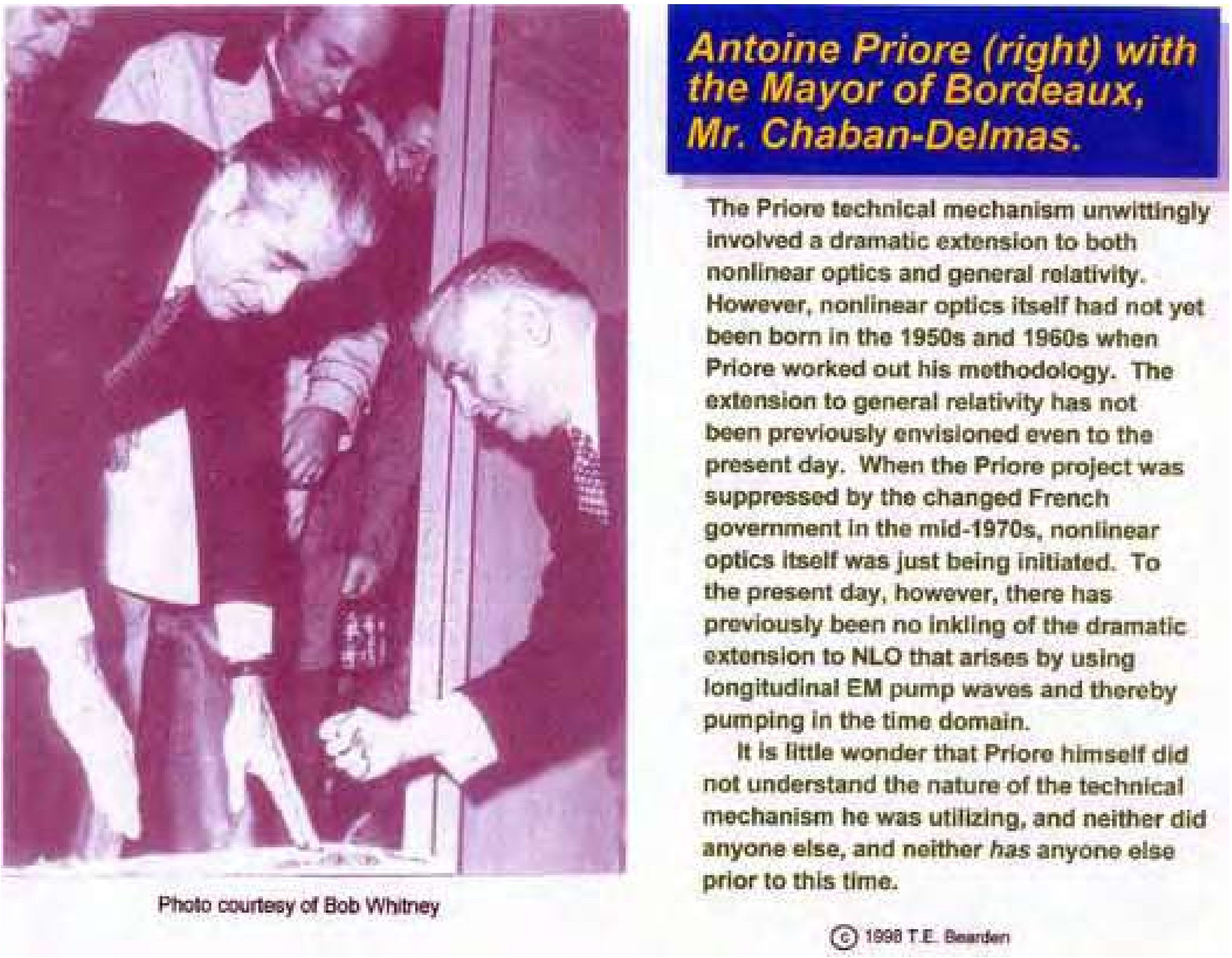

Antoine Priore (right) with the Mayor of Bordeaux, Mr. Chaban-Delmas.

Photo courtesy of Bob Whitney

The Priore technical mechanism unwittingly involved a dramatic extension to both nonlinear optics and general relativity. However, nonlinear optics itself had not yet been born in the 1950s and 1960s when Priore worked out his methodology. The extension to general relativity has not been previously envisioned even to the present day. When the Priore project was suppressed by the changed French government in the mid-1970s, nonlinear optics itself was just being initiated. To the present day, however, there has previously been no inkling of the dramatic extension to NLO that arises by using longitudinal EM pump waves and thereby pumping in the time domain.

It is little wonder that Priore himself did not understand the nature of the technical mechanism he was utilizing, and neither did anyone else, and neither *has* anyone else prior to this time.

© 1998 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

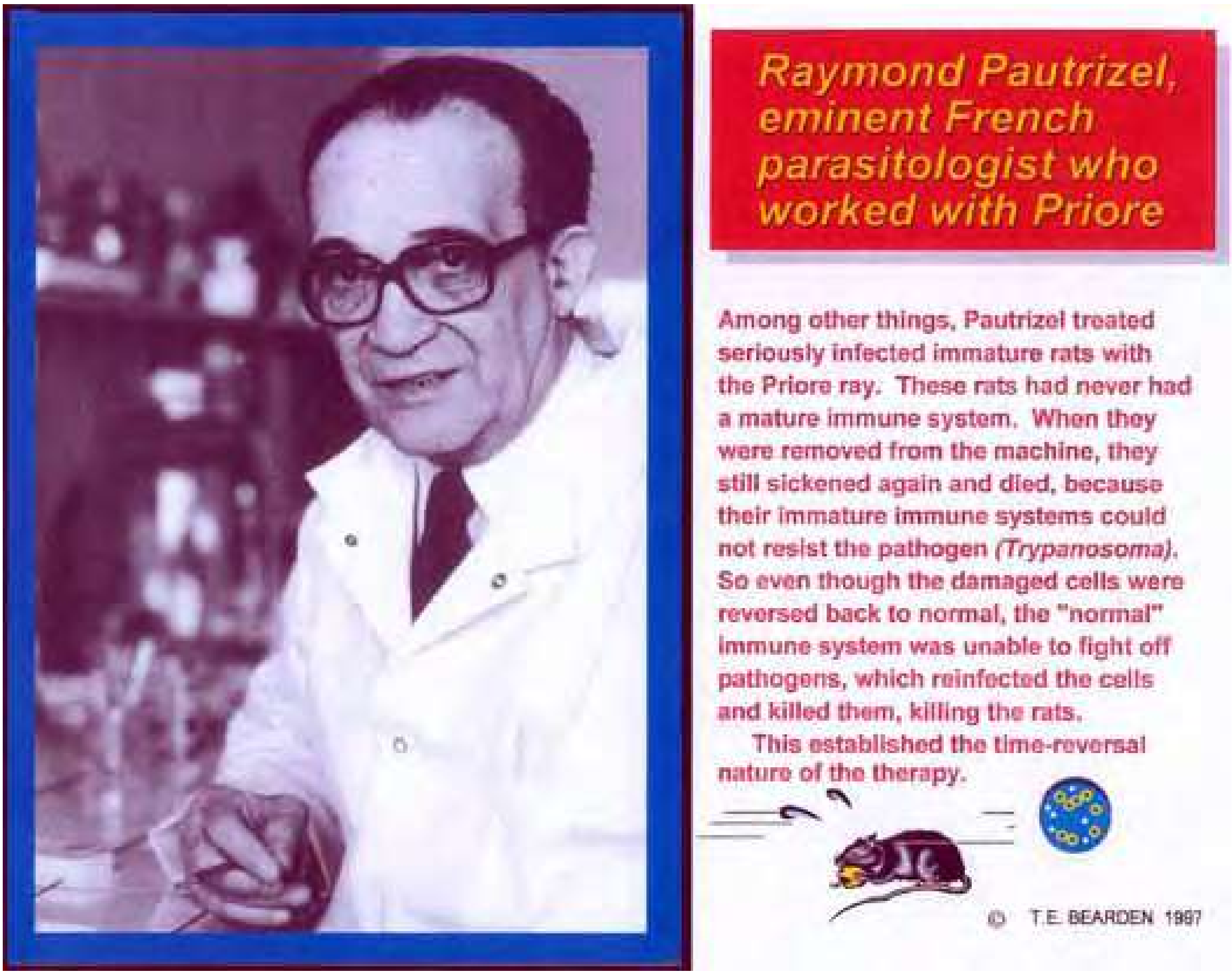

Raymond Pautrizel, eminent French parasitologist who worked with Priore

Among other things, Pautrizel treated seriously infected immature rats with the Priore ray. These rats had never had a mature immune system. When they were removed from the machine, they still sickened again and died, because their immature immune systems could not resist the pathogen *(Trypanosoma)*. So even though the damaged cells were reversed back to normal, the "normal" immune system was unable to fight off pathogens, which reinfected the cells and killed them, killing the rats.

This established the time-reversal nature of the therapy.

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

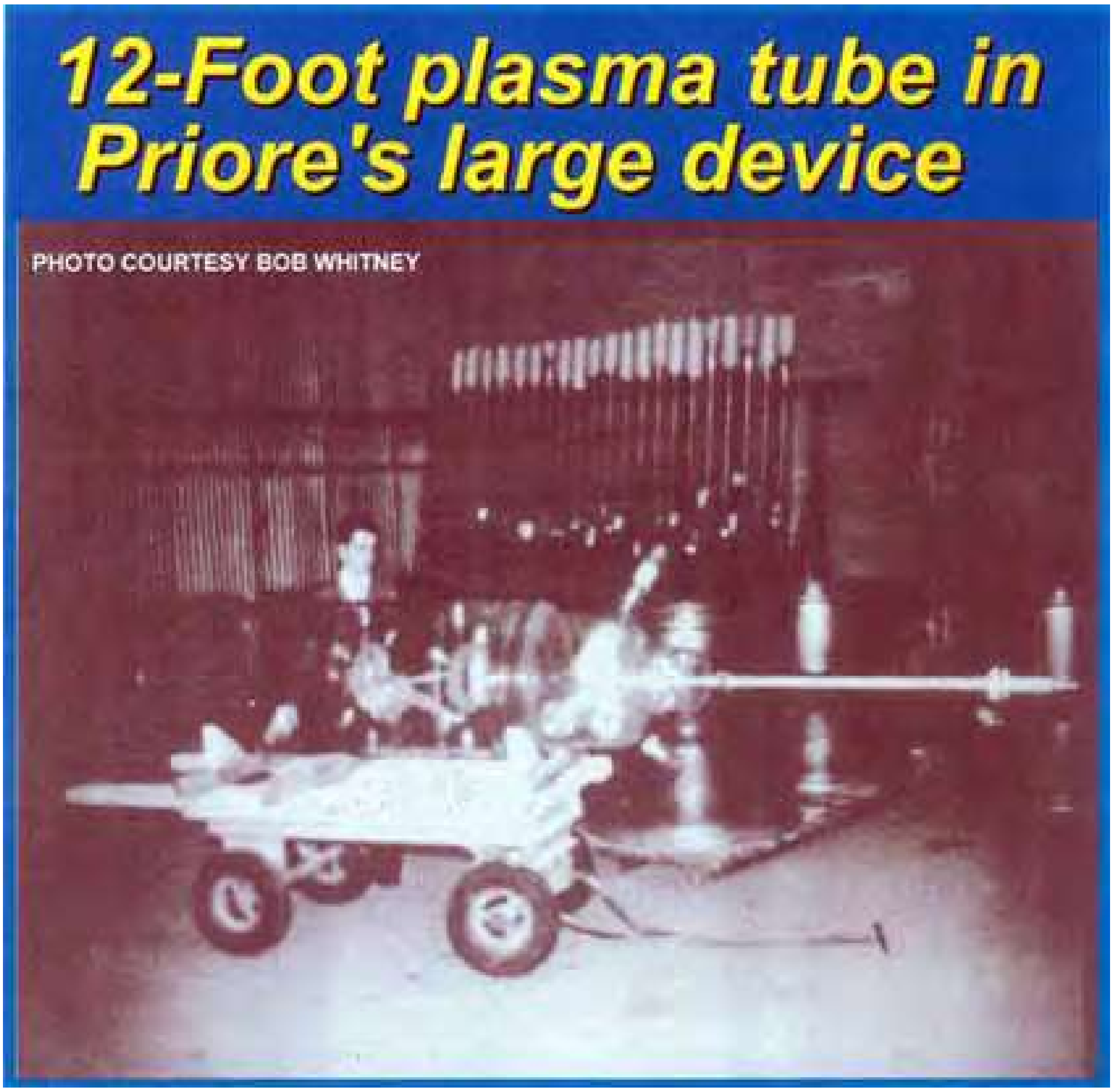

12-Foot plasma tube in Priore's large device

PHOTO COURTESY BOB WHITNEY

Return to Slide Index

Next Slide

Previous Slide

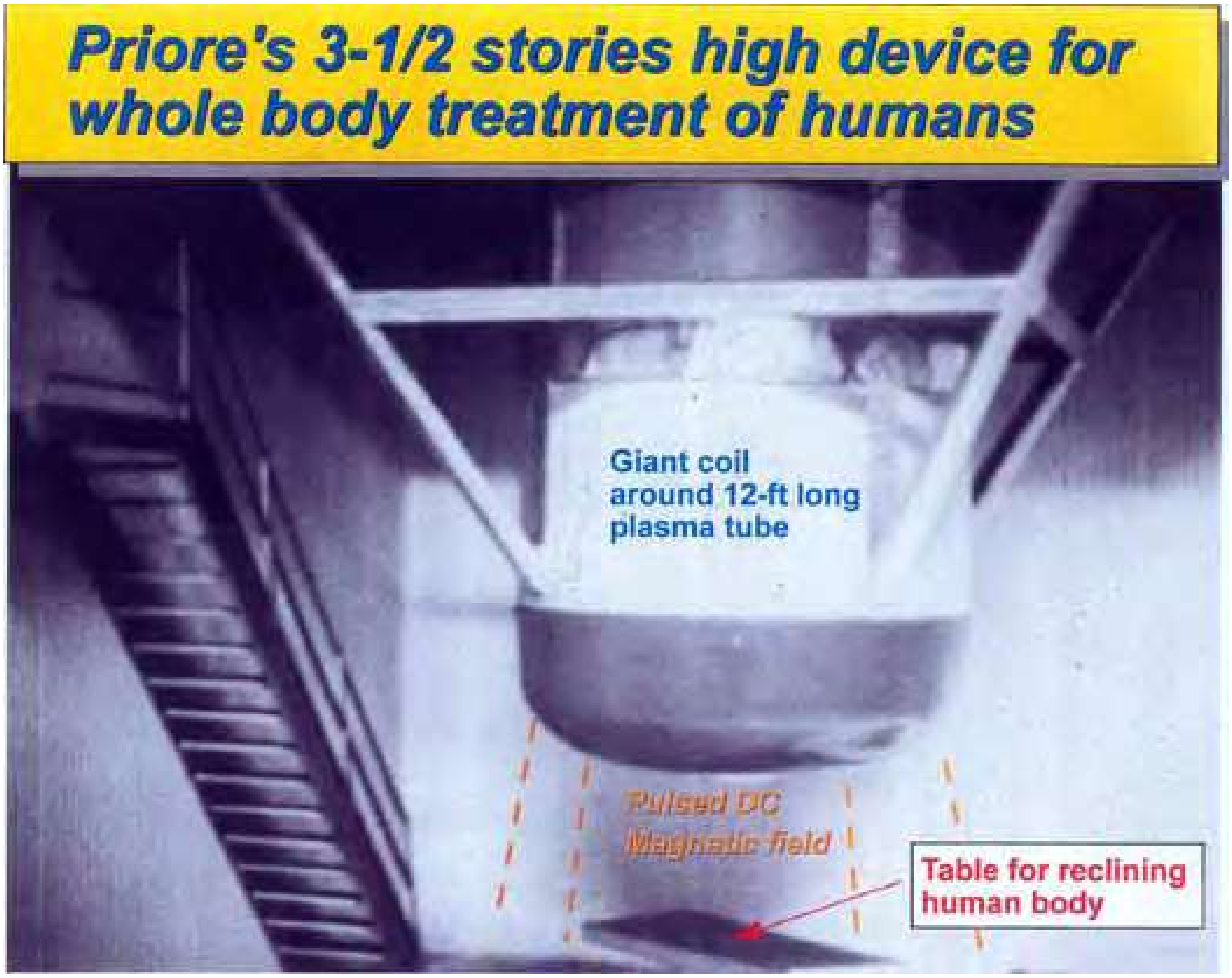

Priore's 3-1/2 stories high device for whole body treatment of humans

Giant coil
around 12-ft long
plasma tube

Pulsed DC
Magnetic field

Table for reclining
human body

Return to Slide Index

Next Slide

Previous Slide

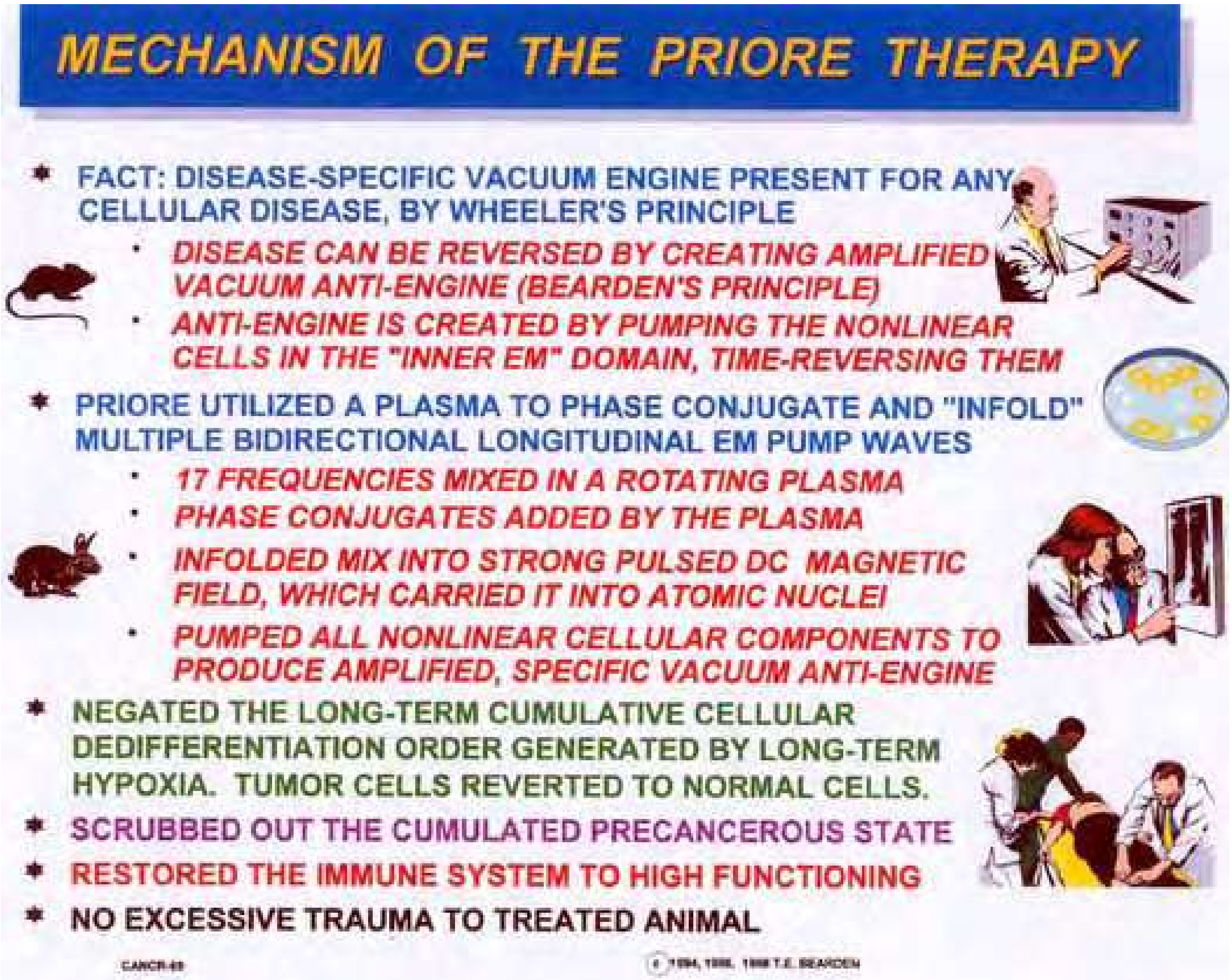

MECHANISM OF THE PRIORE THERAPY

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1998 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

[Previous Slide]

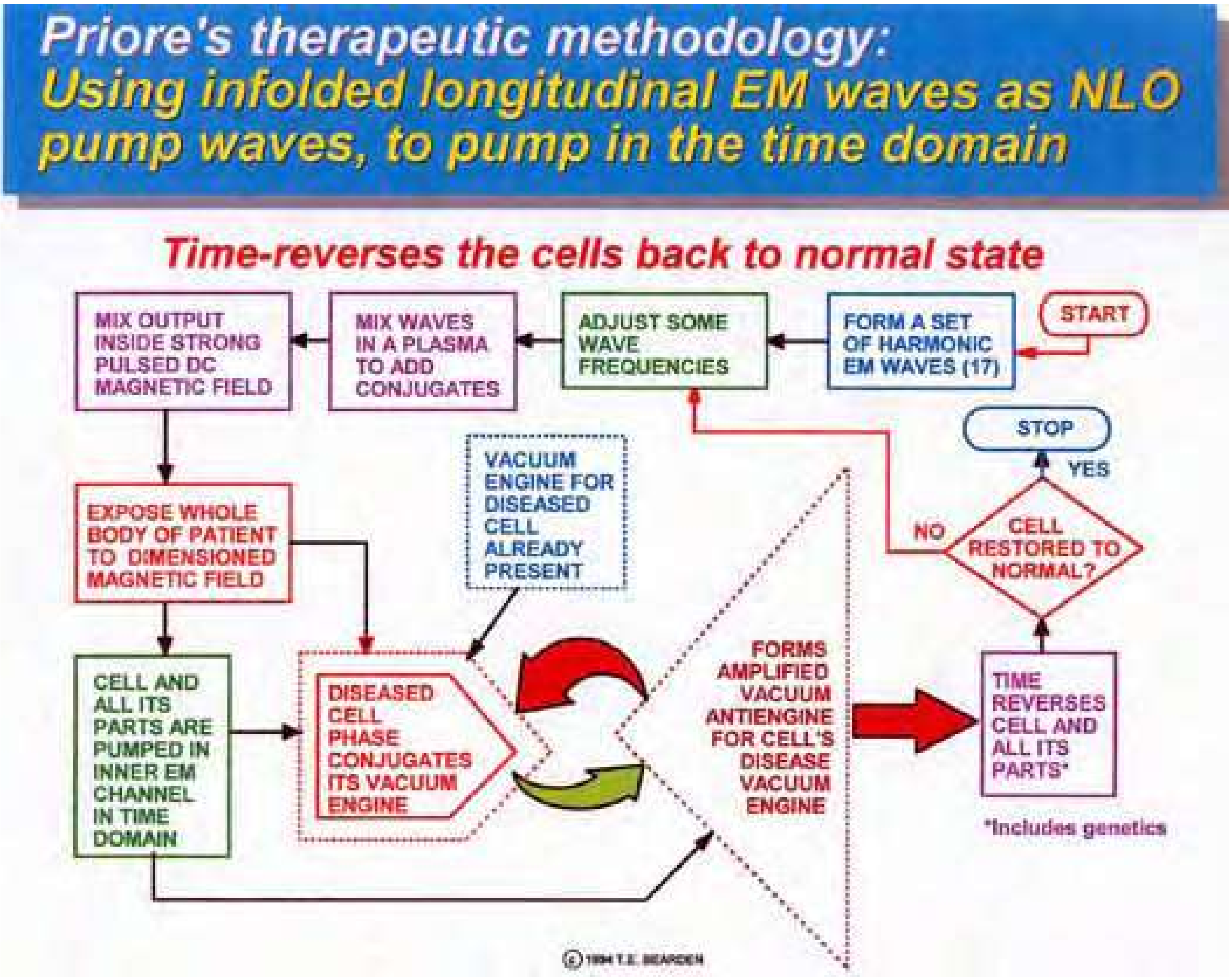

Priore's therapeutic methodology:
Using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain

Time-reverses the cells back to normal state

START

FORM A SET OF HARMONIC EM WAVES (17)

ADJUST SOME WAVE FREQUENCIES

MIX WAVES IN A PLASMA TO ADD CONJUGATES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME REVERSES CELL AND ALL ITS PARTS*

CELL RESTORED TO NORMAL?

NO

YES

STOP

*Includes genetics

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

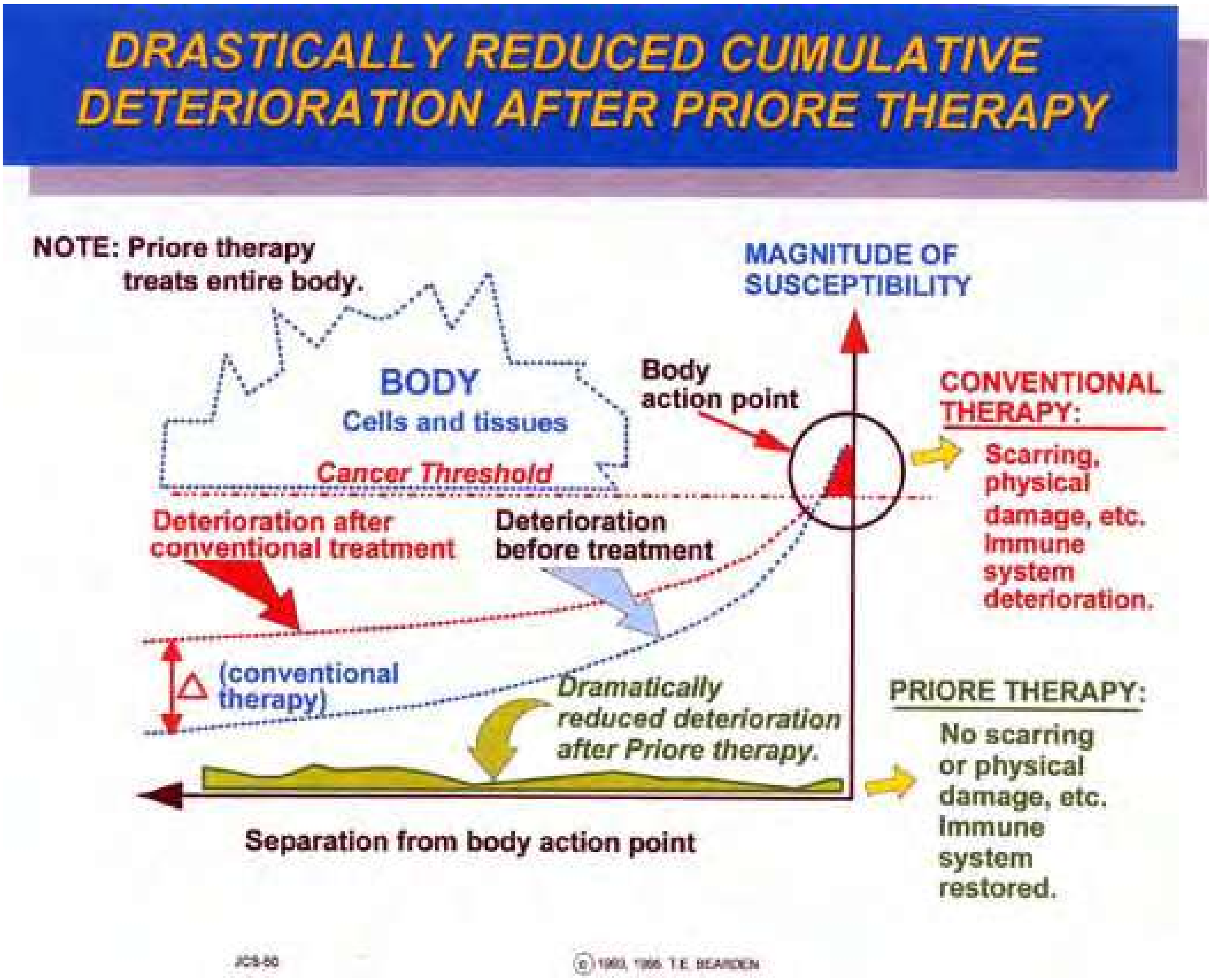

DRASTICALLY REDUCED CUMULATIVE DETERIORATION AFTER PRIORE THERAPY

NOTE: Priore therapy treats entire body.

BODY

Cells and tissues

Cancer Threshold

Body action point

MAGNITUDE OF SUSCEPTIBILITY

Deterioration after conventional treatment

Deterioration before treatment

Δ (conventional therapy)

Dramatically reduced deterioration after Priore therapy.

Separation from body action point

CONVENTIONAL THERAPY:

Scarring, physical damage, etc. Immune system deterioration.

PRIORE THERAPY:

No scarring or physical damage, etc. Immune system restored.

JCS-50

© 1993, 1996 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

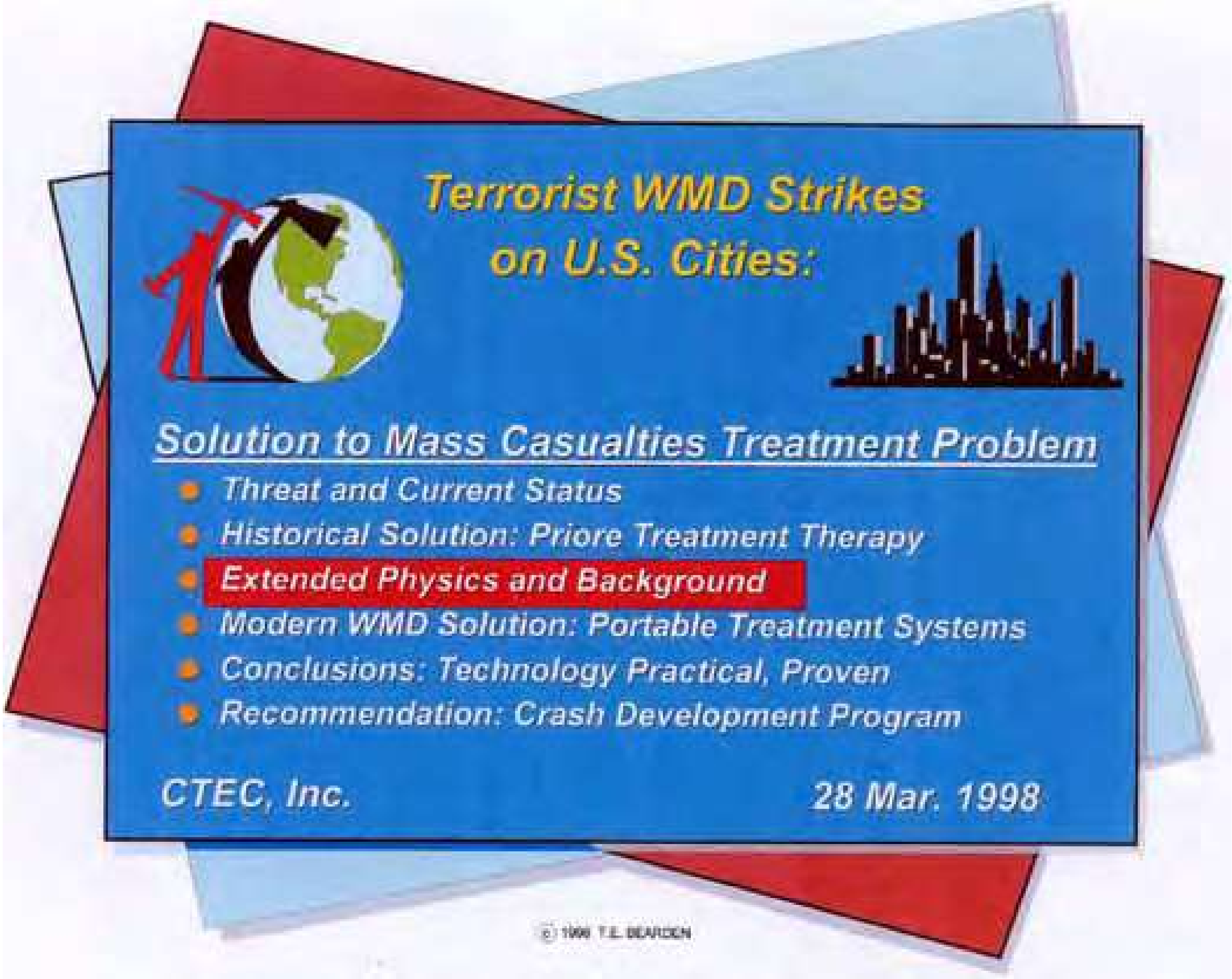

Terrorist WMD Strikes
on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

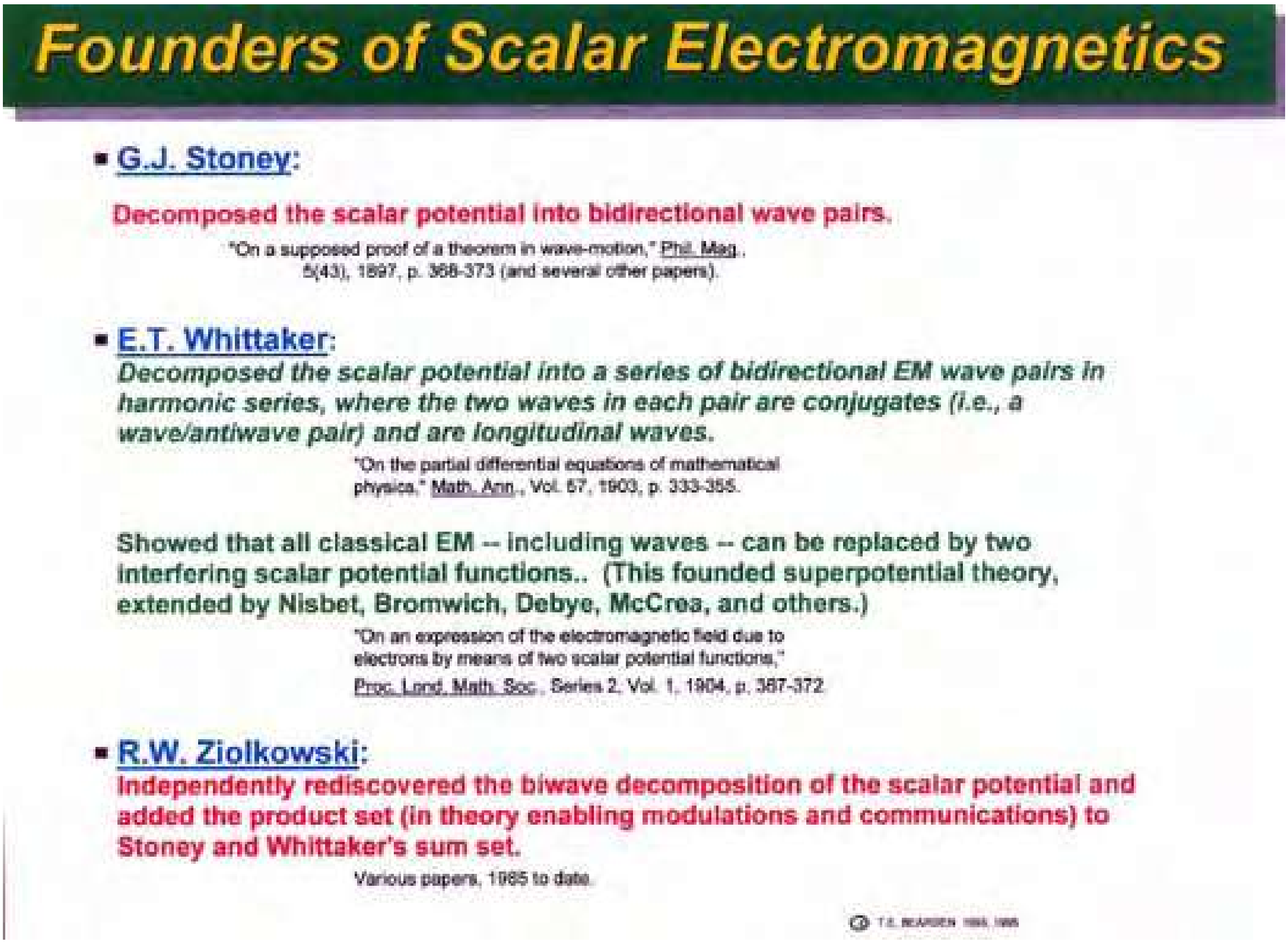

Founders of Scalar Electromagnetics

- **G.J. Stoney:**

  Decomposed the scalar potential into bidirectional wave pairs.

  "On a supposed proof of a theorem in wave-motion," Phil. Mag., 5(43), 1897, p. 368-373 (and several other papers).

- **E.T. Whittaker:**

  Decomposed the scalar potential into a series of bidirectional EM wave pairs in harmonic series, where the two waves in each pair are conjugates (i.e., a wave/antiwave pair) and are longitudinal waves.

  "On the partial differential equations of mathematical physics," Math. Ann., Vol. 57, 1903, p. 333-355.

  Showed that all classical EM -- including waves -- can be replaced by two interfering scalar potential functions.. (This founded superpotential theory, extended by Nisbet, Bromwich, Debye, McCrea, and others.)

  "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372

- **R.W. Ziolkowski:**

  Independently rediscovered the biwave decomposition of the scalar potential and added the product set (in theory enabling modulations and communications) to Stoney and Whittaker's sum set.

  Various papers, 1985 to date.

© T.E. BEARDEN [illegible]

Return to Slide Index

Next Slide

Previous Slide

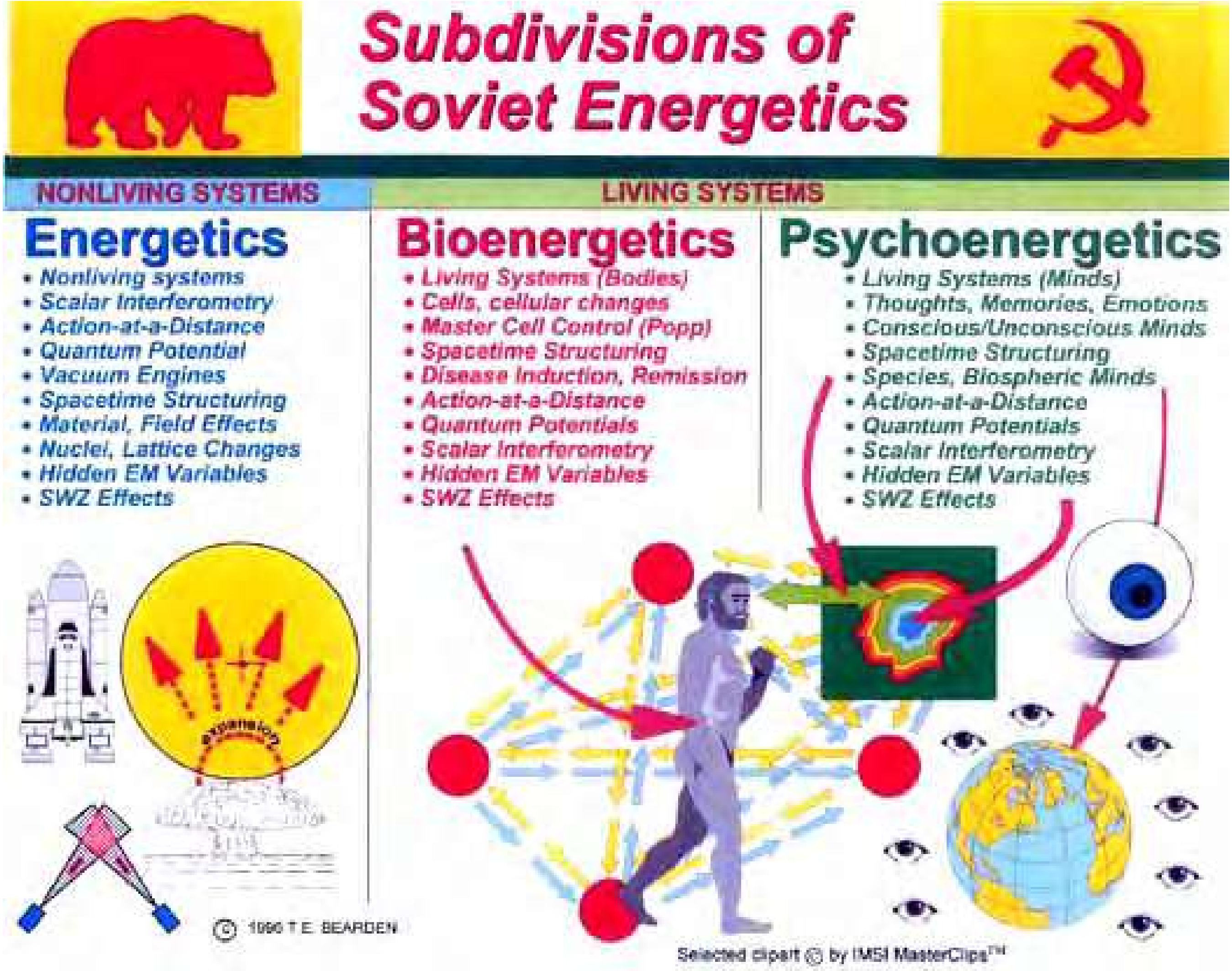

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

LIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1990 T.E. BEARDEN

Selected clipart © by IMSI MasterClips™

Return to Slide Index

Next Slide

Previous Slide

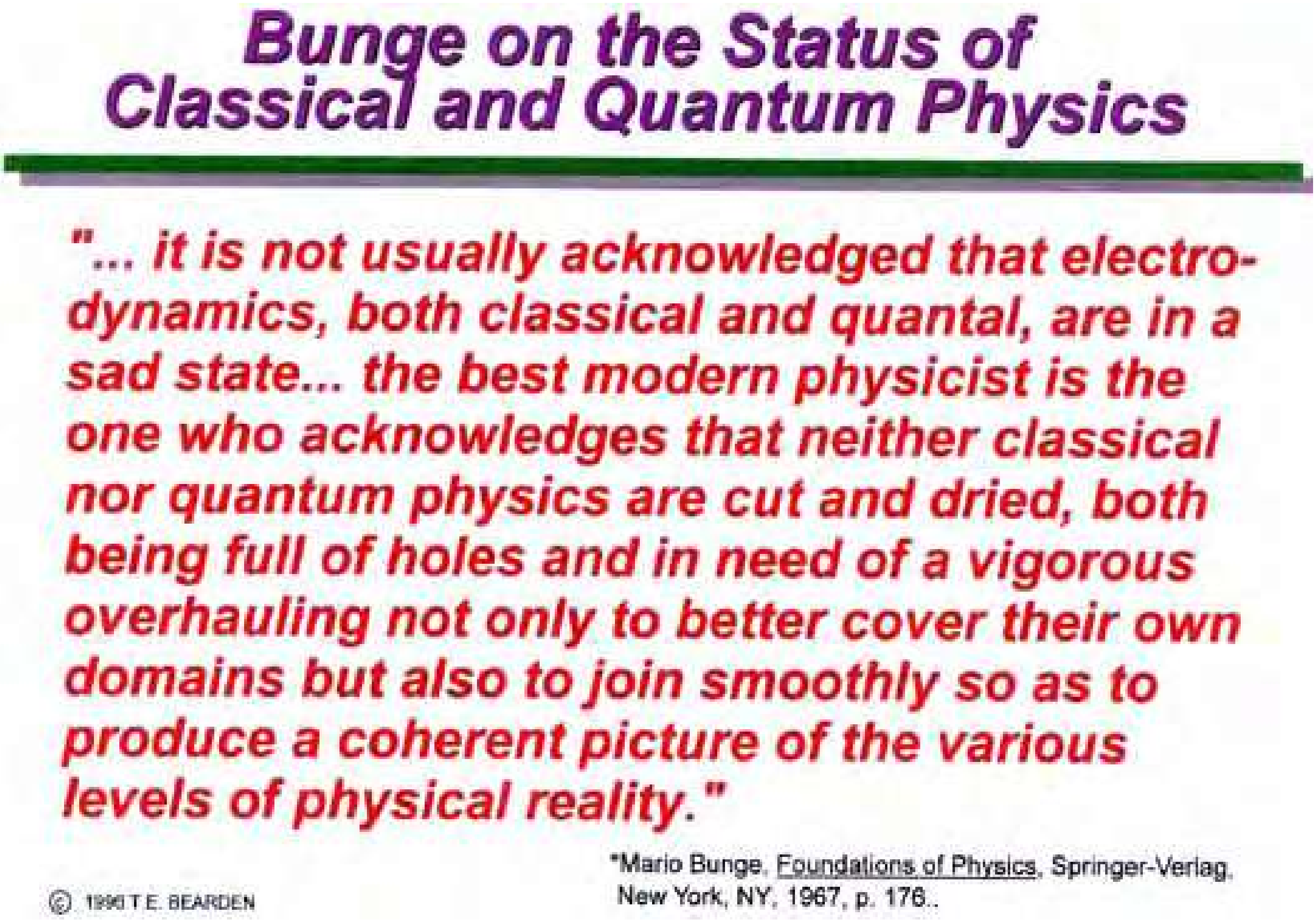

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

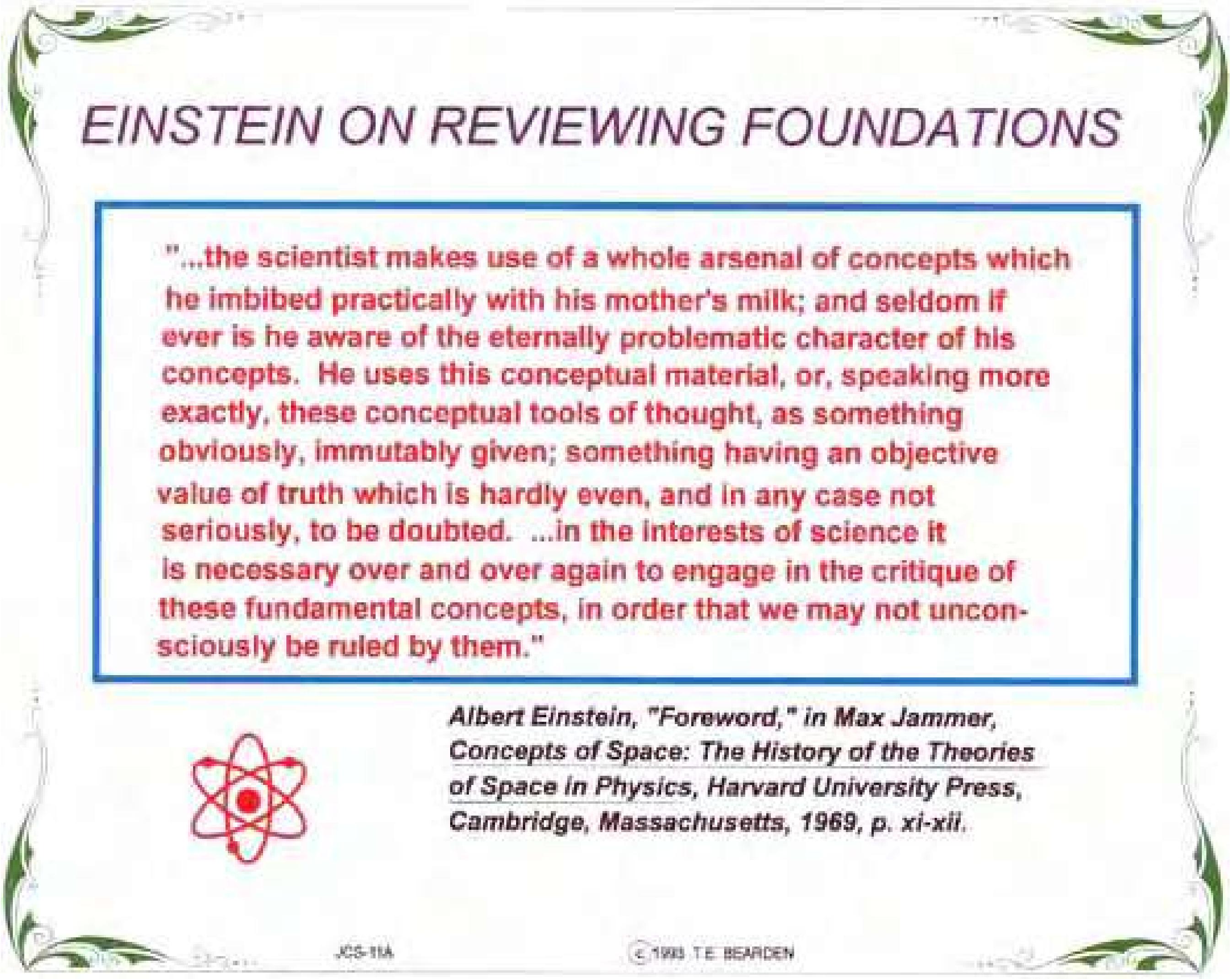

EINSTEIN ON REVIEWING FOUNDATIONS

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

JCS-11A

© 1993 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

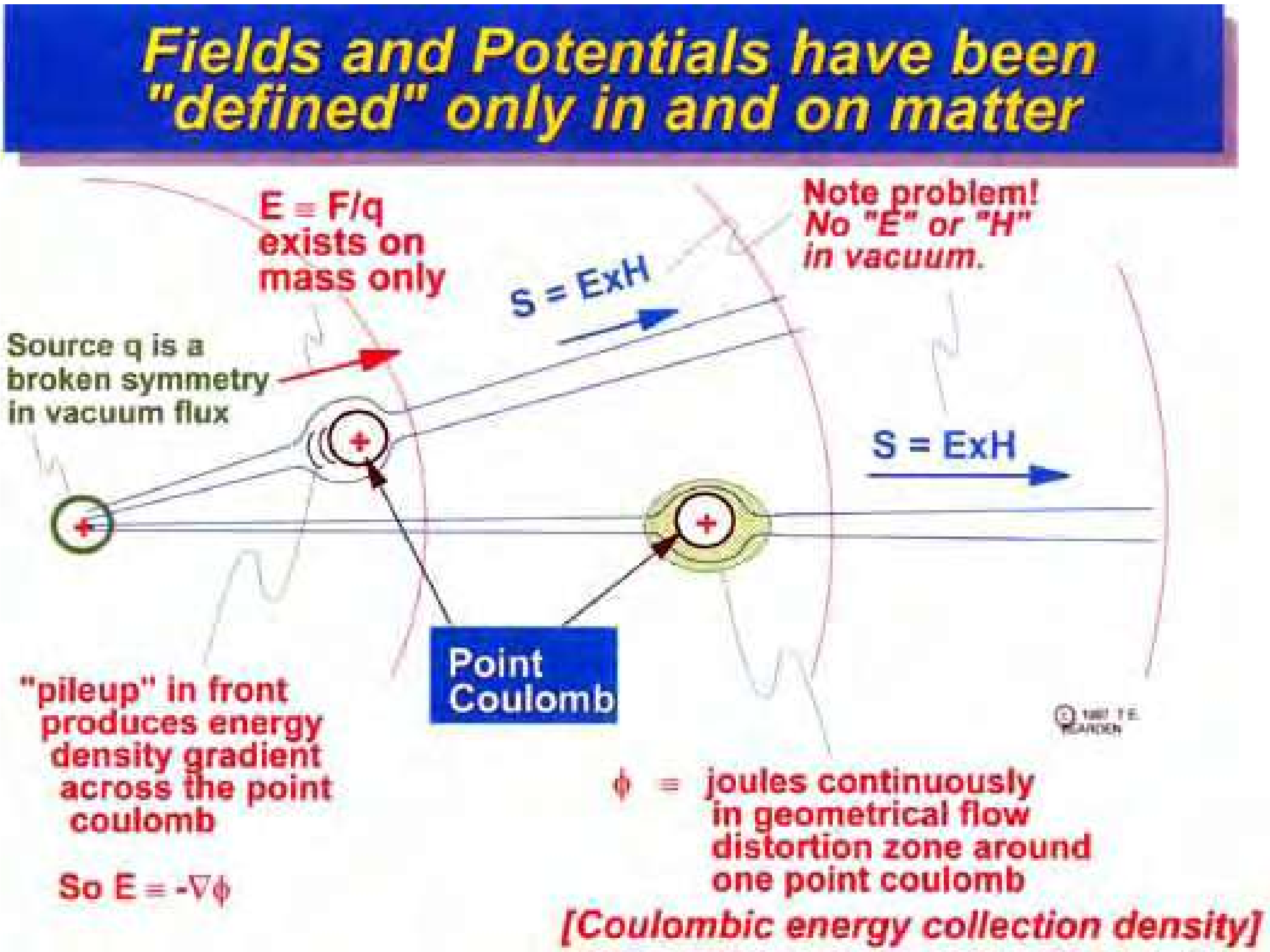

Fields and Potentials have been "defined" only in and on matter

$E \equiv F/q$
exists on
mass only

Note problem!
No "E" or "H"
in vacuum.

$S = ExH$

Source q is a
broken symmetry
in vacuum flux

$S = ExH$

Point
Coulomb

"pileup" in front
produces energy
density gradient
across the point
coulomb

So $E \equiv -\nabla\phi$

$\phi \equiv$ joules continuously
in geometrical flow
distortion zone around
one point coulomb

[Coulombic energy collection density]

Return to Slide Index

Next Slide

Previous Slide

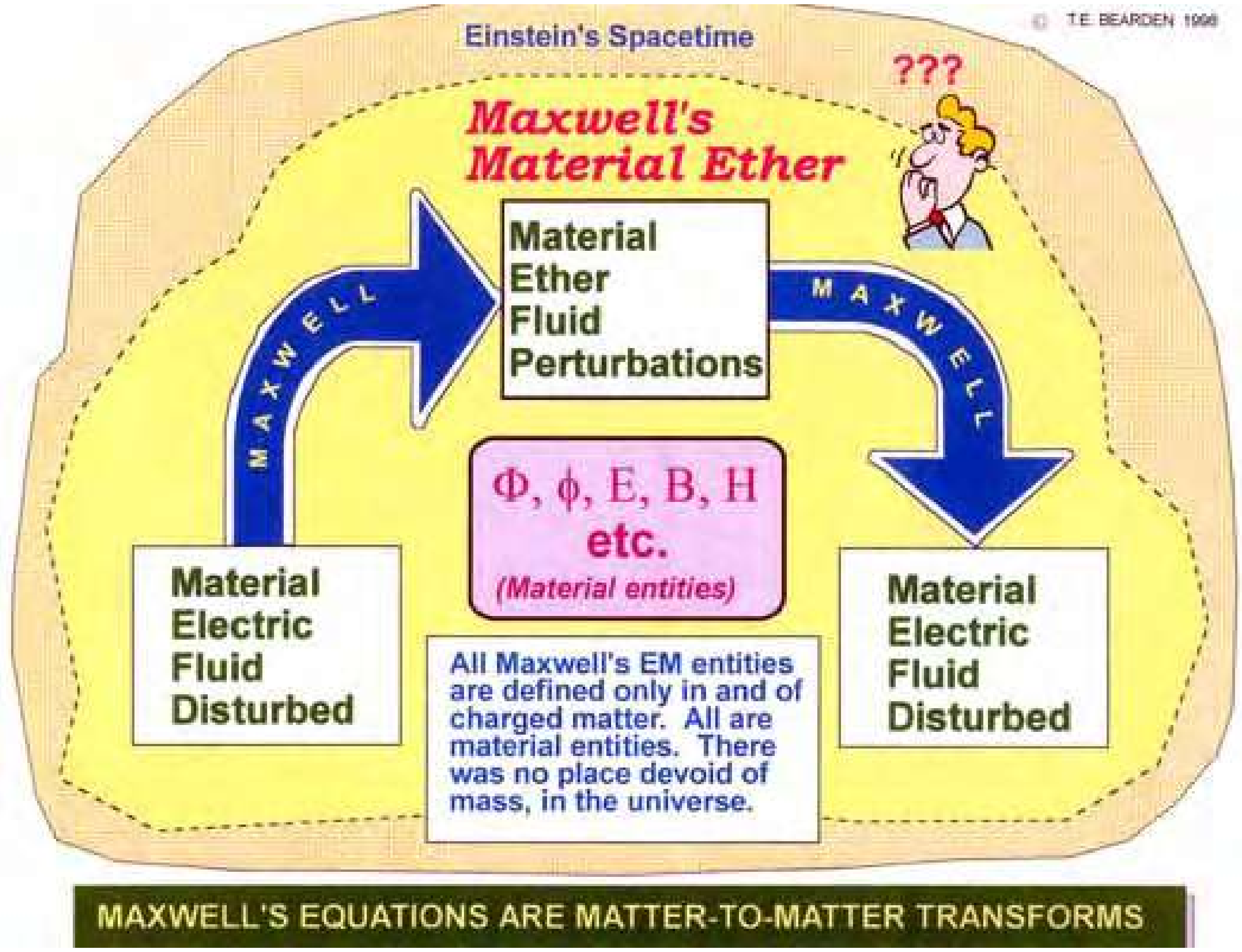
???
MAXWELL
MAXWELL

Einstein's Spacetime

© T.E. BEARDEN 1996

Maxwell's Material Ether

Material Ether Fluid Perturbations

Φ, ϕ, E, B, H etc.

(Material entities)

Material Electric Fluid Disturbed

All Maxwell's EM entities are defined only in and of charged matter. All are material entities. There was no place devoid of mass, in the universe.

Material Electric Fluid Disturbed

MAXWELL'S EQUATIONS ARE MATTER-TO-MATTER TRANSFORMS

Return to Slide Index

Next Slide

Previous Slide

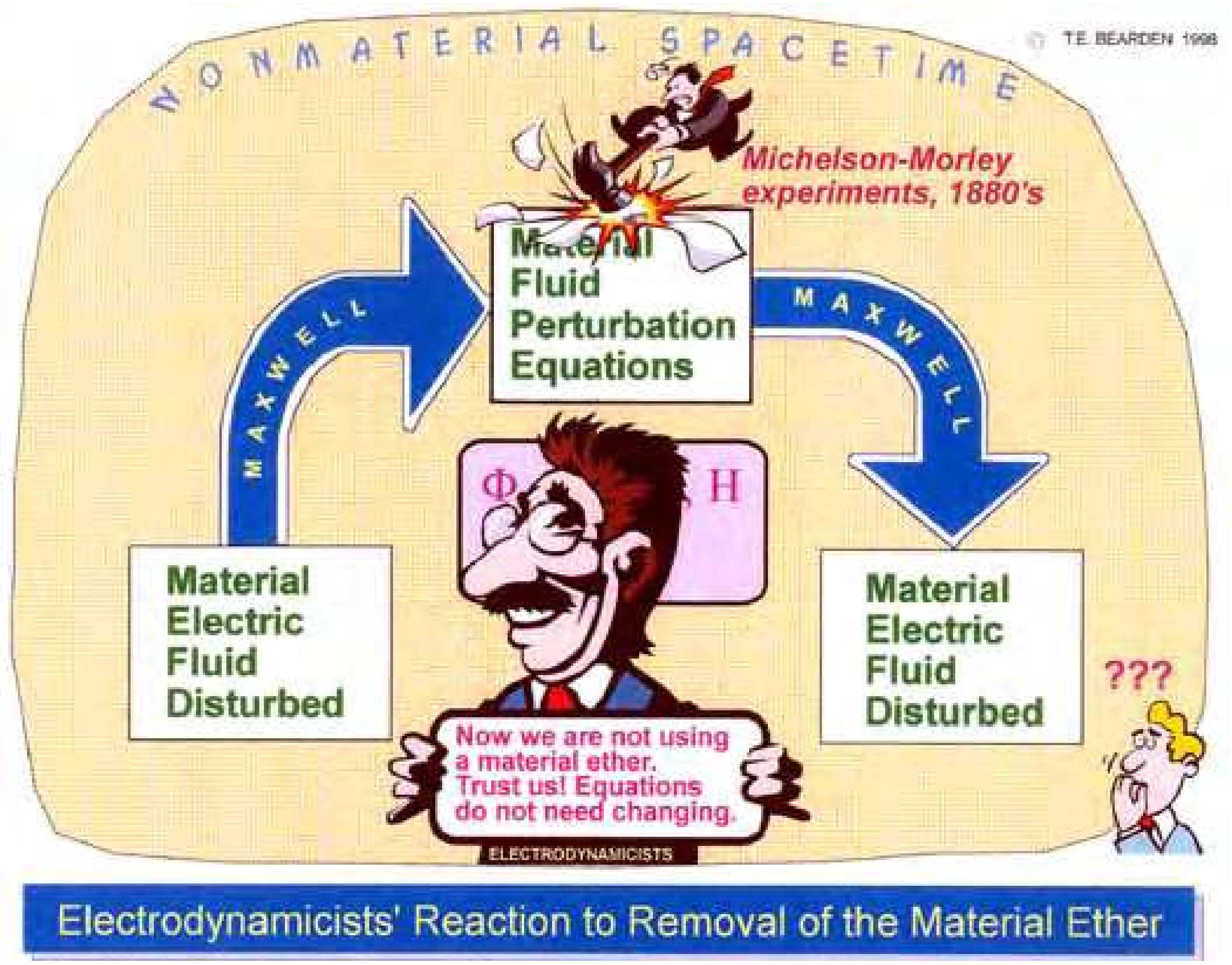
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Michelson-Morley experiments, 1880's
Material Fluid Perturbation Equations
MAXWELL
MAXWELL
Φ
H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! Equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

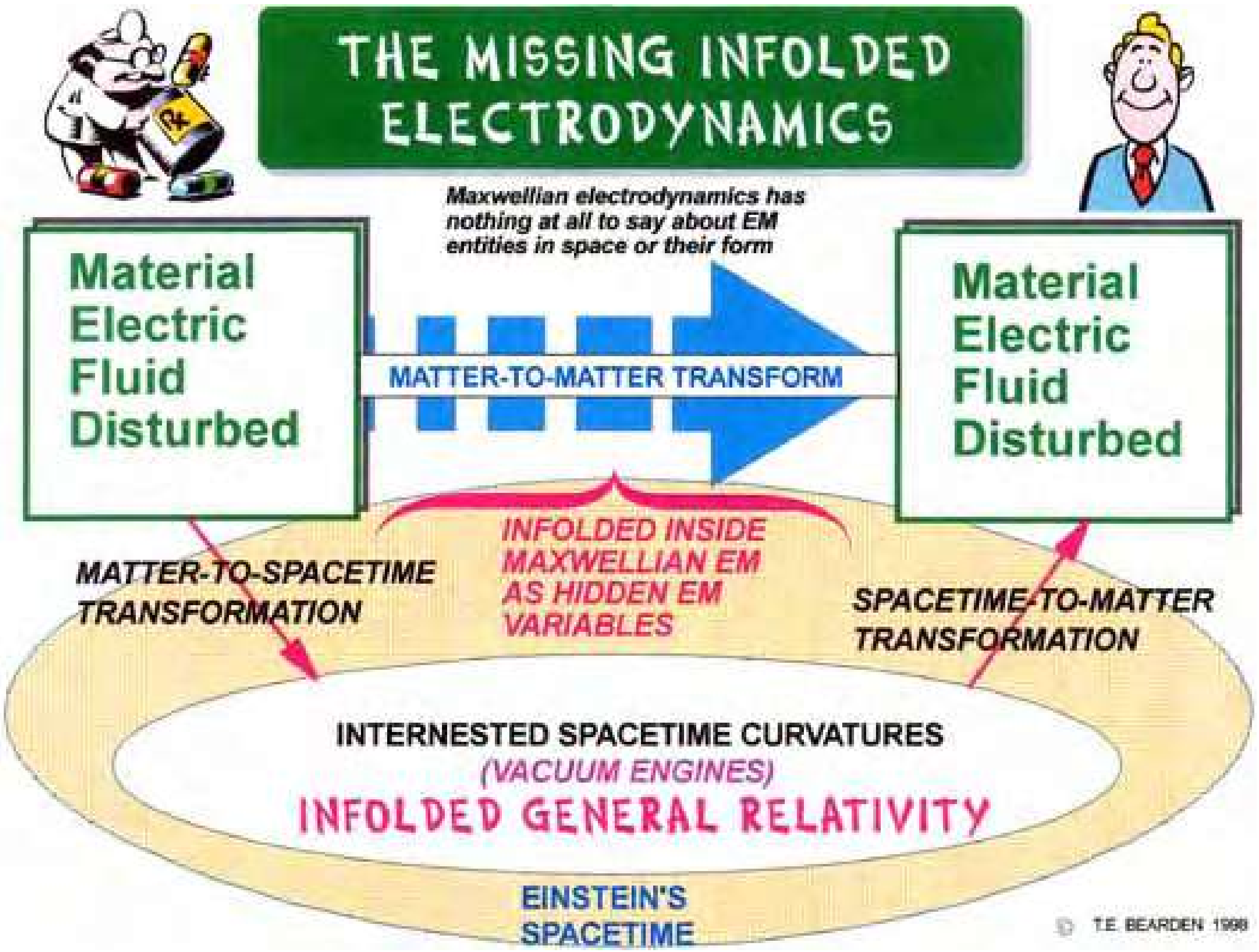

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

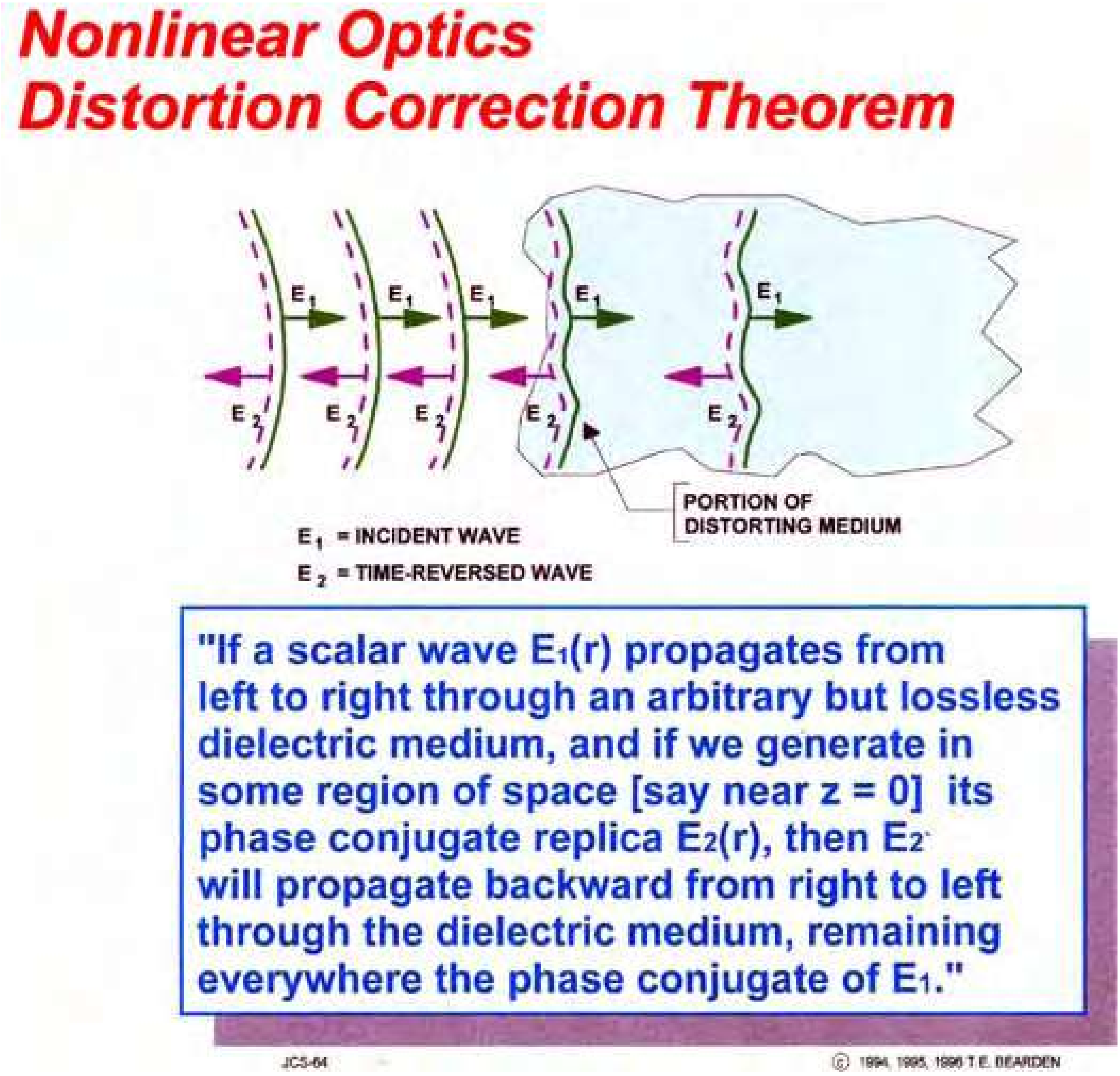

Nonlinear Optics Distortion Correction Theorem

E_1
E_2
PORTION OF DISTORTING MEDIUM

E_1 = INCIDENT WAVE
E_2 = TIME-REVERSED WAVE

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

JCS-64

© 1994, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

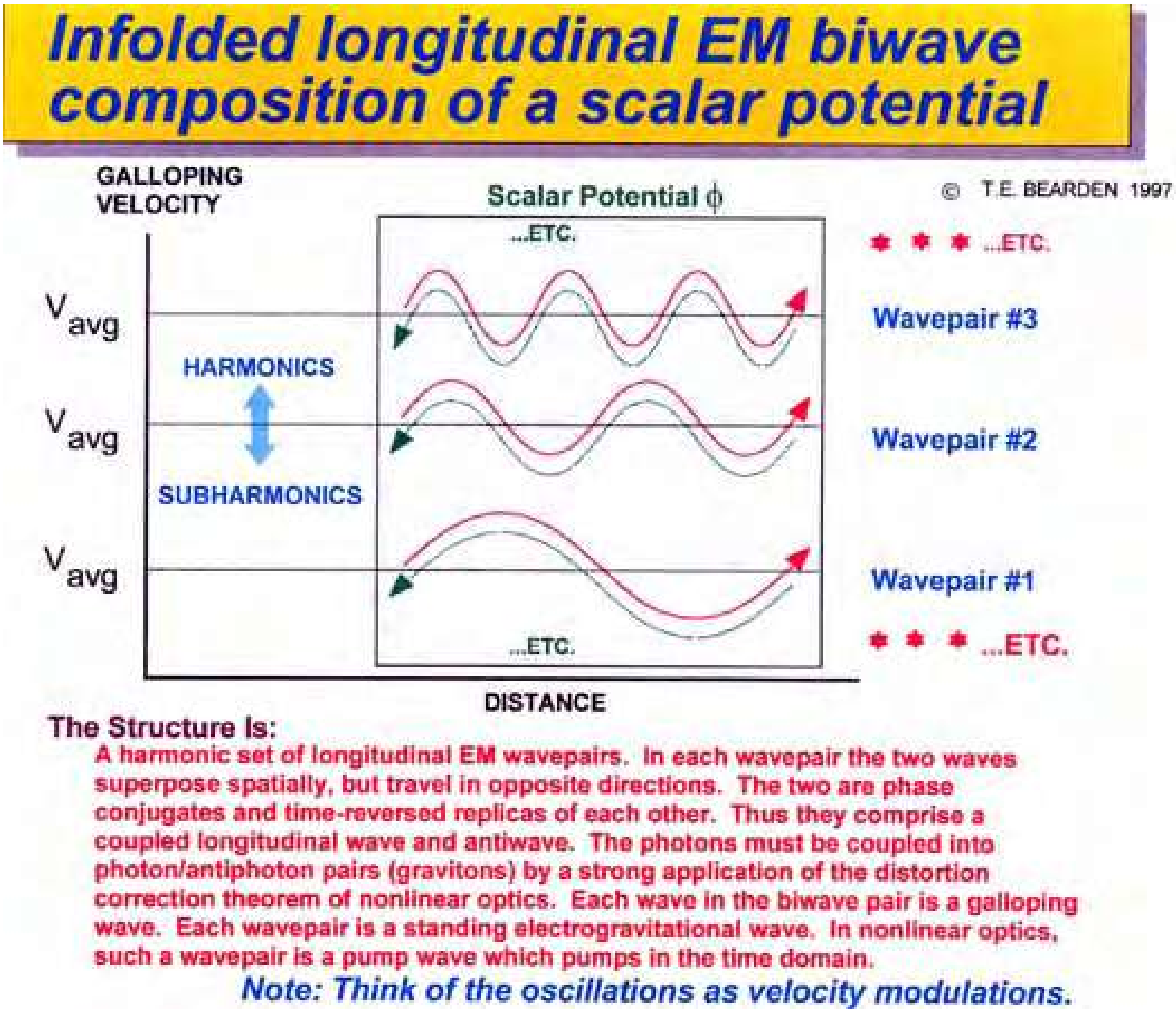

Infolded longitudinal EM biwave composition of a scalar potential

GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC.
* * * ...ETC.
V_{avg}
Wavepair #3
HARMONICS
V_{avg}
Wavepair #2
SUBHARMONICS
V_{avg}
Wavepair #1
...ETC.
* * * ...ETC.
DISTANCE

The Structure Is:

A harmonic set of longitudinal EM wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave. In nonlinear optics, such a wavepair is a pump wave which pumps in the time domain.

Note: Think of the oscillations as velocity modulations.

Return to Slide Index

Next Slide

Previous Slide

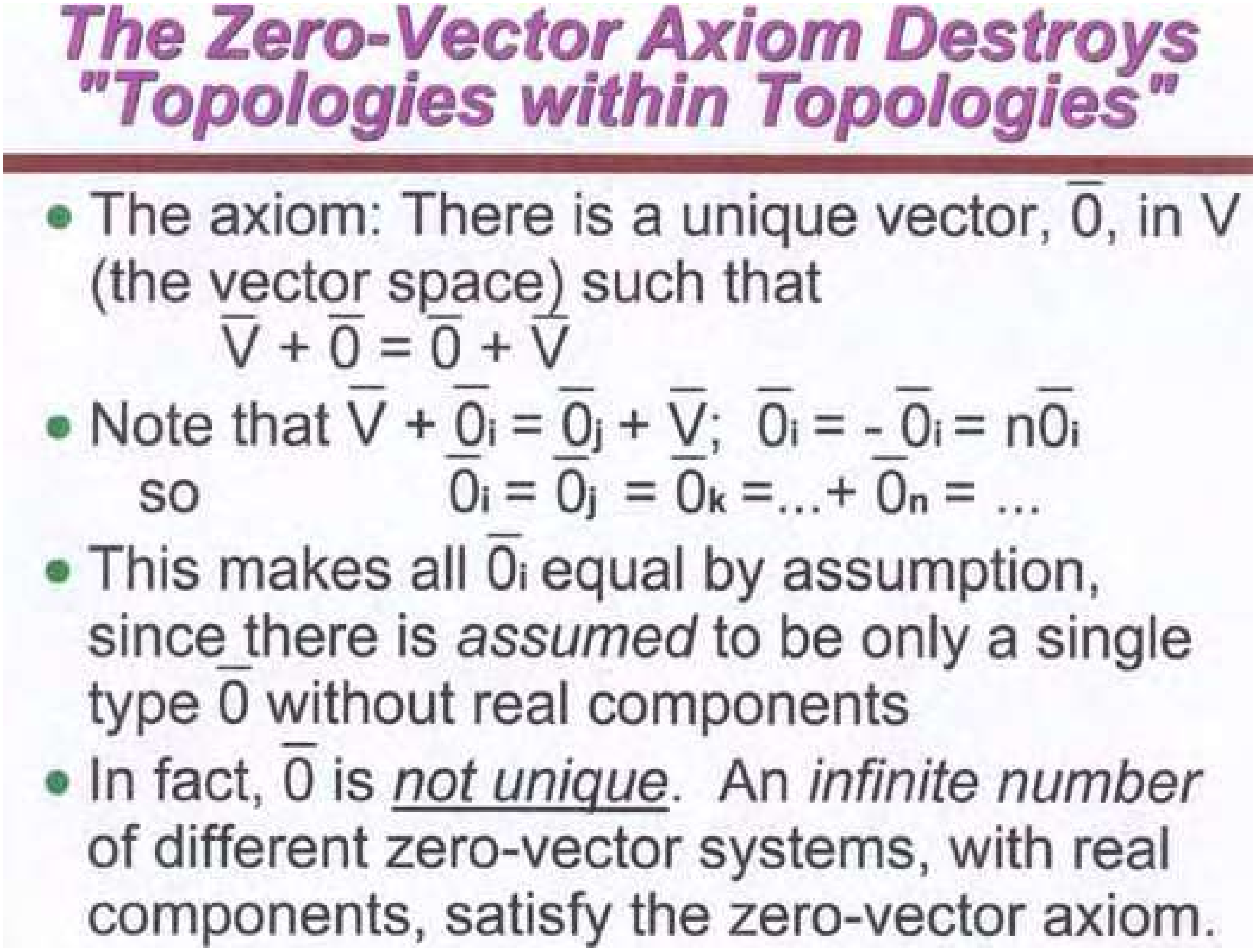

The Zero-Vector Axiom Destroys "Topologies within Topologies"

- The axiom: There is a unique vector, $\overline{0}$, in V (the vector space) such that
  $\overline{V} + \overline{0} = \overline{0} + \overline{V}$
- Note that $\overline{V} + \overline{0}_i = \overline{0}_j + \overline{V}$; $\overline{0}_i = -\overline{0}_i = n\overline{0}_i$
  so $\overline{0}_i = \overline{0}_j = \overline{0}_k = \ldots + \overline{0}_n = \ldots$
- This makes all $\overline{0}_i$ equal by assumption, since there is *assumed* to be only a single type $\overline{0}$ without real components
- In fact, $\overline{0}$ is *not unique*. An *infinite number* of different zero-vector systems, with real components, satisfy the zero-vector axiom.

Return to Slide Index

Next Slide

Previous Slide

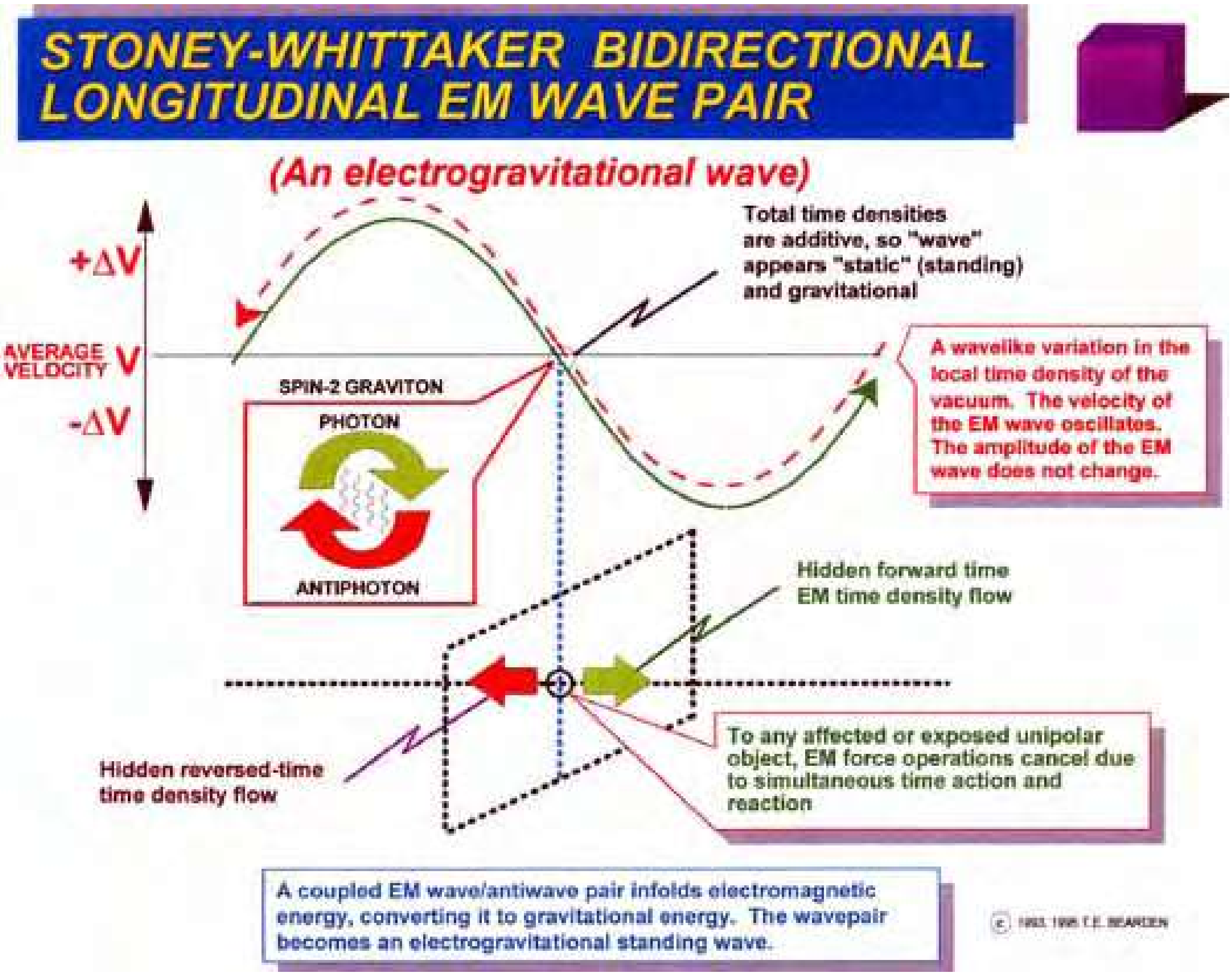

STONEY-WHITTAKER BIDIRECTIONAL LONGITUDINAL EM WAVE PAIR

(An electrogravitational wave)

Total time densities are additive, so "wave" appears "static" (standing) and gravitational

A wavelike variation in the local time density of the vacuum. The velocity of the EM wave oscillates. The amplitude of the EM wave does not change.

Hidden forward time EM time density flow

Hidden reversed-time time density flow

To any affected or exposed unipolar object, EM force operations cancel due to simultaneous time action and reaction

A coupled EM wave/antiwave pair infolds electromagnetic energy, converting it to gravitational energy. The wavepair becomes an electrogravitational standing wave.

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

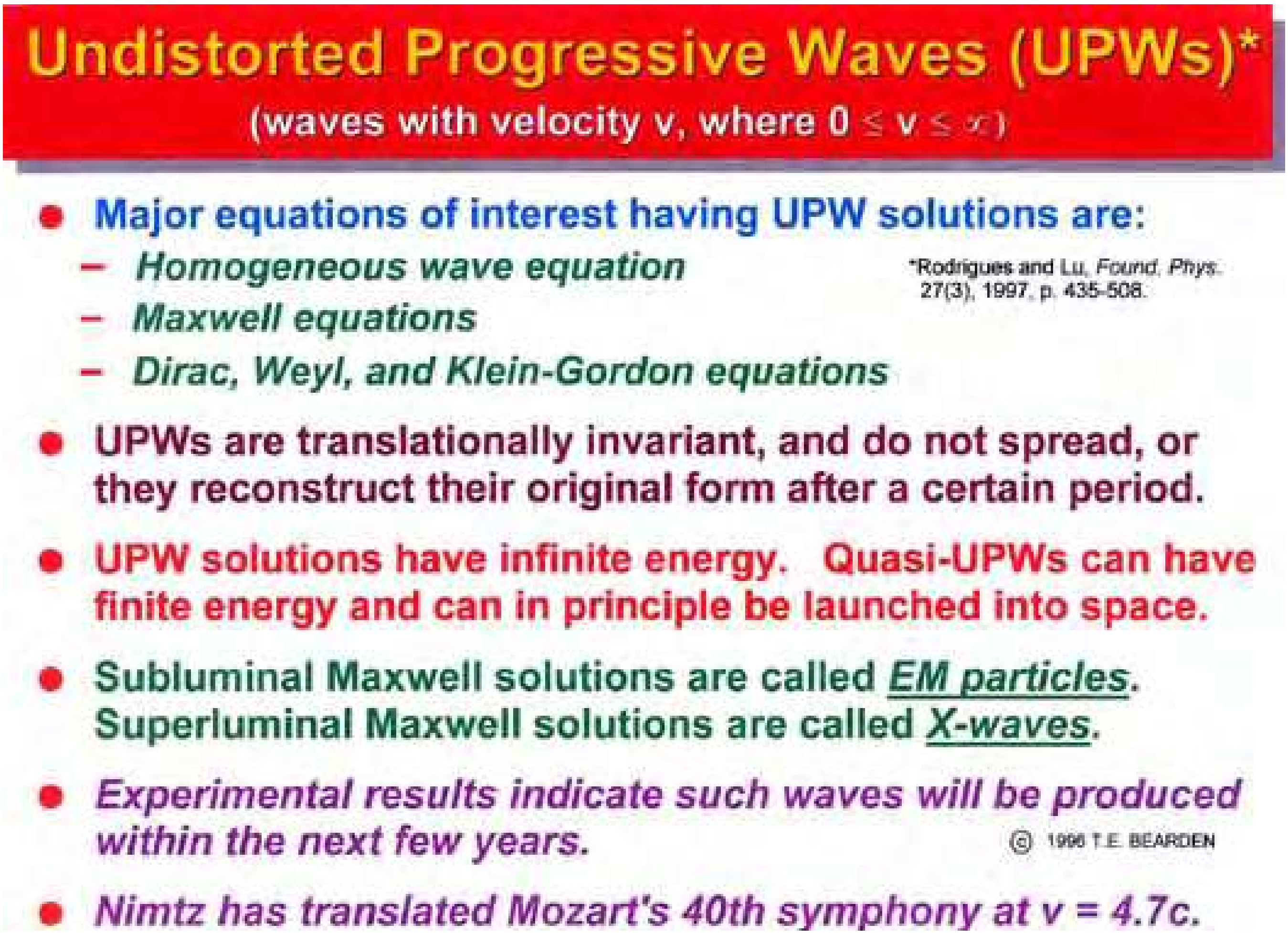

Undistorted Progressive Waves (UPWs)*

(waves with velocity v, where $0 \leq v \leq \infty$)

- Major equations of interest having UPW solutions are:
  - *Homogeneous wave equation*
  - *Maxwell equations*
  - *Dirac, Weyl, and Klein-Gordon equations*

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

- UPWs are translationally invariant, and do not spread, or they reconstruct their original form after a certain period.
- UPW solutions have infinite energy. Quasi-UPWs can have finite energy and can in principle be launched into space.
- Subluminal Maxwell solutions are called *EM particles*. Superluminal Maxwell solutions are called *X-waves*.
- *Experimental results indicate such waves will be produced within the next few years.*

© 1996 T.E. BEARDEN

- *Nimtz has translated Mozart's 40th symphony at v = 4.7c.*

Return to Slide Index

Next Slide

Previous Slide

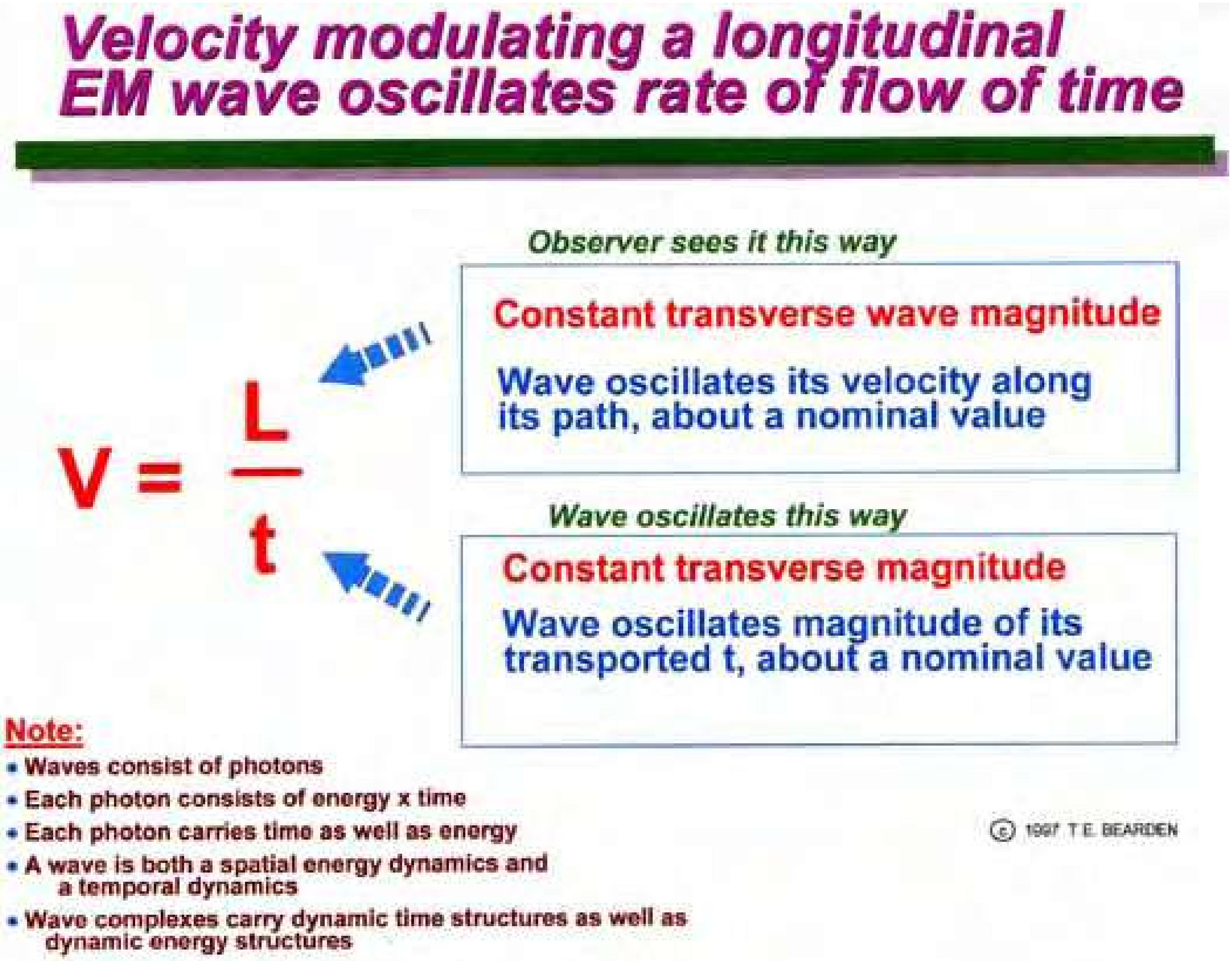

Velocity modulating a longitudinal EM wave oscillates rate of flow of time

$$V = \frac{L}{t}$$

Observer sees it this way

Constant transverse wave magnitude

Wave oscillates its velocity along its path, about a nominal value

Wave oscillates this way

Constant transverse magnitude

Wave oscillates magnitude of its transported t, about a nominal value

Note:

- Waves consist of photons
- Each photon consists of energy x time
- Each photon carries time as well as energy
- A wave is both a spatial energy dynamics and a temporal dynamics
- Wave complexes carry dynamic time structures as well as dynamic energy structures

© 1997 T E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

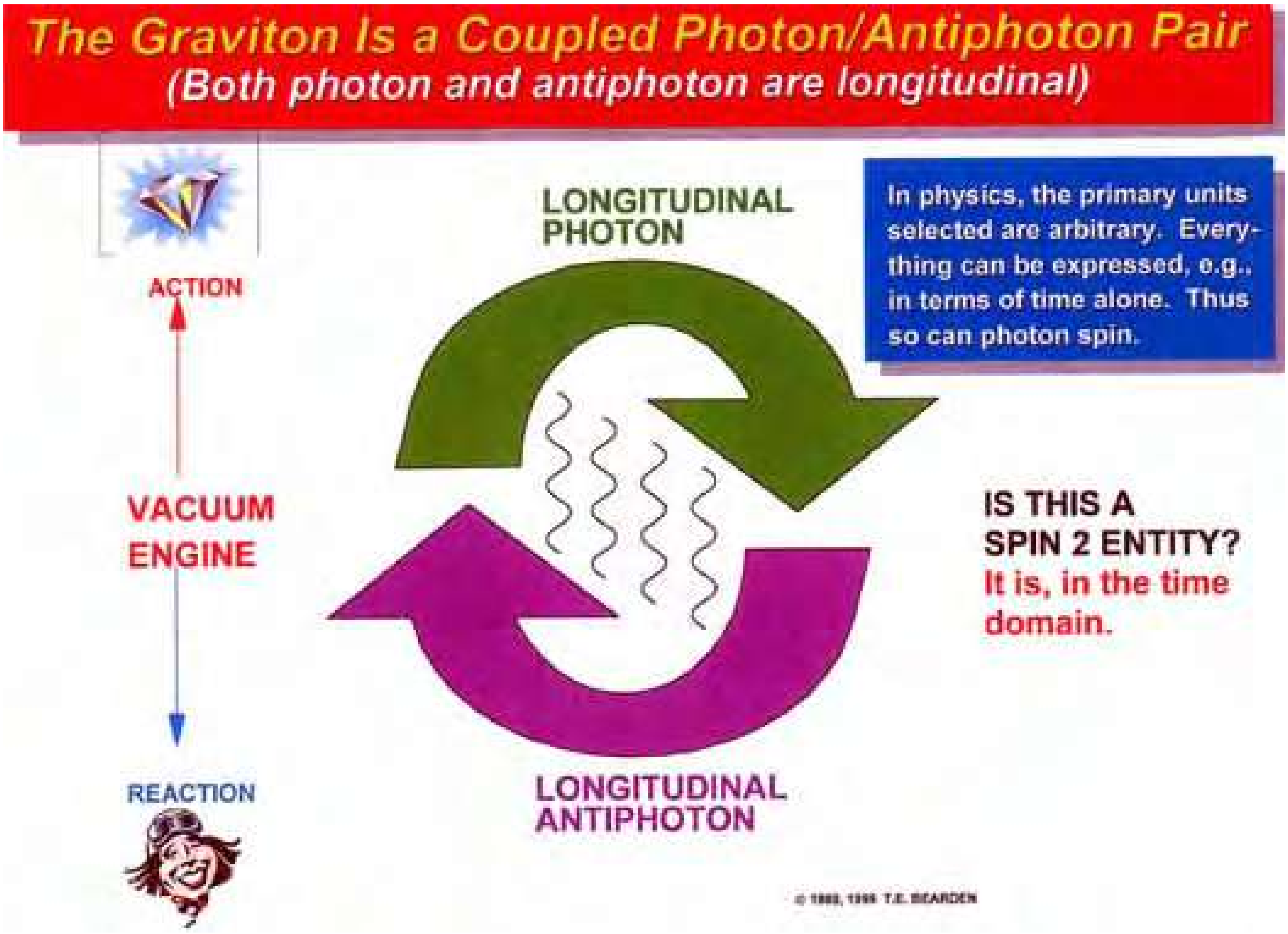

The Graviton Is a Coupled Photon/Antiphoton Pair

(Both photon and antiphoton are longitudinal)

In physics, the primary units selected are arbitrary. Everything can be expressed, e.g., in terms of time alone. Thus so can photon spin.

ACTION

VACUUM ENGINE

REACTION

LONGITUDINAL PHOTON

LONGITUDINAL ANTIPHOTON

IS THIS A SPIN 2 ENTITY?

It is, in the time domain.

© 1988, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

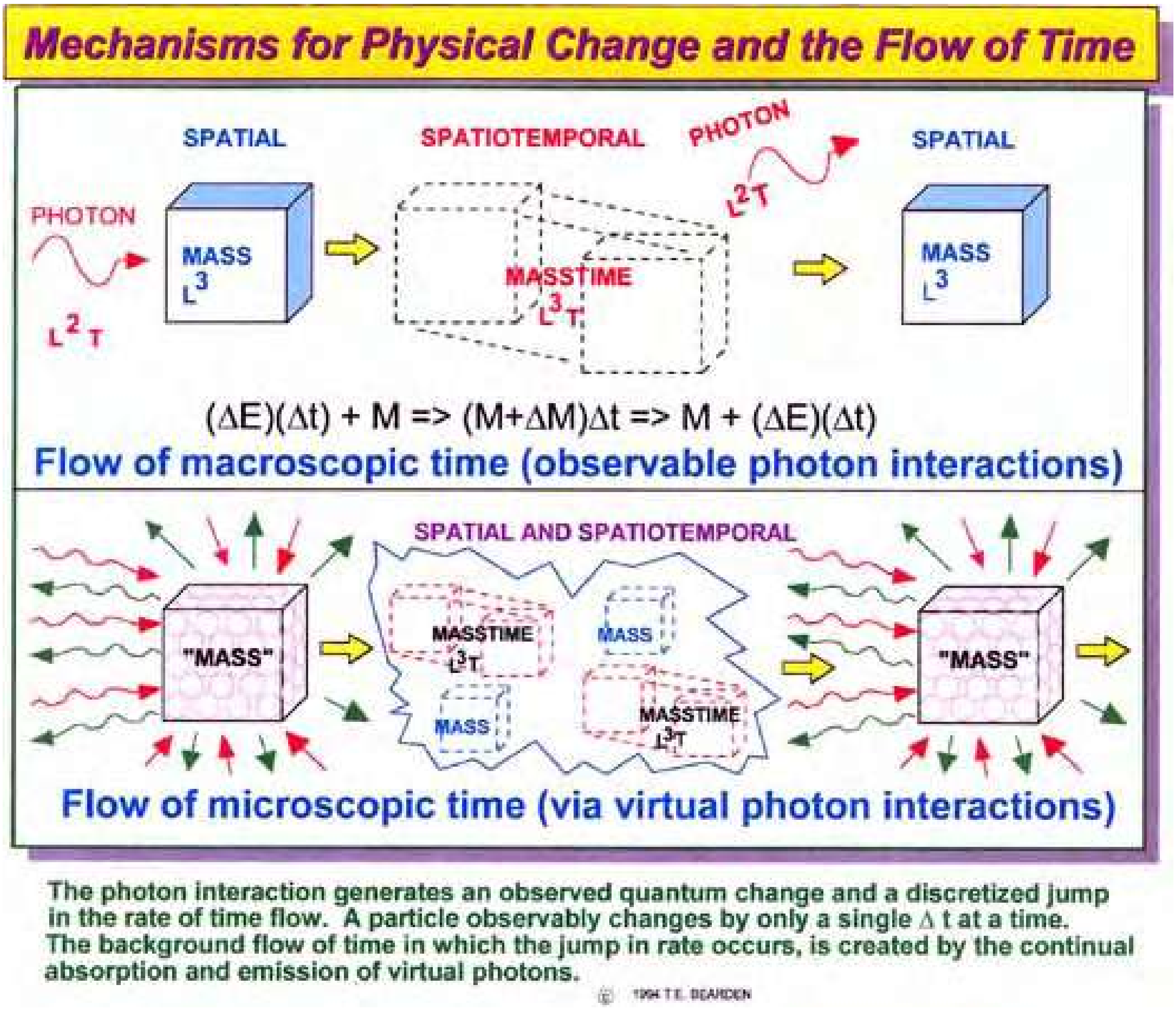

Mechanisms for Physical Change and the Flow of Time

SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS L^3
L^2T
MASSTIME L^3T
L^2T
MASS L^3

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME L^3T
MASS
MASS
MASSTIME L^3T
"MASS"

Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δt at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

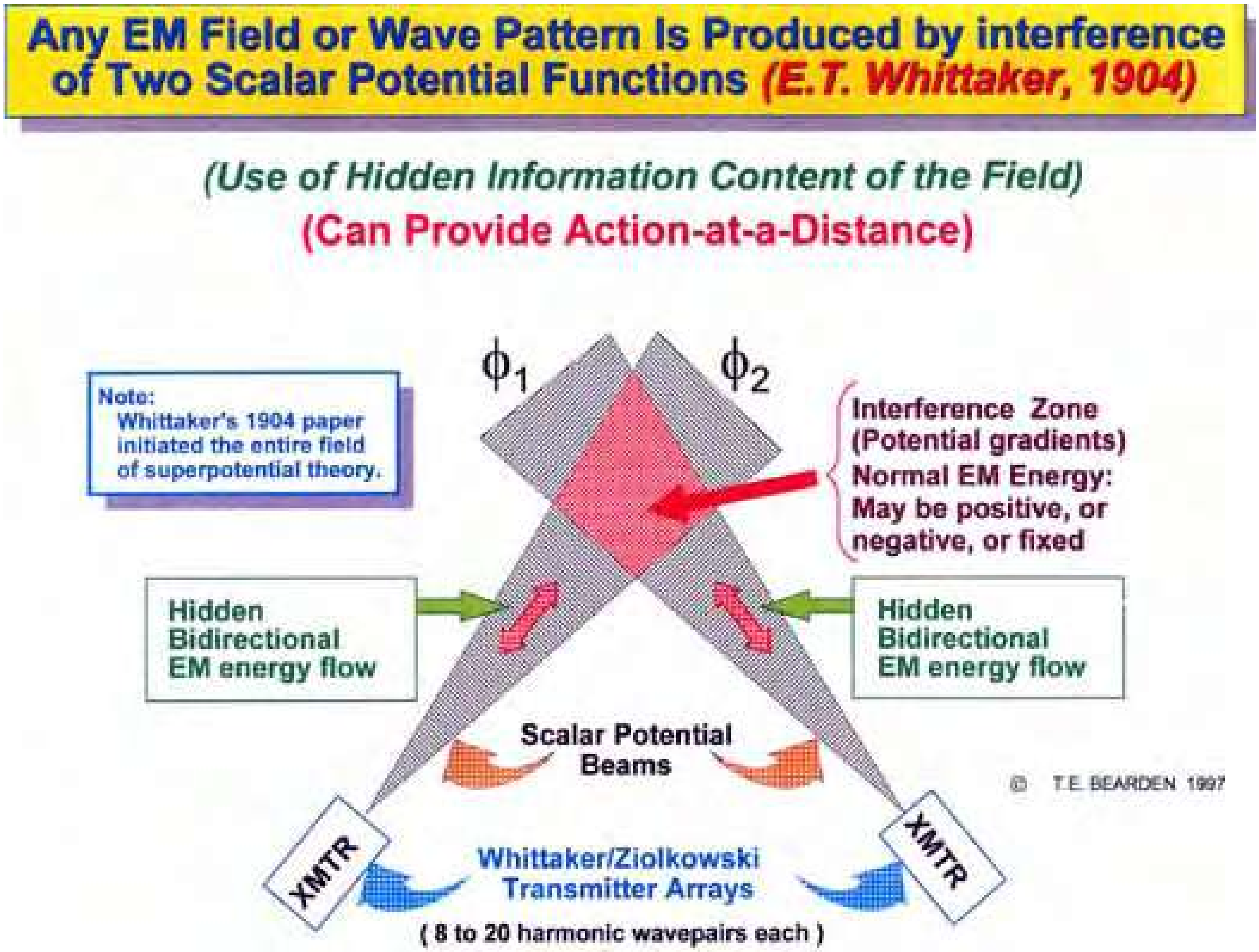

Any EM Field or Wave Pattern Is Produced by interference of Two Scalar Potential Functions *(E.T. Whittaker, 1904)*

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.

Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or negative, or fixed

ϕ_1
ϕ_2
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

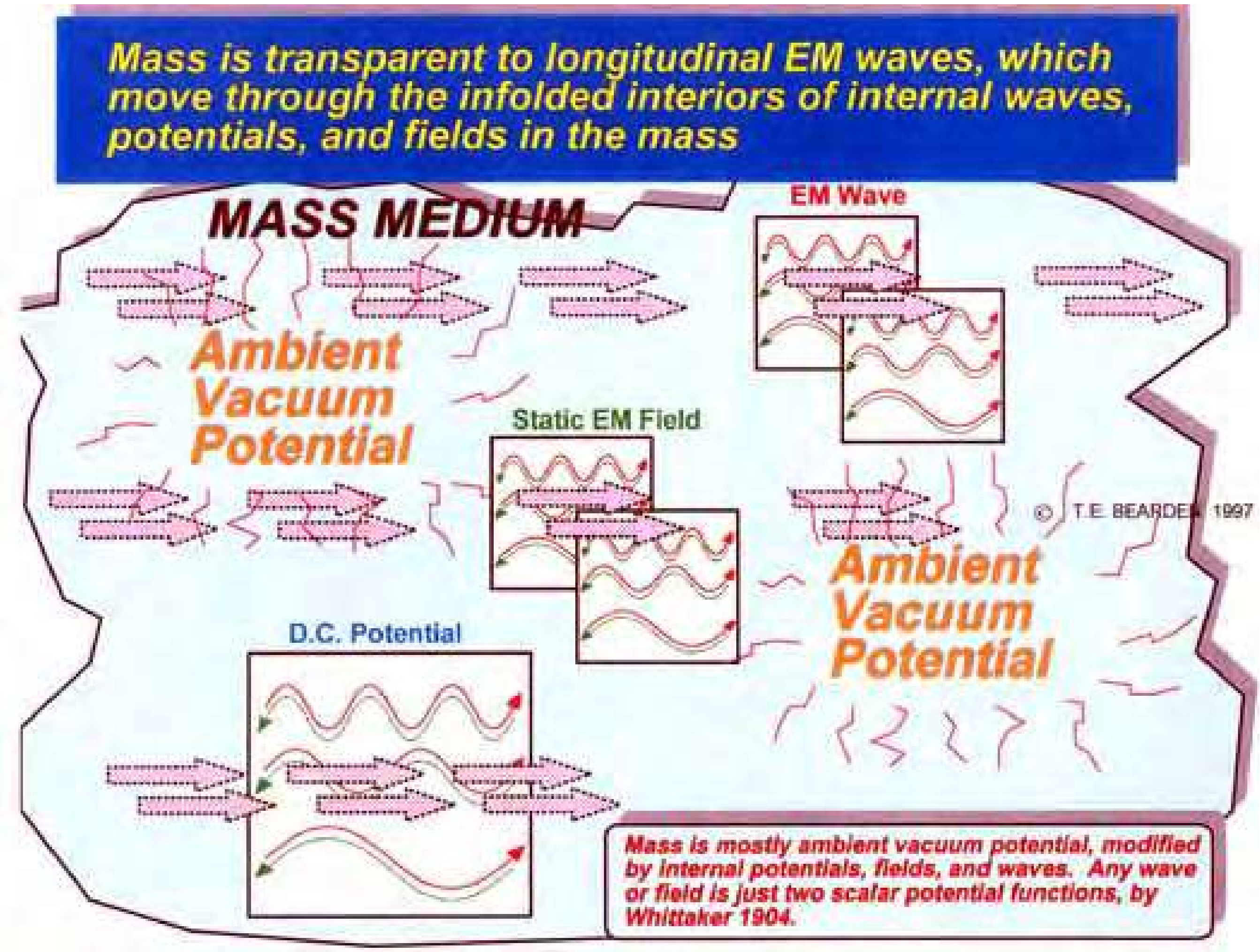

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
Ambient Vacuum Potential
EM Wave
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

Return to Slide Index

Next Slide

Previous Slide

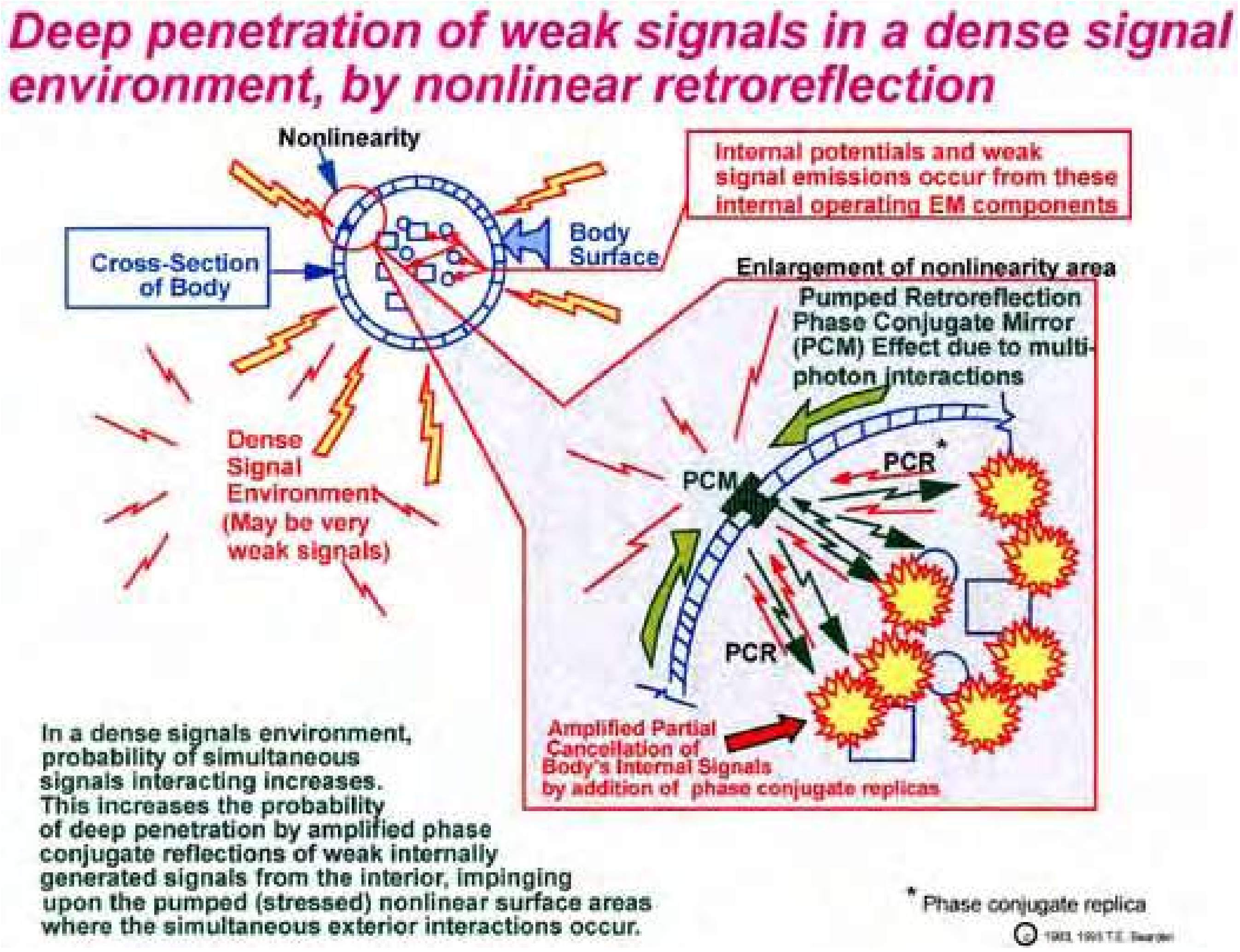

Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection

Nonlinearity
Cross-Section of Body
Body Surface
Dense Signal Environment (May be very weak signals)
PCM
PCR*
PCR

Internal potentials and weak signal emissions occur from these internal operating EM components

Enlargement of nonlinearity area

Pumped Retroreflection Phase Conjugate Mirror (PCM) Effect due to multi-photon interactions

Amplified Partial Cancellation of Body's Internal Signals by addition of phase conjugate replicas

In a dense signals environment, probability of simultaneous signals interacting increases. This increases the probability of deep penetration by amplified phase conjugate reflections of weak internally generated signals from the interior, impinging upon the pumped (stressed) nonlinear surface areas where the simultaneous exterior interactions occur.

* Phase conjugate replica

Return to Slide Index

Next Slide

Previous Slide

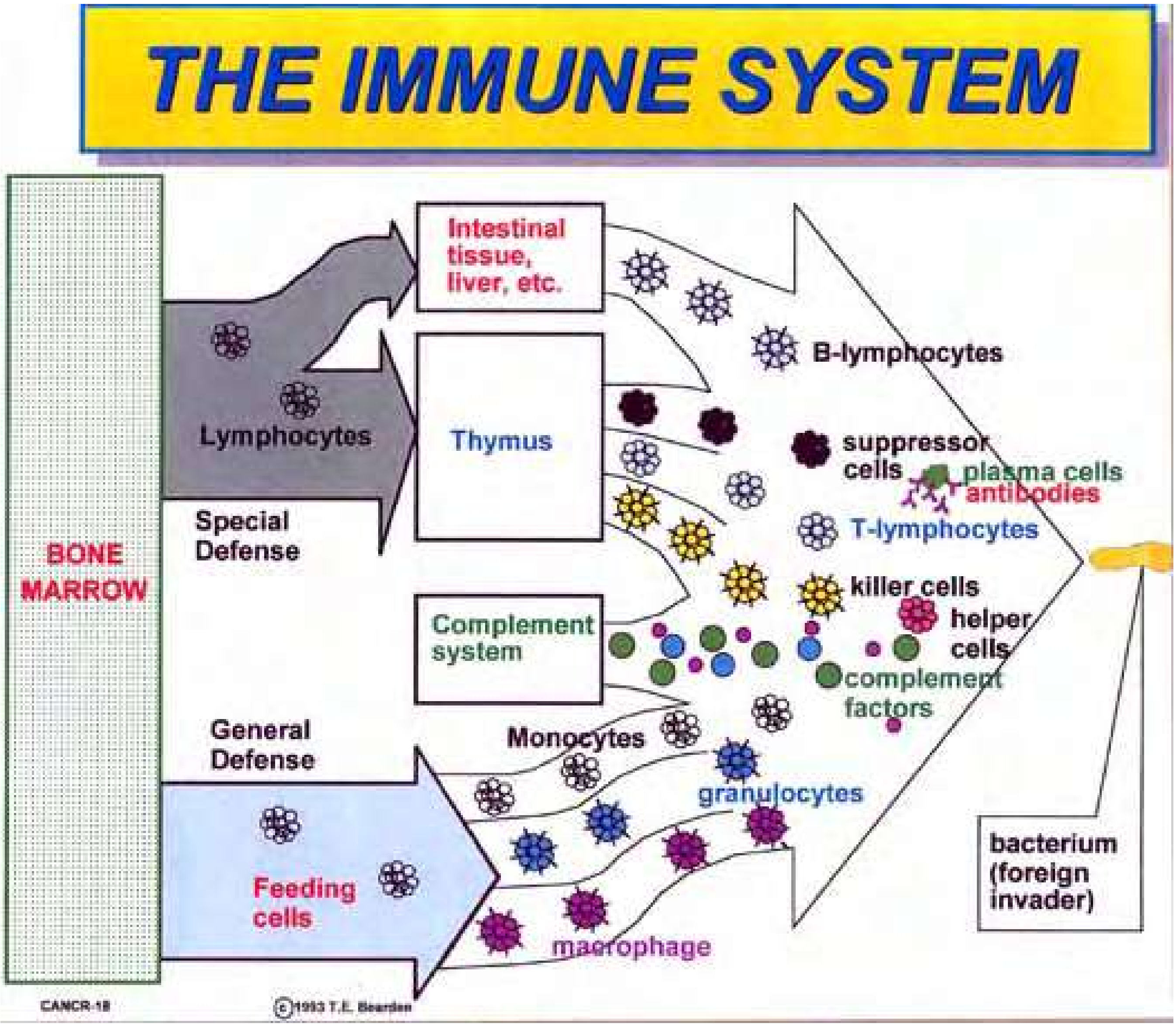

THE IMMUNE SYSTEM

BONE MARROW
Lymphocytes
Special Defense
Intestinal tissue, liver, etc.
Thymus
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
Complement system
complement factors
General Defense
Monocytes
granulocytes
Feeding cells
macrophage
bacterium (foreign invader)

CANCR-18

©1993 T.E. Bearden

[Return to Slide Index]
[Next Slide]
[Previous Slide]

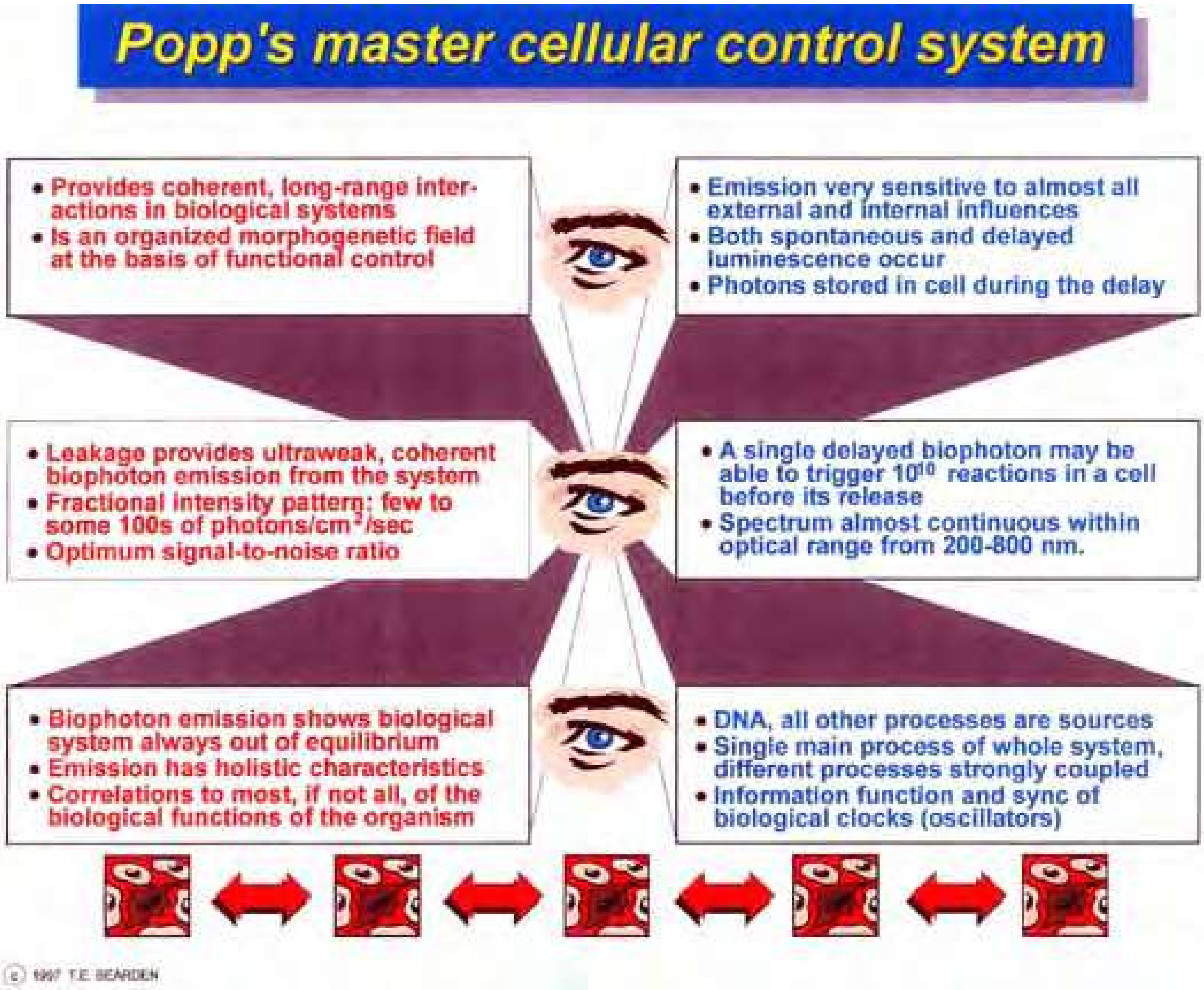

Popp's master cellular control system

- Provides coherent, long-range interactions in biological systems
- Is an organized morphogenetic field at the basis of functional control

- Emission very sensitive to almost all external and internal influences
- Both spontaneous and delayed luminescence occur
- Photons stored in cell during the delay

- Leakage provides ultraweak, coherent biophoton emission from the system
- Fractional intensity pattern: few to some 100s of photons/cm^2/sec
- Optimum signal-to-noise ratio

- A single delayed biophoton may be able to trigger 10^{10} reactions in a cell before its release
- Spectrum almost continuous within optical range from 200-800 nm.

- Biophoton emission shows biological system always out of equilibrium
- Emission has holistic characteristics
- Correlations to most, if not all, of the biological functions of the organism

- DNA, all other processes are sources
- Single main process of whole system, different processes strongly coupled
- Information function and sync of biological clocks (oscillators)

(c) 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

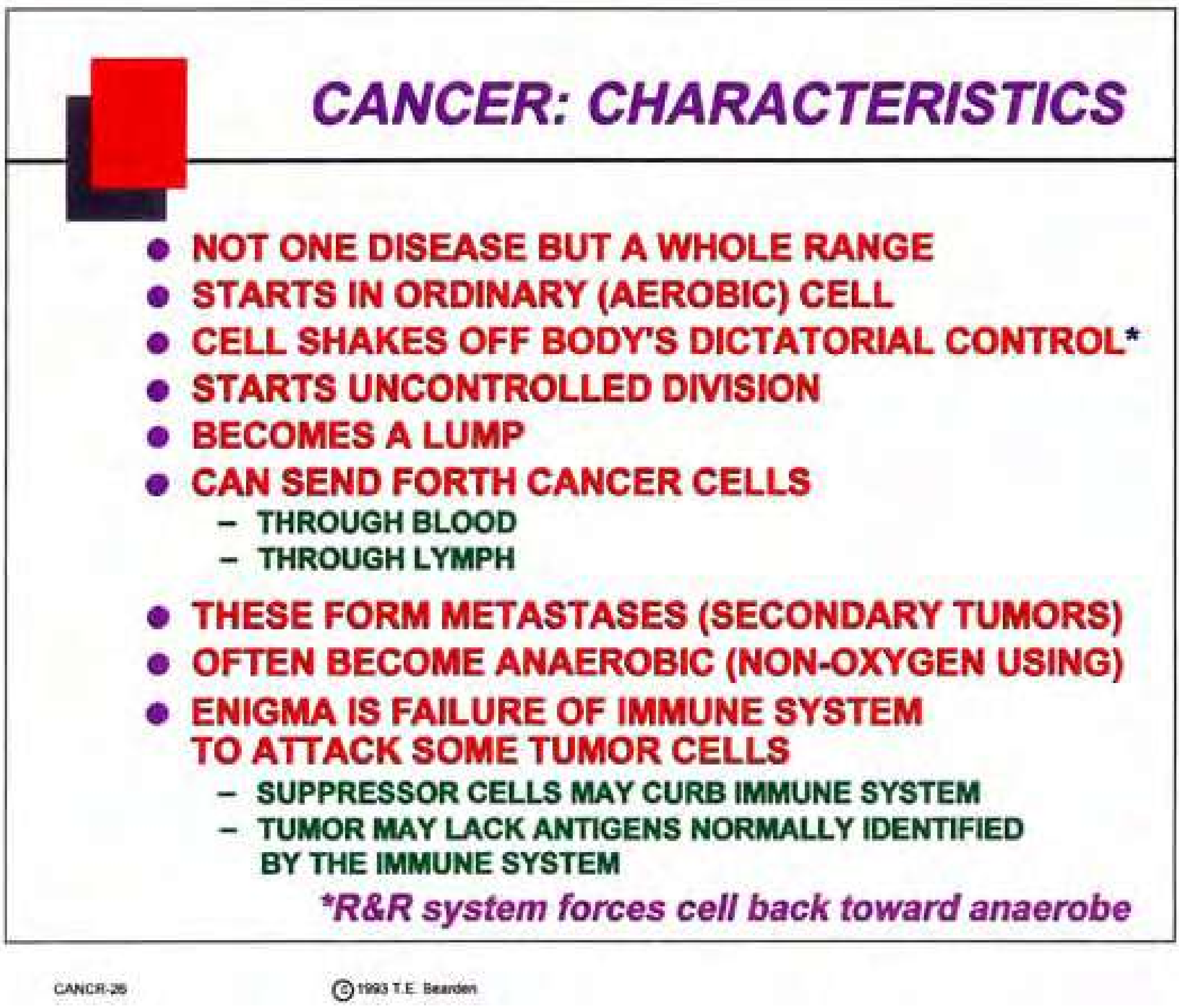

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

**R&R system forces cell back toward anaerobe*

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

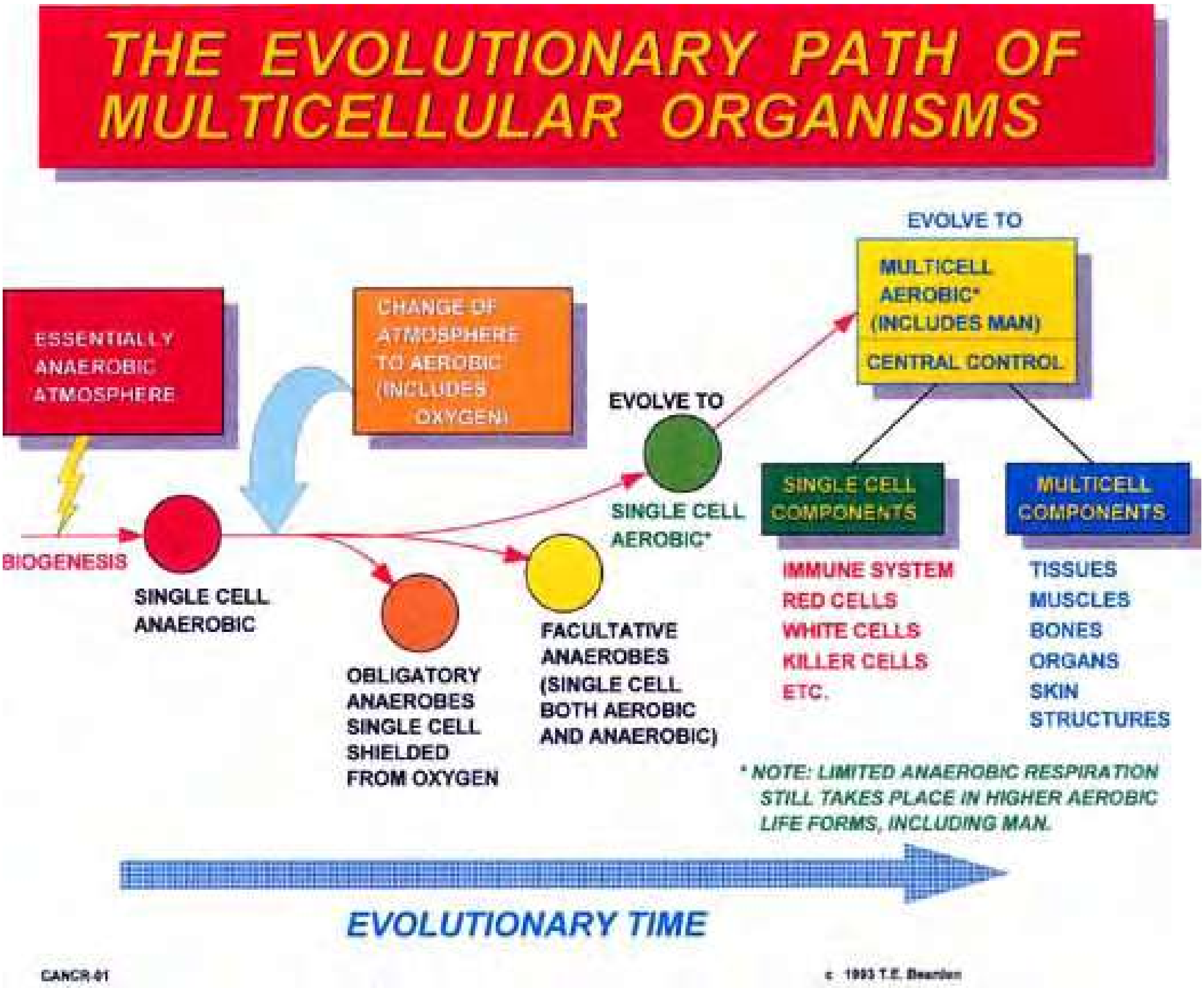

THE EVOLUTIONARY PATH OF MULTICELLULAR ORGANISMS

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

*NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

CANCR-01

c 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

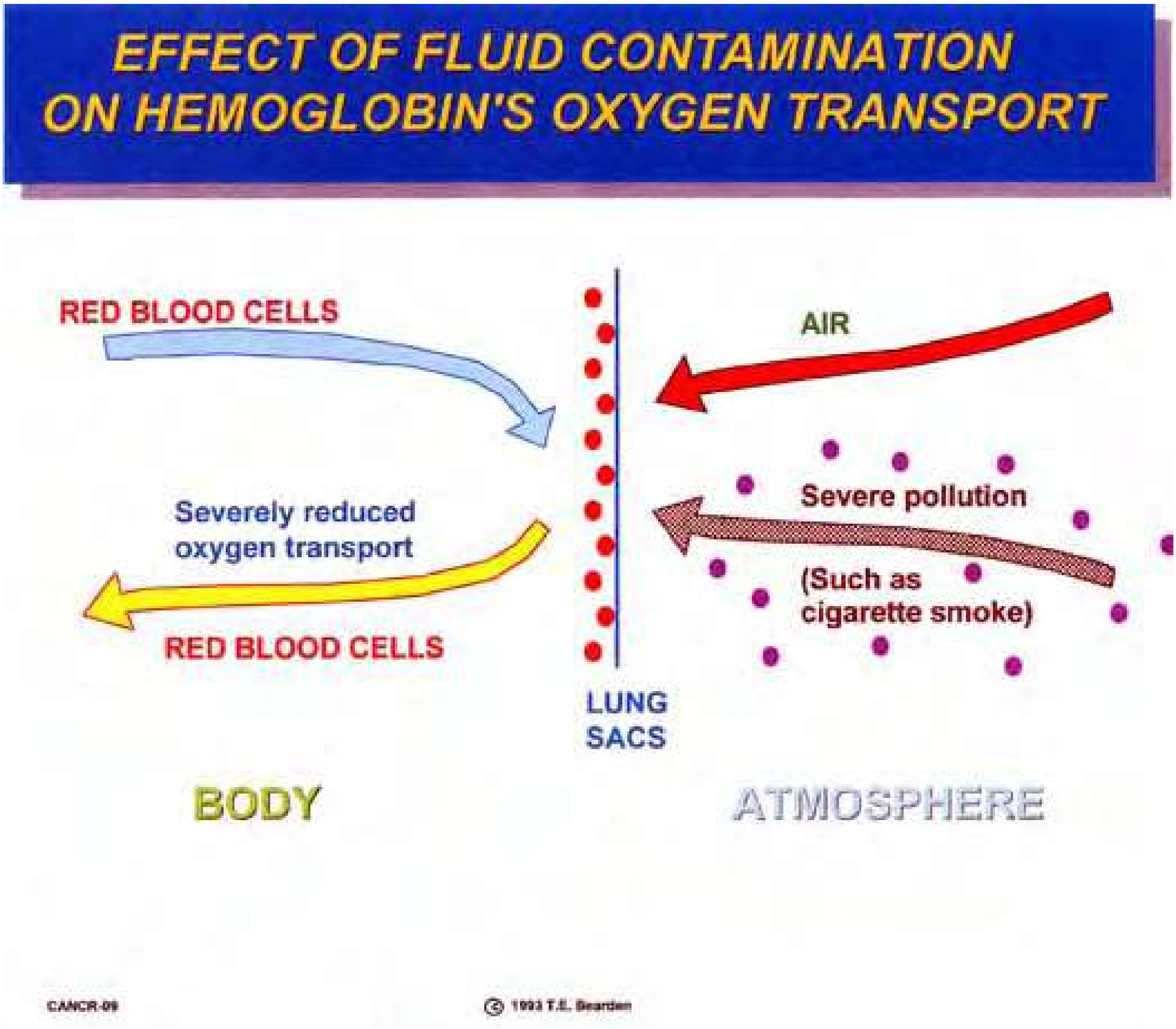

EFFECT OF FLUID CONTAMINATION ON HEMOGLOBIN'S OXYGEN TRANSPORT

RED BLOOD CELLS

AIR

Severe pollution

(Such as cigarette smoke)

Severely reduced oxygen transport

RED BLOOD CELLS

LUNG SACS

BODY

ATMOSPHERE

CANCR-09

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

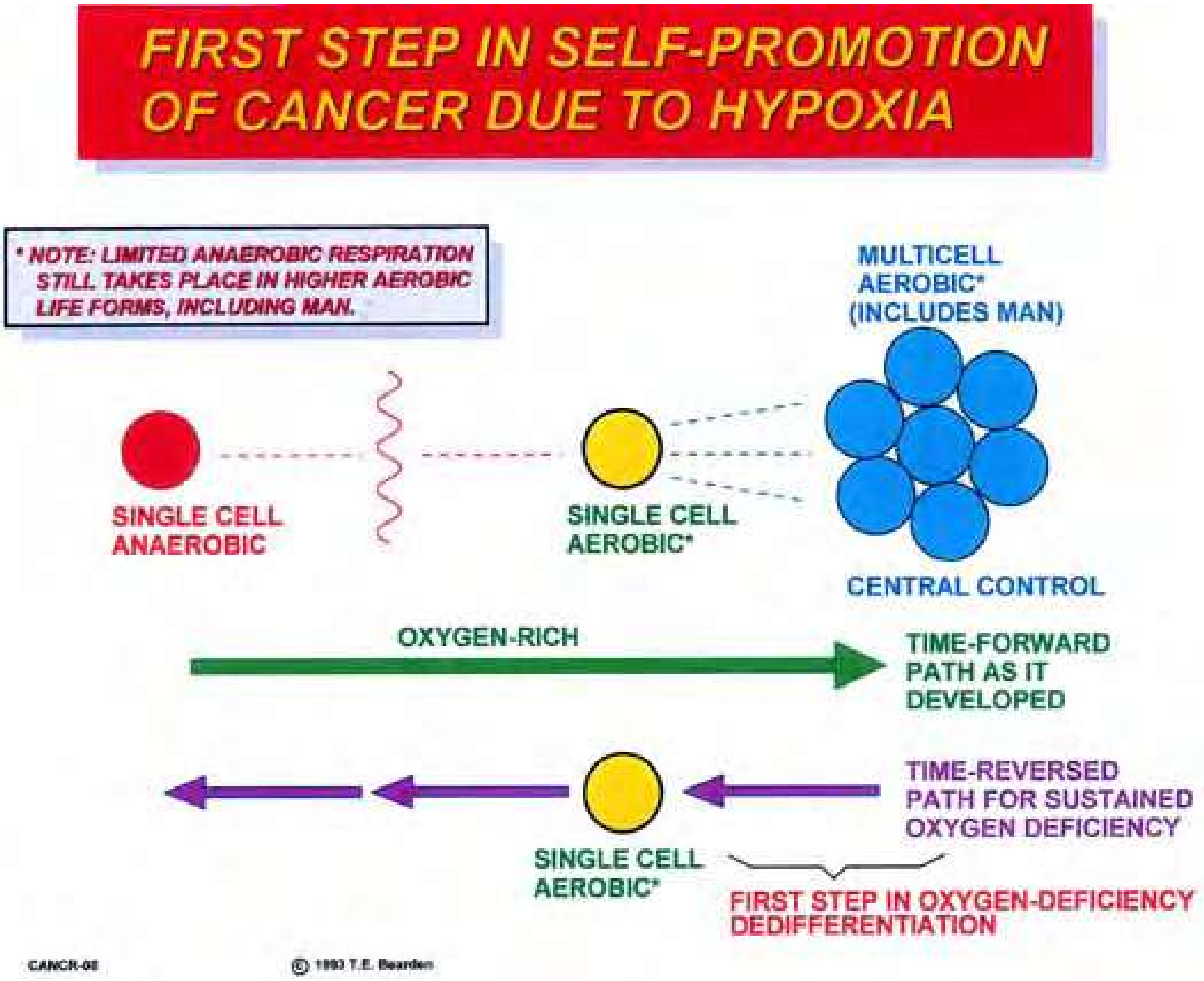

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

SINGLE CELL ANAEROBIC

SINGLE CELL AEROBIC*

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD PATH AS IT DEVELOPED

TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY

SINGLE CELL AEROBIC*

FIRST STEP IN OXYGEN-DEFICIENCY DEDIFFERENTIATION

CANCR-08

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

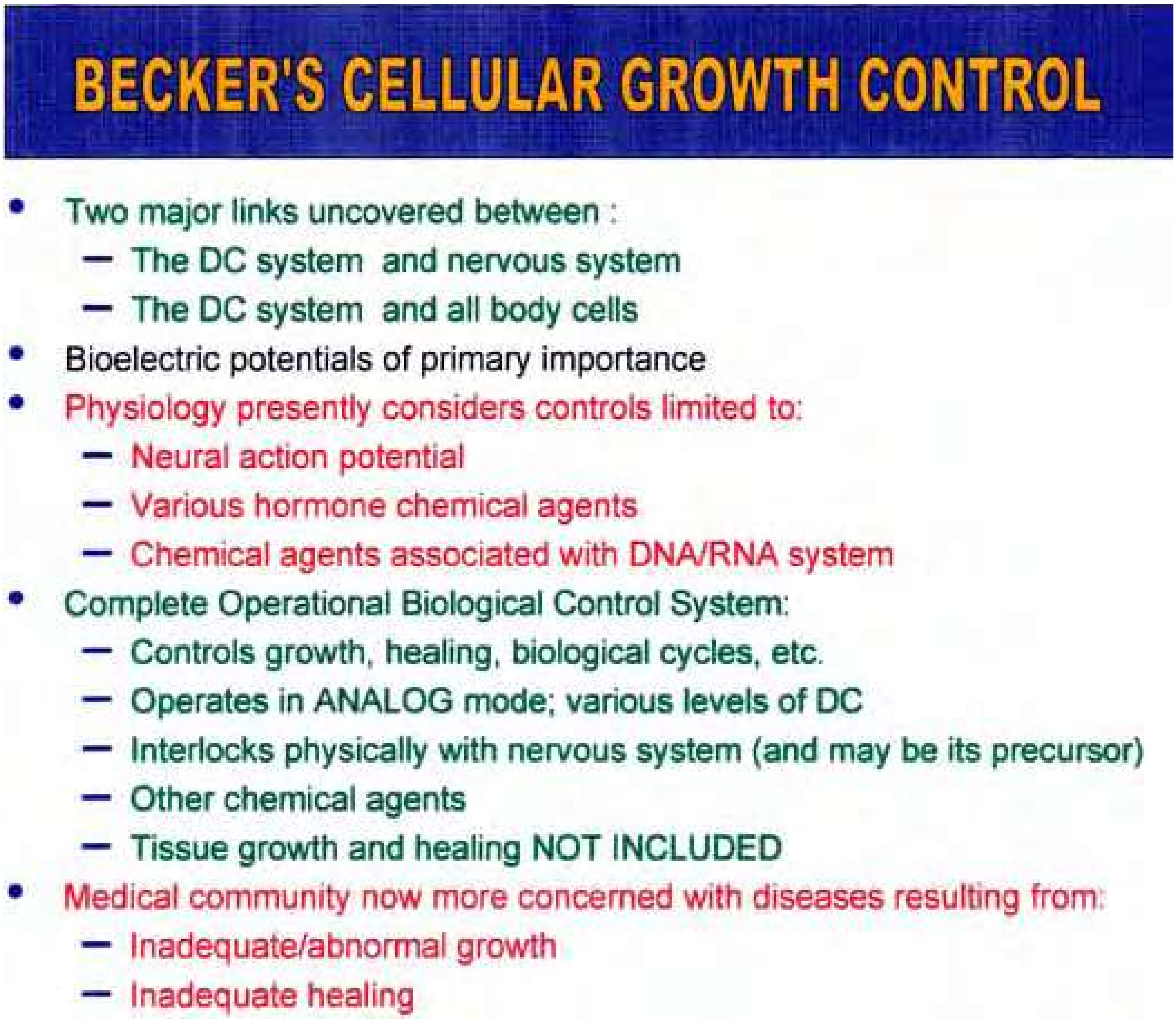

BECKER'S CELLULAR GROWTH CONTROL

- Two major links uncovered between :
  - The DC system and nervous system
  - The DC system and all body cells
- Bioelectric potentials of primary importance
- Physiology presently considers controls limited to:
  - Neural action potential
  - Various hormone chemical agents
  - Chemical agents associated with DNA/RNA system
- Complete Operational Biological Control System:
  - Controls growth, healing, biological cycles, etc.
  - Operates in ANALOG mode; various levels of DC
  - Interlocks physically with nervous system (and may be its precursor)
  - Other chemical agents
  - Tissue growth and healing NOT INCLUDED
- Medical community now more concerned with diseases resulting from:
  - Inadequate/abnormal growth
  - Inadequate healing

Return to Slide Index

Next Slide

Previous Slide

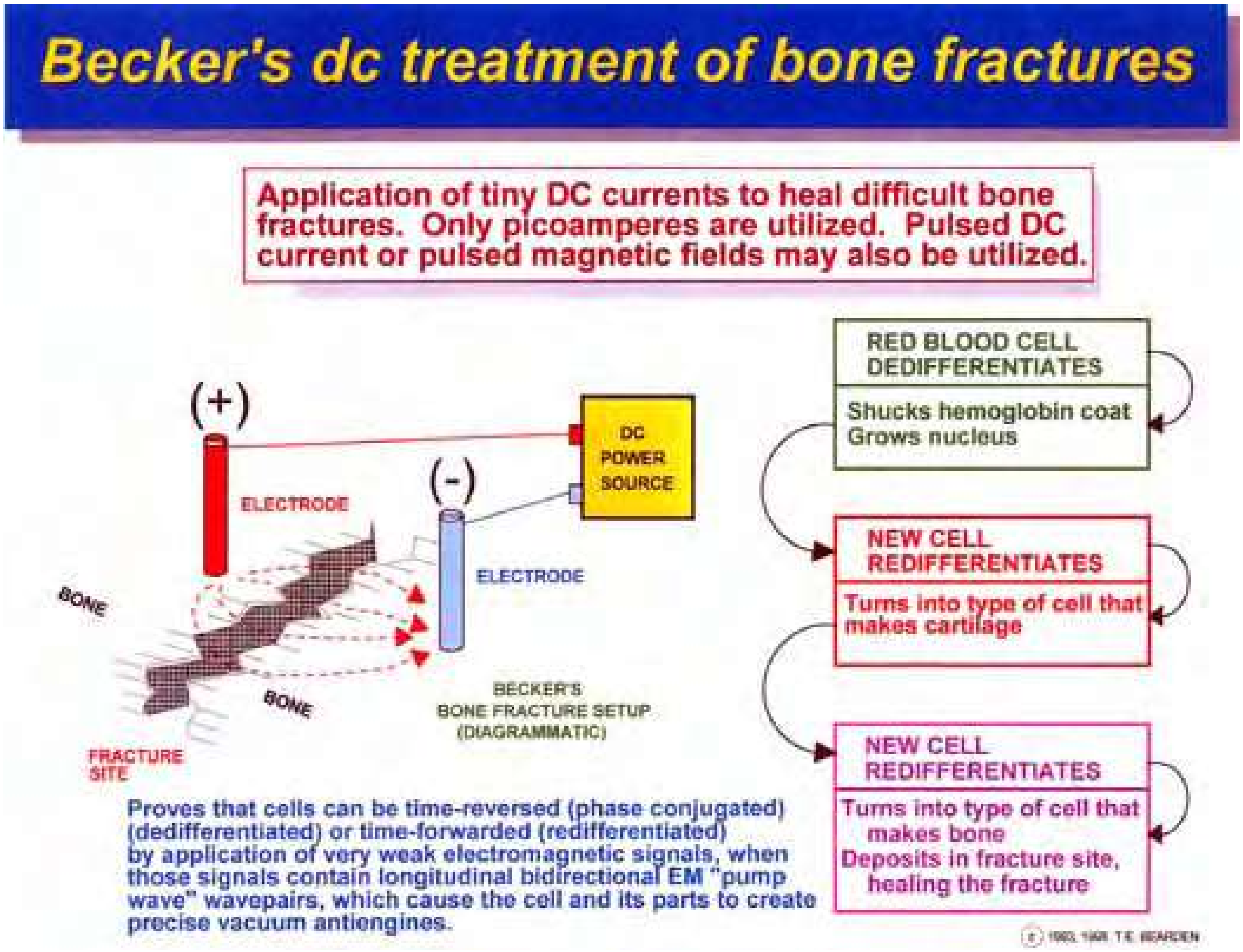

Becker's dc treatment of bone fractures

Application of tiny DC currents to heal difficult bone fractures. Only picoamperes are utilized. Pulsed DC current or pulsed magnetic fields may also be utilized.

(+)
ELECTRODE
(-)
DC POWER SOURCE
ELECTRODE
BONE
BONE
FRACTURE SITE

BECKER'S BONE FRACTURE SETUP (DIAGRAMMATIC)

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Proves that cells can be time-reversed (phase conjugated) (dedifferentiated) or time-forwarded (redifferentiated) by application of very weak electromagnetic signals, when those signals contain longitudinal bidirectional EM "pump wave" wavepairs, which cause the cell and its parts to create precise vacuum antiengines.

© 1993, 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

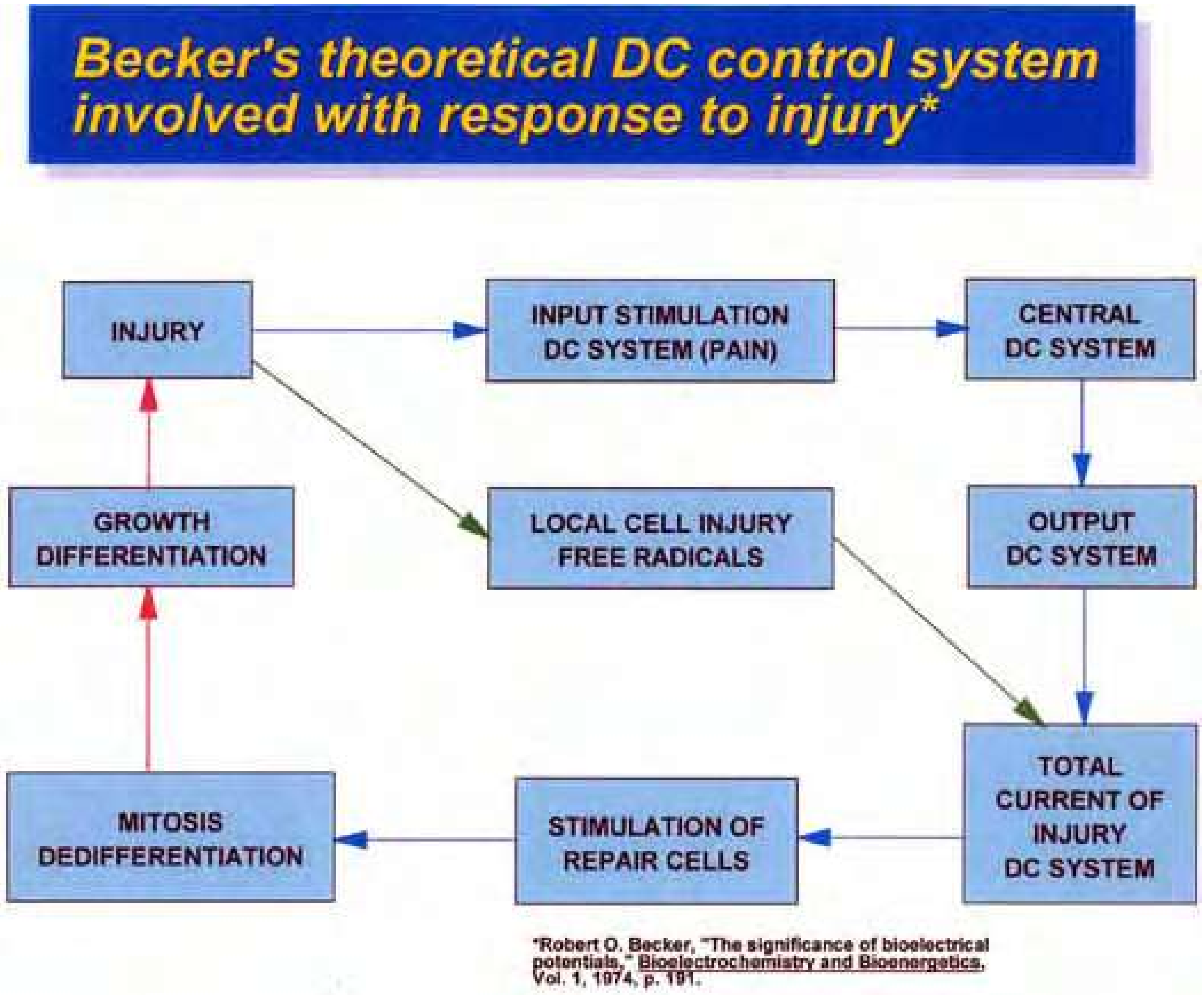

*Becker's theoretical DC control system involved with response to injury**

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

MITOSIS DEDIFFERENTIATION

STIMULATION OF REPAIR CELLS

TOTAL CURRENT OF INJURY DC SYSTEM

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

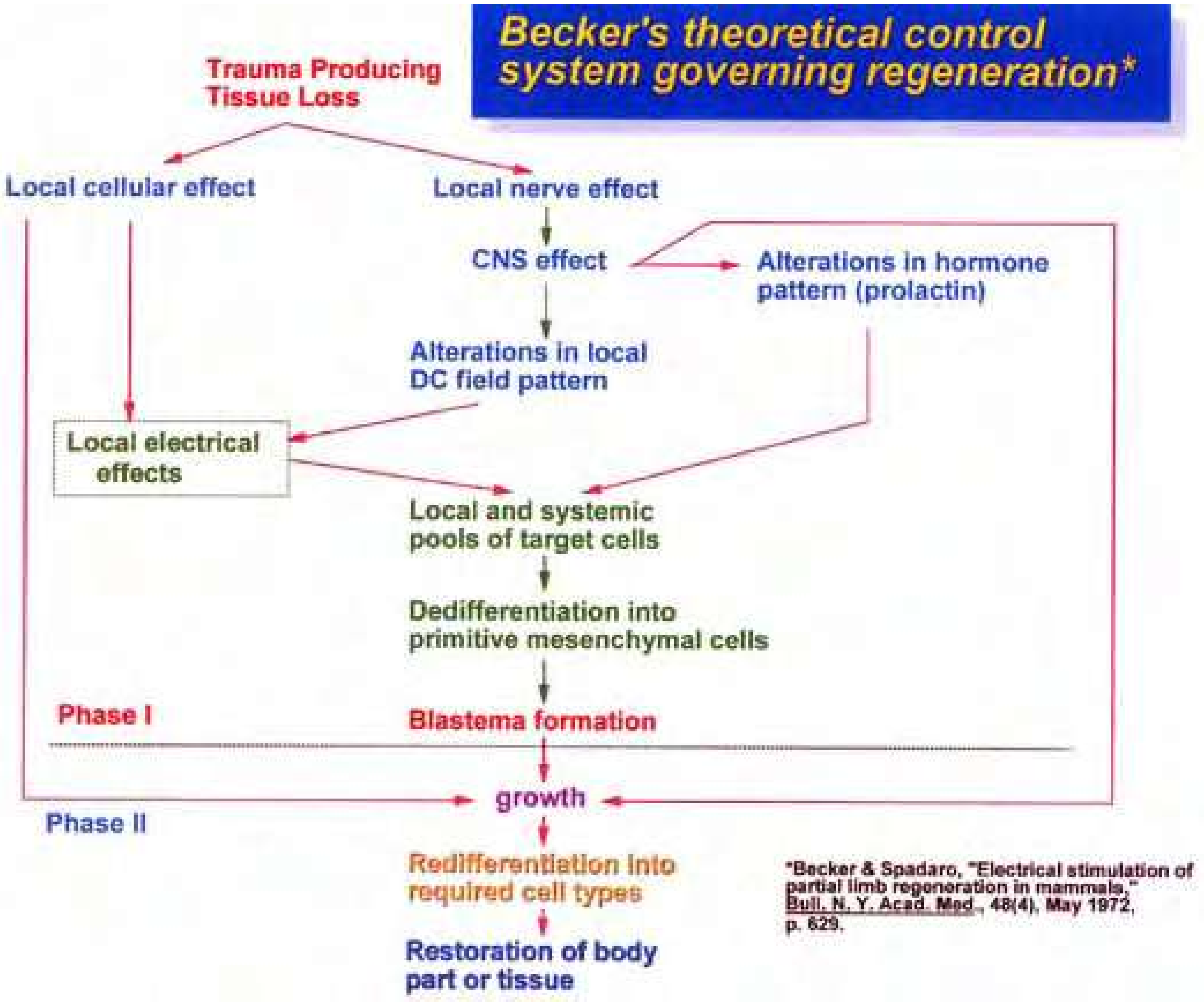

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Return to Slide Index

Next Slide

Previous Slide

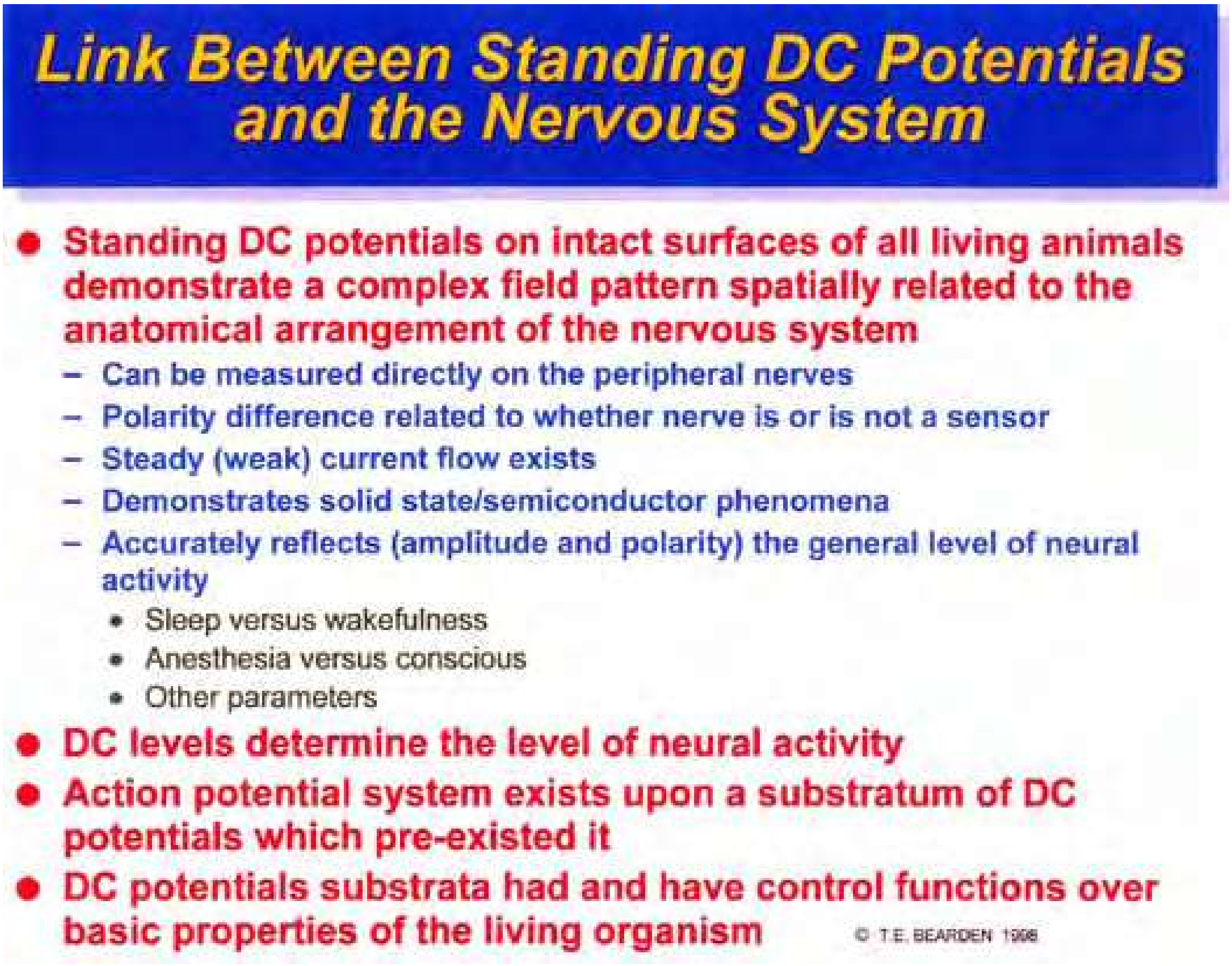

Link Between Standing DC Potentials and the Nervous System

- Standing DC potentials on intact surfaces of all living animals demonstrate a complex field pattern spatially related to the anatomical arrangement of the nervous system
  - Can be measured directly on the peripheral nerves
  - Polarity difference related to whether nerve is or is not a sensor
  - Steady (weak) current flow exists
  - Demonstrates solid state/semiconductor phenomena
  - Accurately reflects (amplitude and polarity) the general level of neural activity
    - Sleep versus wakefulness
    - Anesthesia versus conscious
    - Other parameters
- DC levels determine the level of neural activity
- Action potential system exists upon a substratum of DC potentials which pre-existed it
- DC potentials substrata had and have control functions over basic properties of the living organism

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

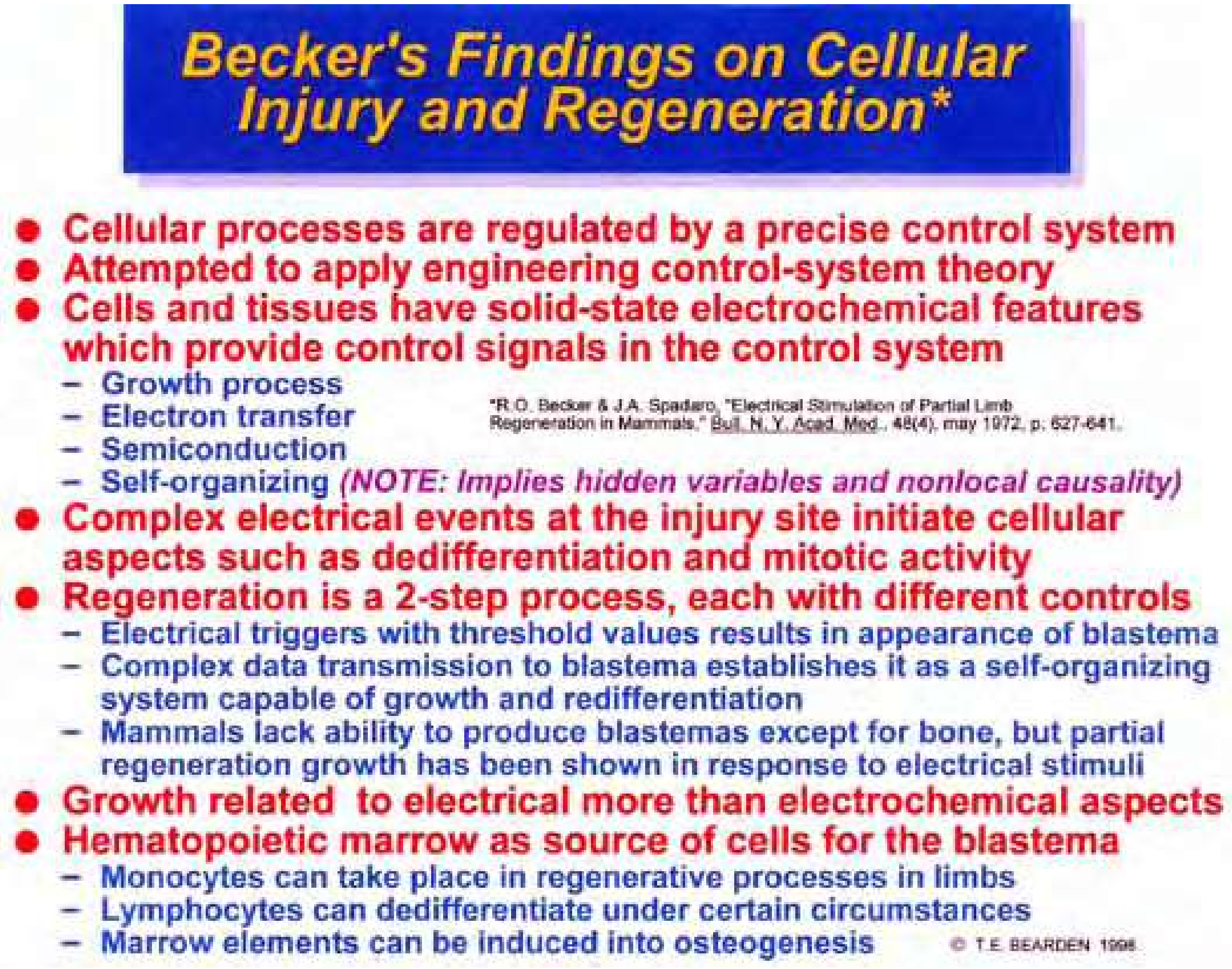

Becker's Findings on Cellular Injury and Regeneration*

- Cellular processes are regulated by a precise control system
- Attempted to apply engineering control-system theory
- Cells and tissues have solid-state electrochemical features which provide control signals in the control system
  - Growth process
  - Electron transfer
  - Semiconduction
  - Self-organizing *(NOTE: Implies hidden variables and nonlocal causality)*
- Complex electrical events at the injury site initiate cellular aspects such as dedifferentiation and mitotic activity
- Regeneration is a 2-step process, each with different controls
  - Electrical triggers with threshold values results in appearance of blastema
  - Complex data transmission to blastema establishes it as a self-organizing system capable of growth and redifferentiation
  - Mammals lack ability to produce blastemas except for bone, but partial regeneration growth has been shown in response to electrical stimuli
- Growth related to electrical more than electrochemical aspects
- Hematopoietic marrow as source of cells for the blastema
  - Monocytes can take place in regenerative processes in limbs
  - Lymphocytes can dedifferentiate under certain circumstances
  - Marrow elements can be induced into osteogenesis

*R.O. Becker & J.A. Spadaro, "Electrical Stimulation of Partial Limb Regeneration in Mammals," Bull. N. Y. Acad. Med., 48(4), may 1972, p. 627-641.

© T.E. BEARDEN 1998

[Return to Slide Index]

[Next Slide]

[Previous Slide]

Corrections For the Present EM Bioeffects Model

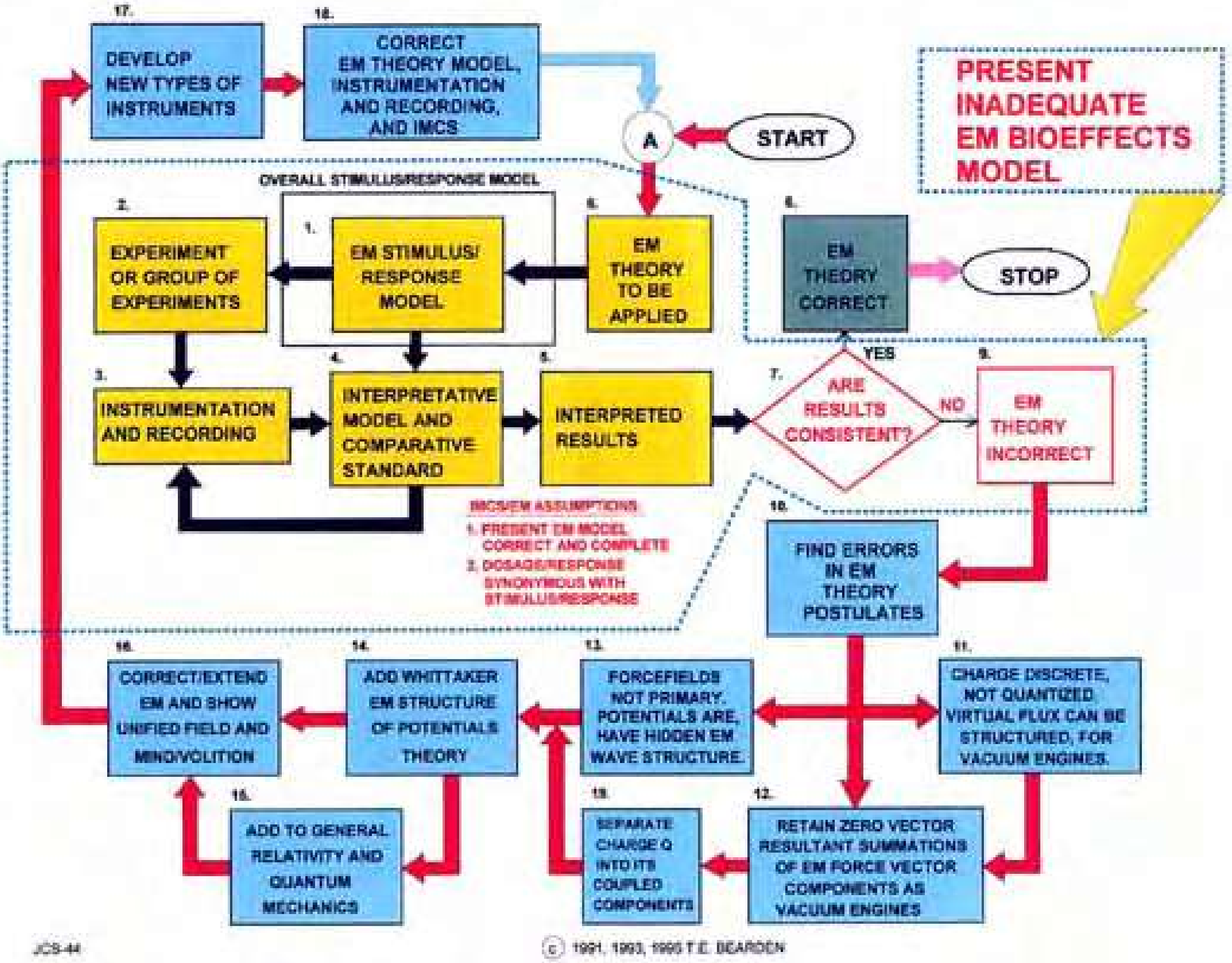

START

A

OVERALL STIMULUS/RESPONSE MODEL

1. EM STIMULUS/ RESPONSE MODEL

2. EXPERIMENT OR GROUP OF EXPERIMENTS

3. INSTRUMENTATION AND RECORDING

4. INTERPRETATIVE MODEL AND COMPARATIVE STANDARD

5. INTERPRETED RESULTS

6. EM THEORY TO BE APPLIED

7. ARE RESULTS CONSISTENT?

YES

NO

8. EM THEORY CORRECT

STOP

9. EM THEORY INCORRECT

PRESENT INADEQUATE EM BIOEFFECTS MODEL

IMCS/EM ASSUMPTIONS

1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE

10. FIND ERRORS IN EM THEORY POSTULATES

11. CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.

12. RETAIN ZERO VECTOR RESULTANT SUMMATIONS OF EM FORCE VECTOR COMPONENTS AS VACUUM ENGINES

13. FORCEFIELDS NOT PRIMARY. POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.

14. ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY

15. ADD TO GENERAL RELATIVITY AND QUANTUM MECHANICS

16. CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION

17. DEVELOP NEW TYPES OF INSTRUMENTS

18. CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS

19. SEPARATE CHARGE Q INTO ITS COUPLED COMPONENTS

JCS-44

© 1991, 1993, 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

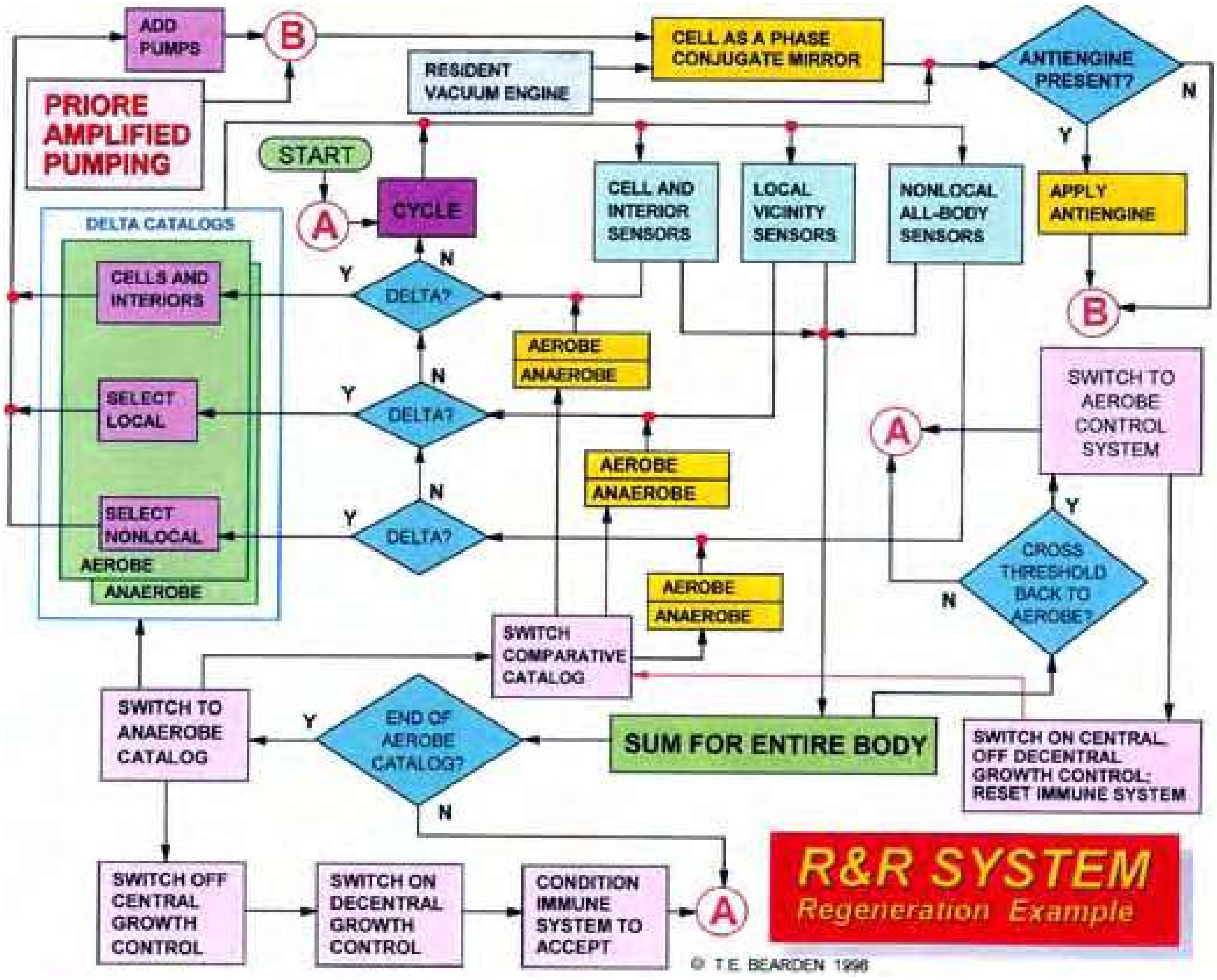
ADD PUMPS
B
CELL AS A PHASE CONJUGATE MIRROR
RESIDENT VACUUM ENGINE
ANTIENGINE PRESENT?
N
Y
APPLY ANTIENGINE
B
PRIORE AMPLIFIED PUMPING
START
A
CYCLE
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NONLOCAL ALL-BODY SENSORS
DELTA CATALOGS
CELLS AND INTERIORS
Y
DELTA?
N
AEROBE
ANAEROBE
SELECT LOCAL
Y
DELTA?
N
SWITCH TO AEROBE CONTROL SYSTEM
A
AEROBE
ANAEROBE
SELECT NONLOCAL
Y
DELTA?
AEROBE
ANAEROBE
AEROBE
ANAEROBE
CROSS THRESHOLD BACK TO AEROBE?
Y
N
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
Y
END OF AEROBE CATALOG?
N
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT
A
R&R SYSTEM
Regeneration Example
© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

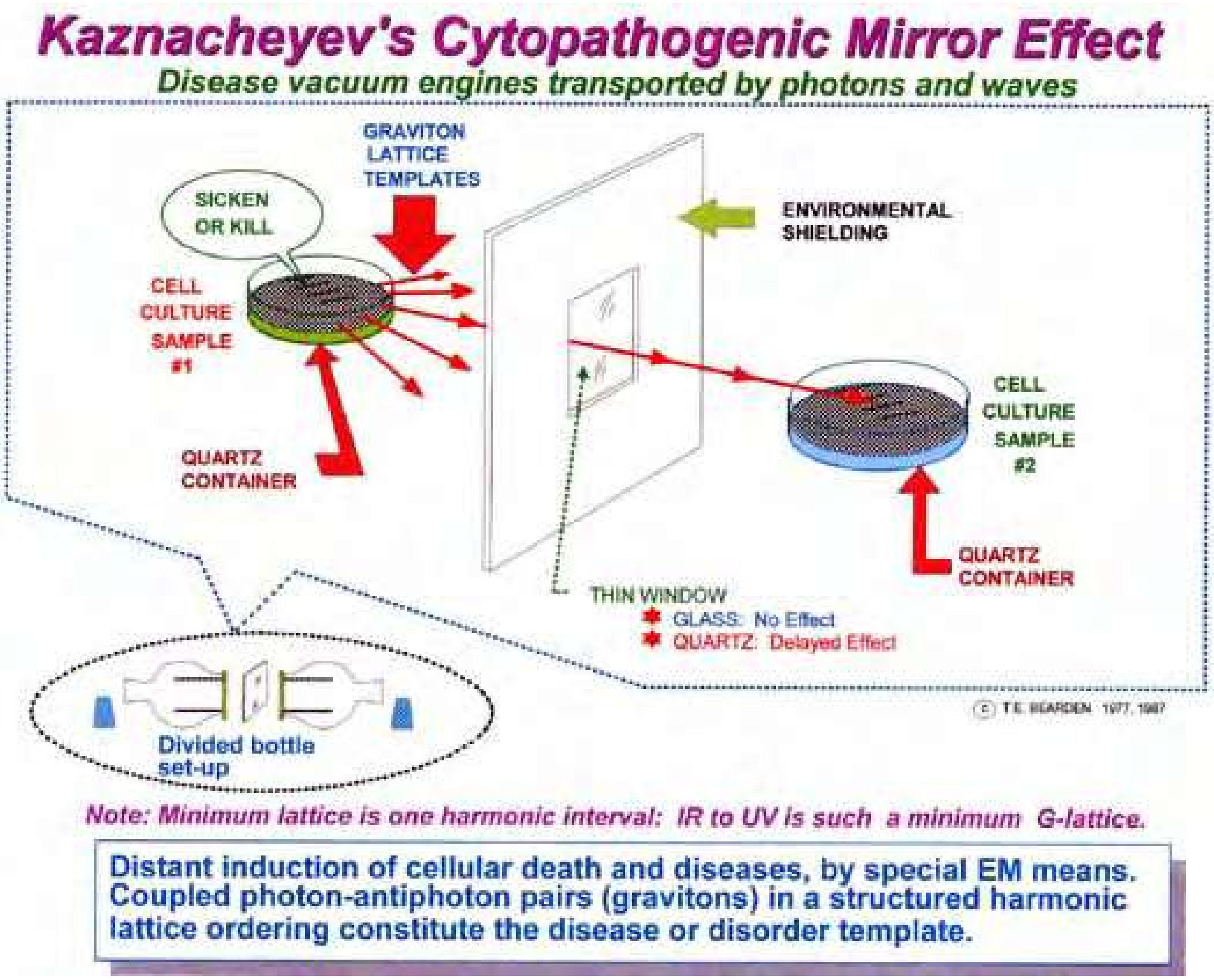

Kaznacheyev's Cytopathogenic Mirror Effect

Disease vacuum engines transported by photons and waves

GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
CELL
CULTURE
SAMPLE
#1
QUARTZ
CONTAINER
ENVIRONMENTAL
SHIELDING
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle
set-up

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Distant induction of cellular death and diseases, by special EM means. Coupled photon-antiphoton pairs (gravitons) in a structured harmonic lattice ordering constitute the disease or disorder template.

Return to Slide Index

Next Slide

Previous Slide

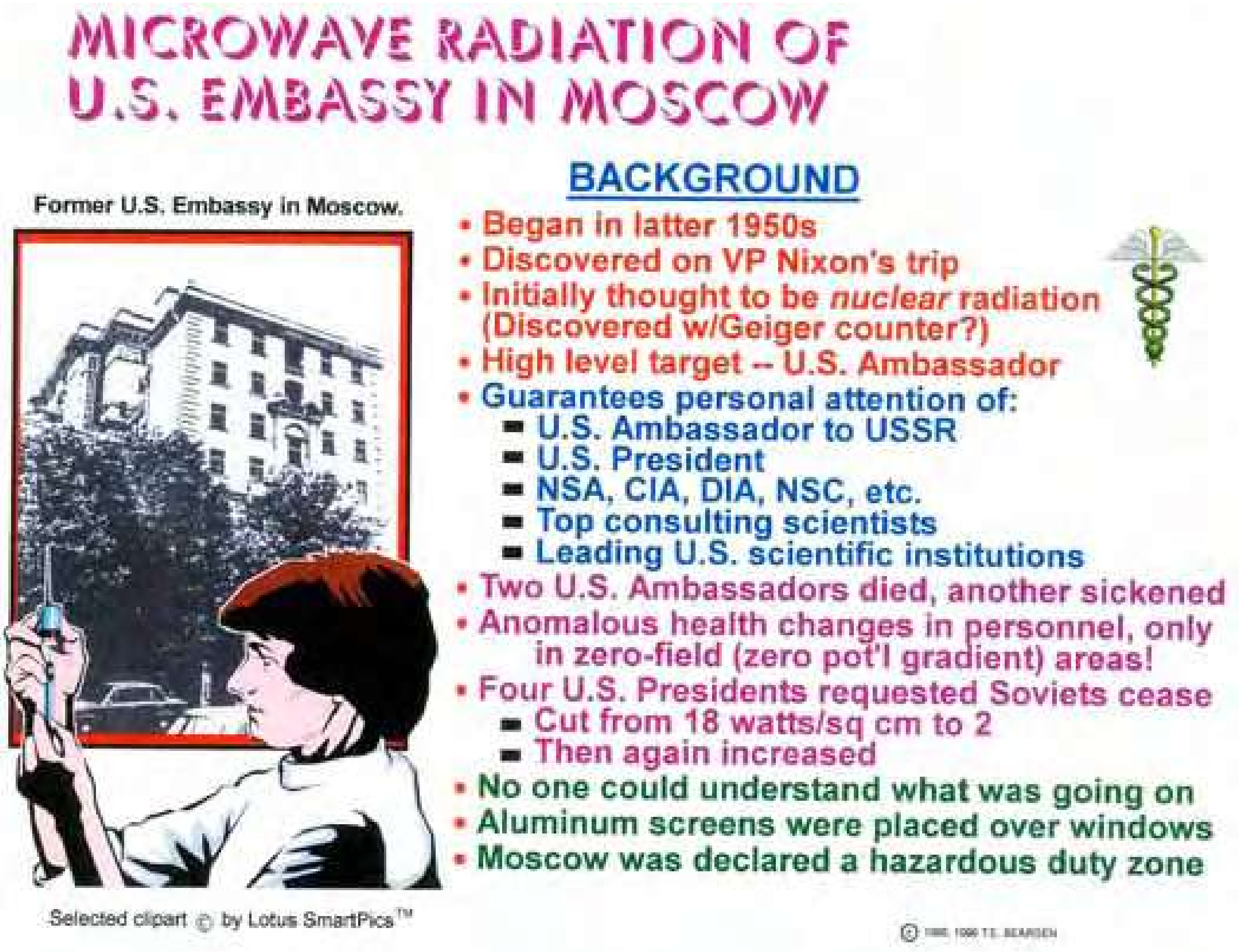

MICROWAVE RADIATION OF U.S. EMBASSY IN MOSCOW

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Selected clipart © by Lotus SmartPics™

© 1990, 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

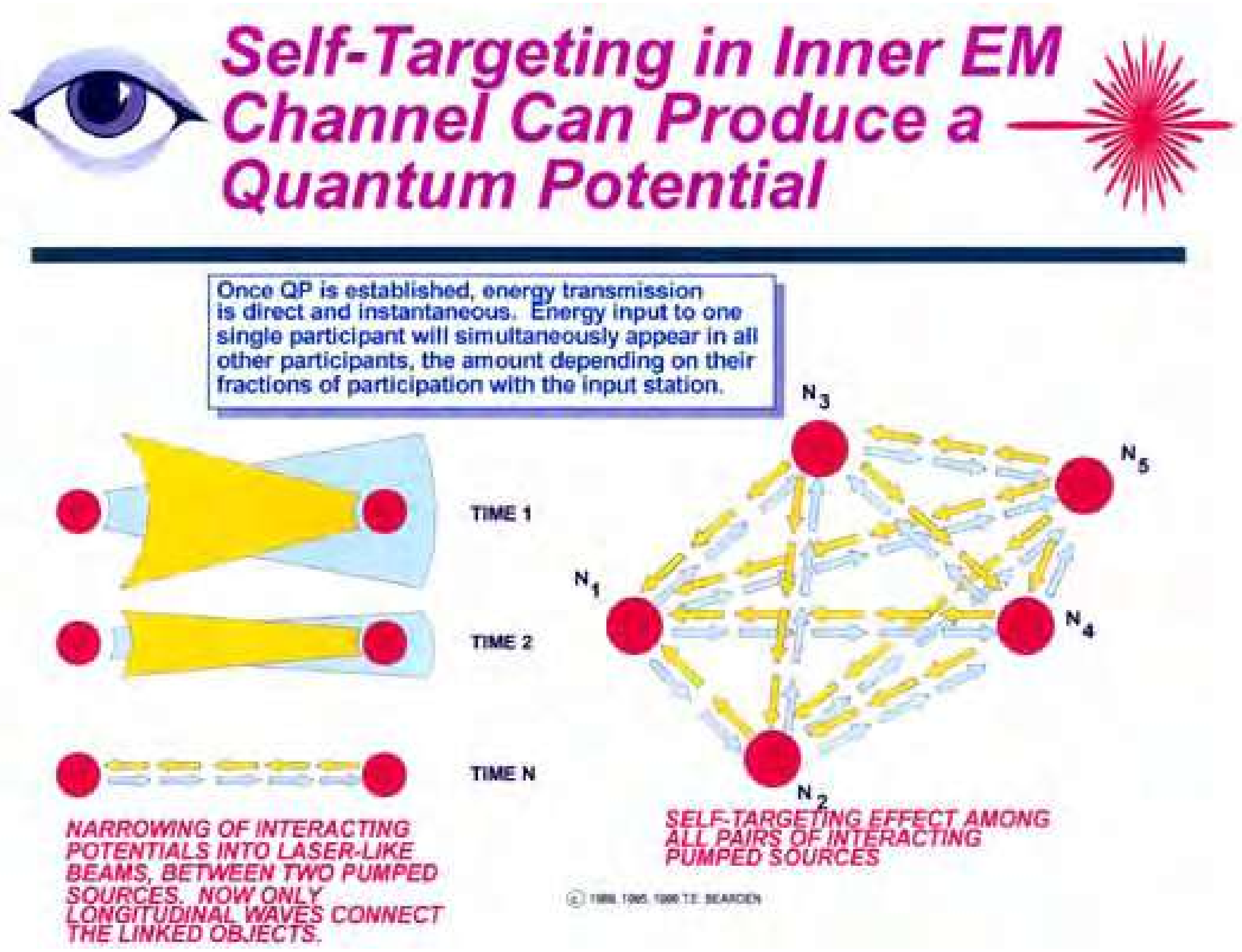

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© [illegible], 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

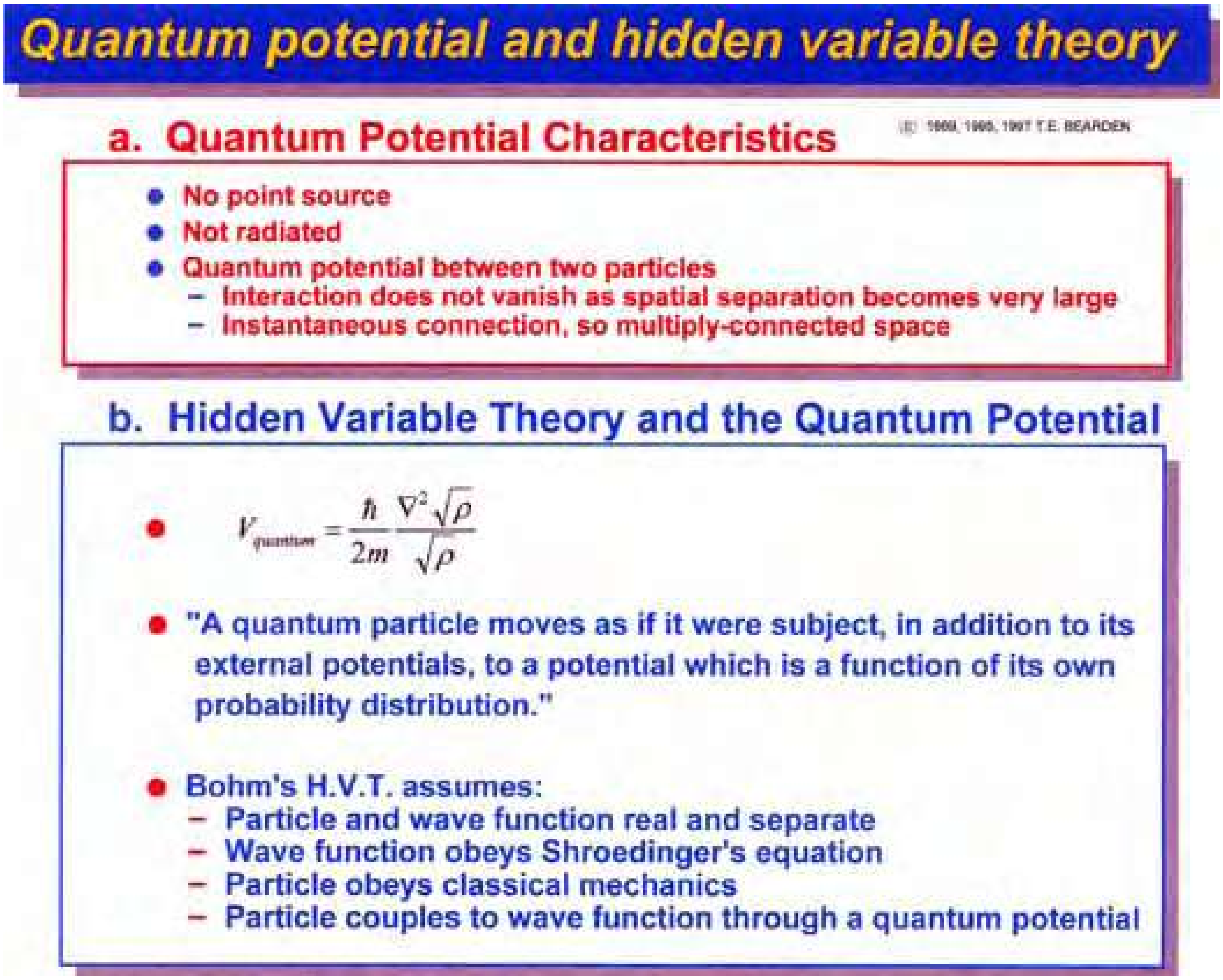

Quantum potential and hidden variable theory

a. Quantum Potential Characteristics

© 1988, 1995, 1997 T.E. BEARDEN

- No point source
- Not radiated
- Quantum potential between two particles
  - Interaction does not vanish as spatial separation becomes very large
  - Instantaneous connection, so multiply-connected space

b. Hidden Variable Theory and the Quantum Potential

- $$V_{quantum} = \frac{\hbar}{2m} \frac{\nabla^2 \sqrt{\rho}}{\sqrt{\rho}}$$
- "A quantum particle moves as if it were subject, in addition to its external potentials, to a potential which is a function of its own probability distribution."
- Bohm's H.V.T. assumes:
  - Particle and wave function real and separate
  - Wave function obeys Shroedinger's equation
  - Particle obeys classical mechanics
  - Particle couples to wave function through a quantum potential

Return to Slide Index

Next Slide

Previous Slide

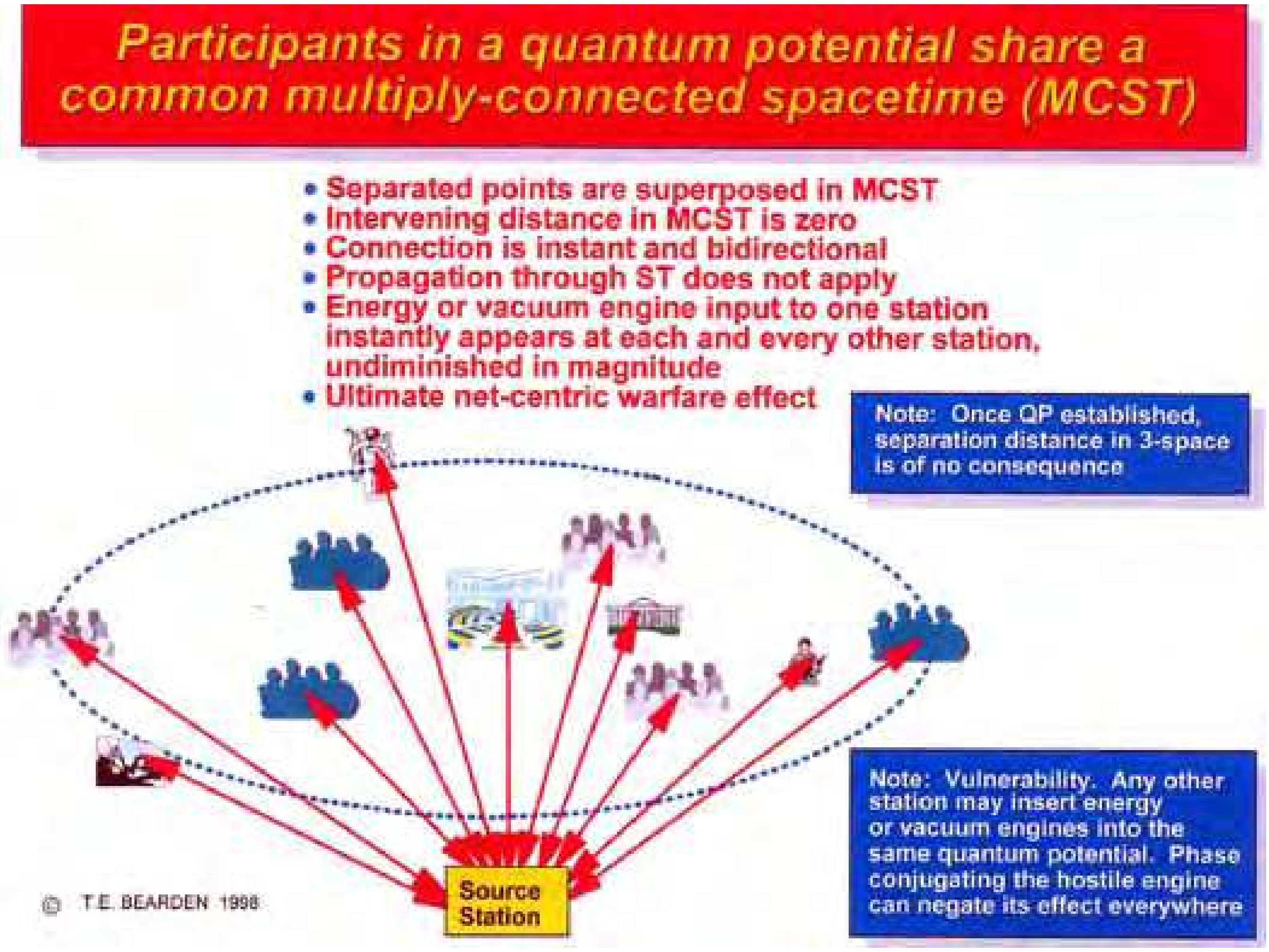

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Intervening distance in MCST is zero
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude
- Ultimate net-centric warfare effect

Note: Once QP established, separation distance in 3-space is of no consequence

Note: Vulnerability. Any other station may insert energy or vacuum engines into the same quantum potential. Phase conjugating the hostile engine can negate its effect everywhere

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

Gulf War Syndrome And Typical Comparative Disease Curves

Note: There is cumulative genetic damage to germ cells also.

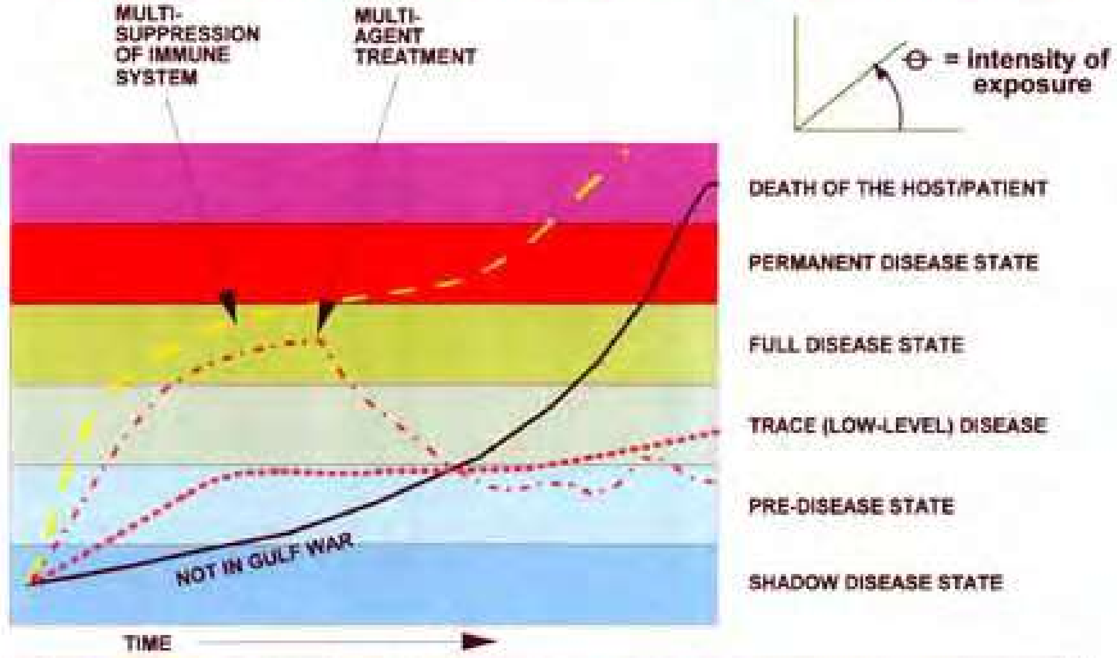

EXPOSED WITH MOST CUMULATIVE DAMAGE
EXPOSED WITH MULTIPLE AGENT INFECTIONS AND RECEIVING MULTITREATMENT
EXPOSED WITH LESS CUMULATIVE DAMAGE
NON-GULF WAR: TYPICAL NORMAL DISEASE WHICH KILLS PATIENT

JCS-100

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

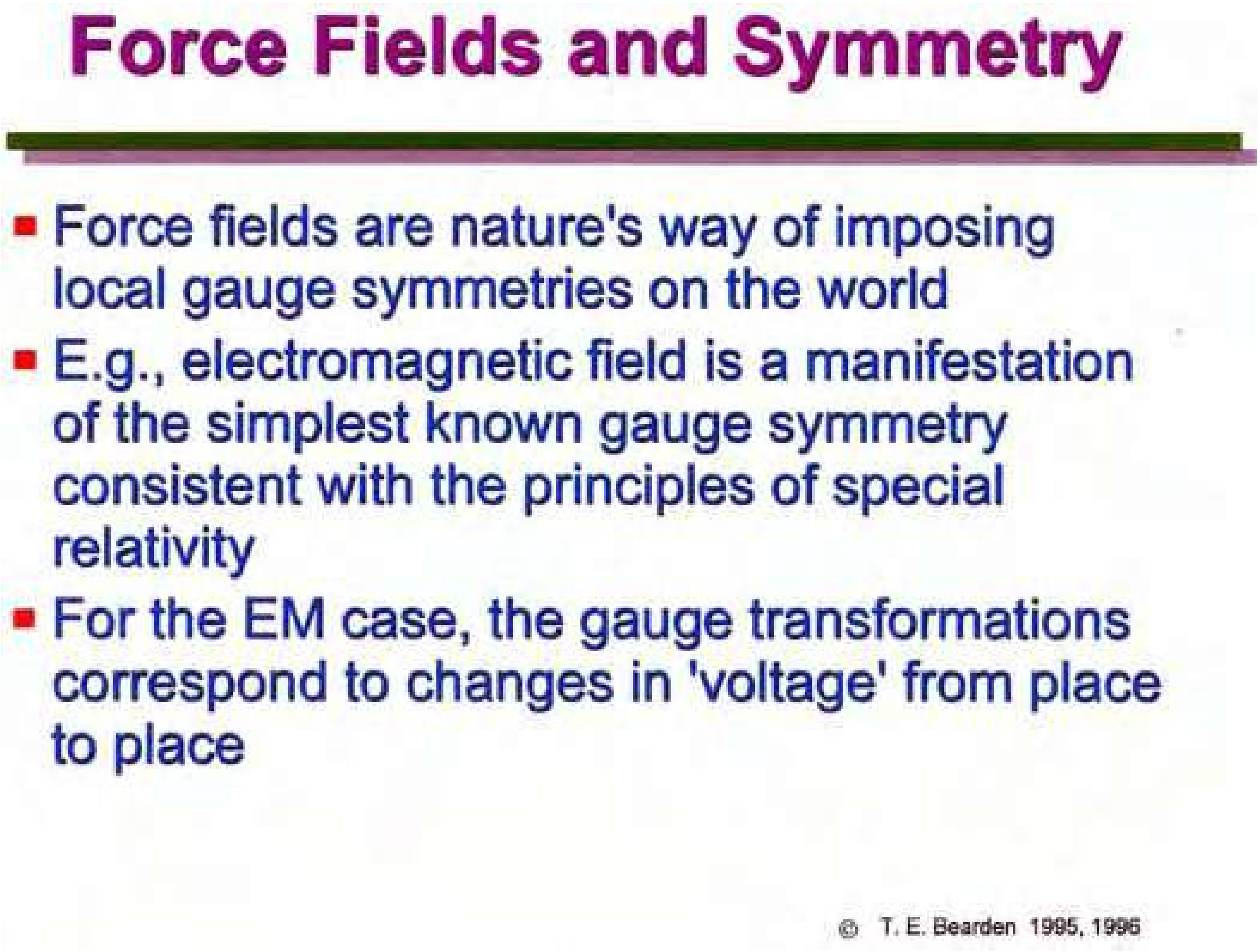

Force Fields and Symmetry

- Force fields are nature's way of imposing local gauge symmetries on the world
- E.g., electromagnetic field is a manifestation of the simplest known gauge symmetry consistent with the principles of special relativity
- For the EM case, the gauge transformations correspond to changes in 'voltage' from place to place

© T. E. Bearden 1995, 1996

Return to Slide Index

Next Slide

Previous Slide

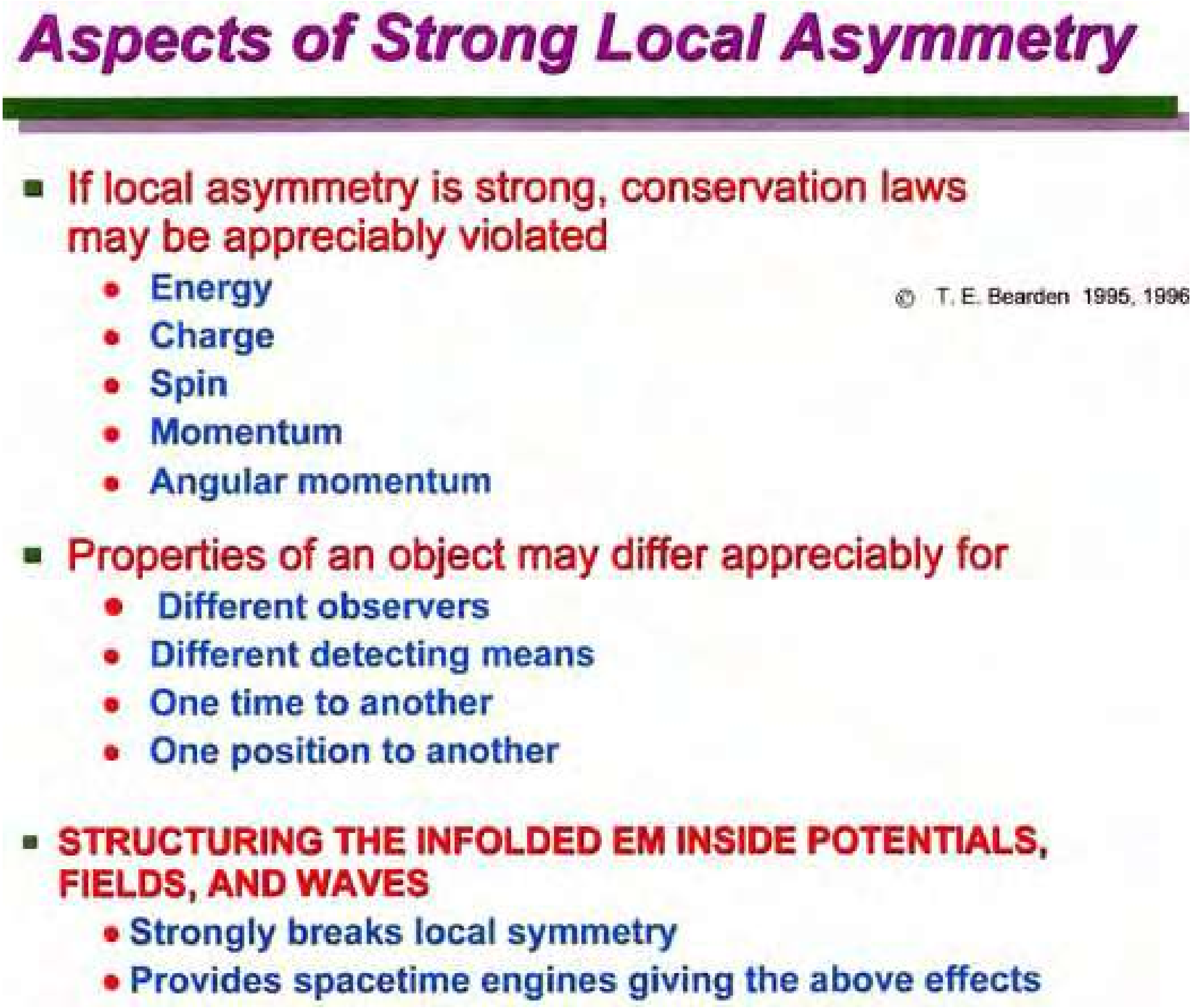

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

[Return to Slide Index]

[Next Slide]

[Previous Slide]

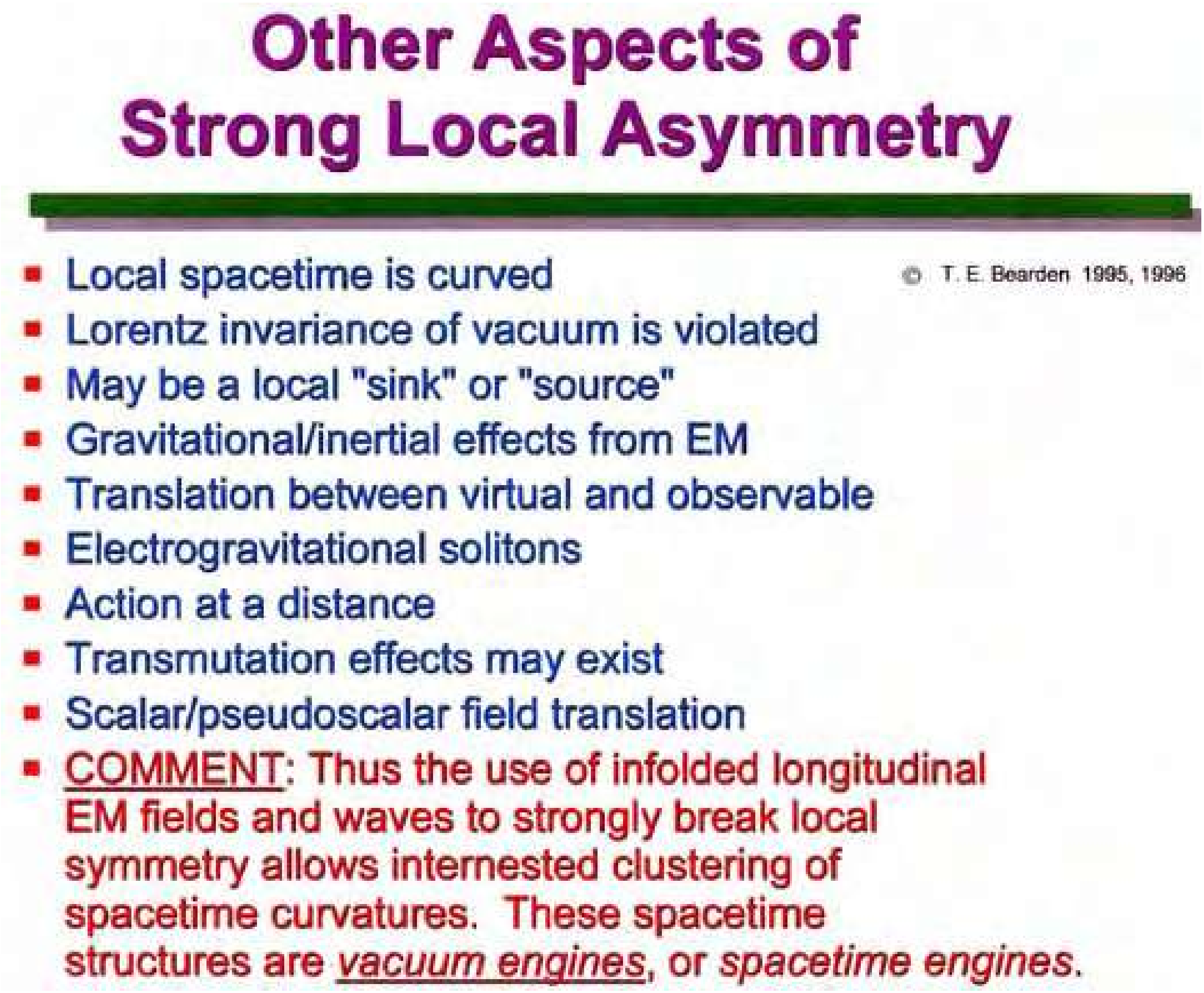

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational/inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

Return to Slide Index

Next Slide

Previous Slide

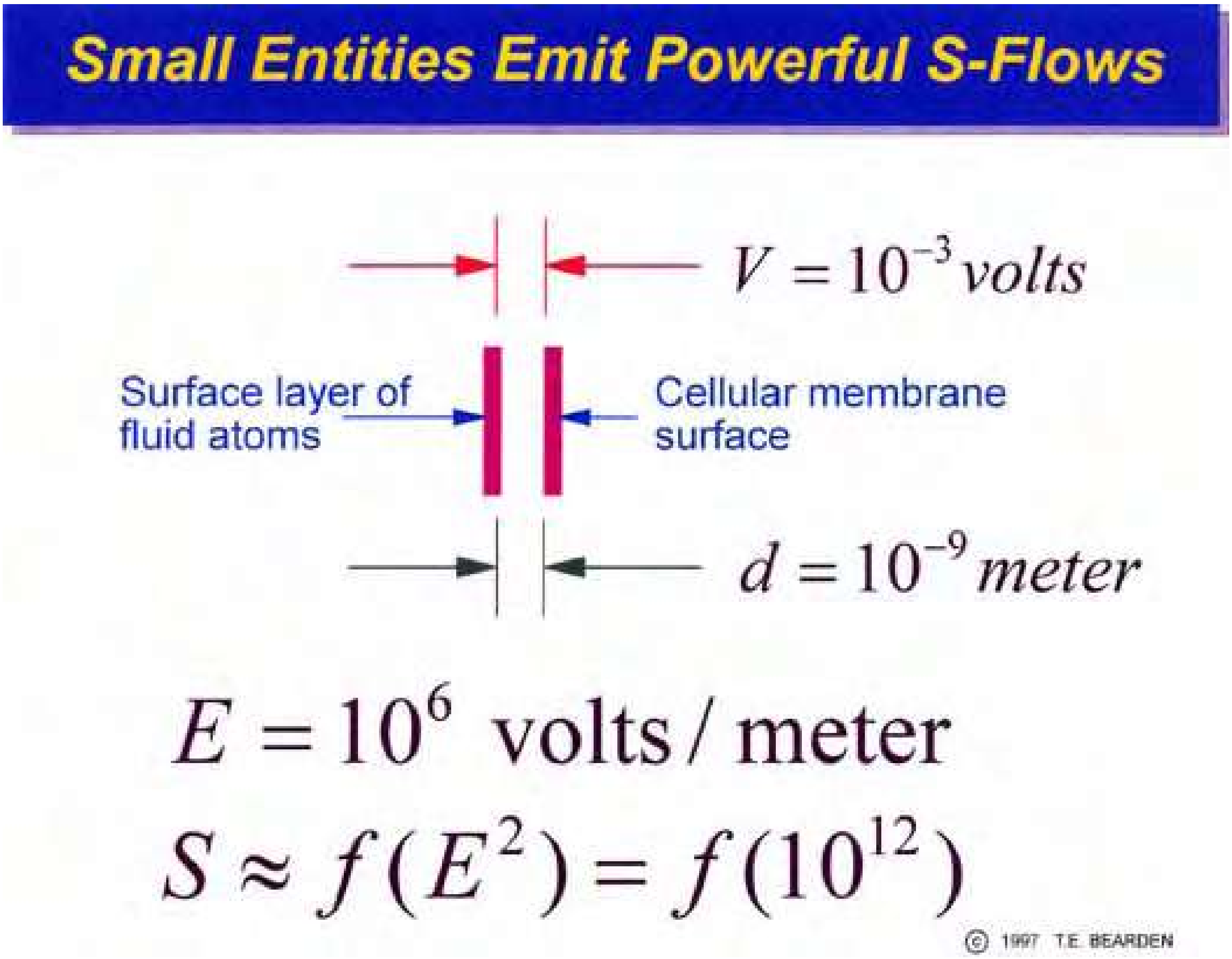

Small Entities Emit Powerful S-Flows

$V = 10^{-3}\,volts$

Surface layer of fluid atoms

Cellular membrane surface

$d = 10^{-9}\,meter$

$$E = 10^{6}\ \text{volts / meter}$$

$$S \approx f(E^{2}) = f(10^{12})$$

© 1997 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

[Previous Slide]

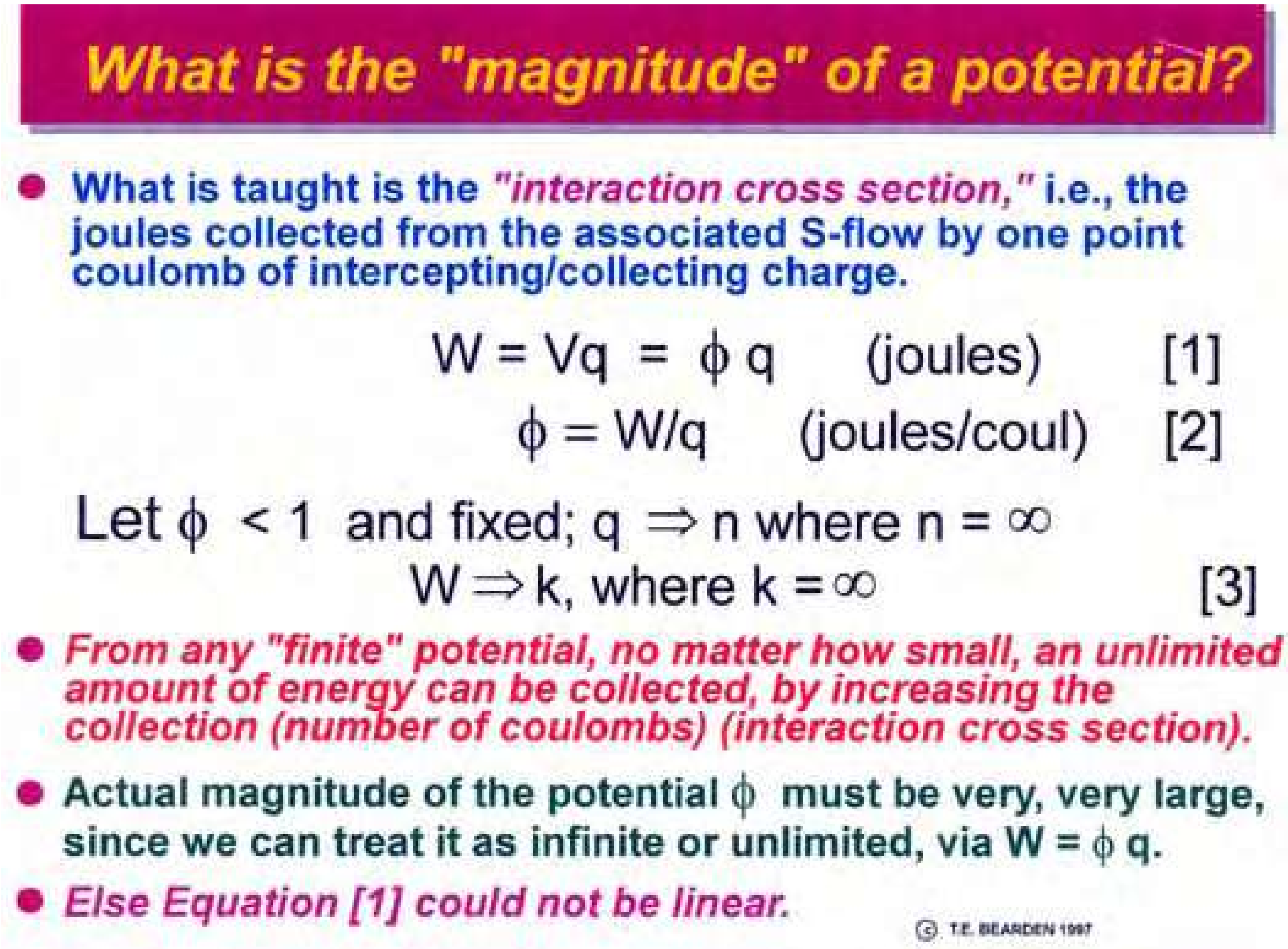

What is the "magnitude" of a potential?

- What is taught is the *"interaction cross section,"* i.e., the joules collected from the associated S-flow by one point coulomb of intercepting/collecting charge.

$$W = Vq = \phi q \qquad \text{(joules)} \qquad [1]$$

$$\phi = W/q \qquad \text{(joules/coul)} \qquad [2]$$

Let $\phi < 1$ and fixed; $q \Rightarrow n$ where $n = \infty$

$$W \Rightarrow k, \text{ where } k = \infty \qquad [3]$$

- *From any "finite" potential, no matter how small, an unlimited amount of energy can be collected, by increasing the collection (number of coulombs) (interaction cross section).*
- Actual magnitude of the potential ϕ must be very, very large, since we can treat it as infinite or unlimited, via $W = \phi q$.
- *Else Equation [1] could not be linear.*

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

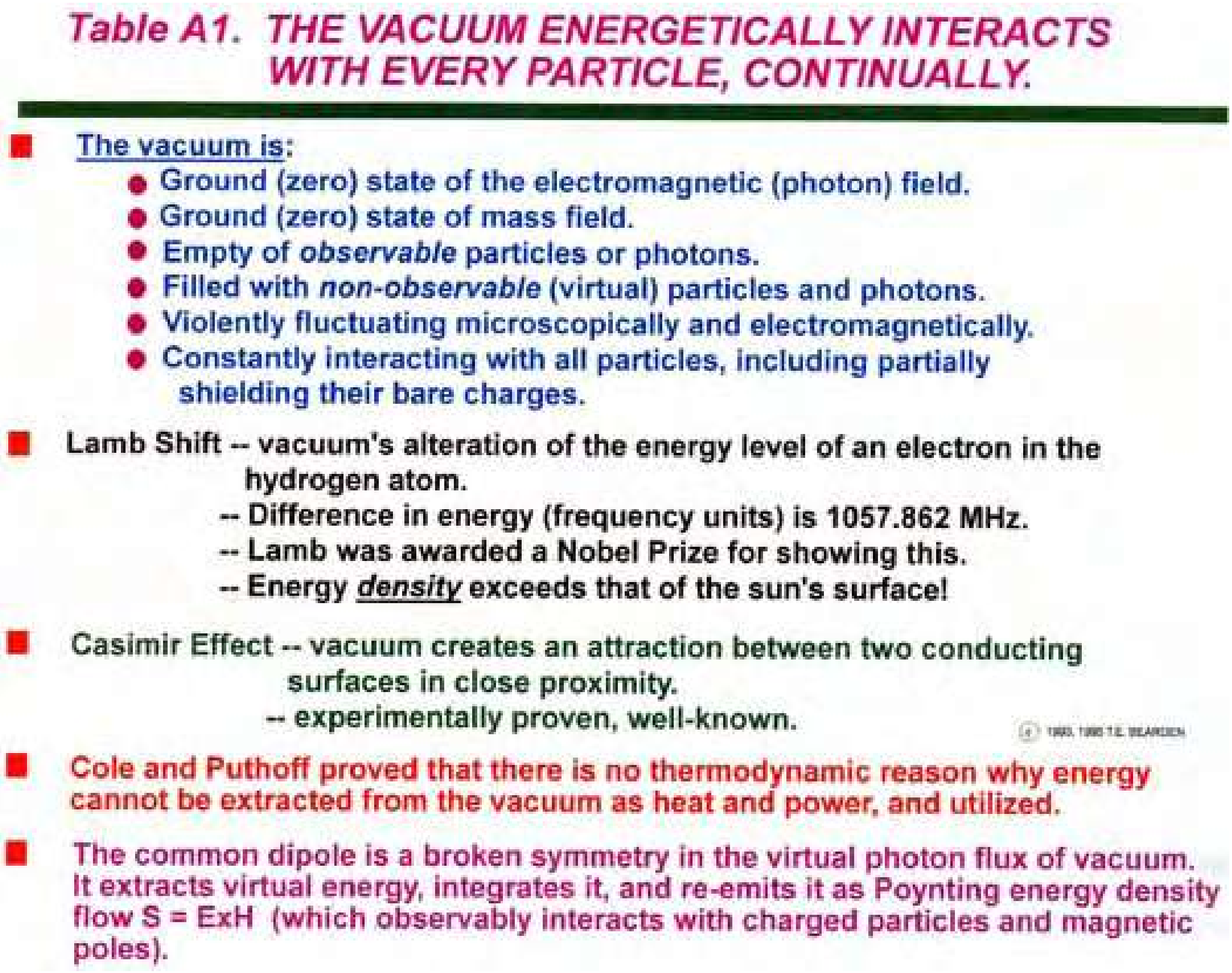

Table A1. THE VACUUM ENERGETICALLY INTERACTS WITH EVERY PARTICLE, CONTINUALLY.

- The vacuum is:
  - Ground (zero) state of the electromagnetic (photon) field.
  - Ground (zero) state of mass field.
  - Empty of *observable* particles or photons.
  - Filled with *non-observable* (virtual) particles and photons.
  - Violently fluctuating microscopically and electromagnetically.
  - Constantly interacting with all particles, including partially shielding their bare charges.
- Lamb Shift -- vacuum's alteration of the energy level of an electron in the hydrogen atom.
  - -- Difference in energy (frequency units) is 1057.862 MHz.
  - -- Lamb was awarded a Nobel Prize for showing this.
  - -- Energy *density* exceeds that of the sun's surface!
- Casimir Effect -- vacuum creates an attraction between two conducting surfaces in close proximity.
  - -- experimentally proven, well-known.
- Cole and Puthoff proved that there is no thermodynamic reason why energy cannot be extracted from the vacuum as heat and power, and utilized.
- The common dipole is a broken symmetry in the virtual photon flux of vacuum. It extracts virtual energy, integrates it, and re-emits it as Poynting energy density flow S = ExH (which observably interacts with charged particles and magnetic poles).

Return to Slide Index

Next Slide

Previous Slide

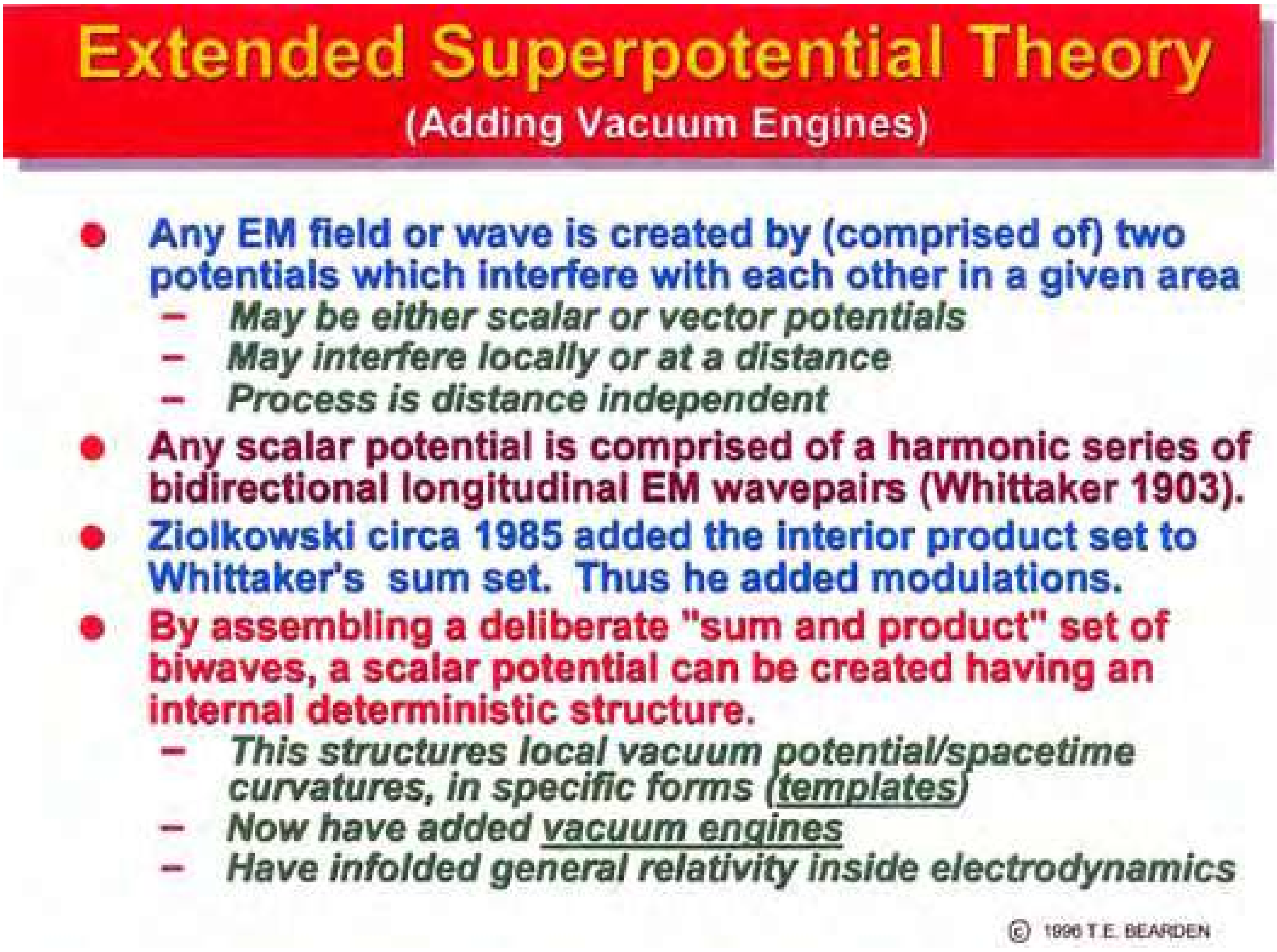

Extended Superpotential Theory

(Adding Vacuum Engines)

- Any EM field or wave is created by (comprised of) two potentials which interfere with each other in a given area
  - *May be either scalar or vector potentials*
  - *May interfere locally or at a distance*
  - *Process is distance independent*
- Any scalar potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs (Whittaker 1903).
- Ziolkowski circa 1985 added the interior product set to Whittaker's sum set. Thus he added modulations.
- By assembling a deliberate "sum and product" set of biwaves, a scalar potential can be created having an internal deterministic structure.
  - *This structures local vacuum potential/spacetime curvatures, in specific forms (templates)*
  - *Now have added vacuum engines*
  - *Have infolded general relativity inside electrodynamics*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

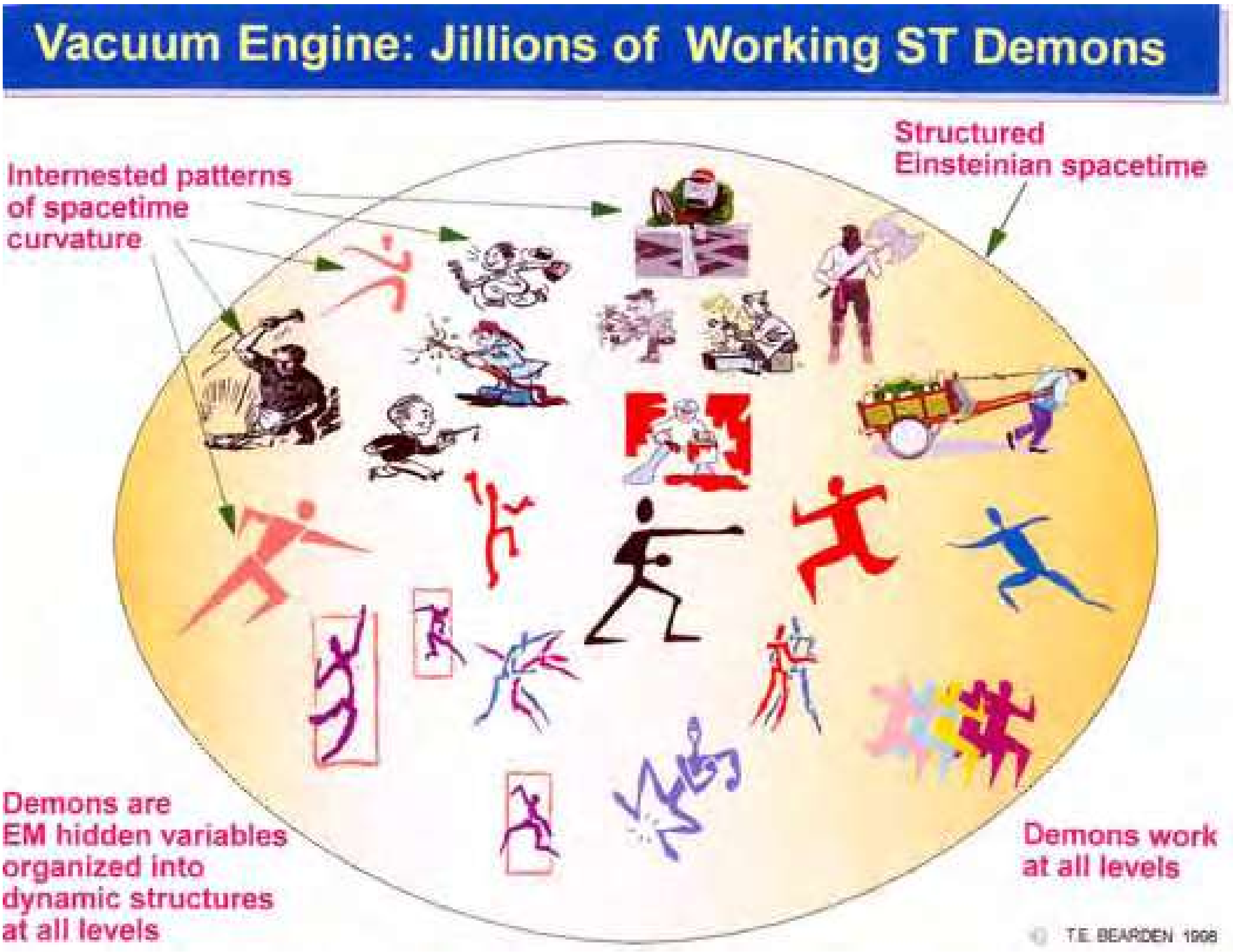

Vacuum Engine: Jillions of Working ST Demons

Internested patterns of spacetime curvature

Structured Einsteinian spacetime

Demons are EM hidden variables organized into dynamic structures at all levels

Demons work at all levels

© T.E. BEARDEN 1998

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

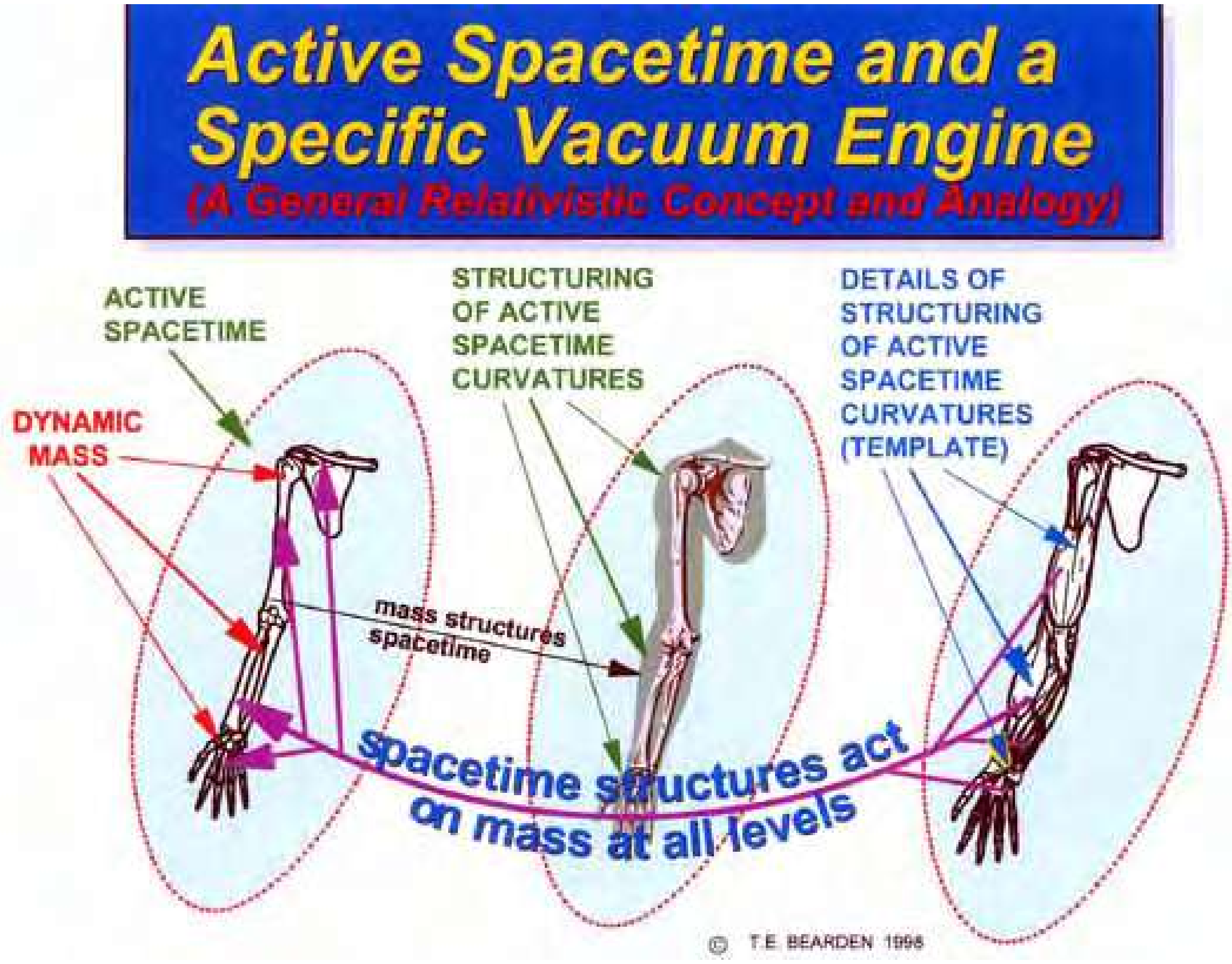

Active Spacetime and a Specific Vacuum Engine

(A General Relativistic Concept and Analogy)

ACTIVE SPACETIME

DYNAMIC MASS

STRUCTURING OF ACTIVE SPACETIME CURVATURES

DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)

mass structures spacetime

spacetime structures act on mass at all levels

© T.E. BEARDEN 1998

[Return to Slide Index]

[Next Slide]

[Previous Slide]

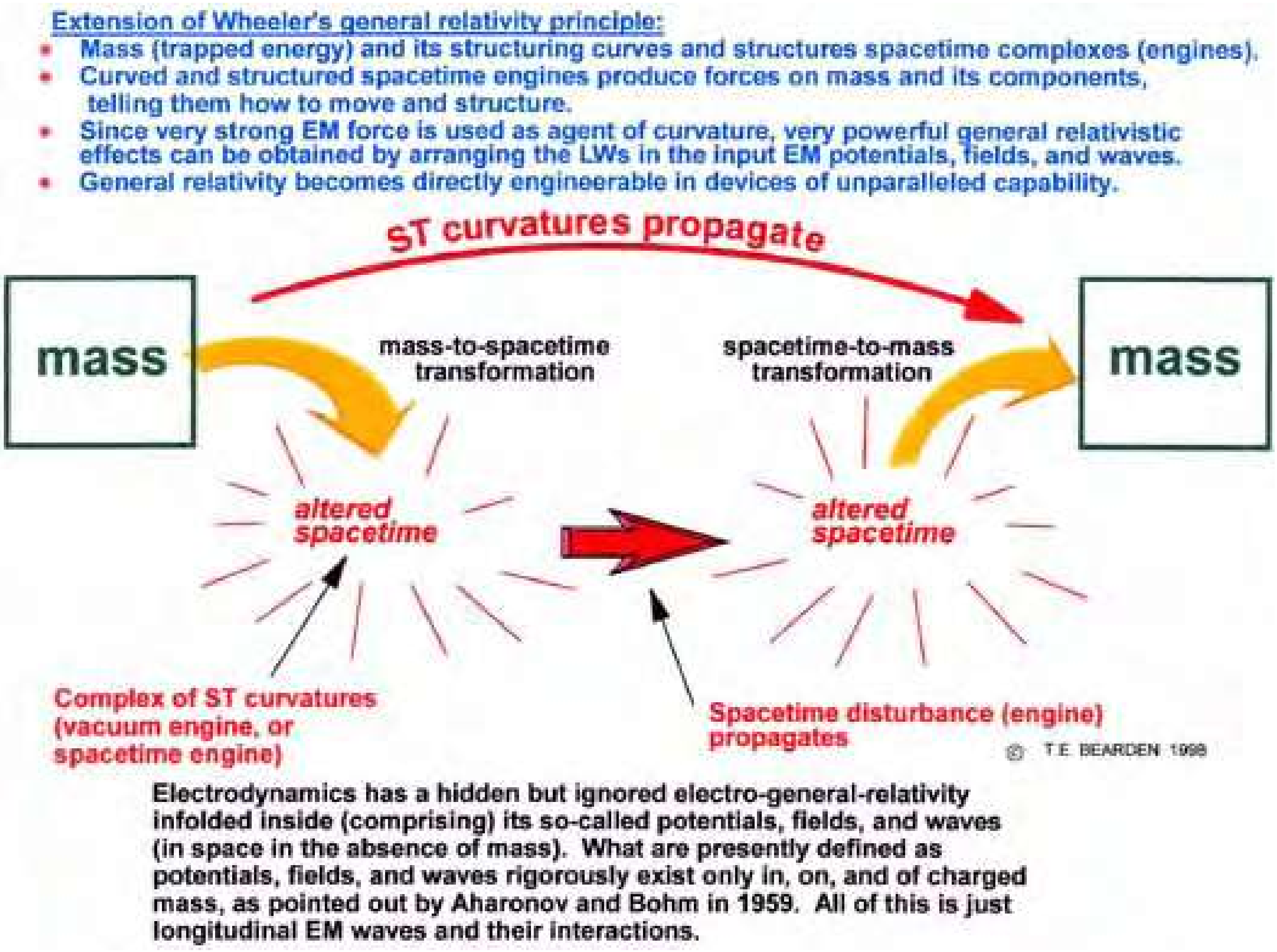

Extension of Wheeler's general relativity principle:

- Mass (trapped energy) and its structuring curves and structures spacetime complexes (engines).
- Curved and structured spacetime engines produce forces on mass and its components, telling them how to move and structure.
- Since very strong EM force is used as agent of curvature, very powerful general relativistic effects can be obtained by arranging the LWs in the input EM potentials, fields, and waves.
- General relativity becomes directly engineerable in devices of unparalleled capability.

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

© T.E. BEARDEN 1998

Electrodynamics has a hidden but ignored electro-general-relativity infolded inside (comprising) its so-called potentials, fields, and waves (in space in the absence of mass). What are presently defined as potentials, fields, and waves rigorously exist only in, on, and of charged mass, as pointed out by Aharonov and Bohm in 1959. All of this is just longitudinal EM waves and their interactions.

Return to Slide Index

Next Slide

Previous Slide

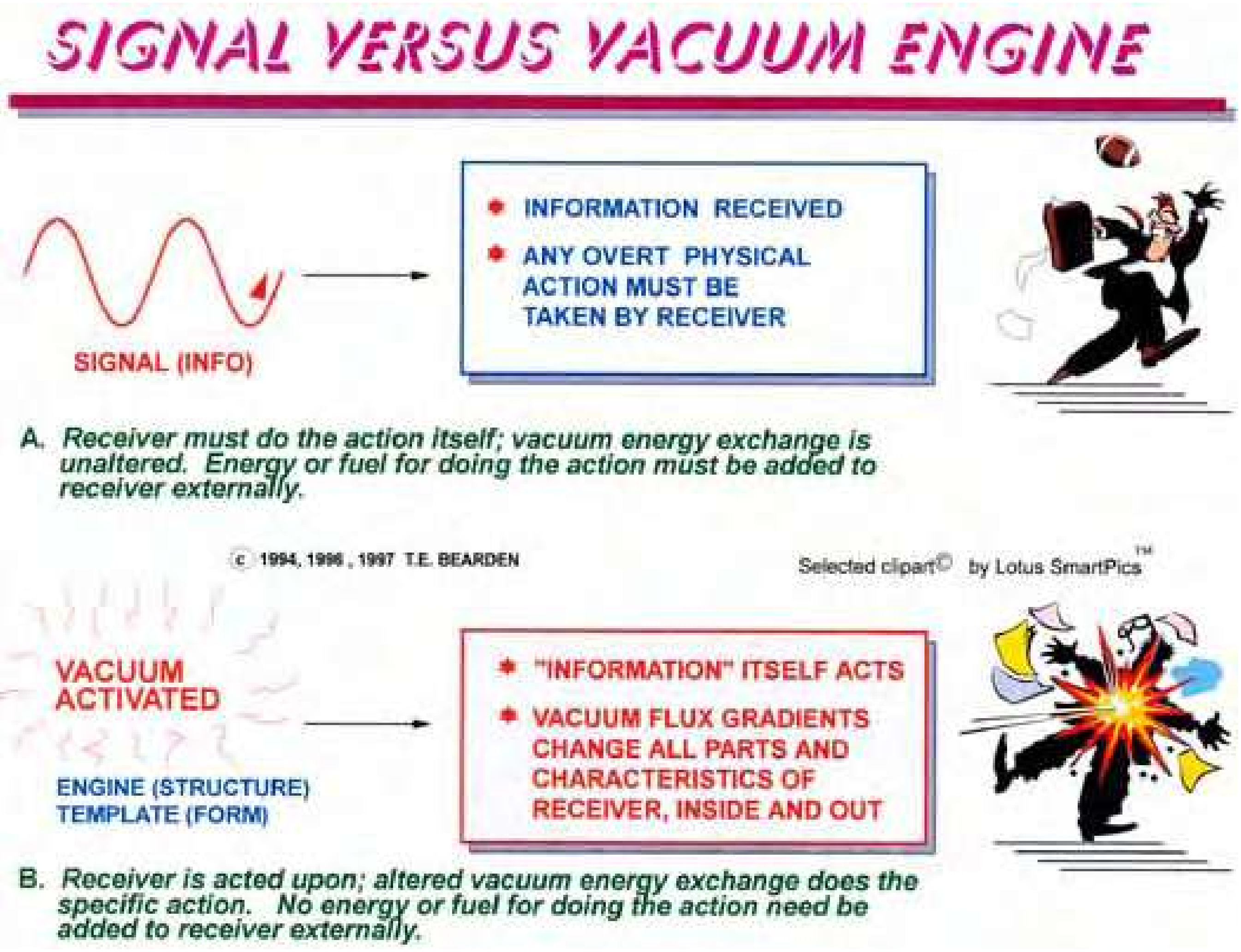

SIGNAL VERSUS VACUUM ENGINE

SIGNAL (INFO)

- INFORMATION RECEIVED
- ANY OVERT PHYSICAL ACTION MUST BE TAKEN BY RECEIVER

A. *Receiver must do the action itself; vacuum energy exchange is unaltered. Energy or fuel for doing the action must be added to receiver externally.*

© 1994, 1996, 1997 T.E. BEARDEN

Selected clipart© by Lotus SmartPics™

VACUUM ACTIVATED

ENGINE (STRUCTURE) TEMPLATE (FORM)

- "INFORMATION" ITSELF ACTS
- VACUUM FLUX GRADIENTS CHANGE ALL PARTS AND CHARACTERISTICS OF RECEIVER, INSIDE AND OUT

B. *Receiver is acted upon; altered vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to receiver externally.*

Return to Slide Index

Next Slide

Previous Slide

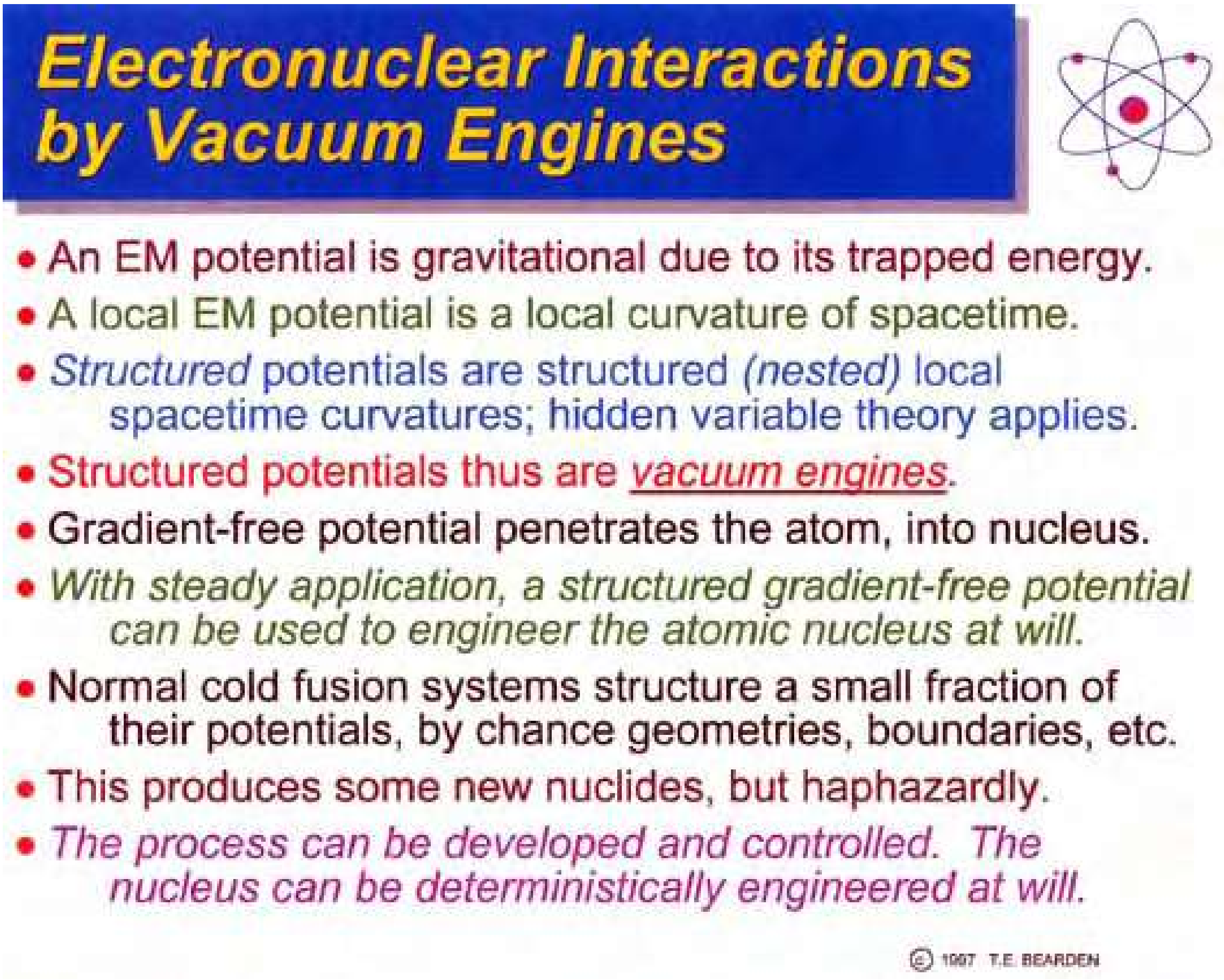

Electronuclear Interactions by Vacuum Engines

- An EM potential is gravitational due to its trapped energy.
- A local EM potential is a local curvature of spacetime.
- *Structured* potentials are structured *(nested)* local spacetime curvatures; hidden variable theory applies.
- Structured potentials thus are *vacuum engines*.
- Gradient-free potential penetrates the atom, into nucleus.
- *With steady application, a structured gradient-free potential can be used to engineer the atomic nucleus at will.*
- Normal cold fusion systems structure a small fraction of their potentials, by chance geometries, boundaries, etc.
- This produces some new nuclides, but haphazardly.
- *The process can be developed and controlled. The nucleus can be deterministically engineered at will.*

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

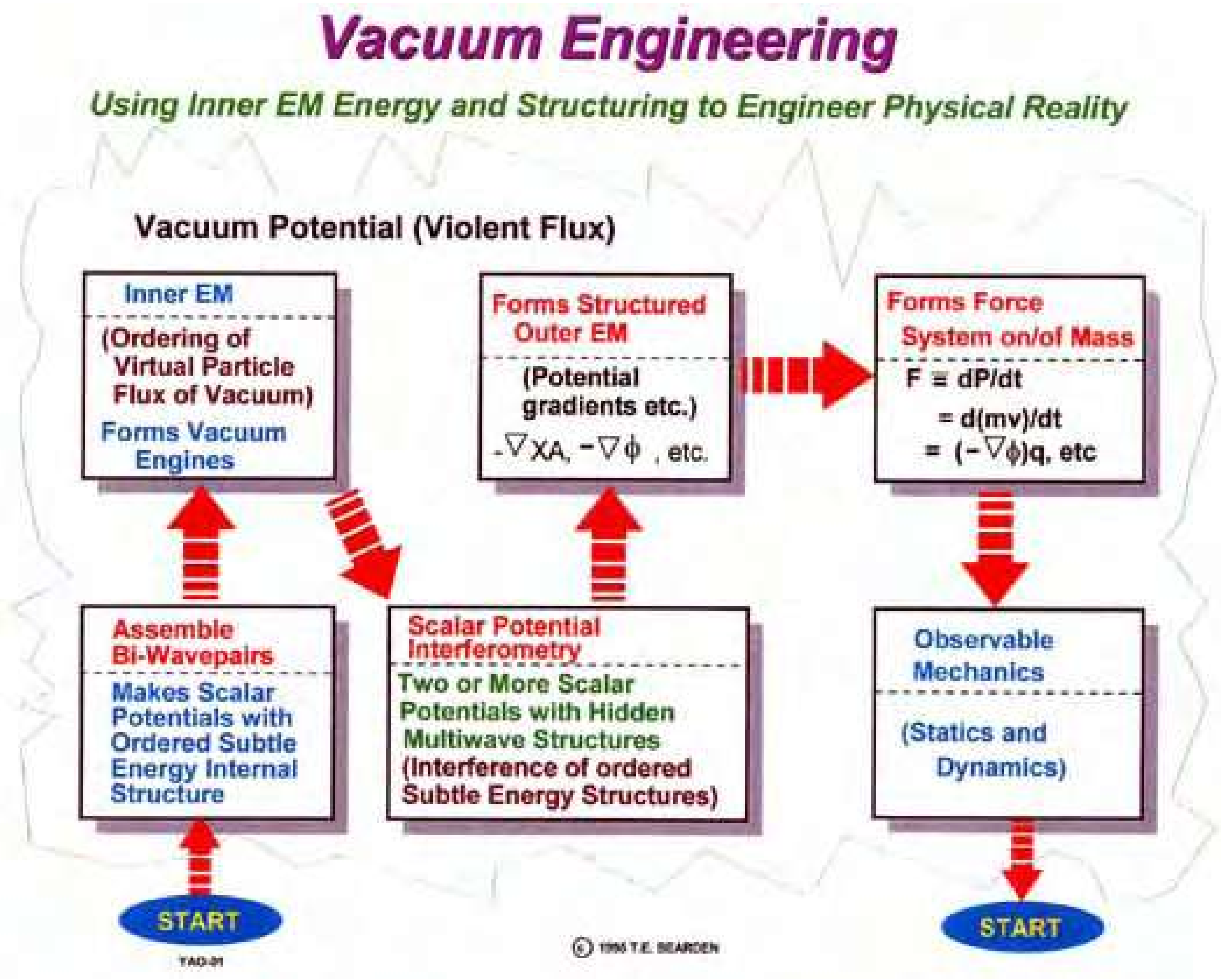

Vacuum Engineering

Using Inner EM Energy and Structuring to Engineer Physical Reality

Vacuum Potential (Violent Flux)

Inner EM

(Ordering of Virtual Particle Flux of Vacuum)

Forms Vacuum Engines

Forms Structured Outer EM

(Potential gradients etc.)

$-\nabla \times A$, $-\nabla \phi$, etc.

Forms Force System on/of Mass

$F \equiv dP/dt$

$= d(mv)/dt$

$= (-\nabla \phi)q$, etc

Assemble Bi-Wavepairs

Makes Scalar Potentials with Ordered Subtle Energy Internal Structure

Scalar Potential Interferometry

Two or More Scalar Potentials with Hidden Multiwave Structures (Interference of ordered Subtle Energy Structures)

Observable Mechanics

(Statics and Dynamics)

START

YAO-01

© 1996 T.E. BEARDEN

START

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

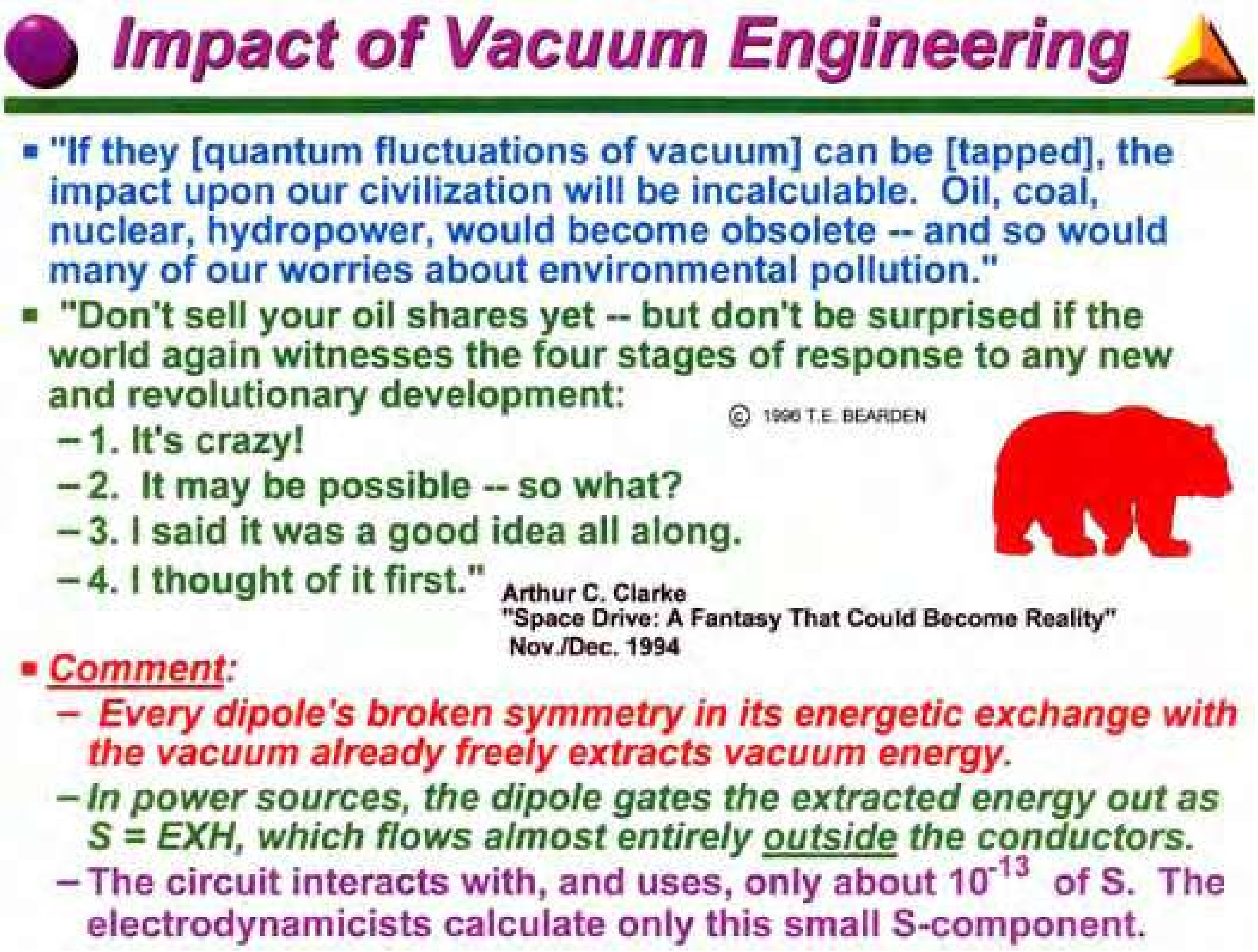

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete -- and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet -- but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

© 1996 T.E. BEARDEN

- *Comment:*
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

Return to Slide Index

Next Slide

Previous Slide

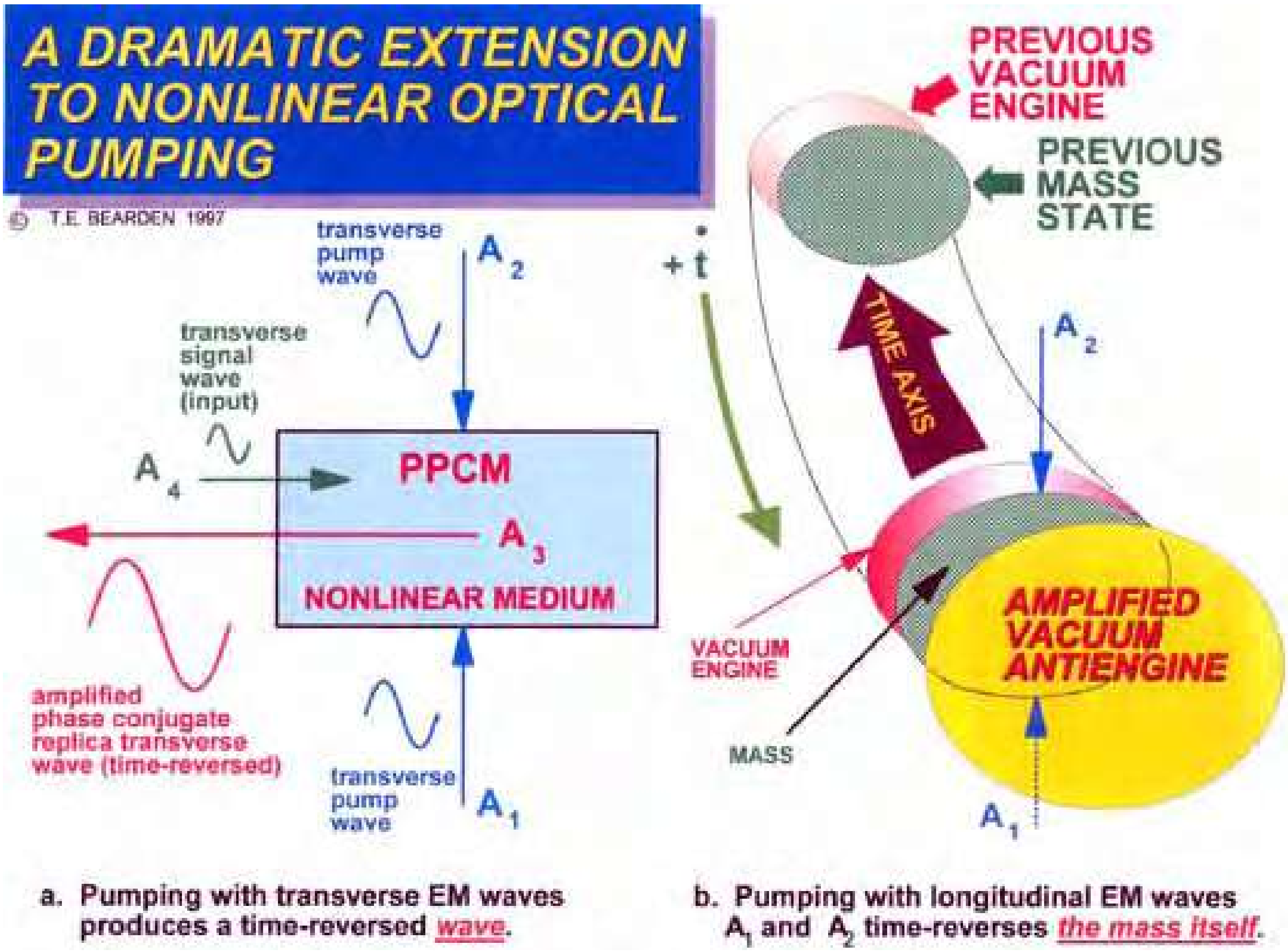

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

transverse pump wave

A_2

transverse signal wave (input)

A_4

PPCM

A_3

NONLINEAR MEDIUM

amplified phase conjugate replica transverse wave (time-reversed)

transverse pump wave

A_1

+ t

PREVIOUS VACUUM ENGINE

PREVIOUS MASS STATE

TIME AXIS

A_2

AMPLIFIED VACUUM ANTIENGINE

VACUUM ENGINE

MASS

A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Return to Slide Index

Next Slide

Previous Slide

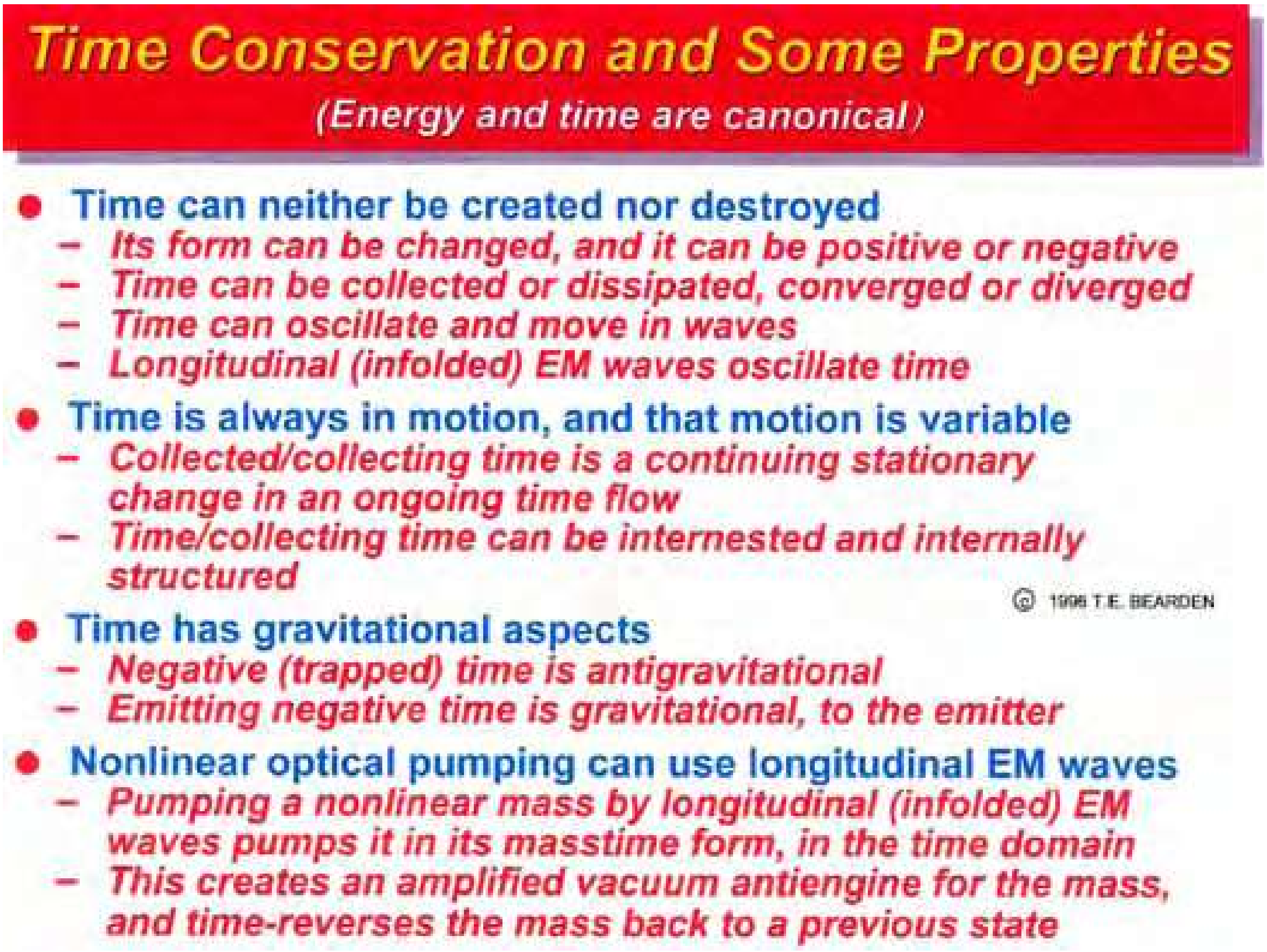

Time Conservation and Some Properties

(Energy and time are canonical)

- Time can neither be created nor destroyed
  - *Its form can be changed, and it can be positive or negative*
  - *Time can be collected or dissipated, converged or diverged*
  - *Time can oscillate and move in waves*
  - *Longitudinal (infolded) EM waves oscillate time*
- Time is always in motion, and that motion is variable
  - *Collected/collecting time is a continuing stationary change in an ongoing time flow*
  - *Time/collecting time can be internested and internally structured*

© 1996 T.E. BEARDEN

- Time has gravitational aspects
  - *Negative (trapped) time is antigravitational*
  - *Emitting negative time is gravitational, to the emitter*
- Nonlinear optical pumping can use longitudinal EM waves
  - *Pumping a nonlinear mass by longitudinal (infolded) EM waves pumps it in its masstime form, in the time domain*
  - *This creates an amplified vacuum antiengine for the mass, and time-reverses the mass back to a previous state*

Return to Slide Index

Next Slide

Previous Slide

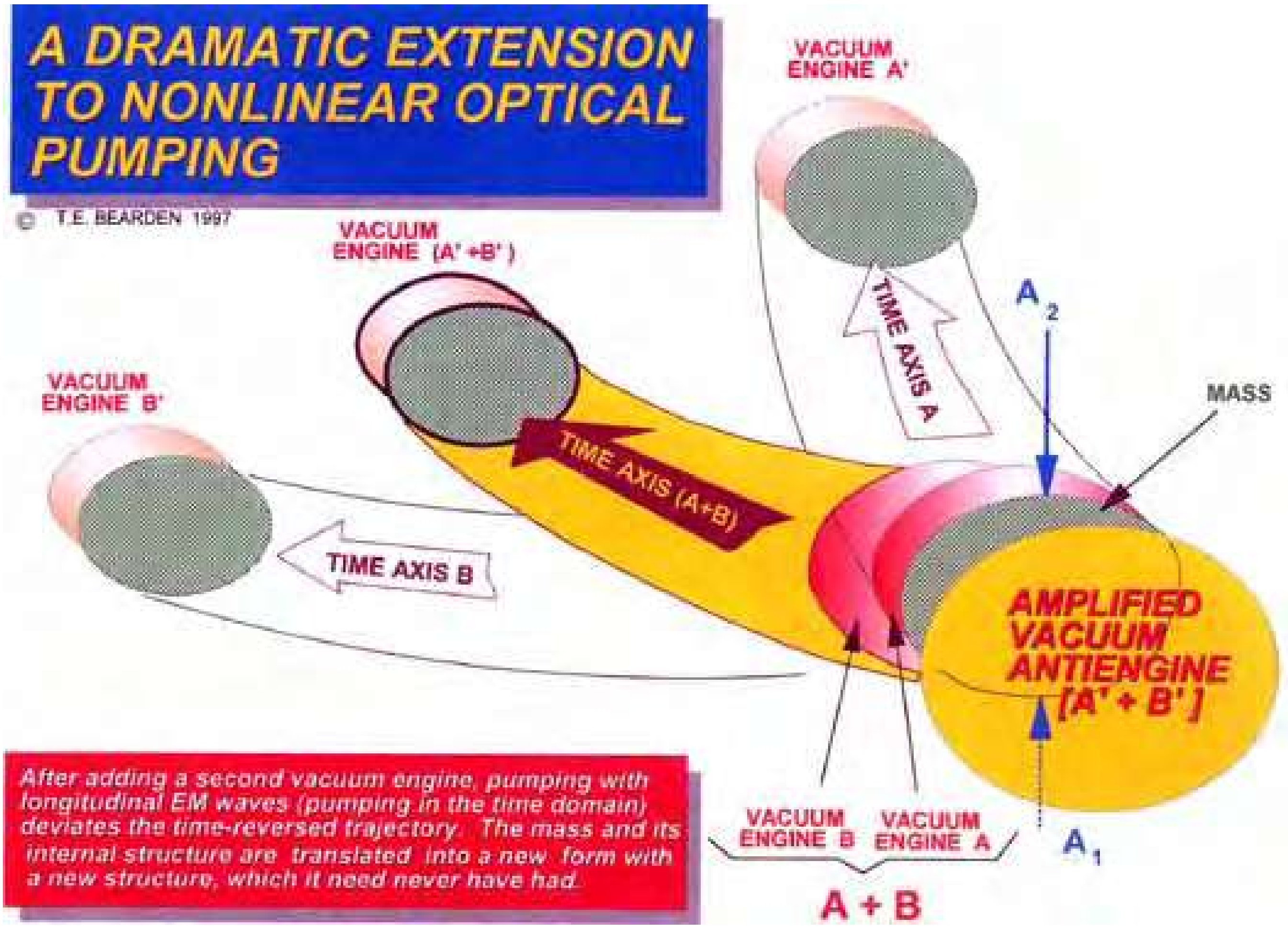

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
A₂
MASS
TIME AXIS (A+B)
TIME AXIS B
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A₁
A + B

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

Return to Slide Index

Next Slide

Previous Slide

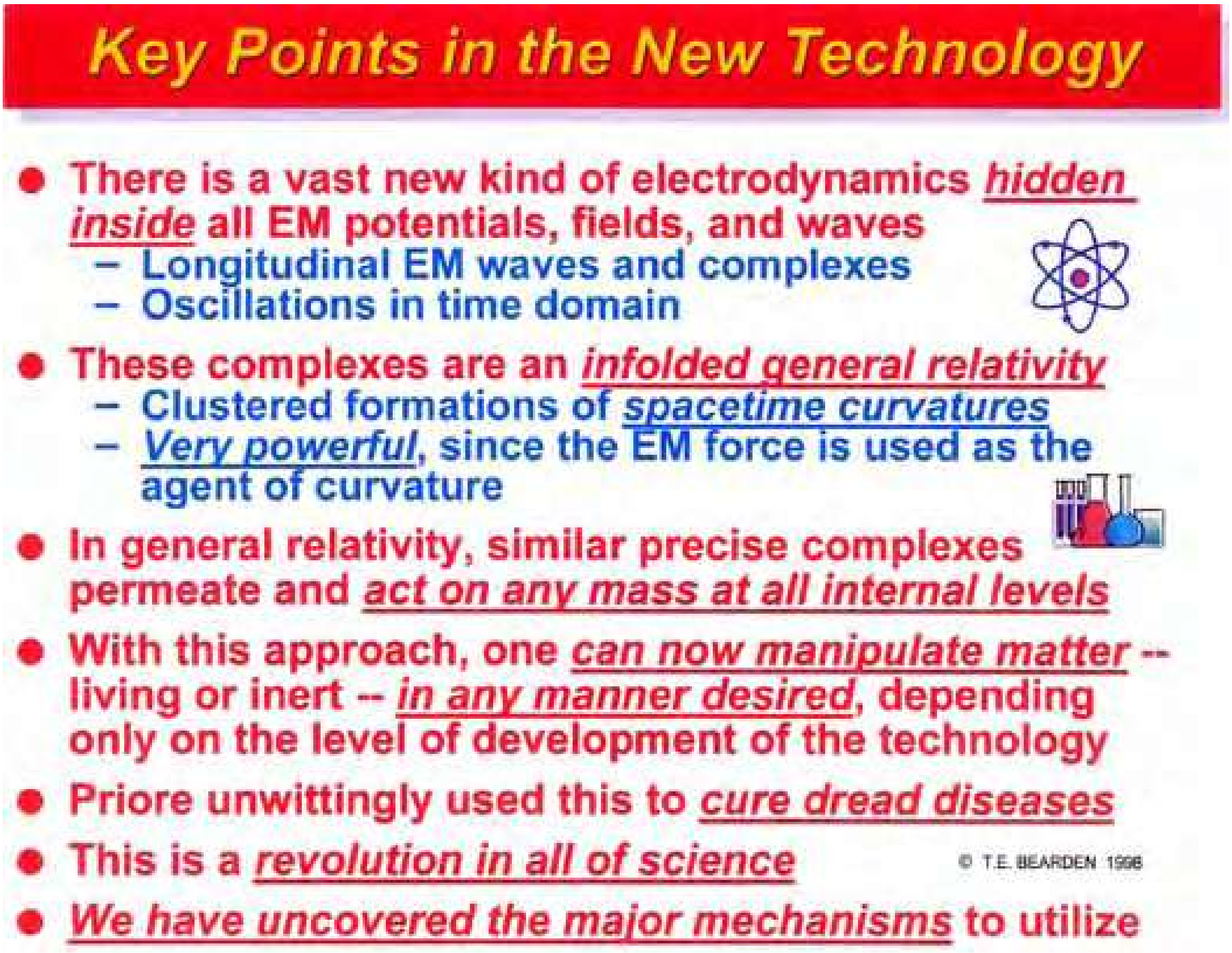

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

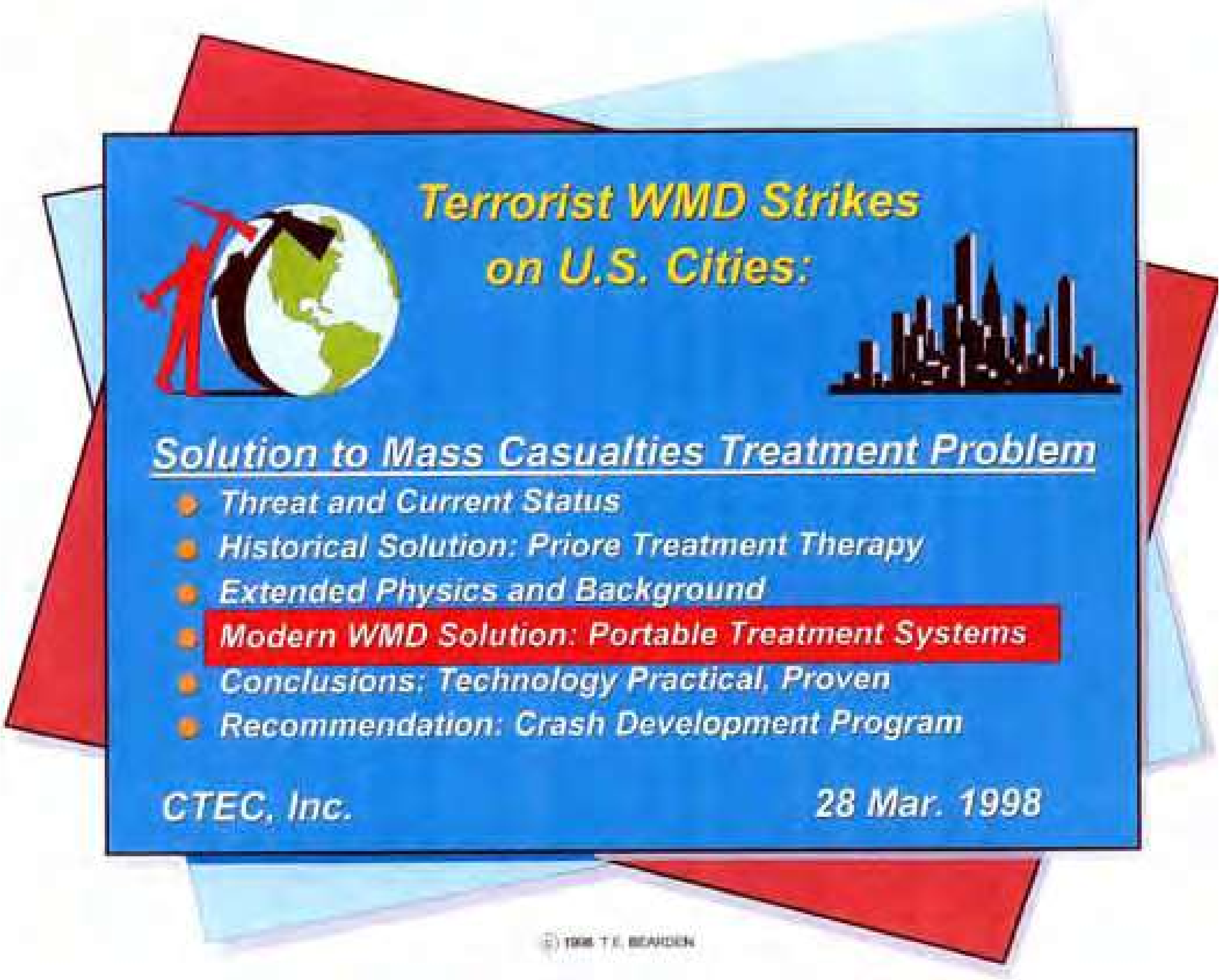

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

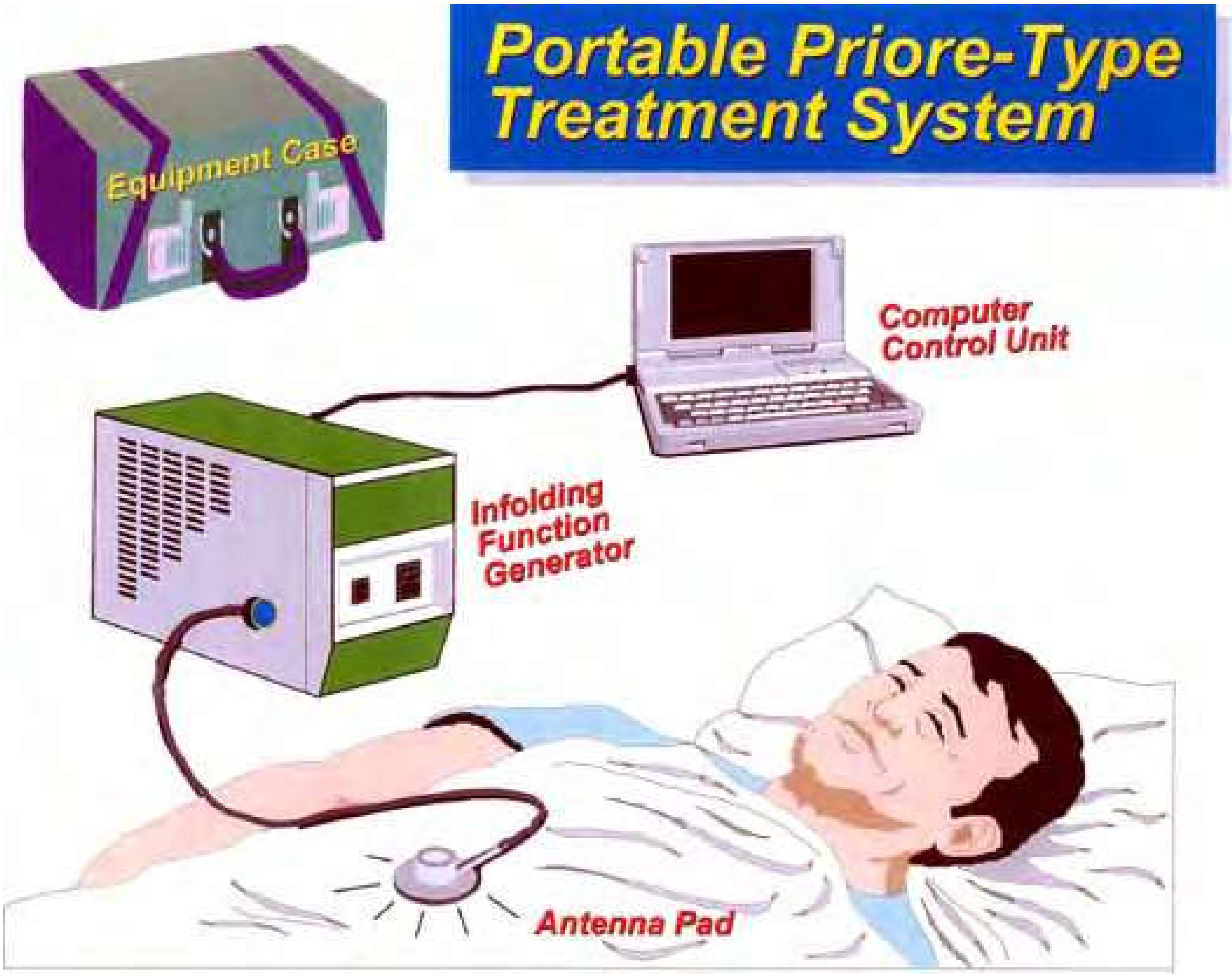

Portable Priore-Type Treatment System

Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Return to Slide Index

Next Slide

Previous Slide

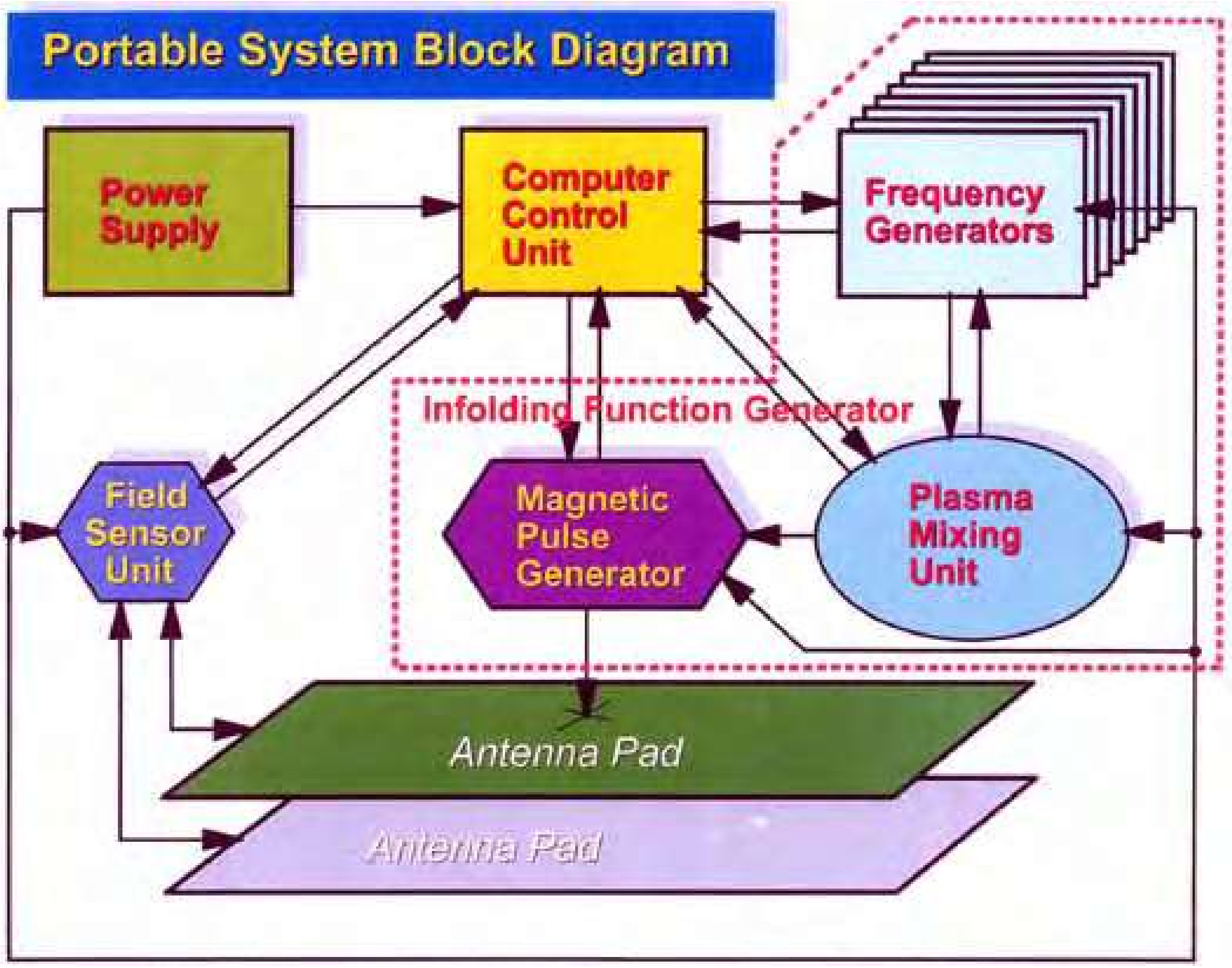

Portable System Block Diagram

Power Supply
Computer Control Unit
Frequency Generators
Infolding Function Generator
Field Sensor Unit
Magnetic Pulse Generator
Plasma Mixing Unit
Antenna Pad
Antenna Pad

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

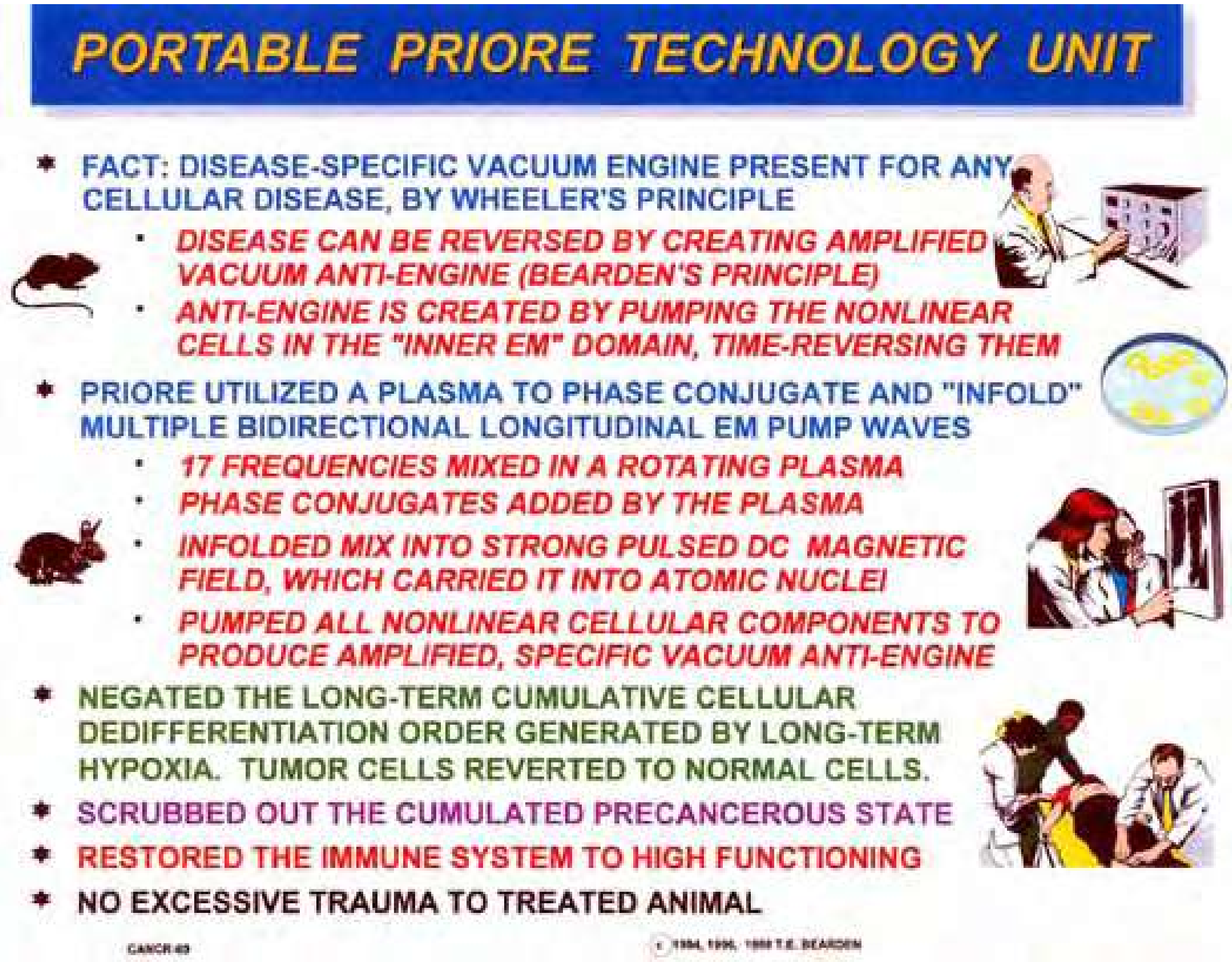

PORTABLE PRIORE TECHNOLOGY UNIT

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1999 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

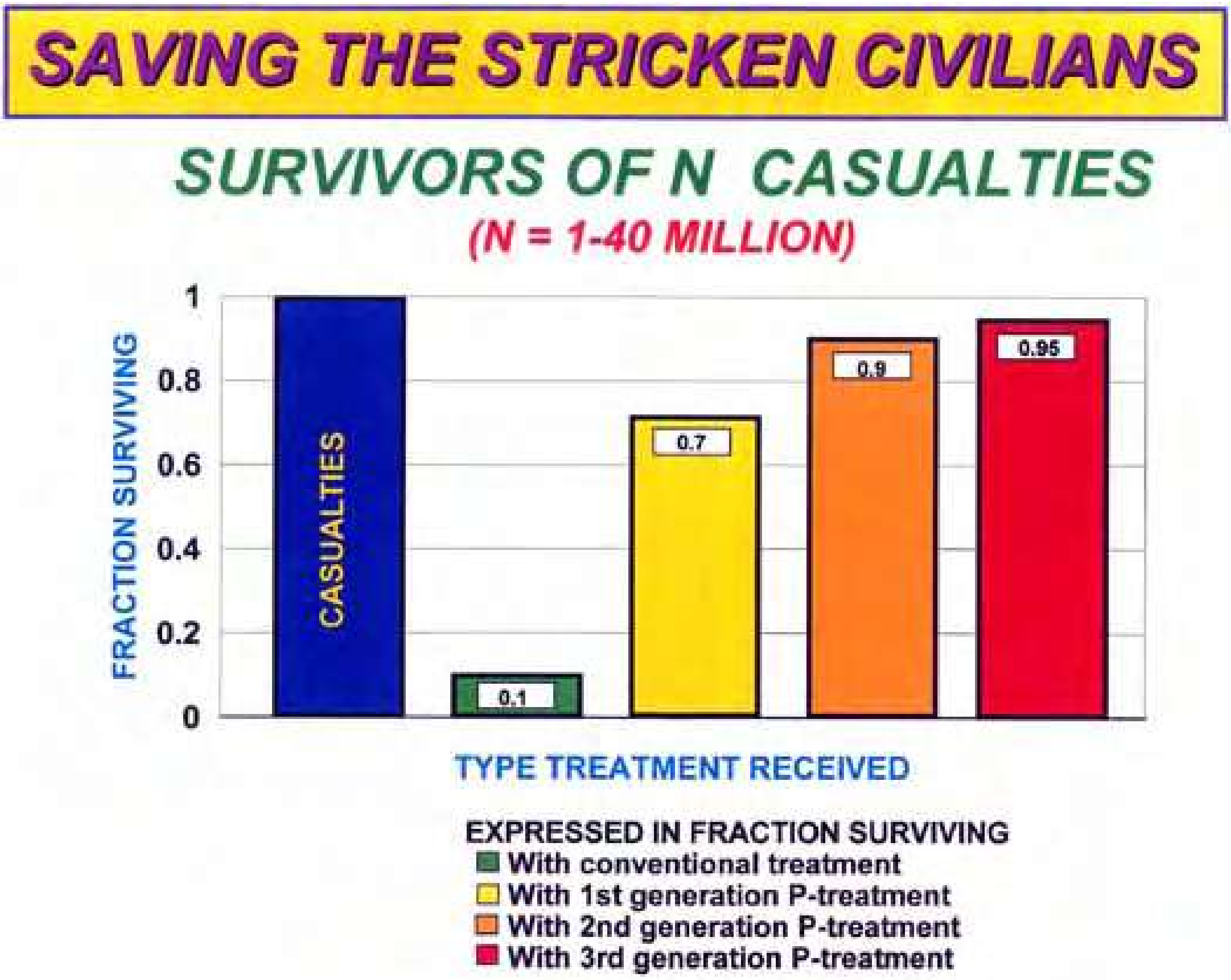

SAVING THE STRICKEN CIVILIANS

SURVIVORS OF N CASUALTIES

(N = 1-40 MILLION)

EXPRESSED IN FRACTION SURVIVING
- 🟩 With conventional treatment
- 🟨 With 1st generation P-treatment
- 🟧 With 2nd generation P-treatment
- 🟥 With 3rd generation P-treatment

Return to Slide Index

Next Slide

Previous Slide

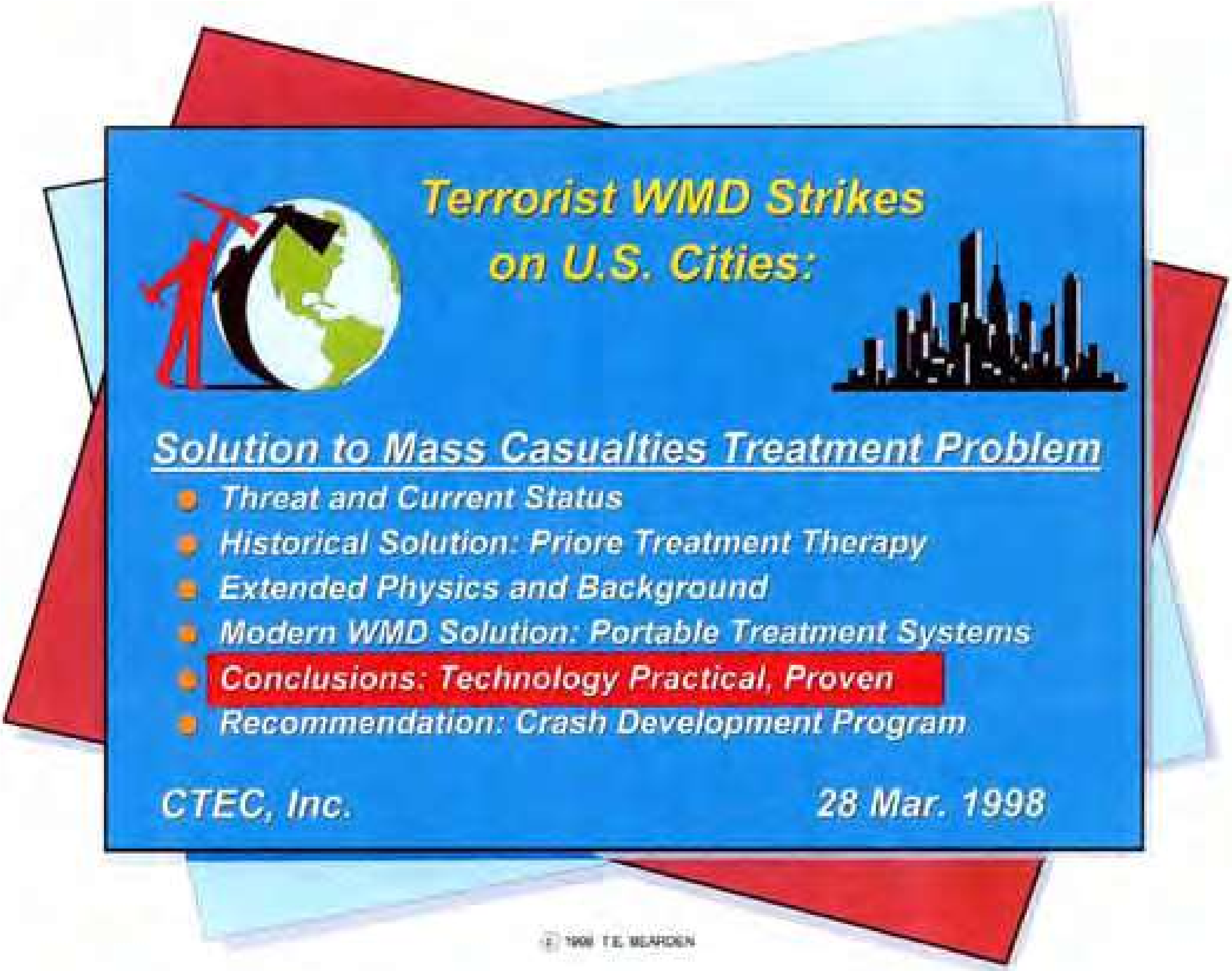

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc. 28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

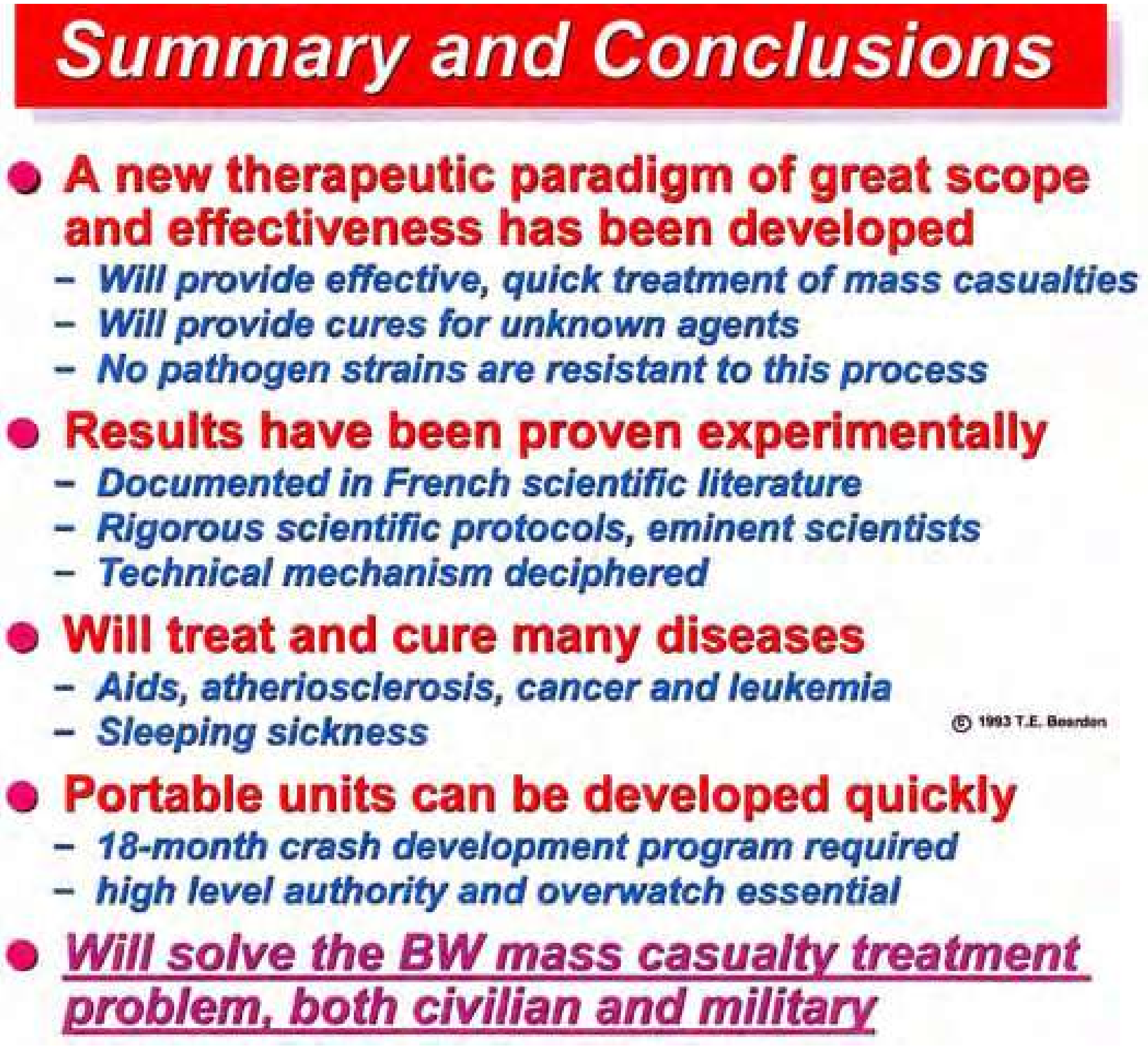

Summary and Conclusions

- A new therapeutic paradigm of great scope and effectiveness has been developed
  - *Will provide effective, quick treatment of mass casualties*
  - *Will provide cures for unknown agents*
  - *No pathogen strains are resistant to this process*
- Results have been proven experimentally
  - *Documented in French scientific literature*
  - *Rigorous scientific protocols, eminent scientists*
  - *Technical mechanism deciphered*
- Will treat and cure many diseases
  - *Aids, atheriosclerosis, cancer and leukemia*
  - *Sleeping sickness*

© 1993 T.E. Bearden

- Portable units can be developed quickly
  - *18-month crash development program required*
  - *high level authority and overwatch essential*
- *Will solve the BW mass casualty treatment problem, both civilian and military*

Return to Slide Index

Next Slide

Previous Slide

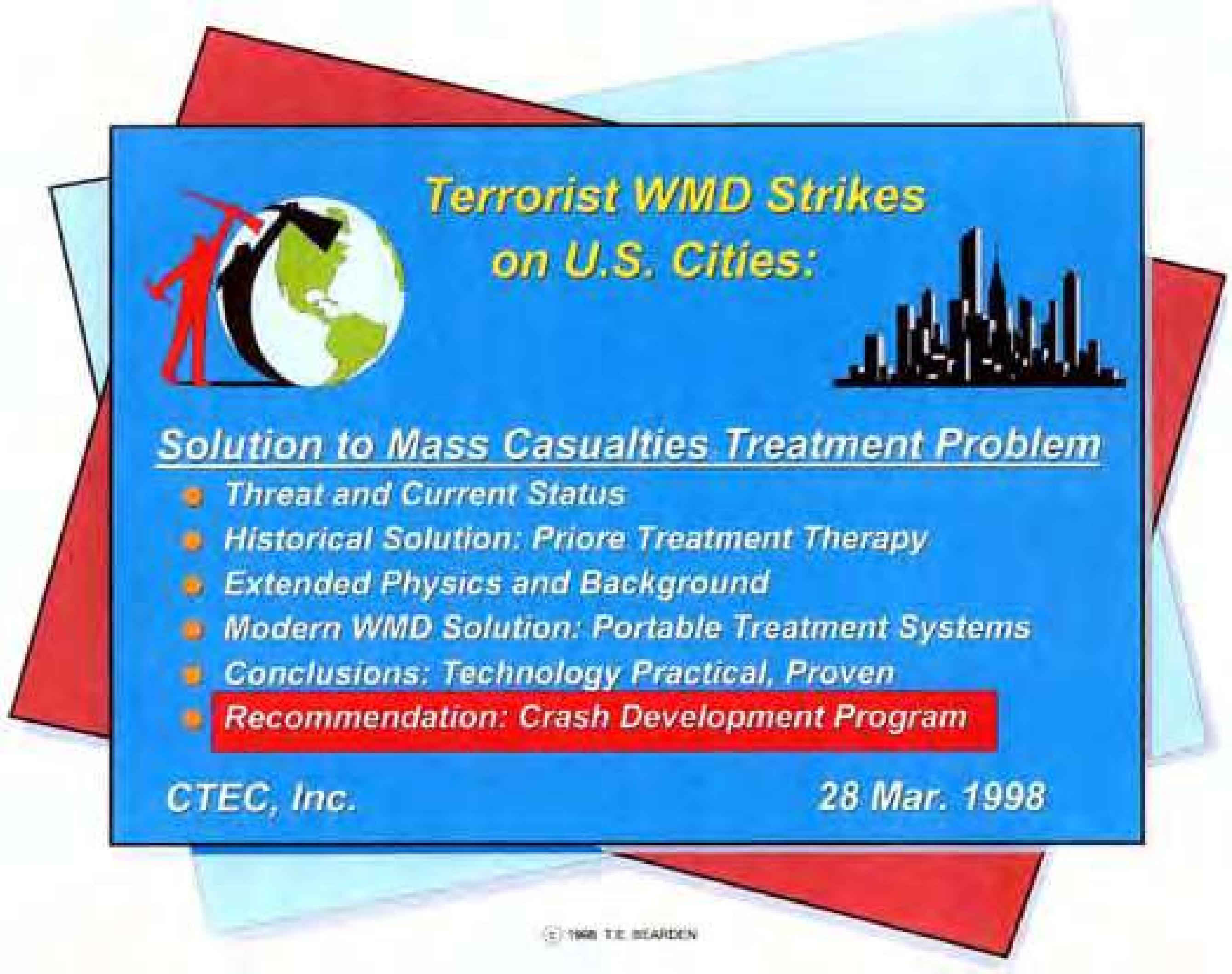

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc. 28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

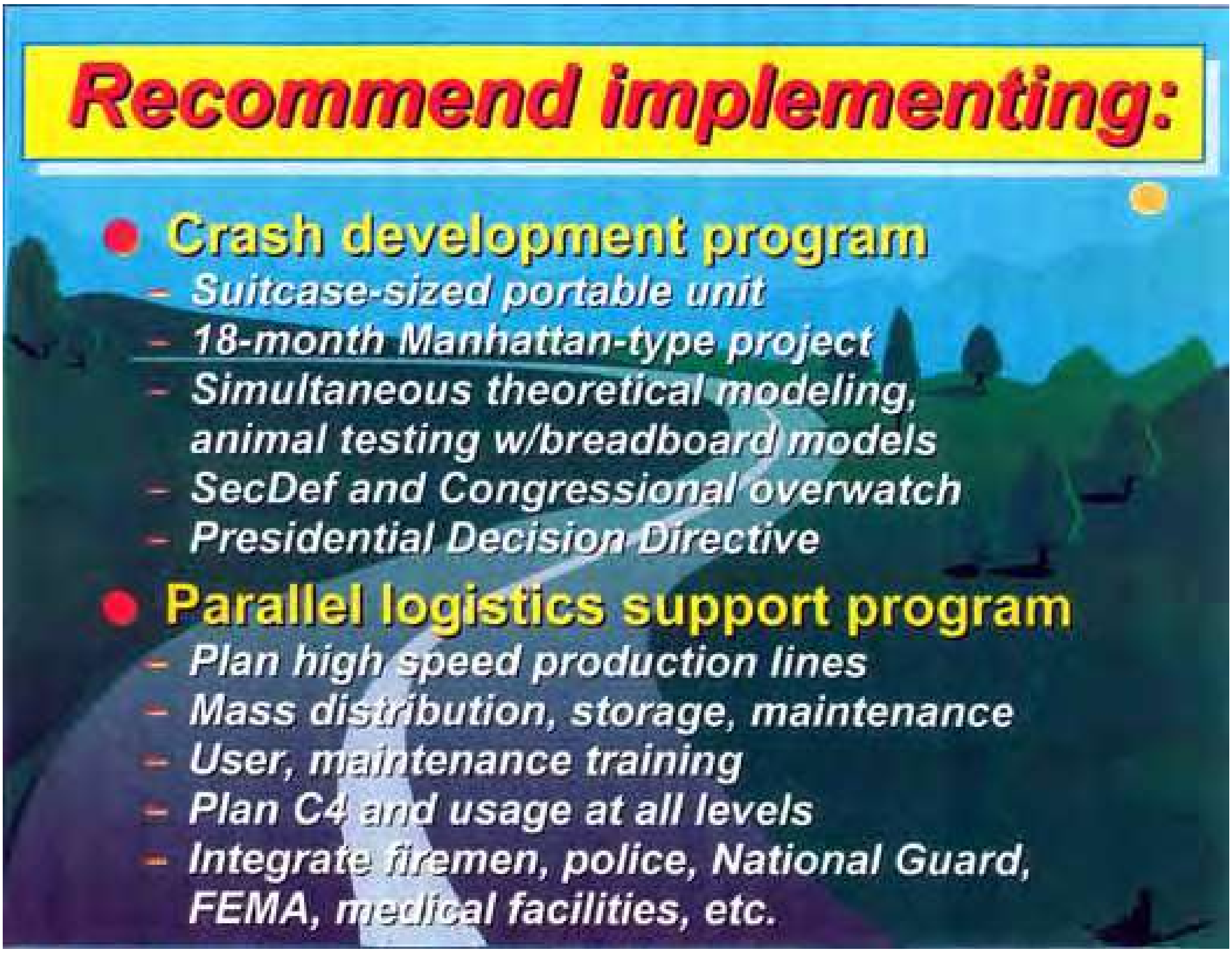

Recommend implementing:

- Crash development program
  - Suitcase-sized portable unit
  - 18-month Manhattan-type project
  - Simultaneous theoretical modeling, animal testing w/breadboard models
  - SecDef and Congressional overwatch
  - Presidential Decision Directive
- Parallel logistics support program
  - Plan high speed production lines
  - Mass distribution, storage, maintenance
  - User, maintenance training
  - Plan C4 and usage at all levels
  - Integrate firemen, police, National Guard, FEMA, medical facilities, etc.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

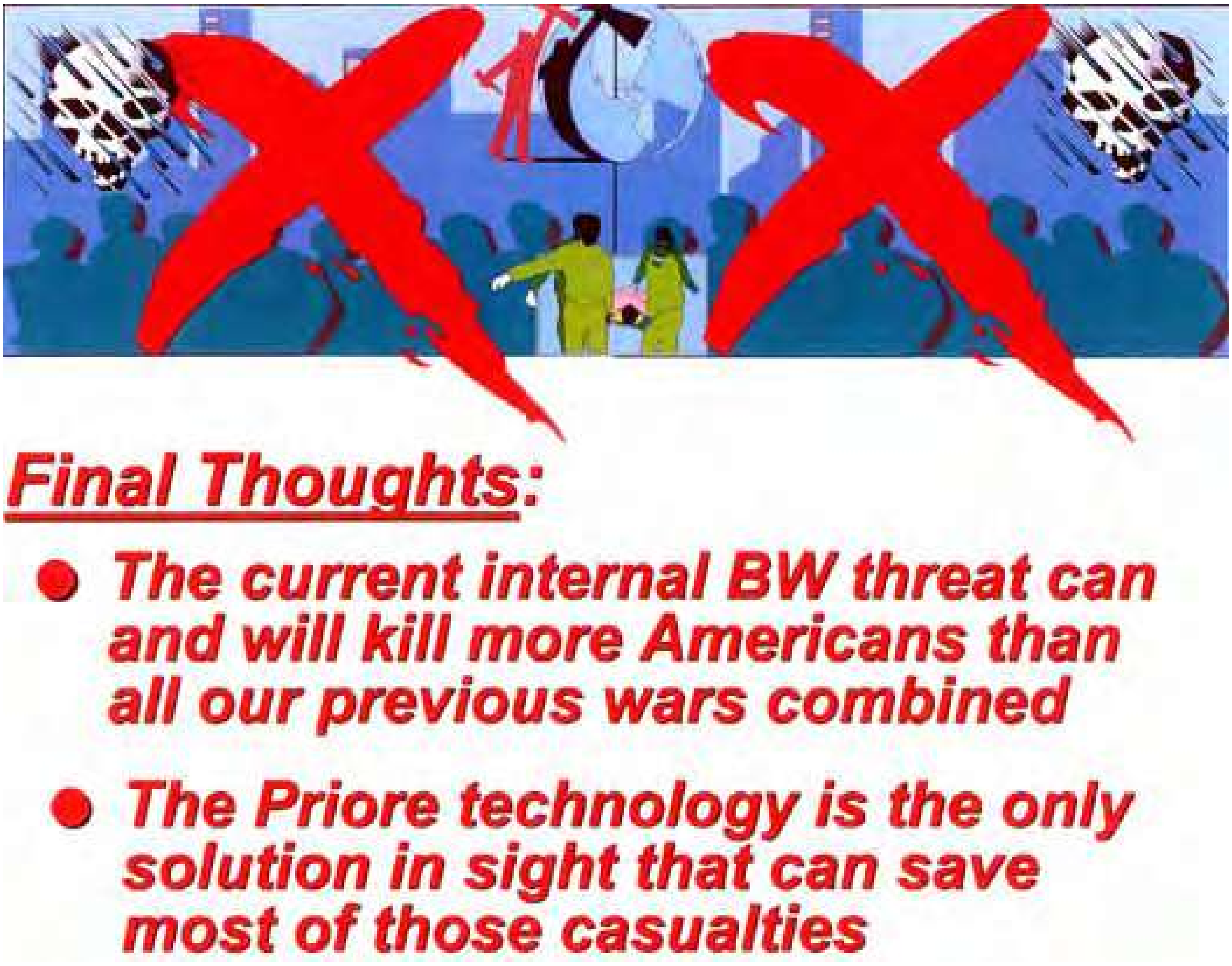

Final Thoughts:

- The current internal BW threat can and will kill more Americans than all our previous wars combined
- The Priore technology is the only solution in sight that can save most of those casualties

Return to Slide Index

Previous Slide

CTEC, Inc.
2311 Big Cove Road
Huntsville, AL 35801
(256) 533-3682 Ph/
(256) 539-5951 Fax

CTEC, Inc.

June 24, 1998

Dr. Jane F. Kinsel, Director
Office of Policy Analysis
Department of Health & Human Services
Bathesda, MD 20892

Dear Dr. Kinsel:

Your letter of June 8 is appreciated. However, the entire action package had already gone to the Assistant SecDef and also to Major General Neary in the USAF.

The proof of a revolution in medical treatment was in the package. The basis for it is already demonstrated experimentally in the scientific literature, both in the U.S. and in France. We are speaking of something already proven, but missed by the entire U.S. medical community. Strong words, but true. Let me be specific.

Any cell in the body can be time-reversed (physics term) or dedifferentiated (biology term) back to an earlier state, by pumping (i.e., "squeezing" in simple terms) that cell with longitudinal EM waves as pump waves. There are considerable papers in the French scientific literature showing the experimental results.

Apply that to cancer. We are stating bluntly that all cancerous cells in the body can be "depromoted" back to normal cells easily. Further, the damage previously done to them (such as by sustained hypoxia due to contaminants reducing the oxygen-carrying ability normally given to it by some 60-to-80 surrounding water molecules per red cell and the consequent H-bonding interactions) will also be reversed. The cells can all be returned to healthy, normal cells again, without "killing" or "burning" or "cutting" anything.

With some development of the technology, you can quickly and easily cure every cancer patient in every hospital in the United States.

Now apply it to AIDS. Since you can time-reverse every cell in the body, you can time-reverse the HIV-infected cells – genetics and all – right back to normal cells with normal genetics. Those cells not infected and healthy, will just get a "little younger" again.

So with that small development, you can quickly and easily cure every AIDS patient in every hospital in the United States.

I would have thought that NIH would certainly be interested in such a revolutionary therapy which has already been demonstrated in laboratory experiments. World-renowned French scientists – such as the eminent parasitologist Pautrizel and research PhDs assigned directly by Robert Courrier, head of the Biology Section of the French Academy of Sciences, worked with Prioré in performing those startling experiments that proved what we are saying.

The method was rigorously shown to cure infectious diseases (such as trypanosomiasis), atherosclerosis (which now afflicts millions of Americans), and terminal tumors. It was shown to reverse suppressed immune systems back to normal vigorous functioning.

Please permit me to disagree with you that the examination of such is not the mission of the NIH. To the contrary, it is precisely the NIH which should have some of its very best scientists looking into this.

The action package to DoD was to propose a crash program to develop this methodology in the form of a small, portable unit for treatment of mass casualties resulting from a terrorist BW attack on our cities (now expected in the next few weeks, as you must be aware). The leading Arab terrorist of the world has already promised just this.

We are speaking of several million Americans who are going to die. They are going to die in spite of everything that NIH and its labs can do, even though I know you will make a heroic effort. With NIH present methods, you cannot do very much for those stricken Americans at all. Everything we have in the "normal" kit bag is almost useless against a knowledgeable spray attack of modified smallpox (as you know, the Russian secret BW labs have made and sold tons of that already, to guess who!). Or a spray of tularemia (you are well aware of the extreme lethality of that), or anthrax (easily obtained). It is also common knowledge that these terrorist teams with their BW agents are already on site in this country, waiting for the word to attack. This has now been officially raised to our primary Strategic Threat.

In the face of such looming strikes on the U.S., I simply cannot believe that the NIH does not feel it within their mission to rigorously check out a proposed method for saving perhaps 70% of those coming deaths (first generation equipment), and up to 90% with second generation equipment.

Perhaps you just did not read the package carefully. I urge you to do so.

Else NIH is going to look very, very bad when (1) the strikes do occur, (2) they can't handle them (and they cannot, as you well know), (3) they had within their hands in advance, a proposed methodology that could have been investigated intensely, to save millions of those stricken Americans.

As a staunch supporter of NIH and especially CDC, I have to believe that there is a different attitude there. If NIH is not interested in a previously experimentally demonstrated method of saving millions of Americans, then the only recourse left is the political channel.

I assure you that my critique of electrodynamics is well-founded, as supported by various leading physicists such as Nobelist Feynman (who stated bluntly that the field concept is wrong), Wheeler, Bunge, and so on. It is inexplicable why all our universities are still teaching a 130 year old abridged (Heaviside's) subset of Maxwell's theory, and one that still contains a material ether even though that was falsified 110 years ago!

You already have seen the misuse of this technology by the Russians, in the decades-long radiation of the U.S. Embassy in Moscow. Is it not NIH's mission to understand how they generated all those diseases and health changes? Have you ever looked into Kaznacheyev's experiments showing that ANY cellular disease or disorder can be induced in cells at a distance by purely EM means? Have you looked into what got eliminated from electrodynamics by ignoring the hidden longitudinal EM inside every potential, field, and wave? Do you realize what ignoring that has done to American chemistry, biochemistry, and medical science?"

I would hope that the NIH would at least be interested in its own mission area!

I challenge you bluntly to admit that no one presently understands the primary operation of the human cellular *regenerative* system. That's because its technical principle is not presently in our recognized physics, biochemistry, and medical science – even though it has been in the hard physics literature since 1903-4. Becker's epochal work came closest; since he only had access to the conventional EM model, it failed him. But he at least showed you rigorously that cells can be changed (both differentiated and redifferentiated) all over the map by laughably weak EM – picoamperes of current, which means persistent DC potentials. In the microwave radiation of the U.S. Embassy, all the health changes occurred in regions absent of fields – i.e., absent of bleed-off of the potentials. Hence in regions of stable DC potentials! With 100% correlation of the health changes to persistent, gradient free potentials, and 100% anticorrelation to changing potentials (i.e., fields), our own scientists concluded totally erroneously that it could not have been the EM radiation causing the changes, when their own results had just proven that it was.

Now I ask you: Can NIH explain how it was that the DC potentials caused those health changes? I think not. If it could, it would understand the action package I wrote, and its importance.

This is not a "political" letter. I'm just an ordinary American citizen who has done about 30 years very hard work on this particular area. The work is good enough now to stand on its own merits. The scientific community will eventually accept it, about 50 years from now. But we cannot wait 50 years! Those millions of Americans are going to die, because of the bureaucracy and inaction of our own scientific organizations. Whether they like it or not, the U.S. scientific community is now the forefront of the struggle for this nation to survive. And they are failing us dramatically.

So as a parent who wants his children to live, and his neighbors' children to live, and those millions of soon-to-be stricken Americans to live also, I urge you to at least have someone competent in physics (and knowledgeable that the foundations of physics itself are in woeful shape!) check the work. Has anybody there *read* any of the references to the Prioré work I cited? Does anyone there realize that *there are no EM forcefields in space?* Probably not.

Dr. Kinsel, there is a time for "business as usual" and there is a time to do some serious new work. Those stricken Americans are going to be lying there. Nothing NIH can do is going to stop that. That's a given, and it is not my own estimate. It's the official estimate. And you *know* you presently do not have the tools to do anything really effective about it.

Frankly, I interpret your letter to me as just the normal "spin control" for easy disposition. I would hope that NIH would seriously rethink the situation, get off the "spin control" and "business as usual" posturing, and do something dramatically different for a change.

What have you got to lose? A little time from perhaps two or three of your best scientists.

What have you got to gain? Saving untold millions of American lives. A new, *already partially demonstrated* cure for most dread diseases such as AIDS and cancer – diseases which you presently cannot do very much about in spite of decades of heroic struggle.

Sincerely,

cc: Senator Shelby, Congr. Cramer

T.E. Bearden
President and CEO

CTEC, Inc.
T.E. Bearden, President & CEO
2311 Big Cove Road
Huntsville, AL 35801-1351
(256)533-3682 ph/(256)536-0411 Fax

CTEC, Inc.

May 12, 1998

Dr. Harold Varmus, Director
National Institute of Health
Bathesda, MD 20892
(301) 496-1766

Dear Dr Varmus:

We have recently sent action documents to the Director of Nuclear and Counterproliferation, Office of the DCS, Air and Space Operations, HQ USAF and to the DoD (General Busbee) Assistant Secretary of Defense, as well as several Senate and House committees, which may be of interest to you. For one thing, we give the technical mechanism used to generate the Gulf War Syndrome. The same weaponry is now being used to rapidly induce new strains of certain pathogens worldwide. I urge you to closely read that section of the letter to MG Neary which explains the GWS, why the French did not get it, and why Southern Iraqis got an increase in cancer and leukemia but not GWS. Also explained is how the "microwave radiation" of the U.S. Embassy in Moscow induced diseases for decades, and how we failed to comprehend it because the standard EM model deceived us.

There is a way to treat and quickly cure most diseases, including the mass casualties resulting from a terrorist BW strike on our civilian population centers, our military bases, or our forces in the field. The method is little known, but was scientifically demonstrated in France in the 1960s and early 1970s before cancellation of the Prioré Project when the French Government changed. Eminent French scientists worked on the project, and the results are fully documented in leading French scientific journals. References are listed in a separate attachment.

Previously the Western scientific community has been unable to understand the technical mechanism responsible for such remarkable cures, because of serious foundations flaws in Western physics and electrodynamics. *Foundations* physicists have long been aware our science is flawed, and have pointed out many of these fundamental errors. However, the operational science establishment has not paid attention, even though physicists such as Feynman and Wheeler pointed out that the notion of "force fields in space" was totally wrong.

As a primary example, there is an "infolded" electrodynamics inside all potentials, fields, and waves that is far more primary than the coarse EM that is in our textbooks. It is actually an infolded general relativity, and this is what the Russians have weaponized. This infolded EM-GR has been ignored in the West (but not in Russia) since 1903.

As another example the potentials, fields, and waves of present electrodynamics are rigorously defined – and exist – only in the presence of mass. They do not exist in such form (force fields and oscillating force fields) at all in space, in the absence of mass. Rigorously, Maxwell's equations are mass-to-mass transforms, since Maxwell and everyone at the time assumed the ubiquitous presence of the material ether. In other words, to the founding electrodynamicists, there was no place in all the universe that was devoid of mass. Hence they defined all EM entities as entities containing mass. Even though Michelson-Morley experiments of more than a century ago destroyed that material ether, not a single Maxwellian equation has ever been changed! They still assume the material ether.

Specifically missing from our electrodynamics are two essential transforms: (1) the mass-to-spacetime transform, and (2) the spacetime-to-mass transform. In other words, these two transforms are "infolded" or hidden inside (comprise) the end point extremum case – the mass-to-mass transform actually written by Maxwell.

But these two missing transforms are just general relativity (they correspond precisely to Wheeler's principle of general relativity). So general relativity has always been infolded inside electrodynamics as we know it, but ignored. Further, this is a very powerful general relativity, since it uses the very strong EM force as the agent of spacetime curvature. Hence now one can engineer a powerful general relativity in the laboratory and in practical devices, and even action-at-a-distance, which one cannot do using merely the staid and weak G-force.

It turns out that longitudinal EM waves as shown by E.T. Whittaker in 1903 and 1904 are far more primary than our present EM theory and even our present GR theory. Both EM and GR are directly engineerable using longitudinal EM waves, including subluminal and superluminal EM waves in vacuum, and including action at a distance. One can engineer complexes of spacetime curvatures as desired, which in turn are *spacetime engines* (Wheeler's term). These engines can act upon mass – either living or inert – in any fashion desired, and at any distance desired.

The little-known and poorly understood Regeneration & Recovery (R&R) system – as contrasted to the immune system – utilizes this exact "spacetime engine" methodology to restore damaged or diseased cells, within its limitations – "dedifferentiating" or "time-reversing" them back to normal. The mechanism is fully explained in the attached package. The proposed program applies this fundamental method to greatly amplify the effect. Thus cancer cells can be reversed back to normal cells, HIV-infected cells (HIV-factories) can be reversed back to normal cells (genetics and all), etc. This represents a dramatic and unparalleled new medical therapeutic methodology of extreme power. The same techniques can simply be conjugated one additional time, and used to produce diseases and cellular changes at a distance, including in entire mass populations, as the KGB is doing right now. Specifically, by impressing an extremely weak "cocktail mix" of disease spacetime engines on the U.S. populace, the immune systems are reacting and being "spread thin" across several "shadow state" infectious pathogenic conditions. Thus when an actual terrorist BW attack occurs, the lethality of the strike will be remarkably enhanced because the immune systems will be "thinned" in their ability to fend off this new pathogen. By doing it this way, BW warfare can actually be conducted upon the U.S. without anyone being the wiser.

Our very survival as a nation is threatened by remarkable KGB weapons using this fully documented, extended electrodynamics. Bluntly, unless we take off our scientific blinders and understand Russian energetics weapon science, GWS and the Embassy health changes in Moscow were just a picnic before what is upon us. After two decades of work on the KGB energetics weapon problem, we have fully deciphered the technical mechanisms and what energetics is, as well as citing many of the Russian weapon tests and how they were done. Again, you can see that for yourself in the package.

Now that we understand the technical mechanisms, we are proposing rapid redevelopment of the Prioré process, in portable suitcase-sized units to be mass produced and filtered down into the entire emergency community for mass treatment of lethal infectious diseases resulting from BW strikes. And we are strongly urging the U.S. scientific community to quickly investigate and come to grips with this vast new *biological engineering* mechanism. Our very survival depends on it.

The inclosed documents provide an overview of the process. There is nothing else on the scientific horizon that can save the majority of all those stricken Americans, once the first foreign-sponsored terrorist teams strike our population centers with BW agents.

Sincerely,

Incls: Full package with table of contents

T.E. Bearden
President & CEO

The Tom Bearden Website

Energy, Observations and Comments

---

1. A Conservative Field Requires a Single-Valued Potential
2. Earth-Electrosphere and Earth-Ionosphere Capacitors
3. Earth's Poynting Energy Flow
4. The Hidden Energy Flow around us
5. Electrostatic Fields of the Earth's Electrosphere Capacitor in Fair Weather and Thunderstorms
6. Flow of Current and Energy Into, Through, and Along a Conductor
7. Energy "collected" on a charge
8. Fiber Fuse Effect
9. Impact of Vacuum Engineering
10. Lawandy's Experiment
11. Field may be nonconservative for a multivalued potential
12. The Ostrich Problem - Part 1
13. The Ostrich Problem - Part 2
14. Russia had Overunity Devices in 1930s
15. Energy output of human body
16. A "static" Poynting S-flow free-energy generator
17. Topology and Energy Conservation
18. Block Diagram: Sweet's Vacuum Triode

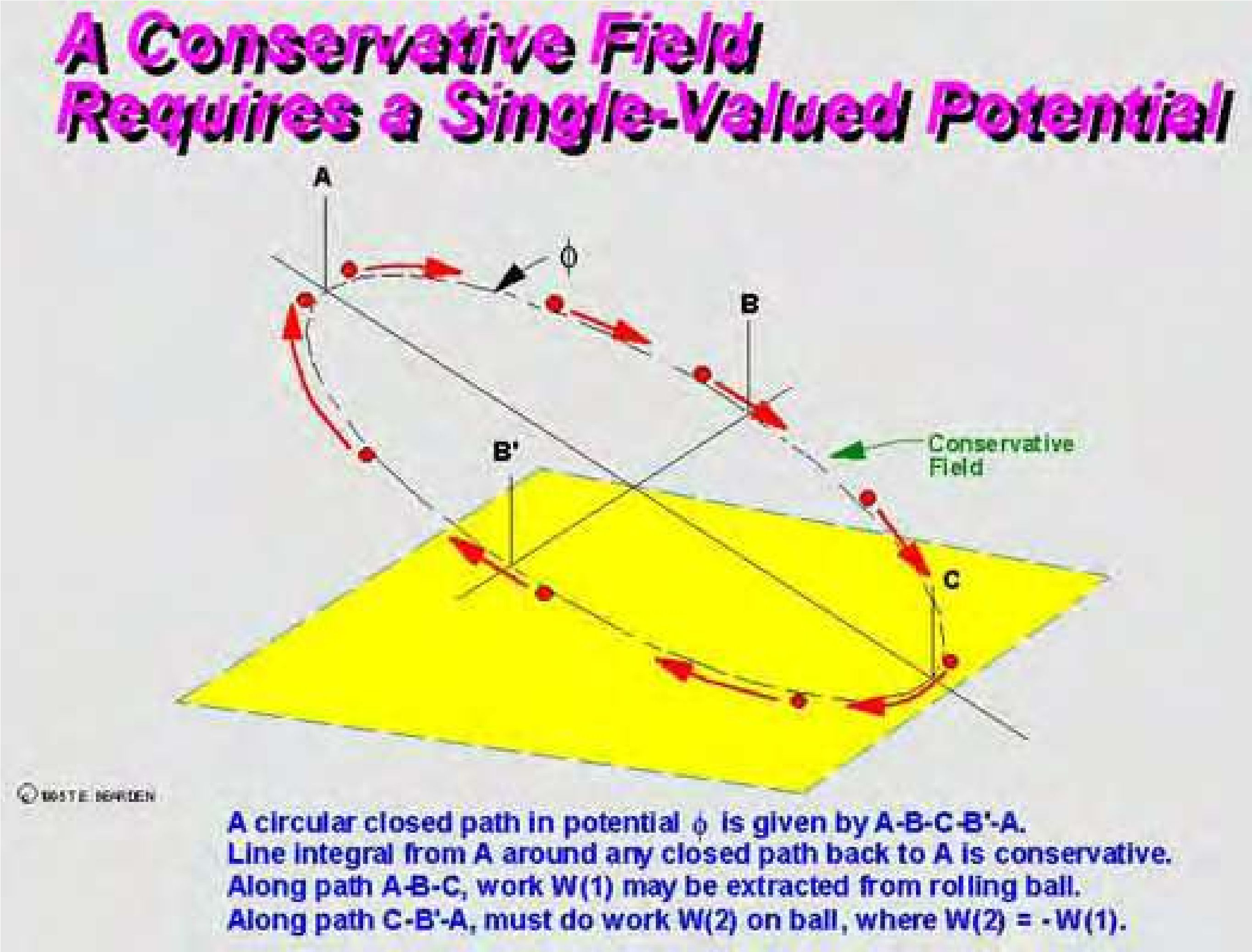

A Conservative Field Requires a Single-Valued Potential

© 1995 T.E. BEARDEN

A circular closed path in potential φ is given by A-B-C-B'-A.
Line integral from A around any closed path back to A is conservative.
Along path A-B-C, work W(1) may be extracted from rolling ball.
Along path C-B'-A, must do work W(2) on ball, where W(2) = - W(1).

Return to Slide Index

Next Slide

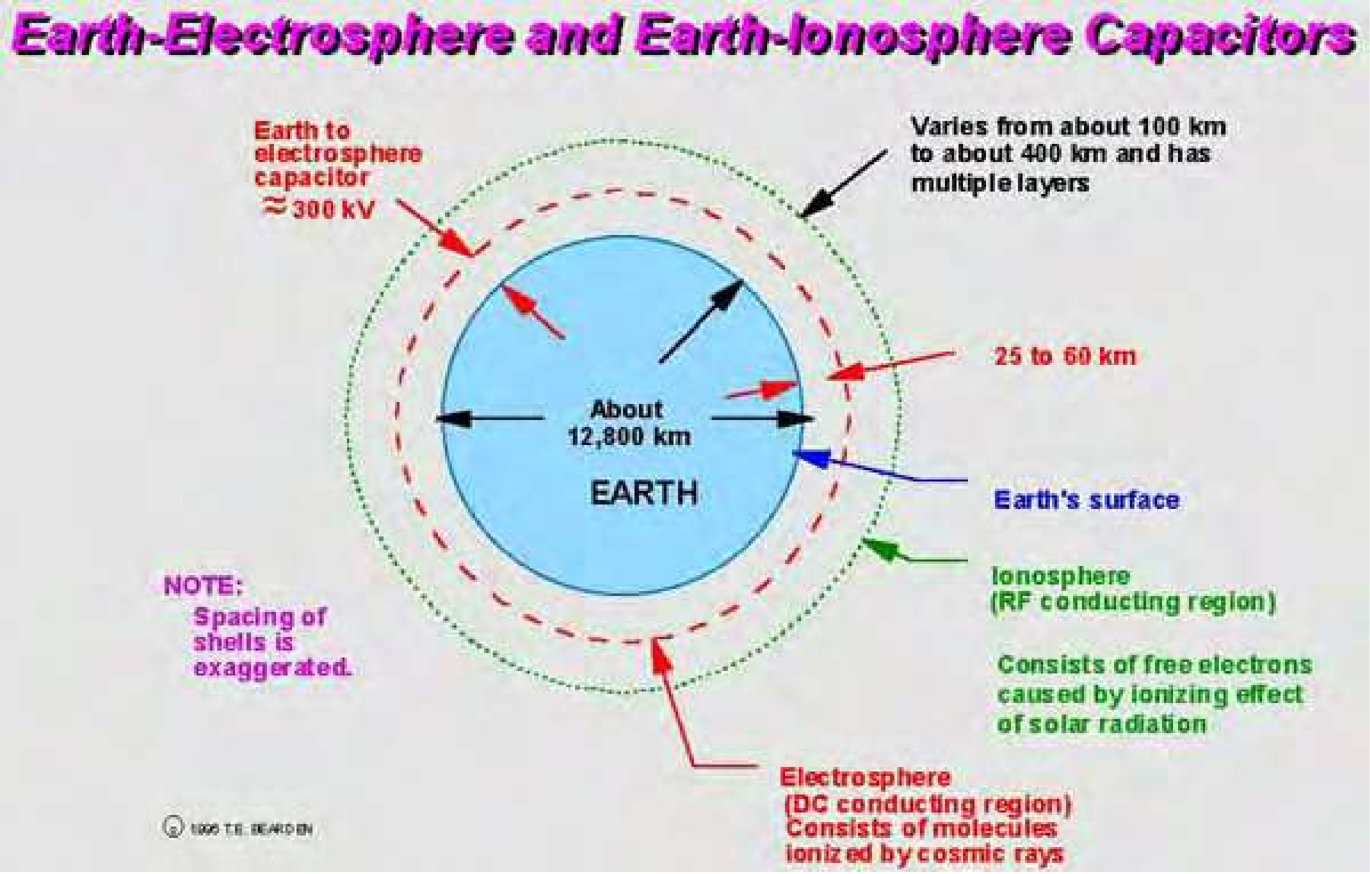

Earth-Electrosphere and Earth-Ionosphere Capacitors

Earth to electrosphere capacitor ≈300 kV

Varies from about 100 km to about 400 km and has multiple layers

Ionosphere (RF conducting region)

Consists of free electrons caused by ionizing effect of solar radiation

NOTE:
Spacing of shells is exaggerated.

Electrosphere (DC conducting region)
Consists of molecules ionized by cosmic rays

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

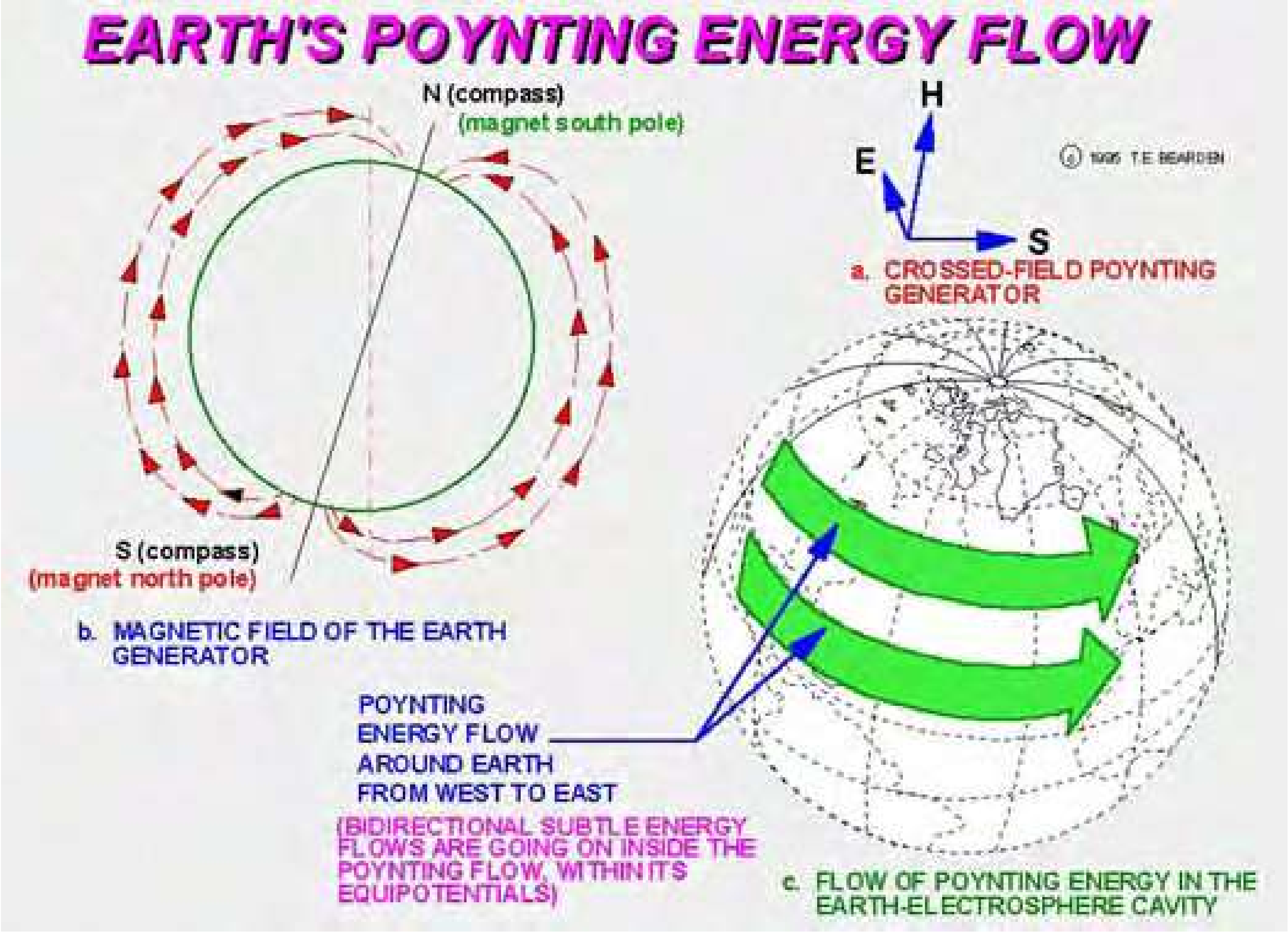

EARTH'S POYNTING ENERGY FLOW

N (compass)
(magnet south pole)
S (compass)
(magnet north pole)
H
E
S
© 1995 T.E. BEARDEN

a. CROSSED-FIELD POYNTING GENERATOR

b. MAGNETIC FIELD OF THE EARTH GENERATOR

POYNTING ENERGY FLOW AROUND EARTH FROM WEST TO EAST

(BIDIRECTIONAL SUBTLE ENERGY FLOWS ARE GOING ON INSIDE THE POYNTING FLOW, WITHIN ITS EQUIPOTENTIALS)

c. FLOW OF POYNTING ENERGY IN THE EARTH-ELECTROSPHERE CAVITY

Return to Slide Index

Next Slide

Previous Slide

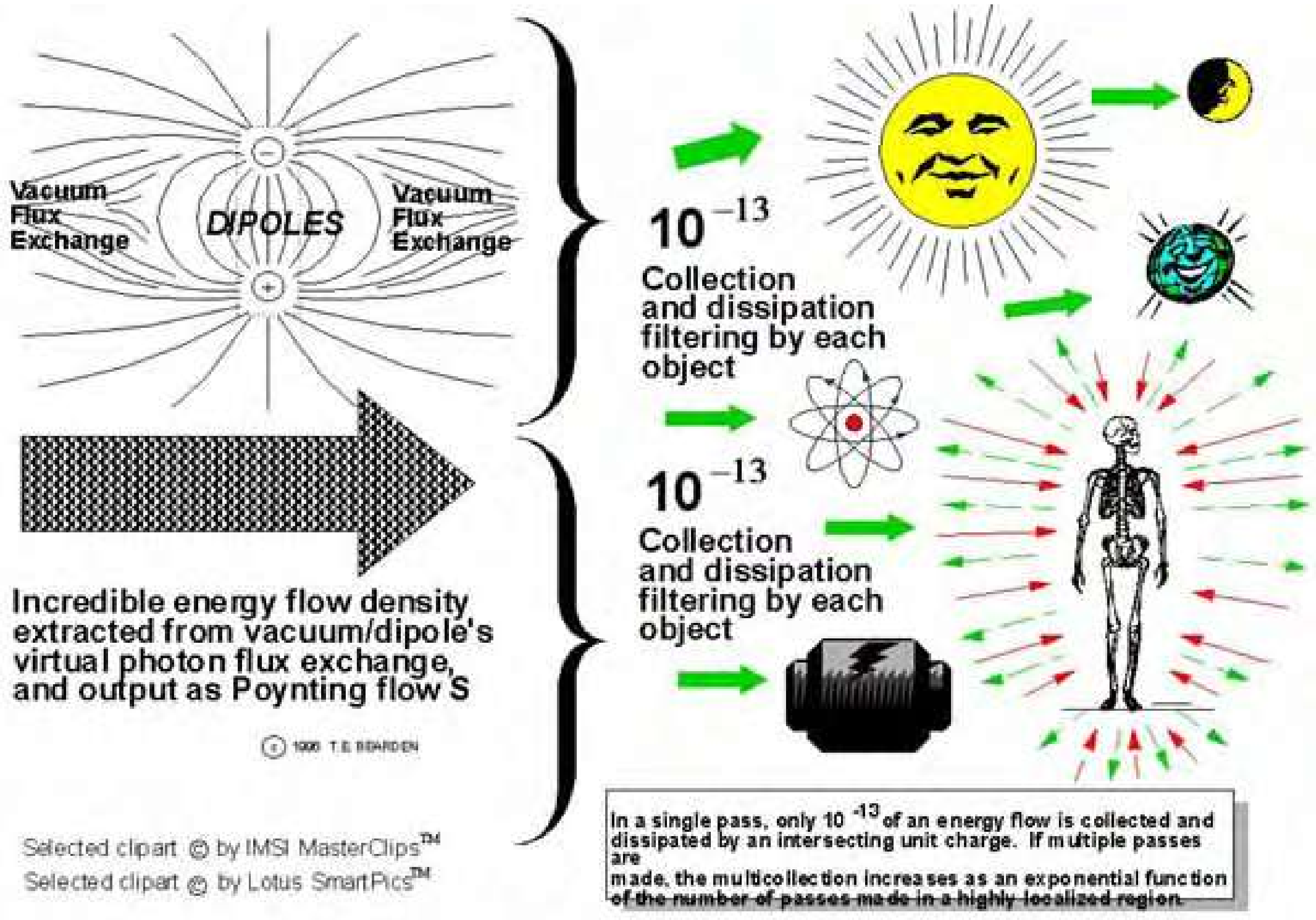
Vacuum Flux Exchange
DIPOLES
Vacuum Flux Exchange
Incredible energy flow density extracted from vacuum/dipole's virtual photon flux exchange, and output as Poynting flow S
© 1996 T.E. BEARDEN
Selected clipart © by IMSI MasterClips™
Selected clipart © by Lotus SmartPics™
10^{-13}
Collection and dissipation filtering by each object
10^{-13}
Collection and dissipation filtering by each object

In a single pass, only 10^{-13} of an energy flow is collected and dissipated by an intersecting unit charge. If multiple passes are made, the multicollection increases as an exponential function of the number of passes made in a highly localized region.

Figure 19. The world of physical matter exists behind a dramatic 10^{-13} filter in a fiery cauldron of fierce and primeval EM energy flow.

Return to Slide Index

Next Slide

Previous Slide

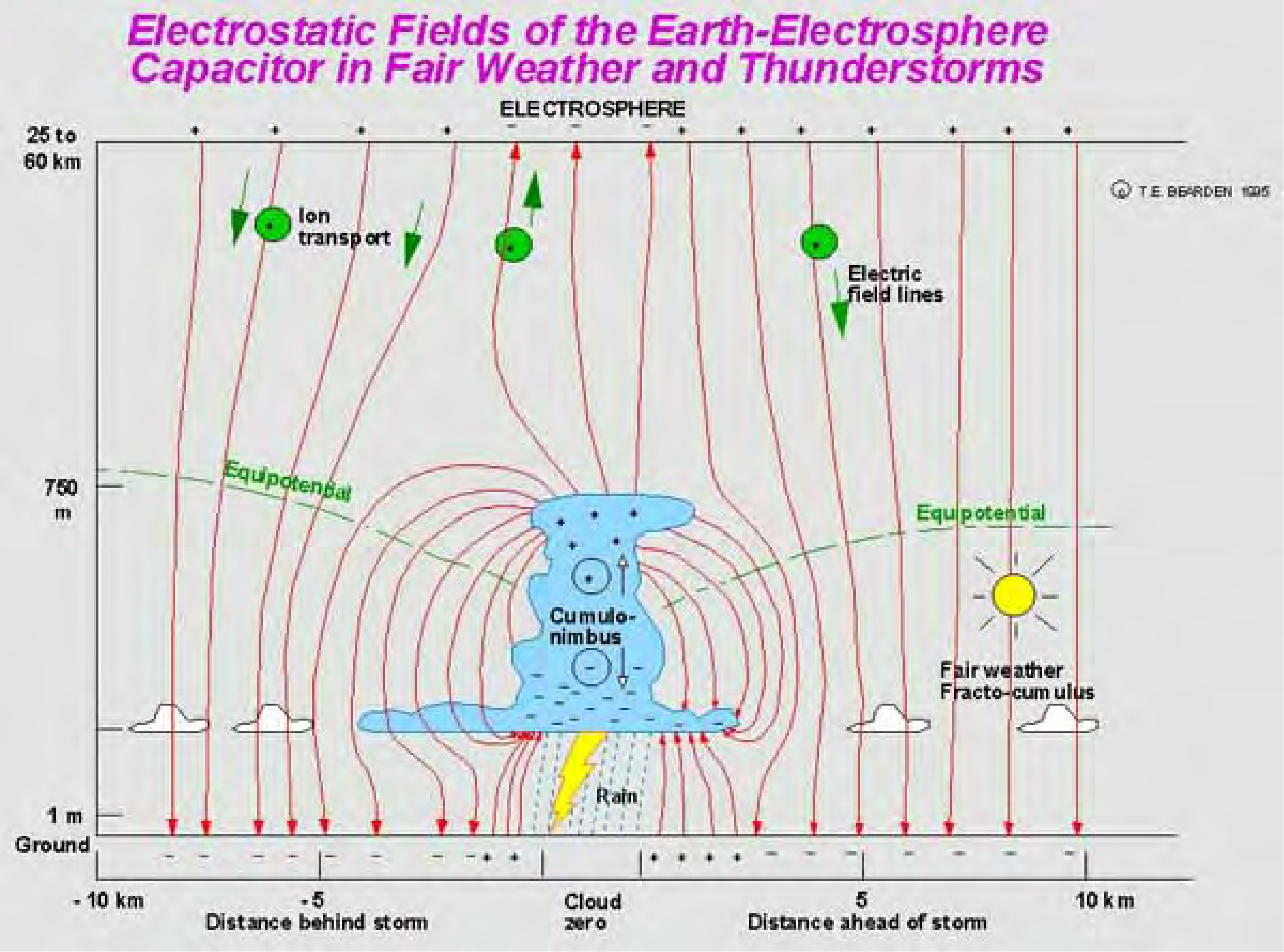

Electrostatic Fields of the Earth-Electrosphere Capacitor in Fair Weather and Thunderstorms

ELECTROSPHERE
25 to 60 km
750 m
1 m
Ground
Ion transport
© T.E. BEARDEN 1995
Electric field lines
Equipotential
Equipotential
Cumulo-nimbus
Rain
Fair weather Fracto-cumulus
- 10 km
- 5
Distance behind storm
Cloud zero
5
Distance ahead of storm
10 km

Return to Slide Index

Next Slide

Previous Slide

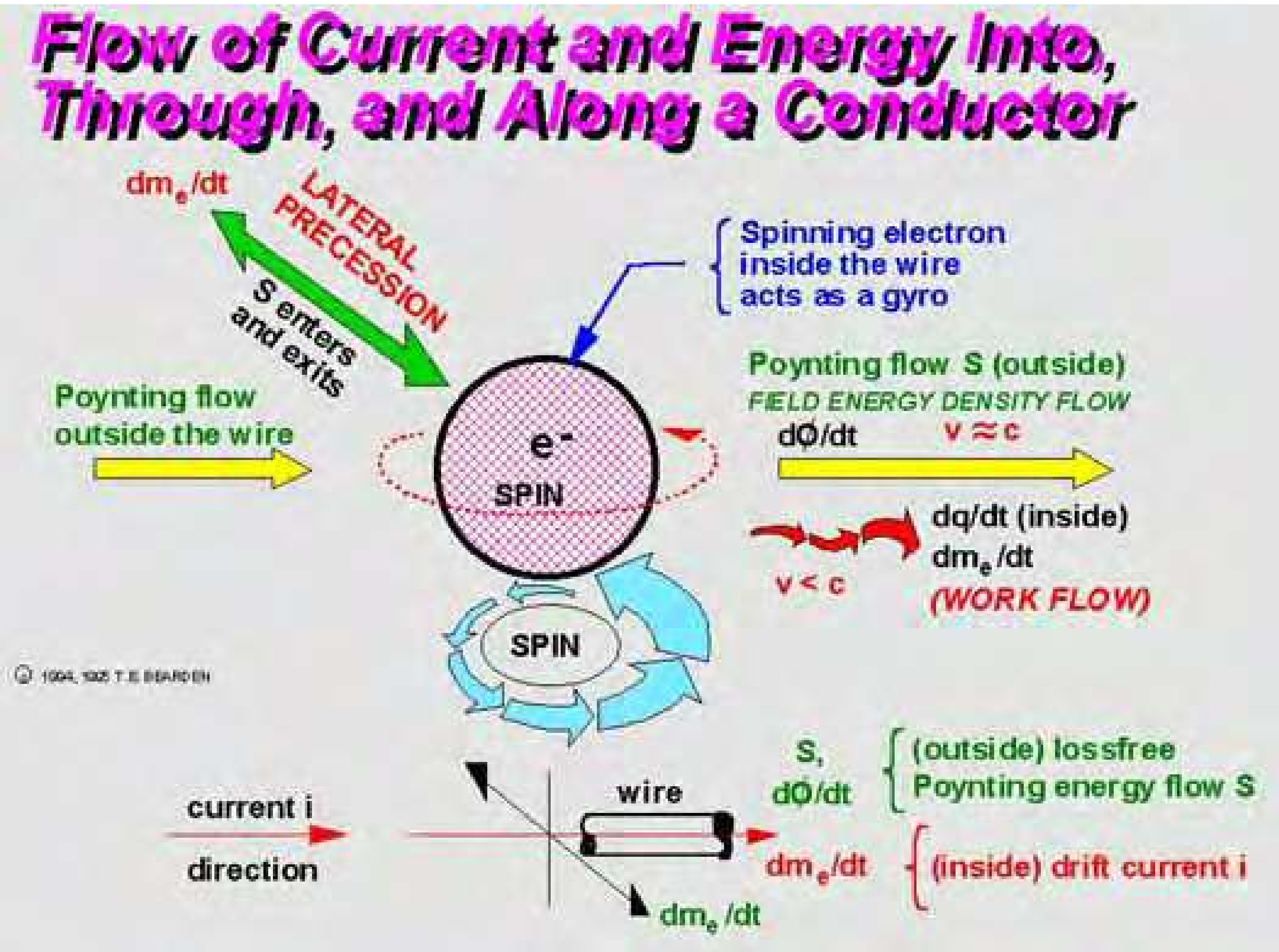

Flow of Current and Energy Into, Through, and Along a Conductor

dm_e/dt
LATERAL PRECESSION
S enters and exits
Spinning electron inside the wire acts as a gyro
Poynting flow outside the wire
e^-
SPIN
Poynting flow S (outside)
FIELD ENERGY DENSITY FLOW
$d\phi/dt$
$v \approx c$
dq/dt (inside)
dm_e/dt
$v < c$
(WORK FLOW)
SPIN
current i
direction
wire
S, $d\phi/dt$
(outside) lossfree Poynting energy flow S
dm_e/dt
(inside) drift current i
dm_e/dt

© 1994, 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

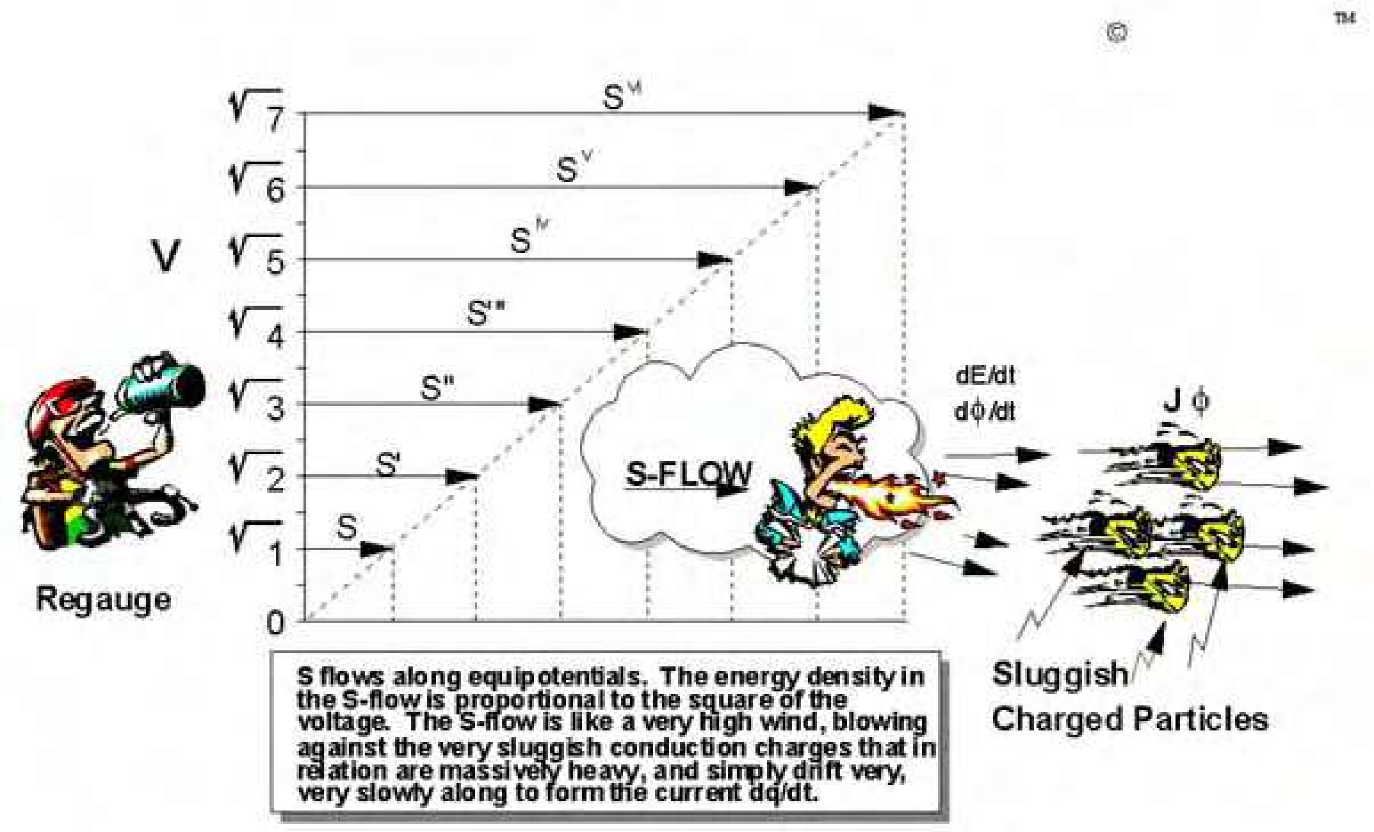
©
TM
V
√7
√6
√5
√4
√3
√2
√1
0
S
S'
S''
S'''
S^IV
S^V
S^VI
Regauge
S-FLOW
dE/dt
dφ/dt
J φ
Sluggish
Charged Particles

S flows along equipotentials. The energy density in the S-flow is proportional to the square of the voltage. The S-flow is like a very high wind, blowing against the very sluggish conduction charges that in relation are massively heavy, and simply drift very, very slowly along to form the current dq/dt.

Figure 17. Energy "collected" on a charge depends upon the charge being in a continuous Poynting S-flow. When S ceases, there is no excess energy "collected" on the charge. Energy collection never occurs in chunks, but is always a dynamic, ongoing process.

Return to Slide Index

Next Slide

Previous Slide

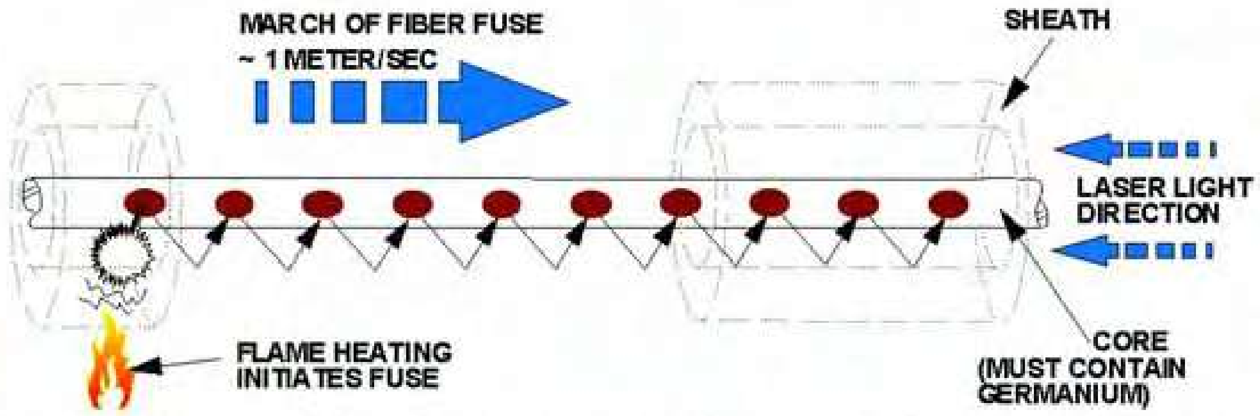
MARCH OF FIBER FUSE
~ 1 METER/SEC
SHEATH
LASER LIGHT DIRECTION
FLAME HEATING INITIATES FUSE
CORE (MUST CONTAIN GERMANIUM)

31A. Fiber fuse destroys core of fiber optics cable of indefinite length, pitting core with holes.

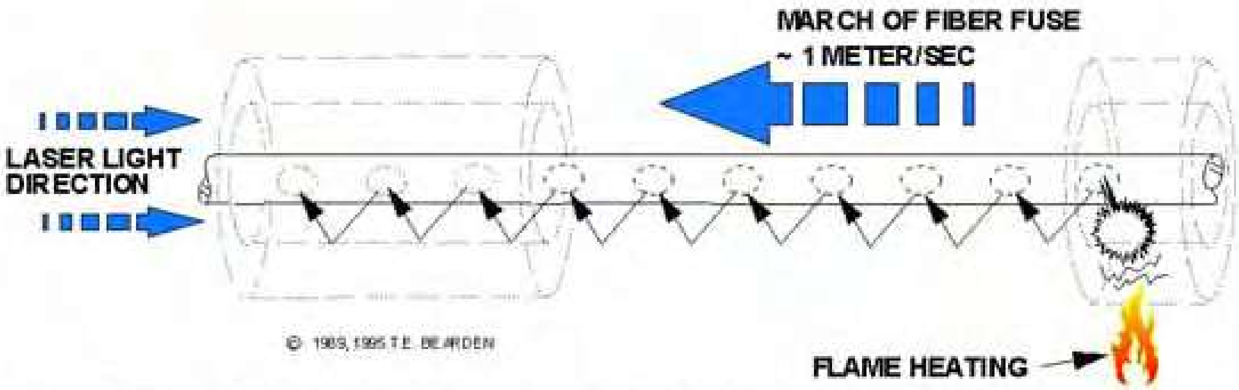
MARCH OF FIBER FUSE
~ 1 METER/SEC
LASER LIGHT DIRECTION
FLAME HEATING

© 1989, 1995 T.E. BEARDEN

31B. Reversed second fiber fuse often *restores* pitted core in cable, filling holes back up.

Figure 31. Fiber fuse effect and its strange anomalies.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

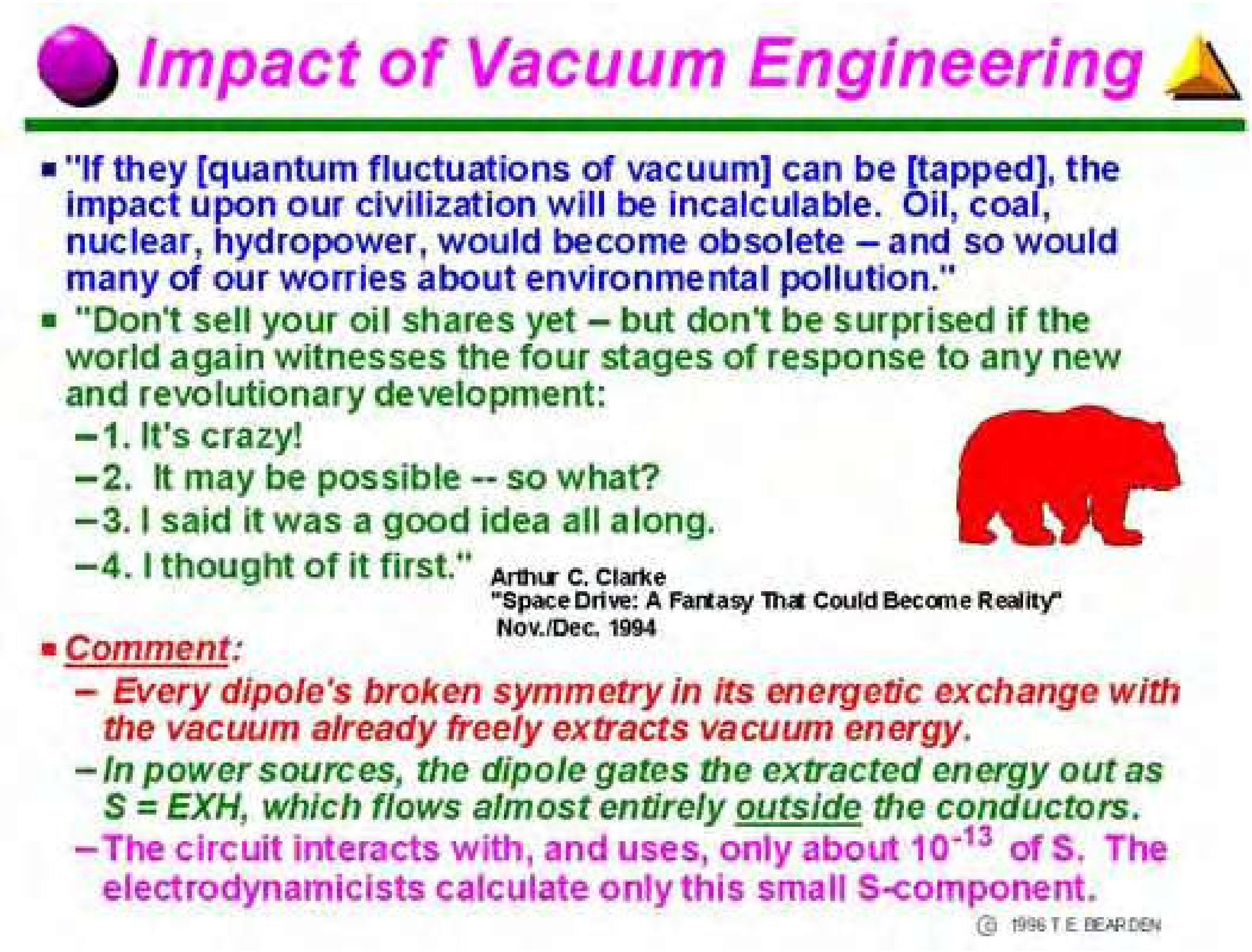

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete – and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet – but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

- *Comment:*
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

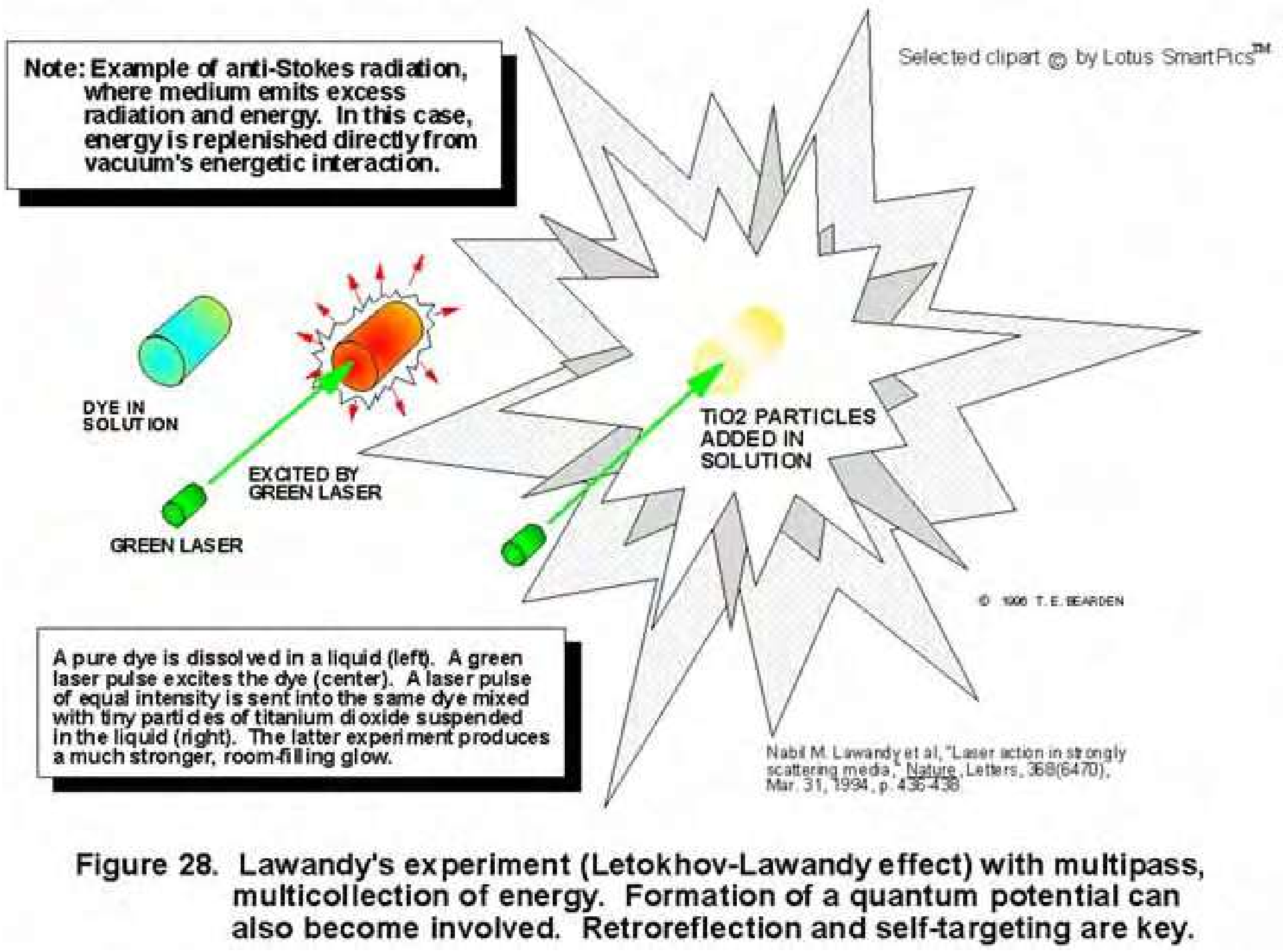
Selected clipart © by Lotus SmartPics™
DYE IN SOLUTION
EXCITED BY GREEN LASER
GREEN LASER
TiO2 PARTICLES ADDED IN SOLUTION
© 1996 T. E. BEARDEN

Note: Example of anti-Stokes radiation, where medium emits excess radiation and energy. In this case, energy is replenished directly from vacuum's energetic interaction.

A pure dye is dissolved in a liquid (left). A green laser pulse excites the dye (center). A laser pulse of equal intensity is sent into the same dye mixed with tiny particles of titanium dioxide suspended in the liquid (right). The latter experiment produces a much stronger, room-filling glow.

Nabil M. Lawandy et al, "Laser action in strongly scattering media," Nature, Letters, 368(6470), Mar. 31, 1994, p. 436-438

Figure 28. Lawandy's experiment (Letokhov-Lawandy effect) with multipass, multicollection of energy. Formation of a quantum potential can also become involved. Retroreflection and self-targeting are key.

Return to Slide Index

Next Slide

Previous Slide

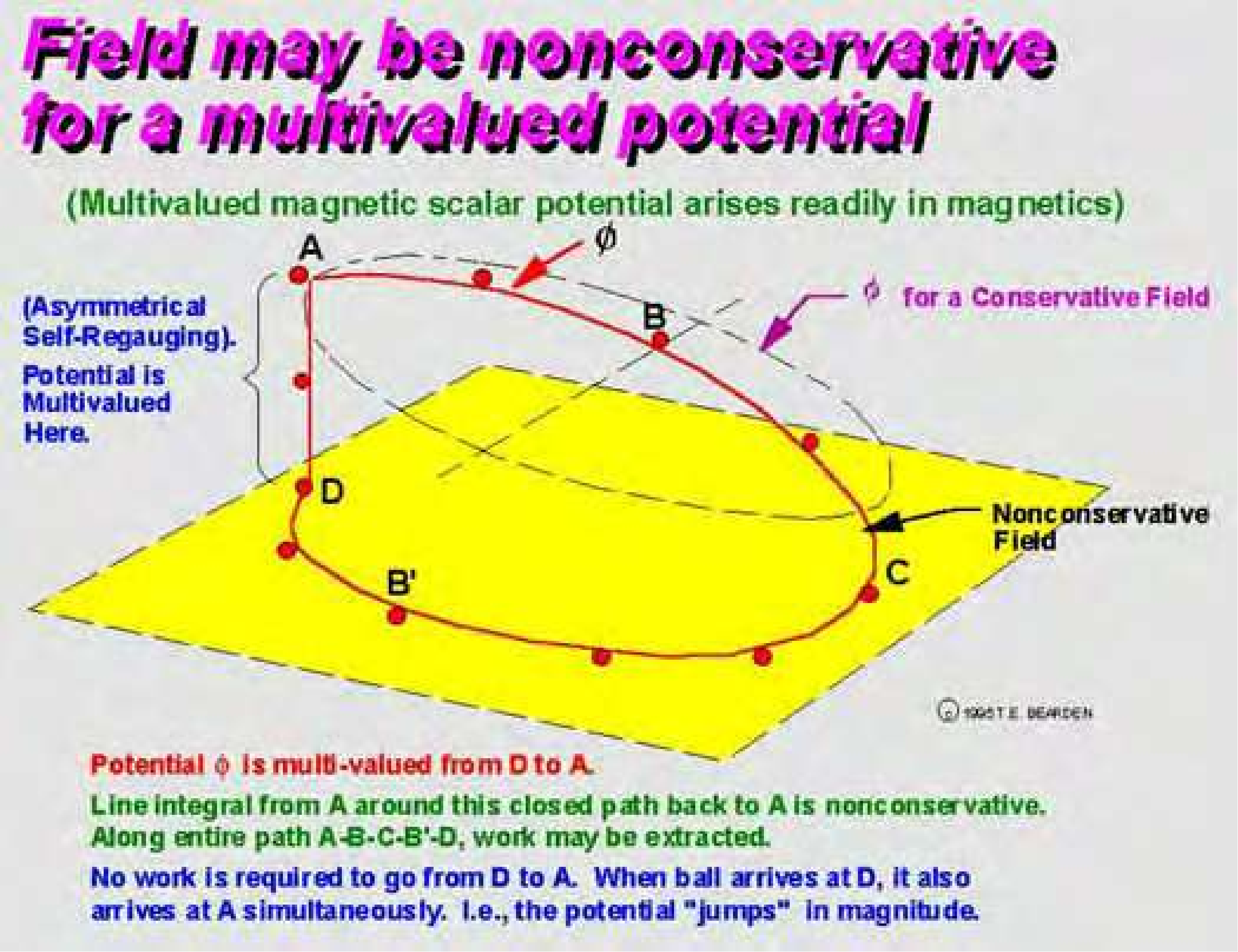

Field may be nonconservative for a multivalued potential

(Multivalued magnetic scalar potential arises readily in magnetics)

A
ϕ
B
ϕ for a Conservative Field
D
B'
C
Nonconservative Field
© 1995 T.E. BEARDEN

(Asymmetrical Self-Regauging). Potential is Multivalued Here.

Potential ϕ is multi-valued from D to A.

Line integral from A around this closed path back to A is nonconservative. Along entire path A-B-C-B'-D, work may be extracted.

No work is required to go from D to A. When ball arrives at D, it also arrives at A simultaneously. I.e., the potential "jumps" in magnitude.

Return to Slide Index

Next Slide

Previous Slide

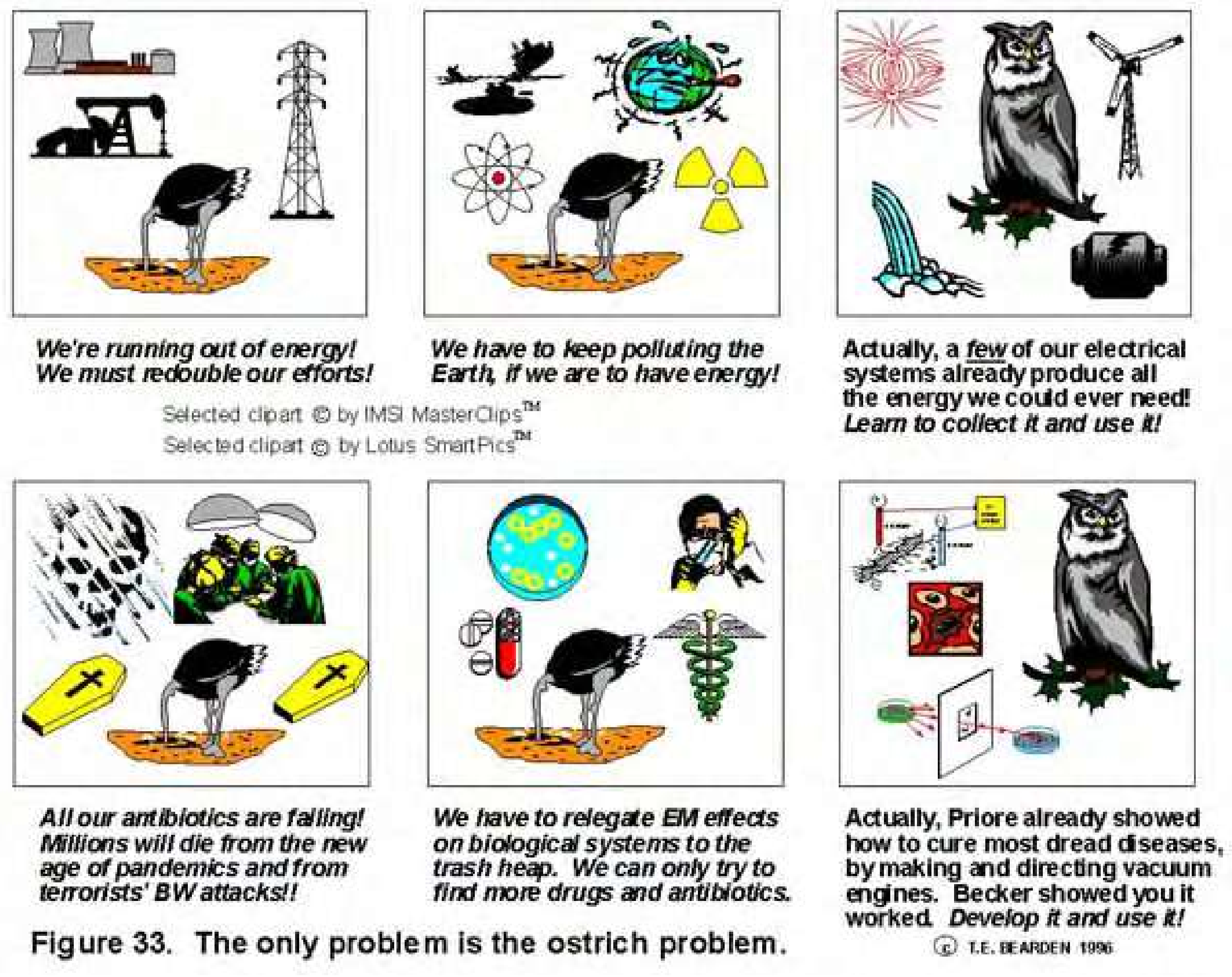

We're running out of energy! We must redouble our efforts!

We have to keep polluting the Earth, if we are to have energy!

Actually, a few of our electrical systems already produce all the energy we could ever need! *Learn to collect it and use it!*

Selected clipart © by IMSI MasterClips™
Selected clipart © by Lotus SmartPics™

All our antibiotics are failing! Millions will die from the new age of pandemics and from terrorists' BW attacks!!

We have to relegate EM effects on biological systems to the trash heap. We can only try to find more drugs and antibiotics.

Actually, Priore already showed how to cure most dread diseases, by making and directing vacuum engines. Becker showed you it worked. *Develop it and use it!*

Figure 33. The only problem is the ostrich problem.

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

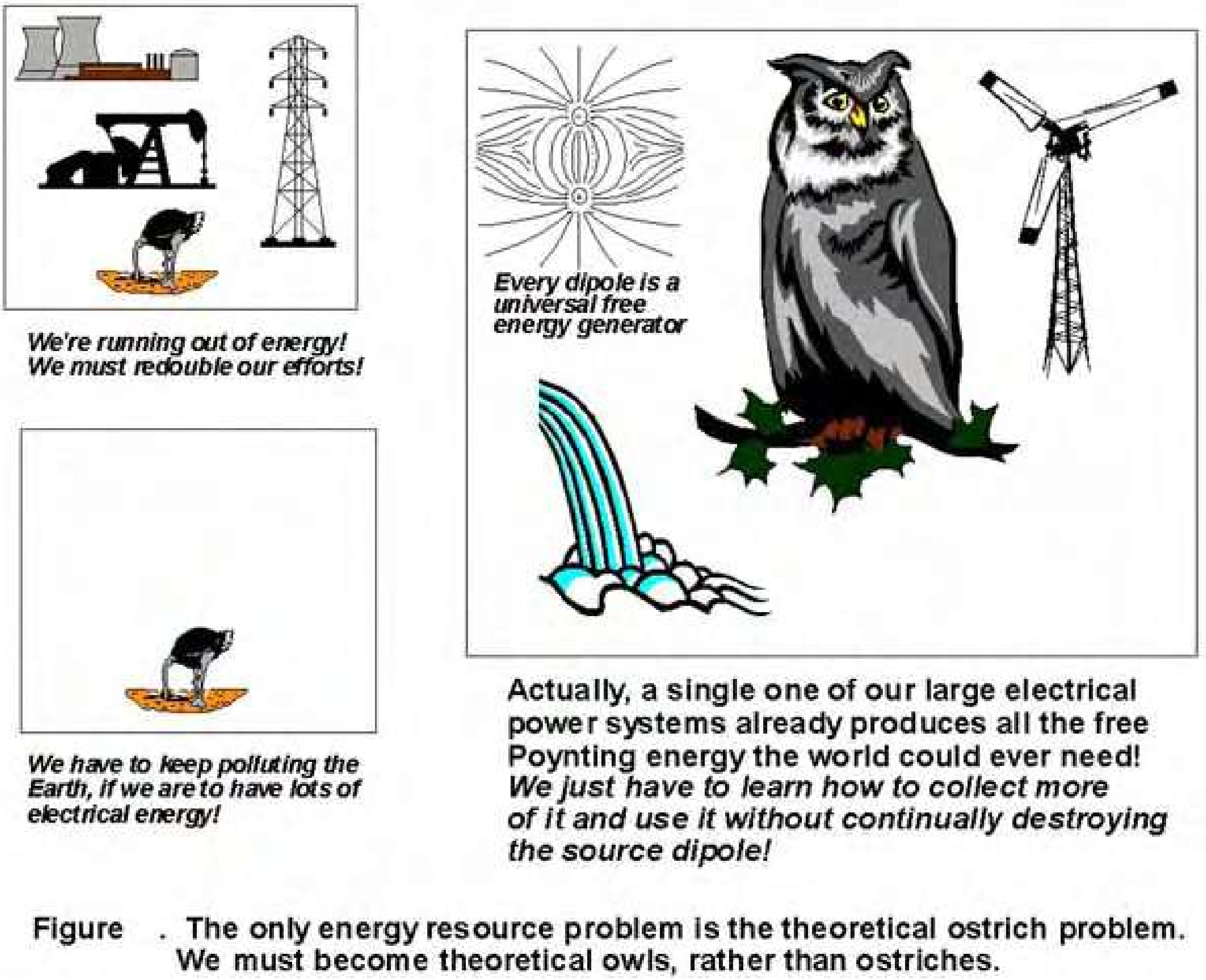

*We're running out of energy!
We must redouble our efforts!*

We have to keep polluting the Earth, if we are to have lots of electrical energy!

Every dipole is a universal free energy generator

Actually, a single one of our large electrical power systems already produces all the free Poynting energy the world could ever need! *We just have to learn how to collect more of it and use it without continually destroying the source dipole!*

Figure . The only energy resource problem is the theoretical ostrich problem. We must become theoretical owls, rather than ostriches.

Return to Slide Index

Next Slide

Previous Slide

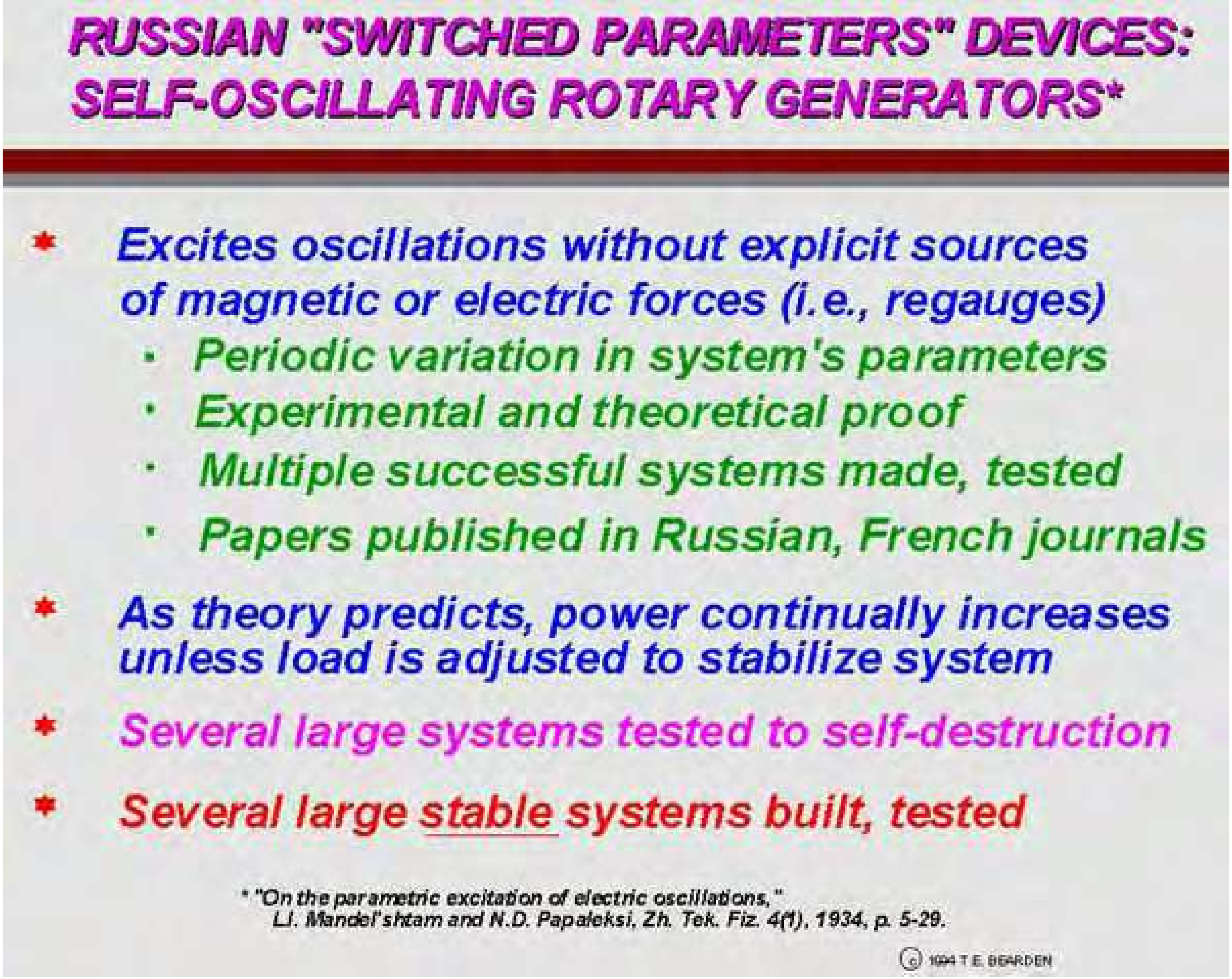

RUSSIAN "SWITCHED PARAMETERS" DEVICES: SELF-OSCILLATING ROTARY GENERATORS*

- Excites oscillations without explicit sources of magnetic or electric forces (i.e., regauges)
  - Periodic variation in system's parameters
  - Experimental and theoretical proof
  - Multiple successful systems made, tested
  - Papers published in Russian, French journals
- As theory predicts, power continually increases unless load is adjusted to stabilize system
- Several large systems tested to self-destruction
- Several large stable systems built, tested

* "On the parametric excitation of electric oscillations," L.I. Mandel'shtam and N.D. Papaleksi, Zh. Tek. Fiz. 4(1), 1934, p. 5-29.

ⓒ 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

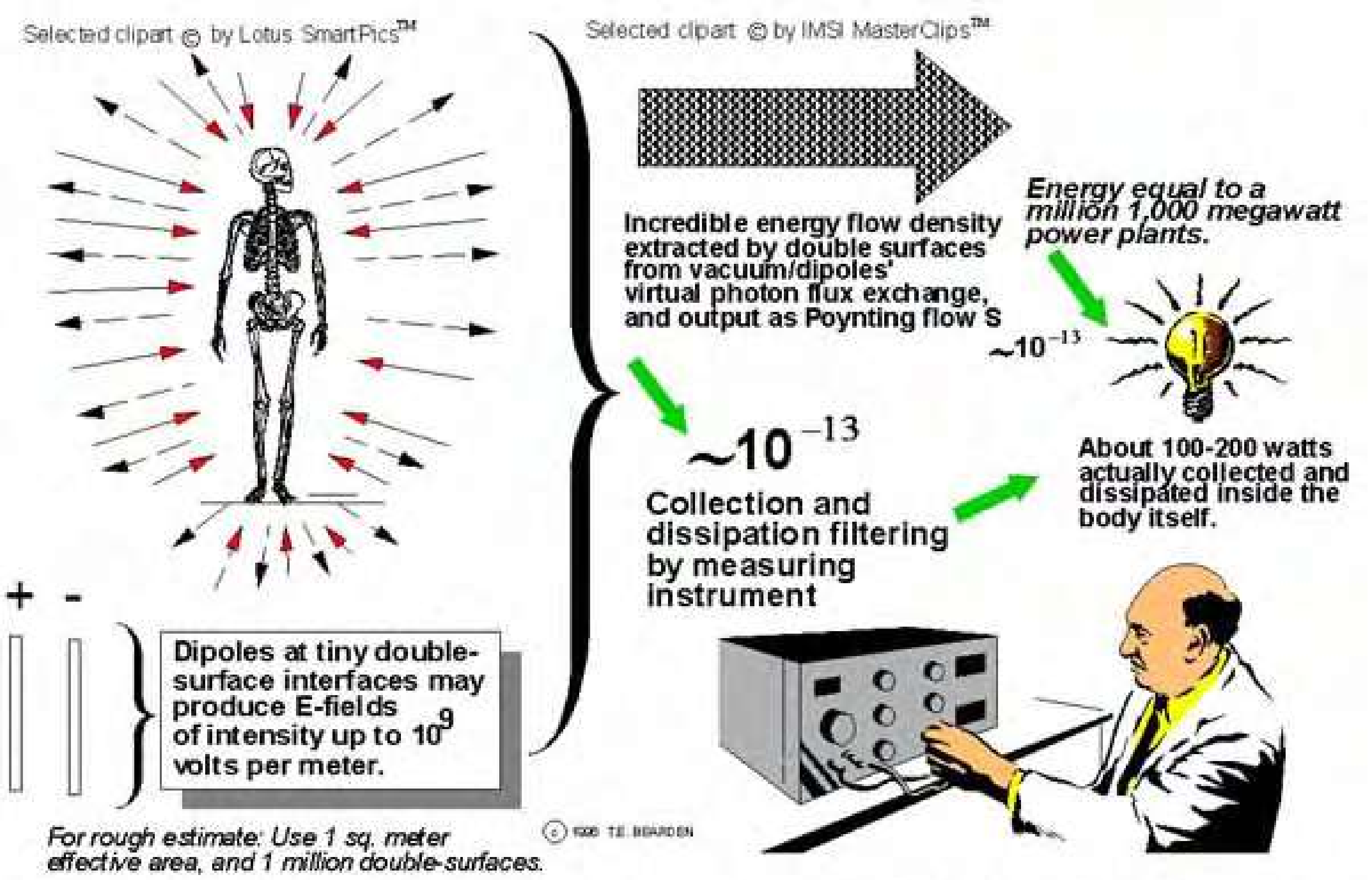
Selected clipart © by Lotus SmartPics™
Selected clipart © by IMSI MasterClips™
+ -
~10^{-13}
~10^{-13}
© 1996 T.E. BEARDEN

Incredible energy flow density extracted by double surfaces from vacuum/dipoles' virtual photon flux exchange, and output as Poynting flow S

Energy equal to a million 1,000 megawatt power plants.

About 100-200 watts actually collected and dissipated inside the body itself.

Collection and dissipation filtering by measuring instrument

Dipoles at tiny double-surface interfaces may produce E-fields of intensity up to 10^9 volts per meter.

For rough estimate: Use 1 sq. meter effective area, and 1 million double-surfaces.

Figure 20. One human body produces more energy flow than all the electrical power loads on earth dissipate. However, only 100-200 watts of its own energy output is collected and dissipated by the body.

Return to Slide Index

Next Slide

Previous Slide

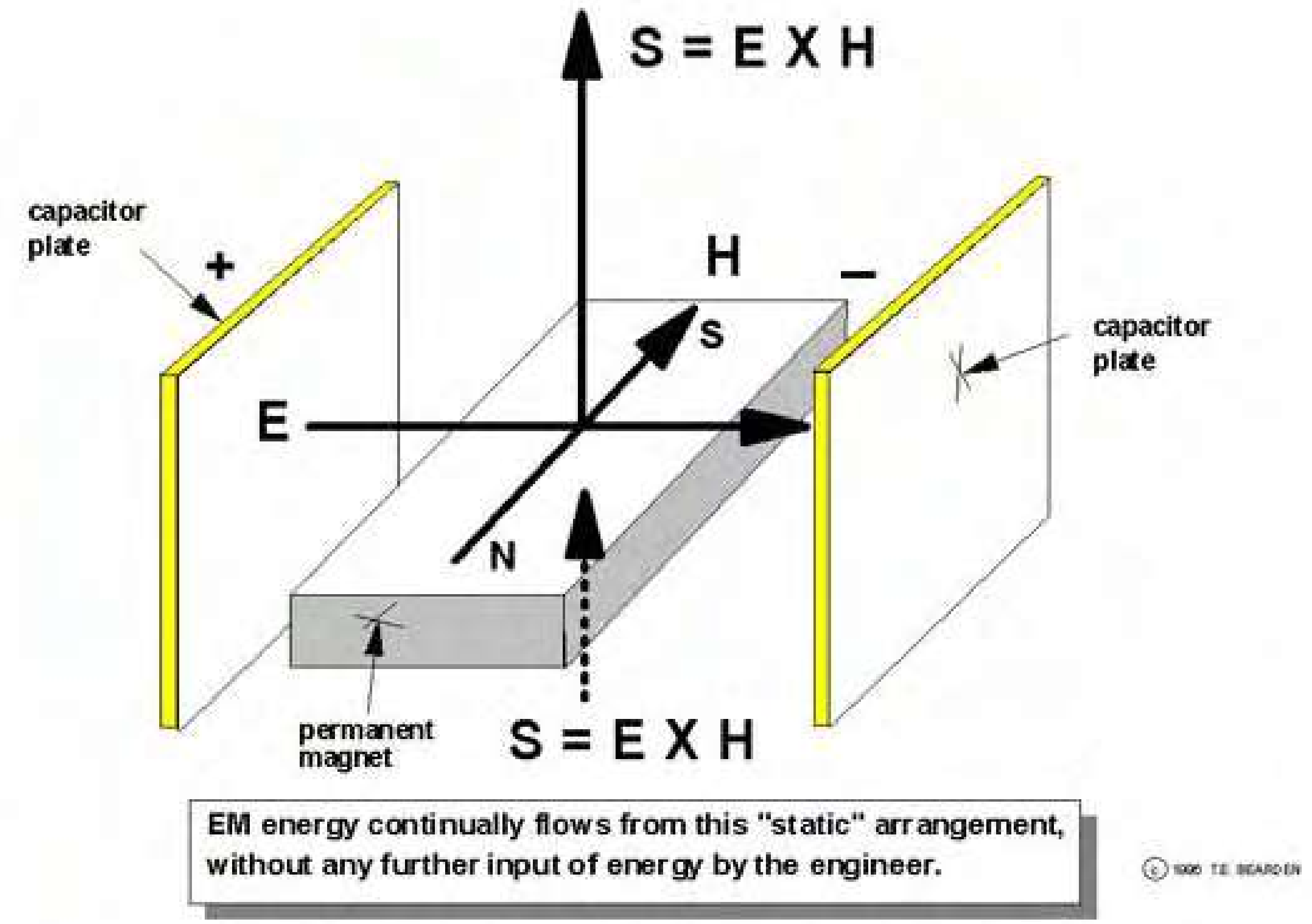
S = E X H
capacitor plate
+
H
–
S
capacitor plate
E
N
permanent magnet
S = E X H
© 1996 T.E. BEARDEN

EM energy continually flows from this "static" arrangement, without any further input of energy by the engineer.

Figure 18. A "static" Poynting S-flow free-energy generator.

Return to Slide Index

Next Slide

Previous Slide

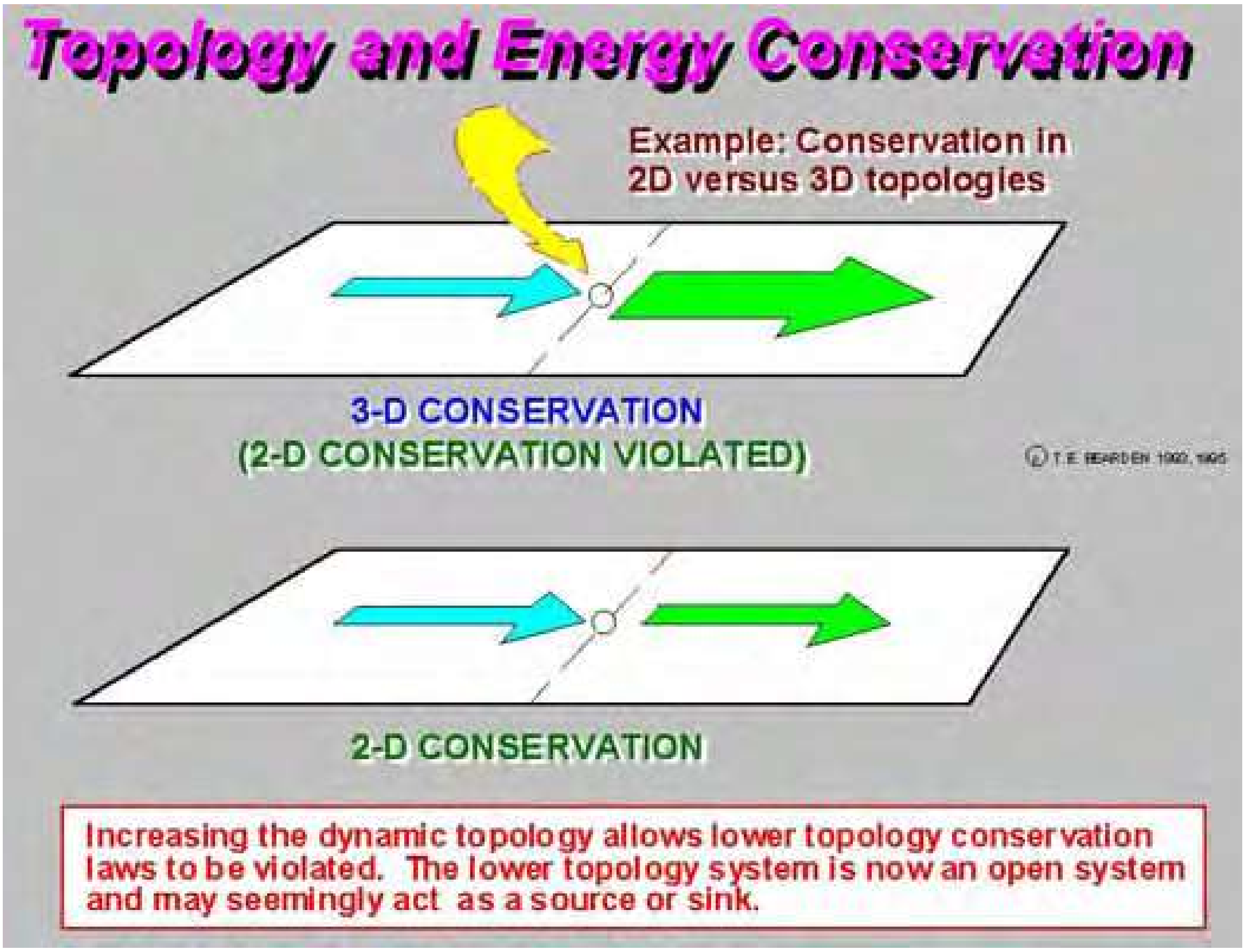

Topology and Energy Conservation

Example: Conservation in 2D versus 3D topologies

3-D CONSERVATION
(2-D CONSERVATION VIOLATED)

© T.E. BEARDEN 1992, 1995

2-D CONSERVATION

Increasing the dynamic topology allows lower topology conservation laws to be violated. The lower topology system is now an open system and may seemingly act as a source or sink.

Return to Slide Index

Next Slide

Previous Slide

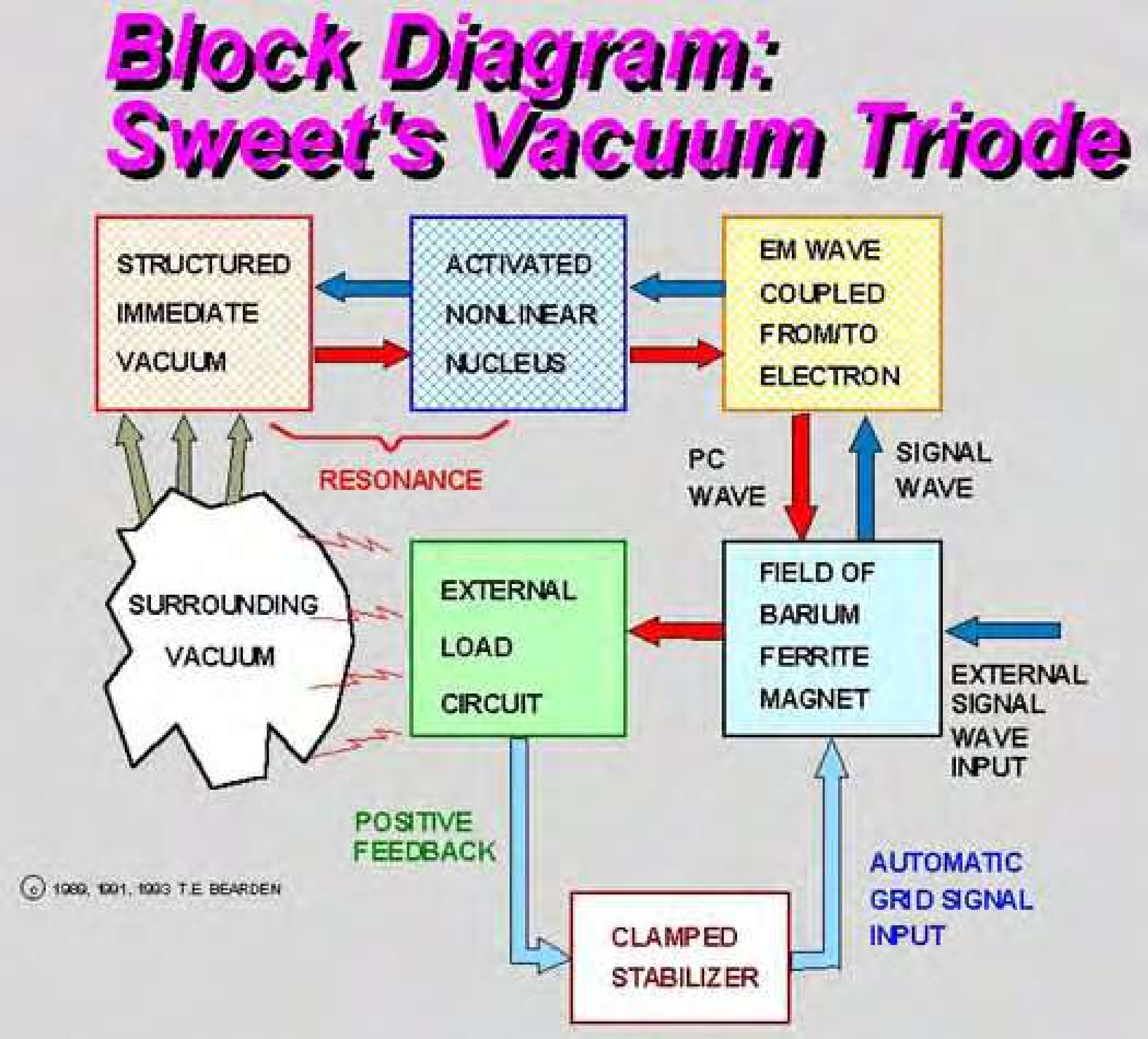

Block Diagram: Sweet's Vacuum Triode

STRUCTURED IMMEDIATE VACUUM

ACTIVATED NONLINEAR NUCLEUS

EM WAVE COUPLED FROM/TO ELECTRON

RESONANCE

PC WAVE

SIGNAL WAVE

SURROUNDING VACUUM

EXTERNAL LOAD CIRCUIT

FIELD OF BARIUM FERRITE MAGNET

EXTERNAL SIGNAL WAVE INPUT

POSITIVE FEEDBACK

AUTOMATIC GRID SIGNAL INPUT

CLAMPED STABILIZER

© 1989, 1991, 1993 T.E. BEARDEN

Return to Slide Index

Previous Slide

The Tom Bearden Website

Help support the research

Foundations of Electro-Magnetism
Slide Series

- Major Points of Presentation
- Chung's Carbon Filament Negative Resistor
- Aharanov-Bohm Effect
- Absorption and emission reactions (1)
- Absorption and emission reactions (2)
- Anti-engine for cell's deviation reverses cell back to normal
- Asymmetric regauging produces excess force, which can be used to do work on the system
- Asymmetric regauging produces excess force, which can be used to do work on the regauging system
- Becker's bone fracture healing experiments
- Becker's theoretical control system governing regeneration
- Becker's theoretical DC control system involved with response to injury
- Geometrical Distortion of Poynting Energy Flow
- Cosmic Bootstrap: Before the Big Bang
- Bridging Enables COP > 1

- Aspects of Strong Local Symmetry
- Other Aspects of Strong Local Symmetry
- Bunge on the Status of Electrodynamics
- Bunge on the Status of Electrodynamics and Physics
- Bunge on the Status of Classical and Quantum Physics
- Fractions of BW Casualties Surviving
- First Step in Self-Promotion of Cancer Due to Hypoxia
- Differentiation and Dedifferentiation Timelines
- Evolution of Cells on Earth
- A Charged Particle q is a Coupled System of m and Ø
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (1)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (2)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (3)
- A Charged Particle is a Coupled System
- The Anomalies in Navy Electrolyte Experiments at China Lake
- Chung's Carbon Filament Negative Resistor

- Typical circuit has about 10^{-13} energy collection efficiency
- Classical View of EM
- Curved Spacetime acts as source or sink
- Chronic Disease Puzzling
- Cancer Characteristics
- Solving a Cancer Enigma
- Cancer: Characteristics
- Nonlinear Optics Distortion Correction Theorem
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)
- In Time-Reversed Zone: Nuclear reactions are biased
- In Time-Reversed Zone: Dynamics may be reversed
- Time-Reversed Zone: Significant only after time-density charging
- A New Conservation of Energy Law
- Questions on Carcinogens and EM

Radiation

- AS DEFINED, fields and potentials only exist in and on charged matter
- Serious flaws and errors in classical EM theory
- Electrodynamicists' Reaction to Removal of the Material Ether
- Reaction to Loss of Material Ether (1)
- Reaction to Loss of Material Ether (2)
- Some EM waves in Spacetime
- Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String
- Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna
- Phase Conjugate Wavepairs Produce New Waves
- Scalar Electromagnetics (Energetics) View of EM (1)
- Scalar Electromagnetics (Energetics) View of EM (2)

- Scalar Electromagnetics (Energetics) View of EM (3)
- Extended General Relativity Principle
- Extension of Work-Energy Theorem
- 34 Flaws in Classical EM Theory
- MTW's General Relativity Principle
- Use of General Relativity (GR) in Particle Physics
- The Missing Infolded Electrodynamics
- Infolded Engines: Excluded by Present Physics
- Mass curves and structures spacetime; ST curvatures change and structure mass (1)
- Mass curves and structures spacetime; ST curvatures change and structure mass (2)
- At Infinite Velocity: Each Is Every Other
- Lisitsyn's Report: Brain Code Broken
- Lorentz Closed Surface Integration
- Lorentz's physically insignificant energy flow can be collected and utilized
- Lorentz surface integration of the Poynting vector around a closed surface
- Longitudinal EM Photon Interaction with Charge
- Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass.

- Pumping with Longitudinal EM waves Time-reverses the Mass
- Maxwell's Quaternion Theory
- Maxwellian Systems before and after Regauging
- Maxwell's four equations reduce to: Page 1
- Maxwell's four equations reduce to: Page 2
- More to come

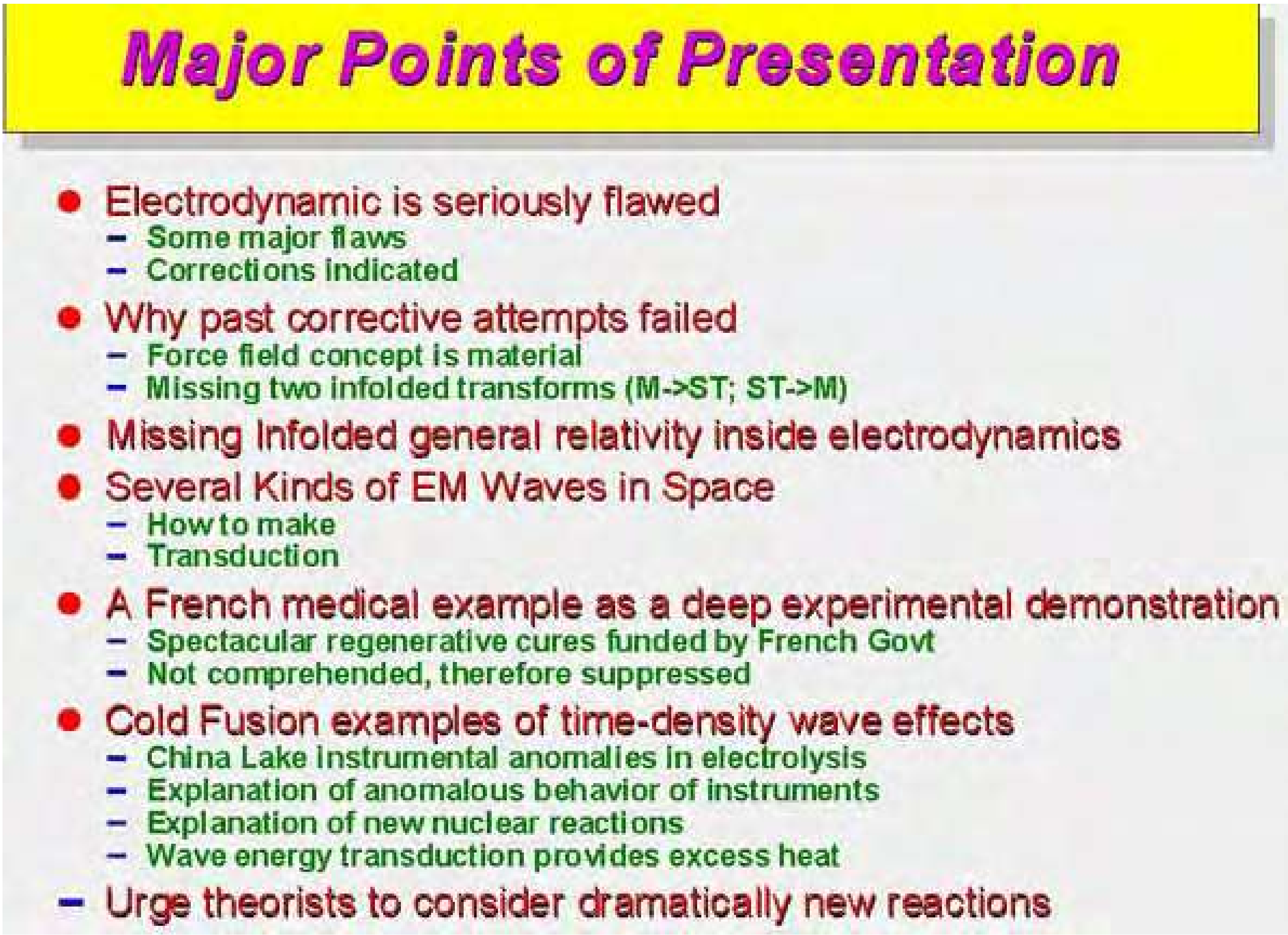

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

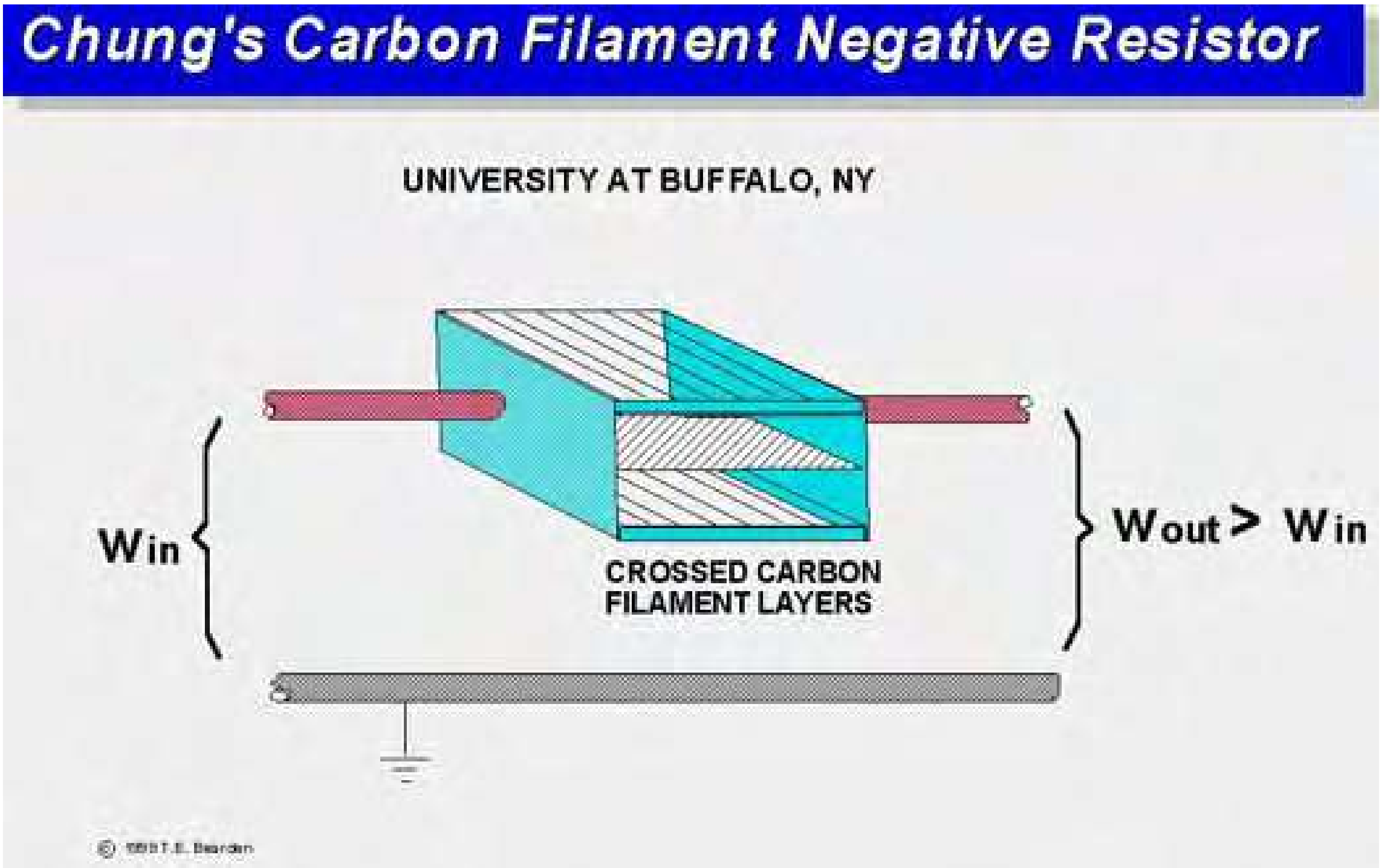

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

W_{in}

CROSSED CARBON
FILAMENT LAYERS

$W_{out} > W_{in}$

© 1998 T.E. Bearden

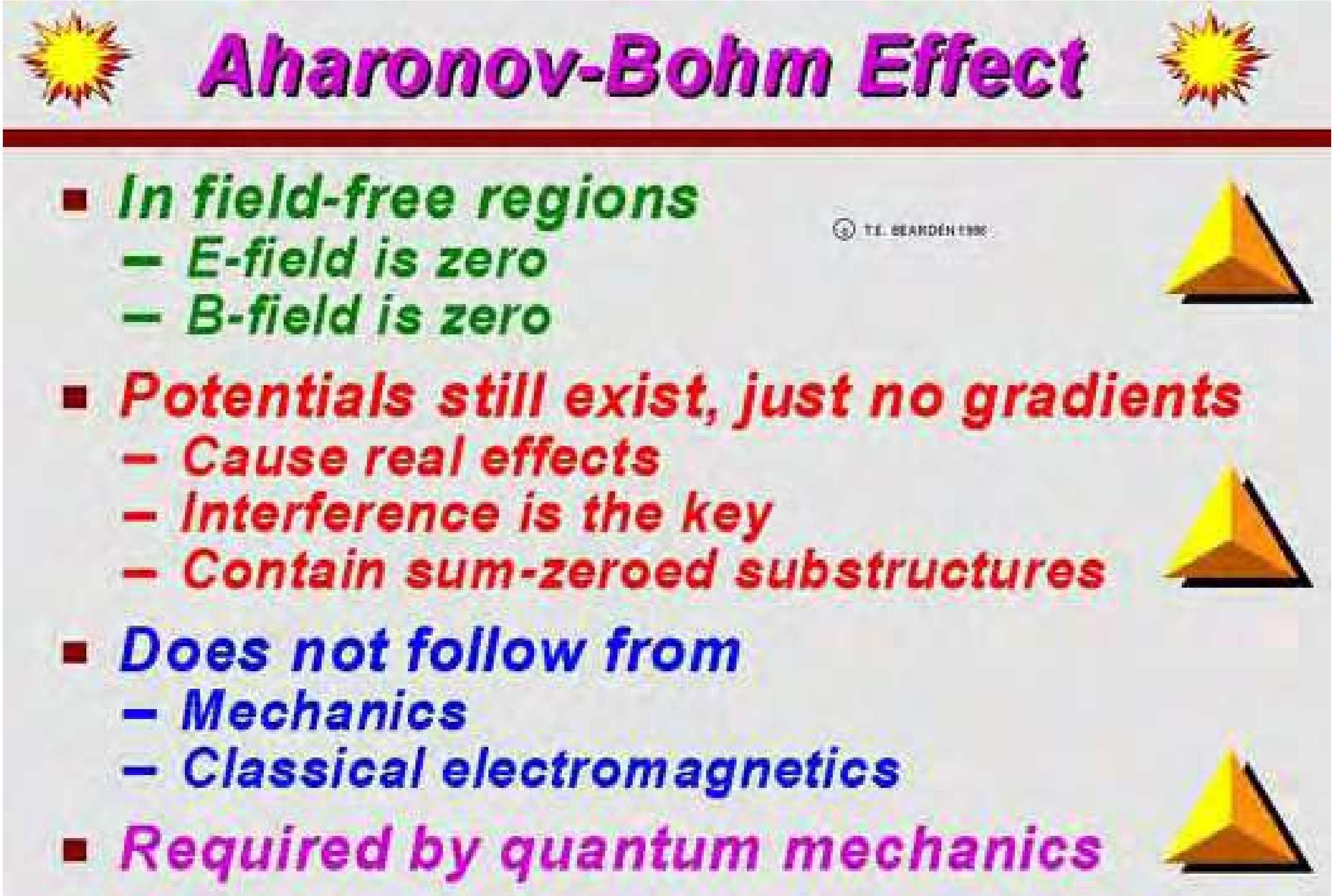

Aharonov-Bohm Effect

© T.E. BEARDEN 1996

- In field-free regions
  - E-field is zero
  - B-field is zero
- Potentials still exist, just no gradients
  - Cause real effects
  - Interference is the key
  - Contain sum-zeroed substructures
- Does not follow from
  - Mechanics
  - Classical electromagnetics
- Required by quantum mechanics

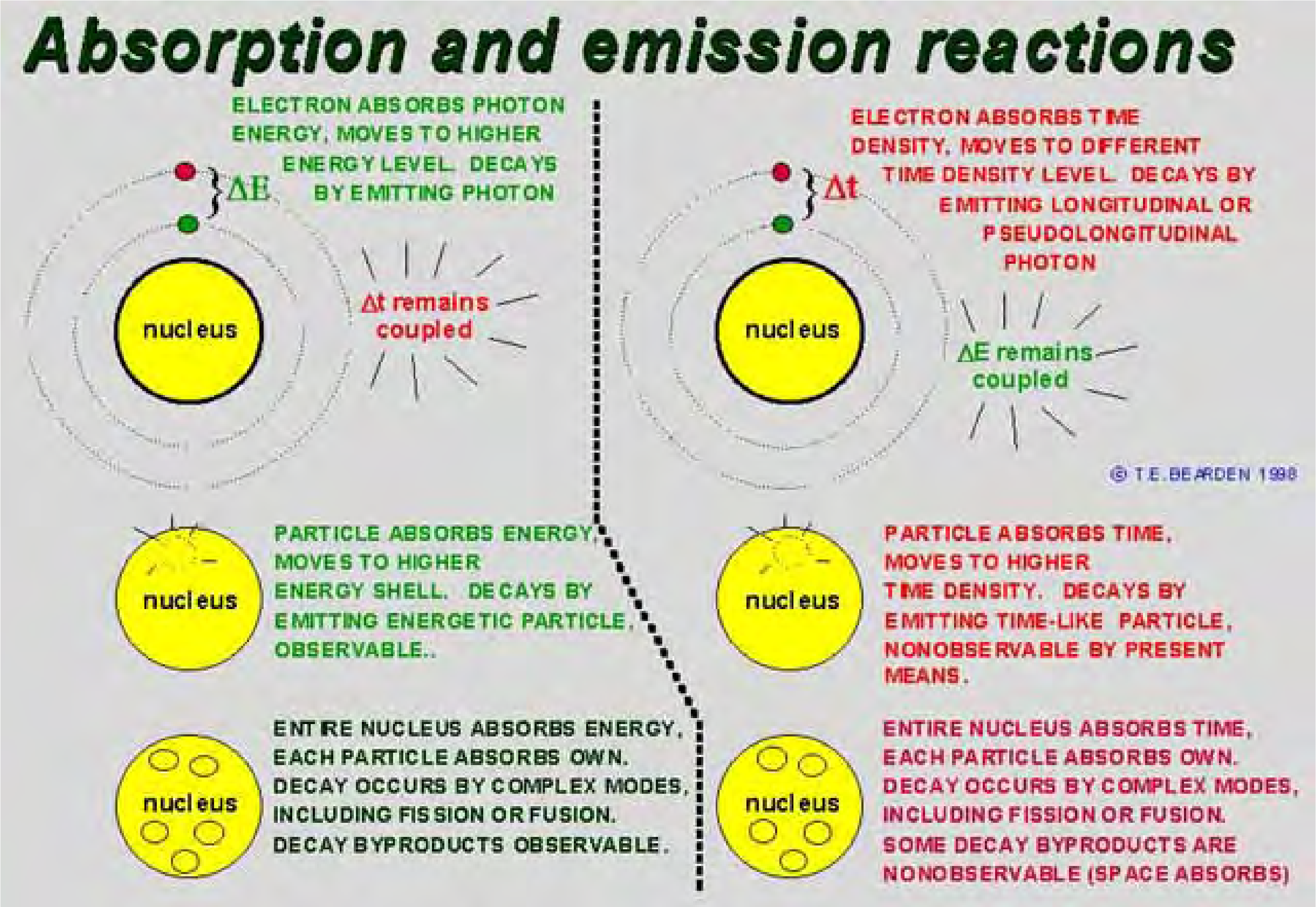

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE. OBSERVABLE..

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

© T.E. BEARDEN 1998

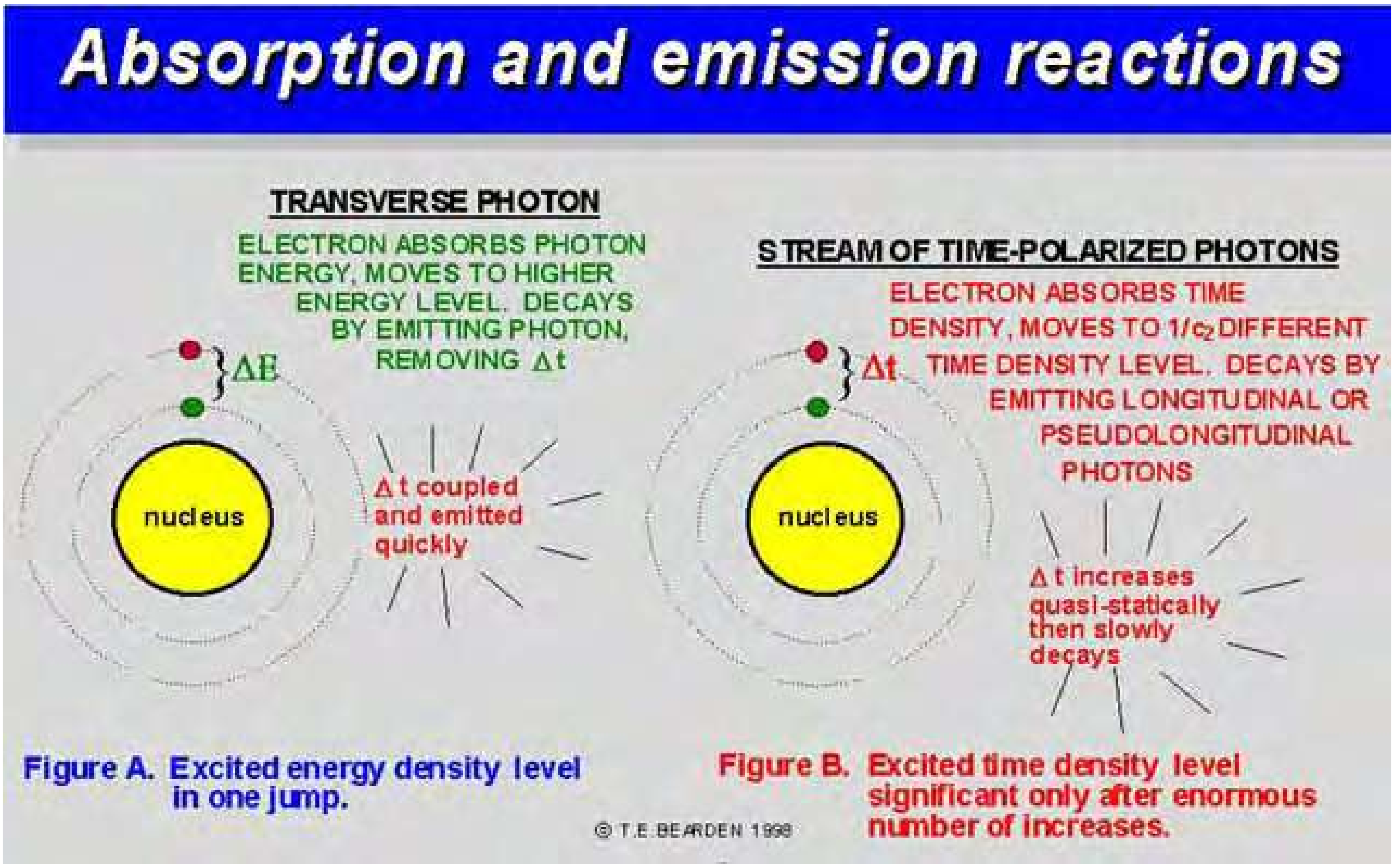

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

Δt coupled and emitted quickly

Figure A. Excited energy density level in one jump.

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c_2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt increases quasi-statically then slowly decays

Figure B. Excited time density level significant only after enormous number of increases.

© T.E.BEARDEN 1998

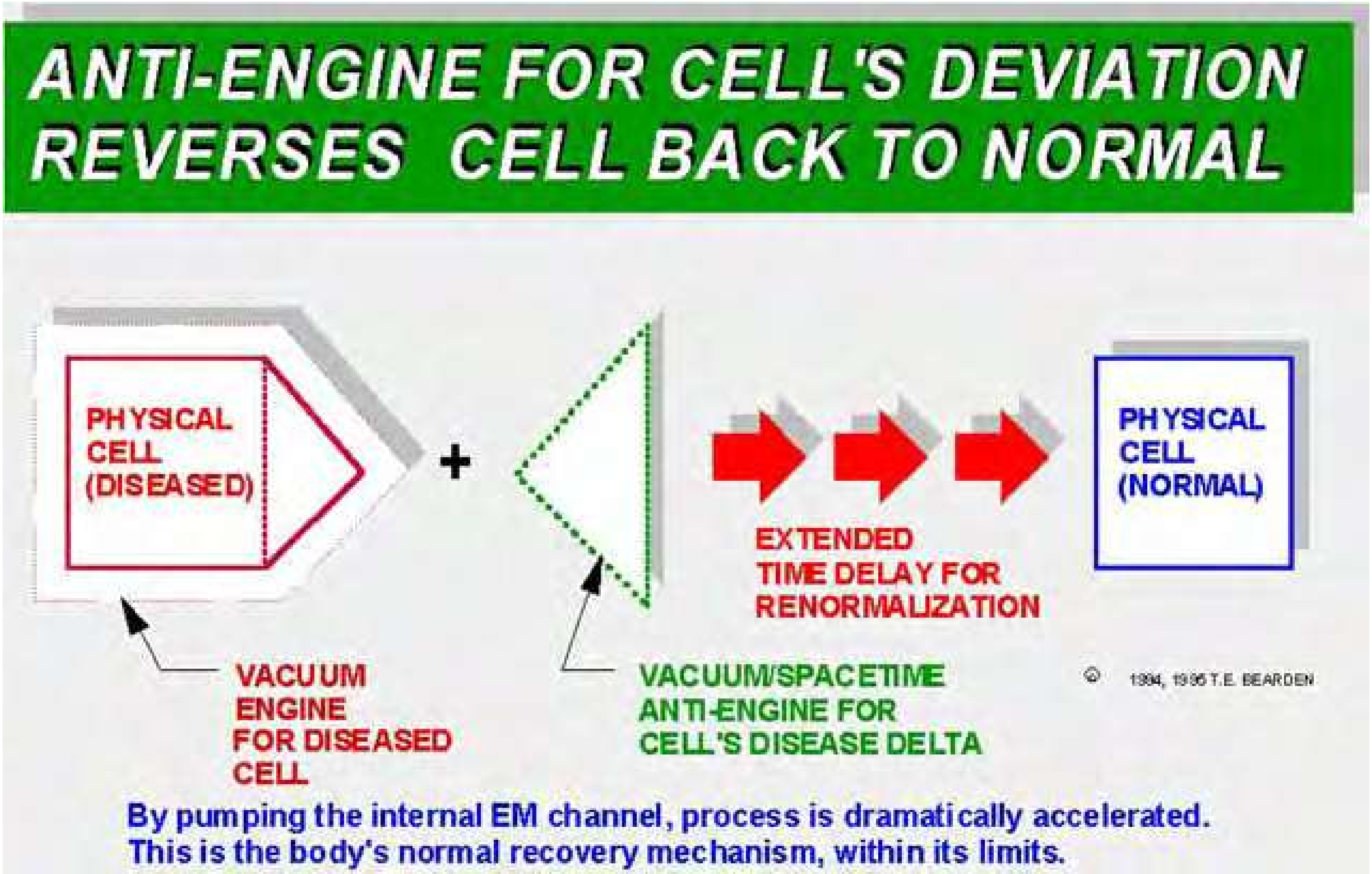

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL CELL (DISEASED)

+

PHYSICAL CELL (NORMAL)

EXTENDED TIME DELAY FOR RENORMALIZATION

VACUUM ENGINE FOR DISEASED CELL

VACUUM/SPACETIME ANTI-ENGINE FOR CELL'S DISEASE DELTA

© 1994, 1995 T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated.
This is the body's normal recovery mechanism, within its limits.

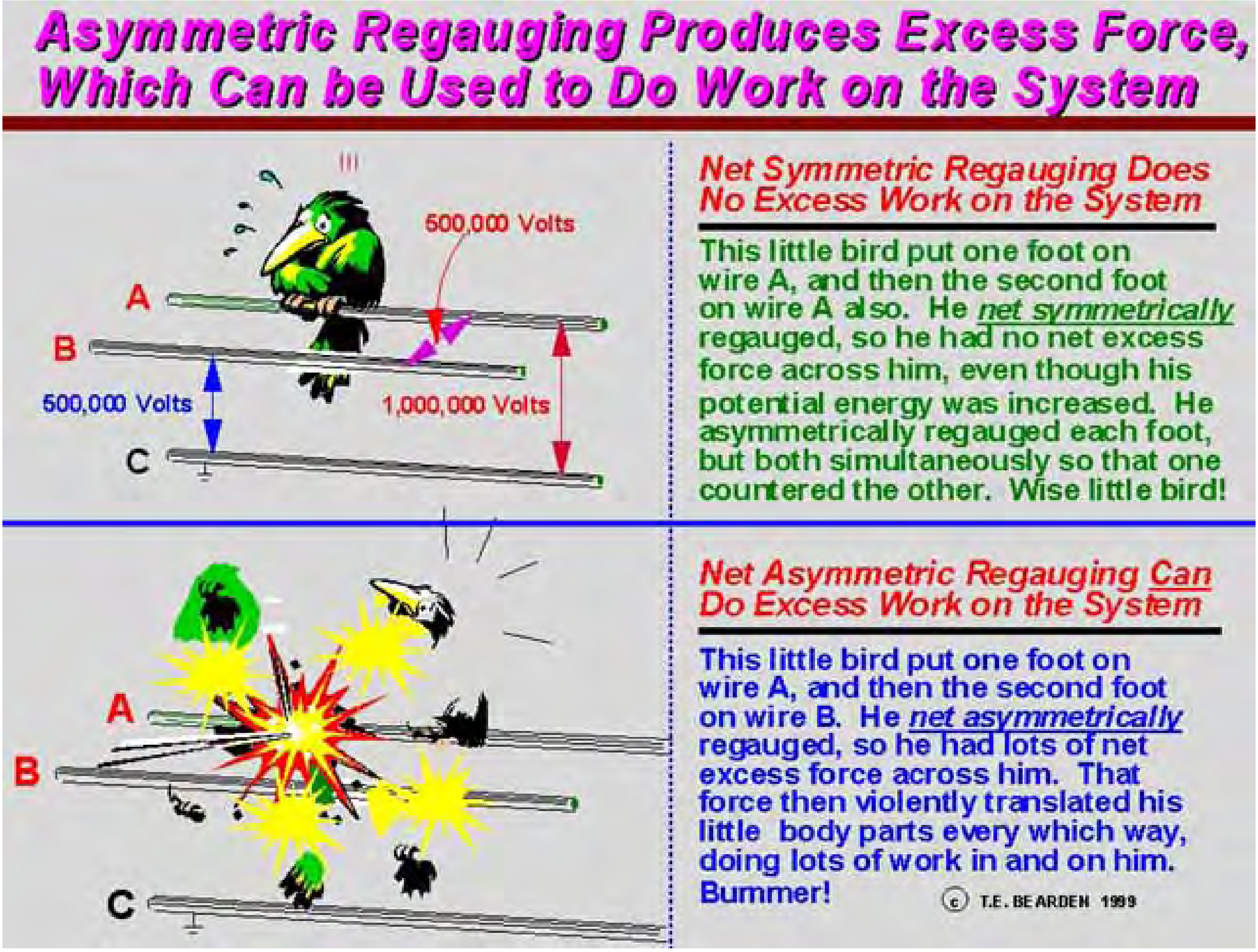

Asymmetric Regauging Produces Excess Force, Which Can be Used to Do Work on the System

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging *Can* Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1999

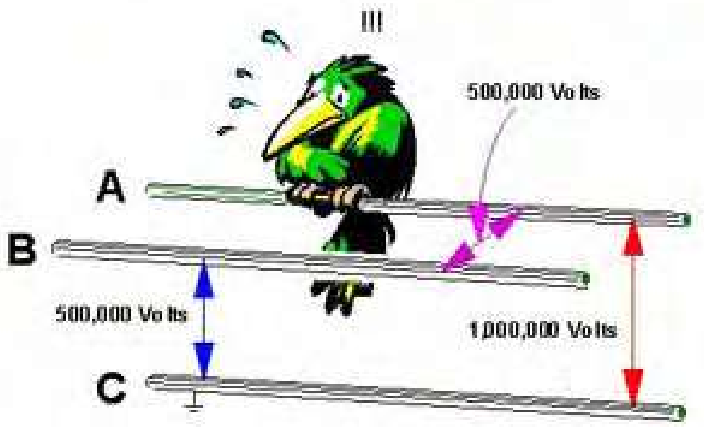

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

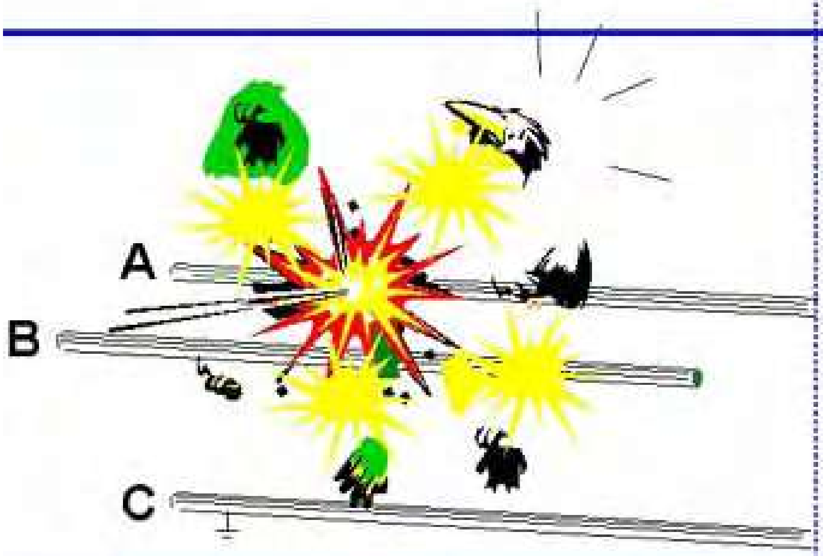

Net Asymmetric Regauging Can Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1996

Figure 32. Asymmetric regauging produces excess force, which can be used to do work on the regauging system.

Selected clipart © by Lotus SmartPics™

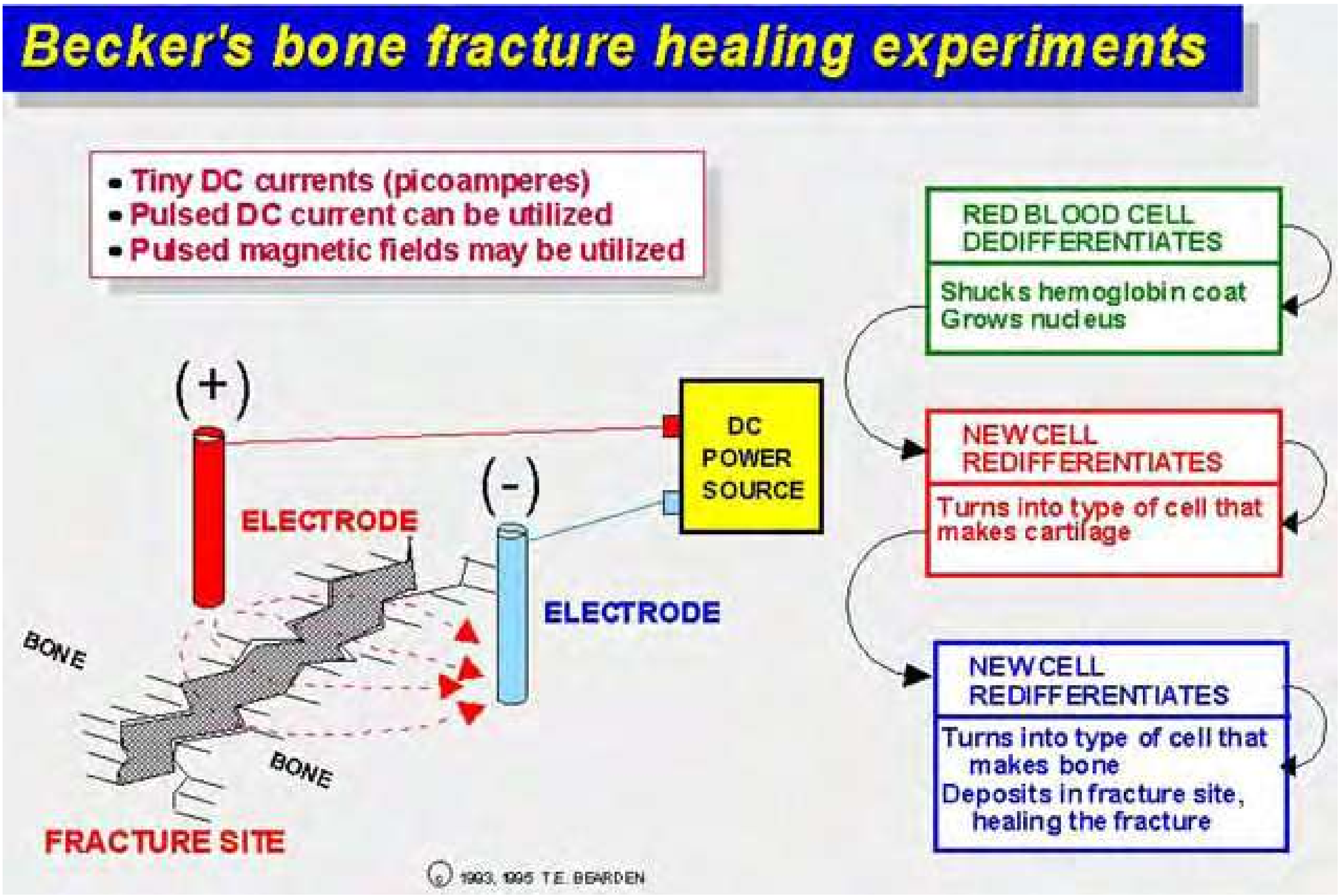

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1995 T.E. BEARDEN

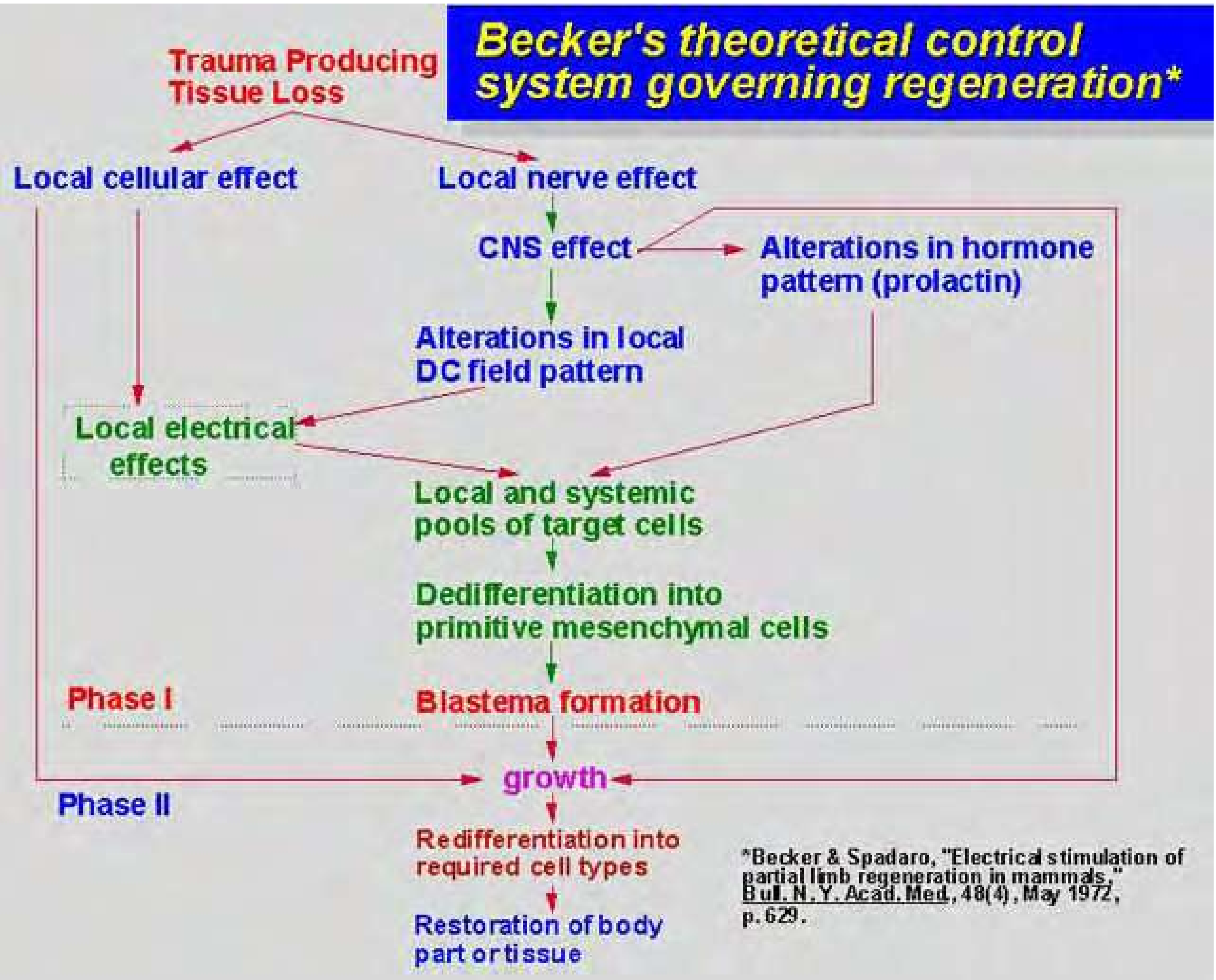

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N.Y. Acad. Med., 48(4), May 1972, p. 629.

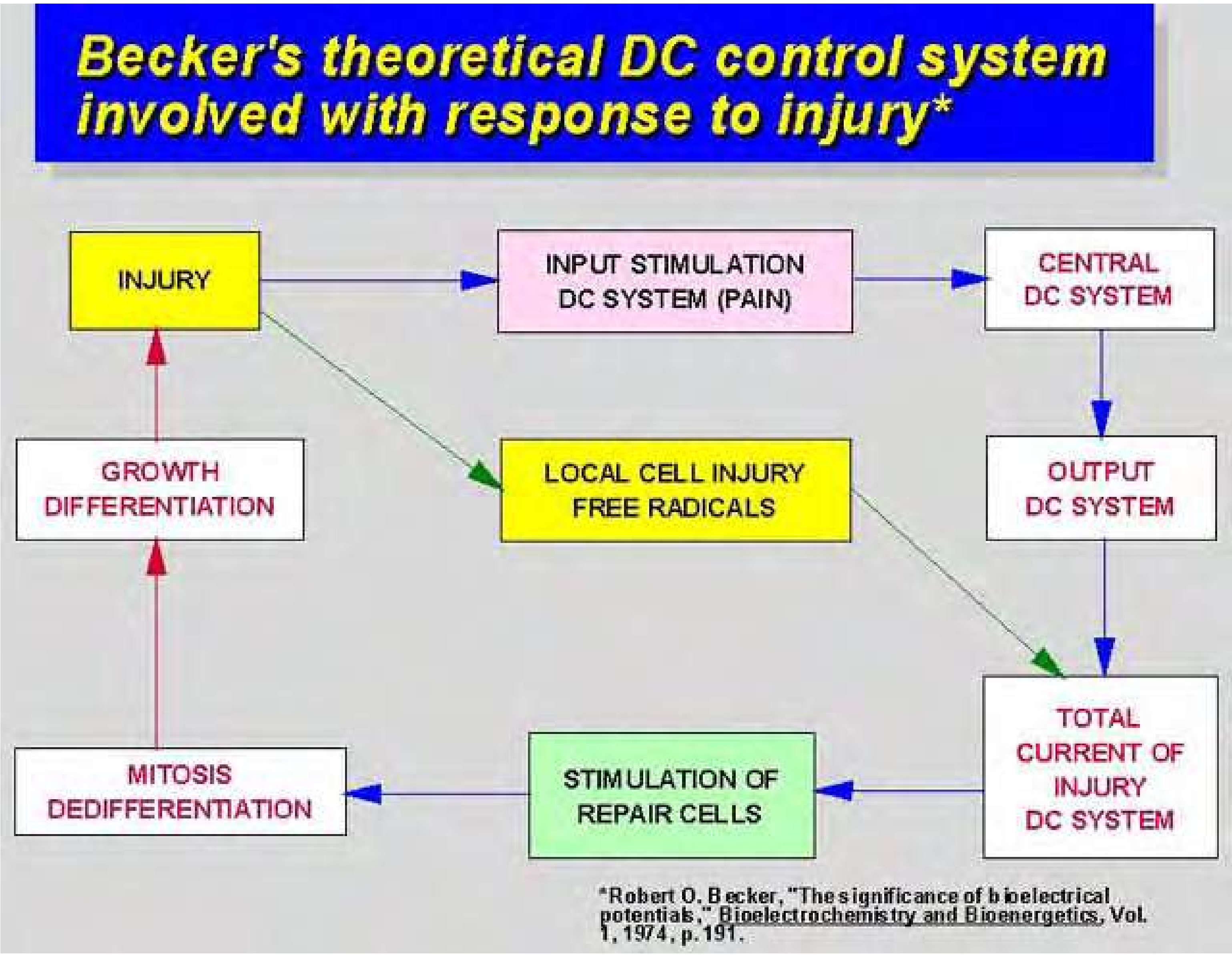

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION
DC SYSTEM (PAIN)

CENTRAL
DC SYSTEM

LOCAL CELL INJURY
FREE RADICALS

OUTPUT
DC SYSTEM

TOTAL
CURRENT OF
INJURY
DC SYSTEM

STIMULATION OF
REPAIR CELLS

MITOSIS
DEDIFFERENTIATION

GROWTH
DIFFERENTIATION

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

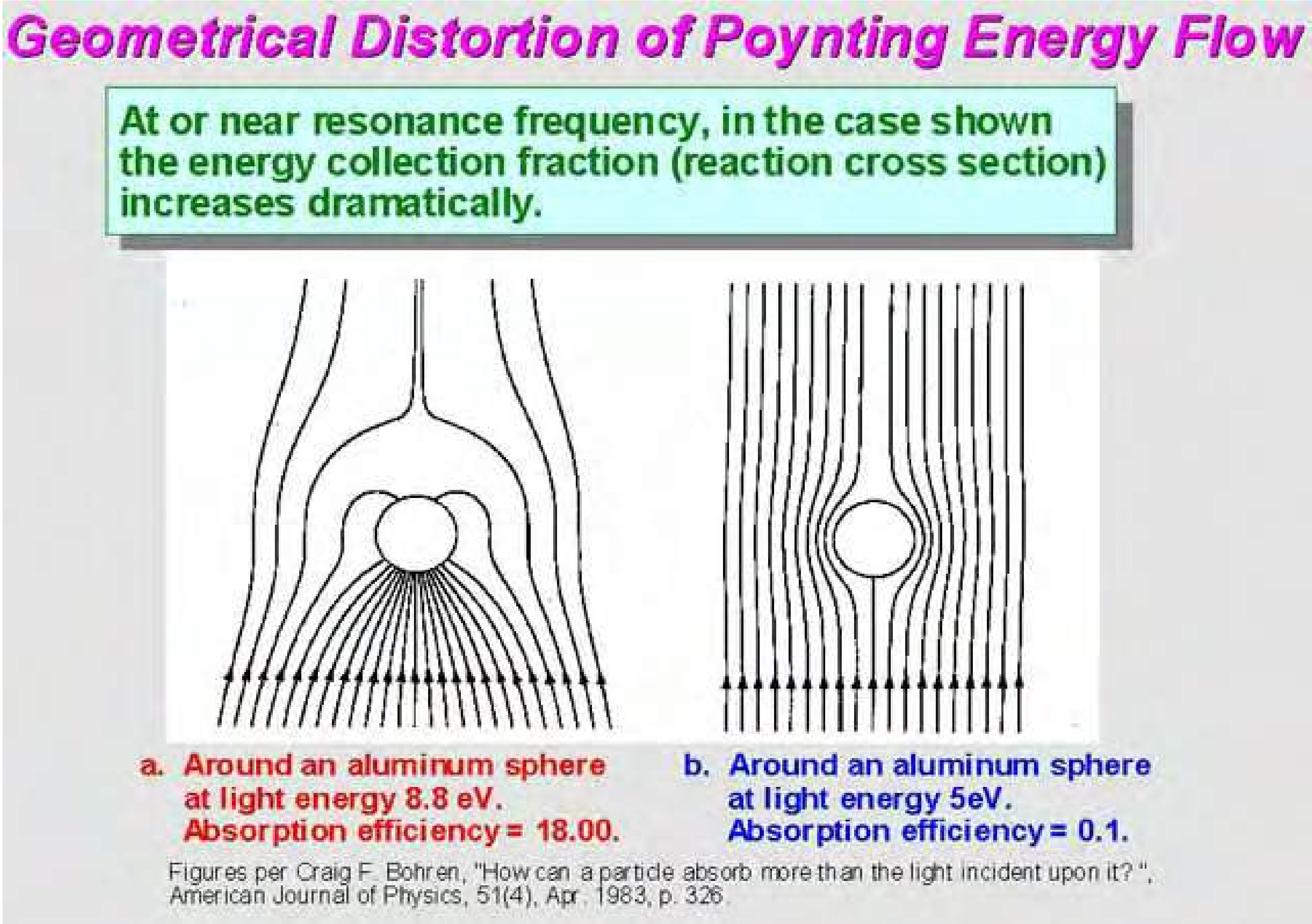

Geometrical Distortion of Poynting Energy Flow

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

a. Around an aluminum sphere at light energy 8.8 eV. Absorption efficiency = 18.00.

b. Around an aluminum sphere at light energy 5eV. Absorption efficiency = 0.1.

Figures per Craig F. Bohren, "How can a particle absorb more than the light incident upon it? ", American Journal of Physics, 51(4), Apr. 1983, p. 326.

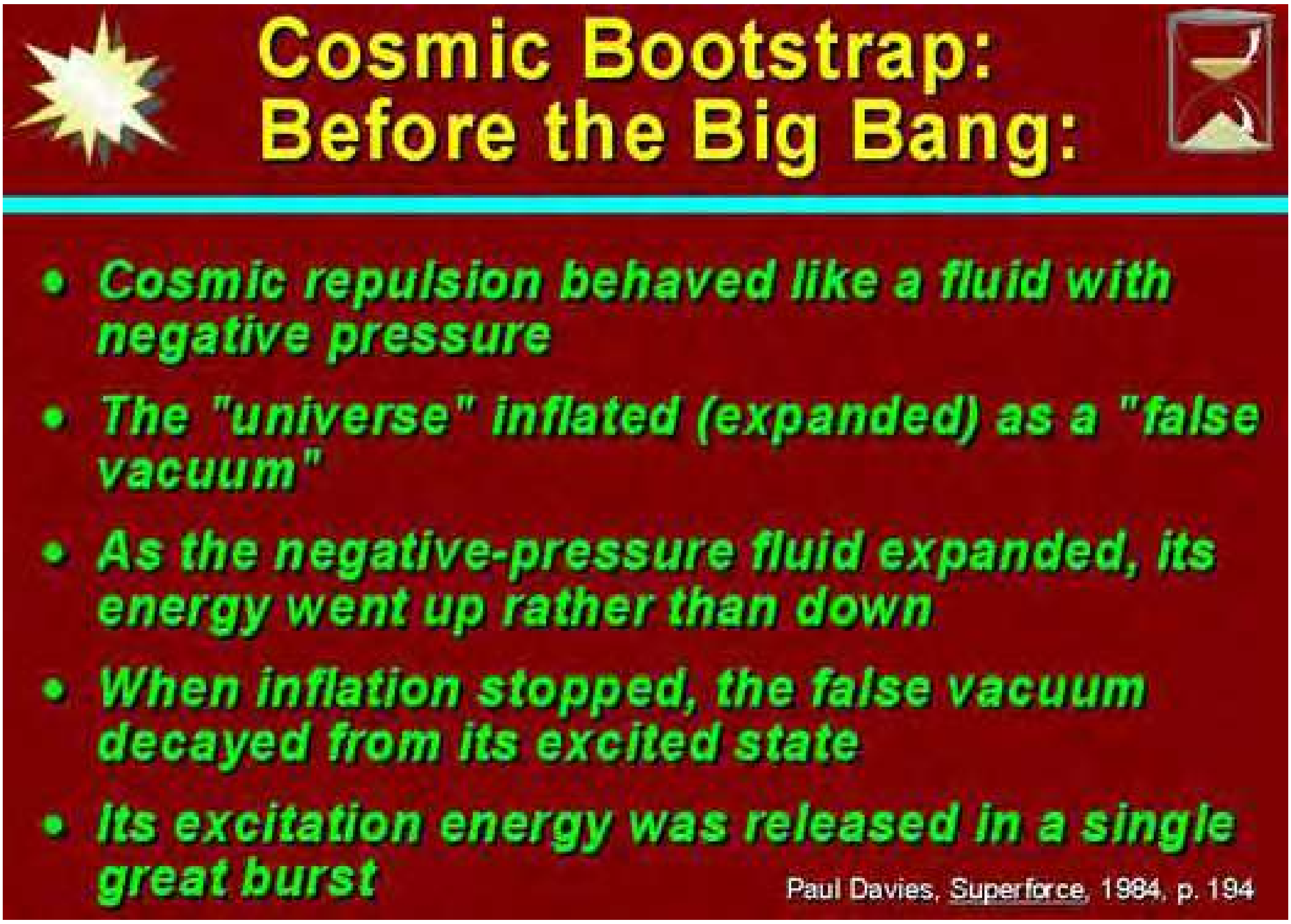

Cosmic Bootstrap: Before the Big Bang:

- *Cosmic repulsion behaved like a fluid with negative pressure*
- *The "universe" inflated (expanded) as a "false vacuum"*
- *As the negative-pressure fluid expanded, its energy went up rather than down*
- *When inflation stopped, the false vacuum decayed from its excited state*
- *Its excitation energy was released in a single great burst*

Paul Davies, Superforce, 1984, p. 194

BRIDGING ENABLES COP > 1.0

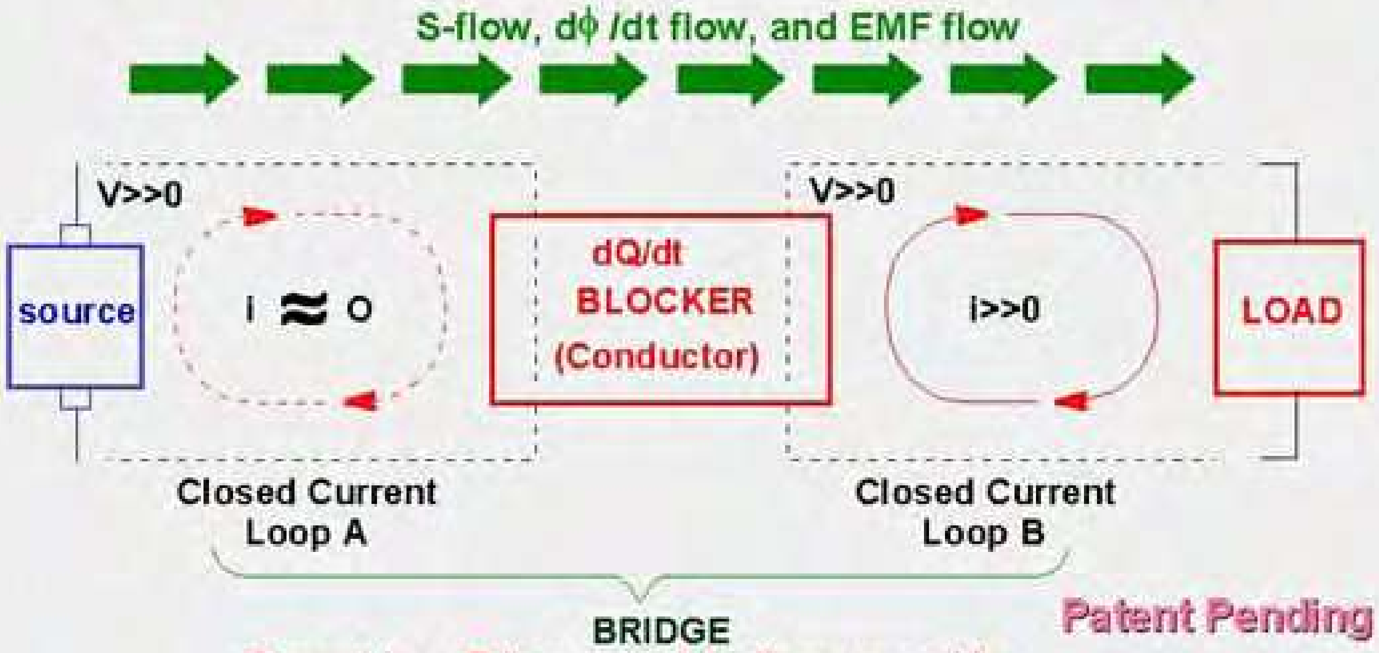
S-flow, dϕ /dt flow, and EMF flow
V>>0
source
i ≈ 0
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD
Closed Current
Loop A
Closed Current
Loop B
BRIDGE
Patent Pending

Current loop B is asymmetrically regauged by S, dϕ /dt, and EMF flow from current loop A.

© 1998 T.E. BEARDEN

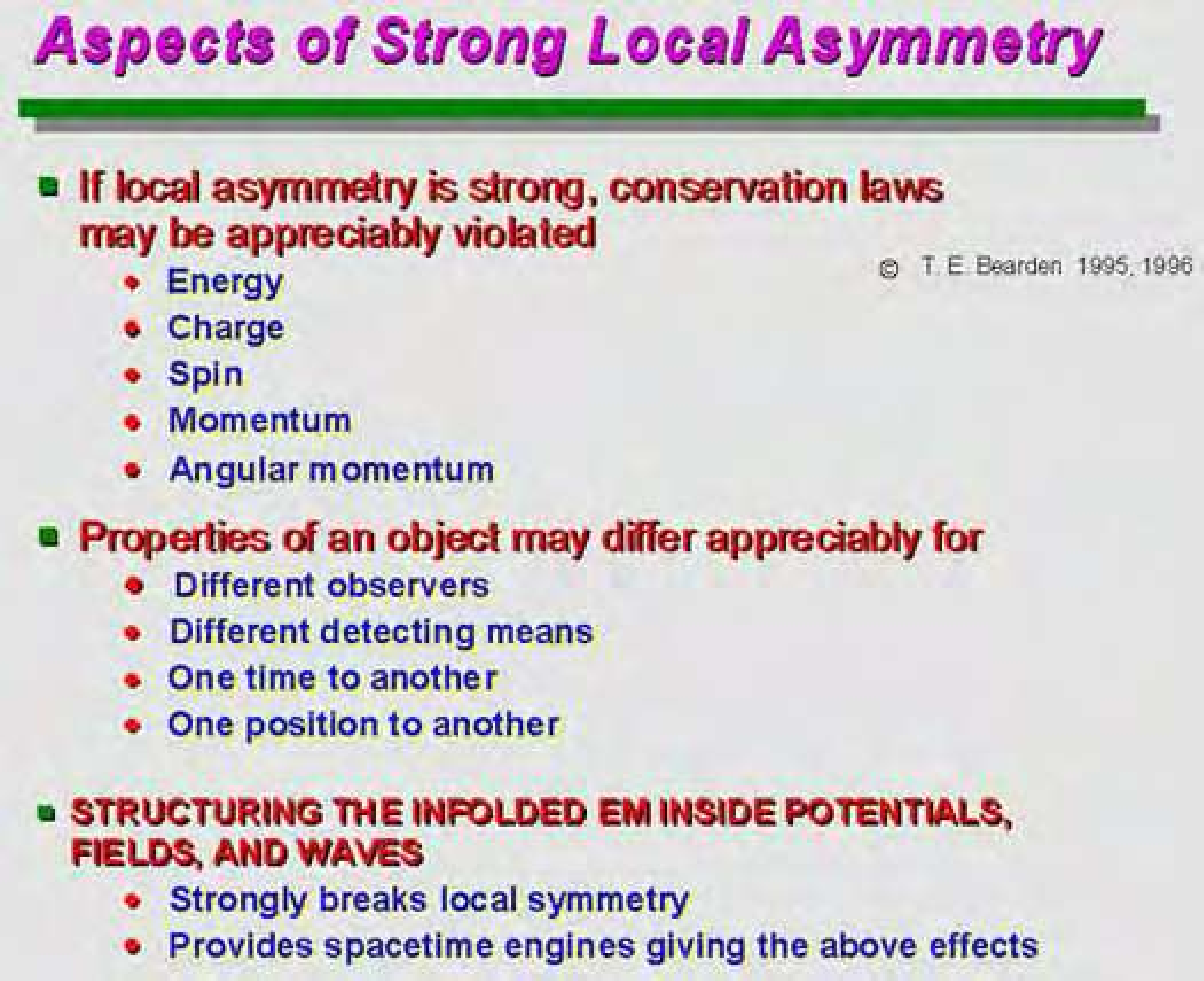

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

© T. E. Bearden 1995, 1996

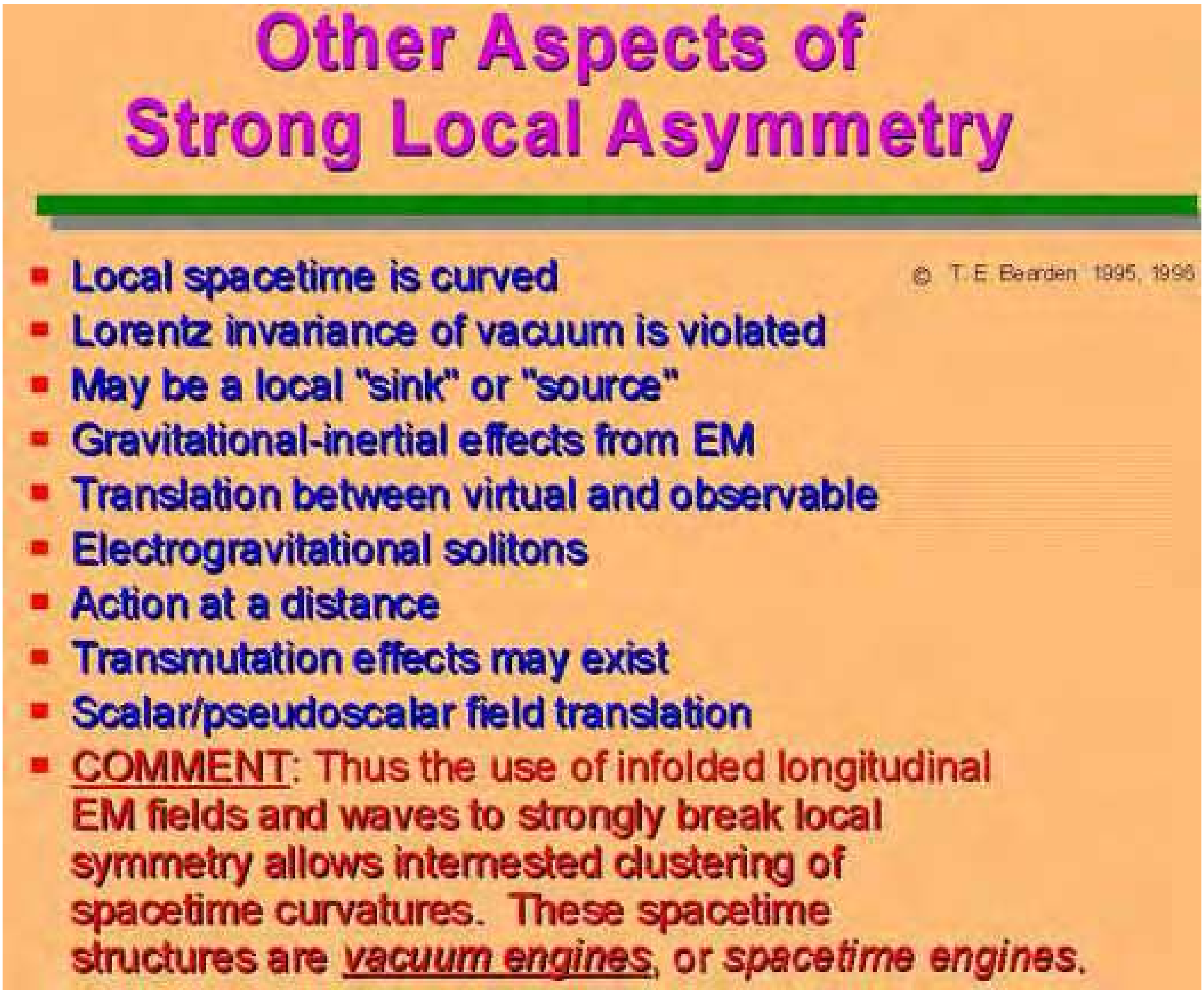

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational-inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

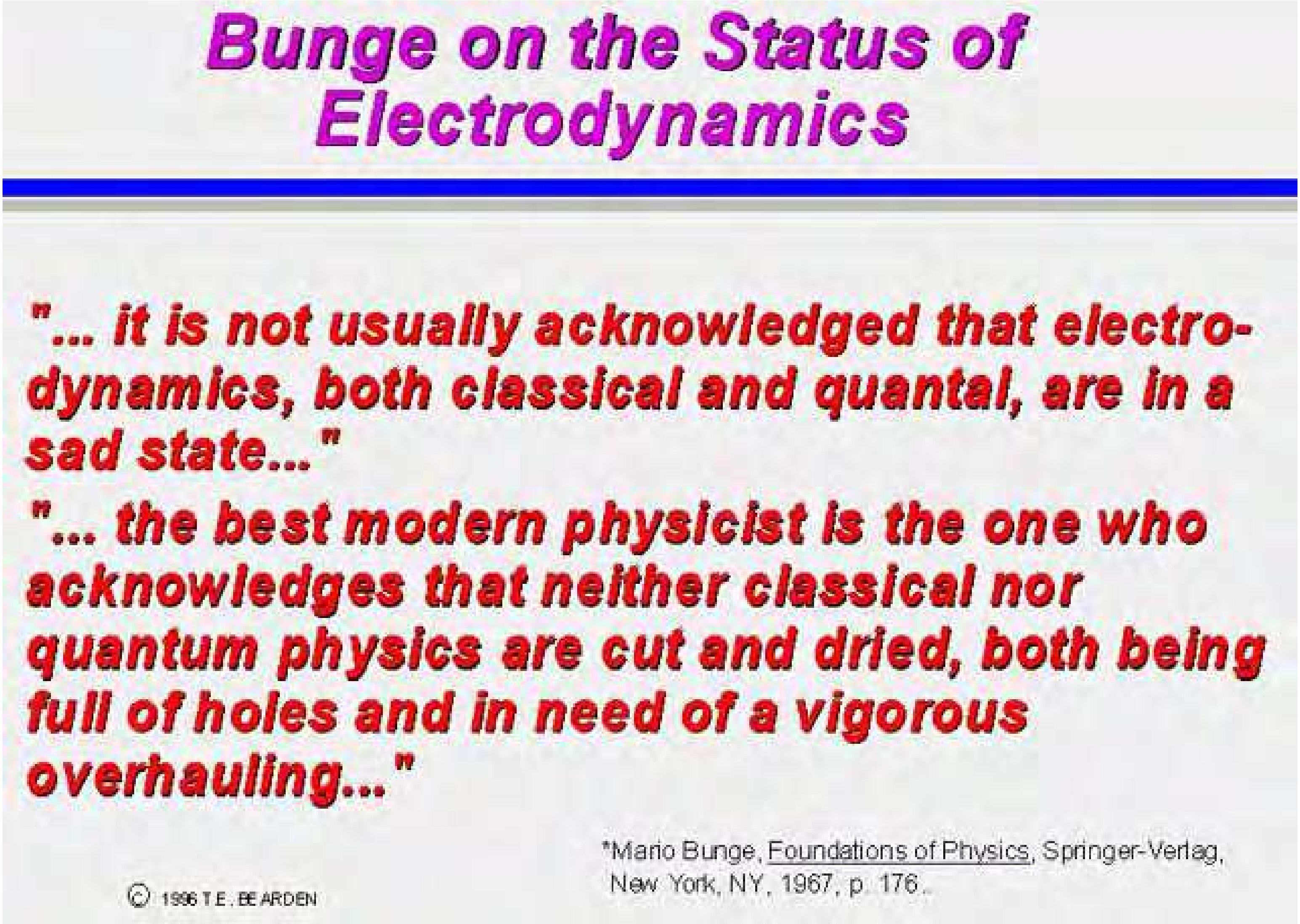

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176.

© 1996 T.E. BEARDEN

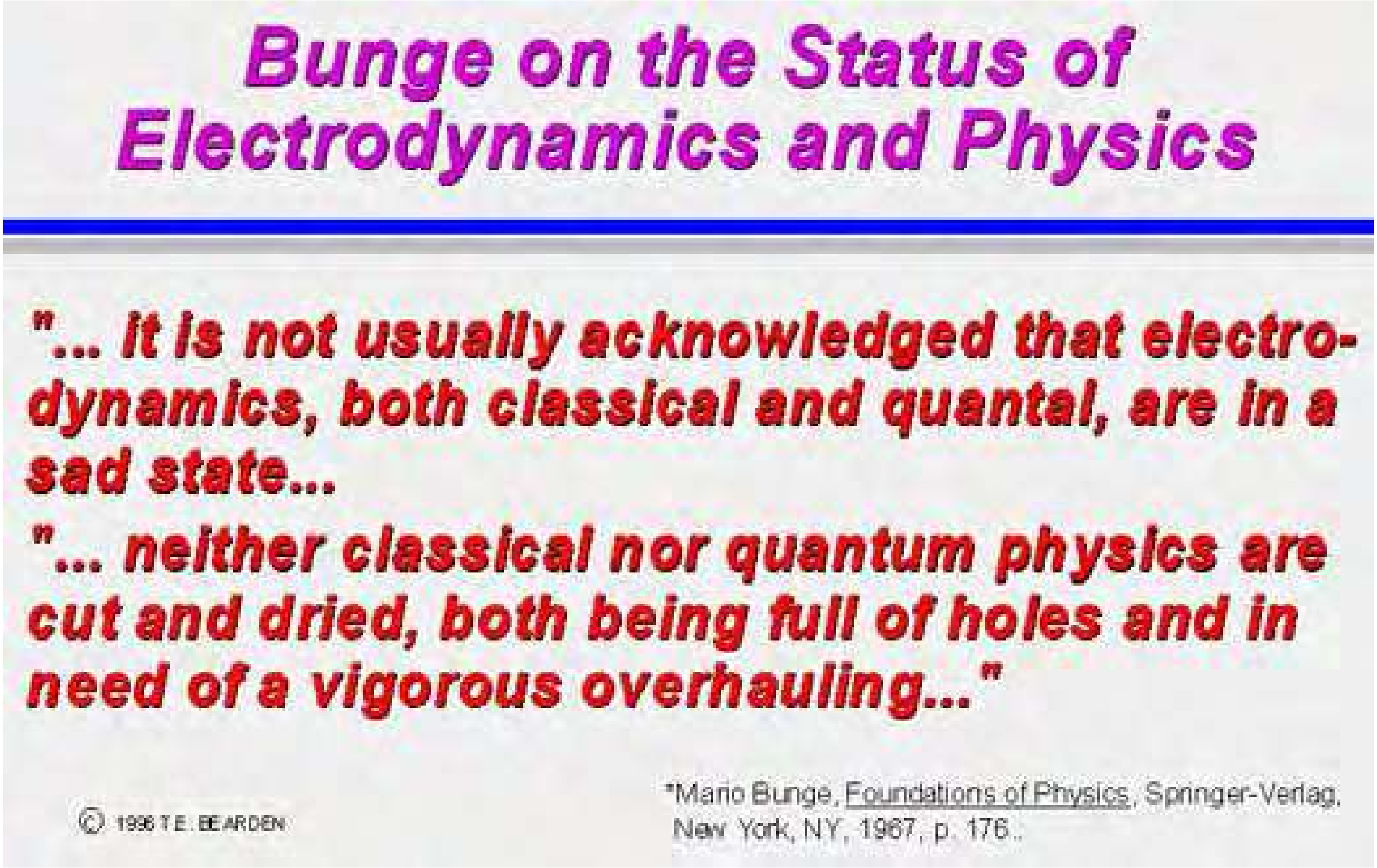

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

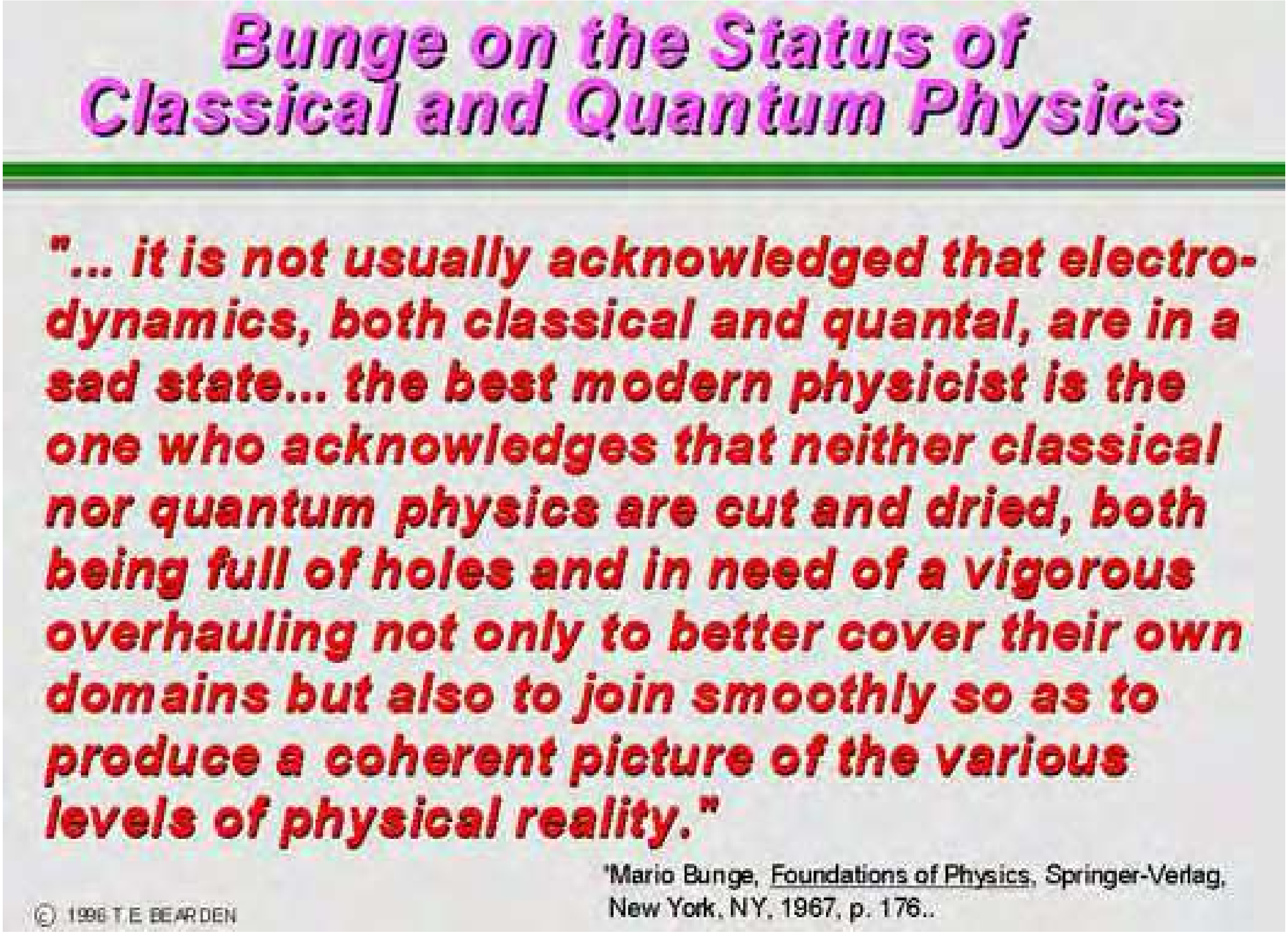

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

'Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

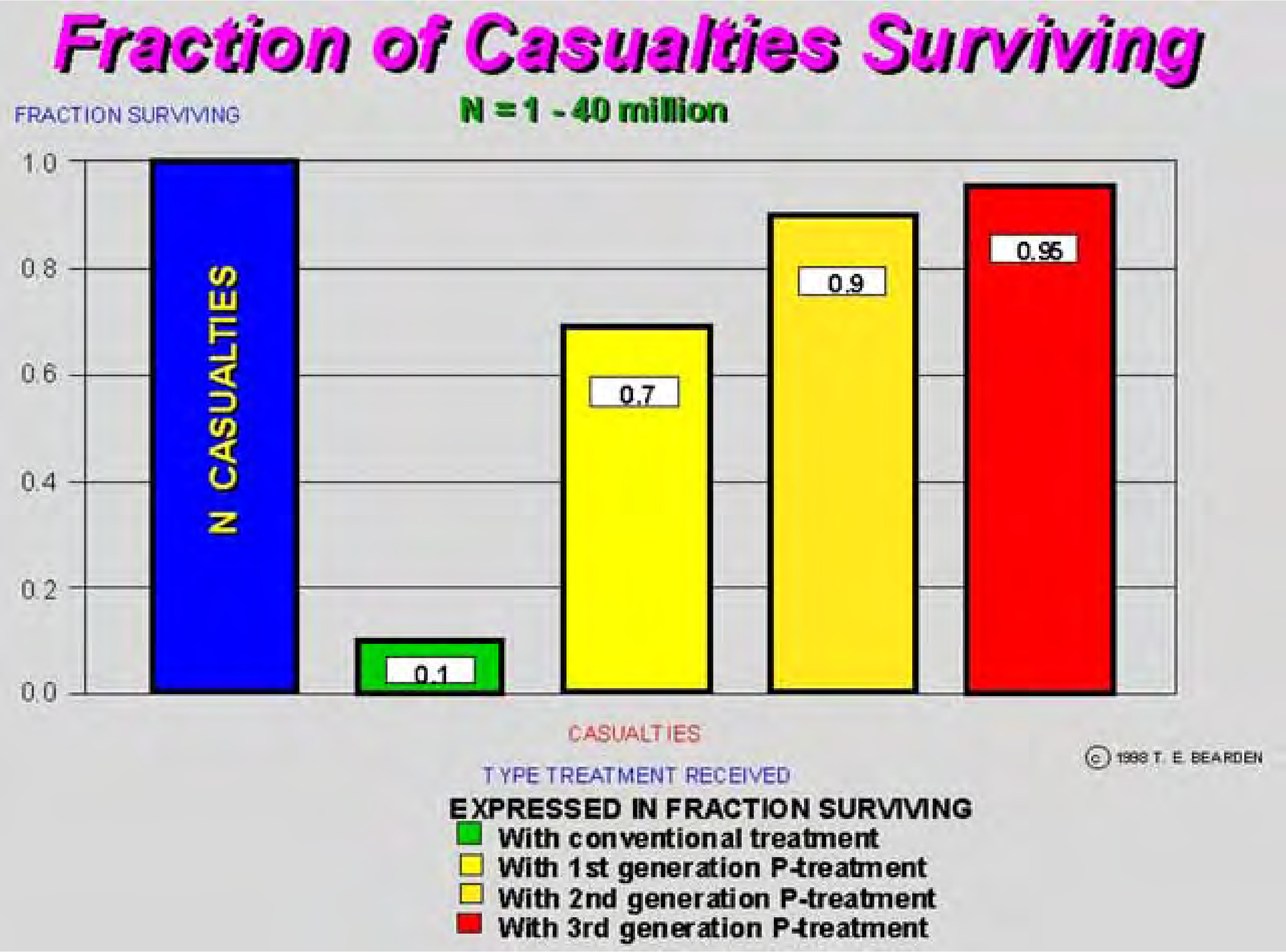
Fraction of Casualties Surviving
FRACTION SURVIVING
N = 1 - 40 million
1.0
0.8
0.6
0.4
0.2
0.0
N CASUALTIES
0.1
0.7
0.9
0.95
CASUALTIES
TYPE TREATMENT RECEIVED
© 1998 T. E. BEARDEN
EXPRESSED IN FRACTION SURVIVING
🟩 With conventional treatment
🟨 With 1st generation P-treatment
🟨 With 2nd generation P-treatment
🟥 With 3rd generation P-treatment

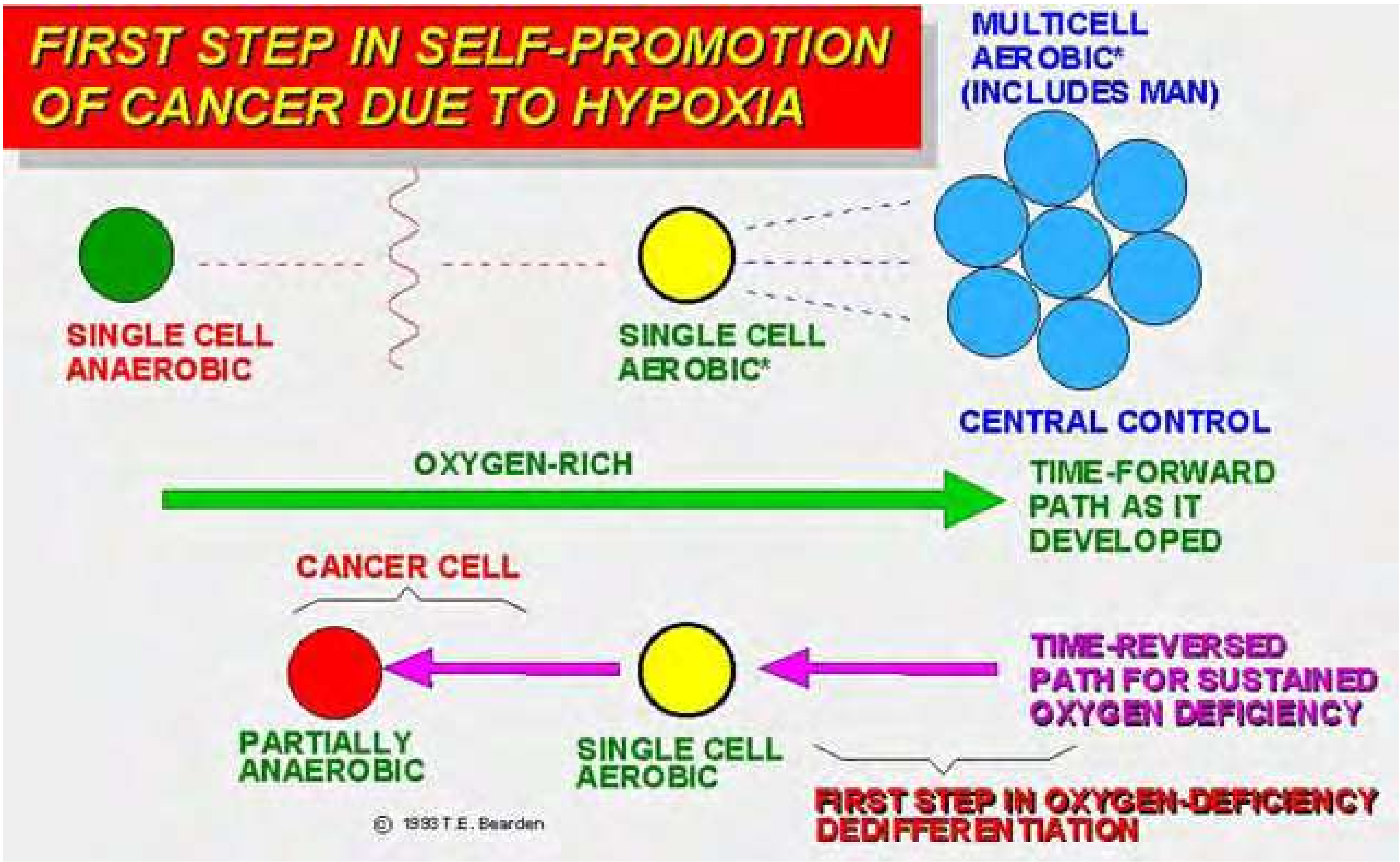

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

MULTICELL
AEROBIC*
(INCLUDES MAN)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

© 1993 T.E. Bearden

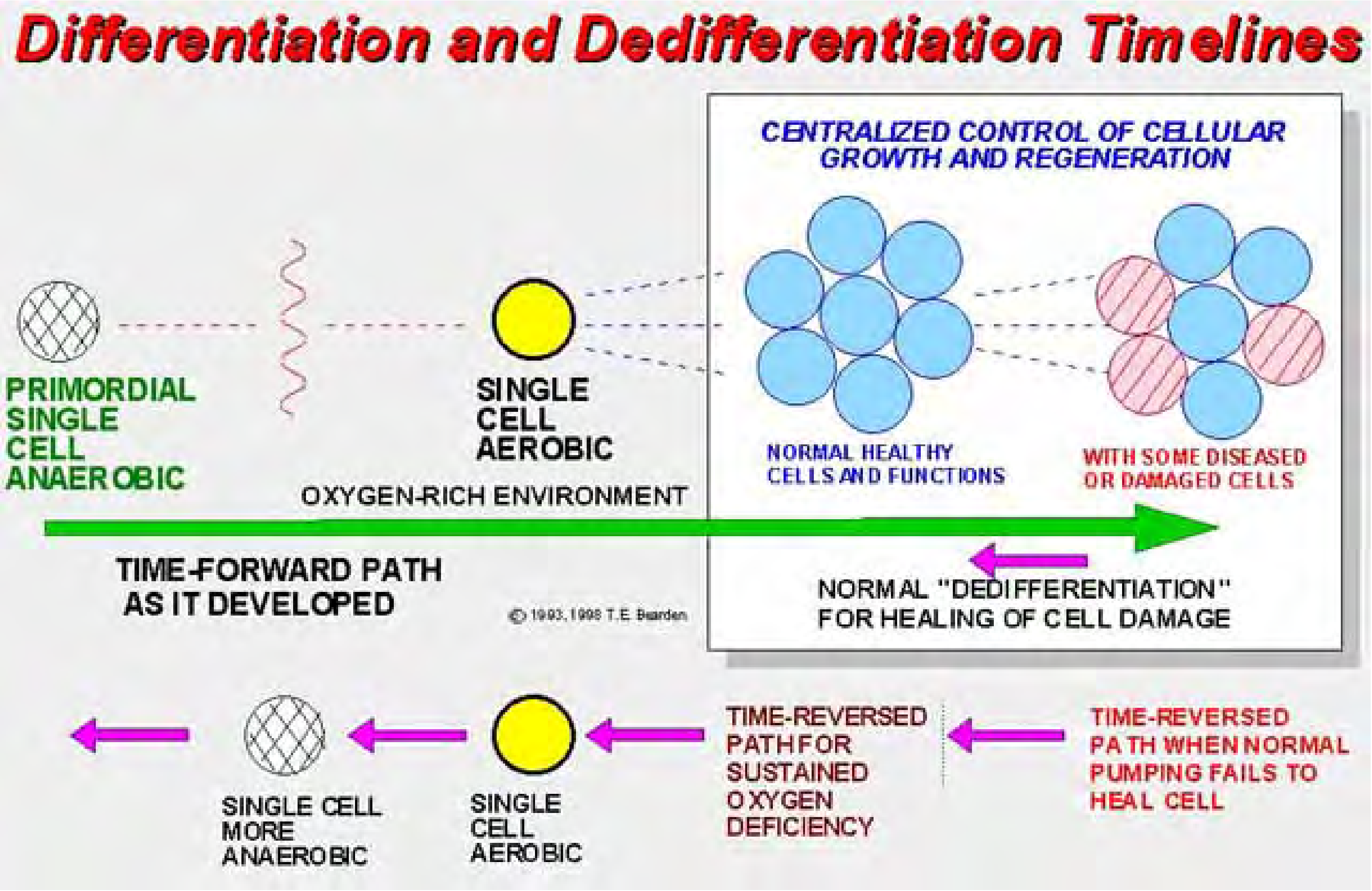

Differentiation and Dedifferentiation Timelines

PRIMORDIAL
SINGLE
CELL
ANAEROBIC

SINGLE
CELL
AEROBIC

OXYGEN-RICH ENVIRONMENT

CENTRALIZED CONTROL OF CELLULAR
GROWTH AND REGENERATION

NORMAL HEALTHY
CELLS AND FUNCTIONS

WITH SOME DISEASED
OR DAMAGED CELLS

TIME-FORWARD PATH
AS IT DEVELOPED

© 1993, 1998 T.E. Bearden

NORMAL "DEDIFFERENTIATION"
FOR HEALING OF CELL DAMAGE

SINGLE CELL
MORE
ANAEROBIC

SINGLE
CELL
AEROBIC

TIME-REVERSED
PATH FOR
SUSTAINED
OXYGEN
DEFICIENCY

TIME-REVERSED
PATH WHEN NORMAL
PUMPING FAILS TO
HEAL CELL

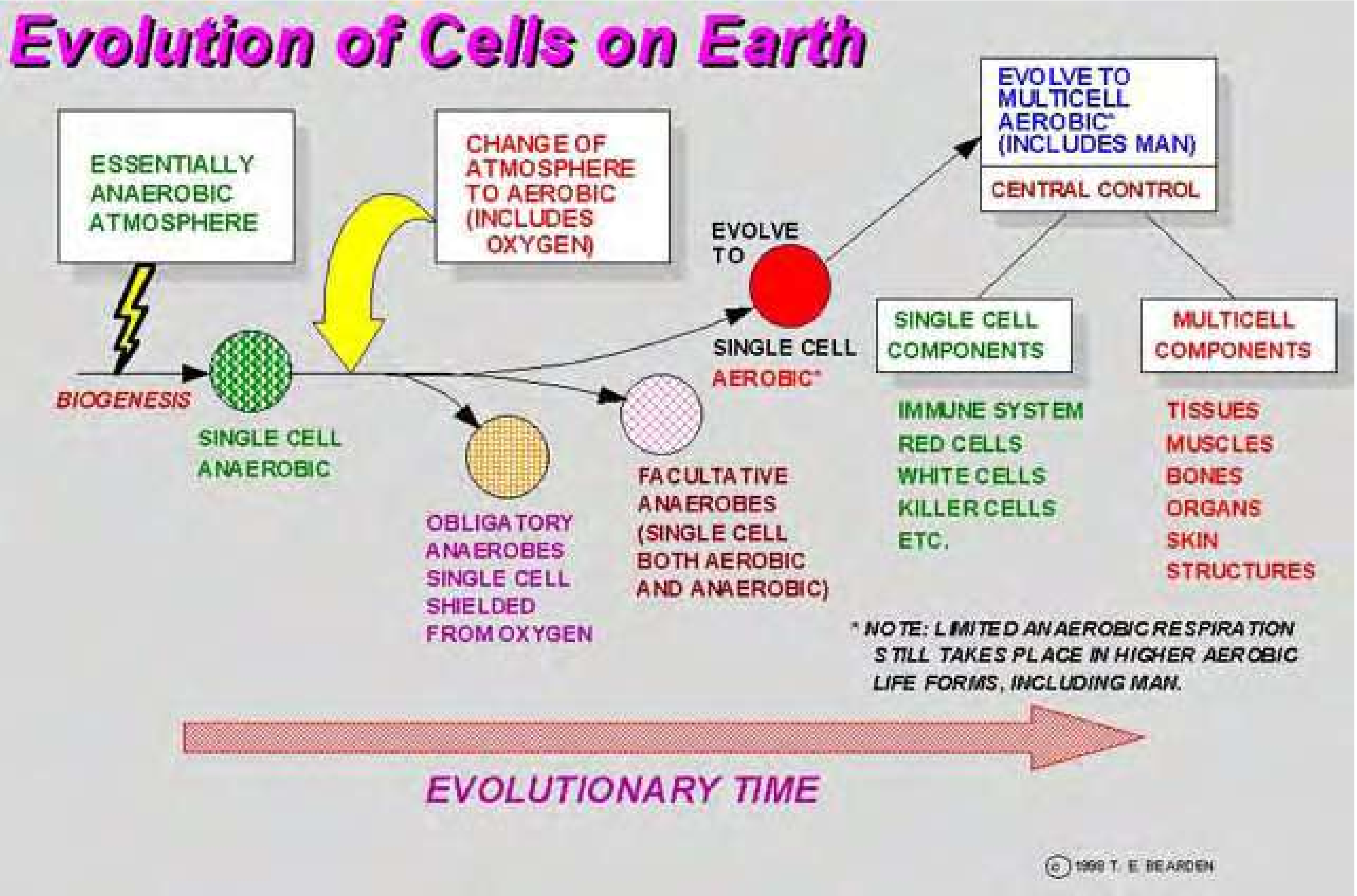

Evolution of Cells on Earth

ESSENTIALLY
ANAEROBIC
ATMOSPHERE

CHANGE OF
ATMOSPHERE
TO AEROBIC
(INCLUDES
OXYGEN)

BIOGENESIS

SINGLE CELL
ANAEROBIC

OBLIGATORY
ANAEROBES
SINGLE CELL
SHIELDED
FROM OXYGEN

FACULTATIVE
ANAEROBES
(SINGLE CELL
BOTH AEROBIC
AND ANAEROBIC)

EVOLVE
TO

SINGLE CELL
AEROBIC*

EVOLVE TO
MULTICELL
AEROBIC*
(INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL
COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL
COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

*NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

(c) 1999 T. E. BEARDEN

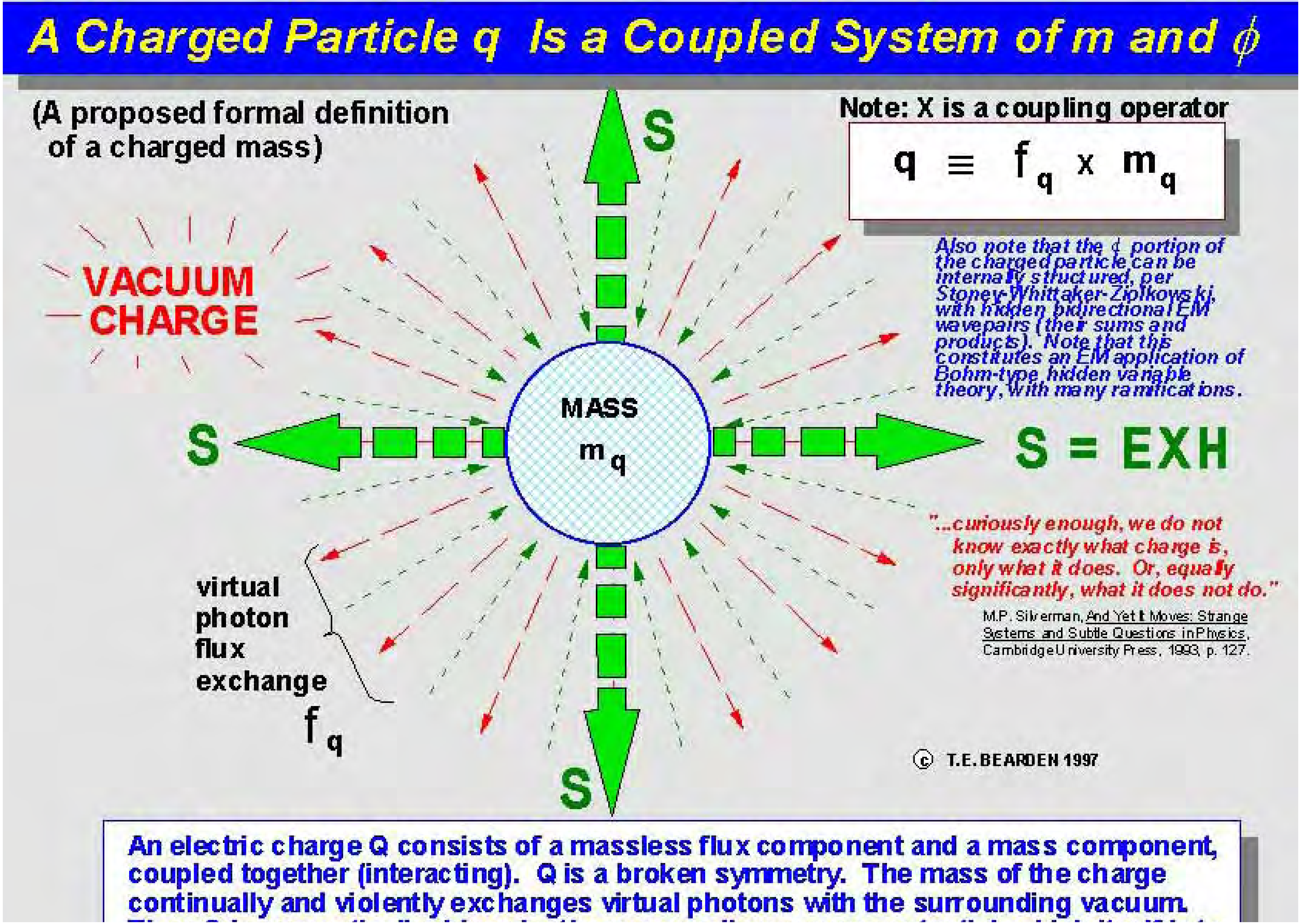

A Charged Particle q Is a Coupled System of m and ϕ

(A proposed formal definition of a charged mass)

Note: X is a coupling operator

$$q \equiv f_q \times m_q$$

Also note that the ϕ portion of the charged particle can be internally structured, per Stoney-Whittaker-Ziolkowski, with hidden bidirectional EM wavepairs (their sums and products). Note that this constitutes an EM application of Bohm-type hidden variable theory, with many ramifications.

$$S = E X H$$

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

© T.E. BEARDEN 1997

An electric charge Q consists of a massless flux component and a mass component, coupled together (interacting). Q is a broken symmetry. The mass of the charge continually and violently exchanges virtual photons with the surrounding vacuum.

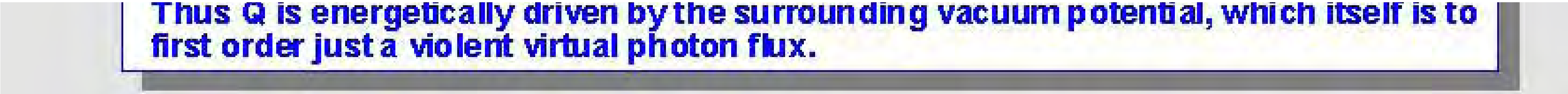

Thus Q is energetically driven by the surrounding vacuum potential, which itself is to first order just a violent virtual photon flux.

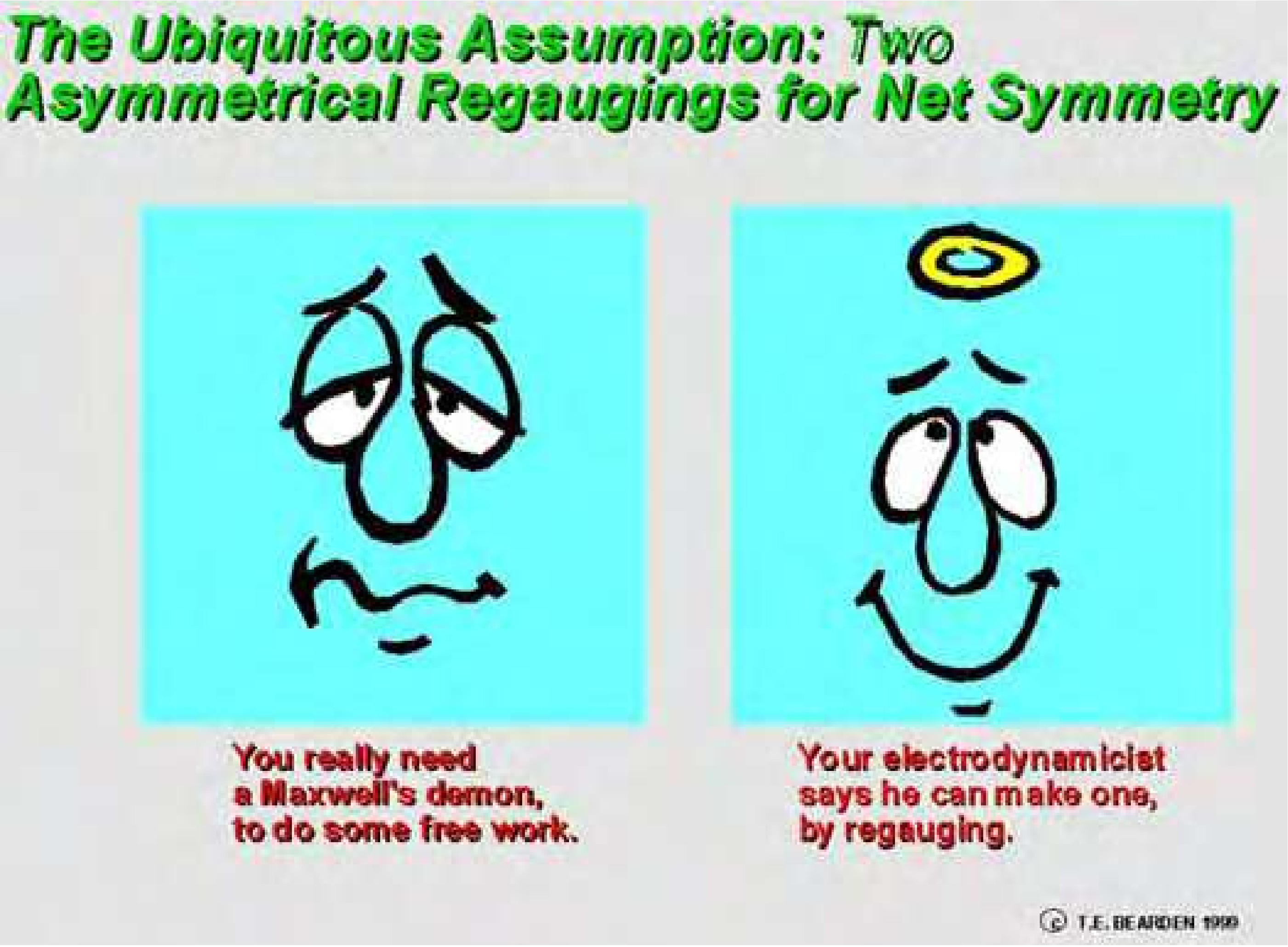

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

You really need
a Maxwell's demon,
to do some free work.

Your electrodynamicist
says he can make one,
by regauging.

© T.E. BEARDEN 1999

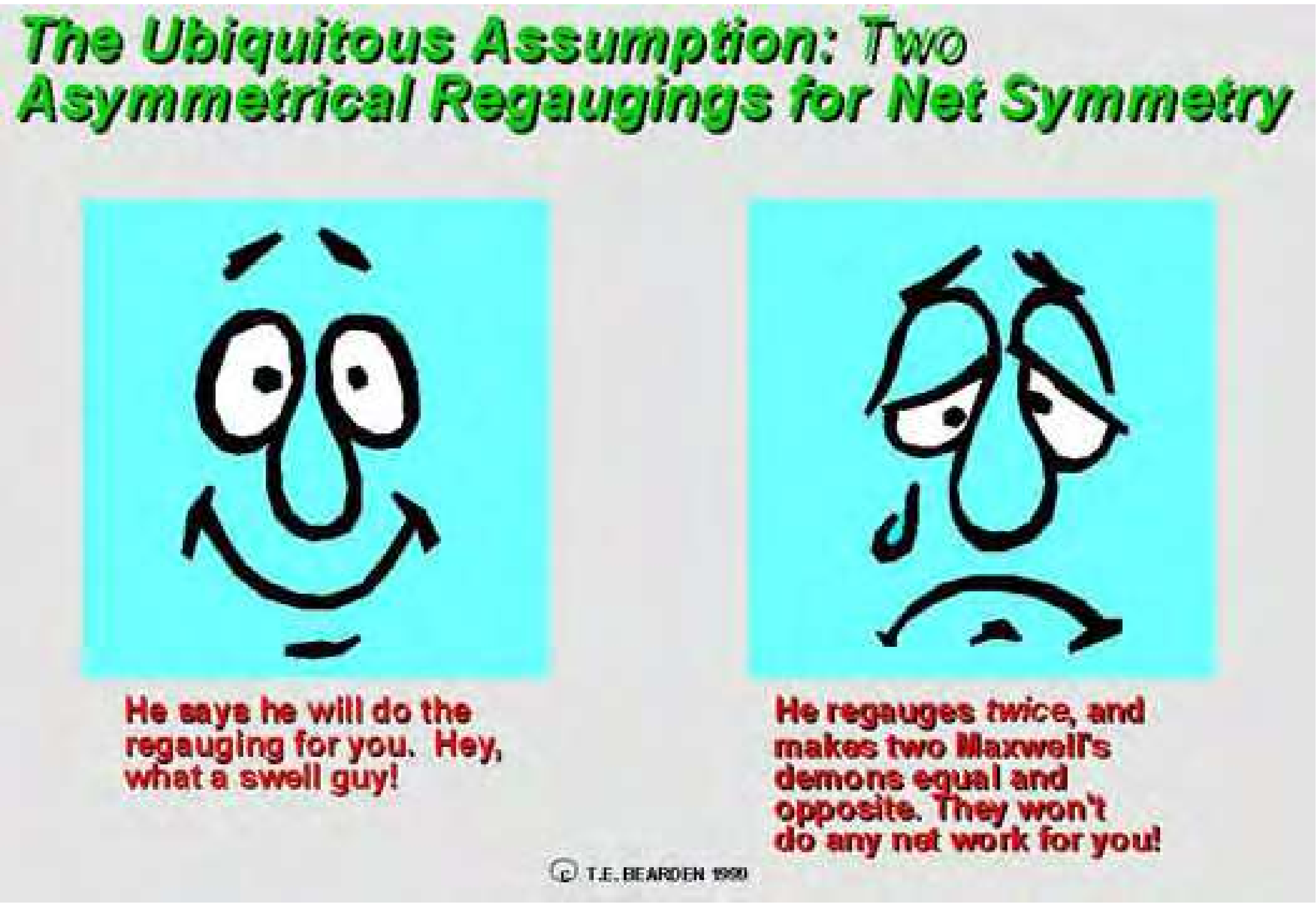

The Ubiquitous Assumption: *Two* Asymmetrical Regaugings for Net Symmetry

He says he will do the regauging for you. Hey, what a swell guy!

He regauges *twice*, and makes two Maxwell's demons equal and opposite. They won't do any net work for you!

© T.E. BEARDEN 1999

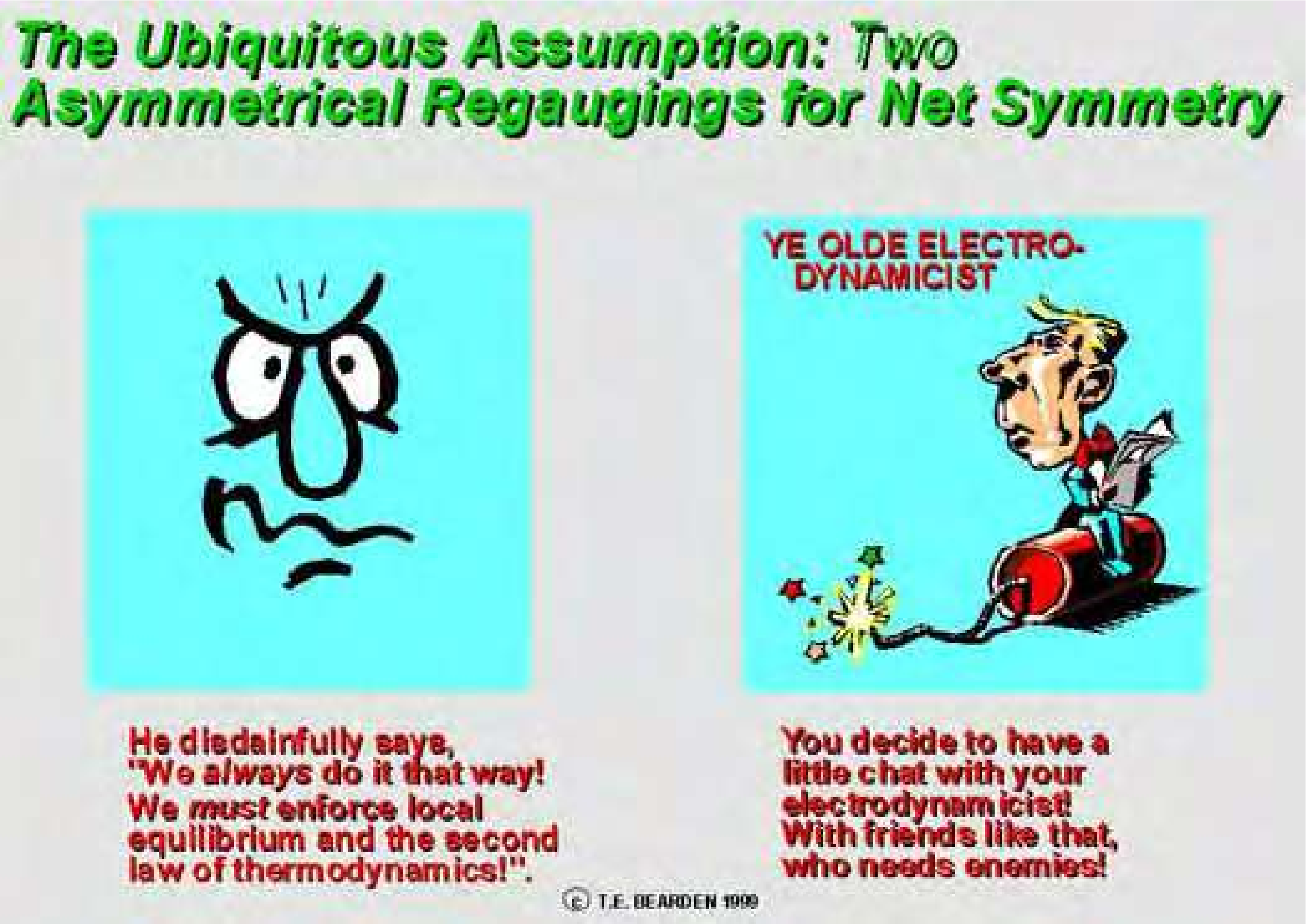

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

YE OLDE ELECTRO-
DYNAMICIST

He disdainfully says,
"We *always* do it that way!
We *must* enforce local
equilibrium and the second
law of thermodynamics!".

You decide to have a
little chat with your
electrodynamicist!
With friends like that,
who needs enemies!

© T.E. BEARDEN 1999

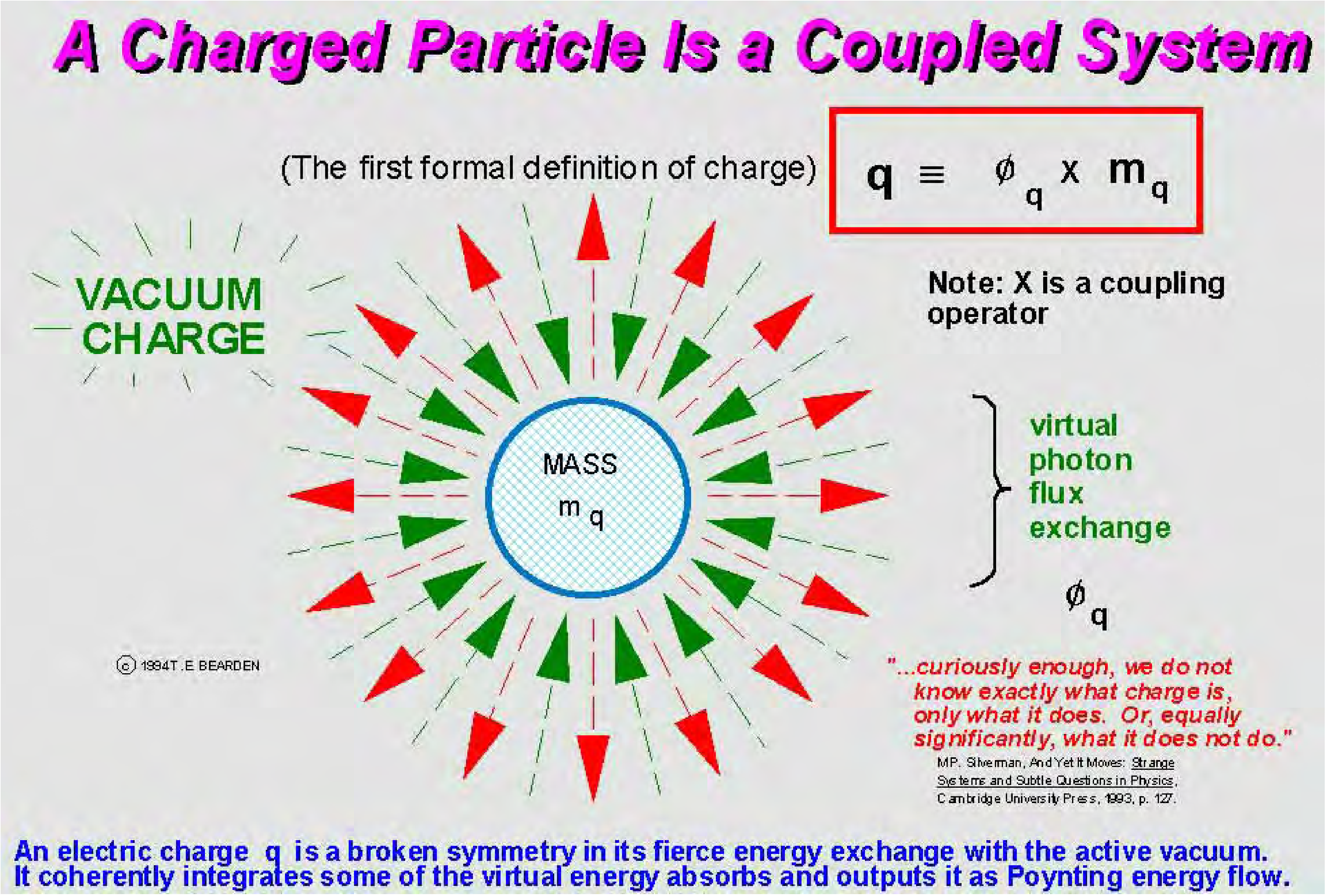

A Charged Particle Is a Coupled System

(The first formal definition of charge)

$$\boxed{q \equiv \phi_q \ \mathrm{x} \ m_q}$$

Note: X is a coupling operator

© 1994 T.E BEARDEN

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge q is a broken symmetry in its fierce energy exchange with the active vacuum. It coherently integrates some of the virtual energy absorbs and outputs it as Poynting energy flow.

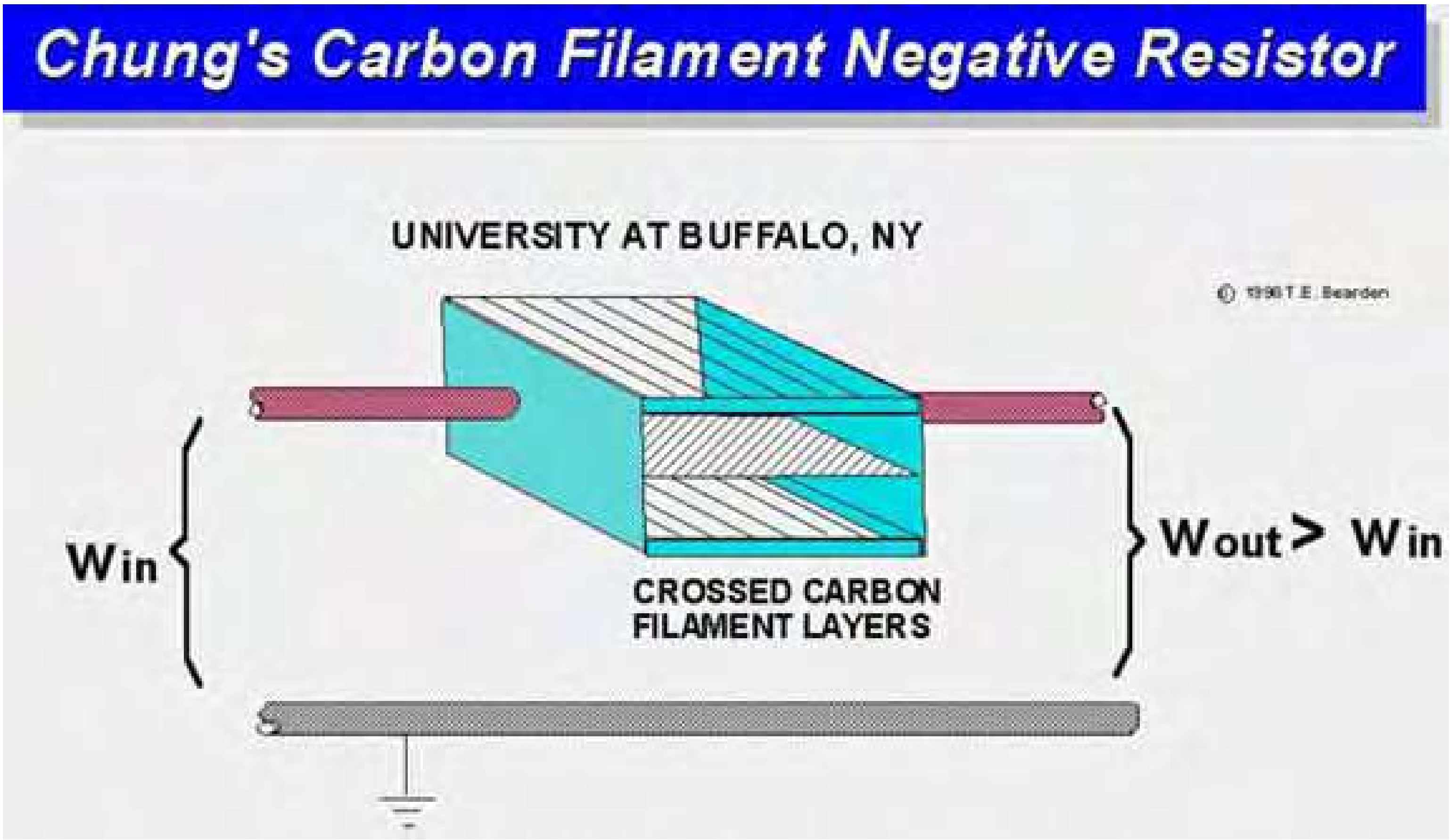

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

© 1996 T.E. Bearden

Win
CROSSED CARBON FILAMENT LAYERS
Wout > Win

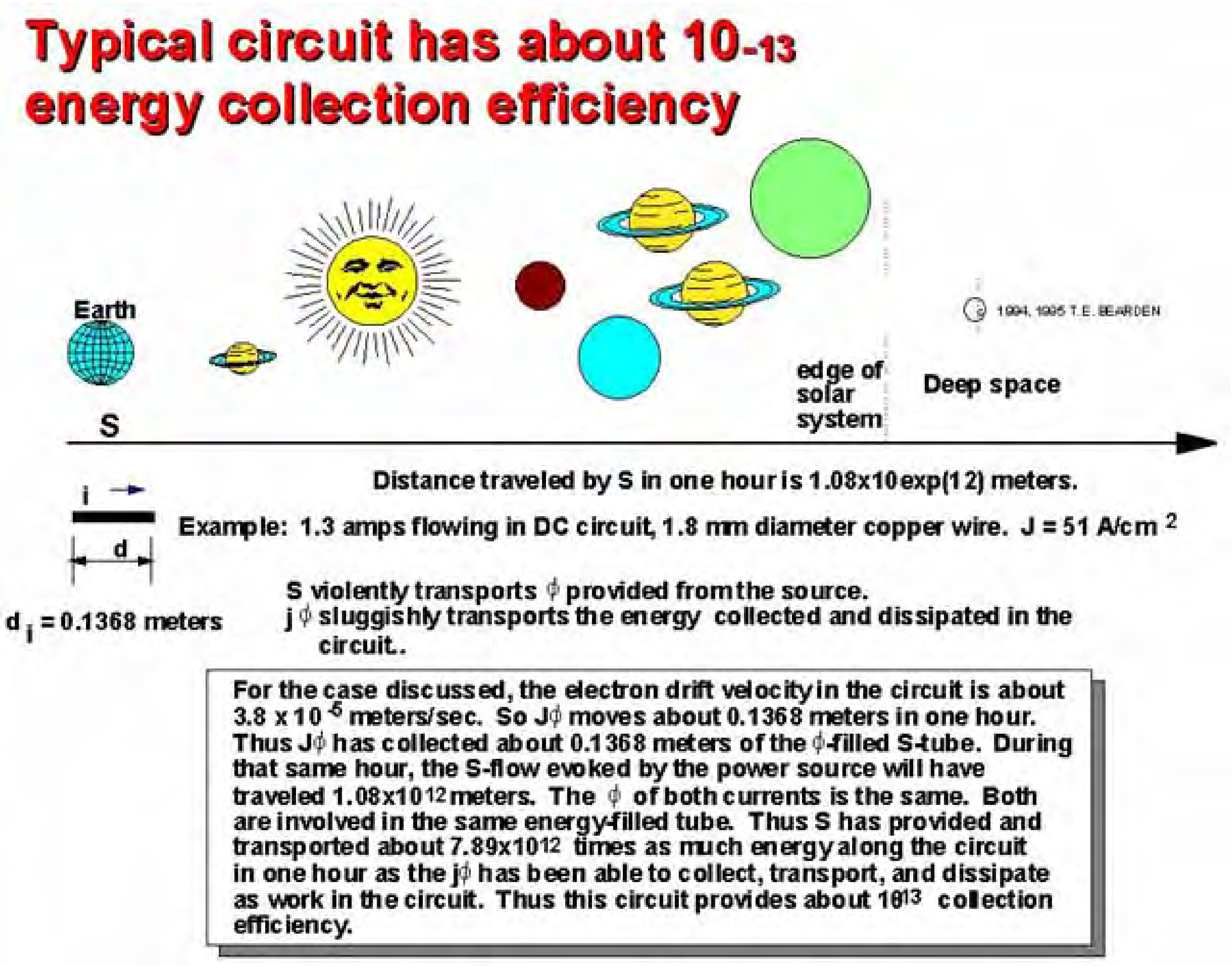

Typical circuit has about 10-13 energy collection efficiency

Distance traveled by S in one hour is 1.08x10exp(12) meters.

Example: 1.3 amps flowing in DC circuit, 1.8 mm diameter copper wire. $J = 51\ A/cm^2$

$d_i = 0.1368$ meters

S violently transports ϕ provided from the source.
j ϕ sluggishly transports the energy collected and dissipated in the circuit..

For the case discussed, the electron drift velocity in the circuit is about 3.8×10^{-5} meters/sec. So Jϕ moves about 0.1368 meters in one hour. Thus Jϕ has collected about 0.1368 meters of the ϕ-filled S-tube. During that same hour, the S-flow evoked by the power source will have traveled 1.08×10^{12} meters. The ϕ of both currents is the same. Both are involved in the same energy-filled tube. Thus S has provided and transported about 7.89×10^{12} times as much energy along the circuit in one hour as the jϕ has been able to collect, transport, and dissipate as work in the circuit. Thus this circuit provides about 10^{13} collection efficiency.

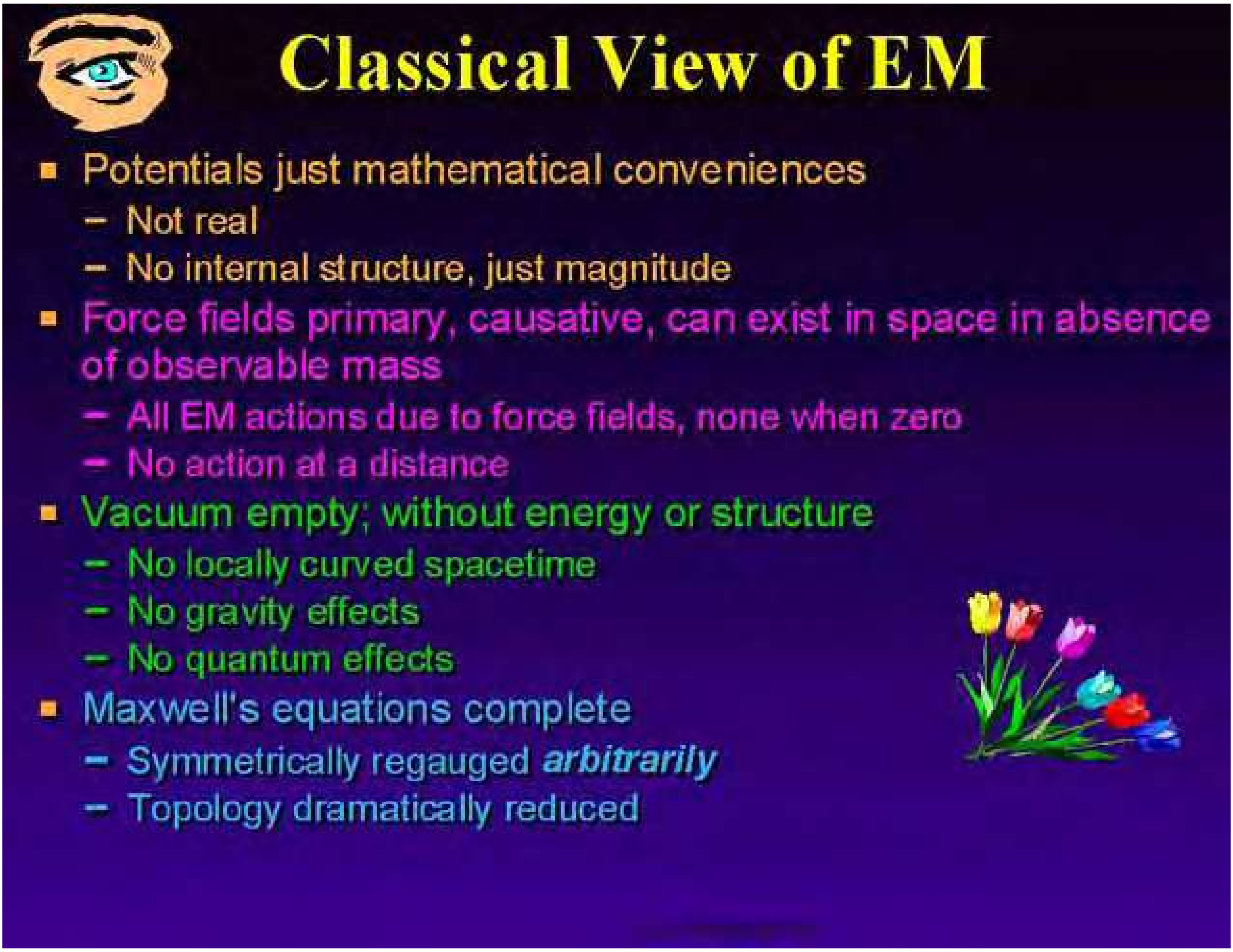

Classical View of EM

- Potentials just mathematical conveniences
  - Not real
  - No internal structure, just magnitude
- Force fields primary, causative, can exist in space in absence of observable mass
  - All EM actions due to force fields, none when zero
  - No action at a distance
- Vacuum empty; without energy or structure
  - No locally curved spacetime
  - No gravity effects
  - No quantum effects
- Maxwell's equations complete
  - Symmetrically regauged ***arbitrarily***
  - Topology dramatically reduced

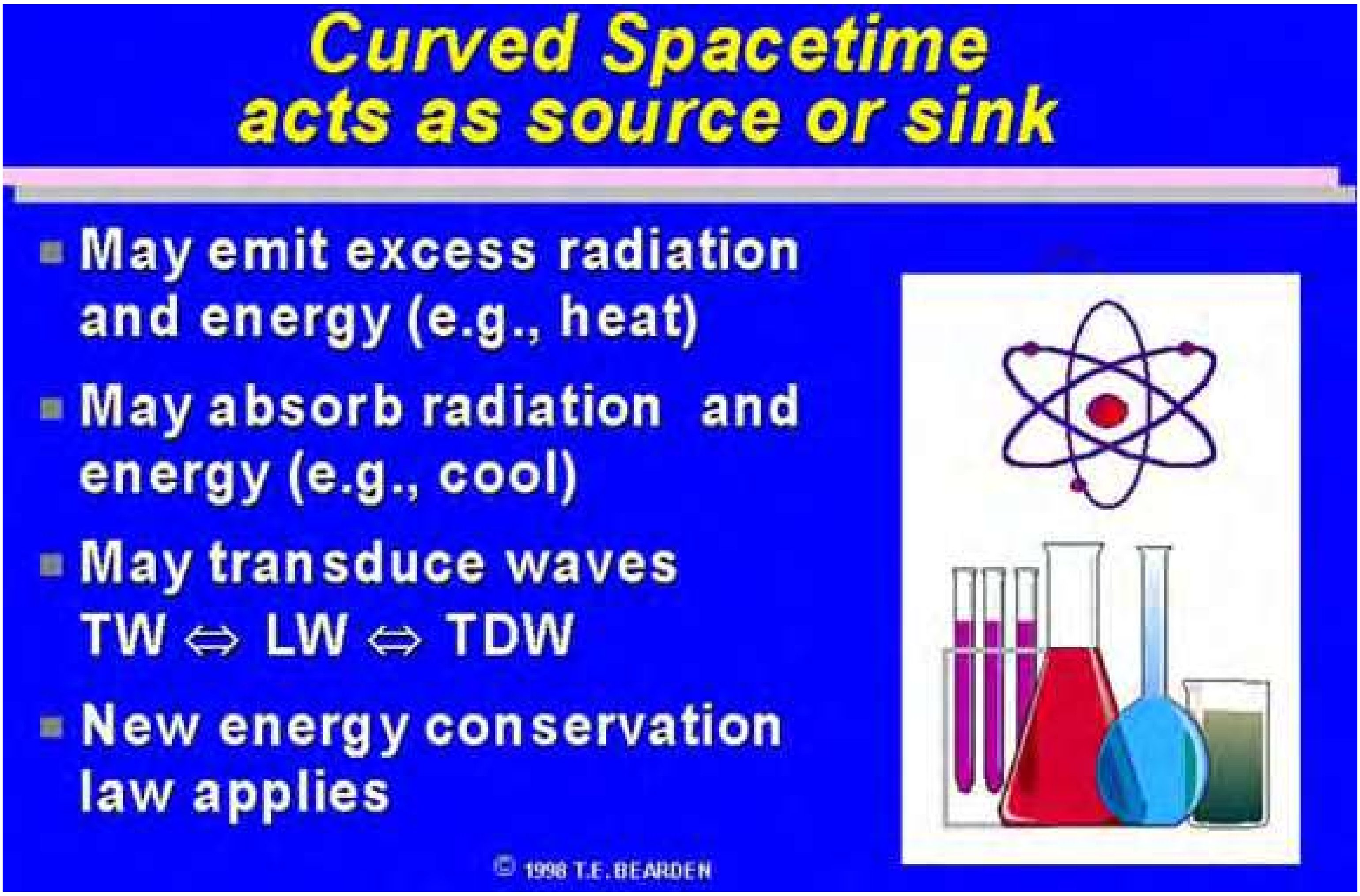

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E. BEARDEN

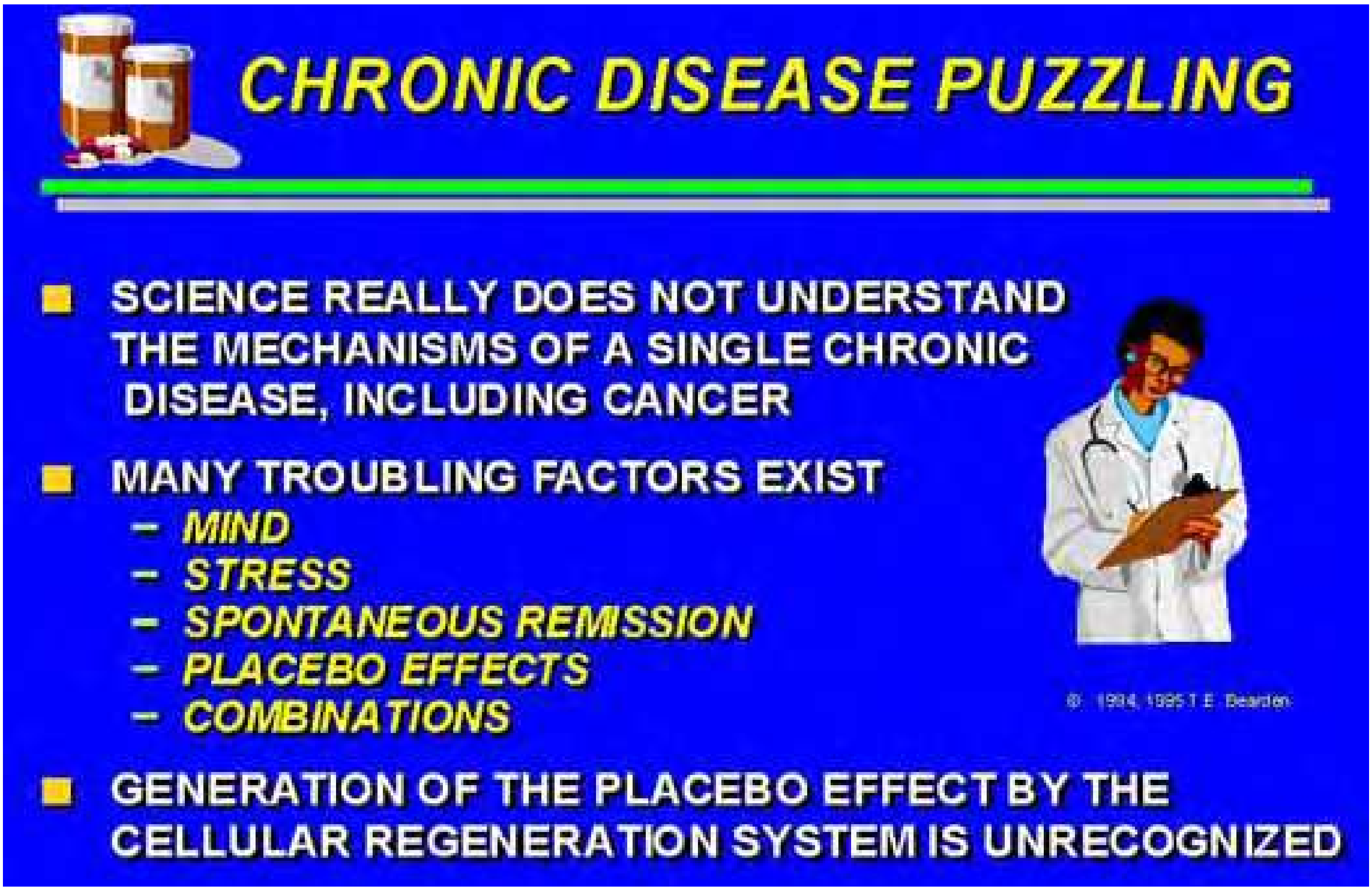

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

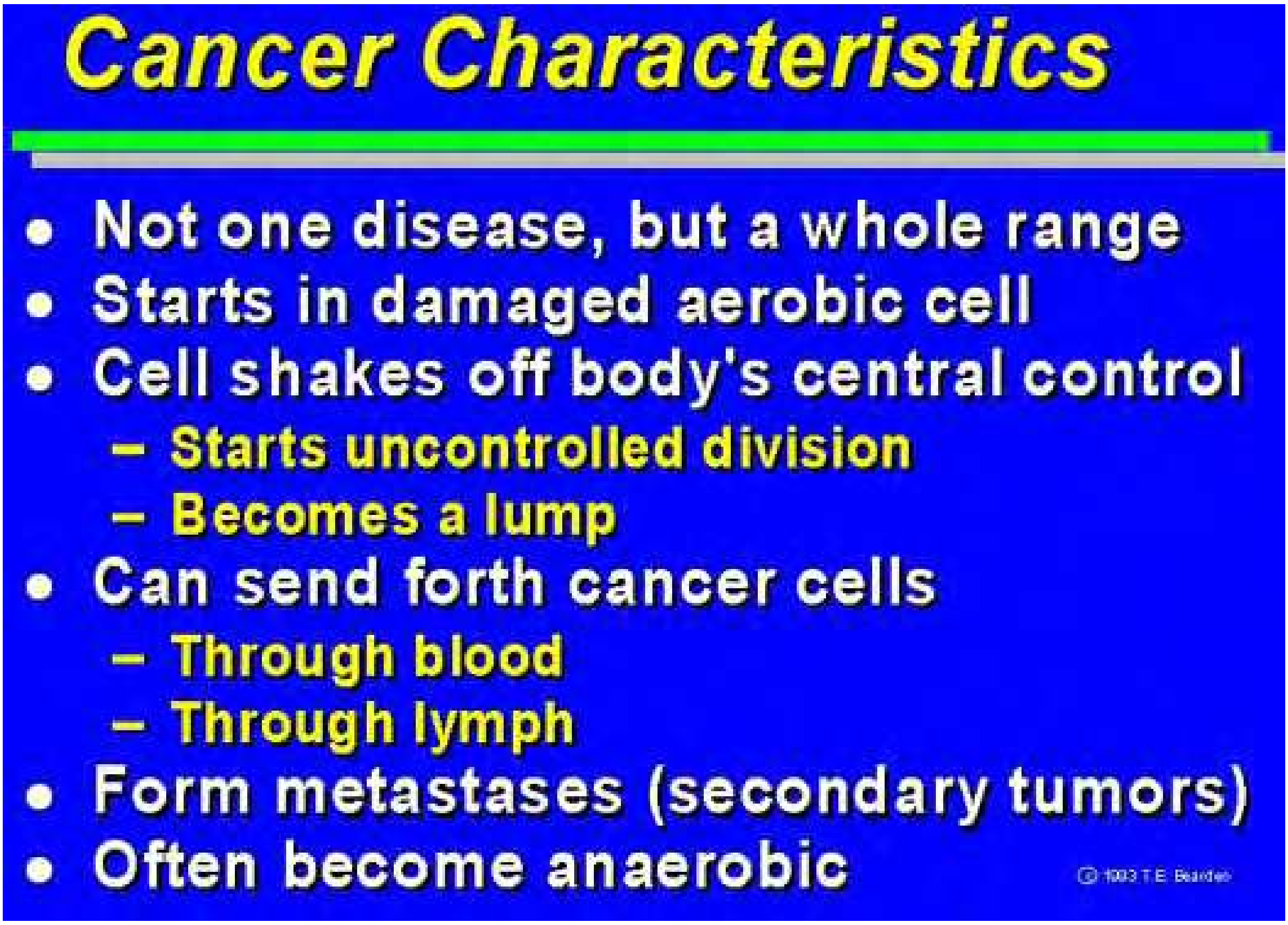

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

© 1993 T.E. Bearden

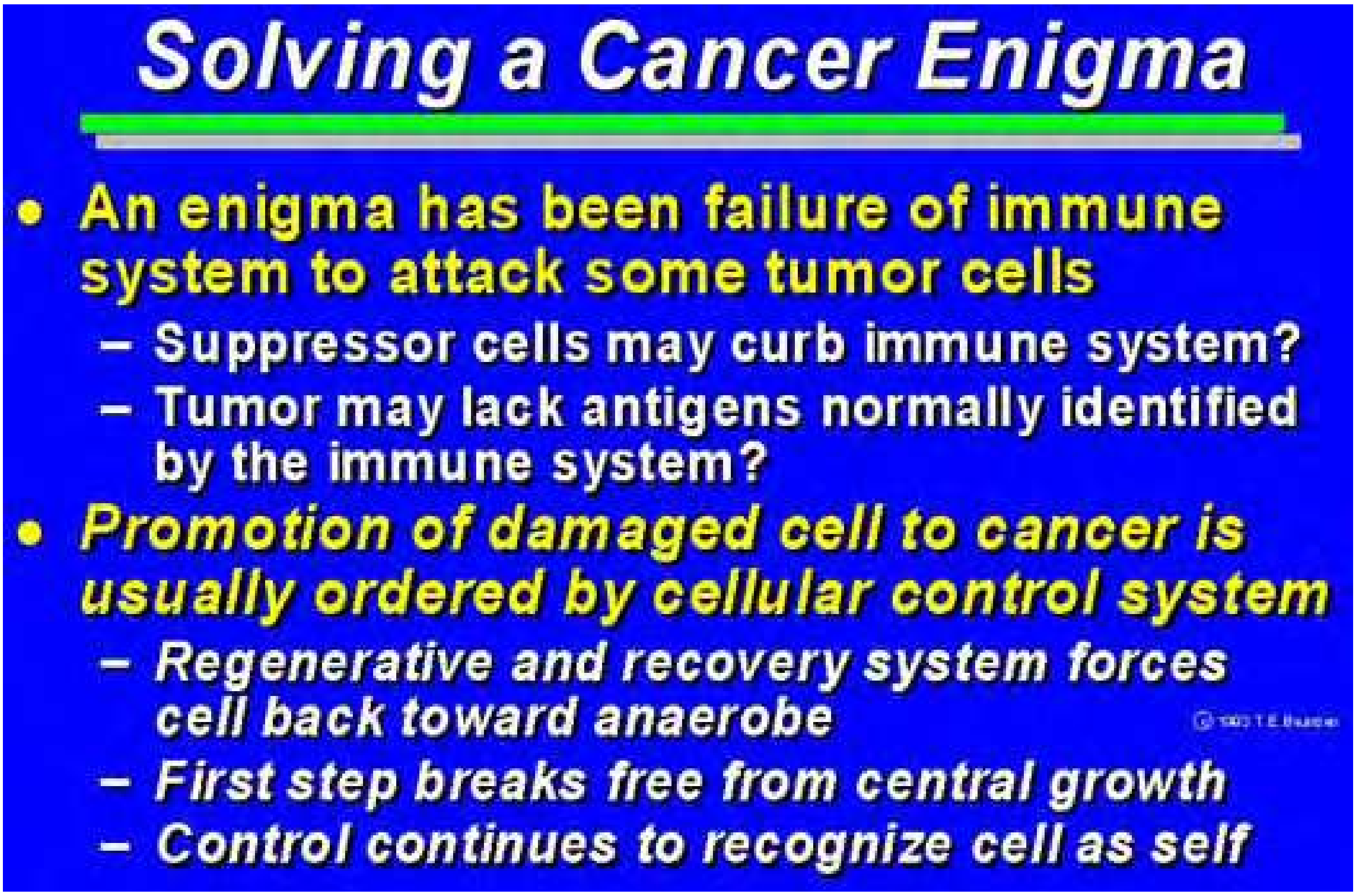

Solving a Cancer Enigma

- **An enigma has been failure of immune system to attack some tumor cells**
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- ***Promotion of damaged cell to cancer is usually ordered by cellular control system***
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

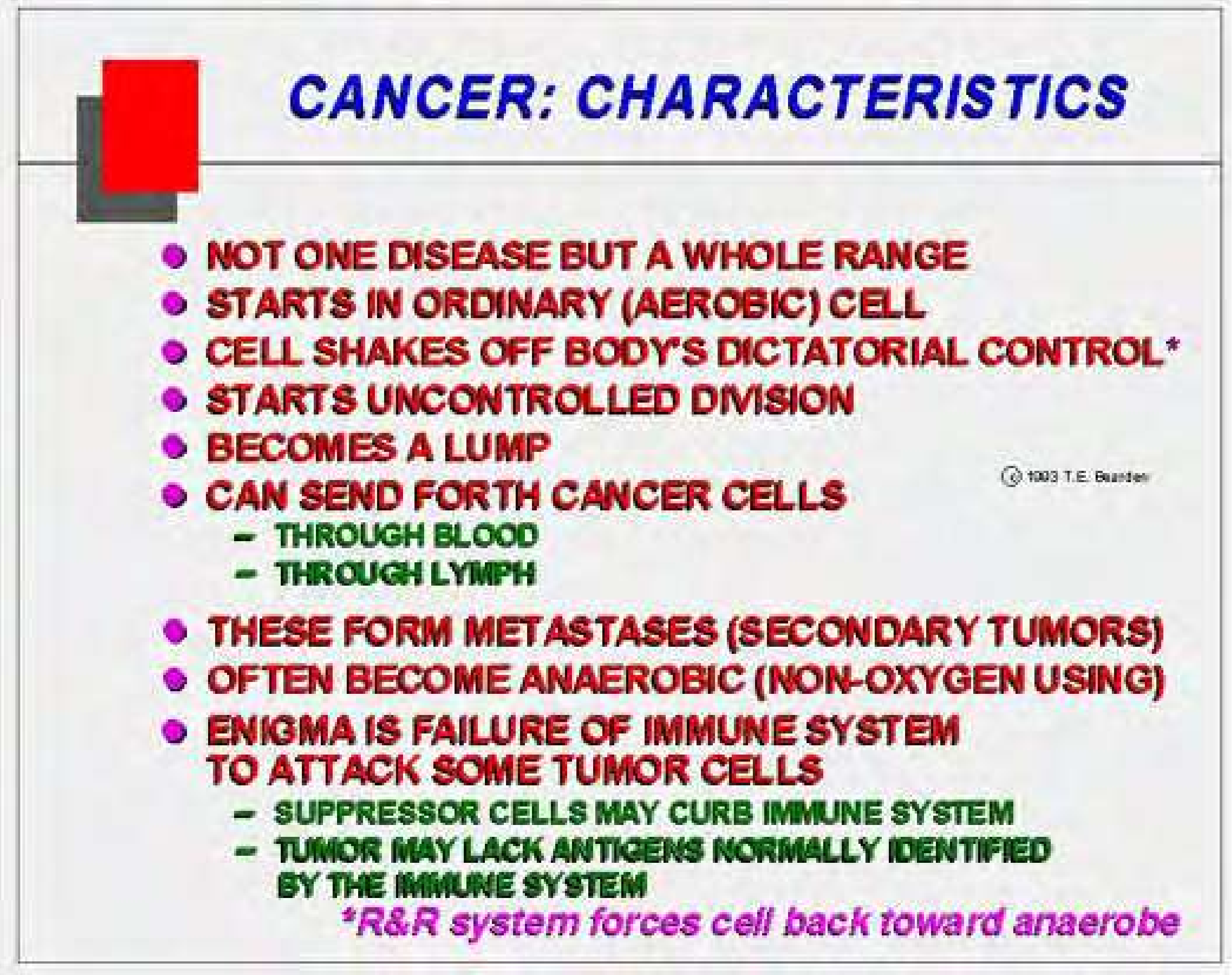

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

*R&R system forces cell back toward anaerobe

© 1993 T.E. Bearden

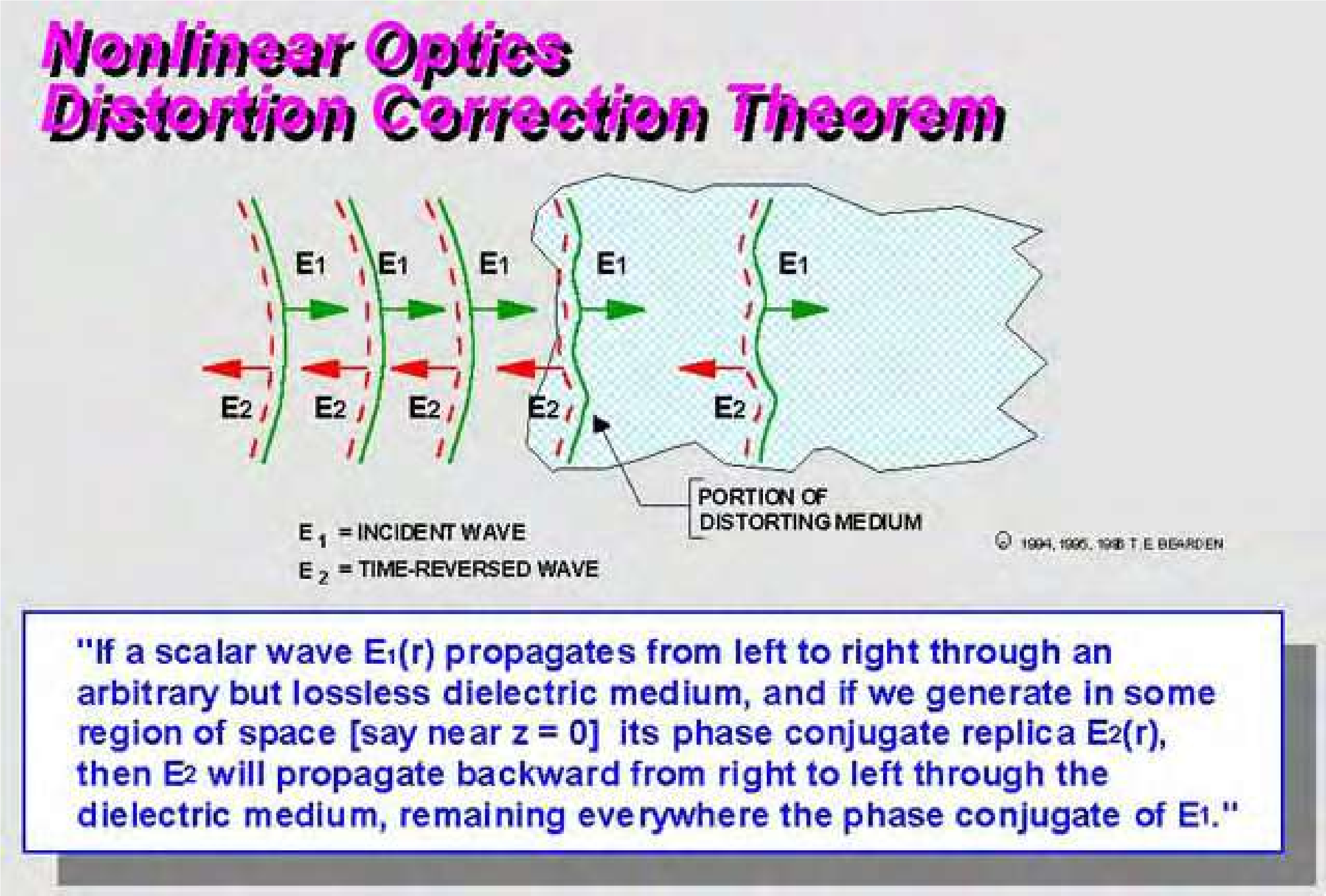

Nonlinear Optics Distortion Correction Theorem

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

© 1994, 1995, 1998 T E BEARDEN

> "If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

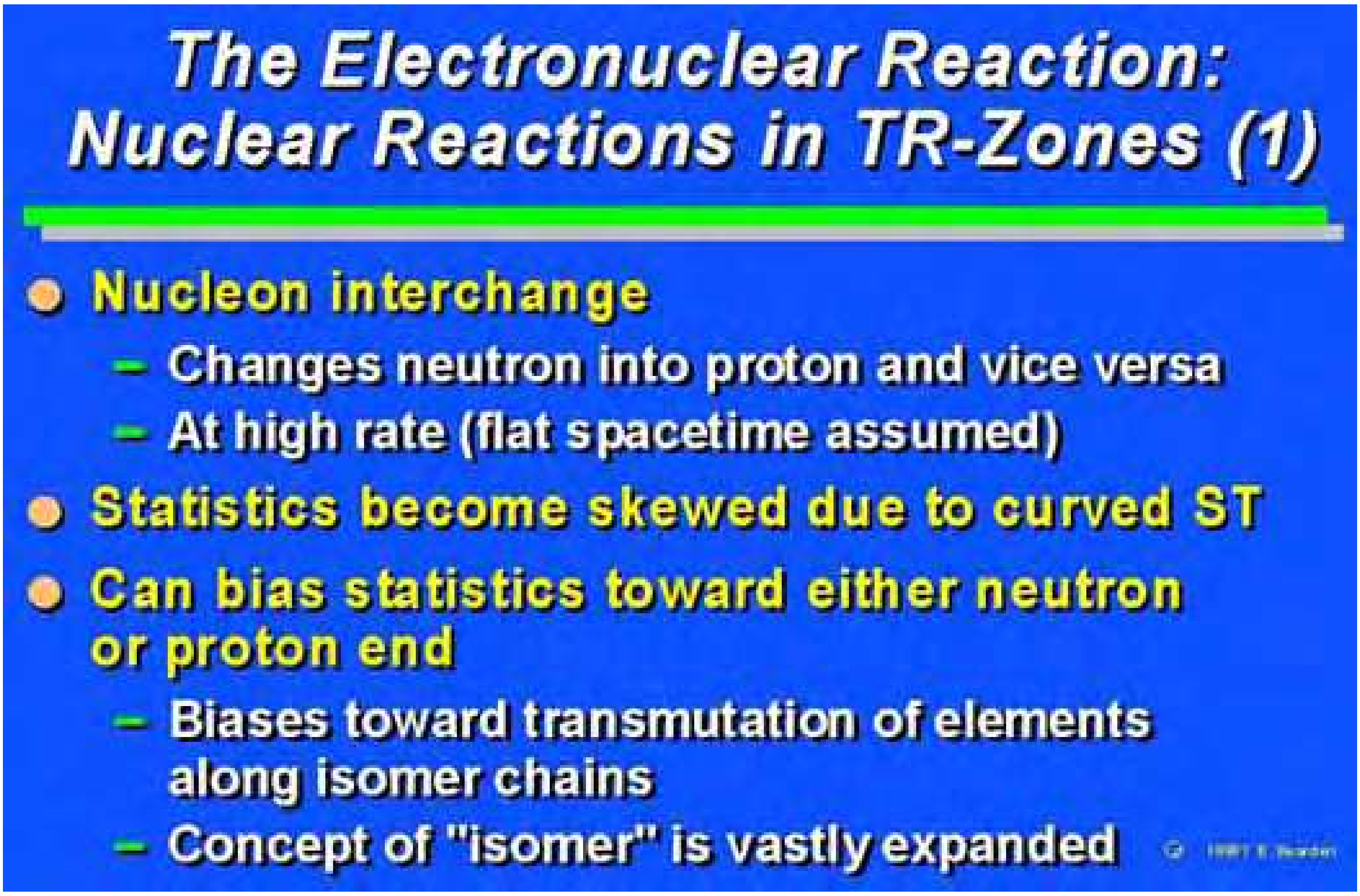

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- Nucleon interchange
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- Statistics become skewed due to curved ST
- Can bias statistics toward either neutron or proton end
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

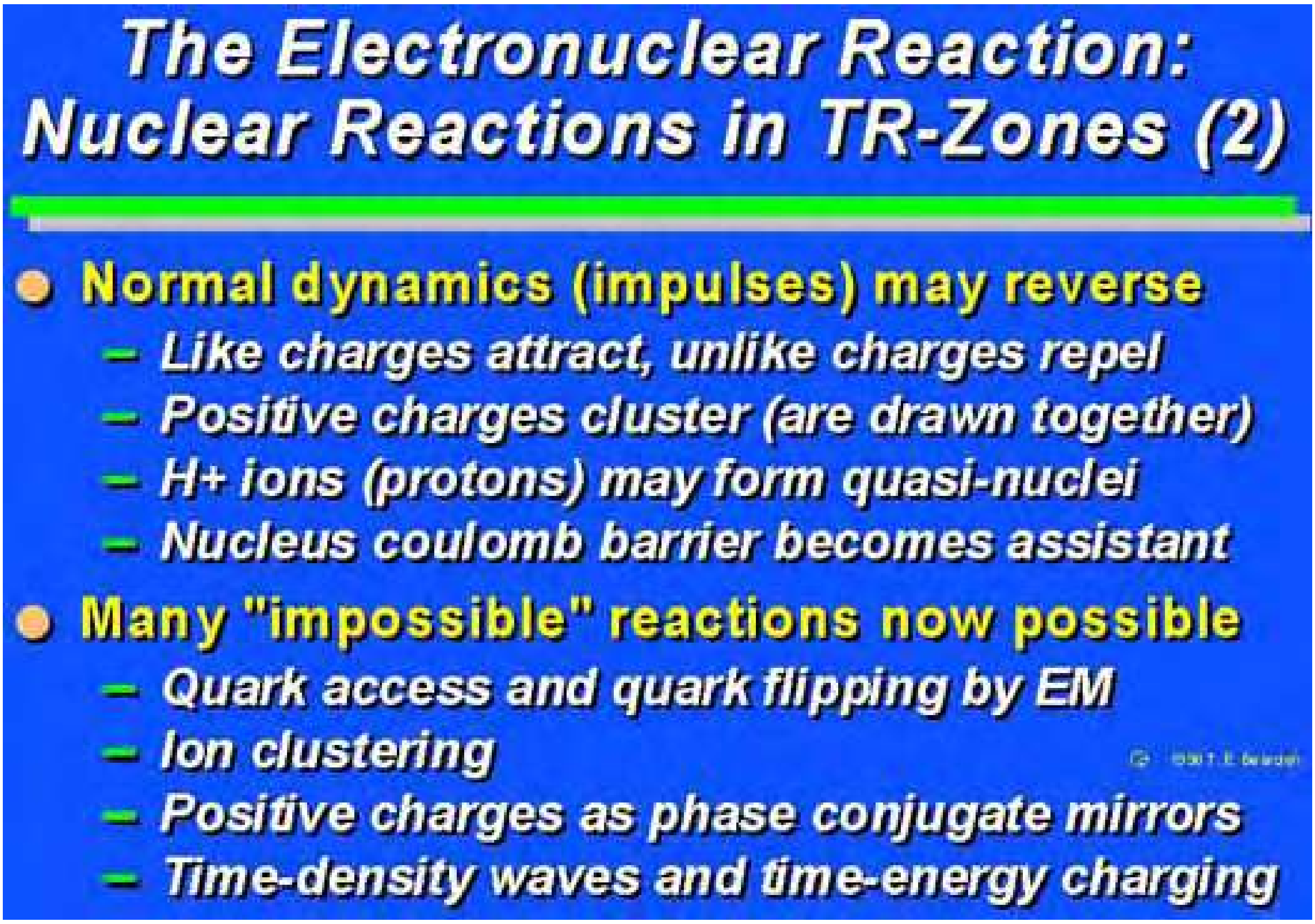

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- **Normal dynamics (impulses) may reverse**
    - *Like charges attract, unlike charges repel*
    - *Positive charges cluster (are drawn together)*
    - *H+ ions (protons) may form quasi-nuclei*
    - *Nucleus coulomb barrier becomes assistant*
- **Many "impossible" reactions now possible**
    - *Quark access and quark flipping by EM*
    - *Ion clustering*
    - *Positive charges as phase conjugate mirrors*
    - *Time-density waves and time-energy charging*

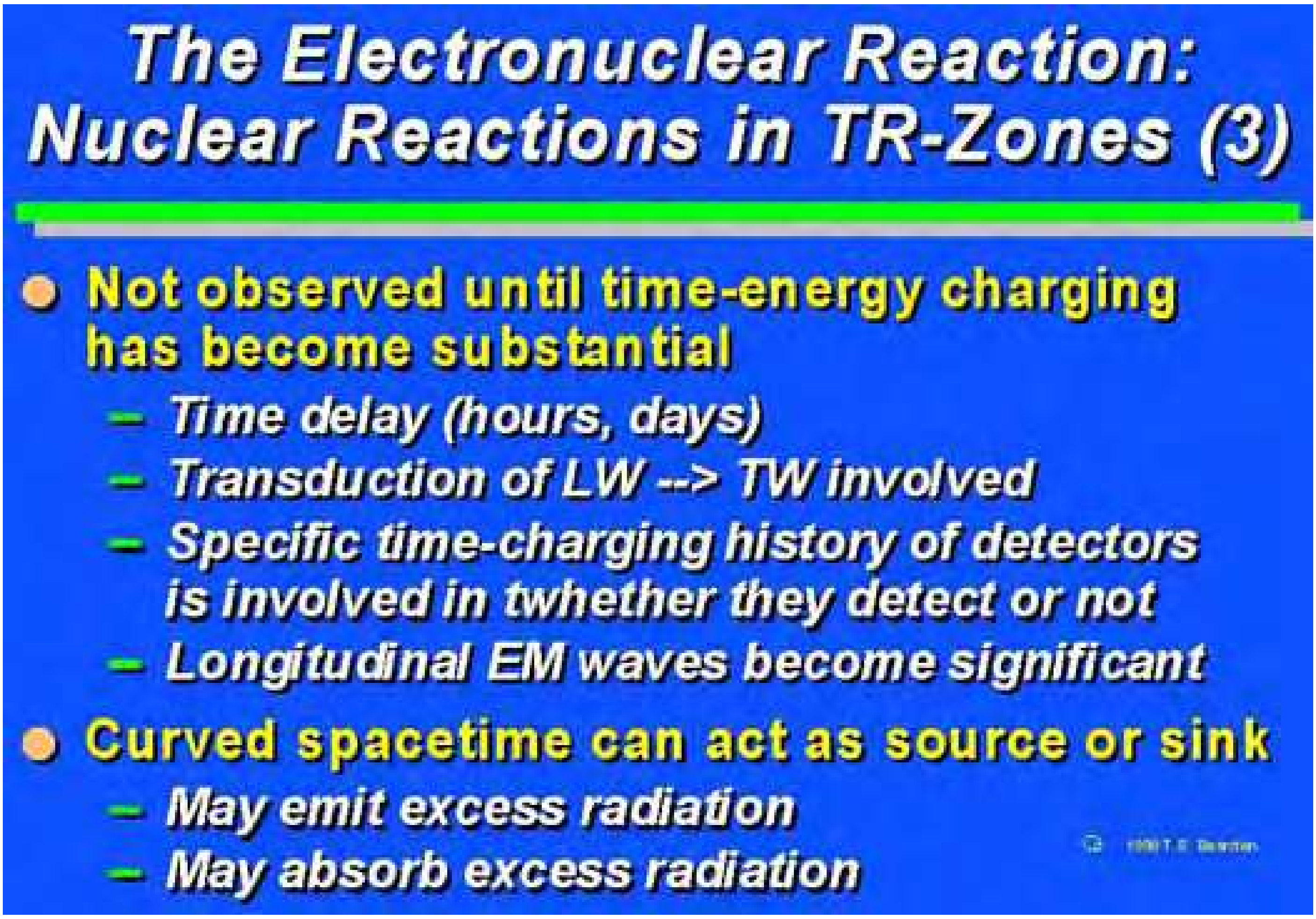

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- **Not observed until time-energy charging has become substantial**
    - *Time delay (hours, days)*
    - *Transduction of LW --> TW involved*
    - *Specific time-charging history of detectors is involved in twhether they detect or not*
    - *Longitudinal EM waves become significant*
- **Curved spacetime can act as source or sink**
    - *May emit excess radiation*
    - *May absorb excess radiation*

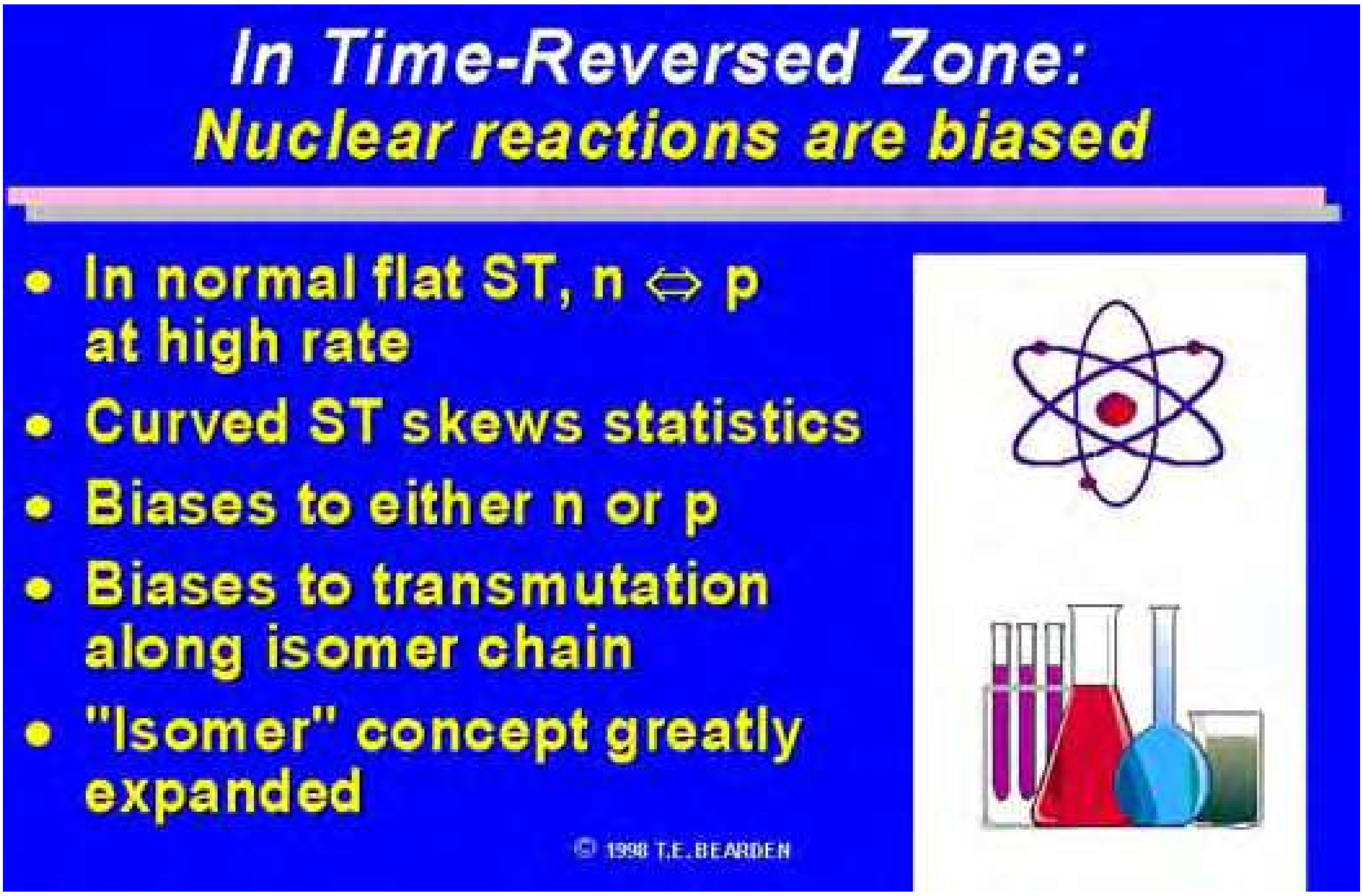

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T.E.BEARDEN

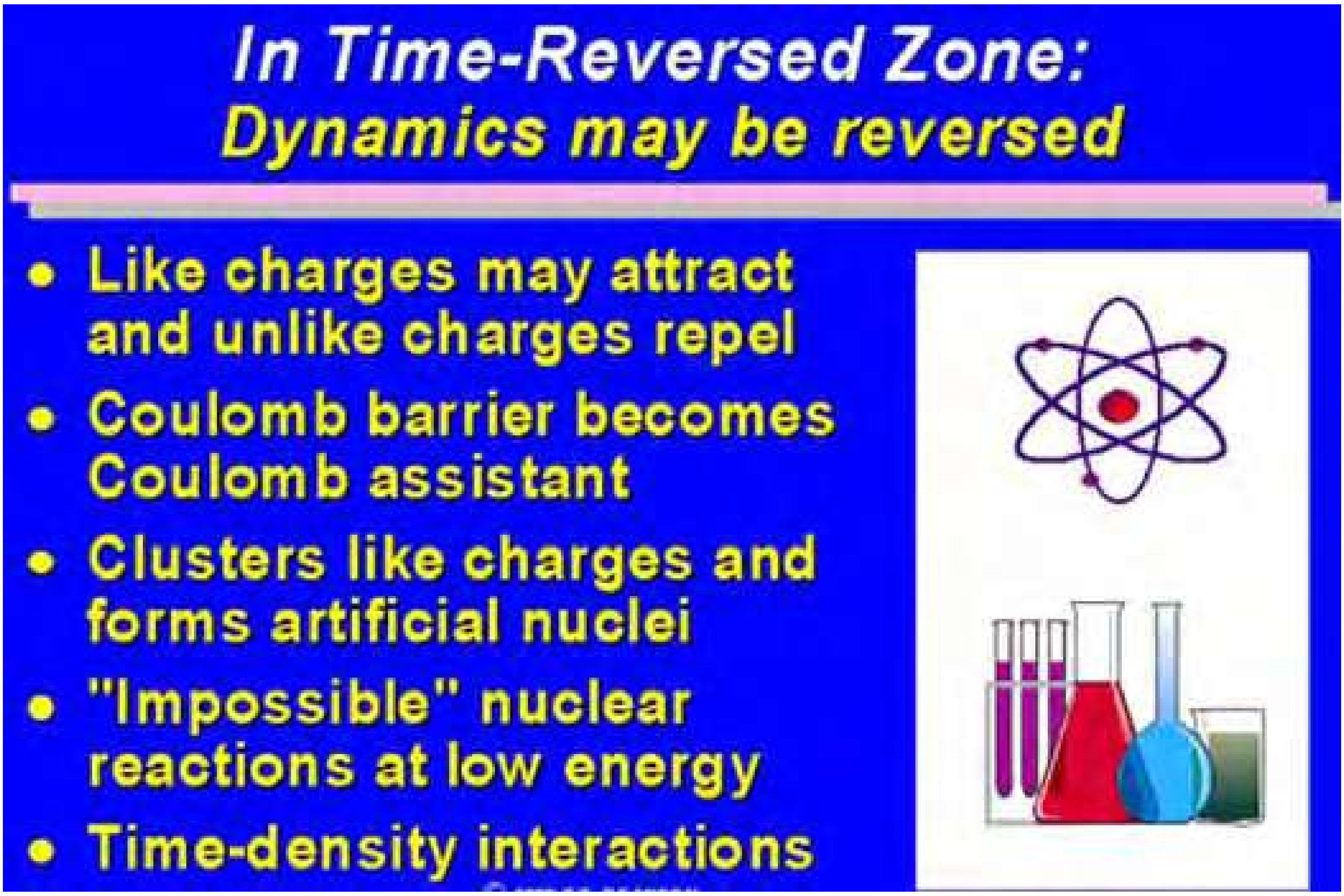

In Time-Reversed Zone:
Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

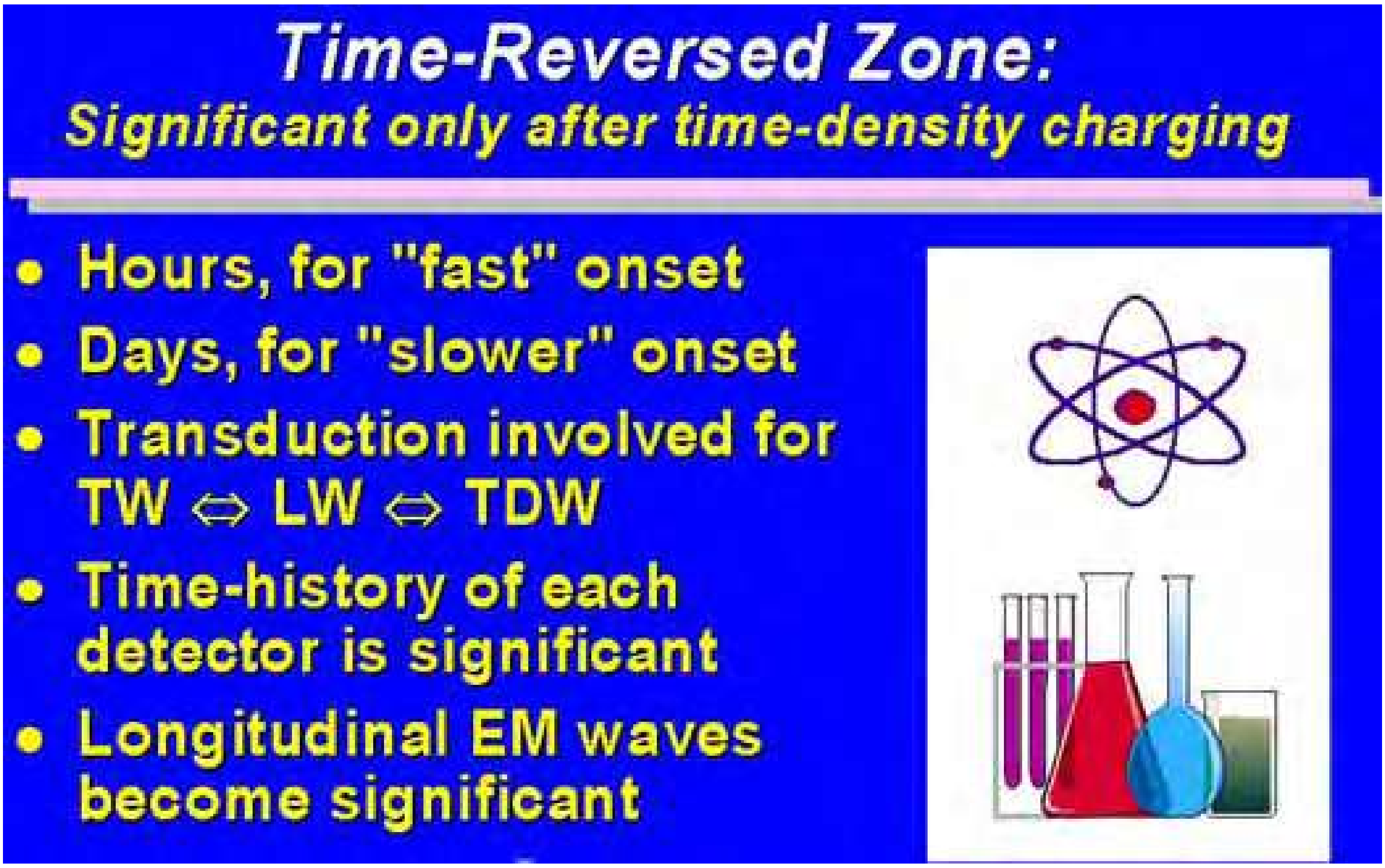

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

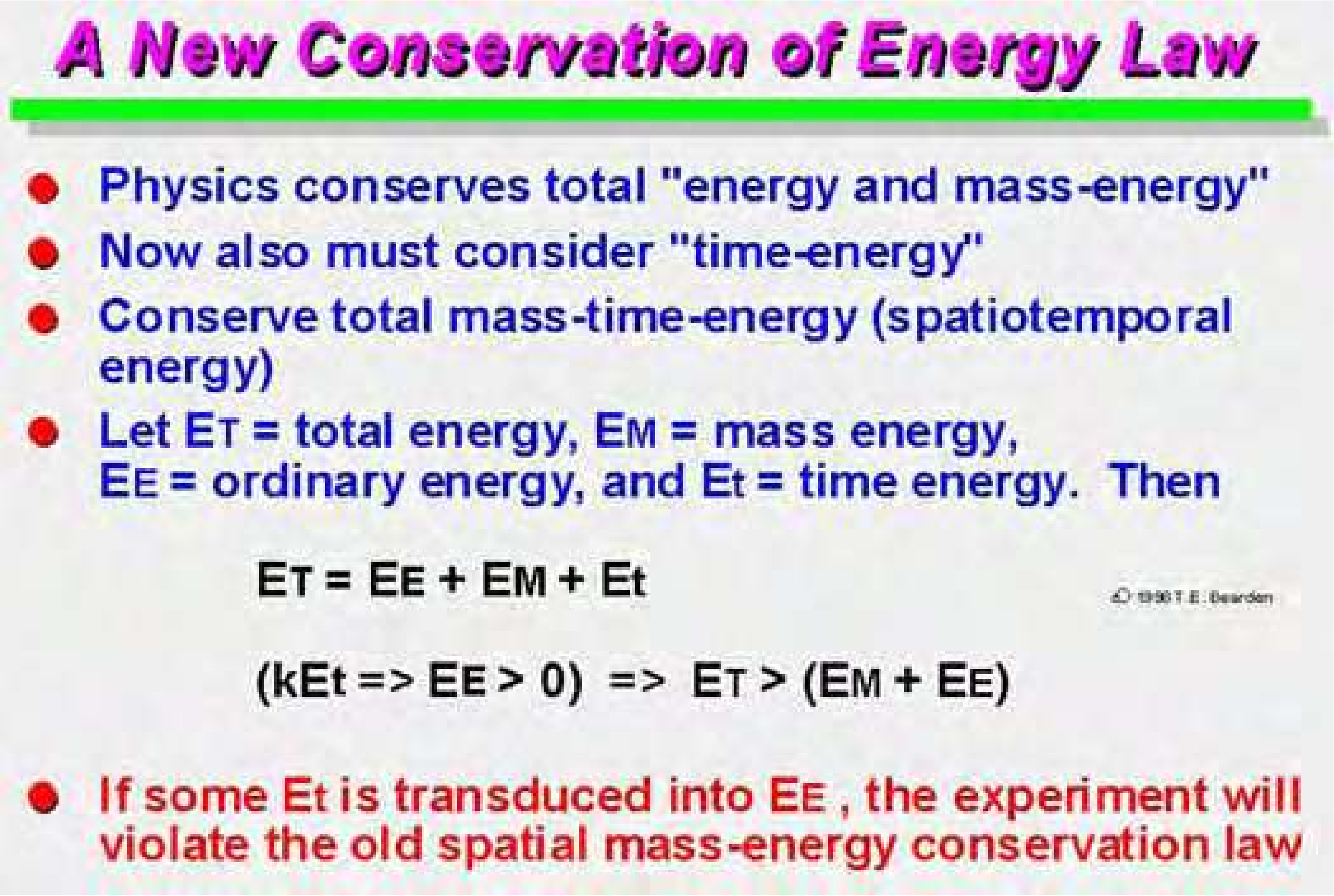

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

© 1998 T.E. Bearden

$$(kE_t \Rightarrow E_E > 0) \Rightarrow E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

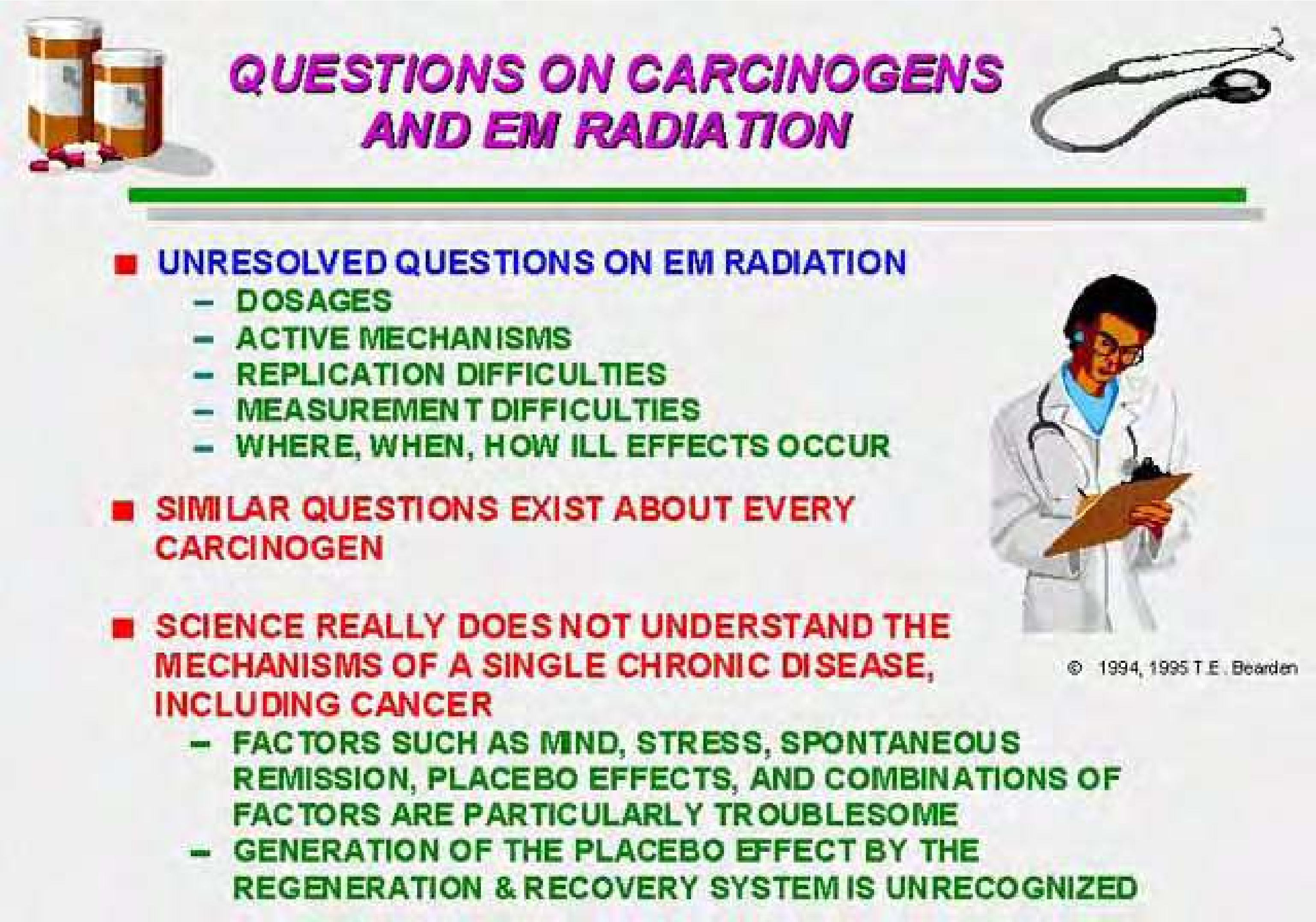

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

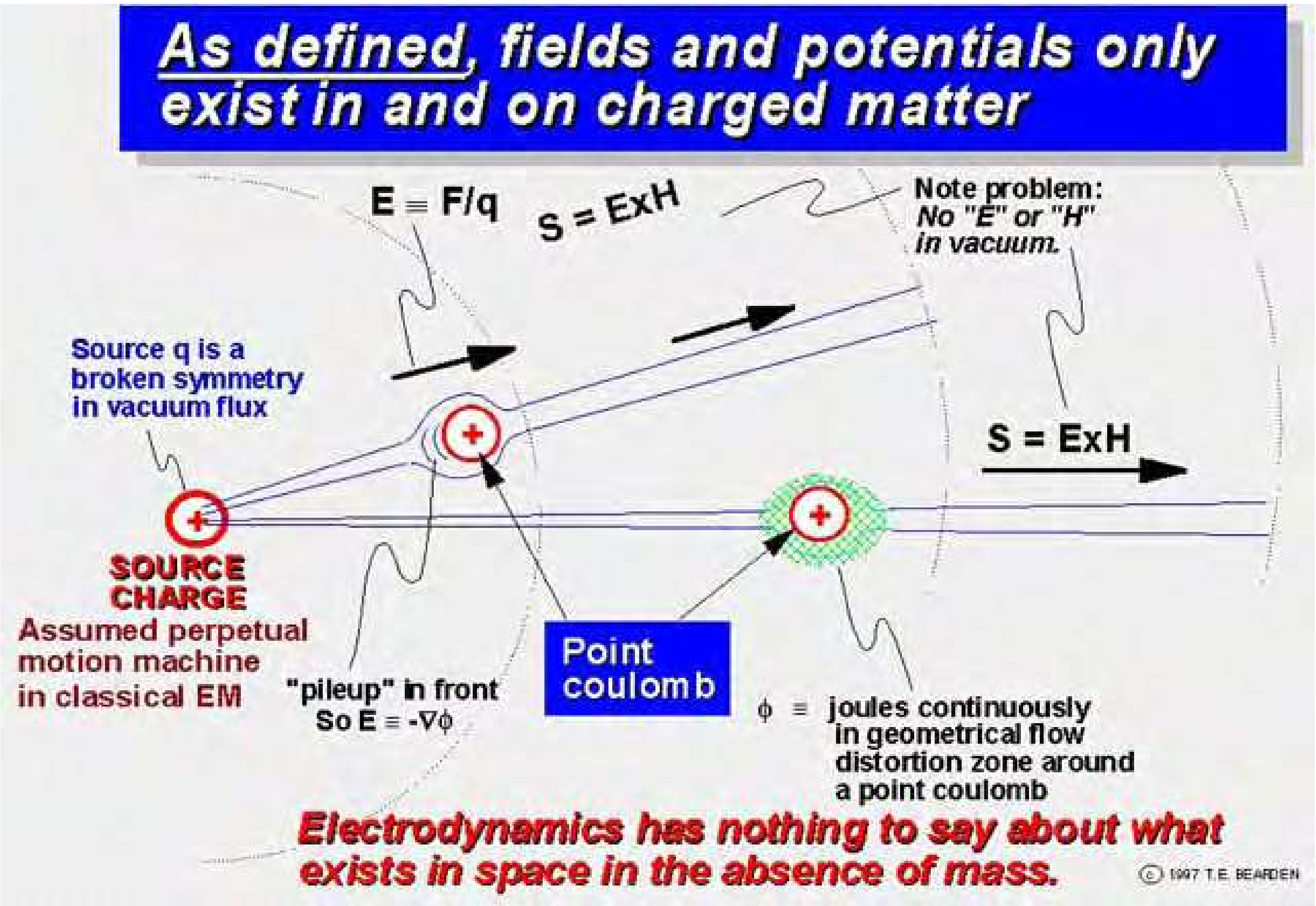

As defined, fields and potentials only exist in and on charged matter

$E \equiv F/q$

$S = E \times H$

Note problem:
No "E" or "H" in vacuum.

Source q is a broken symmetry in vacuum flux

$S = E \times H$

SOURCE CHARGE

Assumed perpetual motion machine in classical EM

"pileup" in front
So $E \equiv -\nabla\phi$

Point coulomb

$\phi \equiv$ joules continuously in geometrical flow distortion zone around a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

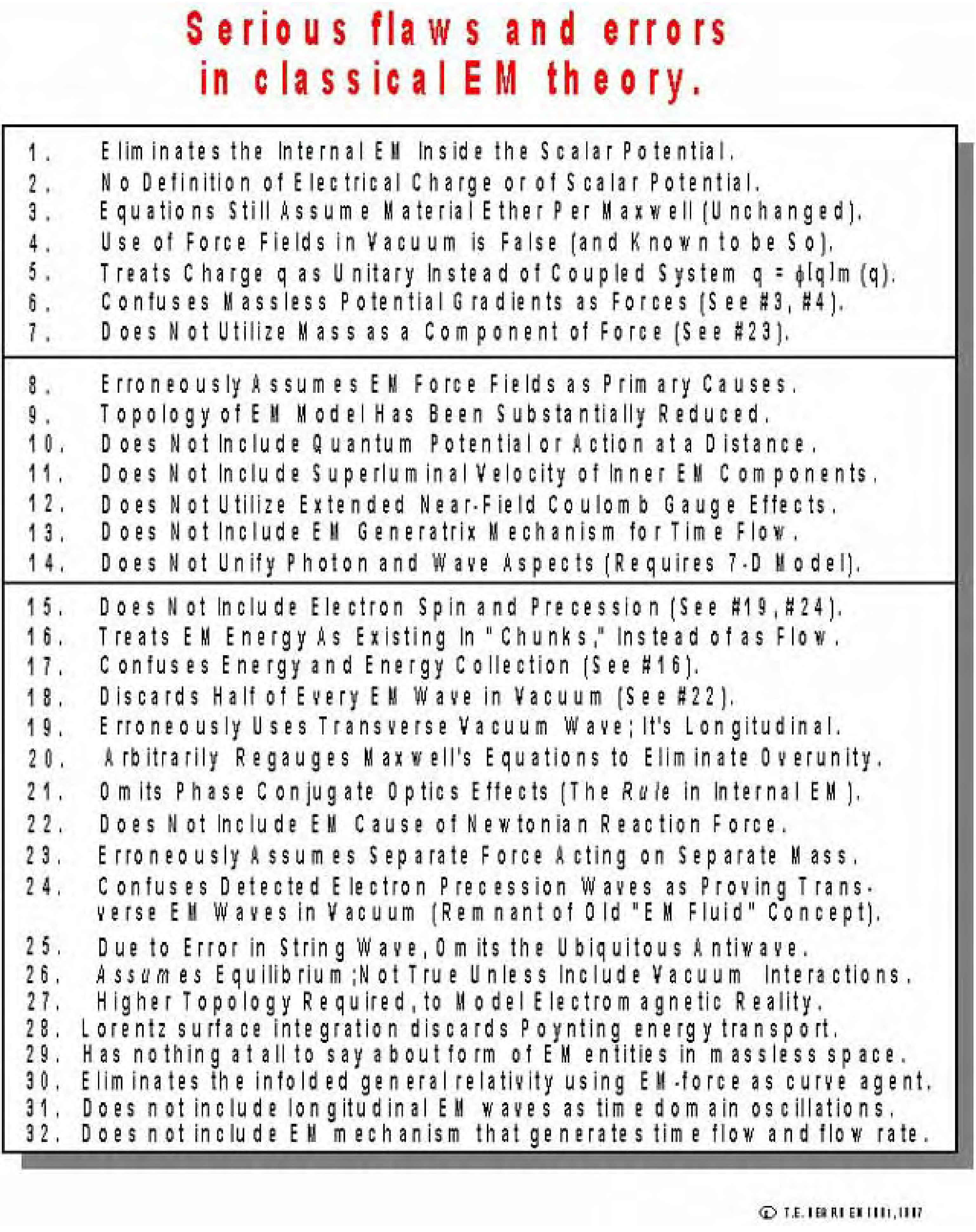

Serious flaws and errors in classical EM theory.

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi[q]m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

© T.E. BEARDEN 1991, 1997

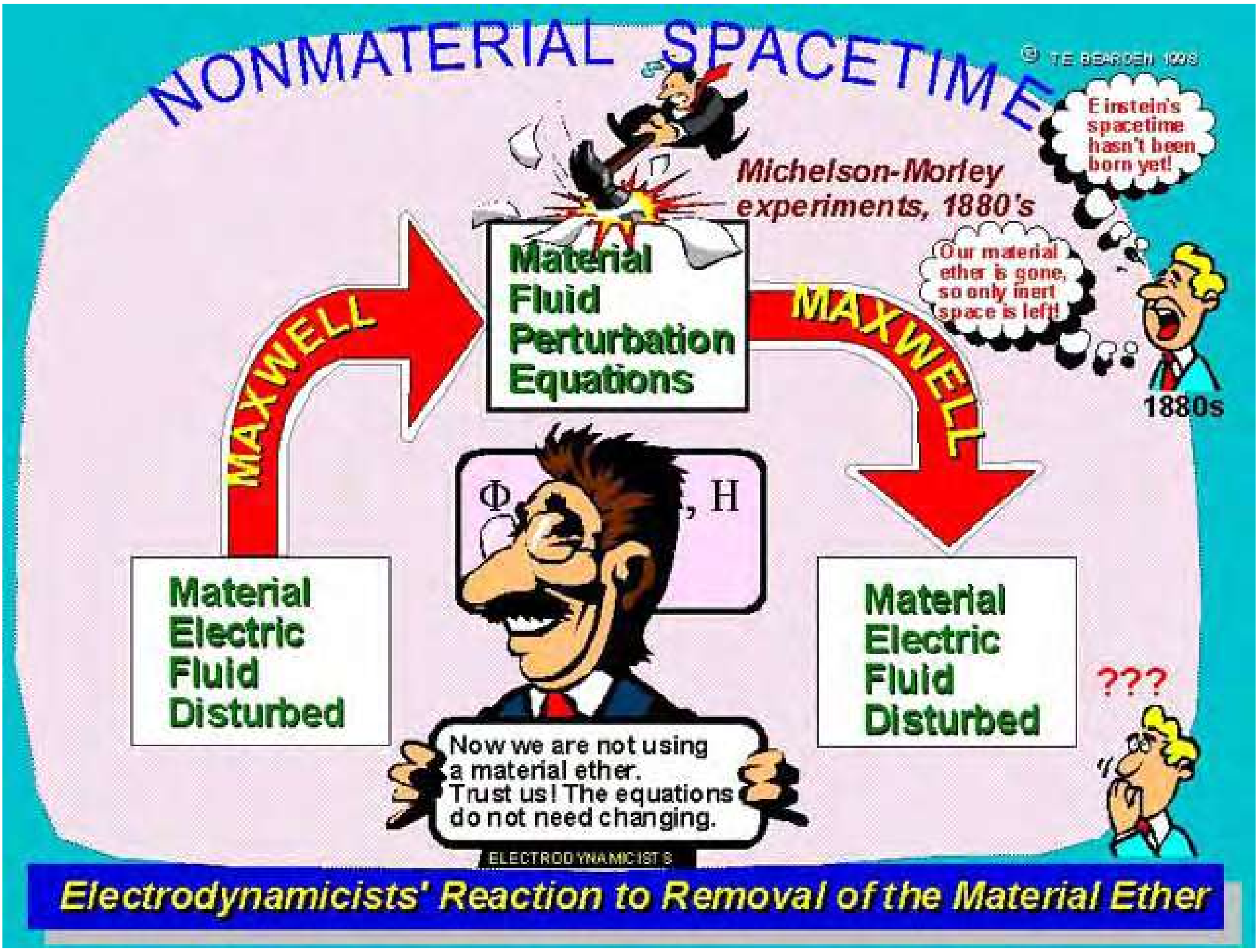
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Einstein's spacetime hasn't been born yet!
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left!
1880s
MAXWELL
Material Fluid Perturbation Equations
MAXWELL
Φ
, H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

REACTION TO LOSS OF MATERIAL ETHER

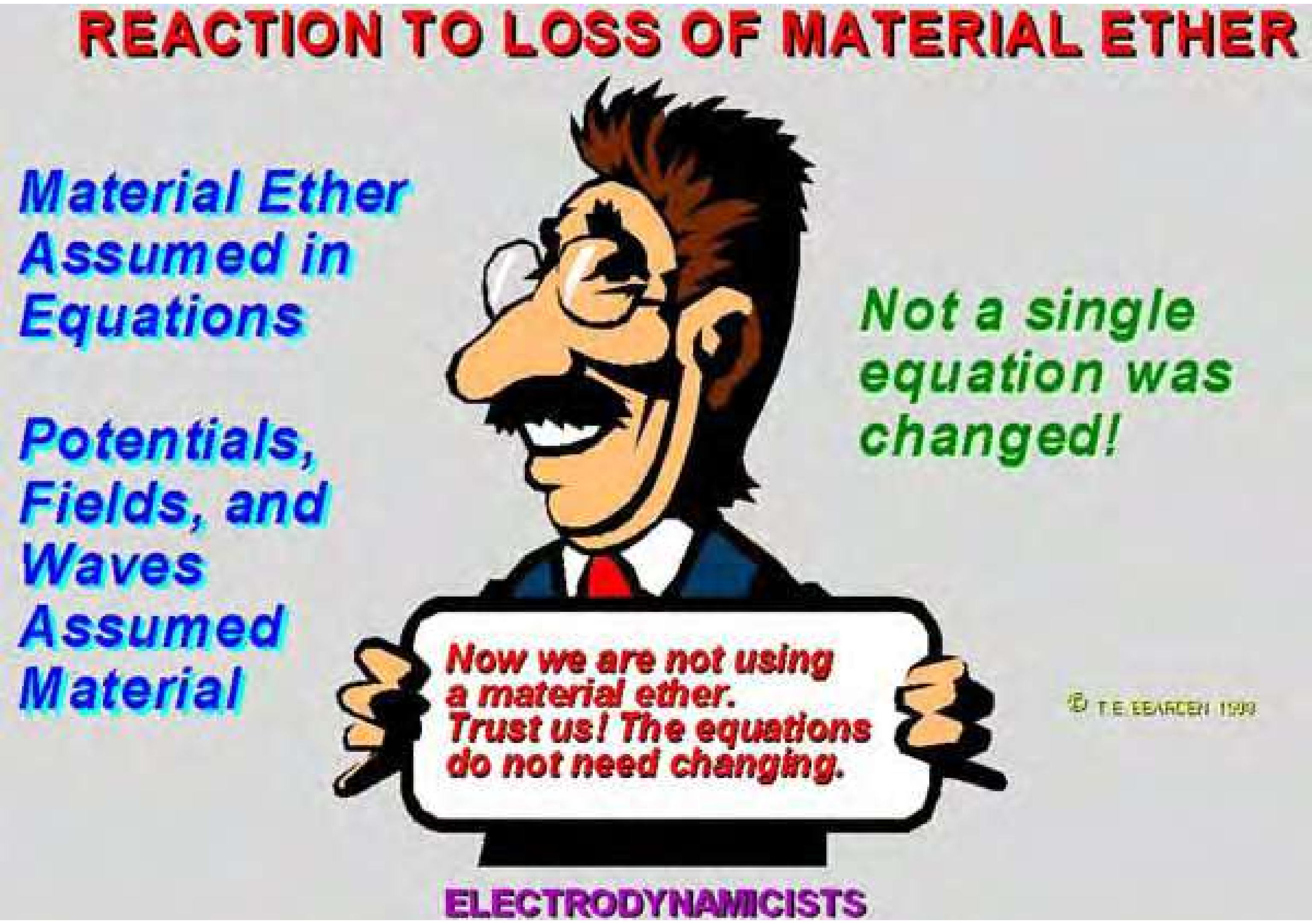

Material Ether
Assumed in
Equations

Potentials,
Fields, and
Waves
Assumed
Material

Not a single
equation was
changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.
ELECTRODYNAMICISTS

© T.E. BEARDEN 1999

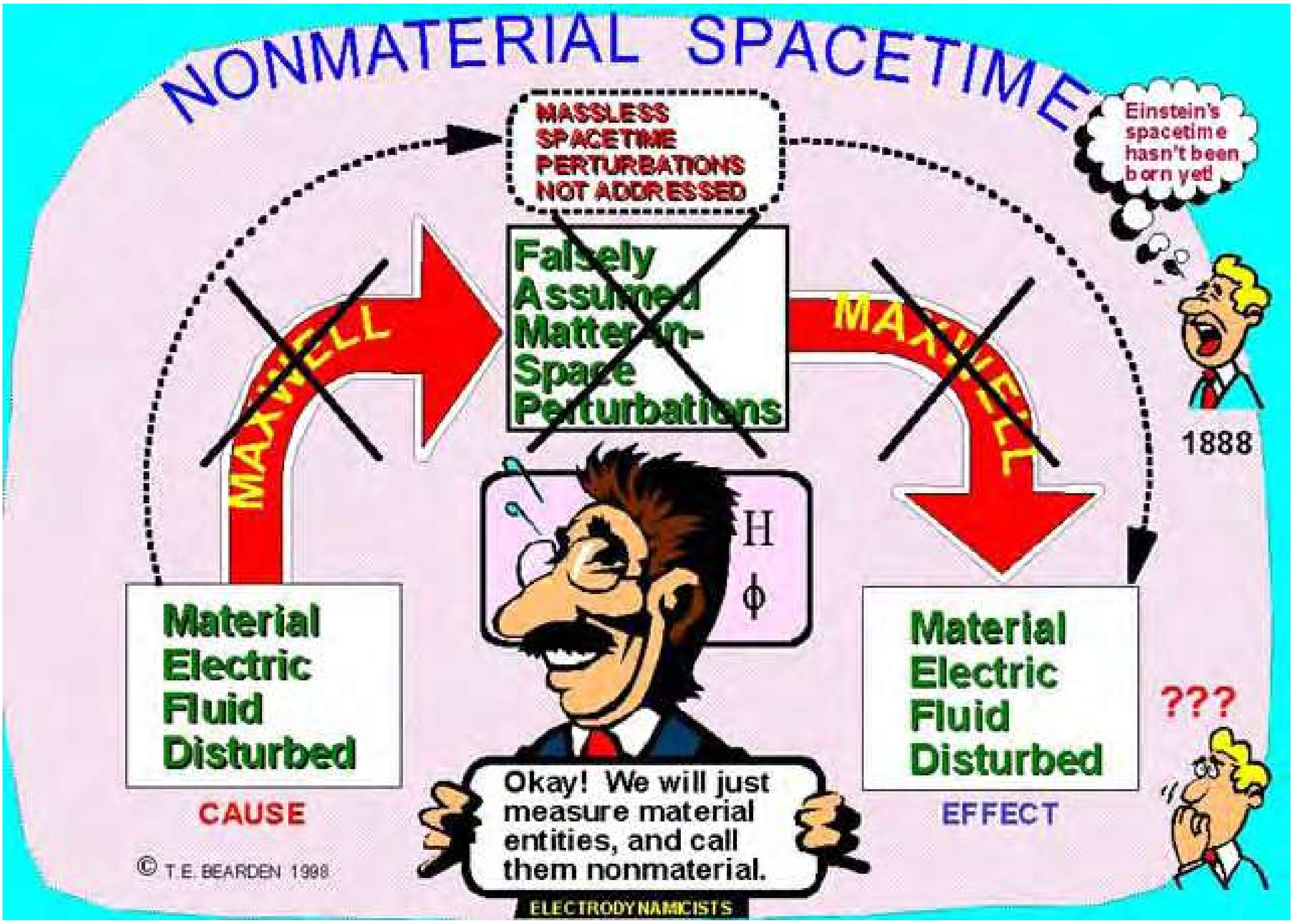
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
MAXWELL
Falsely Assumed Matter-in-Space Perturbations
MAXWELL
1888
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E. BEARDEN 1998

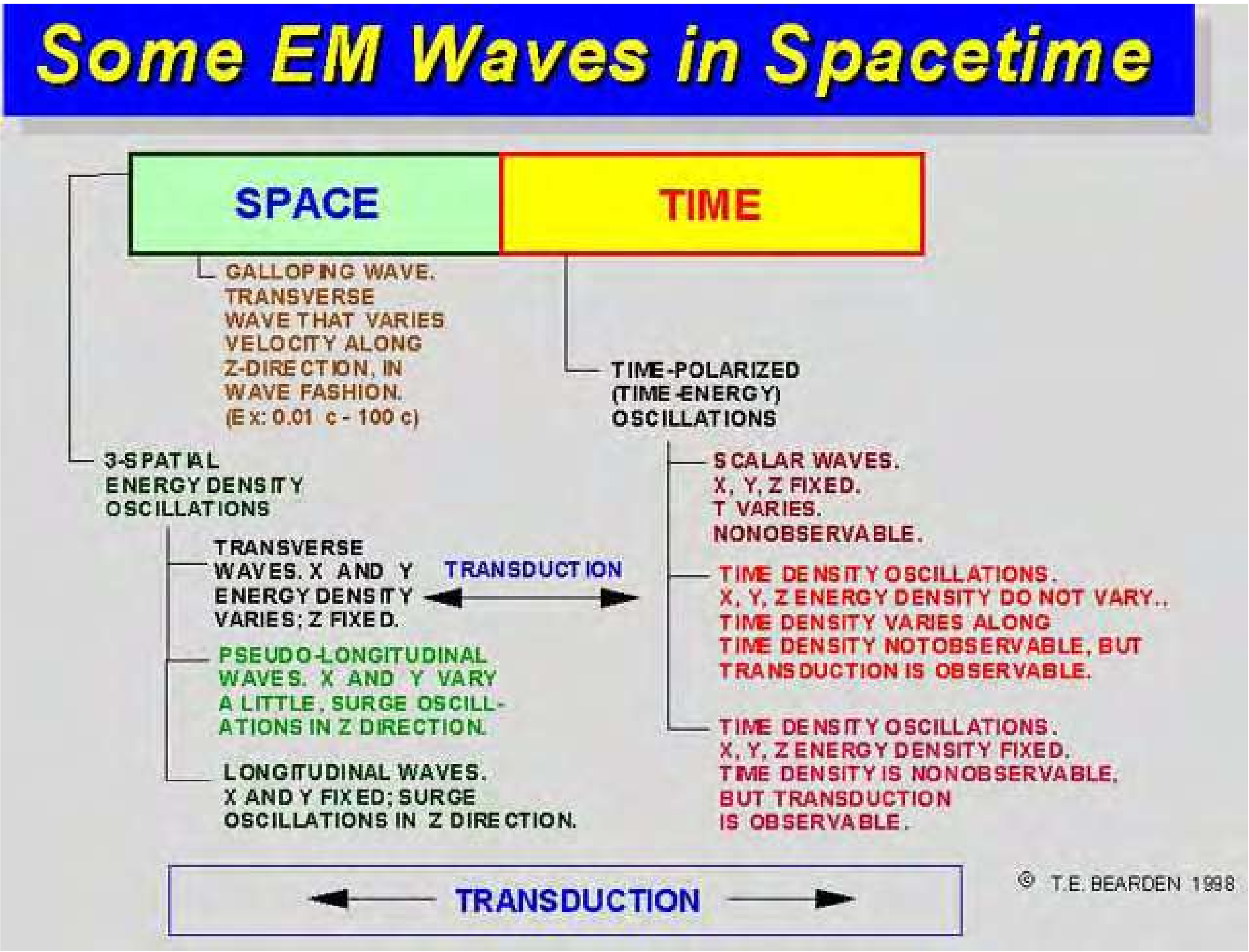

Some EM Waves in Spacetime

SPACE

TIME

GALLOPING WAVE. TRANSVERSE WAVE THAT VARIES VELOCITY ALONG Z-DIRECTION, IN WAVE FASHION. (Ex: 0.01 c - 100 c)

3-SPATIAL ENERGY DENSITY OSCILLATIONS

- TRANSVERSE WAVES. X AND Y ENERGY DENSITY VARIES; Z FIXED.
- PSEUDO-LONGITUDINAL WAVES. X AND Y VARY A LITTLE, SURGE OSCILLATIONS IN Z DIRECTION.
- LONGITUDINAL WAVES. X AND Y FIXED; SURGE OSCILLATIONS IN Z DIRECTION.

TRANSDUCTION

TIME-POLARIZED (TIME-ENERGY) OSCILLATIONS

- SCALAR WAVES. X, Y, Z FIXED. T VARIES. NONOBSERVABLE.
- TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY DO NOT VARY.. TIME DENSITY VARIES ALONG TIME DENSITY NOTOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.
- TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY FIXED. TIME DENSITY IS NONOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.

TRANSDUCTION

© T.E. BEARDEN 1998

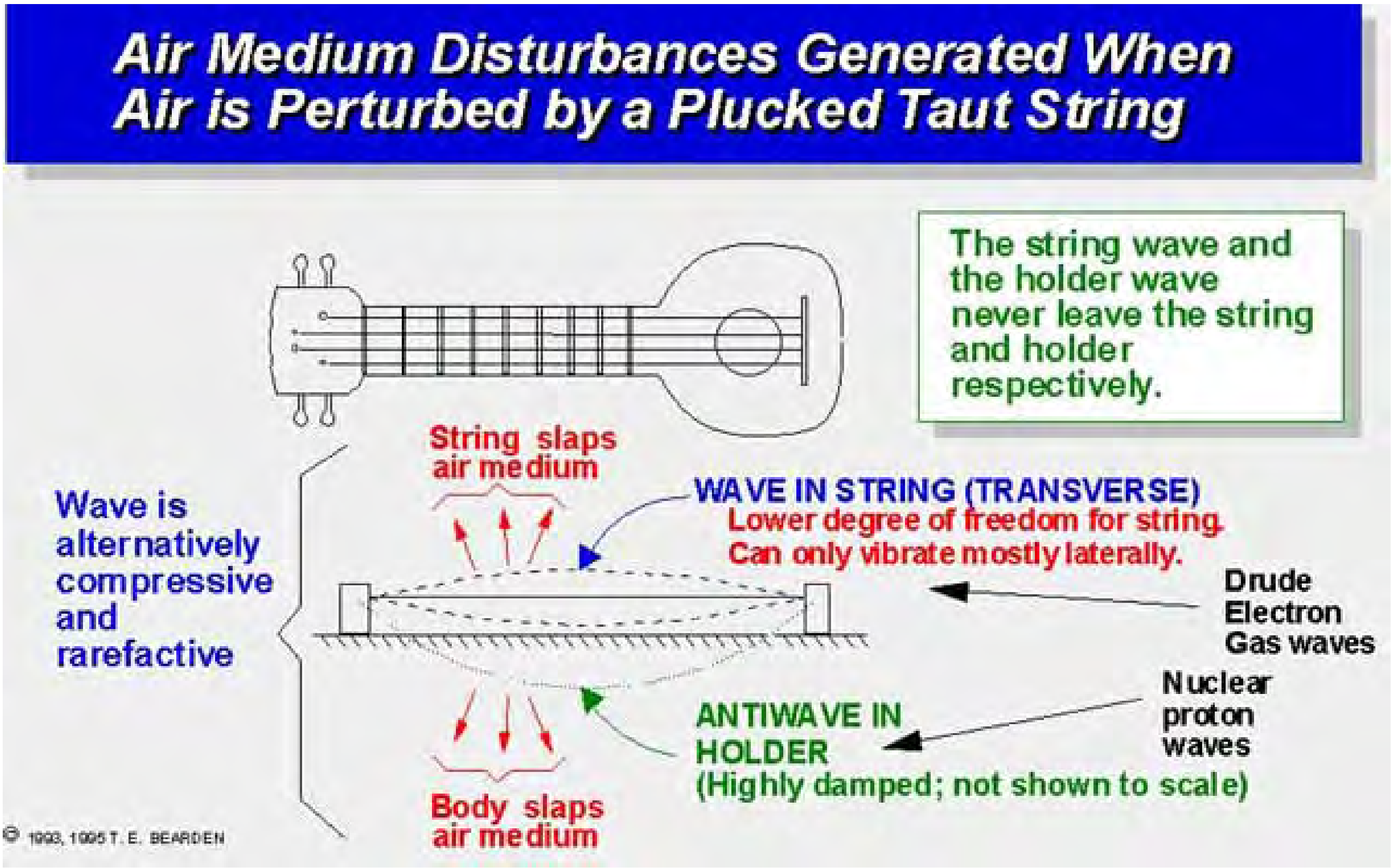

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

String slaps air medium

WAVE IN STRING (TRANSVERSE)
Lower degree of freedom for string.
Can only vibrate mostly laterally.

Drude Electron Gas waves

Wave is alternatively compressive and rarefactive

Nuclear proton waves

ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)

Body slaps air medium

© 1993, 1995 T. E. BEARDEN

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

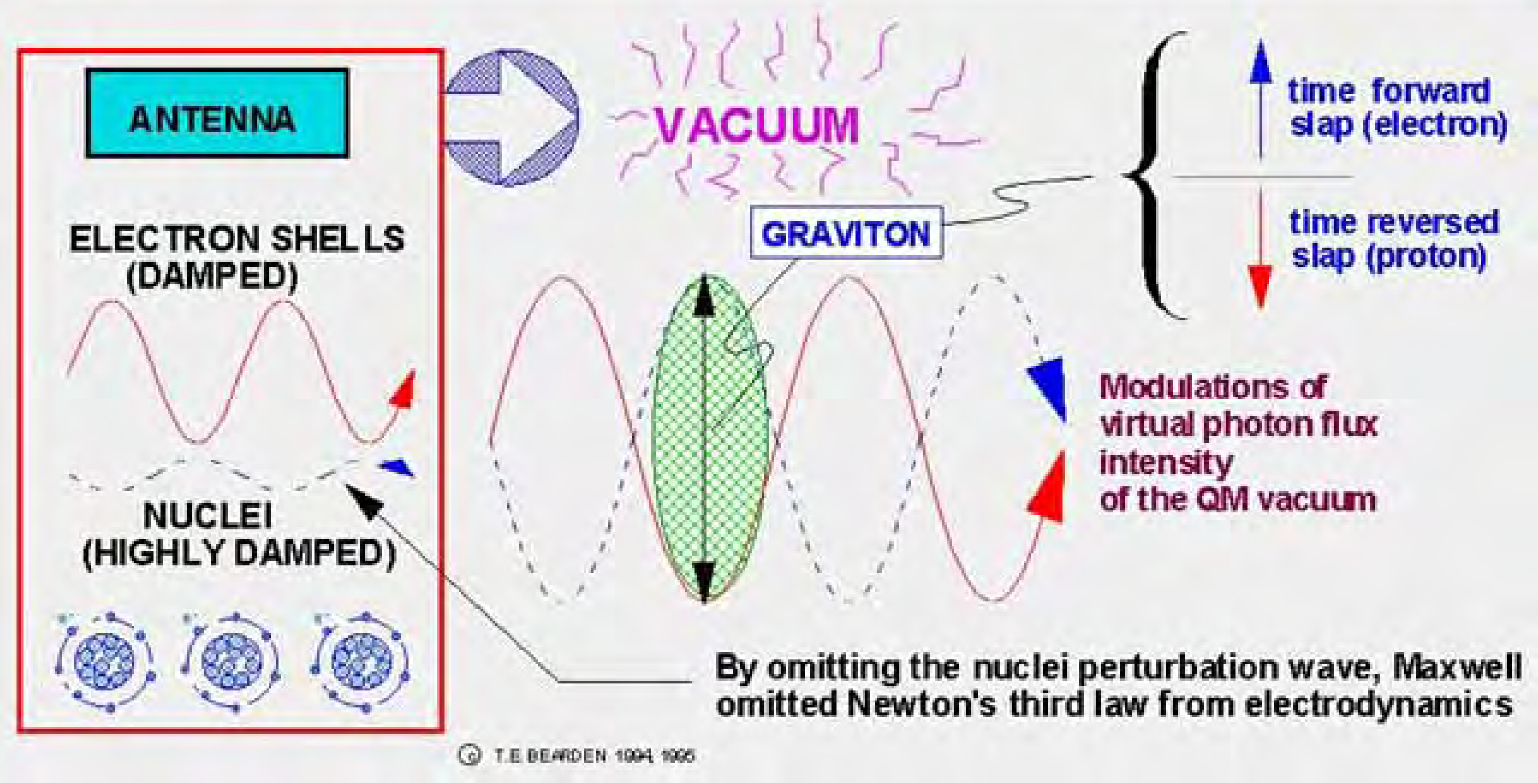
ANTENNA
ELECTRON SHELLS (DAMPED)
NUCLEI (HIGHLY DAMPED)
VACUUM
GRAVITON
time forward slap (electron)
time reversed slap (proton)
Modulations of virtual photon flux intensity of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

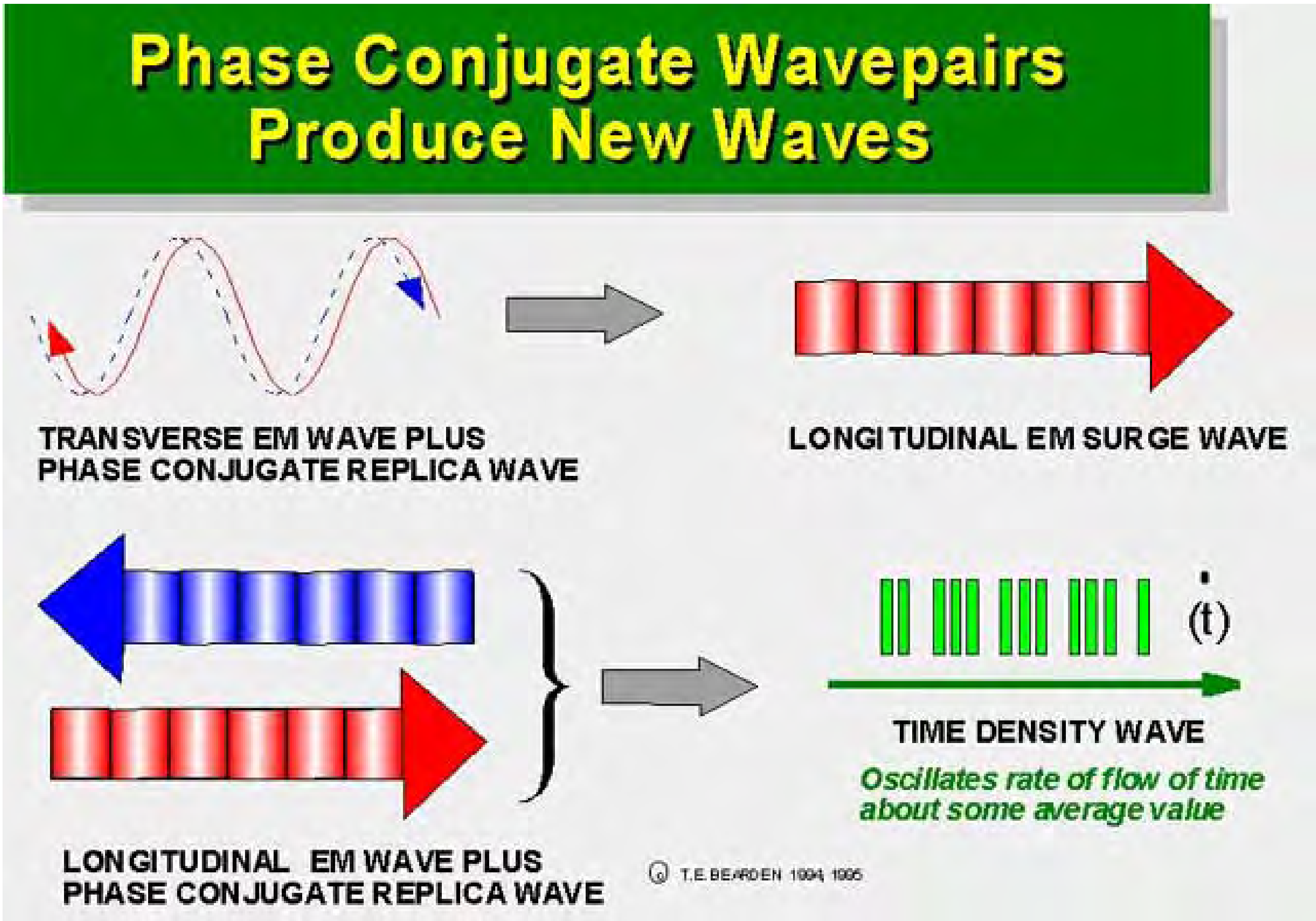

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM SURGE WAVE

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

TIME DENSITY WAVE

Oscillates rate of flow of time about some average value

© T.E. BEARDEN 1994, 1995

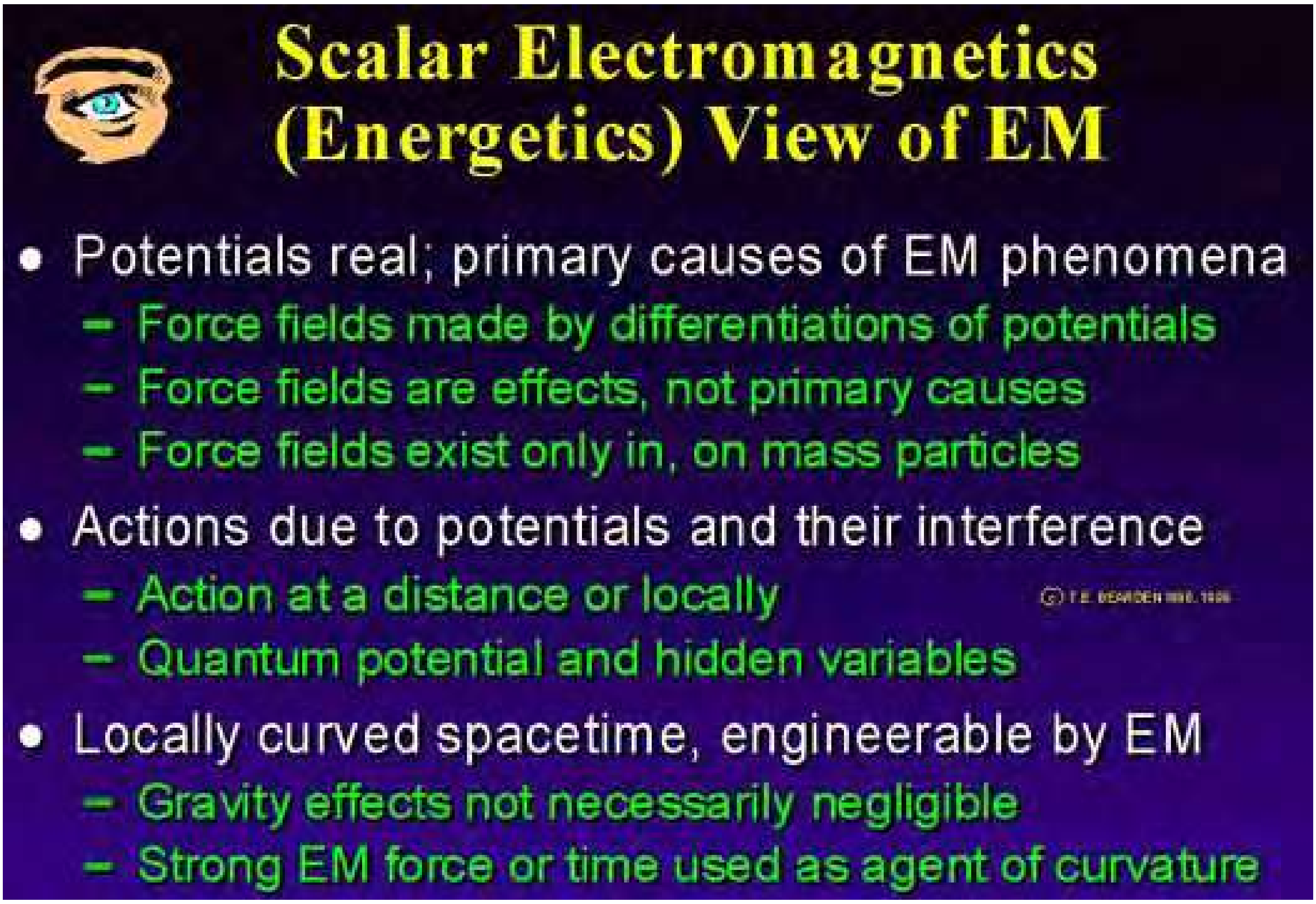

Scalar Electromagnetics (Energetics) View of EM

- Potentials real; primary causes of EM phenomena
  - Force fields made by differentiations of potentials
  - Force fields are effects, not primary causes
  - Force fields exist only in, on mass particles
- Actions due to potentials and their interference
  - Action at a distance or locally
  - Quantum potential and hidden variables
- Locally curved spacetime, engineerable by EM
  - Gravity effects not necessarily negligible
  - Strong EM force or time used as agent of curvature

© T.E. BEARDEN [illegible]

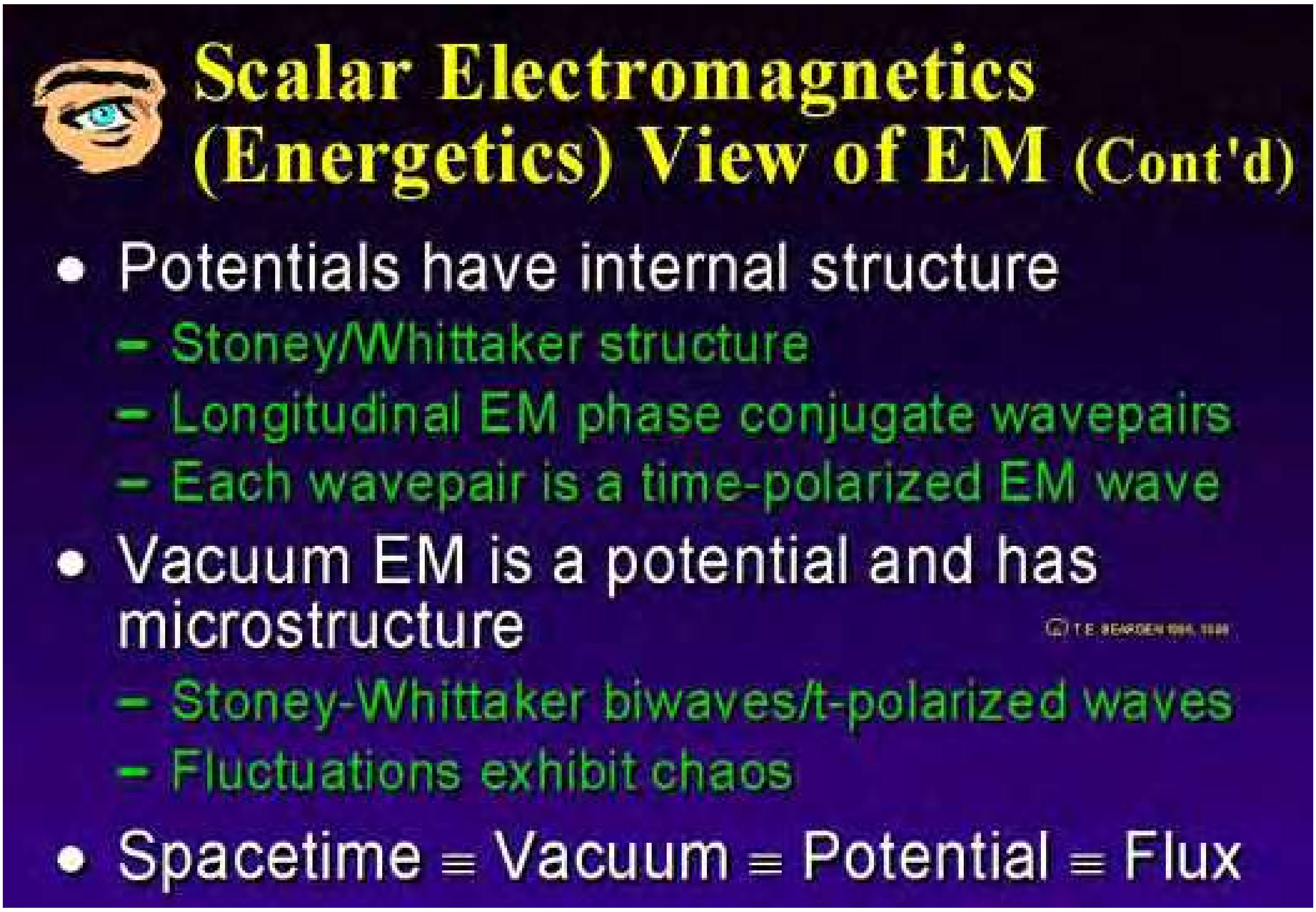

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Potentials have internal structure
  - Stoney/Whittaker structure
  - Longitudinal EM phase conjugate wavepairs
  - Each wavepair is a time-polarized EM wave
- Vacuum EM is a potential and has microstructure
  - Stoney-Whittaker biwaves/t-polarized waves
  - Fluctuations exhibit chaos
- Spacetime ≡ Vacuum ≡ Potential ≡ Flux

© T.E. BEARDEN [illegible]

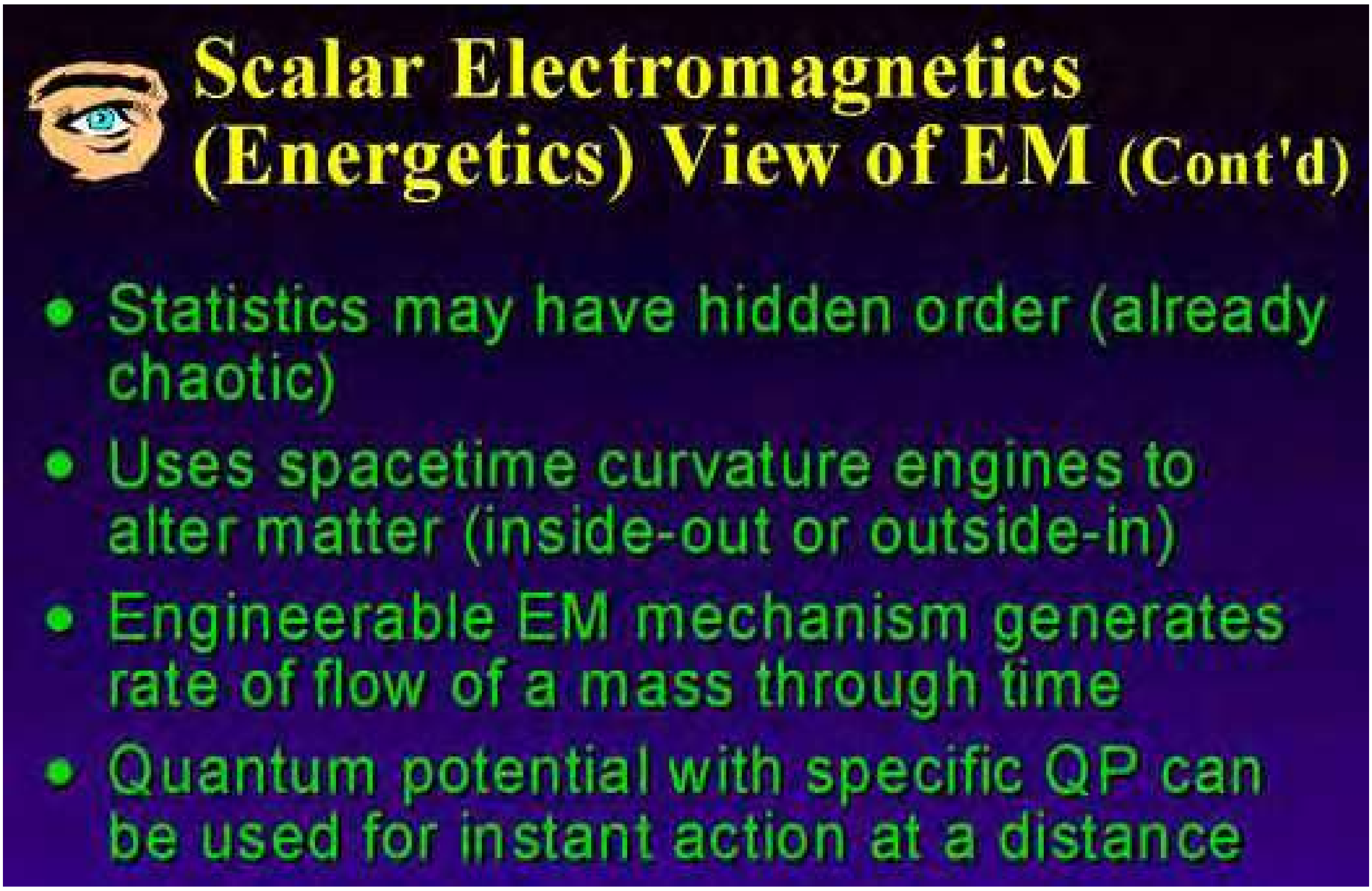

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Statistics may have hidden order (already chaotic)
- Uses spacetime curvature engines to alter matter (inside-out or outside-in)
- Engineerable EM mechanism generates rate of flow of a mass through time
- Quantum potential with specific QP can be used for instant action at a distance

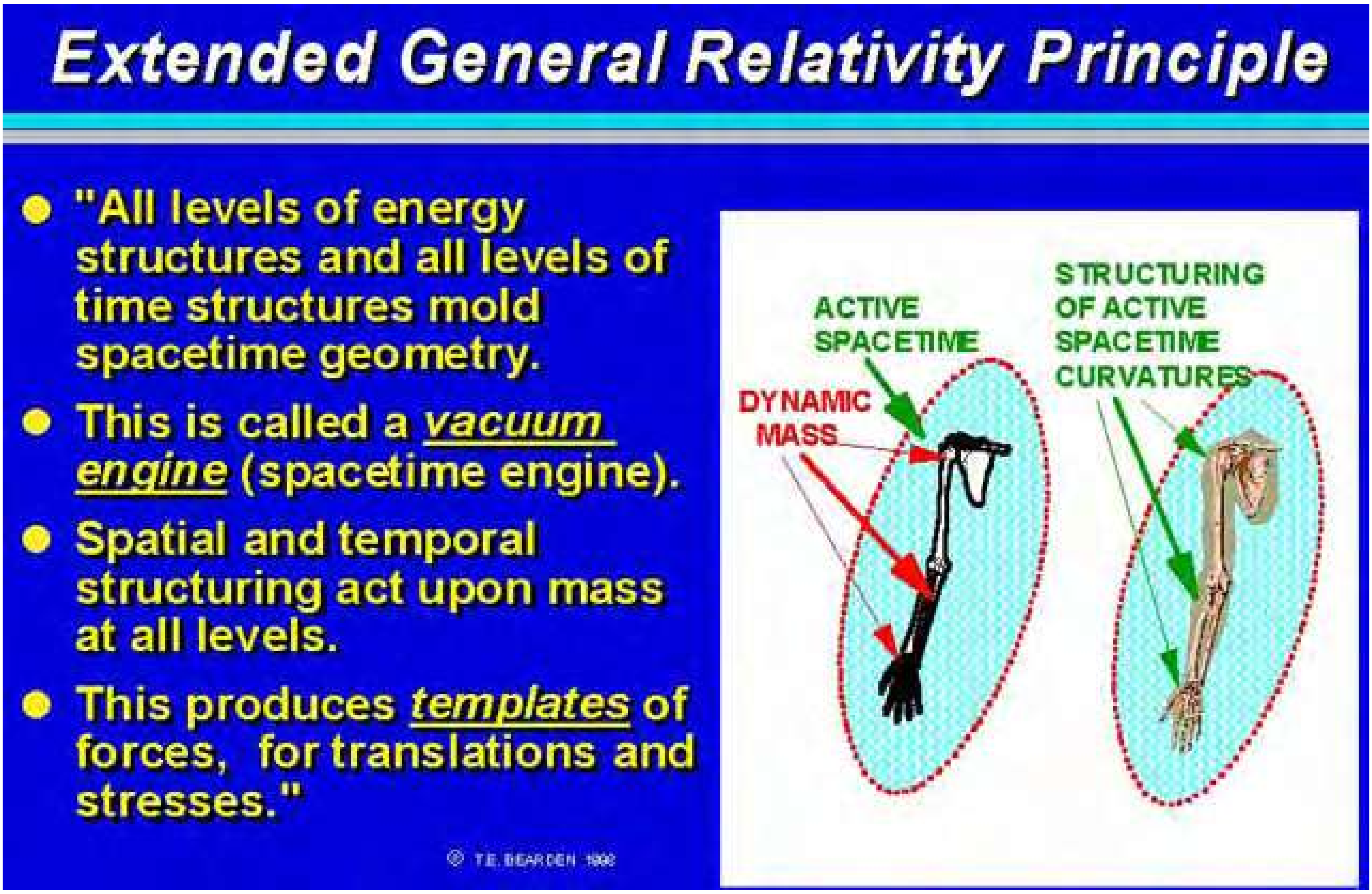

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a *vacuum engine* (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces *templates* of forces, for translations and stresses."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E. BEARDEN 1990

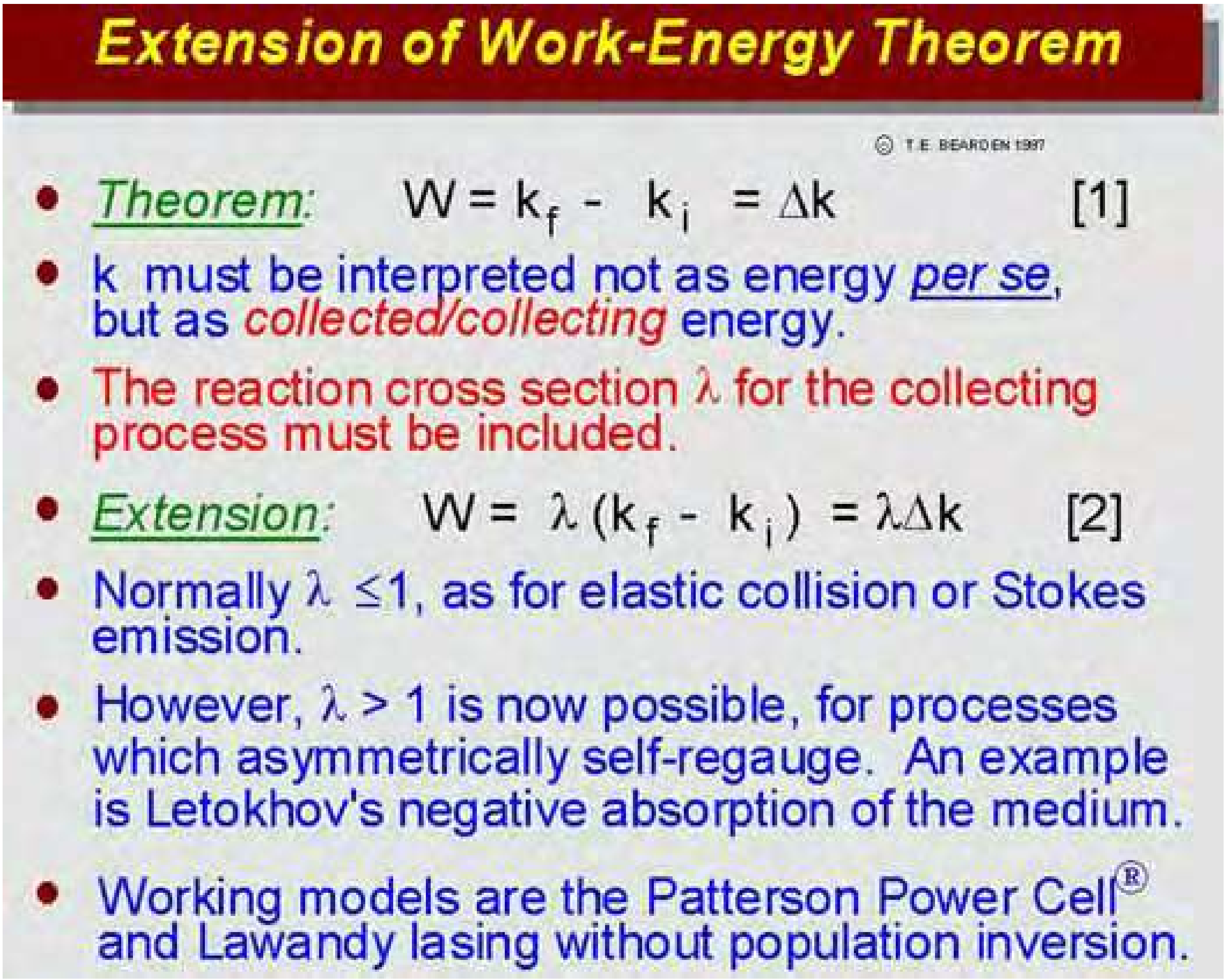

Extension of Work-Energy Theorem

© T.E. BEARDEN 1997

- *Theorem:* $W = k_f - k_i = \Delta k$ [1]
- k must be interpreted not as energy *per se*, but as *collected/collecting* energy.
- The reaction cross section λ for the collecting process must be included.
- *Extension:* $W = \lambda (k_f - k_i) = \lambda \Delta k$ [2]
- Normally $\lambda \leq 1$, as for elastic collision or Stokes emission.
- However, $\lambda > 1$ is now possible, for processes which asymmetrically self-regauge. An example is Letokhov's negative absorption of the medium.
- Working models are the Patterson Power Cell® and Lawandy lasing without population inversion.

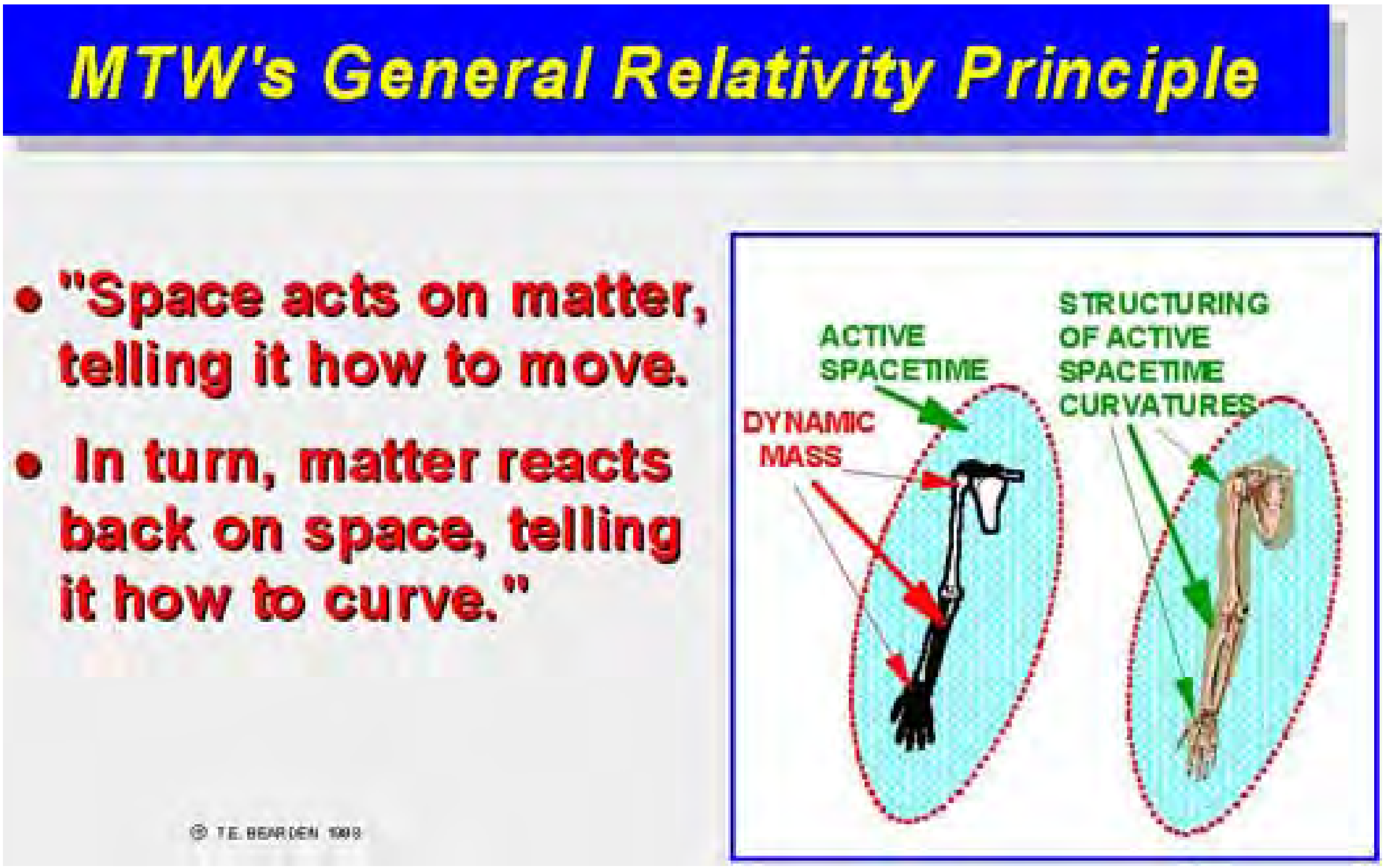

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE SPACETIME
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DYNAMIC MASS

© T.E. BEARDEN 1998

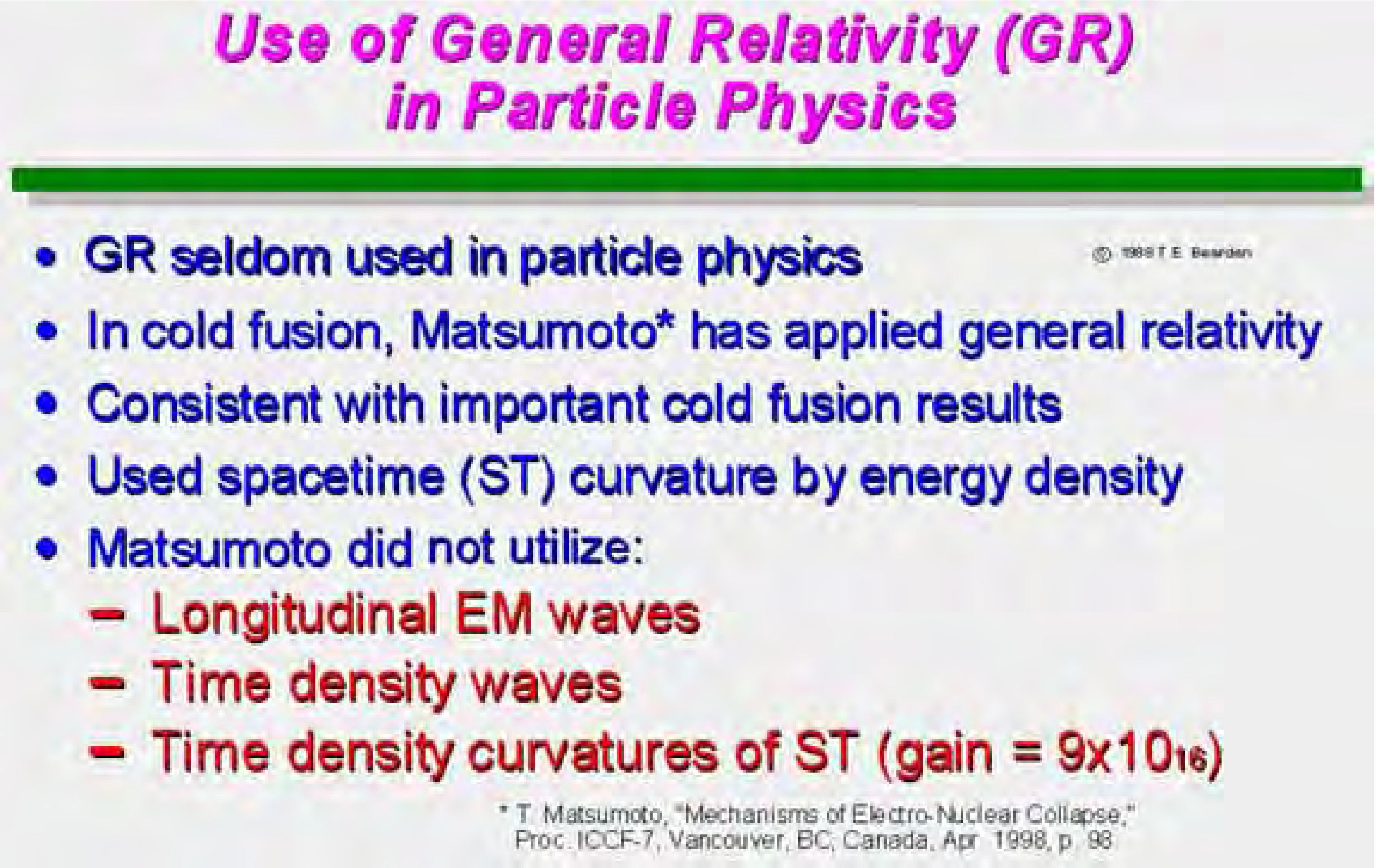

Use of General Relativity (GR) in Particle Physics

© 1998 T.E. Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density waves
  - Time density curvatures of ST (gain = $9x10_{16}$)

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98

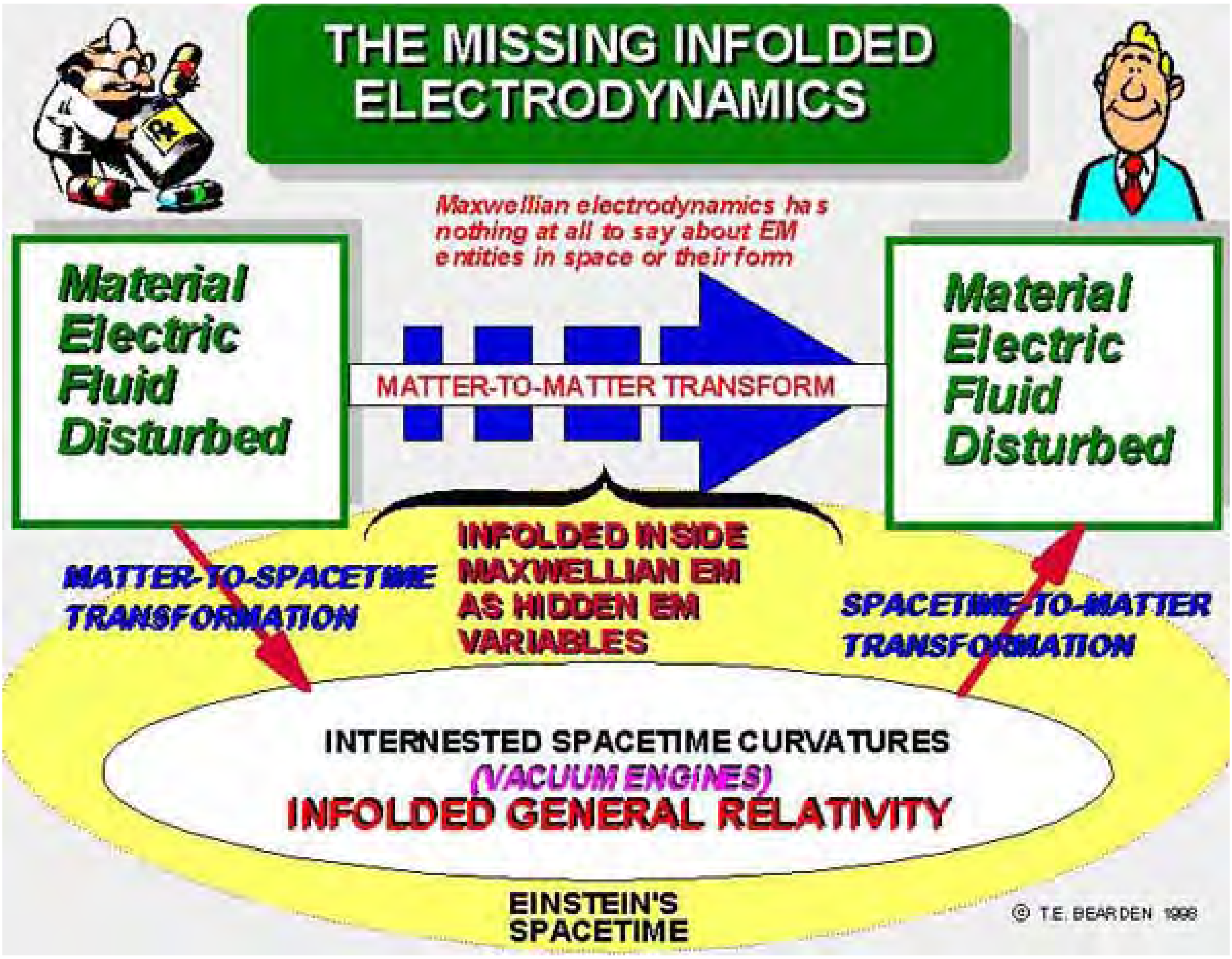

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

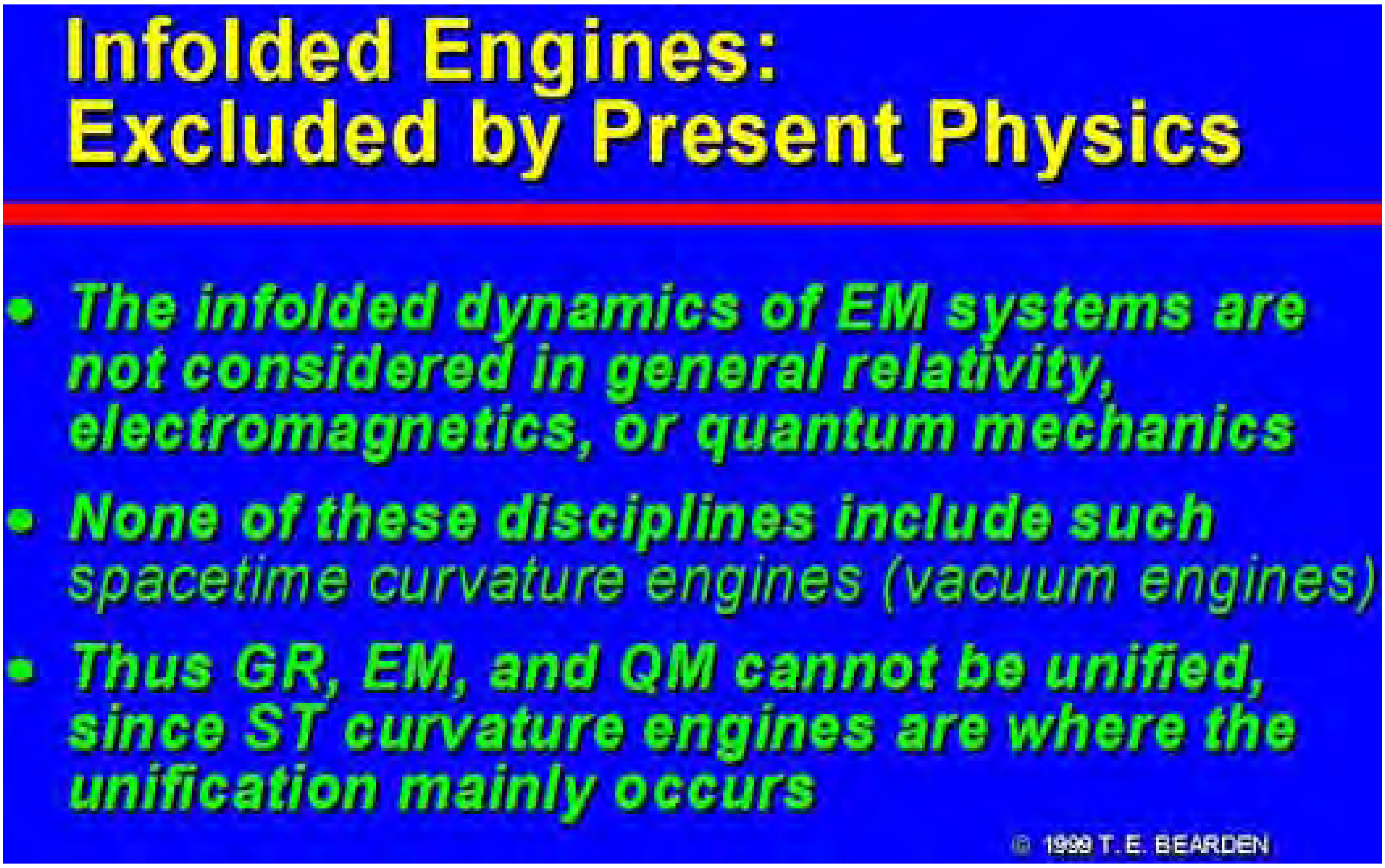

Infolded Engines: Excluded by Present Physics

- *The infolded dynamics of EM systems are not considered in general relativity, electromagnetics, or quantum mechanics*
- *None of these disciplines include such* *spacetime curvature engines (vacuum engines)*
- *Thus GR, EM, and QM cannot be unified, since ST curvature engines are where the unification mainly occurs*

© 1999 T. E. BEARDEN

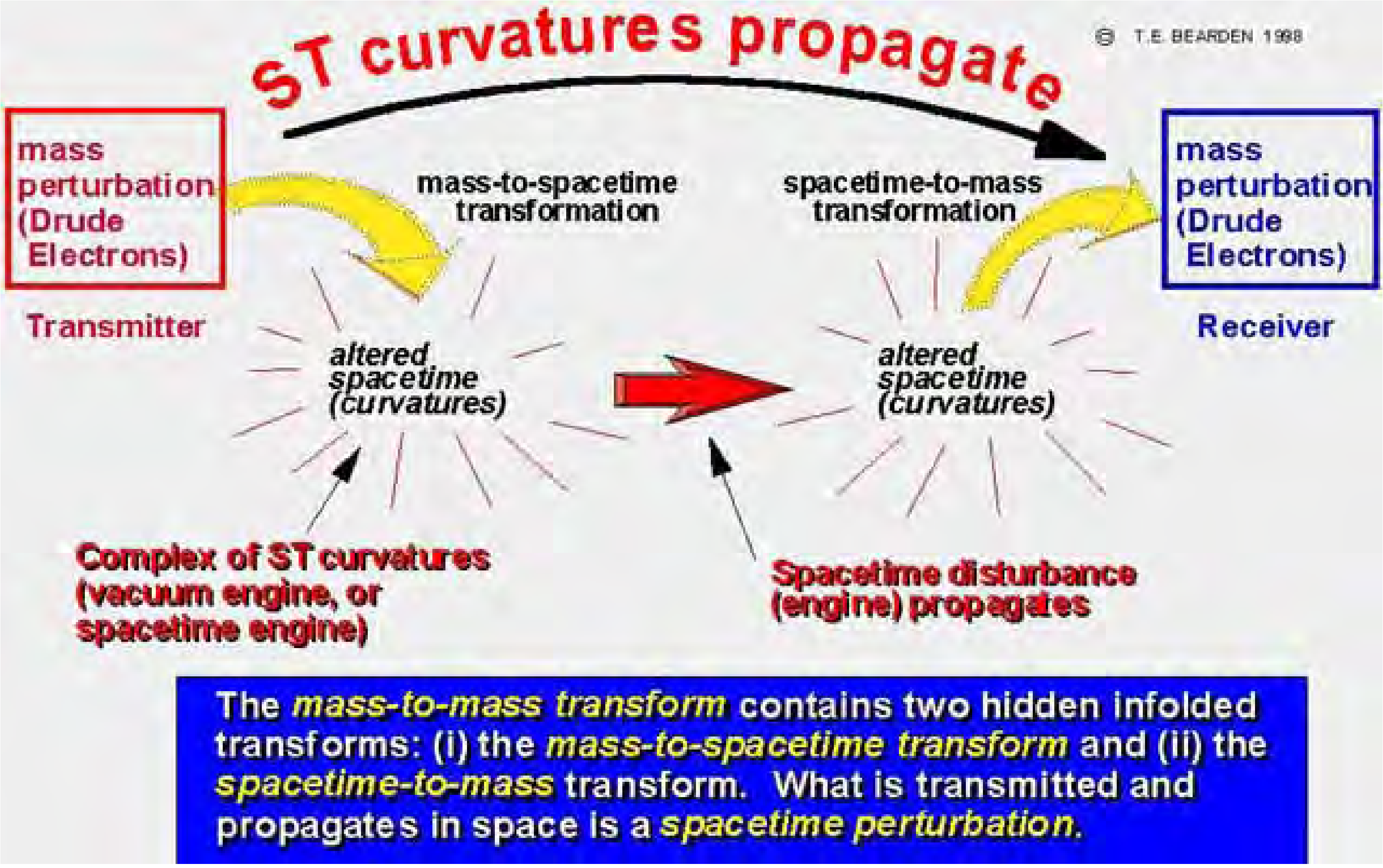

ST curvatures propagate

© T.E. BEARDEN 1998

mass perturbation (Drude Electrons)

Transmitter

mass-to-spacetime transformation

altered spacetime (curvatures)

spacetime-to-mass transformation

altered spacetime (curvatures)

mass perturbation (Drude Electrons)

Receiver

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation.*

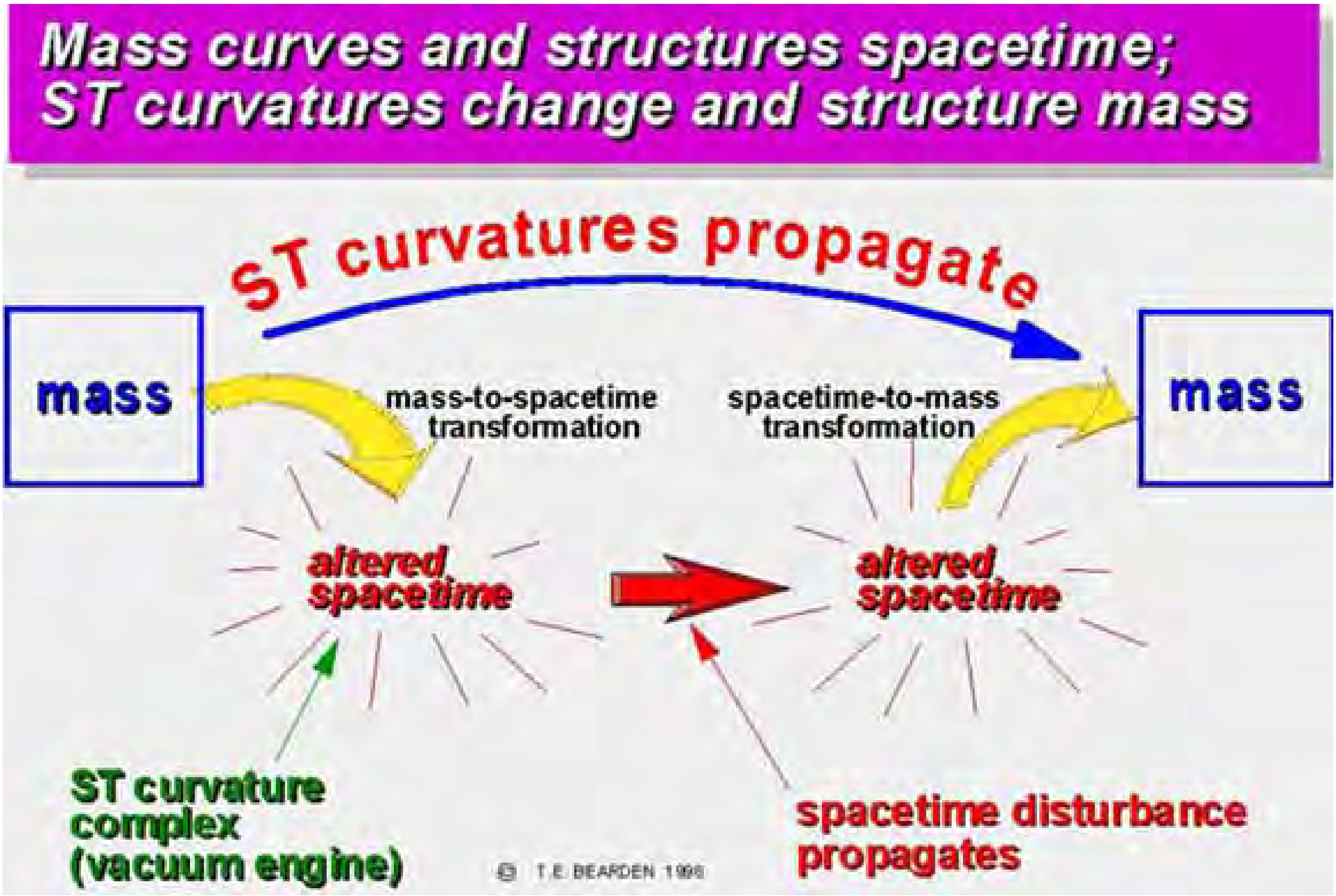

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate

mass

mass-to-spacetime transformation

altered spacetime

spacetime-to-mass transformation

altered spacetime

mass

ST curvature complex (vacuum engine)

spacetime disturbance propagates

© T.E. BEARDEN 1996

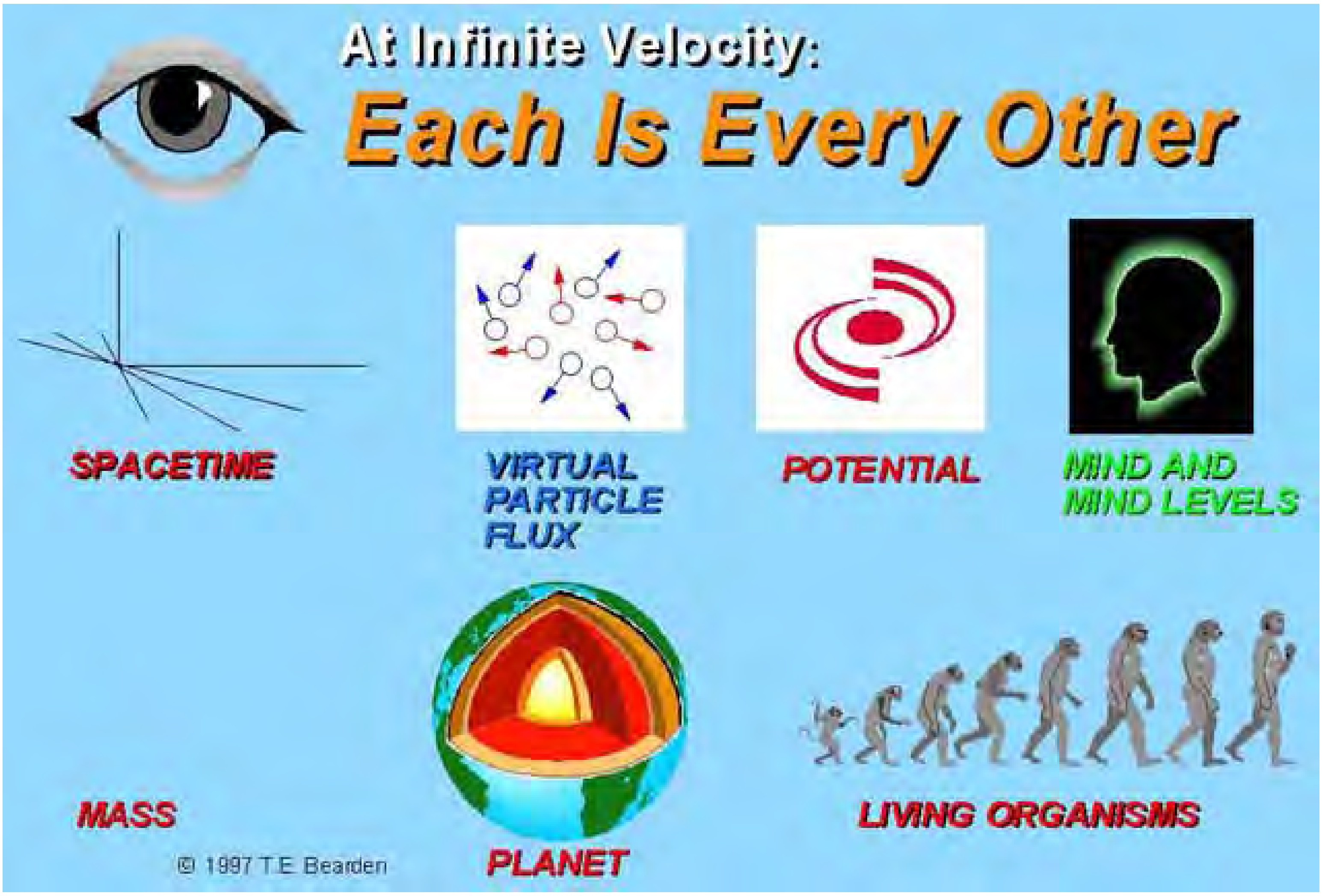

At Infinite Velocity:

Each Is Every Other

SPACETIME

VIRTUAL PARTICLE FLUX

POTENTIAL

MIND AND MIND LEVELS

MASS

PLANET

LIVING ORGANISMS

© 1997 T.E. Bearden

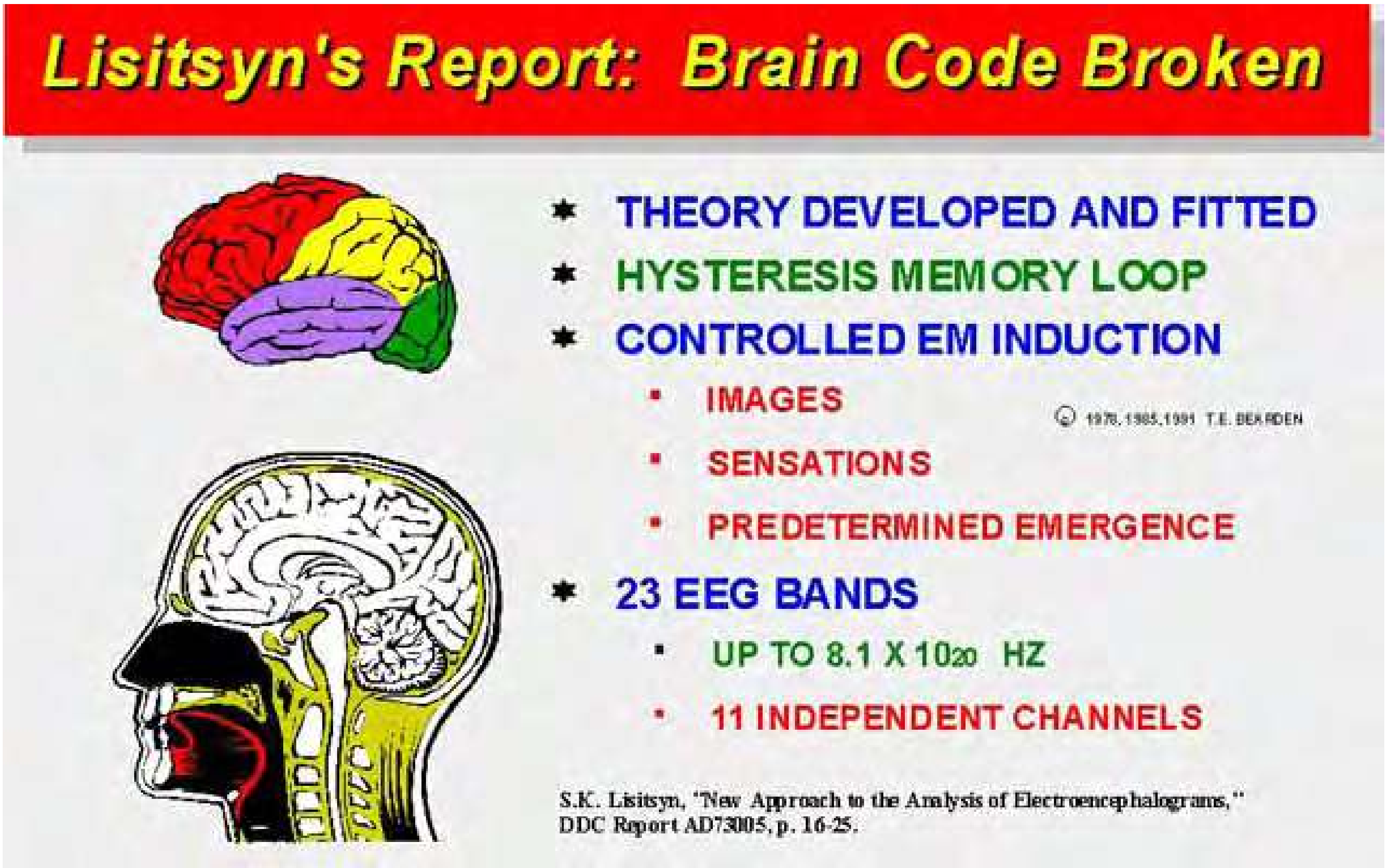

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

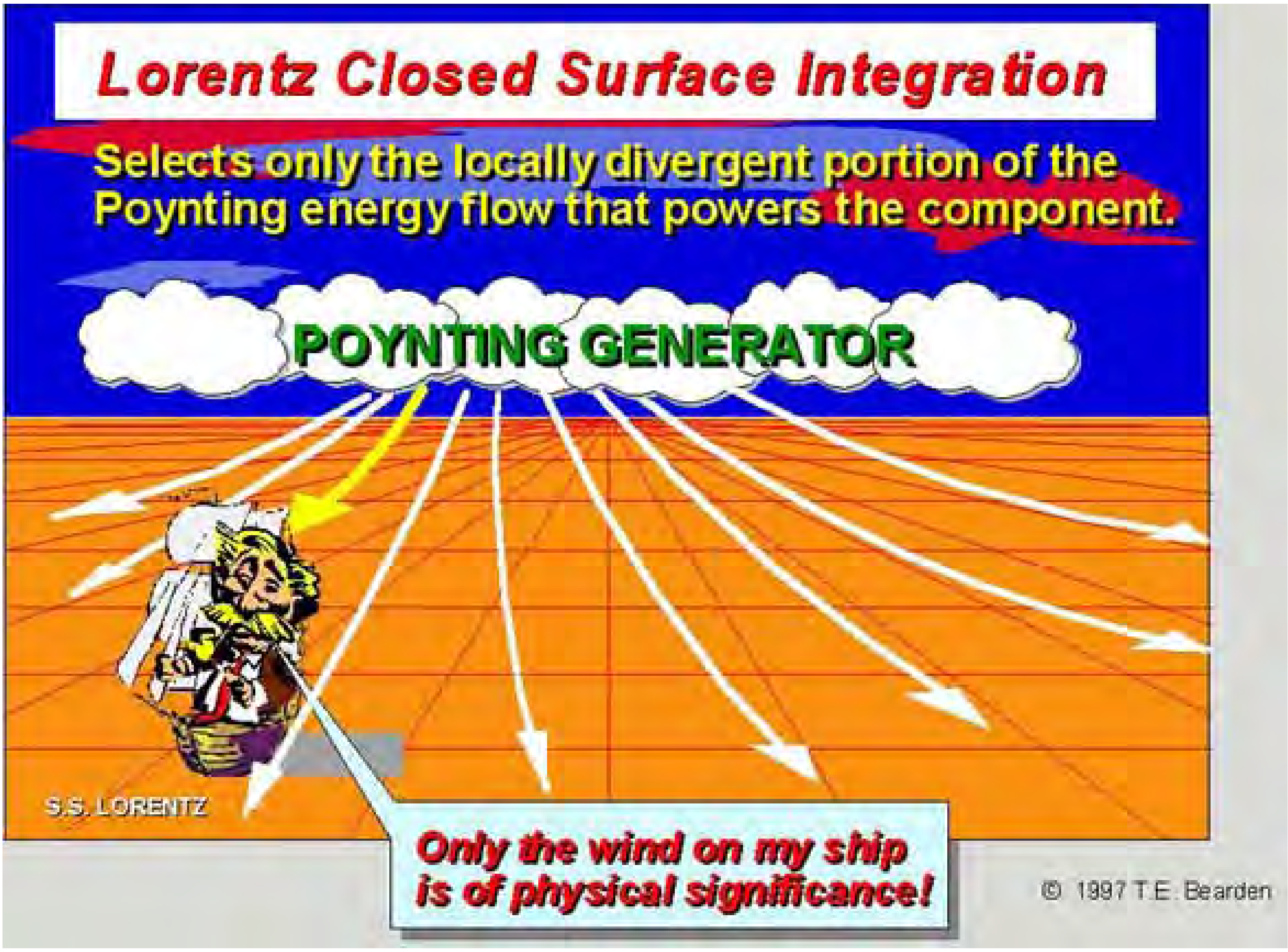

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

POYNTING GENERATOR

S.S. LORENTZ

Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

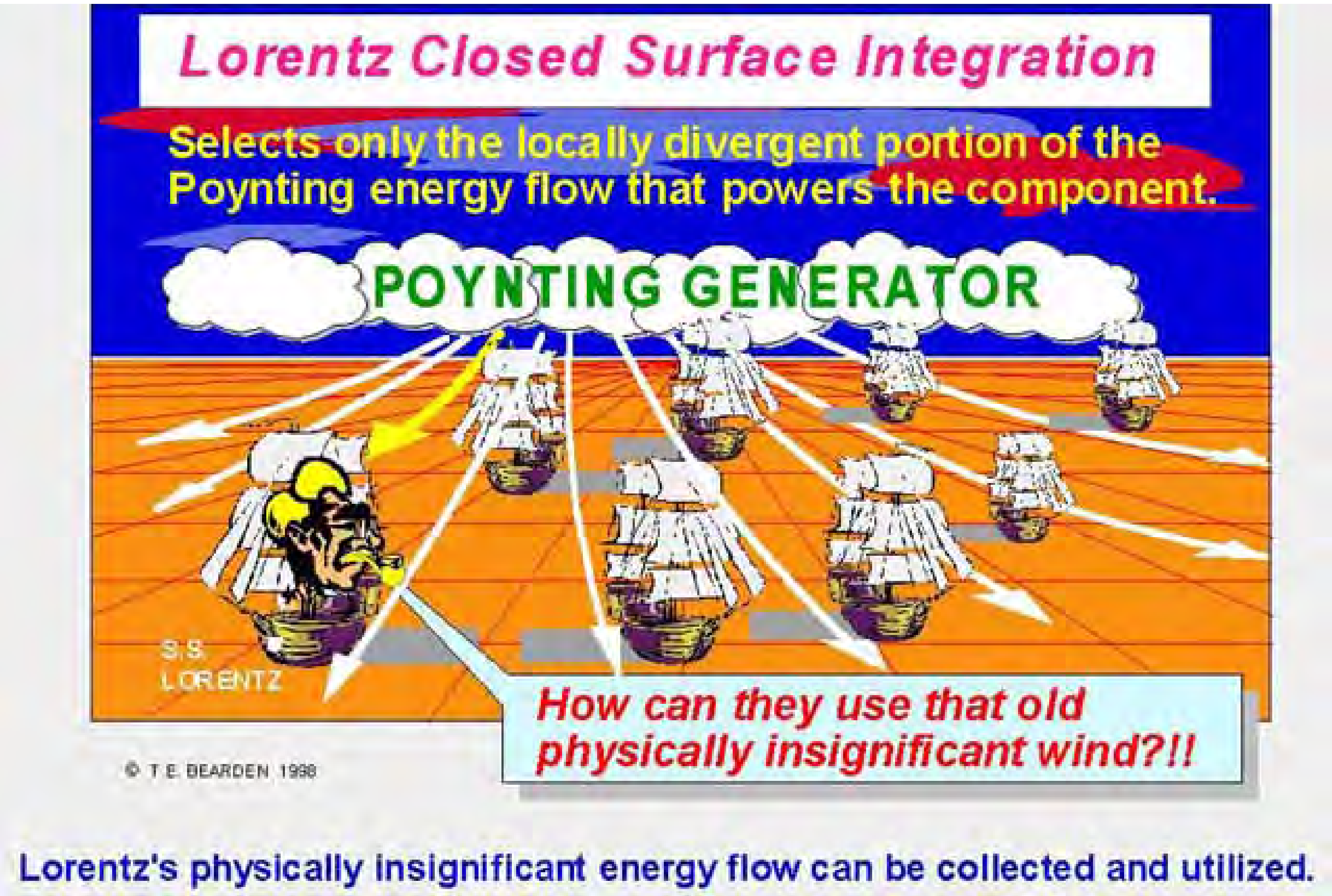

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

How can they use that old physically insignificant wind?!!

© T.E. BEARDEN 1998

Lorentz's physically insignificant energy flow can be collected and utilized.

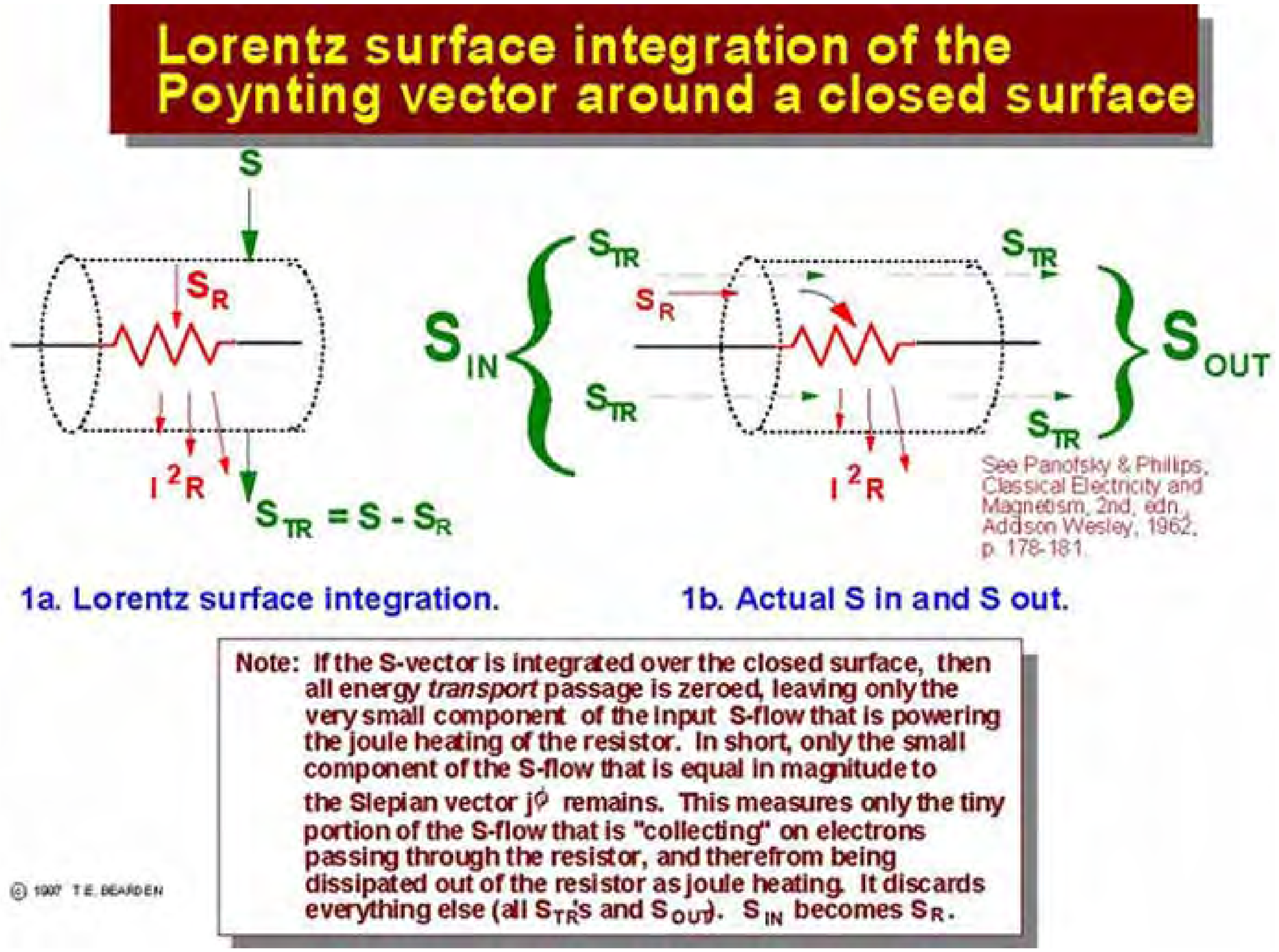

Lorentz surface integration of the Poynting vector around a closed surface

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd, edn., Addison Wesley, 1962, p. 178-181.

1a. Lorentz surface integration.

1b. Actual S in and S out.

Note: If the S-vector is integrated over the closed surface, then all energy *transport* passage is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Slepian vector $j\phi$ remains. This measures only the tiny portion of the S-flow that is "collecting" on electrons passing through the resistor, and therefrom being dissipated out of the resistor as joule heating. It discards everything else (all S_{TR}'s and S_{OUT}). S_{IN} becomes S_R.

© 1997 T.E. BEARDEN

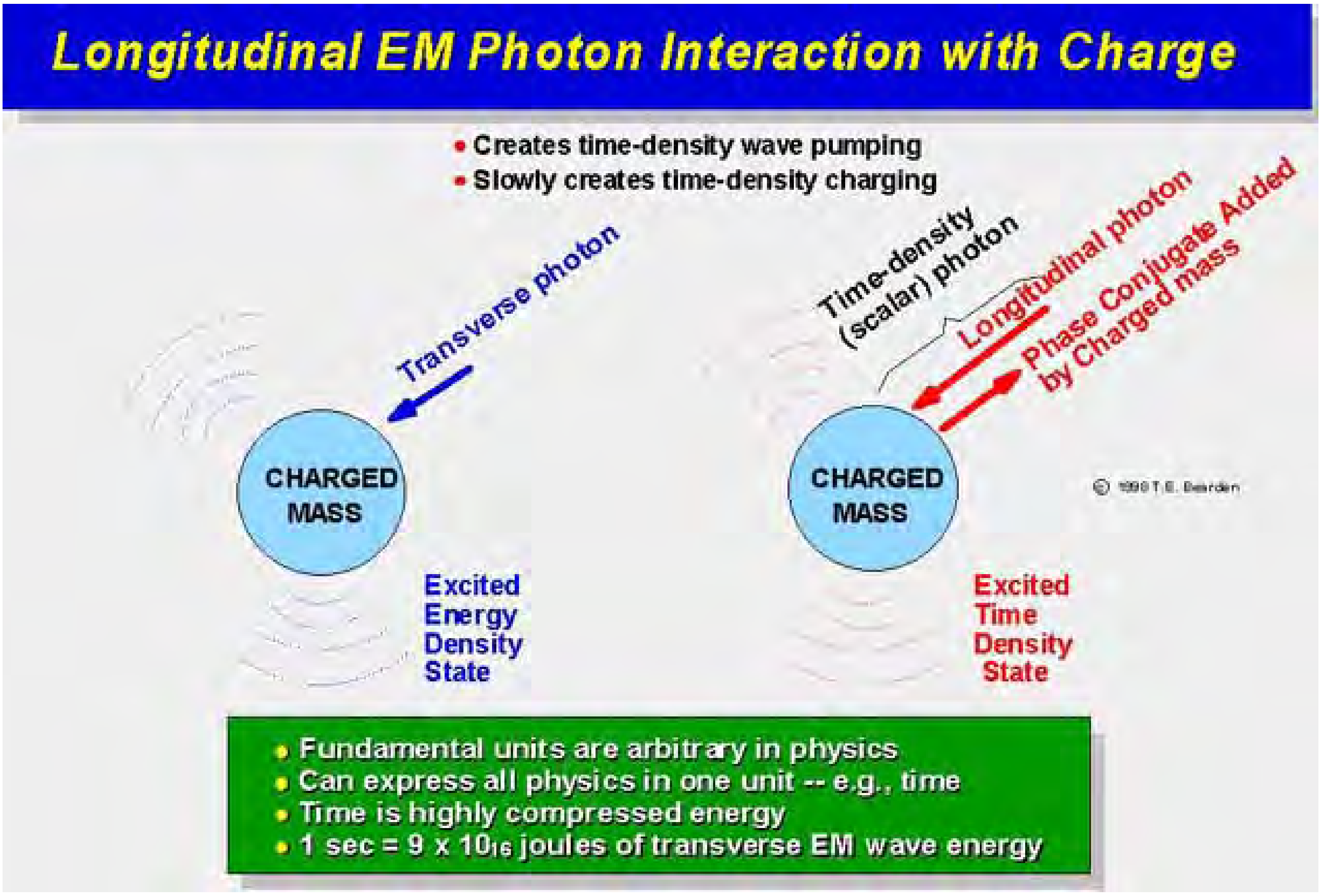

Longitudinal EM Photon Interaction with Charge

- Creates time-density wave pumping
- Slowly creates time-density charging

Transverse photon
CHARGED MASS
Excited Energy Density State
Time-density (scalar) photon
Longitudinal photon
Phase Conjugate Added by Charged mass
CHARGED MASS
Excited Time Density State
© 1990 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9 x 10^{16} joules of transverse EM wave energy

Figure 1-8. Infolded biwave composition of a scalar potential.

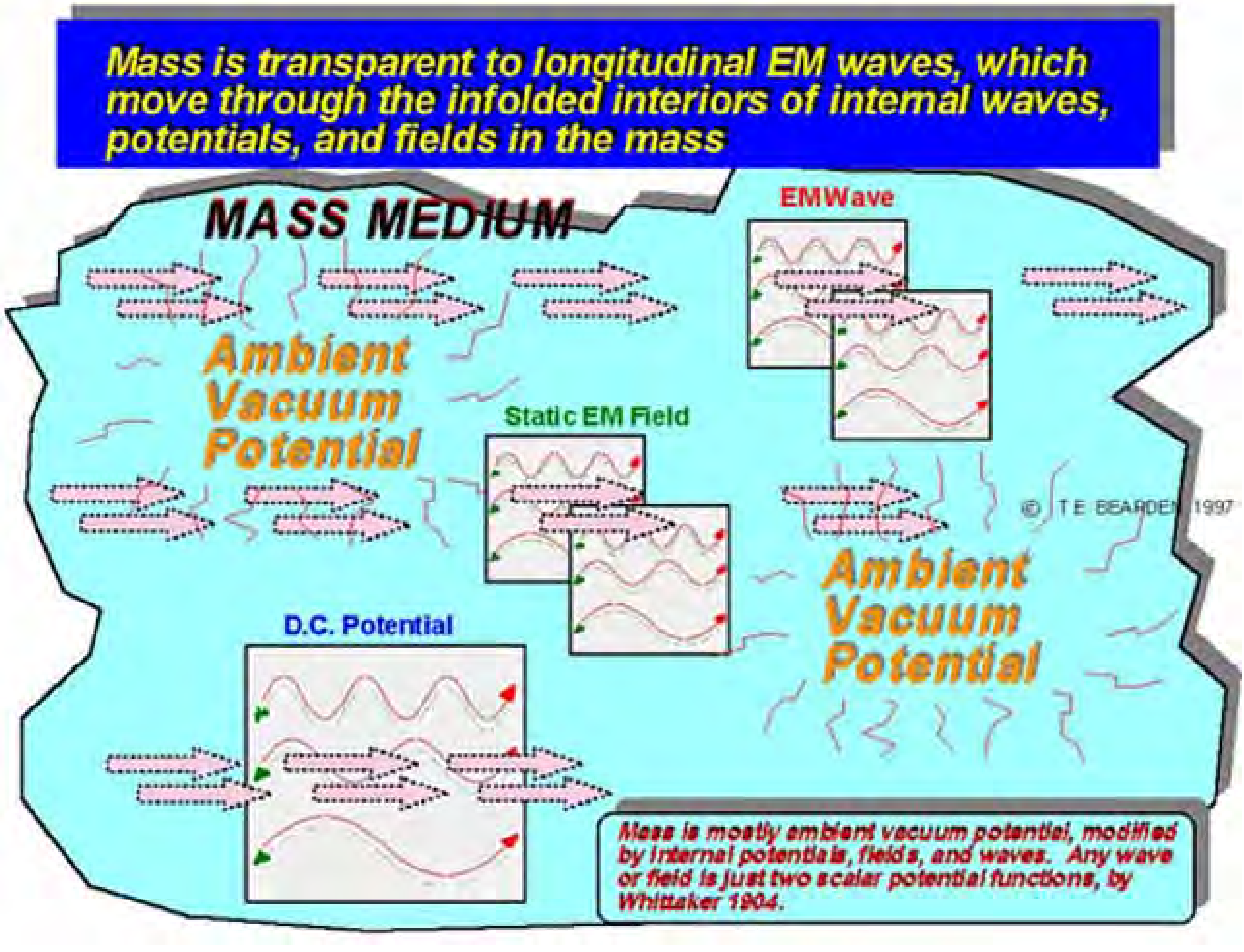

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

MASS MEDIUM
Ambient Vacuum Potential
Static EM Field
EM Wave
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Pumping with Longitudinal EM Waves Time-reverses the Mass

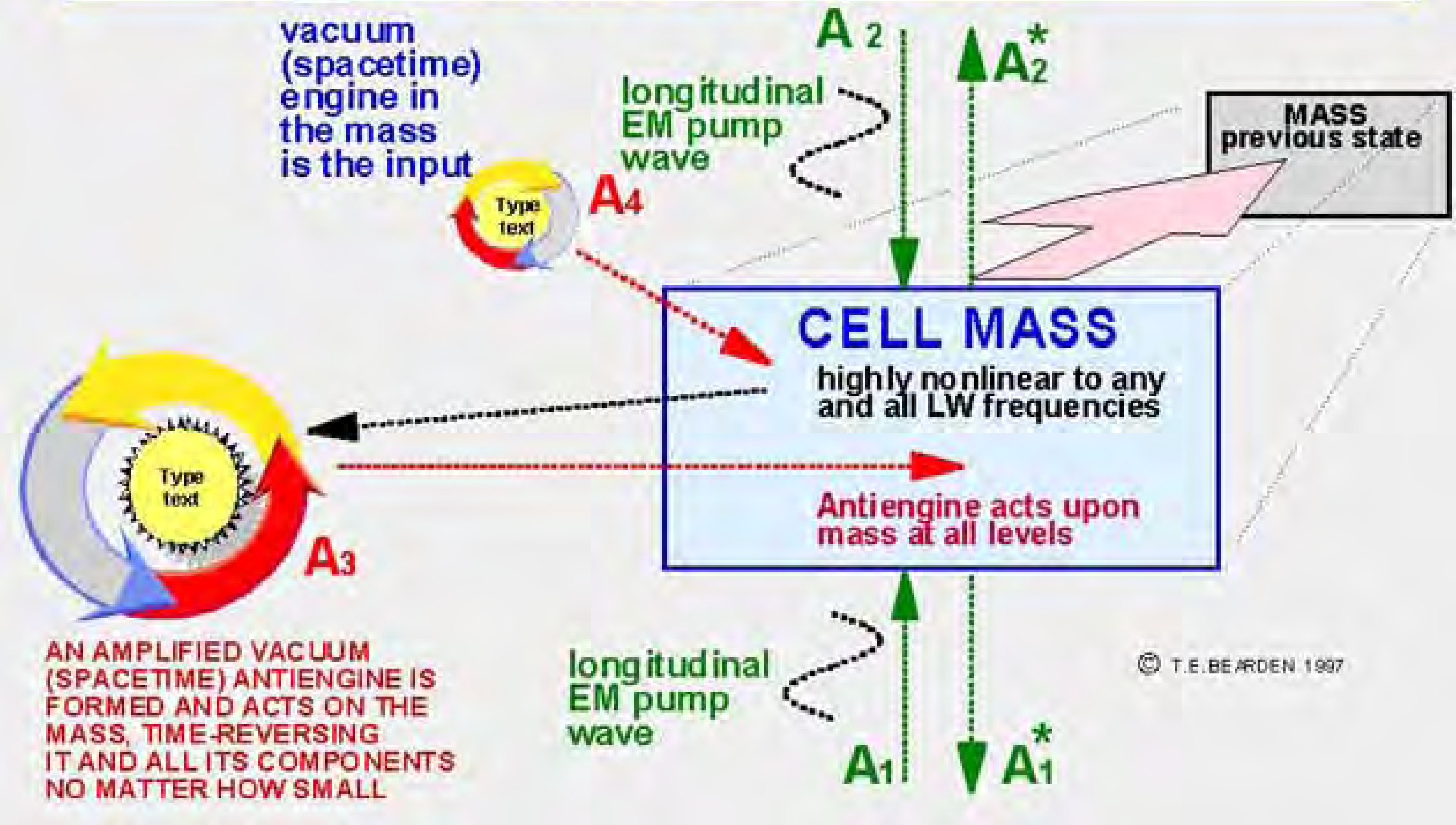

vacuum (spacetime) engine in the mass is the input

longitudinal EM pump wave

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

longitudinal EM pump wave

© T.E.BEARDEN 1997

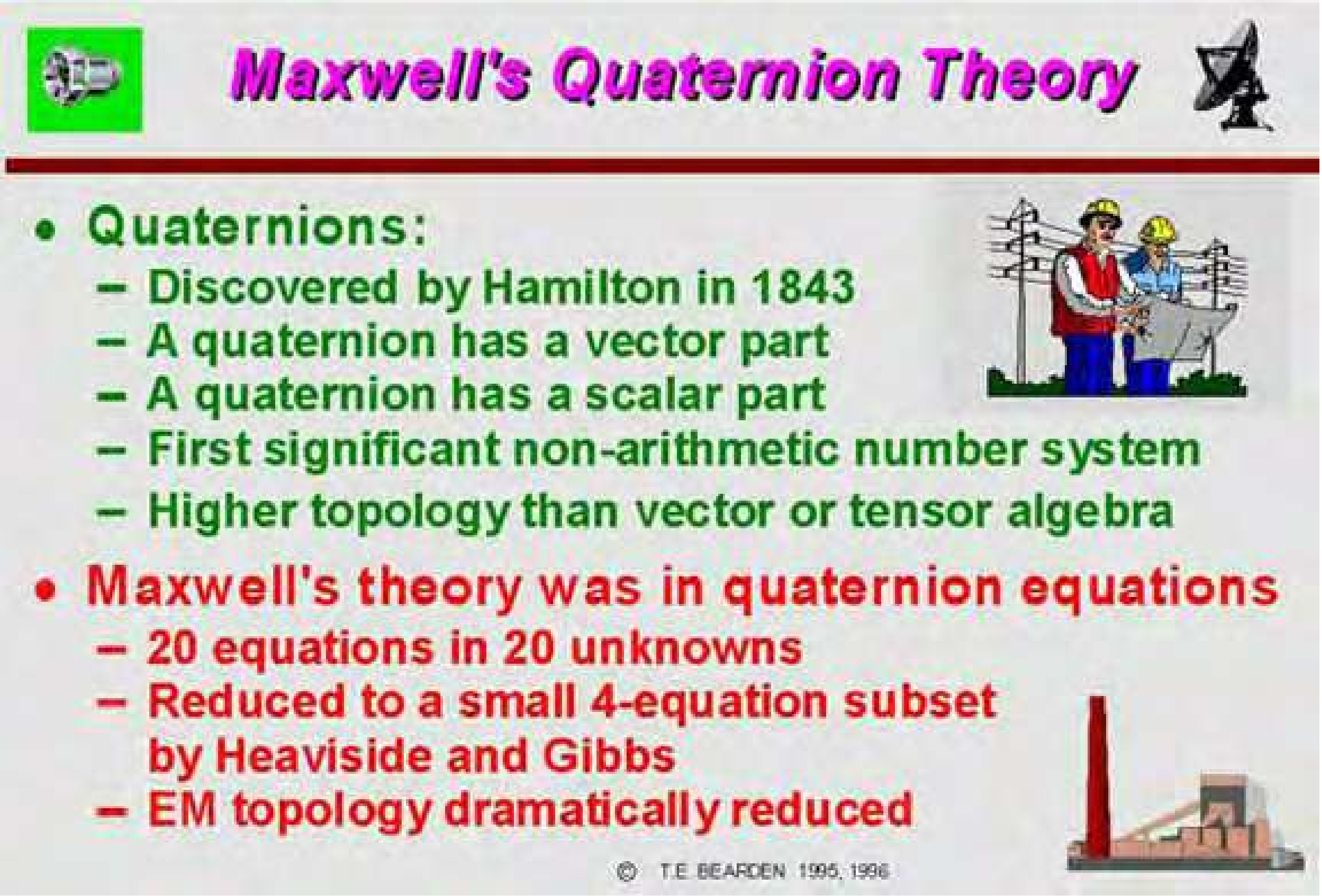

Maxwell's Quaternion Theory

- Quaternions:
  - Discovered by Hamilton in 1843
  - A quaternion has a vector part
  - A quaternion has a scalar part
  - First significant non-arithmetic number system
  - Higher topology than vector or tensor algebra
- Maxwell's theory was in quaternion equations
  - 20 equations in 20 unknowns
  - Reduced to a small 4-equation subset by Heaviside and Gibbs
  - EM topology dramatically reduced

© T.E. BEARDEN 1995, 1996

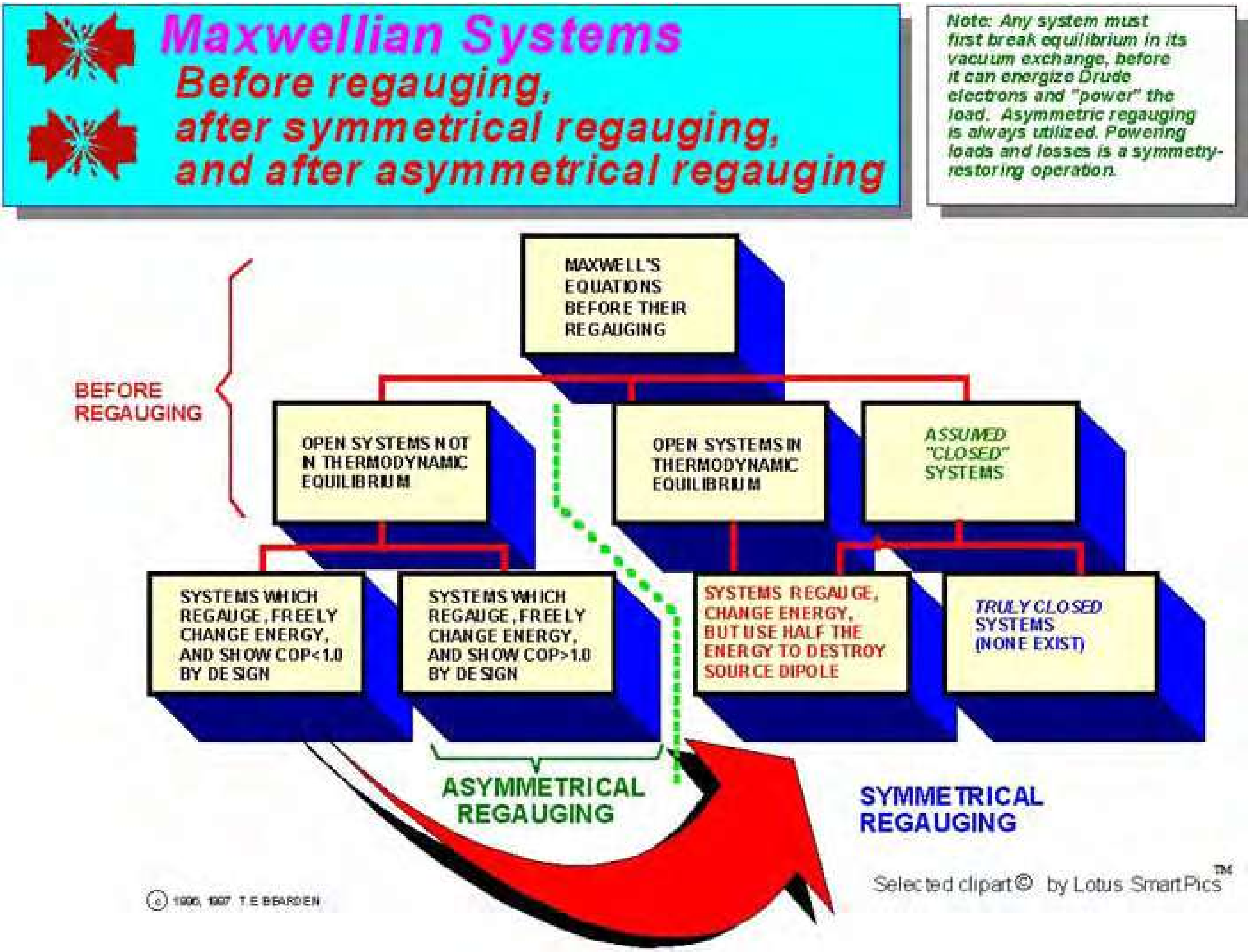

Maxwellian Systems

Before regauging, after symmetrical regauging, and after asymmetrical regauging

Note: Any system must first break equilibrium in its vacuum exchange, before it can energize Drude electrons and "power" the load. Asymmetric regauging is always utilized. Powering loads and losses is a symmetry-restoring operation.

MAXWELL'S EQUATIONS BEFORE THEIR REGAUGING

BEFORE REGAUGING

OPEN SYSTEMS NOT IN THERMODYNAMIC EQUILIBRIUM

OPEN SYSTEMS IN THERMODYNAMIC EQUILIBRIUM

ASSUMED "CLOSED" SYSTEMS

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP<1.0 BY DESIGN

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP>1.0 BY DESIGN

SYSTEMS REGAUGE, CHANGE ENERGY, BUT USE HALF THE ENERGY TO DESTROY SOURCE DIPOLE

TRULY CLOSED SYSTEMS (NONE EXIST)

ASYMMETRICAL REGAUGING

SYMMETRICAL REGAUGING

Selected clipart© by Lotus SmartPics™

© 1996, 1997 T.E BEARDEN

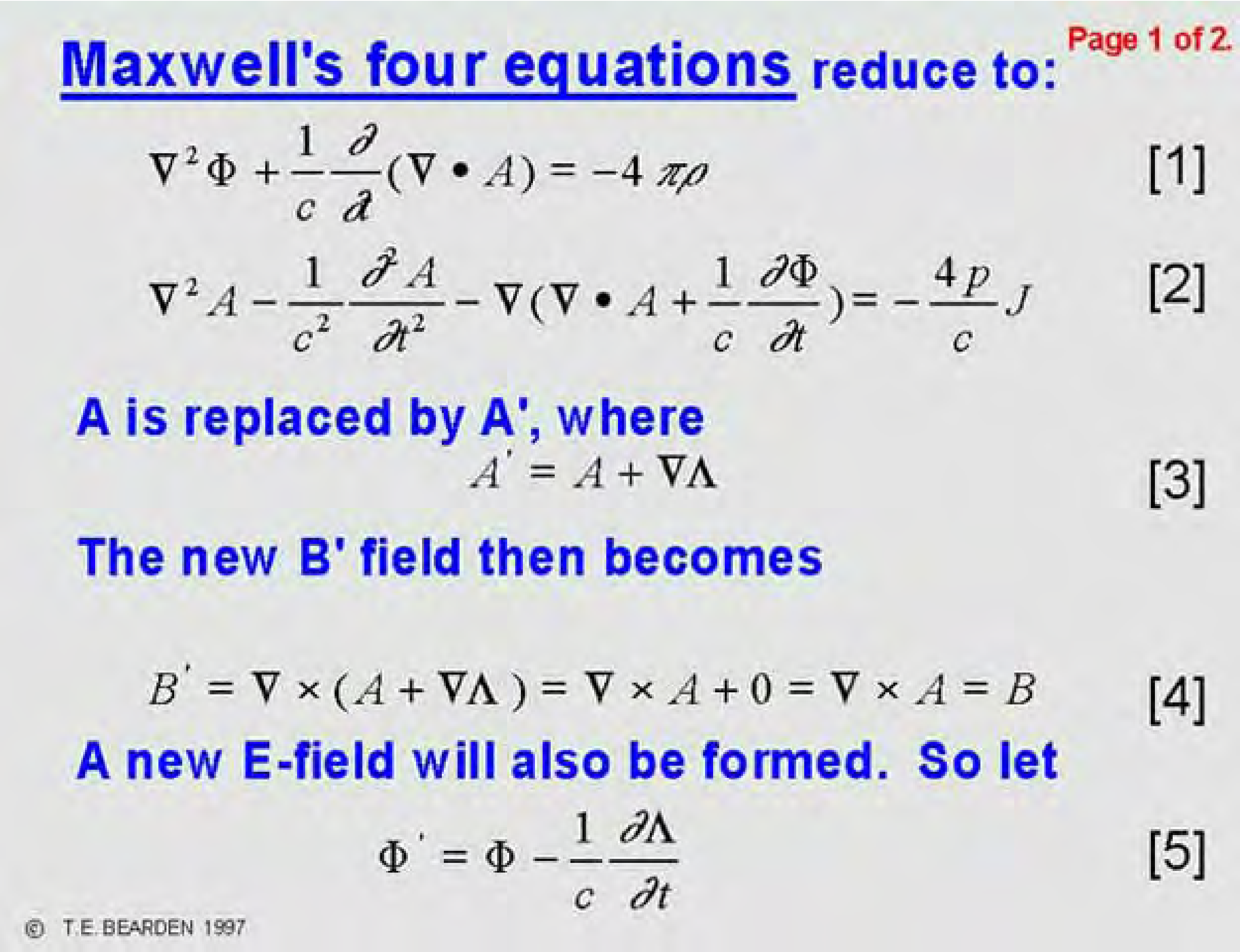

Page 1 of 2.

Maxwell's four equations reduce to:

$$\nabla^2\Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet A) = -4\pi\rho \qquad [1]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} - \nabla\left(\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t}\right) = -\frac{4p}{c}J \qquad [2]$$

A is replaced by A', where

$$A' = A + \nabla\Lambda \qquad [3]$$

The new B' field then becomes

$$B' = \nabla \times (A + \nabla\Lambda) = \nabla \times A + 0 = \nabla \times A = B \qquad [4]$$

A new E-field will also be formed. So let

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t} \qquad [5]$$

© T.E. BEARDEN 1997

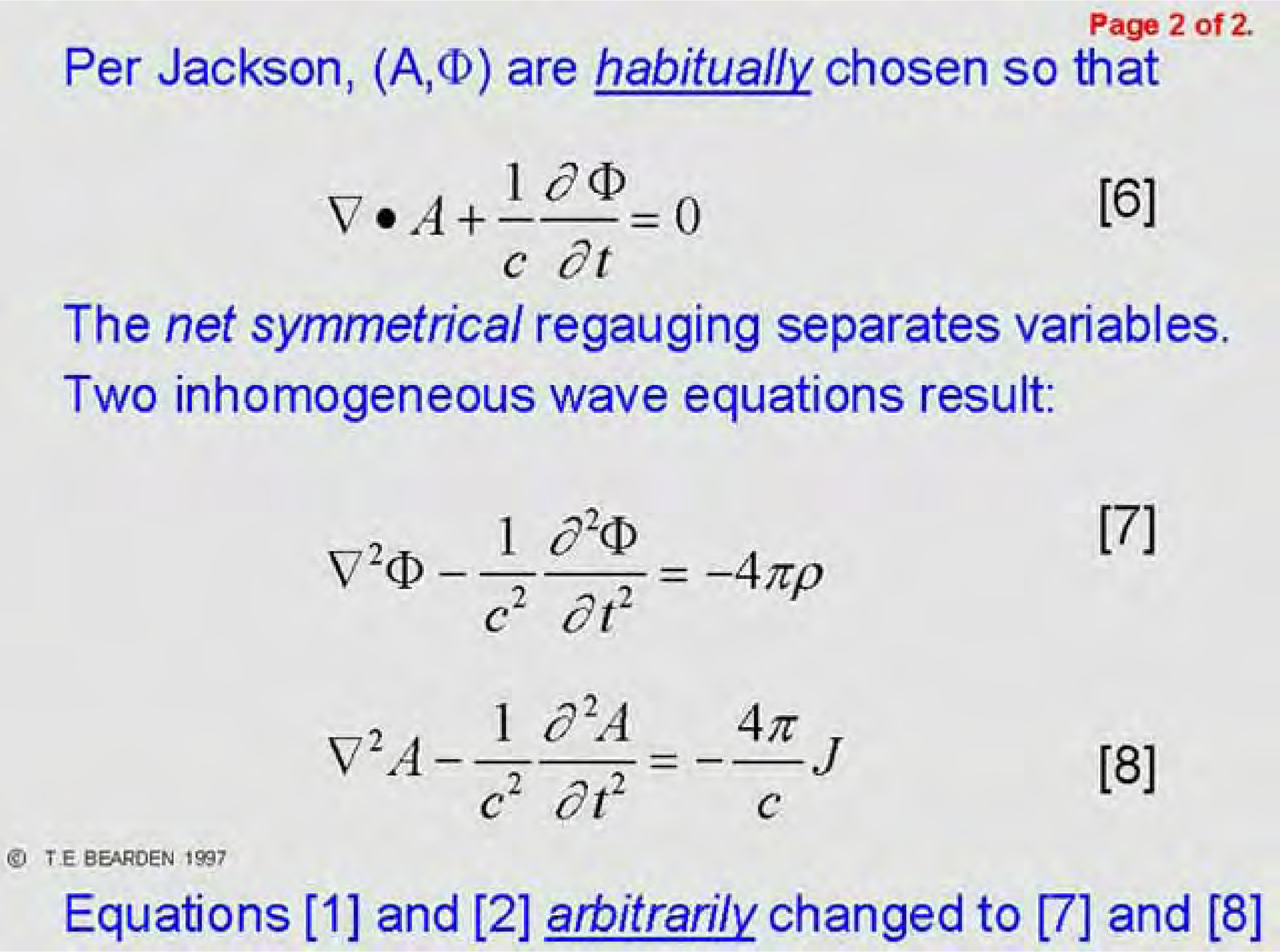

Page 2 of 2.

Per Jackson, (A,Φ) are *habitually* chosen so that

$$\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [6]$$

The *net symmetrical* regauging separates variables.

Two inhomogeneous wave equations result:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [7]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{4\pi}{c} J \qquad [8]$$

© T.E. BEARDEN 1997

Equations [1] and [2] *arbitrarily* changed to [7] and [8]

The Tom Bearden Website

Life Systems

Toward the Beginning of the End

---

1. Toward the Beginning of the End: Capstone Presentations
2. Toward the Beginning of the End
3. Biospheric Evolution
4. A Living Biological System
5. The Evolutionary Path of Multicellular Organisms
6. To Paraphrase Churchill
7. Directed Evolution: Mechanism for Species Reactive Adaptation
8. Earth's Poynting Energy Flow
9. The Immune System
10. Mind and Matter Interaction
11. Levels of Consciousness and Unconsciousness
12. Access to Mind, Memory and Thought
13. US Intelligence Statement on Soviet Psi Research
14. Mind and Matter Interaction
15. Conscious and Unconscious Minds
16. Velocity is Rotation
17. Species Reactive Adaptation: Directed Evolution
18. Technology's Goal
19. Technology's End
20. Technological Species' Problems

The Tom Bearden
Website

Briefings

One of Tom's first public lectures: World Futures Society 1982
Other panelists included James Beale, Carol Rosin, Barbara Marx Hubbard and Glenn Doman

---

The "Porthole" Concept
Extension of the Priore Method of Healing as applicable to the treatment of biological warfare effects and cellular afflictions such as cancers, AIDS etc.
NEW

Hubbert's Peak: The Impending Oil Crisis, by Professor Kenneth S. Deffeyes
Book Review and Slides

A National Report on America's Energy Crisis
by Energy Secretary Spencer Abraham - March 19, 2001

Commentary and Slide Show (66) by Tom Bearden

A Visual Tour of what they don't want you to know about electrical circuits
10 Trillion Percent of the Current Produced is Wasted!

Weapons Slides

The Missing Infolded Electrodynamics
85 slides

A Brief Technical Background to the Extended Priore Healing Methodology
11 slides

Briefing sent by Tom Bearden to Defense Department, National Institute of Health, USAF, National Science Foundation etc. re extended and improved Priore Treatment to counter mass casualties from Biological Warfare

- Letter to DHHS re adoption of Priore technology
  24 June, 1998
- Letter to NIH re adoption of Priore technology
  12 May, 1998

Towards the Beginning of the End - Lecture in Florida for Capstone Presentations, Sept. 1998

TOWARD THE BEGINNING OF THE END

Capstone Presentations
26-27 Sep. 1998

T. E. Bearden
LTC, US Army (Retired)
M.S., Nuclear Engineering
Fellow Emeritus, AIAS

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

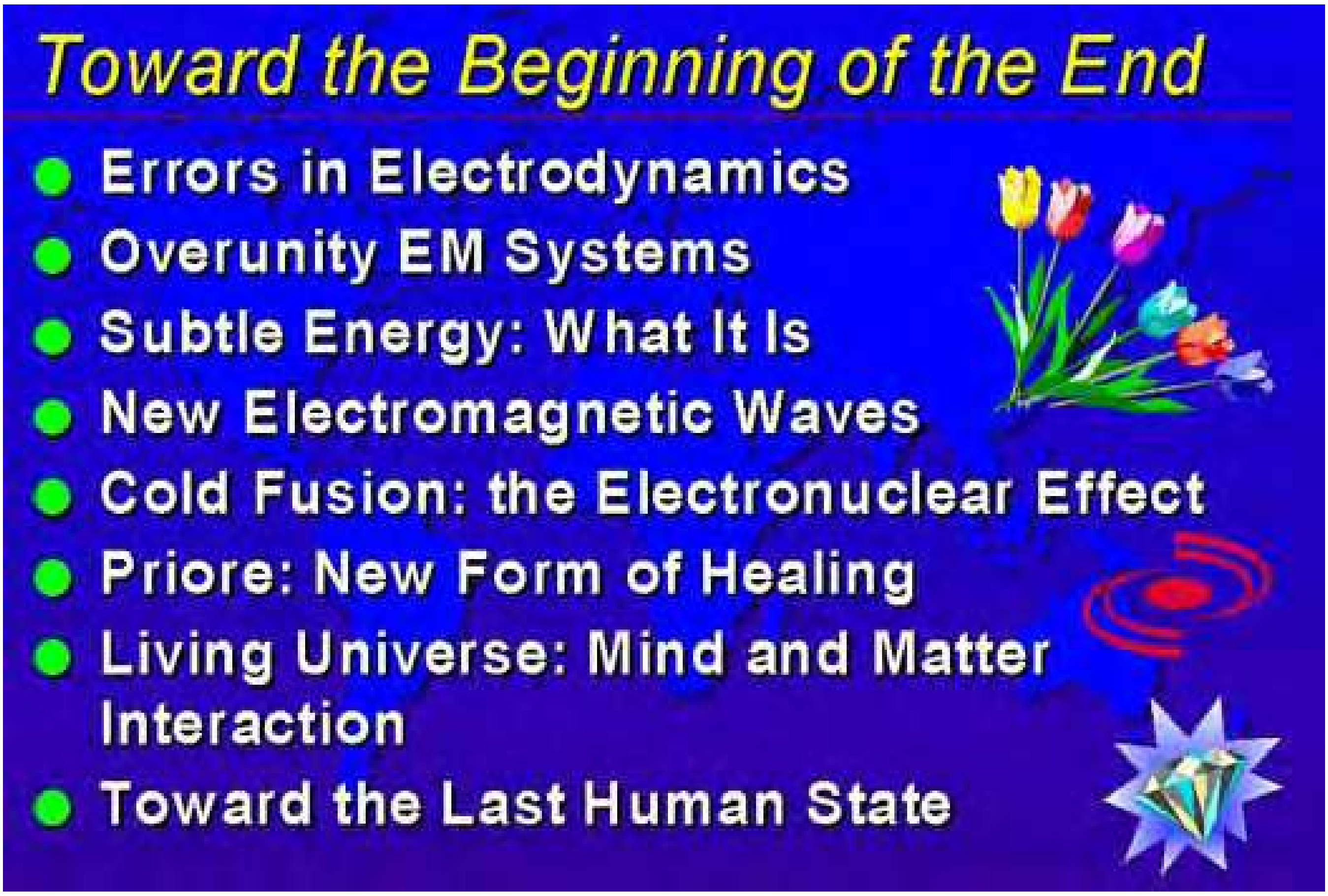

Toward the Beginning of the End

- Errors in Electrodynamics
- Overunity EM Systems
- Subtle Energy: What It Is
- New Electromagnetic Waves
- Cold Fusion: the Electronuclear Effect
- Priore: New Form of Healing
- Living Universe: Mind and Matter Interaction
- Toward the Last Human State

Previous Slide

Return to Slide Index

Next Slide

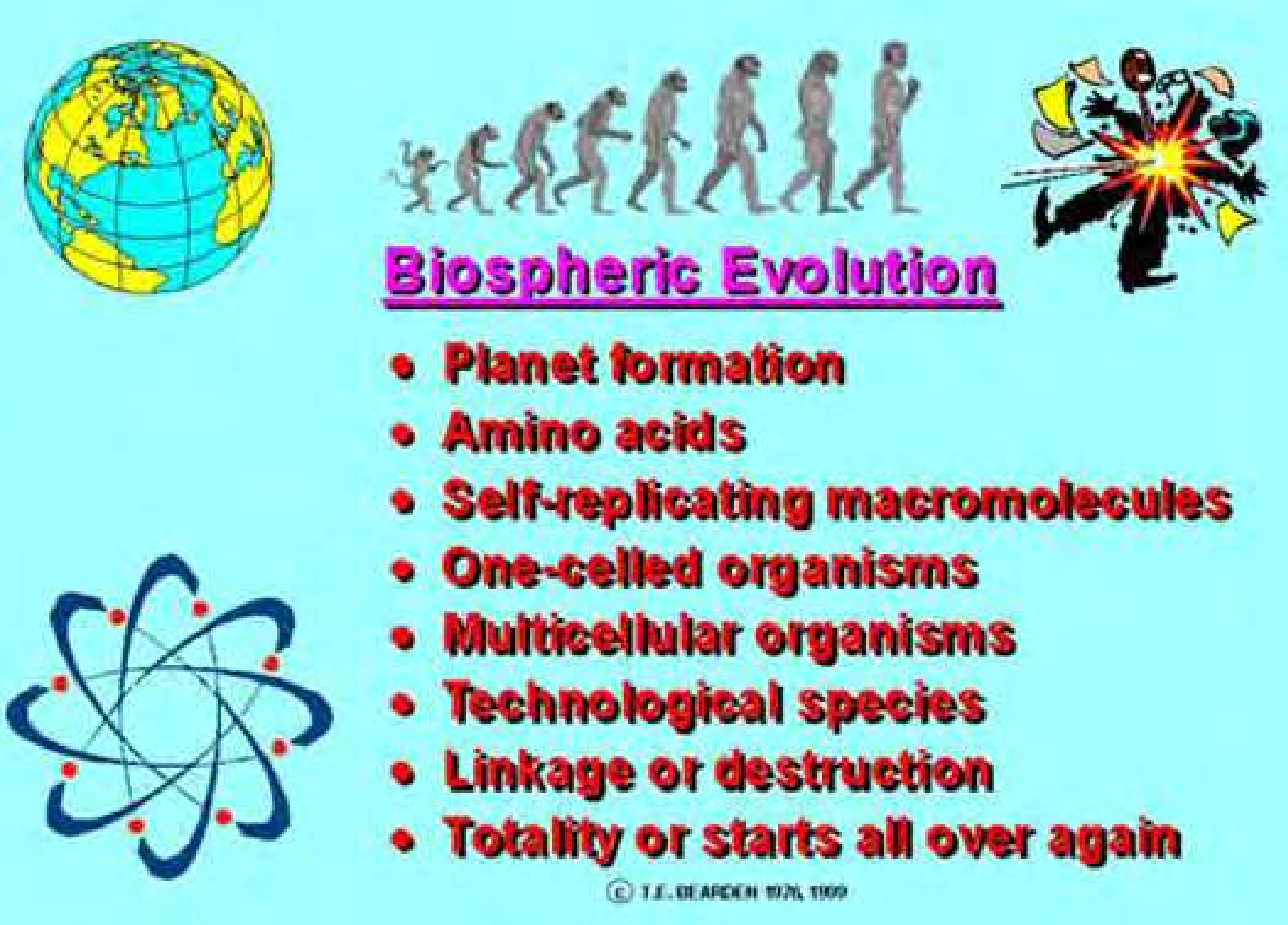

Biospheric Evolution

- Planet formation
- Amino acids
- Self-replicating macromolecules
- One-celled organisms
- Multicellular organisms
- Technological species
- Linkage or destruction
- Totality or starts all over again

© T.E. BEARDEN 1976, 1999

Return to Slide Index

Next Slide

Previous Slide

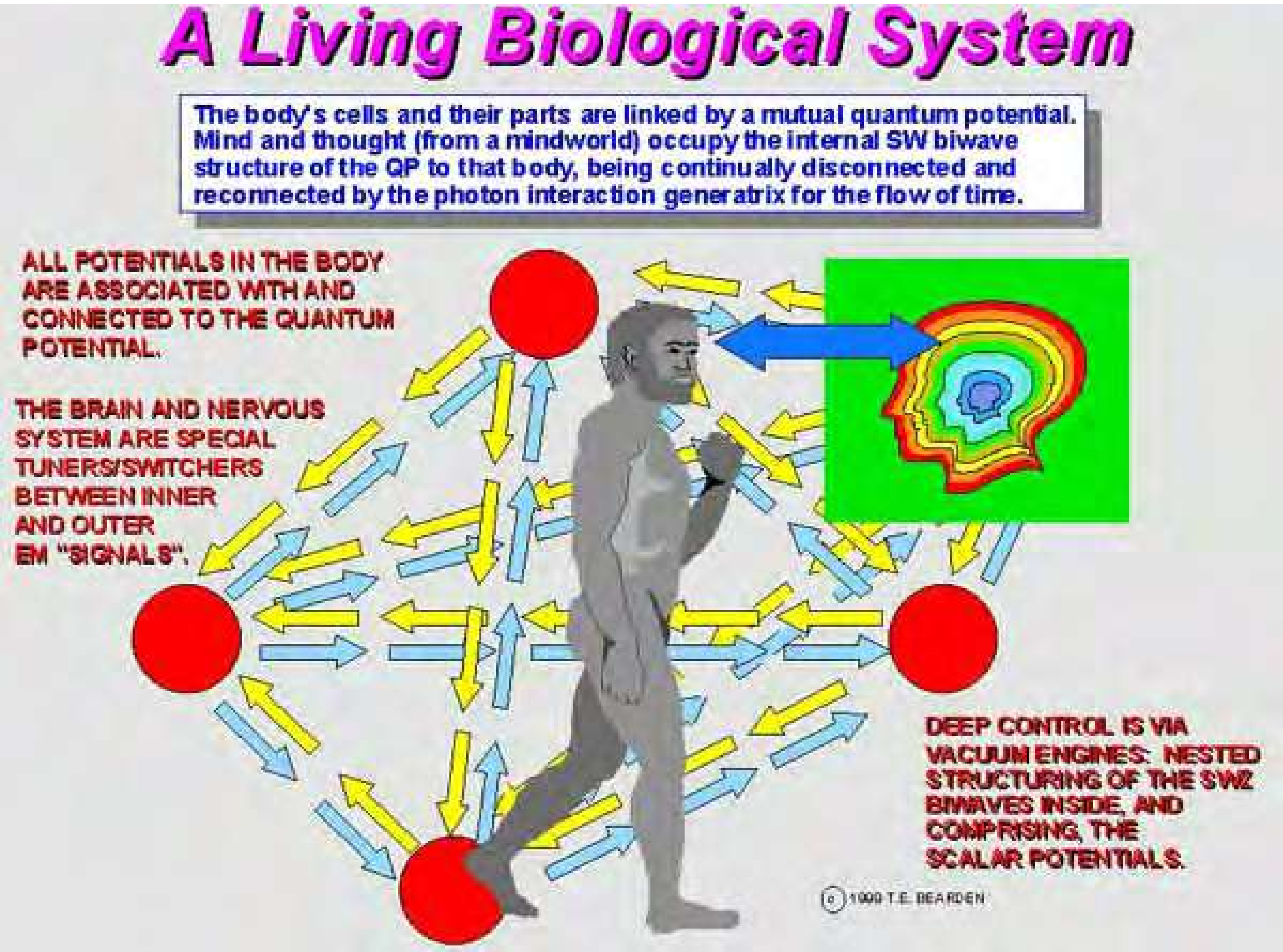

A Living Biological System

The body's cells and their parts are linked by a mutual quantum potential. Mind and thought (from a mindworld) occupy the internal SW biwave structure of the QP to that body, being continually disconnected and reconnected by the photon interaction generatrix for the flow of time.

ALL POTENTIALS IN THE BODY ARE ASSOCIATED WITH AND CONNECTED TO THE QUANTUM POTENTIAL.

THE BRAIN AND NERVOUS SYSTEM ARE SPECIAL TUNERS/SWITCHERS BETWEEN INNER AND OUTER EM "SIGNALS".

DEEP CONTROL IS VIA VACUUM ENGINES: NESTED STRUCTURING OF THE SWZ BIWAVES INSIDE, AND COMPRISING, THE SCALAR POTENTIALS.

(c) 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

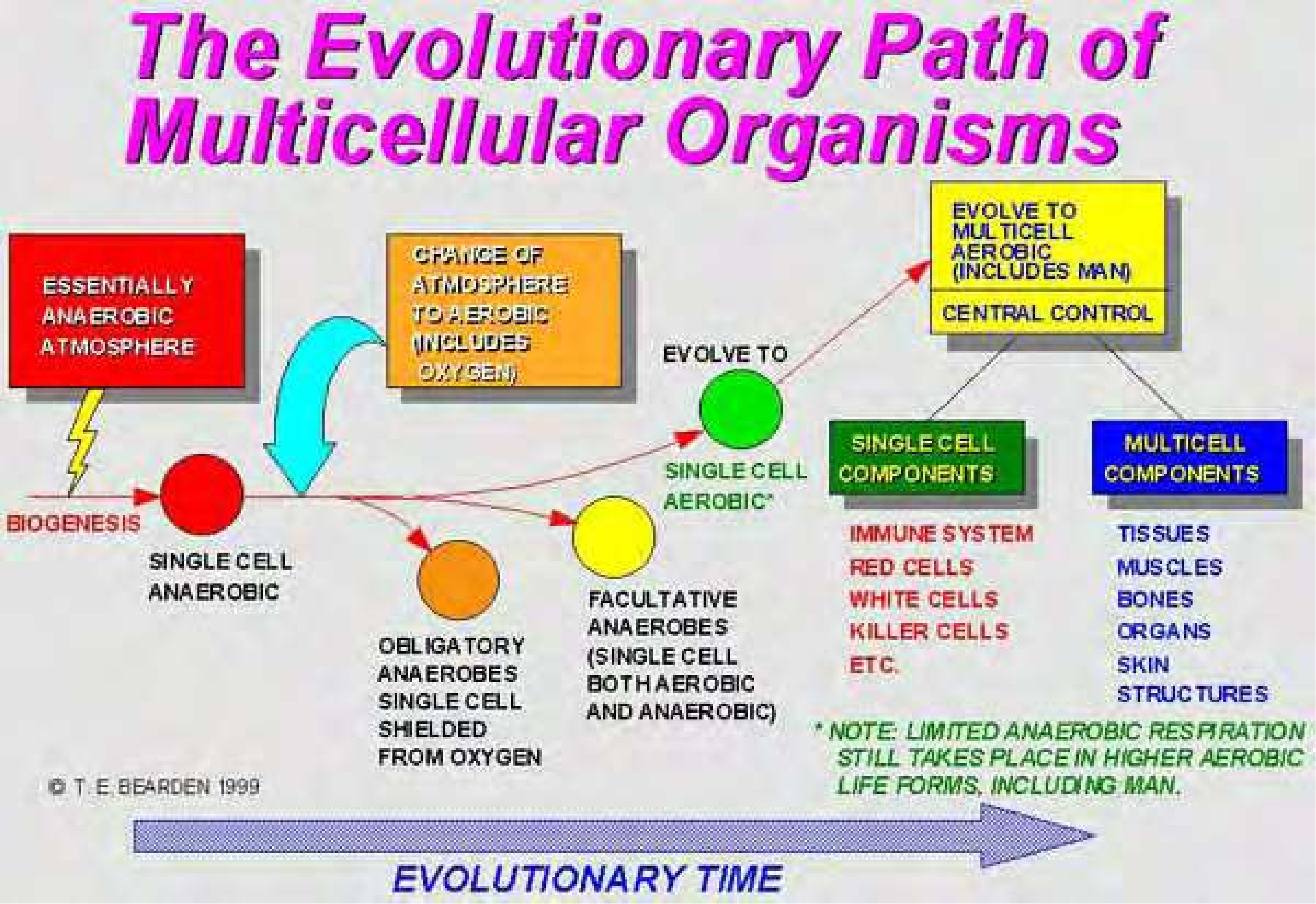

The Evolutionary Path of Multicellular Organisms

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

EVOLVE TO MULTICELL AEROBIC (INCLUDES MAN)

CENTRAL CONTROL

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

© T. E. BEARDEN 1999

EVOLUTIONARY TIME

Return to Slide Index

Next Slide

Previous Slide

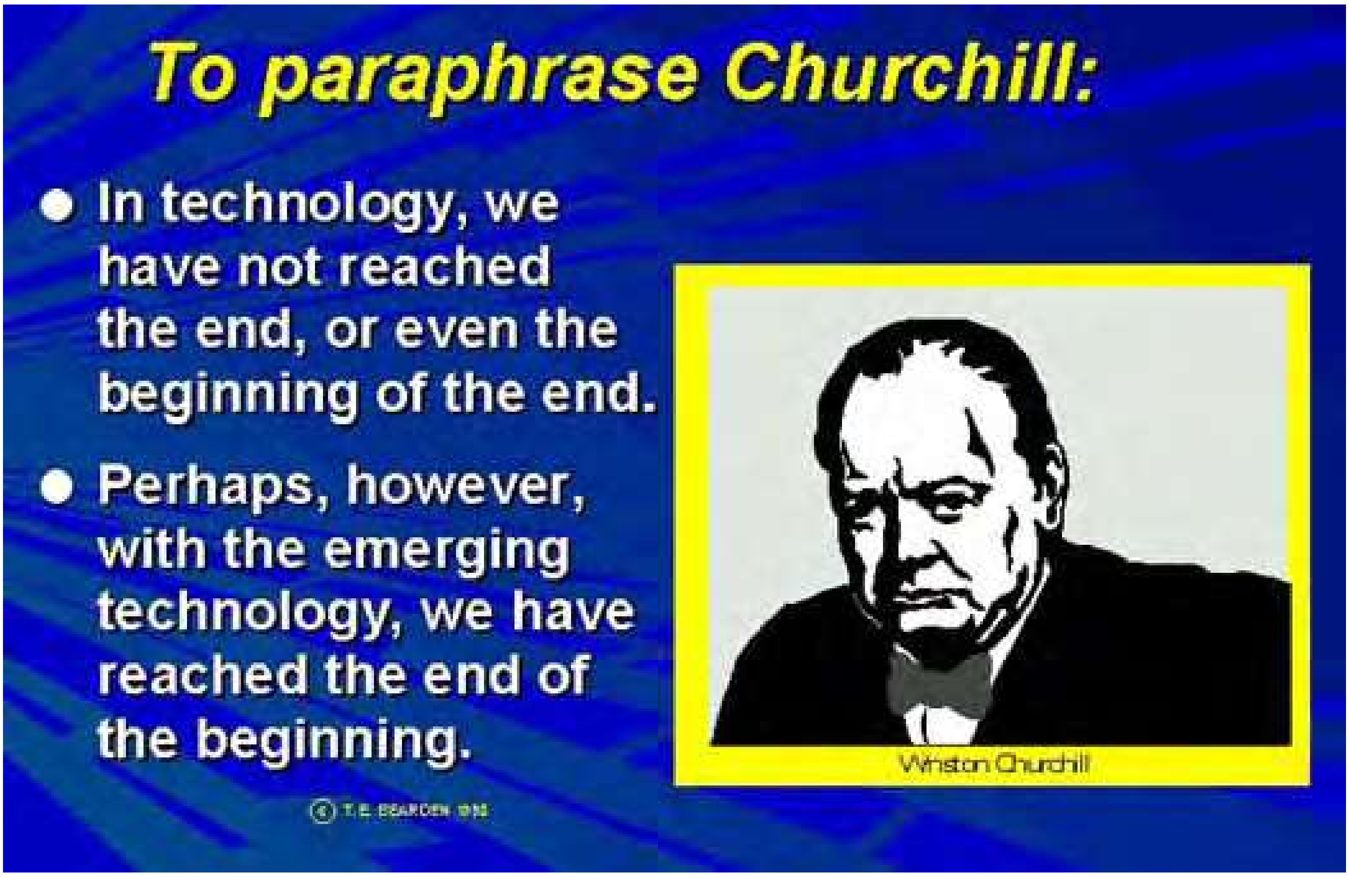

To paraphrase Churchill:

- In technology, we have not reached the end, or even the beginning of the end.
- Perhaps, however, with the emerging technology, we have reached the end of the beginning.

Winston Churchill

© T.E. BEARDEN [illegible]

Return to Slide Index

Next Slide

Previous Slide

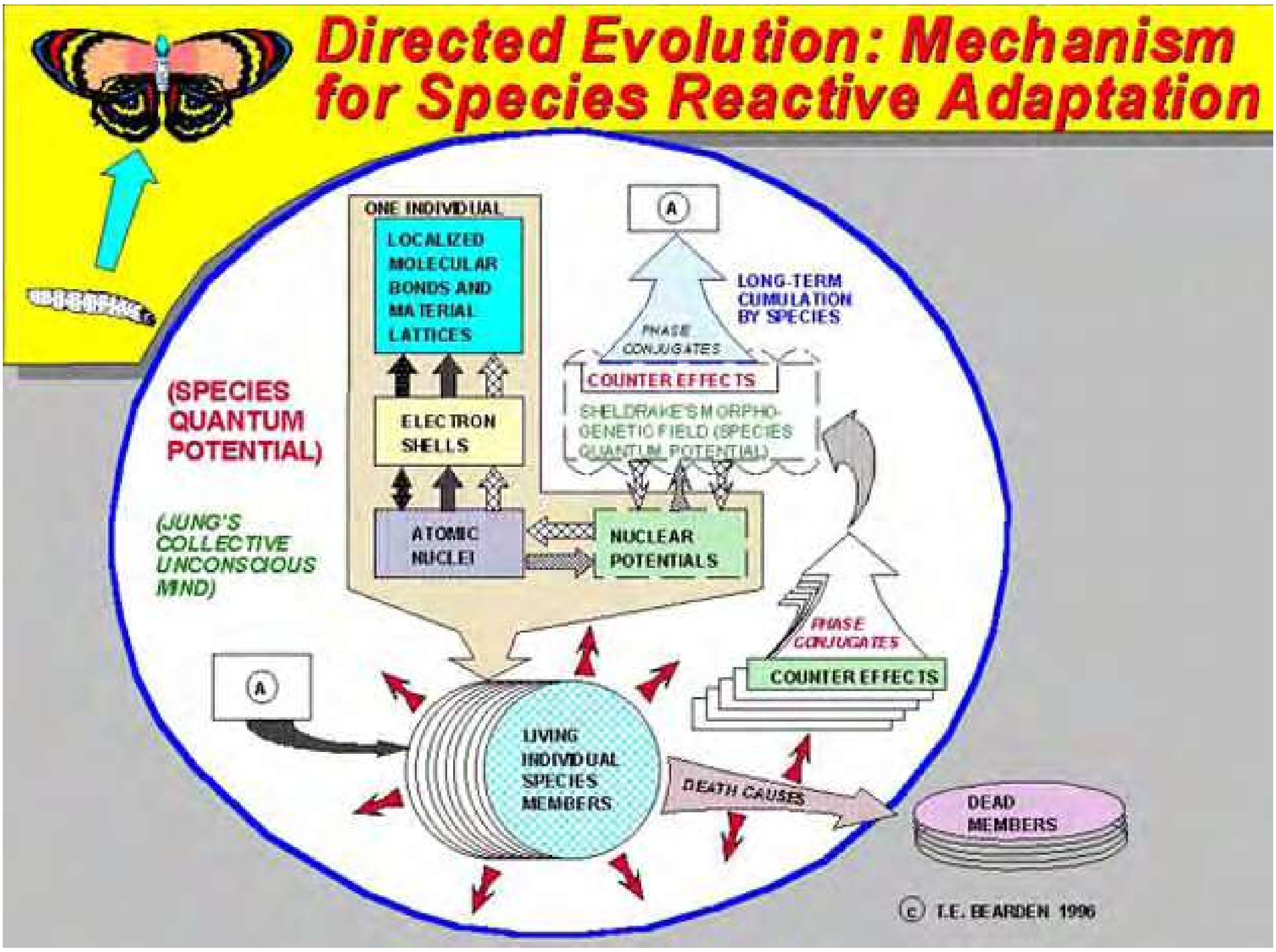

Directed Evolution: Mechanism for Species Reactive Adaptation

ONE INDIVIDUAL

LOCALIZED MOLECULAR BONDS AND MATERIAL LATTICES

ELECTRON SHELLS

ATOMIC NUCLEI

NUCLEAR POTENTIALS

A

LONG-TERM CUMULATION BY SPECIES

PHASE CONJUGATES

COUNTER EFFECTS

SHELDRAKE'S MORPHOGENETIC FIELD (SPECIES QUANTUM POTENTIAL)

(SPECIES QUANTUM POTENTIAL)

(JUNG'S COLLECTIVE UNCONSCIOUS MIND)

A

LIVING INDIVIDUAL SPECIES MEMBERS

DEATH CAUSES

DEAD MEMBERS

PHASE CONJUGATES

COUNTER EFFECTS

ⓒ T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

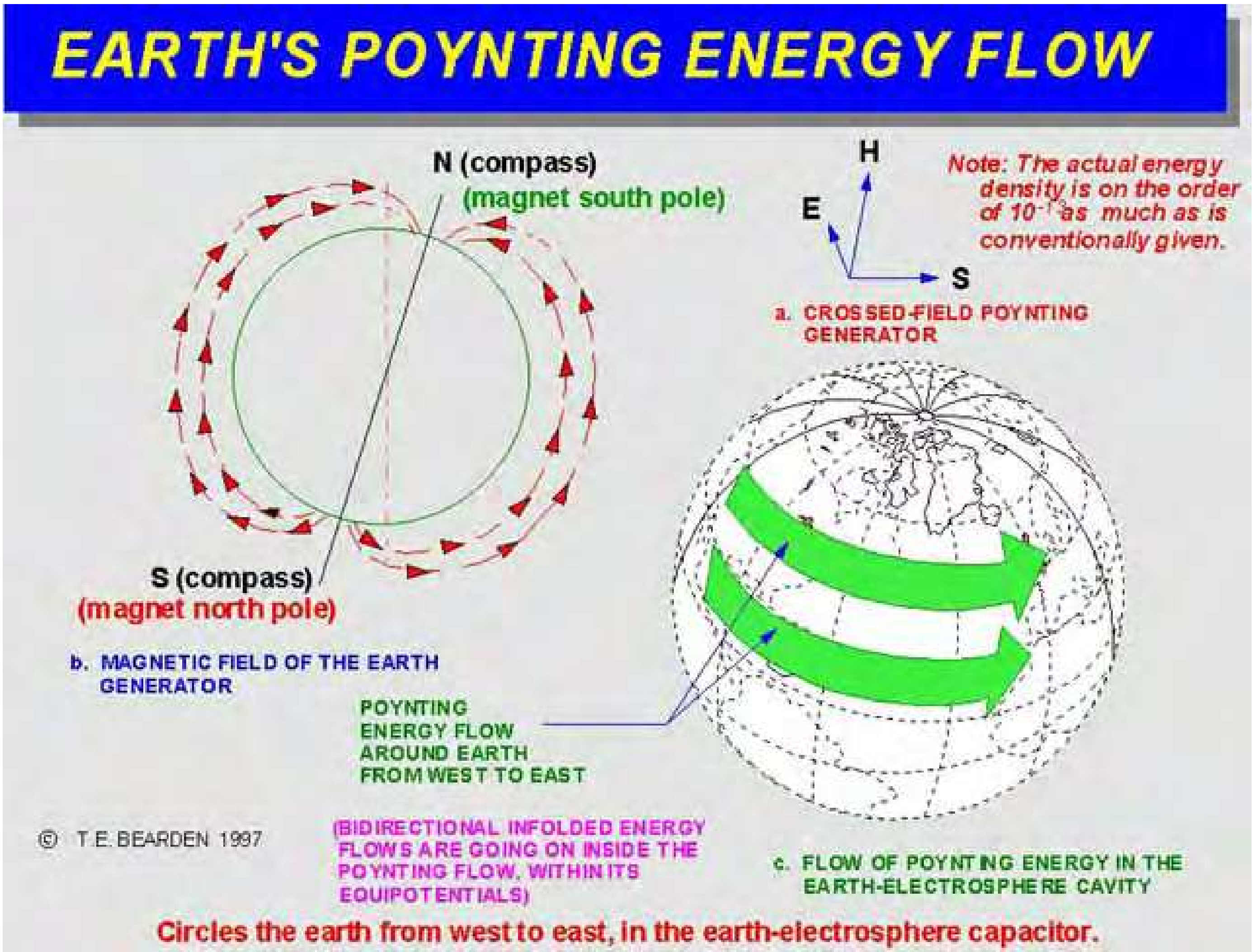

EARTH'S POYNTING ENERGY FLOW

Note: The actual energy density is on the order of 10^{-13} as much as is conventionally given.

a. CROSSED-FIELD POYNTING GENERATOR

b. MAGNETIC FIELD OF THE EARTH GENERATOR

POYNTING ENERGY FLOW AROUND EARTH FROM WEST TO EAST

(BIDIRECTIONAL INFOLDED ENERGY FLOWS ARE GOING ON INSIDE THE POYNTING FLOW, WITHIN ITS EQUIPOTENTIALS)

c. FLOW OF POYNTING ENERGY IN THE EARTH-ELECTROSPHERE CAVITY

© T.E. BEARDEN 1997

Circles the earth from west to east, in the earth-electrosphere capacitor.

Return to Slide Index

Next Slide

Previous Slide

Note: Should be

10^{13}

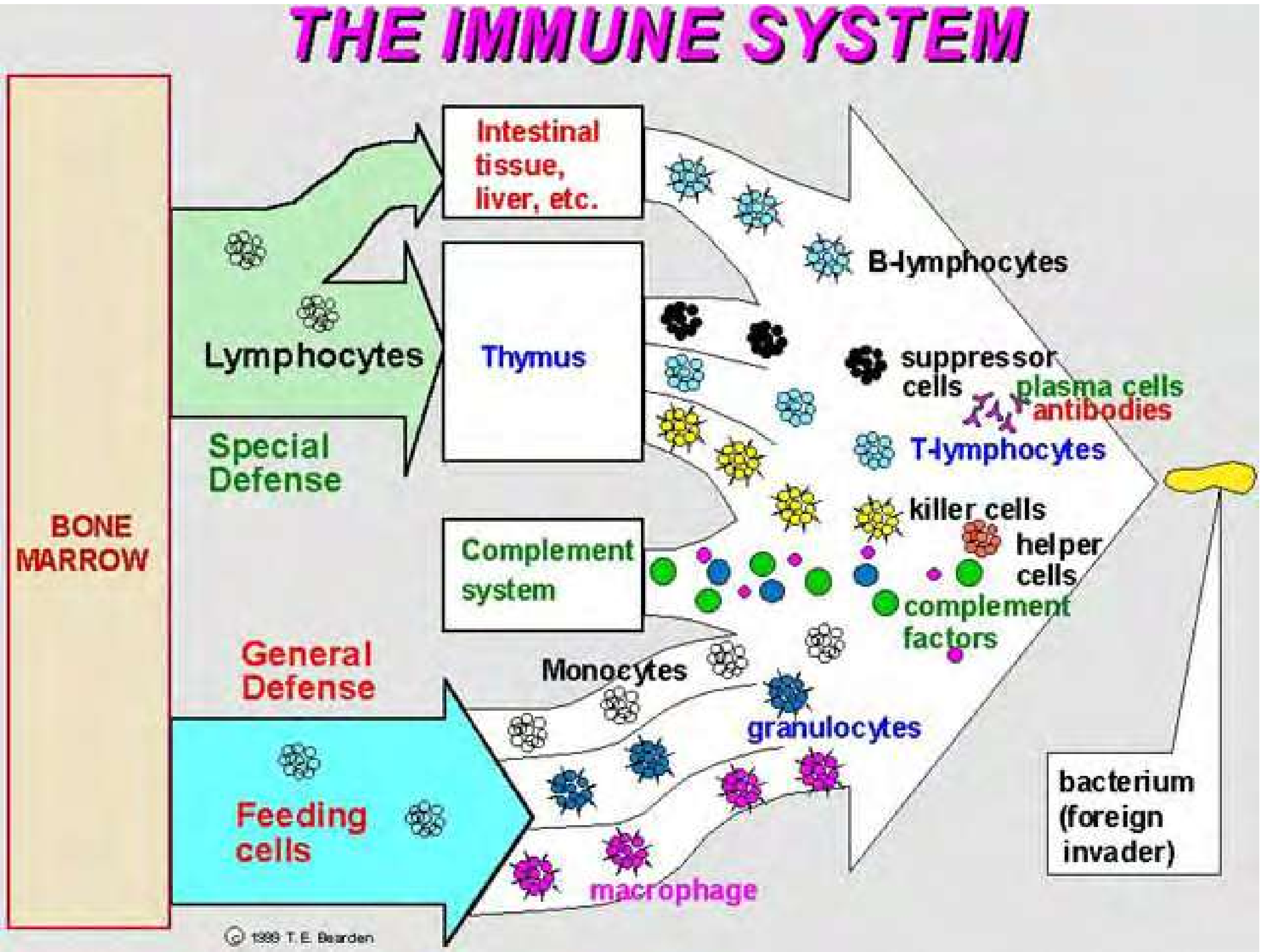
THE IMMUNE SYSTEM
BONE MARROW
Lymphocytes
Special Defense
General Defense
Feeding cells
Intestinal tissue, liver, etc.
Thymus
Complement system
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
Monocytes
granulocytes
macrophage
bacterium (foreign invader)
© 1999 T. E. Bearden

Return to Slide Index

Next Slide

Previous Slide

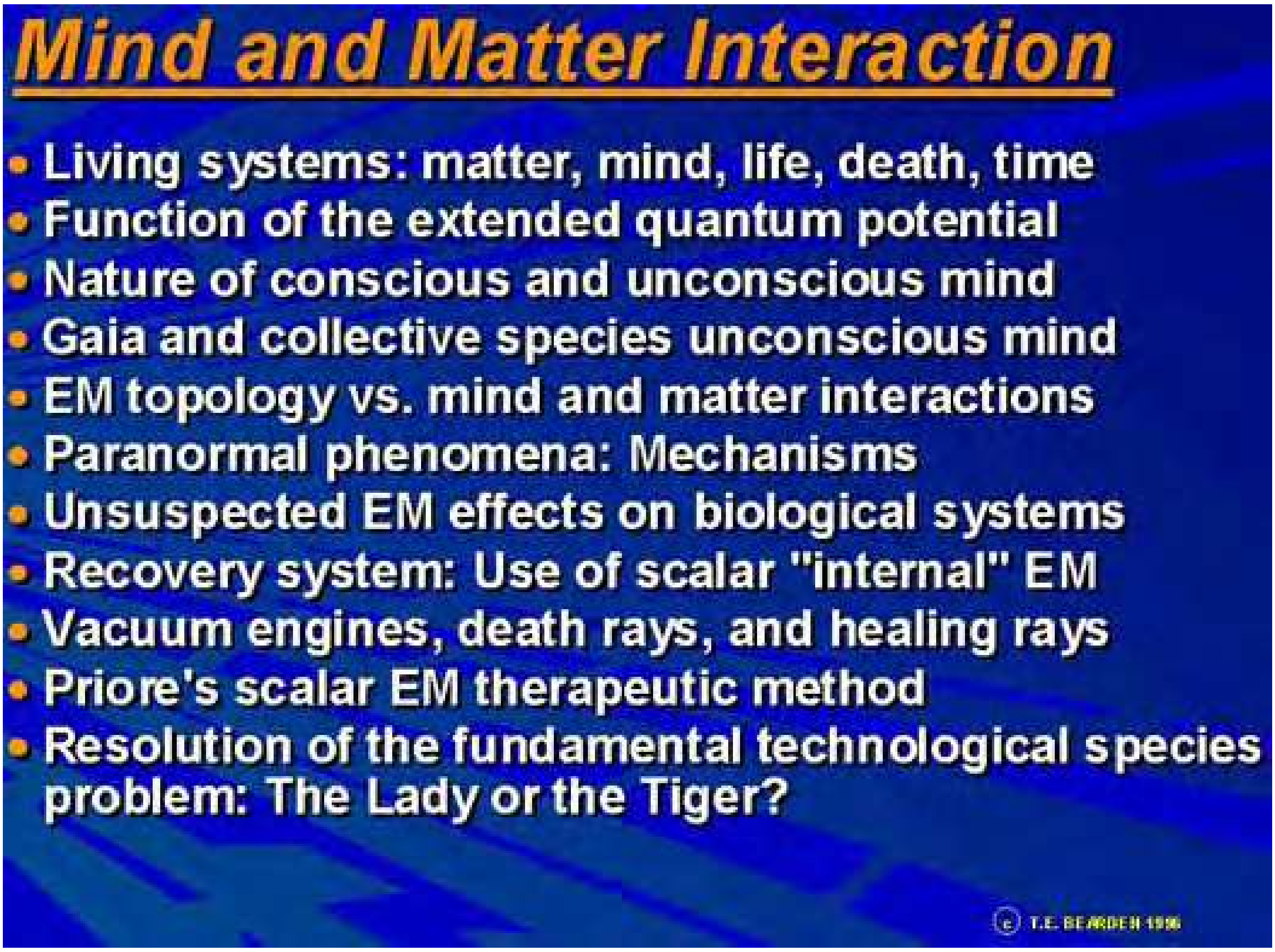

Mind and Matter Interaction

- Living systems: matter, mind, life, death, time
- Function of the extended quantum potential
- Nature of conscious and unconscious mind
- Gaia and collective species unconscious mind
- EM topology vs. mind and matter interactions
- Paranormal phenomena: Mechanisms
- Unsuspected EM effects on biological systems
- Recovery system: Use of scalar "internal" EM
- Vacuum engines, death rays, and healing rays
- Priore's scalar EM therapeutic method
- Resolution of the fundamental technological species problem: The Lady or the Tiger?

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

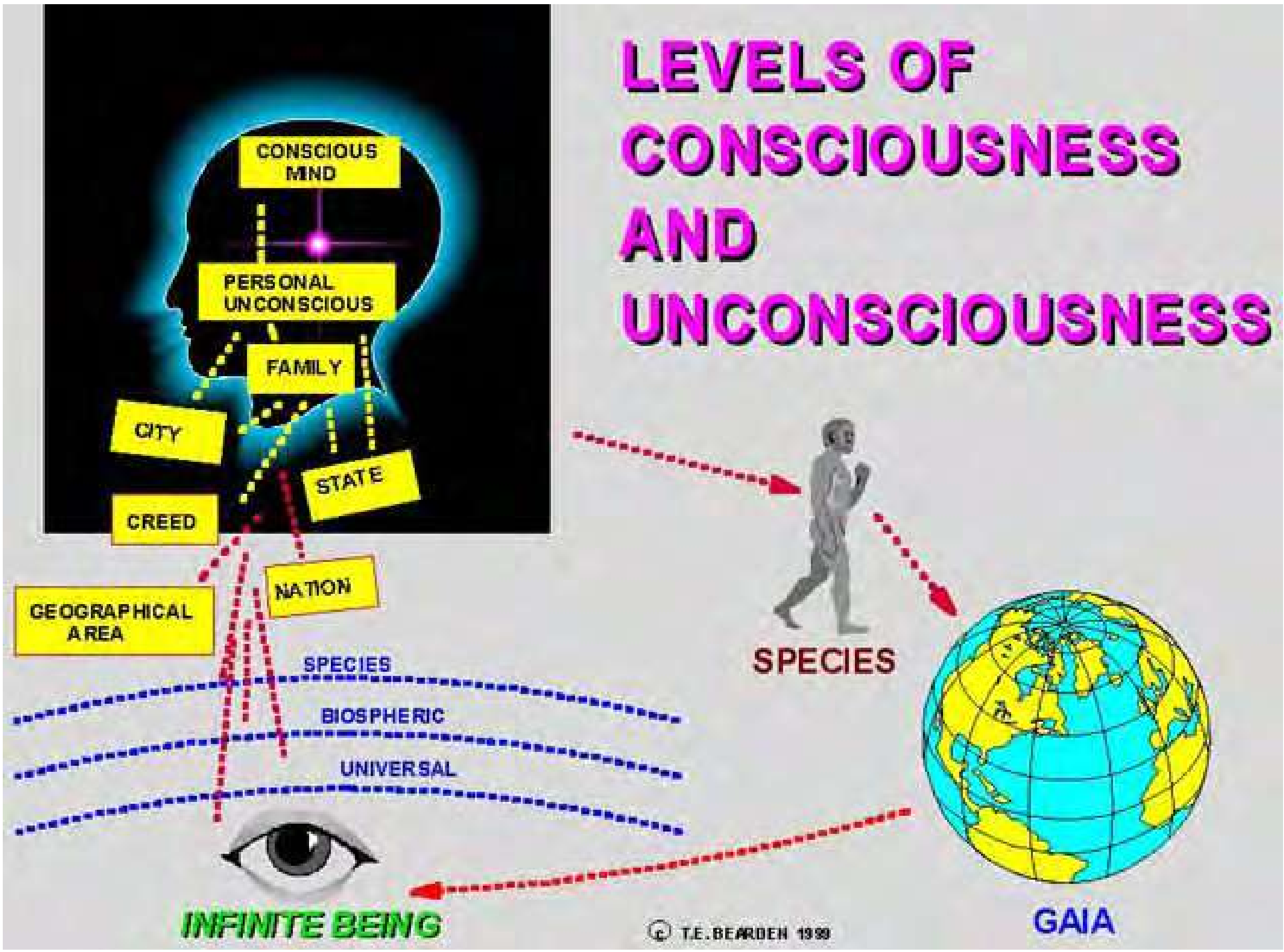

LEVELS OF CONSCIOUSNESS AND UNCONSCIOUSNESS

CONSCIOUS MIND
PERSONAL UNCONSCIOUS
FAMILY
CITY
STATE
CREED
NATION
GEOGRAPHICAL AREA
SPECIES
BIOSPHERIC
UNIVERSAL
INFINITE BEING
SPECIES
GAIA

© T.E. BEARDEN 1999

Return to Slide Index

Next Slide

Previous Slide

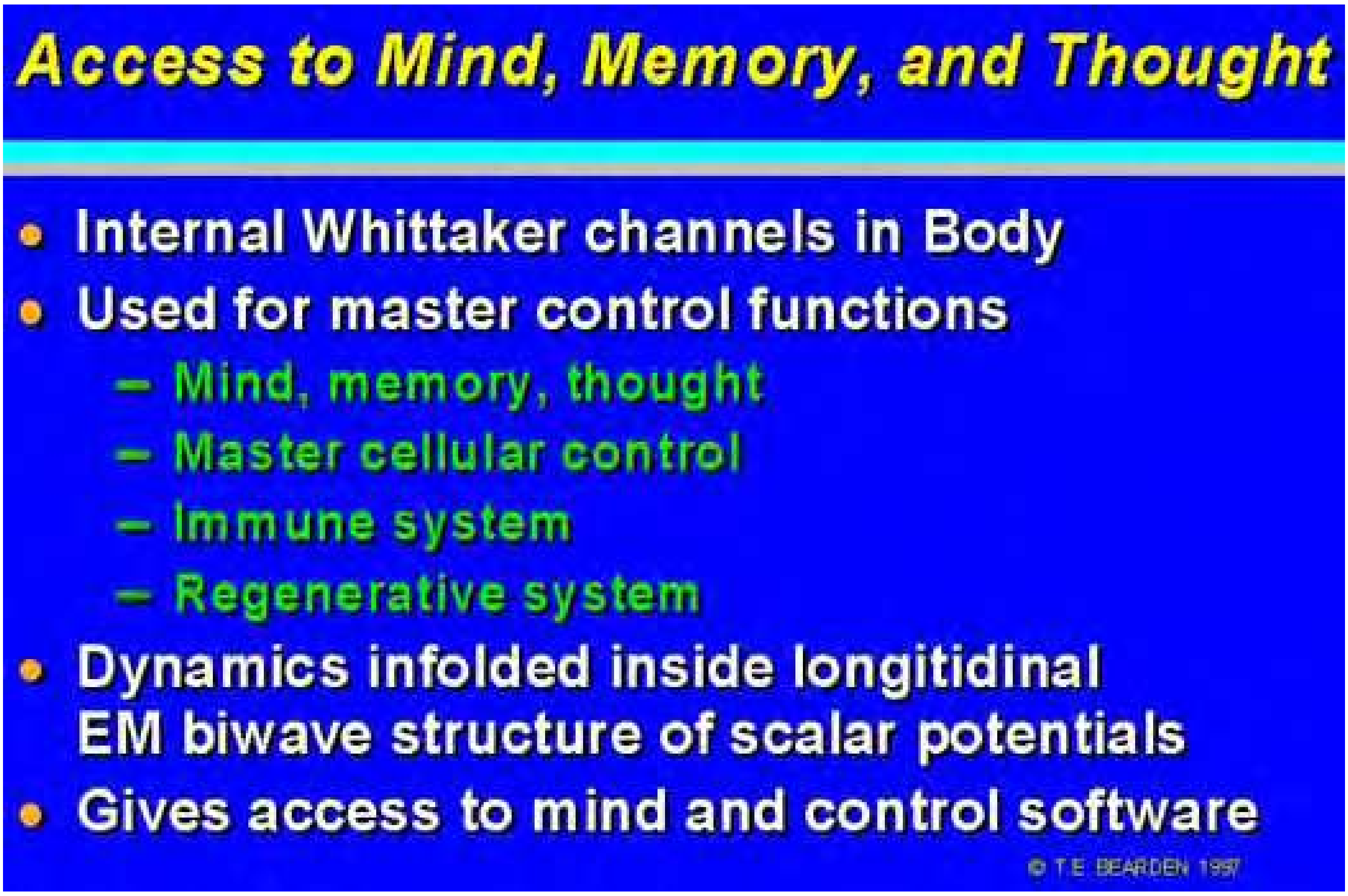

Access to Mind, Memory, and Thought

- Internal Whittaker channels in Body
- Used for master control functions
  - Mind, memory, thought
  - Master cellular control
  - Immune system
  - Regenerative system
- Dynamics infolded inside longitidinal EM biwave structure of scalar potentials
- Gives access to mind and control software

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

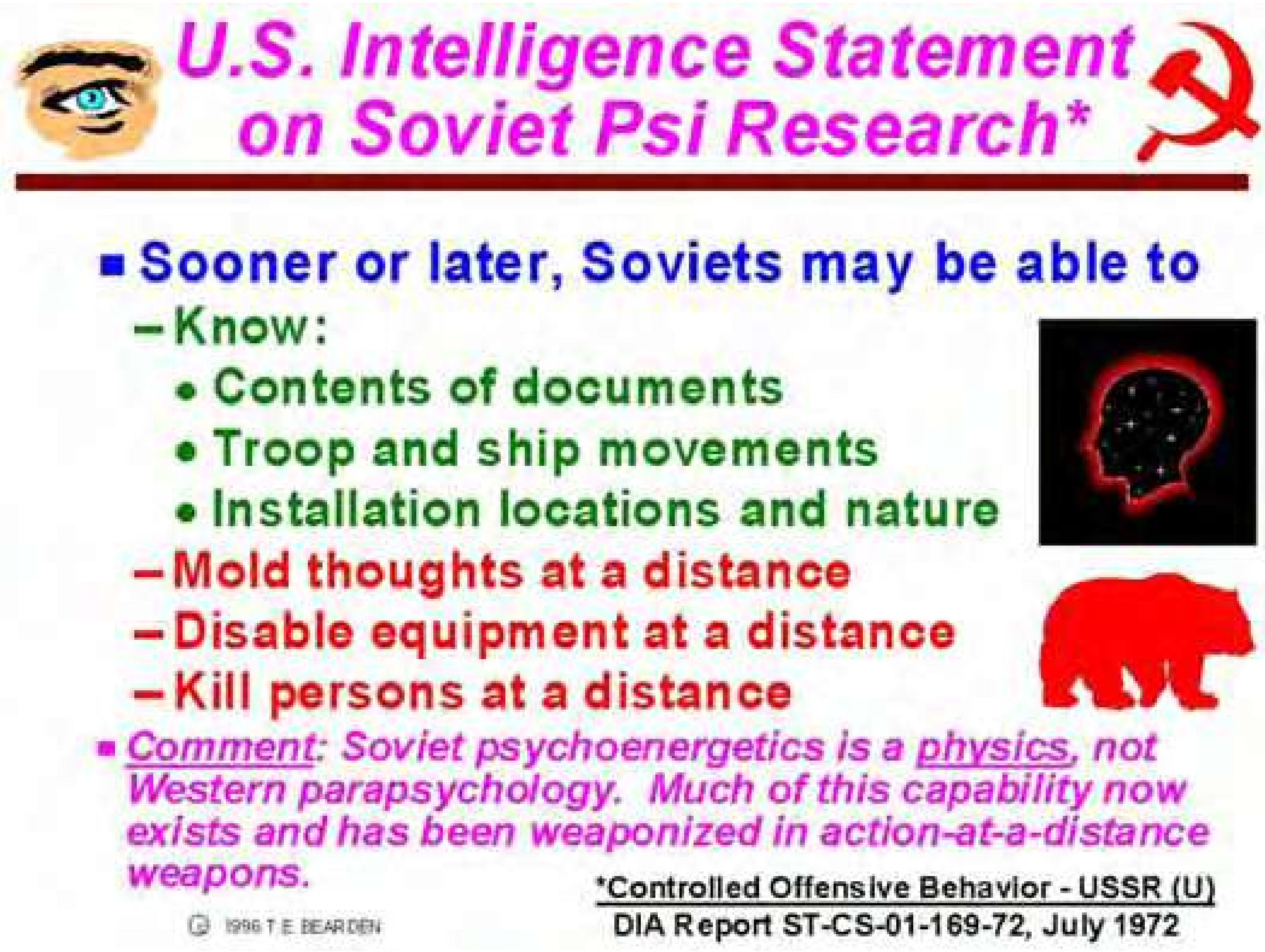

U.S. Intelligence Statement on Soviet Psi Research*

- Sooner or later, Soviets may be able to
  - Know:
    - Contents of documents
    - Troop and ship movements
    - Installation locations and nature
  - Mold thoughts at a distance
  - Disable equipment at a distance
  - Kill persons at a distance
- *Comment: Soviet psychoenergetics is a physics, not Western parapsychology. Much of this capability now exists and has been weaponized in action-at-a-distance weapons.*

*Controlled Offensive Behavior - USSR (U)
DIA Report ST-CS-01-169-72, July 1972

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

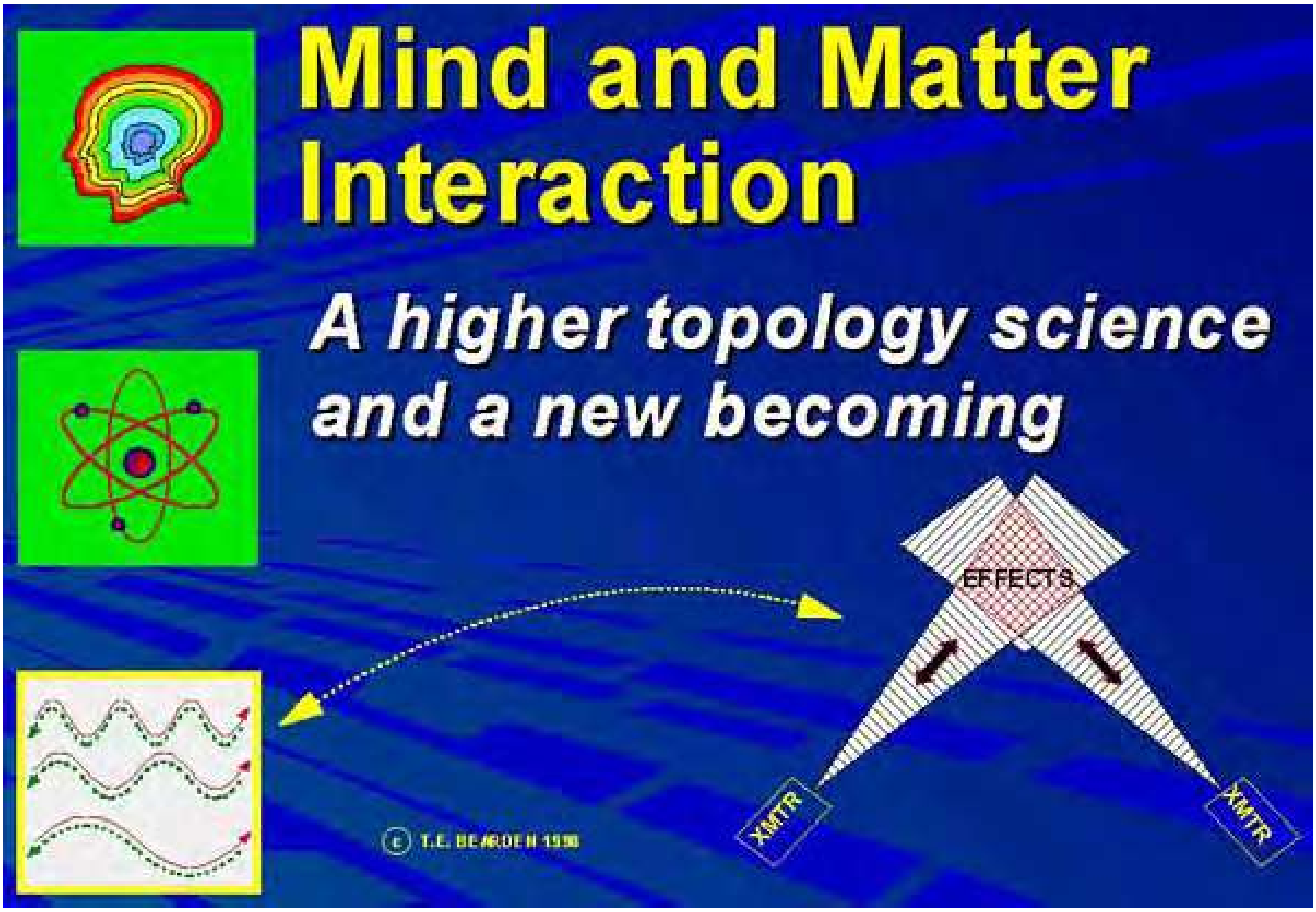

Mind and Matter Interaction

A higher topology science and a new becoming

EFFECTS
XMTR
XMTR

© T.E. BEARDEN 1990

[Return to Slide Index]
[Next Slide]
[Previous Slide]

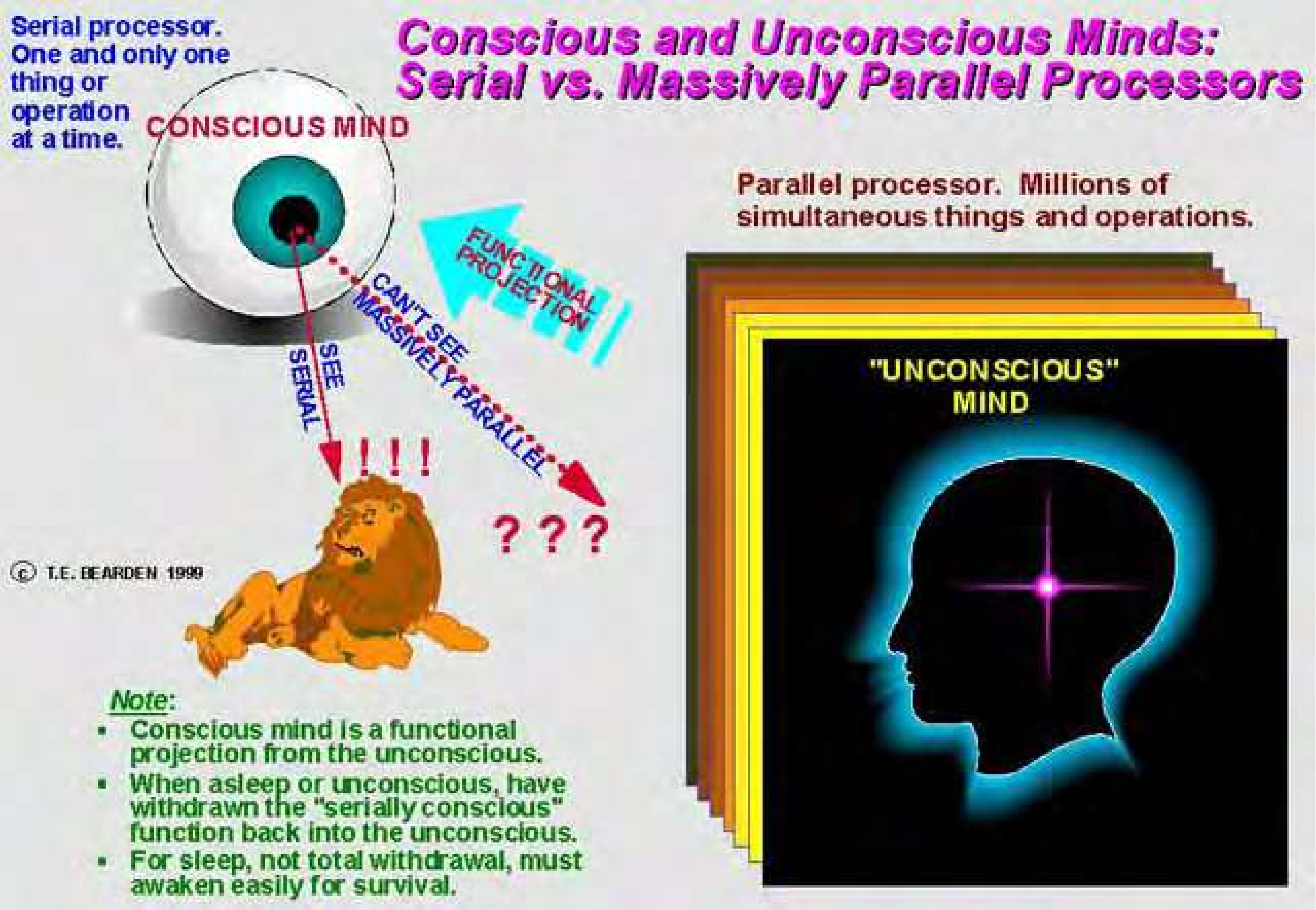

Conscious and Unconscious Minds: Serial vs. Massively Parallel Processors

Serial processor. One and only one thing or operation at a time.

Parallel processor. Millions of simultaneous things and operations.

CONSCIOUS MIND
FUNCTIONAL PROJECTION
CAN'T SEE MASSIVELY PARALLEL
SEE SERIAL
! ! !
? ? ?
"UNCONSCIOUS" MIND

© T.E. BEARDEN 1999

Note:

- Conscious mind is a functional projection from the unconscious.
- When asleep or unconscious, have withdrawn the "serially conscious" function back into the unconscious.
- For sleep, not total withdrawal, must awaken easily for survival.

Return to Slide Index

Next Slide

Previous Slide

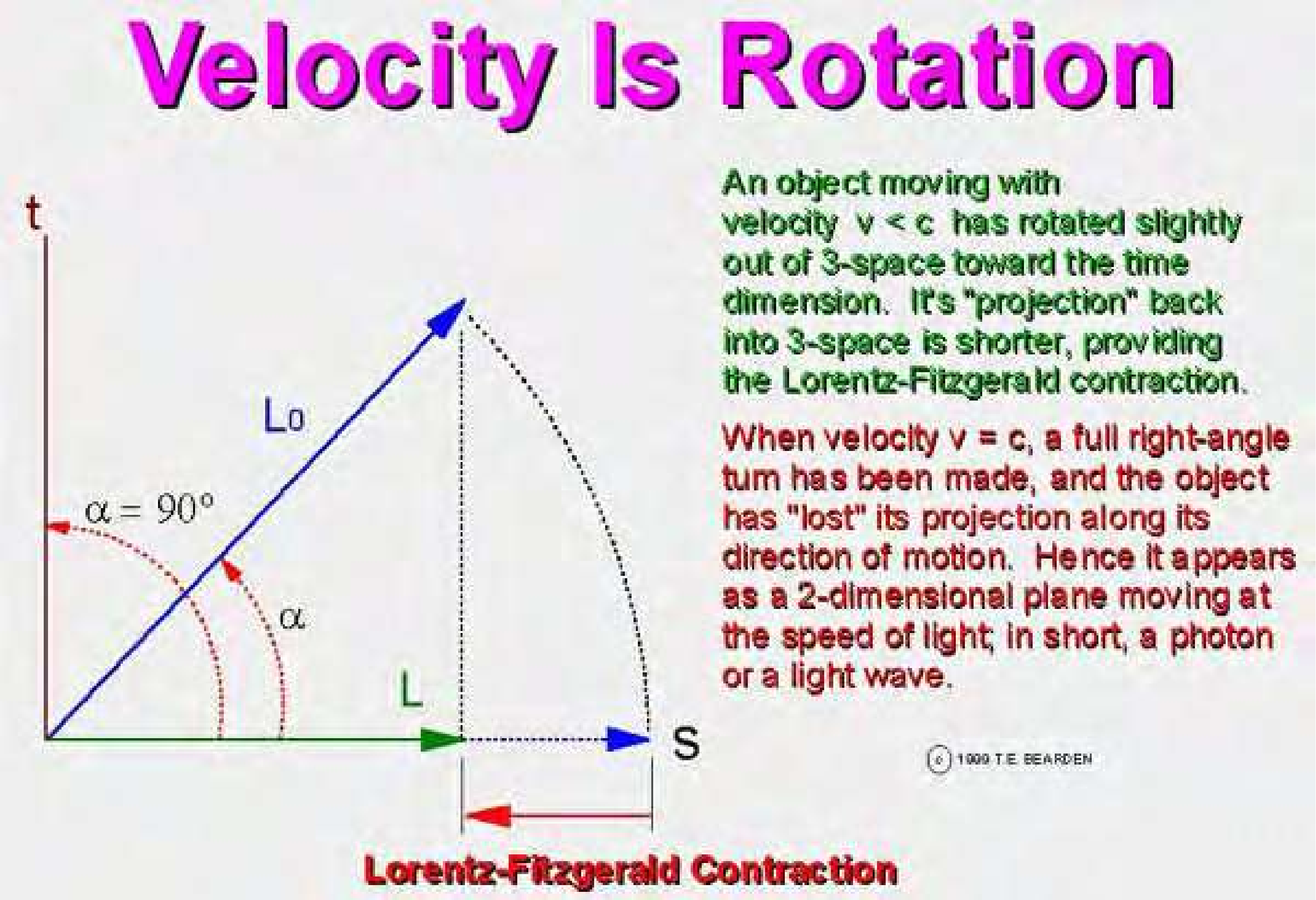

Velocity Is Rotation

An object moving with velocity $v < c$ has rotated slightly out of 3-space toward the time dimension. It's "projection" back into 3-space is shorter, providing the Lorentz-Fitzgerald contraction.

When velocity $v = c$, a full right-angle turn has been made, and the object has "lost" its projection along its direction of motion. Hence it appears as a 2-dimensional plane moving at the speed of light; in short, a photon or a light wave.

© 1999 T.E. BEARDEN

Lorentz-Fitzgerald Contraction

Next Slide

Return to Slide Index

Previous Slide

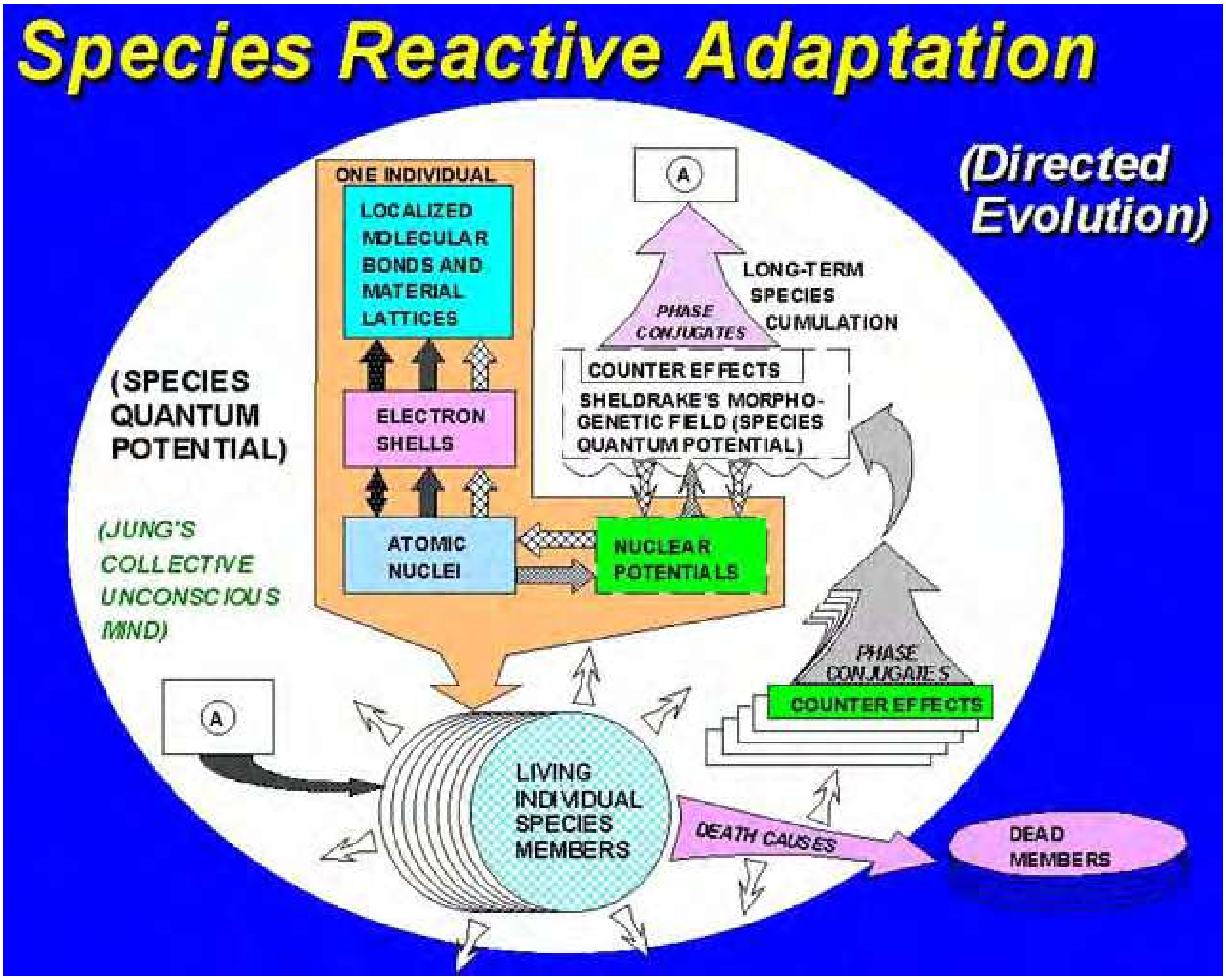

Species Reactive Adaptation

(Directed Evolution)

ONE INDIVIDUAL

LOCALIZED MOLECULAR BONDS AND MATERIAL LATTICES

ELECTRON SHELLS

ATOMIC NUCLEI

NUCLEAR POTENTIALS

A

PHASE CONJUGATES

LONG-TERM SPECIES CUMULATION

COUNTER EFFECTS

SHELDRAKE'S MORPHOGENETIC FIELD (SPECIES QUANTUM POTENTIAL)

(SPECIES QUANTUM POTENTIAL)

(JUNG'S COLLECTIVE UNCONSCIOUS MIND)

A

LIVING INDIVIDUAL SPECIES MEMBERS

DEATH CAUSES

DEAD MEMBERS

PHASE CONJUGATES

COUNTER EFFECTS

Return to Slide Index

Next Slide

Previous Slide

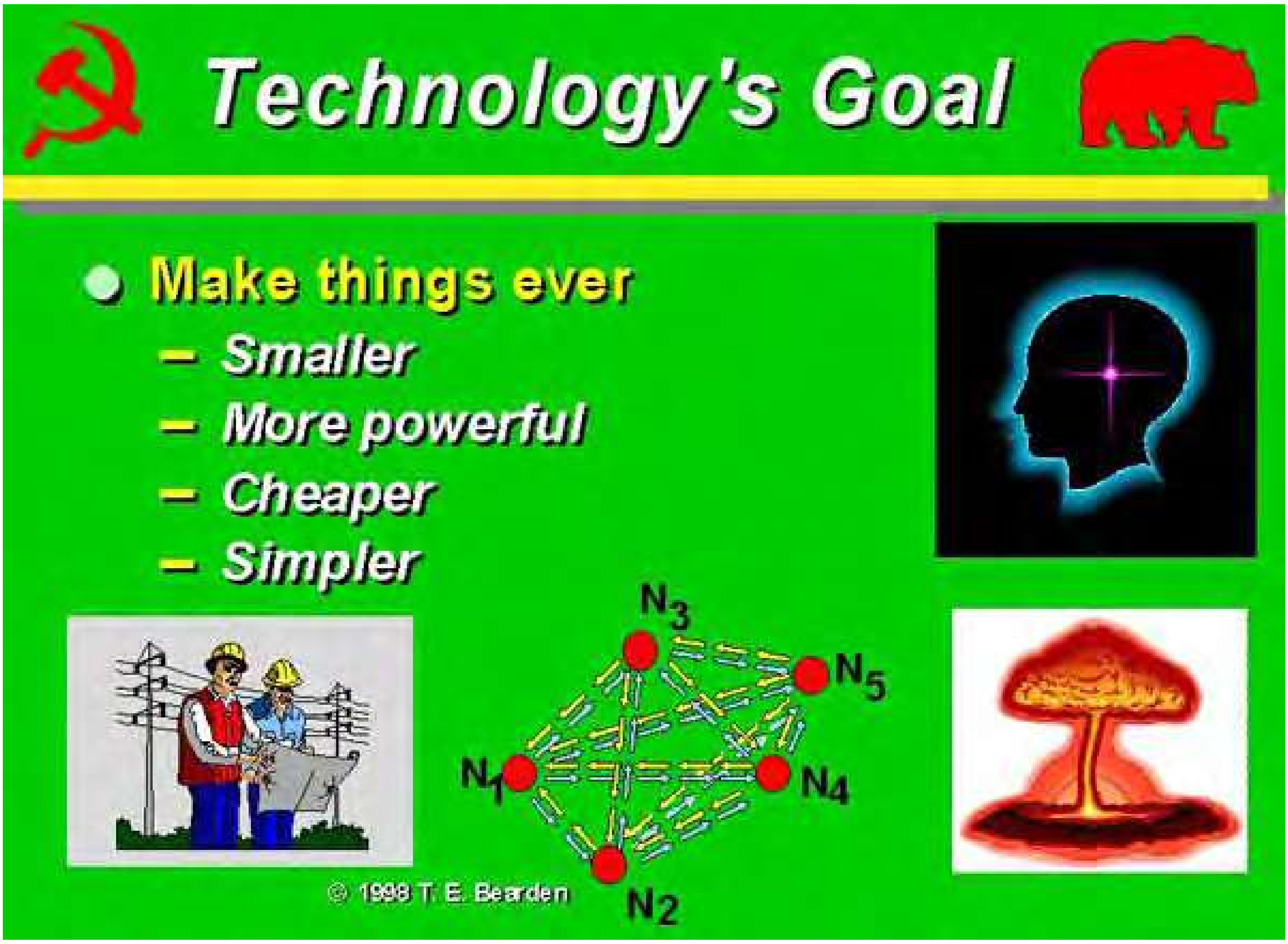

Technology's Goal

- Make things ever
  - Smaller
  - More powerful
  - Cheaper
  - Simpler

N3
N5
N1
N4
N2

© 1998 T. E. Bearden

Return to Slide Index

Next Slide

Previous Slide

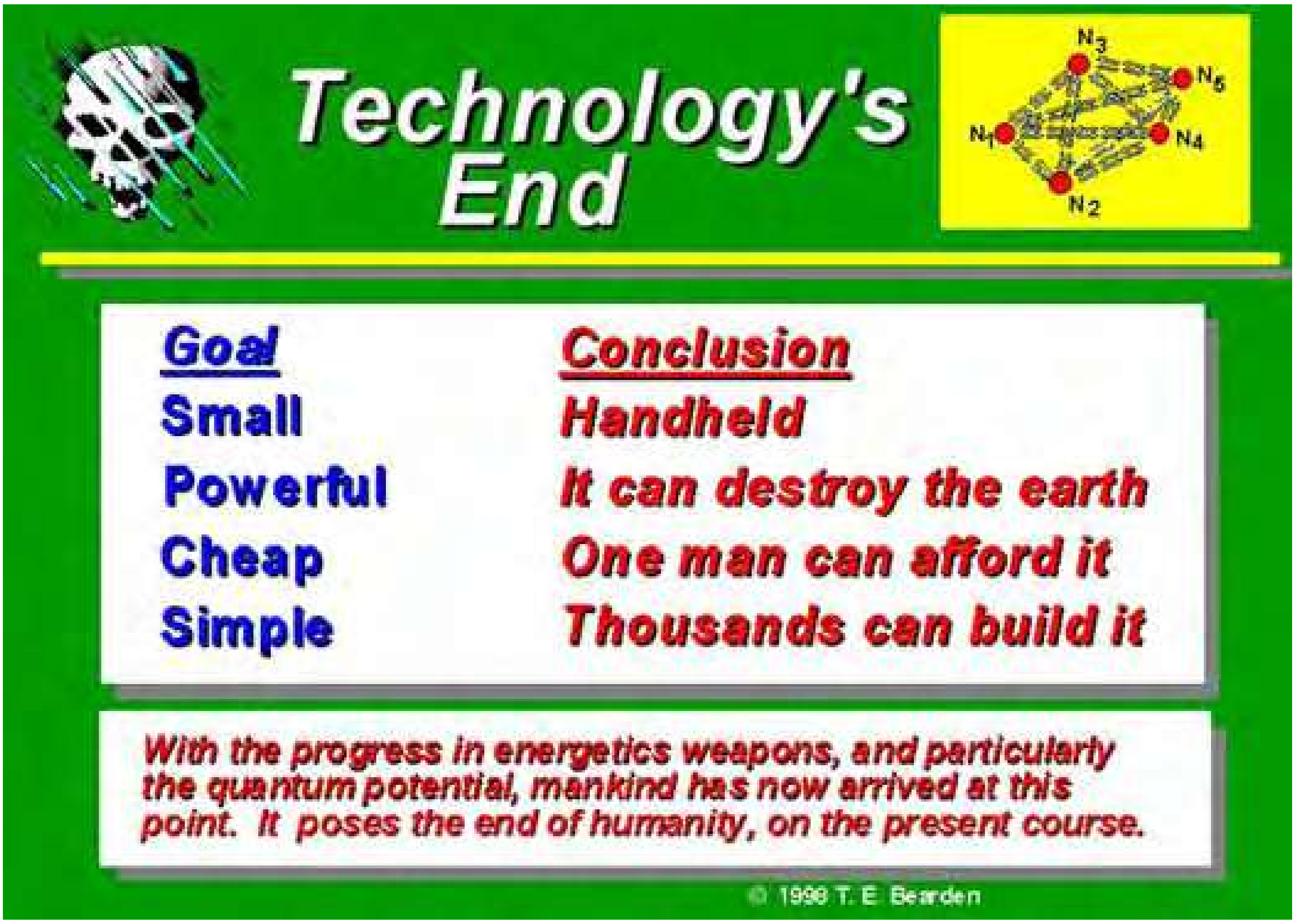
N3
N5
N1
N4
N2

Technology's End

Goal	Conclusion
Small	Handheld
Powerful	It can destroy the earth
Cheap	One man can afford it
Simple	Thousands can build it

With the progress in energetics weapons, and particularly the quantum potential, mankind has now arrived at this point. It poses the end of humanity, on the present course.

© 1998 T. E. Bearden

Return to Slide Index

Next Slide

Previous Slide

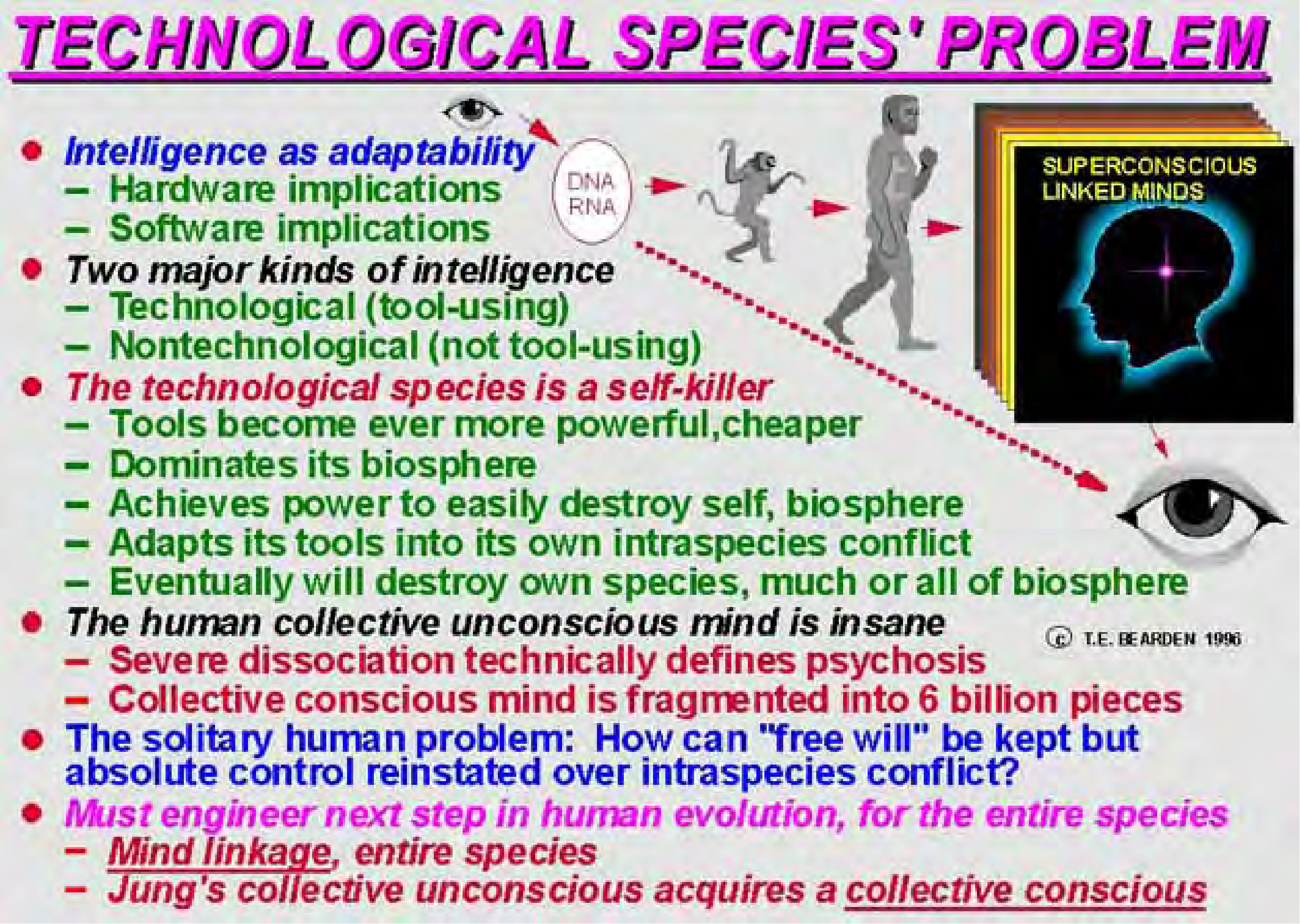

TECHNOLOGICAL SPECIES' PROBLEM

DNA
RNA
SUPERCONSCIOUS
LINKED MINDS

- *Intelligence as adaptability*
  - Hardware implications
  - Software implications
- *Two major kinds of intelligence*
  - Technological (tool-using)
  - Nontechnological (not tool-using)
- *The technological species is a self-killer*
  - Tools become ever more powerful,cheaper
  - Dominates its biosphere
  - Achieves power to easily destroy self, biosphere
  - Adapts its tools into its own intraspecies conflict
  - Eventually will destroy own species, much or all of biosphere
- *The human collective unconscious mind is insane*
  - Severe dissociation technically defines psychosis
  - Collective conscious mind is fragmented into 6 billion pieces
- The solitary human problem: How can "free will" be kept but absolute control reinstated over intraspecies conflict?
- *Must engineer next step in human evolution, for the entire species*
  - *Mind linkage, entire species*
  - *Jung's collective unconscious acquires a collective conscious*

© T.E. BEARDEN 1996

Return to Slide Index

Previous Slide

The Missing Infolded Electro-dynamics

Contents

1. The Missing Infolded Electrodynamics
2. Major Points
3. Major Points (cont.)
4. MTW's General Relativity Principle
5. Extended General Relavity Principle
6. Active Spacetime and Specific Vacuum Engine
7. Electrodynamicists React to Removal of Material Ether
8. Reaction to Loss of Material Ether
9. Massless Spacetime Perturbations Not Addressed
10. Bunge on Status of Electrodynamics
11. Bunge (continued)
12. Poynting Energy Flow Along Two Parallel Conductors
13. Fields and potentials defined only in/on charged matter
14. Wave in String (Transverse)
15. Launching EM Wave from Wire Antenna
16. Phase Conjugate Wavepairs Produce New Waves
17. Time-Polarized (Scalar) EM Wave
18. Mechanism for flow of time
19. Present systems coefficient of performance (COP)<1.0
20. Bridging for Overunity
21. Free Energy Generator
22. Demonstration of Overunity Concept
23. Altered Spacetime Curvatures
24. Interaction of Mass and Spacetime Curvature
25. Composition of Scalar Potential
26. Lorenz Closed Surface Integration
27. How can they use that old physically insignificant wind?!!
28. Vacuum Engines: Spacetime Demons
29. Undistorted Progressive Waves (1)
30. Undistorted Progressive Waves (2)
31. Other Superluminal Waves
32. Becker's Experiments: Healing Bone Fractures
33. Producing a Specific Spacetime Curvature Engine
34. Pumping A Phase Conjugate with Transverse EM

35. Differentiation and Dedifferentiation
36. First Step to Cancer: Hypoxia
37. Longitudinal Photon Interaction with Charge
38. General Relativity in Particle Physics
39. Pepper* on time-reversing mass
40. Pepper* on time-reversing mass (2)
41. Pepper* on time-reversing mass (3)
42. Chung's Negative Resistor
43. China Lake Anomalies
44. Time Reversal (TR) Zone
45. Electronuclear Reactions in TR-Zones (1)
46. Electronuclear Reactions in TR-Zones (2)
47. Electronuclear Reactions in TR-Zones (3)
48. Nuclear Reactions in Time-Reversal Zones
49. Nuclear reactions are biased
50. Dynamics may be reversed
51. Significant only after time-density charging
52. Curved Spacetime as source or sink
53. Anomalous China Lake Phenomena (1)
54. Anomalous China Lake 2
55. Anomalous China Lake 3
56. Anomalous China Lake 4
57. Consistent Anomalies 1
58. Consistent Anomalies 2
59. LW Emission in Electrolysis
60. New Energy Conservation Law
61. Carcinogens & EM Radiation
62. Carcinogens & EM Radiation 1
63. Chronic Disease Puzzle
64. Cancer characteristics
65. Solving Cancer Enigma
66. Cancer Characteristics
67. Portable Prioré Unit
68. Lisitsyn: Brain Code Broken
69. Anti-engine reverses cell
70. Prioré Methodology
71. Prioré Methodology (2)
72. Pumping cell with LW
73. Extension to Nonlinear Optics
74. Time Reversal of Mass to a Form Not in Its Past
75. Cellular R&R System

76. Portable Prioré Treatment System
77. Portable System Block Diagram
78. Absorption and Emission
79. Absorption and Emission (2)
80. EM Waves in Spacetime
81. Cellular Evolution
82. Becker's Regeneration System
83. Becker's Control System Responds to Injury
84. Casualties Surviving Conventional Treatment
85. "Time as fuel" is the future

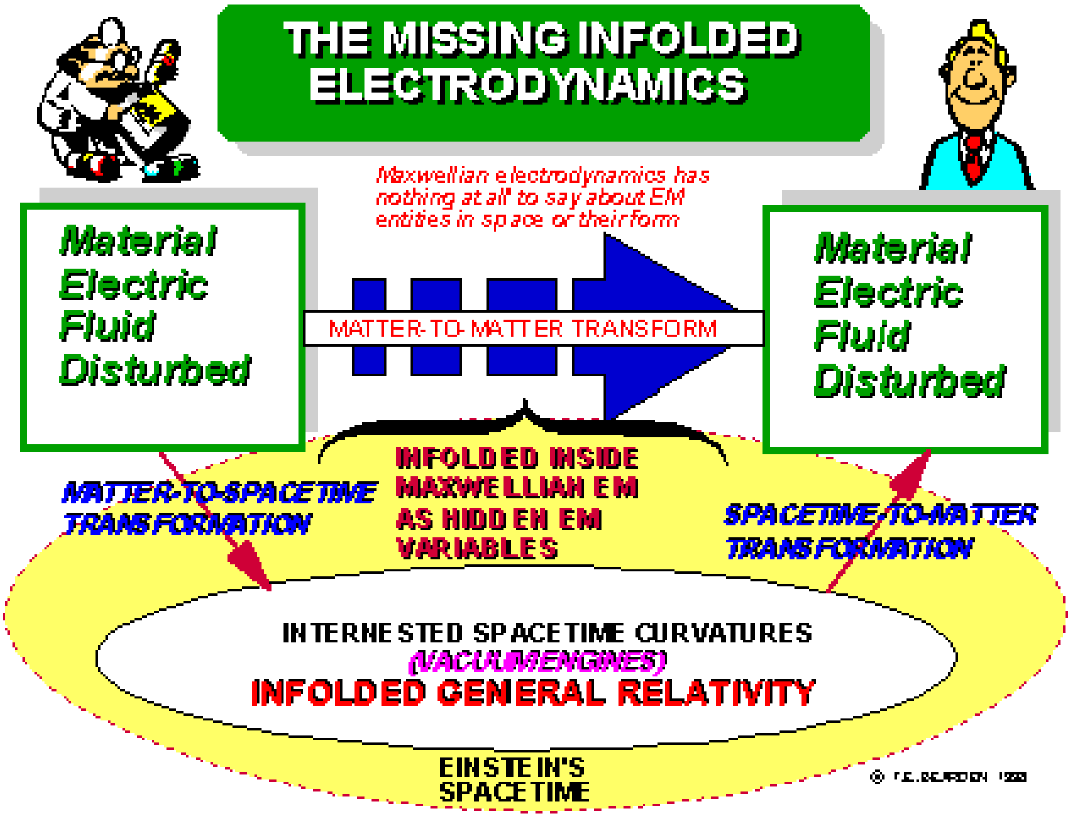

THE MISSING INFOLDED ELECTRODYNAMICS

Material Electric Fluid Disturbed

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN [illegible]

Published by Lotus® Freelance Graphics®

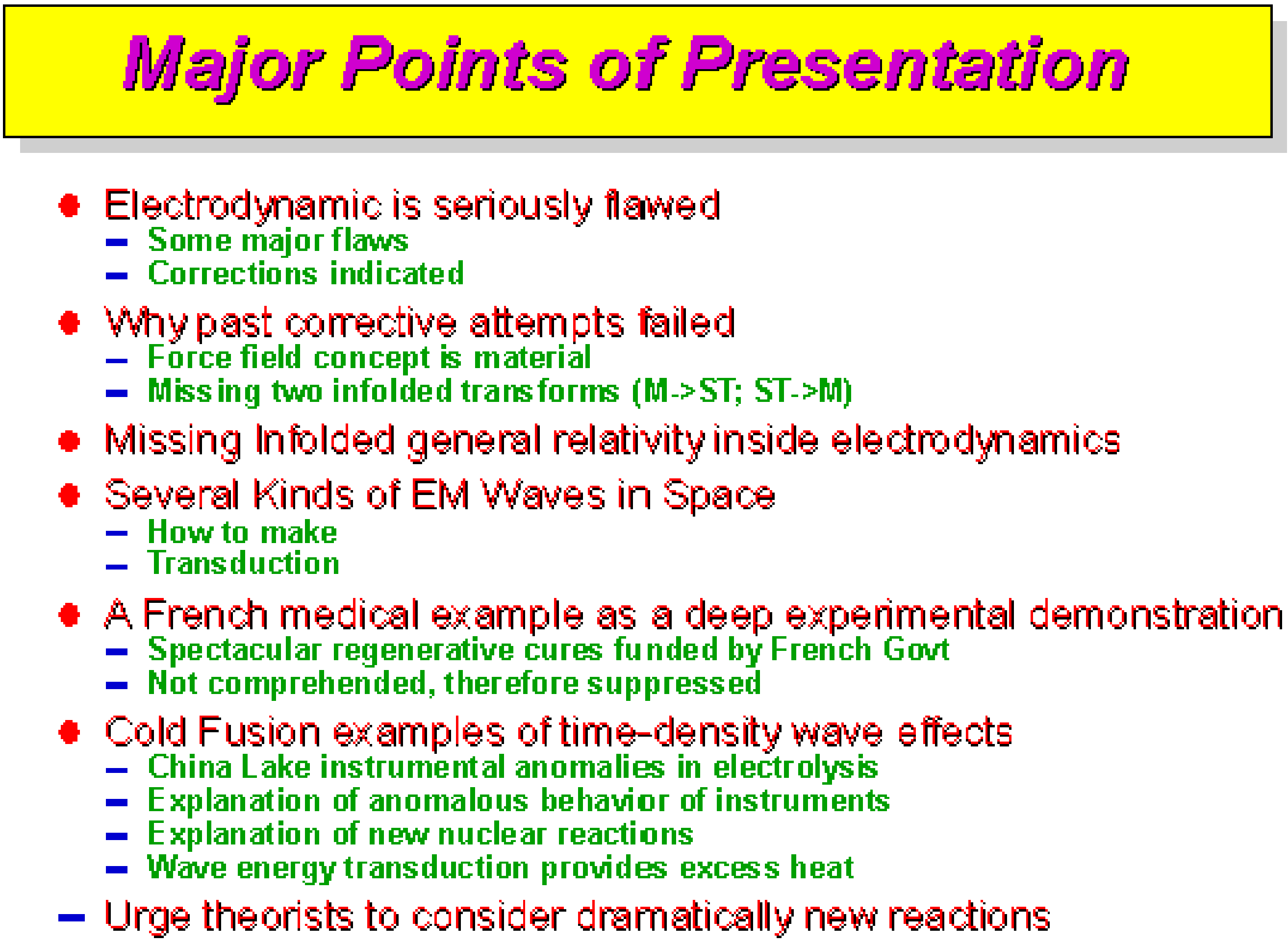

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing Infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

Published by Lotus® Freelance Graphics®

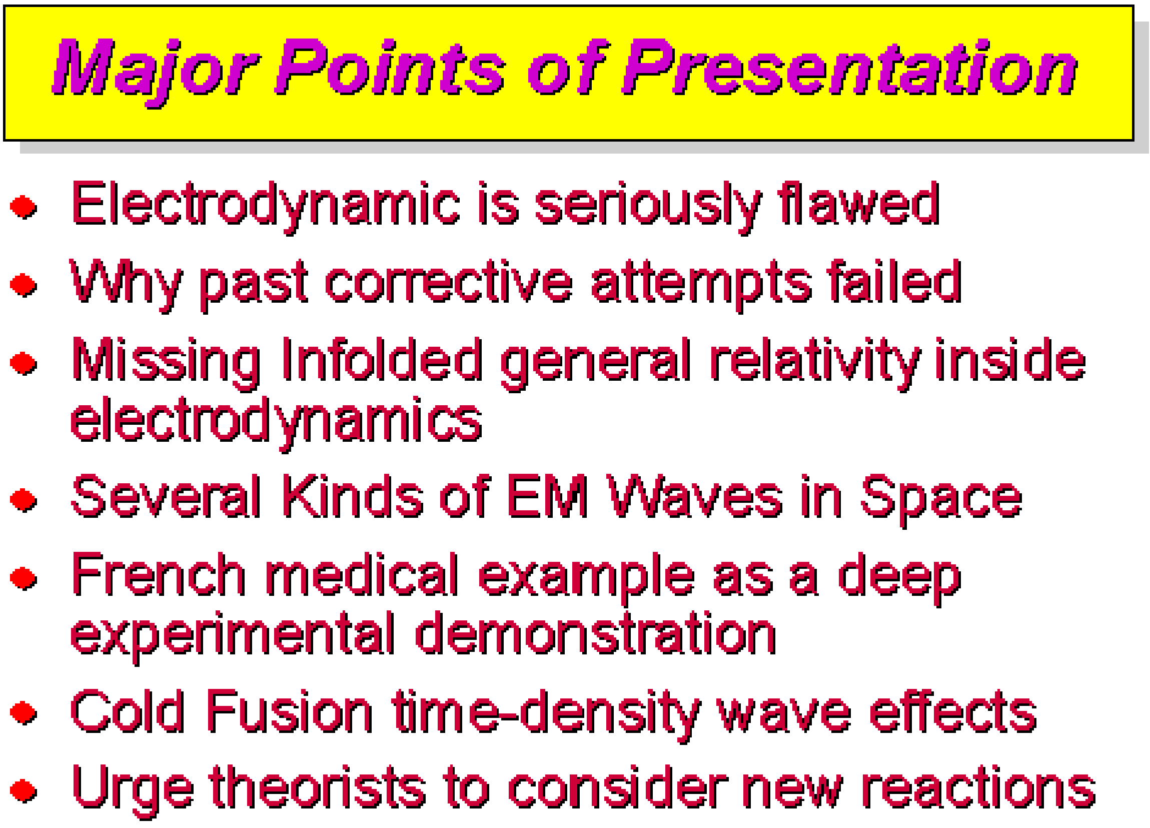

Major Points of Presentation

- Electrodynamic is seriously flawed
- Why past corrective attempts failed
- Missing Infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
- French medical example as a deep experimental demonstration
- Cold Fusion time-density wave effects
- Urge theorists to consider new reactions

Published by Lotus® Freelance Graphics®

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

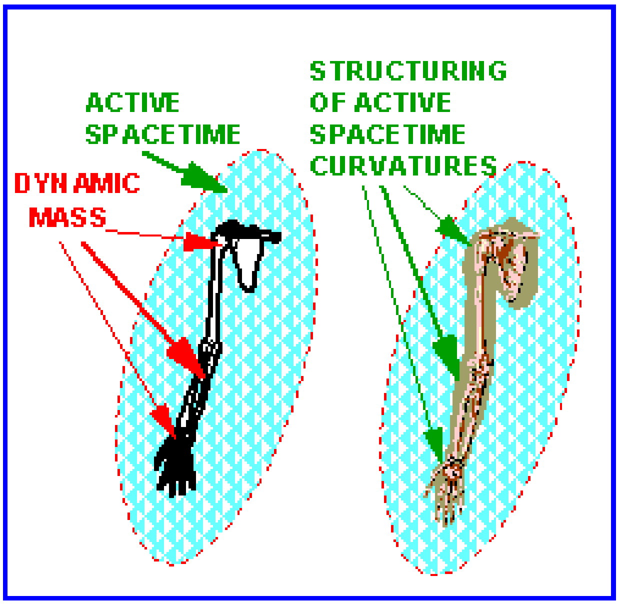
ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E.BEARDEN [illegible]

Published by Lotus® Freelance Graphics®

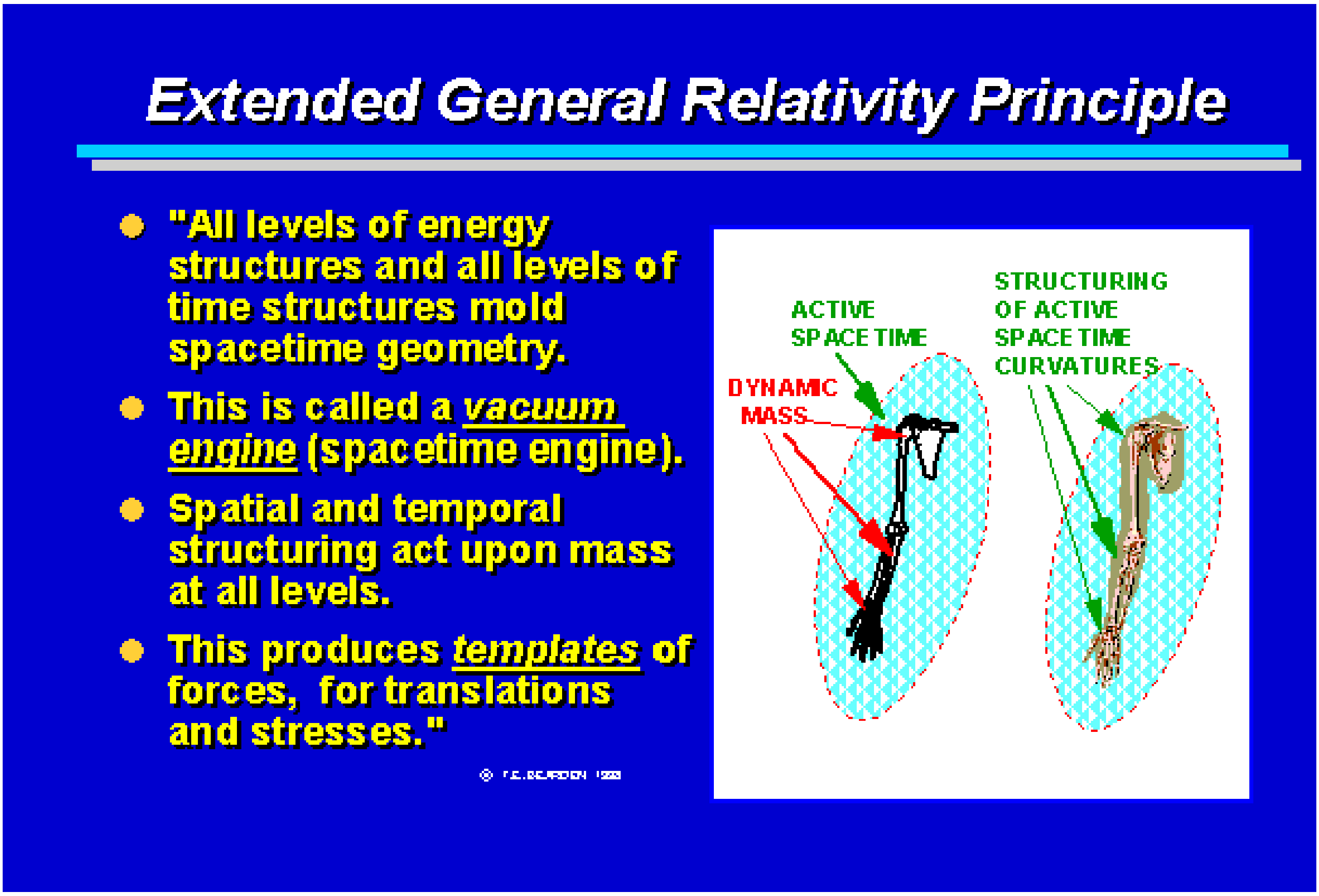

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a *vacuum engine* (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces *templates* of forces, for translations and stresses."

ACTIVE
SPACE TIME
DYNAMIC
MASS
STRUCTURING
OF ACTIVE
SPACE TIME
CURVATURES

Published by Lotus® Freelance Graphics®

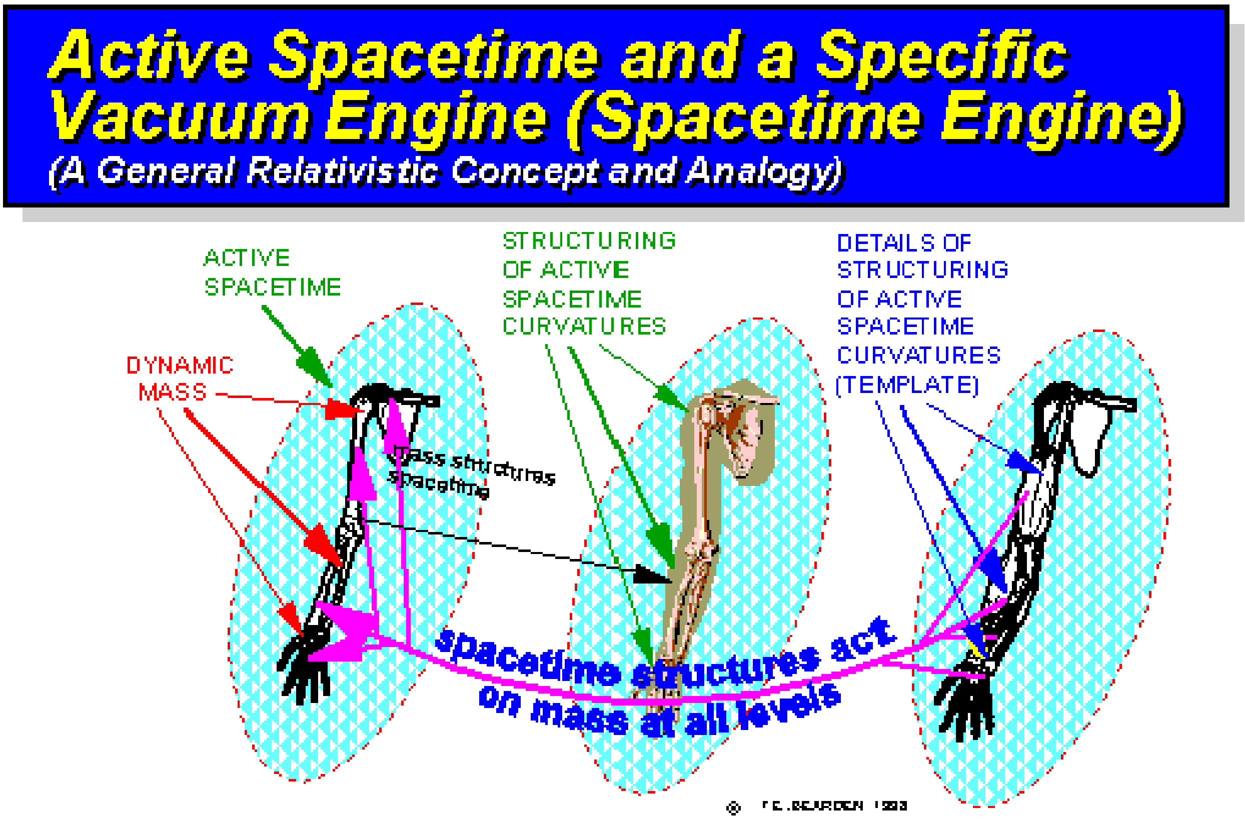

Active Spacetime and a Specific Vacuum Engine (Spacetime Engine)

(A General Relativistic Concept and Analogy)

ACTIVE SPACETIME
DYNAMIC MASS
mass structures spacetime
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)
spacetime structures act on mass at all levels
© T.E. BEARDEN 1993

Published by Lotus® Freelance Graphics®

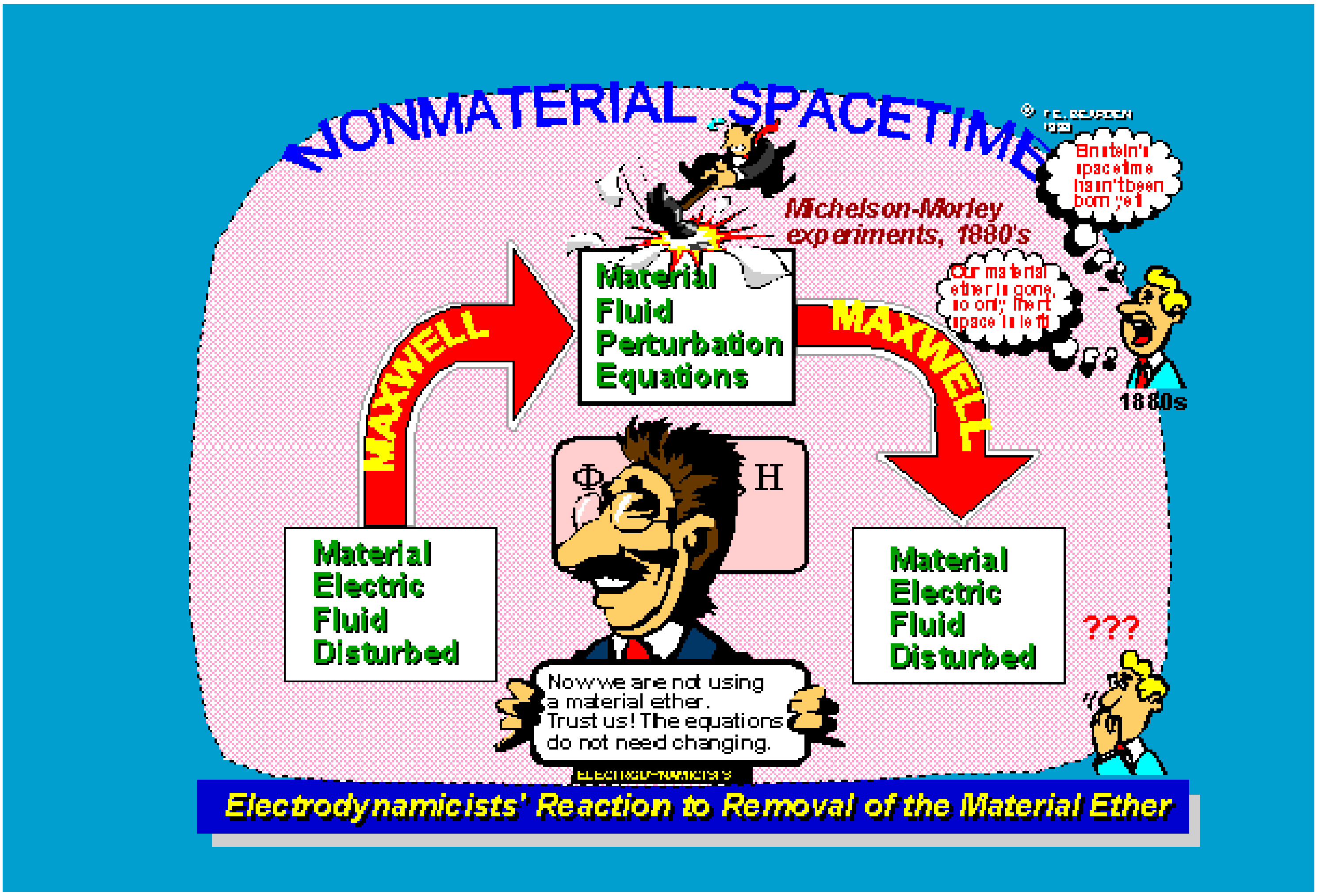
NONMATERIAL SPACETIME
© T.E. BEARDEN
Einstein's spacetime hasn't been born yet!
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left!
1880s
Material Fluid Perturbation Equations
MAXWELL
MAXWELL
Φ
H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS
Electrodynamicists' Reaction to Removal of the Material Ether

---

Published by Lotus® Freelance Graphics®

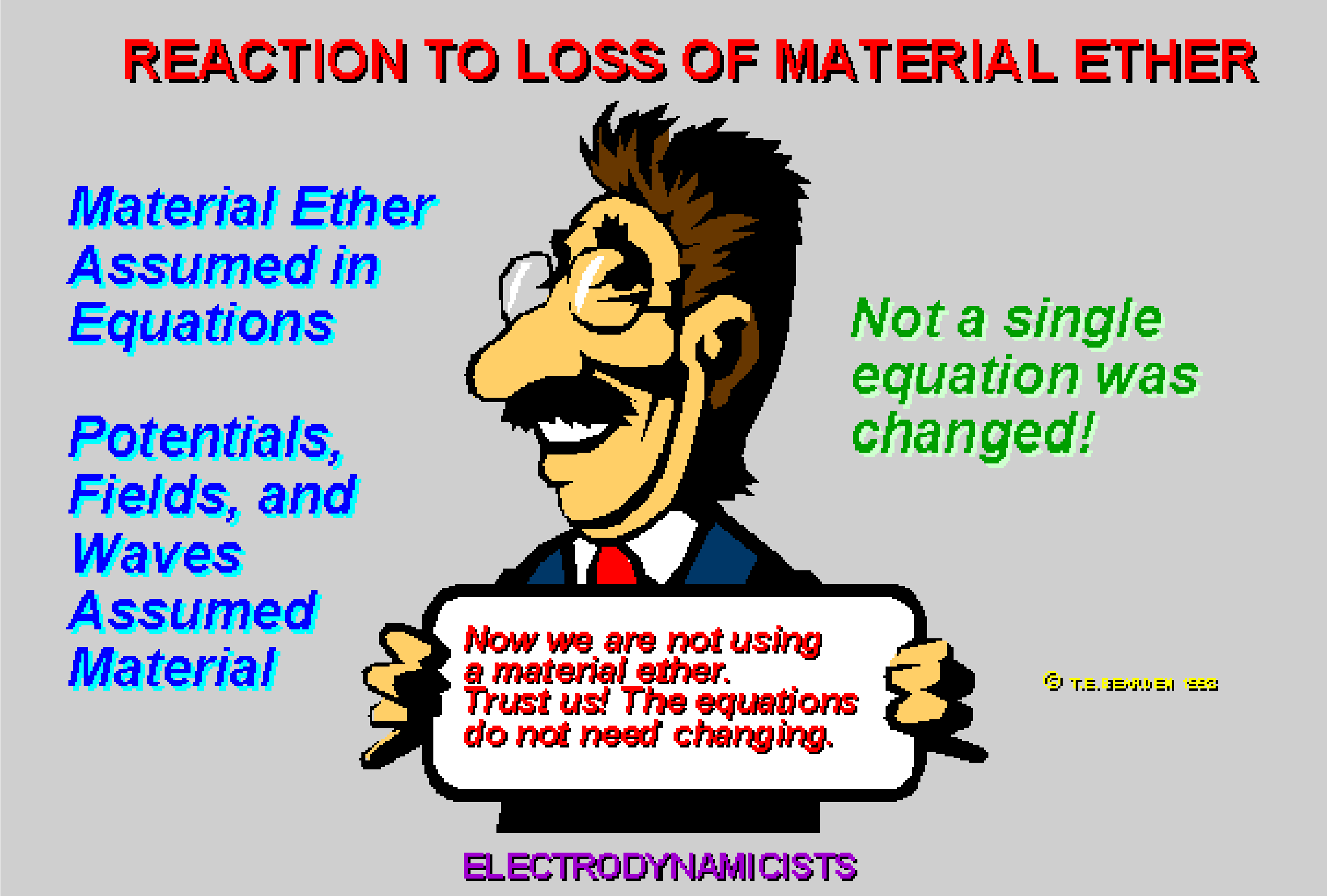

REACTION TO LOSS OF MATERIAL ETHER

Material Ether Assumed in Equations

Potentials, Fields, and Waves Assumed Material

Not a single equation was changed!

Now we are not using a material ether. Trust us! The equations do not need changing.

© T.E. BEARDEN 1993

ELECTRODYNAMICISTS

Published by Lotus® Freelance Graphics®

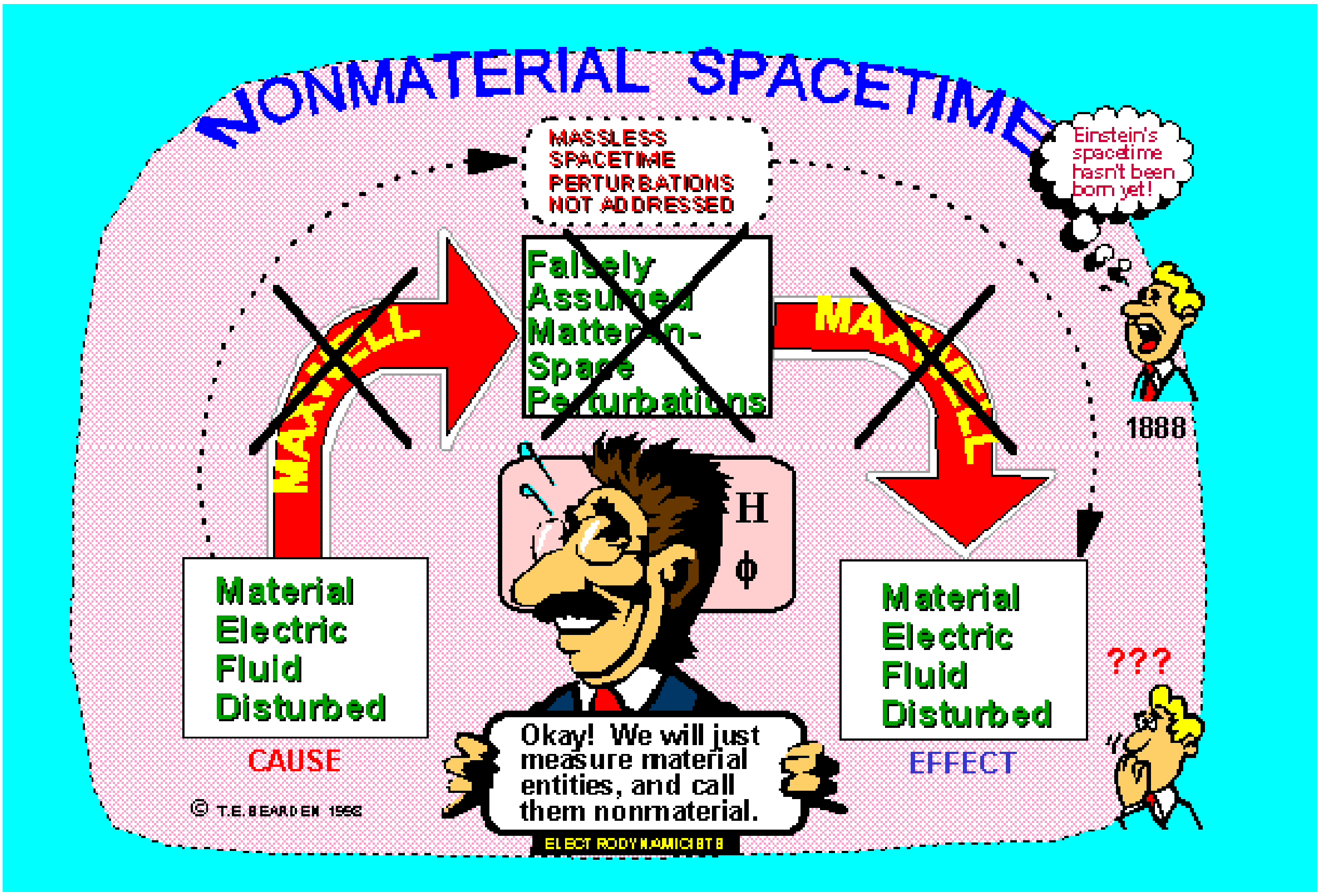
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
Falsely Assumed Matter-in-Space Perturbations
MAXWELL
MAXWELL
1888
H
ϕ
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
CAUSE
EFFECT
© T.E. BEARDEN 1993
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS

Published by Lotus® Freelance Graphics®

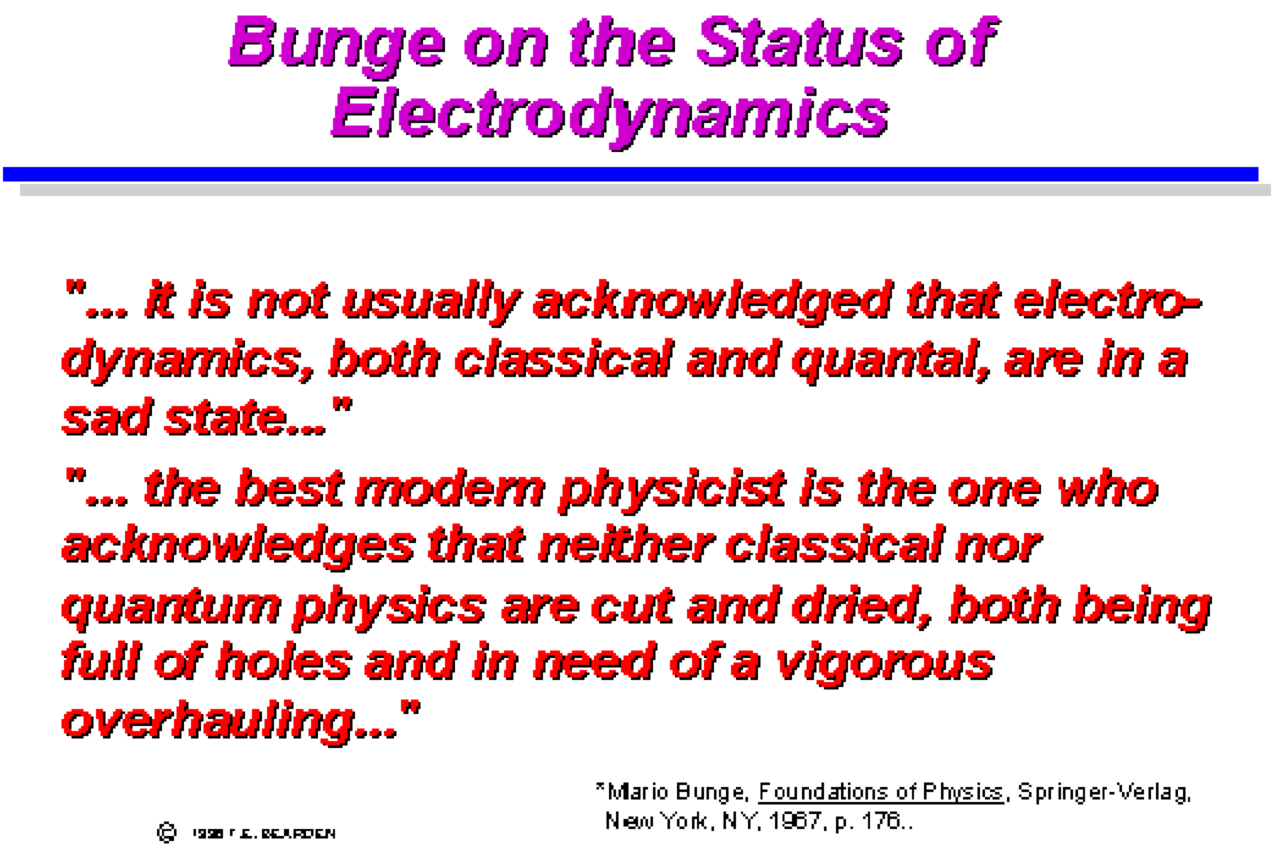

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1998 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

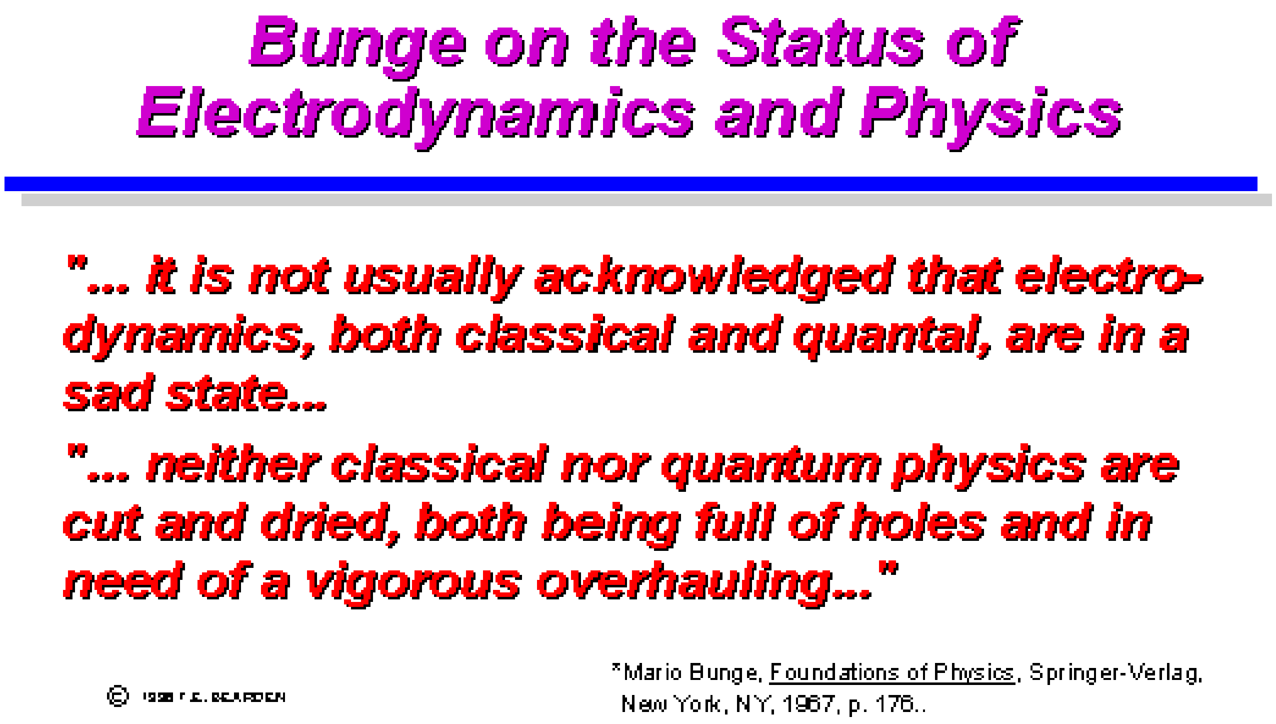

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

© 1998 T.E. BEARDEN

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

Published by Lotus® Freelance Graphics®

Poynting energy flow S along two parallel conductors

Surface charges in conductors form a distributed dipole

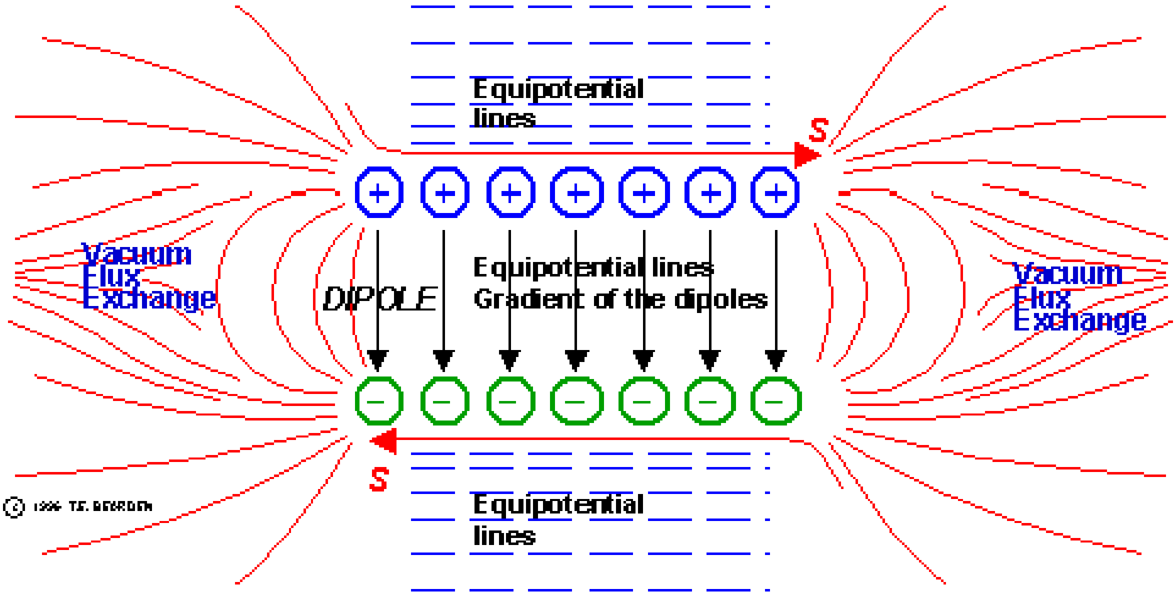
Equipotential lines
S
+
Vacuum Flux Exchange
DIPOLE
Equipotential lines
Gradient of the dipoles
Vacuum Flux Exchange
−
S
© 1996 T.E. Bearden
Equipotential lines

Published by Lotus® Freelance Graphics®

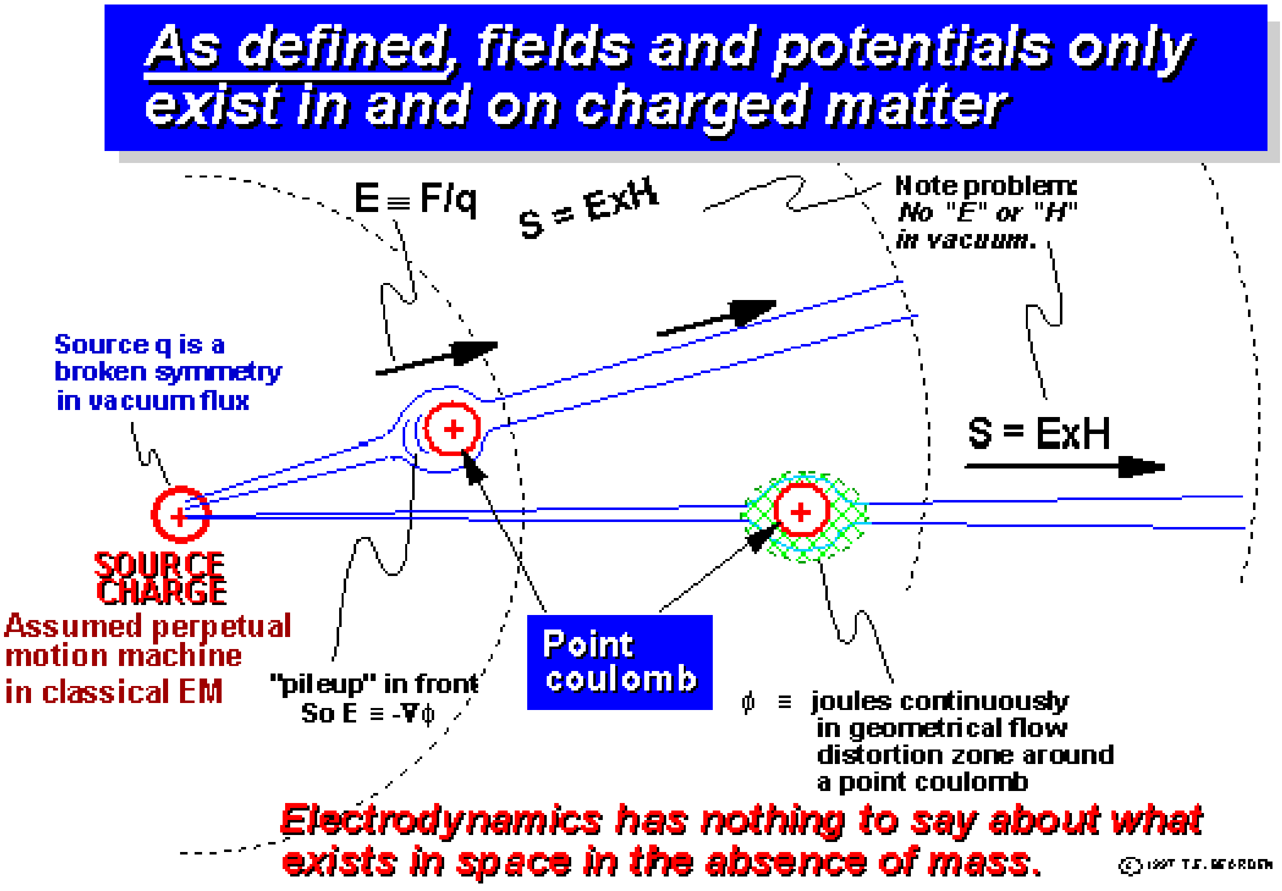

As defined, fields and potentials only exist in and on charged matter

E ≡ F/q
S = ExH
Note problem:
No "E" or "H"
in vacuum.
Source q is a
broken symmetry
in vacuum flux
S = ExH
SOURCE CHARGE
Assumed perpetual
motion machine
in classical EM
"pileup" in front
So E ≡ -∇ϕ
Point
coulomb
ϕ ≡ joules continuously
in geometrical flow
distortion zone around
a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

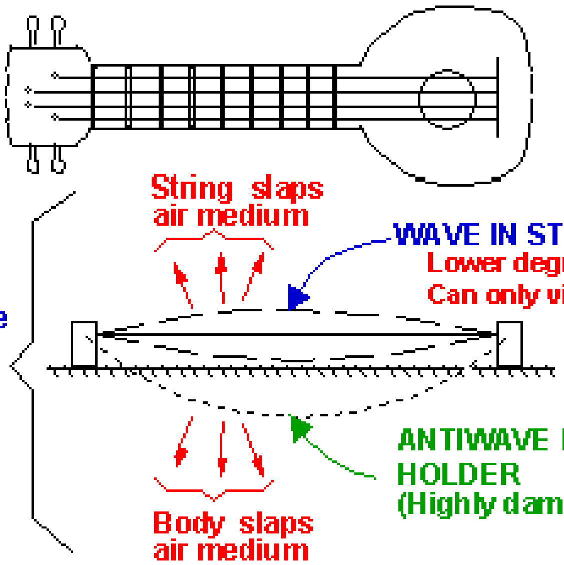

The string wave and the holder wave never leave the string and holder respectively.

Wave is alternatively compressive and rarefactive

String slaps air medium

WAVE IN STRING (TRANSVERSE)
Lower degree of freedom for string.
Can only vibrate mostly laterally.

Drude Electron Gas waves

Nuclear proton waves

ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)

Body slaps air medium

© 1990, 1995 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

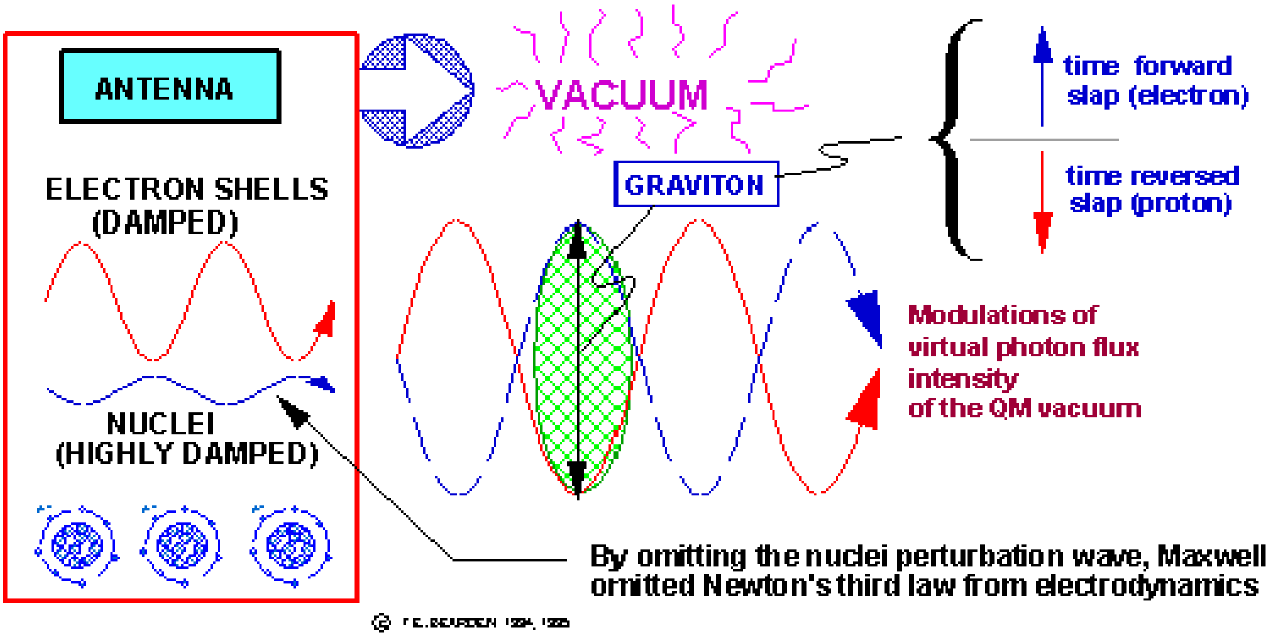
ANTENNA
ELECTRON SHELLS
(DAMPED)
NUCLEI
(HIGHLY DAMPED)
VACUUM
GRAVITON
time forward
slap (electron)
time reversed
slap (proton)
Modulations of
virtual photon flux
intensity
of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

Published by Lotus® Freelance Graphics®

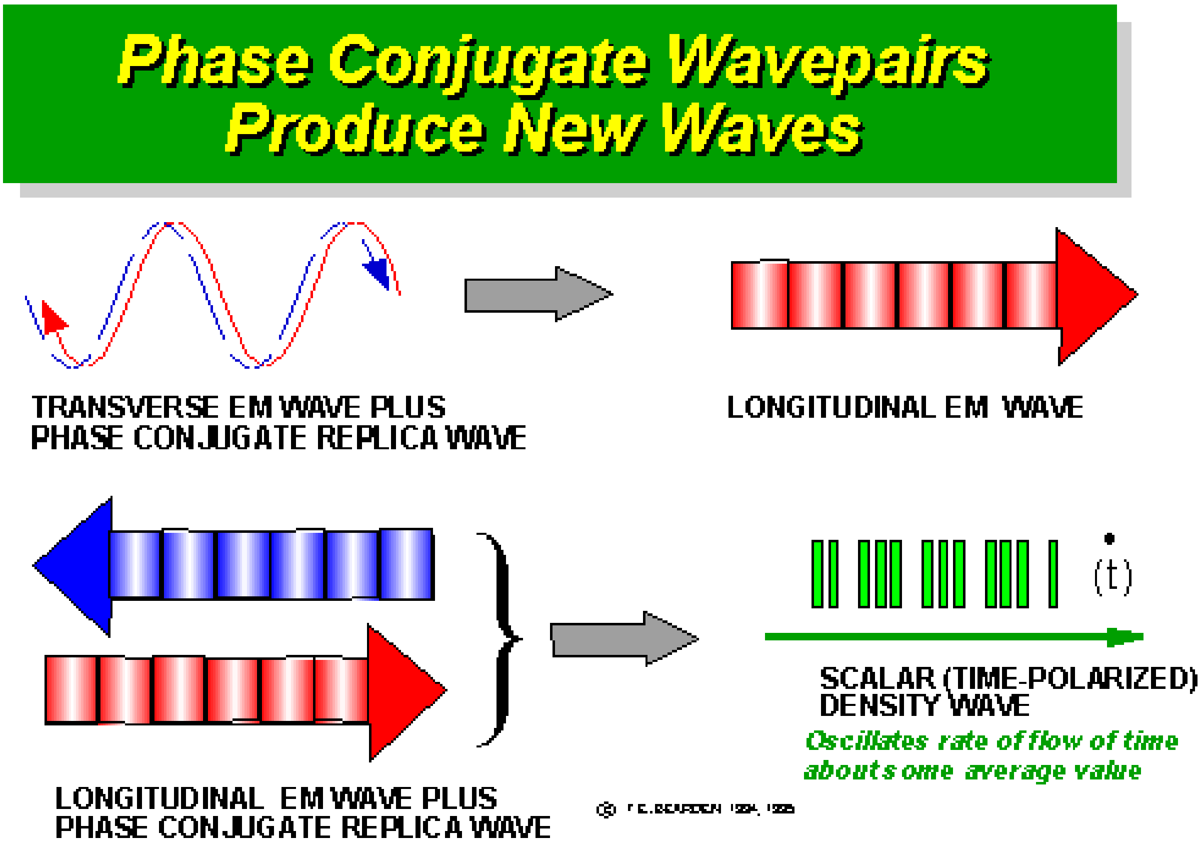

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM WAVE

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

$(\dot{t})$

SCALAR (TIME-POLARIZED)
DENSITY WAVE

*Oscillates rate of flow of time
about some average value*

© T.E. BEARDEN 1994, 1995

Published by Lotus® Freelance Graphics®

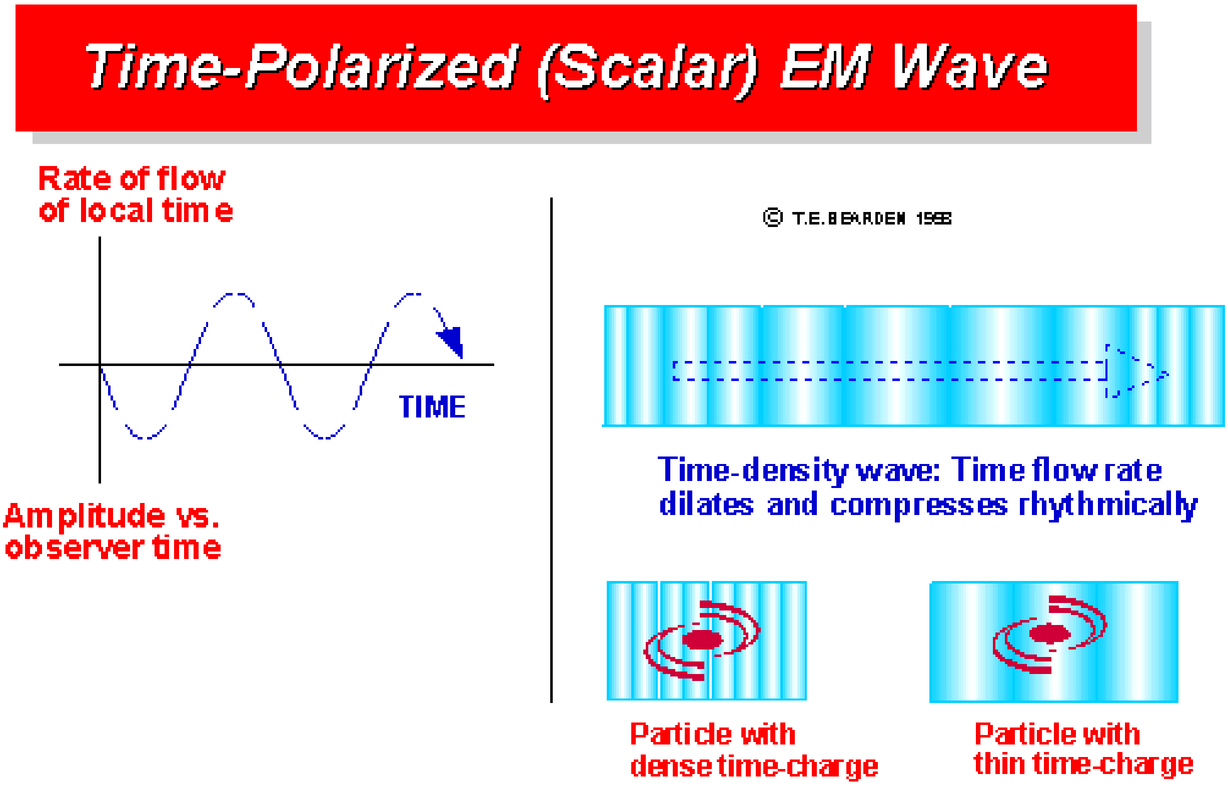

Time-Polarized (Scalar) EM Wave

Rate of flow of local time
TIME
Amplitude vs. observer time

© T.E. BEARDEN 1988

Time-density wave: Time flow rate dilates and compresses rhythmically

Particle with dense time-charge

Particle with thin time-charge

Published by Lotus® Freelance Graphics®

Mechanism for the Flow of Time

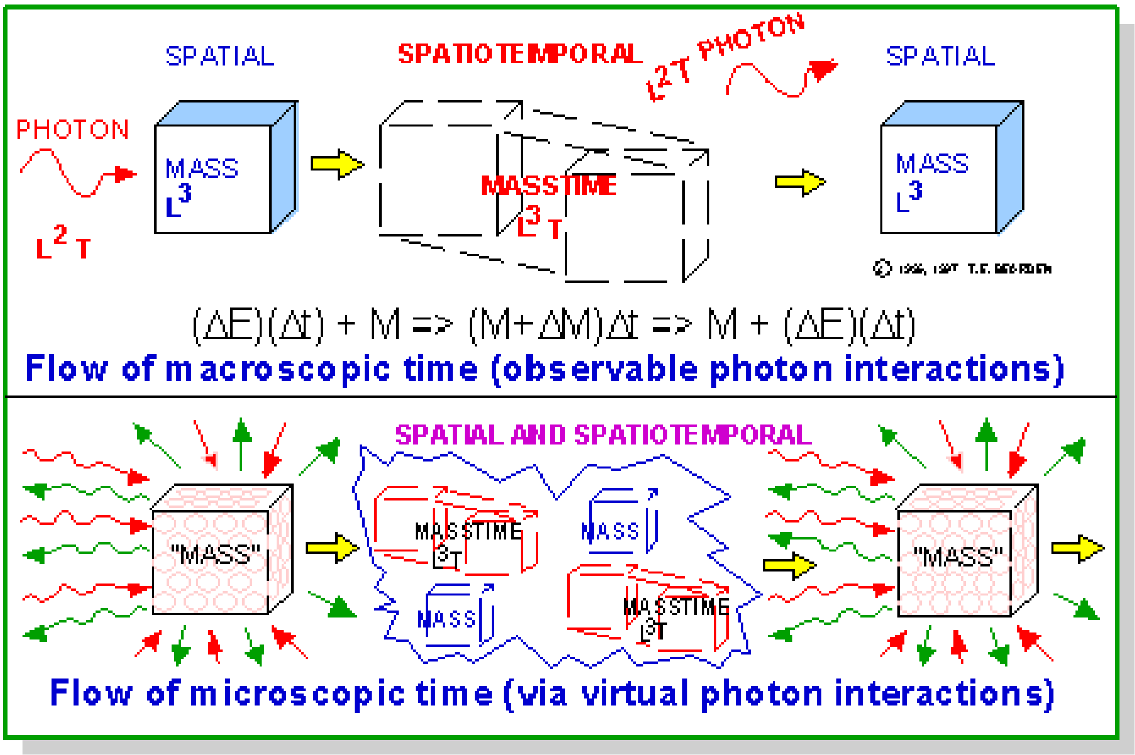
SPATIAL
SPATIOTEMPORAL
L^2T PHOTON
SPATIAL
PHOTON
MASS
L^3
L^2T
MASSTIME
L^3T
MASS
L^3
© 1996, 1997 T.E. BEARDEN
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

Flow of microscopic time (via virtual photon interactions)

Published by Lotus® Freelance Graphics®

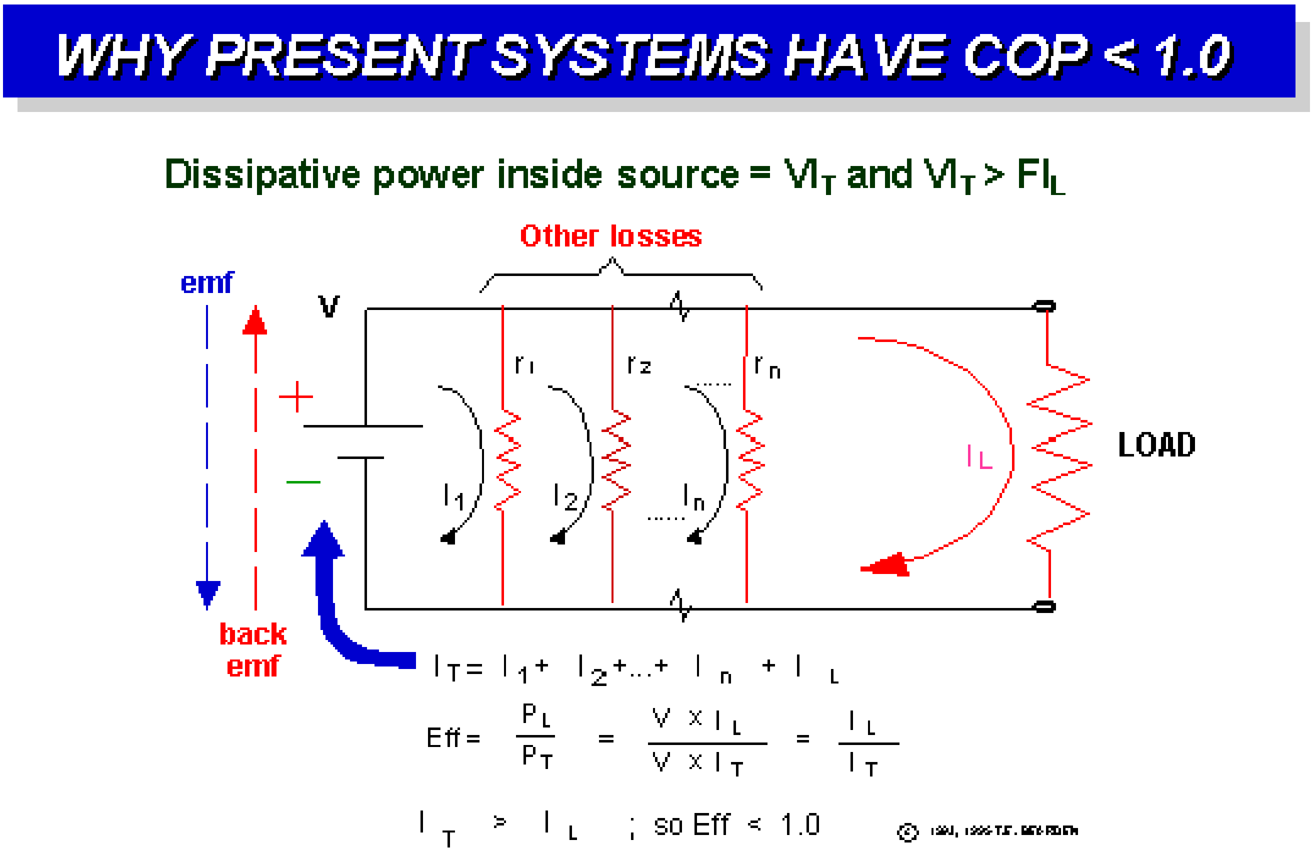

WHY PRESENT SYSTEMS HAVE COP < 1.0

Dissipative power inside source = VI_T and $VI_T > FI_L$

Other losses
emf
V
+
–
back emf
r_1
r_2
r_n
I_1
I_2
I_n
I_L
LOAD

$I_T = I_1 + I_2 + \ldots + I_n + I_L$

$$\text{Eff} = \frac{P_L}{P_T} = \frac{V \times I_L}{V \times I_T} = \frac{I_L}{I_T}$$

$I_T > I_L$; so Eff < 1.0

© 1994, 1996 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

BRIDGING ENABLES COP > 1.0

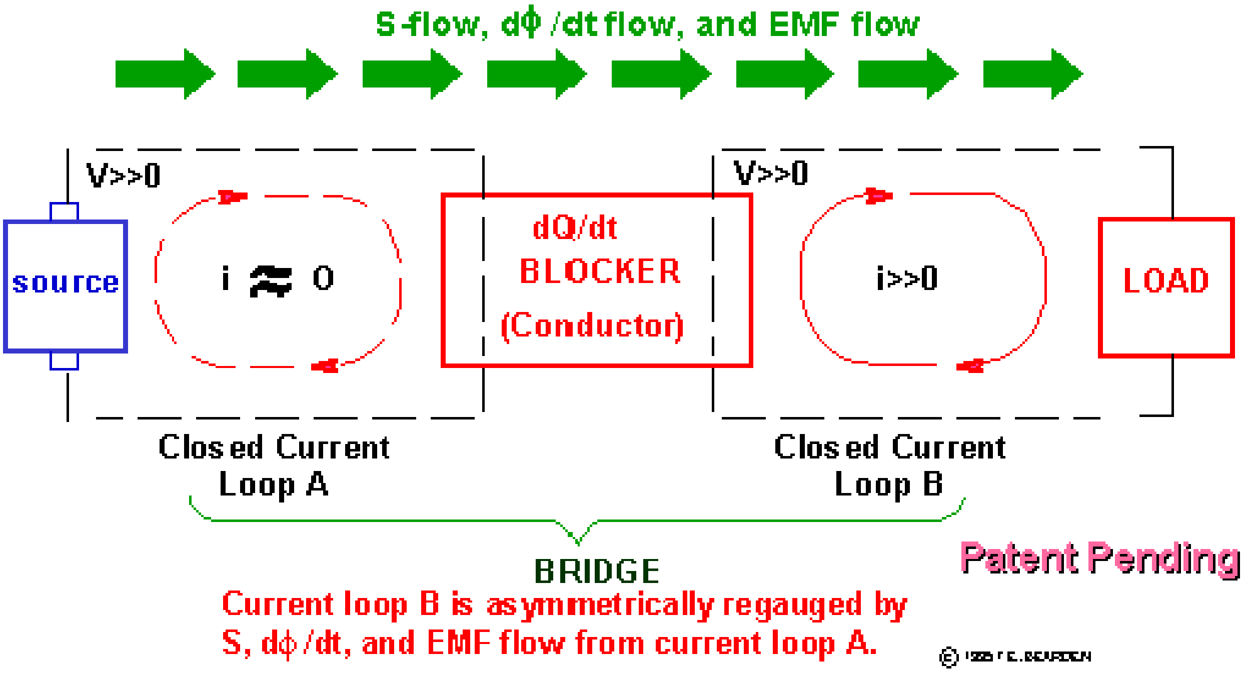
S-flow, dϕ/dt flow, and EMF flow
V>>0
source
i ≈ 0
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD
Closed Current
Loop A
Closed Current
Loop B
BRIDGE
Patent Pending
© [illegible] T.E. BEARDEN

Current loop B is asymmetrically regauged by S, dϕ/dt, and EMF flow from current loop A.

Published by Lotus® Freelance Graphics®

"Static" Poynting free-energy generator

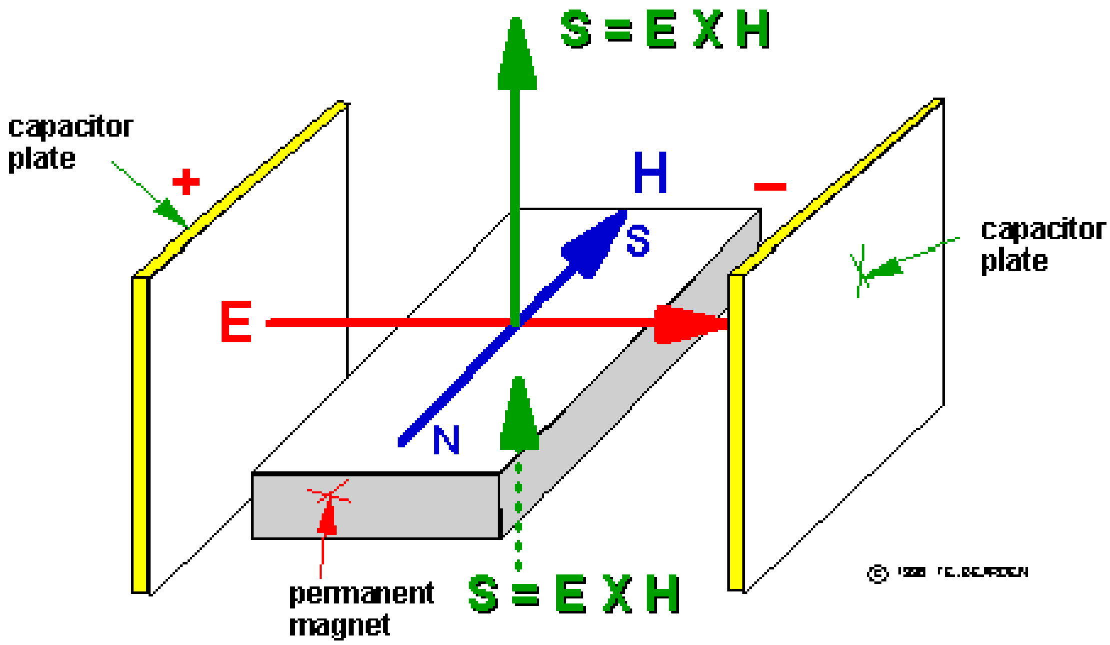
S = E X H
capacitor plate
+
H
–
S
capacitor plate
E
N
permanent magnet
S = E X H
© 1998 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

Demonstration of an Overunity Concept

(By Asymmetrical Self-Regauging)

Patent Pending

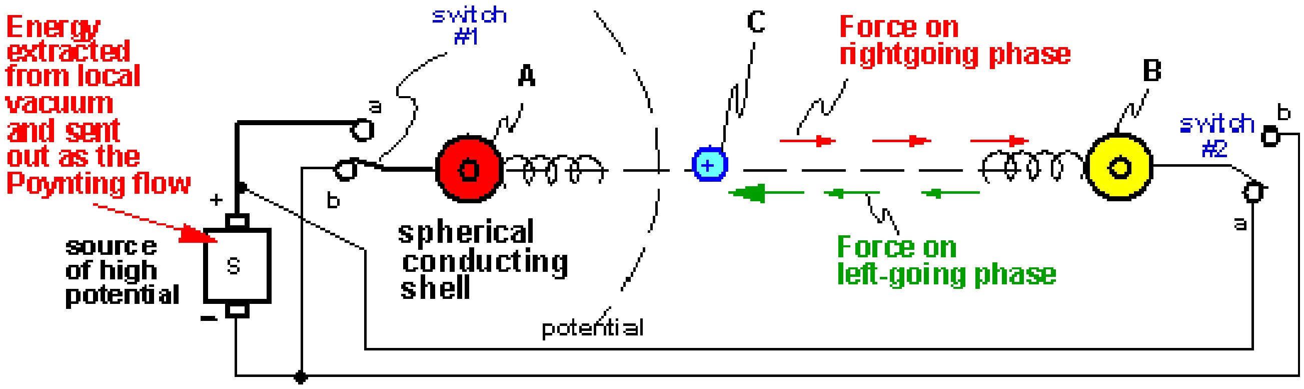
Energy extracted from local vacuum and sent out as the Poynting flow
switch #1
a
b
A
C
Force on rightgoing phase
B
switch #2
b
a
+
source of high potential
S
−
spherical conducting shell
potential
Force on left-going phase

Assumptions: Capacitances of A and B very small;
Charge B large, power-free switching

© T.E. BEARDEN [illegible]

Published by Lotus® Freelance Graphics®

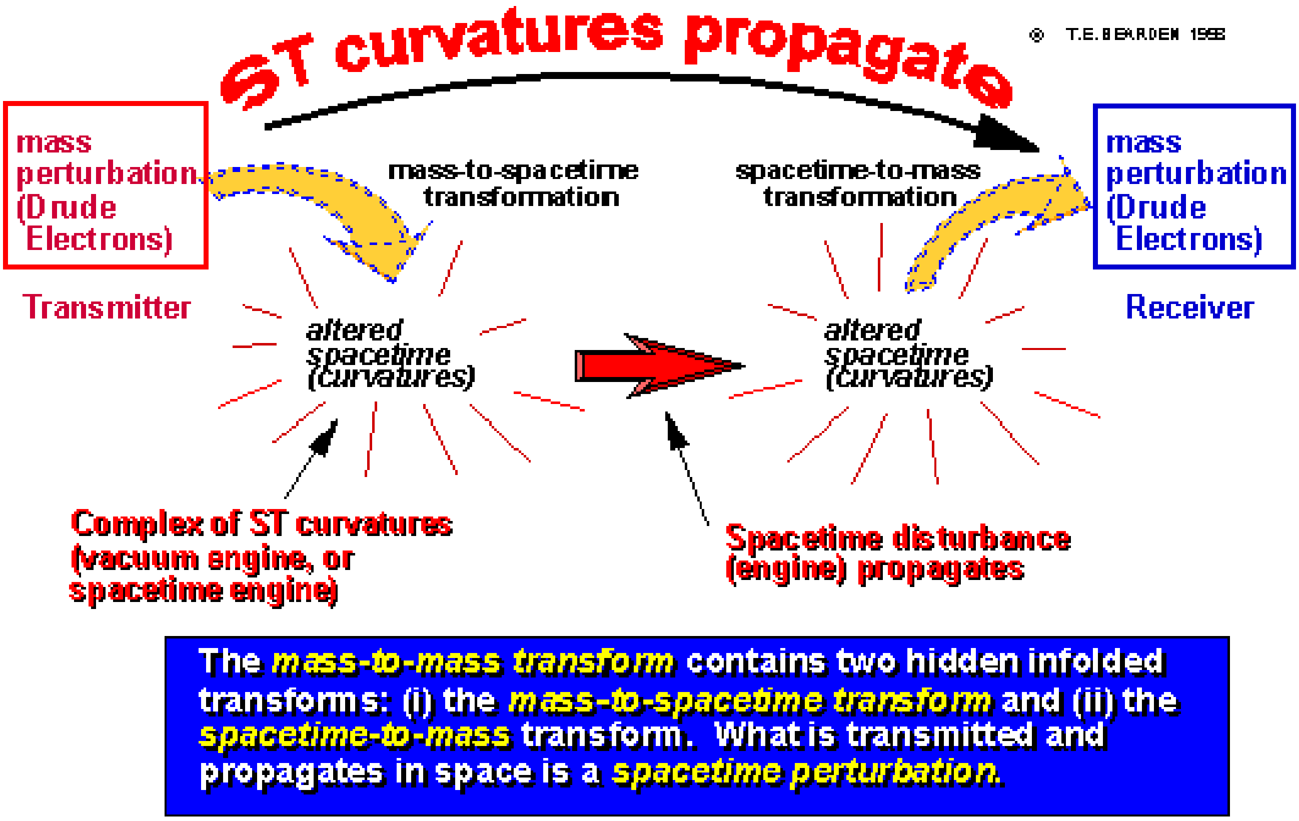
ST curvatures propagate
© T.E. BEARDEN 1998
mass perturbation (Drude Electrons)
Transmitter
mass-to-spacetime transformation
spacetime-to-mass transformation
mass perturbation (Drude Electrons)
Receiver
altered spacetime (curvatures)
altered spacetime (curvatures)
Complex of ST curvatures (vacuum engine, or spacetime engine)
Spacetime disturbance (engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation.*

Published by Lotus® Freelance Graphics®

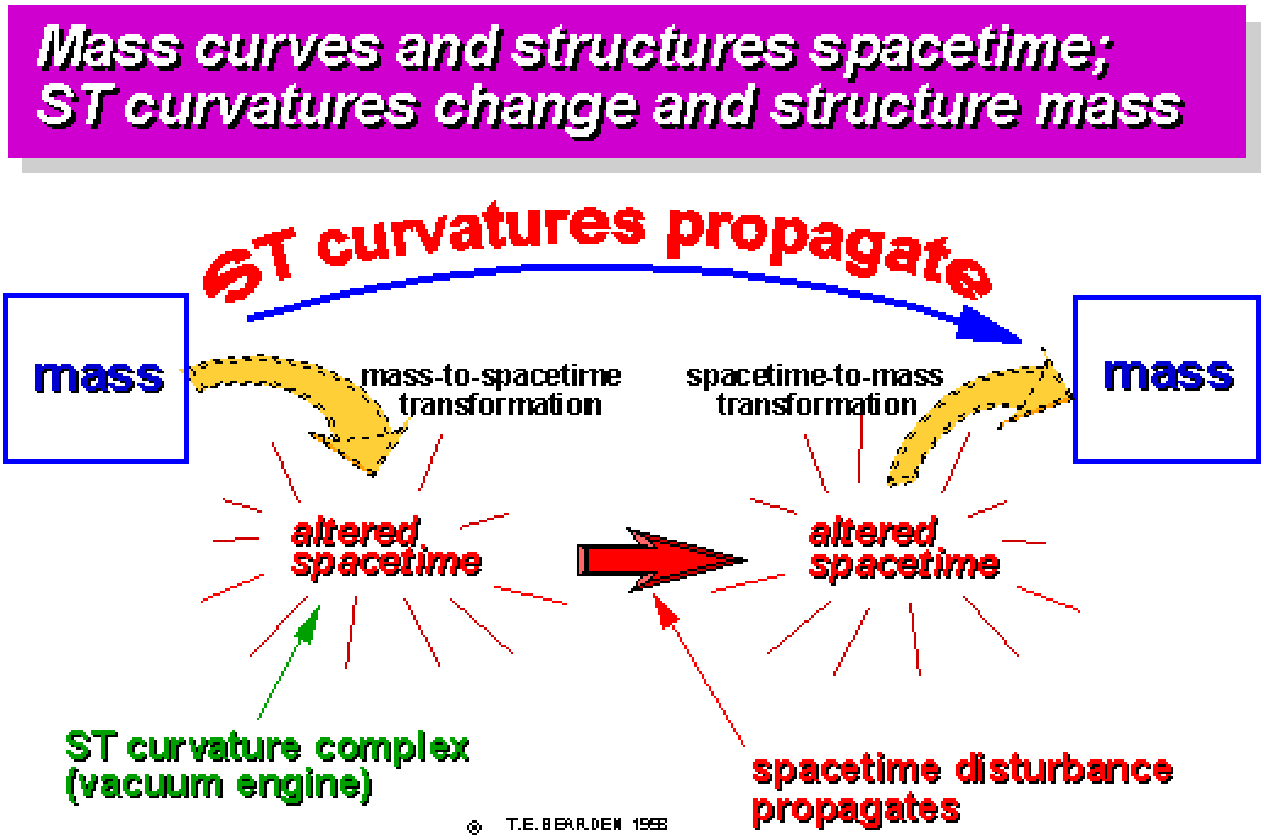

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate
mass
mass-to-spacetime transformation
spacetime-to-mass transformation
mass
altered spacetime
altered spacetime
ST curvature complex (vacuum engine)
spacetime disturbance propagates

© T.E.BEARDEN 1993

Published by Lotus® Freelance Graphics®

Composition of the Scalar Potential ϕ

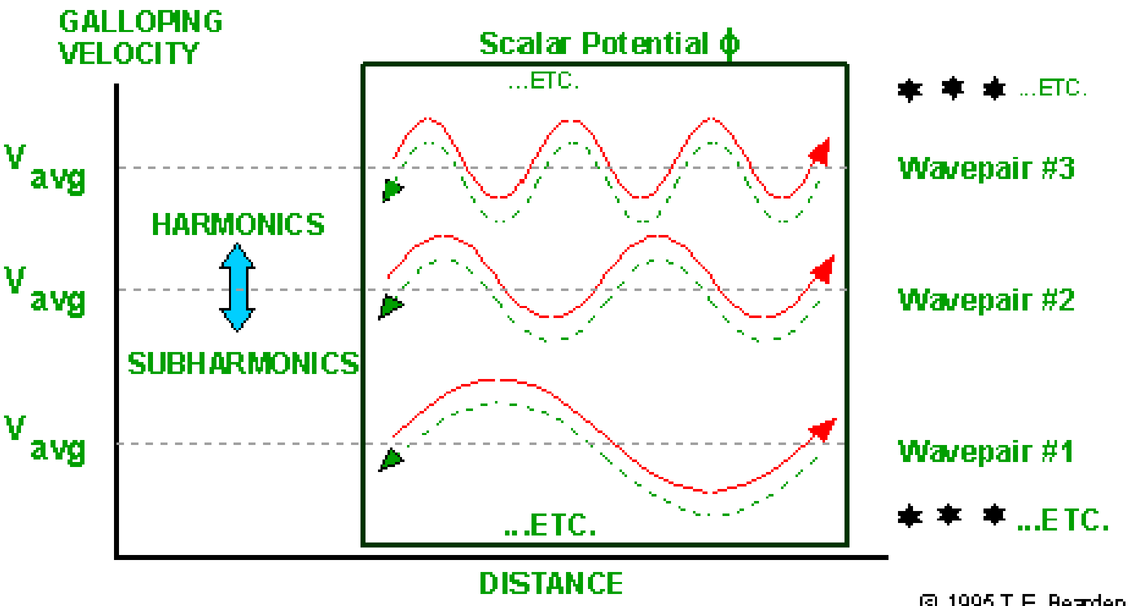
GALLOPING
VELOCITY
Scalar Potential ϕ
...ETC.
✱ ✱ ✱ ...ETC.
V_{avg}
Wavepair #3
HARMONICS
SUBHARMONICS
V_{avg}
Wavepair #2
V_{avg}
Wavepair #1
...ETC.
✱ ✱ ✱ ...ETC.
DISTANCE

A harmonic set of bidirectional longitudinal EM wavepairs.
Each wavepair is also a net time-polarized EM wave.

© 1995 T.E. Bearden

Published by Lotus® Freelance Graphics®

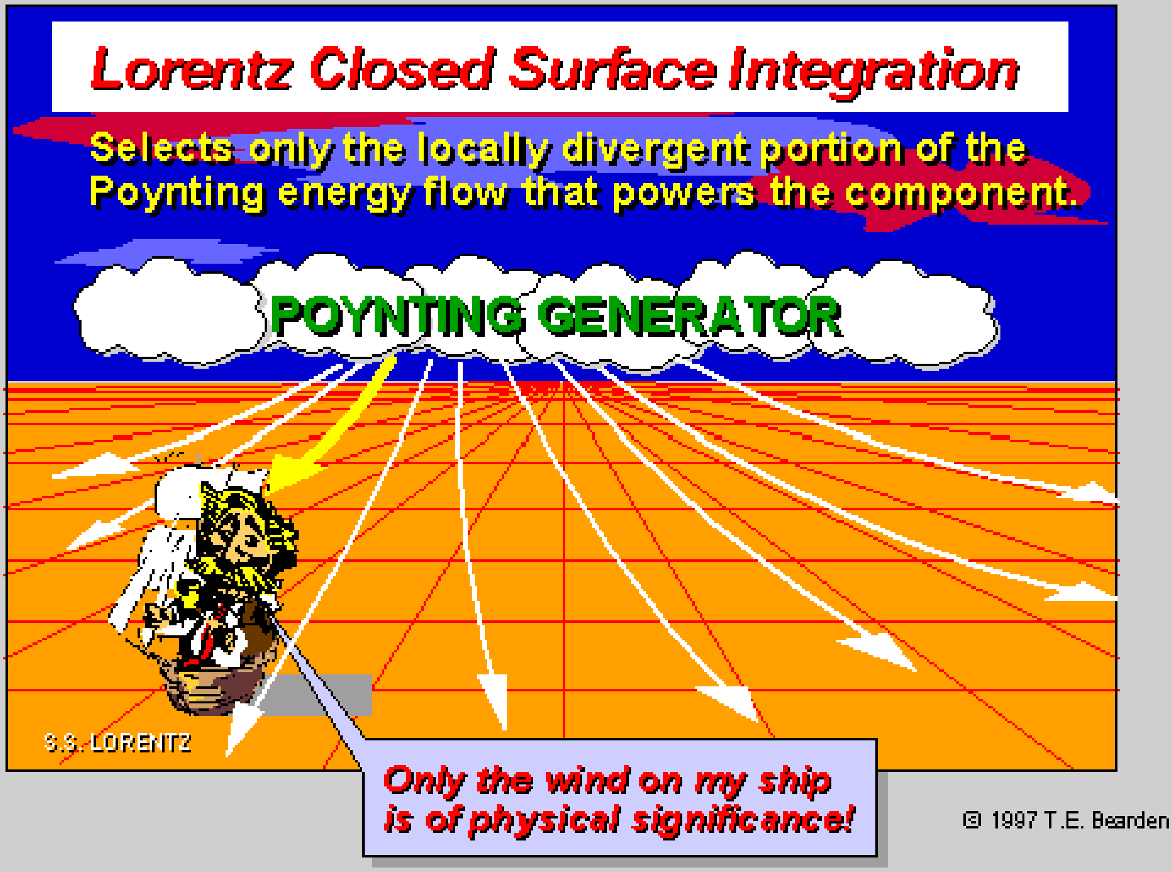

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

Published by Lotus® Freelance Graphics®

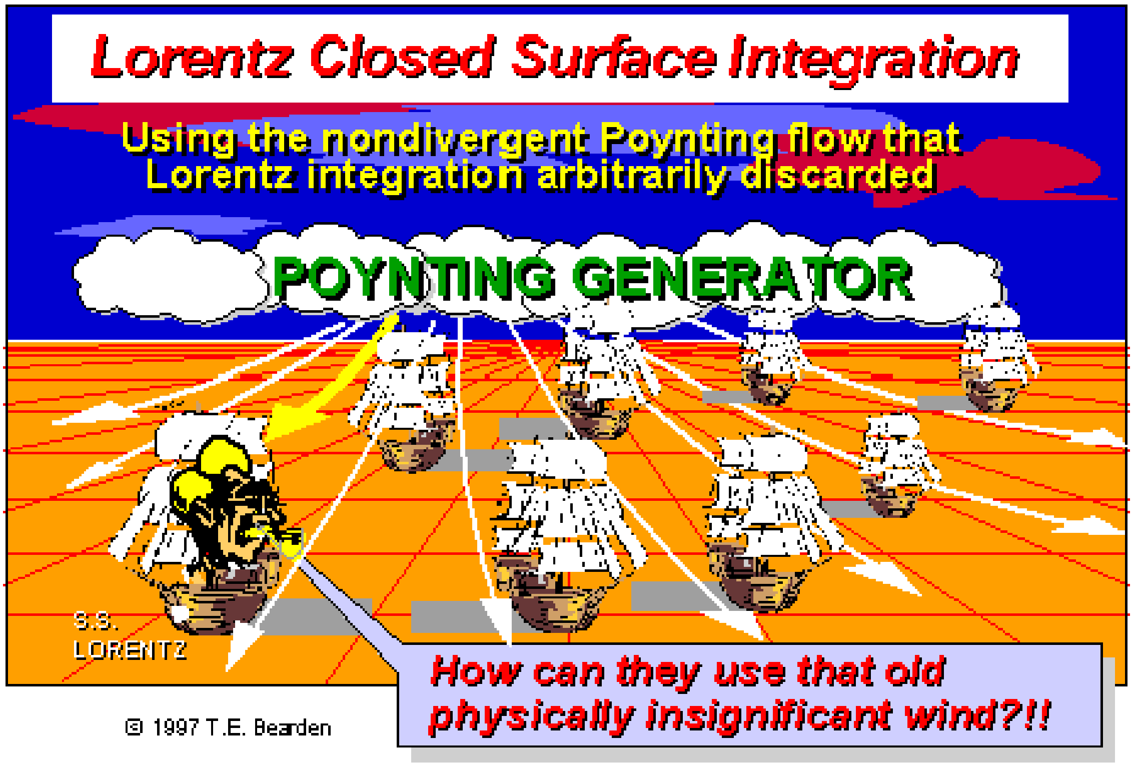

Lorentz Closed Surface Integration

Using the nondivergent Poynting flow that Lorentz integration arbitrarily discarded

How can they use that old physically insignificant wind?!!

© 1997 T.E. Bearden

Published by Lotus® Freelance Graphics®

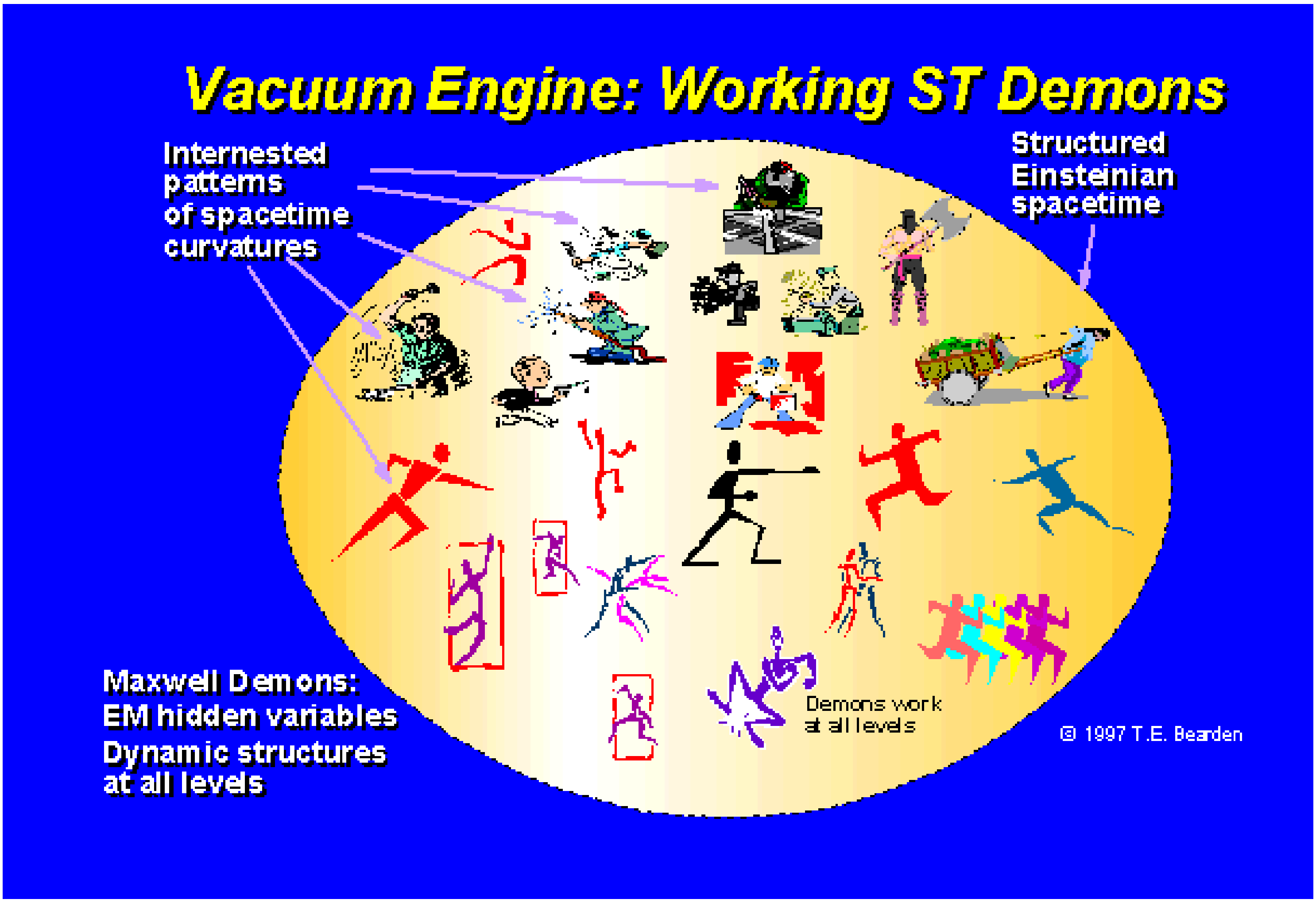

Vacuum Engine: Working ST Demons

Internested
patterns
of spacetime
curvatures

Structured
Einsteinian
spacetime

Maxwell Demons:
EM hidden variables
Dynamic structures
at all levels

Demons work
at all levels

© 1997 T.E. Bearden

Published by Lotus® Freelance Graphics®

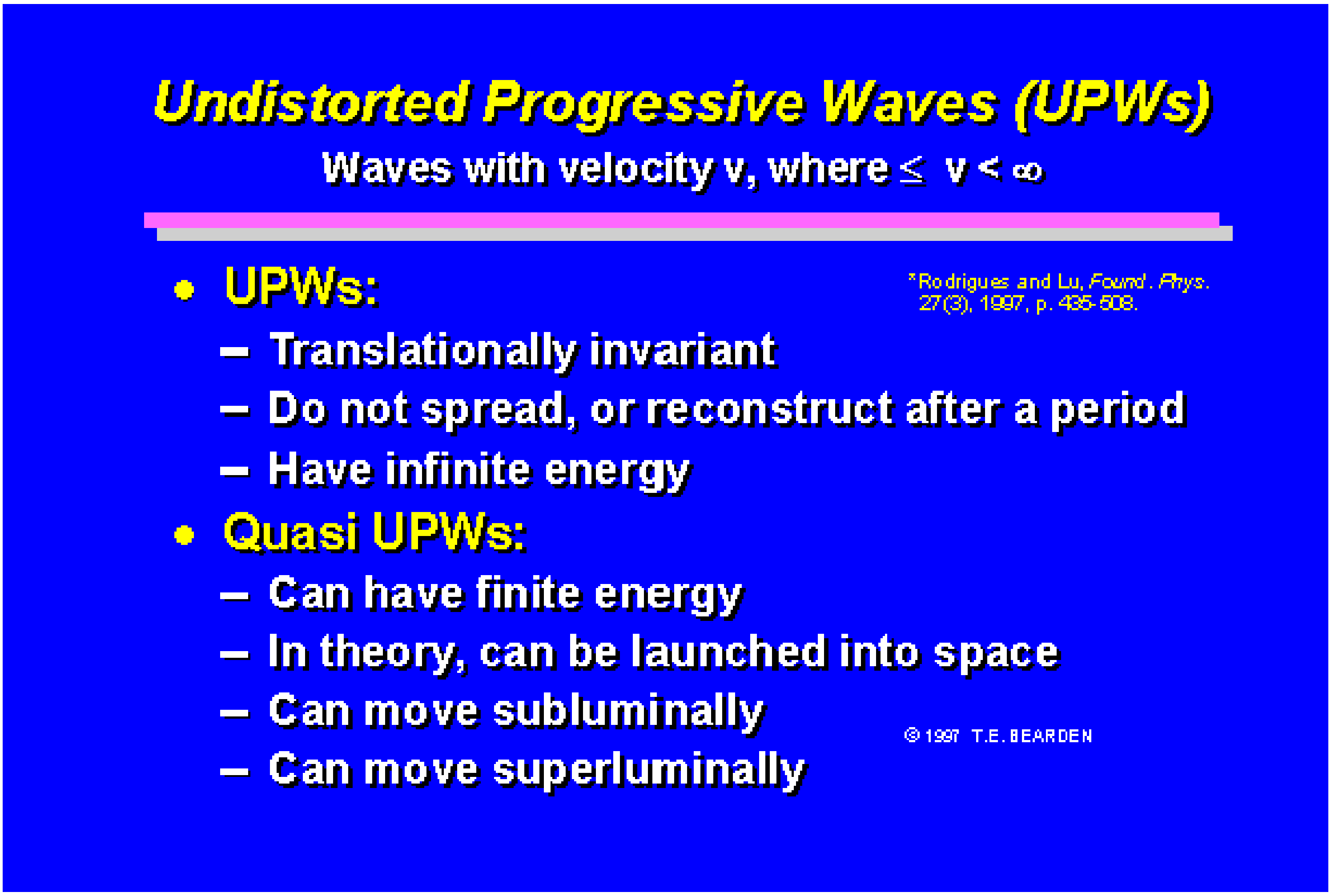

Undistorted Progressive Waves (UPWs)

Waves with velocity v, where $\leq$ v < ∞

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

- **UPWs:**
  - Translationally invariant
  - Do not spread, or reconstruct after a period
  - Have infinite energy
- **Quasi UPWs:**
  - Can have finite energy
  - In theory, can be launched into space
  - Can move subluminally
  - Can move superluminally

© 1997 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

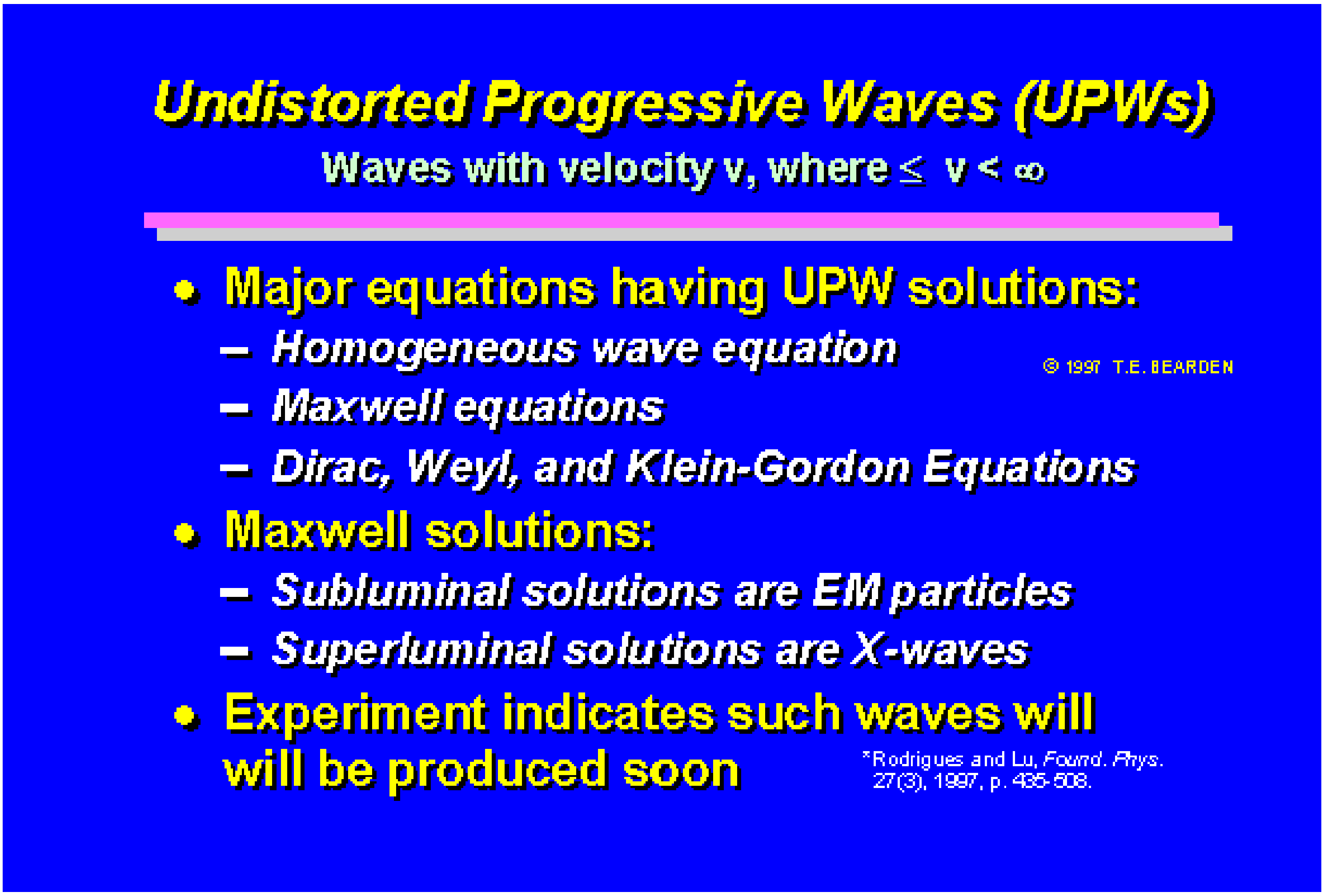

Undistorted Progressive Waves (UPWs)

Waves with velocity v, where $\leq$ v < ∞

- Major equations having UPW solutions:
  - *Homogeneous wave equation*
  - *Maxwell equations*
  - *Dirac, Weyl, and Klein-Gordon Equations*
- Maxwell solutions:
  - *Subluminal solutions are EM particles*
  - *Superluminal solutions are X-waves*
- Experiment indicates such waves will will be produced soon

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

© 1997 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

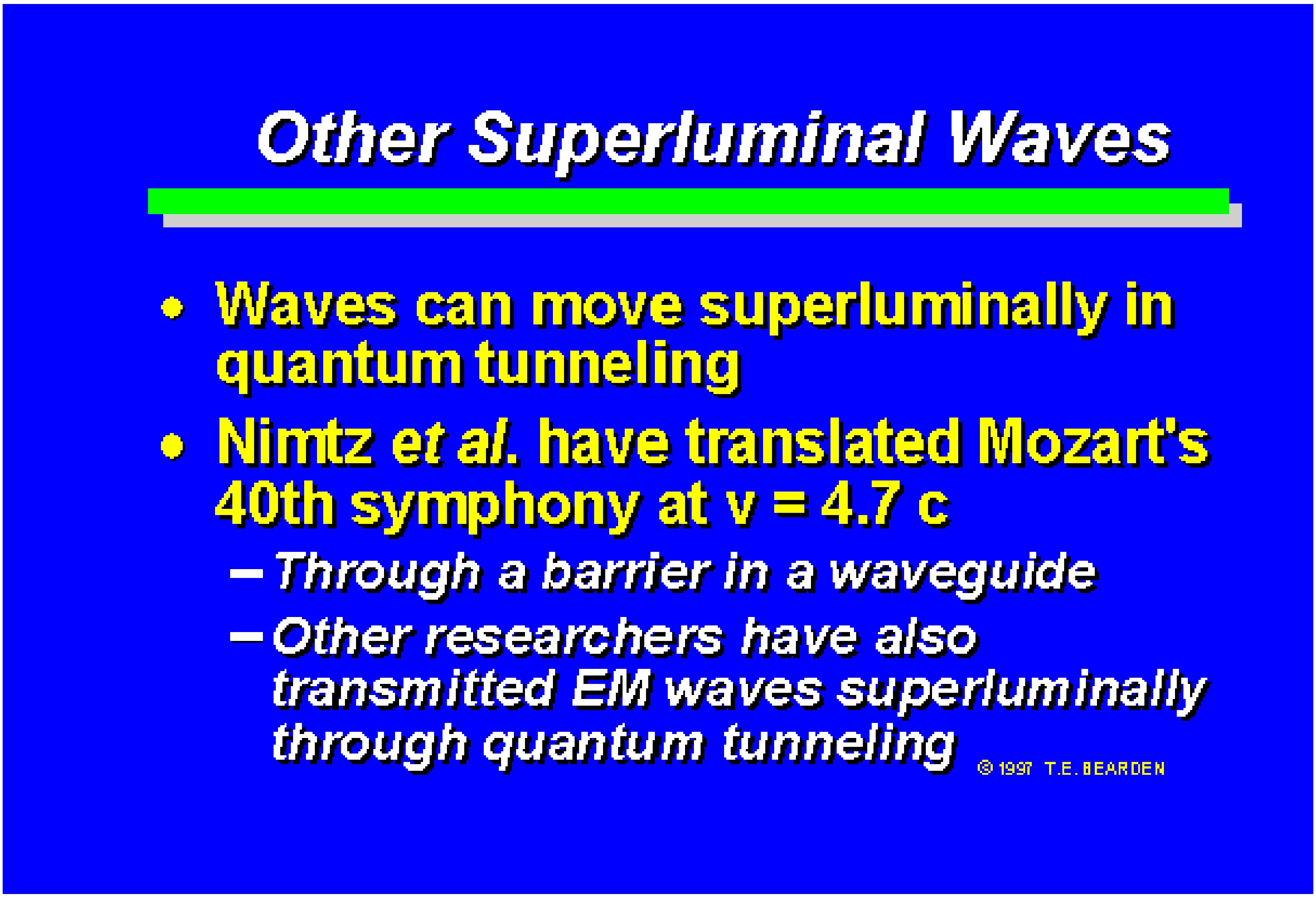

Other Superluminal Waves

- **Waves can move superluminally in quantum tunneling**
- **Nimtz *et al.* have translated Mozart's 40th symphony at v = 4.7 c**
  - *Through a barrier in a waveguide*
  - *Other researchers have also transmitted EM waves superluminally through quantum tunneling*

© 1997 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

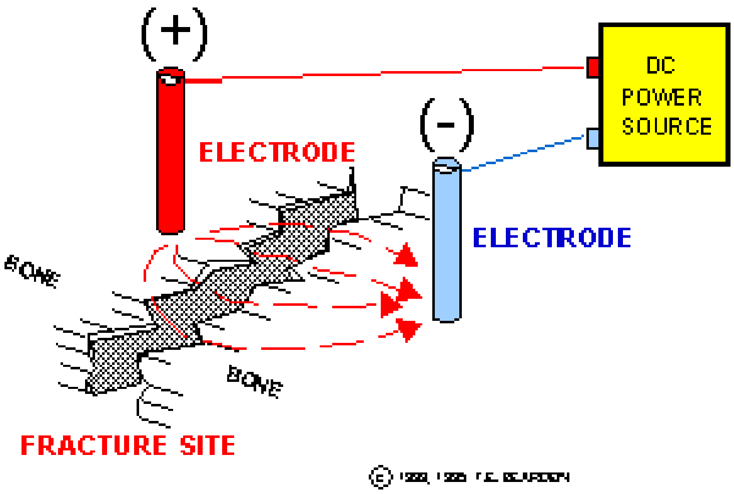
(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE
© 1993, 1995 T.E. BEARDEN

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Published by Lotus® Freelance Graphics®

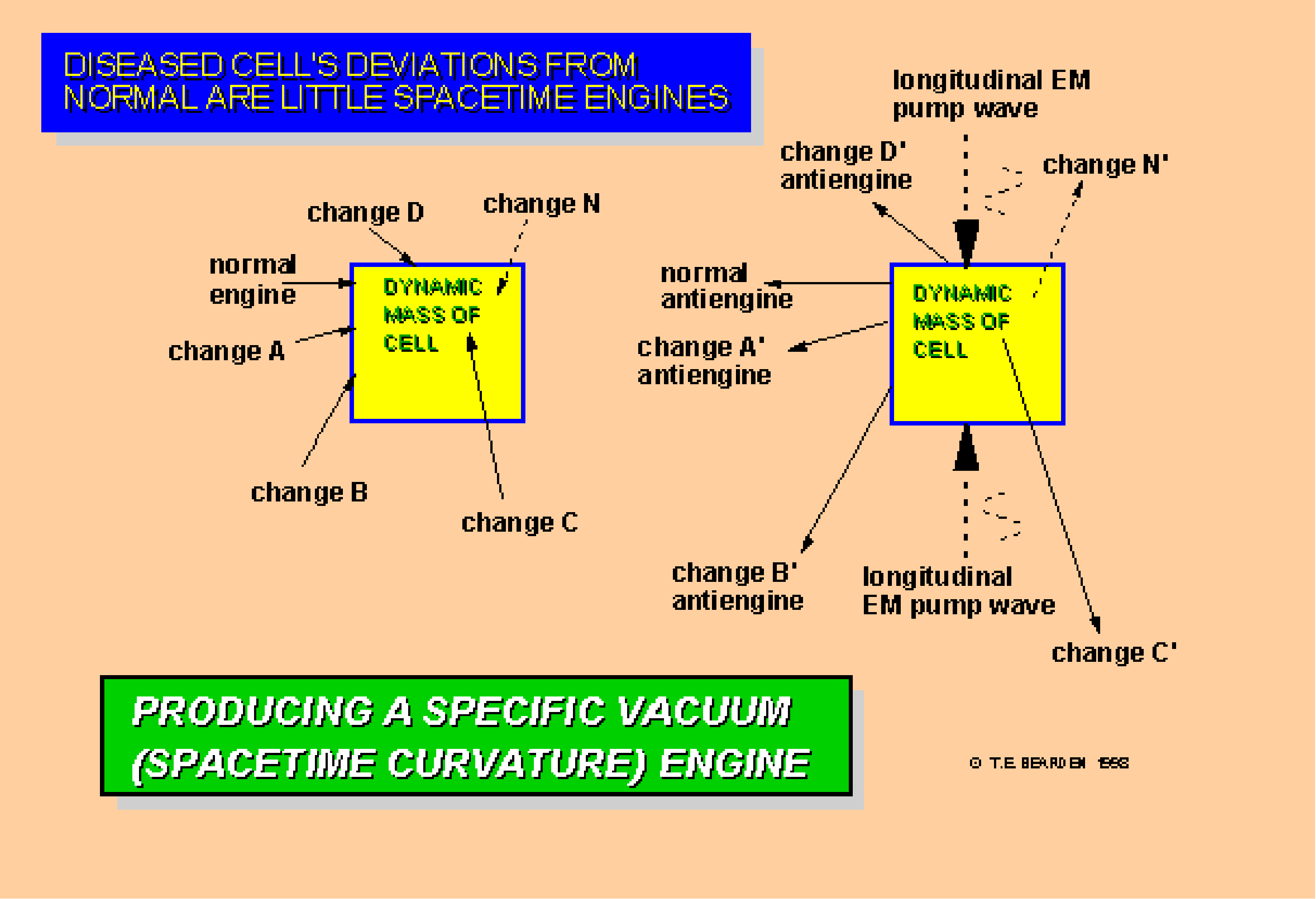

DISEASED CELL'S DEVIATIONS FROM NORMAL ARE LITTLE SPACETIME ENGINES

change D

change N

normal engine

DYNAMIC MASS OF CELL

change A

change B

change C

longitudinal EM pump wave

change D' antiengine

change N'

normal antiengine

DYNAMIC MASS OF CELL

change A' antiengine

change B' antiengine

longitudinal EM pump wave

change C'

PRODUCING A SPECIFIC VACUUM (SPACETIME CURVATURE) ENGINE

© T.E. BEARDEN 1993

Published by Lotus® Freelance Graphics®

Transverse EM Wave Pumping of PCM

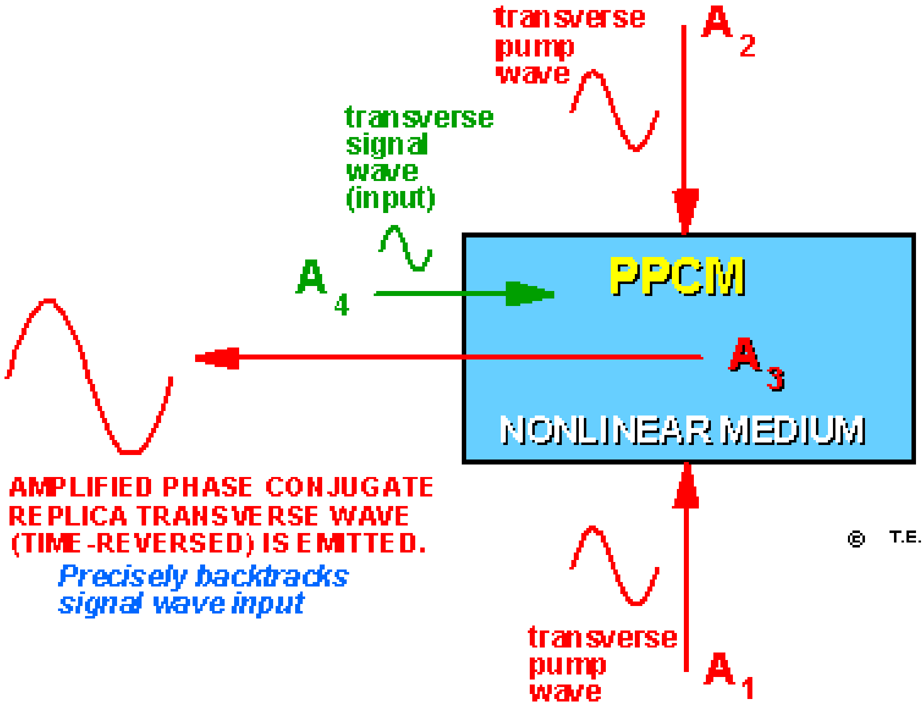
transverse
pump
wave
A 2
transverse
signal
wave
(input)
A 4
PPCM
A 3
NONLINEAR MEDIUM
transverse
pump
wave
A 1

AMPLIFIED PHASE CONJUGATE REPLICA TRANSVERSE WAVE (TIME-REVERSED) IS EMITTED.

Precisely backtracks signal wave input

© T.E. BEARDEN 1997

Published by Lotus® Freelance Graphics®

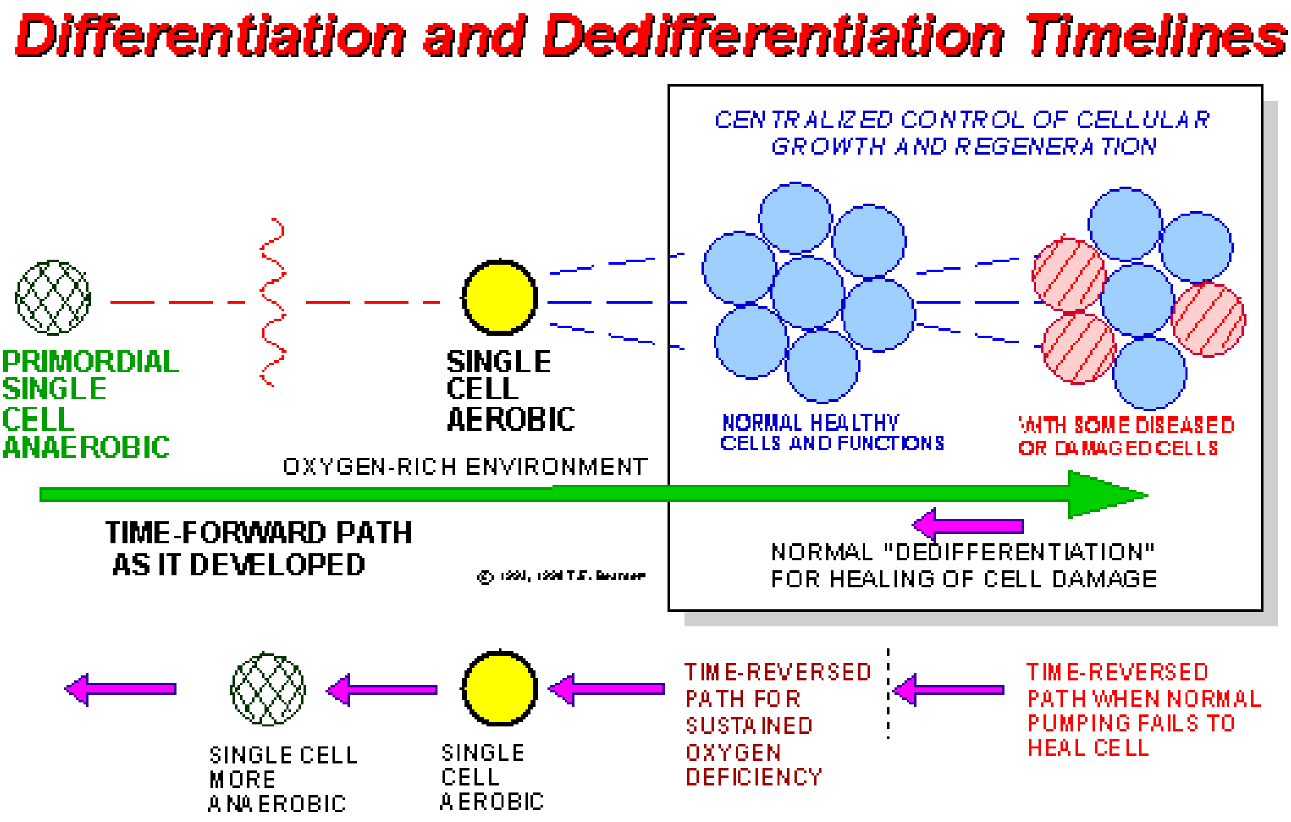
Differentiation and Dedifferentiation Timelines
CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION
PRIMORDIAL SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC
NORMAL HEALTHY CELLS AND FUNCTIONS
WITH SOME DISEASED OR DAMAGED CELLS
OXYGEN-RICH ENVIRONMENT
TIME-FORWARD PATH AS IT DEVELOPED
NORMAL "DEDIFFERENTIATION" FOR HEALING OF CELL DAMAGE
SINGLE CELL MORE ANAEROBIC
SINGLE CELL AEROBIC
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
TIME-REVERSED PATH WHEN NORMAL PUMPING FAILS TO HEAL CELL

Published by Lotus® Freelance Graphics®

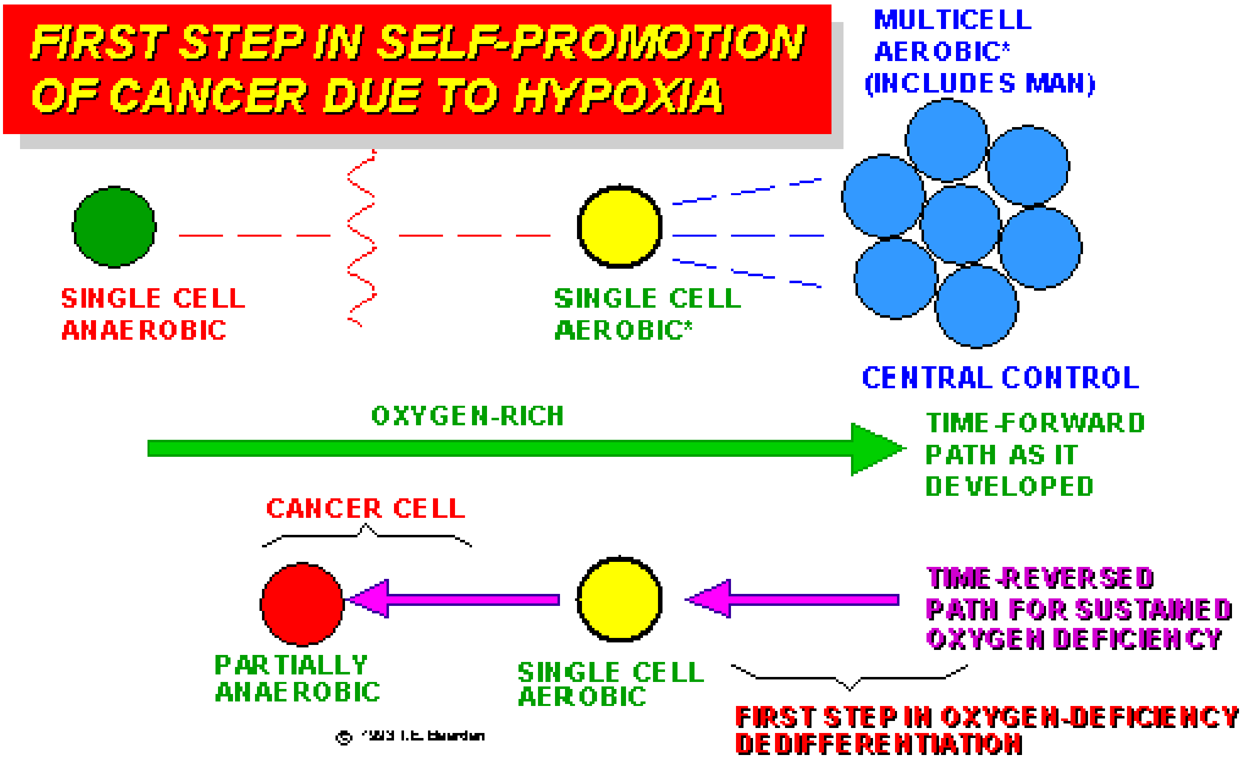
FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA
MULTICELL AEROBIC* (INCLUDES MAN)
SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC*
CENTRAL CONTROL
OXYGEN-RICH
TIME-FORWARD PATH AS IT DEVELOPED
CANCER CELL
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
PARTIALLY ANAEROBIC
SINGLE CELL AEROBIC
FIRST STEP IN OXYGEN-DEFICIENCY DEDIFFERENTIATION
© 1993 T.E. Bearden

Published by Lotus® Freelance Graphics®

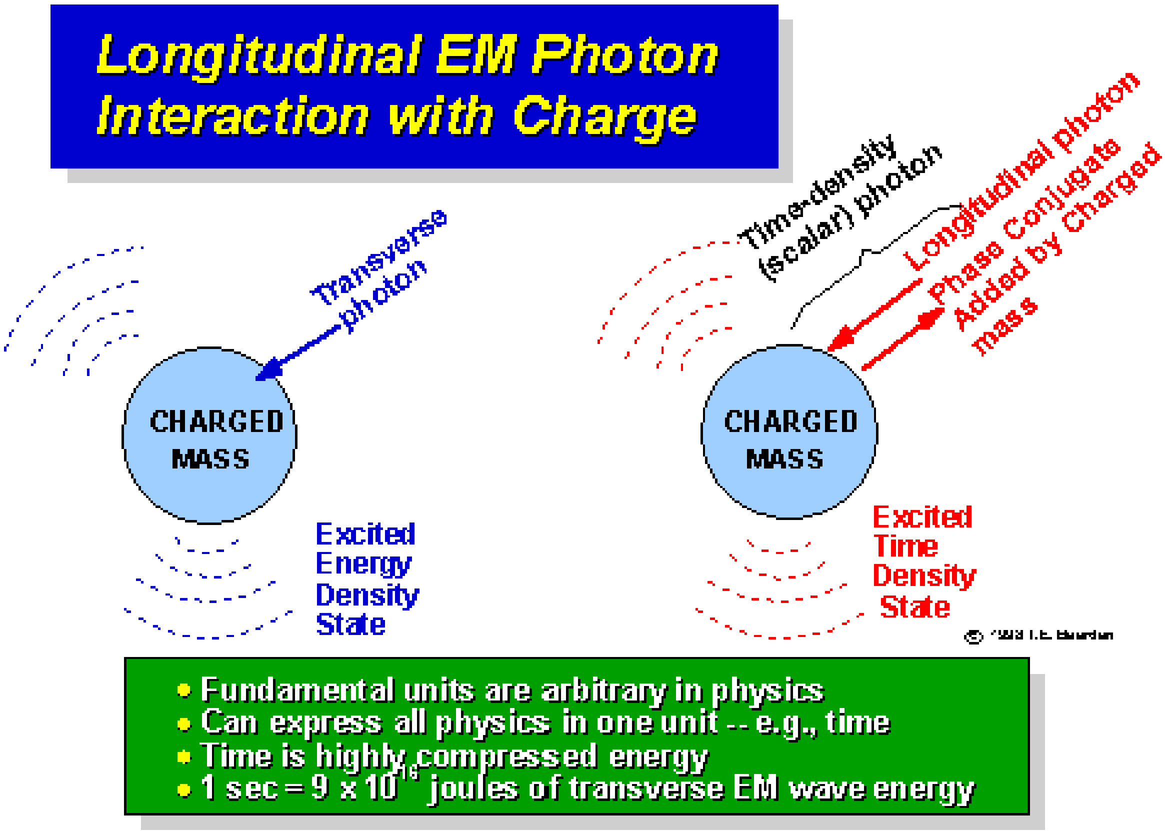

Longitudinal EM Photon Interaction with Charge

Transverse photon
CHARGED MASS
Excited Energy Density State
Time-density (scalar) photon
Longitudinal photon
Phase Conjugate Added by Charged mass
CHARGED MASS
Excited Time Density State

© 1998 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9×10^{16} joules of transverse EM wave energy

Published by Lotus® Freelance Graphics®

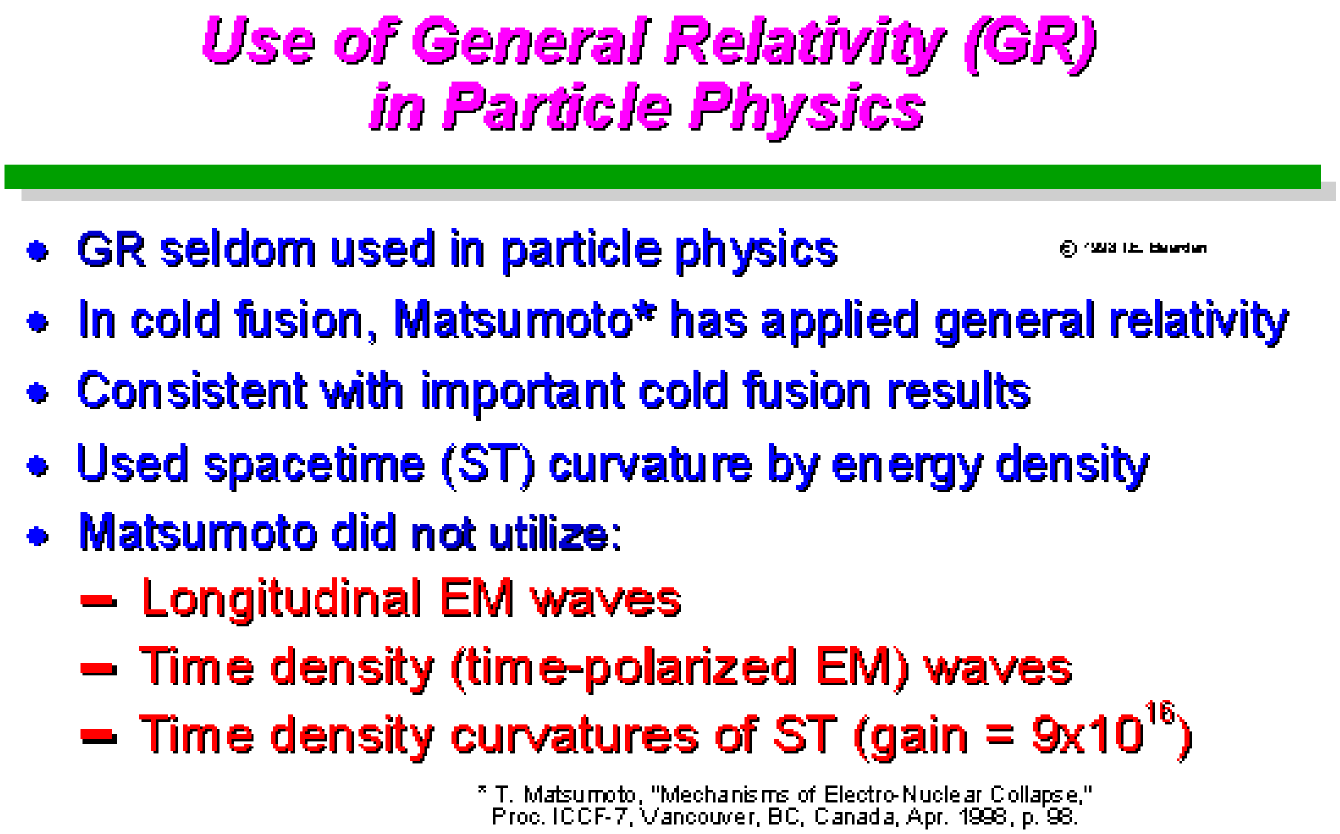

Use of General Relativity (GR) in Particle Physics

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density (time-polarized EM) waves
  - Time density curvatures of ST (gain = $9x10^{16}$)

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98.

Published by Lotus® Freelance Graphics®

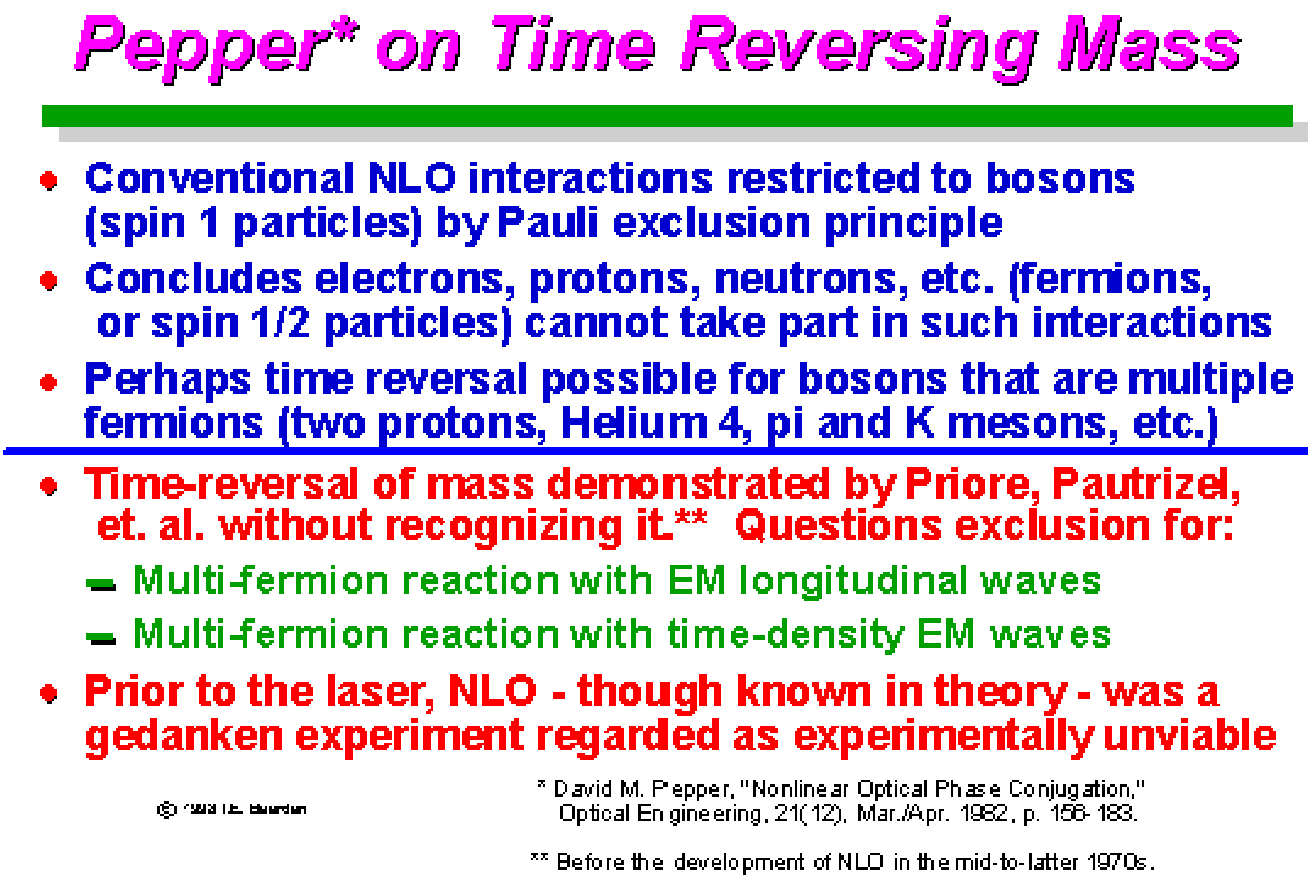

Pepper* on Time Reversing Mass

- Conventional NLO interactions restricted to bosons (spin 1 particles) by Pauli exclusion principle
- Concludes electrons, protons, neutrons, etc. (fermions, or spin 1/2 particles) cannot take part in such interactions
- Perhaps time reversal possible for bosons that are multiple fermions (two protons, Helium 4, pi and K mesons, etc.)

- Time-reversal of mass demonstrated by Priore, Pautrizel, et. al. without recognizing it.** Questions exclusion for:
  - Multi-fermion reaction with EM longitudinal waves
  - Multi-fermion reaction with time-density EM waves
- Prior to the laser, NLO - though known in theory - was a gedanken experiment regarded as experimentally unviable

* David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(12), Mar./Apr. 1982, p. 156-183.

** Before the development of NLO in the mid-to-latter 1970s.

Published by Lotus® Freelance Graphics®

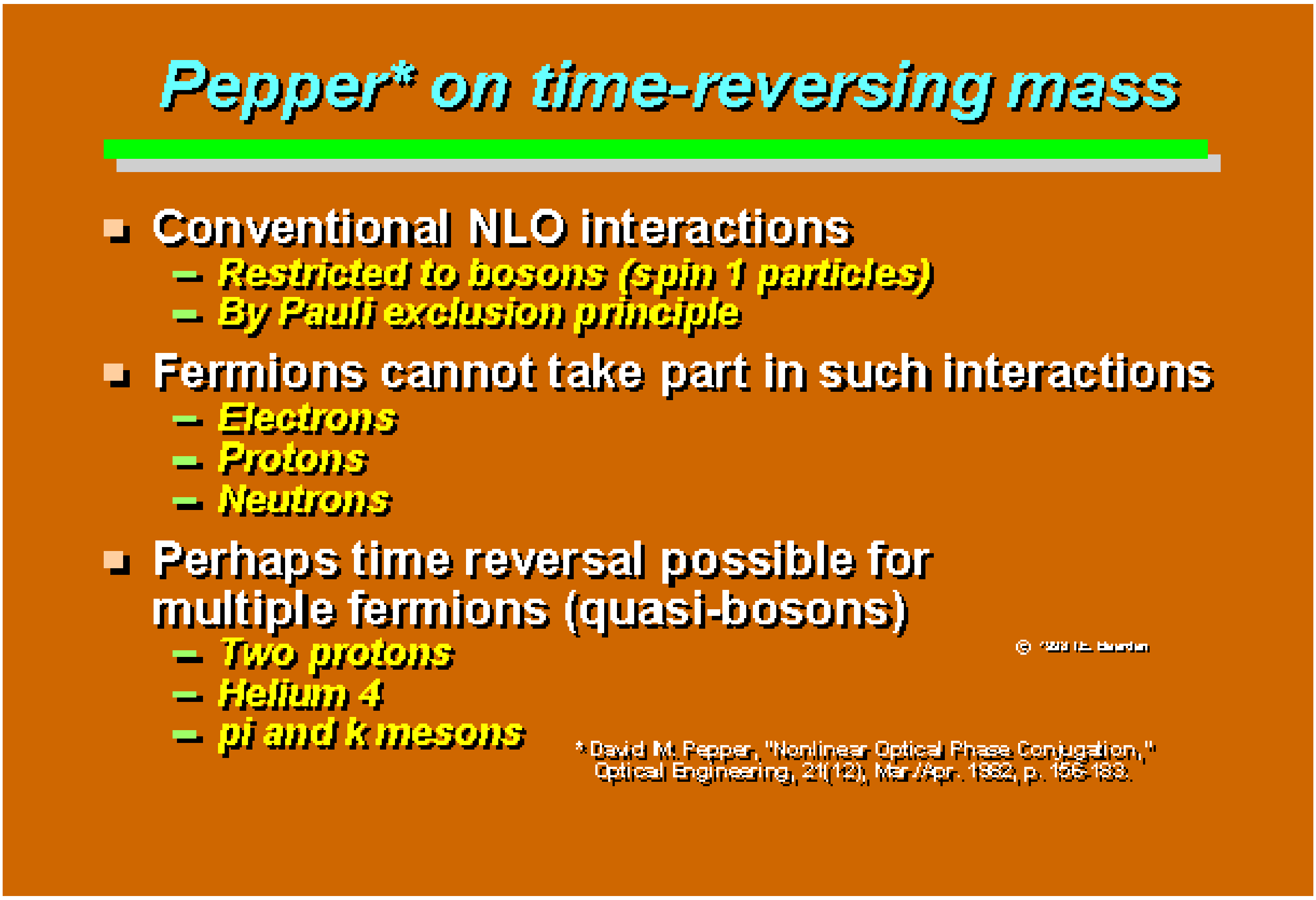

Pepper* on time-reversing mass

- Conventional NLO interactions
  - *Restricted to bosons (spin 1 particles)*
  - *By Pauli exclusion principle*
- Fermions cannot take part in such interactions
  - *Electrons*
  - *Protons*
  - *Neutrons*
- Perhaps time reversal possible for multiple fermions (quasi-bosons)
  - *Two protons*
  - *Helium 4*
  - *pi and k mesons*

* David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(12), Mar./Apr. 1982, p. 156-183.

Published by Lotus® Freelance Graphics®

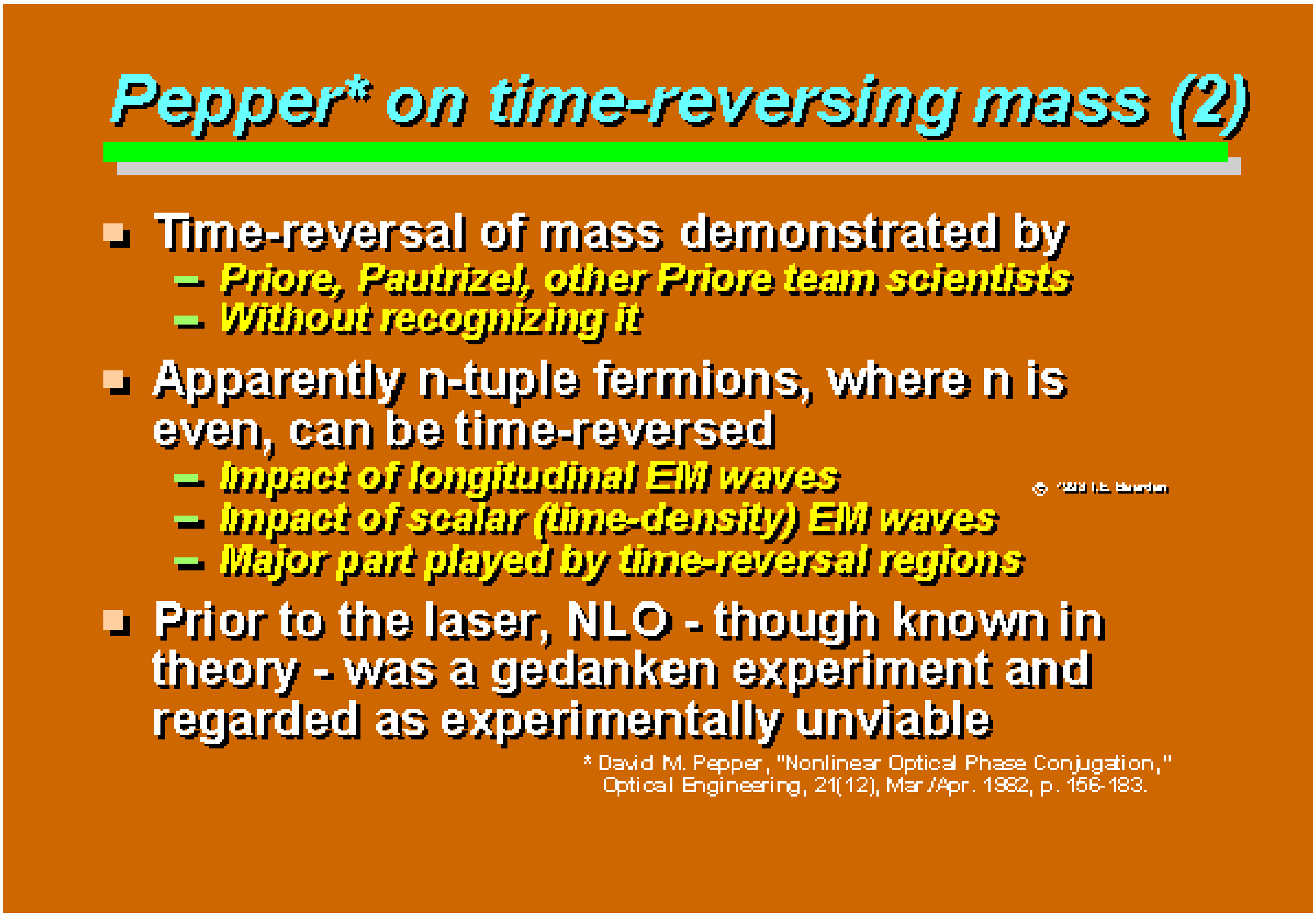

Pepper* on time-reversing mass (2)

- Time-reversal of mass demonstrated by
  - *Priore, Pautrizel, other Priore team scientists*
  - *Without recognizing it*
- Apparently n-tuple fermions, where n is even, can be time-reversed
  - *Impact of longitudinal EM waves*
  - *Impact of scalar (time-density) EM waves*
  - *Major part played by time-reversal regions*
- Prior to the laser, NLO - though known in theory - was a gedanken experiment and regarded as experimentally unviable

* David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(12), Mar./Apr. 1982, p. 156-183.

Published by Lotus® Freelance Graphics®

Chung's Carbon Filament Negative Resistor

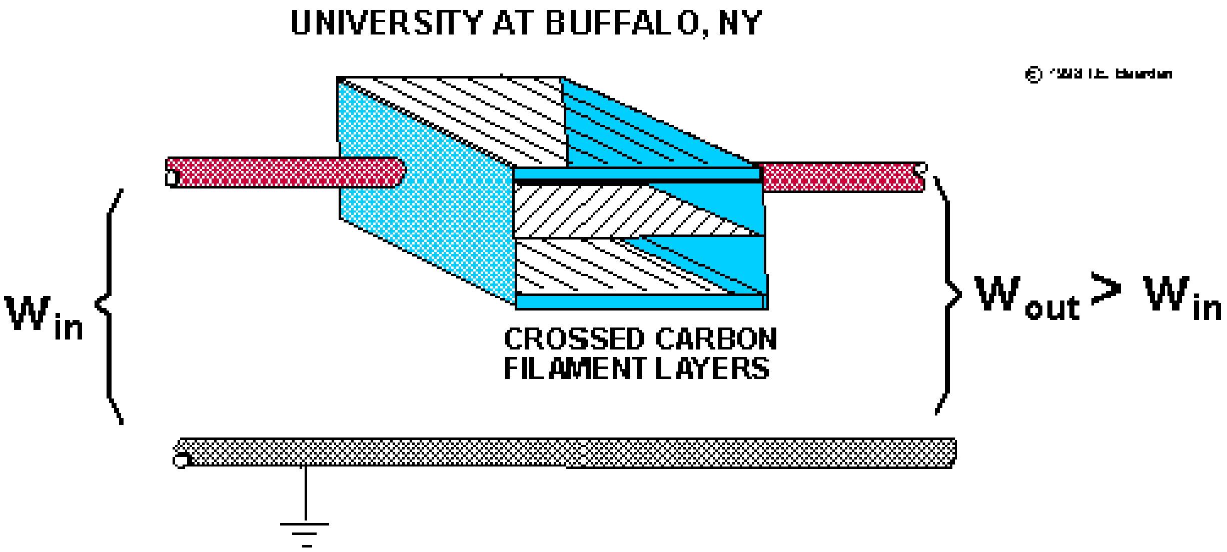
UNIVERSITY AT BUFFALO, NY
© 1993 T.E. Bearden
W_{in}
CROSSED CARBON FILAMENT LAYERS
$W_{out} > W_{in}$

Published by Lotus® Freelance Graphics®

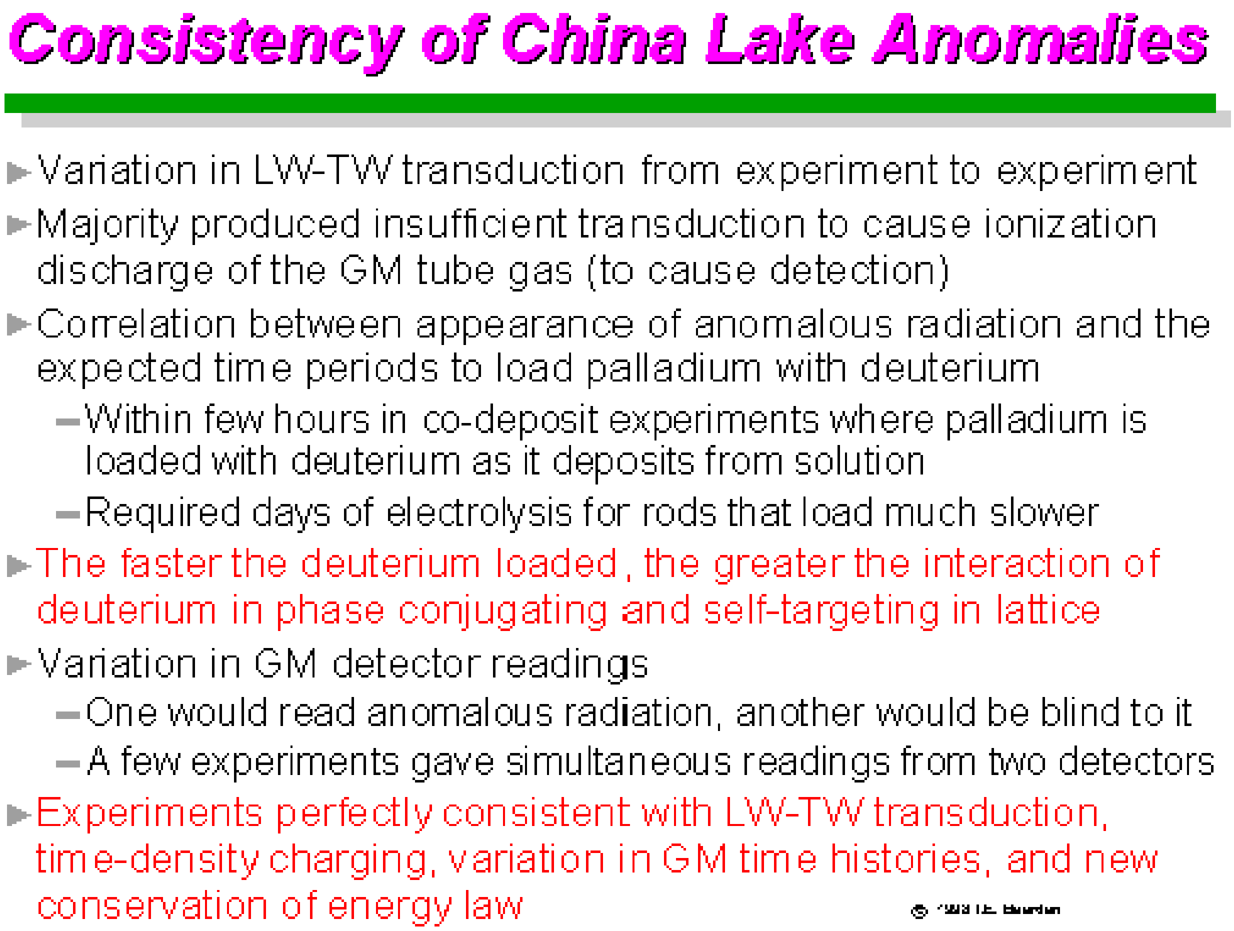

Consistency of China Lake Anomalies

- Variation in LW-TW transduction from experiment to experiment
- Majority produced insufficient transduction to cause ionization discharge of the GM tube gas (to cause detection)
- Correlation between appearance of anomalous radiation and the expected time periods to load palladium with deuterium
  - Within few hours in co-deposit experiments where palladium is loaded with deuterium as it deposits from solution
  - Required days of electrolysis for rods that load much slower
- The faster the deuterium loaded, the greater the interaction of deuterium in phase conjugating and self-targeting in lattice
- Variation in GM detector readings
  - One would read anomalous radiation, another would be blind to it
  - A few experiments gave simultaneous readings from two detectors
- Experiments perfectly consistent with LW-TW transduction, time-density charging, variation in GM time histories, and new conservation of energy law

Published by Lotus® Freelance Graphics®

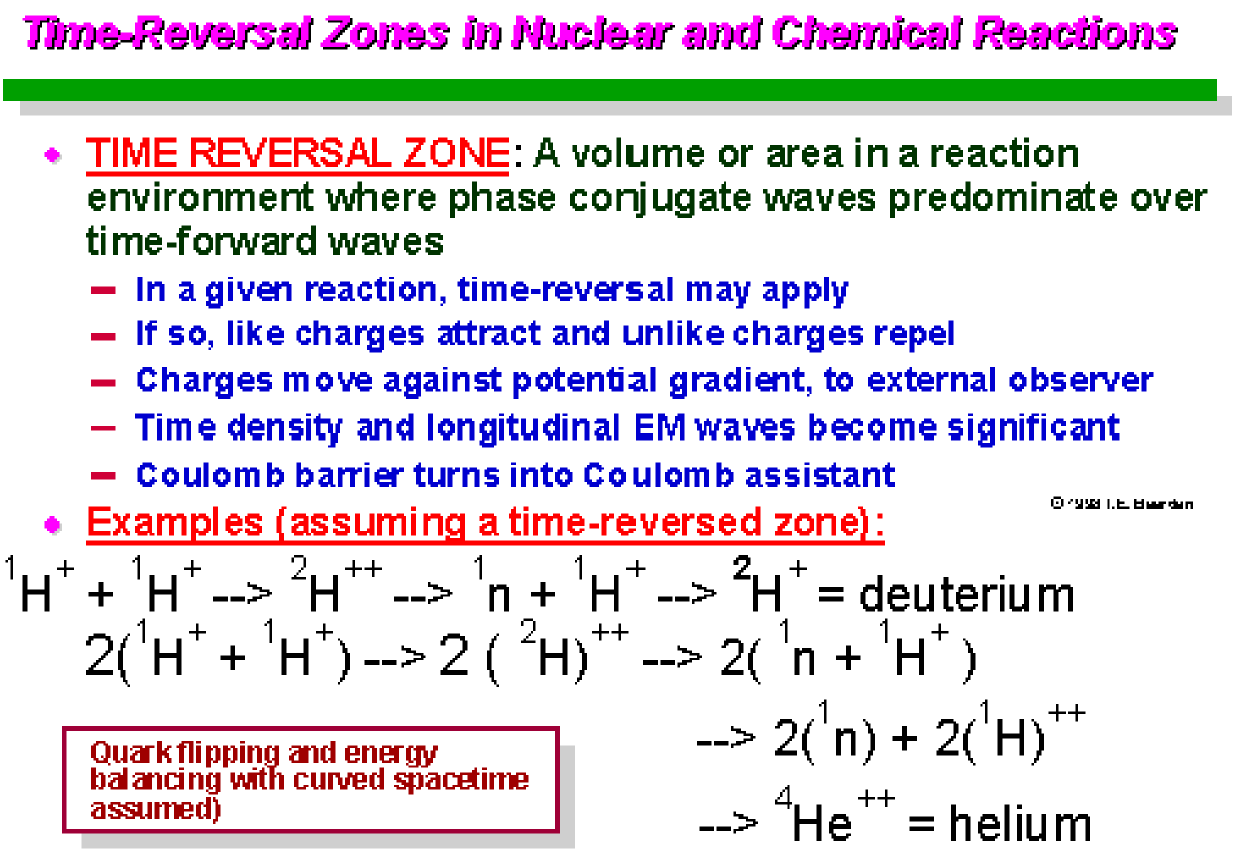

Time-Reversal Zones in Nuclear and Chemical Reactions

- **TIME REVERSAL ZONE**: A volume or area in a reaction environment where phase conjugate waves predominate over time-forward waves
  - In a given reaction, time-reversal may apply
  - If so, like charges attract and unlike charges repel
  - Charges move against potential gradient, to external observer
  - Time density and longitudinal EM waves become significant
  - Coulomb barrier turns into Coulomb assistant
- Examples (assuming a time-reversed zone):

© 1998 T.E. Bearden

$^{1}H^{+} + {}^{1}H^{+}$ --> $^{2}H^{++}$ --> $^{1}n + {}^{1}H^{+}$ --> $^{2}H^{+}$ = deuterium

$2({}^{1}H^{+} + {}^{1}H^{+})$ --> $2\,({}^{2}H)^{++}$ --> $2({}^{1}n + {}^{1}H^{+})$

--> $2({}^{1}n) + 2({}^{1}H)^{++}$

--> $^{4}He^{++}$ = helium

Quark flipping and energy balancing with curved spacetime assumed)

Published by Lotus® Freelance Graphics®

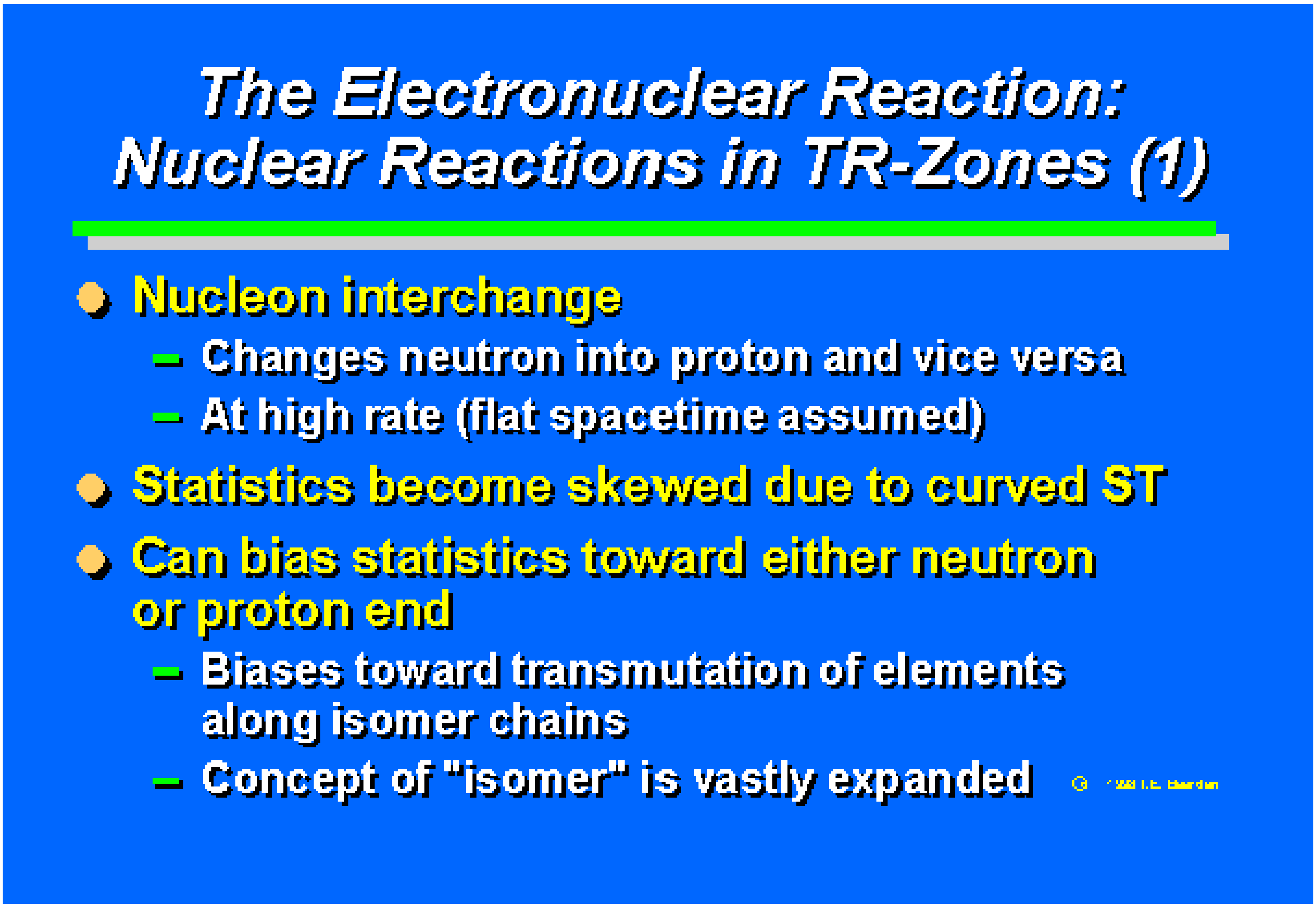

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- Nucleon interchange
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- Statistics become skewed due to curved ST
- Can bias statistics toward either neutron or proton end
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

© [illegible] T.E. Bearden

Published by Lotus® Freelance Graphics®

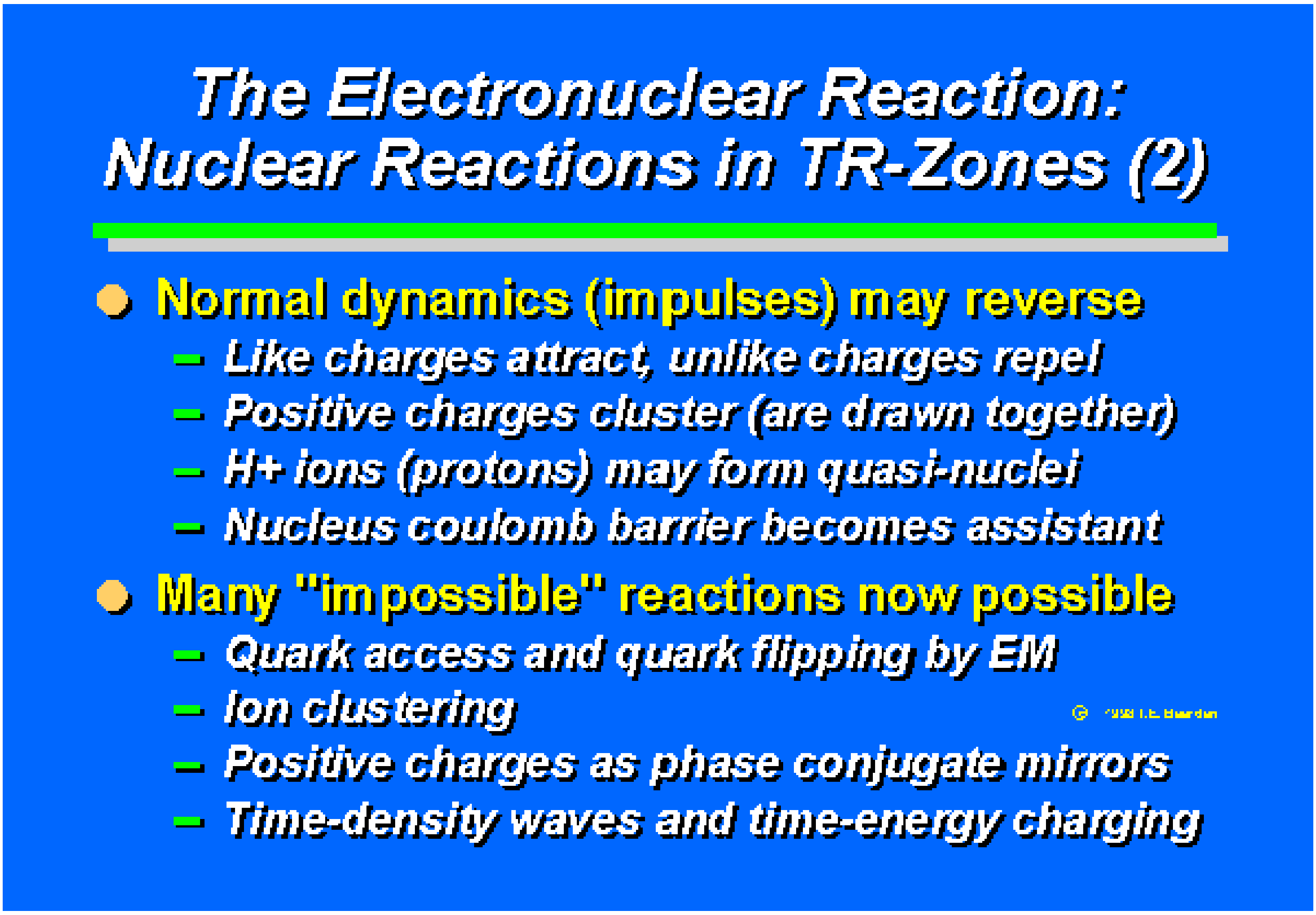

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- Normal dynamics (impulses) may reverse
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- Many "impossible" reactions now possible
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

© 1998 T.E. Bearden

Published by Lotus® Freelance Graphics®

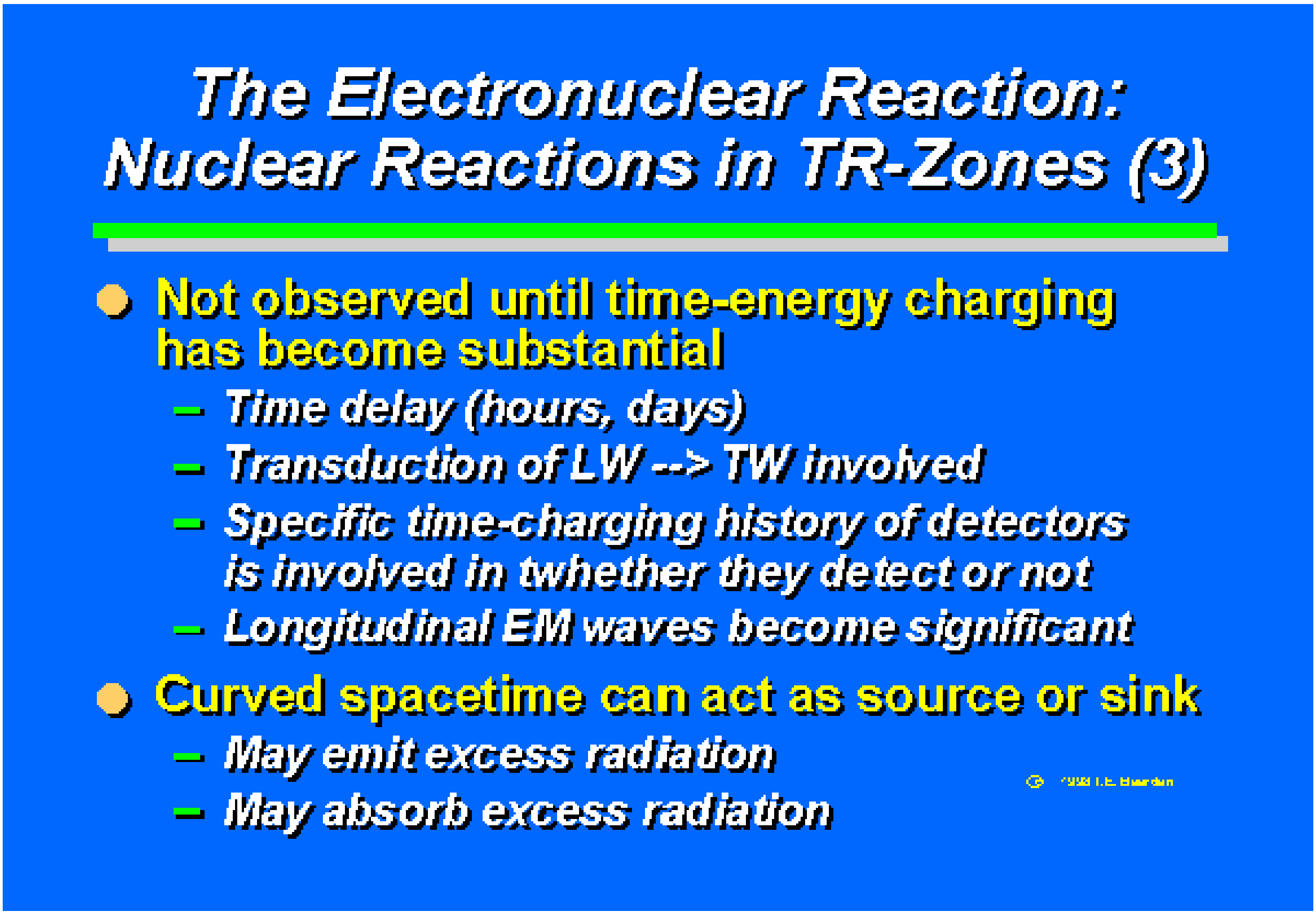

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- Not observed until time-energy charging has become substantial
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- Curved spacetime can act as source or sink
  - *May emit excess radiation*
  - *May absorb excess radiation*

© 1998 T.E. Bearden

Published by Lotus® Freelance Graphics®

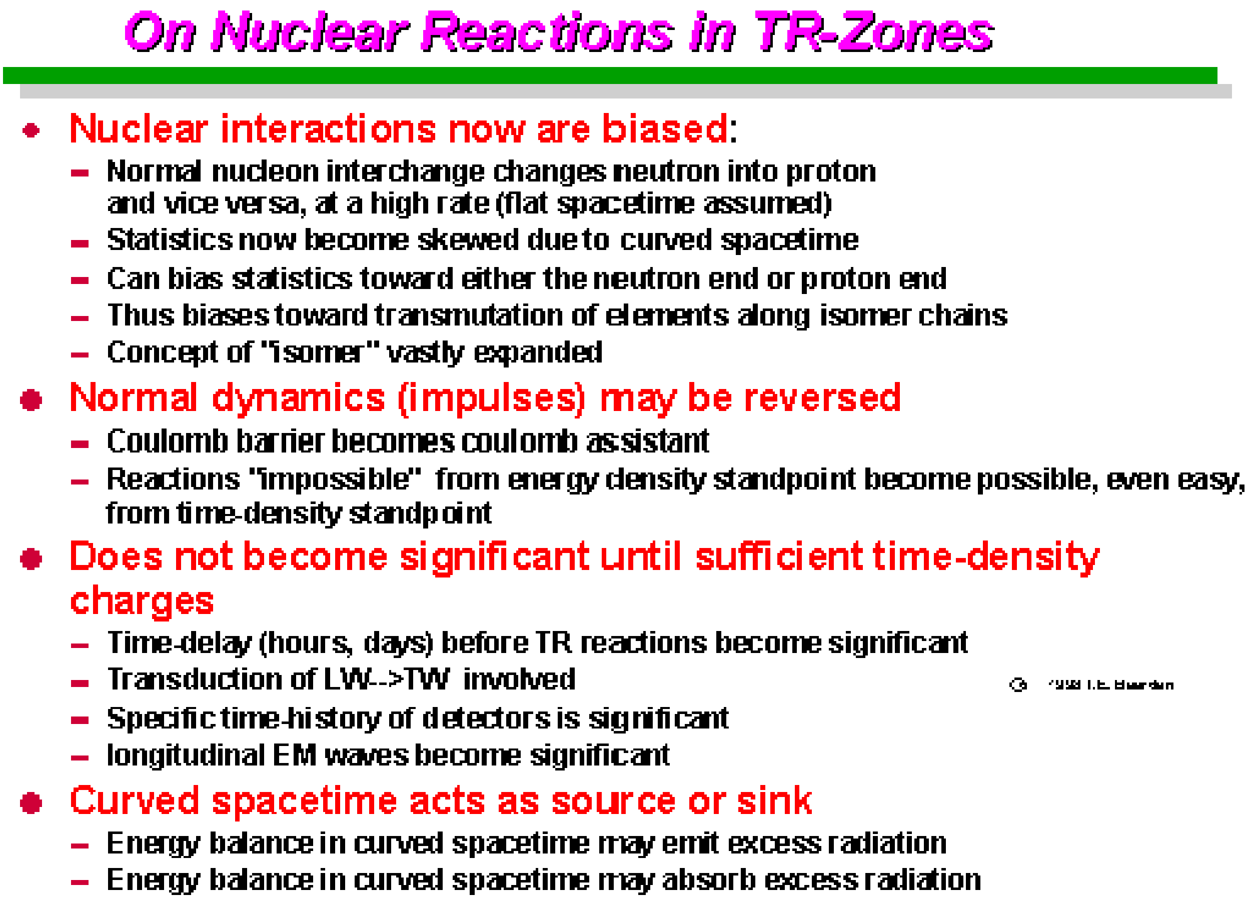

On Nuclear Reactions in TR-Zones

- Nuclear interactions now are biased:
  - Normal nucleon interchange changes neutron into proton and vice versa, at a high rate (flat spacetime assumed)
  - Statistics now become skewed due to curved spacetime
  - Can bias statistics toward either the neutron end or proton end
  - Thus biases toward transmutation of elements along isomer chains
  - Concept of "isomer" vastly expanded
- Normal dynamics (impulses) may be reversed
  - Coulomb barrier becomes coulomb assistant
  - Reactions "impossible" from energy density standpoint become possible, even easy, from time-density standpoint
- Does not become significant until sufficient time-density charges
  - Time-delay (hours, days) before TR reactions become significant
  - Transduction of LW-->TW involved
  - Specific time-history of detectors is significant
  - longitudinal EM waves become significant
- Curved spacetime acts as source or sink
  - Energy balance in curved spacetime may emit excess radiation
  - Energy balance in curved spacetime may absorb excess radiation

© [illegible] T.E. Bearden

Published by Lotus® Freelance Graphics®

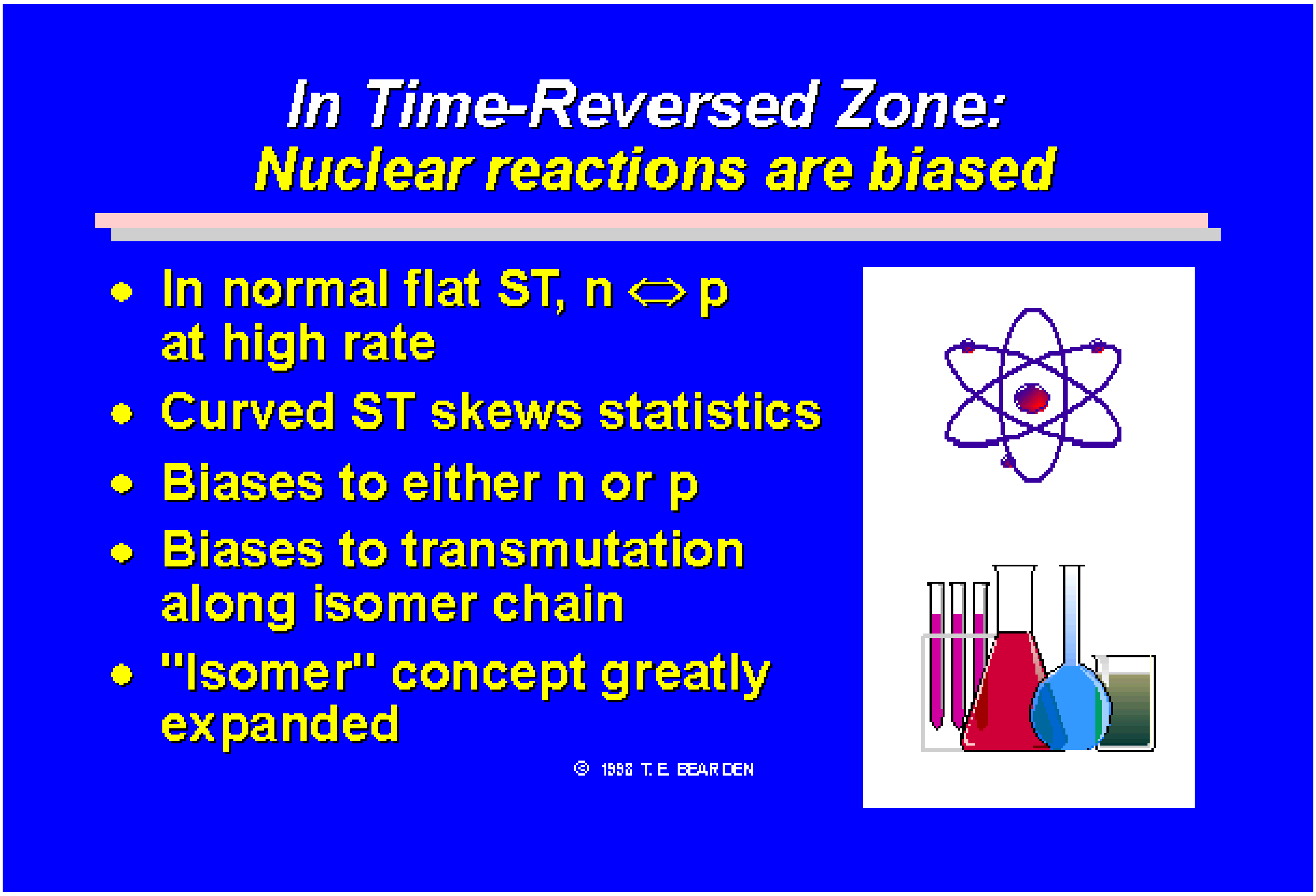

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T. E. BEARDEN

Published by Lotus® Freelance Graphics®

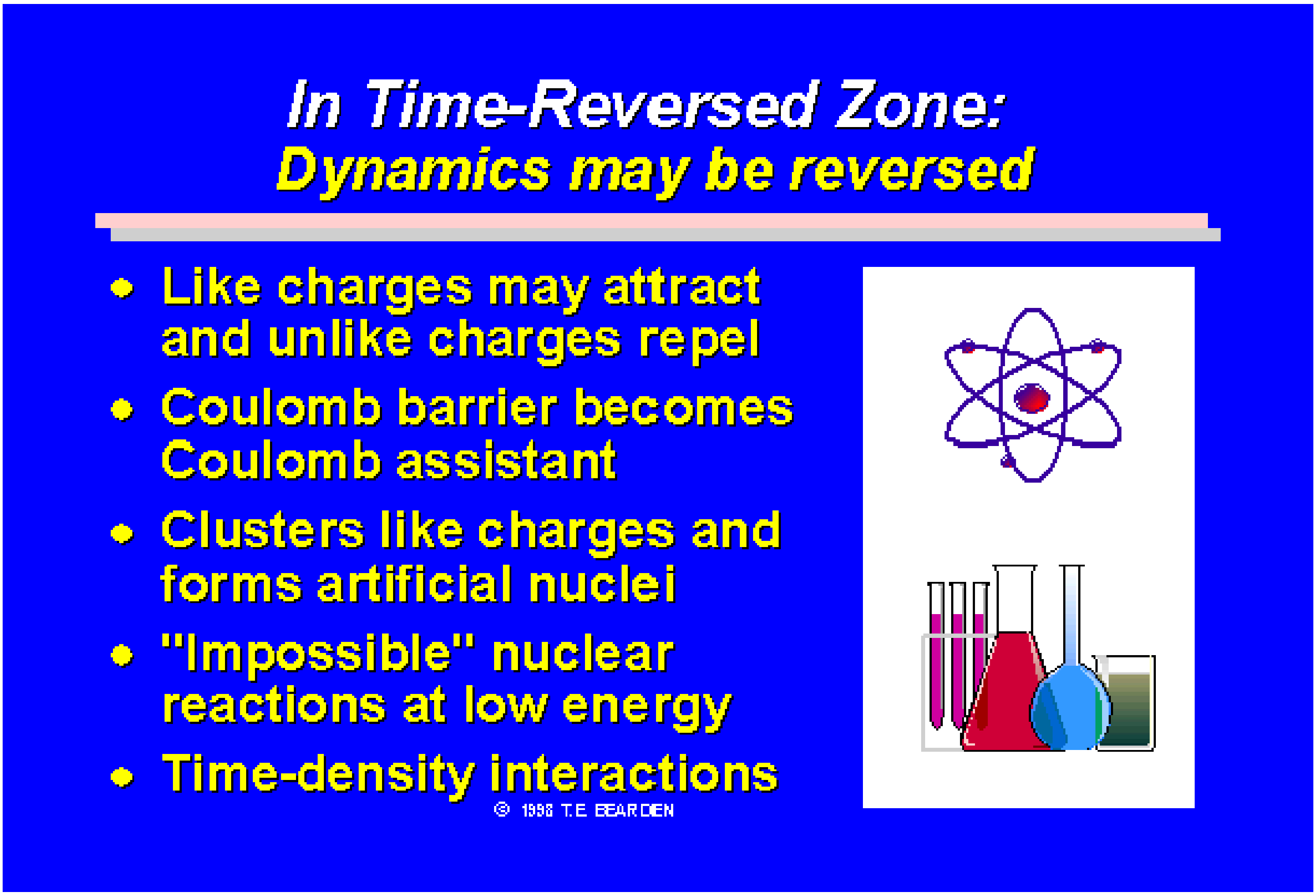

In Time-Reversed Zone:
Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

© 1998 T.E BEARDEN

Published by Lotus® Freelance Graphics®

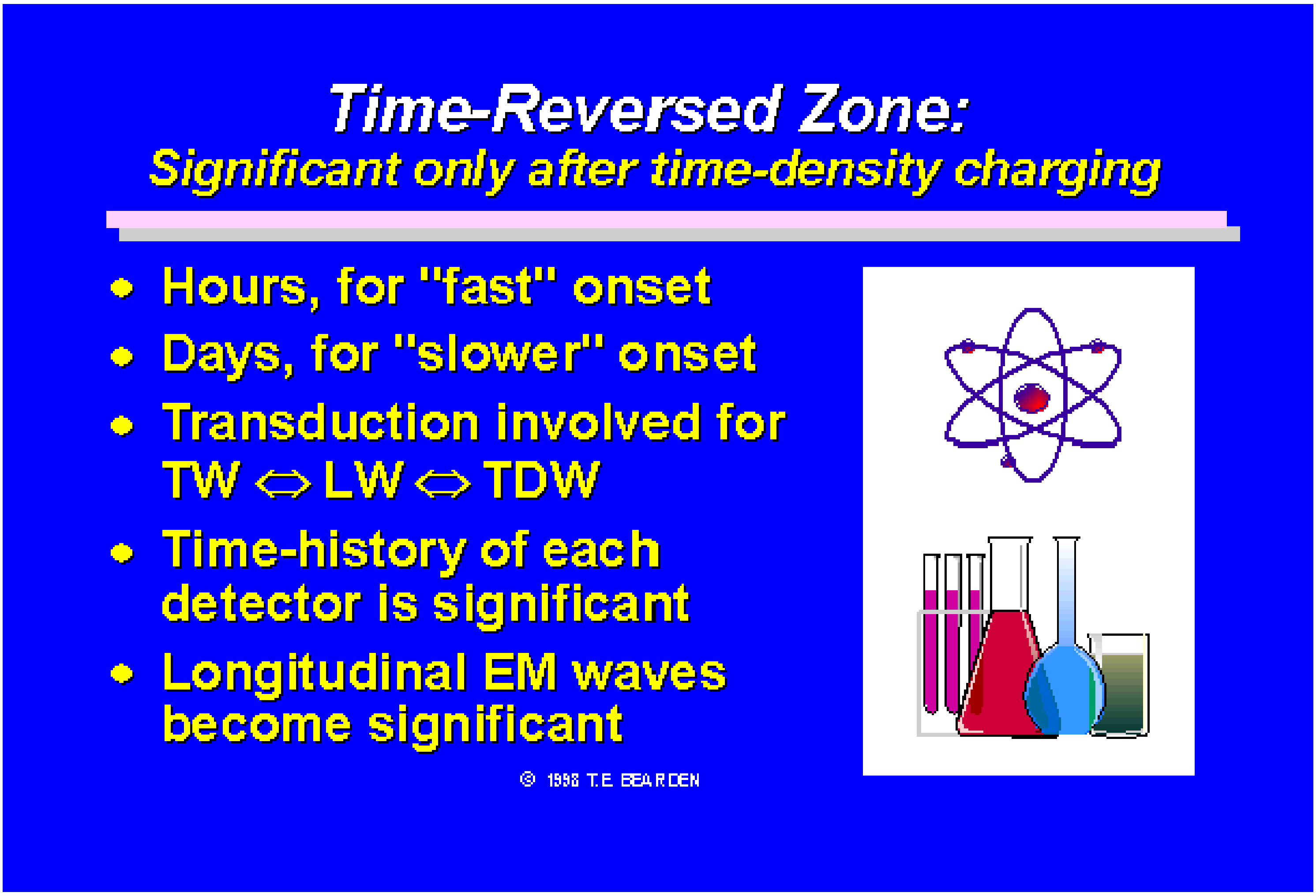

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

© 1998 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

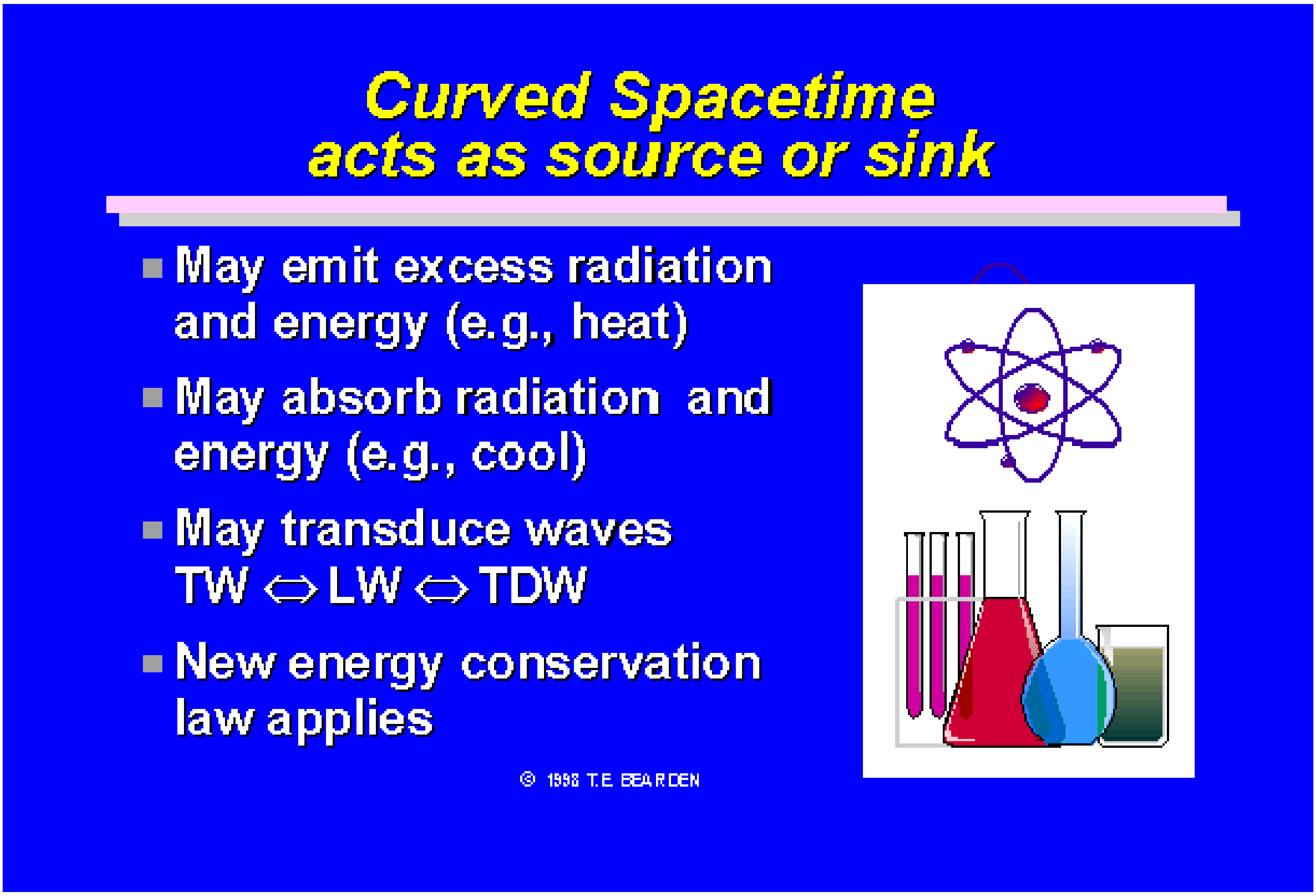

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

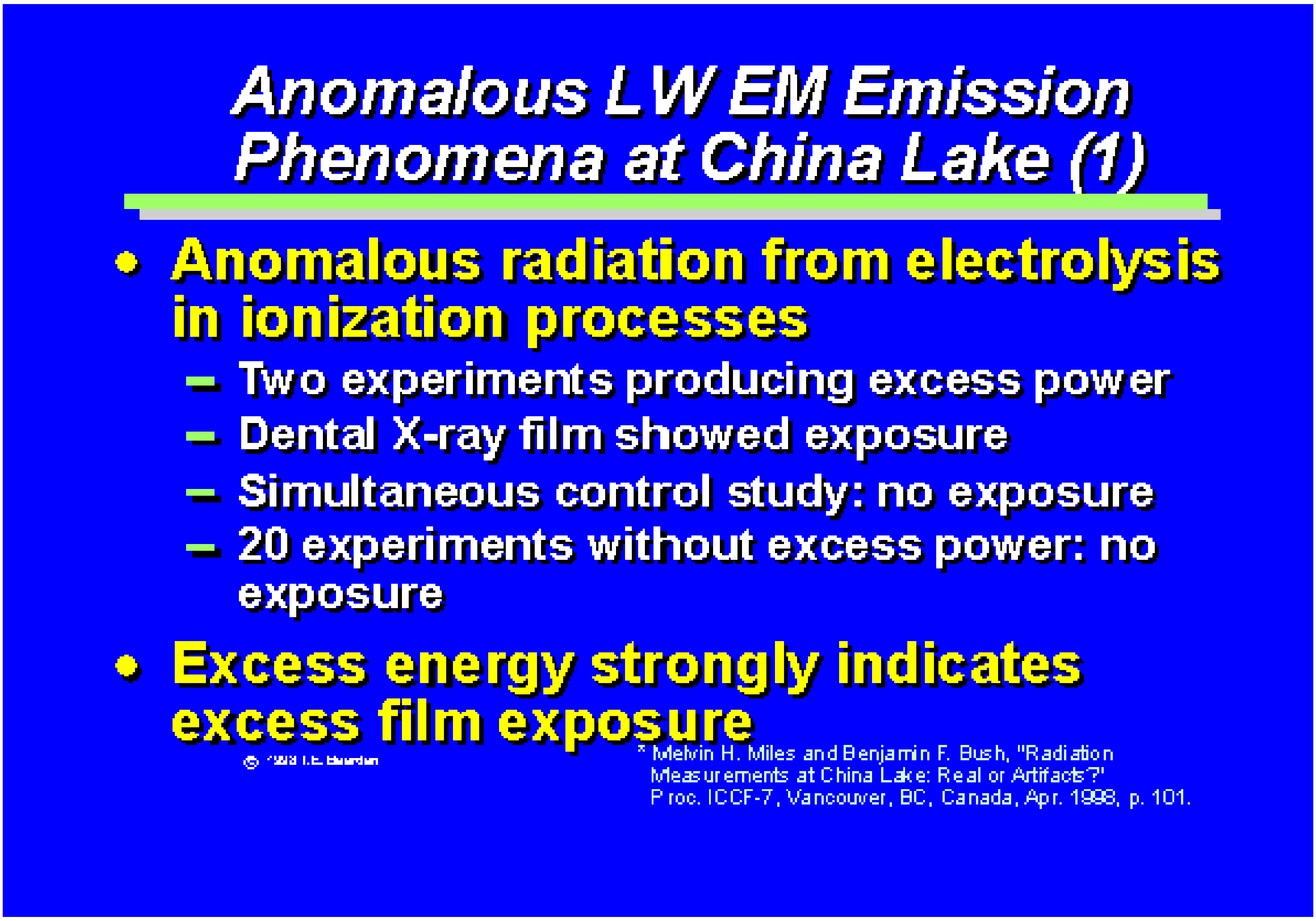

Anomalous LW EM Emission Phenomena at China Lake (1)

- **Anomalous radiation from electrolysis in ionization processes**
  - Two experiments producing excess power
  - Dental X-ray film showed exposure
  - Simultaneous control study: no exposure
  - 20 experiments without excess power: no exposure
- **Excess energy strongly indicates excess film exposure**

© 1998 T.E. Bearden

* Melvin H. Miles and Benjamin F. Bush, "Radiation Measurements at China Lake: Real or Artifacts?" Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 101.

Published by Lotus® Freelance Graphics®

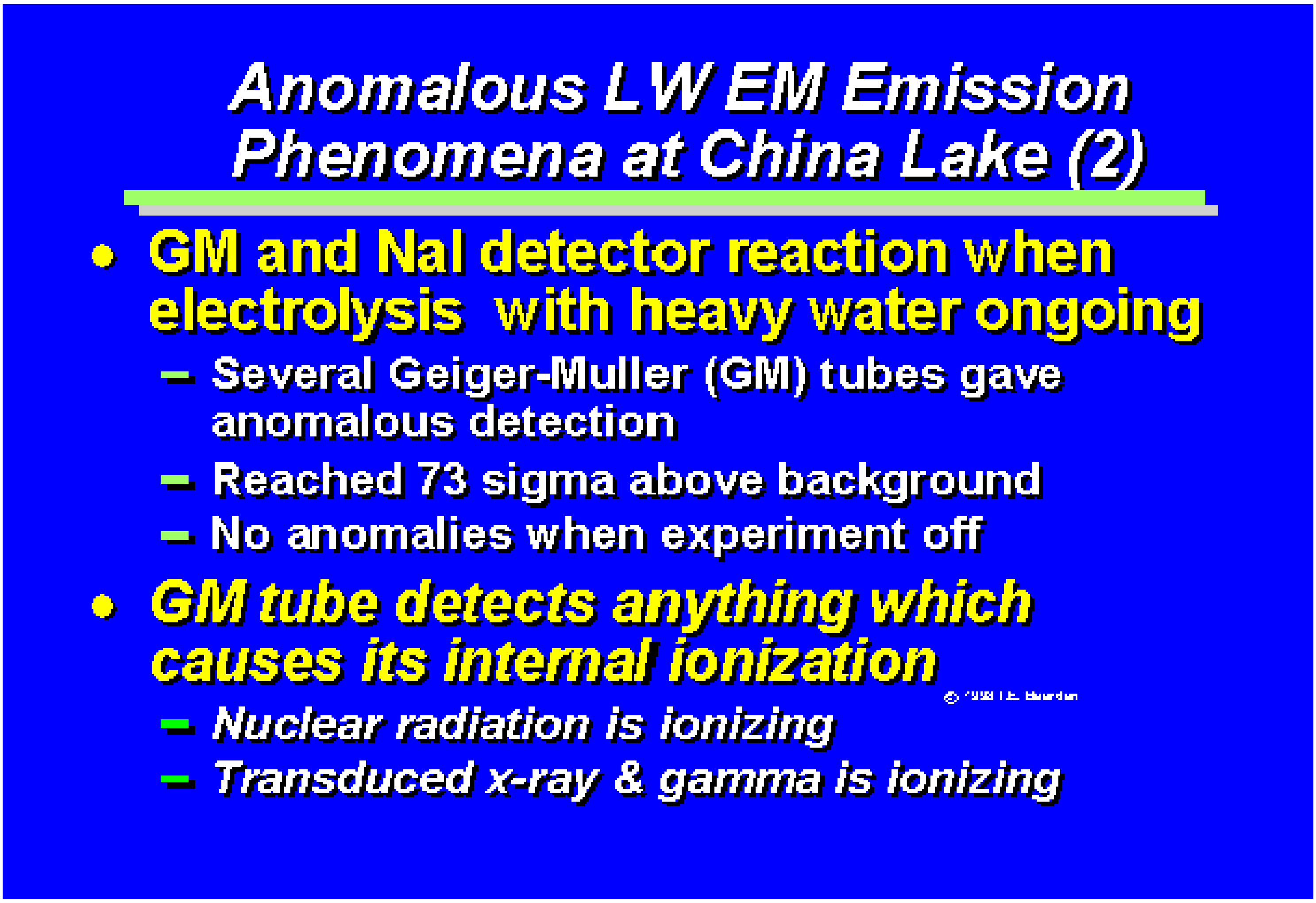

Anomalous LW EM Emission Phenomena at China Lake (2)

- **GM and NaI detector reaction when electrolysis with heavy water ongoing**
  - Several Geiger-Muller (GM) tubes gave anomalous detection
  - Reached 73 sigma above background
  - No anomalies when experiment off
- ***GM tube detects anything which causes its internal ionization***
  - *Nuclear radiation is ionizing*
  - *Transduced x-ray & gamma is ionizing*

Published by Lotus® Freelance Graphics®

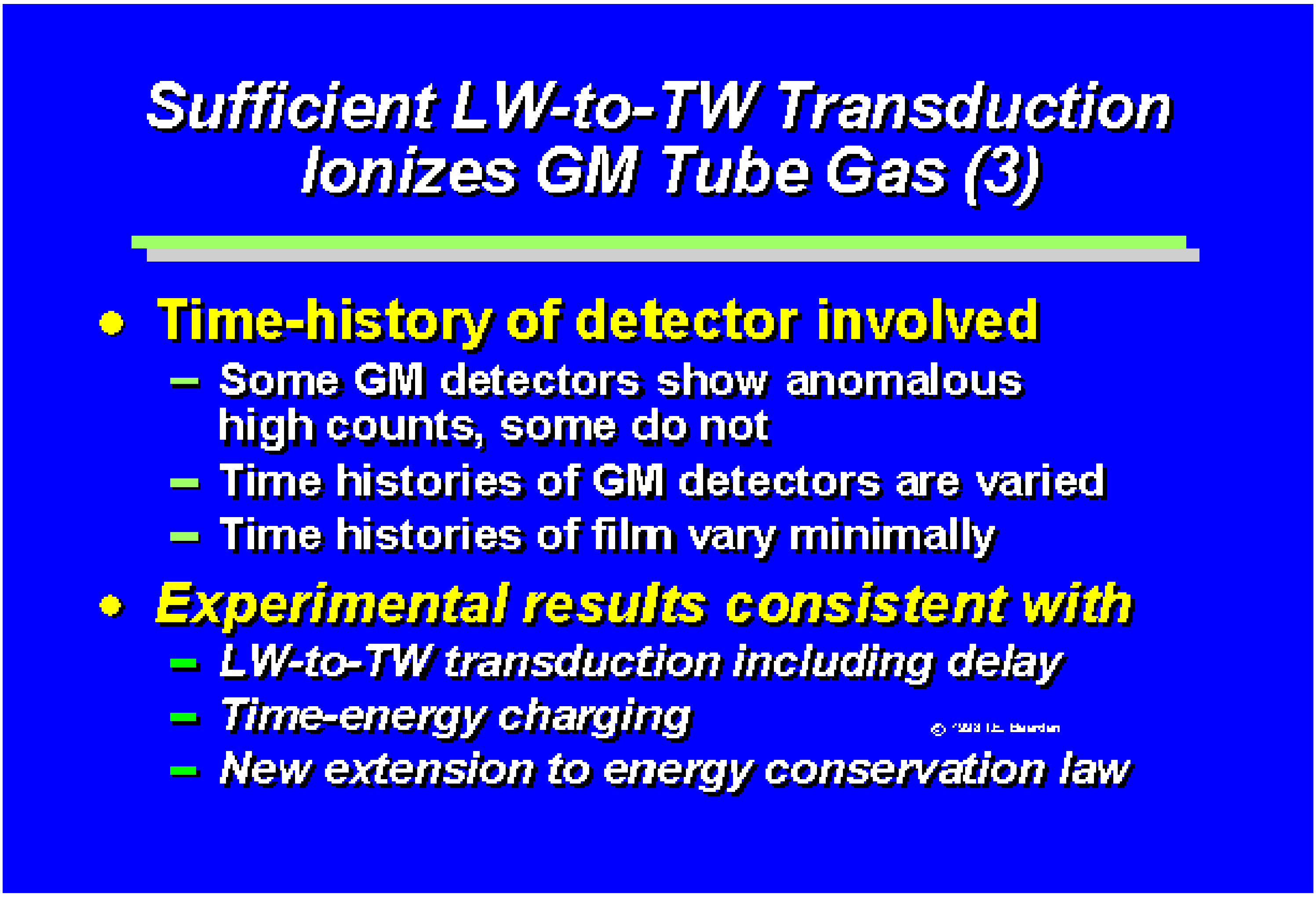

Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (3)

- **Time-history of detector involved**
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors are varied
  - Time histories of film vary minimally
- ***Experimental results consistent with***
  - *LW-to-TW transduction including delay*
  - *Time-energy charging*
  - *New extension to energy conservation law*

Published by Lotus® Freelance Graphics®

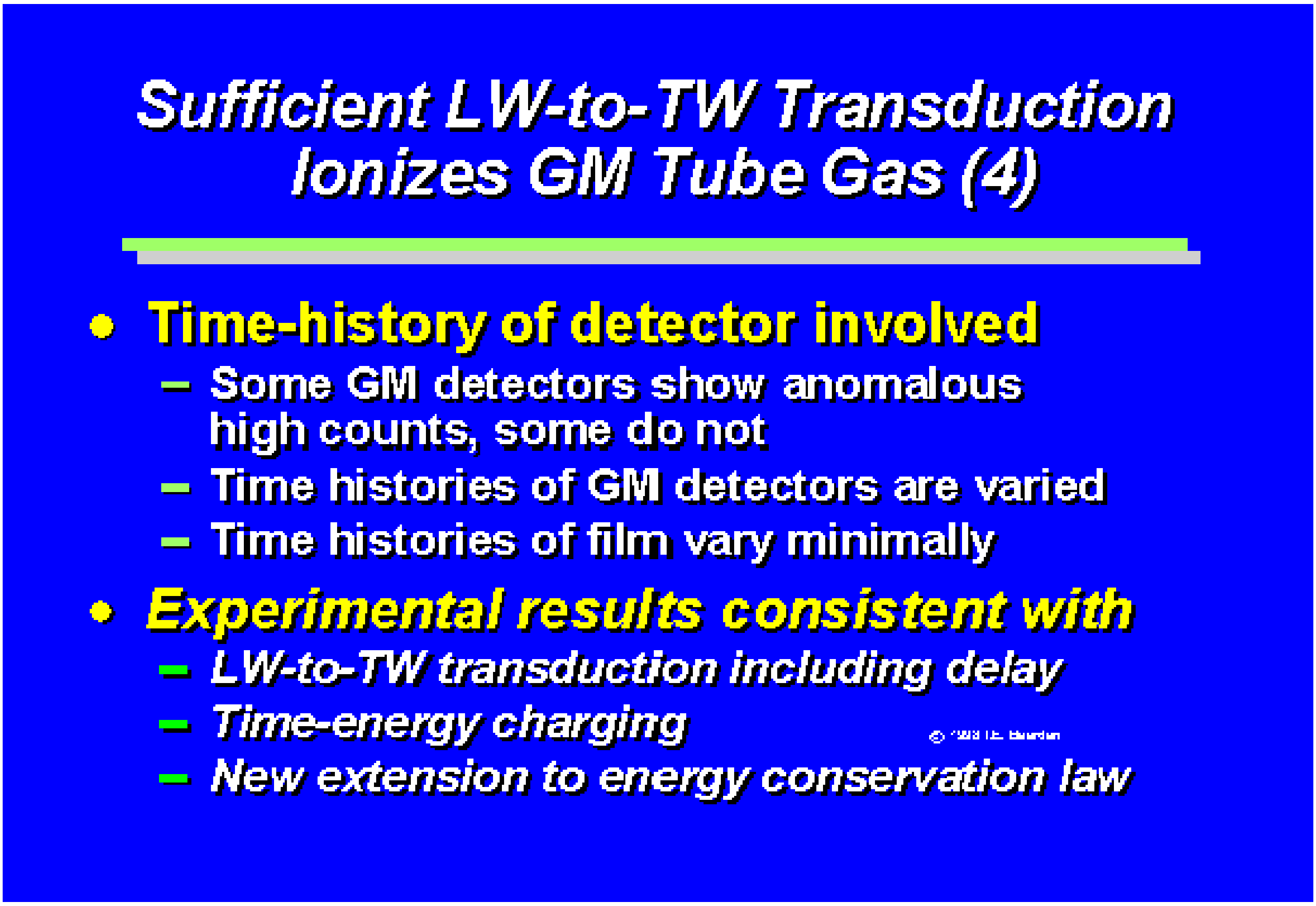

Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (4)

- **Time-history of detector involved**
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors are varied
  - Time histories of film vary minimally
- ***Experimental results consistent with***
  - *LW-to-TW transduction including delay*
  - *Time-energy charging*
  - *New extension to energy conservation law*

© 2003 T.E. Bearden

Published by Lotus® Freelance Graphics®

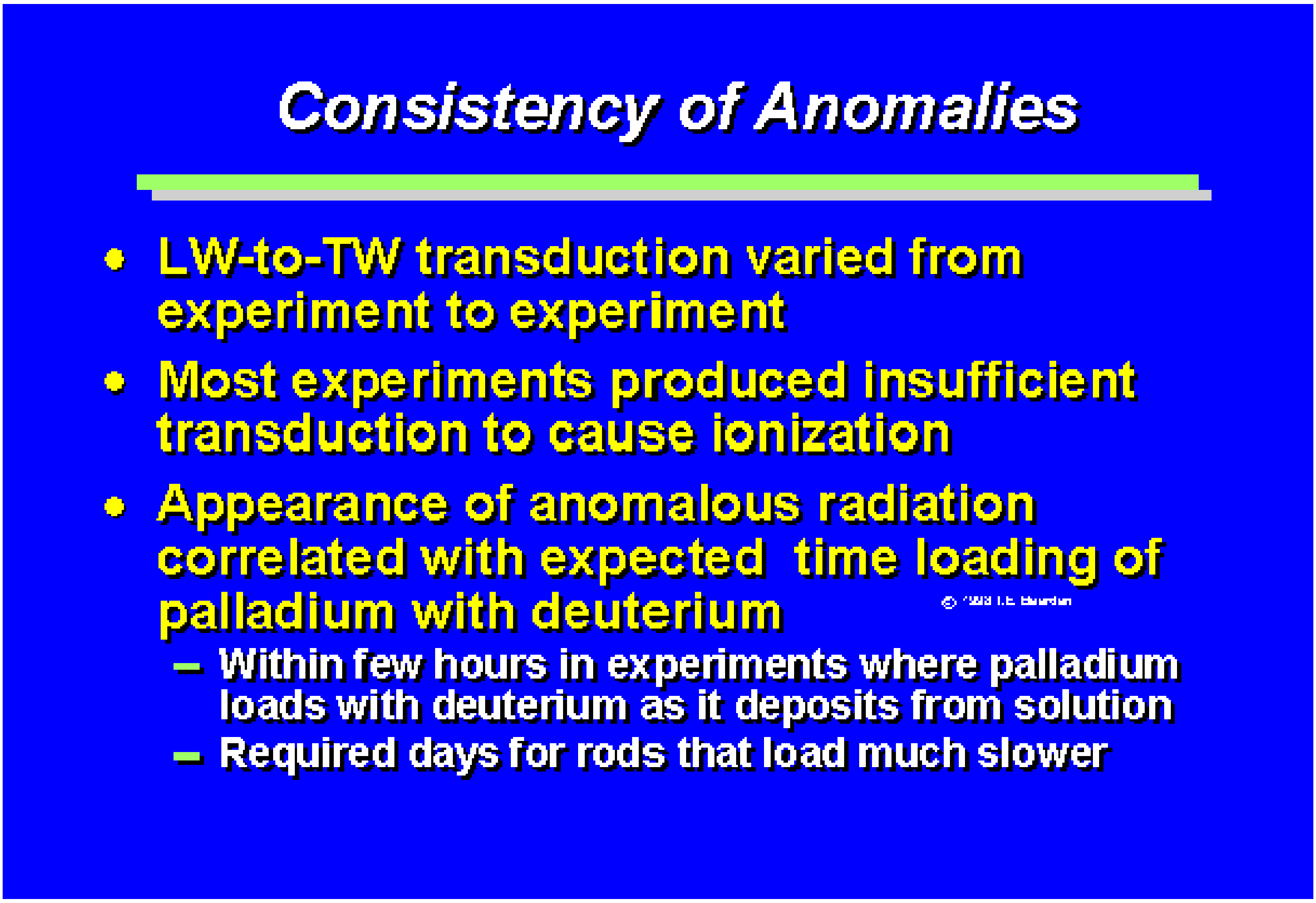

Consistency of Anomalies

- LW-to-TW transduction varied from experiment to experiment
- Most experiments produced insufficient transduction to cause ionization
- Appearance of anomalous radiation correlated with expected time loading of palladium with deuterium
  - Within few hours in experiments where palladium loads with deuterium as it deposits from solution
  - Required days for rods that load much slower

Published by Lotus® Freelance Graphics®

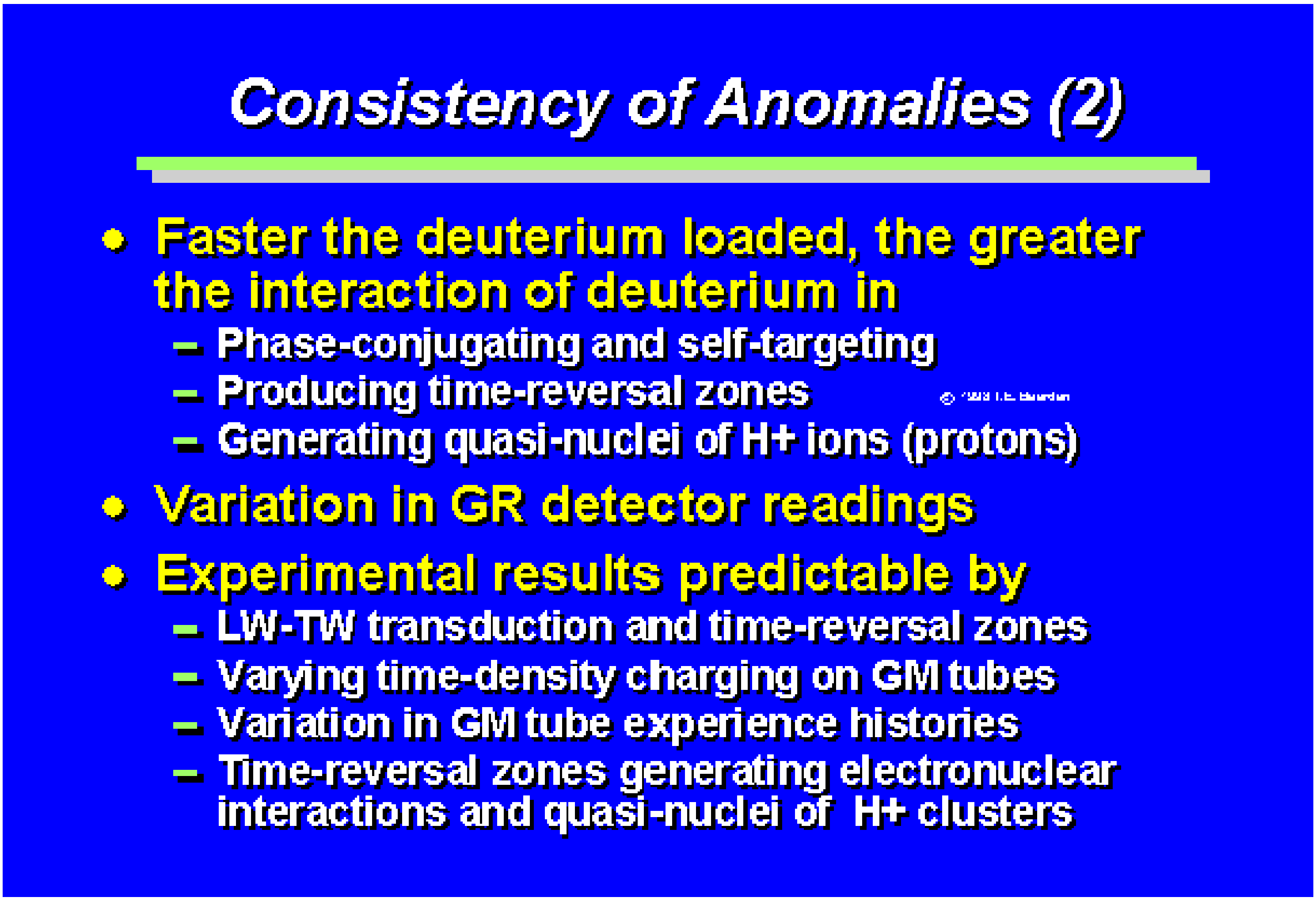

Consistency of Anomalies (2)

- Faster the deuterium loaded, the greater the interaction of deuterium in
  - Phase-conjugating and self-targeting
  - Producing time-reversal zones
  - Generating quasi-nuclei of H+ ions (protons)
- Variation in GR detector readings
- Experimental results predictable by
  - LW-TW transduction and time-reversal zones
  - Varying time-density charging on GM tubes
  - Variation in GM tube experience histories
  - Time-reversal zones generating electronuclear interactions and quasi-nuclei of H+ clusters

Published by Lotus® Freelance Graphics®

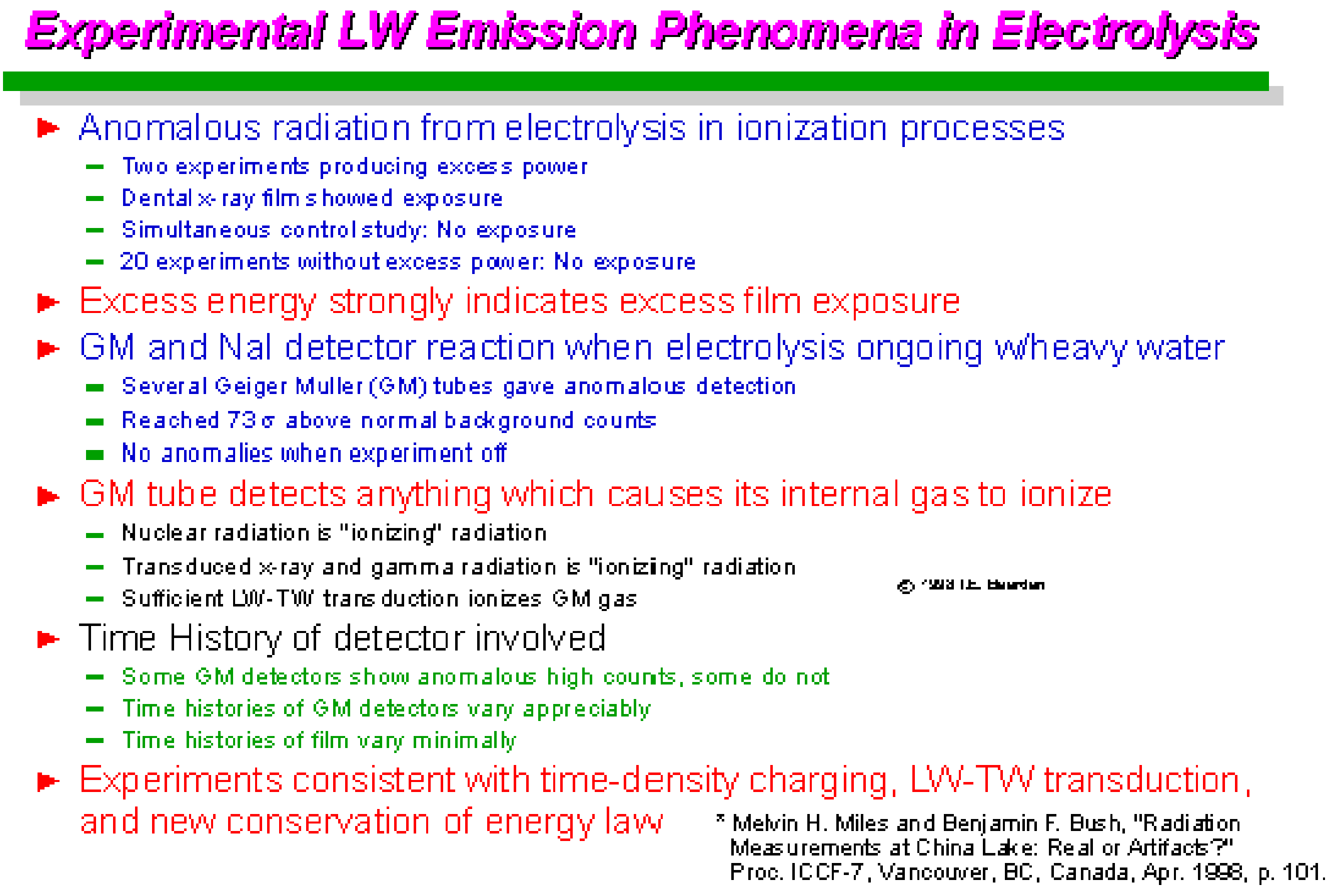

Experimental LW Emission Phenomena in Electrolysis

- Anomalous radiation from electrolysis in ionization processes
  - Two experiments producing excess power
  - Dental x-ray film showed exposure
  - Simultaneous control study: No exposure
  - 20 experiments without excess power: No exposure
- Excess energy strongly indicates excess film exposure
- GM and NaI detector reaction when electrolysis ongoing w/heavy water
  - Several Geiger Muller (GM) tubes gave anomalous detection
  - Reached 73 σ above normal background counts
  - No anomalies when experiment off
- GM tube detects anything which causes its internal gas to ionize
  - Nuclear radiation is "ionizing" radiation
  - Transduced x-ray and gamma radiation is "ionizing" radiation
  - Sufficient LW-TW transduction ionizes GM gas
- Time History of detector involved
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors vary appreciably
  - Time histories of film vary minimally
- Experiments consistent with time-density charging, LW-TW transduction, and new conservation of energy law

© 1998 T.E. Bearden

* Melvin H. Miles and Benjamin F. Bush, "Radiation Measurements at China Lake: Real or Artifacts?" Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 101.

Published by Lotus® Freelance Graphics®

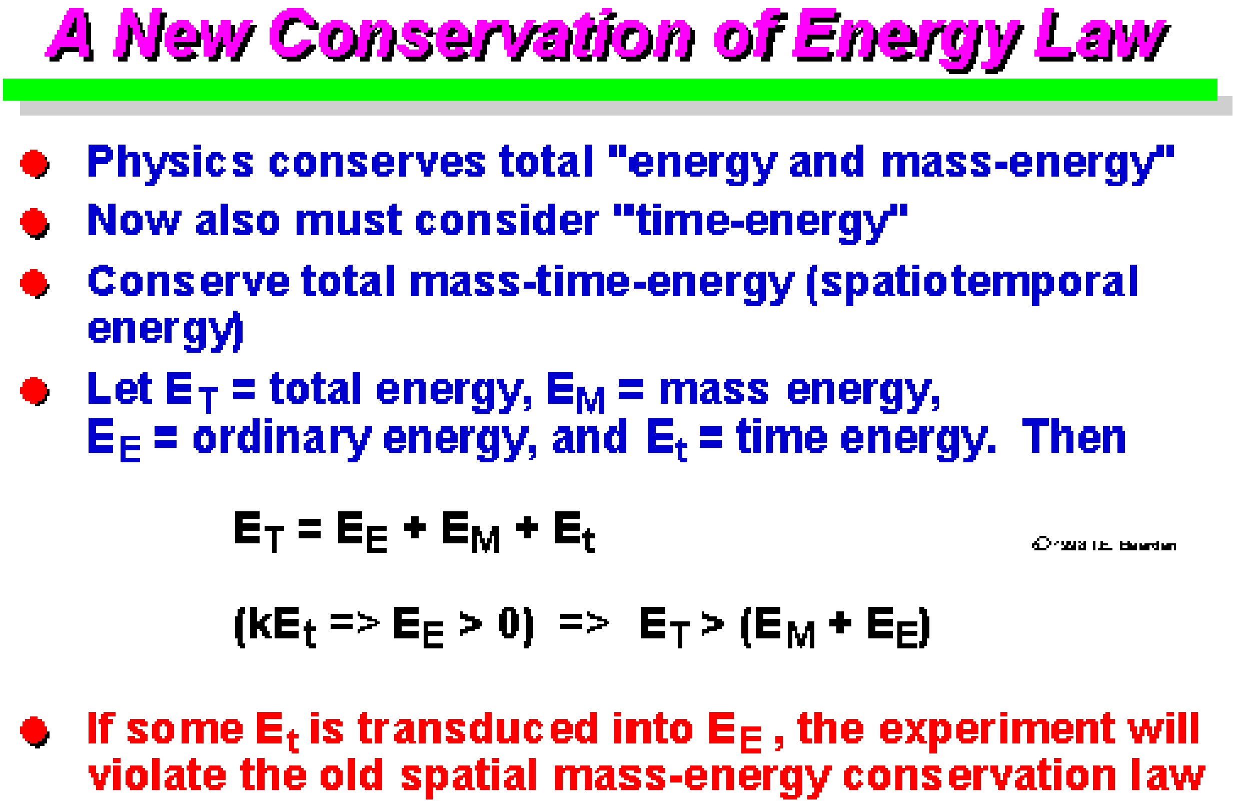

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

$$(kE_t \Rightarrow E_E > 0) \;\; \Rightarrow \;\; E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E , the experiment will violate the old spatial mass-energy conservation law

Published by Lotus® Freelance Graphics®

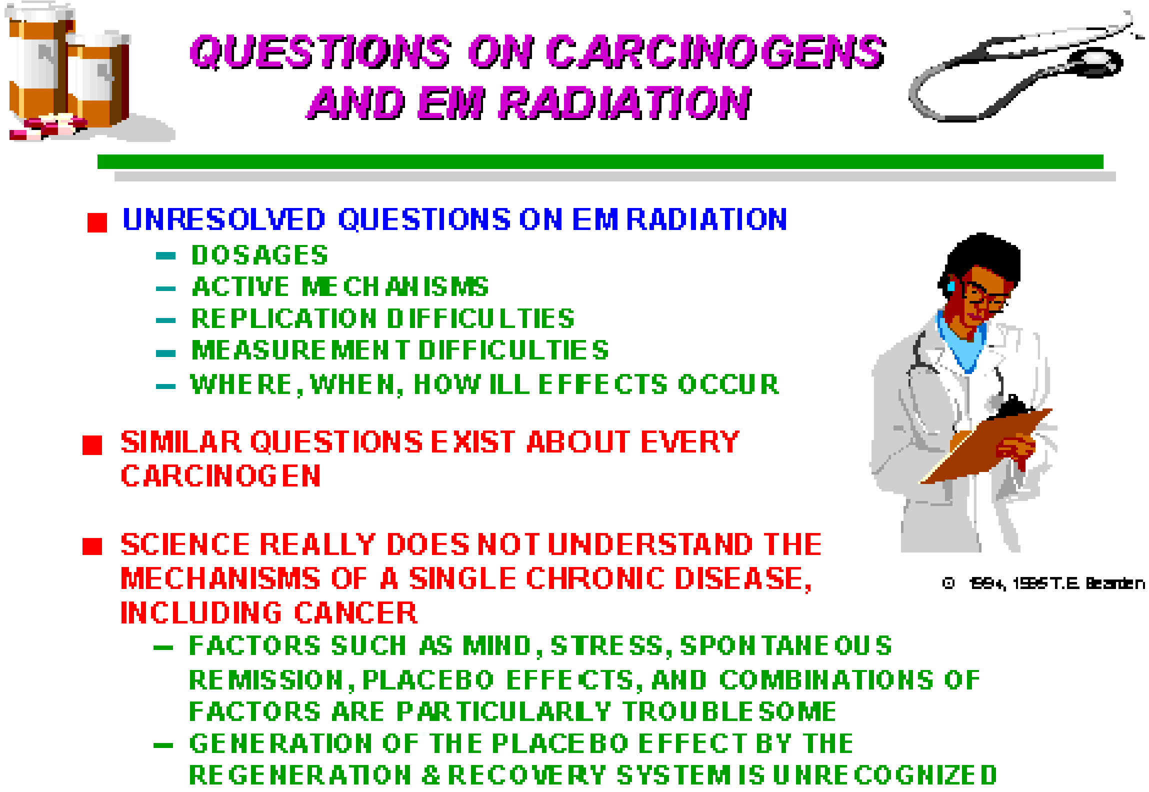

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1996 T.E. Bearden

Published by Lotus® Freelance Graphics®

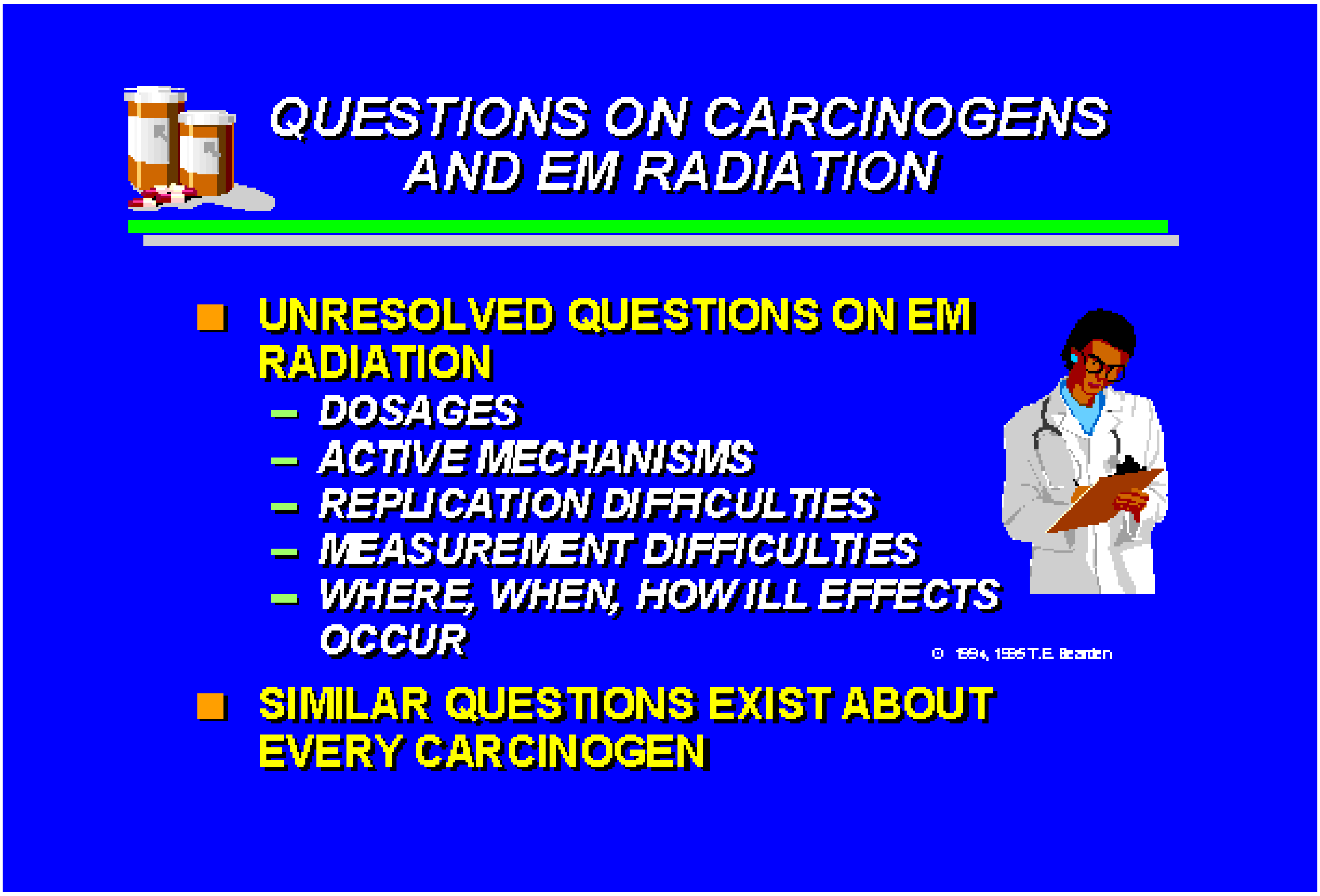

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN

© 1994, 1995 T.E. Bearden

Published by Lotus® Freelance Graphics®

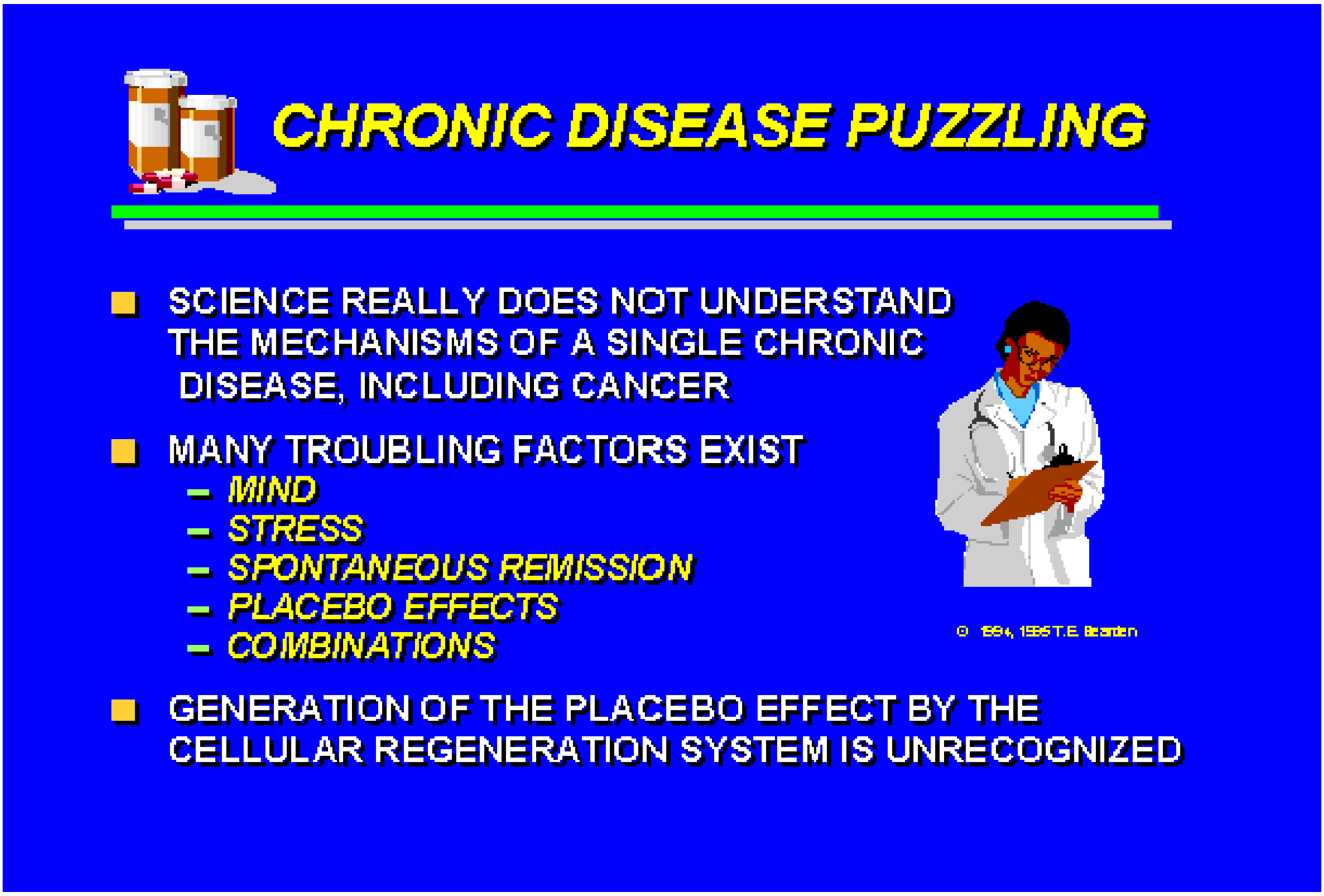

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1996 T.E. Bearden

Published by Lotus® Freelance Graphics®

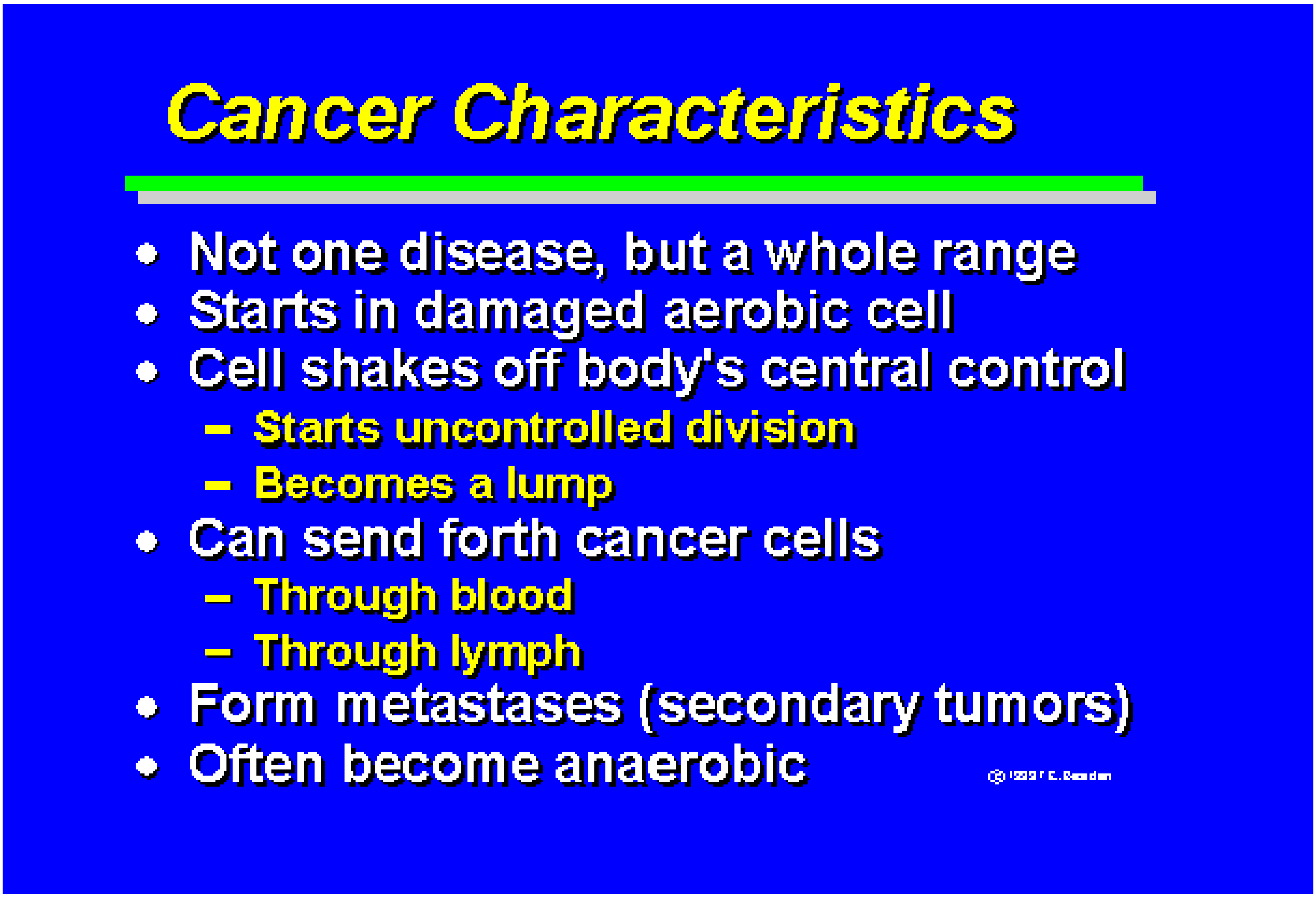

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

Published by Lotus® Freelance Graphics®

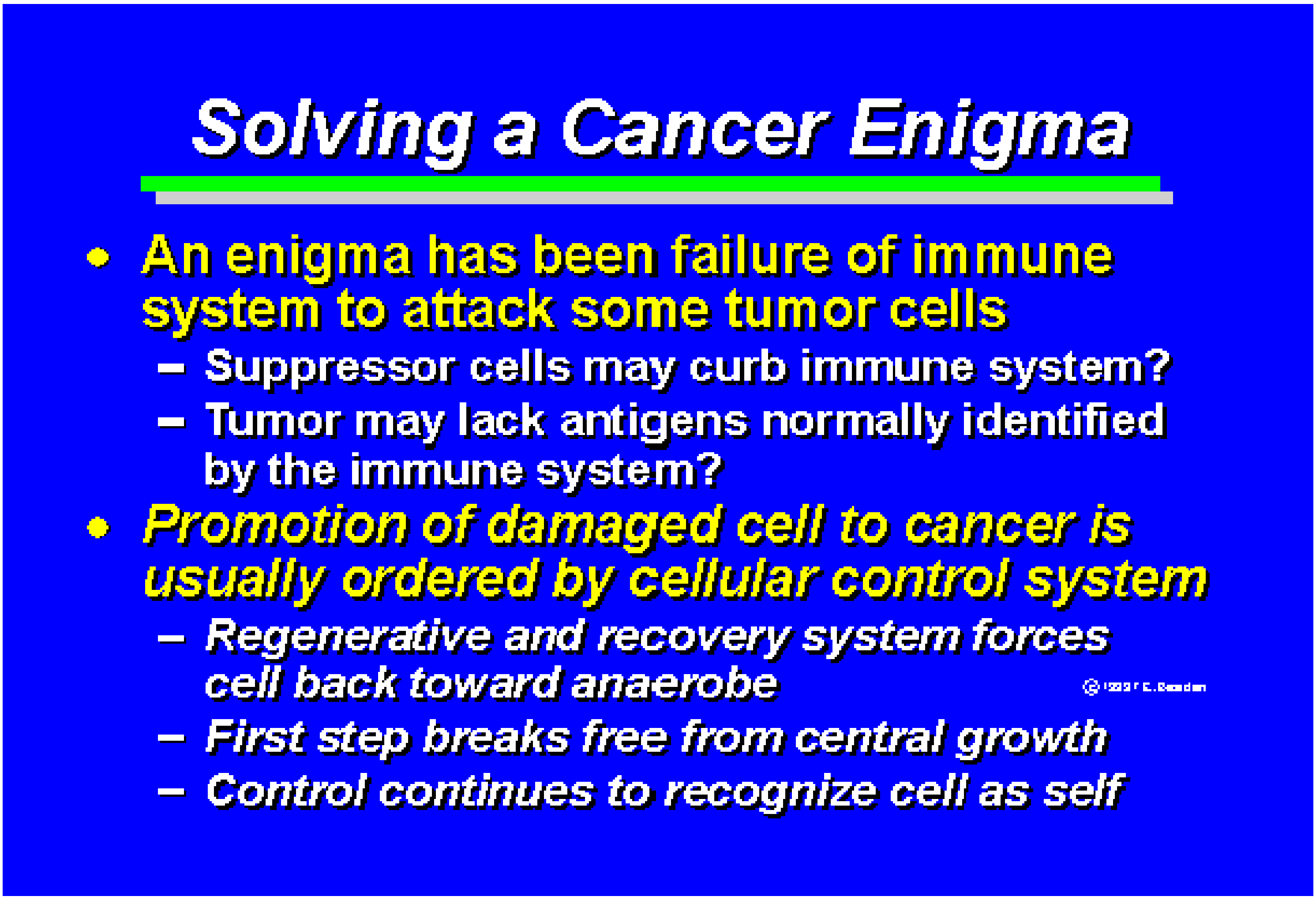

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

Published by Lotus® Freelance Graphics®

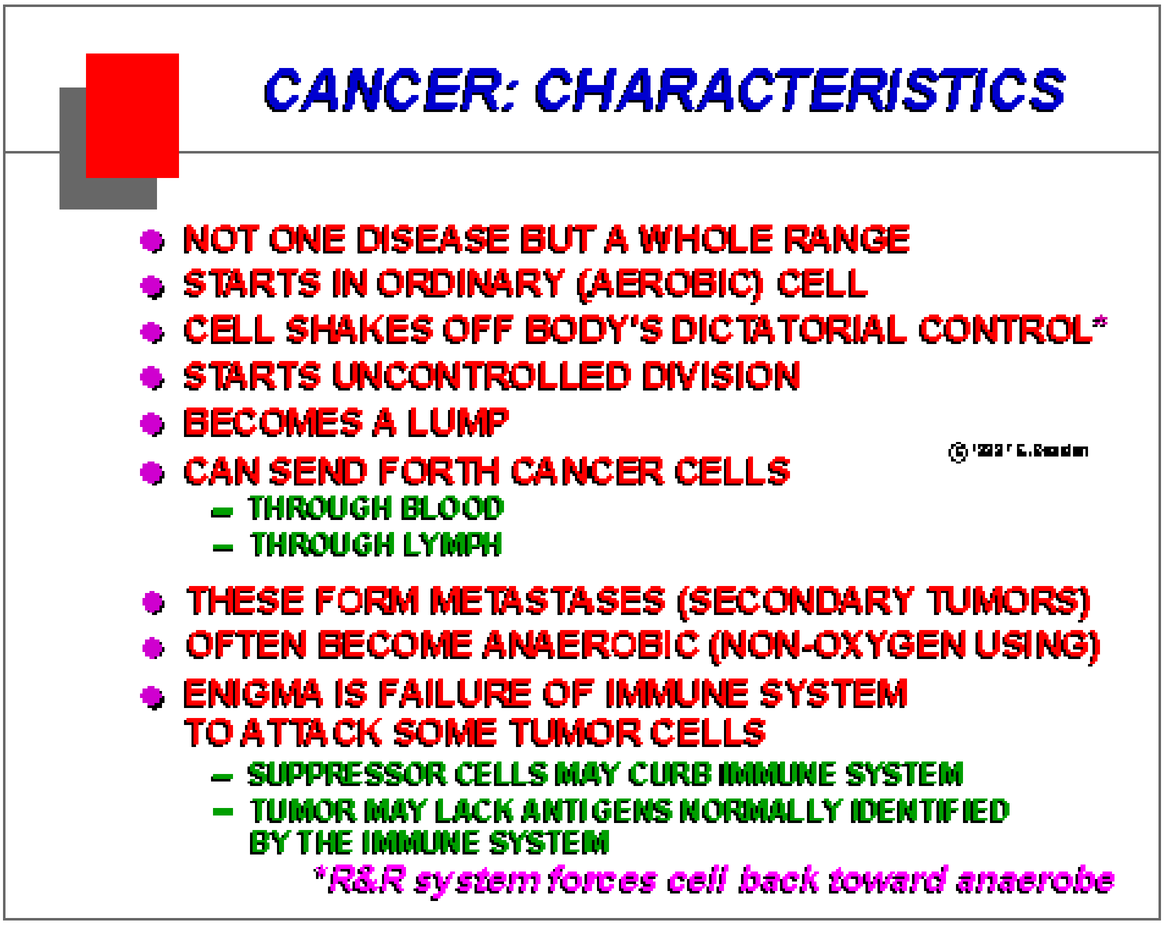

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

*R&R system forces cell back toward anaerobe

Published by Lotus® Freelance Graphics®

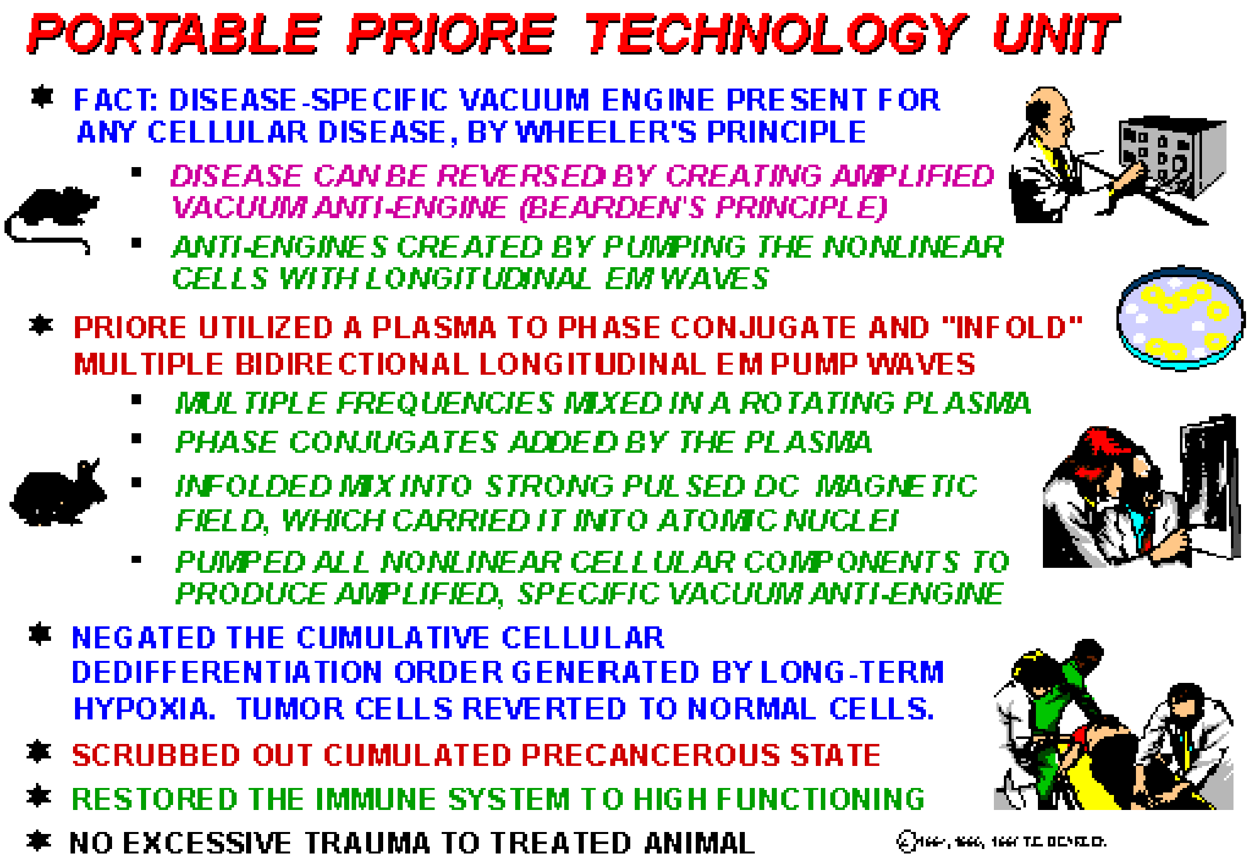

PORTABLE PRIORE TECHNOLOGY UNIT

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINES CREATED BY PUMPING THE NONLINEAR CELLS WITH LONGITUDINAL EM WAVES*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *MULTIPLE FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

Published by Lotus® Freelance Graphics®

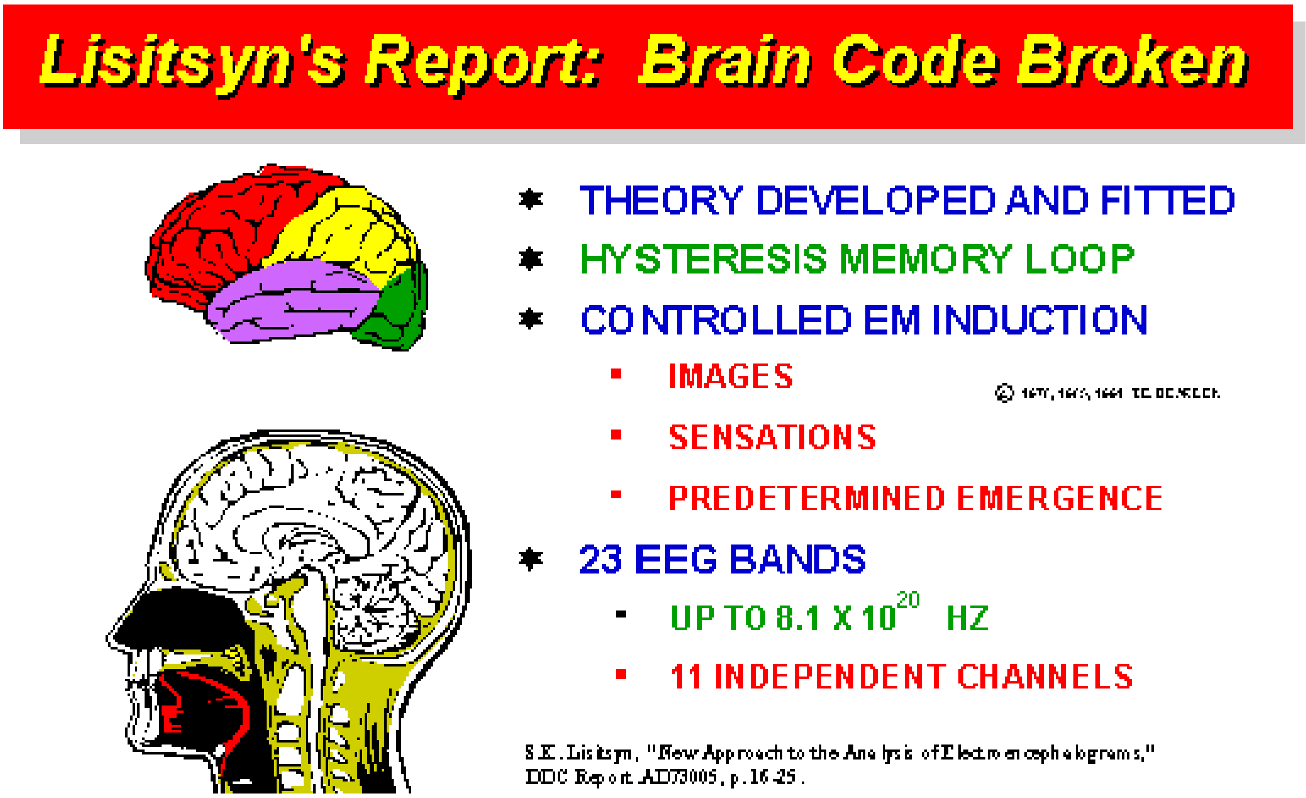

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© [illegible]

S.E. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

Published by Lotus® Freelance Graphics®

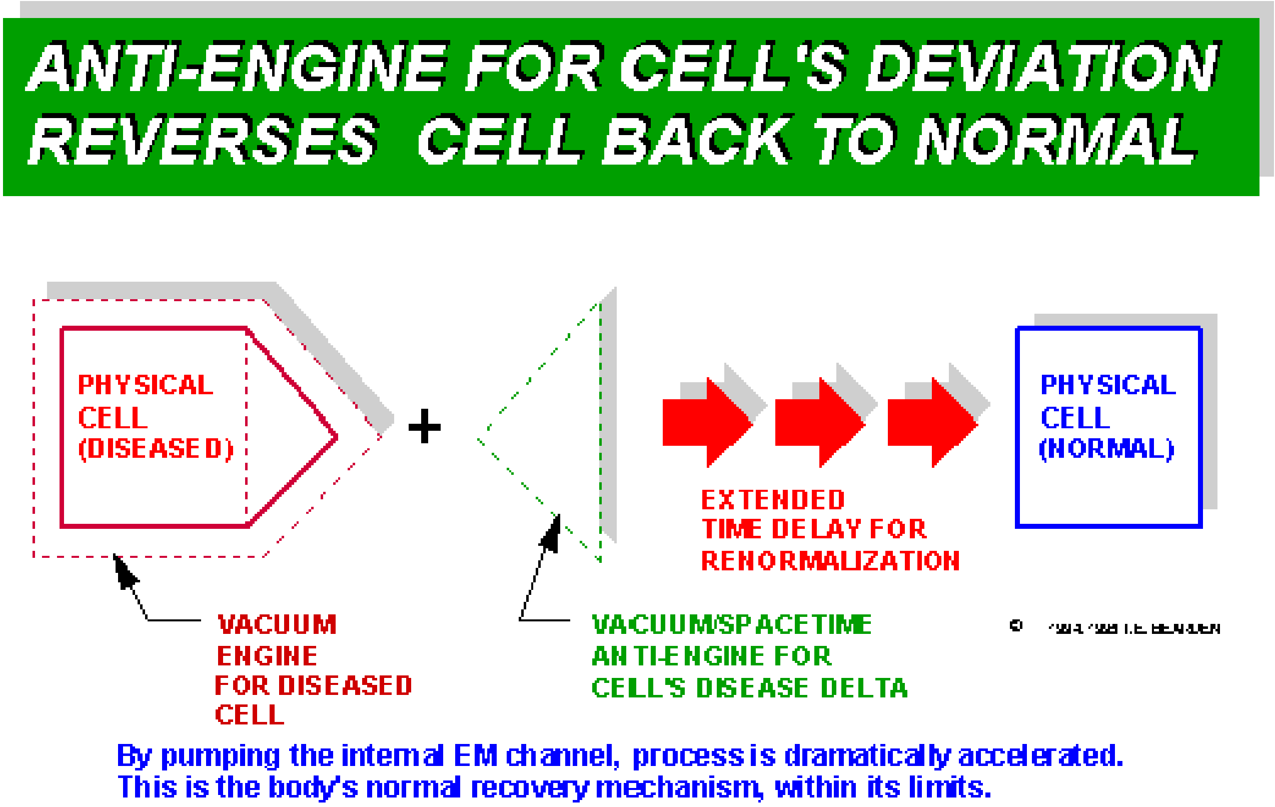

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL
CELL
(DISEASED)
+
EXTENDED
TIME DELAY FOR
RENORMALIZATION
PHYSICAL
CELL
(NORMAL)
VACUUM
ENGINE
FOR DISEASED
CELL
VACUUM/SPACETIME
ANTI-ENGINE FOR
CELL'S DISEASE DELTA
© [illegible] T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated. This is the body's normal recovery mechanism, within its limits.

Published by Lotus® Freelance Graphics®

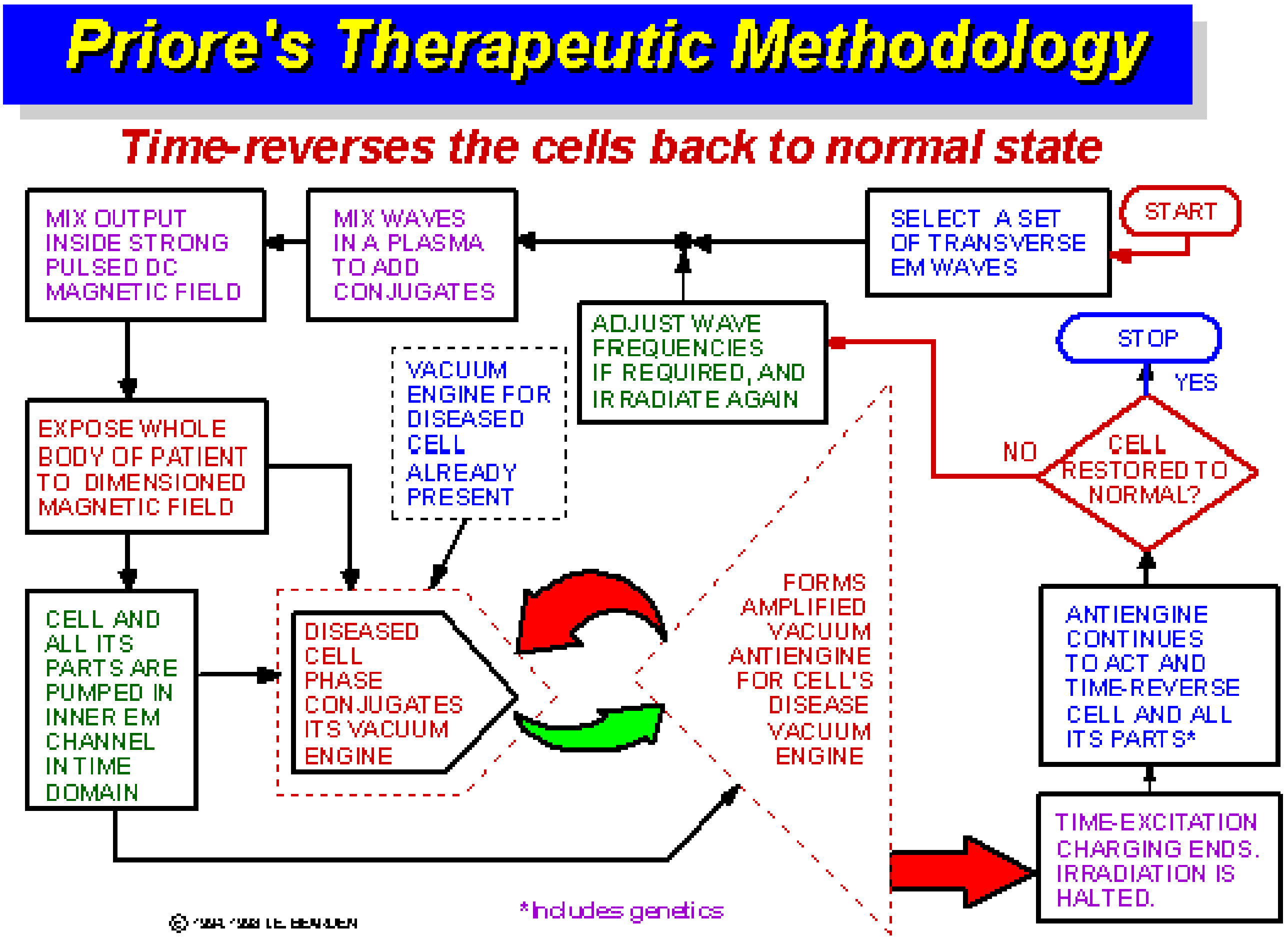

Priore's Therapeutic Methodology

Time-reverses the cells back to normal state

START

SELECT A SET OF TRANSVERSE EM WAVES

MIX WAVES IN A PLASMA TO ADD CONJUGATES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN

TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.

ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*

CELL RESTORED TO NORMAL?

NO

YES

STOP

*Includes genetics

Published by Lotus® Freelance Graphics®

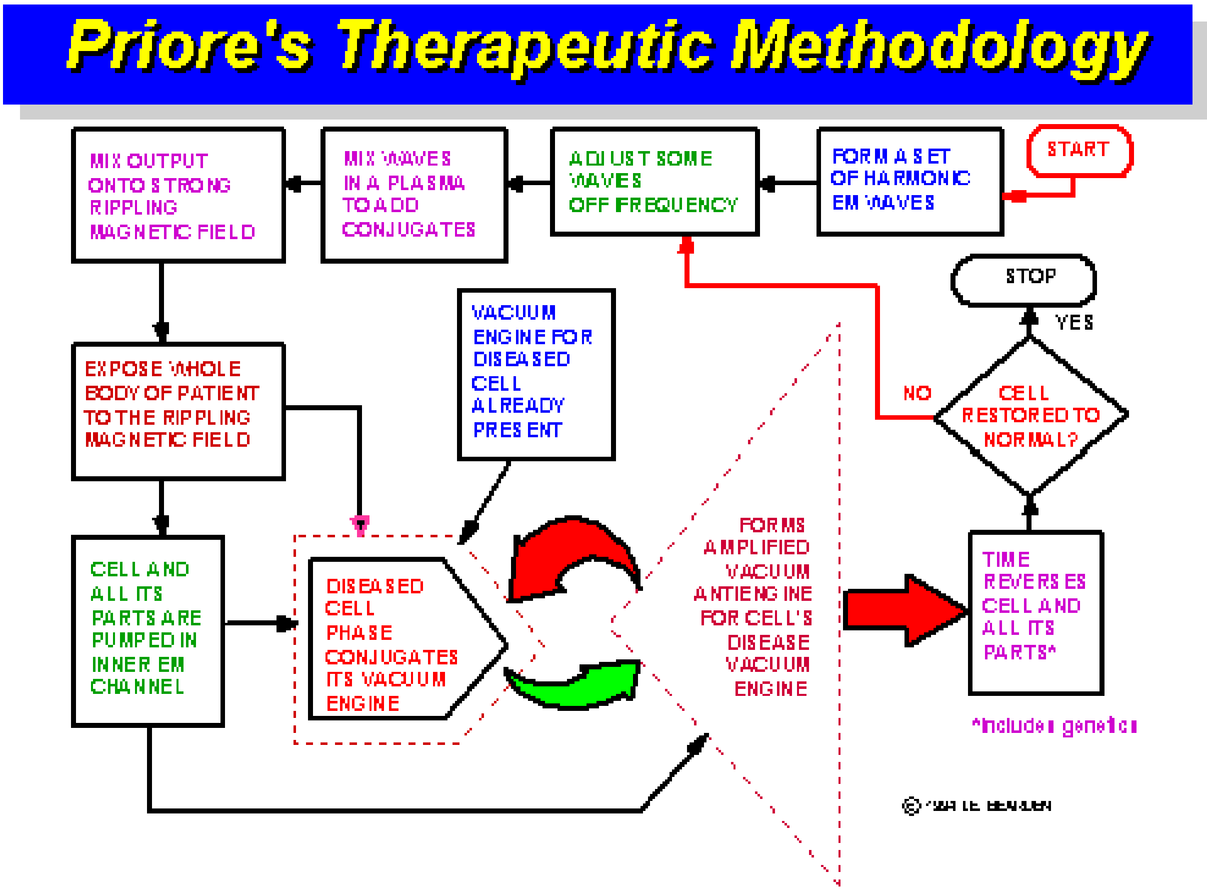
Priore's Therapeutic Methodology
START
FORM A SET OF HARMONIC EM WAVES
ADJUST SOME WAVES OFF FREQUENCY
MIX WAVES IN A PLASMA TO ADD CONJUGATES
MIX OUTPUT ONTO STRONG RIPPLING MAGNETIC FIELD
EXPOSE WHOLE BODY OF PATIENT TO THE RIPPLING MAGNETIC FIELD
CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL
VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT
DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE
FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE
TIME REVERSES CELL AND ALL ITS PARTS*
CELL RESTORED TO NORMAL?
NO
YES
STOP
*Includes genetics
© 1984 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

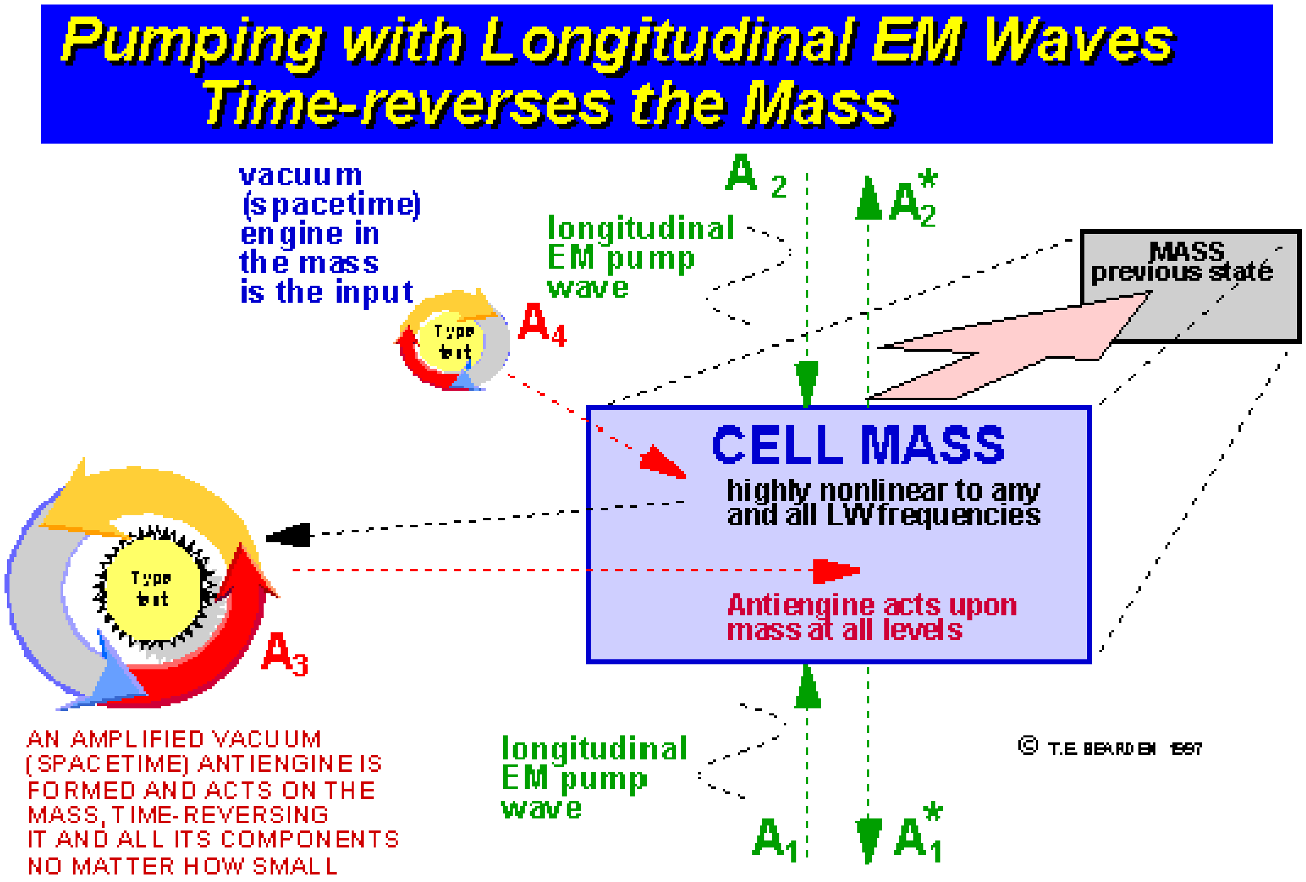
Pumping with Longitudinal EM Waves
Time-reverses the Mass
vacuum (spacetime) engine in the mass is the input
Type text
A4
longitudinal EM pump wave
A 2
A*2
MASS previous state
CELL MASS
highly nonlinear to any and all LW frequencies
Antiengine acts upon mass at all levels
Type text
A3
AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL
longitudinal EM pump wave
A1
A*1
© T.E. BEARDEN 1997

Published by Lotus® Freelance Graphics®

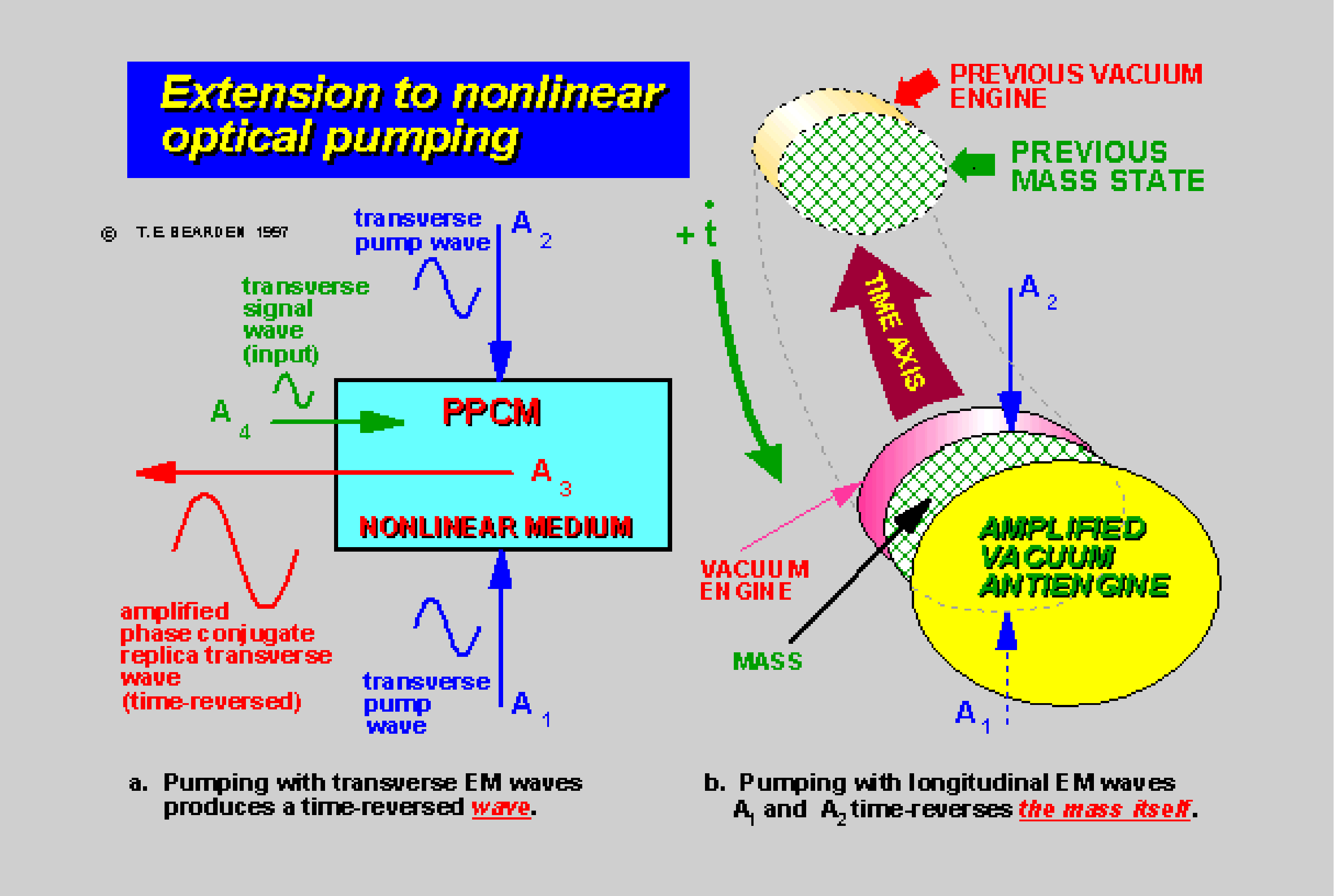

Extension to nonlinear optical pumping

© T.E.BEARDEN 1997
transverse pump wave
A 2
transverse signal wave (input)
A 4
PPCM
A 3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A 1
+ t
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
TIME AXIS
A 2
VACUUM ENGINE
AMPLIFIED VACUUM ANTIENGINE
MASS
A 1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses ***the mass itself***.

Published by Lotus® Freelance Graphics®

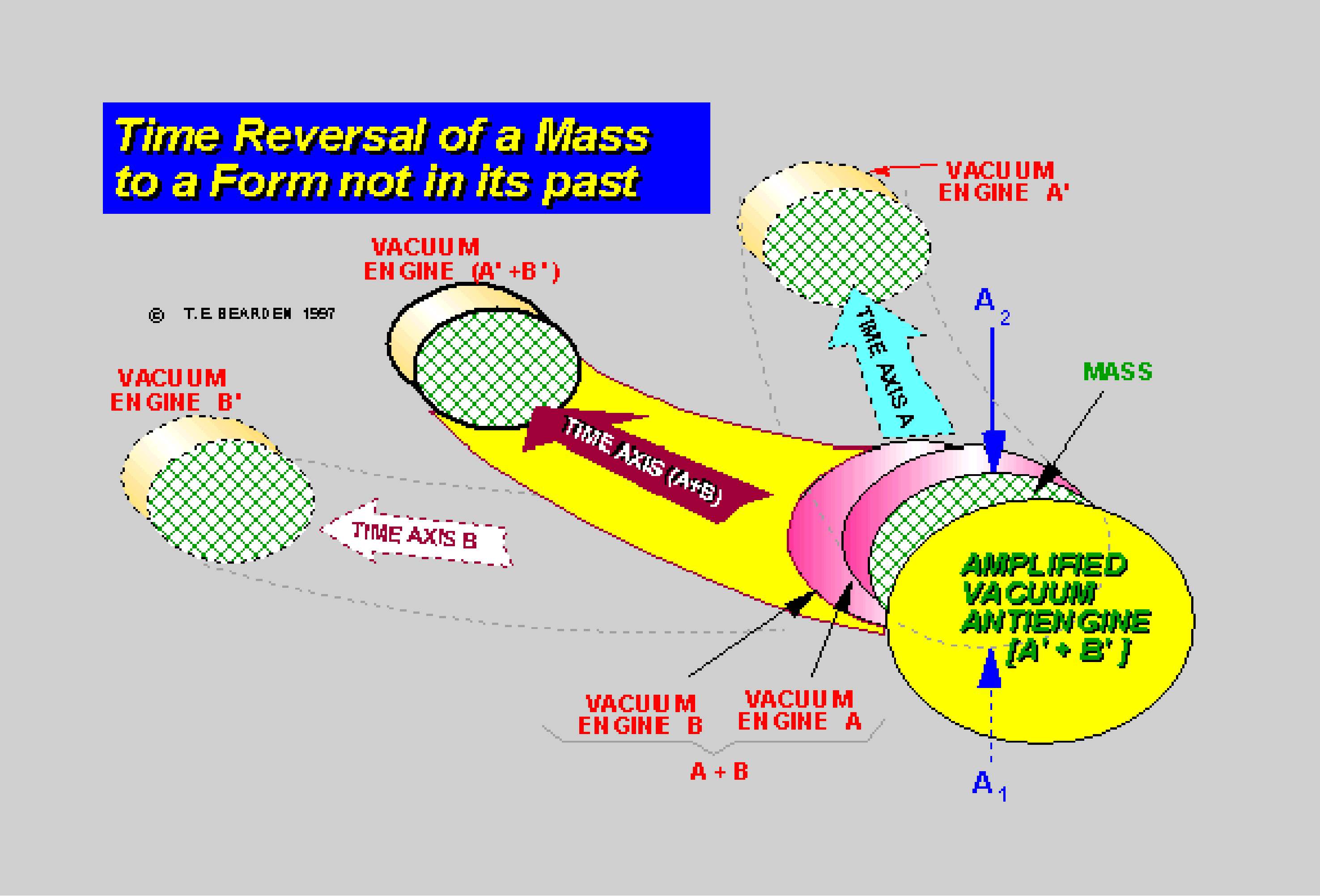

Time Reversal of a Mass to a Form not in its past

© T.E. BEARDEN 1997
VACUUM ENGINE (A' +B')
VACUUM ENGINE A'
VACUUM ENGINE B'
TIME AXIS A
TIME AXIS (A+B)
TIME AXIS B
A 2
MASS
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A + B
A 1

Published by Lotus® Freelance Graphics®

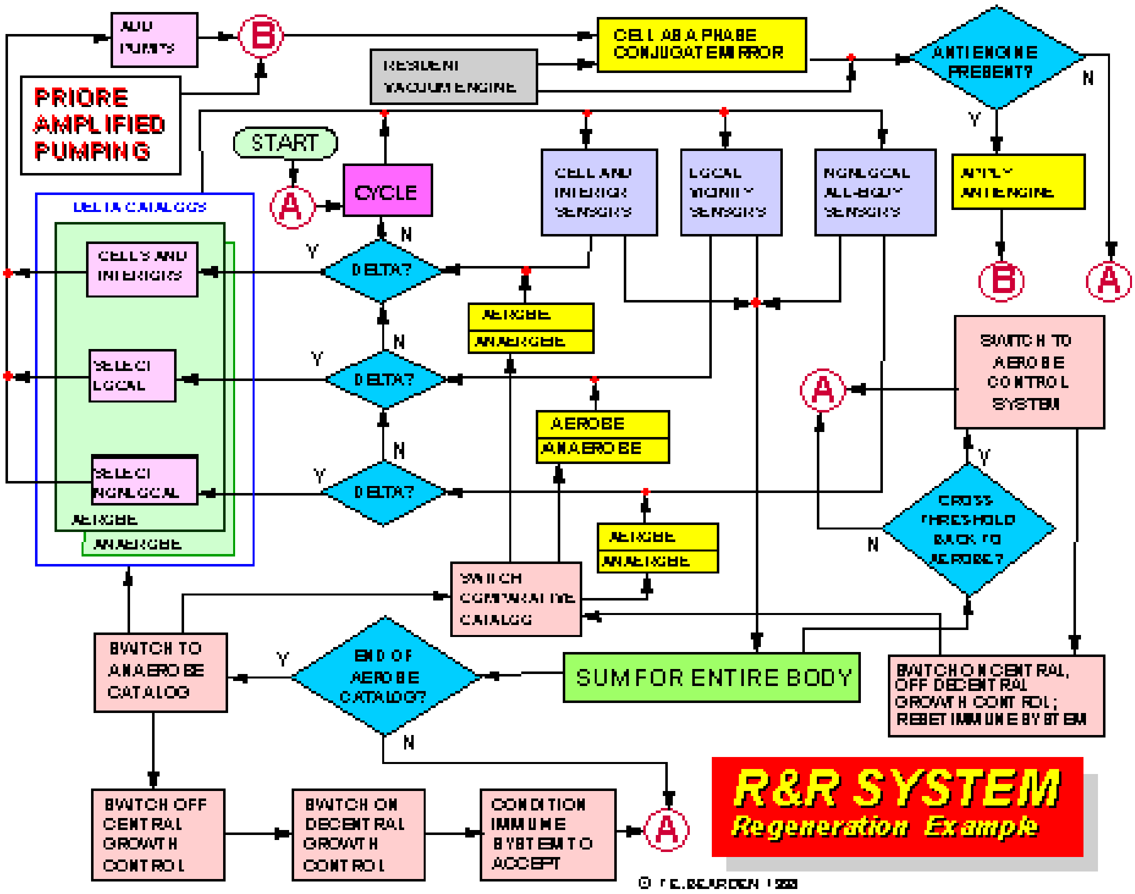
ADD PUMPS
B
PRIORE AMPLIFIED PUMPING
RESIDENT VACUUM ENGINE
CELL AS A PHASE CONJUGATE MIRROR
ANTIENGINE PRESENT?
N
Y
APPLY ANTIENGINE
B
A
START
A
CYCLE
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NONLOCAL ALL-BODY SENSORS
DELTA CATALOGS
CELLS AND INTERIORS
SELECT LOCAL
SELECT NONLOCAL
AEROBE
ANAEROBE
N
Y
DELTA?
N
Y
DELTA?
N
Y
DELTA?
AEROBE
ANAEROBE
AEROBE
ANAEROBE
AEROBE
ANAEROBE
SWITCH TO AEROBE CONTROL SYSTEM
A
Y
CROSS THRESHOLD BACK TO AEROBE?
N
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
Y
END OF AEROBE CATALOG?
N
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT
A
R&R SYSTEM
Regeneration Example
© T.E. BEARDEN

Published by Lotus® Freelance Graphics®

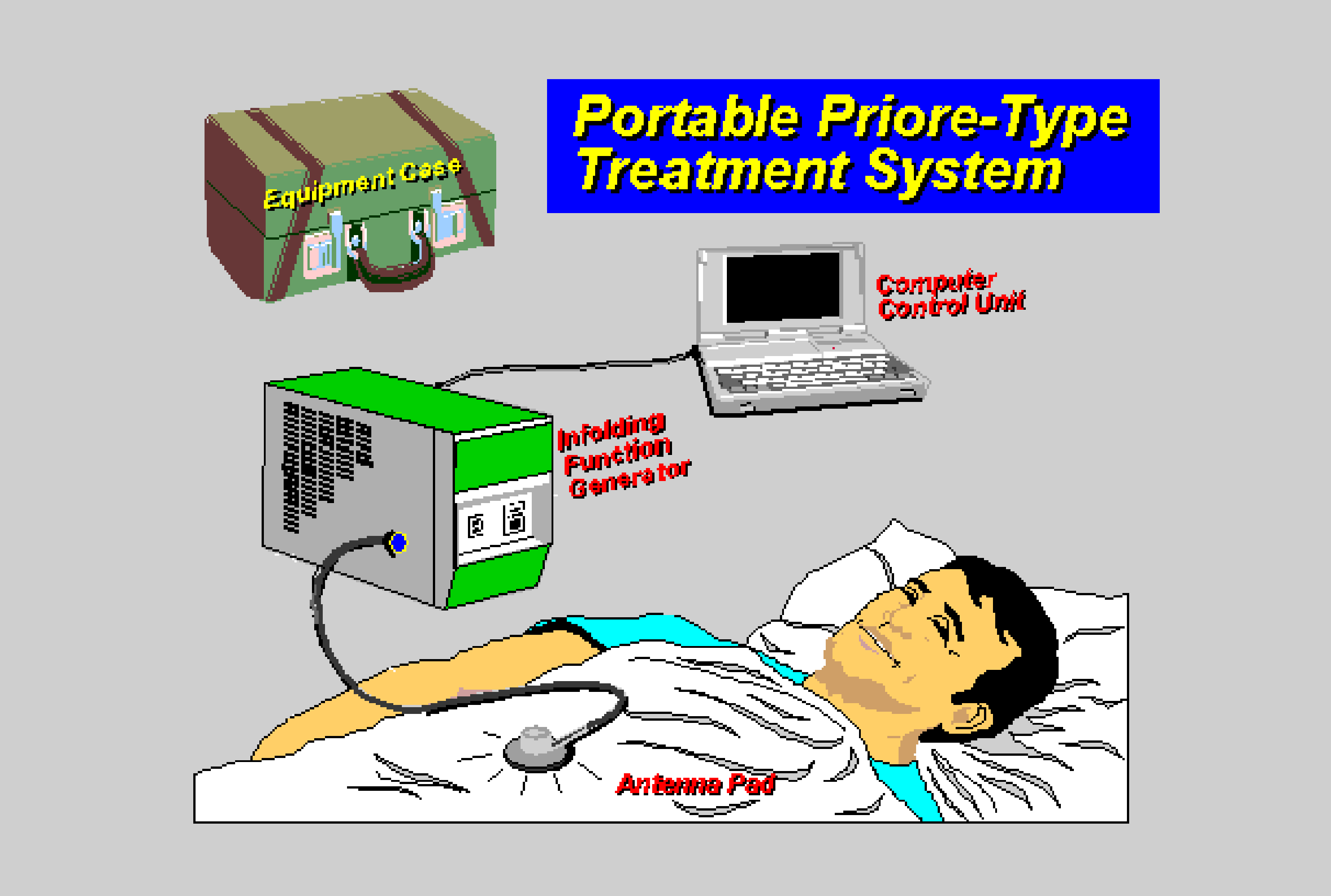
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Published by Lotus® Freelance Graphics®

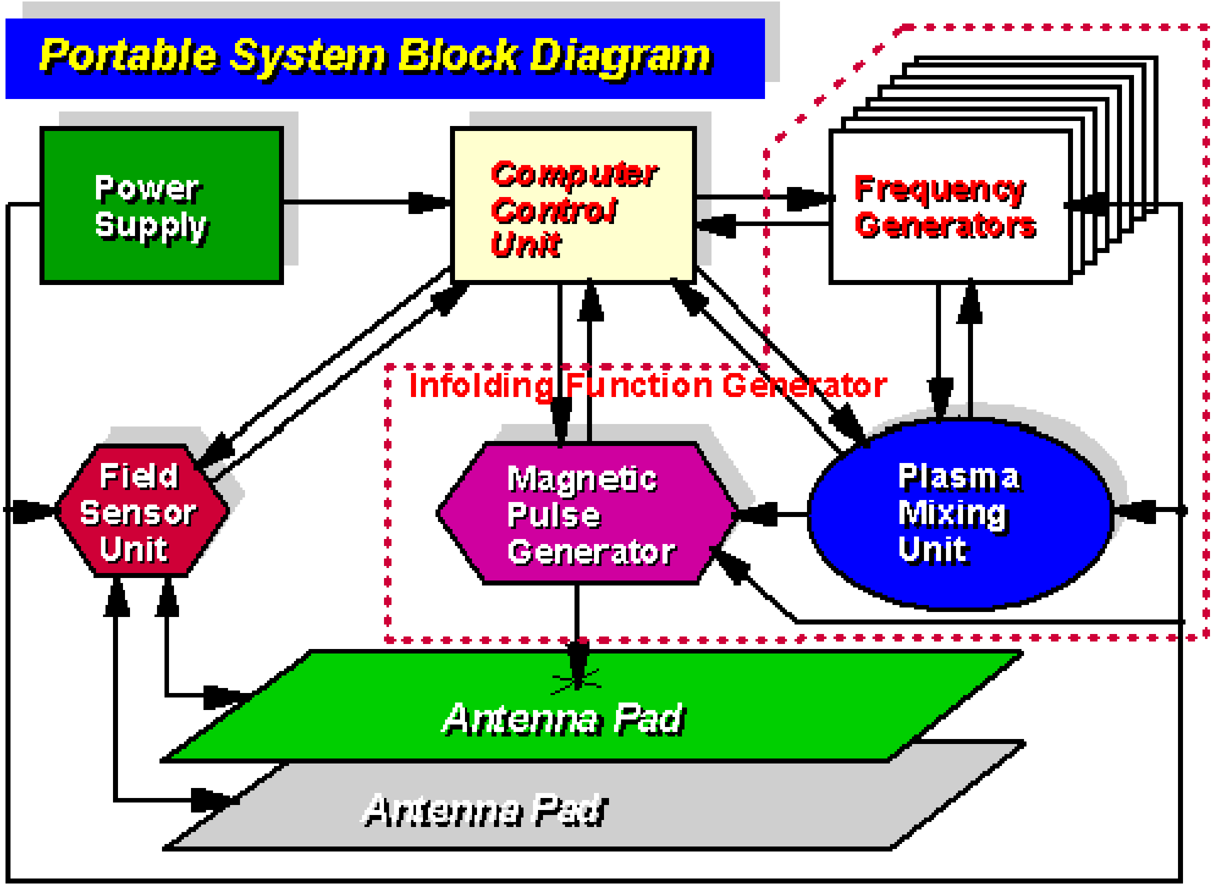
Portable System Block Diagram
Power Supply
Computer Control Unit
Frequency Generators
Infolding Function Generator
Field Sensor Unit
Magnetic Pulse Generator
Plasma Mixing Unit
Antenna Pad
Antenna Pad

Published by Lotus® Freelance Graphics®

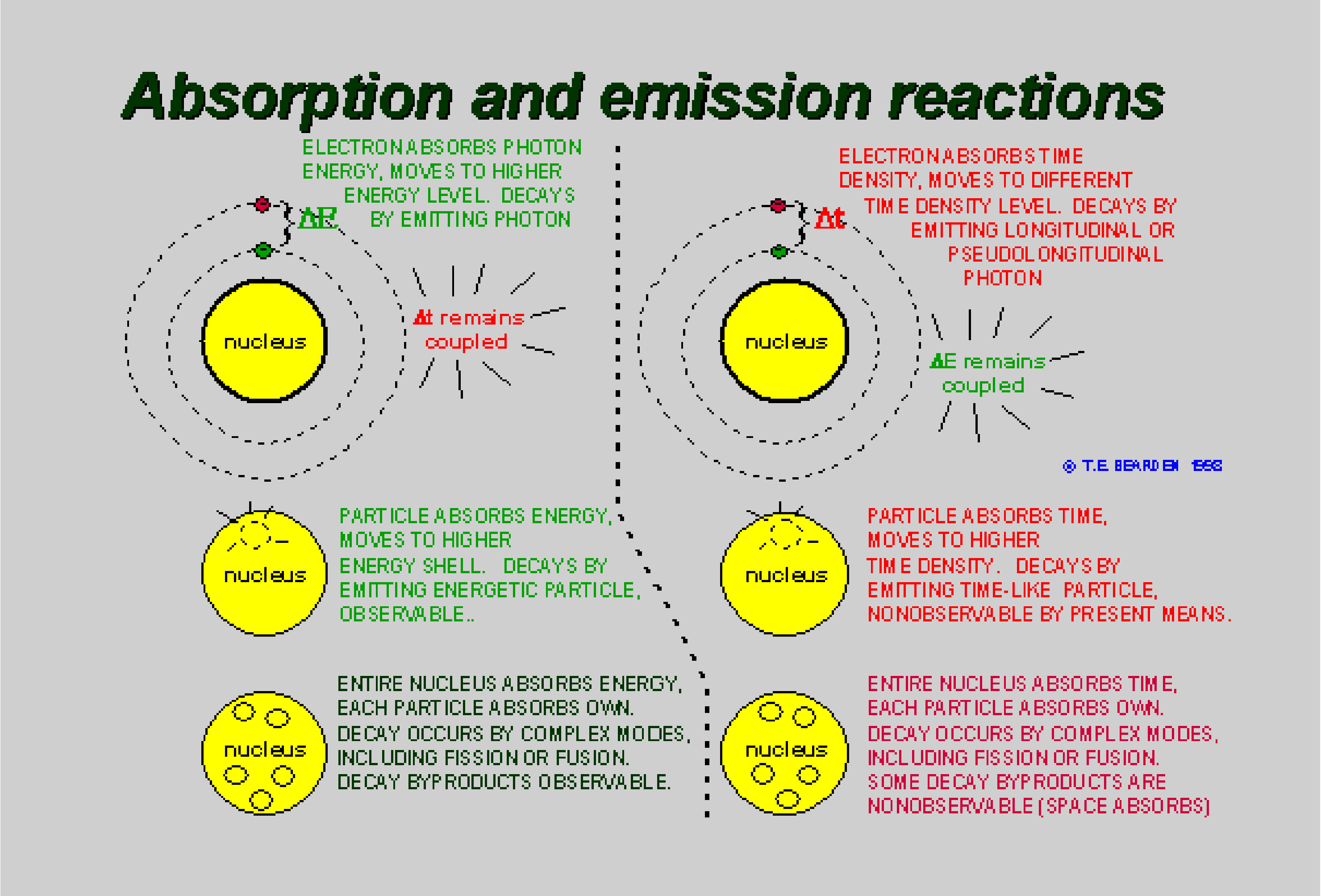

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

© T.E. BEARDEN 1998

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE, OBSERVABLE.

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

Published by Lotus® Freelance Graphics®

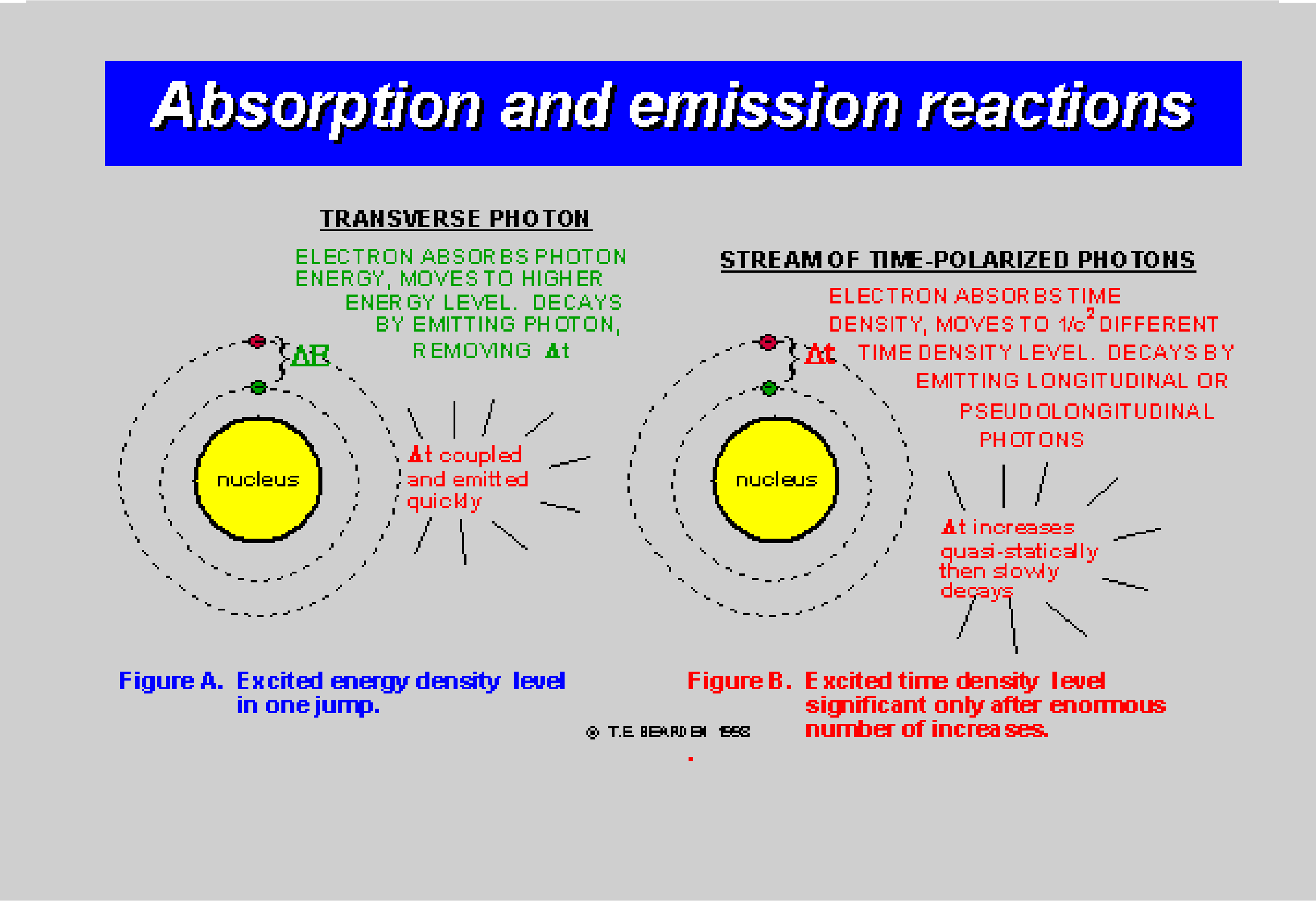

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

ΔE
nucleus
Δt coupled and emitted quickly

Figure A. Excited energy density level in one jump.

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c^2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt
nucleus
Δt increases quasi-statically then slowly decays

Figure B. Excited time density level significant only after enormous number of increases.

© T.E. BEARDEN 1998

Published by Lotus® Freelance Graphics®

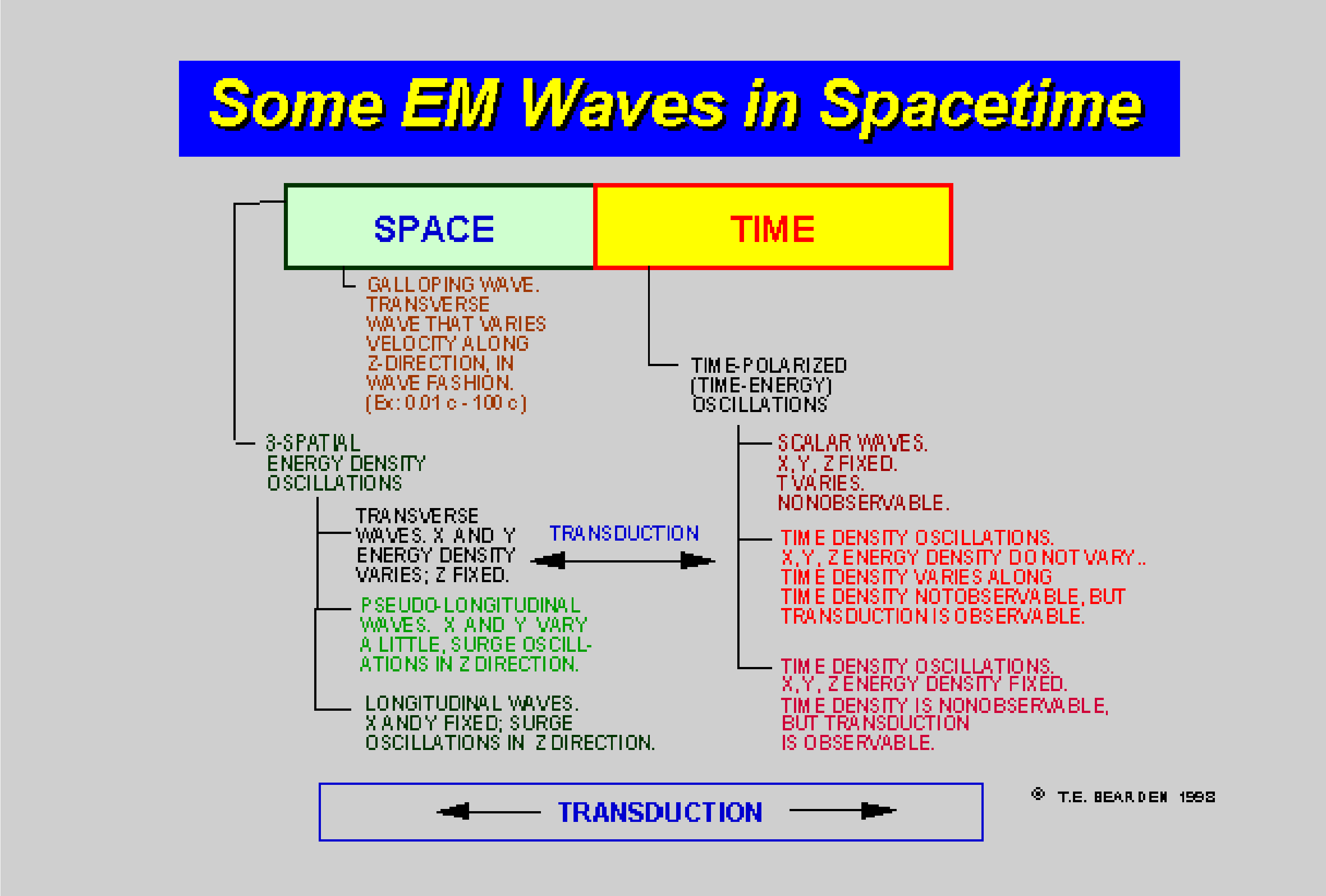

Some EM Waves in Spacetime

SPACE

TIME

GALLOPING WAVE.
TRANSVERSE
WAVE THAT VARIES
VELOCITY ALONG
Z-DIRECTION, IN
WAVE FASHION.
(Ex: 0.01 c - 100 c)

TIME-POLARIZED
(TIME-ENERGY)
OSCILLATIONS

3-SPATIAL
ENERGY DENSITY
OSCILLATIONS

TRANSVERSE
WAVES. X AND Y
ENERGY DENSITY
VARIES; Z FIXED.

TRANSDUCTION

PSEUDO-LONGITUDINAL
WAVES. X AND Y VARY
A LITTLE, SURGE OSCILL-
ATIONS IN Z DIRECTION.

LONGITUDINAL WAVES.
X AND Y FIXED; SURGE
OSCILLATIONS IN Z DIRECTION.

SCALAR WAVES.
X, Y, Z FIXED.
T VARIES.
NONOBSERVABLE.

TIME DENSITY OSCILLATIONS.
X,Y, Z ENERGY DENSITY DO NOT VARY..
TIME DENSITY VARIES ALONG
TIME DENSITY NOTOBSERVABLE, BUT
TRANSDUCTION IS OBSERVABLE.

TIME DENSITY OSCILLATIONS.
X,Y, Z ENERGY DENSITY FIXED.
TIME DENSITY IS NONOBSERVABLE,
BUT TRANSDUCTION
IS OBSERVABLE.

TRANSDUCTION

© T.E. BEARDEN 1998

Published by Lotus® Freelance Graphics®

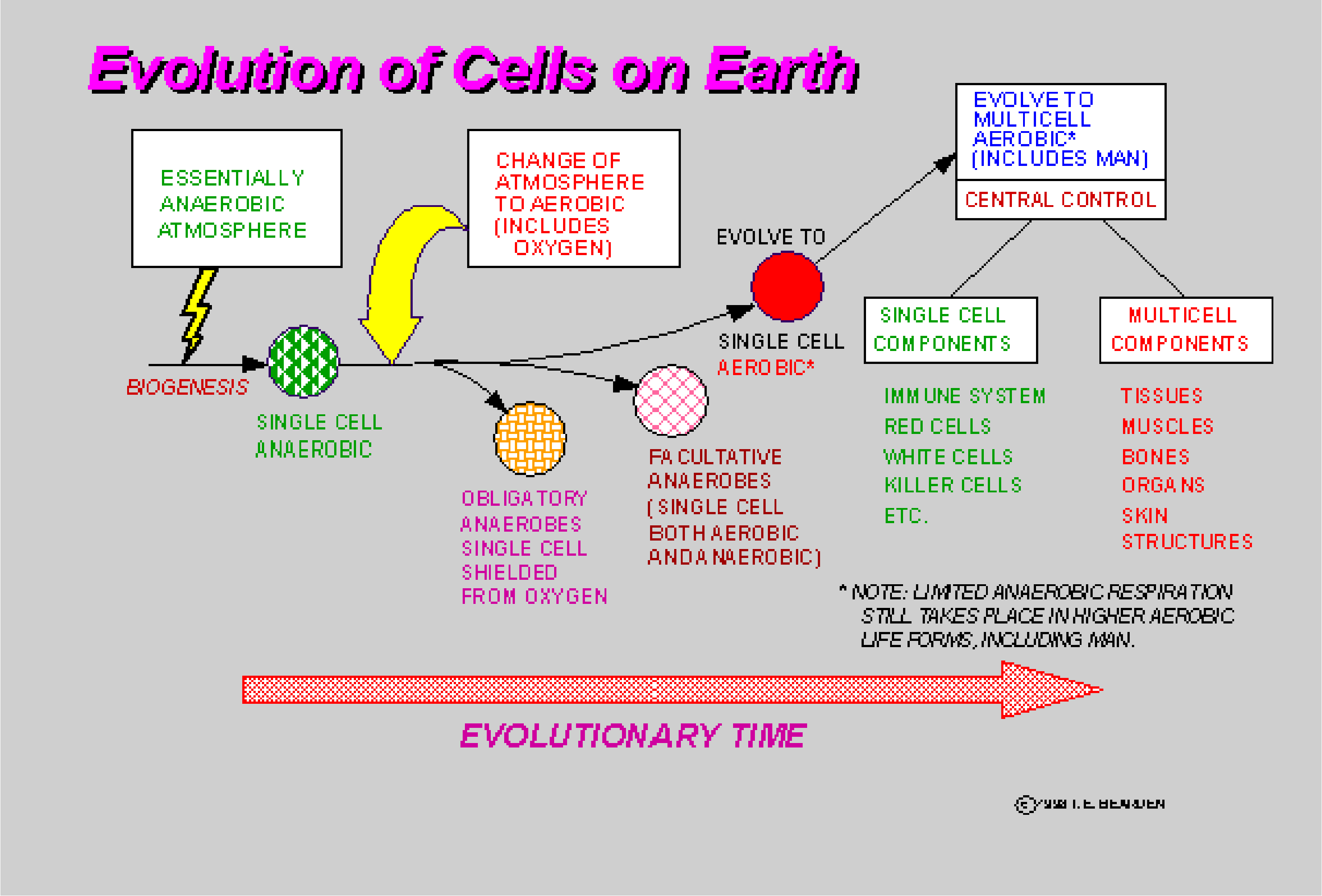

Evolution of Cells on Earth

ESSENTIALLY ANAEROBIC ATMOSPHERE

BIOGENESIS

SINGLE CELL
ANAEROBIC

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

OBLIGATORY
ANAEROBES
SINGLE CELL
SHIELDED
FROM OXYGEN

FACULTATIVE
ANAEROBES
(SINGLE CELL
BOTH AEROBIC
AND ANAEROBIC)

EVOLVE TO

SINGLE CELL
AEROBIC*

EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

*NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

©1998 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

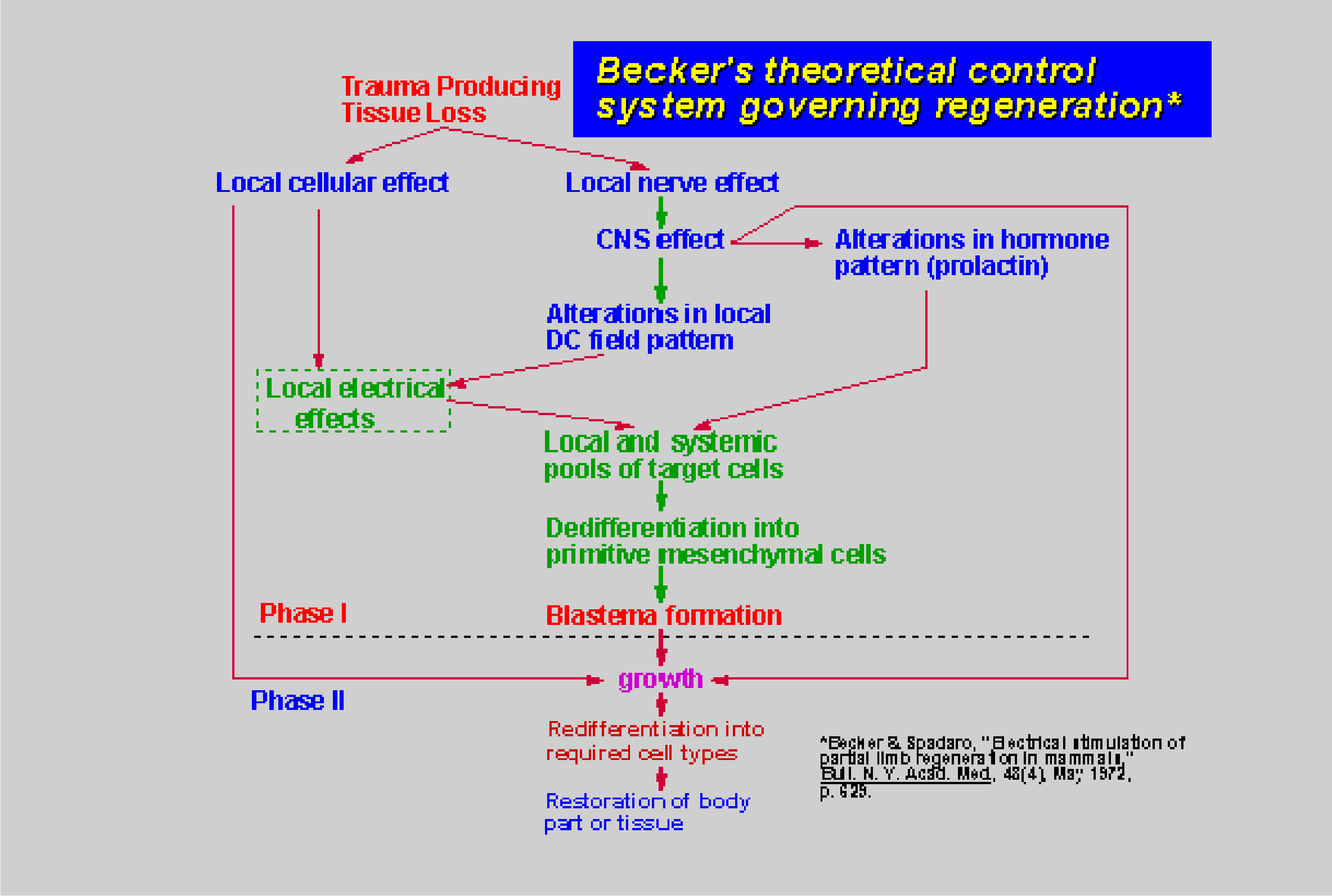

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 625.

Published by Lotus® Freelance Graphics®

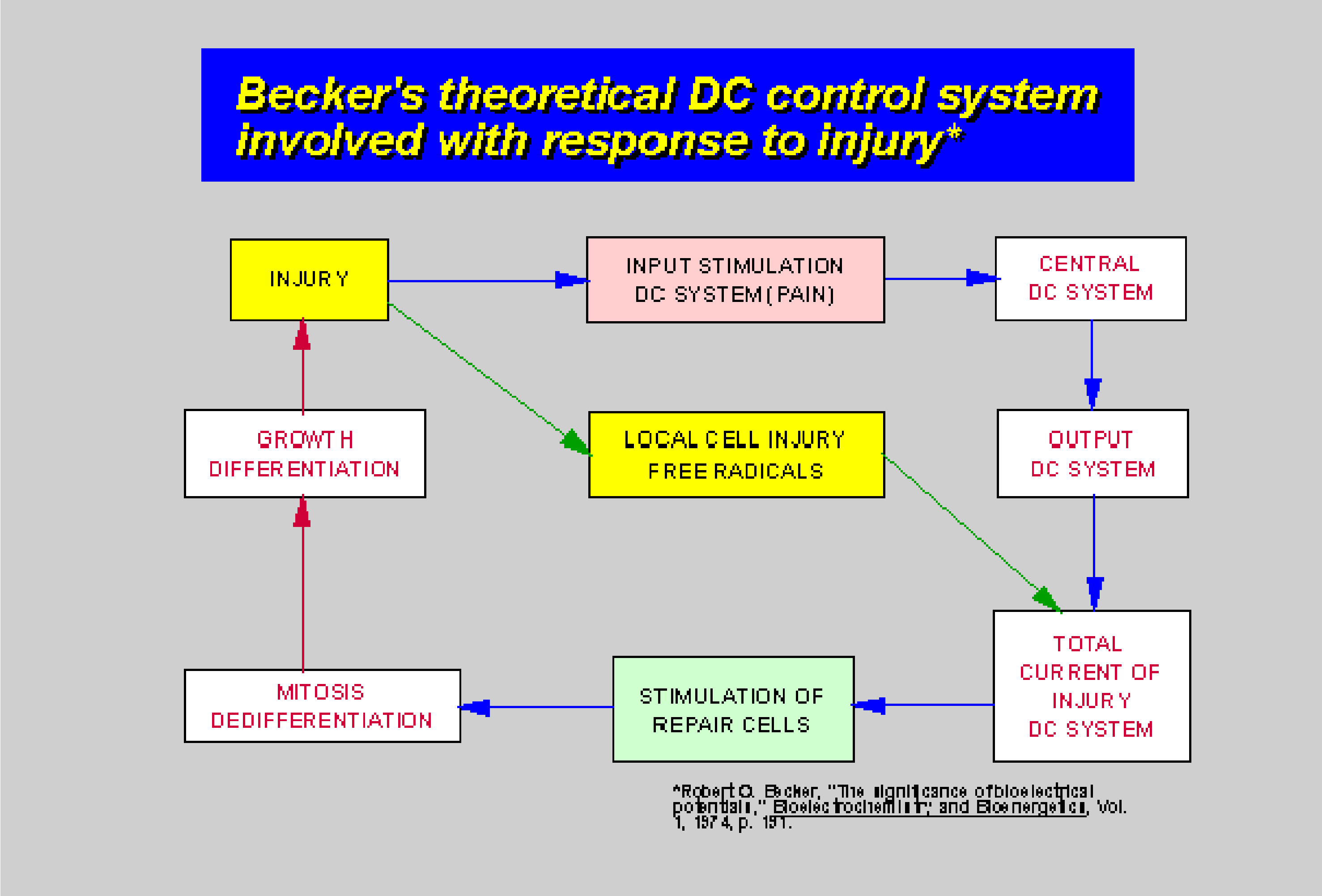

Becker's theoretical DC control system involved with response to injury*

INJURY
INPUT STIMULATION DC SYSTEM (PAIN)
CENTRAL DC SYSTEM
GROWTH DIFFERENTIATION
LOCAL CELL INJURY FREE RADICALS
OUTPUT DC SYSTEM
TOTAL CURRENT OF INJURY DC SYSTEM
MITOSIS DEDIFFERENTIATION
STIMULATION OF REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Published by Lotus® Freelance Graphics®

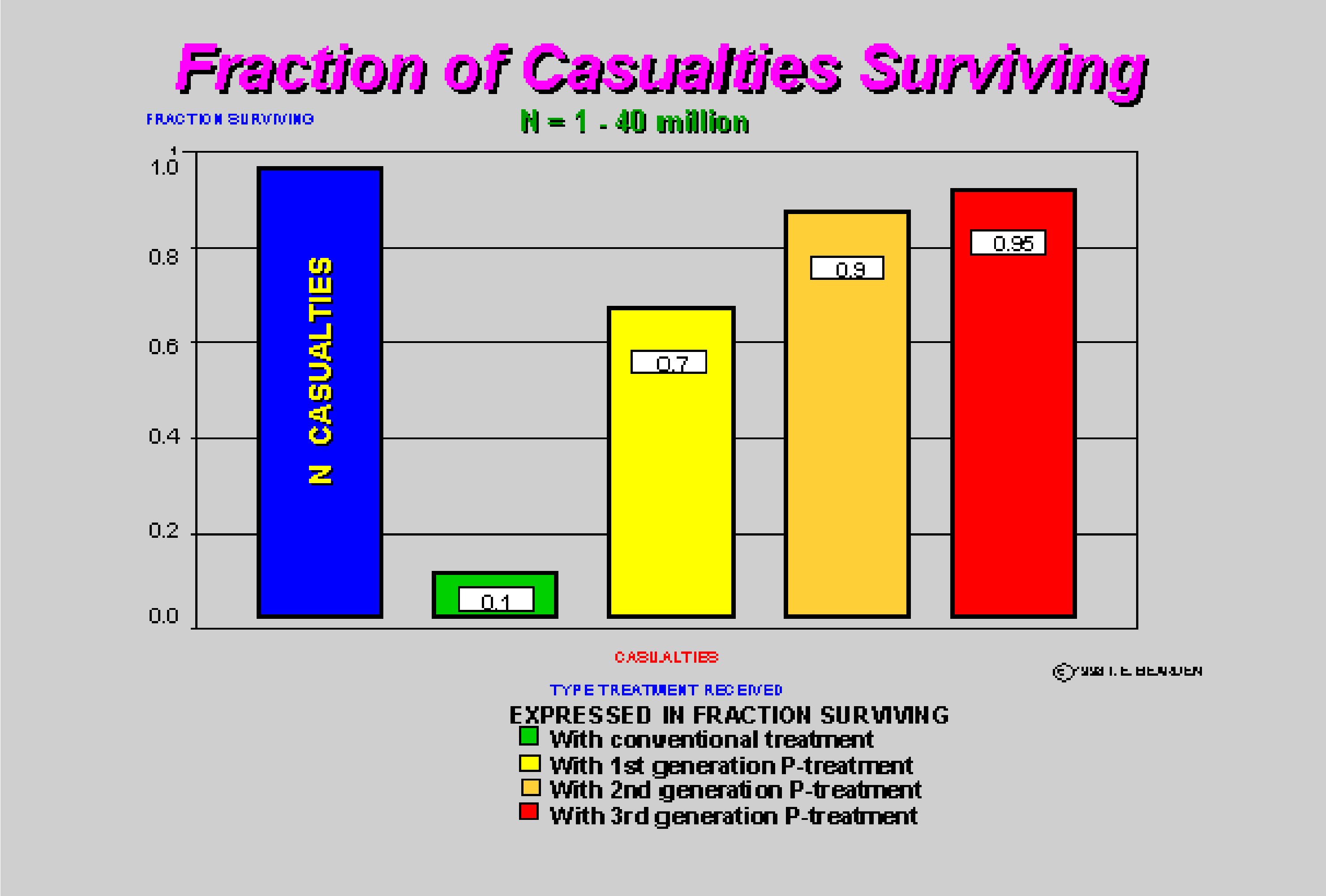

Published by Lotus® Freelance Graphics®

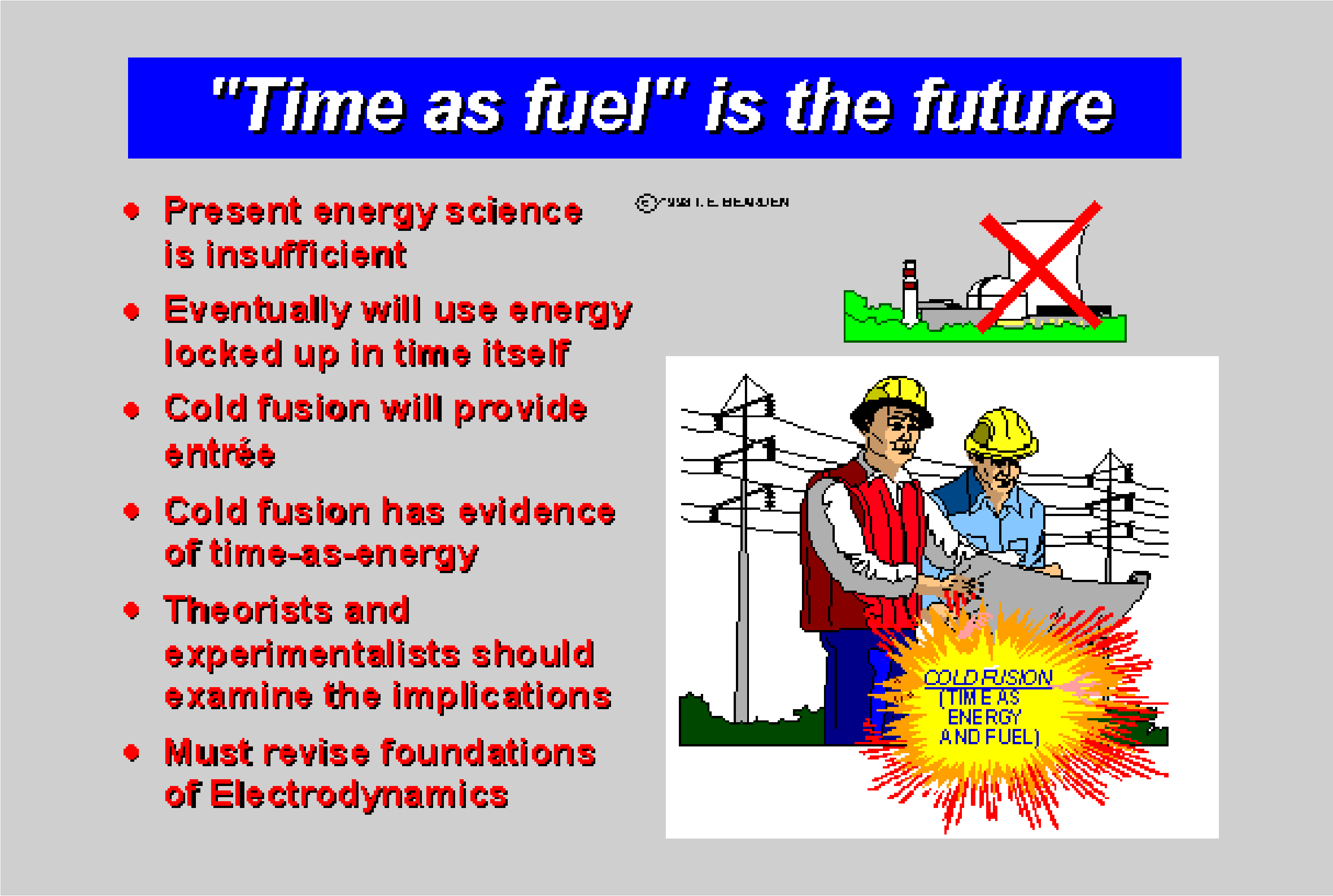

"Time as fuel" is the future

- Present energy science is insufficient
- Eventually will use energy locked up in time itself
- Cold fusion will provide entrée
- Cold fusion has evidence of time-as-energy
- Theorists and experimentalists should examine the implications
- Must revise foundations of Electrodynamics

© 1998 T. E. BEARDEN

COLD FUSION
(TIME AS
ENERGY
AND FUEL)

Published by Lotus® Freelance Graphics®

A National Report on America's Energy Crisis

Remarks by U. S. Secretary of Energy Spencer Abraham

U. S. Chamber of Commerce, National Energy Summit March 19, 2001

Opening Remarks

I would like to congratulate the U. S. Chamber of Commerce for convening this two-day National Energy Summit and I appreciate your invitation to participate. As we all know, the topic of energy is as timely as the day's headlines. In just my first eight weeks as Energy Secretary, California has faced almost daily power alerts. Rising natural gas prices have punished consumers with bills that, in some cases, are double or triple last winter's. And forecasts for this summer suggest the possibility of rolling blackouts.

The good news is that America's energy problems can be solved. The bad news is that the situation in California is not isolated, it is not temporary, and it will not fix itself.

America faces a major energy supply crisis over the next two decades. The failure to meet this challenge will threaten our nation's economic prosperity, compromise our national security, and literally alter the way we live our lives.

America has heard these dire warnings before - in fact, they seem to be a recurring theme in our nation's energy discussion, almost since "Colonel" Drake made the first oil strike near Titusville, Pennsylvania in 1859. "The amazing exhibition of oil," advised the State Geologist of Pennsylvania, "[is] a temporary and vanishing phenomenon-one which young men will live to see come to its natural end. " That was in 1885.

Around the same time, John Archbold - who succeeded John D. Rockefeller as head of Standard Oil - joked about the prospects for oil discoveries in Oklahoma: "Are you crazy?" Archbold said. "I'll drink every gallon produced west of the Mississippi. "

I don't know if anyone ever called Mr. Archbold on that pledge. But for whatever reason, in 1914 the U. S. Geological Survey predicted that the U. S. would soon exhaust its available oil supplies. They issued the same warning in 1926. And again in 1939. And in 1949.

All of these warnings have proven false. Despite all these expert predictions, the world has not run out of oil or other resources. And yet here we are, faced with the most serious energy shortage since the days of oil embargoes and gas lines.

My point is, America's current energy supply crisis is not due to some inevitable neo-Malthusian depletion of resources. The United States - and our North American and hemispheric neighbors-are blessed with a rich abundance of natural resources. It's political leadership that has been scarce.

For the past eight years, Washington sat on the sidelines as our nation's energy needs mounted. During the 1990s, the Clinton Administration employed a policy of taxing demand, limiting supply, and ignoring the rapidly expanding needs of the future.

Their energy strategy boiled down to: you can't find it . . . you can't transport it . . . and even if you get it, we don't want you to use it.

Through neglect or complacency or ideology, this approach has led us to the crisis we face today.

The Bush Administration is fully prepared to respond to the broad set of challenges we inherited. But we must be candid with the American people about the magnitude of the problem - which is what I'd like to talk about today.

America's Energy Supply Challenges

Three overriding facts starkly define the challenge of America's energy needs over the next two decades:

First, demand for energy is rising across the board, but particularly for natural gas and electricity;

Second, supplies are being limited by a regulatory structure that, in many respects, has failed to keep pace with advances in technology and an uncertain political environment that often discourages investment in desperately needed facilities;

And third, our energy infrastructure - that network of the generators, transmission lines, refineries and pipelines that convert raw resources into usable fuel - is woefully antiquated and inadequate to meet our future needs.

Unless these challenges are addressed, America's energy supply will be continually at risk . . . our citizens will encounter blackouts and other lifestyle-altering disruptions . . . and our economy will be hobbled by rising energy prices. Let me briefly outline some of the major issues on the horizon:

Oil: Rising Consumption, Accelerating Dependence

In the next 20 years, according to estimates by the Energy Information Administration, America's demand for oil is projected to increase by 33 percent. Yet as consumption surges, U. S. production continues to drop precipitously. We now produce 39 percent less oil than we did in 1970, losing nearly 4 million barrels a day in the process. And unless energy policy is changed, production will slip further - to just 5.1 million barrels per day by 2020 - down from a high of 9. 4 million a day 30 years ago. This widening gap between demand and domestic supply will make us increasingly dependent upon foreign imports. Back in 1973-at the height of the oil crisis - America imported just 36 percent of its oil from abroad.

Today, we import 54 percent. And, if we allow this trend to continue, we will soon be forced to look abroad for some 64 percent of our oil. This will put more power in the hands of foreign suppliers - power they are not reluctant to use, as we just saw when the OPEC cartel decided to reduce oil output by one million barrels a day.

While this administration does not agree with OPEC's decision, that decision demonstrates the importance of increasing America's production of oil. Securing an affordable, reliable and adequate supply of crude is a critical challenge. But it is only half the oil story.

Since 1980, the number of American refineries has been cut in half. There hasn't been a new refinery built in the United States in over 25 years. New regulatory interpretations limit the ability of existing refineries to expand capacity. Add to that regulations that require the production of more than 15 different types of gasoline-and you have a refining industry strained to capacity, leaving us dangerously vulnerable to regional supply disruptions and price spikes.

Refineries are so constrained that when President Clinton made the politically symbolic gesture of releasing 30 million barrels of oil from the Strategic Petroleum Reserve last fall, that oil had to be shipped overseas to be refined.

Natural Gas: Rapidly Rising Demand; Constraints on Supply

Many of the same issues confront the future of natural gas. America's demand for natural gas is projected to rise even more rapidly than oil. If Department of Energy projections are correct, by 2020 Americans will consume 62 percent more natural gas than we do today. More than 9 out of 10 of the announced new electric generating plants will be fired by natural gas.

This approaching wave of new demand begs the question: Do we have an energy policy and a regulatory structure capable of meeting our natural gas needs?

Consider just a few constraints in this market.

Right now, an estimated 40 percent of potential gas resources in the United States are on federal lands that are either closed to exploration or covered by severe restrictions. The last lease sale in the some areas of the Gulf of Mexico was more than a decade ago. New discoveries of natural gas in the United States have fallen for three straight years, creating increasing pressure for more imports.

The notion that we can rely so heavily on natural gas . . . maintain severe restrictions on exploration . . . and still enjoy low prices is a dangerous assumption.

Even if we find the supplies, moving that gas to market will require an additional 38,000 miles of transmission pipeline and 255,000 miles of distribution lines-at an estimated cost of $120-$150 billion.

Today's pipeline system can hardly handle the supplies we know exist. Alaska's Prudhoe Bay, for example, produces about 8 billion cubic feet of natural gas a day - approximately 13 percent of America's daily consumption.

But that gas never reaches the market. Instead, it is just pumped back into the ground, waiting until a pipeline is built to connect the Alaska fields to the U.S./Canada distribution system.

Electricity: Powering the Information Economy

As everyone knows, we also face a real challenge in generating enough electricity to light out homes and run our businesses. Over the next 20 years, the Department of Energy estimates that electricity demand in the United States will increase by 45 percent. That rising growth rate will require the construction of over 1,300 new power plants-about 65 every year.

Yet, the last time we added that much power was 1985.

Furthermore, there is reason to believe that this could turn out to be a conservative estimate. During the 1990s, electricity consumption far outstripped projections, driven by the energy-hungry information economy. Some experts calculate that the demands of the Internet already consume some 8-13 percent of electricity. If demand grows at just the same pace as during the last decade, we'll need nearly 1,900 new plants by 2020-or more than 90 every year-just to keep pace.

Hundreds of new generating plants will place even greater pressure on our already strained and aging power grid. America's network of transmission lines, substations and transformers was built when utilities were tightly regulated monopolies providing service to assigned regions. Interconnections between suppliers were strictly an emergency backup measure to guard against rare service interruptions. The system was simply not designed for long-haul swapping of power in a highly competitive market.

Consumers are already feeling the impact of a transmission systems stressed by rising demand. Transmission bottlenecks contributed to the blackouts that have swept through California and to price spikes in New York City last summer that cost consumers an estimated $100 million.

Coal, Nuclear, and Hydro-Electric Power

Coal has historically been American's number one source for affordable electricity; it currently powers half of America's electricity generators. And at today's recovery rates, our nation has enough coal to keep those plants running for the next 250 years.

Coal generators have already been called upon to make broad reductions in emissions. The Bush Administration supports those efforts-and we will back it up with greater incentives for investment in clean coal technology.

But the administration will not regulate coal out of existence . . . and we will not support measures that will threaten electricity supplies and significantly raise electricity prices. President Bush made the right decision last week not to impose new federal mandates on the emissions of carbon dioxide. If America is to have reliable electricity over the next 20 years, coal must continue to play a major role.

But, coal is not the only energy source facing an uncertain future. There hasn't been a new nuclear power plant permit granted since 1979. Many of the 103 existing nuclear plants are not even expected to file for a renewal of their licenses as they expire over the next 15 years.

Even hydroelectric power generation is expected to fall sharply.

Re-licensing a hydro facility can take a decade or more and cost millions. And now, even though consumers are faced with potential blackouts and chronic electricity shortages in the West, activists and some political leaders want to breach one or more of the four federal dams on the Snake River to help young salmon, on their trek to the sea.

The Dangers of Complacency

What are the dangers of complacency in light of these challenges? How does it all add up for our economy and our citizens?

This nation's last three recessions have all been tied to rising energy prices - and there is strong evidence that the latest crisis is already having a negative effect.

The National Association of Manufacturers estimates that soaring fuel prices between 1999 and 2000 cost the U.S. economy more than $115 billion-shaving a full percentage point off our Gross Domestic Product. A January survey of its 5,500 members reveals that nearly one quarter were forced to curtail operations.

During a two-week period this past January, Californians lost an estimated $2.3 billion in wages, sales and productivity. Layoffs are already hitting workers in the West as companies shift production to states with more reliable energy sources. Then there are the jobs that will never be created.

Intel's CEO Craig Barrett announced that the world's leading chipmaker won't be expanding in California: "As long as California is a Third World country," Barrett said, "we won't build $2 billion manufacturing plants here."

The Food and Agricultural Policy Institute reports that farmers are likely to see their income drop 20 percent over the next two years due to higher energy costs.

Rising energy costs are hitting every family's checkbook, primarily affecting those who can afford it least. Gas bills for many homeowners in the Washington, DC area more than tripled this year. Some residents are reporting that their heating bills are higher than their food bills this winter.

The power crisis isn't just pinching our wallets, it's changing the way we live our lives. In California, power outages have shut down traffic lights, darkened schools and closed businesses. The governor has ordered local police to patrol the streets-not for criminals, but to make sure businesses keep their lights dimmed.

But California is not the only state facing a mismatch between supply and demand. With electricity shortages predicted for New York City and Long Island this summer . . . low capacity margins threatening electricity reliability in the Midwest, Southeast and Northern Plains states . . . and strained refinery capacity in the Midwest, Americans across the nation are feeling the energy squeeze.

The Need For A National Energy Policy

Rising demand . . . tightening supplies . . . an aging power infrastructure . . . a decade of neglect from Washington: These are the trends that define America's emerging energy needs.

President Bush has committed this administration to meeting these challenges - a job that begins with the urgent task of developing and implementing a long-term national energy policy.

To accomplish this, President Bush created an Energy Task Force headed by Vice President Cheney. He has asked us to define a clear strategy - a strategy that will allow environmentally responsible exploration and recovery of our domestic resources . . . enhance our commitment to conservation and energy efficiency . . . and encourage investment in new technology to further the development of renewable energy sources.

I wish I could say that the energy crisis now sweeping the nation has shocked the political system into action. But like other political discussions in recent years, the debate over energy seems as deeply polarized as ever.

On one end of the spectrum, some activists propose what amounts to a "zero tolerance" policy toward exploration and cling to the quixotic idea that new, undiscovered sources will somehow allow us to meet our energy needs. On the other end, some advocates place an almost limitless faith in special tax breaks for this favored activity or subsidies for that preferred industry.

The two extremes in our energy debate are founded on several enduring myths - myths that today conspire to block any true advance toward a rational and stable energy policy in the United States. Here are a few of the more prominent ones.

Myth Number 1: It is impossible to balance energy exploration and environmental protection.

Advances in technology have brought us a long way from the days when wildcatters punched holes in the ground based on the hunch they might hit a gusher. But from a regulatory standpoint, our view of oil and gas exploration hasn't changed much since we saw Jed Clampett strike "black gold" and split for Beverly Hills.

Today, satellites and computers are the tools of choice in the exploration business. Geologists can bounce acoustic and electrical vibrations off the earth's inner depths, gather the resulting mass of data into powerful computers and then create three dimensional and even four dimensional maps of resource fields miles below the surface.

Armed with these pinpoint accurate images, companies employ advanced equipment to drill vertically, horizontally and around corners-allowing us to access supplies from previously unimaginable depths, up to six miles away.

The marriage of oil and gas exploration with cutting-edge technology means fewer rigs, fewer roads and fewer pipelines. Drilling operations that required 65 acres in the 1970s need only 10 acres today. Technological improvements in just the past 15 years have generated success rate increases of 50 percent. America's national energy policy must reflect these staggering advances that have revolutionized the way we develop our resources.

Myth Number 2: All our current problems are due to an energy industry that is engaged in a massive conspiracy to gouge consumers by limiting supply to drive up prices.

This myth has been punctuated by calls for investigations into everything from last summer's Midwest gasoline price spikes, to recent allegations that power generators in the West have been withholding electricity. We have a fair and objective process for judging these claims-and action will be taken when it is merited. Over the past two weeks, for example, the Federal Energy Regulatory Commission ordered power companies to rebate some $124 million to California utilities. Meanwhile, the Federal Trade Commission recently cleared gasoline suppliers of all charges relating to last summer's price increases.

But charges of price gouging largely miss the point. There is no magic source of supply; no hidden pool of energy that can be turned on and off like a faucet. California-and other power-strapped states-will never solve the power crises they confront until they resolve the conflict between demand and supply.

Earlier this year, one company proposed building a $400 million power plant in California that would have provided enough additional electricity to light 600,000 homes in energy-starved Silicon Valley. The company pledged to plant 800 new trees to beautify the area. They proposed cloaking the power station in a brick facade to make it essentially indistinguishable from a high-rent office complex. They even promised to help maintain the local habitat for the endangered bay checkerspoon butterfly.

Their environmentally-sensitive plans won the support of the Sierra Club, the American Lung Association and the NAACP. But city officials voted unanimously -- 11-0 -- to reject the plan. In an editorial, the local paper called this move "Dumb and Dumber."

Meanwhile, further south, plans to build a 550-megawatt gas-fired generator in a Los Angeles suburb were scrapped after residents voted 2:1 against the project. The local mayor added a much-needed dose of reason and maturity to the debate-by launching a hunger strike in opposition to the plant.

In California, workers are being laid off, companies are leaving the state, farmers and small businesses are losing millions, consumers are threatened with rolling blackouts, but local officials reject power plants with little regard for the consequences. Is it really any mystery why there hasn't been a single new power plant built in California in the last decade?

Myth Number 3: The Bush energy plan is focused almost exclusively on opening the Arctic National Wildlife Refuge (ANWR) to exploration-a move that would buy us only about 6 months worth of American consumption while destroying a pristine natural wilderness, not to mention disrupting the breeding ground of the Porcupine Caribou.

Let's separate fact from fiction when it comes to ANWR.

First, according to estimates by the U. S. Geological Survey, ANWR holds between 5. 7 to 16 billion barrels of recoverable reserves-with a mean estimate of 10. 4 billion barrels. And that assumes the use of drilling technology now nearly a decade old. This represents more than 300 times the amount of the oil President Clinton released from the Strategic Petroleum Reserve last fall. And based on December 2000 figures, it would free us from about 54 years of oil imports from Saddam Hussein and Iraq.

Second, exploration would impact only about 2,000 acres out of more than 19 million. To put that in perspective, the massive Arctic National Wildlife Refuge is about the same size as the entire State of South Carolina; the two thousand acres that would be affected is less than half the size of Dulles airport.

And as for the caribou, the herd in the Prudhoe Bay area grew more than 9-fold over the past 20 years to an estimated 28,000 in 2000-seemingly irrefutable evidence that caribou mating and oil exploration can peacefully coexist.

The decision to open a small portion of ANWR should be made on the merits. But it should not be made on the mistaken assumption that opening ANWR will allow us to produce our way to full energy independence.

America first became a net importer of energy in the 1950s - and our economy will continue to depend, in part, on imported oil. However, closing off virtually every available new source of domestic supply, enhances the leverage and power of an oil cartel that cannot be relied on to put America's interests first. While the resources of ANWR won't make us energy independent, they will help increase America's energy security by ensuring a more diverse supply of oil.

Myth Number 4: Government subsidies and tax breaks are the best way to encourage new exploration and production of energy.

This administration will continue to support funding for energy research and development initiatives. But capital is best allocated to its highest uses through the workings of the free market, not manipulations of the tax code. Government regulatory policy should not be aimed at picking winners and losers in any market, including energy. Neither should tax policy.

Myth Number 5: We can forego traditional sources and instead meet rising energy demand by harnessing wind, geothermal, solar and other forms of renewable power.

Excluding hydro-power, renewable sources currently generate about 2 percent of America's electricity. Billions have been invested in developing renewable energy-and will continue to be invested under the Bush Administration. But renewables have yet to overcome the economic advantages of conventional energy sources.

Even with promising advances in research and development, renewables will only provide, according to Energy Information Administration estimates, about 6 percent of our total electricity consumption by 2020. Even if renewables exceed our most optimistic expectations, they would still supply only a fraction of our needs over the next 20 years.

Myth Number 6: Price controls are the answer to today's energy crisis.

Memories are short, aren't they. So let me remind everyone. America imposed price controls on oil and gas in the 1970s. They were an utter failure. They led to shortages and rationing and the idea that America was gripped by malaise.

Let me be clear: The Bush Administration does not support price controls. Price controls on electricity will lead to more blackouts. Price controls on gasoline will lead to gas lines. Price controls will deepen America's energy crisis, because they won't reduce demand, but they will cripple incentives for desperately needed new investments in energy supply.

Charting a New Policy Course

The challenges are formidable . . . the warning signs are obvious . . . but I am optimistic because I know this administration's commitment is equal to the task.

Our national energy policy will be comprehensive. It will reach across every department that touches the energy marketplace-from the Interior Department and the EPA to the Transportation Department and the DOE.

Our national energy policy will be hemispheric. It will be based on the understanding that our policy cannot stand in isolation from our neighbors throughout the Americas.

Our national energy policy will stress the need to diversify America's energy supply. It will be founded on the understanding that diversity of supply means security of supply . . . and that a broad mix of supply options-from coal to windmills, nuclear to natural gas-will help protect consumers against price spikes and supply disruptions.

And our national energy policy will be balanced. It will leapfrog the myths that stifle change-rejecting the notion that there is no middle ground between environmental protection, regardless of the cost and energy exploration, regardless of the impact.

Soon we will deliver our recommendations to President Bush. Later, we will introduce legislation aimed at winning bipartisan support for a national energy policy that matches the magnitude of the challenge. I am hopeful that men and women of good will-from both ends of the political spectrum . . . from environmental organizations to industry groups-will then come together and transcend the stale debate that has characterized energy policy in recent years.

About 150 years ago, America faced a vastly different energy crisis.

Supplies of whale oil were becoming more and more scarce. Few could afford to pay for the luxury of this or other costly methods of illumination. Sure, crude oil was available. In those days it was soaked up with rags, wrung out into small vials, and then sold as a treatment for toothaches . . . until an entrepreneur lined up an investor and a chemist and launched an energy revolution that would light the world.

In America, resources become scarce only when our imagination languishes. By engaging that imagination, I am confident we can meet the challenges of today. If complacency yields to action. If we resolve to strike a rational balance between our energy needs and our environmental concerns. And if a national energy policy becomes an urgent priority.

Thank you.

Report by Energy Secretary Spencer Abraham

Contents

1. Report by Energy Secretary Spencer Abraham
2. National energy policy for 8 years
3. Energy supply challenges
4. Oil: Rising consumption and accelerating dependence
5. Oil: Rising consumption and accelerating dependence
6. Natural Gas: Rapidly rising demand, supply constraints
7. Natural Gas: Rapidly rising demand, supply constraints
8. Electricity: Needs, problems
9. The aging power grid
10. Coal fired electrical power
11. Nuclear power plants
12. Hydroelectric power generation
13. Dangers of complacency
14. Dangers of complacency (2)
15. Need for national energy policy
16. Extremes of the energy positions
17. Six prominent myths, none of which is true
18. Can we balance exploration and environmental protection?
19. Is there a price gouge conspiracy?
20. Is there a price gouge conspiracy? (2)
21. Is there a price gouge conspiracy? (3)
22. Arctic National Wildlife Refuge Exploration
23. Arctic National Wildlife Refuge Exploration (2)
24. Our increasing oil imports
25. Government subsidies and tax breaks
26. Renewable energy sources
27. Price controls
28. Charting a new national policy

---

29. **Commentary by Tom Bearden, Ph.D.**
30. Commentary
31. Caution: It is stressed that
32. An example of the time problem

33. The pending disaster
34. The bottom line:
35. Here are some events etc.
36. Lowered energy requirements for appliances
37. Easing air conditioner and heat pump standards
38. Navarro's Estimate*
39. Factors driving comparisons
40. Prices likely per kilowatt-hour
41. Covering debt already incurred
42. Deteriorating situation, rising prices
43. Other Factors
44. Paying the piper: California's bond rating reduced
45. Associated Press: Gasoline supplies are critical
46. A peculiar problem
47. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (1)
48. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (2)
49. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (3)
50. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (4)
51. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (5)
52. Negentropic Engineering
53. Rigorous "out-of-the-box" work
54. Rigorous "out-of-the-box" work (2): (Bedini "negative resistor" in a battery)*
55. Initiating a Battery to COP>1.0
56. Secretary alone cannot solve it
57. What is desperately needed
58. Why it can be done in time
59. Cautions!
60. The scientists have failed us
61. 136-yr-old EM Model1
62. All EM from Time Domain1
63. Astonishing but true
64. With plentiful, cheap, clean electrical energy:
65. Recommendations to DOE
66. In conclusion:

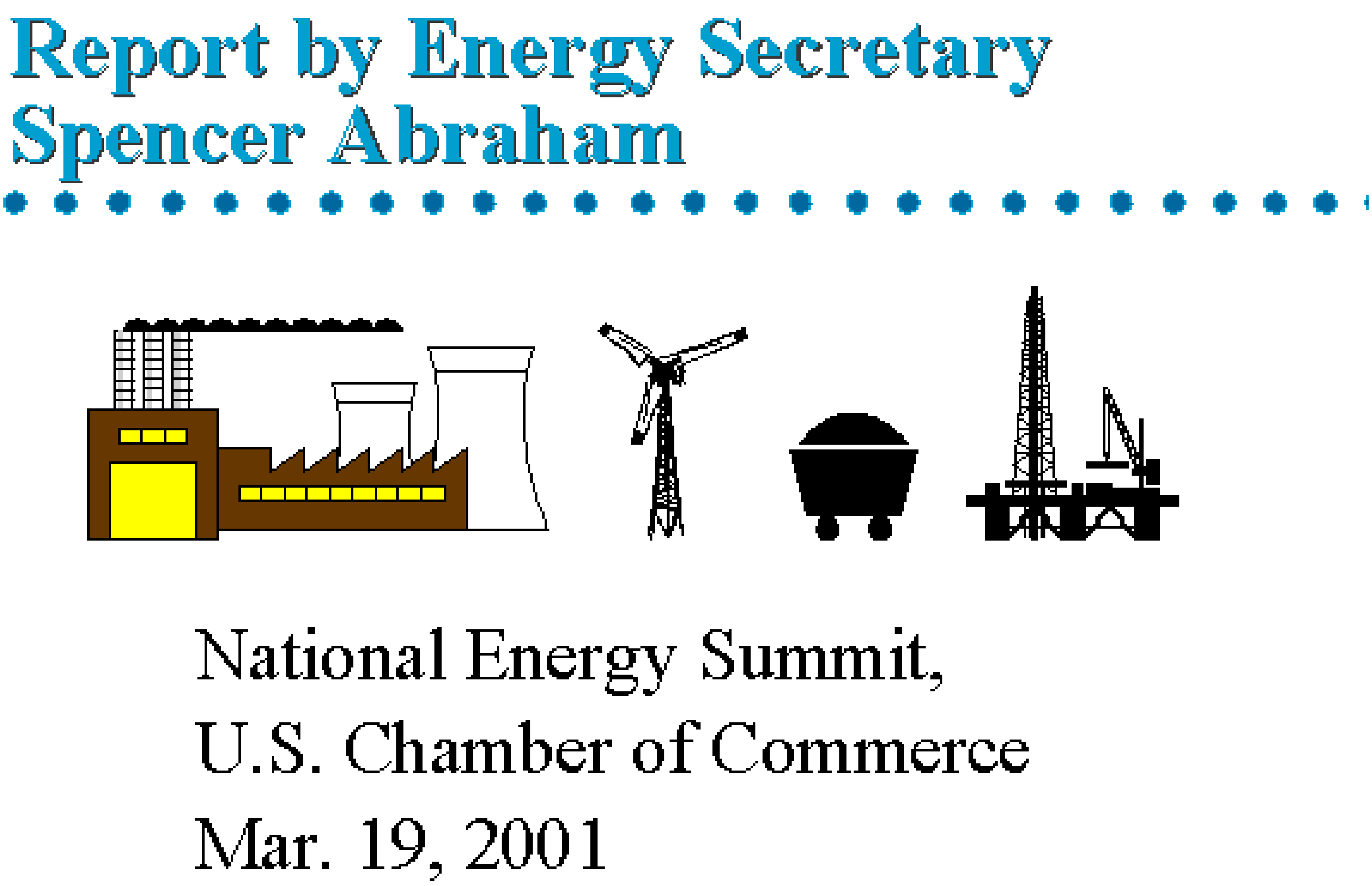

Report by Energy Secretary Spencer Abraham

National Energy Summit,
U.S. Chamber of Commerce
Mar. 19, 2001

Published by Lotus® Freelance Graphics®

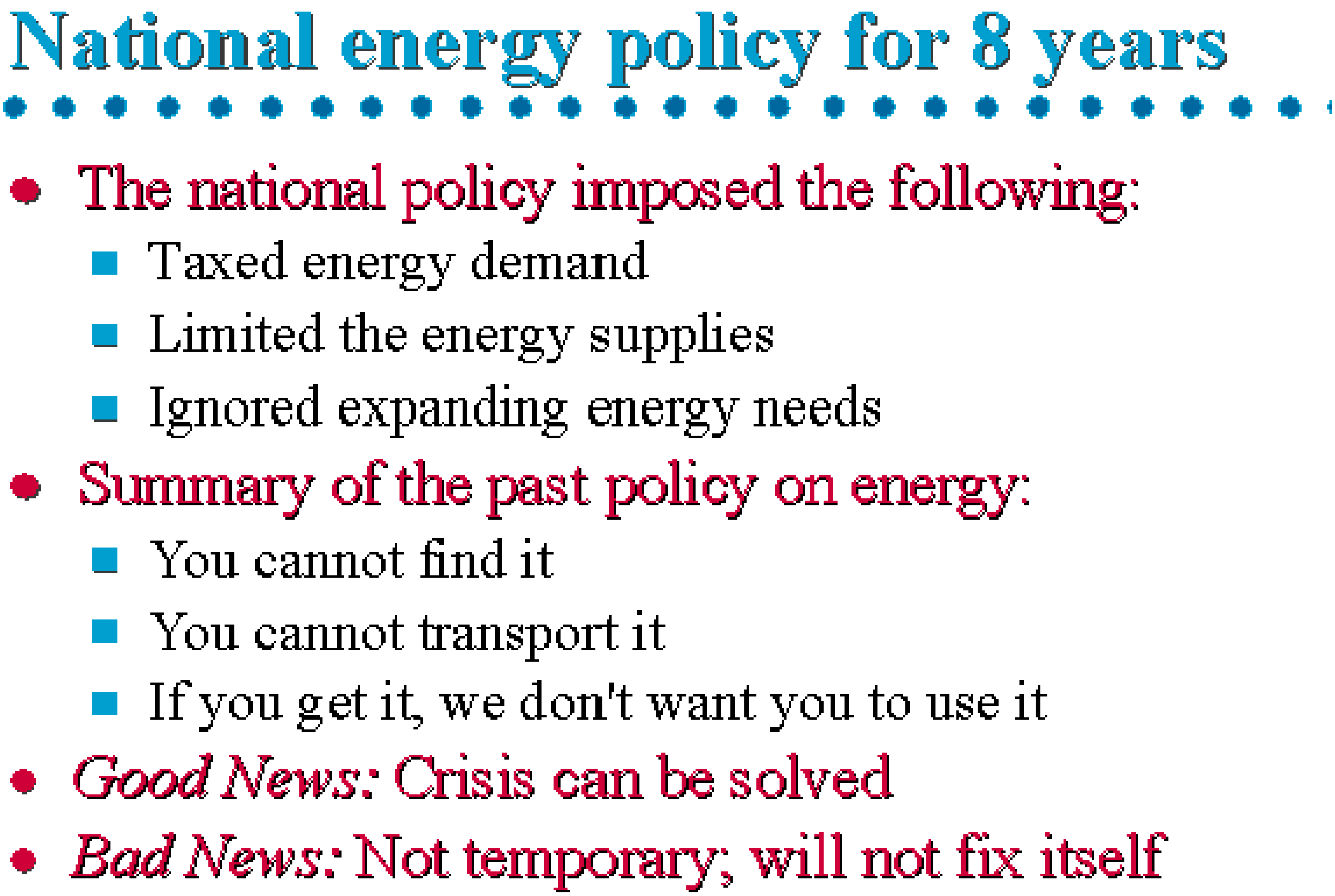

National energy policy for 8 years

- The national policy imposed the following:
  - Taxed energy demand
  - Limited the energy supplies
  - Ignored expanding energy needs
- Summary of the past policy on energy:
  - You cannot find it
  - You cannot transport it
  - If you get it, we don't want you to use it
- *Good News*: Crisis can be solved
- *Bad News*: Not temporary; will not fix itself

Published by Lotus® Freelance Graphics®

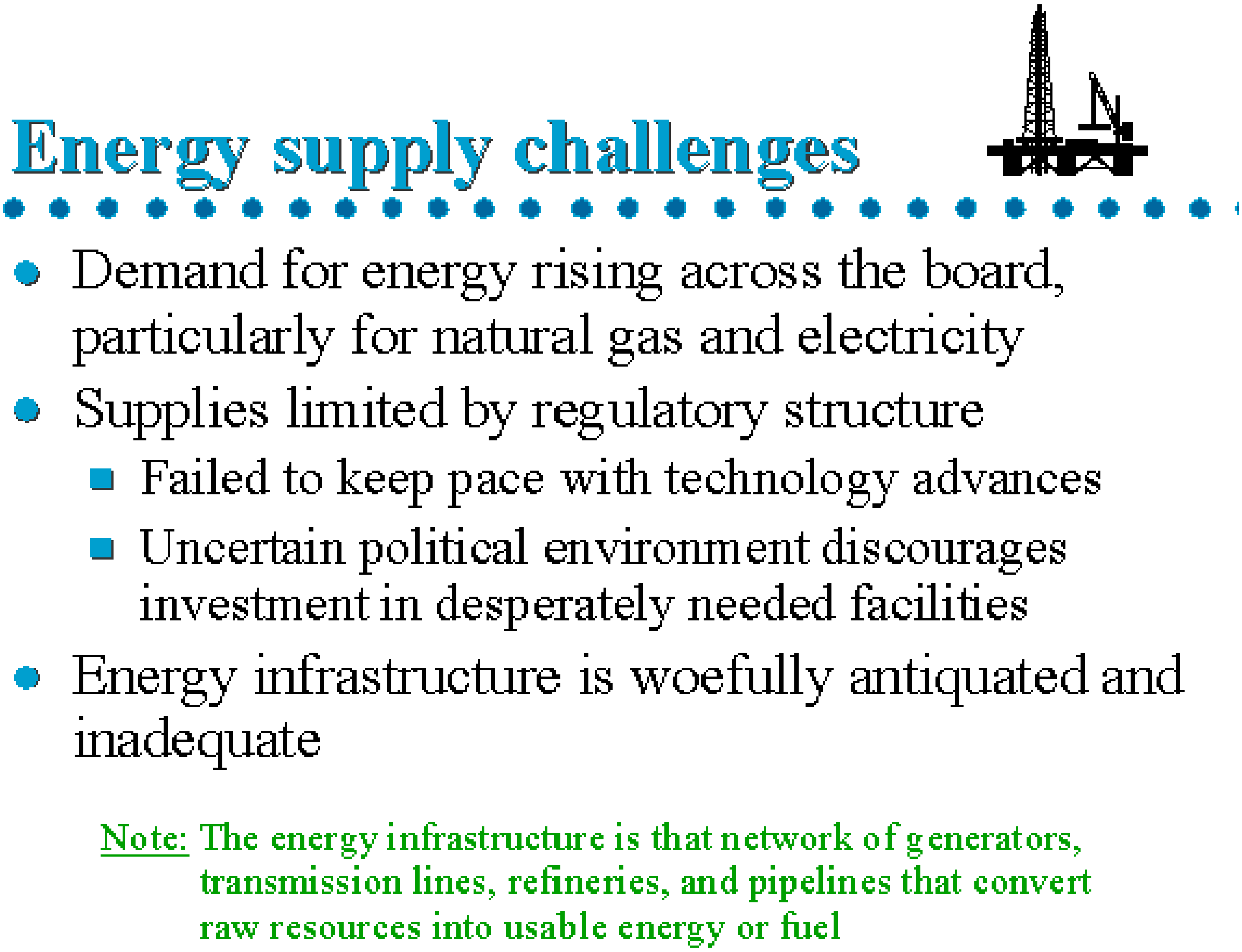

Energy supply challenges

- Demand for energy rising across the board, particularly for natural gas and electricity
- Supplies limited by regulatory structure
  - Failed to keep pace with technology advances
  - Uncertain political environment discourages investment in desperately needed facilities
- Energy infrastructure is woefully antiquated and inadequate

Note: The energy infrastructure is that network of generators, transmission lines, refineries, and pipelines that convert raw resources into usable energy or fuel

Published by Lotus® Freelance Graphics®

Oil: Rising consumption and accelerating dependence

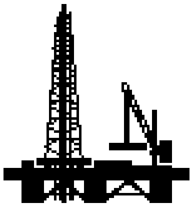

- In U.S., will increase 33 percent in 20 years
- U.S. production continues to drop
  - Produce 39% less now than in 1970
  - At present reduction rate, production will decrease to 5.1 million barrels per day by 2020, down from 9.4 million barrels per day in 1970.
- Imports have steadily risen
  - From 36% in 1973 to 54% today
  - Will soon rise to 64% unless checked
  - Empowers foreign suppliers such as OPEC
- Assuring affordable, reliable, adequate supply of crude oil is a critical national challenge

---

Published by Lotus® Freelance Graphics®

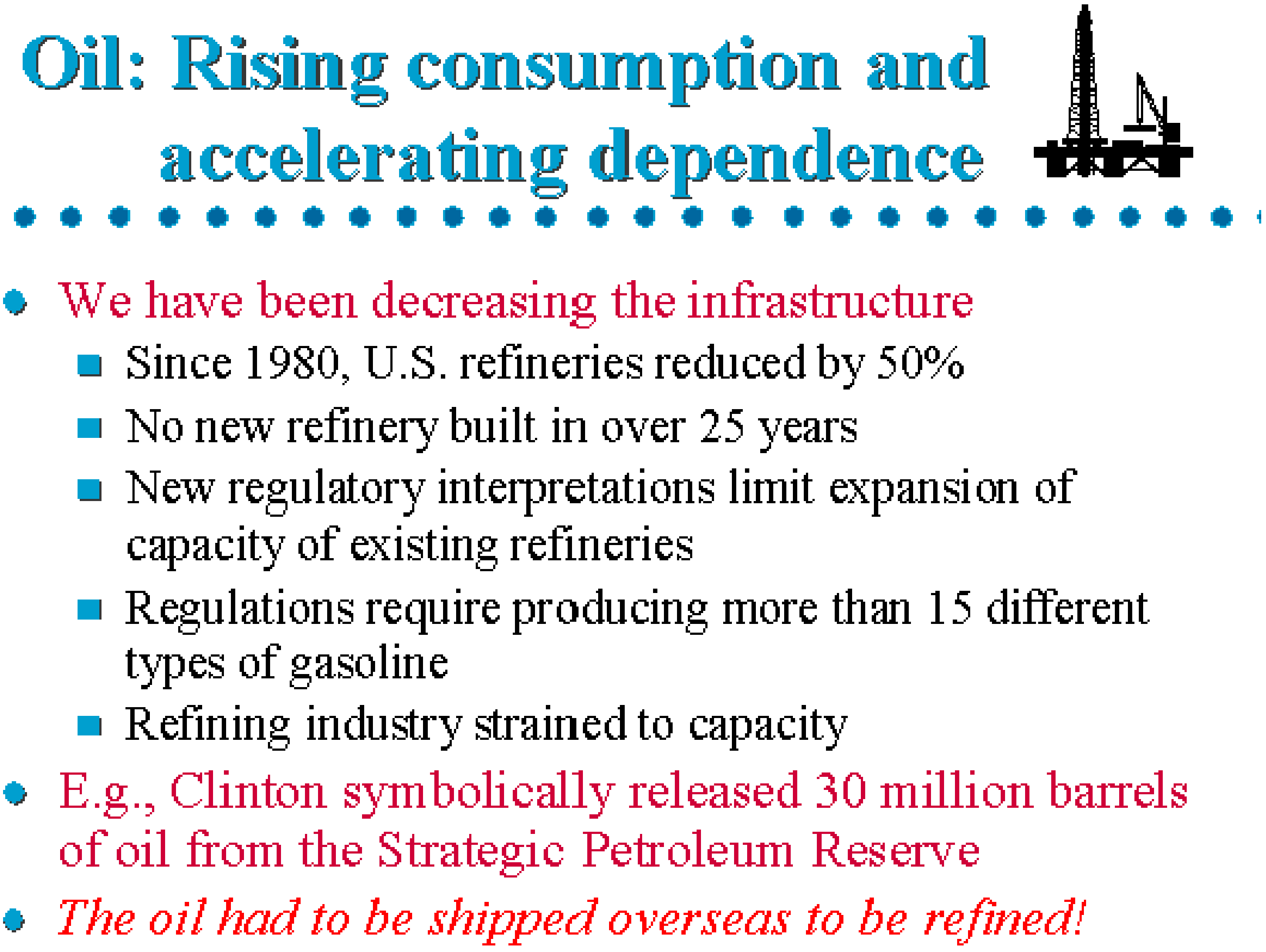

Oil: Rising consumption and accelerating dependence

- We have been decreasing the infrastructure
  - Since 1980, U.S. refineries reduced by 50%
  - No new refinery built in over 25 years
  - New regulatory interpretations limit expansion of capacity of existing refineries
  - Regulations require producing more than 15 different types of gasoline
  - Refining industry strained to capacity
- E.g., Clinton symbolically released 30 million barrels of oil from the Strategic Petroleum Reserve
- *The oil had to be shipped overseas to be refined!*

Published by Lotus® Freelance Graphics®

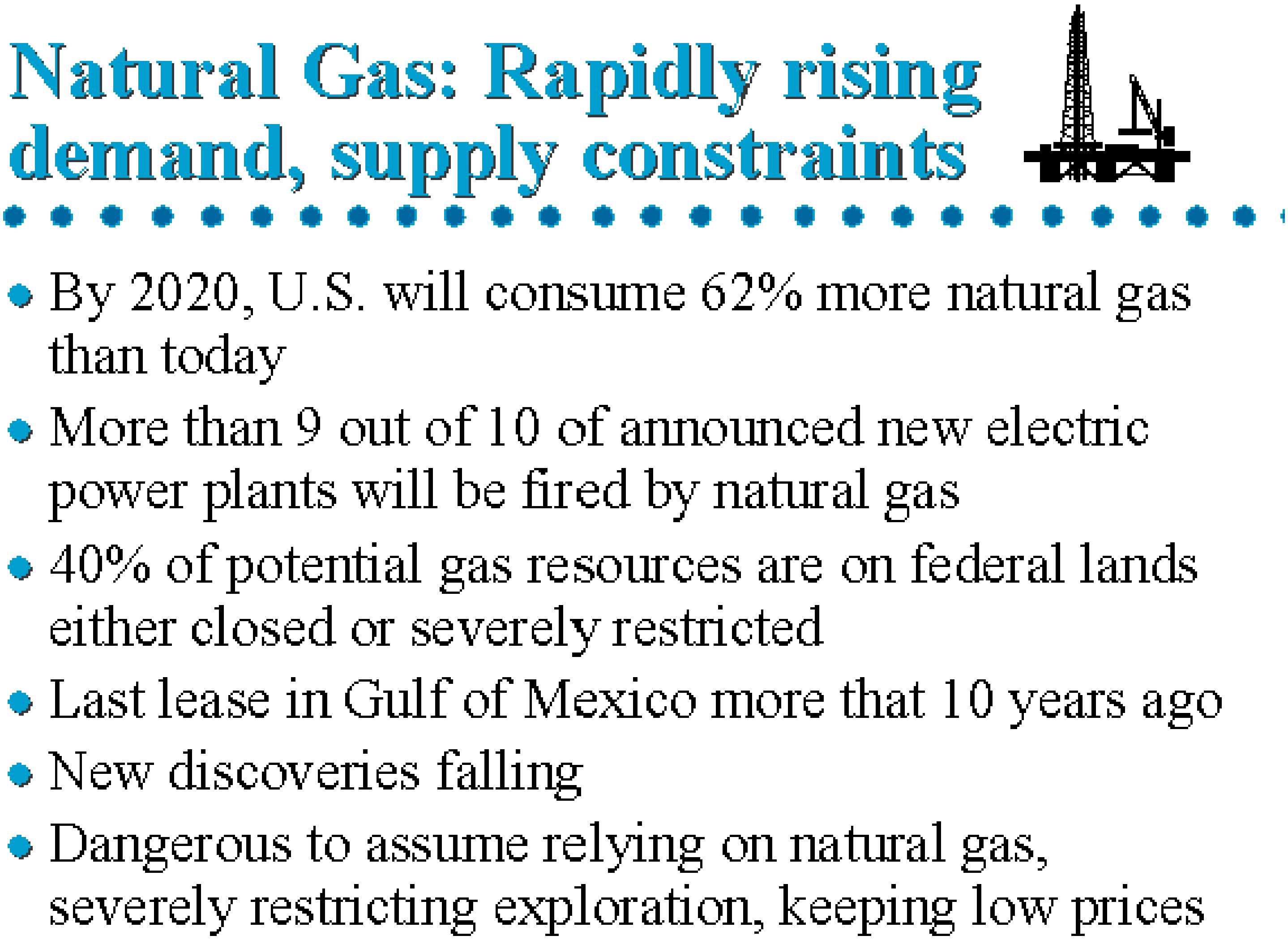

Natural Gas: Rapidly rising demand, supply constraints

- By 2020, U.S. will consume 62% more natural gas than today
- More than 9 out of 10 of announced new electric power plants will be fired by natural gas
- 40% of potential gas resources are on federal lands either closed or severely restricted
- Last lease in Gulf of Mexico more that 10 years ago
- New discoveries falling
- Dangerous to assume relying on natural gas, severely restricting exploration, keeping low prices

Published by Lotus® Freelance Graphics®

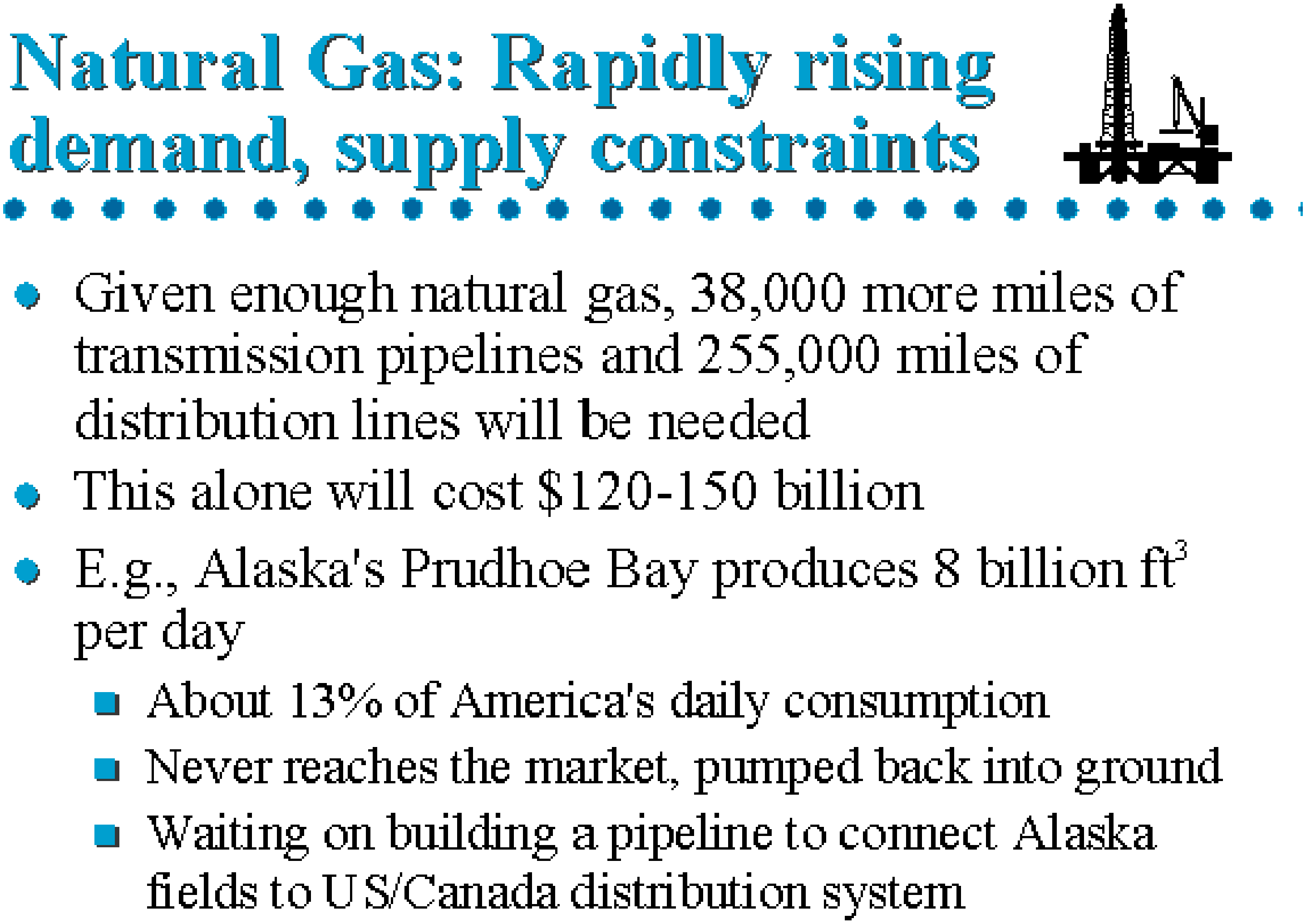

Natural Gas: Rapidly rising demand, supply constraints

- Given enough natural gas, 38,000 more miles of transmission pipelines and 255,000 miles of distribution lines will be needed
- This alone will cost $120-150 billion
- E.g., Alaska's Prudhoe Bay produces 8 billion ft^3 per day
  - About 13% of America's daily consumption
  - Never reaches the market, pumped back into ground
  - Waiting on building a pipeline to connect Alaska fields to US/Canada distribution system

Published by Lotus® Freelance Graphics®

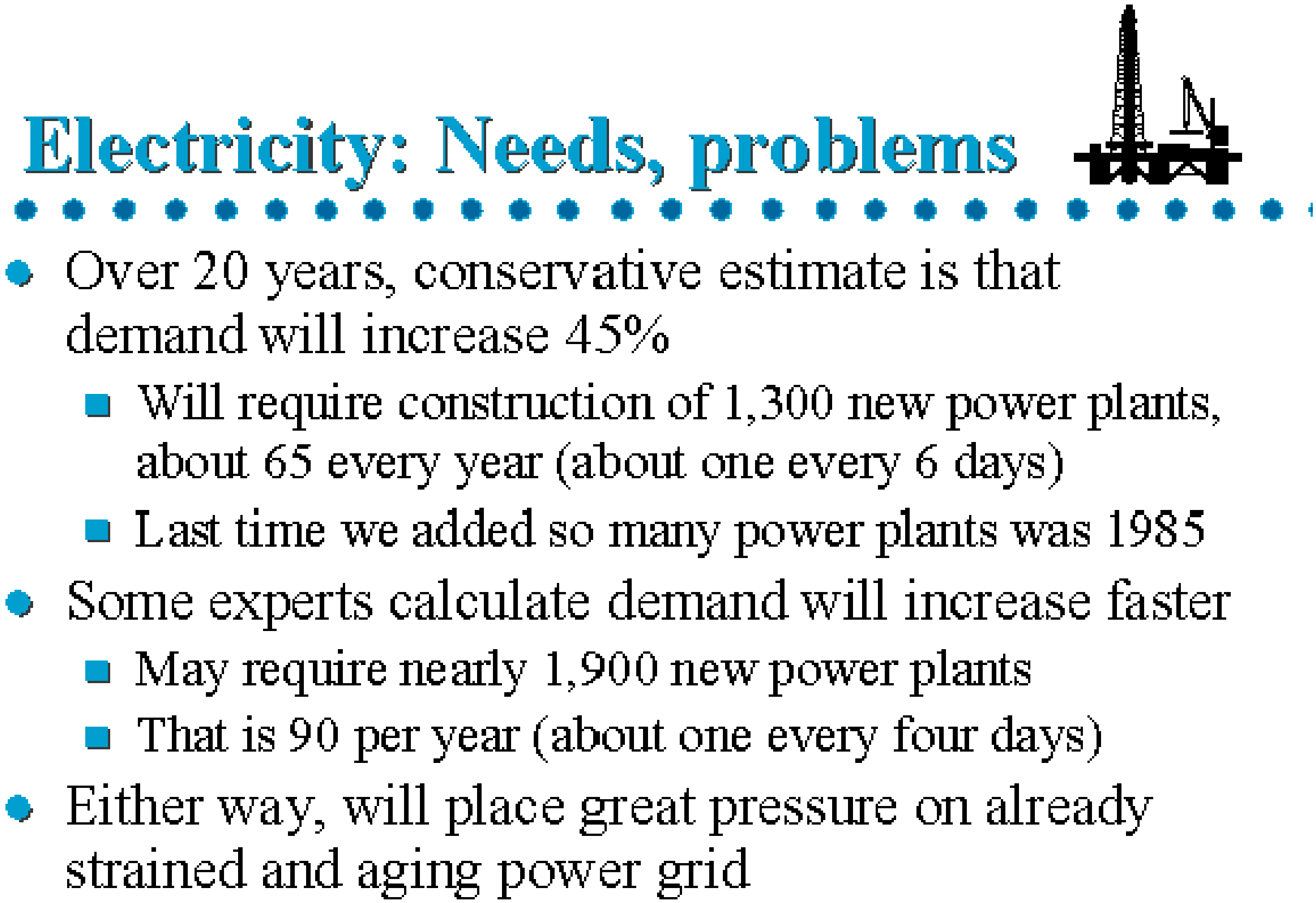

Electricity: Needs, problems

- Over 20 years, conservative estimate is that demand will increase 45%
  - Will require construction of 1,300 new power plants, about 65 every year (about one every 6 days)
  - Last time we added so many power plants was 1985
- Some experts calculate demand will increase faster
  - May require nearly 1,900 new power plants
  - That is 90 per year (about one every four days)
- Either way, will place great pressure on already strained and aging power grid

Published by Lotus® Freelance Graphics®

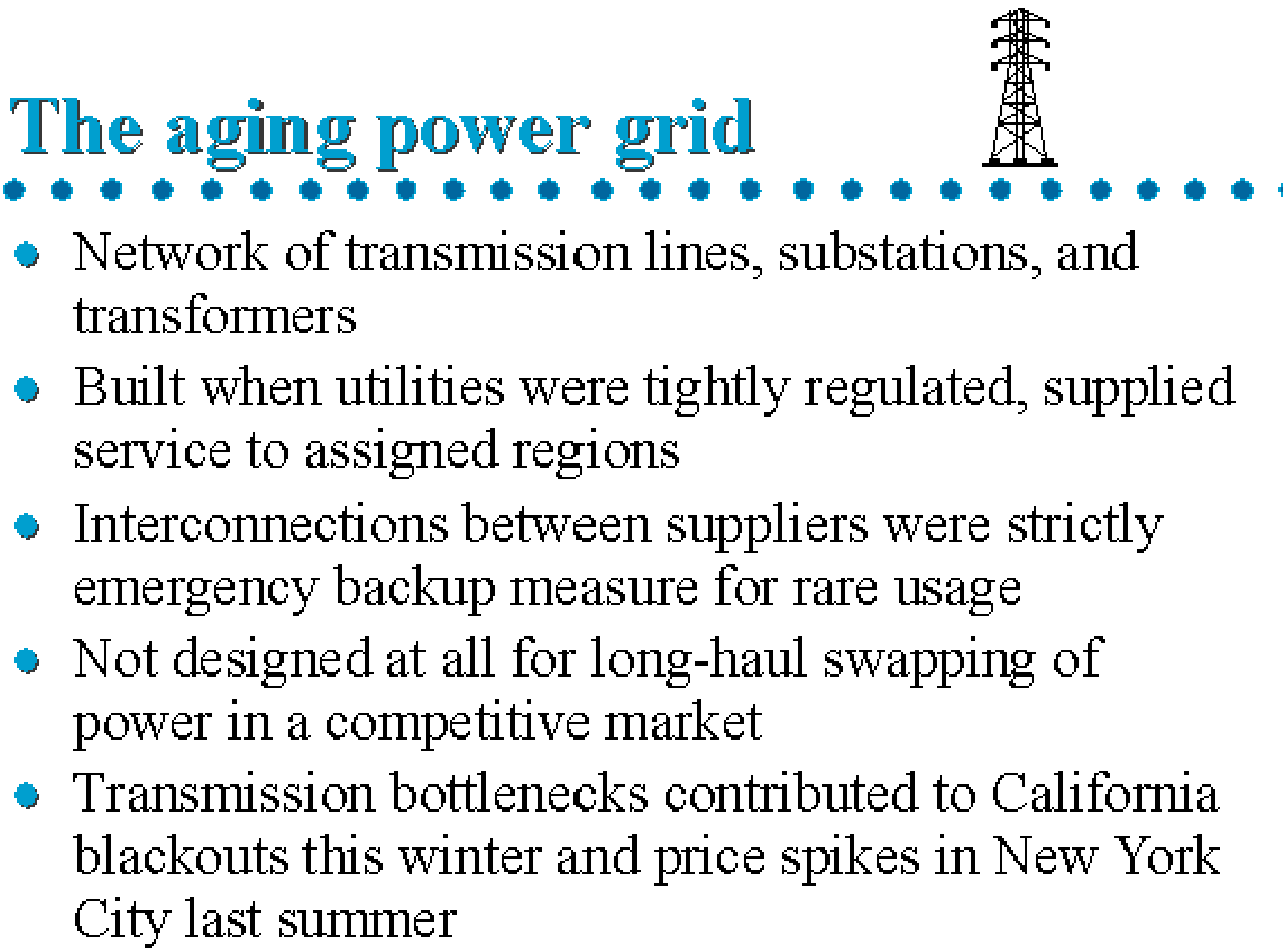

The aging power grid

- Network of transmission lines, substations, and transformers
- Built when utilities were tightly regulated, supplied service to assigned regions
- Interconnections between suppliers were strictly emergency backup measure for rare usage
- Not designed at all for long-haul swapping of power in a competitive market
- Transmission bottlenecks contributed to California blackouts this winter and price spikes in New York City last summer

Published by Lotus® Freelance Graphics®

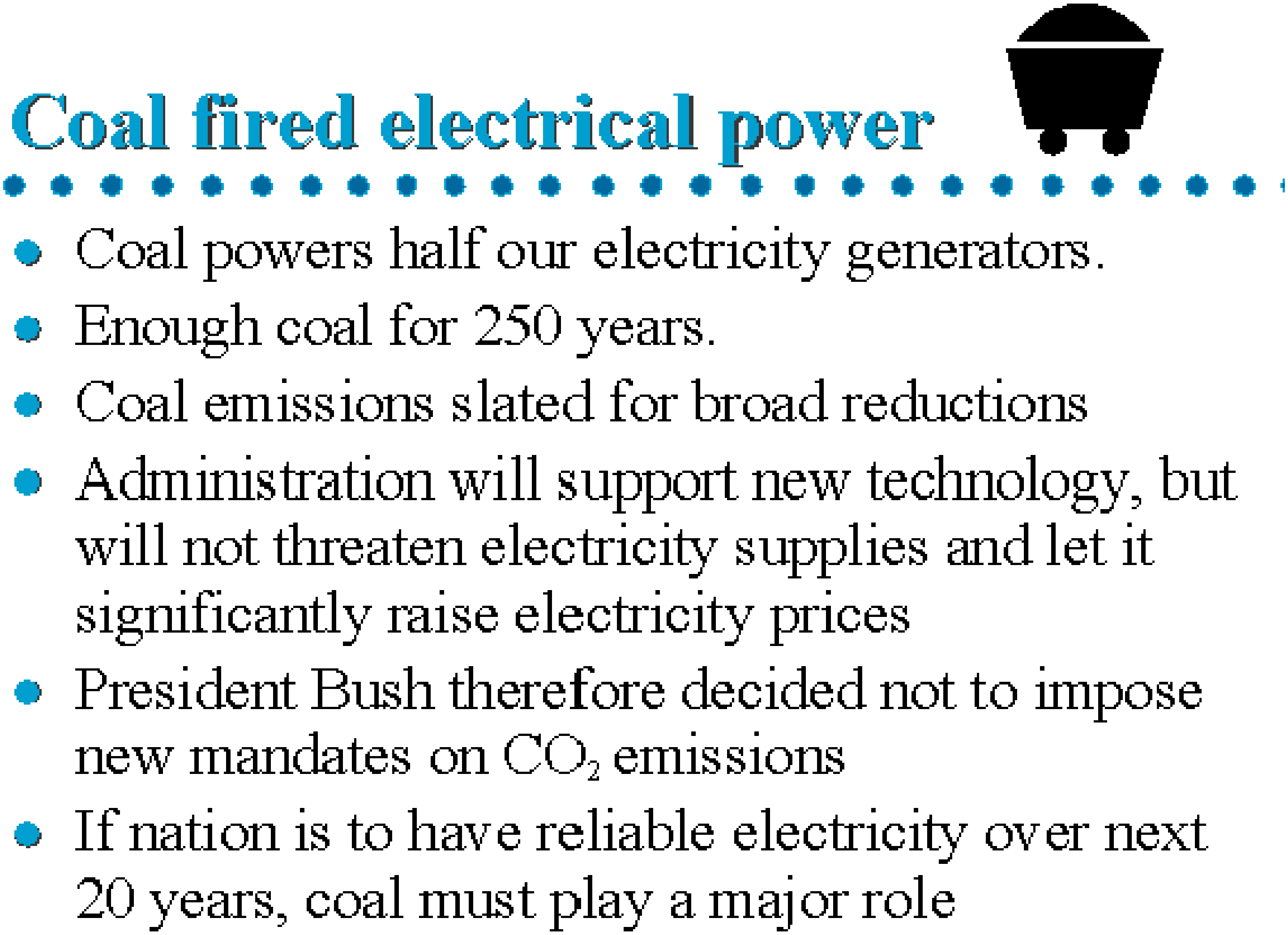

Coal fired electrical power

- Coal powers half our electricity generators.
- Enough coal for 250 years.
- Coal emissions slated for broad reductions
- Administration will support new technology, but will not threaten electricity supplies and let it significantly raise electricity prices
- President Bush therefore decided not to impose new mandates on CO_2 emissions
- If nation is to have reliable electricity over next 20 years, coal must play a major role

Published by Lotus® Freelance Graphics®

Nuclear power plants

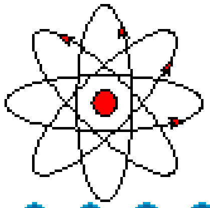

- No new nuclear power plant permit granted since 1979
- 103 existing nuclear power plants
- Many not expected to even file for license renewal as licenses expire over the next 15 years

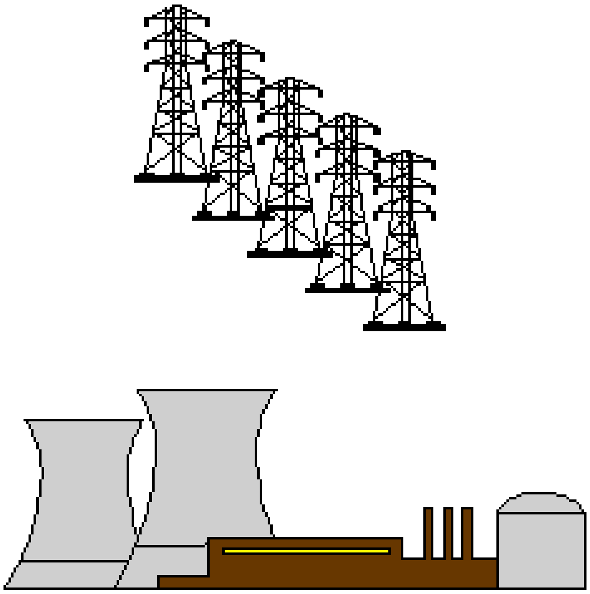

Published by Lotus® Freelance Graphics®

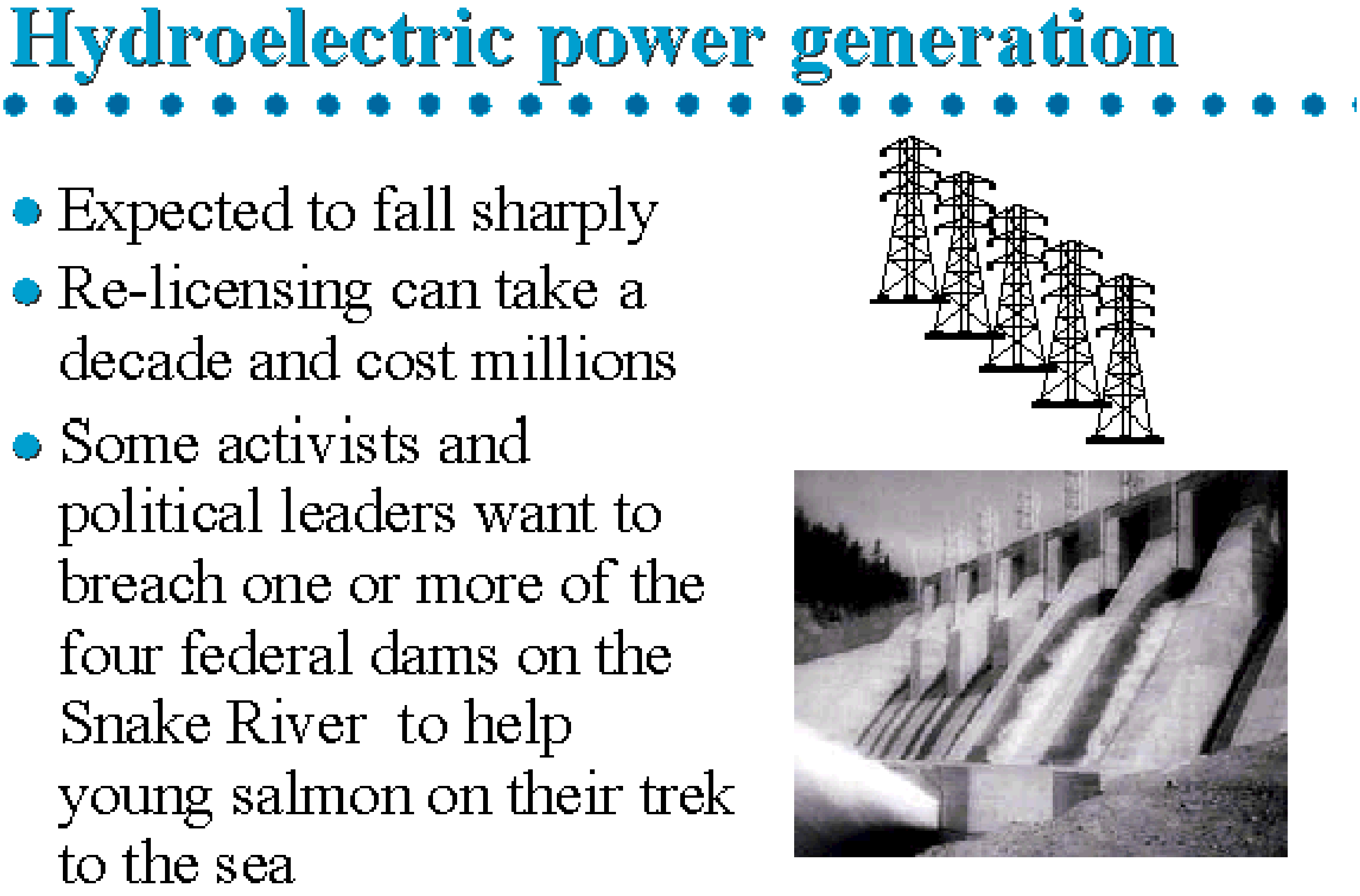

Hydroelectric power generation

- Expected to fall sharply
- Re-licensing can take a decade and cost millions
- Some activists and political leaders want to breach one or more of the four federal dams on the Snake River to help young salmon on their trek to the sea

Published by Lotus® Freelance Graphics®

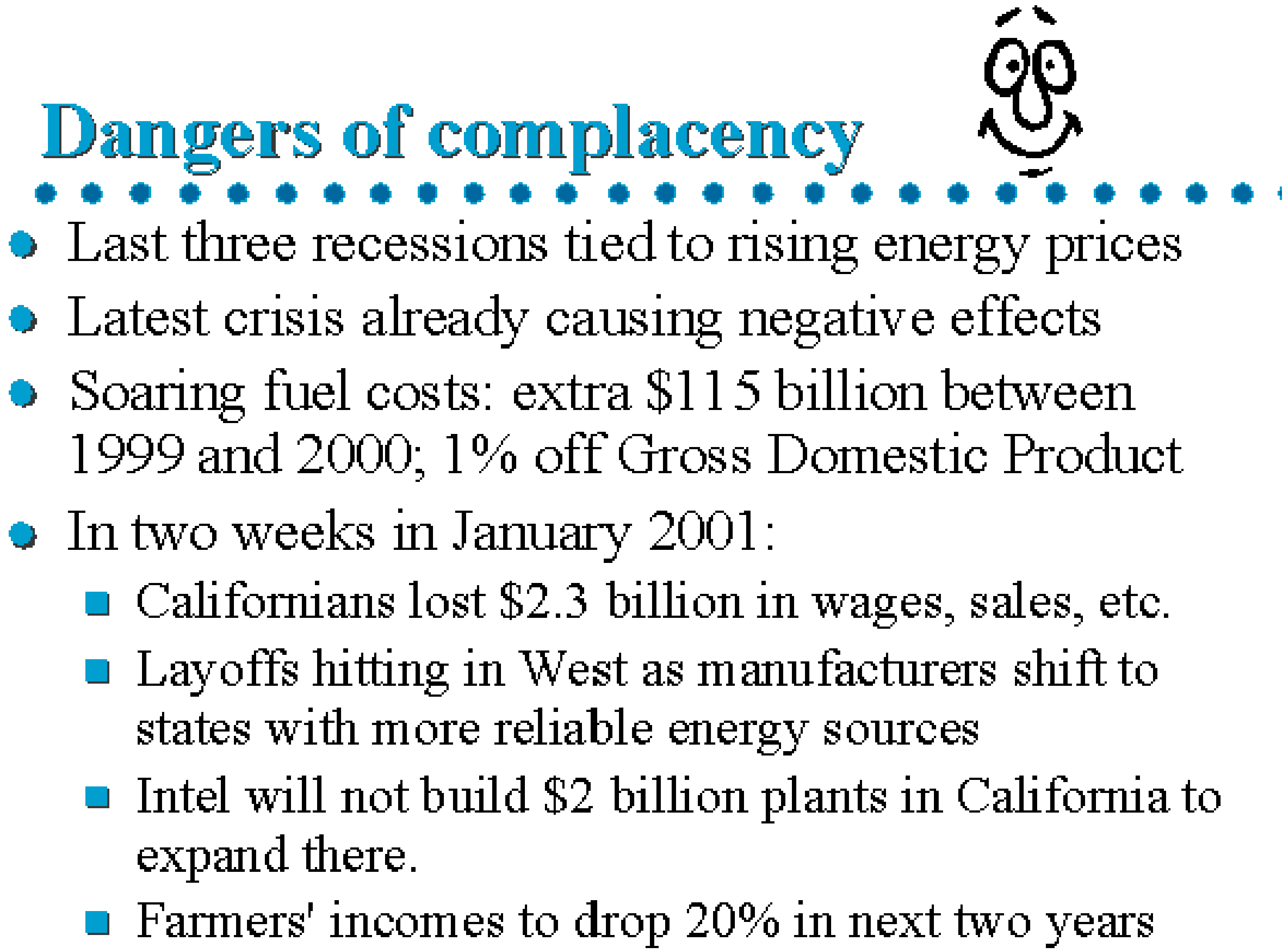

Dangers of complacency

- Last three recessions tied to rising energy prices
- Latest crisis already causing negative effects
- Soaring fuel costs: extra $115 billion between 1999 and 2000; 1% off Gross Domestic Product
- In two weeks in January 2001:
  - Californians lost $2.3 billion in wages, sales, etc.
  - Layoffs hitting in West as manufacturers shift to states with more reliable energy sources
  - Intel will not build $2 billion plants in California to expand there.
  - Farmers' incomes to drop 20% in next two years

Published by Lotus® Freelance Graphics®

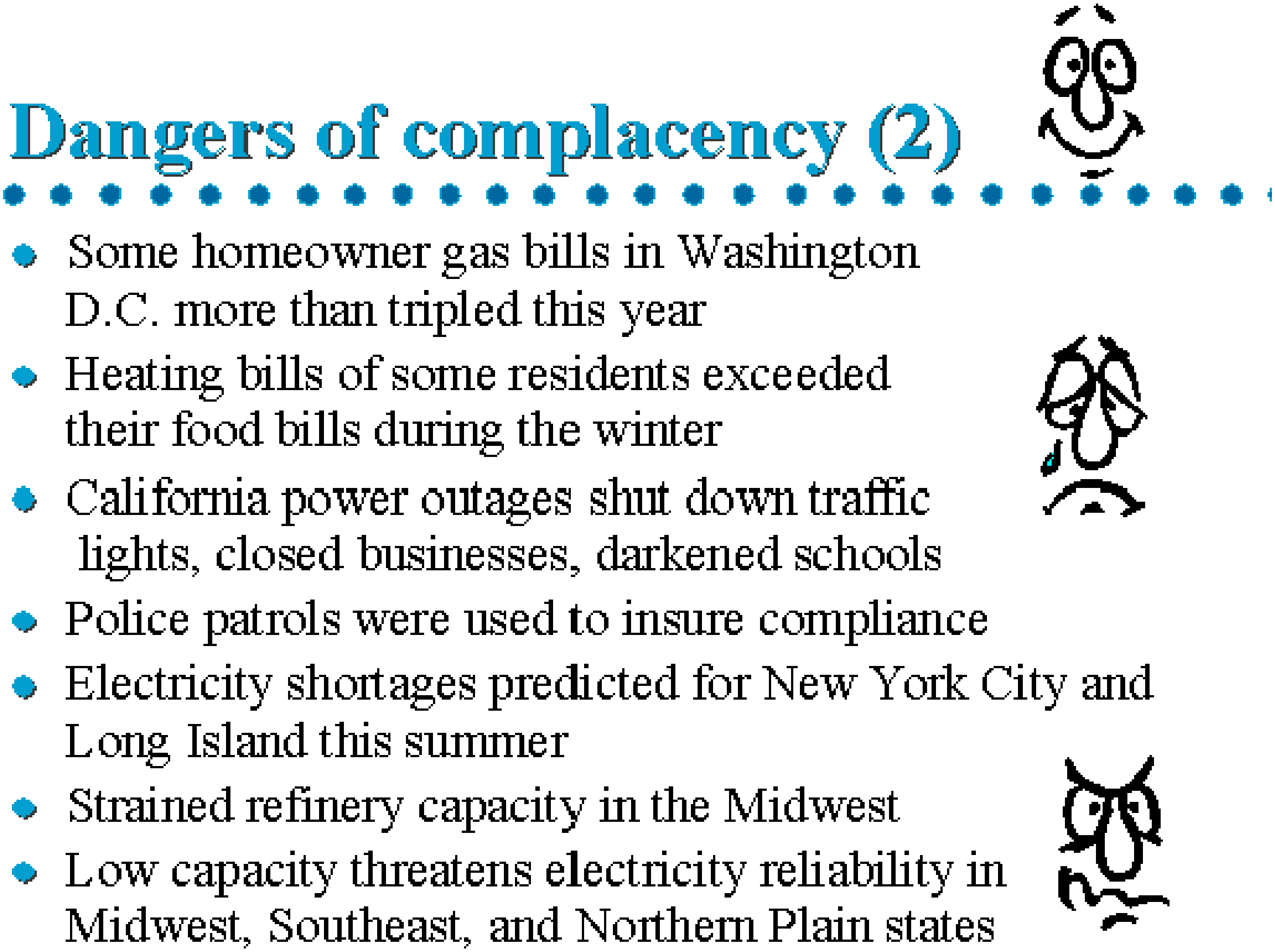

Dangers of complacency (2)

- Some homeowner gas bills in Washington D.C. more than tripled this year
- Heating bills of some residents exceeded their food bills during the winter
- California power outages shut down traffic lights, closed businesses, darkened schools
- Police patrols were used to insure compliance
- Electricity shortages predicted for New York City and Long Island this summer
- Strained refinery capacity in the Midwest
- Low capacity threatens electricity reliability in Midwest, Southeast, and Northern Plain states

Published by Lotus® Freelance Graphics®

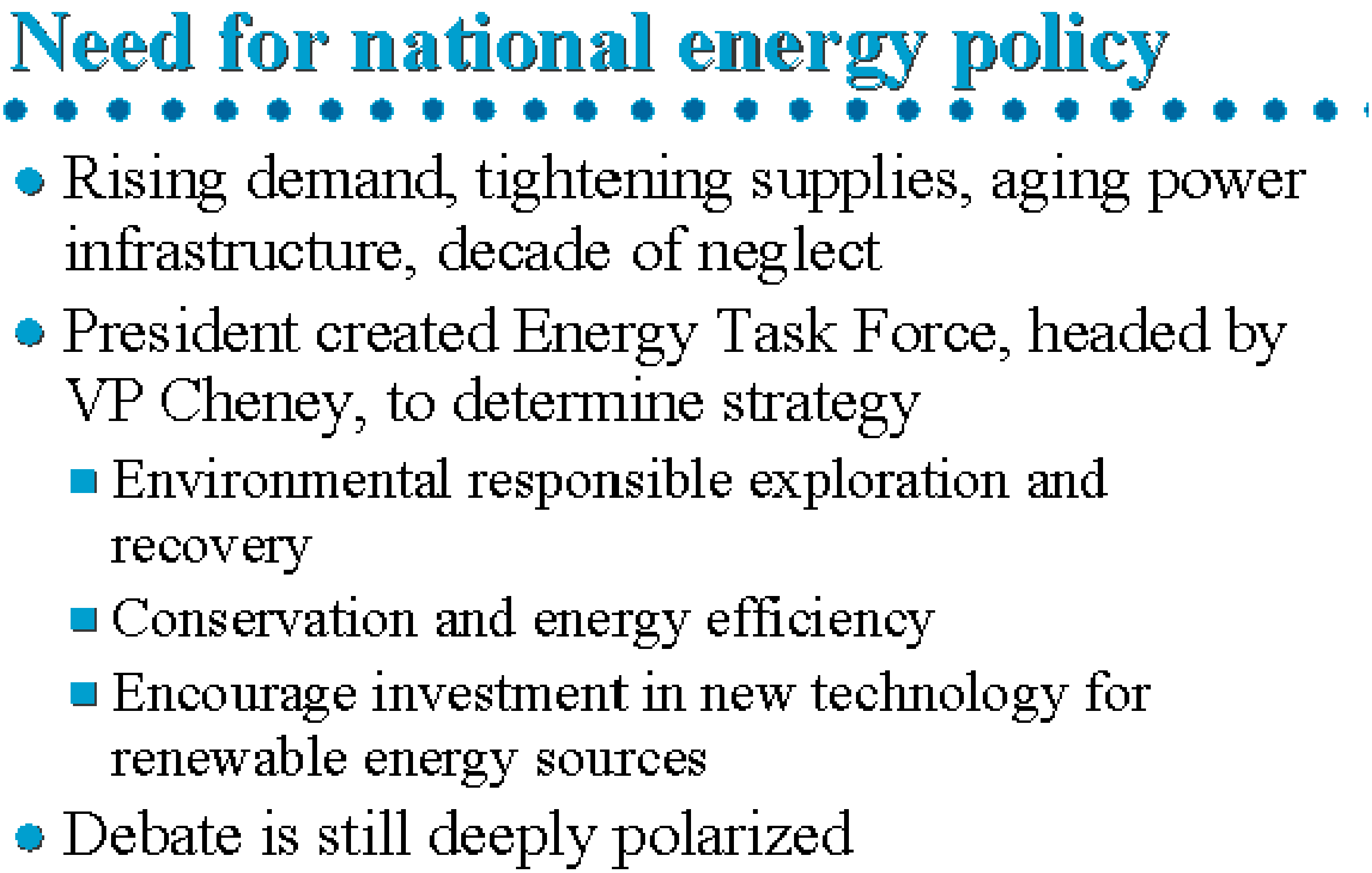

Need for national energy policy

- Rising demand, tightening supplies, aging power infrastructure, decade of neglect
- President created Energy Task Force, headed by VP Cheney, to determine strategy
  - Environmental responsible exploration and recovery
  - Conservation and energy efficiency
  - Encourage investment in new technology for renewable energy sources
- Debate is still deeply polarized

Published by Lotus® Freelance Graphics®

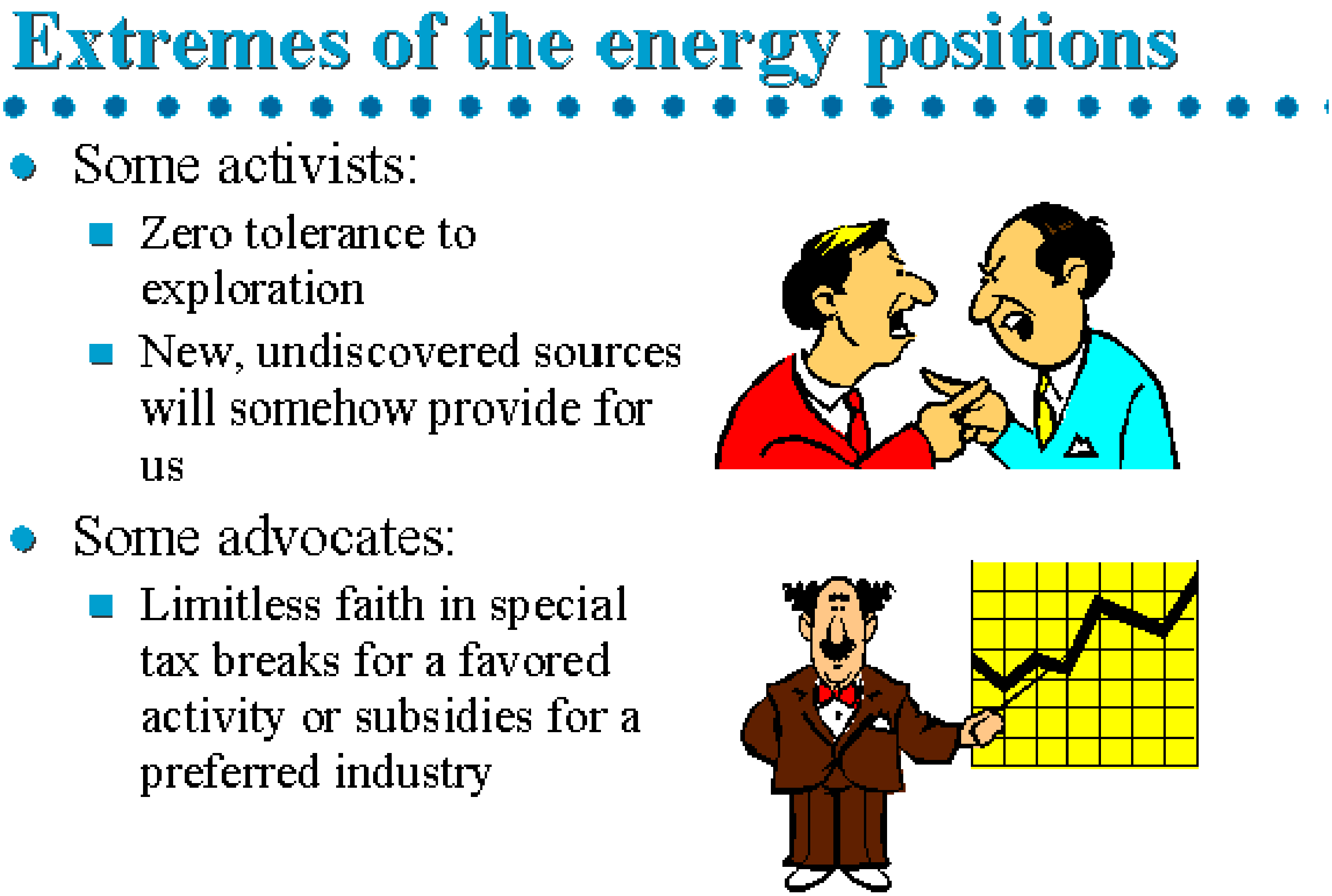

Extremes of the energy positions

- Some activists:
  - Zero tolerance to exploration
  - New, undiscovered sources will somehow provide for us
- Some advocates:
  - Limitless faith in special tax breaks for a favored activity or subsidies for a preferred industry

Published by Lotus® Freelance Graphics®

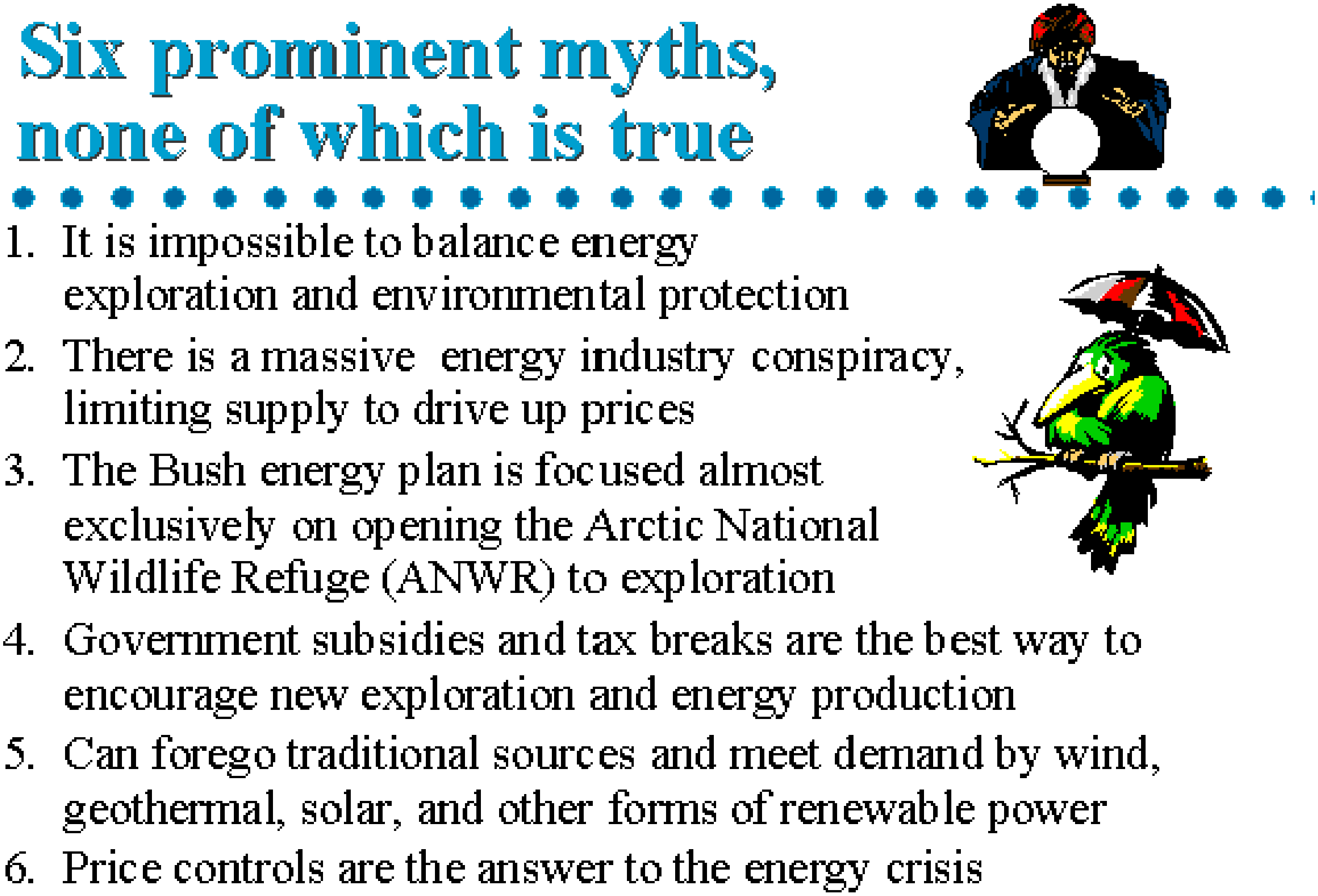

Six prominent myths, none of which is true

1. It is impossible to balance energy exploration and environmental protection
2. There is a massive energy industry conspiracy, limiting supply to drive up prices
3. The Bush energy plan is focused almost exclusively on opening the Arctic National Wildlife Refuge (ANWR) to exploration
4. Government subsidies and tax breaks are the best way to encourage new exploration and energy production
5. Can forego traditional sources and meet demand by wind, geothermal, solar, and other forms of renewable power
6. Price controls are the answer to the energy crisis

Published by Lotus® Freelance Graphics®

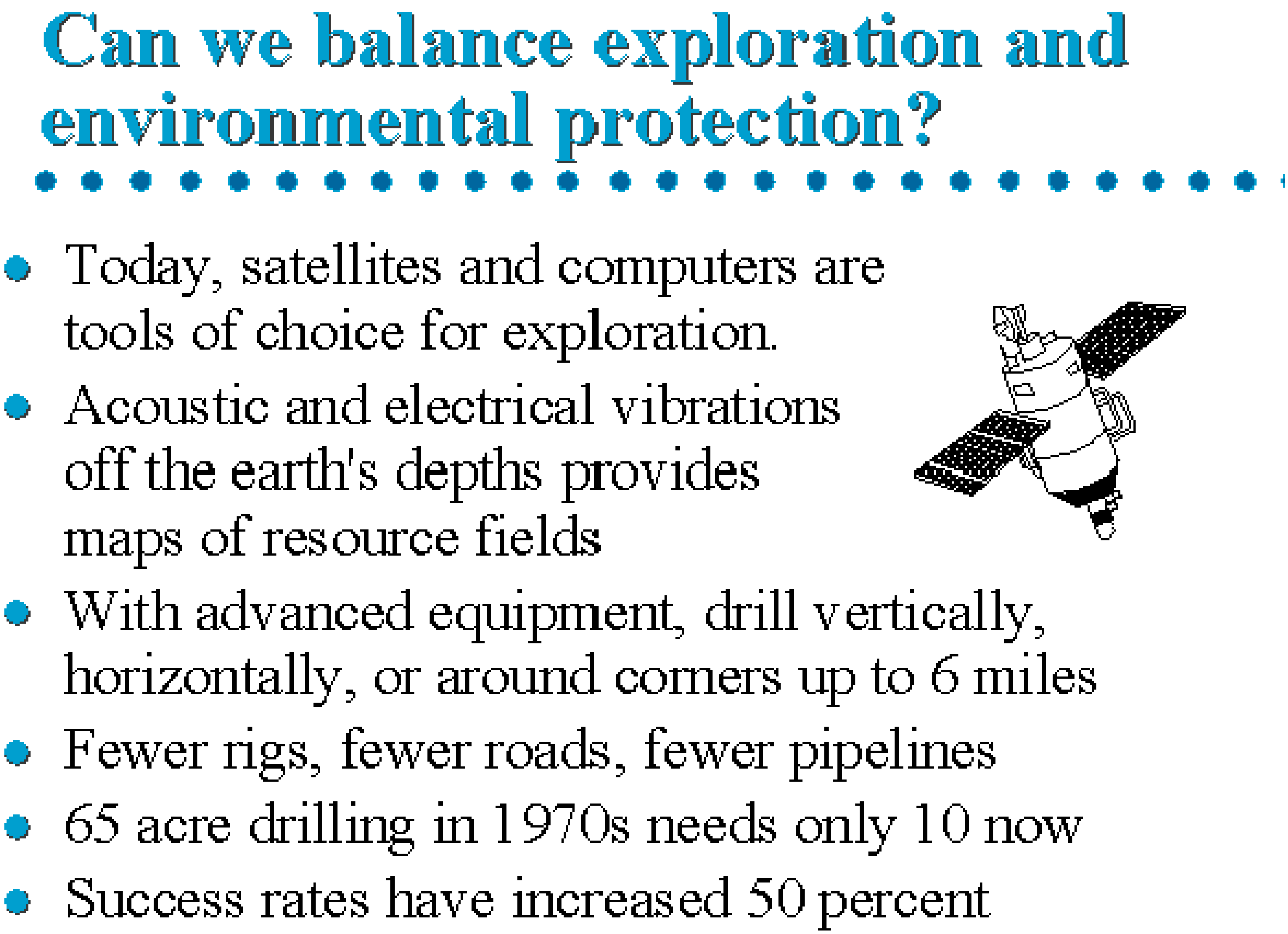

Can we balance exploration and environmental protection?

- Today, satellites and computers are tools of choice for exploration.
- Acoustic and electrical vibrations off the earth's depths provides maps of resource fields
- With advanced equipment, drill vertically, horizontally, or around corners up to 6 miles
- Fewer rigs, fewer roads, fewer pipelines
- 65 acre drilling in 1970s needs only 10 now
- Success rates have increased 50 percent

Published by Lotus® Freelance Graphics®

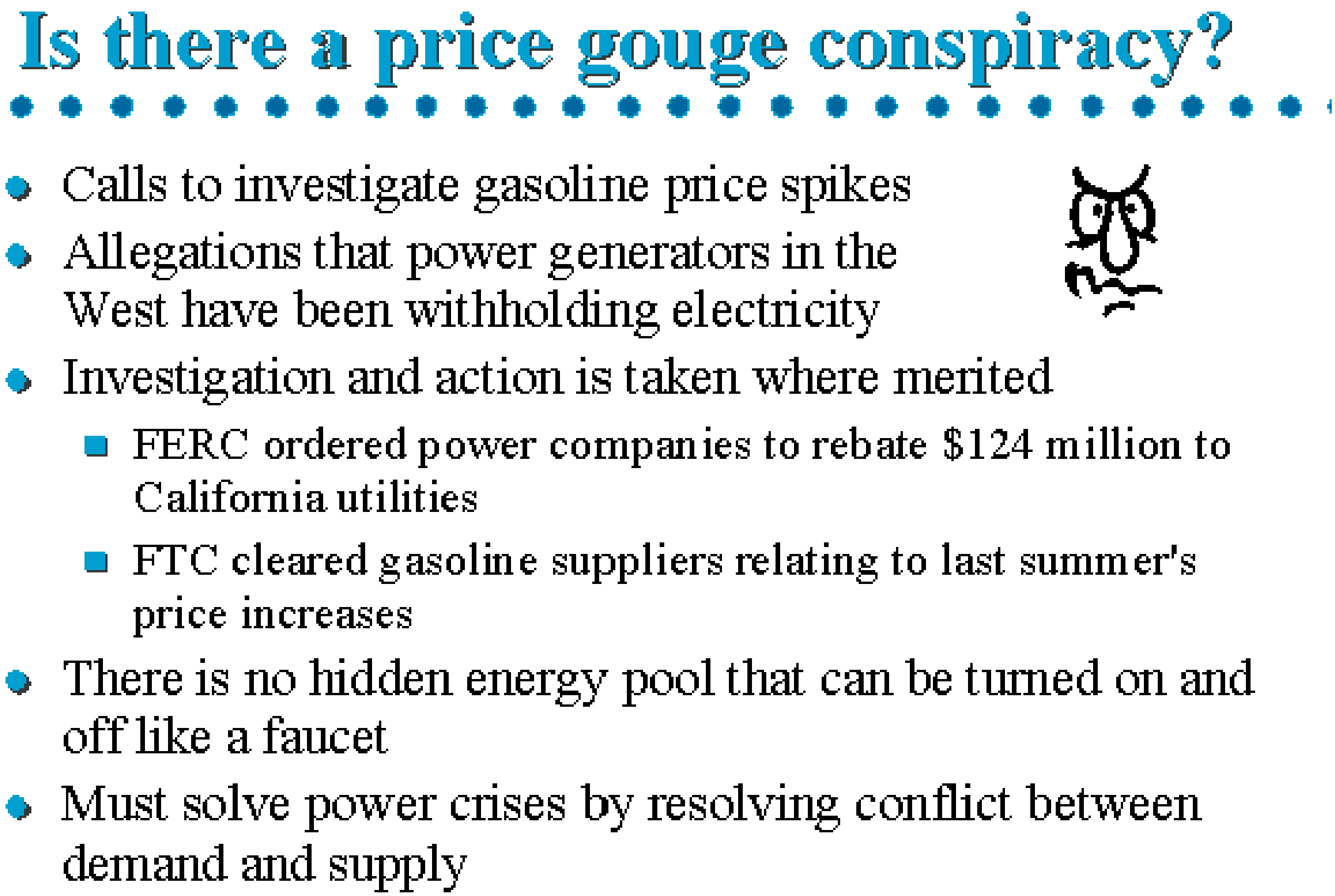

Is there a price gouge conspiracy?

- Calls to investigate gasoline price spikes
- Allegations that power generators in the West have been withholding electricity
- Investigation and action is taken where merited
  - FERC ordered power companies to rebate $124 million to California utilities
  - FTC cleared gasoline suppliers relating to last summer's price increases
- There is no hidden energy pool that can be turned on and off like a faucet
- Must solve power crises by resolving conflict between demand and supply

Published by Lotus® Freelance Graphics®

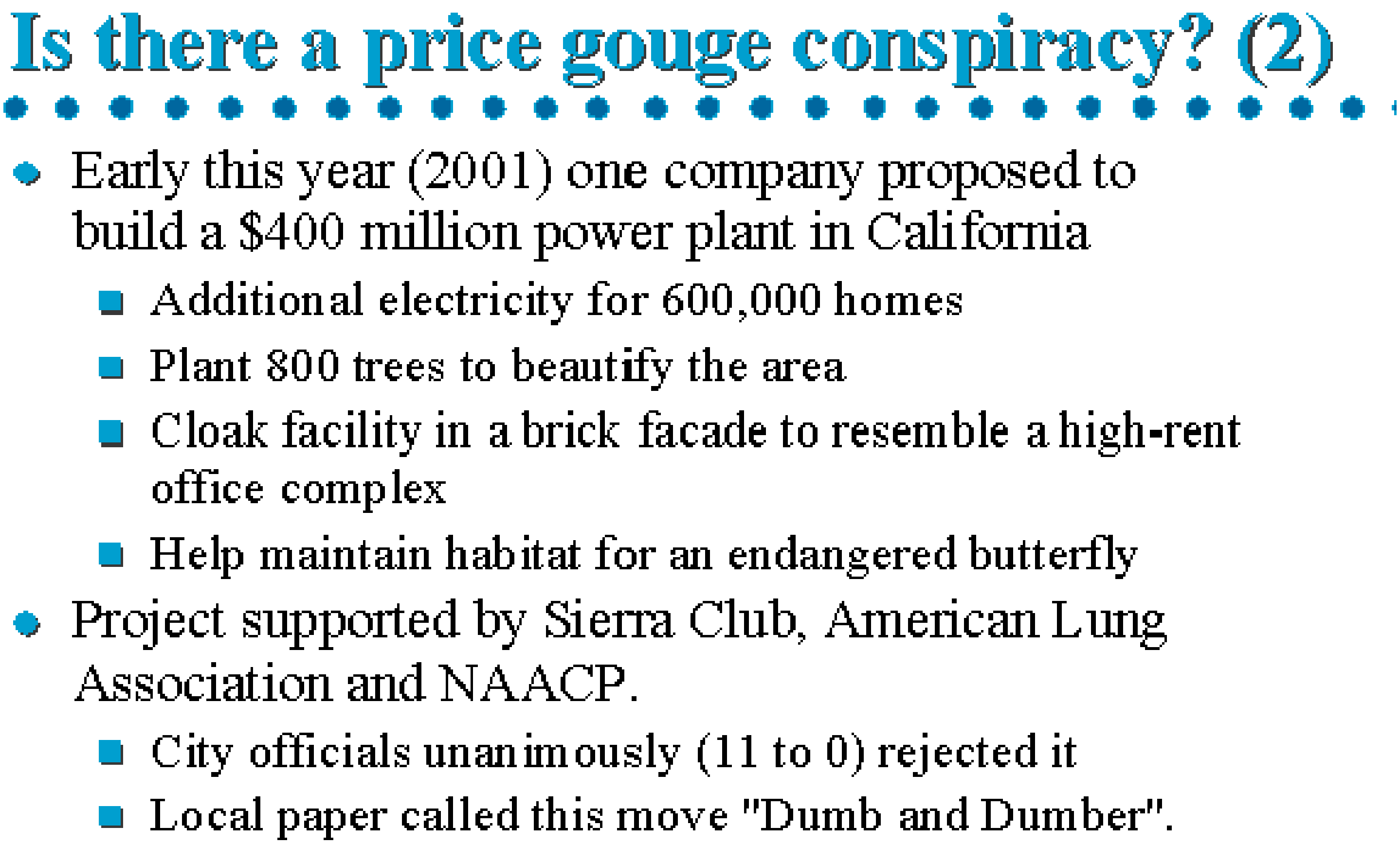

Is there a price gouge conspiracy? (2)

- Early this year (2001) one company proposed to build a $400 million power plant in California
  - Additional electricity for 600,000 homes
  - Plant 800 trees to beautify the area
  - Cloak facility in a brick facade to resemble a high-rent office complex
  - Help maintain habitat for an endangered butterfly
- Project supported by Sierra Club, American Lung Association and NAACP.
  - City officials unanimously (11 to 0) rejected it
  - Local paper called this move "Dumb and Dumber".

Published by Lotus® Freelance Graphics®

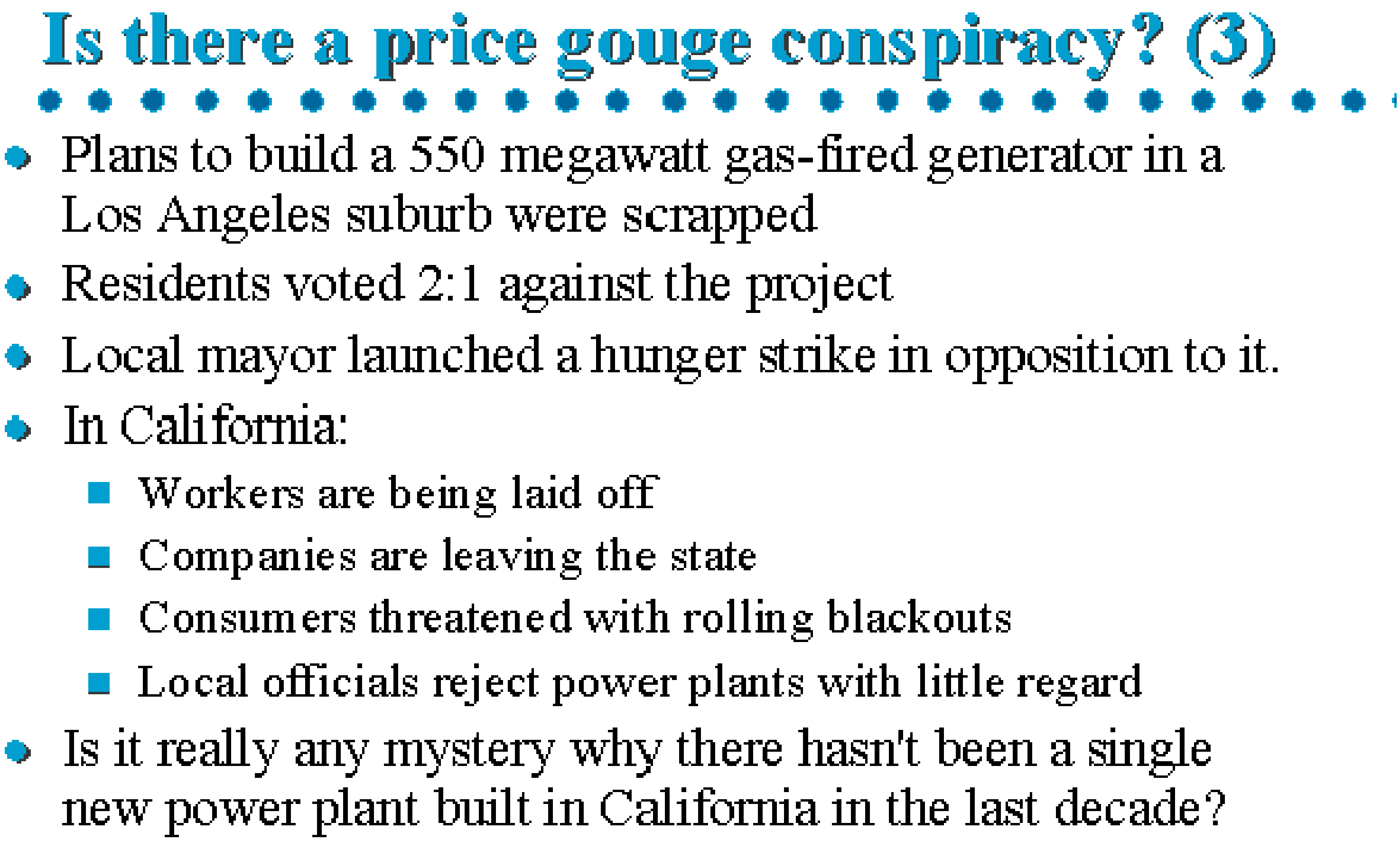

Is there a price gouge conspiracy? (3)

- Plans to build a 550 megawatt gas-fired generator in a Los Angeles suburb were scrapped
- Residents voted 2:1 against the project
- Local mayor launched a hunger strike in opposition to it.
- In California:
    - Workers are being laid off
    - Companies are leaving the state
    - Consumers threatened with rolling blackouts
    - Local officials reject power plants with little regard
- Is it really any mystery why there hasn't been a single new power plant built in California in the last decade?

Published by Lotus® Freelance Graphics®

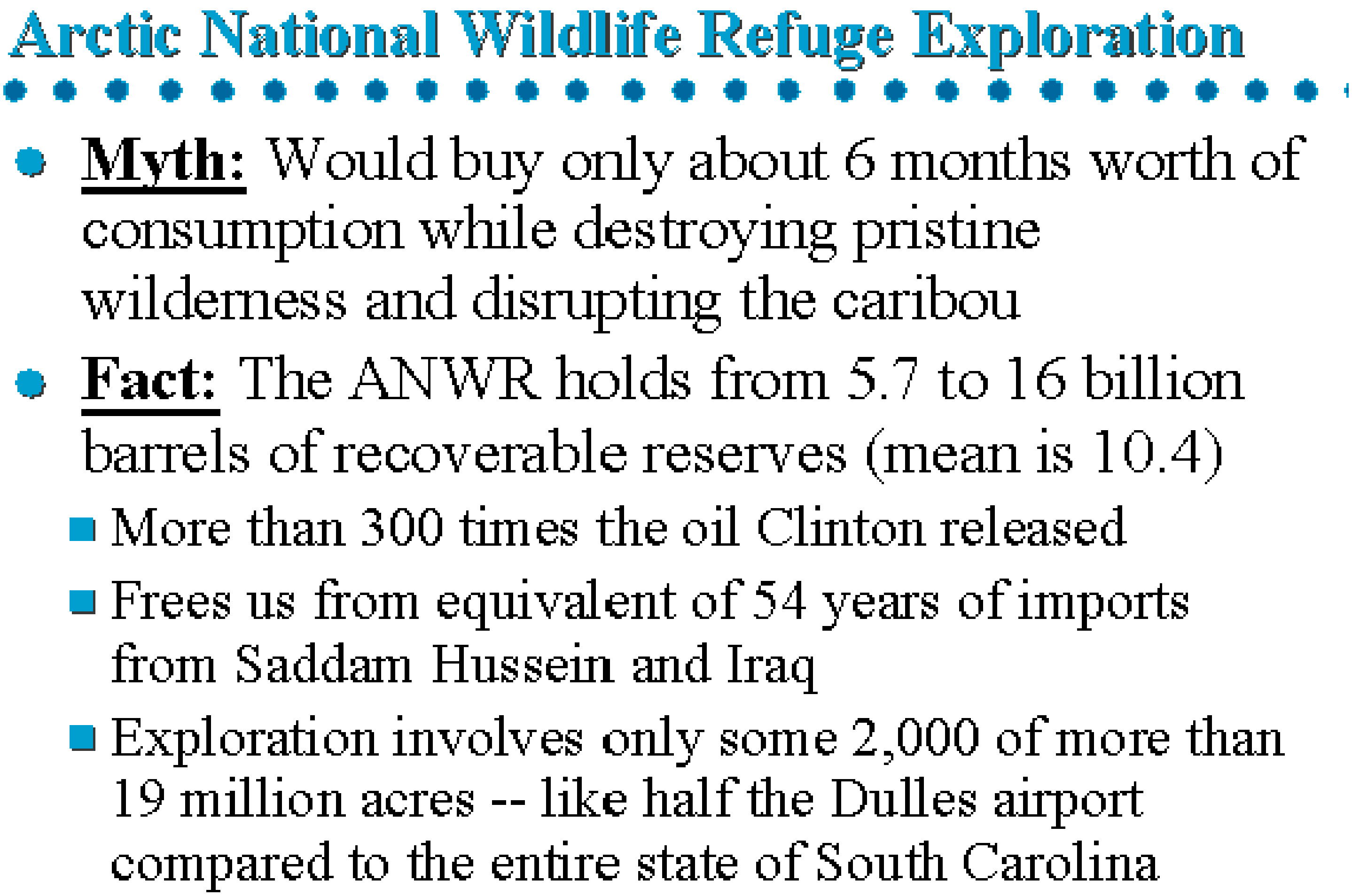

Arctic National Wildlife Refuge Exploration

- **Myth:** Would buy only about 6 months worth of consumption while destroying pristine wilderness and disrupting the caribou
- **Fact:** The ANWR holds from 5.7 to 16 billion barrels of recoverable reserves (mean is 10.4)
  - More than 300 times the oil Clinton released
  - Frees us from equivalent of 54 years of imports from Saddam Hussein and Iraq
  - Exploration involves only some 2,000 of more than 19 million acres -- like half the Dulles airport compared to the entire state of South Carolina

Published by Lotus® Freelance Graphics®

Arctic National Wildlife Refuge Exploration (2)

- The caribou herd in the Prudhoe Bay area grew more than 9-fold over the last 20 years (to 28,000 in 2000)
- Proves that caribou mating and oil exploration can peacefully co-exist
- Exploration should be considered on merits
- Opening ANWR will help, but does not allow full energy dependence

Published by Lotus® Freelance Graphics®

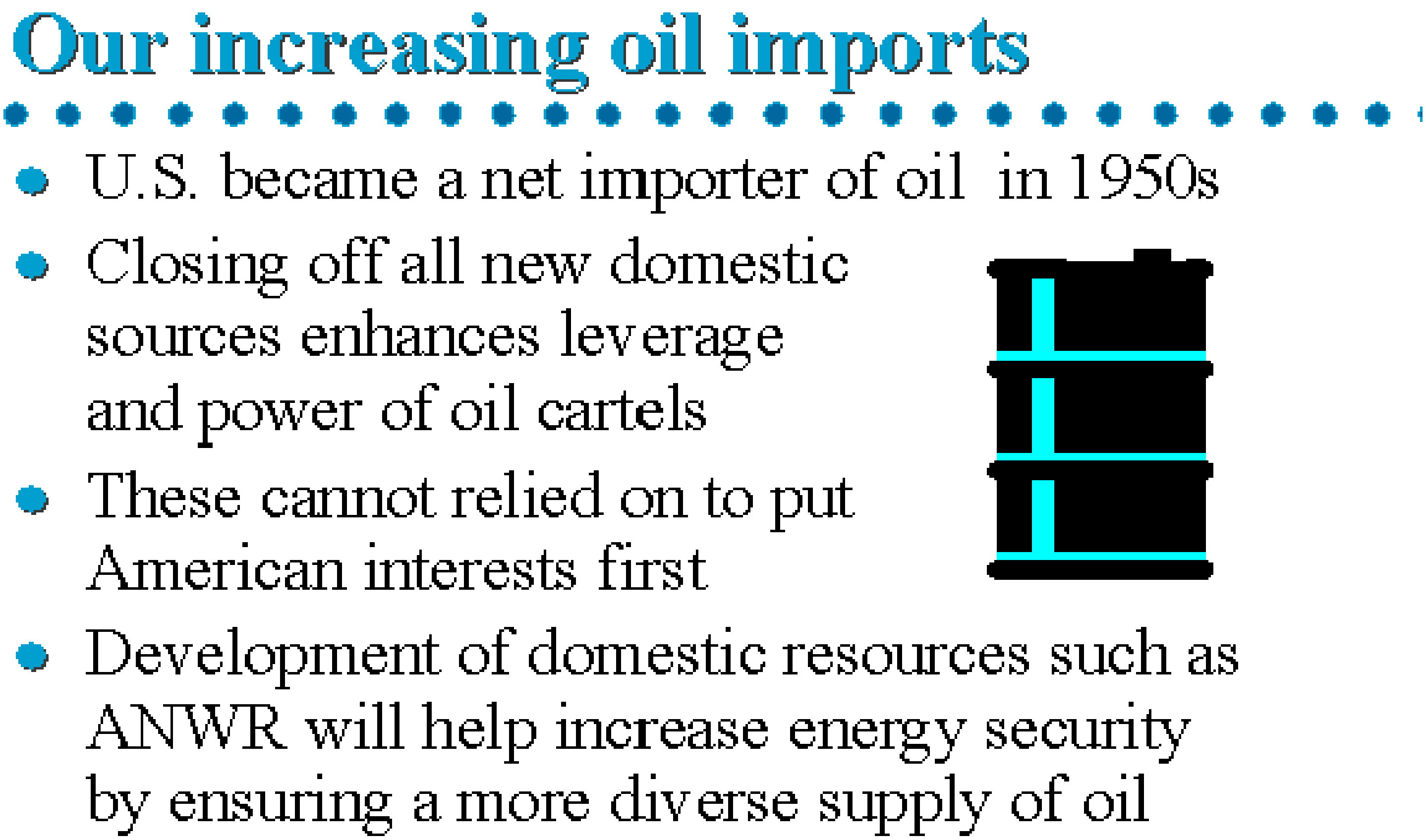

Our increasing oil imports

- U.S. became a net importer of oil in 1950s
- Closing off all new domestic sources enhances leverage and power of oil cartels
- These cannot relied on to put American interests first
- Development of domestic resources such as ANWR will help increase energy security by ensuring a more diverse supply of oil

Published by Lotus® Freelance Graphics®

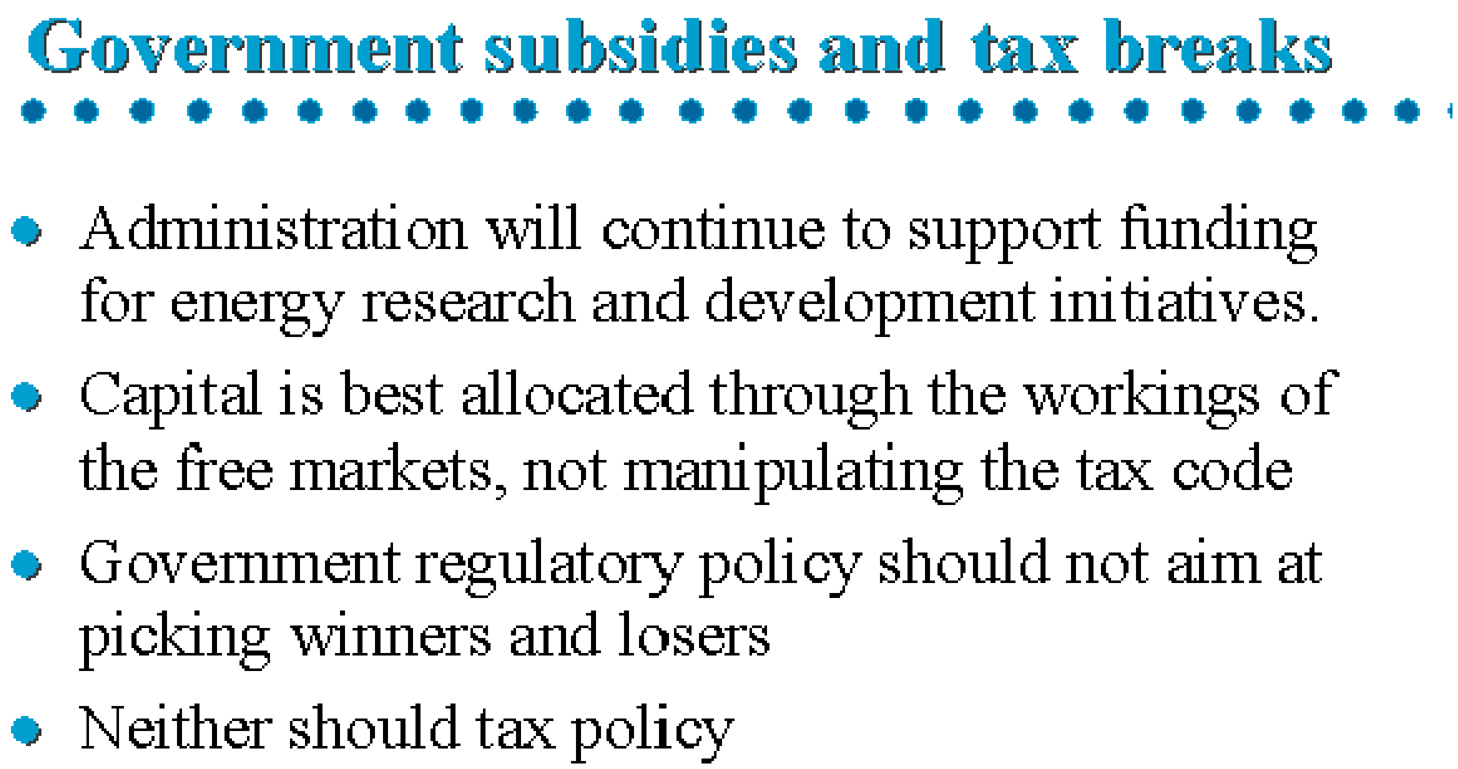

Government subsidies and tax breaks

- Administration will continue to support funding for energy research and development initiatives.
- Capital is best allocated through the workings of the free markets, not manipulating the tax code
- Government regulatory policy should not aim at picking winners and losers
- Neither should tax policy

Published by Lotus® Freelance Graphics®

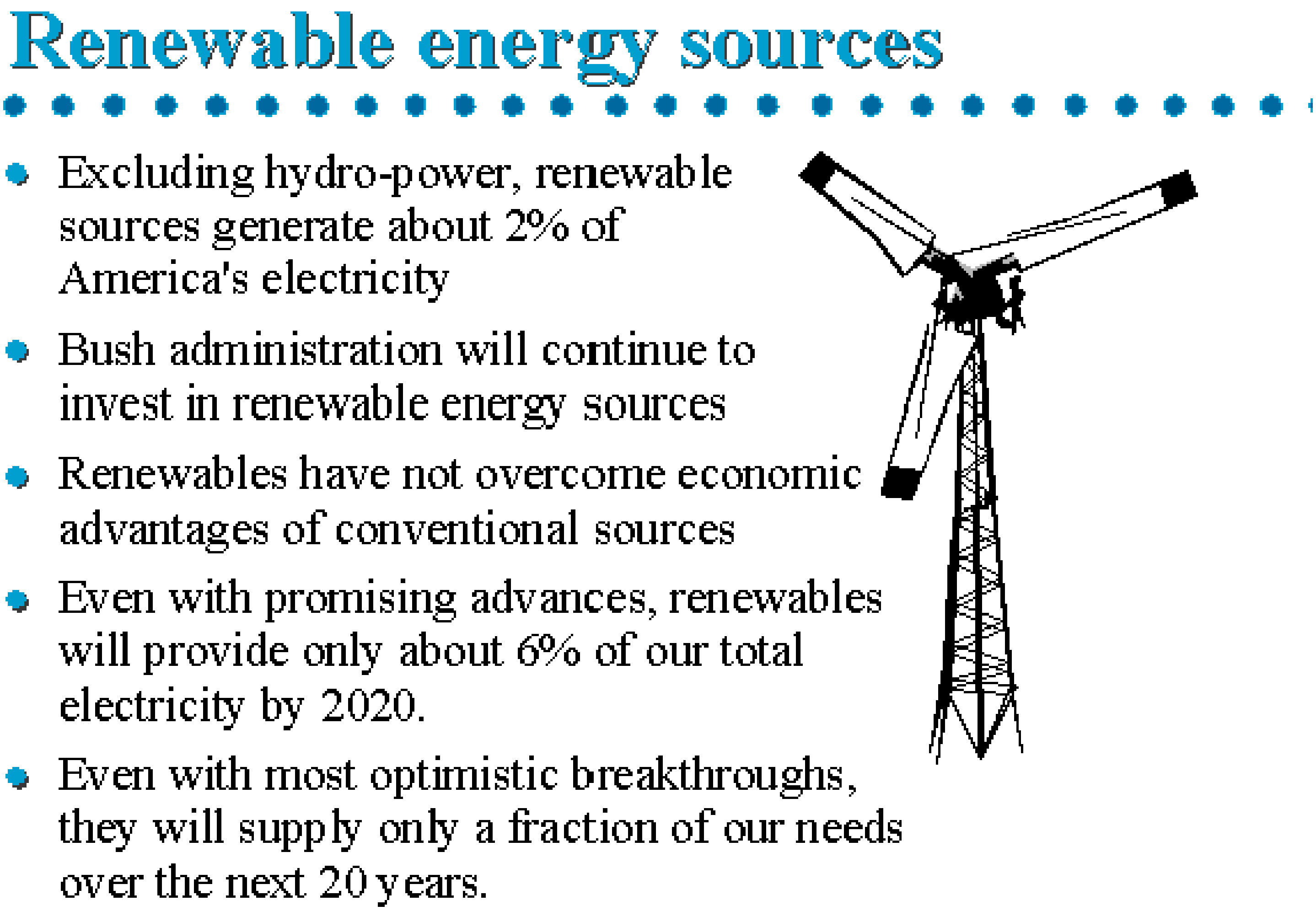

Renewable energy sources

- Excluding hydro-power, renewable sources generate about 2% of America's electricity
- Bush administration will continue to invest in renewable energy sources
- Renewables have not overcome economic advantages of conventional sources
- Even with promising advances, renewables will provide only about 6% of our total electricity by 2020.
- Even with most optimistic breakthroughs, they will supply only a fraction of our needs over the next 20 years.

Published by Lotus® Freelance Graphics®

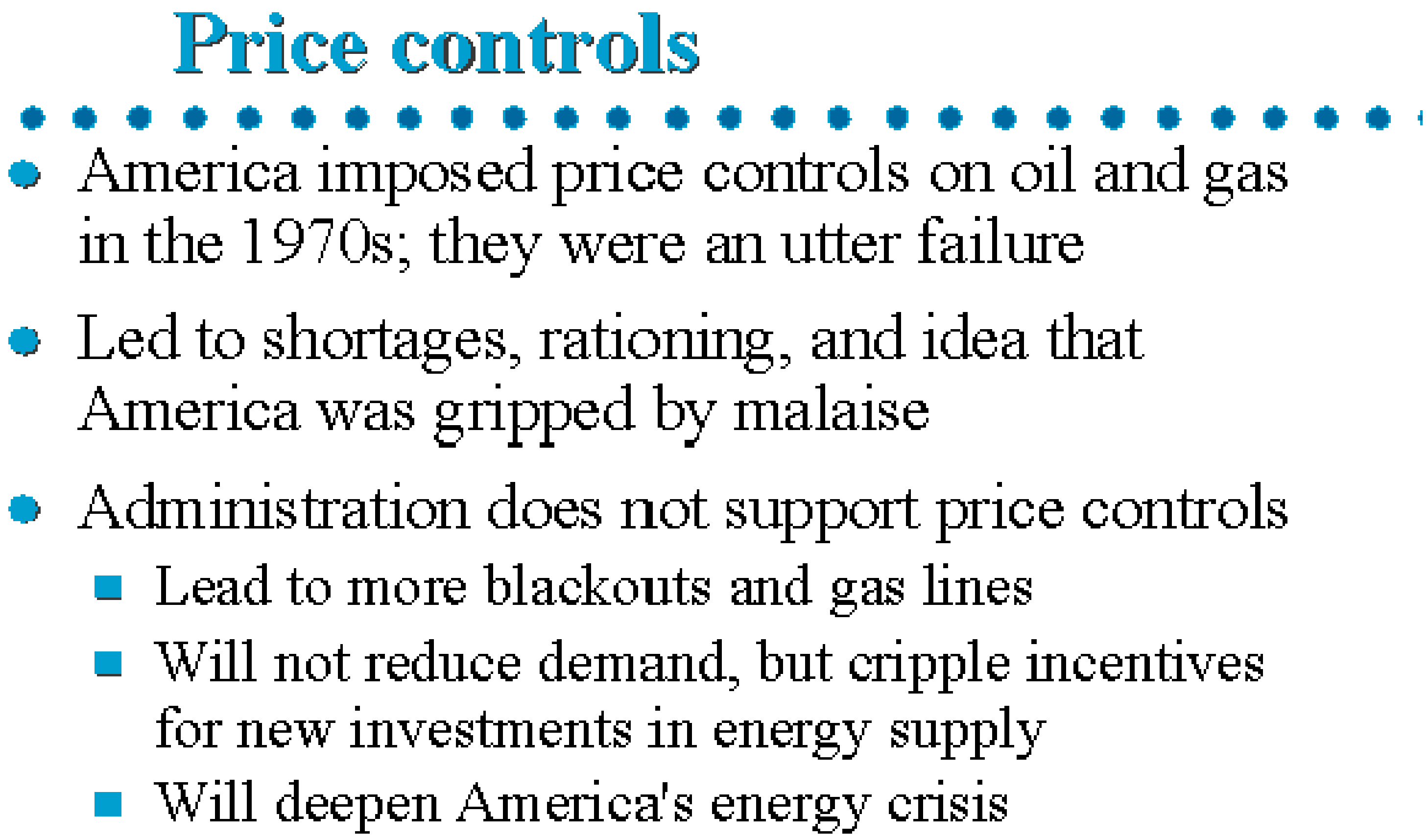

Price controls

- America imposed price controls on oil and gas in the 1970s; they were an utter failure
- Led to shortages, rationing, and idea that America was gripped by malaise
- Administration does not support price controls
  - Lead to more blackouts and gas lines
  - Will not reduce demand, but cripple incentives for new investments in energy supply
  - Will deepen America's energy crisis

Published by Lotus® Freelance Graphics®

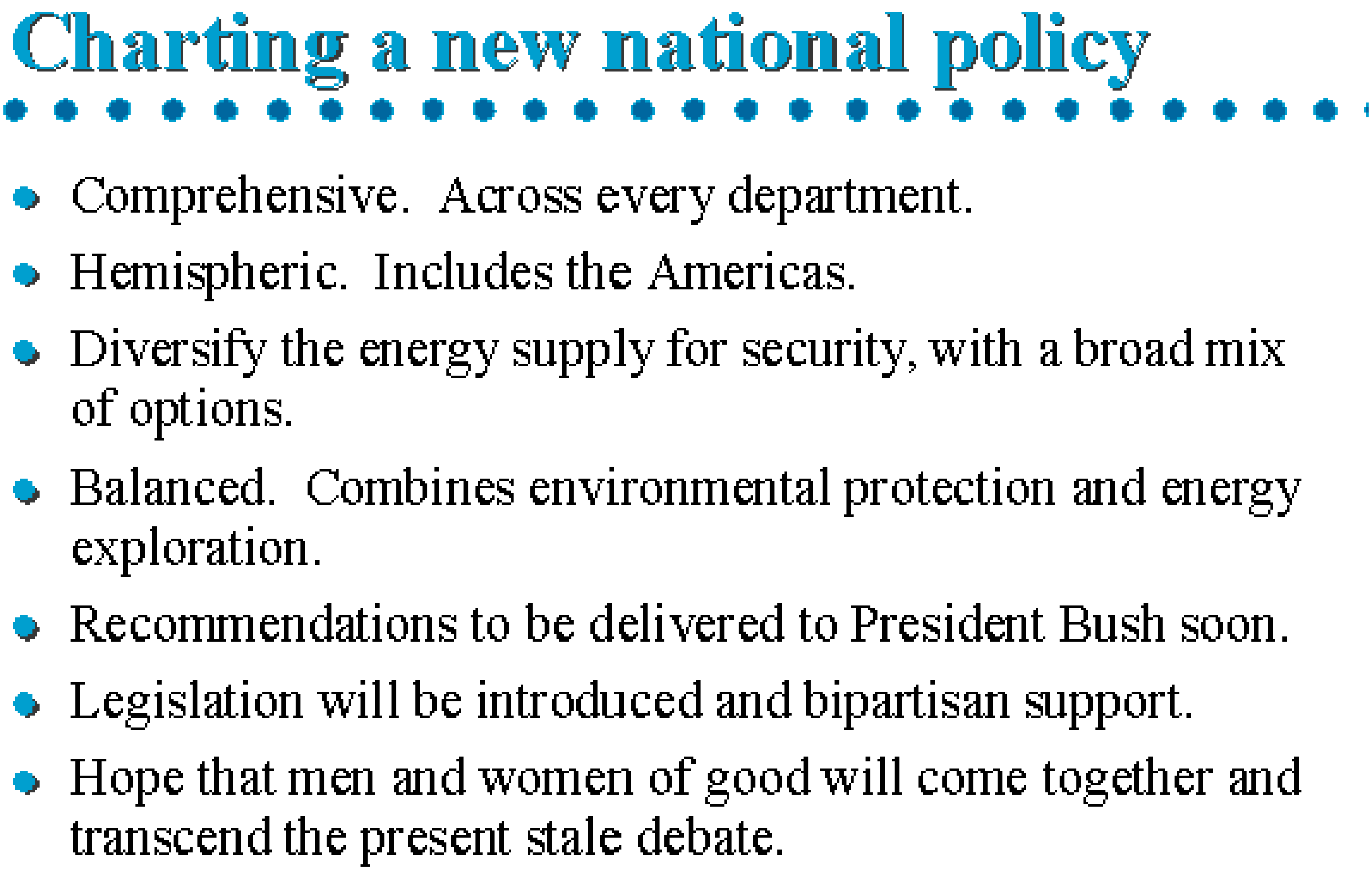

Charting a new national policy

- Comprehensive. Across every department.
- Hemispheric. Includes the Americas.
- Diversify the energy supply for security, with a broad mix of options.
- Balanced. Combines environmental protection and energy exploration.
- Recommendations to be delivered to President Bush soon.
- Legislation will be introduced and bipartisan support.
- Hope that men and women of good will come together and transcend the present stale debate.

Published by Lotus® Freelance Graphics®

Commentary

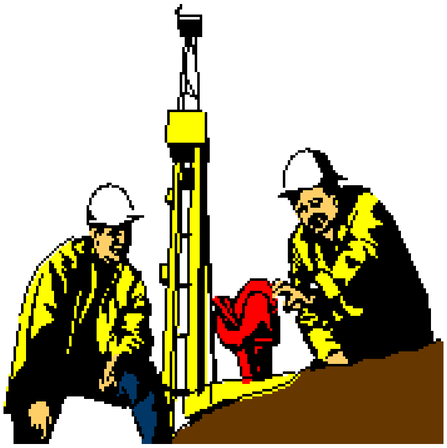

by Tom Bearden, Ph.D.

12 April 2001

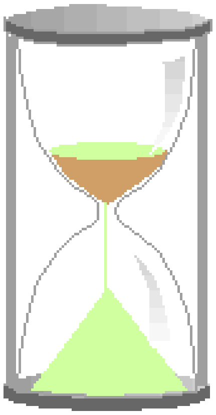

Published by Lotus® Freelance Graphics®

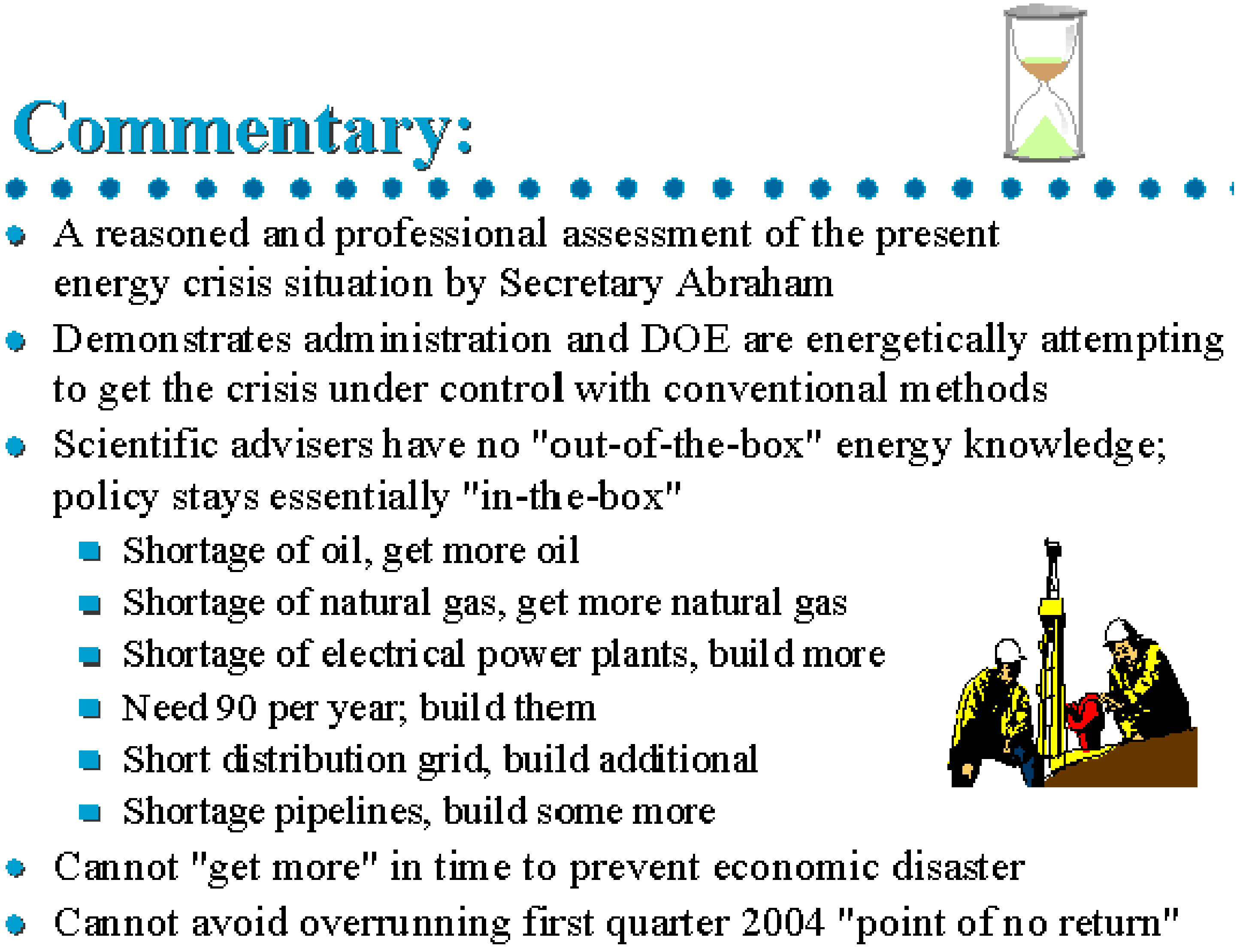

Commentary:

- A reasoned and professional assessment of the present energy crisis situation by Secretary Abraham
- Demonstrates administration and DOE are energetically attempting to get the crisis under control with conventional methods
- Scientific advisers have no "out-of-the-box" energy knowledge; policy stays essentially "in-the-box"
    - Shortage of oil, get more oil
    - Shortage of natural gas, get more natural gas
    - Shortage of electrical power plants, build more
    - Need 90 per year; build them
    - Short distribution grid, build additional
    - Shortage pipelines, build some more
- Cannot "get more" in time to prevent economic disaster
- Cannot avoid overrunning first quarter 2004 "point of no return"

Published by Lotus® Freelance Graphics®

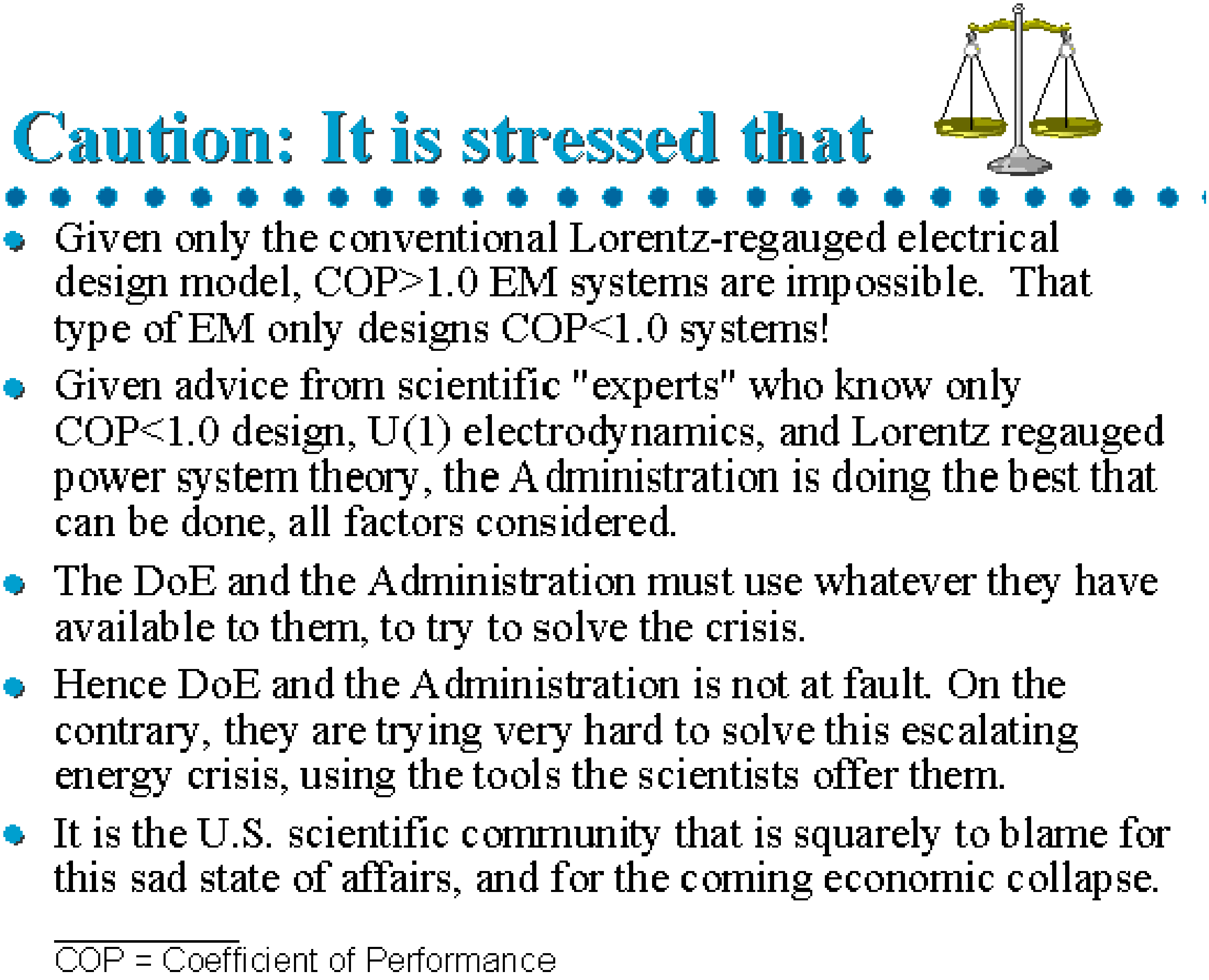

Caution: It is stressed that

- Given only the conventional Lorentz-regauged electrical design model, COP>1.0 EM systems are impossible. That type of EM only designs COP<1.0 systems!
- Given advice from scientific "experts" who know only COP<1.0 design, U(1) electrodynamics, and Lorentz regauged power system theory, the Administration is doing the best that can be done, all factors considered.
- The DoE and the Administration must use whatever they have available to them, to try to solve the crisis.
- Hence DoE and the Administration is not at fault. On the contrary, they are trying very hard to solve this escalating energy crisis, using the tools the scientists offer them.
- It is the U.S. scientific community that is squarely to blame for this sad state of affairs, and for the coming economic collapse.

COP = Coefficient of Performance

Published by Lotus® Freelance Graphics®

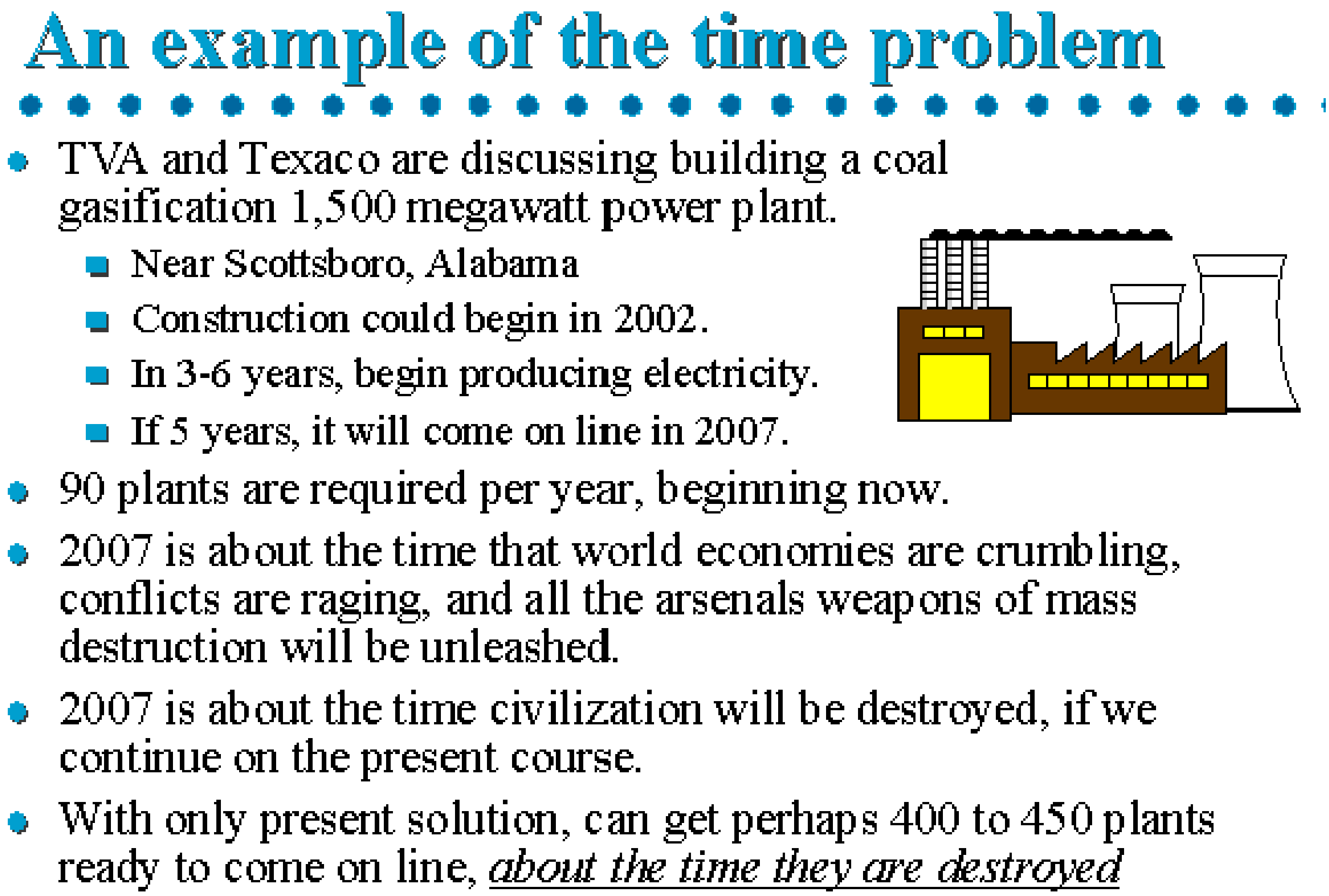

An example of the time problem

- TVA and Texaco are discussing building a coal gasification 1,500 megawatt power plant.
  - Near Scottsboro, Alabama
  - Construction could begin in 2002.
  - In 3-6 years, begin producing electricity.
  - If 5 years, it will come on line in 2007.
- 90 plants are required per year, beginning now.
- 2007 is about the time that world economies are crumbling, conflicts are raging, and all the arsenals weapons of mass destruction will be unleashed.
- 2007 is about the time civilization will be destroyed, if we continue on the present course.
- With only present solution, can get perhaps 400 to 450 plants ready to come on line, *about the time they are destroyed*

Published by Lotus® Freelance Graphics®

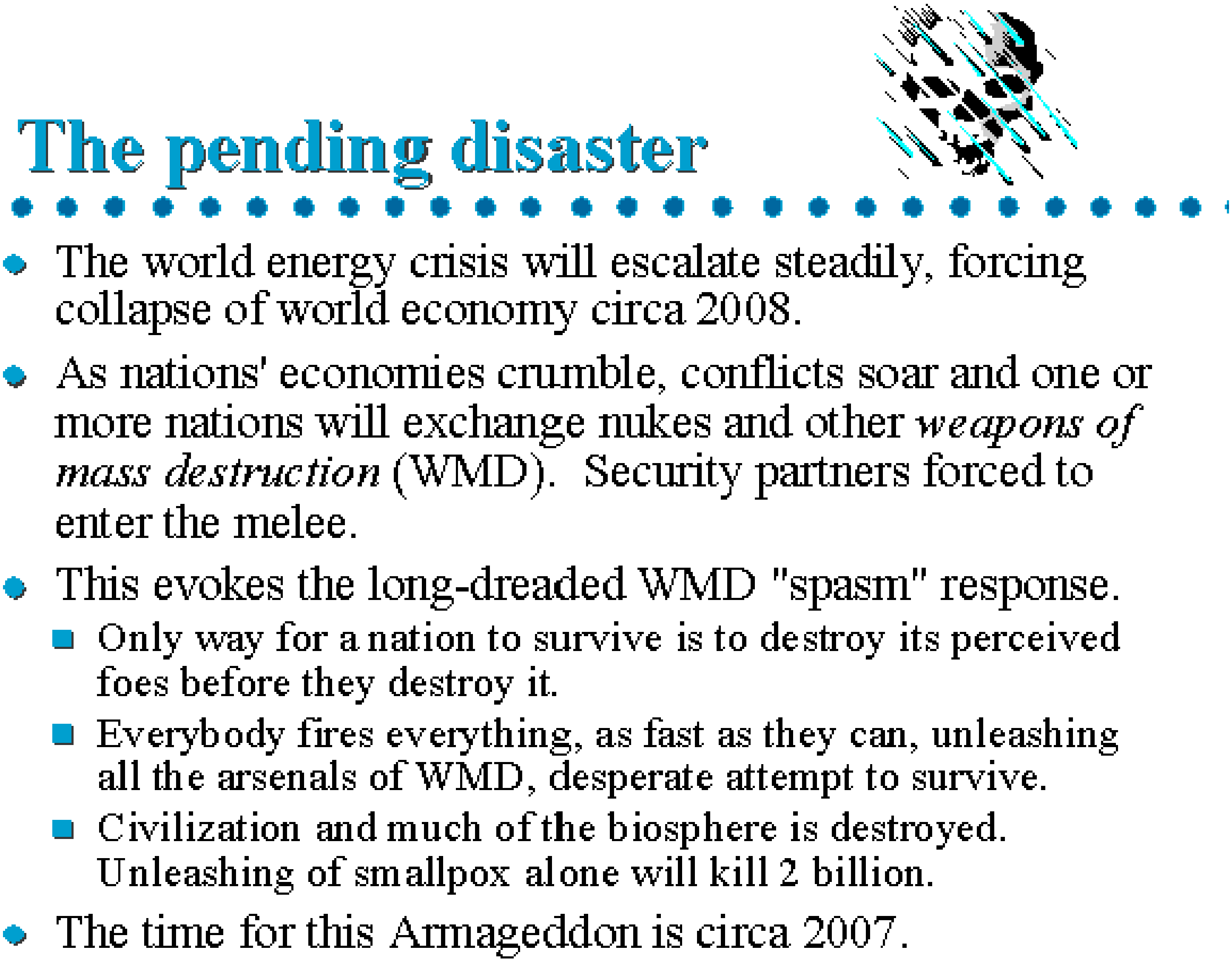

The pending disaster

- The world energy crisis will escalate steadily, forcing collapse of world economy circa 2008.
- As nations' economies crumble, conflicts soar and one or more nations will exchange nukes and other *weapons of mass destruction* (WMD). Security partners forced to enter the melee.
- This evokes the long-dreaded WMD "spasm" response.
    - Only way for a nation to survive is to destroy its perceived foes before they destroy it.
    - Everybody fires everything, as fast as they can, unleashing all the arsenals of WMD, desperate attempt to survive.
    - Civilization and much of the biosphere is destroyed. Unleashing of smallpox alone will kill 2 billion.
- The time for this Armageddon is circa 2007.

Published by Lotus® Freelance Graphics®

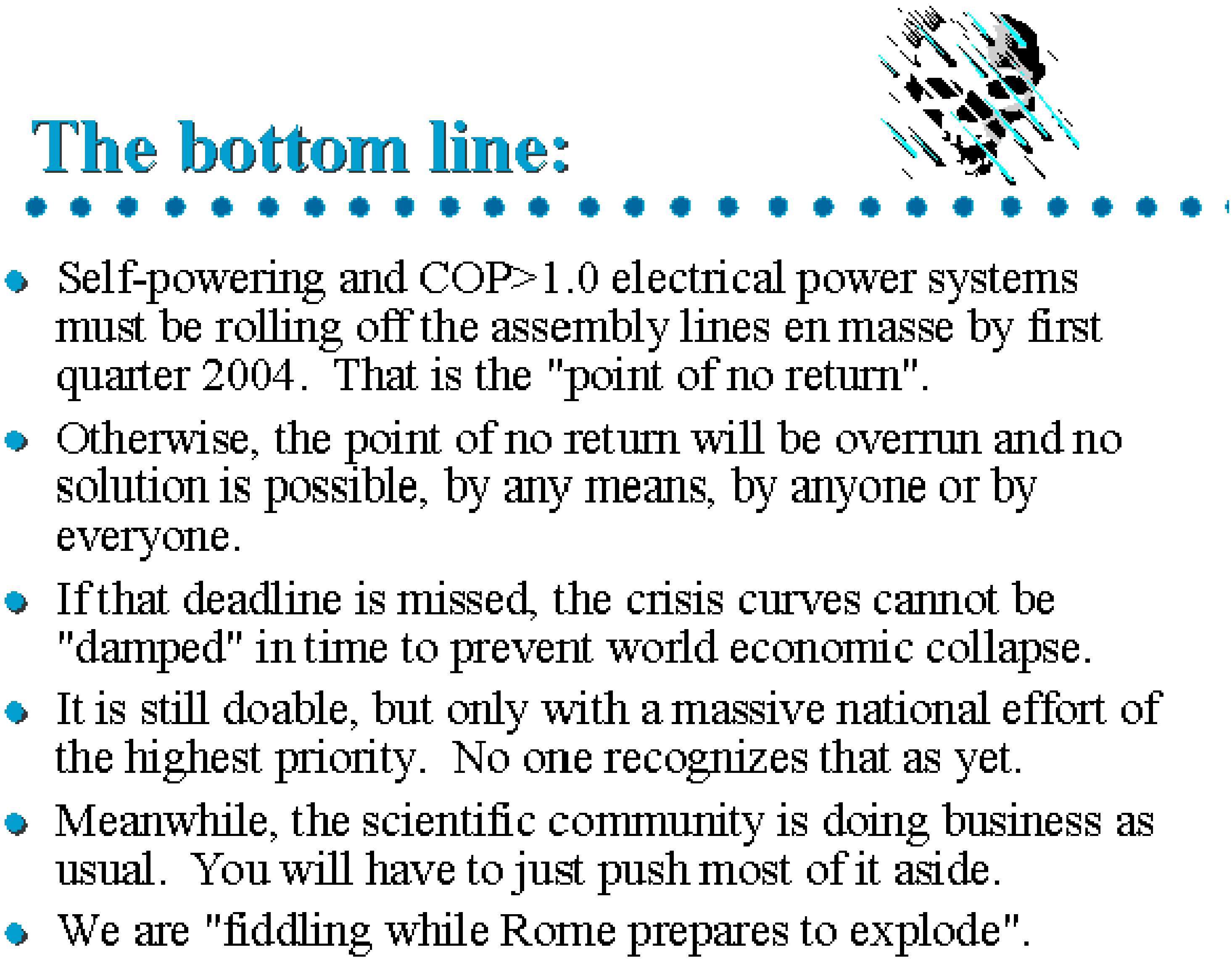

The bottom line:

- Self-powering and COP>1.0 electrical power systems must be rolling off the assembly lines en masse by first quarter 2004. That is the "point of no return".
- Otherwise, the point of no return will be overrun and no solution is possible, by any means, by anyone or by everyone.
- If that deadline is missed, the crisis curves cannot be "damped" in time to prevent world economic collapse.
- It is still doable, but only with a massive national effort of the highest priority. No one recognizes that as yet.
- Meanwhile, the scientific community is doing business as usual. You will have to just push most of it aside.
- We are "fiddling while Rome prepares to explode".

Published by Lotus® Freelance Graphics®

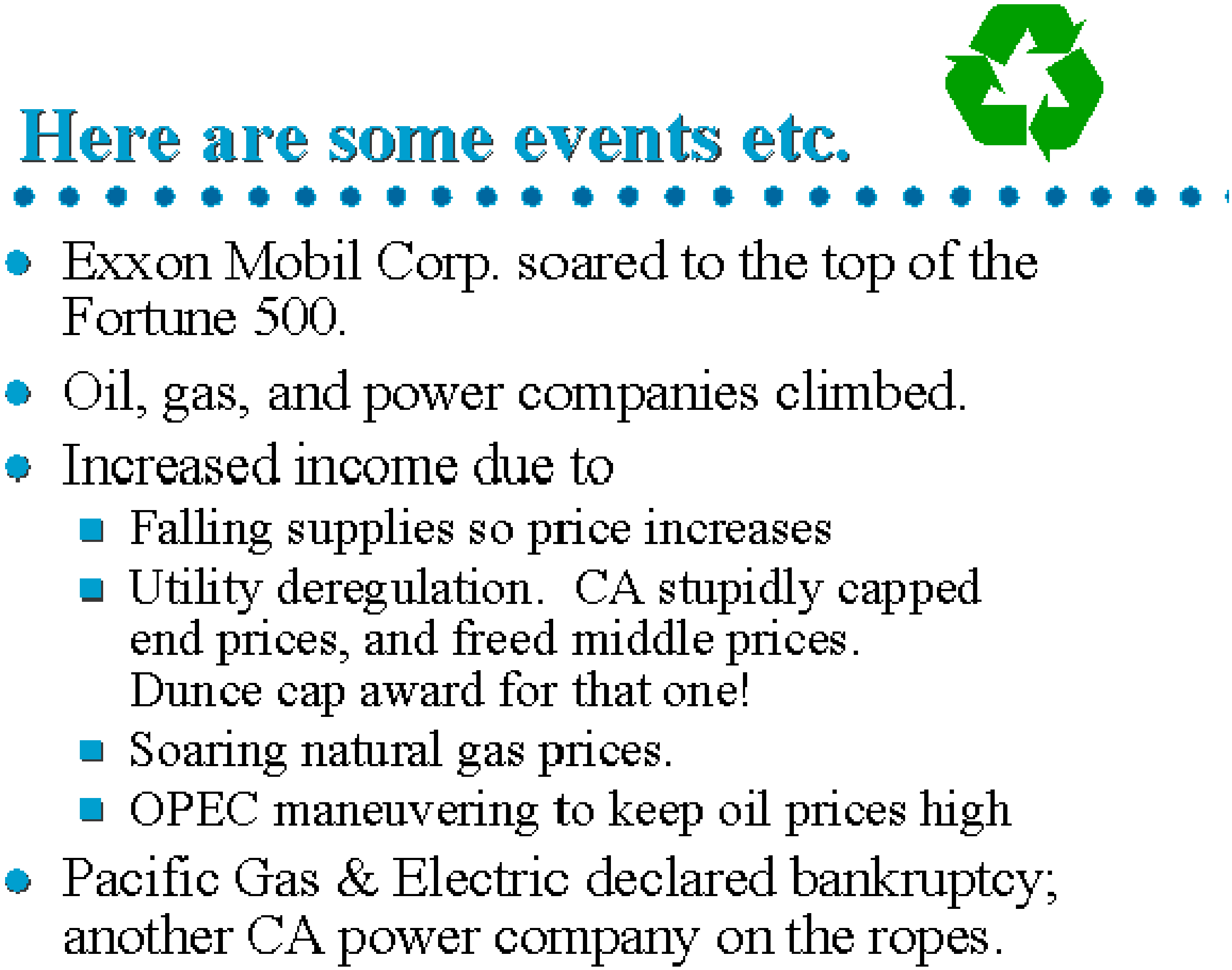

Here are some events etc.

- Exxon Mobil Corp. soared to the top of the Fortune 500.
- Oil, gas, and power companies climbed.
- Increased income due to
  - Falling supplies so price increases
  - Utility deregulation. CA stupidly capped end prices, and freed middle prices. Dunce cap award for that one!
  - Soaring natural gas prices.
  - OPEC maneuvering to keep oil prices high
- Pacific Gas & Electric declared bankruptcy; another CA power company on the ropes.

Published by Lotus® Freelance Graphics®

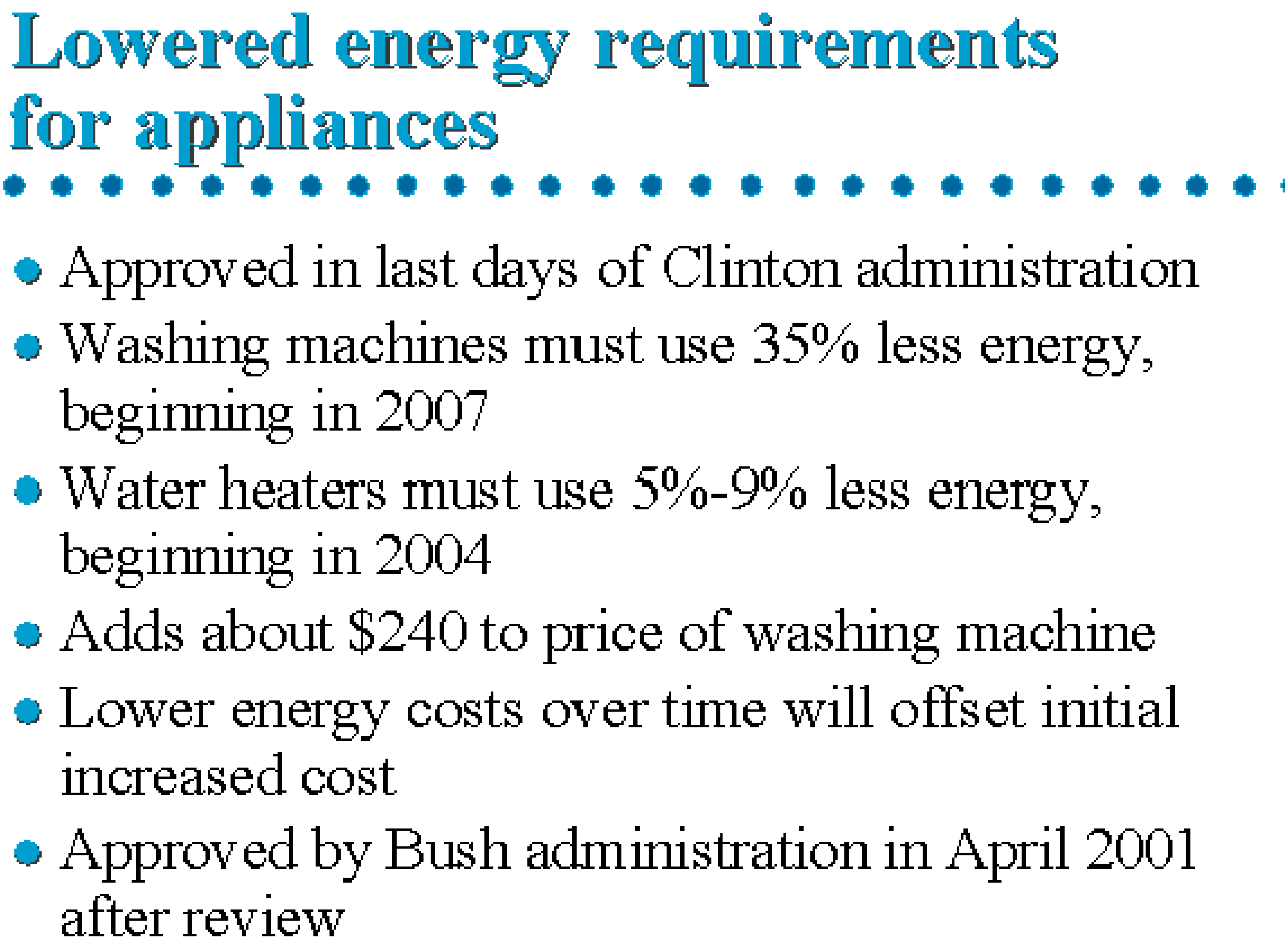

Lowered energy requirements for appliances

- Approved in last days of Clinton administration
- Washing machines must use 35% less energy, beginning in 2007
- Water heaters must use 5%-9% less energy, beginning in 2004
- Adds about $240 to price of washing machine
- Lower energy costs over time will offset initial increased cost
- Approved by Bush administration in April 2001 after review

Published by Lotus® Freelance Graphics®

Easing air conditioner and heat pump standards

- Bush administration eased Clinton's new efficiency standards
  - Still require air conditioners and heat pumps to use 20% less energy than most current models, beginning in 2006
  - Clinton standard was 30% less
- All central home air conditioning units must meet minimum seasonal efficiency ratio (SEER) of 12, compared to present 10
- Replaces Clinton SEER standard of 12
- In California, easing these standards requires two more power plants

Published by Lotus® Freelance Graphics®

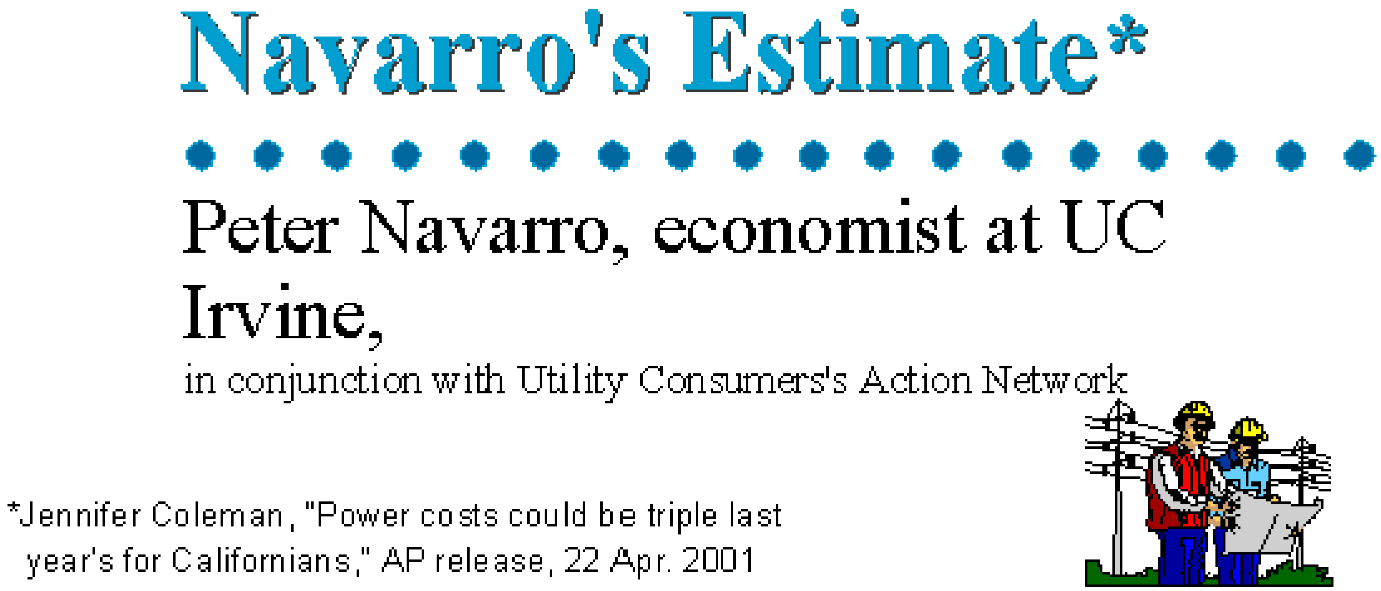

Navarro's Estimate*

Peter Navarro, economist at UC Irvine,

in conjunction with Utility Consumers's Action Network

*Jennifer Coleman, "Power costs could be triple last year's for Californians," AP release, 22 Apr. 2001

Published by Lotus® Freelance Graphics®

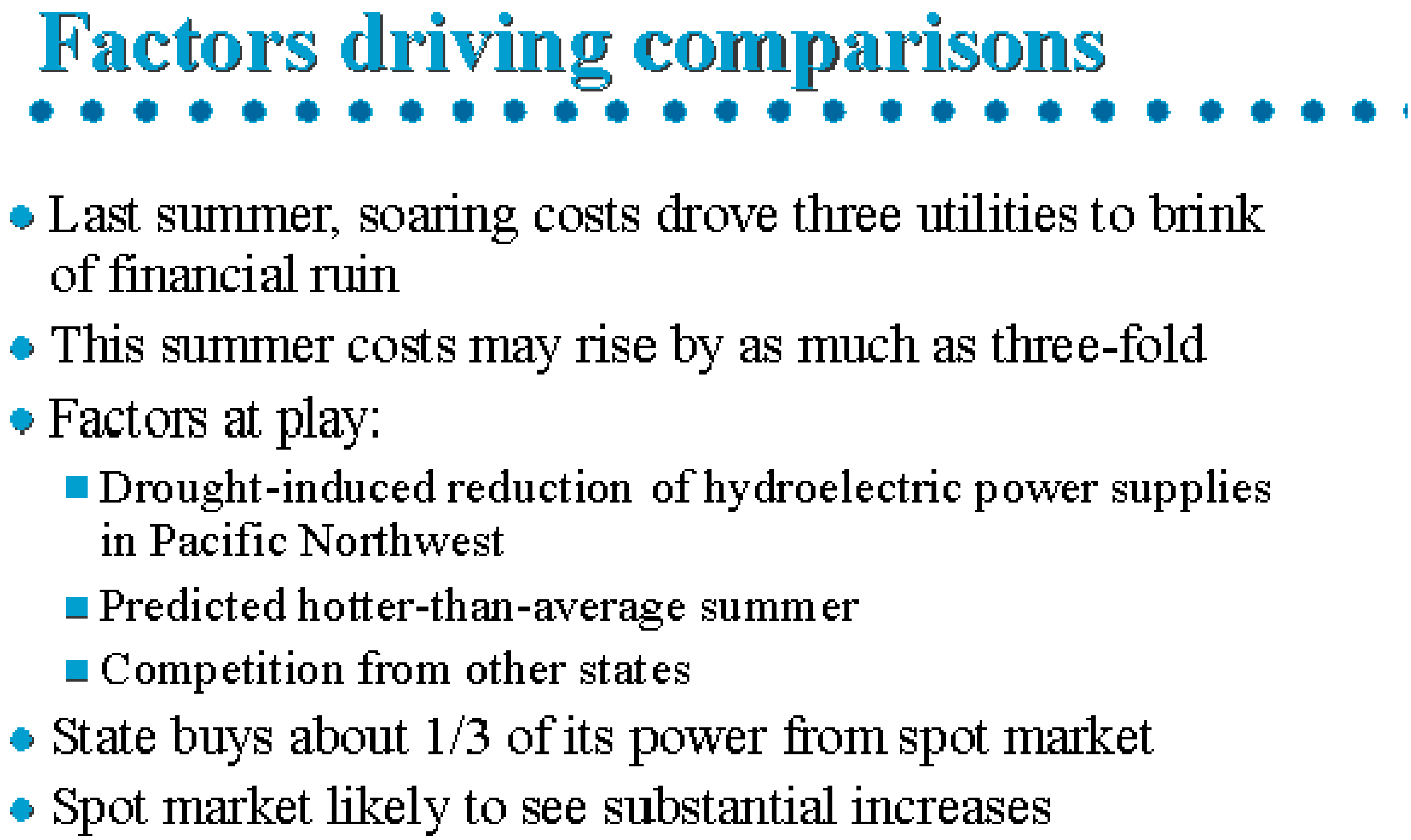

Factors driving comparisons

- Last summer, soaring costs drove three utilities to brink of financial ruin
- This summer costs may rise by as much as three-fold
- Factors at play:
    - Drought-induced reduction of hydroelectric power supplies in Pacific Northwest
    - Predicted hotter-than-average summer
    - Competition from other states
- State buys about 1/3 of its power from spot market
- Spot market likely to see substantial increases

Published by Lotus® Freelance Graphics®

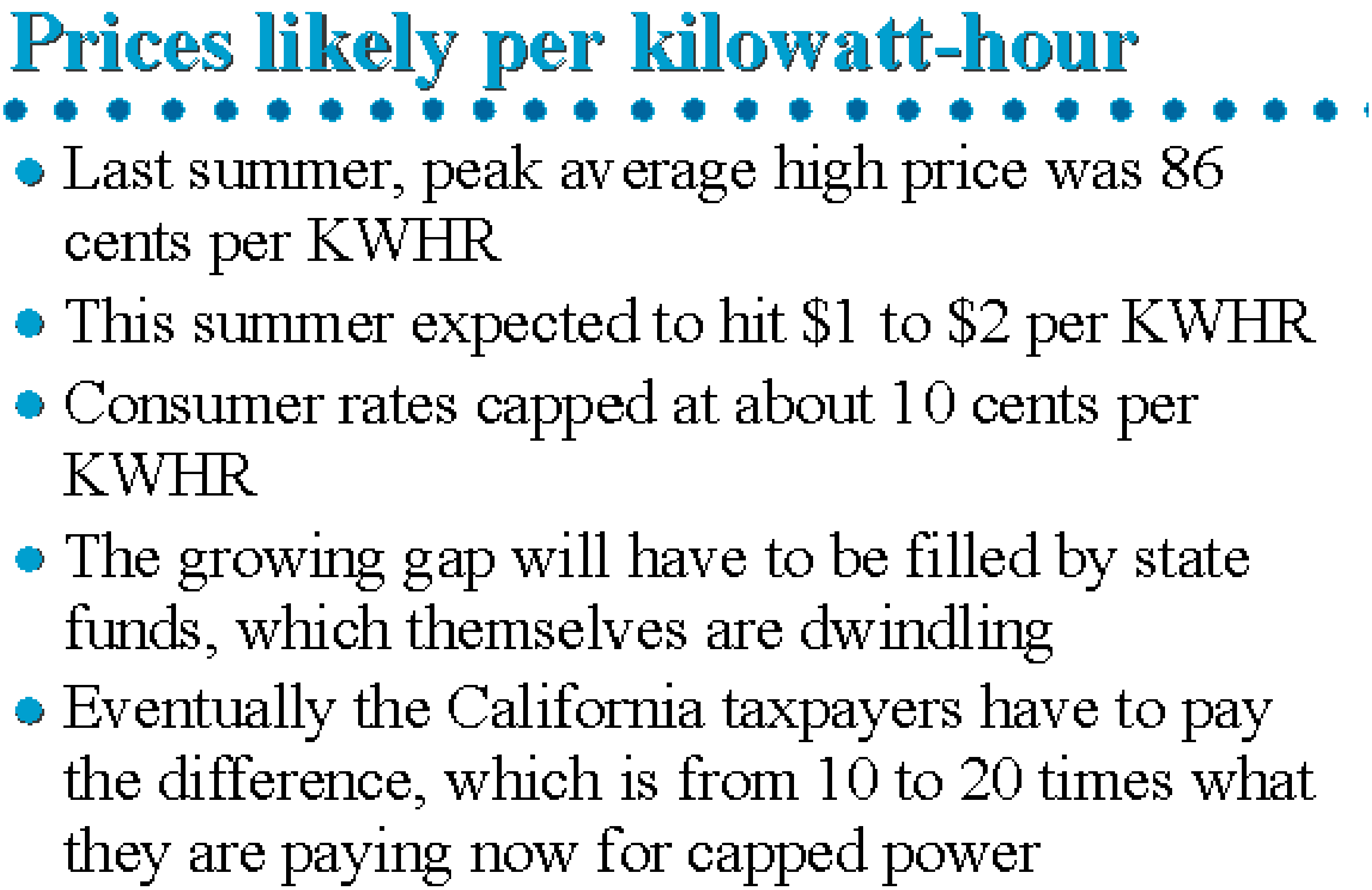

Prices likely per kilowatt-hour

- Last summer, peak average high price was 86 cents per KWHR
- This summer expected to hit $1 to $2 per KWHR
- Consumer rates capped at about 10 cents per KWHR
- The growing gap will have to be filled by state funds, which themselves are dwindling
- Eventually the California taxpayers have to pay the difference, which is from 10 to 20 times what they are paying now for capped power

Published by Lotus® Freelance Graphics®

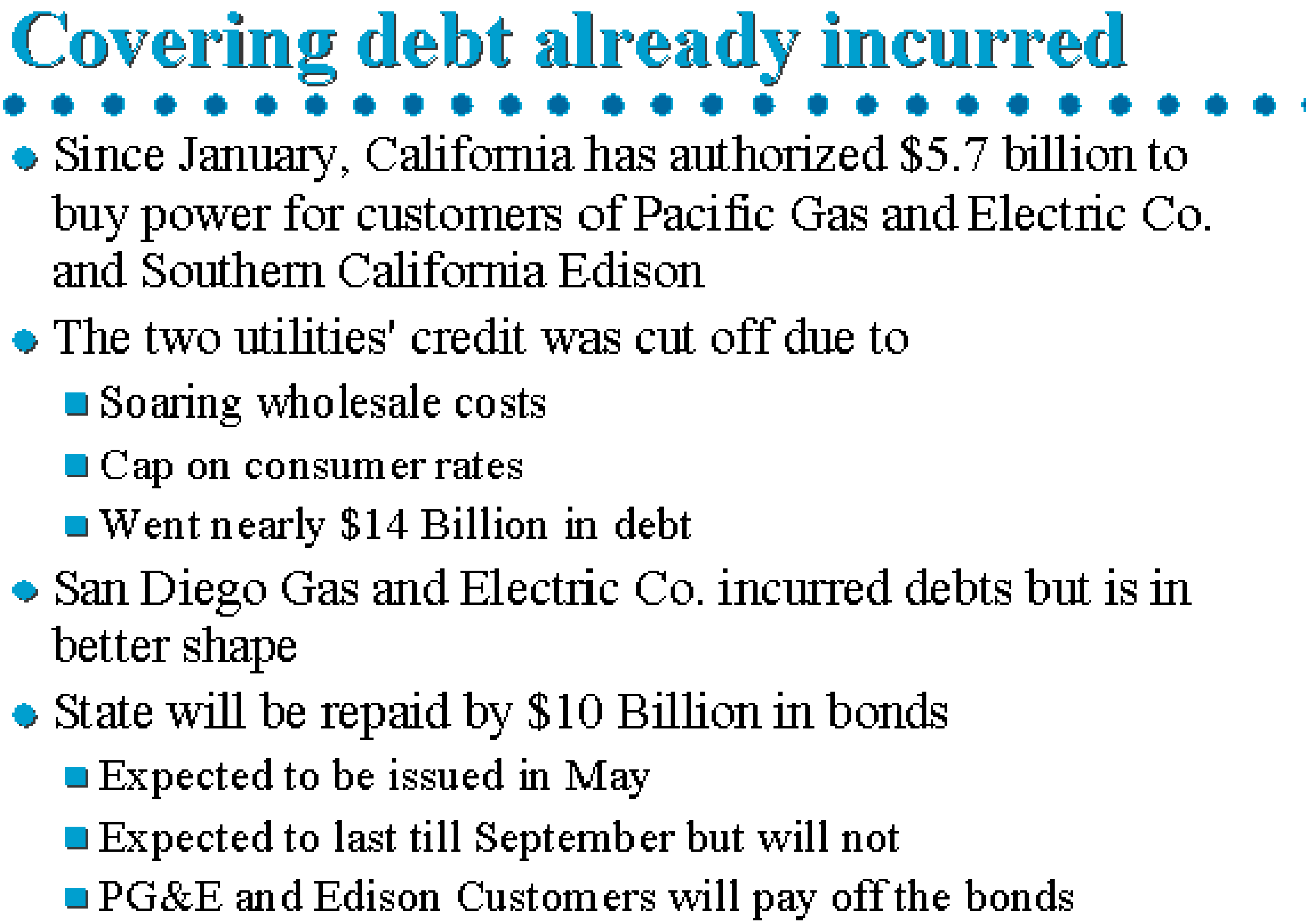

Covering debt already incurred

- Since January, California has authorized $5.7 billion to buy power for customers of Pacific Gas and Electric Co. and Southern California Edison
- The two utilities' credit was cut off due to
  - Soaring wholesale costs
  - Cap on consumer rates
  - Went nearly $14 Billion in debt
- San Diego Gas and Electric Co. incurred debts but is in better shape
- State will be repaid by $10 Billion in bonds
  - Expected to be issued in May
  - Expected to last till September but will not
  - PG&E and Edison Customers will pay off the bonds

Published by Lotus® Freelance Graphics®

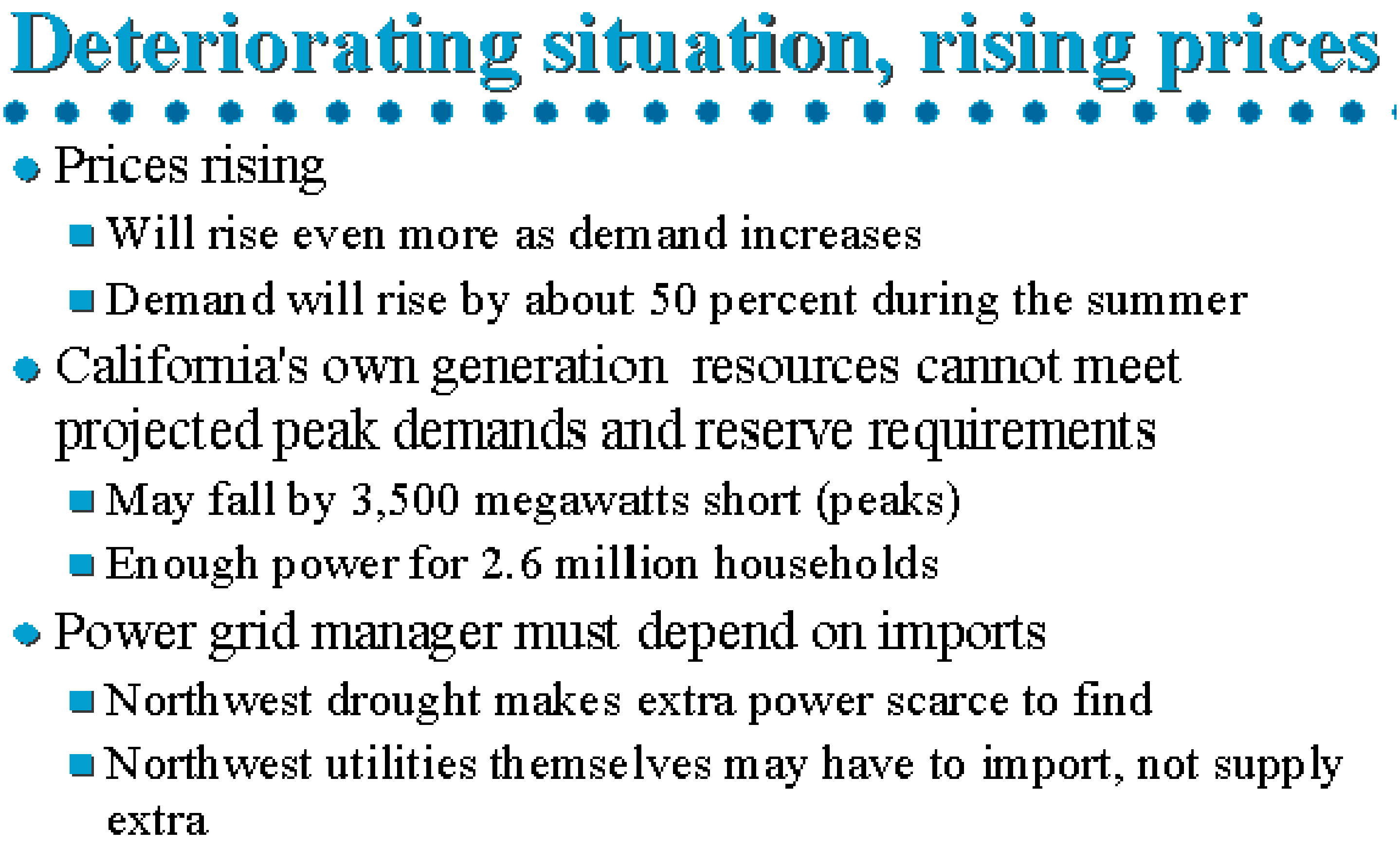

Deteriorating situation, rising prices

- Prices rising
  - Will rise even more as demand increases
  - Demand will rise by about 50 percent during the summer
- California's own generation resources cannot meet projected peak demands and reserve requirements
  - May fall by 3,500 megawatts short (peaks)
  - Enough power for 2.6 million households
- Power grid manager must depend on imports
  - Northwest drought makes extra power scarce to find
  - Northwest utilities themselves may have to import, not supply extra

---

Published by Lotus® Freelance Graphics®

Other Factors

Credit, payback, scarcities

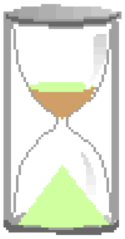

Published by Lotus® Freelance Graphics®

Paying the piper: California's bond rating reduced

- Standard & Poor dropped California's bond rating two notches, from AA to A+
    - Financial drain from continuing energy crisis
    - Mounting uncertainty in costs to state of current electrical power crisis
    - Likely long term detrimental effect on state's economy
    - State's ability to pay its debt has reduced
- Rating not reduced further because of
    - State's diverse economy
    - Proposed revenue bonds to reimburse state treasury
- Can reduce further if California does not make $10 Billion bond issue and pay off its energy-related debts

Published by Lotus® Freelance Graphics®

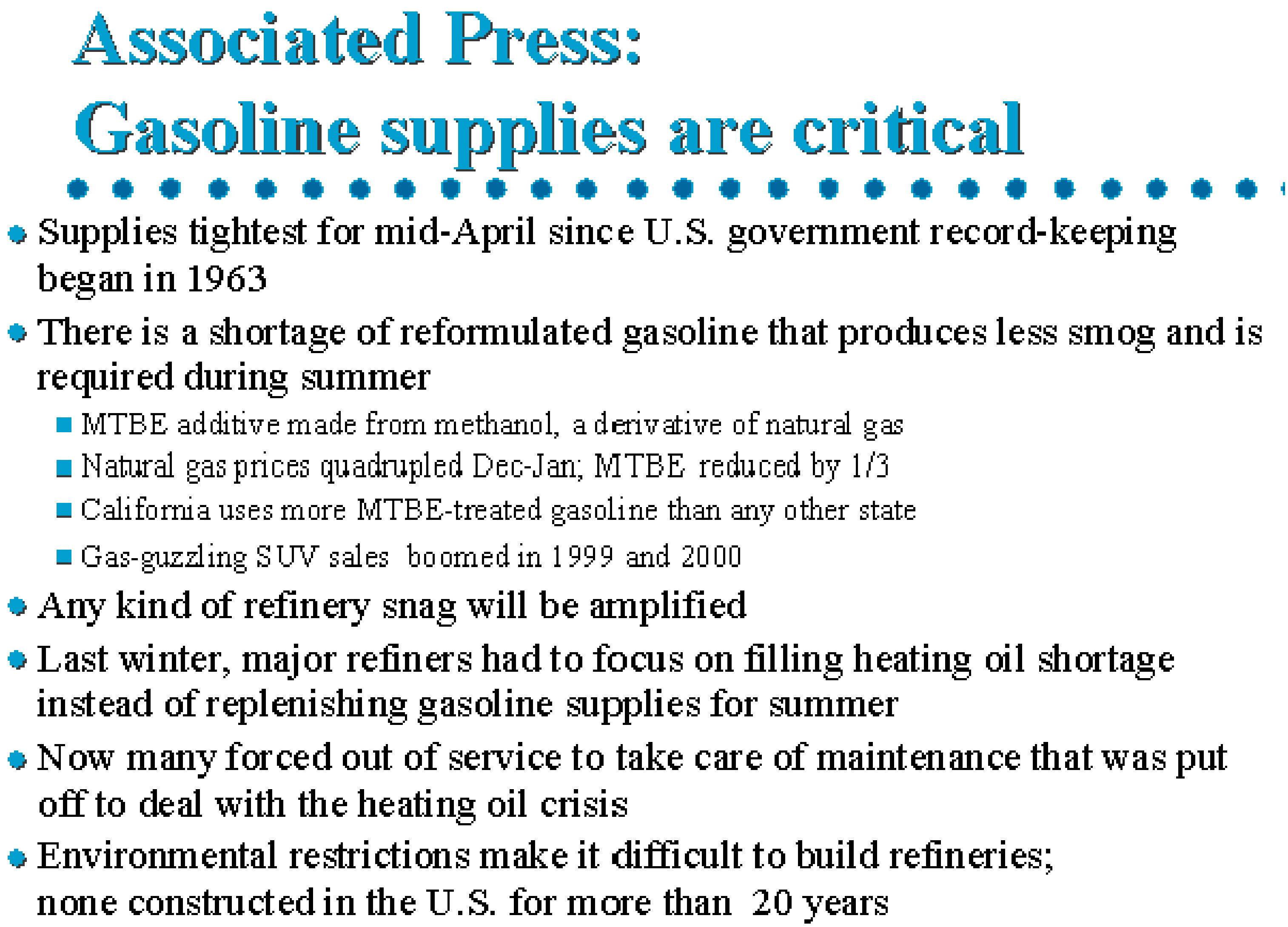

Associated Press: Gasoline supplies are critical

- Supplies tightest for mid-April since U.S. government record-keeping began in 1963
- There is a shortage of reformulated gasoline that produces less smog and is required during summer
  - MTBE additive made from methanol, a derivative of natural gas
  - Natural gas prices quadrupled Dec-Jan; MTBE reduced by 1/3
  - California uses more MTBE-treated gasoline than any other state
  - Gas-guzzling SUV sales boomed in 1999 and 2000
- Any kind of refinery snag will be amplified
- Last winter, major refiners had to focus on filling heating oil shortage instead of replenishing gasoline supplies for summer
- Now many forced out of service to take care of maintenance that was put off to deal with the heating oil crisis
- Environmental restrictions make it difficult to build refineries; none constructed in the U.S. for more than 20 years

Published by Lotus® Freelance Graphics®

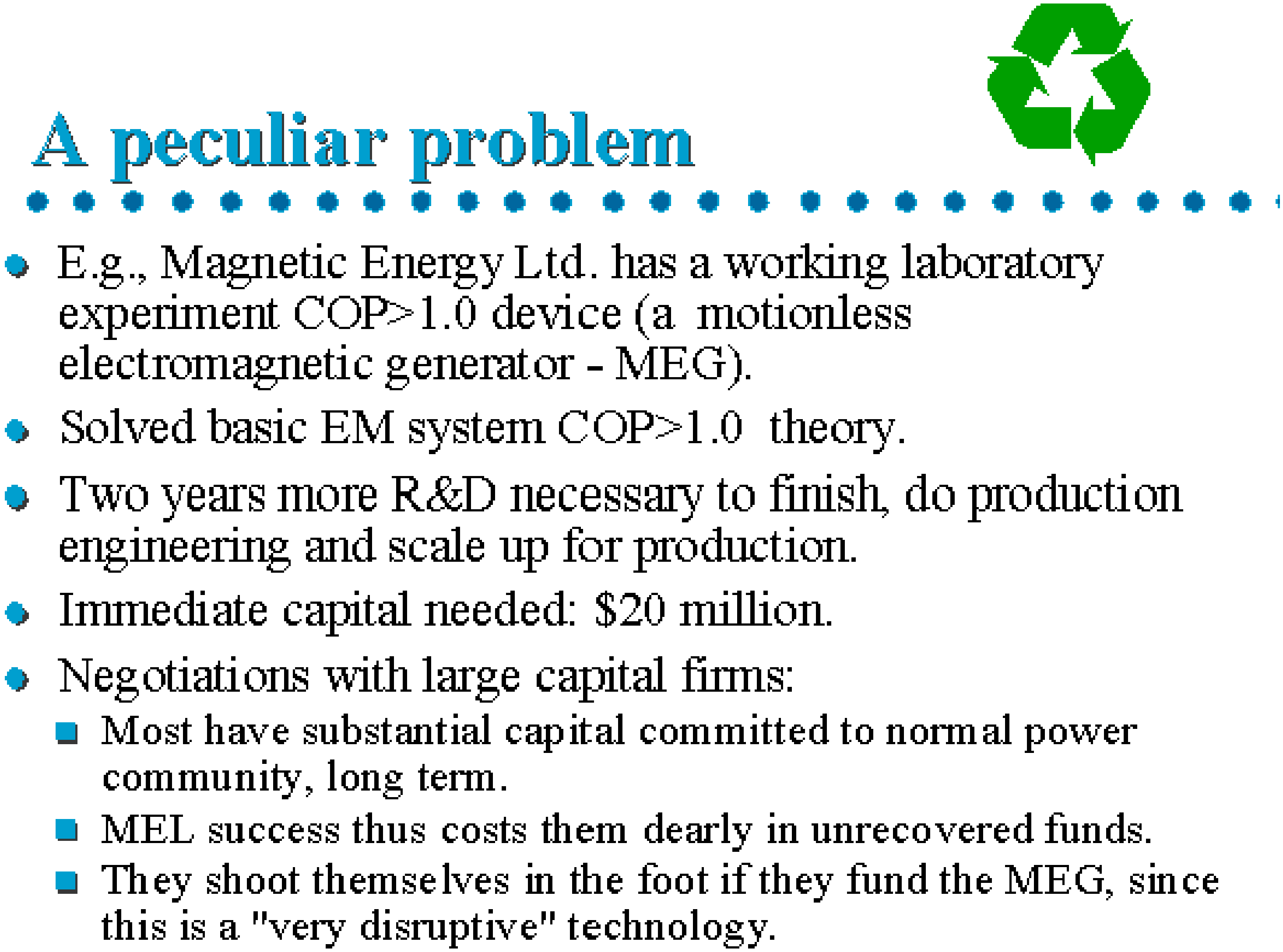

A peculiar problem

- E.g., Magnetic Energy Ltd. has a working laboratory experiment COP>1.0 device (a motionless electromagnetic generator - MEG).
- Solved basic EM system COP>1.0 theory.
- Two years more R&D necessary to finish, do production engineering and scale up for production.
- Immediate capital needed: $20 million.
- Negotiations with large capital firms:
  - Most have substantial capital committed to normal power community, long term.
  - MEL success thus costs them dearly in unrecovered funds.
  - They shoot themselves in the foot if they fund the MEG, since this is a "very disruptive" technology.

Published by Lotus® Freelance Graphics®

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (1)

- The Aharonov-Bohm effect (ABE) is in over 2,000 published physics papers in the hard literature.
- Hundreds of experiments prove it. Physicists themselves would not believe it for 25 years after it was advanced and proven. Now it is universally accepted.
- Free evocation of ABE never appeared for use in power systems, without paying energy to evoke it, until we used ABE in the MEG and *freely* evoked it.
- If the magnetic field **B** of a source is locally confined in a closed path (as in a toroid), the outside local spacetime (ST) where the **B**-field was withdrawn is still curved. Curved ST is energetic. The 3-space energy has been transformed into curl-free magnetic vector potential **A**. But one pays for that **A** around a toroid, since one has to "feed" the toroid.

Published by Lotus® Freelance Graphics®

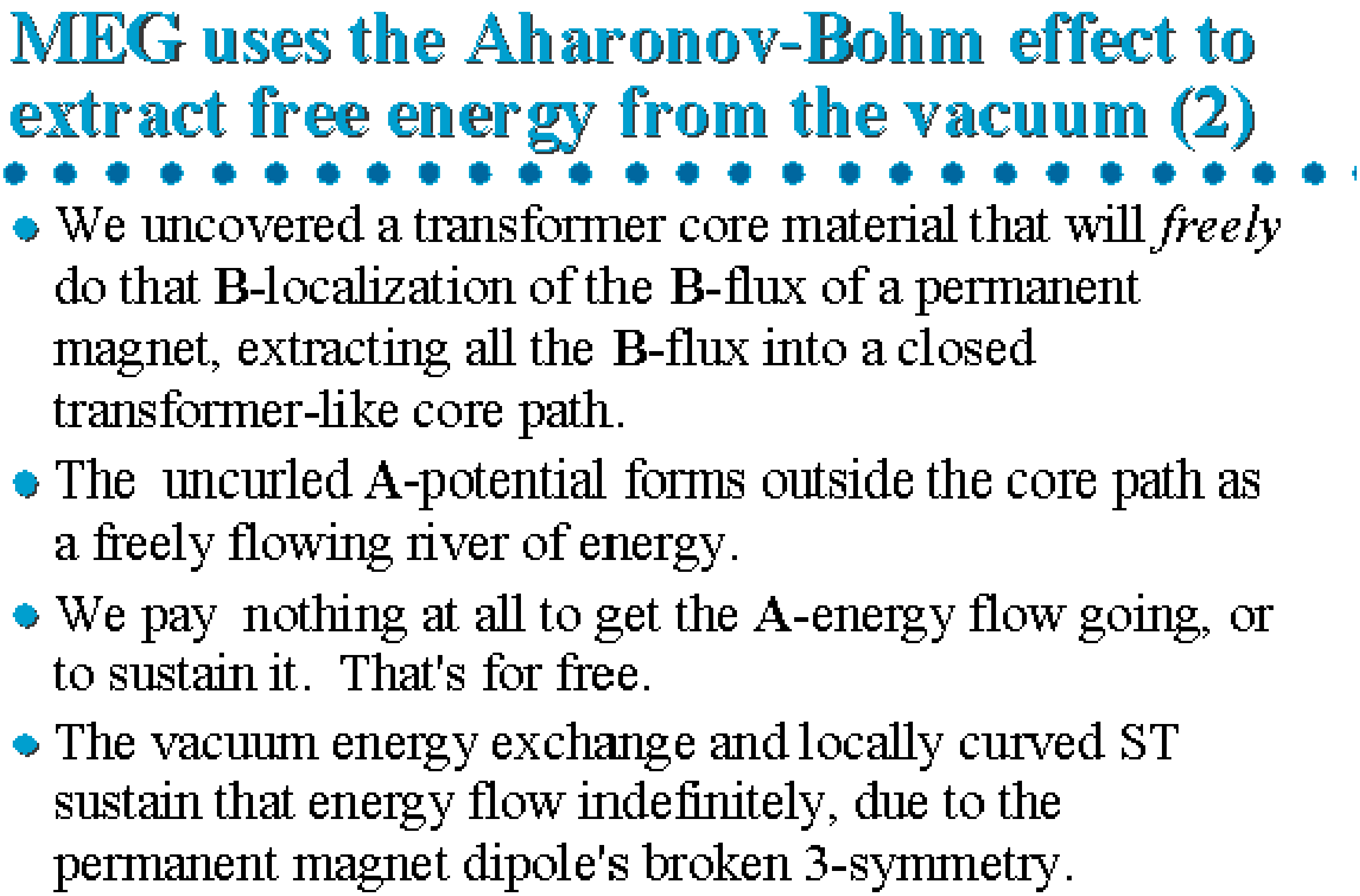

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (2)

- We uncovered a transformer core material that will *freely* do that **B**-localization of the **B**-flux of a permanent magnet, extracting all the **B**-flux into a closed transformer-like core path.
- The uncurled **A**-potential forms outside the core path as a freely flowing river of energy.
- We pay nothing at all to get the **A**-energy flow going, or to sustain it. That's for free.
- The vacuum energy exchange and locally curved ST sustain that energy flow indefinitely, due to the permanent magnet dipole's broken 3-symmetry.

Published by Lotus® Freelance Graphics®

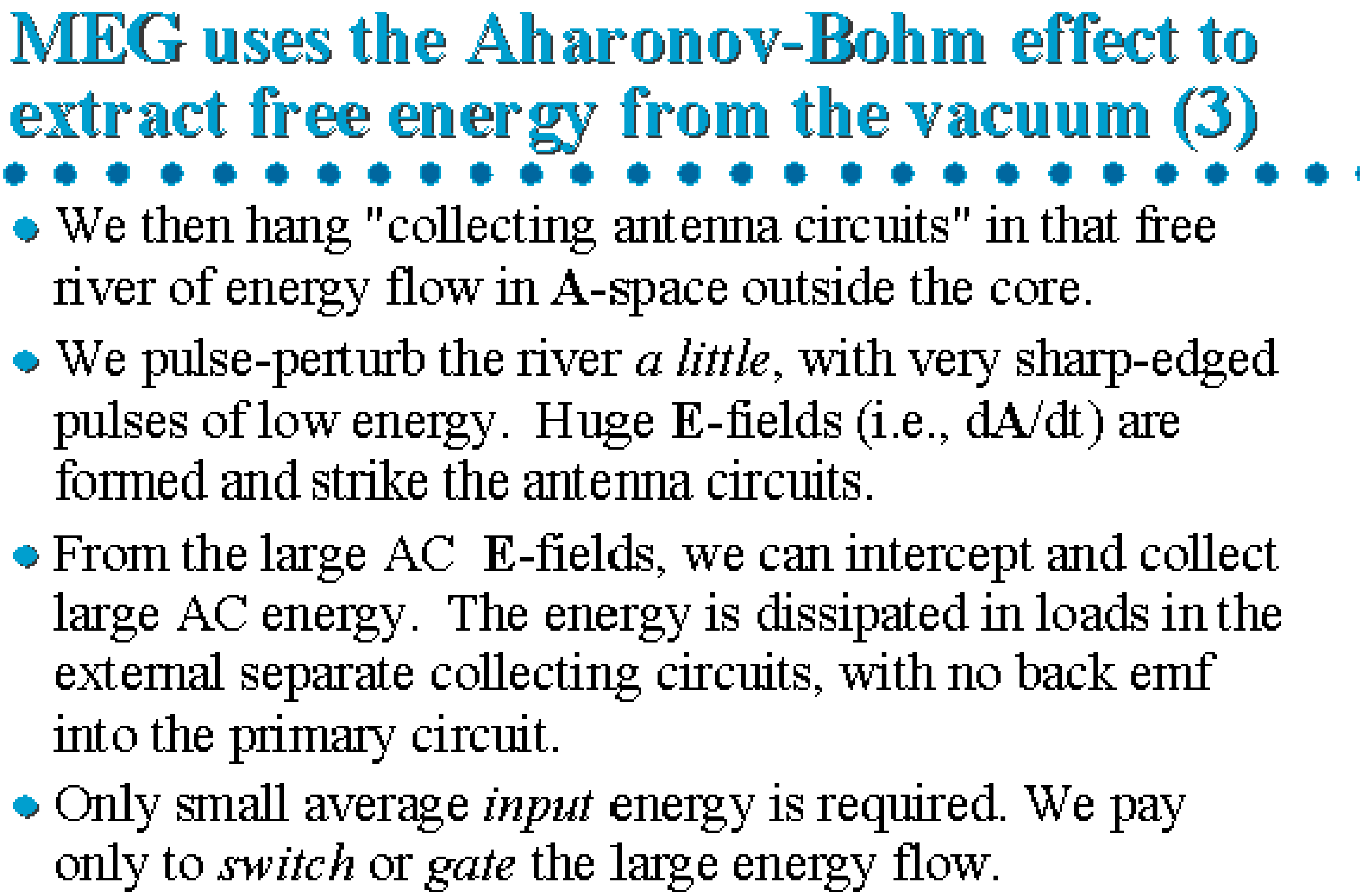

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (3)

- We then hang "collecting antenna circuits" in that free river of energy flow in **A**-space outside the core.
- We pulse-perturb the river *a little*, with very sharp-edged pulses of low energy. Huge **E**-fields (i.e., d**A**/dt) are formed and strike the antenna circuits.
- From the large AC **E**-fields, we can intercept and collect large AC energy. The energy is dissipated in loads in the external separate collecting circuits, with no back emf into the primary circuit.
- Only small average *input* energy is required. We pay only to *switch* or *gate* the large energy flow.

Published by Lotus® Freelance Graphics®

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (4)

- By analogy, imagine an old triode vacuum tube.
- Suppose we trick nature into *giving* us the cathode and electron energy flow, for free, really pouring energy out of there continuously.
- Then we only have to put in a grid to switch it, a plate to intercept and catch it, and a plate circuit to dissipate it in a load.
- In the MEG, the permanent magnet and the special core provide the cathode and free flow.
- We provide the "grid" gating, plate "catching," and "plate circuits" with loads that are powered.

Published by Lotus® Freelance Graphics®

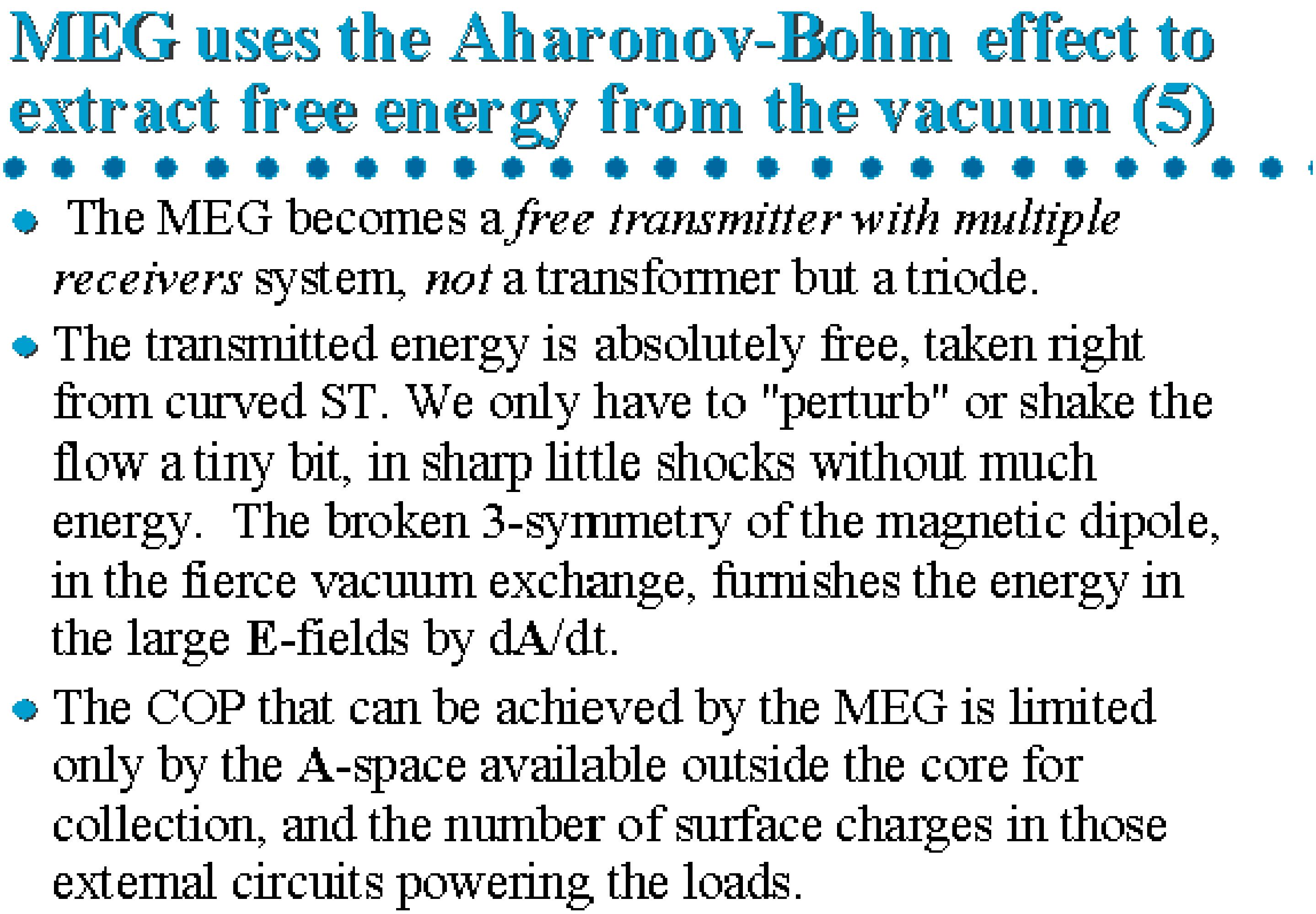

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (5)

- The MEG becomes a *free transmitter with multiple receivers* system, *not* a transformer but a triode.
- The transmitted energy is absolutely free, taken right from curved ST. We only have to "perturb" or shake the flow a tiny bit, in sharp little shocks without much energy. The broken 3-symmetry of the magnetic dipole, in the fierce vacuum exchange, furnishes the energy in the large **E**-fields by d**A**/dt.
- The COP that can be achieved by the MEG is limited only by the **A**-space available outside the core for collection, and the number of surface charges in those external circuits powering the loads.

Published by Lotus® Freelance Graphics®

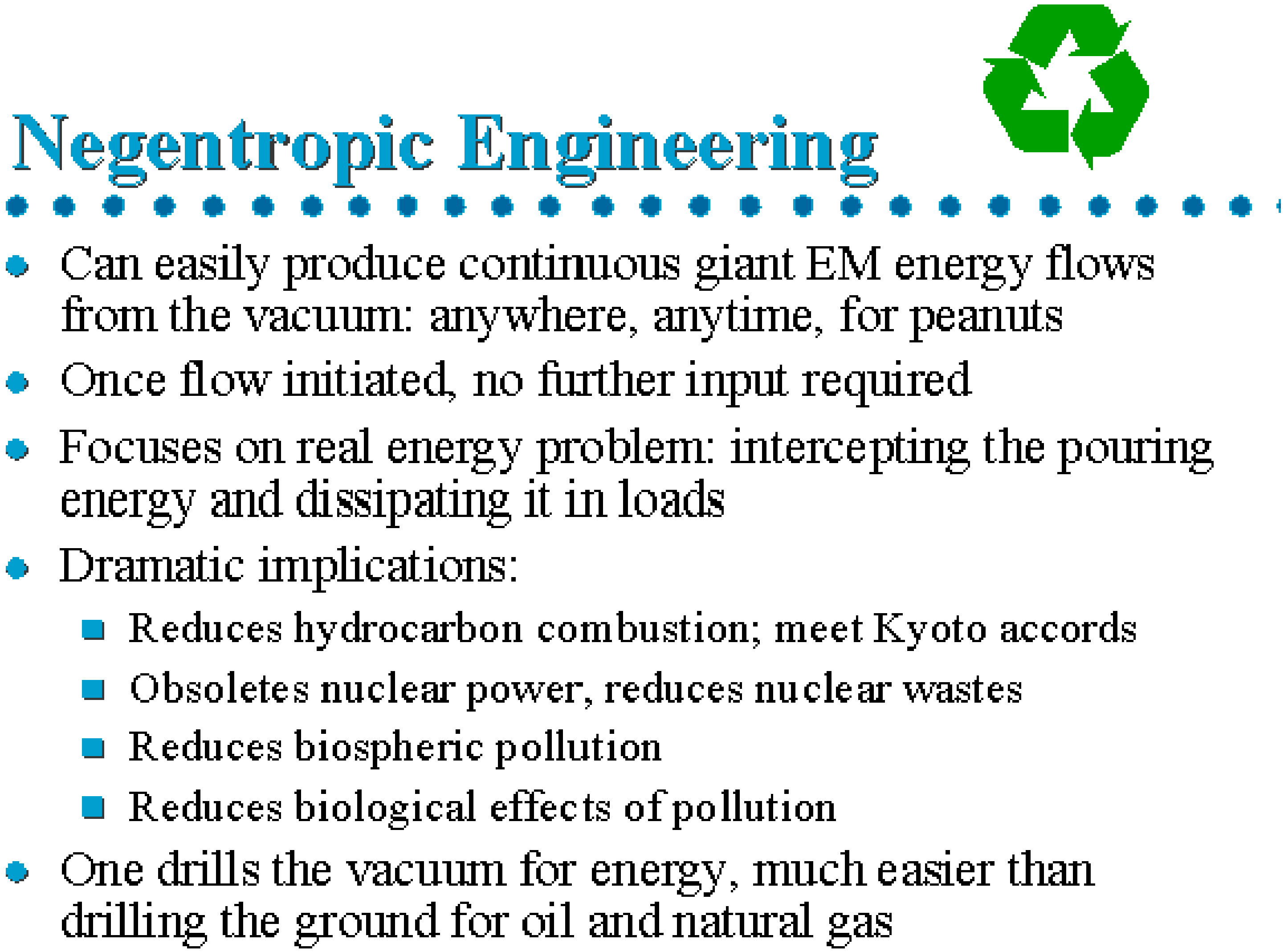

Negentropic Engineering

- Can easily produce continuous giant EM energy flows from the vacuum: anywhere, anytime, for peanuts
- Once flow initiated, no further input required
- Focuses on real energy problem: intercepting the pouring energy and dissipating it in loads
- Dramatic implications:
  - Reduces hydrocarbon combustion; meet Kyoto accords
  - Obsoletes nuclear power, reduces nuclear wastes
  - Reduces biospheric pollution
  - Reduces biological effects of pollution
- One drills the vacuum for energy, much easier than drilling the ground for oil and natural gas

Published by Lotus® Freelance Graphics®

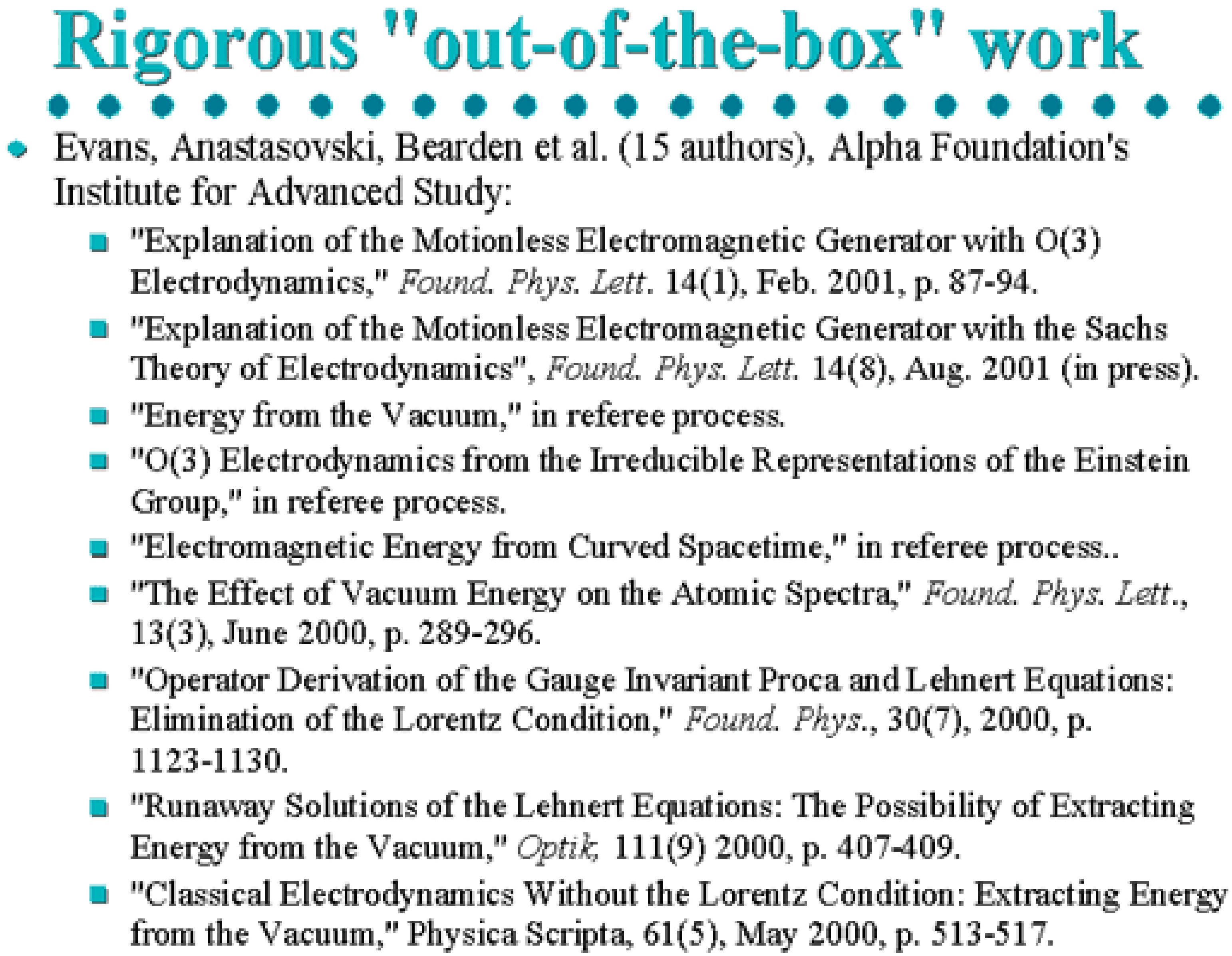

Rigorous "out-of-the-box" work

- Evans, Anastasovski, Bearden et al. (15 authors), Alpha Foundation's Institute for Advanced Study:
  - "Explanation of the Motionless Electromagnetic Generator with O(3) Electrodynamics," *Found. Phys. Lett*. 14(1), Feb. 2001, p. 87-94.
  - "Explanation of the Motionless Electromagnetic Generator with the Sachs Theory of Electrodynamics", *Found. Phys. Lett.* 14(8), Aug. 2001 (in press).
  - "Energy from the Vacuum," in referee process.
  - "O(3) Electrodynamics from the Irreducible Representations of the Einstein Group," in referee process.
  - "Electromagnetic Energy from Curved Spacetime," in referee process..
  - "The Effect of Vacuum Energy on the Atomic Spectra," *Found. Phys. Lett*., 13(3), June 2000, p. 289-296.
  - "Operator Derivation of the Gauge Invariant Proca and Lehnert Equations: Elimination of the Lorentz Condition," *Found. Phys*., 30(7), 2000, p. 1123-1130.
  - "Runaway Solutions of the Lehnert Equations: The Possibility of Extracting Energy from the Vacuum," *Optik*, 111(9) 2000, p. 407-409.
  - "Classical Electrodynamics Without the Lorentz Condition: Extracting Energy from the Vacuum," Physica Scripta, 61(5), May 2000, p. 513-517.

Published by Lotus® Freelance Graphics®

Rigorous "out-of-the-box" work (2):
(Bedini "negative resistor" in a battery)*

- A battery-powered circuit is not a single closed current loop; but has two largely separate current half-loops:
  - Internal ion current between plates.
  - External electron current from outside of one plate, through external circuit and load, to other plate.
- M/q ratios very different: Ion current more than 200,000 times heavier per coulomb than electron current.
- Can dephase the two currents, overpotentialize both.
- Battery can thus be hyper-recharged while external circuit simultaneously powered.
- Bedini has been building successful little prototypes for more than 20 years. Could dramatically improve electric automobiles.

* T.E. Bearden, "Bedini's Method for Forming Negative Resistors in Batteries," *Journal of New Energy*, 5(1), Summer 2000, p. 24-38. http://www.cheniere.org/techpapers/Bedini.pdf

Published by Lotus® Freelance Graphics®

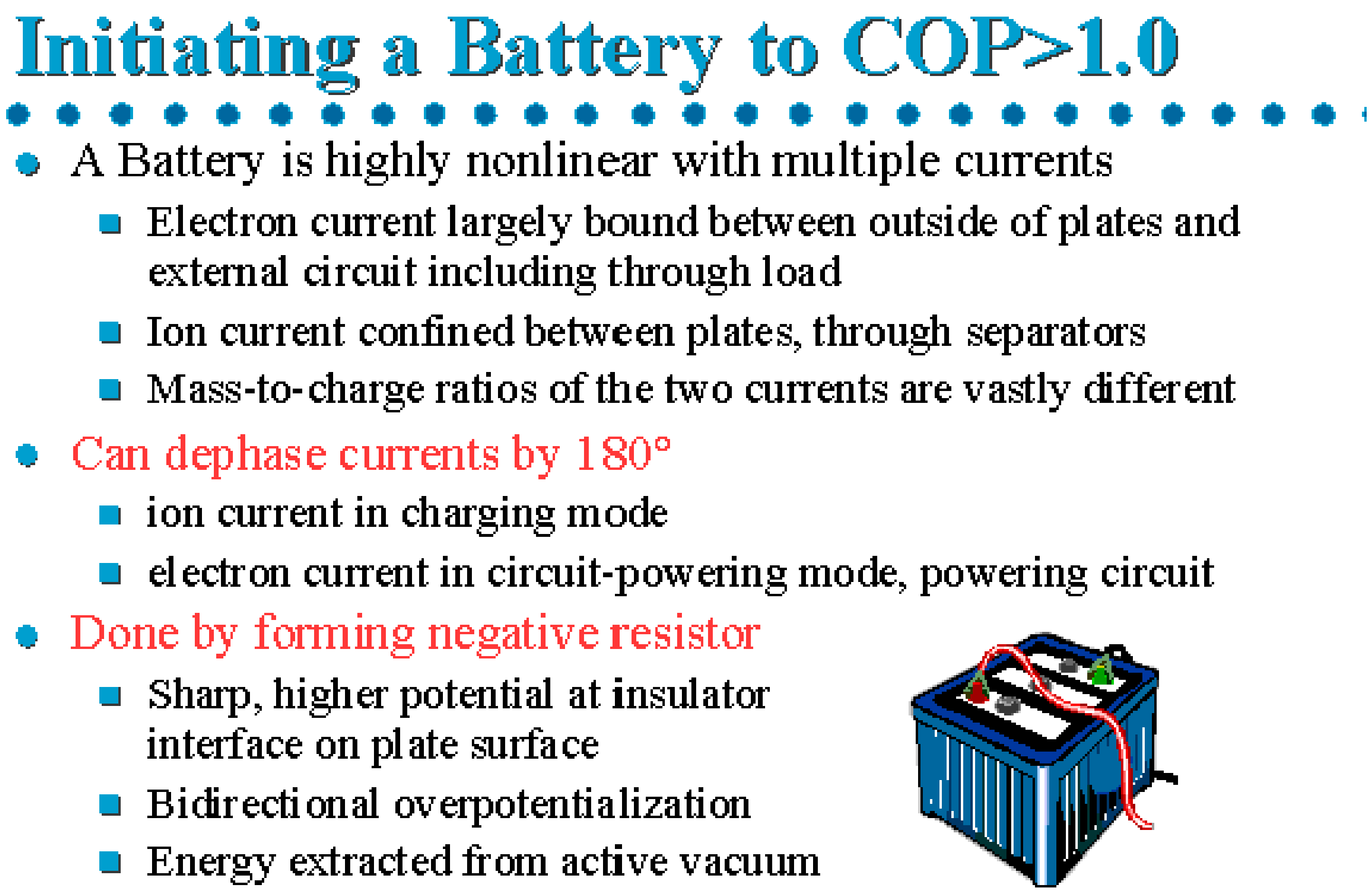

Initiating a Battery to COP>1.0

- A Battery is highly nonlinear with multiple currents
  - Electron current largely bound between outside of plates and external circuit including through load
  - Ion current confined between plates, through separators
  - Mass-to-charge ratios of the two currents are vastly different
- Can dephase currents by 180°
  - ion current in charging mode
  - electron current in circuit-powering mode, powering circuit
- Done by forming negative resistor
  - Sharp, higher potential at insulator interface on plate surface
  - Bidirectional overpotentialization
  - Energy extracted from active vacuum

Published by Lotus® Freelance Graphics®

Secretary alone cannot solve it

- Scientific community, universities, national labs, NSF, NAS, power companies doing business as usual
    - No really out-of-the-box energy research
    - Do not realize all electrical systems and power lines are powered by energy extracted from the vacuum
    - Not correcting flawed electrodynamics, uninterested
    - Not funding any vacuum energy extraction projects
- New policy cannot quickly eliminate the long delay time till new electrical power plants come on line
    - Need 60 to 90 powerplants per year; first one can be operational only after several years even with crash program
    - Skyrocketing energy, trucking and other transport costs will drive the nation into financial disaster during that delay
- Only doable solution: vacuum-energy systems

Published by Lotus® Freelance Graphics®

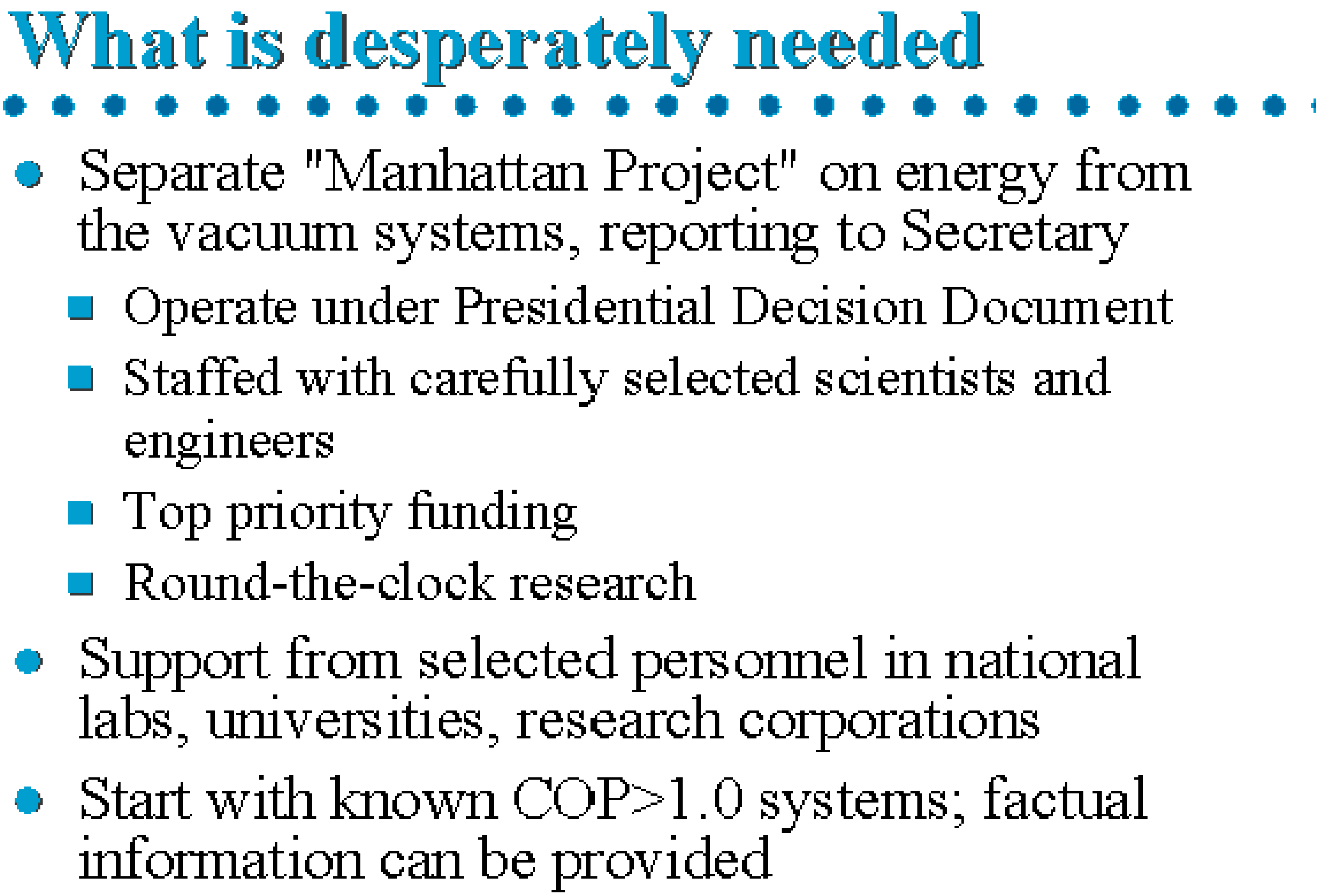

What is desperately needed

- Separate "Manhattan Project" on energy from the vacuum systems, reporting to Secretary
  - Operate under Presidential Decision Document
  - Staffed with carefully selected scientists and engineers
  - Top priority funding
  - Round-the-clock research
- Support from selected personnel in national labs, universities, research corporations
- Start with known COP>1.0 systems; factual information can be provided

Published by Lotus® Freelance Graphics®

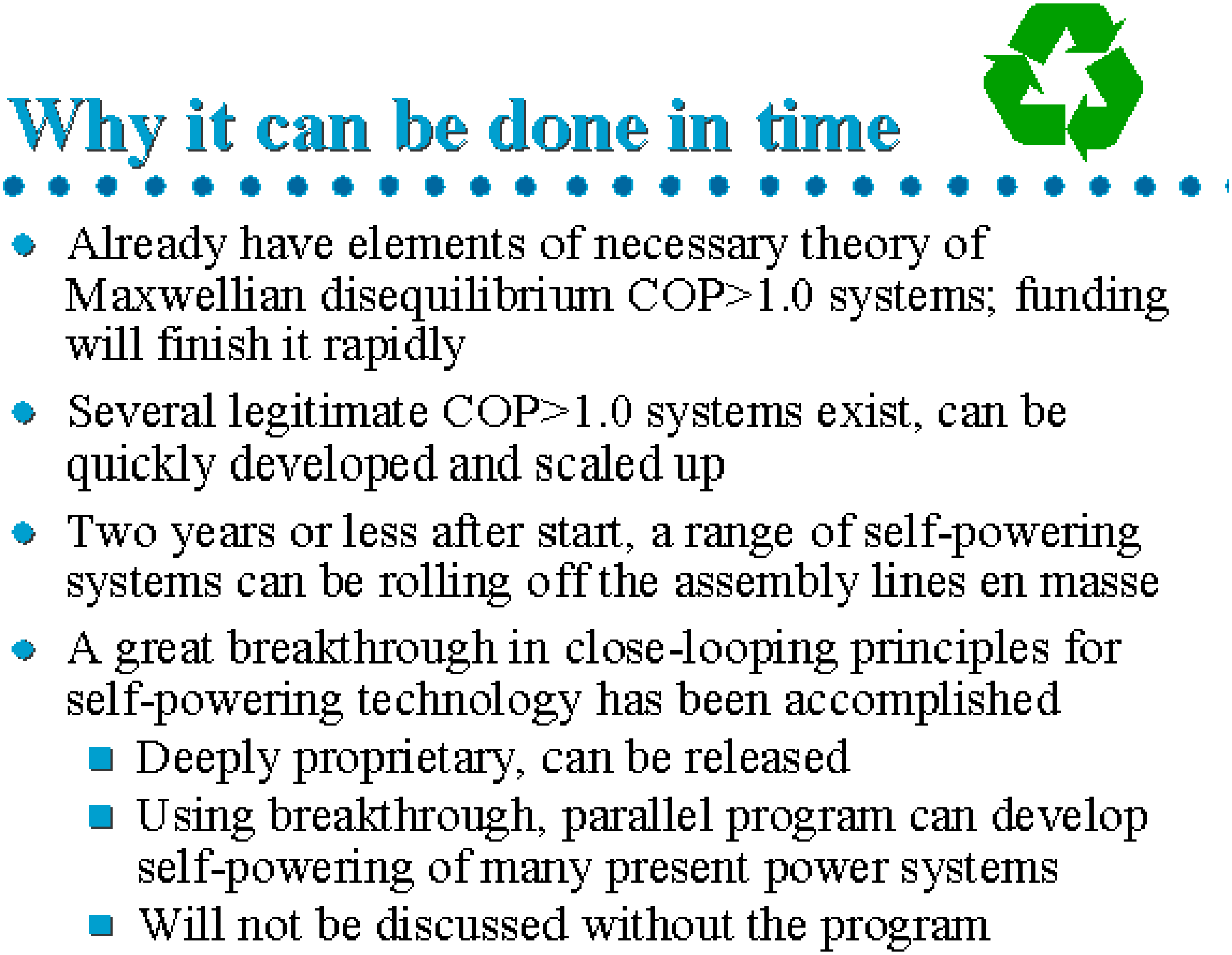

Why it can be done in time

- Already have elements of necessary theory of Maxwellian disequilibrium COP>1.0 systems; funding will finish it rapidly
- Several legitimate COP>1.0 systems exist, can be quickly developed and scaled up
- Two years or less after start, a range of self-powering systems can be rolling off the assembly lines en masse
- A great breakthrough in close-looping principles for self-powering technology has been accomplished
  - Deeply proprietary, can be released
  - Using breakthrough, parallel program can develop self-powering of many present power systems
  - Will not be discussed without the program

Published by Lotus® Freelance Graphics®

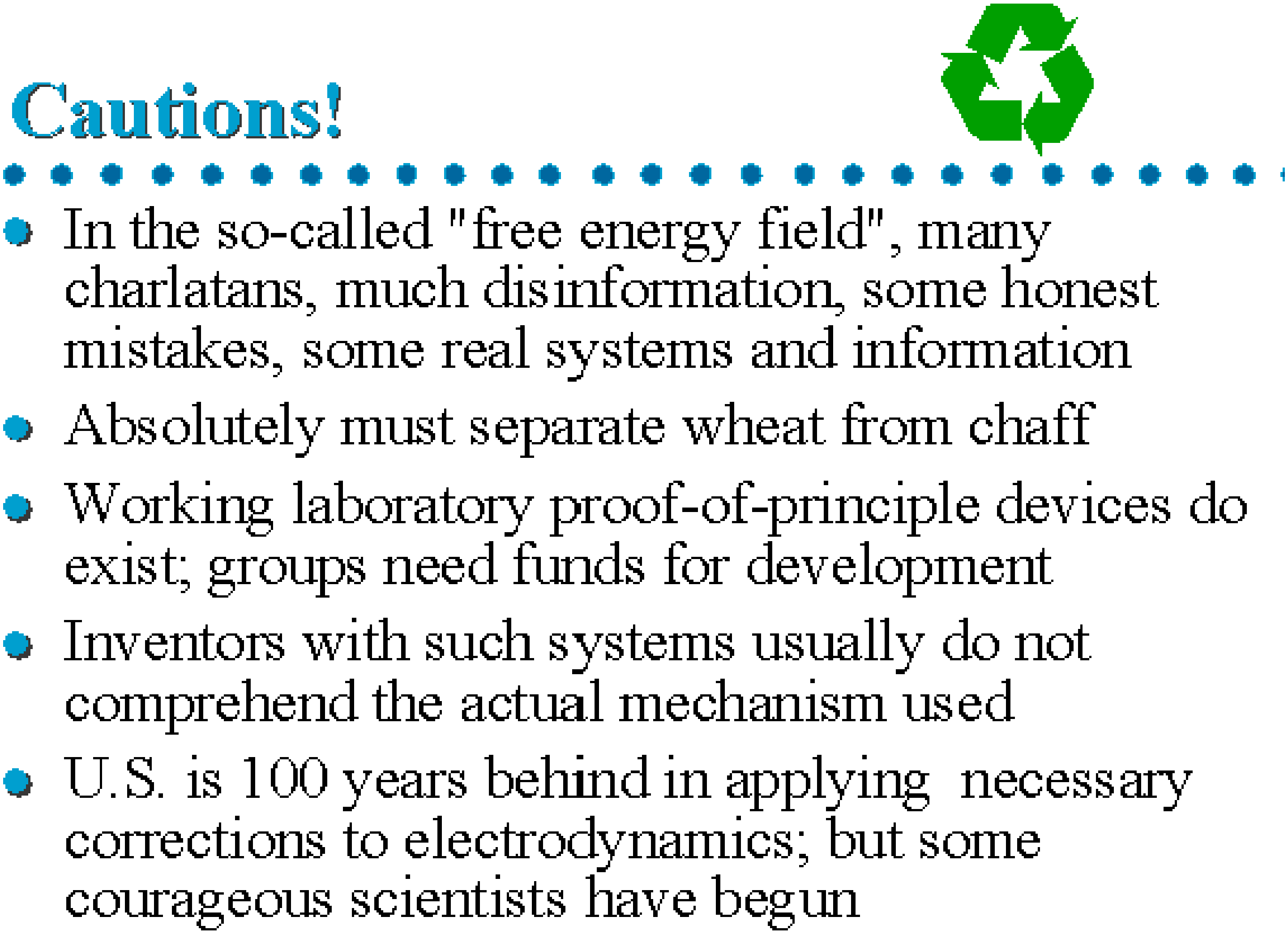

Cautions!

- In the so-called "free energy field", many charlatans, much disinformation, some honest mistakes, some real systems and information
- Absolutely must separate wheat from chaff
- Working laboratory proof-of-principle devices do exist; groups need funds for development
- Inventors with such systems usually do not comprehend the actual mechanism used
- U.S. is 100 years behind in applying necessary corrections to electrodynamics; but some courageous scientists have begun

Published by Lotus® Freelance Graphics®

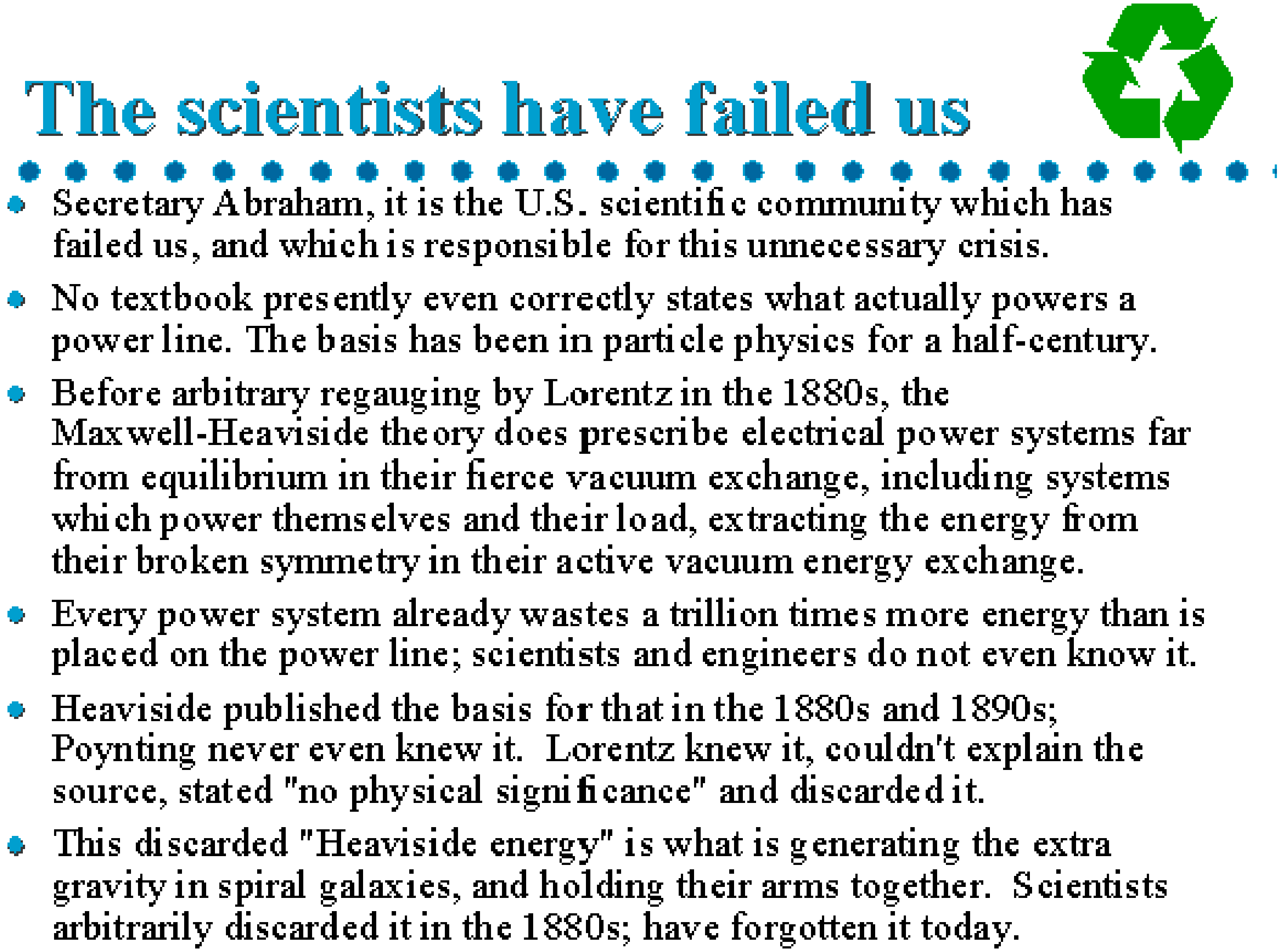

The scientists have failed us

- Secretary Abraham, it is the U.S. scientific community which has failed us, and which is responsible for this unnecessary crisis.
- No textbook presently even correctly states what actually powers a power line. The basis has been in particle physics for a half-century.
- Before arbitrary regauging by Lorentz in the 1880s, the Maxwell-Heaviside theory does prescribe electrical power systems far from equilibrium in their fierce vacuum exchange, including systems which power themselves and their load, extracting the energy from their broken symmetry in their active vacuum energy exchange.
- Every power system already wastes a trillion times more energy than is placed on the power line; scientists and engineers do not even know it.
- Heaviside published the basis for that in the 1880s and 1890s; Poynting never even knew it. Lorentz knew it, couldn't explain the source, stated "no physical significance" and discarded it.
- This discarded "Heaviside energy" is what is generating the extra gravity in spiral galaxies, and holding their arms together. Scientists arbitrarily discarded it in the 1880s; have forgotten it today.

Published by Lotus® Freelance Graphics®

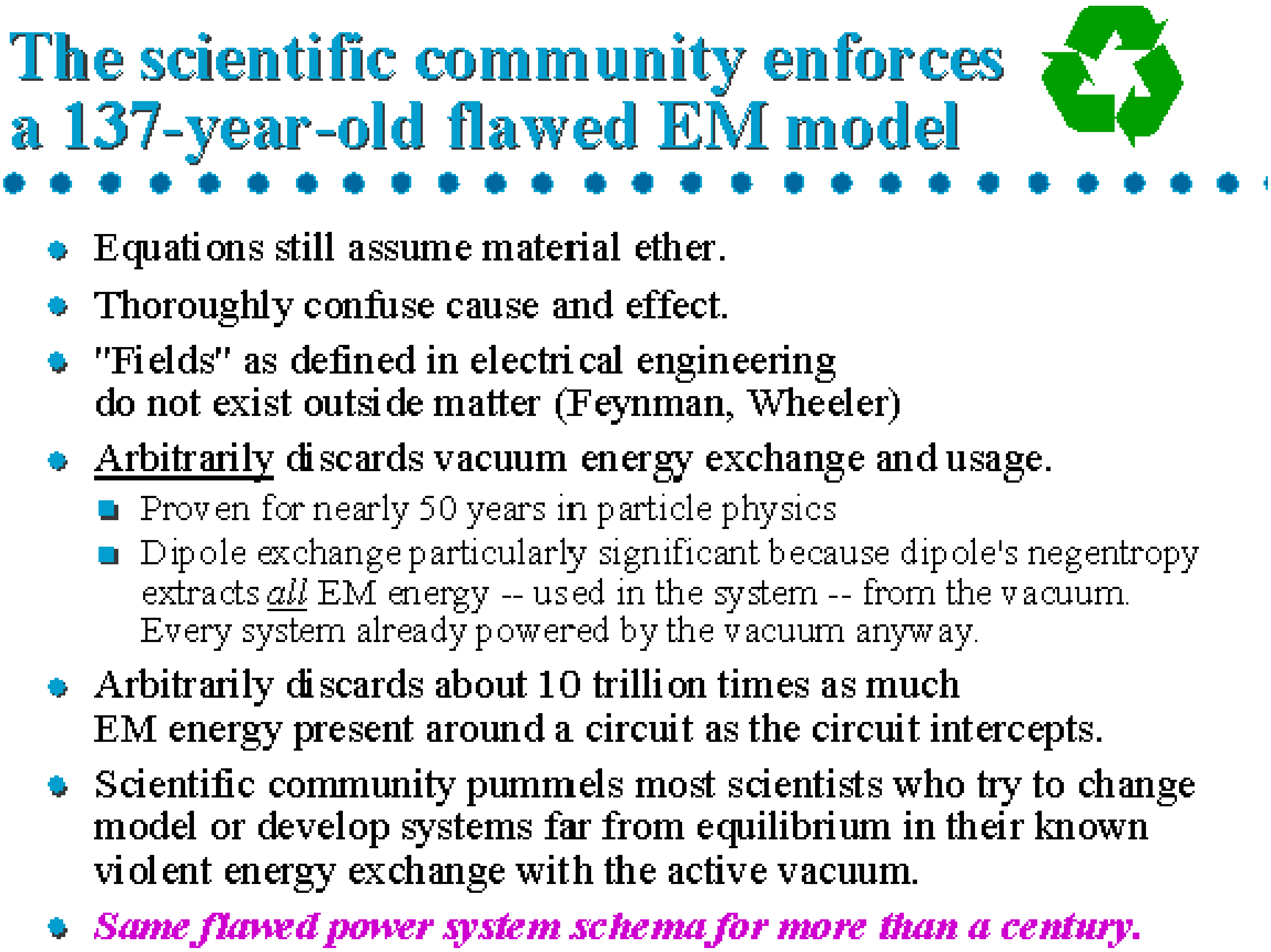

The scientific community enforces a 137-year-old flawed EM model

- Equations still assume material ether.
- Thoroughly confuse cause and effect.
- "Fields" as defined in electrical engineering do not exist outside matter (Feynman, Wheeler)
- Arbitrarily discards vacuum energy exchange and usage.
  - Proven for nearly 50 years in particle physics
  - Dipole exchange particularly significant because dipole's negentropy extracts *all* EM energy -- used in the system -- from the vacuum. Every system already powered by the vacuum anyway.
- Arbitrarily discards about 10 trillion times as much EM energy present around a circuit as the circuit intercepts.
- Scientific community pummels most scientists who try to change model or develop systems far from equilibrium in their known violent energy exchange with the active vacuum.
- ***Same flawed power system schema for more than a century.***

Published by Lotus® Freelance Graphics®

All EM 3-space energy freely comes from the time domain*

* T. E. Bearden, "Giant Negentropy of the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23.
http://www.cheniere.org/techpapers/GiantNegentropy.pdf

Whittaker, Math. Ann., 57, 333 (1903)
shows this rigorously.
Dipole's broken 3-symmetry well-known
in particle physics

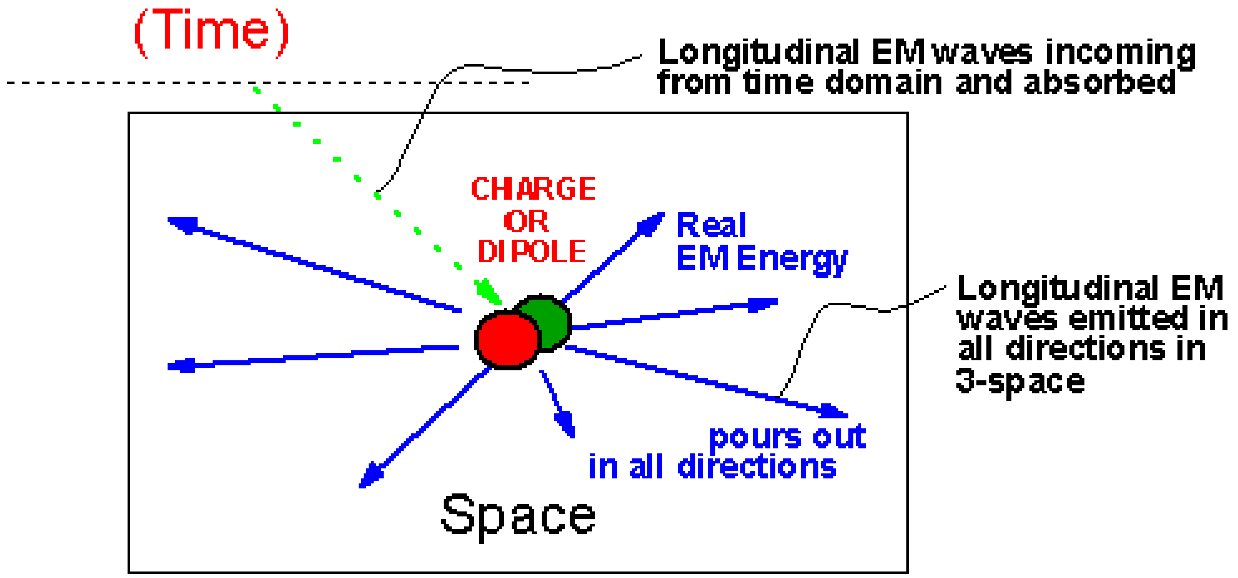
(Time)
Longitudinal EM waves incoming from time domain and absorbed
CHARGE OR DIPOLE
Real EM Energy
Longitudinal EM waves emitted in all directions in 3-space
pours out in all directions
Space

Note: Whittaker (and others) interpreted the phase conjugate half set of LWs after interaction with the charges of the dipole; and as an *effect* rather than the *cause*. This fundamental non sequitur has just been repeated since then. Thus they missed the solution of the source charge and source dipole.

Published by Lotus® Freelance Graphics®

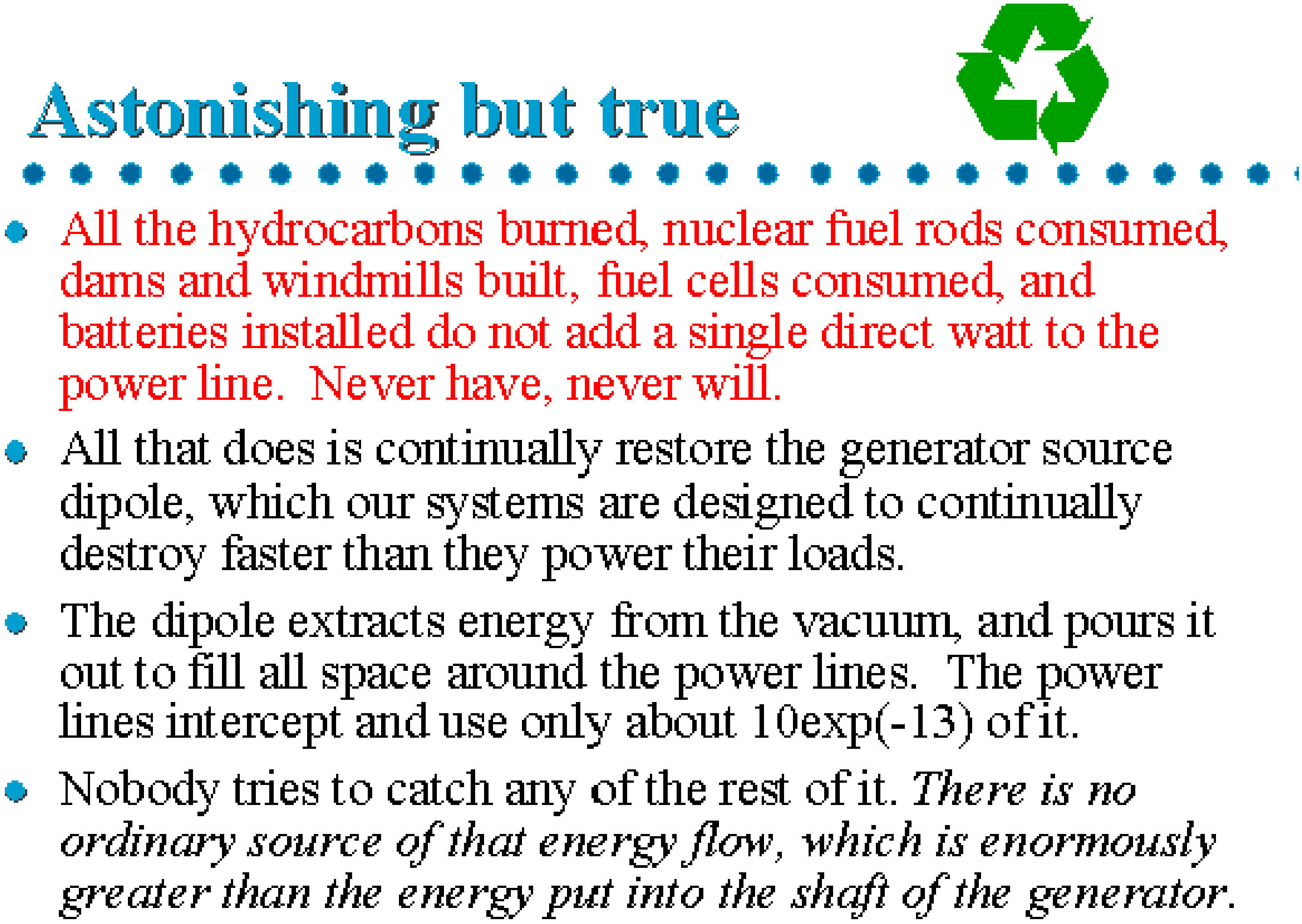

Astonishing but true

- All the hydrocarbons burned, nuclear fuel rods consumed, dams and windmills built, fuel cells consumed, and batteries installed do not add a single direct watt to the power line. Never have, never will.
- All that does is continually restore the generator source dipole, which our systems are designed to continually destroy faster than they power their loads.
- The dipole extracts energy from the vacuum, and pours it out to fill all space around the power lines. The power lines intercept and use only about 10exp(-13) of it.
- Nobody tries to catch any of the rest of it. *There is no ordinary source of that energy flow, which is enormously greater than the energy put into the shaft of the generator.*

Published by Lotus® Freelance Graphics®

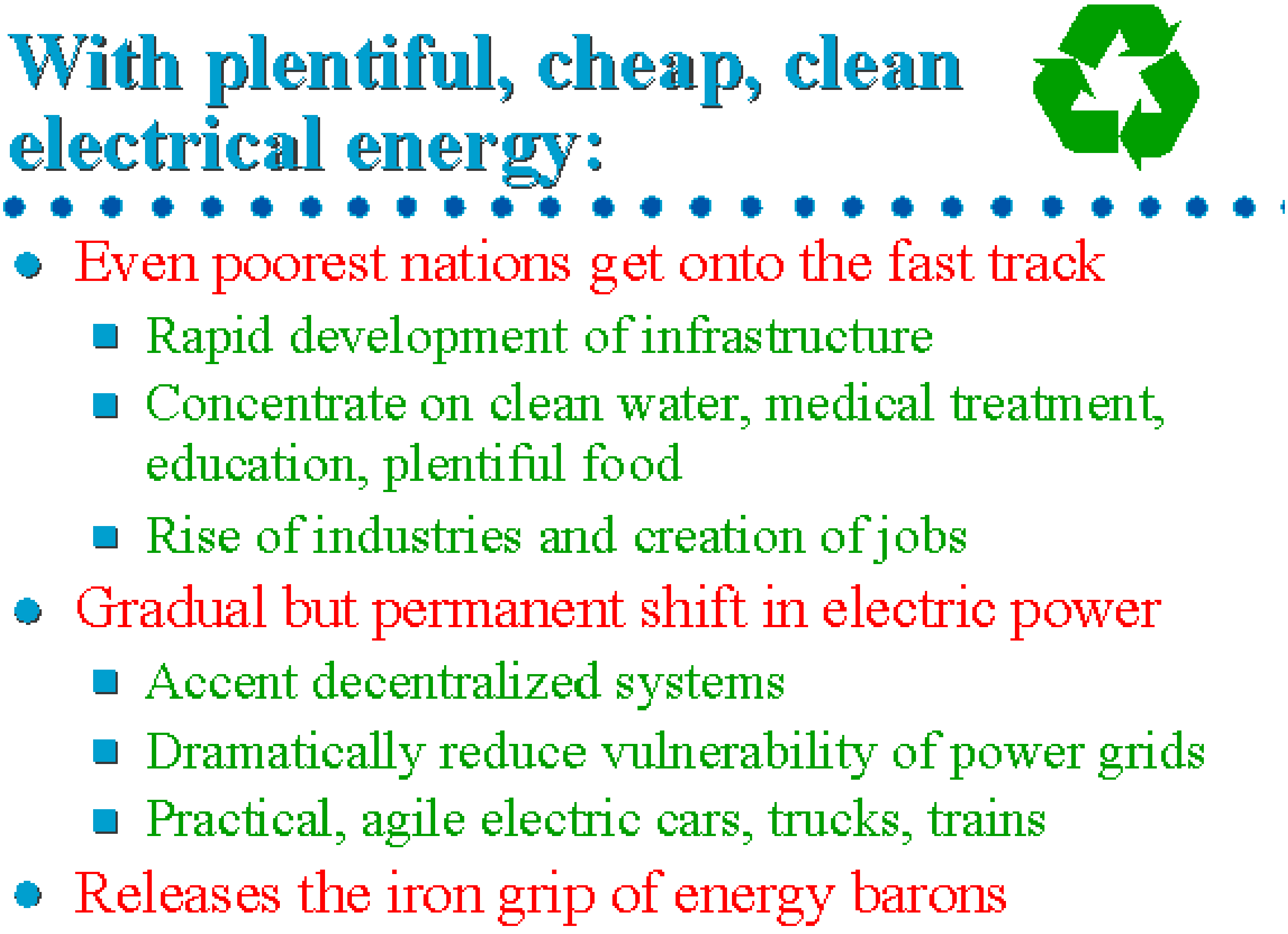

With plentiful, cheap, clean electrical energy:

- Even poorest nations get onto the fast track
  - Rapid development of infrastructure
  - Concentrate on clean water, medical treatment, education, plentiful food
  - Rise of industries and creation of jobs
- Gradual but permanent shift in electric power
  - Accent decentralized systems
  - Dramatically reduce vulnerability of power grids
  - Practical, agile electric cars, trucks, trains
- Releases the iron grip of energy barons

Published by Lotus® Freelance Graphics®

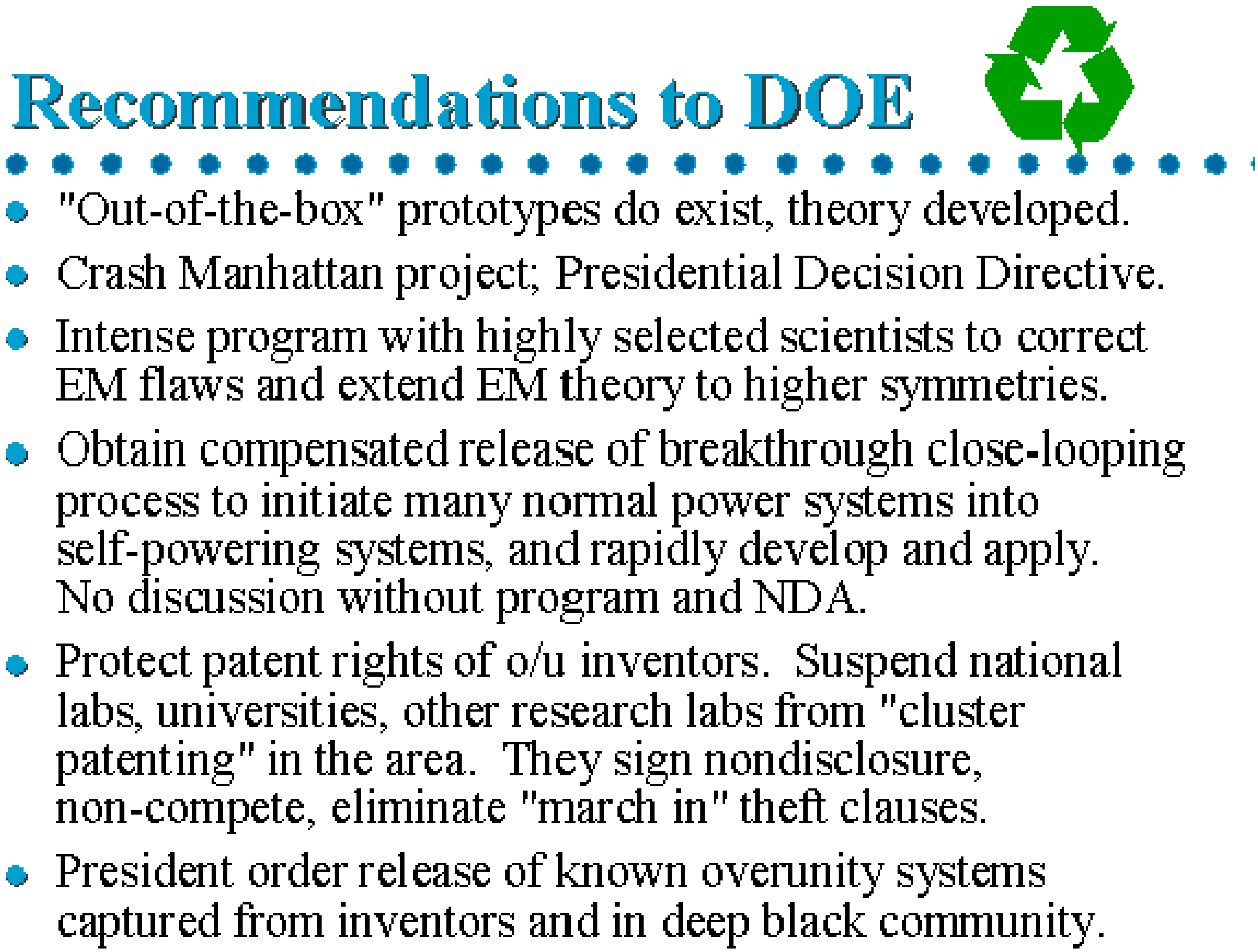

Recommendations to DOE

- "Out-of-the-box" prototypes do exist, theory developed.
- Crash Manhattan project; Presidential Decision Directive.
- Intense program with highly selected scientists to correct EM flaws and extend EM theory to higher symmetries.
- Obtain compensated release of breakthrough close-looping process to initiate many normal power systems into self-powering systems, and rapidly develop and apply. No discussion without program and NDA.
- Protect patent rights of o/u inventors. Suspend national labs, universities, other research labs from "cluster patenting" in the area. They sign nondisclosure, non-compete, eliminate "march in" theft clauses.
- President order release of known overunity systems captured from inventors and in deep black community.

Published by Lotus® Freelance Graphics®

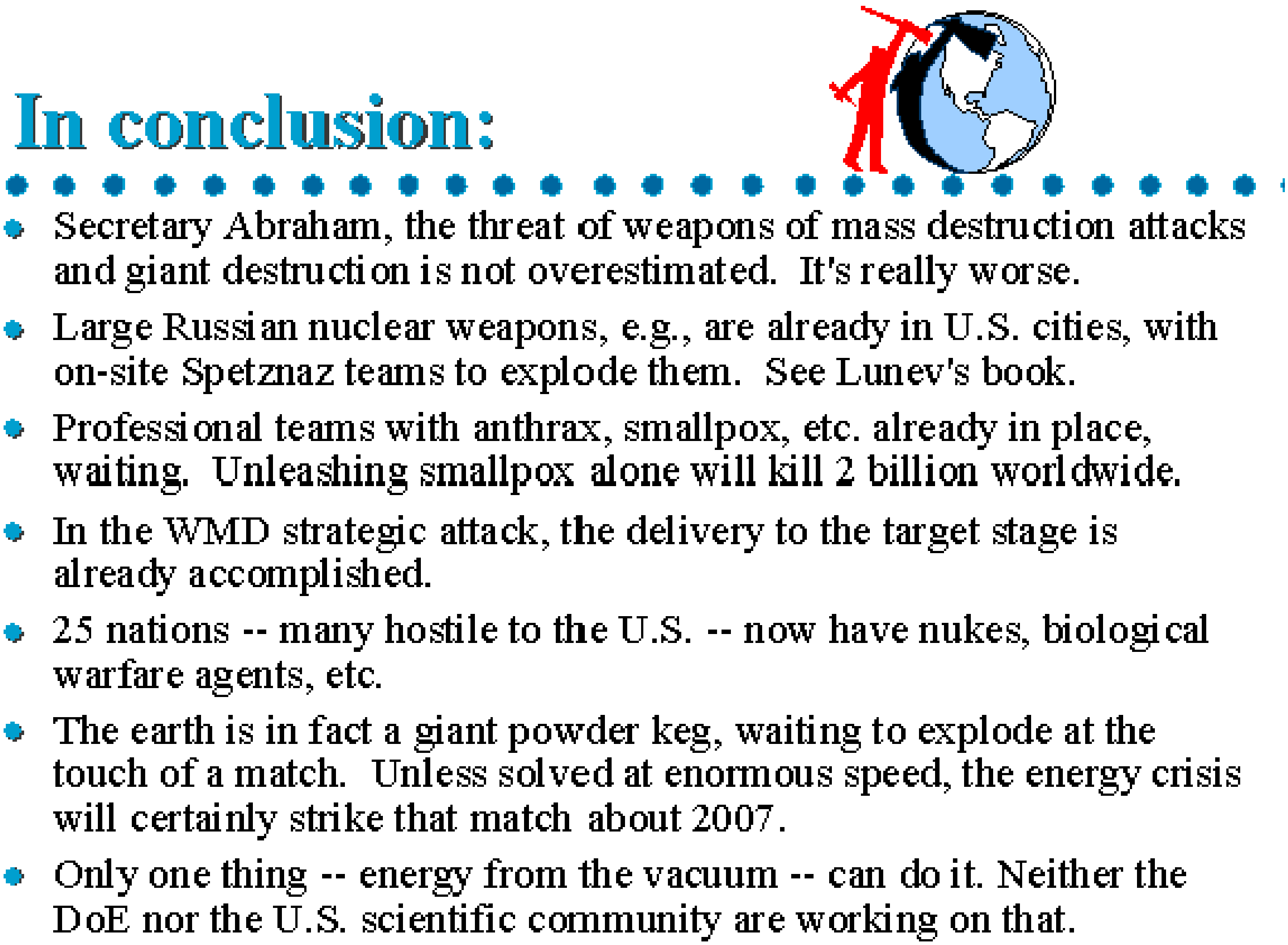

In conclusion:

- Secretary Abraham, the threat of weapons of mass destruction attacks and giant destruction is not overestimated. It's really worse.
- Large Russian nuclear weapons, e.g., are already in U.S. cities, with on-site Spetznaz teams to explode them. See Lunev's book.
- Professional teams with anthrax, smallpox, etc. already in place, waiting. Unleashing smallpox alone will kill 2 billion worldwide.
- In the WMD strategic attack, the delivery to the target stage is already accomplished.
- 25 nations -- many hostile to the U.S. -- now have nukes, biological warfare agents, etc.
- The earth is in fact a giant powder keg, waiting to explode at the touch of a match. Unless solved at enormous speed, the energy crisis will certainly strike that match about 2007.
- Only one thing -- energy from the vacuum -- can do it. Neither the DoE nor the U.S. scientific community are working on that.

Published by Lotus® Freelance Graphics®

The "Porthole" Concept with technical background

Amplifying the cellular regenerative (healing) system of the body for treatment and healing
of mass casualties

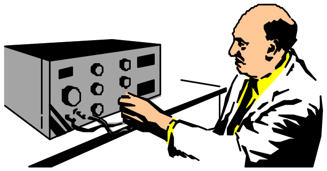

Outline of the presentation

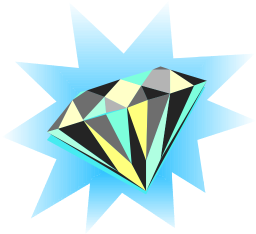

- Mechanism for deep EM exchange of a body with its active EM environment
- Becker's epochal work and its ramifications
- Kaznacheyev's startling results and ramifications
- Weaponization of Kaznacheyev's findings
- Importance of time-energy and the time-domain
- Mechanisms for medical treatment and healing

Outline of the presentation (2)

- Prioré's revolutionary results and their ramifications.
- Proposed new mechanism for cancer promotion.
- Porthole concept: The 2-minute shortcut.
- Terrorist biological warfare capabilities.
- Summary and conclusion.
- Budgetary and Program package.

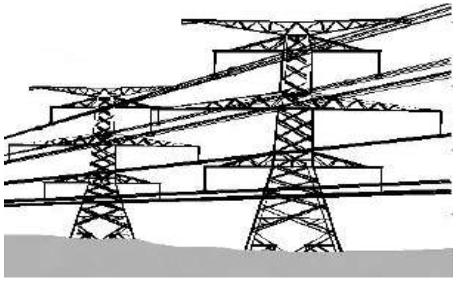

Mechanism for the body's deep EM exchange with its active EM environment

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

And some ramifications

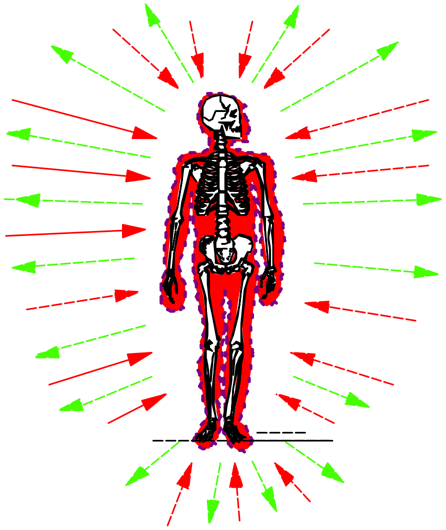

The body continually receives and retransmits EM energy

- The body may be considered as a dielectric.
- Every part of the dielectric participates in the absorption and emission of each photon to and from any point on the surface.*
- Mechanical force effects are also electrical in nature and exchange similarly.

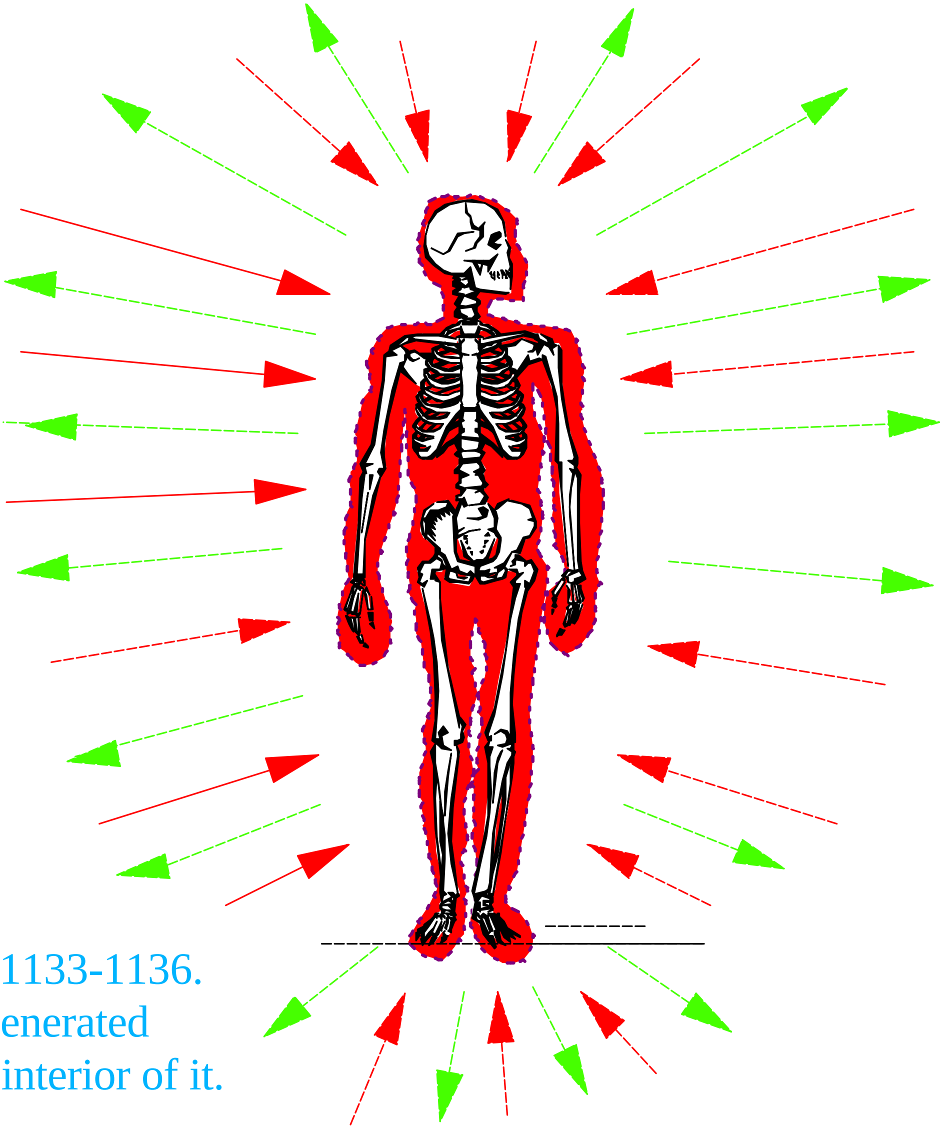

* Reali, G. C., "Reflection from dielectric materials," American Journal of Physics, 50(12), Dec. 1982, p. 1133-1136. The reflected field from a dielectric material is not generated just at its surface but comes from everywhere in the interior of it.

Ramifications:

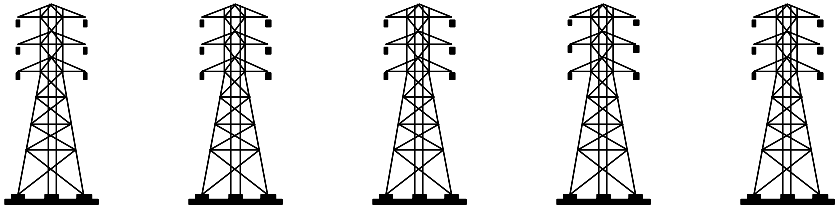

- There is an EM "dielectric transmission path" connecting every part of every cell to every part of the surface of the body.
- All electrical processes in the body, including deep within the cell and its nucleus, participate in the absorption and emission of every photon to and from the body's outer surface (and every mechanical force* experience).
- The EM emission energy of the body is a conglomerate mix of fractions of all processes and functions ongoing in the body or affecting it, including in every cell.
- This includes all electromagnetic, mechanical, chemical, and environmental effects upon and in the body and every tiniest part of it, and provides an extended definition of "stress".

* At base, all normal mechanical forces are generated by electromagnetic forces.

Ramifications (2)

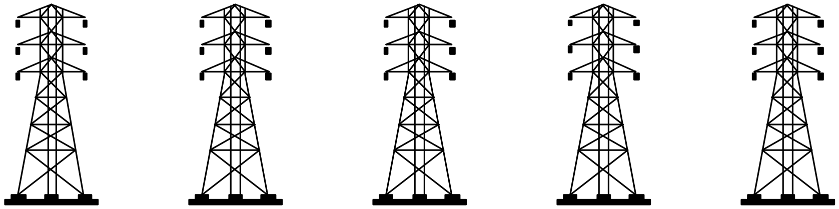

- The reaction of each photon absorbed on any point of the outer surface of the body, connects through the dielectric transmission path to every part of the body, including to every part of every cell*.
- Every emission from the body is participated in by every part of the body, including every part of every cell.
- The body's emission "changes" its environment, adding EM radiation. Some of the internal induced changes (excited states) in the body decay by this external emission. Some do not, but must be otherwise dealt with internally by the body systems.
- The body's EM absorption allows the environment to change or affect every function, operation, and part of the body, by adding EM radiation. All are continually being affected and deviated, both helpfully and harmfully, by these deeply induced EM interactions.
- The body's reaction (healing) must continually readjust and correct deviations from normalcy, in every part of every cell.

* In quantum physics, part of the photon is localized and part is nonlocalized, reaching even across the universe.

Ramifications (3)

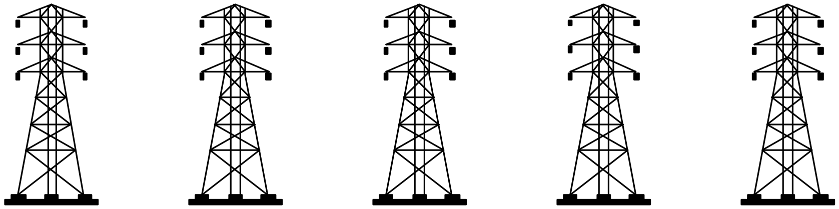

- The sum total of all the impinging external EM radiation upon the body, plus mechanical "special EM" radiation, thus affects and changes the body and all its processes, participating in them.
- That impinging EM radiation can be manipulated to either harm or heal the body and any or all of its cells.
- The body must continually overcome these externally-induced EM changes in its cells and processes, to maintain normalcy. Some EM changes are restored by the subsequent dielectric emission process, relieving the previously-induced deviant EM excitations in its internal processes. Some are not.
- The poorly studied *cellular regenerative system* performs the remainder of the restorative function, which is the fundamental "healing" function of the body.*

*The cellular regenerative system was mostly studied by Becker and a few others. See R. O.Becker, and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, No. 73, Nov.-Dec. 1970, p. 169-198.

Becker's epochal work

And its ramifications

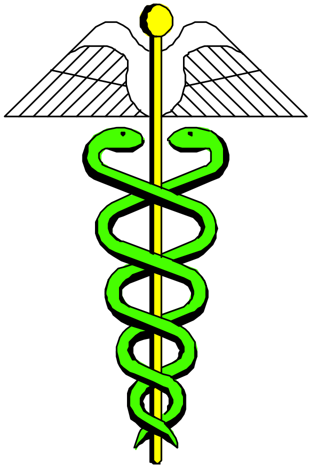

Becker's bone fracture healing

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

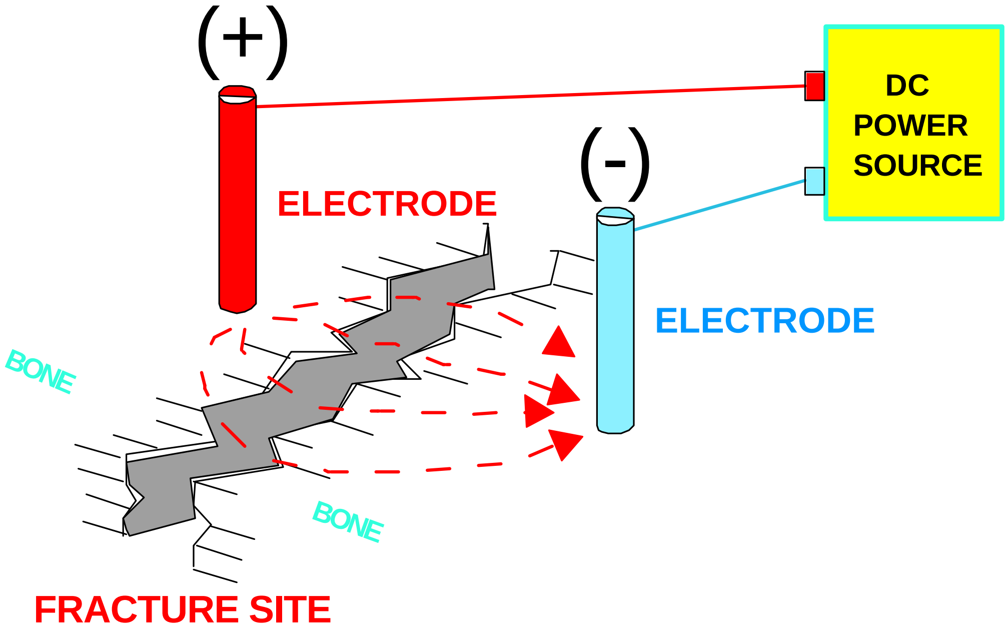
(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Becker's theoretical DC control system involved with response to injury

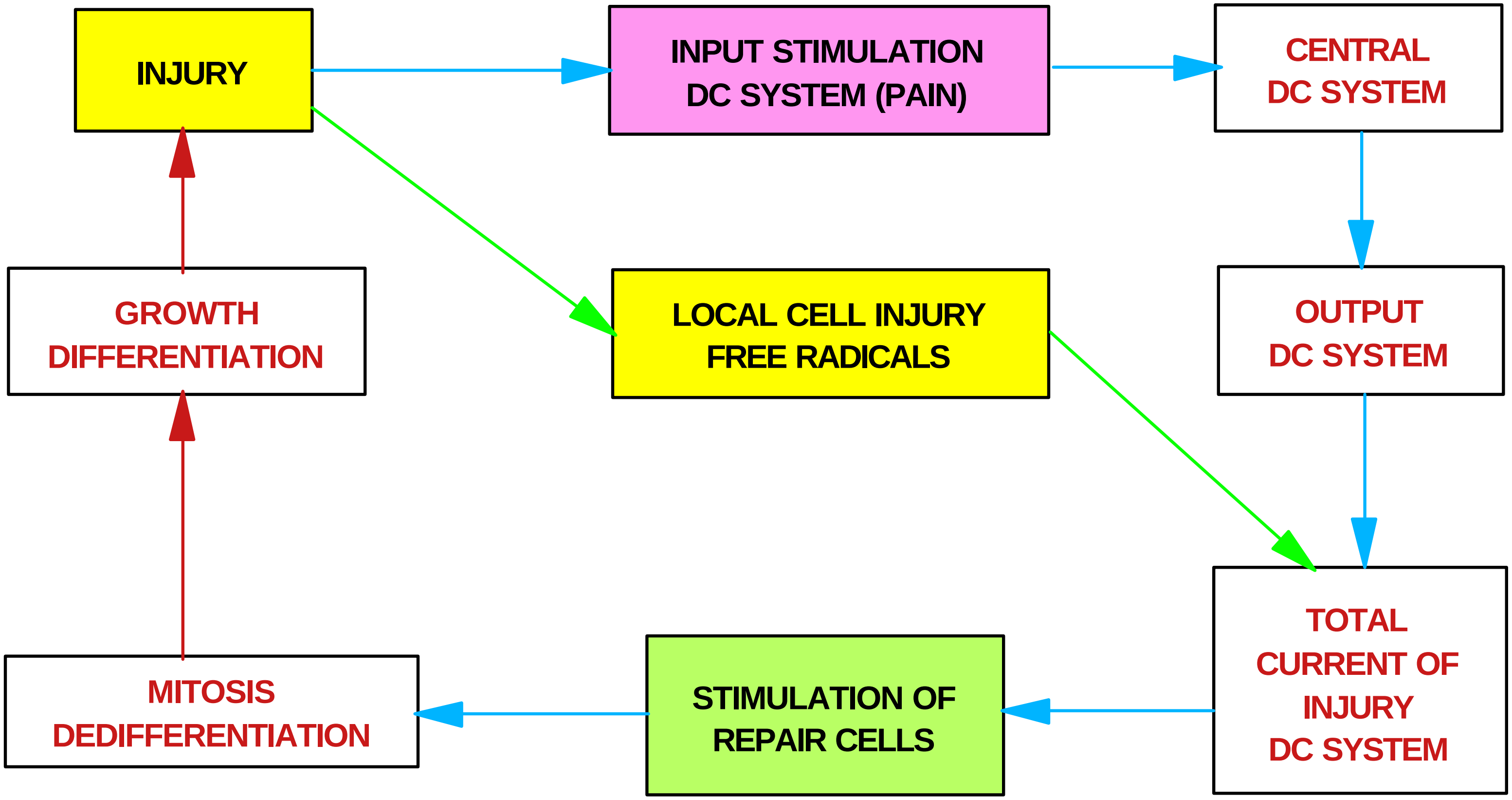

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

MITOSIS DEDIFFERENTIATION

STIMULATION OF REPAIR CELLS

***Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.**

Becker's proposed control system governing regeneration*

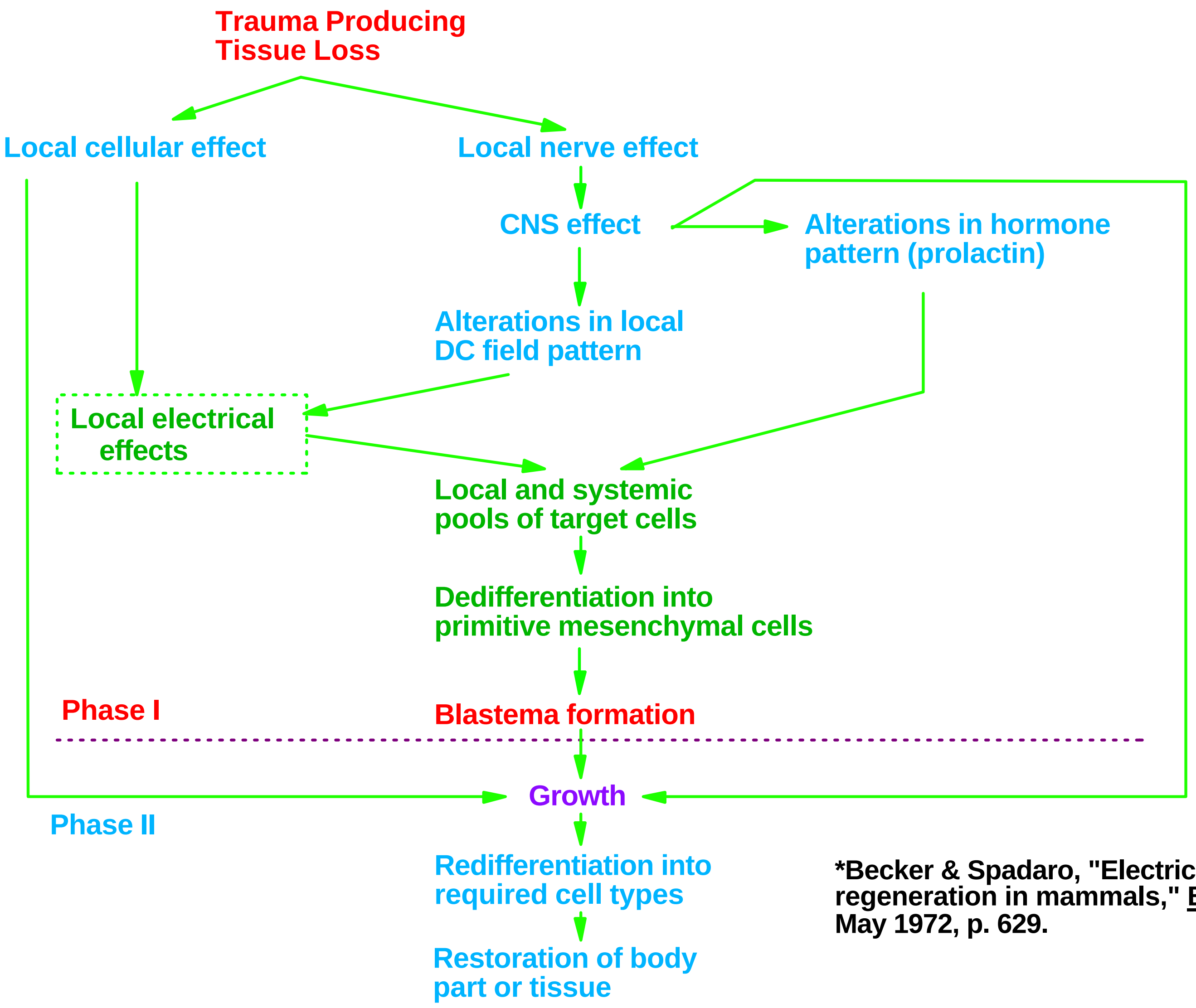

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

Growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

---

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Extending Becker's work

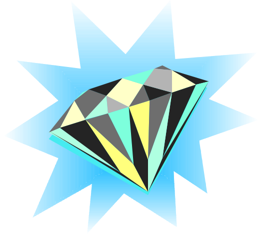

- Becker had only the archaic classical electrodynamics available to him. That EM is grossly inadequate.
  - Insufficient group symmetry to model the process.
  - Already eliminates the "infolded" longitudinal wave EM that structurally and dynamically comprises all "normal" EM potentials, fields, and waves.
  - It is this internal LW EM that is manipulated and used by the cellular regenerative system.
- Nonlinear phase conjugate optics was not yet developed when Becker did his seminal work.
- Successful higher symmetry electrodynamics models -- such as O(3) -- have been developed and are available.

A sad commentary

- Becker was nominated for the Nobel Prize, but he was hounded, his funds were withdrawn, and he was forced to retire early.
- His main "sin" was having the courage to testify against the harmful effects of unrestricted powerline radiation and electronic smog.
- Powerful interests then crushed him, set up their own well-funded institutes, and hired scientists who then predictably found that powerline radiation was no problem.
- Such research is so tainted that recently several important journals have now required the authors of submitted papers to state what their affiliated companies do, and what "company massaging" of their research occurred.
- From a higher symmetry EM view, it is obvious that the dielectric pathway and quantum electrodynamics already show a significant problem due to such radiation, particularly with respect to long-term effects.
- This is particularly significant since "noise" -- the mere presence of more "hash" -- appreciably amplifies the dielectric transmission effect. So the major factor becomes EM smog *signals density*, not signal power.

Kaznacheyev's startling results

And its ramifications

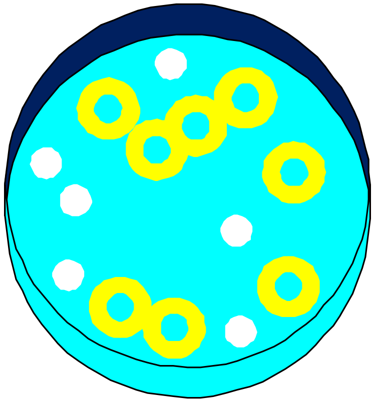

Proof that the cellular condition is radiated from groups of cells

- It follows that a diseased, damaged, or infected cell culture will and does emit its exact internal EM condition, hence the disease or damage "generatrix form" itself (in the internal structuring of the EM radiations).
- Further, these "state" emissions can couple into other targeted cells to produce disease and disorder in them.
- This was clearly proven in some 15,000 Russian experiments by Kaznacheyev.
- Kaznacheyev demonstrated induction of cellular disease and disorder between EM-coupled but otherwise environmentally shielded cell cultures
- Similar experiments have been replicated in the West by Reid *et al*., and by Popp *et al*.

Kaznacheyev's induction of cellular disease and disorder

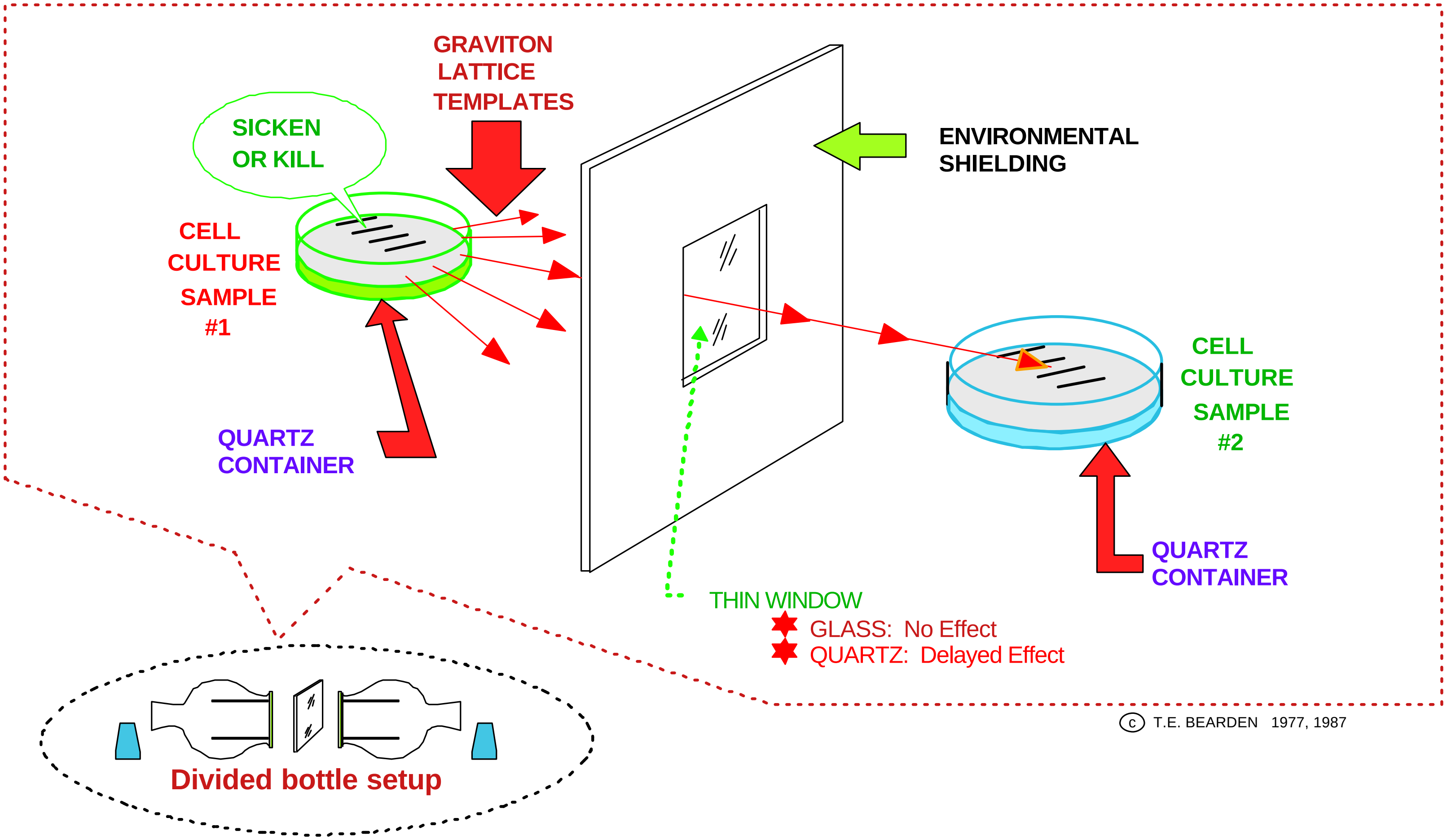
GRAVITON LATTICE TEMPLATES
SICKEN OR KILL
CELL CULTURE SAMPLE #1
QUARTZ CONTAINER
ENVIRONMENTAL SHIELDING
CELL CULTURE SAMPLE #2
QUARTZ CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle setup

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice. Consider particularly the dielectric pathway in the emitting diseased cells and in the receiving targeted healthy cells.

Another surprising feature of Kaznacheyev's experiments

- The experiments do not work if the targeted cells are in normal light, such as sunlight*.
- The reason is that the normal light bandwidth occupies a single harmonic interval between ultraviolet and infrared.
- Hence its difference frequencies also fill such an interval. For a set of n frequencies, the number *d* of difference frequencies is $d >> n$.
- Thus the visible light spectrum represents "jamming" of specific internal EM signal inputs through the dielectric pathway into the body, turning much of the environmental signals into harmless warmth.

* It has also been shown that if bacteria are killed in the "dark" by UV, then 24 hours later placed in sunlight, a substantial fraction of the "dead" bacteria will revive.

Ramifications of Kaznacheyev's experimental results

- At least one harmonic interval is necessary for the effect to be evidenced in the target cells.
- Interestingly, if we assume the body dielectric to be isotropically nonlinear (to first order):
  - Velocities of the actual transmitted frequencies depend upon the particular point on the wave amplitude.
  - The transmitted frequencies overshoot, interfere, breakup, etc. The difference frequency does not.
  - The "difference frequency" is transmitted through the nonlinear dielectric medium as if it were a sine wave passing through a linear medium*.
- To our knowledge, this direct and deep communication of electronic "hash" difference frequencies in dielectric human bodies has not been investigated in the West.

* Owen Flynn, "Parametric arrays: A new concept for sonar," Electronic Warfare Magazine, June 1977, p. 107-112.

Ramifications of Kaznacheyev's experimental results (2)

- A heterodyne signal -- at the difference frequency between an input signal and a reference signal -- can be enhanced by adding noise*.
- Thus the signal of the disorder in the cellular pattern emitted from diseased cells, received by normal cells, can be amplified electronically and rather easily.
- Powerline radiation (electronic smog) thus carries its own amplification mechanism. Eventually the amplification overcomes any squelching by sunlight, etc.
- The Russians promptly (as early as the late 1950s) weaponized and tested these and similar effects.

* Dykman, M. I. et al., "Noise-enhanced heterodyning in bistable systems," Physical Review E, 49(3), Mar. 1994, p. 1935-1942

Weaponization of Kaznacheyev's findings

And its ramifications

"Microwave" radiation of the U.S. Embassy in Moscow.

Former U.S. Embassy in Moscow.

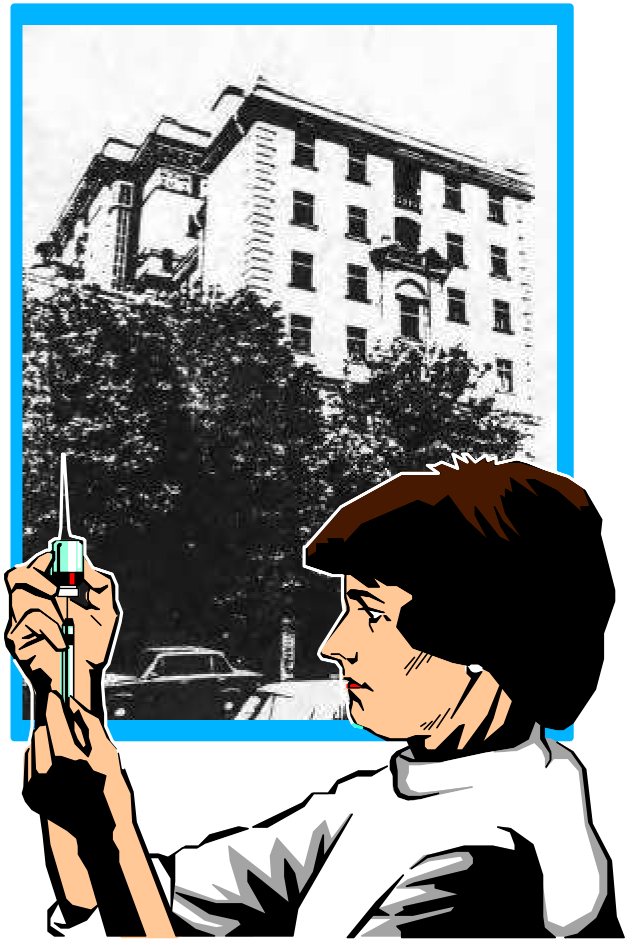

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Ramifications of the microwave radiation of the U.S. Embassy

- The cellular disorder pattern in the EM radiation is carried by the field-free scalar potential.
- Since the disorder pattern is a "conglomerate", it must be carried by an internal EM structure existing inside the scalar potential itself.
- Whittaker decomposed the scalar potential into the appropriate internal longitudinal EM biwave structure in 1903*.

* E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355

The importance of time as energy, and the time-domain

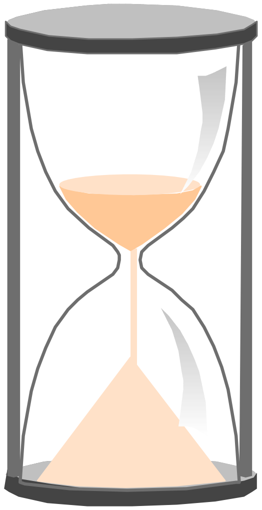

Time as energy

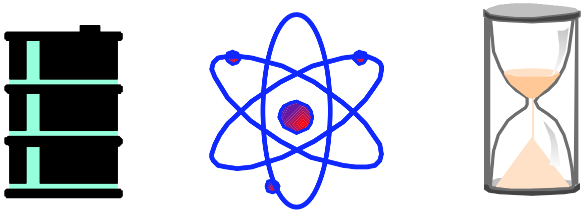

- In physics, the fundamental units one uses in his model are arbitrary.
  - A valid model can be made using only one unit.
  - It is already done in one branch of physics, using length as the single fundamental unit.
- Make a model with a single unit, the joule.
  - All other entities become functions of the joule.
  - We are familiar with mass as highly compressed EM energy, by the factor c^2.
  - It turns out that time is also highly compressed EM energy, by the factor c^2.
  - 1 sec. $\approx 9 \times 10^{16}$ joules.

How to see that "time is compressed spatial energy"

- Take some spatial EM energy and compress it by c^2
- Place it in 3-space, and it is *mass*.
- Place it on the time axis, and it is *time*.

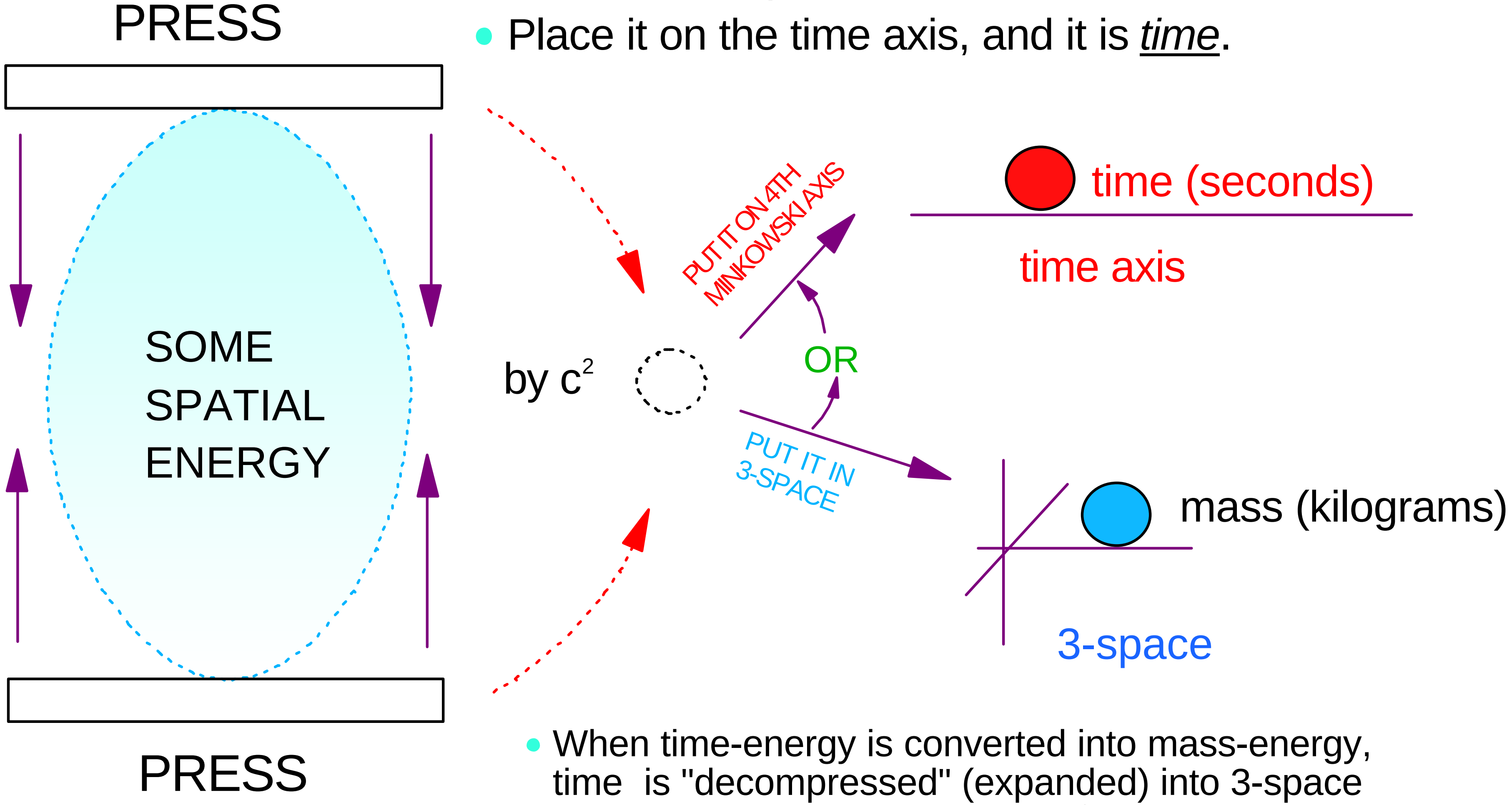

PRESS

SOME
SPATIAL
ENERGY

PRESS

by c^2

PUT IT ON 4TH MINKOWSKI AXIS

OR

PUT IT IN 3-SPACE

time (seconds)

time axis

mass (kilograms)

3-space

- When time-energy is converted into mass-energy, time is "decompressed" (expanded) into 3-space EM energy and 1 sec. $\approx 9 \times 10^{16}$ joules.

Time as EM energy in the time domain, with a density the same as mass-energy

- Bearden* reinterpreted (corrected) the 1903 Whittaker decomposition of the potential in 2000.
- The reinterpretation is consistent with broken symmetry in particle physics, for which Lee and Yang received the Nobel Prize in 1957.
- It is also powerfully supported by quantum field theory, notably by Mandl and Shaw in 1984**.
- Whittaker's pair is a combined time-polarized (scalar) EM wave and a 3-space longitudinal EM wave.

* T. E. "Giant Negentropy from the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23.

** F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5.

All EM energy in 3-space freely comes from the time domain

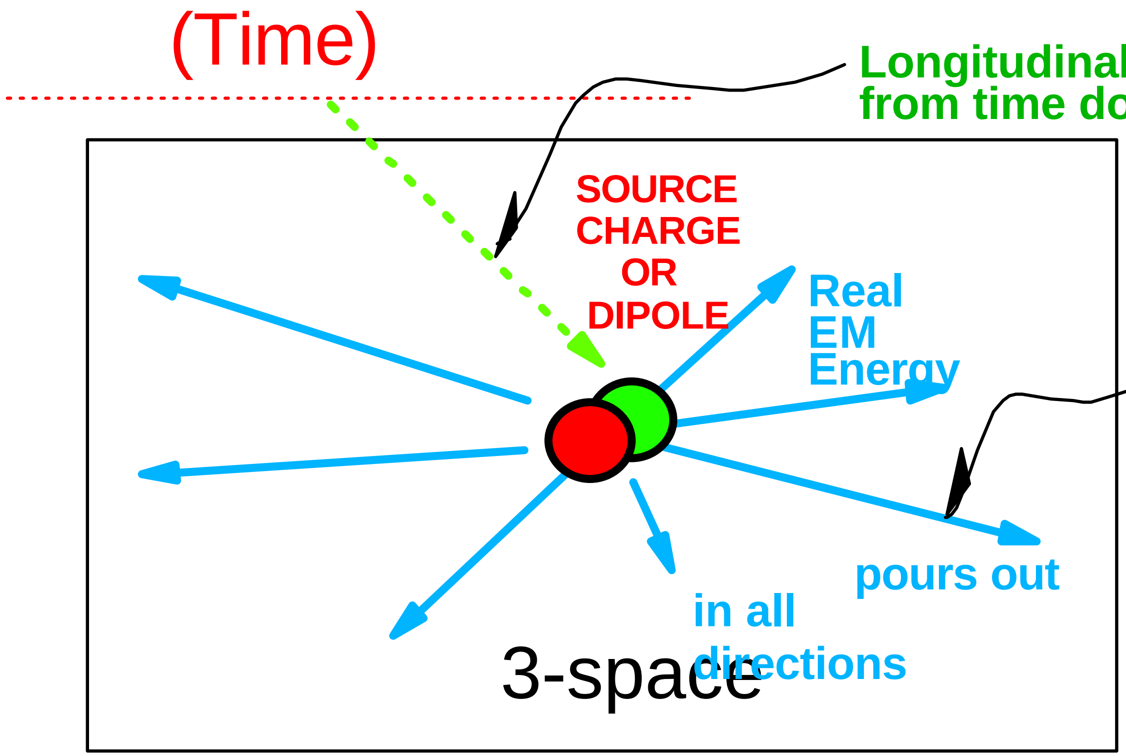

Longitudinal EM wave incoming from time domain and absorbed

Charge's 720º spin transforms time-energy into 3-space energy

Longitudinal EM waves emitted in all directions in 3-space

Whittaker, Math. Ann., 57, 333 (1903) shows this rigorously, when properly reinterpreted. Dipole's broken 3-symmetry is well-known in particle physics

Note: Whittaker (and others) interpreted the phase conjugate half set of LWs after interaction with the charges of the dipole, and as a 3-space *effect* rather than the time-domain *cause*. This fundamental non sequitur was just repeated since then, until corrected by Bearden to be consistent with broken symmetry and quantum field theory.

Strong support from quantum field theory

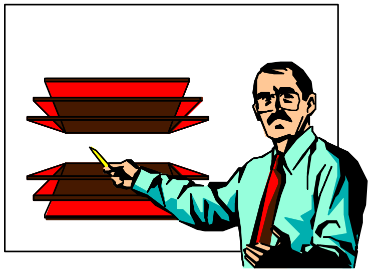

- Quantum field theory recognizes four photon polarizations:
  - Energy along x-, y-, z-, or t- axis.
  - Polarized along x- and/or y-axis, the photon is a *transverse* photon.
  - Polarized along z-axis (along its line of motion in 3-space), the photon is a *longitudinal* photon.
  - Polarized along the time-axis, the photon is a *time-polarized* or *scalar* photon.
- The individual longitudinal or scalar photon is not observable*.
- The combination of a longitudinal photon and a scalar photon is observable as the instantaneous scalar potential*.

* F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5.

Mechanism generating the flow of a mass through time

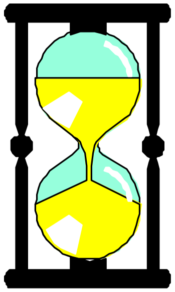

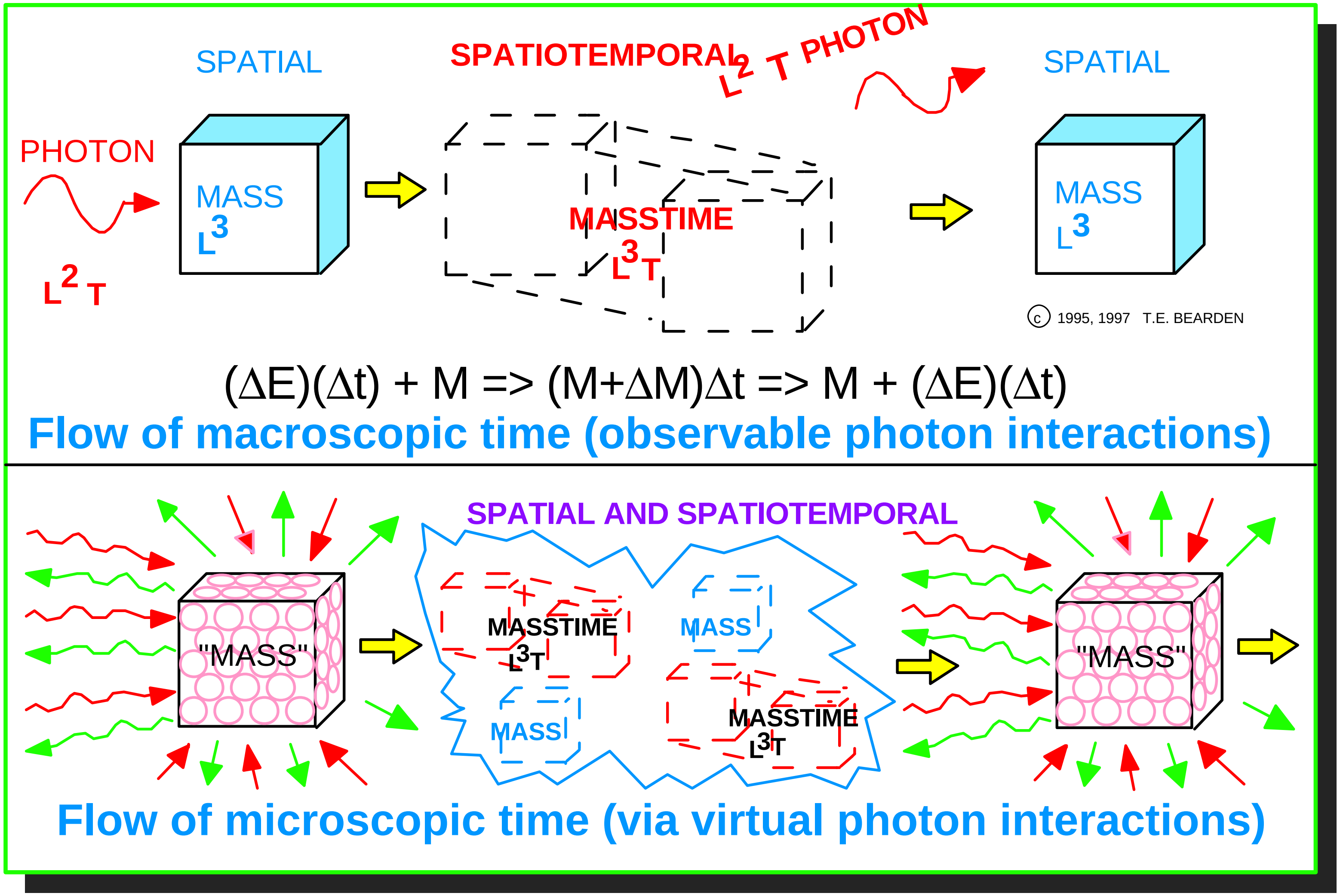
SPATIAL
SPATIOTEMPORAL
L^2T PHOTON
SPATIAL
PHOTON
MASS
L^3
L^2T
MASSTIME
L^3T
MASS
L^3
© 1995, 1997 T.E. BEARDEN
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

Flow of microscopic time (via virtual photon interactions)

Mechanism is engineerable

Pumping a mass in the time-domain will time-reverse it back to an earlier state

- The reinterpreted biwave EM structure of the scalar potential means that time-reversal (phase conjugation by pumping in the time-domain) can be utilized to time-reverse a targeted cell (its mass-energy) in its entirety, including its genetics.
- Becker's bone fracture experiments showed it:
  - Weak scalar potential across the fracture site.
  - LWs in 3-space and time-energy waves in time
  - Red blood cells entered, dedifferentiated by shucking their hemoglobin and growing a nucleus, which is dedifferentiation (time-reversal of the cell and its mass-energy back to a previous physical state).

Principle for time-reversing a time-pumped mass

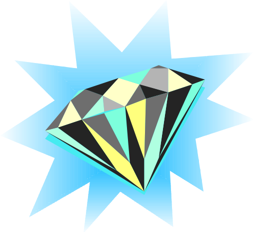

- Becker's cells then redifferentiated and changed -- first to primitive cartilage cells, then to primitive bone cells which were deposited in the fracture, healing it with new bone growth.
- This shows "time-forwarding" (redifferentiation) of the cell also.
- Becker actually showed that the *time-domain pumping* of the cells (by the time-polarized EM waves of Whittaker's reinterpreted wavepairs comprising the scalar potential), and the time-pumping of their mass-energy, operated to continuously and steadily *remove the delta between that cell's present condition and its former "normal" cellular physical condition in that area*.
- This is the long-sought secret of stem-cell research.
- Phase conjugate optics had not been born at the time. Without realizing its full nature, Becker demonstrated the body's own internal mechanism for healing, and the fundamental mechanism for time-reversing a physical mass, whether living or inert.

Mechanisms for use in medical treatment and healing

And ramifications

Successive deep penetration reactions of the cellular dielectric

- Change after change affects the cell and all of its parts, including from the external environment.
- Each successive change affects primarily the internal "LW EM structures and dynamics" comprising the ordinary potentials, fields, and waves.
- Thus in positive time the internal EM structurings produce successive "overlaid layers of change" -- so to speak -- in the internal EM structuring and dynamics of a cell in the body. These layers contain the "delta" from normal state.

Successive deep penetration reactions of the cellular dielectric (2)

- During time reversal, highest layer deltas are successively "peeled away" by the interaction of the mass-energy with the time-domain pumping.
- The result is to return the cell directly back to successive physical states, in the case of a single disease or disorder.
  - A cancer cell is returned to a normal cell prior to its "promotion".
  - An HIV-infected cell would be returned to its normal unaffected condition, free of HIV-induced genetic change.
  - For injuries and damage, the existing "delta" in each successive state-layer is steadily reduced to zero.
- This process thus can involve both cellular differentiation and dedifferentiation, as shown by Becker and by Prioré.

Engines (vacuum engines, spacetime curvature engines)

- Following Whittaker* as reinterpreted**, all normal EM fields, potentials, and waves decompose into sets of longitudinal EM waves (LWs) and their dynamics.
- Since these LWs and their dynamics are dynamics in the local energy density of spacetime (ST), they constitute *ST curvatures*.
- Hence they are general relativity infolded inside electrodynamics.
- By manipulating and changing these LWs, one manipulates and changes spacetime curvatures and their dynamics.
- The result is a unified field model where one directly controls and engineers a set of ST curvatures by use of higher group symmetry electrodynamics.
- Controlled patterns of local curvatures of ST are called "vacuum engines" or "spacetime curvature engines" -- or just "engines".

* E.T. Whittaker, Math. Ann., 57, 333-355 (1903); Proc. Lond. Math. Soc., Series 2, 1, 1367-372 (1904).

** T. E. Bearden, J. New Energy, 5(1), 11-23 (2000).

Pumping with EM time-polarized waves forms *amplified* antiengines

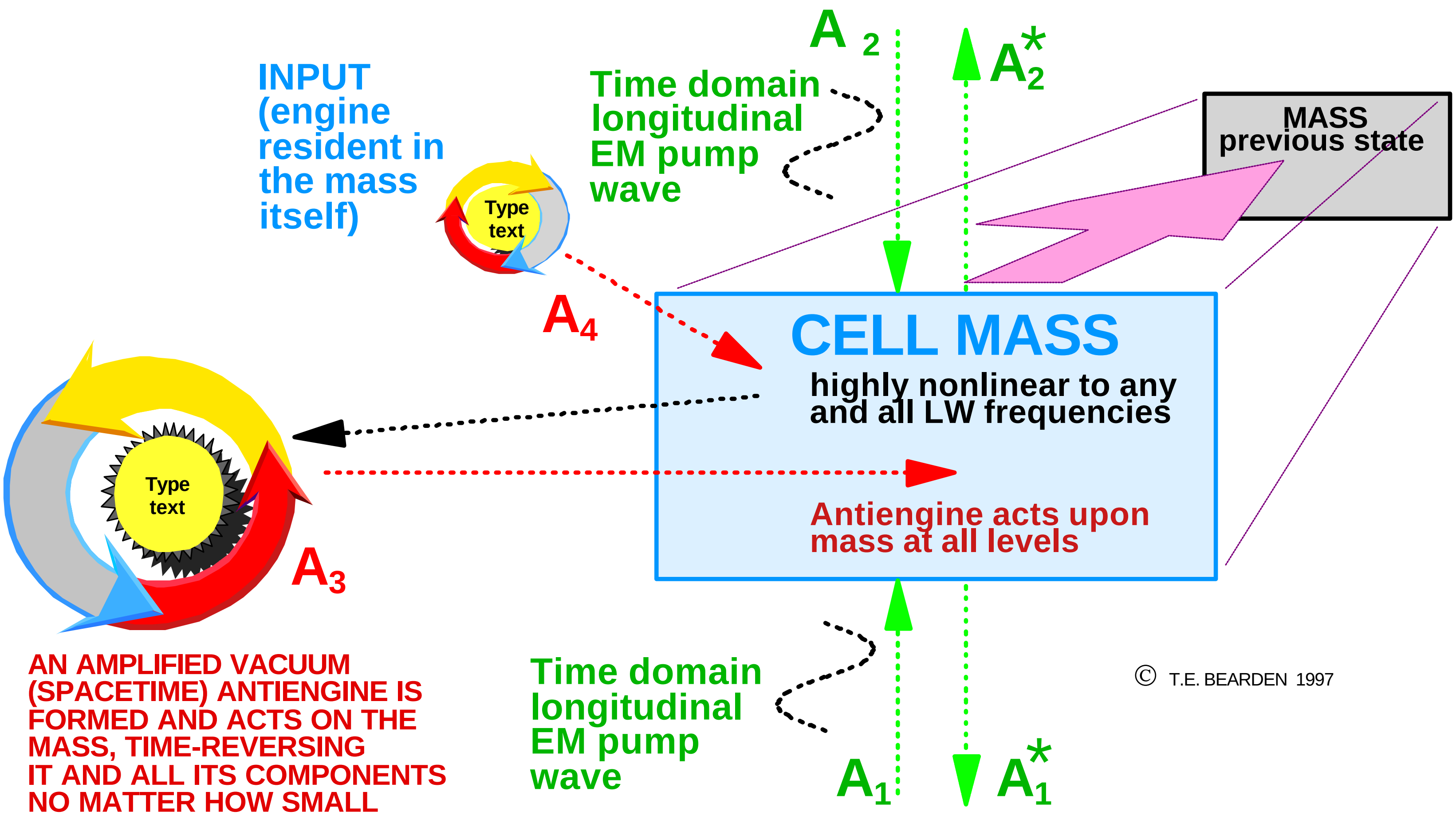

INPUT (engine resident in the mass itself)

Time domain longitudinal EM pump wave

MASS previous state

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

Time domain longitudinal EM pump wave

© T.E. BEARDEN 1997

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

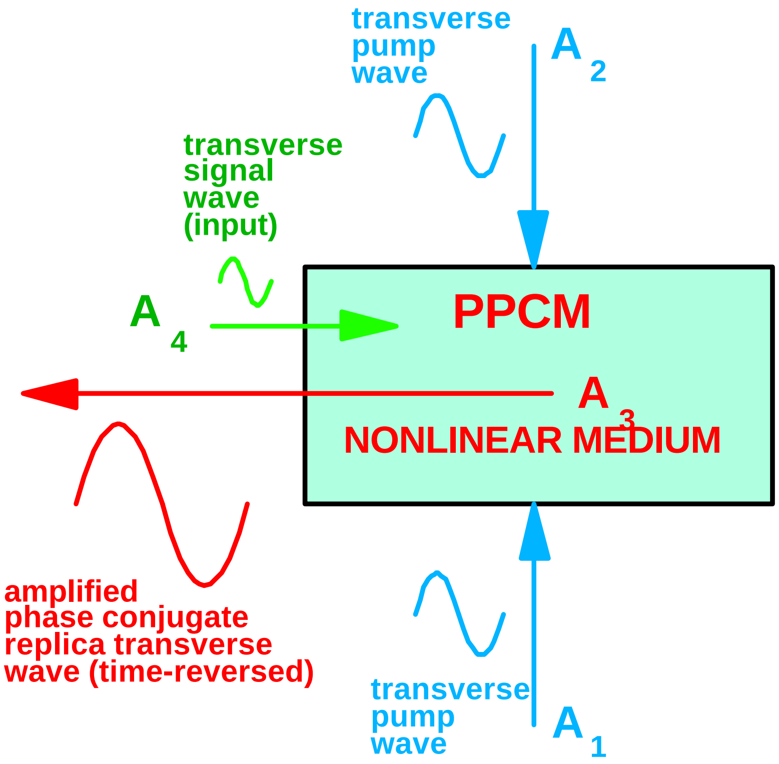

transverse pump wave

A_2

transverse signal wave (input)

A_4

PPCM

A_3

NONLINEAR MEDIUM

amplified phase conjugate replica transverse wave (time-reversed)

transverse pump wave

A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

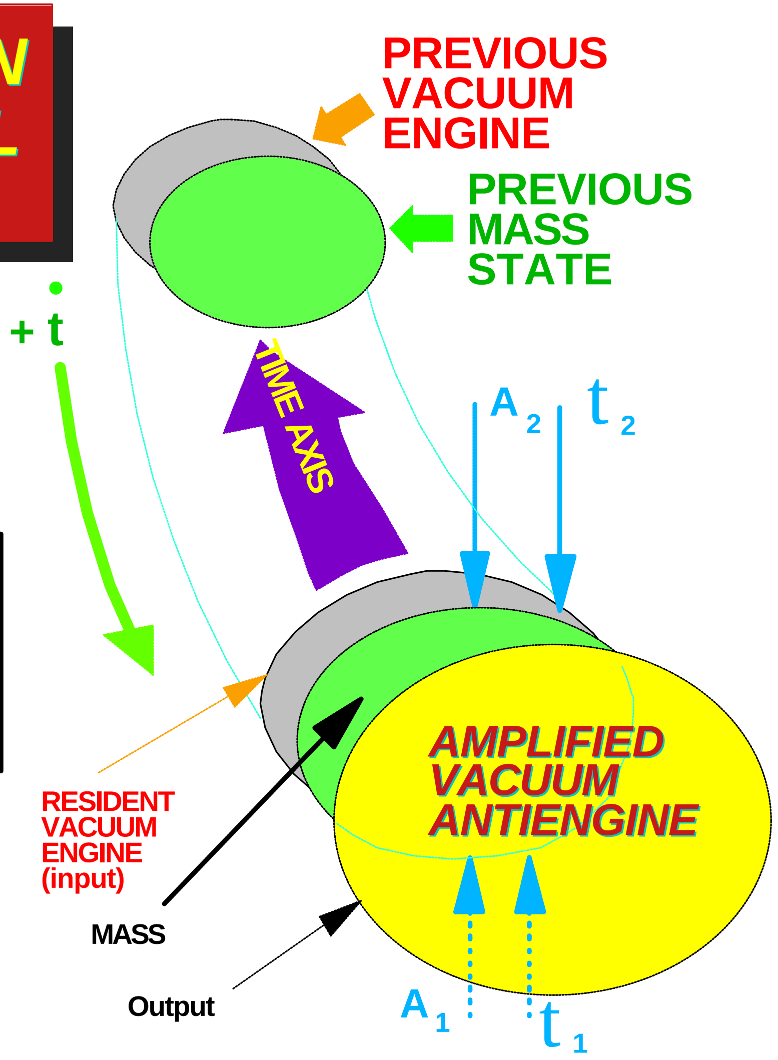

PREVIOUS VACUUM ENGINE

PREVIOUS MASS STATE

+ t

TIME AXIS

A_2 t_2

RESIDENT VACUUM ENGINE (input)

MASS

Output

AMPLIFIED VACUUM ANTIENGINE

A_1 t_1

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Time-reversal of a mass to a new physical state not in its past

- Original engine A'
- New engine B' added
- Time-pumping moves to new "past" that is determined by (A'+B')

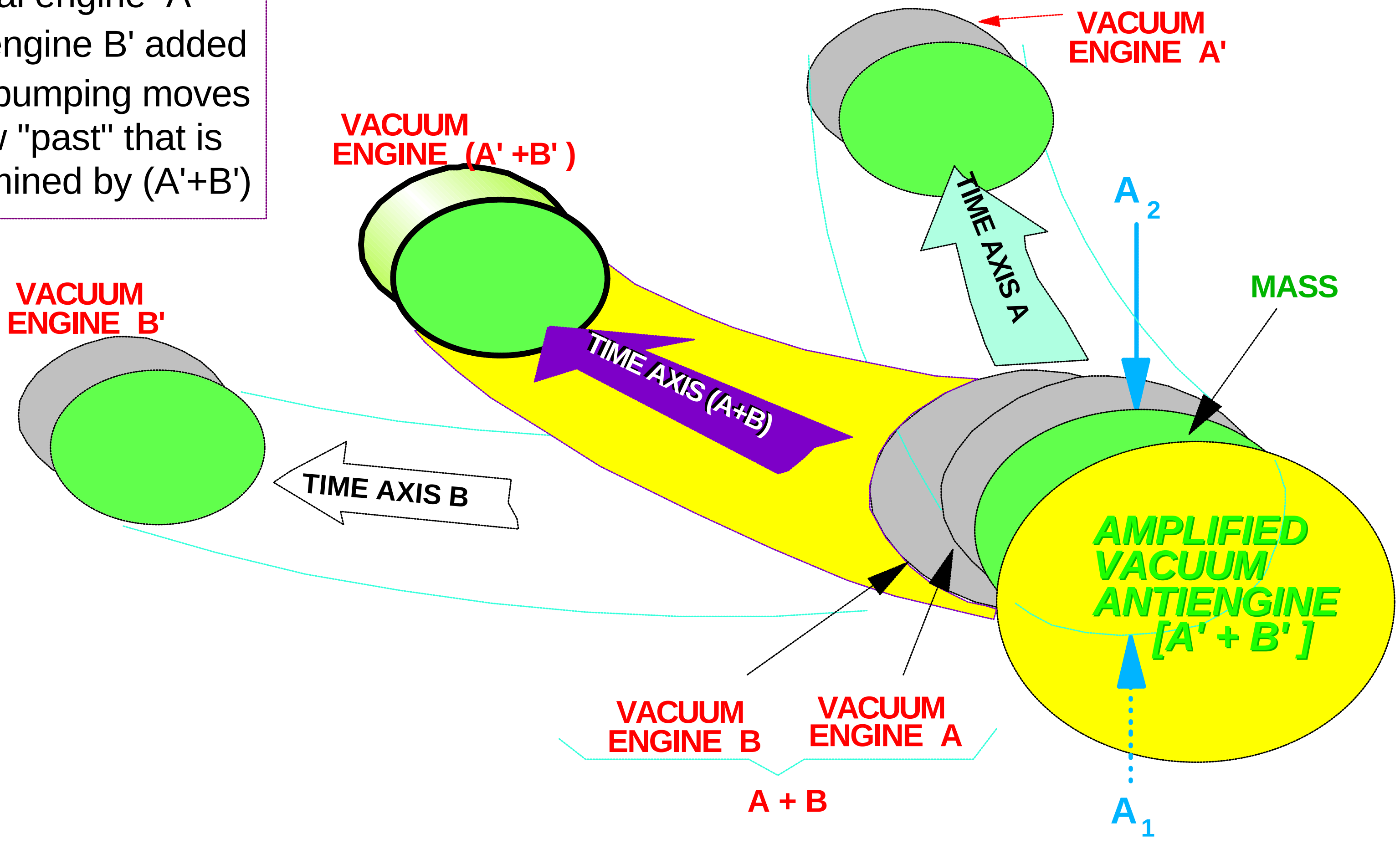
VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
TIME AXIS (A+B)
TIME AXIS B
A_2
MASS
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A + B
A_1

Extended general relativity principle

- All levels of energy structures and all levels of time structures mold spacetime geometry.
- The "pattern" is called a *template* for a vacuum engine (spacetime curvature engine).
- Spatial and temporal structurings of the engine act upon any exposed mass at al levels.
- This produces a template of forces, for precise translations and stresses within the mass at all levels. The forces exist in both time and 3-space.

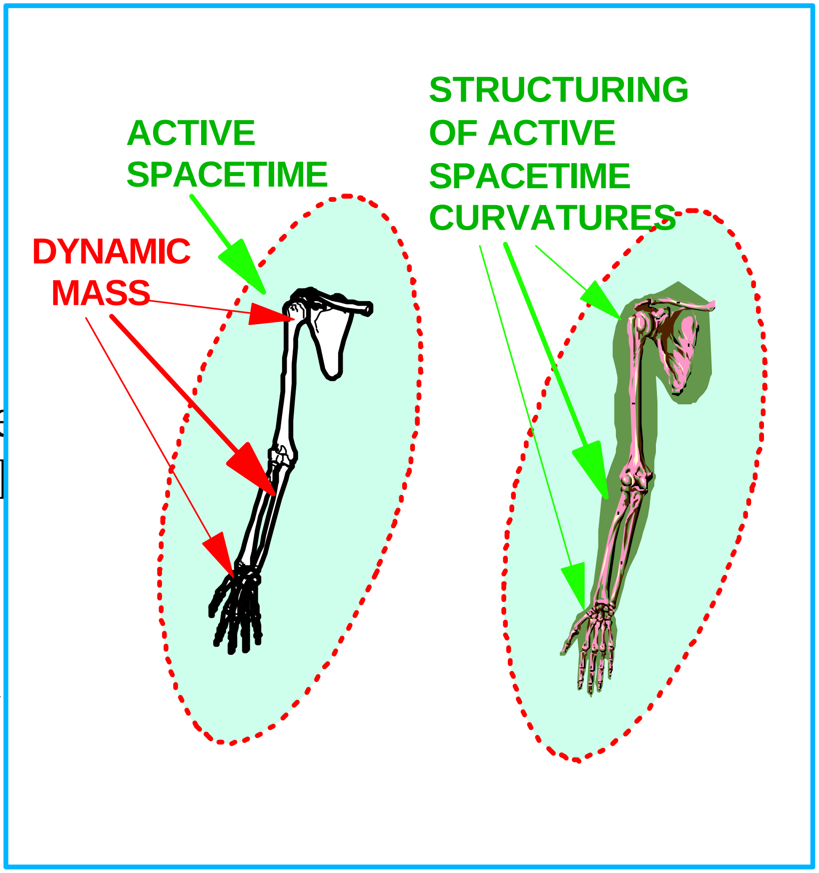
ACTIVE SPACETIME
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DYNAMIC MASS

Engines: Working spacetime demons

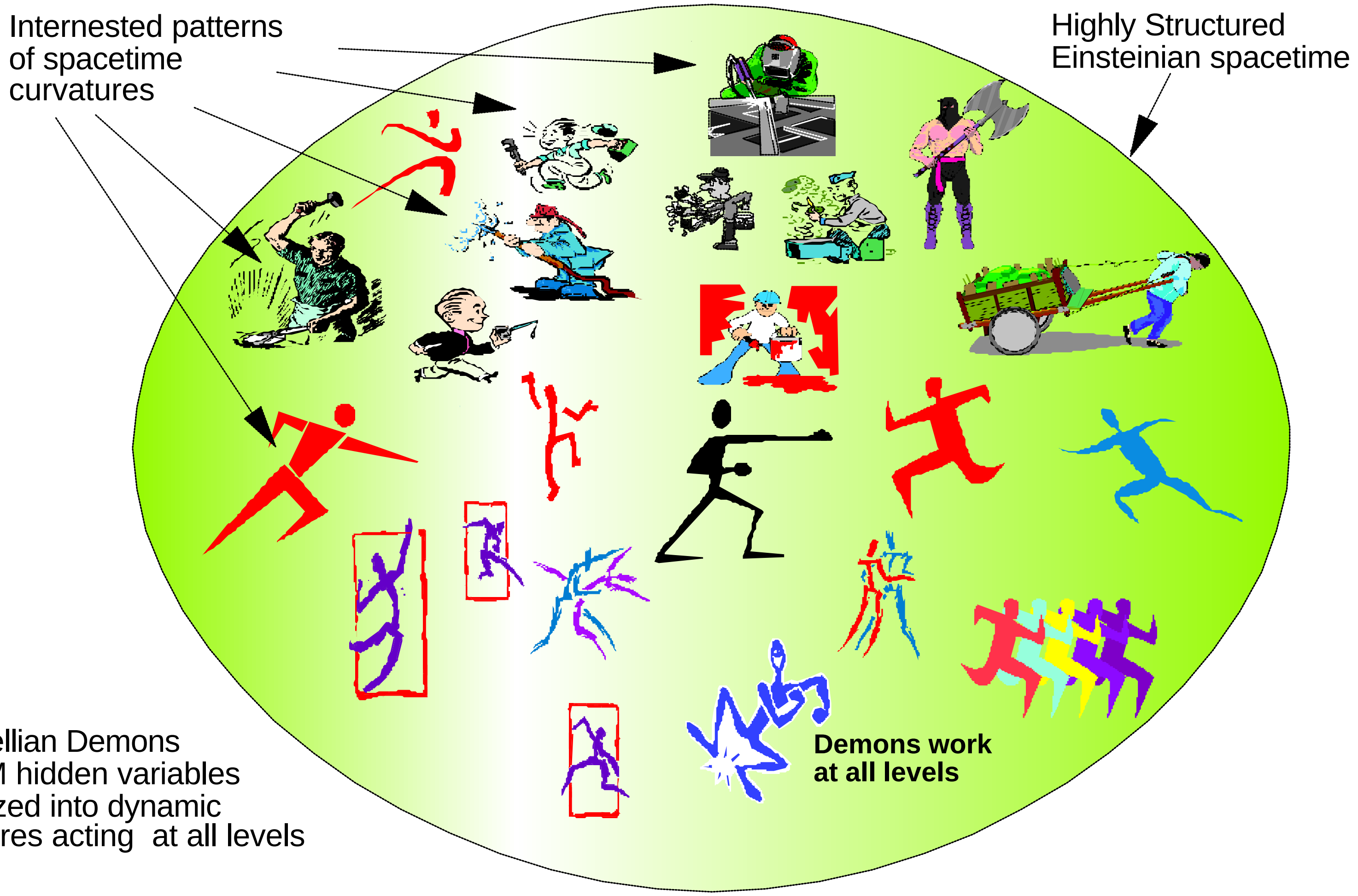

Internested patterns of spacetime curvatures

Highly Structured Einsteinian spacetime

Maxwellian Demons are EM hidden variables organized into dynamic structures acting at all levels

Demons work at all levels

Signal vs. vacuum engine

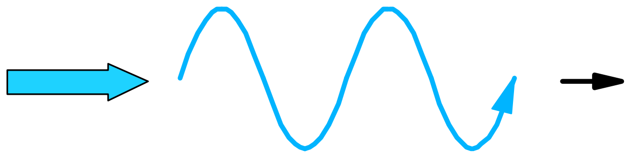

SIGNAL (INFO)

★ Information received
★ Any overt physical action must be taken by cell itself
★ Cell must furnish the energy for any action it takes
★ Passes from outside to in

A. Cell must do the action itself; vacuum energy exchange is passive. Energy or fuel for doing the action must be added to cell externally.

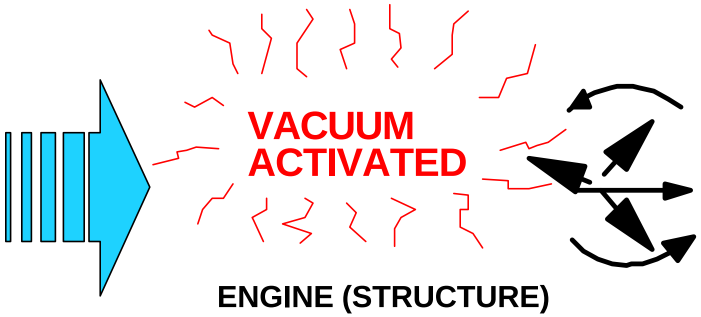

ENGINE (STRUCTURE)
TEMPLATE (FORM)

★ Information itself acts
★ Local spacetime curved, with internal subcurvatures
★ Produces template of forces in local spacetime (ST)
★ Arises from within local ST
★ Acts on and changes all parts of cell, including its genetics

B. Cell is acted upon; vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to cell externally.

Unified field theory requires modeling and use of supersystem

Three components interact with each other:

- System and its dynamics
- Local active vacuum and its dynamics
- Local curvatures of spacetime and their dynamics.

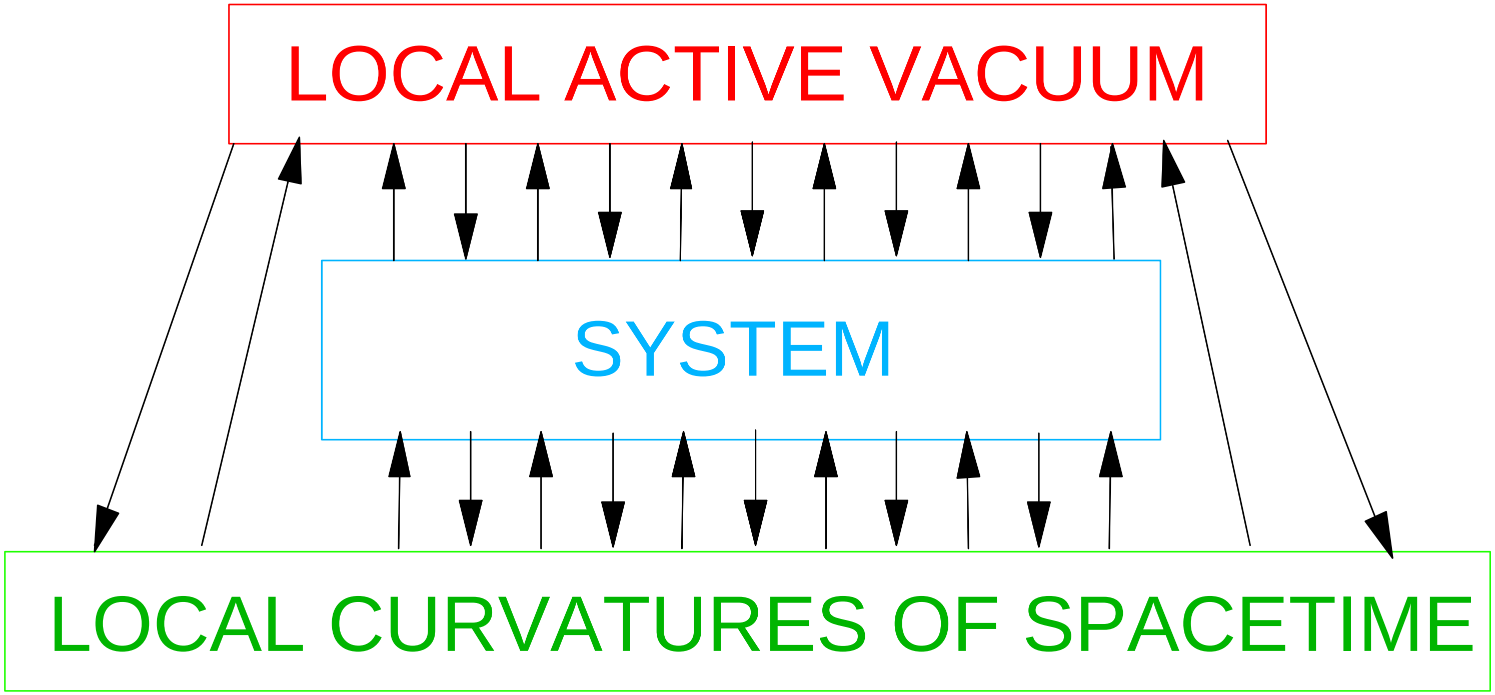
LOCAL ACTIVE VACUUM
SYSTEM
LOCAL CURVATURES OF SPACETIME

Classical EM kills the supersystem, prevents supersystem engineering

- Assumes flat local spacetime
- Assumes inert vacuum

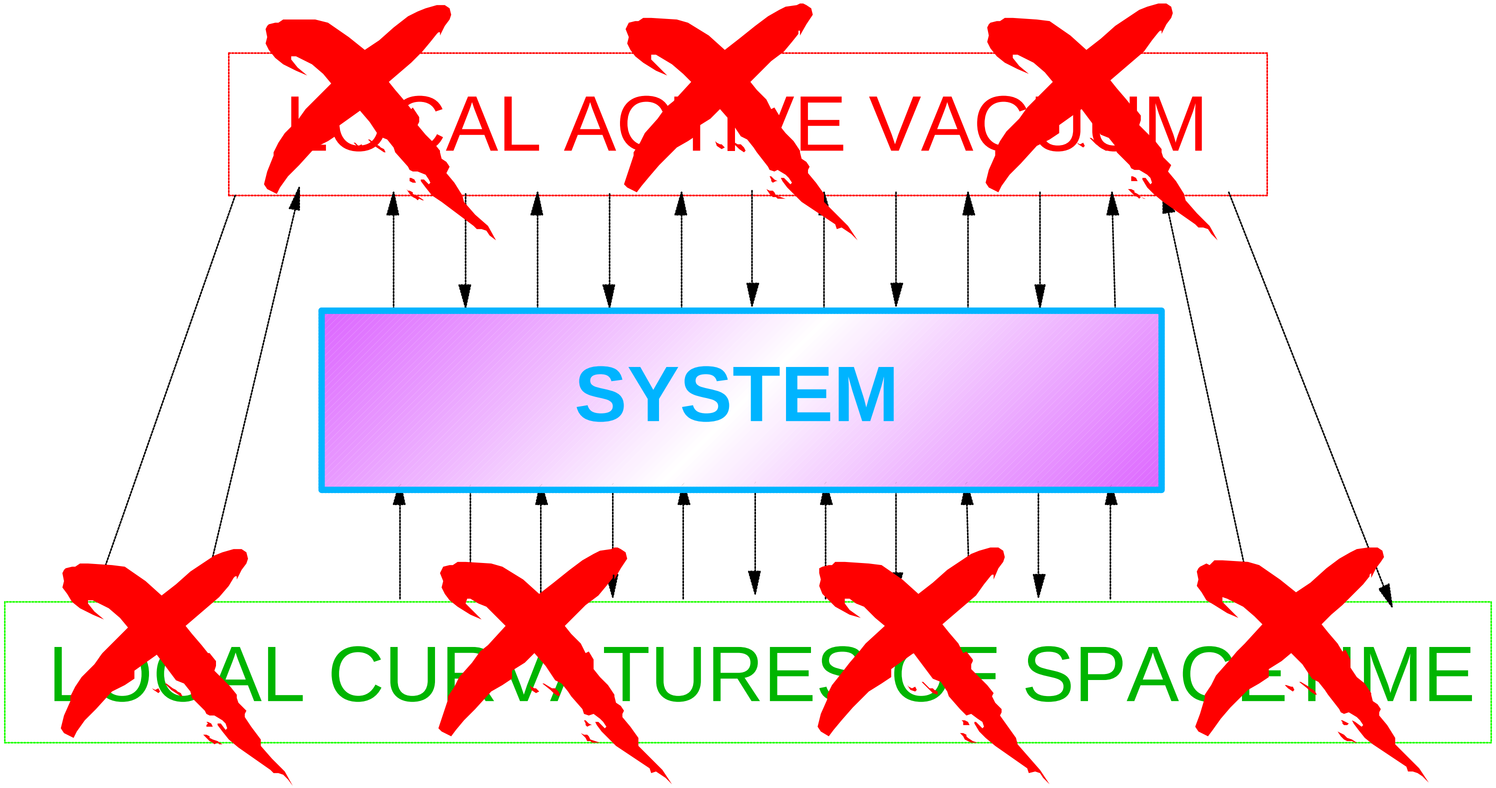

LOCAL ACTIVE VACUUM

SYSTEM

LOCAL CURVATURES OF SPACETIME

Malignant teratoma exhibits both time-reversal and fast-forwarding

- A teratoma cell is a tumor cell that has reverted to stem cells, then proceeds to differentiate toward new kinds of normal cells.
- May produce hair, skin, muscle, heart, bone, and other type cells.
- Sometimes present at birth.
- Demonstrates both "time-reversal" and "time-forwarding".
- The missing key is the engine function(s).

Miller-Fox-Urey biogenesis experiments

- Engines of and from all life forms ever on earth or in the universe exist in the ambient vacuum potential.
- The Miller-Fox-Urey experiments were not in a "sterile" environment.
- Heat (IR) and UV constitute (1) a harmonic interval with a difference frequency, and (2) amplifying by the process previously shown.
- These experiments "kindled" previously living forms from the available gases etc. by amplifying the ancient living engines from their virtual state to the observable state, so that the masses were molded into the physical states of their templates.

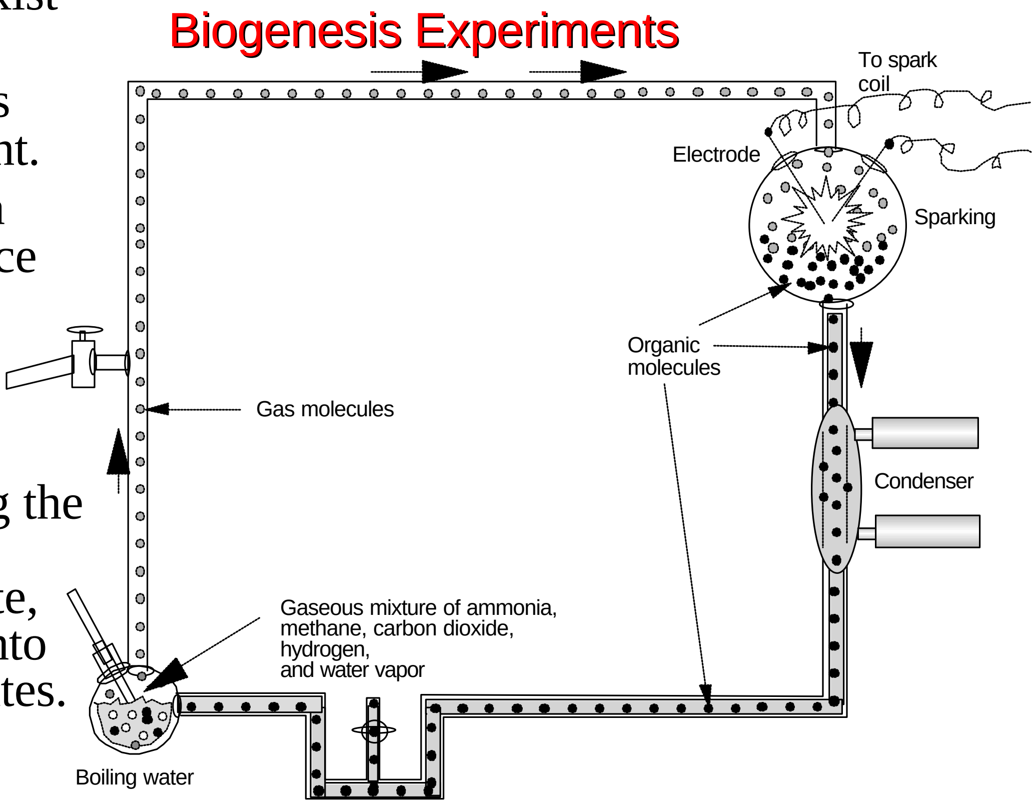
Biogenesis Experiments
To spark coil
Electrode
Sparking
Organic molecules
Gas molecules
Condenser
Gaseous mixture of ammonia, methane, carbon dioxide, hydrogen, and water vapor
Boiling water

Cloning to produce stem cells

- Human DNA can be inserted in an animal cell, and the cell chemically tricked into proceeding as if fertilized.

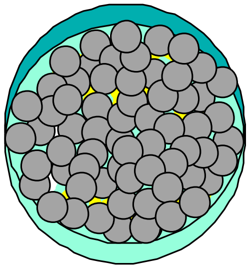

- Formation, increase, and growth of stem cells begins and continues.
- The "engine" carried by the human DNA is changing the matter available into human cells.
- readily demonstrated experimentally.

Regeneration in a strain of rat with part of its immune system missing

- In mammals, the immune system plays a role in reducing the regeneration capability, compared to that of more primitive systems.
- In a strain of rat where that "suppressive" part of the immune system is missing, the regenerative capability is dramatically increased.
  - Can regrow cut away sections of tail.
  - Regrows plug cut out of ear.
  - Reconnects severed optic nerve.
  - Partially restores severed spinal cord.

- The suppressive "delta" engine of the immune system, operating simultaneously with the underlying primitive regenerative engine, sums to the modern much-reduced regenerative engine in the mammal.

An observation for the future

- When the science of engines is developed, physicians will be able to tailor these engine combinations to restore severed human spines, regrow severed limbs, etc.
- They will also be able to direct the body's "normal growth" to reduce and direct birth defects, genetic defects, etc.

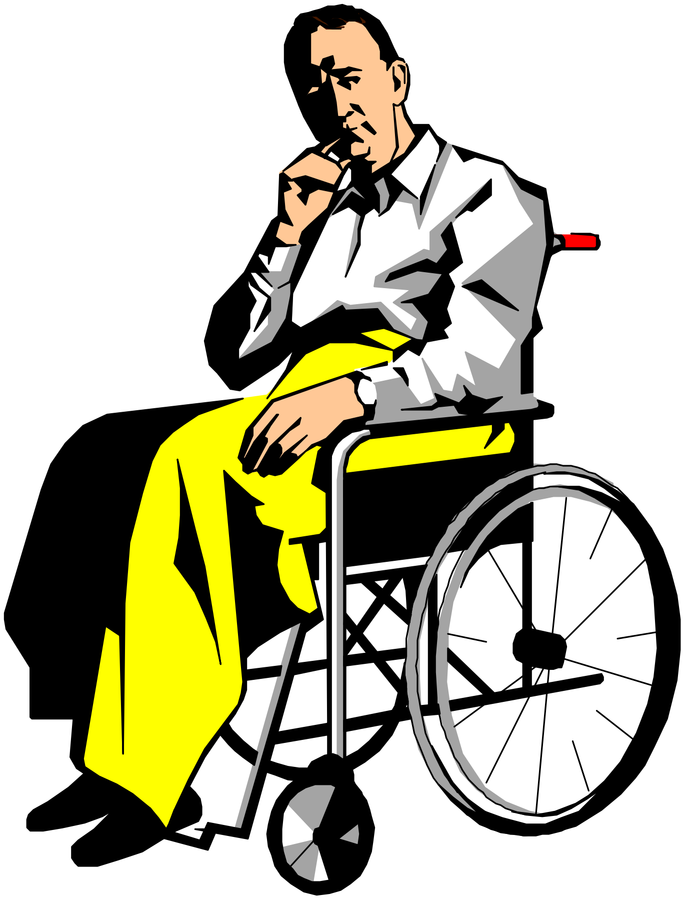

Biochemistry and EM models

- Medical scientists analyze such things as teratomas in terms of biochemistry and the biochemical model.
- An assembly of charges and charge distributions, with their associated fields and potentials, drives all chemistry at base level.
- The charge assembly and its dynamics involves a corresponding assembly of spacetime curvatures and their dynamics. These are *the cause that is driving the charge assembly and its dynamics, and therefore the chemistry*..
  - In unified field theory (such as that of Sachs) and O(3) electrodynamics.
  - Not in classical U(1) EM theory or quantum electrodynamics.
- Conventionally a causative EM field in space is "defined" as the *diversion from* that spatial field by and after its interaction with charge).
  The "definition" substitutes the effect for the cause, a non sequitur.
- This is a known major flaw, pointed out in one fashion or another by many physicists such as Wheeler and Nobelist Feynman.
- Classical EM theory is a 137-year-old model, seriously flawed. These flaws emerge in biochemistry and medical science.

Necessary EM corrections*

- Many of the corrections necessary in electrodynamics modeling either have been done or are underway.
- Higher symmetry, non-Abelian electrodynamics models have long been developed in particle physics.
- The new approach extends these models to also incorporate general relativity, thus producing a unified field theory*.
- In the new approach, spacetime curvature engines are perfectly natural and understandable, and can be directly modeled and engineered by advanced theorists.
- In short, the theory is sufficiently advanced and robust that technology and engineering can now advance.
- This will be a great leap forward for biochemistry and medical science, particularly in therapeutic methods.

* Sachs's unified field theory implemented by O(3) electrodynamics presently is the most suitable to use. Numerical methods are required.

Prioré's revolutionary results in lab animals

And its ramifications

Prioré (right) makes a point to Chaban-Delmas, Mayor of Bordeaux

- Many of the Prioré team's experiments were done in Bordeaux.
- Antoine Prioré patents:
  - "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968.
  - "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966.
  - "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963.

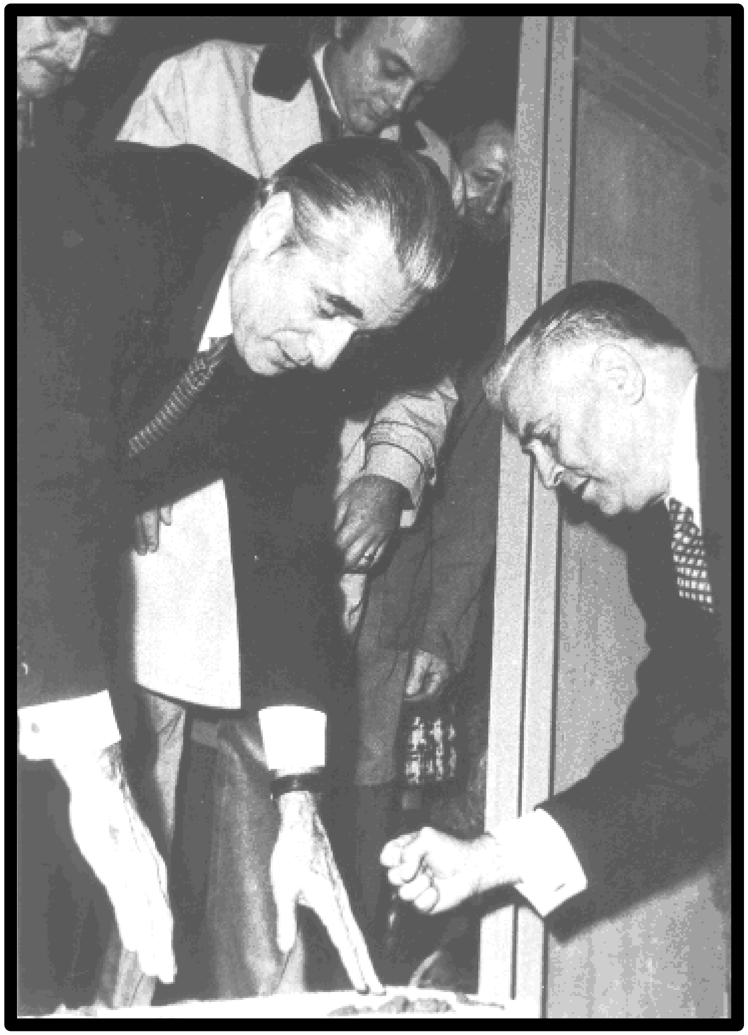

Block diagram of Prioré's method

Time-reversing the cells back to normal state

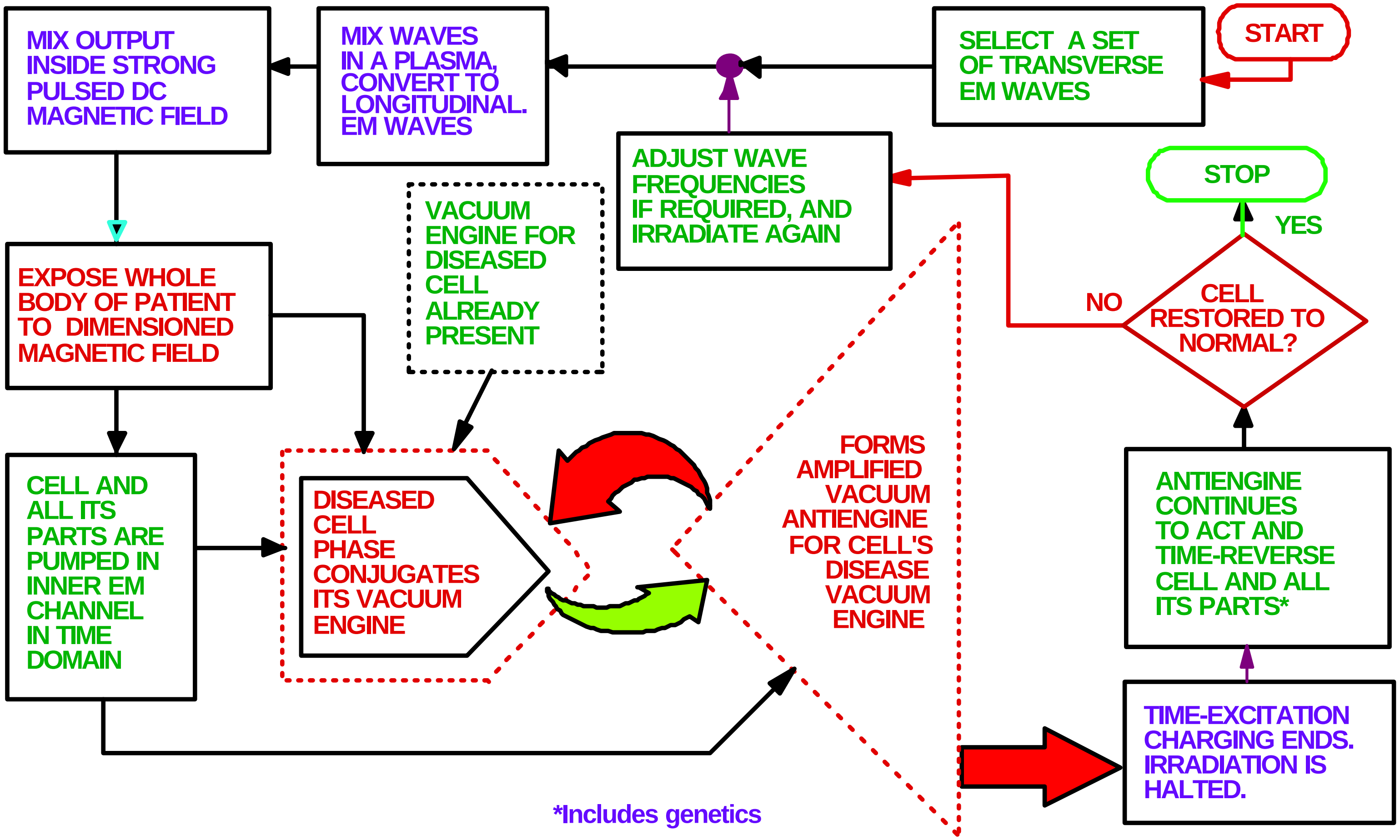

START

SELECT A SET OF TRANSVERSE EM WAVES

MIX WAVES IN A PLASMA, CONVERT TO LONGITUDINAL. EM WAVES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.

ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*

CELL RESTORED TO NORMAL?

YES

STOP

NO

ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN

*Includes genetics

In Laboratory Animals:

Prioré Team Cured

- **Terminal malignant tumors**
- **Clogged arteries (atheriosclerosis)**
- **Suppressed Immune Systems**
- **Infectious trypanosomas**
- ***Results were presented to French Academy of Science by Robert Courrier, Director of the Biology Section of the Academy***
- ***No one understood the mechanism!***

© 1999 T. E. BEARDEN

Some peculiar results by the Prioré team

- Once a rat is cured of a particular disease:
  - A single drop of blood from the cured rat, injected into another rat stricken with same disorder, would result in the injected rat getting well slowly.
  - This ability to transfer immunity and healing in the extracted drop of blood gradually decayed over a period of a few weeks.
- Showed that even the fields and potentials in the blood had been altered internally, in the treated animal.
- Showed that the injected animal's cellular regenerative system possessed the ability to also time-domain pump over the entire body, everywhere reducing the delta between that injected drop of blood now the rat's own, and the rest of the blood in its body.

Proof of cellular time-reversal

- Pautrizel compared identical treatment of immature rats with immature immune systems, and mature rats with mature immune systems.
- Treatment restored the damaged mature system back to full mature functioning, and it promptly recognized and dispatched the pathogens.
- Treatment restored the damaged immature system back to full immature functioning, unable to cope, and the pathogens promptly reinfected and killed the immature rat.
- The lesson: The body is indeed returned to a previous physical state. The pathogens themselves are not killed by the treatment, but by restoring the ability of the immune system to recognize and destroy them.
- Only the *previous ability* of the immune system is restored by the process.

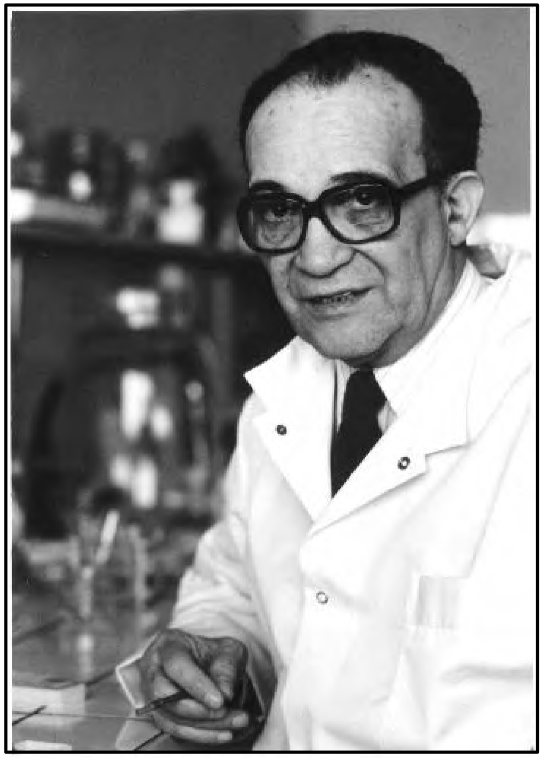

PROF. RAYMOND PAUTRIZEL
(Renowned Parasitologist)

Disadvantages of Prioré method

- Large plasma tube and facility required.
  - 3-stories high building bay if to treat humans.
  - Long irradiation time (2-3 hours).
  - Slow, numerous experimental adjustments.
- Fundamental mechanism was unknown.
  - Phase conjugate optics not yet born in West.
  - No higher symmetry EM modeling used.
  - Prioré thought it was a result of ionization.
  - Importance of difference frequency unknown.
  - Body dielectric transmission path unknown.

3-Story high Prioré device for treating humans

- At the end of the project, Prioré had developed a large unit to treat humans.
- A few were treated.
- Cures replicated what had been done in lab animals
- The massive size of the unit and long treatment time (hours) made the process bulky and very inconvenient -- and also expensive

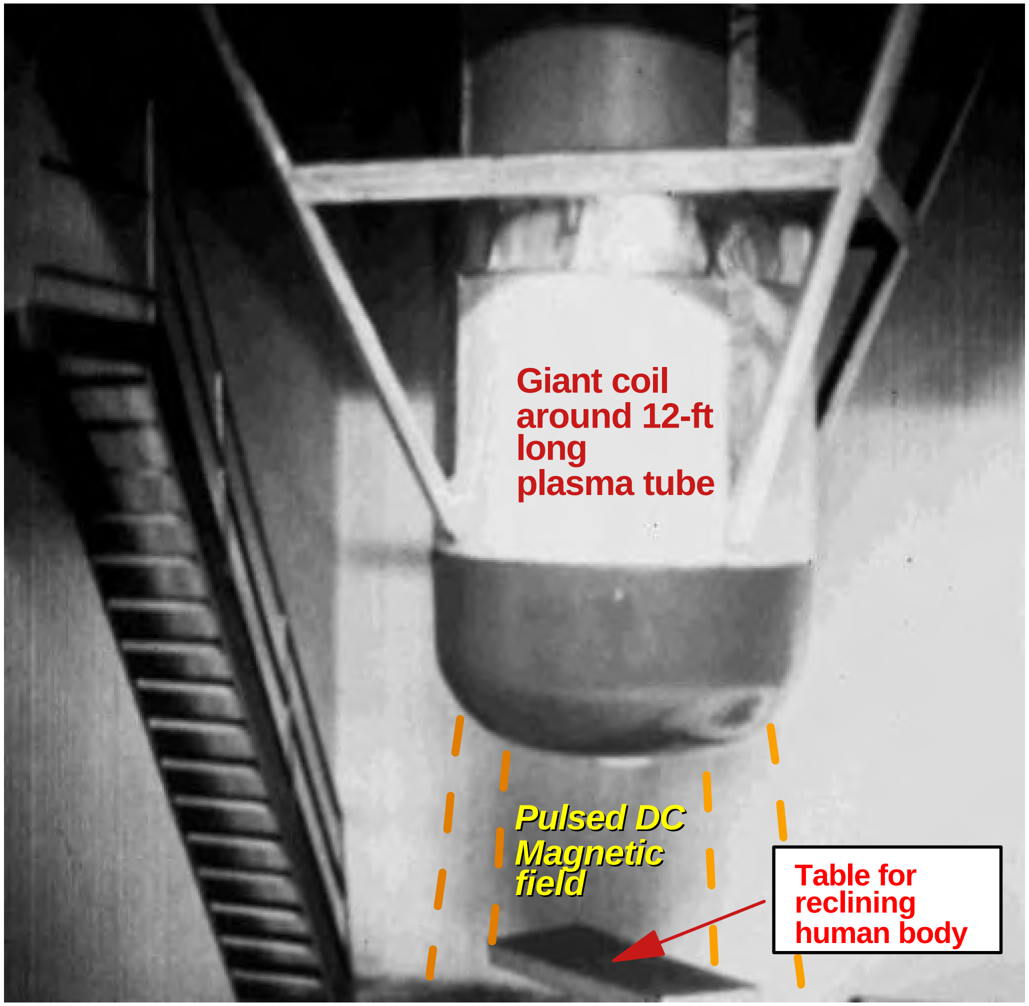
Giant coil around 12-ft long plasma tube
Pulsed DC Magnetic field
Table for reclining human body

17-foot Prioré plasma tube

Used in large machine (3-stories high) to treat a limited number of human patients

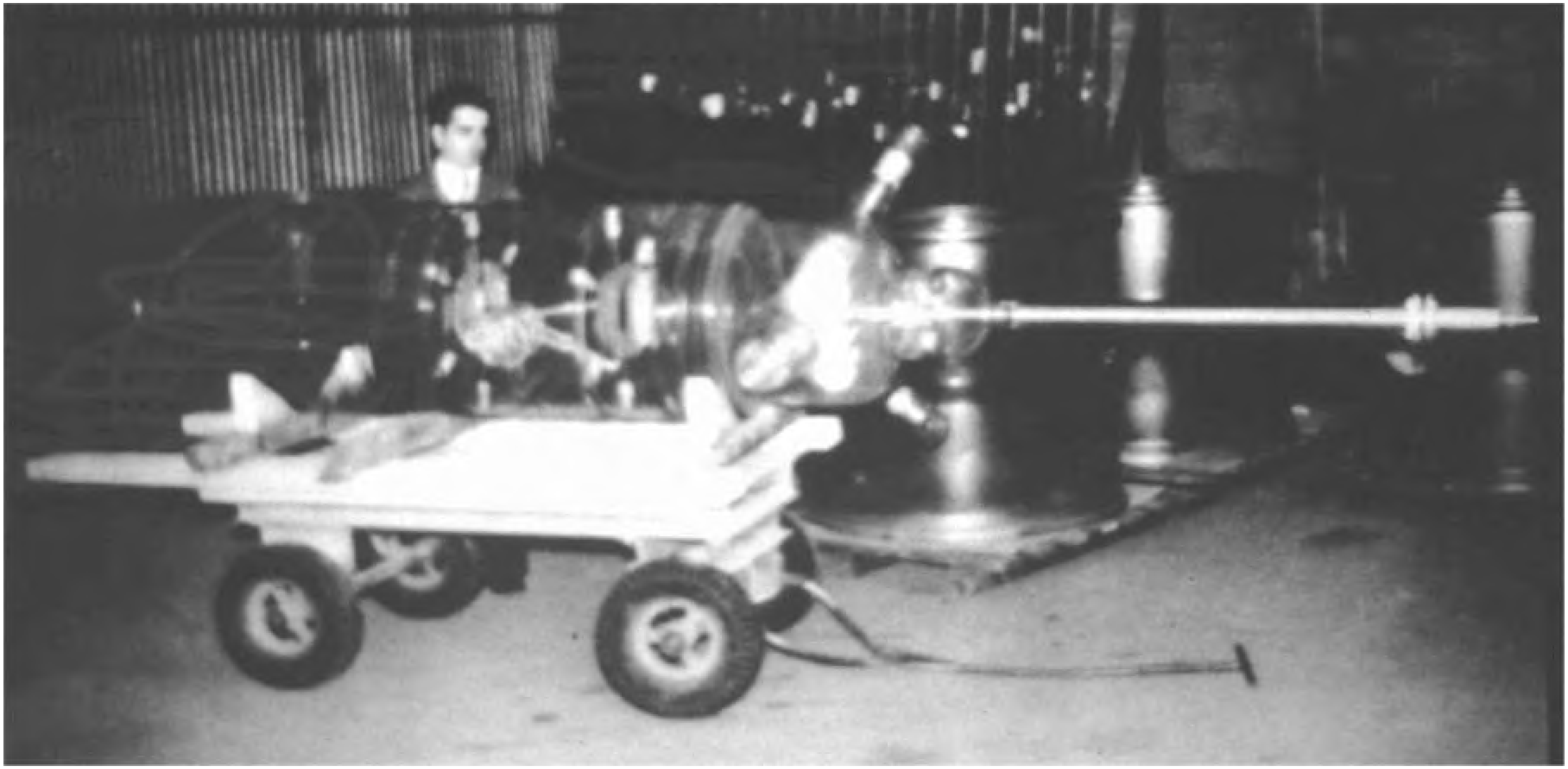

A proposed new mechanism for cancer promotion

And its ramifications

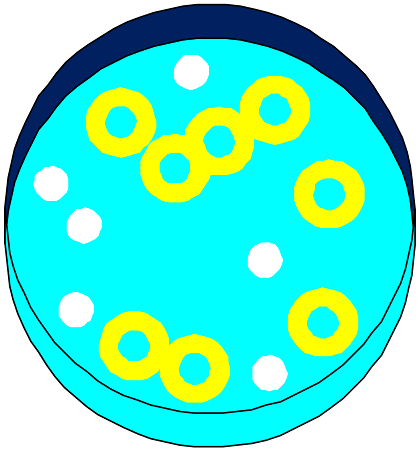

Remarks on cancer from a new perspective

- Two stages: cellular damage and promotion of the damaged cell to a tumorous cell. Promotion presently is poorly understood if at all.
- A master cellular control system (MCS) (studied by Popp et al.) controls both the immune system and the cellular regenerative system, as well as normal functions of the cells.
- The "past history" of ancestral development of aerobic cells and higher multicell aerobic bodies on planet Earth, resides in the "internested engine layers" in every tiniest part of the entire body dielectric, and in all cells.

Remarks on cancer from a new perspective (2)

- Faced with cell damage (such as sustained hypoxia) it cannot handle, the MCS continues to "peel the layers" in taking actions related to ever more distant pasts.
- Faced with continuing hypoxia, e.g., the MCS eventually reaches ancestral layers -- with an engine -- it forces the damaged cells to start moving back toward the original primeval anaerobic form.
- The first step backward is *separation from central growth control*, freeing the cell to multiply independently. At the same time, genetic changes in the cell are forced and produced, starting it back toward becoming an anaerobe.
- These actions comprise the *promotion process*.

The promotion process via engines

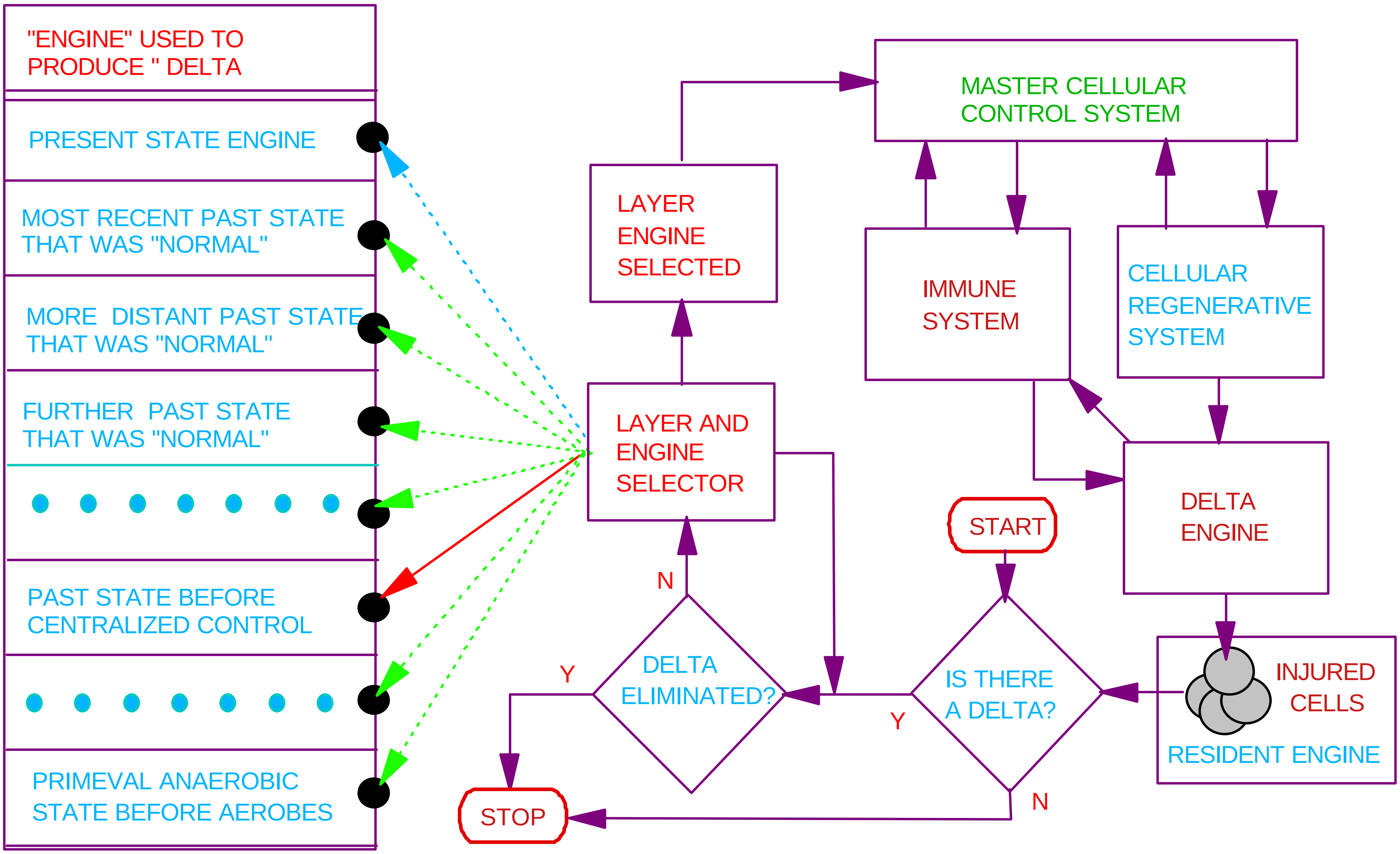

"ENGINE" USED TO PRODUCE " DELTA

PRESENT STATE ENGINE

MOST RECENT PAST STATE THAT WAS "NORMAL"

MORE DISTANT PAST STATE THAT WAS "NORMAL"

FURTHER PAST STATE THAT WAS "NORMAL"

● ● ● ● ● ● ●

PAST STATE BEFORE CENTRALIZED CONTROL

● ● ● ● ● ● ●

PRIMEVAL ANAEROBIC STATE BEFORE AEROBES

MASTER CELLULAR CONTROL SYSTEM

LAYER ENGINE SELECTED

IMMUNE SYSTEM

CELLULAR REGENERATIVE SYSTEM

LAYER AND ENGINE SELECTOR

DELTA ENGINE

START

N

Y

DELTA ELIMINATED?

IS THERE A DELTA?

Y

N

INJURED CELLS

RESIDENT ENGINE

STOP

Evolution of aerobic systems

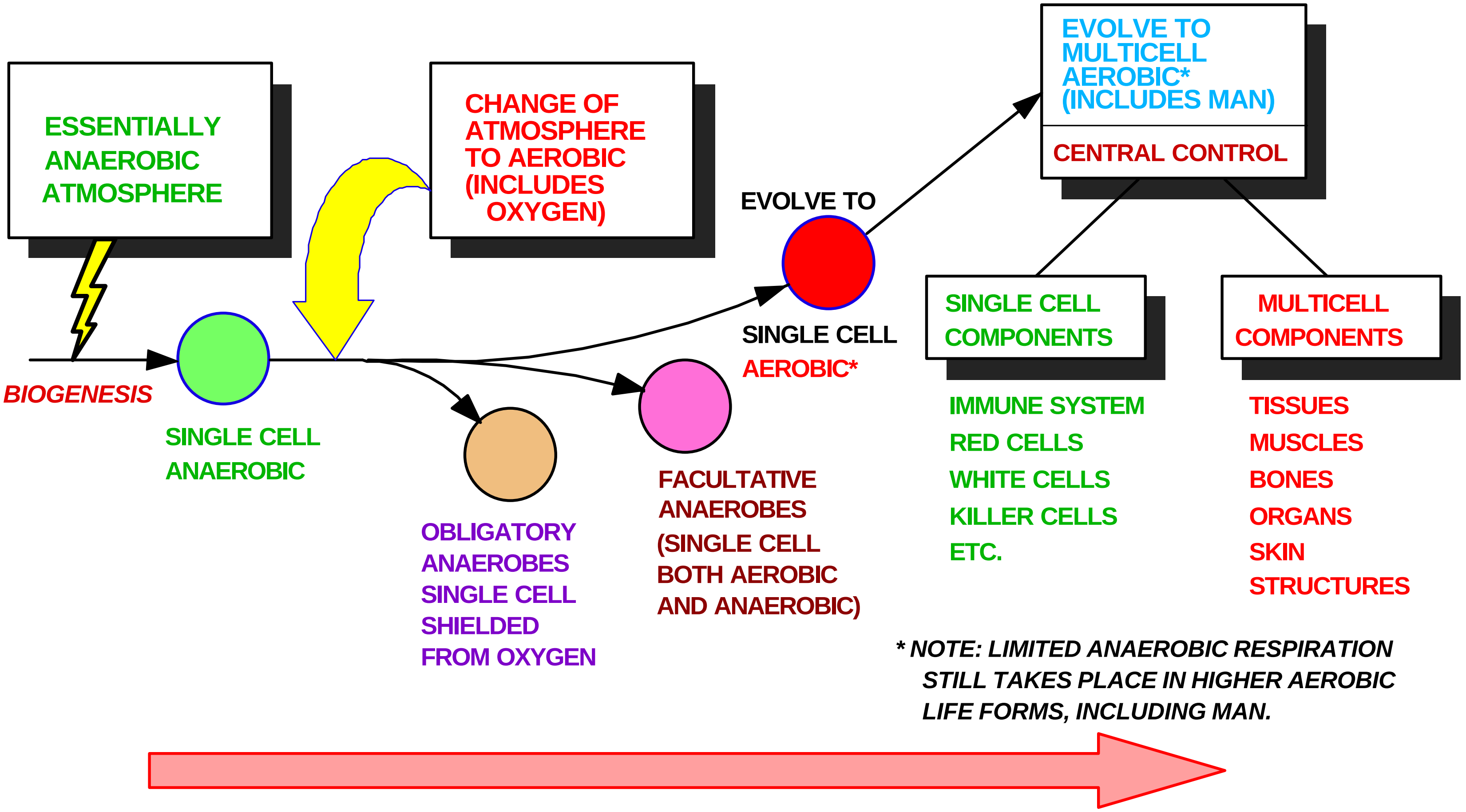

ESSENTIALLY ANAEROBIC ATMOSPHERE

BIOGENESIS

SINGLE CELL ANAEROBIC

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

First step in self-promotion of cancer due to hypoxia

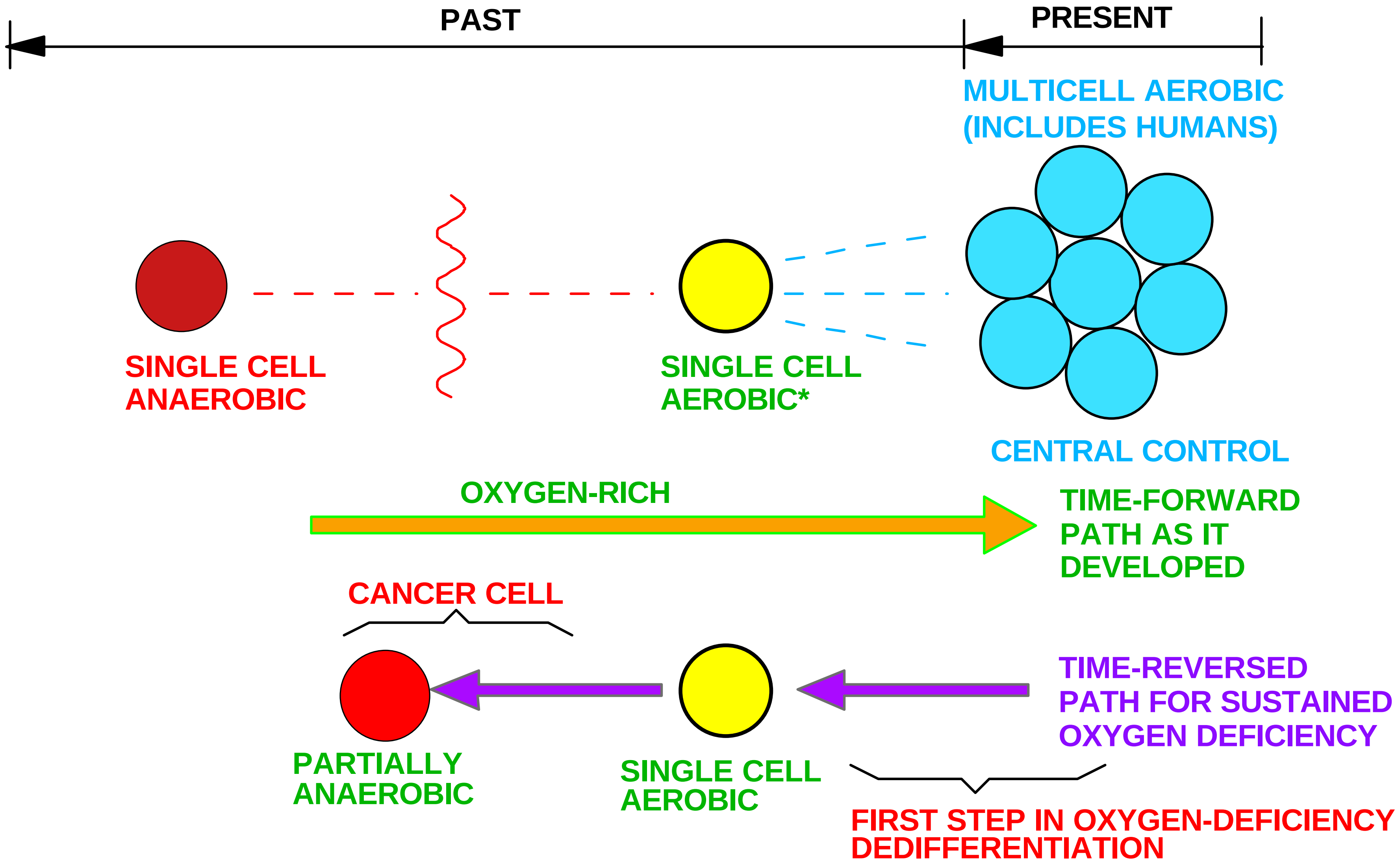

PAST

PRESENT

MULTICELL AEROBIC
(INCLUDES HUMANS)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

Block diagram of the cellular regenerative system

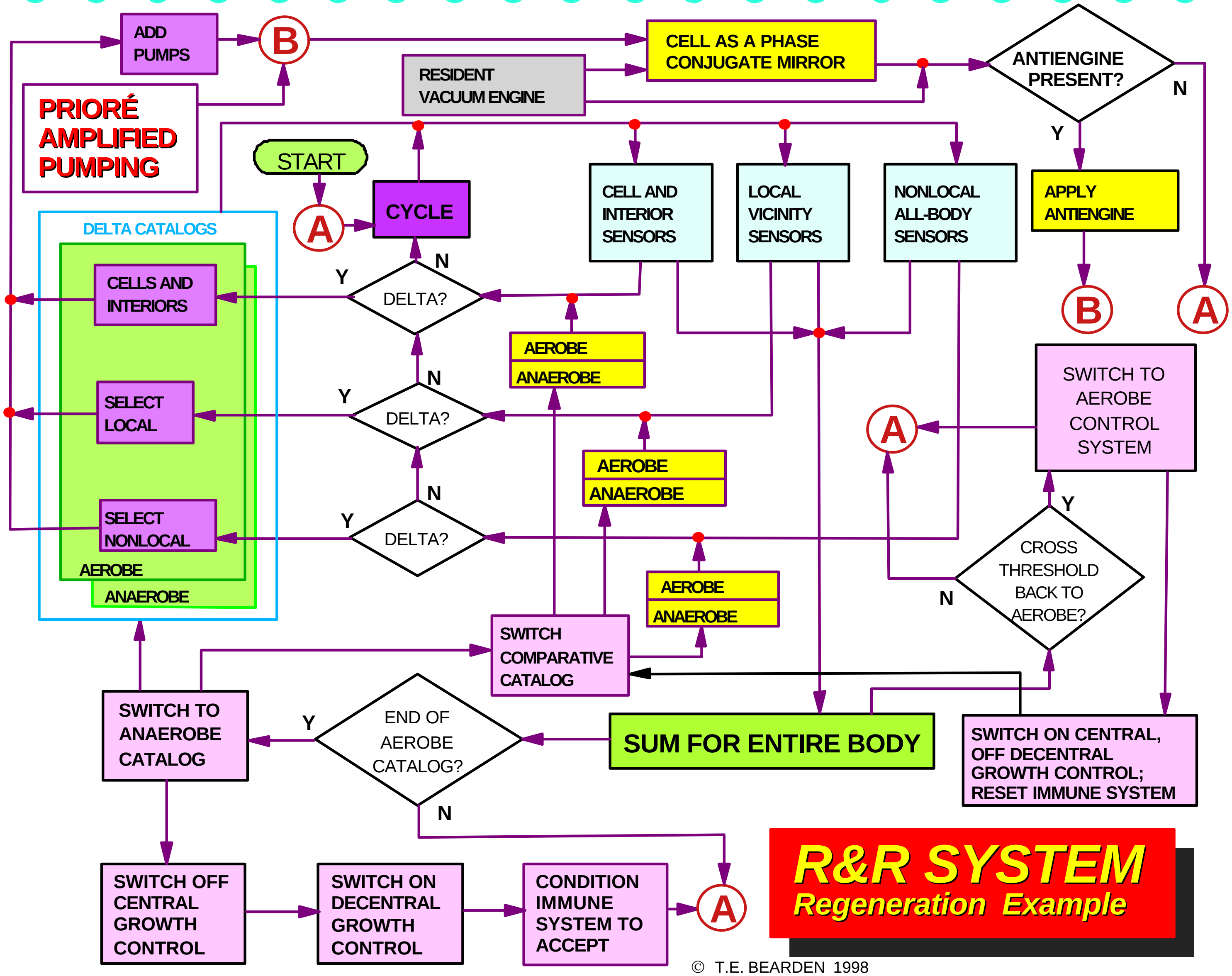

ADD PUMPS

B

CELL AS A PHASE CONJUGATE MIRROR

ANTIENGINE PRESENT?

N

Y

RESIDENT VACUUM ENGINE

PRIORÉ AMPLIFIED PUMPING

START

A

CYCLE

CELL AND INTERIOR SENSORS

LOCAL VICINITY SENSORS

NONLOCAL ALL-BODY SENSORS

APPLY ANTIENGINE

B

A

DELTA CATALOGS

CELLS AND INTERIORS

SELECT LOCAL

SELECT NONLOCAL

AEROBE

ANAEROBE

Y

N

DELTA?

Y

N

DELTA?

Y

N

DELTA?

AEROBE

ANAEROBE

AEROBE

ANAEROBE

AEROBE

ANAEROBE

SWITCH TO AEROBE CONTROL SYSTEM

A

Y

CROSS THRESHOLD BACK TO AEROBE?

N

SWITCH COMPARATIVE CATALOG

SWITCH TO ANAEROBE CATALOG

Y

END OF AEROBE CATALOG?

N

SUM FOR ENTIRE BODY

SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM

SWITCH OFF CENTRAL GROWTH CONTROL

SWITCH ON DECENTRAL GROWTH CONTROL

CONDITION IMMUNE SYSTEM TO ACCEPT

A

R&R SYSTEM

Regeneration Example

© T.E. BEARDEN 1998

Bombshell: Amplified EM healing

- A unified field theory approach and use of supersystem analysis has broken the basic mechanism for healing itself.
  - The Sachs unified field theory applies and has been partially fitted.
  - Evans-Vigier O(3) electrodynamics can properly model it, including both the internal and external electrodynamics and effects of engines.
  - Direct engineering development is now possible.
- Experimental proof of the results obtainable already exists in the literature.
- A medical revolution is in the offing.

Potential Medical Applications

- Non-traumatic cancer treatment
  - *Cure cancers and leukemias of all types*
  - *Remove pre-cancerous state*
  - *Restore system*
- Treat and cure Infectious diseases
- Cure AIDS and restore immune systems
- Defense against biological warfare
  - *Virulent diseases from unknown agents*
  - *Treat, cure prior to symptoms developing*
  - *Both civil populace and armed forces*
- *Dramatically lower health costs*
- *Rejuvenate the aged*

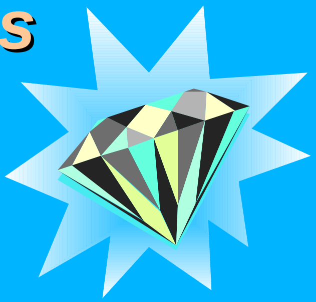

© 1999 T. E. BEARDEN

To treat and save mass casualties: What is urgently needed now:

- A small, portable, relatively inexpensive treatment unit, not requiring plasmas etc.
- Treatment in 2 to 5 minutes rather than 2-hours.
- Simple treatment application, almost anywhere.
- Effective in advanced stages as well as prior to emergence of symptoms.
- Effectiveness against a wide array of illnesses.
- Use is easy, high school student can be trained in 30 minutes.
- Crash program to develop, flood down through emergency systems in great quantities.

The Solution:

- Two blankets have antennas in them.
- Uses the "porthole" concept (discussed shortly)
- Eliminates plasma tube mixing etc.
- Computer-controlled complex signal structure.

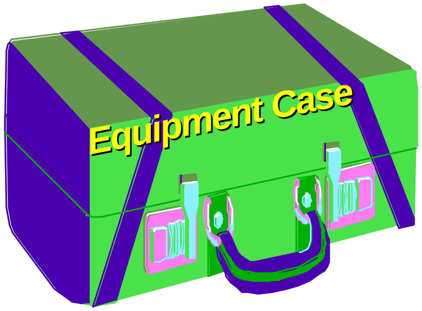
Equipment Case

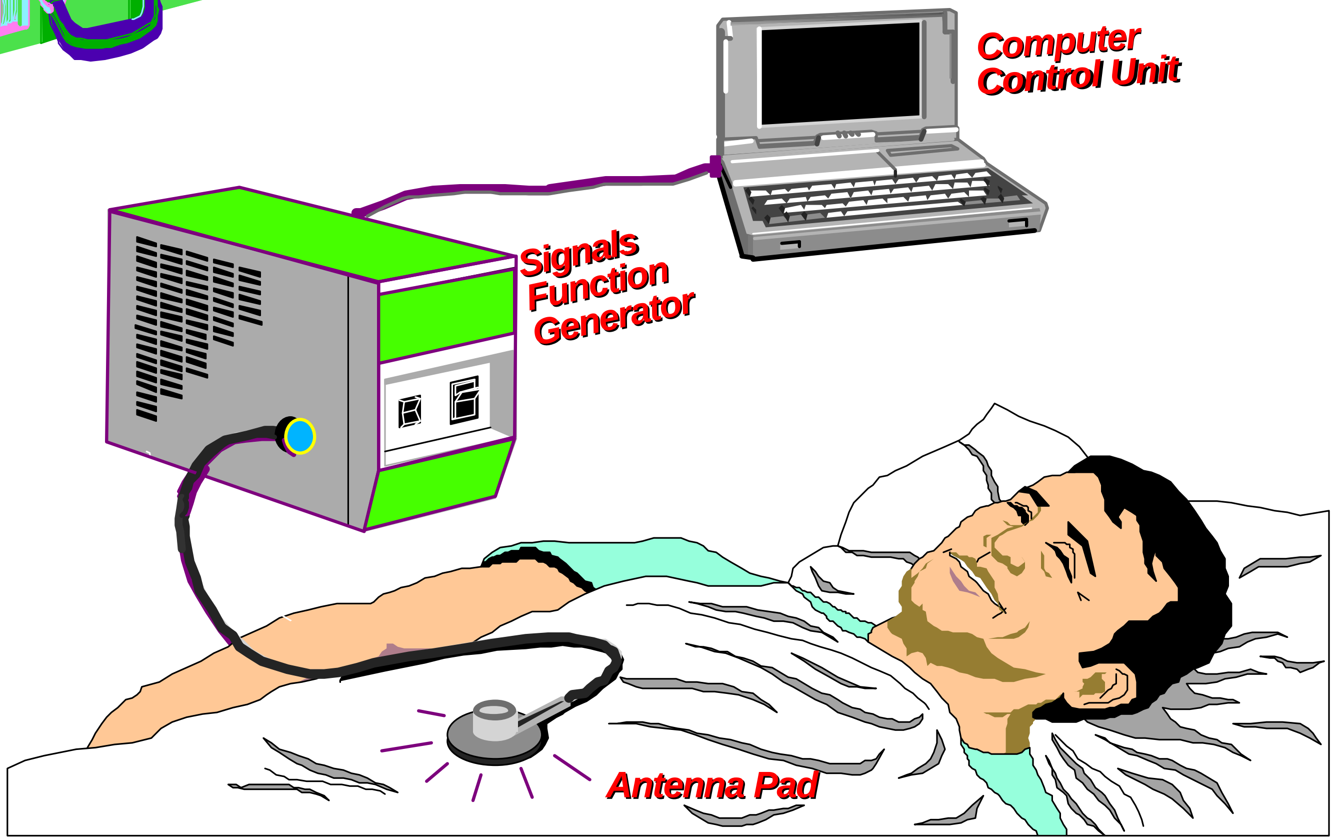
Computer
Control Unit
Signals
Function
Generator
Antenna Pad

The Porthole concept:

The two-minute treatment shortcut

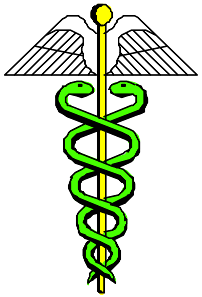

How the porthole concept works

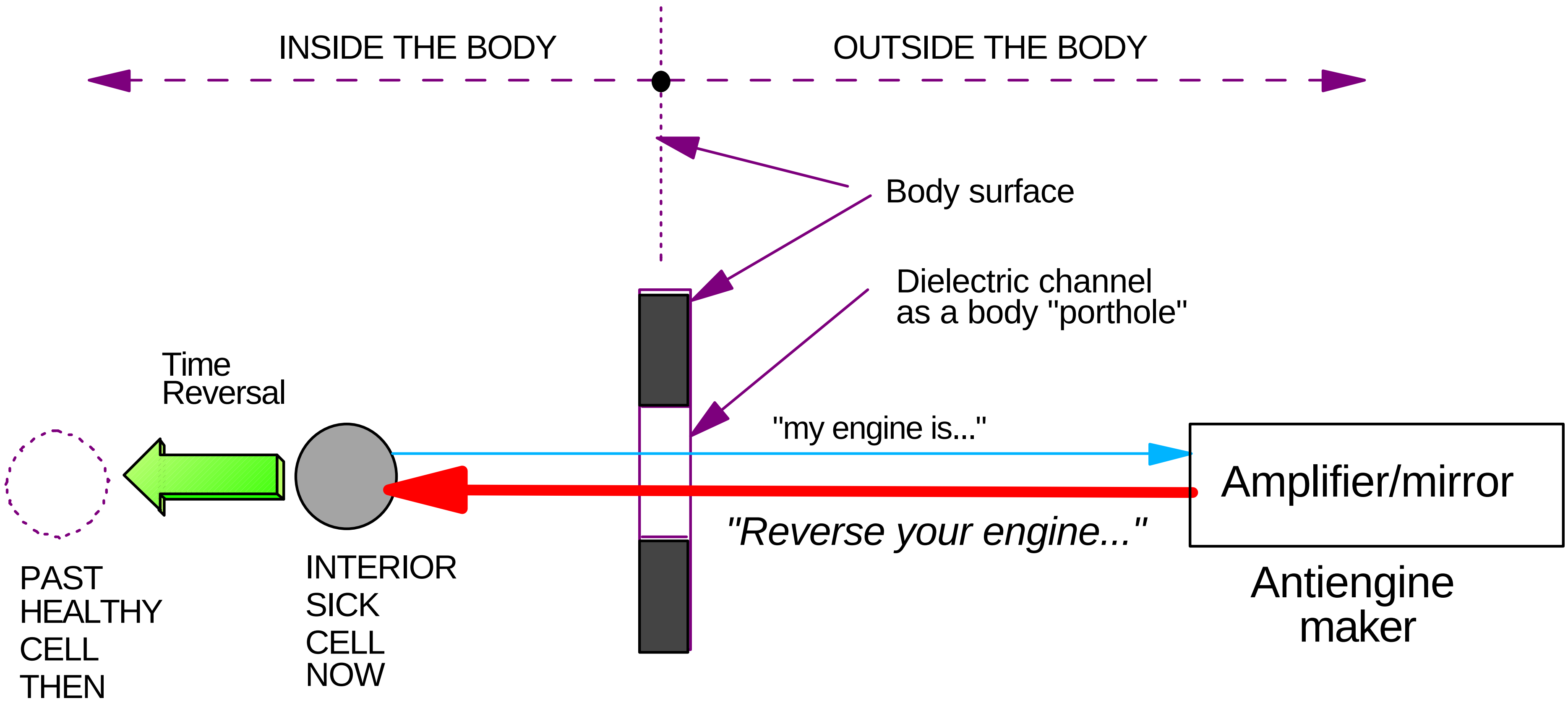
INSIDE THE BODY
OUTSIDE THE BODY
Body surface
Dielectric channel as a body "porthole"
Time Reversal
"my engine is..."
Amplifier/mirror
"Reverse your engine..."
PAST HEALTHY CELL THEN
INTERIOR SICK CELL NOW
Antiengine maker

Porthole concept: An analogy of the dielectric path and inner-outer functional connections

- Imagine perfectly reversible dynamic light processes inside a large "cruise ship", that are also perfect phase conjugate mirrors
- Each process receives, absorbs, and emits light continuously
- Light escaping through the portholes is a composite from all the dynamic processes ongoing inside
- If we record and amplify the escaping light, then force it back through the portholes, it reverses all the internal processes
- The processes revert back, slowly, to a previous ongoing process state
- We cease radiation, and the processes reverse and move forward again

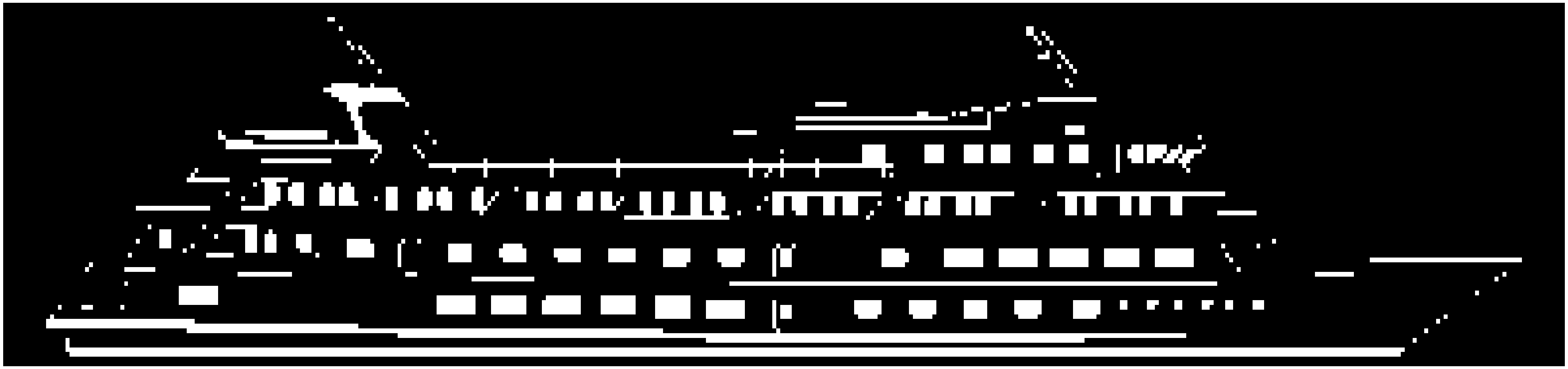

Porthole concept: The discovery

- The processes themselves, *in their own "frame"*, are never actually "reversed". Instead, the frame's motion through time is reversed.
  - Each to itself moves normally -- just receiving, absorbing, and radiating light, regardless of "whatever light" is interacting
  - This is due to a peculiarity of the photon and photon interaction
- Photon interaction creates observed physical reality, including the "frame" and the seeming "passage of a mass through time"
  - The photon is its own antiparticle, which is just a photon traveling in the opposite direction
  - So in this case we can use "light" to precisely time reverse (to the observer) the frame and the apparent operations (to the observer)
- When we cease irradiation, after a decay period to remove the reversed excitation, all functions are now seen to be moving forward normally in observer forward time, but from a previous physical state or condition

Using the porthole concept: Research

- The body's own emitted dense, broadband signals are recorded across a broad bandwidth
  - These signals are amplified, then rebroadcast precisely back into the body
  - This returns an amplified "reversed" signal set back through the dielectric path, invoking and applying the porthole concept in all processes
- The returning photons produce time-charging but in reversed frame direction
  - Tremendous potential energy is absorbed in highly compressed "time-energy" form in seconds
  - This "time-potential" slowly drains off as the body processes and cells reverse back to a previous state

Research using the porthole concept

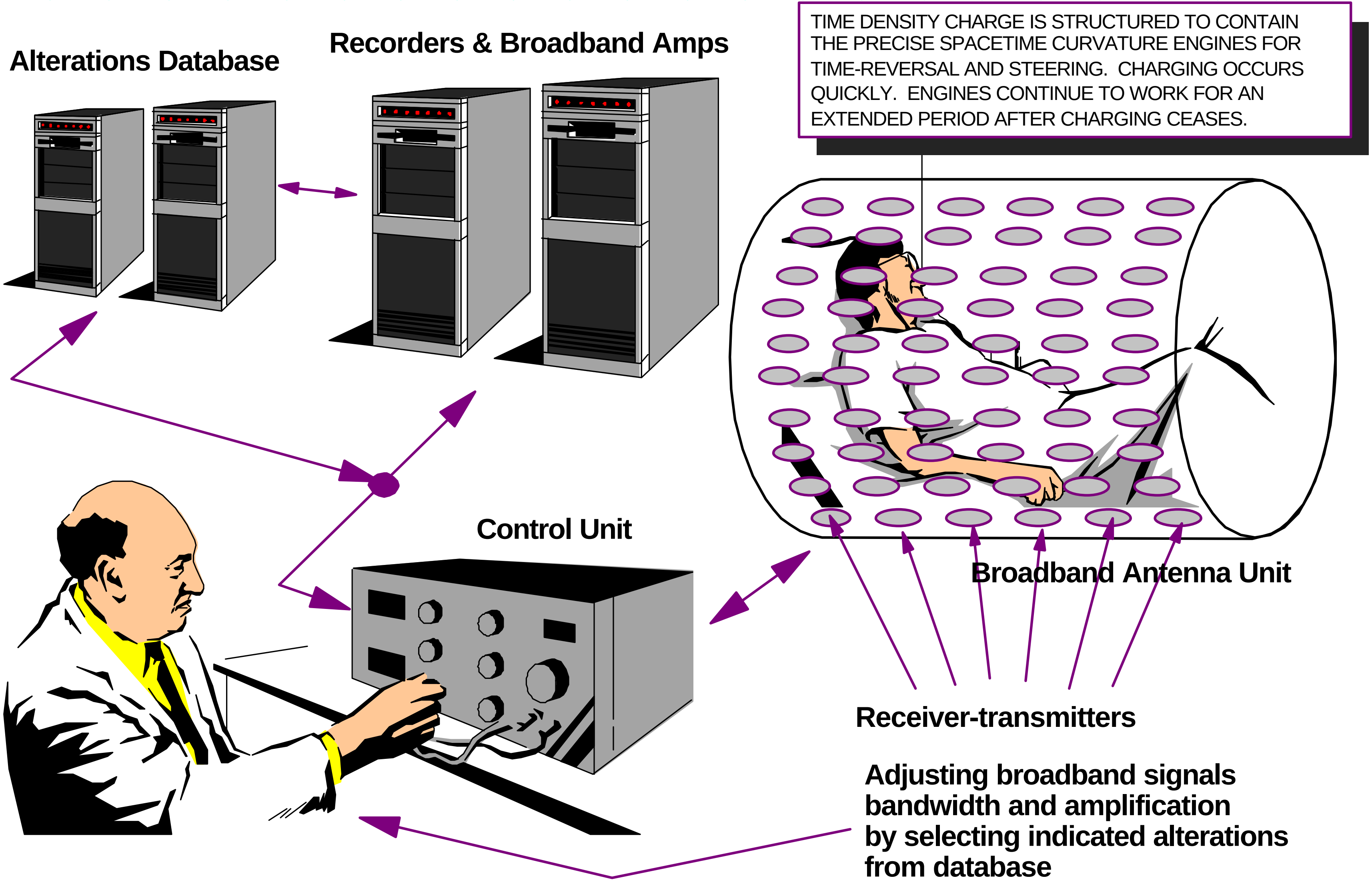

TIME DENSITY CHARGE IS STRUCTURED TO CONTAIN THE PRECISE SPACETIME CURVATURE ENGINES FOR TIME-REVERSAL AND STEERING. CHARGING OCCURS QUICKLY. ENGINES CONTINUE TO WORK FOR AN EXTENDED PERIOD AFTER CHARGING CEASES.

Alterations Database

Recorders & Broadband Amps

Control Unit

Broadband Antenna Unit

Receiver-transmitters

Adjusting broadband signals bandwidth and amplification by selecting indicated alterations from database

© T.E. BEARDEN 1998

Portable unit then developed for rapid treatment of mass BW casualties

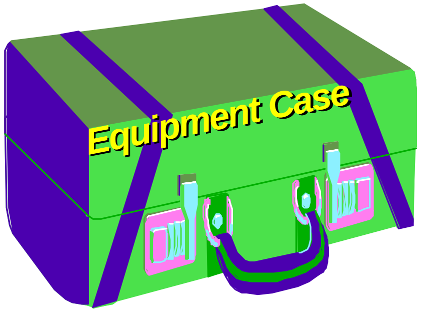
Equipment Case

- Two blankets have antennas in them.
- Uses the "porthole" concept (discussed shortly)
- Eliminates plasma tube mixing etc.
- Computer-controlled complex signal structure.
- Irradiation for 2 minutes.

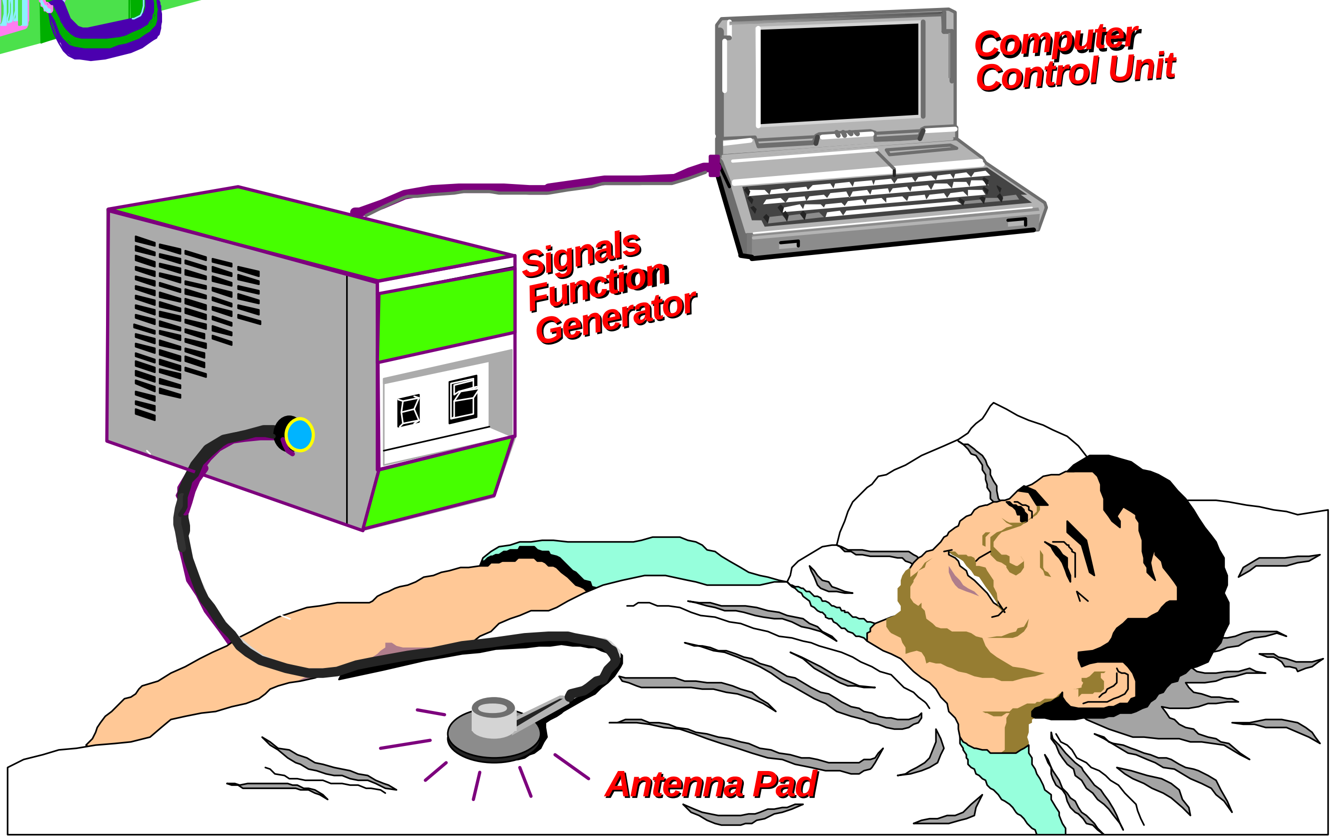
Computer
Control Unit
Signals
Function
Generator
Antenna Pad

Training to use the unit

- High school student level, 30 minutes.
- Patient lies on one blanket, other covers him.
- Set the prescribed standard signal sets from listing.
- Computer-controlled generator produces the signal.
- Irradiation for 2 minutes, uncover patient, bring in another.
- Unit can be self-powering from vacuum energy (MEG*).

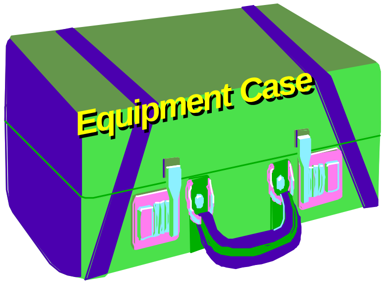
Equipment Case

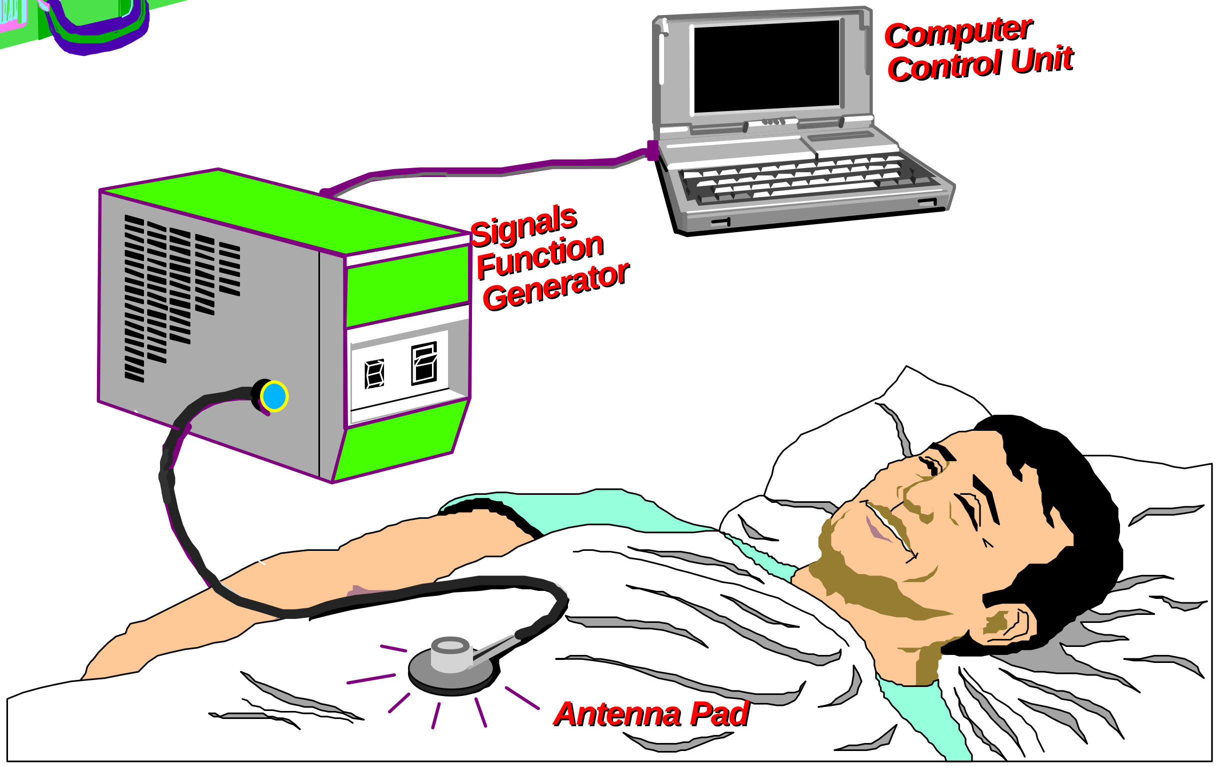
Computer
Control Unit
Signals
Function
Generator
Antenna Pad

Additional advantages

- Can be developed in a crash program in 2-3 years.
- After sunk costs, projected cost per unit in mass production is less than $100,000
- Can be produced by the thousands
- Then flooded down through
  - Fire stations and police stations
  - Red cross units, hospitals, clinics
  - Emergency action teams and armed forces
- Large numbers of operators easily produced
- Emergency training could be watching a 15 minute film, then performing one graded practice run
- First generation equipment can save some 70% of the coming mass casualties from anthrax etc.

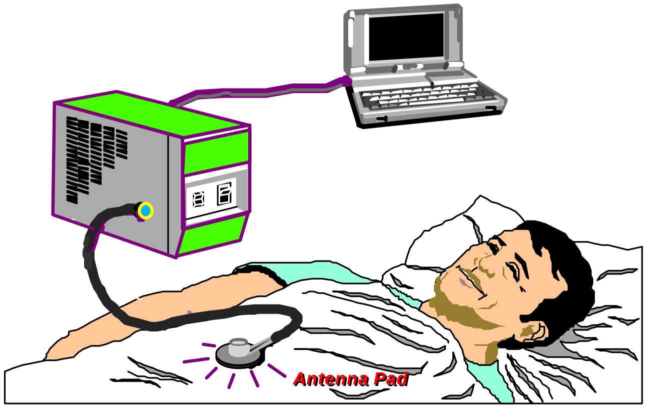

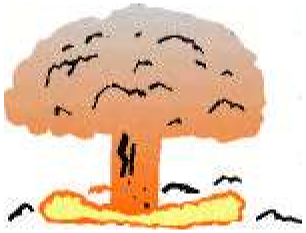

Terrorist BW warfare capabilities against the United States

The growing threat

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- 100 kilograms of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

***Per OTA Report to Congress, 1993**

For decades, infiltration of terrorist teams into the U.S.

- From several foreign nations, not just Bin Laden in Afghanistan
- Several thousand Cuban guerillas, from training camps in Southern Mexico

- Russian Spetznaz teams
- Nuclear warheads (up to 80 KT)
- Biological warfare agents
- Yakuza in oriental populations in our cities
  - Yakuza manufactures Russian energetics weapons in their own facilities in Japan
  - Possesses the new EM weapons referred to by Secretary of Defense Cohen in 1997
- We are now in a very long, hard war and will suffer millions of casualties in our cities. It may last for decades.

Defense Secretary Cohen*

(Involvement of secret superweapons)

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts."

* Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Some typical BW agents

- Smallpox, including different strain developed by Russia (some sold to terrorists)
- Anthrax
  - Over 400 known locations around the Earth have cultures
  - Difficult to decontaminate

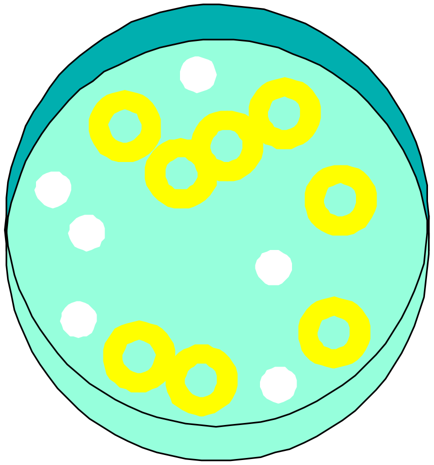

- Bubonic Plague
- Botulism toxin
- Tularemia (rabbit fever)
- Hemorrhagic fevers, such as Ebola

Anthrax

- OTA study: Some 200 pounds released in a spray, on calm night over Washington, D.C. could result in up to 3 million casualties
- Symptoms usually in 2-10 days
  - Sometimes as long as 7 weeks
  - Flu-like symptoms

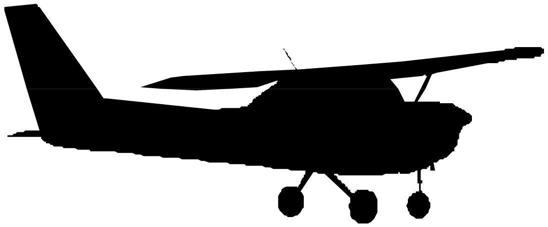

- Then 1-3 days later, shock and breathing problems cause death in up to 100%
- Antibiotics given early (before symptoms) can prevent infected persons from becoming ill
- Limited vaccine, reserved for military

Smallpox

- More than 500 million people killed in the 20th century before "eradicated" in 1977. Vaccination ceased in 1980.
- Russians kept it, developed new strains, sold to terrorists
- Symptoms in about 12 days: fever, headache, nausea. Rash similar to chickenpox then appears.
- Rash turns into hard blisters. *Highly contagious*, kills one-third of its victims.
- No cure. About 12 million doses of vaccine left.
- If smallpox is unleashed in any major city on earth, it will eventually kill 2 billion, or about 1/3 of the human species -- unless a radical new cure is found and utilized.

Bubonic Plague: Historically the most feared (Black Death)

- From 1980-1994, 18,739 cases in 20 countries.
- Symptoms occur in 1-6 days after inhaling the pneumonic form.

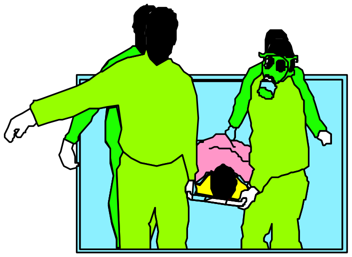

- High fever, cough, labored breathing lead to respiratory failure and death.
- It is contagious.
- Rapid use of antibiotics can be effective.
- Vaccine not currently produced in the U.S.

Botulism toxin (a nerve toxin)

- Single most poisonous BW substance known.
  - Typically borne in food, contacted by eating.
  - Could be developed as aerosol weapon.
- Symptoms in 24-36 hours; blurred vision and difficulty swallowing, speaking.
- Paralyzes muscles, leading to respiratory failure and death.
- Centers for Disease Control and Prevention maintain a botulism antitoxin supply.

Tularemia

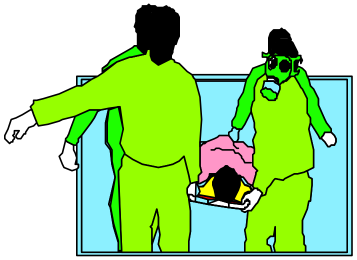

- U.S. studied its weapon potential in 1950s and 1960s.
- Symptoms appear in 3-5 days, including fever, chills, headache, and weakness
- Inflammation and hemorrhaging of the airways can lead to death.
- Without antibiotics, one-third die. Can be 90% if breathed in, in aerosol form.
- Vaccine developed and under review by FDA.

Hemorrhagic fevers, such as Ebola

- E.g., origin of Ebola is unknown. Probably transmitted to humans by animals, to begin a breakout.
- Symptoms in 3-5 days, include fever, aching muscles, diarrhea.
- Hemorrhaging of fluids out of tissues and orifices.
- From 30 to 90 percent die, depending on specific fever type.
- Some of these fevers respond to antiviral drugs.
- The drugs are in short supply.
- Presently *containment* is the primary goal.

A factor the U.S. does not know

(Secret involvement of clandestine superweapons)

- Three nations (not the U.S.) developed quantum potential weapons using Bohm's hidden variable theory of quantum mechanics.
- Can induce disease engines in matter or bodies at a distance.
- Have developed "spreading" of immune systems in a targeted populace (cleverly tested in Gulf War against U.S. troops).
  - Cocktail mix of disease engines for 2 dozen different diseases
  - Targeted area's human immune systems react to these engines when they are still in virtual state, just below quantum threshold.
  - Spread immune system's resources thus "thinned" across response to two dozen shadow diseases at once.
- Hit with a terrorist BW attack, the spread immune systems are overwhelmed much more easily and quickly.
  - Yield (casualties) of the attack is increased by factor of 3-5.
  - Spreading of U.S. immune systems has been accomplished and is still being maintained.

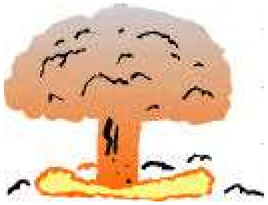

Summary and conclusion

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

The new war

A new Pearl Harbor: Strike on the World Trade Center in NYC

- The North Tower has already been struck and is burning.
- A second jetliner is about to strike the South Tower.

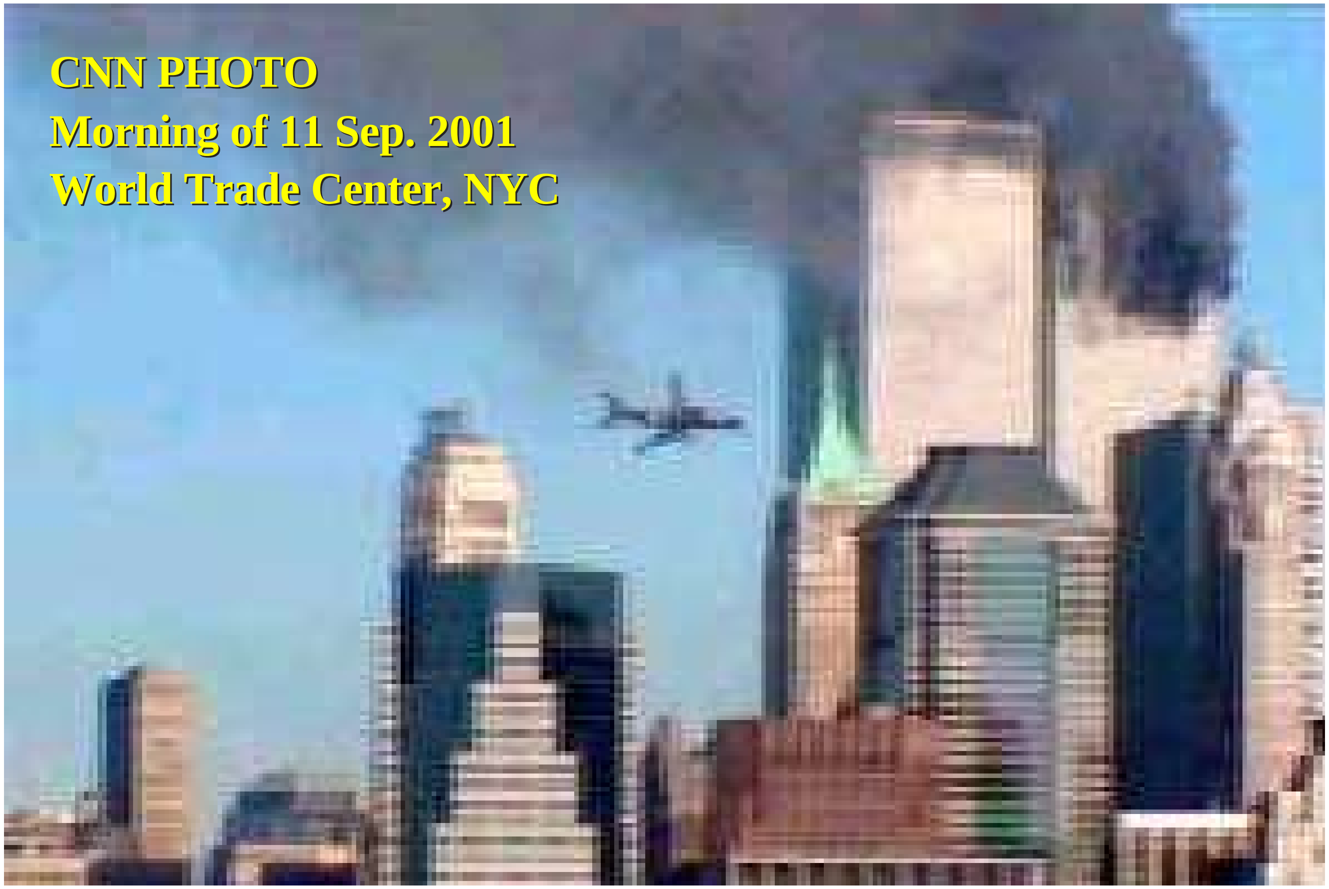

CNN PHOTO
Morning of 11 Sep. 2001
World Trade Center, NYC

Fireball From Impact on South Tower

- The two buildings held some 50,000 occupants
- Evacuation was underway after the first strike
- More than 6,000 persons were killed
- This is more than were killed in the Dec. 7, 1941 strike on Pearl Harbor

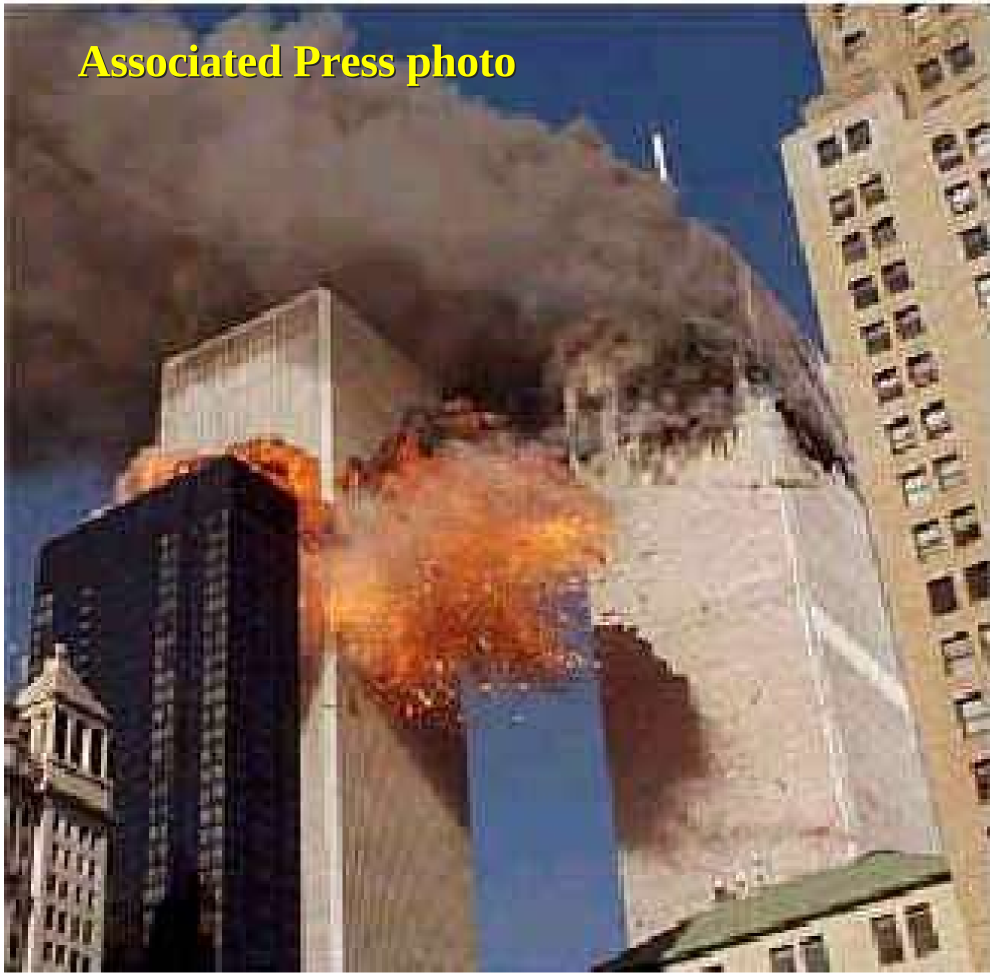

Associated Press photo

Collapse of the South Tower

- The first tower to collapse was the South Tower
- It is still uncertain just how many people were caught inside and killed
- This is a half million tons of debris, falling from the New York skyview.

Photo from CNN

Collapse of the North Tower begins

- The North Tower then collapsed
- This was another half-million tons of debris falling from the NYC skyview

Collapse of North Tower finishes

- Collapse of the North Tower killed many rescue workers, fire-men, and policemen
- At this writing, the total casualties are still unknown

Photos from CNN

Next began the recovery efforts

- Recovery operations will require months
- At this writing, the total casualties are still unknown
- The damage was done by two civilian jetliner "bombs"

The Pentagon was hit on the same day

- Hundreds were killed the same day when another jetliner struck the Pentagon.
- Yet another jetliner, possibly headed for the White House, crashed enroute when the passengers attacked the suicide terrorist hijackers.

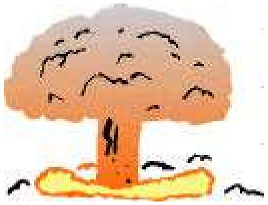

A sobering thought

- Nuclear weapons are already smuggled into the U.S., as are BW agents such as anthrax, smallpox, etc.
- Had a 40-80kT weapon been exploded in New York and one in Washington, millions would have perished and both cities would have been devastated
- There would have been no recovery of those cities for many decades

Present Status of the WMD Threat to the United States

- **Many foreign nations are hostile to the U.S. and sponsor terrorists**
- **Some 25 nations have WMD or are acquiring them**
  - **BW agents and weapons**
  - **Chemical agents and weapons**
  - **Nuclear materials and weapons**
- **Thousands of students and émigrés**
- **Infiltrated teams, BW, other WMD already**
- **Castro guerrillas infiltrated over the years**
- **Can do unacceptable damage to U.S. now**
- ***May reach knockout capability***

© 1998 T.E. BEARDEN

***Per OTA Report to Congress, 1993**

The U.S. is terribly unprepared

- Limited medical supplies, antibiotics, vaccines
- Insufficient facilities, emergency teams
- No civilian shelters (much less BW-filtered)
- Not enough doctors, nurses, masks, etc.
- Present treatment methods woefully inadequate
- Triage will apply
  - First treat those most likely to recover
  - Others (showing symptoms, or advanced) are just set aside to die
- Other attacks are a matter of *when*, not *if*.

Conclusion

- As several cities are struck, millions of U.S. civilian casualties will result. Devastation will be enormous.
- We shall eventually lose more American lives than in all the rest of our wars to date.
- Present medical therapeutic science is totally inadequate to do the job. It cannot save us.

- At some point, when we are sufficiently damaged, the Spetznaz or other terrorists may detonate nuclear weapons in our cities, with appalling carnage.
- A combined superweapon technology and terrorist groups will then destroy the remaining populace.
- On its present course, the inadequacy of Western medicine may doom us.

A difficult truth to consider

- In any strategic strike, the first phase is to get the strategic weapons (weapons of mass destruction) delivered to their distant targets.
- That has already been done.
- Eerily, the first phase of World War III has already been accomplished.
- It was already accomplished before the Sept. 11, 2001 strike on New York and Washington. That was the new "Pearl Harbor", and only the opening round. It already killed more Americans than died at Pearl Harbor on Dec. 7, 1941.
- Unless we improve our medical therapy quickly and dramatically, we will lose -- particularly when the hidden energetics weapons and outside terrorist sponsorship by hostile nations are factored in.

A personal perspective

- There now exists an overwhelming, immediate moral imperative to get this development program done at the utmost speed humanly possible.
- The sheer survival of the United States and Western civilization depends upon it.
- Officialdom is aware that the KGB/Russians have long planned to finish the U.S., after initial nuclear exchanges, by fierce and massive BW strikes.
- It appears that the plan has changed to let the terrorists do the main work, while augmenting U.S. casualty levels by energetics weapons.

China is also involved

- China has declared South China Sea the territorial waters of China
- China will eventually take Taiwan
  - Has massed missile artillery trained on Taiwan
  - Recently held invasion practice on an island near Taiwan
- Recently deployed fearsome EMP weapons utilizing negative energy rather than positive energy (Dirac sea theory)
  - Extinguishes all electron currents
  - Duds electrical and electron systems instantly
  - Instant death to struck bodies
  - Instant kill of missiles, aircraft, electronics in nuclear warheads
  - Ultimate "death ray"; can be made small or large
- Recent 2-carrier U.S. task force maneuvered in South China Sea to caution China
  - The new Chinese EMP weapon was trained on that task force
  - Could have destroyed it within minutes

Final thoughts

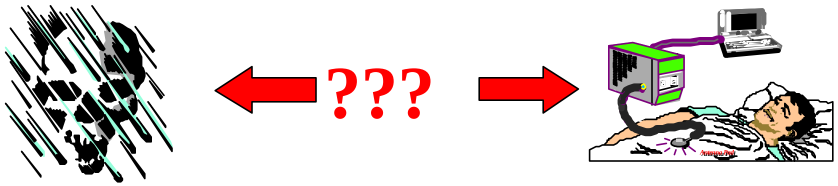

- This is one of the great turning points in history.
- The development program is eminently doable. That it works has been experimentally demonstrated. The long-vexing puzzle of the active mechanism has been solved.
- Our scientific community bitterly opposes any use of electromagnetics -- even an advanced, higher symmetry electrodynamics -- for medical treatment.
- A consortium of the scientific community, the large pharmaceuticals, FDA, and the AMA will oppose this project even while the U.S. is being destroyed wholesale.
- It can only be done in the private sector, and there will likely be no funding unless by philanthropists.

Continuation

Business plan, development program, budget breakdown, proposed schedule etc. go here.

Weapons Slides

The Tom Bearden Website

Help support the research

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts." *

Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Slides
(covers period up to 1999)

- The Coming Strategic Attack on the United States
- Outline
- Loss of two separate A-10 Warthogs in 1997
- Typical ABM Defenses (Widely Deployed by mid-70s)

- Launch Phase Anti-Missile System
- Launch Phase Anti-Bomber System
- HC-130s Engines Quit, Aircraft Crashes
  Nov. 22, 1996
- 1996 Actions by this analyst
- Additional Information Exists: Not Analyzed, Not Integrated in Briefing
- Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985
- Satellite Photos of Plumes in Soviet Arctic Since 1974
  See "Fer de Lance" for additional information
- Aum Shinrikyo ("Supreme Truth" Cult), Shoko Asahara, Leader

  Aum Shinrikyo - Anatomy of a Dangerous Cult
- Ashkhabad Problem and Its Solution
- Strange Characteristics of Iran Quake, Sep. 1978
- Iran Earthquake, September 1978
- Situation Now (Assessment)
- Situation Now (Assessment 2)
- Aspects of Strong Local Asymmetry
- Mass Casualty Problem: Aerial Anthrax

Spray on Washington, D.C.

- Present Status of the WMD Threat to the United States
- Some Present Problems in the U.S. WMD Defense Capabilities
- Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future
- Attempt to Ban Energetics Weapons Worldwide
- Brezhnev's Schedule
- Large glowing Tesla globe witnessed in Red China by hundreds
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  Sequence of Events
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  What They Were
- Cold Explosion -- 9 April 1984 Near Kurils (Plus additional weapons effects)
- Countering Insane MAD Systems:
  Find Loophole, Exploit It
- Next Generation Counter to Both Sane and Insane MAD systems
- Simultaneous Dudding of Nuclear Weapons Worldwide

- Simultaneous Dudding of Nuclear Weapons Worldwide
- Operation Divine Wind II
- Advanced EMBW Applications
- EM Biological Warfare (EMBW)
- Microwave Radiation of U.S. Embassy in Moscow
- Pentagon EMI Studies, 1989
- EM Missile Practice (6 slides)
- Subdivisions of Soviet Energetics Program
- Communism's Goal
- Manuilsky's Goal vs. Status Today (Dec. 1996)
- Professor Hellman, German Scientist Released from Work in Soviet Union
- Hemispheres and Globes of Light (4 slides)
- How Could Such Superweapons Ever Be Concealed?
- Soviets Easily Concealed These Superweapons for Four Decades
- Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes
- Longitudinal Wave Interferometry: Endothermic

- Longitudinal Wave Interferometry: Exothermic
- U.S. Air Attack on Libya, April 1986
- Lightning Wall
  circa 1959-1960
- Lisitsyn's Report: Brain Code Broken
- Large Longitudinal Wave Interferometer (LWI) Accident
- Large Longitudinal Wave Interferometer (LWI) Accident
- Use of Hidden Information Content of the Field Can Provide Action-At-A-Distance
- Two Types of Mutual Assured Destruction (MAD) System
- Psychoenergetics Weapons Teams Mentally Disable All Personnel
- Strikes Against Special Ship and Energetics Weapons Sites
- "Mindsnapper" Attack on Special Ships and Energetics Weapons Sites
- How Much Off Guard Are We?
  (7 Slides)
- Possible Targets: 1995-9 Escalation Phase
- Propagation of Weapon Effects
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15th, 1979

- Countering an Insane QP Weapon
- Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases
- Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
- Mechanism for Producing a Quantum Potential
- Major Principles for Use of Quantum Potential Weapons
- Major Principles for Use of Quantum Potential Weapons (2)
- Instantaneous Communication by a Quantum Potential
- Participants in a quantum potential share a common multiply-connected spacetime (MCST)
- Quake in Tangshan, China
  28 July, 1976
- Exchanging Earthquakes
- Instant Communication by a Quantum Potential
- Westerners Do Not React to Slow Threats
- Characterizing a Given Risk or Risk Factor
- Technical Risk Management
- Aum Shinrikyo's Sarin Facility

- In Scalar Beam Interference Zone
- Physicist Schappeller Auralaz Munster, Austria
- Initial Soviet Plan for Cold War, After WWII
- Josef Stalin: Absolute Dictator
- Exploding any Resistance by Operational Scientists
- Significant Soviet Statements
- Captain Svoboda Dives Headlong to Her Death
- T-polarized Beam Projector
- Possible Tests (1)
- Possible Tests (2)
- Possible Tests (3)
- Possible Tests (4)
- Mushroom Cloud Rising From Sea
- Launches from Cape Canaveral and Vandenberg AFB Provided Practice Targets of Opportunity
- The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth
- Embodiment of an instrument array sufficient to discriminate when time-density charging has occurred, and that transduction

of time-charge decay has occurred.

- Time-Polarized EM Wave Interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone
- Transmutation of Radioactive Nuclei at a Distance
- Transmutation of Radioactive Nuclei
- Typical Tesla Shield, Northern Route of Japan
- A Curious Tide in Human Affairs
- Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958
- The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
- The KGB/Yakuza-Aum War Plan
- Why Western Science is Ignorant of KGB Energetics
- Transmutation of Radioactive Nuclei at a Distance
- TheYakuza
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15, 1979

---

* Note that the Secretary did not speak the words in square blocks; these were added by the writers later,

probably to "soften the tone" of the Secretary's information release. So the Secretary confirmed that some nations of the world -- not just some ragged terrorists, as the writers tried to imply for spin control -- do possess novel electromagnetic weapons and are using them to induce earthquakes, engineer the weather and climate, and trigger volcanic eruptions.

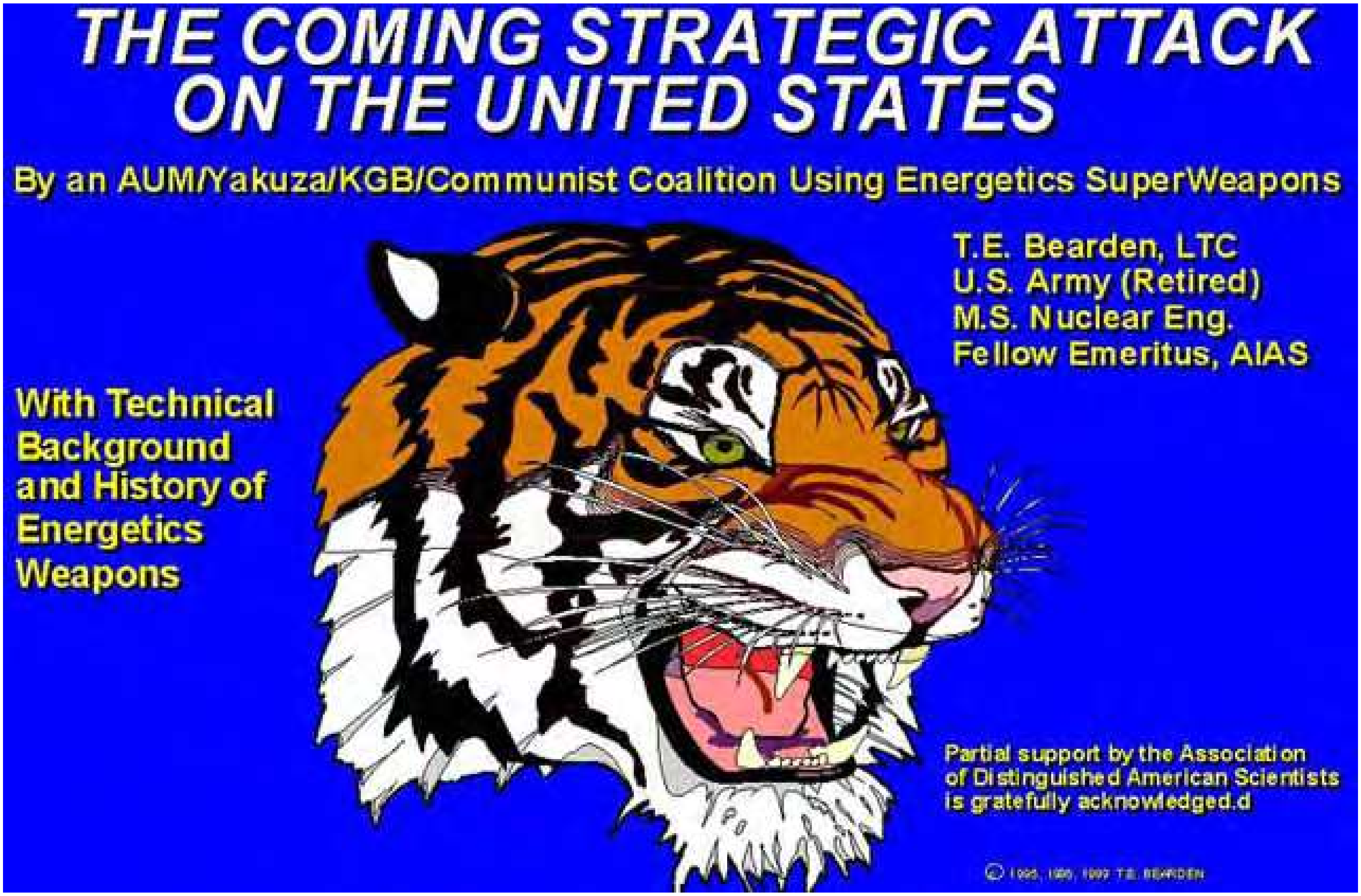

THE COMING STRATEGIC ATTACK ON THE UNITED STATES

By an AUM/Yakuza/KGB/Communist Coalition Using Energetics SuperWeapons

T.E. Bearden, LTC
U.S. Army (Retired)
M.S. Nuclear Eng.
Fellow Emeritus, AIAS

With Technical Background and History of Energetics Weapons

Partial support by the Association of Distinguished American Scientists is gratefully acknowledged.d

© 1995, 1996, 1997 T.E. BEARDEN

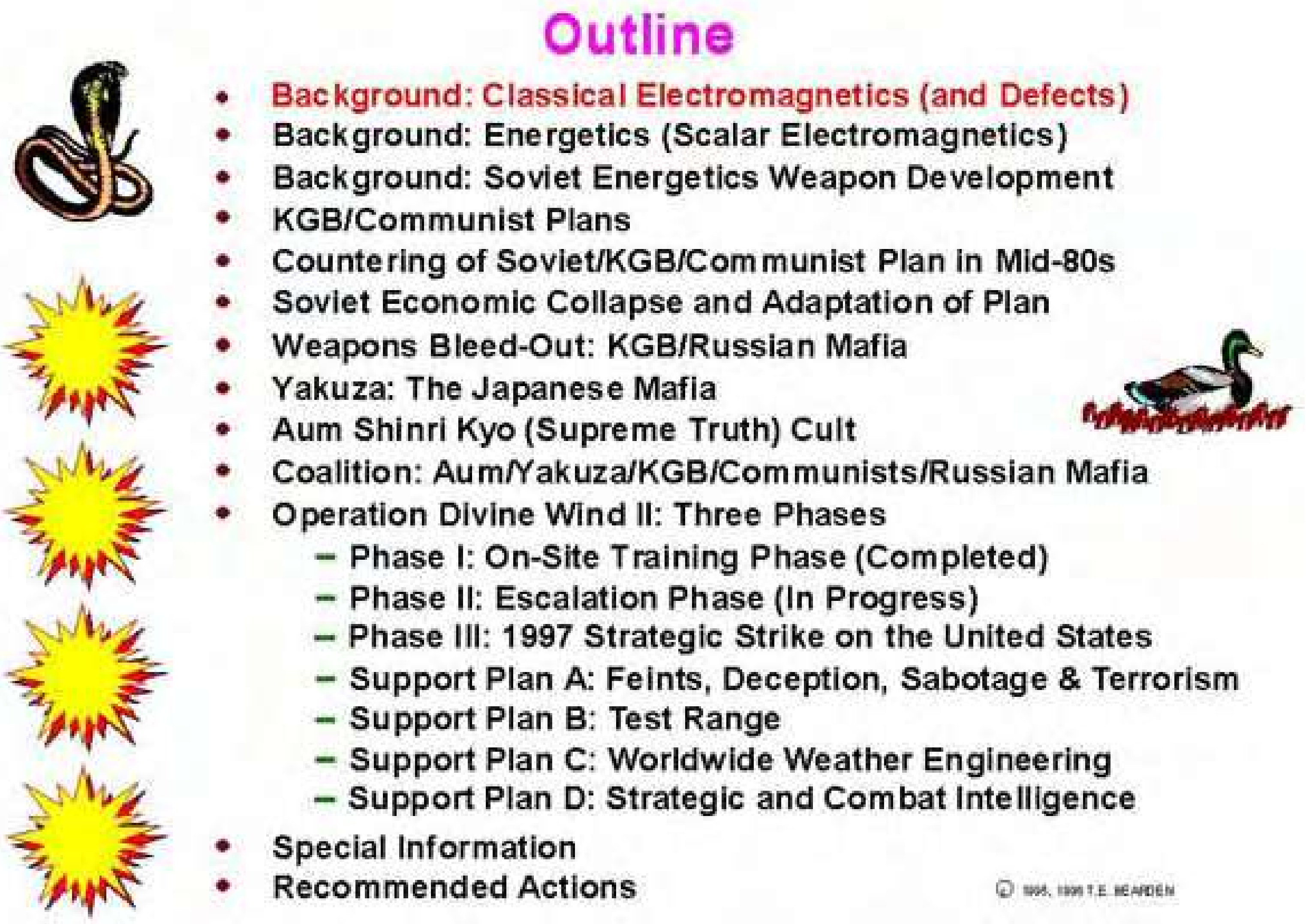

Outline

- Background: Classical Electromagnetics (and Defects)
- Background: Energetics (Scalar Electromagnetics)
- Background: Soviet Energetics Weapon Development
- KGB/Communist Plans
- Countering of Soviet/KGB/Communist Plan in Mid-80s
- Soviet Economic Collapse and Adaptation of Plan
- Weapons Bleed-Out: KGB/Russian Mafia
- Yakuza: The Japanese Mafia
- Aum Shinri Kyo (Supreme Truth) Cult
- Coalition: Aum/Yakuza/KGB/Communists/Russian Mafia
- Operation Divine Wind II: Three Phases
  - Phase I: On-Site Training Phase (Completed)
  - Phase II: Escalation Phase (In Progress)
  - Phase III: 1997 Strategic Strike on the United States
  - Support Plan A: Feints, Deception, Sabotage & Terrorism
  - Support Plan B: Test Range
  - Support Plan C: Worldwide Weather Engineering
  - Support Plan D: Strategic and Combat Intelligence
- Special Information
- Recommended Actions

© [illegible], [illegible] T.E. BEARDEN

The Tom Bearden Website

Help support the research

Loss of two separate A-10 Warthogs in 1997

- A-10 Mysterious Flyaway and Crash - April 2, 1997
- Death of Second A-10 Pilot - May 27, 1997
- Captain Button's Mysterious Flyaway and Eventual Crash
- Death of Second A-10 Pilot - May 27, 1997

A-10 Mysterious Flyaway and Crash

April 2, 1997

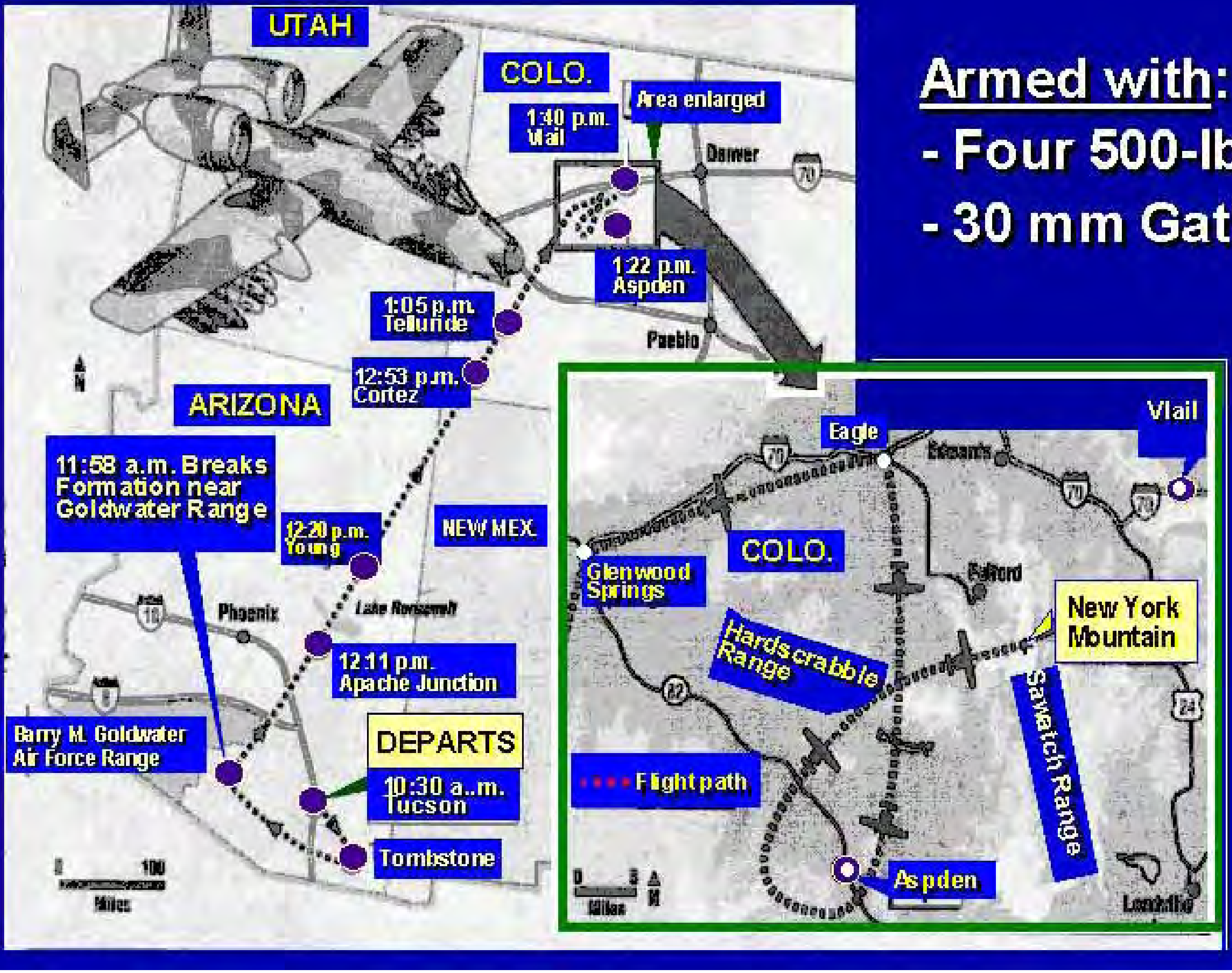
UTAH
COLO.
Area enlarged
1:40 p.m. Vail
Denver
70
1:22 p.m. Aspden
1:05 p.m. Telluride
Pueblo
12:53 p.m. Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
NEW MEX.
Phoenix
Lake Roosevelt
10
12:11 p.m. Apache Junction
8
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a.m. Tucson
Tombstone
0 100
Miles
N
Eagle
Edwards
Vail
70
COLO.
Glenwood Springs
Gilford
New York Mountain
Hardscrabble Range
82
Sawatch Range
24
Flight path
Aspden
0 5
Miles
N
Leadville

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

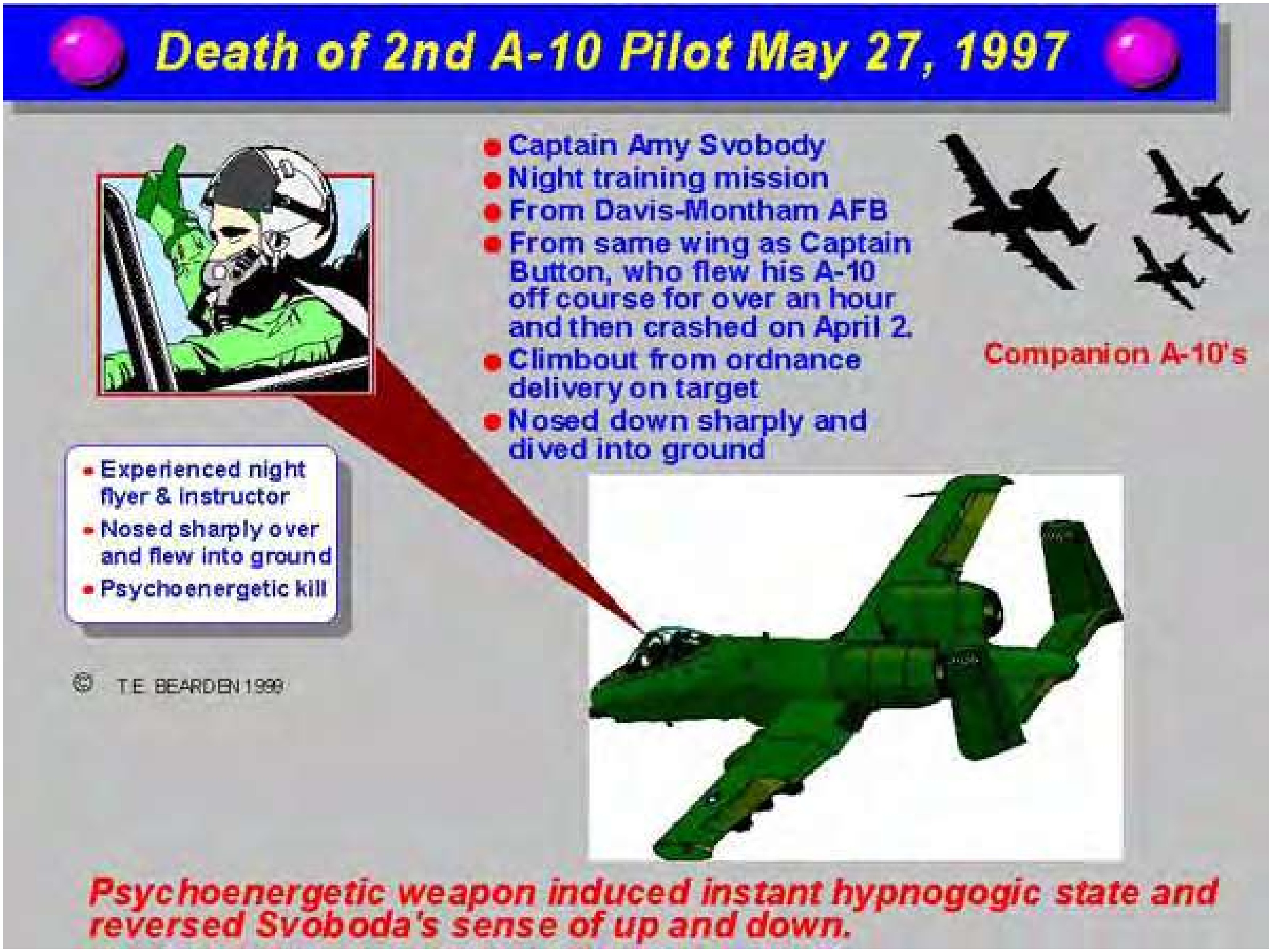

Death of 2nd A-10 Pilot May 27, 1997

- Captain Amy Svobody
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Climbout from ordnance delivery on target
- Nosed down sharply and dived into ground

Companion A-10's

- Experienced night flyer & instructor
- Nosed sharply over and flew into ground
- Psychoenergetic kill

© T.E. BEARDEN 1999

Psychoenergetic weapon induced instant hypnogogic state and reversed Svoboda's sense of up and down.

A-10 Mysterious Flyaway and Crash

April 2, 1997

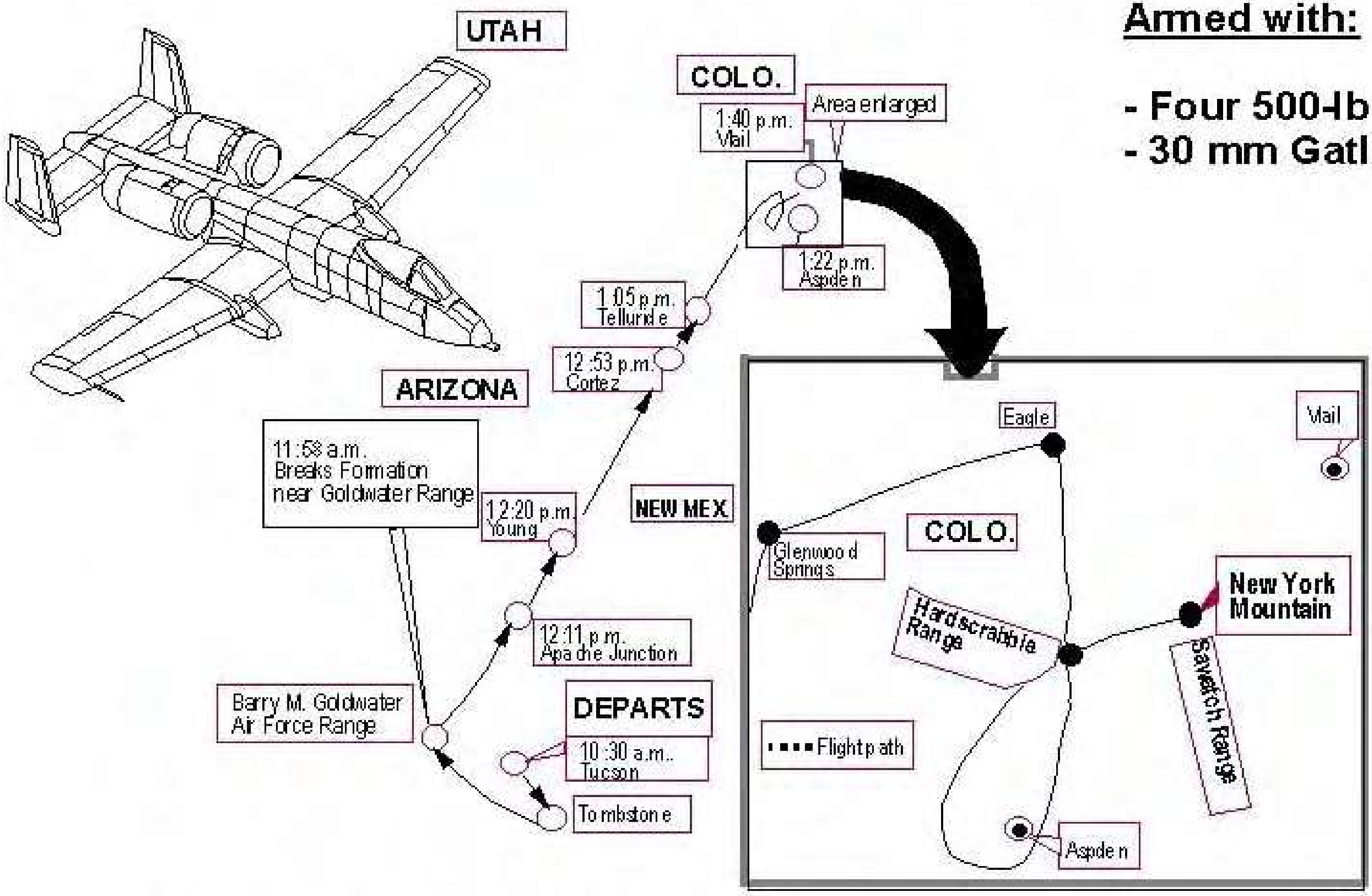
UTAH
COLO.
1:40 p.m.
Vail
Area enlarged
1:22 p.m.
Aspden
1:05 p.m.
Telluride
12:53 p.m.
Cortez
ARIZONA
11:58 a.m.
Breaks Formation
near Goldwater Range
12:20 p.m.
Young
NEW MEX
12:11 p.m.
Apache Junction
Barry M. Goldwater
Air Force Range
DEPARTS
10:30 a.m.
Tucson
Tombstone
Eagle
Vail
Glenwood
Springs
COLO.
Hardscrabble
Range
New York
Mountain
Savatch Range
Flight path
Aspden

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

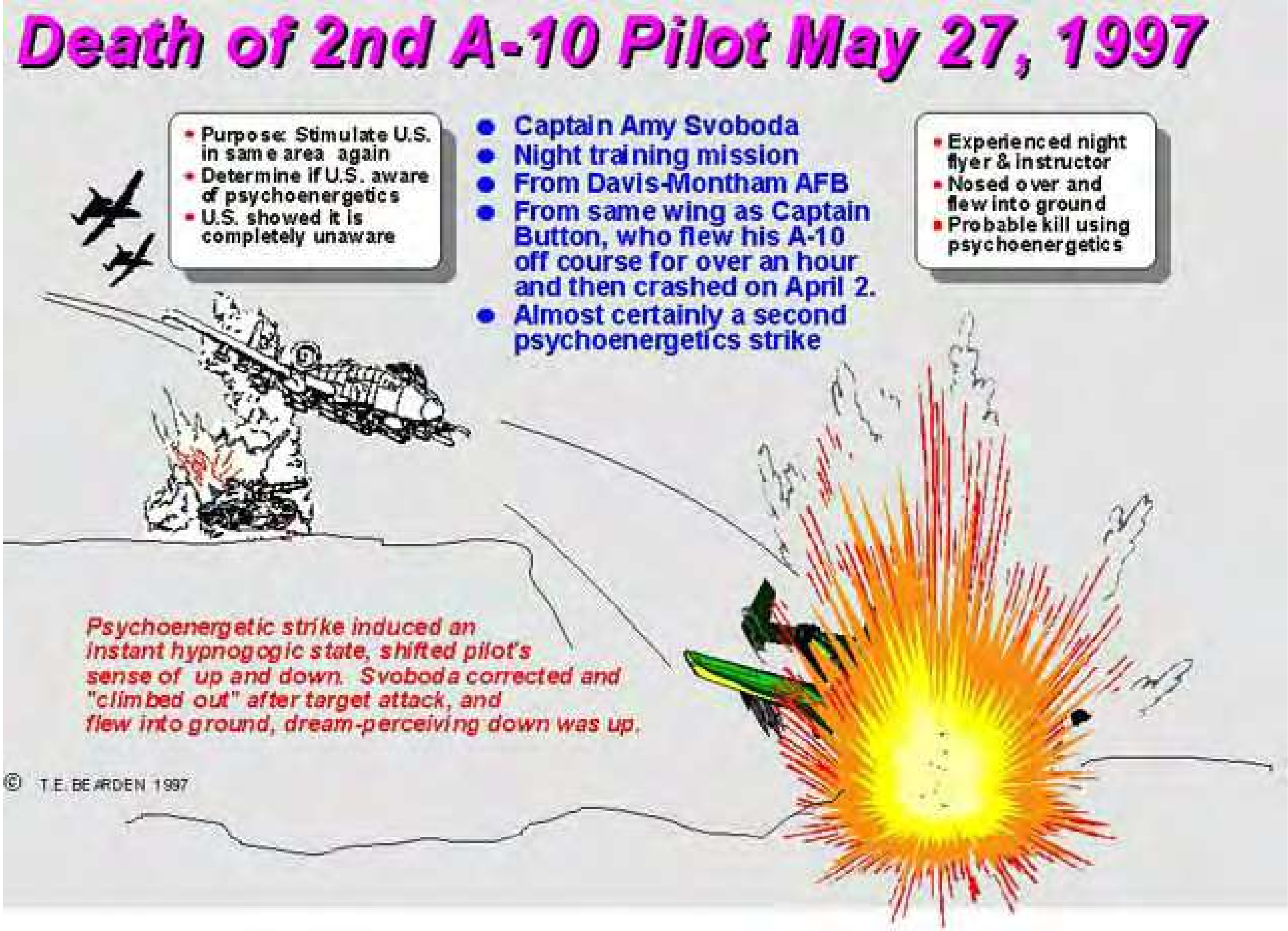

Death of 2nd A-10 Pilot May 27, 1997

- Purpose: Stimulate U.S. in same area again
- Determine if U.S. aware of psychoenergetics
- U.S. showed it is completely unaware

- Captain Amy Svoboda
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Almost certainly a second psychoenergetics strike

- Experienced night flyer & instructor
- Nosed over and flew into ground
- Probable kill using psychoenergetics

Psychoenergetic strike induced an instant hypnogogic state, shifted pilot's sense of up and down. Svoboda corrected and "climbed out" after target attack, and flew into ground, dream-perceiving down was up.

© T.E. BEARDEN 1997

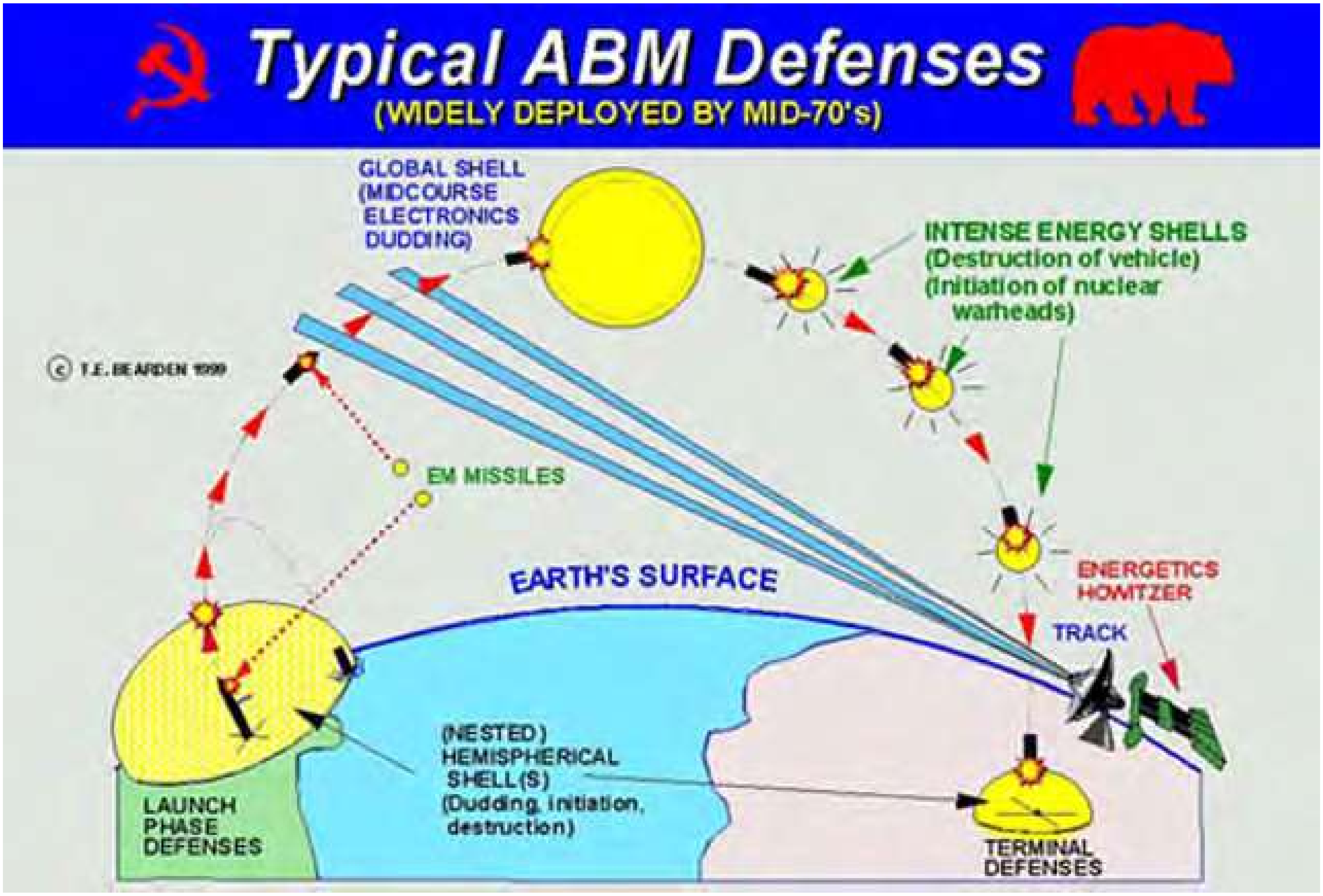
Typical ABM Defenses
(WIDELY DEPLOYED BY MID-70's)
GLOBAL SHELL (MIDCOURSE ELECTRONICS DUDDING)
INTENSE ENERGY SHELLS (Destruction of vehicle) (Initiation of nuclear warheads)
© T.E. BEARDEN 1999
EM MISSILES
EARTH'S SURFACE
ENERGETICS HOWITZER
TRACK
(NESTED) HEMISPHERICAL SHELL(S) (Dudding, initiation, destruction)
LAUNCH PHASE DEFENSES
TERMINAL DEFENSES

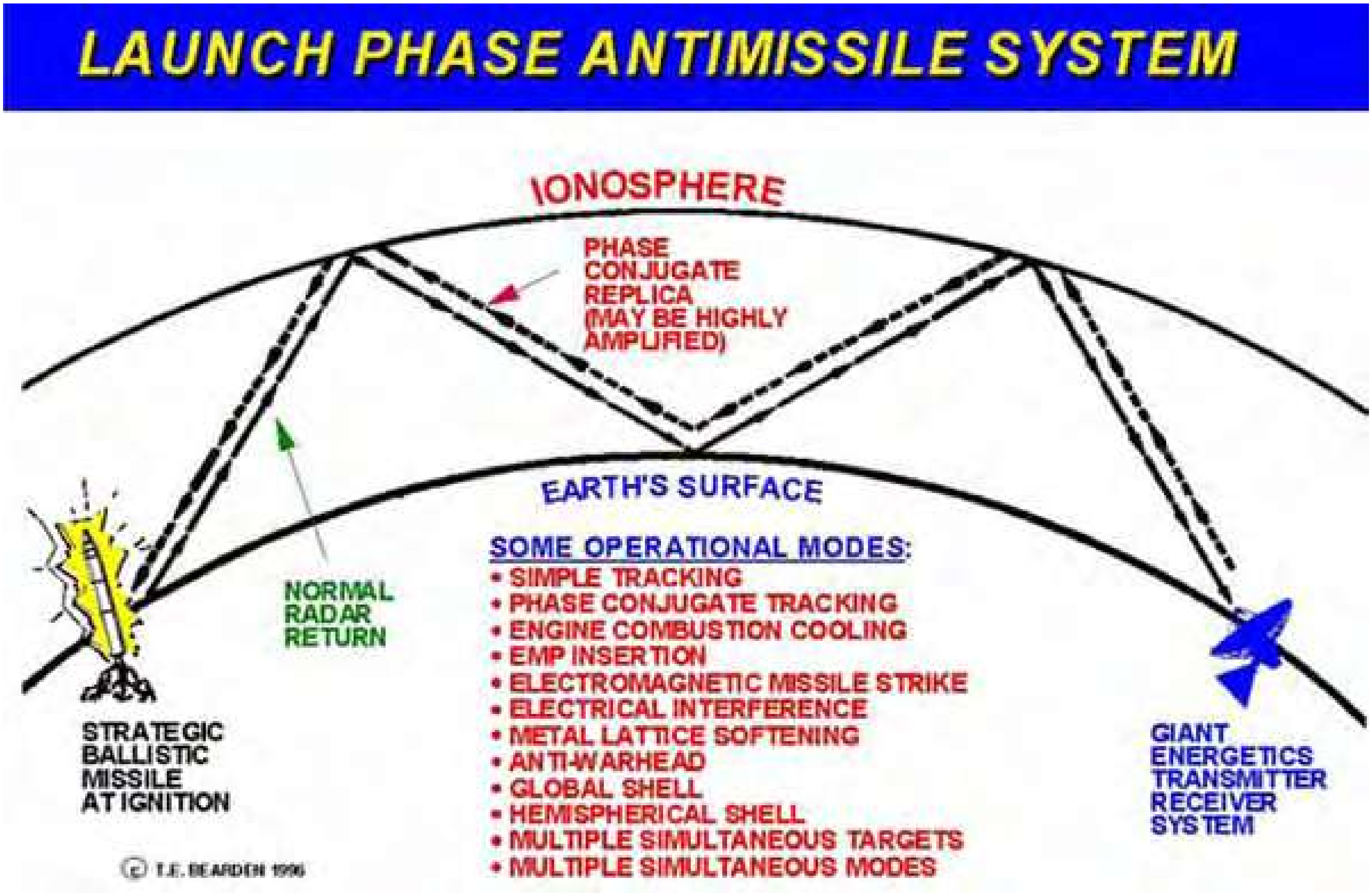

LAUNCH PHASE ANTIMISSILE SYSTEM

IONOSPHERE

PHASE
CONJUGATE
REPLICA
(MAY BE HIGHLY
AMPLIFIED)

EARTH'S SURFACE

NORMAL
RADAR
RETURN

STRATEGIC
BALLISTIC
MISSILE
AT IGNITION

GIANT
ENERGETICS
TRANSMITTER
RECEIVER
SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-WARHEAD
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

© T.E. BEARDEN 1996

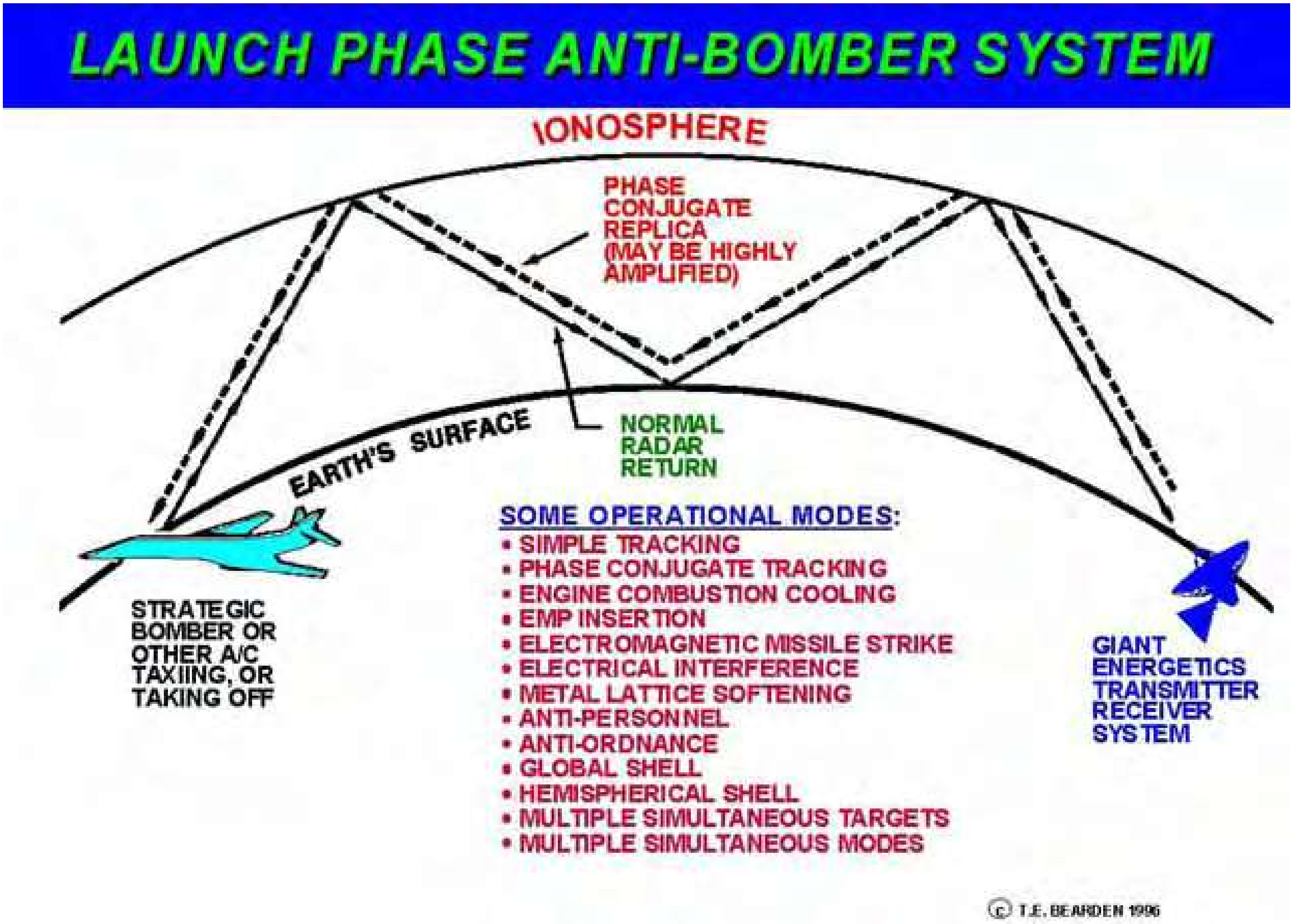

LAUNCH PHASE ANTI-BOMBER SYSTEM

IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC BOMBER OR OTHER A/C TAXIING, OR TAKING OFF
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-PERSONNEL
- ANTI-ORDNANCE
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

© T.E. BEARDEN 1996

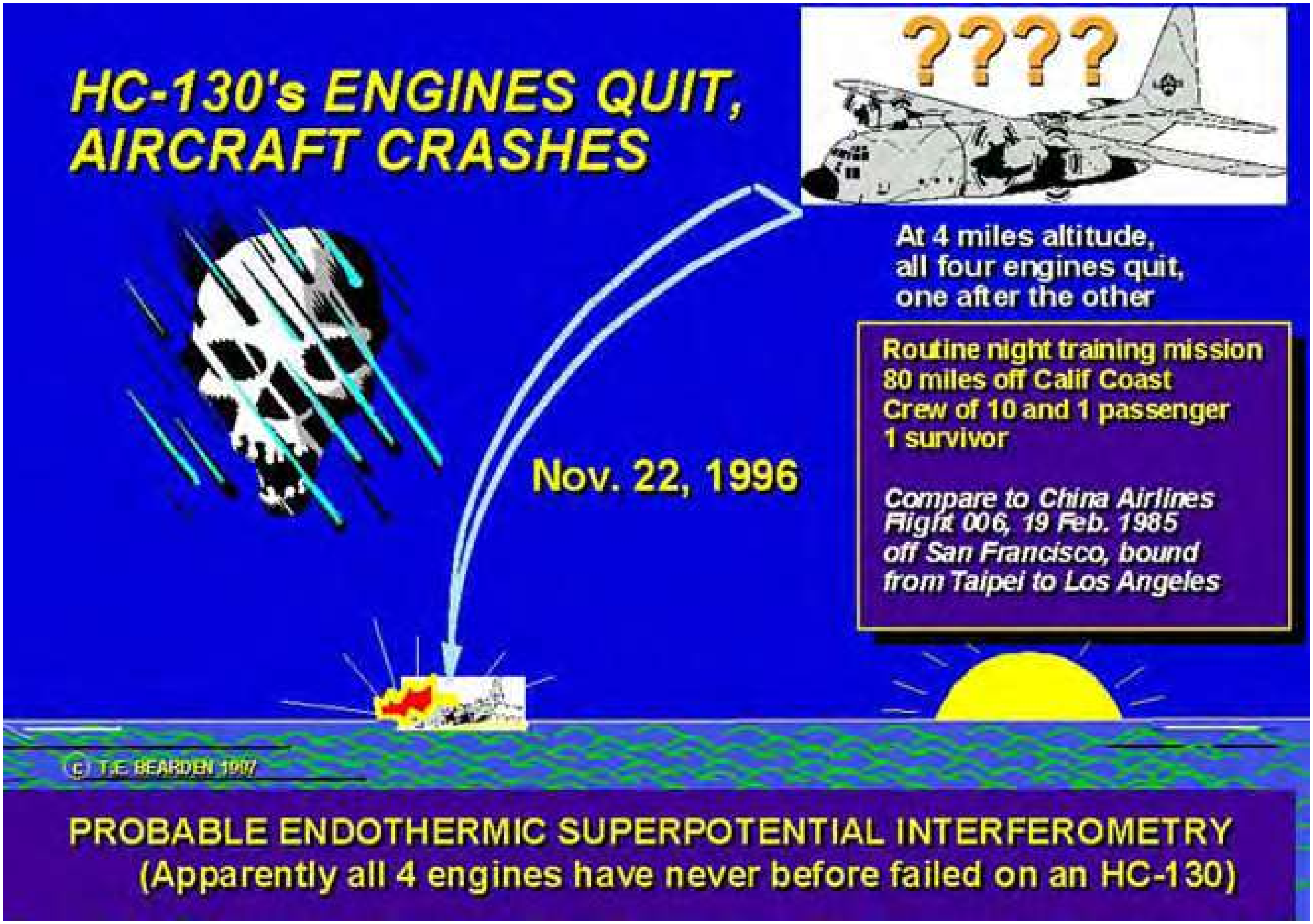

HC-130's ENGINES QUIT, AIRCRAFT CRASHES

At 4 miles altitude,
all four engines quit,
one after the other

Routine night training mission
80 miles off Calif Coast
Crew of 10 and 1 passenger
1 survivor

Compare to China Airlines Flight 006, 19 Feb. 1985 off San Francisco, bound from Taipei to Los Angeles

Nov. 22, 1996

© T.E. BEARDEN 1997

PROBABLE ENDOTHERMIC SUPERPOTENTIAL INTERFEROMETRY
(Apparently all 4 engines have never before failed on an HC-130)

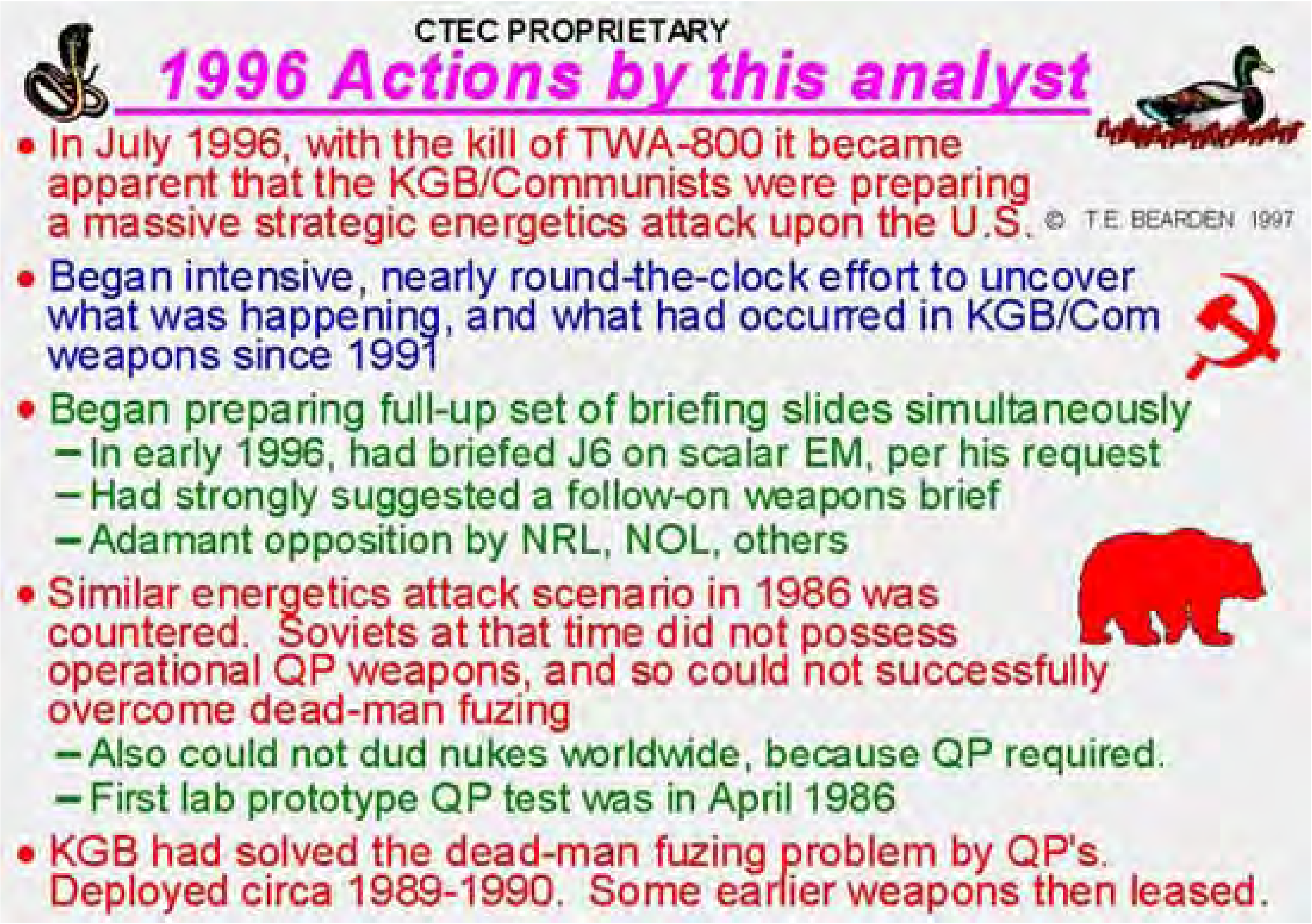

CTEC PROPRIETARY

1996 Actions by this analyst

- In July 1996, with the kill of TWA-800 it became apparent that the KGB/Communists were preparing a massive strategic energetics attack upon the U.S.

© T.E. BEARDEN 1997

- Began intensive, nearly round-the-clock effort to uncover what was happening, and what had occurred in KGB/Com weapons since 1991
- Began preparing full-up set of briefing slides simultaneously
  - In early 1996, had briefed J6 on scalar EM, per his request
  - Had strongly suggested a follow-on weapons brief
  - Adamant opposition by NRL, NOL, others
- Similar energetics attack scenario in 1986 was countered. Soviets at that time did not possess operational QP weapons, and so could not successfully overcome dead-man fuzing
  - Also could not dud nukes worldwide, because QP required.
  - First lab prototype QP test was in April 1986
- KGB had solved the dead-man fuzing problem by QP's. Deployed circa 1989-1990. Some earlier weapons then leased.

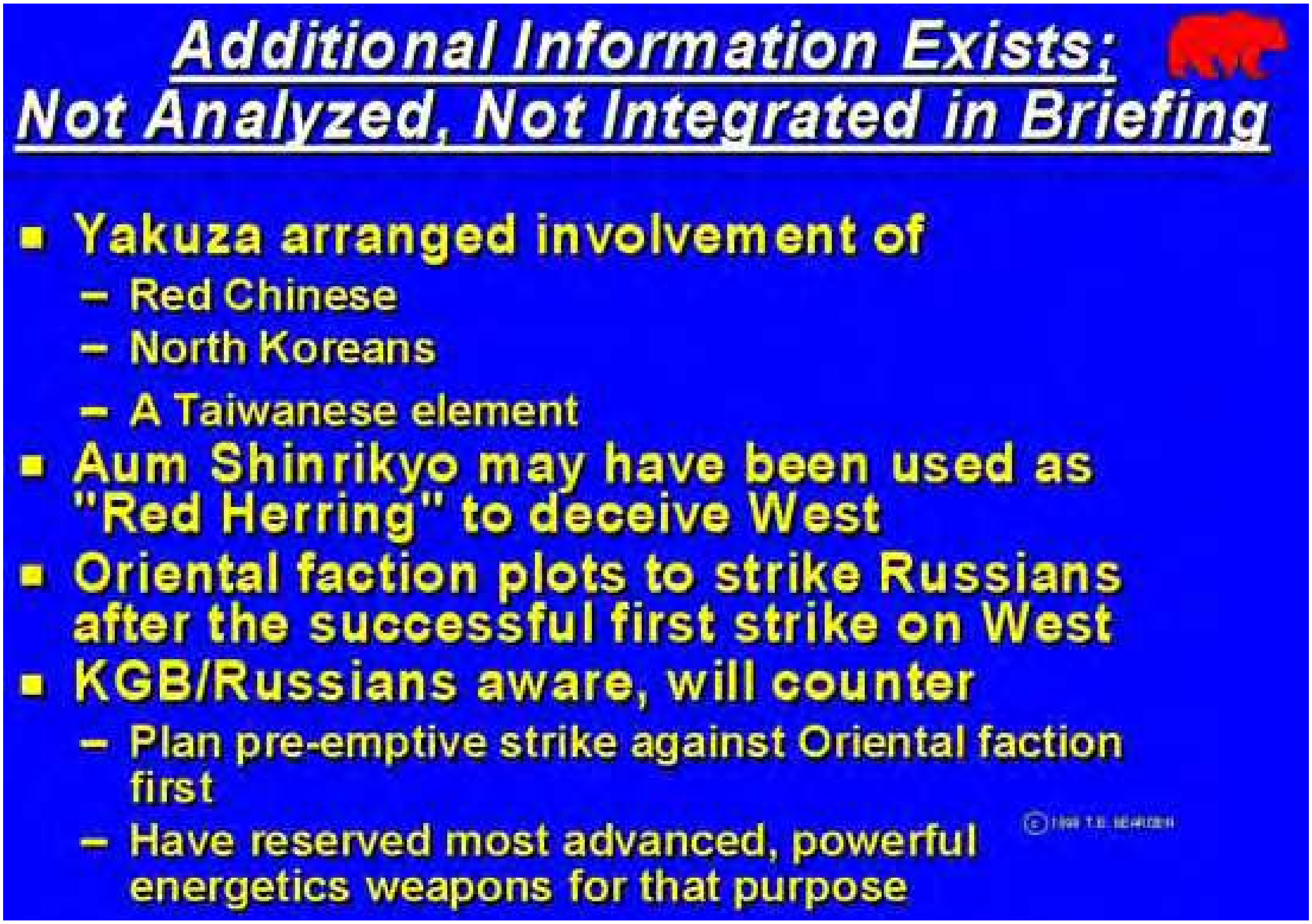

Additional Information Exists; Not Analyzed, Not Integrated in Briefing

- Yakuza arranged involvement of
  - Red Chinese
  - North Koreans
  - A Taiwanese element
- Aum Shinrikyo may have been used as "Red Herring" to deceive West
- Oriental faction plots to strike Russians after the successful first strike on West
- KGB/Russians aware, will counter
  - Plan pre-emptive strike against Oriental faction first
  - Have reserved most advanced, powerful energetics weapons for that purpose

The Tom Bearden Website

Help support the research

Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6

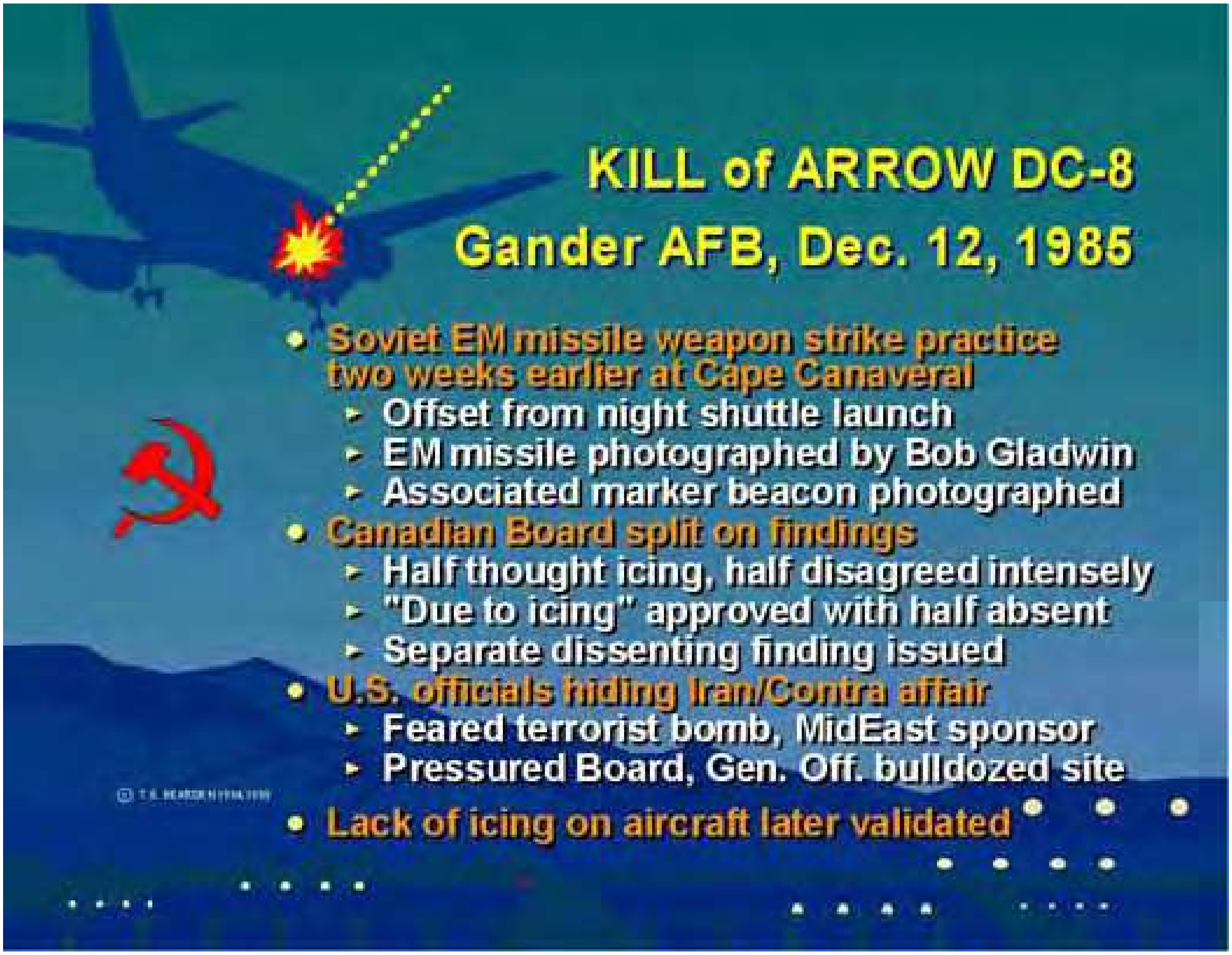

KILL of ARROW DC-8

Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. bulldozed site
- Lack of icing on aircraft later validated

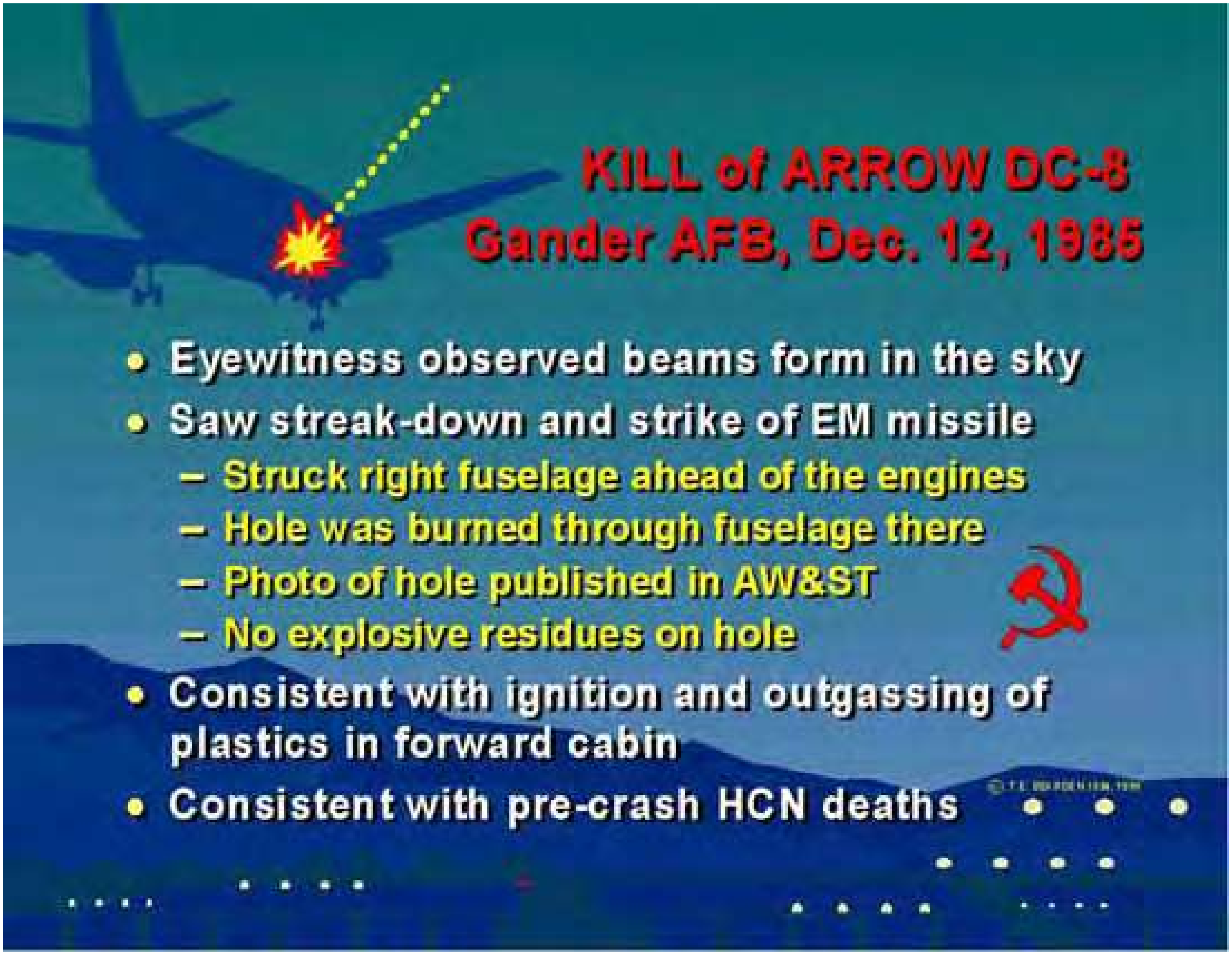

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Eyewitness observed beams form in the sky
- Saw streak-down and strike of EM missile
  - Struck right fuselage ahead of the engines
  - Hole was burned through fuselage there
  - Photo of hole published in AW&ST
  - No explosive residues on hole
- Consistent with ignition and outgassing of plastics in forward cabin
- Consistent with pre-crash HCN deaths

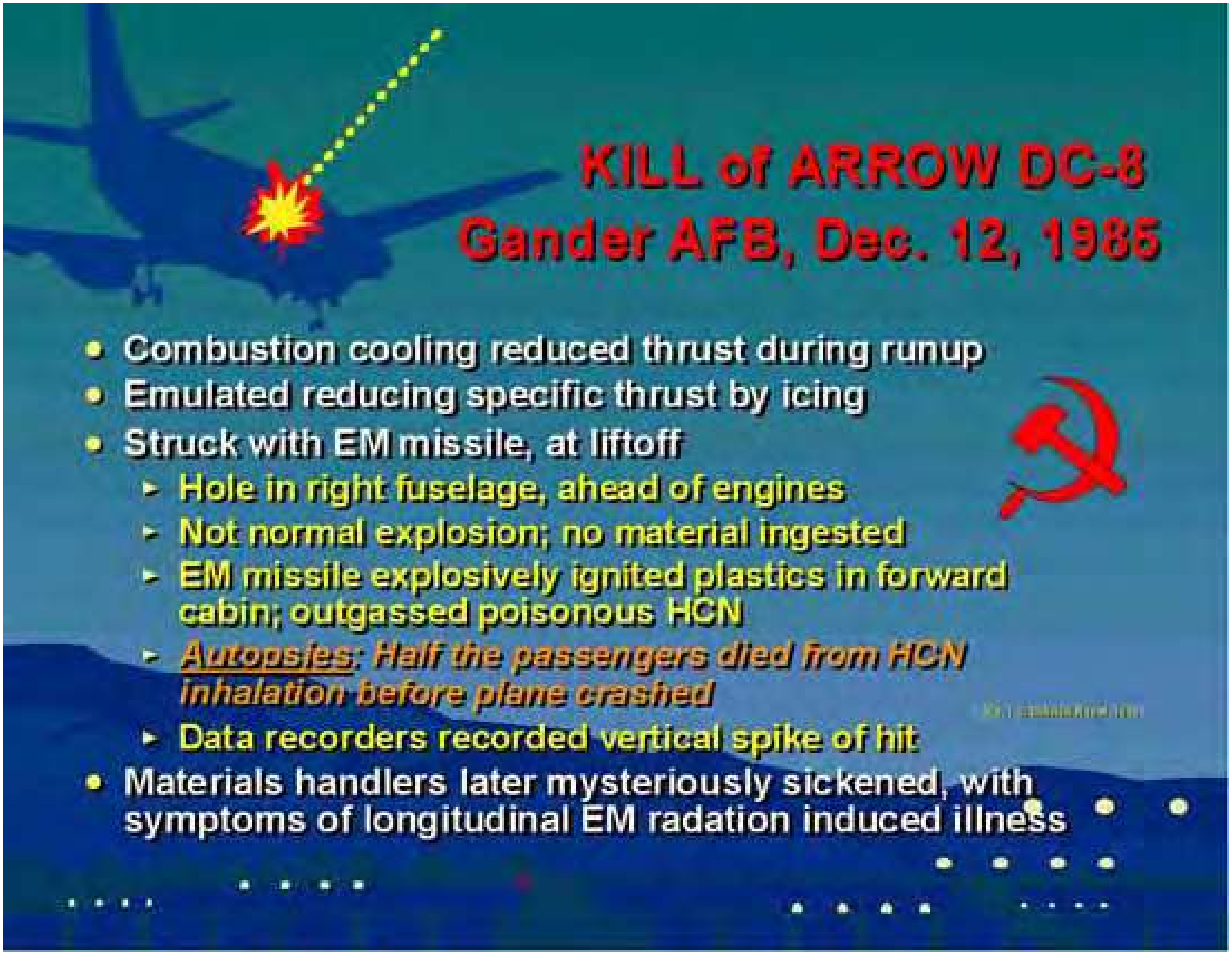

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation before plane crashed*
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of longitudinal EM radation induced illness

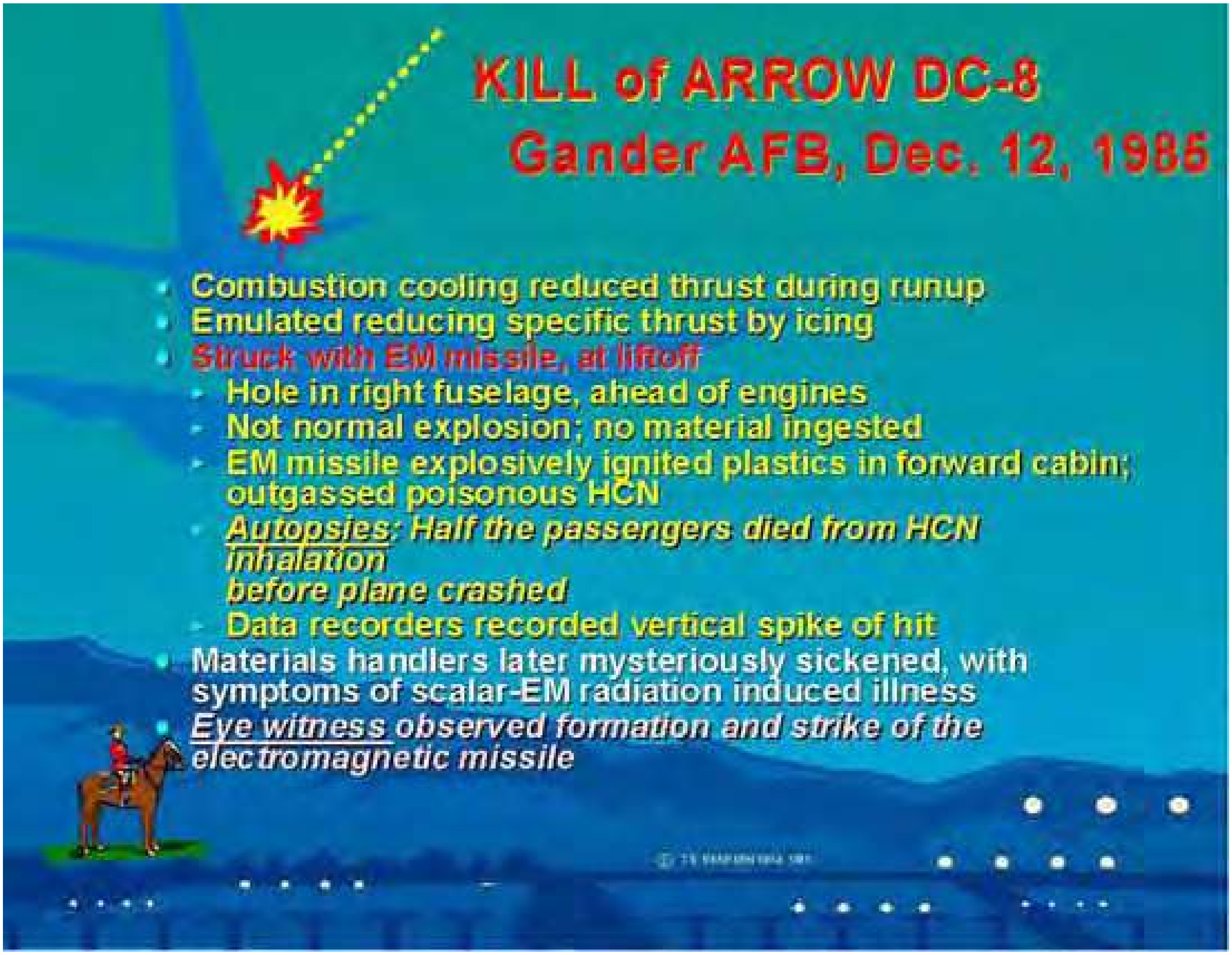

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation*
    before plane crashed
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of scalar-EM radiation induced illness
- *Eye witness observed formation and strike of the electromagnetic missile*

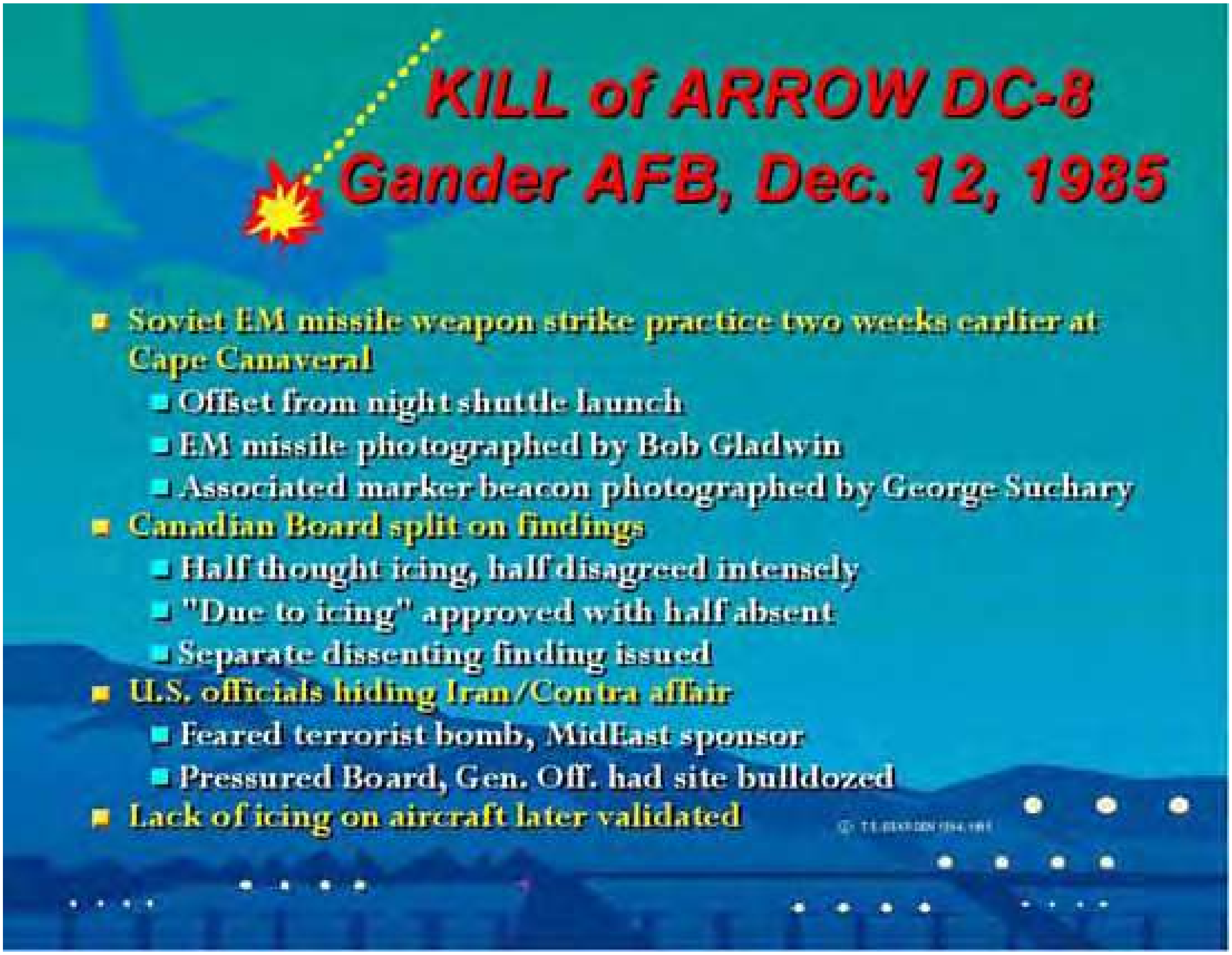

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed by George Suchary
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. had site bulldozed
- Lack of icing on aircraft later validated

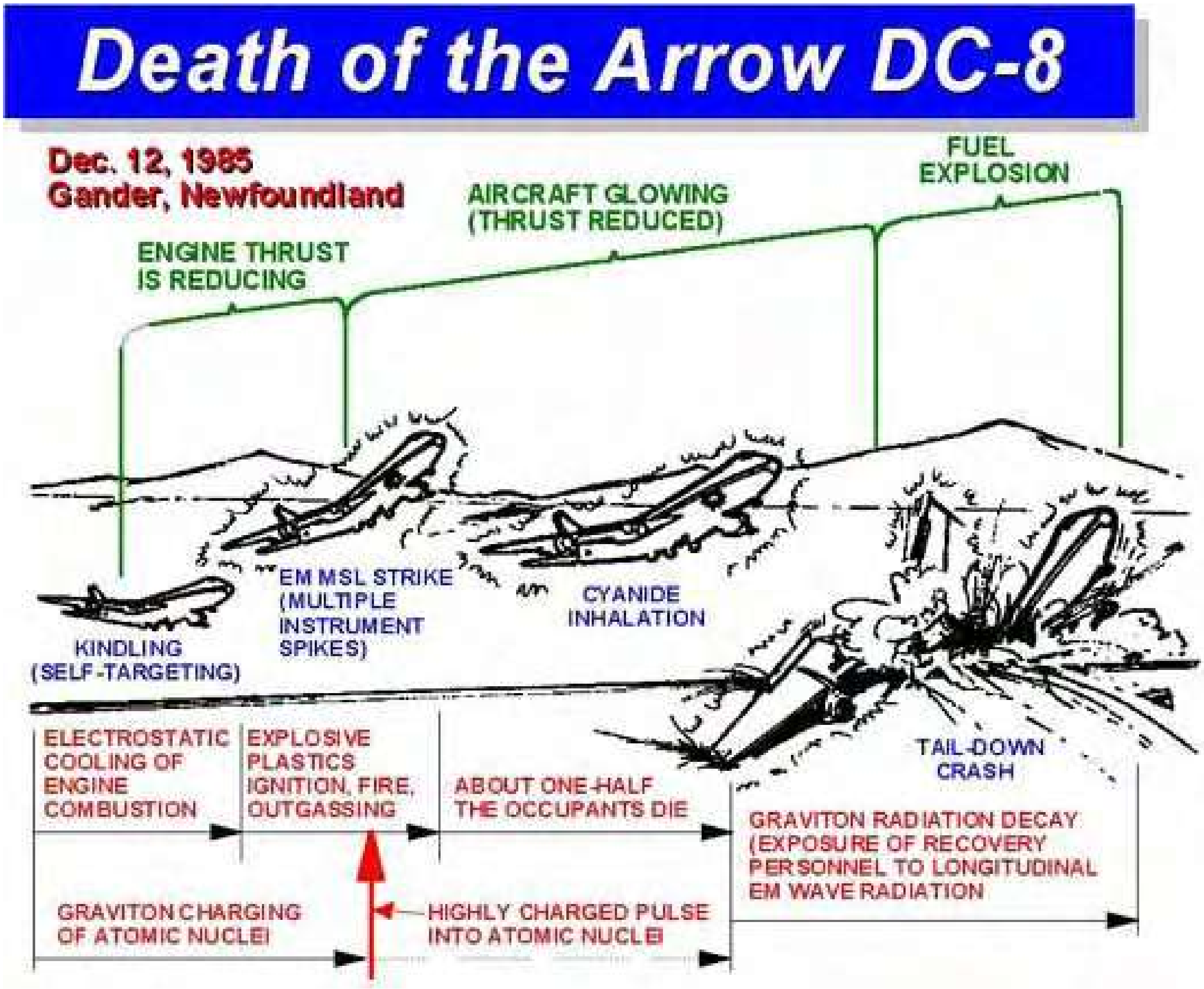
Death of the Arrow DC-8
Dec. 12, 1985
Gander, Newfoundland
ENGINE THRUST IS REDUCING
AIRCRAFT GLOWING (THRUST REDUCED)
FUEL EXPLOSION
KINDLING (SELF-TARGETING)
EM MSL STRIKE (MULTIPLE INSTRUMENT SPIKES)
CYANIDE INHALATION
TAIL-DOWN CRASH
ELECTROSTATIC COOLING OF ENGINE COMBUSTION
EXPLOSIVE PLASTICS IGNITION, FIRE, OUTGASSING
ABOUT ONE-HALF THE OCCUPANTS DIE
GRAVITON RADIATION DECAY (EXPOSURE OF RECOVERY PERSONNEL TO LONGITUDINAL EM WAVE RADIATION
GRAVITON CHARGING OF ATOMIC NUCLEI
HIGHLY CHARGED PULSE INTO ATOMIC NUCLEI

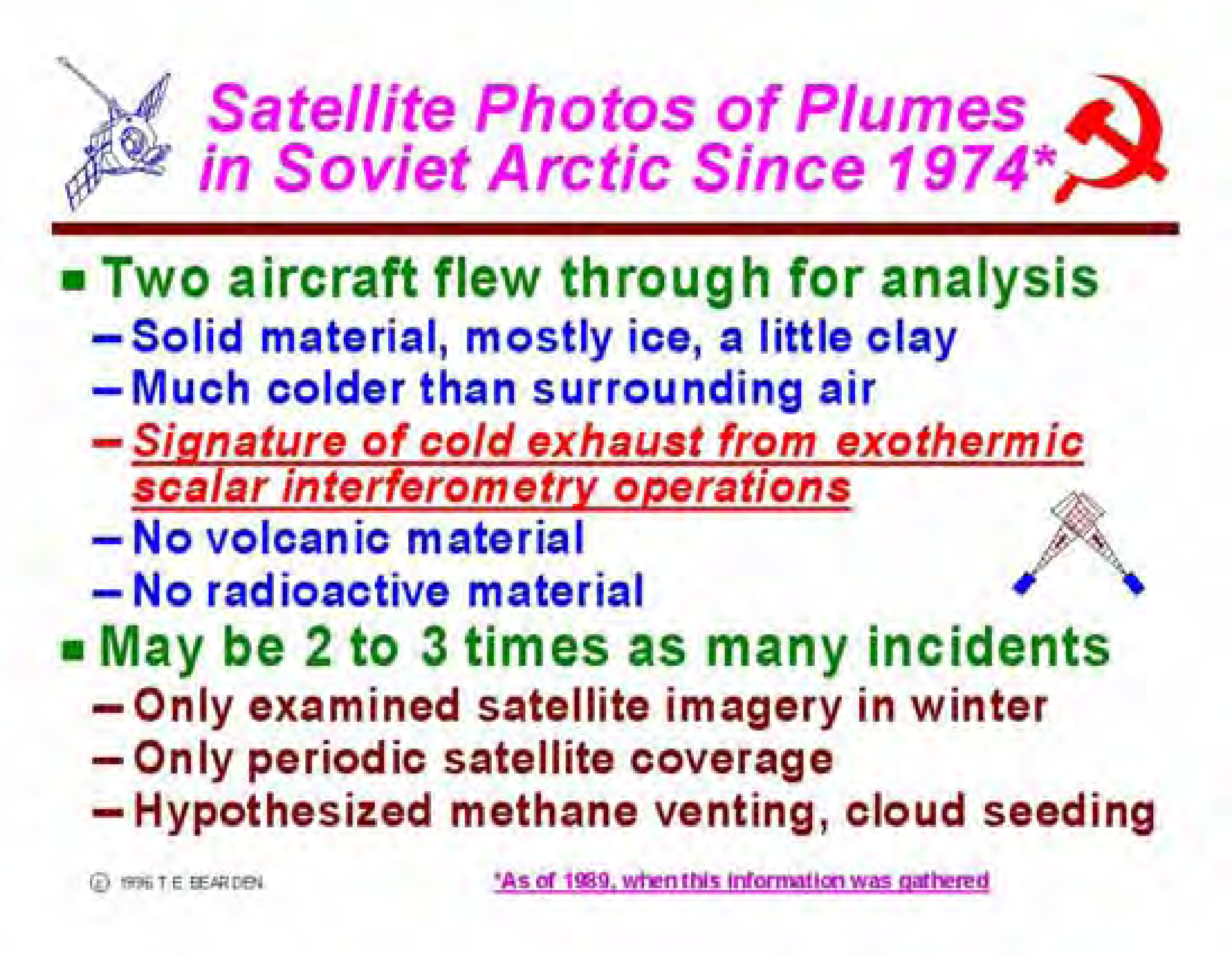

*Satellite Photos of Plumes in Soviet Arctic Since 1974**

- Two aircraft flew through for analysis
  - Solid material, mostly ice, a little clay
  - Much colder than surrounding air
  - *Signature of cold exhaust from exothermic scalar interferometry operations*
  - No volcanic material
  - No radioactive material
- May be 2 to 3 times as many incidents
  - Only examined satellite imagery in winter
  - Only periodic satellite coverage
  - Hypothesized methane venting, cloud seeding

© 1996 T.E. BEARDEN

*As of 1989, when this information was gathered

AUM SHINRIKYO
("SUPREME TRUTH" CULT)

Shoko Asahara, Leader

© 1995, 1998 T.E. BEARDEN

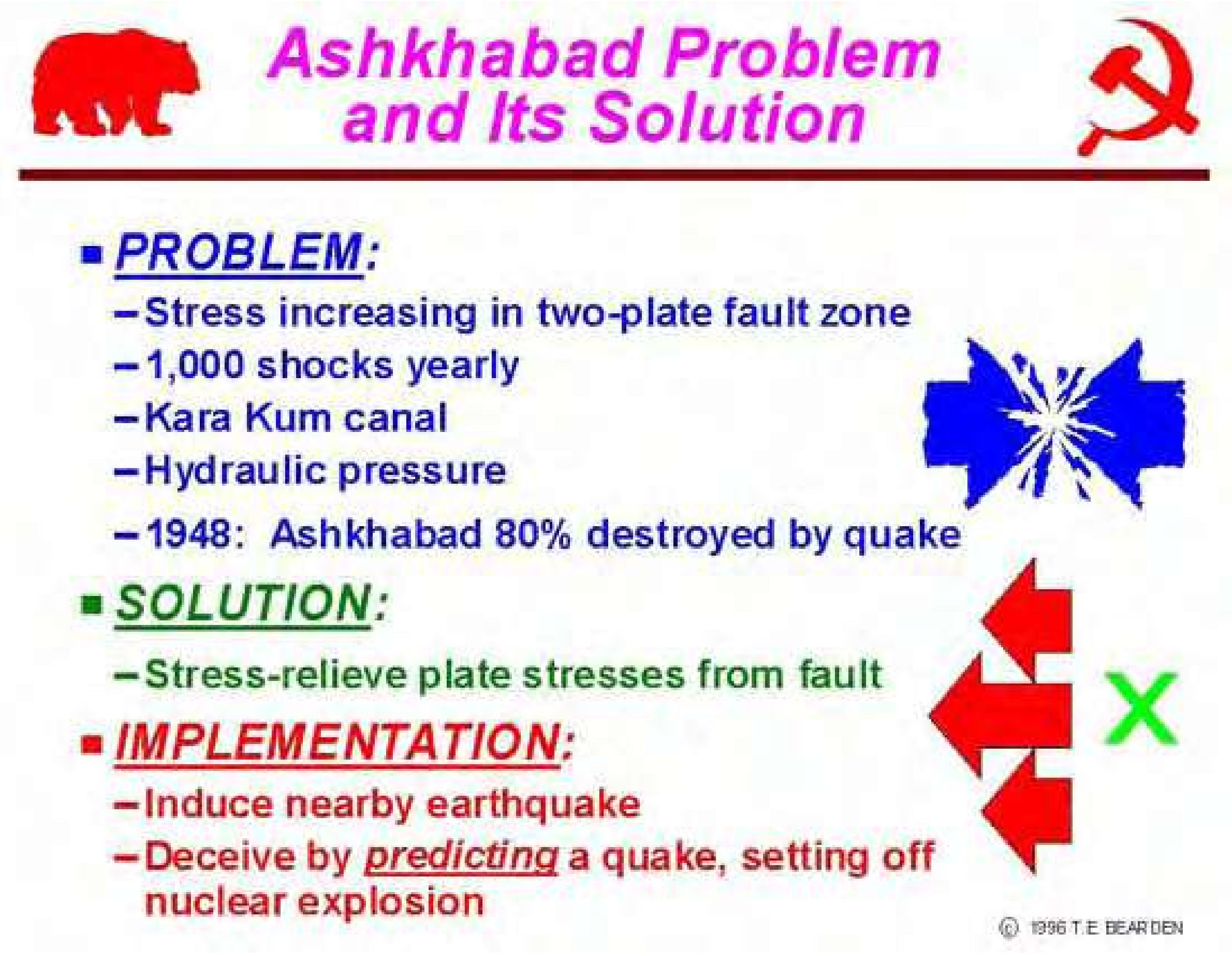

Ashkhabad Problem and Its Solution

- **PROBLEM:**
  - Stress increasing in two-plate fault zone
  - 1,000 shocks yearly
  - Kara Kum canal
  - Hydraulic pressure
  - 1948: Ashkhabad 80% destroyed by quake
- **SOLUTION:**
  - Stress-relieve plate stresses from fault
- **IMPLEMENTATION:**
  - Induce nearby earthquake
  - Deceive by *predicting* a quake, setting off nuclear explosion

© 1996 T.E. BEARDEN

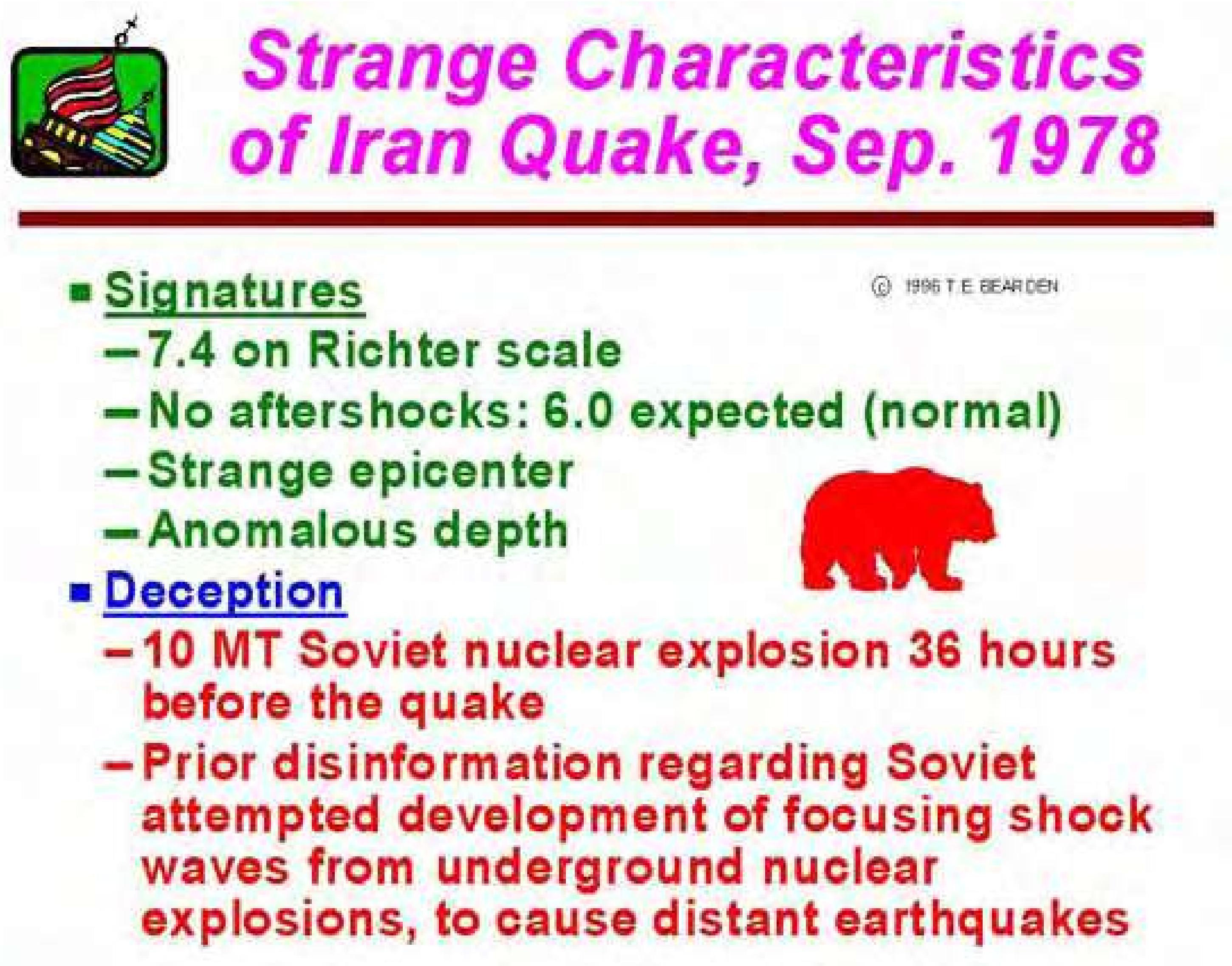

Strange Characteristics of Iran Quake, Sep. 1978

© 1996 T.E. BEARDEN

- Signatures
  - 7.4 on Richter scale
  - No aftershocks: 6.0 expected (normal)
  - Strange epicenter
  - Anomalous depth
- Deception
  - 10 MT Soviet nuclear explosion 36 hours before the quake
  - Prior disinformation regarding Soviet attempted development of focusing shock waves from underground nuclear explosions, to cause distant earthquakes

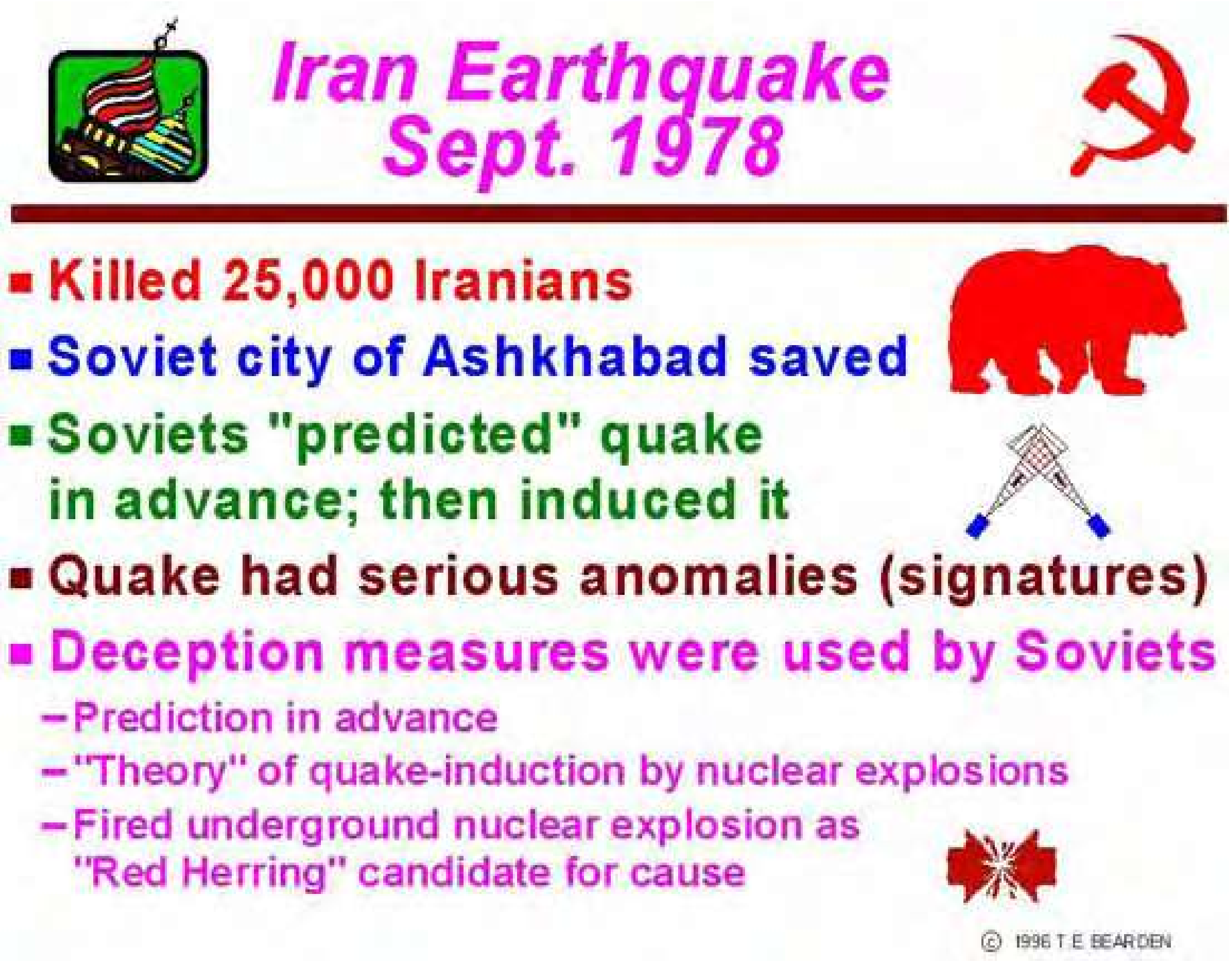

Iran Earthquake Sept. 1978

- Killed 25,000 Iranians
- Soviet city of Ashkhabad saved
- Soviets "predicted" quake in advance; then induced it
- Quake had serious anomalies (signatures)
- Deception measures were used by Soviets
  - –Prediction in advance
  - –"Theory" of quake-induction by nuclear explosions
  - –Fired underground nuclear explosion as "Red Herring" candidate for cause

© 1996 T E BEARDEN

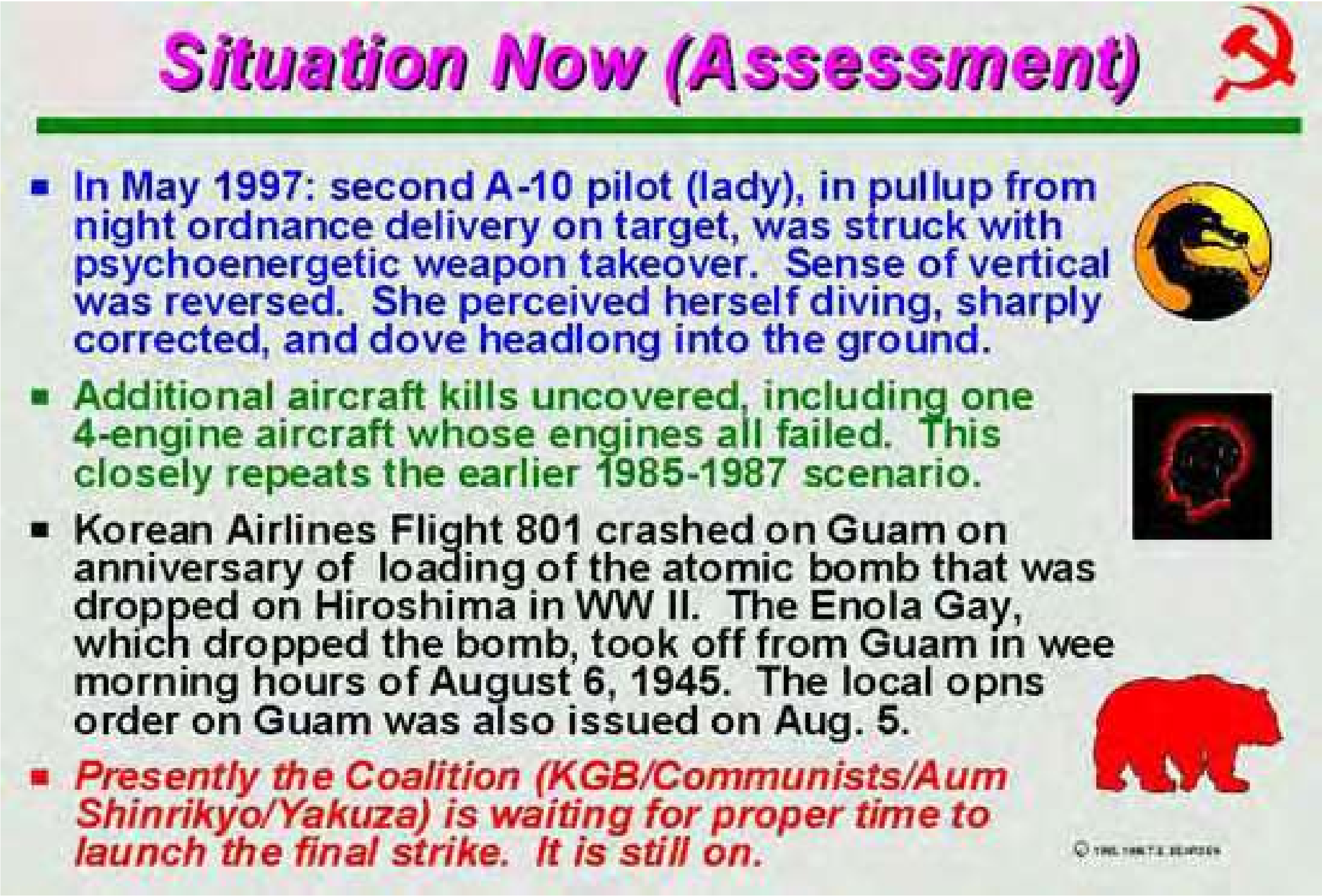

Situation Now (Assessment)

- In May 1997: second A-10 pilot (lady), in pullup from night ordnance delivery on target, was struck with psychoenergetic weapon takeover. Sense of vertical was reversed. She perceived herself diving, sharply corrected, and dove headlong into the ground.
- Additional aircraft kills uncovered, including one 4-engine aircraft whose engines all failed. This closely repeats the earlier 1985-1987 scenario.
- Korean Airlines Flight 801 crashed on Guam on anniversary of loading of the atomic bomb that was dropped on Hiroshima in WW II. The Enola Gay, which dropped the bomb, took off from Guam in wee morning hours of August 6, 1945. The local opns order on Guam was also issued on Aug. 5.
- *Presently the Coalition (KGB/Communists/Aum Shinrikyo/Yakuza) is waiting for proper time to launch the final strike. It is still on.*

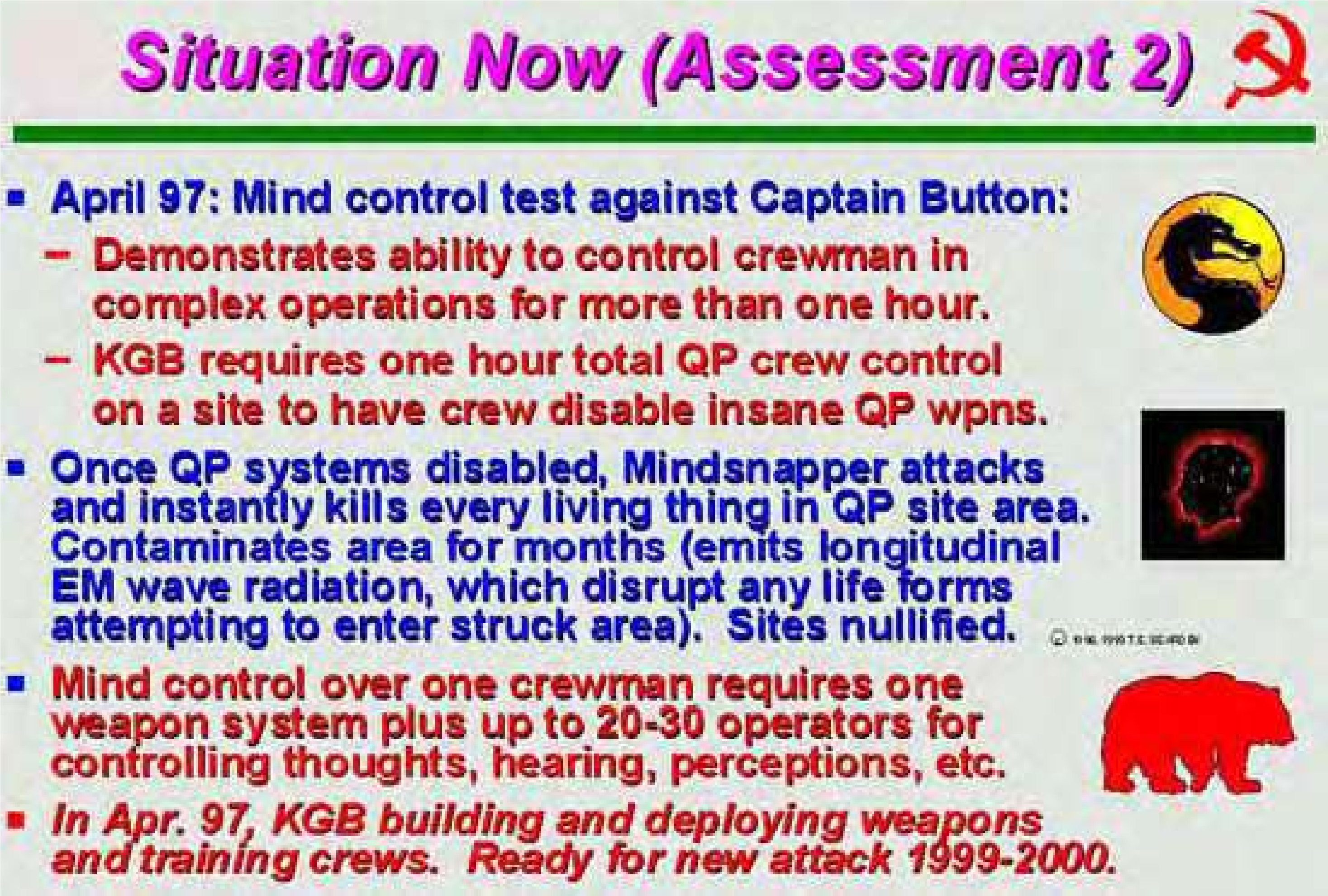

Situation Now (Assessment 2)

- **April 97: Mind control test against Captain Button:**
  - **Demonstrates ability to control crewman in complex operations for more than one hour.**
  - **KGB requires one hour total QP crew control on a site to have crew disable insane QP wpns.**
- **Once QP systems disabled, Mindsnapper attacks and instantly kills every living thing in QP site area. Contaminates area for months (emits longitudinal EM wave radiation, which disrupt any life forms attempting to enter struck area). Sites nullified.**
- **Mind control over one crewman requires one weapon system plus up to 20-30 operators for controlling thoughts, hearing, perceptions, etc.**
- ***In Apr. 97, KGB building and deploying weapons and training crews. Ready for new attack 1999-2000.***

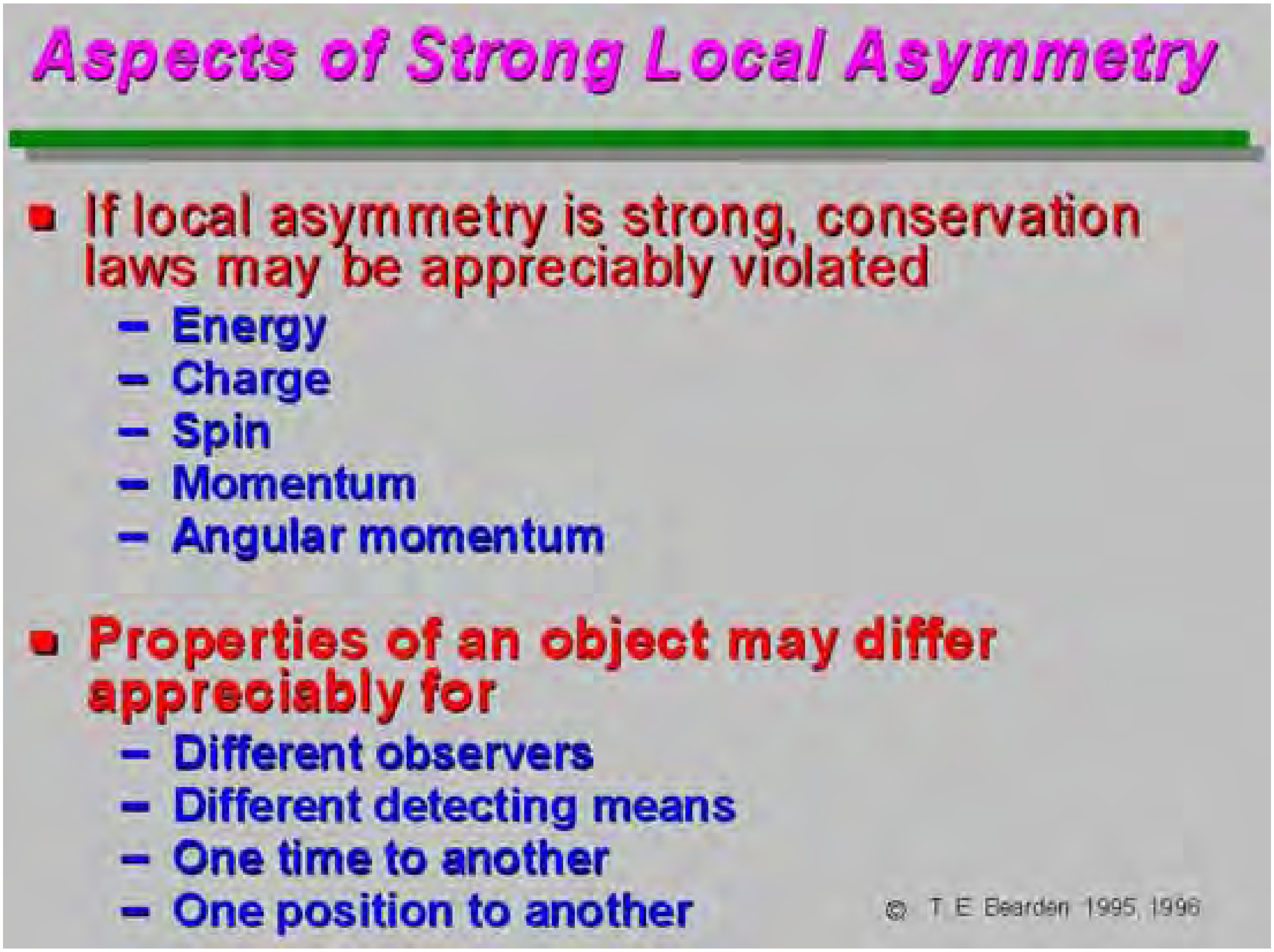

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

© T. E. Bearden 1995, 1996

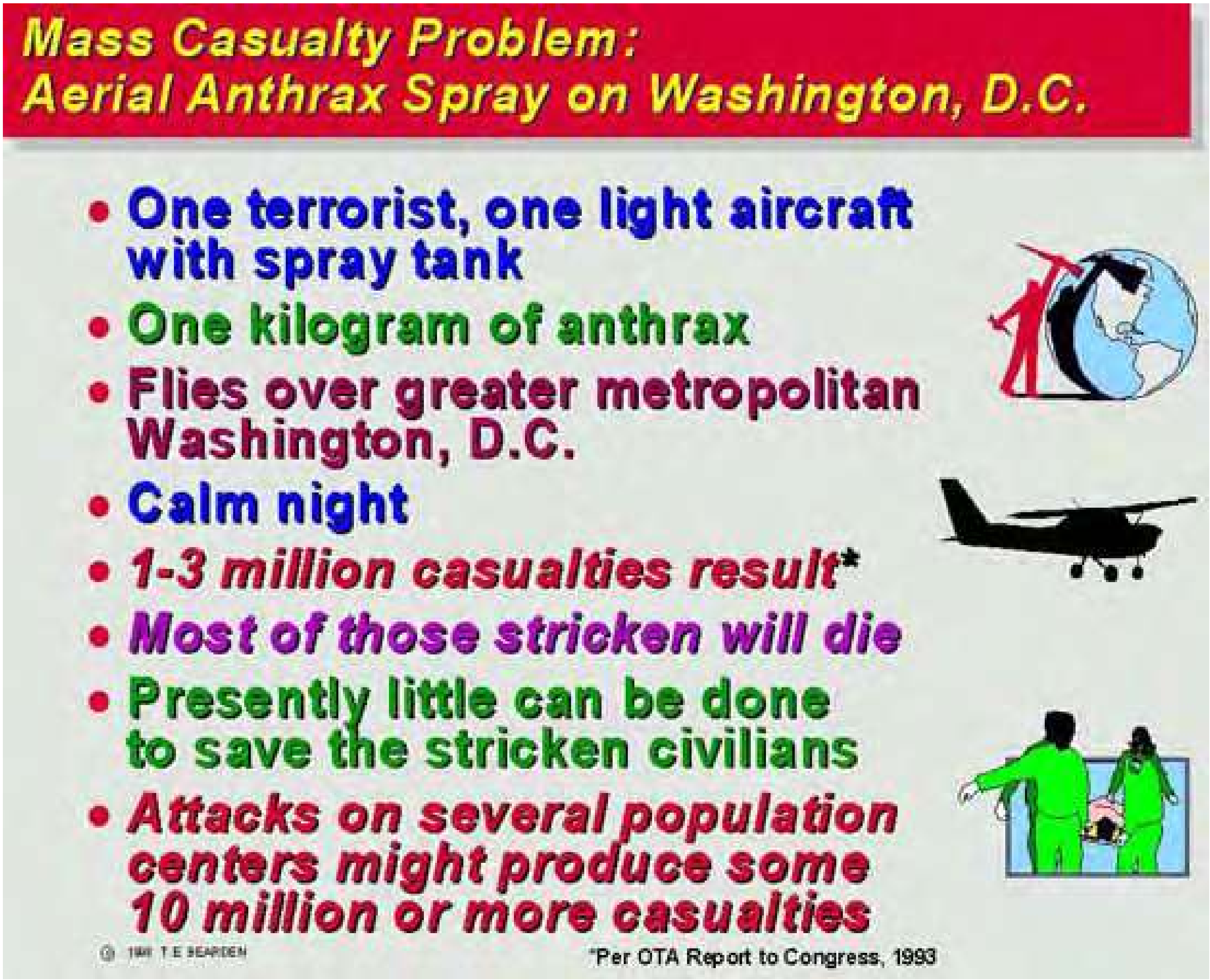

Mass Casualty Problem:
Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- One kilogram of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

© [illegible] T E BEARDEN

*Per OTA Report to Congress, 1993

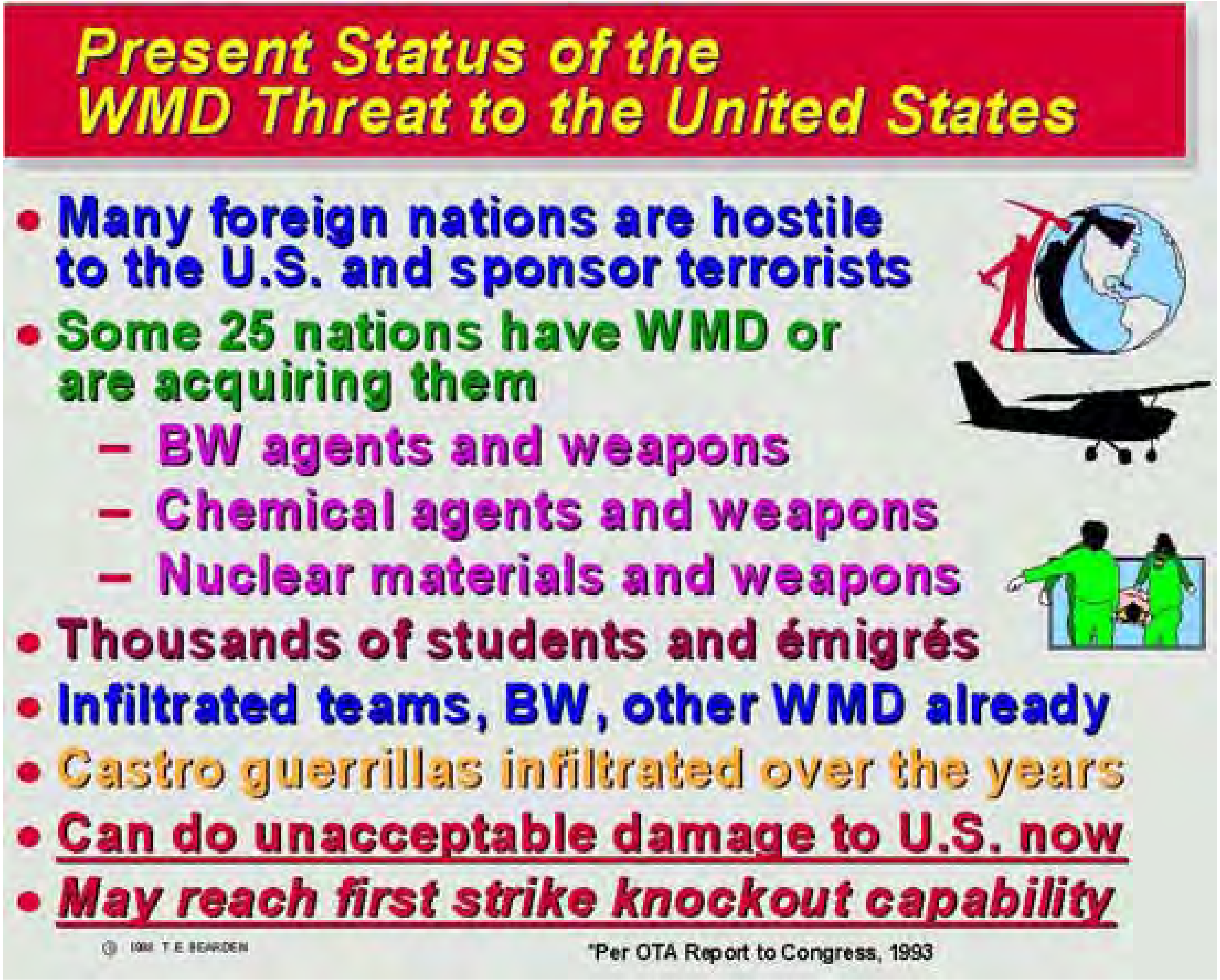

Present Status of the WMD Threat to the United States

- **Many foreign nations are hostile to the U.S. and sponsor terrorists**
- **Some 25 nations have WMD or are acquiring them**
  - **BW agents and weapons**
  - **Chemical agents and weapons**
  - **Nuclear materials and weapons**
- **Thousands of students and émigrés**
- **Infiltrated teams, BW, other WMD already**
- **Castro guerrillas infiltrated over the years**
- **Can do unacceptable damage to U.S. now**
- ***May reach first strike knockout capability***

© [illegible] T.E. BEARDEN

*Per OTA Report to Congress, 1993

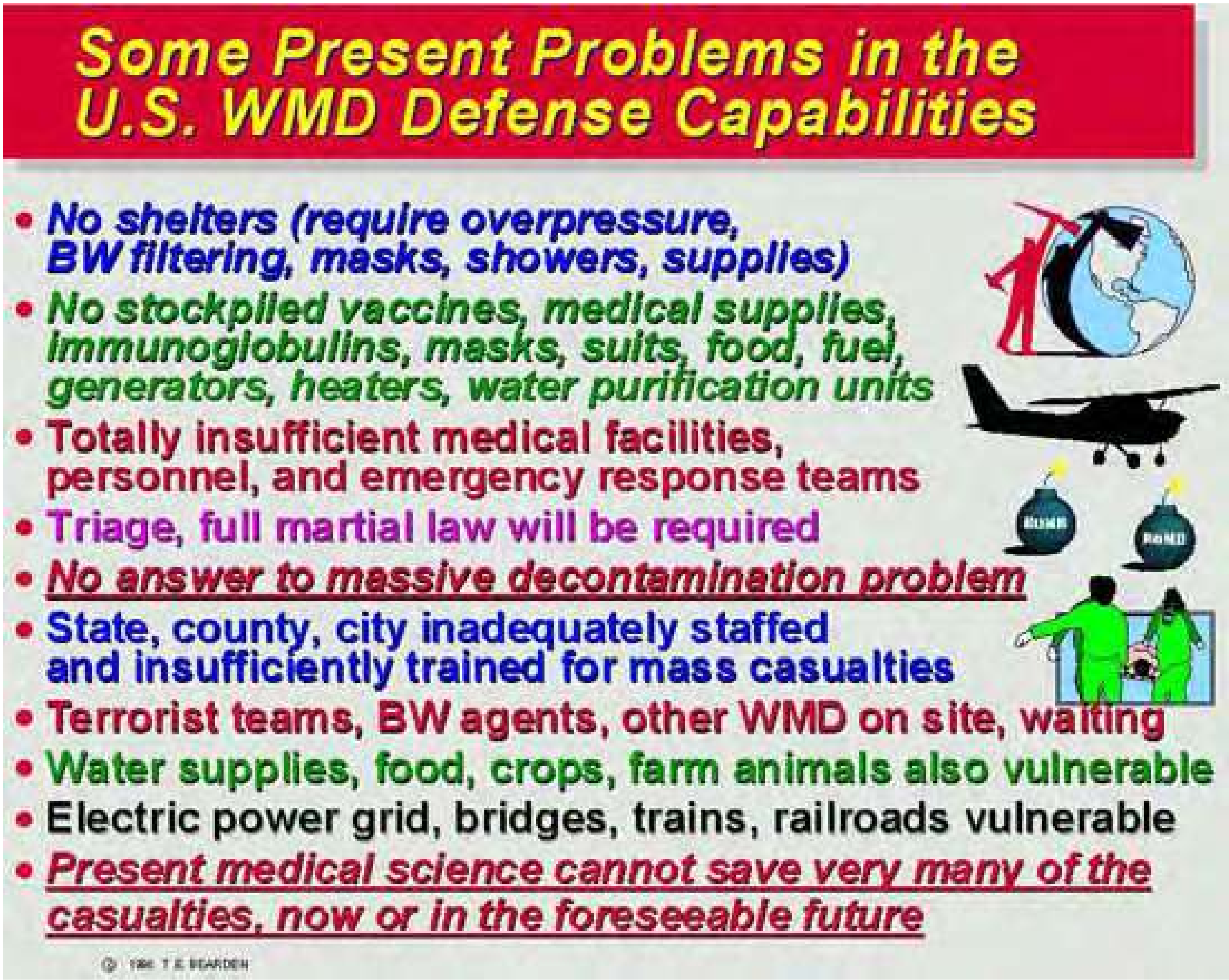

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© [illegible] T.E. BEARDEN

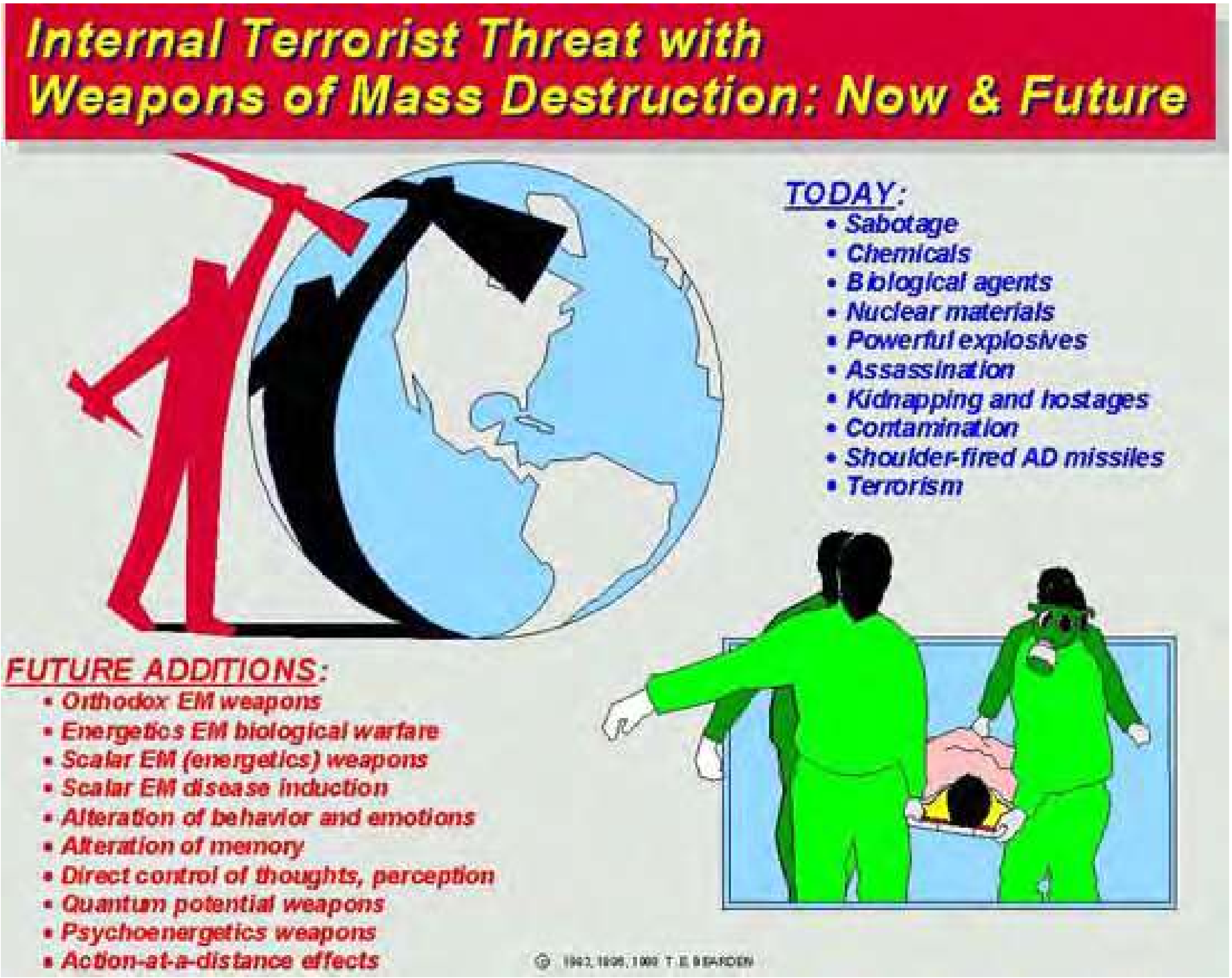

Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- Orthodox EM weapons
- Energetics EM biological warfare
- Scalar EM (energetics) weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct control of thoughts, perception
- Quantum potential weapons
- Psychoenergetics weapons
- Action-at-a-distance effects

© [illegible] T. E. BEARDEN

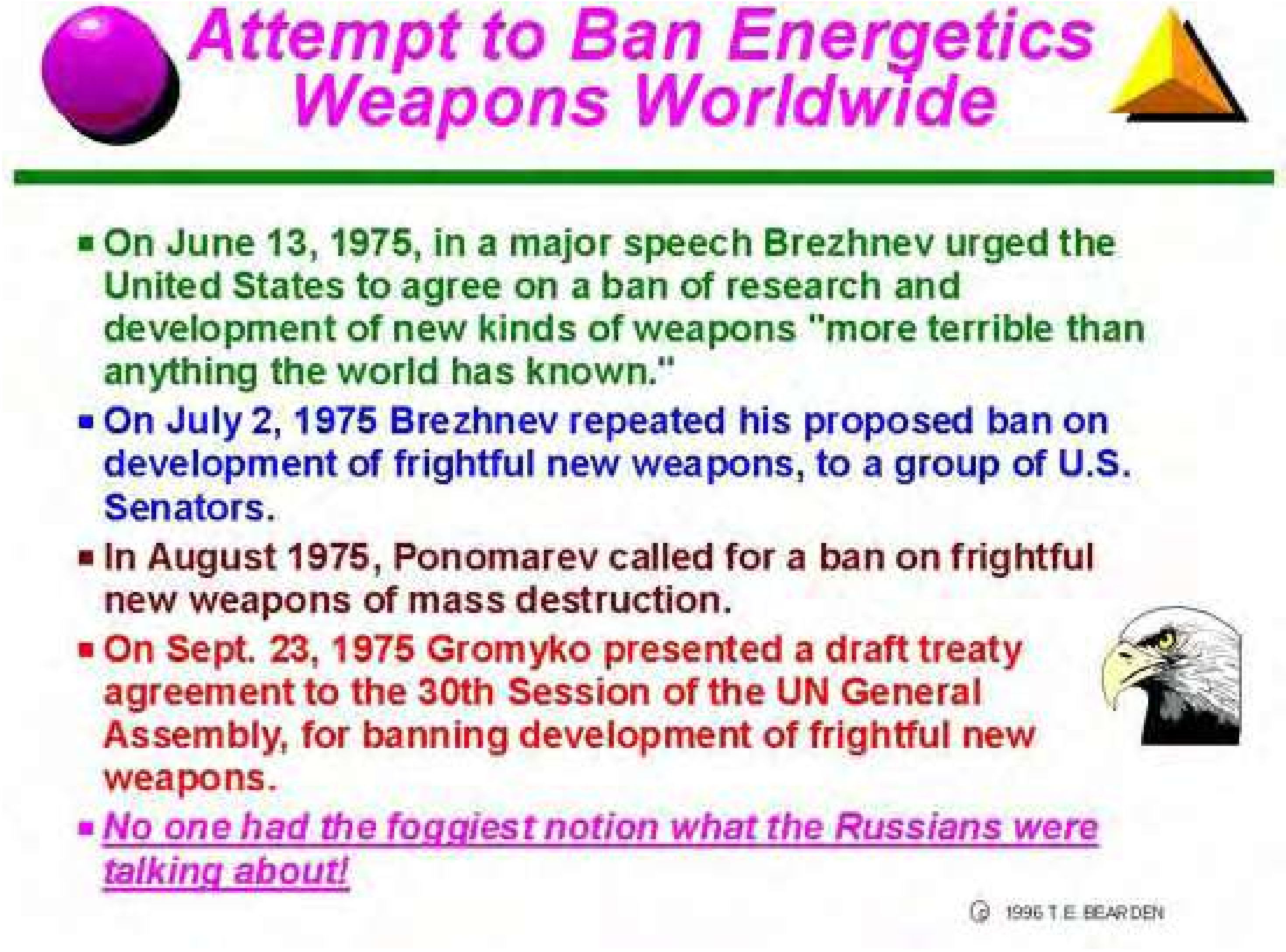

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T.E. BEARDEN

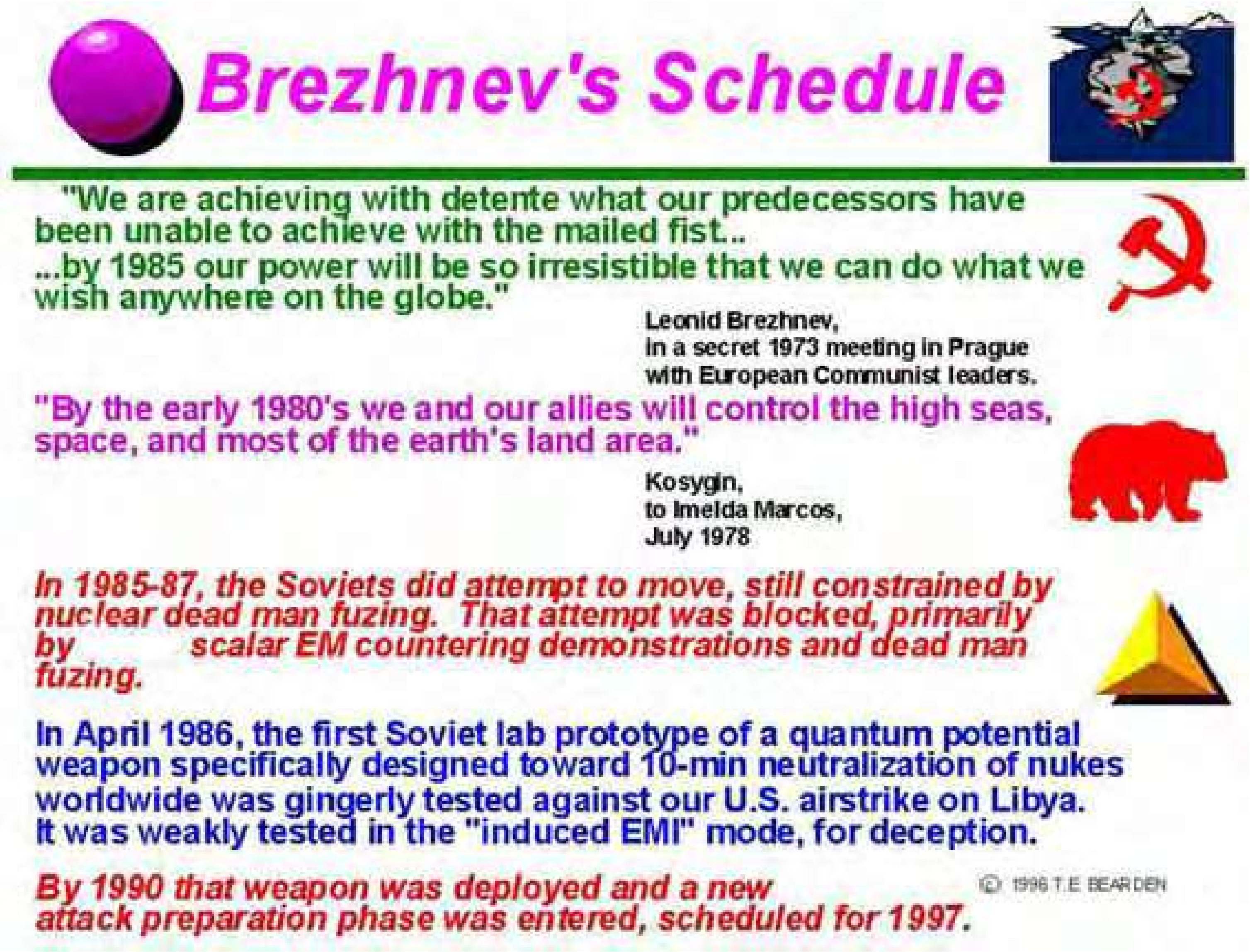

Brezhnev's Schedule

"We are achieving with detente what our predecessors have been unable to achieve with the mailed fist...
...by 1985 our power will be so irresistible that we can do what we wish anywhere on the globe."

Leonid Brezhnev,
in a secret 1973 meeting in Prague
with European Communist leaders.

"By the early 1980's we and our allies will control the high seas, space, and most of the earth's land area."

Kosygin,
to Imelda Marcos,
July 1978

In 1985-87, the Soviets did attempt to move, still constrained by nuclear dead man fuzing. That attempt was blocked, primarily by scalar EM countering demonstrations and dead man fuzing.

In April 1986, the first Soviet lab prototype of a quantum potential weapon specifically designed toward 10-min neutralization of nukes worldwide was gingerly tested against our U.S. airstrike on Libya. It was weakly tested in the "induced EMI" mode, for deception.

By 1990 that weapon was deployed and a new attack preparation phase was entered, scheduled for 1997.

© 1996 T.E. BEARDEN

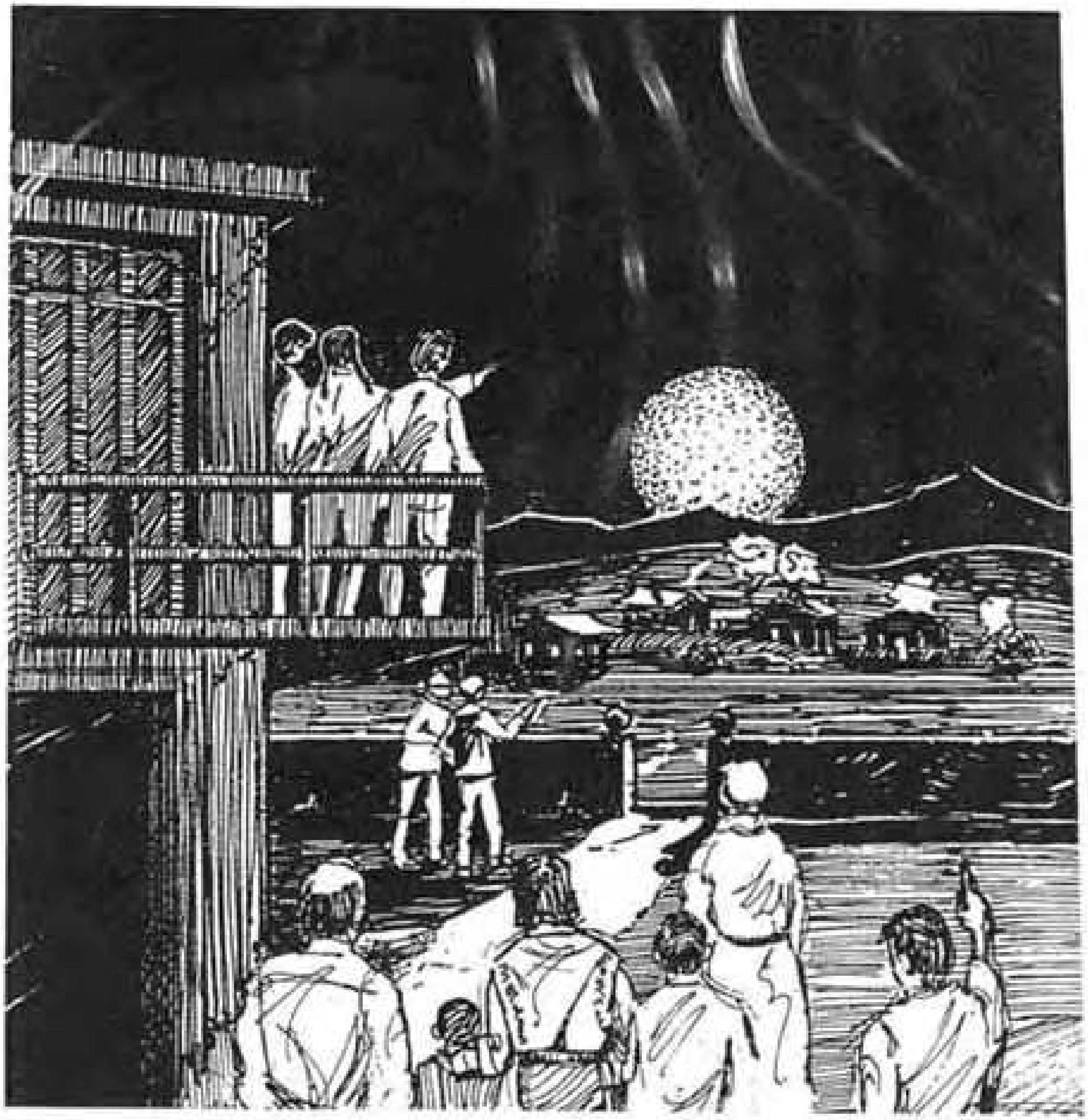

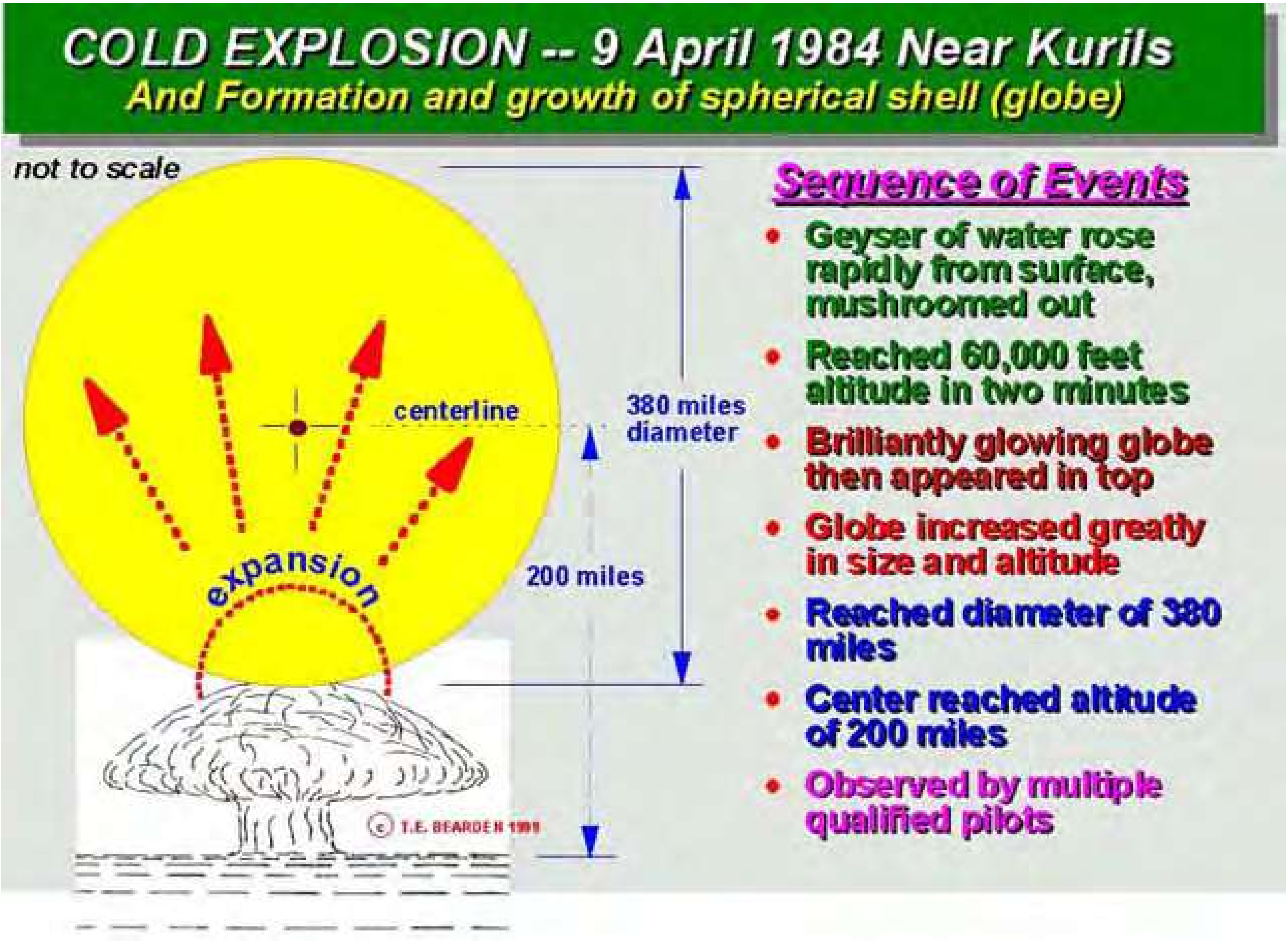

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale

centerline
expansion
380 miles diameter
200 miles
© T.E. BEARDEN 1991

Sequence of Events

- Geyser of water rose rapidly from surface, mushroomed out
- Reached 60,000 feet altitude in two minutes
- Brilliantly glowing globe then appeared in top
- Globe increased greatly in size and altitude
- Reached diameter of 380 miles
- Center reached altitude of 200 miles
- Observed by multiple qualified pilots

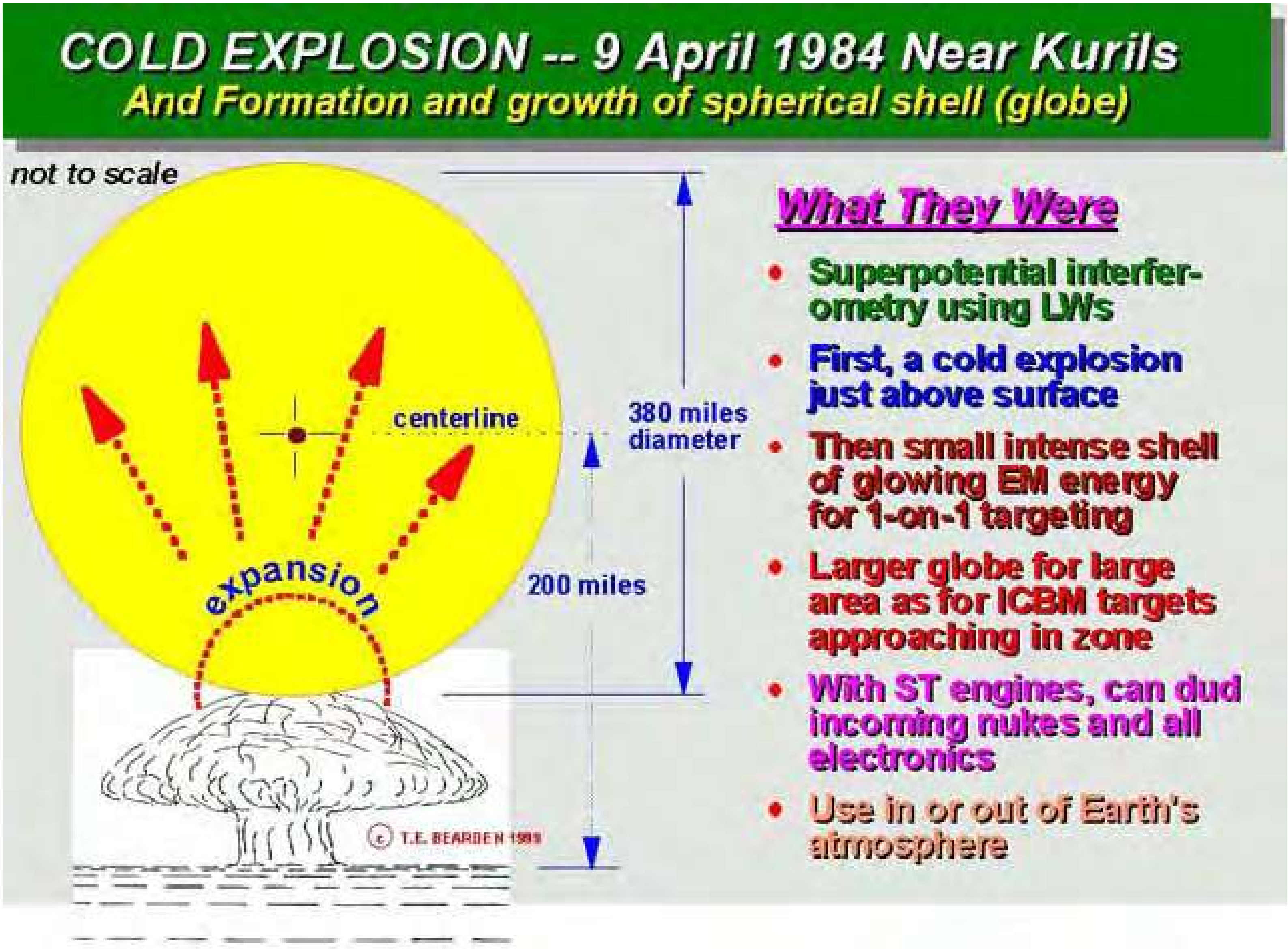

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
expansion
380 miles
diameter
200 miles
© T.E. BEARDEN 1999

What They Were

- Superpotential interferometry using LWs
- First, a cold explosion just above surface
- Then small intense shell of glowing EM energy for 1-on-1 targeting
- Larger globe for large area as for ICBM targets approaching in zone
- With ST engines, can dud incoming nukes and all electronics
- Use in or out of Earth's atmosphere

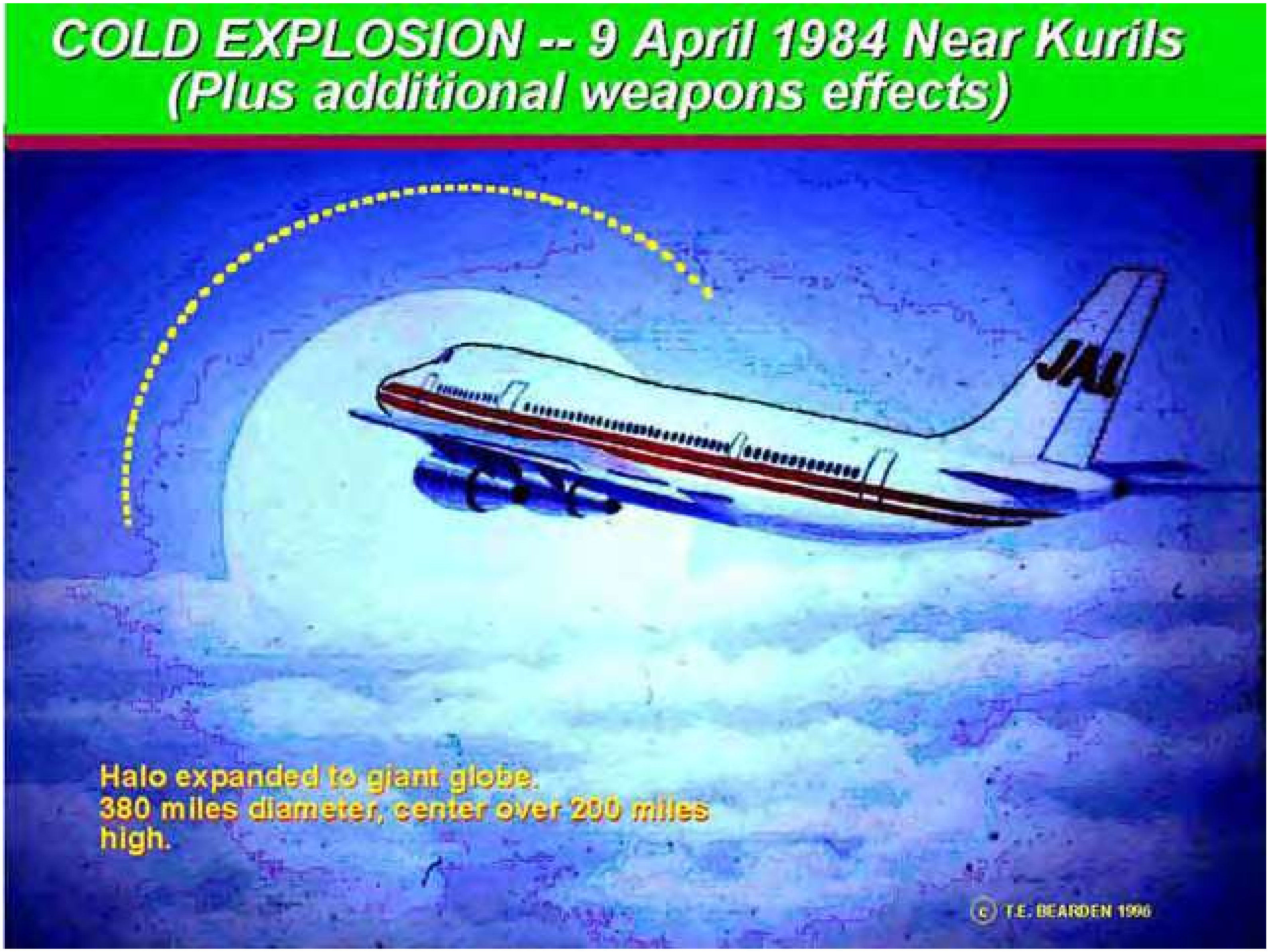

COLD EXPLOSION -- 9 April 1984 Near Kurils
(Plus additional weapons effects)

Halo expanded to giant globe.
380 miles diameter, center over 200 miles high.

© T.E. BEARDEN 1996

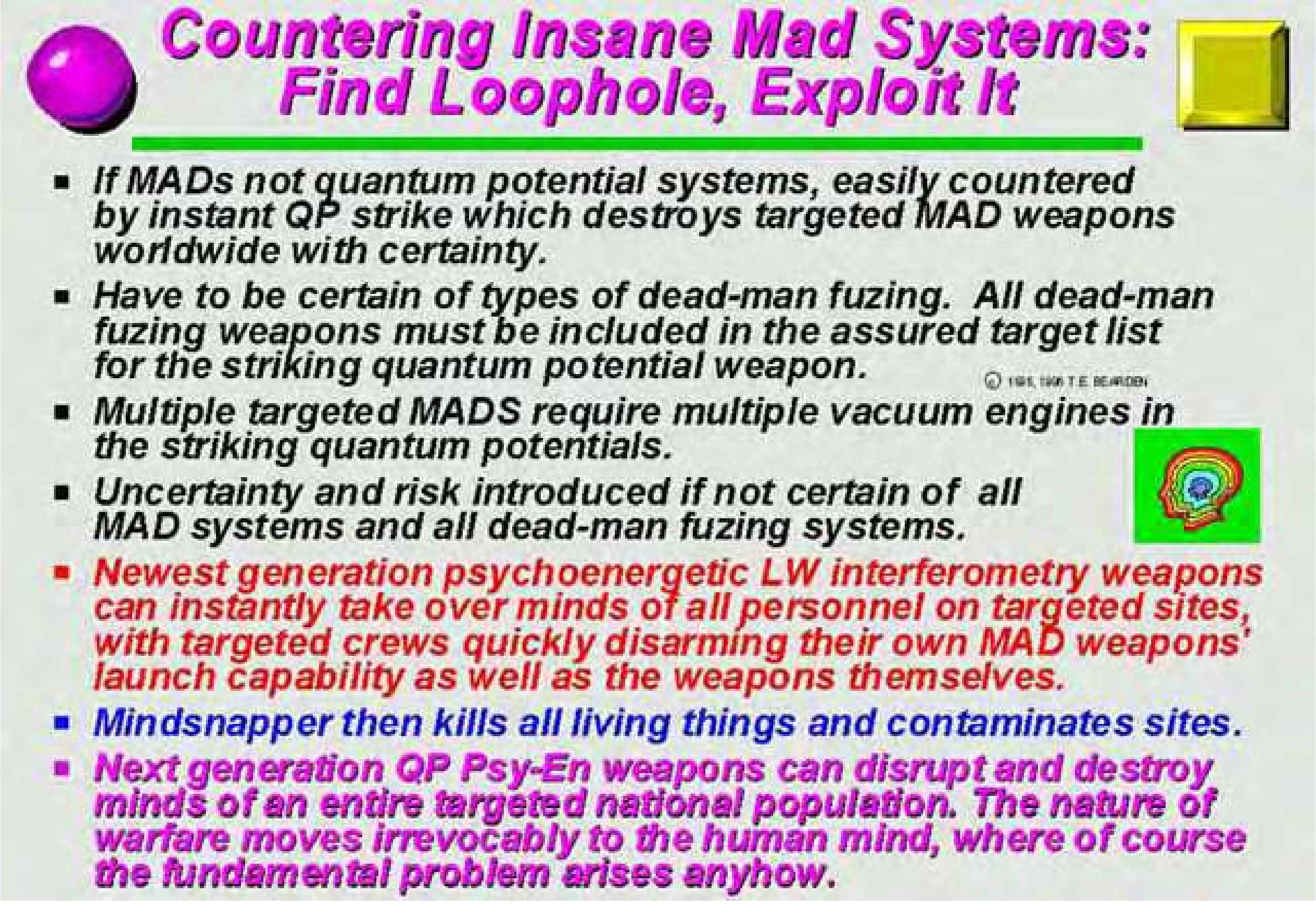

Countering Insane Mad Systems: Find Loophole, Exploit It

- *If MADs not quantum potential systems, easily countered by instant QP strike which destroys targeted MAD weapons worldwide with certainty.*
- *Have to be certain of types of dead-man fuzing. All dead-man fuzing weapons must be included in the assured target list for the striking quantum potential weapon.*
- *Multiple targeted MADS require multiple vacuum engines in the striking quantum potentials.*
- *Uncertainty and risk introduced if not certain of all MAD systems and all dead-man fuzing systems.*
- *Newest generation psychoenergetic LW interferometry weapons can instantly take over minds of all personnel on targeted sites, with targeted crews quickly disarming their own MAD weapons' launch capability as well as the weapons themselves.*
- *Mindsnapper then kills all living things and contaminates sites.*
- *Next generation QP Psy-En weapons can disrupt and destroy minds of an entire targeted national population. The nature of warfare moves irrevocably to the human mind, where of course the fundamental problem arises anyhow.*

© [illegible] T.E. BEARDEN

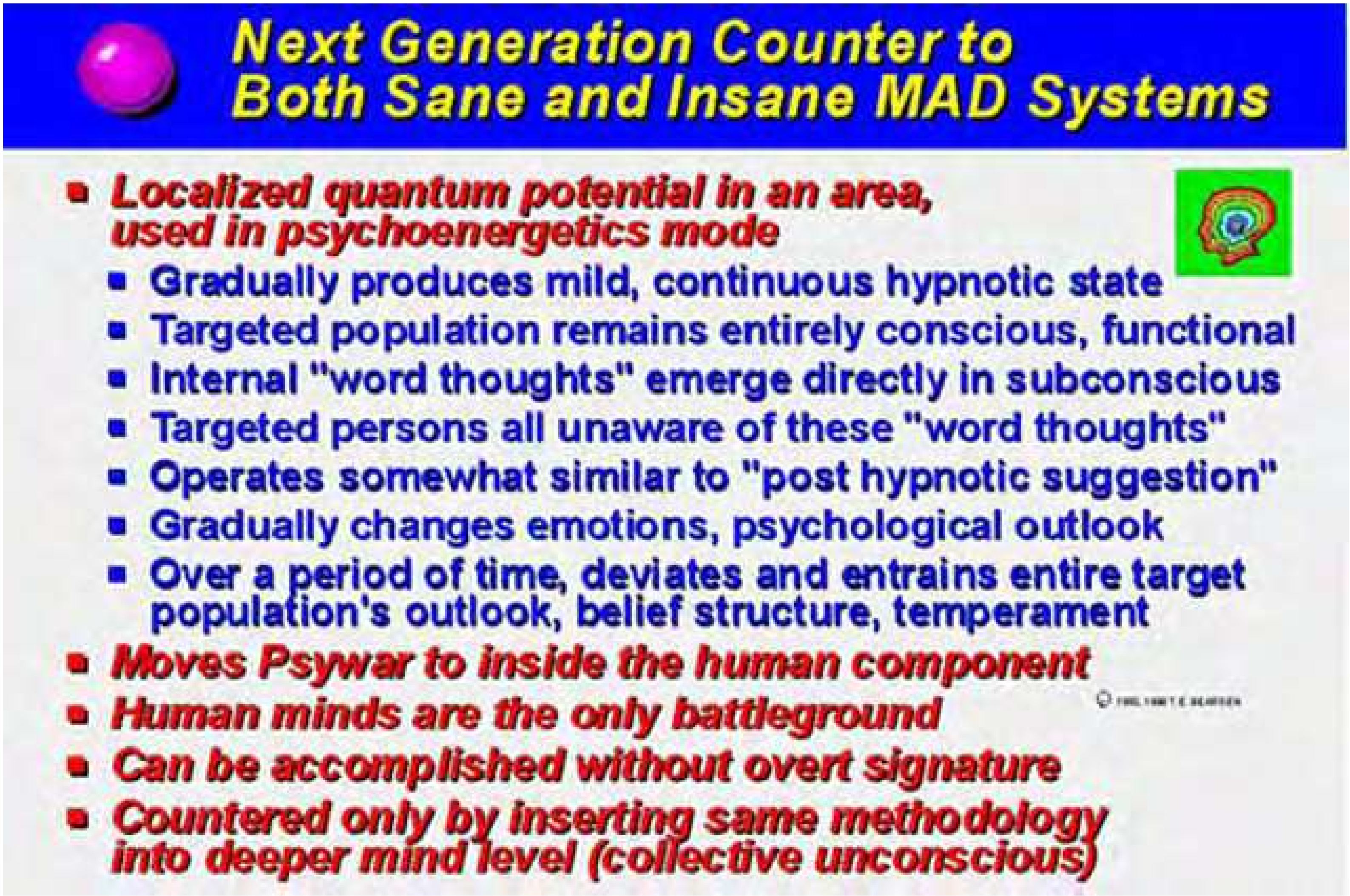

Next Generation Counter to Both Sane and Insane MAD Systems

- *Localized quantum potential in an area, used in psychoenergetics mode*
  - Gradually produces mild, continuous hypnotic state
  - Targeted population remains entirely conscious, functional
  - Internal "word thoughts" emerge directly in subconscious
  - Targeted persons all unaware of these "word thoughts"
  - Operates somewhat similar to "post hypnotic suggestion"
  - Gradually changes emotions, psychological outlook
  - Over a period of time, deviates and entrains entire target population's outlook, belief structure, temperament
- *Moves Psywar to inside the human component*
- *Human minds are the only battleground*
- *Can be accomplished without overt signature*
- *Countered only by inserting same methodology into deeper mind level (collective unconscious)*

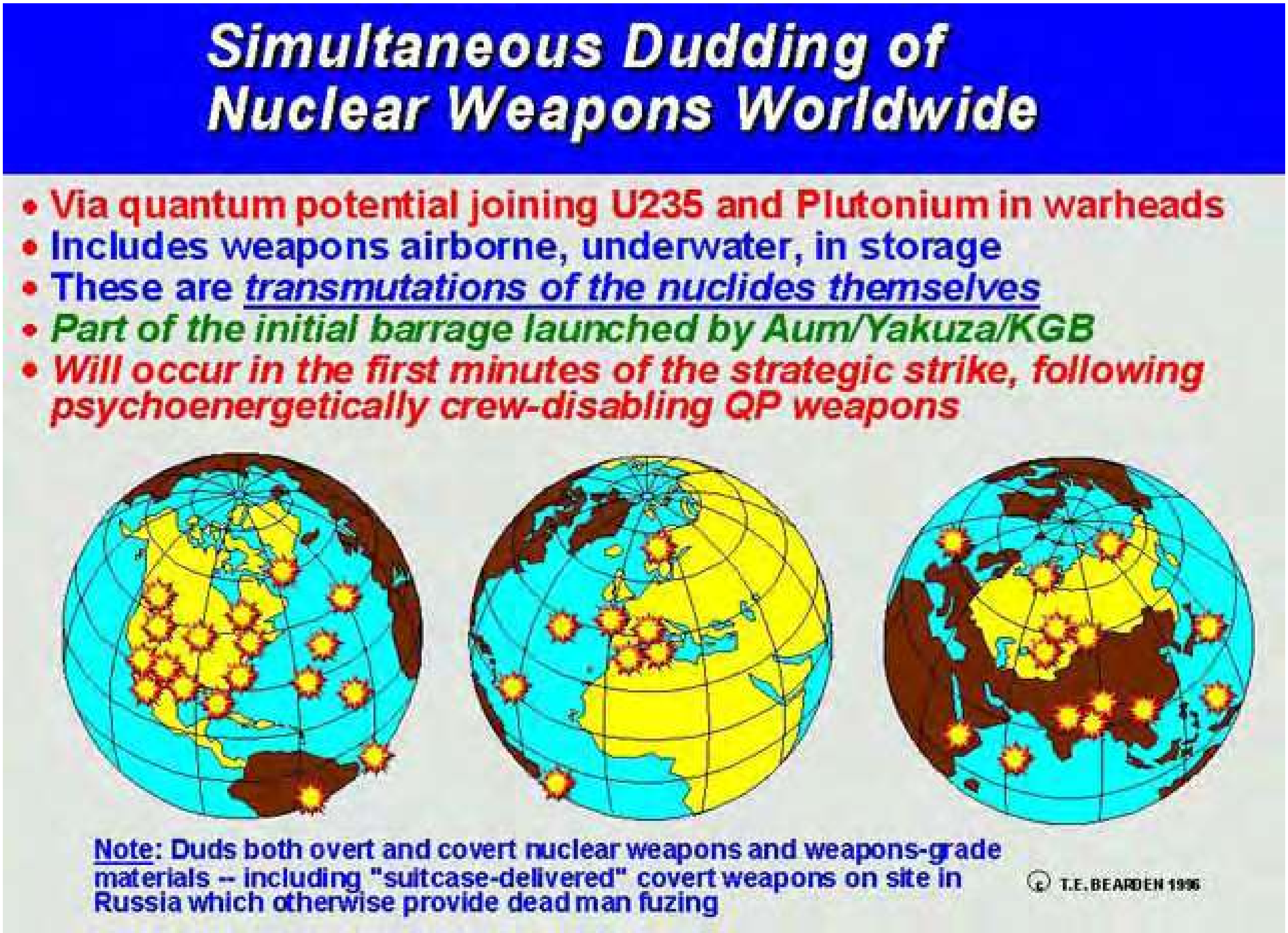

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium in warheads
- Includes weapons airborne, underwater, in storage
- These are *transmutations of the nuclides themselves*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike, following psychoenergetically crew-disabling QP weapons*

Note: Duds both overt and covert nuclear weapons and weapons-grade materials -- including "suitcase-delivered" covert weapons on site in Russia which otherwise provide dead man fuzing

© T.E. BEARDEN 1996

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium warheads
- Includes weapons airborne, underwater, in storage
- These are not explosions, but are *transmutations and increased fissioning rates*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike*

© T.E. BEARDEN 1996

Note: Duds both overt and covert nuclear weapons and weapons-grade materials.

The Tom Bearden Website

Help support the research

Operation Divine Wind II

Slide 1

Slide 2

Slide 3

Slide 4

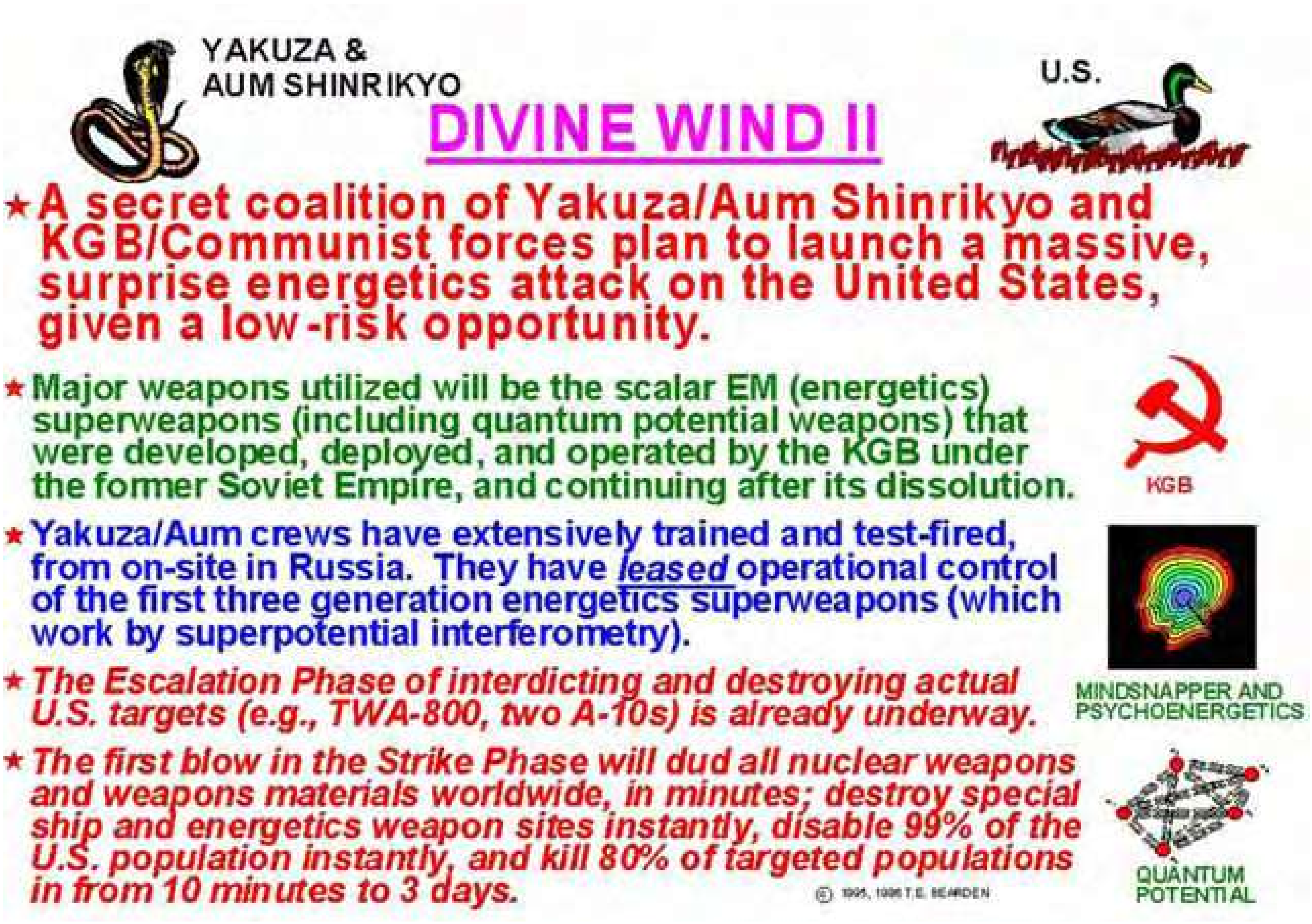
YAKUZA &
AUM SHINRIKYO
U.S.
KGB
MINDSNAPPER AND
PSYCHOENERGETICS
QUANTUM
POTENTIAL

DIVINE WIND II

- ★ **A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plan to launch a massive, surprise energetics attack on the United States, given a low-risk opportunity.**
- ★ Major weapons utilized will be the scalar EM (energetics) superweapons (including quantum potential weapons) that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.
- ★ Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the first three generation energetics superweapons (which work by superpotential interferometry).
- ★ *The Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800, two A-10s) is already underway.*
- ★ *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, disable 99% of the U.S. population instantly, and kill 80% of targeted populations in from 10 minutes to 3 days.*

© 1995, 1996 T.E. BEARDEN

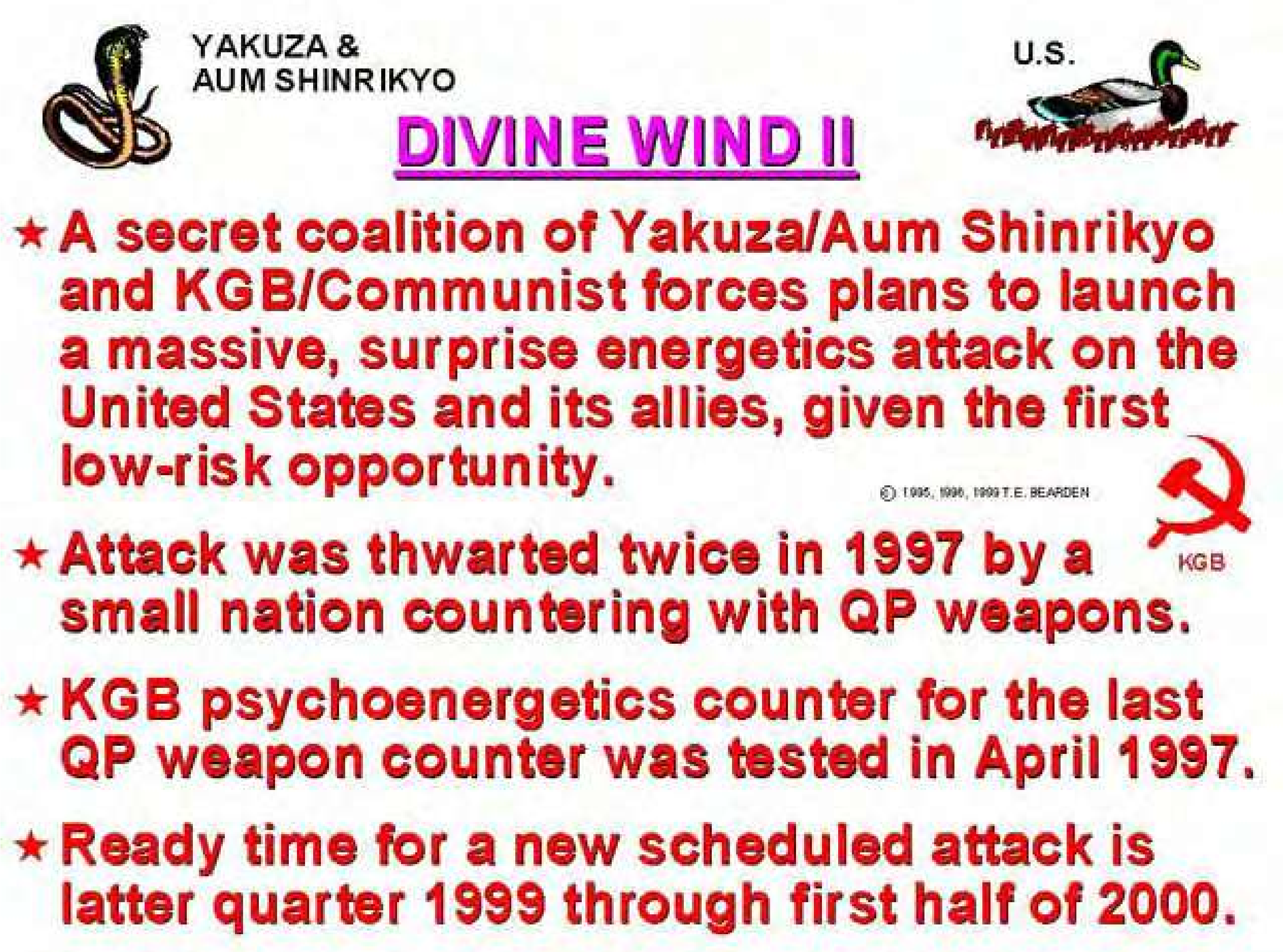

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

- ★ A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plans to launch a massive, surprise energetics attack on the United States and its allies, given the first low-risk opportunity.

© 1995, 1996, 1999 T.E. BEARDEN

KGB

- ★ Attack was thwarted twice in 1997 by a small nation countering with QP weapons.
- ★ KGB psychoenergetics counter for the last QP weapon counter was tested in April 1997.
- ★ Ready time for a new scheduled attack is latter quarter 1999 through first half of 2000.

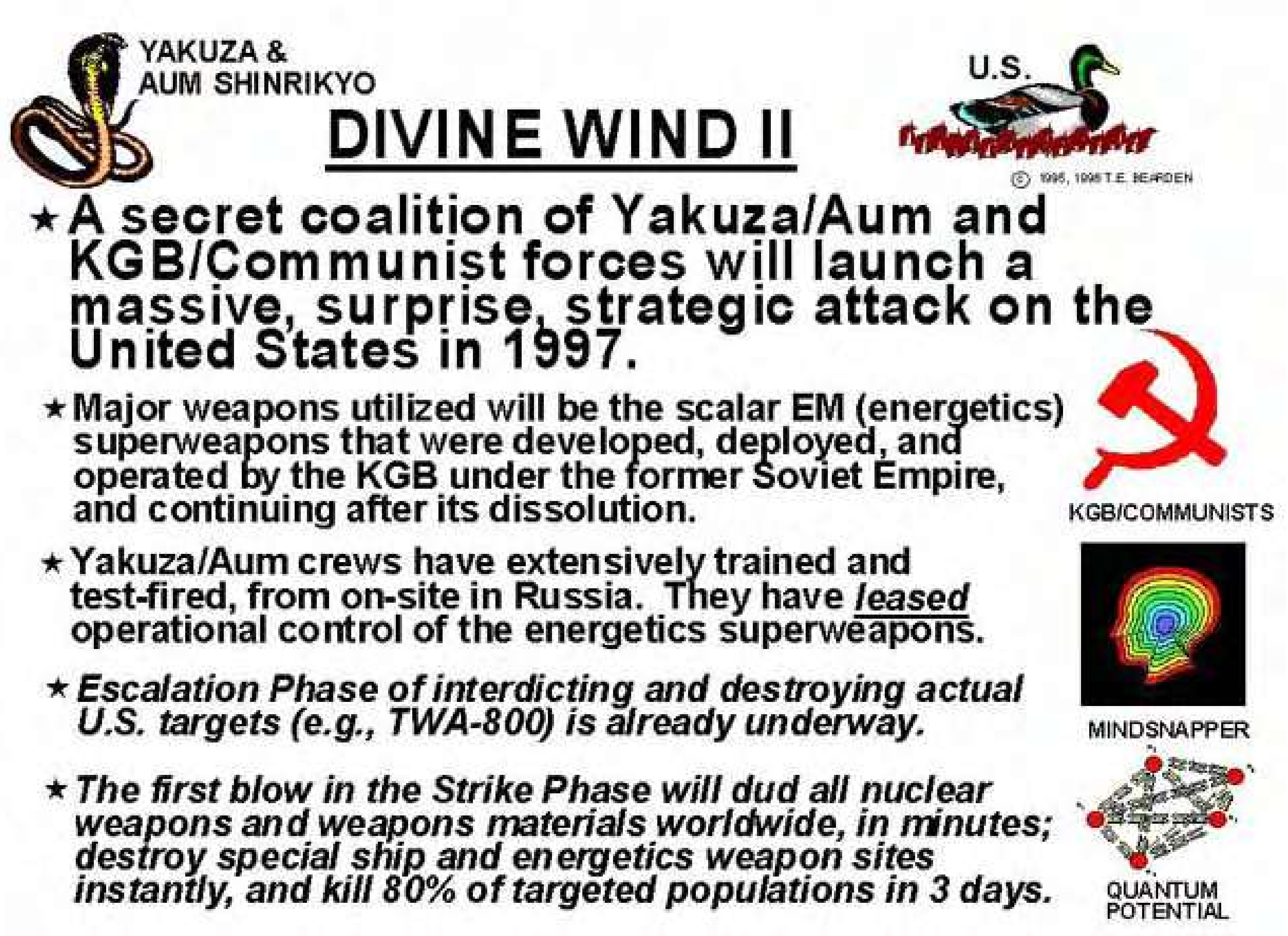
YAKUZA &
AUM SHINRIKYO
U.S.
KGB/COMMUNISTS
MINDSNAPPER
QUANTUM
POTENTIAL

DIVINE WIND II

© 1995, 1996 T.E. BEARDEN

- **A secret coalition of Yakuza/Aum and KGB/Communist forces will launch a massive, surprise, strategic attack on the United States in 1997.**
  - Major weapons utilized will be the scalar EM (energetics) superweapons that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.
  - Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the energetics superweapons.
  - *Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800) is already underway.*
  - *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, and kill 80% of targeted populations in 3 days.*

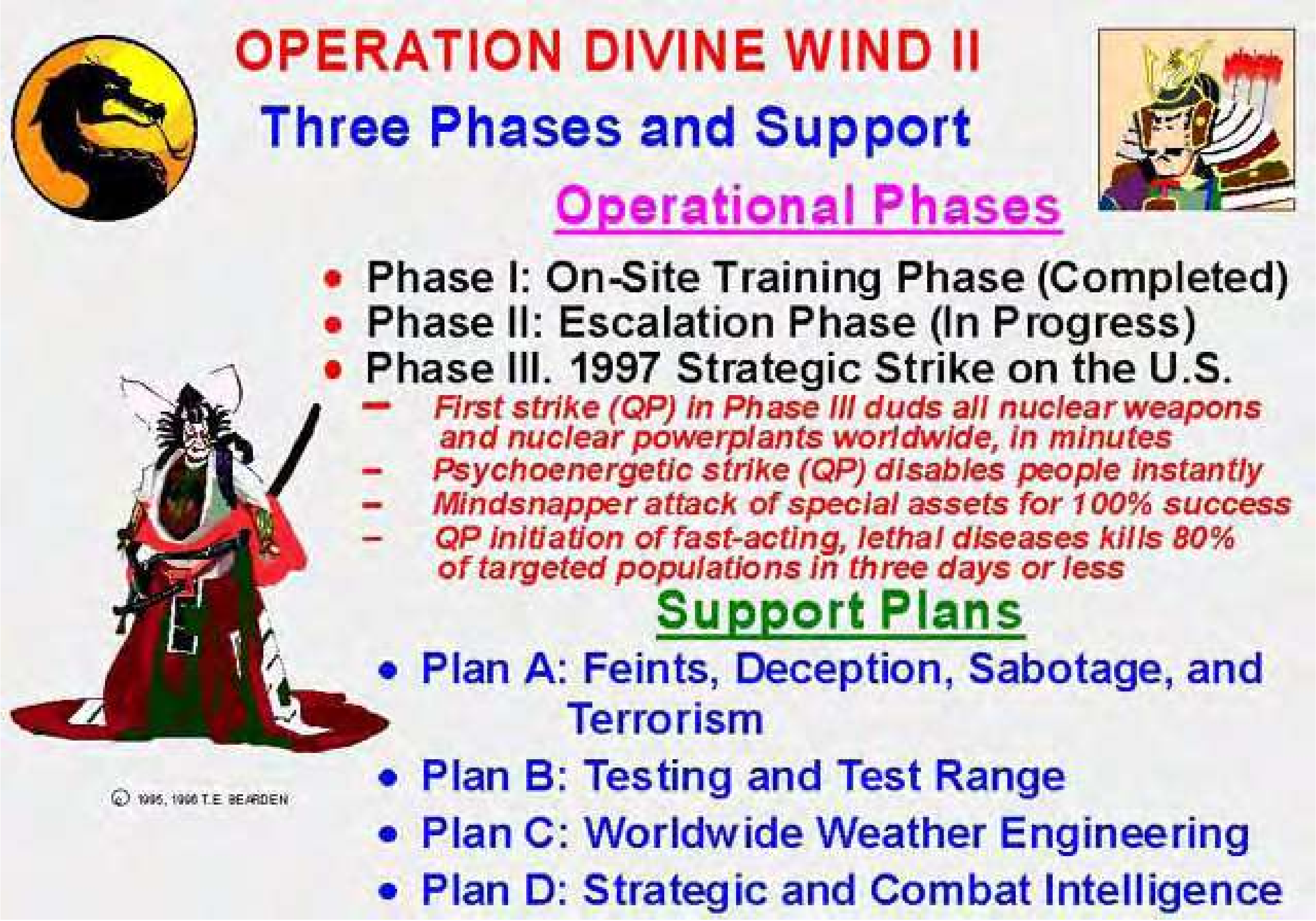

OPERATION DIVINE WIND II

Three Phases and Support

Operational Phases

- Phase I: On-Site Training Phase (Completed)
- Phase II: Escalation Phase (In Progress)
- Phase III. 1997 Strategic Strike on the U.S.
  - *First strike (QP) in Phase III duds all nuclear weapons and nuclear powerplants worldwide, in minutes*
  - *Psychoenergetic strike (QP) disables people instantly*
  - *Mindsnapper attack of special assets for 100% success*
  - *QP initiation of fast-acting, lethal diseases kills 80% of targeted populations in three days or less*

Support Plans

- Plan A: Feints, Deception, Sabotage, and Terrorism
- Plan B: Testing and Test Range
- Plan C: Worldwide Weather Engineering
- Plan D: Strategic and Combat Intelligence

© 1995, 1996 T.E. BEARDEN

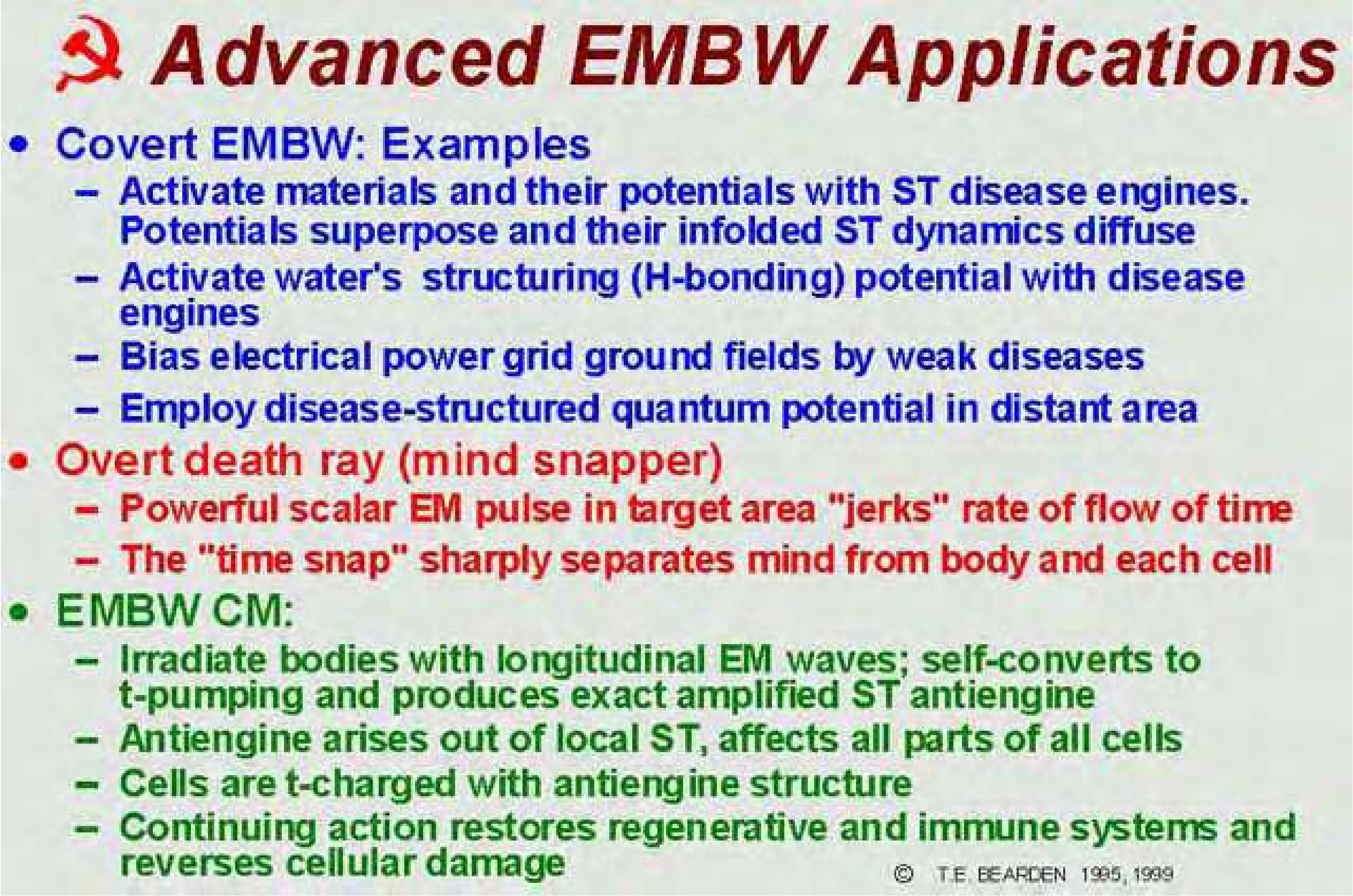

Advanced EMBW Applications

- Covert EMBW: Examples
  - Activate materials and their potentials with ST disease engines. Potentials superpose and their infolded ST dynamics diffuse
  - Activate water's structuring (H-bonding) potential with disease engines
  - Bias electrical power grid ground fields by weak diseases
  - Employ disease-structured quantum potential in distant area
- Overt death ray (mind snapper)
  - Powerful scalar EM pulse in target area "jerks" rate of flow of time
  - The "time snap" sharply separates mind from body and each cell
- EMBW CM:
  - Irradiate bodies with longitudinal EM waves; self-converts to t-pumping and produces exact amplified ST antiengine
  - Antiengine arises out of local ST, affects all parts of all cells
  - Cells are t-charged with antiengine structure
  - Continuing action restores regenerative and immune systems and reverses cellular damage

© T.E. BEARDEN 1995, 1999

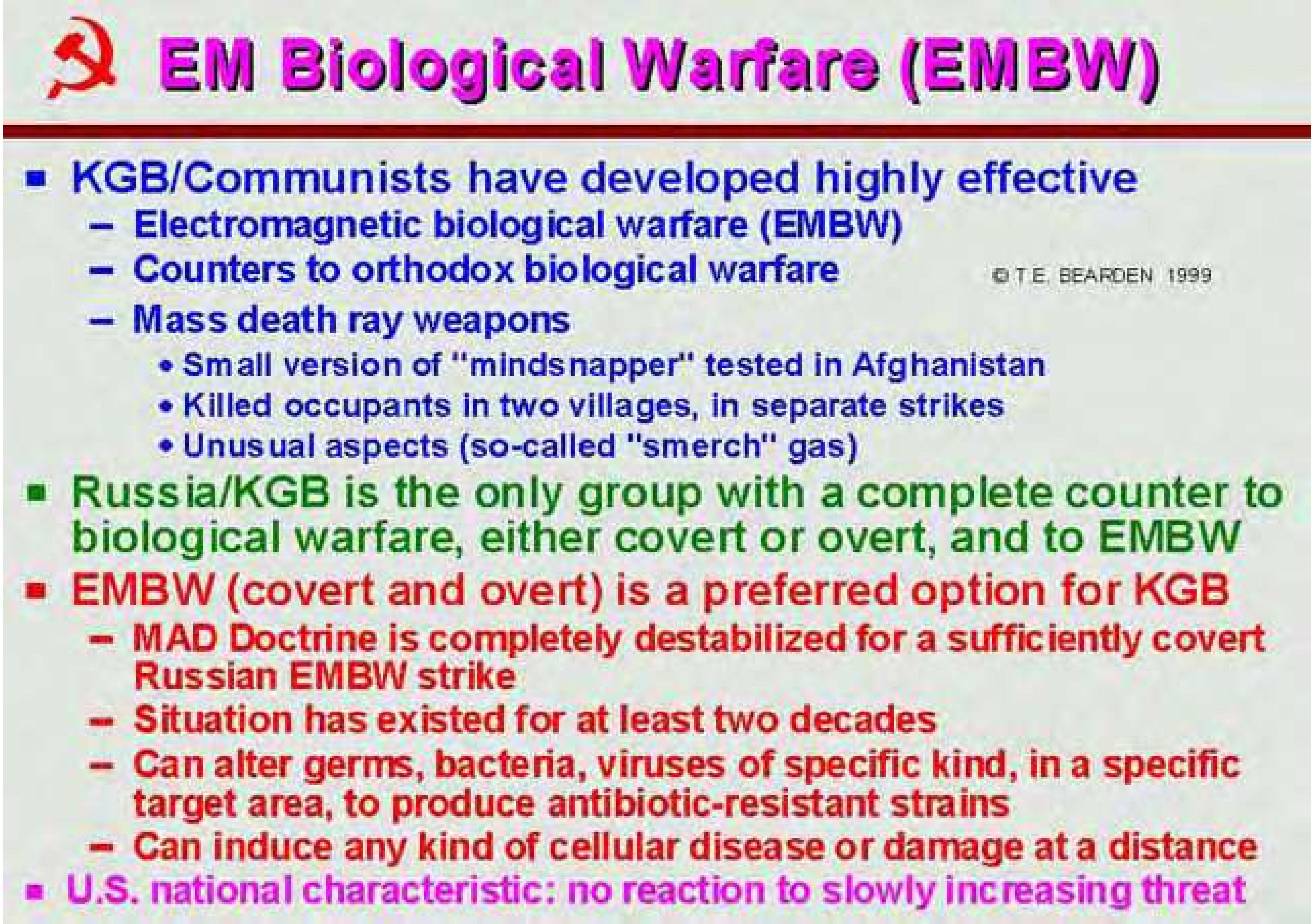

EM Biological Warfare (EMBW)

- KGB/Communists have developed highly effective
  - Electromagnetic biological warfare (EMBW)
  - Counters to orthodox biological warfare
  - Mass death ray weapons
    - Small version of "mindsnapper" tested in Afghanistan
    - Killed occupants in two villages, in separate strikes
    - Unusual aspects (so-called "smerch" gas)
- Russia/KGB is the only group with a complete counter to biological warfare, either covert or overt, and to EMBW
- EMBW (covert and overt) is a preferred option for KGB
  - MAD Doctrine is completely destabilized for a sufficiently covert Russian EMBW strike
  - Situation has existed for at least two decades
  - Can alter germs, bacteria, viruses of specific kind, in a specific target area, to produce antibiotic-resistant strains
  - Can induce any kind of cellular disease or damage at a distance
- U.S. national characteristic: no reaction to slowly increasing threat

© T.E. BEARDEN 1999

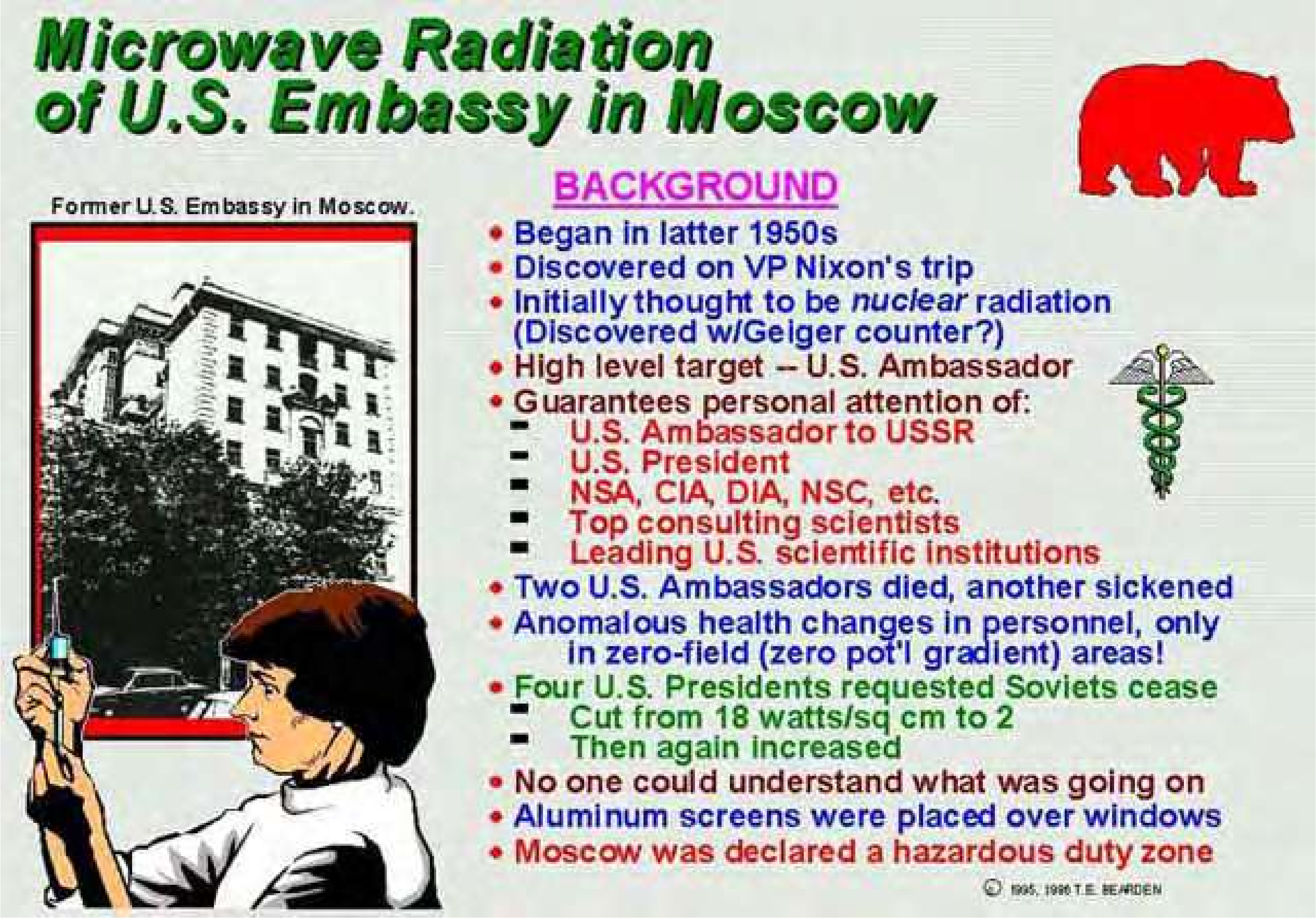

Microwave Radiation of U.S. Embassy in Moscow

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

© 1995, 1996 T.E. BEARDEN

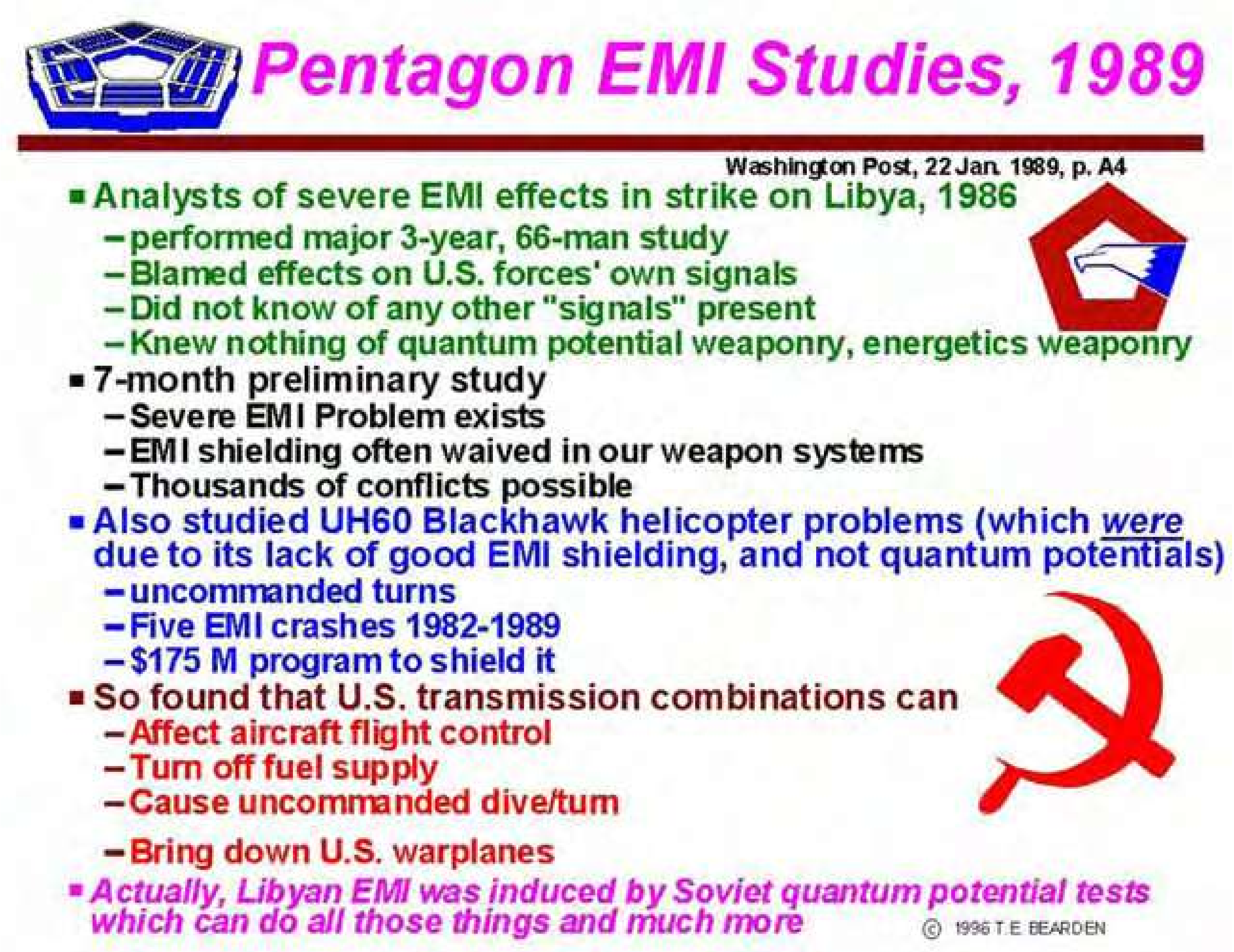

Pentagon EMI Studies, 1989

Washington Post, 22 Jan. 1989, p. A4

- Analysts of severe EMI effects in strike on Libya, 1986
  - performed major 3-year, 66-man study
  - Blamed effects on U.S. forces' own signals
  - Did not know of any other "signals" present
  - Knew nothing of quantum potential weaponry, energetics weaponry
- 7-month preliminary study
  - Severe EMI Problem exists
  - EMI shielding often waived in our weapon systems
  - Thousands of conflicts possible
- Also studied UH60 Blackhawk helicopter problems (which *were* due to its lack of good EMI shielding, and not quantum potentials)
  - uncommanded turns
  - Five EMI crashes 1982-1989
  - $175 M program to shield it
- So found that U.S. transmission combinations can
  - Affect aircraft flight control
  - Turn off fuel supply
  - Cause uncommanded dive/turn
  - Bring down U.S. warplanes
- *Actually, Libyan EMI was induced by Soviet quantum potential tests which can do all those things and much more*

© 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

EM Missile Practice

- EM Missile Practice
  Latter Sept. - Early Nov. 1996
  Melbourne, Australia

- Deception: Using EM Missile to Reinforce Notion of SAM Attack
  Saturday evening, Nov. 16, 1996

- Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry
  Thursday, December 12th, 1996

- EM Missile Strike, Offset from Night Shuttle Launch, Cape Canaveral, Florida
  The Smoking Gun in November 1985

- Distant Operator Onsite in Russia Slew Away Registration Point
  Marker Beacon over Night Shuttle Launch, 26 November, 1985

- EM Missile Practice and Deception
  November 17, 1996

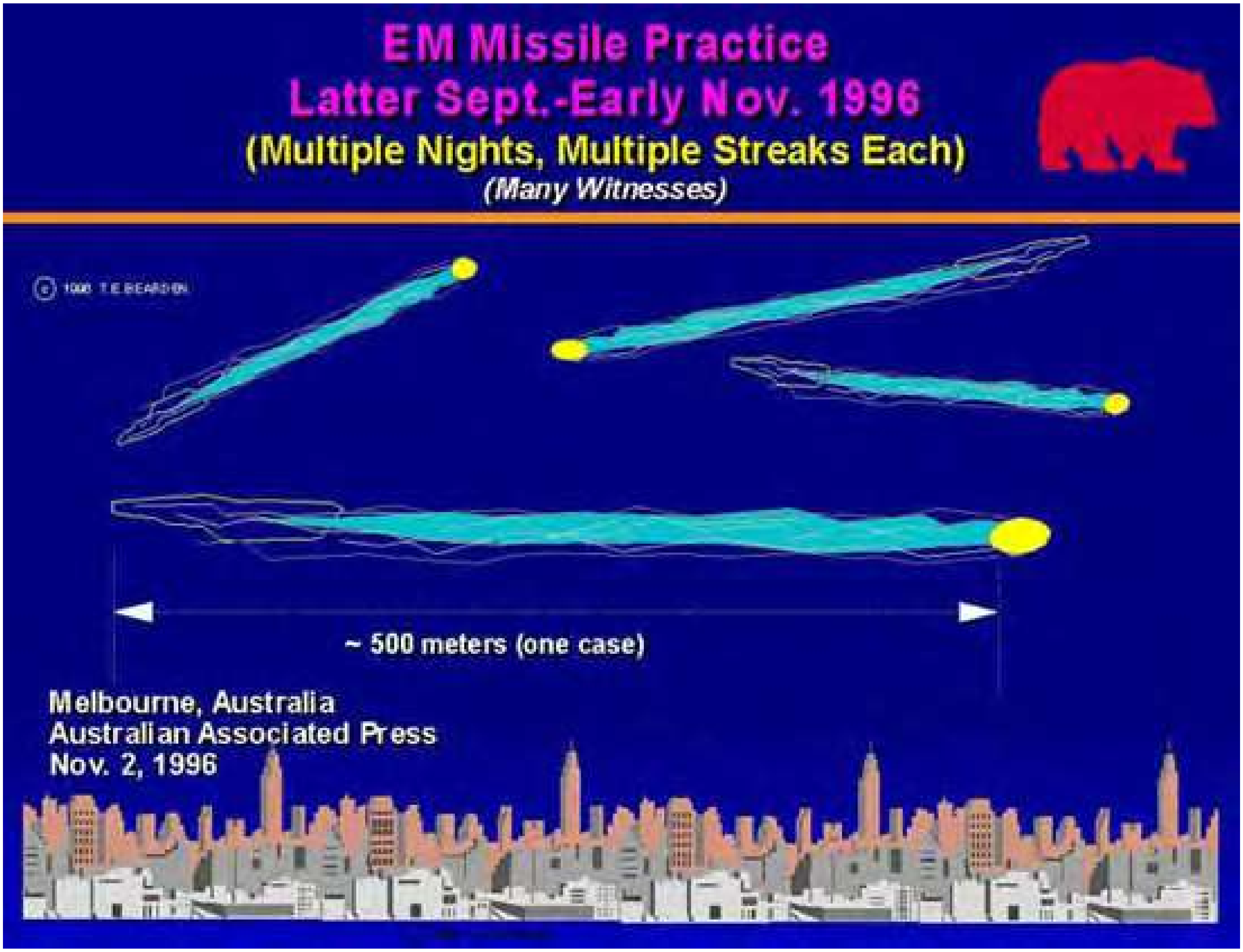

EM Missile Practice
Latter Sept.-Early Nov. 1996
(Multiple Nights, Multiple Streaks Each)
(Many Witnesses)

© 1996 T.E.BEARDEN

~ 500 meters (one case)

Melbourne, Australia
Australian Associated Press
Nov. 2, 1996

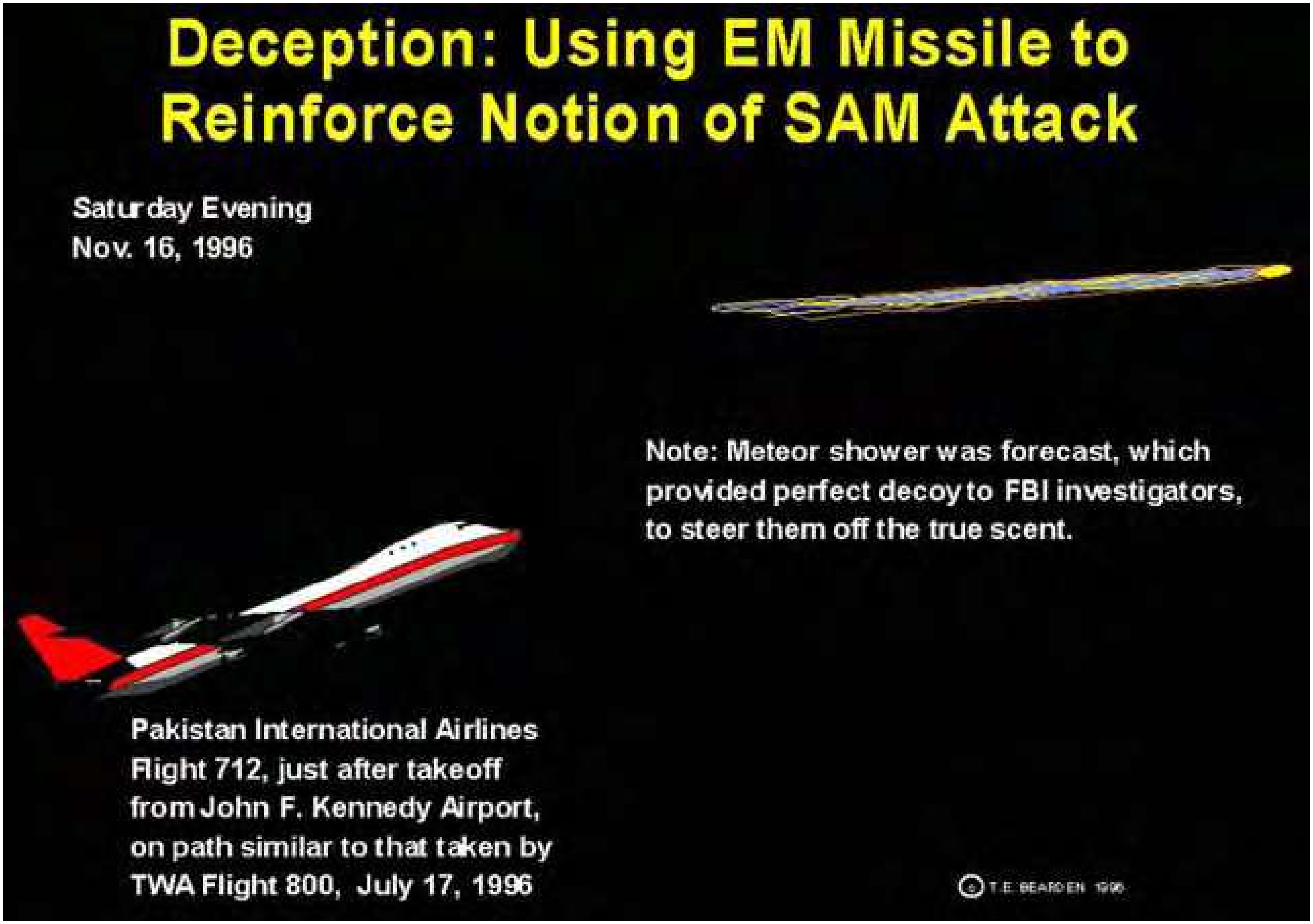

Deception: Using EM Missile to Reinforce Notion of SAM Attack

Saturday Evening
Nov. 16, 1996

Note: Meteor shower was forecast, which provided perfect decoy to FBI investigators, to steer them off the true scent.

Pakistan International Airlines Flight 712, just after takeoff from John F. Kennedy Airport, on path similar to that taken by TWA Flight 800, July 17, 1996

© T.E. BEARDEN 1996

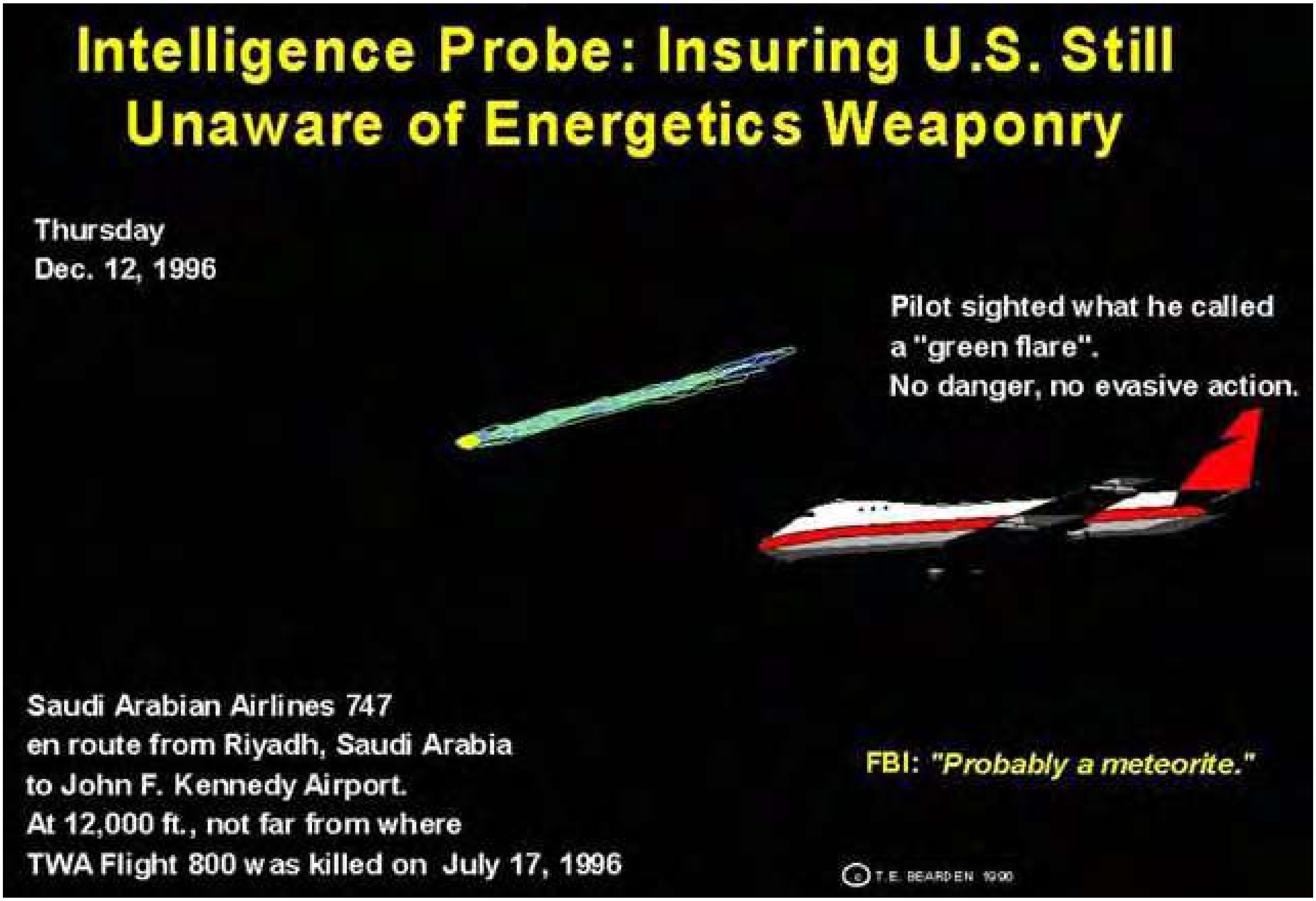

Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry

Thursday
Dec. 12, 1996

Pilot sighted what he called
a "green flare".
No danger, no evasive action.

Saudi Arabian Airlines 747
en route from Riyadh, Saudi Arabia
to John F. Kennedy Airport.
At 12,000 ft., not far from where
TWA Flight 800 was killed on July 17, 1996

FBI: *"Probably a meteorite."*

© T.E. BEARDEN 1990

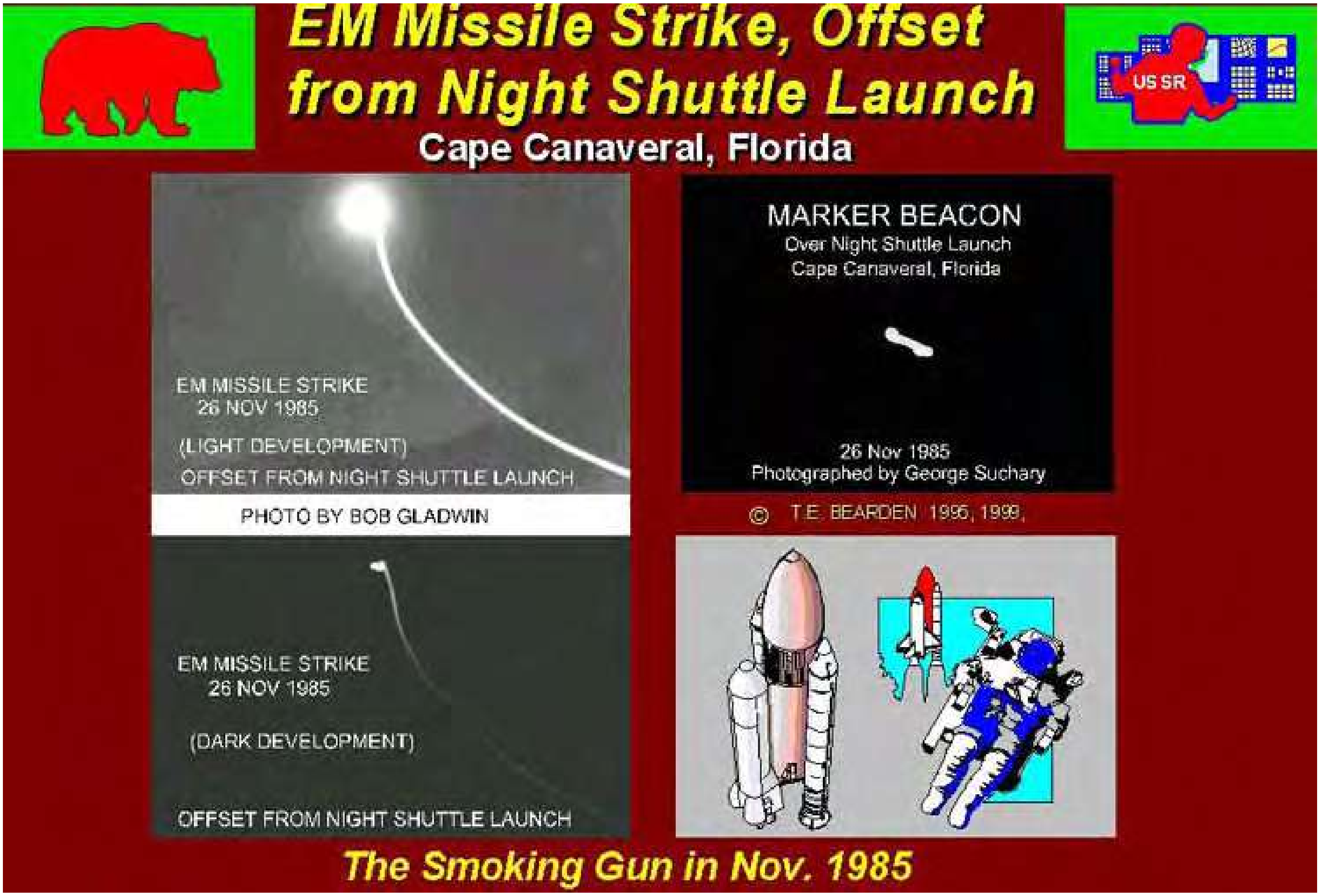

EM Missile Strike, Offset from Night Shuttle Launch

Cape Canaveral, Florida

EM MISSILE STRIKE
26 NOV 1985

(LIGHT DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

PHOTO BY BOB GLADWIN

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

© T.E. BEARDEN 1996, 1999,

EM MISSILE STRIKE
26 NOV 1985

(DARK DEVELOPMENT)

OFFSET FROM NIGHT SHUTTLE LAUNCH

The Smoking Gun in Nov. 1985

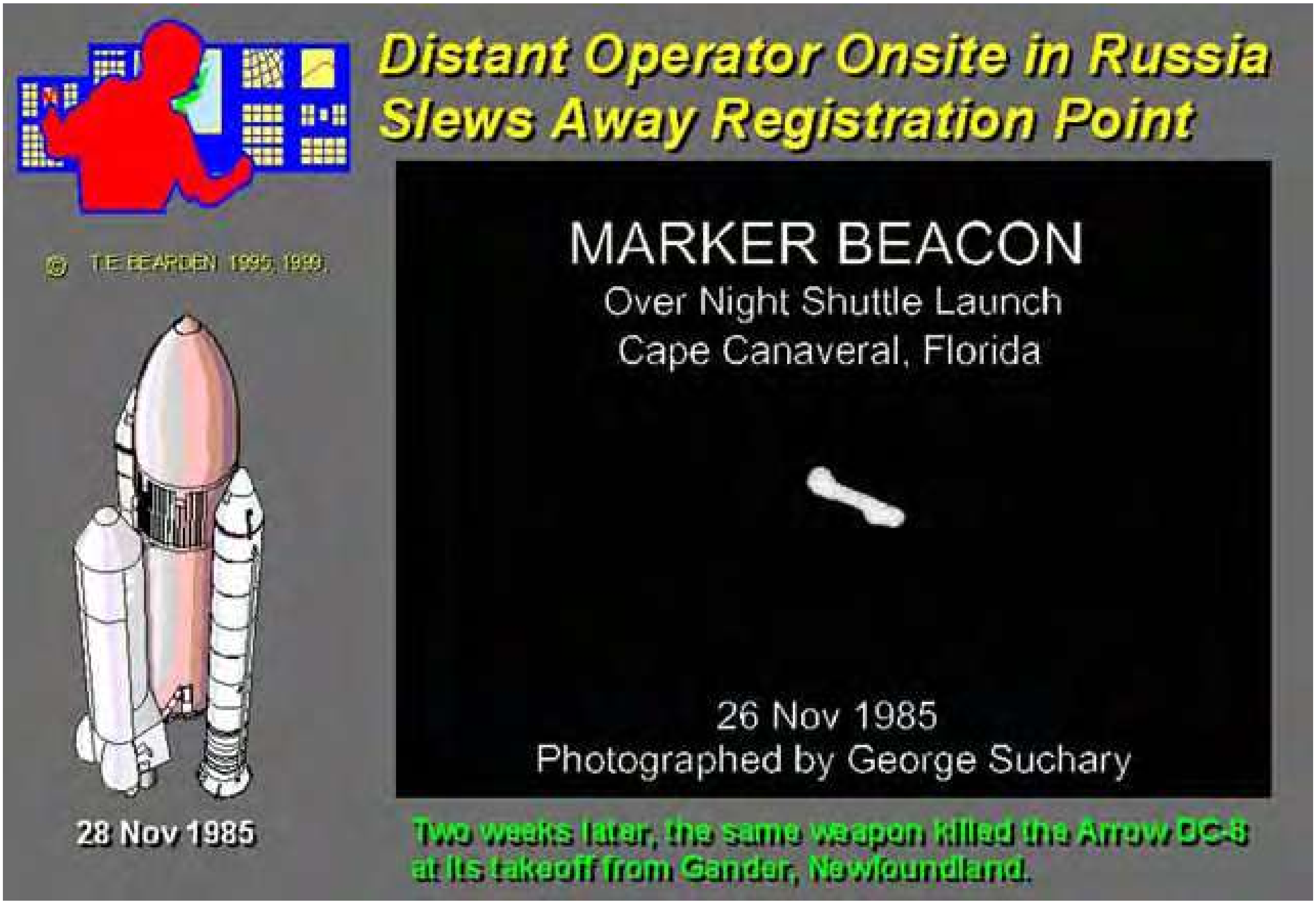

Distant Operator Onsite in Russia Slews Away Registration Point

© T.E. BEARDEN 1995, 1999.

MARKER BEACON

Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

28 Nov 1985

Two weeks later, the same weapon killed the Arrow DC-8 at its takeoff from Gander, Newfoundland.

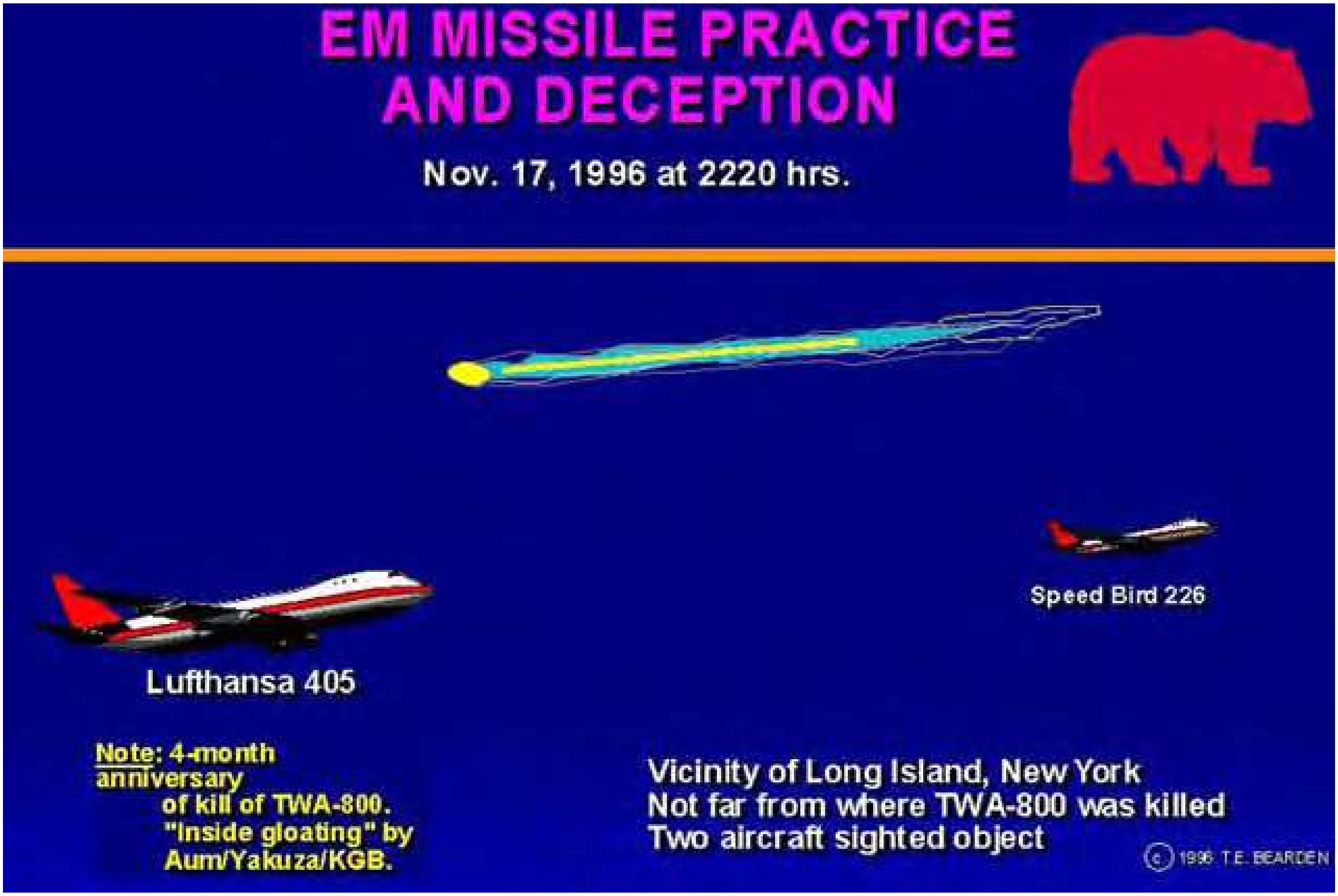

EM MISSILE PRACTICE AND DECEPTION

Nov. 17, 1996 at 2220 hrs.

Speed Bird 226

Lufthansa 405

Note: 4-month anniversary
of kill of TWA-800.
"Inside gloating" by
Aum/Yakuza/KGB.

Vicinity of Long Island, New York
Not far from where TWA-800 was killed
Two aircraft sighted object

(c) 1996 T.E. BEARDEN

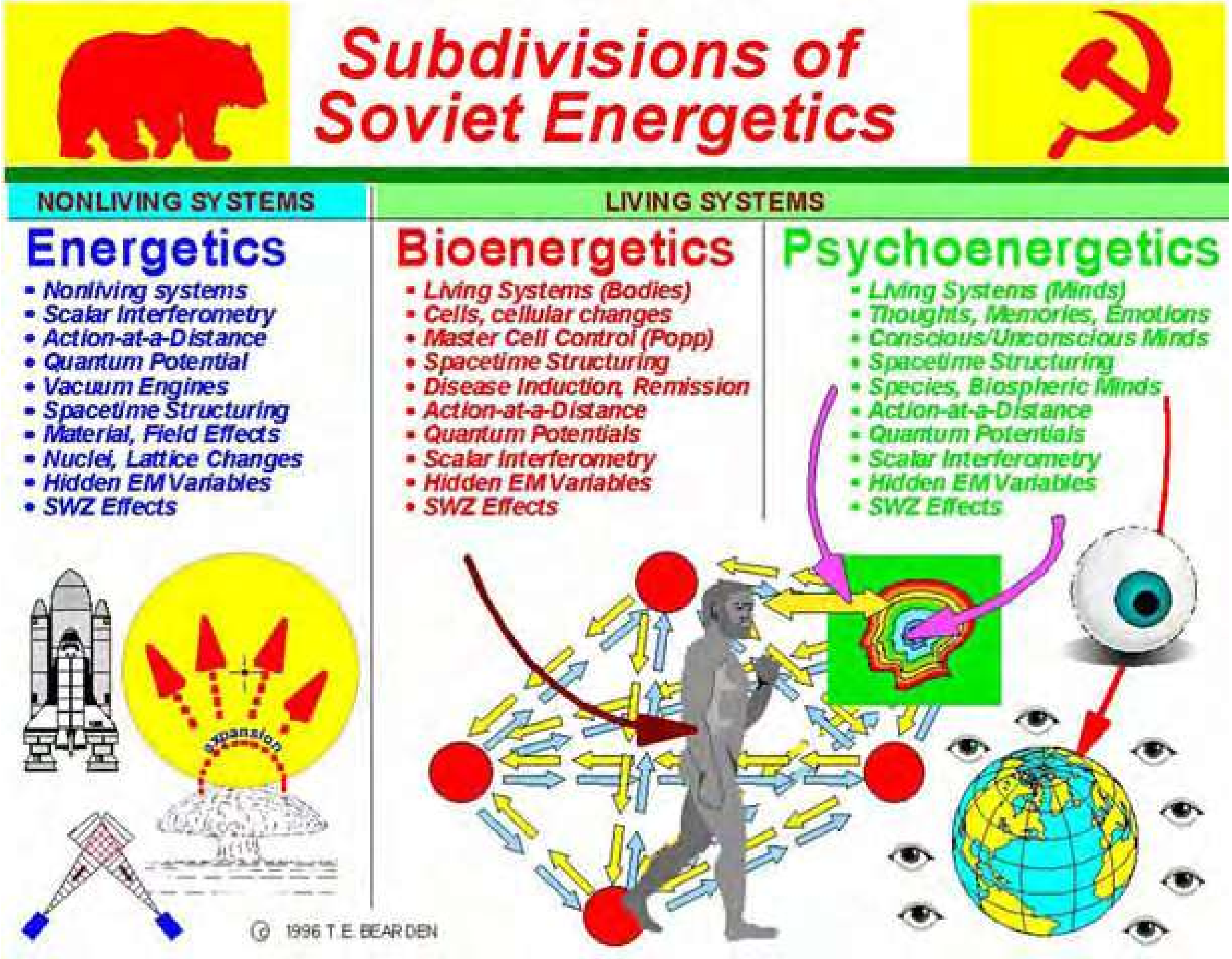

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

LIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T.E. BEARDEN

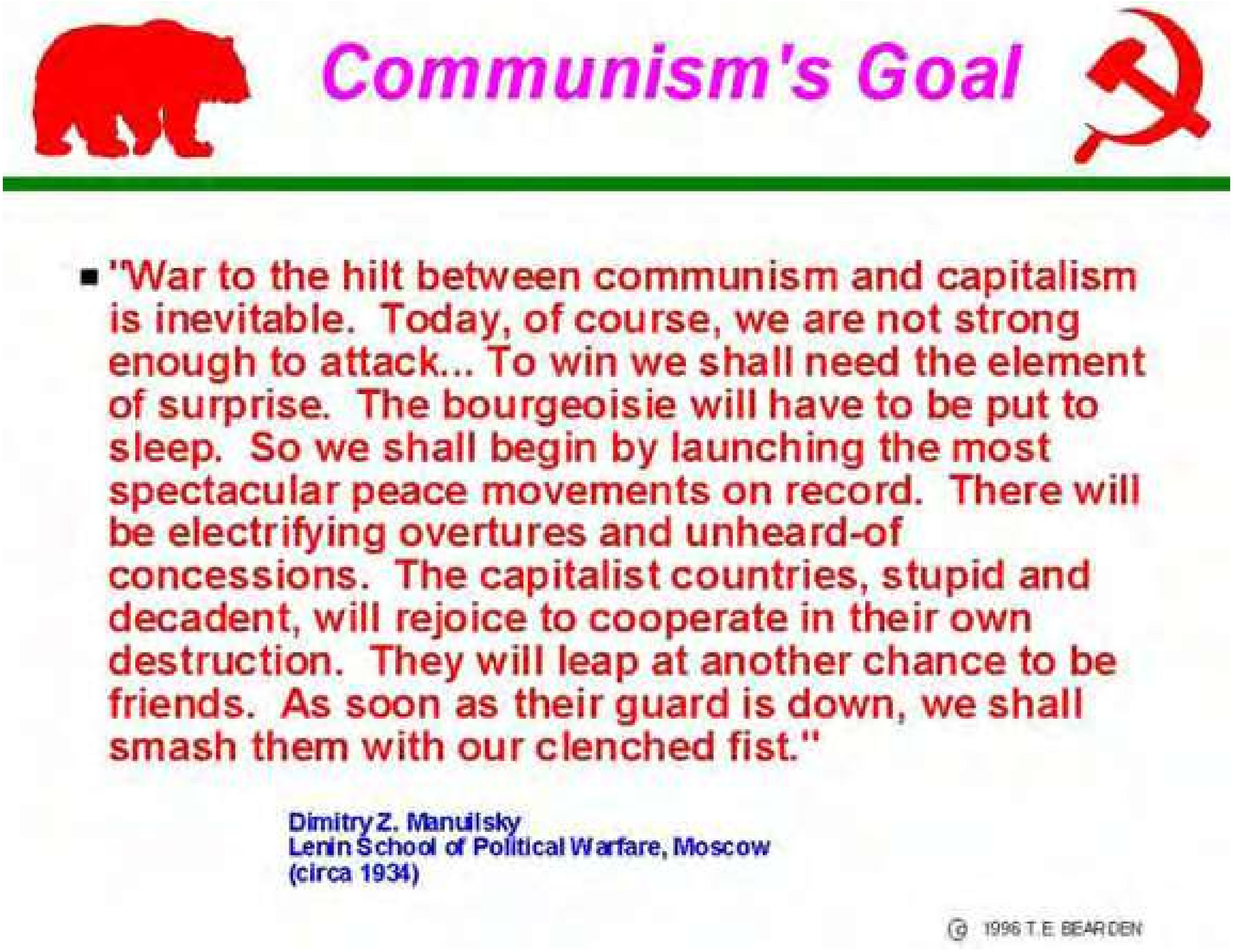

Communism's Goal

- "War to the hilt between communism and capitalism is inevitable. Today, of course, we are not strong enough to attack... To win we shall need the element of surprise. The bourgeoisie will have to be put to sleep. So we shall begin by launching the most spectacular peace movements on record. There will be electrifying overtures and unheard-of concessions. The capitalist countries, stupid and decadent, will rejoice to cooperate in their own destruction. They will leap at another chance to be friends. As soon as their guard is down, we shall smash them with our clenched fist."

Dimitry Z. Manuilsky
Lenin School of Political Warfare, Moscow
(circa 1934)

© 1996 T.E. BEARDEN

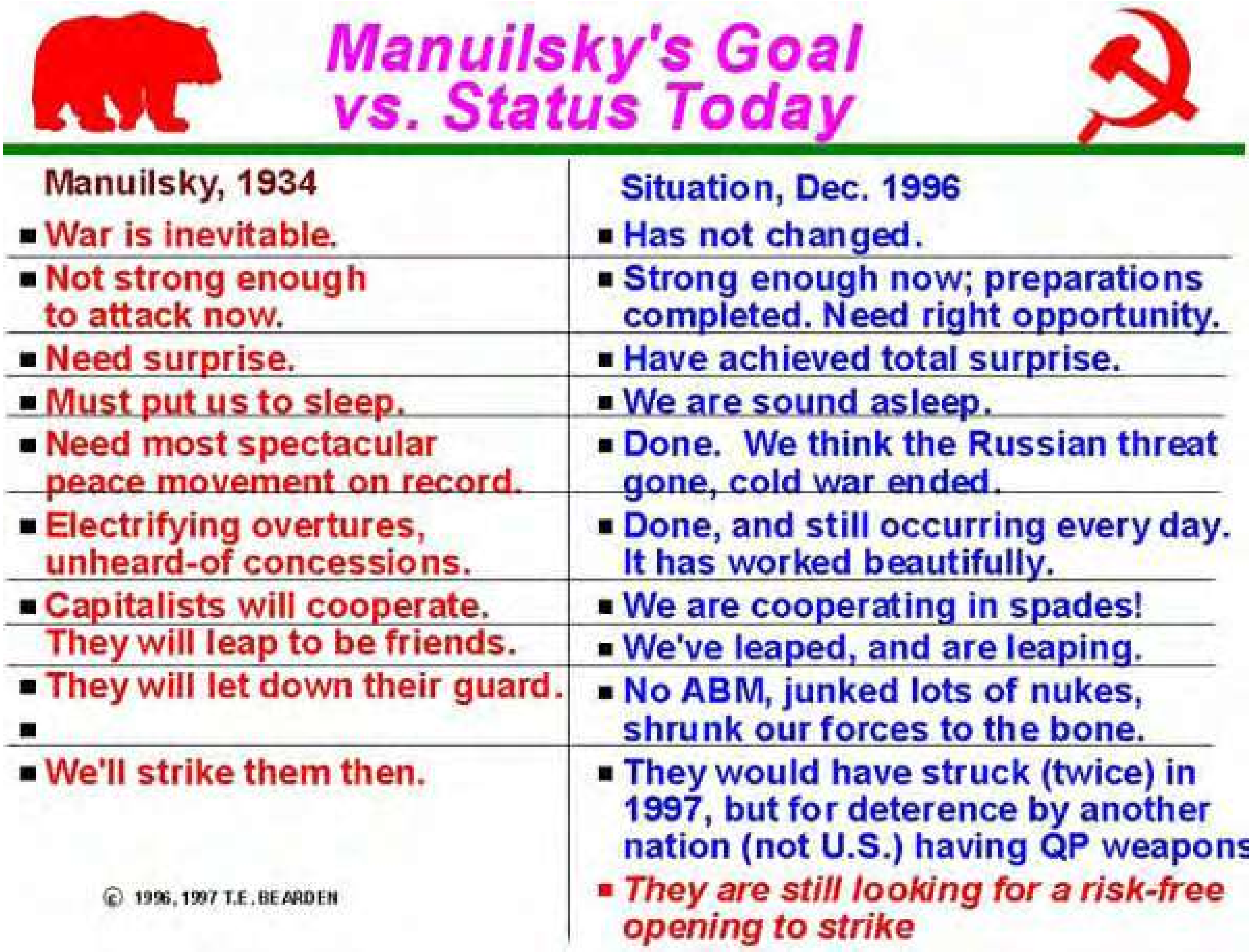

Manuilsky's Goal vs. Status Today

Manuilsky, 1934	Situation, Dec. 1996
▪ War is inevitable.	▪ Has not changed.
▪ Not strong enough to attack now.	▪ Strong enough now; preparations completed. Need right opportunity.
▪ Need surprise.	▪ Have achieved total surprise.
▪ Must put us to sleep.	▪ We are sound asleep.
▪ Need most spectacular peace movement on record.	▪ Done. We think the Russian threat gone, cold war ended.
▪ Electrifying overtures, unheard-of concessions.	▪ Done, and still occurring every day. It has worked beautifully.
▪ Capitalists will cooperate.	▪ We are cooperating in spades!
They will leap to be friends.	▪ We've leaped, and are leaping.
▪ They will let down their guard. ▪	▪ No ABM, junked lots of nukes, shrunk our forces to the bone.
▪ We'll strike them then. © 1996, 1997 T.E. BEARDEN	▪ They would have struck (twice) in 1997, but for deterence by another nation (not U.S.) having QP weapons ▪ *They are still looking for a risk-free opening to strike*

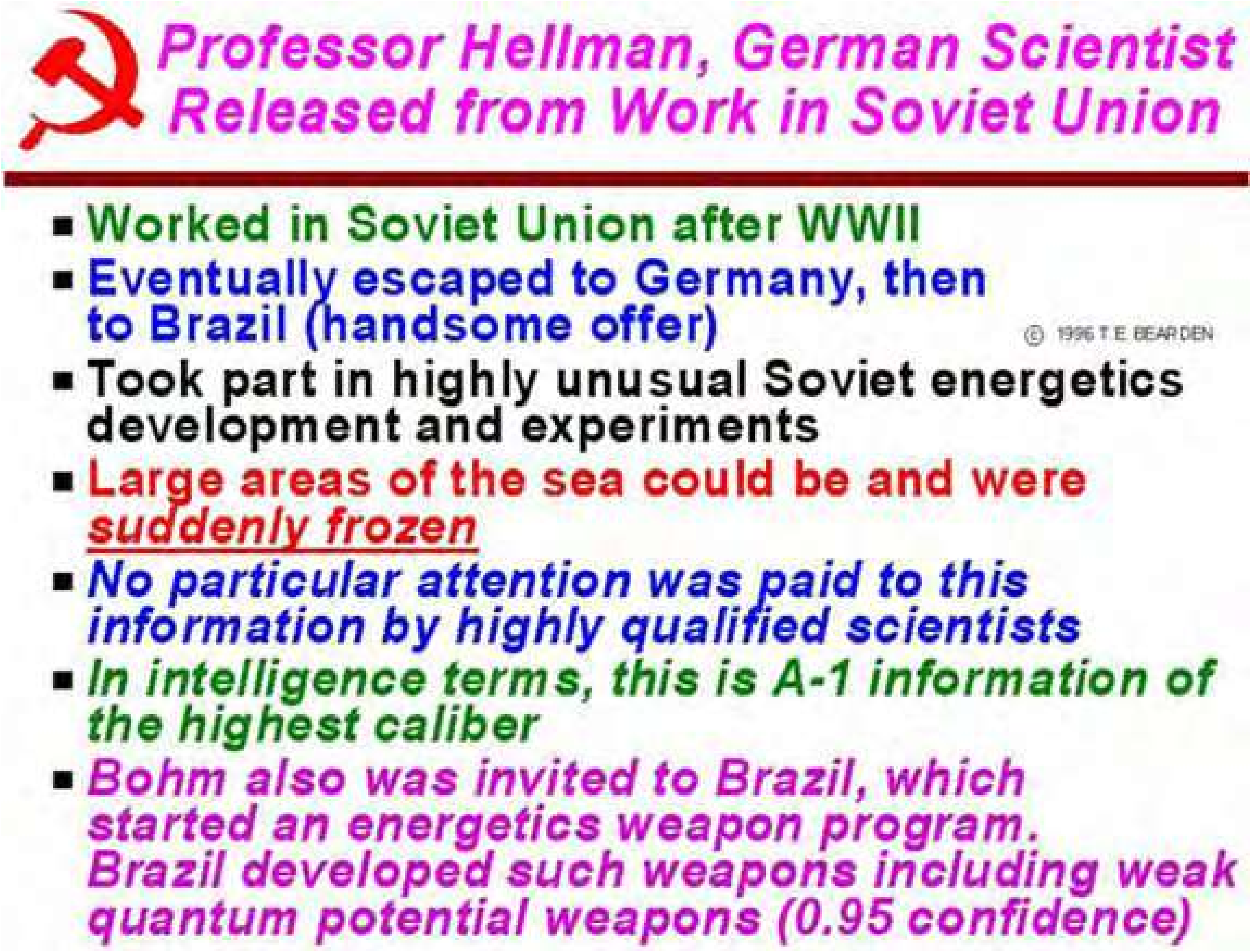

Professor Hellman, German Scientist Released from Work in Soviet Union

- Worked in Soviet Union after WWII
- Eventually escaped to Germany, then to Brazil (handsome offer)
- Took part in highly unusual Soviet energetics development and experiments
- Large areas of the sea could be and were *suddenly frozen*
- *No particular attention was paid to this information by highly qualified scientists*
- *In intelligence terms, this is A-1 information of the highest caliber*
- *Bohm also was invited to Brazil, which started an energetics weapon program. Brazil developed such weapons including weak quantum potential weapons (0.95 confidence)*

© 1996 T.E. BEARDEN

Hemispheres of Light

- March 20th, 1969
  Caribbean and Western North Atlantic
- Hemisphere and Globes, 1977
  24 March, 1977
- Virgin Islands Incident
  August 1969
- Typical Shield Test, Northern Route to Japan
  June, 1982

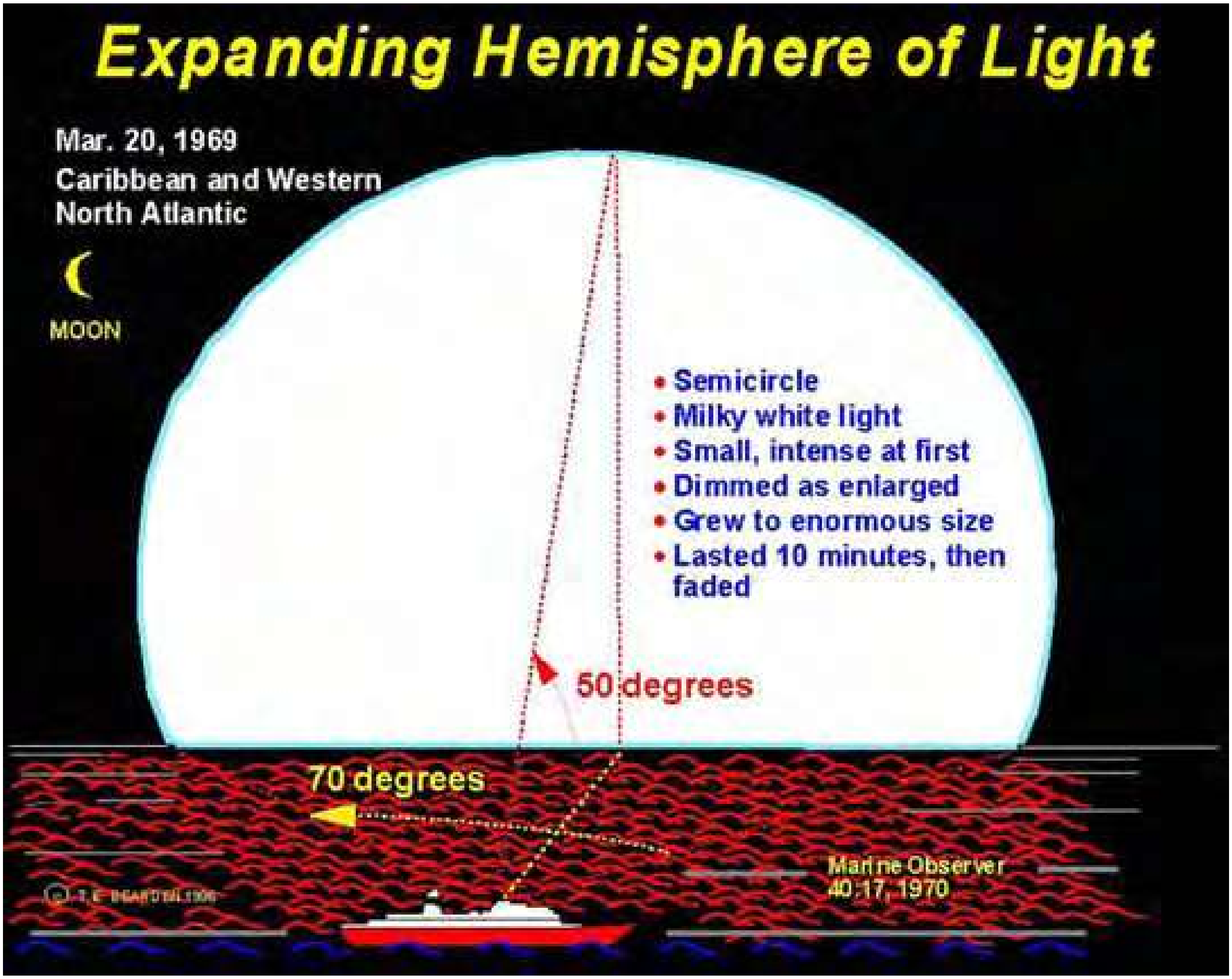

Expanding Hemisphere of Light

Mar. 20, 1969
Caribbean and Western
North Atlantic

- Semicircle
- Milky white light
- Small, intense at first
- Dimmed as enlarged
- Grew to enormous size
- Lasted 10 minutes, then faded

Marine Observer
40:17, 1970

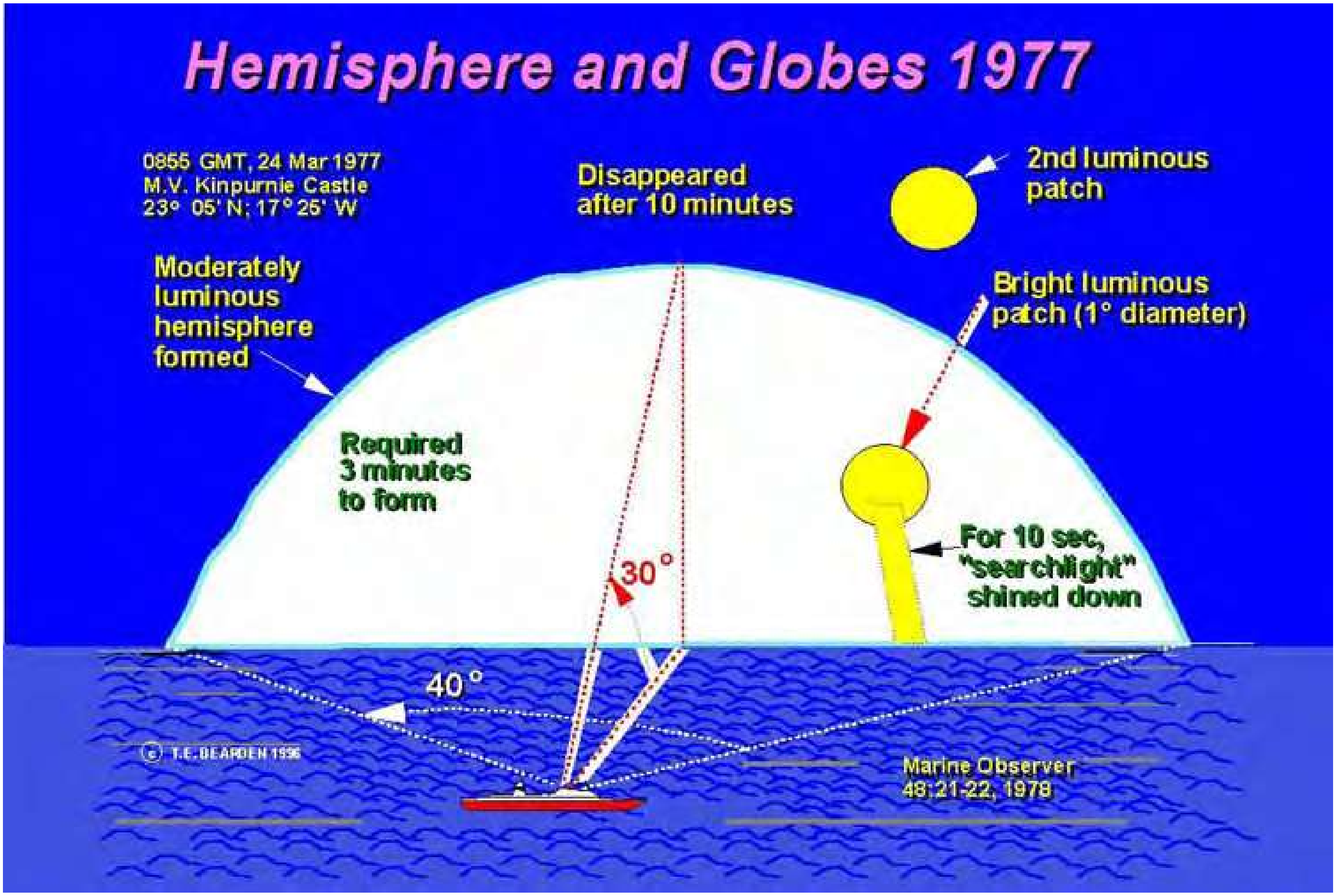
Hemisphere and Globes 1977
0855 GMT, 24 Mar 1977
M.V. Kinpurnie Castle
23° 05' N; 17° 25' W
Disappeared after 10 minutes
2nd luminous patch
Moderately luminous hemisphere formed
Bright luminous patch (1° diameter)
Required 3 minutes to form
For 10 sec, "searchlight" shined down
30°
40°
© T.E. BEARDEN 1996
Marine Observer 48:21-22, 1978

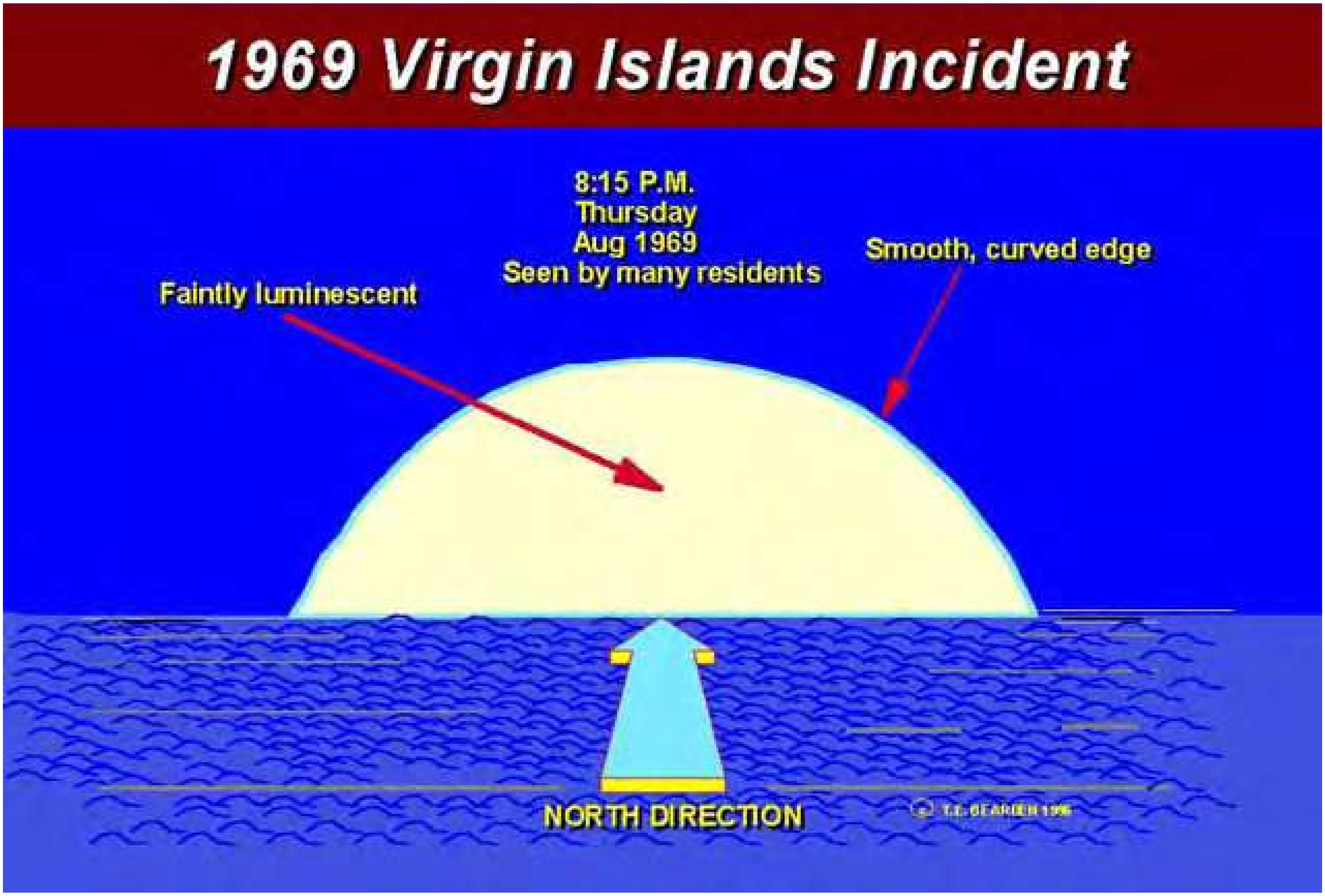

1969 Virgin Islands Incident

8:15 P.M.
Thursday
Aug 1969
Seen by many residents

Faintly luminescent

Smooth, curved edge

NORTH DIRECTION

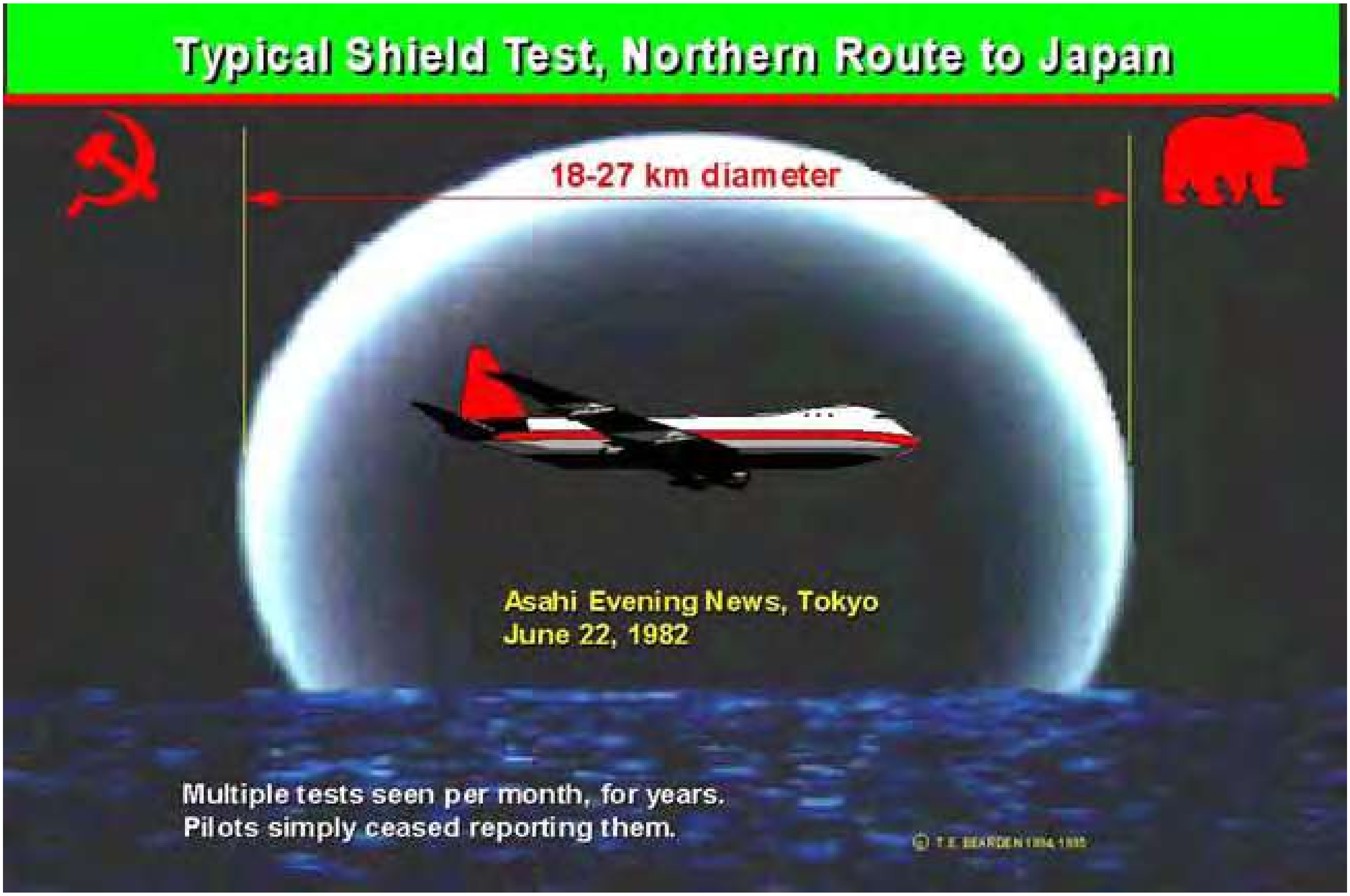

Typical Shield Test, Northern Route to Japan

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T.E. BEARDEN 1994, 1995

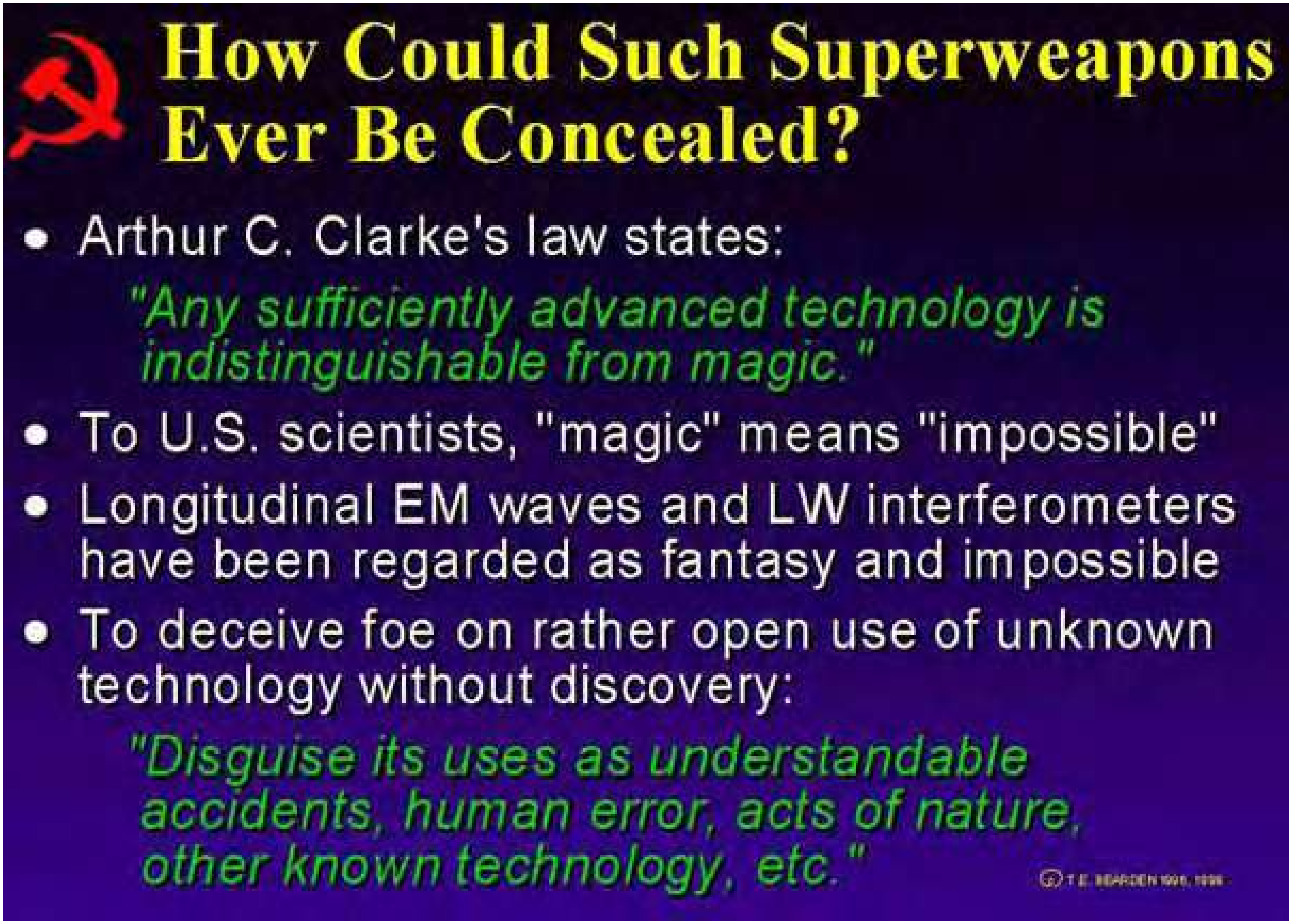

How Could Such Superweapons Ever Be Concealed?

- Arthur C. Clarke's law states:

  "Any sufficiently advanced technology is indistinguishable from magic."

- To U.S. scientists, "magic" means "impossible"
- Longitudinal EM waves and LW interferometers have been regarded as fantasy and impossible
- To deceive foe on rather open use of unknown technology without discovery:

  "Disguise its uses as understandable accidents, human error, acts of nature, other known technology, etc."

© T.E. BEARDEN [illegible]

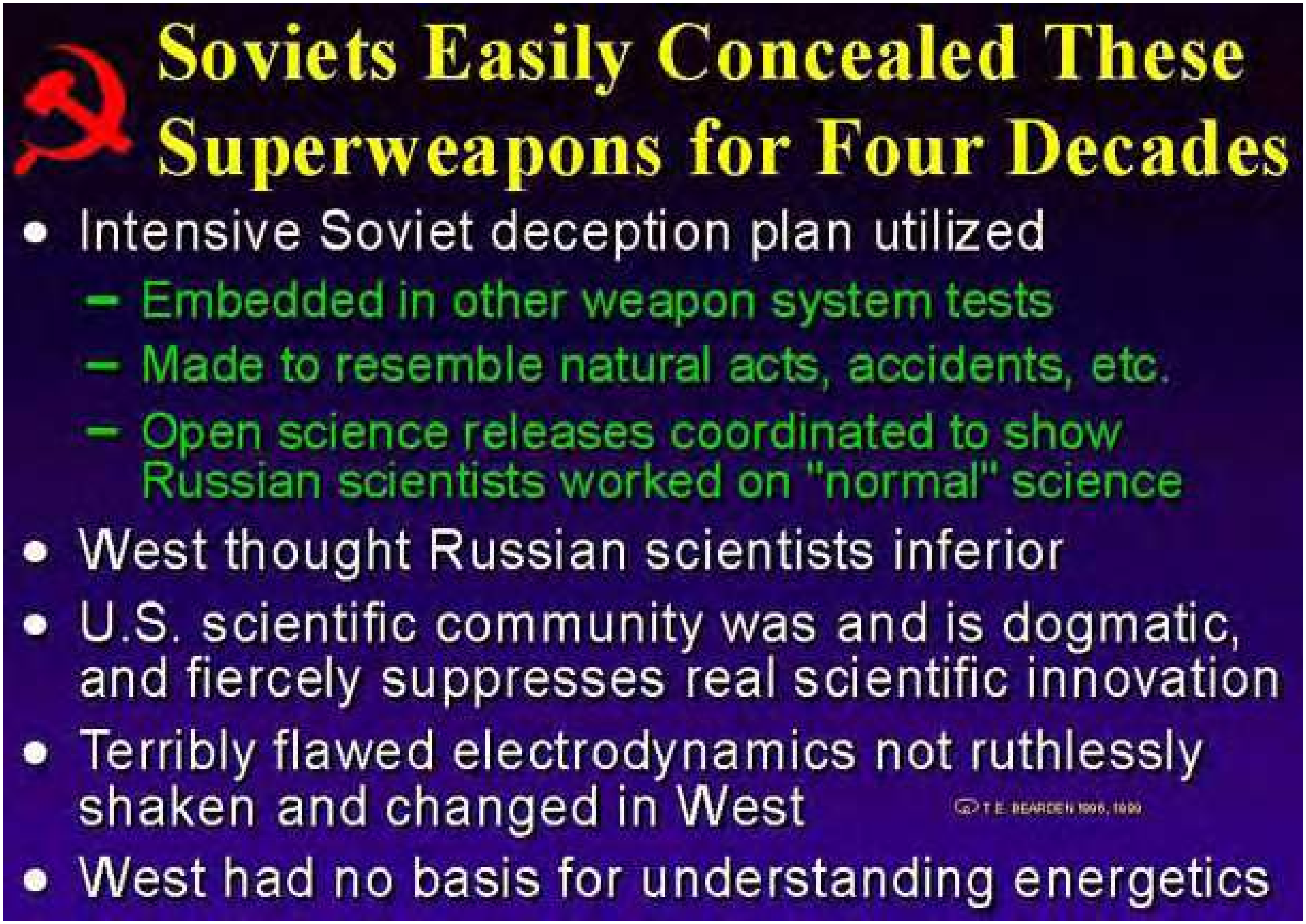

Soviets Easily Concealed These Superweapons for Four Decades

- Intensive Soviet deception plan utilized
  - Embedded in other weapon system tests
  - Made to resemble natural acts, accidents, etc.
  - Open science releases coordinated to show Russian scientists worked on "normal" science
- West thought Russian scientists inferior
- U.S. scientific community was and is dogmatic, and fiercely suppresses real scientific innovation
- Terribly flawed electrodynamics not ruthlessly shaken and changed in West
- West had no basis for understanding energetics

© T.E. BEARDEN 1996, 1999

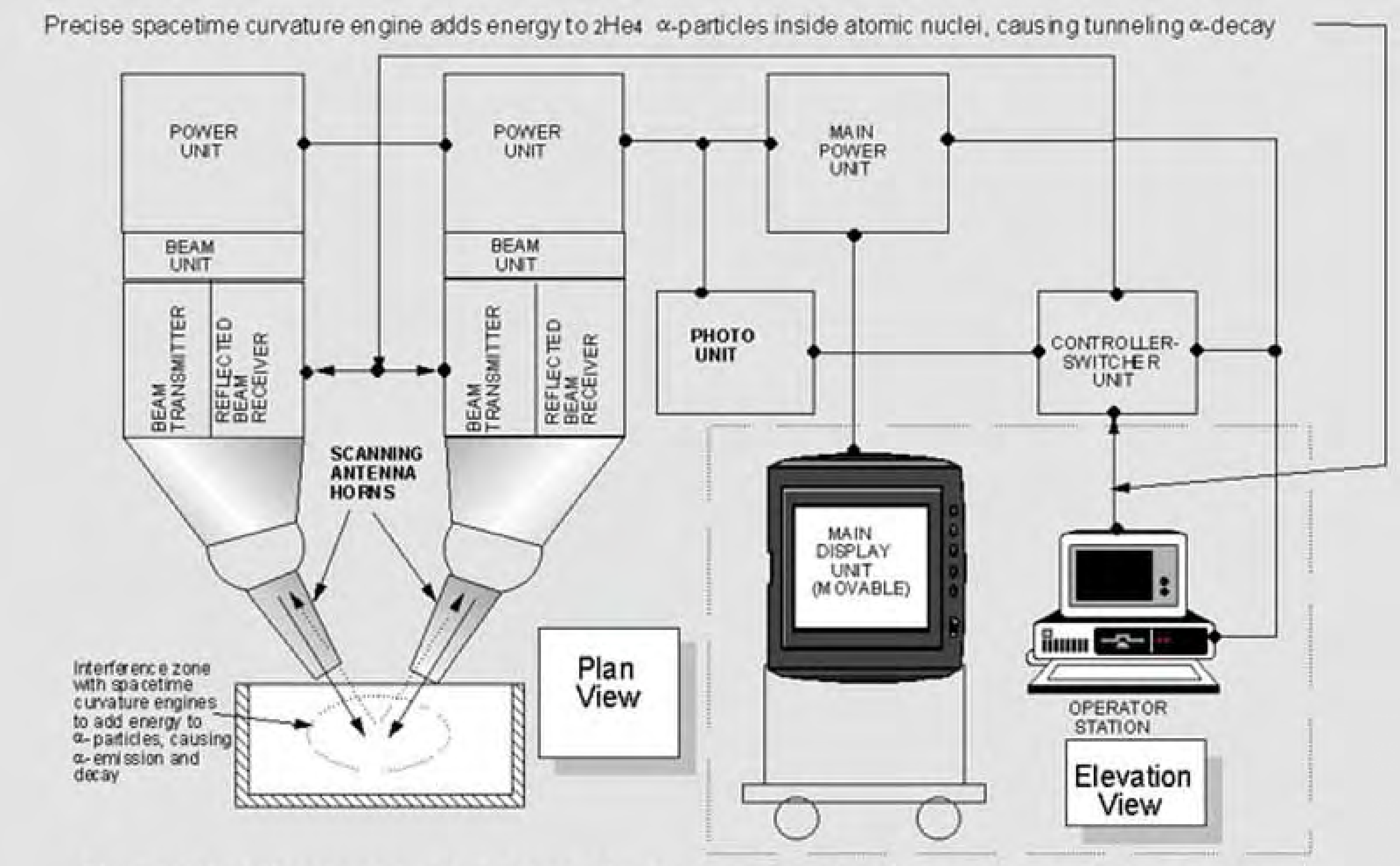
Precise spacetime curvature engine adds energy to 2He4 α-particles inside atomic nuclei, causing tunneling α-decay
POWER UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
POWER UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
MAIN POWER UNIT
PHOTO UNIT
CONTROLLER-SWITCHER UNIT
SCANNING ANTENNA HORNS
MAIN DISPLAY UNIT (MOVABLE)
Interference zone with spacetime curvature engines to add energy to α-particles, causing α-emission and decay
Plan View
OPERATOR STATION
Elevation View

Figure 26. Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes.

© T.E. BEARDEN 1998

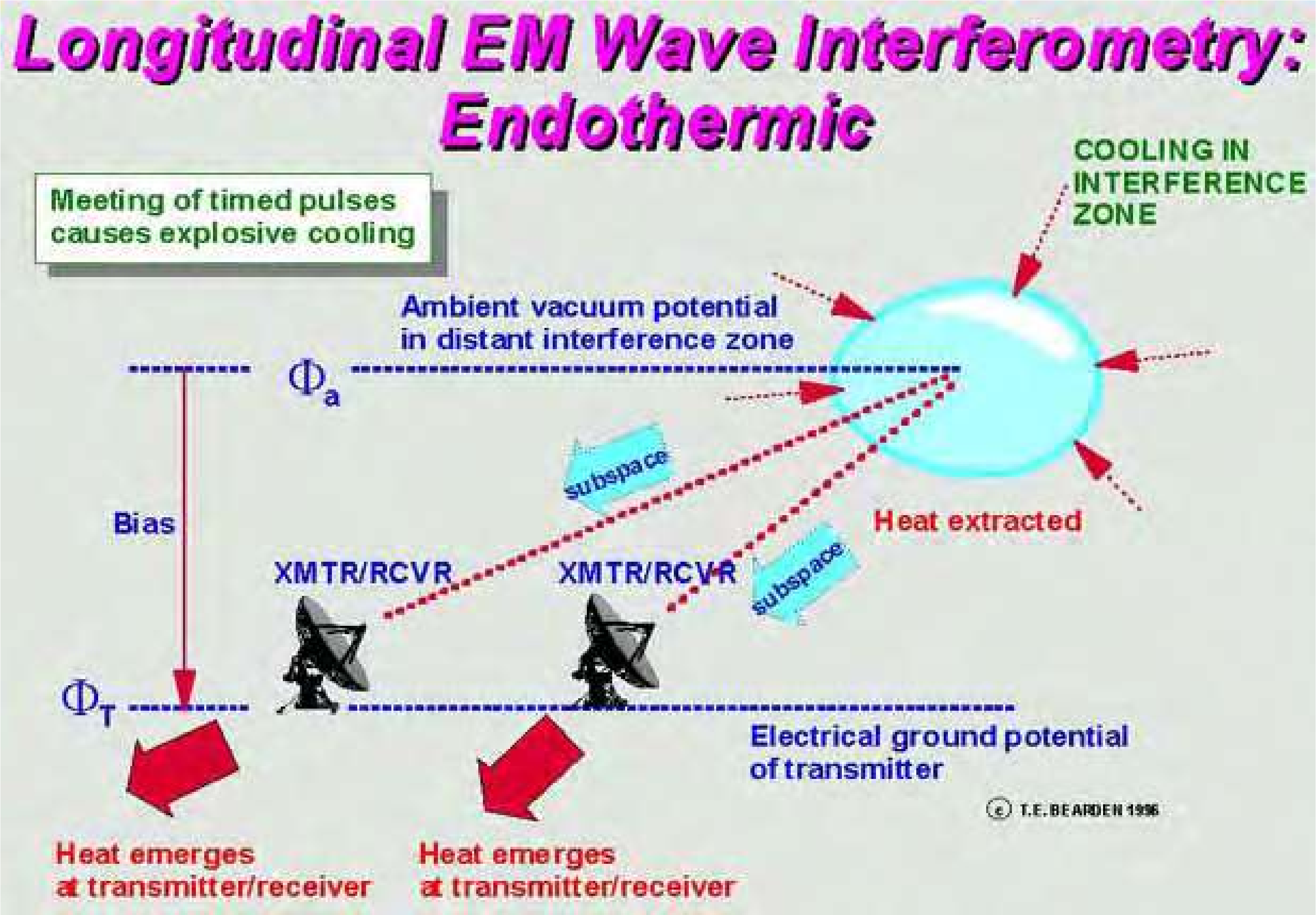

Longitudinal EM Wave Interferometry: Endothermic

Meeting of timed pulses
causes explosive cooling

COOLING IN
INTERFERENCE
ZONE

Ambient vacuum potential
in distant interference zone

Φ_a

subspace

Bias

Heat extracted

XMTR/RCVR

XMTR/RCVR

subspace

Φ_T

Electrical ground potential
of transmitter

© T.E. BEARDEN 1996

Heat emerges
at transmitter/receiver

Heat emerges
at transmitter/receiver

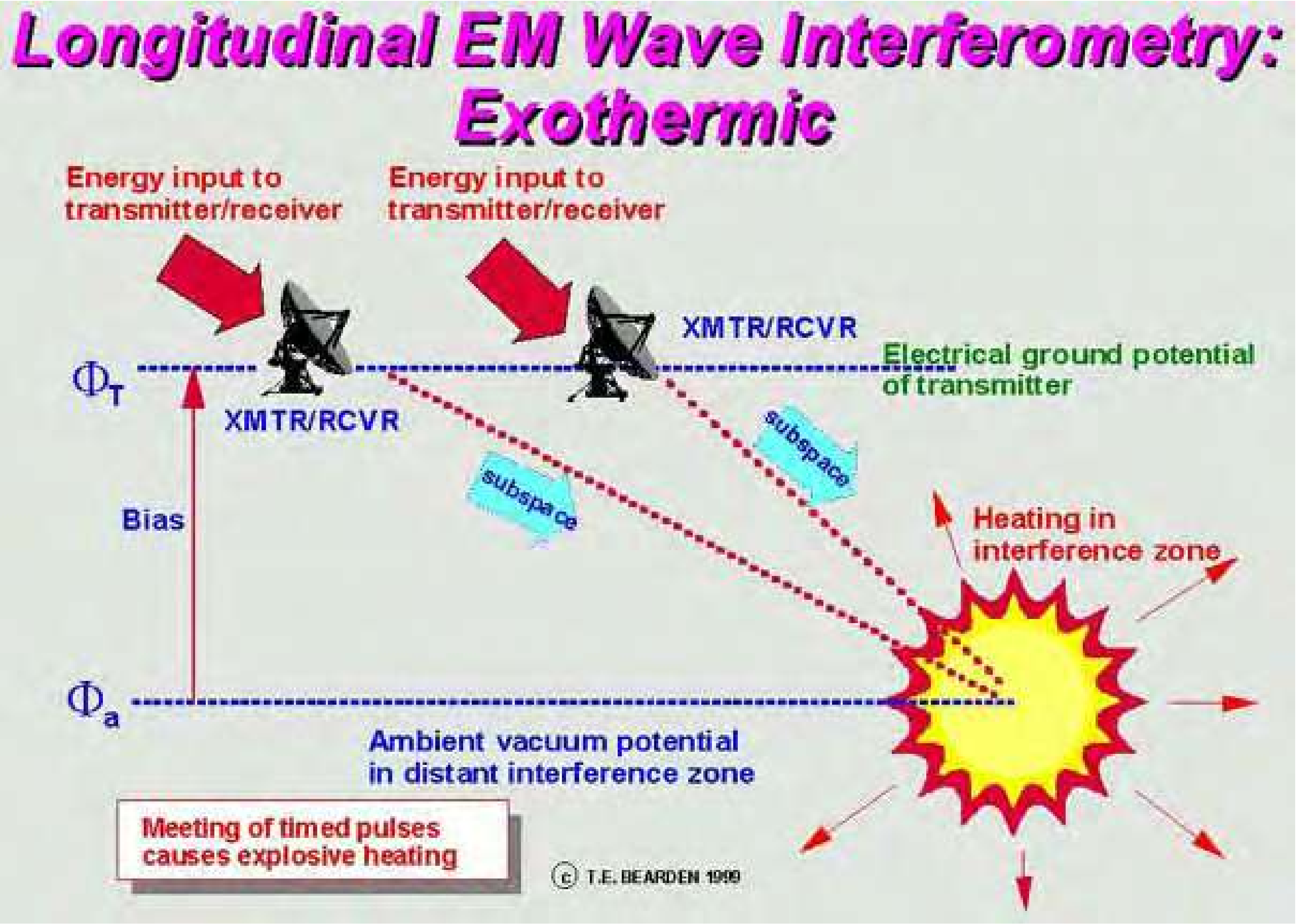

Longitudinal EM Wave Interferometry: Exothermic

Energy input to transmitter/receiver
Energy input to transmitter/receiver
XMTR/RCVR
XMTR/RCVR
Φ_T
Electrical ground potential of transmitter
subspace
subspace
Bias
Heating in interference zone
Φ_a
Ambient vacuum potential in distant interference zone

Meeting of timed pulses causes explosive heating

© T.E. BEARDEN 1999

The Tom Bearden Website

Help support the research

U.S. Air Attack on Libya

April 1986

- Slide 1
- Slide 2
- "Washington Post" Article
  January 22, 1989

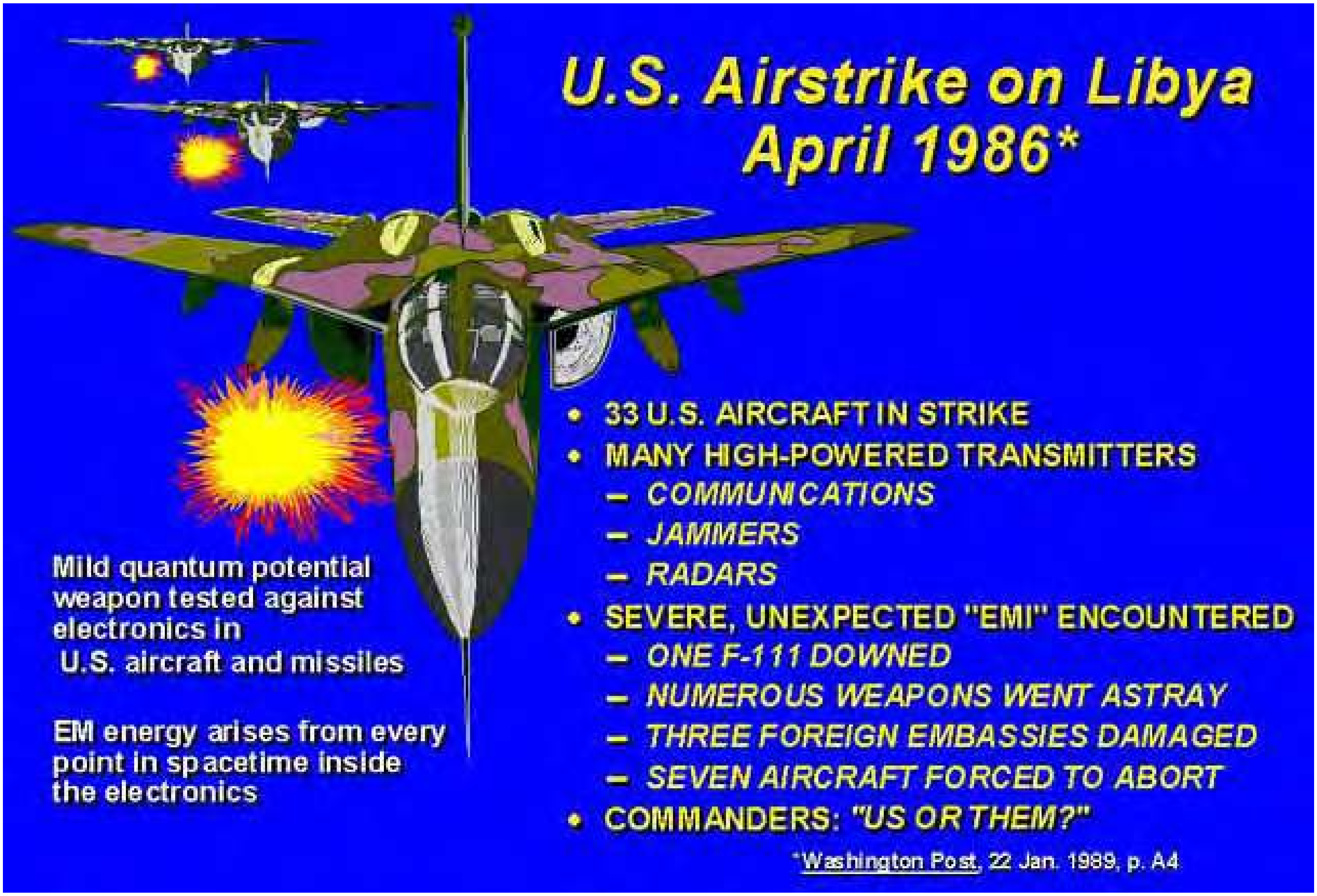

U.S. Airstrike on Libya April 1986*

Mild quantum potential weapon tested against electronics in U.S. aircraft and missiles

EM energy arises from every point in spacetime inside the electronics

- 33 U.S. AIRCRAFT IN STRIKE
- MANY HIGH-POWERED TRANSMITTERS
  - *COMMUNICATIONS*
  - *JAMMERS*
  - *RADARS*
- SEVERE, UNEXPECTED "EMI" ENCOUNTERED
  - *ONE F-111 DOWNED*
  - *NUMEROUS WEAPONS WENT ASTRAY*
  - *THREE FOREIGN EMBASSIES DAMAGED*
  - *SEVEN AIRCRAFT FORCED TO ABORT*
- COMMANDERS: *"US OR THEM?"*

*Washington Post, 22 Jan. 1989, p. A4

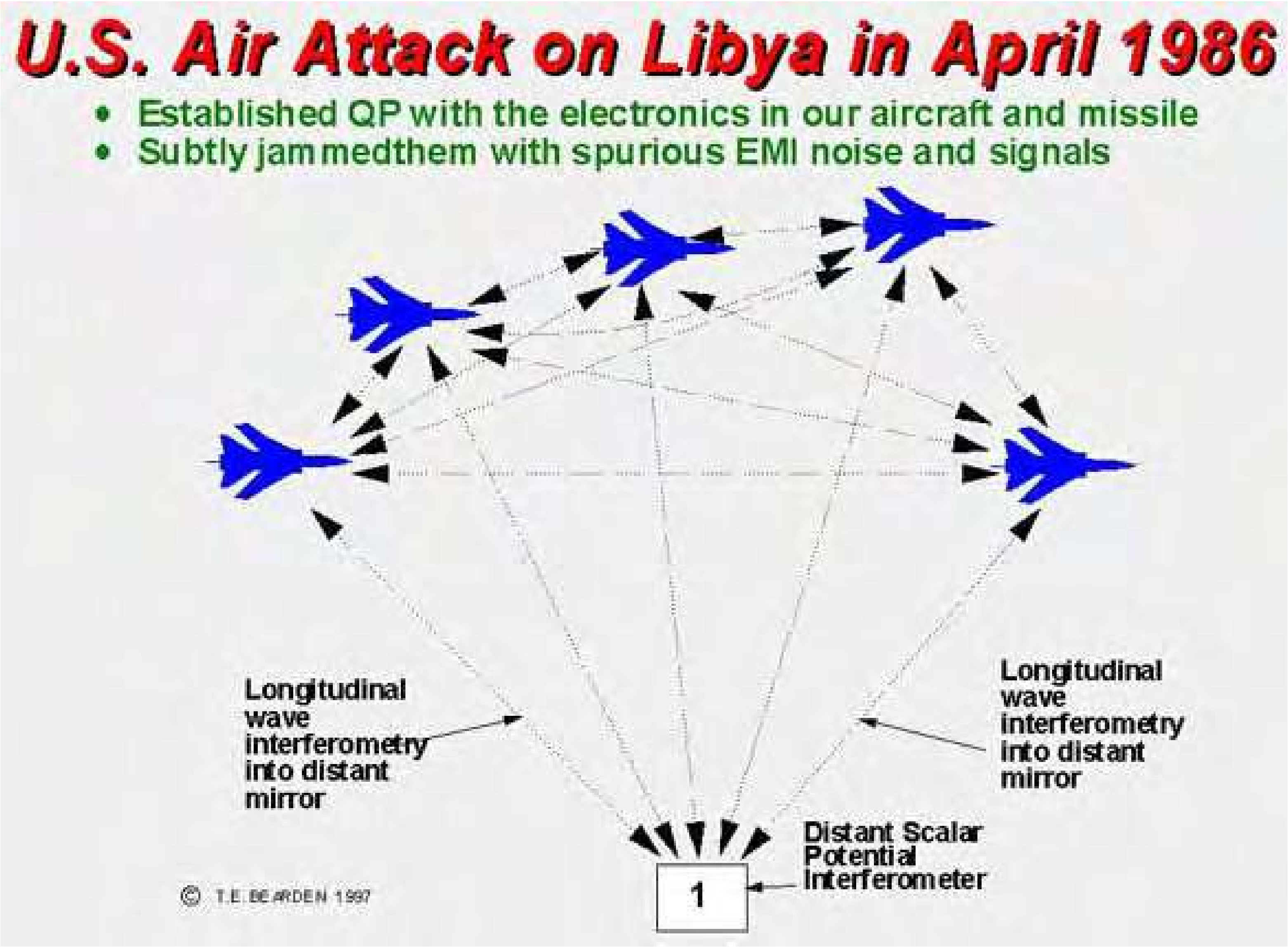

U.S. Air Attack on Libya in April 1986

- Established QP with the electronics in our aircraft and missile
- Subtly jammedthem with spurious EMI noise and signals

Longitudinal wave interferometry into distant mirror
Longitudinal wave interferometry into distant mirror
Distant Scalar Potential Interferometer
1
© T.E. BEARDEN 1997

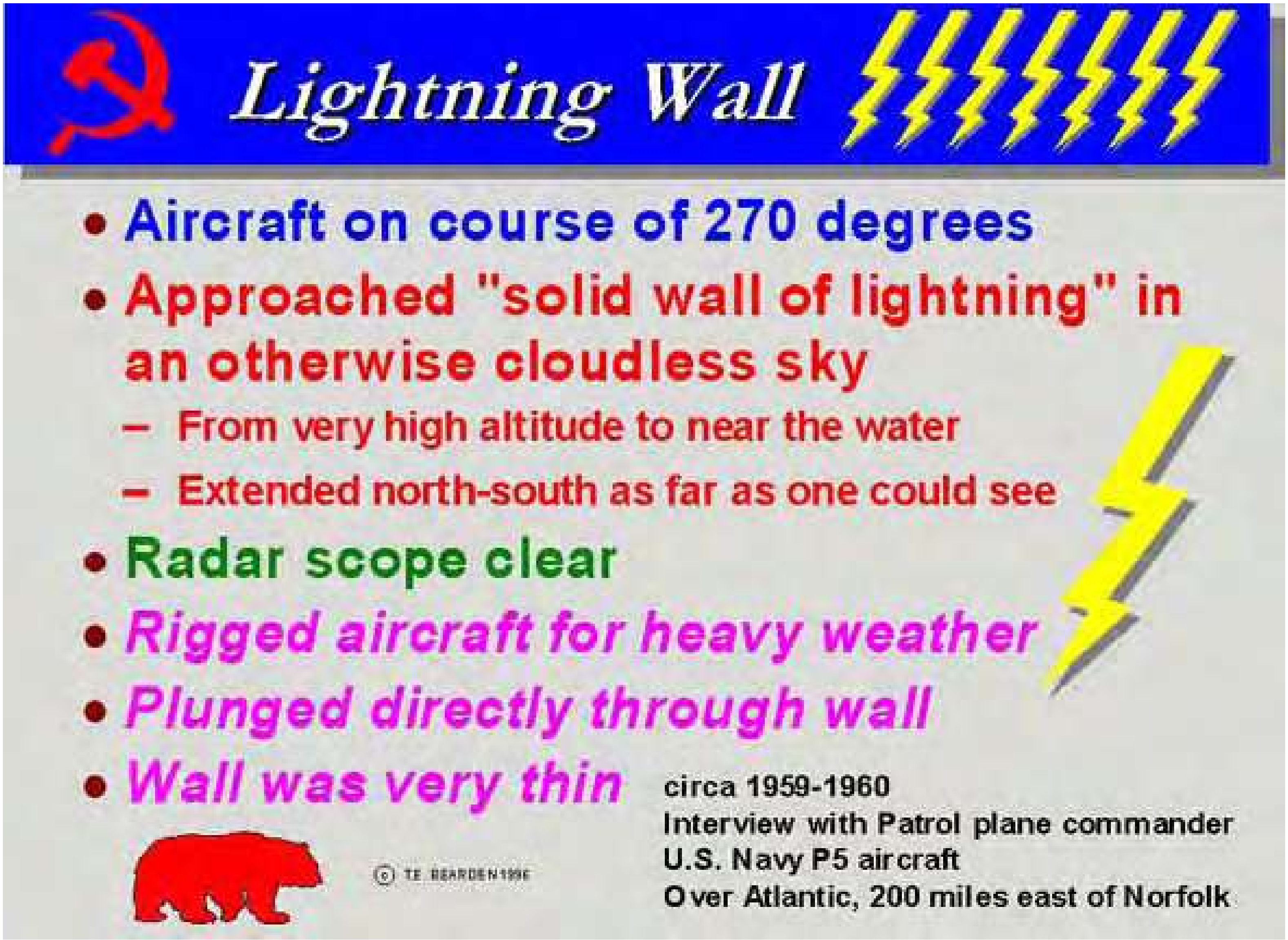

Lightning Wall

- Aircraft on course of 270 degrees
- Approached "solid wall of lightning" in an otherwise cloudless sky
  - From very high altitude to near the water
  - Extended north-south as far as one could see
- Radar scope clear
- *Rigged aircraft for heavy weather*
- *Plunged directly through wall*
- *Wall was very thin*

circa 1959-1960
Interview with Patrol plane commander
U.S. Navy P5 aircraft
Over Atlantic, 200 miles east of Norfolk

© T.E. BEARDEN 1996

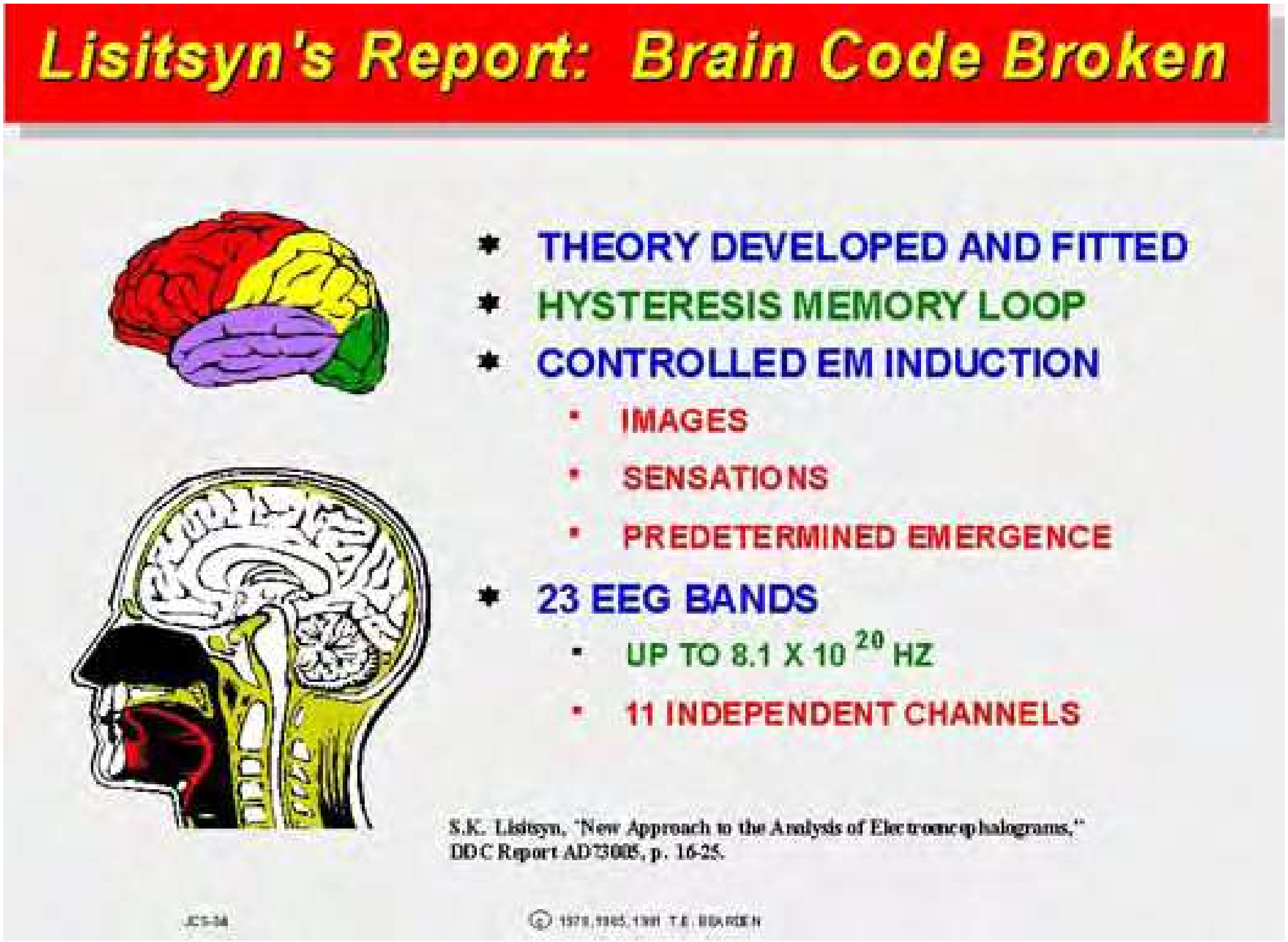

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
    - IMAGES
    - SENSATIONS
    - PREDETERMINED EMERGENCE
- 23 EEG BANDS
    - UP TO 8.1×10^{20} HZ
    - 11 INDEPENDENT CHANNELS

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

JCS-34

© 1978, 1985, 1991 T.E. BEARDEN

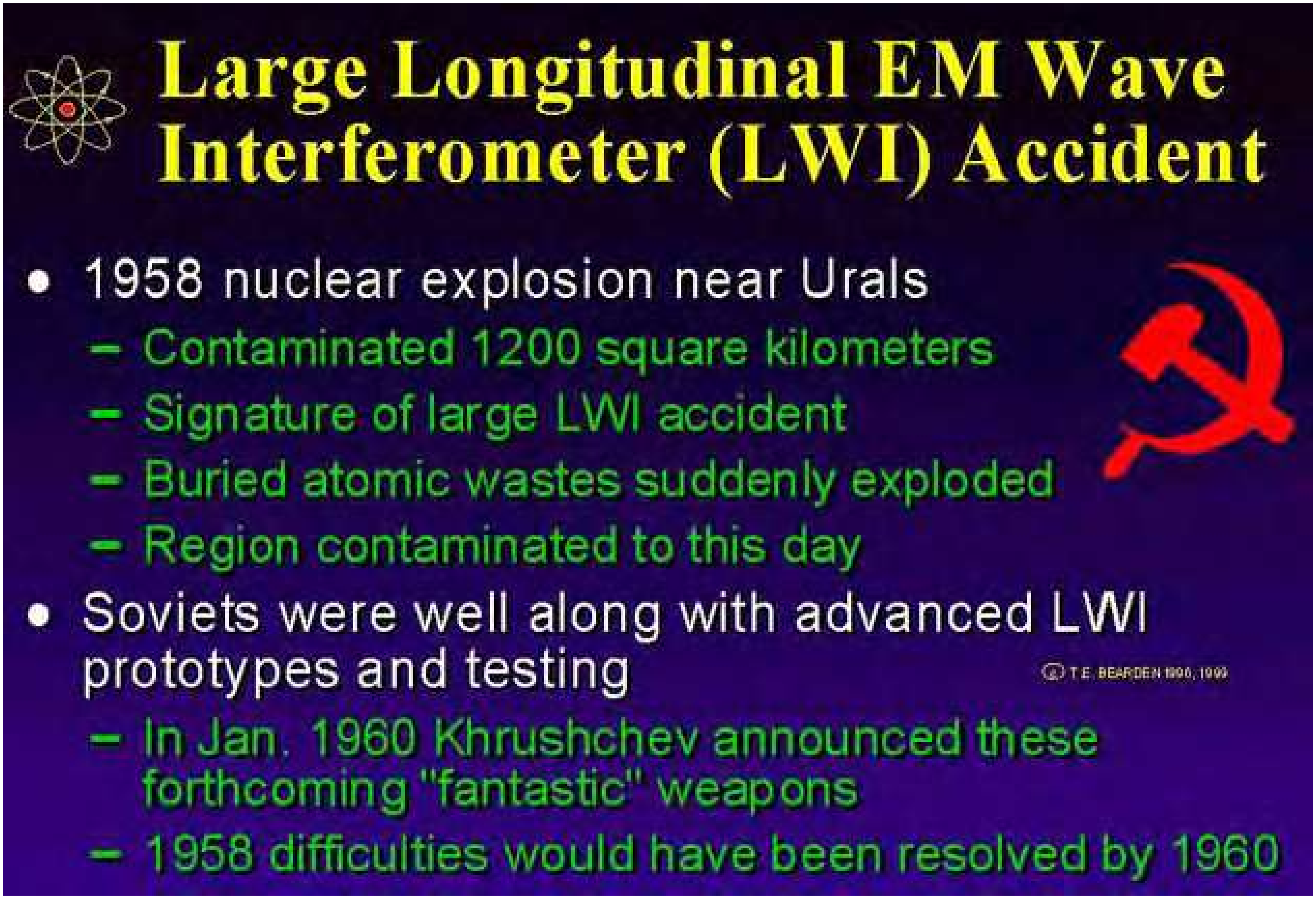

Large Longitudinal EM Wave Interferometer (LWI) Accident

- 1958 nuclear explosion near Urals
  - Contaminated 1200 square kilometers
  - Signature of large LWI accident
  - Buried atomic wastes suddenly exploded
  - Region contaminated to this day
- Soviets were well along with advanced LWI prototypes and testing
  - In Jan. 1960 Khrushchev announced these forthcoming "fantastic" weapons
  - 1958 difficulties would have been resolved by 1960

© T.E. BEARDEN 1996, 1999

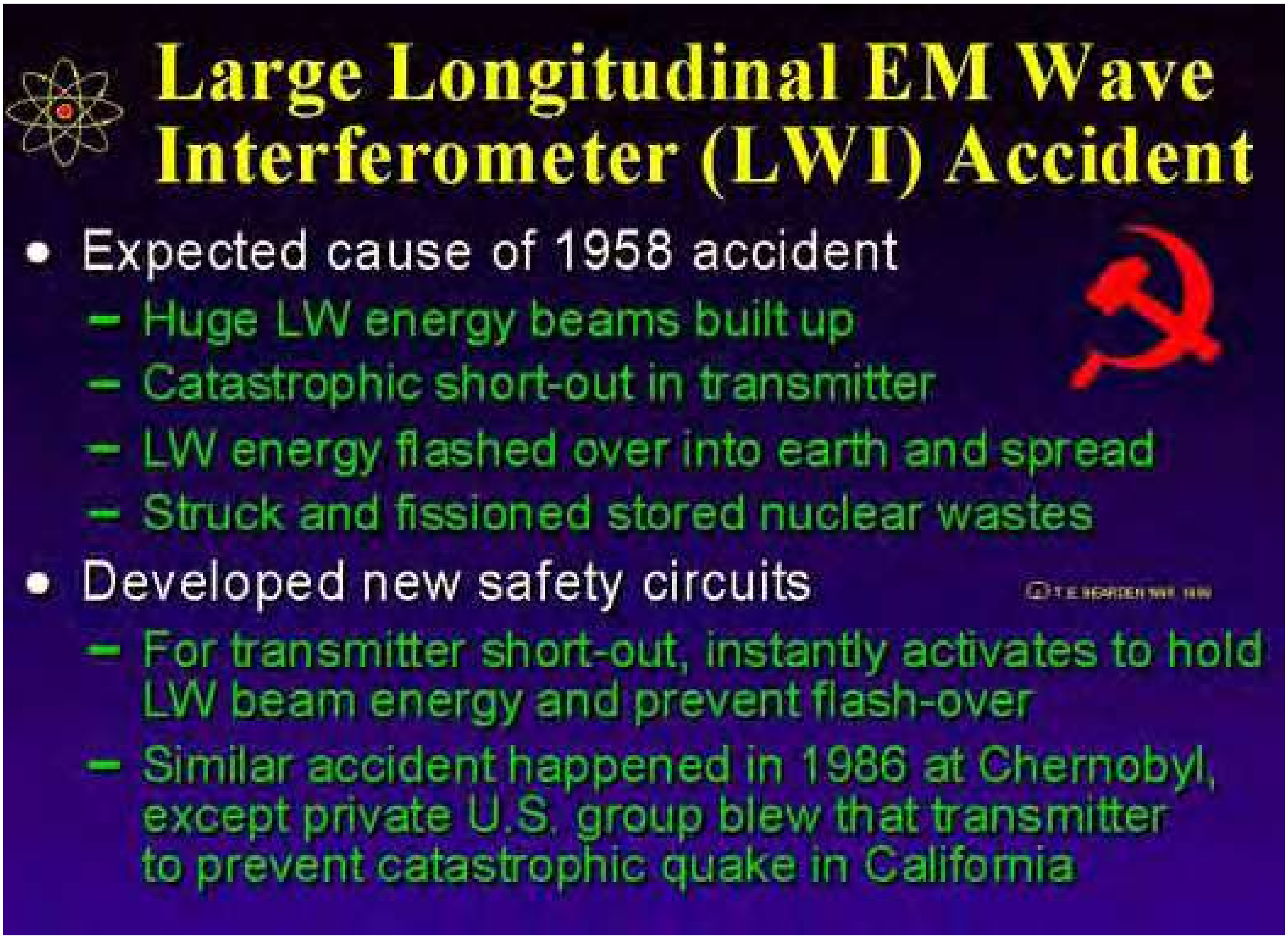

Large Longitudinal EM Wave Interferometer (LWI) Accident

- Expected cause of 1958 accident
  - Huge LW energy beams built up
  - Catastrophic short-out in transmitter
  - LW energy flashed over into earth and spread
  - Struck and fissioned stored nuclear wastes
- Developed new safety circuits
  - For transmitter short-out, instantly activates to hold LW beam energy and prevent flash-over
  - Similar accident happened in 1986 at Chernobyl, except private U.S. group blew that transmitter to prevent catastrophic quake in California

© T.E. BEARDEN [illegible]

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

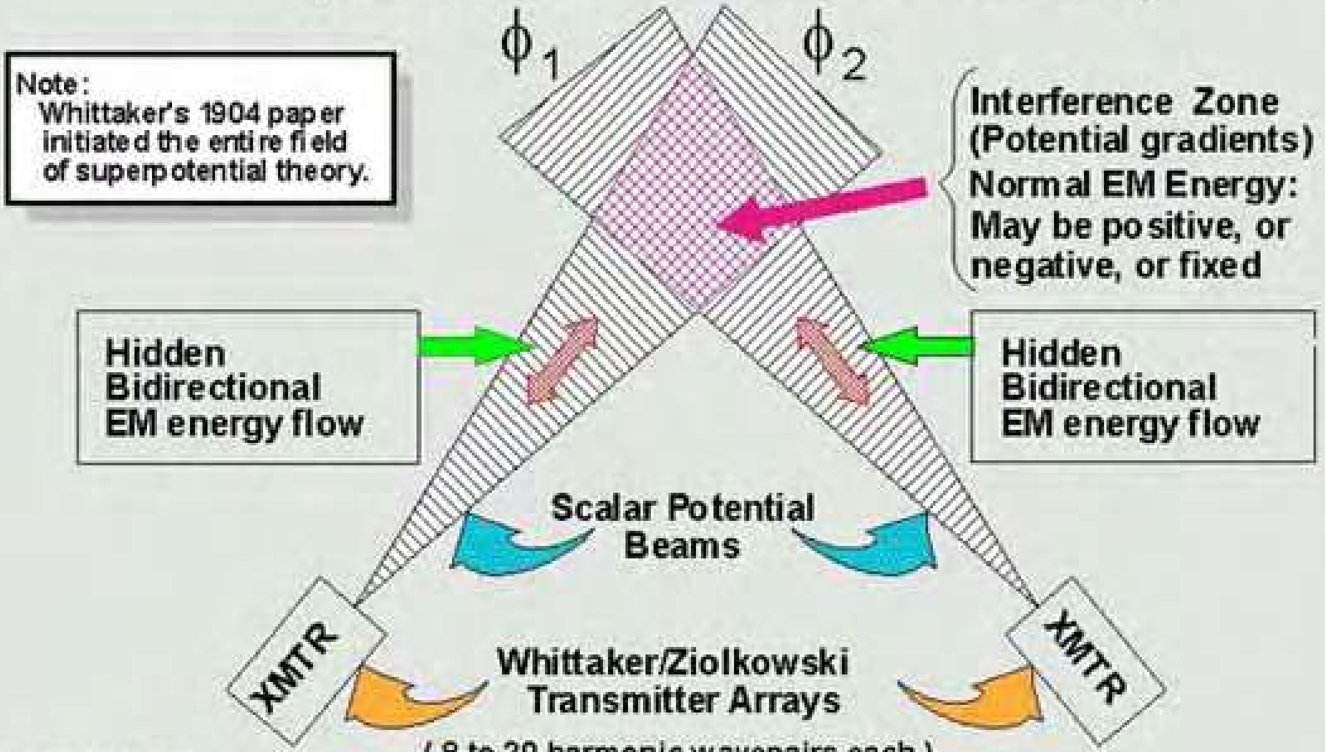
Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.
ϕ_1
ϕ_2
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

© T. E. Bearden 1994, 1999

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepair functions) produces all EM force fields and waves.

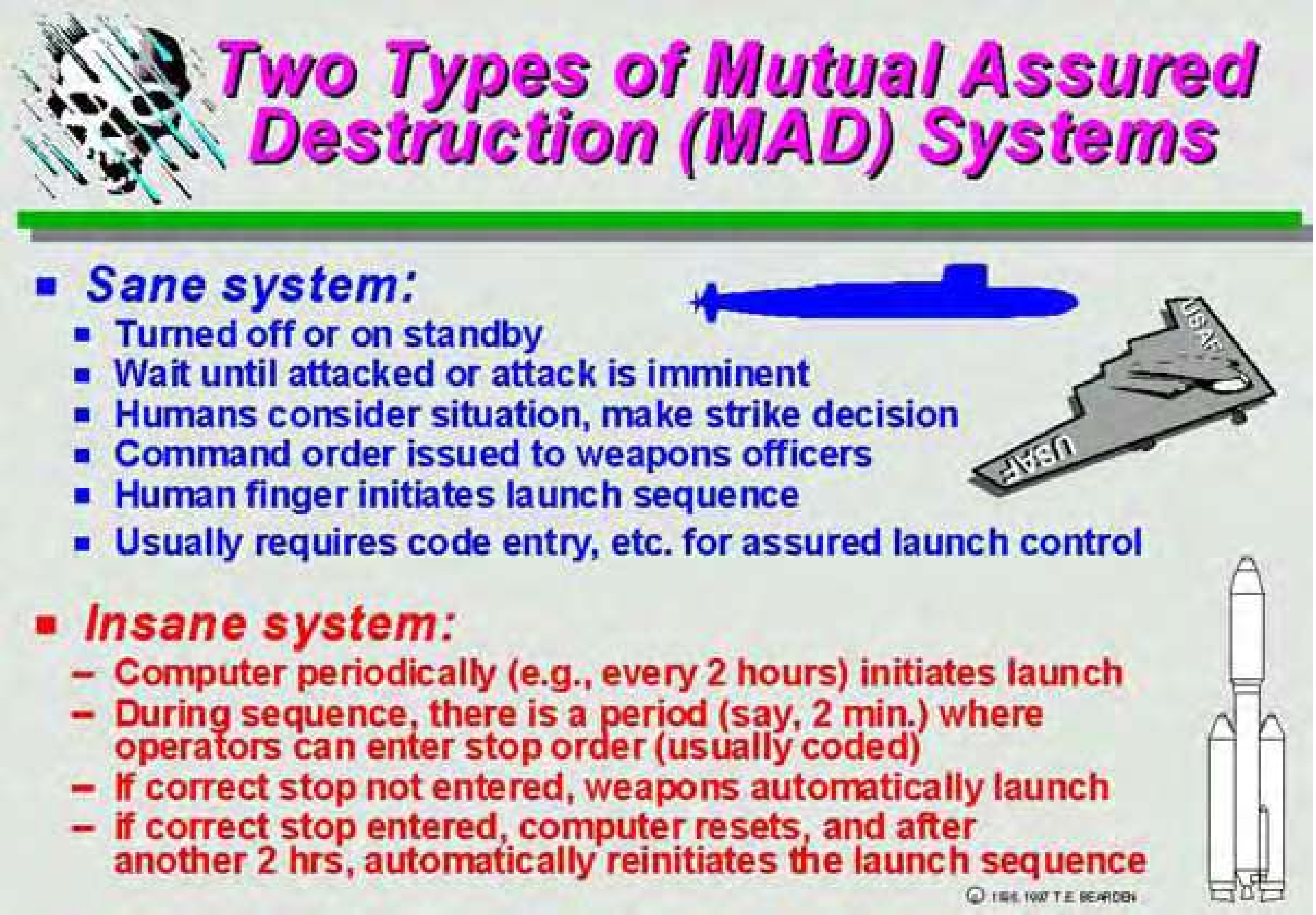

Two Types of Mutual Assured Destruction (MAD) Systems

- ***Sane system:***
  - Turned off or on standby
  - Wait until attacked or attack is imminent
  - Humans consider situation, make strike decision
  - Command order issued to weapons officers
  - Human finger initiates launch sequence
  - Usually requires code entry, etc. for assured launch control
- ***Insane system:***
  - Computer periodically (e.g., every 2 hours) initiates launch
  - During sequence, there is a period (say, 2 min.) where operators can enter stop order (usually coded)
  - If correct stop not entered, weapons automatically launch
  - If correct stop entered, computer resets, and after another 2 hrs, automatically reinitiates the launch sequence

© 1996, 1997 T.E. BEARDEN

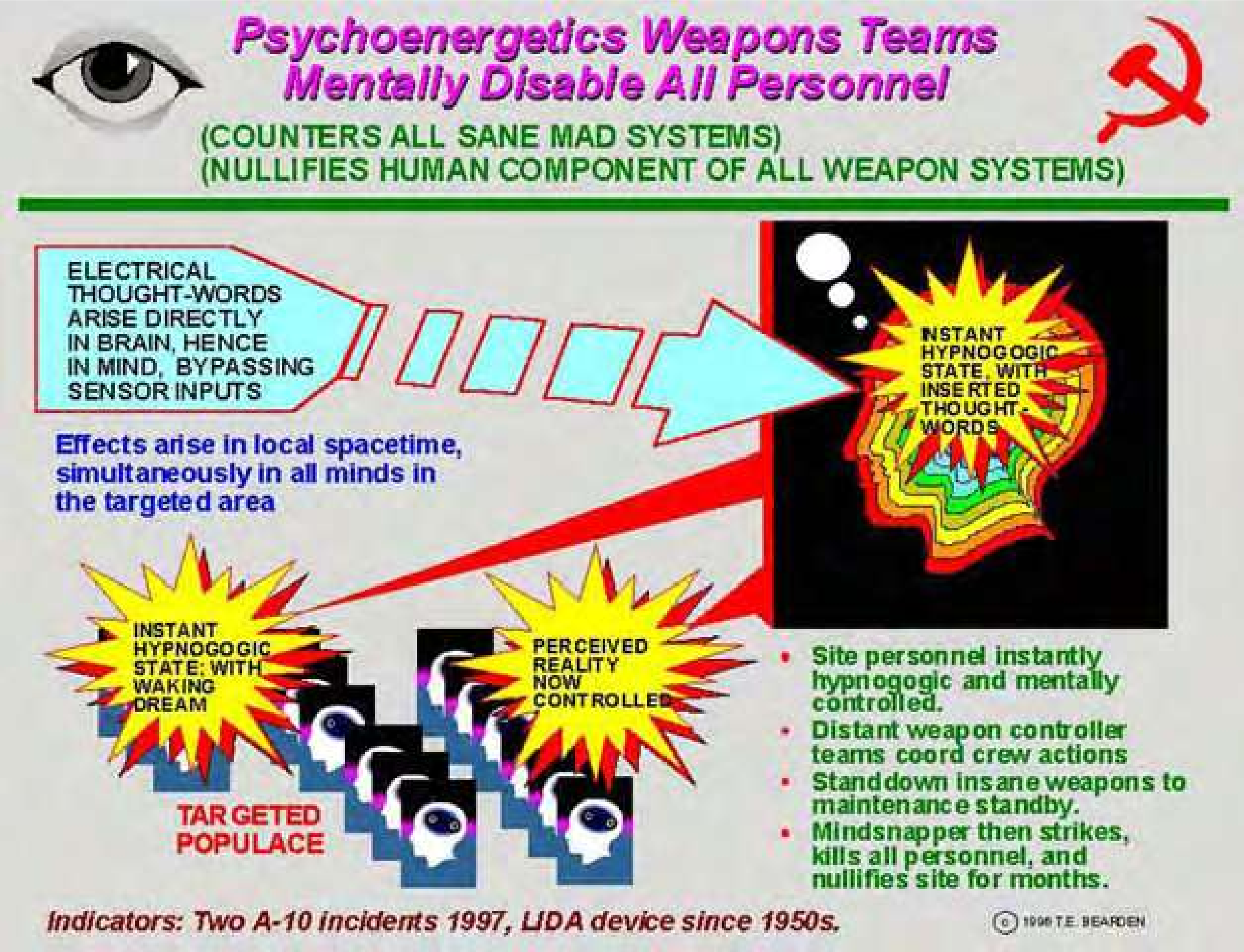

Psychoenergetics Weapons Teams Mentally Disable All Personnel

(COUNTERS ALL SANE MAD SYSTEMS)
(NULLIFIES HUMAN COMPONENT OF ALL WEAPON SYSTEMS)

ELECTRICAL THOUGHT-WORDS ARISE DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED THOUGHT-WORDS

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED POPULACE

- Site personnel instantly hypnogogic and mentally controlled.
- Distant weapon controller teams coord crew actions
- Standdown insane weapons to maintenance standby.
- Mindsnapper then strikes, kills all personnel, and nullifies site for months.

Indicators: Two A-10 incidents 1997, LIDA device since 1950s.

© 1998 T.E. BEARDEN

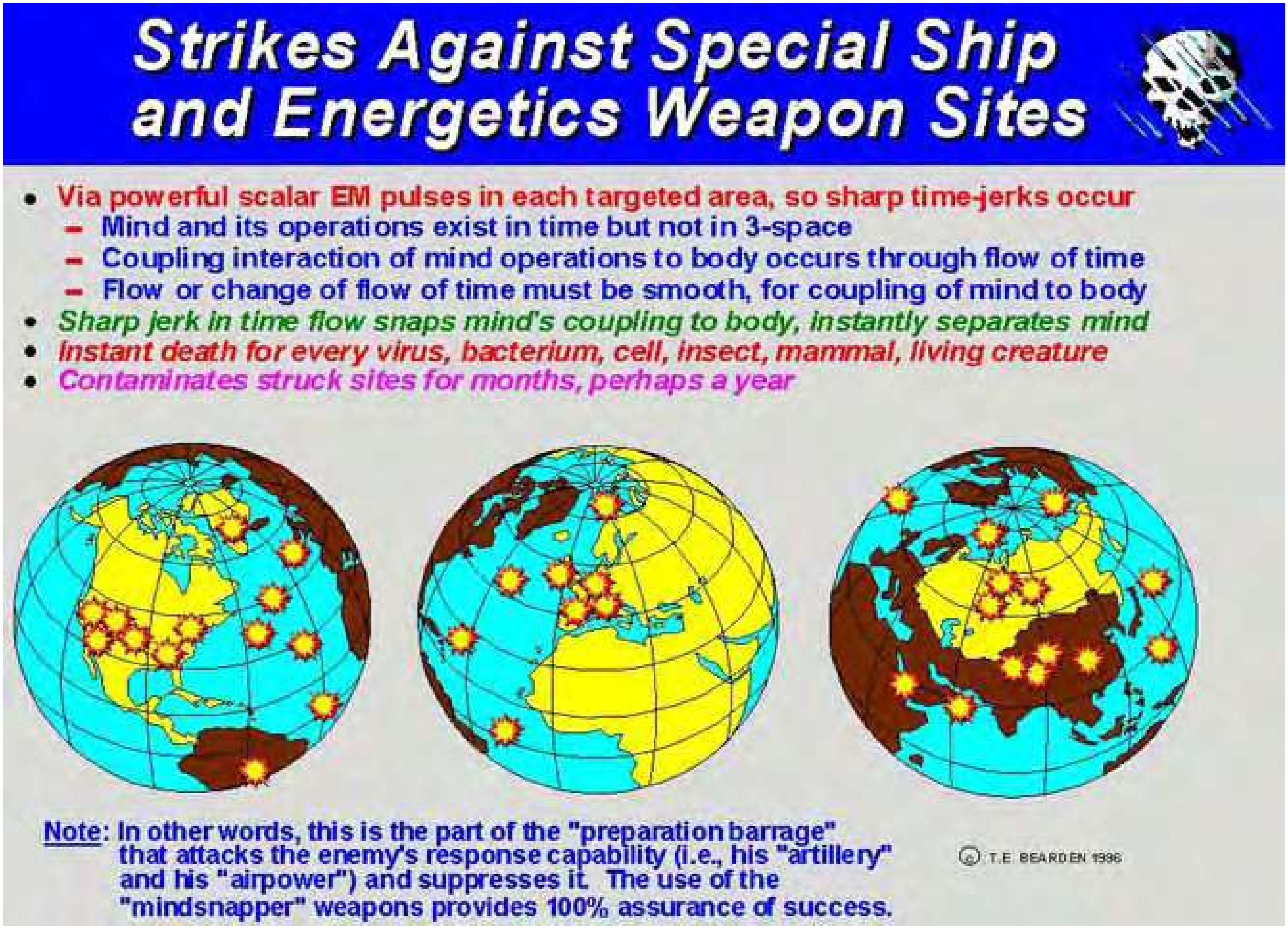

Strikes Against Special Ship and Energetics Weapon Sites

- Via powerful scalar EM pulses in each targeted area, so sharp time-jerks occur
  - Mind and its operations exist in time but not in 3-space
  - Coupling interaction of mind operations to body occurs through flow of time
  - Flow or change of flow of time must be smooth, for coupling of mind to body
- *Sharp jerk in time flow snaps mind's coupling to body, instantly separates mind*
- *Instant death for every virus, bacterium, cell, insect, mammal, living creature*
- *Contaminates struck sites for months, perhaps a year*

Note: In other words, this is the part of the "preparation barrage" that attacks the enemy's response capability (i.e., his "artillery" and his "airpower") and suppresses it. The use of the "mindsnapper" weapons provides 100% assurance of success.

© T.E. BEARDEN 1996

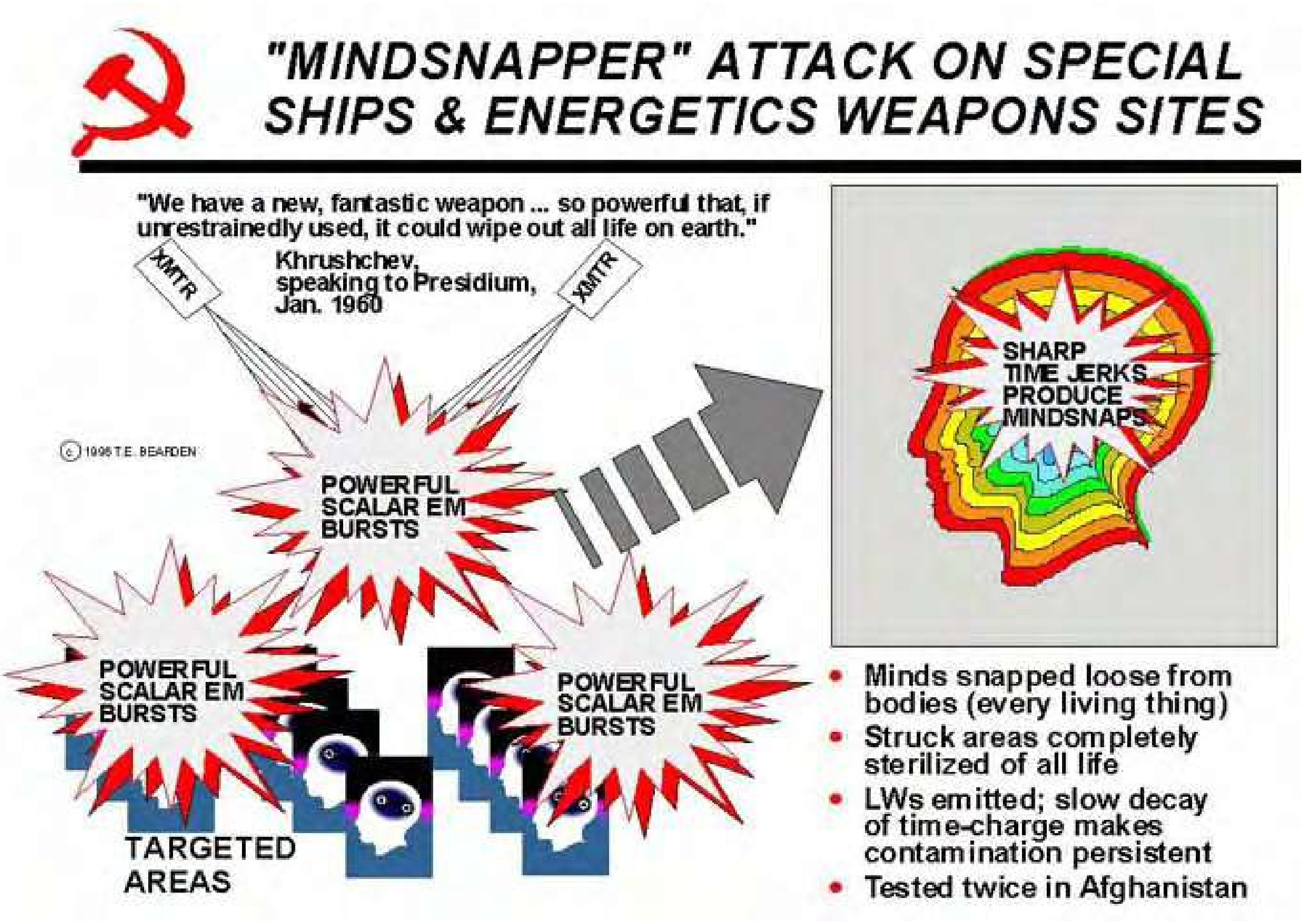

"MINDSNAPPER" ATTACK ON SPECIAL SHIPS & ENERGETICS WEAPONS SITES

"We have a new, fantastic weapon ... so powerful that, if unrestrainedly used, it could wipe out all life on earth."

Khrushchev,
speaking to Presidium,
Jan. 1960

XMTR
XMTR
ⓒ 1996 T.E. BEARDEN
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
TARGETED AREAS
SHARP TIME JERKS PRODUCE MINDSNAPS

- Minds snapped loose from bodies (every living thing)
- Struck areas completely sterilized of all life
- LWs emitted; slow decay of time-charge makes contamination persistent
- Tested twice in Afghanistan

[The Tom Bearden Website](#)

[**Help support the research**](#)

How Much Off Guard Are We?

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6
- Slide 7

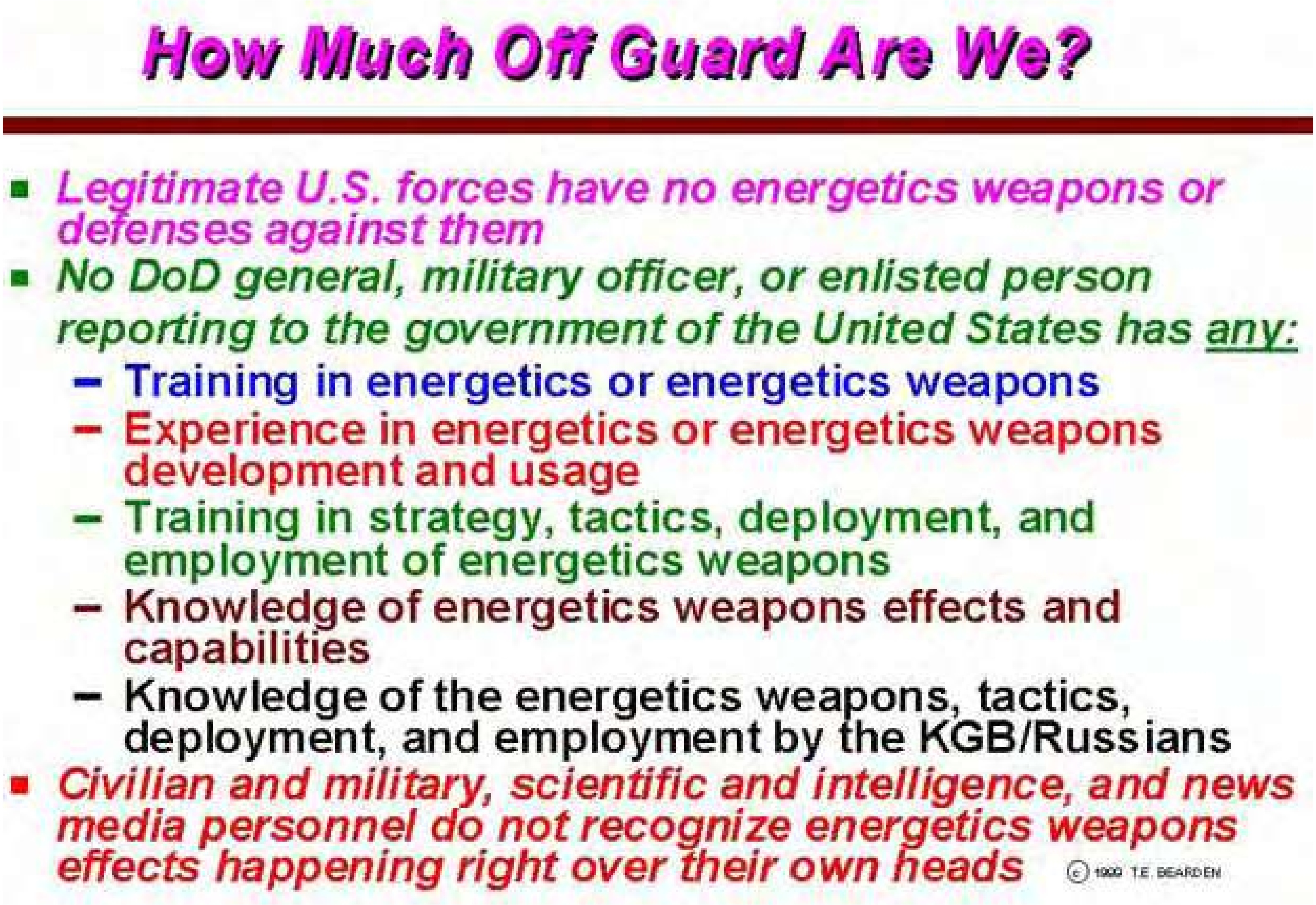

How Much Off Guard Are We?

- *Legitimate U.S. forces have no energetics weapons or defenses against them*
- *No DoD general, military officer, or enlisted person reporting to the government of the United States has any:*
    - Training in energetics or energetics weapons
    - Experience in energetics or energetics weapons development and usage
    - Training in strategy, tactics, deployment, and employment of energetics weapons
    - Knowledge of energetics weapons effects and capabilities
    - Knowledge of the energetics weapons, tactics, deployment, and employment by the KGB/Russians
- *Civilian and military, scientific and intelligence, and news media personnel do not recognize energetics weapons effects happening right over their own heads*

© 1999 T.E. BEARDEN

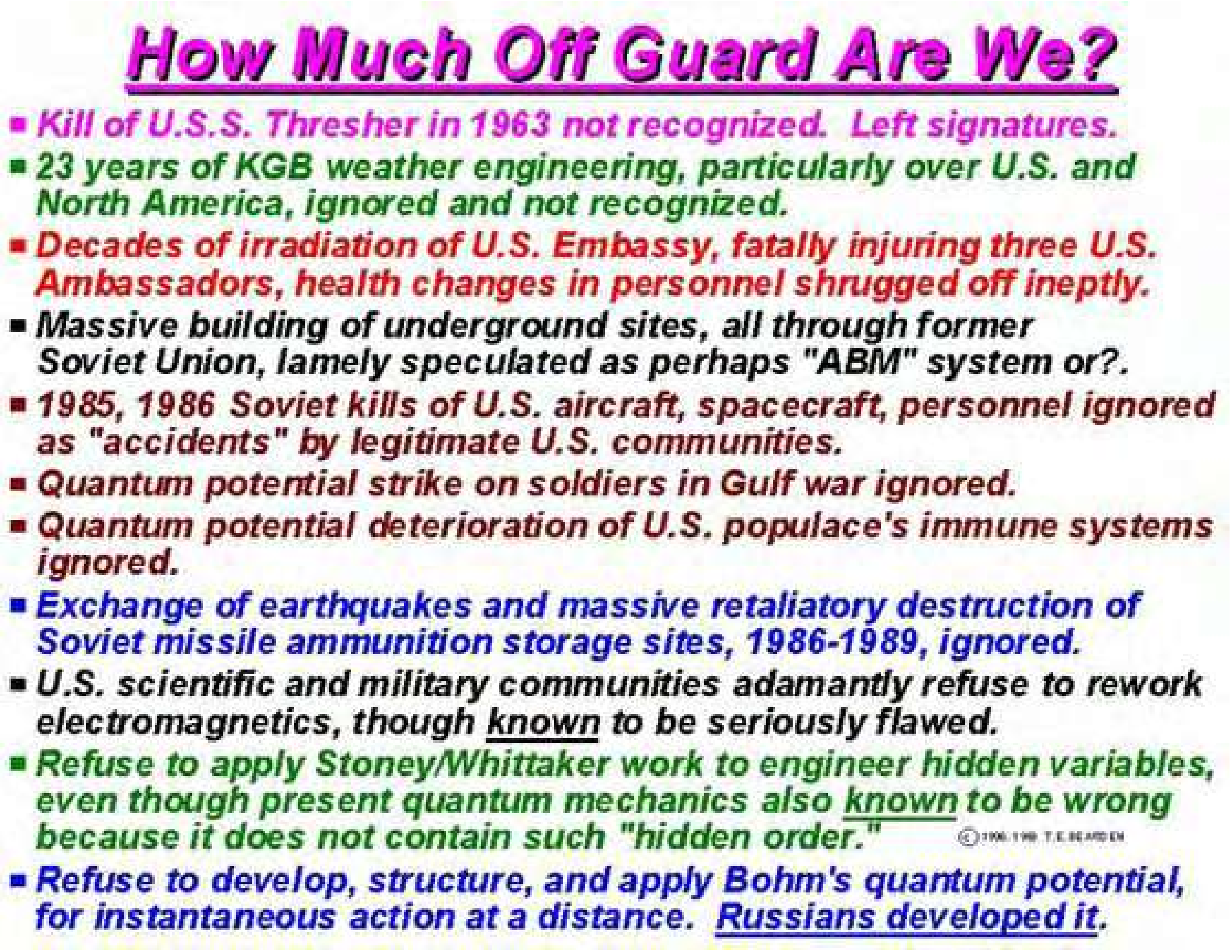

How Much Off Guard Are We?

- Kill of U.S.S. Thresher in 1963 not recognized. Left signatures.
- 23 years of KGB weather engineering, particularly over U.S. and North America, ignored and not recognized.
- Decades of irradiation of U.S. Embassy, fatally injuring three U.S. Ambassadors, health changes in personnel shrugged off ineptly.
- Massive building of underground sites, all through former Soviet Union, lamely speculated as perhaps "ABM" system or?.
- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.
- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."
- Refuse to develop, structure, and apply Bohm's quantum potential, for instantaneous action at a distance. Russians developed it.

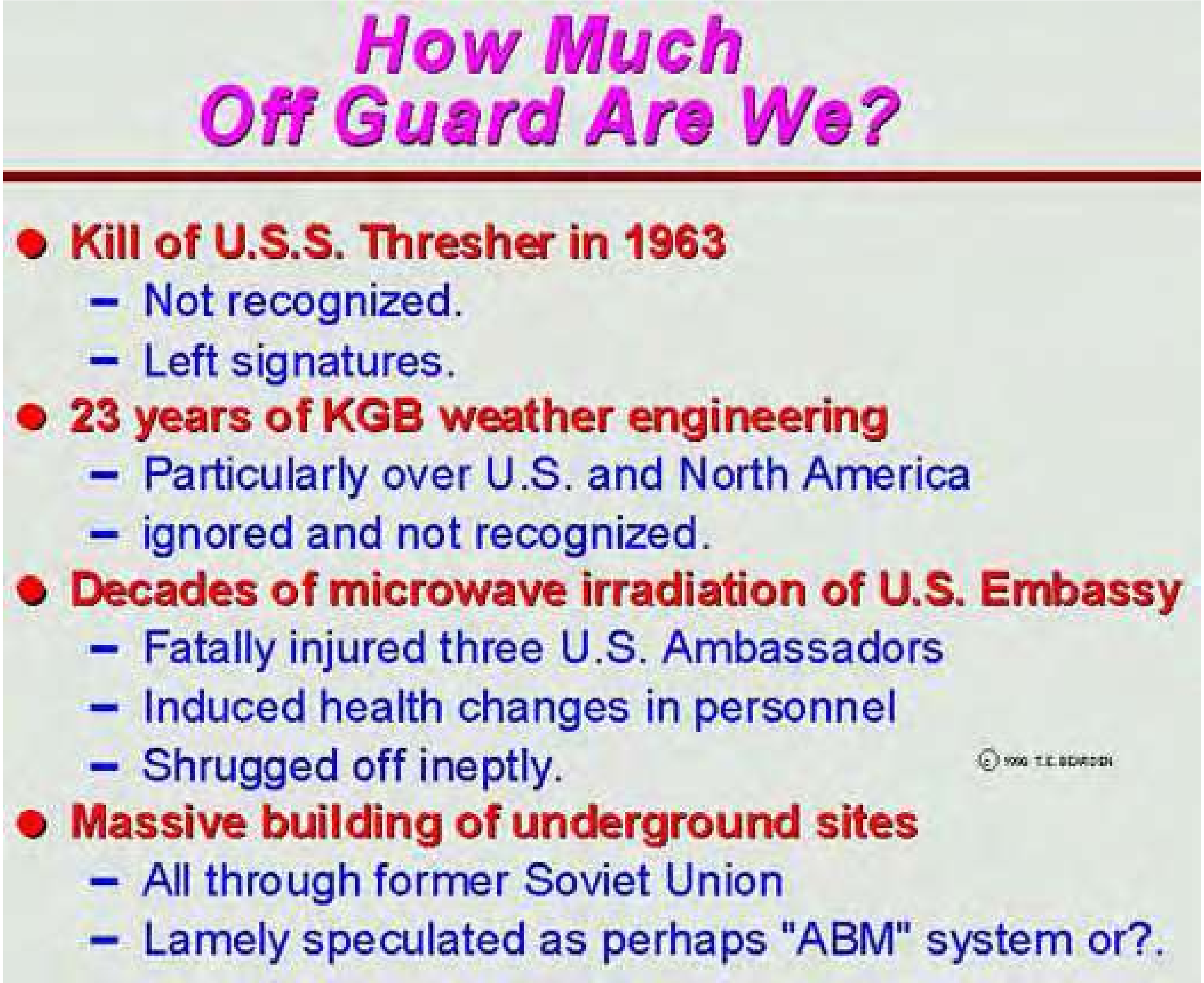

How Much Off Guard Are We?

- **Kill of U.S.S. Thresher in 1963**
  - Not recognized.
  - Left signatures.
- **23 years of KGB weather engineering**
  - Particularly over U.S. and North America
  - ignored and not recognized.
- **Decades of microwave irradiation of U.S. Embassy**
  - Fatally injured three U.S. Ambassadors
  - Induced health changes in personnel
  - Shrugged off ineptly.
- **Massive building of underground sites**
  - All through former Soviet Union
  - Lamely speculated as perhaps "ABM" system or?.

© 1990 T.E. BEARDEN

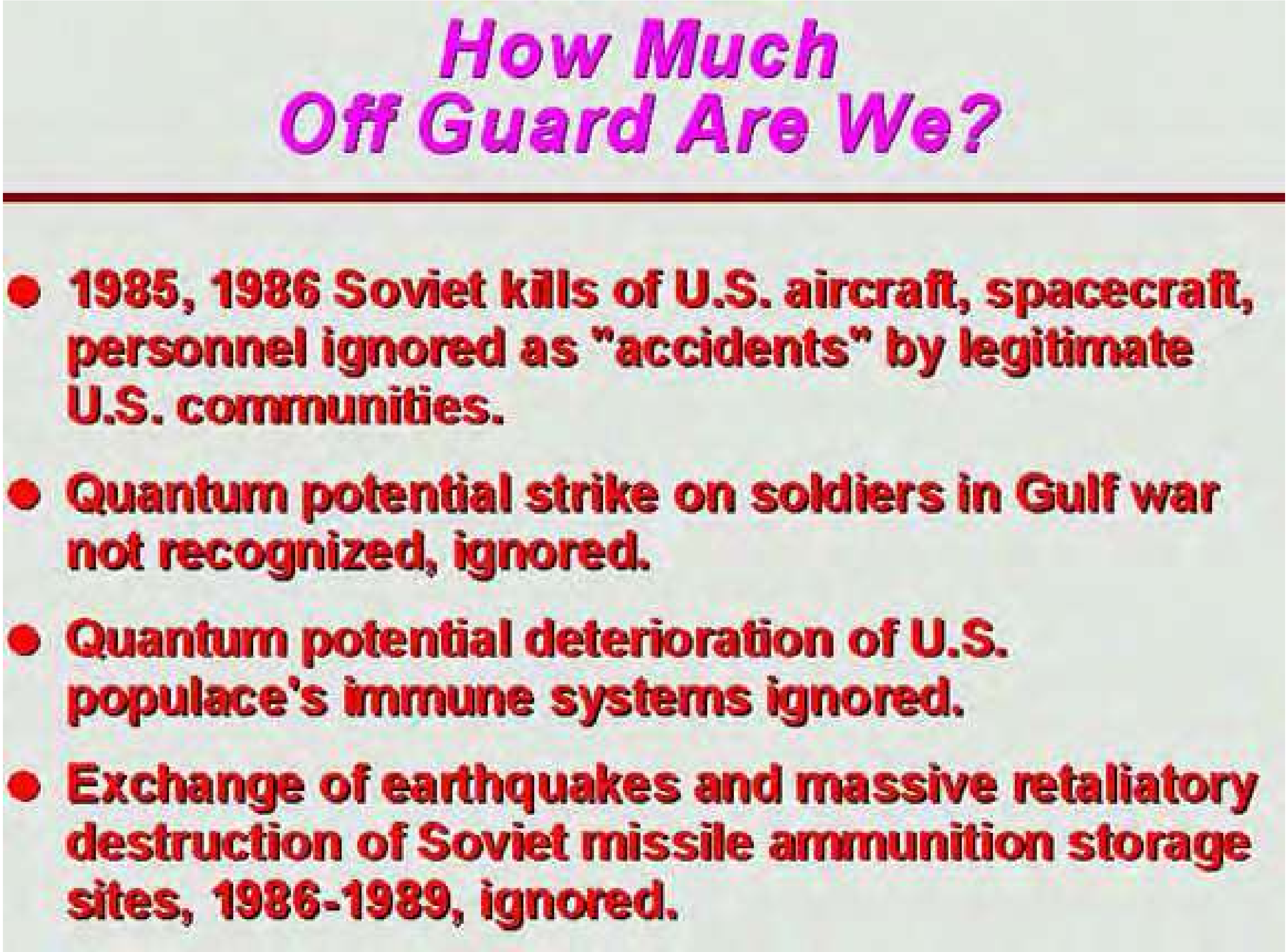

How Much
Off Guard Are We?

- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war not recognized, ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.

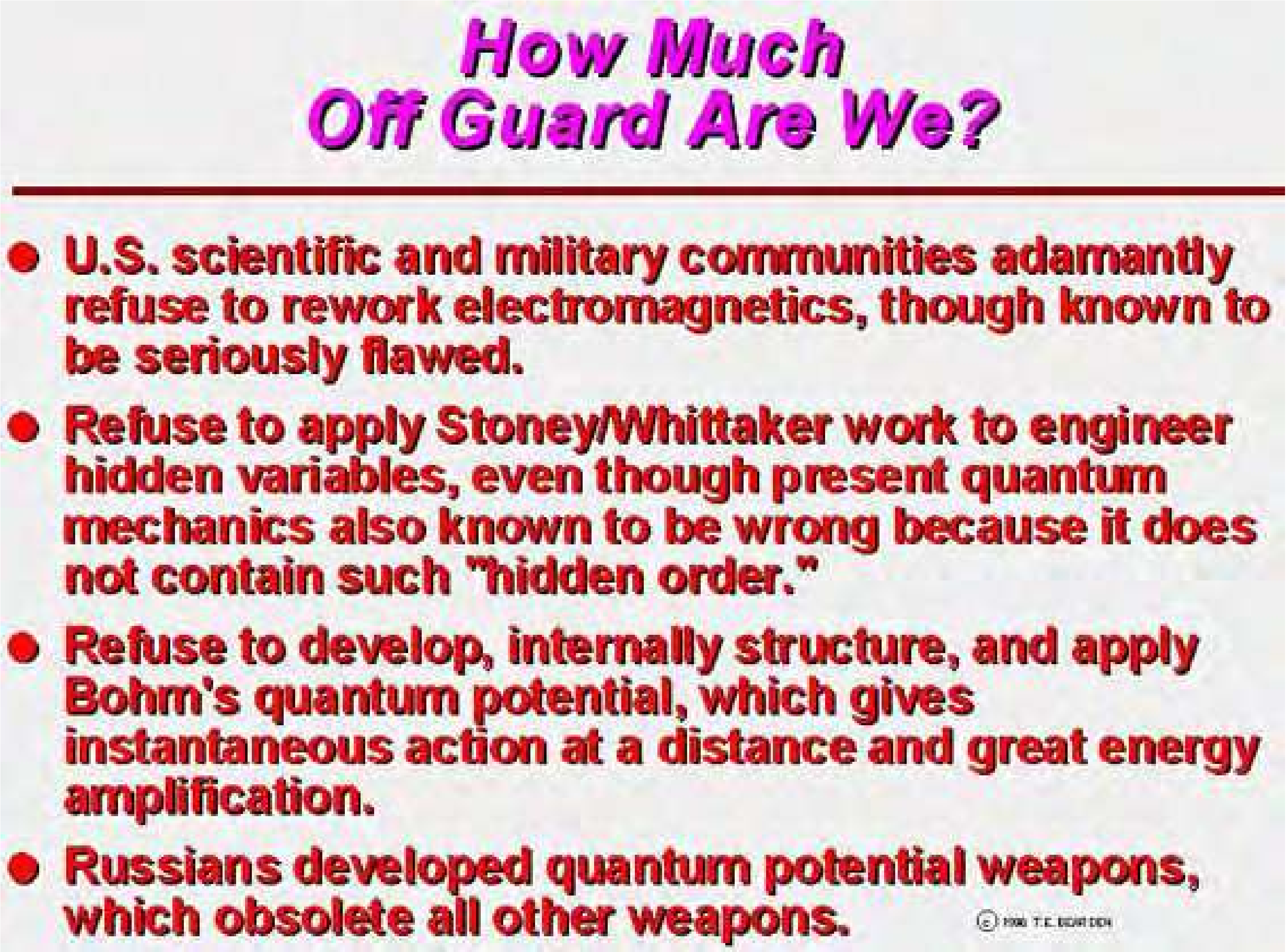

How Much Off Guard Are We?

- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."
- Refuse to develop, internally structure, and apply Bohm's quantum potential, which gives instantaneous action at a distance and great energy amplification.
- Russians developed quantum potential weapons, which obsolete all other weapons.

© 1996 T.E. BEARDEN

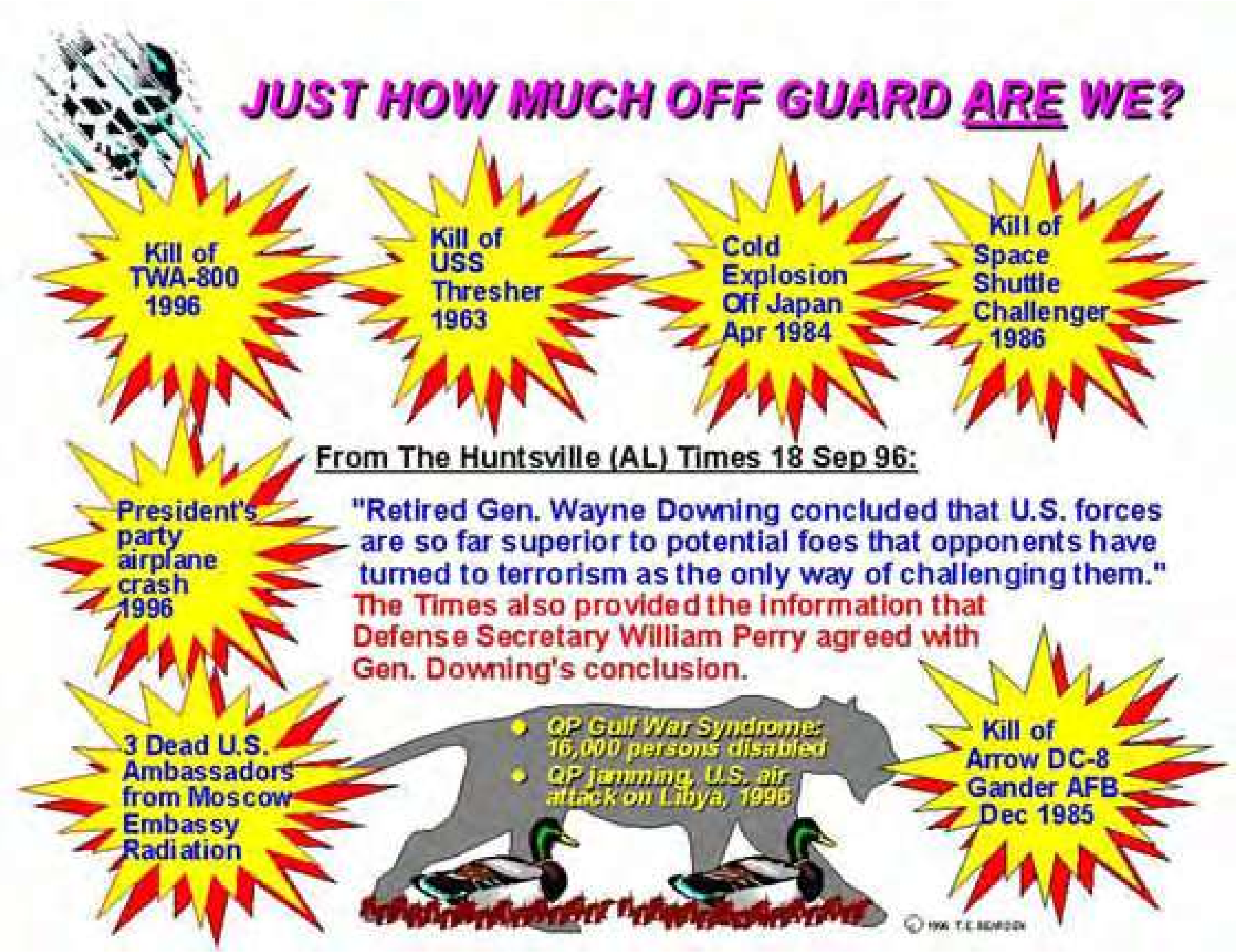

JUST HOW MUCH OFF GUARD ARE WE?

Kill of
TWA-800
1996

Kill of
USS
Thresher
1963

Cold
Explosion
Off Japan
Apr 1984

Kill of
Space
Shuttle
Challenger
1986

President's
party
airplane
crash
1996

From The Huntsville (AL) Times 18 Sep 96:

"Retired Gen. Wayne Downing concluded that U.S. forces are so far superior to potential foes that opponents have turned to terrorism as the only way of challenging them."

The Times also provided the information that Defense Secretary William Perry agreed with Gen. Downing's conclusion.

3 Dead U.S.
Ambassadors
from Moscow
Embassy
Radiation

- *QP Gulf War Syndrome: 16,000 persons disabled*
- *QP jamming, U.S. air attack on Libya, 1996*

Kill of
Arrow DC-8
Gander AFB
Dec 1985

© 1996 T.E. BEARDEN

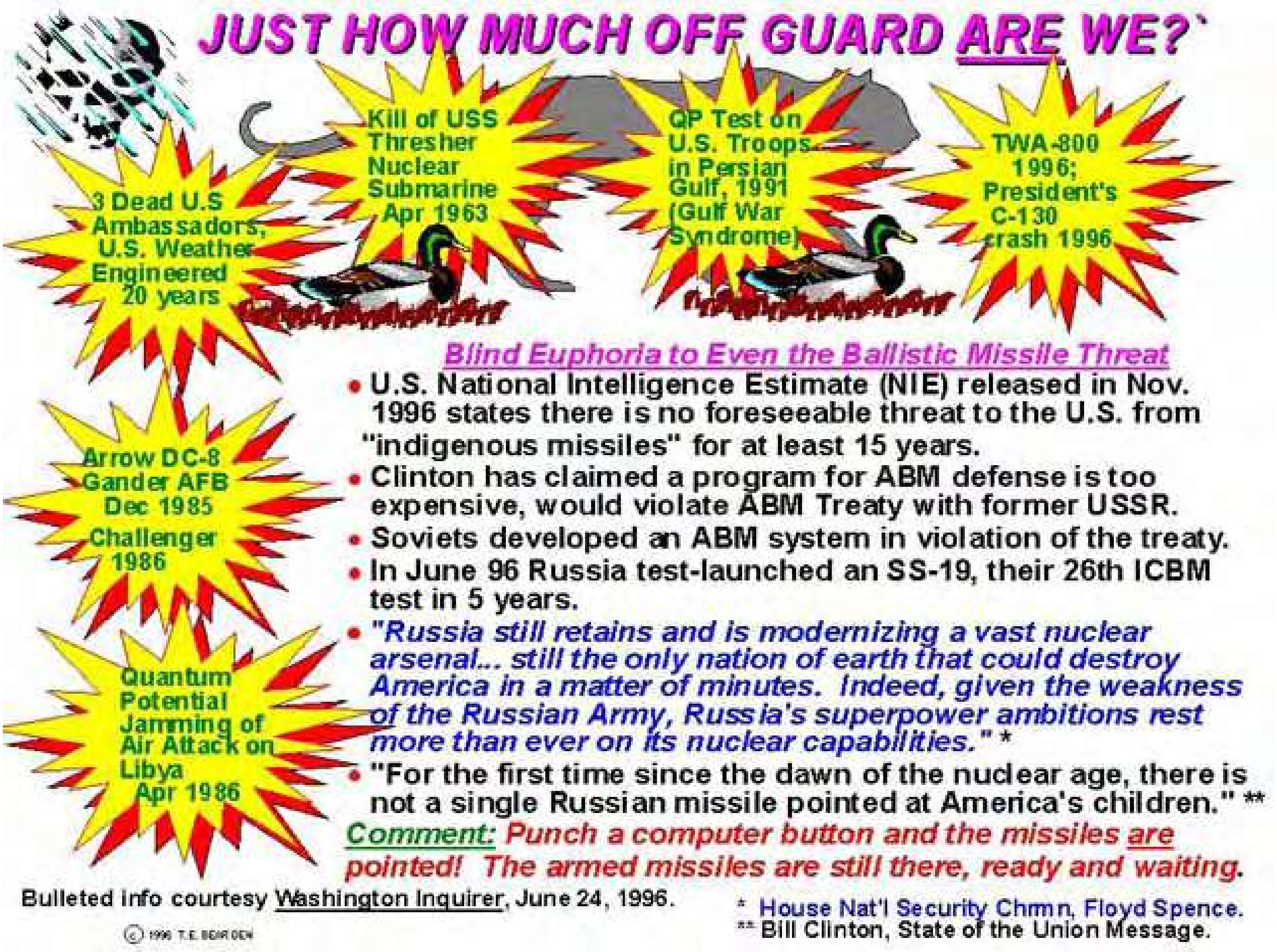

JUST HOW MUCH OFF GUARD ARE WE?

3 Dead U.S Ambassadors, U.S. Weather Engineered 20 years

Kill of USS Thresher Nuclear Submarine Apr 1963

QP Test on U.S. Troops in Persian Gulf, 1991 (Gulf War Syndrome)

TWA-800 1996; President's C-130 crash 1996

Arrow DC-8 Gander AFB Dec 1985

Challenger 1986

Quantum Potential Jamming of Air Attack on Libya Apr 1986

Blind Euphoria to Even the Ballistic Missile Threat

- U.S. National Intelligence Estimate (NIE) released in Nov. 1996 states there is no foreseeable threat to the U.S. from "indigenous missiles" for at least 15 years.
- Clinton has claimed a program for ABM defense is too expensive, would violate ABM Treaty with former USSR.
- Soviets developed an ABM system in violation of the treaty.
- In June 96 Russia test-launched an SS-19, their 26th ICBM test in 5 years.
- *"Russia still retains and is modernizing a vast nuclear arsenal... still the only nation of earth that could destroy America in a matter of minutes. Indeed, given the weakness of the Russian Army, Russia's superpower ambitions rest more than ever on its nuclear capabilities."* *
- "For the first time since the dawn of the nuclear age, there is not a single Russian missile pointed at America's children." **

Comment: Punch a computer button and the missiles are pointed! The armed missiles are still there, ready and waiting.

Bulleted info courtesy Washington Inquirer, June 24, 1996.

© 1996 T.E. BEARDEN

* House Nat'l Security Chmn, Floyd Spence.
** Bill Clinton, State of the Union Message.

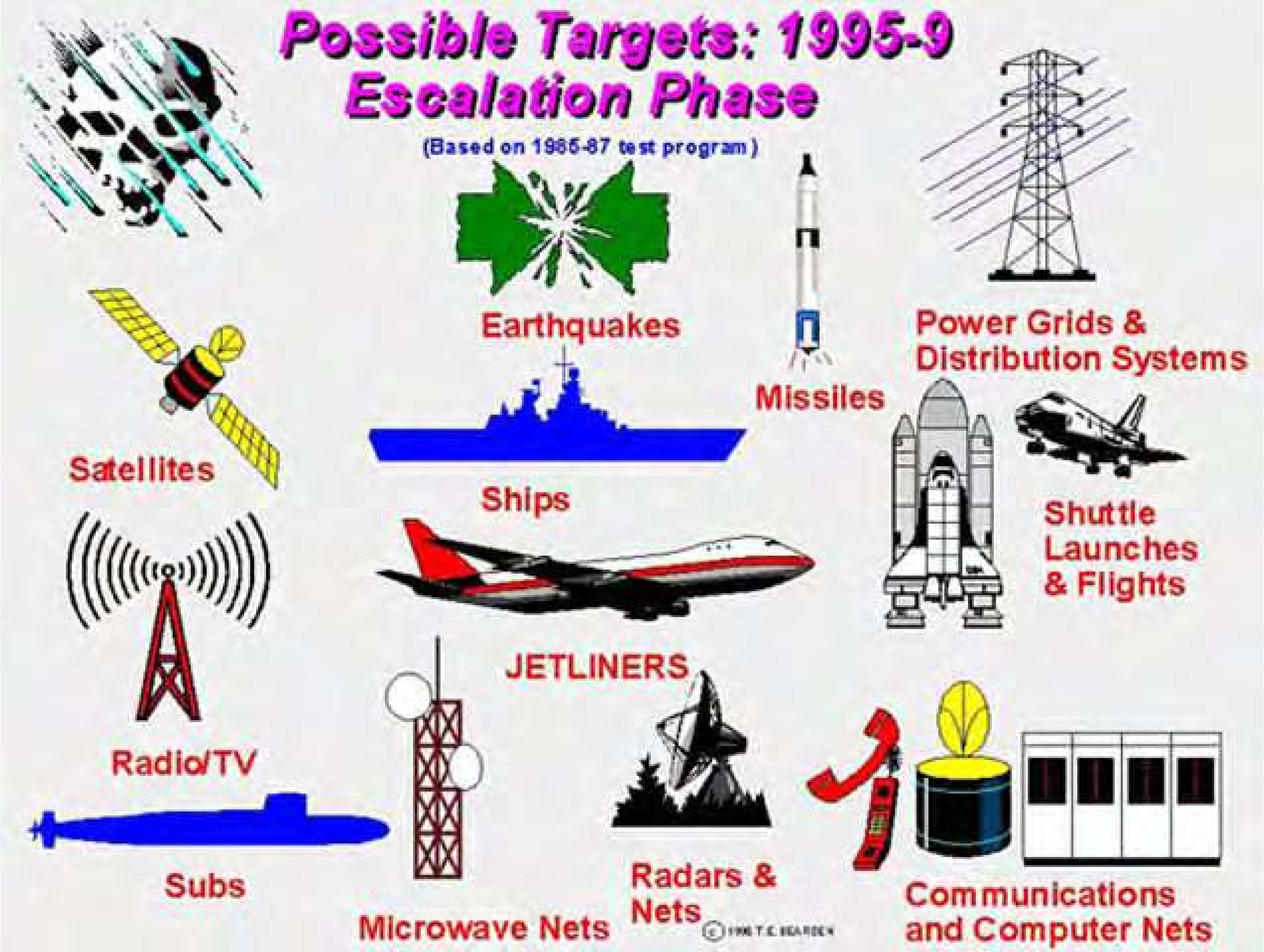
Possible Targets: 1995-9
Escalation Phase
(Based on 1985-87 test program)
Earthquakes
Missiles
Power Grids &
Distribution Systems
Satellites
Ships
Shuttle
Launches
& Flights
JETLINERS
Radio/TV
Subs
Microwave Nets
Radars &
Nets
Communications
and Computer Nets

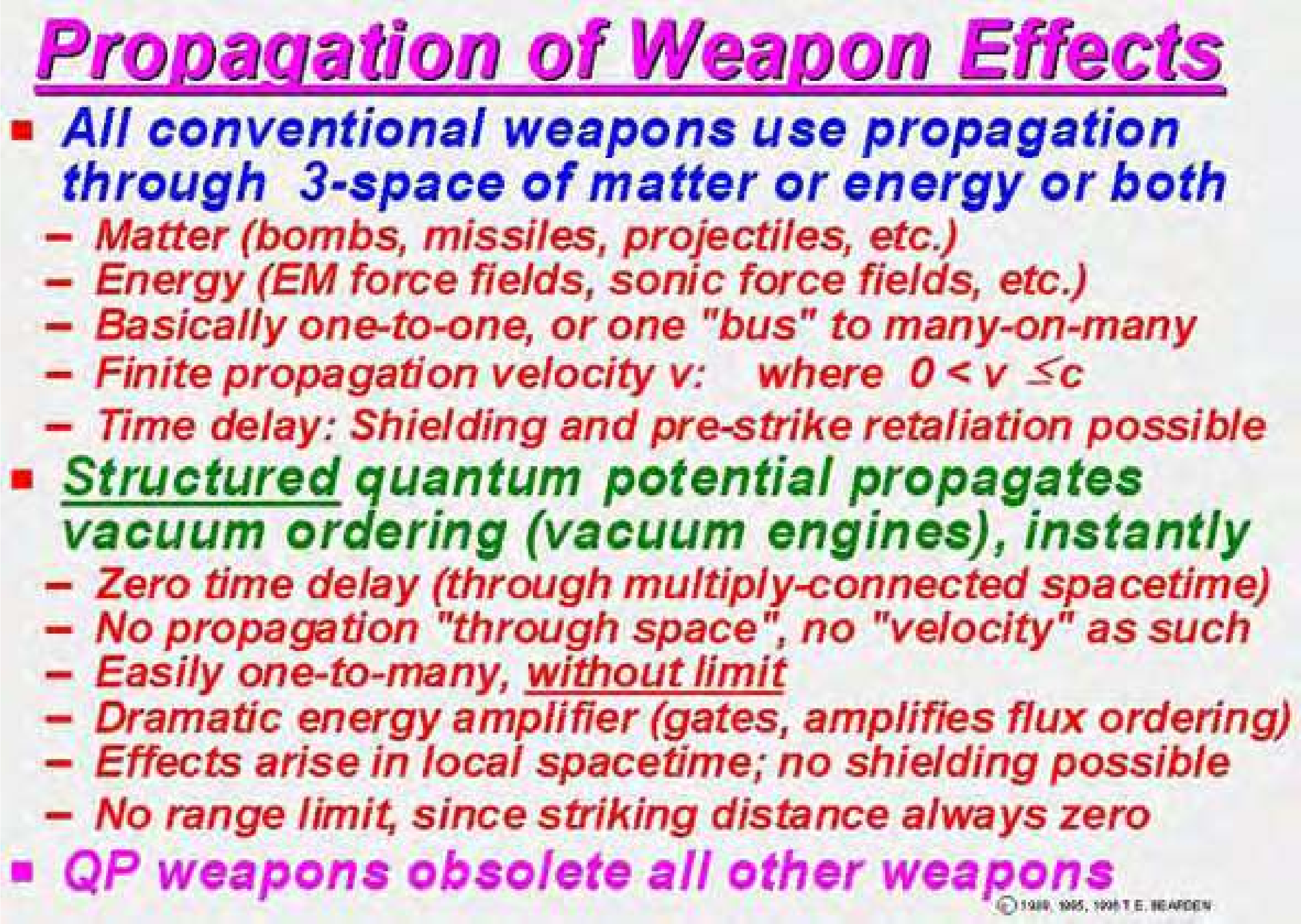

Propagation of Weapon Effects

- All conventional weapons use propagation through 3-space of matter or energy or both
  - Matter (bombs, missiles, projectiles, etc.)
  - Energy (EM force fields, sonic force fields, etc.)
  - Basically one-to-one, or one "bus" to many-on-many
  - Finite propagation velocity v: where $0 < v \leq c$
  - Time delay: Shielding and pre-strike retaliation possible
- Structured quantum potential propagates vacuum ordering (vacuum engines), instantly
  - Zero time delay (through multiply-connected spacetime)
  - No propagation "through space", no "velocity" as such
  - Easily one-to-many, without limit
  - Dramatic energy amplifier (gates, amplifies flux ordering)
  - Effects arise in local spacetime; no shielding possible
  - No range limit, since striking distance always zero
- QP weapons obsolete all other weapons

© 1988, 1995, 1996 T.E. BEARDEN

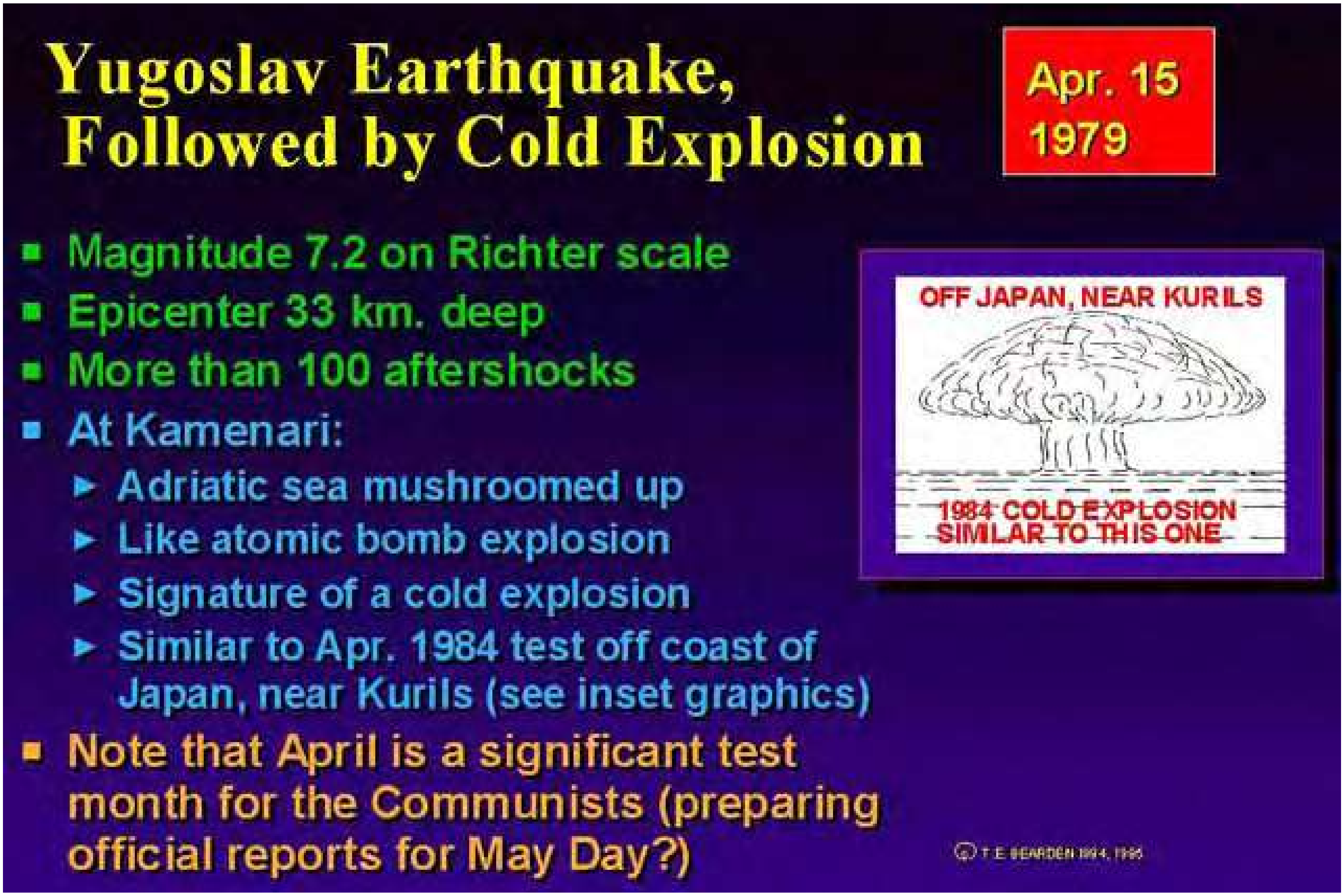

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
    - Adriatic sea mushroomed up
    - Like atomic bomb explosion
    - Signature of a cold explosion
    - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- Note that April is a significant test month for the Communists (preparing official reports for May Day?)

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

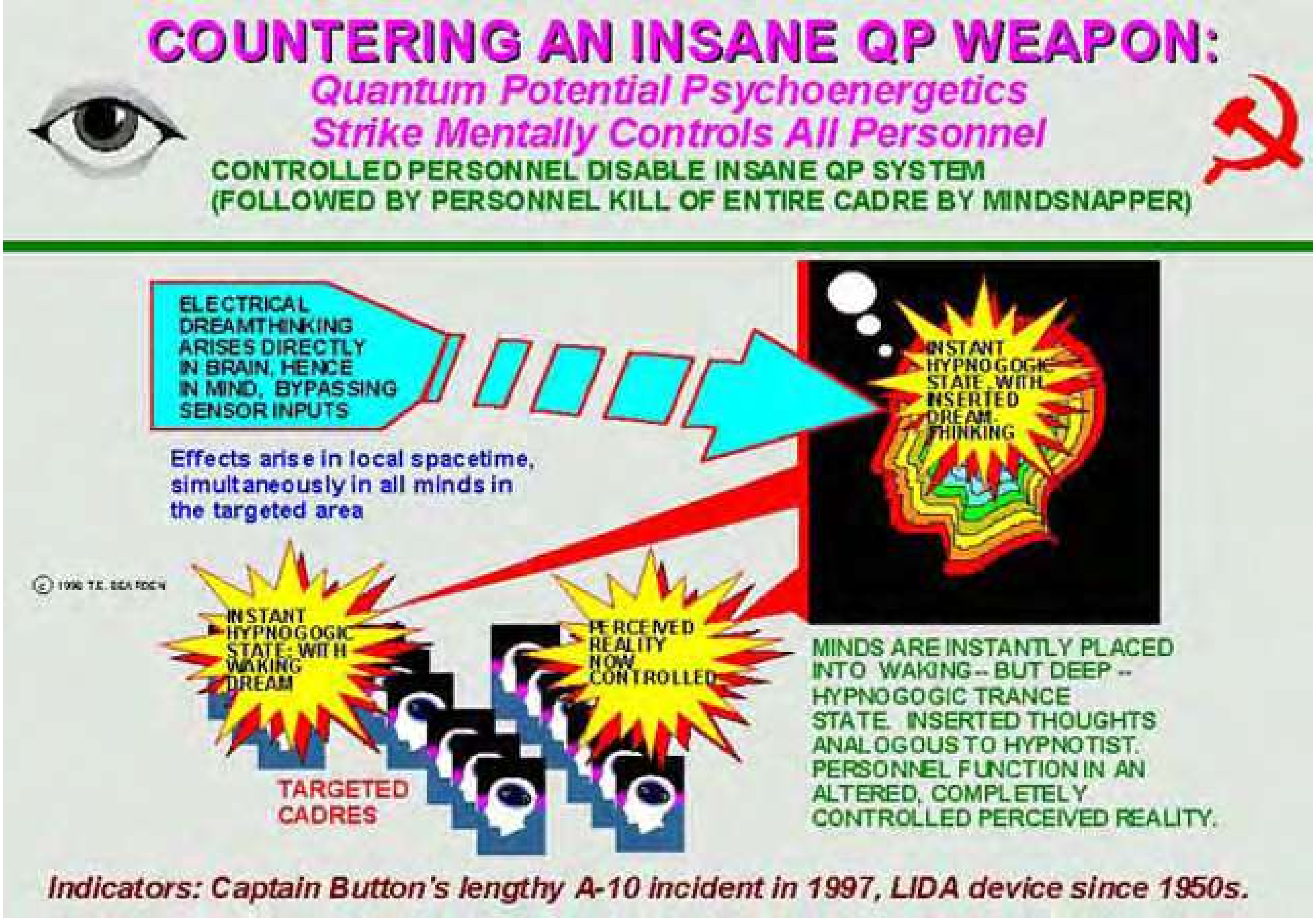

COUNTERING AN INSANE QP WEAPON:

Quantum Potential Psychoenergetics Strike Mentally Controls All Personnel

CONTROLLED PERSONNEL DISABLE INSANE QP SYSTEM
(FOLLOWED BY PERSONNEL KILL OF ENTIRE CADRE BY MINDSNAPPER)

ELECTRICAL DREAMTHINKING ARISES DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED DREAM-THINKING

© [illegible] T.E. BEARDEN

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED CADRES

MINDS ARE INSTANTLY PLACED INTO WAKING -- BUT DEEP -- HYPNOGOGIC TRANCE STATE. INSERTED THOUGHTS ANALOGOUS TO HYPNOTIST. PERSONNEL FUNCTION IN AN ALTERED, COMPLETELY CONTROLLED PERCEIVED REALITY.

Indicators: Captain Button's lengthy A-10 incident in 1997, LIDA device since 1950s.

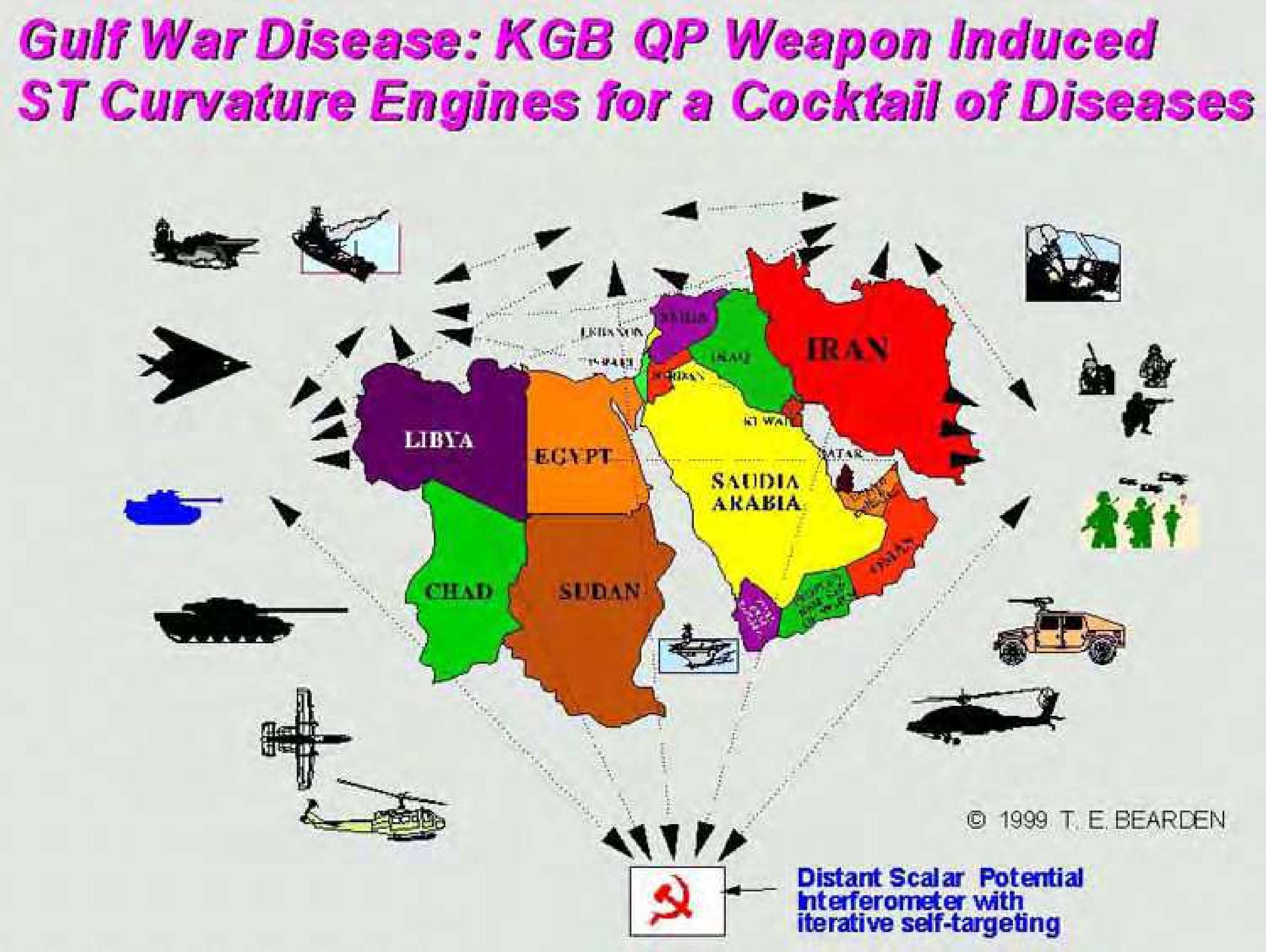

Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases

LEBANON
SYRIA
ISRAEL
IRAQ
JORDAN
IRAN
KUWAIT
QATAR
LIBYA
EGYPT
SAUDIA ARABIA
OMAN
CHAD
SUDAN

© 1999 T. E. BEARDEN

Distant Scalar Potential Interferometer with iterative self-targeting

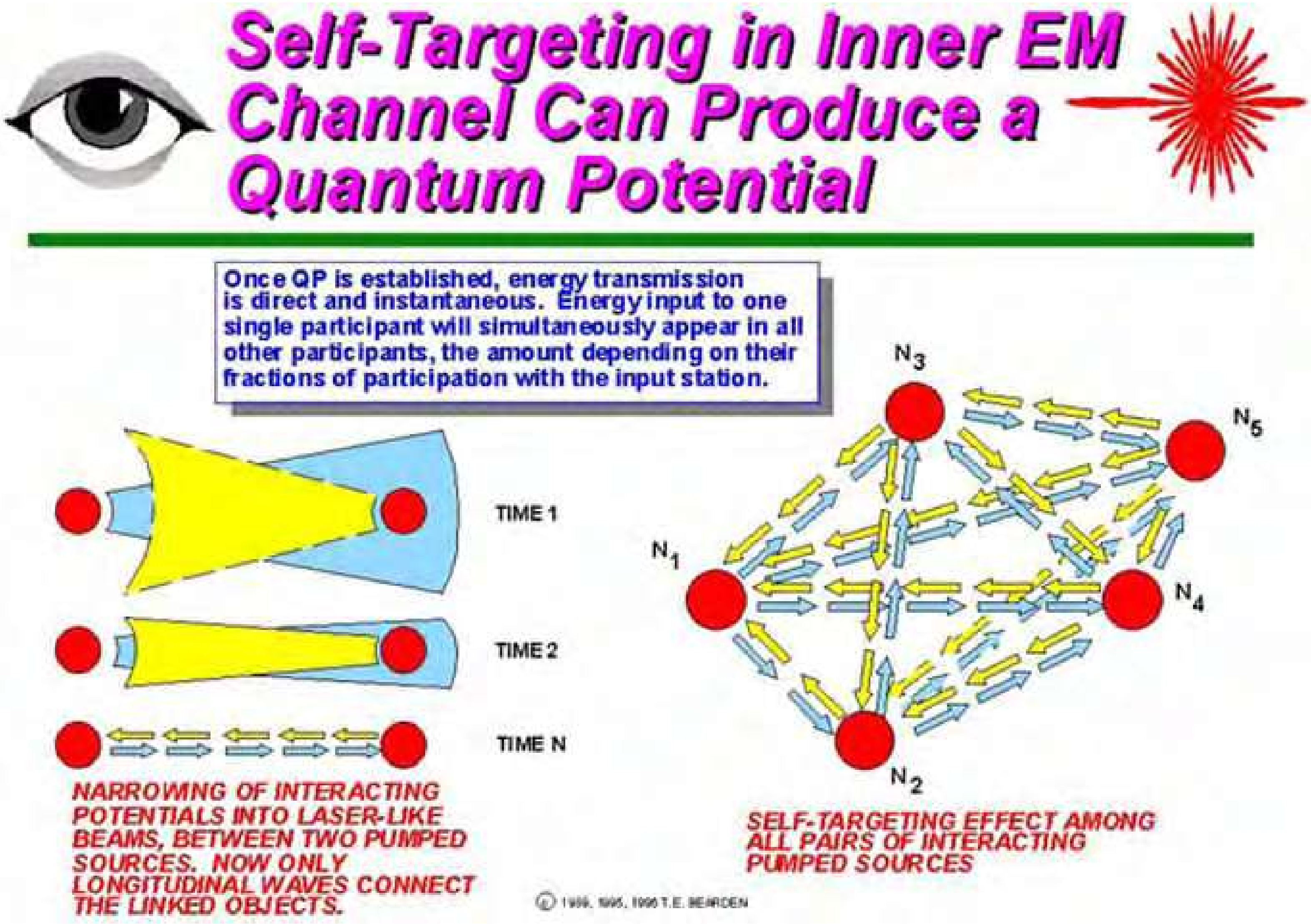

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1989, 1995, 1996 T.E. BEARDEN

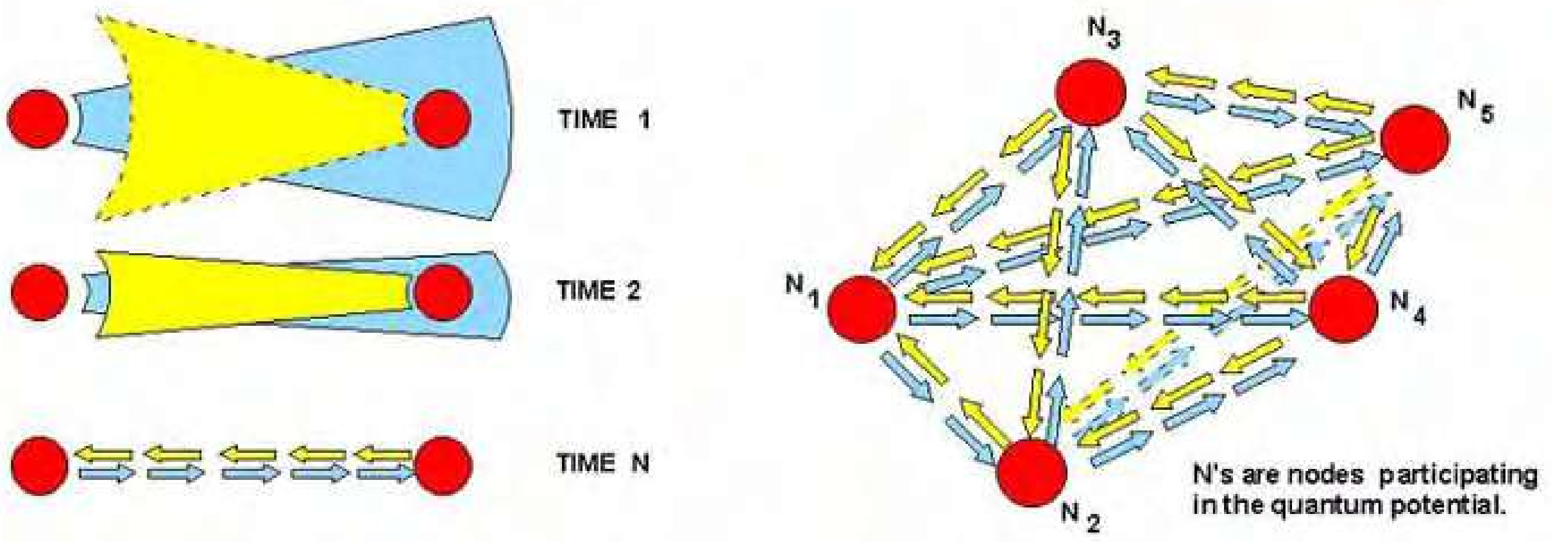

a. Iterative retroreflection between any pair of nodes narrows the interconnecting potentials to a laser-like, bidirectional beam. Iterative phase conjugate reflection provides self-targeting.

b. The quantum potential exists as narrow, laser-like beams between participating nodes, even when the nodes are widely separated. The beam space is multiply-connected; transmission is instantaneous.

© 1989, 1995 T.E. BEARDEN

Consider the hidden Stoney/Whittaker bidirectional EM waves that comprise the potential from each source node. If these nodes have sufficient intersymmetry of dynamic material form, then self-targeting occurs between the two nodes of any nodal pair. By iterative phase conjugate reflection, the scattering is narrowed into interconnecting wavepairs, in laser-like beams, between each two nodes. Thus the interconnecting potentials narrow into laser-like, bidirectional beams. The Coulomb gauge may apply, and a quantum potential may be formed.

Figure 30. Mechanism for generating a quantum potential, by iterative self-targeting induced into a system of multiple scattering nodes. Once a QP is established, spatial separation has little or no effect.

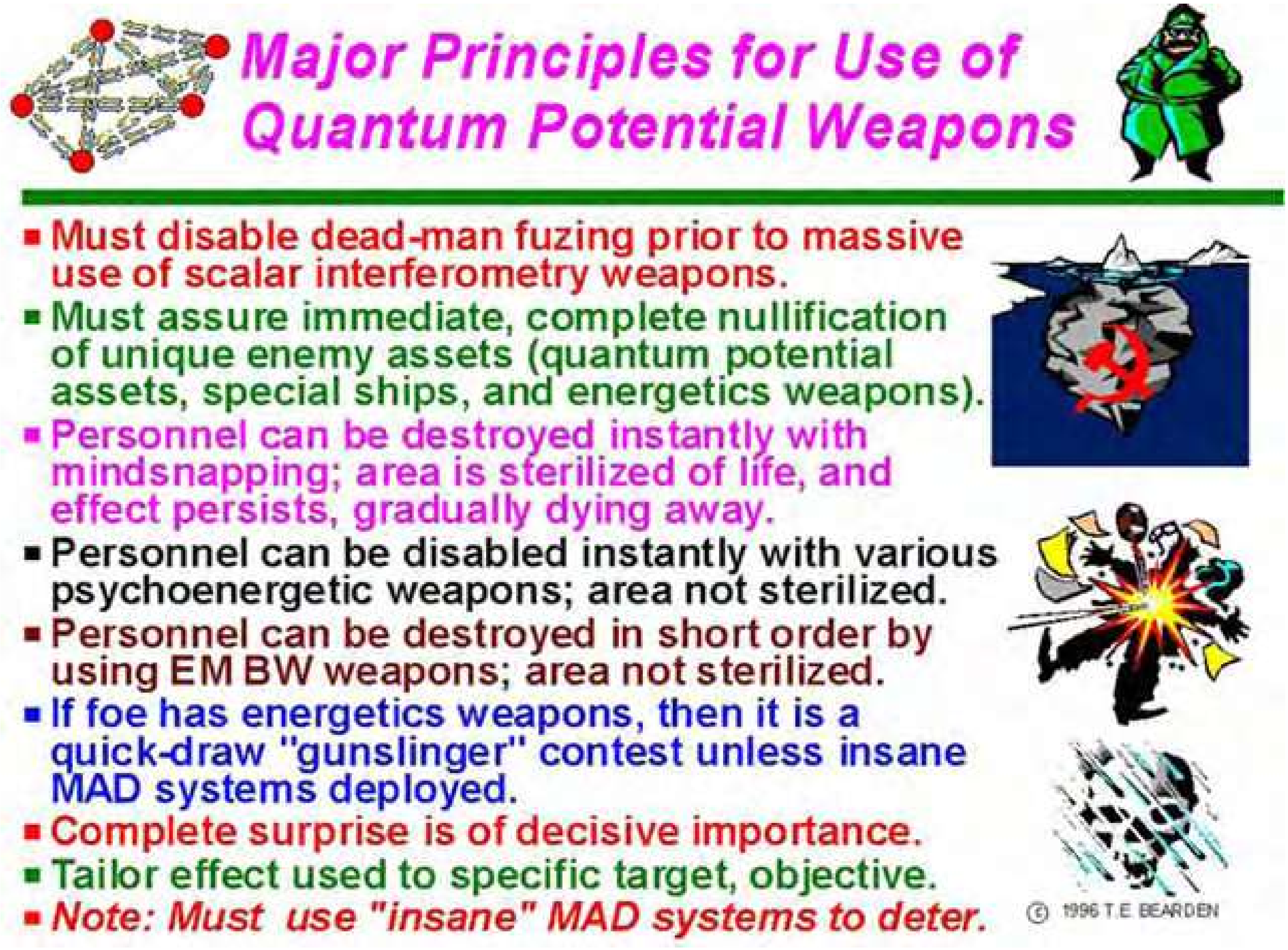

Major Principles for Use of Quantum Potential Weapons

- Must disable dead-man fuzing prior to massive use of scalar interferometry weapons.
- Must assure immediate, complete nullification of unique enemy assets (quantum potential assets, special ships, and energetics weapons).
- Personnel can be destroyed instantly with mindsnapping; area is sterilized of life, and effect persists, gradually dying away.
- Personnel can be disabled instantly with various psychoenergetic weapons; area not sterilized.
- Personnel can be destroyed in short order by using EM BW weapons; area not sterilized.
- If foe has energetics weapons, then it is a quick-draw "gunslinger" contest unless insane MAD systems deployed.
- Complete surprise is of decisive importance.
- Tailor effect used to specific target, objective.
- *Note: Must use "insane" MAD systems to deter.*

© 1996 T.E. BEARDEN

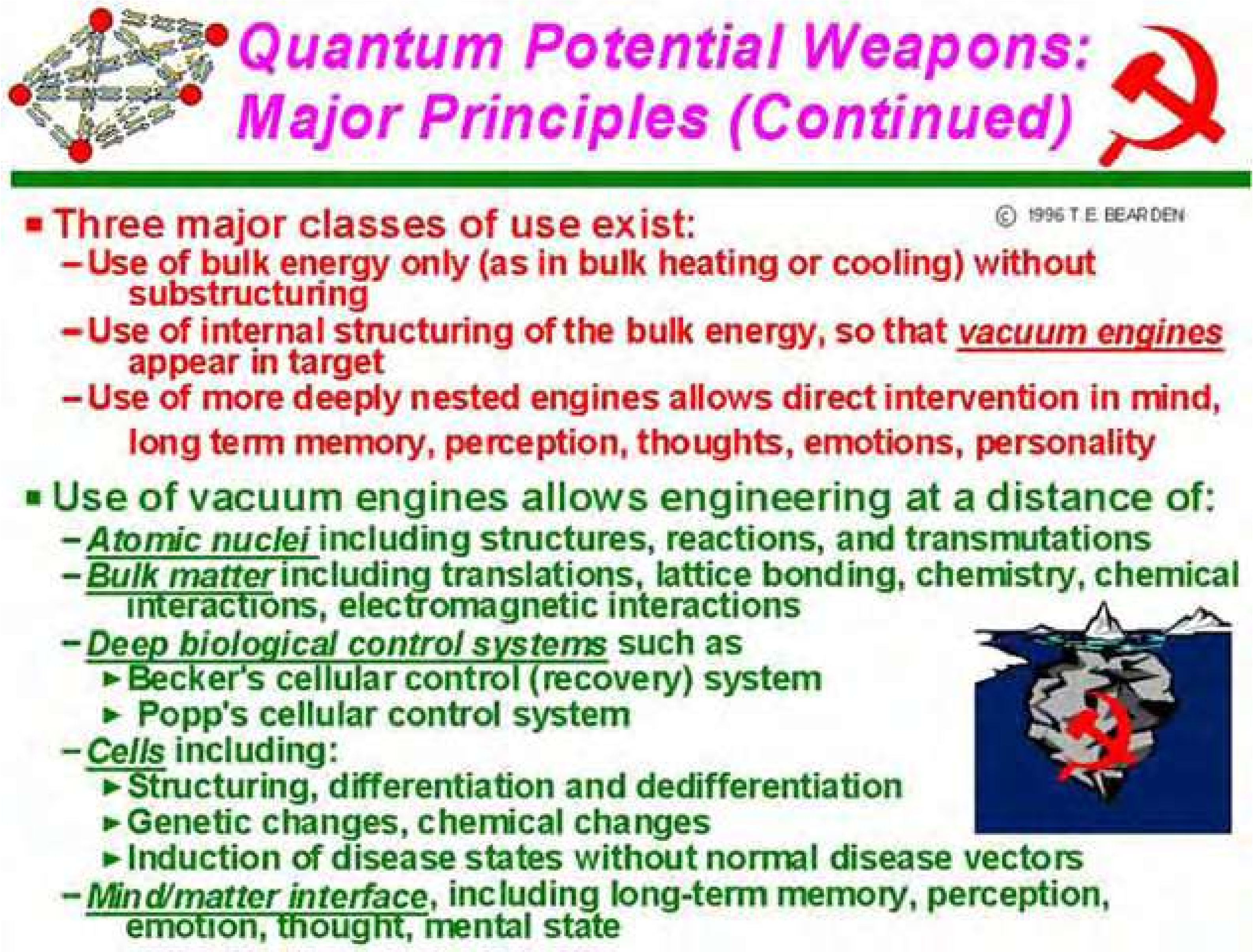

Quantum Potential Weapons: Major Principles (Continued)

© 1996 T. E. BEARDEN

- Three major classes of use exist:
  - –Use of bulk energy only (as in bulk heating or cooling) without substructuring
  - –Use of internal structuring of the bulk energy, so that *vacuum engines* appear in target
  - –Use of more deeply nested engines allows direct intervention in mind, long term memory, perception, thoughts, emotions, personality
- Use of vacuum engines allows engineering at a distance of:
  - –*Atomic nuclei* including structures, reactions, and transmutations
  - –*Bulk matter* including translations, lattice bonding, chemistry, chemical interactions, electromagnetic interactions
  - –*Deep biological control systems* such as
    - ► Becker's cellular control (recovery) system
    - ► Popp's cellular control system
  - –*Cells* including:
    - ► Structuring, differentiation and dedifferentiation
    - ► Genetic changes, chemical changes
    - ► Induction of disease states without normal disease vectors
  - –*Mind/matter interface*, including long-term memory, perception, emotion, thought, mental state

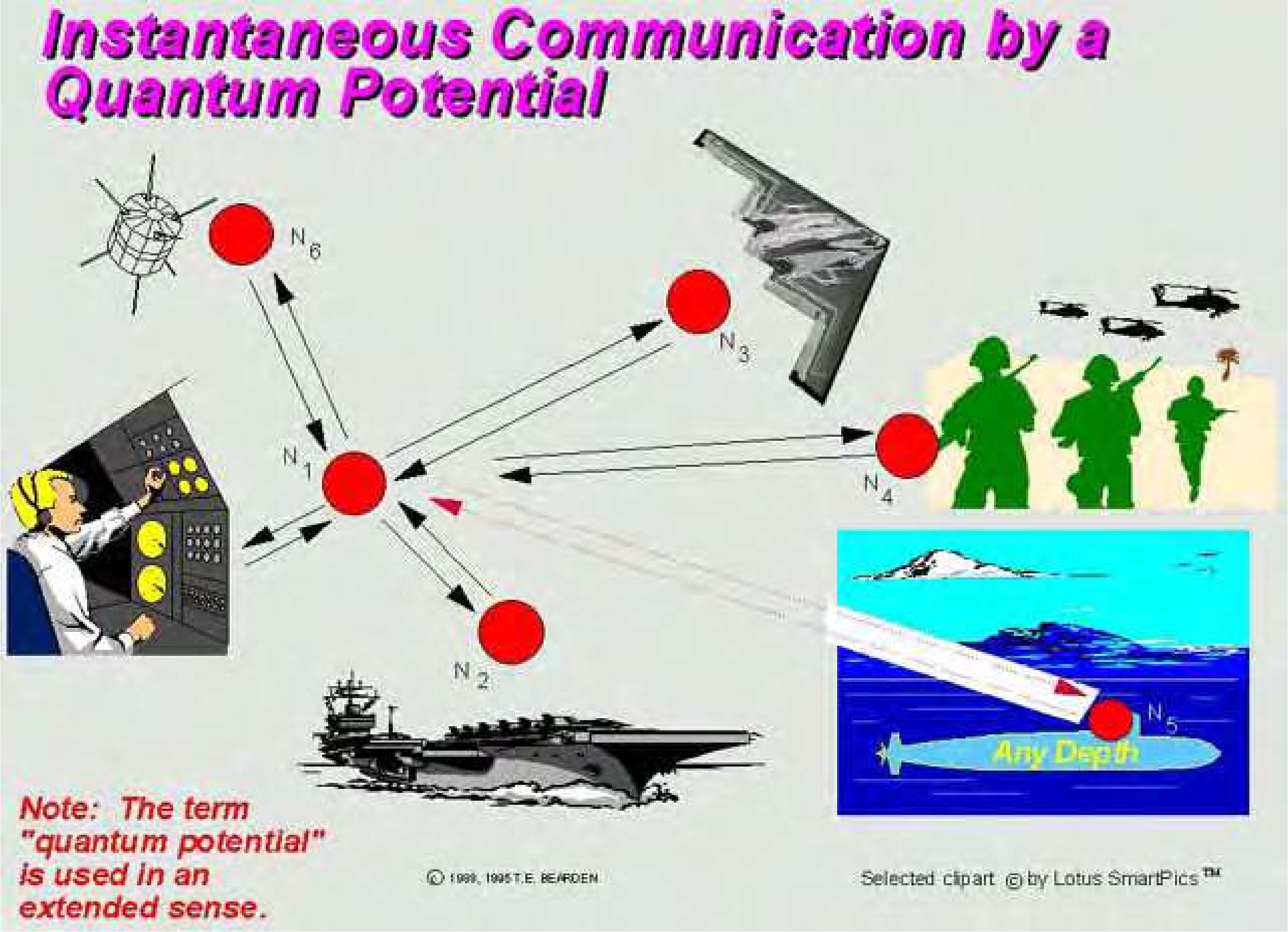

Instantaneous Communication by a Quantum Potential

Note: The term "quantum potential" is used in an extended sense.

© 1989, 1995 T.E. BEARDEN

Selected clipart © by Lotus SmartPics™

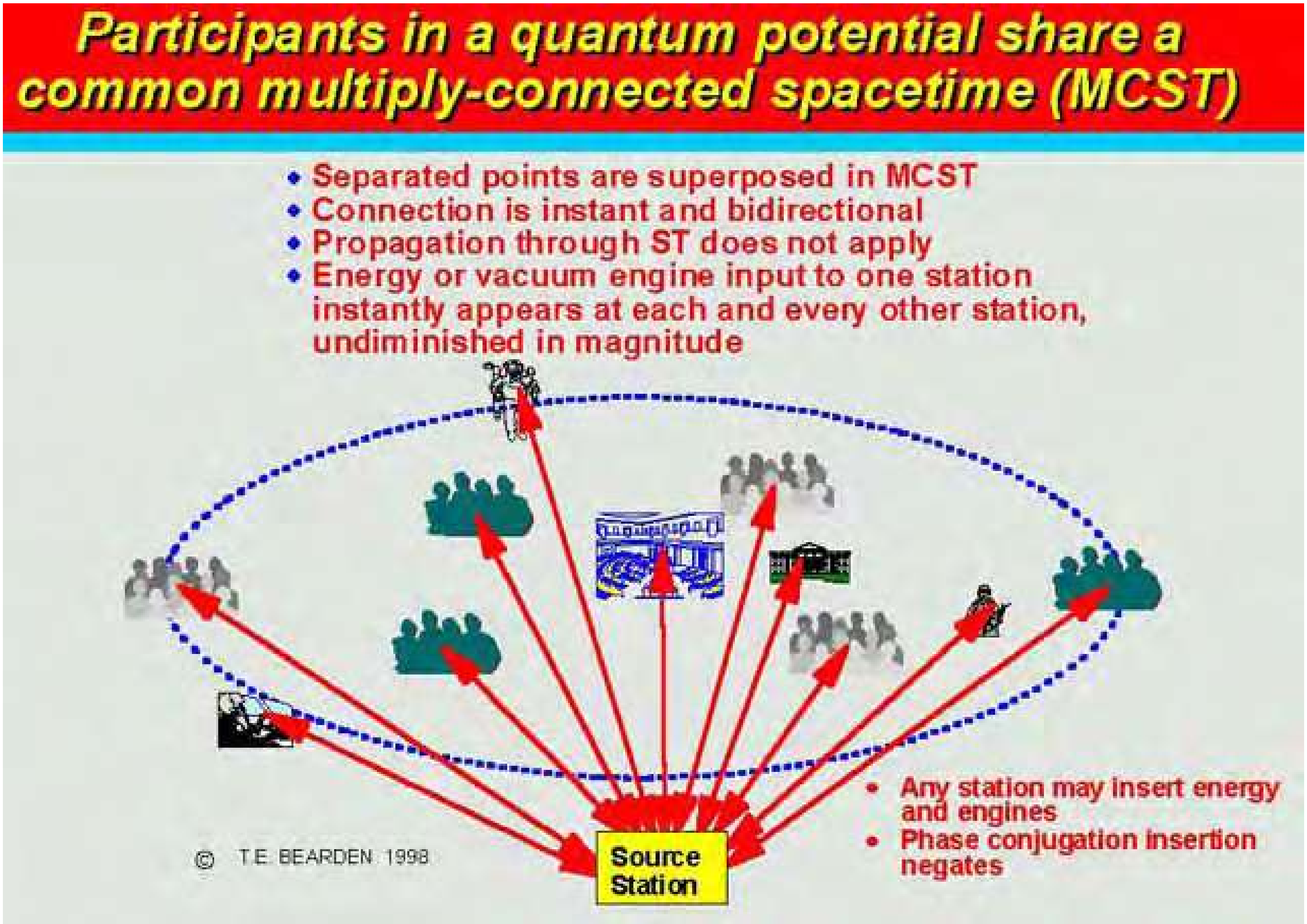

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude

- Any station may insert energy and engines
- Phase conjugation insertion negates

© T.E. BEARDEN 1998

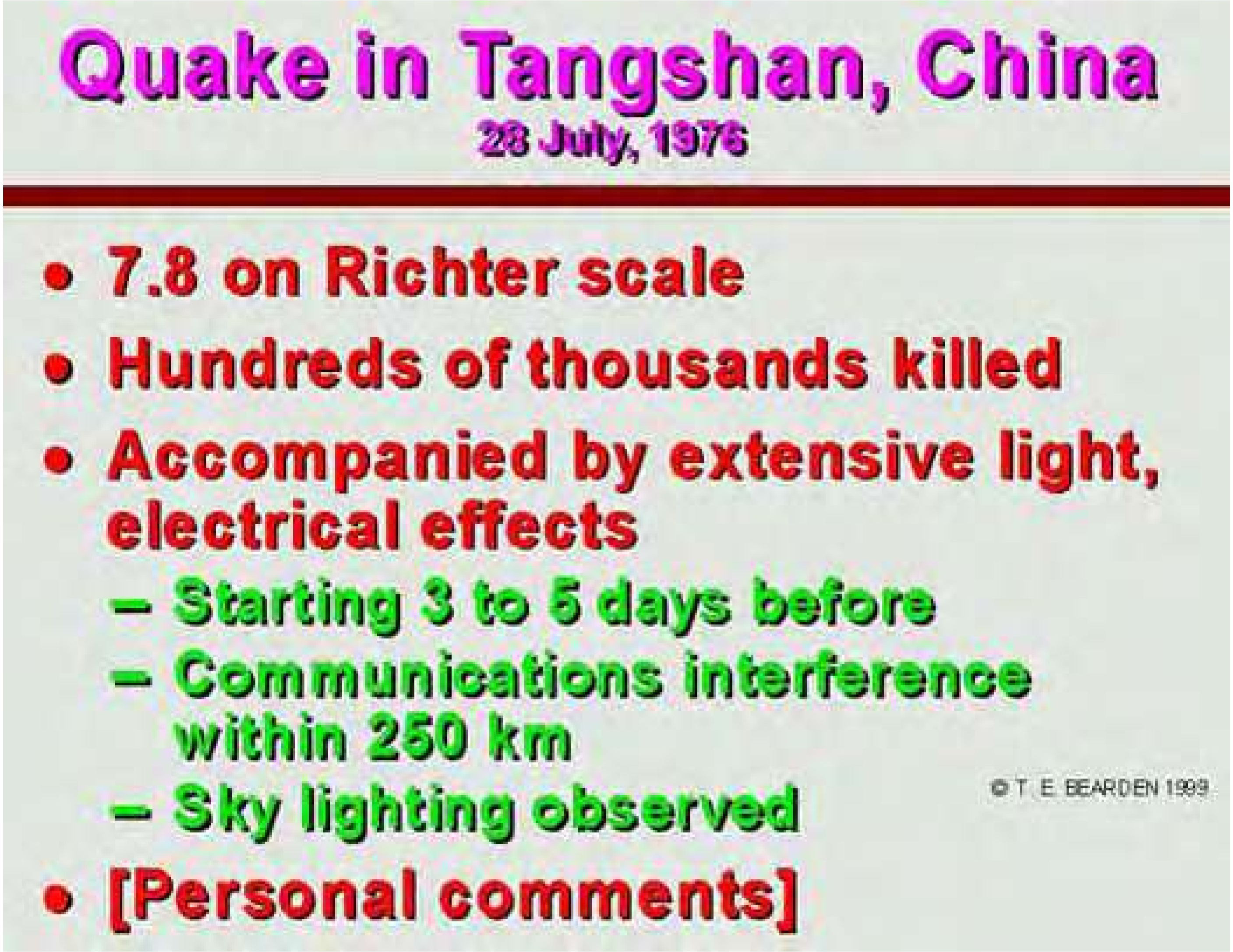

Quake in Tangshan, China

28 July, 1976

- 7.8 on Richter scale
- Hundreds of thousands killed
- Accompanied by extensive light, electrical effects
  - Starting 3 to 5 days before
  - Communications interference within 250 km
  - Sky lighting observed
- [Personal comments]

© T. E. BEARDEN 1999

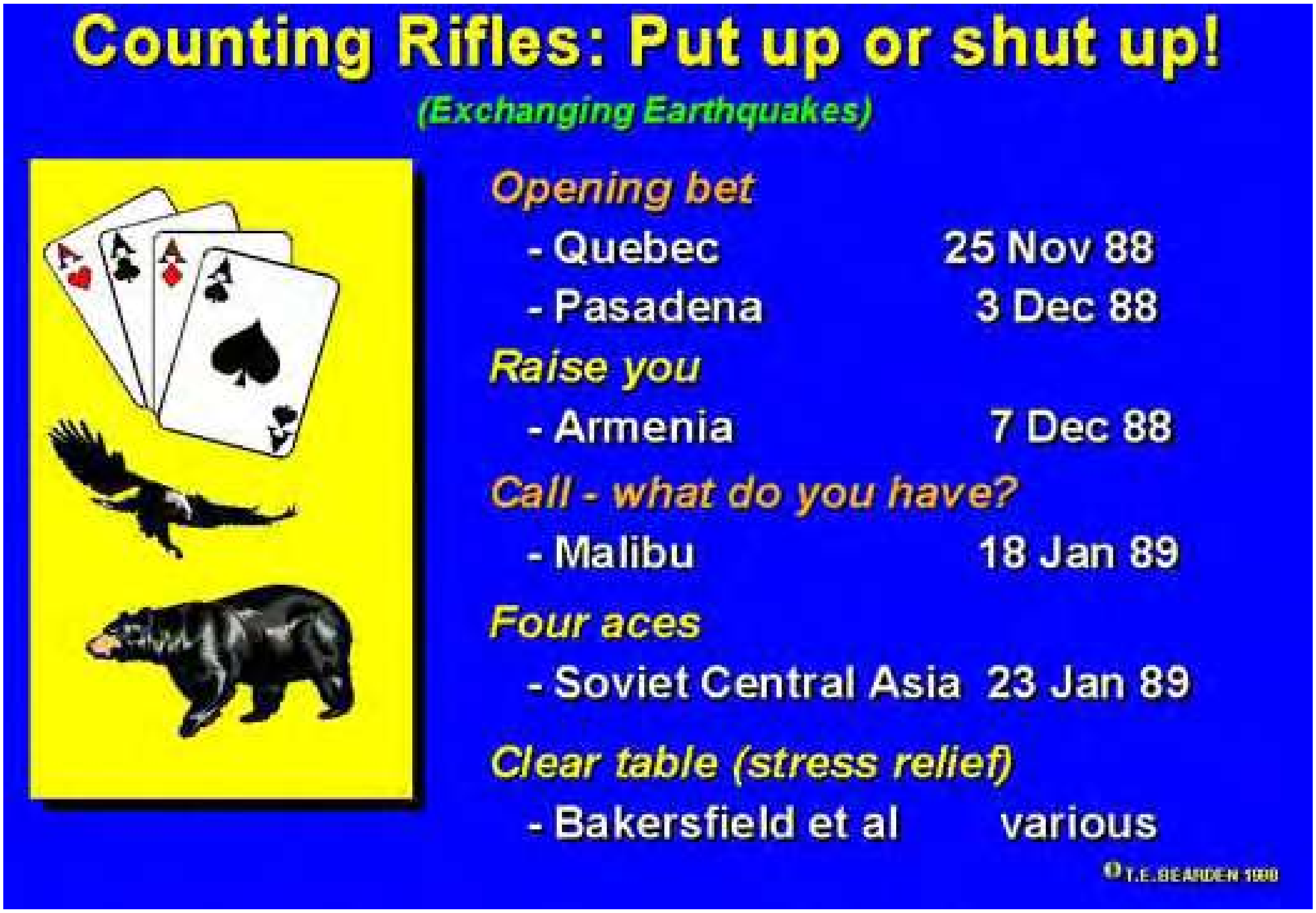

Counting Rifles: Put up or shut up!

(Exchanging Earthquakes)

Opening bet	
- Quebec	25 Nov 88
- Pasadena	3 Dec 88
Raise you	
- Armenia	7 Dec 88
Call - what do you have?	
- Malibu	18 Jan 89
Four aces	
- Soviet Central Asia	23 Jan 89
Clear table (stress relief)	
- Bakersfield et al	various

© T.E.BEARDEN 1990

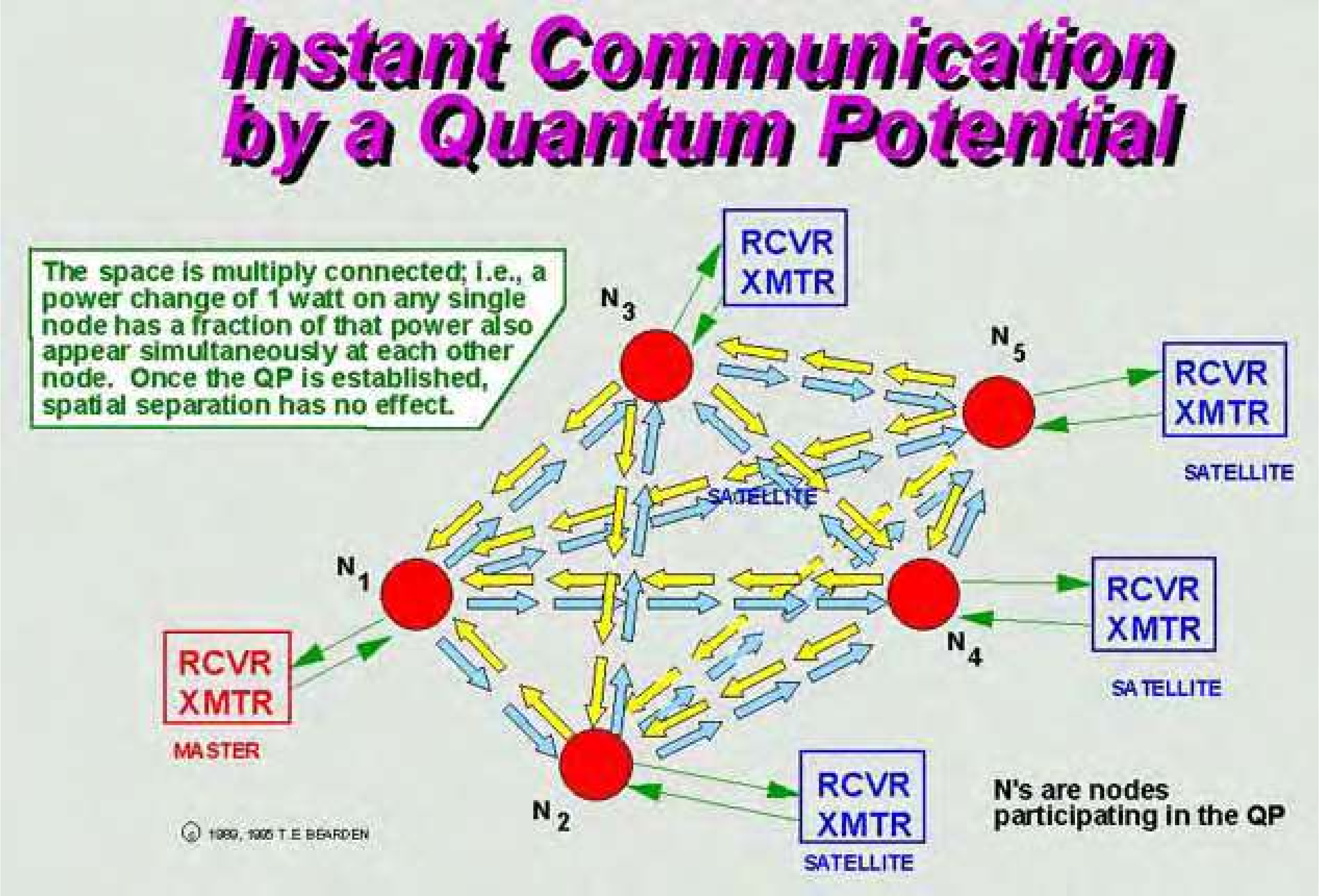

Instant Communication by a Quantum Potential

The space is multiply connected; i.e., a power change of 1 watt on any single node has a fraction of that power also appear simultaneously at each other node. Once the QP is established, spatial separation has no effect.

N's are nodes participating in the QP

N 1
N 2
N 3
N 4
N 5
RCVR XMTR
MASTER
RCVR XMTR
SATELLITE
RCVR XMTR
SATELLITE
RCVR XMTR
SATELLITE
RCVR XMTR
SATELLITE

© 1989, 1995 T.E BEARDEN

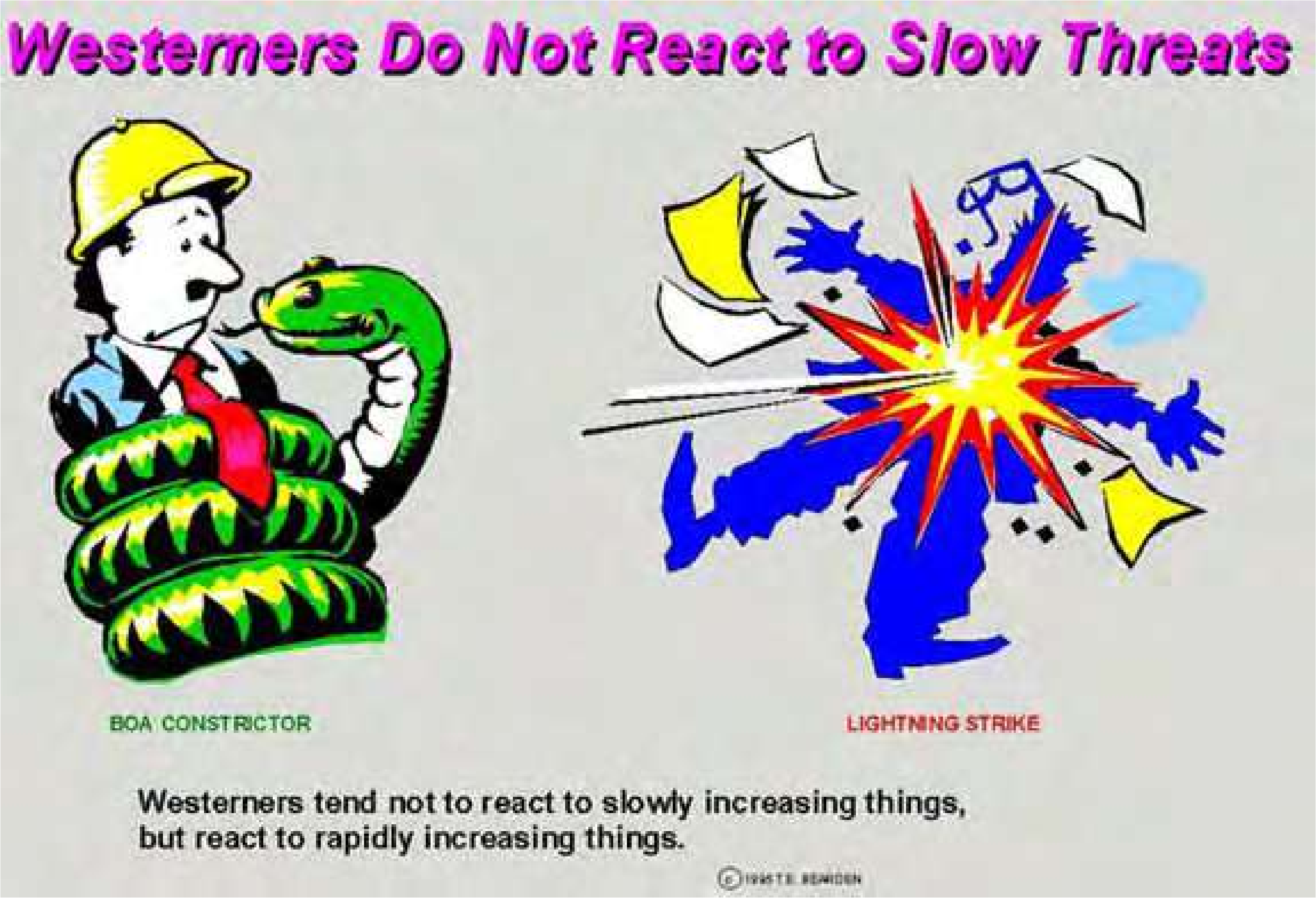

Westerners Do Not React to Slow Threats

BOA CONSTRICTOR
LIGHTNING STRIKE

Westerners tend not to react to slowly increasing things,
but react to rapidly increasing things.

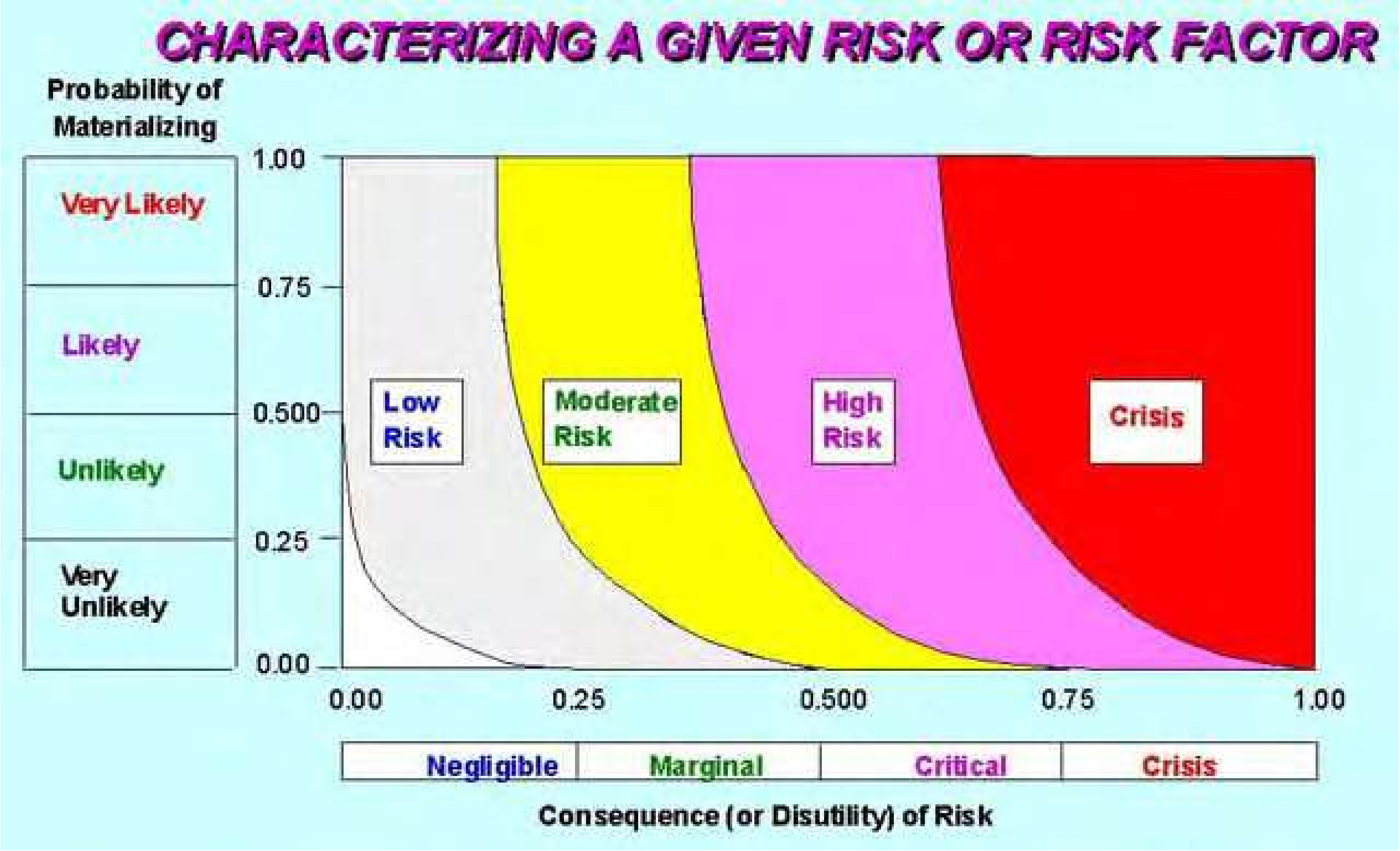
CHARACTERIZING A GIVEN RISK OR RISK FACTOR
Probability of Materializing
Very Likely
Likely
Unlikely
Very Unlikely
1.00
0.75
0.500
0.25
0.00
Low Risk
Moderate Risk
High Risk
Crisis
0.00
0.25
0.500
0.75
1.00
Negligible
Marginal
Critical
Crisis
Consequence (or Disutility) of Risk

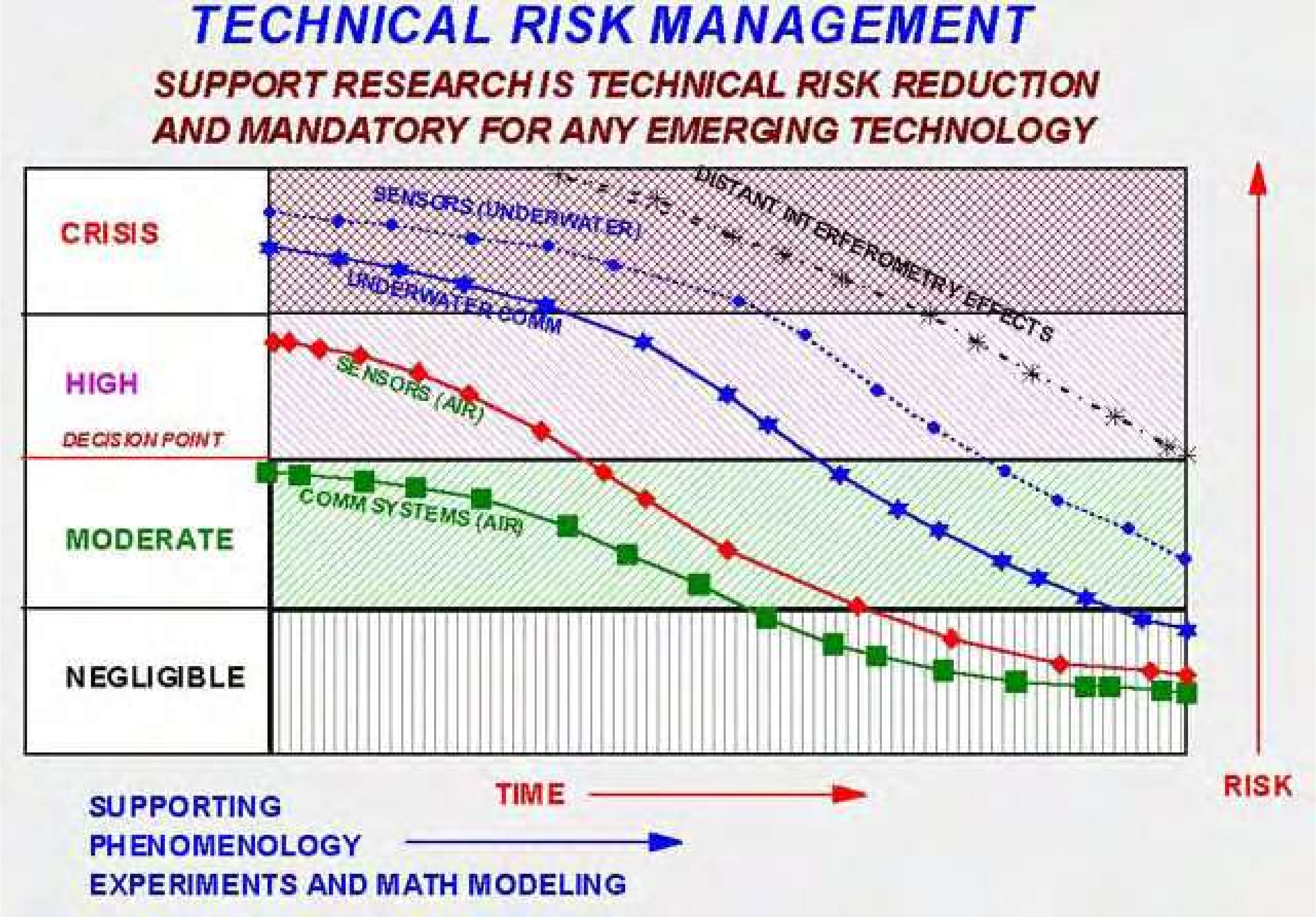

TECHNICAL RISK MANAGEMENT

SUPPORT RESEARCH IS TECHNICAL RISK REDUCTION AND MANDATORY FOR ANY EMERGING TECHNOLOGY

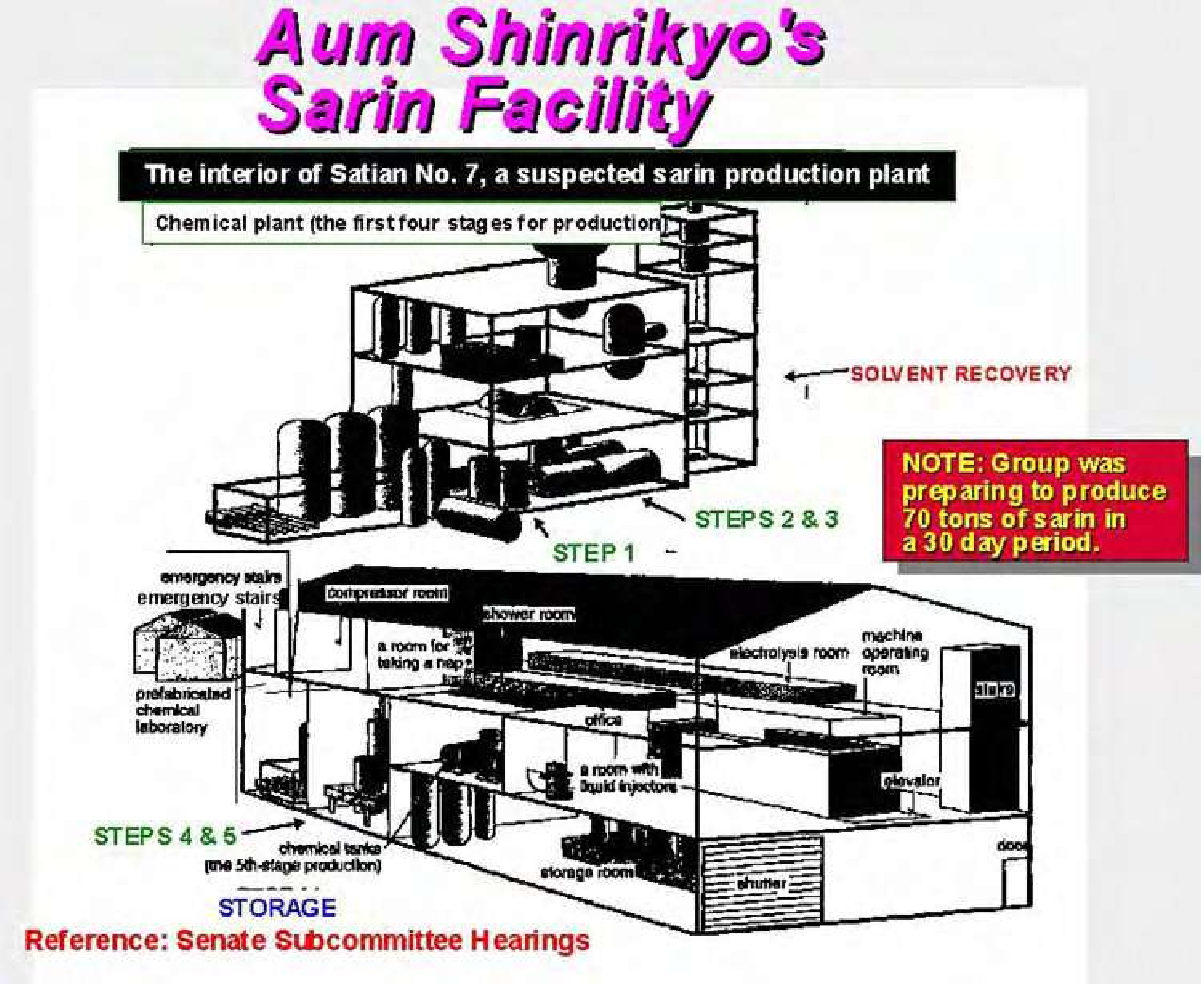

Aum Shinrikyo's Sarin Facility

The interior of Satian No. 7, a suspected sarin production plant

Chemical plant (the first four stages for production)

SOLVENT RECOVERY
STEPS 2 & 3
STEP 1
emergency stairs
emergency stairs
compressor room
shower room
a room for taking a nap
electrolysis room
machine operating room
stairs
prefabricated chemical laboratory
office
a room with liquid injectors
elevator
STEPS 4 & 5
chemical tanks (the 5th-stage production)
storage room
shutter
doo
STORAGE

NOTE: Group was preparing to produce 70 tons of sarin in a 30 day period.

Reference: Senate Subcommittee Hearings

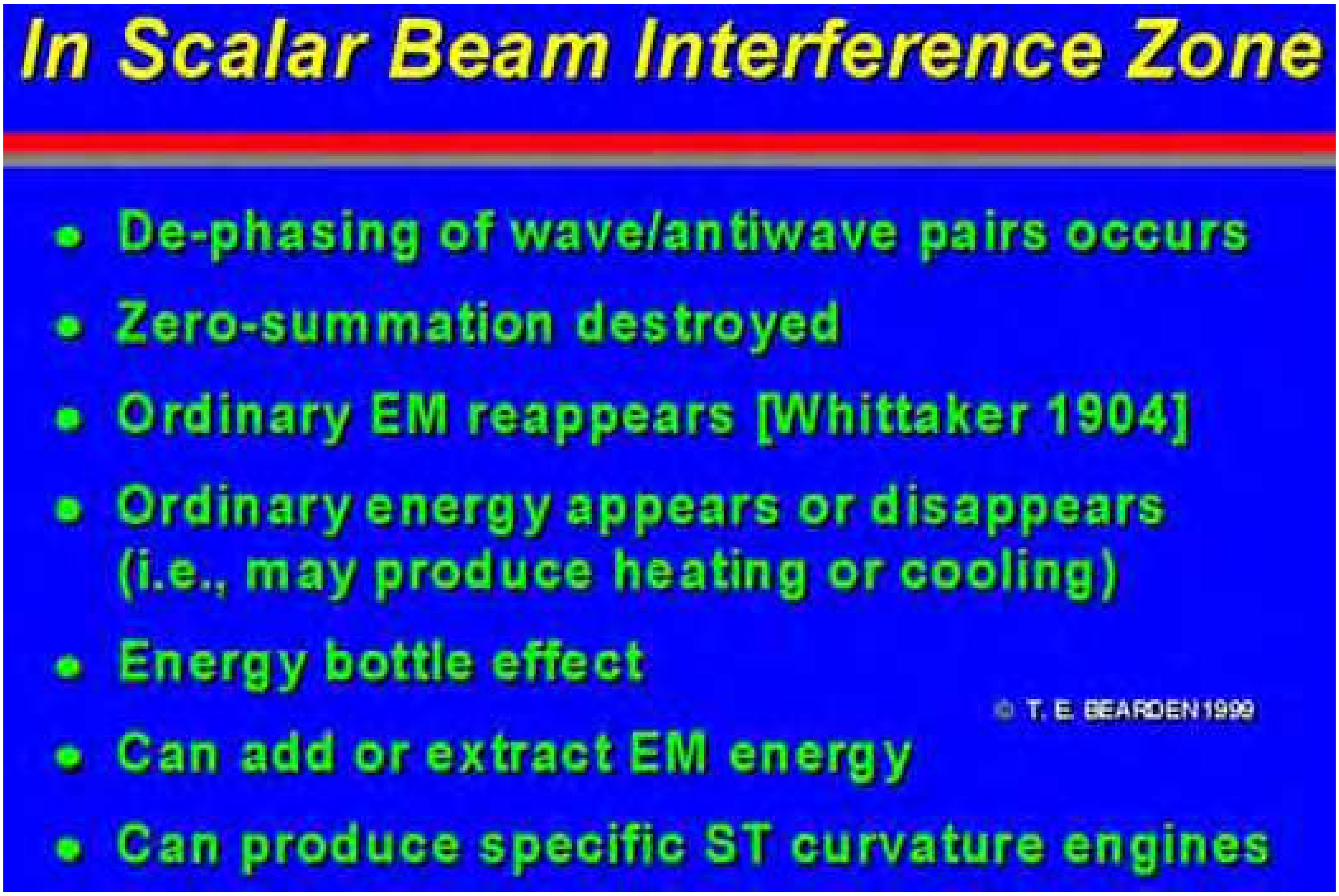

In Scalar Beam Interference Zone

- De-phasing of wave/antiwave pairs occurs
- Zero-summation destroyed
- Ordinary EM reappears [Whittaker 1904]
- Ordinary energy appears or disappears (i.e., may produce heating or cooling)
- Energy bottle effect
- Can add or extract EM energy
- Can produce specific ST curvature engines

© T. E. BEARDEN 1999

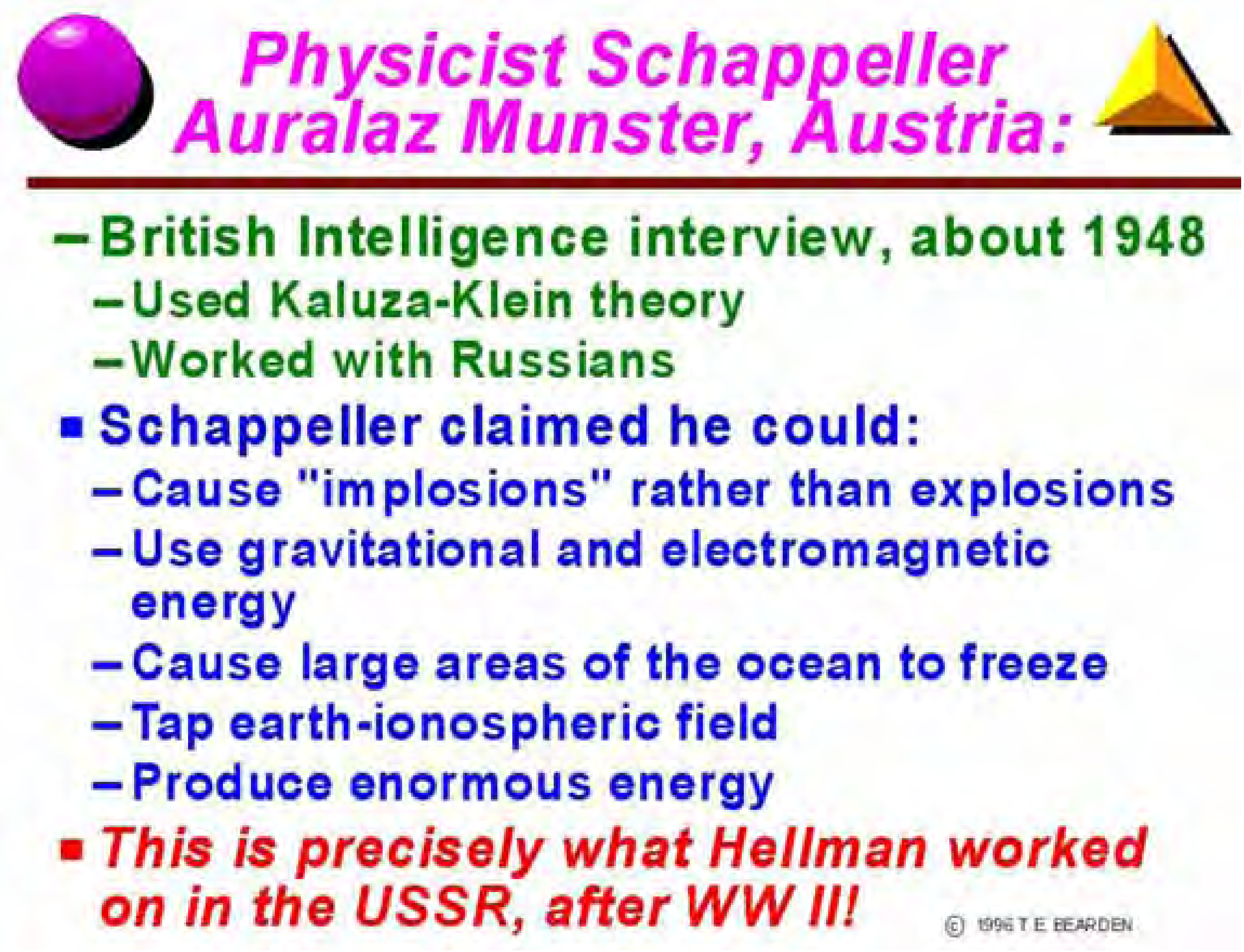

Physicist Schappeller Auralaz Munster, Austria:

- British Intelligence interview, about 1948
  - Used Kaluza-Klein theory
  - Worked with Russians
- Schappeller claimed he could:
  - Cause "implosions" rather than explosions
  - Use gravitational and electromagnetic energy
  - Cause large areas of the ocean to freeze
  - Tap earth-ionospheric field
  - Produce enormous energy
- *This is precisely what Hellman worked on in the USSR, after WW II!*

© 1996 T E BEARDEN

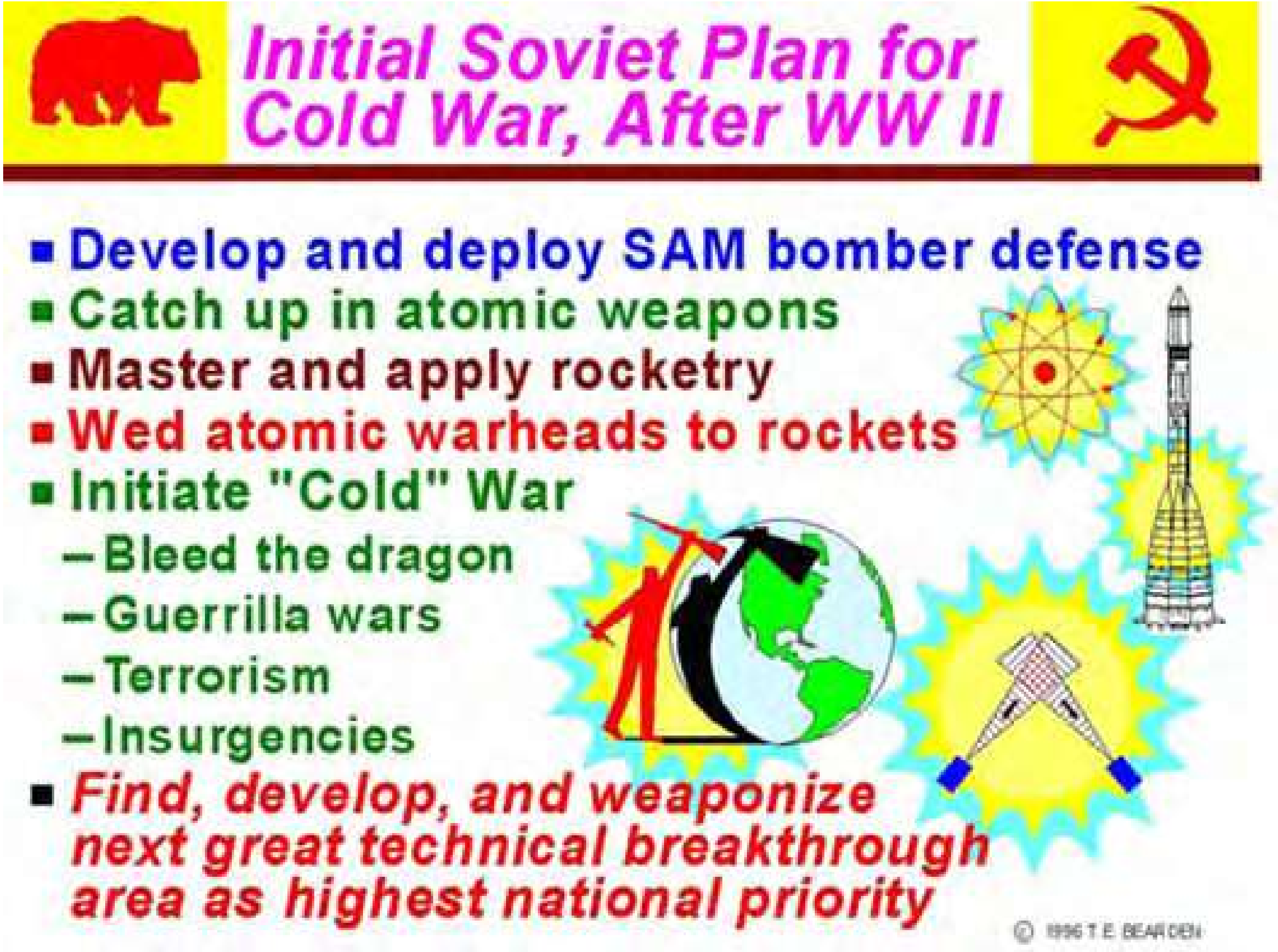

Initial Soviet Plan for Cold War, After WW II

- Develop and deploy SAM bomber defense
- Catch up in atomic weapons
- Master and apply rocketry
- Wed atomic warheads to rockets
- Initiate "Cold" War
  - Bleed the dragon
  - Guerrilla wars
  - Terrorism
  - Insurgencies
- *Find, develop, and weaponize next great technical breakthrough area as highest national priority*

© 1996 T.E. BEARDEN

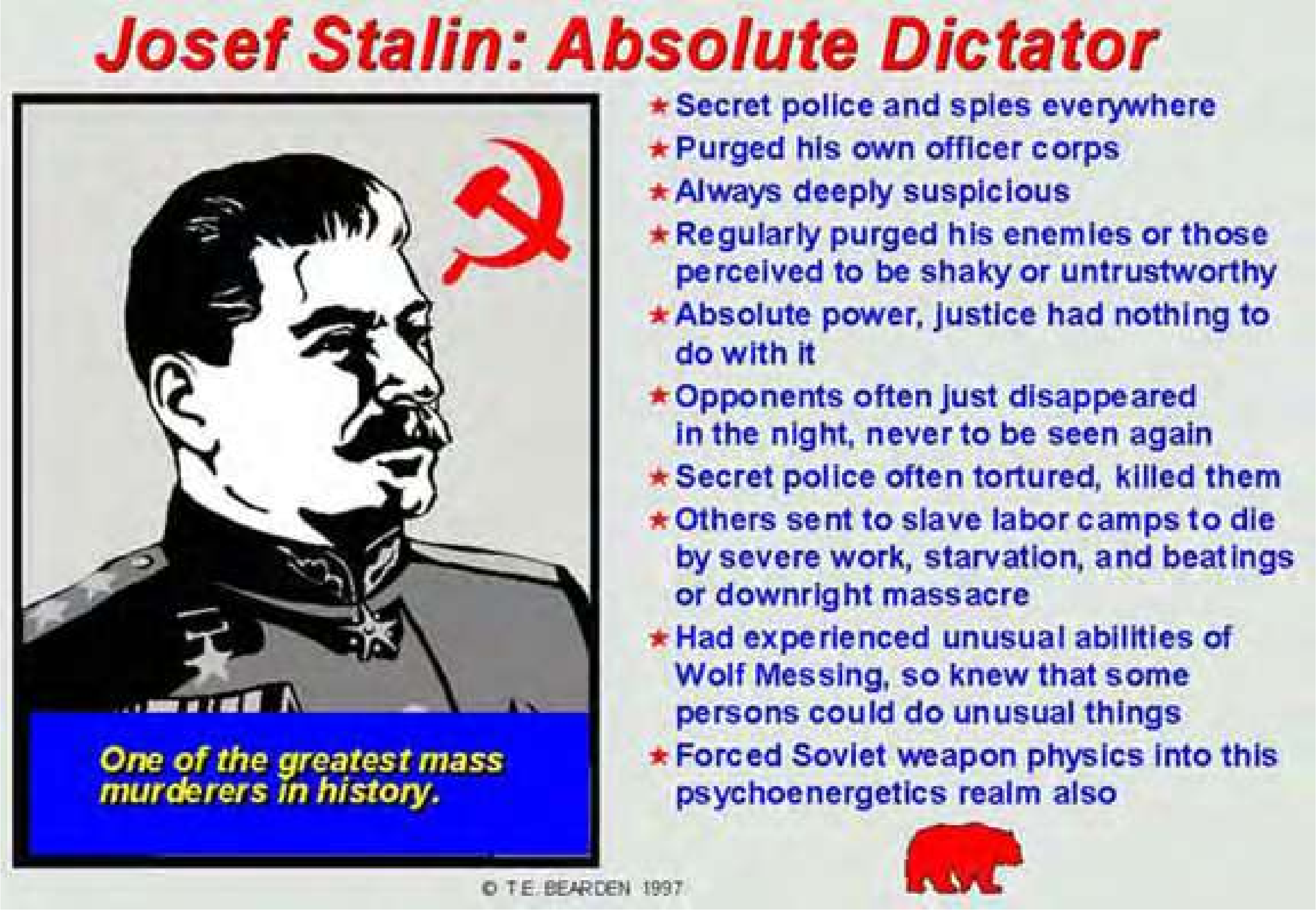

Josef Stalin: Absolute Dictator

- Secret police and spies everywhere
- Purged his own officer corps
- Always deeply suspicious
- Regularly purged his enemies or those perceived to be shaky or untrustworthy
- Absolute power, justice had nothing to do with it
- Opponents often just disappeared in the night, never to be seen again
- Secret police often tortured, killed them
- Others sent to slave labor camps to die by severe work, starvation, and beatings or downright massacre
- Had experienced unusual abilities of Wolf Messing, so knew that some persons could do unusual things
- Forced Soviet weapon physics into this psychoenergetics realm also

One of the greatest mass murderers in history.

© T.E. BEARDEN 1997

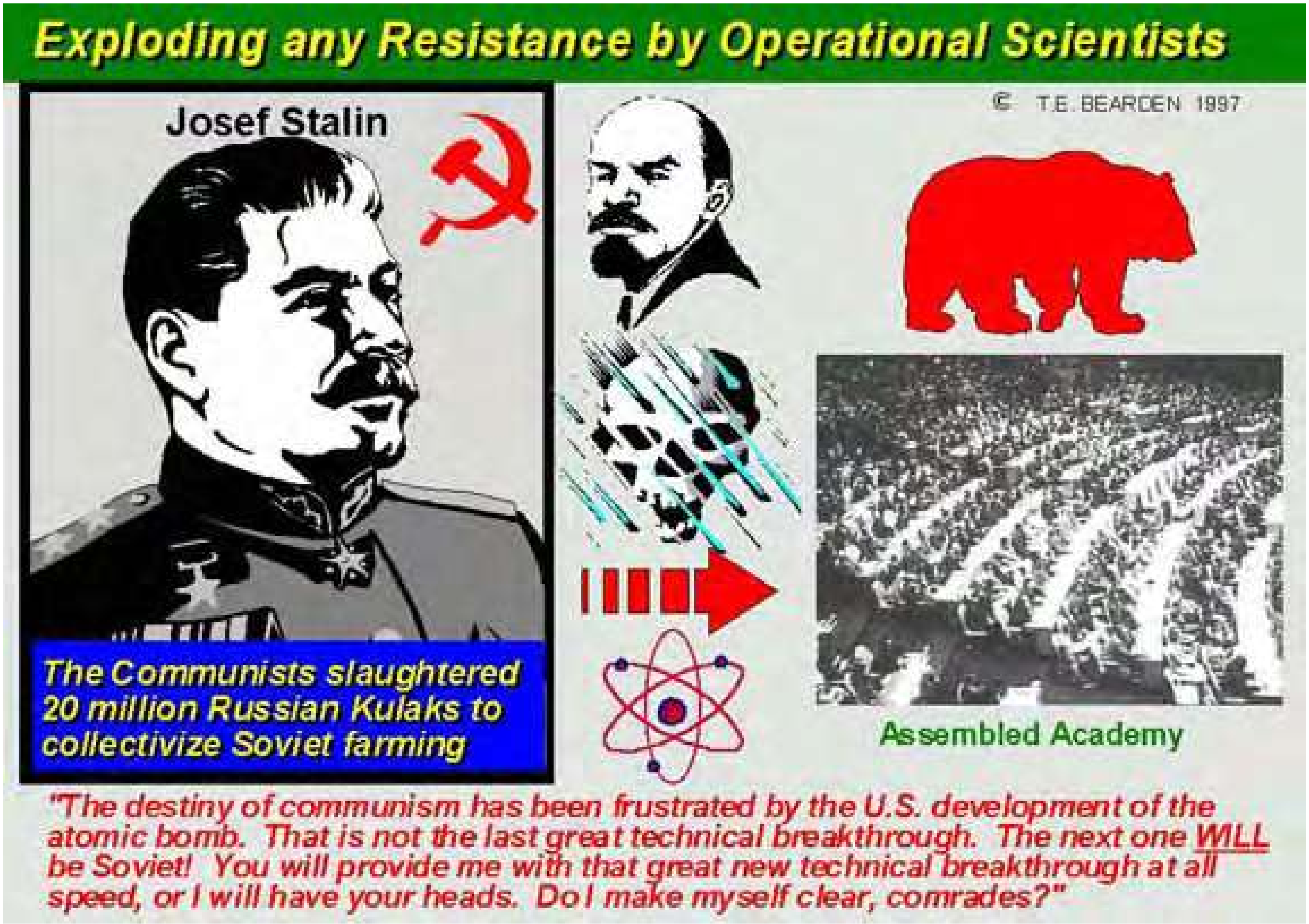

Exploding any Resistance by Operational Scientists

© T.E. BEARDEN 1997

Josef Stalin

The Communists slaughtered 20 million Russian Kulaks to collectivize Soviet farming

Assembled Academy

"The destiny of communism has been frustrated by the U.S. development of the atomic bomb. That is not the last great technical breakthrough. The next one WILL be Soviet! You will provide me with that great new technical breakthrough at all speed, or I will have your heads. Do I make myself clear, comrades?"

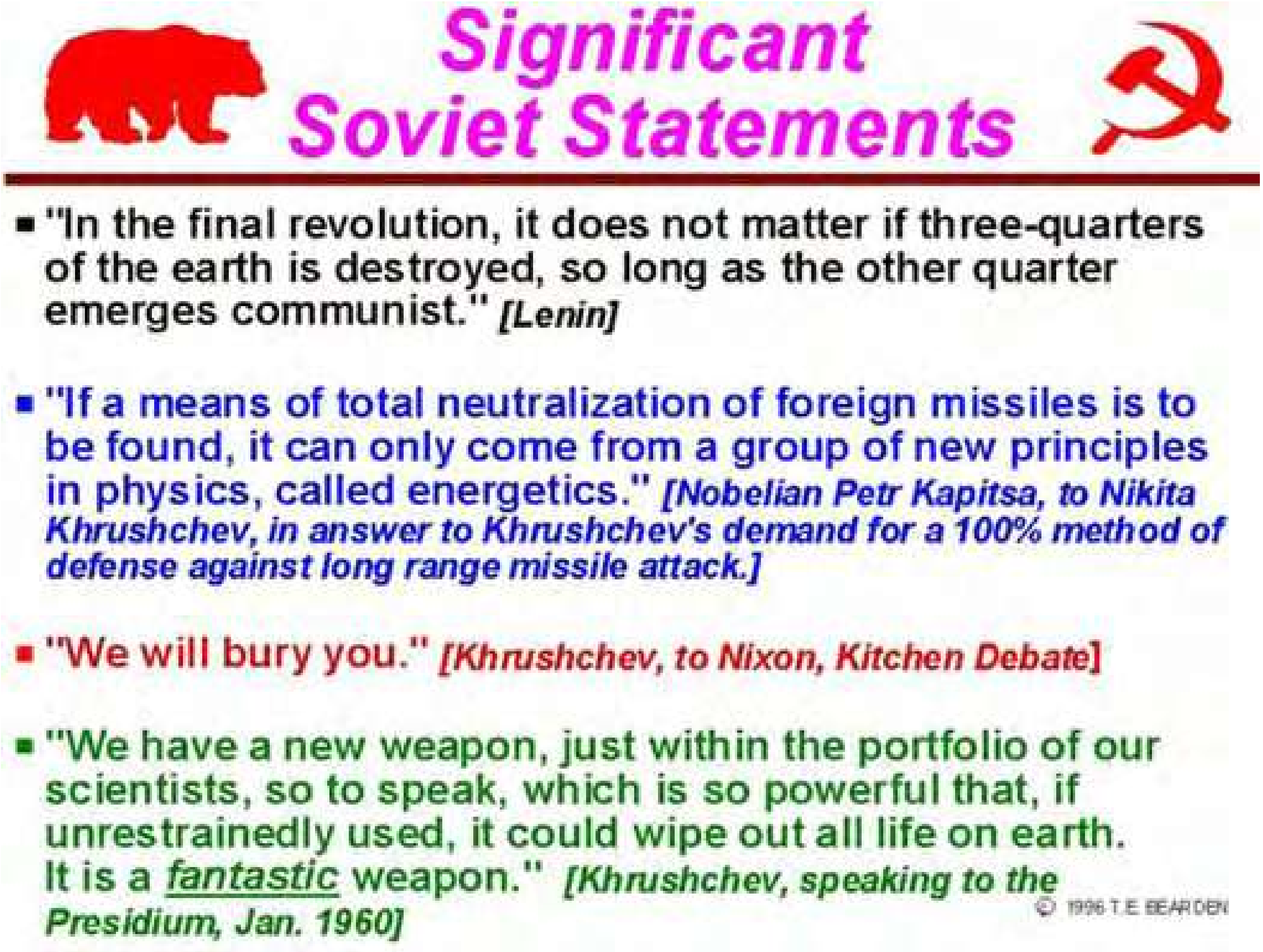

Significant Soviet Statements

- "In the final revolution, it does not matter if three-quarters of the earth is destroyed, so long as the other quarter emerges communist." *[Lenin]*
- "If a means of total neutralization of foreign missiles is to be found, it can only come from a group of new principles in physics, called energetics." *[Nobelian Petr Kapitsa, to Nikita Khrushchev, in answer to Khrushchev's demand for a 100% method of defense against long range missile attack.]*
- "We will bury you." *[Khrushchev, to Nixon, Kitchen Debate]*
- "We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that, if unrestrainedly used, it could wipe out all life on earth. It is a *fantastic* weapon." *[Khrushchev, speaking to the Presidium, Jan. 1960]*

© 1996 T.E. BEARDEN

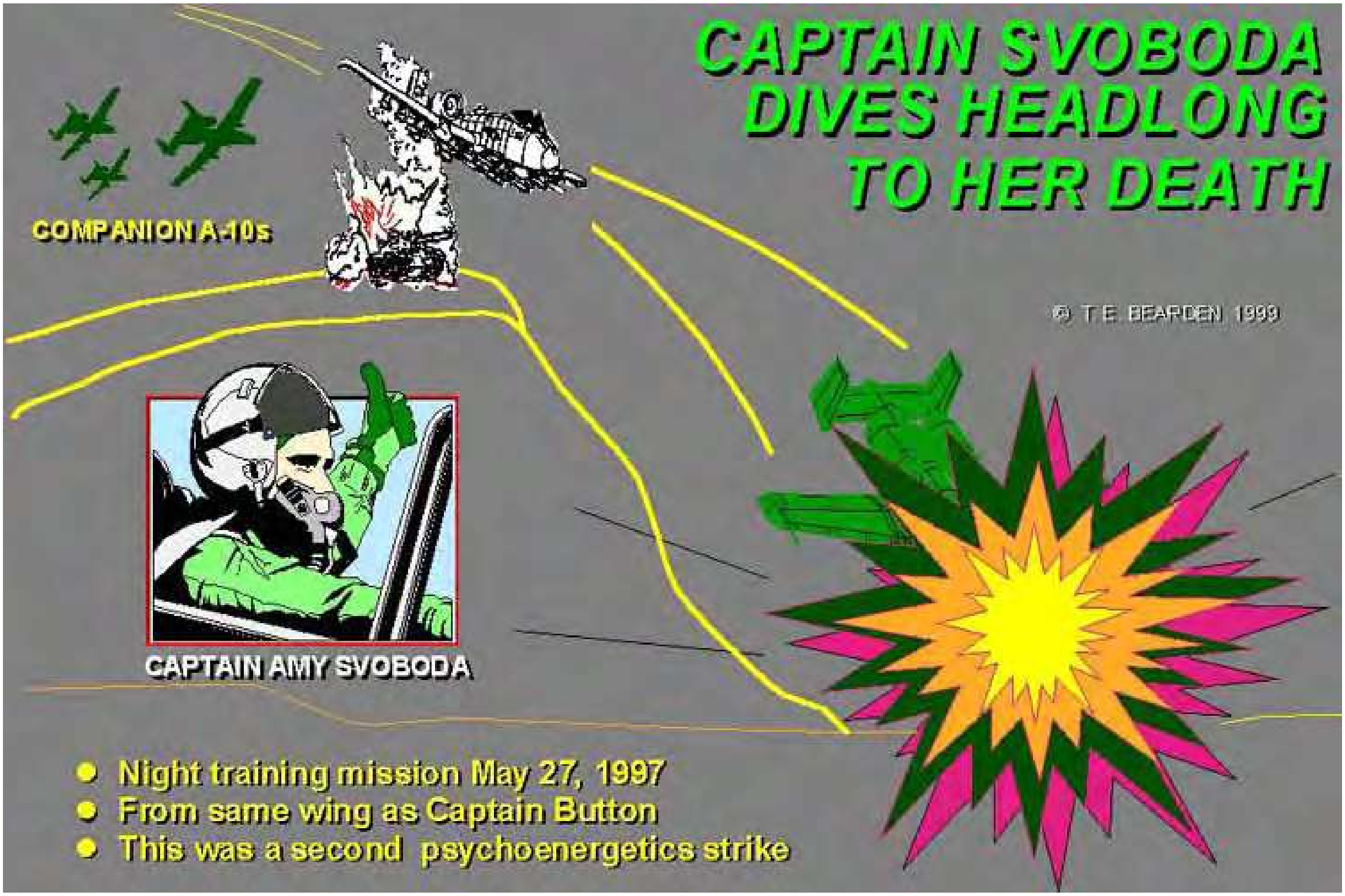

CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH

COMPANION A-10s

© T.E. BEARDEN 1999

CAPTAIN AMY SVOBODA

- Night training mission May 27, 1997
- From same wing as Captain Button
- This was a second psychoenergetics strike

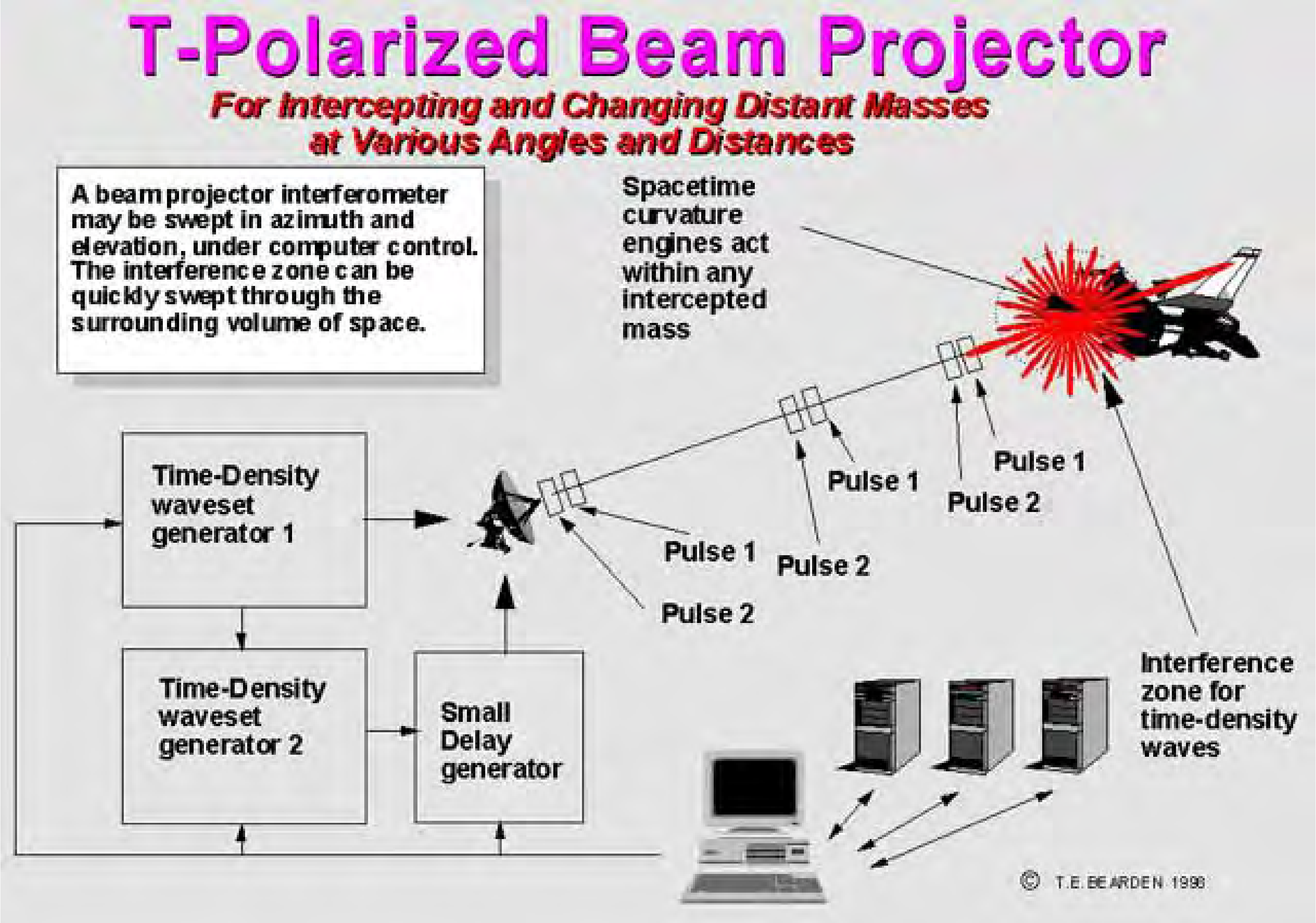

T-Polarized Beam Projector

For Intercepting and Changing Distant Masses at Various Angles and Distances

A beam projector interferometer may be swept in azimuth and elevation, under computer control. The interference zone can be quickly swept through the surrounding volume of space.

Spacetime curvature engines act within any intercepted mass
Time-Density waveset generator 1
Time-Density waveset generator 2
Small Delay generator
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Interference zone for time-density waves
© T.E. BEARDEN 1996

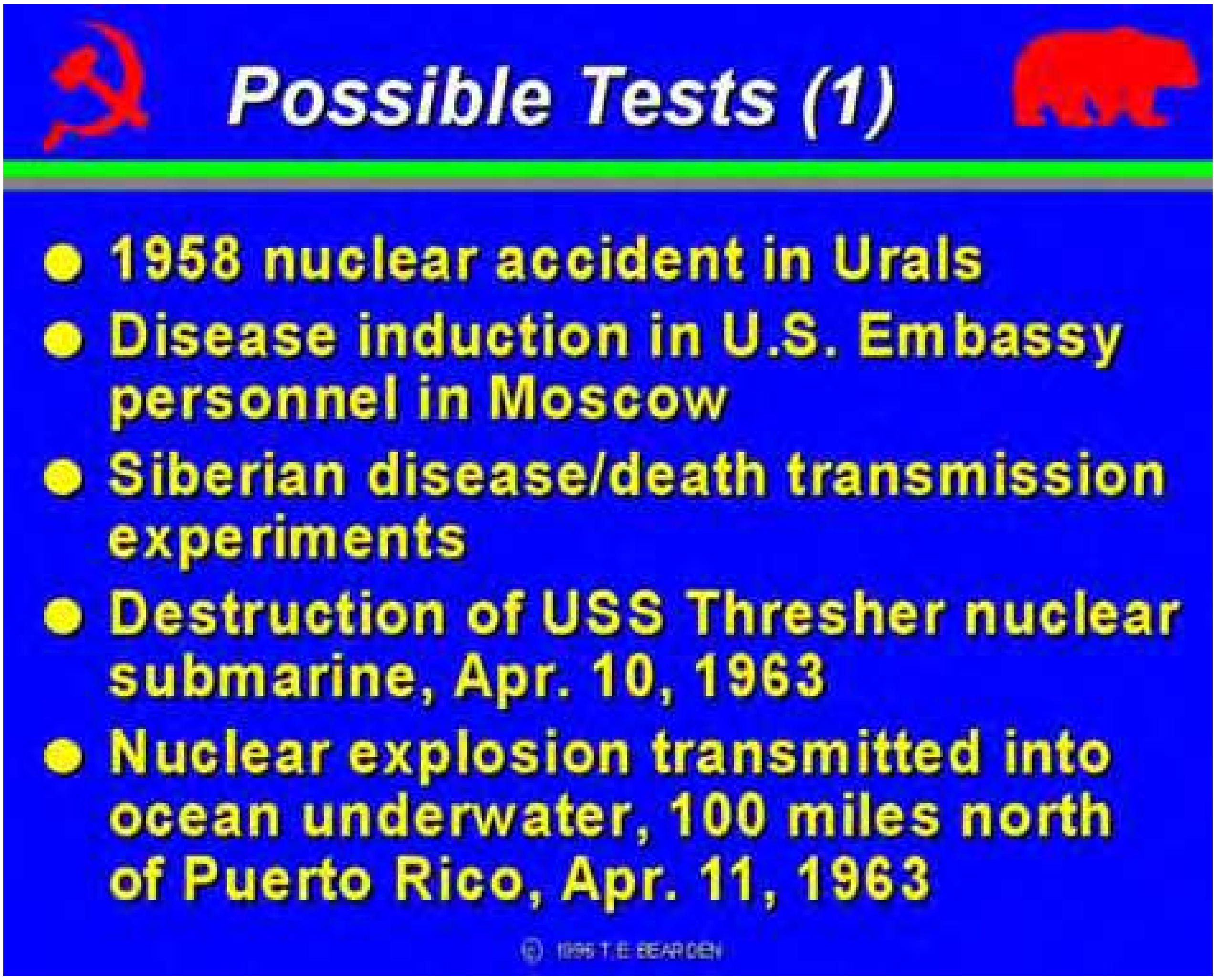

Possible Tests (1)

- 1958 nuclear accident in Urals
- Disease induction in U.S. Embassy personnel in Moscow
- Siberian disease/death transmission experiments
- Destruction of USS Thresher nuclear submarine, Apr. 10, 1963
- Nuclear explosion transmitted into ocean underwater, 100 miles north of Puerto Rico, Apr. 11, 1963

© 1996 T.E. BEARDEN

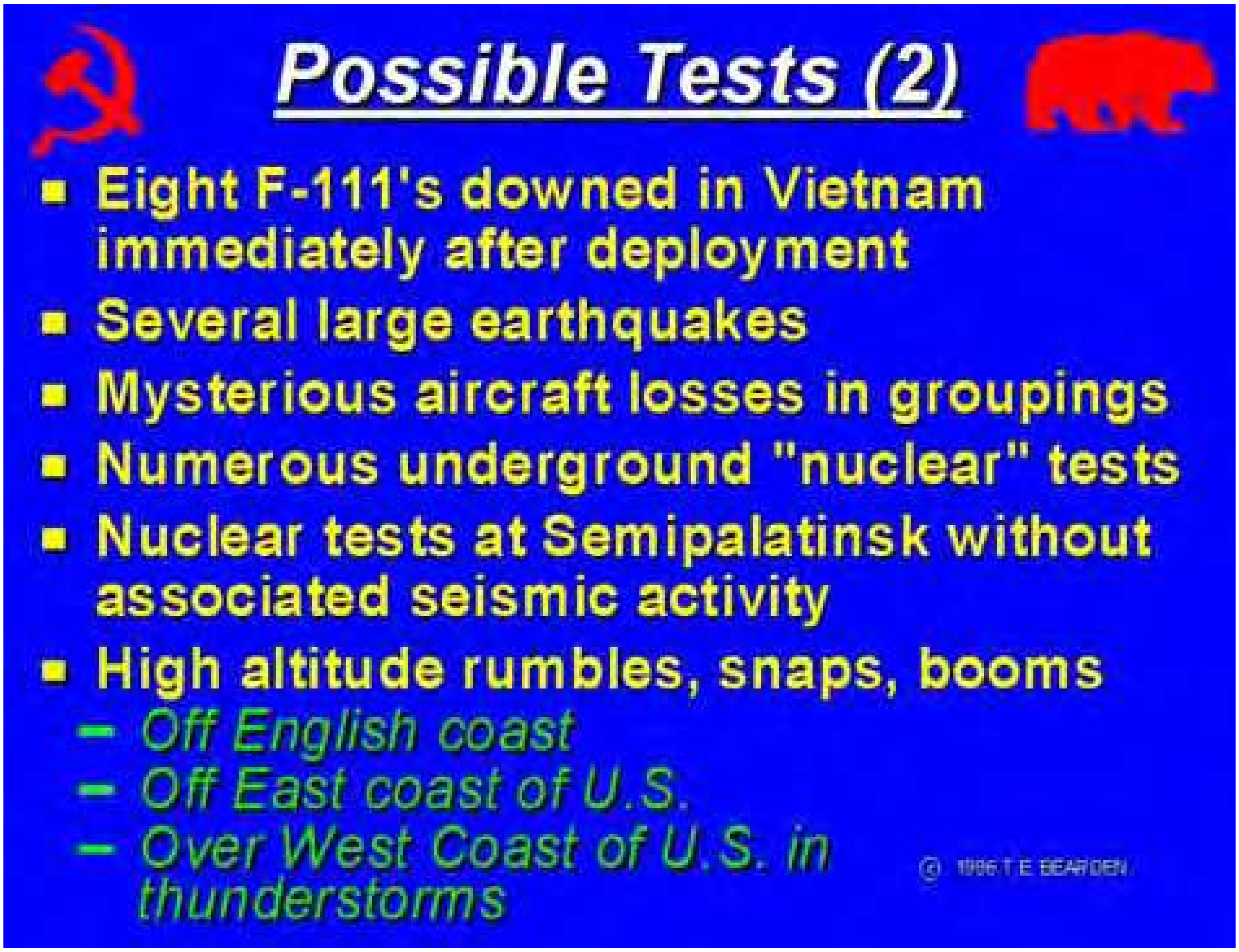

Possible Tests (2)

- Eight F-111's downed in Vietnam immediately after deployment
- Several large earthquakes
- Mysterious aircraft losses in groupings
- Numerous underground "nuclear" tests
- Nuclear tests at Semipalatinsk without associated seismic activity
- High altitude rumbles, snaps, booms
  - *Off English coast*
  - *Off East coast of U.S.*
  - *Over West Coast of U.S. in thunderstorms*

© 1996 T E BEARDEN

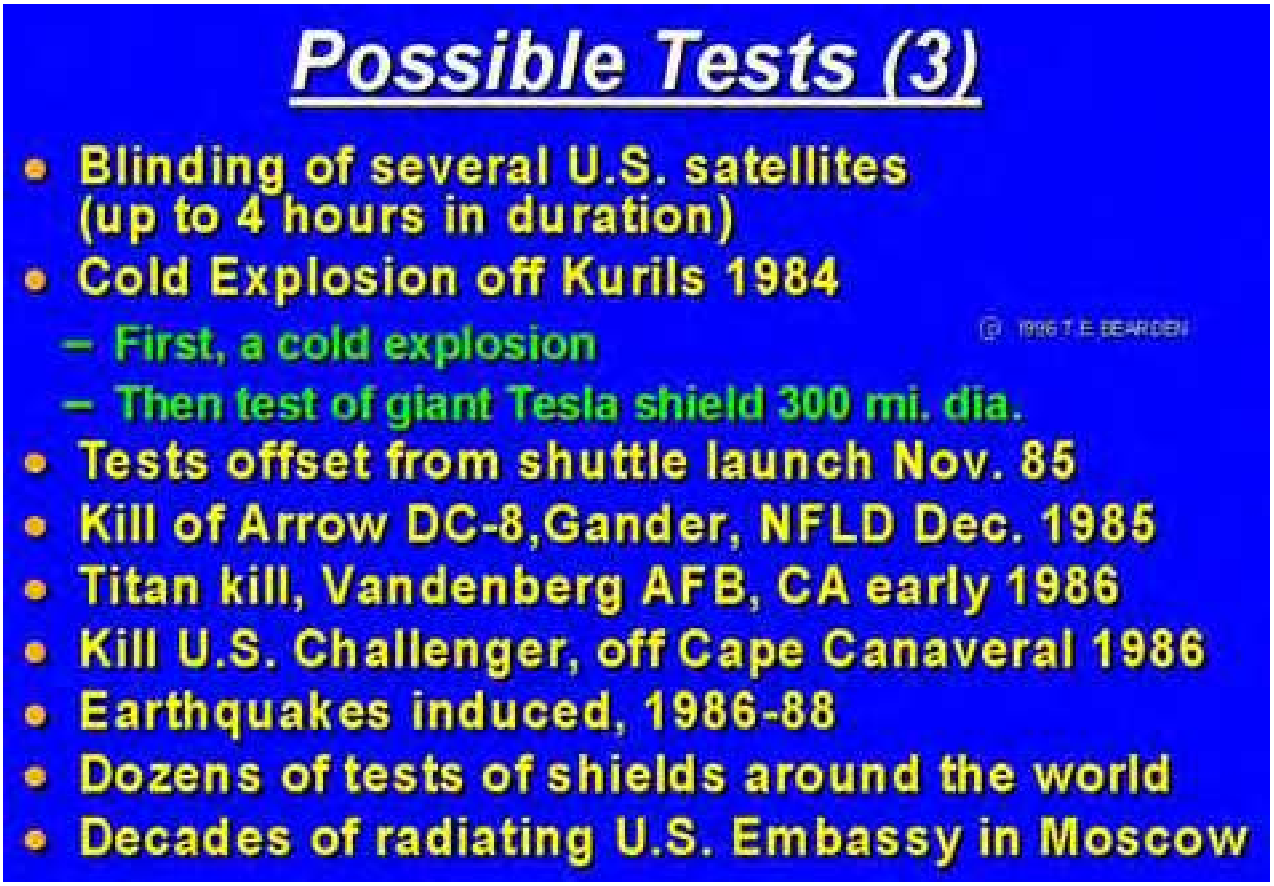

Possible Tests (3)

- Blinding of several U.S. satellites (up to 4 hours in duration)
- Cold Explosion off Kurils 1984
  - First, a cold explosion
  - Then test of giant Tesla shield 300 mi. dia.
- Tests offset from shuttle launch Nov. 85
- Kill of Arrow DC-8,Gander, NFLD Dec. 1985
- Titan kill, Vandenberg AFB, CA early 1986
- Kill U.S. Challenger, off Cape Canaveral 1986
- Earthquakes induced, 1986-88
- Dozens of tests of shields around the world
- Decades of radiating U.S. Embassy in Moscow

© 1996 T. E. BEARDEN

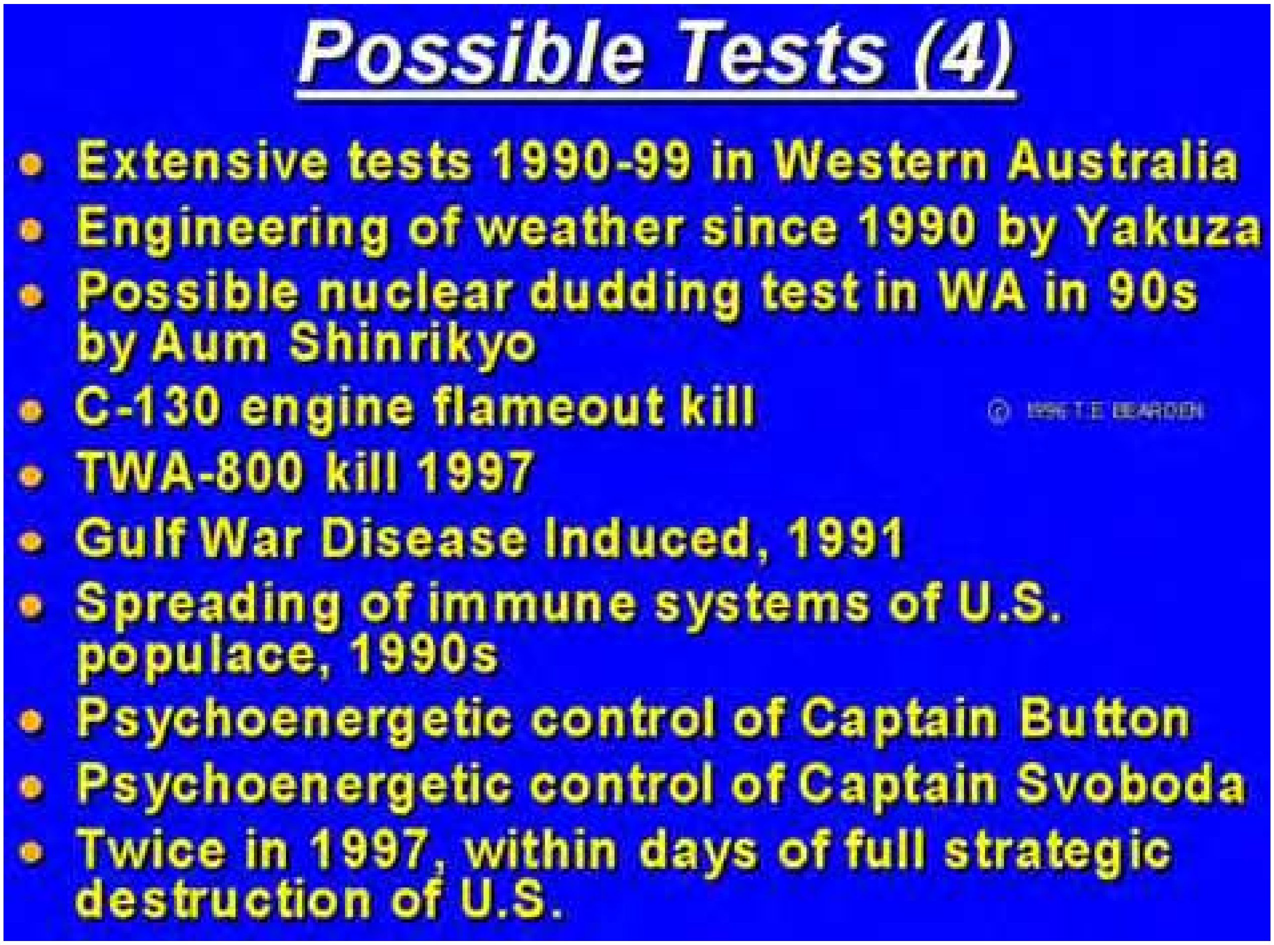

Possible Tests (4)

- Extensive tests 1990-99 in Western Australia
- Engineering of weather since 1990 by Yakuza
- Possible nuclear dudding test in WA in 90s by Aum Shinrikyo
- C-130 engine flameout kill
- TWA-800 kill 1997
- Gulf War Disease Induced, 1991
- Spreading of immune systems of U.S. populace, 1990s
- Psychoenergetic control of Captain Button
- Psychoenergetic control of Captain Svoboda
- Twice in 1997, within days of full strategic destruction of U.S.

© 1996 T.E BEARDEN

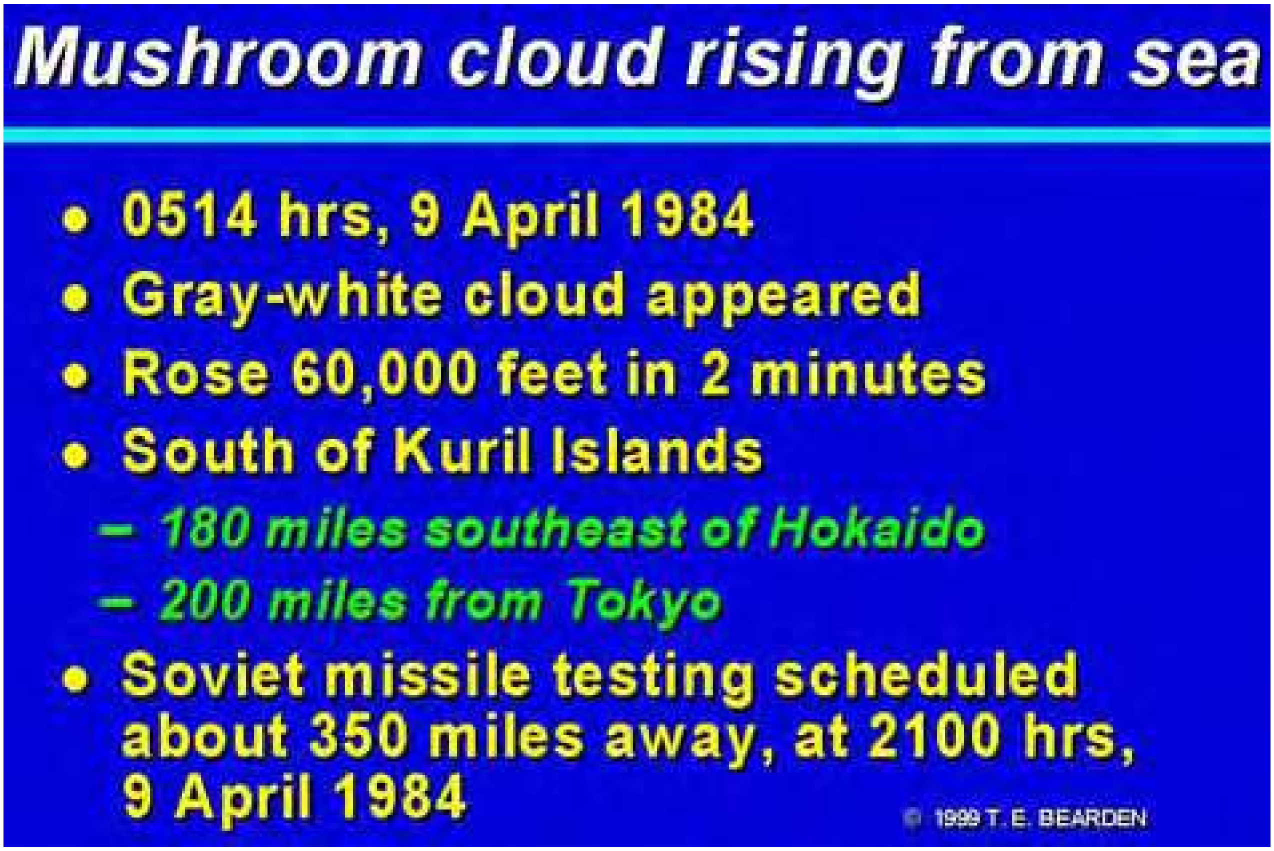

Mushroom cloud rising from sea

- 0514 hrs, 9 April 1984
- Gray-white cloud appeared
- Rose 60,000 feet in 2 minutes
- South of Kuril Islands
  - 180 miles southeast of Hokaido
  - 200 miles from Tokyo
- Soviet missile testing scheduled about 350 miles away, at 2100 hrs, 9 April 1984

© 1999 T. E. BEARDEN

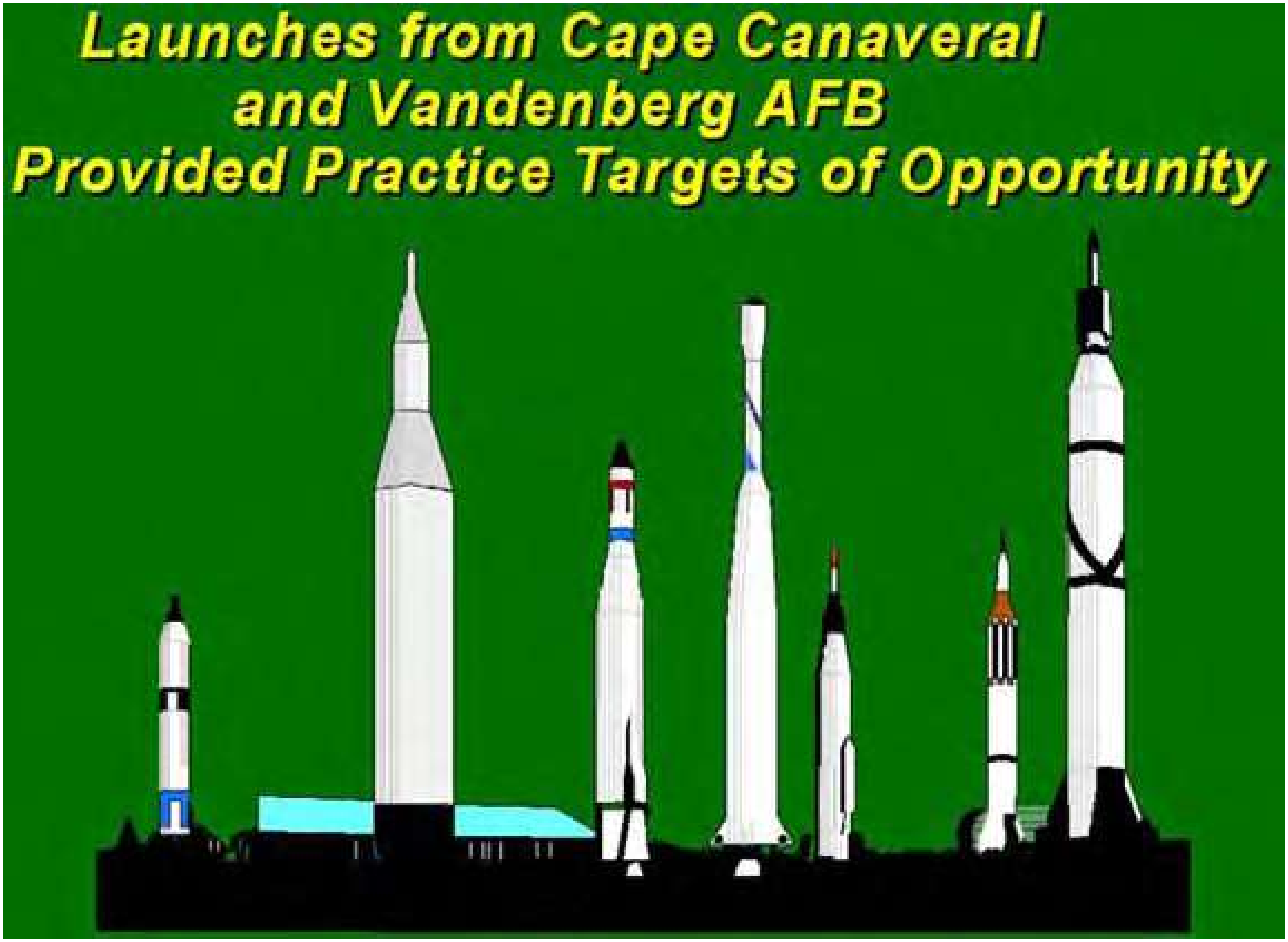
Launches from Cape Canaveral
and Vandenberg AFB
Provided Practice Targets of Opportunity

The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth

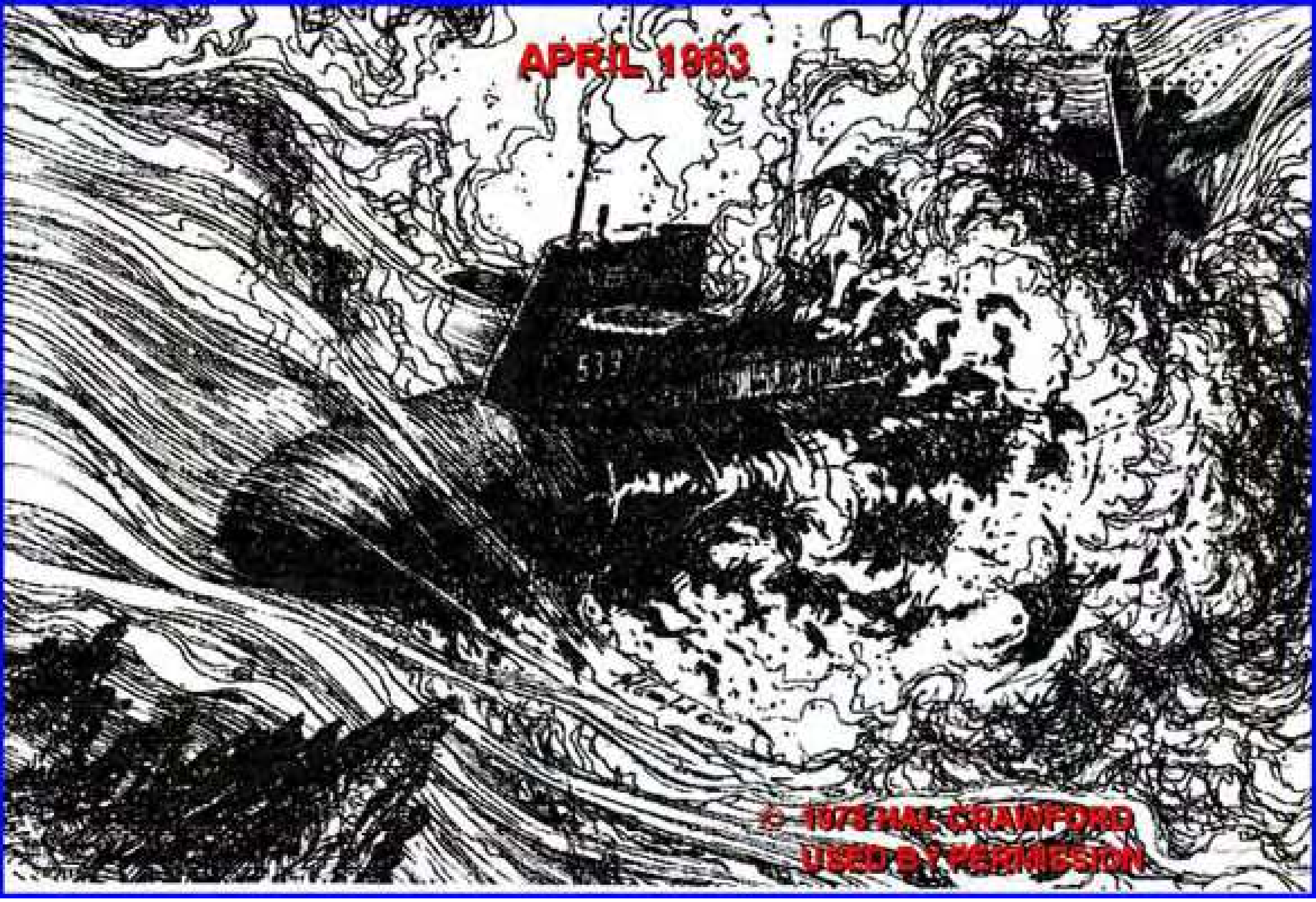
APRIL 1963
533
© 1978 HAL CRAWFORD
USED BY PERMISSION

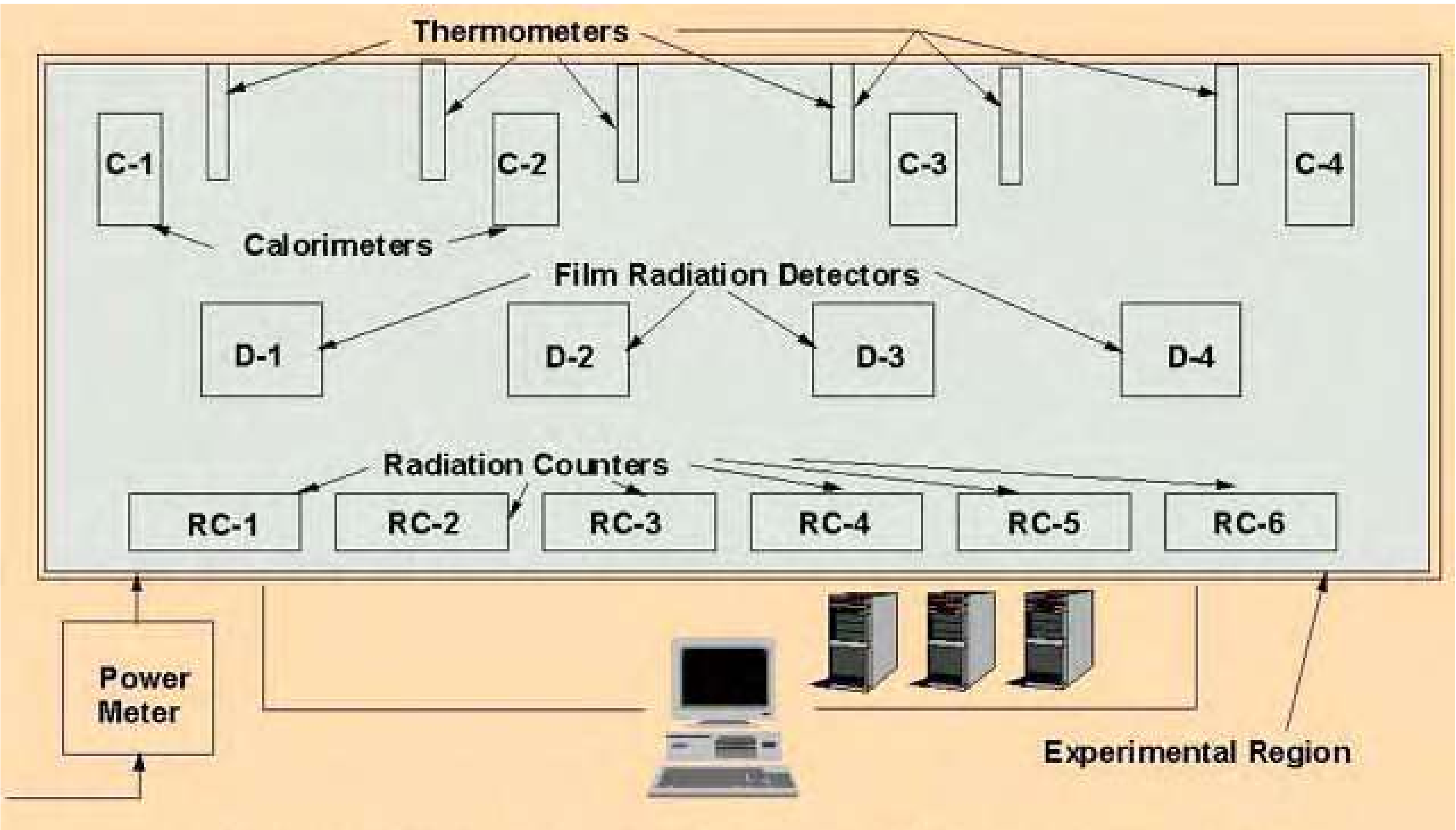
Thermometers
C-1
C-2
C-3
C-4
Calorimeters
Film Radiation Detectors
D-1
D-2
D-3
D-4
Radiation Counters
RC-1
RC-2
RC-3
RC-4
RC-5
RC-6
Power Meter
Experimental Region

Figure 32. Embodiment of an instrumental array sufficient to discriminate when time-density charging has occurred, and that transduction of time-charge decay is occurring.

© T.E. BEARDEN 1998

Time-Polarized EM Wave interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone

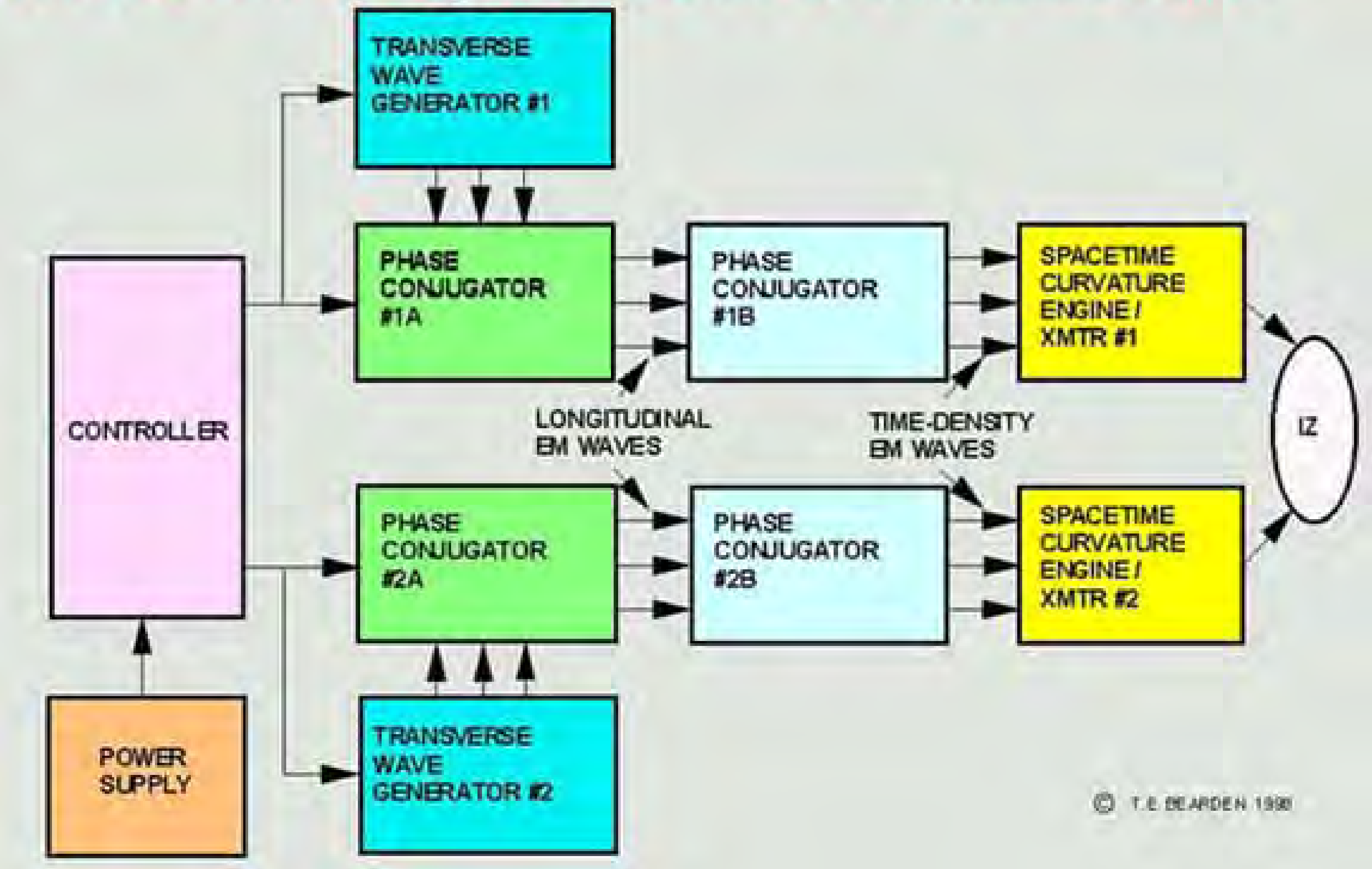
TRANSVERSE WAVE GENERATOR #1
PHASE CONJUGATOR #1A
PHASE CONJUGATOR #1B
SPACETIME CURVATURE ENGINE / XMTR #1
CONTROLLER
LONGITUDINAL EM WAVES
TIME-DENSITY EM WAVES
IZ
PHASE CONJUGATOR #2A
PHASE CONJUGATOR #2B
SPACETIME CURVATURE ENGINE / XMTR #2
POWER SUPPLY
TRANSVERSE WAVE GENERATOR #2
© T.E. BEARDEN 1998

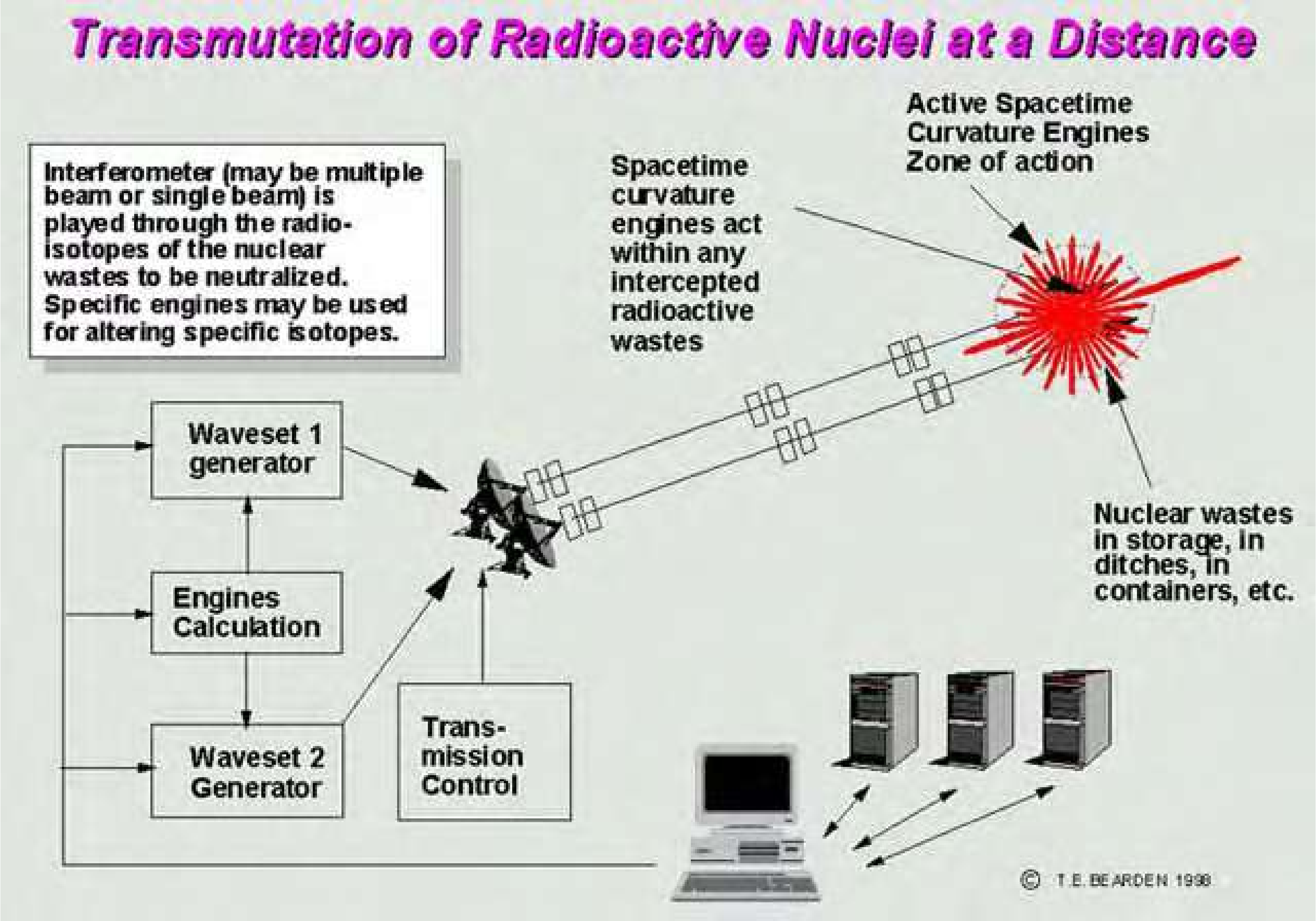

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radioisotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control
Spacetime curvature engines act within any intercepted radioactive wastes
Active Spacetime Curvature Engines Zone of action
Nuclear wastes in storage, in ditches, in containers, etc.

© T.E. BEARDEN 1998

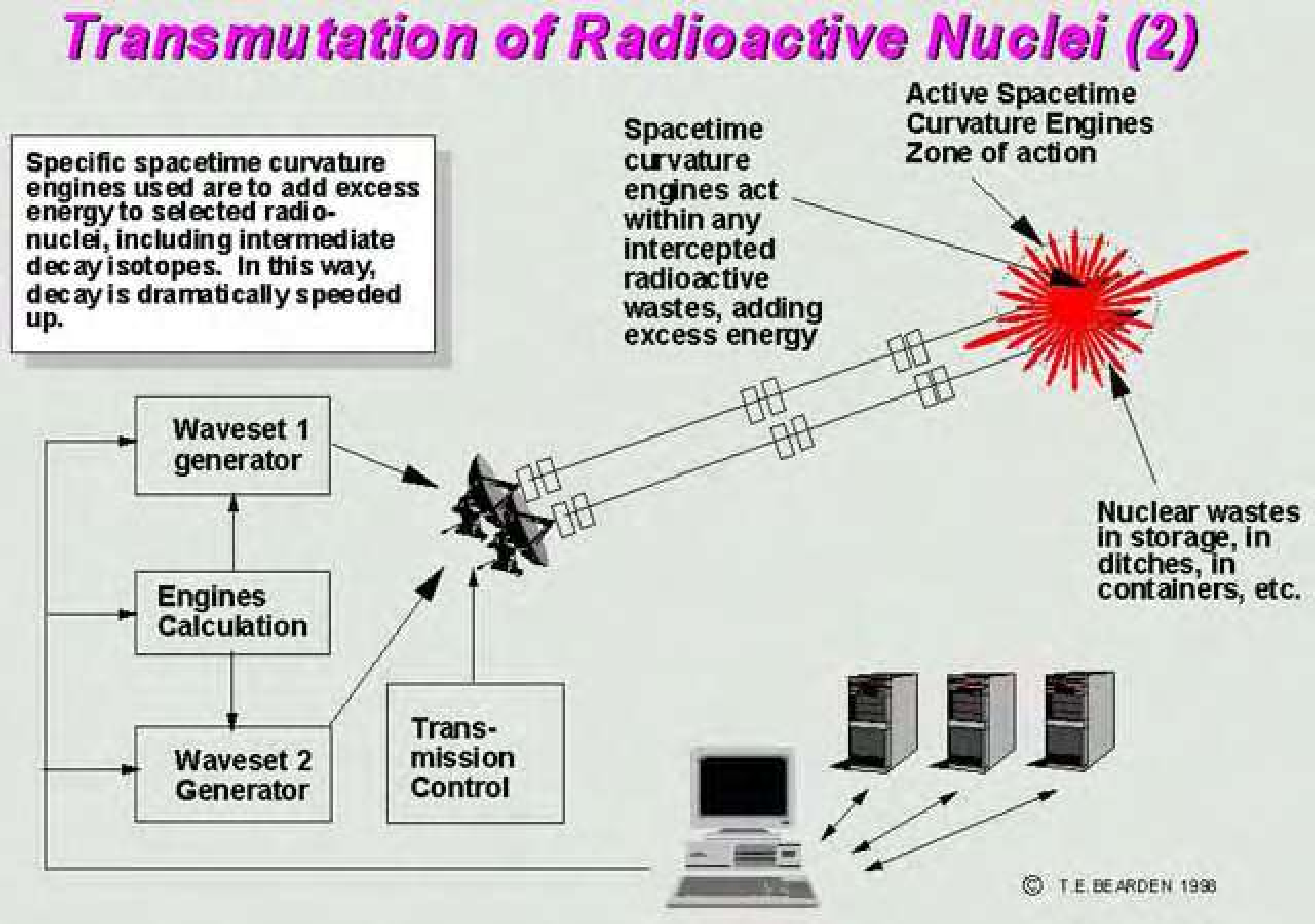

Transmutation of Radioactive Nuclei (2)

Specific spacetime curvature engines used are to add excess energy to selected radio-nuclei, including intermediate decay isotopes. In this way, decay is dramatically speeded up.

Spacetime curvature engines act within any intercepted radioactive wastes, adding excess energy

Active Spacetime Curvature Engines Zone of action

Nuclear wastes in storage, in ditches, in containers, etc.

Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control

© T.E. BEARDEN 1998

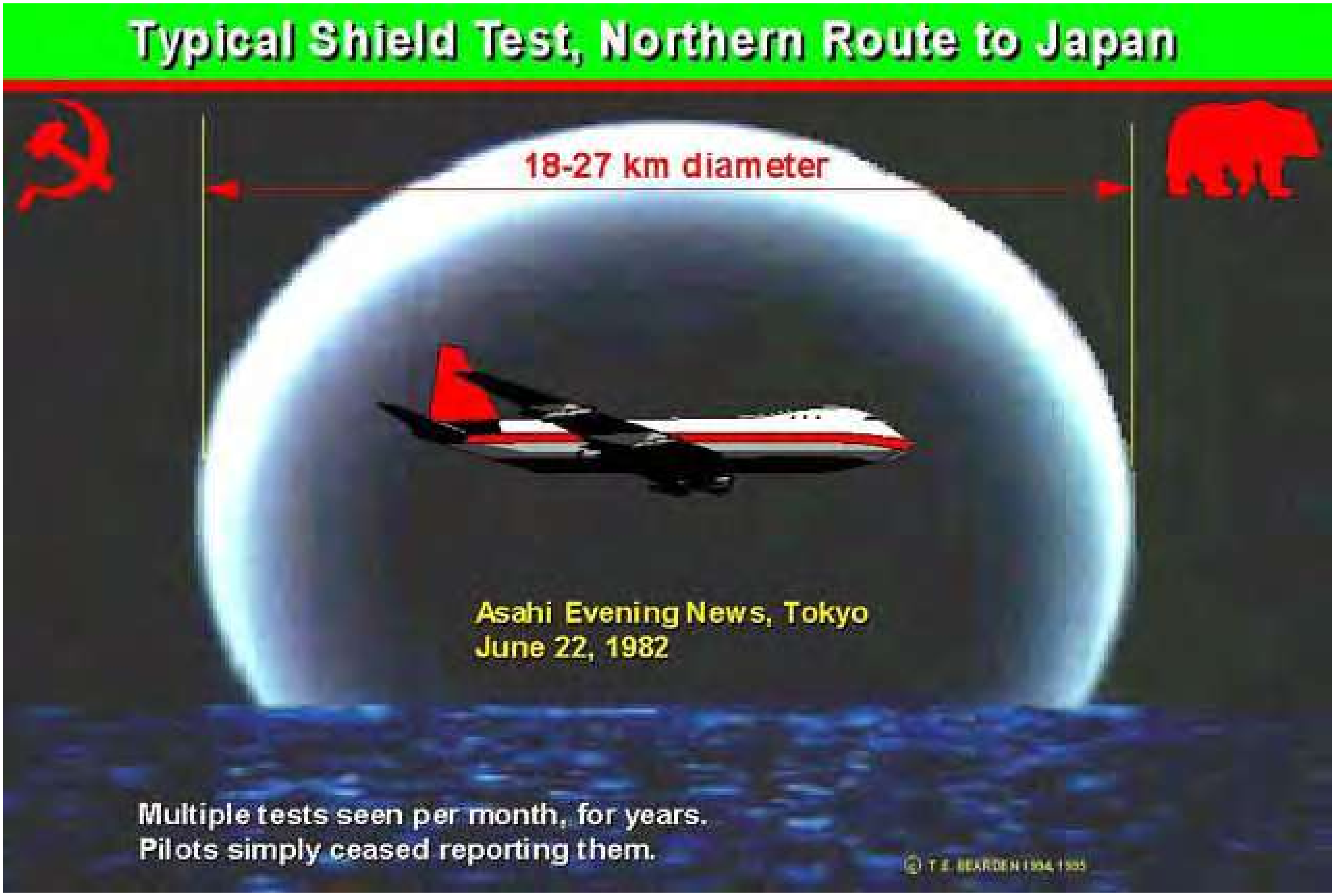

Typical Shield Test, Northern Route to Japan

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T. E. BEARDEN 1994, 1995

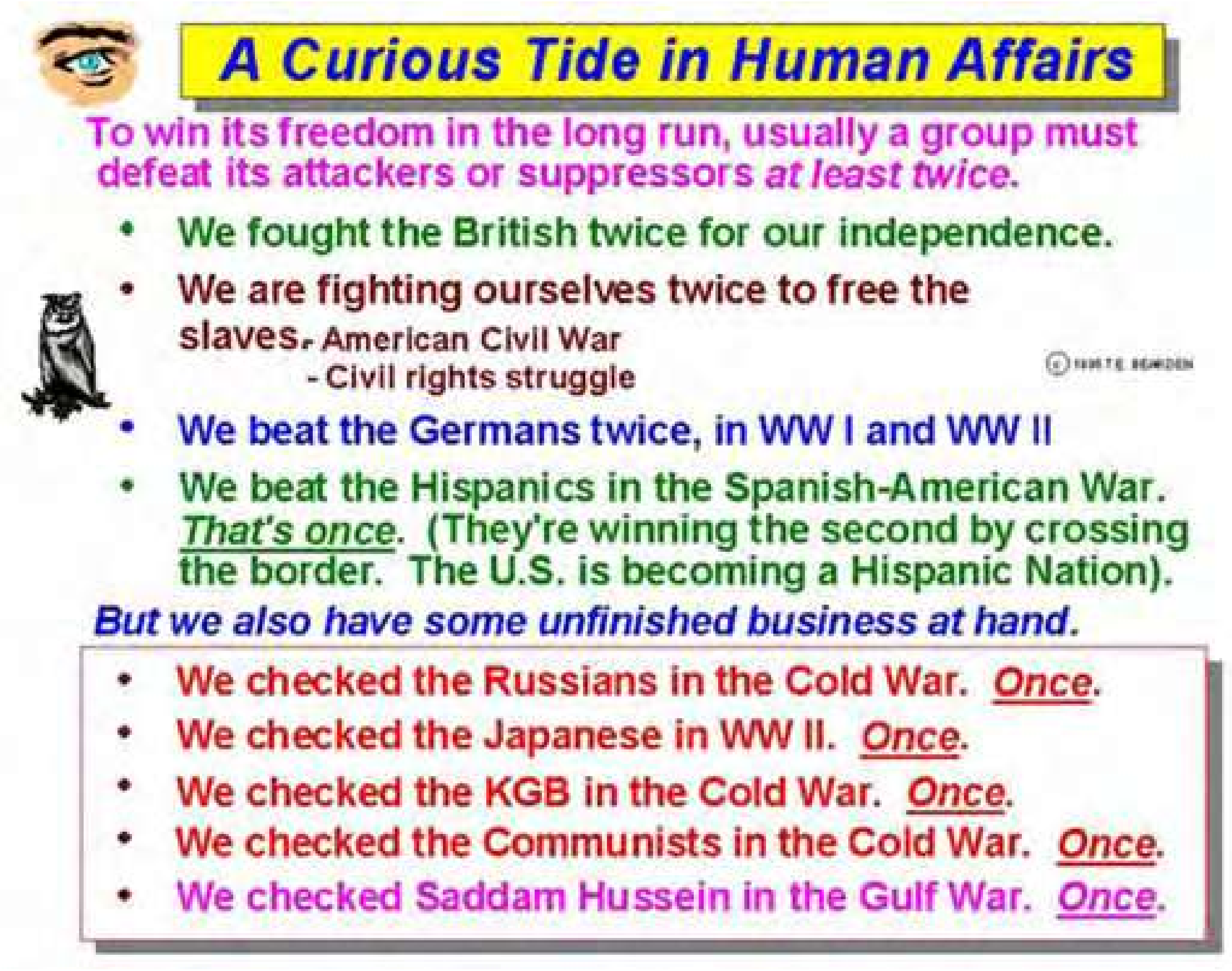

A Curious Tide in Human Affairs

To win its freedom in the long run, usually a group must defeat its attackers or suppressors *at least twice*.

- We fought the British twice for our independence.
- We are fighting ourselves twice to free the slaves. - American Civil War
  - Civil rights struggle
- We beat the Germans twice, in WW I and WW II
- We beat the Hispanics in the Spanish-American War. *That's once*. (They're winning the second by crossing the border. The U.S. is becoming a Hispanic Nation).

But we also have some unfinished business at hand.

- We checked the Russians in the Cold War. *Once*.
- We checked the Japanese in WW II. *Once*.
- We checked the KGB in the Cold War. *Once*.
- We checked the Communists in the Cold War. *Once*.
- We checked Saddam Hussein in the Gulf War. *Once*.

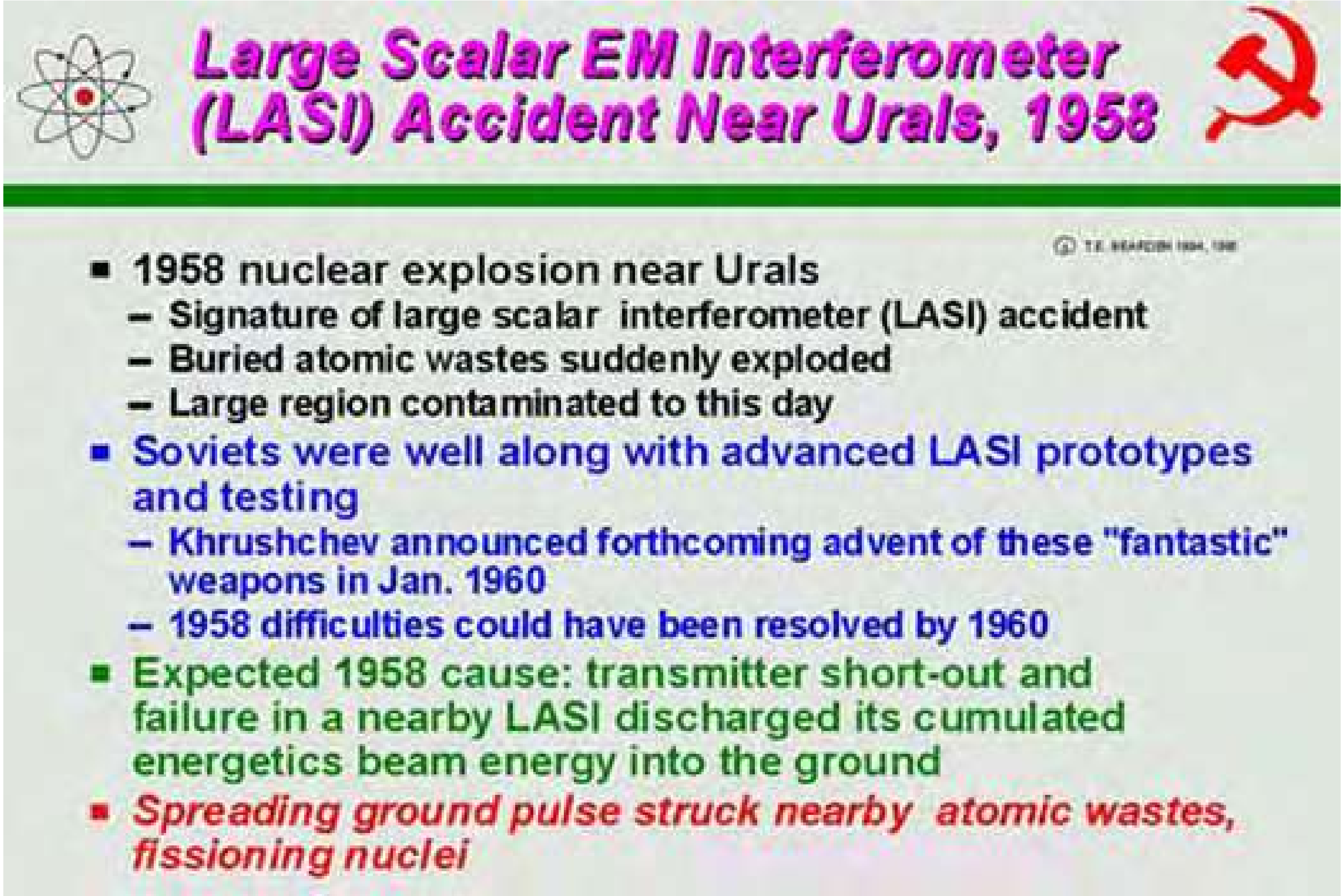

Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958

- 1958 nuclear explosion near Urals
  - Signature of large scalar interferometer (LASI) accident
  - Buried atomic wastes suddenly exploded
  - Large region contaminated to this day
- Soviets were well along with advanced LASI prototypes and testing
  - Khrushchev announced forthcoming advent of these "fantastic" weapons in Jan. 1960
  - 1958 difficulties could have been resolved by 1960
- Expected 1958 cause: transmitter short-out and failure in a nearby LASI discharged its cumulated energetics beam energy into the ground
- *Spreading ground pulse struck nearby atomic wastes, fissioning nuclei*

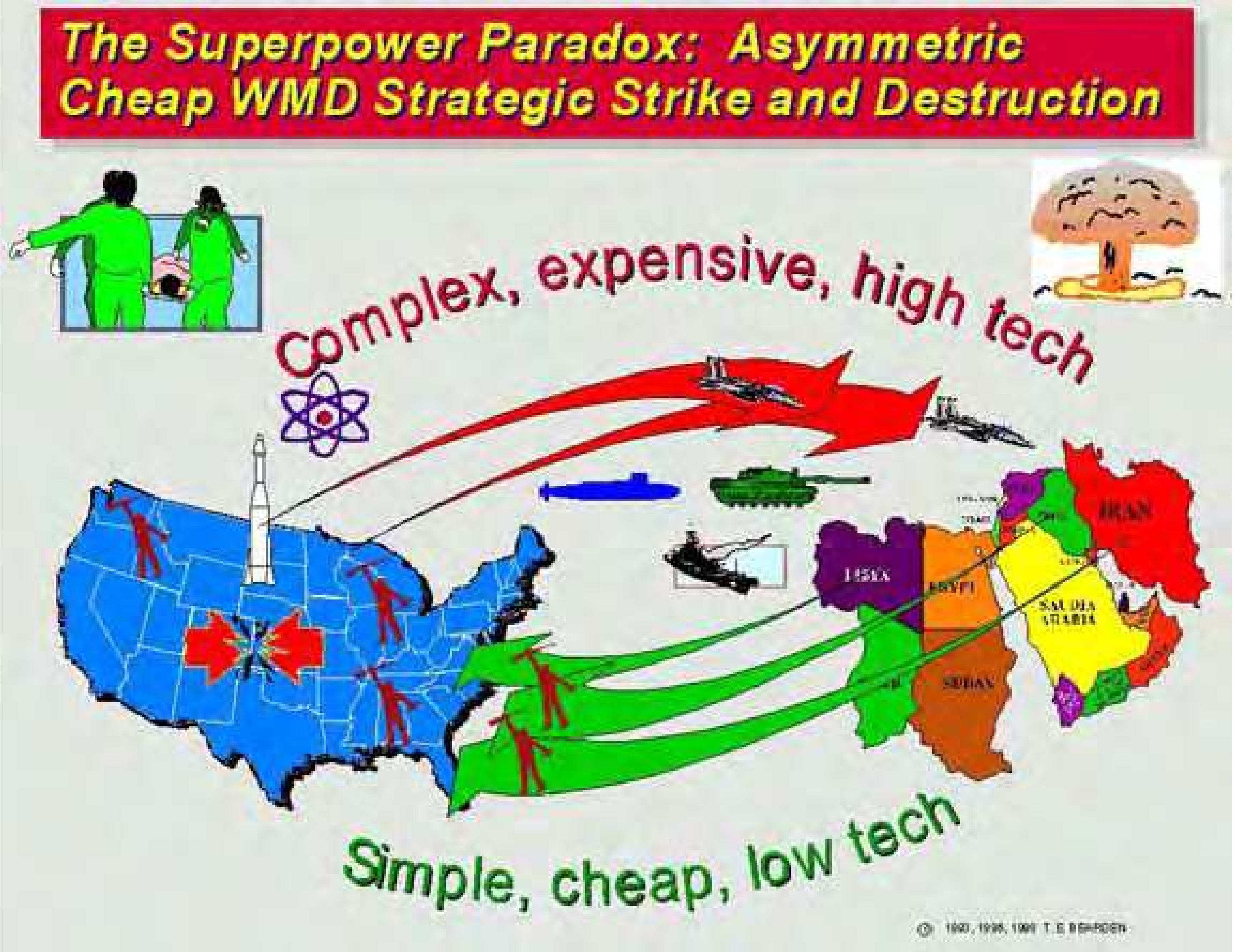

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

Simple, cheap, low tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN

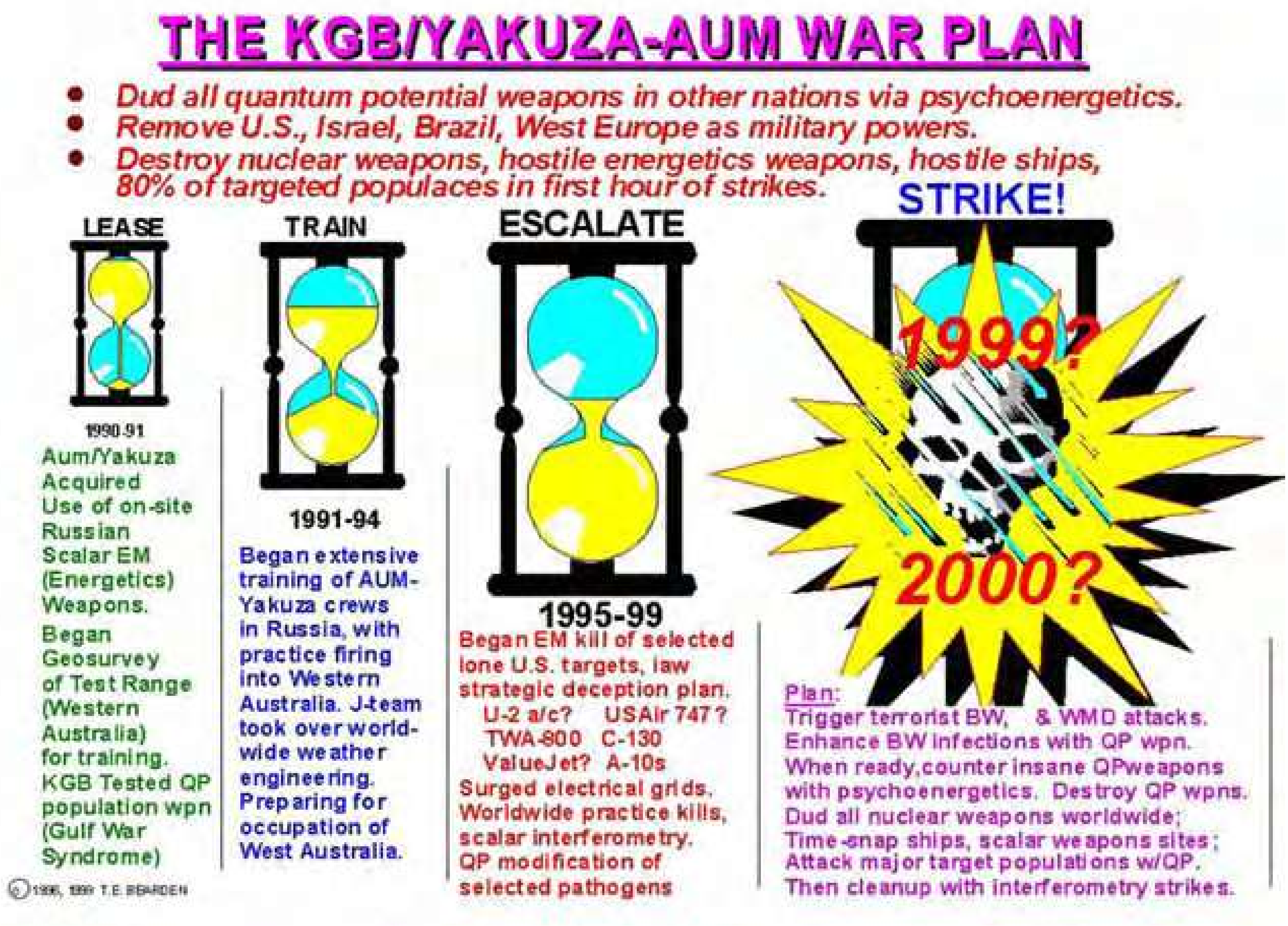

THE KGB/YAKUZA-AUM WAR PLAN

- *Dud all quantum potential weapons in other nations via psychoenergetics.*
- *Remove U.S., Israel, Brazil, West Europe as military powers.*
- *Destroy nuclear weapons, hostile energetics weapons, hostile ships, 80% of targeted populaces in first hour of strikes.*

LEASE

1990-91

Aum/Yakuza Acquired Use of on-site Russian Scalar EM (Energetics) Weapons.

Began Geosurvey of Test Range (Western Australia) for training.

KGB Tested QP population wpn (Gulf War Syndrome)

TRAIN

1991-94

Began extensive training of AUM-Yakuza crews in Russia, with practice firing into Western Australia. J-team took over world-wide weather engineering. Preparing for occupation of West Australia.

ESCALATE

1995-99

Began EM kill of selected lone U.S. targets, law strategic deception plan.

U-2 a/c? USAir 747 ?
TWA-800 C-130
ValueJet? A-10s

Surged electrical grids. Worldwide practice kills, scalar interferometry. QP modification of selected pathogens

STRIKE!

Plan:
Trigger terrorist BW, & WMD attacks.
Enhance BW infections with QP wpn.
When ready, counter insane QPweapons with psychoenergetics. Destroy QP wpns.
Dud all nuclear weapons worldwide;
Time-snap ships, scalar weapons sites;
Attack major target populations w/QP.
Then cleanup with interferometry strikes.

© 1996, 1999 T.E. BEARDEN

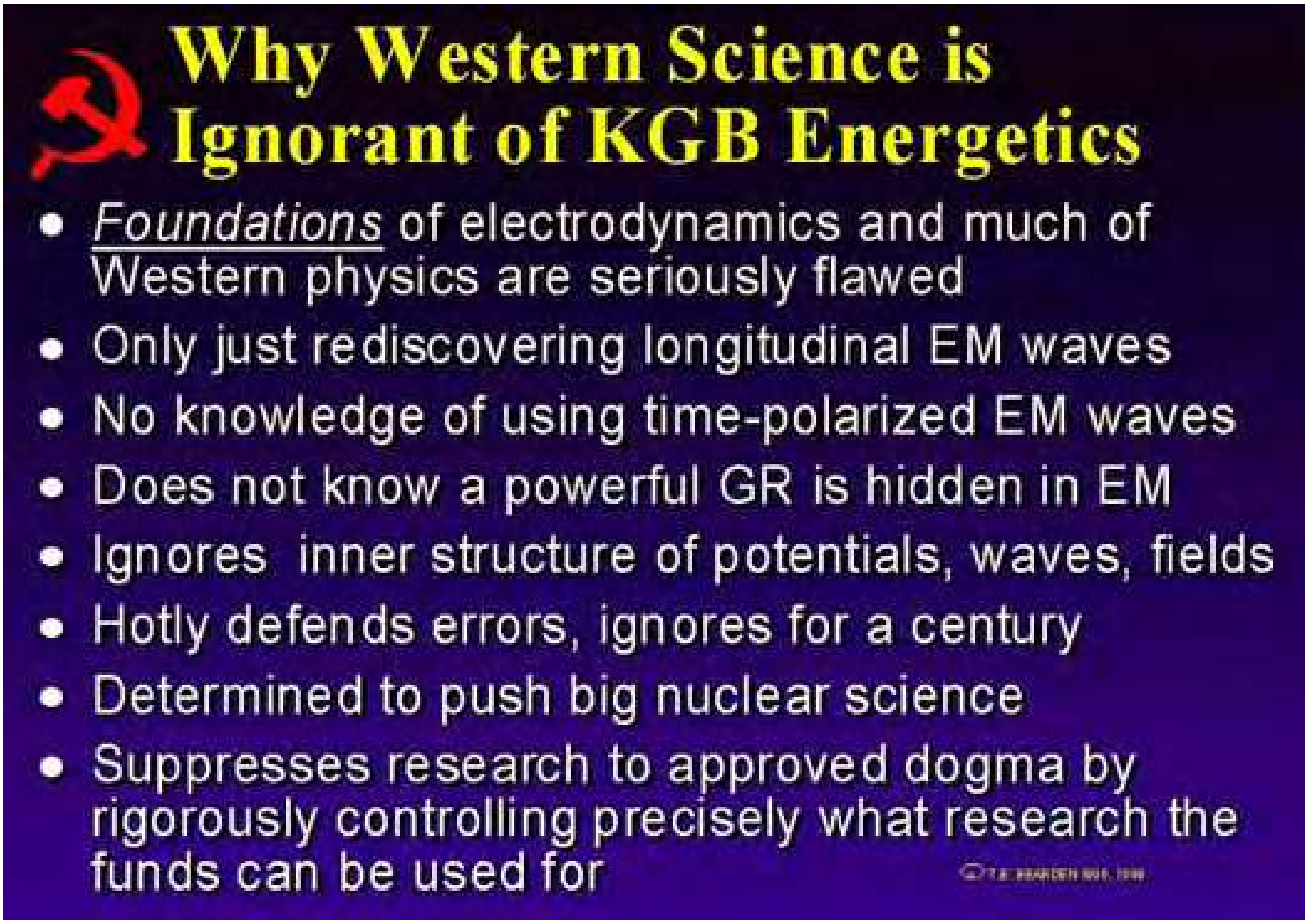

Why Western Science is Ignorant of KGB Energetics

- *Foundations* of electrodynamics and much of Western physics are seriously flawed
- Only just rediscovering longitudinal EM waves
- No knowledge of using time-polarized EM waves
- Does not know a powerful GR is hidden in EM
- Ignores inner structure of potentials, waves, fields
- Hotly defends errors, ignores for a century
- Determined to push big nuclear science
- Suppresses research to approved dogma by rigorously controlling precisely what research the funds can be used for

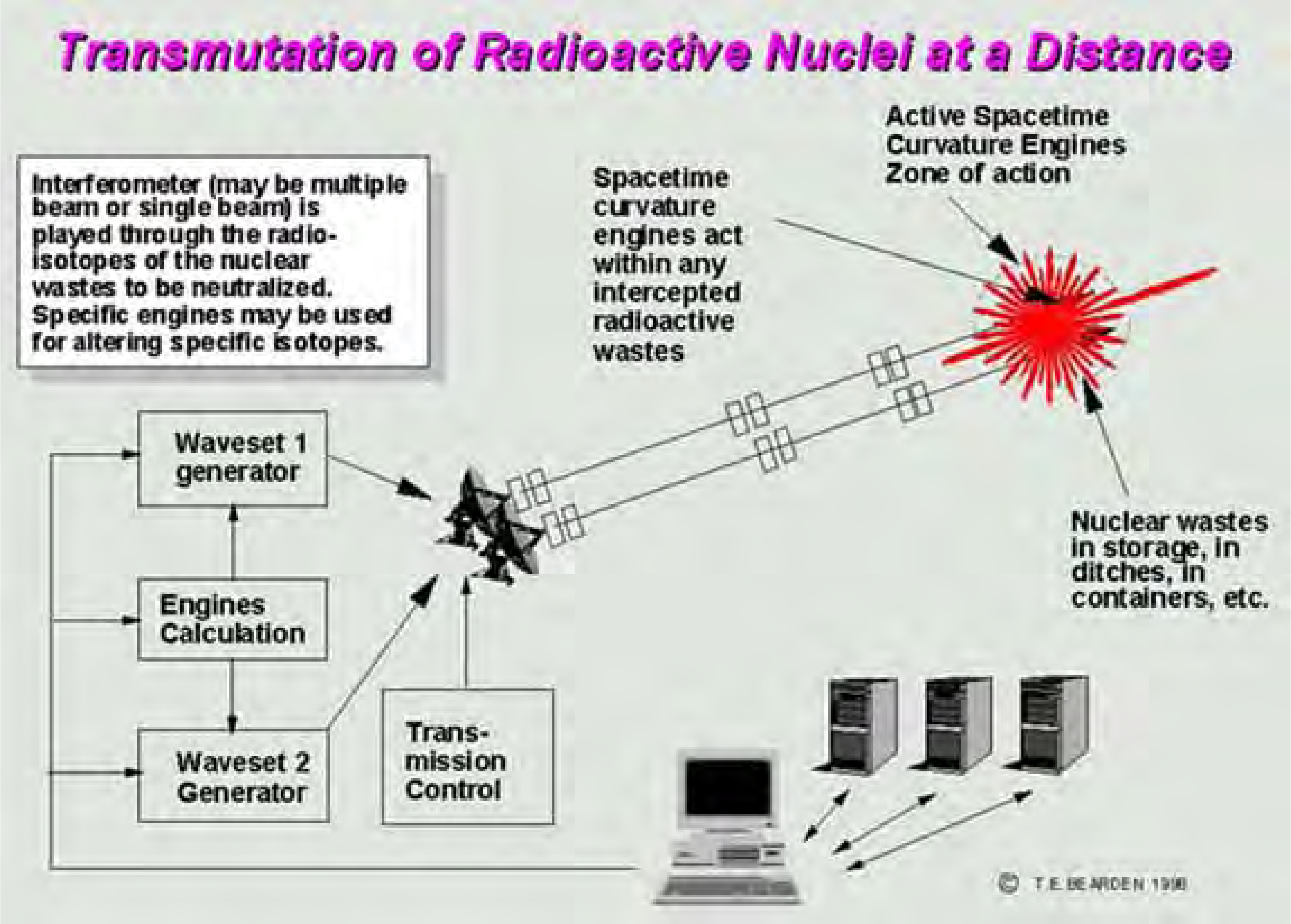

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radioisotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes

Active Spacetime Curvature Engines Zone of action

Nuclear wastes in storage, in ditches, in containers, etc.

Waveset 1 generator

Engines Calculation

Waveset 2 Generator

Trans-mission Control

© T.E. BEARDEN 1998

THE YAKUZA

(JAPANESE MAFIA)

To do business with Japan is to do business with the Yakuza

The Last Samurai, Defending Bushido

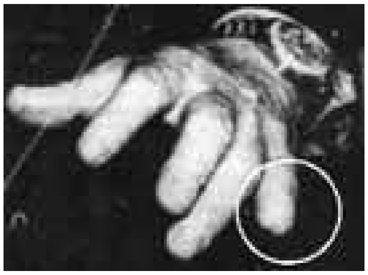

© 1995, 1996, 2001 T.E. BEARDEN

- **3 Major Yakuza Groups, In Coalition.**
- **Penetrate Japanese Government.**
- **Open Offices, Signs on Street.**
- **Public Romanticizes Yakuza as Last Japanese Samurai**
- **In Large High-Tech Corporations.**
- **In Control of Japanese Banks.**
- **Japanese Real-Estate Rip-off.**
- **$500 - 1,000 Billion per Year Income.**
- **Rapidly Spreading in U.S.**
- **Uniting Mafia Groups Worldwide.**
- **Coalition with Aum Shinrikyo.**
- **Leased Russian Energetics Weapons.**
- **Crews On-site in Russia, Trained.**
- **Practice Firings into Western Australia for Three Years. (Dozens of Shots per Month)**
- **Now in Escalation Phase, Killing U.S. Targets.**
- **All-out Strike on U.S. ready 1999-2000.**
- **Now held in check by KGB, dead-man fuzed by small nation friendly to U. S.**

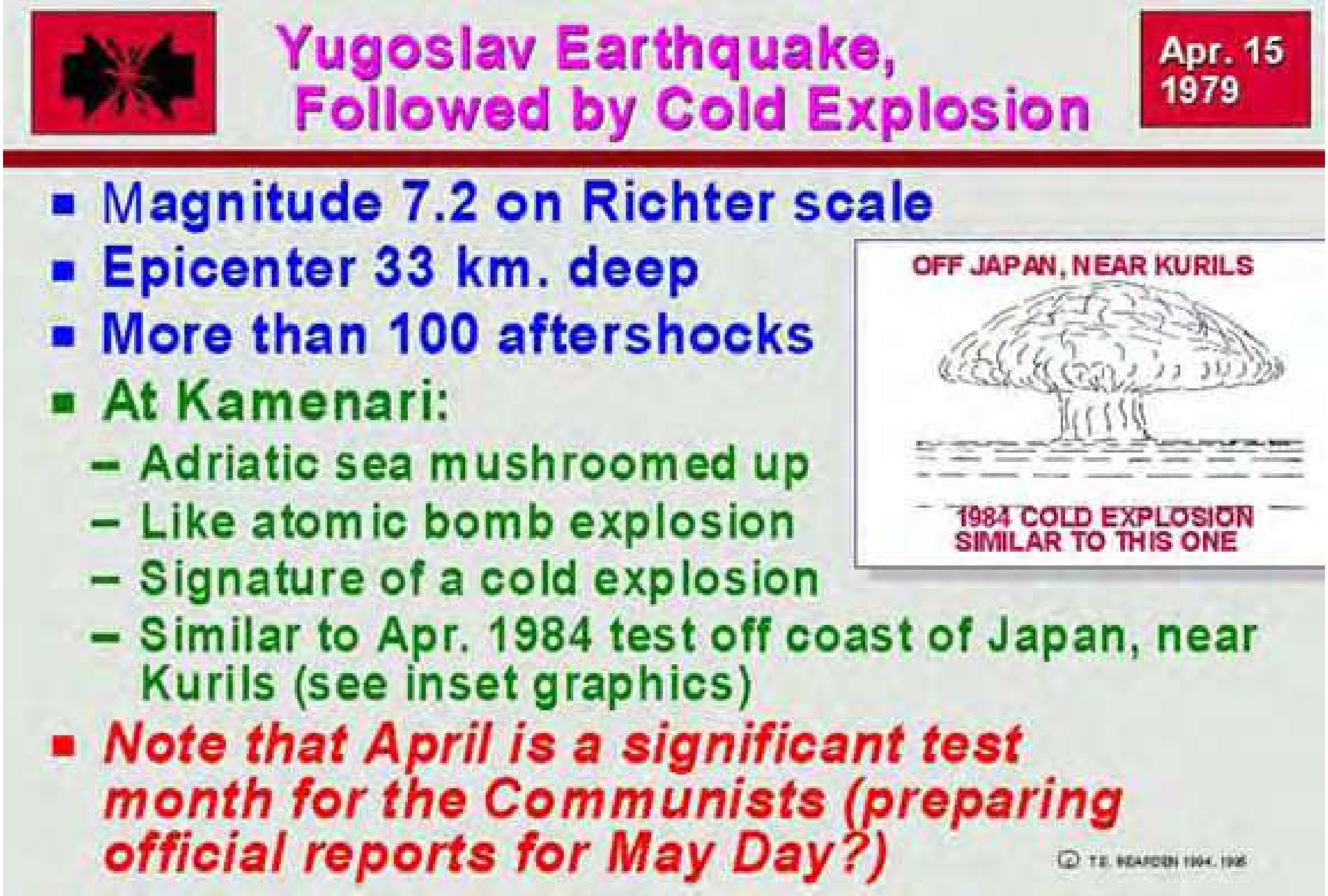

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- *Note that April is a significant test month for the Communists (preparing official reports for May Day?)*

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

Technical Background on the Priore Healing Process

Contents

1. Technical Background on the Priore Healing Process
2. Transduction of EM wave types by successive phase conjugation.
3. Time-charging and Time-density excited states
4. Interference for transduction in slow decay of time-density excited states
5. Infolded longitudinal biwaves composition of a scalar potential
6. Scalar potential interferometry produces all EM force fields and waves
7. Some serious flaws and errors in classical EM theory (#1-16)
8. Some serious flaws and errors in classical EM theory (#17-32)
9. Abnormal penetration depth of weak signals due to photon interaction
10. Amplified replica of body's signal complex to reverse cellular disease & damage
11. Amplified replica of body's signal complex to correct a congenital condition

Technical Background on the Priore Healing Process

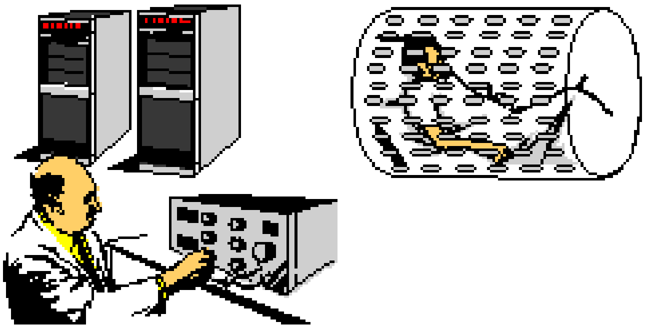

© T.E. BEARDEN 2001

Published by Lotus® Freelance Graphics®

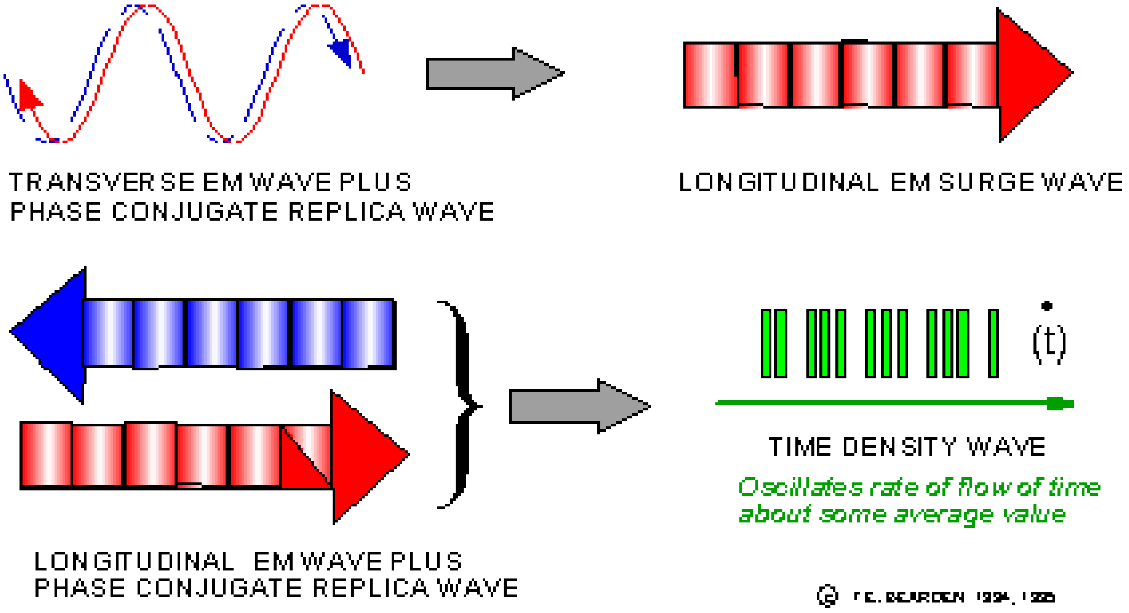
TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
LONGITUDINAL EM SURGE WAVE
LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
$\dot{(t)}$
TIME DENSITY WAVE
Oscillates rate of flow of time
about some average value
© T.E. BEARDEN 1994, 1995

Figure 1. Transduction of EM wave types by successive phase conjugation.

Published by Lotus® Freelance Graphics®

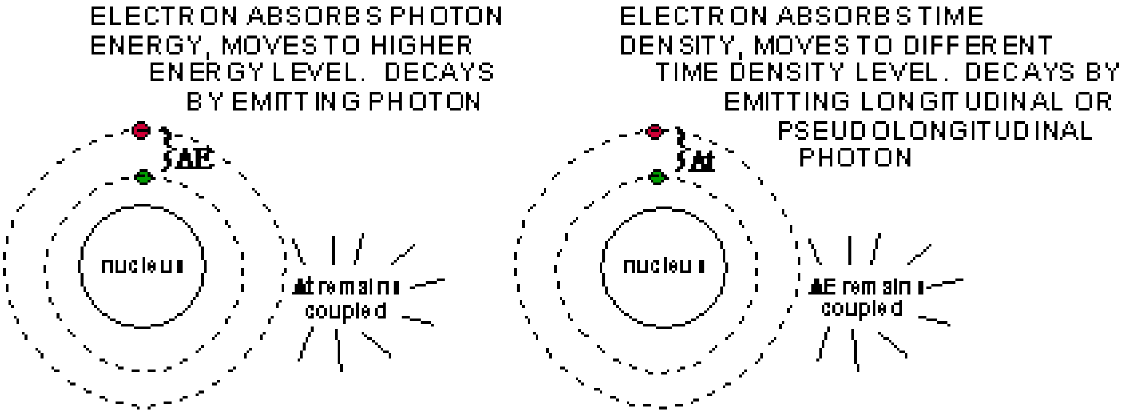
ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON
ΔE
nucleus
Δt remains coupled
ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON
Δt
nucleus
ΔE remains coupled

Figure A. Excited energy density level. Figure B. Excited time density level.

© T.E. BEARDEN 1993

Figure 2. Time-charging and Time-density excited states, contrasted to ordinary spatial energy excited state.

Published by Lotus® Freelance Graphics®

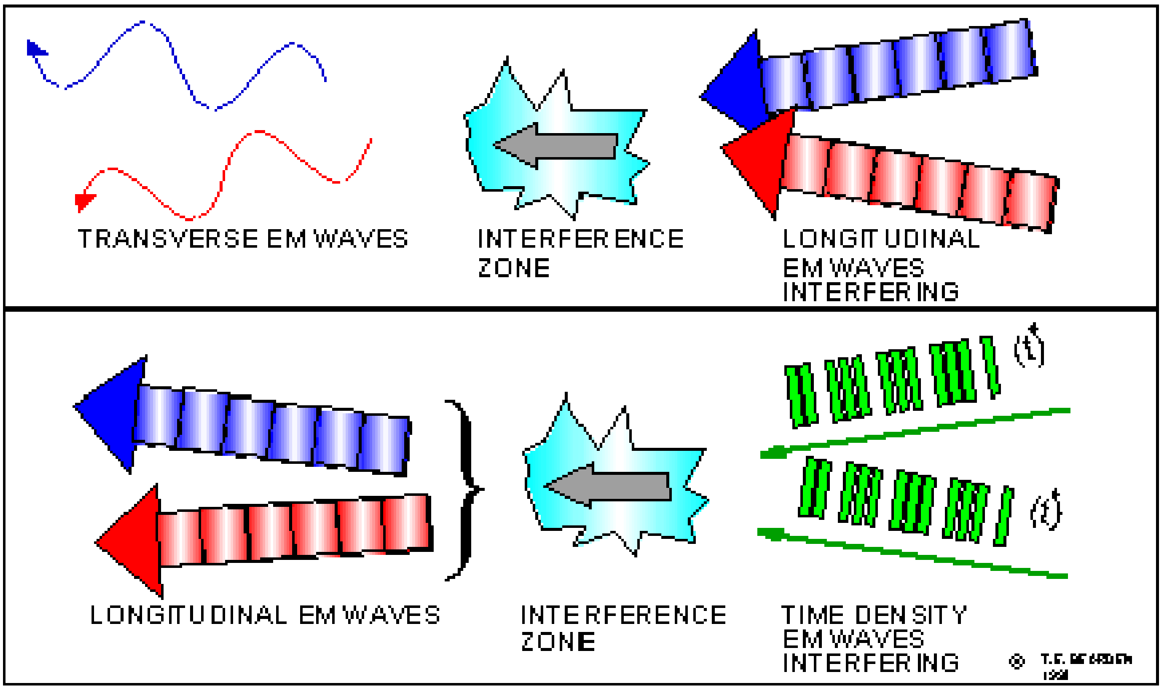
TRANSVERSE EM WAVES
INTERFERENCE ZONE
LONGITUDINAL EM WAVES INTERFERING
(t)
(t)
LONGITUDINAL EM WAVES
INTERFERENCE ZONE
TIME DENSITY EM WAVES INTERFERING

Figure 3. Interference processes for transduction in slow decay of time-density excited states.

Published by Lotus® Freelance Graphics®

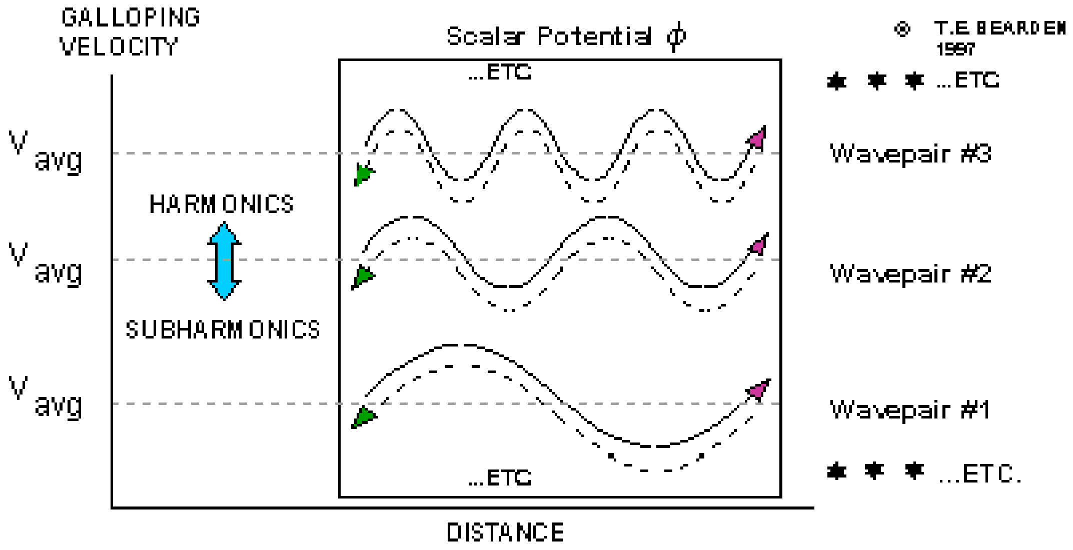
GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC
✱ ✱ ✱ ...ETC
V avg
Wavepair #3
HARMONICS
V avg
Wavepair #2
SUBHARMONICS
V avg
Wavepair #1
...ETC
✱ ✱ ✱ ...ETC.
DISTANCE

The Structure Is:

A harmonic set of longitudinal wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. Each wavepair is a time-density wave and therefore an electrogravitational wave.

Figure 4. Infolded longitudinal biwaves (time-density EM waves) composition of a scalar potential.

Published by Lotus® Freelance Graphics®

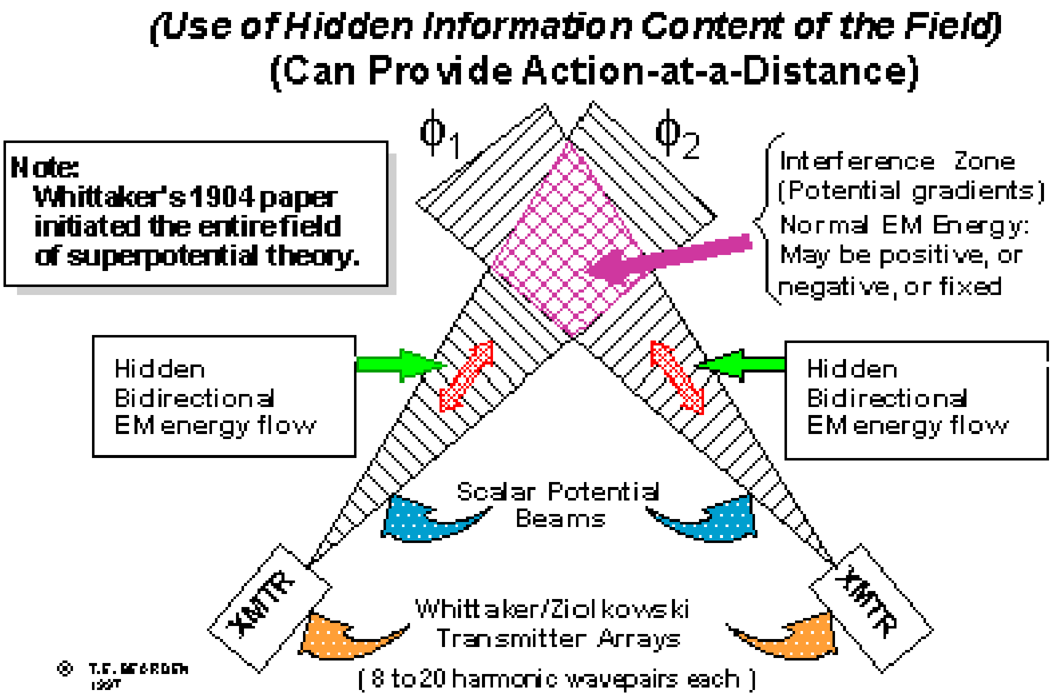
(Use of Hidden Information Content of the Field)
(Can Provide Action-at-a-Distance)
ϕ1
ϕ2
Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.
Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)
© T.E. BEARDEN 1997

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepairfunctions) produces all EM force fields and waves.

Published by Lotus® Freelance Graphics®

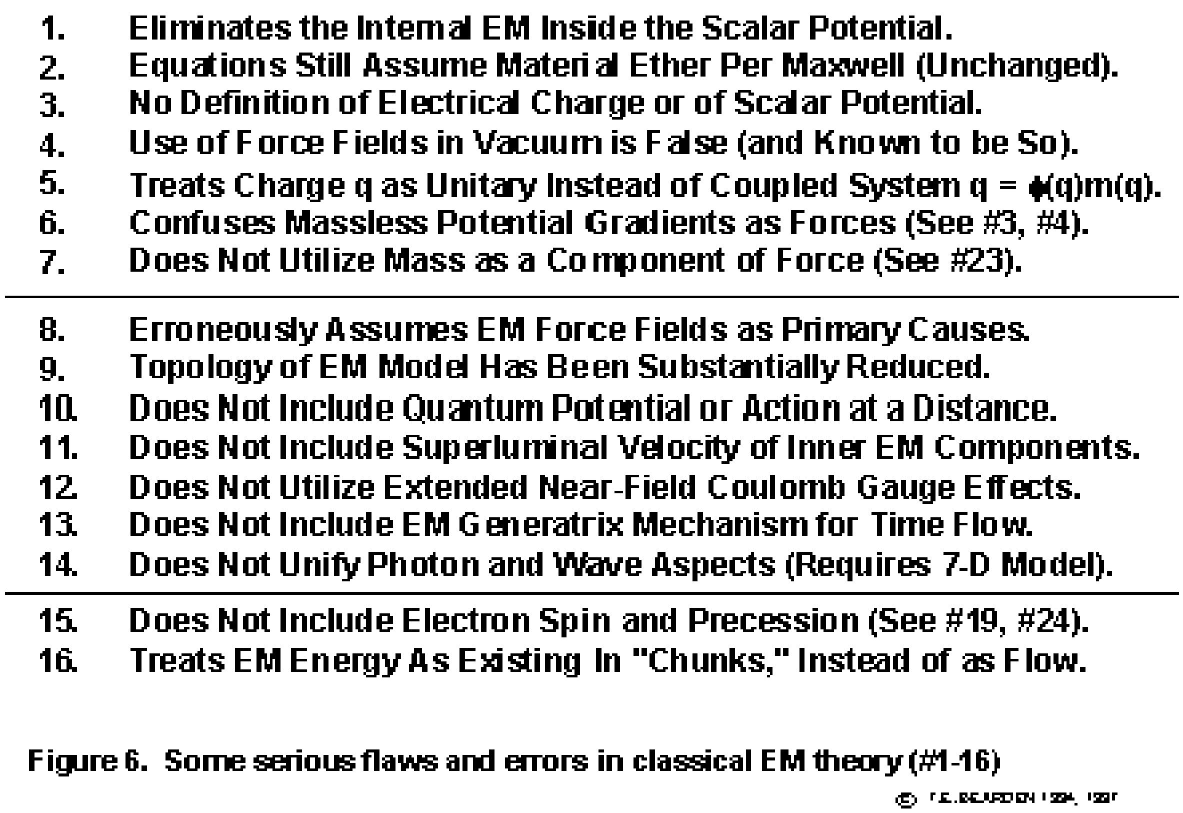

1. Eliminates the Internal EM Inside the Scalar Potential.
2. Equations Still Assume Material Ether Per Maxwell (Unchanged).
3. No Definition of Electrical Charge or of Scalar Potential.
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi(q)m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.

Figure 6. Some serious flaws and errors in classical EM theory (#1-16)

© T.E. BEARDEN 1994, 1997

Published by Lotus® Freelance Graphics®

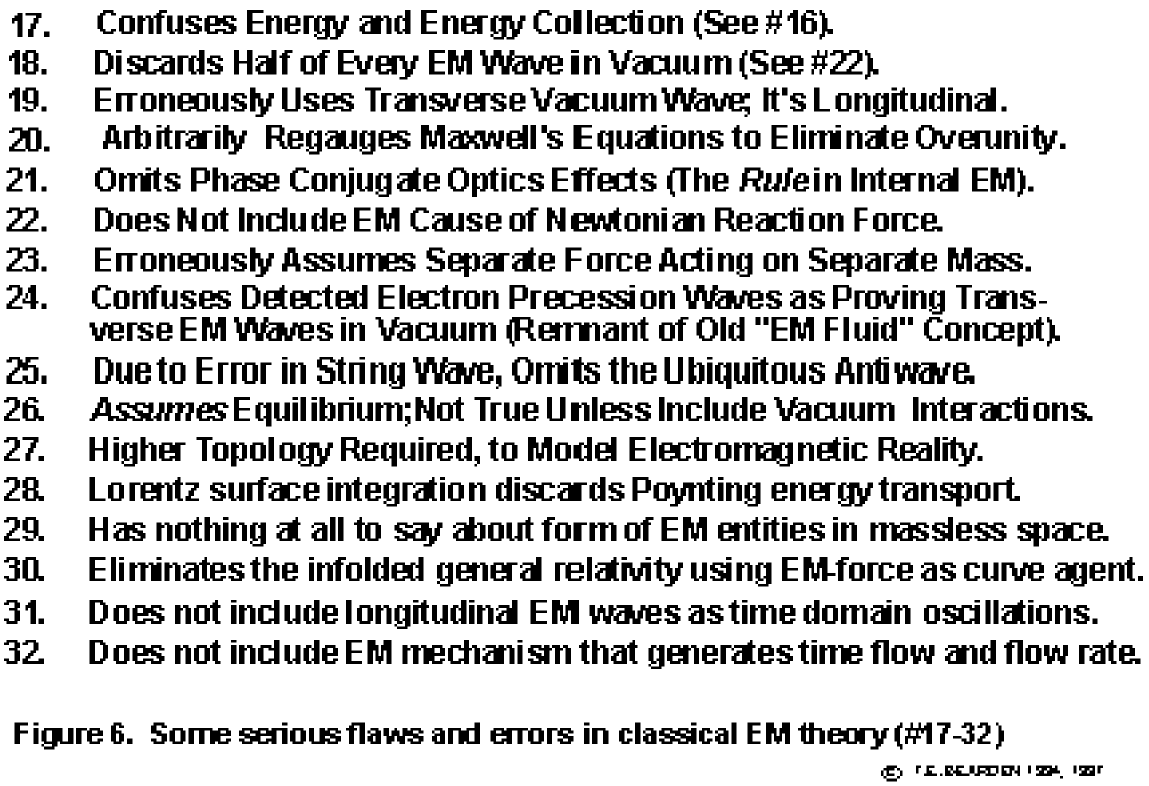

17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

Figure 6. Some serious flaws and errors in classical EM theory (#17-32)

© T.E. BEARDEN 1994, 1997

Published by Lotus® Freelance Graphics®

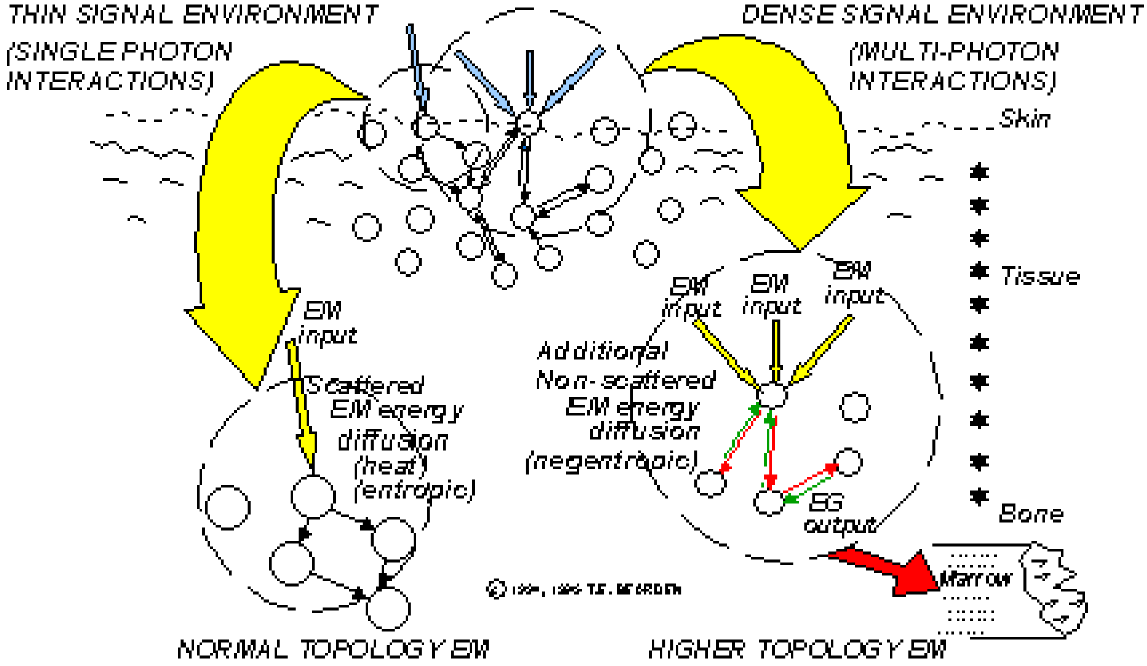
THIN SIGNAL ENVIRONMENT
(SINGLE PHOTON INTERACTIONS)
DENSE SIGNAL ENVIRONMENT
(MULTI-PHOTON INTERACTIONS)
Skin
Tissue
Bone
Marrow
EM input
Scattered EM energy diffusion (heat) (entropic)
EM input
EM input
EM input
Additional Non-scattered EM energy diffusion (negentropic)
EG output
NORMAL TOPOLOGY EM
HIGHER TOPOLOGY EM

In a dense signal environment, phase conjugation effects become significant, generating PCR signals which couple to and backtrack the cells' and tissues' own signals, thus penetrating deeply throughout the body with weak partial cancellation jamming.

Figure 30. Abnormal penetration depth of incident weak signals due to multiphoton interaction in a dense weak signal environment.

Published by Lotus® Freelance Graphics®

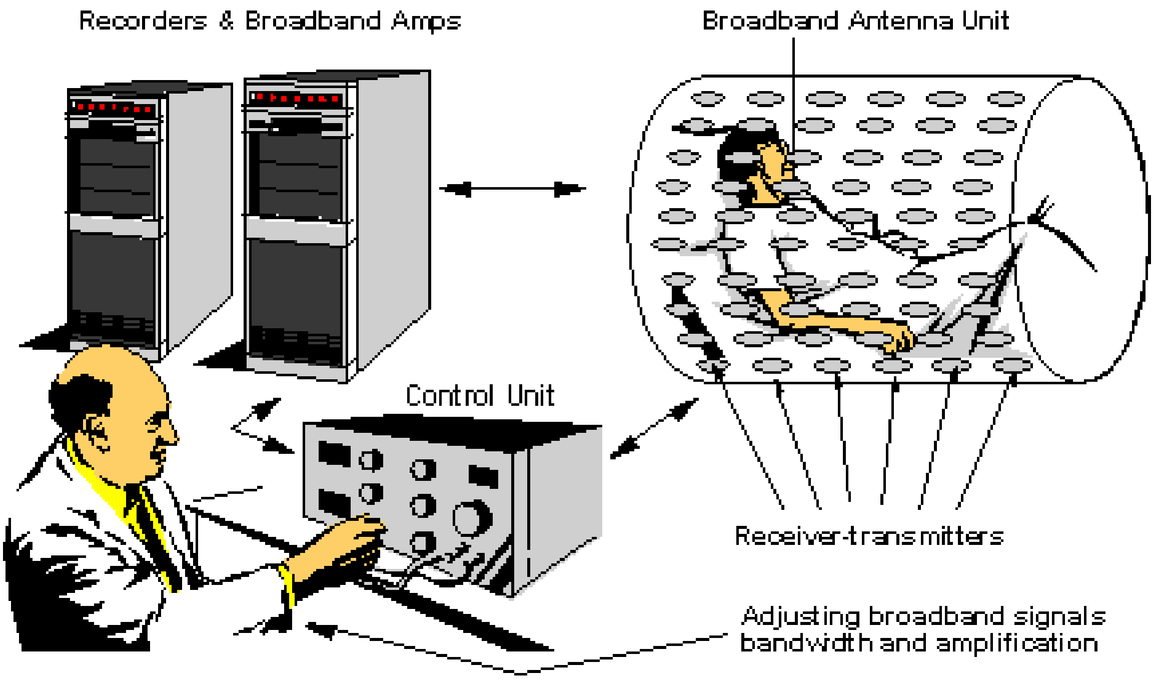
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
Adjusting broadband signals bandwidth and amplification

Figure 31. Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage.

© T.E. BEARDEN 1996

Published by Lotus® Freelance Graphics®

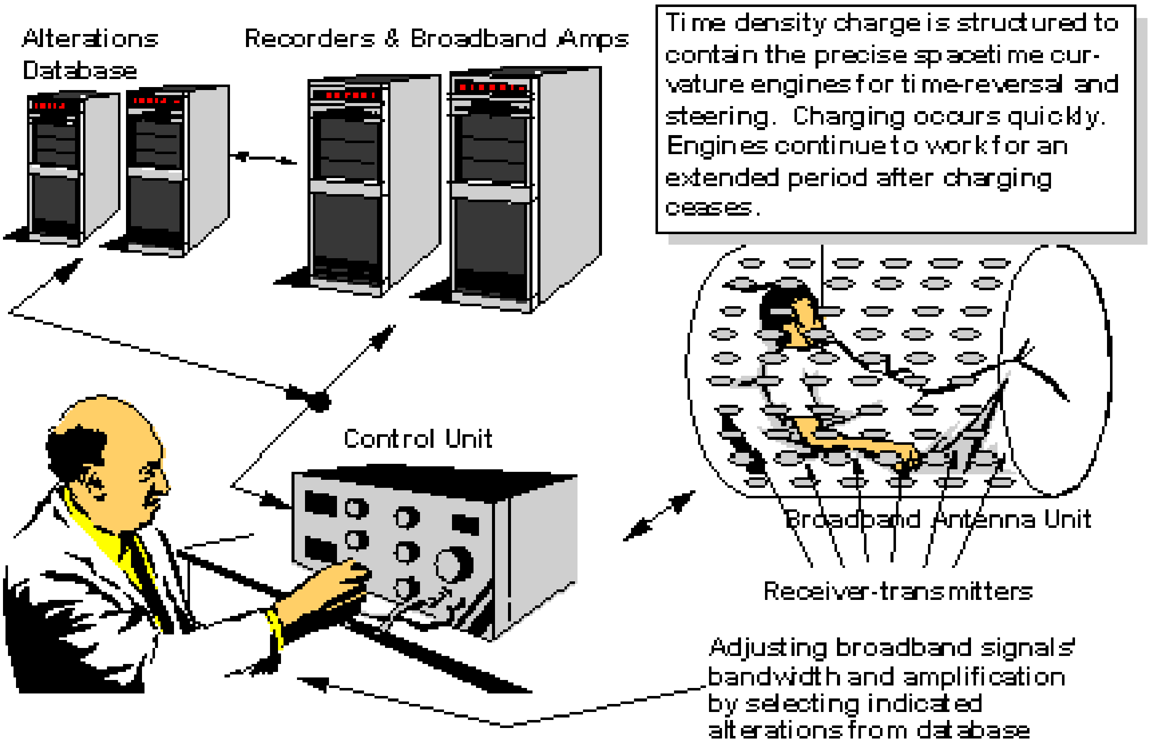
Alterations Database
Recorders & Broadband Amps
Time density charge is structured to contain the precise spacetime curvature engines for time-reversal and steering. Charging occurs quickly. Engines continue to work for an extended period after charging ceases.
Control Unit
Broadband Antenna Unit
Receiver-transmitters
Adjusting broadband signals' bandwidth and amplification by selecting indicated alterations from database

Figure 9. Using an amplified replica of the body's own dense, broadband signal complex containing added alterations, to irradiate the body and produce deeply-penetrating precise time-density charging for cellular time-reversal, while steering the time-reversal trajectory to also correct a congenital condition.

© T.E. BEARDEN [illegible]

Published by Lotus® Freelance Graphics®

San Francisco
R-2508
China Lake
Point Mugu
Sea Range
Ft. Irwin
Edwards AFB
Los Angeles
San Diego
Santa Fe
Albuquerque
Missile Range
White Sands
Las Cruces

The Anomalies in Navy Electrolyte Experiments at China Lake

These novel experimental anomalies, like those of Cold Fusion, were caused by using time as an energy source by transducing time-energy into spatial energy. Transducing one microsecond per second of time into spatial EM energy, yields nearly 10^{11} watts of power

- Melvin H. Miles and Benjamin F. Bush, "Radiation measurements at China Lake: Real or Artifacts?", *Proc. ICCF-7 (International Conference on Cold Fusion—7)*, Vancouver, BC, Canada, Apr. 1998, p. 101.

- Consistency of China Lake Anomalies
- Anomalous LW EM Emission Phenomena at China Lake (1)
- Anomalous LW EM Emission Phenomena at China Lake (2)
- Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (3)
- Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (4)
- Consistency of Anomalies
- Consistency of Anomalies (2)

- Experimental LW Emission Phenomena in Electrolysis

Next Slide Return to Index Previous Slide

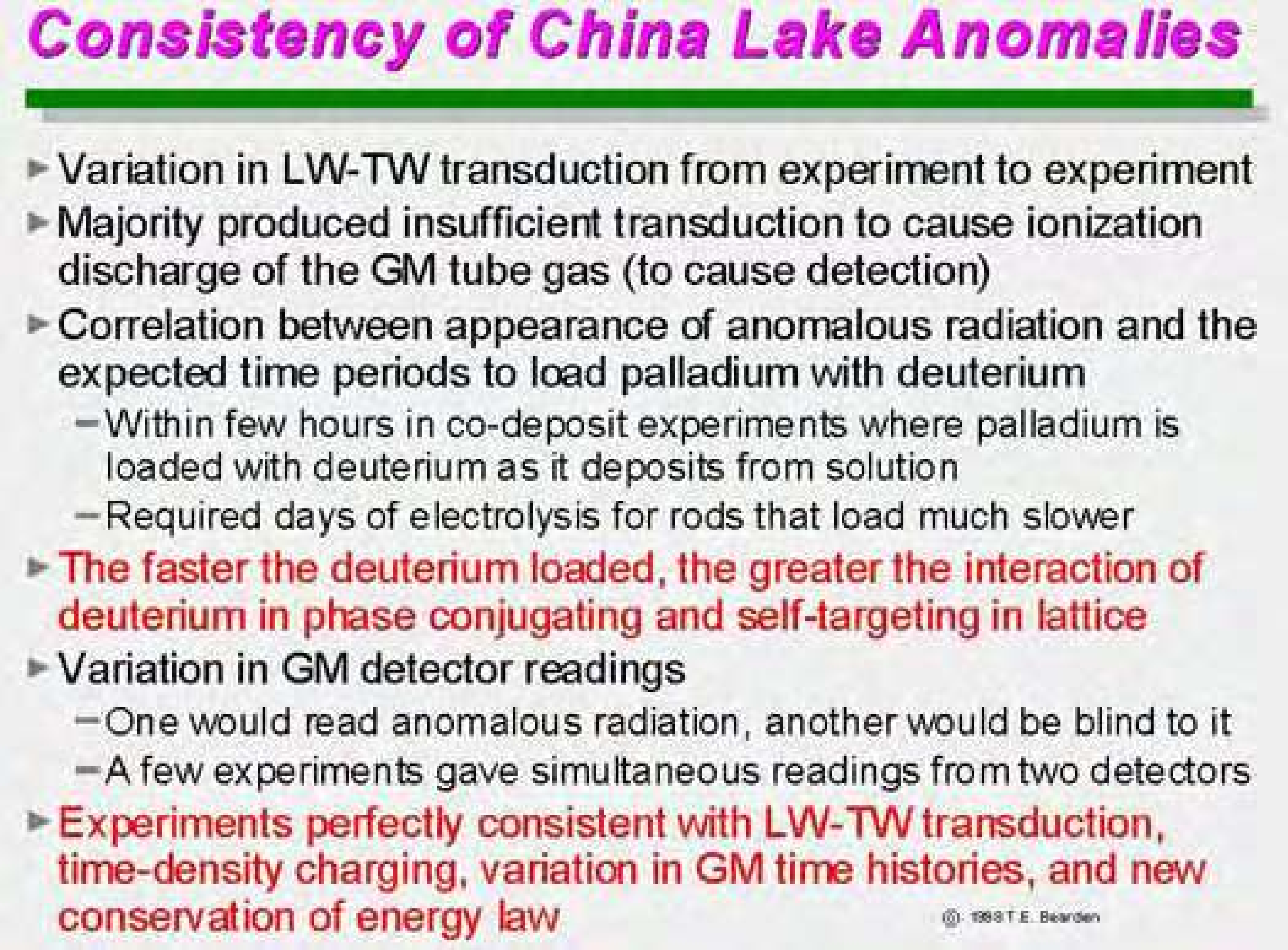

Consistency of China Lake Anomalies

- Variation in LW-TW transduction from experiment to experiment
- Majority produced insufficient transduction to cause ionization discharge of the GM tube gas (to cause detection)
- Correlation between appearance of anomalous radiation and the expected time periods to load palladium with deuterium
  - Within few hours in co-deposit experiments where palladium is loaded with deuterium as it deposits from solution
  - Required days of electrolysis for rods that load much slower
- The faster the deuterium loaded, the greater the interaction of deuterium in phase conjugating and self-targeting in lattice
- Variation in GM detector readings
  - One would read anomalous radiation, another would be blind to it
  - A few experiments gave simultaneous readings from two detectors
- Experiments perfectly consistent with LW-TW transduction, time-density charging, variation in GM time histories, and new conservation of energy law

© 1993 T.E. Bearden

[Next Slide](#)

[Return to Index](#)

Next Slide Return to Index Previous Slide

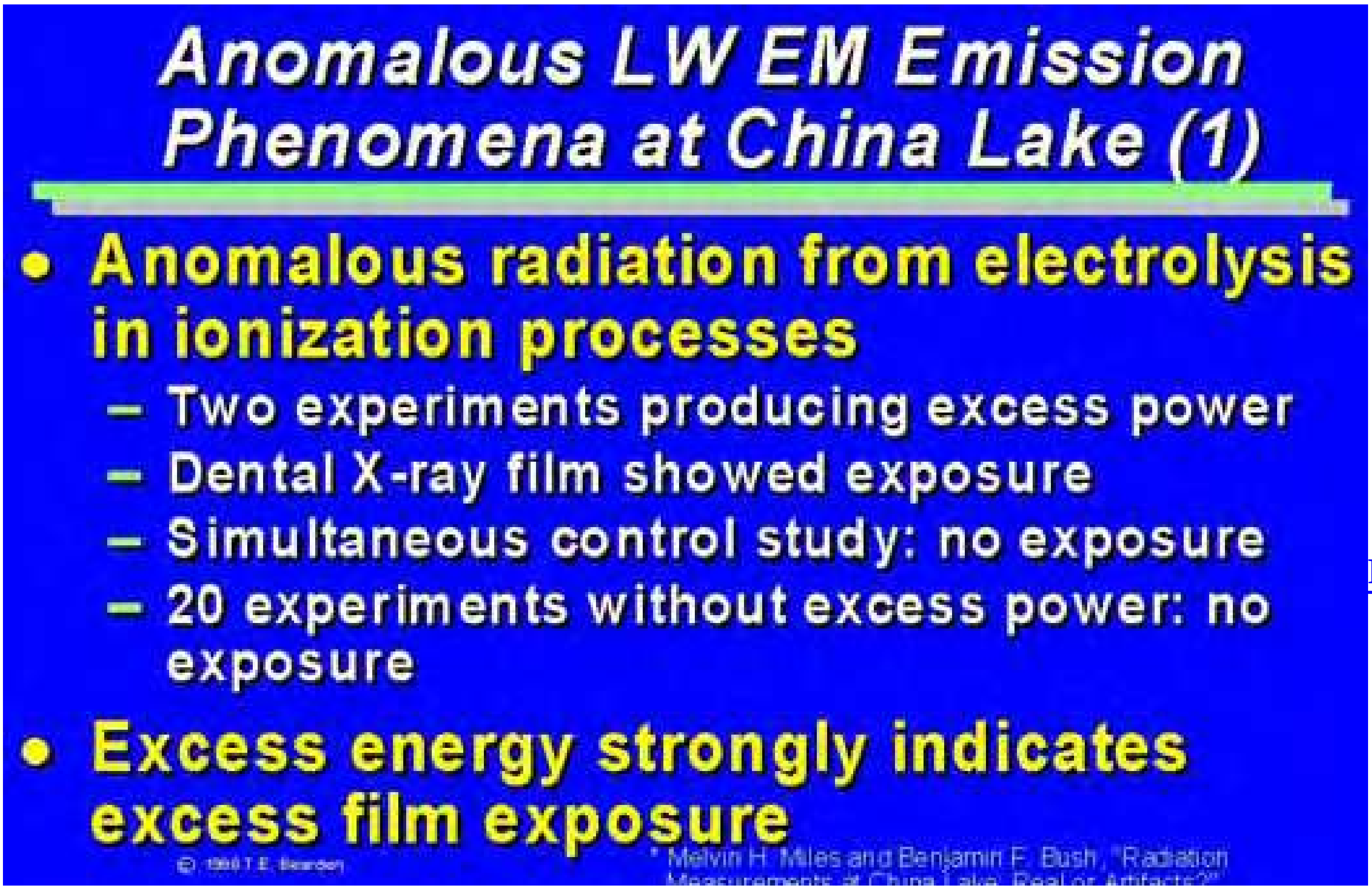

Anomalous LW EM Emission Phenomena at China Lake (1)

- Anomalous radiation from electrolysis in ionization processes
  - Two experiments producing excess power
  - Dental X-ray film showed exposure
  - Simultaneous control study: no exposure
  - 20 experiments without excess power: no exposure
- Excess energy strongly indicates excess film exposure

© 1998 T.E. Bearden

* Melvin H. Miles and Benjamin F. Bush, "Radiation Measurements at China Lake, Real or Artifacts?"

Next Slide

Return to Index

Previous Slide

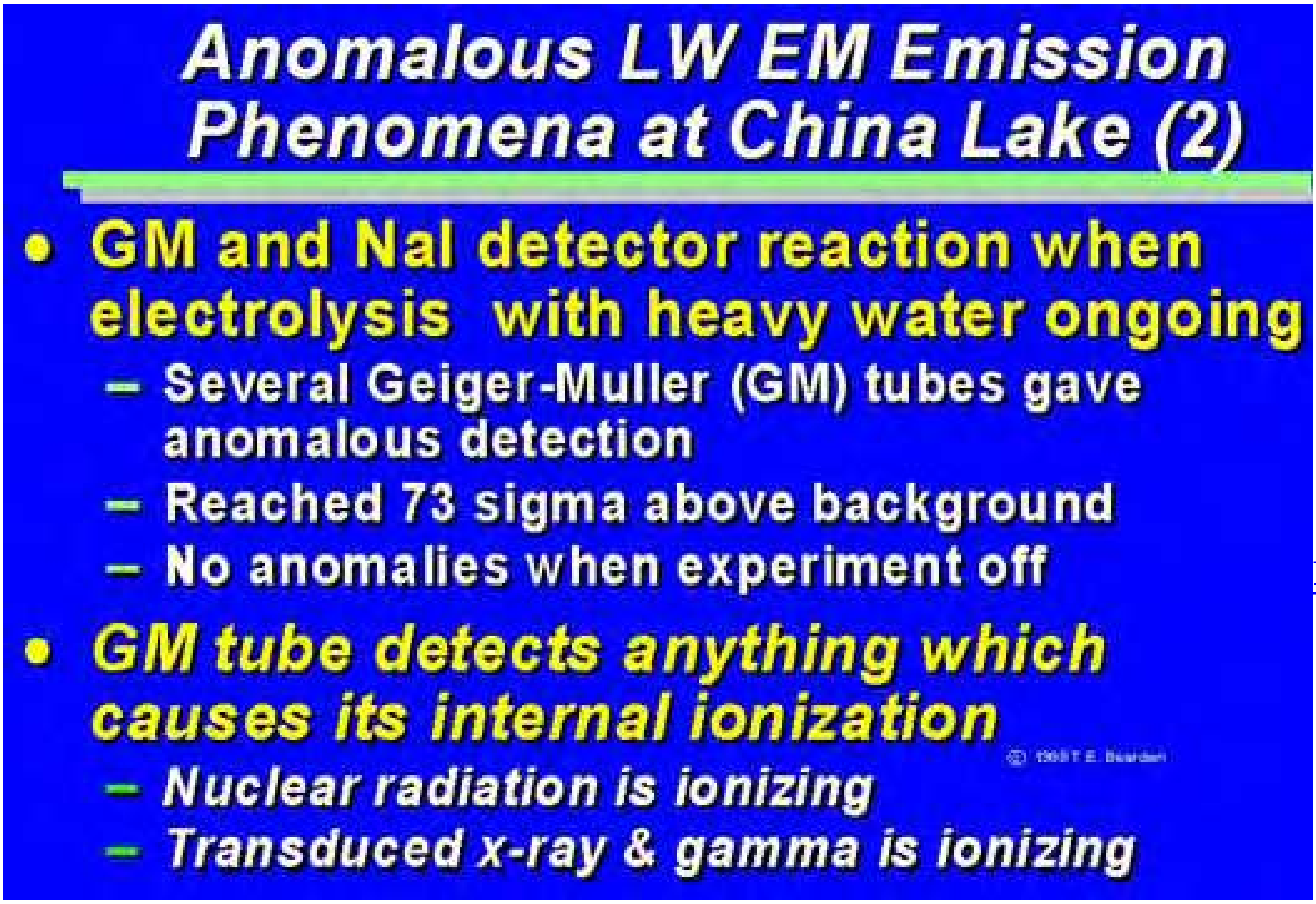

Anomalous LW EM Emission Phenomena at China Lake (2)

- GM and NaI detector reaction when electrolysis with heavy water ongoing
  - Several Geiger-Muller (GM) tubes gave anomalous detection
  - Reached 73 sigma above background
  - No anomalies when experiment off
- *GM tube detects anything which causes its internal ionization*
  - *Nuclear radiation is ionizing*
  - *Transduced x-ray & gamma is ionizing*

© [illegible] T. E. Bearden

Next Slide

Return to Index

Previous Slide

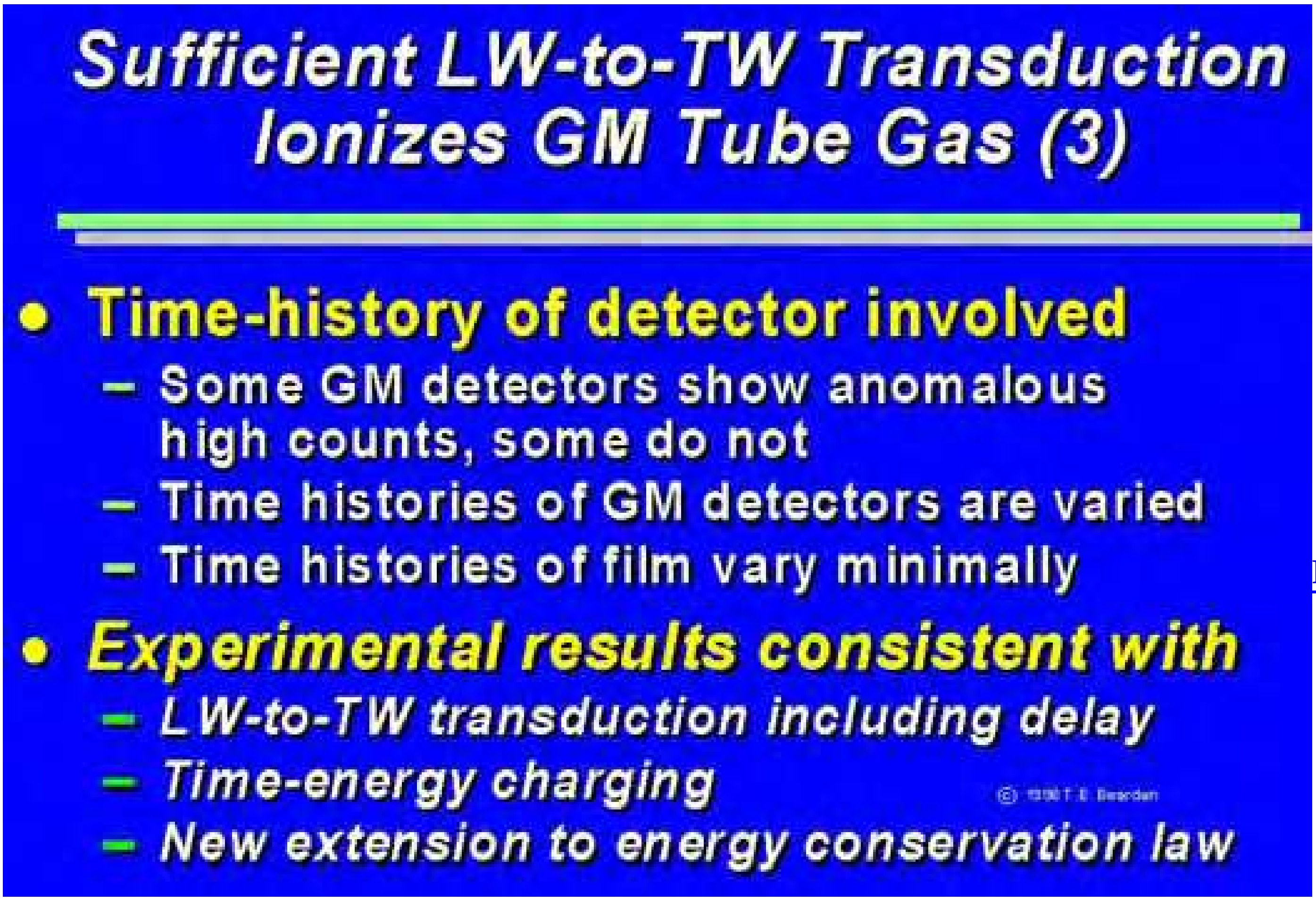

Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (3)

- **Time-history of detector involved**
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors are varied
  - Time histories of film vary minimally
- ***Experimental results consistent with***
  - *LW-to-TW transduction including delay*
  - *Time-energy charging*
  - *New extension to energy conservation law*

© [illegible] T. E. Bearden

Next Slide

Return to Index

Previous Slide

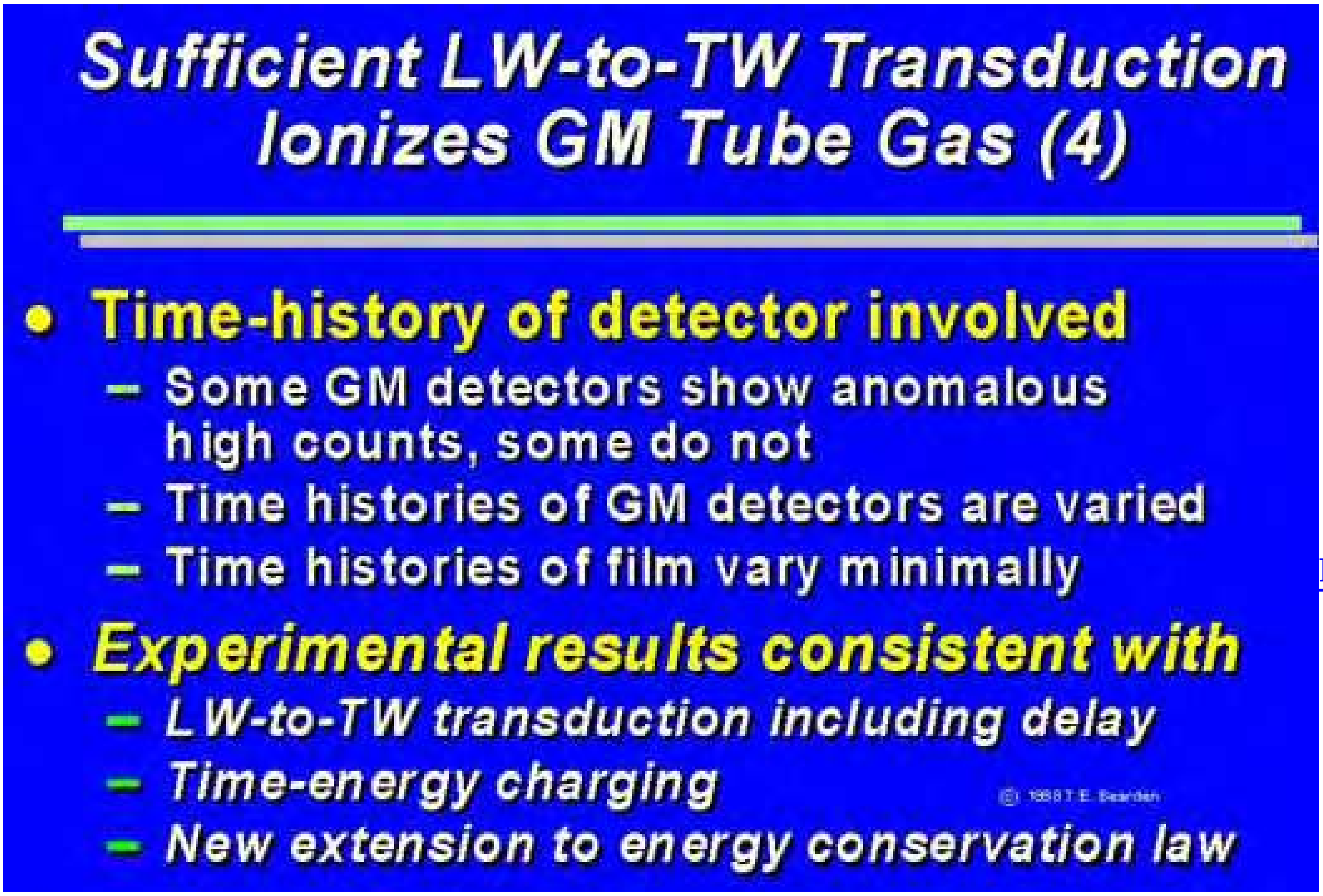

Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (4)

- **Time-history of detector involved**
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors are varied
  - Time histories of film vary minimally
- ***Experimental results consistent with***
  - *LW-to-TW transduction including delay*
  - *Time-energy charging*
  - *New extension to energy conservation law*

© 1988 T.E. Bearden

Next Slide

Return to Index

Previous Slide

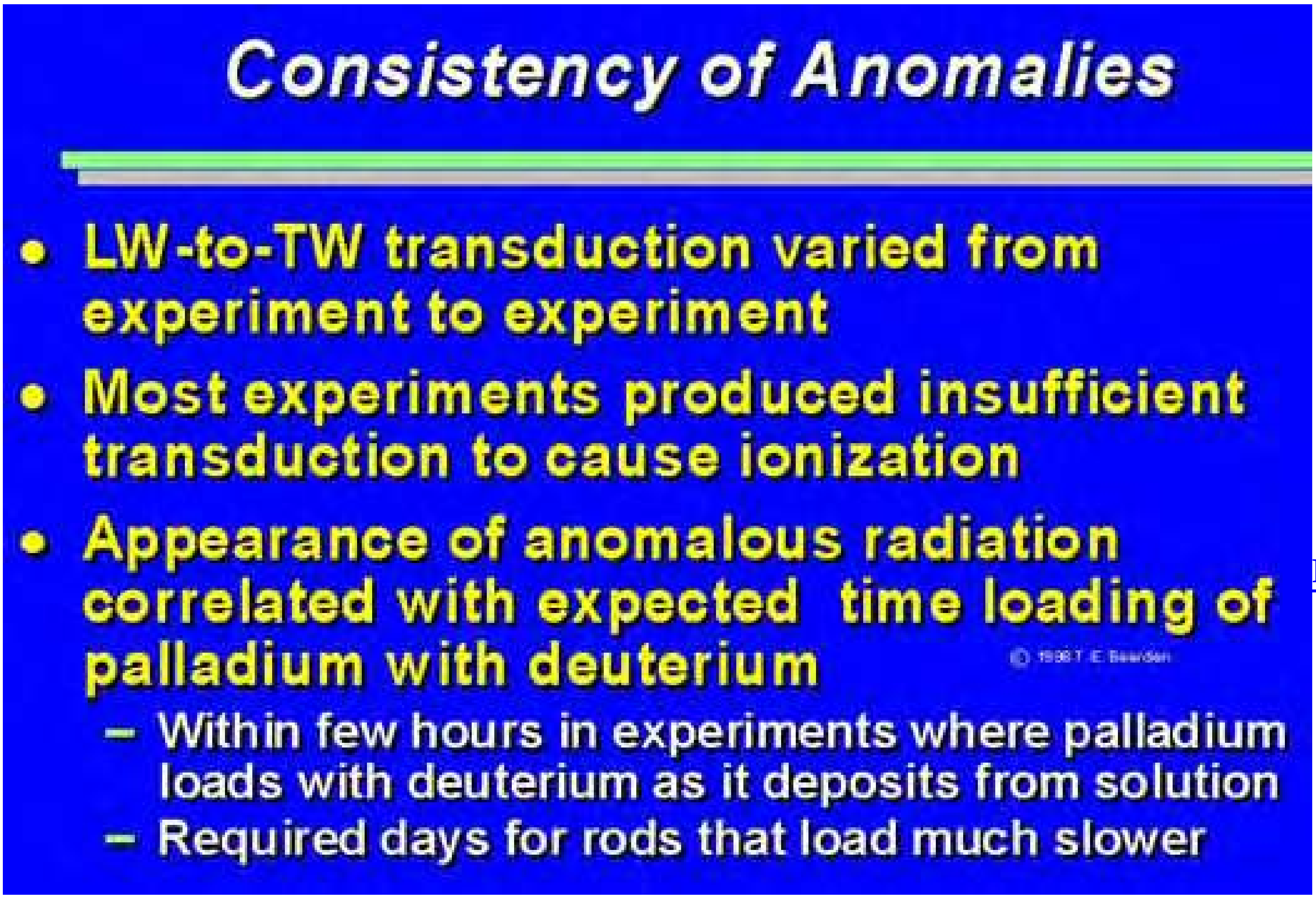

Consistency of Anomalies

- LW-to-TW transduction varied from experiment to experiment
- Most experiments produced insufficient transduction to cause ionization
- Appearance of anomalous radiation correlated with expected time loading of palladium with deuterium
  - Within few hours in experiments where palladium loads with deuterium as it deposits from solution
  - Required days for rods that load much slower

© 1998 T. E. Bearden

Next Slide

Return to Index

Previous Slide

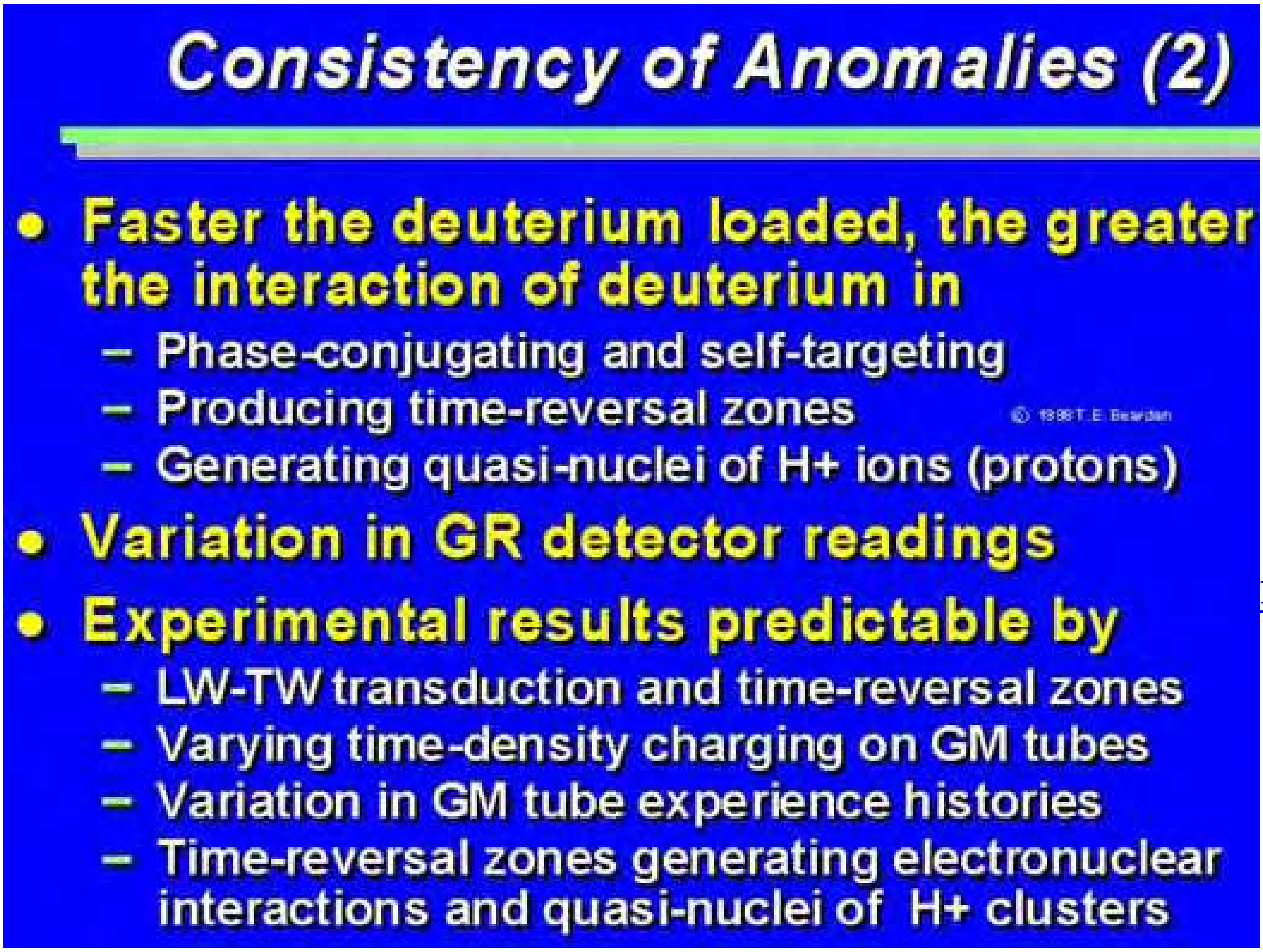

Consistency of Anomalies (2)

- Faster the deuterium loaded, the greater the interaction of deuterium in
  - Phase-conjugating and self-targeting
  - Producing time-reversal zones
  - Generating quasi-nuclei of H+ ions (protons)
- Variation in GR detector readings
- Experimental results predictable by
  - LW-TW transduction and time-reversal zones
  - Varying time-density charging on GM tubes
  - Variation in GM tube experience histories
  - Time-reversal zones generating electronuclear interactions and quasi-nuclei of H+ clusters

© 1998 T.E. Bearden

Next Slide

Return to Index

Previous Slide

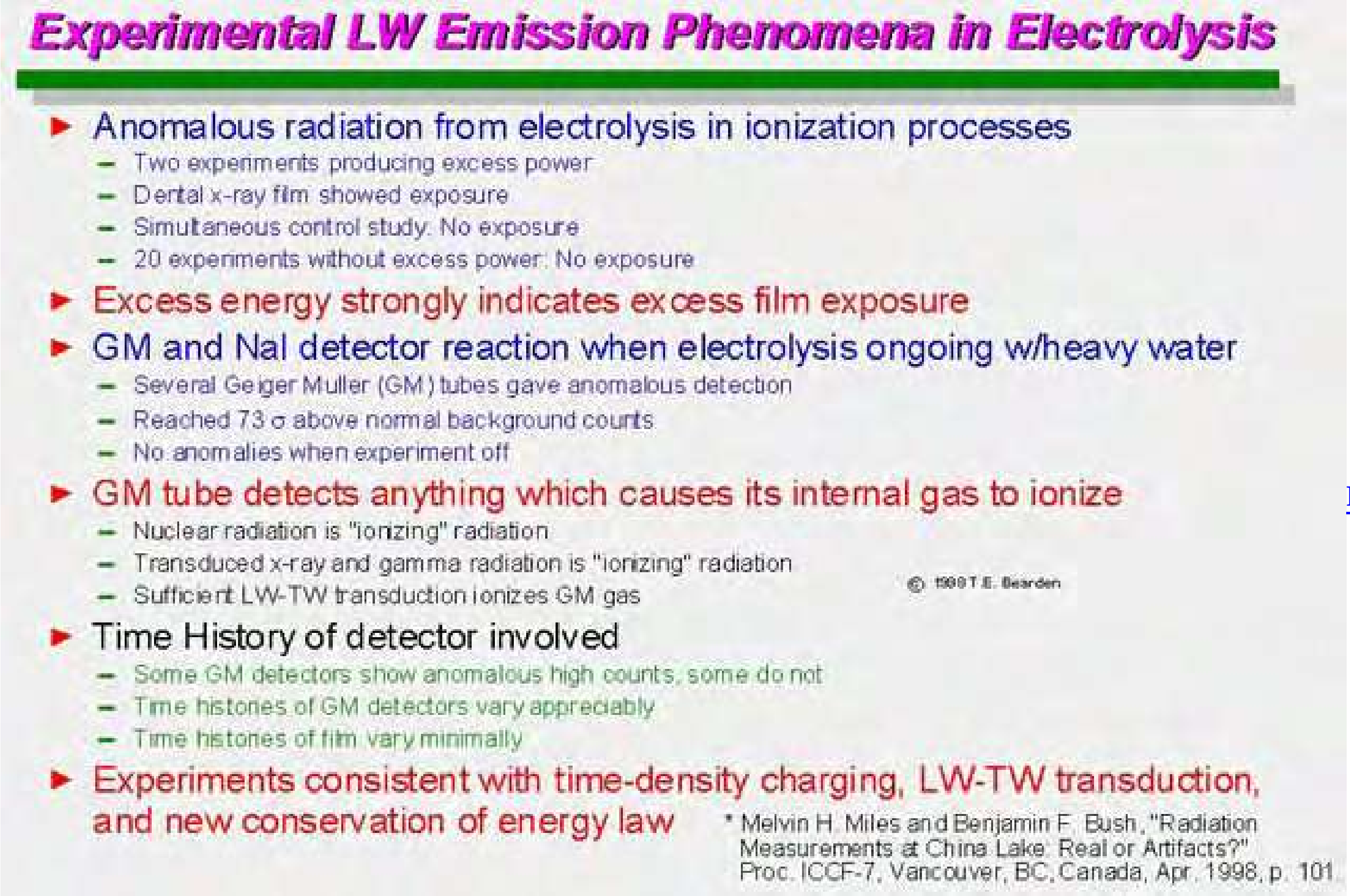

Experimental LW Emission Phenomena in Electrolysis

- Anomalous radiation from electrolysis in ionization processes
  - Two experiments producing excess power
  - Dental x-ray film showed exposure
  - Simultaneous control study: No exposure
  - 20 experiments without excess power: No exposure
- Excess energy strongly indicates excess film exposure
- GM and NaI detector reaction when electrolysis ongoing w/heavy water
  - Several Geiger Muller (GM) tubes gave anomalous detection
  - Reached 73 σ above normal background counts
  - No anomalies when experiment off
- GM tube detects anything which causes its internal gas to ionize
  - Nuclear radiation is "ionizing" radiation
  - Transduced x-ray and gamma radiation is "ionizing" radiation
  - Sufficient LW-TW transduction ionizes GM gas
- Time History of detector involved
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors vary appreciably
  - Time histories of film vary minimally
- Experiments consistent with time-density charging, LW-TW transduction, and new conservation of energy law

© 1998 T.E. Bearden

* Melvin H. Miles and Benjamin F. Bush, "Radiation Measurements at China Lake: Real or Artifacts?" Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 101

Return to Index

Previous Slide

A Visual Tour of what they don't want you to know about electrical circuits

10 TRILLION PERCENT OF THE CURRENT PRODUCED IS WASTED!

---

1. The scalar potential is a harmonic set of phase conjugate longitudinal EM waves.
2. The dipole is a true negative resistor.
3. The dipole's broken 3-symmetry initiates a spreading giant negentropic reordering of a fraction of the vacuum's energy.
4. Energy flow contours surrounding a transmission line.
5. Heaviside and Poynting energy flow components. The Heaviside component is often **10 trillion times** the Poynting component, but **is simply wasted** in ordinary single-pass energy circuits.
6. Negative resistance process v. positive resistance process.
7. Lorentz's integration trick to discard the enormous Heaviside non-diverged energy flow component.

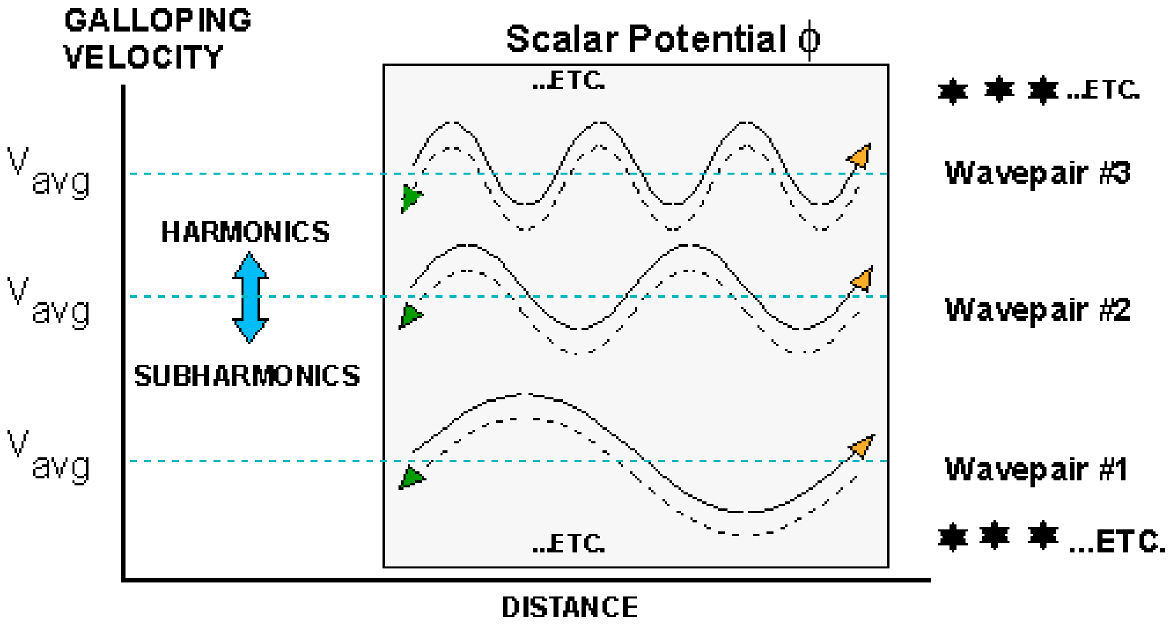

Next Slide

Return to Index

The Structure Is:

- A harmonic set of longitudinal phase conjugate wavepairs.
- In each wavepair the two waves superpose spatially after detection, but travel in opposite directions.
- The convergent wave set is in the imaginary plane, and hence is not observable. It is EM energy incoming to the potential (dipolarity) from the time domain.
- The charge's spin is 720 degrees, 360 in the real plane and 360 in the imaginary plane.
- Hence the charge receives the complex convergent EM energy, transduces it into real EM energy, and emits enormous energy at the speed of light in all directions -- which includes bidirectional pairs in 3-space (after the reaction, being after "observation").
- This produces the fields and potentials from the "source charge or dipole."
- Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus their quantum field theory work strongly supports the "negative resistor" interpretation of the scalar potential and our solution to the source charge and source dipole problem.

Figure 1. The scalar potential is a harmonic set of phase conjugate longitudinal EM waves.

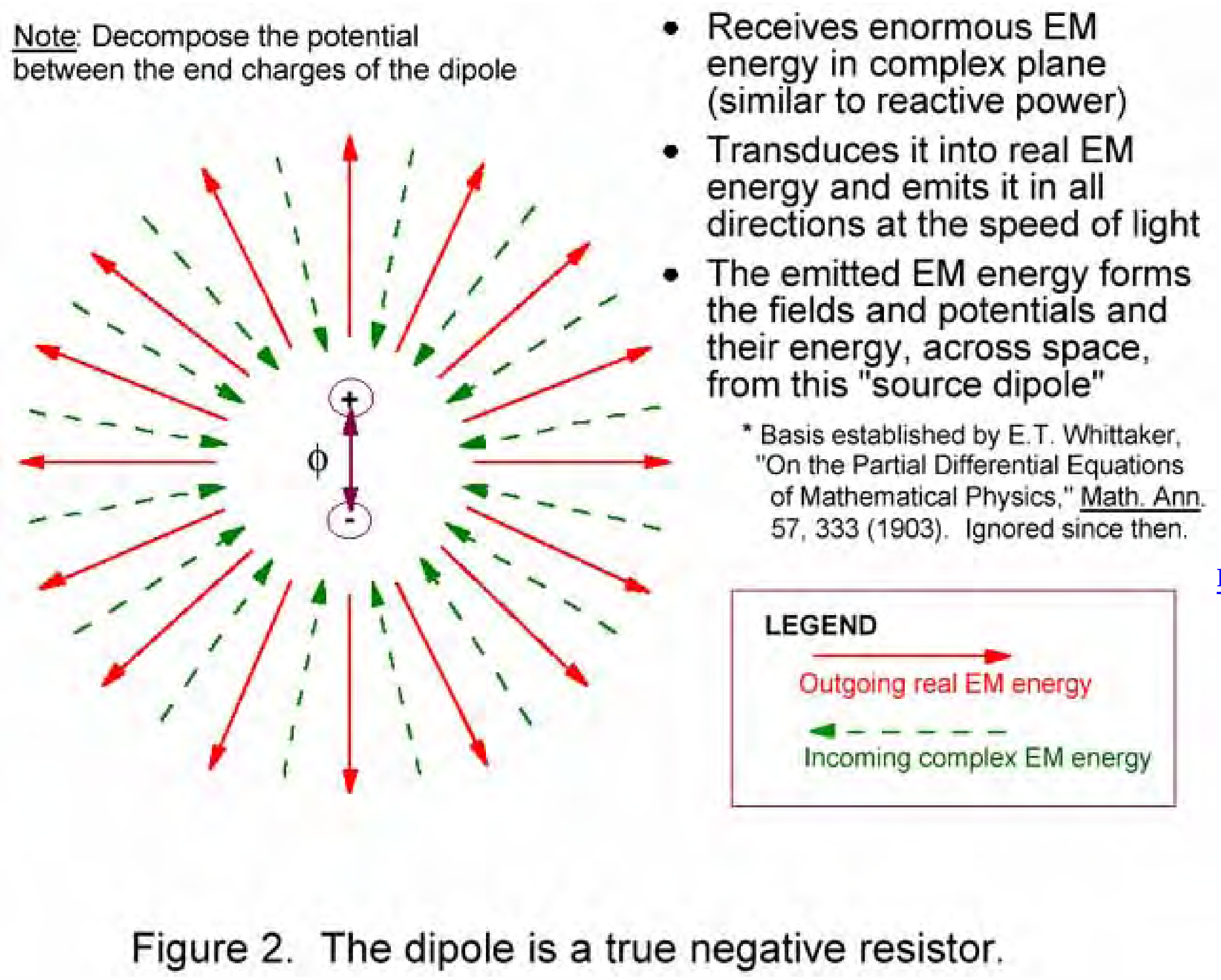

Note: Decompose the potential between the end charges of the dipole

- Receives enormous EM energy in complex plane (similar to reactive power)
- Transduces it into real EM energy and emits it in all directions at the speed of light
- The emitted EM energy forms the fields and potentials and their energy, across space, from this "source dipole"

* Basis established by E.T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Math. Ann. 57, 333 (1903). Ignored since then.

LEGEND

Outgoing real EM energy

Incoming complex EM energy

Figure 2. The dipole is a true negative resistor.

Next Slide

Return to Index

Previous Slide

Broken 3-space symmetry initiates jump to 4-space symmetry between complex plane and real plane. Energy flow is now conserved in 4-space, but not in 3-space. This is the true negative resistor effect, and a negentropic reordering of the vacuum.

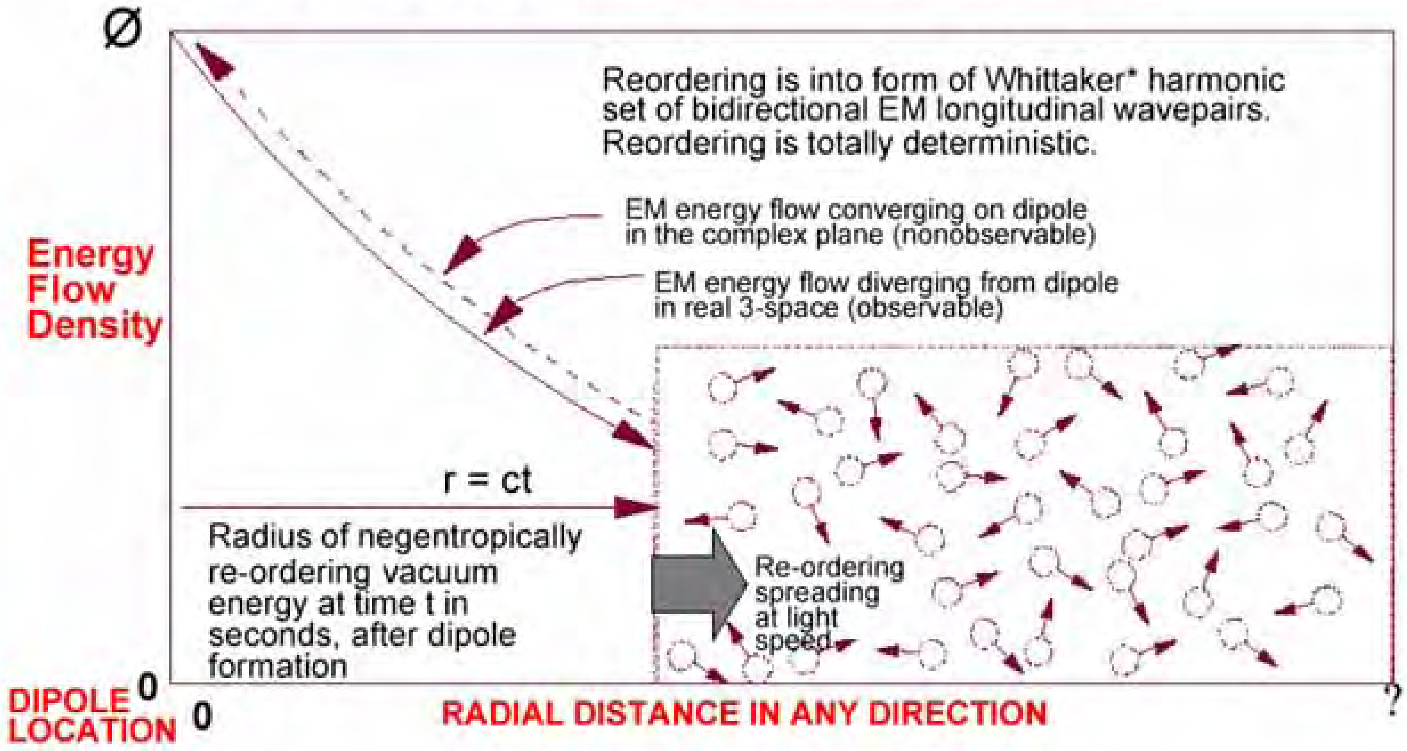

Reordering is into form of Whittaker* harmonic set of bidirectional EM longitudinal wavepairs. Reordering is totally deterministic.

EM energy flow converging on dipole in the complex plane (nonobservable)

EM energy flow diverging from dipole in real 3-space (observable)

Radius of negentropically re-ordering vacuum energy at time t in seconds, after dipole formation

Re-ordering spreading at light speed

Figure 3. The dipole's broken 3-symmetry initiates a spreading giant negentropic reordering of a fraction of the vacuum's energy.

Next Slide

Return to Index

Previous Slide

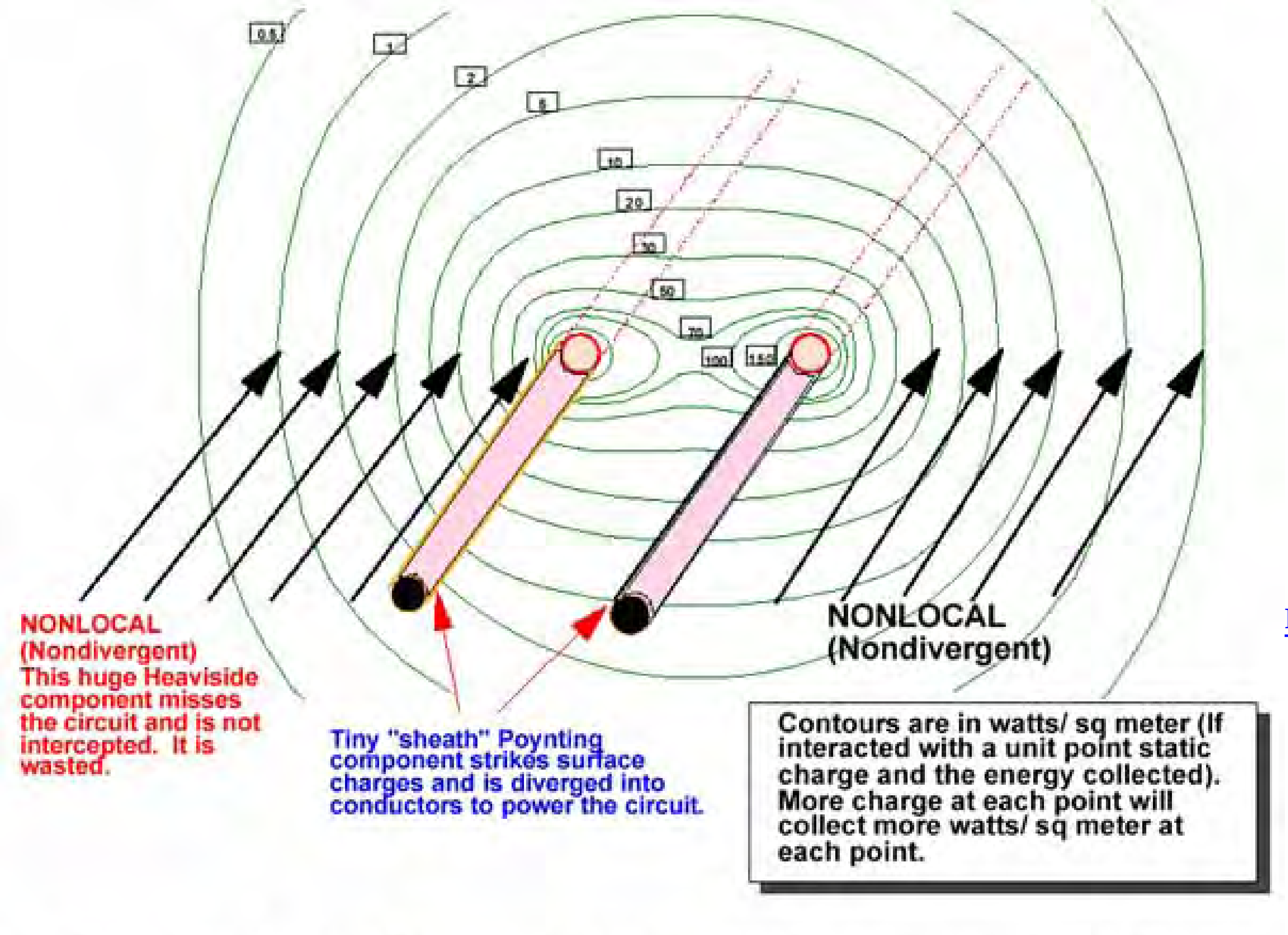

NONLOCAL (Nondivergent) This huge Heaviside component misses the circuit and is not intercepted. It is wasted.

Tiny "sheath" Poynting component strikes surface charges and is diverged into conductors to power the circuit.

NONLOCAL (Nondivergent)

Contours are in watts/ sq meter (If interacted with a unit point static charge and the energy collected). More charge at each point will collect more watts/ sq meter at each point.

Figure 4. Energy flow contours surrounding a transmission line.

Next Slide

Return to Index

Previous Slide

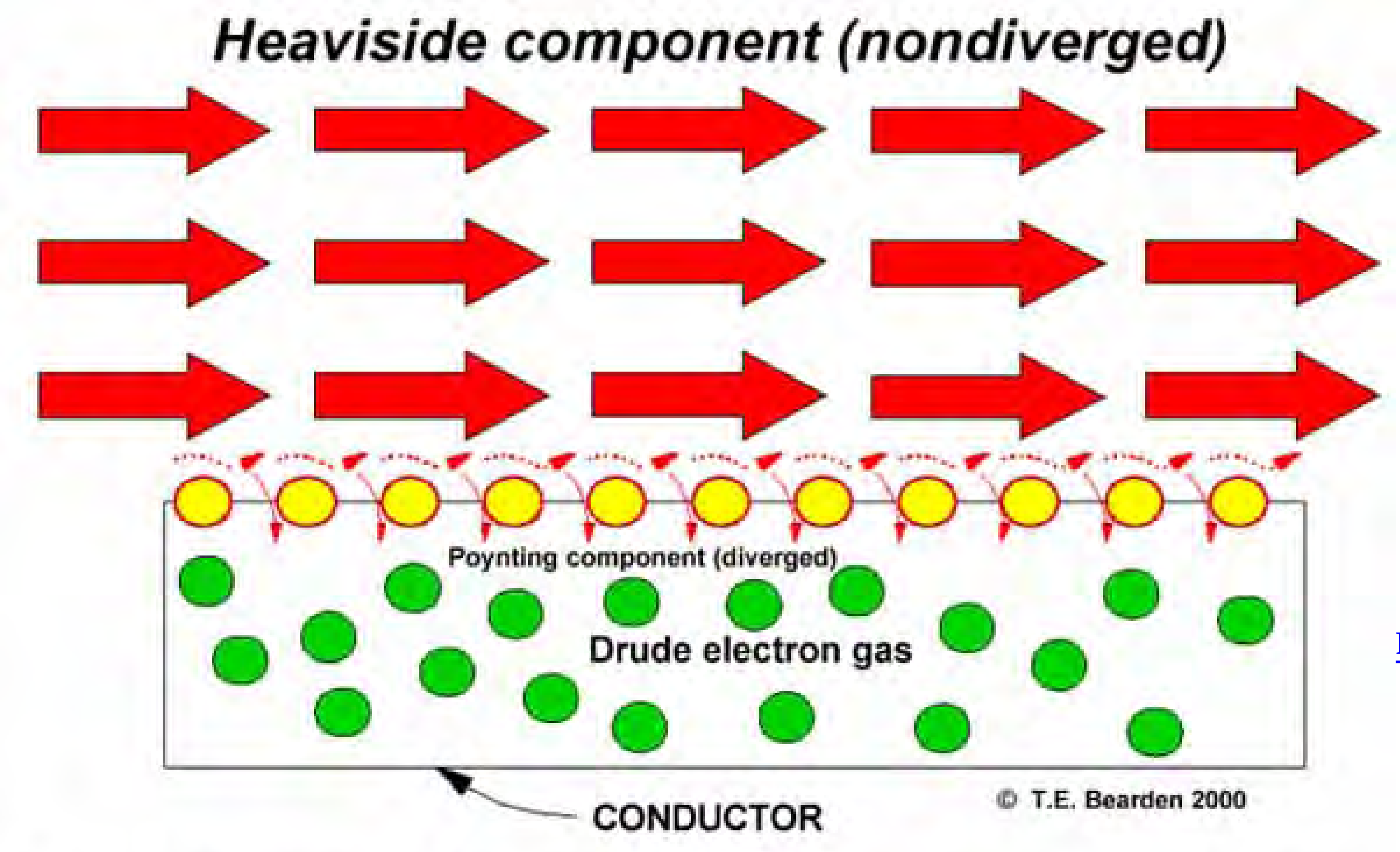
Heaviside component (nondiverged)
Poynting component (diverged)
Drude electron gas
CONDUCTOR
© T.E. Bearden 2000

Figure 5. Heaviside and Poynting energy flow components. The Heaviside component is often 10 trillion times the Poynting component, but is simply wasted in ordinary single-pass energy flow circuits.

Next Slide

Return to Index

Previous Slide

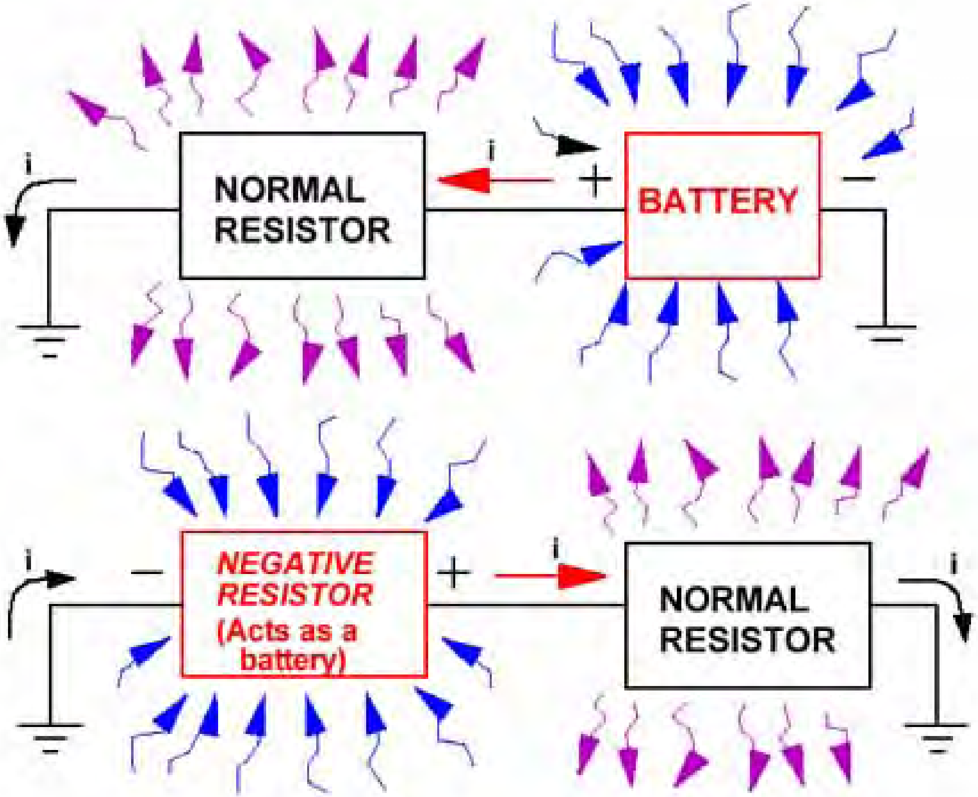
NORMAL RESISTOR
i
i
+
BATTERY
−
NEGATIVE RESISTOR
(Acts as a battery)
−
+
i
i
i
NORMAL RESISTOR

Figure 6. Negative resistance process vs. positive resistance process. A negative resistor receives energy in unusable form, transduces it, and outputs it in usable form. A positive resistor receives energy in usable form, and scatters it into unusable form.

[Next Slide]

[Return to Index]

[Previous Slide]

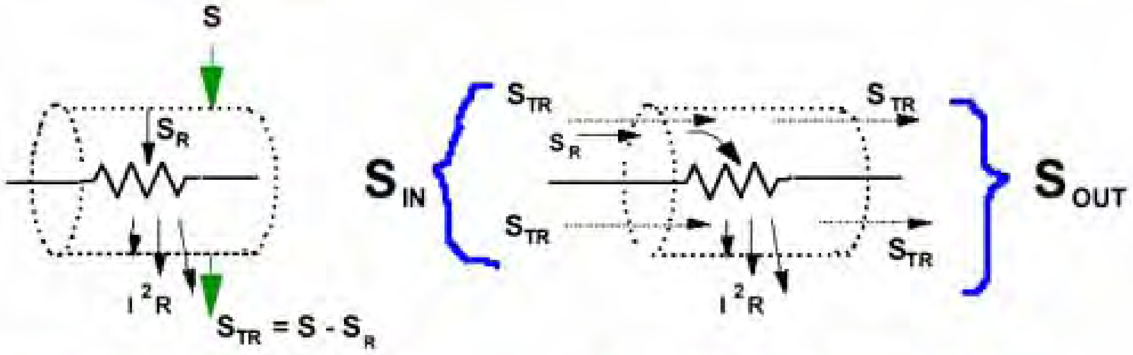

1a. Lorentz surface integration.

1b. Actual S in and S out.

[Return to Index](#)

[Previous Slide](#)

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd. edn., Addison Wesley, 1962, p. 178-181.

Note: If the S-vector is integrated over the closed surface, then all nondiverged energy flow is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Poynting vector remains. This measures only the tiny portion of the S-flow that is intercepted and diverged into the conductors by their surface charges, powering the electrons and then dissipated out of the resistor as joule heating.

The Lorentz procedure arbitrarily discards the enormous Heaviside component that misses the circuit entirely and is wasted. This results in a non sequitur of first magnitude in energy flow theory .

Figure 7. Lorentz's integration trick to discard the enormous Heaviside non-diverged energy flow component.

Page 1	Page 2	Page 3	Page 4	Page 5	Page 6
(Table 1)	(Table 2)	[note 6]	[note 10]	[note 14]	[note 16]
[note 1]	[note 2]	[note 7]	[note 11]	[note 15]	[note 17]
	[note 3]	[note 8]	[note 12]	Figure 11	(Figure 13)
	[note 4]	[note 9]	[note 13]	Figure 12	(Figure 14)
	[note 5]	(Figure 1)	(Figure 7)		(Figure 15)
		(Figure 2)	(Figure 8)		
		(Figure 3)	(Figure 9)		
		(Figure 4)	(Figure 10)		
		(Figure 5)			
		(Figure 6)			

Page 7	Page 9	Page 10	Page 12	Page 13
[note 13]	(Table 3)	[note 23]	(Table 4)	[note 31]
[note 18]	[note 22]	[note 24]	[note 28]	[note 32]
[note 19]	(Figure 17)	[note 25]	[note 29]	[note 33]
[note 20]	(Figure 18)	[note 26]	[note 30]	[note 34]
(Figure 16)		[note 27]	(Figure 21)	[note 35]
		(Figure 19)	(Figure 22)	(Figure 23)
		(Figure 20)		

THE VIRTUAL TIMES

CHASING THE WILD DRAGON:

Foundations of a New Science

By T.E. Bearden
November 12, 1995

It is a pleasure to accept the offer of John and Larry to write a column in their new InterNet magazine. Theirs is a great venture, and with the energy and knowledge they bring to the task, one can be confident of their success in making this a very interesting and informative publication.

There is a new "rebel" science slowly being born, with most of the work being done outside the universities and orthodox scientific community. In the months ahead it will be my job to try to acquaint you with the gist of what's going on, and what all the excitement is about.

This first article is deliberately just introductory, and its purpose is to interest you in the unusual fields we are going to be covering in more scientific detail, in future columns. We will also be listing a variety of references from the technical literature, which bear upon the various points addressed.

In this first article a little of my own background experience [note 1] is given, so you can decide whether you will be interested in what we have to say!

The real purpose of the column will just be to *present information* that you may not be able to readily obtain any other way. We will not argue or debate with skeptics; nothing productive is to be gained by that. If the information we present is useful to you, then our purpose has been accomplished. If it isn't of use to you, then just file it in the old wastebasket. Also, from time to time I may answer a generic question, constructed from multiple queries. Individual responses to correspondence cannot be provided.

The work toward a new breakthrough science is going on in a broad spectrum of "fields," most of which are not yet recognized by academia as even constituting legitimate areas of scientific endeavor (Table 1).

Some of the major areas involved are:

1. theoretical and experimental research in overunity electrical machines,
2. free energy research of other kinds, such as devices utilizing thermodynamics phase changes,
3. rapidly advancing efforts in cold fusion,
4. work on permanent magnet motors,
5. antigravity research, both theoretical and experimental,
6. development and utilization of a "hidden" electrodynamics inside the scalar potential (inside voltage),
7. the "memory" of space and materials; where this "memory" can be deliberately conditioned so that space and the materials themselves become causal actants rather than simple inert entities,

8. direct engineering of quantum change via a new kind of "hidden electrodynamic variables,"
9. the direct engineering of the vacuum's virtual photon flux exchange with matter,
10. action at a distance, including the deliberate formation and usage of quantum potentials,
11. direct engineering of the local curvature of spacetime (yes, the direct engineering of general relativity, using the new hidden kind of electrodynamics),
12. use of the new "hidden" electromagnetics to engineer body cells including the reversal of diseased cells, genetics and all, back to a previous healthy state (cellular dedifferentiation, in biophysical terms, or cellular phase conjugation, in physics terminology),
13. research toward the direct engineering of the mind, including both the conscious and unconscious minds, long-term memory, and the personality itself, and
14. a ruthless re-examination (and correction where necessary and possible) of the definitions, concepts, and postulates that present physics is founded upon.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

[1] For those interested in orthodox qualifications, I have an M.S. in nuclear engineering (1971) from Georgia Institute of Technology and a BS in mathematics (1953) from what is now Northeast Louisiana University with a minor in electronic engineering.

Table 1. Fourteen Emerging New Science Areas

1. Overunity Electrical Machines and Power Systems
2. Overunity Devices Using Thermodynamical Phase Changes
3. Cold Fusion and Non-Nuclear Transmutation
4. Self-Powered Permanent Magnet Motors
5. Antigravity and Antigravity Engineering
6. Use of the "Inner" EM Inside the Scalar Potential
7. Vacuum Engineering, With Vacuum and Materials Memory
8. Direct Engineering of Quantum Change
9. Engineering and Structuring of Massless Charge Flux
10. Action-At-A-Distance, Including Use of Quantum Potentials
11. Direct Engineering of Local Spacetime Curvature
12. "Inner EM" Time Reversal of Diseased Cells Back to Health
13. Direct Engineering of the Mind, Including Software Loading
14. Unified Field Revision of Foundations Concepts of Science

Each of these "major areas" as yet has no single "experts" in the sense that individuals exist who possess a "mastery" of the given area. As an example, in spite of multiple researchers and decades of vigorous research, no one is really an "overall expert" in the overunity electrical power systems field. The "field" itself is still struggling to be born and is not recognized at all by orthodoxy; in fact, many orthodox scientists bitterly resent one even attempting to work in this area. There is as yet no cohesive single "theory" of overunity electrical or magnetic machines, although our own group is making rapid progress in that respect. But there is still no universal agreement amongst even the researchers themselves. Literally the field is still being *discovered*.

This "groping dimly toward a new field" should not be too surprising. After all, the early electromagnetics researchers went through the same tortuous, agonizing kind of development. They were playing with cat fur, glass rods, and pith balls for quite an extended period before understanding progressed to the point where Maxwell could mathematically tie together and model the results of the experimenters such as Faraday and others. It has taken a century and a half for modern electromagnetics and electronics to reach its own present development. It is going to take another decade or so before the overunity researchers can overturn several ill-founded parts of the present electromagnetics and get the emerging field onto a solid foundations footing and into a solid theoretical position.

In the effort to produce a theoretical model of electromagnetics, Maxwell and his predecessors performed a magnificent work. Yet they also made very serious errors (Table 2), which persist to this day. [note 2] Let us list just a few of the more blatant errors in classical EM theory today:

First, Franklin guessed wrong on which way electrical current flowed in a circuit. There is no definition at all of electrical charge in all of physics, and there is no acceptable definition of potential (that is, there is no rigorous definition of just good old "voltage.") There are no "fields of force" in the vacuum; there are just potential gradients __ because the vacuum is just one vast potential (virtual particle flux) in the first place. Force fields actually occur only in and of the matter in a material system. The primary causative agents of electromagnetics are not the so-called "force fields' at all, but are the potentials __ as is well-known in quantum mechanics. [note 3] Maxwell assumed a material ether in his equations. The Michelson-Morley experiment destroyed the material ether, but the Maxwell equations have never been changed accordingly. The electron was not yet discovered when the Maxwell equations were written; the theory does not address electron spin. Maxwell was already dead some two decades when Barus' 1898 paper [note 4] was published, pointing out the strange "backwards-traveling" wave (i.e., time-reversed wave) solution to the wave equation. In addition, Heaviside and Gibbs had already produced their vector truncation __ this truncated theory is the modern so-called "Maxwell's equations" that universities teach today __ of Maxwell's theory. Not a single one of them ever appeared in any book or paper by James Clerk Maxwell. [note 5]

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

Table 2. Some Defects In Electrodynamics

1. Conventional Current Direction Is Reversed
2. No Definition of Charge or Potential
3. Equations Still Assume Material Ether
4. Use of Force Fields in Vacuum is Known to be False
5. Treats Charge as Unitary; lnstead, It is a Coupled System
6. Confuses Massless Potential Gradients as Forces
7. Does Not Utilize Mass as a Component of Force
8. Erroneously Assumes Force Fields as Primary Causes
9. Topology of EM Model Has Been Substantially Reduced
10. Does Not Include Quantum Potential
11. Does Not Include Proven Action At A Distance
12. Does Not Utilize Near Field Coulornb Gauge Effects
13. Does Not Include EM Generatrix for Time Flow
14. Does Not Unify Photon and Wave Aspects
15. Does Not Include Electron Spin and Precession
16. Treats Energy As Existing In "Chunks," instead of as Flow
17. Confuses Energy and Energy Collection
18. Discards Half of Every EM Wave in Vacuum
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal
20. Omits Phase Conjugate Optics Effects
21. Does Not Include EM Cause of Newtonjan Reaction Force
22. Erroneously Assumes Separate Force Acting on Separate Mass

[2] For an independent statement that electrodynamics, both classical and quantal, are in serious disarray, see Mario Bunge, Foundations of Physics, Springer-Verlag, New York, 1967, p. 176. Quoting: "...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state." Bunge points out many errors in electrodynamics. See also Terence W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in Lakhtakia, A. (ed.): Essays on the Formal Aspects of Electromaqnetic Theory, World Scientific, Singapore, 1992, p. 6-86. To find information on what's wrong with EM theory, one must read foundations literature, not the standard textbooks. One of my own contributions has been to point out that the notion of charge q is not unitary. In fact, electrical charge should be defined in terms of q ~ m(sub q)~(sub q). In other words, electric charge q has a massive part m, and a massless part consisting of its potential phi. Further, the magnitude of phi is just the change in the virtual photon flux of the ambient vacuum, due to the [quantum field theoretic] virtual photon exchange with the mass of the q. It follows that the true electrical charge of a particle is just its native potential ~, which is also a dynamic energy exchange with the surrounding vacuum. It also follows that this massless electrical charge changes whenever the particle is placed in a potential (in a different ambient virtual photon flux). After all, potentials superpose; that is their major characteristic. Note that the potential of the test particle is ignored in classical electromagnetics, whenever one speaks of the "E-field" upon the particle. Further, the actual structure of this virtual photon flux that comprises massless electrical charge, can itself be deterministically structured and utilized to generate nonlinear effects that do not appear at all in the present conventional theory. We will be covering many of these effects in future articles.

[3] E.g., see Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), 1959, p. 485-491. Effects of potentials on charged particles exist even in the region where all the fields (and therefore the forces on the particles) vanish, contrary to classical electrodynamics. The quantum effects are due to the phenomenon of interference. These effects occur in spite of Faraday shielding. The Lorentz force does not appear anywhere in the fundamental quantum theory, but appears only as an approximation that holds in the classical limit. In QM, the fundamental physical entities are the potentials, while the fields are derived from them by differentiation.

[4] Carl Barus, "A Curious Inversion in the Wave Mechanism of the Electromagnetic Theory of Light," American Journal of Science, Vol. 5, Fourth Series, 1898, p. 343-348.

[5] E.g., see Paul J. Nahin, Oliver Heaviside: Saqe in Solitude, IEEE Press, New York. Quoting from p. 9: "...nowhere in Maxwell's writings do the equations for the electromagnetic field appear as we write them today. Maxwell used an amalgamation of Cartesian component and quaternion notation, and it was Heaviside who first wrote the electromagnetic field equations in modern vector form." Quoting from p. 92: "In actuality , however, the fields were not the primary reality to Maxwell at all (that is an idea that developed after Maxwell's death and is due to Hertz and Heaviside), but rather it was Faraday's 'electrotonic state' that he thought to be the real thing. Like Faraday, Maxwell believed that electromagnetic effects are observable results of an altered state of the ether. The mathematical formulation of this electrotonic state, for Maxwell, is what we today (as he did) call the vector potential." And again, from p. 96: "To Maxwell, however, the vector potential had a most definite physical meaning."

So the time-reversed EM wave that one finally sees emerging in nonlinear phase conjugate optics in the early 1970s was omitted from the Maxwellian theory and its later modification by Heaviside and Gibbs. Maxwell's actual theory is in a higher topology (quaternion algebra), and involves some 20 quaternion equations. Heaviside and Gibbs created vector algebra, and seriously curtailed Maxwell's theory when they "translated" it into the much lower topology of vectors and some four equations. The later addition of tensors as the basis mathematics did not restore the lost topology provided by the quaternion model.[note 6]

For the EM wave, Maxwell simply assumed the transverse EM wave that had come from the much earlier derivation of the familiar wave equation from the old "plucked taut string" notions. [note 7] Yet when one closely examines and corrects some of the erroneous *assumptions* [note 8] made for the actual mathematical derivation of this "taut string" transverse wave, one finds (Figure 1) that an equal-energy second antiphased transverse "antiwave" is actually produced by the instrument and added to the "plucked string's" transverse wave disturbance of the air medium, to produce a longitudinal wave of compression and rarefaction in the air itself. The "transverse string wave" in fact never leaves the string. The string wave is not the wave that is launched into the medium.

A similar electromagnetic situation exists. In a whip antenna (Figure 2), the Drude electron gas in the conductor does exhibit a transverse "string-type" EM wave of electron precession. However, at the same time, by Newton's third law (which Maxwell neglected because at the time the atom was regarded as just a blob with no internal structure) the positively charged, phase conjugated atomic nuclei have equal energy, highly damped transverse recoil waves created in them. Further, these "nuclei-recoil" waves are phase conjugate replicas of the electron gas waves. The total disturbance that is launched into the vacuum from the whip antenna is (a) the normal transverse disturbance of the virtual photon flux of vacuum caused by the electron gas wave, and (b) the coupled equal-energy transverse antiwave caused by the nuclei-recoil via Newton's third law. The amplitude of the material antiwaves of nuclei vibration are highly damped because of the massiveness of the nuclei; however, they are of equal energy to the electron wave. But when launched into the same vacuum, the two equal-energy wave disturbances experience equal damping because they are in the same medium. Hence the wave and antiwave are (a) locked together in the vacuum, and (b) of equal energy, and (3) of equal amplitude once launched into the vacuum. What results is a vacuum EM wave of rarefaction (reduction in intensity) and compression (increase in intensity) of the virtual photon flux of vacuum __ just exactly like what is launched into the surrounding gaseous air medium from the "plucked taut string" and perturbed body of a physical instrument. The EM wave in vacuum is a longitudinal wave directly analogous to a sound wave __ just as Nikola Tesla stated. [note 9]

The vacuum may be considered a scalar potential, in which case it has an internal biwave structure (Figure 3). Interference of two potentials __ i.e., the hidden interference (Figure 4) of their internal wave structures __ in fact already creates all EM fields and waves.

When the incoming longitudinal "biwave" from the vacuum strikes the atoms of the receiving antenna, each atom may be considered a dynamic assembly of multiple dipoles, where each dipole consists of one of the electrons in the electron shells and one of the positive charges in the nucleus. The receiving dipole in the interacting atom splits the two waves apart again (Figure 5). The negatively charged end of the dipole is time-forward, and the positively charged end is time-reversed.

Hence the separated normal-wave-half of the biwave interacts with the electron, to give again the familiar transverse "electron precession" wave that we detect with our "electron wiggle-detecting" instruments (Figure 6). This detected electron precession wave is what all our textbooks prescribe erroneously as the vacuum wave! The separated time-reversed-wave-half of the incoming vacuum wave is split off and interacts with the positive end of the receiving atomic dipole, in the time-reversed atomic nucleus, thus generating Newton's third law reaction force to provide the recoil of the nucleus. Present physics has ignored this second wave altogether, because of the serious error in the original derivation of the wave equation from plucked strings. Thus physics presently has absolutely no causative mechanism for Newton's reaction force. Instead, scientists just "mystically invoke the appearance of this magic force." Describing that something occurs, and naming its occurrence, does not constitute advancing a causative mechanism!

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

[6] E.g., see T.W. Barrett, "Tesla's Nonlinear Oscillator- Shuttle-Circuit (OSC) Theory," Annales de la Fondation Louis de Broglie, 16(1), 1991, p. 23-41. Barrett shows that a higher topology EM model (e.g., quaternions) allows shuttling and storage of potentials in circuits, and also allows additional EM functioning of a circuit that a conventional EM analysis cannot reveal. Tesla's actual patented circuits in fact exhibit this higher topology, when a quaternion EM analysis of them is performed. No orthodox vector/tensor analysis can show this.

[7] All the early experimenters seemed to assume the transverse string-type wave automatically for the vacuum medium. E.g., Faraday speculated that light and radiant heat might be transverse vibrations propagated along his "lines of force," from which the direct "plucked string" analogy can readily be seen. In short, he considered that his lines of force could be perturbed to vibrate laterally, just like a plucked string. See M. Faraday, "Thoughts of Ray-Vibrations," Philosophical Magazine, (3), vol. xxvii, 1846, p. 345. Maxwell also assumed the transverse EM wave in vacuum without question, as did Heaviside, Hertz, and Gibbs and as has almost every electrodynamicist since then.

[8] The worst is the assumption of infinitely rigid endposts holding the taut string, so that no movement of these endposts occurs. This further assumes that no induced movement of the body of the instrument occurs. The end result is to eliminate all of that wave that is produced in the body of the instrument, and retain only the transverse wave produced in the string. In the first place, this standard procedure has now become a total violation of Newton's third law, which among other things requires that forces be formed in oppositional pairs. So the standard method for the launch of a wave from an instrument is flawed. Specifically, the "transverse string wave" is falsified insolaf as its appearance in the surrounding medium. This is not the wave that actually exits the instrument and that is launched into the medium. This universal oversight on the part of mathematicians and physicists is a colossal joke upon them. Literally, when a physicist writes the transverse wave equation on the board, he has just discarded exactly half the physics. This means that, in his interactions, at the end there will suddenly appear this mysterious, mystical "equal and opposite reaction force" that is created by the erroneously ignored antiwave. So he has to piously invoke this reaction force "from the blue." Yet quantum field theory teaches us already that all mechanical and electromagnetic forces on a mass are generated by absorption and emission of virtual photons. Hence Newton's third law reaction force must also be so generated. This means that there must exist an equal number of virtual antiphotons interacting in the antiforce, as there was virtual photons interacting in the causative force. By symmetry there must also exist an equal and opposite wave, or antiwave, made of these same antiphotons. The actual disturbance that entered from the vacuum must therefore have consisted of both wave and antiwave. QED.

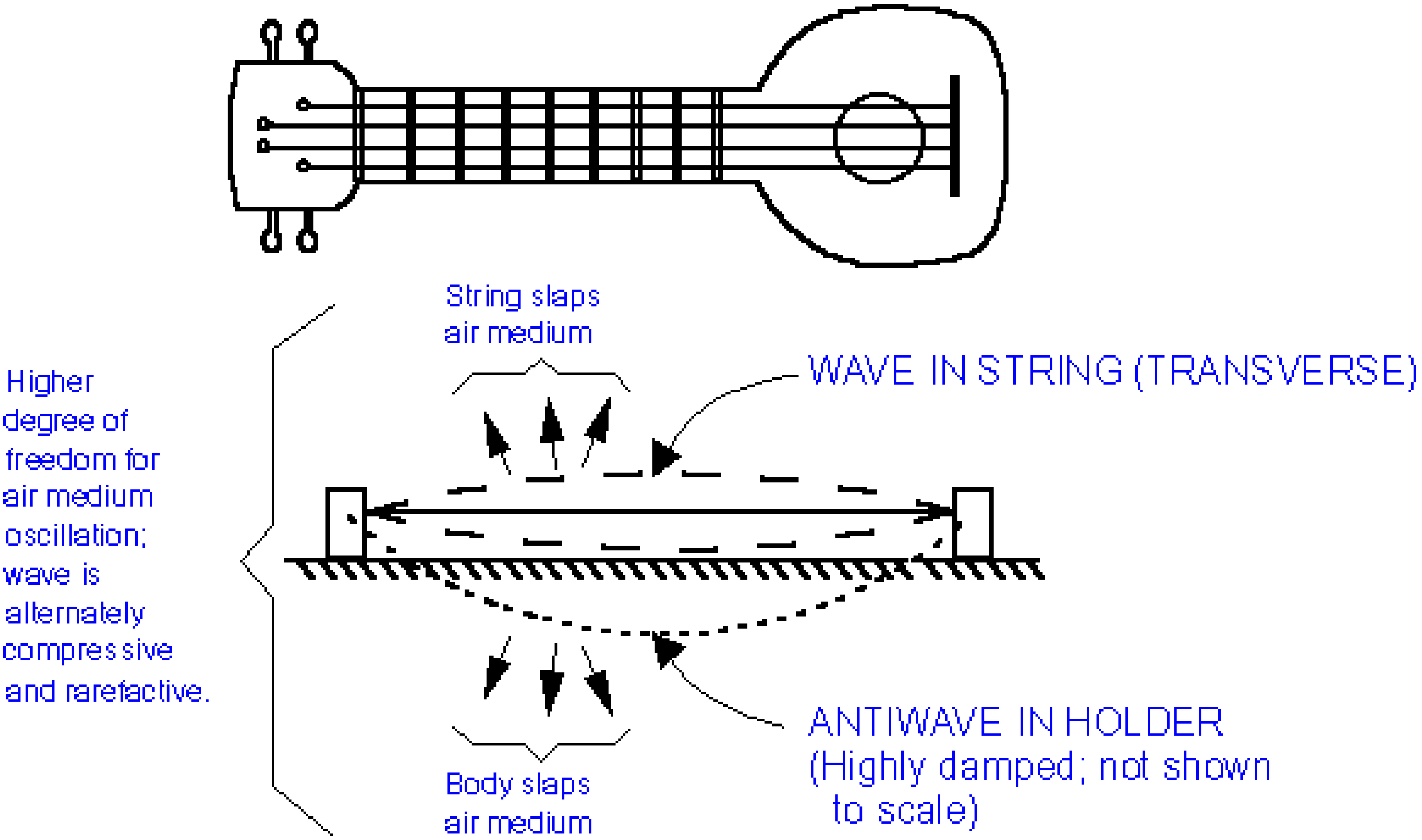
String slaps
air medium
WAVE IN STRING (TRANSVERSE)
Higher
degree of
freedom for
air medium
oscillation;
wave is
alternately
compressive
and rarefactive.
ANTIWAVE IN HOLDER
(Highly damped; not shown
to scale)
Body slaps
air medium

Figure 1. Air medium disturbances generated by plucked string.

Two waves are created, both slapping the air medium at the same time, but in opposite directions. A disturbed medium oscillates according to its own degrees of freedom. The air has more degrees of freedom than the string. A wave of compression and rarefaction is created in the air medium.

D-1-F 1

© 1993 T.E. Bearden

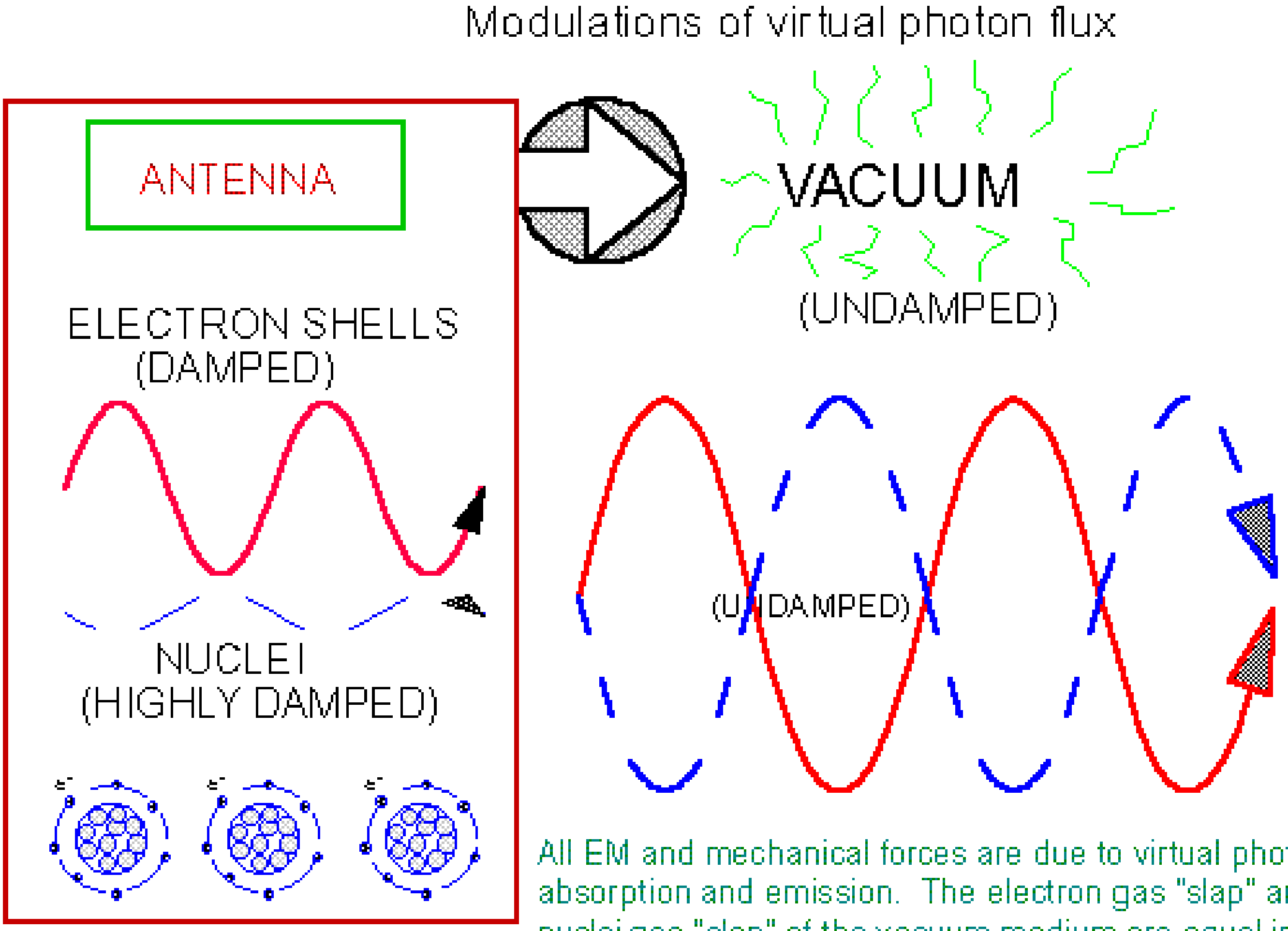
Modulations of virtual photon flux
ANTENNA
ELECTRON SHELLS
(DAMPED)
NUCLEI
(HIGHLY DAMPED)
VACUUM
(UNDAMPED)
(UNDAMPED)

All EM and mechanical forces are due to virtual photon absorption and emission. The electron gas "slap" and the nuclei gas "slap" of the vacuum medium are equal in energy. Hence they inject equal numbers of virtual photons into the vacuum, creating equal and opposite vacuum disturbances. This gives a vacuum wave of alternating compression and rarefaction. This vacuum EM wave is similar to a sound wave, which is modeled longitudinally.

Figure 2. Launching an EM wave from a wire antenna.

D-1-F2

© T.E. BEARDEN 1994, 1996

[9] See Nikola Tesla, "Pioneer Radio Engineer Gives Views on Power," New York Herald Tribune, Sep. 11, 1932. Quoting: "...I showed that the universal medium is a gaseous body in which only longitudinal pulses can be propagated, involving alternating compressions and expansions similar to those produced by sound waves in the air. Thus, a wireless transmitter does not emit Hertz waves which are a myth, but sound waves in the ether, behaving in every respect like those in the air, except that, owing to the great elastic force and extremely small density of the medium, their speed is that of light." Today the QM vacuum has been theoretically and experimentally shown to consist of a "gas" of virtual particles, rather like Tesla stated. Tesla was also correct about the longitudinal EM wave in the vacuum. He was also correct in his statement that a fundamental charged particle can alter its charge, if one considers the massless charge ~ of charge q separately from the mass m of charge q. When the charge q is placed in a nonzero potential V, it follows that the new charge of the electron is simply (~+V).

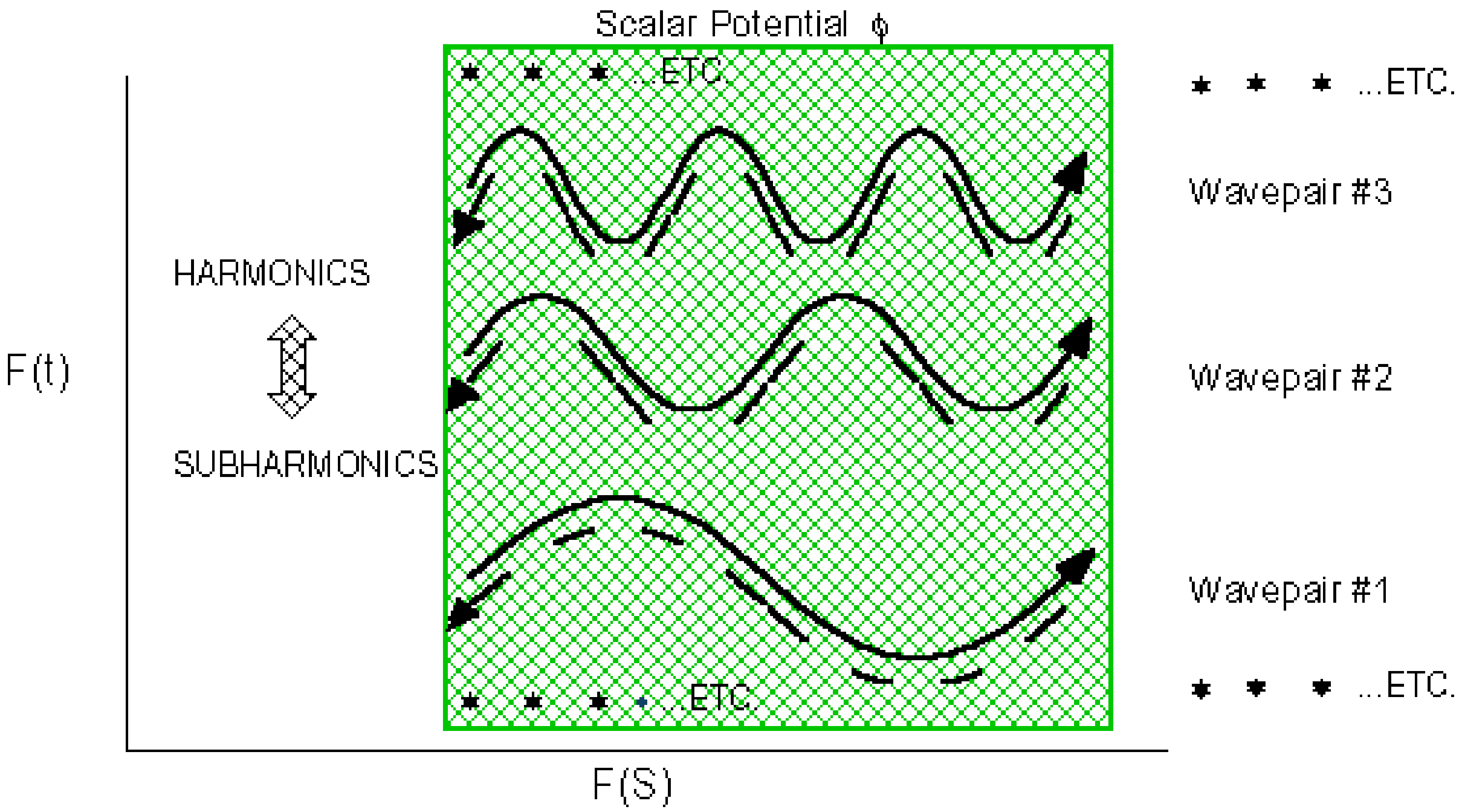
Scalar Potential ϕ
* * * ...ETC.
* * * ...ETC.
Wavepair #3
HARMONICS
SUBHARMONICS
F(t)
Wavepair #2
Wavepair #1
* * * ...ETC.
* * * ...ETC.
F(S)

Figure 3. Internal Wave Structure of the scalar potential.

A harmonic set if wavepairs, where in each wavepair the two waves superpose spatially, but travel in opposite directions. The two waves in each pair are true phase conjugates and time-reversals of each other. Thus they comprise a coupled wave and antiwave. The photons must also be coupled into photon/antiphoton pairs, by a strong application of the distortion correction theorem of nonlinear optics.
Each wavepair is a standing electrogravitational wave.

D-1-F3

© T.E. Bearden 1984, 1985

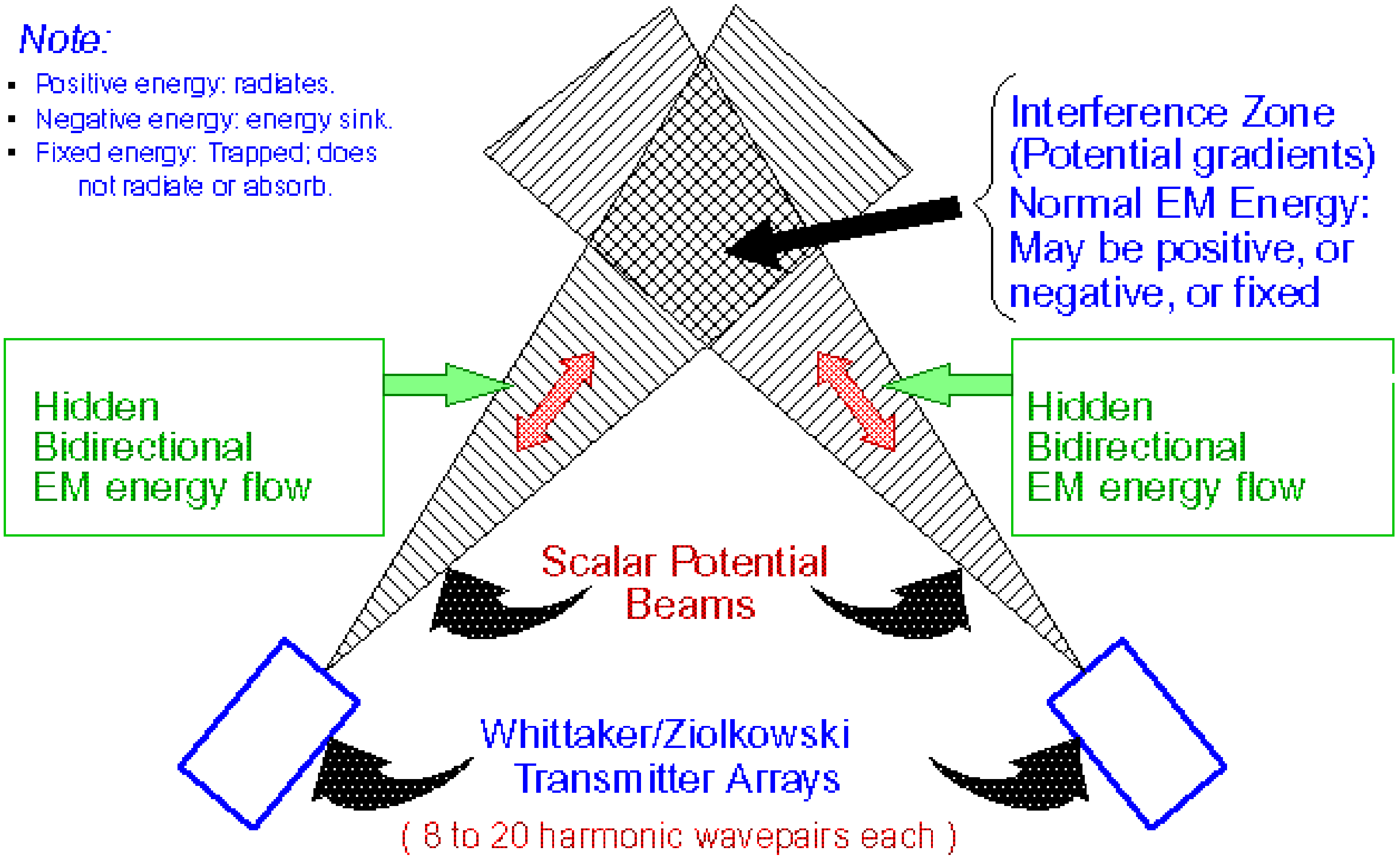
Note:
Positive energy: radiates.
Negative energy: energy sink.
Fixed energy: Trapped; does not radiate or absorb.
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

Figure 4. Scalar potential interferometry provides action-at-a-distance. It already creates normal EM fields and waves.

D-1-F4

© 1994 T.E. Bearden

Tables:

MECHANISM FOR NEWTON'S THIRD LAW

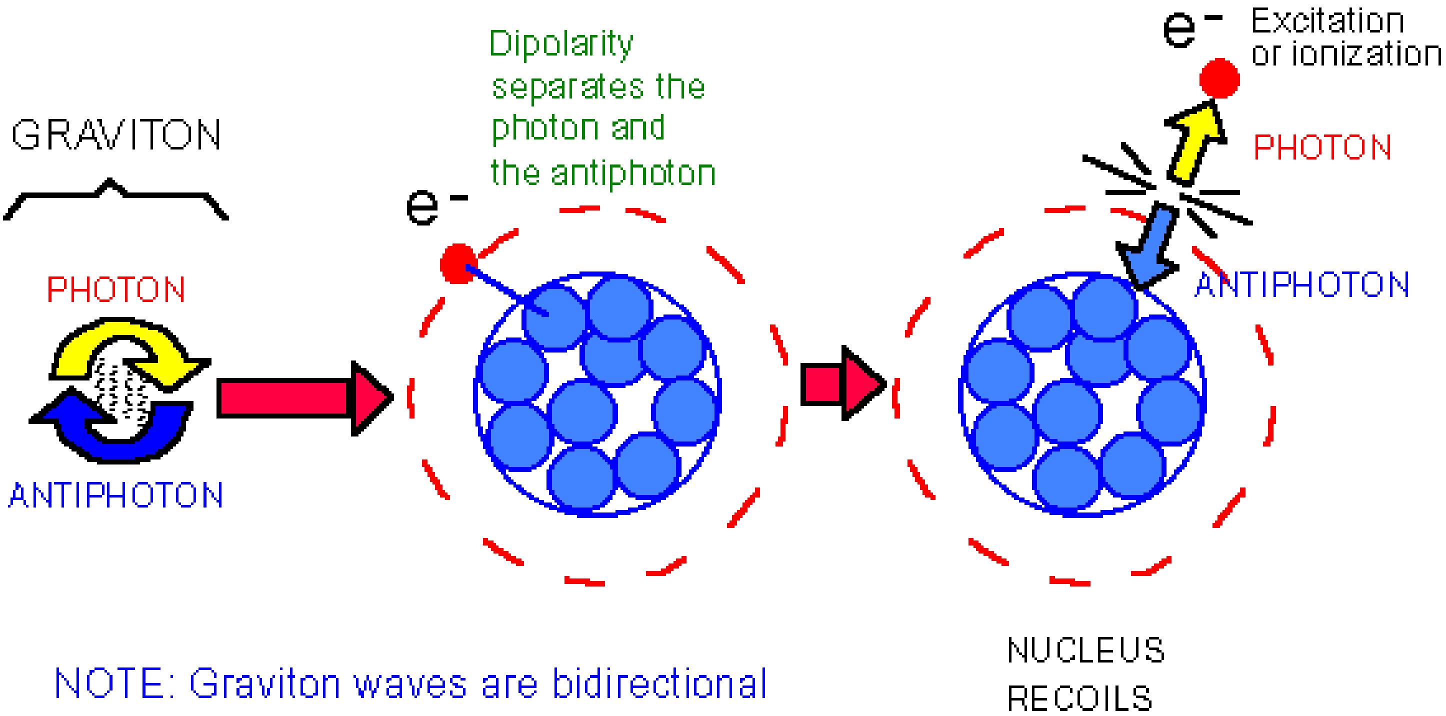
GRAVITON
PHOTON
ANTIPHOTON
Dipolarity separates the photon and the antiphoton
e -
e -
Excitation or ionization
PHOTON
ANTIPHOTON
NUCLEUS RECOILS

NOTE: Graviton waves are bidirectional phase waves of photon/antiphoton coupling and uncoupling in the vacuum.

The atom is an assembly of dynamic dipoles.

A dipole splits the graviton into a photon and an antiphoton.
The photon interacts with the electron, producing the action.
The antiphoton interacts with the nucleus, producing the reaction.

Figure 5. Graviton interaction with matter produces the photon interaction and Newton's third law reaction.

D-1-F5

Tables:

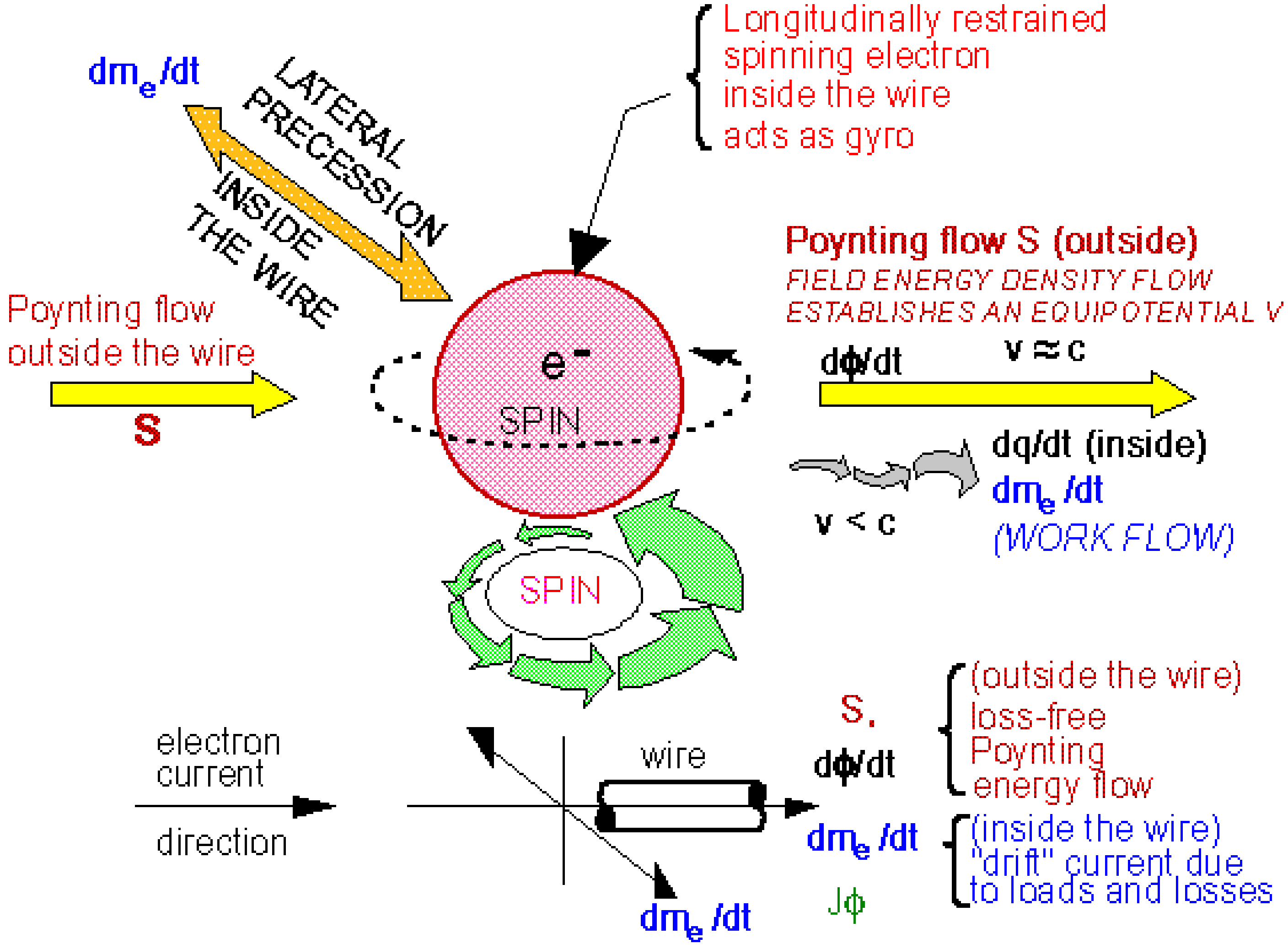
dm_e/dt
LATERAL PRECESSION INSIDE THE WIRE
Longitudinally restrained spinning electron inside the wire acts as gyro
Poynting flow S (outside)
FIELD ENERGY DENSITY FLOW ESTABLISHES AN EQUIPOTENTIAL V
Poynting flow outside the wire
S
e^-
SPIN
$d\phi/dt$
$v \approx c$
dq/dt (inside)
dm_e/dt
(WORK FLOW)
$v < c$
SPIN
electron current
direction
wire
S,
$d\phi/dt$
(outside the wire) loss-free Poynting energy flow
dm_e/dt
(inside the wire) "drift" current due to loads and losses
$J\phi$
dm_e/dt

Figure 6. Currents in and along a conductor, with electron precession.

D-1-F6

© 1993 T.E. Bearden

Further, if we should trick the time-reversed wave half into coming outside the atom instead of going inside to interact with the nucleus (Figure 7), then it follows that the nucleus would not recoil. In that case the interaction would appear to violate Newtonian third law reaction. Phase conjugate mirrors already do just exactly this when they emit a phase conjugate replica wave. They direct the antiwave outside and away from the atom instead of inwards into the nucleus. So they do not recoil, no matter how powerful a phase conjugate replica wave they emit.

Finally, we can also show that the EM wave in vacuum is longitudinal, by yet another approach. The electrons in the electron gas in a conductor are spinning, and they are severely constrained longitudinally to be able to move only very, very slowly down the wire. So when perturbed, they act as gyroscopes, and precess at right angles to the perturbing force. The fact that the gyroelectrons in a wire move radially almost entirely, and only "slip" down the wire very slowly with a "drift" velocity, already proves conclusively that the incoming vacuum disturbance was longitudinal. The precession of a gyro is at right angles to the perturbing force. Therefore rigorously the vacuum EM waves are longitudinal, because the gyroelectrons moved transversely. The present assumption of the transverse EM wave in vacuum contradicts the entire theory of gyroscopes! [note 10] There are many other errors in classical EM that we will point out in future articles; this should give the reader a flavor of what is in store for him in future articles.

So there are no truly satisfactory "experts" in any of the 14 fields. Yet there is still something to be said for having some years of experience in one or more of the areas. I have had the good fortune or misfortune, depending upon one's viewpoint, to be involved in various of those 14 activities for some 30 years or more, and am active in several of the areas now. My colleagues and I are particularly active in researching overunity electrical devices (see Figure 8), to include the effects of controlled chaotic oscillators, charge-blocking oscillators, and oscillators driven by in-shuttled (bridged) Poynting field energy density flow ExH. Parametric oscillation is known to have yielded overunity electrical machines and was fully reported by Russian researchers in the Russian and French technical journals in the mid 1930s. [note 11] We will further cite that little-known work in future articles. Also, we are working mightily upon true negative resistors (Figure 9), wherein the resistor scavenges some of the disorganized energy from the vacuum's energetic exchange with the charges in the system, organizes and collects it, and then outputs that collected energy flow into the external circuit. Such a resistor acts as a power source. Kron, perhaps the greatest of the electrical machinery scientists, is known to have produced such true negative resistors.

I have also proposed a theoretical explanation of the Priore device (Figure 10) in France which in laboratory animals demonstrated remarkable cures of terminal tumors, cured trypanosomias, reversed atheriosclerosis, and restored suppressed immune systems. The Priore work was not anecdotal, but was performed under rigorous scientific protocols, working under the auspices of renowned French scientists such as Courrier and Pautrizel. The work is properly reported in numerous papers in the French medical literature. [note 12] The results were certainly achieved, but neither Priore himself nor the scientists of the day could understand the mechanism involved. The deciphering of that causative technical mechanism utilized by the Priore group required 10 years of intense effort. The results have now been published in the journal Explore! and are there for anyone to read. [note 13]

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

PHOTON INTERACTION WHEN ATOM PHASE CONJUGATES

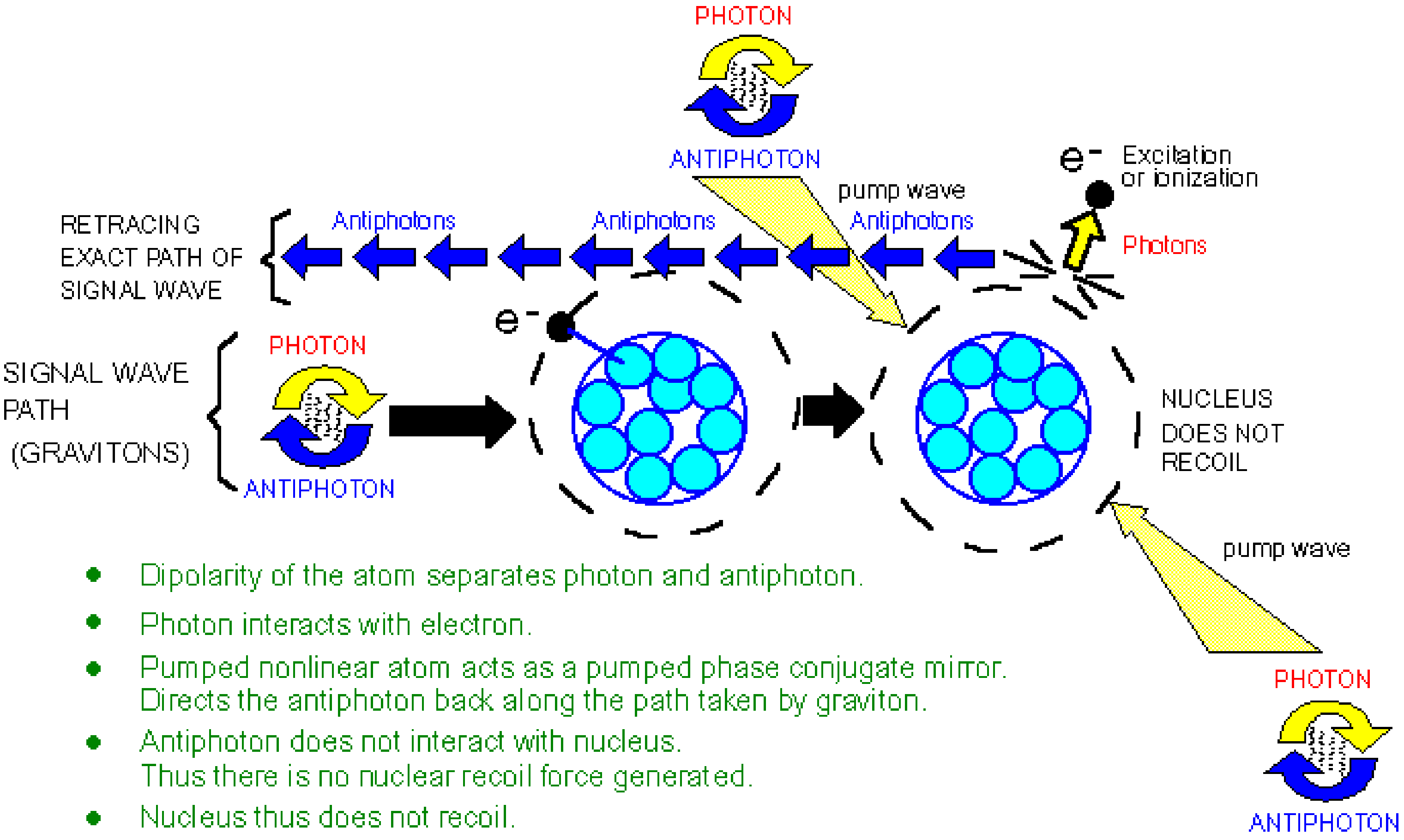
PHOTON
ANTIPHOTON
e⁻
Excitation
or ionization
pump wave
RETRACING
EXACT PATH OF
SIGNAL WAVE
Antiphotons
Antiphotons
Antiphotons
Photons
e⁻
SIGNAL WAVE
PATH
(GRAVITONS)
PHOTON
ANTIPHOTON
NUCLEUS
DOES NOT
RECOIL
pump wave
PHOTON
ANTIPHOTON

- Dipolarity of the atom separates photon and antiphoton.
- Photon interacts with electron.
- Pumped nonlinear atom acts as a pumped phase conjugate mirror. Directs the antiphoton back along the path taken by graviton.
- Antiphoton does not interact with nucleus. Thus there is no nuclear recoil force generated.
- Nucleus thus does not recoil.
- Not a violation of Newton's third law, once one understands the mechanism that generates third law reaction. In the absence of the generating mechanism, there is no third law recoil. In EM field theory, the "missing antiwave" has necessitated third law omission.

Figure 7. Emission of antiphoton; nucleus does not recoil.
Atom acts as a pumped phase conjugate mirror.

D-1-F7

© 1995 T.E. Bearden

[10] When Maxwell's theory was constructed, the electron had not yet been discovered, and so the dipolarity of the atom was unknown. An atom was regarded just as a "blob." Electricity in a conductor was regarded as a thin material fluid Maxwell's equations are in fact hydrodynamic equations. So the founders regarded the detection of induced motion in the electric fluid as induced translation, not induced precession. They had no idea of "spins" of particles of the fluid, nor of gyro precession of those particles. Instead, q was regarded just as a "piece of electric fluid," similar to a cubic centimeter of fluid.

A fundamental AC resonant principle for an overunity electrical power system

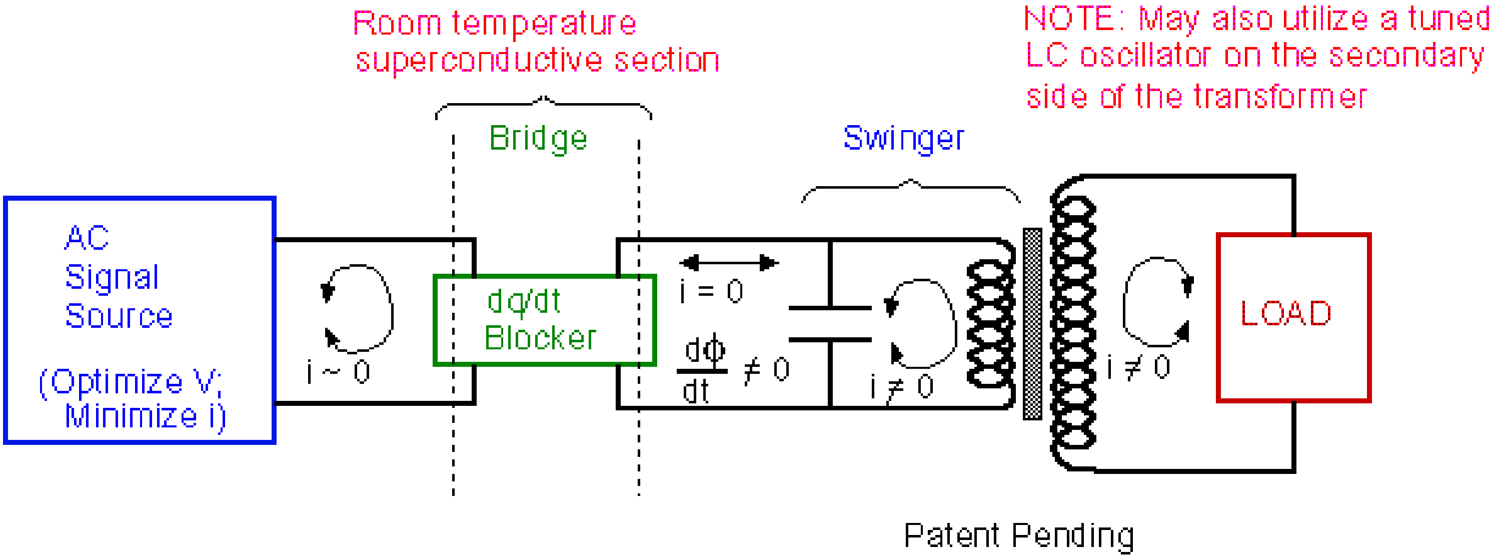
Room temperature
superconductive section
Bridge
Swinger
AC
Signal
Source
(Optimize V;
Minimize i)
i ~ 0
dq/dt
Blocker
i = 0
$\frac{d\phi}{dt} \neq 0$
i ≠ 0
i ≠ 0
LOAD

NOTE: May also utilize a tuned LC oscillator on the secondary side of the transformer

Patent Pending

Figure 8. dq/dt-free driving a swinger by S-flow.

Poynting energy flow S with its accompanying potential V is shuttled from the source to a tuned LC oscillator in resonance at the driving frequency or slightly below. The LC oscillator is driven by S-flow and dV/dt flow alone. The swinger furnishes its own free electrons, which receive S and EMF from the bridge, thereby producing dq/dt.

D-1-F8

© T.E. Bearden 1993, 1994, 1995

[11] E.g., see L. Mandelstam [Mendel'shtam, L.I.], N. Papalexi, A. Andronov, S. Chaikin and A. Witt. (1935) "Report on Recent Research on Nonlinear Oscillations," Translation of "Expose Des Recherches Recentes Sur Les Oscillations Non Lineaires," Technical Physics of the USSR, Leningrad, Vol. 2, 1935, p. 81-134. NASA Translation Doc. TT F-12,678, Nov. 1969. Summarizes lengthy work at the Institute of Physics of the University of Moscow and several of the laboratories on the problem of self-powered, self- oscillating nonlinear oscillators. The paper reports significant work and progress in oscillations created by the oscillating system itself (self-oscillation) without participation of any external forces varying with time and without the expense of a constant source of energy (e.g., a storage battery). Such a system was shown to self-increase in an underdamped manner until system self-destruction. Adding a nonlinear load allowed the systems to stabilize its dynamic operation so that it powered a load under overunity conditions and did not self-destruct. Extensive additional references on this spectacularly successful overunity energy project are listed.

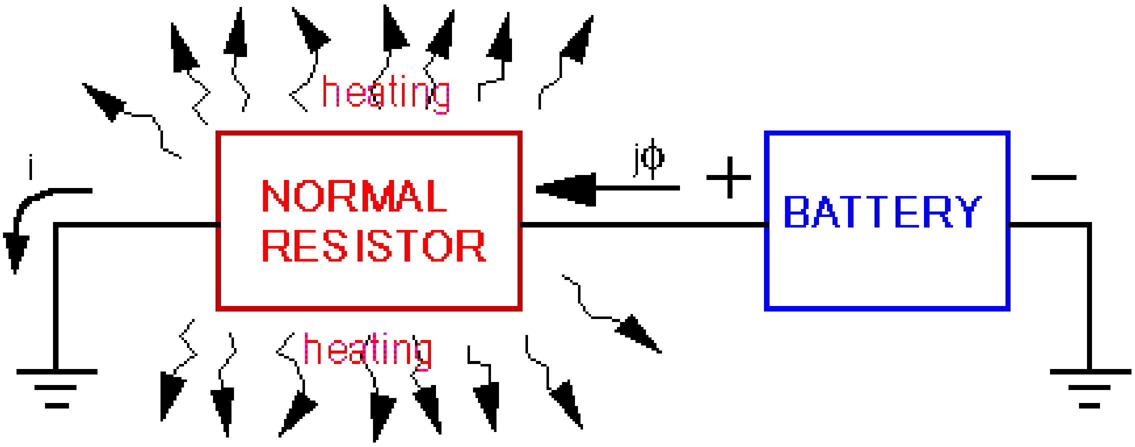
heating
NORMAL RESISTOR
jϕ
+
BATTERY
−
i
heating

Figure 9a. Normal resistor receives ordered energy and scatters it.

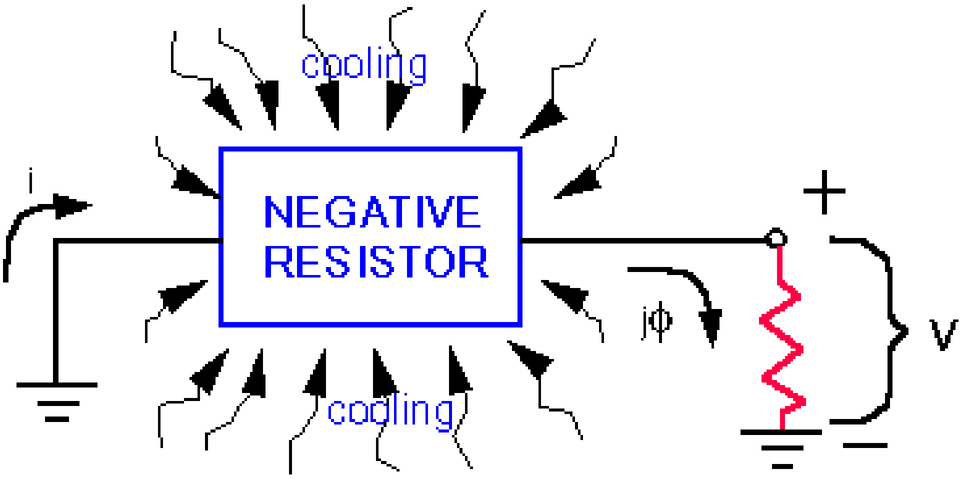
cooling
i
NEGATIVE RESISTOR
+
jϕ
v
−
cooling

Figure 9b. Negative resistor receives, concentrates, and reorders scattered energy.

Figure 9. Normal resistor versus a negative resistor.

D-1-F9

© T.E. Bearden 1993, 1996

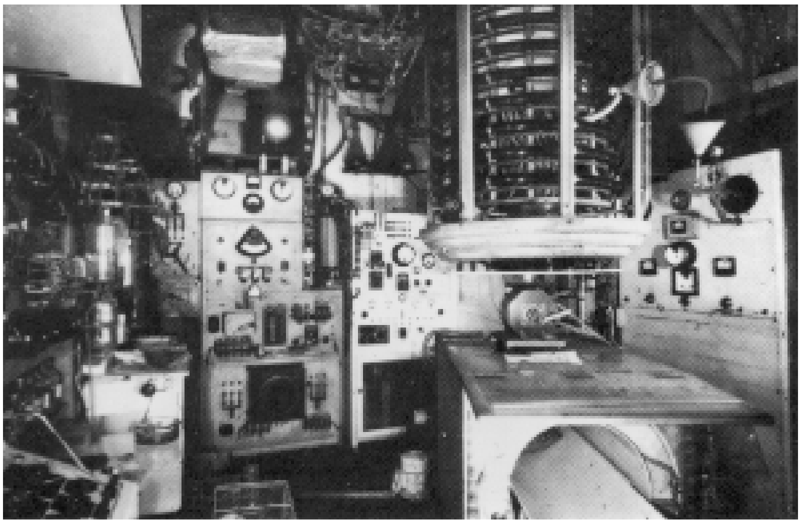

Some 17 signals were fed into a rotating plasma, which added phase conjugates to them. This infolded 17 wavepairs, each pair a coupled wave/antiwave, thereby forming a scalar potential with an artificial Whittaker pump wave internal structure. This infolded pump wave complex was modulated onto a powerful, rippling magnetic field, guaranteeing penetration even into the atomic nuclei. A specific, amplified antiengine for the exact cellular disease state was generated, reversing the diseased cells back to a previous healthy state.

Figure 10. Priore's laboratory apparatus for cellular reversal.

D-1-F10

© T.E. BEARDEN 1993, 1995

[12] For citations and the exposition of our proposed explanation of the Priore mechanism, see T.E. Bearden, "Vacuum engines and Priore's methodology: The true science of energy-medicine." Part I, Explore!, 6(1), 1995, p. 66- 76, Part II, Explore!, 6(2), 1995, p. 50-62.

---

[13] Bearden, "Vacuum engines and...," ibid.

Presently my CTEC colleagues and I have filed three patent applications and a continuation, on various methods for overunity processes and apparatuses, room temperature superconductivity, Poynting generators, etc.

As an example, we took exactly the opposite approach to room temperature superconductivity as has been taken by all the orthodox researchers to date. What exactly is sought, after all, for superconductivity (SC) in a section of an electrical circuit? Well, one has some electrons on one side of that SC section that are transporting excess potential, as given by the Slepian vector J*. [note 14] What is desired is to get some electrons on the other side of the SC section that also are transporting the same amount of excess energy in the form of J*. You can do that in two ways: (1) you can flow the electron carriers through the SC section, carrying their J*, or (2) you can block the J and flow the * itself across as the Poynting flow S. [note 15] Nondivergent Poynting flow flows along an equipotential, which is just another way of saying that, if the S-flow does not diverge, it carries the same potential * along with it. Hence it carries the EMF right along with it as it flows without divergence.

Conventional approaches have all tried to shove the electron carriers through the SC lattice section. Doing that is like trying to fire a very slow bullet through several million rotating fans in a straight line. So cryogenics (to slow the fanblades to a crawl) and massive correlation of the electrons and of the electron-to-lattice interactions is necessary if one is to get the electrons through there without excess collisions that shake off some of the excess * from J*, as scattered photons (heat) or as a change of form of the energy (as in straining the dielectric of a capacitor to convert electrical energy to mechanical strain energy via the piezoelectric effect.)

Our approach is exactly the opposite (Figure 11 and Figure 12). Why not just *stop* the flow of excited electron carriers on one side of the SC section, and continue the nondiverging flow of the Poynting field energy density S across that section at room temperature? Then the cryogenics is not needed at all. After all, circuits already work that way anyhow __ except standard practice is to nullify the process by letting current be driven around the sourcing current loop and back through the back EMF of the primary source. A quantum well (or several other methods) can be used to trap the "sourcing" electrons in the conductor just prior to entering the SC section. The "receiving" electrons on the other side of the SC section, however, must be in their own dq/dt-isolated current loop.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

Comments may be sent to
the **Editor** of The Virtual Times

[14] Briefly, the S-flow outside the wire will be flowing along essentially at the speed of light, at or with the same since S-flow without divergence is along an equipotential. Said another way, a nondiverging Poynting flow S exhibits a constant potential ~ along its streamlines. A tiny, tiny portion of the S-flow is intercepted by the sluggish electrons (which may be moving as dq/dt at 11 feet per hour, in a typical case), to form the Slepian vector J~. The J~ portion is just the minute fraction of the S-flow that is actually intercepted by the electron current and transported to the impedances, for releasing the excess ~ in the loads and losses. [Incidentally, this wee note resolves a century of controversy over exactly how EM energy flows in circuits. The orthodox literature abounds with discussions, differing positions, and debates and with errors.]

[15] Note that this also automatically involves a host of other types of controversial displacement currents, such as d~/dt, dV/dt, dE/dt, dP/dt, dD/dt, etc. The literature is thoroughly floundering as to exactly how all these currents are related, together with the Slepian vector J~ and the Poynting vector S, as well as the "normal" current dq/dt. We will be discussing the "master key" for these relationships in future articles.

Tables:

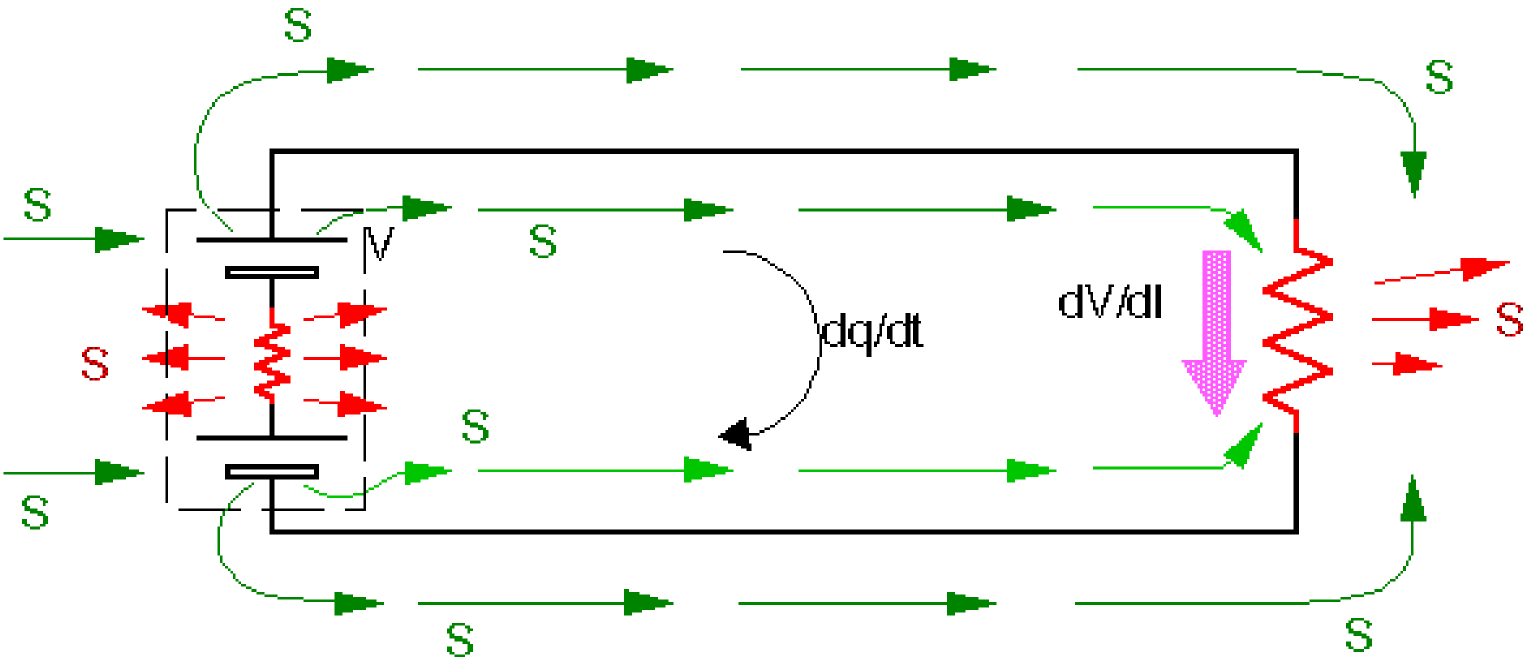

Note: EM texts err in showing the battery "gate" as the source of energy.

Figure 11. Poynting S-flow along a simple circuit.

From particle physics, any electric charge is a broken symmetry in the vacuum's virtual particle flux. Any dipole is thus two locally broken symmetries. Thus the dipole acts as a gate/source, driven by the fierce vacuum flux. So a dipole already "extracts" and gates EM energy flow, directly from the seething vacuum.

Every battery and generator is already a "free energy" source of extracted and gated EM energy from the vacuum.

D-1-F11

© T.E. Bearden 1994

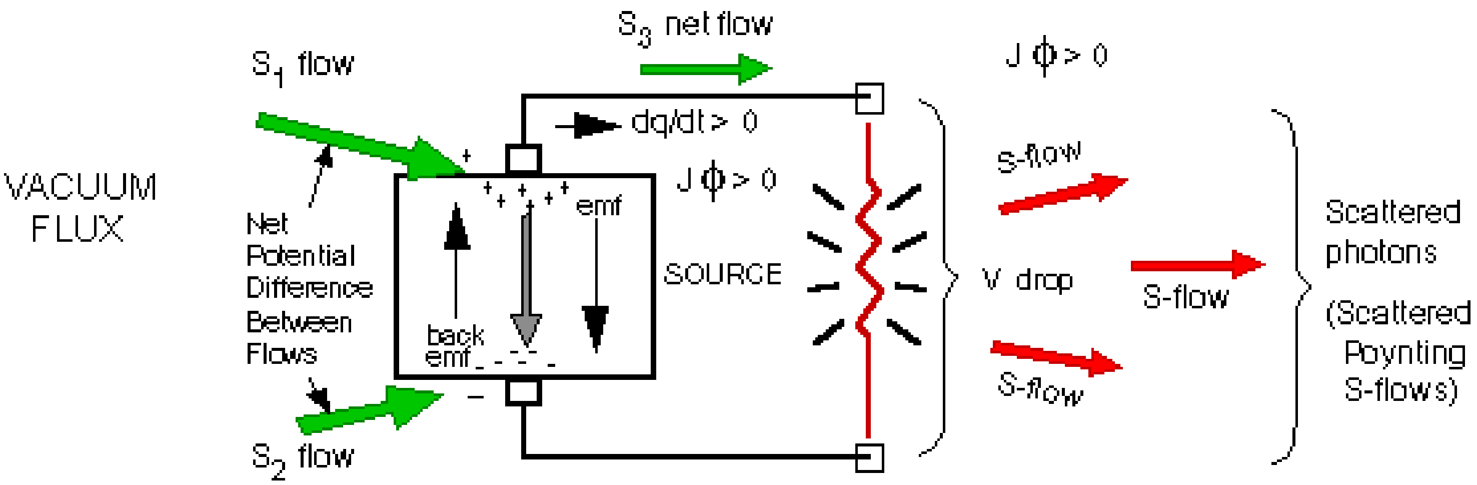
S$_3$ net flow
S$_1$ flow
J ϕ > 0
dq/dt > 0
VACUUM FLUX
+
+ + + + + +
emf
J ϕ > 0
SOURCE
S-flow
Net Potential Difference Between Flows
back emf
- - - - -
V drop
S-flow
Scattered photons
(Scattered Poynting S-flows)
S-flow
−
S$_2$ flow

Figure 12a. Normal current flow; Jϕ is the electron transport of energy, while S is the speed-of-light Poynting energy flow.

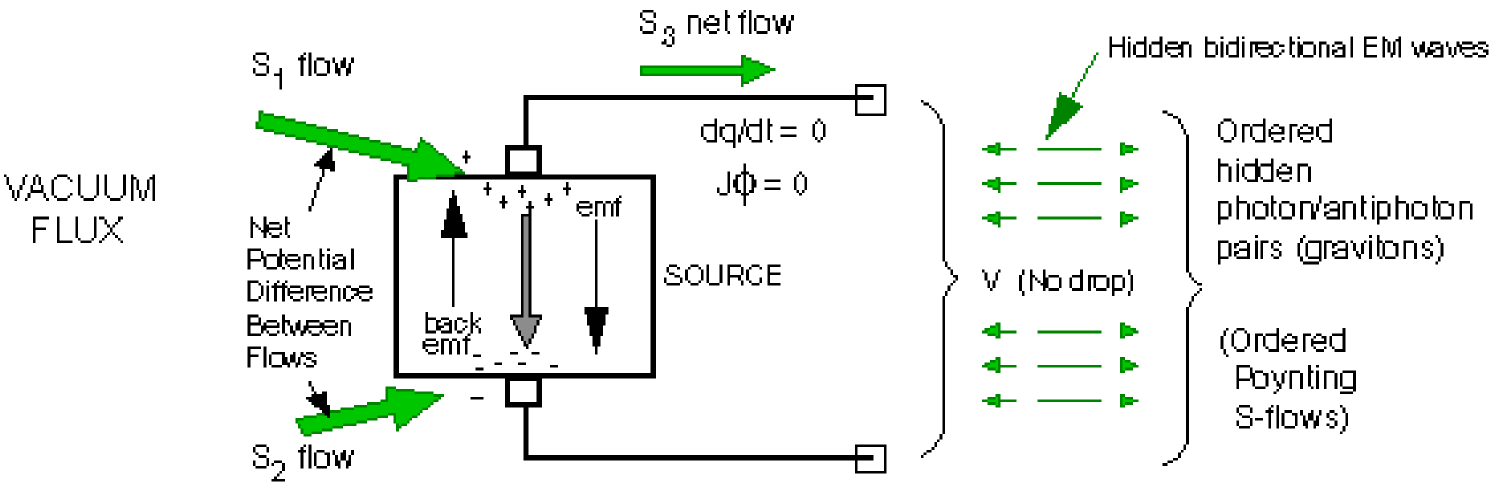
S$_3$ net flow
S$_1$ flow
Hidden bidirectional EM waves
dq/dt = 0
Jϕ = 0
VACUUM FLUX
+
+ + + + + +
emf
SOURCE
Ordered hidden photon/antiphoton pairs (gravitons)
Net Potential Difference Between Flows
V (No drop)
back emf
- - - - -
(Ordered Poynting S-flows)
−
S$_2$ flow

Figure 12b. In a conductor with dq/dt blocked; S-flow and hidden Whittaker-wave flow continue.

Figure 12. Overt and covert field energy density flows.

D-1-F12

© 1994, 1995 T.E. Bearden

Normal sources do not furnish electrons and current to a circuit anyway. Sources just furnish Poynting flow S and EMF. [note 16] In the receiving current loop, the EMF appears automatically once the S-flow flows in and is "locked on to" by the receiving electrons. Given q**, you will immediately have J* if the electrons are free to move in the conducting circuit (see Figure 13). Further, you have eliminated all the loss terms from the standard Poynting equation for energy losses. So all the energy flow S just flows across the SC section, without any current or Cooper pairs flowing through that SC section. The SC section has become a "bridge" which (1) strips off the Poynting field energy density flow S from the electron current dq/dt on the sourcing side of the SC/bridge section, by simply reducing the dq/dt to zero; and (2) continues to pass the S-flow across the dq/dt-blocked SC/bridge section to the other side (Figure 14). The excess S-flow (and EMF) pours into the receiving dq/dt-isolated current loop, exciting the electrons therein to produce dq/dt and J*. *Any closed current loop is self-powered, once it receives S and EMF. It furnishes its own electrons; it only requires excess energy and EMF.* [note 17]

So our approach gives room temperature superconductivity in a very straightforward manner, once you discover how to block the current dq/dt in a conductor. Blocking it in an insulator is not sufficient, because that drops the potential and stops the S-flow and the equipotential * (the EMF). However, a degenerate semiconductor such as the Fogal chip can be used, as can several other processes for blocking dq/dt in a conductor. We will discuss these in a future article.

Another advantage of this approach to room temperature superconductivity is that now one can also have permissible overunity coefficient of performance. Now the load can be placed in its own S-receiving, isolated current loop. With the sourcing current loop furnishing only S and not dq/dt, the load is still powered normally in its own closed dq/dt current loop, but none of the load current is passed back through the back EMF of the primary source in the sourcing circuit.

This principle -- that at least a substantial portion of the load current must not pass back through the primary source -- is the primary principle required for a permissible overunity electrical machine (Figure 15). A permissible overunity electrical machine is one which produces more power in the load than you have to put into the machine to run it.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

[16] E.g., see Robert L. Shrader, Electronic Communication, Sixth Edn., MacMillan/McGraw-Hill, New York, NY, 1991. (1993 Imprint). Quoting from p. 6: "A source of electric energy does not increase the number of ~ree electrons in a circuit; it merely produces a concerted pressure on any aimlessly moving electrons."

Tables:

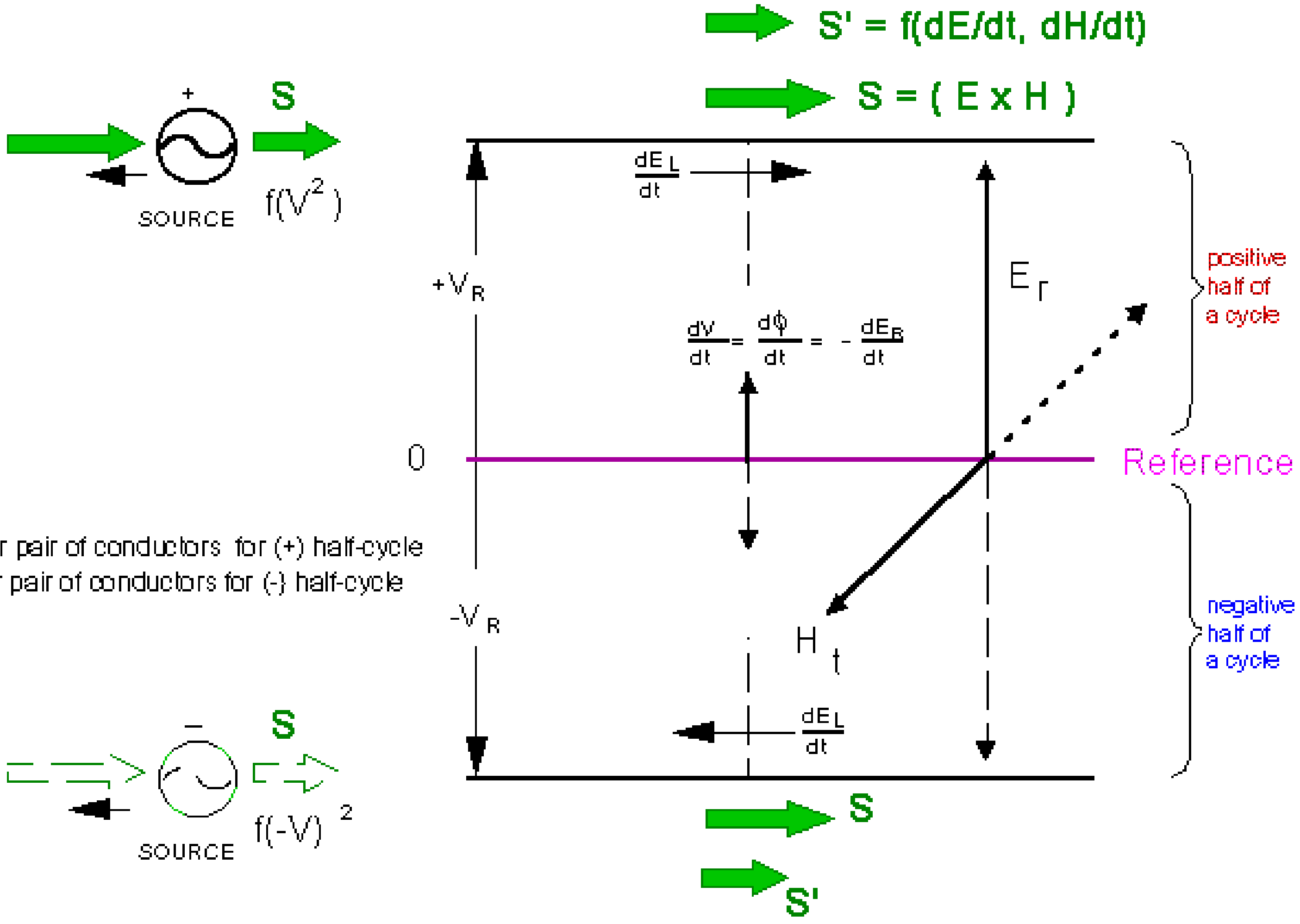
S' = f(dE/dt, dH/dt)
S = (E x H)
+
S
SOURCE
f(V²)
dE L / dt
+V R
E r
positive half of a cycle
dV/dt = dϕ/dt = - dE B/dt
0
Reference
NOTE:
Use upper pair of conductors for (+) half-cycle
Use lower pair of conductors for (-) half-cycle
-V R
H t
negative half of a cycle
–
S
dE L / dt
SOURCE
f(-V) 2
S
S'

Figure 13. Poynting vector currents S and S' versus massless displacement currents.

D-1-F13

© 1995 T.E. Bearden

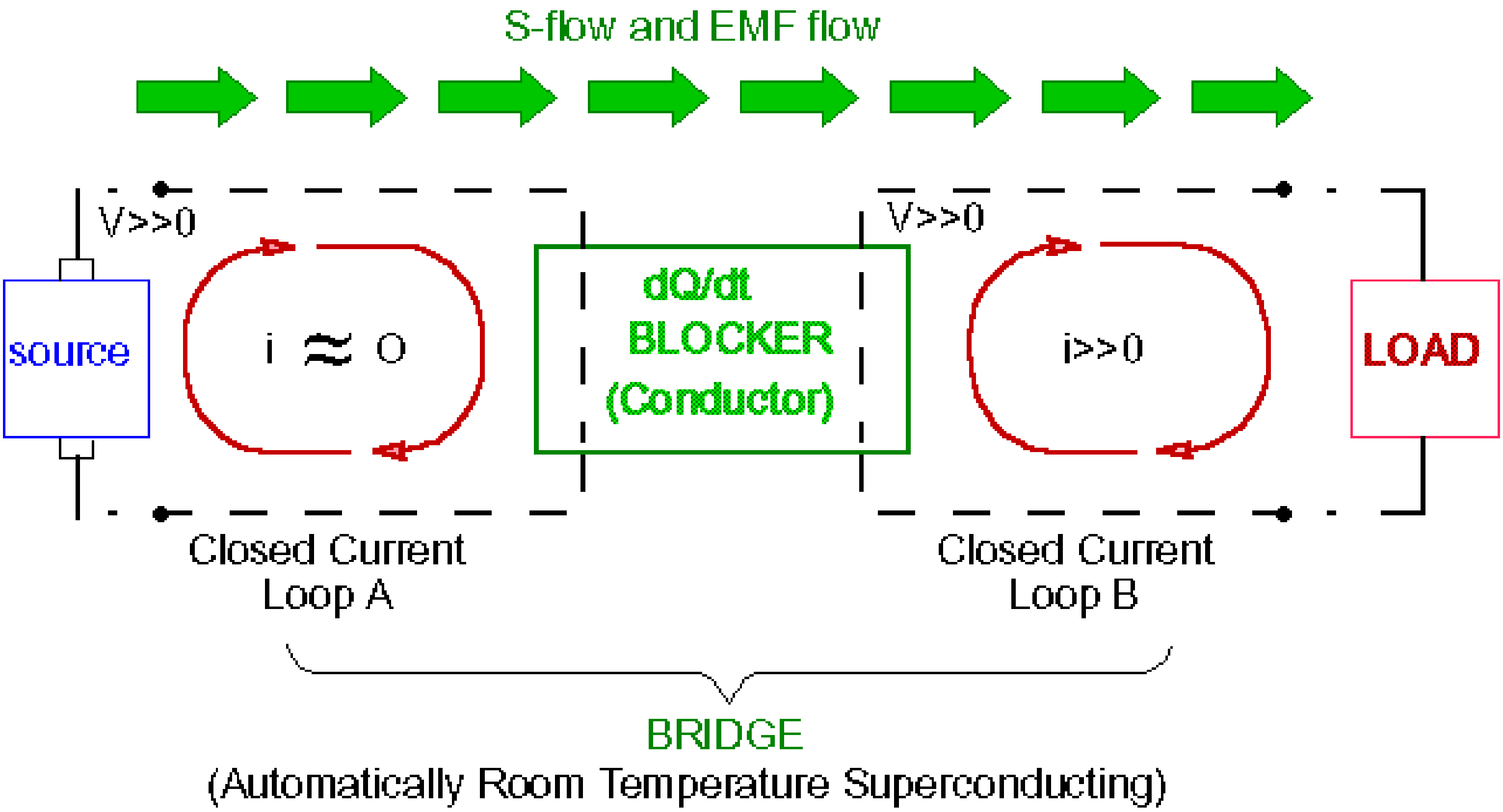
S-flow and EMF flow
V>>0
source
i ≈ 0
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD
Closed Current
Loop A
Closed Current
Loop B
BRIDGE
(Automatically Room Temperature Superconducting)

Figure 14. Bridging to enable room temperature superconductivity.

D-1-F 14

© 1995 T.E. Bearden

[17] My own approach has been, if that's all that the circuits require from our primary power source, then let's simply accept that directly from the source without performing any work in the source. Iœ you do it this way, the source is never dissipated, but will last indefinitely. Sources furnish flowing field energy density, not "energy" per se. If you only demand S-flow, you can have all you wish from a source, without exhausting your source. The trick is to find how to steadily dissipate some of this "artesian well" of flowing S to drive electrons through the load, but without driving electrons back through the back EMF of the source.

Tables:

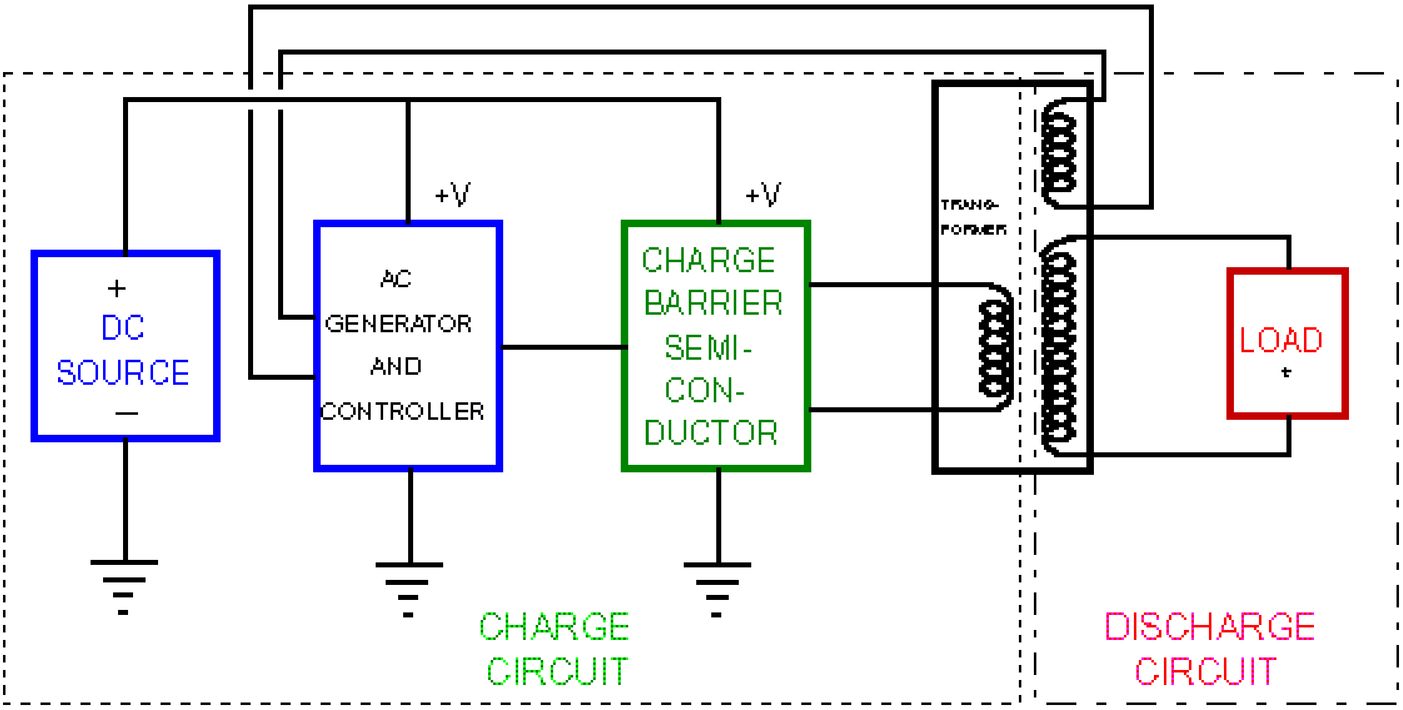
+
DC
SOURCE
—
AC
GENERATOR
AND
CONTROLLER
+V
+V
CHARGE
BARRIER
SEMI-
CON-
DUCTOR
TRANS-
FORMER
LOAD
+
CHARGE
CIRCUIT
DISCHARGE
CIRCUIT

* NOTE:

DISCHARGE CIRCUIT AND LOAD GROUND
MUST BE ISOLATED FROM THE
CHARGE CIRCUIT GROUND.
FEEDBACK LOOP MUST BE RIGIDLY CLAMPED AND LIMITED.

Patent Pending

Figure 15. Use of Fogal semiconductor for overunity (closed system).

D-1-F15

© 1993, 1995 T.E. Bearden

Obviously, the laws of physics and thermodynamics require such an overunity machine to be an open system that extracts and utilizes excess energy from an external source. In the case outlined, we extract the excess energy directly from the vacuum in the form of S-flow; essentially we convert our primary source to a Poynting generator rather than a "power" (energy flow loss rate) generator. The resulting overunity machine accomplishes both room temperature superconductivity and also overunity coefficient of performance (Figure 16). It does not violate the conservation laws of physics. It does not violate the second law of thermodynamics because it is an open system receiving and utilizing excess energy from the vacuum, and the second law need not apply to such systems. [note 18]

You see, all power systems are already free energy density flow sourcing systems, if we but use them correctly. Every dipole is a "gating device" for extracting excess field energy density flow from the violent bidirectional flux exchange between the vacuum and each end of the dipole. [note 19] [note 20] When we attach "conducting leads" to a dipolar source, as in the simple two-wire circuit, all we are doing with the conductors is providing waveguides for the extracted S-flows from the vacuum to flow along (it flows outside the conductors) without loss. In effect, each lead is simply extending the pole (i.e., the terminal) to which it is connected.

My associates and I are also working on several other inventions at present, including two more patent applications to be filed in the near future. One of those is in the field of overunity power systems, while the other is in an entirely different, highly proprietary nuclear effort. I will address all these areas in future columns, and those columns will have real "meat" in them.

Present "big science" proclaims that overunity electrical machines are impossible. The reader should understand that the U.S. Patent and Trademarks Office is never going to grant a patent for an overunity electrical machine unless the machine is independently tested and replicated, and unless it is certified by scientists of impeccable credentials. This is as it should be. Legitimate free energy researchers accept that dictum. Eventually the stuff has to work on the bench in order to be validated. On the other hand, free energy researchers also point out that one must not demand of us the next 100 years of national research by a hundred major laboratories immediately! This is still a shoestring operation, because all conventional sources of research funds are denied to the serious overunity researcher (with the single exception of occasional personal funding of a researcher by a private individual). In other words, we demand at least the same kind of consideration that the orthodox nuclear fusion community has enjoyed. Billions of dollars have been poured into that sinkhole, and not a single controlled fusion power unit is working anywhere in the world after decades of work at enormous expense to the taxpayers.

So we overunity researchers deserve at least a decade or two of our own, to try to accomplish overunity electrical system development. We also painfully point out that "Big Science" will not fund a single overunity researcher, for specific overunity purpose, no matter what his credentials or capability. In fact, "Big Science" is part of the problem, because it fervently seeks to prevent any legitimate research in overunity. To carry out its suppression campaign. it resorts to vilification, character assassination, libel, and slander of the overunity researchers themselves __ again, regardless of their qualifications or legitimate scientific procedures. [note 13]

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

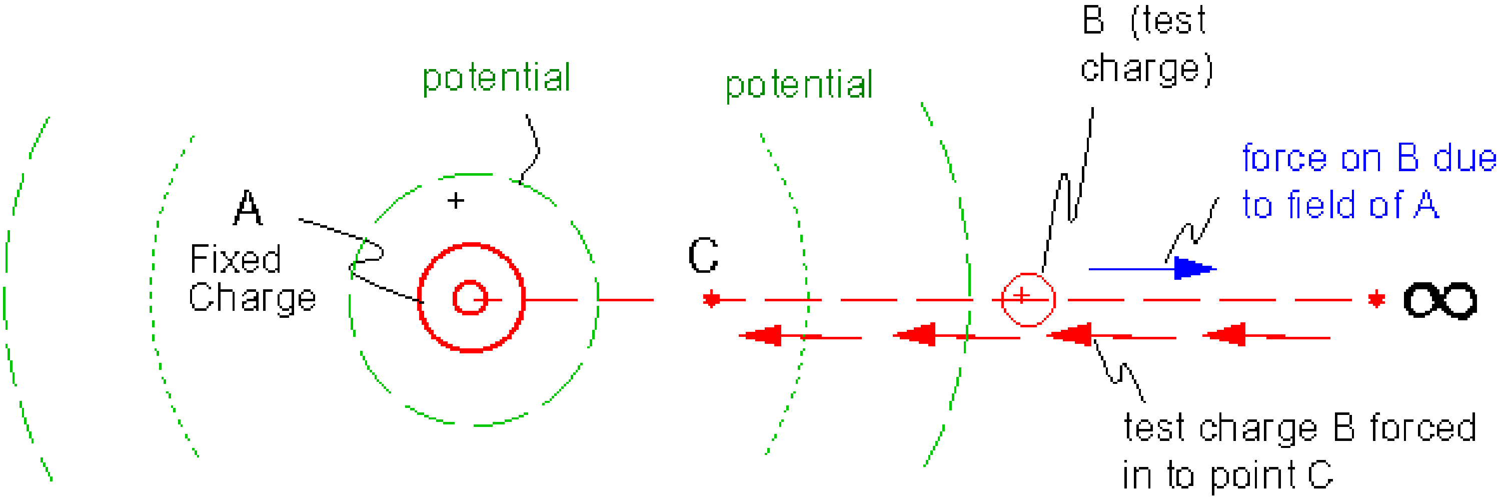
potential
potential
B (test
charge)
force on B due
to field of A
A
Fixed
Charge
+
C
+
∞
test charge B forced
in to point C

Figure 16a. Magnitude of the scalar potential is determined by work required to push in a unit charge from infinity; e.g., against a repelling like charge. If test charge is released, work will be performed upon it to translate it back toward infinity

Figure 16a. Process for overunity coefficient of performance.

D-1-F16A

© T.E. Bearden, 1993, 1994, 1995

Tables:

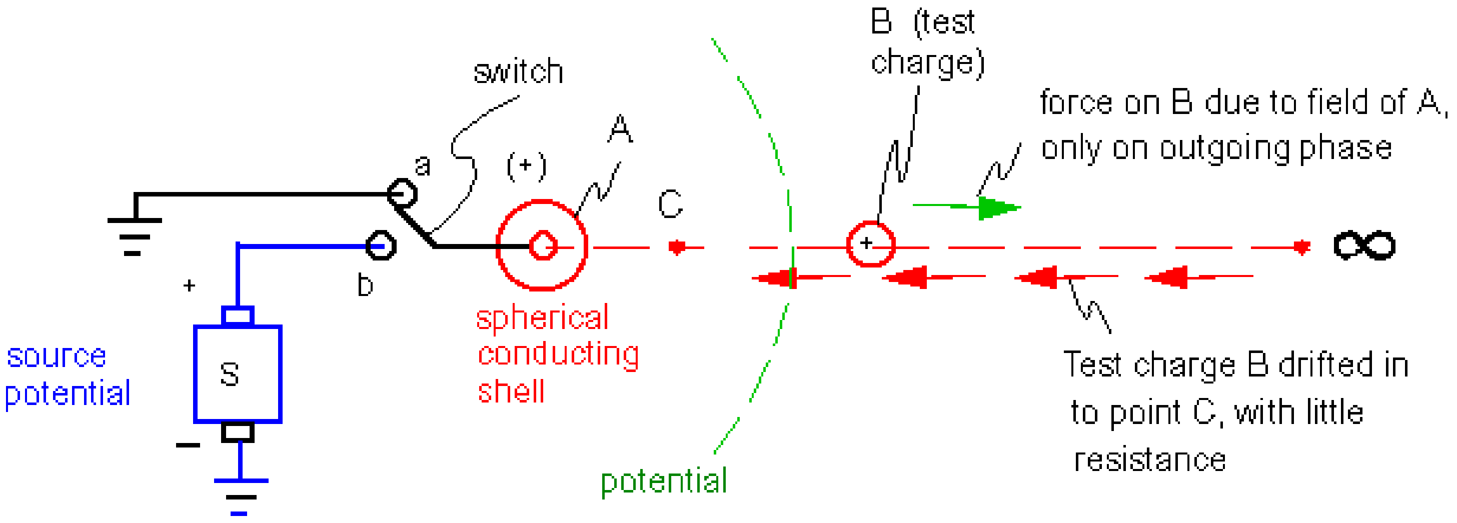
switch
a
b
(+)
A
C
B (test charge)
force on B due to field of A, only on outgoing phase
∞
+
S
−
source potential
spherical conducting shell
potential
Test charge B drifted in to point C, with little resistance

Assumptions: Capacitance of A very small; charge B large, power-free switching

Switch potential off, drift in test charge B, switch potential back on. Can now get free work from repelled test charge B. Get much more free work if A is at a high potential, and charge B is large.

Figure 16 b. Demonstration of overunity efficiency.

D-1-F16B

© T.E. Bearden, 1993, 1994, 1995

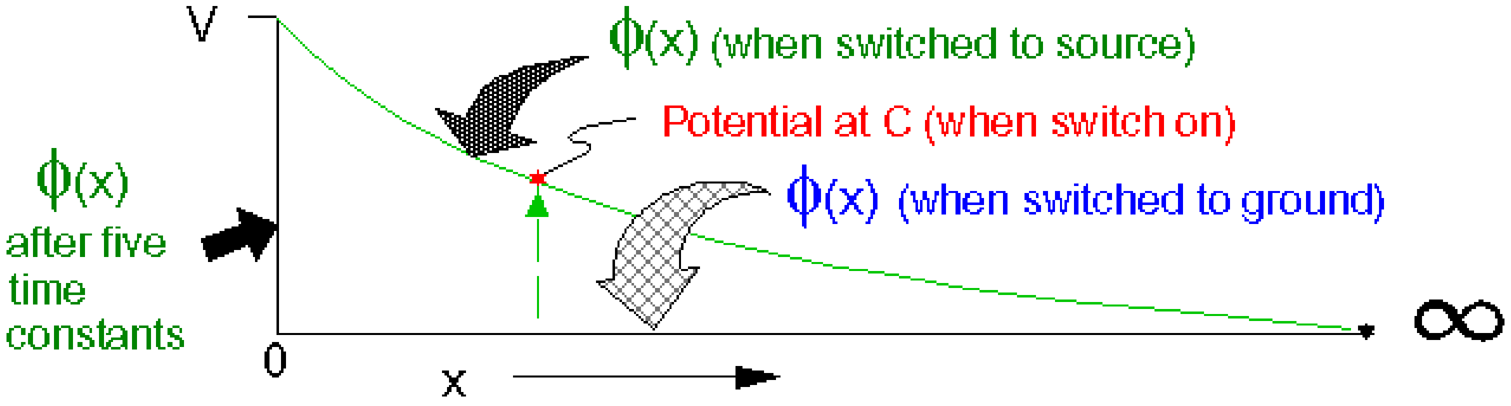

Ideally, switching the potential, rather than flowing charges, requires only minuscule overhead energy dissipation

Figure 16c. Potential to which test charge is exposed.

D-1-F16C

© T.E. Bearden, 1993, 1994, 1995

[18] The system does comply with the nonlinear thermodynamics of a system œar from equilibrium. E.g., Prigogine received the Nobel Prize œor significantly developing this new field. For a detailed technical presentation of nonlinear thermodynamics of a system far from thermodynamic equilibrium, see G. Nicolis and I. Prigogine, Exploringq Complexity, Piper, Munich, 1987. Such systems can exhibit negentropy, self-organization, and other effects outside the confines of classical thermodynamics. Since local symmetry is broken, they can also permissibly exhibit overunity coefficient of perœormance.

[19] Each end 0œ a dipole has its own potential (with respect to the local vacuum), which differs from the potential of the surrounding ambient vacuum. For each of these "end-potentials," one can mathematically decompose that end-potential into a hidden bidirectional set of harmonic wavepairs in harmonic sequence. Each wavepair consists of the wave and its phase conjugate. For the proof, see E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annglen, Vol. 57, 1903, p. 333-355. Since one of the dipole potentials exceeds the local vacuum potential and the other dipole potential is lower than the local vacuum potential, the two bidirectional EM wave flows are at or "carry" different equipotentials. This is the generatrix for the automatic dual œ1ow of energy from the vacuum through the dipolar power source and out along the two leads of the "transmission line" conducting the S-flow to the components of the circuit for powering the loads. So as can be seen, rigorously every dipole is already a "free energy source" if we learn how to properly use it. The energy will flow forever, if we do not allow the destruction of the sourcing dipole.

[20] See also E.T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," Proceedinqs of the London Mathematical Society, Series 2, Vol. 1, 1904, p. 367-372. Here Whittaker shows that the fields of classical electromagnetics can be replaced by scalar potential interferometry of two potentials. Our own comment is to note that by Whittaker 1903, we can with a little difficulty make artificial potentials with deliberate substructures. We can then interfere those potentials (actually, interfere their hidden multiwave structures) even at a distance, to create new EM fields. Or we can dissolve EM fields that exist at a distance, simply by constructing the correct negative EM fields by distant scalar interferometry. My point is that, when we utilize the "inner" electromagnetics inside the potential, we are employing a vastly more powerful electromagnetics of a new kind, which includes and extends the present inadequate EM theory.

Indeed, orthodoxy has a black record with respect to energy. When Mayer advanced the conservation of energy law, orthodoxy hounded him unmercifully. He was universally ridiculed as the very epitome of a fool. He lost his job and suffered a nervous breakdown. Then years later toward the end of his life, scientists came to accept the conservation of energy as a most useful tool that dramatically simplified much of their analyses. So then science began to laud Mayer. There are a thousand other such examples; suffice it to say that "big science" is a bureaucracy and a "bell-shaped distribution curve" just like any other large group of people. Some scientists are near-angels. Others are near-devils. The vast majority are neither, but are just ordinary persons doing a special kind of job. In the scientific bureaucracy, however, the manipulative scientists scurry, slash, and manipulate their way toward the top. So the leadership of any big bureaucracy __ including the scientific bureaucracy __ is always rife with such near-scalawags and power-hungry individuals. Scientists are not exceptions to the same human weaknesses that we all evidence. If you really want to see jealousy, backbiting, and fierce "back-room dealing" and power-struggling, just get into a university research environment! But because the "power merchants" of any bureaucracy always fiercely resist any challenge to their position or superiority, big science has always fiercely resisted anything truly revolutionary. And today they also enlist the power of the state to enforce their dictums. We will devote one of our future columns to pointing out some glaring examples of persecution of independent researchers by big science. It is an area that all unorthodox researchers should be well-aware of. The world of science does not run by sweet reason and lofty ideals; that is just the dogma. It runs strictly by "who's going to be the big monkey," just like everything else. Primate dominance is still a dictum of the functioning of all bureaucratic human systems.

The free energy researchers are not working on the so-called "perpetual motion machine," where a closed system is erroneously stated to produce more output than must be put into it to operate it. That old saw is actually a red herring that dogmatic scientists use to vilify free energy researchers. It is a lie that they have also successfully sold to the U.S. Patent and Trademarks Office. Anyone with a modicum of knowledge of modern physics already knows that there does not exist a single closed system in the universe, anywhere. In particular, every physical system is continuously "open" to the violent exchange of energy with the surrounding vacuum. It's just that most systems are in local equilibrium in these flow exchanges, except for minuscule gating accomplished for such things as the Lamb shift, etc.

By analogy it's something like this. The free energy researcher is standing by the banks of a mighty, rushing river with breathtaking falls, turbulent areas, etc. Big Science is begrudgingly admitting that, yes, there is a river there, but it is an insane delusion to think that you can tap that energy. The free energy researcher is saying that, well, if I can build a gating sluice a little distance upstream, perhaps I can divert a small portion of the river's flow downstream to my waterwheel, and power my mill. The orthodox scientist then vilifies the free energy researcher for even having such a heretical thought. In fact, says the orthodox researcher, any fool can see that the laws of nature force the river's flow to stay firmly in its banks, because it is a closed system, and Moses brought that law down off the mountain with him on a special stone tablet. Then he adds all the other smug remarks such as, "If it could be done, then we at MIT and Harvard and Caltech would already have done it!" So until the orthodox scientist develops a little less dogma and a little more common sense, and learns the difference between a scientific discussion and a dogfight, the free energy researcher can hardly communicate with him.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page 1

Comments may be sent to
the **Editor** of The Virtual Times

The goal of the legitimate free energy researcher is to find a way to break the local equilibrium of this universal energy exchange, and gate a tiny bit of it up into the external circuit as excess electrical energy (Figure 17). Then he must collect the excess energy, and shuttle it from the "gating and collecting circuit" over to a dq/dt-closed current loop containing the load. He must also prevent the load current, or most of it, from passing back through the primary source against its back EMF, because if it does that, he will destroy the separation of charges in the dipolar source, killing the dipole and extinguishing the "excess energy gating." In other words he must find a way to process and transport the energy flow across dq/dt-blocked bridges between the source's closed dq/dt-current loop and the load's closed dq/dt-current loop. Ruthlessly it is an energy transport S-flow problem, not a J* energy transport problem. If he uses J* in his sourcing current loop, he automatically drives the current dq/dt back through the back EMF of his source and kills it. If he does not drive dq/dt back through the source, the source will furnish S-flow for an indefinite period! Every battery and generator we have ever built is already a "free energy" machine in that it already involves broken local symmetry in the vacuum energy exchange. We have just not realized how to use our power sources purely as Poynting S-flow sources.

So we must treat an electrical system as an open system with broken local symmetry (Table 3) in the vacuum flux exchange, so that the system extracts (gates) excess energy from an external source. In this case the source is the system's flux exchange with the vacuum. This approach is no more mysterious or bizarre than putting a waterwheel in a river, or a windmill in a wind, or a bank of solar cells out in the sunshine. The universal "free source of energy" that the overunity researcher seeks to tap is the violent exchange of virtual photons between the electrical charges of the system and the surrounding quantum mechanical vacuum (Figure 18). In quantum physics this powerful, energetic exchange of the vacuum with electrical matter is now proven both experimentally and theoretically. It is already accepted in quantum physics that the vacuum is filled with electromagnetic energy. The researcher doesn't have to prove it anymore; he just has to find out how to properly use it. The artesian well is already there; we just have to learn how to collect and use the flowing water without dynamiting the well!

The remaining objection orthodox scientists usually raise against the notion of extracting vacuum energy has been that "thermodynamics doesn't permit it." However, Cole and Puthoff have rigorously shown that, on the contrary, the theory of thermodynamics does not prohibit the vacuum's energy being extracted and used to produce heat, light, and power. [note 22] So the validity of that final objection has now vanished.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Open system receiving excess energy from an external source

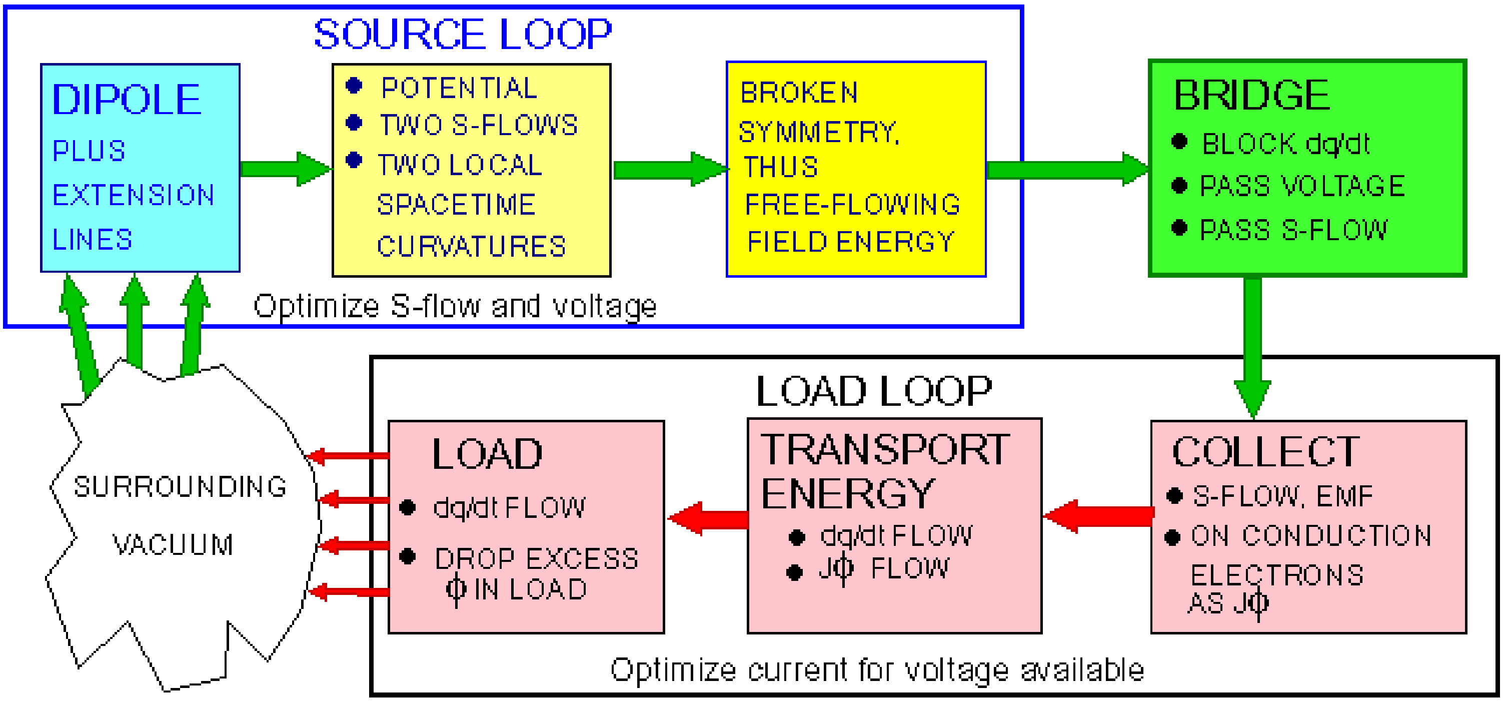
SOURCE LOOP
DIPOLE
PLUS
EXTENSION
LINES
● POTENTIAL
● TWO S-FLOWS
● TWO LOCAL
SPACETIME
CURVATURES
BROKEN
SYMMETRY,
THUS
FREE-FLOWING
FIELD ENERGY
Optimize S-flow and voltage
BRIDGE
● BLOCK dq/dt
● PASS VOLTAGE
● PASS S-FLOW
SURROUNDING
VACUUM
LOAD LOOP
LOAD
● dq/dt FLOW
● DROP EXCESS
ϕ IN LOAD
TRANSPORT
ENERGY
● dq/dt FLOW
● Jϕ FLOW
COLLECT
● S-FLOW, EMF
● ON CONDUCTION
ELECTRONS
AS Jϕ
Optimize current for voltage available

Figure 17. Operation of an overunity electrical power system.

D-1-F 17

© 1992 T.E. Bearden

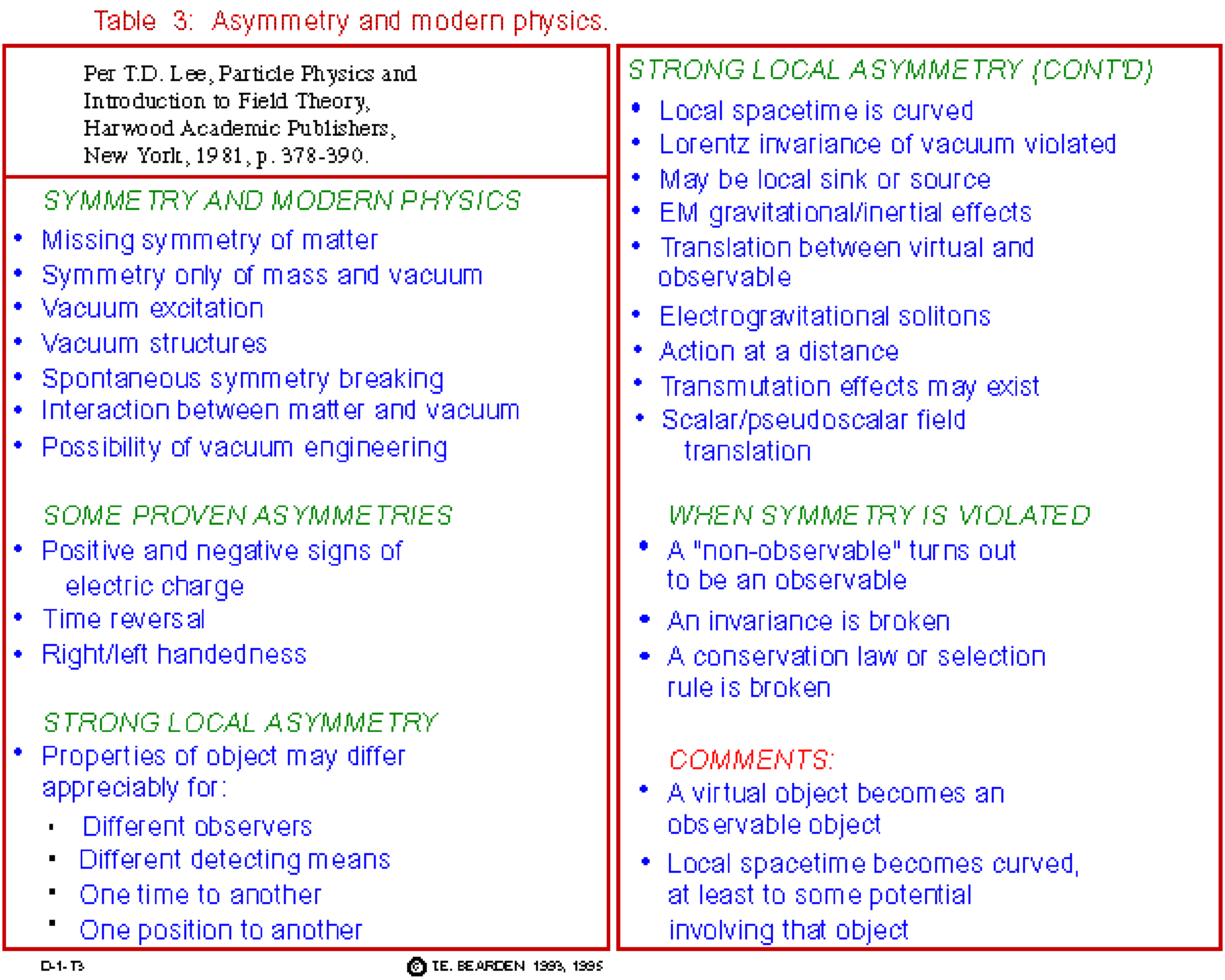

Table 3: Asymmetry and modern physics.

Per T.D. Lee, Particle Physics and
Introduction to Field Theory,
Harwood Academic Publishers,
New York, 1981, p. 378-390.

SYMMETRY AND MODERN PHYSICS

- Missing symmetry of matter
- Symmetry only of mass and vacuum
- Vacuum excitation
- Vacuum structures
- Spontaneous symmetry breaking
- Interaction between matter and vacuum
- Possibility of vacuum engineering

SOME PROVEN ASYMMETRIES

- Positive and negative signs of electric charge
- Time reversal
- Right/left handedness

STRONG LOCAL ASYMMETRY

- Properties of object may differ appreciably for:
    - Different observers
    - Different detecting means
    - One time to another
    - One position to another

STRONG LOCAL ASYMMETRY (CONT'D)

- Local spacetime is curved
- Lorentz invariance of vacuum violated
- May be local sink or source
- EM gravitational/inertial effects
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation

WHEN SYMMETRY IS VIOLATED

- A "non-observable" turns out to be an observable
- An invariance is broken
- A conservation law or selection rule is broken

COMMENTS:

- A virtual object becomes an observable object
- Local spacetime becomes curved, at least to some potential involving that object

D-1-T3

© T.E. BEARDEN 1993, 1995

A Charged Particle Is a Coupled System

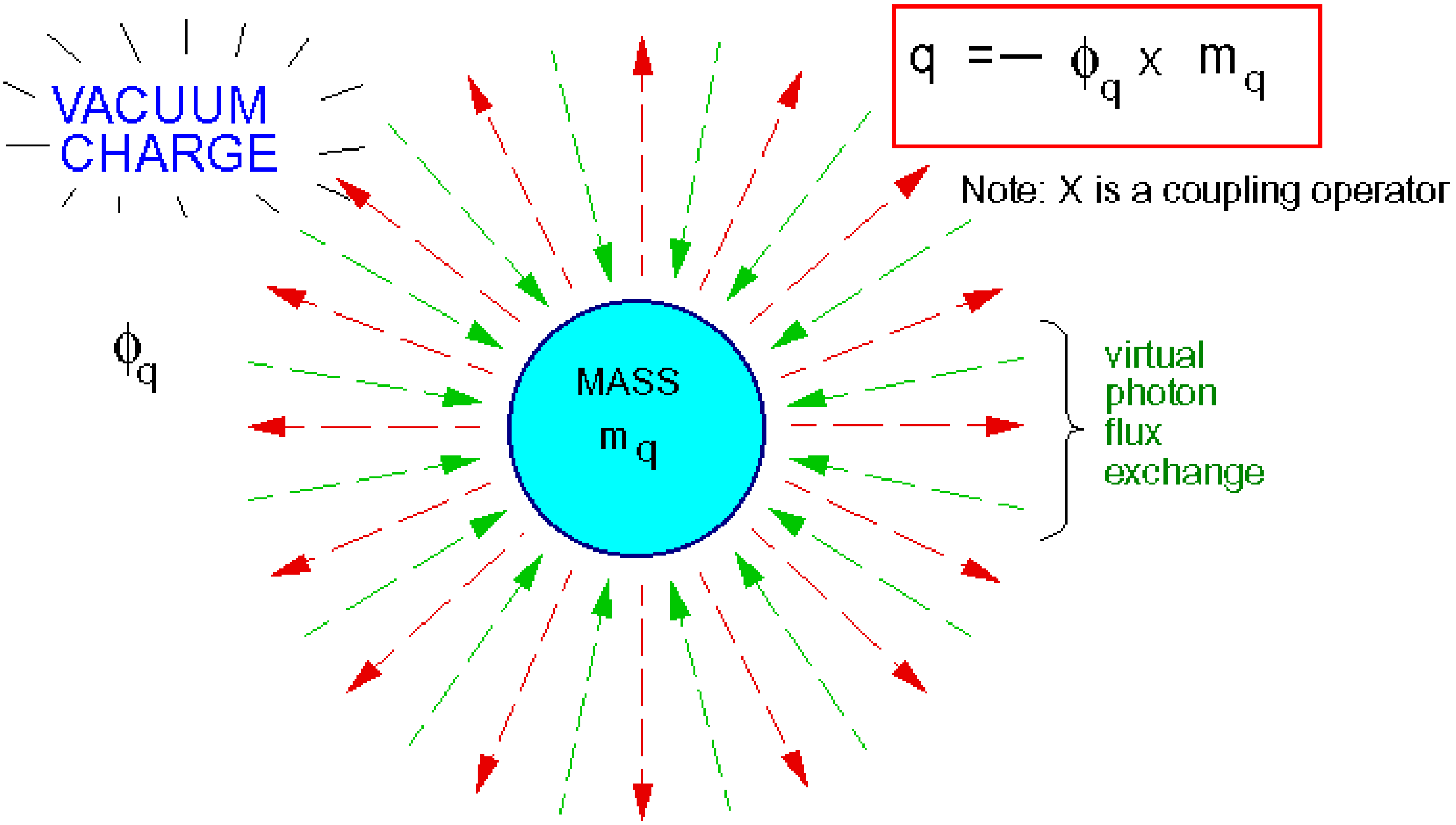
VACUUM CHARGE
$q = - \phi_q \times m_q$
Note: X is a coupling operator
ϕ_q
MASS
m_q
virtual photon flux exchange

Figure 18. An electric charge q is a broken symmetry. It continuously and violently exchanges energy with the vacuum.

D-1-F 18

© 1994 T.E. Bearden

[22] Cole, Daniel C. and Harold E. Putboll. (1993) "Extracting Energy and Heat from the Vacuum," Physical Review E, 48(2), Aug. 1993, p. 1562-1565.

And there is mind-numbing energy density there in the vacuum, waiting to be extracted. The well-known Lamb shift, e.g., in a single hydrogen atom is due to that vacuum exchange. [note 23] Interestingly, the calculated energy density of the exchange causing the Lamb shift is greater than the surface energy density of the sun. Different physicists have made various theoretical estimates of the energy density of this seething vacuum. A conservative ballpark figure is something on the order of 1090 grams per cubic centimeter, expressed in mass units.

It has also been my good fortune to know and work closely with several other free energy researchers. I was a colleague of Floyd ("Sparky") Sweet (recently deceased) for about seven years; Sparky held a Master's degree in EE from MIT and had years of experience with General Electric and other firms. He was a highly skilled researcher who invented the solid state vacuum triode amplifier (Figure 19), a unit which produced 500 watts of output and a coefficient of performance of 1,500,000. I personally tested the machine, activated the magnets on one occasion, and understand how it all worked. [note 24] The paper [note 25] I co-authored in 1991 with Floyd is still the only paper in the orthodox engineering literature that reports (a) successful overunity functioning of a solid state magnetic device, and (b) a successful antigravity experiment on the laboratory bench, reducing the weight of an object by 90%. [note 26]

We stress again that a legitimate overunity electrical machine must be an open system, receiving excess electrical energy from the surrounding vacuum. As such, it must operate in a higher topology than normal electrical machines (Figure 20).

My good friend Frank Golden has long been a close colleague. Years ago Frank developed a motor that produced overunity efficiency, but then to our astonishment we found (courtesy of Bill Tiller, [note 27] with whom we were working at the time) that its operation depended upon a 5-year-long conditioning (structuring) of the local ambient potential in order to work. Previously we had been entirely unaware that one can sometimes condition the local vacuum and the local matter in the area with a particular dynamic form. Given the local conditioning, a motor with that form will then work in that vicinity, but not necessarily elsewhere at an appreciable distance away. We now understand the mechanism for such "conditioning of the vacuum." We also now know what to do if an overunity machine is actually a machine of that type (many of them are not). In other words, we know how to convert it to a machine that will work anywhere, anytime. We will be addressing such little-known things in future articles.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

[23] Lamb, W.E. Jr. and R.C. Retherford. "Fine structure of the hydrogen atom by a microwave method." Physical Review, Vol. 72, 1947, p. 241-243. Lamb received the Nobel prize for this work.

Tables:

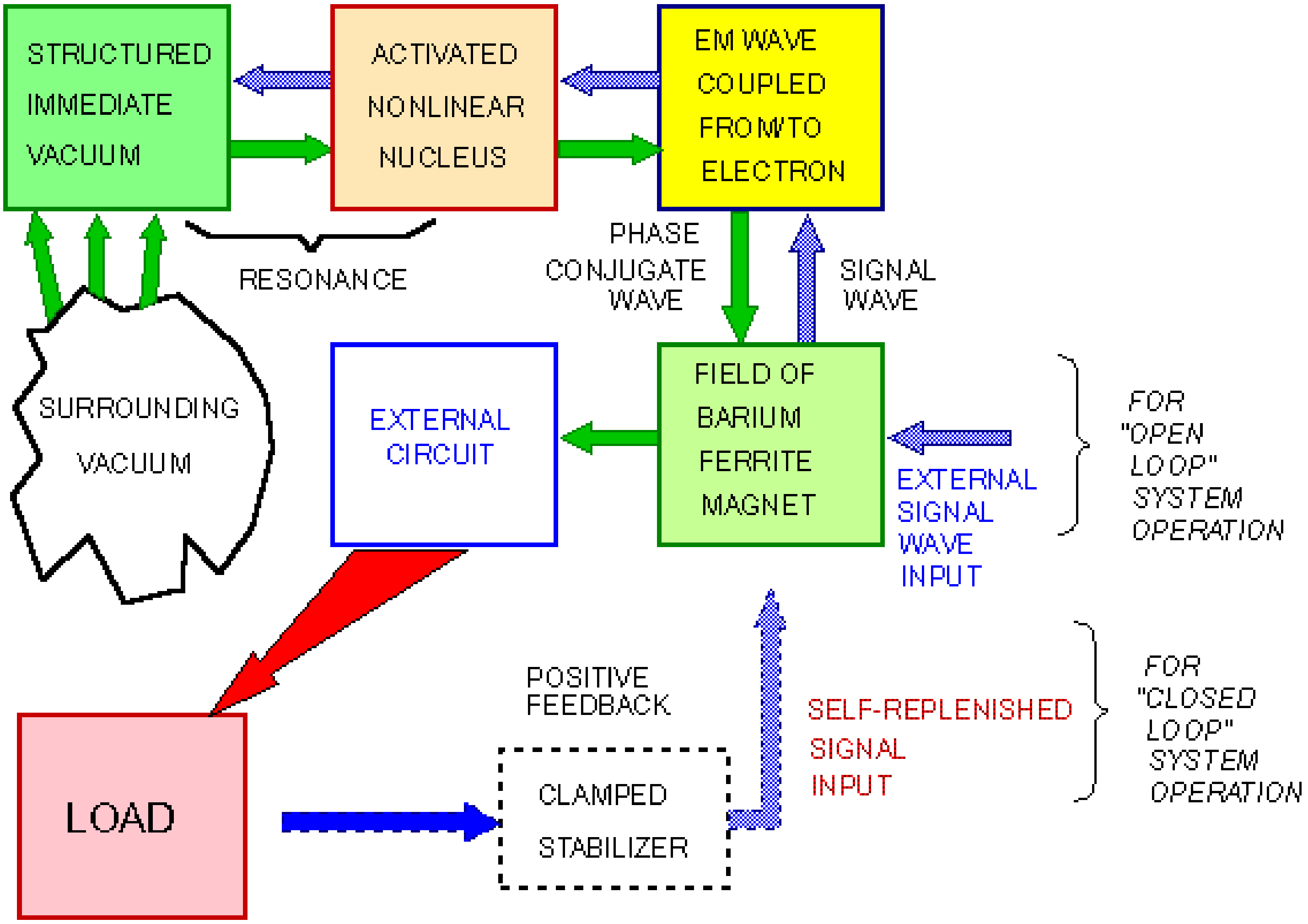
STRUCTURED IMMEDIATE VACUUM
ACTIVATED NONLINEAR NUCLEUS
EM WAVE COUPLED FROM/TO ELECTRON
RESONANCE
PHASE CONJUGATE WAVE
SIGNAL WAVE
SURROUNDING VACUUM
EXTERNAL CIRCUIT
FIELD OF BARIUM FERRITE MAGNET
EXTERNAL SIGNAL WAVE INPUT
FOR "OPEN LOOP" SYSTEM OPERATION
POSITIVE FEEDBACK
SELF-REPLENISHED SIGNAL INPUT
FOR "CLOSED LOOP" SYSTEM OPERATION
LOAD
CLAMPED STABILIZER

Figure 19. Block diagram of the Sweet Vacuum Triode Amplifier.

Adding positive feedback circuit enables self-oscillation.

D-1-F 19

© T.E. Bearden 1991, 1995

[24] Unfortunately Sweet made conflicting agreements with multiple backers, resulting in a most thorough legal foul- up. It appears that nothing can ever be done with this machine, because its status is a legal nightmare.

[25] Floyd Sweet and T. E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," Proceedings of the 26th Intersociety Energy Conversion Engineerinq Conference (IECEC '91), Boston, Massachusetts, 1991, p. 370-375. Paper also distributed over Internet.

[26] The unit was specially rigged and "pushed" to 1,000 watts output to perform the antigravity experiment.

Example: Conservation in 2D versus 3D topologies.

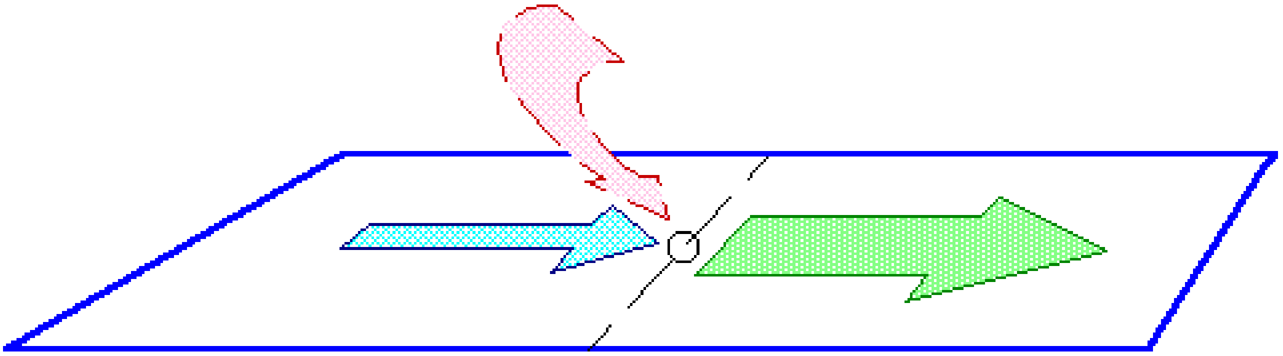

3-D CONSERVATION
(2-D CONSERVATION VIOLATED)

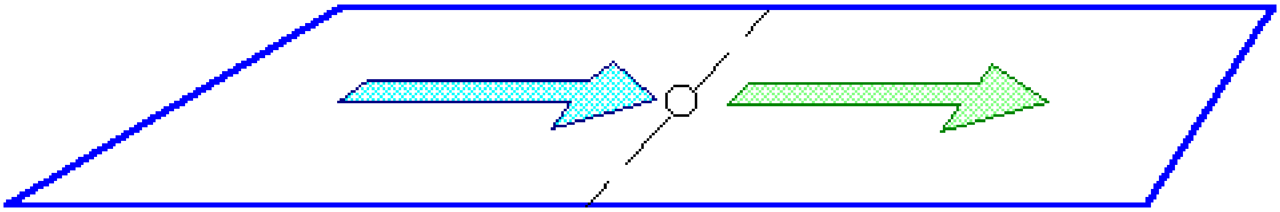

2-D CONSERVATION

Figure 20. Increasing the dynamic topology allows lower topology conservation laws to be violated. Conservation now applies in the higher topology, but not necessarily in the lower. The lower topology system is now an open system and may seemingly act as a source or sink.

D-1-20

© T.E. Bearden 1993, 1995

[27] Tiller was already familiar with the fact that devices that function by more subtle energy often require special conditioning of the vacuum spacetime. He expressed it as "an archetype must first be created for the machine to function efficiently." Tiller had previously experimented with such an effect for several years.

John Bedini is also a close friend and colleague; John built several experimental motors (both electrical and magnetic) in the overunity area, and performed successful trans-mutation experiments. John is a recognized genius in high-end sound amplifier development. Many audiophiles worldwide still swear that the Bedini amplifier is the best and sweetest-sounding audio amplifier ever built. One of John's battery-powered electrical motors (Figure 21), for example, ran continuously off its battery for about five years, and kept the battery charged. When you realize that such a small electric motor is only about 35% efficient, then you realize that about 65% of the energy flowing out of the battery was being dissipated in the motor as heat, core losses, etc. So the unit was continuously performing work for that five years. [note 28] The 1/8 hp motor represented a load in which the continuous rate of work being done (the rate of energy dissipation) was about 0.08 hp. We will have more to say on this motor and its technical process in a future article.

Bill Fogal is a close colleague; Bill has patented the world's first dq/dt-blocking semiconductor, [note 29] which partially blocks the normal current while continuing to pass the flow of voltage (Table 4). I have written a proprietary technical explanation of the Fogal semiconductor, which utilizes an extremely little-known feature called the "overpotential." The overpotential part of the theory is experimentally proven and theoretically explained in an obscure part of electrochemistry, by a few chemists who specialize in electrode effects. J. O'M. Bockris is probably the leading world expert in overpotential theory; he authored the textbook on it. [note 30]

One of the latest overunity developments is the magnetic resonance amplifier (MRA) developed by Joel McLain and Norm Wooten (Figure 22). Earlier testing disputed the overunity functioning. More recent testing at several well-known laboratories has resulted in the instrumental measurements clearly showing overunity. While the test institute scientists themselves may be willing to attest to overunity results of their testing, none of the corporate heads will allow it. I have personally stated that the MRA is capable of overunity. Our own measurements had to be discarded because we did not have the proper instruments available. E.g., all our oscilloscopes were "hard grounded" to building ground, and we needed to make "floating differential ground" measurements in order to settle things once and for all. McLain and Wooten have filed for a patent, and multiple tests with excellent equipment have indeed indicated overunity. What happens next remains to be seen. I suspect that the two inventors may simply accept foreign capital investment and go offshore.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

WARNING: Highly dangerous operation. Pulsed battery outgasses hydrogen gas and explosion is possible.

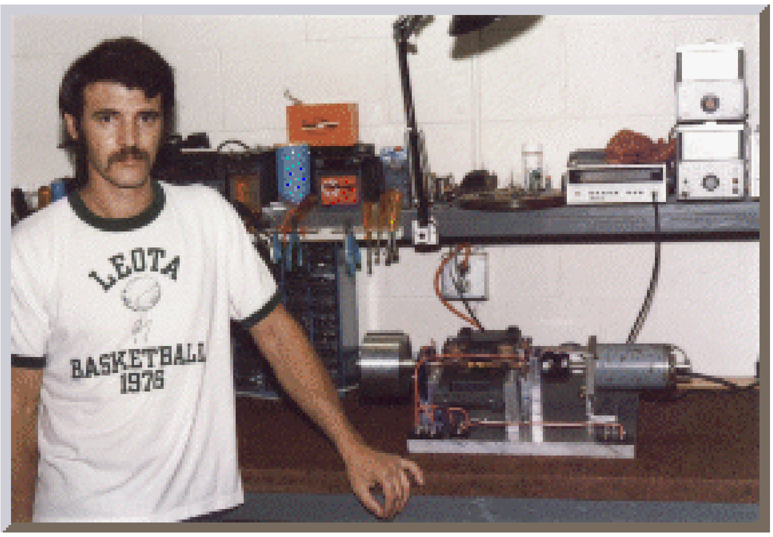

Open system using higher topology extraction of excess vacuum energy. Uses sluggish reversal and recoil time of lead ion onrushing current versus rapid backlash of amplified voltage discharge from a magneto ignition coil, to force the battery to act as an excess S-flow collector during the ion-current charging half-cycle. Thus the inefficient motor is continuously powered and the battery is continually recharged.

Figure 21. Bedini's battery-powered, back-popping electric motor.

D-1-21

© T.E. Bearden 1993, 1995

[28] We strongly emphasize that there is no conservation of work law in physics! It is the overall energy of the universe that must be conserved, not work. More rigorously, the flow of energy must be conserved. But from a free flow of energy, you can continuously extract a little and dissipate it as work, without cutting off the main flow and without violating any of the laws of physics. If this were not so, then waterwheels, windmills, and solar cells would not work. In fact, you can even do better than that. The basic energy conservation law is this: Energy can neither be created nor destroyed. So even when you "use" (i.e., dissipate or scatter) the energy, it is still there. It is only the hierarchical ordering of the photons that constitutes entropy. Each individual photon is perfectly ordered and still contains all its energy. By retroreflection, you can even cause those disordered photons to gather back in your system into perfect order once again and then you can scatter them again to do some more work. You can use the same energy over and over again! The perfect retroreflection is more easily accomplished by phase conjugate reflection. For a vivid example of just such retroreflection of previously scattered EM energy to restore the former order, see David M. Pepper, "Applications of Optical Phase Conjugation," Scientific American, Vol. 254, No. 1, Jan. 1986, p. 74-83. See particularly the striking photographic demonstration of time reversal of disorder on p. 75.

[29] William Jay Fogal, "High gain, low distortion, faster switching transistor." U.S. Patent No. 5,196,809, Mar. 23, 1993. Also William J. Fogal, "High Gain, Low Distortion, Faster Switching Transistor." U.S. Patent No. 5,430,413, July 4, 1995.

Table 4. Fogal's Semiconductor.

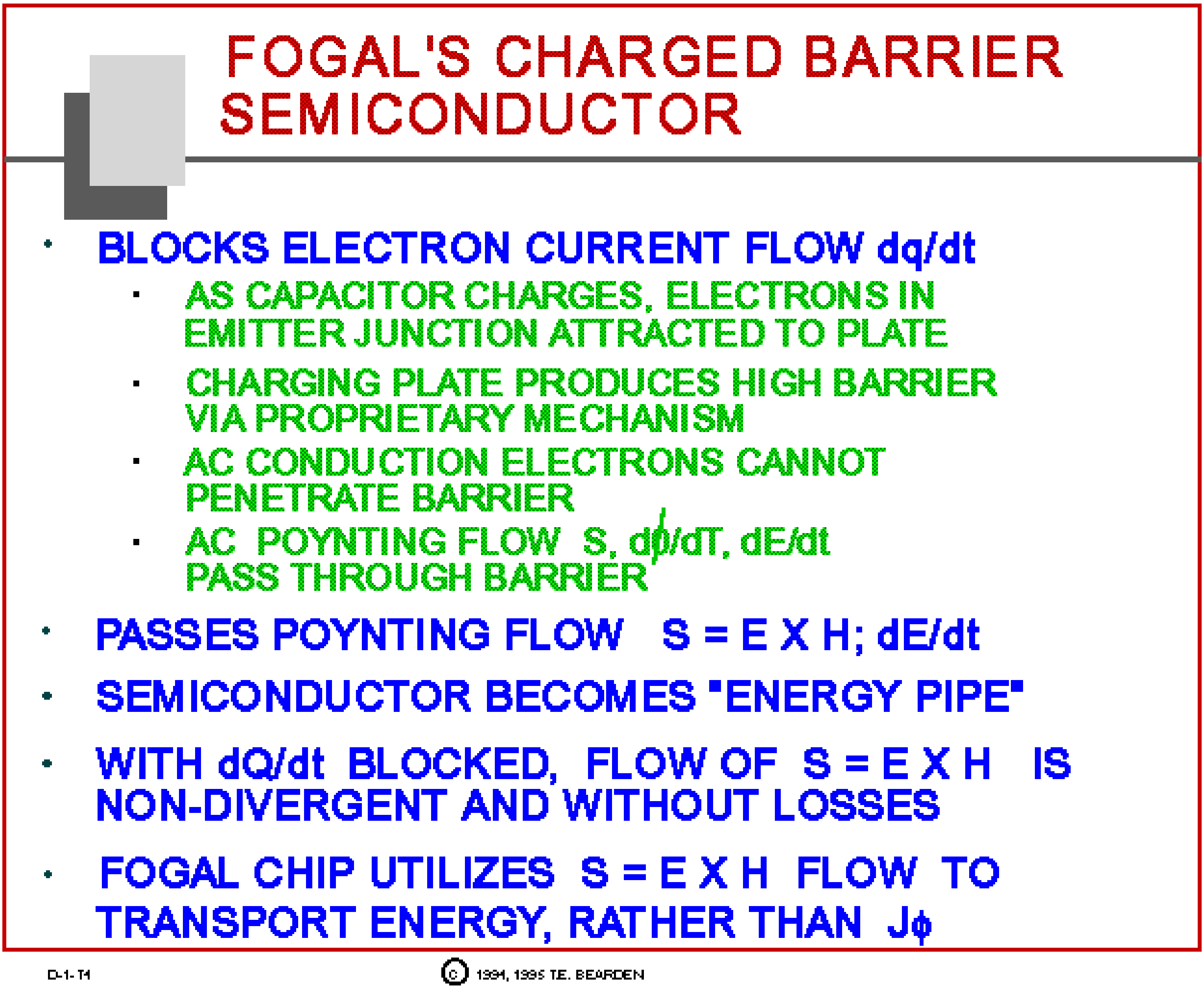

FOGAL'S CHARGED BARRIER SEMICONDUCTOR

- BLOCKS ELECTRON CURRENT FLOW dq/dt
  - AS CAPACITOR CHARGES, ELECTRONS IN EMITTER JUNCTION ATTRACTED TO PLATE
  - CHARGING PLATE PRODUCES HIGH BARRIER VIA PROPRIETARY MECHANISM
  - AC CONDUCTION ELECTRONS CANNOT PENETRATE BARRIER
  - AC POYNTING FLOW S, $d\phi/dT$, dE/dt PASS THROUGH BARRIER
- PASSES POYNTING FLOW S = E X H; dE/dt
- SEMICONDUCTOR BECOMES "ENERGY PIPE"
- WITH dQ/dt BLOCKED, FLOW OF S = E X H IS NON-DIVERGENT AND WITHOUT LOSSES
- FOGAL CHIP UTILIZES S = E X H FLOW TO TRANSPORT ENERGY, RATHER THAN $J\phi$

D-1-T4

© 1994, 1995 T.E. BEARDEN

[30] In a correspondence with Bockris lasting several years, he had graciously furnished me with material on the overpotential, which is how I found out about this esoteric area. Bockris is well-known as a leading scientist on the chemistry of hydrogen, and also as one of the leading scientists in the rapidly developing field of cold fusion. He is a marvelous scientist, and one day his largely unsung work with the overpotential will revolutionize the entire electronics industry.

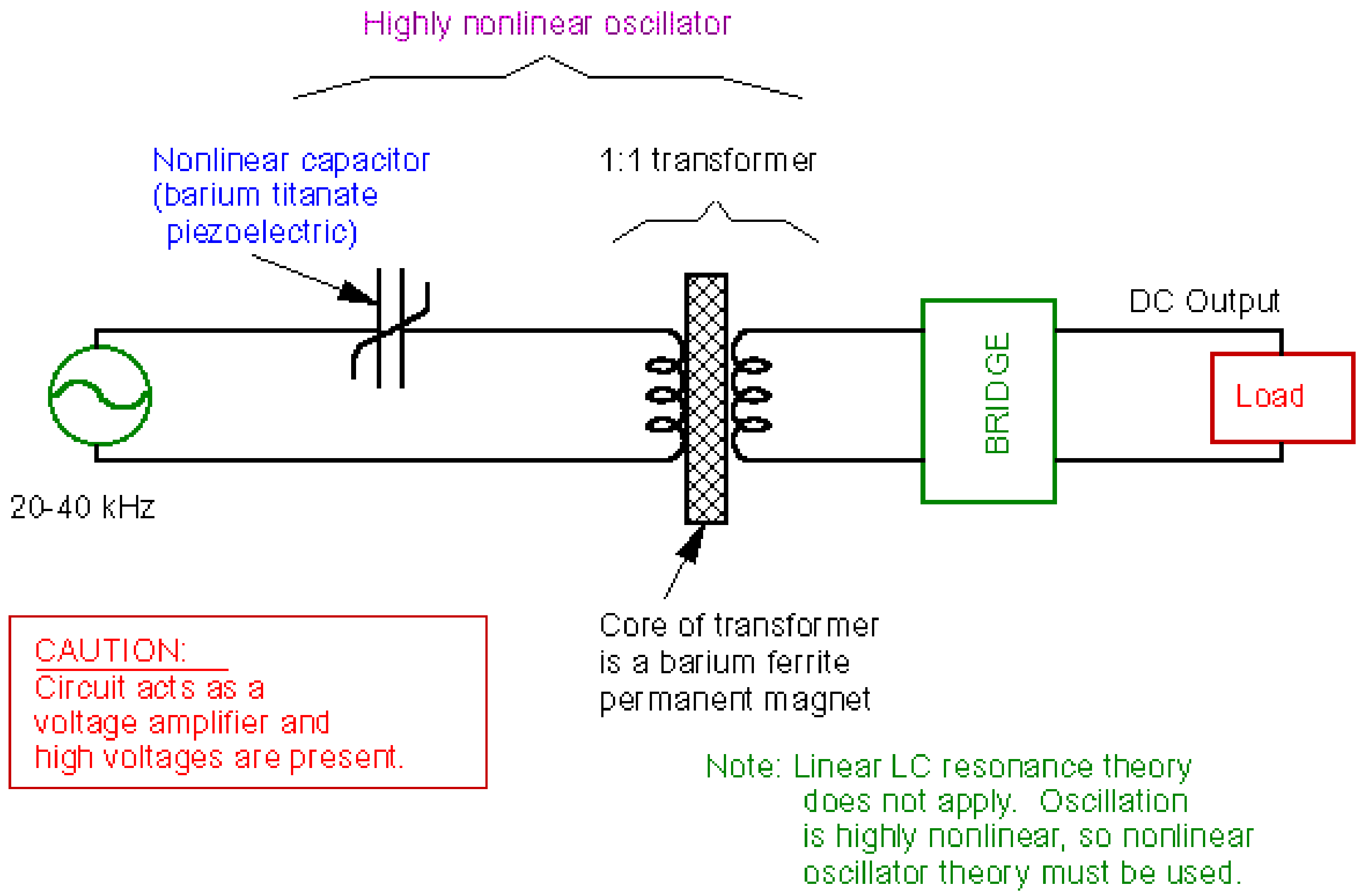
Highly nonlinear oscillator
Nonlinear capacitor (barium titanate piezoelectric)
1:1 transformer
DC Output
BRIDGE
Load
20-40 kHz
Core of transformer is a barium ferrite permanent magnet

CAUTION:
Circuit acts as a voltage amplifier and high voltages are present.

Note: Linear LC resonance theory does not apply. Oscillation is highly nonlinear, so nonlinear oscillator theory must be used.

Figure 22. Schema for the magnetic resonance amplifier (MRA).

D-1-F22

© TE. BEARDEN 1994, 1995

Howard Johnson is also a respected colleague, whom I very much admire. (See Figure 23). Howard has continued to work quietly and patiently upon his patented permanent magnet motor, [note 31] including patenting various magnetic gates, etc. that are necessary to make such a motor work in a rotary configuration. [note 32] Howard employs a two-particle theory of magnetism; i.e., each magnetic flux line is envisioned as having a particle traveling from the north pole to the south pole, and also a particle traveling from the south pole to the north pole. The particles are spinning; the forward-time particle spins in one direction, and the antiparticle spins in the other direction. Howard then slightly separates the two particle flows. [note 33] In other words, Johnson splits the flux lines themselves, into two different pieces. When so separated, the component lines are now curls, so their paths curve. The paths of the two "curl particles" are different; one curls in one direction and the other curls in the opposite direction. Further, a predominance of one form of curl particle gives a "time-forward" aspect, while a predominance of the other form of curl particle gives a "time-reversed" aspect. Johnson is thus able to employ a deeper kind of magnetism than the textbooks presently contain. He demonstrates that a "spin-altered" magnetic assembly exhibiting (to a compass or other such detector) a north polarity can attract another unaltered magnetic assembly exhibiting a north polarity. In short, he can make a north pole attract a north pole. We will give you further insight into Johnson's two-particle theory in a future article. [note 34] We will also explain how and why the physicists missed that antiparticle in the magnetic field's flux lines, and thereby failed to advance the theory of magnetism to a deeper level. Make no mistake, one day when the new theory is done, Johnson may well be awarded a Nobel Prize for his epochal discovery of a deeper structure of magnetism.

Bits and pieces of the new science approach are just now beginning to spill into a few conventional journals and symposia. For some time a few of the rebels spearheading this new science have been doing something that the University scientists should have been doing all along. The rebels have been meticulously examining the concepts, postulates, and definitions that the present scientific models are founded upon, to reveal serious foundations errors. Corrections for some of these errors have been discovered and made. [note 35] Others are still problems yet to be resolved. Yet slow progress is being made, and a "flip-over" of the old science is likely to occur within the next decade if not sooner. By then the new concepts will have become "solid" enough to allow producing an engineering mathematical model. Also the supporting experiments will have become solid enough to justify the new concepts and the model.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

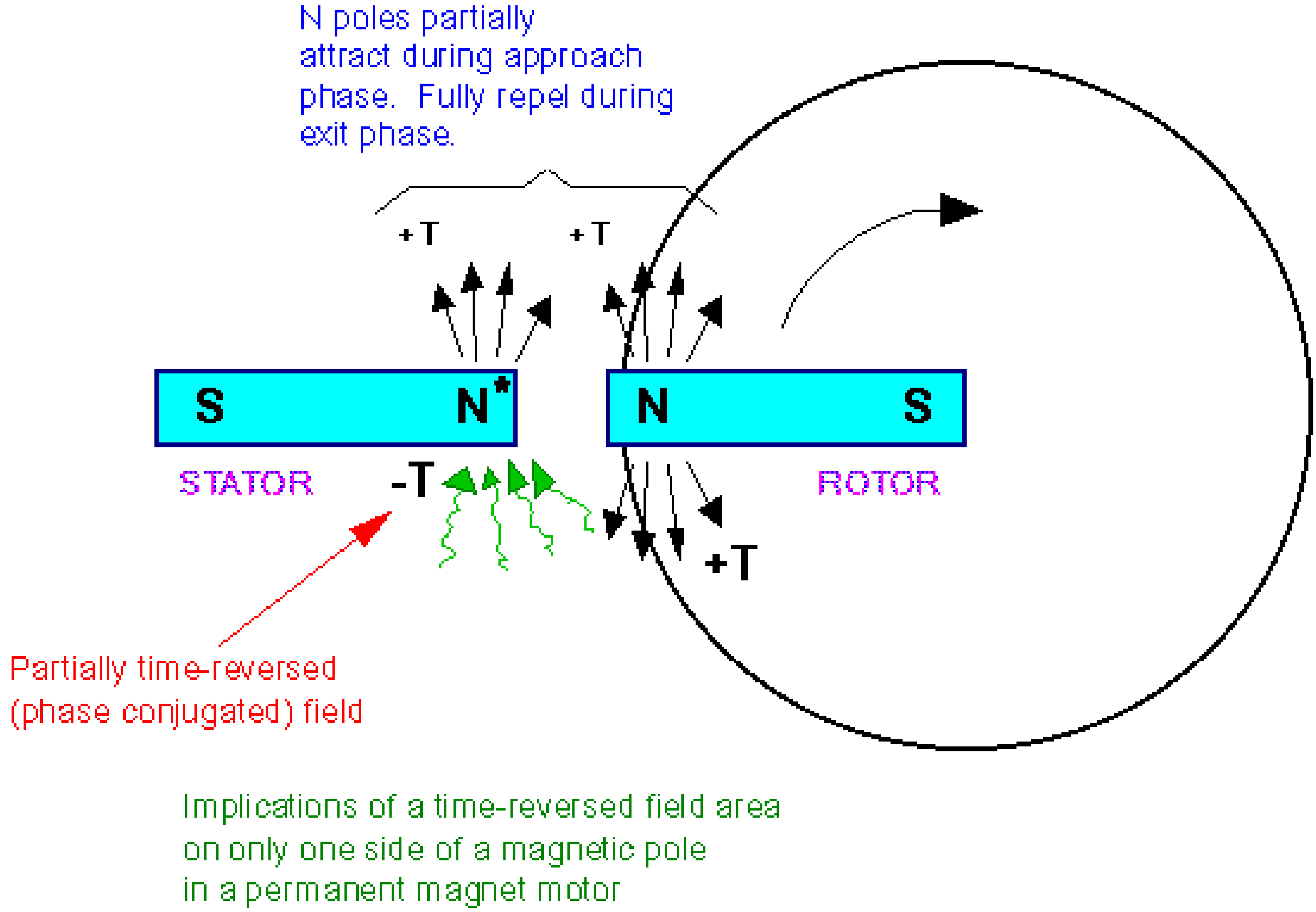
N poles partially attract during approach phase. Fully repel during exit phase.
+T
+T
S
N*
N
S
STATOR
-T
ROTOR
+T
Partially time-reversed (phase conjugated) field
Implications of a time-reversed field area on only one side of a magnetic pole in a permanent magnet motor

Figure 23. Schema for Johnson's permanent magnet motor.

T91H-J-03

© T.E. BEARDEN 1991, 1993, 1995

[31] Howard Johnson, "Permanent Magnet Motor." U.S. Patent No. 4,151,431. Apr. 24, 1979.

[[32] I personally saw and closely examined one demonstration rotary Johnson permanent magnet motor some years ago, and toyed with it for about one hour. It would definitely self- rotate as long as you wished to permit it to turn. It was not a powerful device at all, but just a small laboratory "proof of principle" prototype. It had cost Johnson an enormous amount of time, labor, and optimization to get the critical adjustment required of his two magnet assemblies. But the device had no power source other than the permanent magnet assemblies themselves. Johnson's nonlinear rotor and stator magnets interacted with each other in a manner to break local symmetry. So his machine was an open system and therefore a permissible overunity device; it was not a perpetuum mobile.

[33] As I have pointed out repeatedly in the past, photons also carry time, not just energy. We have previously shown the process and the photon interaction mechanism that creates the flow of time itself; we will discuss this mechanism in the future. So when Johnson separates the particles and antiparticles, not only does he partially separate them according to spin, but he also alters the local character of time flow during which the resulting magnetic field forces must appear. In other words, he accomplishes a partial separation of time-forward and time- reversed polar interactions. A south pole is just a time- reversed magnetic north pole, in the first place! So a north pole of a bar magnet that is slightly time reversed on one side will partially act on that side just like a south pole. On the other side it will continue to act like a normal north pole. By partially time-reversing (phase conjugating) one side of the north magnetic pole piece, Johnson makes that side look and act like a south pole. that way Howard is able to create two north poles, one on a stator and the other on a rotor, and time-reverse part of one face of the stator's north magnetic pole-piece. Therefore when the proper sides of the stator and rotor north poles are facing, they attract each other, contrary to the conventional textbook. The two poles then repel each other normally as soon as the north rotor poles passes the north stator pole. Hence Johnson can make a surrounding north pole stator assembly "draw in" an approaching north pole rotor assembly, and then kick it on out the other side, because he has broken the local magnetic symmetry. In short, Johnson's magnetic gate can provide a legitimate component of unidirectional magnetic thrust, which means that he can indeed make a rotary permanent motor. Simply put, this "partially separating the spin particles," and thereby partially phase conjugating one face of a magnet, is what Johnson calls a "gate," and this is the patented secret by which his magnet assemblies can be made self-powering. The entire process is still very meticulous, and assembly and adjustments are extremely critical. With Johnson's blessings we hope to shed more light on this subject in coming articles.

[34] For a basic article on rotary permanent magnet motors, see T.E. Bearden, "On Rotary Permanent Magnet Motors and 'Free' Energy," Raum&Zeit, 1(3), Aug.-Sep. 1989, p. 43-53.

[35] The most blatant error in all of physics is the erroneous notion that an acting force is separate from its receiving mass. Force is defined as F ~ d/dt(mv). From that definition alone, it can be seen that the affected mass is a component of the force itself. If you remove the mass m, you do not have force but merely acceleration. The notion of force separated from the mass came from ancient and familiar experiences of pushing on stones with one's hand, etc. In this primitive case the force-creating agent (i.e., the hand) is certainly separate from the affected mass (i.e., the stone). But the "force on the stone" is actually the force created of and with the stone. It is "stone-force," not "force upon the stone." Quantum field theory already tells us that it is the absorption and emission of virtual photons by the stone's mass that "creates the force." In fact it should go one step further: It is the coupled ensemble of the virtual photon flux and the mass of the stone that constitutes a "stone-force." In your mind, you should try mentally replacing the word "force" with the phrase "mass-force" until you root out the deeply ingrained erroneous notion that force and mass are separate. Because of this archaic error, today's physics can only define a mass in terms of its resistance to a disturbing force, and it can only define a force in terms of its overcoming a resisting mass. This error has prevented any proper definition of either force or mass! But the problem is solvable. For a proper definition of mass without the notion of force, see T.E. Bearden, Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena, National Technical Information System, Report AD-763210, 1973.

In this first article I wanted to introduce the 14 major areas. We will be discussing these areas in some technical detail in the future. We'll also include current major news items of development in these areas, for the interested lay reader. I'll particularly try to give you the benefit of extensive reference citations painfully gathered over the last 30 years. The articles will be targeted toward both the serious researcher and the interested, educated layman.

So that's the agenda for the months ahead. We hope you find this information and the approach of direct interest and use.

Again, it's a privilege to be aboard. And we invite you to hang on in the future; it's going to be a wild but exciting ride. We really are going to chase this wild dragon with great vigor.

ILLUSTRATION LIST

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

Novel Principles In The Rife Microscope

And Rife's Great Secret [1]

© 1998 T.E. Bearden

Substantially updated Aug. 20, 2001

Foreword

In the 1930s and 1940s, Royal Raymond Rife {1} made revolutionary discoveries in microscopy, including development of novel microscopes that could see far beyond the accustomed one-tenth wavelength limitation according to Rayleigh theory. With his microscope, he was able to "see" ultramicroscopic living forms much smaller than anything considered in the observable state. It would appear that his microscope was somehow capable of peering directly into the virtual state itself, and finding living forms at that level that were far beyond anything ever suspected by conventional science, even to this day.

Obviously this work was dramatically controversial. Rife went on to develop treatment methods and devices based on the new level of living organisms that his microscope revealed. In short, he arrived at essentially a new theory of disease, and did achieve success with his—to conventional scientists—completely unprecedented devices and methods of treating and healing. While we do not discuss his treatment methods in this paper, we do discuss the methods of Prioré, whom we believe to have used similar principles. So these discussions may shed some light on Rife's therapeutic mechanisms also.

Rife and others reported his results in the proper scientific literature. For his efforts, he was ostracized, essentially imprisoned in a medical treatment facility, broken, and rejected by his peers. His findings were harshly ridiculed and discredited. Finally escaping from his enforced confinement, Rife lived out his remaining years and died quietly and unknown.

In the present paper it is not our purpose to address Rife's theory of disease or his treatment devices and methods. It is solely our purpose to propose a higher symmetry electrodynamics functioning of his microscope, as one possible explanation of the startling microscopy results he achieved. We will reveal what we are convinced is Rife's great secret: recursive virtual state magnification (amplification) by iterative difference frequency serial summation of vacuum engines. This is a substantial change from an earlier mechanism proposed by the present author, which proved to be in error {1}.

We accent that this is not a "standard electrodynamics" examination of the Rife microscope; normal U(1) electrodynamics and transverse EM waves contain no mechanisms that can explain the functioning of the instrument. Accordingly, since no such explanation is contained in the usual electrodynamics, one must of necessity turn to those available electrodynamics models that are of much higher symmetry than afforded by vector and tensor algebra. To show the fit to experiments, we must also address an enormous amount of the work performed by Prioré and his colleagues. Only in such EM models and their known fit to actual experiments are the additional phenomena and principles available that can possibly explain Rife's epochal results.

Also, because of the importance of this unified field theory area, we include substantial references and additional extended discussions in the endnotes for further technical clarification. Thirty figures are added to illustrate the material.

The Rife Microscope

The Rife microscope { [1] } (Figure 1) cannot be made to work by normal optical principles at the level utilized by Rife. With normal optics, about one-tenth wavelength is as fine an optical resolution as one can get. Since Rife went far beyond that, it should be obvious that he was using optics and electrodynamics of a *novel kind*, and of a type not yet in the conventional textbook.

We believe we are aware of the major principles used in Rife’s microscope, or at least some of them, but prior to 1998 we shied away from anything to do with the scope. My major interests have been in extracting EM energy from the vacuum and in developing a new kind of medical therapy utilizing higher symmetry electrodynamics—and thus unified field theory—to directly and slowly change cellular disease or disorder back to a normal previous state. Obviously, if one is to produce a microscope that "sees" optically further into the realm of smallness than any other microscope can conventionally do (even in theory), that microscope must be functioning in a higher electrodynamics mode rather than in the conventional one.

In this paper, we will try to summarize some deep physics principles underlying the Rife microscope, at least as we see them terms of a higher symmetry EM model, and we cite some references for the bits and pieces. A more detailed explanation must wait for a future time and assistance from very high-level theorists.

Introduction

A good hint is this: Rife (Figure 2) often painfully adjusted his microscope stages for up to *24 hours* before he finally got the instrument "focused." Immediately that shows it is not standard optical equipment. Well-made standard optics—and Rife’s microscope was beautifully made—can be adjusted far more easily than that (in minutes!).

So far as we are aware, no one ever looked at the Rife work except through essentially rather orthodox electrodynamics and rather orthodox optics eyes. Normal optical theory permits resolution of about one tenth of the wavelength of the light used, and a few microscopes in the last few years have gone just a wee bit further. But none have gone anywhere to the degree of smallness that Rife reported.

In our view, the orthodox approach to explaining the operation of Rife's microscope is doomed to fail from the beginning. Even a very good scientist will fail with replication, if that approach is exclusively taken. An entire scientific team and substantial funding will not do any better, unless something technically deeper is understood and tackled from the outset. We give an overview of our approach to tackling that problem. To begin with, we must present some errors in electrodynamics and an omission in general relativity, so that we can develop the necessary background. The discussion is not simple and will require considerable development. Once the considerable background is developed, the application to the Rife microscope will be straightforward.

Errors in Electrodynamics and Adjustments to General Relativity

First, present electrodynamics itself has serious flaws and shortcomings, and needs substantial revision {[2] }. As foundations scientist Bunge points out {[3]}:

> *"...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state."*

Bunge {[4] } also said:

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

The flaws in electrodynamics pass over into what has been constructed for optical theory—both linear and nonlinear—which just assumes the electrodynamics as a subset.

For our purpose, we need to clarify what modern electrodynamics does and does not say about the massless vacuum, and about the state and form of EM entities—such as potentials, fields, and waves—created in it (in *empty space*) by matter dynamics, or acting from it (from *empty space*) onto matter and causing matter dynamics.

Maxwell and all his peers believed firmly in the material ether, a thin material fluid filling all space and every crevice. To scientists at the time, not one single point in the universe was devoid of this material etheric fluid—this *thin matter*. The atom had not been discovered, the electron had not been discovered, and electricity was also considered a thin material fluid. Charge had no more meaning than “a piece of the electric fluid,” much like a cubic centimeter of water.

Faraday {[5] } believed that the EM influence in space consisted of *material lines of force which acted as taut strings under tension from some mysterious source*. He believed that EM disturbances were just the transverse wave “perturbations” of these taut strings. In his concept, Faraday omitted the string holder or other agent producing the tensile forces upon the string (Figure 3). He thus omitted Newton’s third law from his electrodynamics concept. He also omitted precisely half the energy, by omitting the equal-energy time-reversed EM wave that accompanies the formation of every ordinary EM wave we take account of. In a receiving copper wire antenna, for example, not only do the electrons interact with the incoming fields, but so do the protons in the nuclei—though highly damped in amplitude by the proton's large mass with respect to the mass of a disturbed Drude electron. However, the two energetic interactions (of perturbed protons and perturbed electrons) are of equal *energy* magnitude. So in the conductor, equal energy is expended to move the "lattice holes" (positrons) as is used to move the Drude electrons, else field theory itself is falsified. Except in semiconductors, electrical engineers mostly just ignore the "disturbed lattice hole current" energy.

Maxwell deliberately captured Faraday's thinking, mathematically {[6] }, and so just *assumed* Faraday's lines of force and the "plucked tight string" transverse EM wave in the material ether {[7] } as well as the unexplained tension . The notion of the transverse EM wave in vacuum has essentially no other theoretical justification than that Faraday-Maxwell assumption and the observed "transverse waves" generated in the receiving "electric fluid in the conductor" in and on the Drude electrons. The transverse EM wave notion in vacuum was maintained only by discarding its missing antiwave that is always formed simultaneously in our transmitters and receivers (e.g., when we perturb the Drude electrons in a transmitting wire antenna, we also perturb the protons in the nuclei with the same energy. The two disturbances—proton disturbance in the nucleus and Drude electron disturbance in the Drude electron gas—perturb the local surrounding vacuum/ spacetime with equal energy, so in the vacuum "virtual particle flux gas" the two perturbations are energetically equal. The receiving wire receives both perturbation waves from the vacuum; one perturbs the electrons and the other perturbs the nuclei. Since the second wave is unaccounted, the perturbation of the nuclei is just considered a "mysterious force" that always appears. In a pumped phase conjugate mirror material, e.g., we interact the waves prior to their reaching the nuclei, by multi-wave mixing. Since the "causative" but neglected second wave coming in from space does not reach the nuclei, but are redirected to form the phase conjugate replica wave produced by the pumping, *the mirror does not recoil and it does not exhibit Newtonian third law reaction*. However, since electrodynamics *omits* the EM causative wave that generates Newton's third law forces in the receiver, which in normal detection always appear, then presently electrodynamic fields are erroneously said to be devoid of Newton's third law reaction. To the contrary, the effects of that hidden, neglected EM cause of the Newtonian third law reaction do appear, either as recoil of the physical receiving antenna or as the production of the time-reverse phase conjugate replica wave in optical pumping. But in standard electrodynamics, equal-and-opposite forces in the receiving antenna are assumed to arise "mystically" or from some hidden *electrodynamic demon*. The electrodynamic cause of Newton's third law has been lost since Maxwell wrote his theory.

Maxwell, however, was careful to point out the ad hoc nature of assuming the stress in the vacuum, and that no mechanism for it had been discovered, nor was such a mechanism contained in his theory. Specifically, Maxwell said {[8] }:

> *"It must be carefully borne in mind that we have made only one step in the theory of the action of the medium. We have supposed it to be in a state of stress, but we have not in any way accounted for this stress, or explained how it is maintained. This step, however, seems to me to be an important one, as it explains, by the action of the consecutive parts of the medium, phenomena which were formerly supposed to be explicable only by direct action at a distance. … I have not been able to make the next step, namely, to account by mechanical considerations for these stresses in the dielectric. I therefore leave the theory at this point, merely stating what are the other parts of the phenomenon of induction in dielectrics."*

Indeed, we now know of electrons and their spin. The longitudinally-constrained, spinning Drude electrons in receiving wires will gyroprecess laterally in the wire when an EM disturbance from the vacuum interacts with them. (Figure 4) It is well-known that these Drude electrons are highly restrained longitudinally down the wire; the longitudinal movement is the "drift" velocity and may be nominally only a few inches per hour. The lateral precession of the longitudinally constrained, spinning Drude electrons establishes that the incoming vacuum disturbance, before its interaction with the gyroelectrons, is *longitudinal*. Else one must discard electron spin, electron gyroprecession, and indeed the theory of gyroscopic precession itself. In our view, the EM wave *in vacuum* is longitudinal, always has been, and always will be {[9] }. It also *does not* consist of EM force fields, contrary to present classical theory. Instead, only the potentiality {[10]} for the transverse force fields—*should* the longitudinally perturbed vacuum EM entity be intercepted by a spinning charged particle and interact with it—exists in mass-free space, as pointed out by Feynman. Quoting Feynman again {[11] }:

> . *"We may think of* ***E****(x, y, z, t) and* ***B****(x, y, z, t) as giving the forces that would be experienced at the time t by a charge located at (x, y, z), with the condition that placing the charge there did not disturb the positions or motion of all the other charges responsible for the fields."*

Rigorously the force fields exist only in and of matter. In matter-free space, there is no force at all in what we call the "field as it exists in massfree space" {[12] }. E. g., Jackson states it this way {[13]}:

"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."

But again, Jackson also states {[14] }:

> *"...the thing that eventually gets measured is a force..." "At the moment, the electric field can be defined as the force per unit charge acting at a given point. It is a vector function of position, denoted by E. One must be careful in its definition, however. It is not necessarily the force that one would observe by placing one unit of charge on a pith ball and placing it in position. The reason is that one unit of charge may be so large that its presence alters appreciably the field configuration of the array. Consequently one must use a limiting process whereby the ratio of the force on the small test*

> *body to the charge on it is measured for smaller and smaller amounts of charge. Experimentally, this ratio and the direction of the force will become constant as the amount of test charge is made smaller and smaller. These limiting values of magnitude and direction define the magnitude and direction of the electric field E at the point in question. In symbols we may write F = qE where F is the force, E the electric field, and q the charge. In this equation it is assumed that the charge q is located at a point, and the force and the electric field are evaluated at that point."*

As can be seen, much of the difficulty occurs because physicists continue to erroneously utilize an *equation* as if it were a *definition*. An equation defines nothing at all; it merely states that all the "things on the left" of the equality sign and all the "things on the right" have the same overall magnitudes. It does not *define* any of the things on the left or right, and so it tells us nothing at all about what anything in it *really is*. A definition requires an identity, not an equality. E.g., if we take as an identity the very equation advanced by Jackson's explanation above, we then have a definition:

$$F \equiv qE \Rightarrow E \equiv F/q$$

And now that identity clearly states that charged mass is a *component* of the **E**-field. Similar argument prevails for the **B**-field, etc.

The identity also shows that E is not the *magnitude* of the force field at all, but is its *local intensity* at any given point in the E field, in terms of force per unit point static coulomb—a force formed in, on, and of the charged mass of the coulomb. The misstatement that E represents the *magnitude of the field* is another great non sequitur. No textbook gives the calculation of the *magnitude* of the E-field, though all purport to. Instead, they all give the calculation of the *intensity* of the E-field at each point in it. There is a great difference between the magnitude of an automobile's mounted tire and its pressure per sq. inch on the pavement. In all fairness, some of the better texts do use the term "field intensity", at least occasionally.

As *defined*, the EM force fields are effects produced in, on, and of charged matter and *only* in, on, and of charged matter. They are not at all the same effects produced in matter-free space (in the vacuum/ spacetime). The EM force field (in matter) is not the same as the "massfree" E-field in massfree spacetime. Ultimately, all that exists in spacetime are *spacetime* and *changes to spacetime*.

We may accurately take the mass-free, force-free EM field to be a pattern of spacetime curvatures. The coupling of this spacetime curvature set, or any part of it, with charged mass, then constitutes an EM force field. Again, the problem is the hoary old mechanics notion of having a separate massless force acting upon a separate mass. No such thing exists in all nature, and mechanics should have been rather thoroughly overhauled long ago as a result of this horrendous *foundations of mechanics* error that is centuries old.

In present classical electrodynamics—certainly the kind taught to electrical engineers and used in the electrical engineering discipline—the EM potential {[15] }, field, and wave, as they truly exist in space and in the vacuum, are rigorously not prescribed at all. Instead, the effect of the "causal entity" as it exists in vacuum prior to interaction with charged mass, is used a if it were what exists in and on the charged mass as the effect of the interaction. (See Figure 5, Figure 6, and Figure 7). This substitution of the effect for the cause is arguably the greatest uncorrected error in classical electrodynamics today. It is a hangover from the old days of the material ether, and the refusal to change the Maxwell-Heaviside equations to eliminate that material ether assumption, once the Michelson-Morley experiments falsified it. This has long been known to foundations physicists, and to most of the better electrodynamicists as well. But it still seems to be known to very few electrical engineers!

Many attempts have been made to “modify” the standard classical electrodynamics theory, but most have failed. For some reason, most of these attempted modifications seem to have accepted the present "matter-to-matter transform" approach (Figure 8) of present electrodynamics, failing to see that the two missing transforms of matter-to-spacetime and spacetime-to-matter (Figure 9) are in fact just general relativity itself {[16] , [17] } The solution for adding the missing two transforms into electrodynamics therefore may lie in "infolding" general relativity directly within electrodynamics, rather than placing the two on an equal footing. That requires the analysis of the supersystem rather than the system, a concept which will be discussed shortly. Figure 10 shows the utilization of the "vacuum engine" that results from this "infolded GR" solution.

To our knowledge, the Sachs unified field theory {[18] a,b,c,d,e} is presently the only practical unified field theory that can be directly engineered by higher symmetry electrodynamics, such as the O(3) group symmetry EM model advanced by Evans and Vigier and perfected by Evans {[19] }. The engineering power of Sachs' theory is remarkable {[20] }, and the O (3) electrodynamics has been shown by Evans to be an important subset of the unified electrodynamics in Sachs' theory.

Let us examine that statement a little deeper.

In classical electrodynamics the sources of all EM potentials, fields, waves, and their energies are assumed to be the charges—which are assumed to somehow act as perpetual motion machines that continuously *create and pour out* all that EM energy in the fields and potentials associated with the charge as their "source". This outpouring of 3-space energy comprising and establishing the associated potentials and fields eventually fills all space with energy. All the charges and dipoles present in the universe since shortly after its formation have done this for some 14 billion years.

Yet in classical EM—with all the existing charges in the universe implicitly assumed to continuously pour out energy into all of space, and most having done so for billions of years—inexplicably space is *then* considered to be an emptiness having no interactive energetic exchange with the charge at all {[21] }. That of course is a non sequitur because it involves a total contradiction of the conservation of energy law. If there is no active energy in spacetime for the charged particle to receive and then output in 3-space, then the energy conservation is falsified by every charge and every dipole.

In particle physics the inactive vacuum concept is known to be quite wrong, else the charge creates energy from nothing. That would violate the most sacrosanct law in physics: that *energy can neither be created nor destroyed*. So electrodynamics—both classical and quantum, neither of which contains the solution to this *source charge problem*—is in gross violation of conservation of energy laws, because it assumes that the charge is solely the source of its associated fields and potentials and their energy { [22] } and that it accordingly creates all that energy it has poured out since the beginning of the universe.

Presently the electrical engineer implicitly assumes total violation of the conservation of energy law with every equation he writes, and is blissfully unaware of it.

Particle physicists, however, have done much better and published and incorporated their results, although inexplicably the mainstream electrodynamicists have not changed their own model accordingly. Particle physicists have known and shown for more than 40 years that (i) space is a seething inferno of EM energy { [23]}, (ii) a charge continually undergoes a violent energy exchange (see Figure 11) with the vacuum EM flux. That exchange is what a charged mass is doing, to *make* it a "charged mass"), and (iii) the charge is a *broken symmetry* in that violent energetic flux of the vacuum. "Broken symmetry" means that the charge disrupts and alters part of that virtual particle flux to observable EM energy flux (see again Figures 5, 6. and 7). The charge continually absorbs virtual energy from the vacuum, radiates *some* of that absorbed energy back to the vacuum in virtual form, but *integrates and radiates the remainder* in a flow of potentially observable energy—the energy flow discovered by

Poynting {[24] } and Heaviside {[25] } after Maxwell was already deceased.

So what does the source charge really do? It coheres, transduces, and *gates* some of its virtual energy {[26] } received from the vacuum, into an EM energy flow going out in all directions in 3-space from that source charge. Contrary to present electrodynamics, this energy flow is actually longitudinal and bidirectional {[27] , [28] , [29] , [30]}, with energy *going out in 3-space from the charge* to every point in the universe, and with energy coming from every point in the universe *back to the charge through the time domain*. But as reinterpreted {22 b}, Whittaker {[31] } showed in 1903 that any such bidirectional EM flow of energy —reinterpreted with the input energy to the charge coming from the time-domain, and the output energy from the charge being emitted into 3-space, comprises a scalar potential. This also agrees with quantum field theory {[32] }

See Figure 12. As it really exists in space, a scalar potential is a harmonic set of bidirectional EM energy flows, with the "bidirectional" EM waves being outgoing EM waves in 3-space caused by interaction with the source charge or dipole of incoming time-polarized EM waves. There is a set of outgoing EM longitudinal waves in 3-space, perfectly correlated with a set of incoming time-polarized EM waves in the time domain. Hence one time-polarized EM "causal" wave couples with one outgoing longitudinal EM wave in 3-space. It is this scalar EM wave coupling (combining) with longitudinal EM wave coupling, in a harmonic set of phase conjugate pairs, that constitutes the scalar EM potential.

In short, all EM energy in 3-space comes from the time domain.

After all, no observable persists continuously in time. Any observable is a frozen 3-space momentary snapshot of an ongoing 4-space interaction. It is the result of a d/dt operation imposed upon LLLT spacetime, to yield a momentary LLL by d/dt(LLLT) => LLL

What we see as the "motion of an object in 3-space" is actually the ensemble of a continual serial iteration of such "frozen 3-space snapshots" of each observable part of the object, much as the iterative frames of a movie film are seen as a continuously moving picture.

In Whittaker's bidirectional energy flow associated with the source charge, electrodynamicists—including Whittaker himself—assumed that a point positive coulomb of charge is located at each and every point in space. (Figure 5, Figure 6, and Figure 7). A tiny bit of the impinging EM energy flow “from” the distant source charge {[33] } is geometrically intercepted and diverted by the observing/detecting point charge to streamline around it, creating a “swirl around” of diverted energy flow. As can be seen, only a tiny portion of the impinging and passing EM energy stream in 3-space is diverged into the swirl. This is rather like a small diverted whirlpool of water forming around a stable rock in a flowing river. In electrodynamics the amount of energy in that “diverged whirlpool” is said to be the “collected” energy appearing upon that point charge (upon that fixed rock). That amount of energy swirled by the intercepting point coulomb at a given point in space, is then erroneously said to be the *magnitude of the potential (the magnitude of the entire river)*. It is no such thing. The small fraction of the potential that is *diverted from* the potential's flowing rivers of energy, is not that the magnitude of that potential (the magnitude of the entire set of flowing rivers themselves!) And it is certainly not “identically” the river itself! At best it is a measure of the *intensity* of the potential (the local intensity of all those internal rivers of energy flow) at that point occupied by the intercepting/diverting rock.

The point is, *as defined and used in conventional electrodynamics*, the potential occurs only after the vacuum EM entity’s reaction with the intercepting mass, and is defined only as the *effect* produced upon that coulomb by the causative potential's interaction. Literally, the potential (the cause) is erroneously defined as the “excitation” or “excess energy” it locally produces as an *effect* upon a unit point intercepting/ collecting charge.

Ironically, it appears that no electrodynamicist calculates the magnitude of the actual potential as it exists in vacuum-spacetime, before its interaction with charge {[34], [35] }. Its magnitude is enormous, since the potential may extend over all space, and energy can be intercepted and collected by a charge placed at any point therein, or from charges placed at every point therein.

Certainly no rigorous definition of the potential presently exists in physics; the reader is challenged to check for himself or herself. No legitimate definition *can* exist for the potential, until the little matter is resolved of its present erroneous definition only "in and on and deviated by" charged matter, contradicted by the simultaneous assumption that it does exist in mass-free space in undeviated form. What is *deviated from* the potential, is not the *potential itself*. A part of something torn out of that something at one point in itself, is not the whole. The *local effect* cannot be substituted for the *entire nonlocal cause*—only a small part of which interacts *locally*.

The EM force field is similarly defined—only for charge mass interception, and only as an *effect* produced upon the intercepting charge. (See again Figures 5, 6, and 7.) A flow of water across a rock (or the wind on a rock in the desert) produces a “pileup” of fluid on the windward side, and a lesser “pileup” on the downwind side. The difference (called the “gradient”) in the energy density collected across the standard charge (rock) produces a “difference in pressure” on the rock—in other words, a force per interacting rock. That *effect* is erroneously taken to be the EM field itself, when rigorously it is only the reaction cross section of the EM field. Again, rigorously this EM field *as defined in electrodynamics* only exists as the effect that occurs upon an intercepting standard charge or "rock" {[36]}.

Since the EM wave “in space” is erroneously said to be made of varying EM fields, it too is defined only as the effect upon the intercepting charged mass, after the vacuum EM wave entity—whatever it is—has interacted with the charge and changed the charge translationally and energetically. Specifically, the present “EM wave in vacuum” in the literature is not at all what is presented and taught in undergraduate texts, and even the standard illustration of "oscillating E-field and B-field vectors" lateral to the line of motion of the wave is horribly wrong. E.g., Dr. Robert Romer, former editor of American Journal of Physics, has taken that error to task as follows { [37] }:

> *"…that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. …it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "…perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

The EM field and wave *as commonly regarded* are transfer functions from the source charged masses that “perturbed the vacuum medium and created an unknown kind of disturbance therein”, onto the receiving charged masses that intercept the incoming “unknown vacuum wave entity that was created” and interact with it, being translated and perturbed as a result.

This “detected” translation of the receiving charged masses (usually the interacting Drude electrons in a wire) is what our instruments detect. Indeed, the Drude electrons are highly constrained from moving longitudinally down the wire, usually moving at only a few inches per hour. They are much freer to move laterally, however, and can easily do so.

Reacting with the incoming longitudinal EM wave perturbation of local spacetime, the spin of the electron produces gyroscopic precession laterally in the wire since the electron is terribly restrained in moving longitudinally down the wire. Since our instruments detect these *transverse electron precession waves*, that has been totally confused (since Faraday's "plucked taut string" notion of the EM wave and Maxwell's arbitrary adoption of it) as representing the incoming wave in vacuum. It does not represent the vacuum EM wave at all; the incoming wave is in fact a longitudinal EM wave in 3-space, associated with a precisely correlated longitudinal EM wave in the time-domain.

So we have answered Professor Romer's implied question. That horrible diagram came from the assumption by early electrodynamicists, before discovery of the electron or atom or nucleus, that the "transverse electric fluid waves" in the receiving wires represented actual "intercepted" fluid perturbations arriving in the material ether medium.

In summary: Using a transmitter-receiver analysis, the so-called “EM wave in vacuum” *as presently defined*, actually represents that “detected and measured electron precession wave” in and of the reacting Drude electron gas in the distant receiving wire antenna. The so-called EM force field wave in vacuum has been erroneously defined as observable matter wiggles, not vacuum virtual energy wiggles and not as spacetime-curvature changes and wiggles. So even the present notion of the EM transverse wave in space is wrong, and it has been corrupted as a pure matter-to-matter transform from the beginning { [38] }.

Somewhat less technical explanations are given by Lindsay and Margenau { [39] }, and a rigorous statement and discussion that EM force fields exist only in charged matter is given by Aharonov and Bohm { [40] }. Feynman of course said it bluntly: Only the potential for the field exists in the vacuum, should one place a unit point charge there to allow the force field to be developed (on, of, and containing the source charge itself).

Let us summarize this terrible EM foundations flaw: As classical electrodynamics presently is modeled, Maxwell and his modern followers only include equations which are *mass-to-mass transformations*

(Figure 8). Specifically missing { [41] } are two other transforms: (i) the mass-to-vacuum (mass-to-spacetime) transformation (which obviously is general relativistic in nature), and (ii) the vacuum-to-mass (spacetime-to-mass) transformation (again, obviously general relativistic in nature). (See again Figure 9). When these two missing transforms are added back into electrodynamics, one has suddenly extended EM theory into a new domain. One has also *infolded* general relativity (GR) inside the present EM matter-to-matter transforms, and also included the vacuum interaction including broken symmetry in that exchange, but in an *electromagnetically engineerable manner*. In short, one has a practical, engineerable unified field theory.

Highly creative inventors such as Rife and Prioré intuitively developed apparatuses that unwittingly utilized this presently undeveloped and extended union of EM, particle physics vacuum theory, and GR. The intuition of the theorists in interpreting those revolutionary experiments and their results has very seriously lagged.

The Supersystem Concept

To understand how extraordinary systems such as Rife's microscope works, one must understand the concept of the *supersystem*. (See Figure 13). The supersystem consists of three components, which are (1) the physical system and its dynamics, (2) the local active vacuum and its dynamics, and (3) the local curvatures of spacetime and their dynamics. All three components of the supersystem interact with one another, in nature. The second and third components of the system constitute the active environment in which the system is embedded, and with which it interacts.

Unfortunately, in standard classical electrodynamics either the system's environment is assumed to be inert, or the system is assumed to be in equilibrium in its interaction with that environment. This represents a totally arbitrary assumption that the system cannot receive and utilize any excess energy from its known active environment. Hence the active vacuum's interaction with the system can be and is ignored by classical electromagnetics, as are the local curvatures of spacetime and their interactions with the system.

Extraordinary EM systems invariably involve the supersystem dynamics, and not just the system dynamics. *In nature*, no system analysis is complete until the supersystem interactions and their effects have been determined and analyzed.

The Rife microscope is indeed an extraordinary system using non-negligible interactions between the system and the other two components of its supersystem. Hence in our discussions, we will be examining interactions in that regime, and not just the standard electrodynamics analysis. In these discussions, we will be holding in mind not just the mass-to-mass transforms of standard electrodynamics, but two other transforms: (1) the mass-to-vacuum/spacetime transform, and (2) the vacuum/spacetime to matter transform (previously shown in Figure 9).

The reason that Rife's work has not been understood is that researchers have been unaware of the necessity for supersystem analysis. Almost all analyses of Rife's work have been merely standard classical EM analysis and thus quite insufficient. There are other examples of the standard electrodynamics model completely failing to show higher symmetry phenomenology in an inventor's work, so that his work is not understood even today. An example is provided by Tesla's patented circuits, as rigorously demonstrated by Barrett { [42] }. Barrett went on to improve the new functions discovered in Tesla's patents, and obtained two additional patents related to signals and signal processing { [43] }.

Vacuum-Structurings as “Engines”: Key Part of the Missing Transforms

The missing mass-to-vacuum/spacetime transformation produces a massless structuring of the vacuum medium. That is a *vacuum engine* or “multiplex of specific spacetime curvatures.” (See Figure 14, Figure

15, Figure 16, Figure 17), and Figure 18. Preparing the vacuum engine is accomplished by charged matter dynamics of the source charges producing the necessary potentials, fields and waves. The charges are embedded in continual energetic exchange with the vacuum/spacetime, and moving the charges affects that exchange; it perturbs the local vacuum flux and curves the local spacetime accordingly.

The transform represents producing this vacuum engine by the vacuum-perturbing action of source charge dynamics upon the vacuum, and by *not* assuming a mystical intercepting point-charge at each point in space { [44] }. In the vacuum/spacetime, *only* vacuum/ spacetime entities and changes to them exist. *In pure water, only water and changes to that water exist!*

So a precise dynamic structuring of the energetic vacuum (and spacetime geometry) is what charges produce when they "perturb" the vacuum medium { [45] }, by the presence or dynamics of their fields, longitudinal and time-polarized EM waves, or both. This "perturbed vacuum dynamics" separate from the source charges themselves comprises an engine that consists of altered longitudinal EM wave dynamics inside the internal Whittaker structure { [46] } of the scalar potential of the ambient vacuum. That is the same as also producing internested levels of spacetime curvatures in a *template* or specific pattern { [47] }.

In turn, this vacuum engine region of local curvatures and structurings in spacetime will interact upon any mass placed therein. That action is the missing *electrodynamic vacuum-to-mass transform.*

The missing electrodynamic vacuum-to-mass transformation produces direct changes and alterations in the "receiving" mass, at all levels, by the action of such a vacuum engine. As each action occurs locally in the localized receiving mass, it produces the "local effect" as the result of that "local action." The primary causative chain (i.e., the set of "dominoes" affecting each other serially), started back at the distant source charges. One can say that we therefore have mass-to-mass *action at a distance*, via the matter-to-vacuum and vacuum-to-mass intermediary transforms and the spatial and temporal propagation of the vacuum engines in the middle of the two {[48]}. We will shortly discuss the Whittaker structuring of the potential as a fundamental process for producing vacuum engines.

General relativists mostly concentrate upon the warping and curving of spacetime (i.e., producing a vacuum engine) by the very weak gravitational force—which is only some 10^{-42} times as strong for electrons as the electric force between them. Consequently, relativists must look to huge assemblages of mass, such as in stars and other astronomical objects, for sufficient gravitationally-induced spacetime curvature to produce easily observable results. For that reason, general relativity has resisted laboratory development as an engineerable applied science. By using the far stronger EM force as the agent of spacetime curvature, and including the infolded electrodynamics inside the potential, field, and wave, general relativity then becomes engineerable and usable. But one is engineering the supersystem, not just the system.

The Infolded Electrodynamics Inside Potentials, Fields, and Waves

As previously stated, in 1903 E. T. Whittaker, a well-known mathematical physicist, showed that a scalar potential can be mathematically decomposed into a set of peculiar EM wavepairs in a harmonic set. (Again see Figure 12) These "hidden waves" are longitudinal EM waves, arranged in conjugate pairs {[49]} with the pairs also arranged in a harmonic set. In each wavepair, there is an ordinary (forward-time) longitudinal EM wave (outgoing in 3-space from the interacting/observing charge), coupled to its phase conjugate replica (time-reversed twin).

And here lies a magic secret. Prior to interacting with a charge, the phase conjugate wave in each wavepair is in the complex domain, which means that in Minkowski space it is incoming along the fourth axis, *ict*, where the only variable is *t*. Hence this wave prior to interaction with charge does not even exist in 3-

space. Instead, it exists as a longitudinal EM wave incoming in time, so that it is a time-like flow of EM energy. (See Figure 19). After interaction with a charge, the charge absorbs the incoming time-polarized EM wave, transduces that time-energy to 3-spatial energy, and re-emits the energy as longitudinal EM waves in 3-space in all directions. "In all directions" means that for every outgoing longitudinal EM wave emitted in one direction in 3-space, another wave just like it also emitted in the other direction in 3-space. So the paired "bidirectional" EM longitudinal EM waves in 3-space, as interpreted by Whittaker, do exist as the *effect* waves caused by the incoming causal time-like longitudinal EM waves (scalar EM waves). The pairing of the causal time-polarized EM incoming wave with an outgoing effect longitudinal EM wave was not interpreted by Whittaker, but does agree with more modern quantum field theory interpretations of photon polarizations and their observability or non observability, by Mandl and Shaw {22b}.

We have corrected this "effect substituted for the cause" wave to reinterpret Whittaker's 1903 decomposition of the scalar potential, into combined biwaves where one of the biwaves is incoming in the time domain, and the other is outgoing in 3-space. That re-interpretation then allowed a solution {22a} to the problem of the source charge and its associated fields and potentials, together with their energy. Every source charge and source dipole continuously emits EM energy in 3-space, without any 3-spatial EM energy input. The input is there, but in the time-domain. This problem—resolved by the present author in 2000—has been called the most difficult problem in quantum and classical electromagnetics {30 }.

Because of observed parity reversal accompanying the time-reversal operation, in his forward-time, the observer “sees”—after interacting the incoming wave with charge—the *effect* of the incoming time-reversed or phase conjugate wave (entering from the time domain) as a 3-space longitudinal EM wave leaving in the opposite spatial direction. So the basic 4-space phase conjugate longitudinal wavepair would appear to us in forward time as a “regular” time-forward longitudinal EM wavepair in 3-space, with the two waves going outward in opposite directions. But each is accompanied by a time-polarized EM wave, else it would not be observable.

Further, there is a harmonic series of such bidirectional longitudinal wavepairs comprising the scalar potential. So the staid old electrostatic potential (i) *is not* a "scalar entity" at all { [50] }, (ii) *is* a bundle of longitudinal wavepairs, (iii) *is* an entity with a remarkable internal hidden multiwave structures and dynamics, (iv) *is* a composite of very special waves whose transverse amplitude (i.e., whose spatial energy density, which is a function of transverse amplitude squared) does not change, but whose transported time structure changes, and (v) therefore is a special kind of wave that is velocity modulated about some central velocity, and is not limited to the speed of light c. Indeed, in classical electrodynamics in the Coulomb gauge, the scalar potential is already recognized as having infinite velocity, merely appearing in space instantly at every point it occupies { [51] }.

We also state without elaboration that the active vacuum is *identically* a very huge scalar potential, as indeed is “spacetime geometry”. So a unifying principle is:

vacuum ≡ spacetime ≡ potential ≡ energy
{ [52] }

Further, as a potential, the “ambient” vacuum decomposes into an incredibly rich, internal Whittaker wave structure. So then also must *spacetime* and *energy* similarly decompose. That decomposition structure is even richer when higher symmetry O(3) EM is used instead of the lower symmetry U(1) EM employed by Whittaker.

A revolution in electrodynamics has recently begun, and central to that revolution is the strong appearance of longitudinal EM waves as essentially the basis for the emerging new electrodynamics. Work by Evans { [53] }, Barrett { [54] }, Ziolkowski { [55] }, Evans { [56] }, Rodrigues {57}, and others is fundamental in this respect. One part of this emerging new electrodynamics is advanced under the label of *undistorted progressive waves* (UPWs). A good summary of UPW theory is given by Rodrigues and Lu {[57]}. Even so, to date this body of work still largely ignores the internal Whittaker dynamics {[58]} inside an EM wave. The body of work—particularly the higher symmetry O(3) electrodynamics pioneered by Evans {56} inherently possesses the capability to model such structuring, since much of it utilizes a higher topology algebra such as quaternions {[59]} or even Clifford algebra {[60]} In those higher algebras, many electrodynamic things can be done outside the operations permitted by present vector and tensor electrodynamics.

Superpotentials and Superpotential Theory

Whittaker's 1904 paper {[61]} showed that all the "normal" electrodynamics was based on, *and created by*, the interference of scalar potential functions (i.e., by the interference of those multiwave longitudinal wavepairs and their dynamics {[62]} constituting the interfering scalar potentials.) He showed that any EM field or wave whatsoever can be replaced by two scalar potential functions—in fact, is *comprised of* those two scalar potential functions and is *created by them*. Scalar potential interferometry already creates all the "normal" EM in the textbooks anyway, and a proof has been given by Evans et al. {[63]}.

The 1904 paper by Whittaker initiated a somewhat obscure branch of electrodynamics referred to as *superpotential* theory {[64]}. Whittaker's work was further extended and augmented by later scientists such as Righi {[65]}, Debye {[66]}, Bromwich {[67]}, Nisbet {[68]}, McCrea {[69]}, etc. But even with the extended superpotential theory, without the addition of Whittaker 1903 and the substructuring "internal LW electrodynamics" one cannot understand the mechanism for what Rife was actually doing.

Rife was not using normal potentials and normal **E** and

H fields—which as we discussed, only apply to observable material entities anyhow and thus could not be used to “see” far below the quantum threshold of least detectable material disturbance. His entire protocol was to get beyond those “*material interception of EM energy”* limitations. Unwittingly, Rife was using *vacuum engines*—involving structuring of the active vacuum as well as pure general relativity (GR) and pure structurings of spacetime geometry itself. He was electromagnetically using that part of GR that the GR physicists have mostly only tried to produce by use of the weak gravitational force. In GR, *the ST geometry itself is active, dynamic, energetic, and structuring!* In higher symmetry electrodynamics, *the ST geometry itself is* *powerfully* *active, dynamic, energetic, and structuring!* It is not at all just a "passive spacetime" as classical electrodynamics assumes { [70] }.

Biological Systems Use the Infolded Whittaker-Type Electrodynamics

The problem in the Rife microscope is to be able to observe very subtle ST curvatures and patterns, in an optical manner. These subtle or virtual "disturbances" of local spacetime are far finer and smaller than the one-tenth wavelength optical resolution generally the limit for optical microscopes using conventional optical EM notions and “material” potentials, fields, and waves.

In short, Rife’s twofold problem was to (i) function with vacuum engines (internal structuring of the potentials) so he could penetrate to any degree of smallness, and (ii) output, in the observable state, exact visible summations and analogues of the individual structures detected in the virtual (nonobservable) state { [71] }.

To make the subtle vacuum engines observable on the bench, the ST curvatures and their internested patterning must be accomplished by making and assembling local curvatures of ST, via making longitudinal EM wavepairs in structural pattern assemblies. *Spacetime* is in fact just a powerful scalar potential, or can be modeled as such, and as such Whittaker 1903 and 1904 rigorously applies, along with extensions from higher symmetry electrodynamics. By realizing that ST is identically potential, one has altered one's thinking from the standard GR theory into a correlated EM-GR theory. One has also extended GR: Not only is there a “topology” of things that can be externally *done to* a base spacetime, but there is also a hidden, vast EM topology *within* any overall base spacetime. So we now have “topologies within topologies.” This is in fact what the Russians have long called the *information content of the field*.

However, all living systems already use this "infolded" bidirectional, longitudinal wavepair EM in their ongoing living functions { [72] }. Just as they used frequency modulation, EM signals, EM oscillations, etc. before we even had an electrodynamics or a physics, living systems do use the infolded EM (and vacuum engines) in all their living functions, and particularly in their cellular *regeneration and restoration* (R&R) system { [73] }, as contrasted to their *immune* system. The immune system cells are the fighters and the debris scavengers/cleaners. They go after the hostile invaders, fight them, and usually win—littering the battlefield with the debris. Then the immune system scavenger cells clean up the residue.

But many of the body’s cells will have been damaged in the fight. The immune system itself cannot heal or restore a single damaged cell, even its own! It is not a *healer*! It is a searcher, a marker, a killer, and a residue-cleaner. It contains " the troops and the combat engineer," so to speak, but not the hospitals and the doctors. All healing/restoration of living cells is done by the R&R system, not by the immune system.

A high-level overview of the immune system is given in Figure 20.

No vaccine, drug, herb, vitamin, or mineral *heals* the body, although certainly they can enhance or aid the body's healing process. Each does carry its individual resident vacuum engine, and when absorbed by the cell, this added vacuum engine contributes to the resident vacuum engine in the cell by summing with it. To restore the damaged cells back to normal (i.e., to heal), the R&R system uses a novel kind of extended electrodynamics with infolded vacuum engines, and it uses a novel kind of optical phase conjugate pumping, in the time domain rather than just 3-space. The magic “unified field theory” so long sought by scientists, has long been utilized by the regeneration and restoration system of the body in its minute-to-minute and day-to-day healing and restoring actions.

The magic "universal healing mechanism" so long evading the discovery of medical scientists is none other than the exact cellular regenerative mechanism universally used by every living organism. For the first time, we are explaining the higher symmetry electromagnetic nature of this mechanism.

Robert Becker's epochal work { [74] } (Figure 21, Figure 22) strongly probed the electromagnetic control system directing the body’s cellular regeneration and repair (R&R) system. However, he did not have available the infolded Whittaker electrodynamics. He could only utilize the grossly inadequate standard electrodynamics. Modern nonlinear phase conjugate optics had not yet been born when Becker did most of his seminal work.

Since standard EM does not incorporate infolded GR, vacuum engines, or optical pumping in the time domain, Becker could not formulate the full technical mechanism of the R&R system —which uses the *supersystem*, not just the system. But Becker did strongly point out the importance of the R&R system, both theoretically and experimentally. He found that it operated electrically, with the scalar potentials being key. He modeled the R&R system as closely as can be done with classical U(1) electromagnetics (Figure 22). He proved that cells can be *dedifferentiated* (reversed back to an earlier, more primitive state) and *redifferentiated* (time-forwarded into a more complex state) by weak DC potentials and laughably weak (picoamperes) of current. He was twice nominated for a Nobel prize for his epochal work. His method of electromagnetic healing of otherwise intractable bone fractures (Figure 21) is utilized today in many hospitals.

In his microscope, Rife was in pursuit of optically "seeing" the infolded electrodynamics functioning of the living cells in the most minute detail—i.e., he wished to see the actual vacuum engines and their functions. Using the infolded EM, one can still "see" the cells and all their parts, but one can also go far beyond the one-tenth wavelength optical discrimination limitation and see into the virtual state dynamics. In theory, because the infolded EM is *recursive* as a function of wavelength, one can go to any limit in smallness desired—or at least to whatever limit is determined by the available technology's development at the time { [75] }.

Vlail Kaznacheyev's novel electromagnetic induction of cellular death, disease, and damage, *at a distance*, is instructive in this respect { [76] }. (Figure 23) Indeed, the Russians mastered, extended, and weaponized the Kaznacheyev mechanisms as disease-inducing weaponry, and slyly used it at low level on personnel in the U.S. Embassy in Moscow (Figure 24). This was done to stimulate high level officials in the U.S. government and U.S. scientific community to see if the U.S. knew of this "infolded electrodynamics and vacuum engine" technology { [77] }. The Russians were using GR, but by infolded EM means—so they were using a unified field theory involving higher group symmetry electrodynamics. The puzzled U.S. employed only standard U(1) symmetry electrodynamic analysis { [78] }, so it totally missed the entire physics of what was occurring to cause the health changes and diseases.

In GR, one is very interested in Wheeler's principle which states (slightly paraphrased):

> *Mass (trapped energy) acts upon ST geometry to curve it, and curved ST acts upon mass to produce forces upon it and move it.*

This statement after Wheeler is the very essence of general relativity. It simply states that there is a *mass-to-spacetime* dynamic transform, and also a *spacetime-to-mass* dynamic transform, as we diagrammatically showed in Figure 9.

Understanding Vacuum Engines and Their Action

Wheeler's general relativity principle includes the missing two transforms of electro-dynamics. We are extending Wheeler's principle to the corollary form:

All levels of the structures of mass (trapped energy structures) and all levels of the structures of time (the time aspects of photon structures) act upon ST geometry to structure it in both 3-spatial energy density structuring and in time domain structuring. Producing this multilevel spatial and temporal structuring of spacetime is called "forming a vacuum engine." Both 3-spatial energy density structuring of spacetime at all levels and time structuring of spacetime at all levels, act upon mass at all levels, to produce internested templates of forces and translations and stresses.

In short, energy patterns and time patterns in 3-space act upon spacetime to produce patterned curvatures of ST geometry (vacuum engines). And conversely, patterned curvatures of spacetime geometry (vacuum engines) act upon mass to produce corresponding patterns of forces at all levels in the mass —down to and including upon the quarks and gluons in the nucleons in the atomic nuclei. We note with pleasure that the Sachs unified field theory does apply from the gluons to the entire universe, as has been shown by Sachs himself {18b}.

Muscle and Skeleton Analogy of Interaction of Vacuum Engine and Mass

We call attention again to our analogy (Figures 15, 16, 17, and 18) of the interaction of a vacuum engine and mass, provided by the interaction of arm muscles and the skeletal arm bones to which they are attached. The arm muscles attach to the arm's skeletal bones, and those bones are articulated and free to dynamically move. To the bones, the muscles are mysterious agents (analogous to vacuum engines) exerting forces upon them, and causing them to move or hold. To the muscles, the bones are positioned resisting masses for them to work upon and change their positions, and the bones exert Newtonian reaction forces back upon the muscles as the muscles apply forces to the bones. So the vacuum engines (muscles) create forces upon the resisting masses (bones) to move them, and the masses act back upon the causative vacuum engines to create antiforces in resistance.

Using this analogy, when we create a set of vacuum engines, we create a set of “mysterious muscles” that act directly upon any and all “masses” (bones) that are exposed to them. The bones then change and move until the bidirectional interaction forces are balanced (equal and opposite). This symmetry between the two missing transforms is shown in Figure 9.

The Multilevel Structuring of Time is Important

We also add the extension that dynamic structuring in the time stream also occurs from trapped time domain “time component” structuring and distribution effects, as well as dynamic structuring of the spatial energy density of the ST geometry (the vacuum).

We are speaking *relativistically* of the interaction of mass and spacetime at any and all levels of a mass, down to and including the molecules, the atoms, the atomic nuclei, the nucleons, and even the quarks comprising the nucleons (comprising the protons and neutrons). One must first clearly have this vision in mind: The dynamics of everything ongoing in that mass, down to the most minute particle dynamics and in the most minute detail (and far into the virtual state beyond the one-tenth wavelength limitation of light resolution) consists of the structurings of this "mass-ST curvature" mutual interaction.

So we know from general relativity that collections of energy (such as mass) act directly upon spacetime, its structuring, and its dynamics at any and all levels, down to the tiniest. We know that the variations in local spacetime geometry (including in both energy and time), down to any and all levels and sizes, down to the tiniest conceivable lengths, act directly upon any and all exposed mass at that level, *no matter how tiny*.

There is no “quantum” limitation nor is there any “wavelength of light” limitation to this GR process. Total continuity applies.

Some Characteristics of Vacuum Engines

The great advantage of a vacuum engine is that the "action" it produces upon an exposed mass, arises directly from within each point in the local spacetime in which the mass is embedded. One does not have to "start outside" and move through space and the intervening mass toward the "inside". Every point within an object occupies every point in time, a priori. From a single time-point, the vacuum engine's 3-space action enters everywhere within the object, all-at-once, and the 3-space effects proceed outward from each interior point. (See again Figure 19). Therefore even a weak vacuum engine induced by EM forces can do with ridiculous ease what the most powerful particle accelerators on Earth cannot do. The engine is already everywhere "inside" an object or particle, and working from every point in it, toward an outward direction. On the other hand, a particle accelerator is trying to build tremendous velocity of its particles so that they can "smash" their way in there, by brute force. To get ever deeper, the particle accelerator must be made ever larger, so that a "bigger hammer" is available. Also, the vacuum engine does not smash or damage any of the matter in the object, in any fashion not desired. The particle accelerator, on the other hand, is a great disruptor and may well destroy or serious change the very object it "probes".

Further, contrary to the accelerator's "one shot for one giant amount of energy input", a vacuum engine—once formed—continually exists in that locally altered spacetime and is contained in it and is sustained by it.

Hence that locally altered spacetime—after we pay to alter it once—will continuously expend energy in the form of the action of that vacuum energy, forever, upon any mass emplaced in that local spacetime. The universe itself is furnishing the energy to do the work, via the instantaneous connection in the time domain.

That "multiply connected space" entry everywhere within an object simultaneously, is quite different from "spatial propagation of energy" through singly-connected space from outside an object to the object, striking the outside of a mass and being absorbed, producing a gross translation force upon the mass. In the “energy propagation through space” case, severe energy interaction may occur with only the absorbing part of a macroscopic mass, while the remainder of the mass is little affected or not at all. To wit, one does not get the resulting *electronuclear* alterations of the nucleus (materialization, dematerialization, transmutation, etc.) by absorbed propagated “normal” EM radiation. However, with weak EM-induced vacuum engines one can do things inside the nucleus such as easily flip quarks inside the nucleons, so that transmutations at very weak spatial energy levels are readily obtained. This in fact turned out to be the major secret of the low-energy nuclear transmutations occurring in some 600 cold fusion experiments worldwide { [79] }.

Electrodynamic force fields implicitly define primarily only translation effects upon charges or masses, in their very definitions. With normal EM, usually one will get reradiation of the absorbed energy, or partial reradiation of the absorbed energy accompanied with recoil of the absorbing mass, or no reradiation but recoil of the mass.

We took the term *vacuum engine* from Misner, Thorne, and Wheeler { [80] } and also from Nobelist T. D. Lee { [81] } who spoke of this sort of thing as *vacuum engineering*. The term is not original with us; only the novel higher symmetry electrodynamics application of the concept is original. We call the exact geometrical form of a vacuum engine a *template*. We refer to the process of internally structuring a field, wave, or potential as *dimensioning* or *activating* or *conditioning* or *charging up* the field, wave, or potential. Russian superpotential weapon scientists call the infolded structuring itself the *information content of the field*. For further discussion of some of these terms and related concepts, the reader is referred to our website, www.cheniere.org.

Wheeler's GR principle says that, for every mass with its exact mass structural dynamics condition, there exists a corresponding exact vacuum engine template containing precise spacetime curvature structural dynamics, and interacting back upon the dynamic mass at all levels. That vacuum engine is said to be resident in the mass and continuously interacting with it (see again Figure 13 on the *supersystem*).

Again, the structuring of the vacuum engine has no limitation in fineness, because spacetime is continuous. Hence it freely extends from the observable realm on down into the virtual realm. This is no more audacious than stating that virtual and observable entities occupy the same time domain simultaneously. Obviously, if one is using the structuring of time, one gathers in both those virtual structures and those observable structures, and also the exact correlation between them. This is a direct engineering application of an EM hidden variable theory, somewhat similar to that proposed by David Bohm { [82] }.

Applying Vacuum Engines to Living Cells

Any normal cell already has a normal *template* of overall ST curvature due to its mass and trapped energy structures. No matter how small, every internal part of that cell has its own *subcomponent template* of subcomponent ST curvatures embedded in that overall cellular ST curvature. As the cell functions, the dynamic normal changes of that template into other related templates, is within the range of what we call "normal, healthy cellular functioning."

Any abnormal cell has an *abnormal template* and also an abnormal set of "templating" changes. In other words, there exists a precise *vacuum engine delta* in the abnormal cell, summed to the cell's *normal vacuum engine*. The *sum engine* represents the exact vacuum engine to act on the normal cell and change it to that precise abnormal condition.

But a template depends upon the energy density structuring of the local vacuum and also the local *time* structuring—which latter is still missing from physics books, but is implicit in the nature of the photon. If one produces the necessary energy structuring and density—and the necessary time structuring and density of local ST geometry—one produces a certain vacuum engine and one also produce a certain *template of time structuring*. Time { [83] } has dynamic template structuring and functioning, just as does the 3-spatial energy density of ST geometry.

Indeed, the total (virtual and observable) photon interaction with a mass or particle generates the "flow of that mass or particle through time" (Figure 25).

But one *can* make vacuum engines using EM energy and structuring, rather than with normal gravitational (G) energy and structuring { [84] }. The advantage is that, for electrons, the EM structuring is on the order of 10^{42} times as strong as the G structuring. For protons, the EM force is something like 10^{38} times as strong as the G force. So by using the much stronger EM force as the "agent of ST curvature," one can make vacuum engines that are quite strong at their local level. These engines will affect and profoundly change any exposed mass, including cellular mass—in any fashion desired, given only that one makes the appropriate vacuum engine. Again, *we are simply engineering practical but ultra-strong general relativity on the laboratory bench*, by electromagnetically engineering Wheeler's principle in an extended form.

For any physical cellular deviation from normal—any deviation whatsoever—a corresponding exact deviation (i.e., an exact delta) exists in the cell's otherwise normal vacuum engine template. This includes deviations in the genetics of the cell and everything else, not just the chemistry and gross physical structure.

Now if one could produce the negative (time-reversal) of the sum template of the resident vacuum engine and *amplify* it, one would have a powerful new type of vacuum engine. (Figure 26) It would be an exact *vacuum anti-engine* to act upon that specific diseased or abnormal condition of the cell and gradually erase the delta and bring the cell physically back to normal. The new anti-engine would operate upon all the mass of the cell, at every level no matter how small. It would change the cell's abnormal genetics back to normal genetics.

One could of course reverse cellular damage due to aging, and rejuvenate the body. It is quite doable; all that is needed are the funds, scientific team, and laboratory facilities to develop it. It is doable whenever the U.S. medical science community and the giant pharmaceutical community will allow it to be done.

One could reverse a cancer cell back to a normal cell, by applying the specific summation vacuum antiengine for the exact delta between the cancer cell and a normal cell.

One could reverse an AIDS cell back to a normal cell with normal genetics, where it no longer was an "HIV factory". An AIDS patient could be completely, quickly, and cheaply cured, the moment he or she tested HIV positive. There would be no need to wait for the debilitations of the progressing disease to appear. Not a single HIV-infected cell would be left in the body, even those resistant strains that have developed after the original infection.

One could do precisely the same thing with Ebola infections, and with the “tough” staph infections that now resist everything and are killing more than 120 thousand Americans in our hospitals every year, with the figures rising. *These are patients who did not have the disease that killed them, when they entered the hospital.*

With development of portable treatment units and mass production, one could treat and heal those millions of mass casualties that could well erupt in one or more large U.S. city areas—e.g., from an Iraqi-sponsored terrorist biological warfare (BW) strike, using anthrax { [85]). In 1998, we proposed a crash development of just such portable units { [86] }. With the present world situation, it is not a matter of *if* such terrorist BW strikes will be made in the U.S., it is a matter of *when* { [87] }. Since our 1998 proposal, there has now been time to have had those portable units developed and mass-produced, at a total program cost of perhaps $100 million. Sadly, no one apparently had the foggiest notion of what was actually being proposed, and its vital importance to the survival of the nation. With a looming war in the MidEast—and with hordes of terrorist teams with biological warfare agents already infiltrated into U.S. cities and population centers—this negligence is particularly of concern at this very moment. These officially-forecast attacks on our cities by weapons of mass destruction could now be imminent, if and when the MidEast does explode into war, which appears imminent.

The Process Has Been Experimentally Proven

A very similar thing is also what Becker { [88] } demonstrated, without considering the vacuum engine (general relativistic curvature of spacetime) effects. It is also precisely what Prioré { [89] } unwittingly did in France, with a team of renowned French scientists, to cure terminal cancers, atherosclerosis, lethal infectious diseases, etc. in laboratory animals under rigorous scientific protocols. The team also demonstrated the restoration of suppressed immune systems.

In Germany, Popp { [90] } *et al*. were also involved in studies dealing with the R&R system, but from a quantum electrodynamics viewpoint.

We can also consider the vacuum engine from a nonlinear optics view, and obtain startling extensions of both general relativity and nonlinear phase conjugate optics. That in fact is exactly what Prioré did, though he did not think in those terms. Prioré roughly created a complex set of bidirectional longitudinal EM wavepairs—i.e., a set of “time domain pump waves.” So he produced a kind of vacuum engine that itself is a crude *anti-engine generator* when used to pump a mass such as a living cell.

Prioré unwittingly used an extension to nonlinear optical (NLO) phase conjugation *before nonlinear phase conjugate optics was even born*. Pumping diseased cells in the Prioré manner causes the formation of exact "antiengines" for the specific cellular disease or disorder to be created in the cells themselves.

First, Prioré mixed up to 17 or so EM transverse waves in a rotating plasma { [91] }. Today we know that such a plasma will transform transverse radiations into longitudinal radiations, and it will also phase conjugate the input waves and add their phase conjugate replicas. In short, it is a procedure for somewhat crudely making a set of longitudinal bidirectional EM wavepairs, which is required to produce a scalar potential with a deterministic internal structuring. In short, it produces a scalar potential with an internal vacuum engine template.

These infolded longitudinal biwaves are NLO pump waves of a new kind: They pump the PCM material in the time domain only. This kind of pumping is a *vacuum antiengine-making process*.

The plasma tube was surrounded by a giant coil, and a pulsed magnetic field was produced by the coil. Per Whittaker’s 1904 paper, that pulsed DC magnetic field can be decomposed into two scalar potentials. Hence we have three potentials superposing: (1) the potential function set made by Prioré in his plasma tube, and containing a selected “vacuum antiengine maker” consisting of bidirectional pump waves in the time domain, and (2 and 3) the two Whittaker potential functions comprising the DC pulsed magnetic field of the coil.

The end result of this superposition/mixing of potentials is that their “internal structures”—including their vacuum engines—also diffuse one into the other.

Fundamentally, Prioré produced what to the observer seemed just a pulsed DC magnetic field, but one in which there was a specific “vacuum engine maker” of his deliberate creation and “tailoring.” Originally by trial and error, he adjusted the 17 frequencies introduced to the plasma, thus varying the frequencies of the “infolded bidirectional pump waves” produced and infolded into the pulsed DC magnetic field carrier. In this manner he was “tuning the “vacuum antiengine-maker” content.

Again, Russian energetics weapons scientists call those vacuum engine templates the *information content of the field*. They do not refer to ordinary spectral analysis.

Explaining the Time-Pumping Vacuum Anti-Engine Maker

Irradiating the cell (with non-ionizing radiation) containing infolded specific bidirectional longitudinal waves, precisely corresponds to pumping the cells in the time domain, rather than pumping them in the “spatial energy distribution” domain. Because of the shift of the pumping from the 3-space domain to the time domain, this constitutes a dramatic extension of phase conjugate optics. Because of the amplified phase conjugating action of the “vacuum antiengine maker” upon the vacuum engine in a mass, this also constitutes a dramatic extension of general relativity and of four-wave mixing theory. Let us explain briefly:

When one pumps a nonlinear mass [phase conjugate mirror (PCM) material] with ordinary “transverse” EM waves { [92] } (Figure 27a), one is pumping (stressing, or “squeezing and relaxing”) the mass with dynamic oscillating energy distributions in 3-space. One inputs a similar weak wave (called the “signal” wave) to the mass simultaneously. In that case, multiwave mixing theory tells us that the pumping of the PCM will cause the formation of a fourth EM wave, that is a time-reversed replica of the input “signal” EM wave. The time-reversed wave may contain up to all the energy in the pump waves, so it can be highly amplified in magnitude compared to the input signal wave. The distortion correction theorem tells us that this time-reversed replica wave will then appear precisely superposing in space back along the exact spatial track taken by the input signal wave { [93] }.

When the pumping is shifted to the time domain by pumping with bidirectional *longitudinal* EM waves, we now are pumping the time aspects of the mass—during it iterative existence as masstime; see again Figure 25—rather than just pumping the spatial energy aspects. In other words, we are not pumping *spatially*, but *spatiotemporally*. We are pumping *temporally*—rhythmically squeezing (compressing) and relaxing (extending) the "density of the time interactions" producing the rate of flow of time itself. This is a totally new extension to nonlinear phase conjugate optics.

In this case, the "signal wave input" to the "time-domain-pumped" nonlinear mass PCM is just the resident vacuum engine in the local spacetime in which that PCM mass is embedded (see Figure 27b). The sickened cell's abnormal vacuum engine—consisting of the sum of the normal cellular vacuum engine and the disease delta vacuum engine—is present in the cell and is the input to the time-pumped phase conjugate mirror (cellular masstime). The output is the formation of an amplified time-reversed replica of that vacuum engine. In other words, an exact, amplified *vacuum antiengine* for reversing the precise abnormal condition of the mass, down to and including even the quarks in the atomic nuclei, is formed by time-domain (longitudinal EM biwave) pumping. The result or "output" is that *the entire cellular PCM mass and all its constituent parts are "time-reversed" back along their previous "track through time", back to a previous normal physical state* { [94] }. Figure 27b illustrates this process.

So when Prioré pumped the abnormal cells, he simply time-reversed them—genetics and all—right back to their normal physical state before the abnormality arose.

The beauty of the method is that, by adjusting the input waves to the rotating plasma, Prioré could “tailor” and “tune” the pumping waves to what was needed for a particular class of disease. He used trial and error methods to arrive at the proper settings for the disease to be treated. In theory, by the Prioré method one can time reverse any mass back to an earlier physical state. This mechanism is used by the cellular control portion of the body’s regenerative and restorative (R&R) system to restore damaged cells back to normal. It is the basis for the body’s ability to heal itself!

Applied to diseased cells, the method causes the formation of specific amplified antiengines for the exact specific cellular abnormality, no matter how complex, and no matter what it is. *In applied medicine this is a revolutionary new healing and therapeutic modality*, accomplished by novel EM means, and justified by general relativity, higher symmetry electrodynamics, and longitudinal EM wave pumping in the time domain rather than classical electrodynamics and transverse EM wave pumping. Prioré’s discovery, once understood, may just be the greatest medical discovery of all time, and the discoveries of Rife and of Becker may rank a close second. The Prioré process is shown in Figure 28.

The body's cellular regeneration system uses this same methodology and process. A block diagram of the operation of the body's cellular regeneration system is shown in Figure 29.

Further Principles Needed for Rife’s Microscope

In his microscope (Figures 1 and 2), Rife just wanted to observe the templates [2] corresponding to the cells, germs, viruses, etc. and all their parts, without any smallness limit. To do that, we need to (1) amplify the tiny, tiny amplitudes (actually, in the virtual state and subquantal) of the tiny vacuum engines accompanying tiny spatial locations containing living, functioning entities (no matter how small), and (2) "transduce" that amplified result into the optical spectrum, so that we see a perfect optical analog and thus are "looking" visually at localized ST regions far smaller than the one-tenth wavelength optical resolution limit.

We need another principle to do that. Suppose that, to first order, we have an isotropic nonlinear medium. Then if we transmit a sine wave into it, pandemonium sets loose with our wave. The speed of each part of the wave in this medium is a function of the location of that part upon the wave amplitude, so the peaks move faster than the intermediate parts of the wave of lesser magnitude. The wave has "peak overtake", breakup, collapse, and all sorts of things happening to it. So a simple sine wave is not going to pass smoothly into that medium and "pass back out" coherently and nicely, and ordinary light is not going to give stable reflections from that medium!

However, if we transmit *two* sine waves simultaneously, some frequency interval apart, and then if we *pretend* that we actually transmitted their *difference* frequency, a very strange thing happens. The two individual sine waves themselves still break up and go bananas. However, that difference frequency acts as if it were a simple sine wave transmitted into a *linear* medium! So we get back out a very nice, smooth, well-behaved sine wave of the *difference* frequency, even though terrible things happened to the individual waves themselves. We can detect and use that "difference frequency" nicely.

Several very neat sonar systems presently use that sort of transmission schema in underwater work { [95] }. The principle applies for waves of all frequencies, for all media which can be approximated as nonlinear isotropic in nature. As we shall see, it applies to the Rife microscope also.

Rife's Secret: Recursive Coherent Increase of the Signal-to-Noise Ratio (SNR)

To first order, without too much error we can pretend that the molecules and atoms etc. in the cells and body tissues, together with their finer components, comprise such an isotropic nonlinear medium. That's not entirely true, of course, but it does work to an extent sufficient for our purposes. Fortunately for the Rife process, the parts that *do not* obey this isotropic rule add randomly, while the parts that *do* obey the rule add coherently. So the "signal-to-noise ratio" (SNR) increases with the increasing amplification, and one "grows" the desired signal right up out of the noise (SNR) of the parts that do not obey the isotropic medium assumption. With sufficient number of amplifier stages, in theory the SNR of even the tiniest parts can be made as large as desired { [96] }.

With optical functioning as in Rife's microscope, one may substitute "light gathering power" or "magnification" for the electrical term "amplifier gain".

As we stated, Whittaker in 1904 (and superpotential theory thereafter) shows us that—if we wish—any EM field, wave, and pattern can be represented by two scalar potentials. Each of these two scalar potentials is further comprised of bidirectional EM longitudinal waves, in a harmonic series, in the Whittaker 1903 manner as reinterpreted. So we can reduce any pattern, field, or wave from any assemblage of charge sources, into two potentials, and then further into two sets of bidirectional EM waves interfering with each other. That process already *makes* all conventional EM waves, fields, etc.

An EM structuring of the local vacuum-spacetime occurs by all the electrical charges in the mass of the living cell, down to and including the virtual charges exchanged in the nuclear interactions in the nucleus of the atoms. By Whittaker 1903 and superpotential theory, all of that structuring itself is a potential with an internal set of bidirectional harmonic longitudinal EM waves interfering with each other.

Further this potential is recursively organized, because the inner bidirectional EM wavepair structuring of the potential is already recursive, as can be shown by applying Whittaker-1903 and Whittaker-1904 successively and iteratively { [97] }. This means we get deeper and deeper internested levels of patterns that still precisely correspond to the operational mechanisms of the lower levels. It's similar to an infinitely fractal { [98] } type of process.

Because this conglomerate is recursive as a function of harmonic intervals, it means that we can pick any harmonic interval we choose in the "big world" where we observe things, and we can have an exact observational representation of the dynamics of that ongoing hidden longitudinal EM wavepair interferometry in the two waves bounding that harmonic interval. Of course, the more additional harmonics we add, the better dynamics we have. The greater the number of harmonic intervals applied, the greater the gain. However, these harmonic intervals can be recursively applied using just one "macroscopic" harmonic interval, iteratively and serially applied over and over. The more iterative serial recursive application of one harmonic interval, the mote the "gain" or "amplification"—in the case of a microscope, the more the *amplification*. That was Rife's real secret, finally revealed. See Figure 30.

The Difference Frequency Provides Visible Output

Rife wanted to look at very, very small things *visually*, even though those “small things” were far below the one-tenth wavelength visible light barrier. In other words, to understand the Rife microscope’s visual representation of the virtual state, we need to choose the harmonic interval of the throughput so as to span the visible spectrum. Well, for optical purposes, luckily the IR and UV do exactly that. Examine the frequencies, and you will see that a harmonic interval exists from a spot in the IR to a spot in the UV, and of course the interval in between includes the entire visible spectrum. The visible spectrum in between, however, represents the observable noise that would block what we seek to have emerge in that region in our "Rife microscope", so the presence of the visible spectrum is highly undesirable. Various filters can be used to block the visible spectrum while transmitting the UV and IR regions, or one can just work in the dark or in suitably dim visible light.

Rife then applied his iterative multiple "difference frequency" lenses in serial order, one after the other. While it is doubtful that he understood the theoretical method being utilized, he certainly persisted and found how to use it. This method of iteratively summing a recursive virtual state vacuum engine "template" results in the formation of an observable state vacuum engine template. In short, it magnifies the virtual state entities and dynamics directly to observable state analogues that capture all the virtual state entities and their dynamics.

That is why, if we use special filters on your camera and flash lenses, where the filters are open in the IR and UV but opaque in the visible spectrum in between, we can often photograph the ongoing "longitudinal EM hidden biwave interferometry"—*the subtle energy or vacuum engine functioning*—of nature and of living systems, *even displaced in time*, as Trevor Constable { [99] } and Joe Gambill { [100] } (a researcher friend now deceased) demonstrated, as George Meek { [101] } demonstrated, and as Fogal { [102] } has successfully experimented with.

Occasionally when one gets things just right, one can get pictures of the subtle invisible reality in the virtual state, going on around oneself and in one's surroundings. With Rife-like iterative serial amplification, one can get them all the time, dependably, as with a Fogal semiconductor specially adapted for such operation. One can in fact build instruments that indirectly observe the so-called "unobservable" virtual energy state and its entities and dynamics with one-to-one correspondence.

That's a whole new "subtle physics" reality. Present physics with its assumed quantum observation limitations just has not even gone into it—primarily because physics has been focusing upon the quantized aspects of the photon (which is quantized in one aspect only—its magnetic vector potential—and not in its unquantized aspect—its scalar electrical potential!). Physics has rather ignored the unquantized (and therefore continuous from the smallest to the largest scale) time aspect of the photon { [103] }. Using the continuous aspect of the photon, one can assemble photon “time interval” structures—and in fact one does so whenever one assembles a 3-spatial energy density structuring such as in ordinary EM theory. Simultaneously, one has assembled a “structuring of the time domain” also, because each photon in the spatial energy density structuring consists of both time and energy as its canonical components.

All EM fields and waves, being comprised of photons, thus also carry (1) energy and energy internal structuring, and (2) time and internal time structuring. Again, this is the Russian *information content of the field*. The Russians simply understand 4-vector theory and Whittaker 1903 and 1904, and how to combine them with quantum mechanics and general relativity in a unified field theory.

The above explanation is why Kaznacheyev's { [104] } mitogenetic radiation { [105] } from his control (deliberately damaged or infected) cells would pass through a quartz window and induce the same damage (or disease) in targeted cells on the other side and otherwise environmentally shielded, but would not induce the changes through a normal window glass window. (See again Figure 23) Quartz passes the full optical spectrum from IR to UV, while window glass does not. So quartz passes the *vacuum engines* for the disease or damage, while window glass breaks them up and does not pass them as such. We stress that the vacuum engines are part of the “continuous” aspect of the photons, and so are not bound to any limitation on magnitude (either fineness or largeness) as are the quantized aspects of the photon.

For such a "far beyond the one-tenth wavelength" microscope as Rife’s, one must use vacuum engines (time domain structuring, and longitudinal EM waves) and extraordinary iterative unified field theory optics, rather than the ordinary transverse EM waves and ordinary optics. The lenses must all pass at least one harmonic interval (and the same one!), to provide and utilize the “difference frequency band” needed for recursive *visual* connection to the subquantal state (the hidden virtual structures in terms of the continuity aspects of photons). Successive stages thus "serial iteration" use the recursive principle to coherently amplify and show the vacuum engines, their dynamics, and the dynamics of deeper organizations and dynamics in the virtual state itself. Since Rife wished to “see” the results in the optical spectrum, then the two frequencies chosen for “differencing” had to bound the visible spectrum.

Rife used quartz lenses or at least very good optical glass that is open from IR to UV. That bounds the visible spectrum with a harmonic interval. Rife used multiple stages for successive coherent recursive amplification and increasing the SNR magnitude by coherent integration. One must also have some scheme for filtering out (or avoiding) too much visible spectrum light. One such scheme is to just use enough stages so that, in the new theory where the wavelength limitation is bypassed, one has amplified ("resolved") to the depth or fineness desired. The process long ago removed itself from the "normal visible light" domain, well before it reached the deeper resolution limit utilized by the definition state of the new method.

By using sufficient multiple stages in this fashion, the "tiny" virtual state vacuum engine templates passing through the IR-UV harmonic interval in a lens can be recursively summed and magnified many, many times—essentially without limit, because the phenomenon being magnified is recursive and bidirectional longitudinal in nature. It is definitely *not* normal visible light optics, and it *is not* conventional classical electrodynamics.

Rife’s Recursive “Vacuum Engine” Microscope

The Rife microscope coherently sums (and thus coherently amplifies) the virtual state vacuum engine patterns with the increasing number of stages. For the *subtle* vacuum engine templates, such an arrangement becomes a direct amplifier (magnifier), for amplifying (magnifying) the internested virtual state vacuum engine templates and their dynamic functioning, into the observable state. With a properly designed “vacuum engine” recursive microscope, one can see inside a molecule, inside the atomic nucleus, and even inside a nucleon, given development and perfection of the process. However, at this initial stage of understanding, one can forecast that finely adjusting the various stages of an early laboratory prototype would indeed prove to be a tedious and formidable task, just as Rife reported. With further development, however, that obstacle can be removed.

What comes out the output end of such a "vacuum engine" recursive microscope, is a direct visible light *analogue* of the much more subtle ST functioning in the microscope's highly localized viewing area. This output imaging is put together "electrically" in ways analogous to the construction of a video picture on the screen of a TV tube. Since the result emerges through an output end that still passes an entire harmonic interval bounding the visible spectrum, the amplified virtual state recursive template is not destroyed but is highly amplified and produced as an observable image and dynamics in the visible spectrum. This final "interferometry" of two *internally structured* potential functions now produces "visible light" EM fields and waves and patterns, in accord with Whittaker 1904. These visible patterns are direct analogues representing exactly the very, very minute regions far below the one-tenth wavelength limitation of ordinary optical resolution. Those regions were the regions "being focused onto" when Rife so arduously "adjusted" his microscope's multiple stages.

With Rife's recursive magnifying microscope process, one can see direct visible analogs of what is actually far, far smaller than anything ever before seen in normal optical microscopes, or even in electron microscopes etc. Indeed, one can see directly into the virtual state itself, and observe virtual state structures. While this is unthinkable in quantum mechanics if one considers only quantized aspects of light, it is perfectly feasible if one considers the concomitant unquantized (continuous) aspects of light, and shifts to general relativity and vacuum engines to utilize the unquantized structuring of the time domain.

Conclusion

Rife unquestionably found a way to produce visible recursive amplification and magnification of invisible virtual state dynamics and functions. We hope we have shed at least some light on the extraordinary technical mechanisms involved. This is good conceptual science, consistent with theory and some known experiments, and it certainly can be made a rigorous science and technology and mathematically modeled, using higher symmetry electrodynamics that is part of a unified field theory such as that of Sachs. The basis for all of it is in the literature already, but it is scattered widely in uncorrelated bits and pieces. It is not in the normal electrodynamics courses, nor is it in the normal optical science curriculum. And it is not taught in the normal physics curriculum. Needless to add, it is absent from the medical curriculum as well.

As an additional proposal, we tentatively suggest that this principle also may be the long-sought "hidden chaos" or "hidden order" in quantum mechanics that is known to be present, but has not yet been explained { [106] }. If so, then it becomes a new law of physics, with the experimental proof first demonstrated by Royal Raymond Rife.

With the conceptual technical matrix of Rife’s mechanism in hand, it should be possible for a proper scientific multi-disciplinary team to redevelop and extend the Rife microscope. Marvelous new instruments will emerge, once the quantum physics conceptual blockage of continuous physics is constrained to its place as only “one view of the physics reality elephant.” With the ability to directly “measure”, “observe,” and engineer the virtual state, a breathtaking expansion of physics will occur, dramatically extending the vacuum engineering suggested by Nobelist Lee.

If so, then Rife’s tragic life will have served its noble purpose, and his intuitive insight will have been resoundingly verified. The follow-on results will have untold benefit to science and to all the human generations to come.

NOTES AND REFERENCES

---

[1] For work on the theoretical aspects of this paper, partial financial support by A. Dreyer, R. Francis, H. Ledger, M. Peters, T. Craddock, and the Association of Distinguished American Scientists (ADAS) is gratefully acknowledged, as are helpful discussions with Dr. Bob Flowers and Dr. Myron Evans.

[2] We emphasize that Rife himself did not think in these terms.

---

[1] . (a) Some Rife-related references are: Royal R. Rife, "The filterable virus of carcinoma," Apr. 26, 1939; — "The Rife Microscopes," Apr. 28, 1939; — "Bacteria," Apr. 28, 1939; — "History of the development of a successful treatment for cancer and other viruses, bacteria, and fungi," Allied Industries, Dec. 1, 1953.

(b) Arthur Isaac Kendall and Royal Raymond Rife, "Observations on Bacillus Typhosus in its filterable state," California and Western Medicine, Ca. Med. Assoc., Dec. 1931.

(c) Arthur Isaac Kendall, "The filtration of bacteria," Science, Mar. 18, 1932.

(d) See particularly E. C. Rosenow, "Observations on filter-passing forms…," Proc. Staff Meeting of the Mayo Clinic, July 13, 1932; — "Observations with the Rife microscope of Filter-passing forms of micro-organisms," Science, Vol. 76, Aug. 26, 1932, p. 192-193; — "Microdiplococci in filtrates of natural and experimental poliomyelitic virus compared under the electron and light microscopes," Proc. Staff Meeting of the Mayo Clinic, 17(7), Feb. 18, 1942, p. 99-106; — "Observations of filter-passing forms of Eberthella Typhi (Bacillus Typhosus) and of the Streptococcus from Poliomyelitis," Proc. Staff Mtg.

Mayo Clinic 7, July 13, 1932, p. 408-413; — "Transmutations within the Streptococcus-Pneumonococcus Group," J. Infective Diseases, Vol. 14, 1941.

(e) R. E. Seidel and M. Elizabeth Winter, “The new microscopes,” Journal of the Franklin Institute, Feb. 1944, p. 103-129; — "Filterable bodies seen with the Rife microscope," Science Supplement, Science, Dec. 11, 1932, p. 10-11.

(f) "Is a new field about to be opened in the science of bacteriology?", Editorial, California and Western Medicine, Ca. Med. Assoc., Dec. 1931; "Powerful microscope makes very tiny objects visible," Science Newsletter, Vol. 33, Jan. 22, 1938, p. 55; "The Rife microscope or 'facts and their fate'," Reprint No. 4l7, The Lee Foundation for Nutritional Research, Milwaukee, Wisconsin, undated.

(g) A good lay overview is given by Christopher Bird, “What has become of the Rife microscope?”, New Age Journal, Mar. 1976.

(h) See also T. E. Bearden, AIDS: Biological Warfare, Tesla Book Co., 1988, p. 267-279. The author gives a short introduction to the Rife microscope and includes some pictures. The proposed evanescent wave explanation proved to be wrong. Much additional work by the author eventually culminated in the proposed mechanism presented in the present paper.

[2]. We have one viewgraph listing some 32 flaws, errors, and shortcomings of classical electrodynamics, and there are even more than that. Additional errors are listed, for example, by important scientists such as Feynman, Wheeler, Evans, Bunge, Cornille, Barrett, and others. So a substantial number of errors in classical U(1) EM theory are well known to foundations scientists. We discuss only a minimal few in this paper.

[3]. Mario Bunge, Foundations of Physics, Springer-Verlag, New York, 1967, p. 176.

[4] . *Ibid*., p. 182.

[5] . M. Faraday, Experimental Researches in Electricity and Magnetism. Vol. 1, Taylor and Francis, London, 1839.; Vol. 2, Richard & John E. Taylor, London, 1844; Vol. 3, Taylor and Francis, London, 1855. Reprinted by Dover in 1965. Also see M. Faraday, "On Static Electrical Inductive Action," Phil. Mag., 1843; — "Remarks on Static Induction," Proc. Roy. Inst., Feb. 12, 1858.

[6] . James Clerk Maxwell, Preface to the First Edition, A Treatise on Electricity and Magnetism, third edition, Vol. 1, Dover Publications, New York, 1954, p. viii-ix.

[7] . J. C. Maxwell, Preface to the First Edition, A Treatise on Electricity and Magnetism, third edition, Vol. 1, Dover Publications, New York, 1954, p. viii-ix. Quoting: *"…before I began the study of electricity I resolved to read no mathematics on the subject till I had first read through* Faraday's Experimental Researches in Electricity. *…I perceived that his method was also a mathematical one, though not exhibited in the conventional form of mathematical symbols. I also found that these methods were capable of being expressed in the ordinary mathematical forms…."*

[8] . James Clerk Maxwell, A Treatise on Electricity and Magnetism, Third Edition, Vol. 1, Dover Publications, NY, 1954, p. 165-166.

[9] . Else the precessing electrons in a signal-receiving wire antenna would not precess laterally. Since their precession is at right angles to the disturbing force, the disturbance arriving from the vacuum was a longitudinal disturbance. This also agrees with Whittaker's 1903 decomposition of the scalar potential, and Mandl and Shaw's note that the combined scalar photon and longitudinal photon are observable as the instantaneous scalar potential,

while neither of them is observable individually. See citations for Whittaker and for Mandl and Shaw elsewhere in these references.

[10] . This word "potentiality" (by Feynman) is given flesh and blood in the present paper. What actually exists in the vacuum is potential and structures of potential. Vacuum, spacetime, and potential are identities. Hence we may say that the structured vacuum potentials are indeed structurings of spacetime curvature. We refer to these as *vacuum engines*, which are explained in this paper.

[11] . Richard P. Feynman, Robert B. Leighton and Matthew Sands, The Feynman Lectures on Physics, Addison-Wesley, New York, Vol. II, p. 1-3.

[12] . This is easily shown. $F \equiv d/dt(mv)$, which expands to $F = (m)dv/dt + v(dm/dt)$. Let the mass m be zero continuously, and both terms on the right go to zero, hence $F = 0$. That is, the force disappears because mass is a component of force. It is the use in mechanics of the separate massless force acting on a mass that is the non sequitur and to blame for this giant error in electrodynamics as a hold-over from the hoary old mechanics theory.

[13] . J. D. Jackson, Classical Electrodynamics, Second Edition, Wiley, 1975, p. 249.

[14] . Jackson, ibid. p. 28.

[15] . Melba Phillips, "Classical Electrodynamics," Vol. IV: Principles of Electrodynamics and Relativity, in Encyclopedia of Physics, Edited by S. Flugge, Springer-Verlag, Berlin, 1962, p. 4. Quoting: "The utility for electrical problems of a scalar potential, first introduced in the theory of gravitation, was pointed out by Simeon Denis Poisson, Bull. Soc.

Philomatheique **3**, 388(1813). See also Mem. Institut, Pt. **1**, 1, Pt. **2**, 163 (1811). The word *potential* was introduced by George Green in 1828."

[16] . E.g., see W. Misner, K. S. Thorne, and J. A. Wheeler, Gravitation, W. H. Freeman and Co., San Francisco, 1973, p. 5.

[17] . Two methods of dealing with this problem, and rectifying it, are the use of a topological definition of the fields in mass-free spacetime, and the use of a unified field theory where the "field in space" is not a forcefield per se, but merely a curvature or set of curvatures of spacetime. As an example, the O(3) electrodynamics largely pioneered by Evans takes the latter approach, and it is usable in directly engineering unified field theory effects.

[18] . (a). Mendel Sachs, The Field Concept in Contemporary Science, Charles C. Thomas Publishers, 1973;

(b) — General Relativity and Matter: A Spinor Field Theory from Fermis to Light-Years (Fundamental Theories of Physics), Reidel (now Kluwer), 1982 provides a great generalization of general relativity and electrodynamics reaching from the quarks and gluons to the entire universe. See also

(c) Mendel Sachs, Quantum Mechanics from General Relativity: An Approximation for a Theory of Inertia, Reidel (now Kluwer), 1986;

(d) — "Relativistic Implications in Electromagnetic Field Theory," in T. W. Barrett and D. M. Grimes, eds., Advanced Electromagnetism, World Scientific, 1995, p. 541-559. In the latter, Sachs shows that the most general expression for the field theory is in terms of spinor and quaternion variables, rather than the vector and tensor variables of the conventional expression of Maxwell's theory. This generalized expression leads to extra conservation laws and invariants, thus increasing the predictive capacity of the theory.

(e) For a marvelous treatise for the educated layman, see M. Sachs, Relativity in Our Time: From Physics to Human Relations, Taylor & Francis, 1993.

[19] . E.g., see the several Evans' reviews in Modern Nonlinear Optics, Second Edition, Edited by M.W. Evans, Wiley, 2002 (in press). Many other O(3) electrodynamics papers are also carried in this set of three volumes, as well as in the standard literature.

[20] . Mendel Sachs, "Relativistic Implications in Electromagnetic Field Theory," ibid.

[21] . Without interaction with the vacuum, mass does not exhibit perfect symmetry. For a discussion, see T. D. Lee, Particle Physics and Introduction to Field Theory, Harwood Academic Publishers, New York, p. 378-389.

[22] . In both classical and quantal electrodynamics, the problem of the source charge and its associated fields and potentials—and their energy—has been considered to be the most difficult problem.

(a) For a solution to that problem, see T. E. Bearden, "Giant Negentropy from the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23. On DoE open website http://www.ott.doe.gov/electromagnetic/papersbooks.html and this author's website www.cheniere.org.

(b) For very powerful support of that proposed solution, see F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5. Neither the time-polarized photon nor the longitudinal photon is individually observable, yet the combination of the two is observable as the instantaneous scalar potential. The Mandl and Shaw combination of a scalar (time-polarized) photon and a longitudinal photon is explained by Bearden as the interaction of the causal incoming scalar photon with a charge, the charge subsequently transducing the absorbed energy from the time-domain into 3-space, and re-emission of the energy as an emitted longitudinal photon in 3-space. The actual interaction provides the coupling, and clarifies which is the *cause* and which is the *effect*.

[23] . As the electrodynamicists implicitly assume, by assuming fields and potentials from all charges in the universe penetrating and existing and superposing and interfering in each cubic centimeter of space, and in fact at every point in space. But then the electrodynamicists continue to utilize a material ether assumption in their equations, embedded implicitly upon the notion of an "empty space" filled with this material ether. Tesla called it "*… one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history*." (Nikola Tesla, "The True Wireless," Electrical Experimenter, May 1919). While that biting remark certainly did not endear him to struggling electrical scientists, it did strike uncomfortably close to the mark, and it still does.

[24] . J. H. Poynting, "On the transfer of energy in the electromagnetic field," Phil. Trans. Roy. Soc. Lond., Vol. 175, Part II, 1885, p. 343-361.

[25] . Oliver Heaviside, "Electromagnetic Induction and Its Propagation," The Electrician, 1885, 1886, 1887, and later. This is a series of 47 sections, published section by section in numerous issues of The Electrician during 1885, 1886, and 1887. See also Oliver Heaviside, "On the Forces, Stresses, and Fluxes of Energy in the Electromagnetic Field," Phil. Trans. Roy. Soc. London, 183A, 1893, p. 423-480.

[26] . By a crude analogy, one may think of the "virtual" energy of the vacuum as energy that has been disintegrated into incredibly tiny bits—each so small that the bits fade in and out of existence at an incredible rate. The spinning action of the charge may be thought of as similar to that of an old spinning wheel. The charge spins, so that it coheres some of the "virtual fibers" continually fed to it by the vacuum, into "thread-sized" portions that are emitted radially outward. So the charge "splatters back" some of the fibers as fibers, but "spins" the remainder into useful quantum-sized "threads" which are emitted in observable form.

[27] . Presently electrodynamicists are likely to specify the Poynting energy flow **S** as $\mathbf{S} = \mathbf{E}\times\mathbf{H}$. However, in the vacuum (spacetime) neither **E** nor **H** exists as such *observed force fields*, which requires presence of charged mass. Only potentials and potential structuring of the virtual particle flux exist in the vacuum (which is after all just potential comprised of virtual particle flux). Those structurings or their analogues also exist as sets of curved spacetime regions (i.e., spacetime structurings comprising the elements of the "vacuum engine"). See E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355. Whittaker that any scalar potential decomposes into a multiplicity of longitudinal bidirectional flows of energy. This decomposition must also comprise the vacuum scalar potential, as well as the "spacetime curvature" potential.

[28] . However, Whittaker misinterpreted the *causal* phase conjugate wave (that half of his phase conjugate wavepair) as the *effect* after the wave interacts with charged mass, in which case parity is broken and the *effect* wave produced by the interaction is a longitudinal EM wave going in the opposite direction in 3-space from its "biwave" 3-space partner. The correct interpretation is that,

prior to interaction, the phase conjugate wave exists in the fourth axis of Minkowski 4-space, hence is in the time domain. It is literally a longitudinal EM wave moving on the fourth axis (in the time domain) to the interacting charge, and the other Whittaker biwave partner is the effect longitudinal EM wave emitted in 3-space by the interacting charge after it absorbs the incoming time-polarized EM wave. We point out that a longitudinal EM wave along the time axis is a time-polarized EM wave, corresponding to the known time-polarized photon in quantum field theory.

[29] . After publication of my "Giant Negentropy" paper, I discovered very strong support for my reinterpretation of Whittaker's 1903 decomposition of the scalar potential. This support was by F. Mandl and G. Shaw, ibid. Mandl and Shaw give a deeper coverage of the photon polarizations. They argue that the longitudinal and scalar polarizations individually are not directly observable, but only in combination, wherein they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. Our comment is that this argument, translated from particle terminology to wave terminology, directly fits my re-interpretation of Whittaker's 1903 decomposition of the scalar potential. However, Mandl and Shaw fail to account for their assumed interaction of the detecting/observing unit point charge to perform the perfectly correlated "combining", and thus fail to account for the absorption of the incoming *causal* time-polarized wave or photon, the transduction of that excitation energy of the charge into longitudinal EM wave/photon energy in 3-space, and the subsequent emission of that excitation energy as a perfectly correlated longitudinal EM wave in 3-space.

Thus Mandl and Shaw missed the time-excitation charging of the source charge via absorption of the "coupled" time-polarized EM wave/photon, and the subsequent decay of the time-charge excited state by emission of a 3-space longitudinal EM wave/photon. This omitted interaction represents the discovery of nature's giant negentropy mechanism in EM energy flow, whenever 3-space energy flow symmetry is broken, as with a common dipole. So Mandl and Shaw do not account for photon (or wave) polarization transduction, where the "causal" time-polarized EM wave or photon comes in and is absorbed by the detecting charge or dipole to energetically excite it in the time-domain, then the excitation energy is re-emitted as the longitudinally polarized EM wave or photon in 3-space.

Recognition of these missing functions allowed at last a solution to the long-vexing problem of the source charge. We actually solved the problem for a source dipole, then pointed out that any "isolated" charge is a set of dipoles when its clustering virtual charges of opposite sign are considered.

[30] . The source charge problem was recognized as the most difficult problem in electrodynamics. E.g., see D. K. Sen, Fields and/or Particles, Academic Press, London and New York, 1968, p. viii. Quoting: *"The connection between the field and its source has always been and still is the most difficult problem in classical and quantum electrodynamics."* Since our proposed solution is consistent with particle physics (broken 3-symmetry of the dipole) and with quantum field theory (combination of the time-polarized and longitudinal photon comprises the scalar EM potential), it is robust and we believe it will stand up to meticulous scrutiny.

[31] . E. T. Whittaker, 1903, *ibid*.

[32] . Mandl and Shaw, *ibid*.

[33] . Poynting only assumed that diverged part of the energy flow. Heaviside, on the other hand, showed that there is a much larger energy

flow component that misses the intercepting charges (or circuit) altogether, and is wasted. Since this means that far more EM energy flows from the terminals of a generator than the external circuit intercepts and catches, it means that all generators produce more energy flow output than the textbook states, or that Poynting assumed. In fact, as *transducers of vacuum energy*, all dipolar generators already exhibit COP>1.0. Since no one in the 1880s could explain where such an astounding extra energy flow was coming from, Lorentz originated an integration trick to discard the nondiverged Heaviside EM energy flow component.

(a) See J. H. Poynting, "On the transfer of energy in the electromagnetic field," Philosophical Transactions of the Royal Society of London, Vol. 175, Part II, 1885, p. 343-361.

(b) See Oliver Heaviside, "Electromagnetic Induction and Its Propagation," The Electrician, 1885, 1886, 1887, and later—a series of 47 sections, published section by section in numerous issues of The Electrician during 1885, 1886, and 1887.

(c) See also Oliver Heaviside, "On the Forces, Stresses, and Fluxes of Energy in the Electromagnetic Field," Phil. Trans. Roy. Soc. London, 183A, 1893, p. 423-480.

(d) For Lorentz' trick to discard the nondiverged component, see H. A. Lorentz, Vorlesungen über Theoretische Physik an der Universität Leiden, Vol. V, Die Maxwellsche Theorie (1900-1902), Akademische Verlagsgesellschaft M.B.H., Leipzig, 1931, "Die Energie im elektromagnetischen Feld," p. 179-186. Figure 25 on p. 185 shows the Lorentz concept of integrating the Poynting vector around a closed cylindrical surface surrounding a volumetric element. This is the procedure which arbitrarily selects only a small component of the energy flow associated with a circuit—specifically, the small Poynting component striking the surface charges and being diverged into the circuit to power it—and then treats that tiny component as the "entire" Poynting energy flow. Thereby Lorentz arbitrarily discarded all that extra Heaviside energy transport component which does not strike the circuit at all, and is just wasted.

[34] . In fact, the simple equation W = ϕq, where W is the diverged/collected energy, q is the intercepting coulomb, and ϕ is the reaction cross section of the potential, shows that electrodynamicists implicitly assume the “true” causal potential to contain unlimited energy. E.g., using that equation, any desired amount W of energy can be collected from any nonzero potential ϕ, no matter how small its intensity, simply by adding sufficient intercepting/collecting charges.

[35] . A back-of-the-envelope calculation for a nominal simple circuit has been made and previously presented by the present author. In the nominal circuit used, the source dipole extracts from the vacuum and sends down the outside of the attached circuit about 10^{13} times as much energy as the circuit itself intercepts, diverges, and utilizes. The huge remainder is not intercepted, remains nondivergent and just passes on out into space at the speed of light.

[36] . Solving the standard equation **E** = **F**/q for **F**, one has **F** = **E**q. Again, from any nonzero **E**-field, no matter how small, as much force can be “gathered” and produced as one desires, merely by increasing the intercepting/collecting charge q.

[37] . Robert H. Romer, "Heat is not a noun," American Journal of Physics, 69(2), Feb. 2001, p. 107-109. This is an editorial discussion by the Editor of AJP of the concept of heat in thermodynamics. Heat is not a substance, not a thermodynamic function of state, and should not be used as a noun. The quotation castigating the standard "field diagram" of an EM wave in space is

from endnote 24, p. 109.

[38] . This is readily understood when one realizes that the founders of electromagnetics assumed a material ether. So to them, there was not a single point in all the universe where mass was absent! Hence the field, potential, and wave "in space" were actually the field, potential, and wave "in the material ether medium", and therefore perfectly material. Even though the Michelson-Morley experiments falsified the material ether in the 1880s, the electrodynamicists (i) never changed their classical EM equations which still to this day assume the material ether filling all space, and (ii) still implicitly mistake the transverse electron precession wave in the Drude gas in a receiving antenna as representing the assumed interception of the incoming "transverse EM wave" in that external material ether medium.

[39] . Robert Bruce Lindsay and Henry Margenau, Foundations of Physics, Dover Publications, New York, 1963, p. 283-287, and passim.

[40] . Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), 1959, p. 485-491. However, Aharonov and Bohm seem to have missed the similar mass interpretation of the potential, so that they utilized the potential as massless and therefore the vacuum EM entity that is the primary cause of all EM phenomena. If one truly eliminates only the potential’s effect upon the intercepting charged mass, and utilizes the non-intercepted EM entity in vacuum, one has arrived at the direct alteration of the massless spacetime geometry of general relativity. In short, suddenly one has latched onto a unified field theory, and one that is engineerable by surprising electromagnetic means.

[41] . But erroneously assumed to be present, because of Maxwell's ubiquitous use of the material ether and the continued implied material ether in the electrodynamics equations to this day.

[42] . T. W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," Annales de la Fondation Louis de Broglie, 16(1), 1991, p. 23-41. Barrett analyzed Tesla's patented circuits in terms of quaternion electrodynamics, which is of higher symmetry than either vector or tensor algebra electrodynamics. Lo and behold, Tesla could shuttle potential energy around in his circuits at will, and that functioning would not even show up in a tensor or vector electrodynamic analysis.

[43] . T. W. Barrett, "Active Signalling Systems," U.S. Patent No. 5,486,833, Jan. 23, 1996; — Oscillator-Shuttle-Circuit (OSC) Networks for Conditioning Energy in Higher-Order Symmetry Algebraic Topological Forms and RF Phase Conjugation," U.S. Patent No. 5,493,691, Feb. 20, 1996.

[44] . As the reader can appreciate, this represents a dramatic change to the foundations of electrodynamics. It is closely allied to Sachs's unified field theory; see appropriate Sachs references in this paper.

[45] . As Dirac pointed out, the Michelson-Morley experiments of the 1880s destroyed the *material* ether assumed by Faraday and Maxwell and subsequent electrodynamicists; it *did not* destroy the ether per se. Dirac emphasized that any nonmaterial Lorentz-invariant ether is perfectly consistent with the Michelson-Morley experiments. General relativity describes the shaping and structuring of the vacuum as curvatures of spacetime geometry and clustered patterns thereof (although the latter part is essentially neglected). GR consists of the changes

of this spacetime due to the interaction of mass/ trapped energy, and also the forces upon mass/ trapped energy created by the interaction of such curvature-patterns of spacetime. Essentially, electrodynamics is missing just this "mass to spacetime" transform and this "spacetime to mass" transform. It is in fact missing its "infolded general relativity" dynamic internal structuring of the potential, field, and wave. KGB energetics weapon scientists call these missing transforms *the information content of the field*—which our fellows mistakenly interpret in their own flawed electrodynamics as "spectral content." Nothing could be further from the truth.

[46] . In 1903 E. T. Whittaker showed that any scalar potential is comprised of an infinite harmonic series of bidirectional longitudinal EM waves, arranged in bidirectional wavepairs. See E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355.

[47] . General relativity usually stops short of including an "infolded" topology within its basic spacetime, and only applies an "exterior" topology to it. However, it does not exclude that capability. Spacetime is just potential identically, and also just vacuum identically. Since GR omits the infolding structure of spacetime/potential, and so does electrodynamics, it has not been possible to unify the two completely. The "bridge" across the chasm between them has been eliminated by both disciplines.

[48] . With finite propagation, this action-at-a-distance is retarded so that some time passes between the effect and the cause. Each Mandl-Shaw photon interaction—the receipt of the scalar (time-polarized) photon, and emission of 3-space longitudinal photon—is retarded by (requires) the amount of time that is transformed into 3-space energy. [Time may be regarded as 3-space

electromagnetic energy compressed by the factor c^2, with the compressed energy placed on the fourth Minkowski axis to appear as time (seconds), rather than being left in 3-space to appear as mass).

Transformation from time-domain EM energy to 3-space energy is a function of the spin of the charge, which "spins" 720 degrees, being 360 degrees in the time domain while absorbing the time-polarized EM photon energy, the being followed by 360 degrees in 3-space while it is emitting the 3-space longitudinal photon. This retardation due to transformation of the energy from the time domain to the 3-space domain is what limits the speed of light in the ambient vacuum to the finite velocity c.

However, when infinite velocity propagation occurs (as is permissible with longitudinal EM waves alone, which a priori implies no retardation), then the notion vanishes of a “cause” and an “effect” separated by a time interval. In that case, one is operating in a multiply-connected spacetime, where in theory there is no observable “length” or “time” between the “cause” event and the “effect” event. Further, one “causative” initiation point can be superposed simultaneously with multiple distant points widely separated in ordinary spacetime. In that case, putting a joule of energy into the “initiating cause” point (to the normal 3-space observer) results in the simultaneous appearance at each and every distant participating point, of one joule. To the normal "singly-connected spacetime" observer, this is a direct amplification of energy, directly from the vacuum energy itself. We have proposed this mechanism as an extension of Bohm’s quantum potential utilized in his hidden variable theory of quantum mechanics. See David Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," Physical Review, 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II). Unfortunately three nations of the world have already seen fit to highly weaponize this effect, and a fourth nation is very close to deploying such a weapon. Presently quantum potential weapons are the dominant weapons on earth, and nothing can stand against them.

[49] . By conjugate pair, we mean that the wavepair consists of an EM longitudinal wave and its time-reversed twin, precisely superposed spatially but antiphased in time. In such a wave, the transverse energy density amplitude cannot vary; instead, the wave consists of (and carries) an oscillation of the rate of flow of time, about some median speed. The median speed of a longitudinal EM wave is not confined to the speed of light in vacuum.

[50] . The value of the energy continuously collected from a steady potential, in a "swirl" around a fixed intercepting charge, is of course a fixed scalar. That, however, has nothing to do with the nature of the potential as an EM entity in massless vacuum, prior to interaction with charged mass.

[51] . J. D. Jackson, Classical Electrodynamics, Second Edition, Wiley, 1975, p. 223.

[52] . Interestingly, one also cannot produce a clear definition of *energy* in physics. A good discussion of the energy concept is given by R. B. Lindsay, "The concept of energy and its early historical development." Foundations of Physics, 1 (4), 1971, p. 383-393. Lindsay investigates the concept of energy from its early historical origin, and covers from ancient times through the 18th century. The root of the concept is the notion of invariance in the midst of change. In short, the concept consists of considering "the changing" as being "nonchanging." In pursuing a rigorous definition of anything, one eventually arrives at a contradiction of Aristotelian logic. Ultimately one can only define something as identical to its own opposite. Geometry, e.g., models the presence of

mass as the absence of mass. Ultimately a "hard definition" is an "accursed necessity for the identity of opposites"—the centuries-old bane of the philosophers. In physics the problem is avoided by the "duality principle"—just treat an entity as either a particle or wave as necessary, and quit arguing about it! The problem is the incompleteness of Aristotelian logic itself. A solution to the "identity of opposites" problem is obtained by explicitly adding a fourth law of logic that has always been implicitly assumed in Aristotle's three laws. For a discussion, see "A Conditional Criterion For Identity, Leading to a Fourth Law of Logic," Appendix III to T. E. Bearden, Aids: Biological Warfare, Tesla Book Co., 1988, p. 428-443. A simple proof of the fourth law of logic is given in *ibid*., p. 439-440.

[53] . In addition to reference 19, see Myron W. Evans, Enigmatic Photon, Vol. II, Kluwer, ISBN 0792332881. See also Apeiron, 4(2-3), Apr.-Jul. 1997, a Special $B^{(3)}$ Field Issue, Guest editor: Myron W. Evans. In this same issue, see the Introduction by Evans, and also his "Helicity and the Electromagnetic Field." In the same issue, see also Jean-Pierre Vigier, "Relativistic Interpretation (with Non-Zero Photon Mass) of the Small Ether Drift Velocity Detected by Michelson, Morley, and Miller." Barrett has also rigorously shown the reality of the $B^{(3)}$ field, and so the conventional closure against it is doomed to failure. The emergence of $B^{(3)}$ field theory will dramatically alter electrodynamics, and elevate the importance of the longitudinal EM wave. We are also much indebted to Bob Flowers for his penetrating elucidation of the background of the $B^{(3)}$ field in private communications, and to Myron Evans for his patient and encouraging private communications.

[54] . See particularly T. W. Barrett, "Sagnac Effect: A Consequence of Conservation of Action Due to Gauge Field Global Conformal Invariance in a Multiply-Joined Topology of Coherent Fields," in T. W. Barrett and D. M Grimes, [Eds.], Advanced Electromagnetism: Foundations, Theory,

& Applications. World Scientific, River Edge, New Jersey, 07661, 1995, p. 278-313; T. W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in A. Lakhtakia, (ed.): Essays on the Formal Aspects of Electromagnetic Theory, World Scientific Publishing, River Edge, NJ, 1993, p. 6-86.

[55]. Richard W. Ziolkowski, "Exact Solutions of the Wave Equation With Complex Source Locations," Journal of Mathematical Physics, 26 (4), April 1985, p. 861-863; I.M. Besieris, A.M. Shaarawi, and R. W. Ziolkowski, "A bidirectional travelling plane wave representation of exact solutions of the scalar wave equation," Journal of Mathematical Physics, 30(6), 1989, p. 1254-1269; Rod Donnelly and Richard Ziolkowski, "A method for constructing solutions of homogeneous partial differential equations: localized waves," Proceedings of the Royal Society of London A., Vol. 437, 1992, p. 673-692; — "Electromagnetic field generated by a moving point charge: A fields-only approach," American Journal of Physics, 62 (10), Oct. 1994, p. 916-922.

[56]. M. W. Evans, editor, Modern Nonlinear Optics, Wiley, 2002, 3 volumes (in press).

[57]. W. A. Rodrigues Jr. and J.-Y. Lu, “On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 \leq v < \infty$ in nature,” Foundations of Physics, 27(3), 1997, p. 435-508; W. A. Rodrigues Jr. and J. E. Maiorino, “A unified theory for construction of arbitrary speeds ($0 \leq v \leq \infty$) solutions of the relativistic wave equations,” Random Operators and Stochastic Equations, Vol. 4, 1996, p. 355-400.

[58]. E. T. Whittaker, 1903, *ibid.*

[59]. Ironically, Maxwell’s actual theory was some 20 quaternion-like equations in 20

unknowns, as given in "A Dynamical Theory of the Electromagnetic Field," Royal Society Transactions, Vol. CLV, 1865, p 459. Read Dec. 8, 1864. Also contained in The Scientific Papers of James Clerk Maxwell, 2 vols. bound as one, edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-597. Heaviside dramatically curtailed Maxwell's theory to a vector subset of some 4 equations in 4 unknowns. Lorentz further curtailed it by effectively throwing out all Maxwell-Heaviside systems that were in disequilibrium with their active vacuum and curved spacetime environments. So after over a century, the long slumbering *environmental* dynamics for electrodynamical systems is awakening and chasing its own roots again. Long live the rise of the supersystem!

[60] . E.g., see J. Vaz Jr. and W. A. Rodrigues Jr., "Maxwell and Dirac Theories as an Already Unified Theory," Advances in Applied Clifford Algebras, Vol. 7 (S), 1997, p. 369-386. See W. A. Rodrigues, Jr. and J. Vaz Jr., "Subluminal and Superluminal Solutions in Vacuum of the Maxwell Equations and the Massless Dirac Equation," Advances in Applied Clifford Algebras, Vol. 7(S), 1997, p. 457-466.

[61] . E. T. Whittaker, 1904, *ibid*.

[62] . We strongly stress that longitudinal EM waves and their dynamics, so ignored in all except the most advanced higher symmetry electrodynamics texts, are actually the key to everything. Longitudinal EM waves already comprise, create, and structure the feeble "envelope EM waves" we presently deal with. A longitudinal EM wave is also an electrogravitational wave, and oscillates the rate of flow of time itself because of its combination with a time-polarized wave. The two combined photons have a total spin two, which is a graviton. Thus internally a graviton is revealed as a particular

dynamic between time-domain EM energy and 3-space EM energy. The longitudinal EM wave in the time domain (that is combining with the longitudinal EM wave in 3-space) is thus an oscillation of time contraction and expansion, or "pulsation" in the rate of flow of time. Very powerful gravitational effects as well as very powerful and unique electrodynamics effects can be obtained by use of longitudinal EM waves, because of the additional dynamics between the time and 3-space domains. Unwitting use of such extraordinary EM waves is how Rife and Prioré achieved such remarkable results.

[63] . M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," Journal of New Energy, 4 (3), Special Issue, Winter 1999, p. 76-78.

[64] . Melba Phillips, “Classical Electrodynamics,” Vol. IV: Principles of Electrodynamics and Relativity, in Encyclopedia of Physics, Edited by S. Flugge, Springer-Verlag, Berlin, 1962, p. 1-108 gives a good overall summary of the superpotential theory.

[65] . A. Righi, Nuovo Cimento, Vol. 2, No. 5, 1901, p. 104.

[66] . P. Debye, Ann. Phys., Leipzig, Vol. 30, 1909, p. 57.

[67] . Thomas John I'Anson Bromwich, An Introduction to the Theory of Infinite Series, MacMillan, London, 1908.

[68] . A. Nisbet, Physica, Vol. 21, 1955, p. 799; Proc. Roy. Soc. Lond. A, Vol. 231, 1955, p. 250.

[69] . W. H. McCrea, “Hertzian electromagnetic

potentials," Proc. Roy. Soc. Lond. A, Vol. 240, 1957, p. 447-457.

[70] . Sachs has achieved the necessary unified field theory, broadly unifying physics from the quarks and gluons to the entire universe, and including both the active vacuum and active, locally curved spacetime. See Mendel Sachs, General Relativity and Matter: A Spinor Field Theory from Fermis to Light-Years (Fundamental Theories of Physics), Reidel (now Kluwer), 1982; — Quantum Mechanics from General Relativity: An Approximation for a Theory of Inertia, Reidel (now Kluwer), 1986. It should be obvious that, unless electrodynamics theory and general relativity theory utilize the same background spacetime, it is impossible to completely merge them, from first principles. Since classical electrodynamics artificially discards its active environment, it is electrodynamics which must be changed in order to be successfully united with general relativity. Sachs has accomplished this long-sought *engineerable* unification.

[71] . Proof that the virtual state can be coherently integrated is already furnished by every source charge in the universe, which does precisely that.

[72] . We note without elaboration that all the mind functions, even of the deepest and most unconscious nature, and even of collective species unconsciousness domains—exist electromagnetically in the infolded EM domain. The "conscious mind" is a serial processor, though extremely rapid. The unconscious is a massively parallel processor. The reason the conscious mind cannot "directly see" the unconscious is that "seeing" innumerable things in the "single snapshot" at once, just obliterates all singular discrimination because of the intense summation. Mind functions are timelike, so consist of specialized structures, functioning, and dynamics

inside the time domain. Since the infolded structuring captures this entire domain, it also involves a physics of mind and matter interaction, but as a science instead of as a mysticism. In theory vacuum engines are also possible that engineer the mind at any and all levels. We have addressed this extended physics of mind and matter reactions elsewhere and do not address it here. The Russian weapon scientists refer to this mind area (which can also be engineered by novel electrodynamic means) as psychoenergetics.

[73] . R. O. Becker and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, Vol. unk, No. 73, Nov.-Dec. 1970, p. 169-198. This is the definitive technical exposition by Becker of his electrical findings about the cellular control system.

[74] . See particularly Robert O. Becker, “The direct current control system. A link between environment and organism,” New York State Journal of Medicine, Vol. 62, 1962, p. 1169-1176; — and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bulletin of the New York Academy of Medicine, Second Series, 48(4), May 1972, p. 627-641; — and Carlton F. Hazlewood, Abraham R. Liboff, and Jan Walleczek, Electromagnetic Applications In Medicine," NIH-OAM Electromagnetics Panel Report, Jan. 15, 1993; — "A technique for producing regenerative healing in humans," Frontier Perspectives, 1(2), Fall/Winter 1990, p. 1-2; — and Andrew A. Marino. Electromagnetism and Life, State University of New York Press, Albany, NY, 1982; — and Gary Selden, The Body Electric, William Morrow and Company, New York, 1985; — Cross Currents, Jeremy P. Tarcher, Inc., Los Angeles, 1990; — "The bioelectric field pattern in the salamander and its stimulation by an electronic analog," IRE Transactions on Medical Electronics, Vol. ME-7, July 1960, p. 202-207; —

"Some observations indicating the possibility of longitudinal charge-carrier flow in peripheral nerves," Biological Prototypes and Synthetic Systems, Plenum Press, New York, 1962, p. 31-37; — and W. Slaughter, "The longitudinal direct current gradients of spinal nerves," Nature, Vol. 196, Nov. 17, 1962, p. 675-676; — and Charles H. Bachman and Howard Friedman, The direct current (control) system: A link between the environment and the organism," New York State Journal of Medicine, Vol. 62, April 15, 1962, p. 1169-1176; — "The direct current field: A primitive control and communication system related to growth processes," Proceedings of the. XVI International. Congress of Zoology, Washington, D.C., Vol. 3, 1963, p. 179-183; — "The neutral semiconduction control system and its interaction with applied electrical current and magnetic fields," Proc. 11th International. Cong. Radiol., Amsterdam, Excerpta Medica Foundation, 1966, Series 105, p, 1753-1759; — "The bioelectric factors in amphibian limb regeneration," The Journal of Bone and Joint Surgery, Vol. 43-A, No. 5, July 1961, p. 643-656; — and David G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents," Transactions, New York Academy of Sciences, 29(5), Mar. 1967, p. 606-615.

[75] . The ultimate tool for use in seeing and working with this "infolded" or "internal electrodynamics inside normal EM fields, waves, and potentials will probably be the Fogal semiconductor when it goes into production in the near future. Fogal's transistors are capable of (1) infolding the desired longitudinal EM waves inside normal EM signals (he demonstrates infolding TV signals inside pure DC potentials, and transmitting them over conductors as pure DC potentials for appreciable distance, then using his semiconductors to again extract the "infolded" hidden TV signals.

[76] . See Vlail Kaznacheyev, "Electromagnetic

Bioinformation in Intercellular Interactions," Psi Research, 1(1), Mar. 1982, p. 47-76; — *et al*., "Distant intercellular interactions in a system of two tissue cultures," Psychoenergetic Systems, 1 (3), Mar. 1976, p. 141-142; "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144, — *et al*., "Apparent information transfer between two groups of cells," Psychoenergetic Systems, 1(1), Dec. 1974. See particularly Vlail P. Kaznacheyev and L. P. Mikhailova, Ultraweak Radiations in Intercellular Interactions [in Russian], Novosibirsk, 1981; "Electromagnetic Bioinformation in Intercellular Interactions," Novosibirsk, 1981 [in Russian]; V. P. Kaznacheyev, "Information Function of Ultraweak Light flows in Biological Systems," Problems in Biophysics, Novosibirsk, 1967, p. 7-18 [in Russian].

[77] . Some years ago we documented this disease-inducing "microwave radiation of the U.S. Embassy, explained its purpose and operational mechanism, and published it. E.g., see T. E. Bearden, AIDS: Biological Warfare, Tesla Book Co., Chula Vista, CA, 1988.

[78] . See Robert C. Mallieu, A Model of the Microwave Intensity Distribution Within the U.S. Embassy, Moscow 1966-1967, Report FS-80-166, Applied Physics Laboratory, Johns Hopkins University, Baltimore, Maryland, Aug. 1980. Also see Microwave Radiation at the U.S. Embassy in Moscow and Its Biological Implications: An Assessment, U.S. Department of Commerce, Doc. # NTIA-SP081-12, Mar. 1981. Also see Captain John D. LaMothe, Controlled Offensive Behavior—USSR, Report ST-CS-01-169-72, Defense Intelligence Agency, Washington, D.C. (released under FOIA). See also Captain John D. LaMothe and Mr. Louis Maire, Soviet and Czechoslovakian Parapsychology Research, Report DST-1810S-387-75, Defense Intelligence Agency, Washington, D.C. (Released under FOIA).

[79] . T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," Journal of New Energy, 3(2/3), 1998, p. 12-28. A slightly updated and extended version is published in Explore, 8(6), 1998, p. 7-16.

[80] . C. W. Misner, K. S. Thorne and J. A. Wheeler, Gravitation, W.H. Freeman and Co., San Francisco, 1973, p. 5, 19-21, 71-72, 163-165, 390-391, 399; 367-369.

[81] . T. D. Lee, Particle Physics and Introduction to Field Theory, Harwood, New York, 1981. See particularly Chapter 25: Outlook, "Possibility of Vacuum Engineering," p. 824-828.

[82] . David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," Physical Review, 85 (2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II).

[83] . I.e., the rate of flow of time, or more specifically, the rate of the EM interactions with a mass or particle that generate the flow of the mass or particle through time. See Figure 25. Nobelist Lee has also shown that time is a discrete dynamical variable across all of physics, from quantum mechanics to nonrelativistic field theory, to relativistic quantum field theory. See T. D. Lee, "Can Time Be a Discrete Dynamical Variable?", Physics Letters, 122B(3, 4), Mar. 10, 1983, p. 217-220. Also in T. D. Lee, Selected Papers, Gerald Feinberg, Ed., Birkhauser, Boston, 1986, Vol. 3, p. 77-80. So we have not hesitated to consider time as any other variable, and the fourth Minkowski axis as any of the other three axes with respect to EM energy, fields, potentials, and waves.

[84] . Indeed, there is no such unique thing as “the gravitational field” anyway! The gravitational field is a conglomerate of other field energies and

functions, not a separate field of nature. See A. A. Logunov and Yu. M. Loskutov, "Nonuniqueness of the predictions of the general theory of relativity," Sov. J. Part. Nucl., 18(3), May-June 1987, p. 179. See C. W. Misner, K. S. Thorne and J. A. Wheeler, Gravitation, W.H. Freeman and Co., San Francisco, 1973, p. 399. Quoting*: "The terms 'gravitational field' and 'gravity' refer in a vague, collective sort of way to all of these entities. Another, equivalent term for them is the 'geometry of spacetime'.“* See also A. D. Sakharov, "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," Soviet Physics Doklady, Vol. 12, No. 11, 1968, p. 1040-1041 [English translation.]. Finally, see our remarks under reference 62 above, particularly with respect to the proposed internal electromagnetic dynamics of the graviton.

[85] . A 1993 study by the Office of Technology Assessment, delivered to Congress, showed that a single terrorist in a small plane, with a tank of anthrax spray, and dispersing the spray as he flew around the greater Washington D.C. area on a calm night, would produce between 1 and 3 million casualties. Most of those casualties would die. So a single biological warfare (BW) strike in one U.S. metropolitan area could easily kill two million Americans. As an example, Saddam Hussein is known to have the terrorist teams and the anthrax, and he has had a decade since the Gulf war to insert additional teams and biological warfare agents into the U.S. Presently this is rated by our government as the strongest strategic threat to the U.S., after which comes nuclear strikes, followed by chemical agent strikes. Redevelopment of the Prioré approach, e.g., could easily save the lives of 90% of that nominal two million dead Americans. As matters stand now, triage would apply and almost all of those two million stricken Americans would die.

[86] . In 1998 we proposed a crash development of this kind of device in 1998, with copies to DoD,

NIH, CDC, the USAF, etc.; in T. E. Bearden, Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ page inclosure to CTEC Letter, “Saving the Lives of Mass BW Casualties from Terrorist BW Strikes on U.S. Population Centers,” to Major General Thomas H. Neary, Director of Nuclear and Counterproliferation, Office of the Deputy Chief of Staff, Air and Space Operations, HQ USAF, May. 4, 1998. Also to Gen. (Ret.) Walter Busby, Deputy Secretary of Defense for Counterproliferation and Chemical and Biological Defense, March 21, 1998. Also to NIH, CDC, and several other U.S. government agencies.

DoD replied that it was not documented in the literature (a specific list of exact scientific papers reporting the results of the Prioré team's experiments and the experiments of Becker was included; no one seemed to read it). CDC and the USAF did not reply. NIH first responded (from its Policy section—read, *spin control* section) that, since I mentioned the cause of the Gulf War Disease and how it was induced, it was not a matter of their interest but of DoD's interest, and so they just shipped it to DoD and got rid of it. A second response evoked from NIH on query from my congressman (again, a response from their *policy* section) directed me to their proposals section. I don't believe a single qualified scientist even read it. DoD weakly replied that it was not documented in the literature (a specific list of citations from the hard French scientific literature, reporting the astounding results of the Prioré treatments, was included, which apparently no one read).

Not a single scientist from any agency called me to discuss the revolutionary and proven results obtained by Prioré, and documented in the attachments and in the cited literature. Sadly, it seems that no one had the foggiest notion what I was talking about, and no one much cared. My conclusion was and is that my own government bureaucrats do not give a tinker's dam about saving millions of American casualties in the coming debacle, *if it has to be done by "out-of-the-box" methods not approved by the big drug industry*. Yet these are the same officials that readily admit the exact threat I referred to, admit that it is a matter of *when* the biological warfare attacks will occur and not *if*, and admit that right now they have no effective way to cope with it or save most of those coming millions of Americans that are going to just be dragged aside to die.

My response is that the American taxpayers are certainly not getting their money's worth from the billions they shower on our medical research community. It is unconscionable to ignore for over 30 years the demonstrated and astounding results achieved by eminent French scientists—such as Pautrizel, a world-renowned parasitologist and Courier, head of the French Academy's Biological Section, and also *Secretaire Perpetuel* of the Academy) working with the Prioré team, in curing terminal tumors etc. in thousands of lab animals, all properly documented in the hard French scientific literature.

In my view, considering the certainty of those future American mass deaths, it is akin to high treason. Whether our scientific community realizes it or not, today's strategic war of massive strategic strikes has shifted to our civilian populace, and to our cities and mass population centers. It is our civilian populace that will sustain the greatest and most prompt casualties, and probably will receive the first massive strikes, totally by surprise.

The first phase of WW III—delivery of the strategic weapons to their strategic targets—has already been accomplished. Large Russian nuclear weapons are already hidden in our major cities and population centers (read Lunev's book, he tells us how they brought in these weapons), the Spetznaz teams are here to detonate them on call, multiple nations such as Iraq, Libya, and Iran already have terrorists agents in country in the U.S. with anthrax and smallpox in hand to be released, and Castro guerrillas and trained saboteurs have been trained in Southern Mexico and infiltrated across the border into the U.S. for decades.

Whether they like it or not, U.S. scientists have thus become the major front line "soldiers" and "defenders" of our nation. And frankly, they have failed abominably in their responsibility, and most do not seem to care. For a startling and even shocking characterization of our present scientific community, the reader is referred to Daniel S. Greenberg, Science, Money, and Politics: Political Triumph and Ethical Erosion, University of Chicago Press, 2001.

[87] . As we stated, just such an attack is now officially the greatest strategic threat to the United States. Second is a nuclear attack, followed by chemical attack as a distant third. E.g., see Richard K. Betts, “The New Threat of Mass Destruction,” Foreign Affairs, 77(1), Jan./Feb. 1998, p. 26-41. For the ineffectiveness of our defenses against the coming mass biological warfare strikes, see Laurie Garrett, The Coming Plague: Newly Emerging Diseases in a World Out of Balance, Farrar, Straus and Giroux, New York, 1994; — "The Nightmare of Bioterrorism," Foreign Affairs, 80(1), Jan./Feb. 2001, p. 76-89; — Betrayal of Trust: The Collapse of Global Health, Hyperion, New York, 2000.

[88] . Becker, *ibid*.

[89] . A full account of the entire Prioré affair is given in Jean-Michel Graille, Dossier Prioré: Une Nouvelle Affaire Pasteur [The Prioré Dossier: A New Pasteur Affair?], De Noel, Paris, 1984. [in French]. A summary account is given by Christopher Bird, "The Case of Antoine Prioré and His Therapeutic Machine: A Scandal in the Politics of Science," Appendix I in T.E. Bearden, AIDS: Biological Warfare, Tesla Book Co., P.O. Box 121873, Chula Vista, California 91912, 1988. Prioré's doctoral thesis (in my personal possession courtesy of Professor Courrier and the late Chris Bird), was rejected by the University of Bordeaux

when the French government suppressed the Prioré program in the mid-70s; it is A. Prioré, Guérison de la Trypanosomiase Expérimentale Aiguë et Chronique par L'action Combinée de Champs Magnétiques et D'Ondes Electromagnétiques Modulés. [Healing of intense and chronic experimental trypanosomiasis by the combined action of magnetic fields and modulated electromagnetic waves], 1973. Prioré's patents are: A. Priore, "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968; — "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; — "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963. Eleven years after Priore's thesis was rejected, the same university awarded a doctoral thesis to a young graduate. It is Eric Perisse, Effets des Ondes Electromagnètiques et des Champs Magnètiques sur le Cancer et la Trypanosomiase Experimentale [Effects of Electromagnetic Waves and Magnetic Fields on Cancer and Experimental Trypanosomias], Doctoral thesis, University of Bordeaux No. 83, March 16, 1984.

[90] . F. A. Popp *et al.*, "Biophoton Emission (Multi-Author Review)," Experientia, Vol. 44, 1988, p. 543-600; — "Biophoton emission: New evidence for coherence and DNA as a source," Cell. Biophys., Vol. 6, 1984, p. 33-52; — "Physical aspects of biophotons." Experientia, Vol. 44, 1988, p. 576-585; — Recent Advances in Biophoton Research and its Applications, World Scientific Publishing Co., New York, NY, 1992; F. A. Popp, "Photon Storage in Biological Systems," in Fritz Albert Popp *et al.*, eds., Electromagnetic Bio-Information: Proceedings of the Symposium,

Marburg, Federal Republic of Germany, Sep. 5, 1977; Urban & Schwarzenberg, Baltimore, 1979, p. 123-149.

[91] . David M. Rorvik, "Do the French have a Cure for Cancer?", Esquire Magazine, July 1975, p. 110-111, 142-149.

[92] . Arguably the best introduction to nonlinear optical phase conjugation is David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(2), March/April 1982, p. 156-183. Another very good presentation is Amnon Yariv, Optical Electronics, 3rd Edition, Holt, Rinehart, and Winston, New York. 1985, Chapter 16: “Phase Conjugate Optics – Theory and Applications.”

[93] . A beautiful example of this “precise backtracking” of the time-reversed wave is given in David M. Pepper, "Applications of Optical Phase Conjugation," Scientific American, 254(1), Jan. 1986, p. 74-83. See particularly the striking photographic demonstration of time reversal of disorder on p. 75.

[94] . We presented some general elements of this dramatic new extension to nonlinear optical phase conjugate pumping and to general relativity, at a graduate seminar, “On the Theory and Principles of EM Systems with COP>1.0”, given to interested graduates and faculty of the University of Louisville, on Nov. 20, 1997.

[95] . E.g., see Owen Flynn, "Parametric arrays: A new concept for sonar," Electronic Warfare Magazine, June 1977, p. 107-112.

[96] . This method of increasing signal-to-noise ratio and pulling hidden signals right up out of the noise, is widely used in signal processing in modern radars. Often signals buried more than 100

dB below the noise level are routinely "pulled out of the noise" and detected, using various forms of coherent summation.

[97] . Any potential decomposes into multiple wavepair sets. In each hidden wavepair, each of the two waves further decomposes into two potentials. Each of *those* potentials then further decomposes into multiple wavepair sets. And so on *ad infinitum*. Or one can start with any wave, decompose it into two potentials, each of those potentials into multiple biwave sets, and so on.

[98] . "Fractal" means self-similarity at all levels, so that patterning or "decomposition" at one level is similar to that at the higher and lower levels. For good overview discussions and applications, see particularly Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," and Joseph Ford, "What is chaos, that we should be mindful of it?", both in The New Physics, Paul Davies, editor, Cambridge University Press, Cambridge and New York, 1989, p. 316-347 and 348-371 respectively. More detailed technical references are cited in each article.

[99] . See Trevor James Constable, "You, Too, Can Photograph UFOs from Airliners," Psychic Observer and Chimes, 37(3), May-June 1977; — The Cosmic Pulse of Life, Merlin Press, Santa Ana, CA, 1976.

[100] . The present author has two beautiful ultraphotographs taken by Joe Gambill using the special filtering approach. One is a photograph of a room, but with its contents of *two weeks earlier*, showing the room prior to the time the photograph was taken. The other is of a "ghost" form—very similar to a photograph obtained by George Meek under rigorous laboratory conditions—and clearly showing the energy form of a "haunted" lady's dead husband. The *interpretation* of these photographs

is of course open to scientific dispute, but the *photographs themselves* are impeccable and real.

[101] . Meek used filters open to UV and IR but opaque to the visible spectrum, on the fluorescent lights in his laboratory. He also used similar filtering on the camera lenses and camera flashes. Under rigorous laboratory conditions, in several thousand photos of the living body Meek was able to take a few highly anomalous photos completely inexplicable by normal science. E.g., he photographed a purported "ghost" and obtained a clear shot of energy streaming associated with a living person. He also photographed the legs of a chair, right through the suddenly "missing" legs of the person sitting in the chair. Again, interpretation is open to dispute, but the photographs are rigorous and real.

[102] . Bill Fogal's patented charge-blocking semiconductor—U.S. Patent Numbers 5,196,809 and 5,430,413—is capable under certain circumstances—and by clever adaptation—of amplifying virtual state information, in a manner somewhat similar to Rife's methodology, but far more easily engineered. When the process works, Bill reports that a clear photograph or video camera shot can be taken right through a solid brick wall, of the hidden scenery beyond. In electrodynamics, in a dielectric absorbing or emitting EM radiation (energy), the entire dielectric participates in each absorption and emission. Emitted light from an object thus carries an astoundingly complete set of little virtual state vacuum engines rigorously representing every part of the object, no matter how small, and all the internal dynamics. This means that the light from the scene beyond the wall strikes the wall on the other side, affecting all the material in the wall, including upon the observer's side, and inserting "dimensioning" from the scene beyond the wall. When light is emitted from the observer's side, the information content of the field representing the scene on the far side, is also part of

the information content of the field in the light emitted from the observer side of the wall and striking the camera and the observers eyes. This “dimensioning” continues in the signals in the camera and on through attached electronic circuitry. Since Fogal’s semiconductor can process such information content of the field in the proper circuitry and under favorable circumstances, the phenomenon is therefore permissible and can be technically explained. Using his transistor to transduce the dimensioned information into the visual band, Fogal *can* view a distant scene right through an intervening wall, when he gets everything “just right.” At this stage of development, however, the process and adjustment are difficult, just as Rife experience with his microscope. We are convinced that the forthcoming advent of Fogal semiconductors will revolutionize the study and treatment of disease and bodily disorders.

[103] . However, highly advanced electrodynamics does consider the photon with far greater complexity than nuclear engineers such as the present author were introduced to! E.g., a deep discussion of photon polarization is given in F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984 under the heading "covariant quantization of the photon propagator" in Chapter 5. As pointed out by Bob Flowers in a private communication, Mandl and Shaw argue that the longitudinal and scalar polarizations are not directly observable, but only in combination, where the combination manifests as the "instantaneous" Coulomb (i.e., electrostatic) potential. Most such books do argue that the extra components of the photon remain hidden, as does the virtual state. In this paper we have revealed how Rife discovered a process by means of which the virtual state can be observed in great detail.

[104] . V. P. Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," Psi

Research, 1(1), Mar. 1982, p. 47-76; — "Information Function of Ultraweak Light Flows in Biological Systems," in Problems in Biophysics, Novosibirsk, 1967, p. 7-18 [in Russian]; — and L. P. Mikhailova, Ultraweak Radiation in Intercellular Interactions, [in Russian], Novosibirsk, 1981; — *et al.*, "Apparent information transfer between two groups of cells," Psychoenergetic Systems, 1(3), Mar. 1976; — *et al.*, "Distant intercellular interactions in a system of two tissue cultures," Psychoenergetic Systems, 1(3), Mar. 1976, p. 141-142; "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144.

[105] . Mitogenetic radiation was studied prior to WW II by Gurwich (spelling varies). See A. Gurwitsch and L. Gurwitsch, "L'Analyse mitogenetique spectrale [The Analysis of the mitogenetic spectrum]," in Andre Mayer, Exposés de Physiologie, [Studies in Physiology], Hermann & Cie, Paris, 1934. Part IV. [In French]. Studies have continued by Popp and others. E.g., see F. A. Popp *et al.*, "Biophoton Emission (Multi-Author Review)," Experientia, Vol. 44, 1988, p. 54 3-600; — "Physical aspects of biophotons." Experientia, Vol. 44, 1988, p. 576-585; — "Biophoton emission: New evidence for coherence and DNA as a source," Cell. Biophys., Vol. 6, 1984, p. 33-52. See Fritz Albert Popp; Ke-hsueh Li, and Qiao Gu, [eds.], Recent Advances in Biophoton Research and its Applications, World Scientific Publishing Co., New York, NY, 1992.

[106] . The conventional quantum mechanics assumes that, at the microscopic level, variation (e.g., in the probabilities assigned to interpret what is propagating in the Schrödinger equation) is assumed to be random. However, there is a problem. Big things must be made of little things. If we sum a set of "random" things based on Gibbs thermodynamics statistics, we cannot and will not get an organized "big thing" because we have

assumed total disorder (randomness) *a priori*. Adding heat does not produce cooling! So one falsifies the prevailing interpretation of quantum mechanics, whenever one observes a tree, a flower, the sun, or a human being. In short, there must exist a great unifying principle in nature of a coherent summation, from the "unobservable" virtual state to the observable state. Only in that fashion can there be an macroscopic order at all—and such macroscopic order is universally observed. We believe that the principle we have elucidated, as experimentally produced by Rife, may well be that long-sought universal ordering principle that is so sorely needed by quantum mechanics. Readers interested in the "missing quantum chaos" problem are referred to P. V. Elyutin, "The Quantum Chaos Problem," Sov. Phys. Usp., Vol. 31, No. 7, 1988, p. 597-622. Elyutin discusses the crisis in quantum mechanics because of the missing hidden order. Quantum mechanics is known to be wrong unless this order can be found, because otherwise it does not predict the ordered macroscopic universe. Other useful references are Ilya Prigogine, with T. Petrosky, "Quantum Chaos, Complex Spectral Representation and Time-Symmetry Breaking," Chaos, Solitons, and Fractals, Vol. 4, 1994, p. 311-359; and Robert Pool, "Quantum Chaos: Enigma Wrapped in a Mystery," Science, 243(4893), 1989, p. 893-895.

Slide Index

Figure 1. The Rife microscope.

Figure 2. Royal Raymond Rife adjusts his microscope.

Figure 3. Faraday omitted the string holder for his taut string, thus omitting half the electrodynamics and half the energy. An equal-and-opposite Newtonian third law reaction wave always appears in our transmitters and receivers, caused by the omitted wave.

Figure 4. Drude gas gyroelectrons in the receiving antenna precess laterally, providing the detected transverse wave and proving the vacuum disturbance wave to be longitudinal.

Figure 5. The magnitude of the causal potential in massfree space is erroneously defined.

Figure 6. Fields and potentials are erroneously defined as what is diverted from them.

Figure 7. The defined magnitude of the field is erroneous and is the field intensity.

Figure 8. Matter-to-matter transform.

Figure 9. The missing matter-to-spacetime and spacetime-to-matter transforms.

Figure 10. Vacuum engine utilization.

Figure 11. A charge is continuously in violent energy exchange with the active vacuum.

Figure 12. Whittaker's biwave decomposition of the scalar potential, and its reinterpretation.

Figure 13. The Supersystem and its components.

Figure 14. Vacuum engine (cartoon representation).

Figure 15. First vacuum engine analogy of arm muscle and bone.

Figure 16. Second vacuum engine analogy of arm bone and muscle.

Figure 17. Third vacuum engine analogy.

Figure 18. Fourth vacuum engine analogy.

Figure 19. Time-like flow of EM energy into 3-space.

Figure 20. Block diagram of the immune system.

Figure 21. Becker's electromagnetic healing of otherwise intractable bone fractures.

Figure 22. Becker's block diagram of the cellular control system.

Figure 23. Kaznacheyev's electromagnetic induction of cellular death and disease, at a distance.

Figure 24. Microwave radiation of the U.S. Embassy in Moscow.

Figure 25. Mechanism generating the flow of a mass through time.

Figure 26. Vacuum antiengine for a specific disease condition.

Figure 27. Pumping with transverse and longitudinal EM waves.

Figure 28. Block diagram of the Prioré process.

Figure 29. Block diagram of the operation of the body's cellular regenerative system.

Figure 30. Rife's great secret: Recursive iterative serial amplification by a single harmonic interval difference frequency.

Help support this work

Rife's Microscope

Table of Photographs

(Thanks to the donor of these photographs, who wishes to remain anonymous)

1. Rife's Universal Microscope
2. Eyepiece Details
3. Chart for Calculating Parallax of Lens System
4. Diagram - Wide Angle Apochromatic Microscope Objective Lens
5. Universal Focus - Photographic Lens
6. Light Passage of Dove Prism

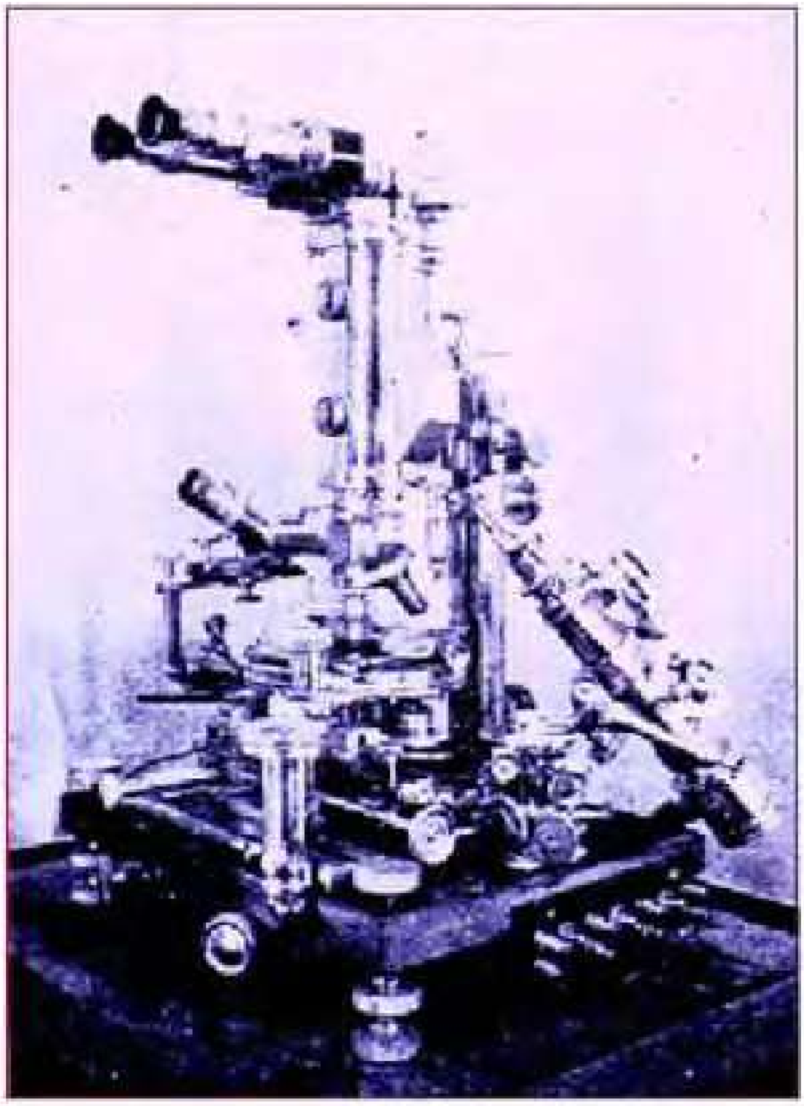

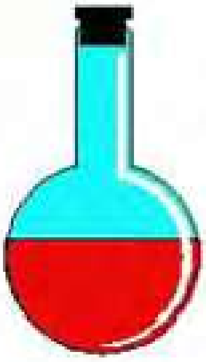

Figure 1. Rife's microscope.

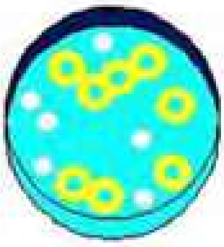

Figure 2. Royal Raymond Rife adjusts his microscope.

Air medium disturbances generated when air is perturbed by a plucked taut string

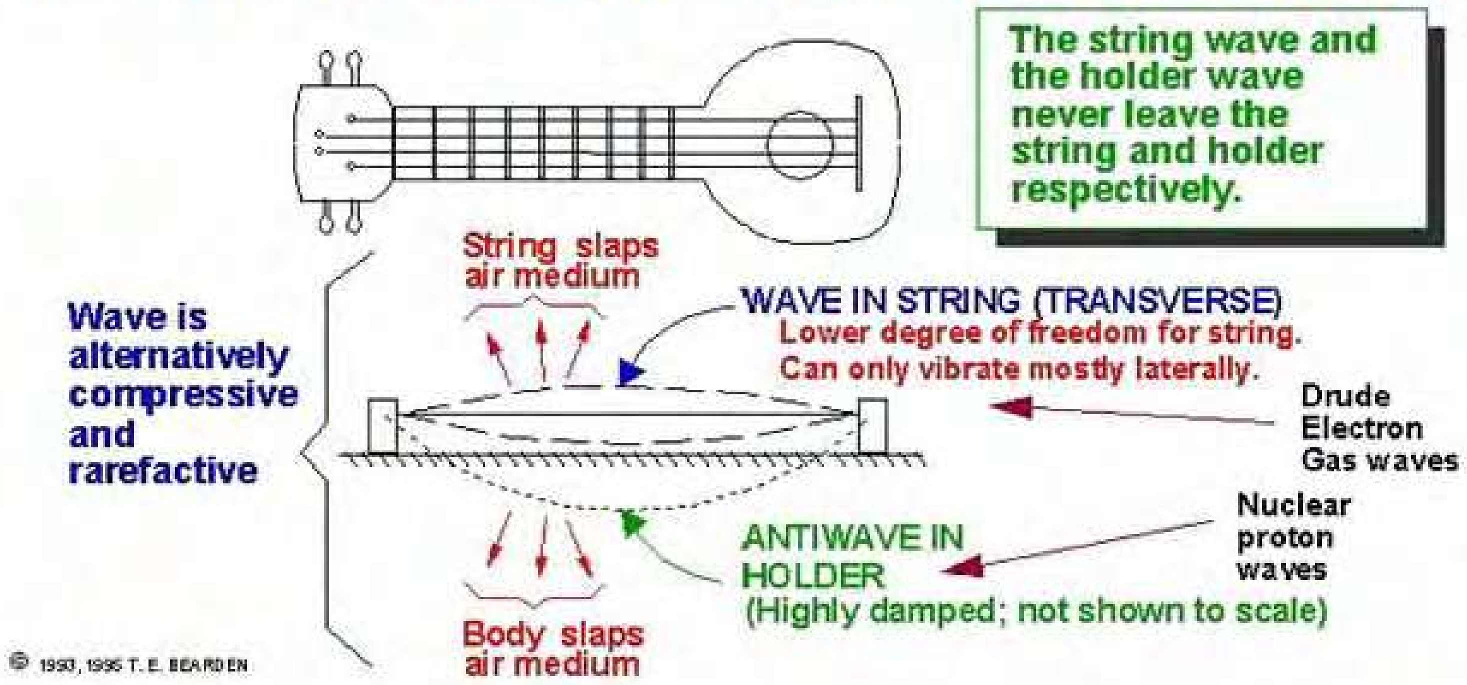
String slaps air medium
WAVE IN STRING (TRANSVERSE)
Lower degree of freedom for string.
Can only vibrate mostly laterally.
Wave is alternatively compressive and rarefactive
Drude Electron Gas waves
Nuclear proton waves
ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)
Body slaps air medium

The string wave and the holder wave never leave the string and holder respectively.

© 1993, 1996 T. E. BEARDEN

Figure 3. Analogy of the "taut string holder" omitted by Faraday. This omitted half the electrodynamics (including in Maxwell's capture of Faraday's notion) and half the EM energy. An equal and opposite Newtonian 3rd law reaction does appear in our transmitters and receivers.

Flow of current and energy into, through, and along a conductor

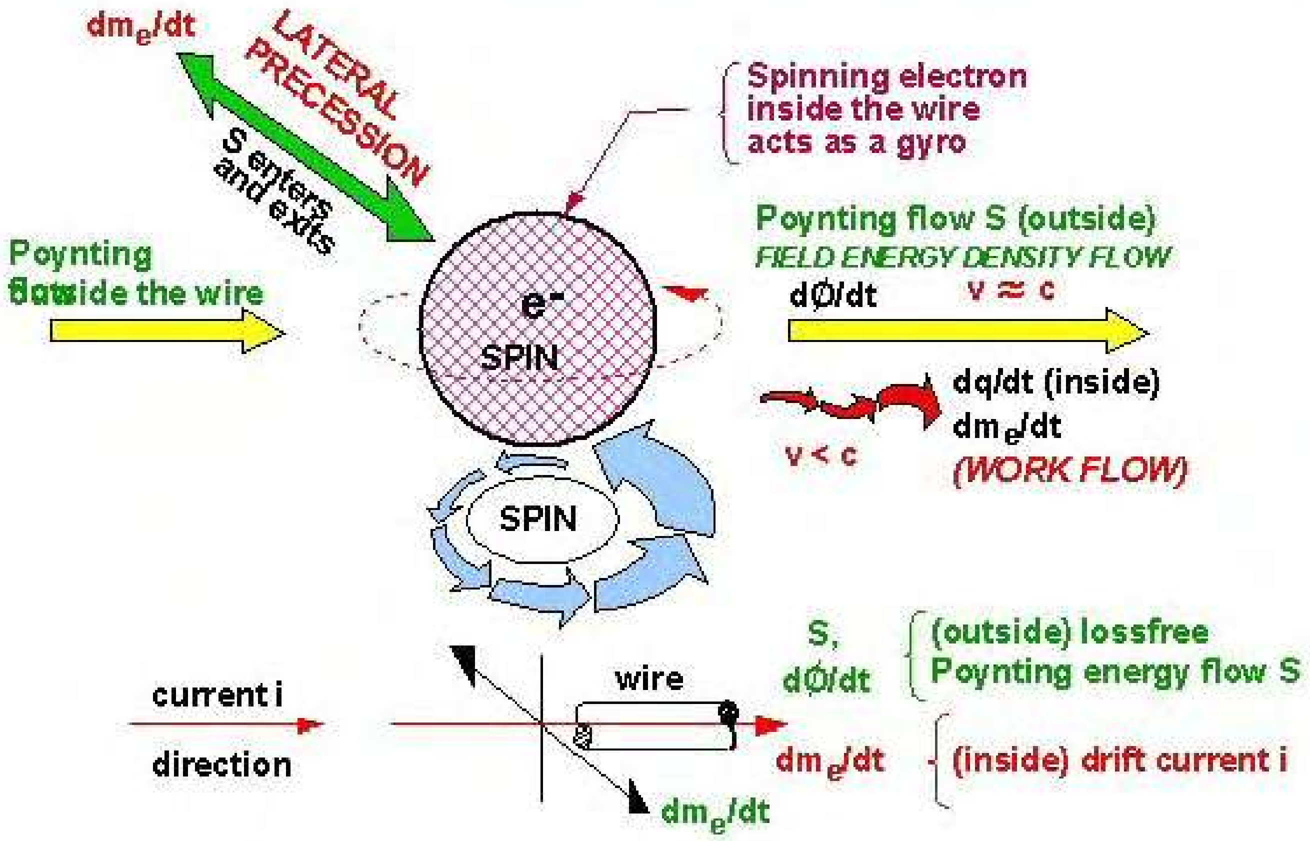
dm$_e$/dt
LATERAL PRECESSION
S enters and exits
Spinning electron inside the wire acts as a gyro
Poynting flow outside the wire
e⁻
SPIN
Poynting flow S (outside)
FIELD ENERGY DENSITY FLOW
dØ/dt
v ≈ c
dq/dt (inside)
dm$_e$/dt
(WORK FLOW)
v < c
SPIN
S,
dØ/dt
(outside) lossfree
Poynting energy flow S
current i
direction
wire
dm$_e$/dt
(inside) drift current i
dm$_e$/dt

Figure 4. Receiving gyroelectrons precess laterally, providing the detected transverse wave and proving the incoming vacuum disturbance to be longitudinal.

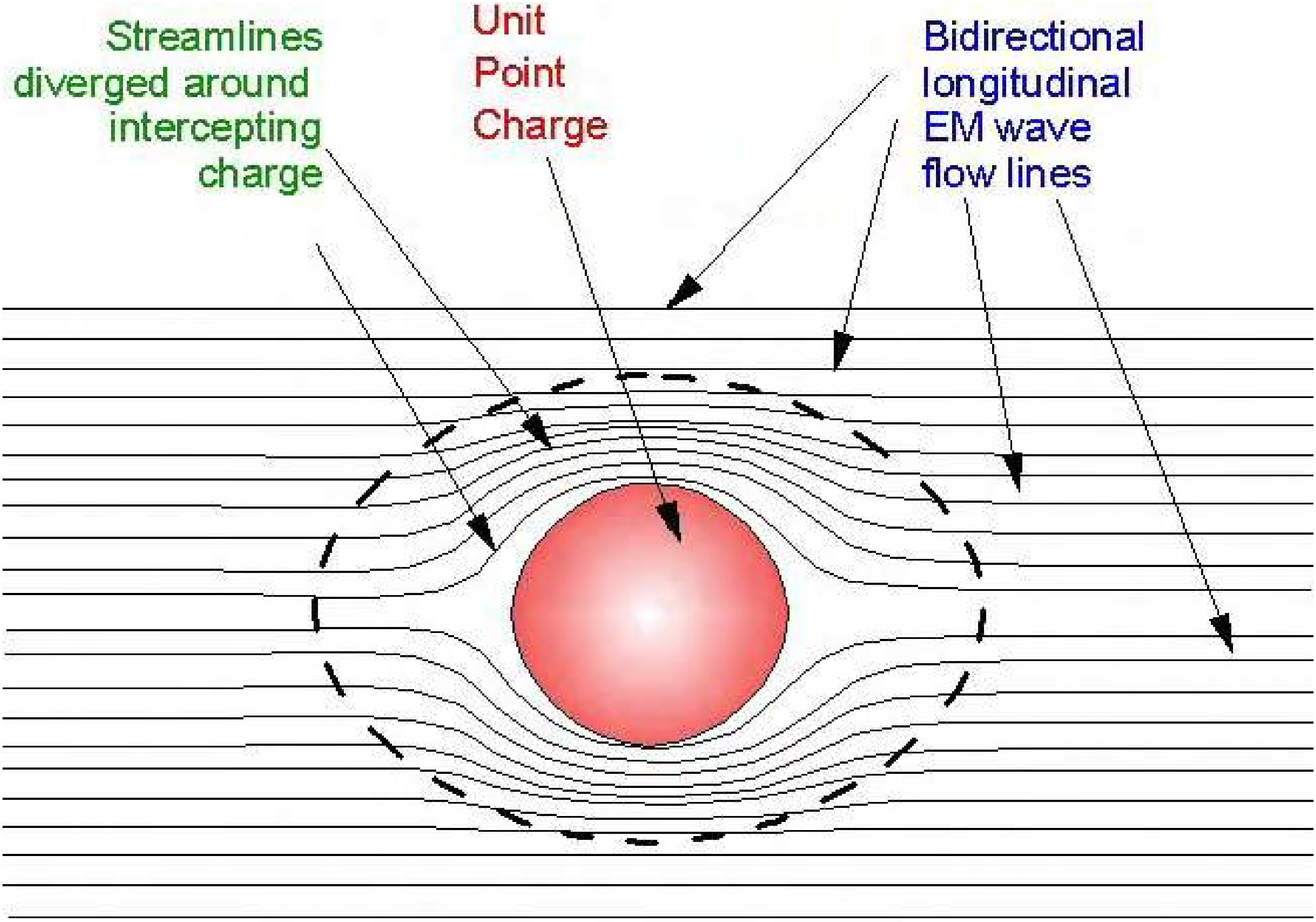
Streamlines diverged around intercepting charge
Unit Point Charge
Bidirectional longitudinal EM wave flow lines

Figure 5. The defined "magnitude of the potential" is not such at all, but is the magnitude of the energy in the diverged flow streamlining around a unit point charge at each point in space where the potential is present.

Electrodynamics has nothing to say about what exists in the vacuum in the absence of mass.

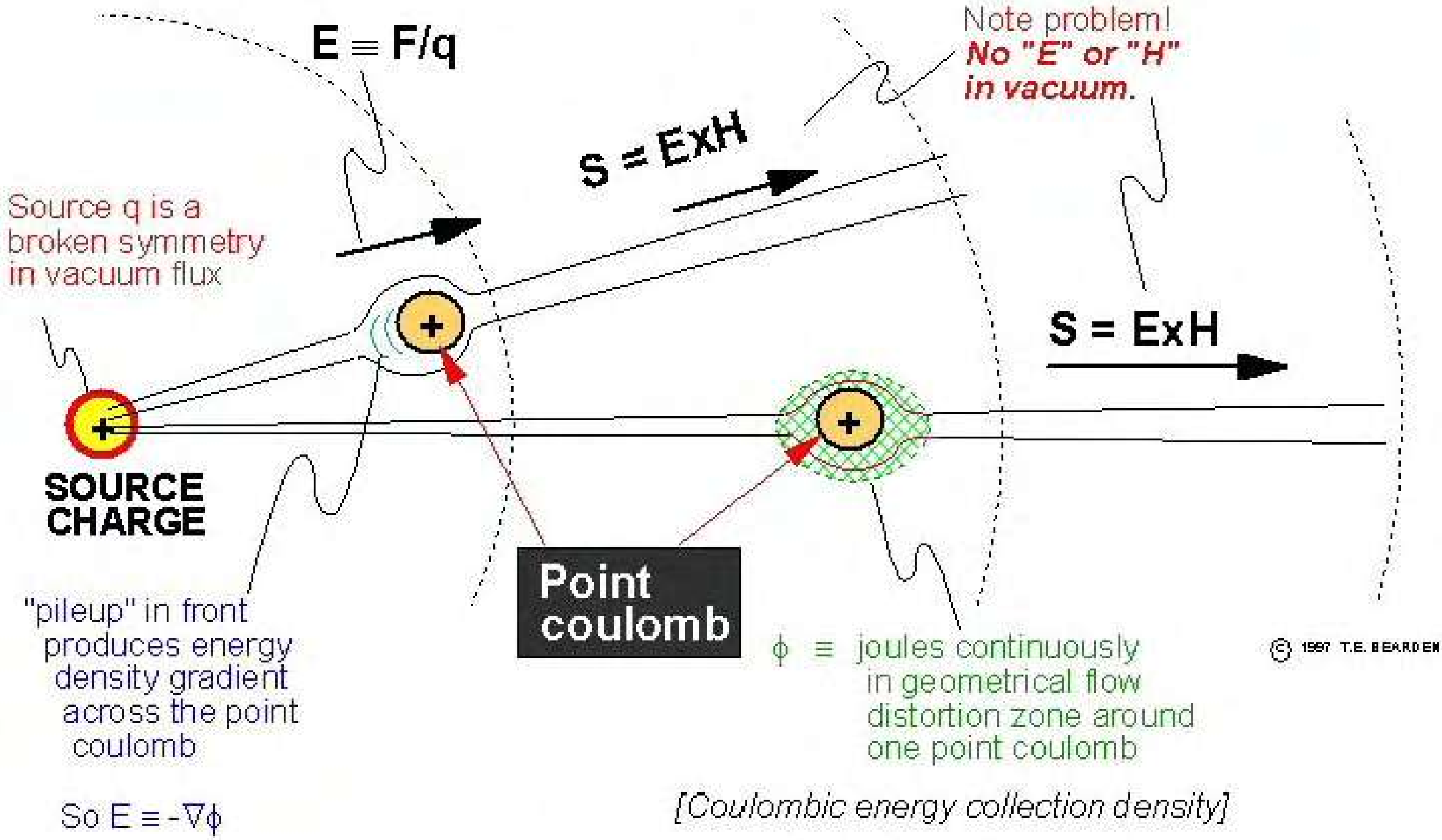
E ≡ F/q
Note problem!
No "E" or "H" in vacuum.
S = ExH
Source q is a broken symmetry in vacuum flux
S = ExH
SOURCE CHARGE
Point coulomb
"pileup" in front produces energy density gradient across the point coulomb
So E ≡ -∇ϕ
ϕ ≡ joules continuously in geometrical flow distortion zone around one point coulomb
[Coulombic energy collection density]
© 1997 T.E. BEARDEN

Figure 6. Fields and potentials are defined in terms of the Poynting disruption of the overall energy flow from the source charge, about an intercepting and "collecting" charge. Rigorously, *as defined* E and H exist only in and on charged matter. The magnitude of the diverted water swinging around a fixed rock in a river, is not the magnitude of the river.

]

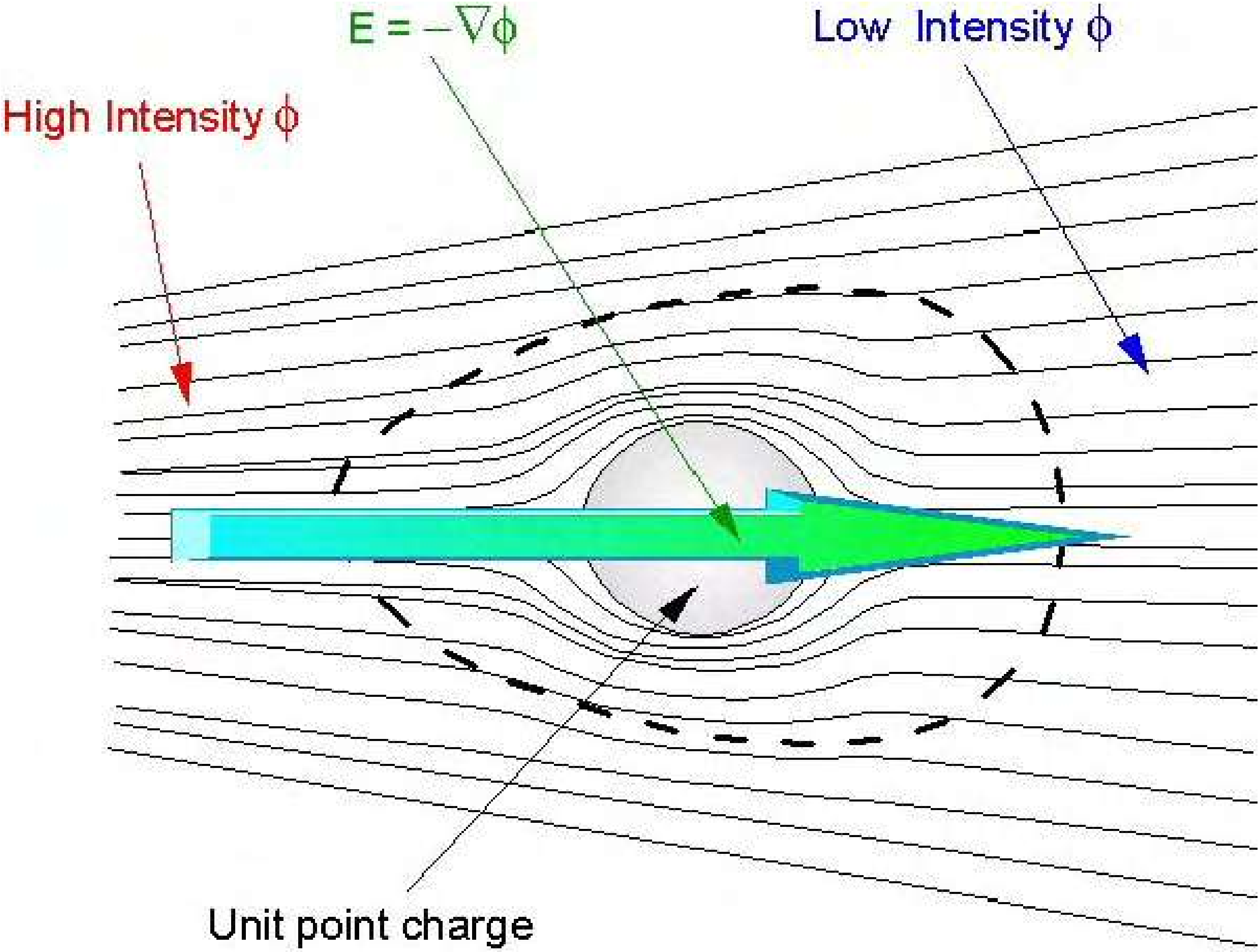
E = −∇ϕ
Low Intensity ϕ
High Intensity ϕ
Unit point charge

Figure 7. The defined "magnitude of the field" depends on the change in intensity of the flows comprising the potential. The E-field is oriented from high (pressure) in the flow to low (pressure) in energy density over a unit length.

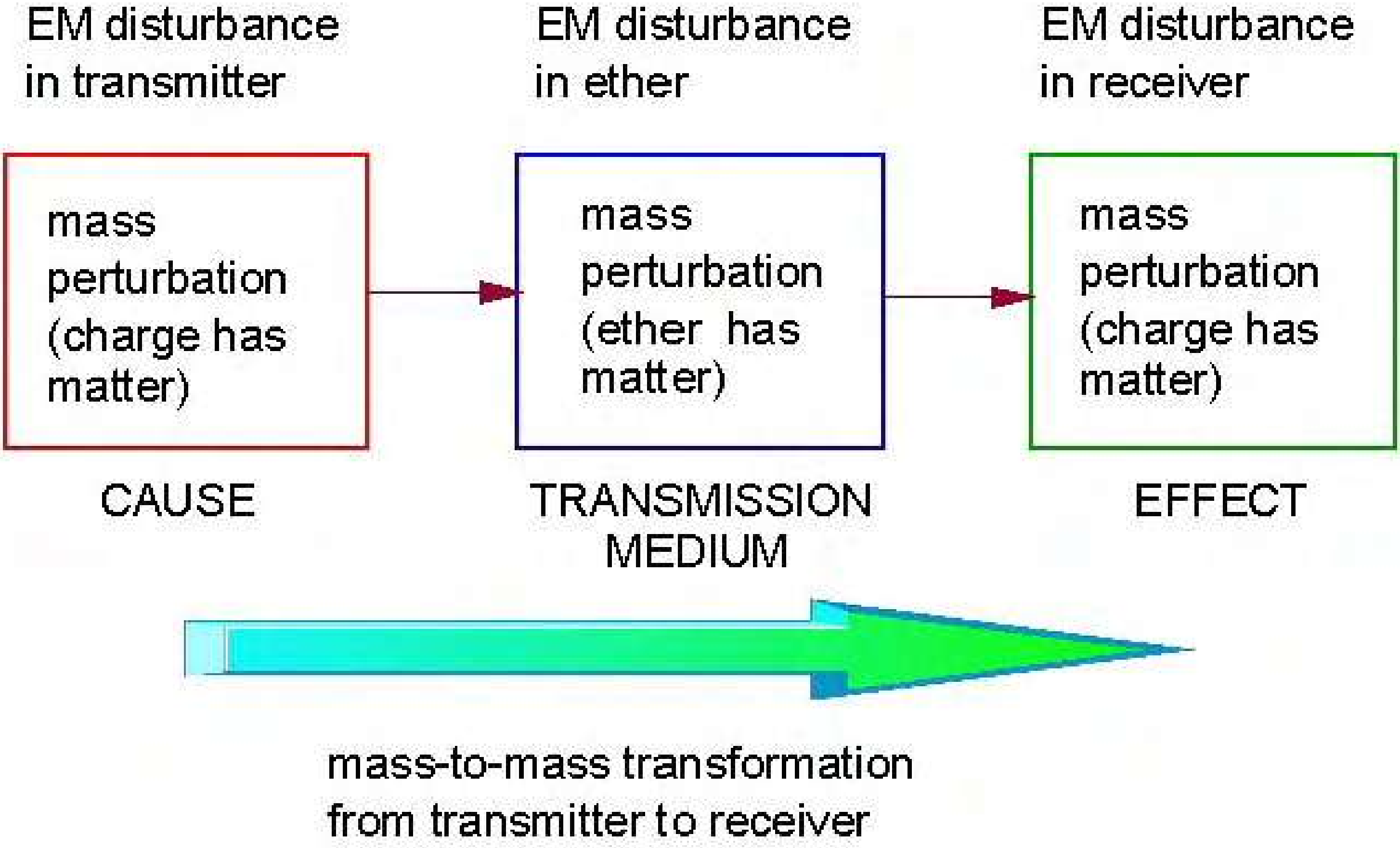
EM disturbance
in transmitter
mass
perturbation
(charge has
matter)
CAUSE
EM disturbance
in ether
mass
perturbation
(ether has
matter)
TRANSMISSION
MEDIUM
EM disturbance
in receiver
mass
perturbation
(charge has
matter)
EFFECT
mass-to-mass transformation
from transmitter to receiver

Figure 8. Maxwell's theory assumed a matter-to-matter transform from cause to effect.

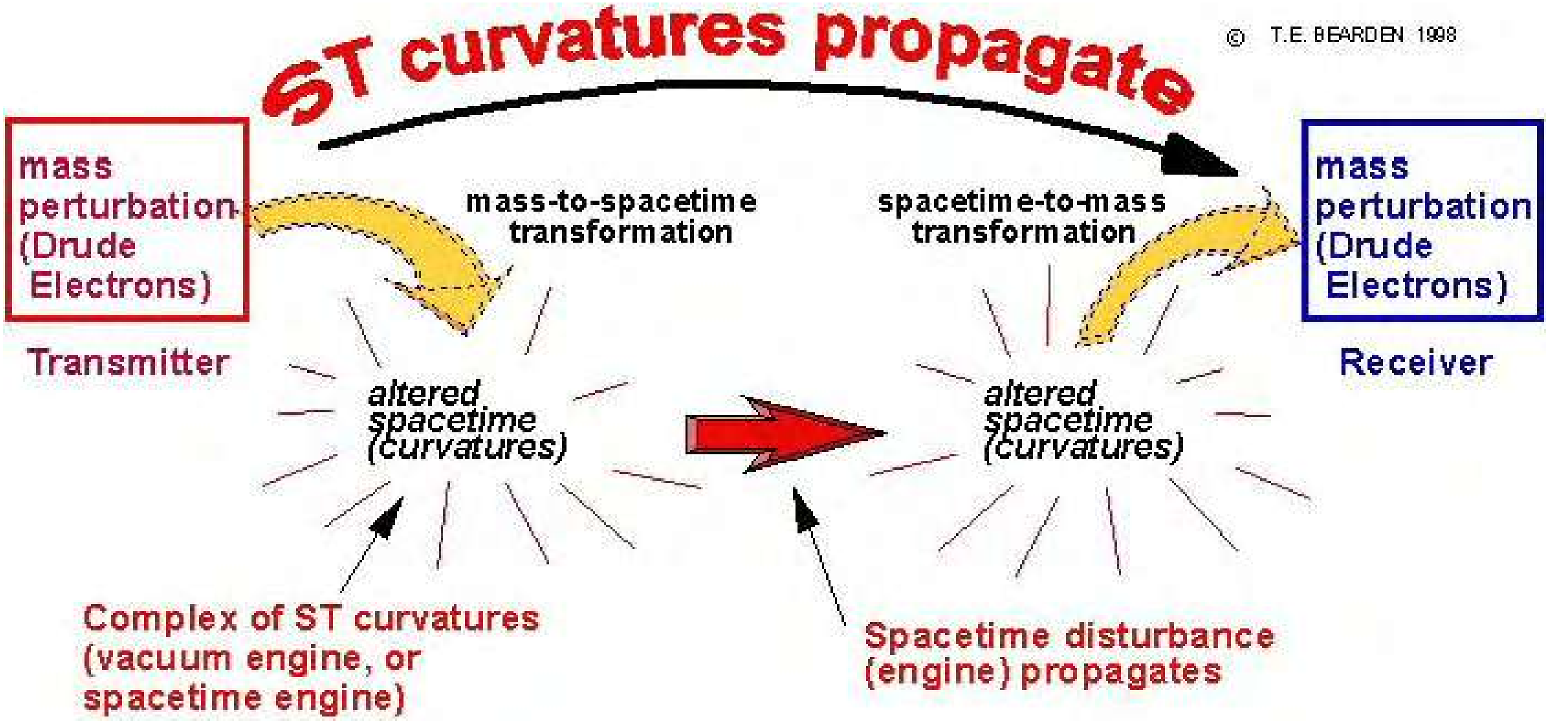
© T.E. BEARDEN 1998
ST curvatures propagate
mass perturbation (Drude Electrons)
Transmitter
mass-to-spacetime transformation
spacetime-to-mass transformation
mass perturbation (Drude Electrons)
Receiver
altered spacetime (curvatures)
altered spacetime (curvatures)
Complex of ST curvatures (vacuum engine, or spacetime engine)
Spacetime disturbance (engine) propagates

The mass-to-mass transform contains two hidden infolded transforms: (i) the mass-to-spacetime transform and (ii) the spacetime-to-mass transform. What is transmitted and propagates in space is a *spacetime perturbation*.

Figure 9. Maxwell's matter-to-matter transform contains two hidden transforms: (1) matter-to-spacetime, and (2) spacetime-to-matter.

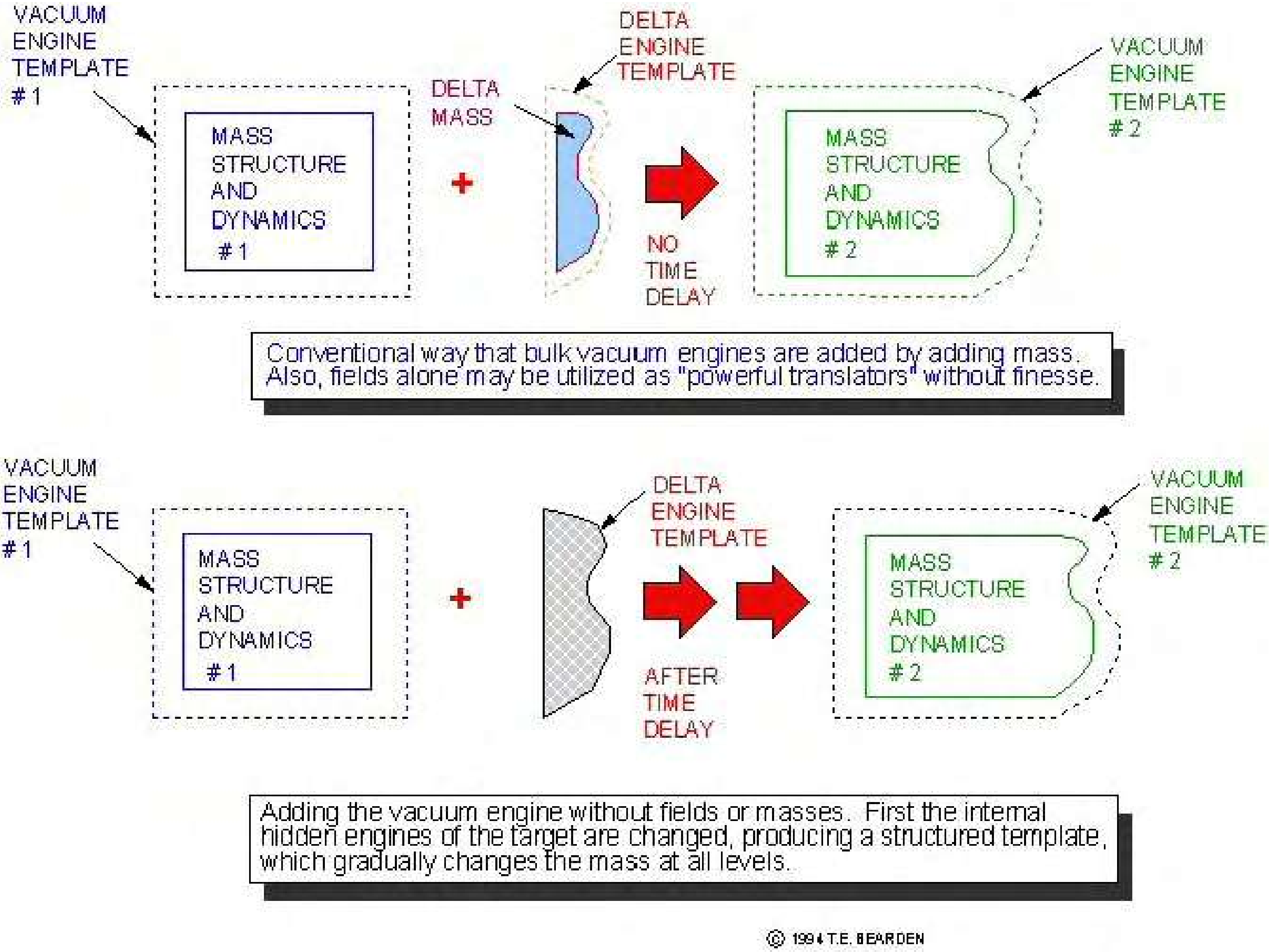
VACUUM ENGINE TEMPLATE # 1
MASS STRUCTURE AND DYNAMICS # 1
+
DELTA MASS
DELTA ENGINE TEMPLATE
NO TIME DELAY
MASS STRUCTURE AND DYNAMICS # 2
VACUUM ENGINE TEMPLATE # 2
VACUUM ENGINE TEMPLATE # 1
MASS STRUCTURE AND DYNAMICS # 1
+
DELTA ENGINE TEMPLATE
AFTER TIME DELAY
MASS STRUCTURE AND DYNAMICS # 2
VACUUM ENGINE TEMPLATE # 2
© 1994 T.E. BEARDEN

Conventional way that bulk vacuum engines are added by adding mass. Also, fields alone may be utilized as "powerful translators" without finesse.

Adding the vacuum engine without fields or masses. First the internal hidden engines of the target are changed, producing a structured template, which gradually changes the mass at all levels.

Figure 10. Vacuum engines and their utilization.

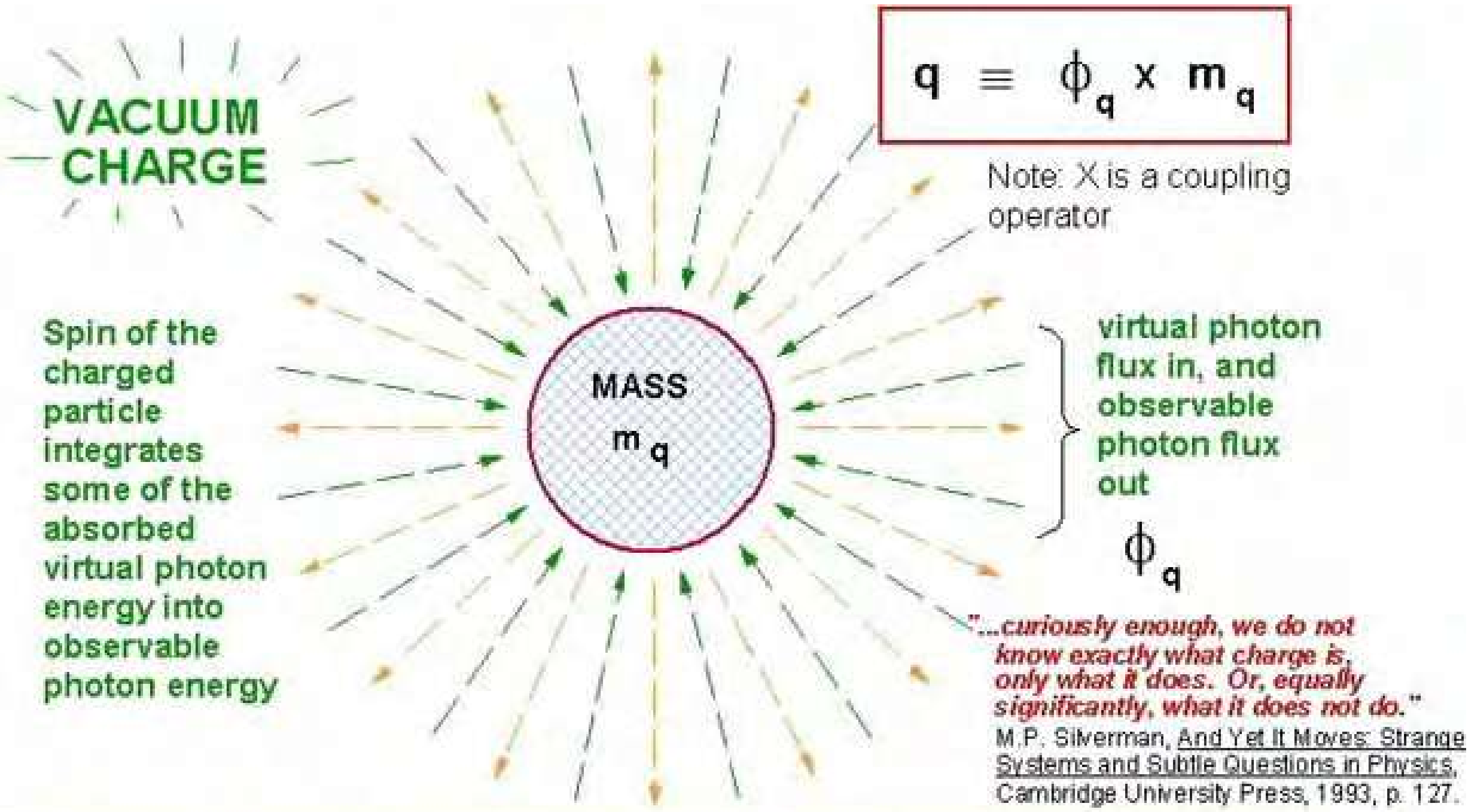
VACUUM CHARGE
$q \equiv \phi_q \times m_q$
Note: X is a coupling operator
Spin of the charged particle integrates some of the absorbed virtual photon energy into observable photon energy
MASS
m_q
virtual photon flux in, and observable photon flux out
ϕ_q

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."
M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge Q is a broken 3-symmetry. The mass of the charge continuously and violently absorbs virtual photons from the surrounding vacuum, and integrates some of them to observable EM longidutinal photons which it emits in 3-space.

Figure 11. A charged particle continuously absorbs virtual energy from the vacuum, integrates some of it into observable energy, and pours out this observable EM energy in 3-space in all directions.

Composition of the scalar potential φ

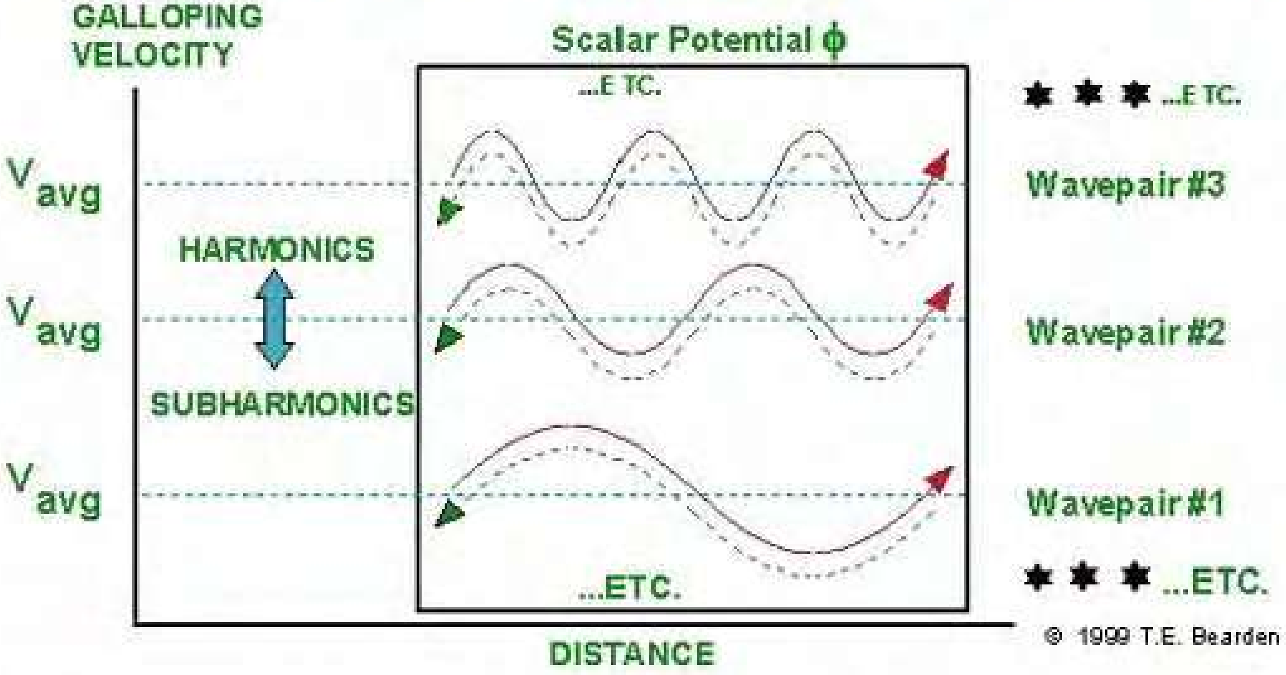
GALLOPING VELOCITY
Scalar Potential φ
...ETC.
Vavg
Vavg
Vavg
HARMONICS
SUBHARMONICS
Wavepair #3
Wavepair #2
Wavepair #1
✱ ✱ ✱ ...ETC.
✱ ✱ ✱ ...ETC.
...ETC.
DISTANCE

© 1999 T.E. Bearden

A harmonic set of bidirectional longitudinal EM wavepairs in 3-space. Unseen here is the time-polarized EM wave in the time domain, which reacts with the source charge to produce the 3-space biwave potential.

Figure 12. The potential as observed or detected is a harmonic set of bidirectional longitudinal EM waves in 3-space. That is, this potential is the "effect" of transduction of an incoming time-polarized EM wave interacting with the source charge.

The supersystem concept

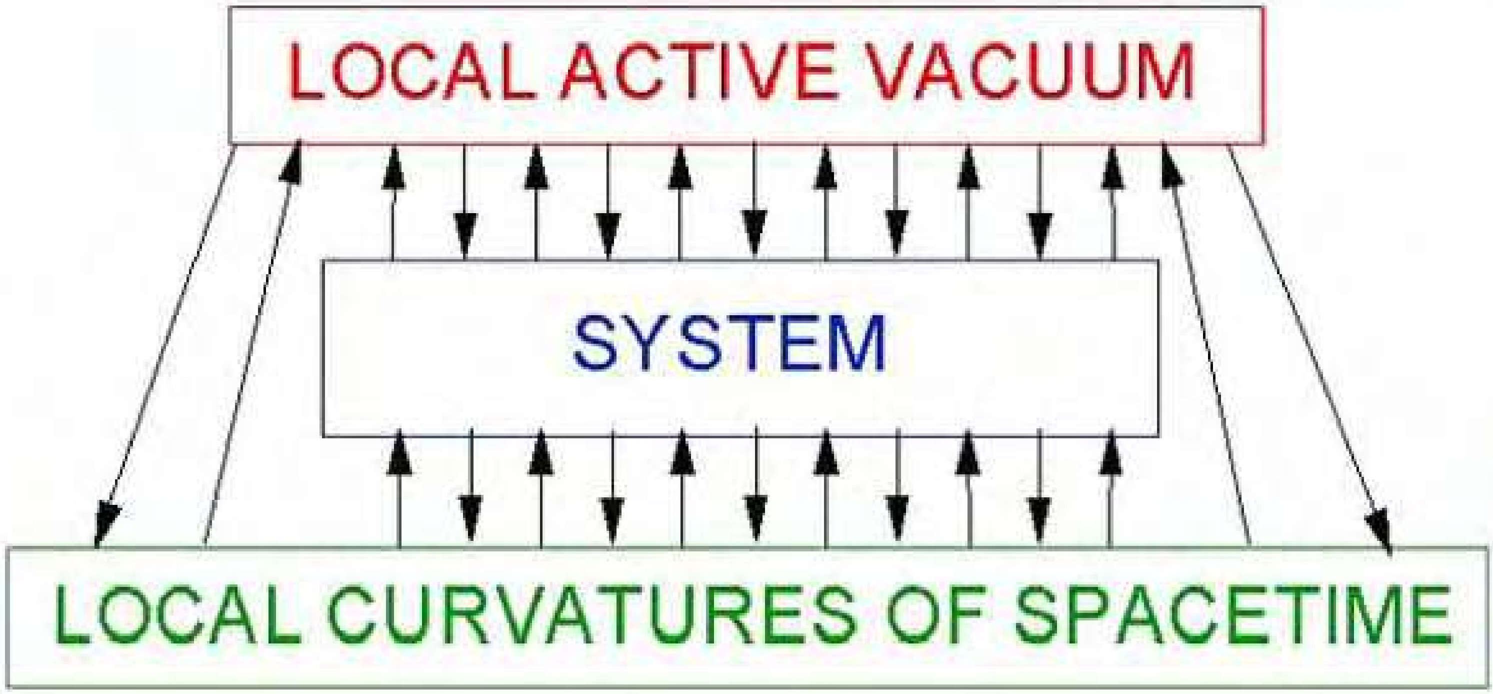
LOCAL ACTIVE VACUUM
SYSTEM
LOCAL CURVATURES OF SPACETIME

Figure 13. Concept of the supersystem and its interacting components.

Vacuum engine: Working spacetime demons

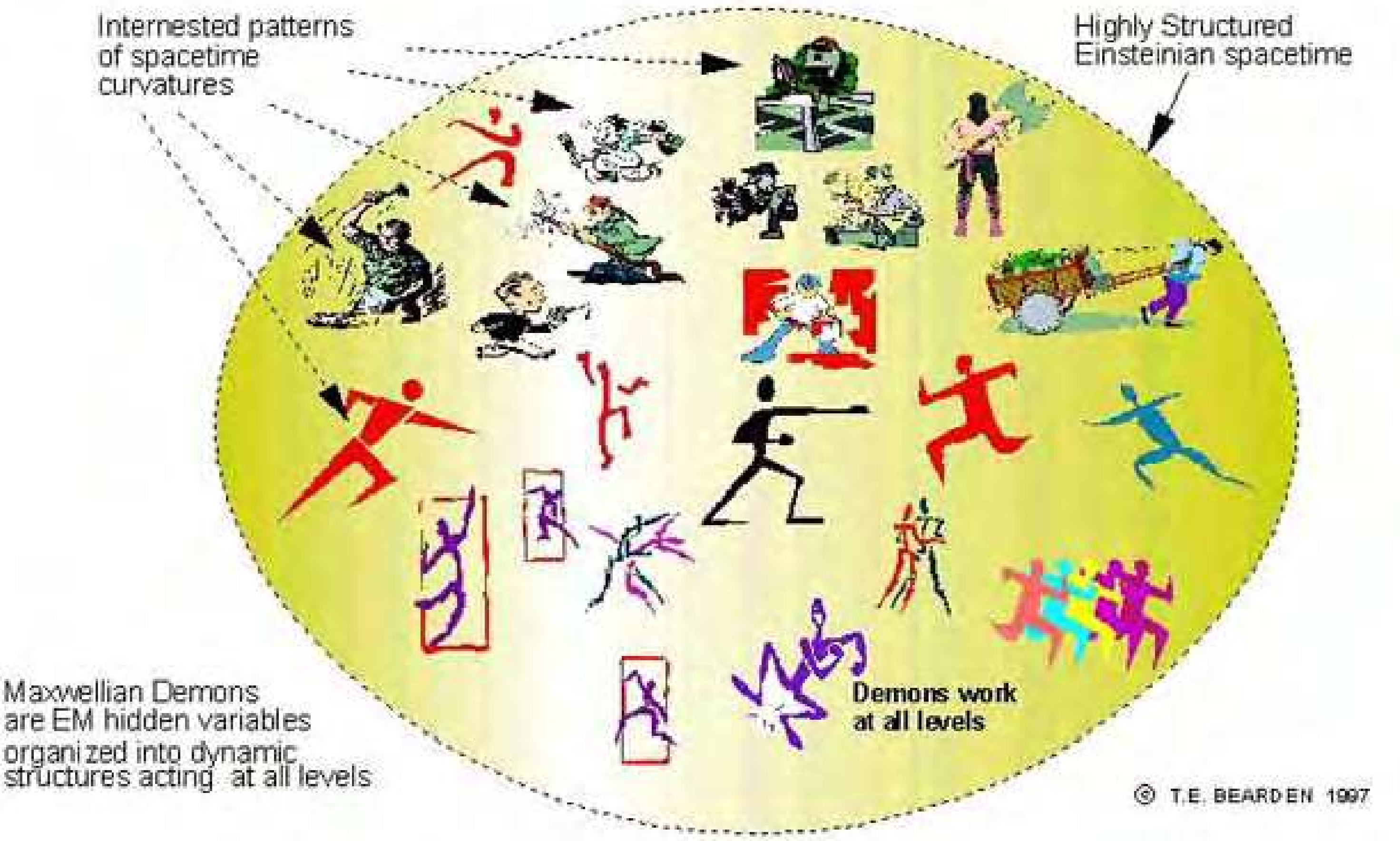
Internested patterns of spacetime curvatures
Highly Structured Einsteinian spacetime
Maxwellian Demons are EM hidden variables organized into dynamic structures acting at all levels
Demons work at all levels
© T.E. BEARDEN 1997

Figure 14. A vacuum engine is a set of spacetime demons working upon a mass system and its motion through time, at every level.

MTW's* general relativity principle

- Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

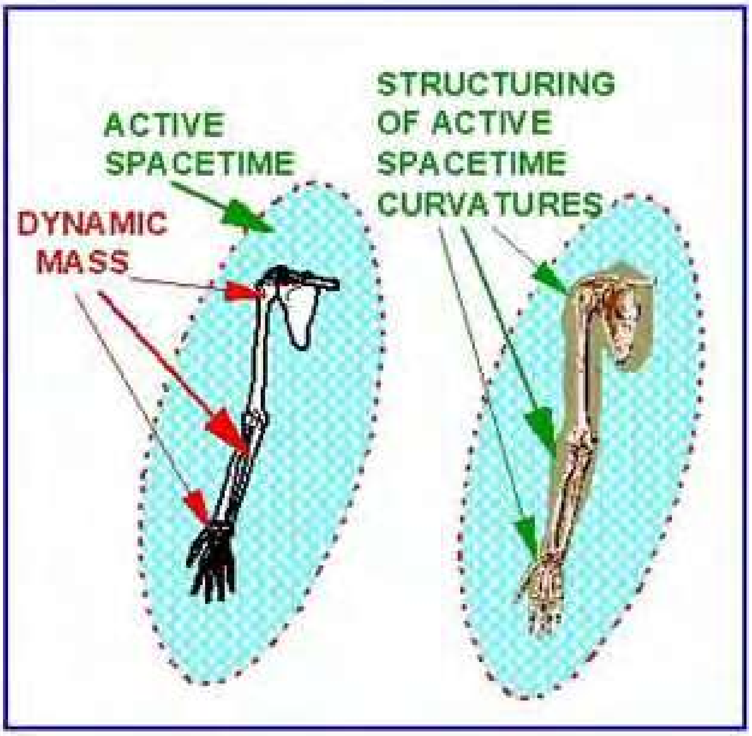
ACTIVE
SPACETIME
DYNAMIC
MASS
STRUCTURING
OF ACTIVE
SPACETIME
CURVATURES

Figure 15. General relativity principle of Misner, Thorne, and Wheeler (MTW).

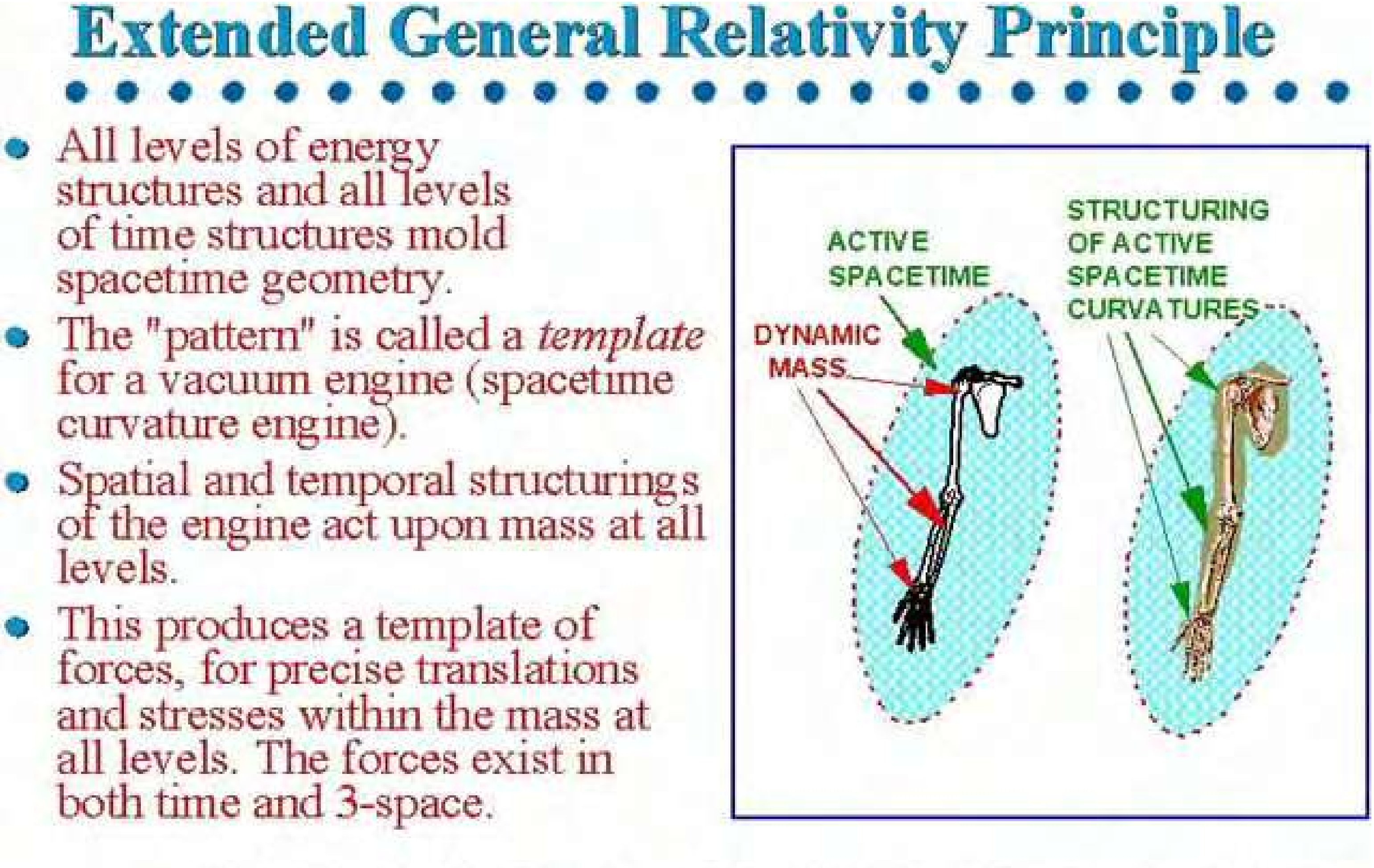

Extended General Relativity Principle

- All levels of energy structures and all levels of time structures mold spacetime geometry.
- The "pattern" is called a *template* for a vacuum engine (spacetime curvature engine).
- Spatial and temporal structurings of the engine act upon mass at all levels.
- This produces a template of forces, for precise translations and stresses within the mass at all levels. The forces exist in both time and 3-space.

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

Figure 16. The principle of general relativity extended to include the vacuum engine.

Active spacetime and a specific vacuum engine

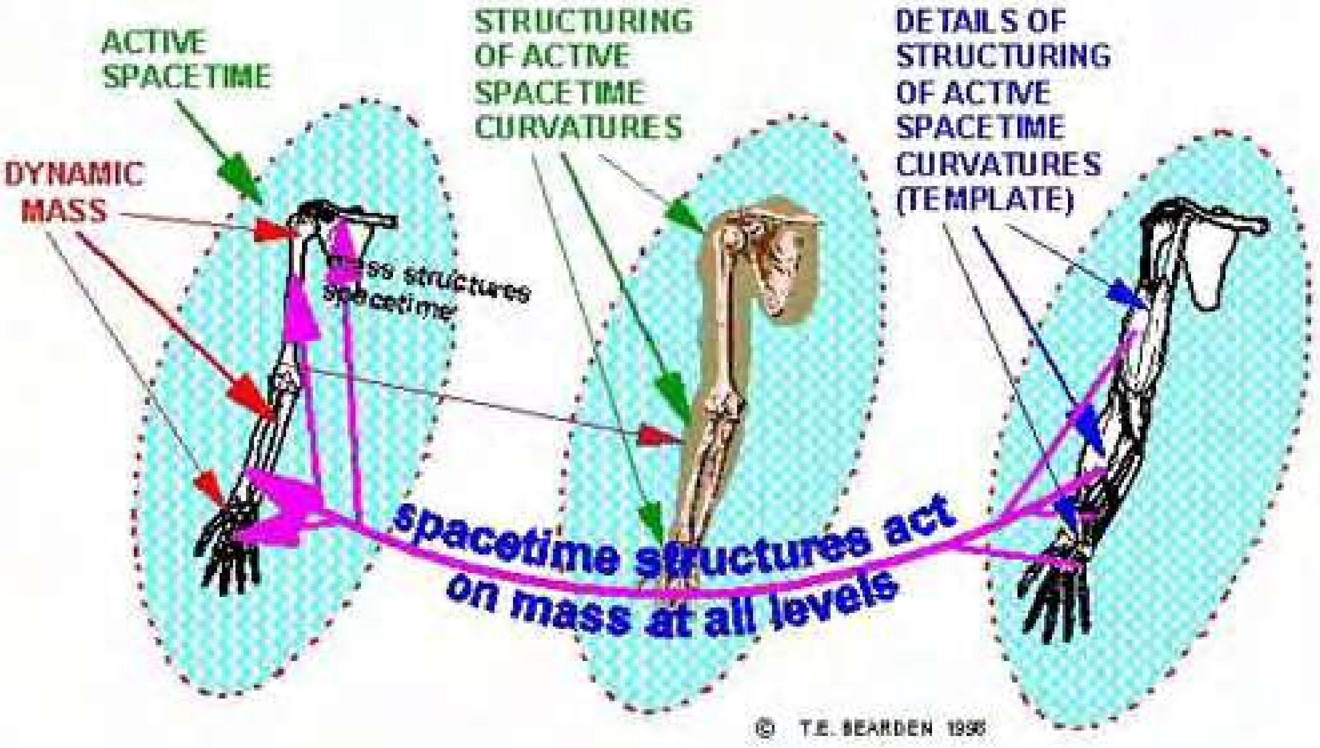
ACTIVE SPACETIME
DYNAMIC MASS
mass structures spacetime
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)
spacetime structures act on mass at all levels
© T.E. BEARDEN 1996

Figure 17. Active spacetime and a specific vacuum engine (spacetime curvature engine).

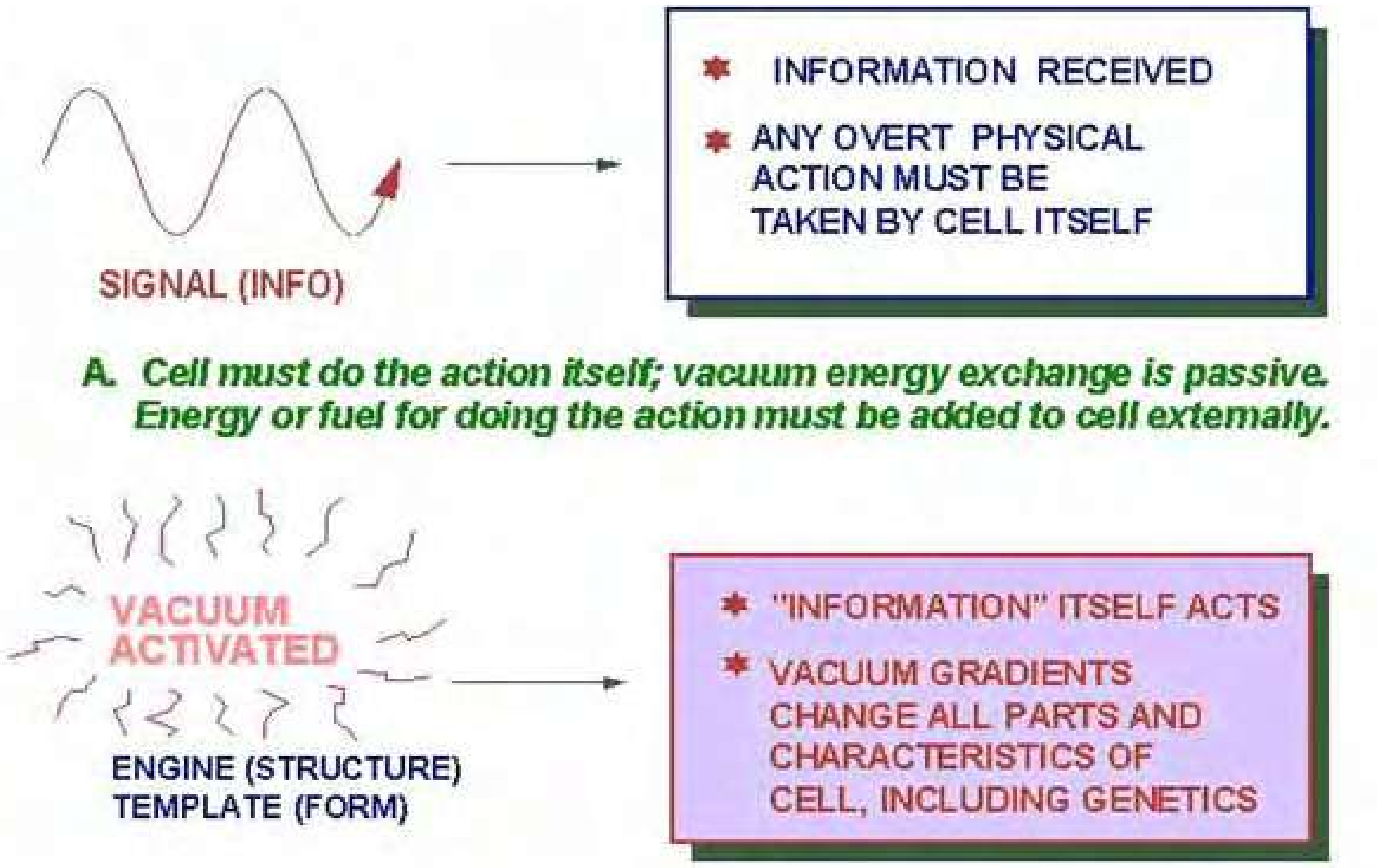
SIGNAL (INFO)
INFORMATION RECEIVED
ANY OVERT PHYSICAL ACTION MUST BE TAKEN BY CELL ITSELF
VACUUM ACTIVATED
ENGINE (STRUCTURE)
TEMPLATE (FORM)
"INFORMATION" ITSELF ACTS
VACUUM GRADIENTS CHANGE ALL PARTS AND CHARACTERISTICS OF CELL, INCLUDING GENETICS

A. *Cell must do the action itself; vacuum energy exchange is passive. Energy or fuel for doing the action must be added to cell externally.*

B. *Cell is acted upon; vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to cell externally.*

Figure 18. Signal versus vacuum engine, as in Priore's cellular reversal.

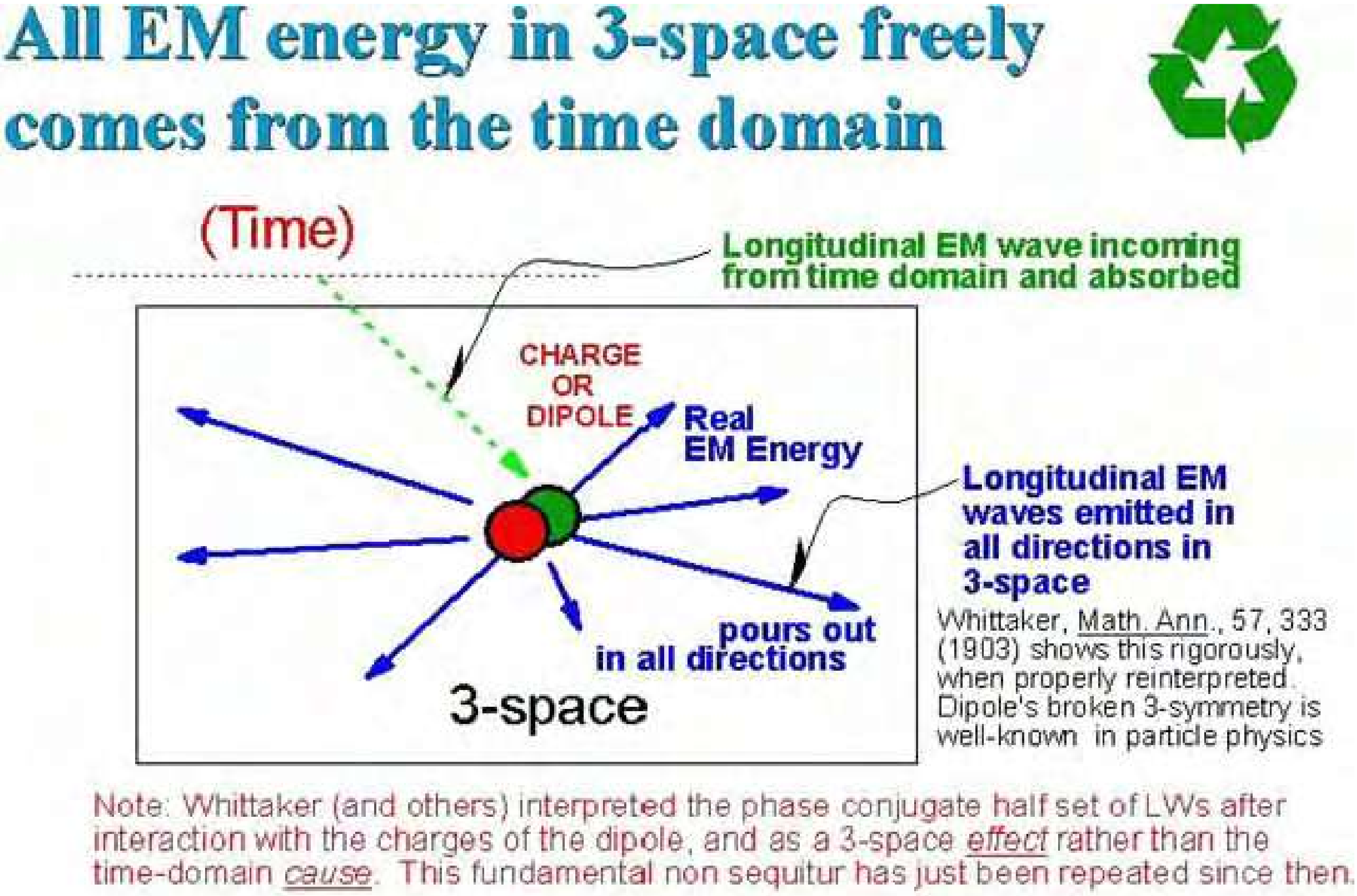

All EM energy in 3-space freely comes from the time domain

(Time)
Longitudinal EM wave incoming from time domain and absorbed
CHARGE OR DIPOLE
Real EM Energy
Longitudinal EM waves emitted in all directions in 3-space
pours out in all directions
3-space

Whittaker, Math. Ann., 57, 333 (1903) shows this rigorously, when properly reinterpreted. Dipole's broken 3-symmetry is well-known in particle physics

Note: Whittaker (and others) interpreted the phase conjugate half set of LWs after interaction with the charges of the dipole, and as a 3-space *effect* rather than the time-domain *cause*. This fundamental non sequitur has just been repeated since then.

Figure 19. All EM energy in the 3-space potential comes from the time-domain.

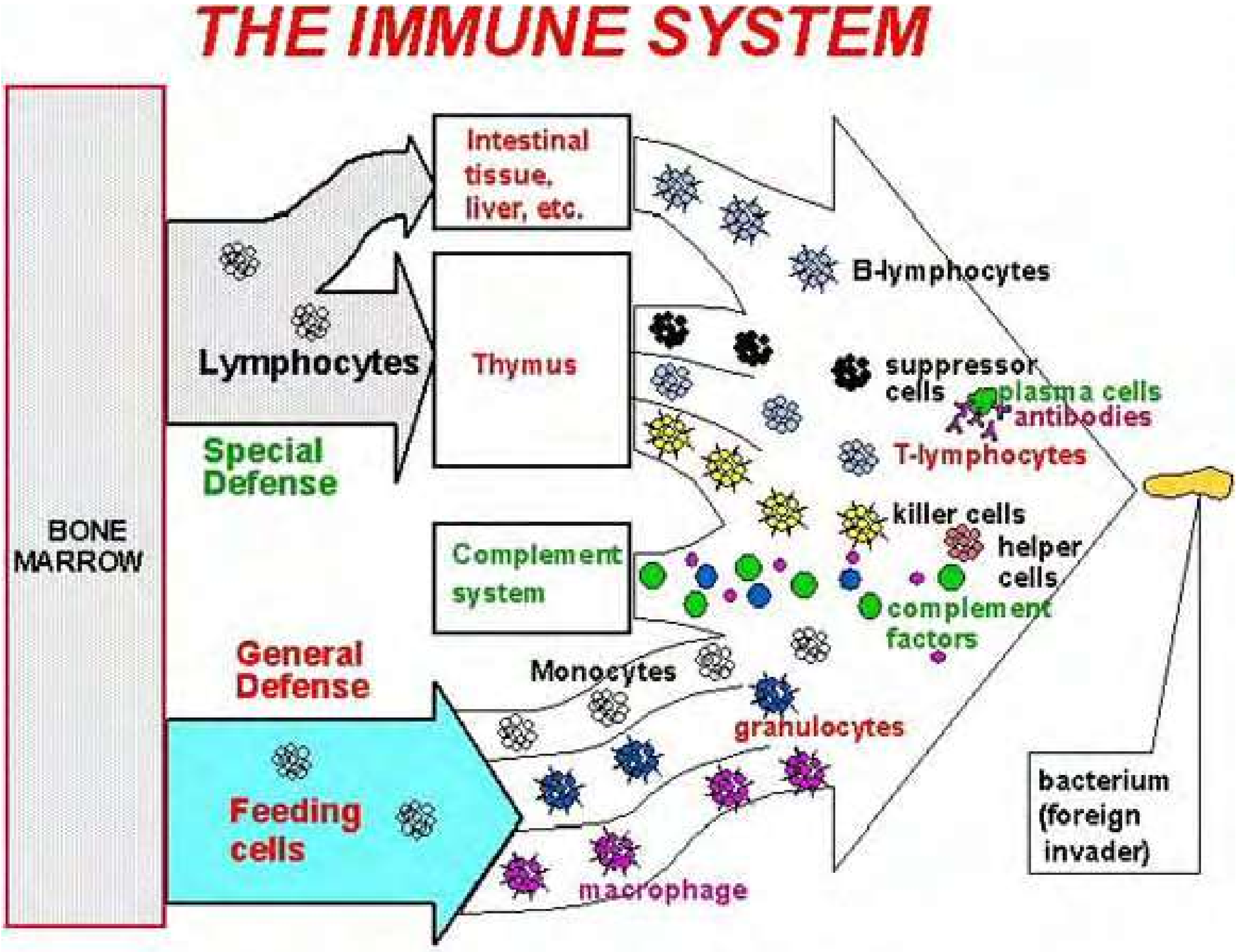
THE IMMUNE SYSTEM
BONE MARROW
Lymphocytes
Special Defense
General Defense
Feeding cells
Intestinal tissue, liver, etc.
Thymus
Complement system
Monocytes
granulocytes
macrophage
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
bacterium (foreign invader)

Figure 20. Immune system block diagram.

Becker's bone fracture healing

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

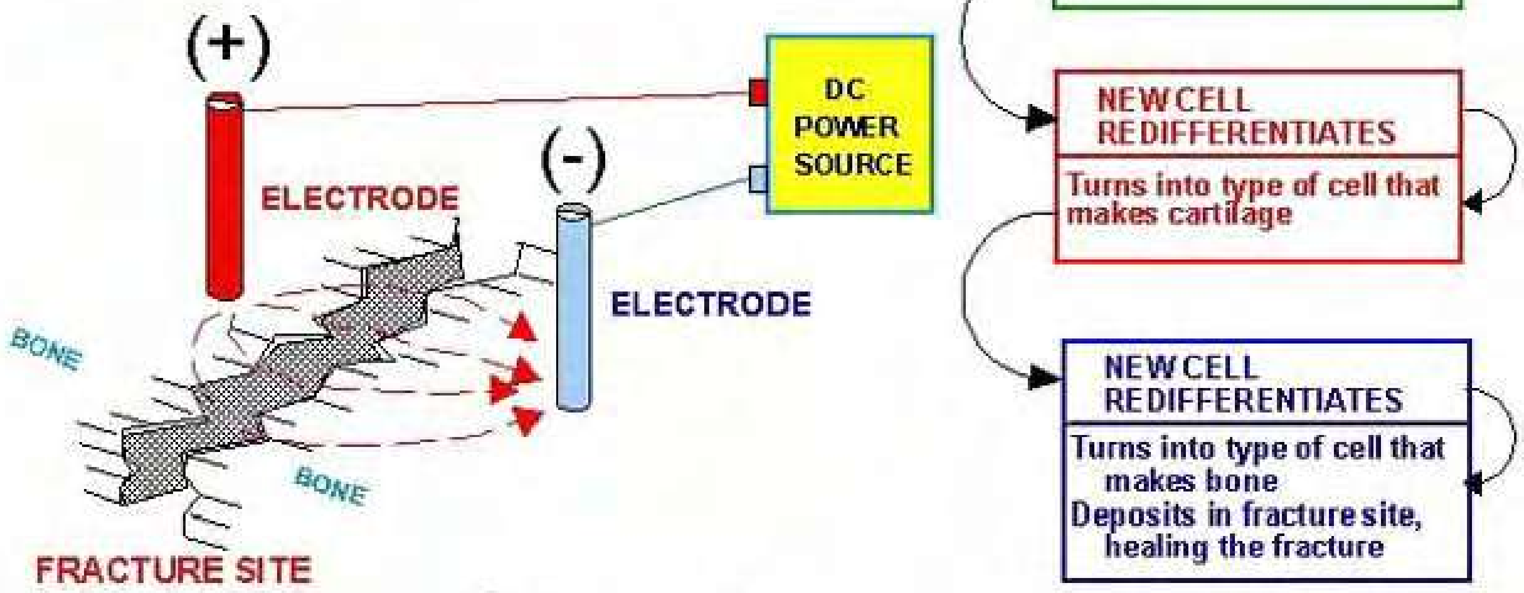

RED BLOOD CELL DEDIFFERENTIATES
Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES
Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES
Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1996 T.E. BEARDEN

Figure 21. Becker's electrodynamic healing of otherwise intractable bone fractures, using weak potentials across the fracture site.

Becker's theoretical DC control system involved with response to injury

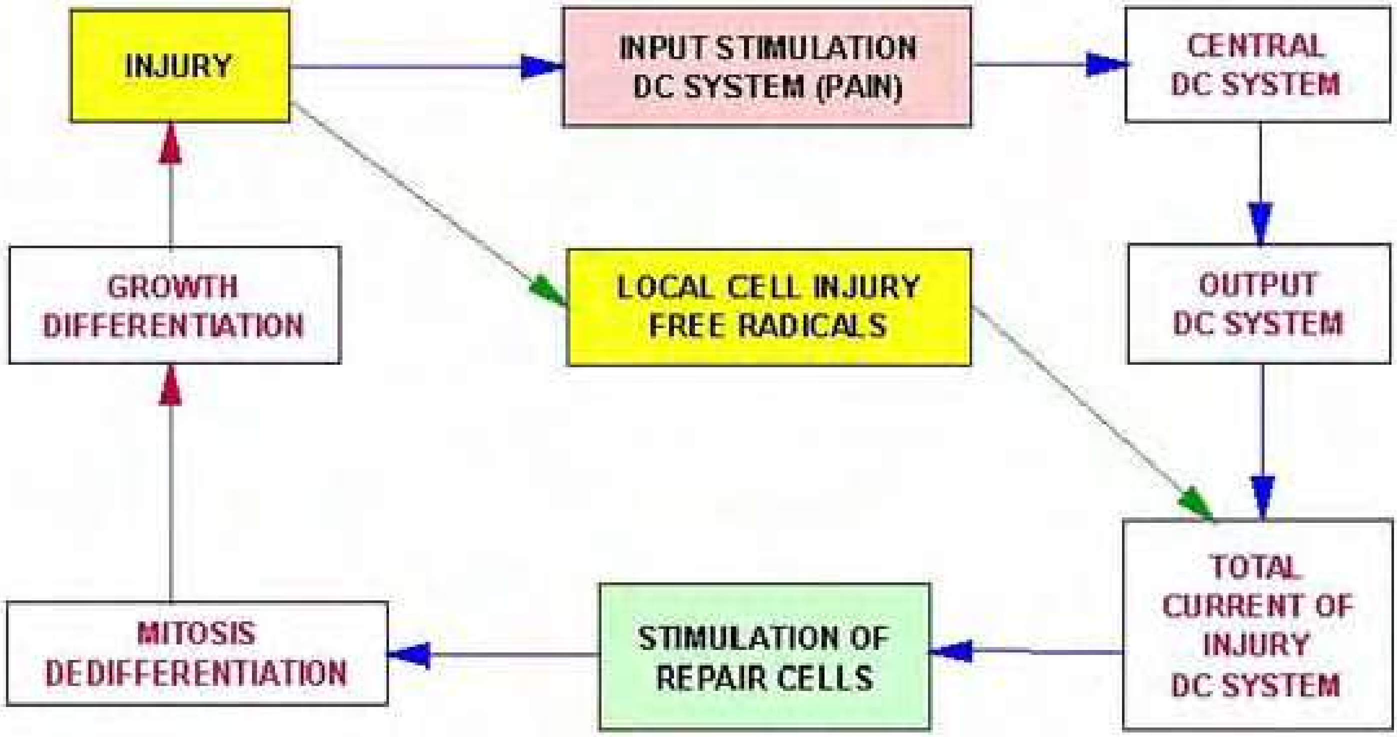
INJURY
INPUT STIMULATION DC SYSTEM (PAIN)
CENTRAL DC SYSTEM
GROWTH DIFFERENTIATION
LOCAL CELL INJURY FREE RADICALS
OUTPUT DC SYSTEM
TOTAL CURRENT OF INJURY DC SYSTEM
MITOSIS DEDIFFERENTIATION
STIMULATION OF REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Figure 22. Becker's block diagram of DC control system involved with response to injury.

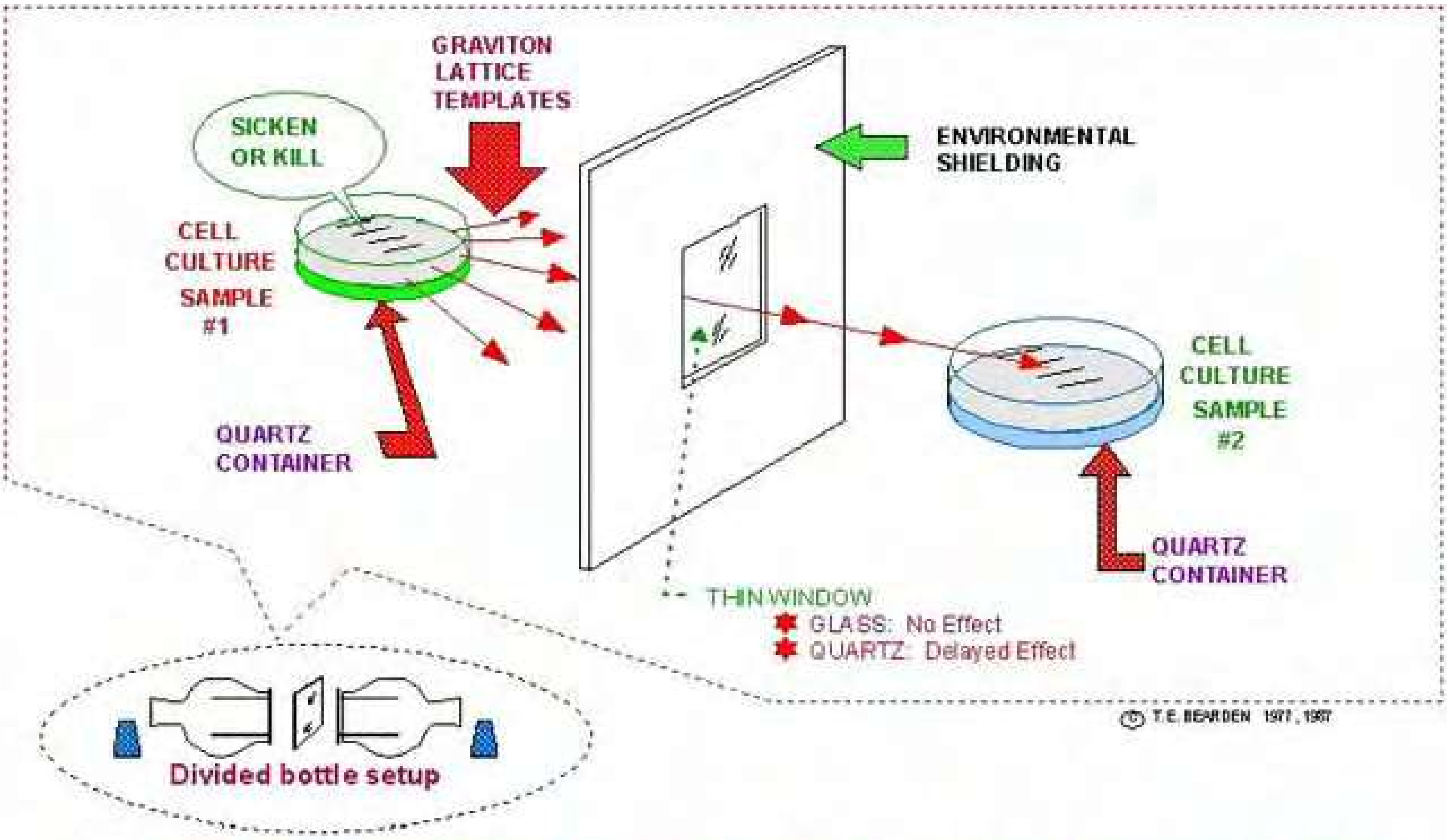
GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
CELL
CULTURE
SAMPLE
#1
QUARTZ
CONTAINER
ENVIRONMENTAL
SHIELDING
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle setup

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Figure 23. Kaznacheyev's induction of cellular disease and disorder at a distance, using novel electromagnetic means.

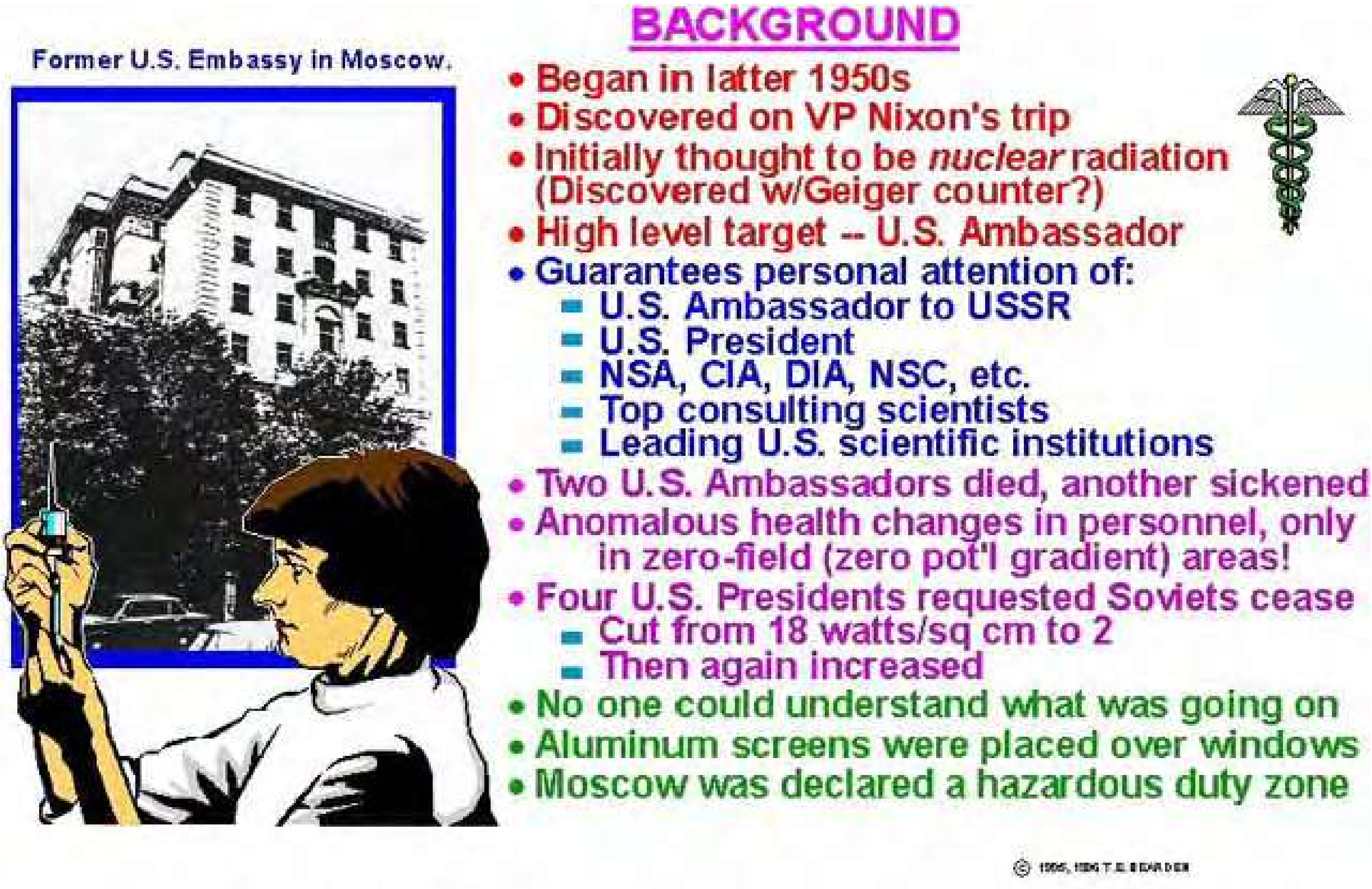

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

© 1995, 1996 T.E. BEARDEN

Figure 24. Russian "microwave" radiation of the U.S. Embassy in Moscow.

Mechanism generating the flow of a mass through time

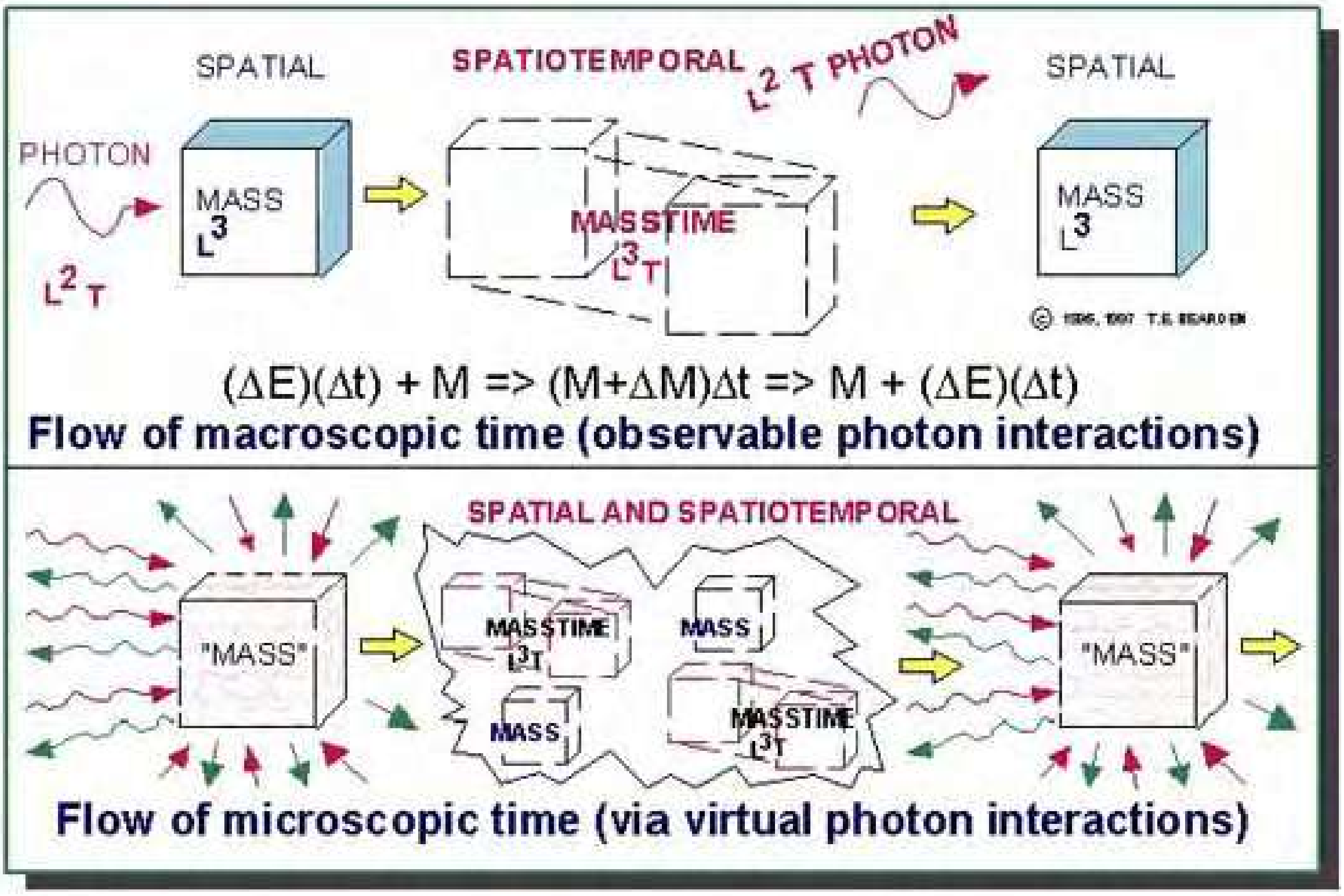
SPATIAL
SPATIOTEMPORAL
L^2 T PHOTON
SPATIAL
PHOTON
MASS L^3
L^2 T
MASSTIME L^3T
MASS L^3
© 1996, 1997 T.E. BEARDEN
$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$
Flow of macroscopic time (observable photon interactions)
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME L^3T
MASS
MASS
MASSTIME L^3T
"MASS"
Flow of microscopic time (via virtual photon interactions)

Figure 25. Mechanism generating the flow of a mass through time.

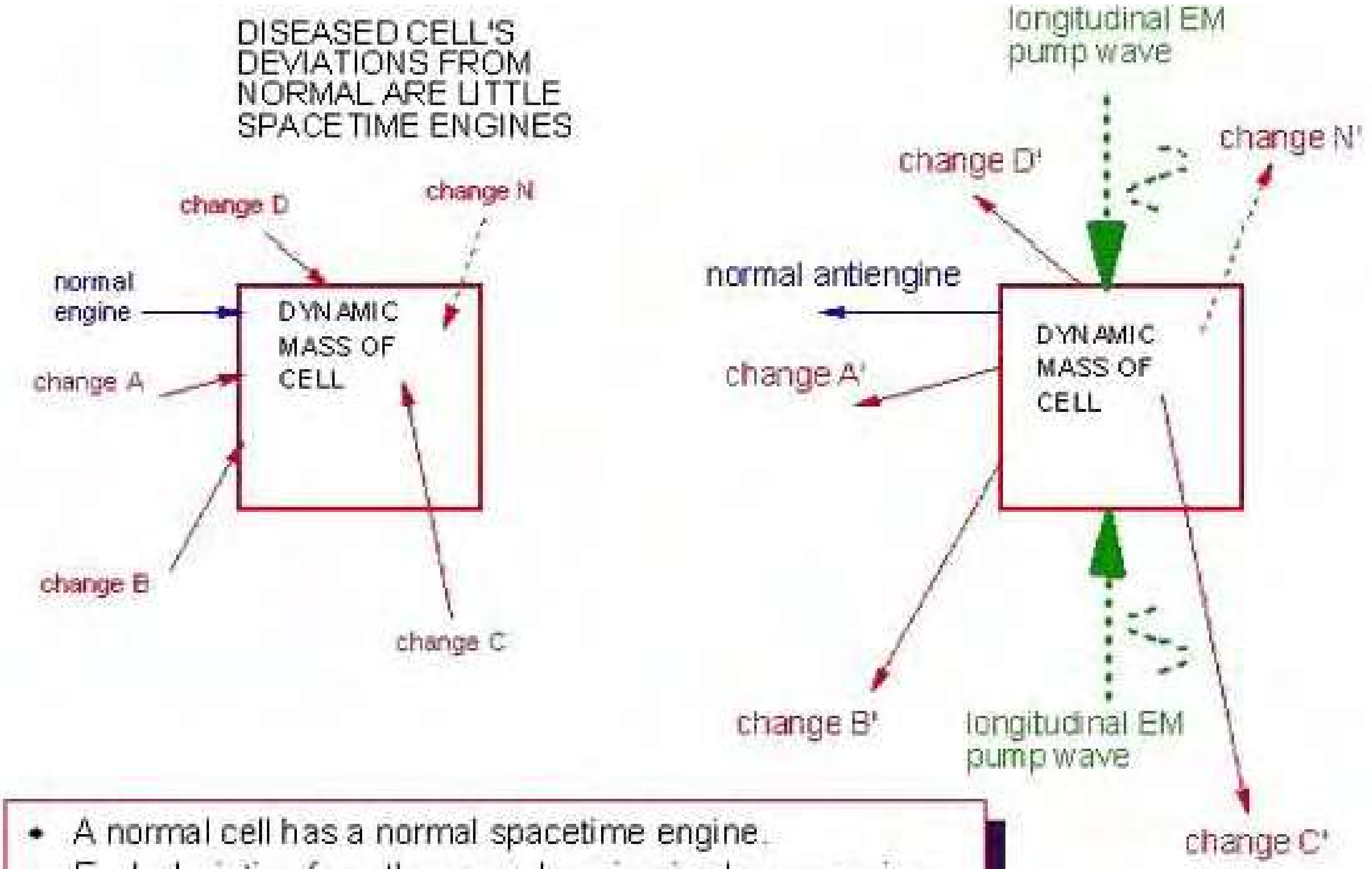
DISEASED CELL'S DEVIATIONS FROM NORMAL ARE LITTLE SPACETIME ENGINES
change D
change N
normal engine
DYNAMIC MASS OF CELL
change A
change B
change C
longitudinal EM pump wave
change D'
change N'
normal antiengine
DYNAMIC MASS OF CELL
change A'
change B'
longitudinal EM pump wave
change C'

- A normal cell has a normal spacetime engine.
- Each deviation from the normal engine is also an engine.
- Pumping with longitudinal EM waves creates amplified antiengines for the normal engine and all deviant engines.
- This set of anti-engines precisely reverses the cell back to a previous physical state, prior to the deviations.

© T.E. BEARDEN 2001

Figure 26. Producing a specific anti-engine for a specific cellular disease condition.

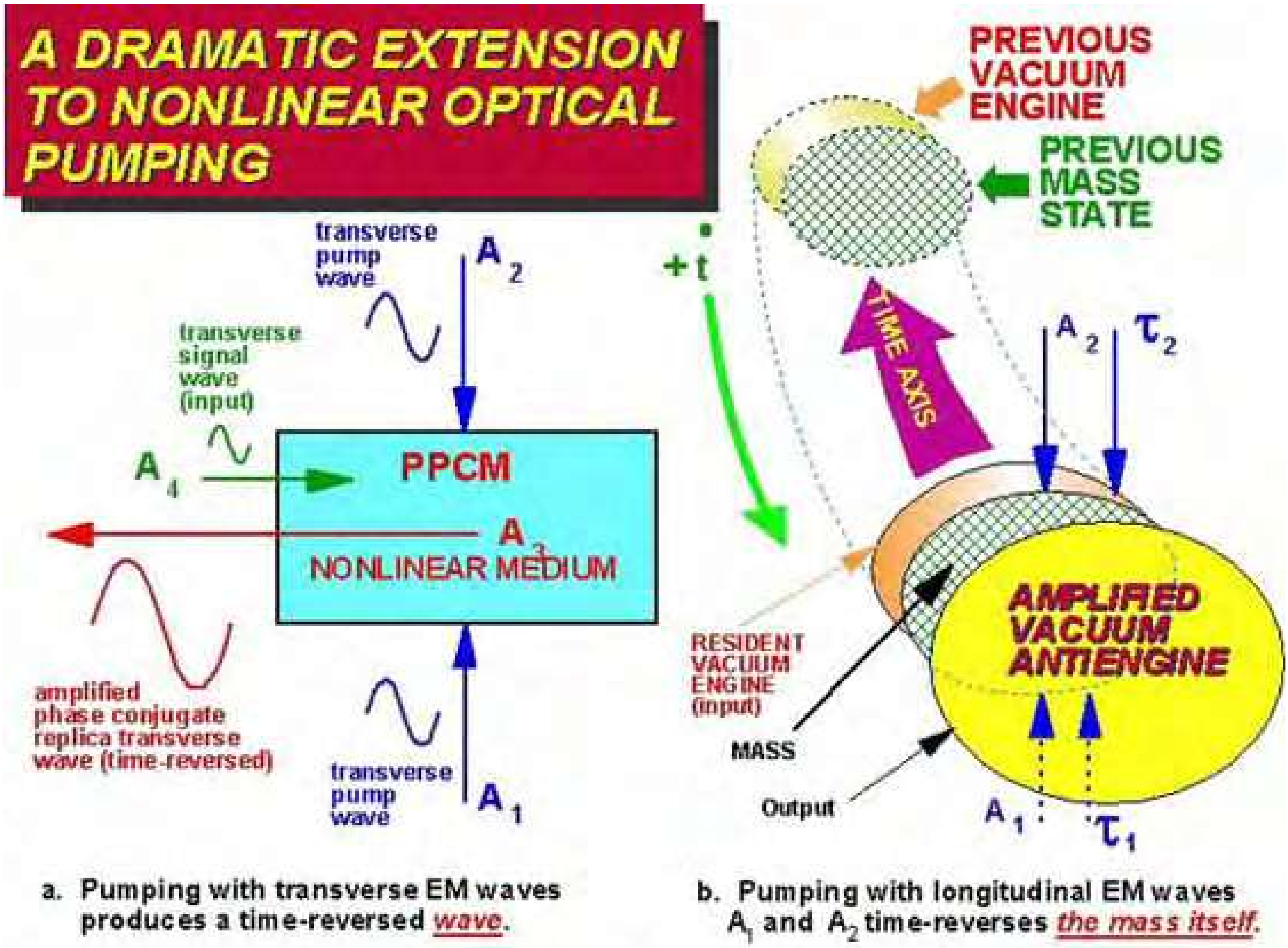
A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING
transverse pump wave
A_2
transverse signal wave (input)
A_4
PPCM
A_3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A_1
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
$+\dot{t}$
TIME AXIS
A_2
τ_2
AMPLIFIED VACUUM ANTIENGINE
RESIDENT VACUUM ENGINE (input)
MASS
Output
A_1
τ_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Figure 27. Pumping with longitudinal EM waves A also pumps in the time domain with time-polarized EM waves τ. This time-reverses a macroscopic mass and its dynamics.

Prioré's therapeutic methodology

Time-reverses the cells back to normal state

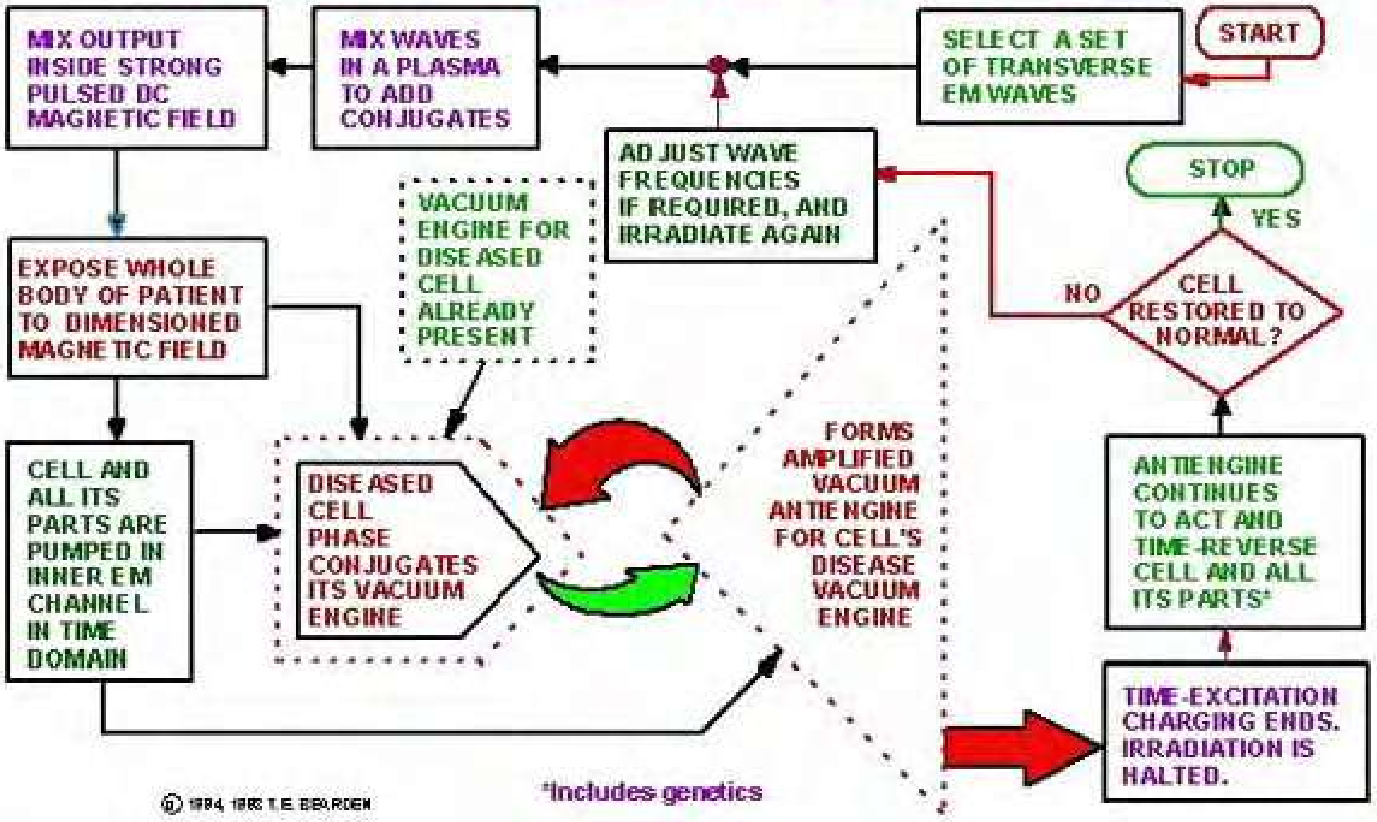
START
SELECT A SET OF TRANSVERSE EM WAVES
MIX WAVES IN A PLASMA TO ADD CONJUGATES
MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD
ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN
STOP
YES
NO
CELL RESTORED TO NORMAL?
VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT
EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD
CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN
DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE
FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE
ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*
TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.
© 1994, 1996 T.E. BEARDEN
*Includes genetics

Figure 28. Block diagram of operation of Prioré's methodology.

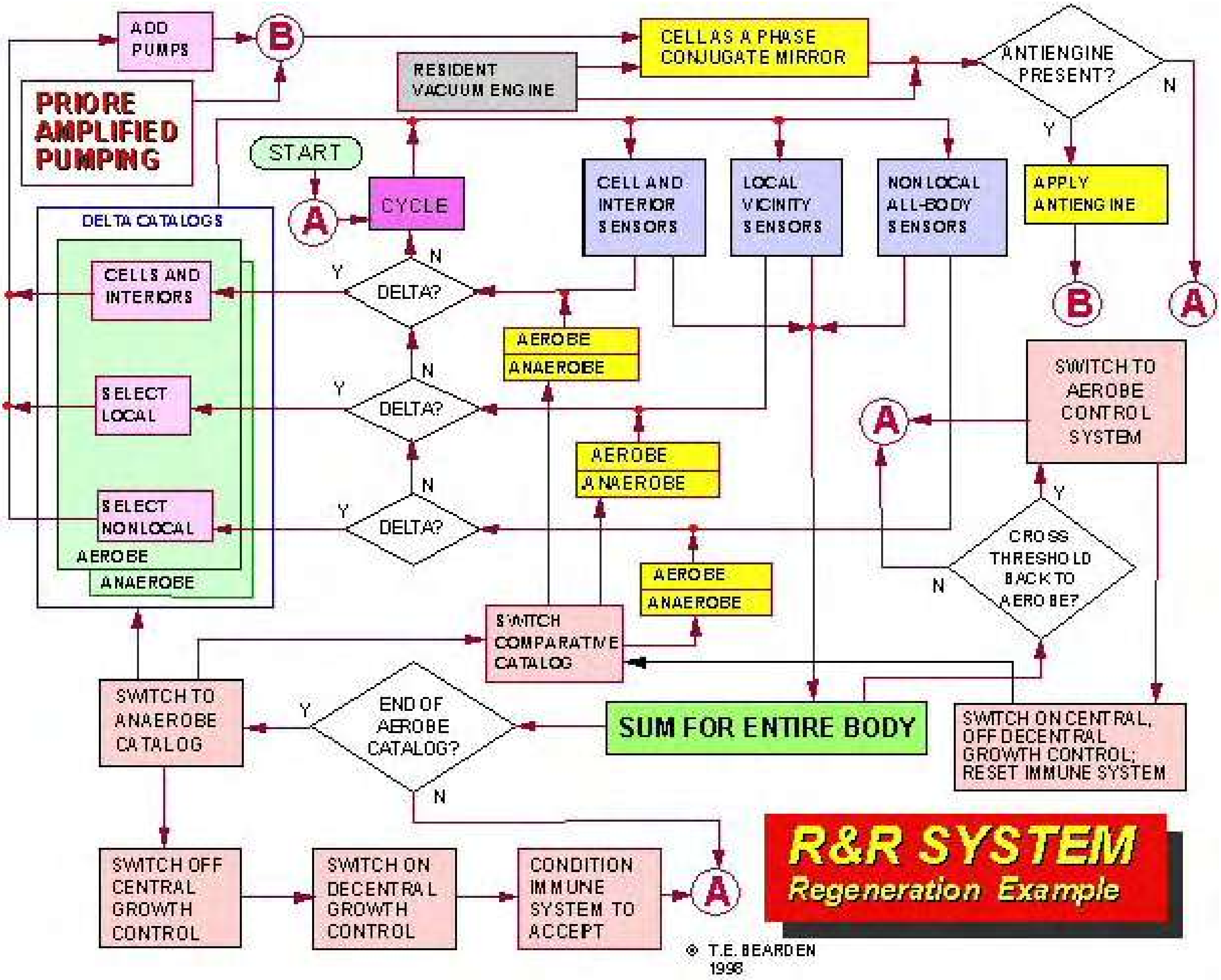
ADD PUMPS
B
CELL AS A PHASE CONJUGATE MIRROR
ANTIENGINE PRESENT?
N
Y
RESIDENT VACUUM ENGINE
PRIORE AMPLIFIED PUMPING
START
A
CYCLE
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NON LOCAL ALL-BODY SENSORS
APPLY ANTIENGINE
B
A
DELTA CATALOGS
CELLS AND INTERIORS
Y
DELTA?
N
AEROBE
ANAEROBE
SELECT LOCAL
Y
DELTA?
N
SWITCH TO AEROBE CONTROL SYSTEM
A
AEROBE
ANAEROBE
SELECT NONLOCAL
Y
DELTA?
N
Y
CROSS THRESHOLD BACK TO AEROBE?
N
AEROBE
ANAEROBE
AEROBE
ANAEROBE
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
Y
END OF AEROBE CATALOG?
N
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM
R&R SYSTEM
Regeneration Example
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT
A
© T.E. BEARDEN 1998

Figure 29. Block diagram of the cellular regeneration system and its overall operation (by the present author).

Rife's great secret: Recursive magnification by iterative serial summation of vacuum engines

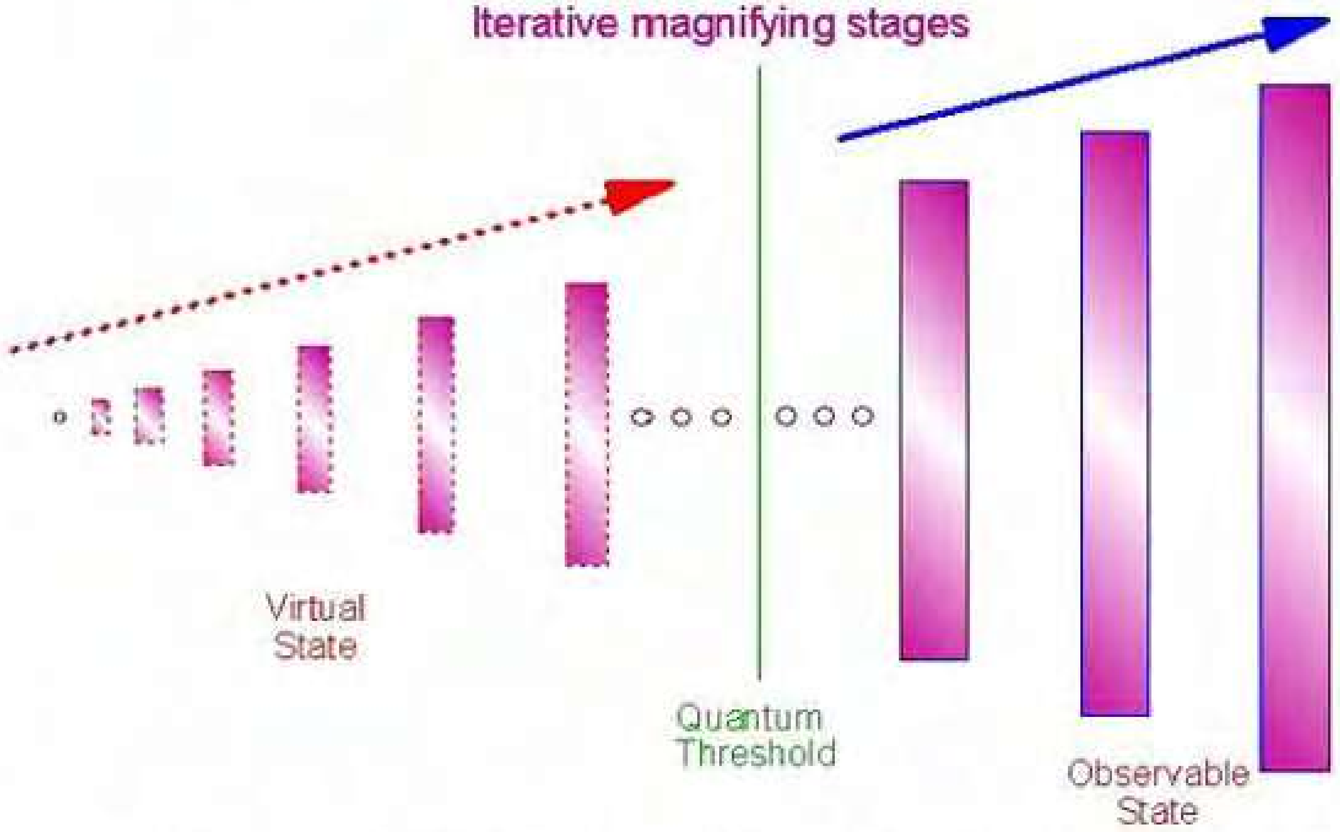
Iterative magnifying stages
Virtual
State
Quantum
Threshold
Observable
State

Figure 30. Rife's great secret: Recursive magnification of the virtual state vacuum engines and their dynamics to the observable state.

Utilizing Scalar Electromagnetics To Tap Vacuum Energy

Floyd Sweet, Association of Distinguished American Scientists
2311 Big Cove Road, Huntsville, Alabama 358010

T. E. Bearden, Association of Distinguished American Scientists
2311 Big Cove Road, Huntsville, Alabama 35801

(Also published in: Proceedings of the 26th Intersociety Energy Conversion Engineering Conference (IECEC Conf),
August 4-9, 1991, Boston, Massachusetts. Vol 4, Advanced Energy Concepts, pp. 370-375.)

Copyright © 1991 by T.E. Bearden & F. Sweet. All Rights Reserved.

ABSTRACT

Based on E.T. Whittaker's previously unnoticed 1903-1904 papers which established a hidden bidirectional EM wave structure in a standing forcefield free scalar potential, a method of directly engineering the ambient potential of the vacuum has been developed and realized experimentally.

Adding Whittaker's engineerable hidden variable theory to classical electro-magnetics, quantum mechanics, and general relativity produces supersets of each discipline. These supersets are joined by the common Whittaker subset, producing a unified field theory that is engineerable and tested.

By treating the nucleus of the atom as a pumped phase conjugate mirror, several working model energy units have been produced which excite and organize the local vacuum, increase the local virtual photon flux between local vacuum and nucleus, establish coherent self-oscillations between the local excited vacuum and the affected nuclei, utilize the self-oscillating standing wave for self-pumping of the nuclei/mirrors, introduce a very tiny signal wave to the mirrors, and output into an external load circuit a powerful, amplified, time-reversed phase conjugate replica wave at 60 Hertz frequency and nominal 120 volt sine wave power.

Several models have been built, ranging from 6 watts early on to one of 5 kilowatts. Both closed batteryless systems with damped positive feedback and open loop systems with battery-powered input have been successfully built. Open loop power gains of from 5×10^4 to 1.5×10^6 have been achieved.

Antigravity experiments have also been successfully conducted where the weight of the unit was reduced by 90% in controlled experiments, with a signal wave input of 175 microwatts and an output of 1 kilowatt.

The basic theory of the device is briefly explained and experimental results presented. In the demonstration session, a videotape of one operating open-loop unit with a 1.5×10^6 power gain is planned, as is the demonstration of an actual working model closed-loop system with a nominal rating of 500 watts, and without external power input of any kind.

The units are solid state, with no moving parts. Each of them comprises a unique form of self-powered vacuum triode of extraordinary gain, where the cathode power and plate power are freely furnished by the vacuum, and only a small grid signal need be furnished either from an external power source or by clamped positive feedback from the device's output. The output is negative energy, and some of its unique characteristics are pointed out.

Implications of the experimental application of the Sweet vacuum triode, the Bearden approach to the nucleus as a pumped phase conjugate mirror, and the unified field theory based on Whittaker's engineerable hidden variable scalar EM potential theory are also briefly addressed.

SCALAR ELECTROMAGNETICS

In 1837 Sir W.R. Hamilton said,

"The notion of time may be unfolded into an independent pure science... a science of pure time is possible."

As is well-known, the fundamental units utilized in physics are arbitrary. It is even possible to construct all of physics on a single unit, time. This oddity shows the truth in Hamilton's statement; it is even more odd, because quantum mechanically time is not an observable. This means that the observable world can be modeled completely in terms of the nonobservable, which is essentially what modern quantum mechanics is now doing.

Hamilton viewed his magnificent quaternions as essentially having accomplished the mathematical structuring of time. Maxwell's original EM theory, as is well-known, was modeled in Hamilton's *quaternions*, not in the highly curtailed Heaviside/Hertz vectors erroneously taught today as "Maxwell's theory."

Not a single one of the present so-called "Maxwell's" vector equations ever appeared in a book or paper by James Clerk Maxwell.

For some years the author has worked on an extended electromagnetics theory, involving the scalar component of the quaternion. [Ref. 1]

In Maxwell's original *quaternion* theory, this scalar component often remains when the directional components zero. Further, it then enfolds vectors and functions of vectors inside, in a hidden variable manner. Specifically, the author has patterned a unified field theory concept upon the previously unnoticed but remarkable early work of E.T. Whittaker. [Ref. 2]

In two fundamental papers in 1903 and 1904, Whittaker showed that all present vector EM can be replaced by scalar potential interferometry, and that bidirectional harmonic EM plane wave sets could be used to produce a standing wave of force-field-free potential (Figures 1 and 2).

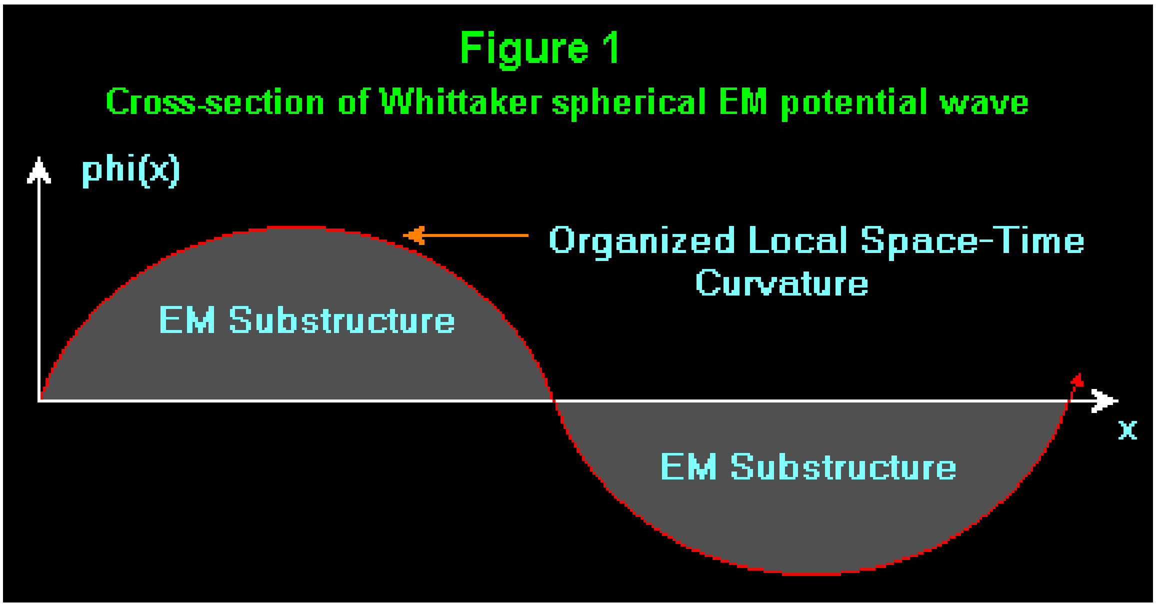
Figure 1
Cross-section of Whittaker spherical EM potential wave
phi(x)
Organized Local Space-Time Curvature
EM Substructure
EM Substructure
x

NOTE: SHAPE MAY BE ALL ABOVE OR ALL BELOW AMBIENT VACUUM POTENTIAL

Thus Whittaker anticipated the quantum mechanical Aharonov/Bohm effect by 55 years, including extending it to the macroscopic world instead of the mesoscopic realm where it has been established to date. [Ref. 3]

In modern terms, Whittaker showed how to turn EM wave energy into electrogravitational potential energy, then how to interfere two such scalar potential waves to recover electromagnetic energy, even at a distance. [Ref. 4]

This unrecognized work is of great importance: when applied to modern physics, it produces supersets of quantum mechanics (QM), classical electromagnetics (EM), and general relativity (GR).

Further, all three extended disciplines unify on their common Whittaker subset, in a testable and engineerable fashion. [Ref. 5]

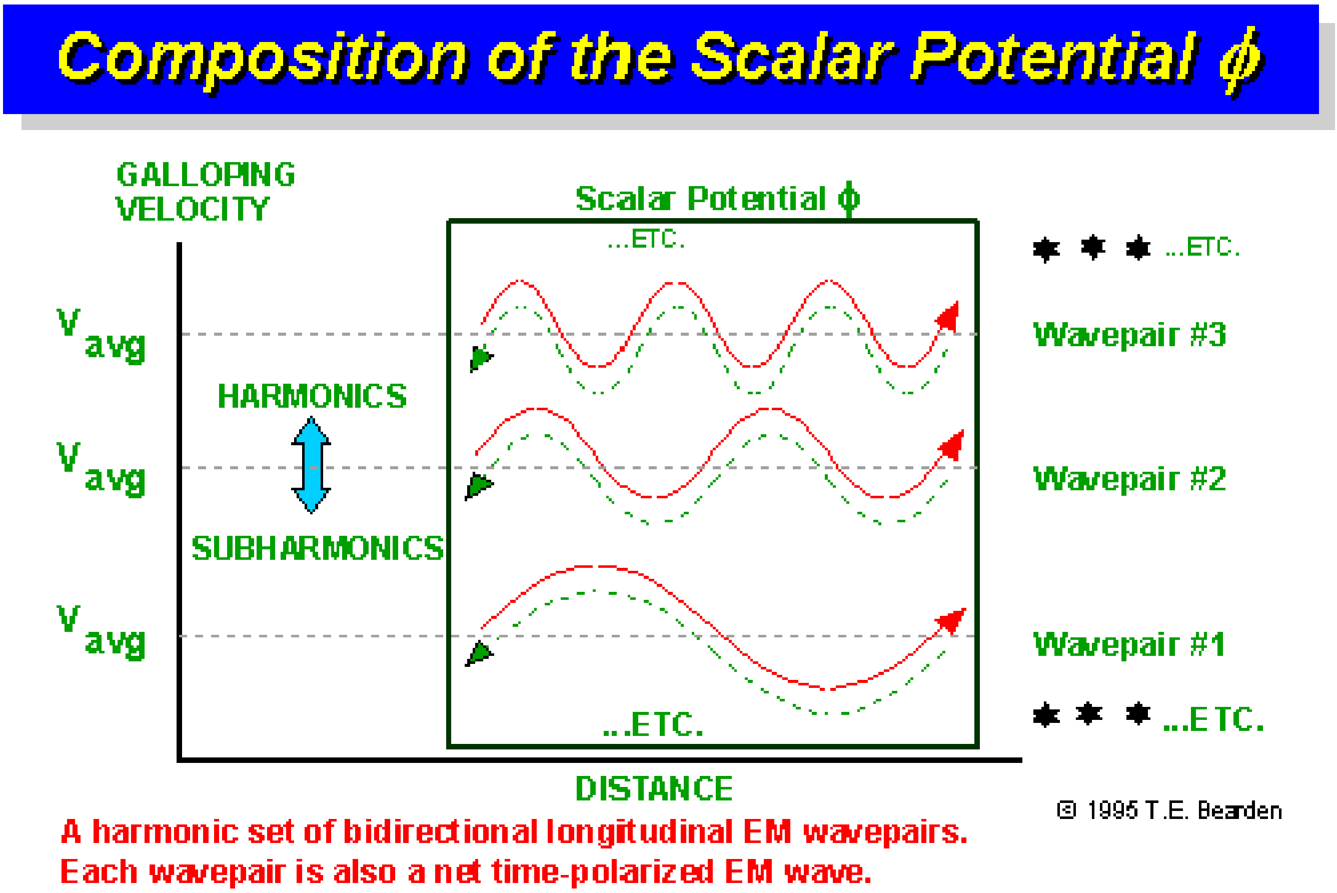
Composition of the Scalar Potential ϕ
GALLOPING VELOCITY
Scalar Potential ϕ
...ETC.
✱ ✱ ✱ ...ETC.
V avg
Wavepair #3
HARMONICS
SUBHARMONICS
V avg
Wavepair #2
V avg
Wavepair #1
...ETC.
✱ ✱ ✱ ...ETC.
DISTANCE
© 1995 T.E. Bearden
A harmonic set of bidirectional longitudinal EM wavepairs.
Each wavepair is also a net time-polarized EM wave.

Figure 2. Infolded EM plane wave structure of a Whittaker wave.

The Nucleus As a PPCM and Triode

The author has also considered the highly nonlinear nucleus of the atom as a pumped phase conjugate mirror (PPCM), having found no other consideration of same in the literature.

The author also dubbed a PPCM a "triode," since the amplified phase conjugate replica of the signal wave is much like the amplified output of a triode, and the signal wave input to a PPCM is much like a triode's grid signal input. The PPCM pump wave then corresponds roughly to the power input to the cathode and plate of a triode. [Ref. 6]

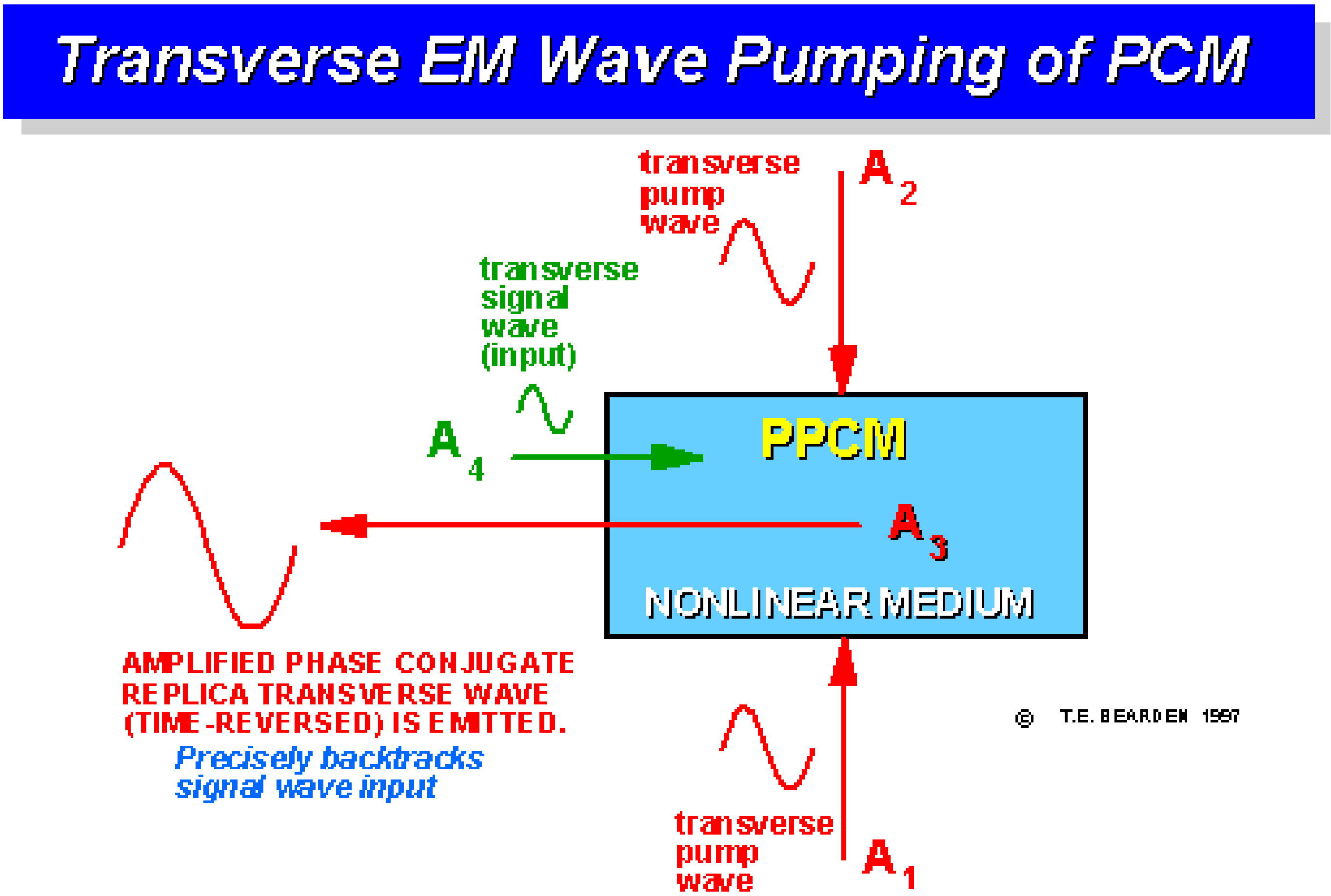
Transverse EM Wave Pumping of PCM
transverse pump wave
A 2
transverse signal wave (input)
A 4
PPCM
A 3
NONLINEAR MEDIUM
AMPLIFIED PHASE CONJUGATE REPLICA TRANSVERSE WAVE (TIME-REVERSED) IS EMITTED.
Precisely backtracks signal wave input
© T.E. BEARDEN 1997
transverse pump wave
A 1

Figure 3. A pumped phase conjugate mirror.

VACUUM TRIODE BACKGROUND

About seven years ago, the author was privileged to see and examine an invention of Mr. Floyd Sweet, that produced about 6 watts of electrical power from the vacuum itself.

This remarkable device, which the author dubbed a *vacuum triode*, in a single unit utilized most of the scalar EM concepts the author had so painfully and slowly formulated over the years.

Sweet, a brilliant inventor with a remarkable knowledge of magnetics, had utilized barium ferrite magnets and special coils to produce a solid-state device that successfully tapped the vacuum energy.

The author quickly put together a theoretical concept for the energy-tapping mechanism, based on treatment of the nucleus as a PPCM and Whittaker's scalar EM potential unified field theory. [Ref. 7]

The author furnished the technical concept, treating the nucleus as a pumped phase conjugate mirror, to the inventor along with copies of Whittaker's papers.

Sweet's Synthesis and Extension

Sweet is also a brilliant EM theoretician, working in four, five, or even six dimensions with ease. He immediately synthesized the entire PPCM and Whittaker theory, and developed a complete theoretical treatment of the device. [Ref. 8]

He also increased the nuclear potential utilized in the activated nuclei of the device, which increased the pumping energy and hence the energy output. He next produced an open-loop vacuum triode (VT) with an output of 500 watts, for an input of 33 microwatts.

Thereafter he produced several other models, including closed-loop systems and one with 5-KW output.

Purpose of This Paper

Our purpose is to explain the detailed scalar EM concept of the operation of the vacuum triode, since it is a universal method for cohering and tapping useful EM energy from the vacuum.

The author believes that this mechanism is the fundamental mechanism that must be invoked in any over-unity device that electromagnetically extracts vacuum energy as electromagnetic effects.

We also intend to demonstrate a full working model of the device at this conference. Sweet's detailed theoretical treatment will be completed and submitted to a major journal shortly, to complete the scientific exposition of the new methodology.

I must also express my deep admiration for my brilliant inventor colleague. It has been a privilege to work with him, though under great difficulties and at a distance. He has developed several other related devices that are of great importance to the emerging new physics of *vacuum engineering*, in the sense referred to by Lee. [Ref. 9]

To mention just one, he has produced a magnetic lens which apparently can directly display the vacuum's virtual particle flux, or a good analog of it. So far as I am aware, this is the only extant instrument today that can perform this feat.

When the vacuum triode has been proven to the scientific community, it is my intention to nominate Sweet for the Nobel Prize he so richly deserves, and seek high scientific endorsements for the recommendation.

INTRODUCTION

Entropy

As is well known, in any closed dynamic system the order existing in it will gradually be dissipated, as more and more interactions occur. This leads to the notion of *entropy* as the increasing disorder in such systems. The assumptions are

1. a closed system, and
2. a positive flow of time for the components of the system.

Actually no such thing as a completely closed system exists in nature. Every mass system is open to virtual particle flux exchange with the vacuum, for example, particularly in the nucleus of its atoms, where the bulk of its mass is located.

However, the closed system assumption is reasonably approximated by a great many systems which are in stable thermodynamic equilibrium, or nearly so.

On the other hand, in an open system far from thermodynamic equilibrium, the second law of thermodynamics does not necessarily apply, because the system violates both the closed system assumption and its equilibrium approximation. [Ref. 10]

Time Reversal

Since being discovered in 1972 in the open Soviet literature, the time-reversed (phase conjugate) EM wave has also been known. The phase conjugate EM wave is truly time-reversed, as has been shown experimentally. Since the time-reversed EM wave violates the second major assumption, the second law of thermodynamics need not necessarily hold for time-reversed entities.

Putting all this together, if one wishes negentropy and hence increased energy in a system, the candidate suggested would appear to be a system that was strongly

1. open loop,
2. time-reversed, and
3. far from thermodynamic equilibrium.

A good overview of time-reversal in physics has been provided by Sachs. [Ref. 11]

Engineering the Nucleus

Since the nucleus already provides a myriad of time-reversed processes, engineering the nucleus of an atom is a very good candidate for practical negentropy.

To engineer the nucleus directly, a Whittaker potential is first artificially constructed, by composing a harmonic set of phase-locked EM wave/antiwave pairs. It is accented that the antiwaves must be true phase conjugates; otherwise they will not constitute a gradient-free Whittaker standing potential wave.

In addition, at least one harmonic interval must be used, and additional harmonic sets are most desirable. The reason is that a space-time lattice must be formed in the vacuum, where the energy is additive spatially but oppositive in t-dot, the rate of flow of time, in the fourth dimension.

So a time-structure is required as well as a spatial structure, which is what is provided by ***n*** bidirectional harmonic Whittaker EM wave sets, where ***n*** is an integer greater than 1.

Once a specific Whittaker structure has been chosen, the local lattice of space-time is established. This establishes phase-locked lattice groupings of coupled photon/antiphoton pairs, or of *gravitons*. In turn, this Whittaker-structured vacuum now contains specific graviton *vacuum engines*, which directly engineer and structure the vacuum's virtual particle flux (VPF) exchange with the nucleus. [Ref. 12]

The nucleus is highly nonlinear, hence strongly phase conjugative, or time-reversed. The ambient potential of the local vacuum surrounding the nucleus is in a violent virtual photon exchange with it, accounting for its electrical charge.

Since the preponderant charge is positive, from the viewpoint of the ordinary light observer whose light interacts with electron shells, the nucleus may be taken to exhibit time reversal (phase conjugation).

Energy, Time, and Gravitons

We take the definition of "energy" to be fundamentally an ordering imposed upon the VPF of vacuum. We take photon scattering from the electron shells of atoms to be the fundamental exterior mechanism producing forward flow of external observer time.

It then follows that "time's arrow" for the EM observer is due to the universal scattering of photons from electron shells.

In this view, forward (positive) time flow and entropy are due to the same primary action: photon scattering from electron shells. It is unfortunate that the concept of "positive" energy has been tied to, and defined in terms of, the scattering and dissipation of VPF order as *work*, or energy expended.

Via the standard labeling, then, negative energy is the reconstitution of order in the vacuum VPF. It should be noted that, in a PPCM, dissipative or external pump wave stress energy can be scavenged and re-emitted in perfect order as the phase conjugate replica. This is a *negentropic* process, for it is capable of turning disorder into order. [Ref. 13]

In the time-reversed PPCM nucleus, we should expect to see appreciable *negative energy*—that is, energy removed from the EM scattering domain. This includes the binding energy of the nucleus, and the gravitational (G) potential energy of the EM energy removed from the "scattering interaction realm" and locked into the mass.

Following Sakharov, we hold that the G-field is not a fundamental field of nature, but a composite caused by, or made from, other fields. [Ref. 14]

To the first order, we assume the G-potential is comprised of coupled photon/antiphoton pairs, on the average, where the statistical coupled spin-2 photon/antiphoton pair is a *graviton*. [Ref. 15]

We follow the modern view of the field: because of vacuum fluctuations, rigorously one no longer speaks of "the" field, but of the *probability of a particular field configuration*. [Ref. 16]

We also hold the vacuum to be composed of potentials, and regard the three terms *space-time, vacuum*, and *scalar potential* as essentially synonymous.

Newton's Third Law and the Detection Process

The VPF EM stress of the local vacuum immediately surrounding the nucleus may be decomposed a la Whittaker into opposing bidirectional EM plane waves/forces. Thus the nonlinear nucleus may be regarded as a pumped phase conjugate mirror, normally with a gain of one.

In this view, Newton's third law reaction force is generated because the so-called "photon" interaction with an atom is in fact a graviton reaction involving a photon/antiphoton pair that is decoupled.

The decoupled photon normally is absorbed and reradiated by an orbital electron, while the decoupled antiphoton interacts with the nucleus, producing a time-reversed twin of the external force—or, in other words, Newton's third law reaction force, which gives a slight recoil of the nucleus.

Half of every measurement physicists normally make is discarded, with the missing half accounting only for Newtonian reaction in the meter or instrument, which is usually ignored.

The fact that half of our measurement interactions are ignored is occasionally discovered and noted by physicists, who may even write a paper pointing it out, but no change is instituted in the foundations. [Ref. 17 and Ref 18]

Semiconducting Vacuum and Self-Oscillation

The vacuum immediately surrounding the nucleus is structured by the nucleus, and is itself nonlinear and capable of acting as a *semiconductor*. [Ref. 19]

Since both this immediately local semiconductor vacuum and the nucleus it surrounds are highly nonlinear, then *nonlinear resonance* can conceivably be established between them.

Further, since the resonating system in such case is an open system away from thermodynamic equilibrium, the oscillation can be self-sustaining. Such self-oscillation of the pumping of a PCM is already well-known in the nonlinear optical literature, particularly with compounds containing barium. [Ref. 20]

THE VACUUM TRIODE

The Basic Concept

As is well-known, a stress can be decomposed into opposing sets of forces. But quantum mechanically, the forces we are interested in with our work here are all caused *electromagnetically*, by the exchange of virtual photons. Even mechanical force, according to QM, is caused in this manner. Thus opposing electromagnetic or mechanical "stress" sets of bidirectional EM forces are microscopically equivalent to the notion of *pump waves* in nonlinear optics.

Hence under the proper conditions, it follows that trapped EM stress energy of the vacuum can be utilized to "pump" the nucleus.

Treating the stress-pumped nonlinear nucleus as a PPCM, it follows that the stress energy of the vacuum can be tapped by a 4-wave mixing mechanism in the atomic nucleus, to provide amplified phase conjugate EM wave outputs from the atom in response to small signal wave inputs. [Ref. 21]

In the proper nonlinear material, the material may act as a PPCM, in which case there exists a suitable connection between the material's atomic nuclei and its external electromagnetic lattice bonds, and the amplified phase conjugate replica wave generated in the nucleus will be emitted from the material as an EM wave field. This field can then be tapped by suitable means and output to an external load circuit.

Block Diagram of the Vacuum Triode

Figure 4 shows a basic block diagram of the vacuum triode process, utilized by Sweet in several laboratory vacuum energy devices. These devices have ranged from a nominal 500 watt output in a 6-lb. device to 5 kilowatts for a heavier unit. Gains have ranged from 50,000 to 1,500,000 for open-loop systems. Both open-loop and closed-loop systems have been built and tested.

Figure 4 shows a combined block diagram for either a closed-loop or open-loop system. In the open-loop system, a barium ferrite magnetic material is used as a pumped phase conjugate mirror.

In the "standard" design, two opposing PPCMs are used. The advantage of this dual combination is the use of *self-targeting* (repetitive phase conjugation, signal by signal). This has the effect of

1. stabilizing the Whittaker field, and
2. producing a quantum potential between the two mirrors, so the mirrors and the Whittaker potential between them are essentially one single space-time entity.

Discussion of a quantum potential is beyond the scope of this paper, but the technical mechanism for creating one has been previously presented by the author on several occasions. [Ref. 22]

First we will explain the open-loop operation of the vacuum triode. In Figure 4, on the right an external 60 Hz, nominal 10 volt AC sine wave of several tens of microwatts in power is input into the stabilized field of the barium ferrite magnet structure, where it modulates the field, producing a signal wave input into the atoms of the material.

In the top right block, the EM signal wave interacts with the electron shell of an atom, which in turn is EM-coupled to the nucleus.

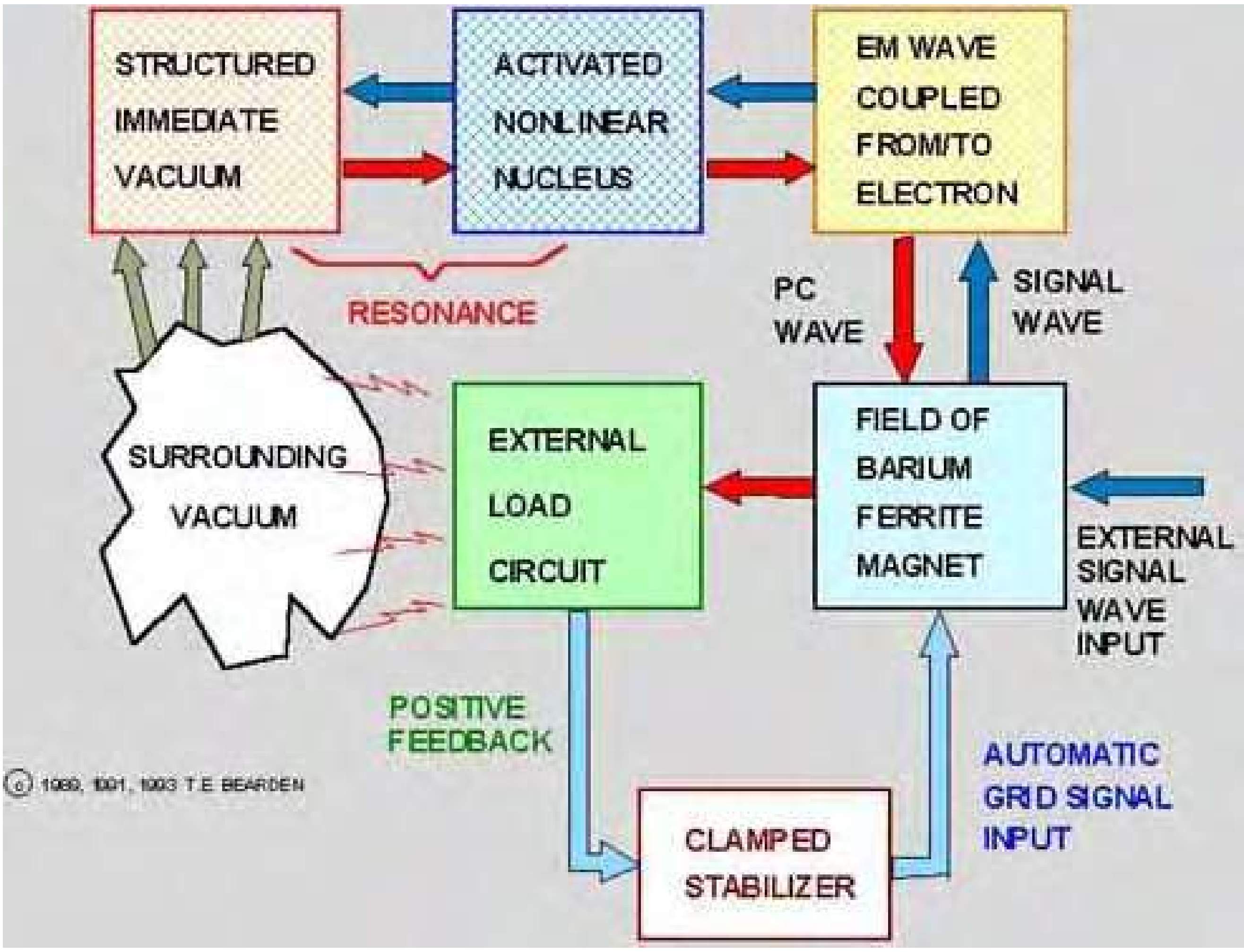
STRUCTURED IMMEDIATE VACUUM
ACTIVATED NONLINEAR NUCLEUS
EM WAVE COUPLED FROM/TO ELECTRON
RESONANCE
PC WAVE
SIGNAL WAVE
SURROUNDING VACUUM
EXTERNAL LOAD CIRCUIT
FIELD OF BARIUM FERRITE MAGNET
EXTERNAL SIGNAL WAVE INPUT
POSITIVE FEEDBACK
AUTOMATIC GRID SIGNAL INPUT
CLAMPED STABILIZER
© 1989, 1991, 1993 T.E. BEARDEN

Figure 4. The Vacuum Triode Concept.

Thus an EM signal wave is input to the nucleus of the atom, which is highly nonlinear (middle top block). Earlier, Sweet had specifically conditioned the atomic nuclei with a proprietary process, wherein in the barium nucleus a trapped 60 Hz scalar EM spherical wave resonance (self-oscillation) exists between the structured semiconductor vacuum immediately surrounding the Ba nucleus.

In the same *activation* process, the ambient potential of the surrounding vacuum was raised and stabilized, in the two leftmost blocks.

At this point the nonlinear nucleus is effectively self-pumped by the trapped, excited, spherical scalar wave oscillation between the structured semiconductor vacuum and the nucleus. The nucleus is now a strongly pumped phase conjugate mirror.

Consequently, when the signal wave input arrives, the PPCM nucleus emits an amplified phase conjugate replica (PCR) wave, which precisely backtracks the input signal wave. This precise backtracking (perfect retroreflection) is referred to as the "distortion correction theorem."

In short, the powerful PCR wave returns precisely toward the external source, passing through the electron shells and arriving in the perturbed barium ferrite magnet assembly field, where it perturbs the field.

A transformer-like system then extracts this magnetic field perturbation and conducts it to the external load circuit.

However, the PCR contains *negative* energy. Short of the load, the internal circuits run cool, rather than heating. This is a signature of a true vacuum energy tapping device.

Indeed, if the output leads of the Sweet vacuum triode are physically shorted together, a brilliant flash occurs, and the leads instantly ice as if dipped in liquid oxygen. This is another signature of the true negentropic over-unity vacuum tap.

Note that the energy extracted from the semiconducting vacuum adjacent to the nucleus is just instantly replaced by the surrounding vacuum's inexhaustible energy pool. This is an open-loop system, with a hidden energy source: the intense virtual particle flux of the vacuum's ambient charge.

It is not possible to exhaust that flux, which is often calculated to have an energy density of some 10^{100} or more grams per cm^3, if the energy were cohered and condensed into mass.

As can be seen, even a VT gain of 1.5×10^6 represents a "vacuum tap" of an incredibly small efficiency, on the order of 10^{-100} or so. However, the vacuum "river" is so energetic that such efficiencies are quite sufficient.

We accent that the barium ferrite magnetic material must be *activated* so that stable self-oscillation between the barium nucleus and the surrounding semiconductor vacuum exists. Although self-oscillating/self-pumped PCMs are known at optical frequencies, Sweet has discovered and perfected a brilliant methodology for activating PPCM nuclei at ELF frequencies.

In a resistive load such as light bulbs, the resistive material accomplishes repetitive phase conjugation. Thus in the resistor, half the total energy is expressed as photon or dissipative energy in the external (electron shell) level.

As the excited electrons decay, they emit scattered EM energy as light and heat. This is an *exothermic* interaction. The other half of the total energy reacts in the atomic nuclei, as a phase conjugative or *endothermic* interaction.

We strongly accent that, Whittaker-wise, there are two electromagnetic channels and two kinds of EM:

1. *external* EM, the common electron-shell interacting, entropic, scattering, time-forward kind, and

2. the *internal*, unsuspected, hidden variable, nuclei interacting, negentropic, reordering/ convergent kind.

Internal EM travels strictly between atomic nuclei, normally not reacting with electron shells unless a pumped phase conjugate mirror reaction is invoked in the nucleus to produce a gain somewhat greater than unity.

Thus if we wish to communicate with atomic nuclei directly, and engineer them directly, we must utilize the internal EM channel via applied Whittaker methods.

Antigravity Tests

Inherent in the preceding discussions is the possibility to turn EM energy into gravitational energy of either sign. In other words, one should be able to utilize Sweet's vacuum triode to produce and demonstrate antigravity.

Indeed this is the case. Sweet has also discovered the special alterations necessary to perform straightforward transformation of the internal energy in the nucleus to antigravitational energy, producing a unilateral thrust upward.

Note that the bulk of G-potential gradient (G-force-field) occurs Whittaker-wise at ELF frequencies. This explains why nonlinear phase conjugate opticians do not notice direct antigravity effects.

At the optical frequencies at which they work, the effects are so miniscule that they are negligible. This is readily explained as follows: In QM, the quantum (photon) is comprised of action (angular momentum), not just energy. It is rather like a "piece of energy welded to a piece of time, with no seam in the middle."

Since quantum change occurs in quanta, the decoupling of the energy and time components, in the continual interaction of photons with matter, exchanges energy between G-potential of vacuum and trapped mass of the atom or particle. In this exchange, small increments of time are continually being formed (and unformed, as photon emission occurs).

Consequently, each mass is moving forward in time in small incremental jumps, usually of exceedingly small magnitude. However, the energy and time trapped in a photon are canonical. The greater the piece of energy, the smaller the piece of time, and vice versa.

So if one wishes to stress the "rate of flow of time" significantly, one needs to produce large amounts of photons that have very large pieces of time, and consequently little pieces of energy.

Since the energy of the photon is directly proportional to its frequency, this means that the lower frequency photons have larger time increments, and hence endure over many "regular-sized photon absorption/emission changes" to appreciably stress the rate of time flow/production.

The bottom line is that the standard pumped phase conjugate mirror can be adapted to produce antigravity at ELF frequencies, but precisely the same adaptation at optical frequencies will have negligible effect.

With this in mind, the author requested Sweet to perform an antigravity experiment to prove the thesis. With Sweet's proprietary adaptation of his vacuum triode/PPCM, the experiment produced rather straightforward but spectacular results, as shown in Figure 5.

The experiment was performed as follows: Rigged for antigravity, the 6-lb. device was placed on a scale so that its weight could be continuously monitored. A special external load box was utilized in which multiple electric light sockets were connected in parallel.

Then the external load draw was adjusted by merely screwing in 100-watt lamps, one at a time, with measurement and observation pauses in between. The output of the device was 120 volt, negative AC sine-wave power at 60 Hz.

For each 100-watt increment, the load power was recorded and the weight was carefully recorded. The results are shown in the rather smooth, classic curve shown in Figure 5.

At 1,000 watts load draw, the previously 6-lb. device had reduced its weight due to gravity by 90 percent. At that point the signal-wave (grid) input to the open-loop vacuum triode was only 175.4 microamps at 10 volts, or just under two milliwatts.

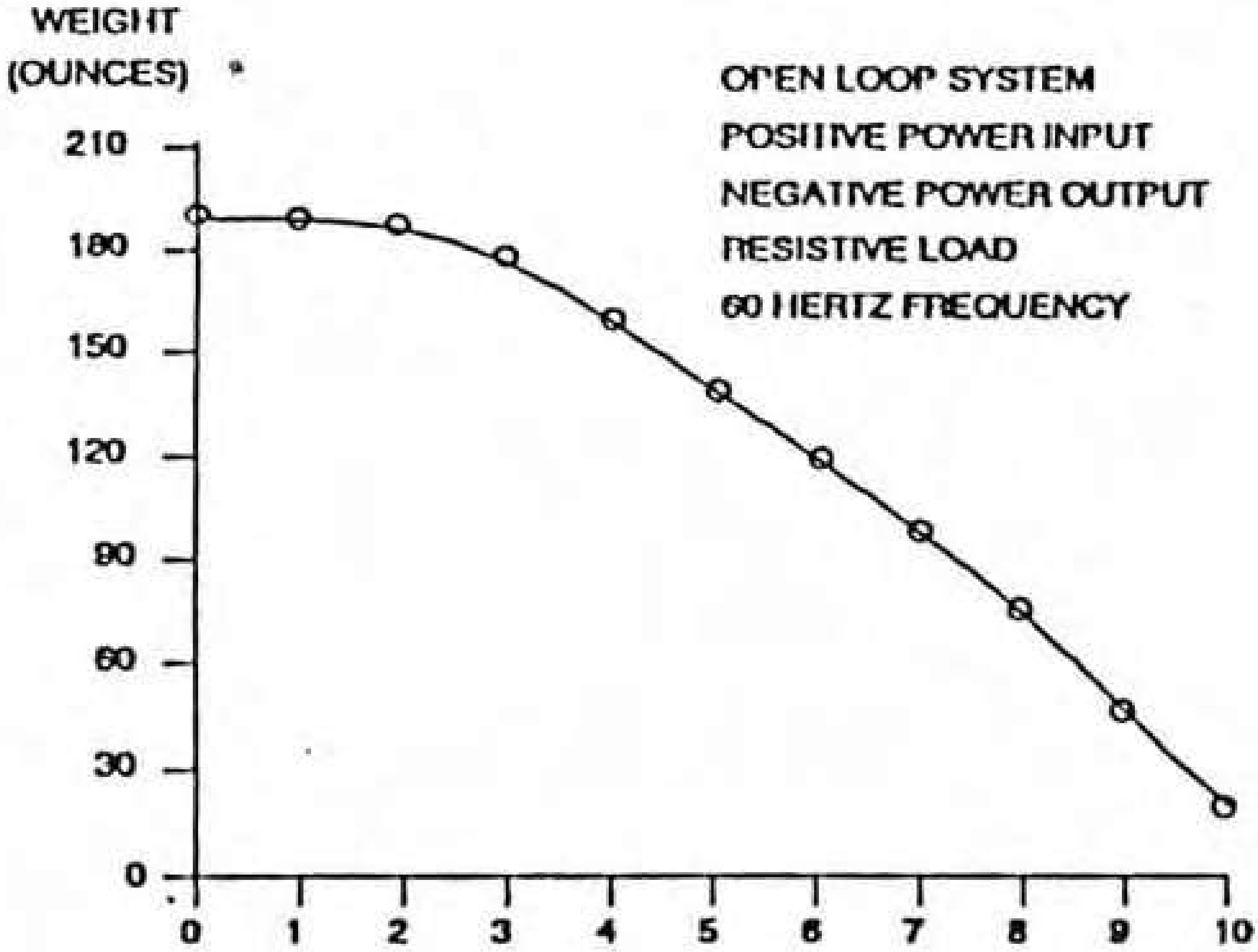

EXTERNAL ELECTRICAL LOAD

(HUNDREDS OF WATTS)

Figure 5. Antigravity Test of Sweet's Adapted Vacuum Triode.

We accent that the nominal two milliwatt input is only a gating signal. It is the organized, gated vacuum energy that is performing the action.

The experiment was stopped short of actual hovering and flying due to safety considerations. With the specific adaptation, magnetic monopoles are deposited in the magnet materials, producing internal tensile stress.

Since no explosive controlled facilities were available, and there was no wish to destroy the VT, the experiment was terminated at 90% antigravity performance.

It was completely successful, and adequately demonstrated the validity of the unified field theory concepts utilized in our approach.

CONCLUSIONS

A new unified field theory has been developed, tested, and at least partially verified experimentally.

The concepts of the theory have been applied by Sweet in a series of inventions that produce readily usable, safe electromagnetic power directly from the vacuum.

The methodology lends itself to formulation of power devices without moving parts. Antigravity, predicted by the concepts of the theory, has been demonstrated in actual practical demonstrations on the laboratory bench.

Though not discussed in this paper, application of the concepts and methodology to a large variety of other fields, such as medical reversal of aging and curing of almost the entire range of present debilitating diseases, has been previously pointed out. [Ref. 23]

We have also pointed out the mechanism for Kaznacheyev's cytopathogenic effect, or the induction of cellular pathology at a distance by electromagnetic means. [Ref. 24]

We have also pointed out the specific mechanism involved in Priore's device, which in rigorous laboratory animal testing under the auspices of eminent French scientists, demonstrated nearly 100 percent cures for terminal cancers and leukemias, sleeping sickness, artherosclerosis, and other debilitating diseases.[Ref. 25]

We believe we have also produced the concepts enabling the direct engineering and therapeutic manipulation of Popp's master cellular control system. [Ref. 26]

We conclude that the concepts we have utilized and experimentally demonstrated are universal, as implied by any notional unified field theory.

Our conclusion is that the concepts, theory, and experiments, taken together, are sufficient for investigation and replication by the scientific community.

If replicated and fully substantiated, we believe the work will directly point the way to, and usher in, a new unified field theory physics of universal application.

SPECULATED IMPLICATIONS

As can be seen, the implications of the new approach are profound. The authors believe they have ushered in the forerunner of a vast new physics, one which will change our lives, and our view of physical reality, in ways previously undreamed of.

By mastering, controlling, and gating the vast, incredible energy of the seething vacuum, we can power our automobiles, flying machines, and technology inexhaustibly. Further, it can be done absolutely cleanly; there are no noxious chemical pollutants to poison the biosphere. With practical antigravity, ships can be developed to cross the solar system as readily as one crosses the ocean today.

And the ships, automobiles, and technology will never run out of fuel; the inexhaustible vacuum fills every system, everywhere, to overflowing.

Not discussed in this paper, it turns out that living systems, faced with the problem of achieving negentropy so as to maintain their form in a dissipative external physical reality, have always used the hidden internal channel for such things as mind, thought, cell control, and living functions.

With the new methodology, one now faces the advent of access and engineering of the mind and life of the observer as readily as the observer's physical body.

Transmutation of the elements, control of the weather, lighting and powering our cities and homes cheaply and cleanly, and provision of plenty for everyone is the vista for the future. We can in fact clean up the radioactive wastes, rid ourselves of coarse nuclear and petroleum powerplants.

We strongly stress that, with the ability to engineer the Schroedinger equation itself, the new methodology allows the direct engineering and control of quantum change, and hence of physical reality itself.

The methodology is extendable to hyperdimensions; nested virtual levels of the vacuum are already precisely that. The author has already pointed out the application of this emerging technology to the absolute cure of diseases such as AIDS, cancer, leukemia, etc., and shown that the Priore device in France already proved the efficacy of the application in the 60s and 70s.

We shall be able to rid ourselves and our descendents of diseases. With direct access to the actual software of life and mind, in the future we should be able to achieve levels of education previously unattainable, by directly inputting the relevant software.

Previously we have also pointed out that four nations of the world are already embarked on weaponization of scalar EM unified field technology. It is sobering to think that, in addition to having the ability to make our planet a paradise for humankind, we also will have the ability to make it a hades.

For that reason, we are doing our best to clarify the technical concept and the theory in this 1991, hopefully with the view that humankind will seize upon the positive aspects, and develop and apply this technology for the betterment of all people everywhere.

Long ago, Albert Einstein said these words:

"It would of course be a great step forward if we succeeded in combining the gravitational field and the electromagnetic field into a single structure. Only so could the era in theoretical physics inaugurated by Faraday and Clerk Maxwell be brought to a satisfactory close."

And Teilhard de Chardin wrote:

"Someday, after we have mastered the winds, the waves, the tides and gravity, we shall harness for God the energies of love. Then for the second time in the history of the world man will have

discovered fire."

The authors fervently believe they have come upon fire for the second time, as allegorized by de Chardin. If so, let us all use the knowledge wisely.

REFERENCES

1. Maxwell's original quaternion EM theory is contained in some 200 quaternion equations and differs extensively from the restricted Heaviside/Gibbs vector interpretation universally taught today as "Maxwell's Theory." See James Clerk Maxwell, *A Treatise on Electricity and Magnetism*, Oxford University Press, Oxford, 1873. The third edition is by Dover, 1954.

2. E.T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," *Math. Ann.*, Vol. 57, 1903, p. 333-355; "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," *Proc. Lond. Math. Soc.*, Series 2, Vol. 1, 1904, pp. 367-372. The first paper was in fact a detailed theory of the scalar standing potential wave that Nikola Tesla discovered four years earlier, on the night of July 3-4 1899, being radiated from thunderstorms, which he entered in his *Colorado Springs Notebook* on the morning of the 4th. More recently the Whittaker structure (WS) inside potentials, including the Schroedinger potential, has been shown by V.K. Ignatovich, "The Remarkable Capabilities of Recursive Relations," *Am. J. Phys.*, 57(10), Oct. 1989, p. 873-878, without credit to Whittaker or to the presence of such structured scalar potentials in Maxwell's original quaternion EM theory. These WSs are universal to scalar potentials; e.g., for the same in acoustics, see Richard W. Ziolkowski, "Localized Transmission of Wave Energy," *Proc. SPIE, Vol. 1061, Microwave and Particle Beam Sources and Directed Energy Concepts*, Jan. 1989, p. 396-397.

3. See Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," *Phys. Rev.*, Second Series, 115(3), Aug. 1, 1959, p. 458-491. This paper pointed out the primacy of the potentials. Instead of being causative agents, the force fields are actually effects generated from the potentials. This is in complete violation of both classical electromagnetics and classical dynamics, but it is absolutely required by quantum mechanics. For an extensive discussion of the Aharonov-Bohm effect and an extensive list of references, see S. Olariu and I. Iovitzu Popescu, "The Quantum Effects of Electromagnetic Fluxes," *Rev. Mod. Phys*. 57(2), Apr. 1985. See Bertram Schwarzschild, "Currents in Normal-Metal Rings Exhibit Aharonov-Bohm Effect," Physics Today, 39(1), Jan. 1986, p. 17-20 for confirmation that the Aharonov/Bohm effect has indeed been firmly proven experimentally.

4. It is pointed out that today all potentials are well-known to be gravitational entities. However, this was not known in Whittaker's time, and so he himself did not realize that he had actually produced an engineerable, testable unified theory of electromagnetics and gravitation.

5. T.E. Bearden, "Possible Whittaker Unification of Electromagnetics, General Relativity, and Quantum Mechanics: Part I: Background," Presented to *Ala. Acad. Sci. Annual Symp.*, Univ. Jacksonville, Mar. 1991.

6. For the theory of a pumped phase conjugate mirror, see David M. Pepper, "Nonlinear Optical Phase Conjugation," *Opt. Eng*., 21(2), Mar./Apr. 1982, p. 156-183; Amnon Yariv, *Optical Electronics*, 3rd Edn., Holt, Rinehart, and Winston, New York, 1985. In a normal triode tube, the amplified plate signal is 180 degrees out of phase *spatially* with the grid input, but in-phase with it in respect to *rate of flow through time*. In a PPCM, the amplified phase conjugate replica wave is in phase spatially with the signal wave, but 180 degrees out of phase with it in respect to its rate of flow through time. It is pointed out that a PCM with a gain of unity produces a coupled EM wave/antiwave pair whose energy is additive in 3-space but subtractive in the first derivative of the fourth dimension, time. Hence it is no longer an electromagnetic wave as such, but an oscillatory wave of stress upon the local rate of flow of time. Hence it is a powerful *electrogravitational wave*, whose EM nature is hidden in the guise of a scalar EM potential. It is also a wave of variation in the local gauge, and of variation in the local ST curvature.

7. See T.E. Bearden, *The Phase Conjugate Vacuum Triode*, Apr. 23, 1987, privately published. An earlier pencil draft was initially produced.

8. Sweet has continued to extend his theoretical treatment. A formal paper providing the complete mathematical theory of the vacuum triode has been drafted, is in final review, and will be submitted to a leading journal by mid-1991.

9. See Chapter 25: Outlook, "Possibility of Vacuum Engineering," T.D. Lee, *Particle Physics and Introduction to Field Theory*, Harwood, New York, 1981, pp. 824-828.

10. In 1977 Ilya Prigogine received the Nobel Prize for extending thermodynamics; in particular, for the theory of dissipative structures in nonequilibrium thermodynamics. In Prigogine systems, negentropy is known to be possible.

11. Robert G. Sachs, *The Physics of Time Reversal*, Univ. Chi. Press., Chicago, 1987.

12. Note that this moves the entire notion of the charge of a fundamental particle to a deeper and more extended level. Now the charge may be *discretized*, but it is not quantized in the hard conventional sense. Further, the internal Whittaker structure of the massless VPF photon exchange of vacuum and mass, which, quantum mechanically, is what the electrical charge of the particle is in the first place, is deterministically structured. Note that this violates the present assumption that all like charged particles are identical; now two electrons may have either the same or different magnitudes of charge, and even when the magnitudes are the same, their internal charge structures (Whittaker structures) and VPF exchange with the vacuum may differ. Also note that this resolves the severe QM problem of missing chaos (hidden order) in quantum change. The reason for the problem was the use of Gibbs statistics with its assumption of random variable change, which a priori excluded hidden order (and hence chaos) from QM. That was only a special case, albeit an important one. There are now three QM cases: (1) the conventional case, where there is no hidden order; (2) the case where there is some hidden order, and the statistics is chaotic, not random; and (3) the case where the QM change is deterministic, with essentially total hidden order. Note that the Whittaker

methodology allows one to directly engineer cases (2) and (3), including the Schroedinger equation itself.

13. The present second law of thermodynamics is written only for time-forward entities, and need not apply for the time-reversed case. Merely viewing the energy-dissipating forward time case in reverse allows an appreciation of the time-reversed case. In other words, the second law of thermodynamics is incomplete as presently stated. The complete law has a corollary to cover the increase in order as the time-reversal of the system increases. Thus the complete law consists of two parts: (1) the entropic, time-forward case, and (2) the negentropic time-reversed case. Since the re-ordering can be amplified at will by a PPCM process, the correct distinction between the two subsets of the complete law is important, and applies to real systems.

14. A.D. Sakharov, *Theor. Math. Phys*., Vol. 23, 1975, p. 435.

15. T. E. Bearden, *Gravitobiology: A New Biophysics*, Tesla Book Co., 604 Date Ave., Chula Vista, CA 91912, 1991.

16. C.f. Charles W. Misner, Kip S. Thorne, and John Archibald Wheeler, *Gravitation*, W.H. Freeman and Co., San Francisco, 1973, p. 1191.

17. E.g., see Richard Kidd et al, "Evolution of the Modern Photon," *Am. J. Phys*., 57(1), Jan. 1989, pp. 27-35. See also R. Chen, "Cancellation of Internal Forces," *Am. J. Phys*. 49(4), Apr. 1981, p. 372.

18. A nonlinear material may simply emit a photon, or it may act as a phase conjugate mirror (PCM) and emit a phase conjugate replica of the absorbed photon. When the material emits a normal photon, it measurably recoils. When it emits a time-reversed photon, it does not recoil, as already experimentally established in nonlinear phase conjugate optics. The solution to the mystery is this: When emitting a normal photon, the material does not act as a PCM. In that case the matching antiphoton which split from the interacting graviton (the graviton that yielded the external photon) interacts with the nucleus, producing a recoil action with a gain of one. Thus Newtonian third-law recoil of the nucleus occurs. On the other hand, when the material acts as a PCM, it also emits the antiphoton *outside* the atom to "backtrack" the absorbed "signal wave" photon. In that case there is no Newtonian recoil of the nucleus, because the *agent* for causing recoil did not interact with the nucleus to produce it.

19. Richard E. Prange and Peter Strance, "The Semiconducting Vacuum," *Am. J. Phys*. 52(1), Jan. 1984, p. 19-21. Also, under nonlinear conditions, a particle can absorb more energy than is in the light incident on it, absorbing the energy from the vacuum VPF. C.f. Craig F. Bohren, "How Can a Particle Absorb More Than the Light Incident on It?" *Am. J. Phys*. 51(4), Apr. 1983, p. 323-327.

20. C.f. Pepper, ibid. and Yariv, ibid. For a specific example, see Mary J. Miller et al, *Appl. Phys. Lett*. 41(8), Oct. 15, 1982, p. 689-691.

21. Again, for the theory of the PPCM, see Pepper, ibid. and Yariv, ibid.

22. E.g., see T.E. Bearden, *Gravitobiology: A New Biophysics*, Tesla Book Co., Chula Vista, CA, 1991, p. 33-36.

23. Bearden, *AIDS: Biological Warfare*, Tesla Book Co., 1988; *Gravitobiology: A New Biophysics*, Tesla Book Co., 1991; *Analysis of Scalar Electromagnetics*, Tesla Book Co., 1990.

24. C.f. Vlail P. Kaznacheyev and L.P. Mikhailova, *Ultraweak Radiations in Intercellular Interactions*, [in Russian], Novosibirsk, 1981; Vlail P. Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," *Psi Research*, 1(1), Mar. 1982, p. 47-76; N.D. Devyatkov, Ed., *Applications of Low-Intensity Millimeter Wave Radiation in Biology and Medicine*, [in Russian], IRE Akad. Nauk. SSSR, Moscow, 1985.

25. C.f. Antoine Priore, "Method of Producing Radiations for Penetrating Living Cells," U.S. Patent No. 3,280,816; Jean-Michel Graille, *Le Dossier Priore*, De Noel, Paris, 1984 [in French]; Christopher Bird, "The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science," Appx. I to Bearden, *AIDS: Biological Warfare*, 1988.

26. C.f. Fritz Albert Popp, "Photon Storage in Biological Systems," in Fritz Albert Popp et al, Eds., *Electromagnetic Bio-Information: Proceedings of the Symposium*, Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979, p. 123-149; also *Biophotonen. Ein neuer weg zur Losung des Krebsproblems*, Verlag fur Medizin, Heidelberg, 1976 [in German].

Defense Department Briefing on Priore Treatment to counter Biological Warfare

---

SLIDE LIST

Solutions to Mass Casualties Treatment Problem

Threat and Current Status

1. Title Page
2. Key Points in the New Technology
3. Title Page
4. Title Page
5. The Terrorist Threat
6. The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
7. Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.
8. Present Status of the WMD Threat to the United States
9. Some Present Problems in the WMD Defense Capabilities

Historical Solution: Priore Treatment Therapy

10. Title Page
11. Priore's laboratory device for treating small animals
12. Antoine Priore with the Mayor of Bordeaux
13. Raymond Pautrizel, eminent French parasitologist, who worked with Priore
14. 12-Foot plasma tube in Priore's large device
15. Priore's 3 1/2 stories high device for whole body treatment of humans
16. Mechanism of the Priore therapy
17. Priore's therapeutic methodology: using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain
18. Drastically reduced cumulative deterioration after Priore therapy

Extended Physics and Background

19. Title Page
20. Founders of Scalar Electromagnetics
21. Subdivisions of Soviet Energetics

22. Bunge on the Status of Classical and Quantum Physics
23. Einstein on Reviewing Foundations
24. Fields and Potentials have been "defined" only in and on matter
25. Maxwell's Equations are Matter-To-Matter Transforms
26. Electrodynamicists' Reaction to Removal of the Material Ether
27. The Missing Infolded Electrodynamics
28. Nonlinear Optics Distortion Correction Theorem
29. Infolded longitudinal EM biwave composition of a scalar potential
30. The Zero-Vector Axiom Destroys "Topologies Within Topologies"
31. Stoney-Whittaker Bidirectional Longitudinal EM Wave-Pair
32. Undistorted Progressive Waves (UPWs)
33. Velocity modulating a longitudinal EM wave oscillates rate of flow of time
34. The Graviton is a Coupled Photon/Antiphoton Pair
35. Mechanism for Physical Change and the Flow of Time
36. Any EM Field or Wave Pattern is Produced by interference of Two Scalar Potential Functions (E.T. Whittaker, 1904)
37. Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass
38. Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection
39. The Immune System
40. Popp's master cellular control system
41. Cancer: characteristics
42. The evolutionary path of multicellular organisms
43. Effect of fluid contamination on hemoglobin's oxygen transport
44. First step in Self-Promotion of Cancer due to Hypoxia
45. Becker's Cellular Growth Control
46. Becker's dc treatment of bone fractures
47. Becker's theoretical DC control system involved with responses to injury
48. Becker's theoretical control system governing regeneration
49. Link Between Standing DC Potentials and the Nervous System
50. Becker's Findings on Cellular Injury and Regeneration
51. Corrections For the Present EM Bioeffects Model
52. R&R System - Regeneration Example
53. Kaznacheyev's Cytopathogenic Mirror Effect
54. Microwave Radiation of U.S. Embassy in Moscow

55. Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
56. Quantum potential and hidden variable theory
57. Participants in a quantum potential share a common multiply-connected spacetime (MCST)
58. Gulf War Syndrome and Typical Comparative Disease Curves
59. Force Fields and Symmetry
60. Aspects of Strong Local Asymmetry
61. Other Aspects of Strong Local Asymmetry
62. Small Entities Emit Powerful S-Flows
63. What is the "magnitude" of a potential?
64. Table A1. The Vacuum Energetically Interacts With Every Particle, Continually
65. Extended Superpotential Theory (Adding Vacuum Engines)
66. Vacuum Engine: Jillions of Working ST Demons
67. Active Spacetime and a Specific Vacuum Engine (A General Relativistic Concept and Analogy)
68. Extension of Wheeler's general relativity principle
69. Signal versus Vacuum Engine
70. Electronuclear Interactions by Vacuum Engines
71. Vacuum Engineering
72. Impact of Vacuum Engineering
73. A Dramatic Extension to Nonlinear Optical Pumping
74. Time Conservation and Some Properties (Energy and time are canonical)
75. A Dramatic Extension to Nonlinear Optical Pumping
76. Key Points in the New Technology

Modern WMD Solution: Portable Treatment Systems

77. Title Page
78. Portable Priore-Type Treatment System
79. Portable System Block Diagram
80. Portable Priore Technology Unit
81. Saving the Stricken Civilians

Conclusions: Technology Practical, Proven

82. Title Page
83. Summary and Conclusions

Recommendation: Crash Development Program

84. Title Page
85. Recommend implementing:
86. Final Thoughts

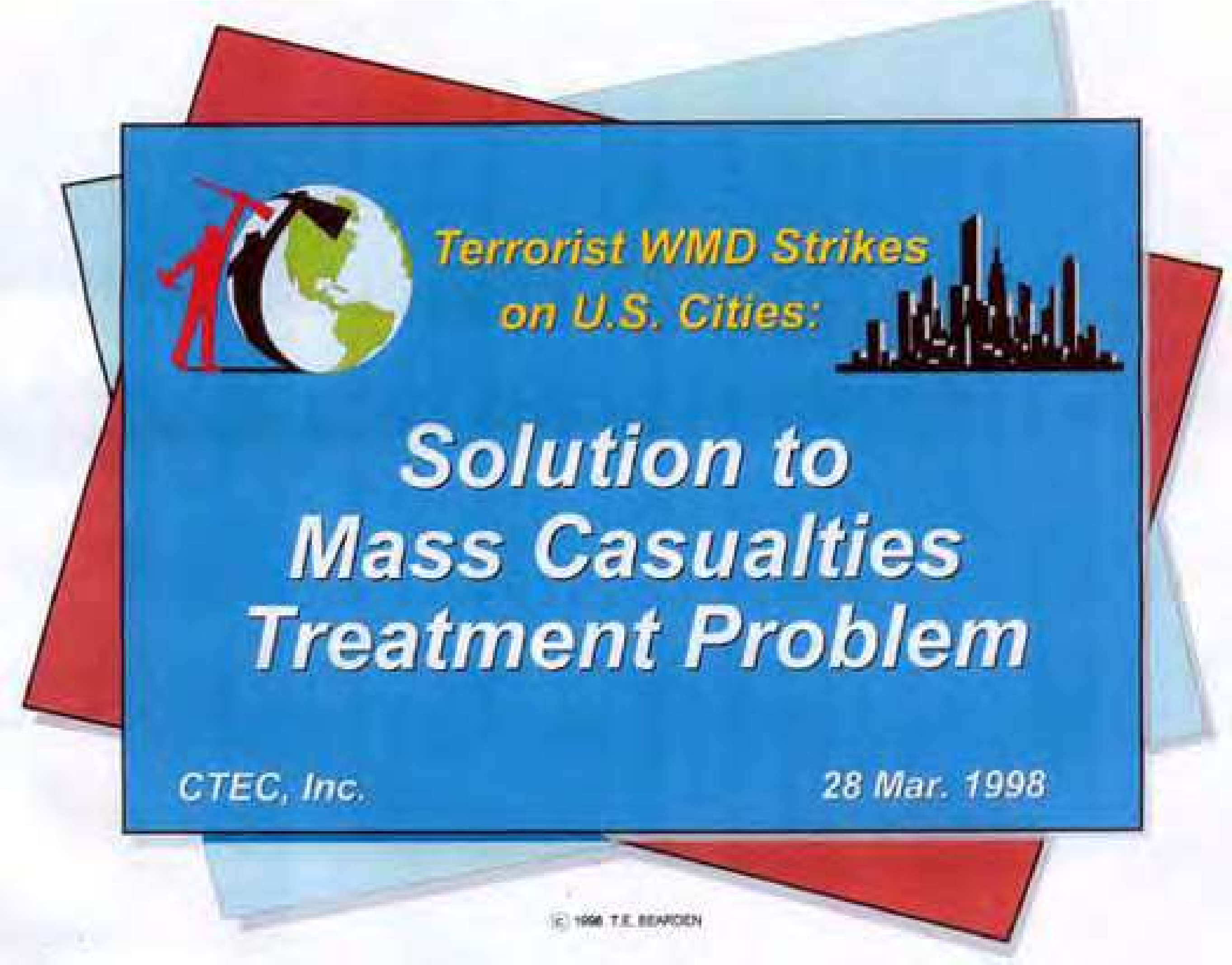

Terrorist WMD Strikes
on U.S. Cities:

Solution to Mass Casualties Treatment Problem

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

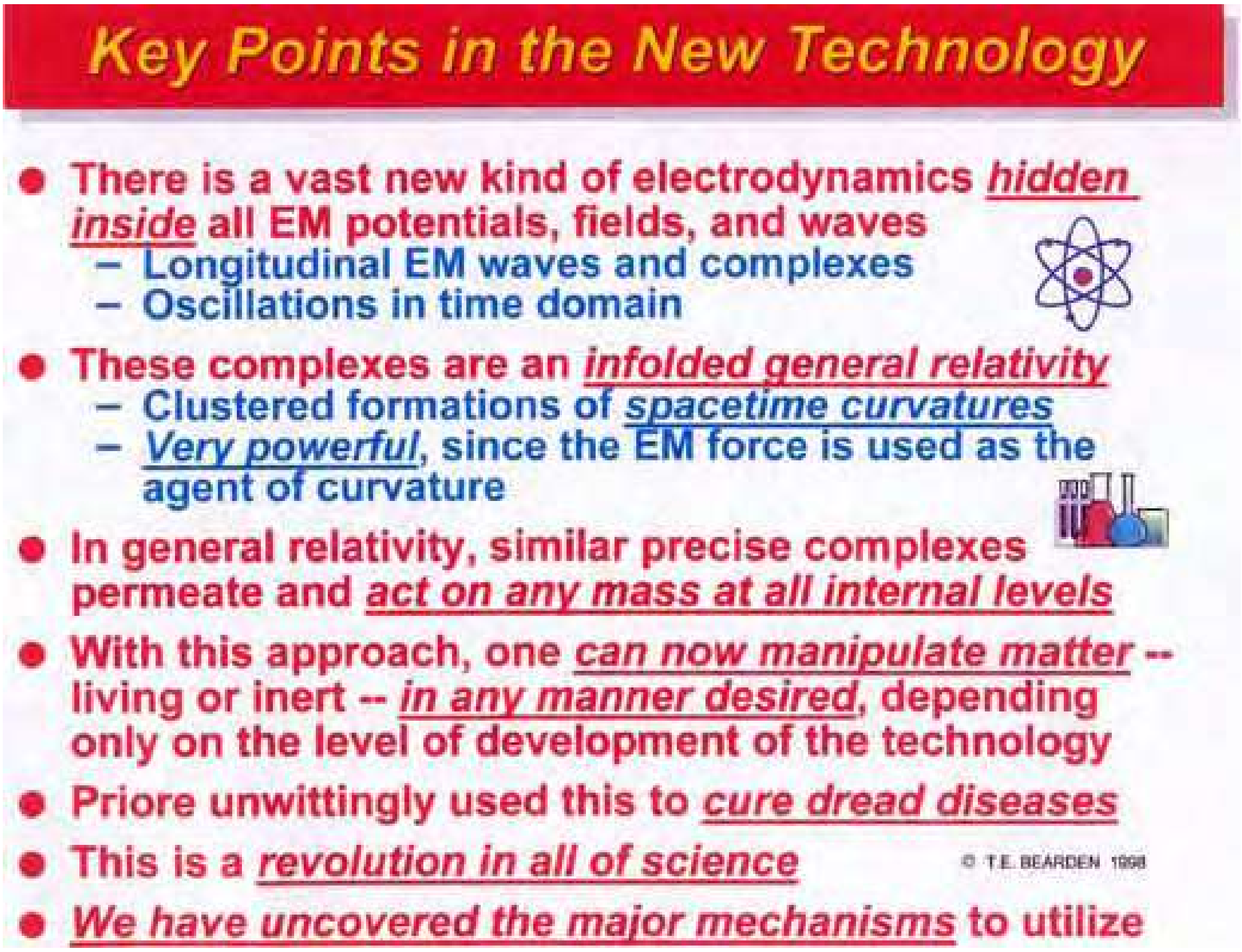

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

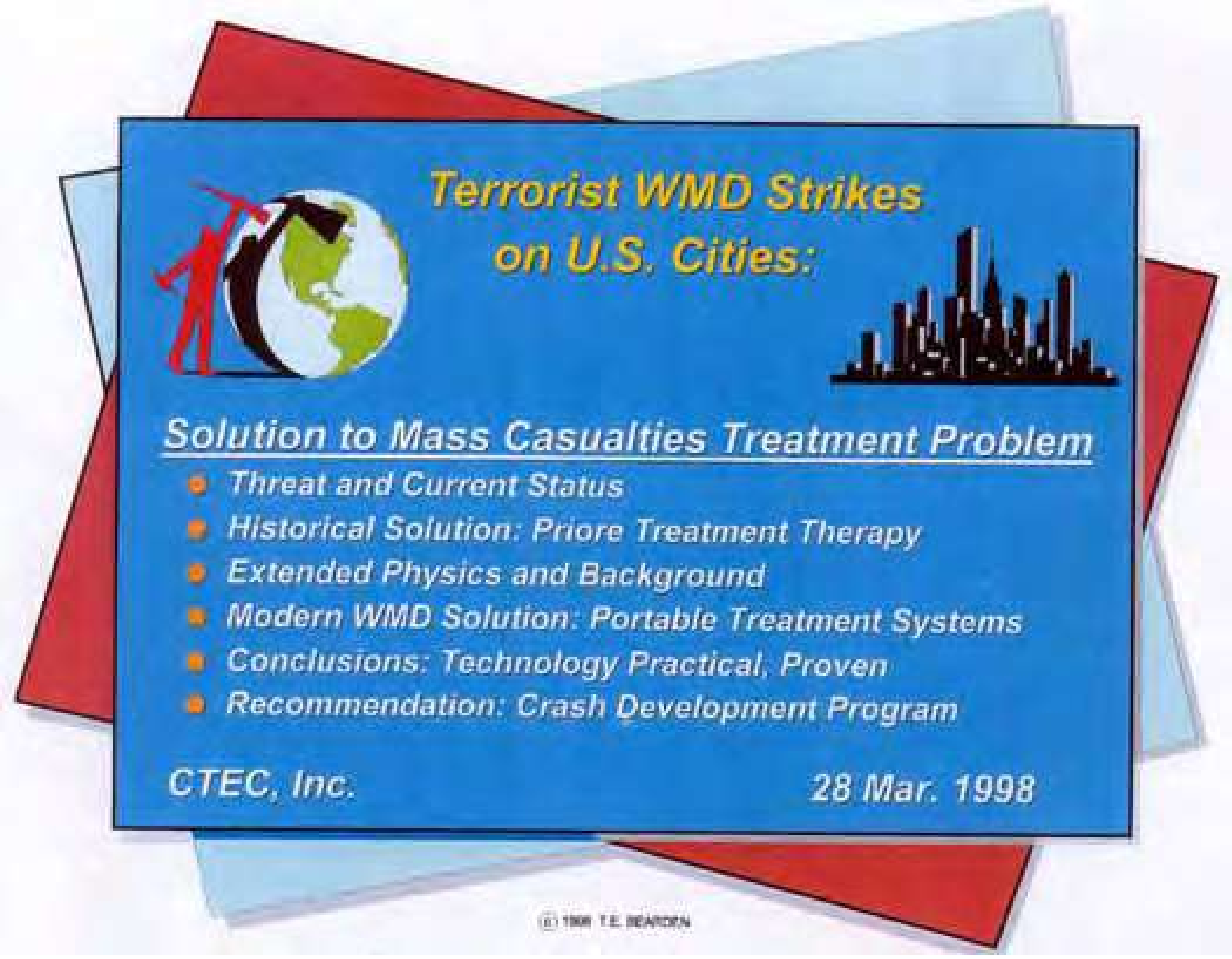

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

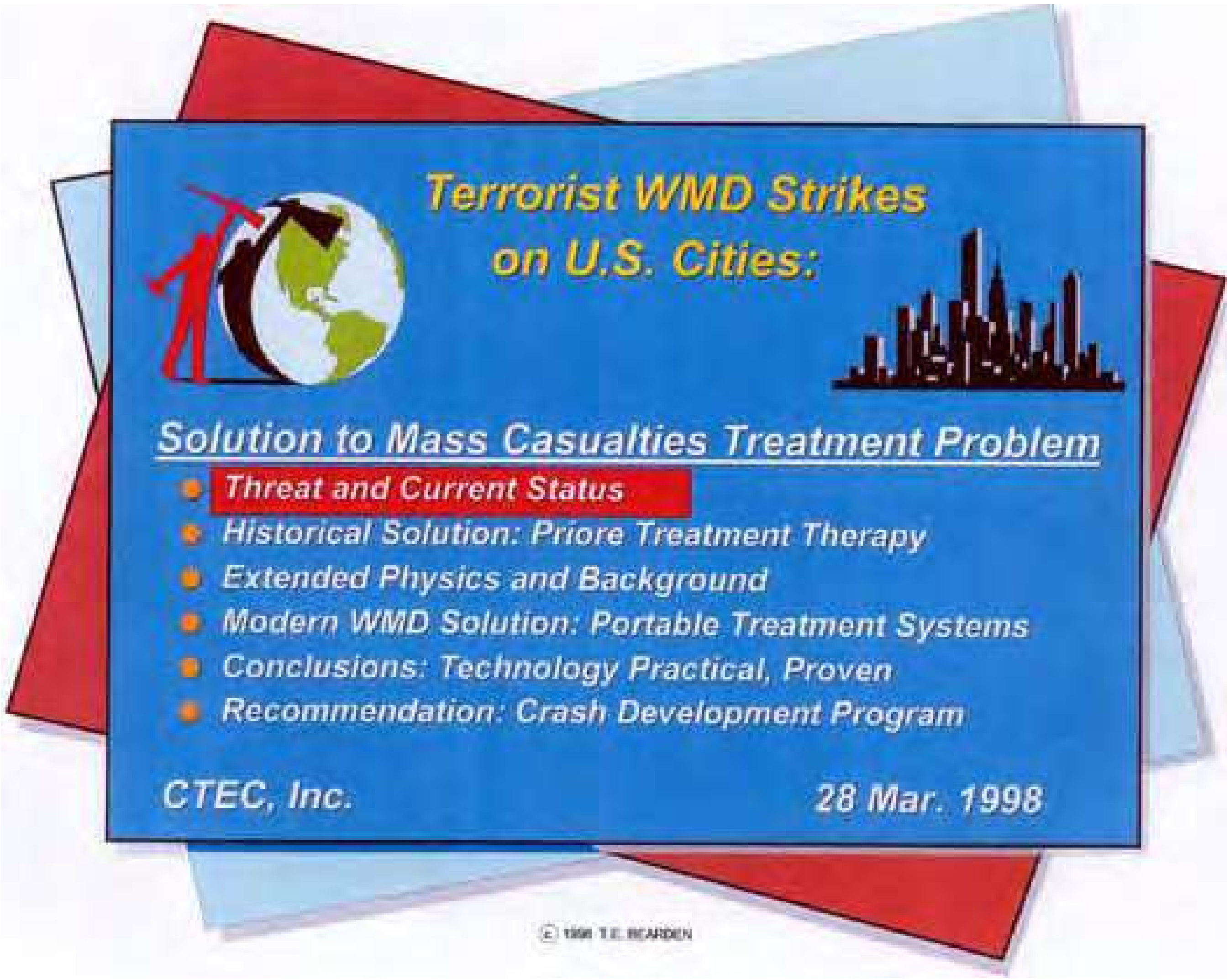

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

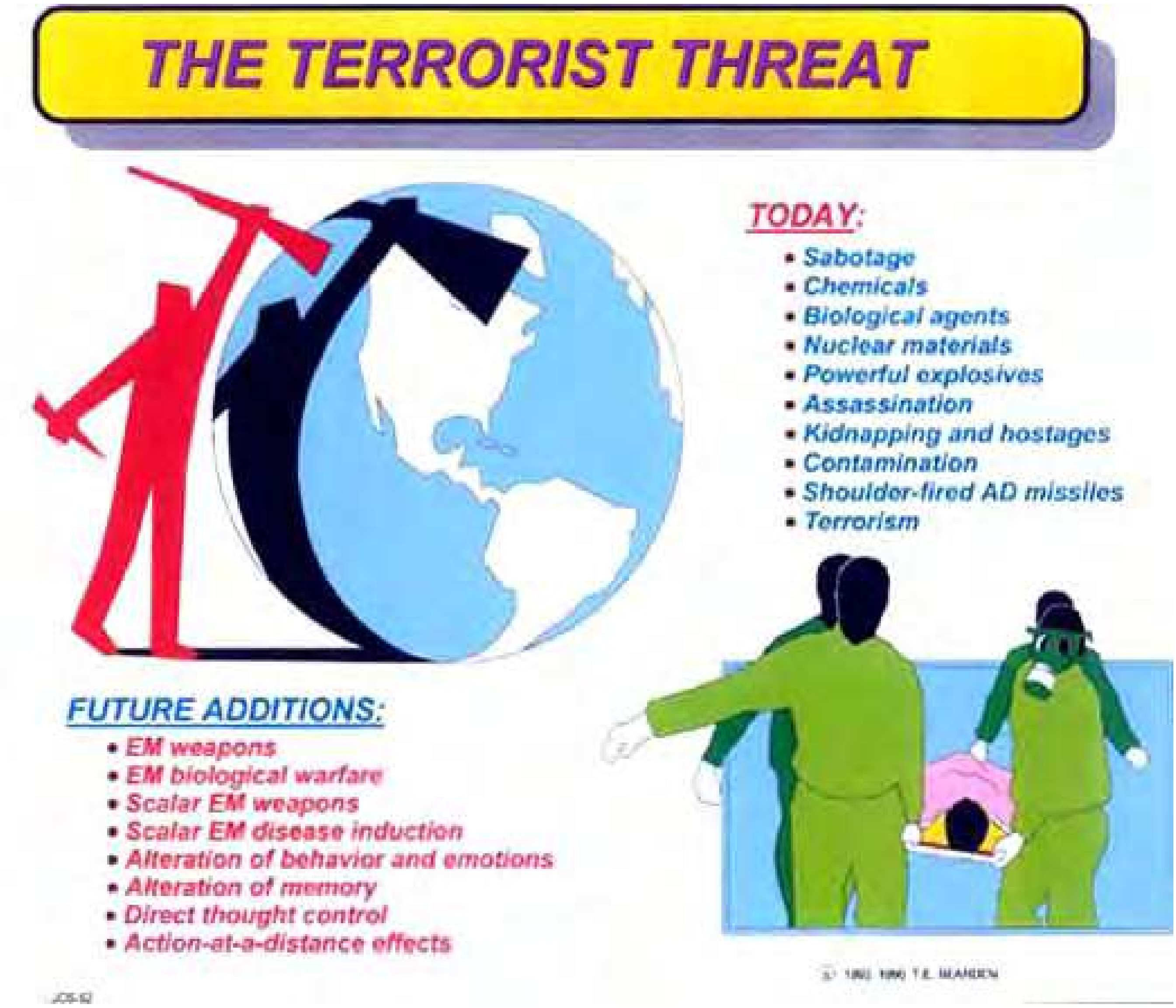

THE TERRORIST THREAT

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- EM weapons
- EM biological warfare
- Scalar EM weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct thought control
- Action-at-a-distance effects

© 1982, 1990 T.E. BEARDEN

JCS-52

Return to Slide Index

Next Slide

Previous Slide

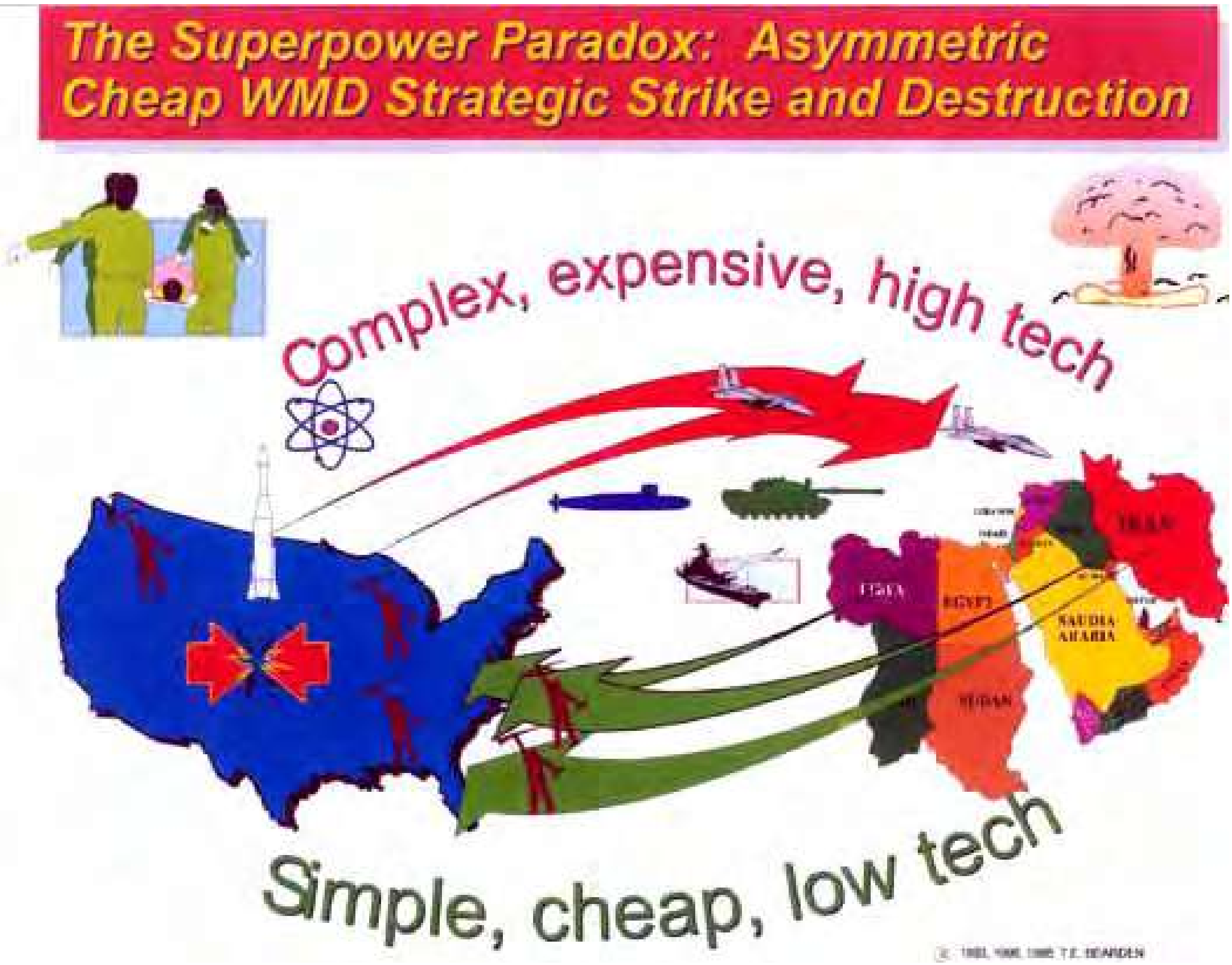

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

Simple, cheap, low tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN

© [illegible] T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

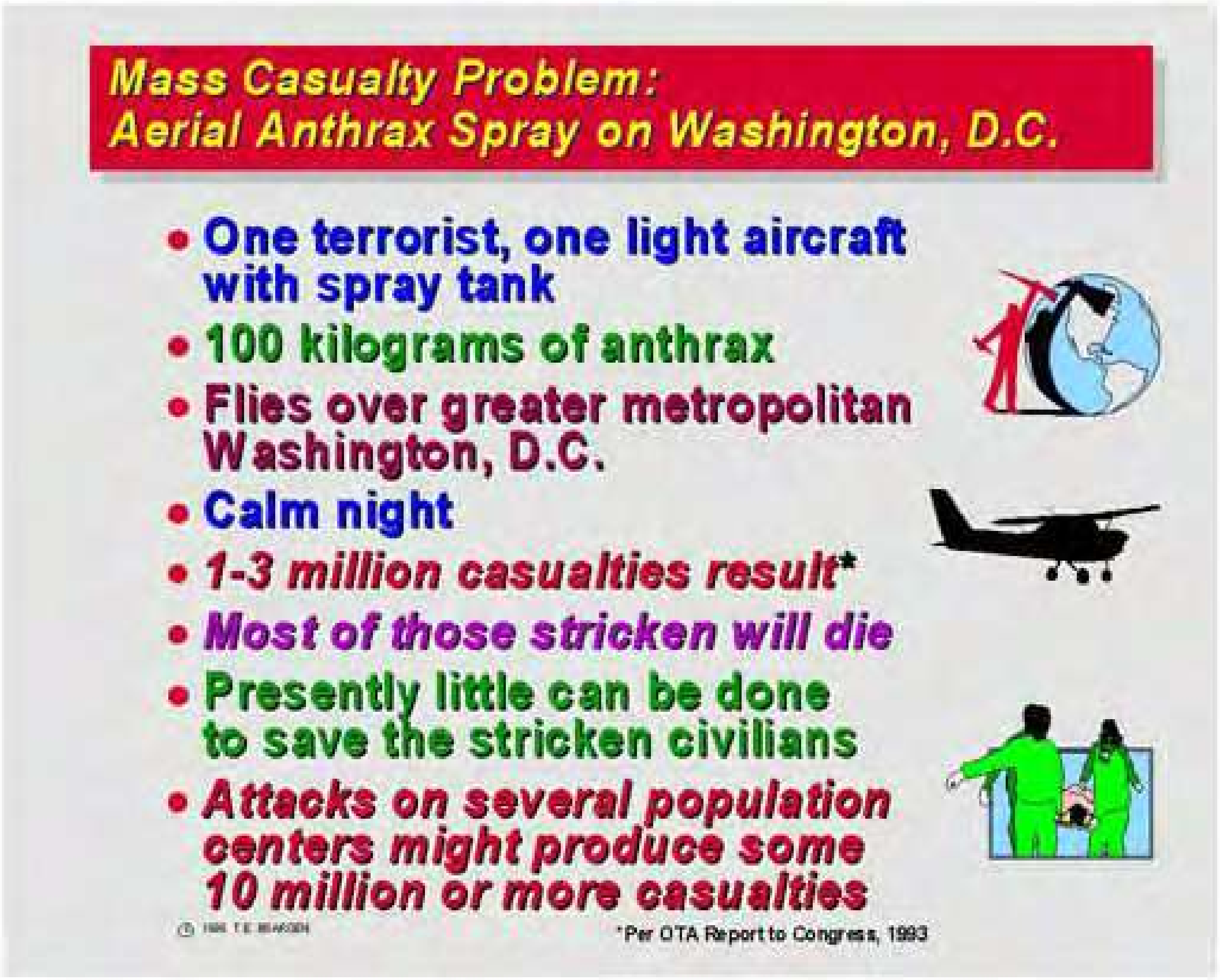

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- 100 kilograms of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

*Per OTA Report to Congress, 1993

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

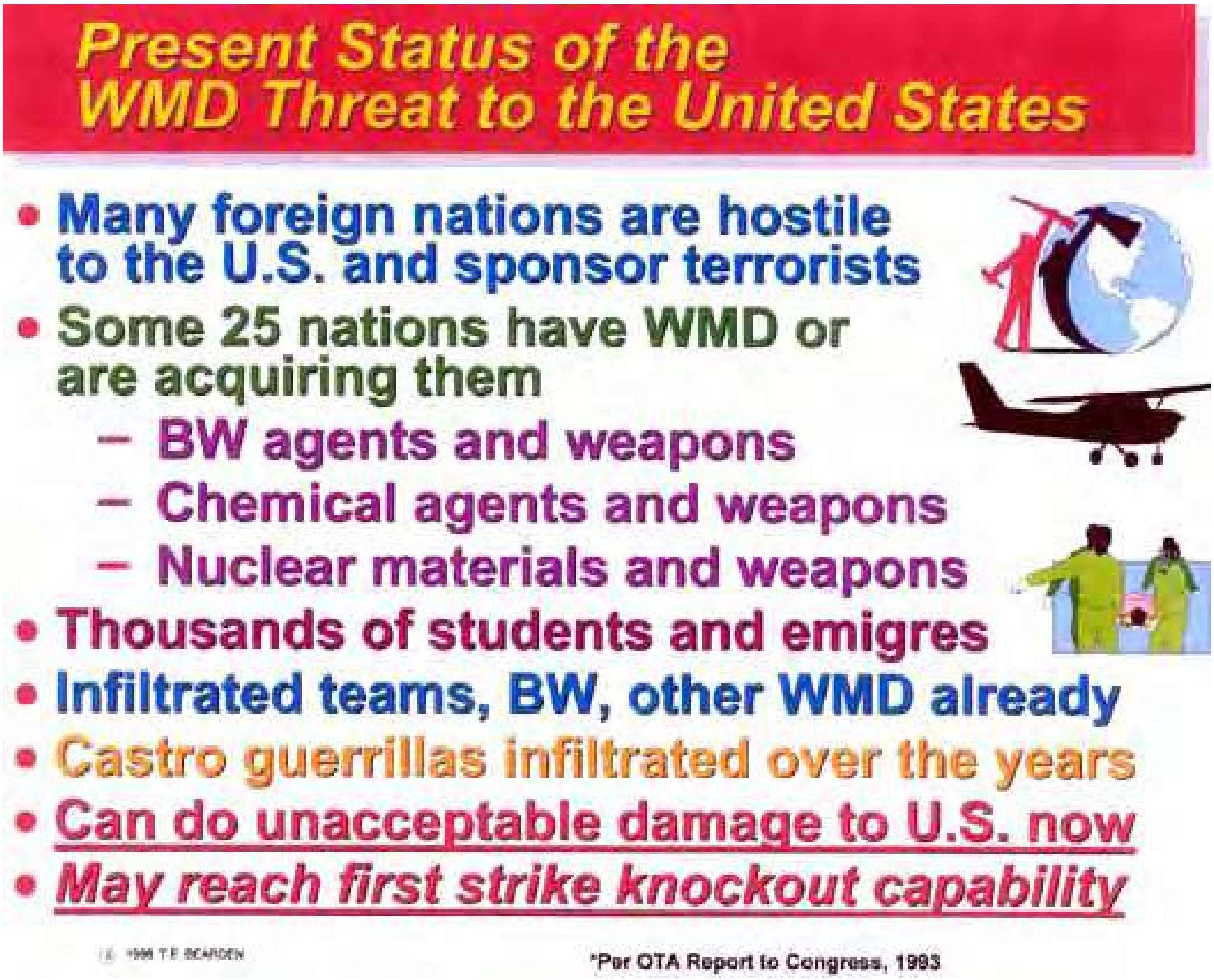

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
    - BW agents and weapons
    - Chemical agents and weapons
    - Nuclear materials and weapons
- Thousands of students and emigres
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© 1998 T.E. BEARDEN

*Per OTA Report to Congress, 1993

Return to Slide Index

Next Slide

Previous Slide

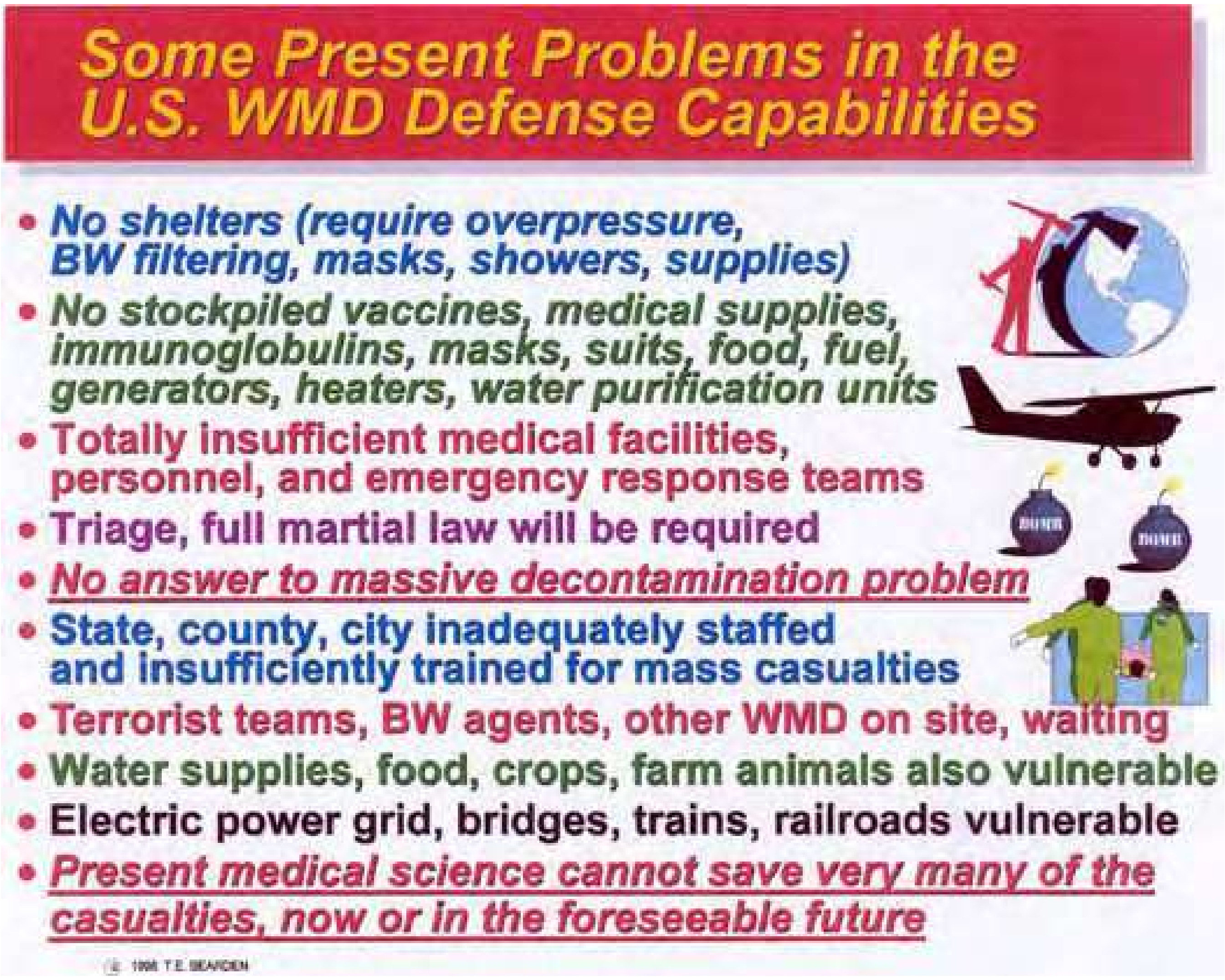

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

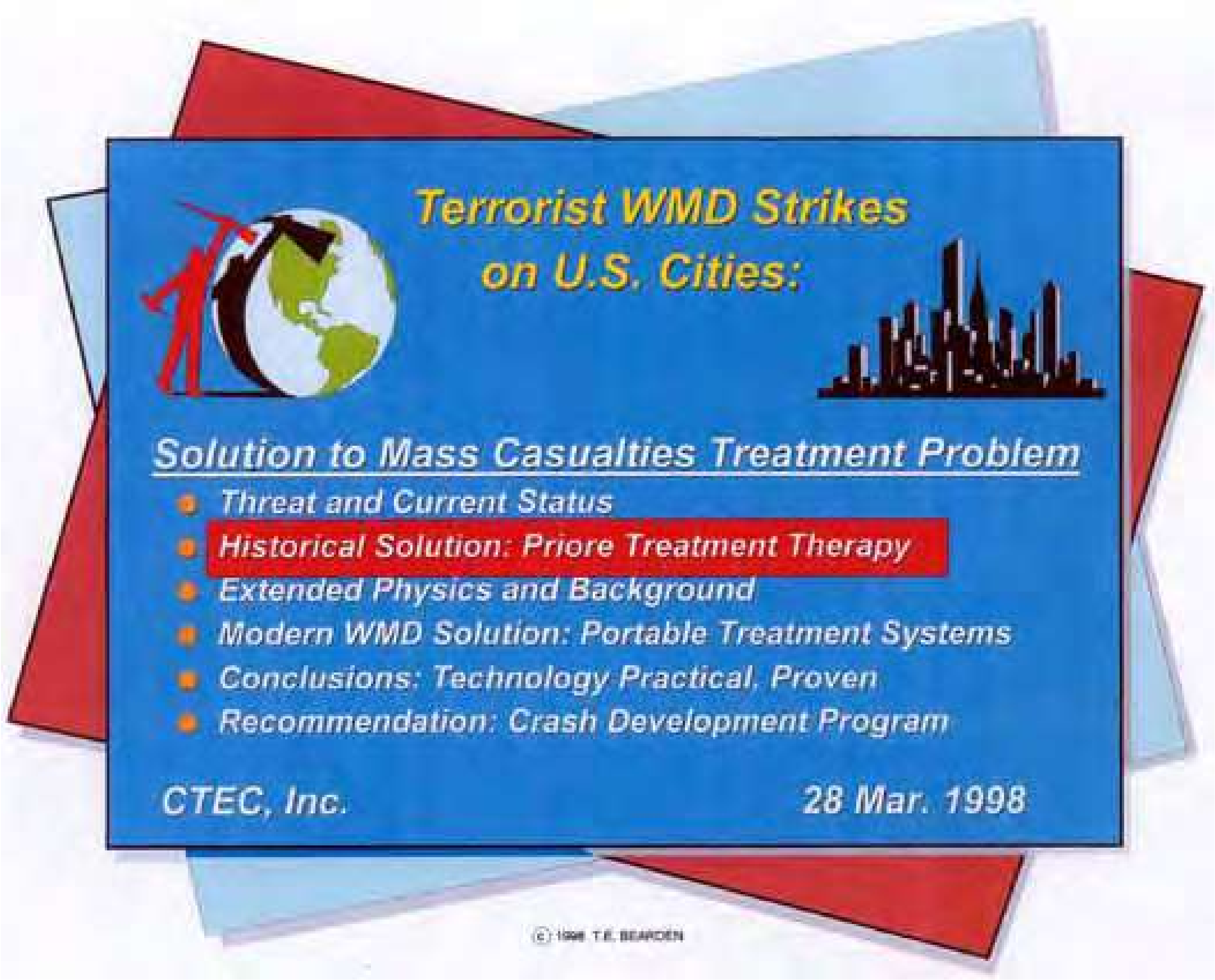

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

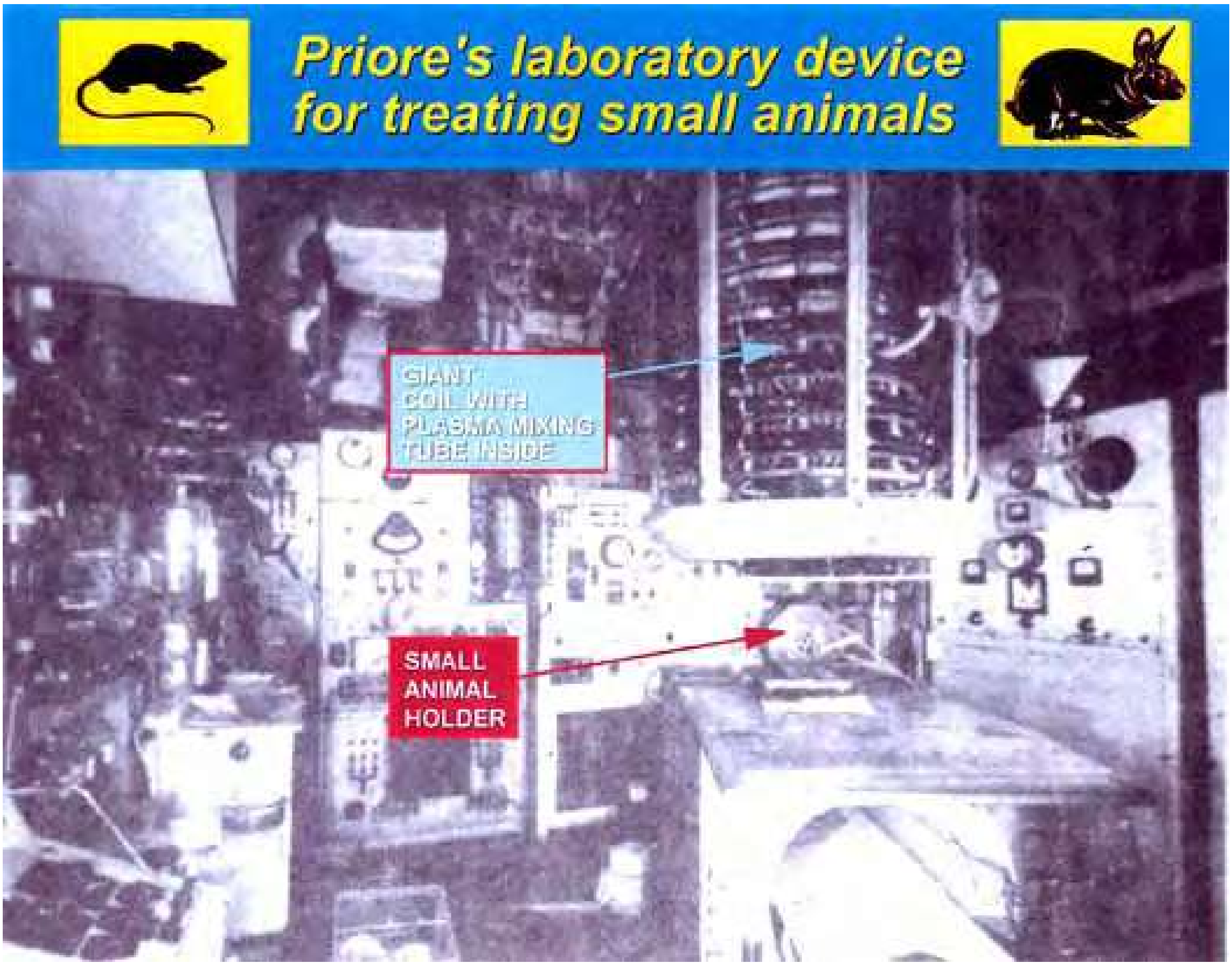

Priore's laboratory device for treating small animals

GIANT COIL WITH PLASMA MIXING TUBE INSIDE

SMALL ANIMAL HOLDER

Return to Slide Index

Next Slide

Previous Slide

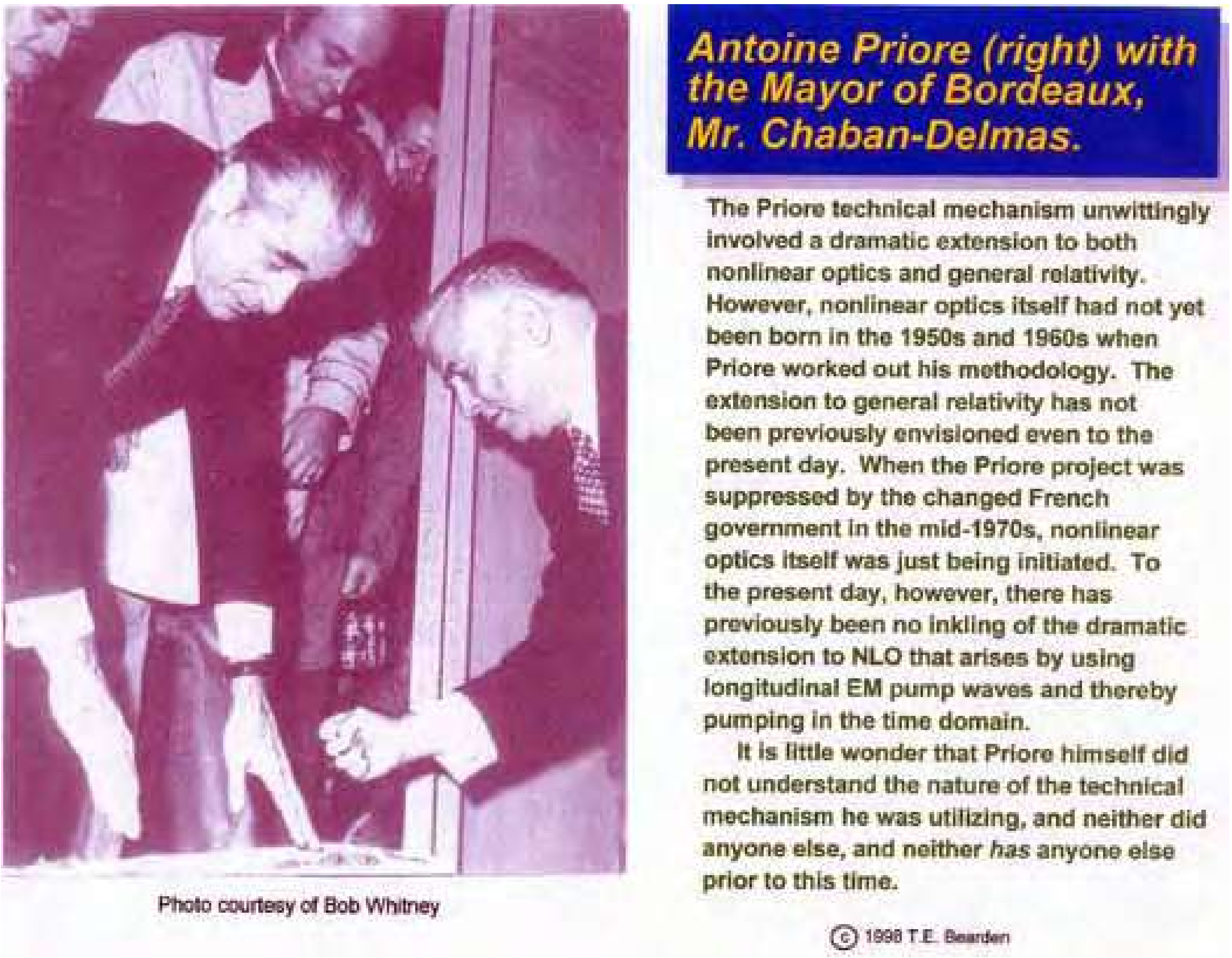

Antoine Priore (right) with the Mayor of Bordeaux, Mr. Chaban-Delmas.

The Priore technical mechanism unwittingly involved a dramatic extension to both nonlinear optics and general relativity. However, nonlinear optics itself had not yet been born in the 1950s and 1960s when Priore worked out his methodology. The extension to general relativity has not been previously envisioned even to the present day. When the Priore project was suppressed by the changed French government in the mid-1970s, nonlinear optics itself was just being initiated. To the present day, however, there has previously been no inkling of the dramatic extension to NLO that arises by using longitudinal EM pump waves and thereby pumping in the time domain.

It is little wonder that Priore himself did not understand the nature of the technical mechanism he was utilizing, and neither did anyone else, and neither *has* anyone else prior to this time.

Photo courtesy of Bob Whitney

ⓒ 1998 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

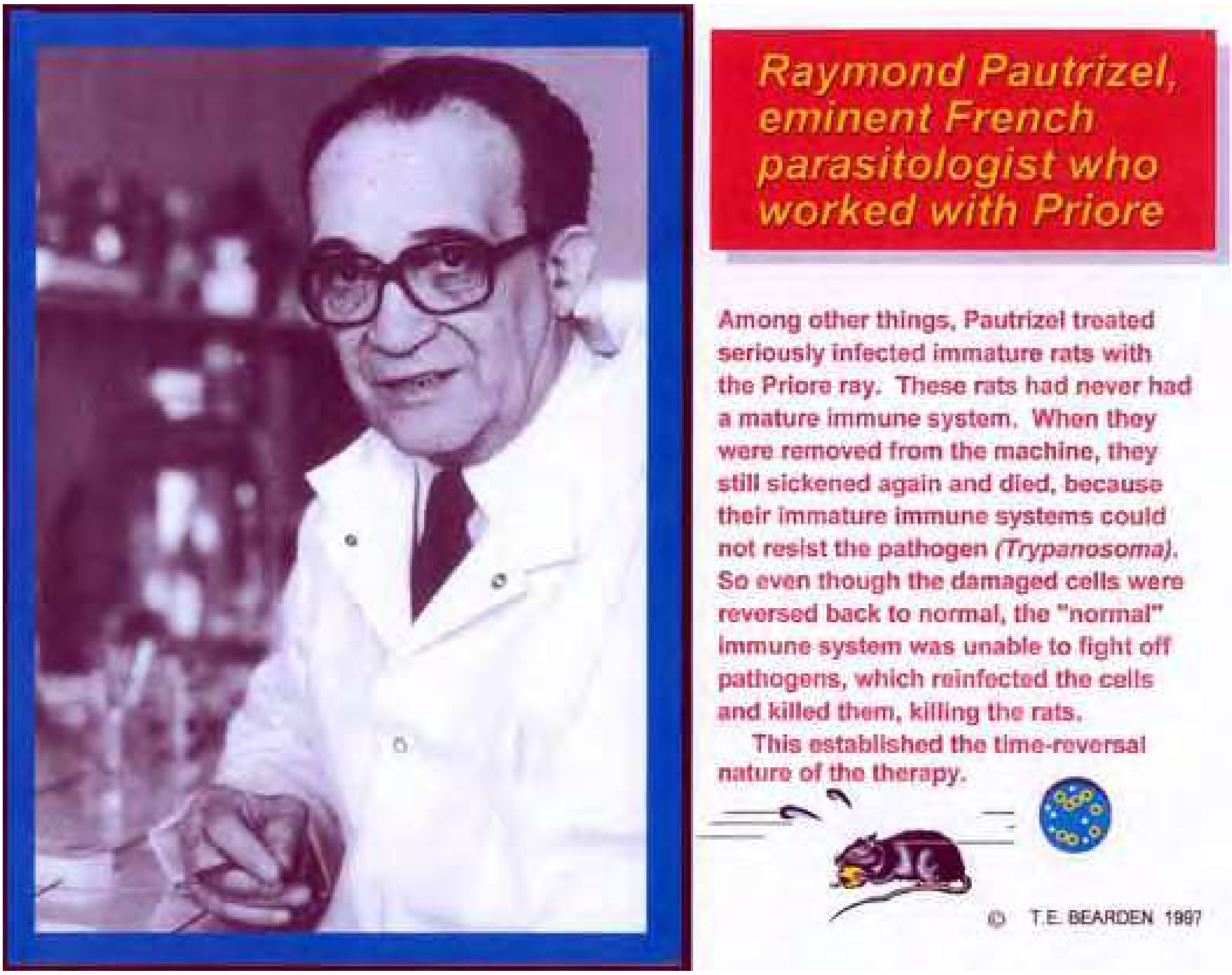

Raymond Pautrizel, eminent French parasitologist who worked with Priore

Among other things, Pautrizel treated seriously infected immature rats with the Priore ray. These rats had never had a mature immune system. When they were removed from the machine, they still sickened again and died, because their immature immune systems could not resist the pathogen *(Trypanosoma)*. So even though the damaged cells were reversed back to normal, the "normal" immune system was unable to fight off pathogens, which reinfected the cells and killed them, killing the rats.

This established the time-reversal nature of the therapy.

© T.E. BEARDEN 1997

[Return to Slide Index]

[Next Slide]

[Previous Slide]

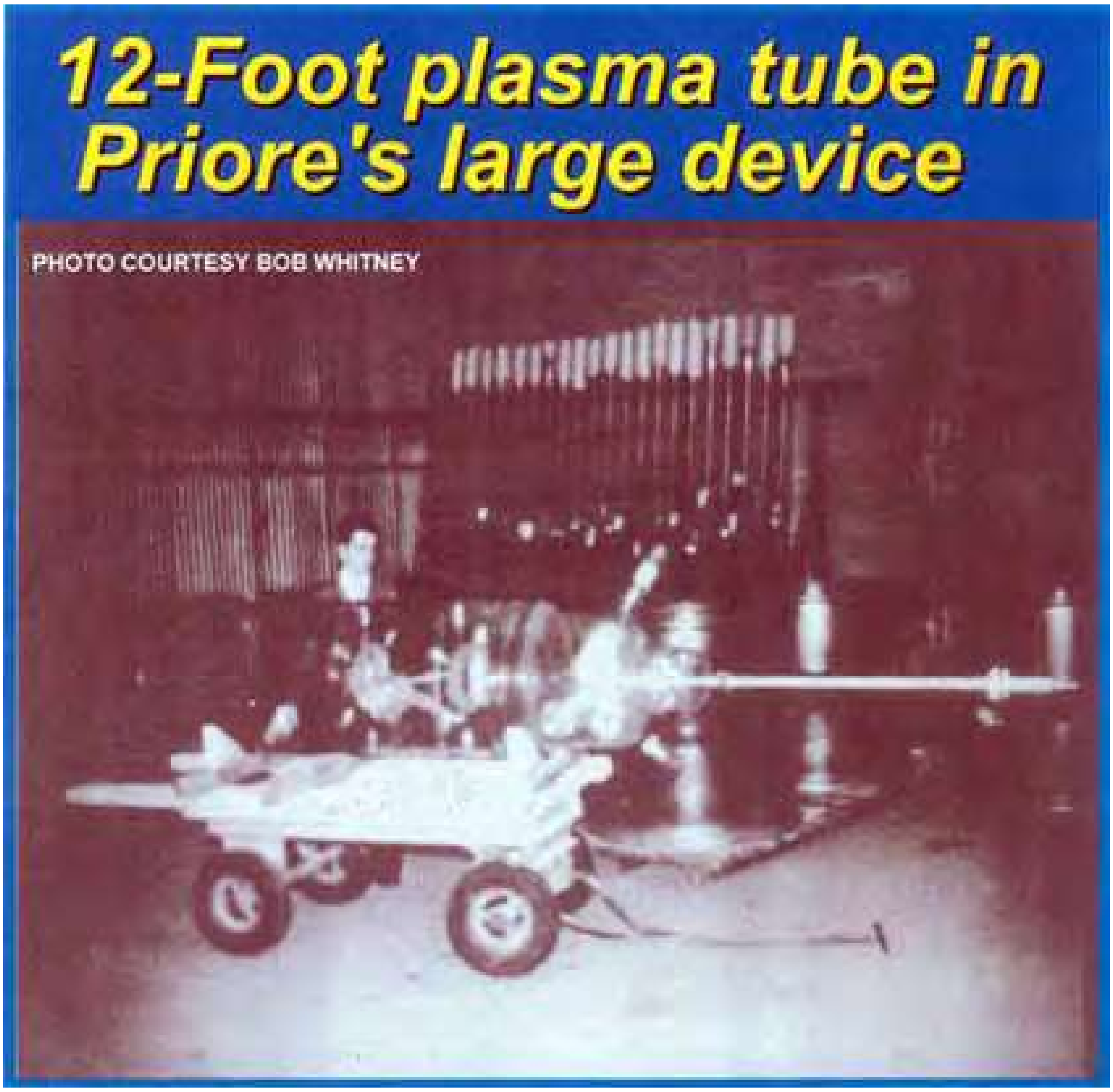

12-Foot plasma tube in Priore's large device

PHOTO COURTESY BOB WHITNEY

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

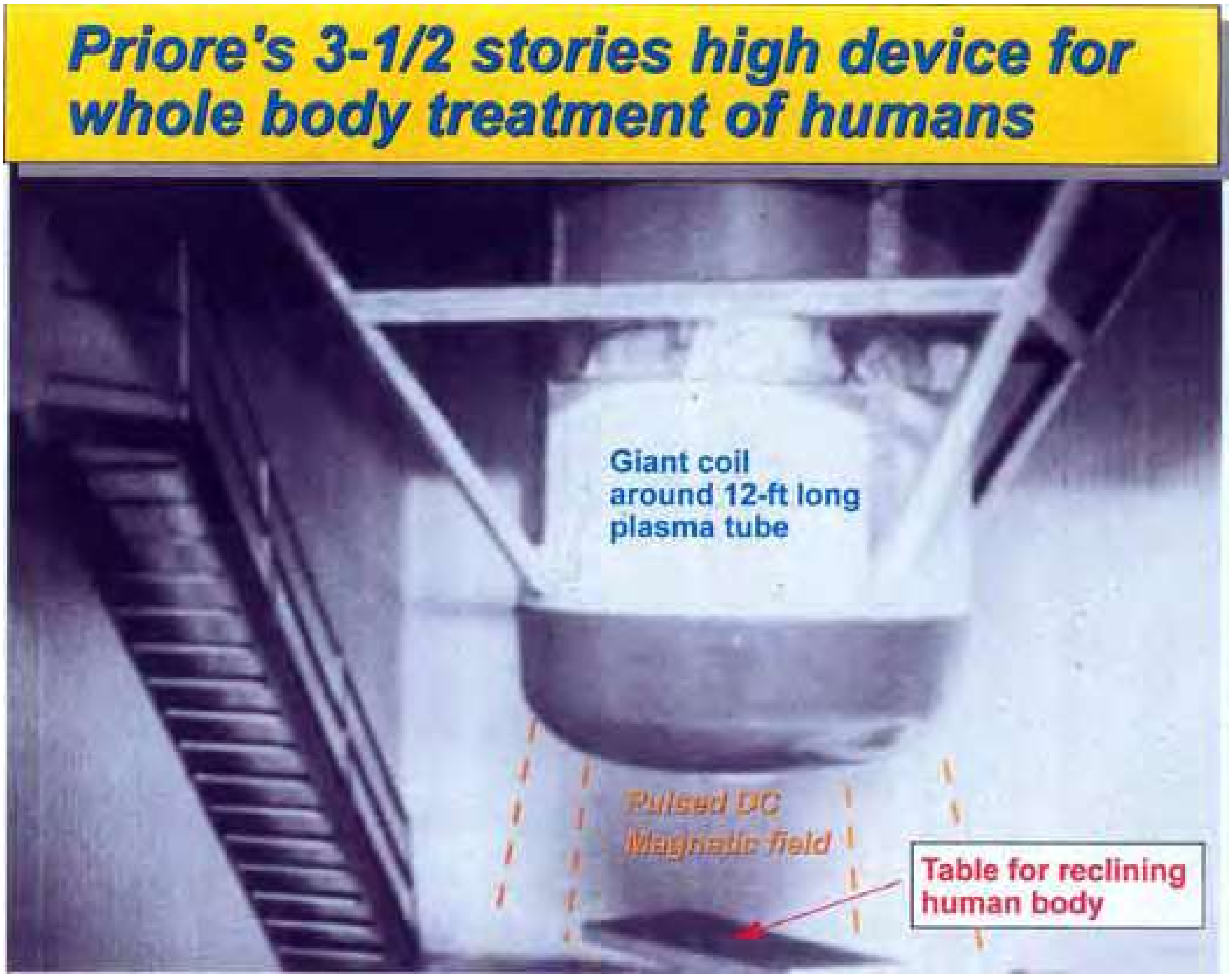

Priore's 3-1/2 stories high device for whole body treatment of humans

Giant coil
around 12-ft long
plasma tube

Pulsed DC
Magnetic field

Table for reclining
human body

Return to Slide Index

Next Slide

Previous Slide

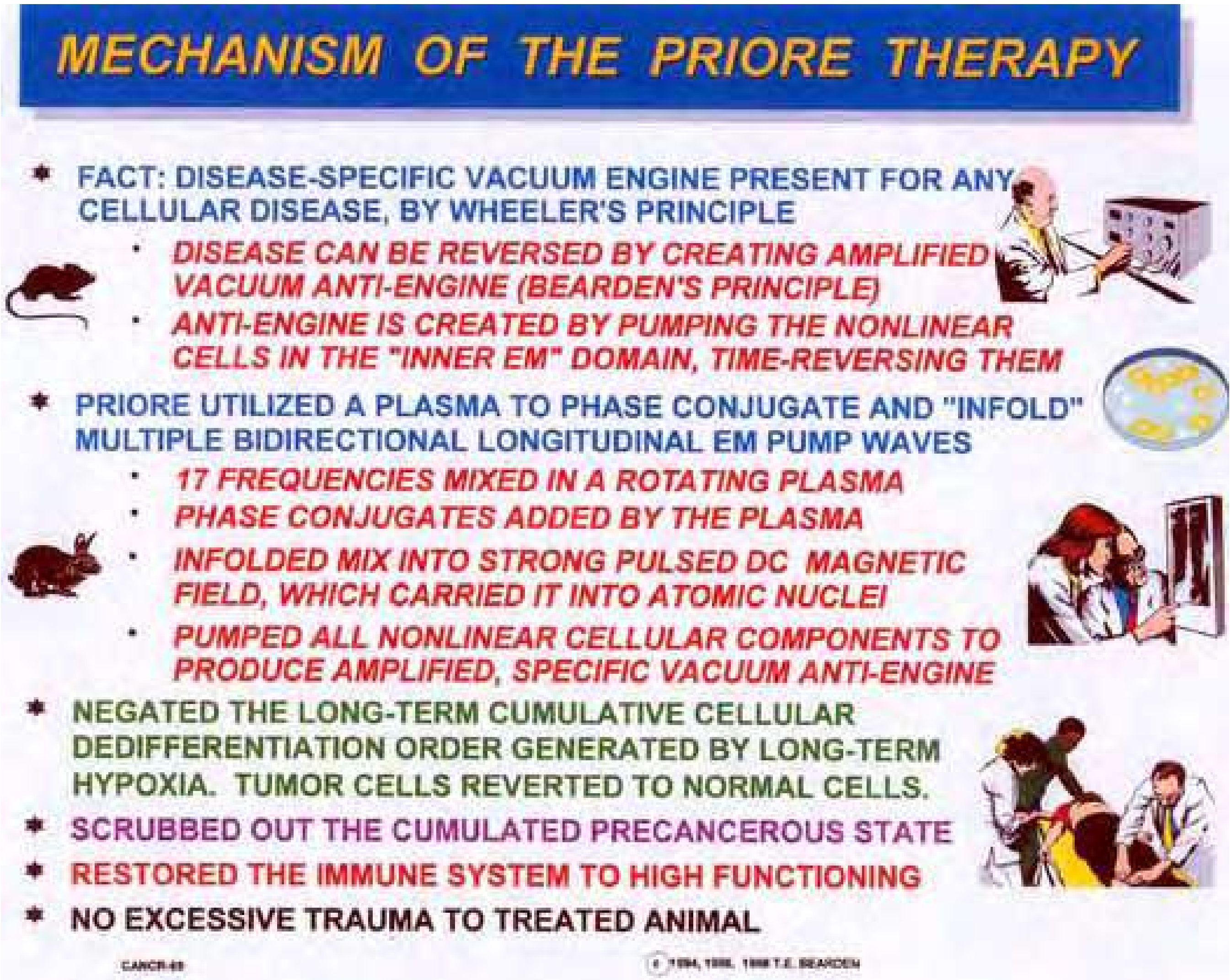

MECHANISM OF THE PRIORE THERAPY

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

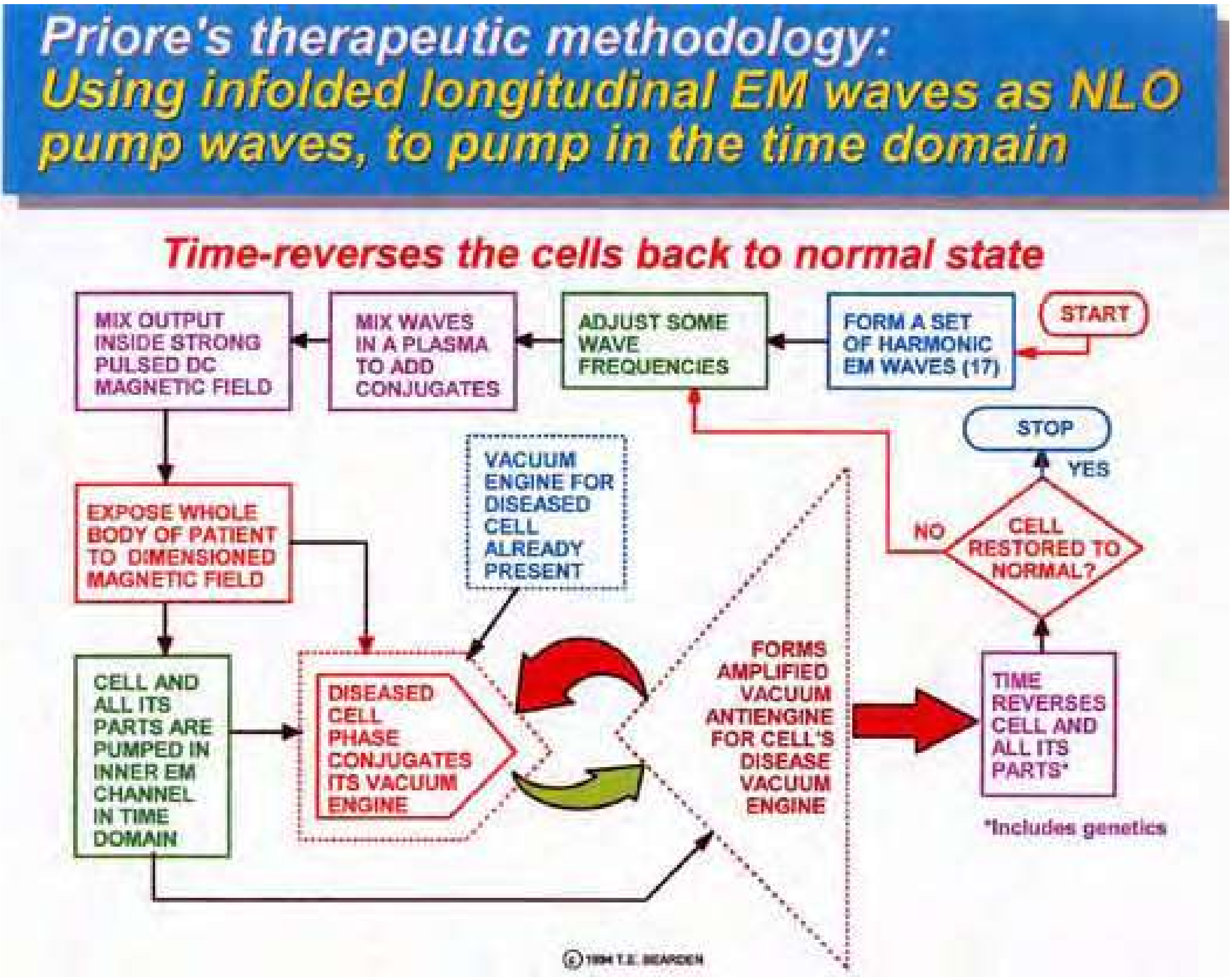

Priore's therapeutic methodology: Using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain

Time-reverses the cells back to normal state

START

FORM A SET OF HARMONIC EM WAVES (17)

ADJUST SOME WAVE FREQUENCIES

MIX WAVES IN A PLASMA TO ADD CONJUGATES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME REVERSES CELL AND ALL ITS PARTS*

CELL RESTORED TO NORMAL?

NO

YES

STOP

*Includes genetics

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

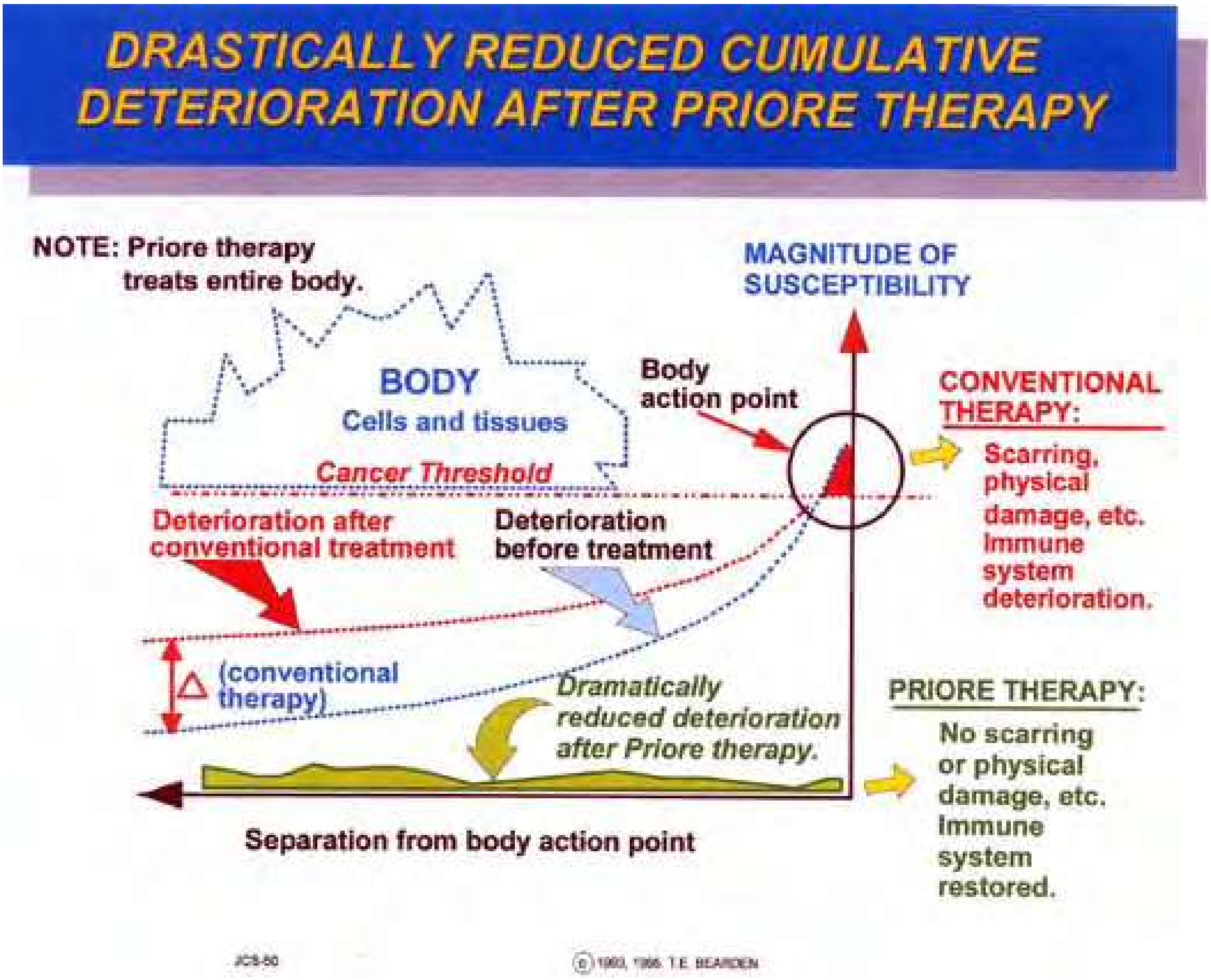

DRASTICALLY REDUCED CUMULATIVE DETERIORATION AFTER PRIORE THERAPY

NOTE: Priore therapy treats entire body.

BODY

Cells and tissues

Cancer Threshold

Body action point

MAGNITUDE OF SUSCEPTIBILITY

Deterioration after conventional treatment

Deterioration before treatment

Δ (conventional therapy)

Dramatically reduced deterioration after Priore therapy.

Separation from body action point

CONVENTIONAL THERAPY:

Scarring, physical damage, etc. Immune system deterioration.

PRIORE THERAPY:

No scarring or physical damage, etc. Immune system restored.

JC9-50

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

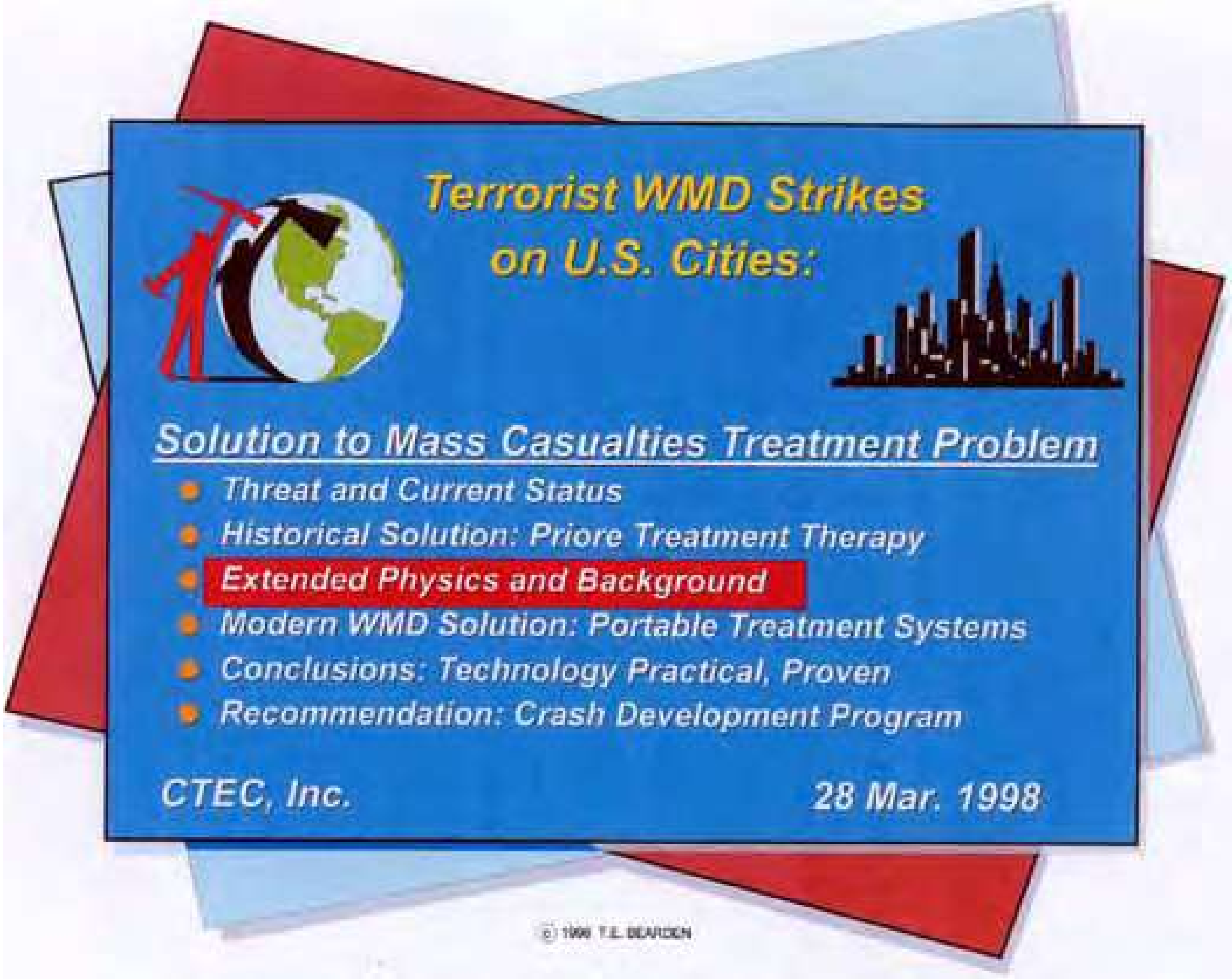

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

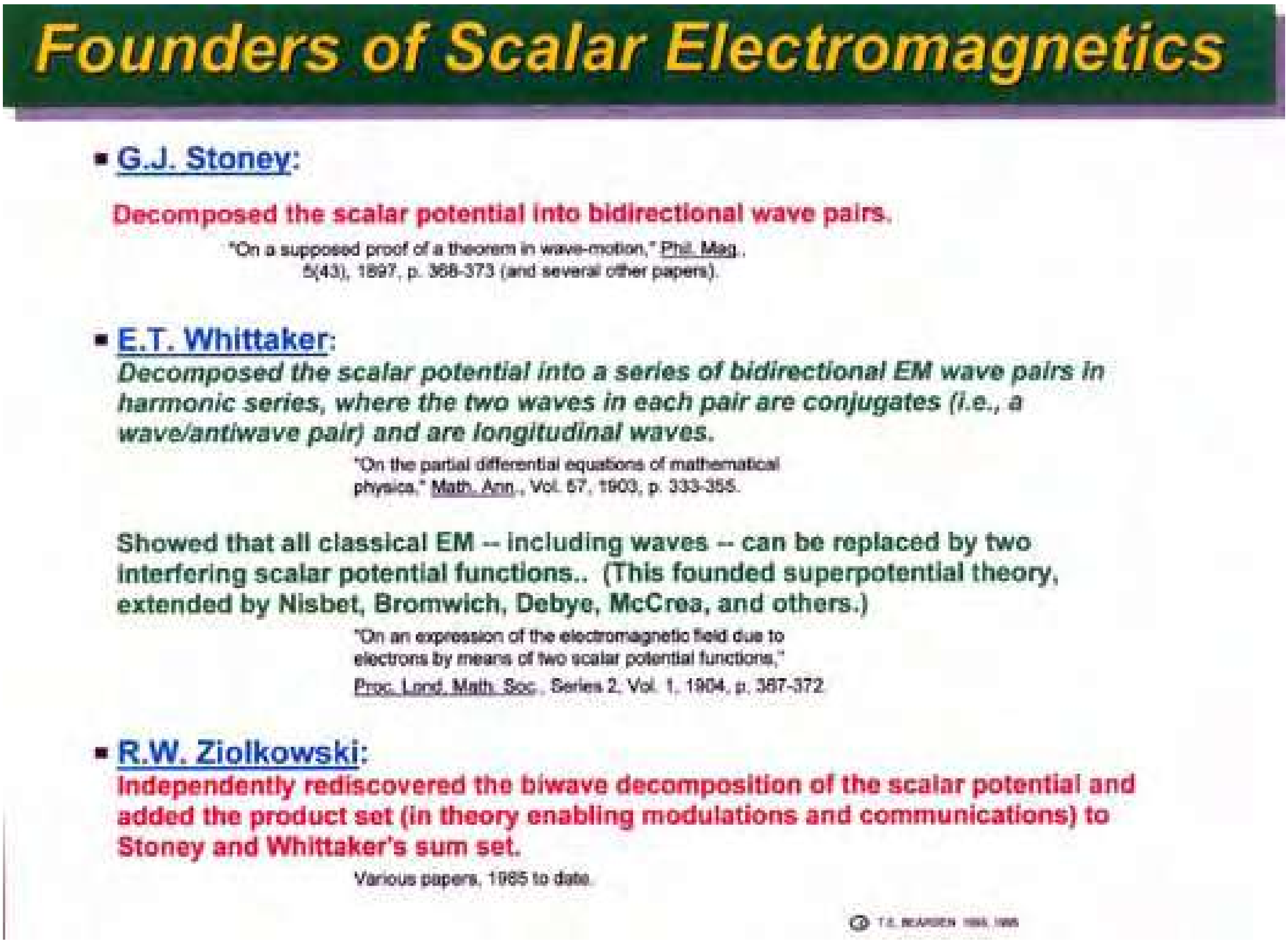

Founders of Scalar Electromagnetics

- **G.J. Stoney:**

Decomposed the scalar potential into bidirectional wave pairs.

"On a supposed proof of a theorem in wave-motion," Phil. Mag.,
5(43), 1897, p. 368-373 (and several other papers).

- **E.T. Whittaker:**

Decomposed the scalar potential into a series of bidirectional EM wave pairs in harmonic series, where the two waves in each pair are conjugates (i.e., a wave/antiwave pair) and are longitudinal waves.

"On the partial differential equations of mathematical
physics," Math. Ann., Vol. 57, 1903, p. 333-355.

Showed that all classical EM -- including waves -- can be replaced by two interfering scalar potential functions.. (This founded superpotential theory, extended by Nisbet, Bromwich, Debye, McCrea, and others.)

"On an expression of the electromagnetic field due to
electrons by means of two scalar potential functions,"
Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372

- **R.W. Ziolkowski:**

Independently rediscovered the biwave decomposition of the scalar potential and added the product set (in theory enabling modulations and communications) to Stoney and Whittaker's sum set.

Various papers, 1985 to date.

© T.E. BEARDEN [illegible]

Return to Slide Index

Next Slide

Previous Slide

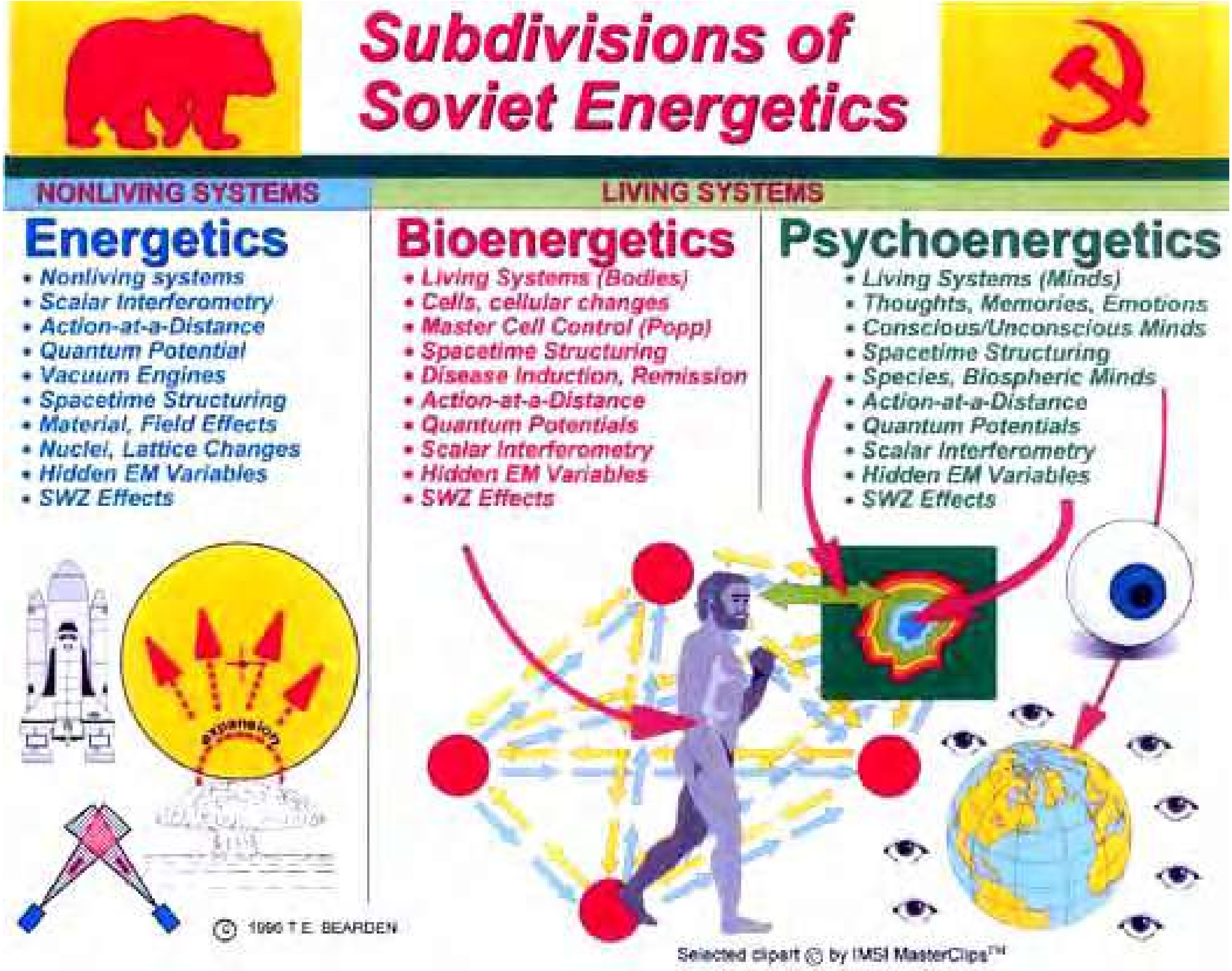

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

LIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1990 T.E. BEARDEN

Selected clipart © by IMSI MasterClips™

Return to Slide Index

Next Slide

Previous Slide

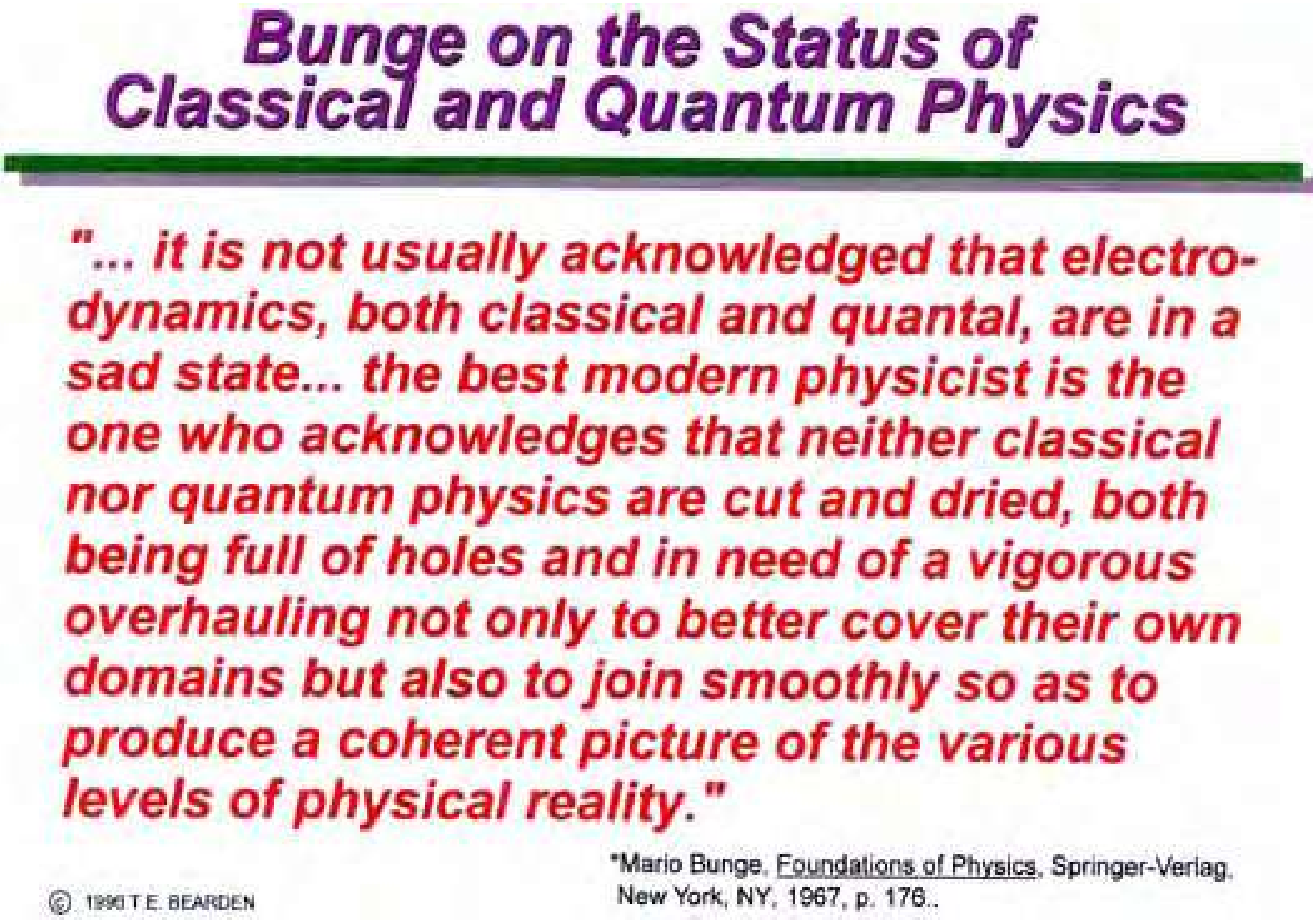

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

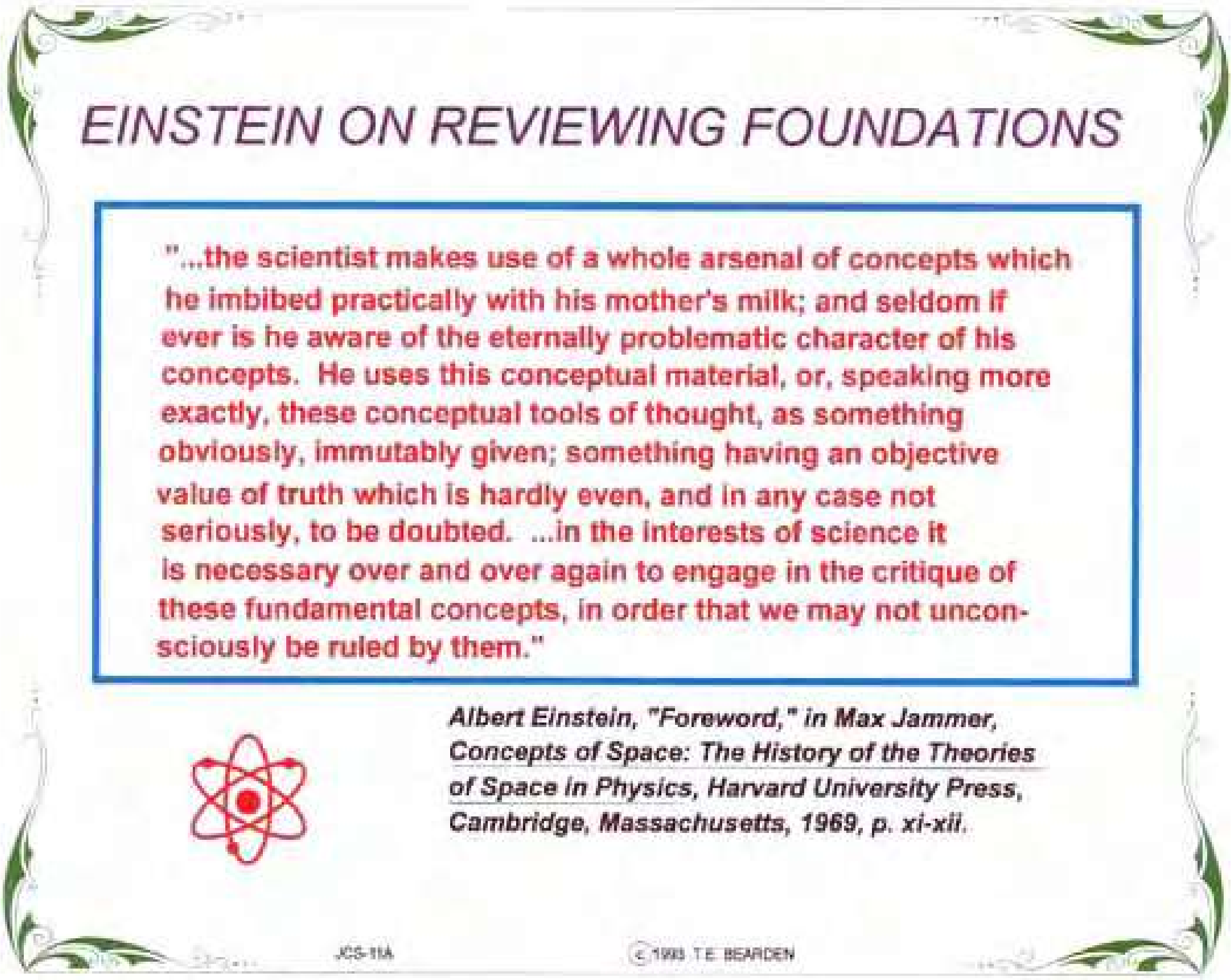

EINSTEIN ON REVIEWING FOUNDATIONS

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

JCS-11A

© 1993 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

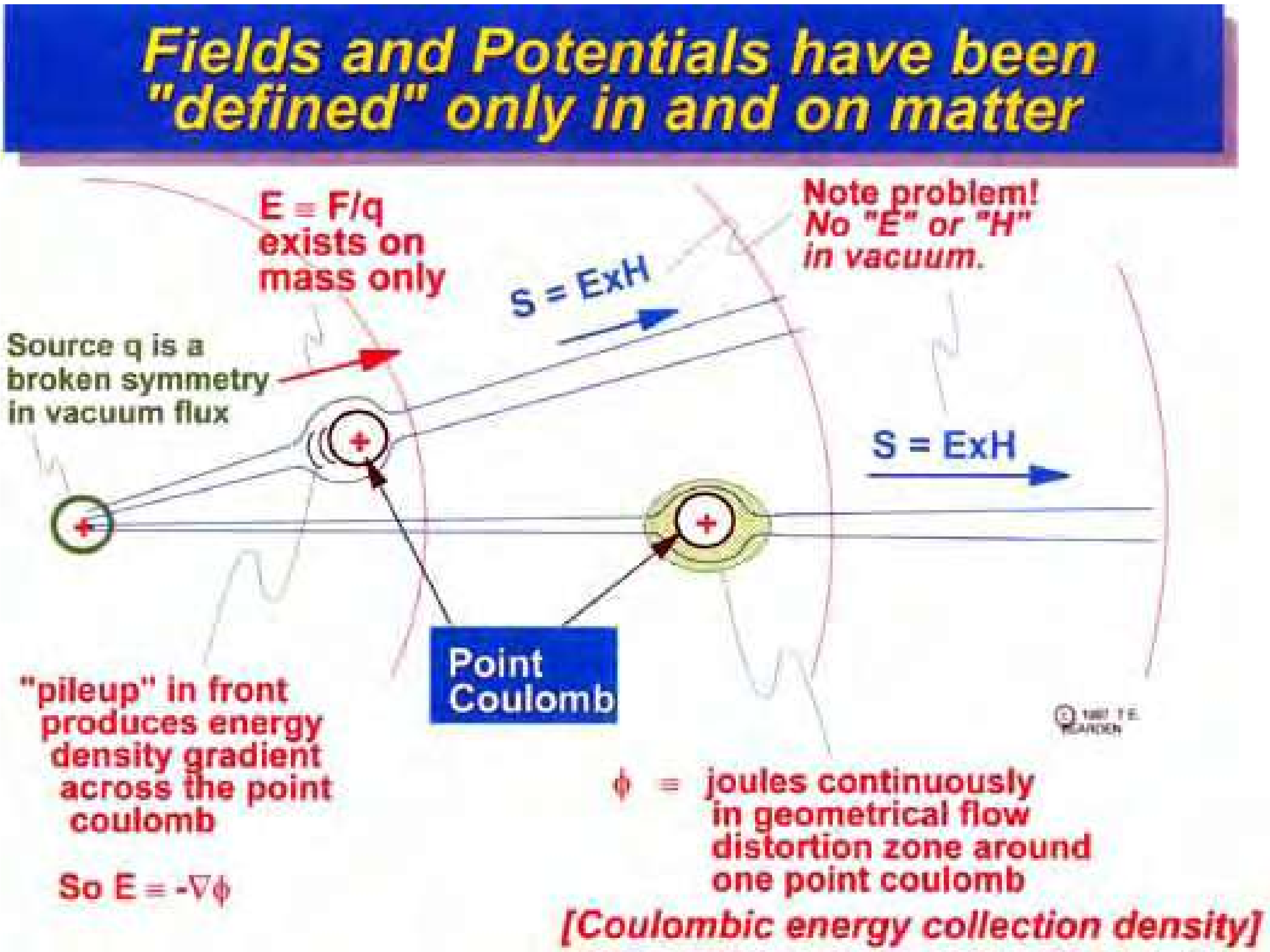

Fields and Potentials have been "defined" only in and on matter

$E \equiv F/q$
exists on
mass only

Note problem!
No "E" or "H"
in vacuum.

Source q is a
broken symmetry
in vacuum flux

S = ExH

S = ExH

Point
Coulomb

"pileup" in front
produces energy
density gradient
across the point
coulomb

So $E \equiv -\nabla\phi$

ϕ $\equiv$ joules continuously
in geometrical flow
distortion zone around
one point coulomb

[Coulombic energy collection density]

© 1987 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

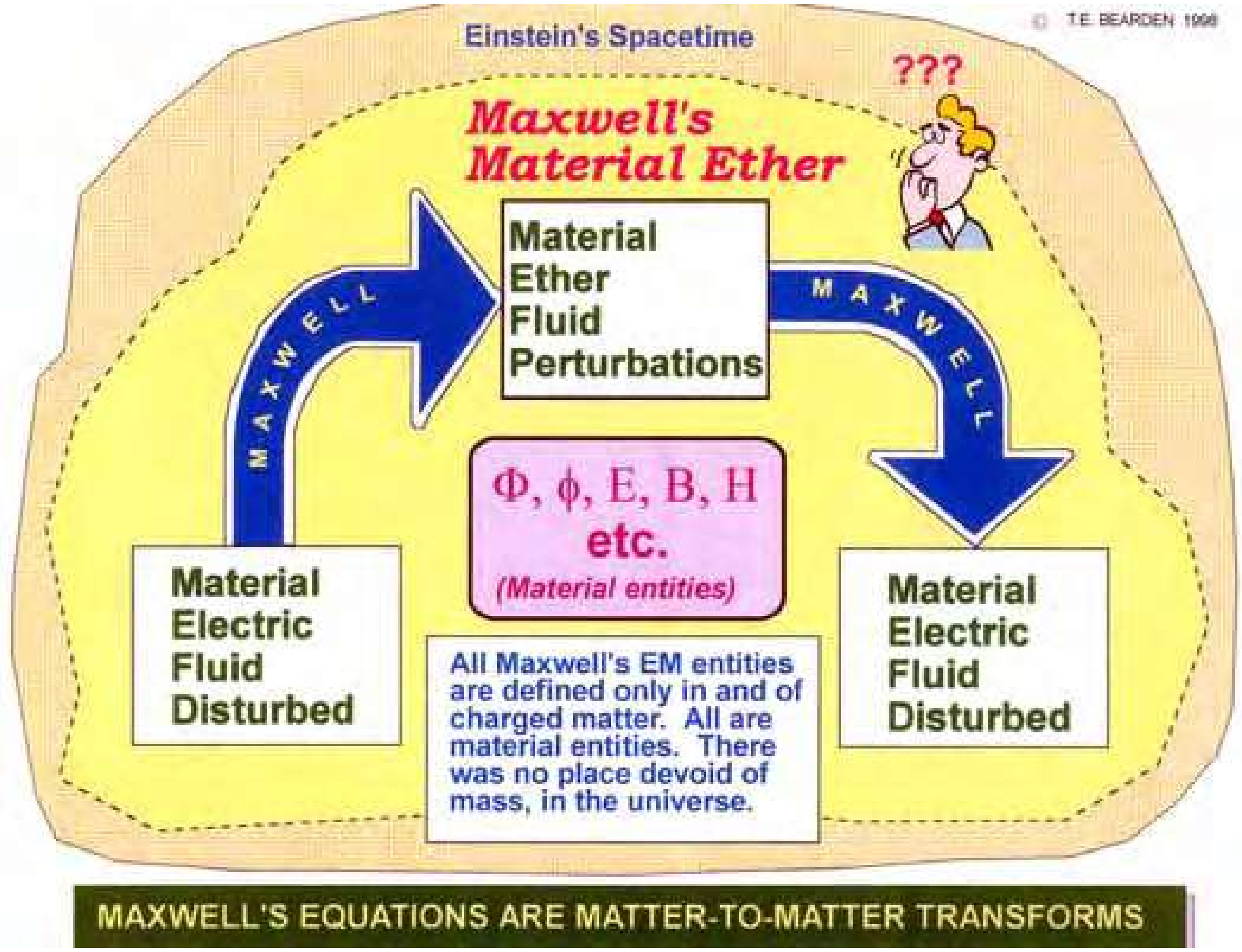

© T.E. BEARDEN 1998
Einstein's Spacetime
???
Maxwell's Material Ether
MAXWELL
MAXWELL

Material Ether Fluid Perturbations

Φ, ϕ, E, B, H etc.

(Material entities)

Material Electric Fluid Disturbed

Material Electric Fluid Disturbed

All Maxwell's EM entities are defined only in and of charged matter. All are material entities. There was no place devoid of mass, in the universe.

MAXWELL'S EQUATIONS ARE MATTER-TO-MATTER TRANSFORMS

[Return to Slide Index]

[Next Slide]

[Previous Slide]

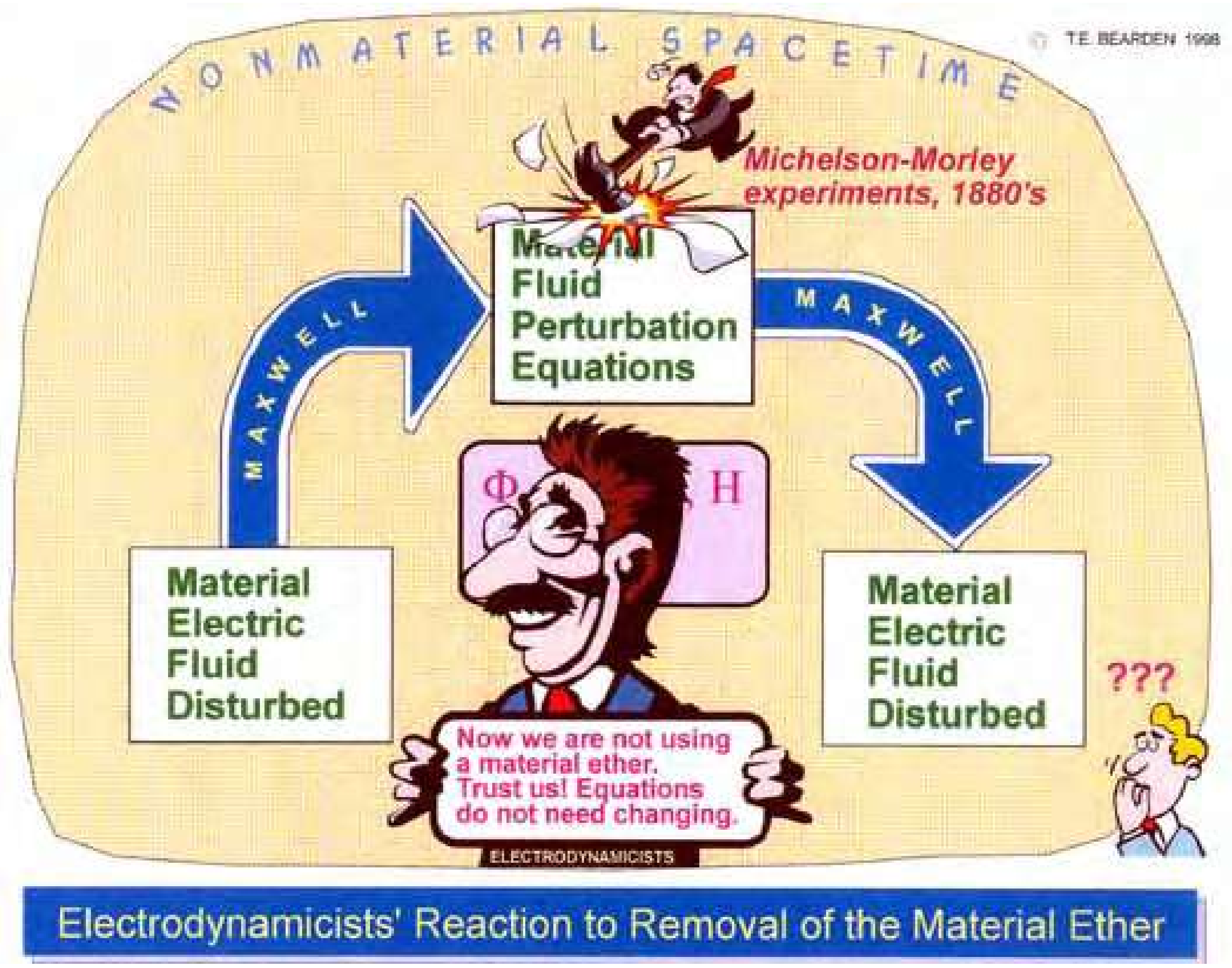
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Michelson-Morley experiments, 1880's
Material Fluid Perturbation Equations
MAXWELL
MAXWELL
Φ
H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! Equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

Return to Slide Index

Next Slide

Previous Slide

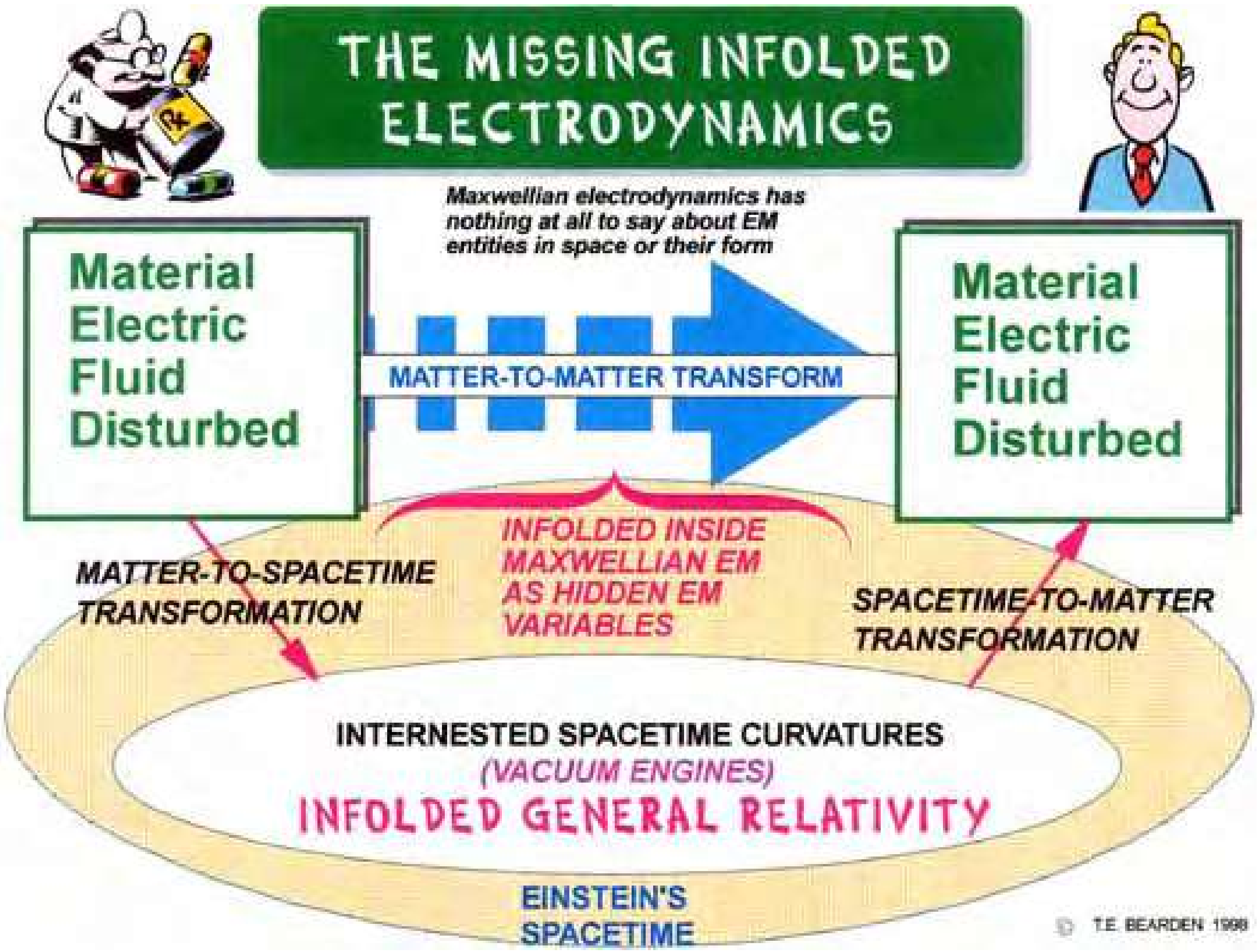

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

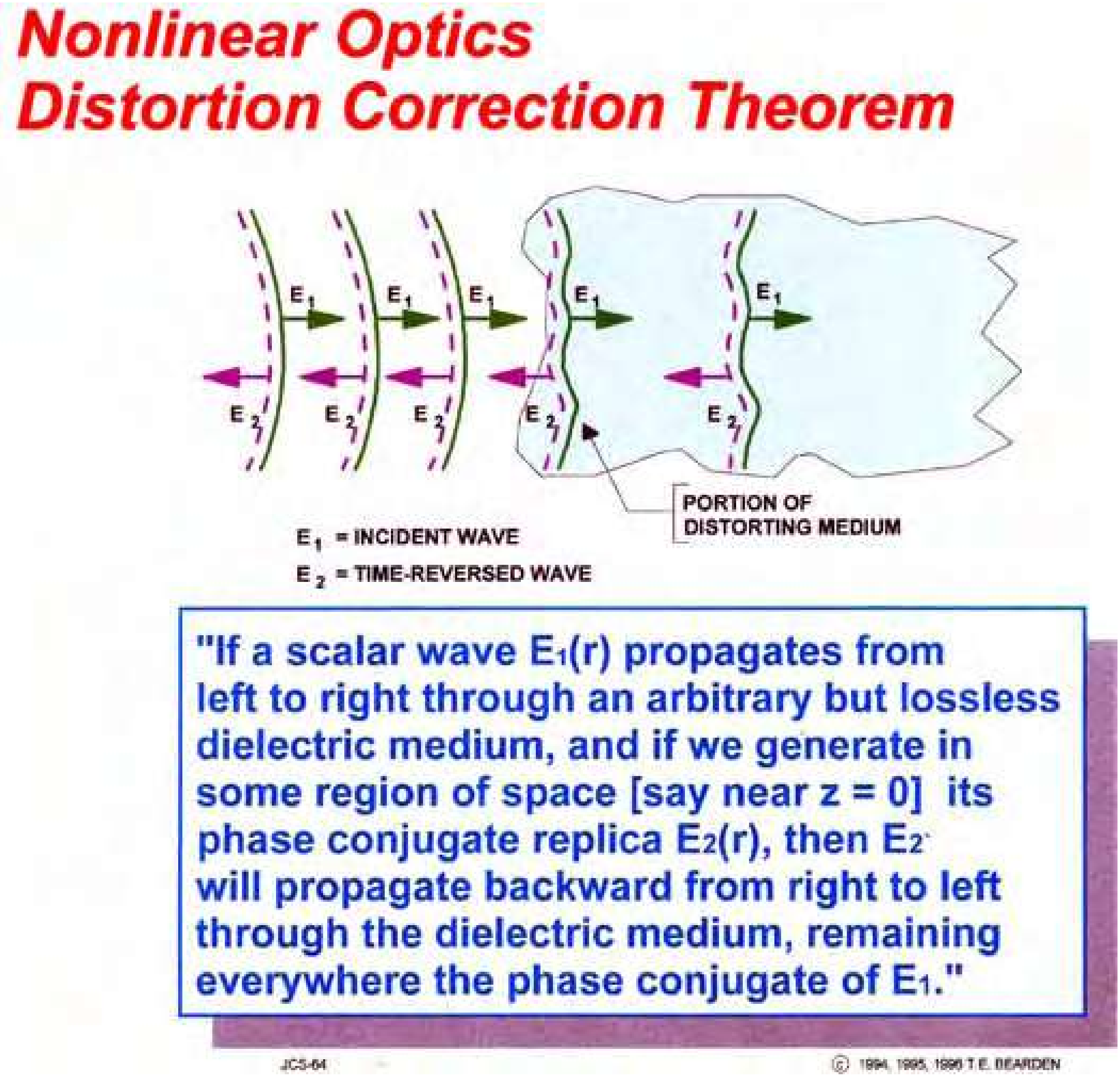

Nonlinear Optics Distortion Correction Theorem

E_1
E_2
PORTION OF DISTORTING MEDIUM

E_1 = INCIDENT WAVE
E_2 = TIME-REVERSED WAVE

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

JCS-64

© 1994, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

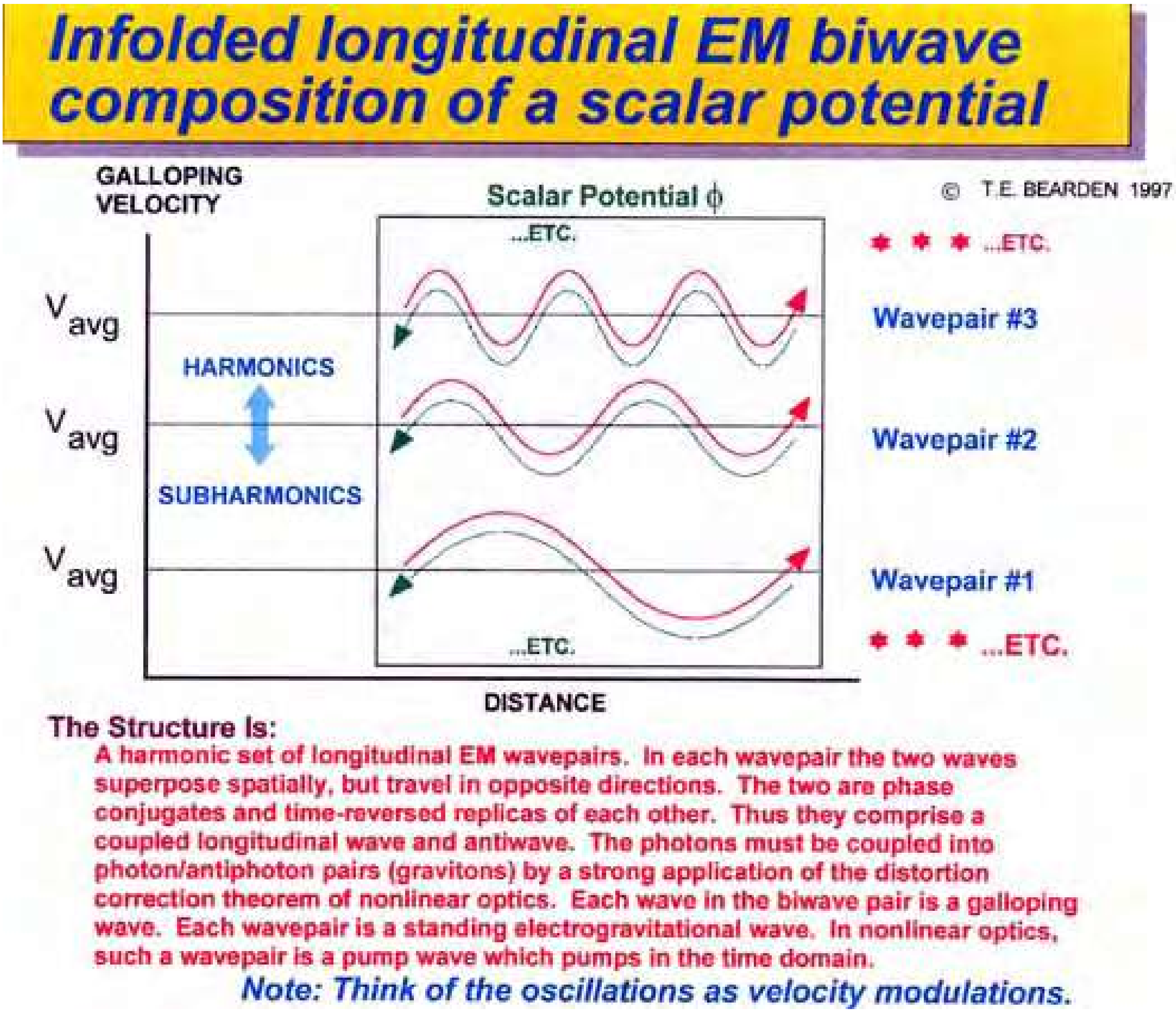

Infolded longitudinal EM biwave composition of a scalar potential

GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC.
* * * ...ETC.
V_{avg}
Wavepair #3
HARMONICS
V_{avg}
Wavepair #2
SUBHARMONICS
V_{avg}
Wavepair #1
...ETC.
* * * ...ETC.
DISTANCE

The Structure Is:

A harmonic set of longitudinal EM wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave. In nonlinear optics, such a wavepair is a pump wave which pumps in the time domain.

Note: Think of the oscillations as velocity modulations.

Return to Slide Index

Next Slide

Previous Slide

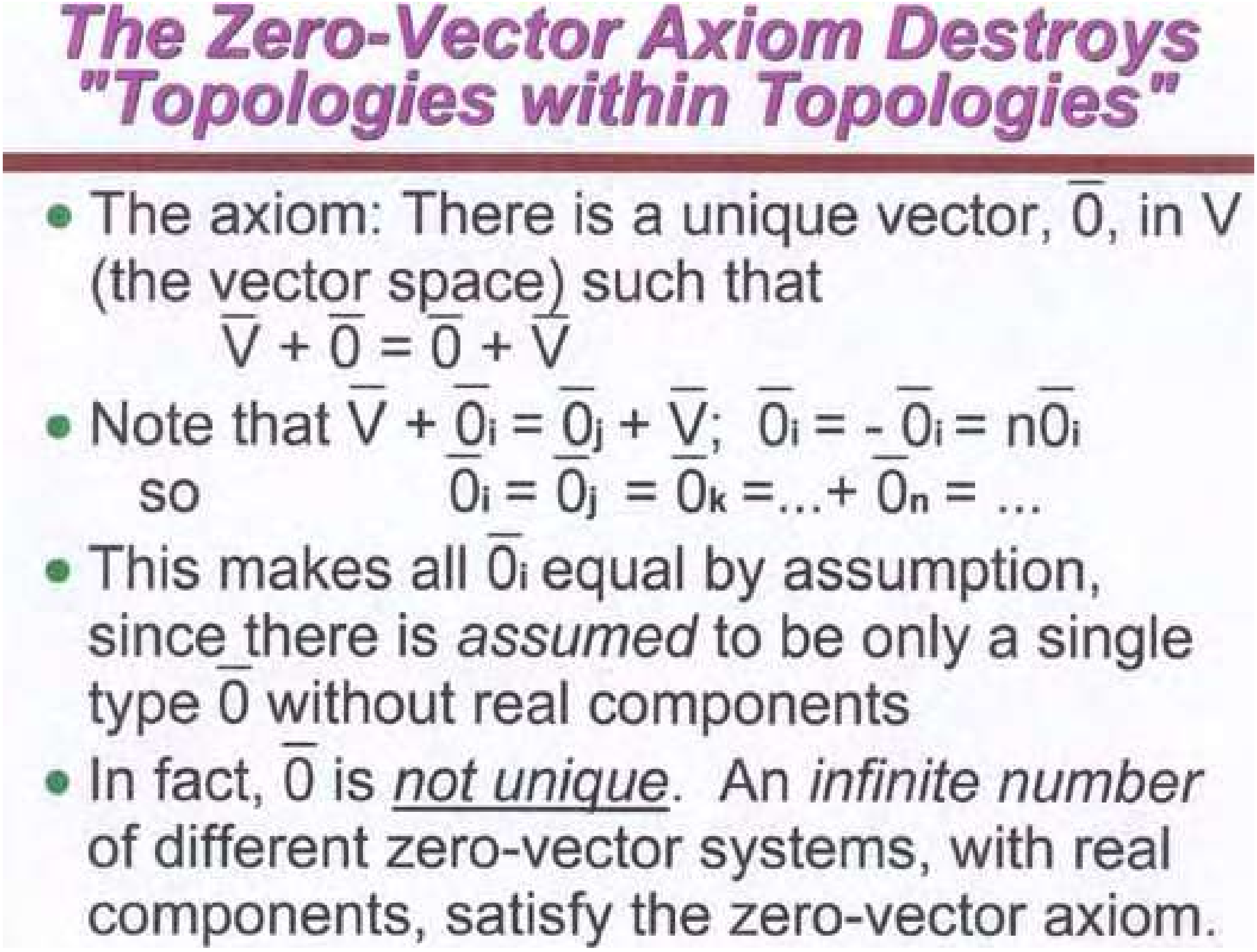

The Zero-Vector Axiom Destroys "Topologies within Topologies"

- The axiom: There is a unique vector, $\overline{0}$, in V (the vector space) such that
  $\overline{V} + \overline{0} = \overline{0} + \overline{V}$
- Note that $\overline{V} + \overline{0}_i = \overline{0}_j + \overline{V}$; $\overline{0}_i = -\overline{0}_i = n\overline{0}_i$
  so $\overline{0}_i = \overline{0}_j = \overline{0}_k = \ldots + \overline{0}_n = \ldots$
- This makes all $\overline{0}_i$ equal by assumption, since there is *assumed* to be only a single type $\overline{0}$ without real components
- In fact, $\overline{0}$ is *not unique*. An *infinite number* of different zero-vector systems, with real components, satisfy the zero-vector axiom.

Return to Slide Index

Next Slide

Previous Slide

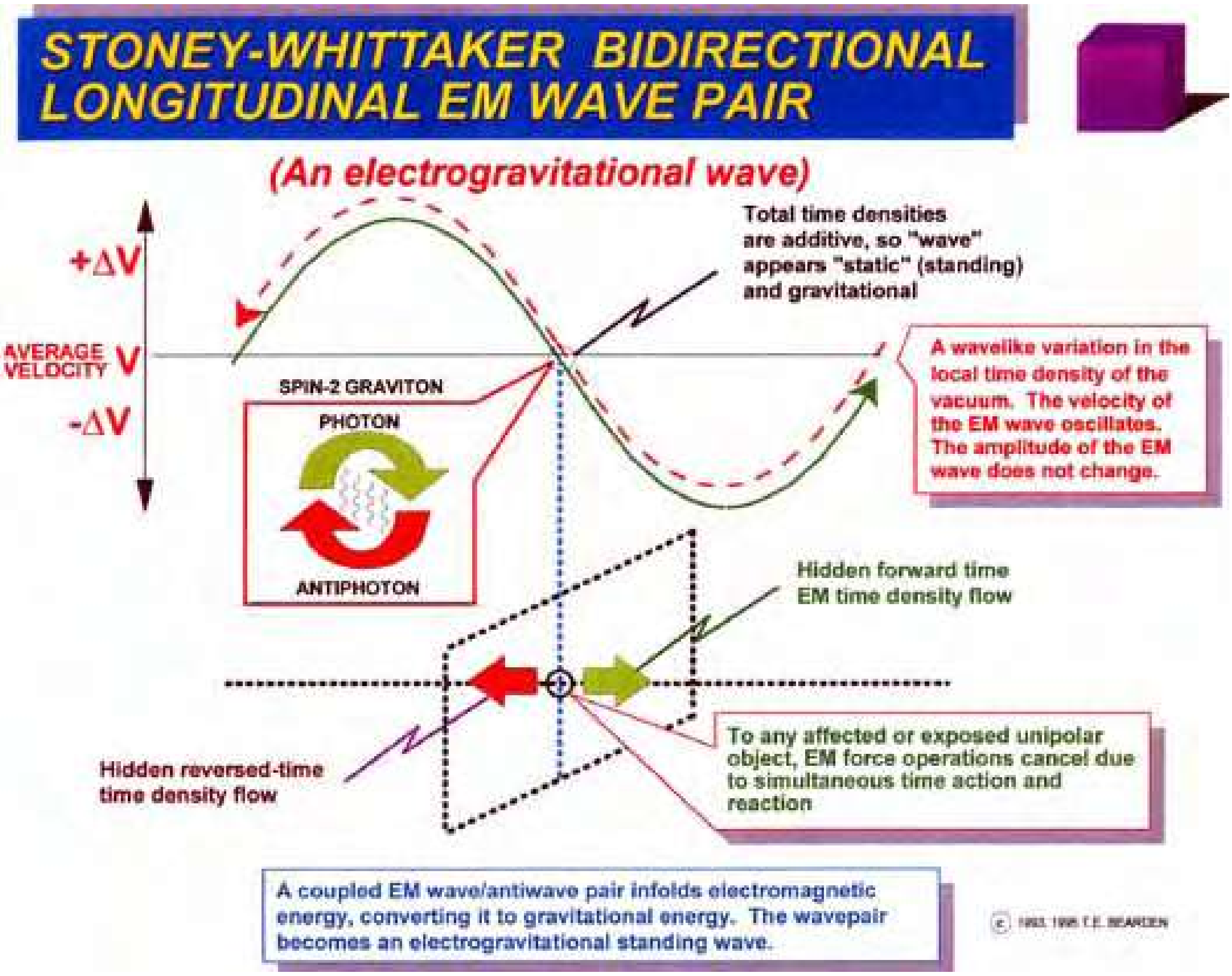

STONEY-WHITTAKER BIDIRECTIONAL LONGITUDINAL EM WAVE PAIR

(An electrogravitational wave)

Total time densities are additive, so "wave" appears "static" (standing) and gravitational

A wavelike variation in the local time density of the vacuum. The velocity of the EM wave oscillates. The amplitude of the EM wave does not change.

Hidden forward time EM time density flow

Hidden reversed-time time density flow

To any affected or exposed unipolar object, EM force operations cancel due to simultaneous time action and reaction

A coupled EM wave/antiwave pair infolds electromagnetic energy, converting it to gravitational energy. The wavepair becomes an electrogravitational standing wave.

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

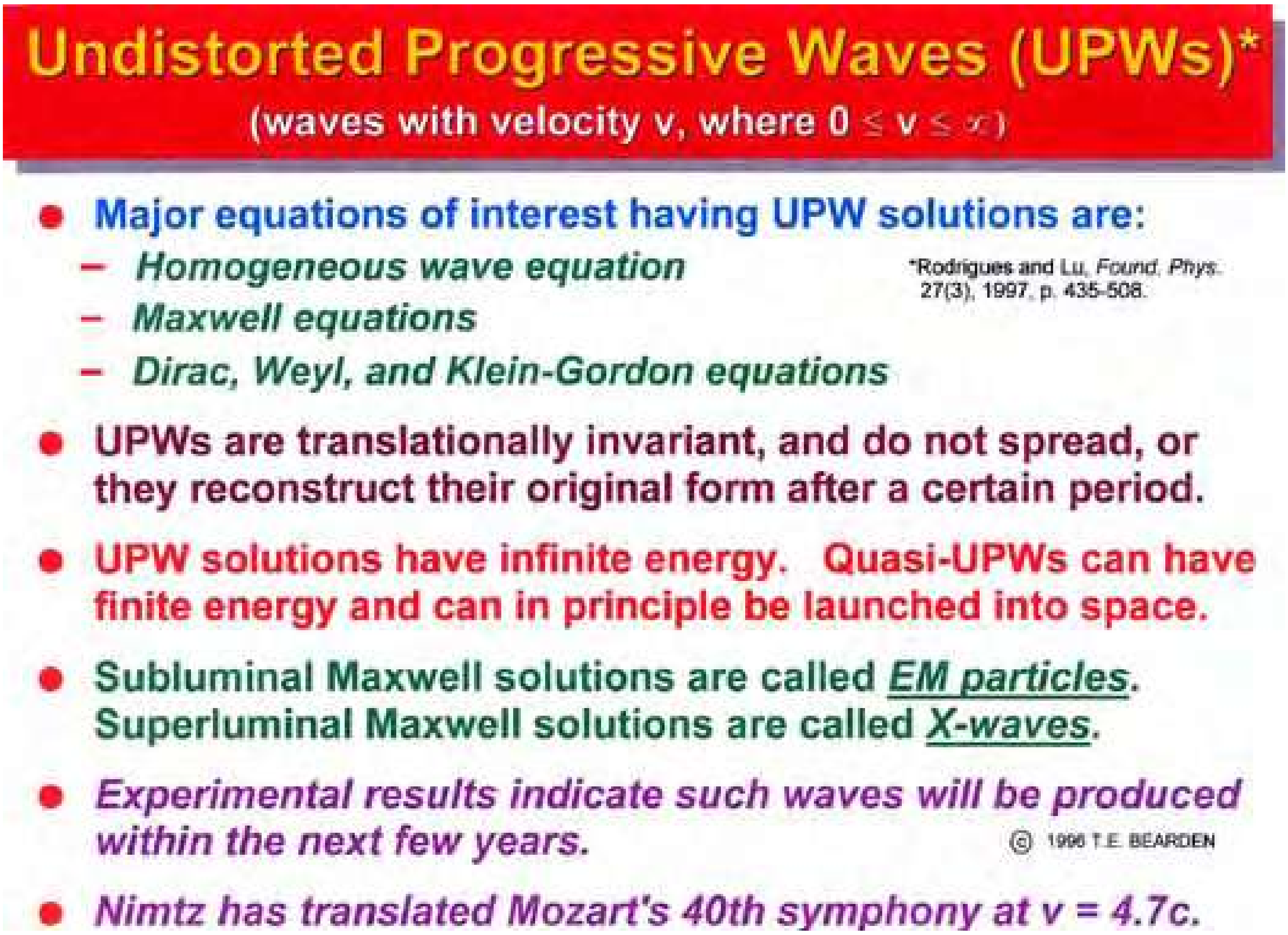

Undistorted Progressive Waves (UPWs)*

(waves with velocity v, where $0 \le v \le \infty$)

- Major equations of interest having UPW solutions are:
  - *Homogeneous wave equation*
  - *Maxwell equations*
  - *Dirac, Weyl, and Klein-Gordon equations*
- UPWs are translationally invariant, and do not spread, or they reconstruct their original form after a certain period.
- UPW solutions have infinite energy. Quasi-UPWs can have finite energy and can in principle be launched into space.
- Subluminal Maxwell solutions are called *EM particles*. Superluminal Maxwell solutions are called *X-waves*.
- *Experimental results indicate such waves will be produced within the next few years.*
- *Nimtz has translated Mozart's 40th symphony at v = 4.7c.*

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

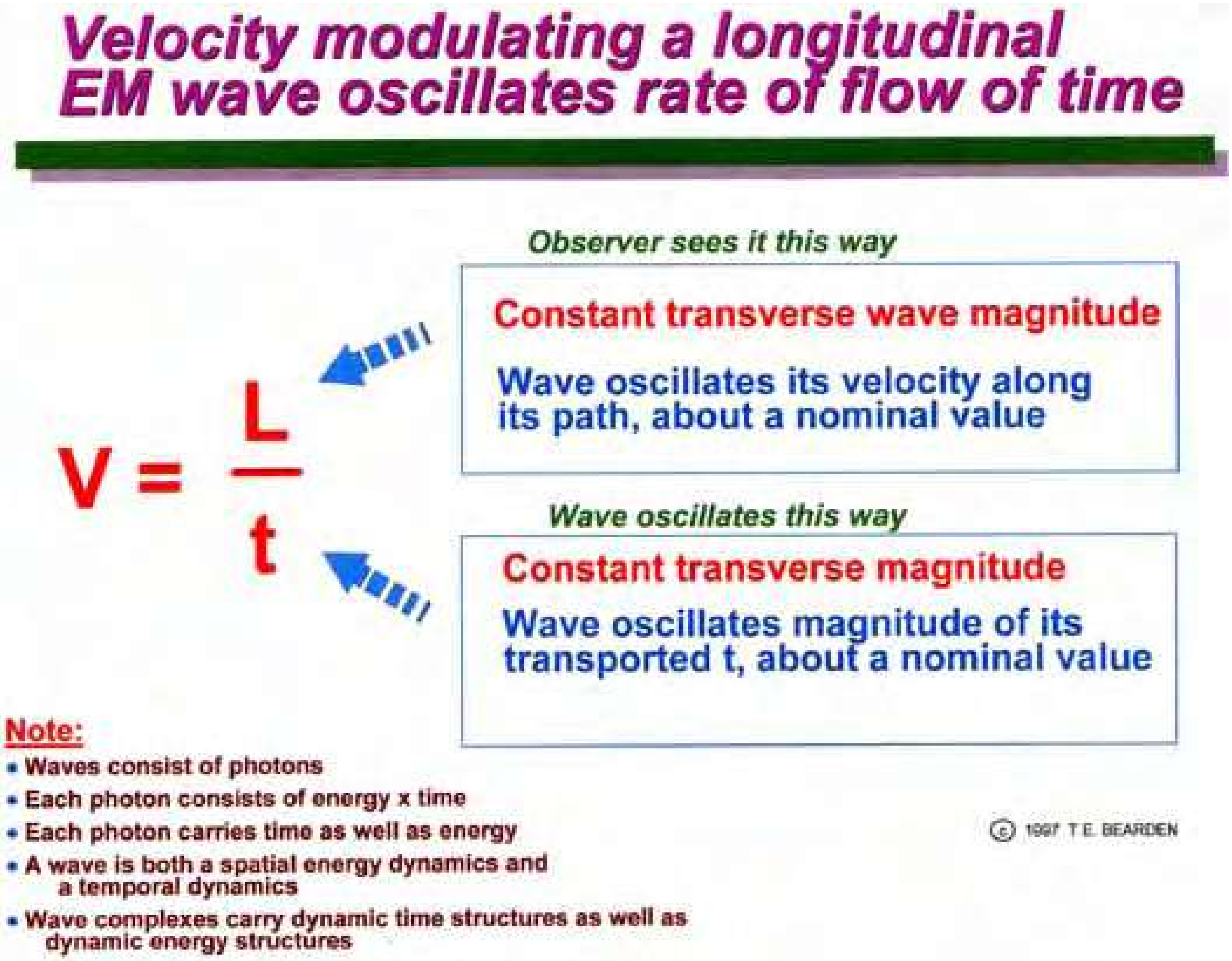

Velocity modulating a longitudinal EM wave oscillates rate of flow of time

$$V = \frac{L}{t}$$

Observer sees it this way

Constant transverse wave magnitude

Wave oscillates its velocity along its path, about a nominal value

Wave oscillates this way

Constant transverse magnitude

Wave oscillates magnitude of its transported t, about a nominal value

Note:

- Waves consist of photons
- Each photon consists of energy x time
- Each photon carries time as well as energy
- A wave is both a spatial energy dynamics and a temporal dynamics
- Wave complexes carry dynamic time structures as well as dynamic energy structures

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

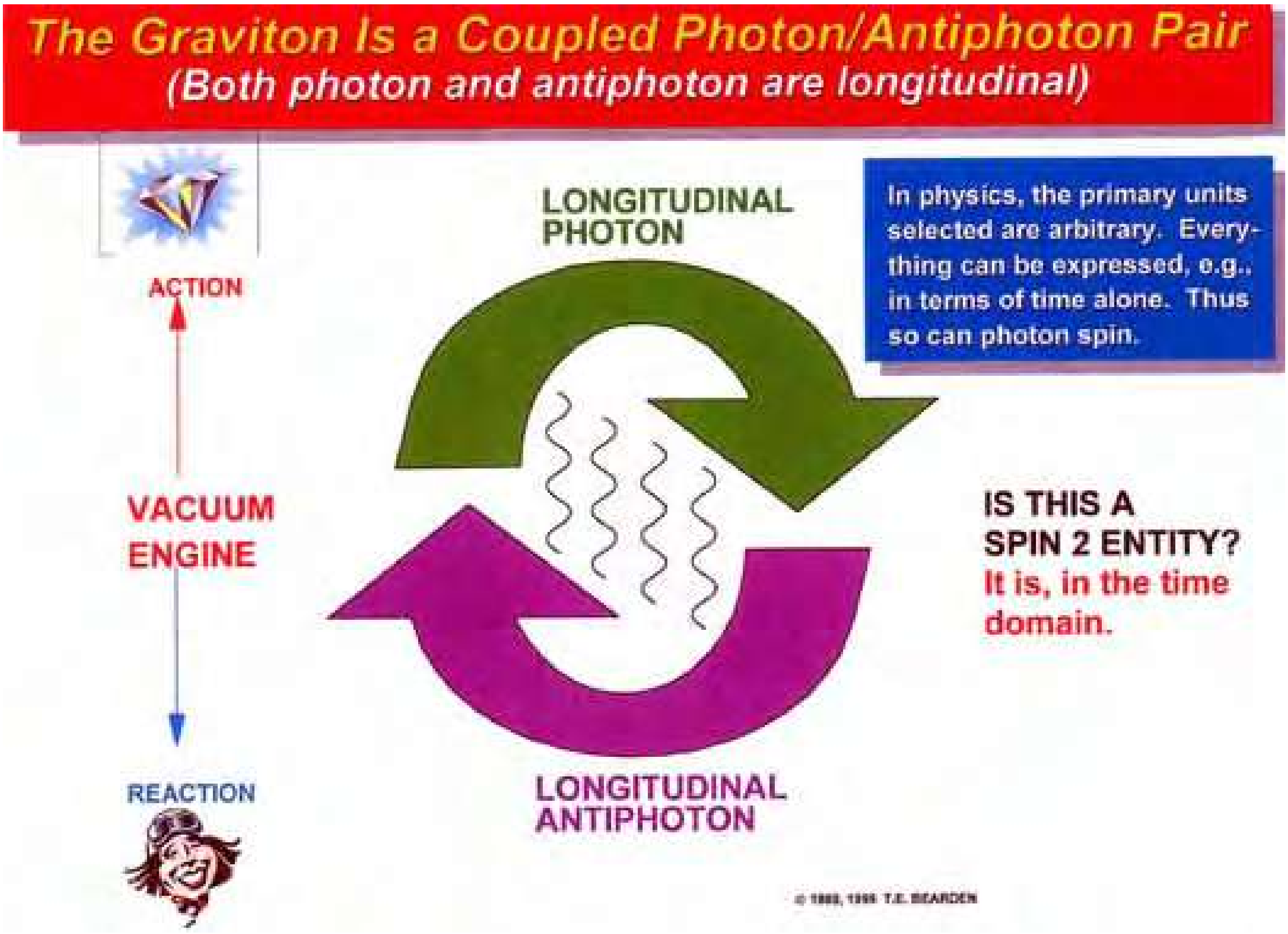

The Graviton Is a Coupled Photon/Antiphoton Pair
(Both photon and antiphoton are longitudinal)

In physics, the primary units selected are arbitrary. Everything can be expressed, e.g., in terms of time alone. Thus so can photon spin.

IS THIS A SPIN 2 ENTITY?
It is, in the time domain.

ACTION
VACUUM ENGINE
REACTION
LONGITUDINAL PHOTON
LONGITUDINAL ANTIPHOTON

© 1988, 1998 T.E. BEARDEN

[Return to Slide Index]
[Next Slide]
[Previous Slide]

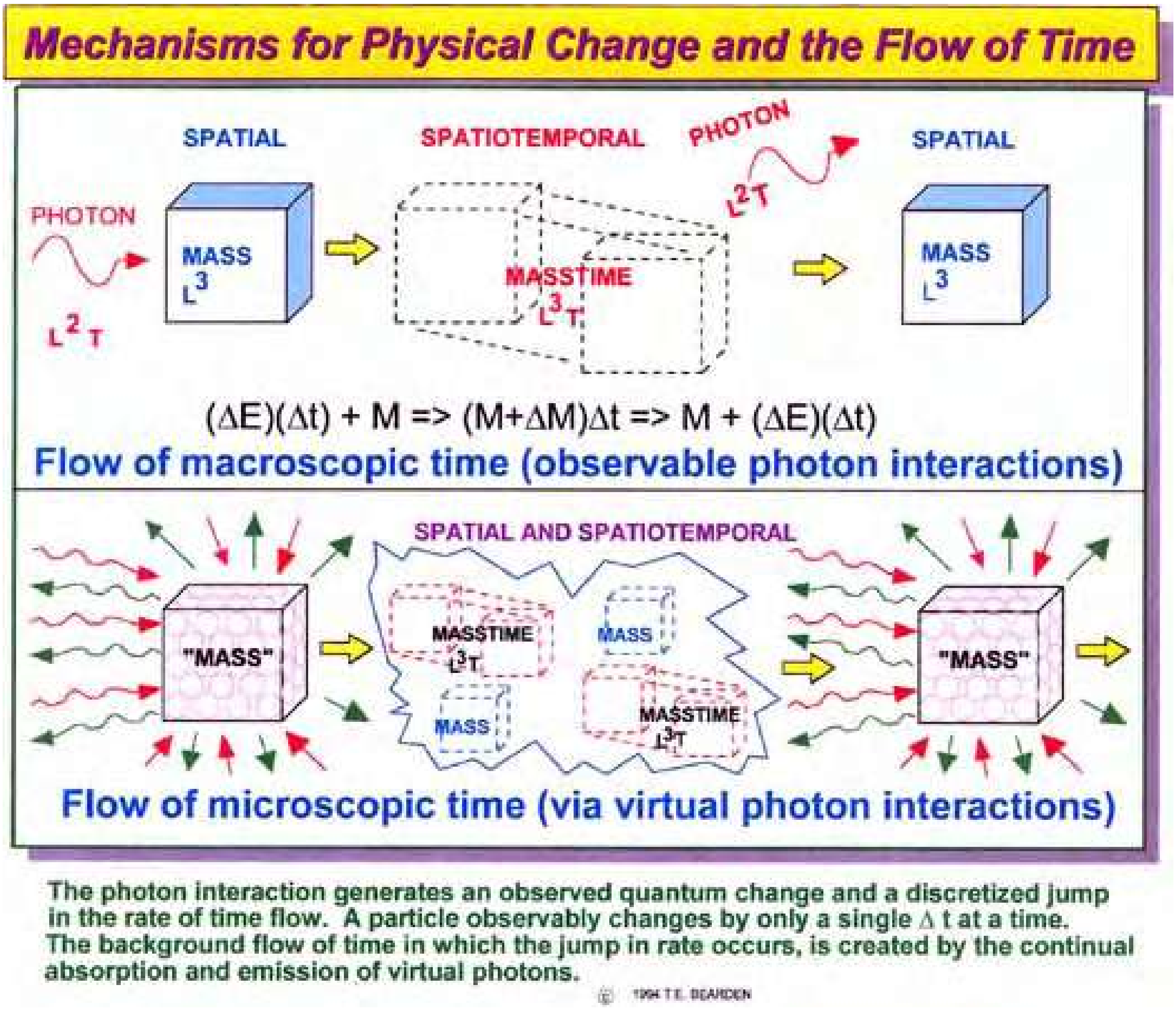

Mechanisms for Physical Change and the Flow of Time

SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS L^3
L^2T
MASSTIME L^3T
L^2T
MASS L^3
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME L^3T
MASS
MASS
MASSTIME L^3T
"MASS"

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δt at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

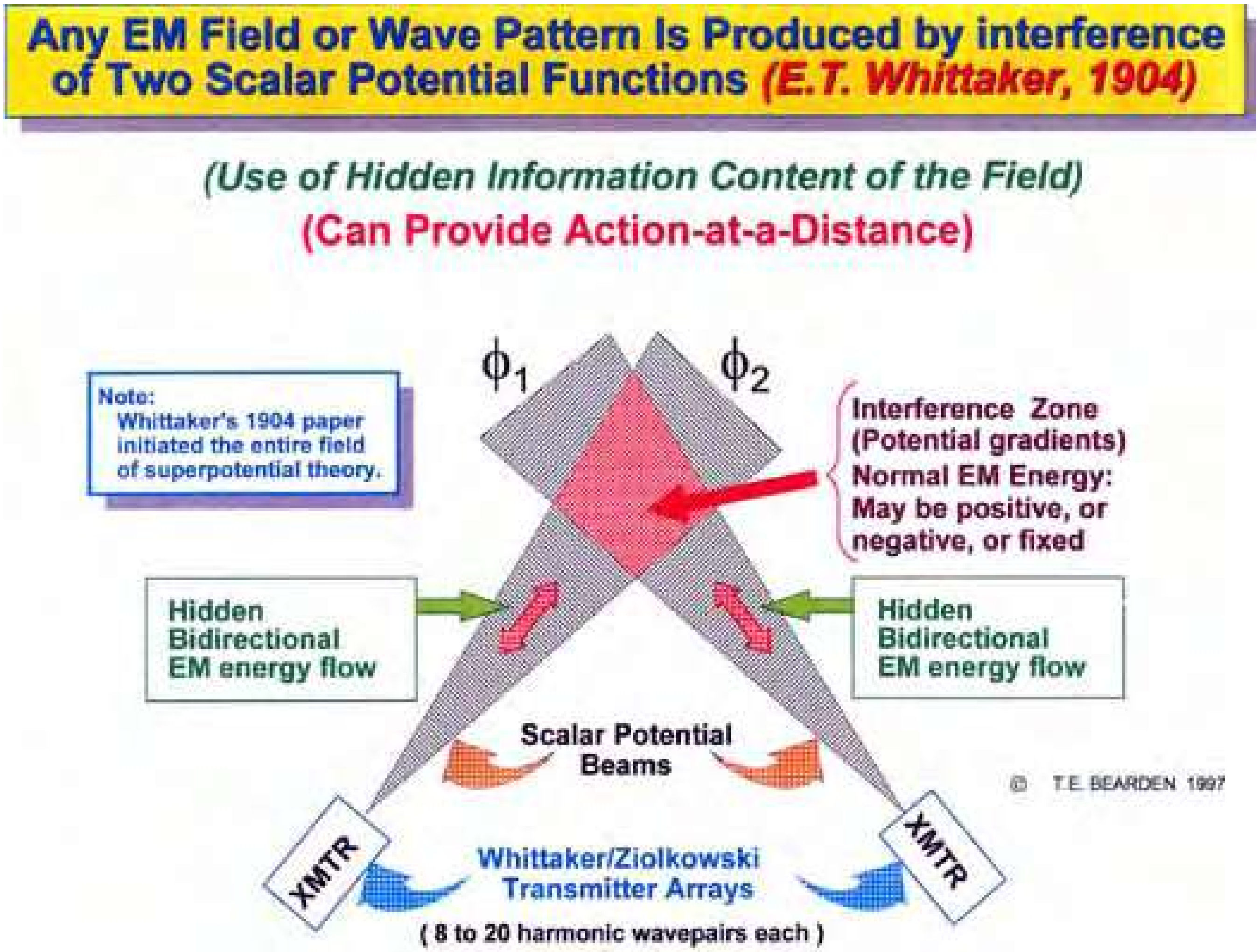

Any EM Field or Wave Pattern Is Produced by interference of Two Scalar Potential Functions *(E.T. Whittaker, 1904)*

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.

ϕ_1

ϕ_2

Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or negative, or fixed

Hidden
Bidirectional
EM energy flow

Hidden
Bidirectional
EM energy flow

Scalar Potential
Beams

© T.E. BEARDEN 1997

XMTR

XMTR

Whittaker/Ziolkowski
Transmitter Arrays

(8 to 20 harmonic wavepairs each)

Return to Slide Index

Next Slide

Previous Slide

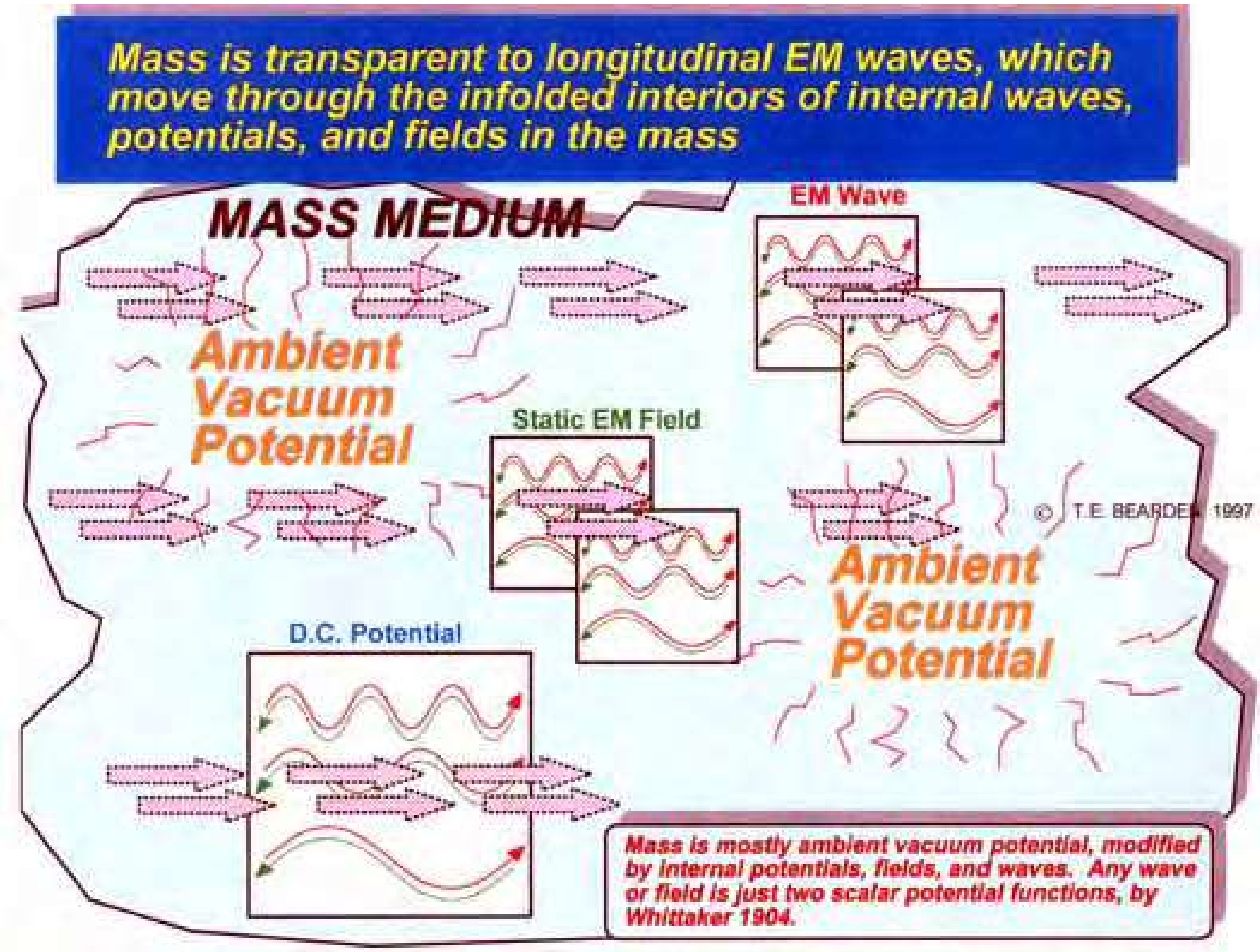

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
Ambient Vacuum Potential
EM Wave
Static EM Field
Ambient Vacuum Potential
D.C. Potential

© T.E. BEARDEN 1997

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

Return to Slide Index

Next Slide

Previous Slide

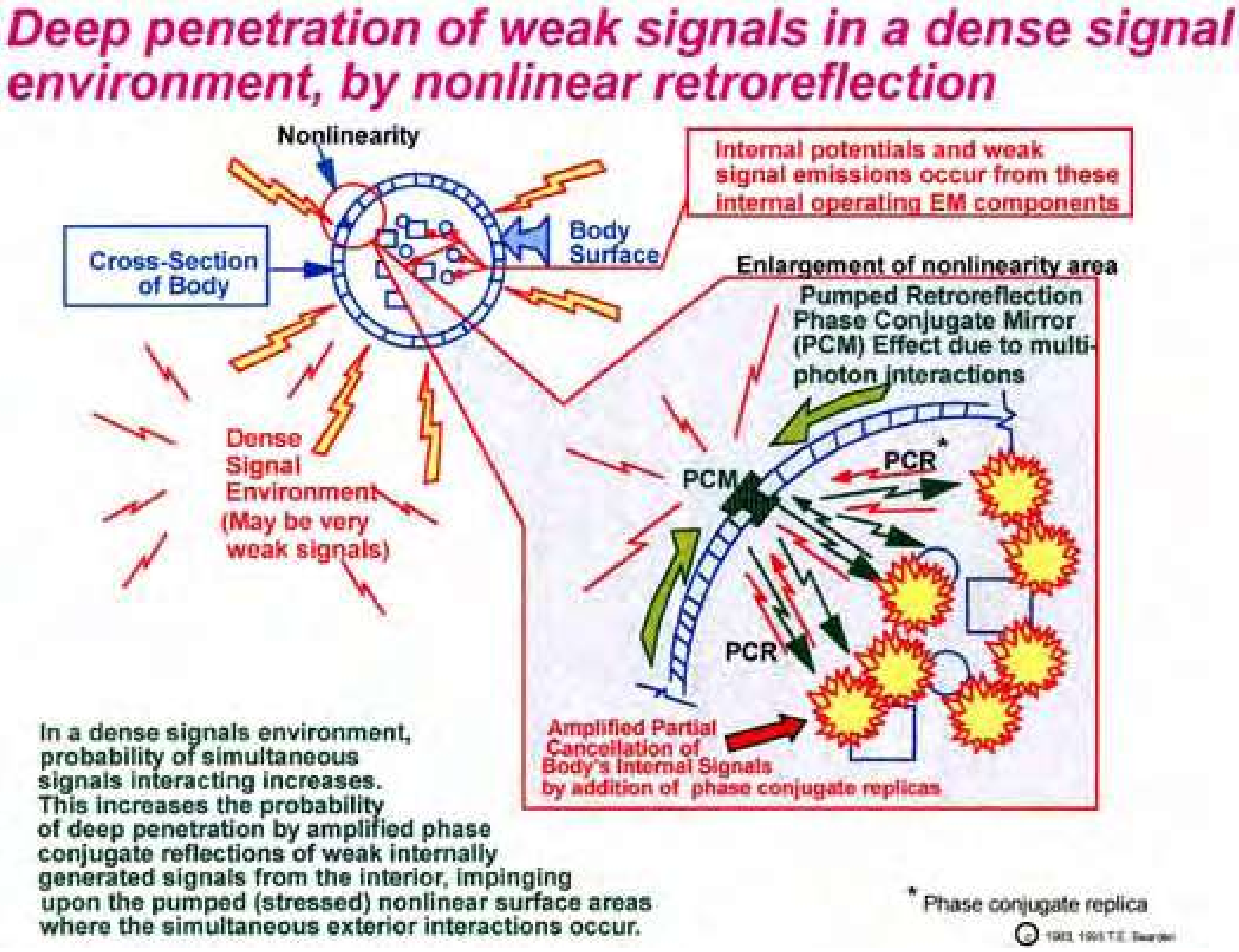

Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection

Nonlinearity

Internal potentials and weak signal emissions occur from these internal operating EM components

Cross-Section of Body

Body Surface

Enlargement of nonlinearity area

Pumped Retroreflection Phase Conjugate Mirror (PCM) Effect due to multi-photon interactions

PCM

PCR*

PCR

Dense Signal Environment (May be very weak signals)

Amplified Partial Cancellation of Body's Internal Signals by addition of phase conjugate replicas

In a dense signals environment, probability of simultaneous signals interacting increases. This increases the probability of deep penetration by amplified phase conjugate reflections of weak internally generated signals from the interior, impinging upon the pumped (stressed) nonlinear surface areas where the simultaneous exterior interactions occur.

* Phase conjugate replica

© 1993, 1995 T.E. Bearden

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

THE IMMUNE SYSTEM

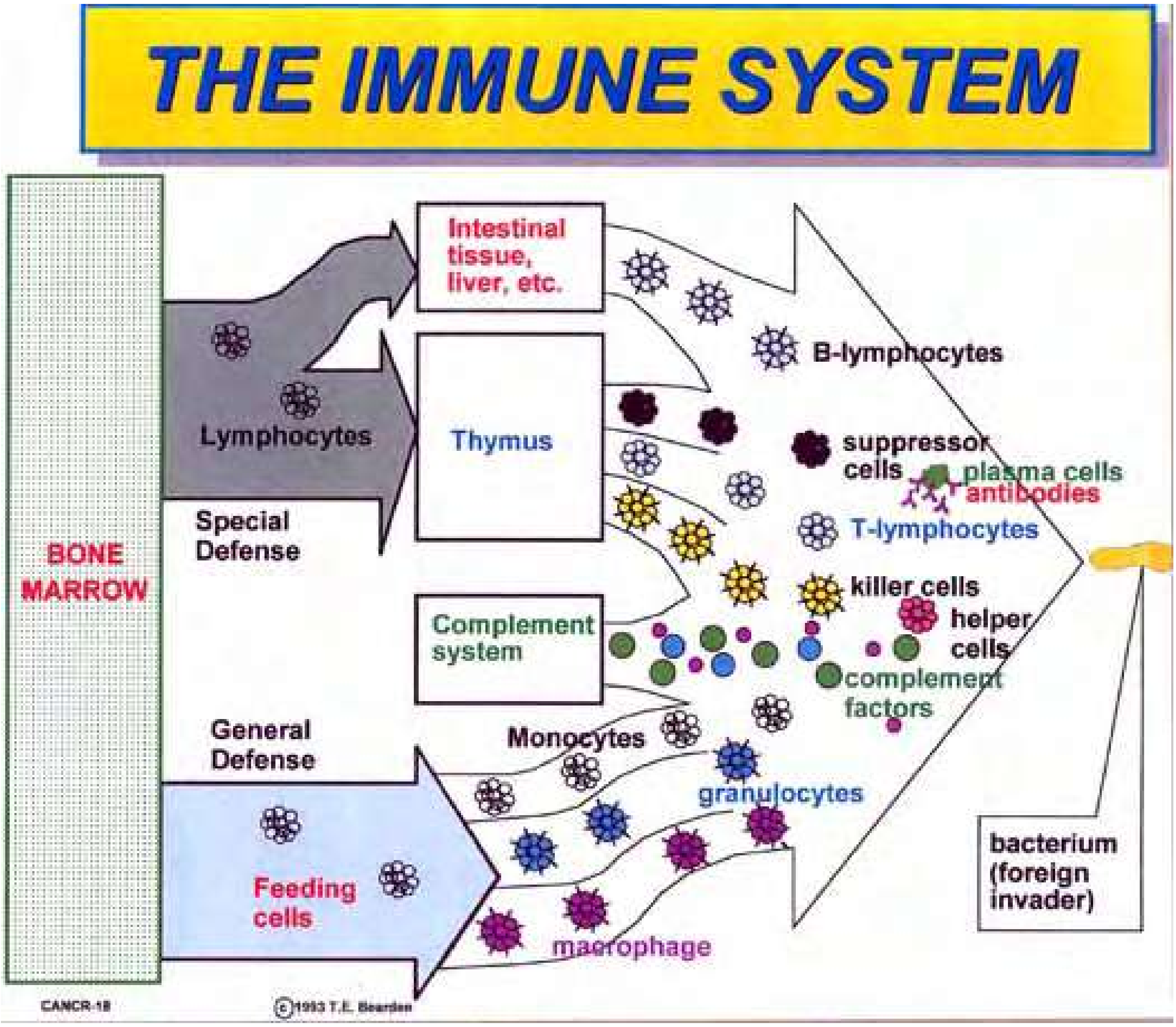
BONE MARROW
Lymphocytes
Special Defense
Intestinal tissue, liver, etc.
Thymus
Complement system
General Defense
Feeding cells
Monocytes
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
granulocytes
macrophage
bacterium (foreign invader)
CANCR-18
©1993 T.E. Bearden

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

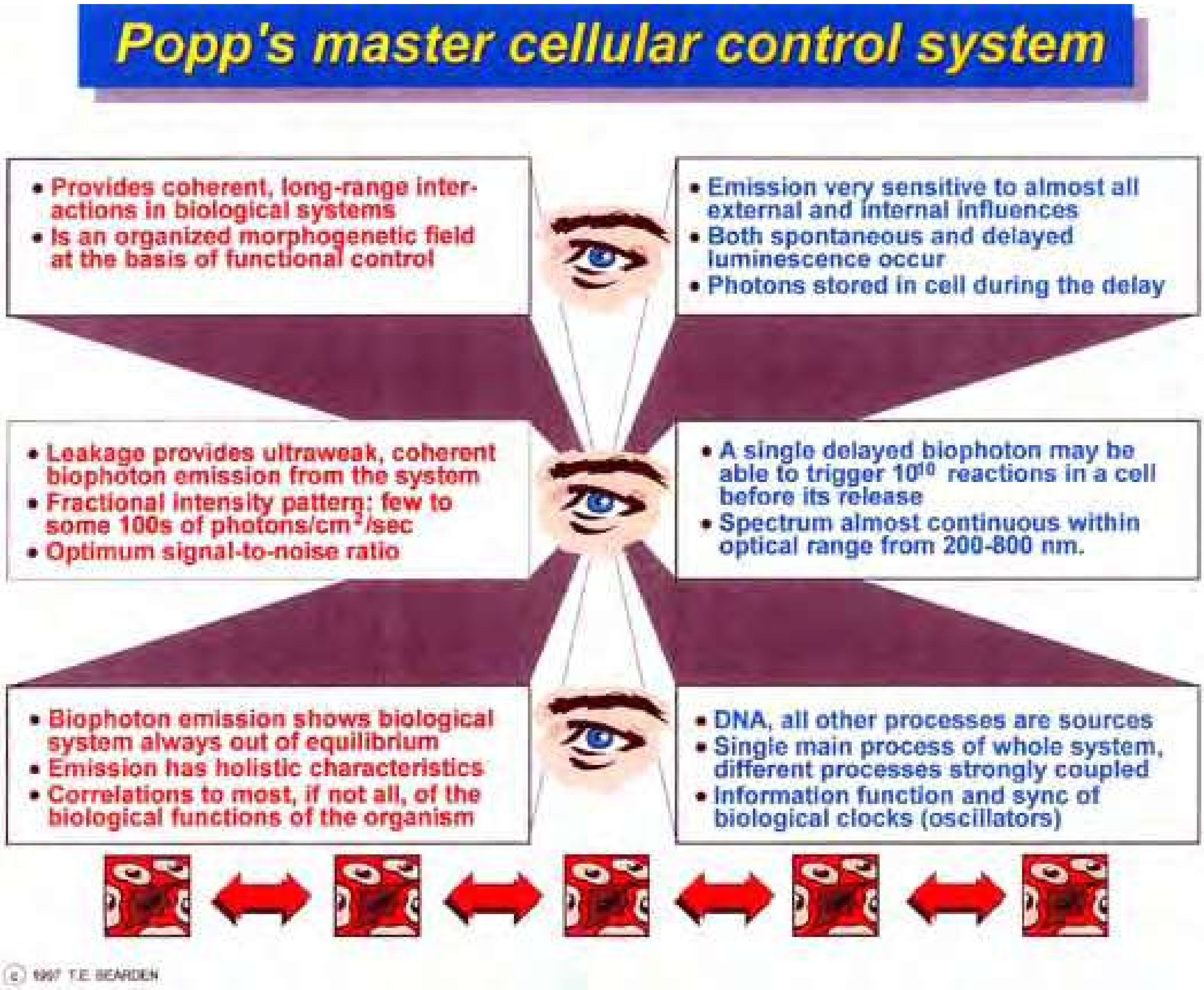

Popp's master cellular control system

- Provides coherent, long-range interactions in biological systems
- Is an organized morphogenetic field at the basis of functional control

- Emission very sensitive to almost all external and internal influences
- Both spontaneous and delayed luminescence occur
- Photons stored in cell during the delay

- Leakage provides ultraweak, coherent biophoton emission from the system
- Fractional intensity pattern: few to some 100s of photons/cm^2/sec
- Optimum signal-to-noise ratio

- A single delayed biophoton may be able to trigger 10^{10} reactions in a cell before its release
- Spectrum almost continuous within optical range from 200-800 nm.

- Biophoton emission shows biological system always out of equilibrium
- Emission has holistic characteristics
- Correlations to most, if not all, of the biological functions of the organism

- DNA, all other processes are sources
- Single main process of whole system, different processes strongly coupled
- Information function and sync of biological clocks (oscillators)

(c) 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

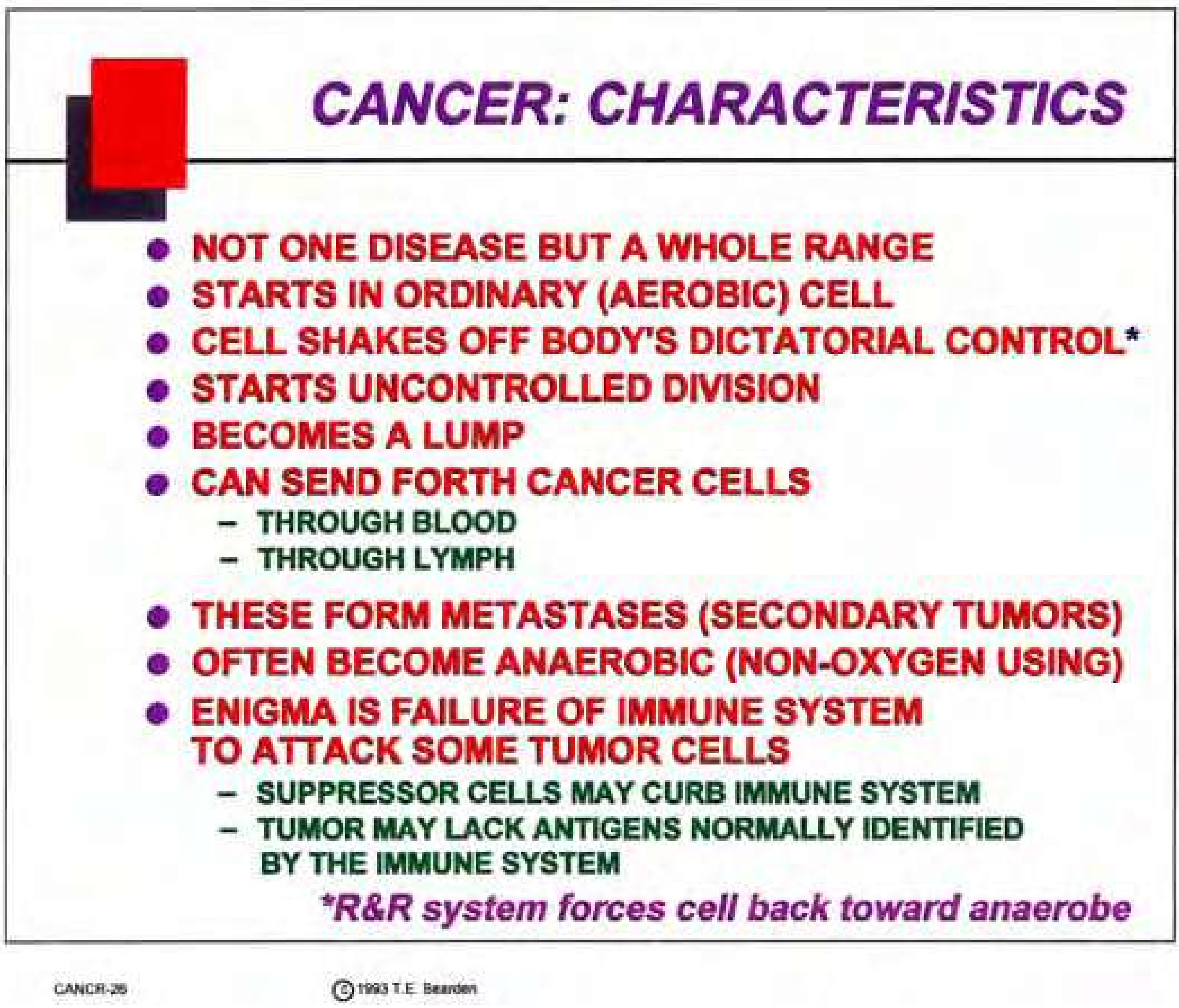

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

**R&R system forces cell back toward anaerobe*

CANCR-26

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

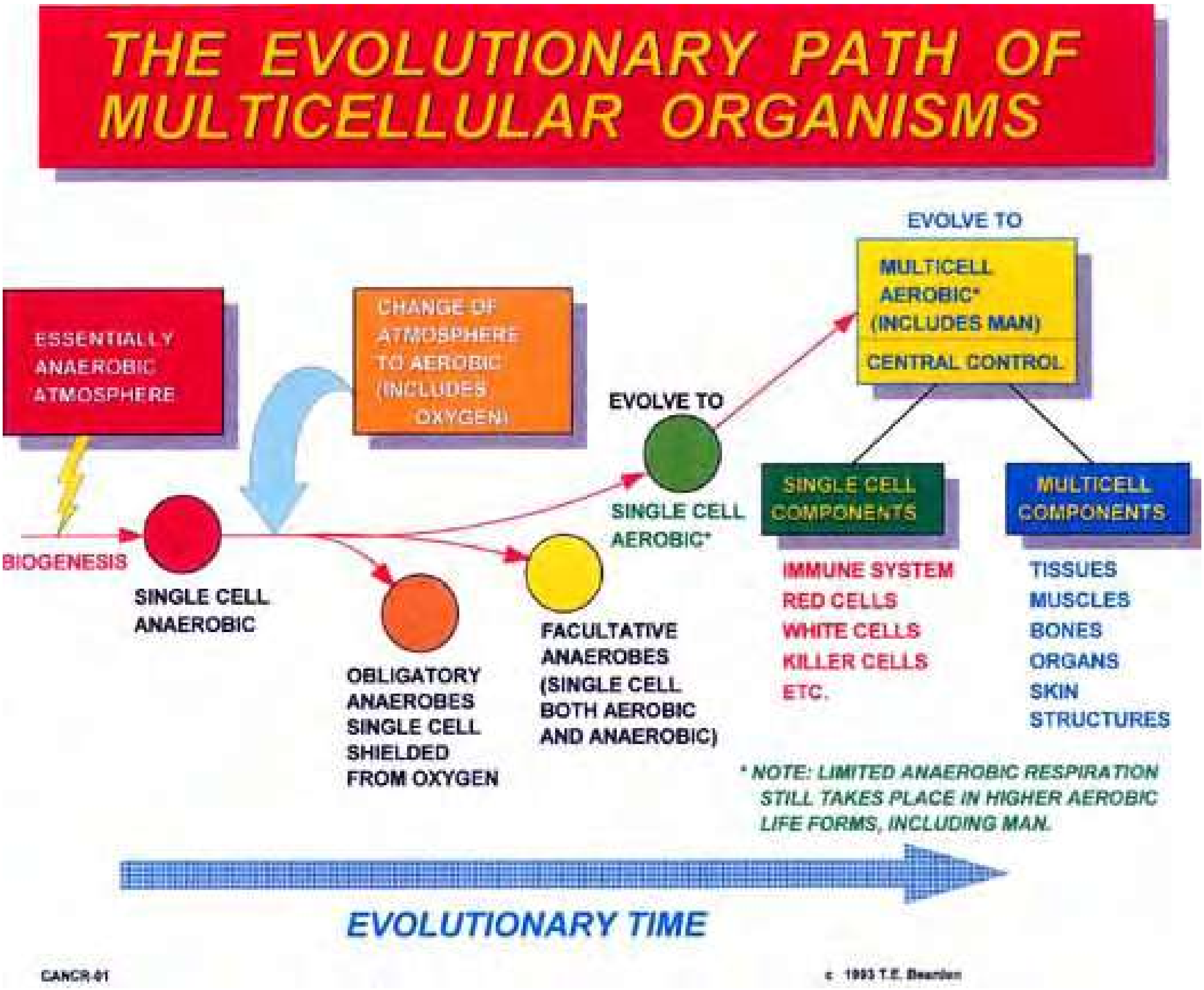

THE EVOLUTIONARY PATH OF MULTICELLULAR ORGANISMS

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

EVOLVE TO

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

BIOGENESIS

SINGLE CELL ANAEROBIC

EVOLVE TO

SINGLE CELL AEROBIC*

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

SINGLE CELL COMPONENTS

IMMUNE SYSTEM
RED CELLS
WHITE CELLS
KILLER CELLS
ETC.

MULTICELL COMPONENTS

TISSUES
MUSCLES
BONES
ORGANS
SKIN
STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

CANCR-01

c 1993 T.E. Bearden

[Return to Slide Index]

[Next Slide]

[Previous Slide]

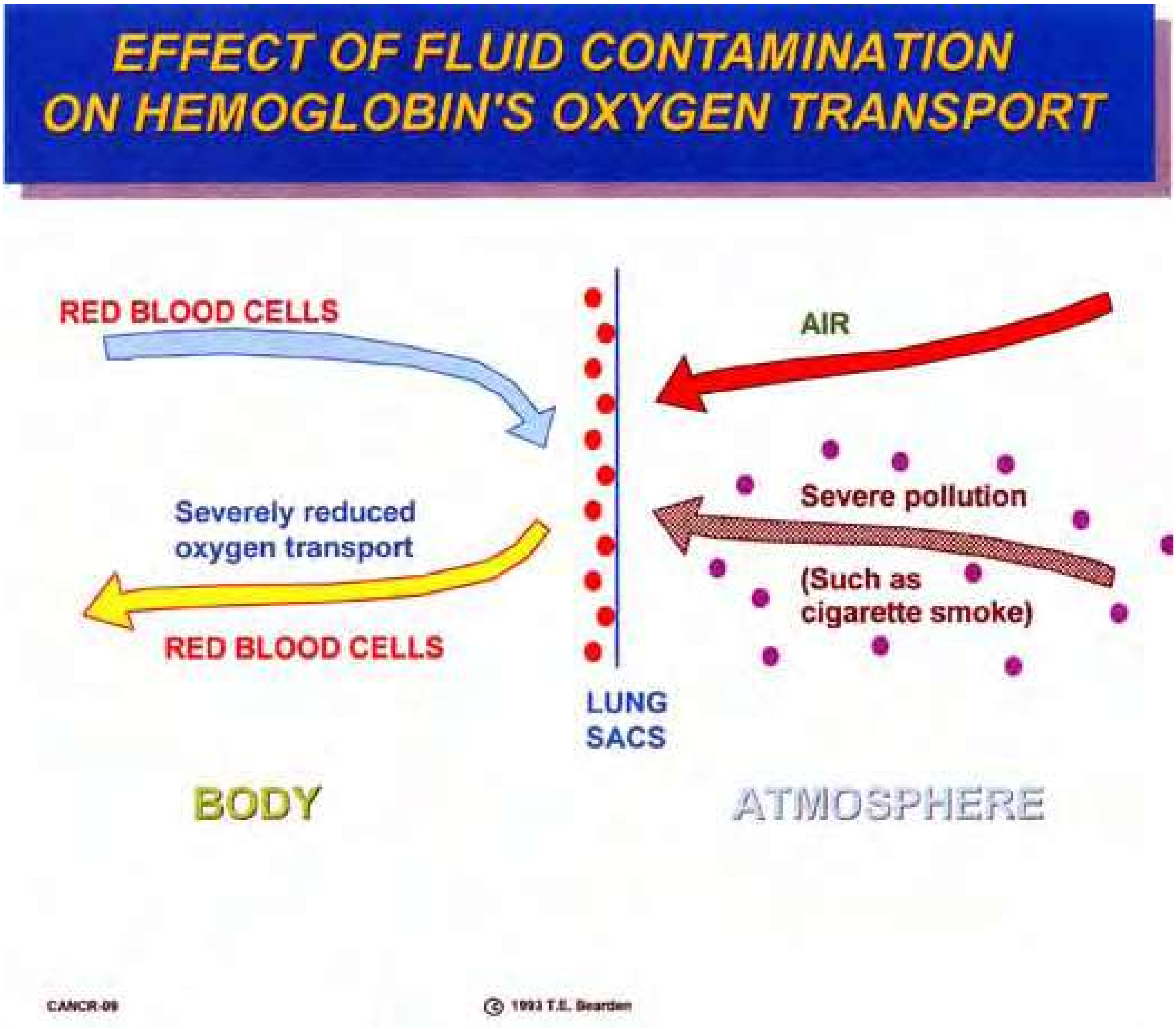

EFFECT OF FLUID CONTAMINATION ON HEMOGLOBIN'S OXYGEN TRANSPORT

RED BLOOD CELLS

AIR

Severe pollution

(Such as cigarette smoke)

Severely reduced oxygen transport

RED BLOOD CELLS

LUNG SACS

BODY

ATMOSPHERE

CANCR-09

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

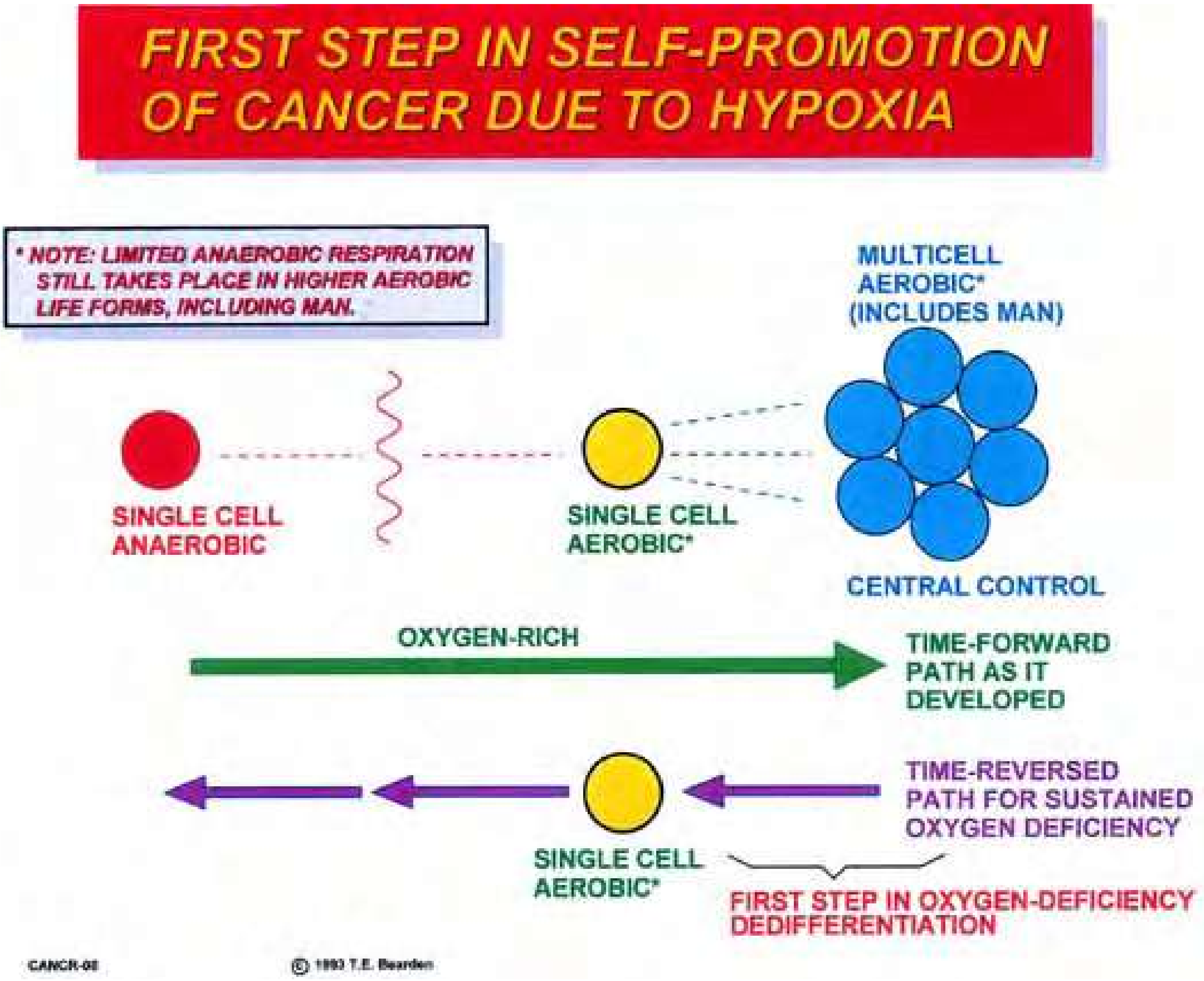

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

SINGLE CELL ANAEROBIC

SINGLE CELL AEROBIC*

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD PATH AS IT DEVELOPED

TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY

SINGLE CELL AEROBIC*

FIRST STEP IN OXYGEN-DEFICIENCY DEDIFFERENTIATION

CANCR-08

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

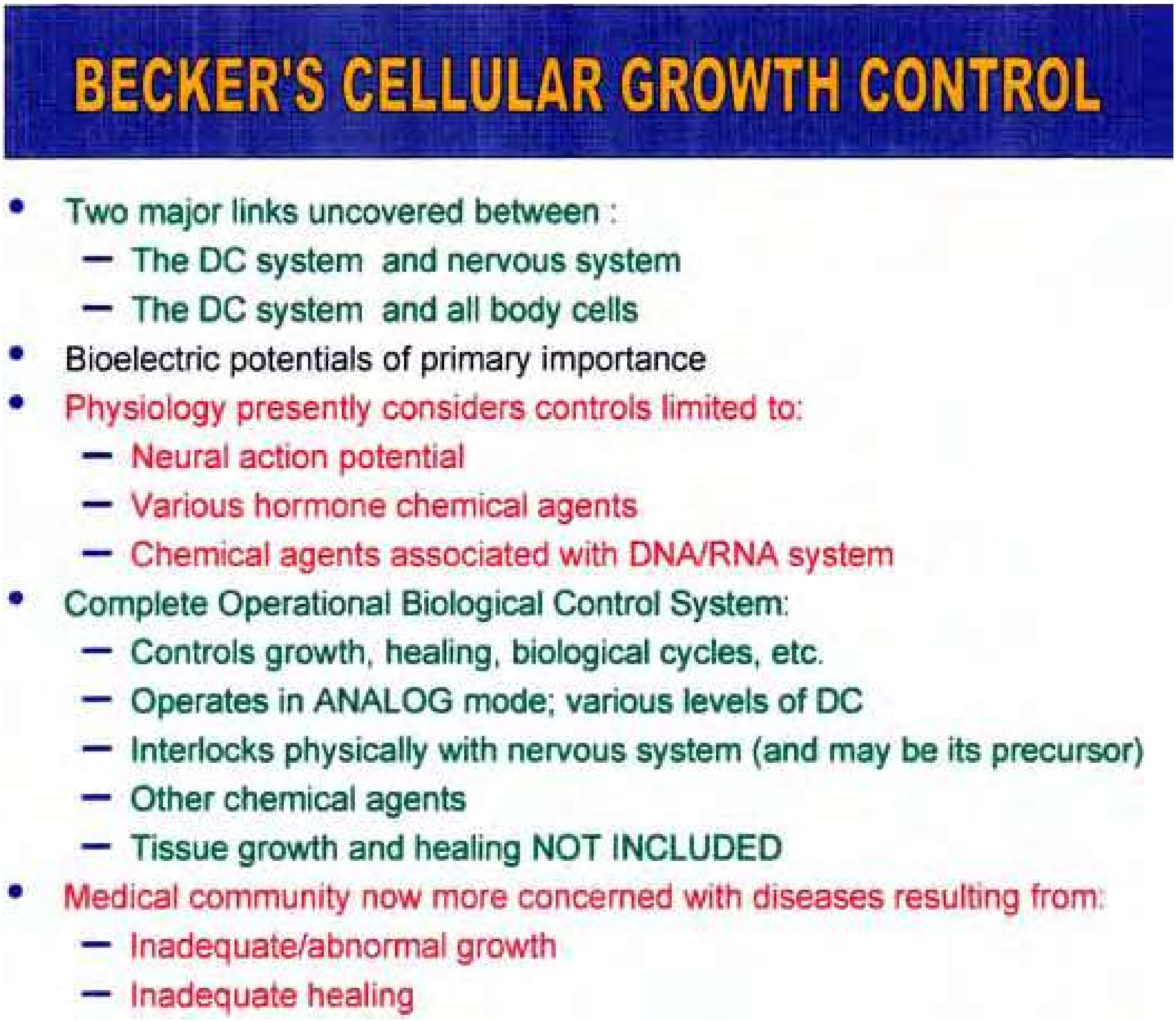

BECKER'S CELLULAR GROWTH CONTROL

- Two major links uncovered between :
  - The DC system and nervous system
  - The DC system and all body cells
- Bioelectric potentials of primary importance
- Physiology presently considers controls limited to:
  - Neural action potential
  - Various hormone chemical agents
  - Chemical agents associated with DNA/RNA system
- Complete Operational Biological Control System:
  - Controls growth, healing, biological cycles, etc.
  - Operates in ANALOG mode; various levels of DC
  - Interlocks physically with nervous system (and may be its precursor)
  - Other chemical agents
  - Tissue growth and healing NOT INCLUDED
- Medical community now more concerned with diseases resulting from:
  - Inadequate/abnormal growth
  - Inadequate healing

Return to Slide Index

Next Slide

Previous Slide

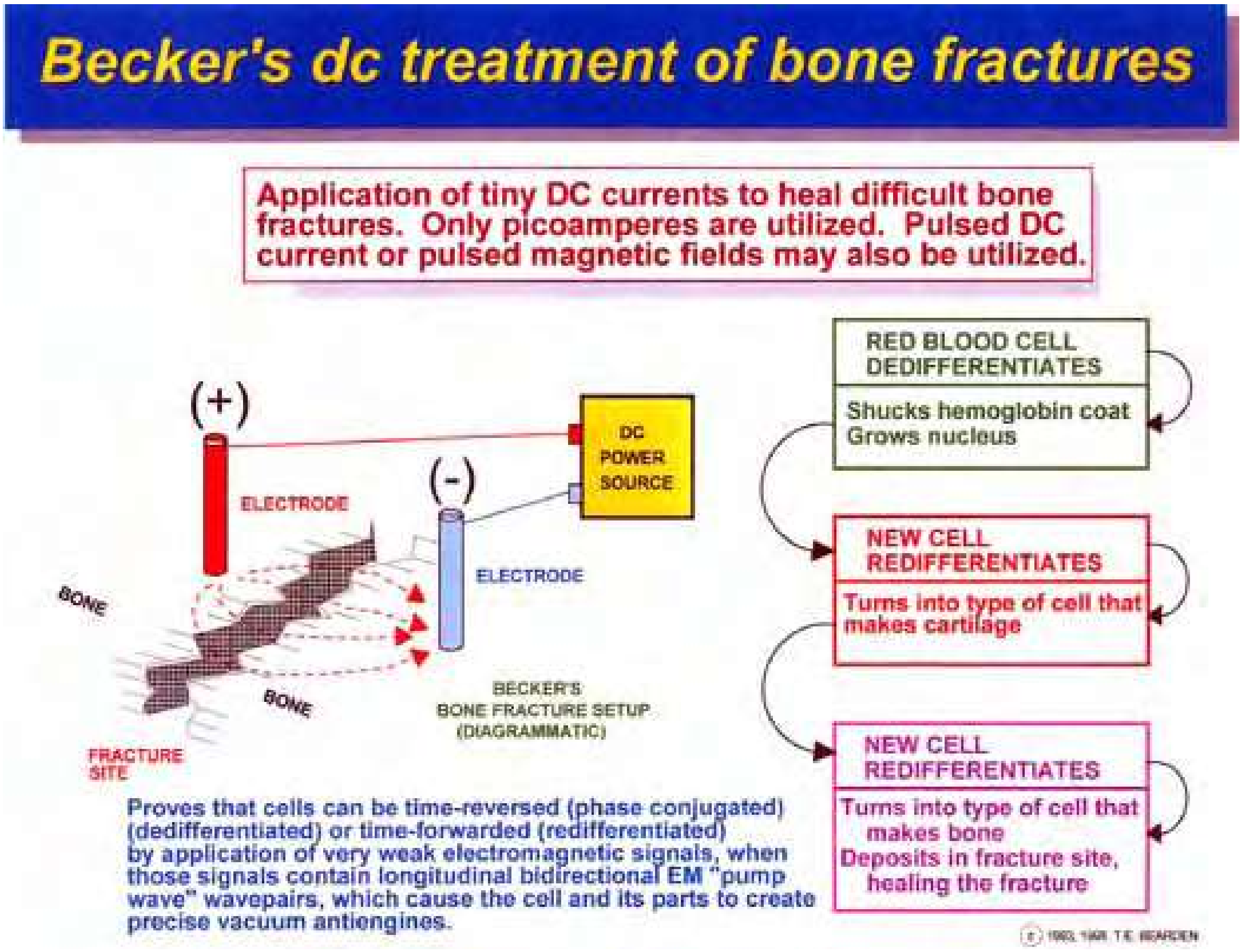

Becker's dc treatment of bone fractures

Application of tiny DC currents to heal difficult bone fractures. Only picoamperes are utilized. Pulsed DC current or pulsed magnetic fields may also be utilized.

BECKER'S
BONE FRACTURE SETUP
(DIAGRAMMATIC)

Proves that cells can be time-reversed (phase conjugated) (dedifferentiated) or time-forwarded (redifferentiated) by application of very weak electromagnetic signals, when those signals contain longitudinal bidirectional EM "pump wave" wavepairs, which cause the cell and its parts to create precise vacuum antiengines.

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

Becker's theoretical DC control system involved with response to injury*

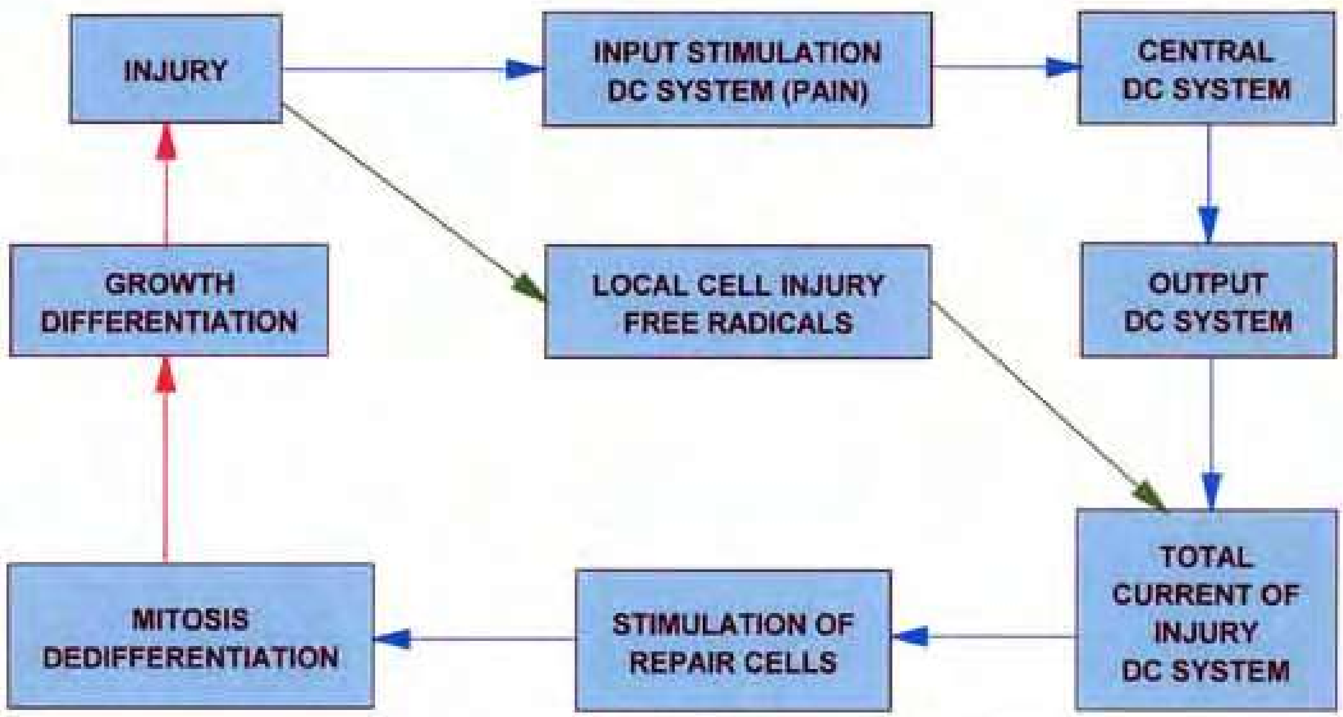

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

STIMULATION OF REPAIR CELLS

MITOSIS DEDIFFERENTIATION

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Return to Slide Index

Next Slide

Previous Slide

*Becker's theoretical control system governing regeneration**

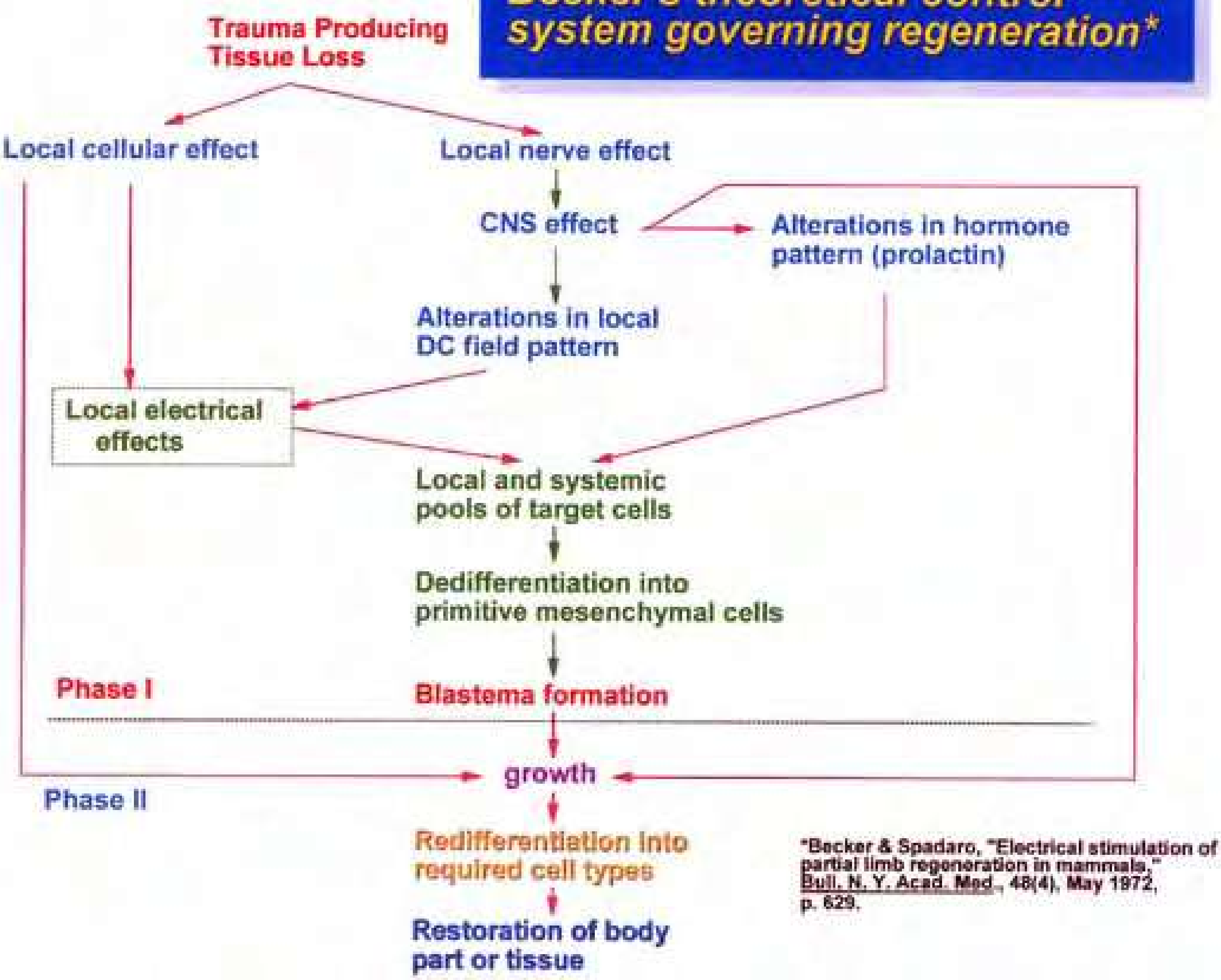

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

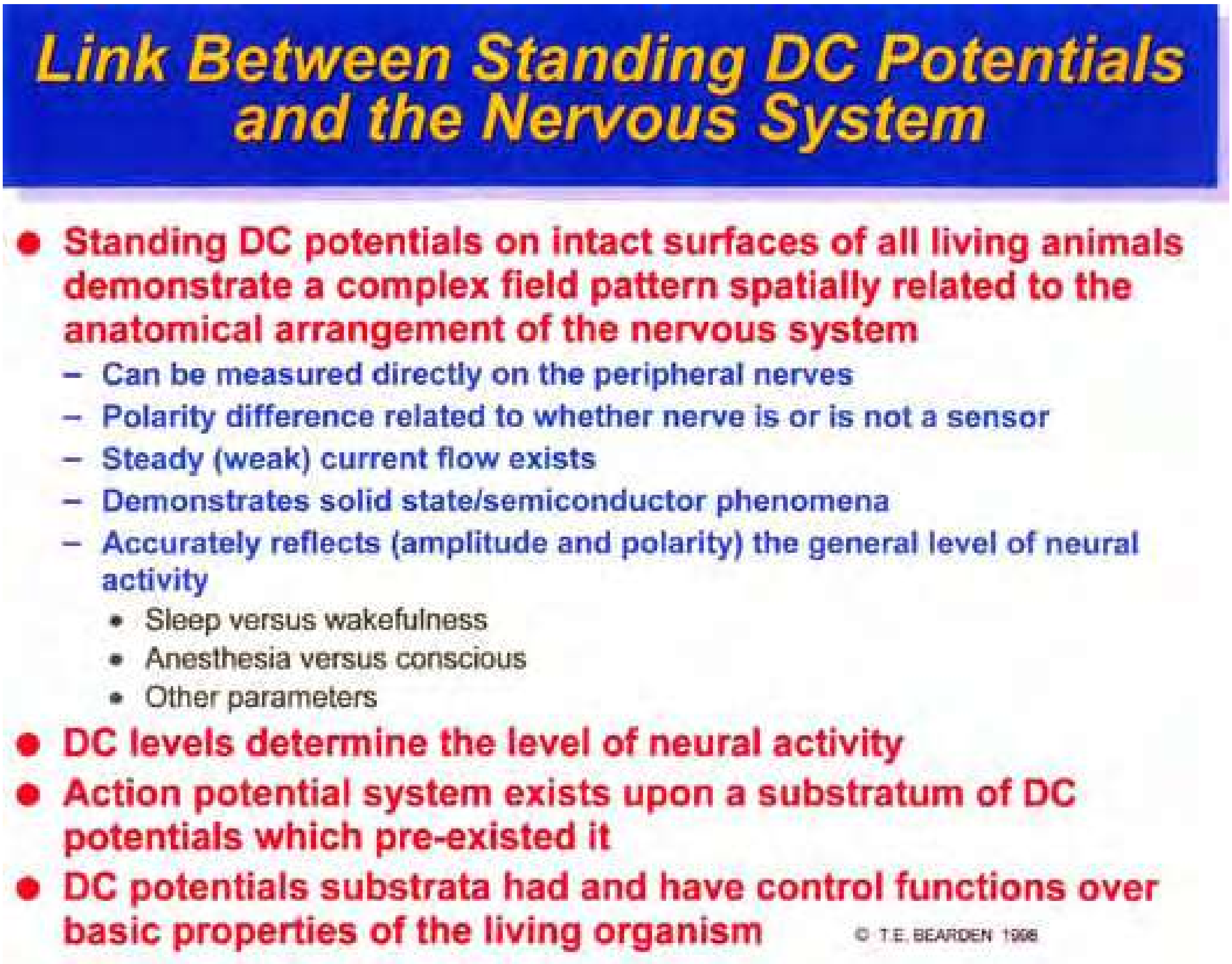

Link Between Standing DC Potentials and the Nervous System

- Standing DC potentials on intact surfaces of all living animals demonstrate a complex field pattern spatially related to the anatomical arrangement of the nervous system
  - Can be measured directly on the peripheral nerves
  - Polarity difference related to whether nerve is or is not a sensor
  - Steady (weak) current flow exists
  - Demonstrates solid state/semiconductor phenomena
  - Accurately reflects (amplitude and polarity) the general level of neural activity
    - Sleep versus wakefulness
    - Anesthesia versus conscious
    - Other parameters
- DC levels determine the level of neural activity
- Action potential system exists upon a substratum of DC potentials which pre-existed it
- DC potentials substrata had and have control functions over basic properties of the living organism

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

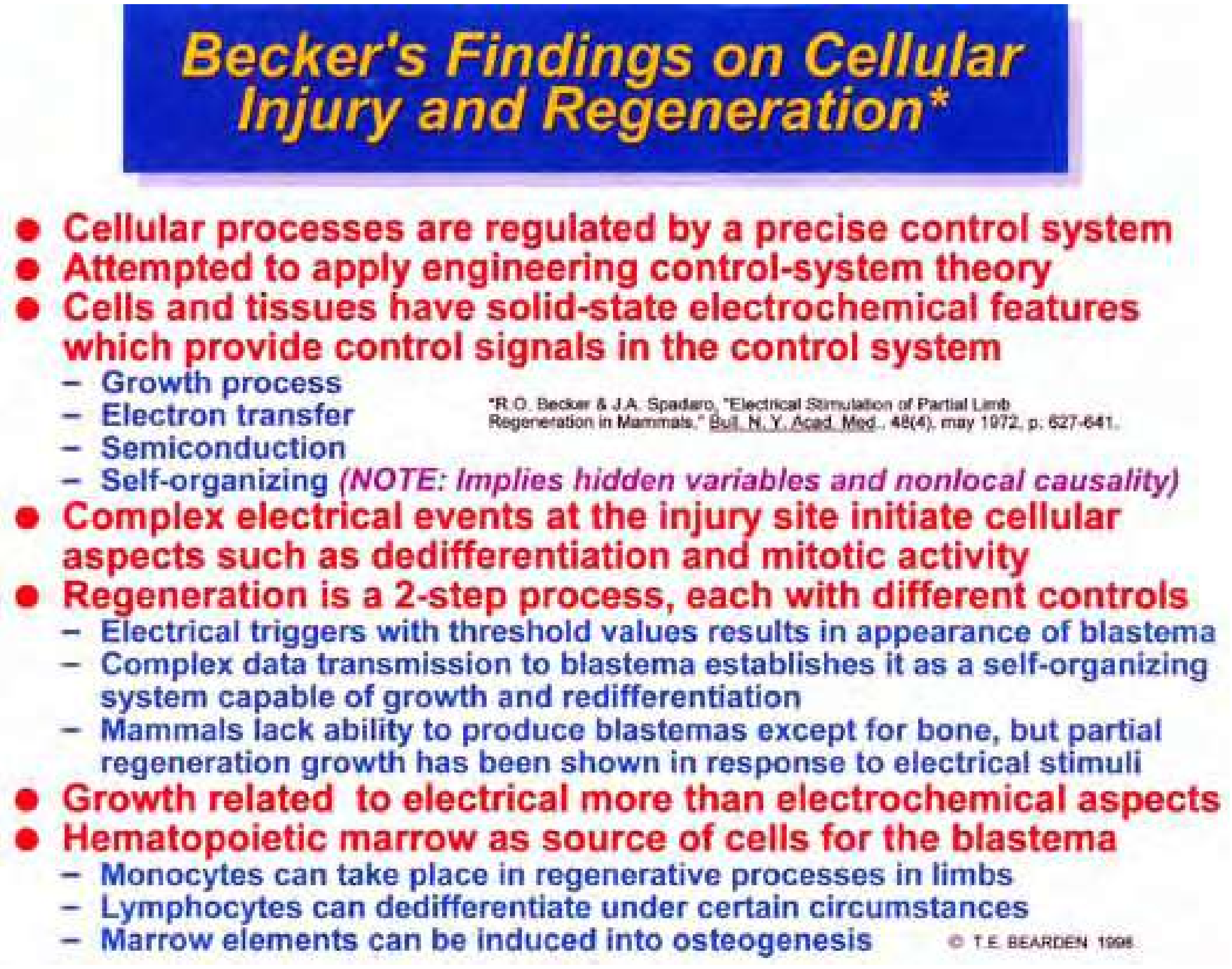

Becker's Findings on Cellular Injury and Regeneration*

- Cellular processes are regulated by a precise control system
- Attempted to apply engineering control-system theory
- Cells and tissues have solid-state electrochemical features which provide control signals in the control system
  - Growth process
  - Electron transfer
  - Semiconduction
  - Self-organizing *(NOTE: Implies hidden variables and nonlocal causality)*
- Complex electrical events at the injury site initiate cellular aspects such as dedifferentiation and mitotic activity
- Regeneration is a 2-step process, each with different controls
  - Electrical triggers with threshold values results in appearance of blastema
  - Complex data transmission to blastema establishes it as a self-organizing system capable of growth and redifferentiation
  - Mammals lack ability to produce blastemas except for bone, but partial regeneration growth has been shown in response to electrical stimuli
- Growth related to electrical more than electrochemical aspects
- Hematopoietic marrow as source of cells for the blastema
  - Monocytes can take place in regenerative processes in limbs
  - Lymphocytes can dedifferentiate under certain circumstances
  - Marrow elements can be induced into osteogenesis

*R.O. Becker & J.A. Spadaro, "Electrical Stimulation of Partial Limb Regeneration in Mammals," Bull. N. Y. Acad. Med., 48(4), may 1972, p. 627-641.

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

Corrections For the Present EM Bioeffects Model

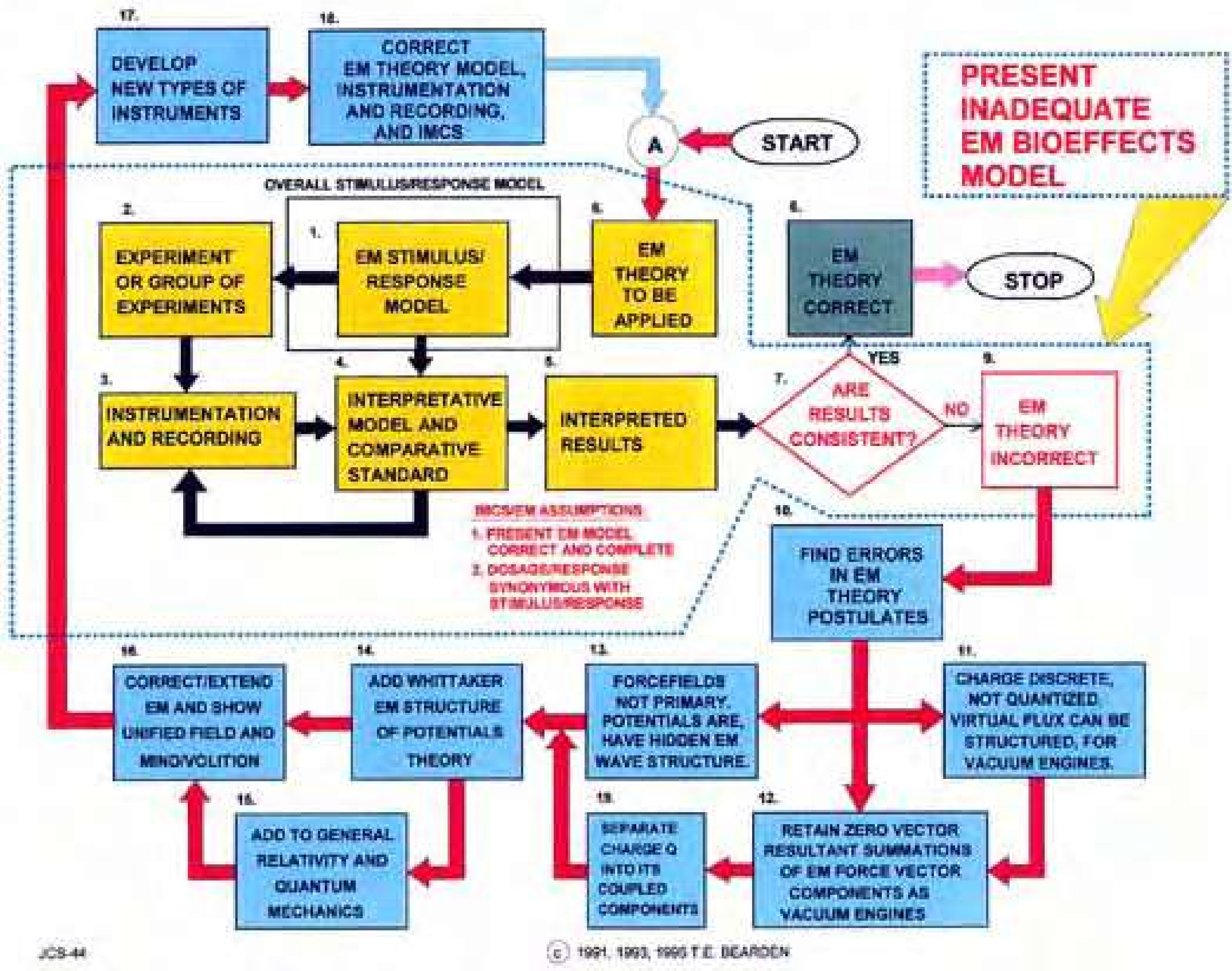

PRESENT INADEQUATE EM BIOEFFECTS MODEL

START

A

OVERALL STIMULUS/RESPONSE MODEL

1. EM STIMULUS/ RESPONSE MODEL

2. EXPERIMENT OR GROUP OF EXPERIMENTS

3. INSTRUMENTATION AND RECORDING

4. INTERPRETATIVE MODEL AND COMPARATIVE STANDARD

5. INTERPRETED RESULTS

6. EM THEORY TO BE APPLIED

7. ARE RESULTS CONSISTENT?

YES

8. EM THEORY CORRECT

STOP

NO

9. EM THEORY INCORRECT

IMCS/EM ASSUMPTIONS

1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE

10. FIND ERRORS IN EM THEORY POSTULATES

11. CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.

12. RETAIN ZERO VECTOR RESULTANT SUMMATIONS OF EM FORCE VECTOR COMPONENTS AS VACUUM ENGINES

13. FORCEFIELDS NOT PRIMARY. POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.

14. ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY

15. ADD TO GENERAL RELATIVITY AND QUANTUM MECHANICS

16. CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION

17. DEVELOP NEW TYPES OF INSTRUMENTS

18. CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS

19. SEPARATE CHARGE Q INTO ITS COUPLED COMPONENTS

JCS-44

© 1991, 1993, 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

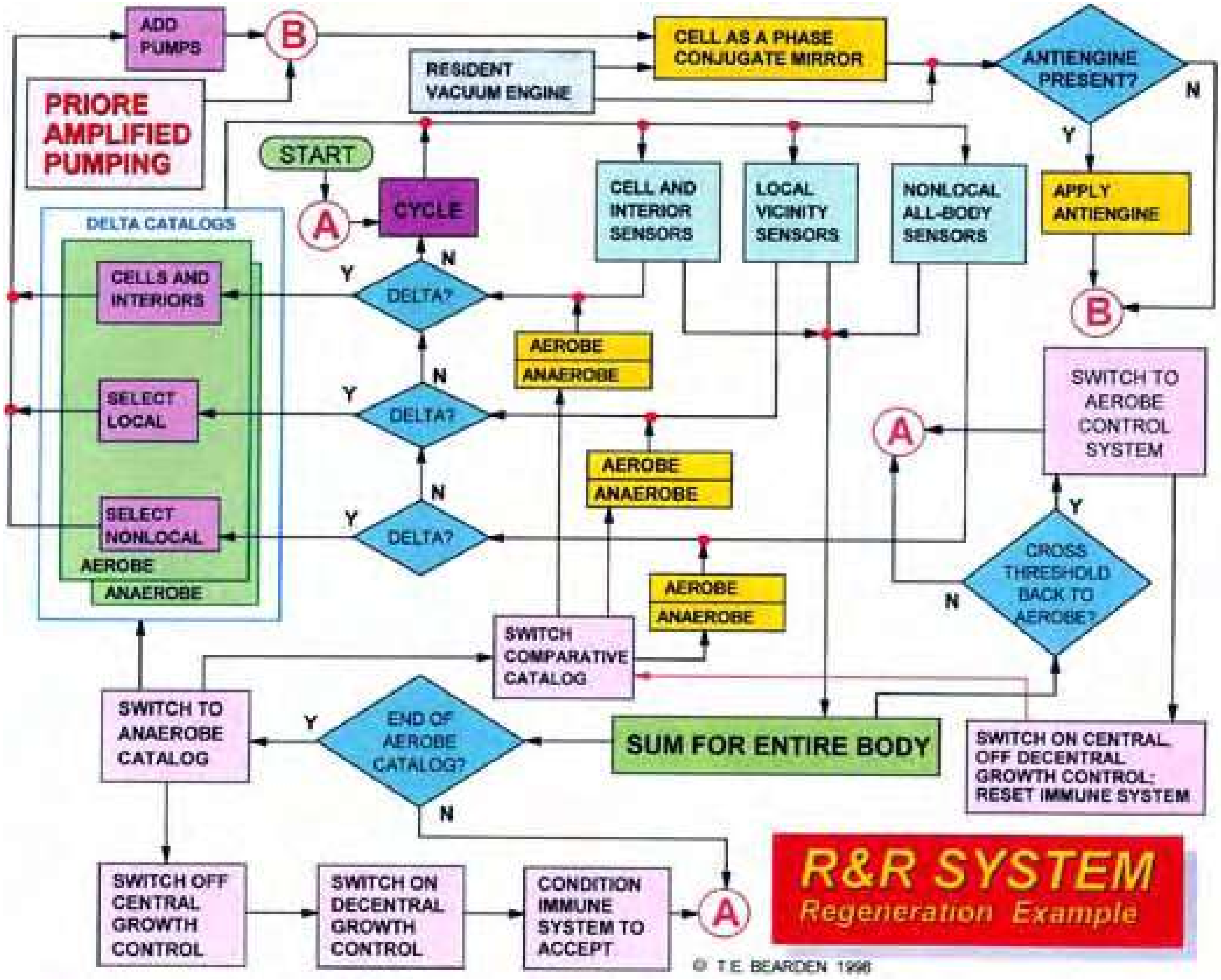

ADD PUMPS

B

CELL AS A PHASE CONJUGATE MIRROR

RESIDENT VACUUM ENGINE

ANTIENGINE PRESENT?

N

Y

PRIORE AMPLIFIED PUMPING

START

A

CYCLE

CELL AND INTERIOR SENSORS

LOCAL VICINITY SENSORS

NONLOCAL ALL-BODY SENSORS

APPLY ANTIENGINE

DELTA CATALOGS

CELLS AND INTERIORS

Y

DELTA?

N

AEROBE

ANAEROBE

B

SELECT LOCAL

Y

DELTA?

N

SWITCH TO AEROBE CONTROL SYSTEM

A

AEROBE

ANAEROBE

SELECT NONLOCAL

Y

DELTA?

N

AEROBE

ANAEROBE

CROSS THRESHOLD BACK TO AEROBE?

Y

N

AEROBE

ANAEROBE

SWITCH COMPARATIVE CATALOG

SWITCH TO ANAEROBE CATALOG

Y

END OF AEROBE CATALOG?

N

SUM FOR ENTIRE BODY

SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM

SWITCH OFF CENTRAL GROWTH CONTROL

SWITCH ON DECENTRAL GROWTH CONTROL

CONDITION IMMUNE SYSTEM TO ACCEPT

A

R&R SYSTEM

Regeneration Example

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

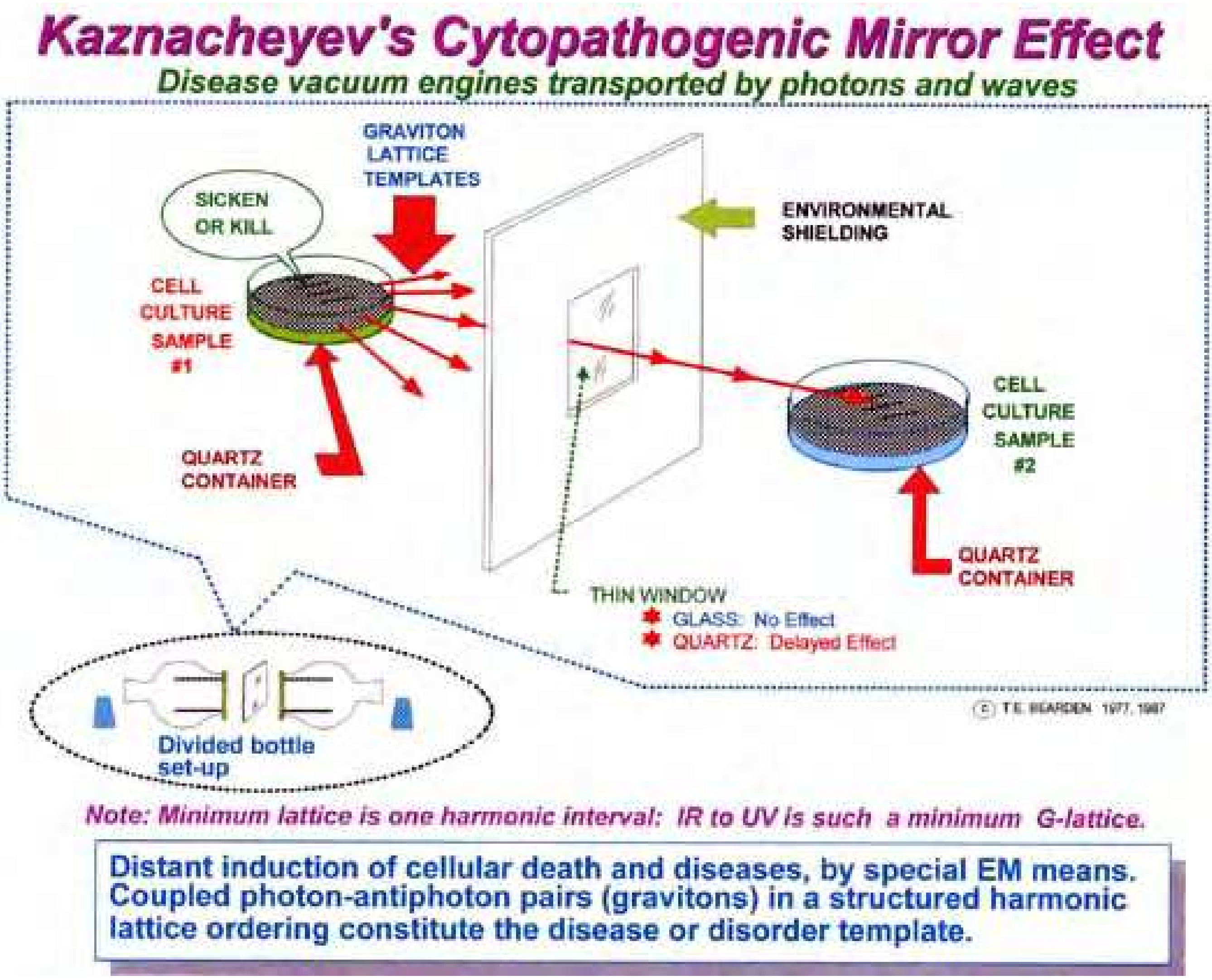

Kaznacheyev's Cytopathogenic Mirror Effect

Disease vacuum engines transported by photons and waves

GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
CELL
CULTURE
SAMPLE
#1
QUARTZ
CONTAINER
ENVIRONMENTAL
SHIELDING
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle
set-up
© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Distant induction of cellular death and diseases, by special EM means. Coupled photon-antiphoton pairs (gravitons) in a structured harmonic lattice ordering constitute the disease or disorder template.

Return to Slide Index

Next Slide

Previous Slide

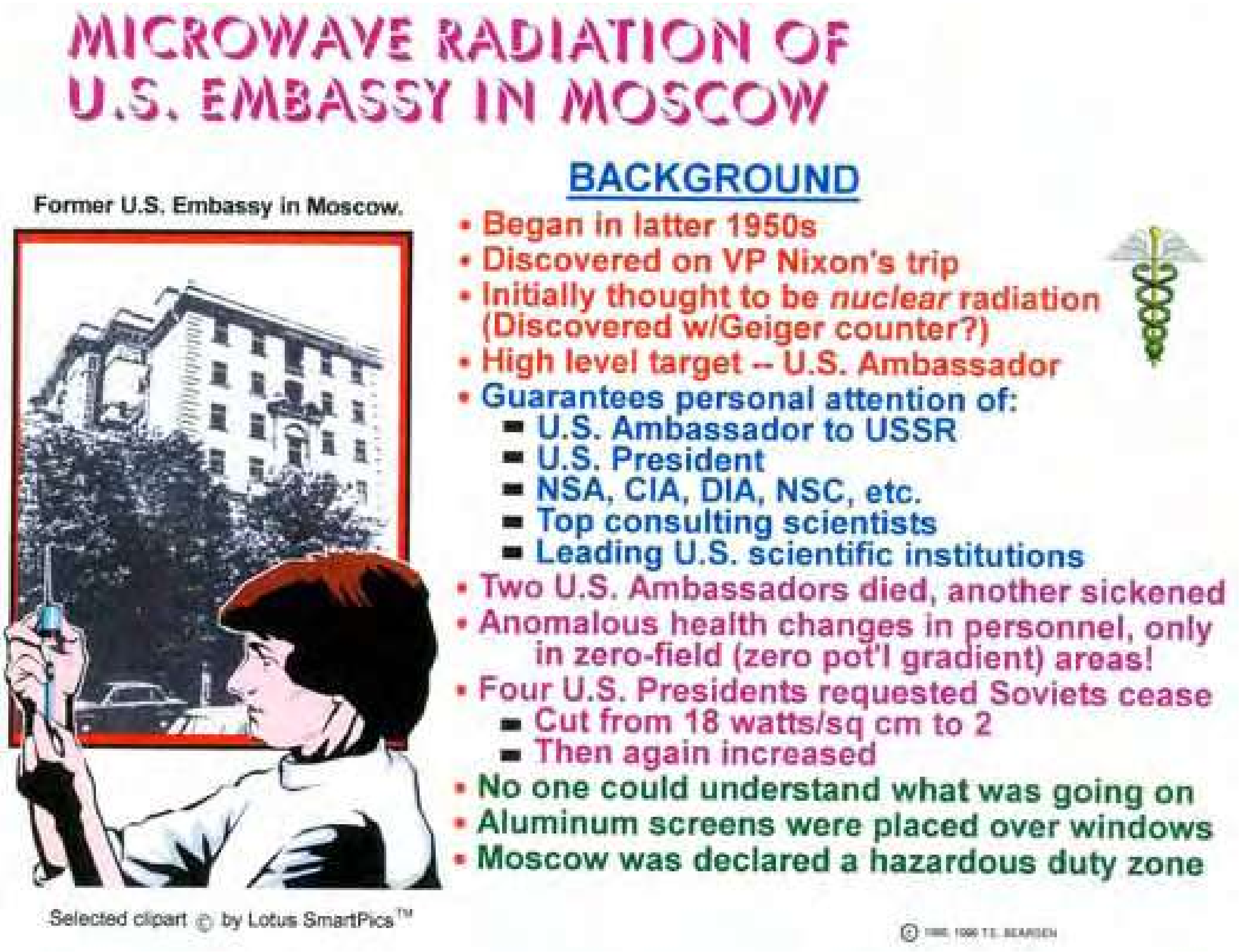

MICROWAVE RADIATION OF U.S. EMBASSY IN MOSCOW

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Selected clipart © by Lotus SmartPics™

© 1986, 1990 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

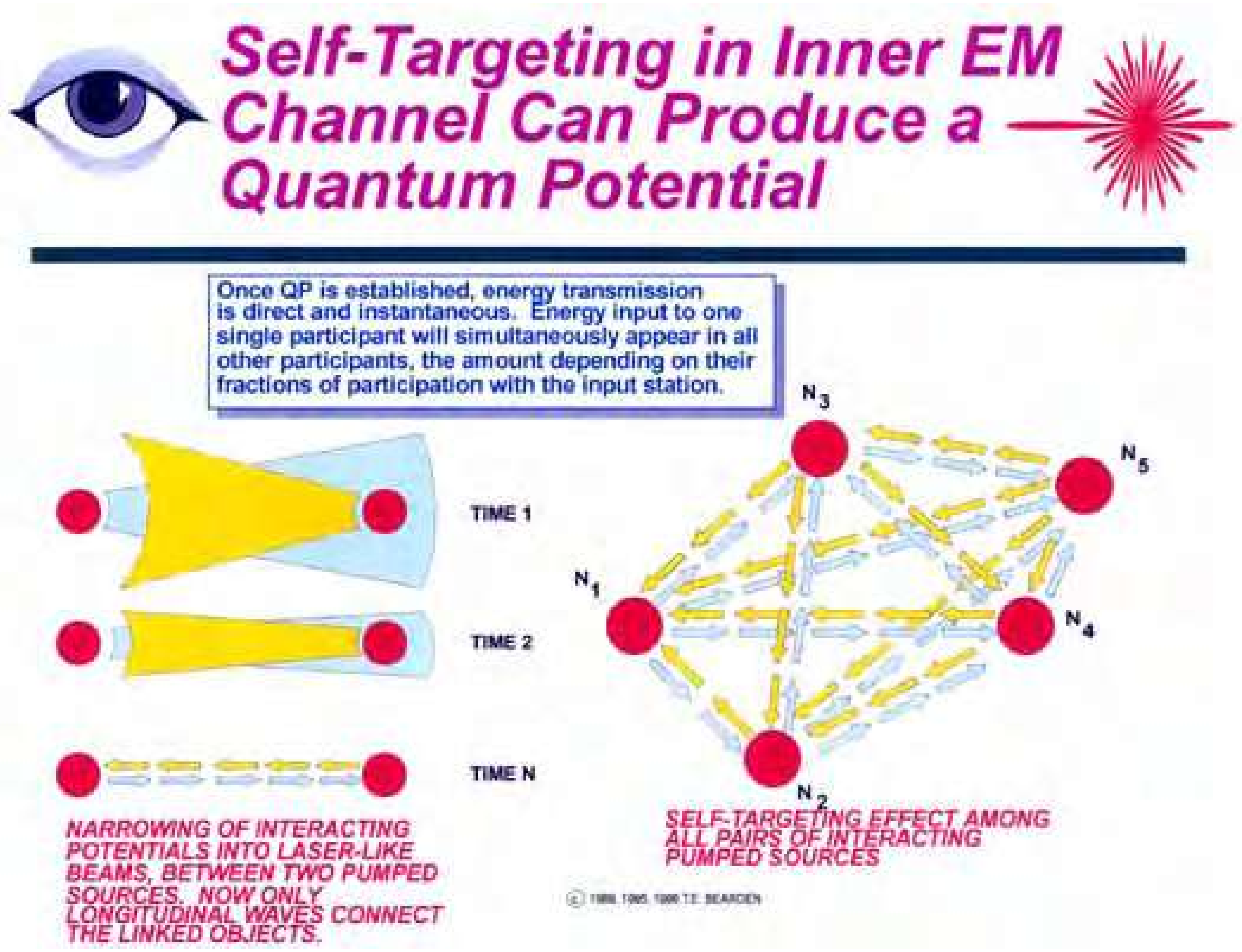

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1988, 1995, 1996 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

[Previous Slide]

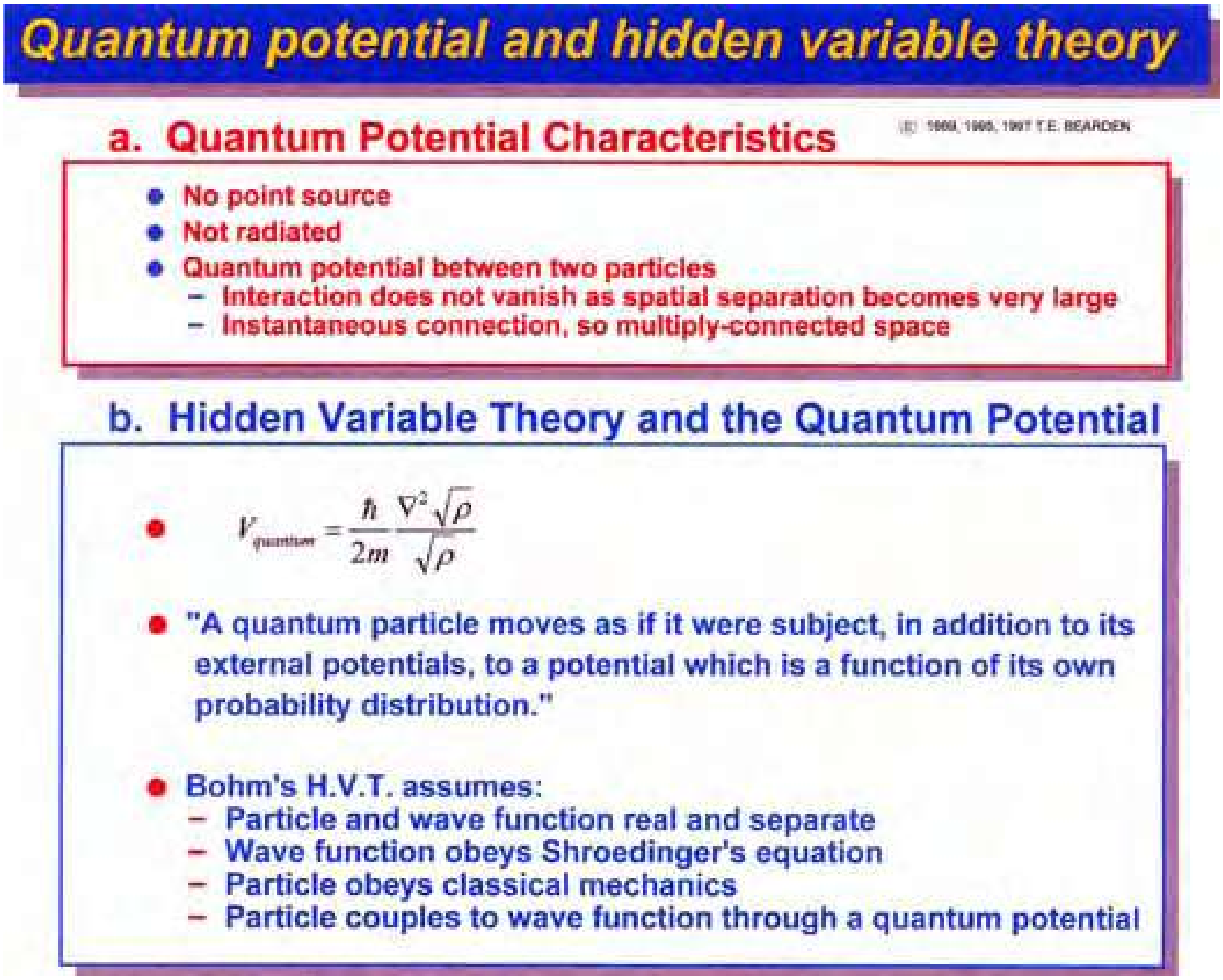

Quantum potential and hidden variable theory

a. Quantum Potential Characteristics

© 1988, 1995, 1997 T.E. BEARDEN

- No point source
- Not radiated
- Quantum potential between two particles
  - Interaction does not vanish as spatial separation becomes very large
  - Instantaneous connection, so multiply-connected space

b. Hidden Variable Theory and the Quantum Potential

- $$V_{quantum} = \frac{\hbar}{2m} \frac{\nabla^2 \sqrt{\rho}}{\sqrt{\rho}}$$
- "A quantum particle moves as if it were subject, in addition to its external potentials, to a potential which is a function of its own probability distribution."
- Bohm's H.V.T. assumes:
  - Particle and wave function real and separate
  - Wave function obeys Shroedinger's equation
  - Particle obeys classical mechanics
  - Particle couples to wave function through a quantum potential

Return to Slide Index

Next Slide

Previous Slide

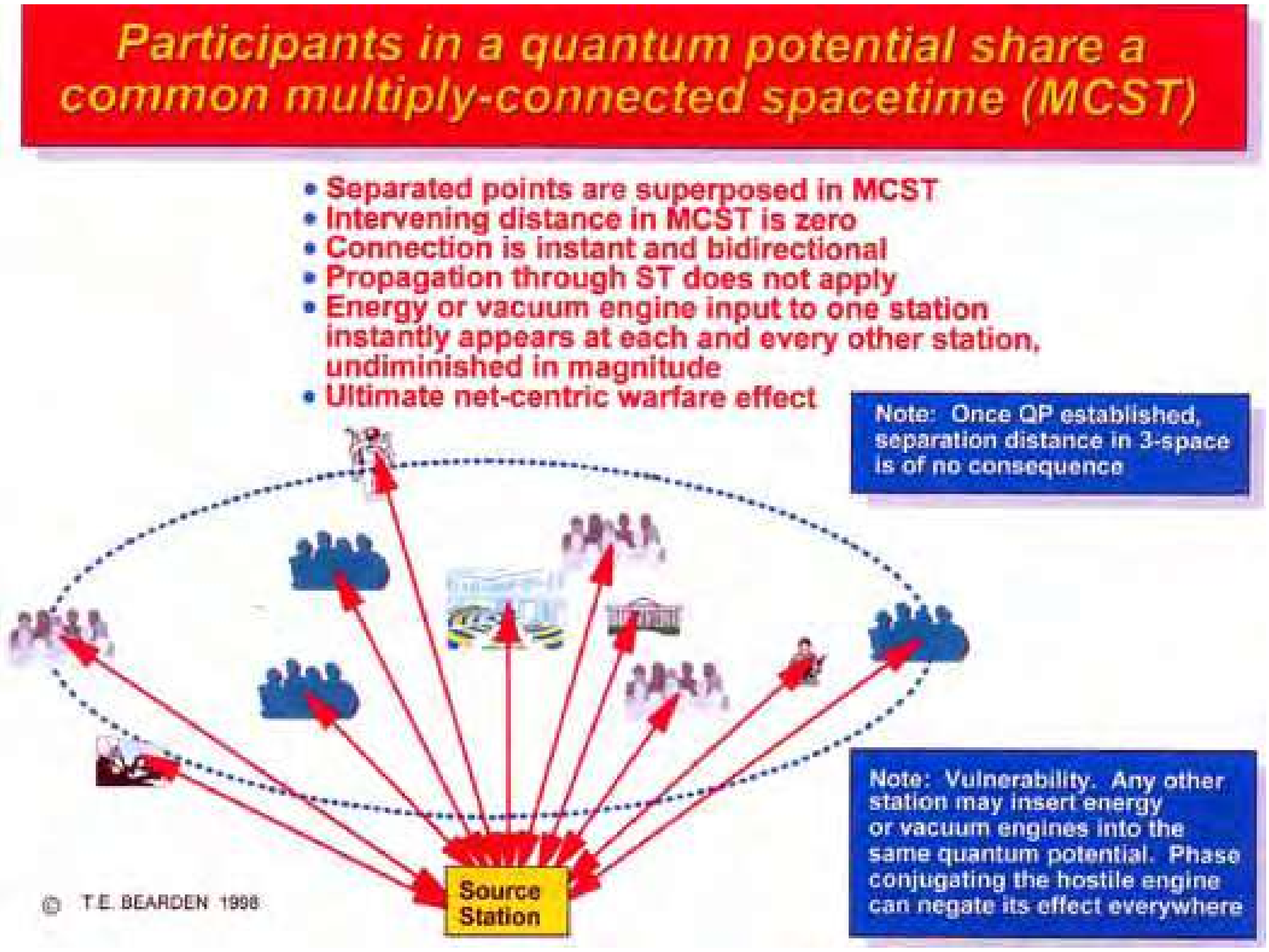

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Intervening distance in MCST is zero
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude
- Ultimate net-centric warfare effect

Note: Once QP established, separation distance in 3-space is of no consequence

Note: Vulnerability. Any other station may insert energy or vacuum engines into the same quantum potential. Phase conjugating the hostile engine can negate its effect everywhere

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

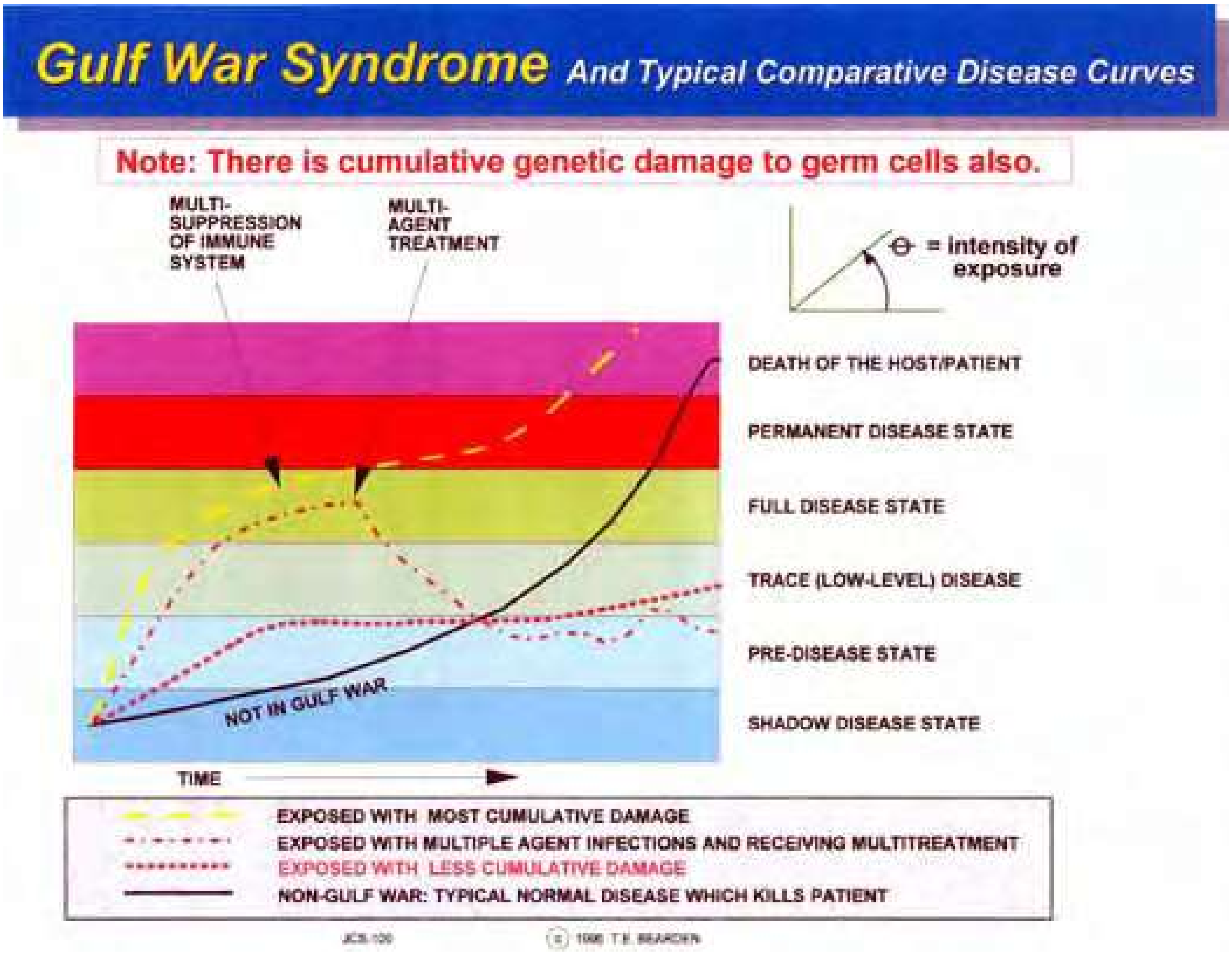

Gulf War Syndrome And Typical Comparative Disease Curves

Note: There is cumulative genetic damage to germ cells also.

- EXPOSED WITH MOST CUMULATIVE DAMAGE
- EXPOSED WITH MULTIPLE AGENT INFECTIONS AND RECEIVING MULTITREATMENT
- EXPOSED WITH LESS CUMULATIVE DAMAGE
- NON-GULF WAR: TYPICAL NORMAL DISEASE WHICH KILLS PATIENT

JCS-100

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

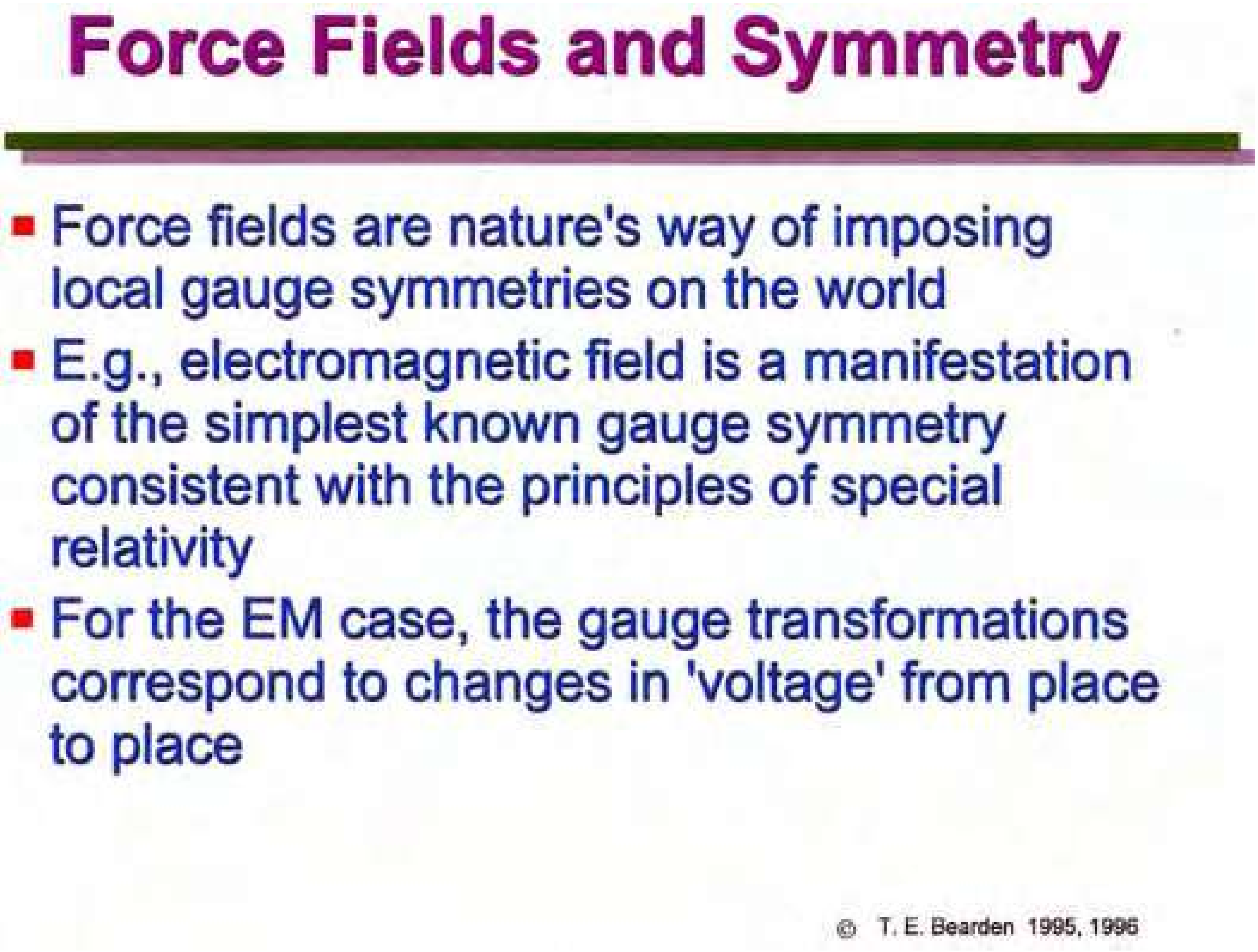

Force Fields and Symmetry

- Force fields are nature's way of imposing local gauge symmetries on the world
- E.g., electromagnetic field is a manifestation of the simplest known gauge symmetry consistent with the principles of special relativity
- For the EM case, the gauge transformations correspond to changes in 'voltage' from place to place

© T. E. Bearden 1995, 1996

Return to Slide Index

Next Slide

Previous Slide

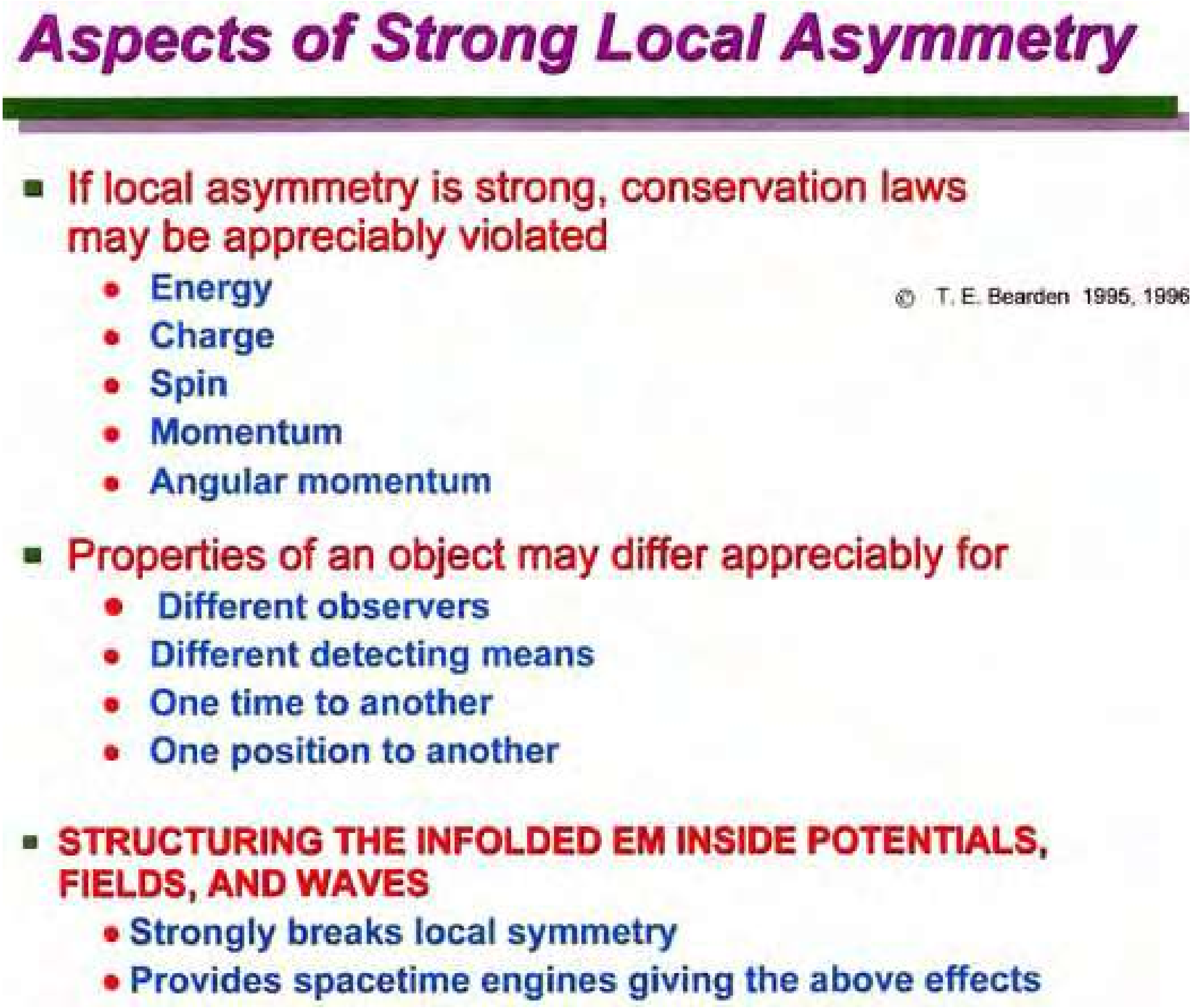

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

Return to Slide Index

Next Slide

Previous Slide

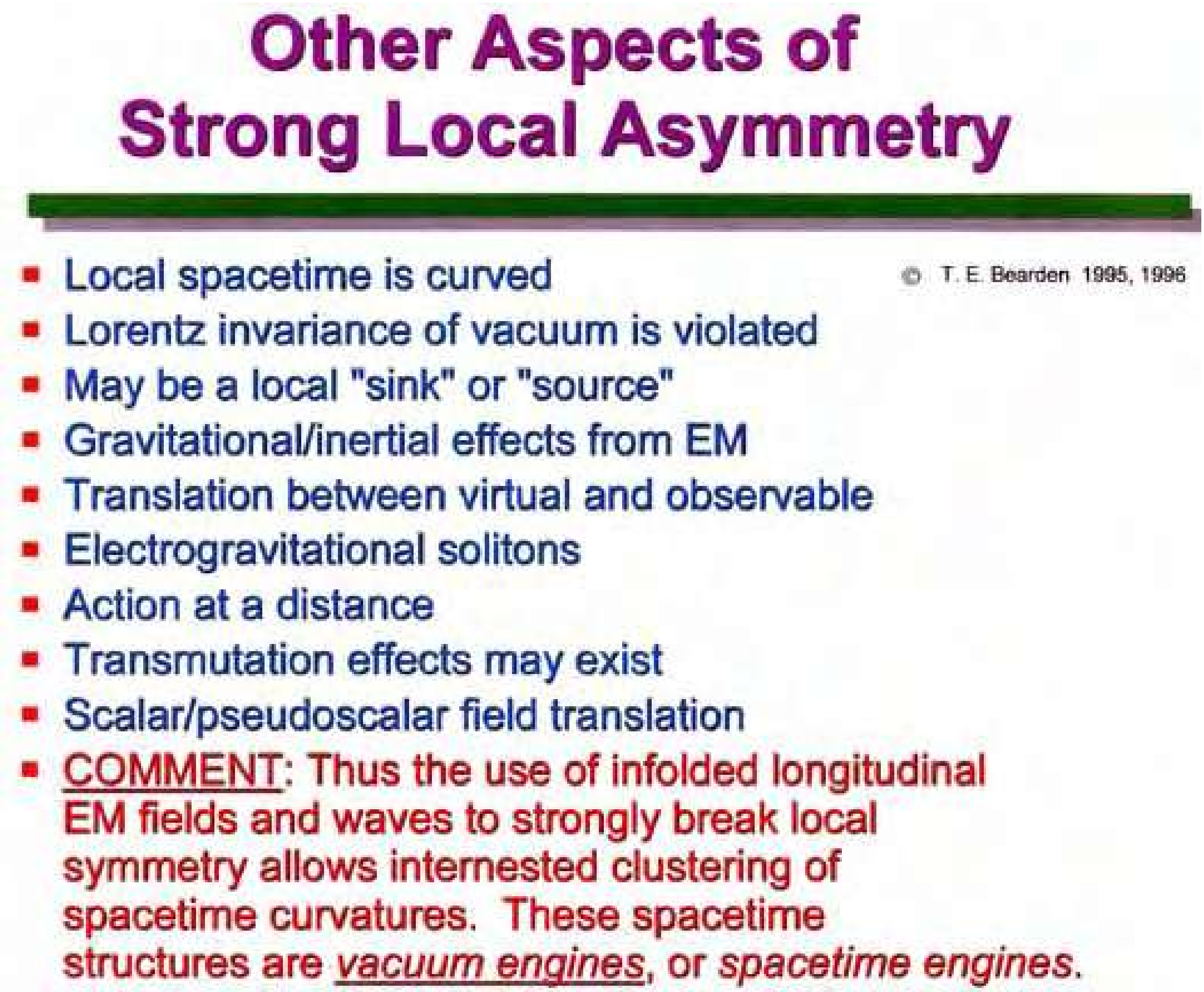

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational/inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

Return to Slide Index

Next Slide

Previous Slide

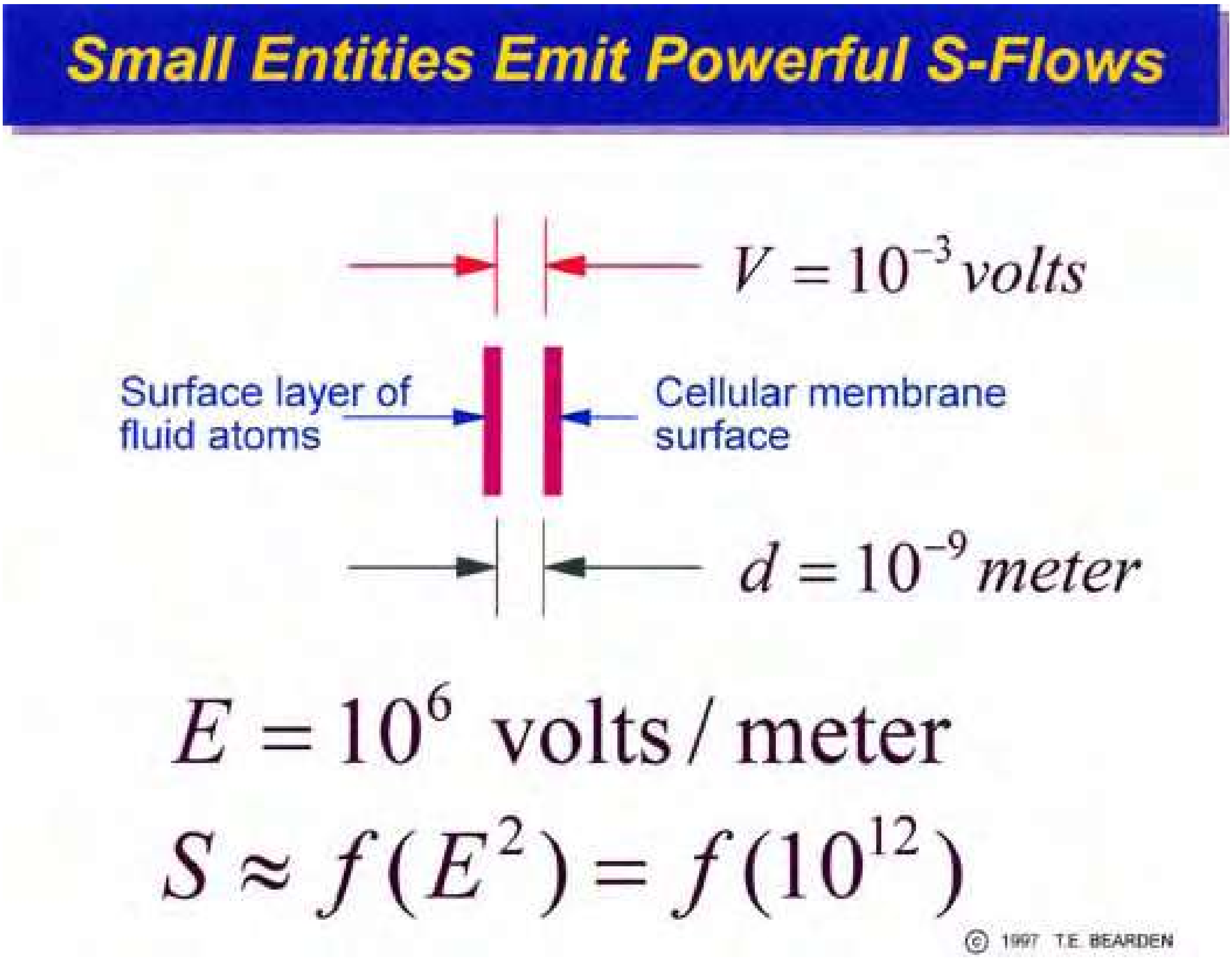

Small Entities Emit Powerful S-Flows

$V = 10^{-3}\, volts$

Surface layer of fluid atoms

Cellular membrane surface

$d = 10^{-9}\, meter$

$$E = 10^{6} \text{ volts / meter}$$

$$S \approx f(E^{2}) = f(10^{12})$$

© 1997 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

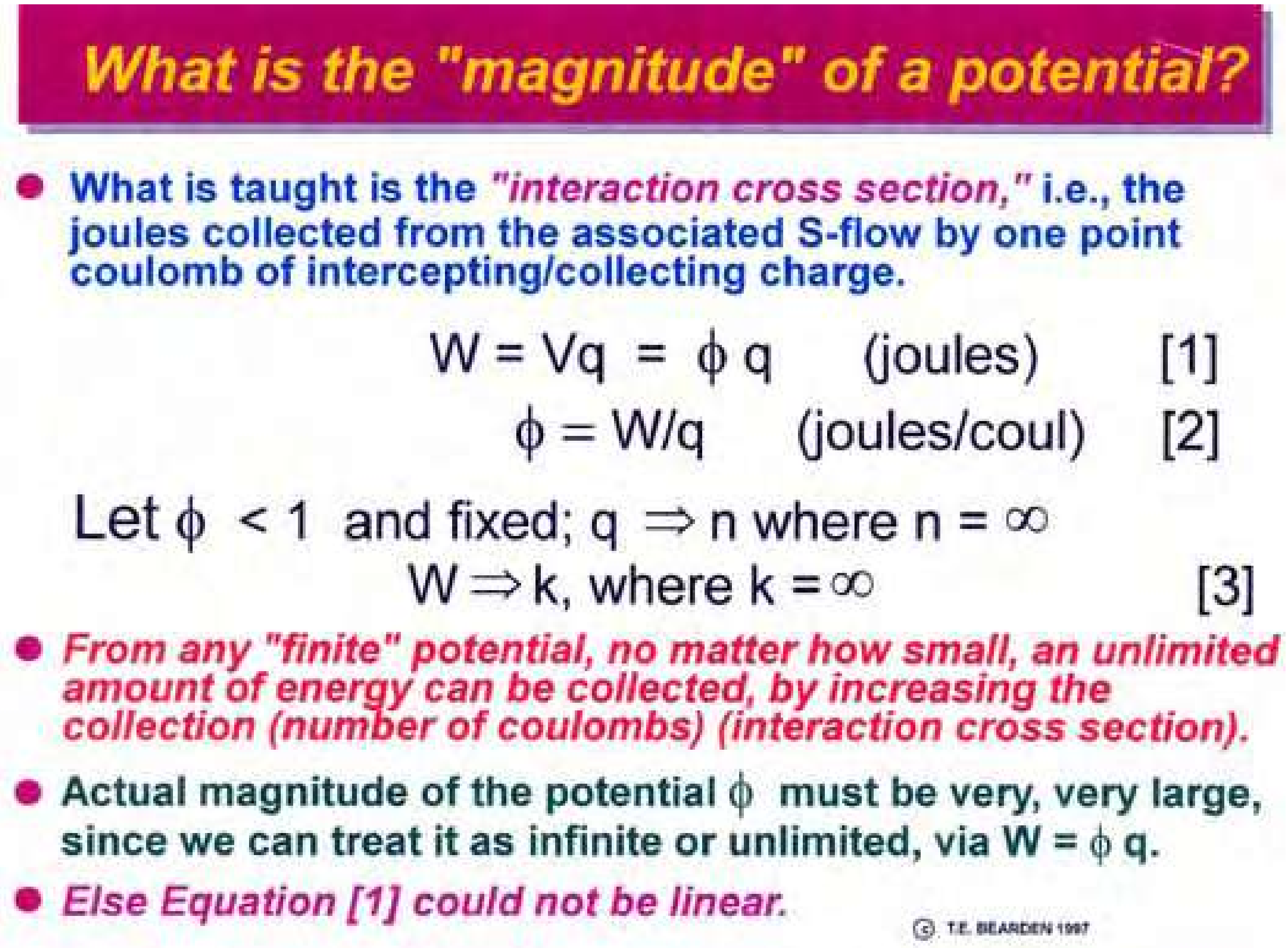

What is the "magnitude" of a potential?

- What is taught is the *"interaction cross section,"* i.e., the joules collected from the associated S-flow by one point coulomb of intercepting/collecting charge.

$$W = Vq = \phi q \quad \text{(joules)} \qquad [1]$$

$$\phi = W/q \quad \text{(joules/coul)} \qquad [2]$$

Let $\phi < 1$ and fixed; $q \Rightarrow n$ where $n = \infty$

$$W \Rightarrow k, \text{ where } k = \infty \qquad [3]$$

- *From any "finite" potential, no matter how small, an unlimited amount of energy can be collected, by increasing the collection (number of coulombs) (interaction cross section).*
- Actual magnitude of the potential ϕ must be very, very large, since we can treat it as infinite or unlimited, via $W = \phi q$.
- *Else Equation [1] could not be linear.*

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

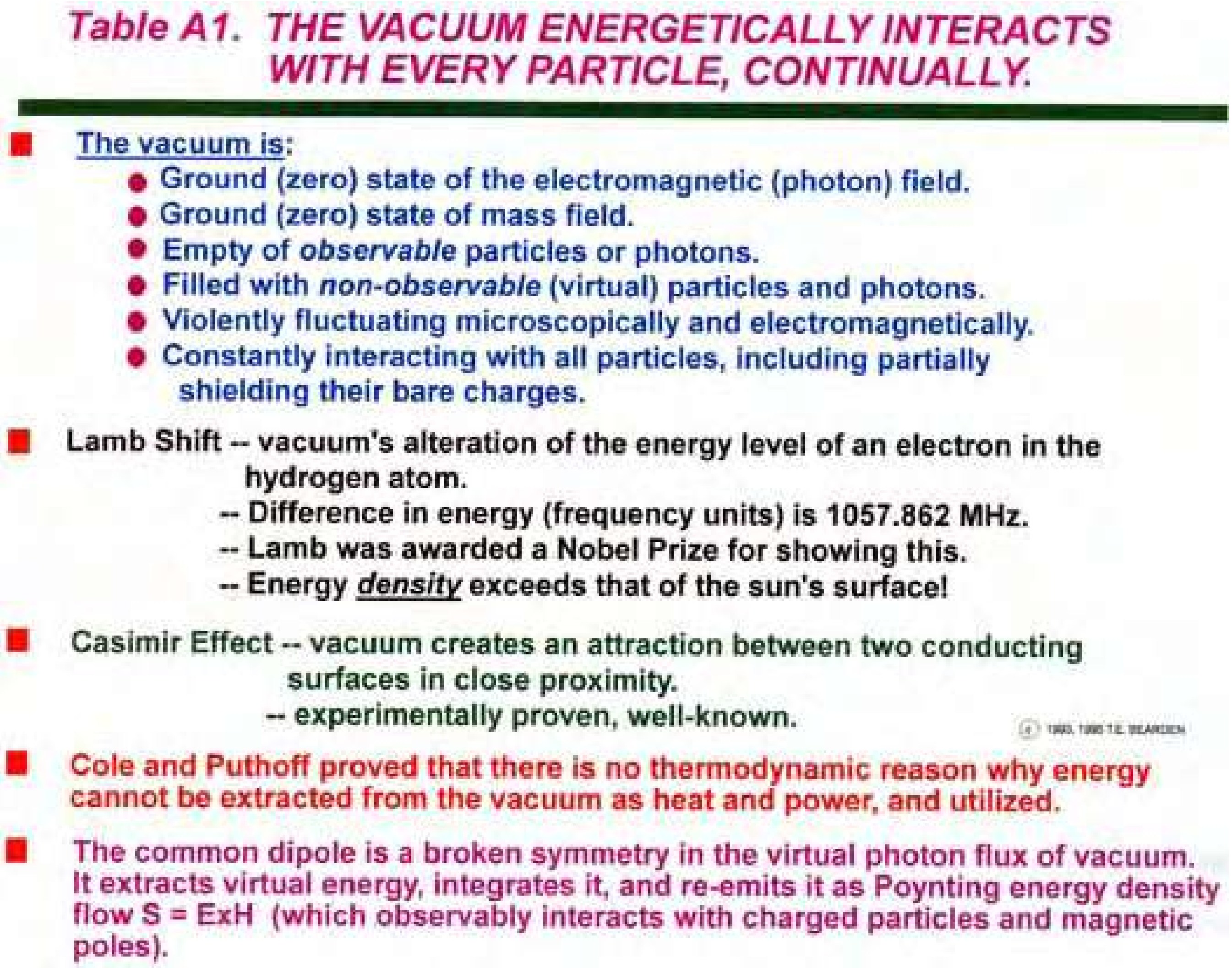

Table A1. THE VACUUM ENERGETICALLY INTERACTS WITH EVERY PARTICLE, CONTINUALLY.

- The vacuum is:
  - Ground (zero) state of the electromagnetic (photon) field.
  - Ground (zero) state of mass field.
  - Empty of *observable* particles or photons.
  - Filled with *non-observable* (virtual) particles and photons.
  - Violently fluctuating microscopically and electromagnetically.
  - Constantly interacting with all particles, including partially shielding their bare charges.
- Lamb Shift -- vacuum's alteration of the energy level of an electron in the hydrogen atom.
  - -- Difference in energy (frequency units) is 1057.862 MHz.
  - -- Lamb was awarded a Nobel Prize for showing this.
  - -- Energy *density* exceeds that of the sun's surface!
- Casimir Effect -- vacuum creates an attraction between two conducting surfaces in close proximity.
  - -- experimentally proven, well-known.
- Cole and Puthoff proved that there is no thermodynamic reason why energy cannot be extracted from the vacuum as heat and power, and utilized.
- The common dipole is a broken symmetry in the virtual photon flux of vacuum. It extracts virtual energy, integrates it, and re-emits it as Poynting energy density flow S = ExH (which observably interacts with charged particles and magnetic poles).

Return to Slide Index

Next Slide

Previous Slide

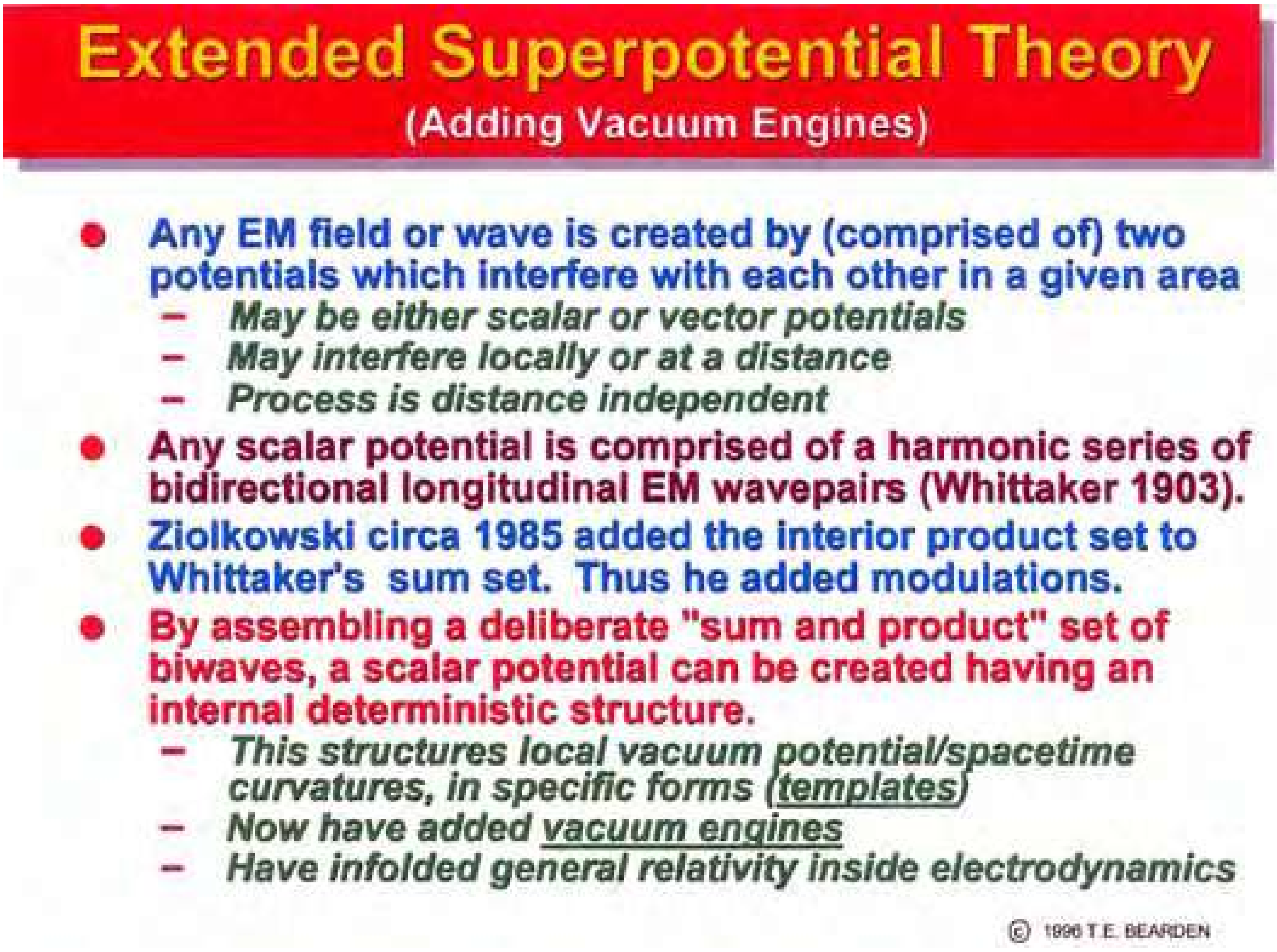

Extended Superpotential Theory

(Adding Vacuum Engines)

- Any EM field or wave is created by (comprised of) two potentials which interfere with each other in a given area
  - *May be either scalar or vector potentials*
  - *May interfere locally or at a distance*
  - *Process is distance independent*
- Any scalar potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs (Whittaker 1903).
- Ziolkowski circa 1985 added the interior product set to Whittaker's sum set. Thus he added modulations.
- By assembling a deliberate "sum and product" set of biwaves, a scalar potential can be created having an internal deterministic structure.
  - *This structures local vacuum potential/spacetime curvatures, in specific forms (templates)*
  - *Now have added vacuum engines*
  - *Have infolded general relativity inside electrodynamics*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

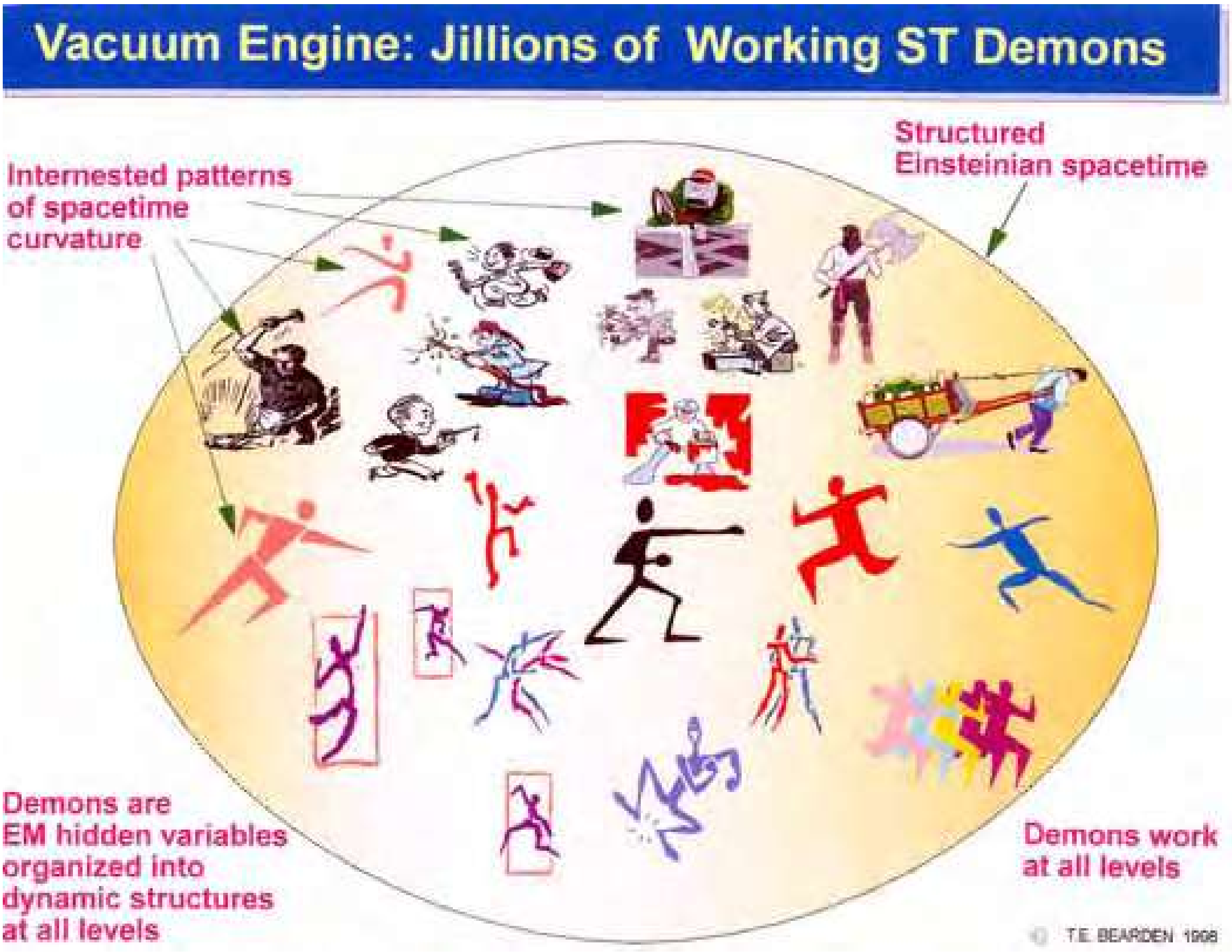

Vacuum Engine: Jillions of Working ST Demons

Internested patterns of spacetime curvature

Structured Einsteinian spacetime

Demons are EM hidden variables organized into dynamic structures at all levels

Demons work at all levels

© T.E. BEARDEN 1998

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

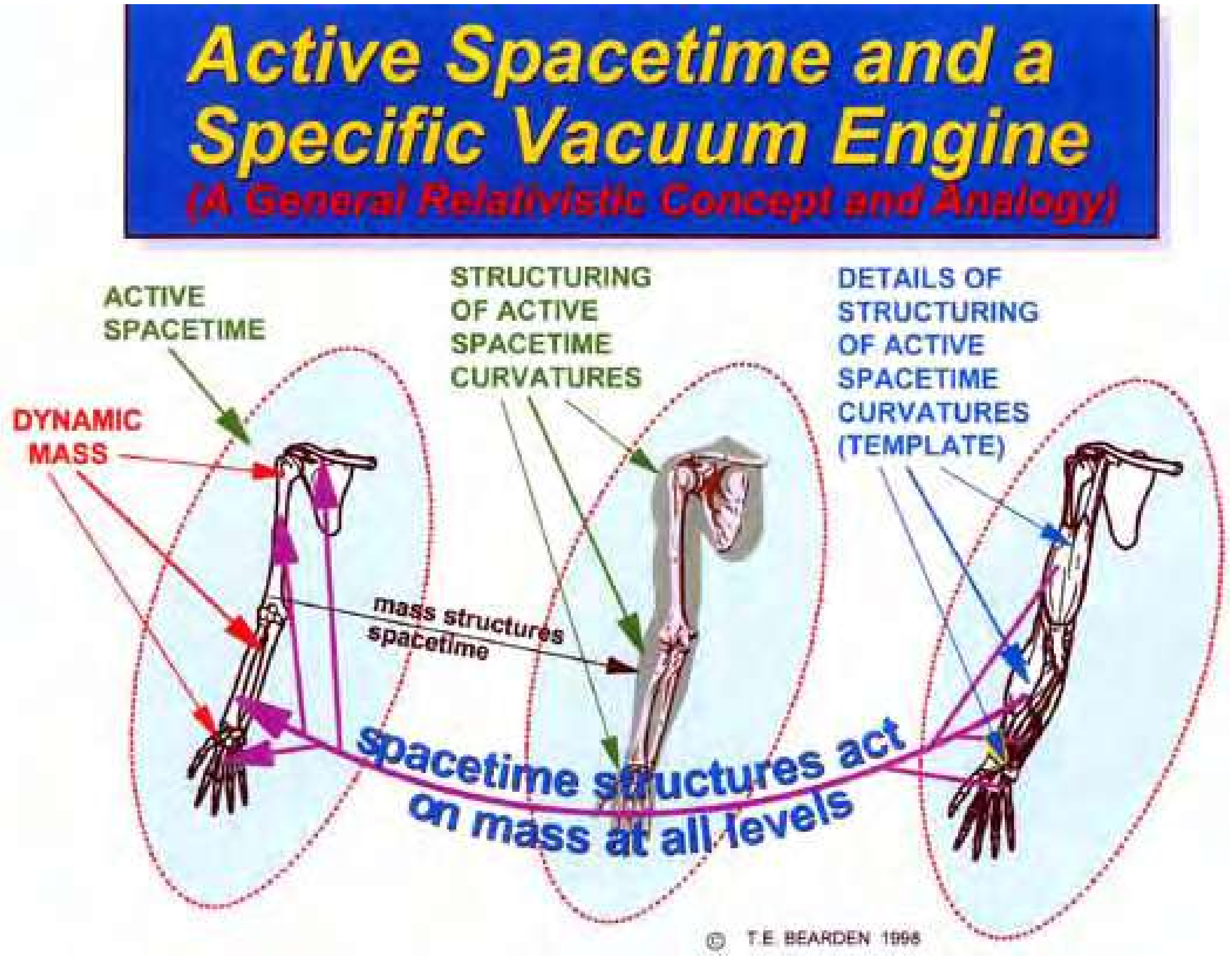

Active Spacetime and a Specific Vacuum Engine

(A General Relativistic Concept and Analogy)

ACTIVE SPACETIME

DYNAMIC MASS

STRUCTURING OF ACTIVE SPACETIME CURVATURES

DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)

mass structures spacetime

spacetime structures act on mass at all levels

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

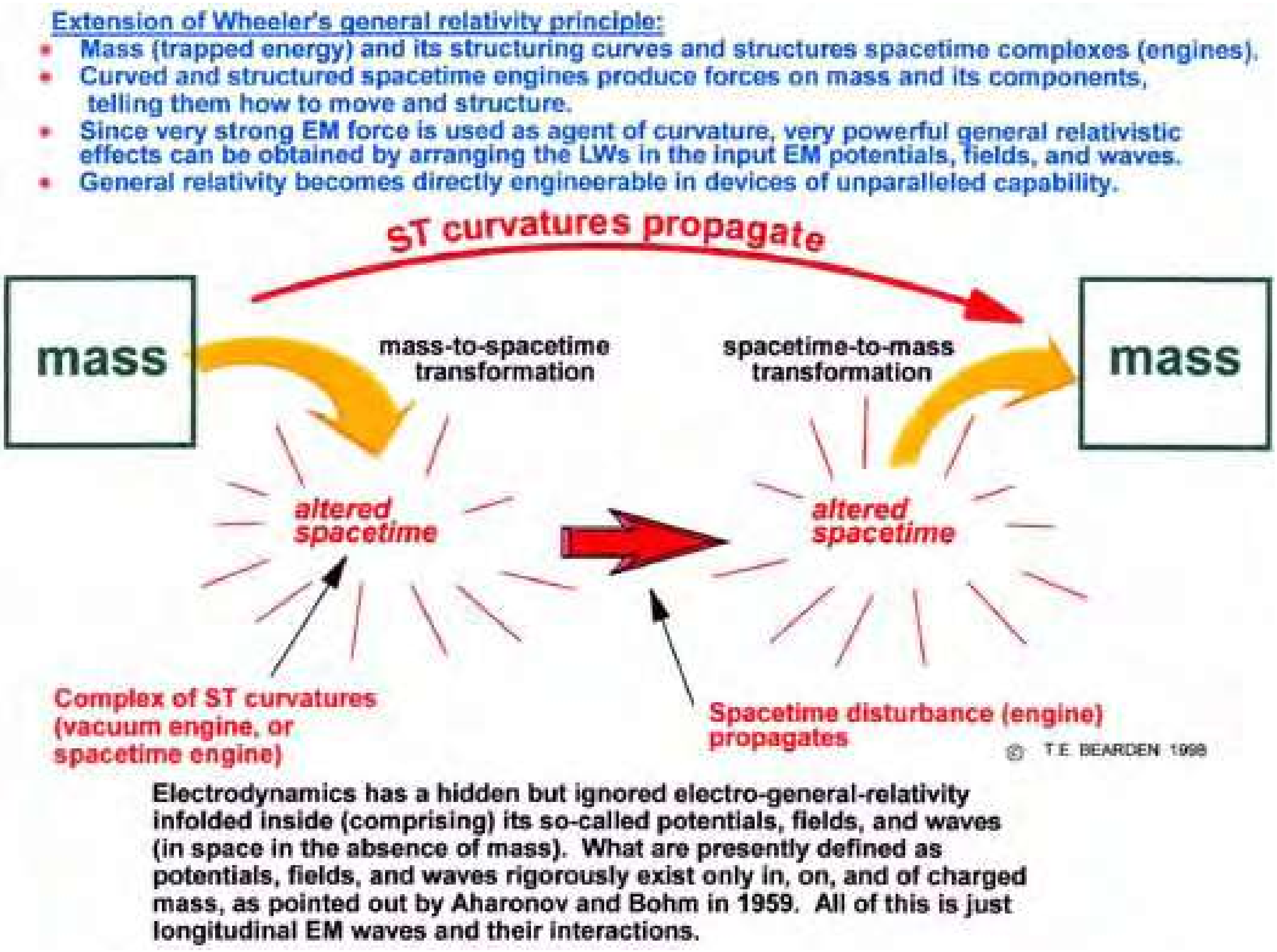

Extension of Wheeler's general relativity principle:

- Mass (trapped energy) and its structuring curves and structures spacetime complexes (engines).
- Curved and structured spacetime engines produce forces on mass and its components, telling them how to move and structure.
- Since very strong EM force is used as agent of curvature, very powerful general relativistic effects can be obtained by arranging the LWs in the input EM potentials, fields, and waves.
- General relativity becomes directly engineerable in devices of unparalleled capability.

ST curvatures propagate
mass
mass-to-spacetime transformation
spacetime-to-mass transformation
mass
altered spacetime
altered spacetime
Complex of ST curvatures (vacuum engine, or spacetime engine)
Spacetime disturbance (engine) propagates

© T.E. BEARDEN 1998

Electrodynamics has a hidden but ignored electro-general-relativity infolded inside (comprising) its so-called potentials, fields, and waves (in space in the absence of mass). What are presently defined as potentials, fields, and waves rigorously exist only in, on, and of charged mass, as pointed out by Aharonov and Bohm in 1959. All of this is just longitudinal EM waves and their interactions.

Return to Slide Index

Next Slide

Previous Slide

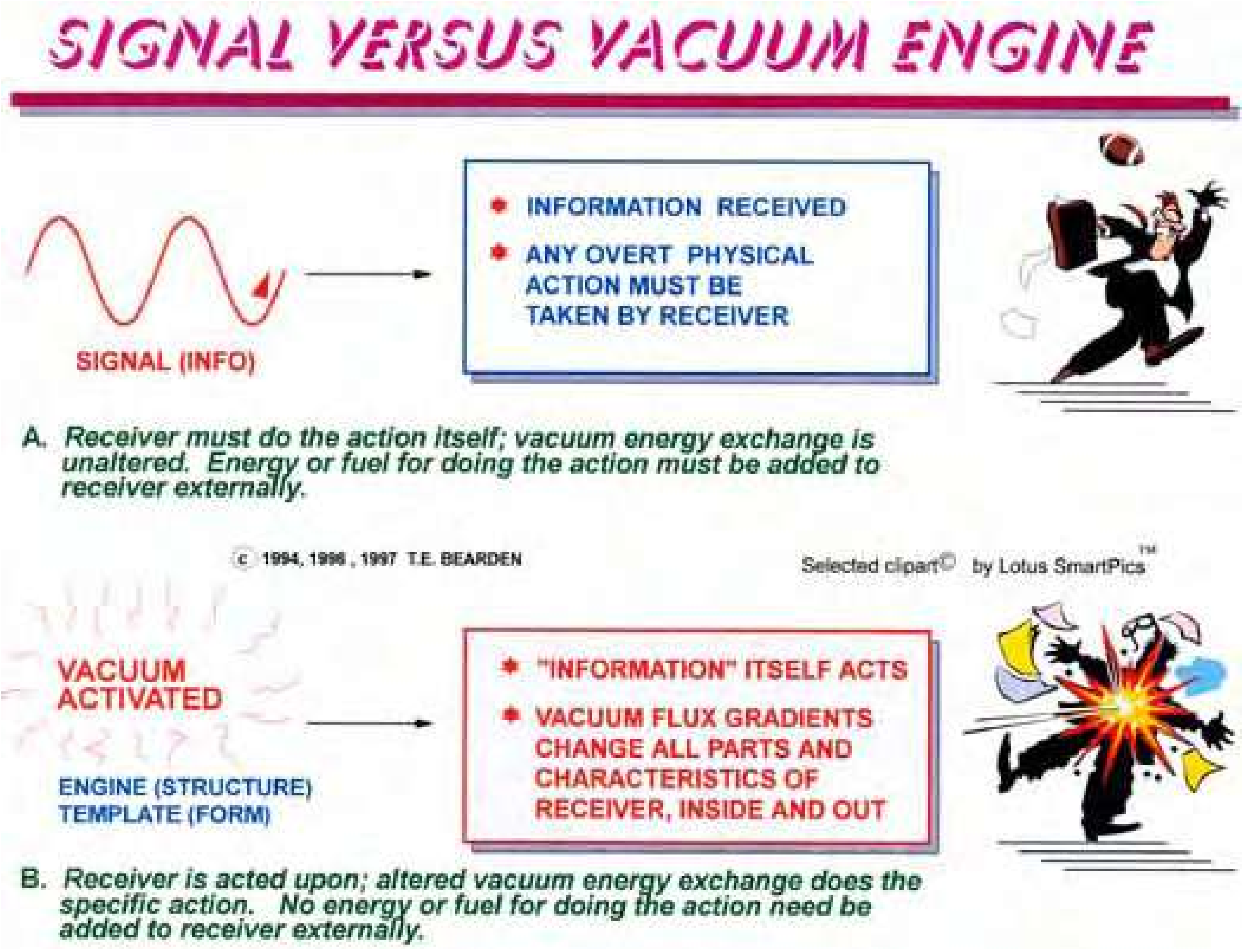

SIGNAL VERSUS VACUUM ENGINE

SIGNAL (INFO)

- INFORMATION RECEIVED
- ANY OVERT PHYSICAL ACTION MUST BE TAKEN BY RECEIVER

A. *Receiver must do the action itself; vacuum energy exchange is unaltered. Energy or fuel for doing the action must be added to receiver externally.*

© 1994, 1996, 1997 T.E. BEARDEN

Selected clipart© by Lotus SmartPics™

VACUUM ACTIVATED

ENGINE (STRUCTURE)
TEMPLATE (FORM)

- "INFORMATION" ITSELF ACTS
- VACUUM FLUX GRADIENTS CHANGE ALL PARTS AND CHARACTERISTICS OF RECEIVER, INSIDE AND OUT

B. *Receiver is acted upon; altered vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to receiver externally.*

Return to Slide Index

Next Slide

Previous Slide

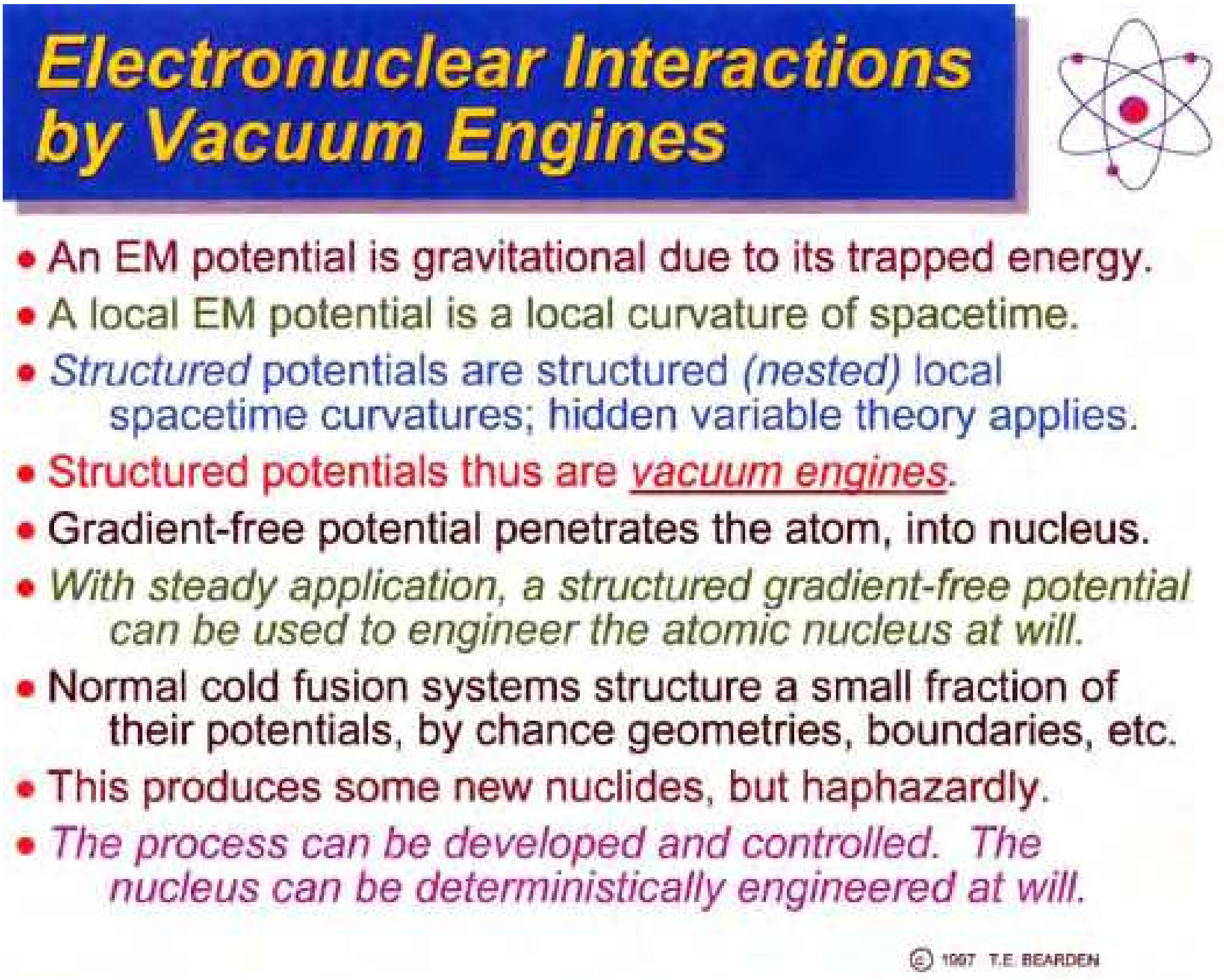

Electronuclear Interactions by Vacuum Engines

- An EM potential is gravitational due to its trapped energy.
- A local EM potential is a local curvature of spacetime.
- *Structured* potentials are structured *(nested)* local spacetime curvatures; hidden variable theory applies.
- Structured potentials thus are *vacuum engines*.
- Gradient-free potential penetrates the atom, into nucleus.
- *With steady application, a structured gradient-free potential can be used to engineer the atomic nucleus at will.*
- Normal cold fusion systems structure a small fraction of their potentials, by chance geometries, boundaries, etc.
- This produces some new nuclides, but haphazardly.
- *The process can be developed and controlled. The nucleus can be deterministically engineered at will.*

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

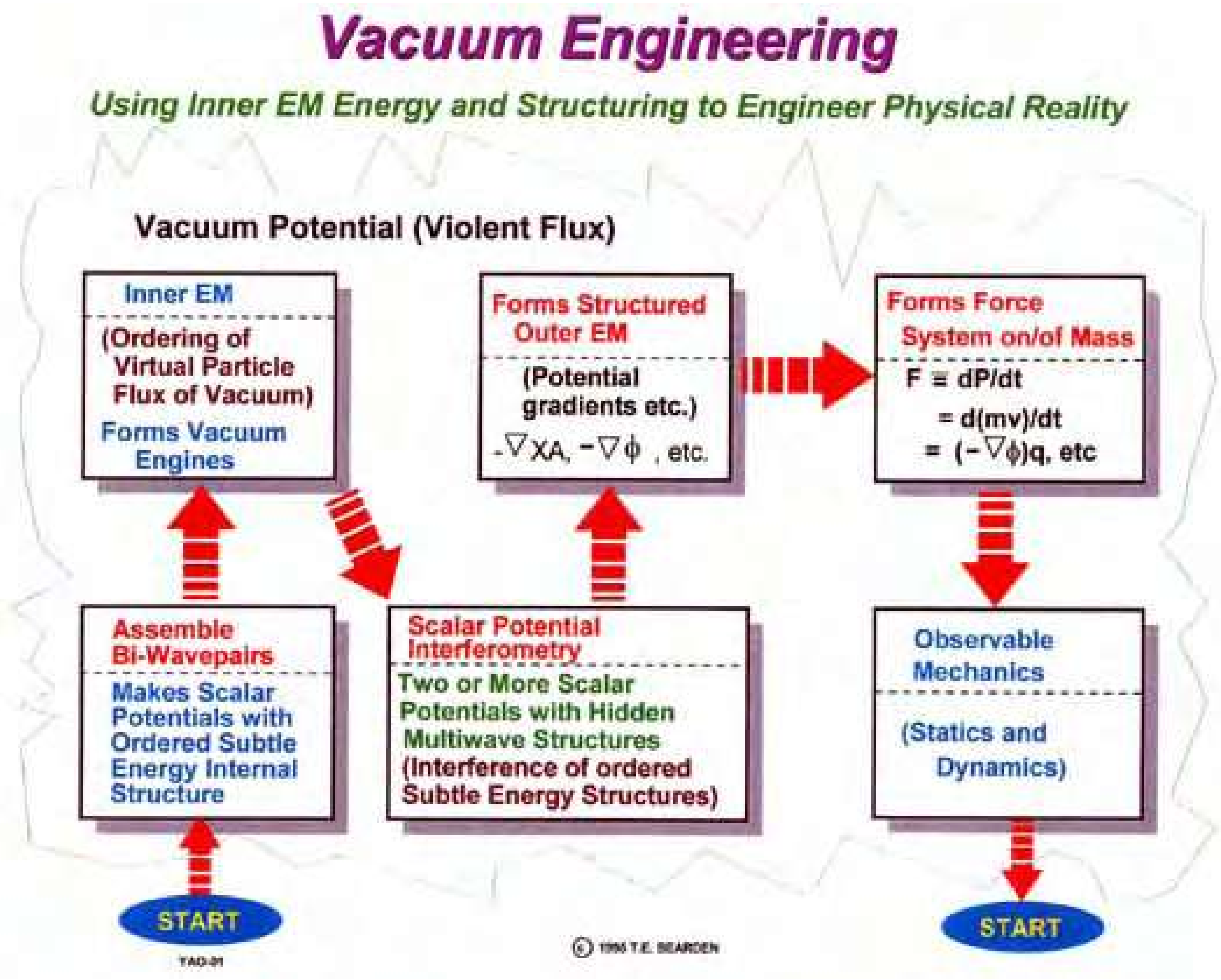

Vacuum Engineering

Using Inner EM Energy and Structuring to Engineer Physical Reality

Vacuum Potential (Violent Flux)

Inner EM

(Ordering of Virtual Particle Flux of Vacuum)

Forms Vacuum Engines

Forms Structured Outer EM

(Potential gradients etc.)

$-\nabla \times A, -\nabla \phi$, etc.

Forms Force System on/of Mass

$F \equiv dP/dt$

$= d(mv)/dt$

$\equiv (-\nabla \phi)q$, etc

Assemble Bi-Wavepairs

Makes Scalar Potentials with Ordered Subtle Energy Internal Structure

Scalar Potential Interferometry

Two or More Scalar Potentials with Hidden Multiwave Structures (Interference of ordered Subtle Energy Structures)

Observable Mechanics

(Statics and Dynamics)

START

YAO-01

© 1996 T.E. BEARDEN

START

Return to Slide Index

Next Slide

Previous Slide

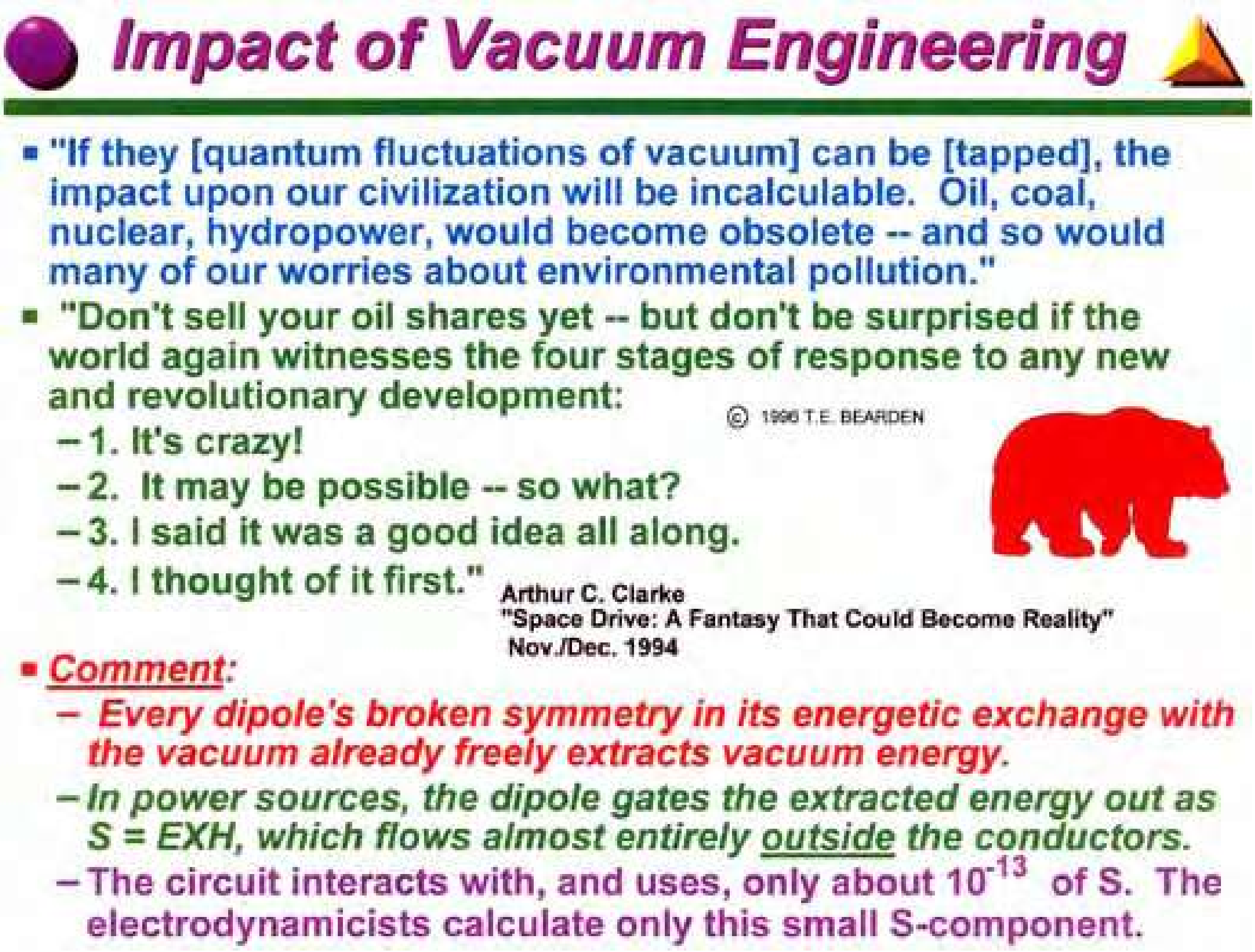

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete -- and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet -- but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

© 1996 T.E. BEARDEN

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

- *Comment*:
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

Return to Slide Index

Next Slide

Previous Slide

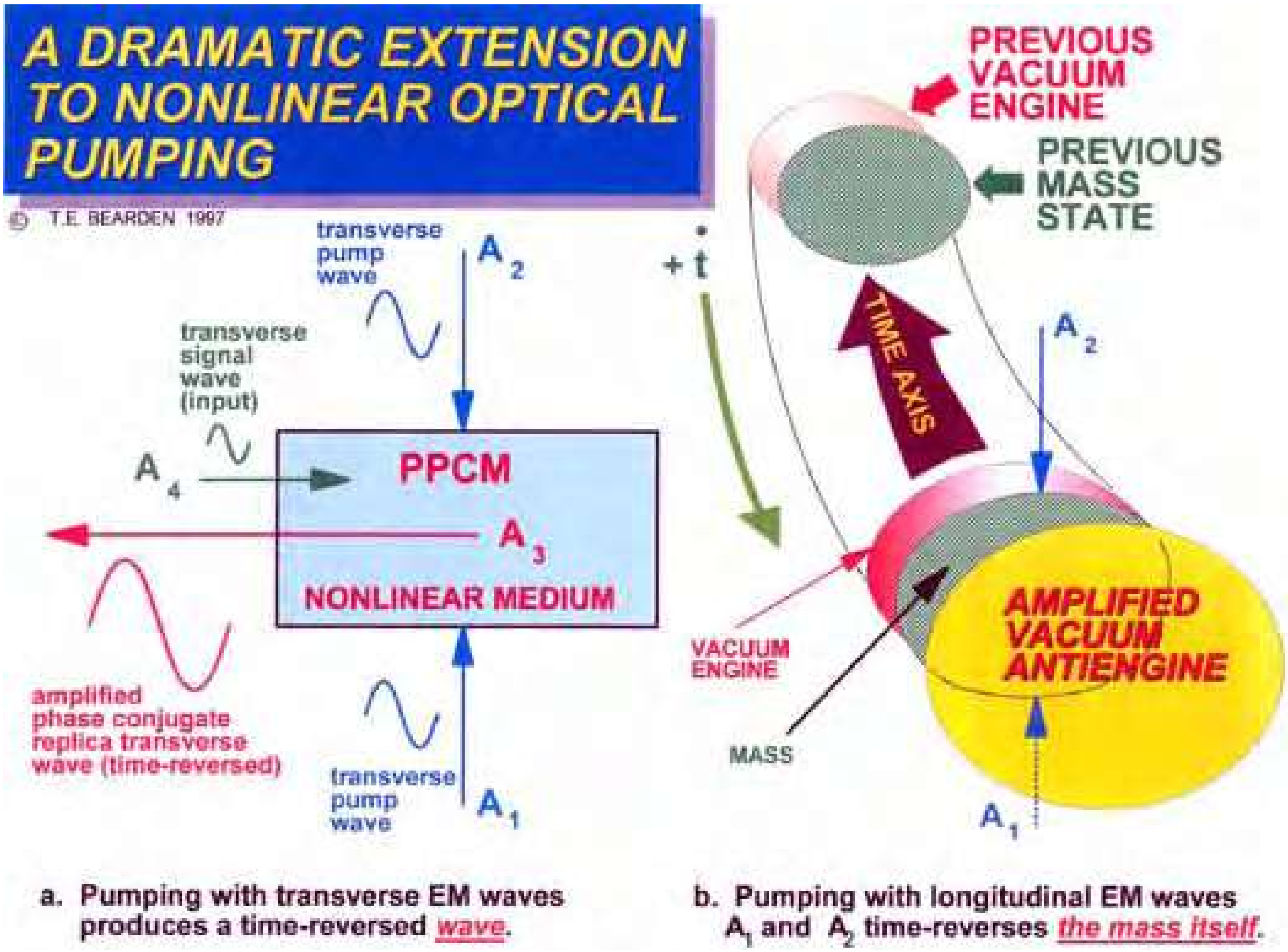

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

transverse pump wave
A_2
transverse signal wave (input)
A_4
PPCM
A_3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A_1
+ t
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
TIME AXIS
A_2
VACUUM ENGINE
AMPLIFIED VACUUM ANTIENGINE
MASS
A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

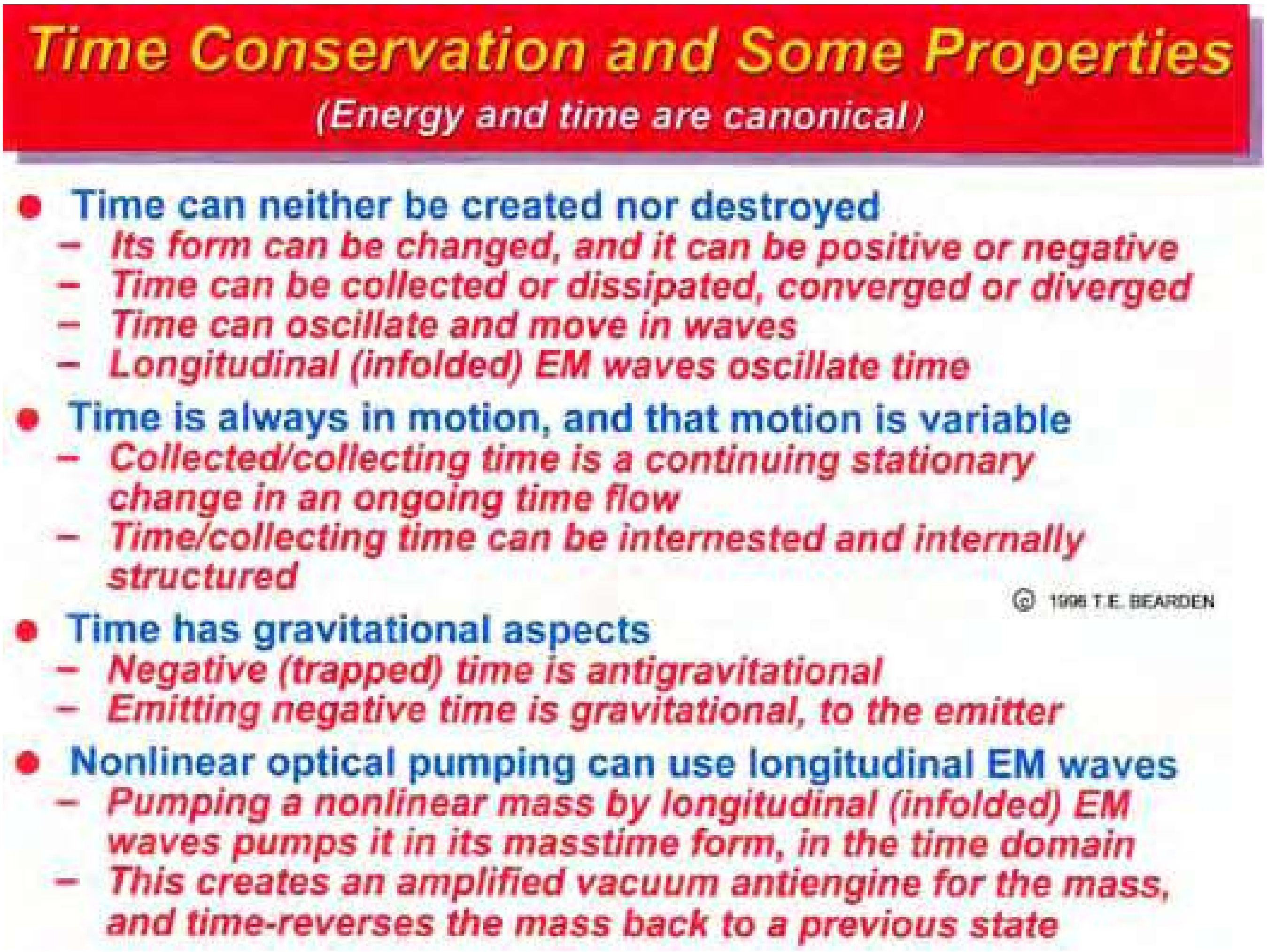

Time Conservation and Some Properties

(Energy and time are canonical)

- Time can neither be created nor destroyed
  - *Its form can be changed, and it can be positive or negative*
  - *Time can be collected or dissipated, converged or diverged*
  - *Time can oscillate and move in waves*
  - *Longitudinal (infolded) EM waves oscillate time*
- Time is always in motion, and that motion is variable
  - *Collected/collecting time is a continuing stationary change in an ongoing time flow*
  - *Time/collecting time can be internested and internally structured*
- Time has gravitational aspects
  - *Negative (trapped) time is antigravitational*
  - *Emitting negative time is gravitational, to the emitter*
- Nonlinear optical pumping can use longitudinal EM waves
  - *Pumping a nonlinear mass by longitudinal (infolded) EM waves pumps it in its masstime form, in the time domain*
  - *This creates an amplified vacuum antiengine for the mass, and time-reverses the mass back to a previous state*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

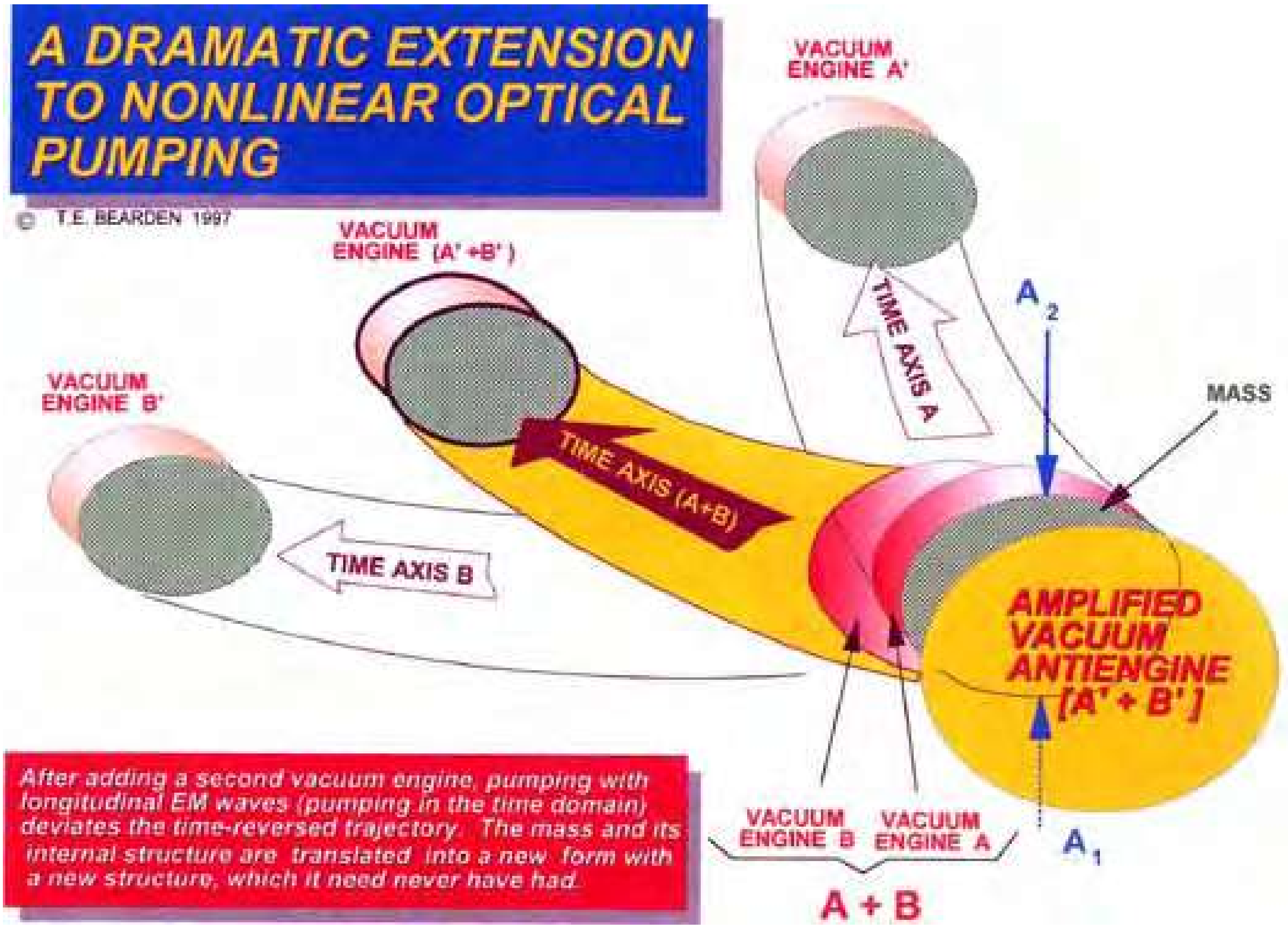

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
A_2
MASS
TIME AXIS (A+B)
TIME AXIS B
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A + B
A_1

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

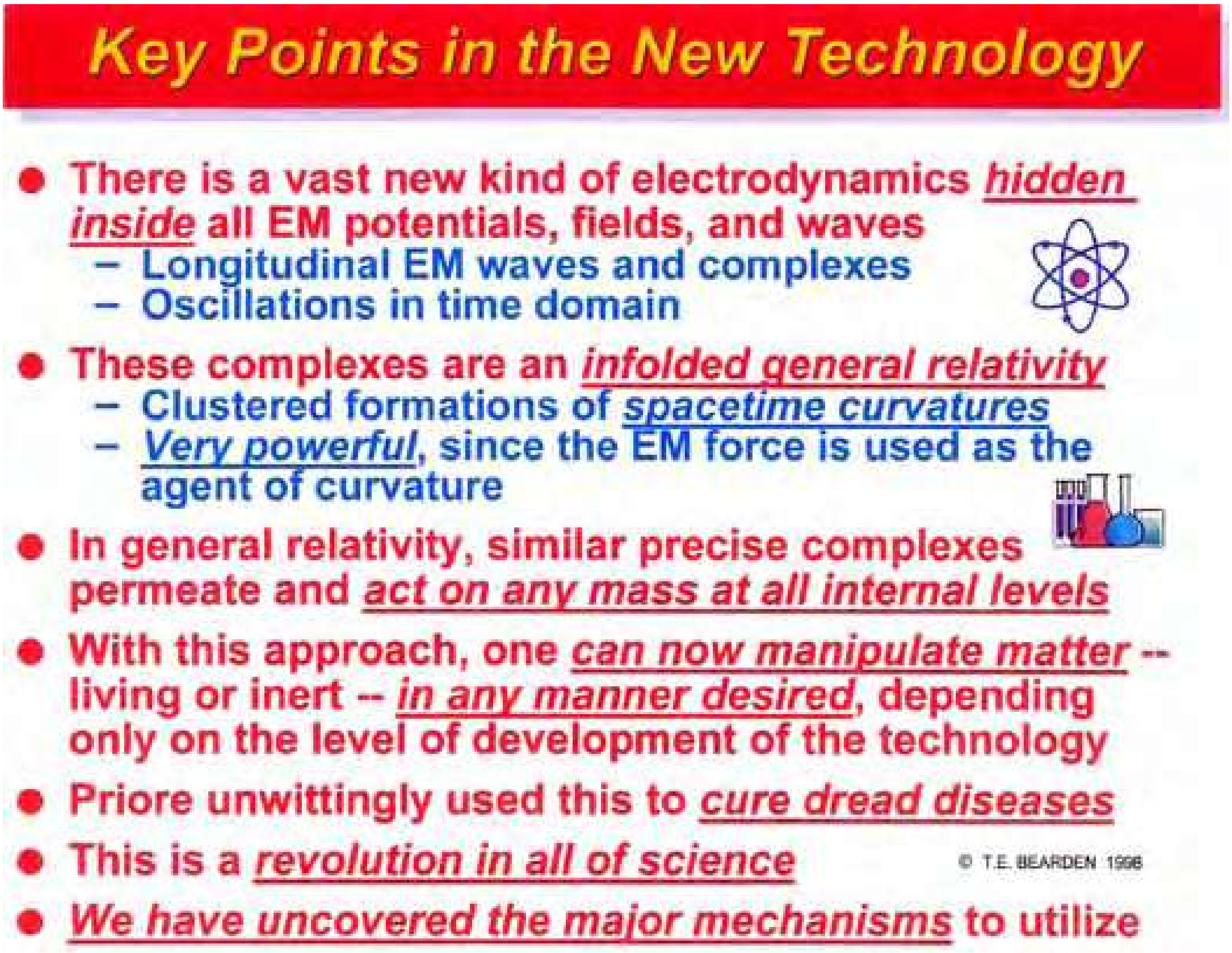

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

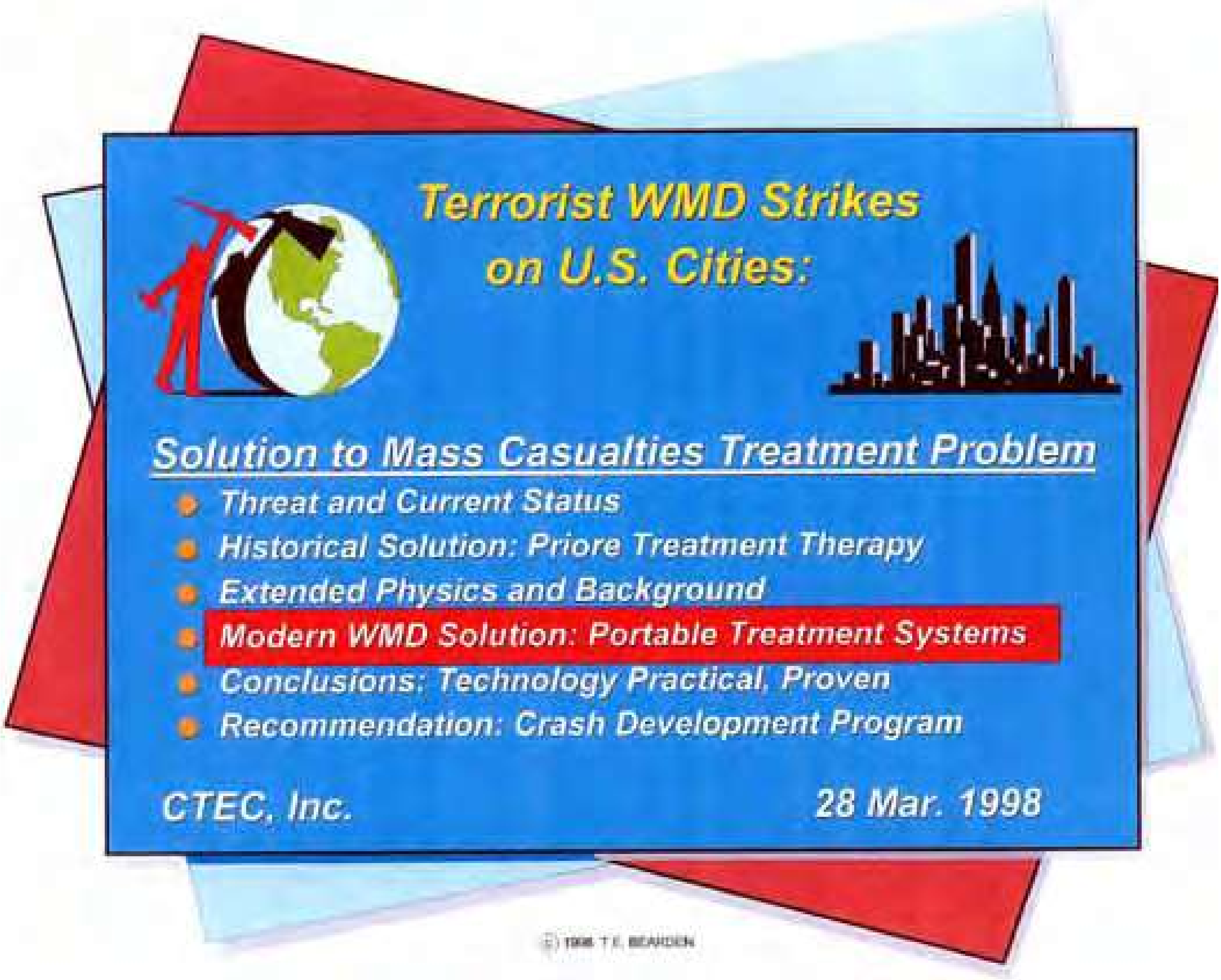

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

Portable Priore-Type Treatment System

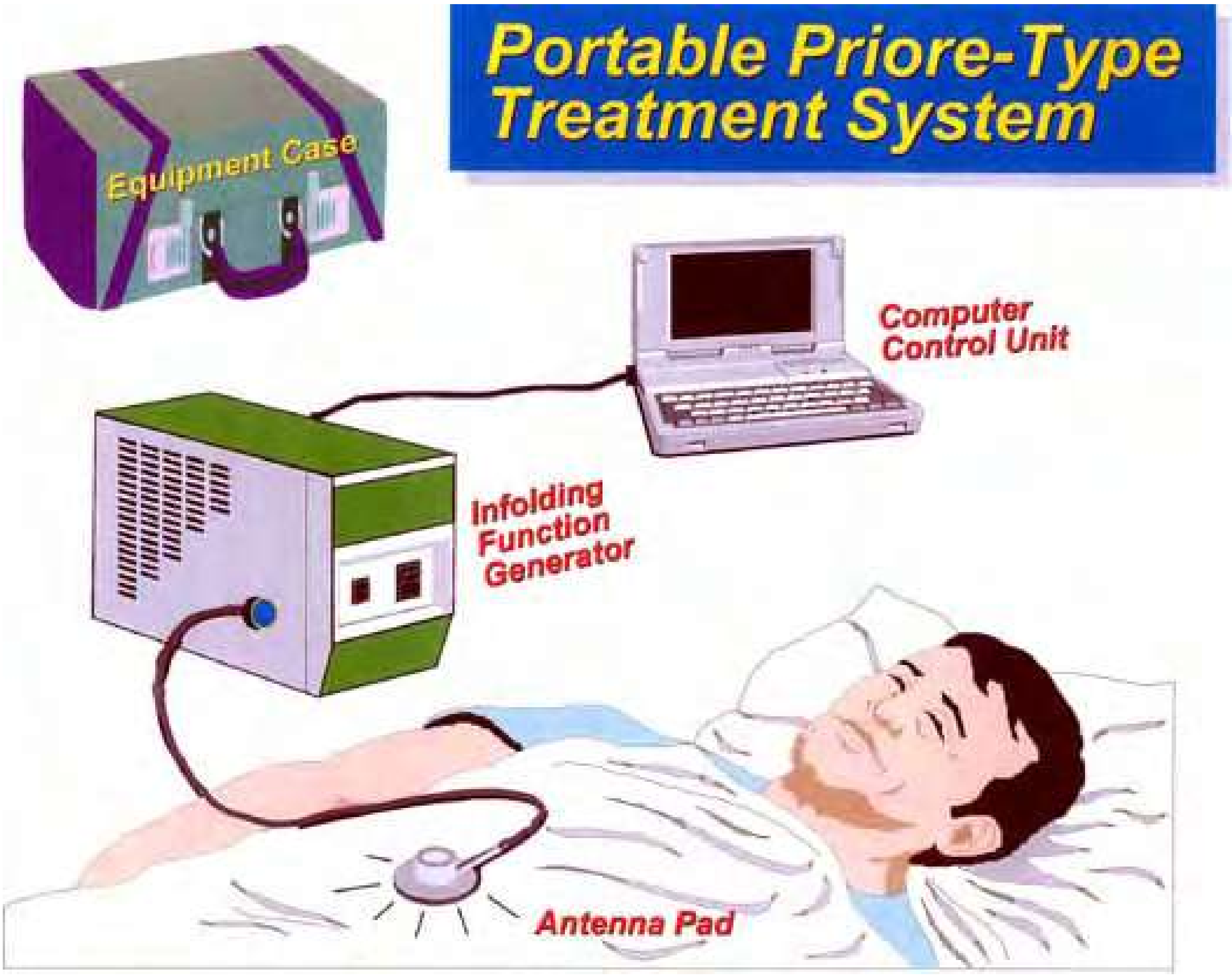
Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

Portable System Block Diagram

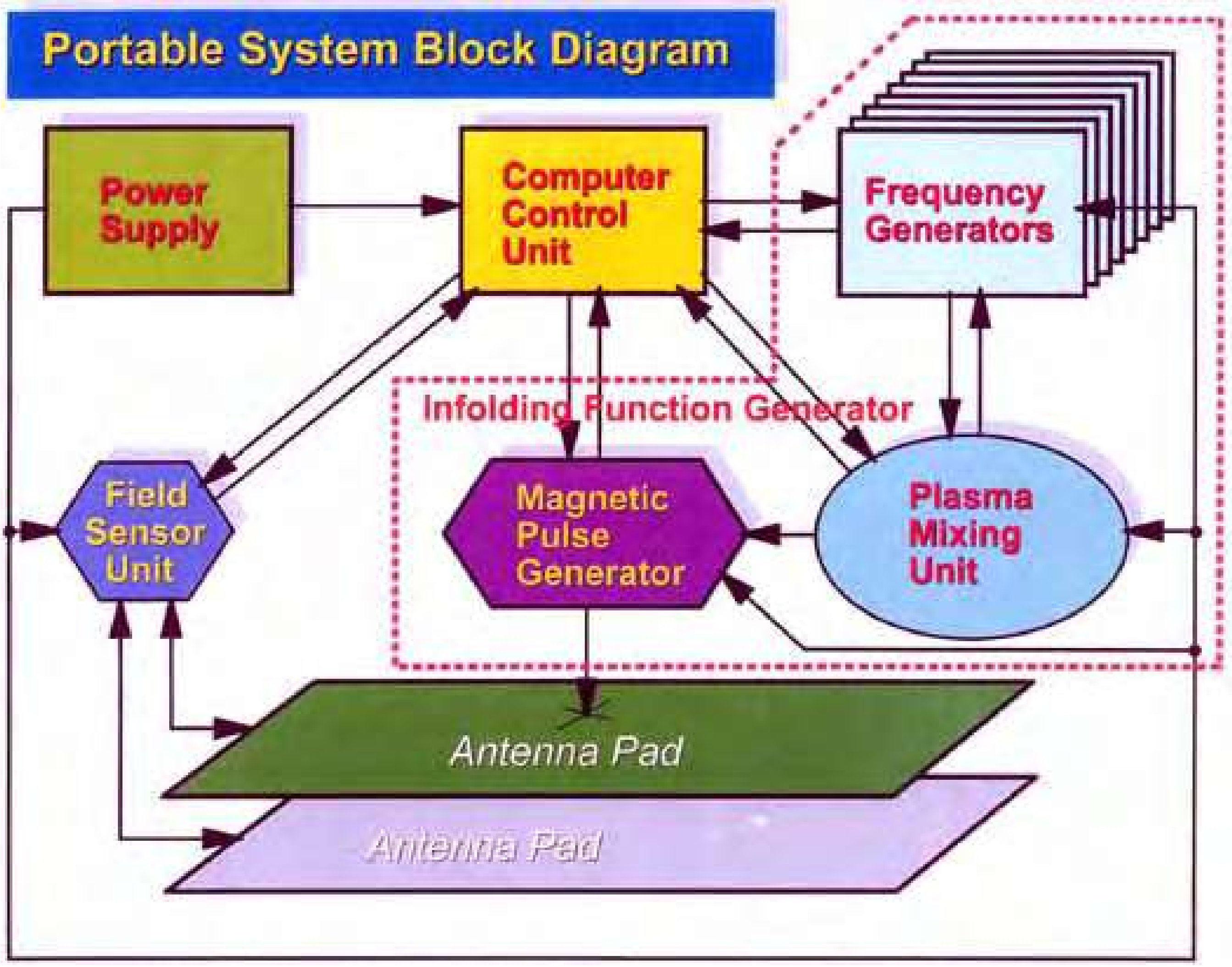
Power Supply
Computer Control Unit
Frequency Generators
Infolding Function Generator
Field Sensor Unit
Magnetic Pulse Generator
Plasma Mixing Unit
Antenna Pad
Antenna Pad

[Return to Slide Index]
[Next Slide]
[Previous Slide]

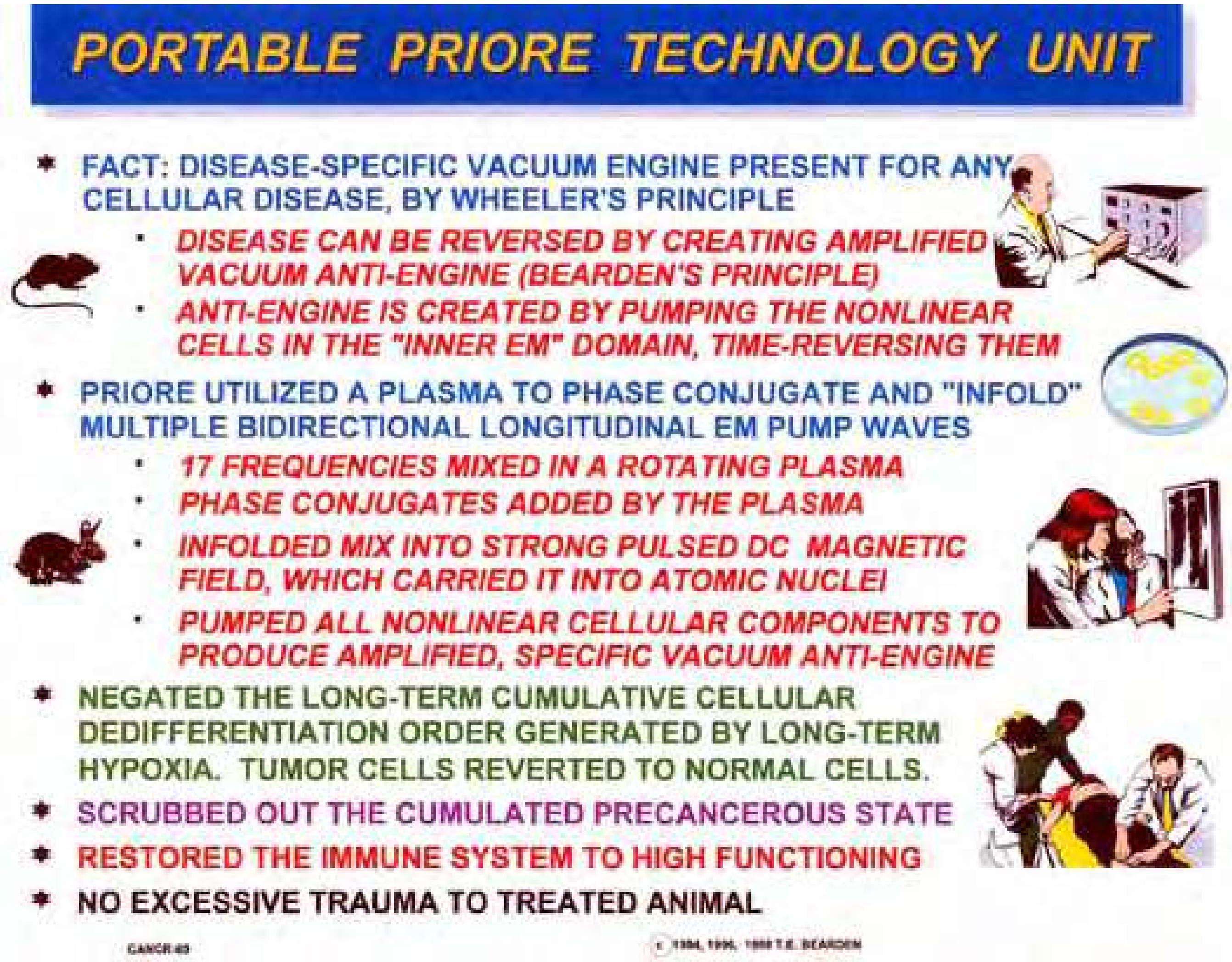

PORTABLE PRIORE TECHNOLOGY UNIT

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

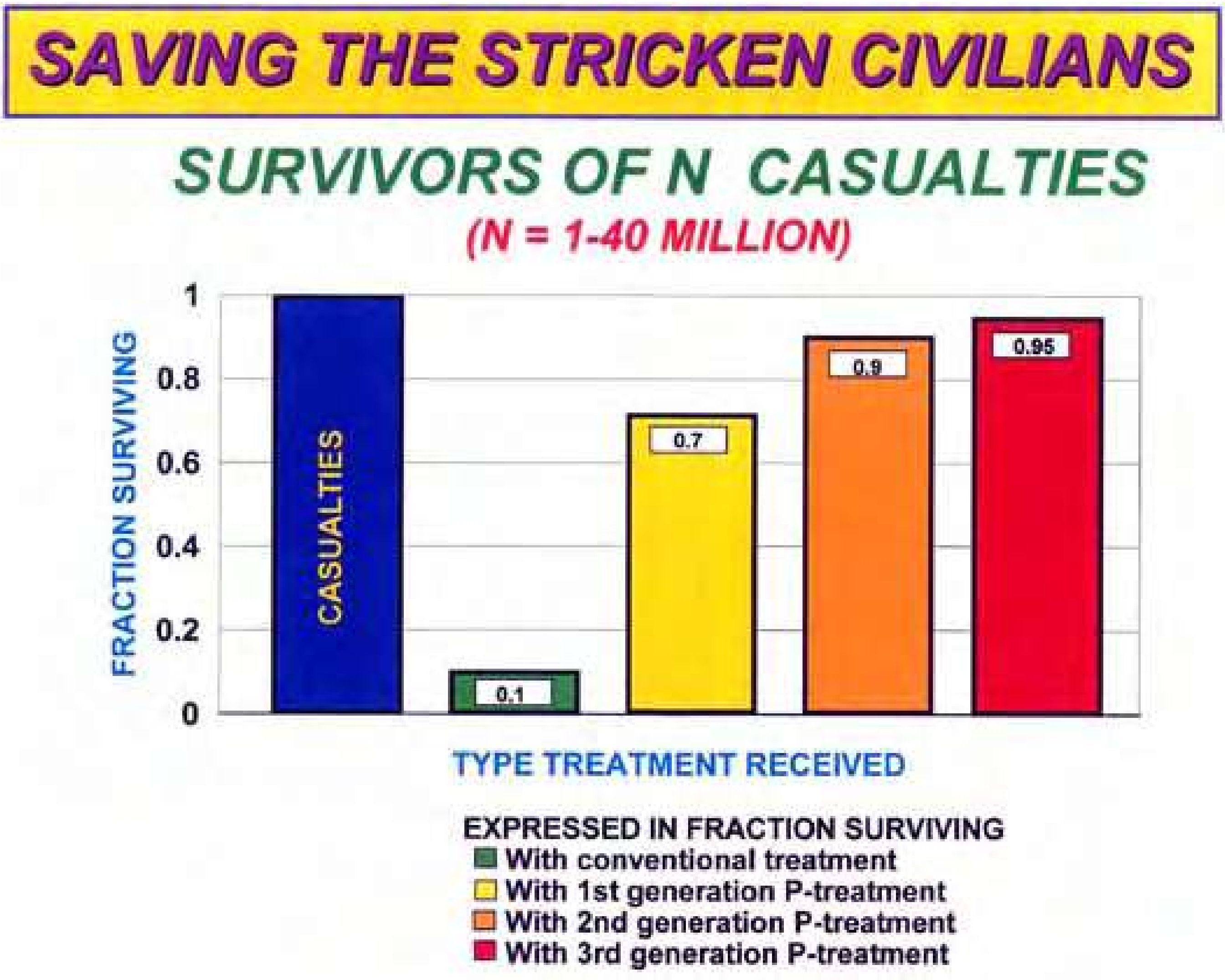

SAVING THE STRICKEN CIVILIANS

SURVIVORS OF N CASUALTIES

(N = 1-40 MILLION)

EXPRESSED IN FRACTION SURVIVING

- 🟩 With conventional treatment
- 🟨 With 1st generation P-treatment
- 🟧 With 2nd generation P-treatment
- 🟥 With 3rd generation P-treatment

Return to Slide Index

Next Slide

Previous Slide

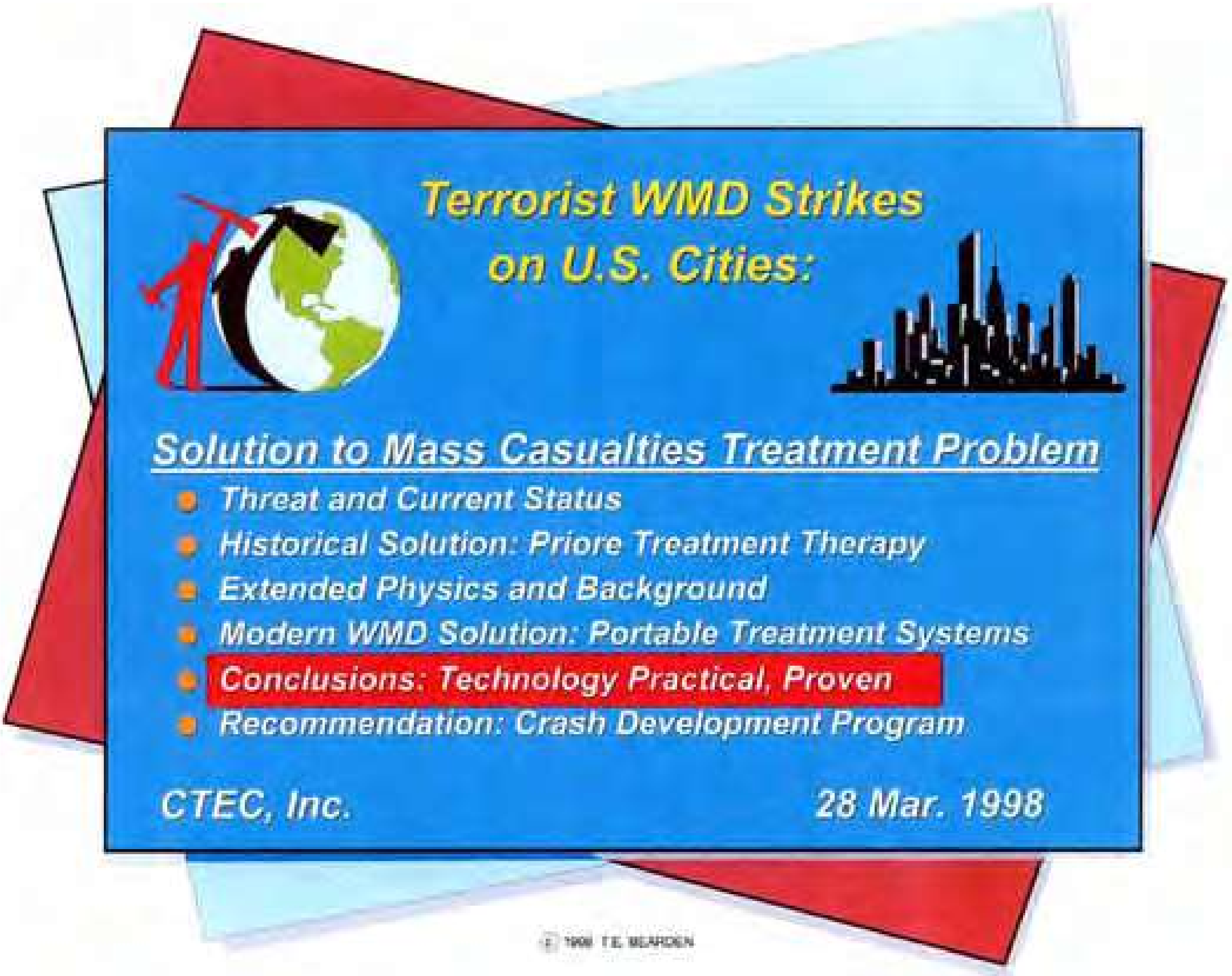

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

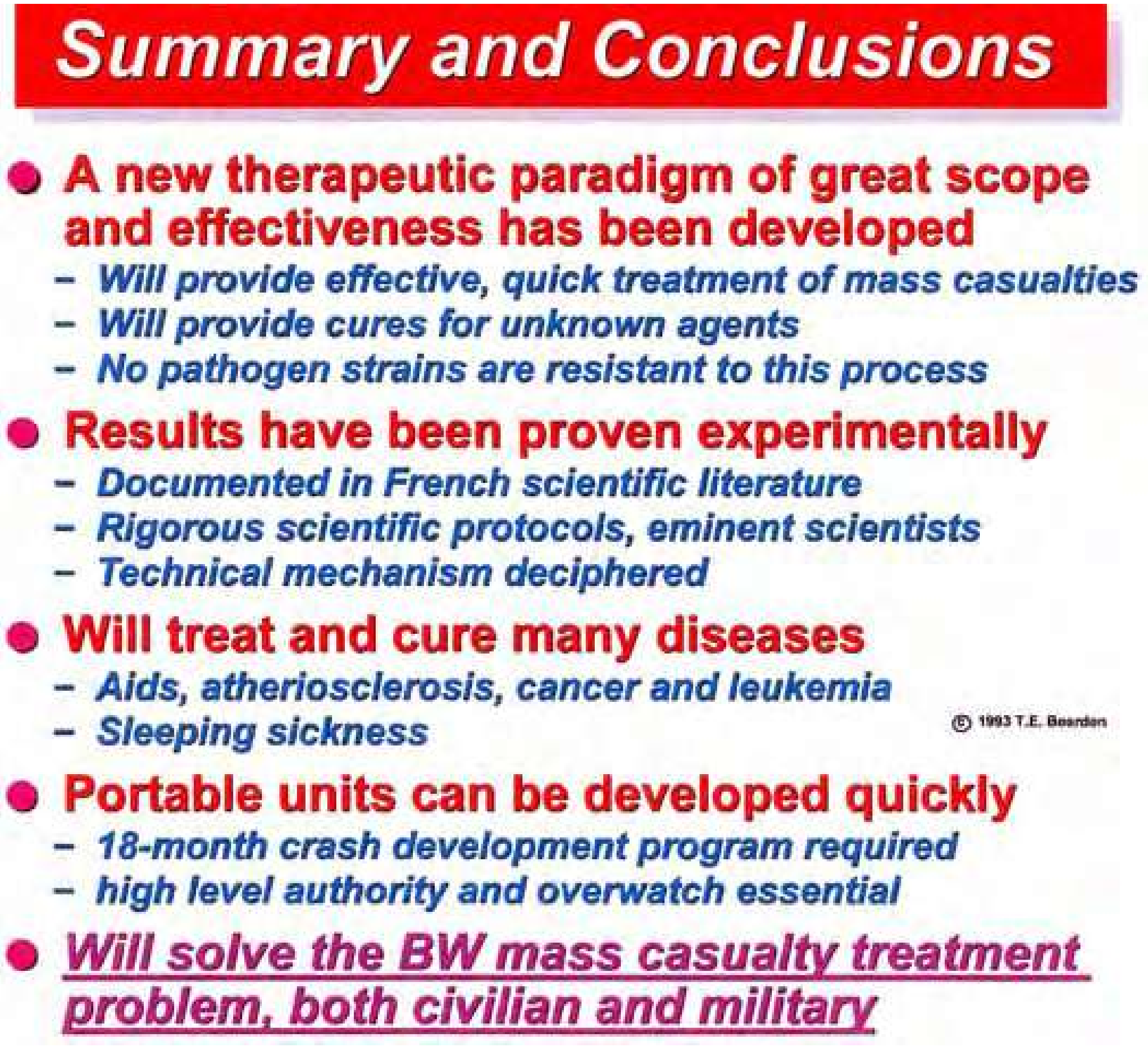

Summary and Conclusions

- A new therapeutic paradigm of great scope and effectiveness has been developed
  - *Will provide effective, quick treatment of mass casualties*
  - *Will provide cures for unknown agents*
  - *No pathogen strains are resistant to this process*
- Results have been proven experimentally
  - *Documented in French scientific literature*
  - *Rigorous scientific protocols, eminent scientists*
  - *Technical mechanism deciphered*
- Will treat and cure many diseases
  - *Aids, atheriosclerosis, cancer and leukemia*
  - *Sleeping sickness*
- Portable units can be developed quickly
  - *18-month crash development program required*
  - *high level authority and overwatch essential*
- *Will solve the BW mass casualty treatment problem, both civilian and military*

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

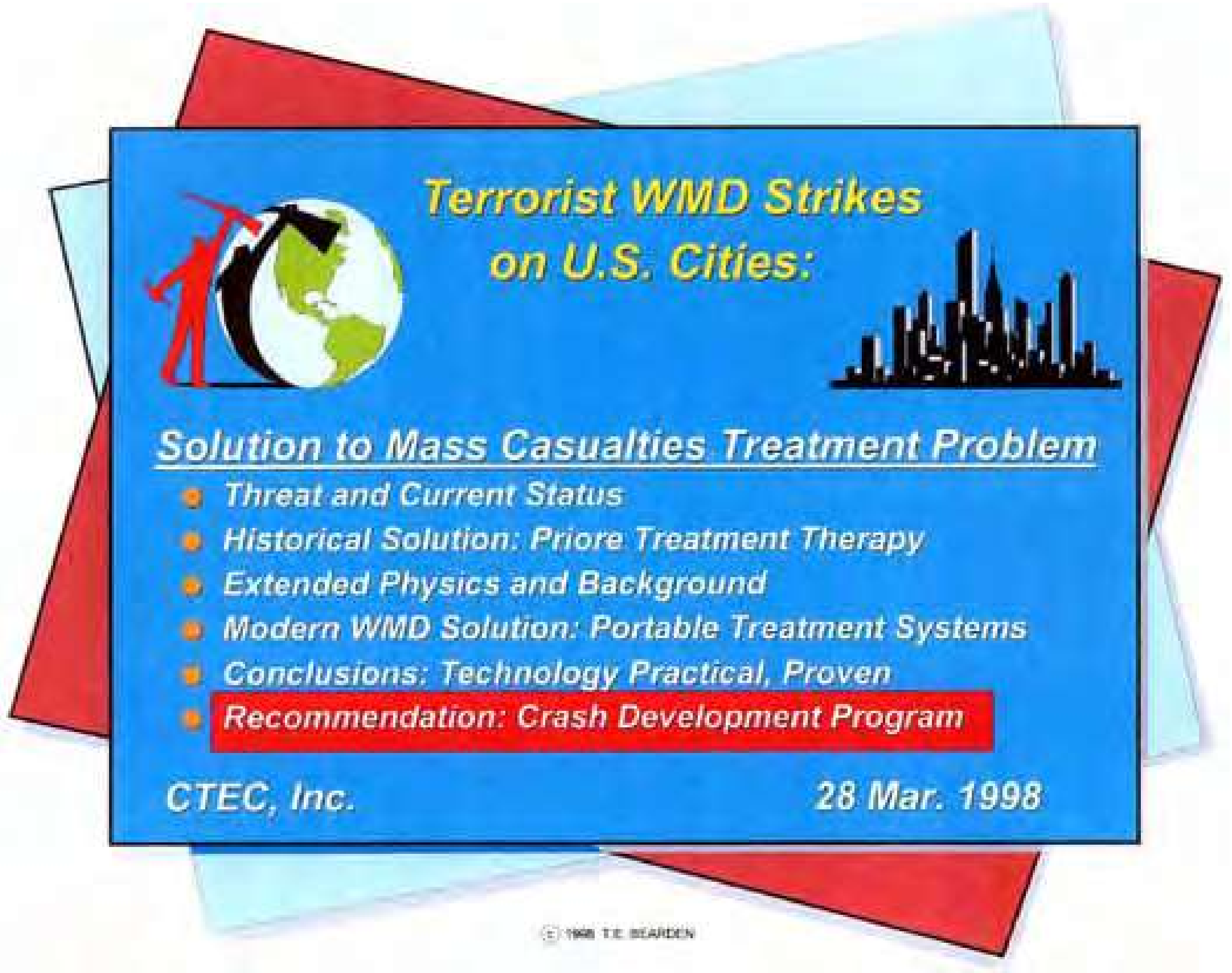

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

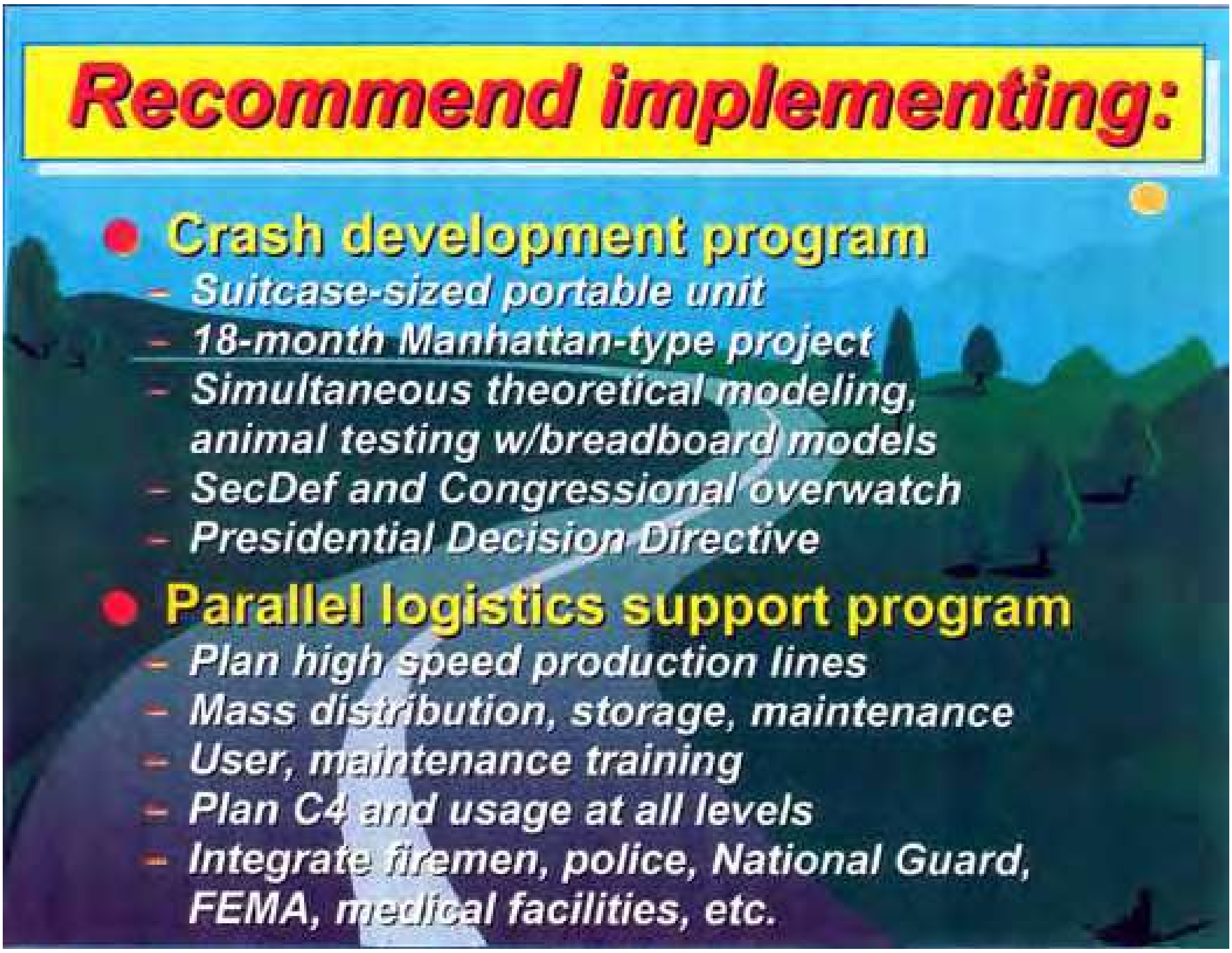

Recommend implementing:

- Crash development program
  - Suitcase-sized portable unit
  - 18-month Manhattan-type project
  - Simultaneous theoretical modeling, animal testing w/breadboard models
  - SecDef and Congressional overwatch
  - Presidential Decision Directive
- Parallel logistics support program
  - Plan high speed production lines
  - Mass distribution, storage, maintenance
  - User, maintenance training
  - Plan C4 and usage at all levels
  - Integrate firemen, police, National Guard, FEMA, medical facilities, etc.

Return to Slide Index

Next Slide

Previous Slide

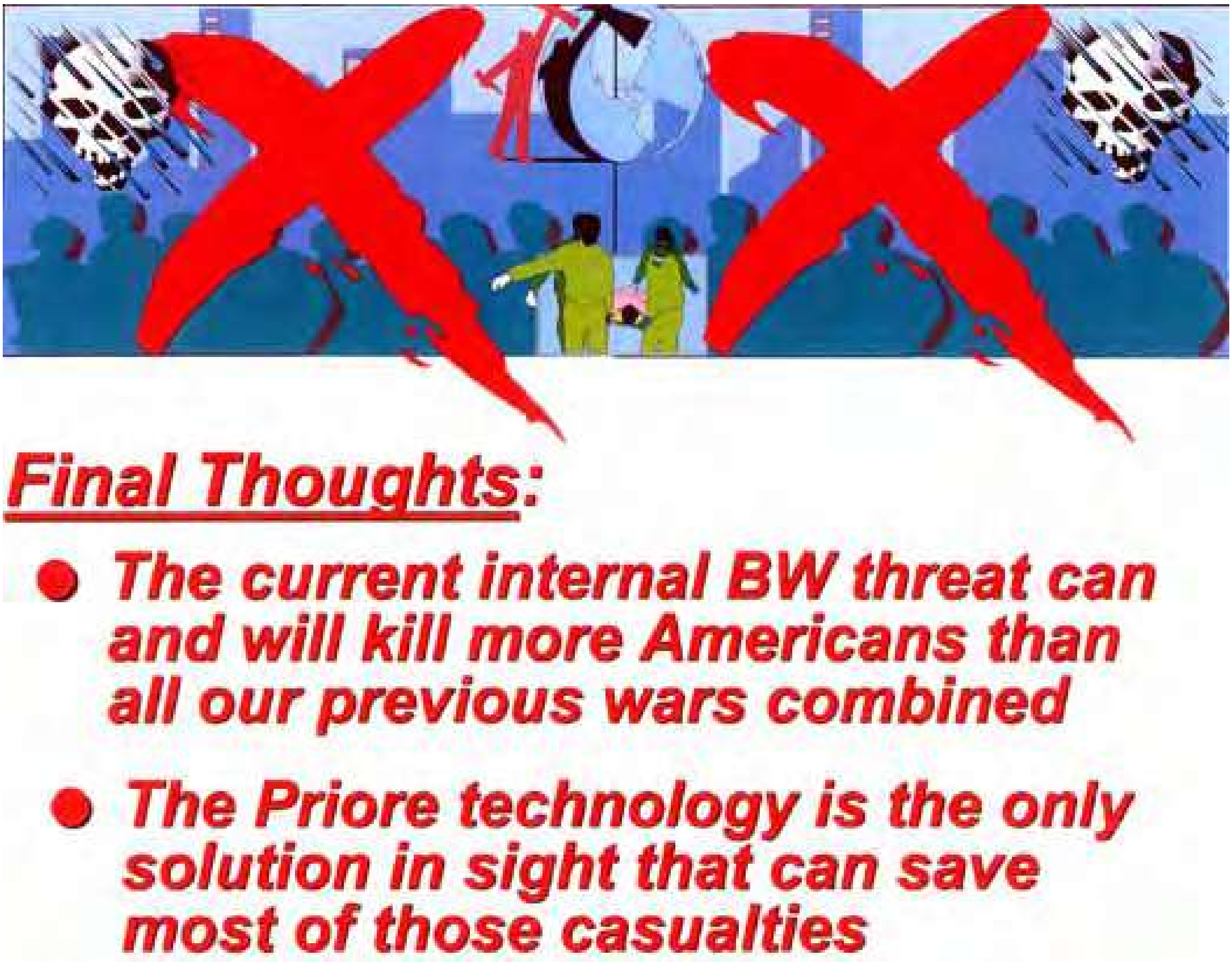

Final Thoughts:

- The current internal BW threat can and will kill more Americans than all our previous wars combined
- The Priore technology is the only solution in sight that can save most of those casualties

Return to Slide Index

Previous Slide

CTEC, Inc.
2311 Big Cove Road
Huntsville, AL 35801
(256) 533-3682 Ph/

CTEC, Inc.

June 24, 1998

Dr. Jane F. Kinsel, Director
Office of Policy Analysis
Department of Health & Human Services
Bathesda, MD 20892

Dear Dr. Kinsel:

Your letter of June 8 is appreciated. However, the entire action package had already gone to the Assistant SecDef and also to Major General Neary in the USAF.

The proof of a revolution in medical treatment was in the package. The basis for it is already demonstrated experimentally in the scientific literature, both in the U.S. and in France. We are speaking of something already proven, but missed by the entire U.S. medical community. Strong words, but true. Let me be specific.

Any cell in the body can be time-reversed (physics term) or dedifferentiated (biology term) back to an earlier state, by pumping (i.e., "squeezing" in simple terms) that cell with longitudinal EM waves as pump waves. There are considerable papers in the French scientific literature showing the experimental results.

Apply that to cancer. We are stating bluntly that all cancerous cells in the body can be "depromoted" back to normal cells easily. Further, the damage previously done to them (such as by sustained hypoxia due to contaminants reducing the oxygen-carrying ability normally given to it by some 60-to-80 surrounding water molecules per red cell and the consequent H-bonding interactions) will also be reversed. The cells can all be returned to healthy, normal cells again, without "killing" or "burning" or "cutting" anything.

With some development of the technology, you can quickly and easily cure every cancer patient in every hospital in the United States.

Now apply it to AIDS. Since you can time-reverse every cell in the body, you can time-reverse the HIV-infected cells – genetics and all – right back to normal cells with normal genetics. Those cells not infected and healthy, will just get a "little younger" again.

So with that small development, you can quickly and easily cure every AIDS patient in every hospital in the United States.

I would have thought that NIH would certainly be interested in such a revolutionary therapy which has already been demonstrated in laboratory experiments. World-renowned French scientists – such as the eminent parasitologist Pautrizel and research PhDs assigned directly by Robert Courrier, head of the Biology Section of the French Academy of Sciences, worked with Prioré in performing those startling experiments that proved what we are saying.

The method was rigorously shown to cure infectious diseases (such as trypanosomiasis), atherosclerosis (which now afflicts millions of Americans), and terminal tumors. It was shown to reverse suppressed immune systems back to normal vigorous functioning.

Please permit me to disagree with you that the examination of such is not the mission of the NIH. To the contrary, it is precisely the NIH which should have some of its very best scientists looking into this.

The action package to DoD was to propose a crash program to develop this methodology in the form of a small, portable unit for treatment of mass casualties resulting from a terrorist BW attack on our cities (now expected in the next few weeks, as you must be aware). The leading Arab terrorist of the world has already promised just this.

We are speaking of several million Americans who are going to die. They are going to die in spite of everything that NIH and its labs can do, even though I know you will make a heroic effort. With NIH present methods, you cannot do very much for those stricken Americans at all. Everything we have in the "normal" kit bag is almost useless against a knowledgeable spray attack of modified smallpox (as you know, the Russian secret BW labs have made and sold tons of that already, to guess who!). Or a spray of tularemia (you are well aware of the extreme lethality of that), or anthrax (easily obtained). It is also common knowledge that these terrorist teams with their BW agents are already on site in this country, waiting for the word to attack. This has now been officially raised to our primary Strategic Threat.

In the face of such looming strikes on the U.S., I simply cannot believe that the NIH does not feel it within their mission to rigorously check out a proposed method for saving perhaps 70% of those coming deaths (first generation equipment), and up to 90% with second generation equipment.

Perhaps you just did not read the package carefully. I urge you to do so.

Else NIH is going to look very, very bad when (1) the strikes do occur, (2) they can't handle them (and they cannot, as you well know), (3) they had within their hands in advance, a proposed methodology that could have been investigated intensely, to save millions of those stricken Americans.

As a staunch supporter of NIH and especially CDC, I have to believe that there is a different attitude there. If NIH is not interested in a previously experimentally demonstrated method of saving millions of Americans, then the only recourse left is the political channel.

I assure you that my critique of electrodynamics is well-founded, as supported by various leading physicists such as Nobelist Feynman (who stated bluntly that the field concept is wrong), Wheeler, Bunge, and so on. It is inexplicable why all our universities are still teaching a 130 year old abridged (Heaviside's) subset of Maxwell's theory, and one that still contains a material ether even though that was falsified 110 years ago!

You already have seen the misuse of this technology by the Russians, in the decades-long radiation of the U.S. Embassy in Moscow. Is it not NIH's mission to understand how they generated all those diseases and health changes? Have you ever looked into Kaznacheyev's experiments showing that ANY cellular disease or disorder can be induced in cells at a distance by purely EM means? Have you looked into what got eliminated from electrodynamics by ignoring the hidden longitudinal EM inside every potential, field, and wave? Do you realize what ignoring that has done to American chemistry, biochemistry, and medical science?"

I would hope that the NIH would at least be interested in its own mission area!

I challenge you bluntly to admit that no one presently understands the primary operation of the human cellular ***regenerative*** system. That's because its technical principle is not presently in our recognized physics, biochemistry, and medical science – even though it has been in the hard physics literature since 1903-4. Becker's epochal work came closest; since he only had access to the conventional EM model, it failed him. But he at least showed you rigorously that cells can be changed (both differentiated and redifferentiated) all over the map by laughably weak EM – picoamperes of current, which means persistent DC potentials. In the microwave radiation of the U.S. Embassy, all the health changes occurred in regions absent of fields – i.e., absent of bleed-off of the potentials. Hence in regions of stable DC potentials! With 100% correlation of the health changes to persistent, gradient free potentials, and 100% anticorrelation to changing potentials (i.e., fields), our own scientists concluded totally erroneously that it could not have been the EM radiation causing the changes, when their own results had just proven that it was.

Now I ask you: Can NIH explain how it was that the DC potentials caused those health changes? I think not. If it could, it would understand the action package I wrote, and its importance.

This is not a "political" letter. I'm just an ordinary American citizen who has done about 30 years very hard work on this particular area. The work is good enough now to stand on its own merits. The scientific community will eventually accept it, about 50 years from now. But we cannot wait 50 years! Those millions of Americans are going to die, because of the bureaucracy and inaction of our own scientific organizations. Whether they like it or not, the U.S. scientific community is now the forefront of the struggle for this nation to survive. And they are failing us dramatically.

So as a parent who wants his children to live, and his neighbors' children to live, and those millions of soon-to-be stricken Americans to live also, I urge you to at least have someone competent in physics (and knowledgeable that the foundations of physics itself are in woeful shape!) check the work. Has anybody there *read* any of the references to the Prioré work I cited? Does anyone there realize that *there are no EM forcefields in space?* Probably not.

Dr. Kinsel, there is a time for "business as usual" and there is a time to do some serious new work. Those stricken Americans are going to be lying there. Nothing NIH can do is going to stop that. That's a given, and it is not my own estimate. It's the official estimate. And you *know* you presently do not have the tools to do anything really effective about it.

Frankly, I interpret your letter to me as just the normal "spin control" for easy disposition. I would hope that NIH would seriously rethink the situation, get off the "spin control" and "business as usual" posturing, and do something dramatically different for a change.

What have you got to lose? A little time from perhaps two or three of your best scientists.

What have you got to gain? Saving untold millions of American lives. A new, *already partially demonstrated* cure for most dread diseases such as AIDS and cancer – diseases which you presently cannot do very much about in spite of decades of heroic struggle.

Sincerely,

cc: Senator Shelby, Congr. Cramer

T.E. Bearden
President and CEO

CTEC, Inc.
T.E. Bearden, President & CEO
2311 Big Cove Road
Huntsville, AL 35801-1351
(256)533-3682 ph/(256)536-0411 Fax

CTEC, Inc.

May 12, 1998

Dr. Harold Varmus, Director
National Institute of Health
Bathesda, MD 20892
(301) 496-1766

Dear Dr Varmus:

We have recently sent action documents to the Director of Nuclear and Counterproliferation, Office of the DCS, Air and Space Operations, HQ USAF and to the DoD (General Busbee) Assistant Secretary of Defense, as well as several Senate and House committees, which may be of interest to you. For one thing, we give the technical mechanism used to generate the Gulf War Syndrome. The same weaponry is now being used to rapidly induce new strains of certain pathogens worldwide. I urge you to closely read that section of the letter to MG Neary which explains the GWS, why the French did not get it, and why Southern Iraqis got an increase in cancer and leukemia but not GWS. Also explained is how the "microwave radiation" of the U.S. Embassy in Moscow induced diseases for decades, and how we failed to comprehend it because the standard EM model deceived us.

There is a way to treat and quickly cure most diseases, including the mass casualties resulting from a terrorist BW strike on our civilian population centers, our military bases, or our forces in the field. The method is little known, but was scientifically demonstrated in France in the 1960s and early 1970s before cancellation of the Prioré Project when the French Government changed. Eminent French scientists worked on the project, and the results are fully documented in leading French scientific journals. References are listed in a separate attachment.

Previously the Western scientific community has been unable to understand the technical mechanism responsible for such remarkable cures, because of serious foundations flaws in Western physics and electrodynamics. *Foundations* physicists have long been aware our science is flawed, and have pointed out many of these fundamental errors. However, the operational science establishment has not paid attention, even though physicists such as Feynman and Wheeler pointed out that the notion of "force fields in space" was totally wrong.

As a primary example, there is an "infolded" electrodynamics inside all potentials, fields, and waves that is far more primary than the coarse EM that is in our textbooks. It is actually an infolded general relativity, and this is what the Russians have weaponized. This infolded EM-GR has been ignored in the West (but not in Russia) since 1903.

As another example the potentials, fields, and waves of present electrodynamics are rigorously defined – and exist – only in the presence of mass. They do not exist in such form (force fields and oscillating force fields) at all in space, in the absence of mass. Rigorously, Maxwell's equations are mass-to-mass transforms, since Maxwell and everyone at the time assumed the ubiquitous presence of the material ether. In other words, to the founding electrodynamicists, there was no place in all the universe that was devoid of mass. Hence they defined all EM entities as entities containing mass. Even though Michelson-Morley experiments of more than a century ago destroyed that material ether, not a single Maxwellian equation has ever been changed! They still assume the material ether.

Specifically missing from our electrodynamics are two essential transforms: (1) the mass-to-spacetime transform, and (2) the spacetime-to-mass transform. In other words, these two transforms are "infolded" or hidden inside (comprise) the end point extremum case – the mass-to-mass transform actually written by Maxwell.

But these two missing transforms are just general relativity (they correspond precisely to Wheeler's principle of general relativity). So general relativity has always been infolded inside electrodynamics as we know it, but ignored. Further, this is a very powerful general relativity, since it uses the very strong EM force as the agent of spacetime curvature. Hence now one can engineer a powerful general relativity in the laboratory and in practical devices, and even action-at-a-distance, which one cannot do using merely the staid and weak G-force.

It turns out that longitudinal EM waves as shown by E.T. Whittaker in 1903 and 1904 are far more primary than our present EM theory and even our present GR theory. Both EM and GR are directly engineerable using longitudinal EM waves, including subluminal and superluminal EM waves in vacuum, and including action at a distance. One can engineer complexes of spacetime curvatures as desired, which in turn are *spacetime engines* (Wheeler's term). These engines can act upon mass – either living or inert – in any fashion desired, and at any distance desired.

The little-known and poorly understood Regeneration & Recovery (R&R) system – as contrasted to the immune system – utilizes this exact "spacetime engine" methodology to restore damaged or diseased cells, within its limitations – "dedifferentiating" or "time-reversing" them back to normal. The mechanism is fully explained in the attached package. The proposed program applies this fundamental method to greatly amplify the effect. Thus cancer cells can be reversed back to normal cells, HIV-infected cells (HIV-factories) can be reversed back to normal cells (genetics and all), etc. This represents a dramatic and unparalleled new medical therapeutic methodology of extreme power. The same techniques can simply be conjugated one additional time, and used to produce diseases and cellular changes at a distance, including in entire mass populations, as the KGB is doing right now. Specifically, by impressing an extremely weak "cocktail mix" of disease spacetime engines on the U.S. populace, the immune systems are reacting and being "spread thin" across several "shadow state" infectious pathogenic conditions. Thus when an actual terrorist BW attack occurs, the lethality of the strike will be remarkably enhanced because the immune systems will be "thinned" in their ability to fend off this new pathogen. By doing it this way, BW warfare can actually be conducted upon the U.S. without anyone being the wiser.

Our very survival as a nation is threatened by remarkable KGB weapons using this fully documented, extended electrodynamics. Bluntly, unless we take off our scientific blinders and understand Russian energetics weapon science, GWS and the Embassy health changes in Moscow were just a picnic before what is upon us. After two decades of work on the KGB energetics weapon problem, we have fully deciphered the technical mechanisms and what energetics is, as well as citing many of the Russian weapon tests and how they were done. Again, you can see that for yourself in the package.

Now that we understand the technical mechanisms, we are proposing rapid redevelopment of the Prioré process, in portable suitcase-sized units to be mass produced and filtered down into the entire emergency community for mass treatment of lethal infectious diseases resulting from BW strikes. And we are strongly urging the U.S. scientific community to quickly investigate and come to grips with this vast new *biological engineering* mechanism. Our very survival depends on it.

The inclosed documents provide an overview of the process. There is nothing else on the scientific horizon that can save the majority of all those stricken Americans, once the first foreign-sponsored terrorist teams strike our population centers with BW agents.

Sincerely,

Incls: Full package with table of contents

T.E. Bearden
President & CEO

Scalar Wars

The Brave New World of Scalar Electromagnetics

by Bill Morgan

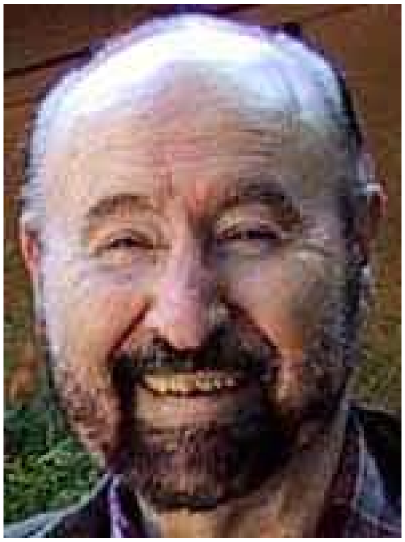

For the past six months I have been undergoing the greatest **paradigm shift** I have ever had to go through. It has rattled my nerves and shaken my bones. This intense adjustment of my "world" has come about by studying the information given by Col. Tom Bearden at his website Cheniere. The new knowledge there has necessitated a total revision of my ideas about physical reality, the world we live in, and the future of humanity. This paradigm shifting even actually made me dizzy on certain days as I tried to absorb and digest Bearden's vast amount of information. I am not a scientist at all, just a layman, and I have little comprehension of the math and high physics of this new science called "**Scalar Electromagnetics**." But there is a great deal of information at Cheniere which needs to become common knowledge as fast as possible, for the sake of the survival of life on earth. To that end I have put together this small primer of Bearden's ideas, as a kind of "beginner's guide" to his website. (I have tried to underline certain words and phrases which are part of the new "jargon" of talking about these "new" phenomena. Also I have taken the liberty of bolding certain words and phrases to help them stick in the mind).

(Note: Throughout "**EM**" means "**electromagnetic**," and "**LW**" means "**longitudinal wave**")

This article has six sections:

New Waves 2. Tapping the Waves 3. Weaponization 4. Healing 5. Psychoenergetics 6. As It Stands

1. New waves discovered

Longitudinal EM energy fills vacuum of space, the time domain of spacetime, time as compressed energy, E=tc2, waves of time, phase conjugate wave pairs.

"Electric power is everywhere present in unlimited quantities and can drive the world's machinery without the need of coal, oil, gas, or any other of the common fuels." Nikola Tesla

"At any point and at any time, one can freely and inexpensively extract enormous EM energy flows directly from the active vacuum itself." Tom Bearden

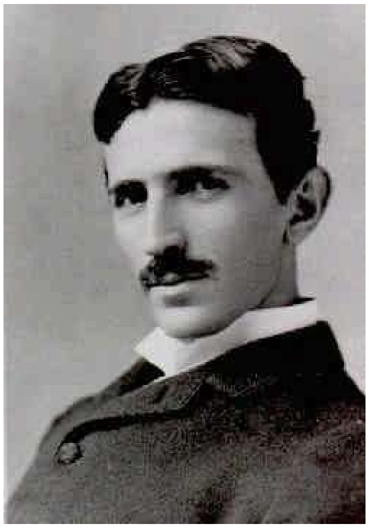

I guess the first thing to try to comprehend is that a "new" kind of electromagnetic (EM) wave has been discovered in the empty vacuum of space called “longitudinal” waves, which, when engineered, can be an inexhaustible supply of energy in great magnitude at any place in the universe. The word "new" is in quotes because the discovery really goes back to **Nikola Tesla** and his discovery of what he called "radiant energy." It is also not "new" because the Russians (KGB) have been working on this technology for over 30 years and have **weaponized** these "new" longitudinal scalar waves to a great degree.

These are the very weapons **Nikita Khrushchev** spoke of in January, 1960.

> "By 1957-8 the Soviets had progressed to the point of a giant scalar EM accident in the Urals which exploded nearby atomic wastes, devastating the area. They had also progressed to development of great new superweapons using their new energetics - weapons to which Khrushchev referred in 1960 when he informed the Soviet Presidium of a new, fantastic weapon in development, a weapon so powerful that it could wipe out all life on earth if unrestrainedly employed."' Tom Bearden

After over 30 years of development, and extensive testing around the globe, these new scalar electromagnetic weapons are up and running and ready to go. Tom Bearden, at his website Cheniere, discusses the history of these new scalar electromagnetic weapons in his paper "Historical Background of Scalar EM Weapons."

Some Immediate Implications

> The implications of successful engineering of the longitudinal waves are enormous, and will change the world as we know it, one way or another. Among other things, these discoveries mean that:
>
> 1. The **solutions to the energy crisis and the "oil problem" are in hand**. These oil wars are unnecessary. There is endless energy available freely from the domain of time.

2. **Unbelievably powerful weapons** are not only possible, but are already operating in several nations. The many powers of these weapons are unprecedented and mind-boggling.

3. The **cure of diseases such as cancer and AIDS**, in fact nearly any disease, has become possible within a few years of sufficient funding. Everyone can be made healthy and stay healthy.

4. **Mind control on a mass scale** has now become possible, and the machines to do it are already in place in certain nations. It has become possible to mentally enslave whole populations with the twist of a few dials.

So the layman will need to understand that there is a new kind of electromagnetic energy that is altogether different from what he knows, e.g. radio, TV, cell phones, etc. The ordinary EM waves that we have known about are called transverse EM waves, to distinguish them from the new longitudinal EM waves. These scalar waves do not actually exist in our "material" world, but exist only in the vacuum of empty space, or the time domain. And we must keep in mind that this vacuum of space we speak of exists all through everything. Even our bodies are mostly empty space between atoms and molecules. So the gateway to this seething ocean of energy can be there at every point in the universe. This seething ocean of energy is all around us and all through us.

Emptiness is Full

This amazing discovery announces that the "emptiness" of empty space is in fact not empty, but a great ocean of seething energy!

Col. Bearden refers to this ocean of energy as being of the "time domain." It seems like something from Star Trek but this is where scalar electromagnetics has come to. And where it is going may be beyond anything Star Trek could have dreamed of.

We live in a 3-dimensional world, which physics calls "3-space." But there is also spacetime, or 4-space, or the "4th dimension." Then suddenly comes this amazing new knowledge that time itself is actually compressed energy. And it is energy which is compressed by exactly the same factor by which matter is considered compressed energy: the speed-of-light-squared!

So we have a new companion to the famous E=mc2. It is now paired with E=tc2 (where t is actually "delta-t," or change in time). Or (amaze your family and friends and) say:

"E equals em-cee-squared"

and

"E equals delta-tee-cee-squared"

It has a nice ring to it, like some freedom bell announcing a new era.

And as the atomic bomb released the compressed energy in matter, so can we now unleash the tremendous energy that is compressed into time itself. It gives a completely new meaning to the term "time bomb."

Some New Terminology of Scalar Electromagnetics

A new science, several new names

> "Those [Russian] weapon scientists resurrected an old term from the history of electrodynamics, called "energetics". That is their approach to a unified field theory, where everything is based on "energetics". This model as its foundations uses a very similar approach to that "single fundamental unit" model, where energy is the unit. If one makes the energy EM in nature, then one has the Russian energetics approach. This unified approach gathers everything in, including all energy actions and relations in inert matter (the first branch of energetics, called by the same name), all field and matter interactions in living matter (the second branch of energetics, called "bioenergetics"), and all mind operations and mind-matter interactions (the third branch of energetics, called "psychoenergetics"). [ed: sometimes also called "psychotronics" by the Russians]

Rick Andersen in an article called "What is Scalar Electromagnetics?" describes the new science this way:

> "Scalar EM is the brainchild of Lt. Col. (retired) Thomas E. Bearden, a systems analyst and wargames specialist who has been advocating a view of electromagnetics which is based on the notion of a vast, unseen background of scalar energies (as opposed to vector energies) which underlie all physical reality.
>
> "If Bearden is correct in his Scalar EM theory, then we can build devices which would enable us to alter gravity, time, inertia, and the apparent mass of an object. This of course has ENORMOUS implications for military applications, space-vehicle drives, time-travel, teleportation, paranormal phenomena, and just about every other area one can think of." http://twm.co.nz/Beard_scalem.html

> "Time-energy, time-currents, and time-structuring play the dominant role in electromagnetics. Time-as-energy eventually becomes engineerable, as easily as is spatial energy now. We are always dealing with spacetime and with spacetime curvature."
>
> Bearden: http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

The scalar energy discovered in the vacuum of space is sometimes called "zero-point" energy. Thomas Valone explains how this term came about.

What does "Zero Point" mean?

"Boyer traces the historical “creation of the vacuum” as proceeding in stages in parallel with the historical development of ideas about the vacuum. To paraphrase, he says that in the 17th century, it was thought that a totally empty volume of space could be created by simply removing all matter and, in particular, all gases. That was our first concept of the vacuum. Just get rid of all the gas. Late in the 19th century, it became apparent that the region still contained thermal radiation. But it seemed that the radiation might be eliminated by cooling. So the second concept of getting a real vacuum is to cool it down to zero temperature. Just go all the way to absolute zero. Then we’ve got a real vacuum. Right? Well, since then, both theory and experiment have shown that there is a non-thermal radiation in the vacuum and that it persists even if the temperature could be lowered to absolute zero. Therefore, it was simply called the “zero point” radiation. "Thomas Valone http://www.seaspower.com/InsideZeroPoint.htm

> "**For our very survival, it is absolutely imperative that informed citizens be aware of this dramatic change, which is just now starting. The powerful new science and engineering must be controlled and used for humanity's benefit, not its detriment. Else it will eventually be let loose unrestrainedly, to destroy all life on earth - a possibility indicated by Nikita Khrushchev in 1960**." - Bearden
>
> http://www.cheniere.org/explore%20articles/priore1/p4.jpg

2. Tapping the waves

Endless free energy, Tesla's radiant energy, the dipole as the gateway to the energy of the void, MEG in one year, Bedini, Yakuza, etc. End of Oil, geopolitical chaos, rogue groups

Move Over G.E. & Big Oil, Energy is Everywhere!

Third-world nations set to rise

> "Certain powerful interests did not wish the extended electromagnetics to be discovered or taught. That, after all, would lead to free energy and loss of economic control of the citizens. That was precisely why Nikola Tesla had already been suppressed ... Today it is still being suppressed by the orthodox establishment."
> http://www.cheniere.org/books/ferdelance/s25.htm

The vast seething ocean of energy of the vacuum, the longitudinal scalar EM waves of the time domain, can now be tapped and "transduced" into ordinary (transverse) EM energy in our 3-Space world. The process can be compared to putting a paddlewheel into a river. The energy acquired is free, since the river is there flowing whether we tap it or not. And it is a mighty river, and is not diminished by our paddlewheel.

This means that a permanent solution to the "energy crisis" is at hand, if only those who hold it so secret would give it up to humankind.

It means our species' insane use of oil is no longer necessary. We can make automobiles that need no fuel at all. We can build motors that power themselves. It means that unlimited electrical power can be available for free anywhere in the world. Every house and even shack can have its own power supply.

Of course there is the cost of the equipment at first, but the energy flow is free for the presumably very long life (no moving parts) of the generator. It will probably take decades to shift over to scalar power.

But the ability to tap the scalar (longitudinal) EM waves of the vacuum (which exists everywhere) also means that unbelievably powerful weapons can be, and already have been developed. In fact these weapons are so dangerous that the fear engendered becomes a psychological obstacle to the grasping of the facts.

I struggled with this myself for months, and even though Tom Bearden had completely convinced me already, it is hard to accept the fact that for everyone on earth the "worst that might happen" has suddenly gotten much worse. It takes time to deal with that, to realize it and get over it.

Indeed, the possibilities of these new weapons, which are tapping the huge energies of time, are so terrifying it is natural for the mind to simply say: "I don't want to hear about that."

And the government, likewise, pretends these weapons do not exist and tries to keep it secret from the people. Because most people would indeed be truly alarmed to know the actual situation. Col. Tom Bearden has decided that the people need to know the truth anyway. As a People on a Planet we must face this now, and ban the use of scalar weapons in war or in mass mind control. The current deployment of these weapons is a world emergency which we must all face.

The physics and mathematics of the longitudinal waves and the time domain are clearly beyond the layman, but I recommend scanning through them anyway just to get a sense of the wonder and the elegance of it all and some of the exotic terms which are used. Two of the most important papers by Col. Bearden are:

Giant Negentropy from the Common Dipole

The Final Secret of Free Energy.

Despite the complexity of the science involved, Bearden sums up the essential secret of it all in a few words.

> "There is no problem at all in extracting all the energy one wishes from the active vacuum, anywhere in the universe, at any time. Just make a dipole.

> The problem is in (1) catching some of that freely gushing EM energy in a circuit containing a load, and (2) dissipating the caught and collected EM energy in that load to power it, without using half the caught energy to destroy the source dipole's).
>
> That is the ONLY real energy problem on the planet, and always has been."
>
> http://www.cheniere.org/correspondence/022502.htm

A dipole could be a battery, which is polarized into negative and positive. Or the terminals of a generator. Or any magnet with its two poles. Or an atom with its positive and negative charges. Or the earth's magnetic field, the sun, and all heavenly bodies. Even a galaxy must be some kind of mega-dipole. Wherever there is a dipole there is already immense scalar EM energy ushering out of and back into the vacuum. And there have already been demonstrated a number of diverse methods for tapping that energy as it flows. (John Bedini, Edwin Gray, Bruce de Palma, and others). Bearden provides what he calls a "partial list" on his website.

> "In other words, nature then gladly gives us as much EM energy flow as we need, indefinitely — just for paying a tiny little bit initially to "make the little dipole." After that, we never have to pay anything again, and nature will happily keep on pouring out that 3-flow of EM energy for us. This is the giant negentropy mechanism I uncovered, performed in the simplest way imaginable: just make an ordinary little dipole."
>
> Bearden http://www.cheniere.org/techpapers/Unnecessary%20Energy%20Crisis.doc

How to Buy a Free Energy Machine

Answer: Just wait one more year! 2003. Then get off the Grid!

Perhaps the most eagerly awaited event now is the marketing of Bearden's Motionless Electromagnetic Generator (MEG) scheduled to begin in about a year's time. This is when the proof will hit the fan. Many people have said to me, when talking about free energy, "I'll believe it when I can go to the store and buy one." Hopefully in a year's time they will be able to order their own MEG. And then they will believe, because it works and is working.

Says Bearden: "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat."

(Bearden : http://www.cheniere.org/correspondence/011202.htm)

Soon the cat will be fully out of the bag. The first MEG units will supply 2.5 kilowatts of free electricity. Forever. Units can be connected together to double or triple the wattage. After that gets going, 10 kilowatt generators are planned. Energy will flow freely from the time domain indefinitely, and there are no moving parts to wear out. This manufacturing project is a heroic effort and will change the world as we know it. For once people are in possession of such devices there can be no further denial from anyone, including the government and/or "secret-government."

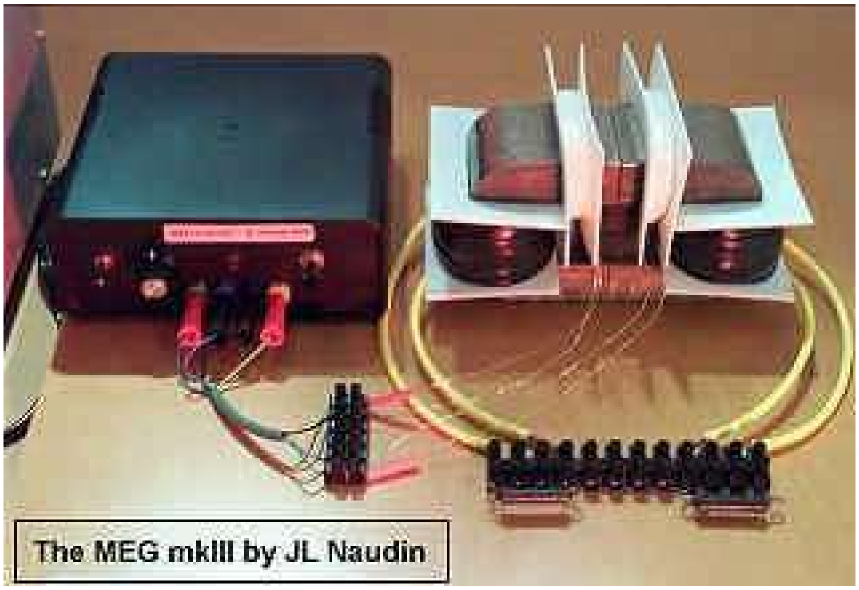

The Motionless Electromagnetic Generator

(Successfully replicated here by J. L. Naudin)

The physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum."

The device was successfully replicated by J.L. Naudin in France, and others, like Steve Utne, are also working on their own models. But Magnetic Energy Limited's MEG is actually scheduled for production soon and looks to be the first commercially available free-energy generator in history!

I imagine it will not take long for tinkerers to put them into their electric cars and make the first fuelless automobiles since Tesla's unheralded (and unheeded) experiments. Or for builders to begin building houses powered by MEG's. Houses which will never see electric bills. How can they ever make enough of these to meet the demand that will suddenly be there? Bearden is right, it needs to be an all-out crash program of the government on the order of the Manhattan Project.

I must include here that Bearden is quick to point out that he is not himself the inventor of the MEG, but is a partner and advisor in bringing it to the public market through Magnetic Energy Limited, Inc. I must also admit that I am praying for the success of this venture, and the ending of the absurd and criminal withholding of this technology from the people by the black-ops military-industrial people. This secrecy has cost the planet over 30 years of unnecessary pollution.

The first units off the assembly line, which is being set up in a "friendly" unnamed nation, will usher the public at last into the new era of scalar electromagnetics.

As of December, 2001, the situation with the MEG was described by Bearden in a letter.

> "What we presently have with the MEG is a successful laboratory experiment. At least a year's very hard research will have to be done before we will be ready to put a commercial power supply into production.
>
> "Consequently, we have made an agreement with a foreign partner (the National Materials Science Lab of the National Academy of Sciences of a friendly foreign nation) to do that year's research. At the same time, we are trying to make an agreement with one or more large financial partners here in the U.S."
>
> Bearden http://www.cheniere.org/correspondence/121901.htm
>
> "So we are hopeful that we will be able to start introducing units on the market about a year from now. Quite simply, we will either succeed or we will fail. But we will give it our very best effort."
>
> Bearden http://www.cheniere.org/correspondence/012202a.htm

Bearden has given another technical explanation of how the MEG works here in a letter.

Motionless Electromagnetic Generator patented, 3-26-2002

Click here to read the patent

"In Magnetic Energy Limited, all business discussions and arrangements are taken care of by Dr. Lee Kenny, Managing Partner. There are five of us who are co-inventors of the MEG, and two of the others are really the principal inventors." Bearden http://www.cheniere.org/megstatus.htm

Free Energy has been Suppressed

Secret Government Aware of Free Energy

> "In this world of contemporary times, all the agencies, CIA, FBI, KGB, NSA, Electric Power Research Institute, DARPA, Brookings Institution, Henry Kissinger, Edward Teller, et al., all are aware of [free-energy researcher Bruce] De Palma and his ideas.
>
> "Because these individuals and institutions are employed by the ruling elite to forecast the future and satisfy present needs and demands, it is clear that free energy is a threat to the world order constructed by business and the mindset of those who want to own the world.
>
> "Free energy represents Man's aspirations and dreams of freedom and equality, uniform division of resources and the ability to choose one's own future.
>
> "The fact that free energy is suppressed speaks to the greed and self-interest of a ruling elite which, even in the face of an emergency of global starvation, resource depletion and environmental pollution, will not give one inch if it means loss of control. This is an

> attitude of paranoid delusion and fantasy which can only arise from the alienation of a class of elitists who, through their money, are completely insulated from the reality of the day to day fight for existence of the common man."
>
> http://depalma.pair.com/Absurdity/Absurdity07/ProblemOfFreeEnergy.html

Bearden describes one case of suppression when the Yakuza killed the Kawai engine in his very presence in Huntsville, Alabama.

> "That night a jet arrived from Los Angeles, with a Yakuza on board. The next morning Kawai no longer controlled his company, his invention, or his own fate. The Kawai party was in fear and trembling. The Yakuza coldly dissolved the agreement, they packed up the two Kawai engines we had, and left. And that was that."

So it is not only governments who are keeping this revolutionary technology from seeing the light of day. The Yakuza? KGB? The Aum Shinrikyo cult? Our "leaders" need to be honest about what is going on, otherwise they look foolish, acting out a kind of "Alice In Wonderland" script before the public, a script that has become completely unbelievable.

Concerning the criminal withholding of this scalar technology by the military black-budget people, and the cover-up of the actual situation regarding scalar weapons throughout the world, Tom has this to say:

> "The high government officials of the United States of America grossly violated our constitution and their sworn oaths of office. Our constitution specifically charges the government of the United States to defend the people of the United States. That means our officials are specifically charged with defending our people. Yet they deliberately agreed not to do so, and they deliberately sacrificed us in advance.
>
> "They committed treason of the highest kind, by deliberately surrendering us to total destruction anytime the big balloon goes up."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Energy Crisis Solved

One of the "must read" documents at Cheniere is "The Unnecessary Energy Crisis: How to Solve It Quickly," especially since the "oil wars," which Bearden has foreseen for decades, have now already begun.

Here are some clips from that Document:

> "To avoid the impending collapse of the world economy and/or the destruction of civilization and the biosphere, we must quickly replace much of the "electrical energy from oil" heart of the crisis at great speed, and simultaneously replace a significant part of the "transportation using oil products" factor also.

> "The technical basis for that solution and part of the prototype technology required, are now at hand. We discuss that solution in this paper."

And although the solution to the energy problem and the oil wars may be at hand, Bearden suggests there is still an urgent rush that needs to be made to save human civilization from collapse.

> "Once the technology hardware solutions are ready for mass production, even with a massive worldwide deployment effort some five years are required to deploy the new systems sufficiently to contain the problem of world economic collapse. This means that, by the end of 2003, those hardware technology solutions must have been completed, and the production replacement power systems must be ready to roll off the assembly lines en masse.
>
> "The 2003 date appears to be the critical "point of no return" for the survival of civilization as we have known it.
>
> "Reaching that point, say, in 2005 will not solve the crisis in time, and the collapse of the world economy as well as the destruction of civilization and the biosphere will still almost certainly occur, even with the solutions in hand."
>
> "The good news is that we now know how to easily initiate continuous and powerful "electromagnetic winds" from the vacuum at will. Once initiated, each free EM energy wind flows continuously so long as the simple initiator is not deliberately destroyed."

On page 9 he explains the concept of the dipole, and how there is already huge scalar energy flowing in the time domain around any dipole. Hold a magnet or a battery in your hand and you hold a stress in time and a beautiful flowing loop of longitudinal waves!

Bearden decries the scientific community in the West for being unable to let go of their dogma long enough to see that the greatest discovery in human history lies at their feet, but for their stubborn ignorance.

> "The present energy crisis then is due totally to that "conspiracy of ignorance" we referred to, that is maintained by the scientific community, and that has been maintained by it for more than 100 years.
>
> "This is the real situation that the environmentalists must become aware of, if they are to see the correct path into which their energies and efforts should be directed — to solve both the energy crisis and the problem of gigantic pollution of the biosphere. "As far as anyone in the legislative branch of government ever understanding the profundity of the discovery and engineering of scalar waves, Tom Bearden says they are at a great disadvantage in doing so.
>
> "The nontechnical person — e.g., a Senator or a Congressperson — is operating under a distinct disadvantage. If he or she takes the stance that much better electrical power systems can readily be achieved, he or she is in fact opposing almost the entire set of University, Government Technical, Power Company, Battery Company, and Organized

> Science communities. Further, in most cases his technical advisors are themselves from one or the other of those communities, and likely to go back into that community or those communities when the Senator or Congressperson leaves office, or even before. So the Congress and the non-technical government community at large operate at a great disadvantage...Bearden National Emergency Declaration is Needed
>
> "The Government Non-Technical community (the Senate and the Congress, in particular) are in far better shape than the Government Technical community, to appreciate the world implications of the pending economic disaster. I am hopeful that both the environmentalists and the Government Non-Technical community will rapidly unite in a common goal to get this vacuum energy program launched, under a National Emergency declaration. If so, then they can solve the energy crisis and the pending economic crisis, in fairly short order, and permanently."
>
> unnecessary_energy_crisis.htm#congress_disadvantage

In the conclusion of his paper The Unnecessary Energy Crisis: How to Solve It Quickly Bearden again states the urgency of replacing our centralized power grids with highly dispersed and local taps on the freely flowing energy of the active vacuum. Given the abilities of the Tesla howitzers, the power grids of all nations are sitting ducks.

> "Personally, the present author regards the increasing energy crisis as the greatest strategic threat to the United States in its entire history. I will do anything within my power to help prevent what I perceive to be the looming economic collapse of the Western world, preceded or accompanied by a sudden, explosive, all-out and continuing exchange of the WMD arsenals of most of the world.
>
> ...In the name of all humanity, let us begin! Else by the time this first decade of the new millennium ends, much of humanity may not remain to see the second decade."

3. Weaponization

Scalar interferometry, Tesla howitzer, Tesla dome, heat/cold, dudding, 10 nations have them, dead-man fusing & Insane Systems, scalar strategy, earthquakes and weather wars, solar storms, mind attacks, leased to Yakuza, many examples of testing, quantum potential weapons.

> "Indeed, Soviet energetics weapons are now capable of destroying both our triad shield, our homeland, our armed forces in the field, and our population, quickly and efficiently. We have a new "gap" of monumental proportions: not a missile gap, not a submarine or bomber gap, and not even a particle beam or laser gap. We have a scalar electromagnetics or electrogravitation gap." Bearden http://www.cheniere.org/books/ferdelance/intro.htm

"**We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that,**

if unrestrainedly used, it could wipe out all life on earth. It is a fantastic weapon." Khrushchev, to the Presidium, Jan. 1960

> "We've totally missed the most incredible "secret superweapons" development program of all times, and the worldwide testing of the weapons themselves.
>
> "Hiroshima and Nagasaki should serve as object lessons in the price of ignorance of such technological breakthroughs.
>
> "History does repeat itself. But this time it's not we Americans who got the great new superweapon first."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Warfare has been changed forever by the development of these scalar energy longitudinal wave howitzers. To get a basic understanding of scalar waves is to have the imagination suddenly run wild as all the implications and possibilities regarding warfare fall into place. One realizes with a certain horror that the world has totally changed, and that there are some very fearsome possibilities. Remember, the power for these weapons comes from the time domain, longitudinal EM waves in the vacuum of empty space, and the power is tremendous and mind-boggling. Being able to blast away at any target from a distant control booth is something that has never happened before. This is incredible power to be in control of and it divides the history of weaponry into "before" and "after." And the destructive power of these weapons is delivered instantaneously to the target from the local vacuum at the place of the target.

From an interview:

> Questioner: "Are you saying that real energy can be transmitted faster than the speed of light?"
>
> Bearden: "Yes. But hyperspatially. Not through 3-space, but "around" it. Again, this means that it's possible to produce energetic changes in a distant system at a distant place, without transmitting energy "through space" in the normal sense. You transmit "around" space, so to speak, and directly in time or through a higher dimension, depending on the model you're using to understand this. Further, the speed of the internal EM energy transmission is not limited to the speed of light."
>
> http://twm.co.nz/beard_interview.htm

Here are some initial considerations about scalar weapons and scalar wars

1. Tesla Howitzer 2. Tesla Domes 3. Rogue groups 4. WW III 5. Russian "UFOs?"

3a. The Tesla Howitzer and its modes of action

Source: Tom Bearden's weapons slides and Fer-de-Lance Briefing

For one thing there is no delivery of any thing to any place. Bombs are obsolete. So are the planes to get them there. No helicopters necessary. All current nuclear devices are rendered nearly useless by electronic "dudding." All distant-destruction can be done from a control room, and to any point on earth. The destructive power of "men" has just increased by orders of magnitude. The new weapons could even cause storms on the sun! Thus there is great urgency to make these new facts public and known, with the hope that all nations would come together to ban the use of scalar weapons.

Two scalar antennae together along with the computers to control them, make up a scalar interferometer, and according to Col. Bearden the Russians have hundreds of such installations already. These interferometers are called "Tesla Howitzers." They can deliver a giant blast of energy to the distant target site, true "action-at-a-distance".

The first howitzer mode is called the "exothermic" mode of operation because immense EM energy blasts outward at the target site. The blast of a scalar howitzer can be of near nuclear level in destructiveness, and can be repeated easily, at that place, or nearby, or anywhere.

The howitzer can use a lesser exothermic power setting and simply destroy all electronics in the target area. Thus they can render our nuclear missiles inoperable as they sit in their silos by "frying" the electronic circuits that guide them. They can bring down any airplane, anywhere in the world, at any time. Any person anywhere, if their exact position is known, can be assassinated without a shot being fired.

The howitzers can also bring down power grids anywhere in the world, at any time. Thus they could bring Wall Street to an instant halt by disabling all its electronics. Even this one little easy tweak of the howitzer could cripple the U.S. economy in a single day. No wonder the "leaders" don't want you to know! This is a new era we are in. The discovery and development of scalar interferometry has put us in a new world, whether we like it or not. Anyone could be expected to go into denial, the befuddled senators and congressmen included. But we have to wake up.

In a second howitzer mode called the "endothermic" mode, the howitzer sucks energy out of the target area, essentially creating a blast of cold at the distant target. It is even capable of freezing parts of the ocean. Bearden gives a number of cases where these "cold explosions" have been witnessed, mostly by airline pilots. April 9, 1984: 1 2 3) A huge mushroom cloud of mostly water is seen rising miles into the sky from out of the ocean, an awesome sight.

In the endothermic mode the sucked-out energy must go somewhere, so it is vented out at some other chosen spot on the earth. These endothermic plumes have been photographed by satellite.

> "The exhaust in the left picture is about 150 miles long, and is nearly horizontal (about 1.5 degrees above the horizontal). It is entirely consistent with the continuous exhaust from a "dumping transfer" scalar EM howitzer in the continuous exothermic mode. (The primary howitzer, of course, was activated in the endothermic mode.)" Bearden More Plumes

By using both of these heat (exothermic) and cold (endothermic) modes together the weather can be altered anywhere. Warm the air over here, cool it down over there, put a curl in the jet stream, dissipate clouds, create clouds, whip up a tornado. Bearden speaks about using the endothermic howitzer mode as a weapon of war:

"Cold explosions can be used to freeze tanks, personnel, and equipment. The equipment and tanks thaw out. The personnel thaw out too, but they are dead when they do." Bearden, Fer-de-Lance

(This endothermic mode of the howitzers makes me think that there might one day be a cure for the problem of global warming after all. Could not a big device, situated in space or on the moon, be able to vent a great deal of excess heat outside the atmosphere into space, thus functioning as a kind of planetary thermostat?)

While the United States has experimented with electomagnetic weapons using ordinary transverse EM waves, the longitudinal [LW] weaponized waves are truly of another order altogether.

For one thing they do not have to travel through space, for they come from the time domain, which is everywhere. The energy of the blast, the heat or light or whatever, comes from the vacuum of space at the location of the target itself. This is astonishing. The weapons actually just trigger the release of immense energy from the vacuum at the target location.

> "... it is possible to focus the potential for the effects of a weapon through spacetime itself, in a manner so that mass and energy do not "travel through space" from the transmitter to the target at all. Instead, ripples and patterns in the fabric of spacetime itself are manipulated to meet and interfere in and at the local spacetime of some distant target. There interference of these ripple patterns creates the desired energetic effect (hence the term energetics) directly in and through the target itself, emerging from the very spacetime (vacuum) in which the target is imbedded at its distant location." Bearden Fer-de-Lance

The big Tesla howitzers are aimed at their targets by using a worldwide electromagnetic pattern called the "Woodpecker Grid," begun by the Russians in 1976.

You can hear the Woodpecker signal here: http://www.cheniere.org/misc/woodpeckersound.htm

The woodpecker grid is just that, a grid of EM waves which provide a channel for the LW waves to any point on earth. Computers use the woodpecker grid information to hone down the aiming of the howitzers to a pinpoint location on earth. (For that matter, it could be under the earth, in space, or under the sea).

A third howitzer mode is what Col. Bearden calls "Mindsnapper" mode. This is a truly frightening mode which affects the electromagnetic mind-body connection. Being mind-snapped at a low level would cause you to lose consciousness. Being mind-snapped at a slightly higher power would "entrain" all minds in the target area into a kind of hypnogogic trance, a state in which they would all be highly susceptible to suggestions and orders.

Being mind-snapped at high level would so disrupt the mind-body connection that everyone in the target area would fall down, instantly dead. Not only that, everything around you and inside you would be killed, so that your sterilized body would not rot for quite a long time.

The Mindsnapper can be used for a small area, or to kill an entire population. In fact, using the scalar waves to affect the human mind has become a new field in its own right, and has been dubbed "psychoenergetics." (This is discussed below in section 5).

Bearden describes the use of the howitzer in mind-snapping mode against people, to cause unconsciousness or death:

> "From a military viewpoint, these scalar EM weapons are really quite all-around weapons.
>
> "For example, they are very lethal against personnel.
>
> "Set in the "high intensity pulse mode," one blast and a person's nervous system is destroyed instantly. Total, instantaneous death results. One can blast away at entire groups of infantrymen, for example. It's the peak power that's important, not the average power. So the weapon doesn't require too much battery power.
>
> "Set on a lower intensity pulse mode, one zap can simply knock out a person, rendering him instantly unconscious. Quite useful if one needs prisoners to interrogate. Also very silent, which is useful in operations employing stealth and surprise." Bearden http://www.cheniere.org/books/ferdelance/s80.htm
>
> "Those hit by the scalar EM weapon, however, have a most peculiar death mode.
>
> "Death comes-instantly and totally. There is no convulsion, no response. The entire nervous system is destroyed instantly. Every living cell in the body is killed instantly, including all bacteria, germs, etc.
>
> "A body hit with this thing falls like a limp rag and lies where it falls. It doesn't decay in even 30-45 days. In a macabre fashion, it's been reduced to something like food irradiated with nuclear radiation; everything is killed, so the material is preserved for an extended period before any decay can set in."
>
> Bearden http://www.cheniere.org/books/ferdelance/s81.htm

He goes on to say that the same "bazooka-sized scalar EM pulse weapon" can also be used to knock out a tank with one shot (or pulse). So tank warfare is obsolete. Helicopters? Bearden points out: "... if he's attacked by a helicopter or a fixed-wing aircraft, he can bring the same weapon to bear on it. He can knock down a chopper or a jet. With one shot."

Obituary for Tank Warfare

Tanks useless against portable LW weapons

> "For example, suppose one attacks a tank. [with an LW weapon] With scalar EM pulse, the personnel die instantly. Total personnel kill is achieved.
>
> All electronic systems of the tank are dudded. Total systems kill is achieved. The ammunition in the tank explodes. Total kill is achieved.
>
> The fuel explodes. This is another total kill mechanism.
>
> That tank has been killed totally, by a variety of mechanisms, all simultaneously.
>
> That's a k-kill in any analyst's book.
>
> And don't worry about retrieval and repair. That tank is finished permanently.
>
> Even the most modern tank is just as vulnerable as the most obsolete."
>
> Bearden http://www.cheniere.org/books/ferdelance/s88.htm

One begins to get the idea how profoundly the conventional means of warfare have been obsoleted. All the usual weapons of war have been rendered nearly useless! And Star Trek's "Phaser," with its "stun" and "kill" modes is here at last.

Although it is the weapons aspect of scalar electromagnetic which has received the most developmental effort, Bearden points out that a great many other amazing things that may be accomplished in this new field of science. Beam me up Scotty!

> "Superluminal communications systems, hyperspace drive, and materialization and dematerialization are all hypothetically possible, using scalar electromagnetics. As the technology develops, we should see the development of many of the systems long thought impossible except in science- fiction."
>
> Bearden http://www.cheniere.org/books/ferdelance/s15.htm

Bearden describes a huge test of the Russian scalar weapons systems which occurred in 1985, a test that went completely unnoticed by western scientists:

> "Just before May 1, 1985 the Soviet Union performed a "full-up" operational test of their entire strategic scalar EM weapons complex. This test was detected and monitored periodically by Frank Golden. Some 27 of these giant Soviet "power taps" were locked into the molten core of the earth, producing forced (entrained) scalar resonance of the entire planet on 54 controlled frequencies under our very feet. The remainder of the scalar frequency spectrum was ablaze with literally hundreds of Soviet scalar transmitters: probably the entire strategic scalar command and control system to

> underwater submarines, higher command centers, distant commands, etc. was activated in the giant exercise.
>
> "For several days, the system was exercised on a mind-boggling scale, apparently as part of the Soviet Union's highly accentuated 40th anniversary celebration of the end of World War II.
>
> "Ironically, not a single U.S. intelligence agency, laboratory, or scientist detected this monstrous exercise for the new Soviet leader, Gorbachev. Not a one of them had a detector for scalar EM radiation, and not a one of them officially believes the exercise ever happened."
>
> Bearden http://www.cheniere.org/books/ferdelance/s38.htm
>
> "The Woodpecker grid/howitzer weapon system can be placed over the ocean and used against cruise missiles, naval surface-to-air and surface-to-surface missiles, submarine-launched ballistic and cruise missiles, etc. Placed over a carrier task force, it can also take care of the aircraft launched by the carrier as fast as they are launched. It can also handily take care of the missiles launched by guided-missile cruisers of the accompanying task force."
>
> (Bearden http://www.cheniere.org/books/ferdelance/s65.htm)

The whole series of the Fer-de-Lance briefing slides and comments is a "must read" for anyone wanting to be aware of the actual state of things in this world right now. It is a new world and it takes some getting used to. Fer-de-Lance speaks clearly, for the layman, of the way things stand.

What is it like to operate these Tesla Howitzers? Col. Bearden gives us an idea in his "Fer-de-Lance" briefing, Slides 66-71.

Who needs armies anymore? A entire war can be fought and won by a handful of men sitting in a comfortable control booth! The scalar electromagnetics revolution has made all previous forms of warfare nearly obsolete.

Why haven't the Russians used the scalar weapons yet?

Bearden gives his thoughts on this in "Fer-de-Lance." The power of these weapons is so great that an accident or war could wipe out all life on earth, and even affect the sun.

> "As can be appreciated, this entire concept [MAD] is now totally obsoleted by the Soviet possession of scalar EM weapons.
>
> 'The "mutual assured destruction" capability -- or MAD concept as it was so aptly named! -- is no longer mutual. And for two reasons: (1) the Soviet surprise attack could destroy essentially all our Triad forces wherever they are, in the first attack. (2) Even if a very small number of our strategic forces survive, the puny strike we could attempt to launch could be repeatedly decimated from launch to terminal phase.

> 'The restraint on the Soviets is not because of our Triad forces. The restraint is twofold: (1) a second nation has scalar EM weapons, and would undoubtedly unleash them upon the Soviet Union with great violence if such a massive world-wide destruction of U.S. forces occurred or was threatened (is it really "accidental" that six major Soviet missile ammunition storage sites mysteriously exploded within seven months?); (2) the closed-loop scalar radiation exchange system of Earth-Sun could be inadvertently "tweaked" in the feedback loop from Earth to Sun, so that a large solar scalar resonance was stimulated. In that case the Sun could emit a mighty resonant "burp" that would simply wipe out all life on Earth in a fiery shower.
>
> 'The "terror" in the "balance of terror" between nations has just increased by many, many orders of magnitude.
>
> 'The biblical prophecy that the Earth will be wiped out by "fire and brimstone from heaven" is very near fulfillment now.
>
> 'Put another way, the pucker factor -- even on the Russians who may contemplate unleashing Armageddon -- is higher than any Western strategic analyst has ever imagined, even in his wildest nightmares.
>
> 'Truly these weapons are "more frightening than the mind of man has imagined," as Brezhnev put it in 1975.
>
> 'The frightening scalar EM weapons can be used, but only very, very gingerly indeed. If a slight mistake is made, everybody loses everything."
>
> Bearden http://www.cheniere.org/books/ferdelance/s74.htm

But the Russians (KGB) are not the worst of it. Bearden claims that some of the scalar weapons have been leased to the Yakuza for about a billion a year, and that only the KGB is holding the Yakuza in check from making a full-scale all out scalar attack on the west.

U.S. Defense Secretary William Cohen

Warns about eco-terrorism using scalar electromagnetic weapons

> "Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our [counter terrorism] efforts." - Defense Secretary William Cohen, 1997
>
> In a letter to a writer named "Russell" Bearden says,
>
> "In short, Russell, the Secretary of Defense of the United States confirmed that there are indeed novel kinds of EM weapons, right now and have been for some time, which have been and are being used to (1) initiate earthquakes, (2) engineer the weather and climate,

> and (3) initiate the eruption of volcanoes. We wrote about those exact uses of the weaponry decades ago. Several nations now have such weapons. Three of them (two on one side and the other on a hostile side) are even firing practice shots into Western Australia, as a convenient test range."
>
> http://www.earthchangestv.com/ufo/0209gandor.htm

Tesla Howitzers in Weather mode

How to Make Weather

> "Engineering the weather is duck soup; they tested that over the U.S. in 1967, and entered upon continuing operations over North America on July 4, 1976 as a grim kind of KGB "Bicentennial Gift" to the United States. Here's how it is done with several interferometers.
>
> "First, the interferometers can deliberately make "high pressure areas" (cool the air so it shrinks and its "footprint" pressure on the ground increases because its density increases) and "low pressure areas" (heat the air so that expands and its "footprint" pressure on the ground decreases because its density decreases). Well, if one makes the highs and lows where one wishes them, and judiciously and somewhat slowly moves them along a given path, these highs and lows will entrain the jet streams and thus "steer" the weather.
>
> "So if you wish a very cold snap or attack, go up into Canada and start these actions to steer down some streams that bring some very frigid air. Establish other highs and lows judiciously to "block" or "slow" other jet streams and flows as desired. In this way, one can pull the "large cold air masses coming down from Canada" deep into the southern U.S. at will. They do it regularly. If you wish an ice storm, add an additional current of warm moist air you bring up (steer up by using artificial highs and lows) from the ocean. Where they meet, you will get freezing rain, then sleet, then severe icing accumulating. This type of attack does rather large damage to the struck area."
>
> http://www.earthchangestv.com/ufo/0209gandor.htm

Whole websites have now sprung up showing weather radar anomalies which may well be indications of scalar weather engineering. One list of over 500 images is here. Cheniere has its own collection here.

In another section at Cheniere are many photos of cloud anomalies which reveal the underlying grid along which the clouds are forming or un-forming. Bearden relates a startling incident in these clandestine ongoing "weather wars."

> "During the spring of 1986, abnormally strong Soviet weather engineering occurred over the U. S., causing a drastic drought in the southeastern U.S. This drought was broken by a colleague who used an extremely powerful scalar EM device to redirect jetstreams. A most unusual and unique signature of the "blocking" against the Soviet scalar EM actions

resulted: Two huge circulations developed in the atmosphere, clearly showed as two adjacent giant "holes" in the swirling cloud cover over the middle and eastern U.S.

"Between these two giant holes, the cloud circulations formed a stream of clouds, moving to the south, looking very: much like a giant vertical "bar" of a huge "Y-shaped" cloud flow.* Several national weathermen commented on this unusual pattern, which had not been observed before. The pattern continued, day after day, as the blocking continued."

* On the weather maps.

http://www.cheniere.org/books/analysis/history.htm

Tesla Howitzers in Earthquake mode

How to Make an Earthquake

"Here's how you initiate a very large earthquake with such weapons. Take a convenient fault zone of set of them. Focus the interferometry on the fault zone, in the "diverging" mode, and deposit EM energy there in the rocks on both sides, increasing (slowly) the stress in the rocks by the reverse piezoelectric effect (deposit excess energy, get crystal mechanical movements).

"Do it slowly, and the stress will build up to large pressures well-above a plate slip minimum energy required. At some point, the rocks yield and one or both sides "slip" and move rather sharply, giving a very large earthquake in that zone.

"Do the same thing down in the earth (remember, LWs easily penetrate right through the earth and ocean at will, and so the "interference zone" focus can be inside the earth or beneath the ocean, at will.

"Anyway, focus this thing down to where the active part of the volcano is still slumbering, down where the hole in the plate has been made. Keep increasing the deposition of energy in the magma itself, and eventually the increasing pressure from deep within that volcano, underground, will cause an eruption. Build the energy slow, and the eruption will likely be much larger."

http://www.earthchangestv.com/ufo/0209gandor.htm

Yugoslav Earthquake, Followed by Cold Explosion

Strange Characteristics of Iran Quake, Sep. 1978

Tesla Howitzers in Tesla's "Big Eye" mode

How to make a Big Eye in the Sky

By using longitudinal wave interferometry in a weakly endothermic mode a new kind of "radar" is now possible, one which can look right through the earth and oceans. Because the endothermic mode extracts energy from the distant target, a pattern of the target area can be scanned for in the returned energy.

> "Thus the receiver produces a representation of the energy extracted from various locations within the distant endothermic zone. By displaying the received signals on an appropriately scanned screen, a representation of the distant scene in the intersection zone can be created. Indeed this is a special kind of "microwave interferometry," and -- with modern techniques -- the imagery obtained might be surprisingly good. With development, it might even become as good as the image presently obtained by side-looking acquisition radars.
>
> "Interestingly enough, since scalar EM beams will easily penetrate the earth or the ocean, one can also look beneath the earth or beneath the ocean with this type of scanning scalar interferometer.
>
> "The importance of this capability to strategic and tactical reconnaissance is obvious. Camouflage, cover, and concealment have no effect on such a system. One can easily look inside buildings and into underground facilities. With a small system such as this, the U.S. Marines at Khe Sahn would have had little difficulty locating the tunnels continually dug under the perimeter by the Viet Cong. And targets under jungle canopies are directly visible."
>
> Bearden http://www.cheniere.org/books/ferdelance/s42.htm

Tesla Howitzers can destroy all life on earth

How to Make a Storm on the Sun!!

(Just make a big mistake with your Tesla howitzers)

The deepest dangers of the use of scalar weapons are frighteningly pointed out by Bearden here where he considers that the sun, earth and moon (actually the whole solar system) is a delicately balanced arrangement of scalar (longitudinal) EM radiation. There is a scalar connection between the earth and the sun, which are both giant dipoles gushing energy in the time domain. Big scalar events on the earth could alter this balance causing true catastrophe.

"Indeed, a solar response could be stimulated so that the Sun would violently belch and destroy our biosphere, among other effects."

Unrestrained scalar warfare could cause huge solar storms? To realize that human activity could actually affect the sun itself is truly mind boggling.

Bearden goes on to say:

> "Accordingly, use of huge scalar EM weapons is a double-edged sword. Unless carefully employed, use of the weapons could cause a terrible backlash to the user as well as the victim, and even accidentally cause the destruction of the earth itself. It is not accidental

that in 1960 Khrushchev stated that his new fantastic weapon could -- if unrestrainedly used -- destroy all life on earth.

"This appalling backlash potential is apparently what prompted Brezhnev in 1975 to make so strong an effort to get agreement to outlaw the development of "new weapons of mass destruction" more frightful than the mind of man had even dreamed of. Gromyko even introduced such a draft agreement into the proceedings of the United Nations. The first article of the draft agreement indicated that the nature of the weapons referred to would be negotiated. Sadly, the entire Western world did not even know what the Russians were talking about."

http://www.cheniere.org/books/ferdelance/s24.htm

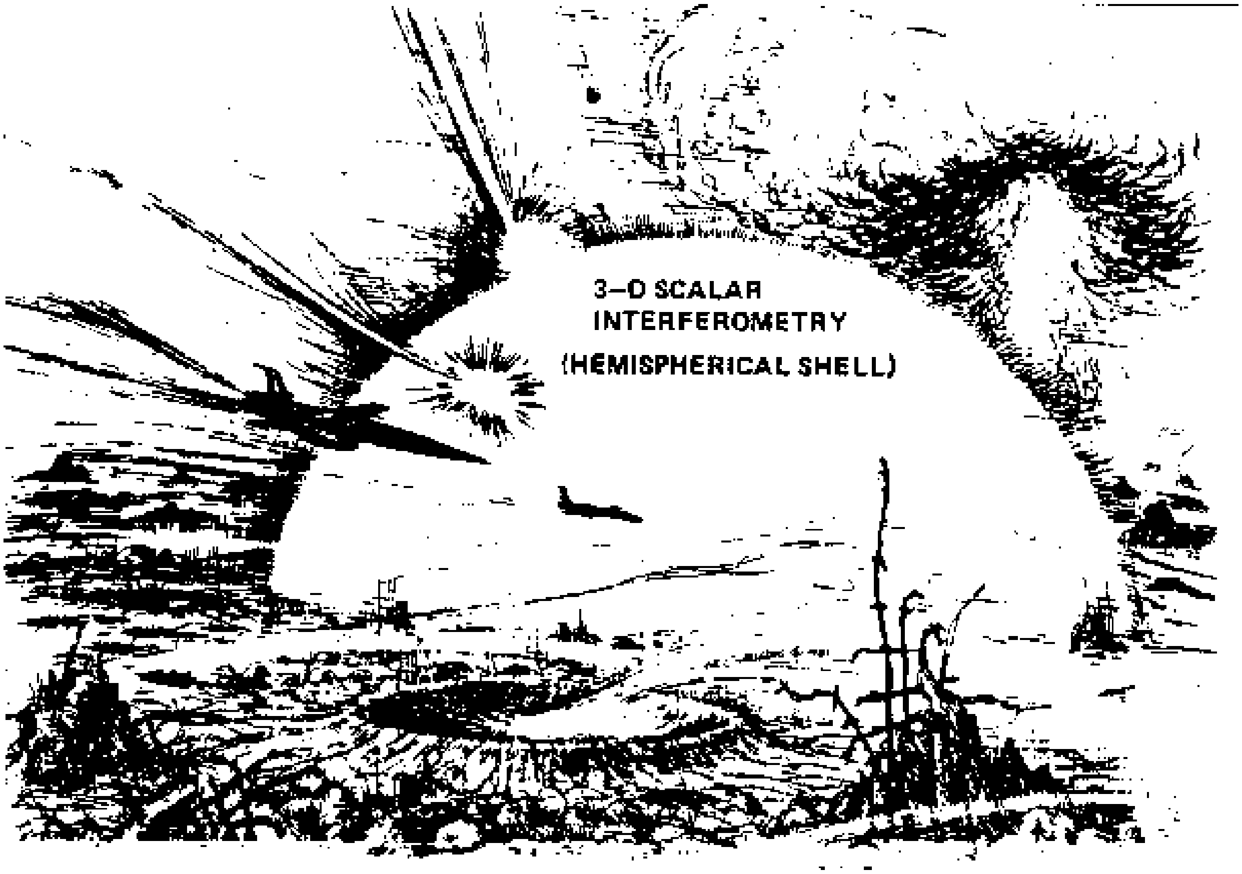

3b. Tesla domes provide near-perfect defense shield.

(Slide from Cheniere.org)

The Tesla howitzers can be used in a nearly impregnable defensive mode whereby they throw up a dome (**Tesla Dome**) or a sphere (**Tesla Globe**) of highly powerful electromagnetic energy, enough to "dud" or destroy missiles which try to penetrate them. Once again, this energy is not going through space from the howitzer, but being made to emerge from the local vacuum at the location of the shell. KGB tests of these domes have been witnessed by airline pilots around the world. Bearden gives many examples of the Tesla domes being sighted around the world in his briefing paper "Fer-de-Lance." (Russian dome test, another globe incident, yet another globe test). These giant electromagnetic domes can be hundreds of miles across, or narrowed down more to total impermeability.

> "Such a shell may be several hundred miles in diameter at the base. The enormous energy required to form such a defense shell is obtained by a "scalar power tap" into the molten core of the earth itself, as previously explained. In late April/early May of 1985, 27 such "power taps" were placed in the earth by the Soviets. If each tap is capable of powering four to six large scalar EM weapons, then the Soviet strategic scalar EM arsenal contains over 100 monstrous superweapons capable of generating exothermic explosions, endothermic explosions, engineering the weather, locating and destroying underwater submarines, detecting and destroying ballistic missiles shortly after launch, detecting and destroying long range strategic bombers as soon as they are airborne, etc."
>
> http://www.cheniere.org/books/ferdelance/s52.htm

By using nested domes one can protect the domed area even against nuclear radiation itself. Tesla globes can be used to hit airborne targets by simply placing a globe of any chosen size in the flight path of the incoming missile. It does double duty because the missile hits it going in, and then the rubble hits it again going out the other side.

While the U.S. persists in its useless boondoggle "missile defense system" by shooting a missile with a missile (missiles which would already have been "dudded" by the Tesla howitzers), between the domes and the globes the Russians (KGB) have in hand (operational) a nearly perfect missile defense system. Someone should ask the Secretary of Defense if the U.S. Tesla domes are in place. And over which cities? How powerful are our American Tesla domes?

Bearden cites a number of examples of these domes being tested around the world:

March 20, 1969 | March 24, 1977 | August 17, 1980 | June 17, 1966 | June 18, 1982 | June 22, 1976

In a slide from the Fer-de-Lance briefing it is shown how the "woodpecker grid" itself also acts as a sensor, detecting any missile launch anywhere, and how the howitzers can respond instantly by delivering a blast to exactly the right coordinates, right along the channels of that same grid which is feeding the location information about the missile launch. This certainly beats "trying to hit a bullet with a bullet." Especially since YOUR bullet would probably be instantly inoperable in a real scalar war situation.

3c. Rogue Groups in the government

Bearden fears that these scalar electromagnetic secrets may have fallen into the hands of what he called secret "rogue groups" within the labyrinthine depths of the black-ops special-access domain of the secret government. "In the extreme, such a deep black program can even become a "captured" program," he says, "which is totally in the hands of rogues and no longer reports to or is bound by the dictates of the parent government."

> "They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification. ...
>
> "In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old 'Power corrupts, and absolute power corrupts absolutely' routine. ... Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control."
>
> "The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. "
>
> http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

3d. World War III has already begun

Because it is now possible to cripple the power grids and electronics of any nation, that nation could be reduced to economic ruin without ever firing an actual "shot." Without electricity Wall Street cannot operate. Without electricity you quickly need martial law. One no longer needs to wage a conventional war anymore at all. There are now lots of new ways to bring any nation to its knees.

> "We are at war, whether we like it or not. And we are going to have strategic strikes on an unparalleled scale for us. Pearl Harbor was nothing compared to what is coming down, once we consider the "augmentation" and such.
>
> "When that factor is considered, then it's a totally new ball game on replacing the power system. There isn't going to be much choice; the old one is going to be destroyed or so badly crippled that it will collapse the U.S. economy in a way we have not seen since 1929. Of course, in such a case hitting us, we will be under martial law, the works, everywhere. For one thing, it's the only way you can keep the cities from imploding from within, by the criminal element, looting, robbing, killing, etc. Simply wait and see; it's on the way."

> Although Bearden has for years decried the "scalar-gap," and chided the military for falling so far behind the Russian KGB in the field of scalar electromagnetics, recent statements indicate that things may at last have changed somewhat.
>
> "In the new war, the first phase of WW III has already been completed. Hence President Bush's recognition that this is a total war, and has to be taken to the enemy no matter where he may be hiding or in what sanctuary or in what foreign nation. This is an entirely different war than anything the U.S. has ever fought in its history. And before it is over, it is going to kill more Americans than have died in all the previous wars in our history. The news media and the public have not yet recognized that awful situation and predicament. Most are already thinking that, well, we won in Afghanistan so it's just about over. Instead, it has only just begun.
>
> "Anyway, that's a brief uptake on some of the weapon situation today. Some 10 nations or more now have LWIs, and five have the QPs. I'm still working on how many have the appalling negative EMP weapons, but would estimate 6 or 7 nations have them in one stage or the other, either deployed or in at least advanced development. And in my estimation, the Yakuza will have them within three to five years if they have to build them themselves, and quicker if they can buy them from the Russians, Chinese, or???"
>
> Bearden http://www.cheniere.org/correspondence/030202a.htm

When he was asked the question '**Does the government listen to Tom Bearden?**" Bearden said:

> "Let me speak carefully. I am very satisfied with how I've been listened to in the quarters where the rubber meets the road. The West is no longer defenseless, not by any means. Cannot amplify it further than that."
>
> http://www.cheniere.org/correspondence/030202a.htm

Bearden has briefed the government a number of times and the Cheniere website contains a great number of his graphic military briefing slides.

3e. Are the Russians making "UFOs?"

Scalar electromagnetics engineering opens the possibility of antigravity vehicles and "hyperspace travel." In other words we can now make our own UFO's! Of such scalar-powered antigravity vehicles, which have now become possible, Bearden has this to say:

> "We might expect to see such vehicles glowing. Their various surface features and mechanisms might appear to be glowing or revolving lights, etc.
>
> "They could exhibit incredible "aerodynamic performance," seemingly in the atmosphere. Actually they would not be moving "through" the atmosphere at all, but through a higher space outside each particle of atmosphere.

> "They could seem to materialize and dematerialize.
>
> "They could seem to plunge into the ocean or rise out of it.
>
> "They could even seem to operate under the ocean or inside the earth itself.
>
> "Such anomalous vehicle performances have been seen all over the world, particularly since a few years after WWII.
>
> "It appears that one or more nations of Earth are operating such vehicles now in great secrecy. Our own "government" may even be one of these.
>
> "If so, it isn't our true "government/government." Instead, it's the "control group/government." It's government at the operational level, but at higher level it belongs to certain control groups who have penetrated our government's vitals and taken over all such projects."
>
> Bearden http://www.cheniere.org/books/ferdelance/slide32.htm

In an excellent article on her Earthfiles website, Linda Moulton Howe reports on the utter bafflement of Norwegian scientists over mysterious lights appearing in the valley of Hessdalen, Norway.

She writes of the research there: "The results can be broken down into two groups: 95% are thermal plasmas and 5% are unidentified solid objects. The plasmas emit long wave radio frequencies and strangely, their temperatures do not vary with change in size or brightness." She quotes the scientists' research summary:

> "1) most of the luminous phenomenon is a thermal plasma;
>
> 2) the light-balls are not single objects but are constituted of many small components which are vibrating around a common barycenter;
>
> 3) the light-balls are able to eject smaller light-balls;
>
> 4) the light-balls change shape all the time;
>
> 5) the luminosity increase of the light balls is due to the increase of the radiating area. But the cause, and the physical mechanism with which radiation is emitted, is currently unknown."

This would almost certainly seem to be another example of the testing of the scalar howitzers. It is baffling because, as Bearden points out repeatedly, Western scientists have limited knowledge of the scalar electromagnetic principles by which the howitzers operate. The Hessdalen scientists were completely baffled as to where the energy of these plasma balls was coming from.

Ms. Howe interviews Massimo Teodorani, Ph.D., Astrophysicist, who released the report on the 2001 Hessdalen sightings. The scientist speaks of the baffling phenomenon observed.

> "I don't know how it is possible that Nature is spontaneously able to do that. Anyway, we deduce that the plasma is trapped inside a sort of magnetic cage and the magnetic cage closes around the plasma and keeps it fixed in some way, prevents it from expanding. But where does it come from? We don't know."

By what we have been considering here, we can recognize the signatures of scalar electromagnetic engineering. If it is a Tesla howitzer, then the energy for the plasma balls is coming from the vacuum of spacetime at the very location of the balls themselves, triggered by scalar interferometers aimed through the woodpecker grid.

These kind of balls can be used as marker beacons giving feedback for precision aiming of the howitzers. The energy of the marker beacon can be read back into the computers giving precise location information for pinpoint aiming. The target area can be very small, or widened out.

There is a little movie of one of the Hessdalen plasma balls here. It would seem to be a clear documentation of a scalar howitzer marker-beacon. How many other current "UFO" sightings are actually the witnessing of Tesla howitzer tests?

In fact there are a great many anomalous events of the past few decades which might very well be explained by the Russian testing of their longitudinal electromagnetic weapons. It would seem that the mysterious appearance of crop circles around the world could be accomplished by feeding a precise mathematical graphic pattern into the computerized aiming software and change the very molecular structure of the crops themselves along the lines of that pattern. The

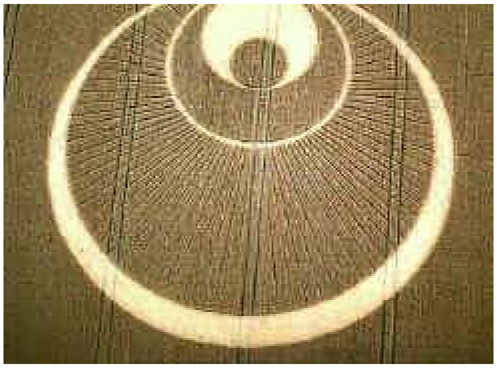

stalks fall over from the localized effect of the longitudinal waves. A plasma ball tracing patterns? Are they slyly sending a message to see if we are "getting it" yet?

The "Cambridge Angel" - Made by Russian scalar plasma balls?

Also see: Interview with Tom Bearden

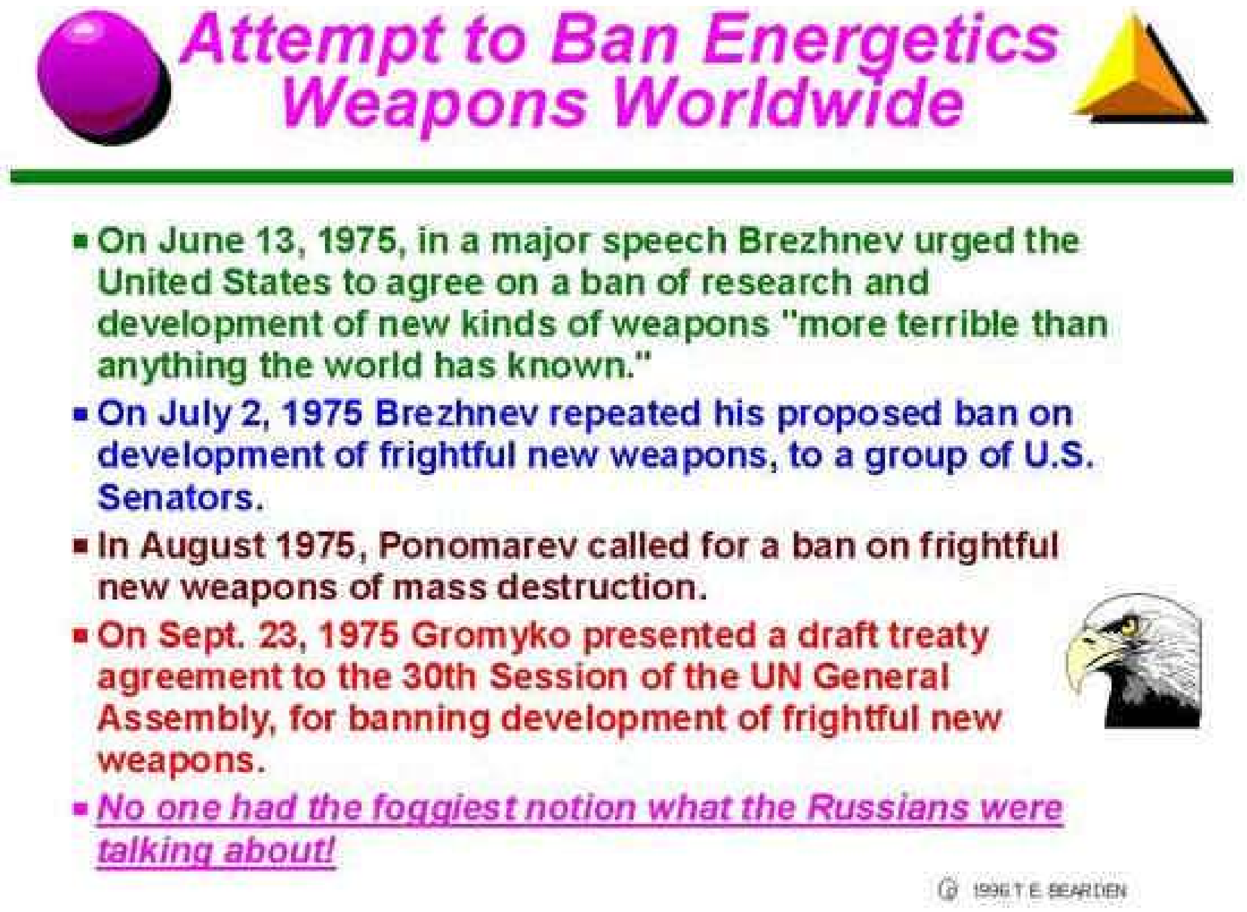

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T E BEARDEN

(Slide from Cheniere.org)

4. Healing powers of longitudinal waves

Healing via longitudinal waves, time reversal of cells (de-differentiation), Priore's machine, Bearden's proposed "blanket" version, using the electric power grid for healing waves or for broadcasting disease patterns. Scalar Biowar.

The good news!

> "Any sort of disease whatsoever - physical or mental - will be rapidly curable, simply and cheaply. It will be possible to reverse aging and rejuvenate the person. It will be possible to regrow limbs and straighten misshapen spines - and directly remove the causes of mental diseases and cure them as well. Lifespan will be staggeringly increased, without 'old age's debilitation.' AIDS, cancer, leukemia, and genetic diseases will be completely conquered. The potential for a true golden era is upon us, for all humanity."
>
> Bearden- http://twm.co.nz/beard_interview.htm

Proof of the healing powers of longitudinal waves from the vacuum is seen in the work of Antoine Priore who produced longitudinal scalar waves using a gigantic glass tube filled with plasma. He himself did not understand why it worked or how it cured cancer in animals.

> "The Priore team cured terminal tumors, trypanosomias, and other dread diseases in laboratory animals in France, in the 1960s and 1970s, under rigorous scientific protocols by eminent scientists. The Priore project was funded by the French Government to the tune of several millions of dollars (equivalent).
> http://www.explorepub.com/articles/energetics.html
>
> ". . . Prioré's machines concretely demonstrated a nearly 100% cure of all kinds of terminal cancers and leukemias, in thousands of rigorous laboratory tests with animals. These results were shown to medical scientists as early as 1960." http://www.cheniere.org/priore/index.html

Bearden has stated that the longitudinal EM waves can have a healing effect because they always come in pairs called "phase conjugate pairs" and one of the pair is time-reversed. This enables engineering of the waves to "pump" the cells of the body (with waves from the time domain) and actually time-reverse the cells back to a previous healthy condition. The scalar cure of cancer does not involve killing the cancer cells but rather pumping them with time energy into a previous condition. I know, this seems miraculous, and is it not?

Priore and his giant plasma tube

This healing power in the longitudinal waves of the time domain is a mind-boggling thing to face. It took me months to get through it, and to suffer the fact that the cure for cancer and AIDS was right there, in this new scalar electromagnetics technology, and that it has been kept back from the people for so many decades now.

Bearden has a great deal to say about the work of Priore, who himself did not understand why his machine worked. Priore did not know about the longitudinal scalar waves of the time domain. In fact, Tom Bearden has supplied, at last, the missing explanation as to how that big giant two-story glass tube consistently cured cancer in animals.

"**The medical implications are enormous**. By just reversing the damaged cells back to a previous physical state, this gives a physics mechanism for controlled cellular dedifferentiation, in biology terms. Physicians will time-reverse diseased, damaged or aged cells back to an earlier healthy condition, including all the cellular genetics. Normal cells are just reversed to a slightly younger condition."

> "We believe this mechanism explains the revolutionary but baffling electromagnetic cures achieved by a team of outstanding French scientists in the 1960s and early 70s, working with inventor Antoine Prioré. In lab animals, cures were achieved for terminal tumors, infectious diseases such as trypanosomiasis, and atheriosclerosis. In addition, deliberately suppressed immune systems were dramatically restored.
>
> "We hypothesize that this is the long-sought mechanism of how the living body heals itself. In short, the body's cellular control system induces cellular reversal by weak time-domain pumping of the damaged cells and the affected area. The cellular regeneration system performs all healing of cellular damage in the body; the immune system heals nothing, not even its own damaged cells) .
>
> "For the first time medical science can develop as a science of healing rather than a science of intervention. And it will develop as a science of unlimited healing, since no pathogen can resist action by curved spacetime engines." Bearden
>
> http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

See Priore's machine See Slides of how it works.

Bearden has extended the Priore work, with the previously missing knowledge now in place, and designed a revised system (for irradiating the body with longitudinal waves) which gets rid of Priore's giant glass plasma tube and uses instead an antenna unit which surrounds the body. He has also provided the missing explanation of how it works. There is an article from "Explore" magazine, Part I and Part II.

In fact, Bearden has presented to the U.S. government a plan to use a portable version of a Scalar Healing Blanket where wire channels in the blanket act as a scalar antenna, which can both send and receive scalar electromagnetic waves. A two minute treatment of the waves is sufficient to restore the body to its earlier healthy condition. Thus even thousands could be treated easily and quickly in a biowar mass attack situation.

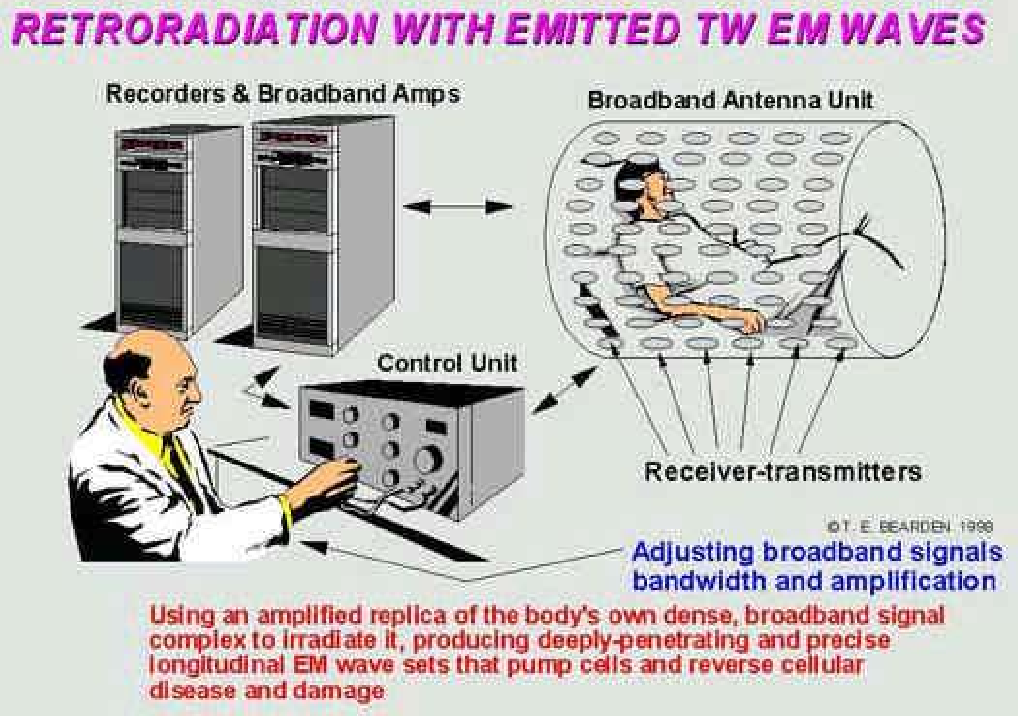
RETRORADIATION WITH EMITTED TW EM WAVES
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
© T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification
Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

"To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S. Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines could be developed cheaply and quickly, and present government facilities could be used to massively attack the correlate database production problem. "The portable treatment units could be cranked out by the hundreds of thousands and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals,

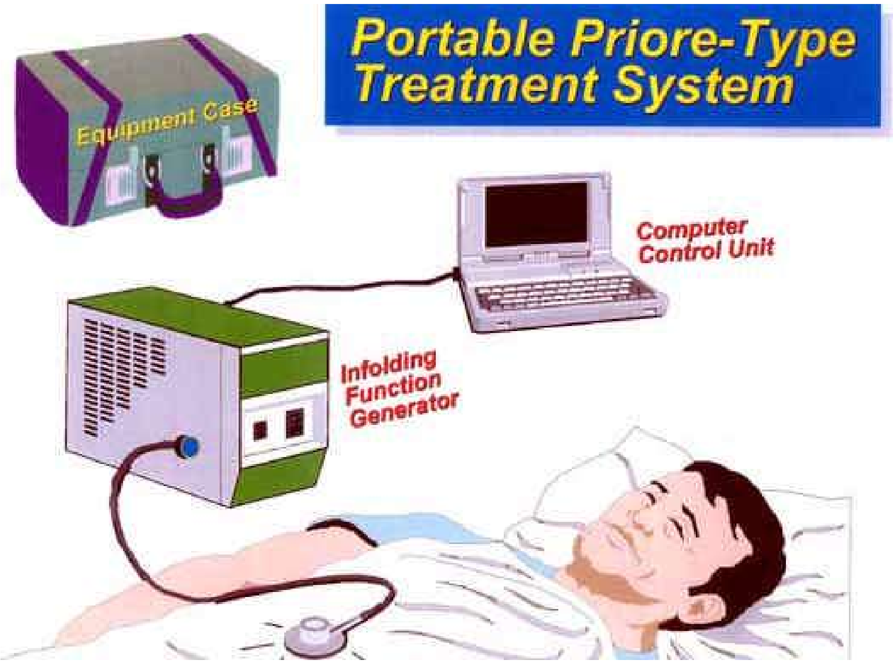
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator

> emergency response teams, etc."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p04.htm

The Scalar Healing Blanket would fit into large suitcase. It has three basic parts, the longitudinal wave generator, a laptop computer, and the antenna-blanket. The plan was a stroke of Bearden's genius, but unfortunately it fell on befuddled government ears. "Unfortunately, so far the DOD [Dept. of Defense] doesn't even understand what the heck we're talking about!"

Article "**Healing with Energy**": http://www.explorepub.com/articles/energetics.html

> "We tried very hard in 1998 to call attention to the extreme national need for the most urgent possible development of a portable computer-controlled treatment device based on the experimentally proven Prioré methodology, but extended to a much simpler and far more rapid method of treatment.
>
> "Quick development of this technology could lead to a suitcase size device, capable of treating a patient in less than one minute, and stopping the symptoms and disease progress cold, then reversing it. Three "less than a minute" treatments one week apart would be required for complete cure. As an added advantage, partial reversal of aging in older patients would also be accomplished as a bonus, in addition to time-reversing the damaged and diseased cells...
>
> "The method proposed to the DoD in 1998 used normal EM in a peculiar way to force the body itself to make its own time-polarized EM waves, and pump every part of its cells -- including the genetics -- in the time domain.
>
> "Quite simply, we are going to be struck this way (QP spreading the immune systems of the populace plus professional smallpox, anthrax, etc. BW attacks on our population centers) and by other means we've spoken of in other papers and briefings."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

The Dark Side of the Force:
Creating and broadcasting disease

Speaking of the scalar wave induction of disease by longitudinal wave patterns Bearden refers to work by French scientist Kervran Russian scientist Kaznacheyev, who proved that any disease could be transmitted electromagnetically.

> "It works this way: the Kaznacheyev experiments (and others) demonstrated that any cellular disease or disorder can be initiated into cells at a distance, by electromagnetic means (albeit unusual EM in nature). The decades of microwave radiation of the U.S. Embassy in Moscow, was an ongoing set of "stimuli" to ping the U.S. system and, by its response, ascertain how far along (or if it was "along") in its knowledge of such weapons. The radiation was responsible for many health changes in personnel and for the eventual deaths of three U.S. Ambassadors."

> Bearden http://www.cheniere.org/correspondence/021702a.htm

In a letter he writes,

> "Yes, the Kaznacheyev experiments were quite real; 15,000 or so in military research institutes in Siberia. As with so many other novel areas investigated and developed by the Soviets, most U.S. investigators just assumed normal EM, which is not what the Kaznacheyev experiments involved, and not what was carrying disease induction patterns from one cell culture to another in rigorous lab tests."

How to "Broadcast" Diseases

Longitudinal EM Biowars

Bearden explains how the new quantum potential weapons can be used to induce disease-at-a-distance in a population, or to "spread" the immune system so thin that a conventional bio-attack would be greatly potentiated. By "broadcasting" virtual disease patterns over a population, the immune system is so overloaded by the presence of the "shadow" diseases that its effectiveness is thinned out.

> "In short, alter the internal wavestructure, and one creates a curved spacetime 'engine' that acts on mass in any fashion one chooses to design -- including initiate diseases . . . Now visualize one of these 'vacuum engines' or 'spacetime curvature engines' that acts on mass to generate the exact effects produced by anthrax. (Or any other disease one wishes) . . . Call the virtual state of a disease engine the disease pattern in the "shadow" state, just prior to becoming observable. . . .
>
> "And the master cellular control system responds to "coming events that cast their shadows before". In short, it responds to the upper level of the "shadow state" of a disease, still in the virtual state! (Hey, virtual particles are real, virtual photons are real, virtual ST curvatures are real --- just very fleeting, only for another to immediately arise). The exchange of virtual particles is known to generate all forces in physics anyway! . . .
>
> "So the cellular control system responds to "shadow state" disease patterns. Enter a QP weapon application. Place a quantum potential weapon "area" on the United States, so that the entire populace is in it. The slowly introduce and bring up in the shadow state, the necrotizing fasciitis disease engine. At some point, the cellular control systems will react, and order the immune system into action. In other words, immune system resources will be committed against this "immediately coming" enemy. Bring it on up a little, but still just in the shadow state. Voila! Humans form a bell-shaped distribution curve. And even if the structure of the QP fluctuates a little, fluctuating a bit stronger and a bit weaker, etc. A few of those targeted bodies will now actually develop necrotizing fasciitis. So far, just like the U.S. Embassy targeting, except a different "carrier".
>
> "But now there is an exact signature that this is what is doing it, and not normal vector carriers of disease and contact or exposure infection. The few cases of necrotizing fasciitis that break out will be most puzzling. They will be randomly scattered across the

> entire nation, and so few that the vector carrier and contact or exposure explanation completely fails."
>
> "And that one happened, right here in the U.S., not so long ago. Check it out.
>
> "So carefully adjust the patterns for, say, 12 major terrible diseases, to that "adjusted and desired shadow state level". Now add all 12 patterns into the internal structure of the quantum potential. Now it has 12 shadow state disease engines in it.
>
> "Place that one on a populace. The cellular control systems order the immune systems into action, and they now have to commit their finite resources to 12 different areas. This greatly thins the response an immune system can make to any one of those 12 disease patterns! In short, it "spreads" the capabilities of the immune system, which can be galvanized to great effort, but only can perform a small bit against each area.
>
> "Suppose one of those "galvanized" areas in the targeted populace's immune systems is anthrax. And suppose the bad guys now hit one or more of those cities with a professional anthrax attack. A former OTA study showed that a professional anthrax aerial spray attack against Washington D.C., dispensing 100 kilograms of anthrax spores in the spray, will account for 1 to 3 million casualties. And that's in an "unthinned" set of immune systems. In the "thinned" set, one can at least double or triple, and even up to five times, that estimate. So conservatively, now there would be from 2 to 6 million casualties. And even mild anthrax spray attacks in other "thinned" cities would also be amplified.
>
> "That one is being set up, or already set up, right now. There are other such "augmentation" setups also under way.
>
> "So the weapons alluded to by the SecDef in 1997 are already being employed. WW III has already begun, and our populace (and our scientific community) is blissfully unaware of it."
>
> http://www.cheniere.org/correspondence/021702a.htm

Most of us can remember the strange outbreaks of the so-called "flesh eating disease." It seemed to break out here and there randomly with no known cause. Bearden believes this was a test of the newer quantum potential weapons which go even beyond the capabilities of the Tesla howitzers.

> "The Kaznacheyev "disease induction by novel electrodynamics" work is also the basis for the spacetime curvature engines (not EM signals!) carried by quantum potential EM biological weapons for distant induction of diseases etc. into the population of an entire nation or area, such as the U.S.
>
> "This quantum potential disease induction weapon -- capable of attacking an entire population of a nation or area -- has been tested several times in the U.S. at very low levels, in the shadow state only (we explain shortly) and it has a most peculiar signature. E.g., one test used necrotizing fasciitis (flesh-rotting disease) engines.

> "The disease 'engine' was deliberately kept down in the 'shadow state' (virtual state, just below the observable state) so that only a tiny bit of the populace with depressed thresholds would "breach the observable threshold' due to the population forming a bell-shaped curve as to the threshold level. The major signature of the test was that the cases of the disease broke out totally statistically and randomly, spread all over, without any 'disease vector' in between.
>
> "In short, it was not spread in any manner required by normal deterministic disease vectoring, but was a totally random set of occurrences. Several other diseases were also tested this way in the American populace, with precisely the same signature."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

"Apparently one planned use of such an augmented clandestine BW weapon will be smallpox, e.g. As you know, the Russians did develop smallpox for biological warfare, and even developed a new strain against which previous vaccination is ineffective.

> "The vast numbers of expected casualties, coupled with immune system spreading, are absolutely irresistible to the Russian mind. You can count on it that they have seen that the smallpox is already on site here, in the U.S. and in the hands of professional terrorist teams. Of course, the Russians/KGB will protest they have nothing to do with that! Yet likely the KGB will have flatly arranged it or at least assisted it. The capability for mass destruction of the nation is so lucrative and easy to achieve in advance, that they would never be able to resist doing it."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

Bearden believes the Gulf War Syndrome involved the use of scalar technology to induce a disease state, as he stated in a 1997 interview:

> "I'll say this flatly, and I won't answer any further questions about it; the Gulf War Syndrome was induced... It was induced with this stuff. All the rest... was contributing factors that everybody's talking about... There are some deeper signatures, if you look into it very deeply, that show you exactly how it was done and the fact that it was induced. It was a test. It was a test of a very special kind of weapon I have not talked about yet. Anyway you can use it to create diseases."
>
> Bearden http://twm.co.nz/beard_weapon.htm

Besides using the woodpecker grid to target a population, the entire electrical grids of nations can be used as "channels" for disease patterns to be broadcast to all people near that grid. Your own house wiring can become an antenna, broadcasting these "shadow" patterns of disease, weakening your immune system.

And here one faces the frustrating fact that the same power grids could be set to channel patterns which would keep the entire population healthy, rejuvenated, continuously bathing in the time-reversing old-age-defying energy of the spacetime vacuum. We have no idea yet of how long the human life span could be increased with this new technology. People might be able to live two centuries given full humane development of this technology.

5. Psychoenergetics

Mental entrainment, inducing hypnogogic trance, mind control, broadcasting thoughts and feelings, the Ant World, education via EM [LW] mental download, the "Feelies," revolution in spirituality, "enlightenment machine?" Scalar churches?

Psychoenergetics is simply using the scalar interferometers to manipulate and engineer the human psyche. To me it is the most mind-boggling and frightening aspect of longitudinal wave engineering.

The mind is electromagnetic in nature, and itself of the time-domain. It is not observable in 3-space. There is no thing you can point to that is the mind. As Bearden says**, "The mind is time-like**."

The longitudinal scalar EM waves are also of the time-domain. By overlaying certain infolded patterns (oscillations in time) on the longitudinal waves, a thought or feeling may be made to arise in a person or persons who are in the **interference zone** (where the wave beams cross).

The person will not notice anything, feeling that this thought or idea is his or her own idea. The longitudinal wave pattern might be a wave of panic or fear that spontaneously sweeps over you, perhaps inexplicably. It could even be in the form of a sudden intense patriotism. It could be explicit words and phrases, which everyone would think they had thought of by themselves. Or it could be rage and hatred, angry and violent. On the other hand It could also be a continuous sense of docility and placidity. This is the brave new world of psychoenergetics and the coming "Mind Wars."

Scalar Psychoenergetics in its most primitive form simply "entrains" all minds in the target area into a deep hypnogogic trance. In this state of mind people would be suddenly extremely suggestible, and would likely believe anything they are told, and would obey any orders given. Talk about winning the hearts and minds of the American People!

Tom Bearden makes me think of **Paul Revere**, who heroically conveyed important warnings for the sake of the new nation. But the words of the famous phrase have now strangely and curiously mutated into the once-inconceivable: '**One if by land, Two if by mind**."

Indeed one strategy in a Mind War would be to simply take over the minds of the "leaders" of the enemy nation. The targeted leader would not be aware that anything was amiss, although he might begin to make unexpected changes in policy.

And are, in fact, any of our "leaders" already having secret thoughts that are not their own? Who knows? Frighteningly, we can no longer be sure. The new sciences of scalar electromagnetics and psychoenergetics are even now plunging humanity into a sudden science-fiction-like world beyond anything in humanity's previous imaginings.

> "The Russians reached the point in the mid 1990s that they could take over control of a person's mind, with modified longitudinal EM waves including some time-polarized EM waves, and with a team of specialists (estimate 25-30 per transmitter, and one transmitter and team per controlled person)." Bearden http://www.cheniere.org/misc/time.htm

In a slide Bearden illustrates how a Scalar War would involve a psychoenergetics attack on the operators of the enemy scalar installation, entraining their minds into hypnogogic trance and getting them to shut down their systems.

And ultimately psychoenergetic warfare goes to the very heart of human identity itself. For if my thoughts might no longer be "mine," then who and what am I? Can my very sense of being "me" be hijacked by some nefarious psychoenergetic scheme? Will the "secret government" eventually dictate directly into your mind how you feel about yourself? Or what you think you are? Or what you should do?

How Does Longitudinal EM Mind Control Work?

The physics of scalar psychoenergetics as expounded by Col. Bearden seems even more arcane and over-my-head than that of extracting energy from the vacuum. I can't understand it, but I look through it anyway. Each time another tiny understanding might occur, or a term might fall into place. I know the mathematics is forever beyond me.

One paper to begin with is “Mind Control and EM Wave Polarization Transductions” This is such serious stuff that Bearden includes a strong warning about misuse of this knowledge. Psychoenergetics weapons can mentally maim and physically kill.

Warning!

(From "Mind Control and EM Wave Polarization Transductions")

> "This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p01.htm

After the serious warning comes the description of the paper. This is heady stuff for the layman.

> "Calling full attention to the special note above, in this paper we present a high-level overview of the novel electromagnetic nature of mind operations, mind and body coupling, and intent - the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transaction mechanisms whereby differing EM wave polarization's can be transformed one-into-the-other are presented."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p01.htm

I will leave it to the technically astute to try to understand the mechanisms of operation which are described in scientific language. But there are tidbits for the layman. Here he speaks of the difficulty Western scientists have in opening up to the "immaterial" (longitudinal wave) nature of the human mind.

Western Science Remains Largely Materialistic

> "Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system. This serious self-limitation exists because in the body we measure only weak TW [transverse wave] EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.
>
> "With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.
>
> "Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW [longitudinal wave] functions exist in the body and in all of nature in general."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p05.htm

Western methods of influencing the mind with EM waves have only used transverse waves, not the longitudinal waves of the vacuum. If only transverse waves are used one pretty well has to hit the target mind with a sledgehammer of waves.

Western Science knows only transverse EM waves

Brute Forcing Time Functions versus Fine Control Methods

> "Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines

> (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.
>
> "While KGB scientists also use TW EM "brute force" TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired... Bearden
>
> "These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design for example of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gail's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p09.htm

The seemingly magical operations of psychoenergetics, sometimes called "psychotronics," cannot be achieved with ordinary transverse EM wave energy. But they are possible with longitudinal scalar waves because the mind itself is scalar in nature.

> "We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain..."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p08.htm

The engineering of the mind can only be accomplished with the waves of time.

The course of psychoenergetics development necessitates the building of a vast database of settings and patterns which bring about given mental and/or emotional states. A kind of "gnome" of the psyche. Bearden describes the process of the decoding of this "gnome."

> "Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.
>
> "In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions."

> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p01.htm
>
> "A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p02.htm

Summing up his paper Bearden speculates on what the actual situation the science of psychoenergetics probably is. The intense secrecy with which this knowledge has been kept from the public has led to the current farce of the "government" pretending it all does not exist, and this secrecy may have now put us at great risk.

> "The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.
>
> "In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians..."
>
> "Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification."
>
> "The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U(l) electrodynamics. Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW [longitudinal wave] interferometers, through intervening mass."

> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

Bearden cites two examples where he believes the Russian KGB tested the psychoenergetics capabilities of their scalar interferometers. One was the case of Captain Button in his A-10 "Warthog" aircraft on April 2, 1997.

> "Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona."
>
> "At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—dream-thinking until he died that everything was normal. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour. "Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p06.htm

The paper "Mind Control and EM Wave Polarization Transductions" ends with the most chilling and frightening possibility of all: the psychoenergetic engineering of the human species as a whole.

The Ant World

The Final Race for Direct Mind Control of the Entire Human Species

> "The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs [longitudinal waves] for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials. "The KGB scientists also know that something very like Gaia — a collective unconscious mind for all species on earth — also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials.""Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.
>
> "I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can

> then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society" — communist style — is still a variation of feudalism (all systems tend to feudalism or some form thereof). "The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now...
>
> "Some versions of Russian mind control devices were used on Russian soldiers in the Afghan War to condition them for performance of merciless acts." 13
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p07.htm

More about "The Ant World"

Whether it is the Russians planning an "ant world" society or rogue U.S. black-ops groups withholding a great boon from human race, the secrecy on all sides around has allowed largely criminal elements to take over this technology while leaving the civilized elements in the dark. Openness and the spreading of this knowledge among good and decent people is essential, and for that we will have to overcome our frightened denial.

> "Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.
>
> "Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p08.htm

Like the huge energy of the Tesla howitzers, the ability to engineer the mind with longitudinal waveforms is a seeming "magical" thing which can be used for good or ill. Bearden envisions that further developments will allow something like the "downloading" of knowledge via scalar interferometry. Everyone could be highly educated, and quickly and easily.

In the conclusion of the paper "Mind Control and EM Wave Polarization Transductions" Bearden makes a plea to the nations of the world regarding the use of psychoenergetic engineering.

> "Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth.

> "Yet it can revolutionize medical science, education, communication, and psychology. We foresee the day -- perhaps 30 years hence -- when education will be accomplished by directly loading the software into the mind. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks on will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an MD And so on. When that happens then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.
>
> "We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p08.htm

(Perhaps after so many frightening topics a slight break is in order here to consider the lighter side of psychoenergetics. For example, what will be the ultimate impact on the entertainment industry?)

The Feelies at Last!

Psychoenergetics will be able to bring about at last, for better or for worse, Aldous Huxley's fanciful idea in his novel "Brave New World," an entertainment format which he called "The Feelies." Those were something you went to, like going to the movies, but you would experience all the actual feelings themselves, as if the depicted events were actually happening to you. Although our "leaders" seem too dim to grasp what is going on in the world of scalar electromagnetics, it should not take some young Hollywood genius too long to understand that The Feelies are now possible, and that the immense megabucks of potential profit more than justify heavy investment in bringing it to market.

One can imagine that first "Feelie" theater, "The Huxley" perhaps, a large comfortable area surrounded by nine giant screens (not just one screen!), which become 3D if you put on your headset. And The Huxley is fitted with the finest sound, and oh, by the way, a modest sized scalar interferometer with computer. And there are one or more new tracks now recorded on the film next to the sound track. These are the digital patterns of thoughts, feelings, and ideas to be fed into the theater's interferometer. And these thoughts and feelings and ideas will be felt by the audience as if they were their own thoughts and ideas. It is an entirely new art form. Psychevision.

As important (and no doubt famous) as the actors would be the "feelers," those from whose minds the feelings were lifted in the original programming of the Feelie tracks. Because, for example, Brad Pitt might be a good and handsome actor, but can he really feel it? Like the dubbing of foreign language films, it might be necessary to dub the "emotional track" using a deeper and more profound person's feelings, thoughts and ideas.

It might even be engineered so that if you sat on one side of the theater you would feel like the hero, and on the other side you would feel exactly like the bad guy! What would that lead to?

We have all cried at a movie some time, most people have, though they might not want to admit it. But this would be deep sobbing and a flood of tears, if the director so desired it. You would feel it at your core. Or, take a big chase scene, you would be terrified. They are breathing down your neck! You would be gripped by panic.

For that matter, the whole genre of Horror-Feelie would be too much for many people to take. Do I really want to feel that an actual vampire is tearing at my throat with his fangs?

A Religious Feelie would make you feel the exaltation of a saint at his prayers, or give you the definite feeling that you are Christ suffering on the cross. All the pain could be there too. The stickiness of the blood. The very feeling itself: "Why hast thou forsaken me?" Who among us could come up that despair for scanning? Had this been developed a bit earlier we might watch a documentary of, say, Mother Theresa, containing her own actual deep sense of compassion, which we would feel arising in us as our own.

Then again, a Psychedelic Feelie could induce an LSD-like state and simply blow the audience's minds.

Even in its simplest most primitive mode, the hypnogogic trance induction with its increased suggestibility, would completely alter the experience of even ordinary current movie going.

6. As it Stands - 2002

"Anyway, it's presently a "Mexican stand-off" with the various parties maneuvering with check and countercheck. That is the real "balance of terror" that is held by such a fragile balance."

"You can see the implications if such weapons fall into the hands of the radical terrorists. Ordinary positive energy EMP shooters can be made fairly readily and cheaply from nearly just Radio Shack parts or surplus parts. It just takes some know-how... When and if the Yakuza [Japanese mafia] gets those (and they will, it's just a question of time), you can begin to appreciate some of the turns this present terrorism and our war on it will take."

Bearden http://www.cheniere.org/correspondence/021702a.htm

The very thought of the building of one of these howitzers by "some lone-nut gunman" from Radio Shack parts makes one realize how serious it is for humanity to face the crisis it has come to.

Once again, as at the dawn of the nuclear era, humanity finds itself at the crossroads of Heaven and Hell. As a species, we must stop and consider. And there is at least one real difference between this visit to the crossroads and the last: this time there is a very real possibility of a true "earthly heaven" on the Heaven side.

The poor nations could come to a good life. Disease could be largely eliminated. Fossil fuel pollution could end. Global warming could be solved, and the weather could be managed in a humane way, bringing water where there is drought, and sunny skies where there is flood. Scalar interferometry would be there to avert any possible hits by asteroids or comets. Mars could be colonized and terraformed. Everyone would be highly educated through psychoenergetic downloading of knowledge. And the wisdom of the wise could be realized directly through scalar connection to their minds and hearts.

These stunning sudden possibilities give me pause to think: if our species has reached this crossroads in such a short span of centuries, then surely other species in this vast universe have passed this way before and long ago. And of those who chose the path to "planetary heaven," what is their life like now? It might be good if some of them were with us now, to guide us through this last struggle over this pesky old good-and-evil problem. It is our last chance to choose the path which benefits the whole species, and not just some elite group or faction.

The Disclosure Project

One person working tirelessly to bring to bring these technologies into the open is Dr. Stephen Greer of the Disclosure Project, which is calling for congressional hearings into the issues of UFOs and the free energy technology that was, perhaps in part, reverse-engineered from them.

> "Once abundant and nearly free energy is available in impoverished areas for agriculture, transportation, construction, manufacturing and electrification, there is no limit to what humanity can achieve. It is ridiculous — obscene even — that mind-boggling poverty and famine exists in the world while we sit on classified technologies that could completely reverse this situation. "So why not release these technologies? Because the social, economic and geo-political order of the world would be greatly altered. Every deep insider with whom I have met has emphasized that this would be the greatest change in known human history. The matter is so highly classified not because it is so silly, but because its implications are so profound and far reaching. By nature, those who control such projects do not like change. And here we are talking about the biggest economic, technological, social and geo-political change in known human history. Hence, the status quo is maintained, even as our civilization hurtles towards oblivion...
>
> "With the types of weapons currently in the covert arsenal — weapons more fearsome even than thermonuclear devices — there is no possibility of a survivable conflict. Yet in the darkness of secrecy, actions have been taken on behalf of every human that may

> endanger our future. Only a full, honest disclosure will correct this situation. It is not possible for me to convey in words the urgency of this."
>
> Dr. Stephen Greer http://www.disclosureproject.org/ES-DisclosureImplications-2.htm

Greer and his team have assembled hundreds of witnesses, many of whom are military or ex-military, who are ready to testify to congressional hearings what they know about the covert black projects which have kept free-energy and antigravity propulsion technology secret from the people.

Greer points out the need to regulate these technologies so they are limited to peaceful uses, but this is of course difficult when the government simply chooses to pretend they do not exist. In the meantime, as the President and the Defense Department continue the charade of war-as-usual, those scalar (longitudinal wave) installations which DO exist continue to fall into ever more sinister hands. And there is no reporter in the press brave enough to ask, "Mr. Secretary of Defense, what steps are you taking to defend our MINDS from attack by the big Russian longitudinal interferometers?"

Longitudinal Interferometers are Proliferating

In a later statement Bearden reveals that in fact the Yakuza [Japanese mafia] has already leased some of the interferometers with a nearly $1 billion "down payment:"

> "After the collapse of the Soviet Union's economy, lots of things got available for lease or purchase that would never have been under the old system. That happened in the resulting economic chaos. The Russian Mafia rose to prominence, as did some wheeling and dealing entrepreneurs who played both ends against the middle.
>
> "In that atmosphere, and in the need for money, the KGB allowed leasing of the earlier longitudinal EM wave interferometers to a consortium of the Yakuza and Aum Shinrikyo. They did this only after they had tested and deployed their great new quantum potential weapons. That class of weapon is the dominant weapon on earth. Five nations now have it: Brazil, Russia (KGB), China, and two nations friendly to the U.S. The deal to lease many of the earlier LWIs [longitudinal wave interferometers] to the Japanese Yakuza and Aum Shinrikyo was consummated at the end of 1989, with a down payment of $900 million in gold bullion as "up front" money. I don't know what the lease per year was and is, but probably something like a billion dollars a year. So the rogue Japanese acquired longitudinal EM weapons and that technical know-how. In fact, the Yakuza is producing certain kinds of those in its own facilities in Japan now."
>
> Bearden http://www.cheniere.org/correspondence/030202a.htm

Personally, I am a little uncomfortable with the knowledge that the Japanese mafia may now be able to insert things into my mind. I'd like to sleep at night knowing that my dreams are my own, such as they may be, and not some weird fantastic visions cooked up by a criminal organization. This is a world emergency.

And in Bearden's analysis, it all comes back to the issue of energy. It is the energy crisis which is now provoking the current "oil wars," since oil production has hit the "Hubbert Peak" and will decline. And the electrical power grids must be replaced with scalar power as fast as possible, for these grids are now totally vulnerable to any possessor of longitudinal interferometers. They can all be wiped out in minutes. By anyone who possesses or controls a Tesla howitzer.

> "Anyway, in my personal view, we should have a national Manhattan Project right now, to develop decentralized electrical power systems taking their energy from the vacuum . . . The MEG is real and so are several other systems invented by private inventors and colleagues. All these efforts should be massively funded and massively and quickly developed. The survival of this nation is going to depend upon it. . . . There are many other facets, but the major point is that our scientific apparatus and the public are still unaware of what a desperate struggle we have entered, since it was thrust upon us. 9-11 was a wake-up call. It was certainly bad enough, and one grieves for those lost American lives. But I fear the worst is yet to come, and it is coming at us like a runaway train hurtling right down the track at us."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm
>
> "The electromagnetic weapons mentioned by the Secretary of Defense in April 1997 are in terrorist hands also, including the Yakuza and Aum Shinrikyo. That rogue group leased those earlier weapons on site in Russia, from the KGB, at the end of 1989. The SecDef stated that such weapons were being used to initiate earthquakes, initiate volcanoes into eruption, and control and engineer the weather. Quite true. And presently there is a truly massive set of weather engineering operations going on over North America, from that rogue Japanese group manning the weapons in Russia. Some 10 nations of the world now have the type of longitudinal EM wave interferometer weapons (which are what the SecDef was most probably referring to). There are other even more fearsome weapons, possessed by five nations. So a great deal of the state of the world is not covered in the news at all, and will not be . . .
>
> "We have been in an undeclared war of an eerie kind for some decades. That war seems now to have started to heat up also. The destructive capability of some of these weapons is awesome, and far greater than nuclear."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm
>
> "In that sense, the first phase of WW III is already completed. Now we are only waiting for Phase II.
>
> "It is against that backdrop that I think one must "analyze" such things as how fast we will be replacing the power grid (it's going to go anyway, and catastrophically, the terrorists will see to that). And everything we have is critically dependent upon energy, from the economy to jobs to mobility, to the millions of trucks that transport our goods every day, etc.
>
> "As far as I am concerned, we are already late in declaring a great national emergency in energy, because of the coming destruction of the present energy infrastructure. The

> scientific community should lead, follow, or get out of the way. The emergency is not in seeking "alternative but conventional" systems. It is in developing "energy from the vacuum" on an emergency, crash basis.
>
> "Unfortunately, everywhere I look I just see business as usual. Most Americans (and news media) seem to think, hey, Afghanistan is finished except for a bit of cleanup, and it's just about over).
>
> "It isn't. It has only just begun."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm

Beyond the Howitzers: Quantum Potential Weapons

Do not use EM propagation through space

As more and more nations acquire the scalar interferometers, Bearden speculates on the next generation of longitudinal weapons called "quantum potential" [QP] weapons.

> "The quantum potential weapons are presently only held by three nations: Russia (in the hands of the KGB, which is now know by its new name), Brazil, and the "friendly little nation." Red China may be working on them, but is just at the beginning...
>
> "The QP weapons have been highly reserved by the Russians. However, the complete extent of their capabilities is still being figured out...
>
> "The quantum potential weaponry (possessed by only three nations) is supreme, because it uses multiply connected spacetime, and DOES NOT involve propagation of EM energy through space as does the longitudinal EM wave interferometers.
>
> Bearden http://www.cheniere.org/misc/qp.htm
>
> "The operational deployment of full-scale strategic QP weapons on site in Russia occurred at the end of 1989, which is when the formal weapons first went operational. These weapons are not in the inventory of the regular Russian armed forces, but all research and development, manufacturing, deployment, an manning and employment is under the ruthless control of the KGB. Shortly thereafter (within weeks), the KGB leased many of their earlier longitudinal EM wave interferometer (LWI) weapons to a rogue Japanese group comprised of the Aum Shinrikyo and Yakuza."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

In another paper, "Time Polarization - Significance and Weaponization," Bearden tells the story of how a KGB longitudinal wave attack was averted by the longitudinal weapons of a "friendly small nation."

So as it stands, it is precarious. Not until the people demand an end of what author Jim Marrs has dubbed the "Rule by Secrecy," can the people's representatives tackle the big problems of a worldwide ban on Scalar Wars and the shifting over to an essentially oil-free economy.

Latest Update from Tom Bearden - March 2002

The whole situation is changing so fast that when I asked Tom to look over this article he mentioned the seeming contradiction when some of the above statements are taken together. At different times the information he has is updated so one has to keep a kind of running track. He was kind enough to clear up the confusion here with the situation as it stands as of March, 2002.

> "We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy." Tom Bearden

Emptiness is Full of Energy

"The Tao is an empty vessel;
it is used, but never filled.
Oh, unfathomable source
of ten thousand things!

"Oh, hidden deep but ever present!
I do not know from whence it comes.
It is the forefather of the gods."

Lao Tzu -Tao te Ching

Afterward

Because of the extreme seriousness of the topics brought forward in this article, I asked Tom Bearden to look it over to make sure I had not gotten anything too terribly wrong or mixed up. I did not want to pass on any misunderstanding of my own. I felt a little nervous writing to an eminent scientist, and hoped I had not presumed upon his time.

But Tom was so kind as to answer immediately, and to my relief I had not gotten anything too terribly wrong. Here is his letter, in which he clears up a misunderstanding, and reports on his current health treatment, and gives a call to the young grad students to turn themselves loose in the whole new field of scalar electromagnetics and free energy.

Subject: RE: Article on Tom Bearden and "Scalar Wars"
Date: Fri, 22 Mar 2002 22:54:01 -0600

Bill,

Well, I must say I'm astounded at your persistence and hard work in congealing all that together. My compliments on a hard job quite well done.

Couple tiny things: In the area mentioning Kervran and Kaznacheyev: Kervran was French; Kaznacheyev was and is Russian. Little typo of "spredding" the immune system. Should be "spreading".

We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy.

Otherwise, it's pretty accurate.

Just now I'm letting the weapons take care of themselves; have to, as I will be recovering now for nearly a year from now after the heart attack last year and discovering that I had contracted mycoplasia (the BW modified kind) in Canada in 1968. Finally got a test which confirmed the chronic mycoplasia (which I had for 33 years!). The stuff burrows inside one's red corpuscles, then draws nutrients from the hemoglobin, hardening it and reducing its ability to take on oxygen (by as much as 50% or more). The resulting symptom is loss of endurance (which I had for that 33 years) -- the so-called "chronic fatigue syndrome".

Many of the Gulf War Veterans who are sick also have it, though they apparently have other complicating factors as well. A very high percentage of persons with chronic fatigue syndrome also have mycoplasia. The treatment for the long-standing mycoplasma infection of that nature is to stay on antibiotics for a year. The only time the mycoplasma is vulnerable is when one wears out a red cell in normal usage, and the body makes a new red cell to replace it. The mycoplasma comes out of the dying cell to infect the new replacement cell, so it is exposed and vulnerable at that specific time and that time only. The antibiotics kills it at that time. Hence the necessity to stay on antibiotics for such a long time so it is continuously in your bloodstream (I've already been on antibiotics for over two months, and have at least 10 more months to go). We also take a little medical oxygen every day.

Another symptom of the mycoplasma infection is the association of runaway fibrillation of the heart. To stop that, there are pills one takes, and they stop it. But they also further reduce the volume of blood pumped by the heart, so that this projects one right back into the hypoxia, dizzy, loss of balance, etc. Catch 22 situation; the fibrillation will build and cause heart attack, stroke, etc. if not stopped, and if stopped it puts you back in the hypoxia which can cause many things! With the conventional treatment, you're damned if you do and damned if you don't. Hence the need for a little extra medical oxygen, so you can take the "heart clamp" pill and stop the fibrillation, but also augment the oxygen enough to continue to survive without a stroke or some such. Interestingly, it's very difficult to even obtain a mycoplasma test, and Medicare fights you tooth and nail over paying for the oxygen. Either they will pay for it or I'll pay for it myself; it is absolutely necessary.

After a heart attack, one is treated in the heart clinic by specialists who are very caring and excellent practitioners, but know nothing of mycoplasma infection, so not consider it even a part of "heart disease", and who are rather puzzled when one has no clogging of the arteries, has the right count of red corpuscles (largely nonfunctional, of course, but there), pumps the right amount of blood, etc. So they identify and treat only the runaway fibrillation, do not prescribe the oxygen -- and essentially place you on a regime as best they know, but one that is guaranteed to kill or maim you or turn you into a totally disabled person or human vegetable via stroke, paralysis, etc. Fortunately my family doctor will prescribe the oxygen anyway, though I may have to pay for it myself. Whatever works!

However, I'm slowly improving little by little, working as much as I'm able to and trying to finish my book: Energy from the Vacuum: Concepts and Principles, and get it to the publisher for publication later this year. Hopefully the book will turn all the interested young grad students and post-docs loose in the free-energy area. They will not have to spend 30 agonizing years to get to where I am, and where a few others are, but can simply start here and go forward. Since speed to develop energy from the vacuum is of the essence because of the world situation, I'm moving on it as fast as I am physically able to.

Meanwhile, working with the AIAS (Alpha Foundation's Institute for Advanced Study), the AIAS (15 authors) has now gotten about 20 scientific papers published in leading journals (Foundations of Physics, Foundations of Physics Letters, Physical Scripta, Optik, etc.) dealing with electromagnetic energy from the vacuum. Several more are either approved for publication or in the review process also. So slowly we are getting it into the mainstream scientific literature. Hopefully it will be in time, but it's cutting it close.

Very best wishes,

Tom Bearden

Late update

Massive display of longitudinal interferometers?

UFO Fleet Taped Over Bedhampton, England

Is this sighting a case of the testing of all the Russian (KGB) longitudinal interferometers at once? And coordinating their targeting through the woodpecker grid to a common place, as an exercise in making a massive scalar attack upon a given region? Was this a confluence of Russian Tesla howitzer marker-beacons?

Bedhampton, England - March 22, 2002 - Video Frames by: Anthony Woods

Certainly for a massive all-out scalar attack you would want coordinated movements of the target areas of many howitzers at once, for you could then just cut a bloody swath across a wide area, all the howitzer blasts marching across the land in a kind of scalar version of the "scorched-earth policy." All in a line, blasting, and heat, and mind waves. Or you might want to put the blasts all in the same place for a truly horrendous destruction. And on top of that biowar on citizens whose immune systems have been weakened with the quantum potential weapons. This is the latest version of Armageddon.

And should any madmen gain control of these weapons, any rogue group at all, and should they have some agenda to radically depopulate the world, to simply kill by the billion, the means will be in their hands to do so.

Update: MEG patent granted

Motionless Electromagnetic Generator Patent Granted

This message is from Jean-Louis Naudin, the French scientist who successfully replicated the Motionless Electromagnetic Generator (MEG) in France. He announces the granting of the patent to the MEG Builder's group at Yahoo.

----- Original Message -----

From: jnaudin509@aol.com
To: jlnlabs@yahoogroups.com
Sent: Tuesday, March 26, 2002 2:20 PM
Subject: [jlnlabs] (Info) GOOD NEWS : The Tom Bearden's MEG IS PATENTED

Dear ALL,

GOOD NEWS !!!!

The Motionless Electromagnetic Generator from Tom Bearden is now PATENTED US 6362718 granted on March 26, 2002

US Patent 6,362,718 :
Motionless Electromagnetic Generator (MEG)

Abstract: An electromagnetic generator without moving parts includes a permanent magnet and a magnetic core including first and second magnetic paths. A first input coil and a first output coil extend around portions of the first magnetic path, while a second input coil and a second output coil extend around portions of the second magnetic path. The input coils are alternatively pulsed to provide induced current pulses in the output coils. Driving electrical current through each of the input coils reduces a level of flux from the permanent magnet within the magnet path around which the input coil extends. In an alternative embodiment of an electromagnetic generator, the magnetic core includes annular spaced-apart plates, with posts and permanent magnets extending in an alternating fashion between the plates. An output coil extends around each of these posts. Input coils extending around portions of the plates are pulsed to cause the induction of current within the output coils.

Inventors: Patrick; Stephen L. (2511 Woodview Dr. SE., Huntsville, AL 35801); Bearden; Thomas E. (2211 Cove Rd., Huntsville, AL 35801); Hayes; James C. (16026 Deaton Dr. SE., Huntsville, AL 35803); Moore; Kenneth D. (1704 Montdale Rd., Huntsville, FL 35801); Kenny; James L. (925 Tascosa Dr., Huntsville, AL 35802)

Appl. No.: 656313
Filed: September 6, 2000

For some technical info see my web site at : http://jnaudin.free.fr/html/meg.htm

Best Regards
Jean-Louis Naudin
Email: JNaudin509@aol.com
Main Web site : http://jlnlabs.org
Site France : http://jlnlabs.multimania.com

Click here to view patent # 6,362,718

Patent Story on Rense.com:

Rense.com

MEG Scalar Energy Device
Patented - Production Starts Next Year

From Bill Morgan
wmorgan@nycap.rr.com
3-28-2

Patent was granted on March 26, 2002 for "The Motionless Magnetic Generator," (MEG) US Patent 6,362,718, which is likely to become the first commercially available free energy device in history in about one year from now. The machine will provide free electricity from the vacuum, for the life of the device, which should be a very long life since it has no moving parts. You can see a picture of scientist Jean-Louis Naudin's MEG replication model at: http://jnaudin.free.fr/html/meg.htm.

It has strong magnets, coils, and a controller unit with the electronics. Naudin made the announcement to the MEG-builder's Yahoo group. The announcement has significance since the patent office has always been skeptical of devices which seem to "get-something-for-nothing." But according to the new science of scalar electromagnetics, the MEG does not break the law of conservation of energy. It's just that the energy is conserved in the fourth dimension, time, and not our 3-space world.

The MEG provides electrical energy by tapping the longitudinal electromagnetic (EM) waves which exist in almost infinite abundance in the vacuum of space. This ocean of energy which permeates everything is sometimes called the "zero point" energy, since it remains there even at absolute zero temperature.

Four inventors are listed: Stephen L. Patrick, Thomas E. Bearden, James C. Hayes, Kenneth D. Moore.

Tom Bearden has explained the operation of the MEG on his website Cheniere.org, and also speaks about the new fearful weapons that can and have been made using the same "longitudinal waves" of the vacuum.

The complicated physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum," which can be found at Tom Bearden's website:

cheniere.org.(http://www.help4all.de/energy/MEGpaper.pdf)

The first MEG units to be produced for sale will output 2.5 kilowatts of free electricity. Forever. They should be in production about a year from now. Facilities for manufacturing the device are being set up in an unnamed "friendly nation."

This free electricity will flow indefinitely, without much, or any maintenance. The units may be hooked together to provide more wattage, so four of them would provide 10 kilowatts. After some production experience units will be made which output 10 kilowatts each. With a couple of those units a house could get off the electrical grid.

Tom Bearden, one of the inventors has said "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat." This cat, it would seem, is out of the bag now.

The ordinary EM waves we are familiar with are called "transverse" waves, to distinguish them from the new "longitudinal" EM waves of the vacuum. Bearden has explained in depth on his website cheniere.org that wherever there is a dipole (battery, generator, magnet) there is an unseen flow of longitudinal EM waves in that local vacuum, the only problem is in tapping that energy and "transducing" it to electricity. Bearden says that the problem with all the electrical circuits we have is that they are two-wire circuits, a loop by which half the energy goes back to destroy the dipole. In the MEG that closed loop is never made. So the dipole is not being destroyed.

Patents are not granted on devices which do not work, so in a sense this announcement proclaims a new era. The MEG device itself proclaims and proves that energy is abundantly available everywhere, for free. We only have to build the devices to tap it. MEG type devices could be put into a car with an electric engine to make a truly fuel-less automobile. Electricity can be available in very remote places. People will be able to get off our very terrorist-vulnerable power grids. In an era of terrorism a highly dispersed power system would be most desirable rather than our centralized systems.

What is somewhat astonishing to me is the discovery by Bearden et. al. that time itself is actually compressed energy, and that this free energy is actually coming from the time domain, the ocean of longitudinal EM waves which fill the empty vacuum of spacetime. In fact, time is energy compressed by the same factor that matter is compressed energy: the speed-of-light squared. Thus there is a new companion to Einstein's E=mc2. Can you say "E equals delta-tee-cee-squared?" The "tee" is time and delta-tee is change in time.

A company has been set up to manufacture the MEG called Magnetic Energy Limited.

The ramifications of free energy are enormous. The oil wars are not necessary. If we threw as much money at this technology as we are spending on the oil wars we would be free of the need for oil in less than a decade. With fuel-less cars air pollution will be greatly lessened. Third world nations can raise their standard of living eventually. And the energy is free. And it never runs out.

I have put together a kind of "Bearden for Beginners" article which explains some of the basic concepts of the current state of scalar technology.

http://www.prahlad.org/pub/bearden/scalar_wars.htm

The granting of the MEG patent is the herald of the new era of scalar electromagnetics, and the free energy which flows forever and never runs out.

Bill Morgan
wmorgan@nycap.rr.com

Do the magnets run down after a while?

Further developments: Tom Bearden clears up a question

When the MEG patent was released many questioned the patent language which seemed to imply that permanent magnets of the MEG would "wear out" after awhile. I myself wondered about the same thing so I put the question to Tom, and here is his reply.

> Dear Bill,
>
> Thanks for the kind words; much appreciated.
>
> I'm not up to going on radio or TV, and won't be for quite some time. Still suffering from hypoxia, though very slowly improving, and will still be on antibiotics for 10 more months at least (maybe longer).
>
> Obviously an extensive dialog developed between the Patent office and our patent attorneys. As best I can understand it, the Patent office accepts the fact that a permanent magnet produces and emits energy, or has a history of accepting it. They do not appear to be at all cognizant of the broken symmetry of opposite charges, and hence of a dipole. So in their opinion (which seems to be vague), the magnet has to use itself up very gradually, or some such. So some rewrites were essentially required by the Patent office.
>
> All I can say is that a magnet is a very durable thing, so long as you don't heat it up too much or shock it to much. Certainly one will make a good 20 years, and of course there are plenty of magnets that have been around awhile longer.
>
> That said, in our earliest experiments we did "measure" some such effect, or so we thought, until we found that a particular instrument was bad and giving erroneous readings. So with a new instrument, we have not observed since then any such readings.
>
> Out of all that fuzziness came the inclusion of "using up the magnets gradually". So far as we are aware (after correcting those early readings on a defective instrument), we have seen no such measurements or effects since then at all. Nonetheless, just in case there is some very small effect in that respect, at least it's in there.
>
> Actually, as you are aware, capacitors and solid state components are the major concerns for failure, and any system does require maintenance, both periodic and whenever malfunction or failure occurs. The MEG is no different from any other device in that respect, so of course it will have a component failure rate and maintenance required from time to time, just as any other electromagnetic device. But without moving parts, it can be ruggedized and made very durable indeed.
>
> Please give Jeff Rense my regrets, and I do admire his work and his show. He's laying it in there.
>
> The depth of the scientific mindset against COP>1.0 is inexplicable, particularly when physics itself already has COP>1.0 validated experiments (such as Bohren's experiment,

which outputs 18 times as much as one inputs by Poynting calculations), and it works every time. Any competent university optics lab can perform the experiment. Also, even the conventional texts admit that a charged capacitor or electret laid on a permanent magnet, so that the E of the cap or electret and the H of the magnet are at right angles, is a "free energy machine". That silly thing will just sit there and pour out Poynting energy S = EXH indefinitely. Years and years and years. If you wait just a single year, that silly thing has changed the energy density of a volume of space a light year in radius, reaching out well beyond the solar system. Now that's a WHALE of a lot of EM energy that beast poured out in that year, and it will still be going strong and pouring it out at a steady rate. The charges and dipoles in original matter in the universe have been pouring out EM energy in that fashion for some 14 billion years, give or take a billion or so. It's also easy to see that, prior to Lorentz's ARBITRARY symmetrical regauging of the Maxwell-Heaviside equations, those equations prescribe both EM systems in equilibrium with their active environment, and EM systems not in equilibrium with it. The latter systems are permitted by the thermodynamics of open disequilibrium systems to perform five magic functions: (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator inputs (the active environment inputs the rest of it), (4) power itself and its load (the active environment inputs all the energy, like a windmill), and (5) exhibit negentropy. That is a well-established thermodynamics. Lorentz arbitrarily discarded all such open disequilibrium EM systems in order to get simpler equations having analytical solutions and not requiring numerical methods. So he simply scrapped all the really interesting and challenging EM systems in Maxwell's theory. ARBITRARILY!

That our scientific establishment continues to teach only that "half" of the theory, and not point out the rest to the young students, and not fund research into developing such systems that take energy from the vacuum, is simply inexplicable. It's a matter of total dogma, not science at all.

But hopefully it is changing a little bit. At least now we cite chapter and verse when these things were done, and what they are, and the young grad students and post docs can go check the papers and make up their own minds.

That is the real hope of the future: that we get a generation of young fellows, with excellent skills, who now know where all the skeletons got hidden in the closet, and understand that the electrical engineering model arbitrarily assumes an inert vacuum environment (falsified for more than a half century in particle physics) and a local flat spacetime (falsified for nearly a century by general relativity).

Once they know that past history, and go read into what broken symmetry of opposite charges means for a dipole and for all dipolar EM circuits, they will set to and change that more than a century of error. In that case, we shall have numerous solutions to the energy crisis rather quickly, popping up all over.

Good thing. It will certainly displace some great energy barons, but it will also free the populace. One can extract all the EM energy from the vacuum one wishes, anywhere, anytime, with ridiculous ease. One can produce that "electrical wind of energy" at will, by making a simple dipole and then letting it alone. The only energy problem is in how to then intercept some of that energy flow and "catch" it in a circuit, discharge it in a load to power it, and not use half the "caught" energy to kill that dipole that is gushing out the extracted EM energy from the vacuum.

Very best wishes,

Tom Bearden

END

The Tom Bearden Website

Help support the research

Foundations of Electro-Magnetism
Slide Series

- Major Points of Presentation
- Chung's Carbon Filament Negative Resistor
- Aharanov-Bohm Effect
- Absorption and emission reactions (1)
- Absorption and emission reactions (2)
- Anti-engine for cell's deviation reverses cell back to normal
- Asymmetric regauging produces excess force, which can be used to do work on the system
- Asymmetric regauging produces excess force, which can be used to do work on the regauging system
- Becker's bone fracture healing experiments
- Becker's theoretical control system governing regeneration
- Becker's theoretical DC control system involved with response to injury
- Geometrical Distortion of Poynting Energy Flow
- Cosmic Bootstrap: Before the Big Bang
- Bridging Enables COP > 1

- Aspects of Strong Local Symmetry
- Other Aspects of Strong Local Symmetry
- Bunge on the Status of Electrodynamics
- Bunge on the Status of Electrodynamics and Physics
- Bunge on the Status of Classical and Quantum Physics
- Fractions of BW Casualties Surviving
- First Step in Self-Promotion of Cancer Due to Hypoxia
- Differentiation and Dedifferentiation Timelines
- Evolution of Cells on Earth
- A Charged Particle q is a Coupled System of m and Ø
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (1)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (2)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (3)
- A Charged Particle is a Coupled System
- The Anomalies in Navy Electrolyte Experiments at China Lake
- Chung's Carbon Filament Negative Resistor

- Typical circuit has about 10^{-13} energy collection efficiency
- Classical View of EM
- Curved Spacetime acts as source or sink
- Chronic Disease Puzzling
- Cancer Characteristics
- Solving a Cancer Enigma
- Cancer: Characteristics
- Nonlinear Optics Distortion Correction Theorem
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)
- In Time-Reversed Zone: Nuclear reactions are biased
- In Time-Reversed Zone: Dynamics may be reversed
- Time-Reversed Zone: Significant only after time-density charging
- A New Conservation of Energy Law
- Questions on Carcinogens and EM

Radiation

- AS DEFINED, fields and potentials only exist in and on charged matter
- Serious flaws and errors in classical EM theory
- Electrodynamicists' Reaction to Removal of the Material Ether
- Reaction to Loss of Material Ether (1)
- Reaction to Loss of Material Ether (2)
- Some EM waves in Spacetime
- Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String
- Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna
- Phase Conjugate Wavepairs Produce New Waves
- Scalar Electromagnetics (Energetics) View of EM (1)
- Scalar Electromagnetics (Energetics) View of EM (2)

- Scalar Electromagnetics (Energetics) View of EM (3)
- Extended General Relativity Principle
- Extension of Work-Energy Theorem
- 34 Flaws in Classical EM Theory
- MTW's General Relativity Principle
- Use of General Relativity (GR) in Particle Physics
- The Missing Infolded Electrodynamics
- Infolded Engines: Excluded by Present Physics
- Mass curves and structures spacetime; ST curvatures change and structure mass (1)
- Mass curves and structures spacetime; ST curvatures change and structure mass (2)
- At Infinite Velocity: Each Is Every Other
- Lisitsyn's Report: Brain Code Broken
- Lorentz Closed Surface Integration
- Lorentz's physically insignificant energy flow can be collected and utilized
- Lorentz surface integration of the Poynting vector around a closed surface
- Longitudinal EM Photon Interaction with Charge
- Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass.

- Pumping with Longitudinal EM waves Time-reverses the Mass
- Maxwell's Quaternion Theory
- Maxwellian Systems before and after Regauging
- Maxwell's four equations reduce to: Page 1
- Maxwell's four equations reduce to: Page 2
- More to come

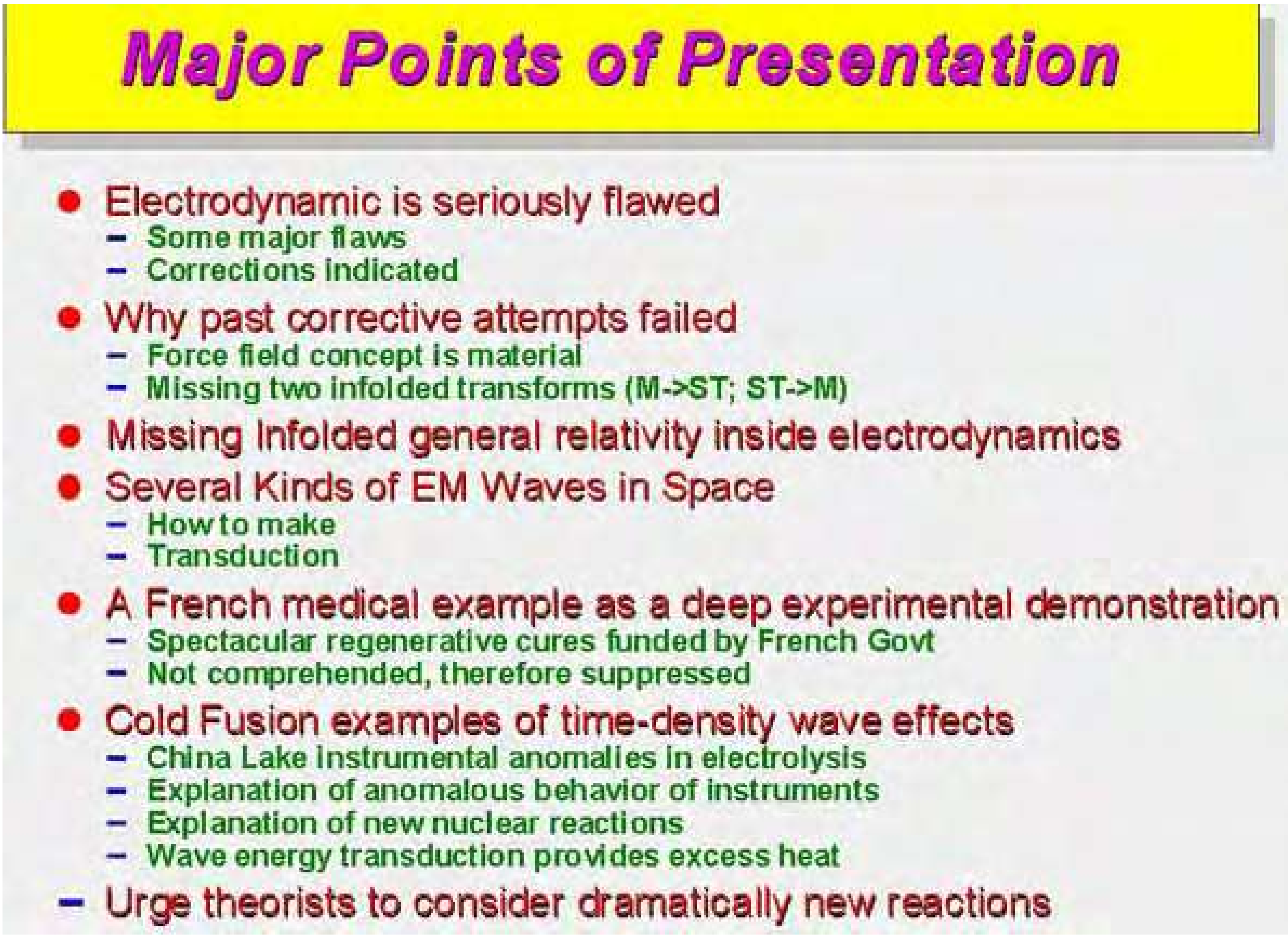

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

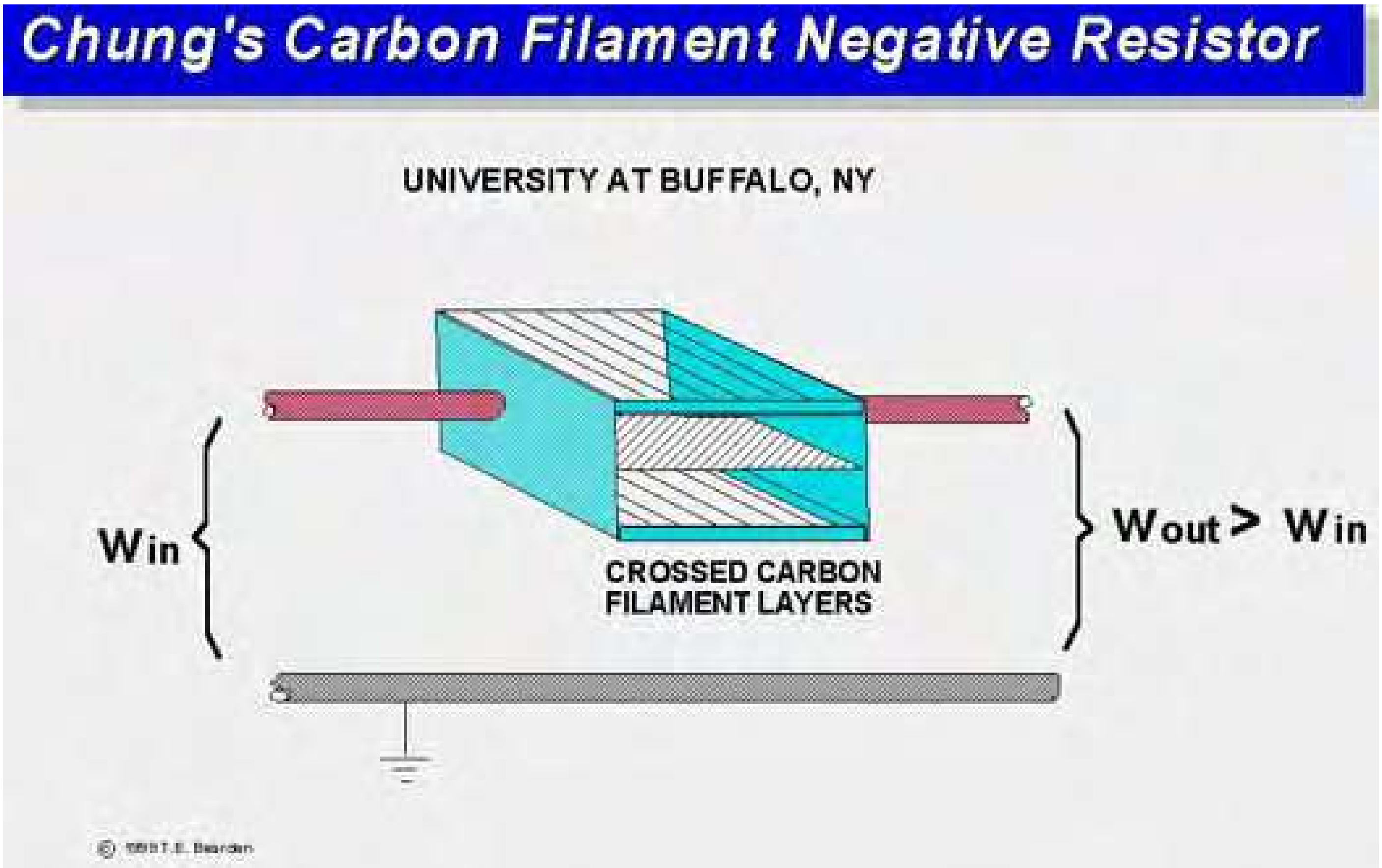

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

Win
CROSSED CARBON FILAMENT LAYERS
Wout > Win

© 1998 T.E. Bearden

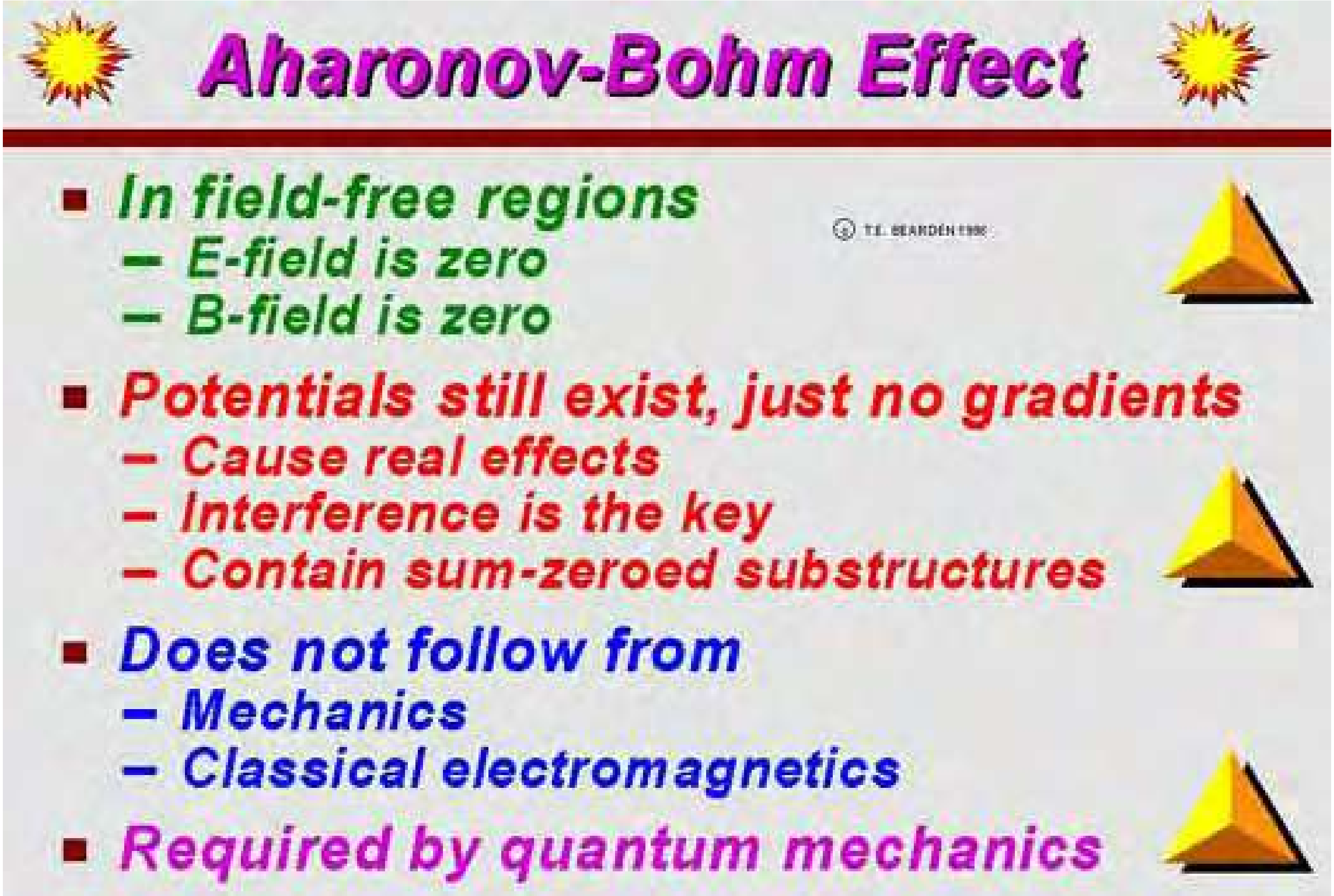

Aharonov-Bohm Effect

© T.E. BEARDEN 1996

- In field-free regions
  - E-field is zero
  - B-field is zero
- Potentials still exist, just no gradients
  - Cause real effects
  - Interference is the key
  - Contain sum-zeroed substructures
- Does not follow from
  - Mechanics
  - Classical electromagnetics
- Required by quantum mechanics

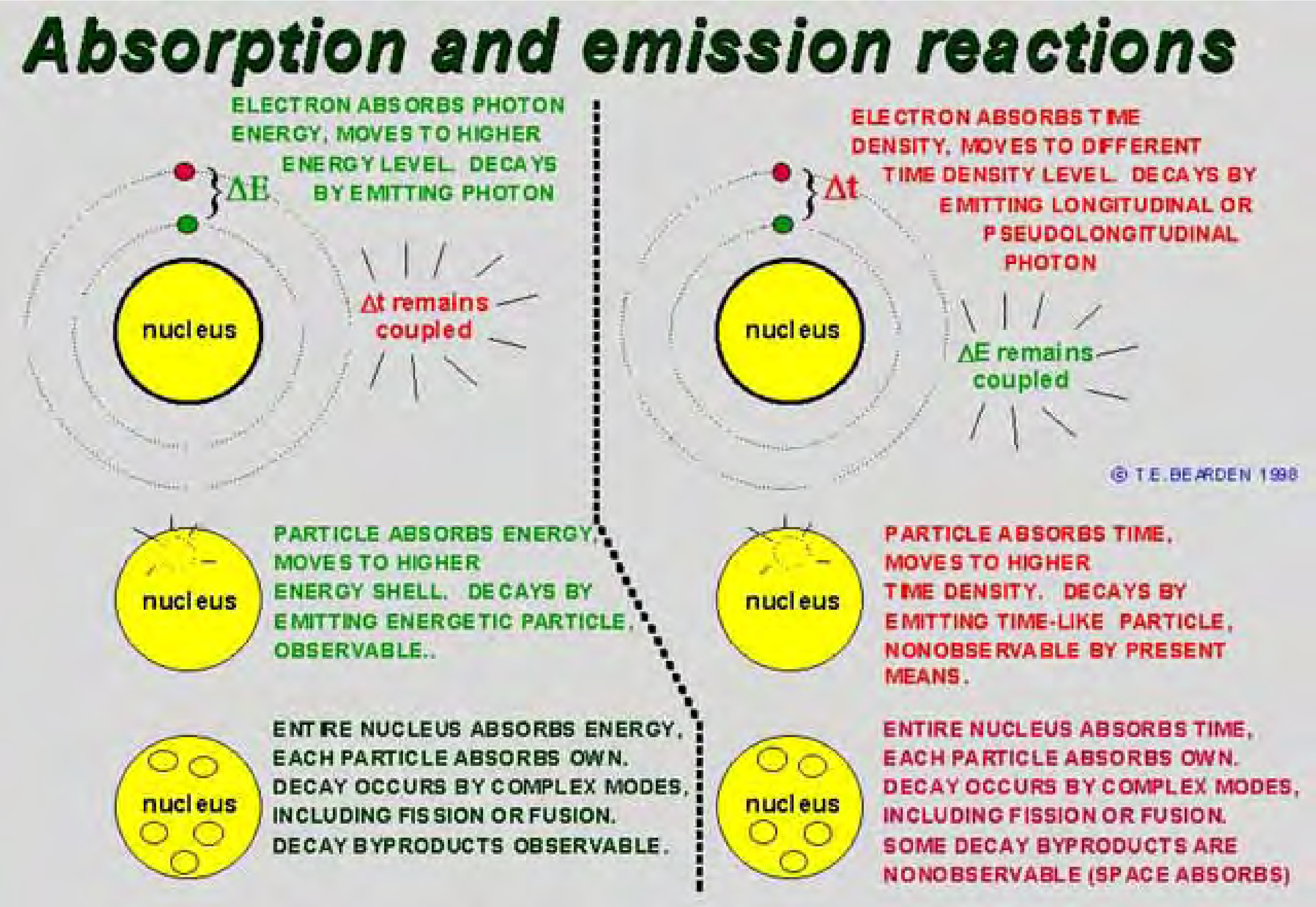

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ΔE
nucleus
Δt remains coupled

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

Δt
nucleus
ΔE remains coupled

© T.E. BEARDEN 1998

nucleus

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE. OBSERVABLE..

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

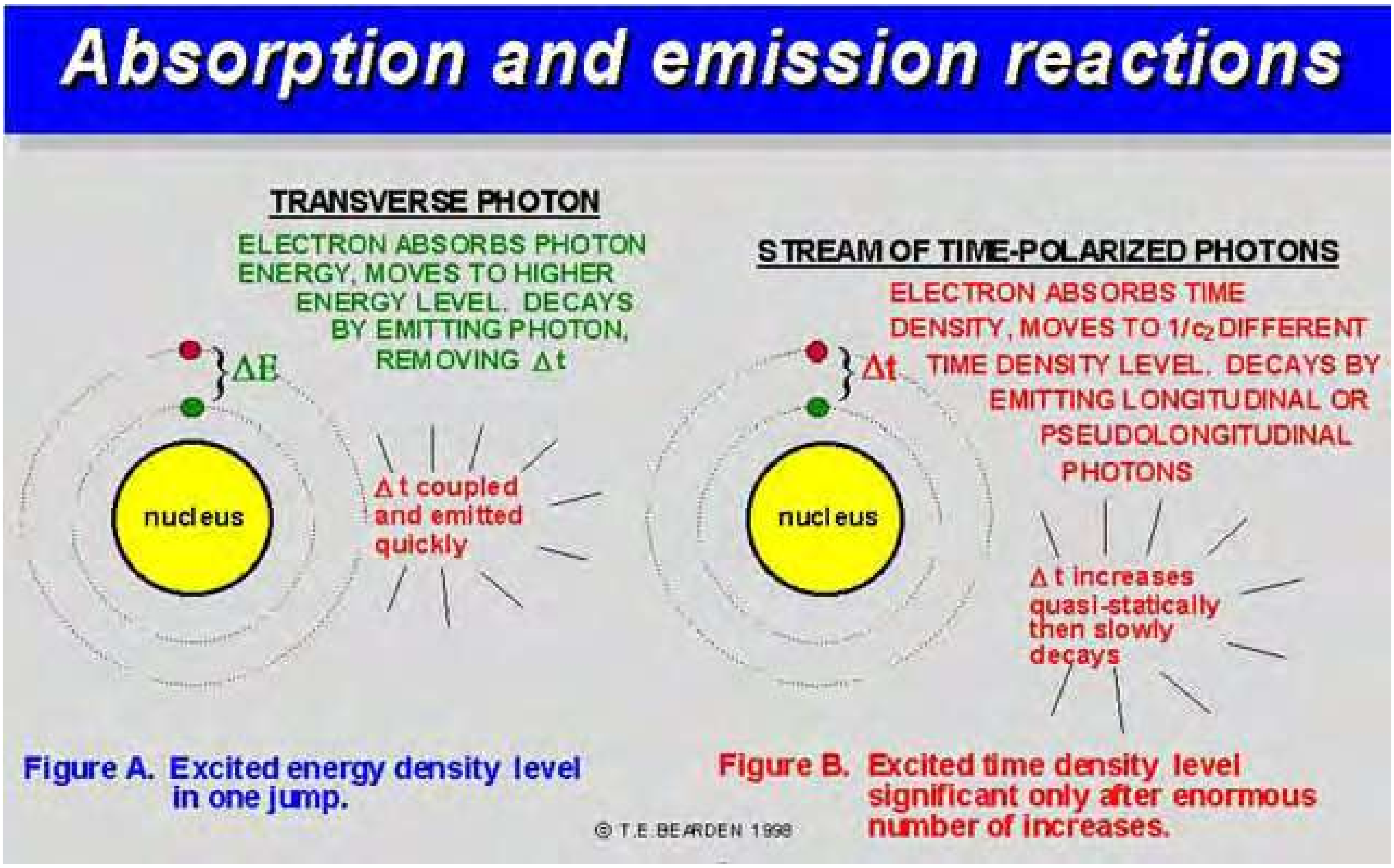

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

Δt coupled and emitted quickly

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c_2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt increases quasi-statically then slowly decays

Figure A. Excited energy density level in one jump.

Figure B. Excited time density level significant only after enormous number of increases.

© T.E.BEARDEN 1998

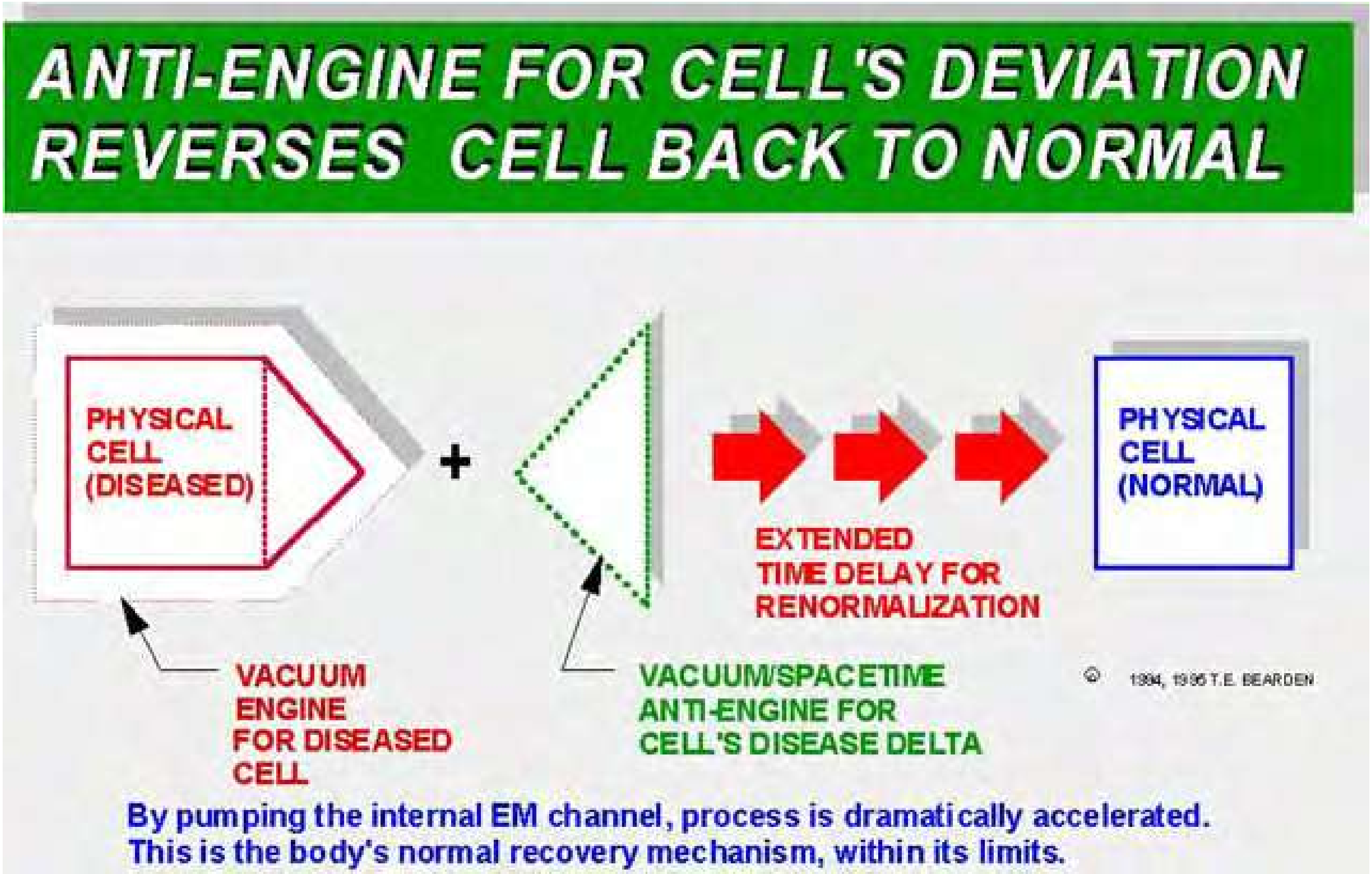

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL
CELL
(DISEASED)

+

PHYSICAL
CELL
(NORMAL)

EXTENDED
TIME DELAY FOR
RENORMALIZATION

VACUUM
ENGINE
FOR DISEASED
CELL

VACUUM/SPACETIME
ANTI-ENGINE FOR
CELL'S DISEASE DELTA

© 1994, 1995 T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated.
This is the body's normal recovery mechanism, within its limits.

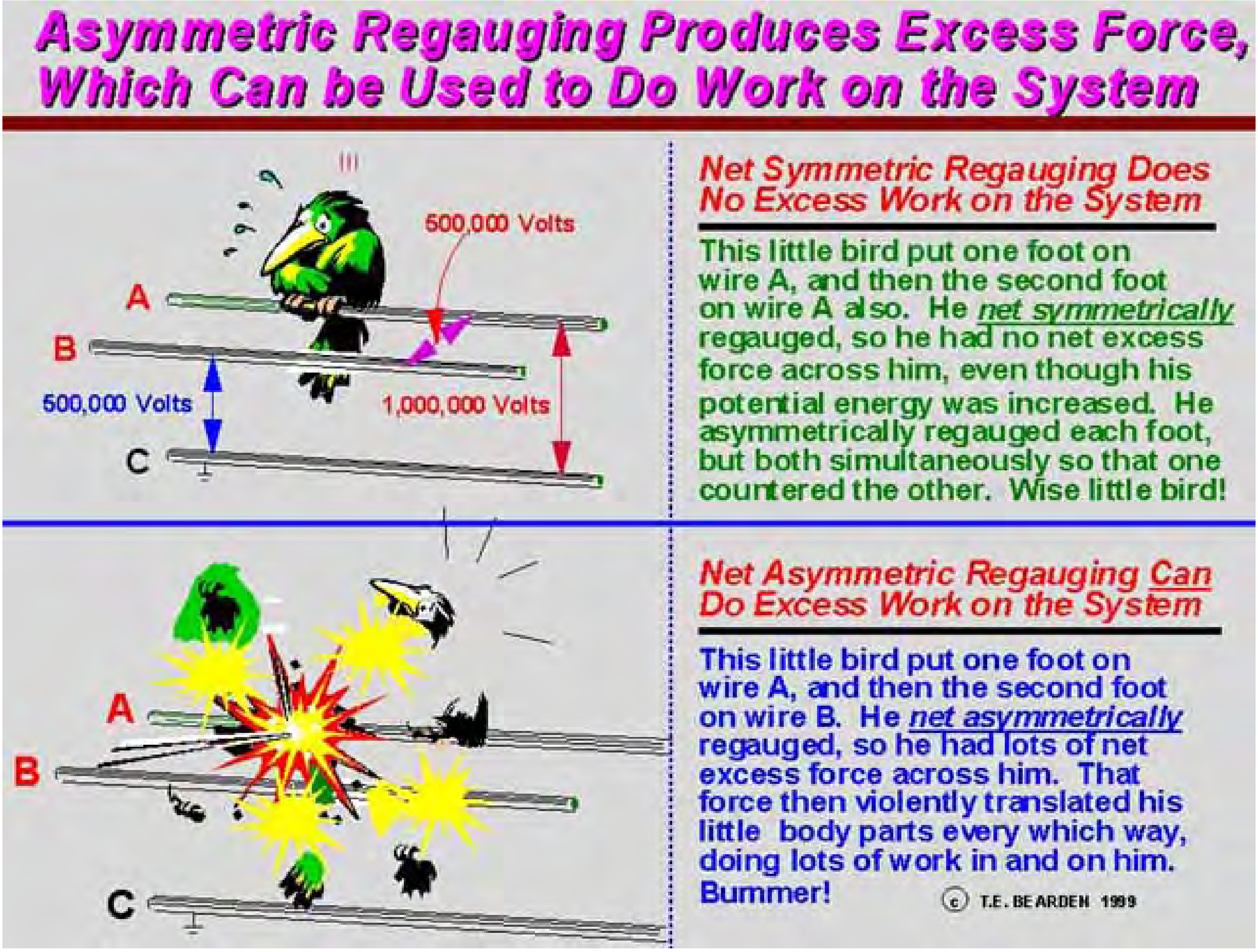

Asymmetric Regauging Produces Excess Force, Which Can be Used to Do Work on the System

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging *Can* Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1999

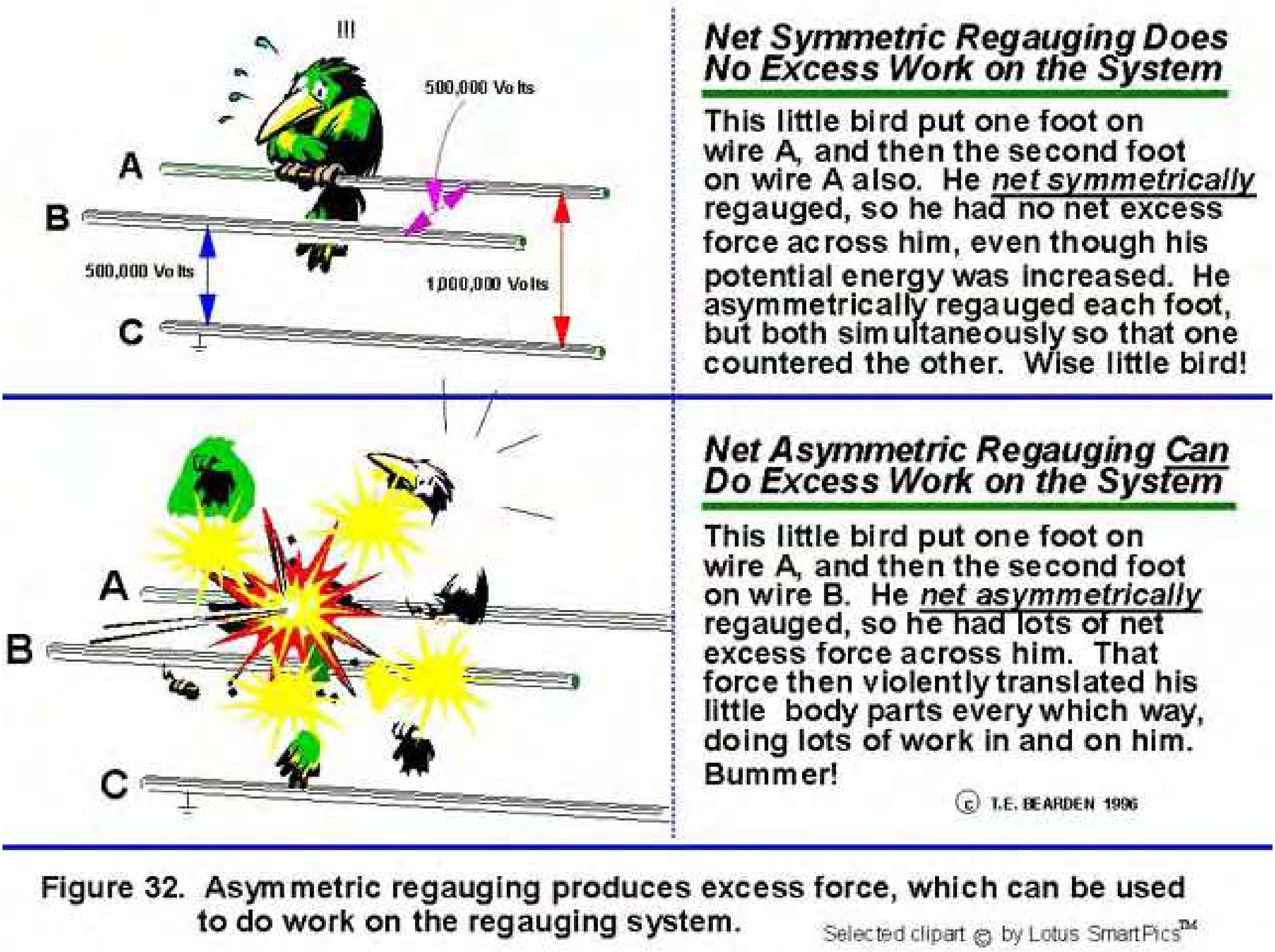

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging Can Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

ⓒ T.E. BEARDEN 1996

Figure 32. Asymmetric regauging produces excess force, which can be used to do work on the regauging system.

Selected clipart © by Lotus SmartPics™

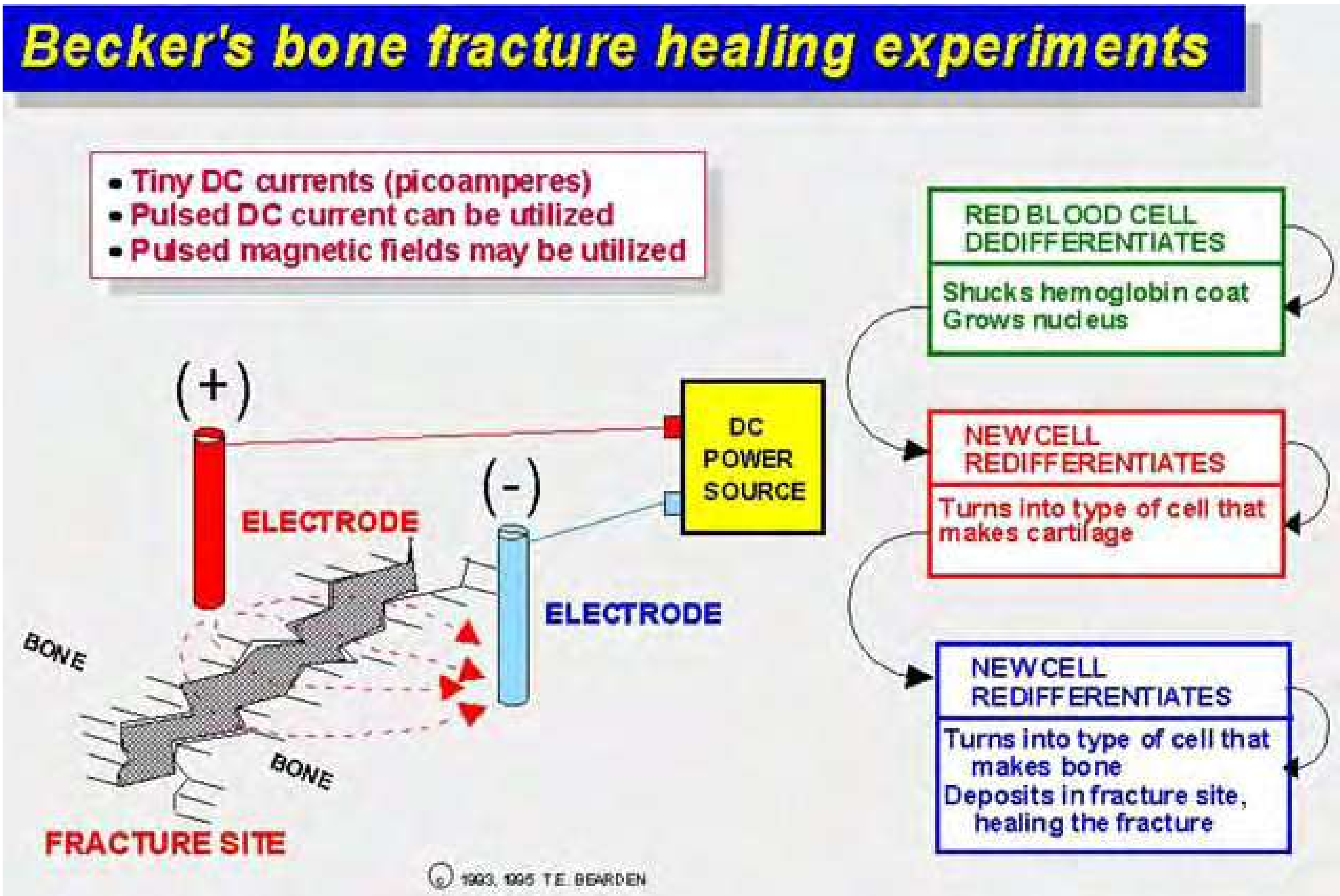

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1995 T.E. BEARDEN

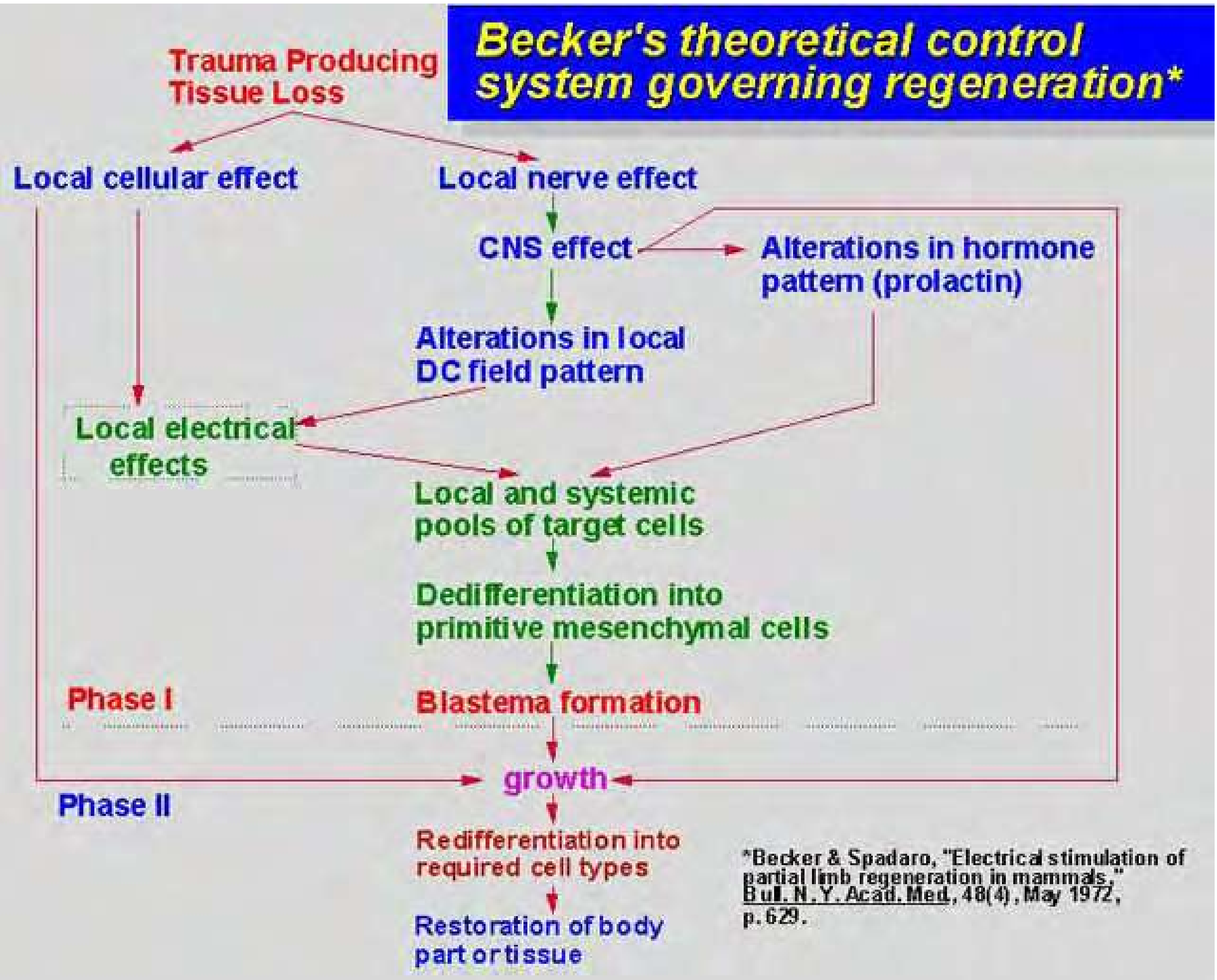

*Becker's theoretical control system governing regeneration**

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N. Y. Acad. Med., 48(4), May 1972, p. 629.

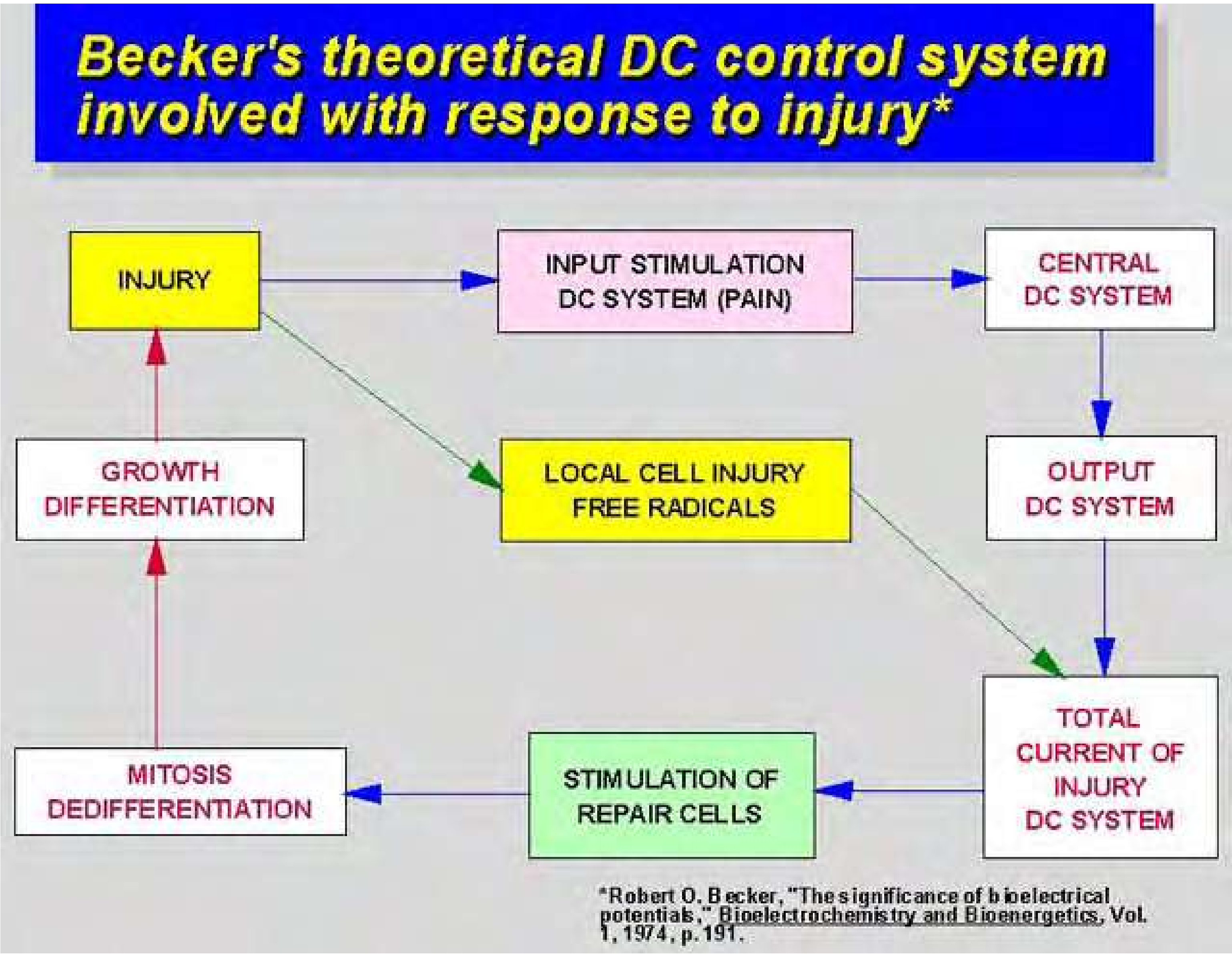

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION
DC SYSTEM (PAIN)

CENTRAL
DC SYSTEM

GROWTH
DIFFERENTIATION

LOCAL CELL INJURY
FREE RADICALS

OUTPUT
DC SYSTEM

MITOSIS
DEDIFFERENTIATION

STIMULATION OF
REPAIR CELLS

TOTAL
CURRENT OF
INJURY
DC SYSTEM

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p.191.

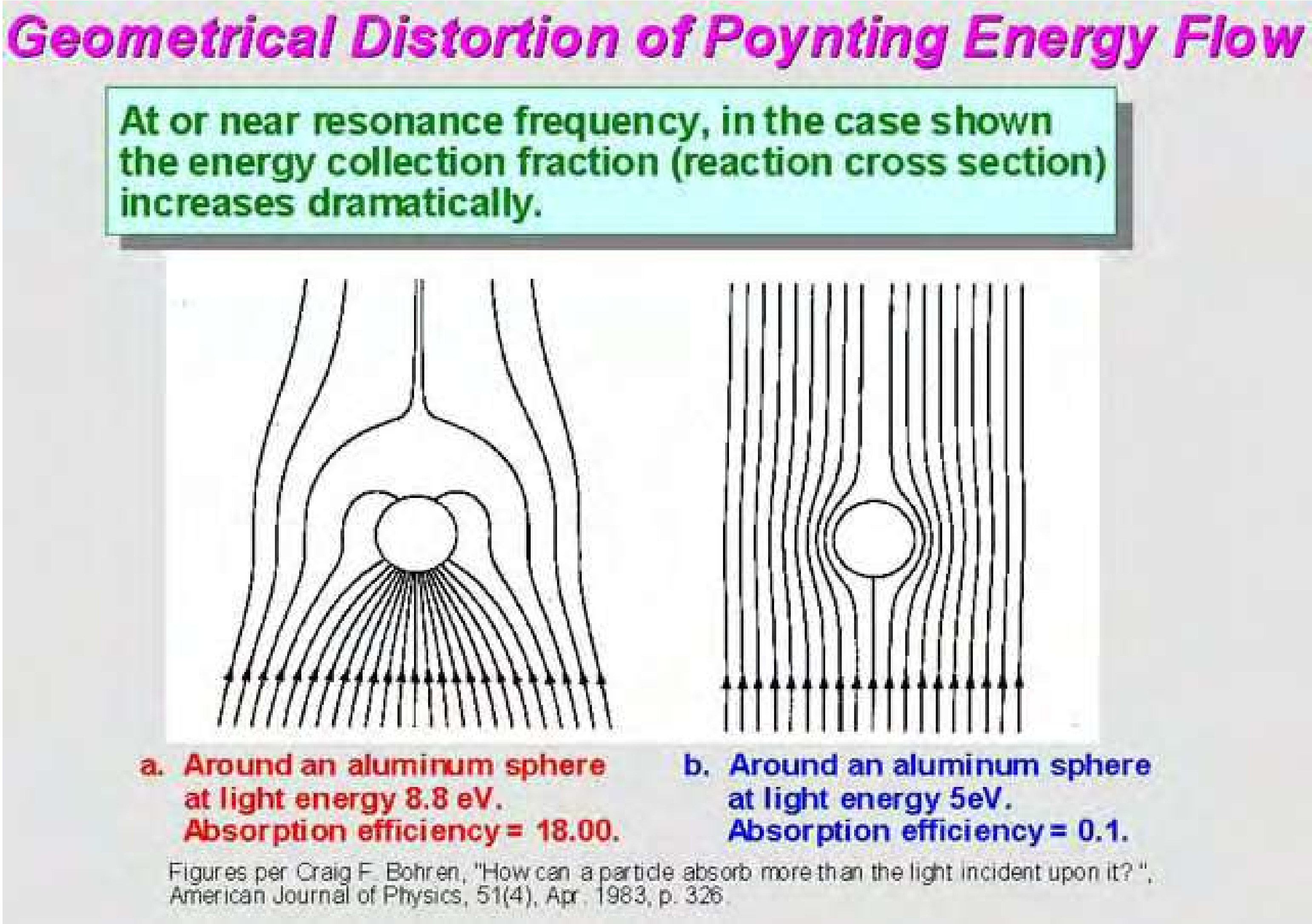

Geometrical Distortion of Poynting Energy Flow

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

a. Around an aluminum sphere at light energy 8.8 eV. Absorption efficiency = 18.00.

b. Around an aluminum sphere at light energy 5eV. Absorption efficiency = 0.1.

Figures per Craig F. Bohren, "How can a particle absorb more than the light incident upon it?", American Journal of Physics, 51(4), Apr. 1983, p. 326.

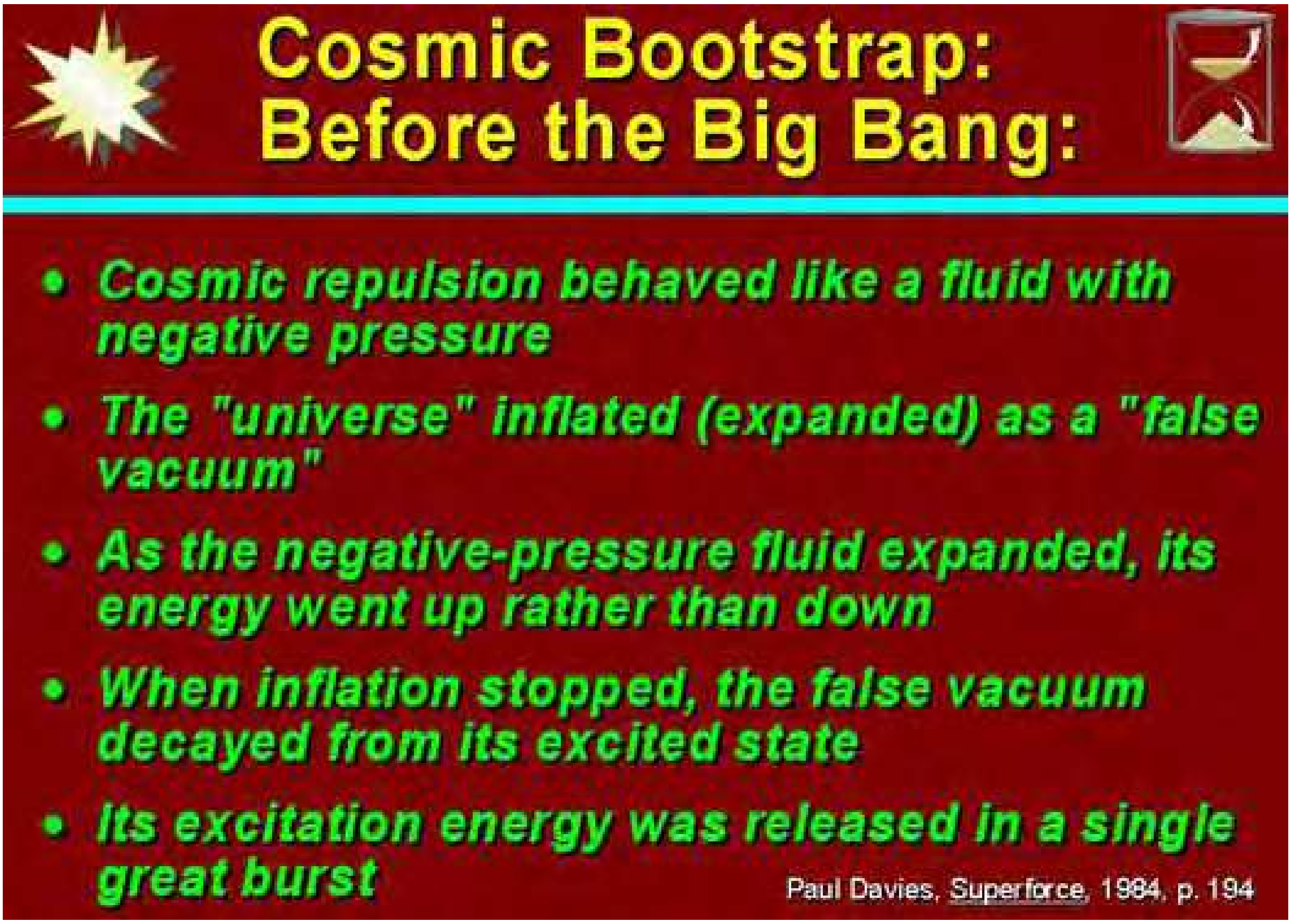

Cosmic Bootstrap: Before the Big Bang:

- *Cosmic repulsion behaved like a fluid with negative pressure*
- *The "universe" inflated (expanded) as a "false vacuum"*
- *As the negative-pressure fluid expanded, its energy went up rather than down*
- *When inflation stopped, the false vacuum decayed from its excited state*
- *Its excitation energy was released in a single great burst*

Paul Davies, Superforce, 1984, p. 194

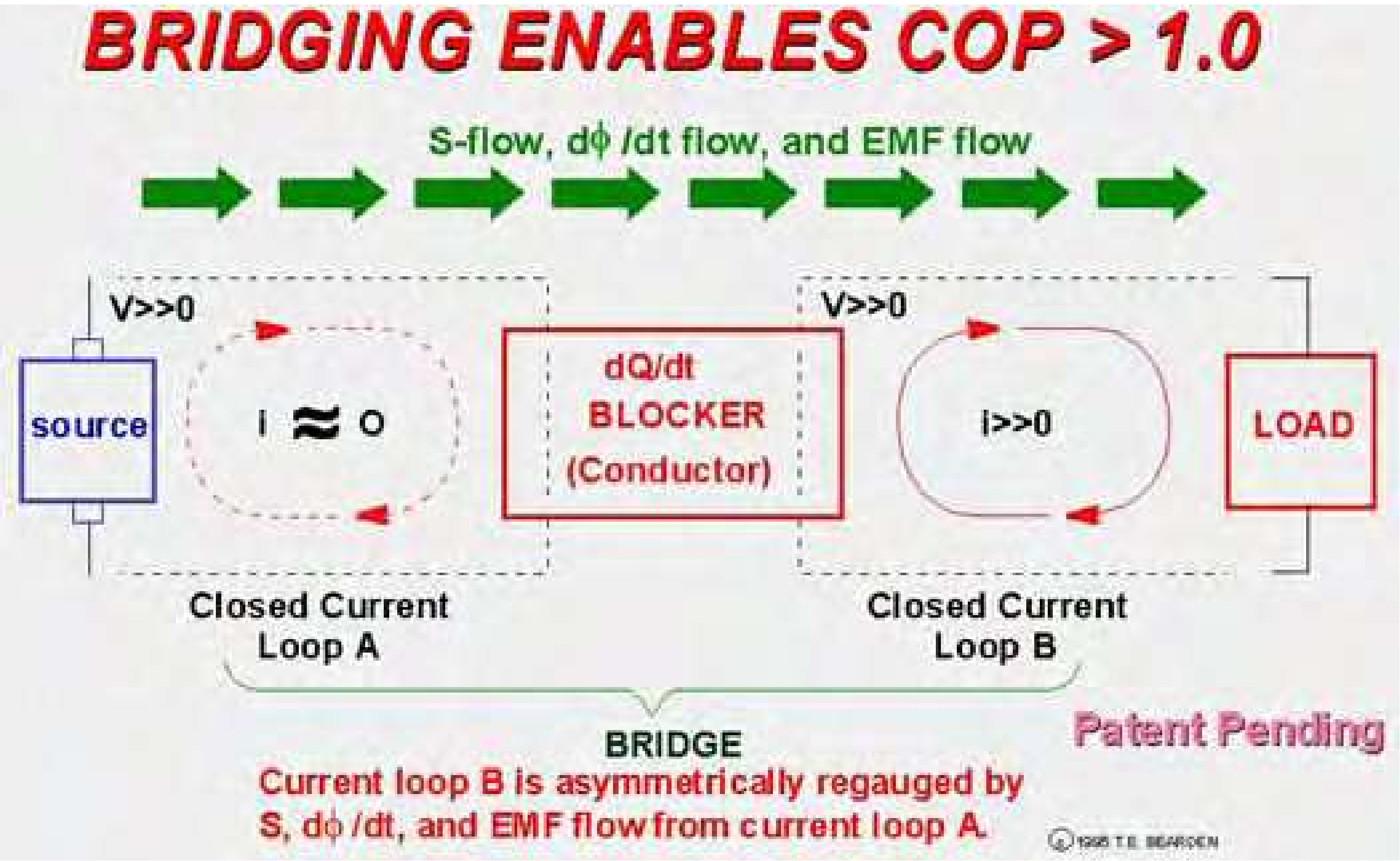

BRIDGING ENABLES COP > 1.0

S-flow, dφ /dt flow, and EMF flow
V>>0
source
i ≈ 0
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD
Closed Current
Loop A
Closed Current
Loop B
BRIDGE
Patent Pending

Current loop B is asymmetrically regauged by
S, dφ /dt, and EMF flow from current loop A.

©1998 T.E. BEARDEN

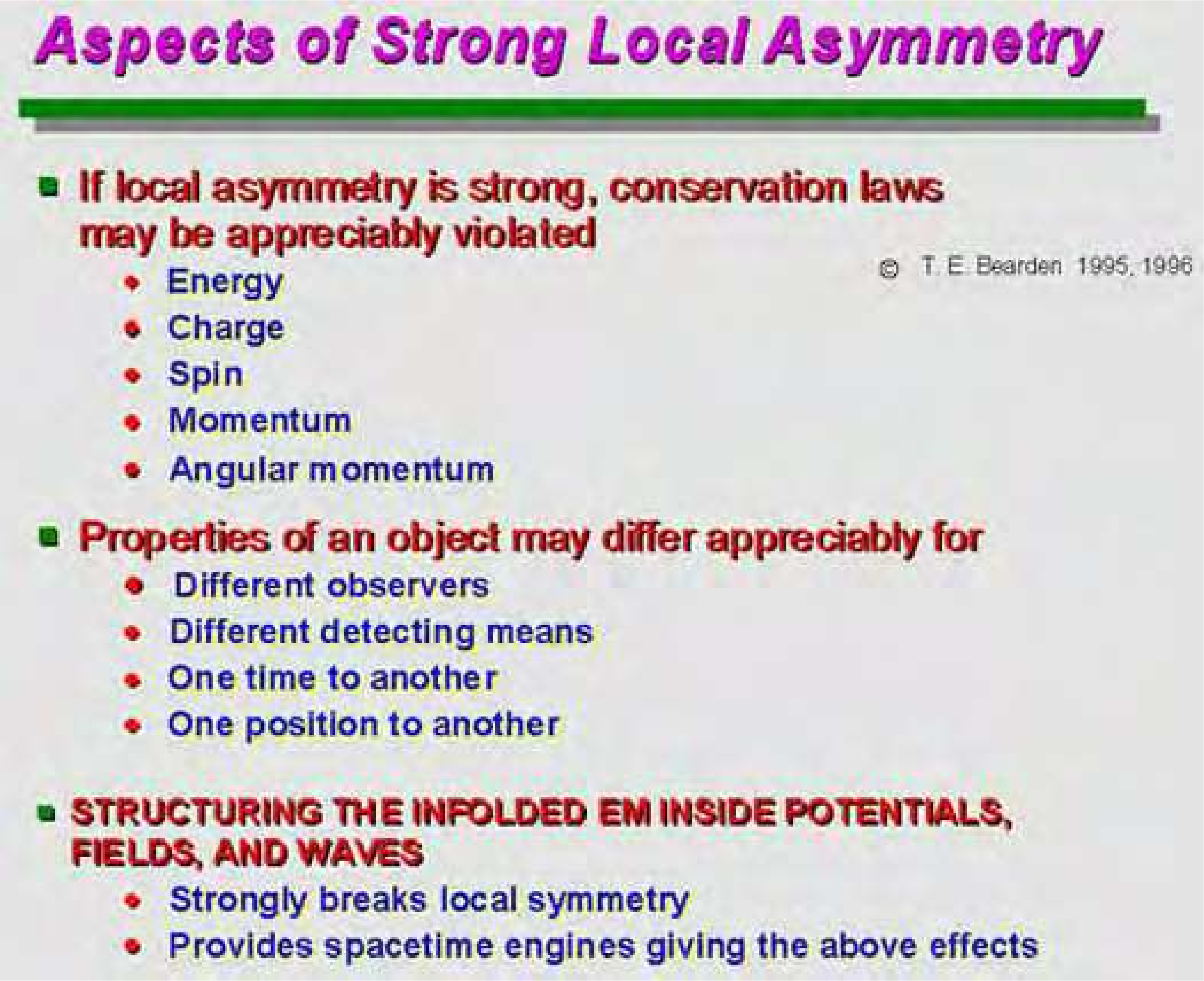

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
    - Energy
    - Charge
    - Spin
    - Momentum
    - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
    - Different observers
    - Different detecting means
    - One time to another
    - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
    - Strongly breaks local symmetry
    - Provides spacetime engines giving the above effects

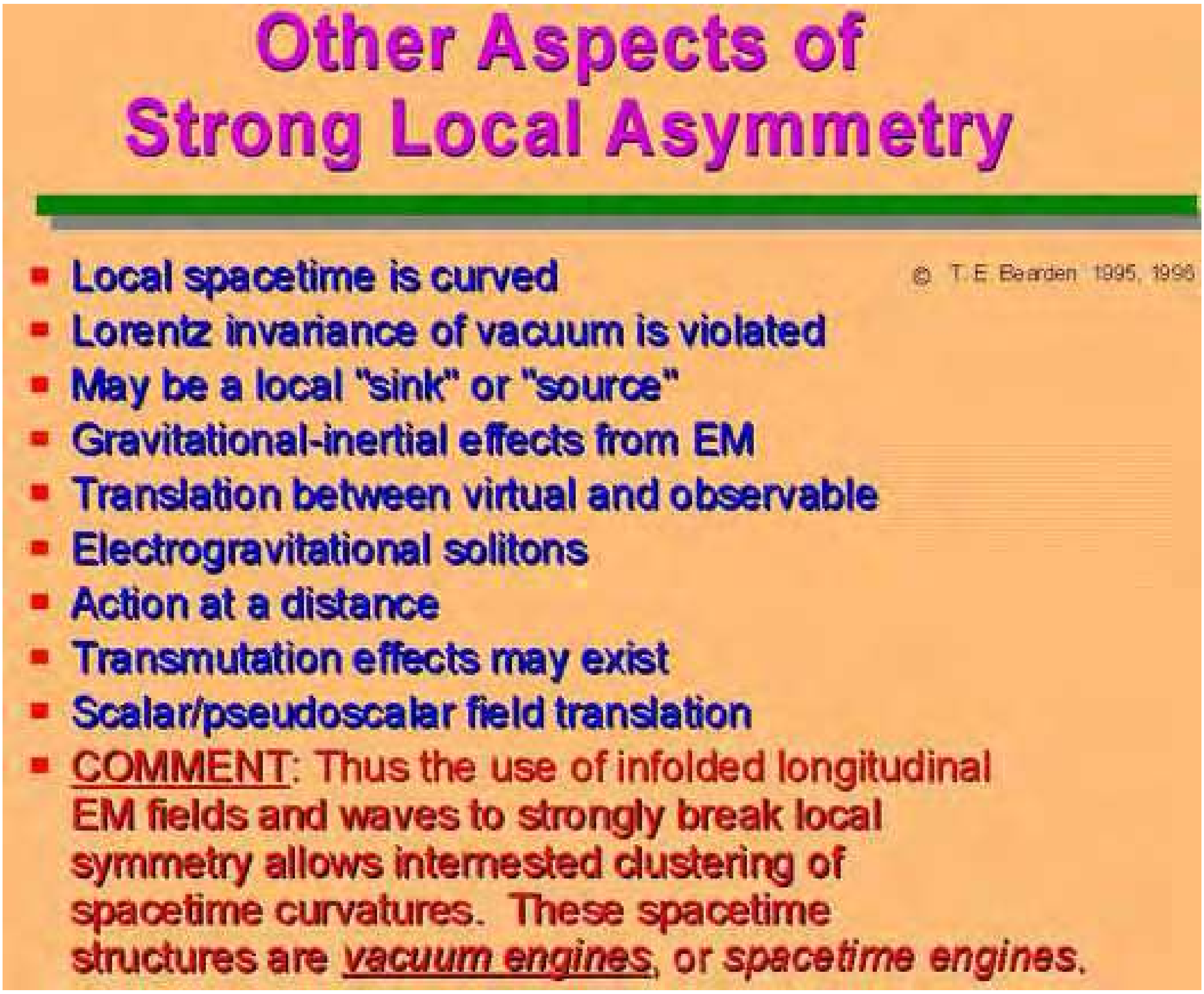

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational-inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

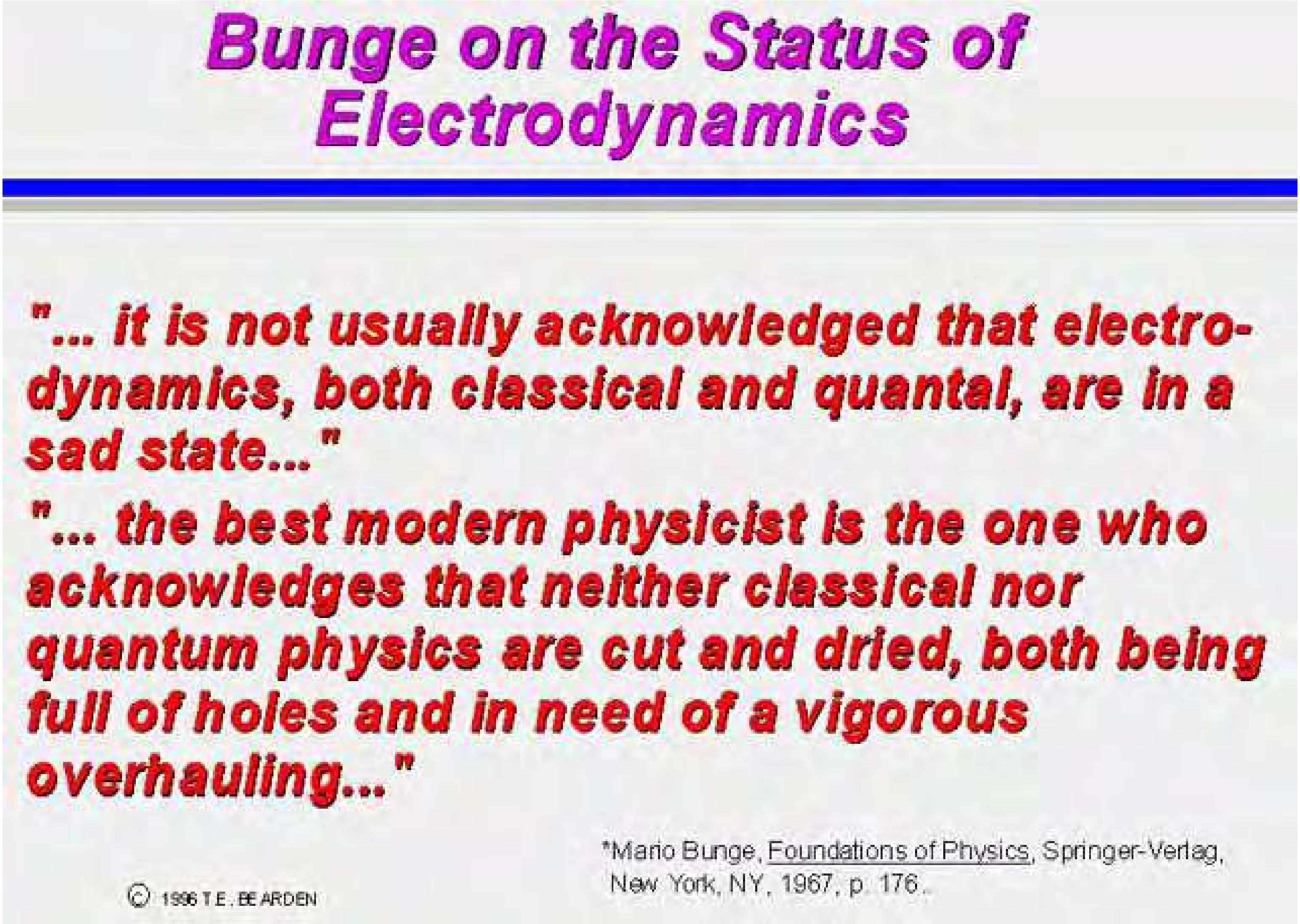

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176.

© 1996 T.E. BEARDEN

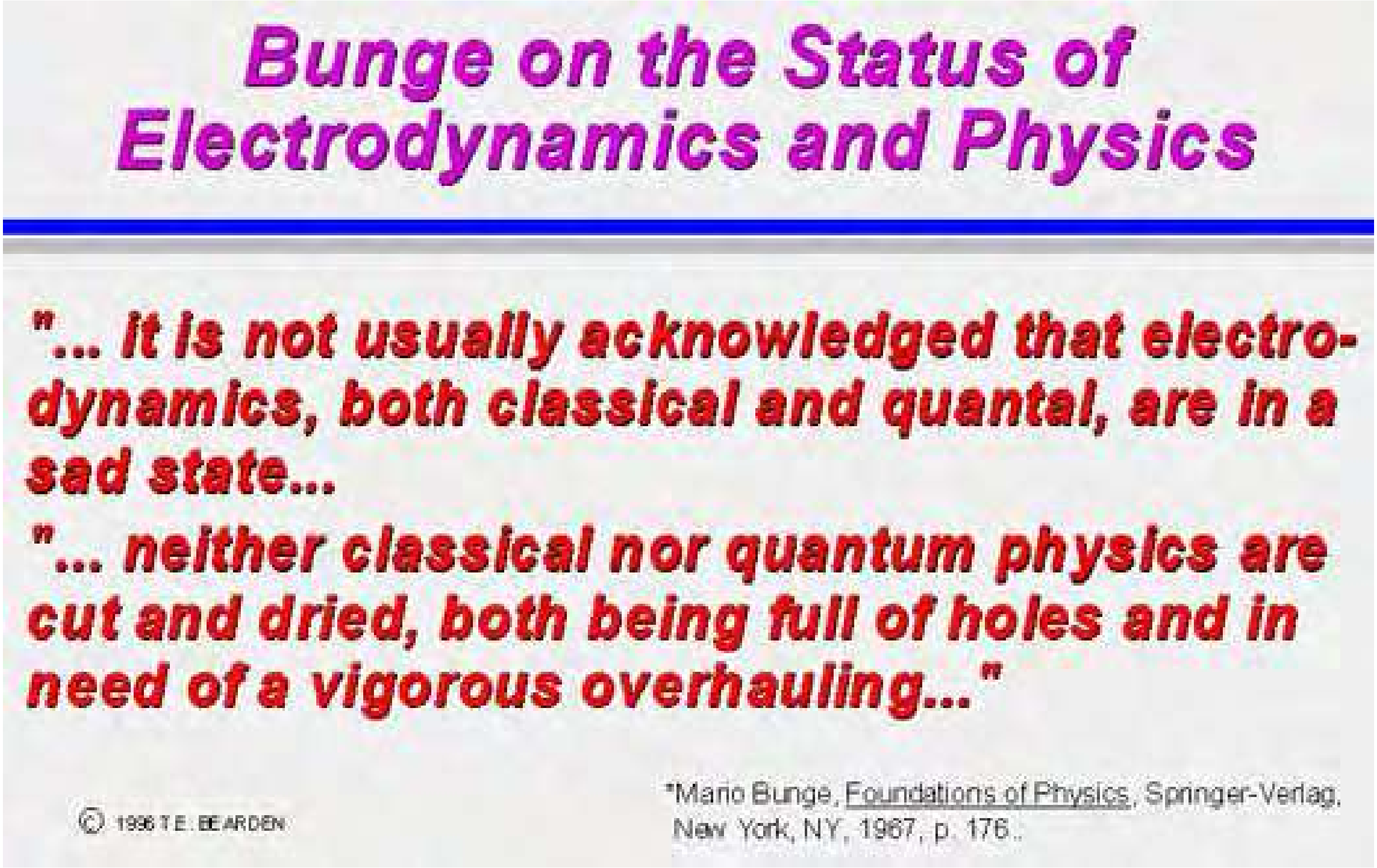

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

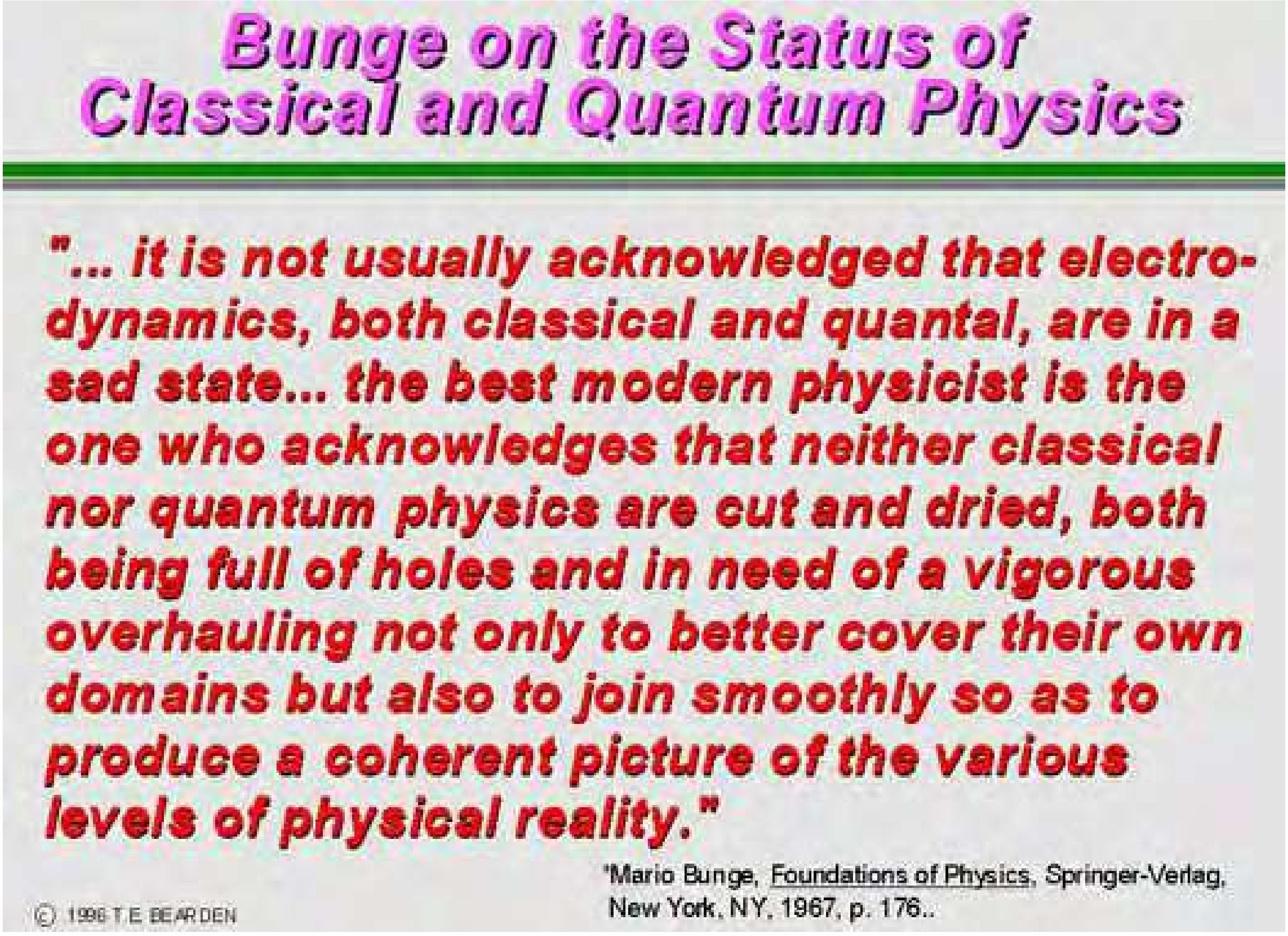

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

'Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

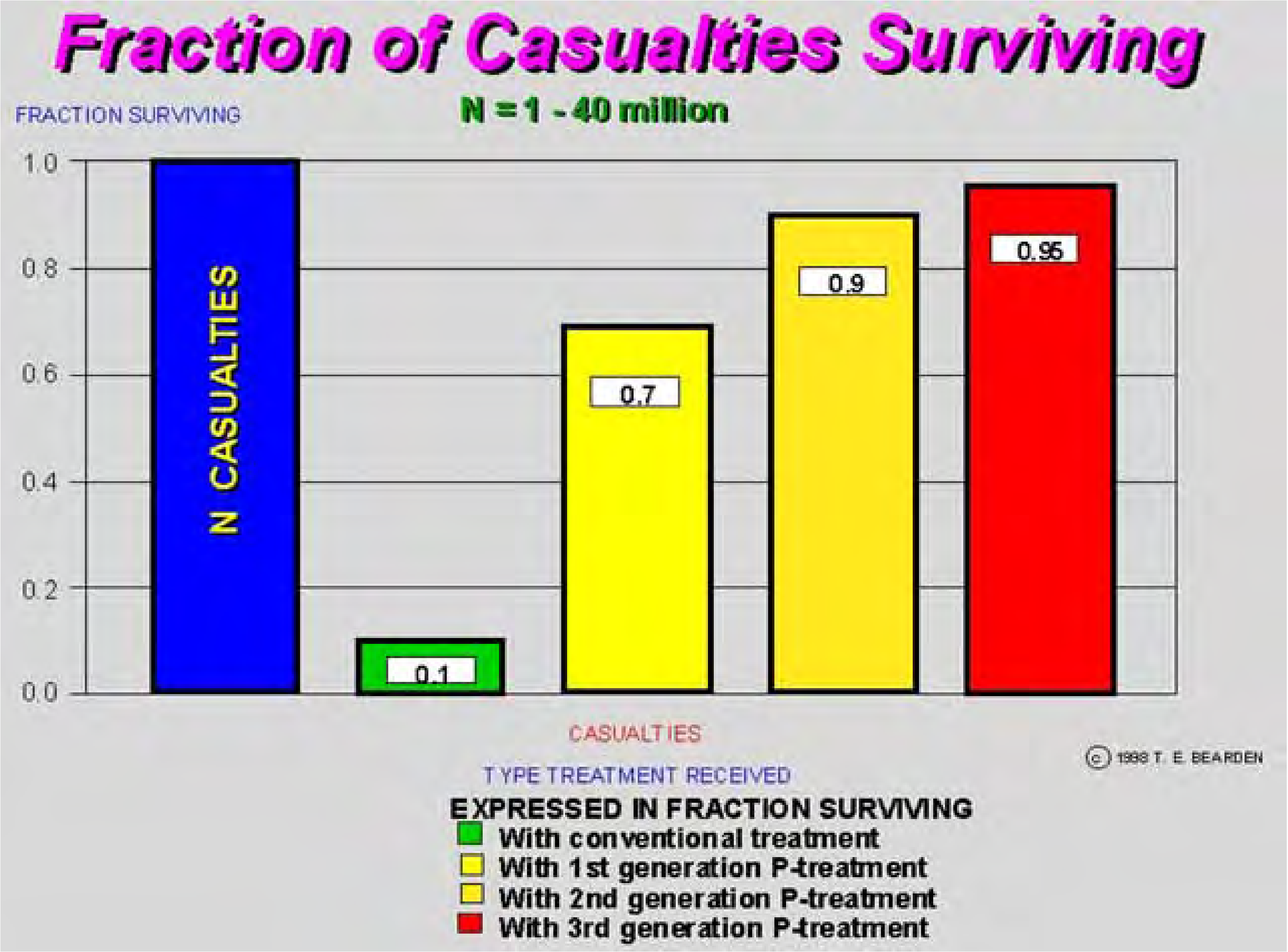
Fraction of Casualties Surviving
FRACTION SURVIVING
N = 1 - 40 million
1.0
0.8
0.6
0.4
0.2
0.0
N CASUALTIES
0.1
0.7
0.9
0.95
CASUALTIES
TYPE TREATMENT RECEIVED
© 1998 T. E. BEARDEN
EXPRESSED IN FRACTION SURVIVING
🟩 With conventional treatment
🟨 With 1st generation P-treatment
🟨 With 2nd generation P-treatment
🟥 With 3rd generation P-treatment

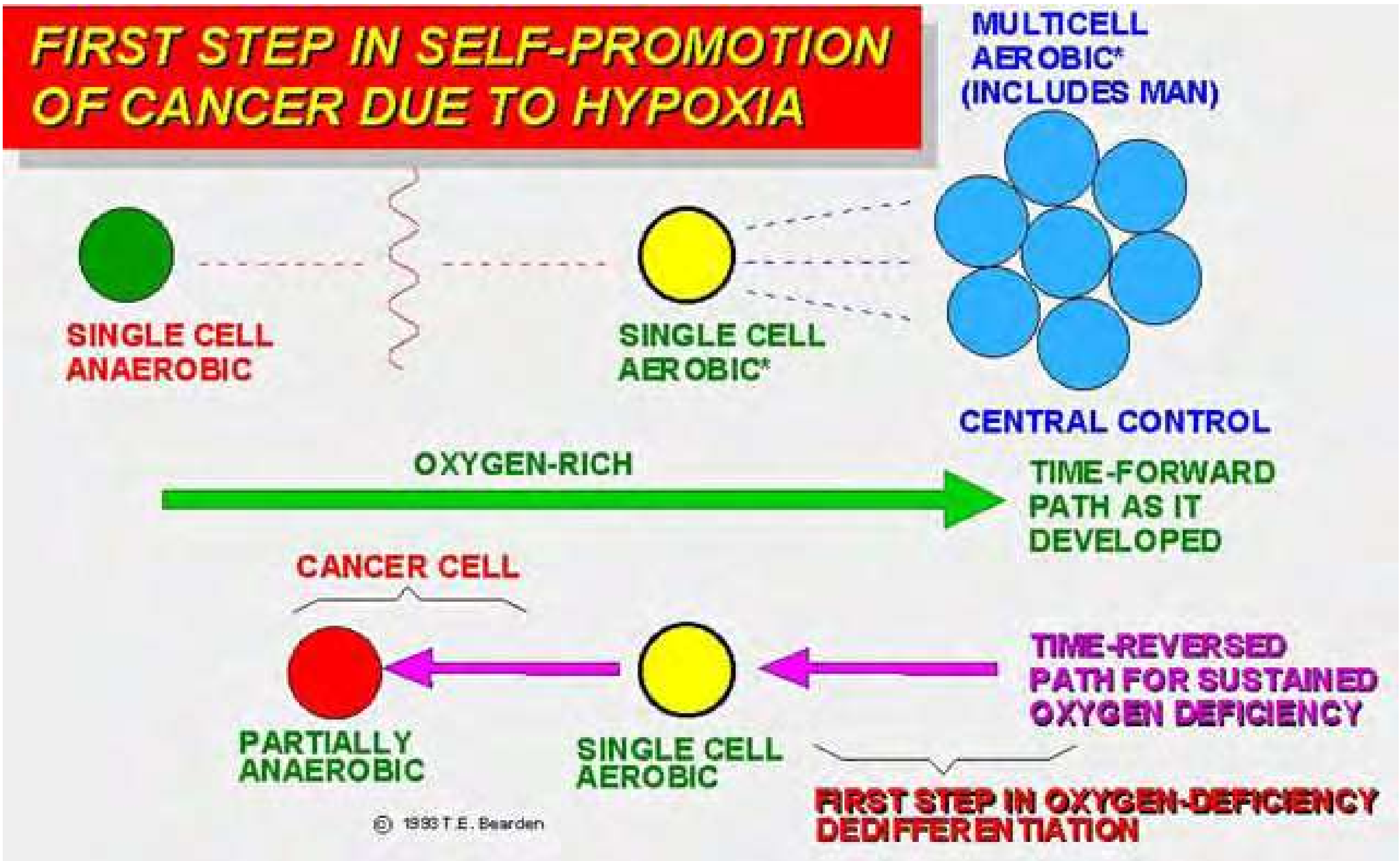

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

MULTICELL
AEROBIC*
(INCLUDES MAN)

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

© 1993 T.E. Bearden

Differentiation and Dedifferentiation Timelines

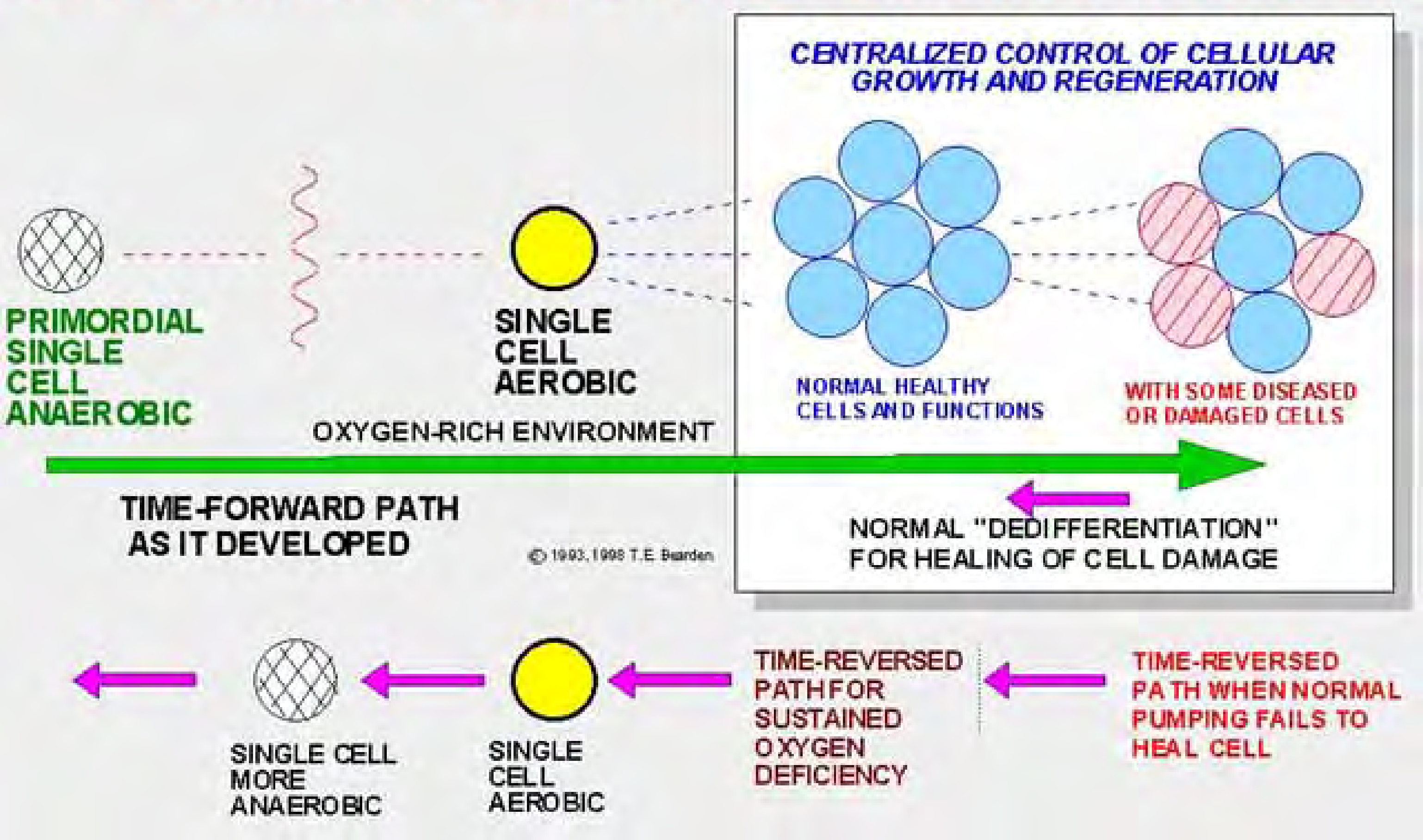

CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION

PRIMORDIAL SINGLE CELL ANAEROBIC

SINGLE CELL AEROBIC

OXYGEN-RICH ENVIRONMENT

NORMAL HEALTHY CELLS AND FUNCTIONS

WITH SOME DISEASED OR DAMAGED CELLS

TIME-FORWARD PATH AS IT DEVELOPED

NORMAL "DEDIFFERENTIATION" FOR HEALING OF CELL DAMAGE

© 1993, 1998 T.E. Bearden

SINGLE CELL MORE ANAEROBIC

SINGLE CELL AEROBIC

TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY

TIME-REVERSED PATH WHEN NORMAL PUMPING FAILS TO HEAL CELL

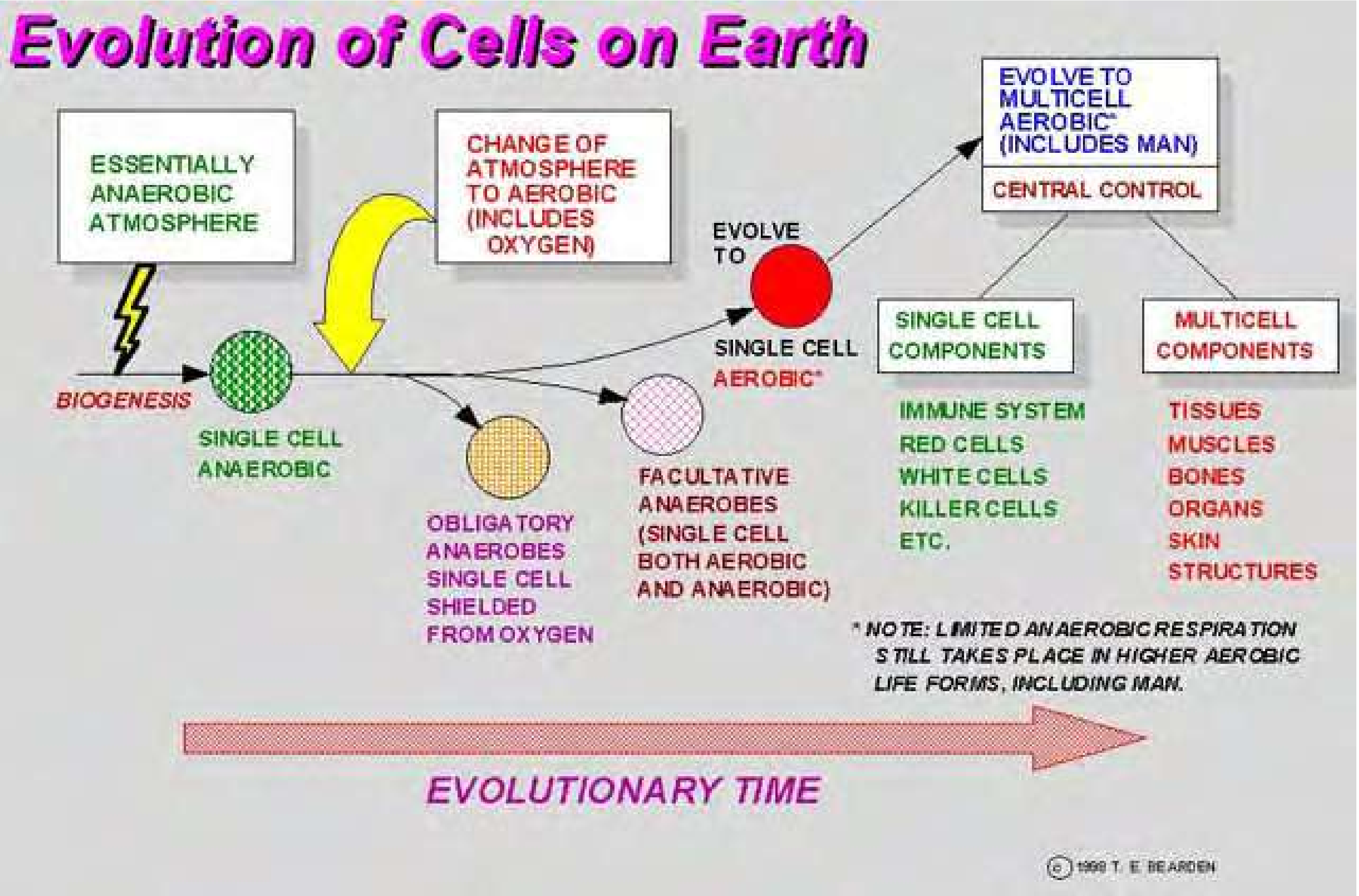
Evolution of Cells on Earth
ESSENTIALLY ANAEROBIC ATMOSPHERE
CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)
EVOLVE TO
EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)
CENTRAL CONTROL
BIOGENESIS
SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC*
SINGLE CELL COMPONENTS
MULTICELL COMPONENTS
OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN
FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)
IMMUNE SYSTEM
RED CELLS
WHITE CELLS
KILLER CELLS
ETC.
TISSUES
MUSCLES
BONES
ORGANS
SKIN
STRUCTURES
* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.
EVOLUTIONARY TIME
(c) 1999 T. E. BEARDEN

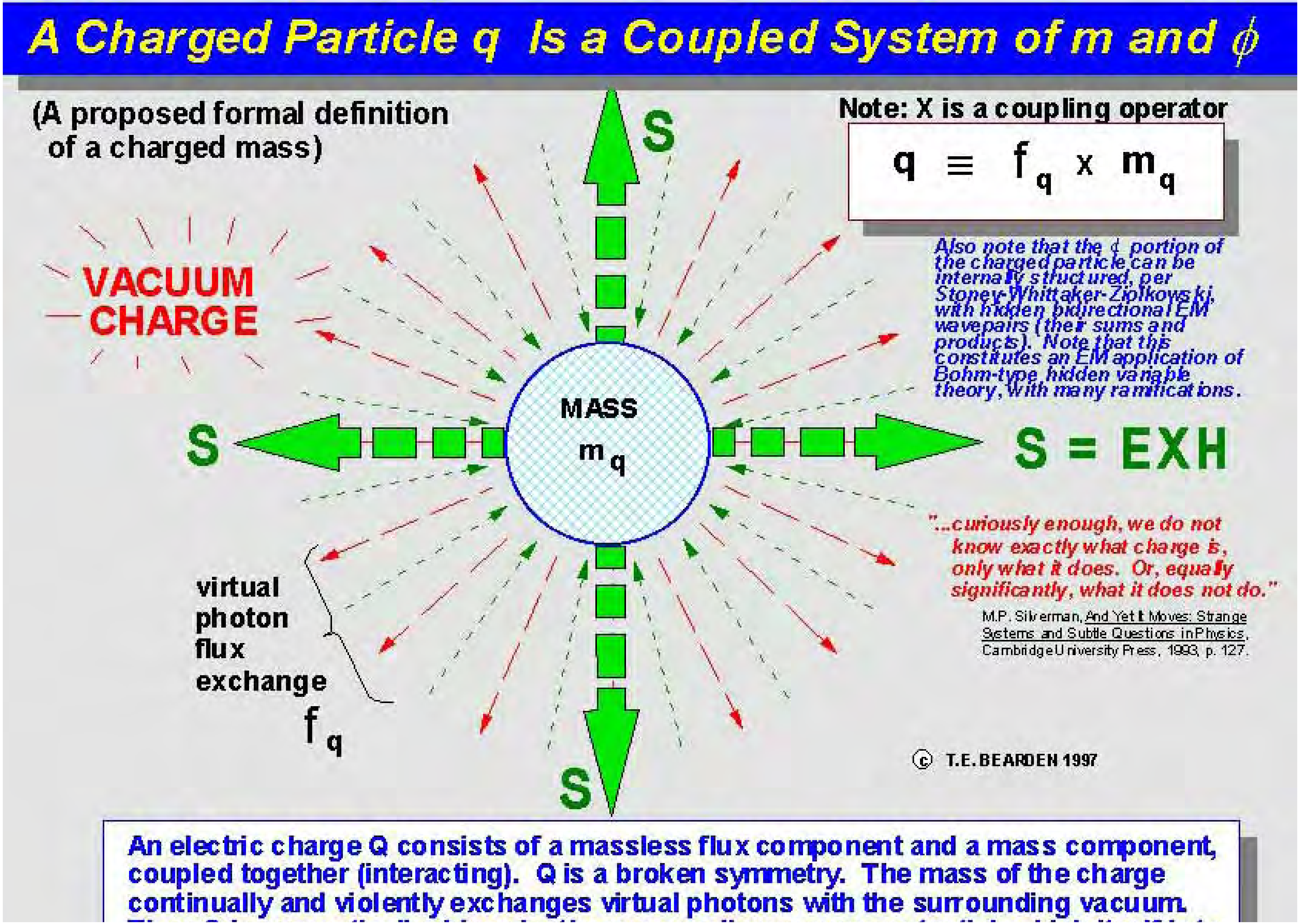

A Charged Particle q Is a Coupled System of m and ϕ

(A proposed formal definition of a charged mass)

Note: X is a coupling operator

$$q \equiv f_q \times m_q$$

Also note that the ϕ portion of the charged particle can be internally structured, per Stoney-Whittaker-Ziolkowski, with hidden bidirectional EM wavepairs (their sums and products). Note that this constitutes an EM application of Bohm-type hidden variable theory, with many ramifications.

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

ⓒ T.E. BEARDEN 1997

An electric charge Q consists of a massless flux component and a mass component, coupled together (interacting). Q is a broken symmetry. The mass of the charge continually and violently exchanges virtual photons with the surrounding vacuum.

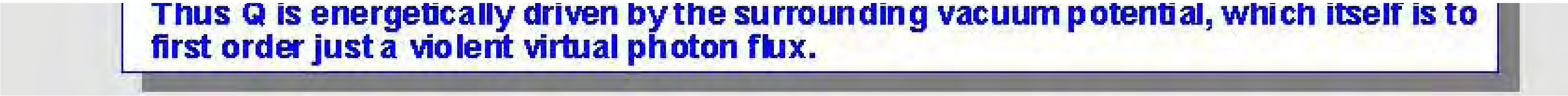

Thus Q is energetically driven by the surrounding vacuum potential, which itself is to first order just a violent virtual photon flux.

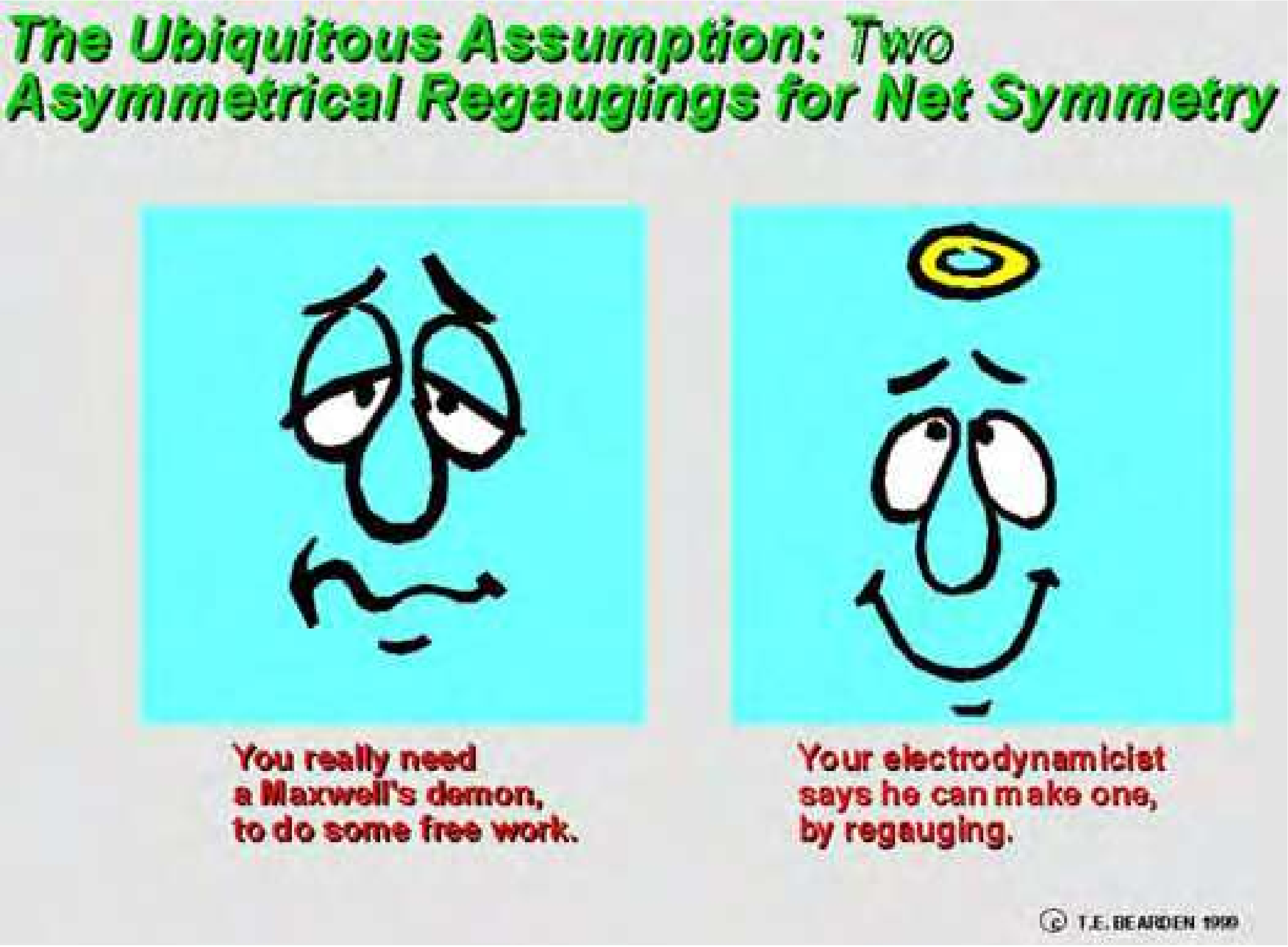

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

You really need
a Maxwell's demon,
to do some free work.

Your electrodynamicist
says he can make one,
by regauging.

© T.E. BEARDEN 1999

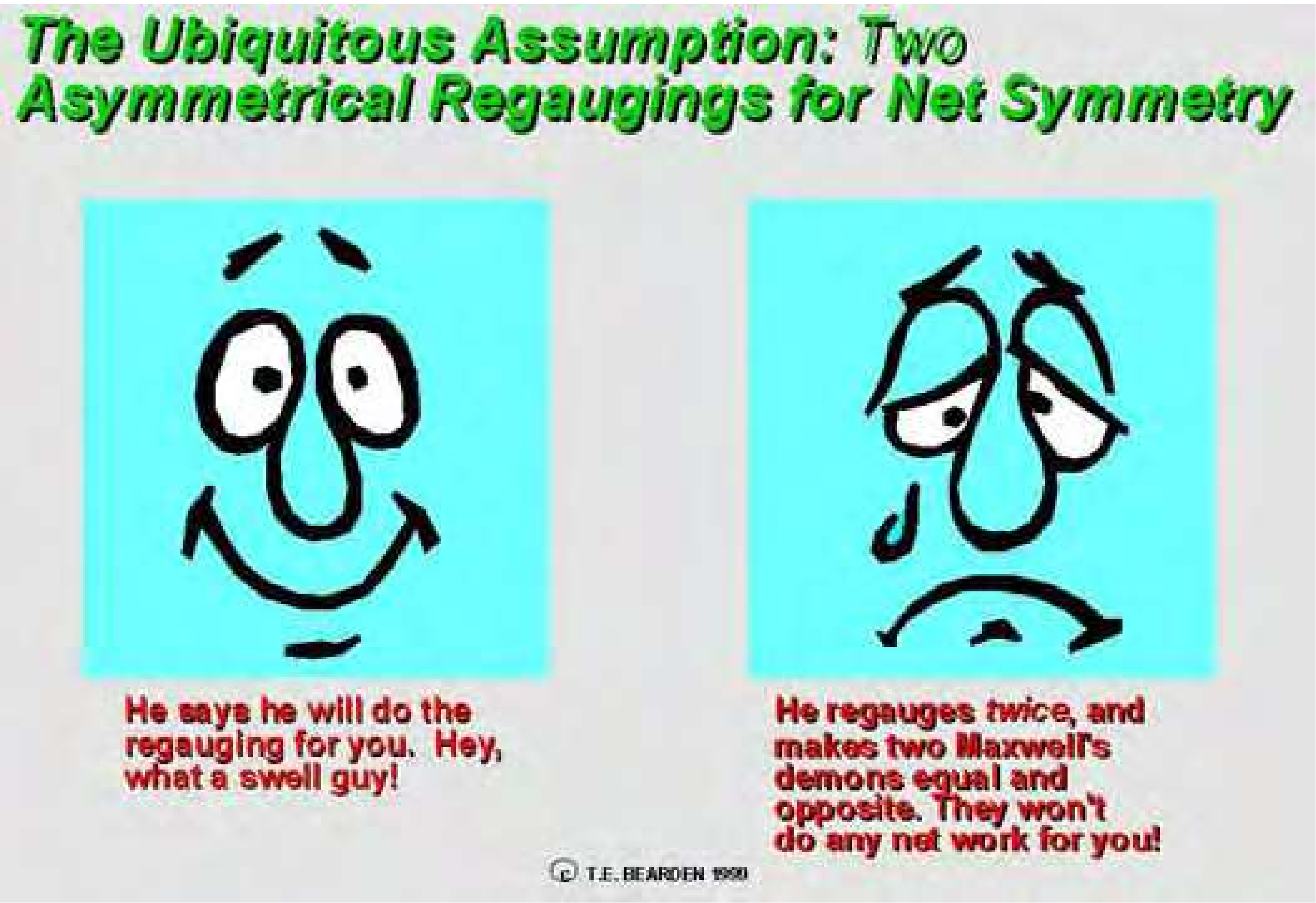

The Ubiquitous Assumption: *Two* Asymmetrical Regaugings for Net Symmetry

He says he will do the regauging for you. Hey, what a swell guy!

He regauges *twice*, and makes two Maxwell's demons equal and opposite. They won't do any net work for you!

© T.E. BEARDEN 1999

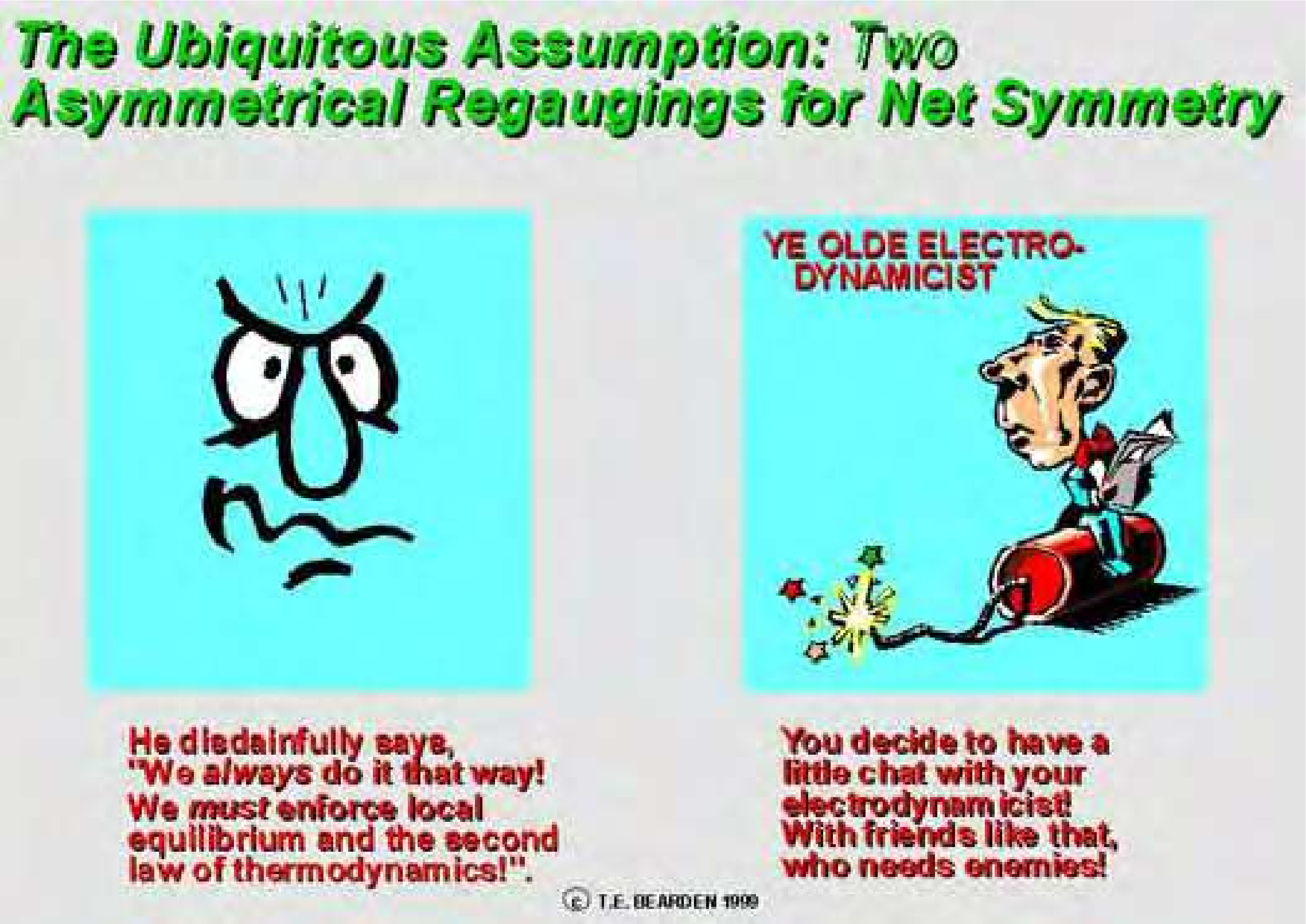

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

He disdainfully says,
"We *always* do it that way!
We *must* enforce local
equilibrium and the second
law of thermodynamics!".

You decide to have a
little chat with your
electrodynamicist!
With friends like that,
who needs enemies!

© T.E. BEARDEN 1999

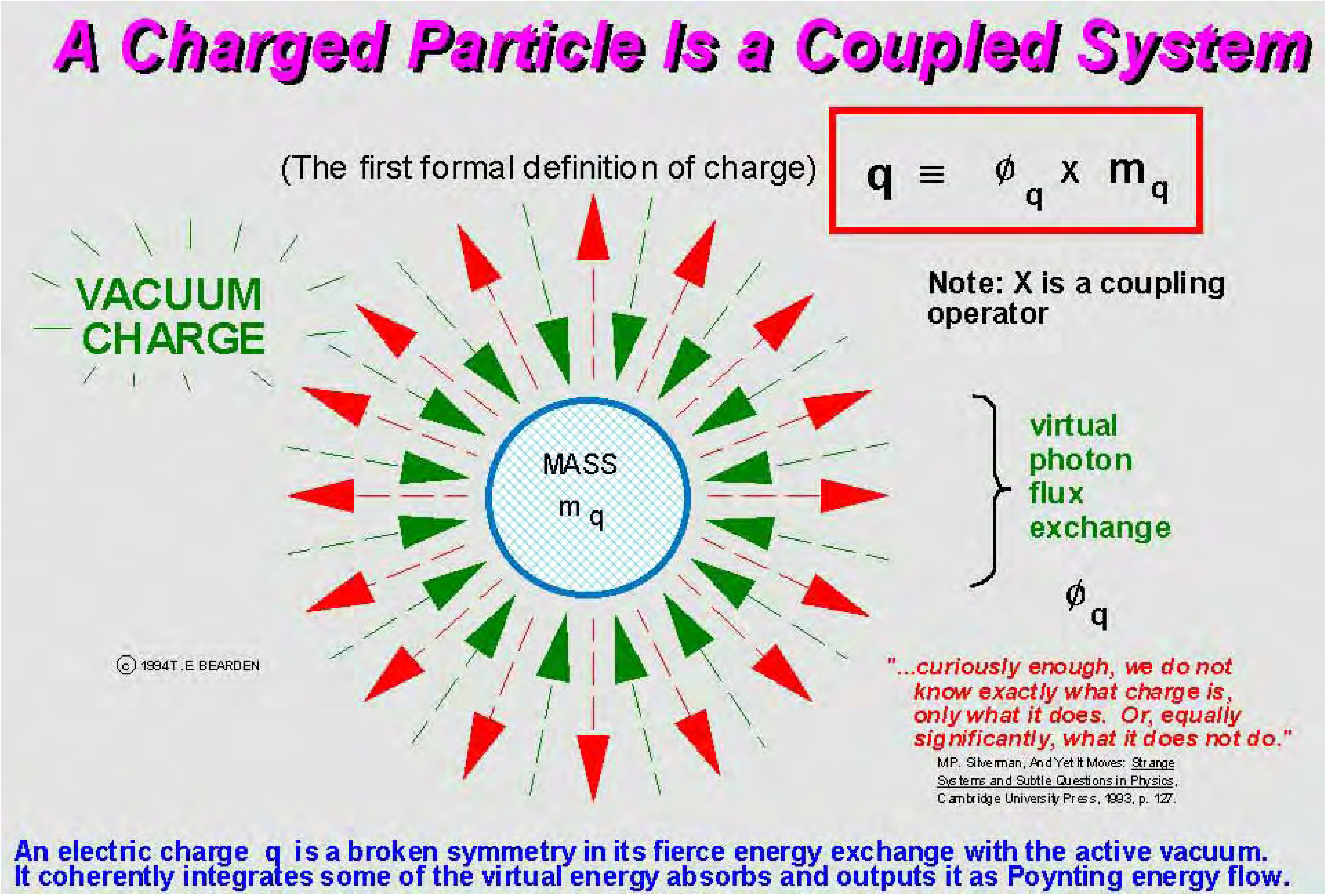

A Charged Particle Is a Coupled System

(The first formal definition of charge)

$$\boxed{q \equiv \phi_q \times m_q}$$

Note: X is a coupling operator

© 1994 T.E BEARDEN

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge q is a broken symmetry in its fierce energy exchange with the active vacuum. It coherently integrates some of the virtual energy absorbs and outputs it as Poynting energy flow.

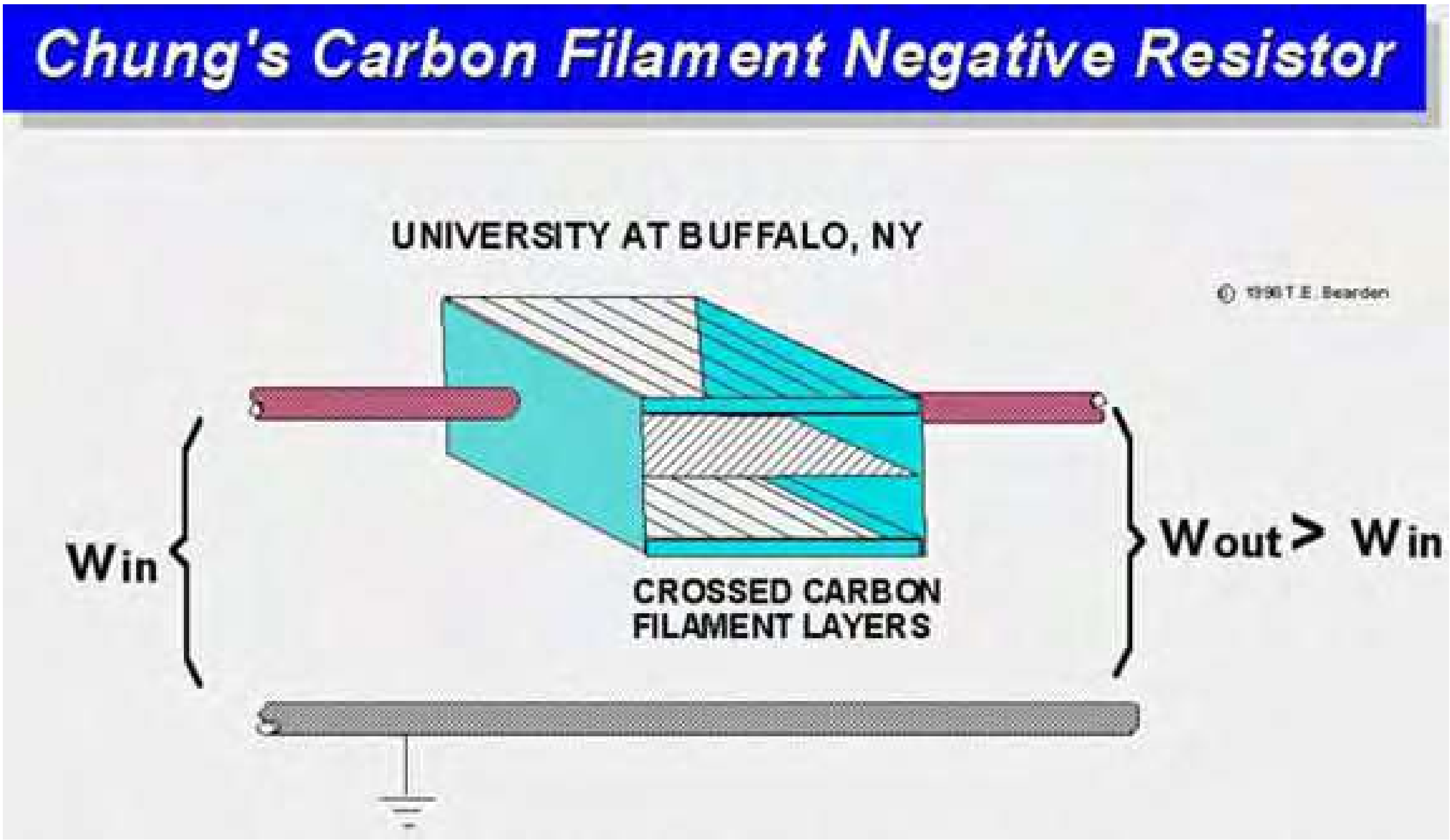

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

© 1998 T.E. Bearden

W_{in}

CROSSED CARBON
FILAMENT LAYERS

$W_{out} > W_{in}$

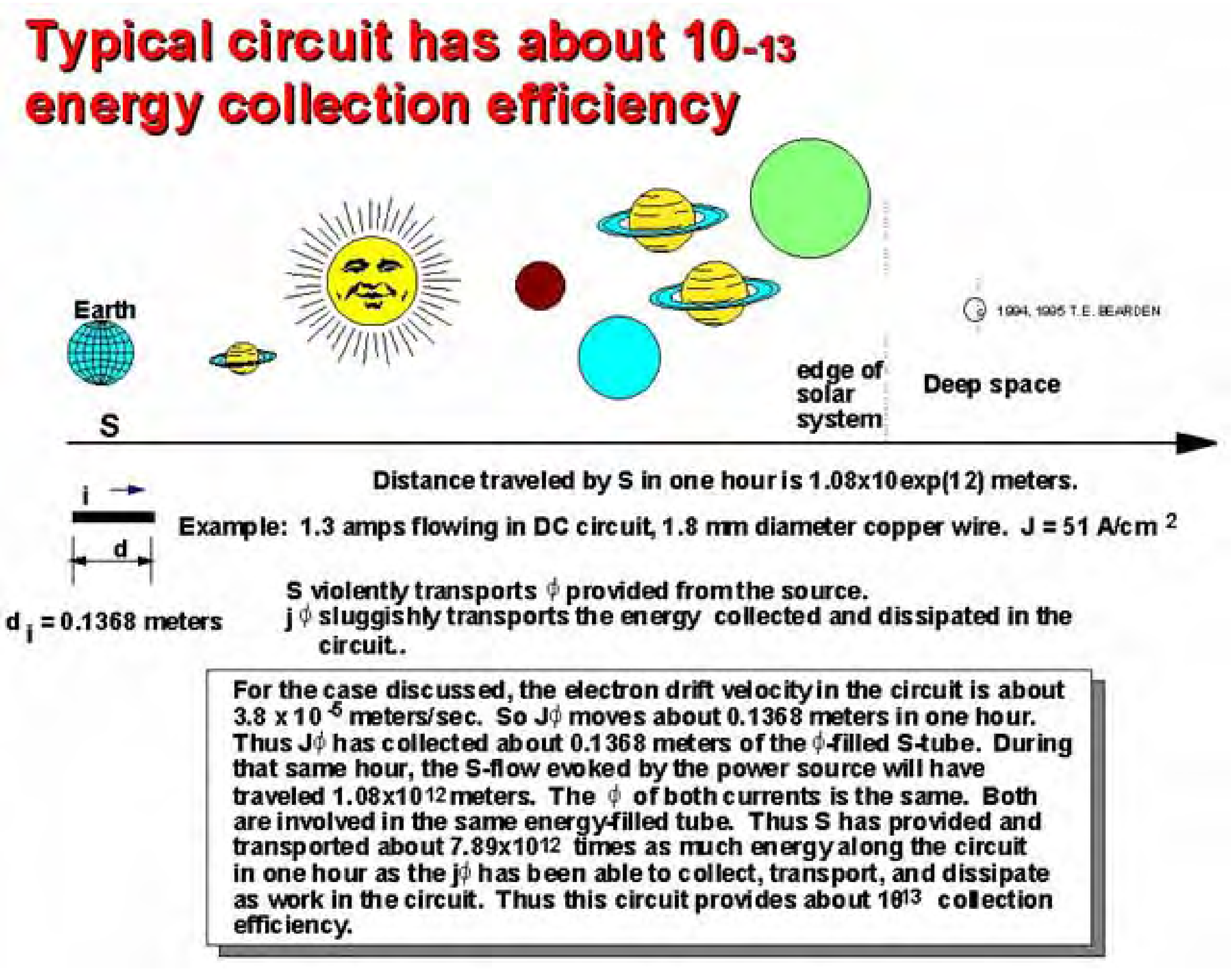

Typical circuit has about 10^{-13} energy collection efficiency

Distance traveled by S in one hour is 1.08x10exp(12) meters.

Example: 1.3 amps flowing in DC circuit, 1.8 mm diameter copper wire. J = 51 A/cm^2

d_i = 0.1368 meters

S violently transports ϕ provided from the source.
jϕ sluggishly transports the energy collected and dissipated in the circuit..

For the case discussed, the electron drift velocity in the circuit is about 3.8 x 10^{-5} meters/sec. So Jϕ moves about 0.1368 meters in one hour. Thus Jϕ has collected about 0.1368 meters of the ϕ-filled S-tube. During that same hour, the S-flow evoked by the power source will have traveled $1.08\text{x}10^{12}$ meters. The ϕ of both currents is the same. Both are involved in the same energy-filled tube. Thus S has provided and transported about $7.89\text{x}10^{12}$ times as much energy along the circuit in one hour as the jϕ has been able to collect, transport, and dissipate as work in the circuit. Thus this circuit provides about 10^{13} collection efficiency.

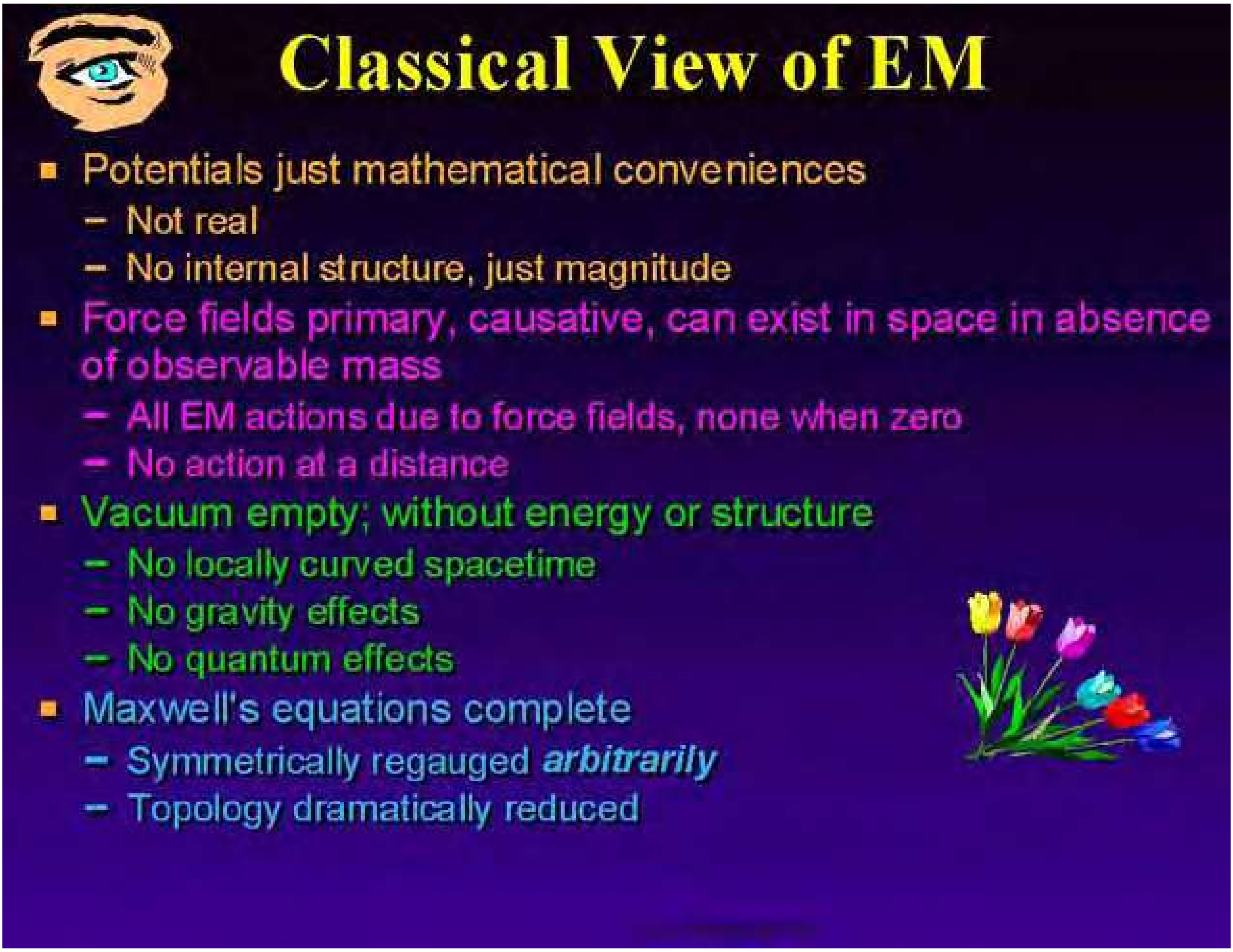

Classical View of EM

- Potentials just mathematical conveniences
  - Not real
  - No internal structure, just magnitude
- Force fields primary, causative, can exist in space in absence of observable mass
  - All EM actions due to force fields, none when zero
  - No action at a distance
- Vacuum empty; without energy or structure
  - No locally curved spacetime
  - No gravity effects
  - No quantum effects
- Maxwell's equations complete
  - Symmetrically regauged ***arbitrarily***
  - Topology dramatically reduced

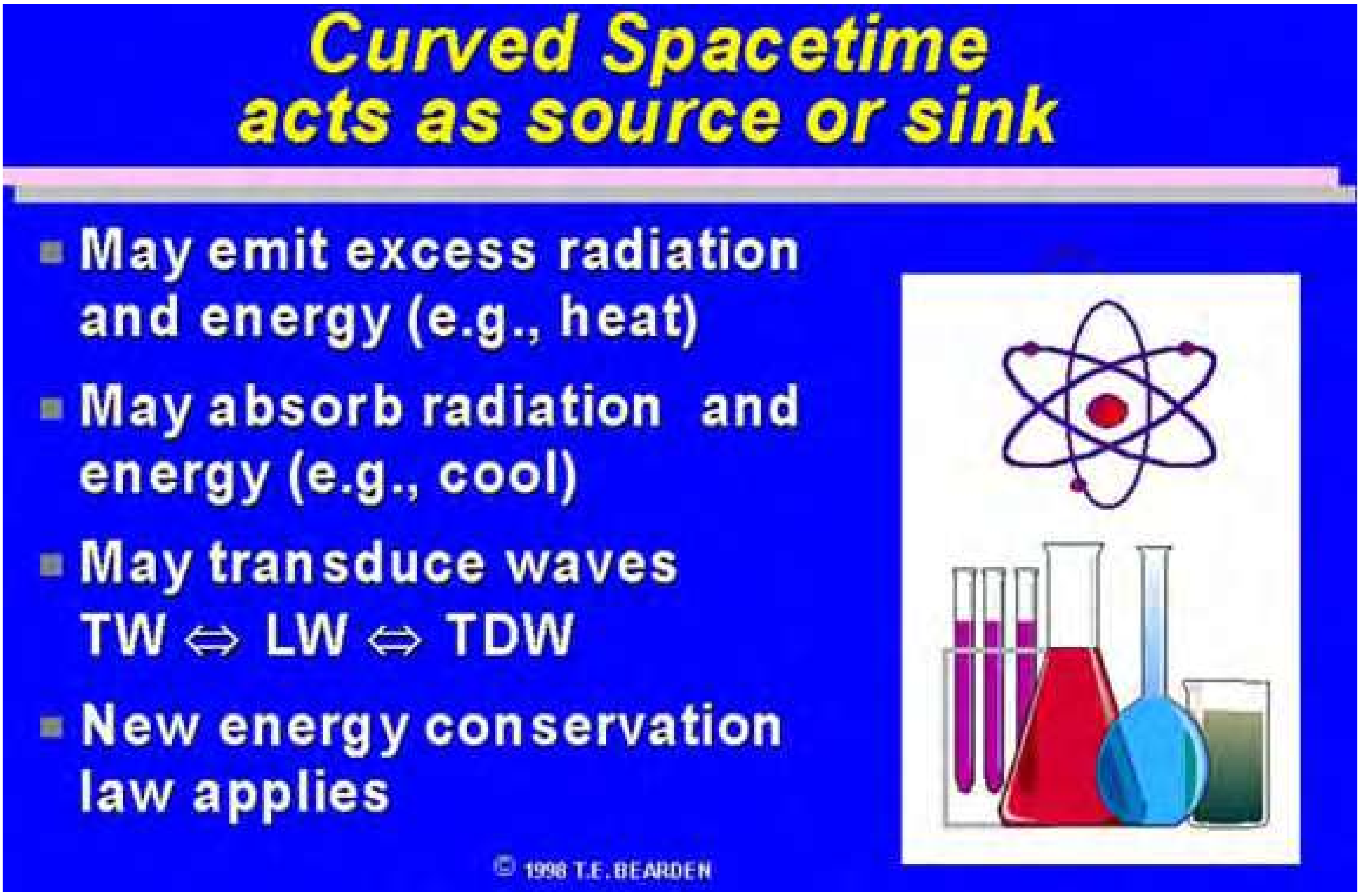

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E.BEARDEN

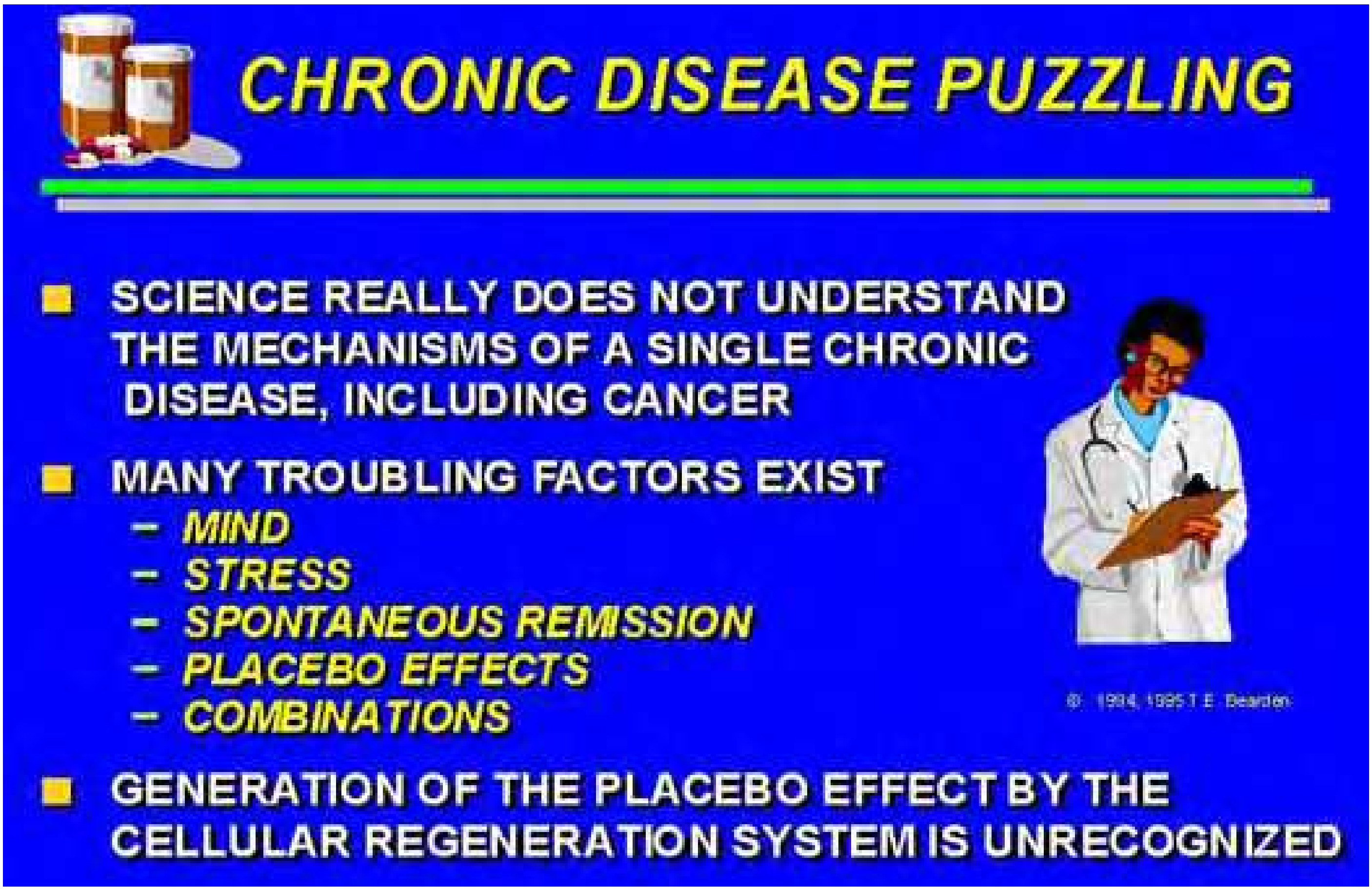

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

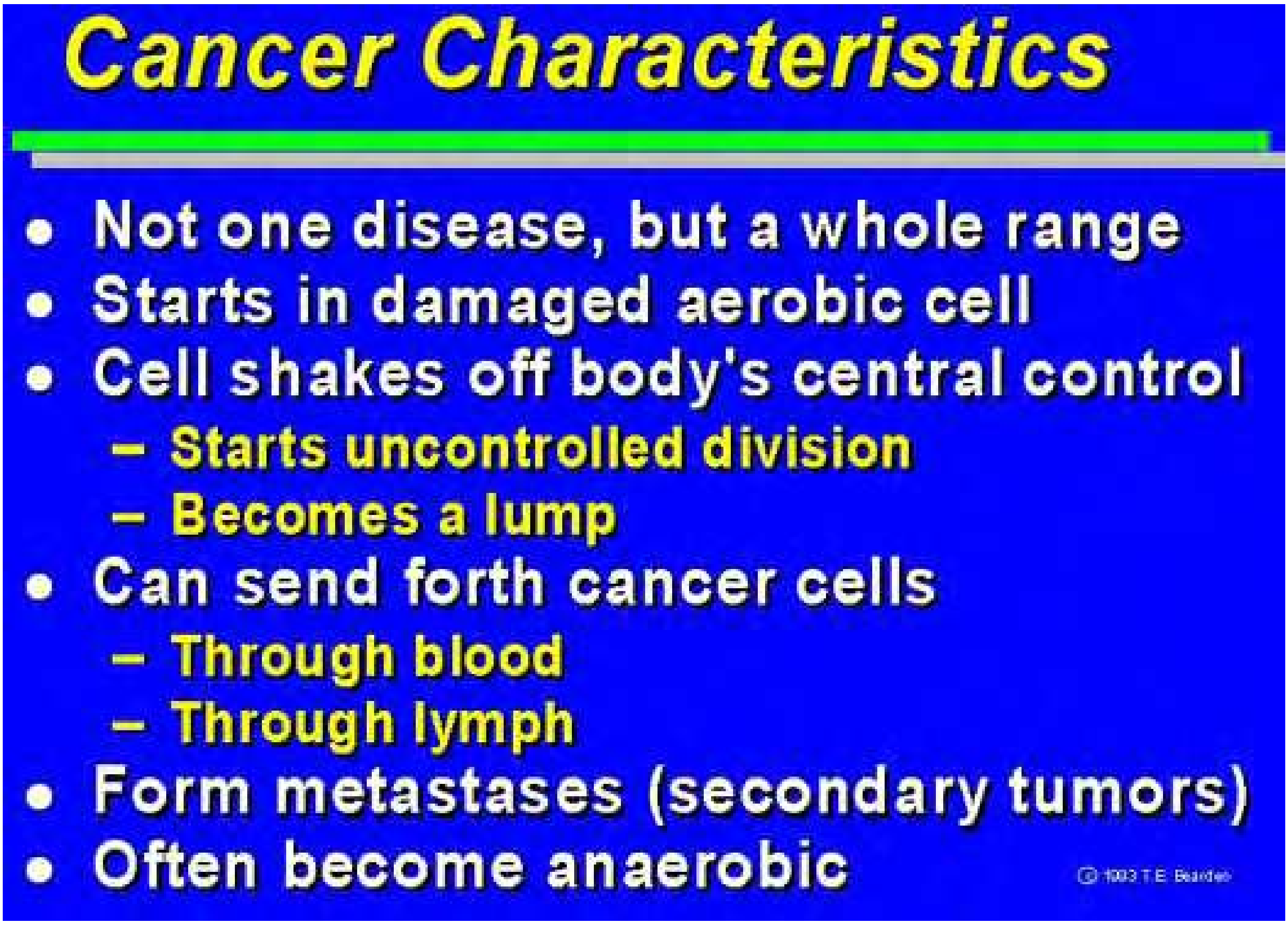

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

© 1993 T.E. Bearden

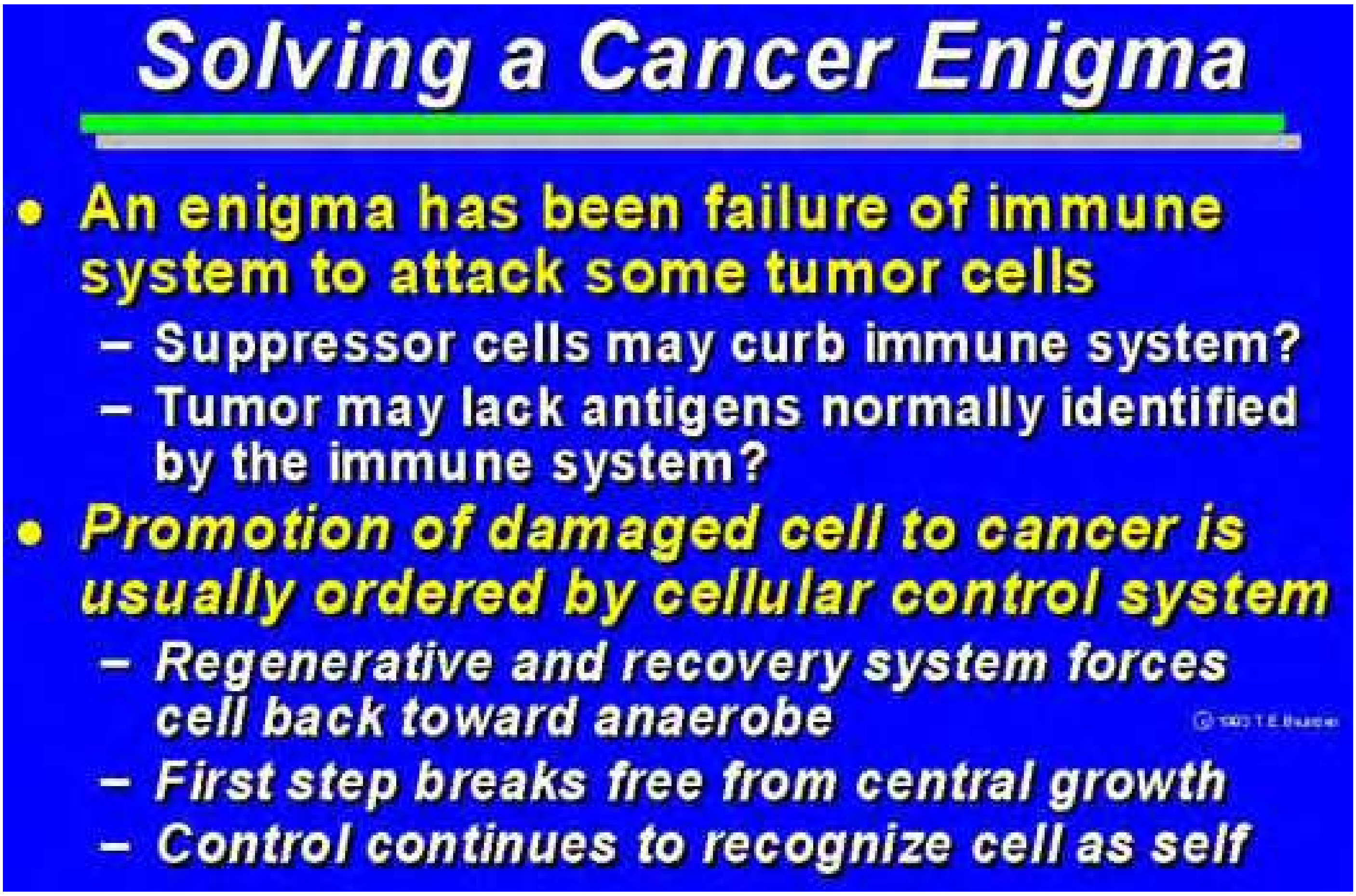

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

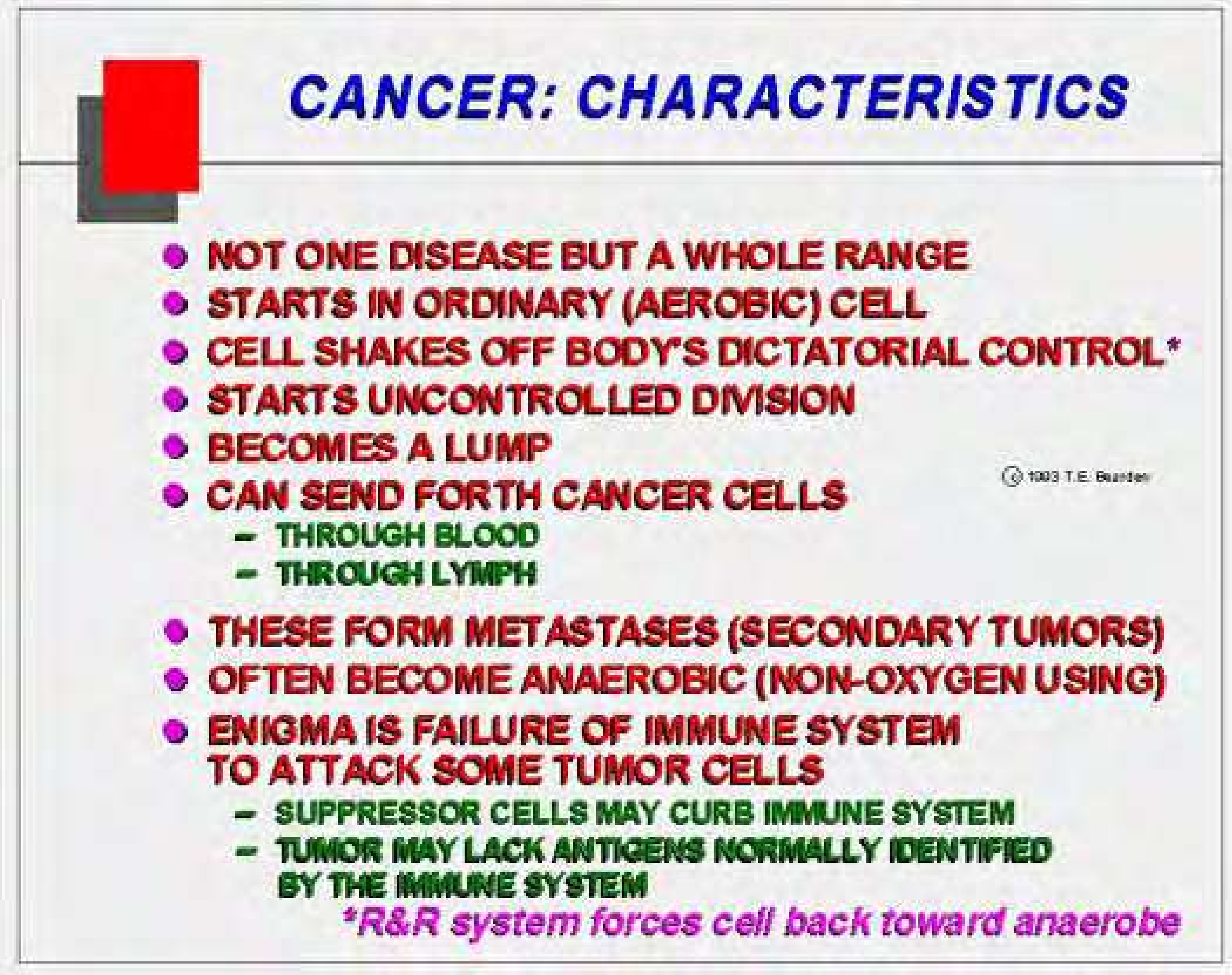

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

*R&R system forces cell back toward anaerobe

© 1993 T.E. Bearden

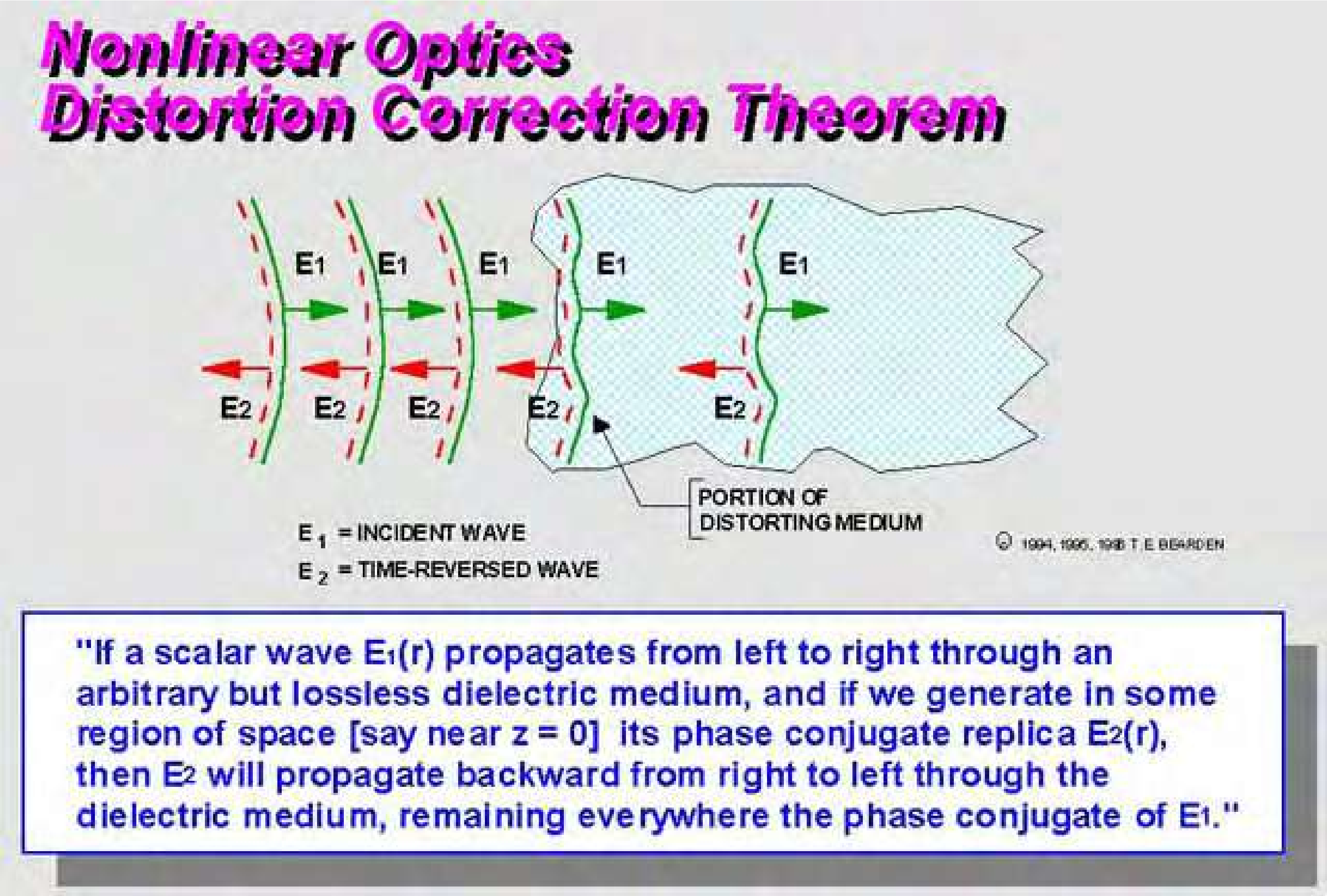

Nonlinear Optics
Distortion Correction Theorem

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

© 1994, 1995, 1998 T E BEARDEN

> "If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

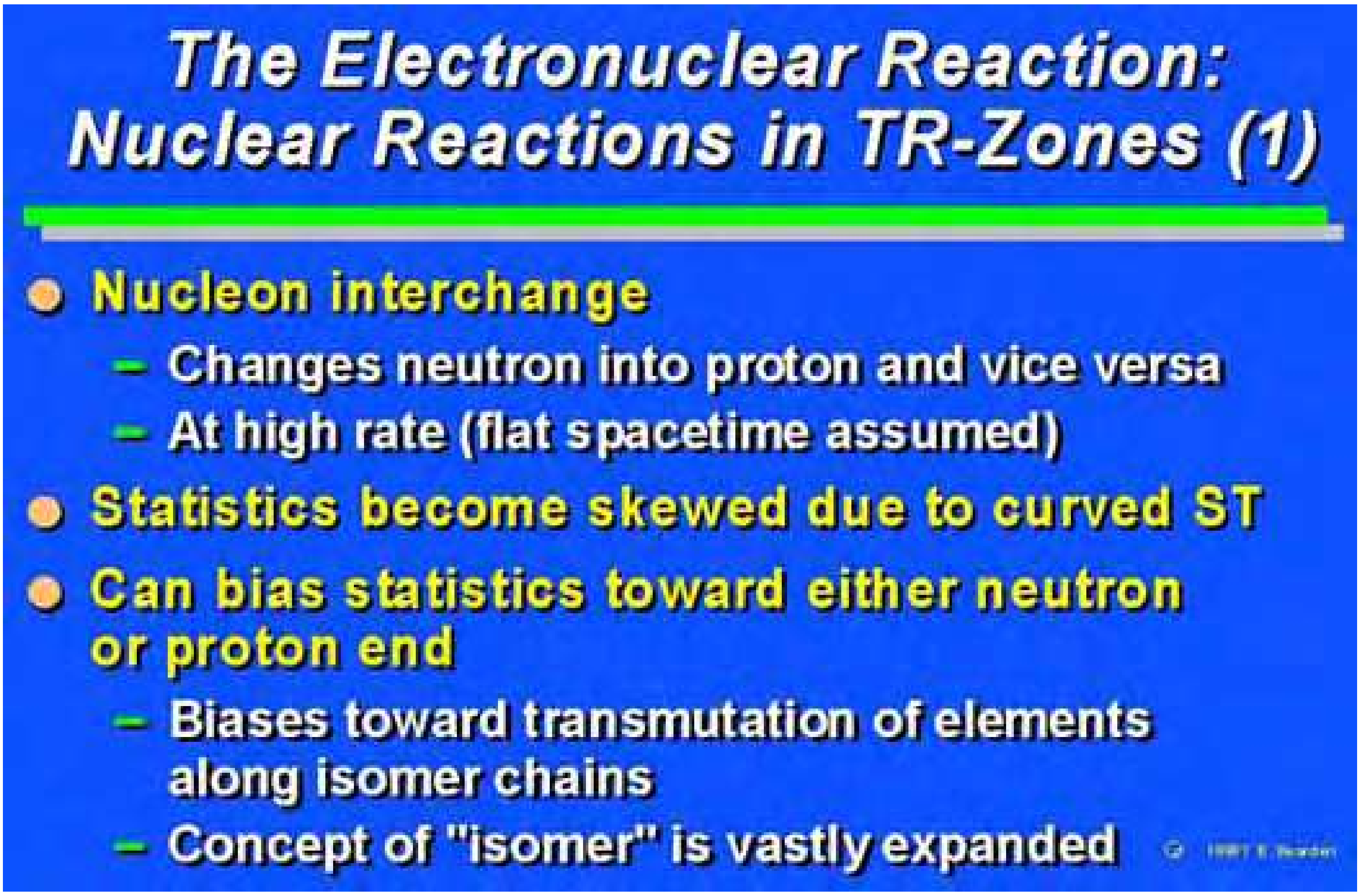

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- Nucleon interchange
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- Statistics become skewed due to curved ST
- Can bias statistics toward either neutron or proton end
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

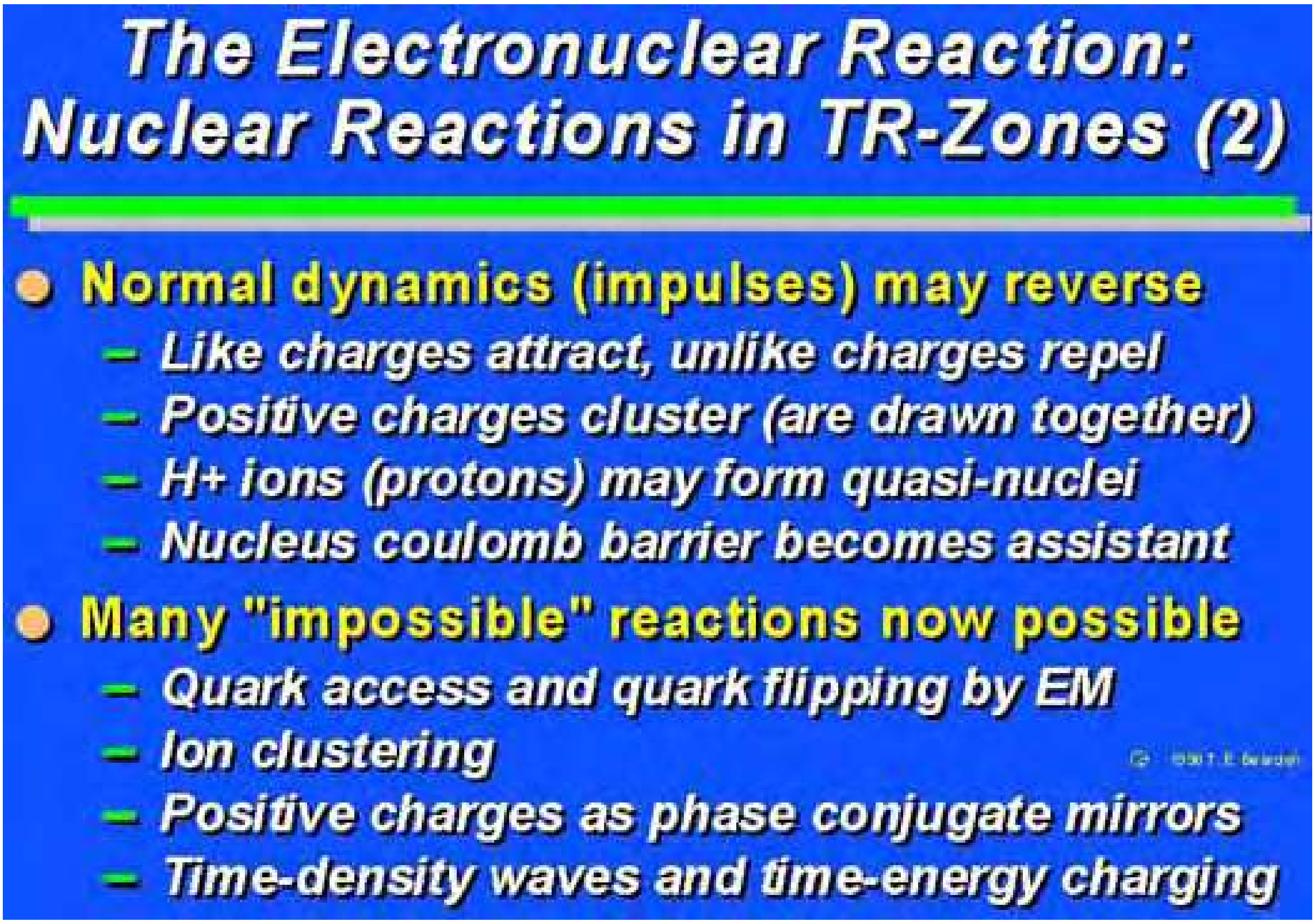

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- **Normal dynamics (impulses) may reverse**
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- **Many "impossible" reactions now possible**
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

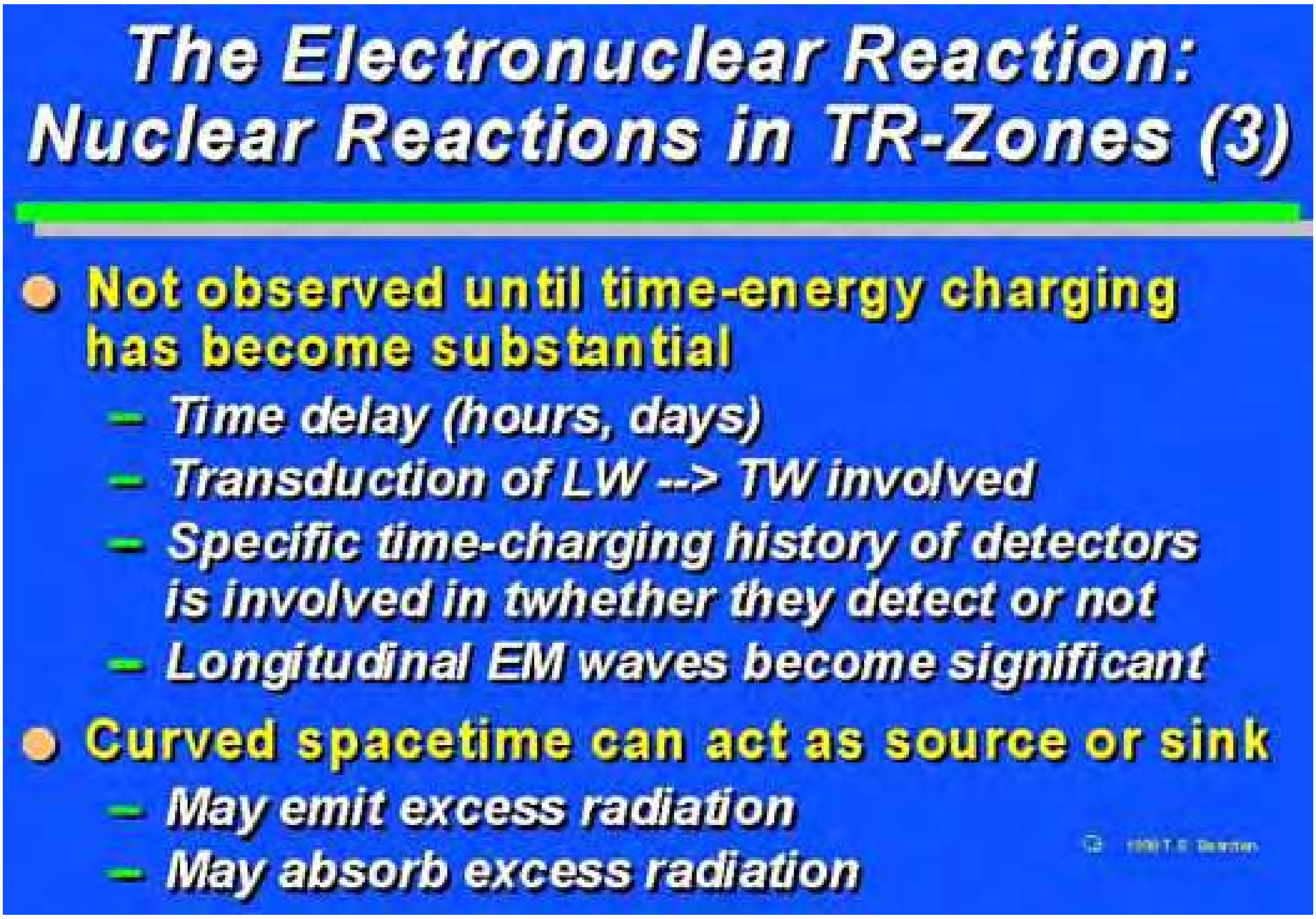

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- **Not observed until time-energy charging has become substantial**
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- **Curved spacetime can act as source or sink**
  - *May emit excess radiation*
  - *May absorb excess radiation*

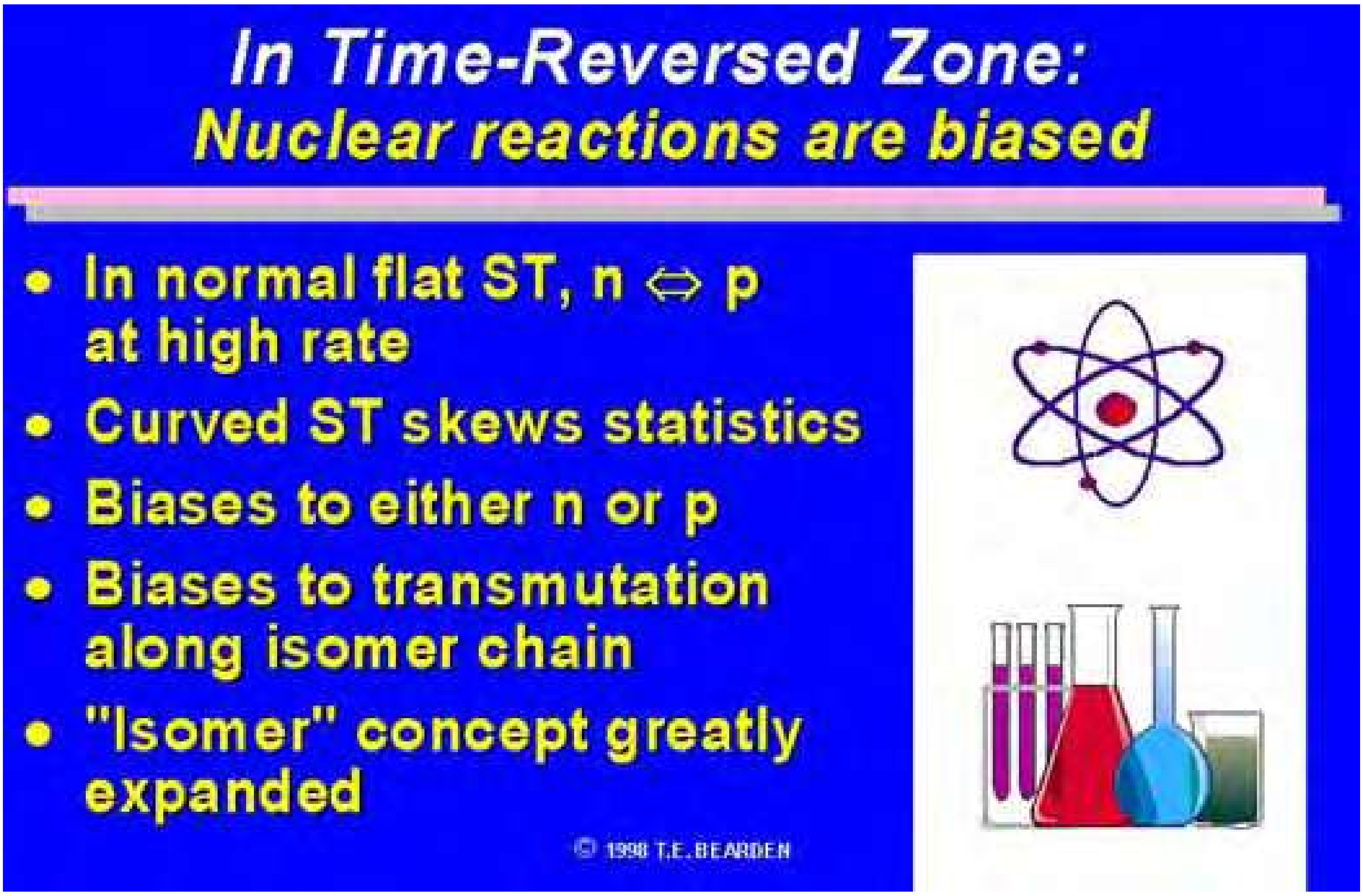

In Time-Reversed Zone: Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T.E.BEARDEN

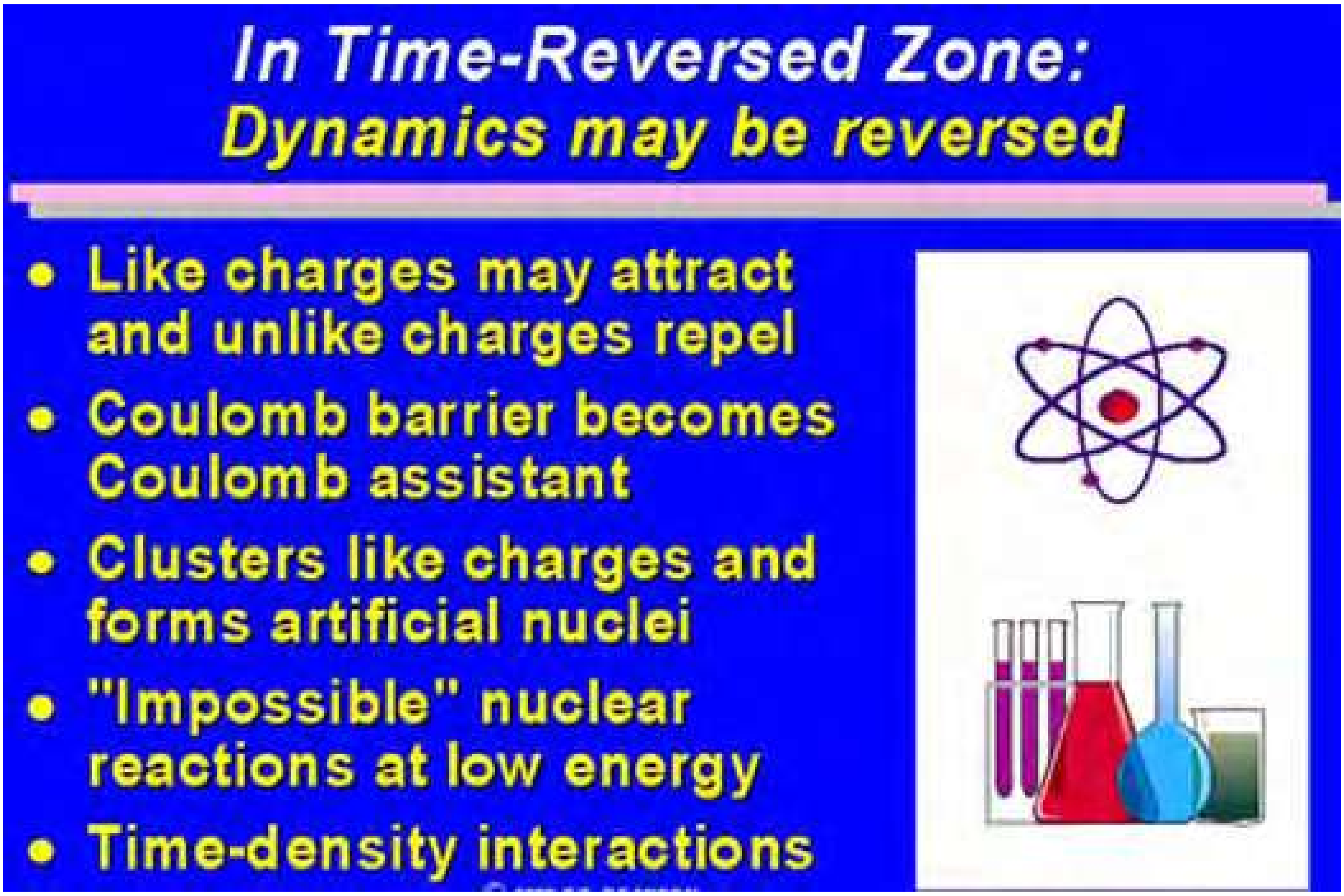

In Time-Reversed Zone:
Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

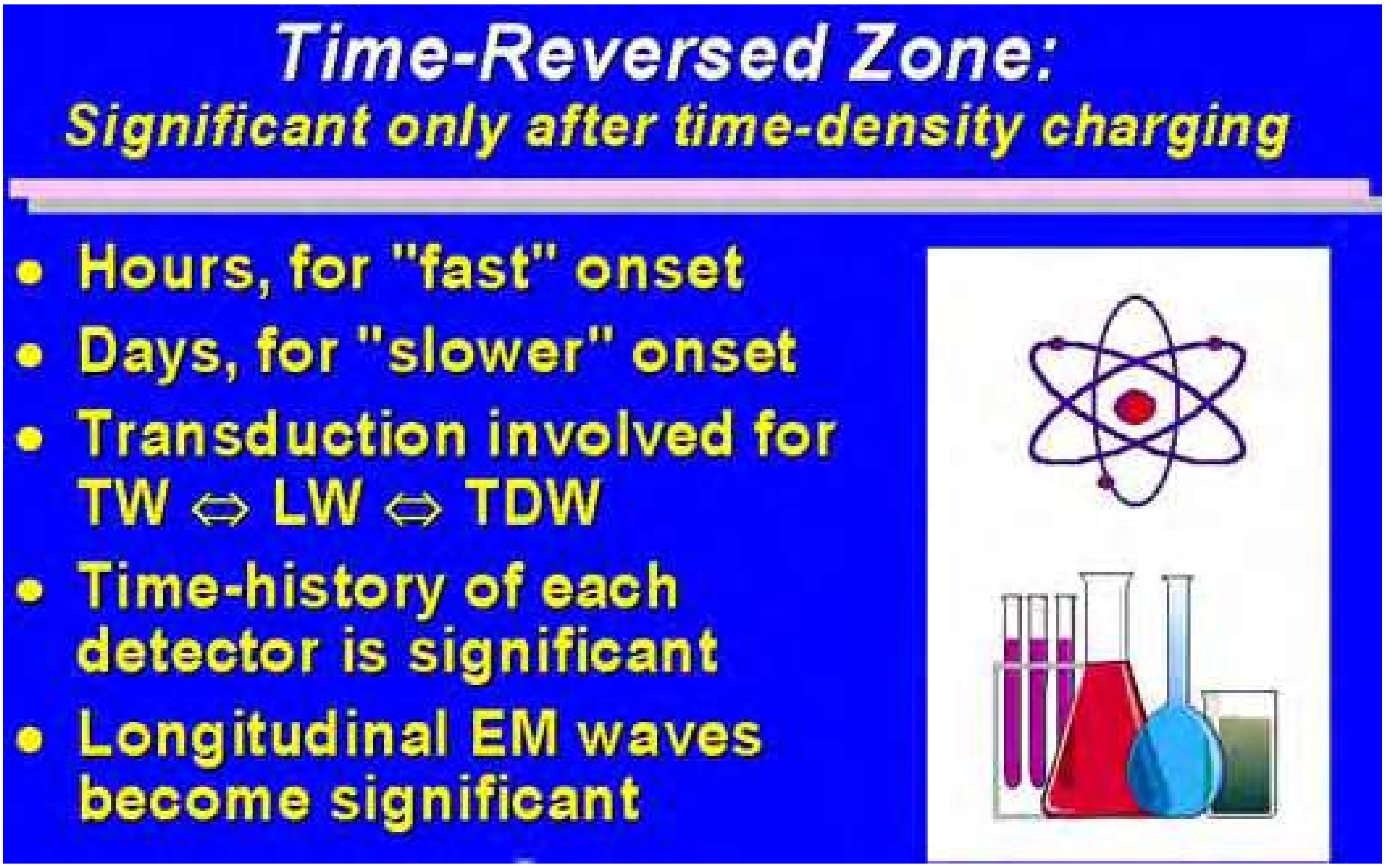

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

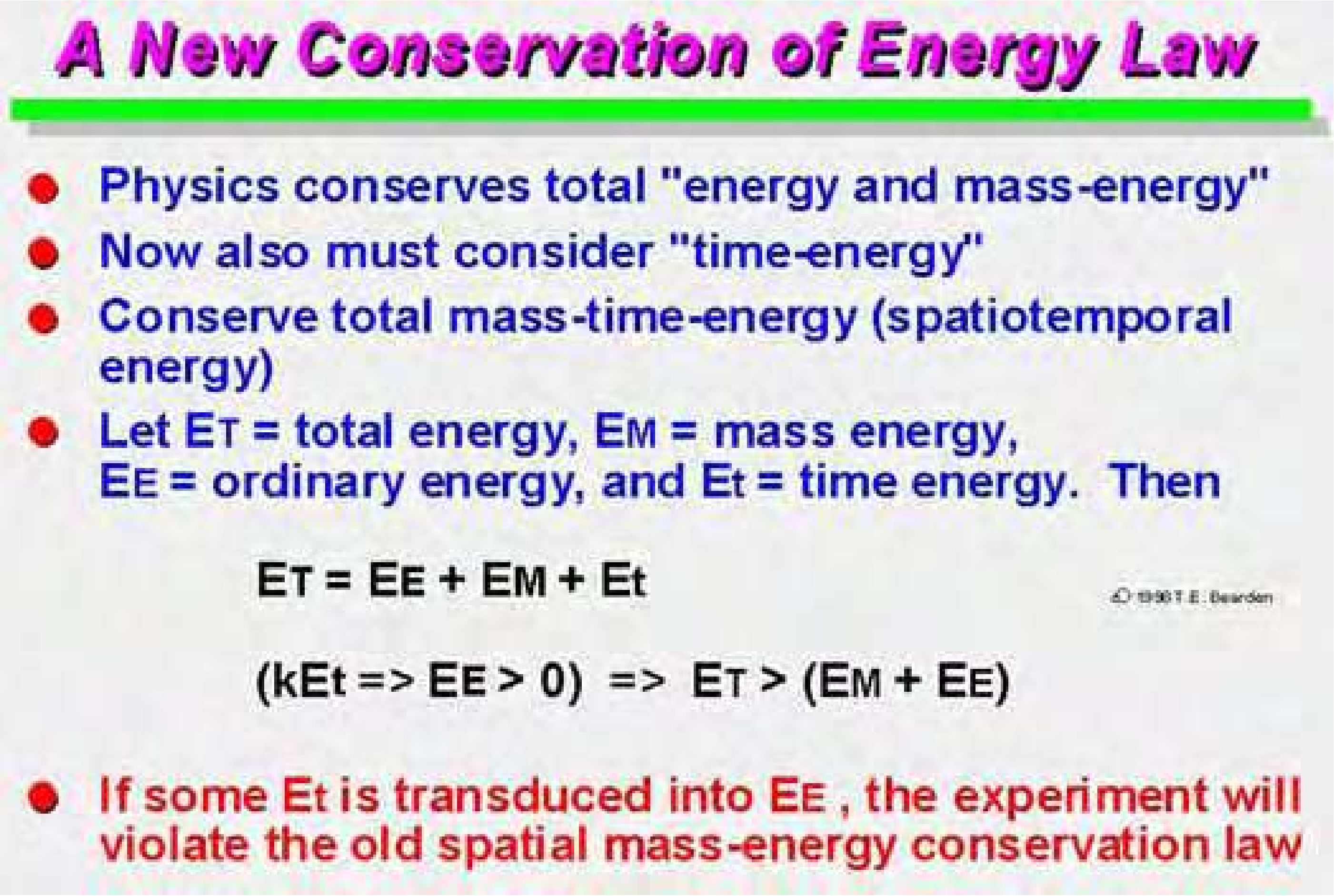

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

© 1998 T.E. Bearden

$$(kE_t \Rightarrow E_E > 0) \Rightarrow E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

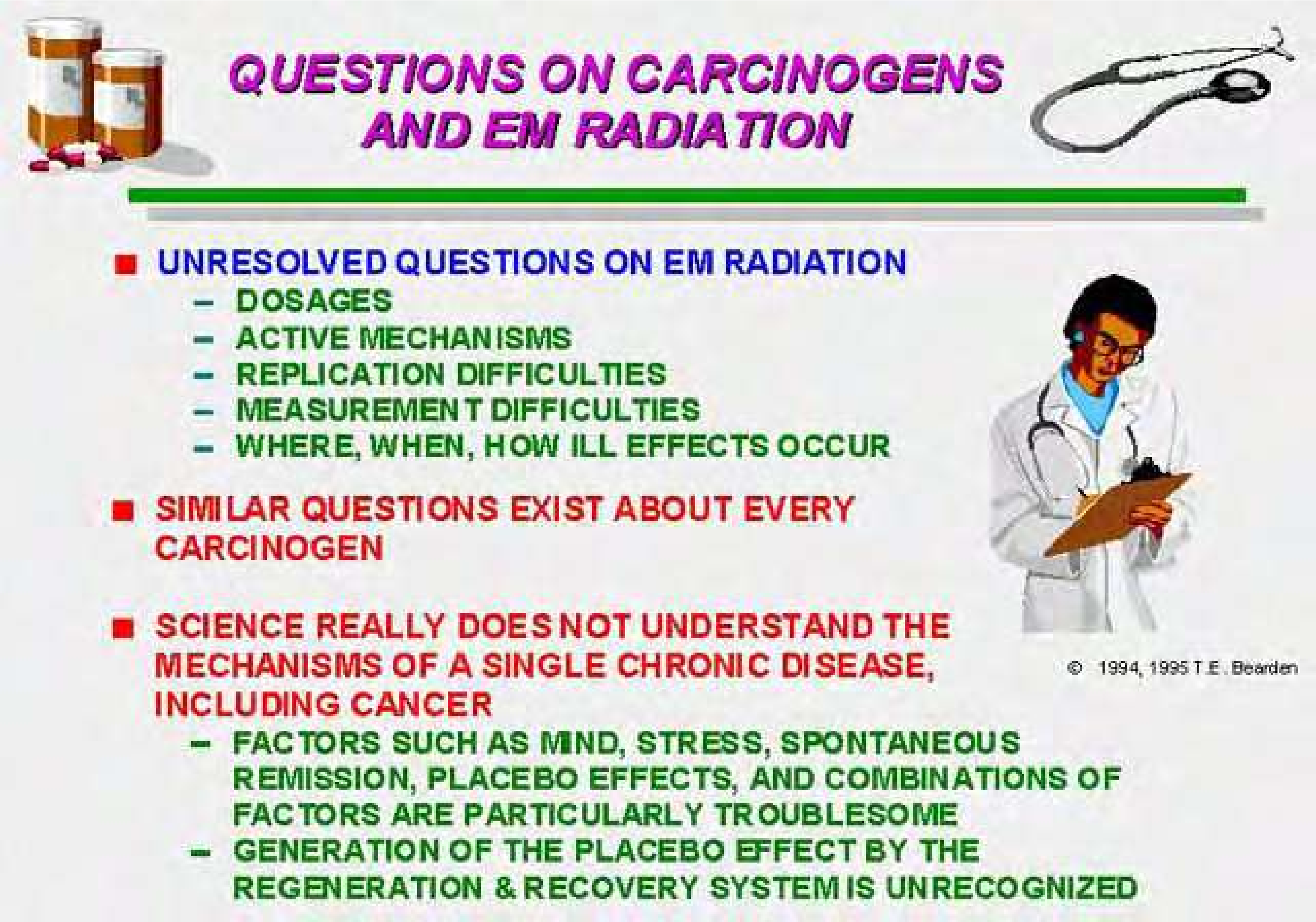

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

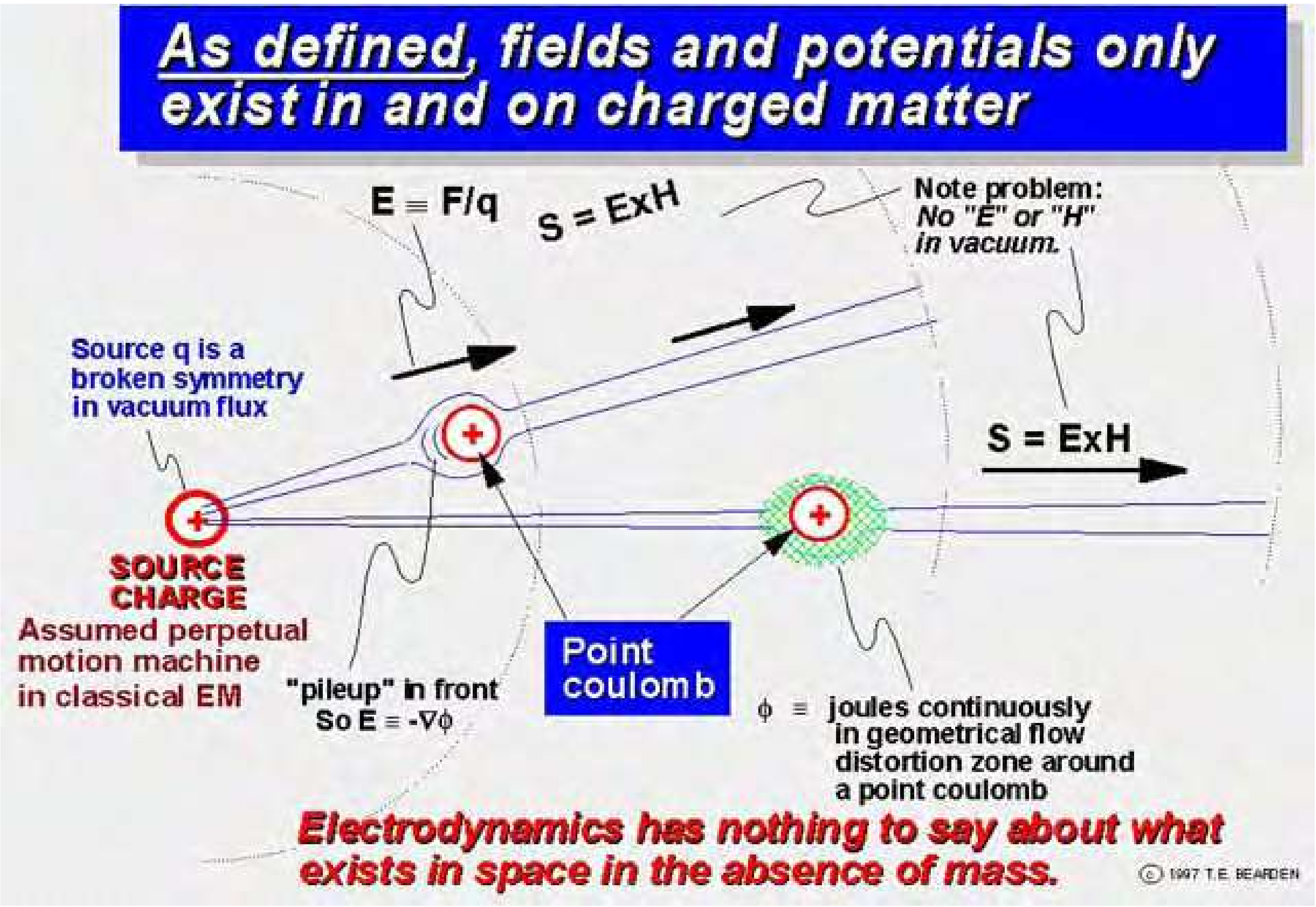

As defined, fields and potentials only exist in and on charged matter

$E \equiv F/q$

$S = E \times H$

Note problem:
No "E" or "H" in vacuum.

Source q is a broken symmetry in vacuum flux

$S = E \times H$

SOURCE CHARGE

Assumed perpetual motion machine in classical EM

"pileup" in front
So $E \equiv -\nabla\phi$

Point coulomb

$\phi \equiv$ joules continuously in geometrical flow distortion zone around a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

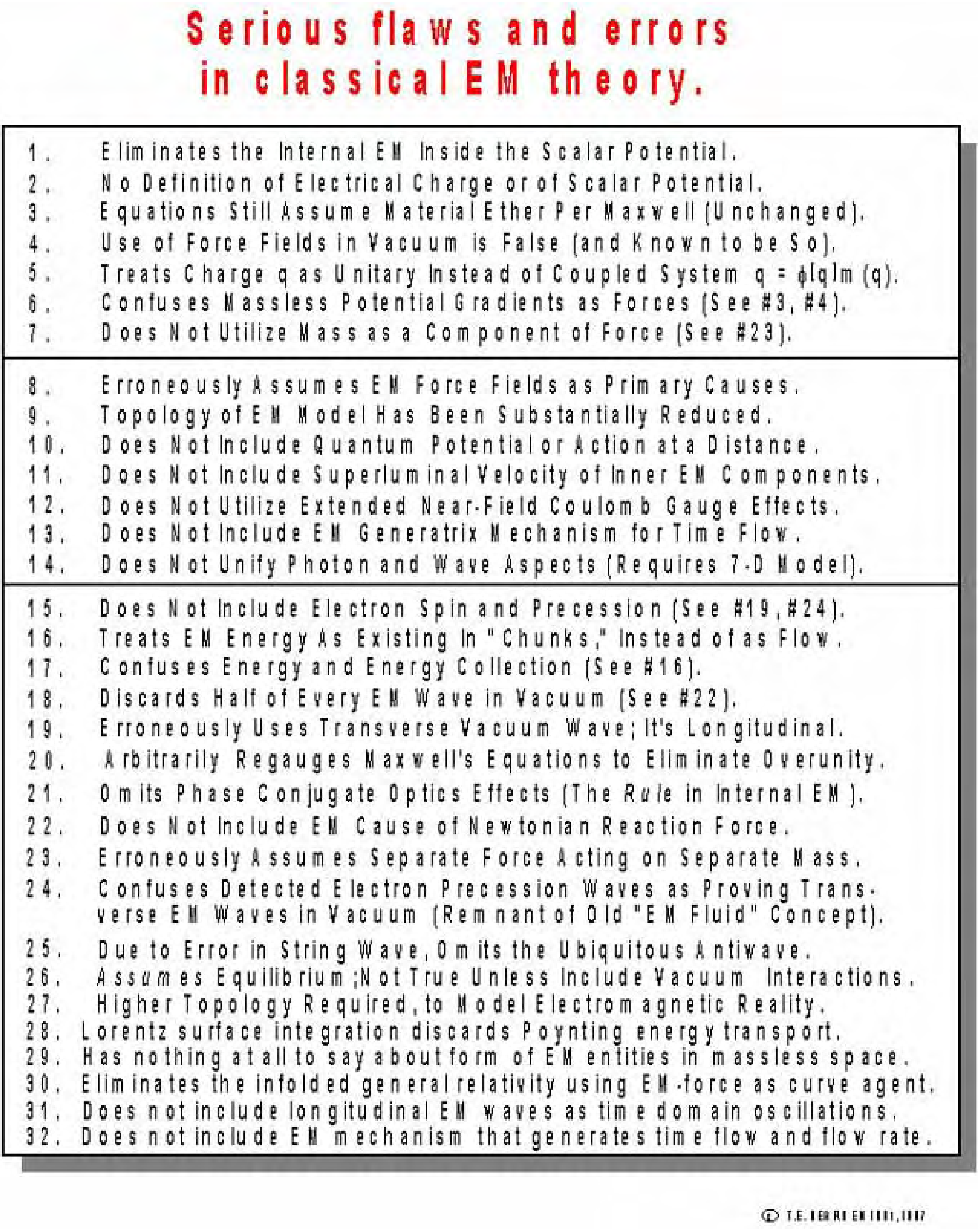

Serious flaws and errors in classical EM theory.

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi[q]m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

© T.E. BEARDEN 1991, 1997

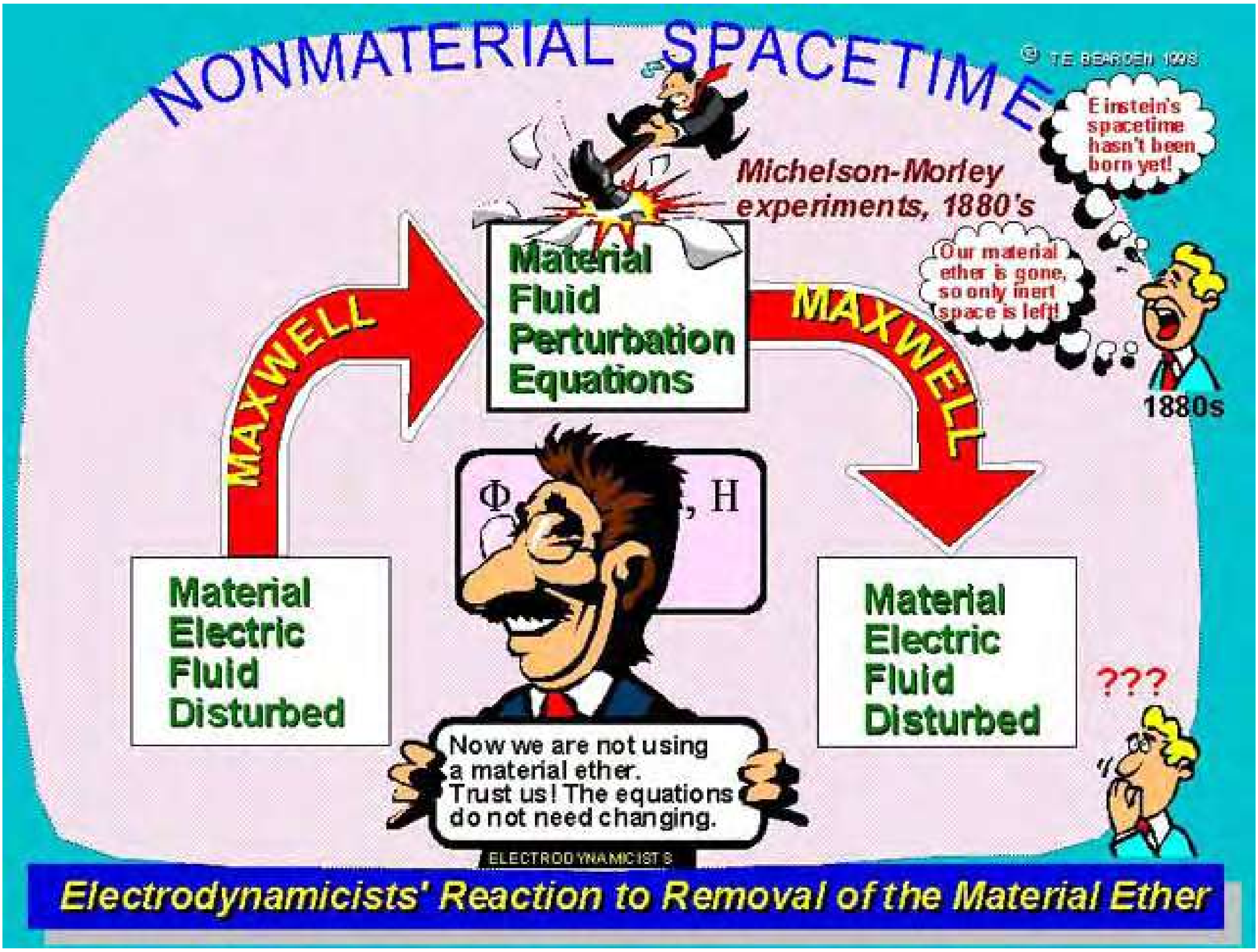
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Einstein's spacetime hasn't been born yet!
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left!
1880s
MAXWELL
Material Fluid Perturbation Equations
MAXWELL
Φ
, H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

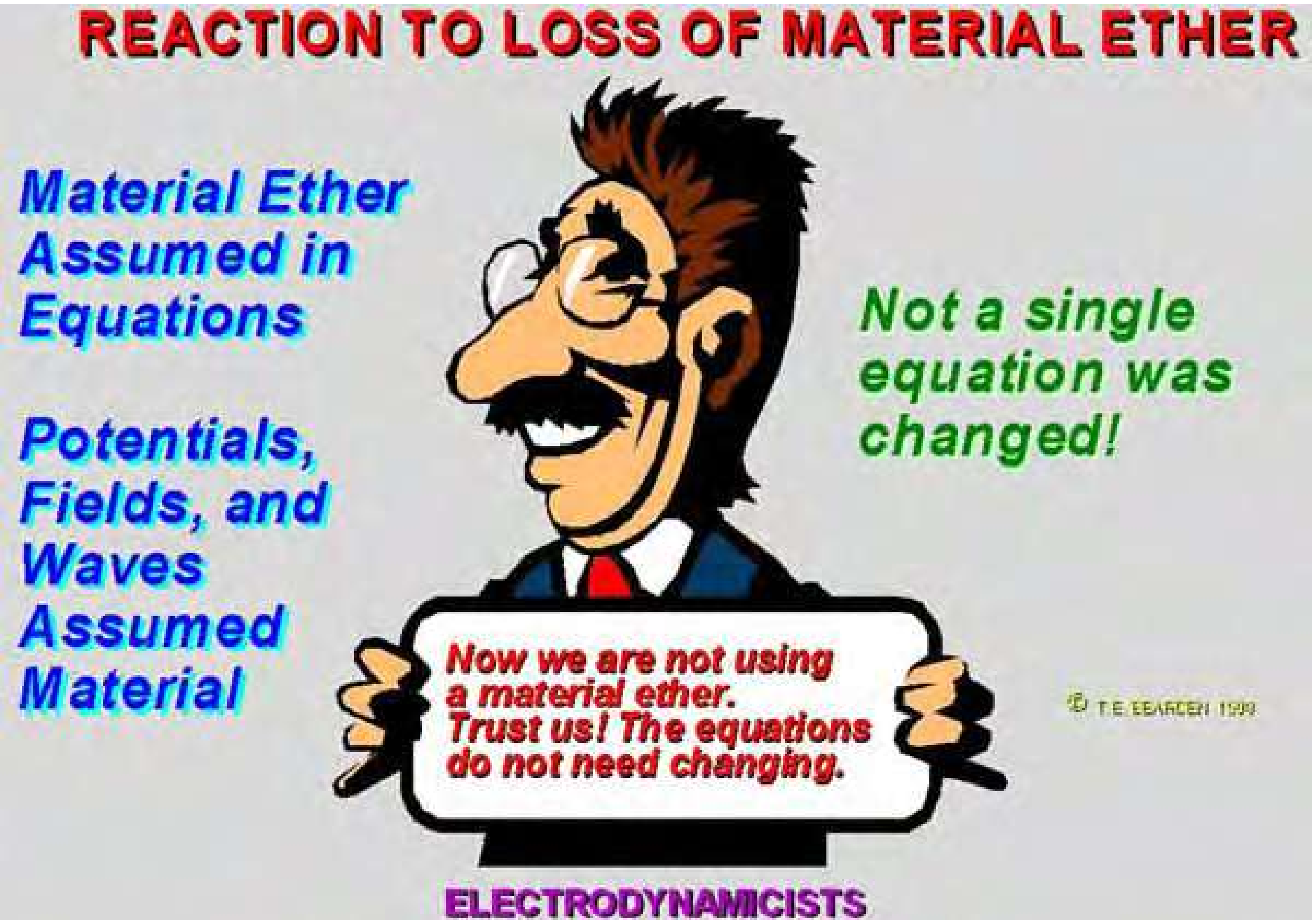

REACTION TO LOSS OF MATERIAL ETHER

Material Ether
Assumed in
Equations

Potentials,
Fields, and
Waves
Assumed
Material

Not a single
equation was
changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.

© T.E. BEARDEN 1999

ELECTRODYNAMICISTS

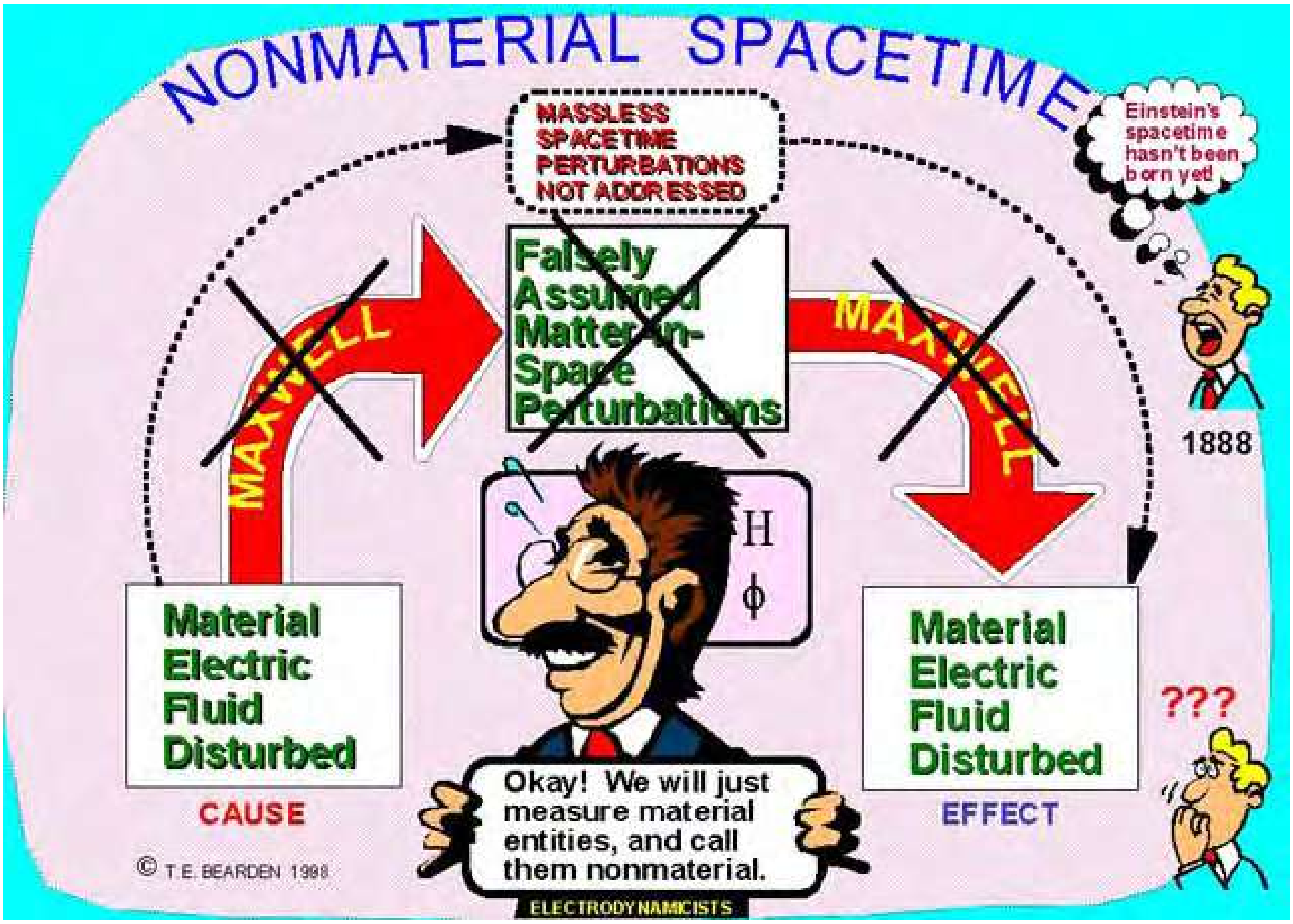
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
1888
~~MAXWELL~~
~~Falsely Assumed Matter-in-Space Perturbations~~
~~MAXWELL~~
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E. BEARDEN 1998

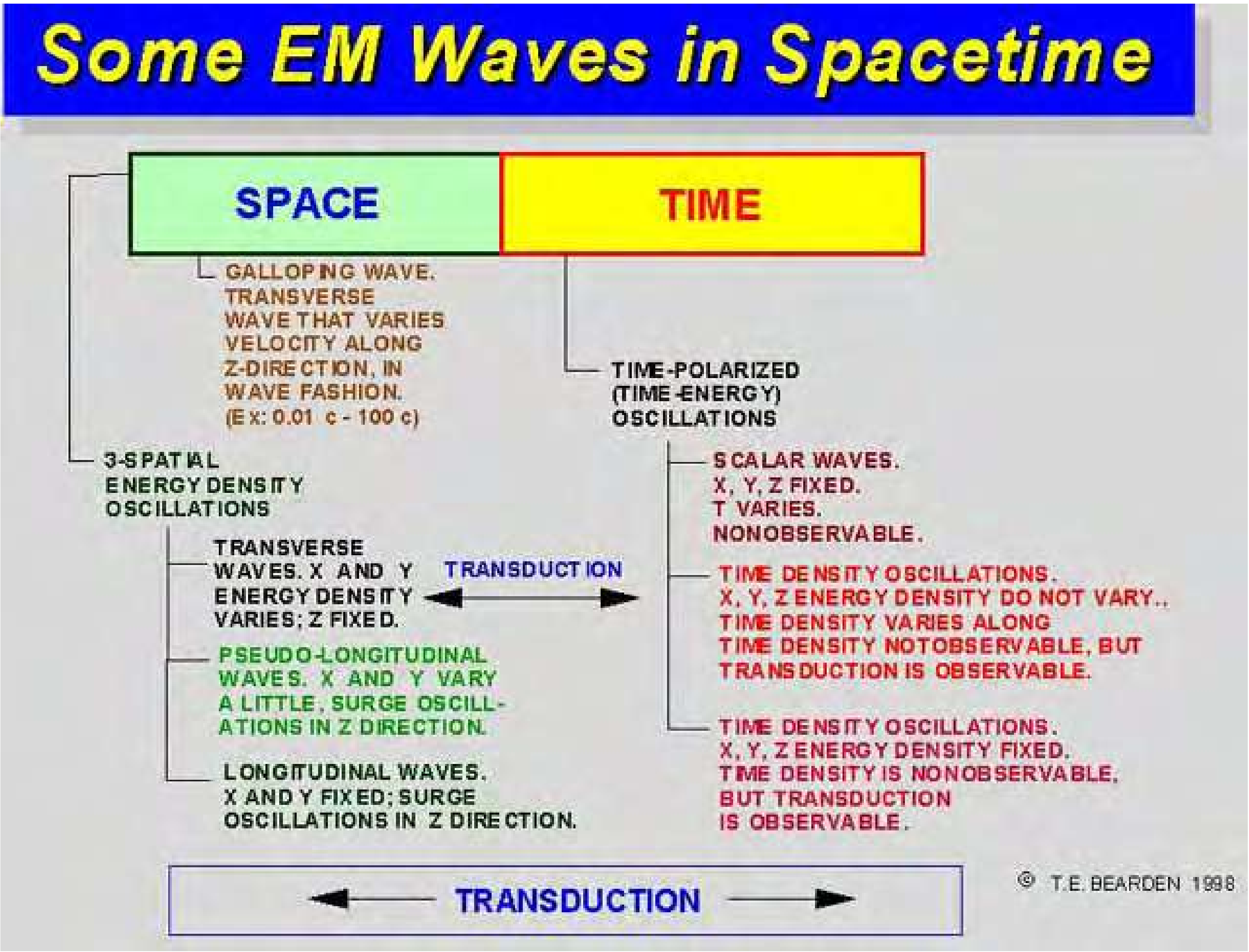

Some EM Waves in Spacetime

SPACE

TIME

GALLOPING WAVE. TRANSVERSE WAVE THAT VARIES VELOCITY ALONG Z-DIRECTION, IN WAVE FASHION. (Ex: 0.01 c - 100 c)

TIME-POLARIZED (TIME-ENERGY) OSCILLATIONS

3-SPATIAL ENERGY DENSITY OSCILLATIONS

SCALAR WAVES. X, Y, Z FIXED. T VARIES. NONOBSERVABLE.

TRANSVERSE WAVES. X AND Y ENERGY DENSITY VARIES; Z FIXED.

TRANSDUCTION

TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY DO NOT VARY.. TIME DENSITY VARIES ALONG TIME DENSITY NOTOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.

PSEUDO-LONGITUDINAL WAVES. X AND Y VARY A LITTLE, SURGE OSCILLATIONS IN Z DIRECTION.

TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY FIXED. TIME DENSITY IS NONOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.

LONGITUDINAL WAVES. X AND Y FIXED; SURGE OSCILLATIONS IN Z DIRECTION.

TRANSDUCTION

© T.E. BEARDEN 1998

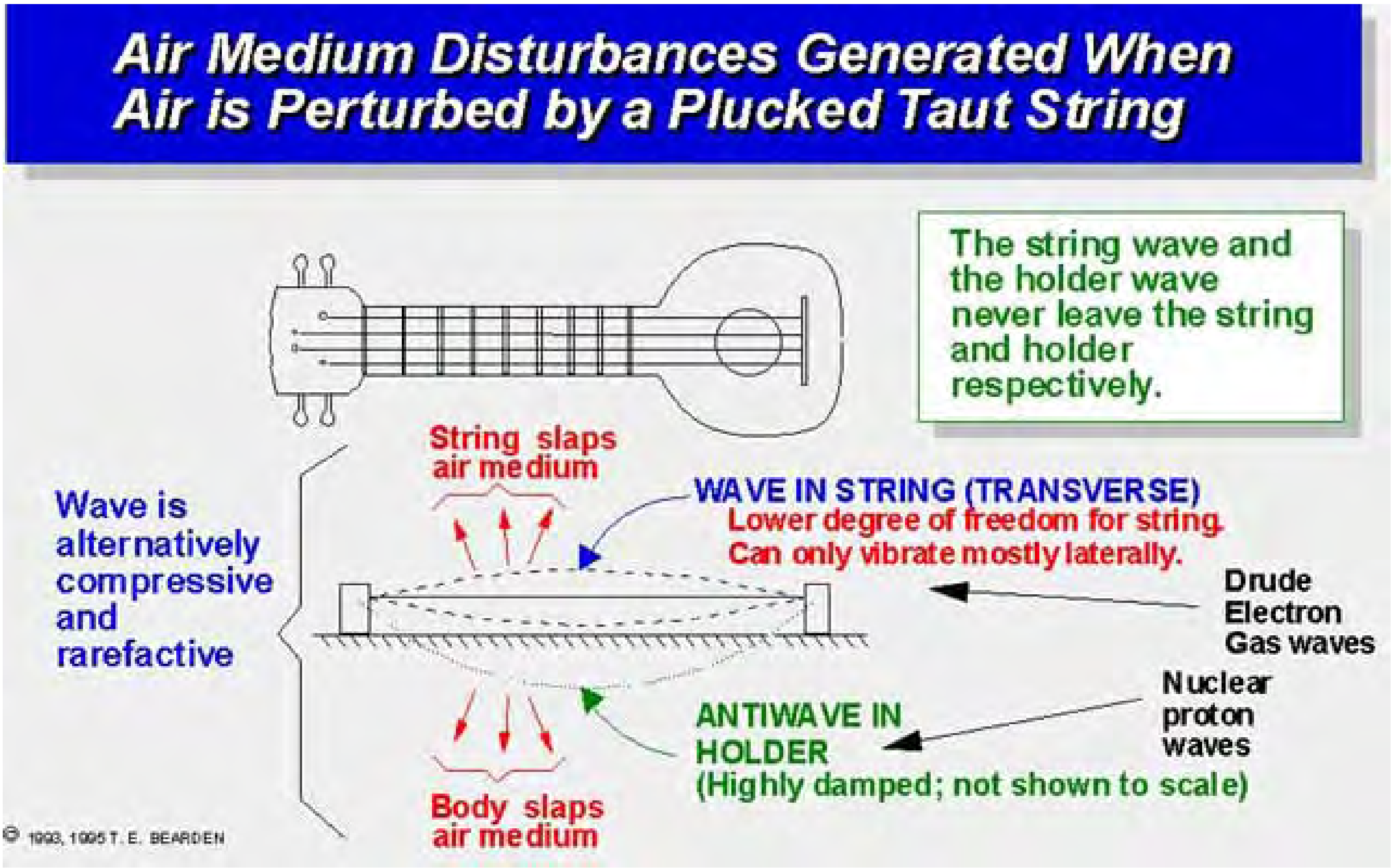

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

Wave is alternatively compressive and rarefactive

String slaps air medium

WAVE IN STRING (TRANSVERSE)

Lower degree of freedom for string. Can only vibrate mostly laterally.

Drude Electron Gas waves

Nuclear proton waves

ANTIWAVE IN HOLDER (Highly damped; not shown to scale)

Body slaps air medium

© 1993, 1995 T. E. BEARDEN

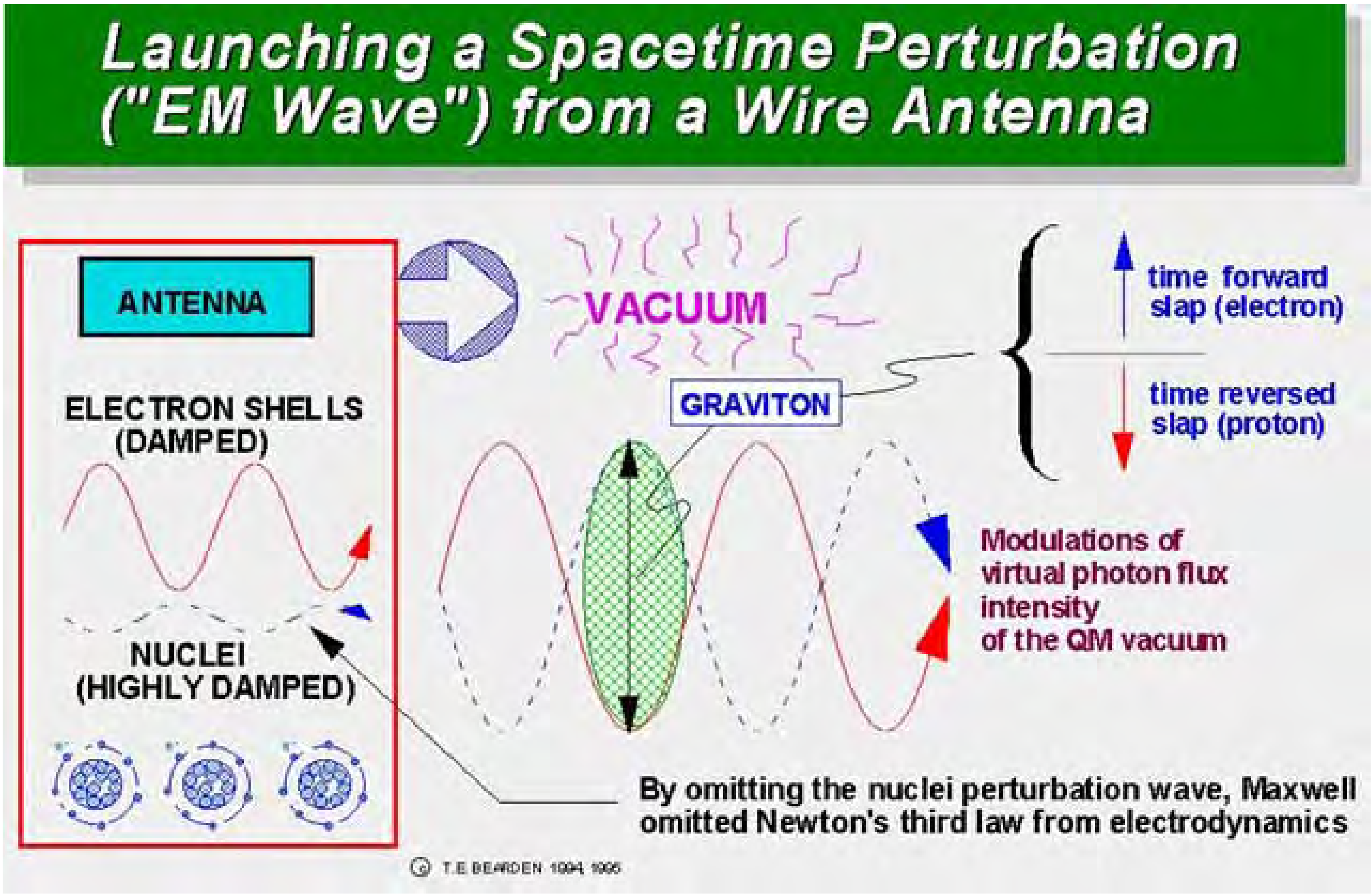

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

ANTENNA
ELECTRON SHELLS (DAMPED)
NUCLEI (HIGHLY DAMPED)
VACUUM
GRAVITON
time forward slap (electron)
time reversed slap (proton)
Modulations of virtual photon flux intensity of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

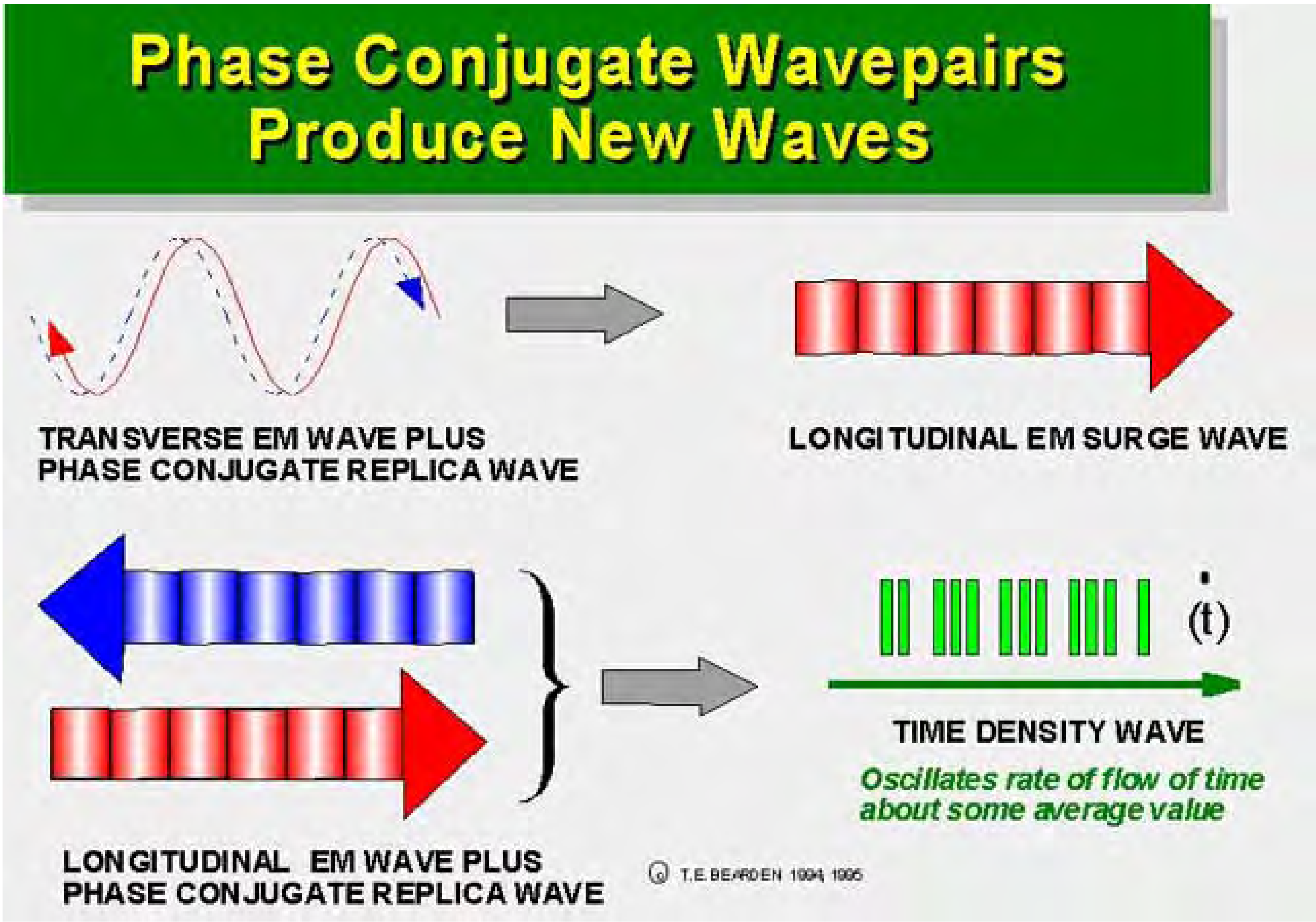

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM SURGE WAVE

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

$(\dot{t})$

TIME DENSITY WAVE

Oscillates rate of flow of time about some average value

© T.E. BEARDEN 1994, 1995

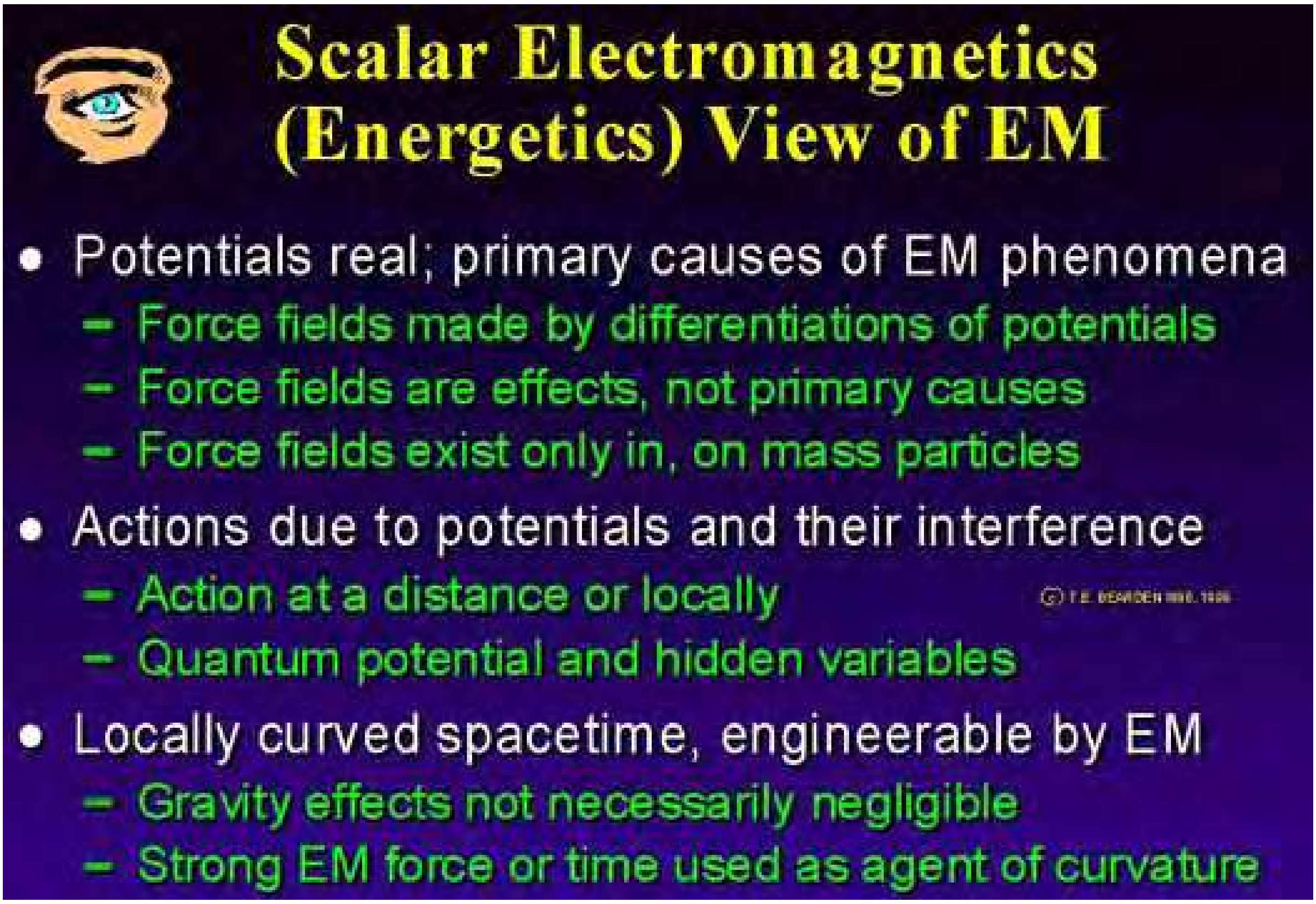

Scalar Electromagnetics (Energetics) View of EM

- Potentials real; primary causes of EM phenomena
  - Force fields made by differentiations of potentials
  - Force fields are effects, not primary causes
  - Force fields exist only in, on mass particles
- Actions due to potentials and their interference
  - Action at a distance or locally
  - Quantum potential and hidden variables
- Locally curved spacetime, engineerable by EM
  - Gravity effects not necessarily negligible
  - Strong EM force or time used as agent of curvature

© T.E. BEARDEN [illegible]

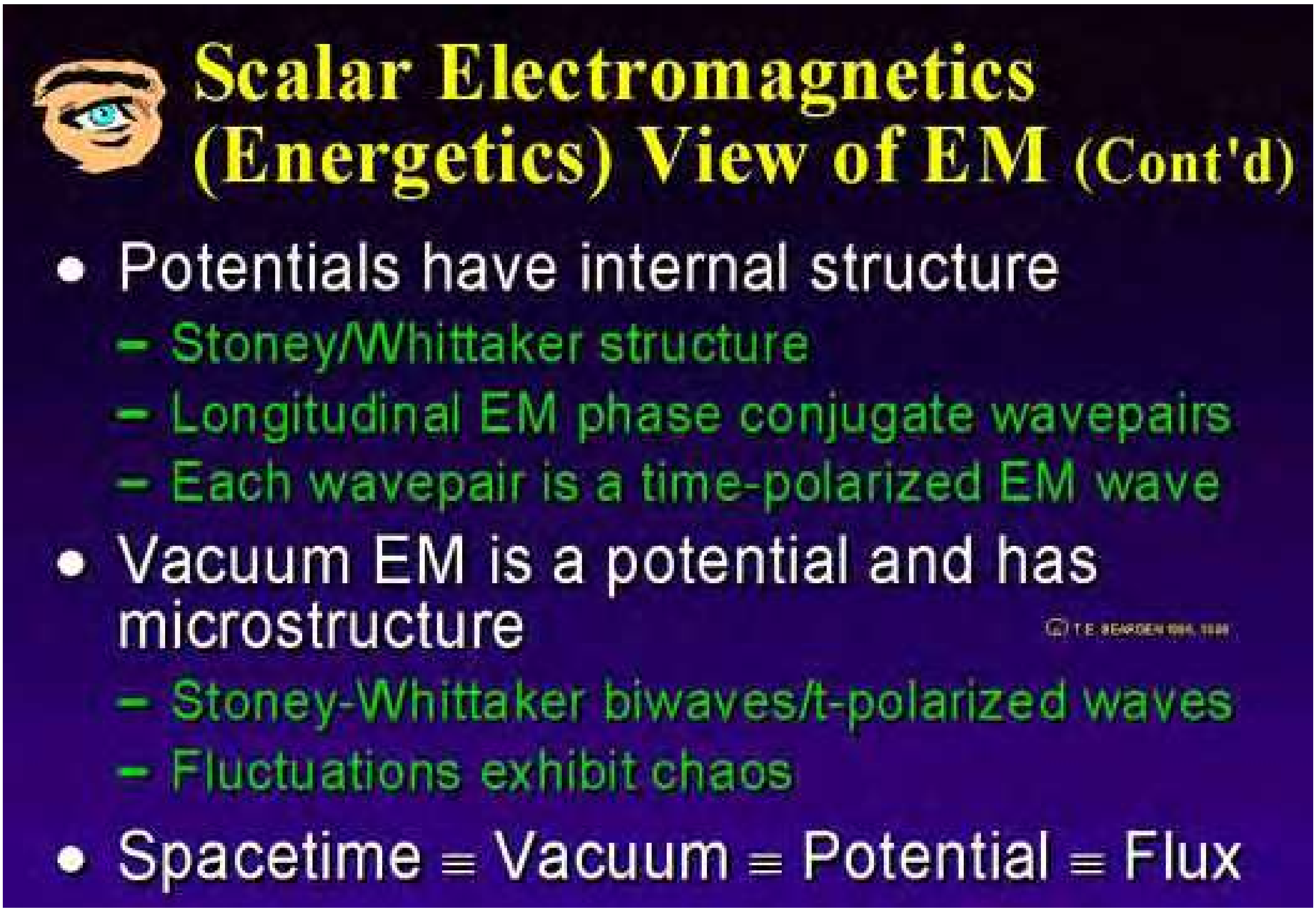

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Potentials have internal structure
  - Stoney/Whittaker structure
  - Longitudinal EM phase conjugate wavepairs
  - Each wavepair is a time-polarized EM wave
- Vacuum EM is a potential and has microstructure
  - Stoney-Whittaker biwaves/t-polarized waves
  - Fluctuations exhibit chaos
- Spacetime ≡ Vacuum ≡ Potential ≡ Flux

© T.E. BEARDEN [illegible]

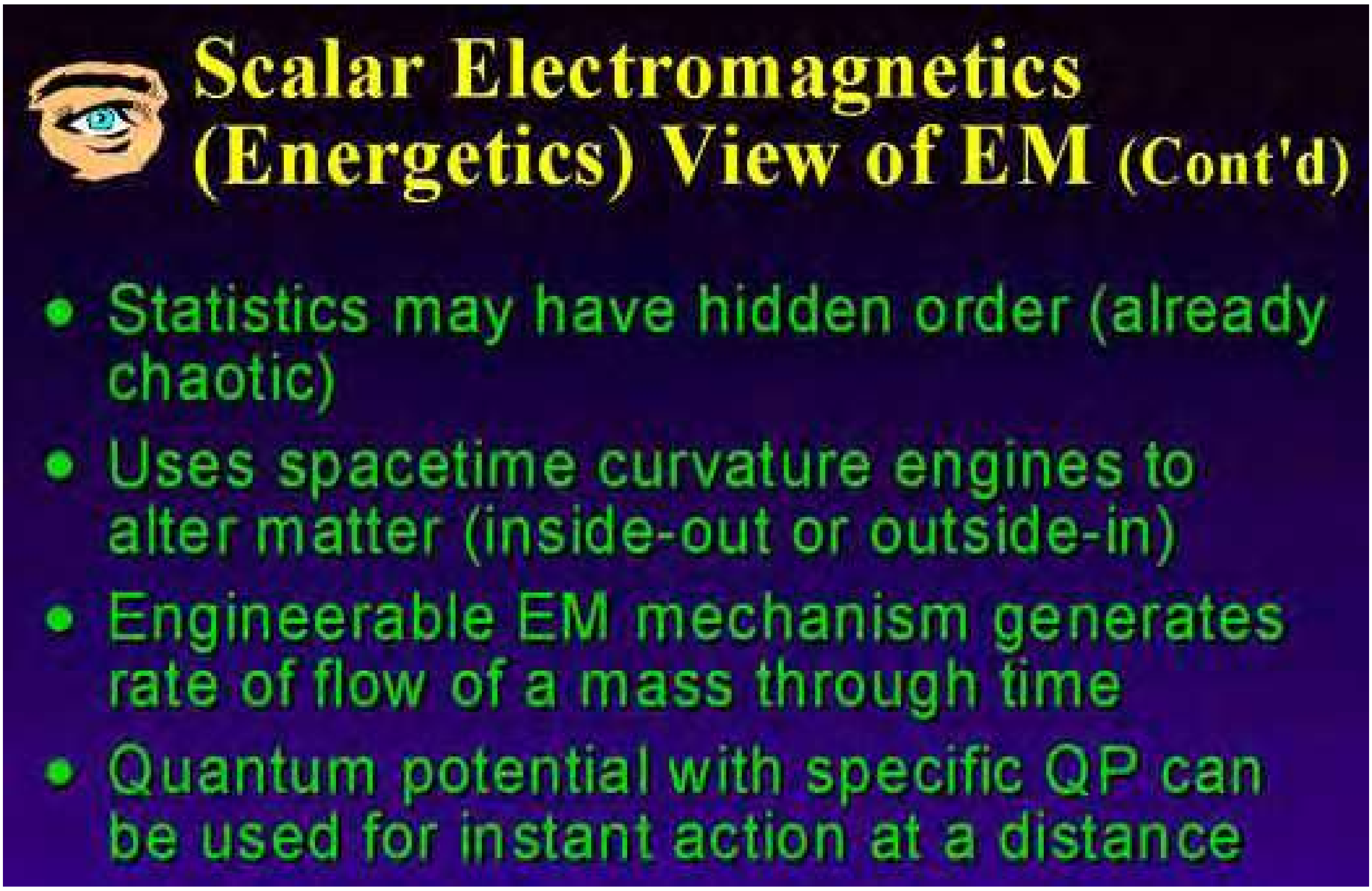

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Statistics may have hidden order (already chaotic)
- Uses spacetime curvature engines to alter matter (inside-out or outside-in)
- Engineerable EM mechanism generates rate of flow of a mass through time
- Quantum potential with specific QP can be used for instant action at a distance

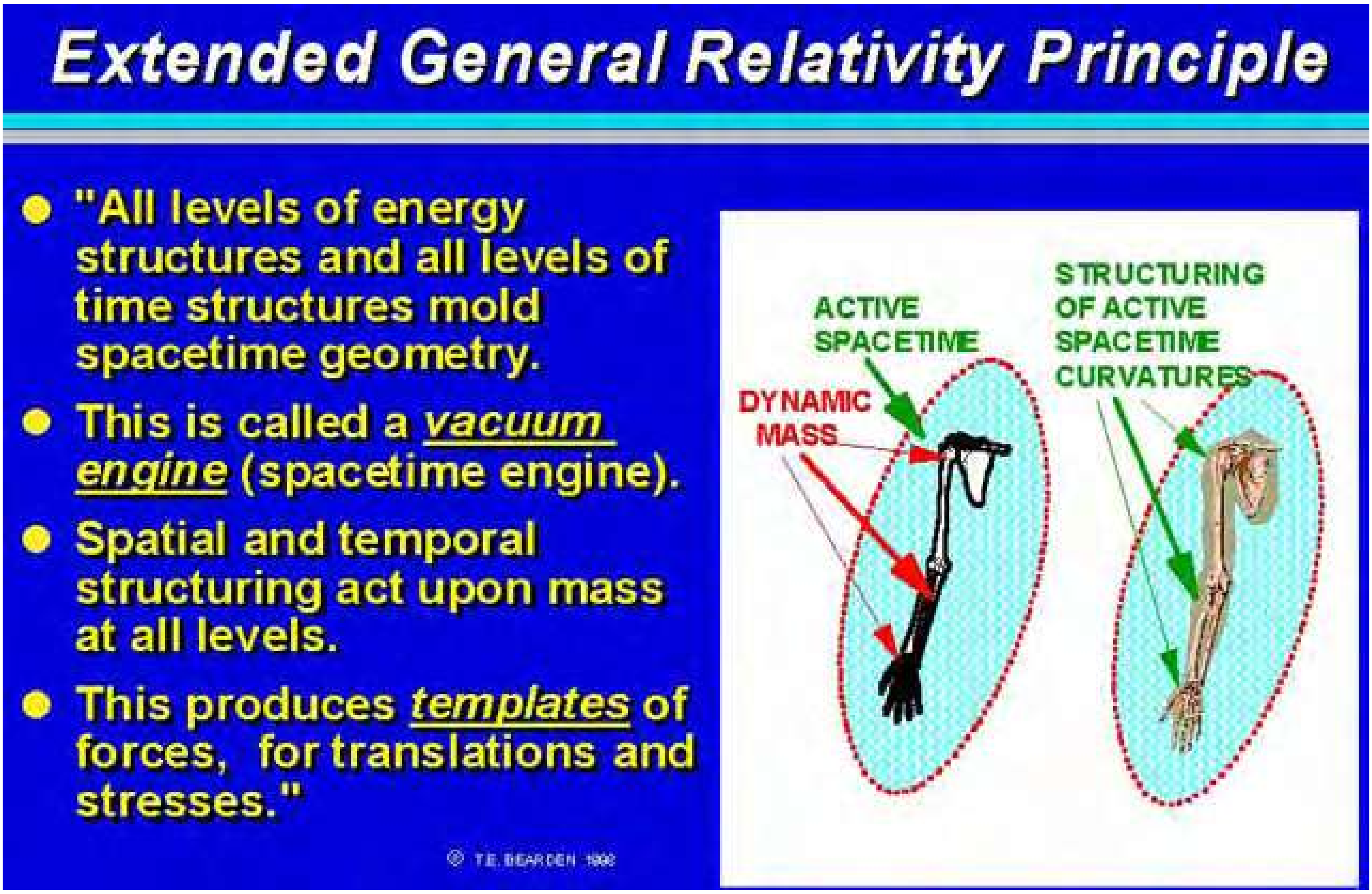

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a *vacuum engine* (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces *templates* of forces, for translations and stresses."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E. BEARDEN 1990

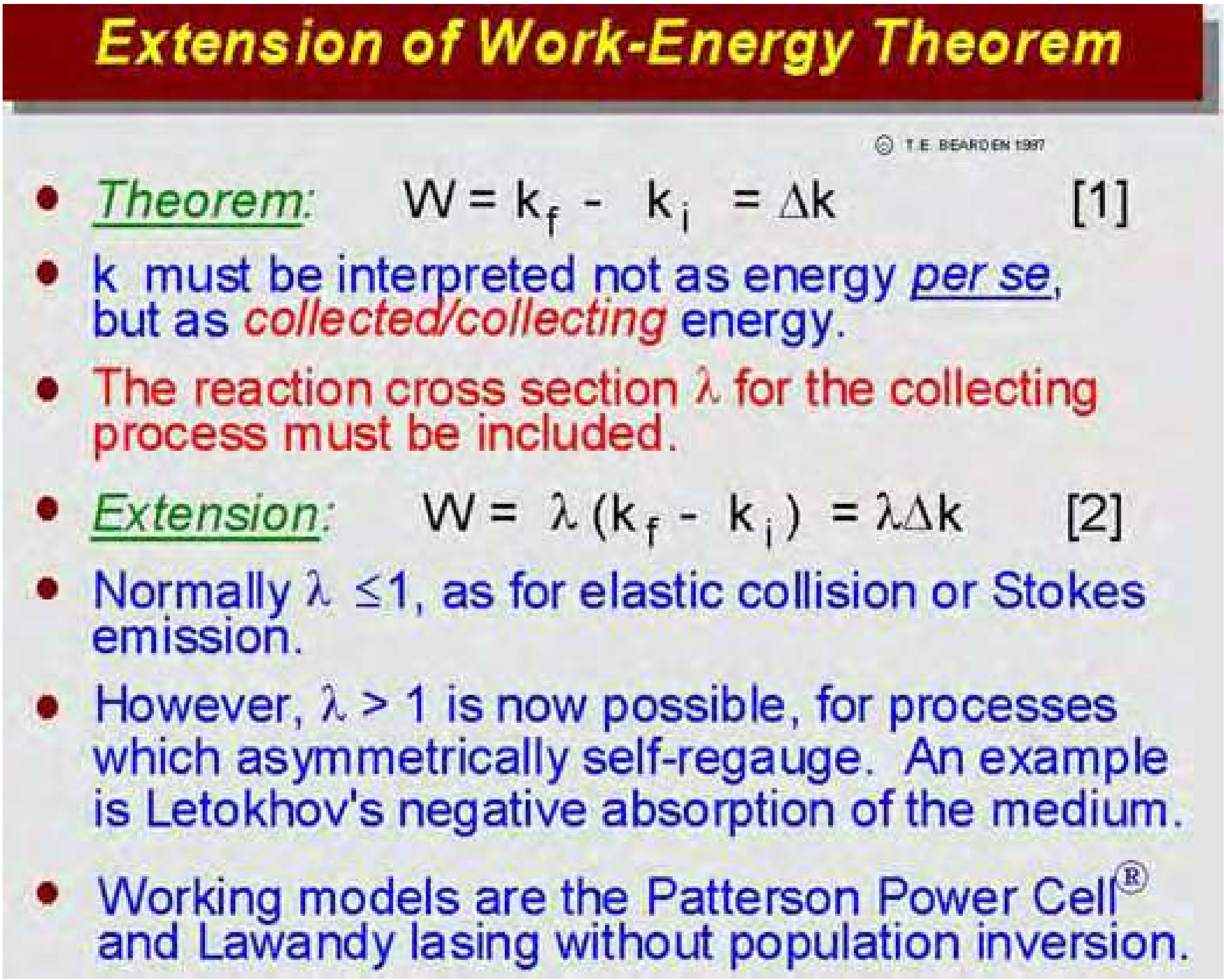

Extension of Work-Energy Theorem

© T.E. BEARDEN 1997

- *Theorem*: $W = k_f - k_i = \Delta k$ [1]
- k must be interpreted not as energy *per se*, but as *collected/collecting* energy.
- The reaction cross section λ for the collecting process must be included.
- *Extension*: $W = \lambda (k_f - k_i) = \lambda \Delta k$ [2]
- Normally $\lambda \leq 1$, as for elastic collision or Stokes emission.
- However, $\lambda > 1$ is now possible, for processes which asymmetrically self-regauge. An example is Letokhov's negative absorption of the medium.
- Working models are the Patterson Power Cell® and Lawandy lasing without population inversion.

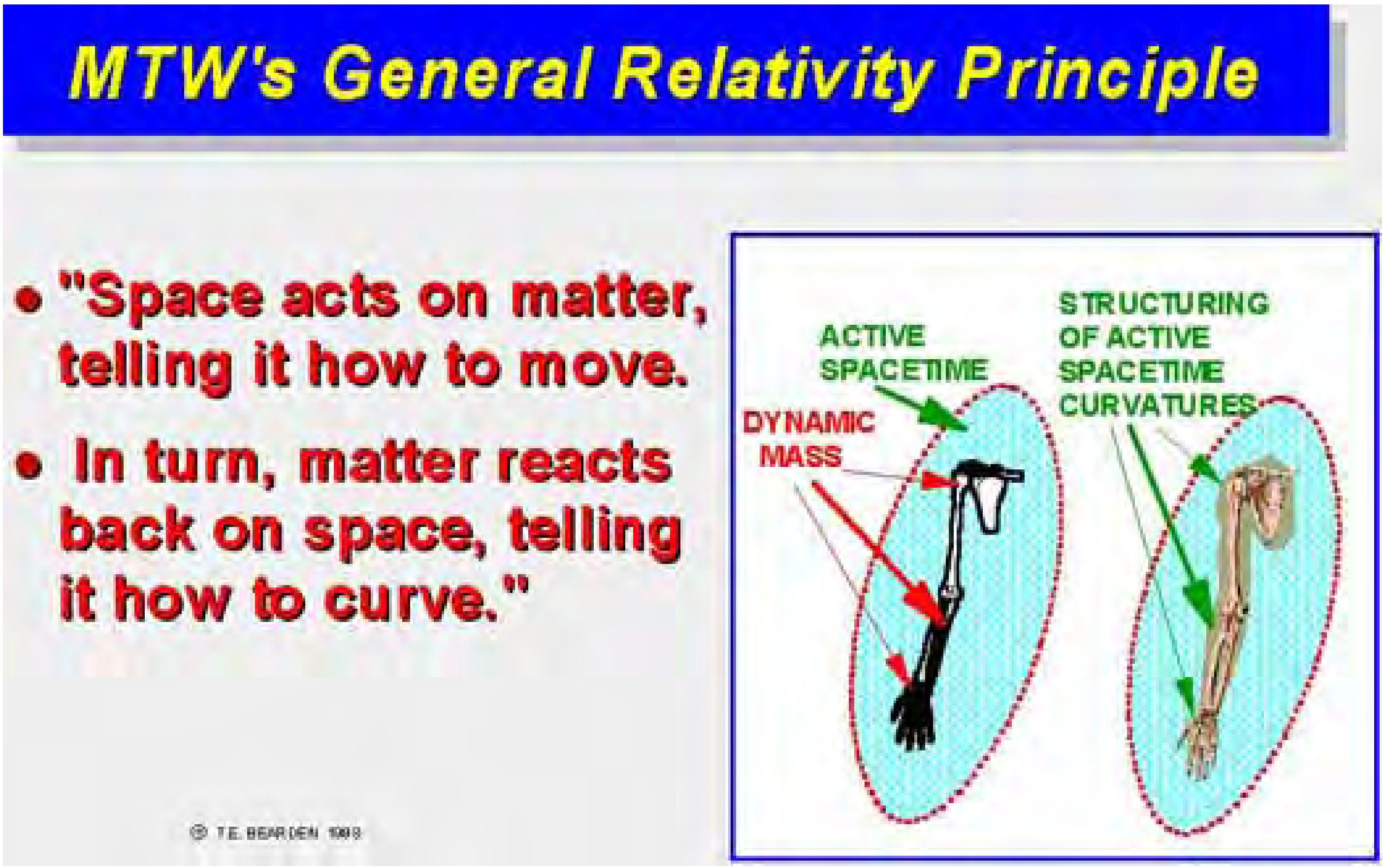

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

© T.E. BEARDEN 1998

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

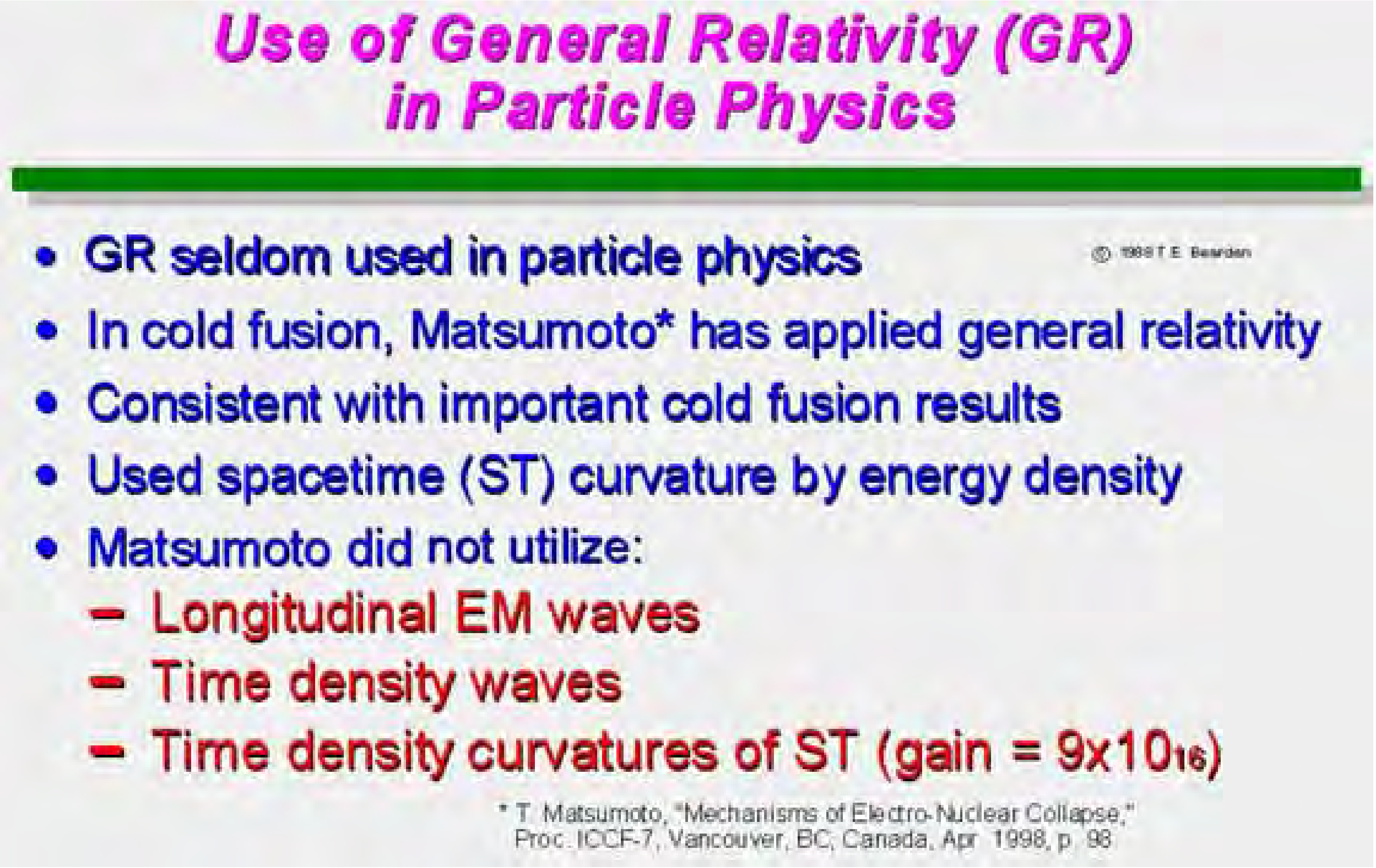

Use of General Relativity (GR) in Particle Physics

© 1998 T.E. Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density waves
  - Time density curvatures of ST (gain = 9×10^{16})

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98.

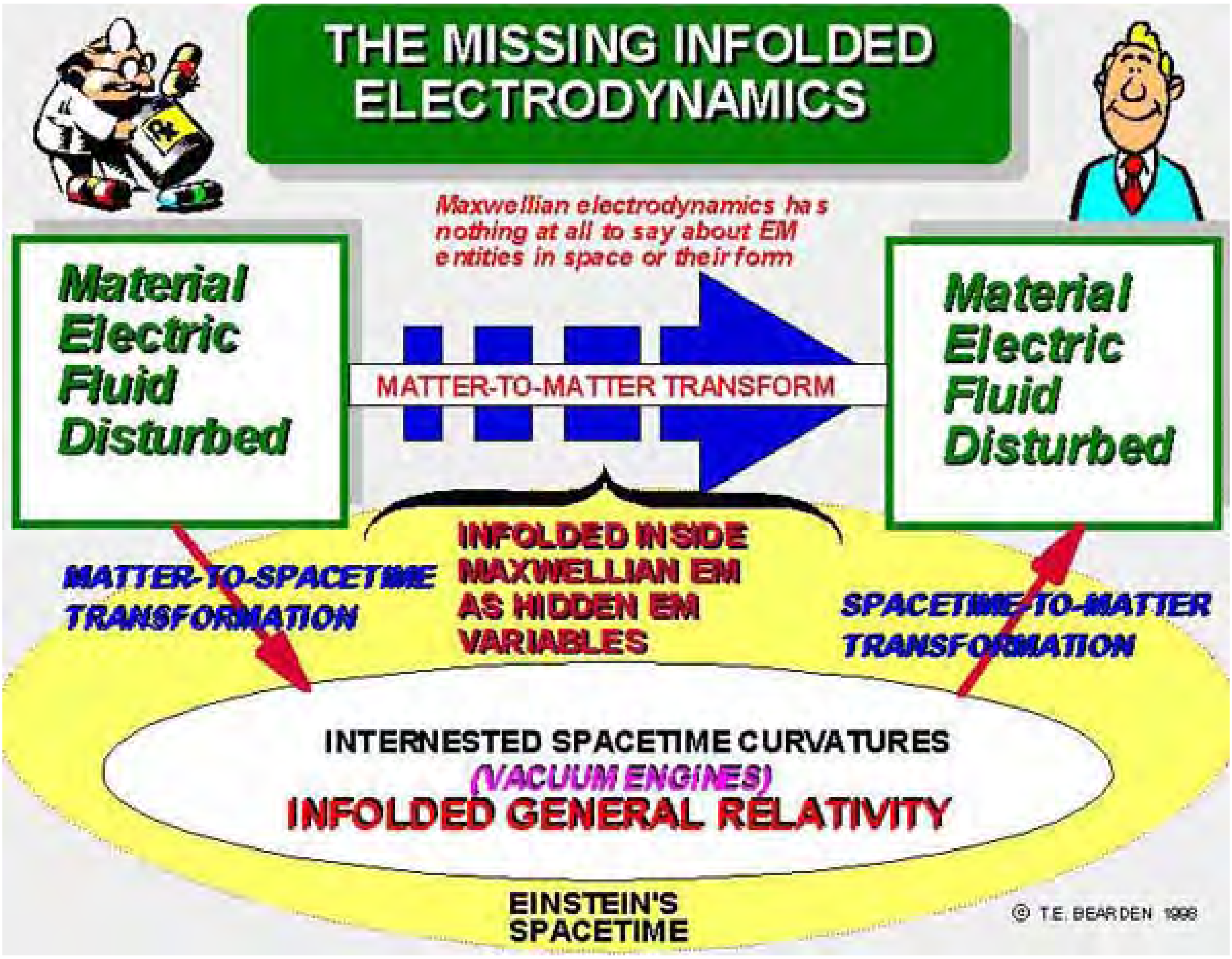

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

MATTER-TO-SPACETIME TRANSFORMATION

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

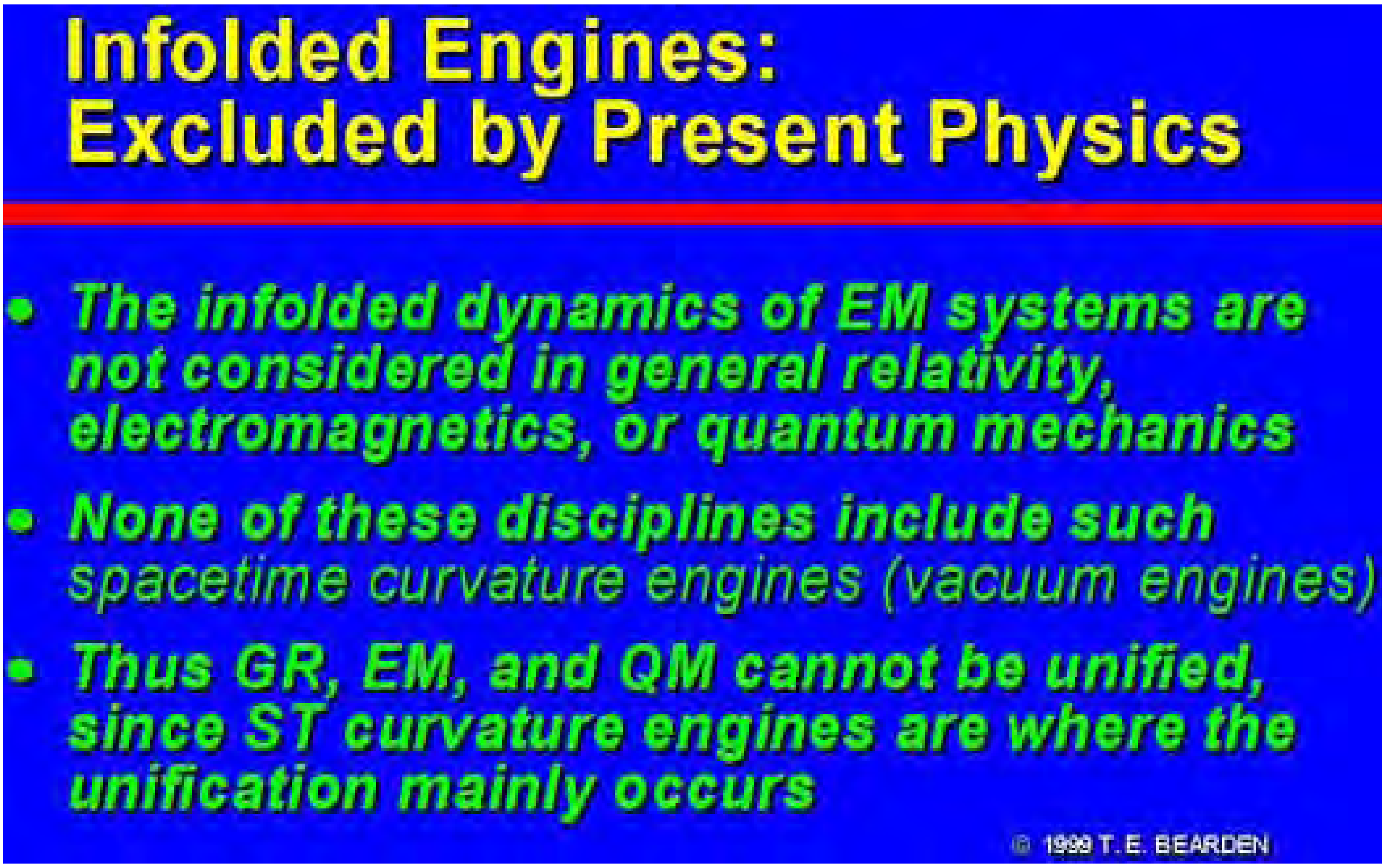

Infolded Engines: Excluded by Present Physics

- *The infolded dynamics of EM systems are not considered in general relativity, electromagnetics, or quantum mechanics*
- *None of these disciplines include such spacetime curvature engines (vacuum engines)*
- *Thus GR, EM, and QM cannot be unified, since ST curvature engines are where the unification mainly occurs*

© 1999 T. E. BEARDEN

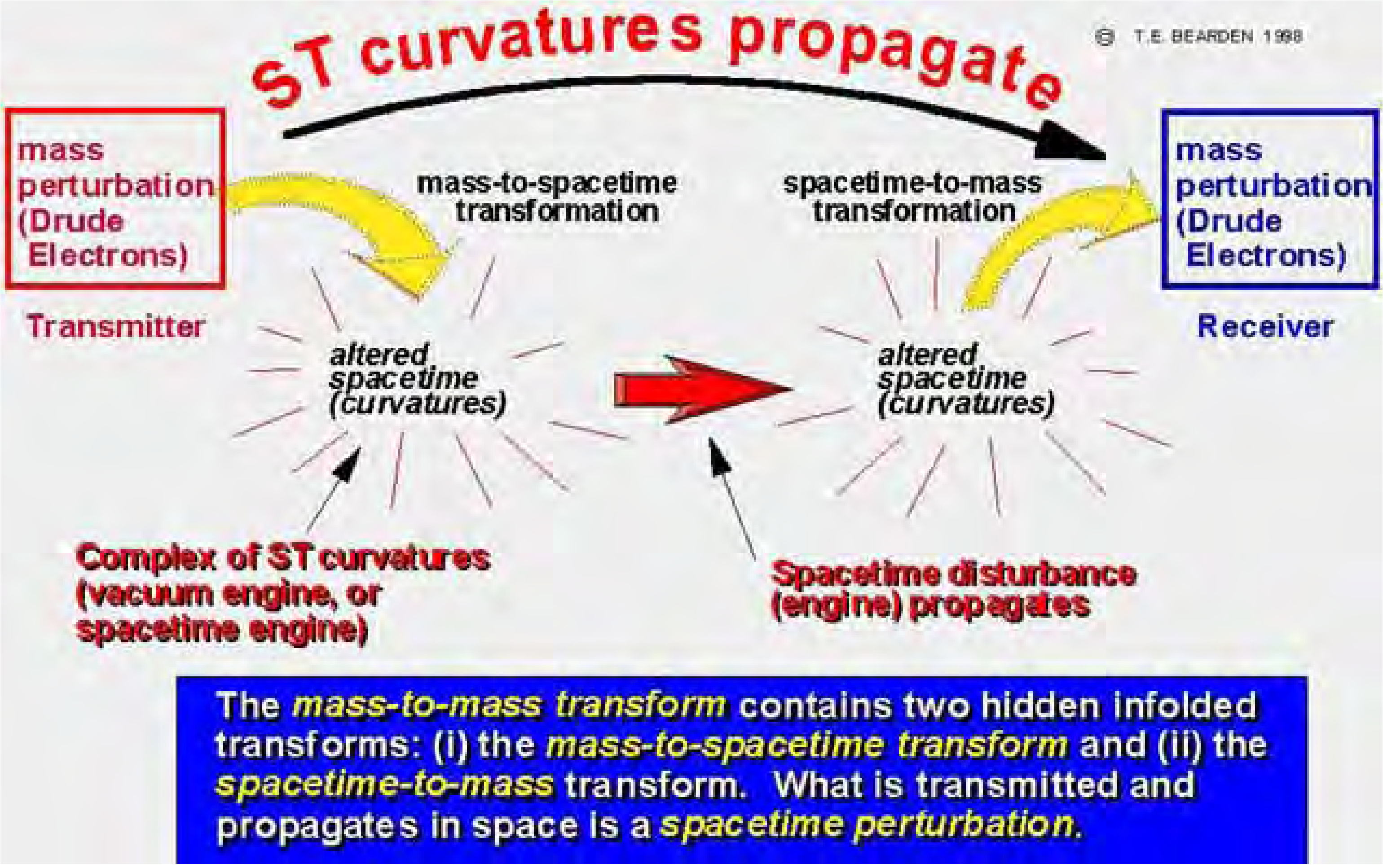

© T.E. BEARDEN 1998

ST curvatures propagate

mass perturbation (Drude Electrons)

Transmitter

mass-to-spacetime transformation

altered spacetime (curvatures)

spacetime-to-mass transformation

altered spacetime (curvatures)

mass perturbation (Drude Electrons)

Receiver

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation*.

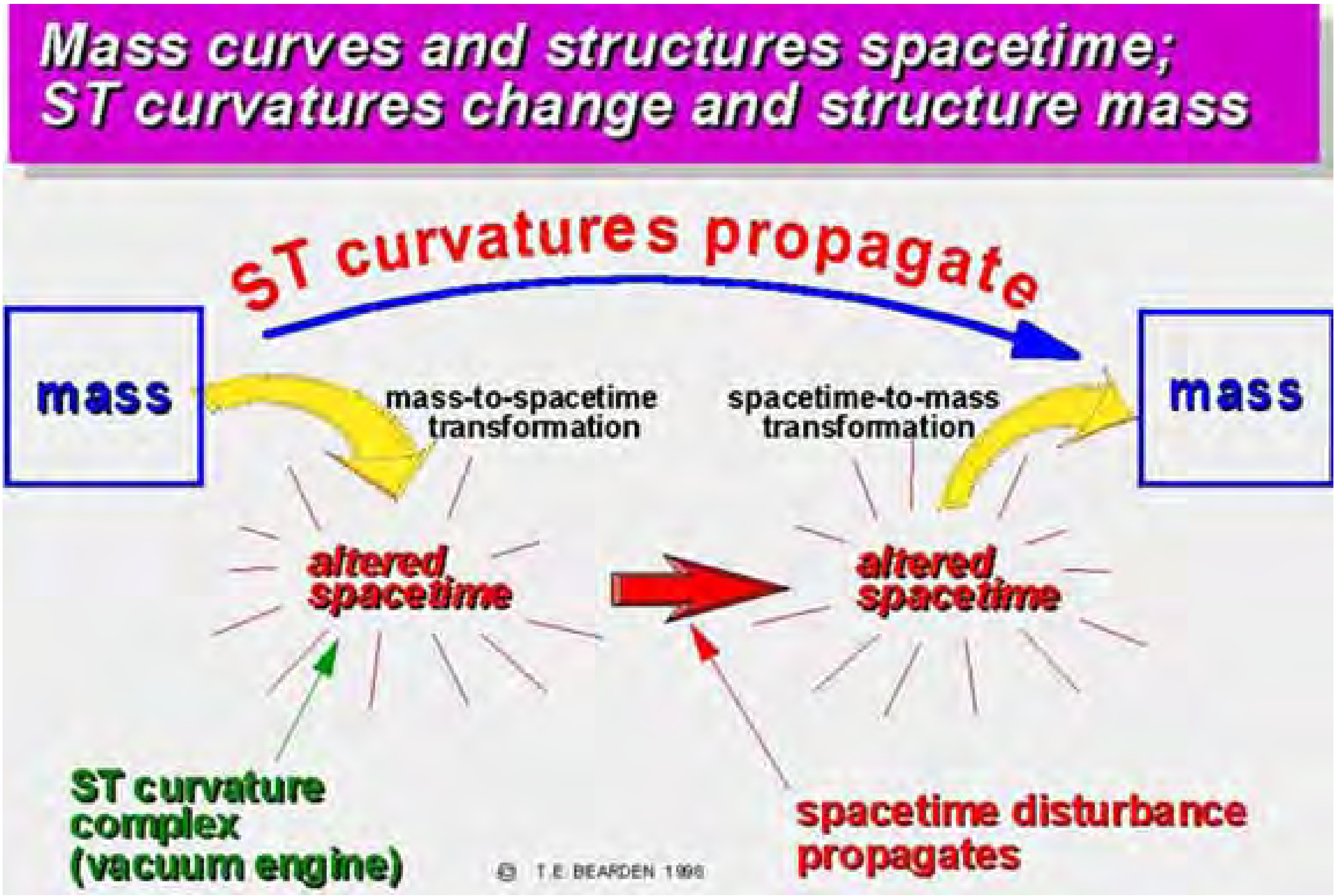

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

ST curvature complex (vacuum engine)

spacetime disturbance propagates

© T.E. BEARDEN 1996

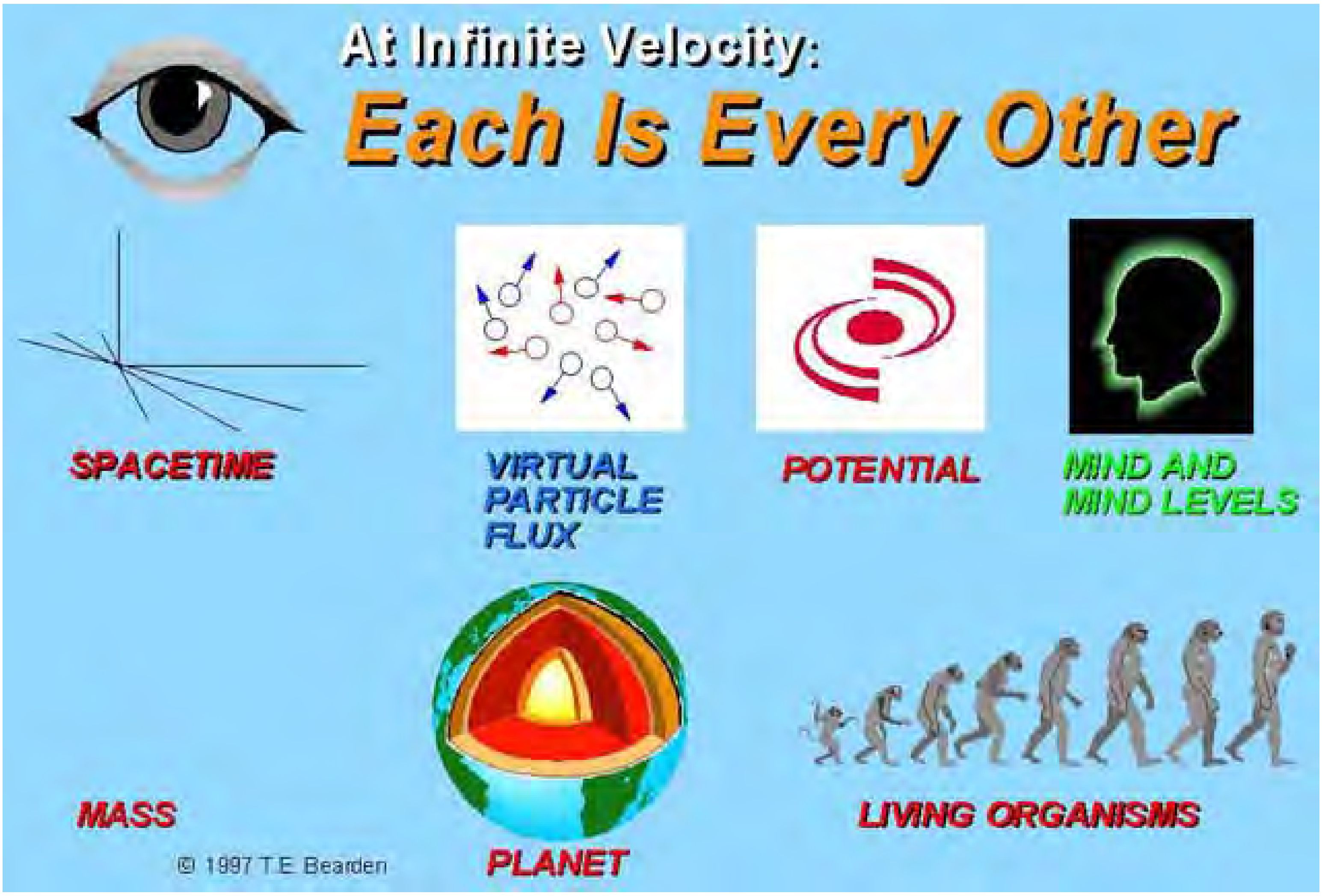

At Infinite Velocity:

Each Is Every Other

SPACETIME

VIRTUAL PARTICLE FLUX

POTENTIAL

MIND AND MIND LEVELS

MASS

PLANET

LIVING ORGANISMS

© 1997 T.E. Bearden

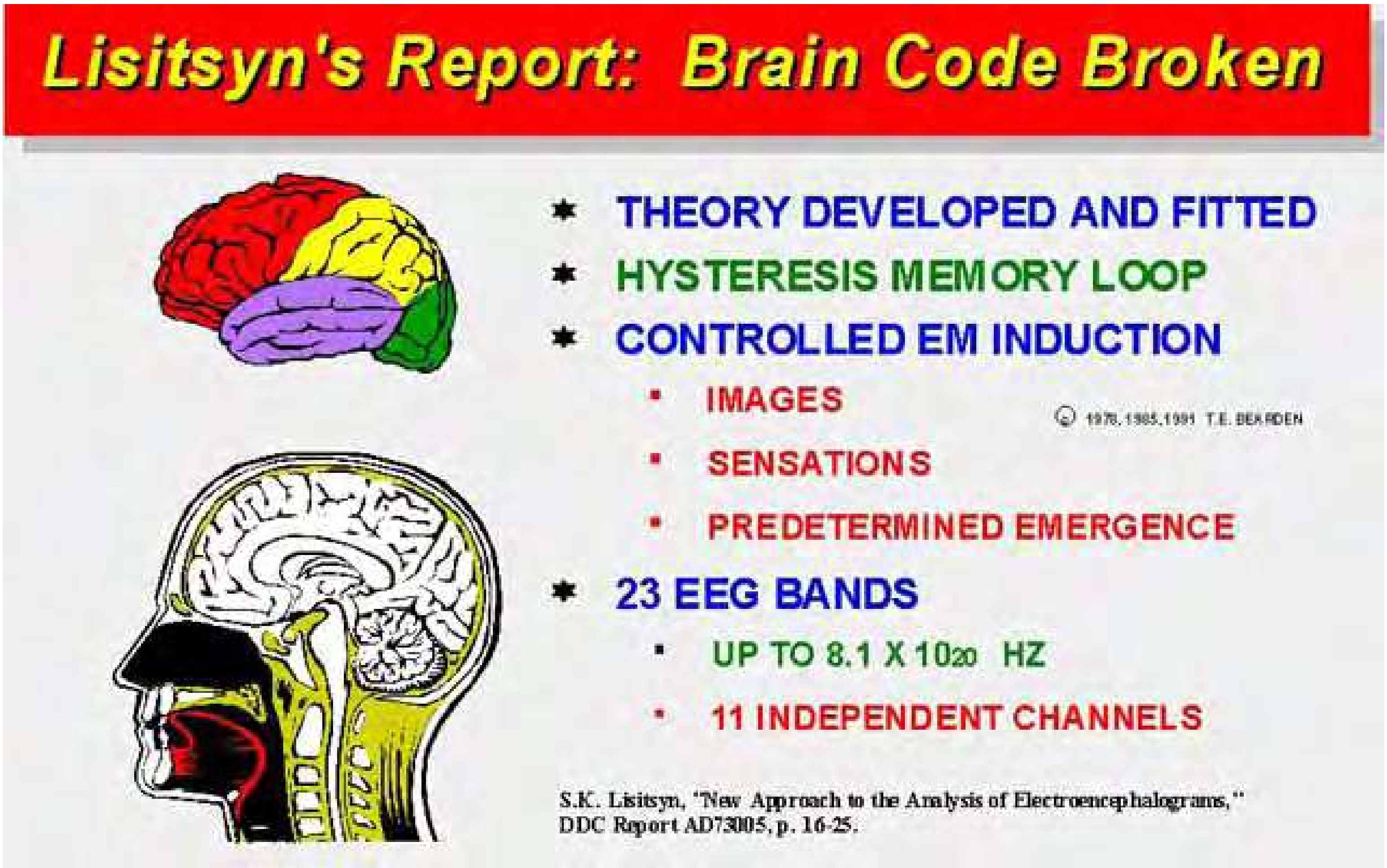

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

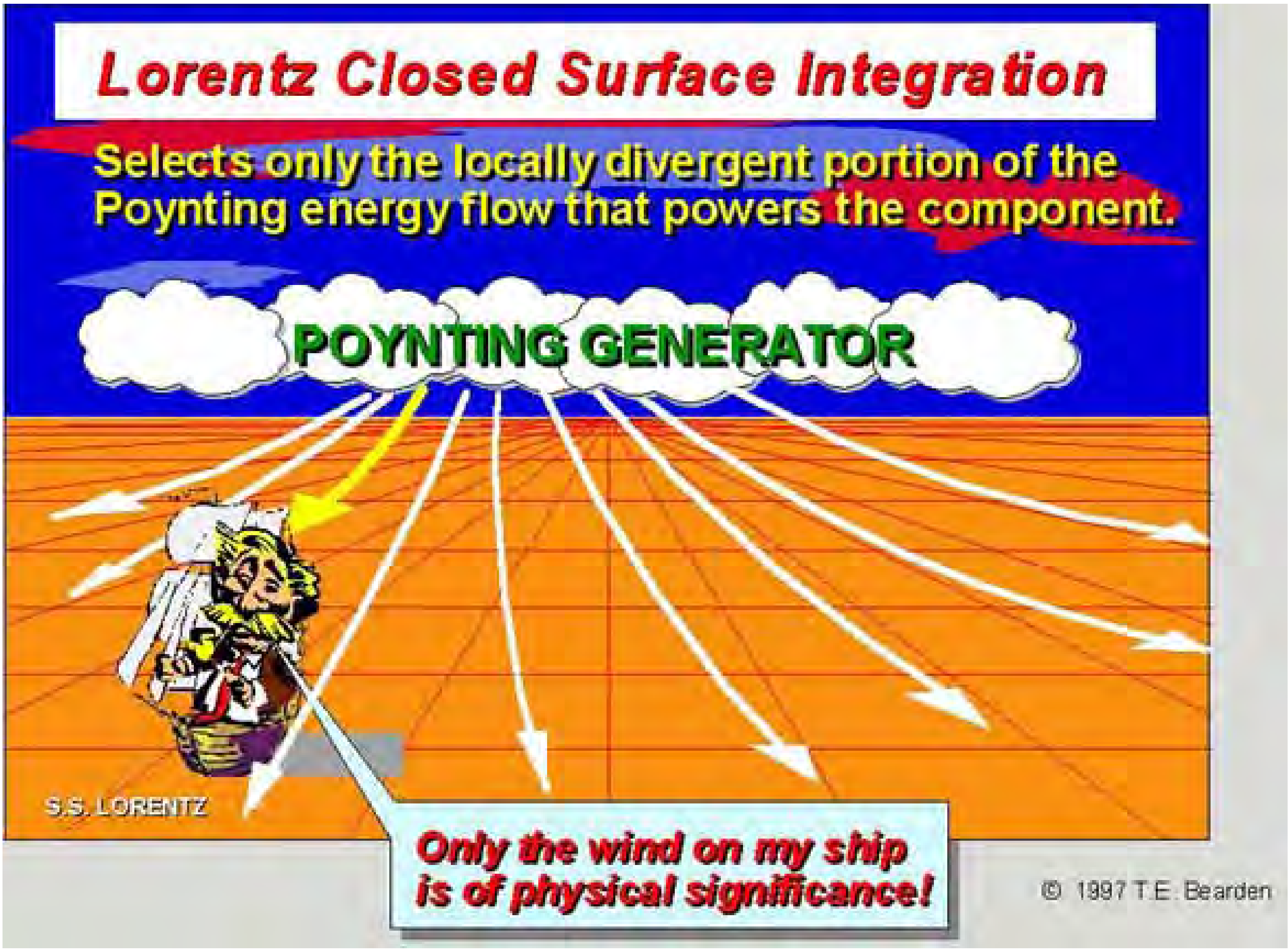

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

POYNTING GENERATOR

Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

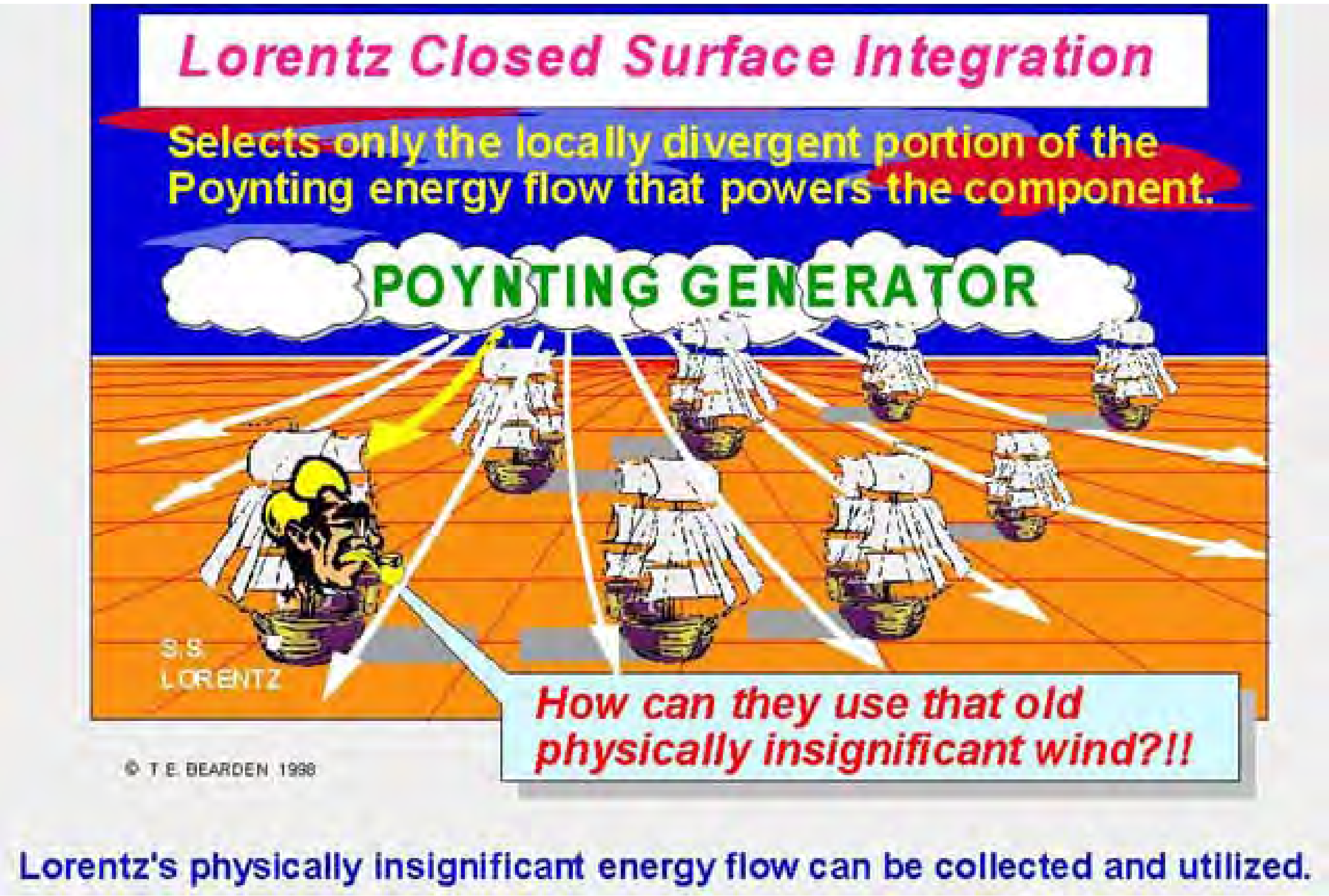

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

POYNTING GENERATOR
S.S. LORENTZ
How can they use that old physically insignificant wind?!!

© T E BEARDEN 1998

Lorentz's physically insignificant energy flow can be collected and utilized.

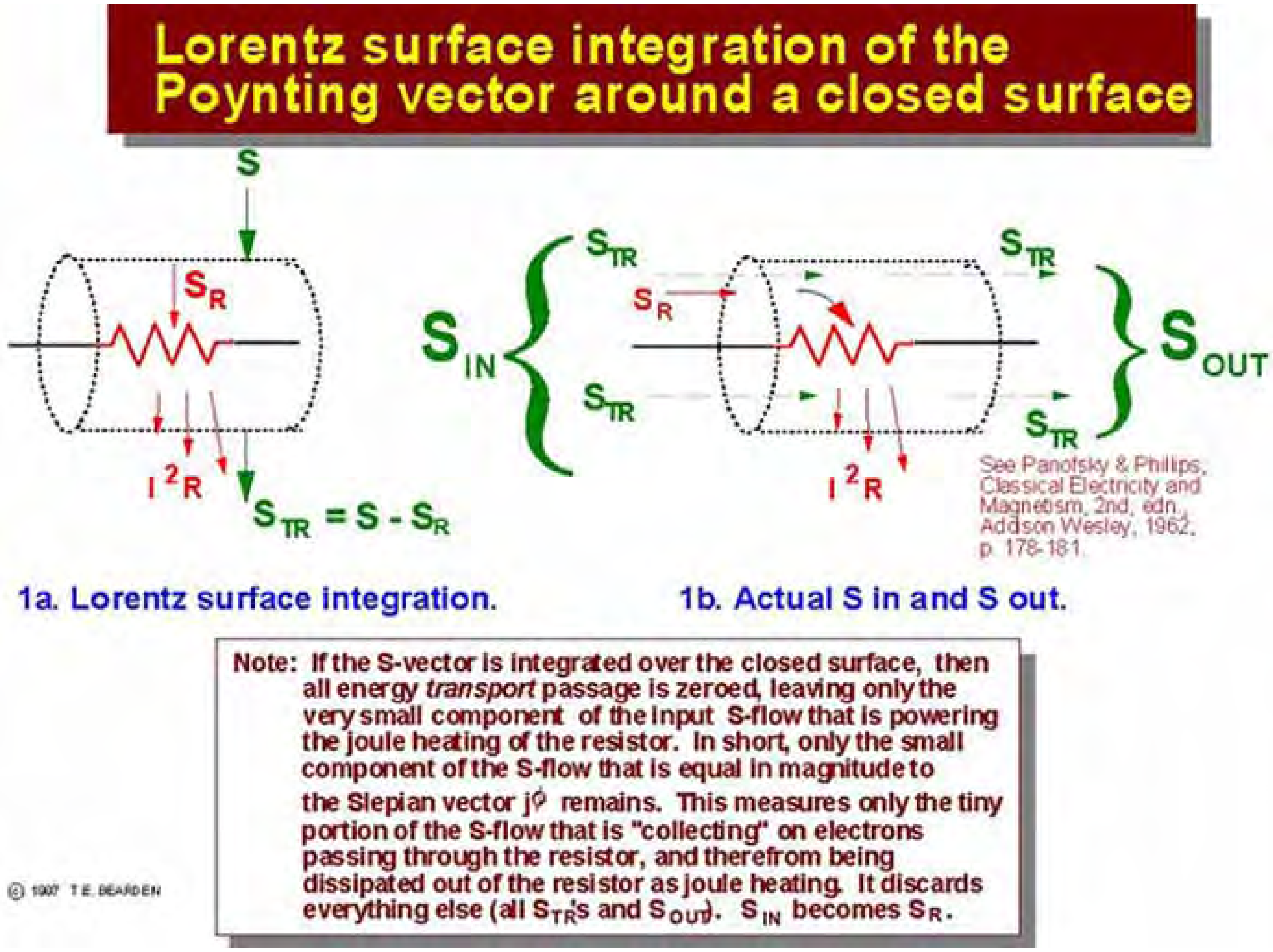

Lorentz surface integration of the Poynting vector around a closed surface

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd, edn., Addison Wesley, 1962, p. 178-181.

1a. Lorentz surface integration.

1b. Actual S in and S out.

Note: If the S-vector is integrated over the closed surface, then all energy *transport* passage is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Slepian vector $j\phi$ remains. This measures only the tiny portion of the S-flow that is "collecting" on electrons passing through the resistor, and therefrom being dissipated out of the resistor as joule heating. It discards everything else (all S_{TR}'s and S_{OUT}). S_{IN} becomes S_R.

© 1997 T.E. BEARDEN

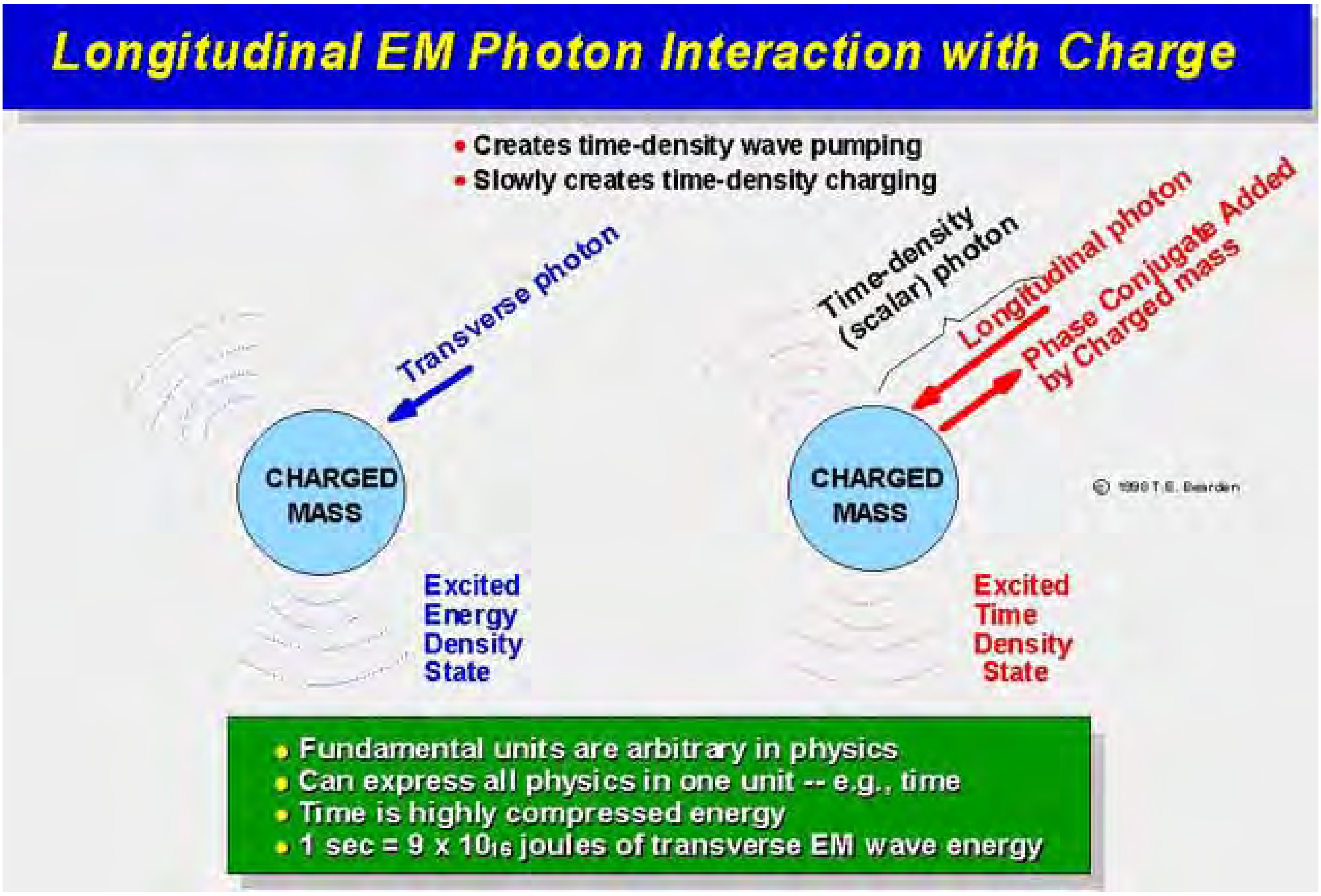

Longitudinal EM Photon Interaction with Charge

- Creates time-density wave pumping
- Slowly creates time-density charging

Transverse photon
CHARGED MASS
Excited Energy Density State
Time-density (scalar) photon
Longitudinal photon
Phase Conjugate Added by Charged mass
CHARGED MASS
Excited Time Density State

© 1990 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = $9 \times 10_{16}$ joules of transverse EM wave energy

Figure 1-8. Infolded biwave composition of a scalar potential.

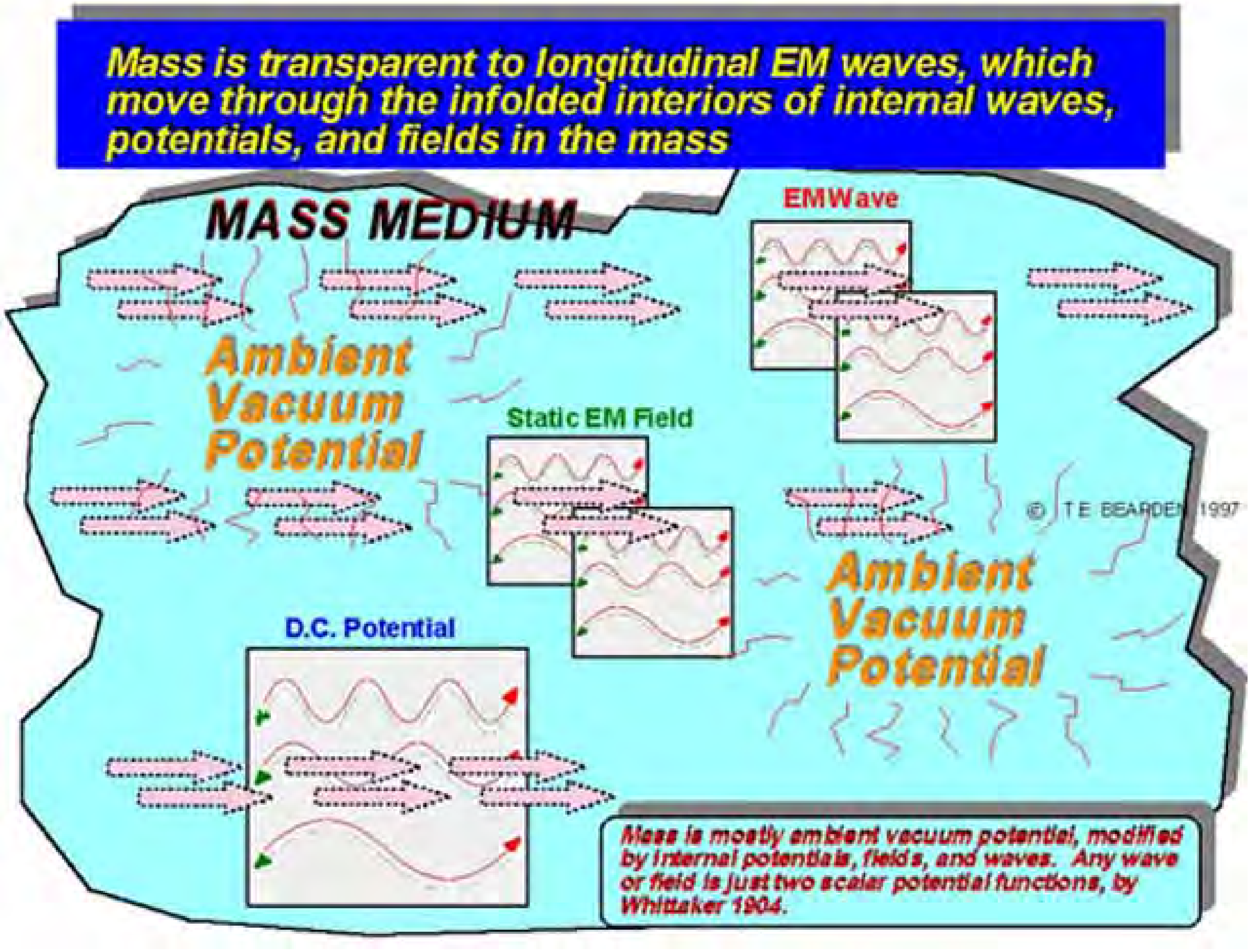

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
EMWave
Ambient Vacuum Potential
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

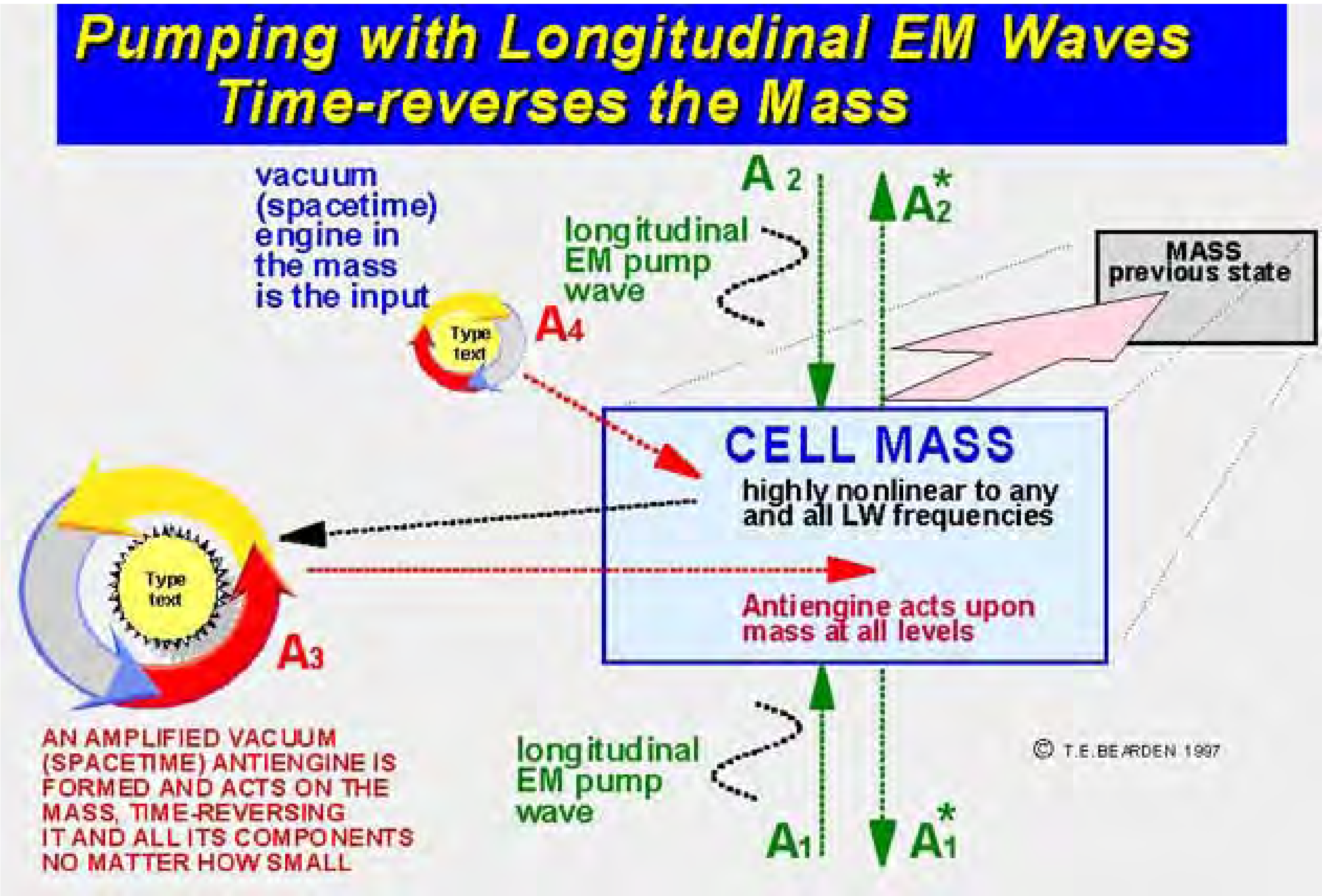

Pumping with Longitudinal EM Waves Time-reverses the Mass

vacuum (spacetime) engine in the mass is the input

longitudinal EM pump wave

MASS previous state

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

longitudinal EM pump wave

© T.E.BEARDEN 1997

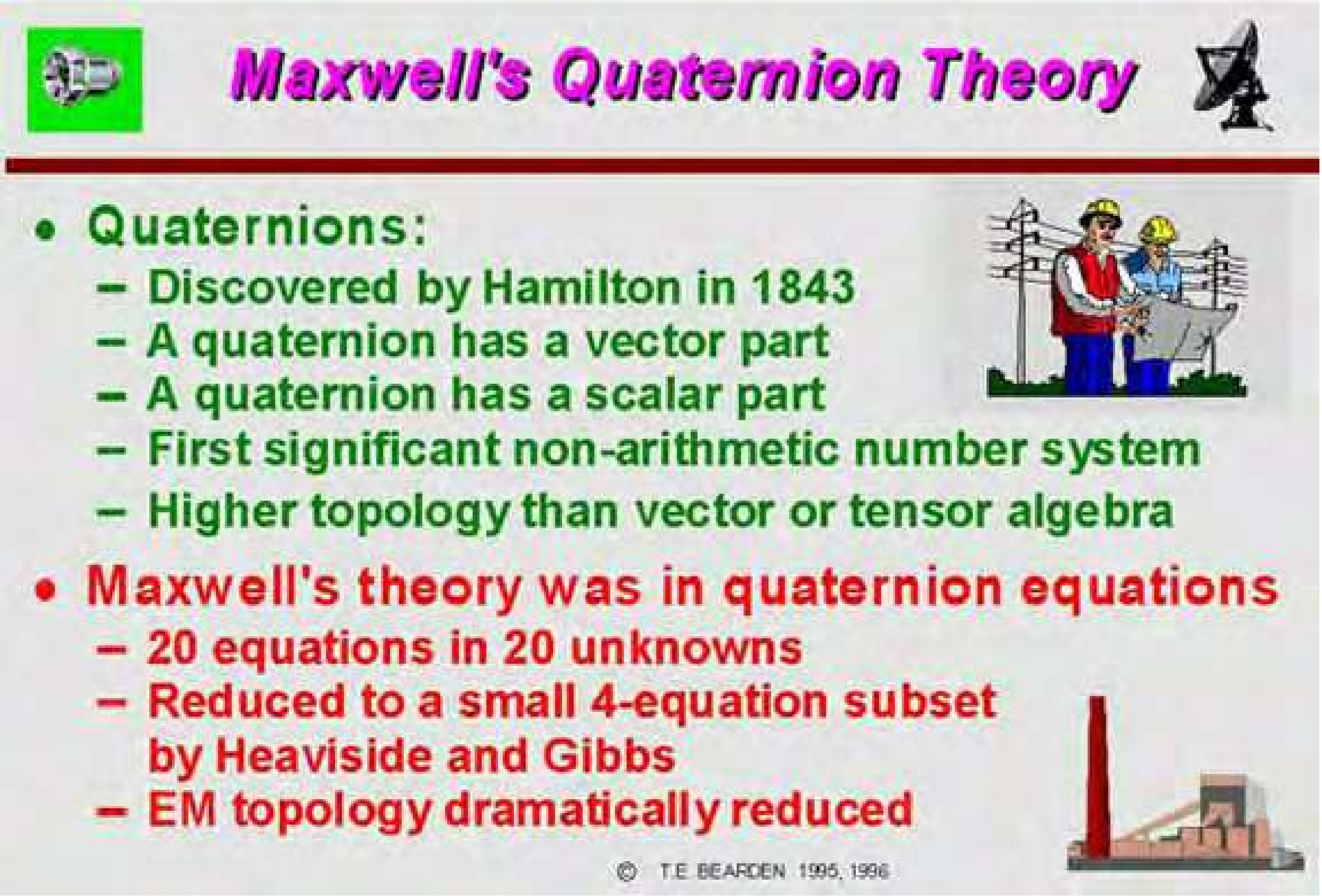

Maxwell's Quaternion Theory

- Quaternions:
  - Discovered by Hamilton in 1843
  - A quaternion has a vector part
  - A quaternion has a scalar part
  - First significant non-arithmetic number system
  - Higher topology than vector or tensor algebra
- Maxwell's theory was in quaternion equations
  - 20 equations in 20 unknowns
  - Reduced to a small 4-equation subset by Heaviside and Gibbs
  - EM topology dramatically reduced

© T.E. BEARDEN 1995, 1996

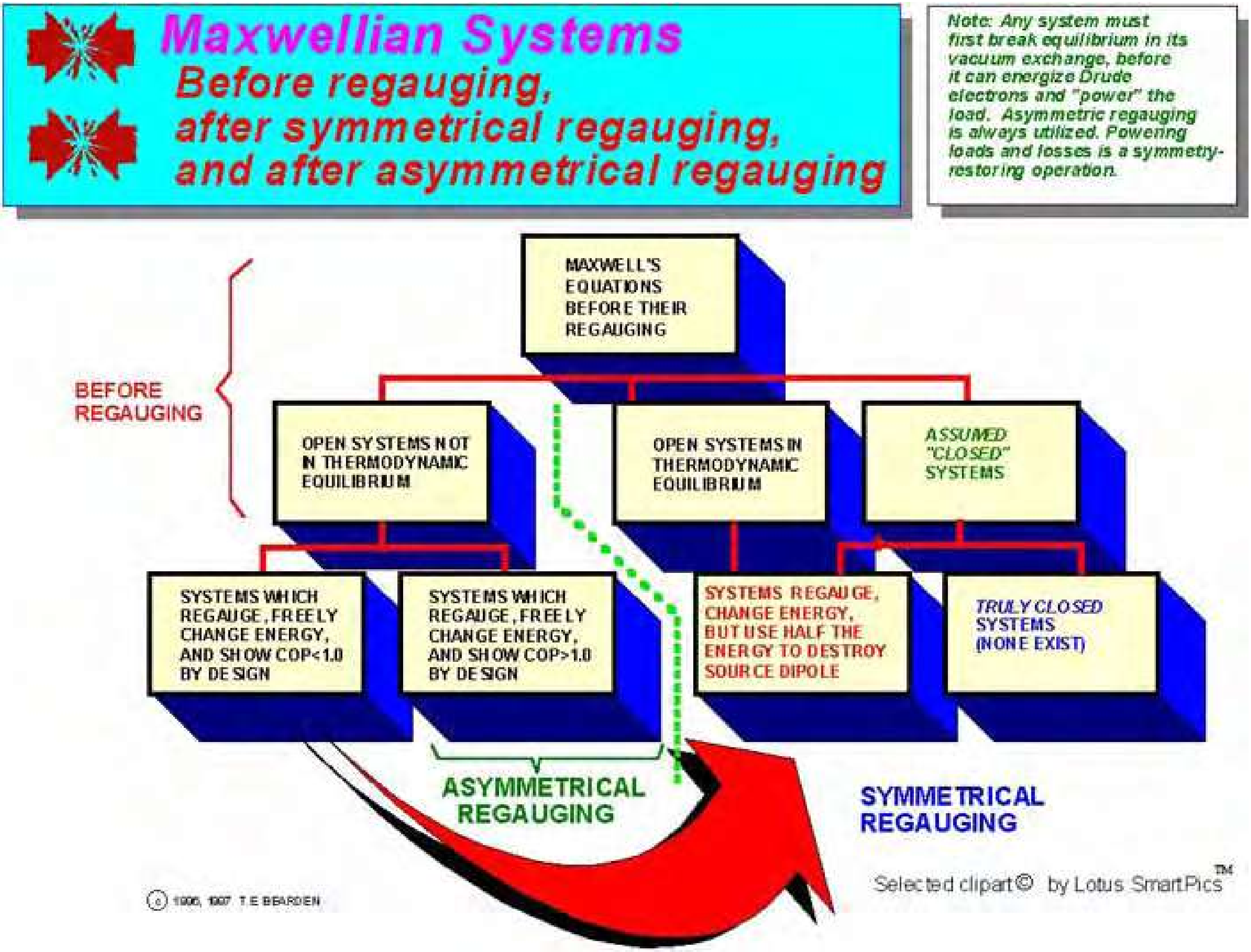

Maxwellian Systems

Before regauging, after symmetrical regauging, and after asymmetrical regauging

Note: Any system must first break equilibrium in its vacuum exchange, before it can energize Drude electrons and "power" the load. Asymmetric regauging is always utilized. Powering loads and losses is a symmetry-restoring operation.

MAXWELL'S EQUATIONS BEFORE THEIR REGAUGING

BEFORE REGAUGING

OPEN SYSTEMS NOT IN THERMODYNAMIC EQUILIBRIUM

OPEN SYSTEMS IN THERMODYNAMIC EQUILIBRIUM

ASSUMED "CLOSED" SYSTEMS

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP<1.0 BY DESIGN

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP>1.0 BY DESIGN

SYSTEMS REGAUGE, CHANGE ENERGY, BUT USE HALF THE ENERGY TO DESTROY SOURCE DIPOLE

TRULY CLOSED SYSTEMS (NONE EXIST)

ASYMMETRICAL REGAUGING

SYMMETRICAL REGAUGING

Selected clipart© by Lotus SmartPics™

ⓒ 1996, 1997 T.E. BEARDEN

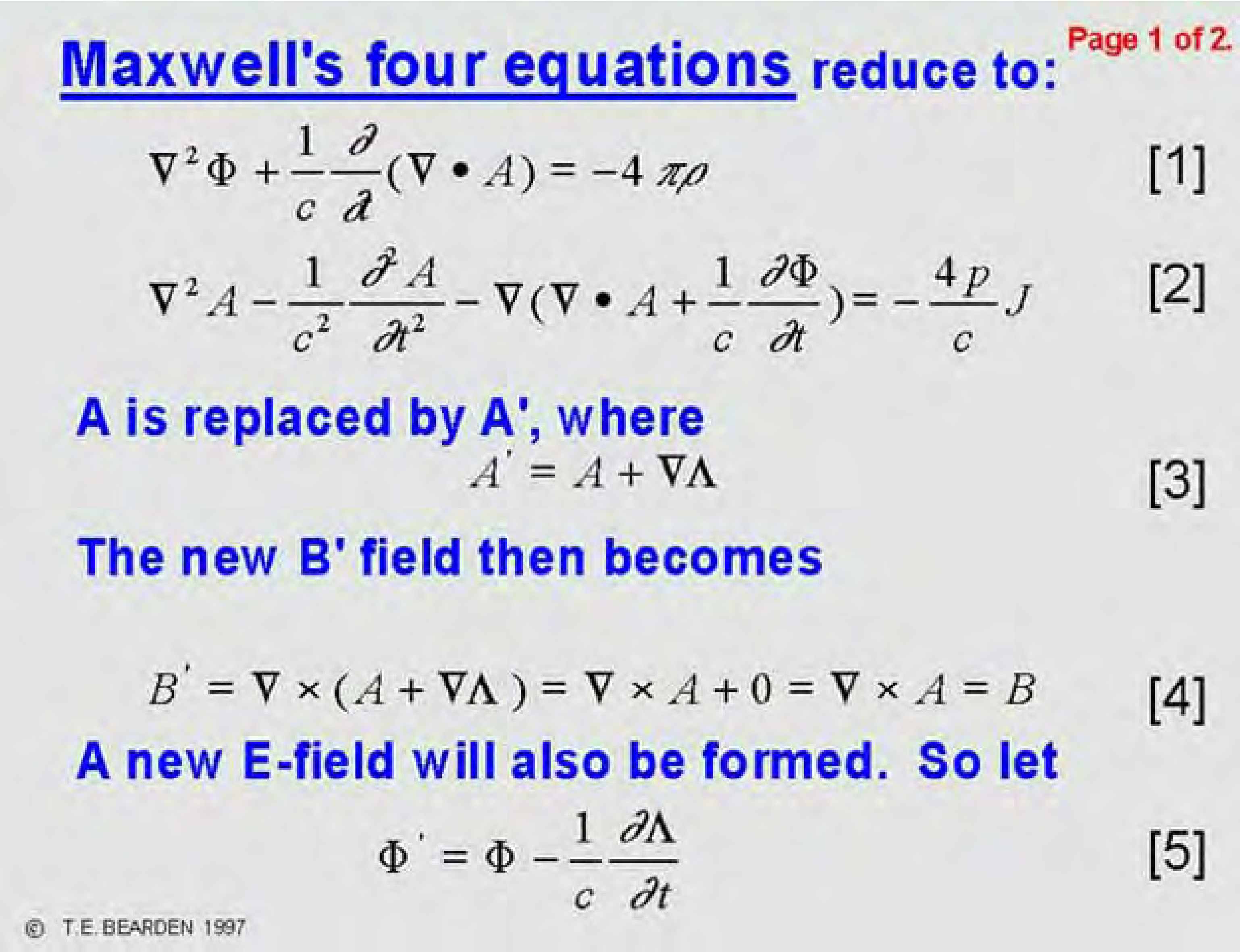

Page 1 of 2.

Maxwell's four equations reduce to:

$$\nabla^2 \Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet A) = -4\pi\rho \qquad [1]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} - \nabla\left(\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t}\right) = -\frac{4p}{c} J \qquad [2]$$

A is replaced by A', where

$$A' = A + \nabla\Lambda \qquad [3]$$

The new B' field then becomes

$$B' = \nabla \times (A + \nabla\Lambda) = \nabla \times A + 0 = \nabla \times A = B \qquad [4]$$

A new E-field will also be formed. So let

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t} \qquad [5]$$

© T.E. BEARDEN 1997

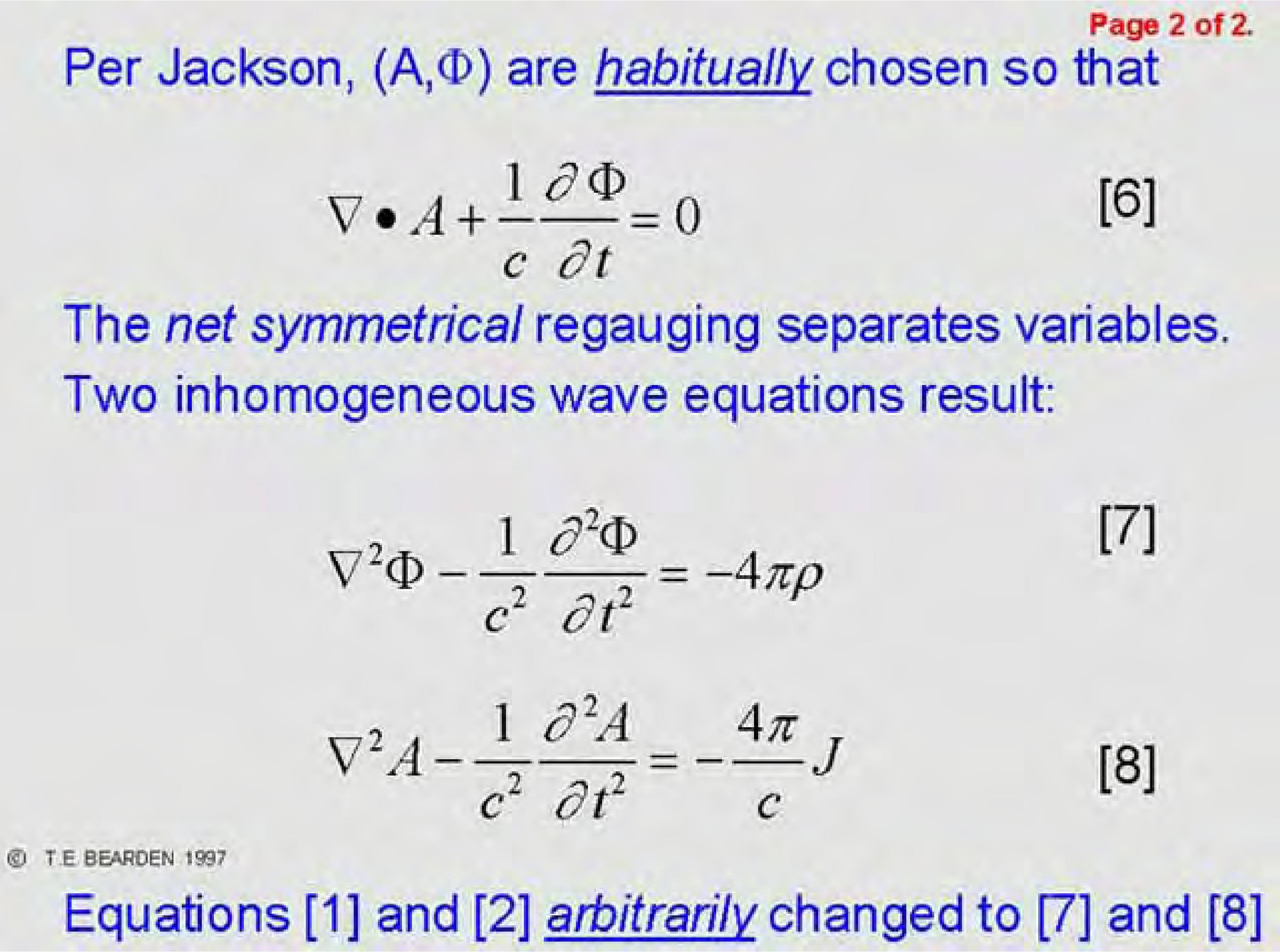

Per Jackson, (A,Φ) are *habitually* chosen so that

$$\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [6]$$

The *net symmetrical* regauging separates variables.

Two inhomogeneous wave equations result:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [7]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{4\pi}{c} J \qquad [8]$$

© T.E. BEARDEN 1997

Equations [1] and [2] *arbitrarily* changed to [7] and [8]

The Tom Bearden Website

Energy, Observations and Comments

---

1. A Conservative Field Requires a Single-Valued Potential
2. Earth-Electrosphere and Earth-Ionosphere Capacitors
3. Earth's Poynting Energy Flow
4. The Hidden Energy Flow around us
5. Electrostatic Fields of the Earth's Electrosphere Capacitor in Fair Weather and Thunderstorms
6. Flow of Current and Energy Into, Through, and Along a Conductor
7. Energy "collected" on a charge
8. Fiber Fuse Effect
9. Impact of Vacuum Engineering
10. Lawandy's Experiment
11. Field may be nonconservative for a multivalued potential
12. The Ostrich Problem - Part 1
13. The Ostrich Problem - Part 2
14. Russia had Overunity Devices in 1930s
15. Energy output of human body
16. A "static" Poynting S-flow free-energy generator
17. Topology and Energy Conservation
18. Block Diagram: Sweet's Vacuum Triode

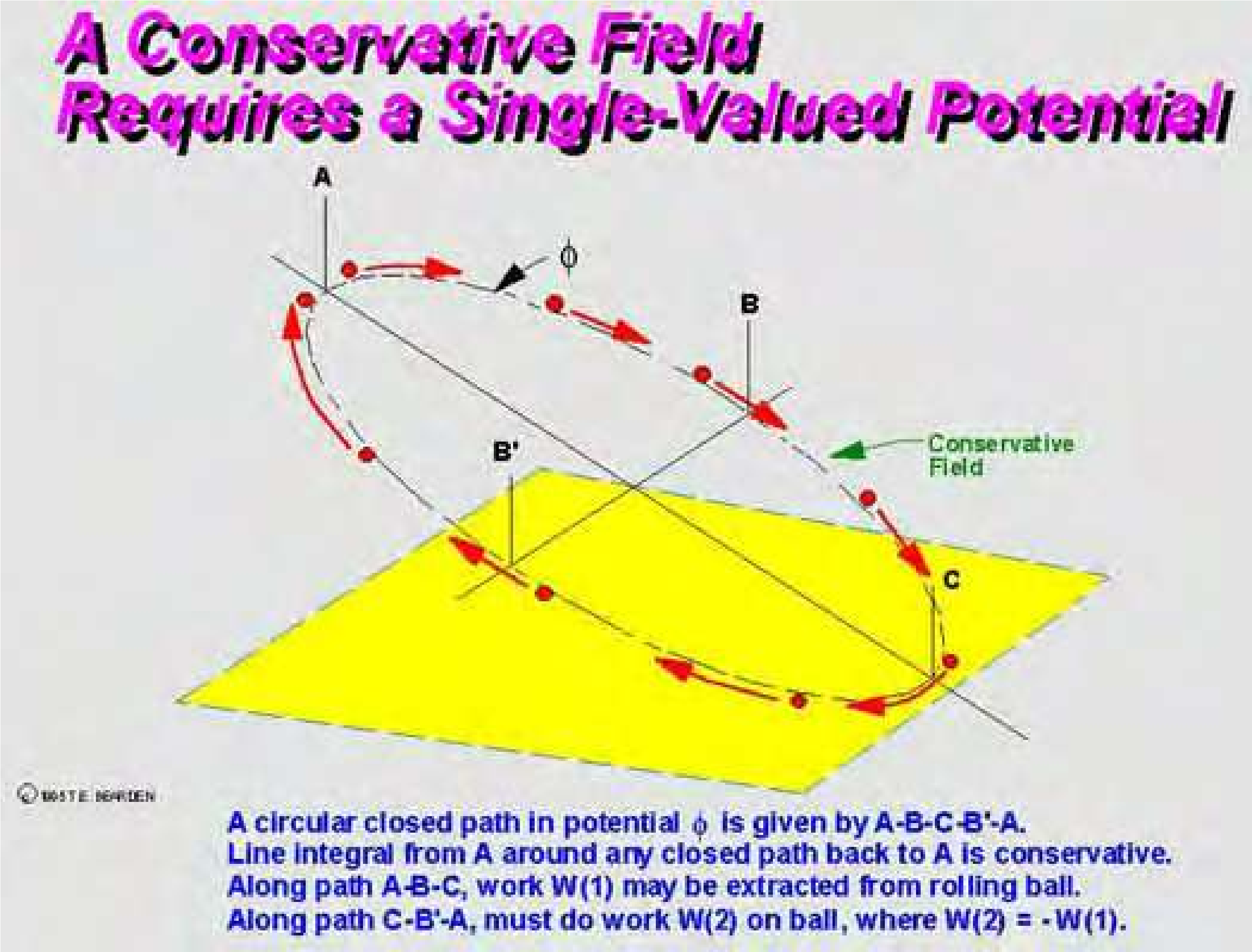

A Conservative Field Requires a Single-Valued Potential

© 1995 T.E. BEARDEN

A circular closed path in potential ϕ is given by A-B-C-B'-A.
Line integral from A around any closed path back to A is conservative.
Along path A-B-C, work W(1) may be extracted from rolling ball.
Along path C-B'-A, must do work W(2) on ball, where W(2) = - W(1).

Return to Slide Index

Next Slide

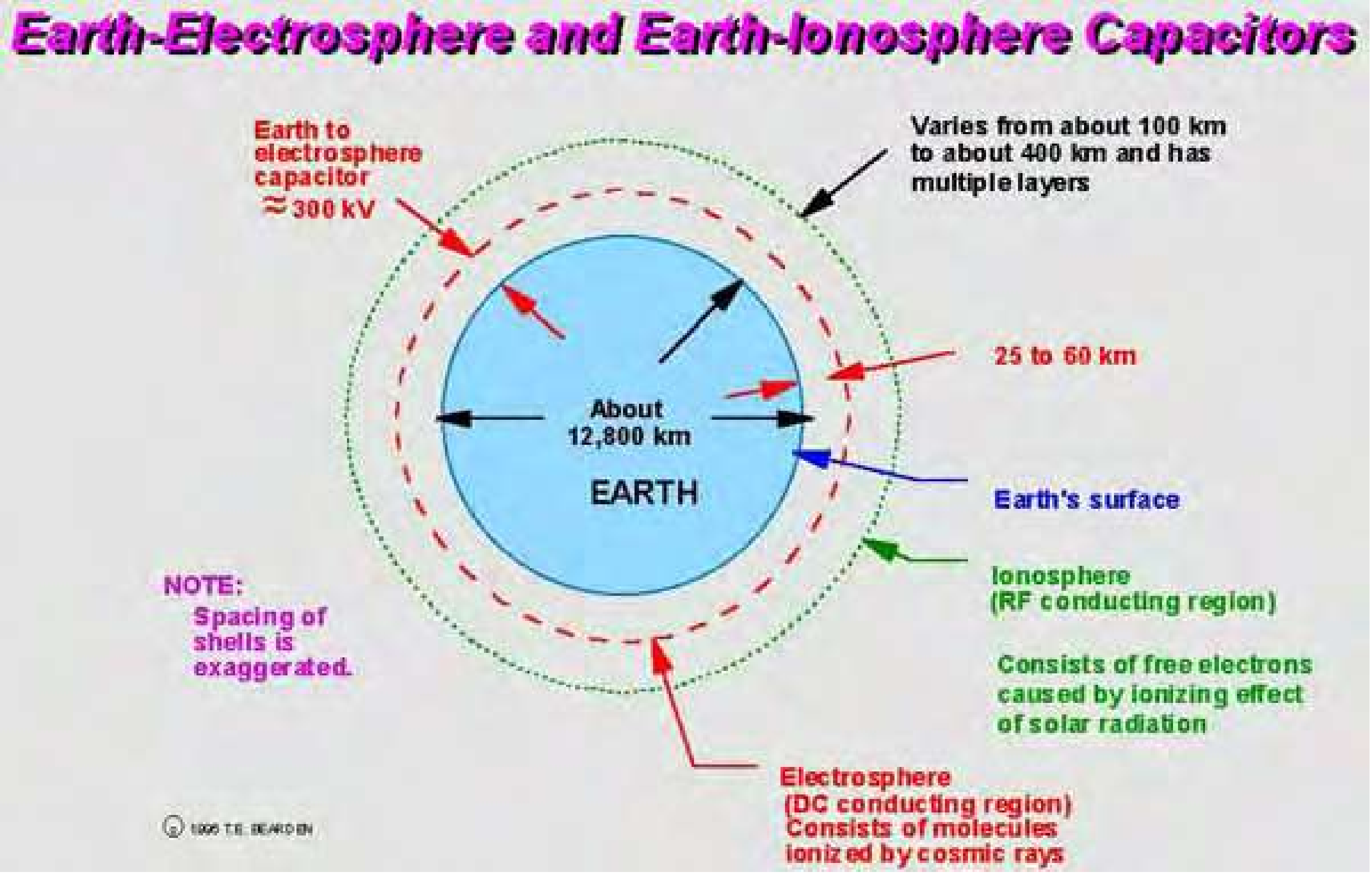

Earth-Electrosphere and Earth-Ionosphere Capacitors

Earth to electrosphere capacitor
≈ 300 kV

Varies from about 100 km to about 400 km and has multiple layers

About 12,800 km

EARTH

25 to 60 km

Earth's surface

Ionosphere (RF conducting region)

Consists of free electrons caused by ionizing effect of solar radiation

NOTE:
Spacing of shells is exaggerated.

Electrosphere (DC conducting region)
Consists of molecules ionized by cosmic rays

© 1996 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

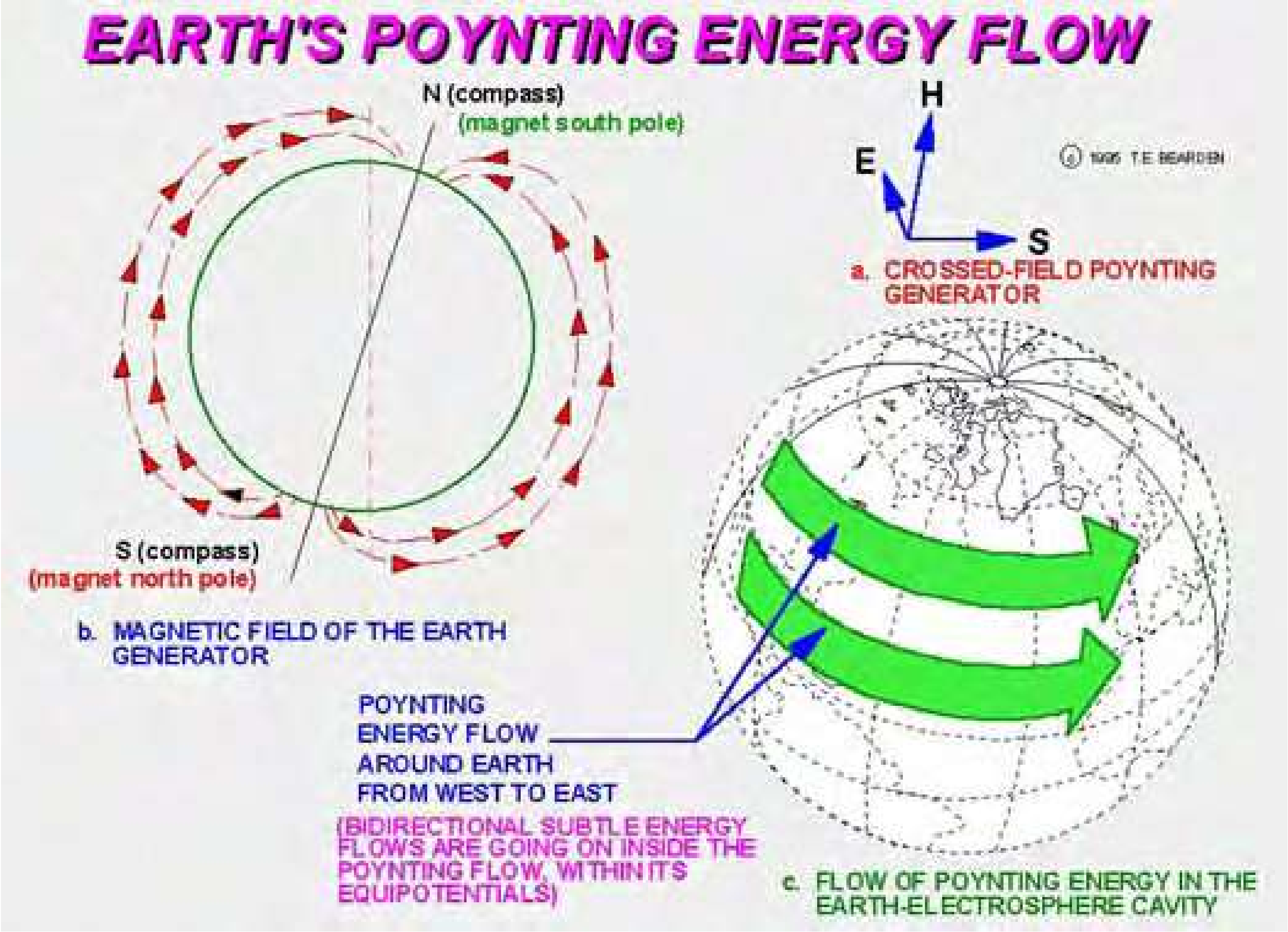
N (compass)
(magnet south pole)
H
E
S
© 1995 T.E. BEARDEN
S (compass)
(magnet north pole)

EARTH'S POYNTING ENERGY FLOW

a. CROSSED-FIELD POYNTING GENERATOR

b. MAGNETIC FIELD OF THE EARTH GENERATOR

POYNTING ENERGY FLOW AROUND EARTH FROM WEST TO EAST

(BIDIRECTIONAL SUBTLE ENERGY FLOWS ARE GOING ON INSIDE THE POYNTING FLOW, WITHIN ITS EQUIPOTENTIALS)

c. FLOW OF POYNTING ENERGY IN THE EARTH-ELECTROSPHERE CAVITY

Return to Slide Index

Next Slide

Previous Slide

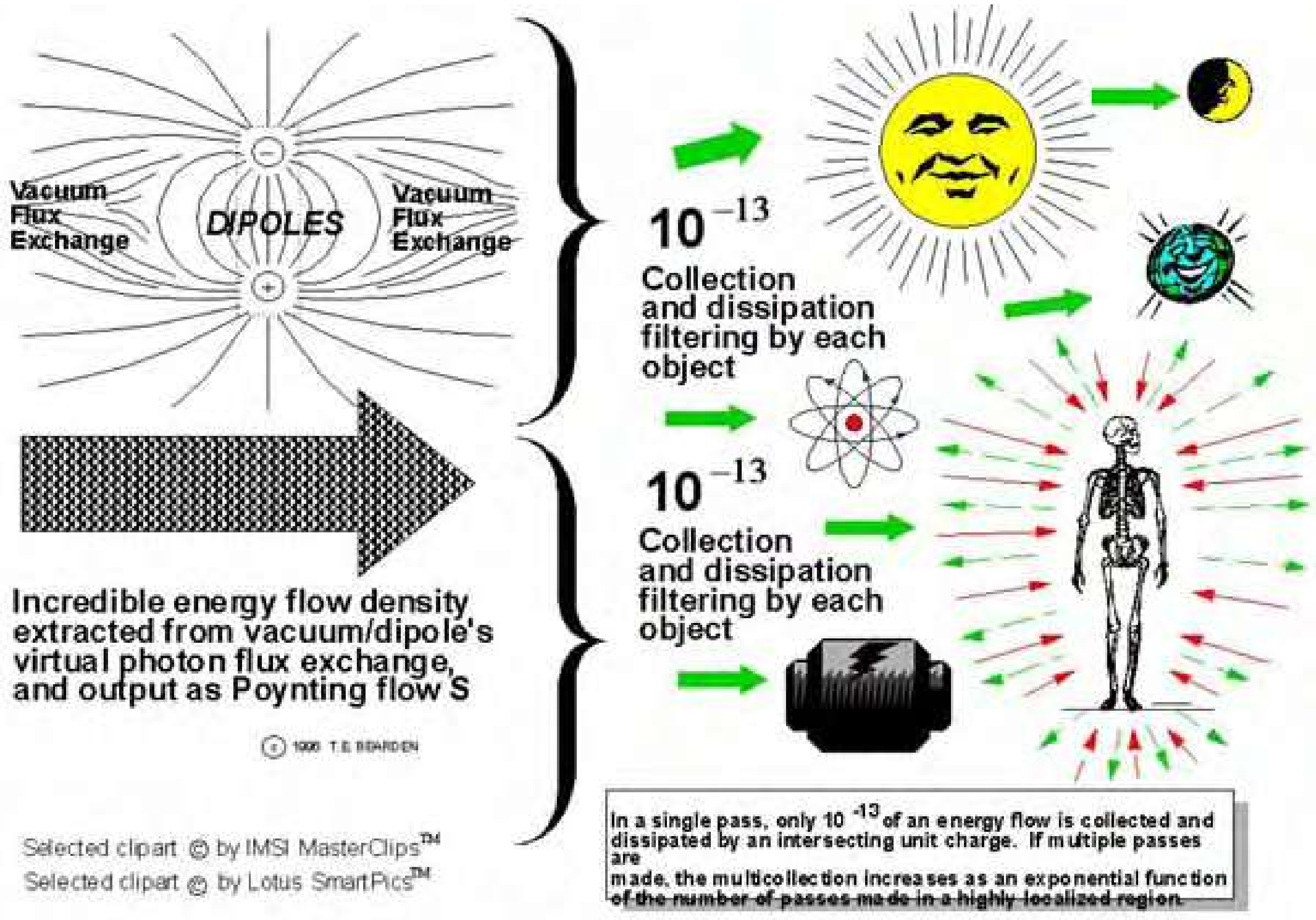
Vacuum
Flux
Exchange
DIPOLES
Vacuum
Flux
Exchange
10^{-13}
Collection and dissipation filtering by each object
10^{-13}
Collection and dissipation filtering by each object
© 1996 T.E. BEARDEN
Selected clipart © by IMSI MasterClips™
Selected clipart © by Lotus SmartPics™

Incredible energy flow density extracted from vacuum/dipole's virtual photon flux exchange, and output as Poynting flow S

In a single pass, only 10^{-13} of an energy flow is collected and dissipated by an intersecting unit charge. If multiple passes are made, the multicollection increases as an exponential function of the number of passes made in a highly localized region.

Figure 19. The world of physical matter exists behind a dramatic 10^{-13} filter in a fiery cauldron of fierce and primeval EM energy flow.

Return to Slide Index

Next Slide

Previous Slide

Electrostatic Fields of the Earth-Electrosphere Capacitor in Fair Weather and Thunderstorms

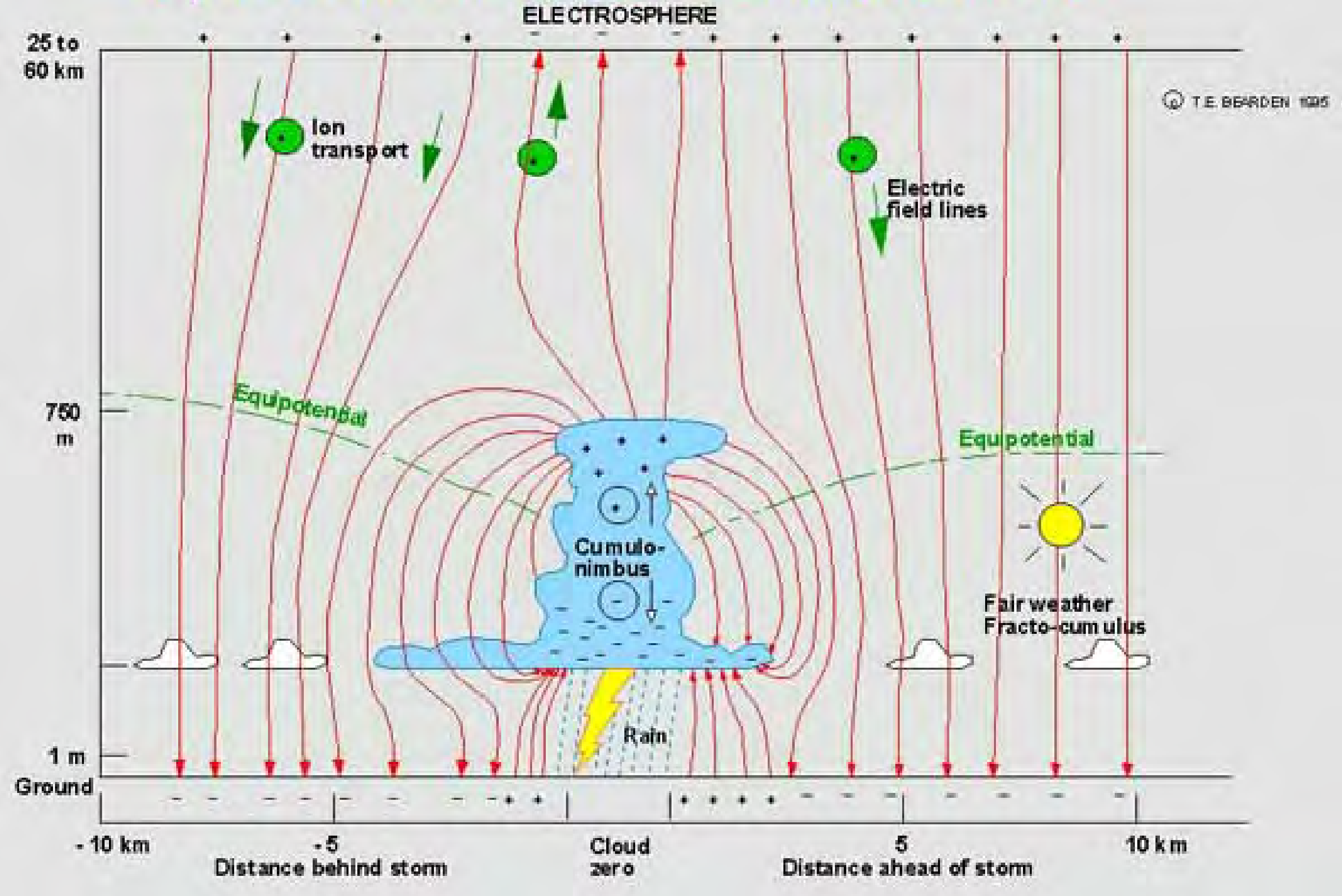
ELECTROSPHERE
25 to
60 km
750
m
1 m
Ground
Ion
transport
Electric
field lines
Equipotential
Equipotential
Cumulo-
nimbus
Rain
Fair weather
Fracto-cumulus
© T.E. BEARDEN 1995
- 10 km
- 5
Distance behind storm
Cloud
zero
5
Distance ahead of storm
10 km

[Return to Slide Index]

[Next Slide]

[Previous Slide]

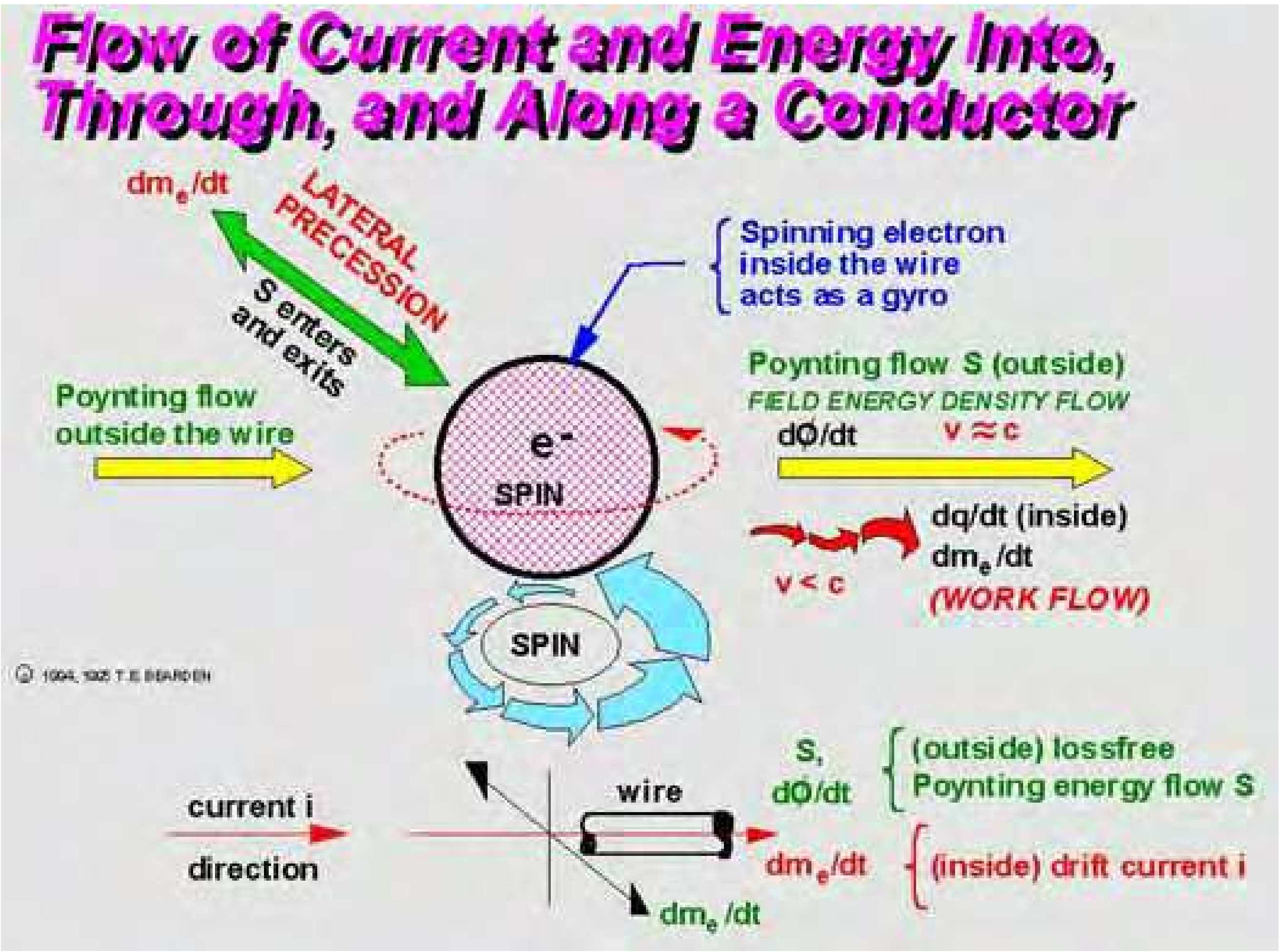

Flow of Current and Energy Into, Through, and Along a Conductor

dm_e/dt

LATERAL PRECESSION

S enters and exits

Spinning electron
inside the wire
acts as a gyro

Poynting flow
outside the wire

e^-

SPIN

Poynting flow S (outside)

FIELD ENERGY DENSITY FLOW

$d\phi/dt$ $v \approx c$

dq/dt (inside)

dm_e/dt

(WORK FLOW)

$v < c$

SPIN

© 1994, 1995 T.E. BEARDEN

current i

direction

wire

S, $d\phi/dt$ (outside) lossfree
Poynting energy flow S

dm_e/dt (inside) drift current i

dm_e/dt

Return to Slide Index

Next Slide

Previous Slide

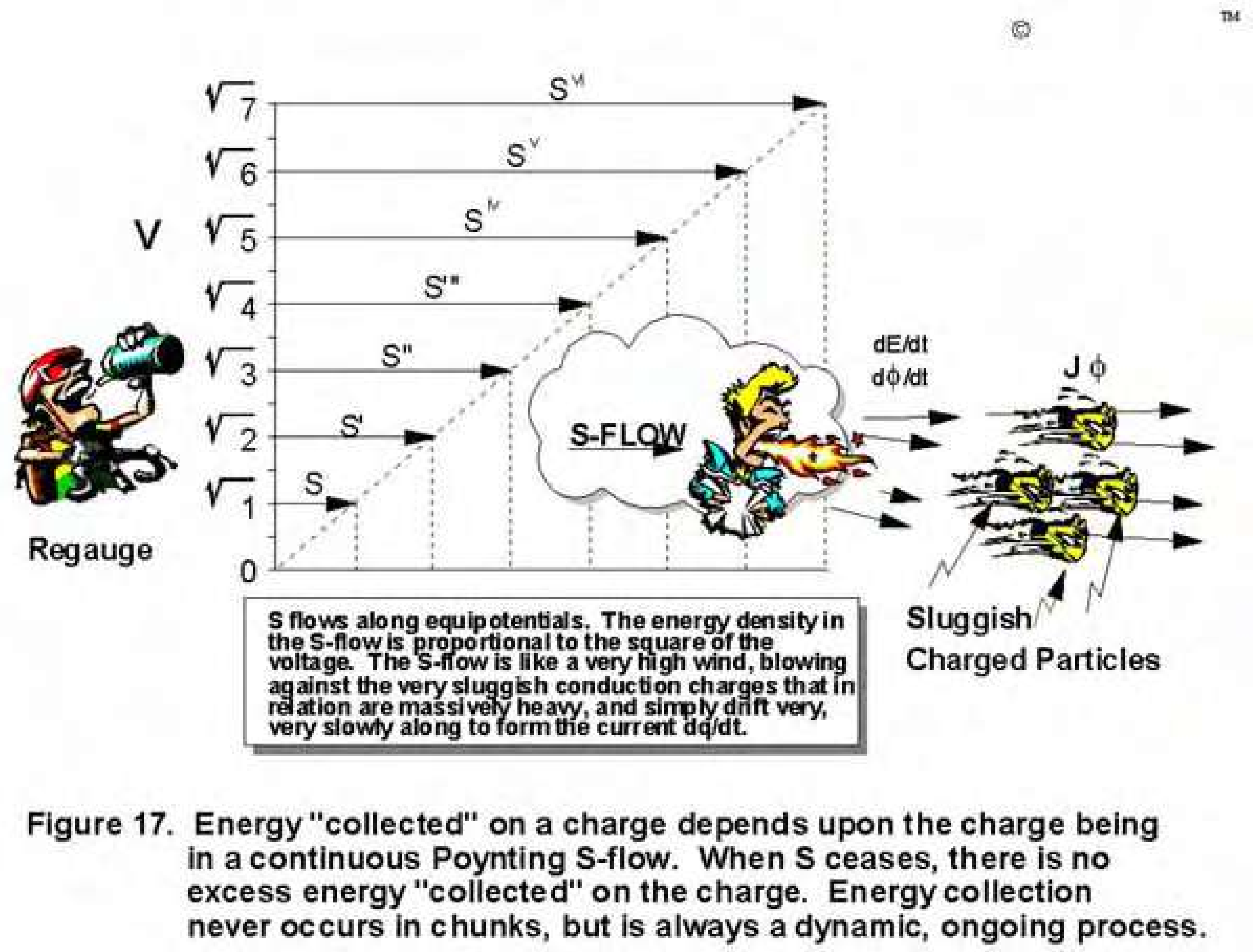
©
™
V
$\sqrt{7}$
$\sqrt{6}$
$\sqrt{5}$
$\sqrt{4}$
$\sqrt{3}$
$\sqrt{2}$
$\sqrt{1}$
0
S^{vi}
S^{v}
S^{iv}
S'''
S''
S'
S
Regauge
S-FLOW
dE/dt
dϕ/dt
J ϕ
Sluggish
Charged Particles

S flows along equipotentials. The energy density in the S-flow is proportional to the square of the voltage. The S-flow is like a very high wind, blowing against the very sluggish conduction charges that in relation are massively heavy, and simply drift very, very slowly along to form the current dq/dt.

Figure 17. Energy "collected" on a charge depends upon the charge being in a continuous Poynting S-flow. When S ceases, there is no excess energy "collected" on the charge. Energy collection never occurs in chunks, but is always a dynamic, ongoing process.

Return to Slide Index

Next Slide

Previous Slide

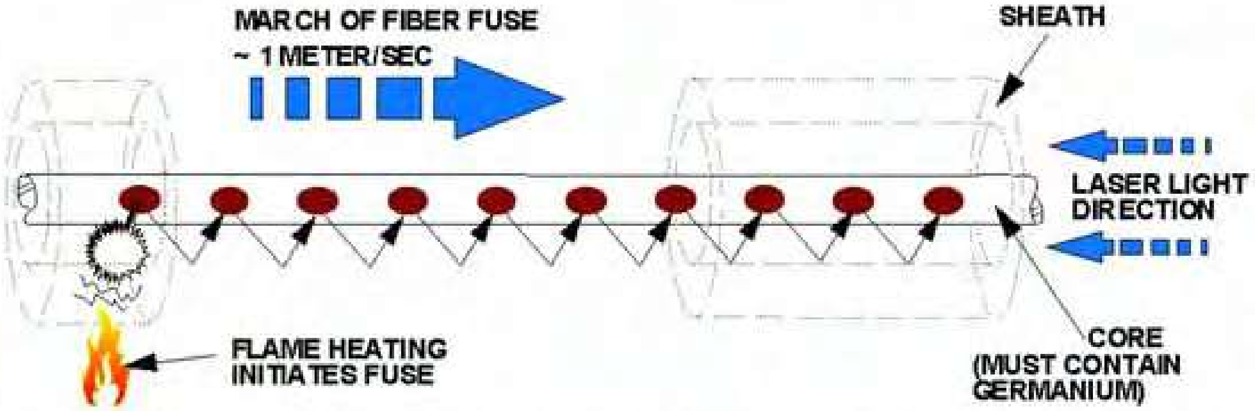
MARCH OF FIBER FUSE
~ 1 METER/SEC
SHEATH
LASER LIGHT DIRECTION
FLAME HEATING INITIATES FUSE
CORE (MUST CONTAIN GERMANIUM)

31A. Fiber fuse destroys core of fiber optics cable of indefinite length, pitting core with holes.

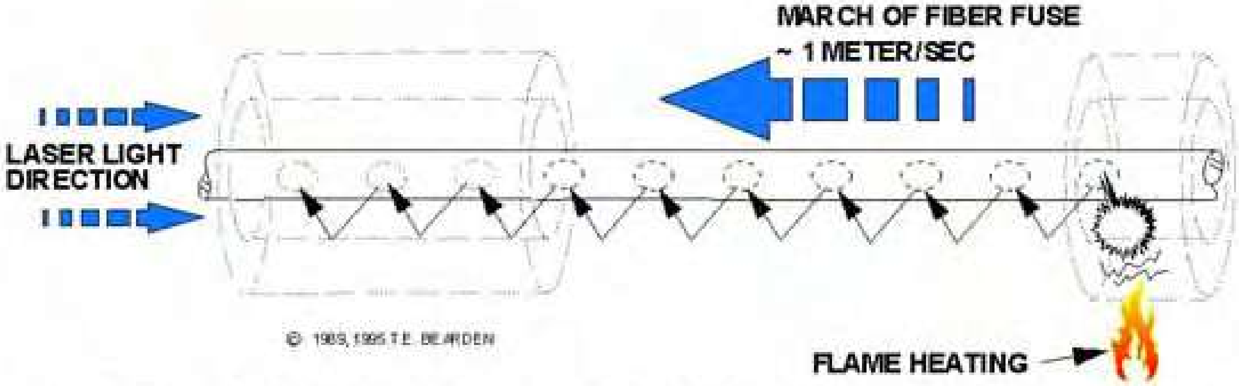
MARCH OF FIBER FUSE
~ 1 METER/SEC
LASER LIGHT DIRECTION
© 1989, 1995 T.E. BEARDEN
FLAME HEATING

31B. Reversed second fiber fuse often *restores* pitted core in cable, filling holes back up.

Figure 31. Fiber fuse effect and its strange anomalies.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

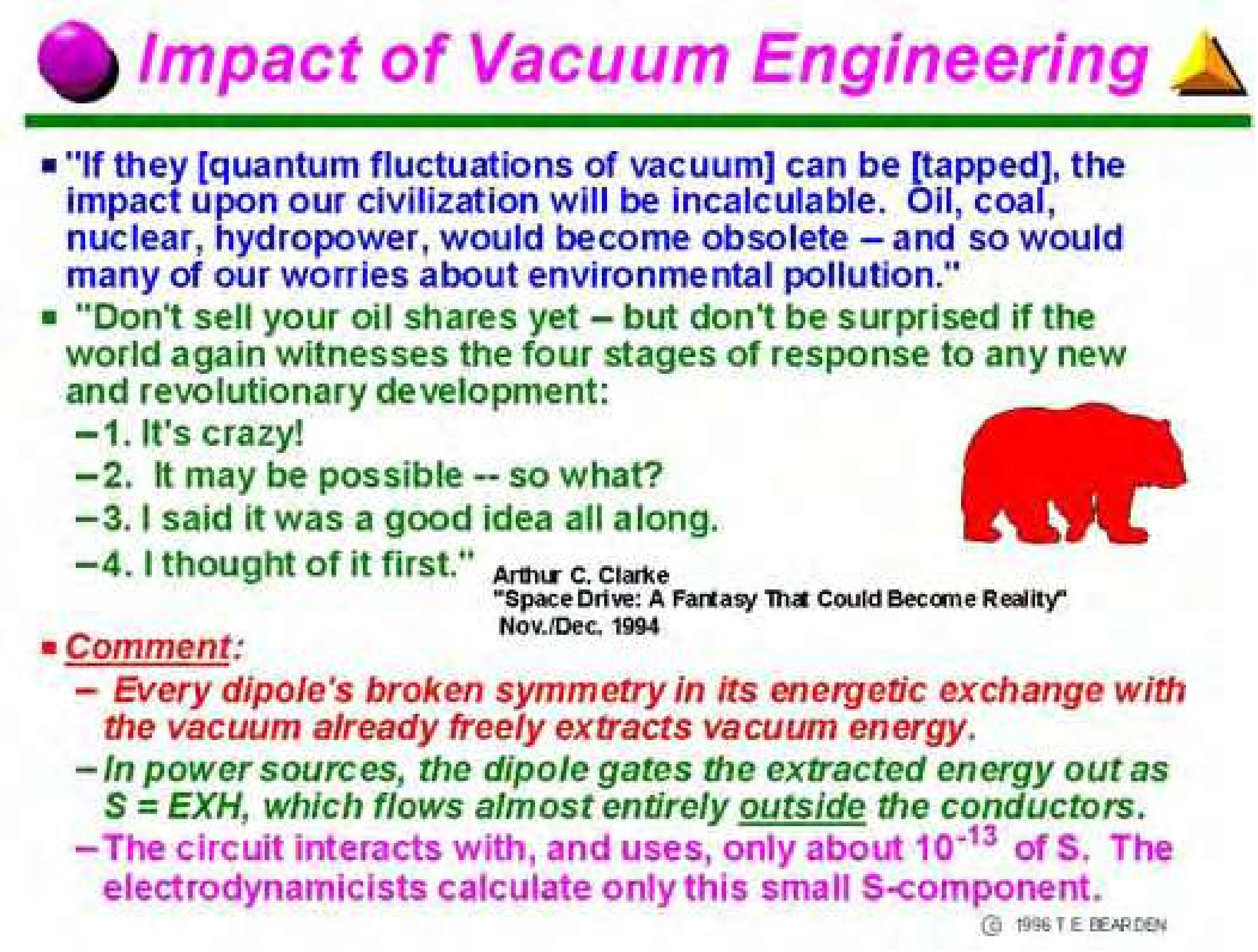

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete – and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet – but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

- *Comment:*
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

© 1996 T E BEARDEN

Return to Slide Index

Next Slide

Previous Slide

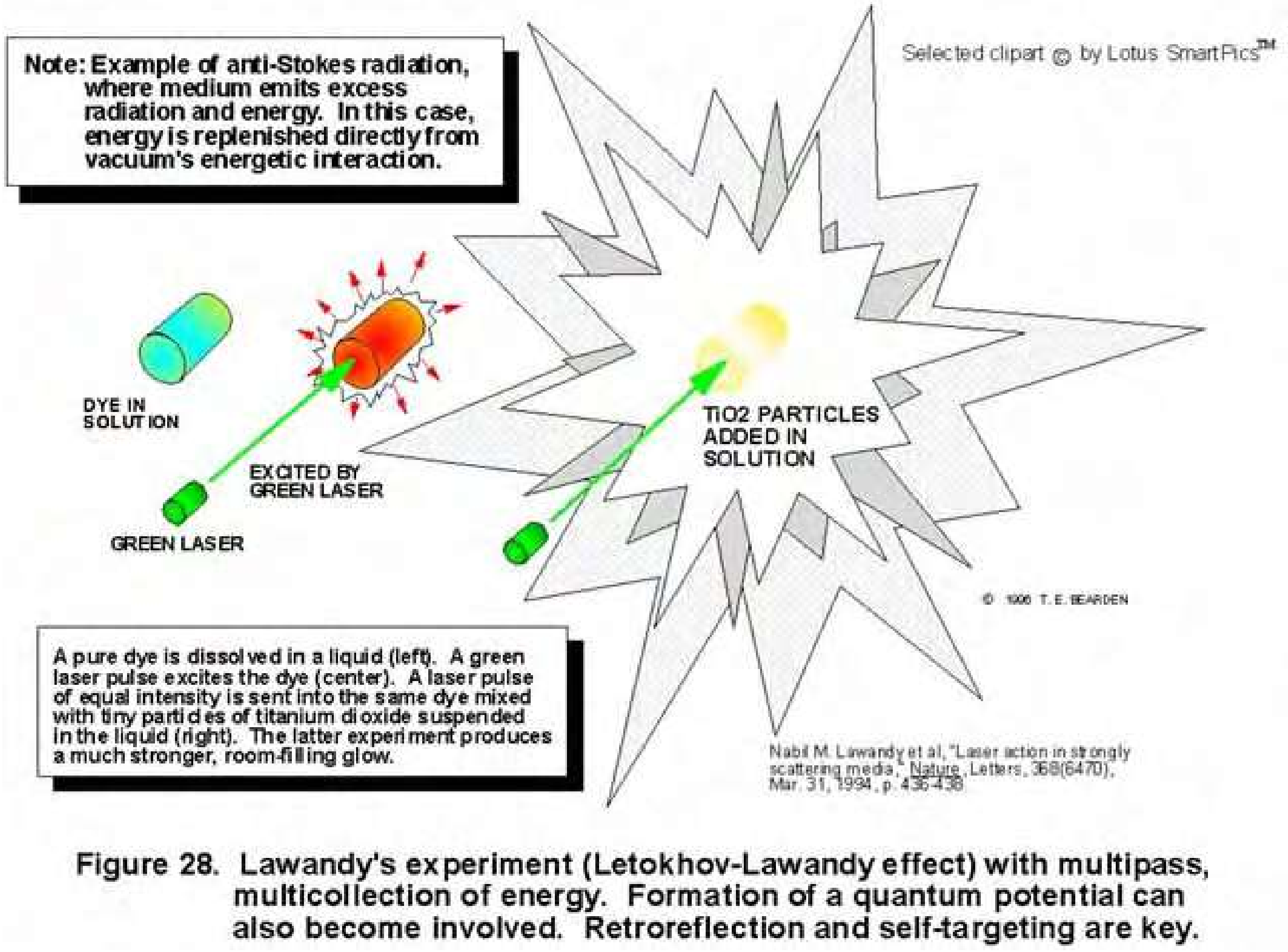
Selected clipart © by Lotus SmartPics™
DYE IN SOLUTION
EXCITED BY GREEN LASER
GREEN LASER
TiO2 PARTICLES ADDED IN SOLUTION
© 1996 T. E. BEARDEN

Note: Example of anti-Stokes radiation, where medium emits excess radiation and energy. In this case, energy is replenished directly from vacuum's energetic interaction.

A pure dye is dissolved in a liquid (left). A green laser pulse excites the dye (center). A laser pulse of equal intensity is sent into the same dye mixed with tiny particles of titanium dioxide suspended in the liquid (right). The latter experiment produces a much stronger, room-filling glow.

Nabil M. Lawandy et al, "Laser action in strongly scattering media," Nature, Letters, 368(6470), Mar. 31, 1994, p. 436-438.

Figure 28. Lawandy's experiment (Letokhov-Lawandy effect) with multipass, multicollection of energy. Formation of a quantum potential can also become involved. Retroreflection and self-targeting are key.

Return to Slide Index

Next Slide

Previous Slide

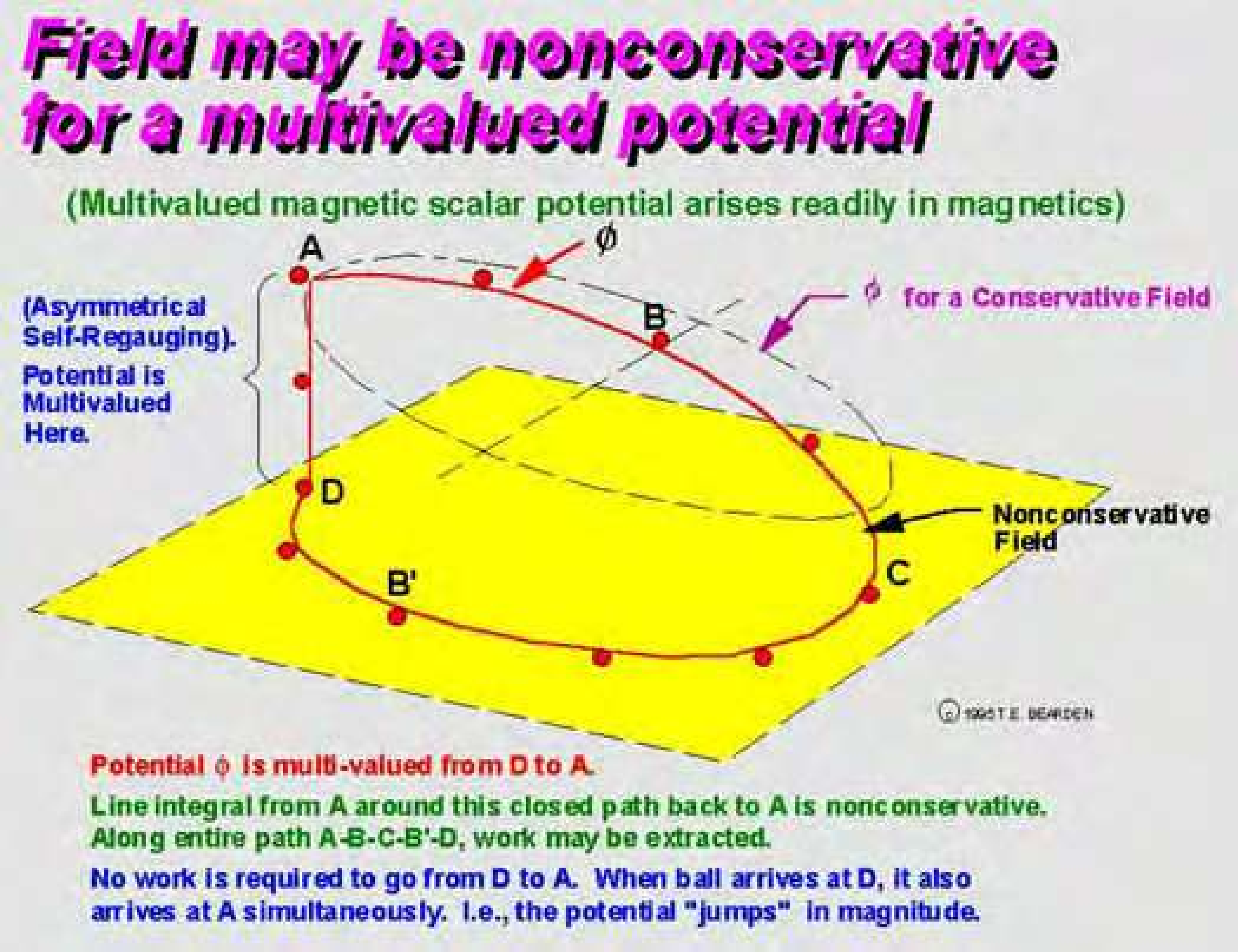

Field may be nonconservative for a multivalued potential

(Multivalued magnetic scalar potential arises readily in magnetics)

(Asymmetrical Self-Regauging). Potential is Multivalued Here.

ϕ for a Conservative Field

Nonconservative Field

© 1996 T.E. BEARDEN

Potential ϕ is multi-valued from D to A.

Line integral from A around this closed path back to A is nonconservative. Along entire path A-B-C-B'-D, work may be extracted.

No work is required to go from D to A. When ball arrives at D, it also arrives at A simultaneously. I.e., the potential "jumps" in magnitude.

Return to Slide Index

Next Slide

Previous Slide

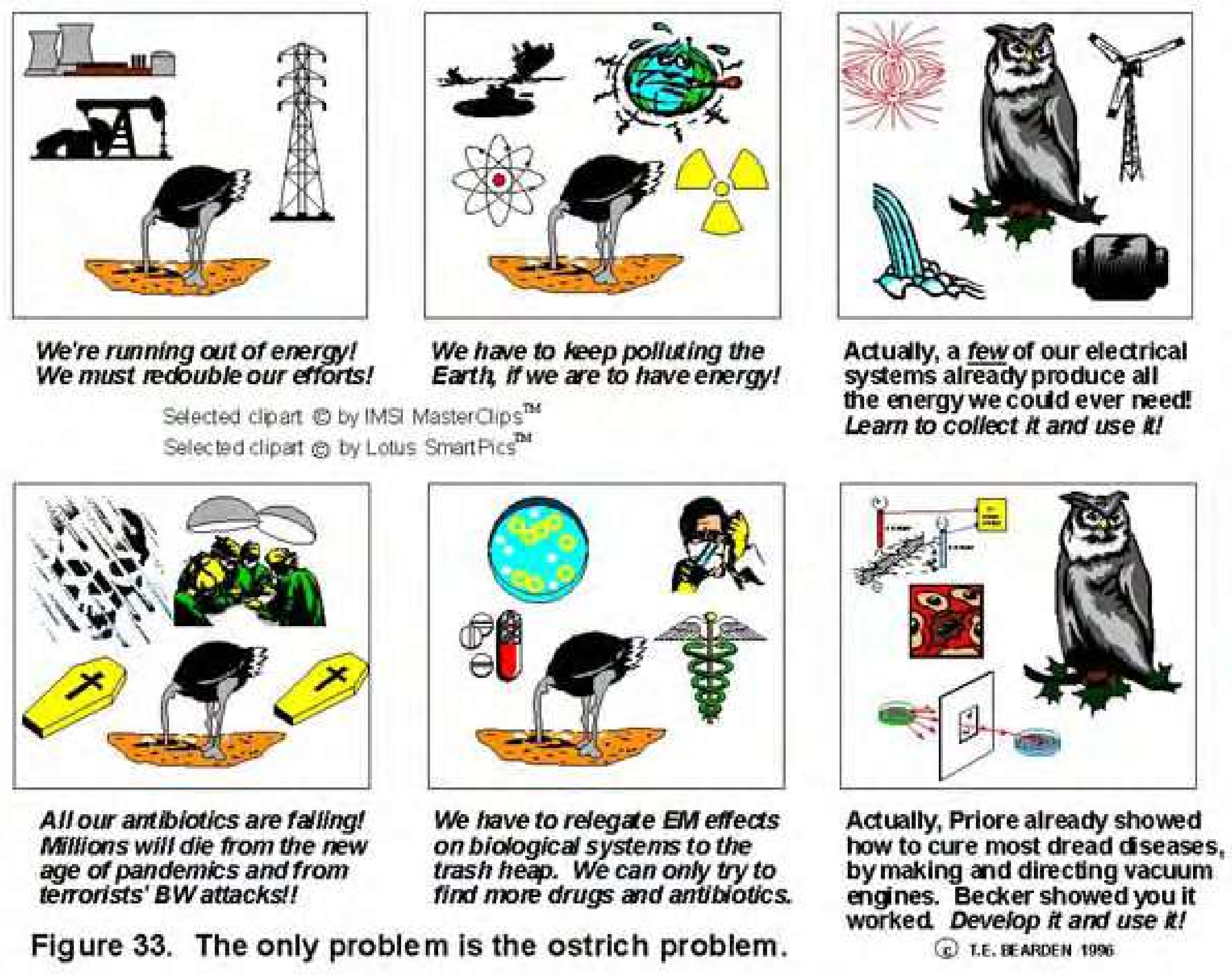

We're running out of energy! We must redouble our efforts!

We have to keep polluting the Earth, if we are to have energy!

Actually, a few of our electrical systems already produce all the energy we could ever need! *Learn to collect it and use it!*

Selected clipart © by IMSI MasterClips™
Selected clipart © by Lotus SmartPics™

All our antibiotics are failing! Millions will die from the new age of pandemics and from terrorists' BW attacks!!

We have to relegate EM effects on biological systems to the trash heap. We can only try to find more drugs and antibiotics.

Actually, Priore already showed how to cure most dread diseases, by making and directing vacuum engines. Becker showed you it worked. *Develop it and use it!*

Figure 33. The only problem is the ostrich problem.

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

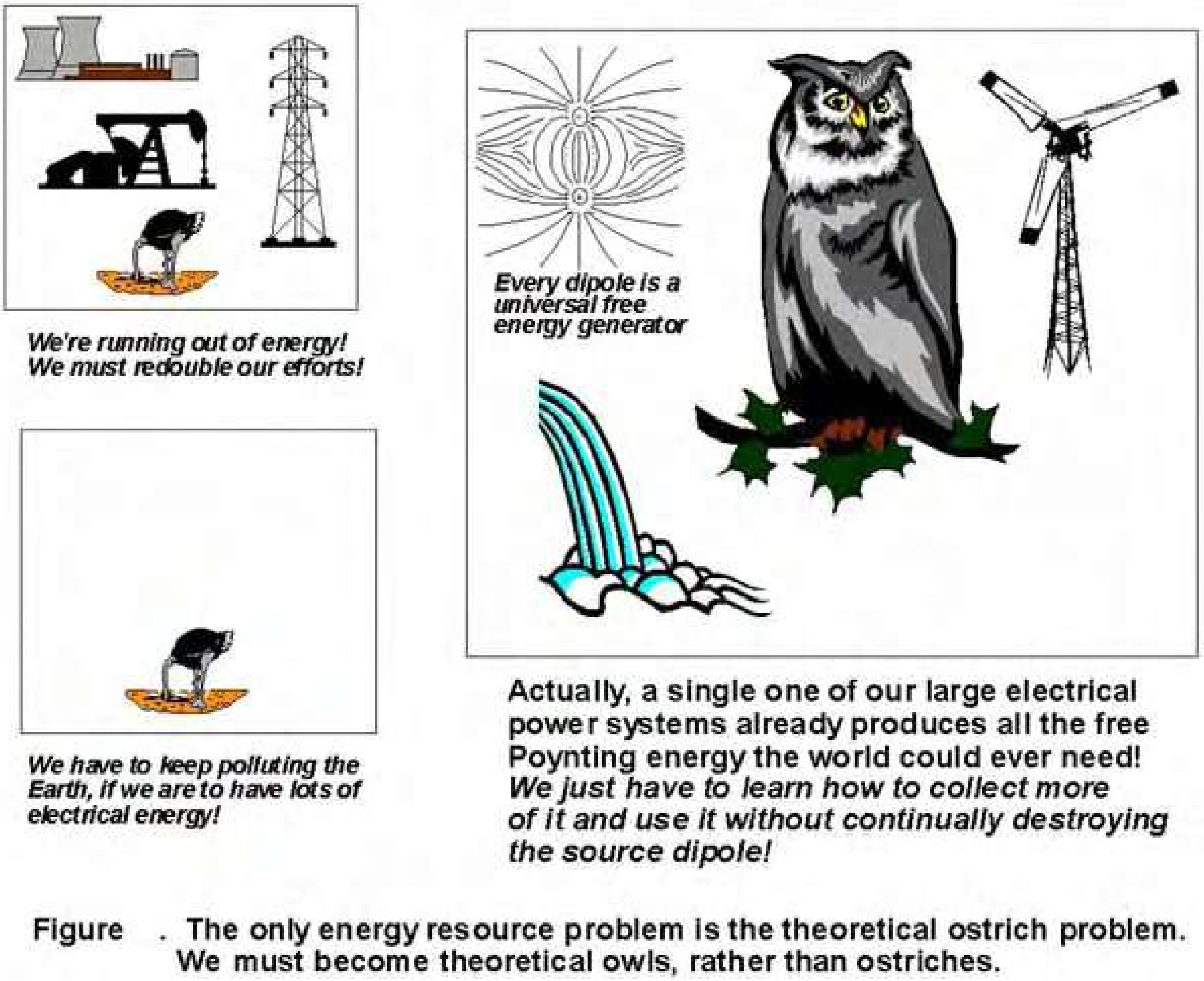

We're running out of energy!
We must redouble our efforts!

We have to keep polluting the Earth, if we are to have lots of electrical energy!

Every dipole is a universal free energy generator

Actually, a single one of our large electrical power systems already produces all the free Poynting energy the world could ever need! *We just have to learn how to collect more of it and use it without continually destroying the source dipole!*

Figure . The only energy resource problem is the theoretical ostrich problem. We must become theoretical owls, rather than ostriches.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

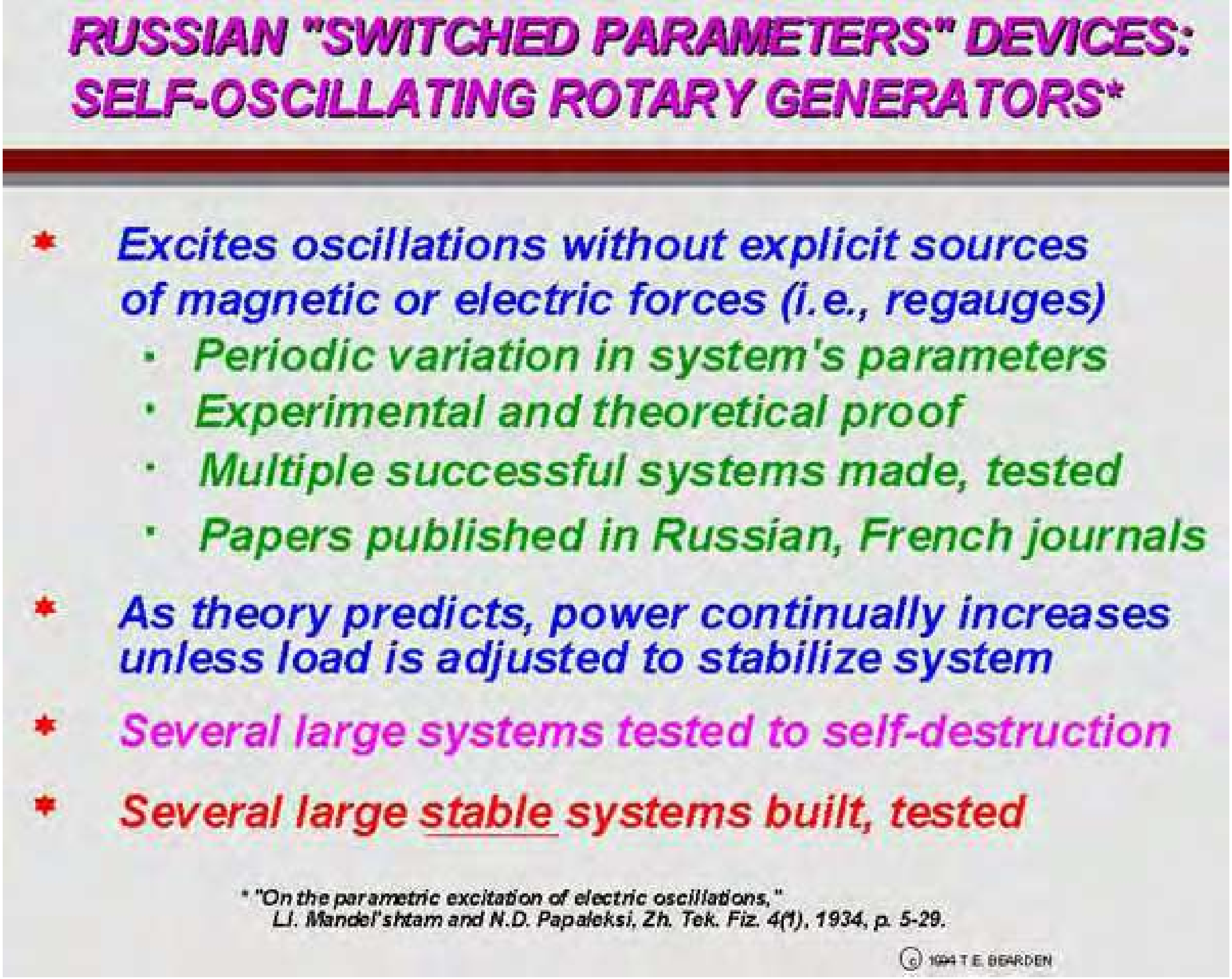

RUSSIAN "SWITCHED PARAMETERS" DEVICES: SELF-OSCILLATING ROTARY GENERATORS*

- Excites oscillations without explicit sources of magnetic or electric forces (i.e., regauges)
  - Periodic variation in system's parameters
  - Experimental and theoretical proof
  - Multiple successful systems made, tested
  - Papers published in Russian, French journals
- As theory predicts, power continually increases unless load is adjusted to stabilize system
- Several large systems tested to self-destruction
- Several large stable systems built, tested

* "On the parametric excitation of electric oscillations," L.I. Mandel'shtam and N.D. Papaleksi, Zh. Tek. Fiz. 4(1), 1934, p. 5-29.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

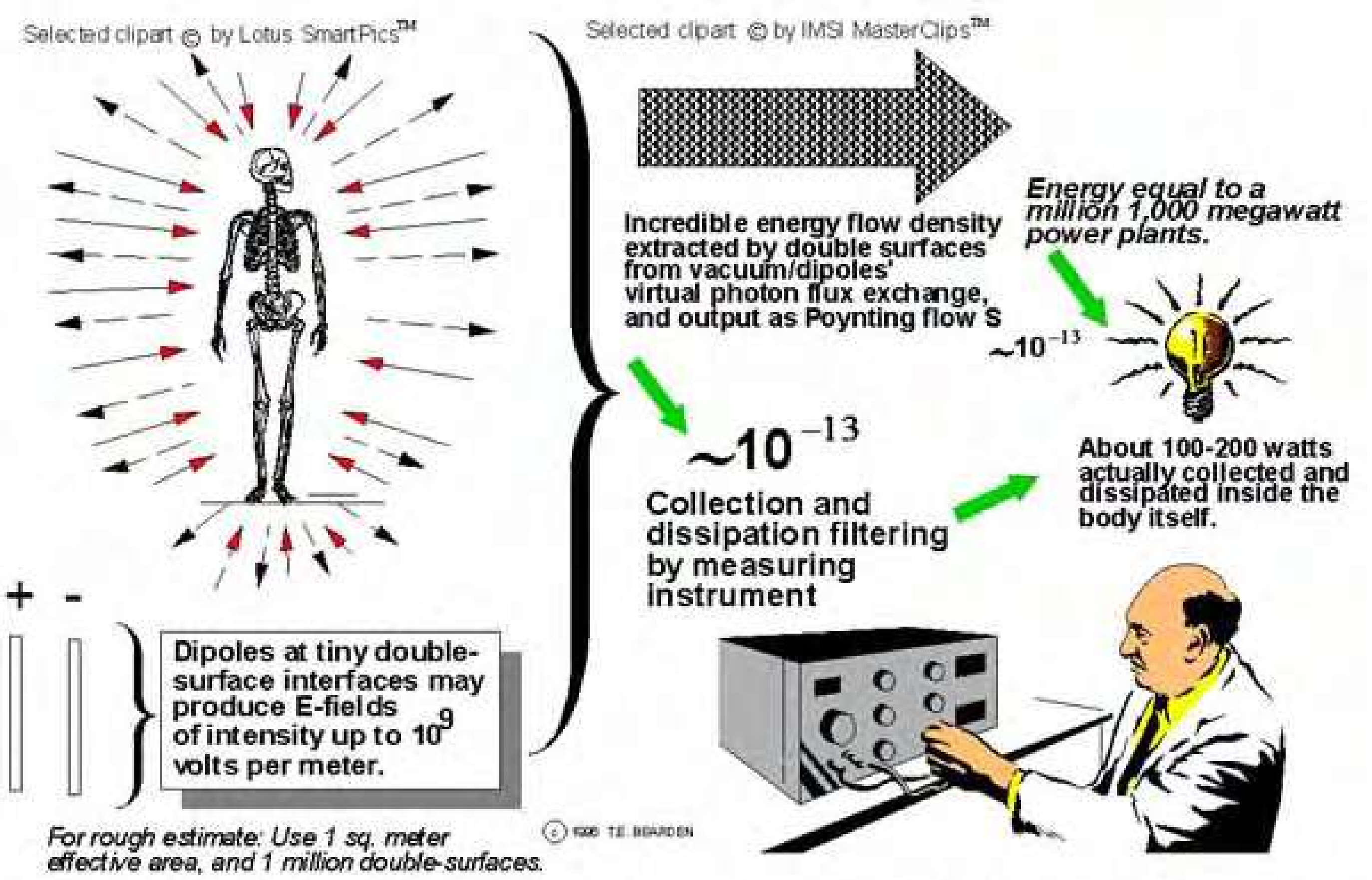

Selected clipart © by Lotus SmartPics™

Selected clipart © by IMSI MasterClips™

Incredible energy flow density extracted by double surfaces from vacuum/dipoles' virtual photon flux exchange, and output as Poynting flow S

Energy equal to a million 1,000 megawatt power plants.

About 100-200 watts actually collected and dissipated inside the body itself.

Collection and dissipation filtering by measuring instrument

Dipoles at tiny double-surface interfaces may produce E-fields of intensity up to 10^9 volts per meter.

For rough estimate: Use 1 sq. meter effective area, and 1 million double-surfaces.

© 1996 T.E. BEARDEN

Figure 20. One human body produces more energy flow than all the electrical power loads on earth dissipate. However, only 100-200 watts of its own energy output is collected and dissipated by the body.

Return to Slide Index

Next Slide

Previous Slide

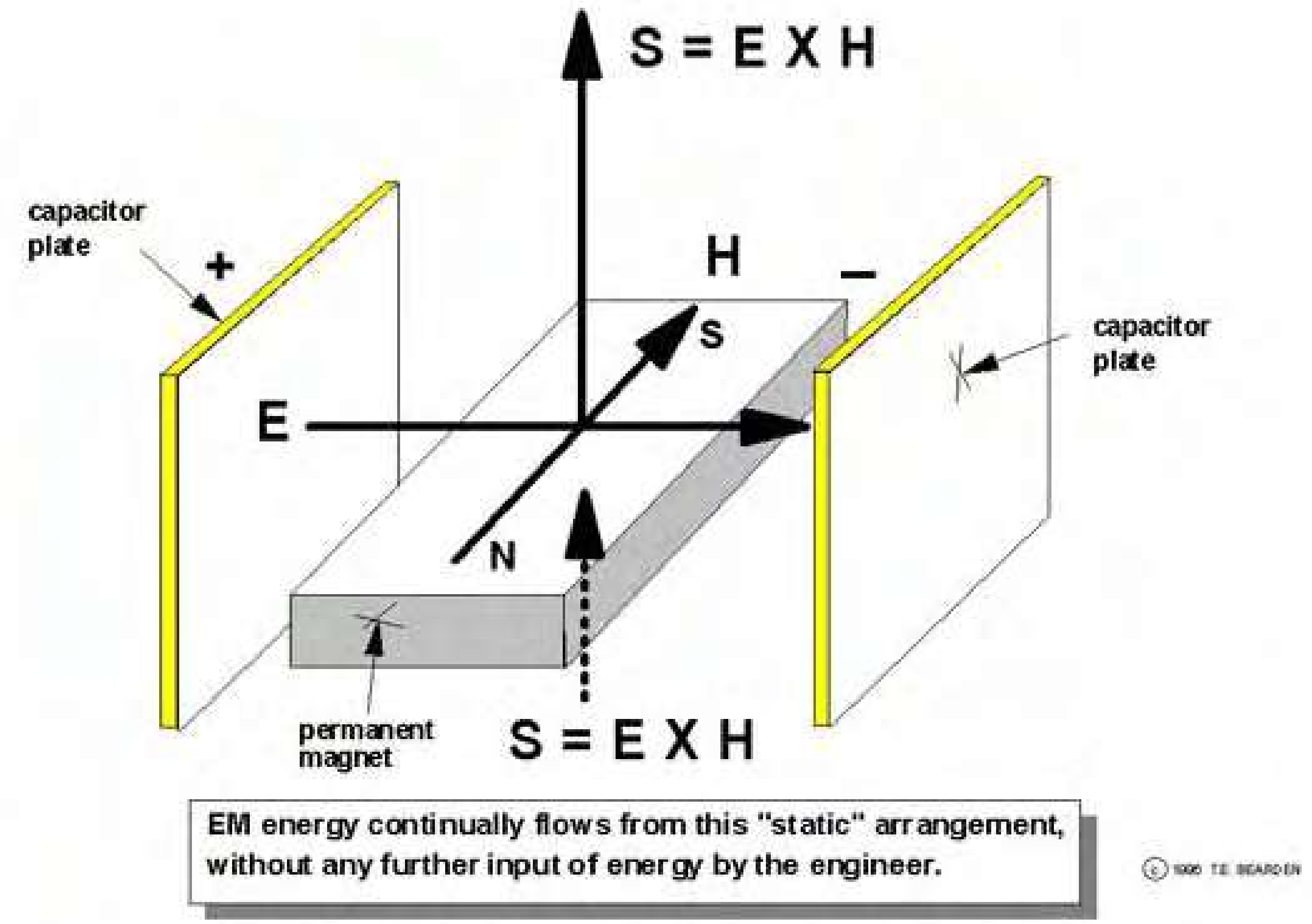
S = E X H
capacitor plate
+
H
–
S
capacitor plate
E
N
permanent magnet
S = E X H
© 1996 T.E. BEARDEN

EM energy continually flows from this "static" arrangement, without any further input of energy by the engineer.

Figure 18. A "static" Poynting S-flow free-energy generator.

Return to Slide Index

Next Slide

Previous Slide

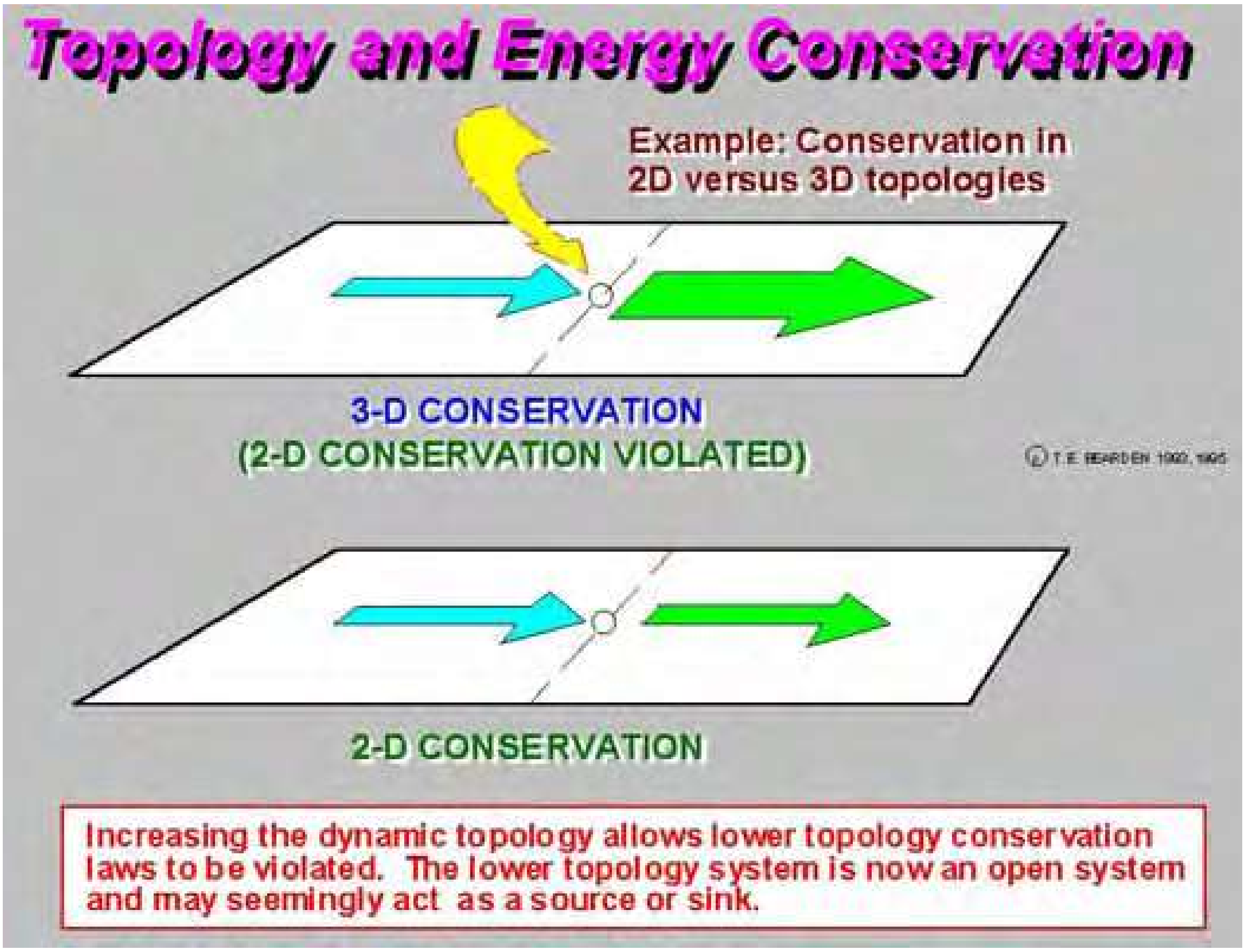

Topology and Energy Conservation

Example: Conservation in 2D versus 3D topologies

3-D CONSERVATION
(2-D CONSERVATION VIOLATED)

© T.E. BEARDEN 1992, 1995

2-D CONSERVATION

Increasing the dynamic topology allows lower topology conservation laws to be violated. The lower topology system is now an open system and may seemingly act as a source or sink.

Return to Slide Index

Next Slide

Previous Slide

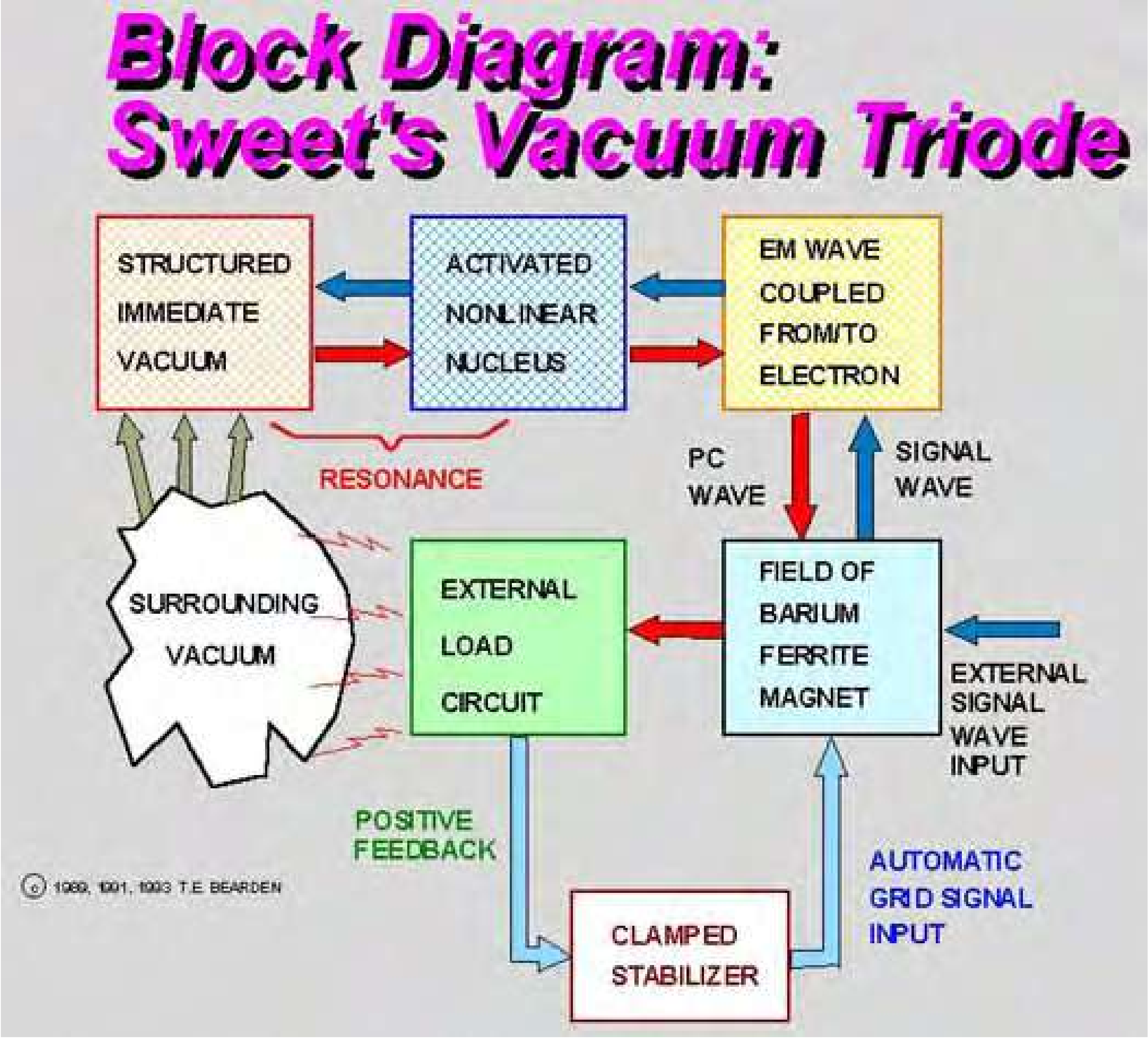

Block Diagram: Sweet's Vacuum Triode

STRUCTURED IMMEDIATE VACUUM

ACTIVATED NONLINEAR NUCLEUS

EM WAVE COUPLED FROM/TO ELECTRON

RESONANCE

PC WAVE

SIGNAL WAVE

SURROUNDING VACUUM

EXTERNAL LOAD CIRCUIT

FIELD OF BARIUM FERRITE MAGNET

EXTERNAL SIGNAL WAVE INPUT

POSITIVE FEEDBACK

AUTOMATIC GRID SIGNAL INPUT

CLAMPED STABILIZER

© 1989, 1991, 1993 T.E. BEARDEN

[Return to Slide Index]

[Previous Slide]

Defense Department Briefing on Priore Treatment to counter Biological Warfare

---

SLIDE LIST

Solutions to Mass Casualties Treatment Problem

Threat and Current Status

1. Title Page
2. Key Points in the New Technology
3. Title Page
4. Title Page
5. The Terrorist Threat
6. The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
7. Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.
8. Present Status of the WMD Threat to the United States
9. Some Present Problems in the WMD Defense Capabilities

Historical Solution: Priore Treatment Therapy

10. Title Page
11. Priore's laboratory device for treating small animals
12. Antoine Priore with the Mayor of Bordeaux
13. Raymond Pautrizel, eminent French parasitologist, who worked with Priore
14. 12-Foot plasma tube in Priore's large device
15. Priore's 3 1/2 stories high device for whole body treatment of humans
16. Mechanism of the Priore therapy
17. Priore's therapeutic methodology: using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain
18. Drastically reduced cumulative deterioration after Priore therapy

Extended Physics and Background

19. Title Page
20. Founders of Scalar Electromagnetics
21. Subdivisions of Soviet Energetics

22. Bunge on the Status of Classical and Quantum Physics
23. Einstein on Reviewing Foundations
24. Fields and Potentials have been "defined" only in and on matter
25. Maxwell's Equations are Matter-To-Matter Transforms
26. Electrodynamicists' Reaction to Removal of the Material Ether
27. The Missing Infolded Electrodynamics
28. Nonlinear Optics Distortion Correction Theorem
29. Infolded longitudinal EM biwave composition of a scalar potential
30. The Zero-Vector Axiom Destroys "Topologies Within Topologies"
31. Stoney-Whittaker Bidirectional Longitudinal EM Wave-Pair
32. Undistorted Progressive Waves (UPWs)
33. Velocity modulating a longitudinal EM wave oscillates rate of flow of time
34. The Graviton is a Coupled Photon/Antiphoton Pair
35. Mechanism for Physical Change and the Flow of Time
36. Any EM Field or Wave Pattern is Produced by interference of Two Scalar Potential Functions (E.T. Whittaker, 1904)
37. Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass
38. Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection
39. The Immune System
40. Popp's master cellular control system
41. Cancer: characteristics
42. The evolutionary path of multicellular organisms
43. Effect of fluid contamination on hemoglobin's oxygen transport
44. First step in Self-Promotion of Cancer due to Hypoxia
45. Becker's Cellular Growth Control
46. Becker's dc treatment of bone fractures
47. Becker's theoretical DC control system involved with responses to injury
48. Becker's theoretical control system governing regeneration
49. Link Between Standing DC Potentials and the Nervous System
50. Becker's Findings on Cellular Injury and Regeneration
51. Corrections For the Present EM Bioeffects Model
52. R&R System - Regeneration Example
53. Kaznacheyev's Cytopathogenic Mirror Effect
54. Microwave Radiation of U.S. Embassy in Moscow

55. Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
56. Quantum potential and hidden variable theory
57. Participants in a quantum potential share a common multiply-connected spacetime (MCST)
58. Gulf War Syndrome and Typical Comparative Disease Curves
59. Force Fields and Symmetry
60. Aspects of Strong Local Asymmetry
61. Other Aspects of Strong Local Asymmetry
62. Small Entities Emit Powerful S-Flows
63. What is the "magnitude" of a potential?
64. Table A1. The Vacuum Energetically Interacts With Every Particle, Continually
65. Extended Superpotential Theory (Adding Vacuum Engines)
66. Vacuum Engine: Jillions of Working ST Demons
67. Active Spacetime and a Specific Vacuum Engine (A General Relativistic Concept and Analogy)
68. Extension of Wheeler's general relativity principle
69. Signal versus Vacuum Engine
70. Electronuclear Interactions by Vacuum Engines
71. Vacuum Engineering
72. Impact of Vacuum Engineering
73. A Dramatic Extension to Nonlinear Optical Pumping
74. Time Conservation and Some Properties (Energy and time are canonical)
75. A Dramatic Extension to Nonlinear Optical Pumping
76. Key Points in the New Technology

Modern WMD Solution: Portable Treatment Systems

77. Title Page
78. Portable Priore-Type Treatment System
79. Portable System Block Diagram
80. Portable Priore Technology Unit
81. Saving the Stricken Civilians

Conclusions: Technology Practical, Proven

82. Title Page
83. Summary and Conclusions

Recommendation: Crash Development Program

84. Title Page
85. Recommend implementing:
86. Final Thoughts

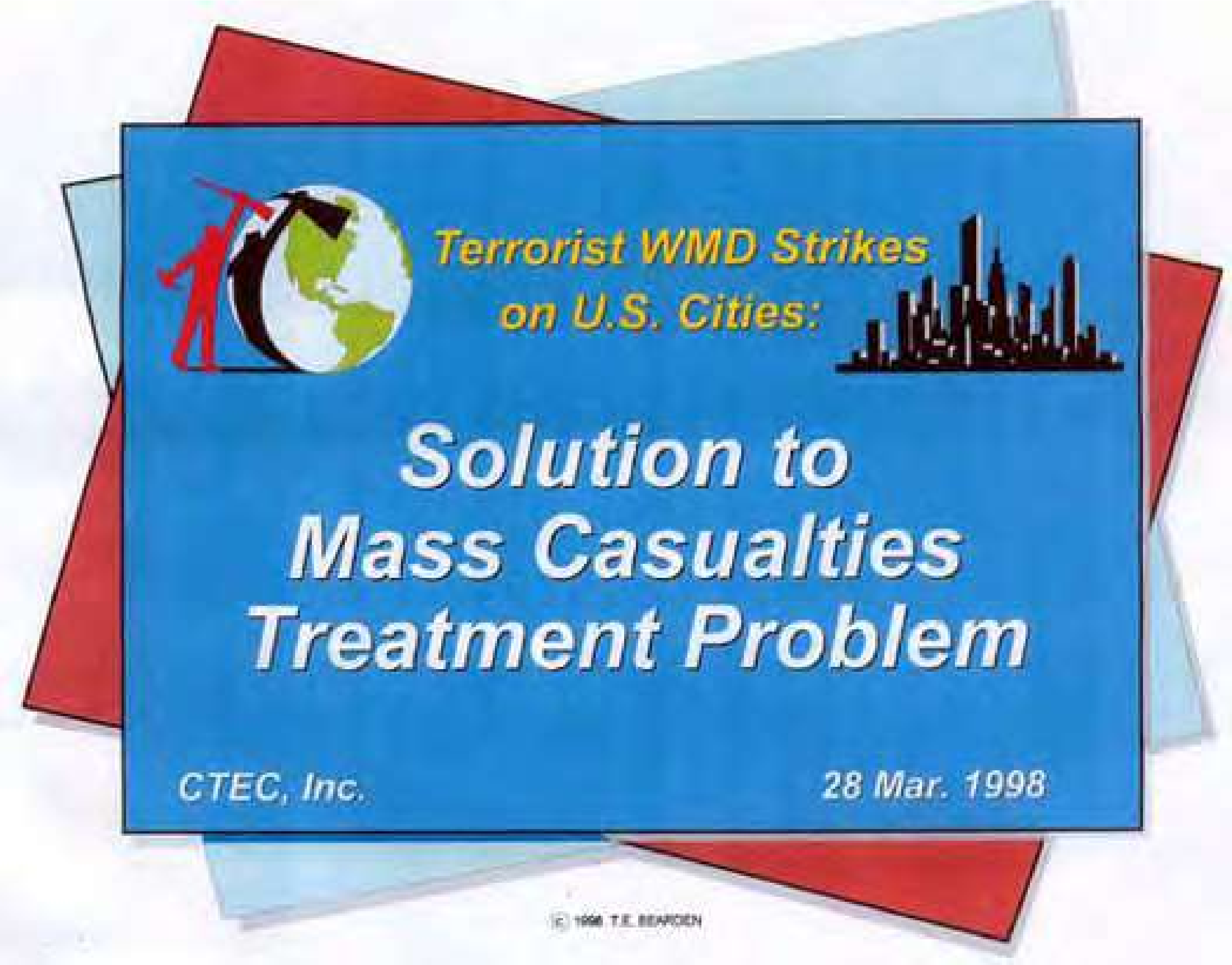

Terrorist WMD Strikes
on U.S. Cities:

Solution to
Mass Casualties
Treatment Problem

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

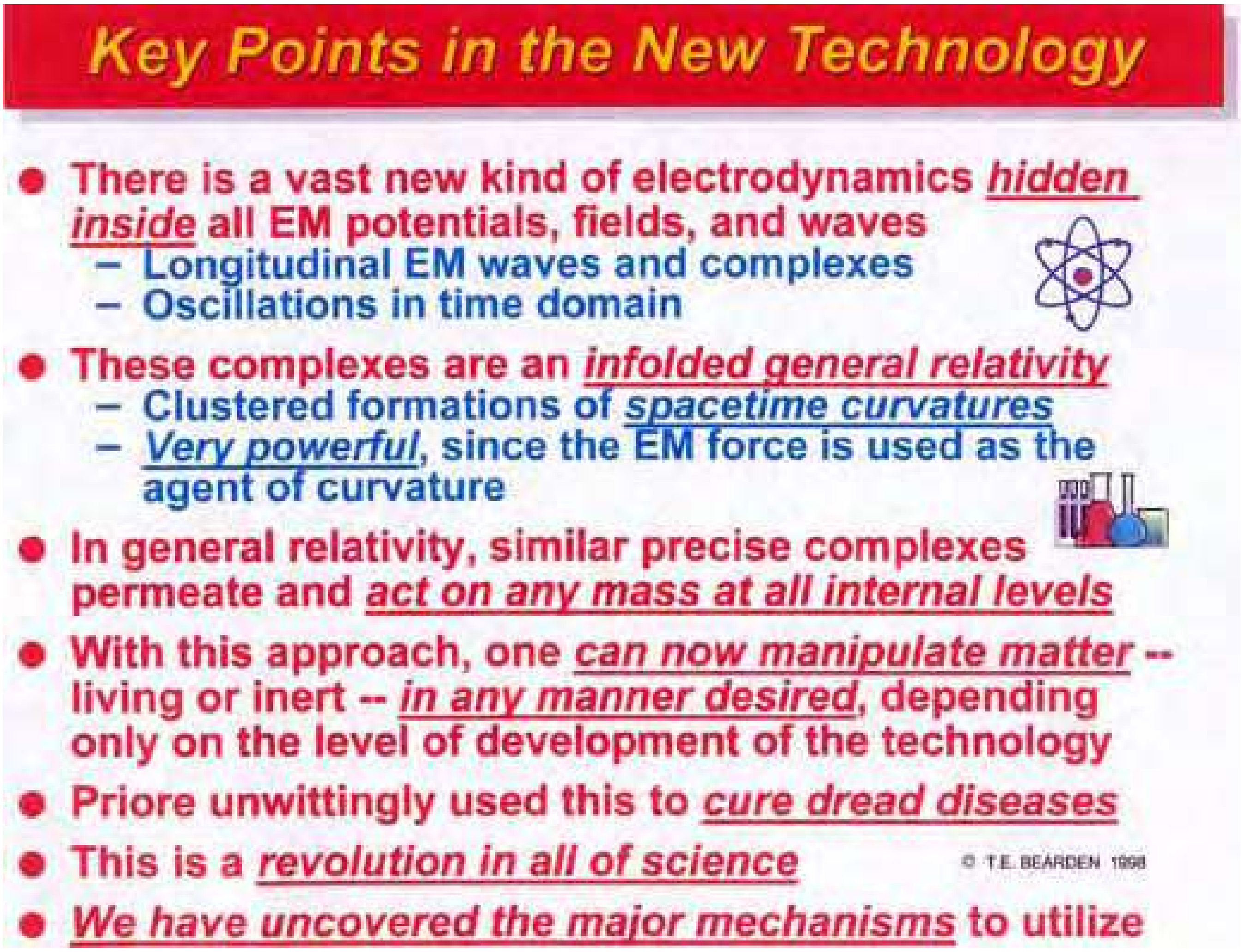

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

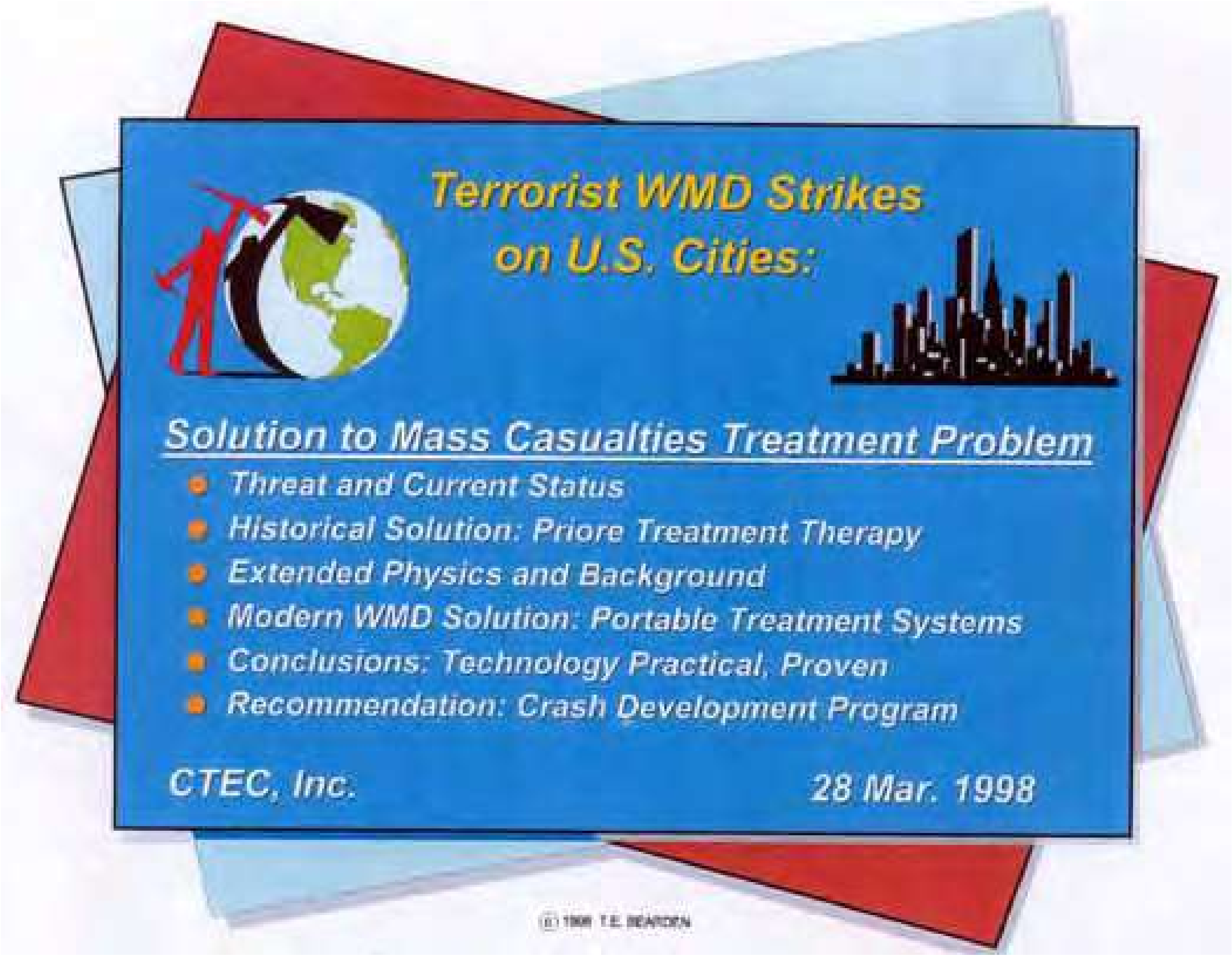

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

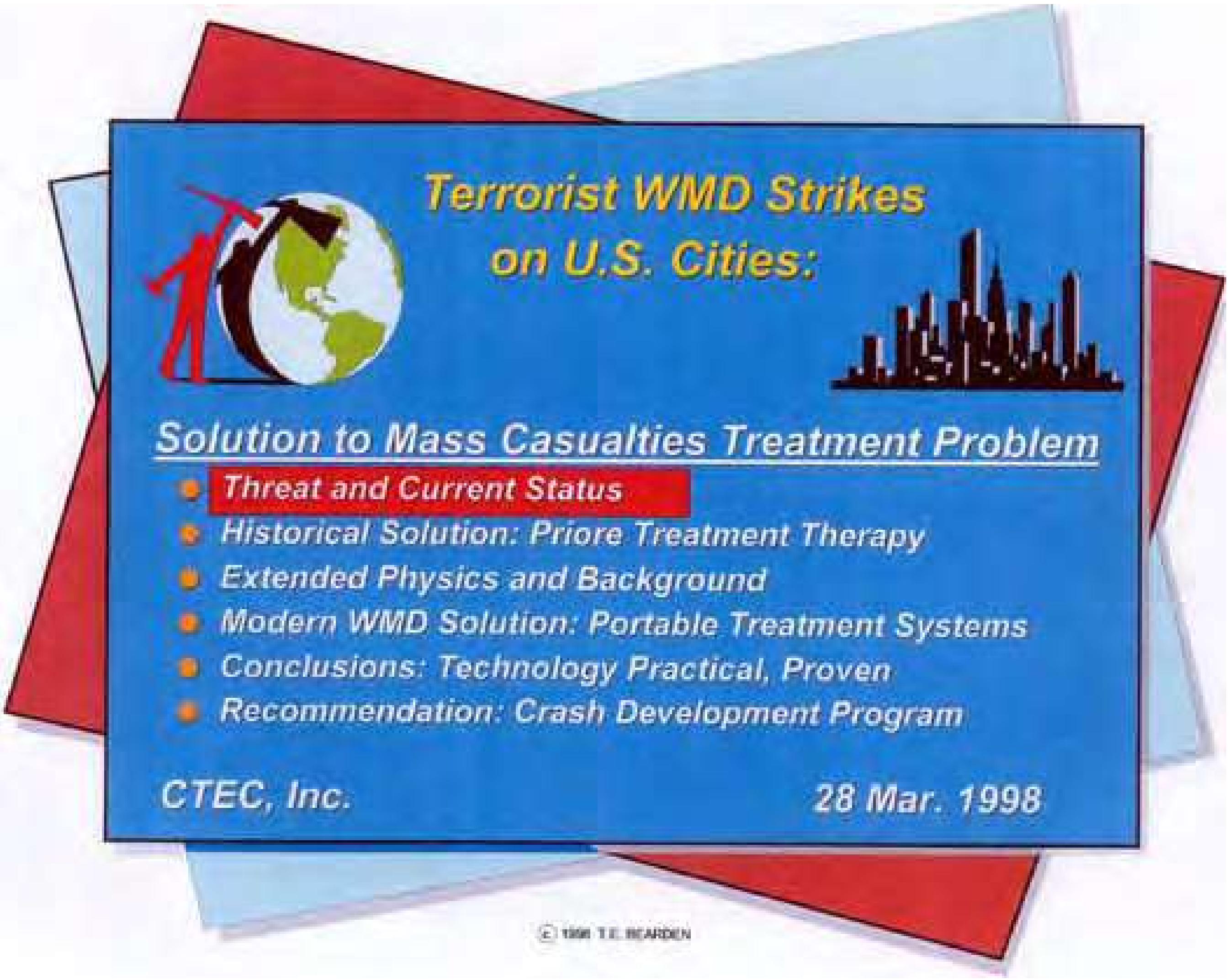

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

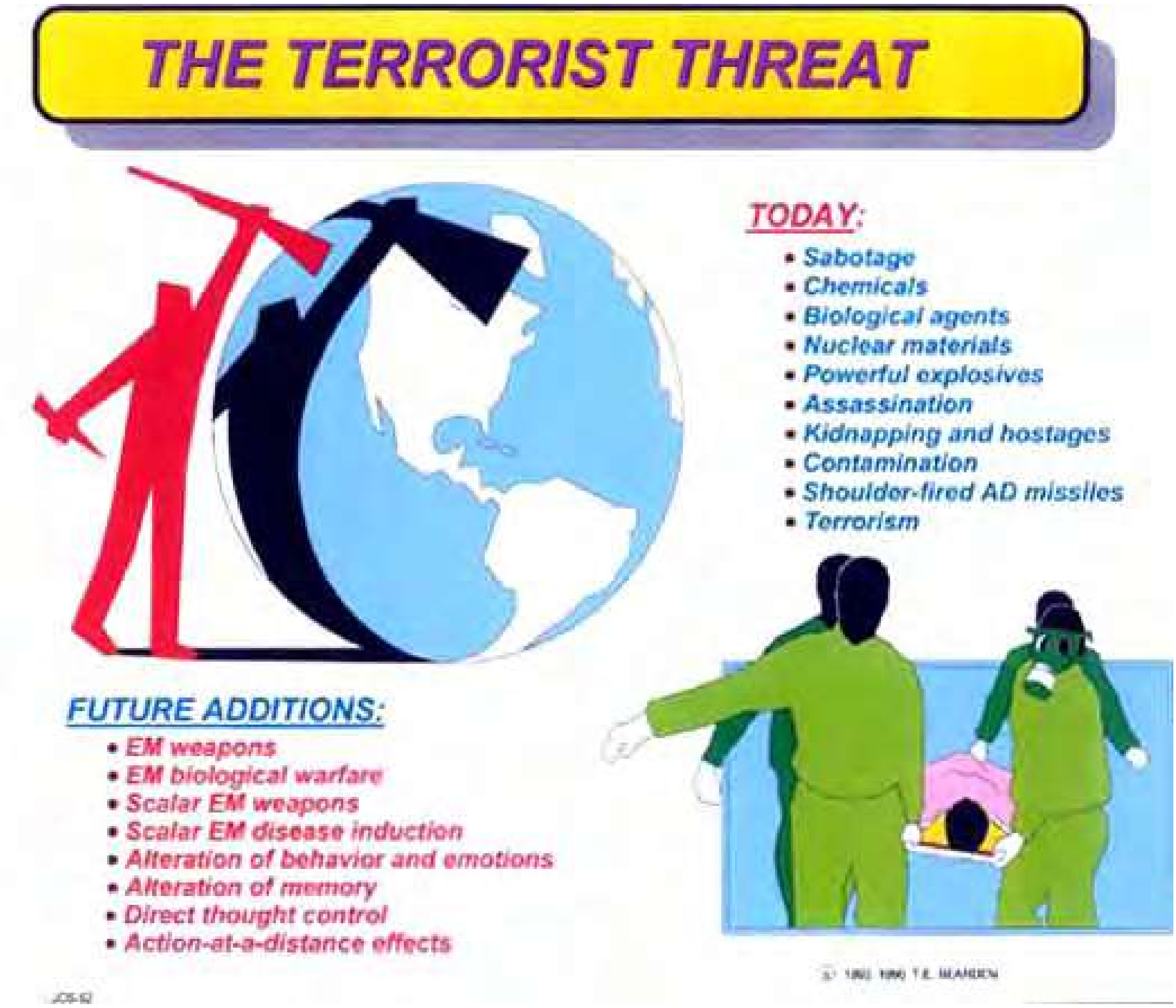

THE TERRORIST THREAT

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- EM weapons
- EM biological warfare
- Scalar EM weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct thought control
- Action-at-a-distance effects

© 1982, 1990 T.E. BEARDEN

JCS-62

Return to Slide Index

Next Slide

Previous Slide

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

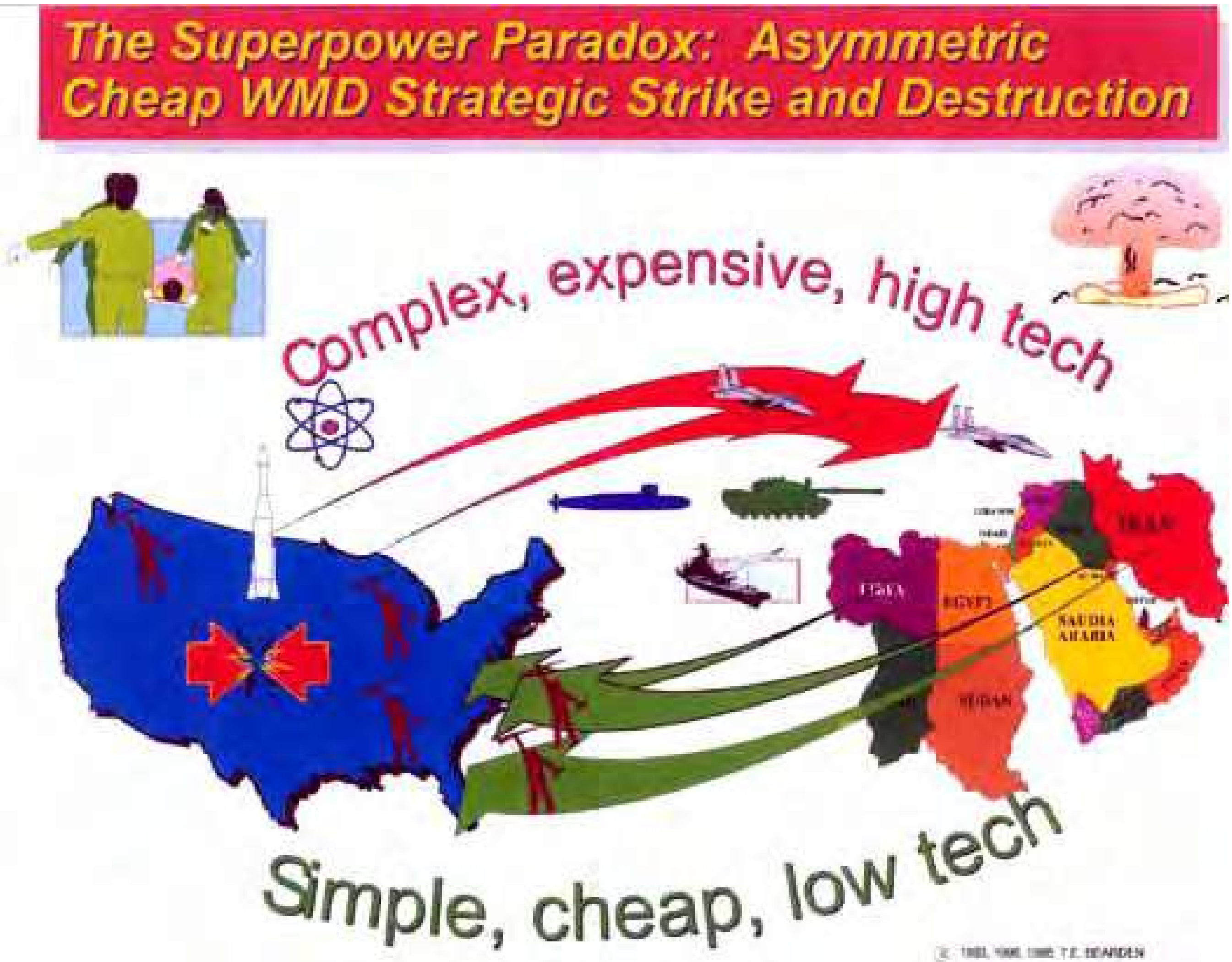
LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN

Complex, expensive, high tech

Simple, cheap, low tech

© 1993, 1996, 1998 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

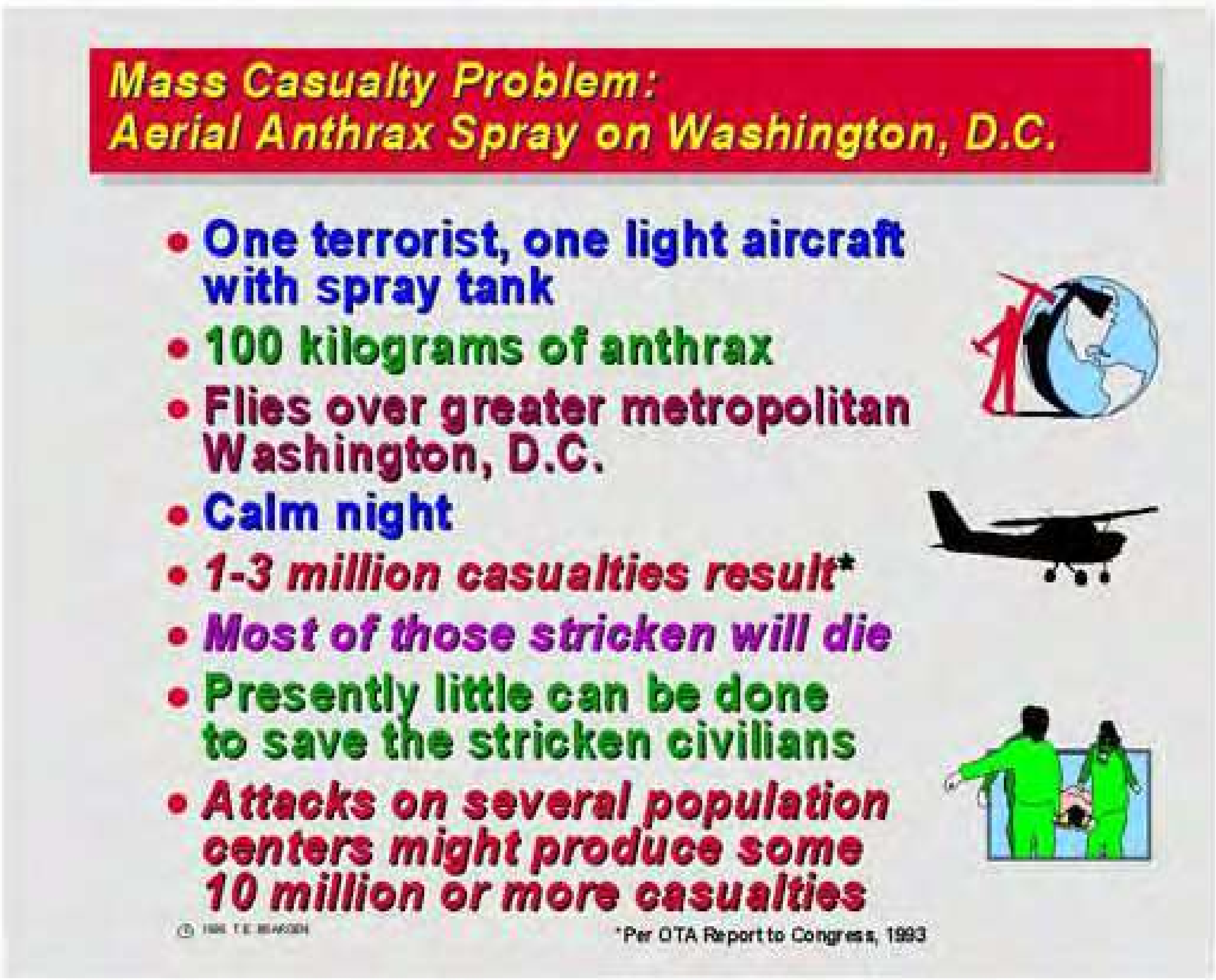

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- 100 kilograms of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

*Per OTA Report to Congress, 1993

[Return to Slide Index]

[Next Slide]

[Previous Slide]

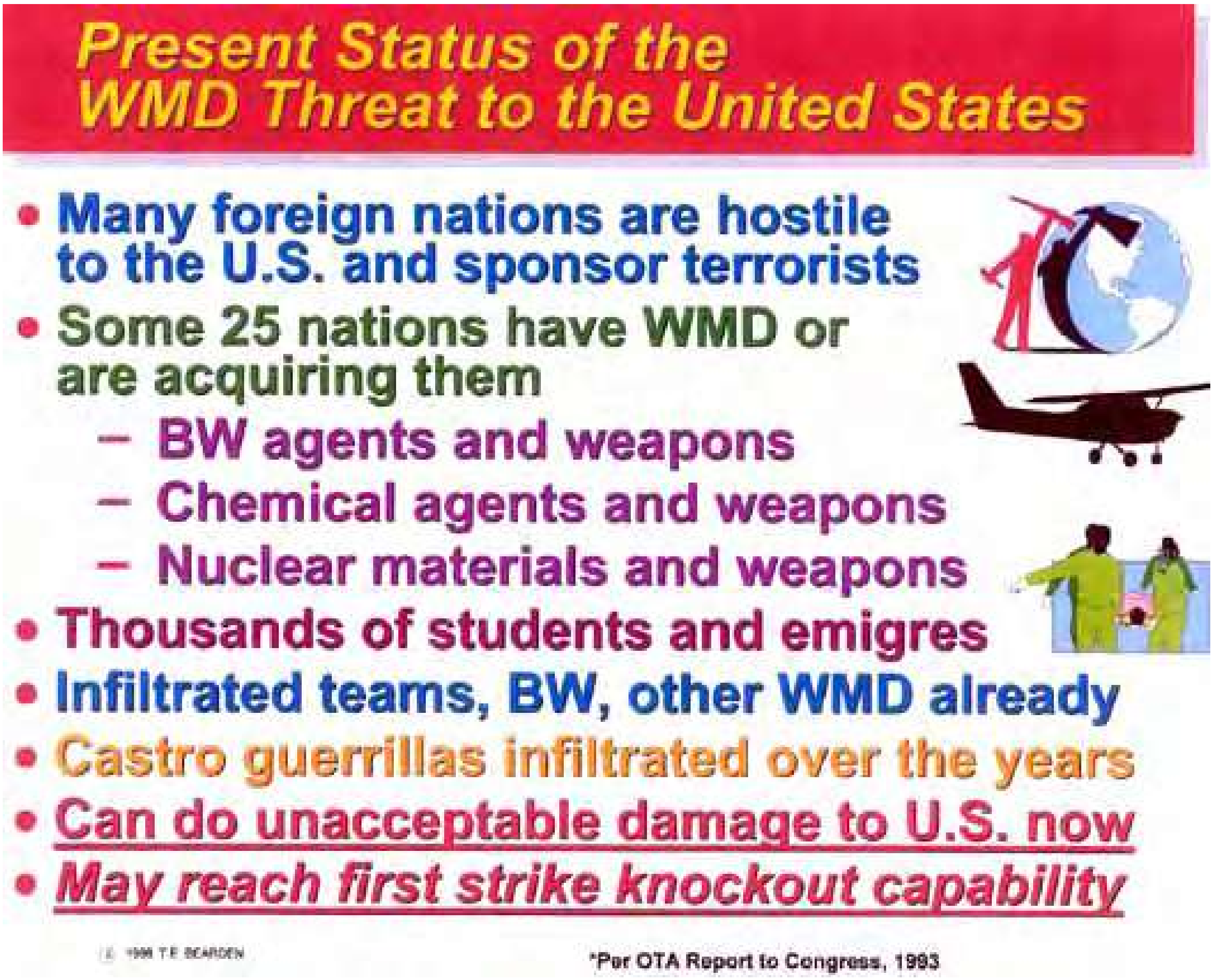

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
  - BW agents and weapons
  - Chemical agents and weapons
  - Nuclear materials and weapons
- Thousands of students and emigres
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© 1998 T.E. BEARDEN

*Per OTA Report to Congress, 1993

Return to Slide Index

Next Slide

Previous Slide

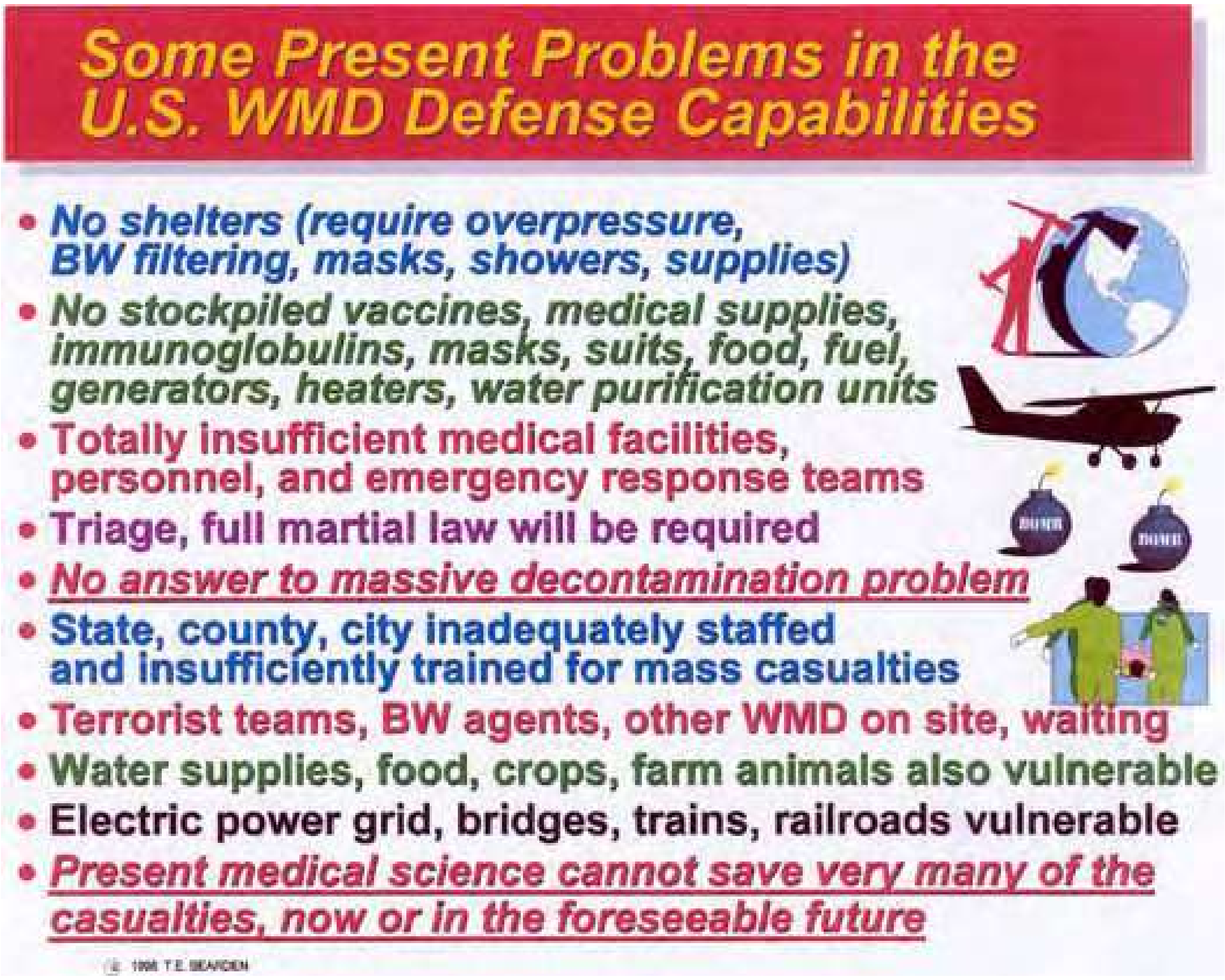

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

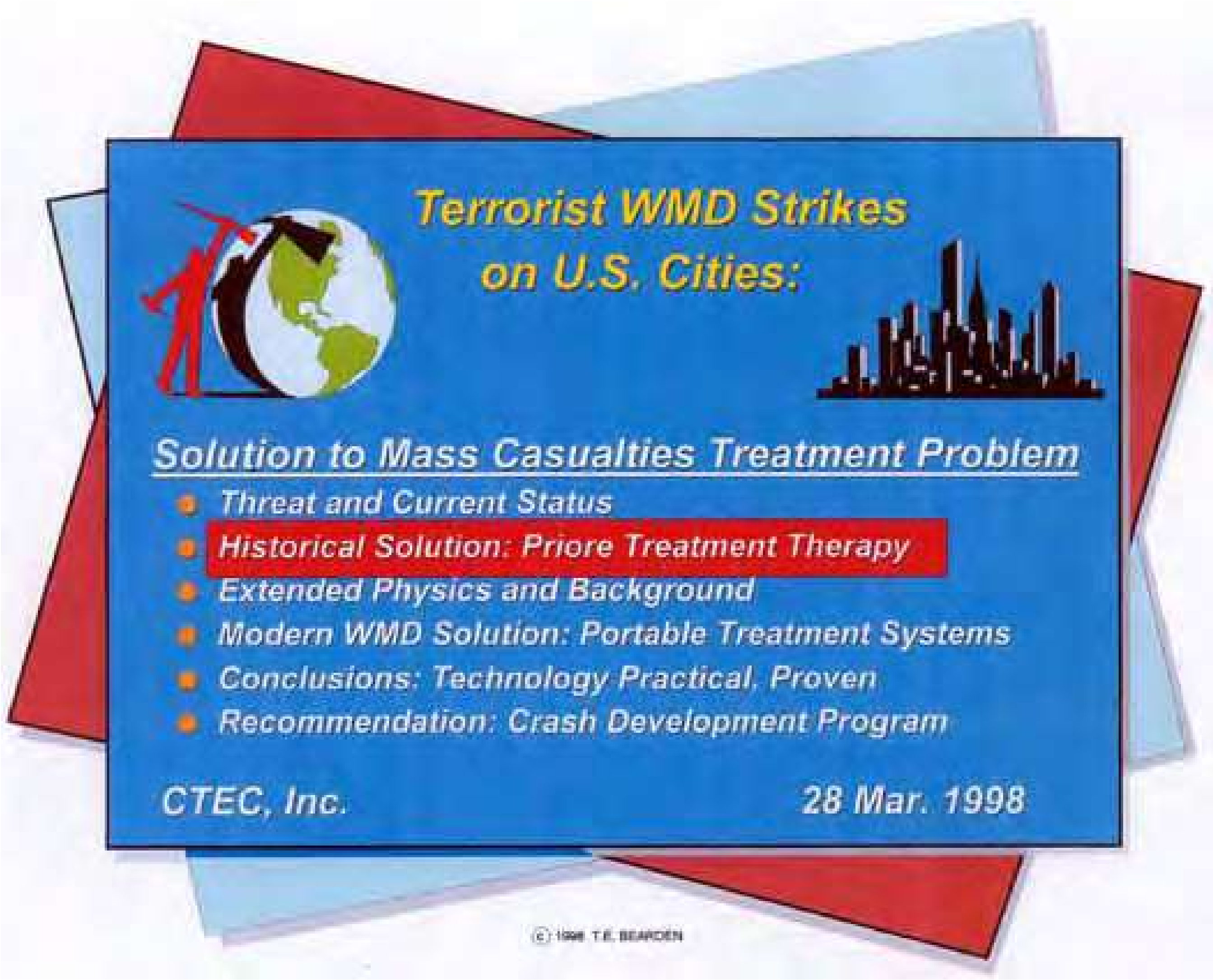

Terrorist WMD Strikes
on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

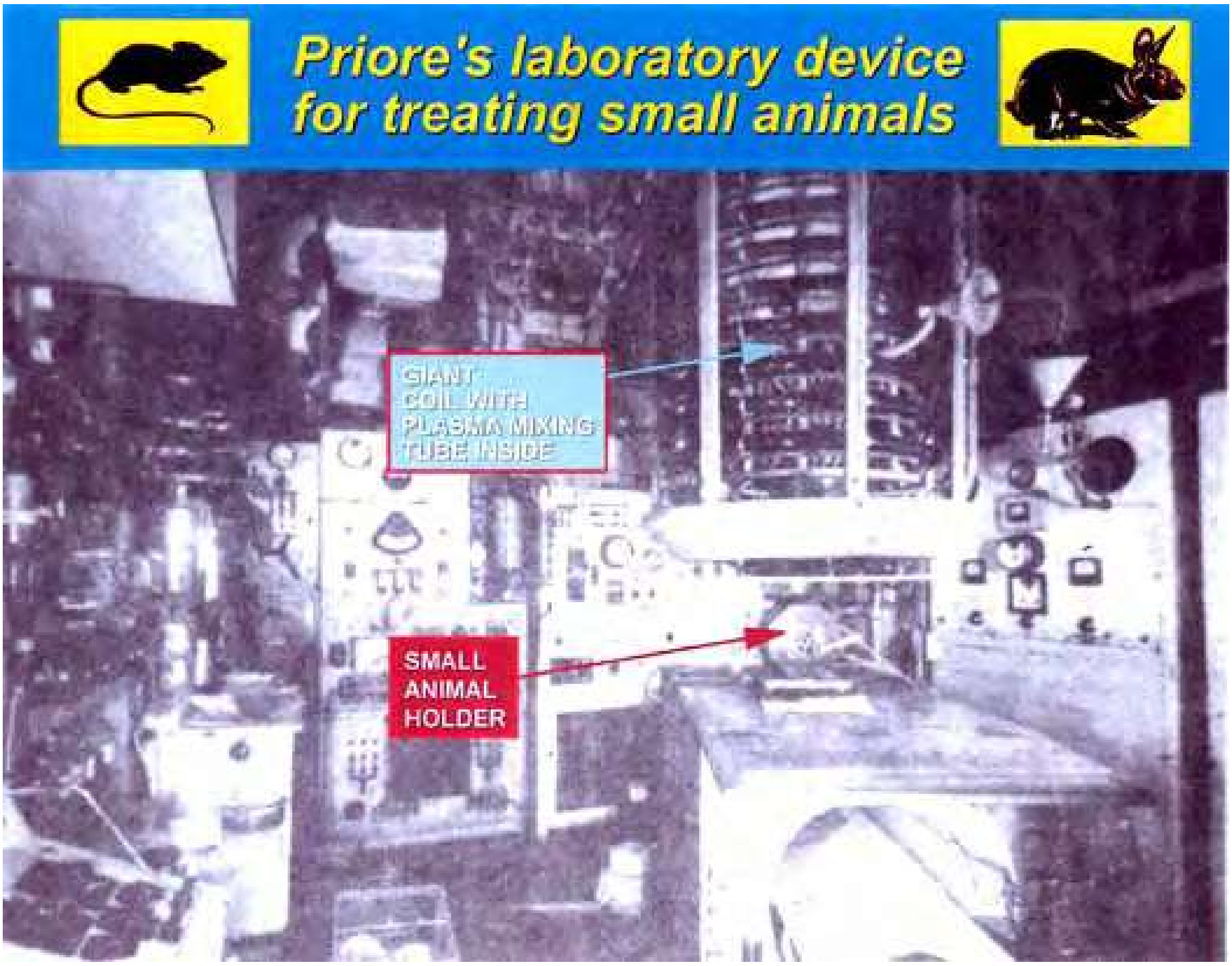

Priore's laboratory device for treating small animals

GIANT COIL WITH PLASMA MIXING TUBE INSIDE
SMALL ANIMAL HOLDER

Return to Slide Index

Next Slide

Previous Slide

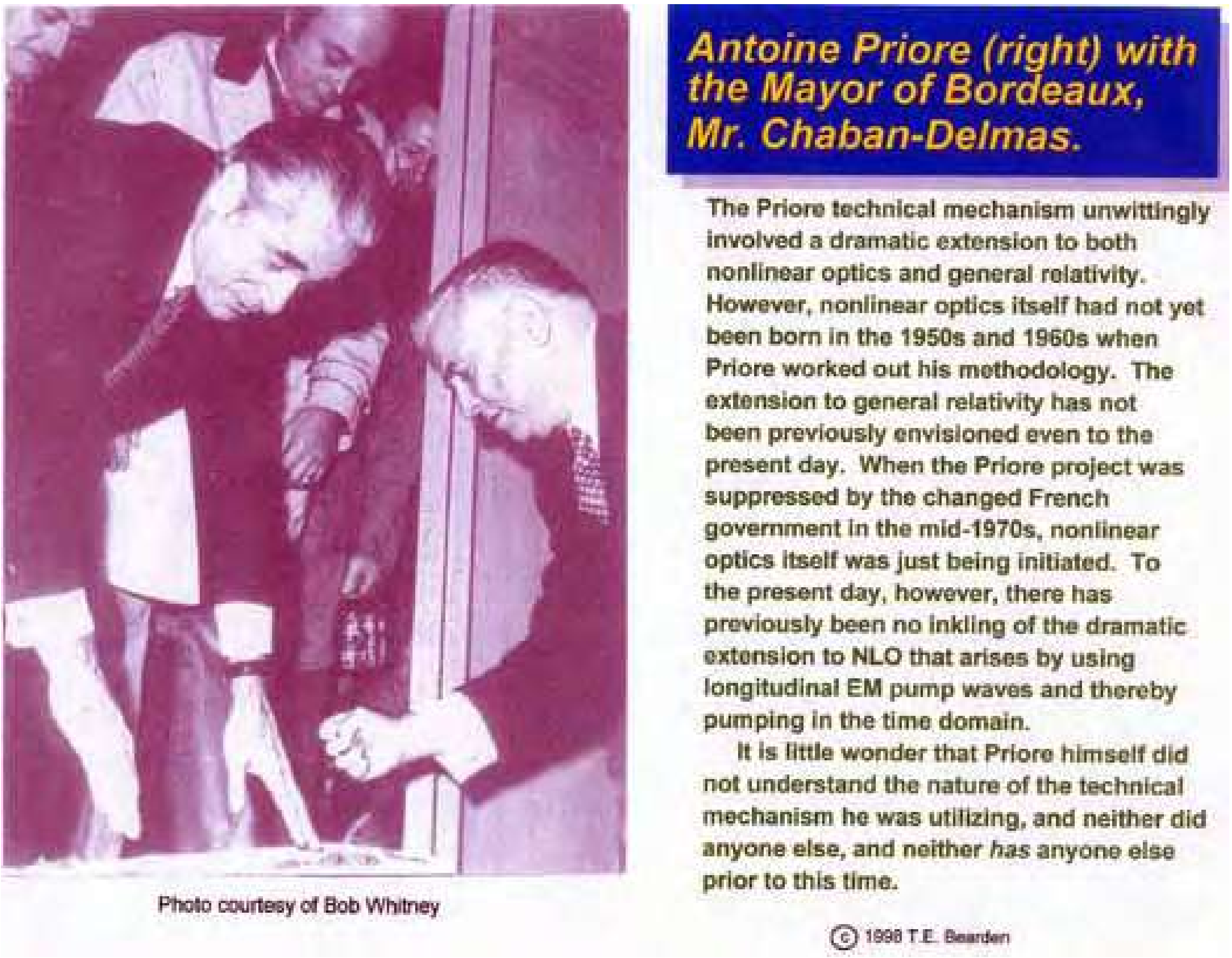

Antoine Priore (right) with the Mayor of Bordeaux, Mr. Chaban-Delmas.

Photo courtesy of Bob Whitney

The Priore technical mechanism unwittingly involved a dramatic extension to both nonlinear optics and general relativity. However, nonlinear optics itself had not yet been born in the 1950s and 1960s when Priore worked out his methodology. The extension to general relativity has not been previously envisioned even to the present day. When the Priore project was suppressed by the changed French government in the mid-1970s, nonlinear optics itself was just being initiated. To the present day, however, there has previously been no inkling of the dramatic extension to NLO that arises by using longitudinal EM pump waves and thereby pumping in the time domain.

It is little wonder that Priore himself did not understand the nature of the technical mechanism he was utilizing, and neither did anyone else, and neither *has* anyone else prior to this time.

© 1998 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

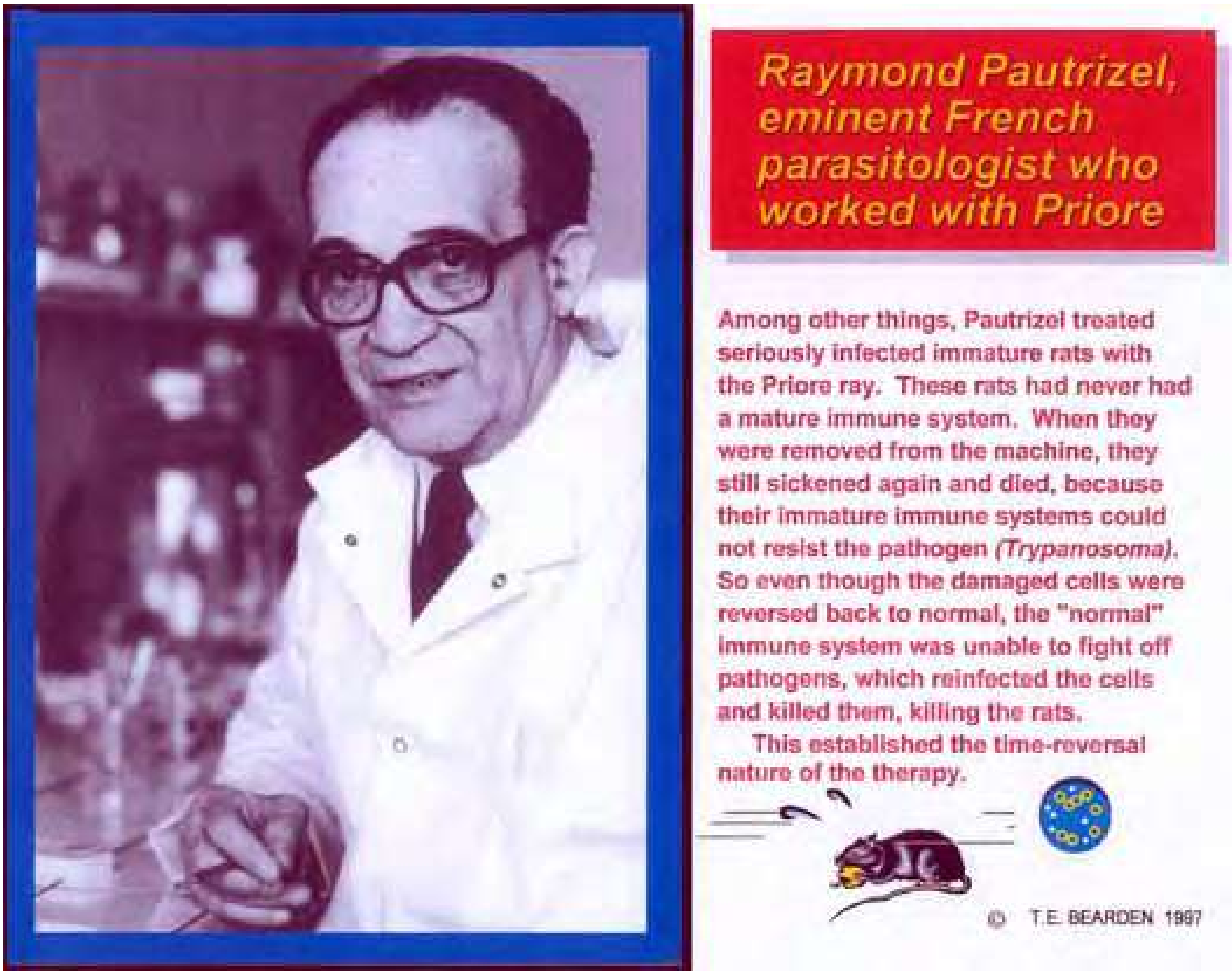

Raymond Pautrizel, eminent French parasitologist who worked with Priore

Among other things, Pautrizel treated seriously infected immature rats with the Priore ray. These rats had never had a mature immune system. When they were removed from the machine, they still sickened again and died, because their immature immune systems could not resist the pathogen *(Trypanosoma)*. So even though the damaged cells were reversed back to normal, the "normal" immune system was unable to fight off pathogens, which reinfected the cells and killed them, killing the rats.

This established the time-reversal nature of the therapy.

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

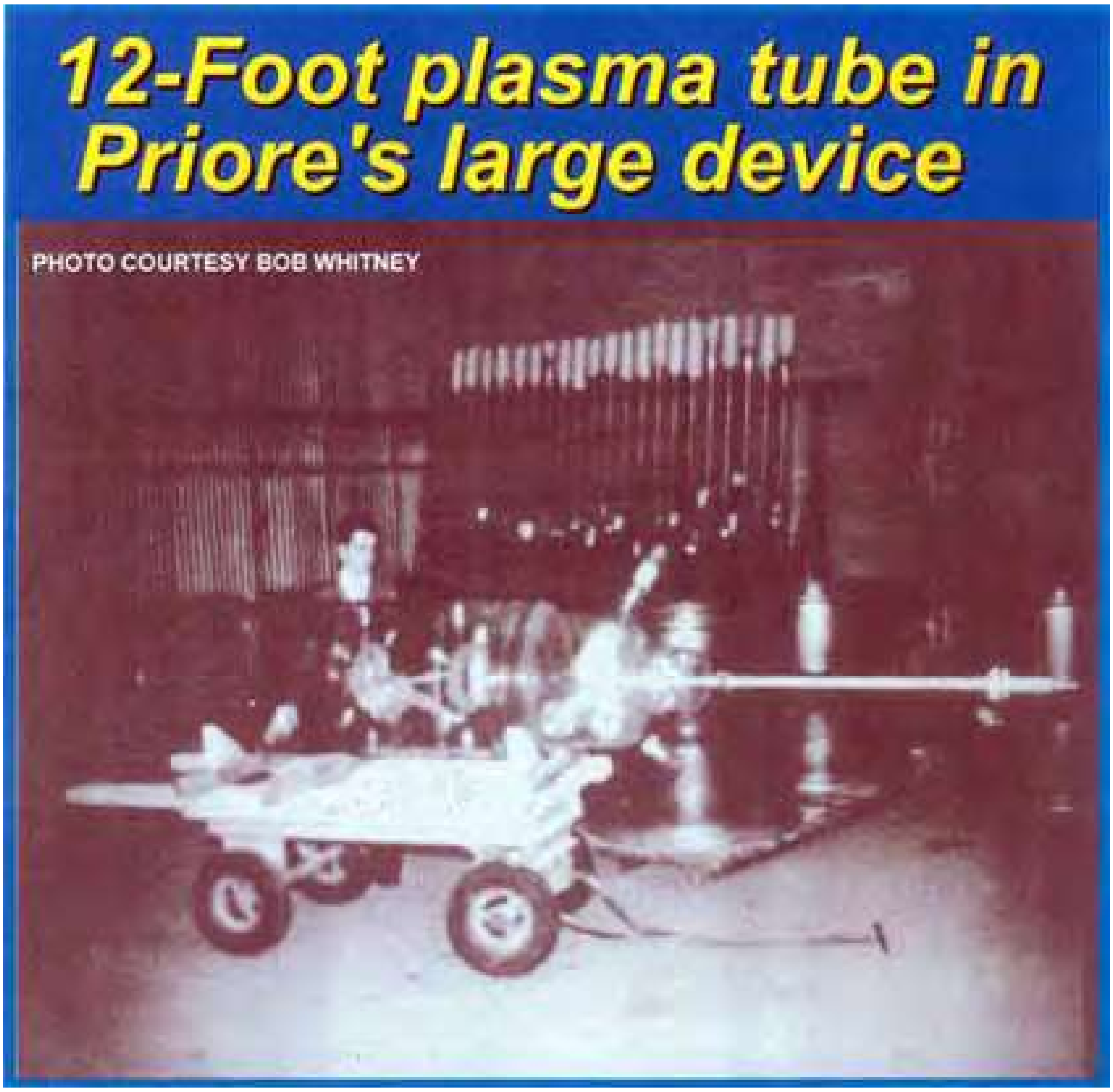

12-Foot plasma tube in Priore's large device

PHOTO COURTESY BOB WHITNEY

Return to Slide Index

Next Slide

Previous Slide

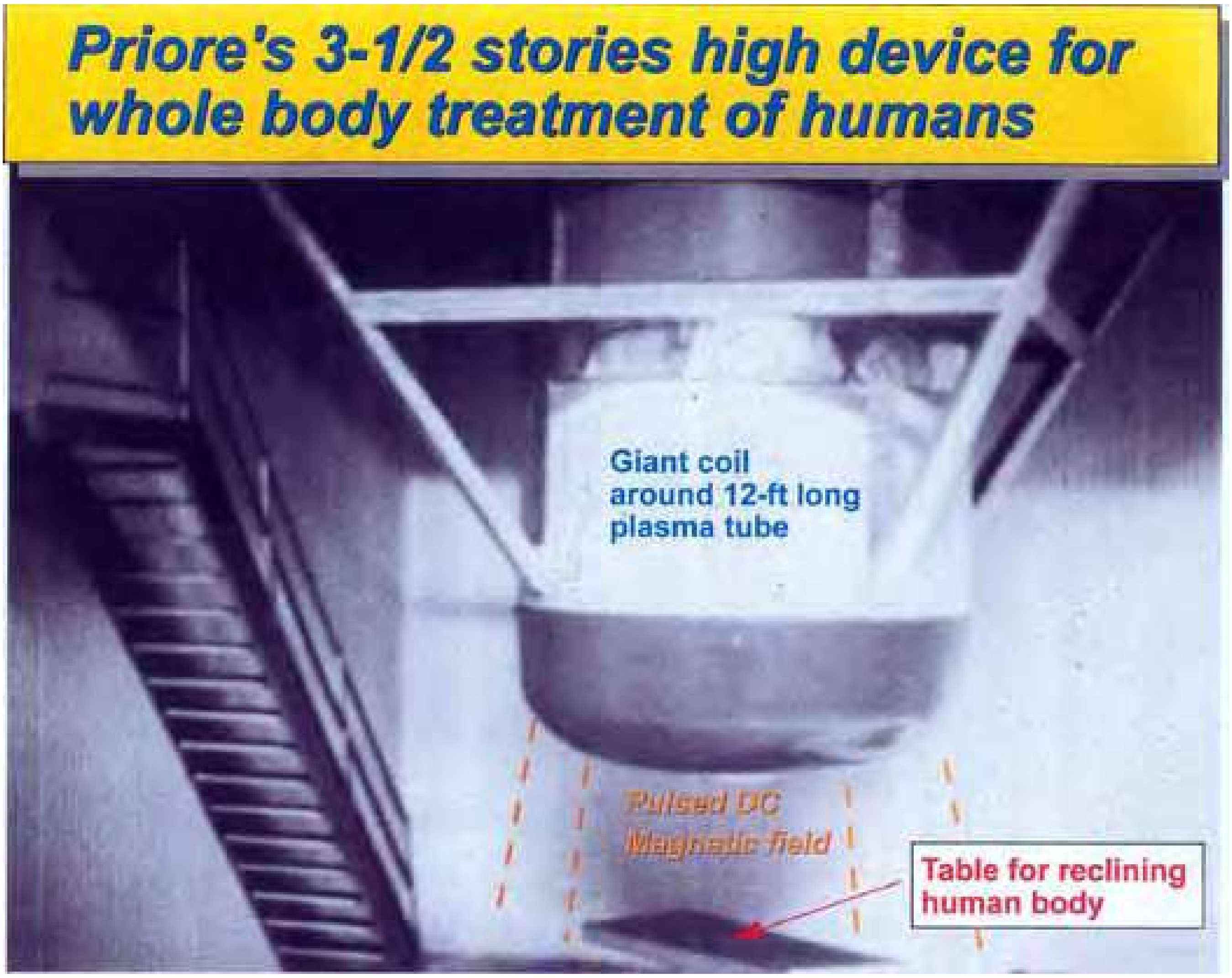

Priore's 3-1/2 stories high device for whole body treatment of humans

Giant coil
around 12-ft long
plasma tube

Pulsed DC
Magnetic field

Table for reclining
human body

[Return to Slide Index]

[Next Slide]

[Previous Slide]

MECHANISM OF THE PRIORE THERAPY

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1998 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

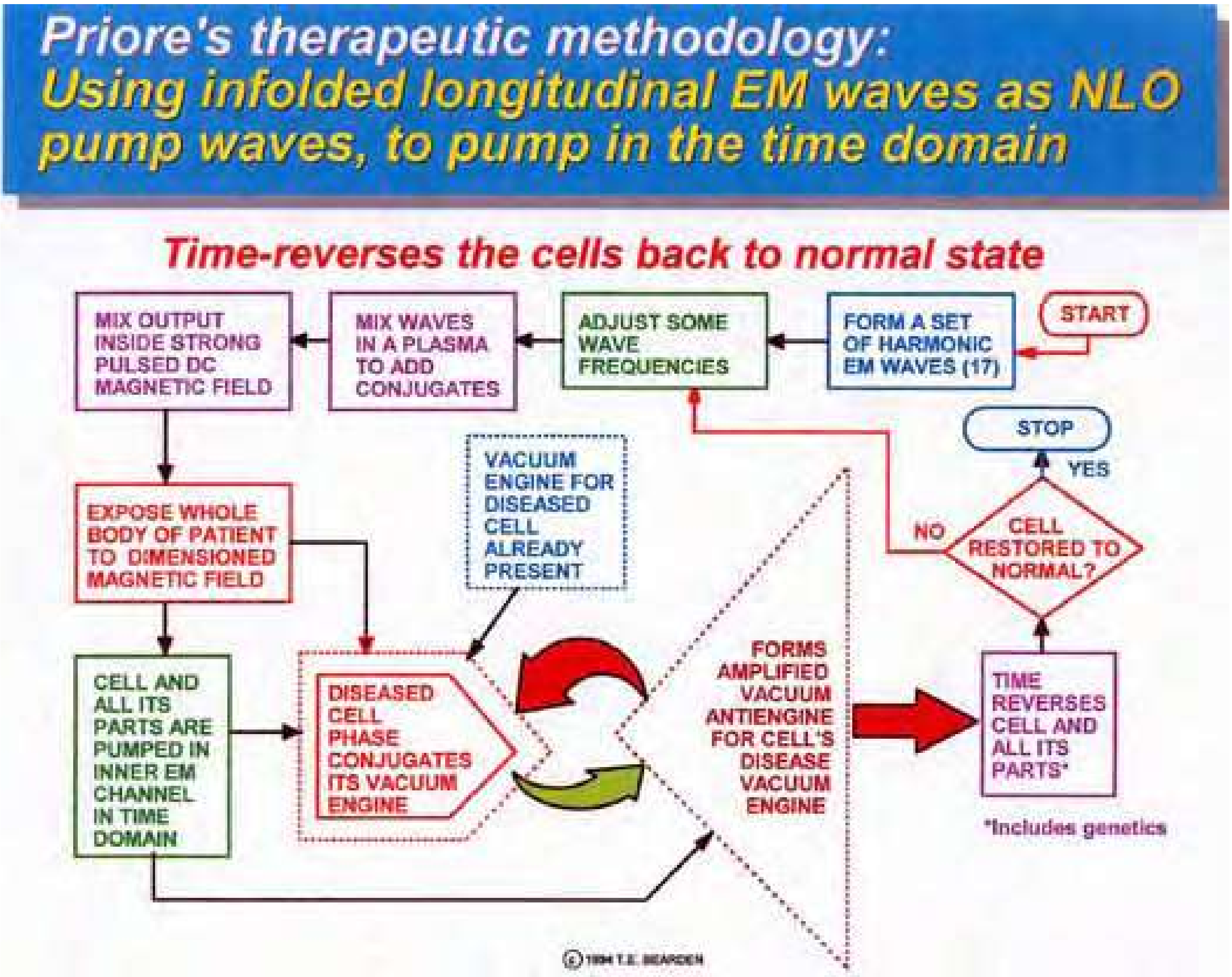

Priore's therapeutic methodology:
Using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain

Time-reverses the cells back to normal state

START

FORM A SET OF HARMONIC EM WAVES (17)

ADJUST SOME WAVE FREQUENCIES

MIX WAVES IN A PLASMA TO ADD CONJUGATES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME REVERSES CELL AND ALL ITS PARTS*

*Includes genetics

CELL RESTORED TO NORMAL?

NO

YES

STOP

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

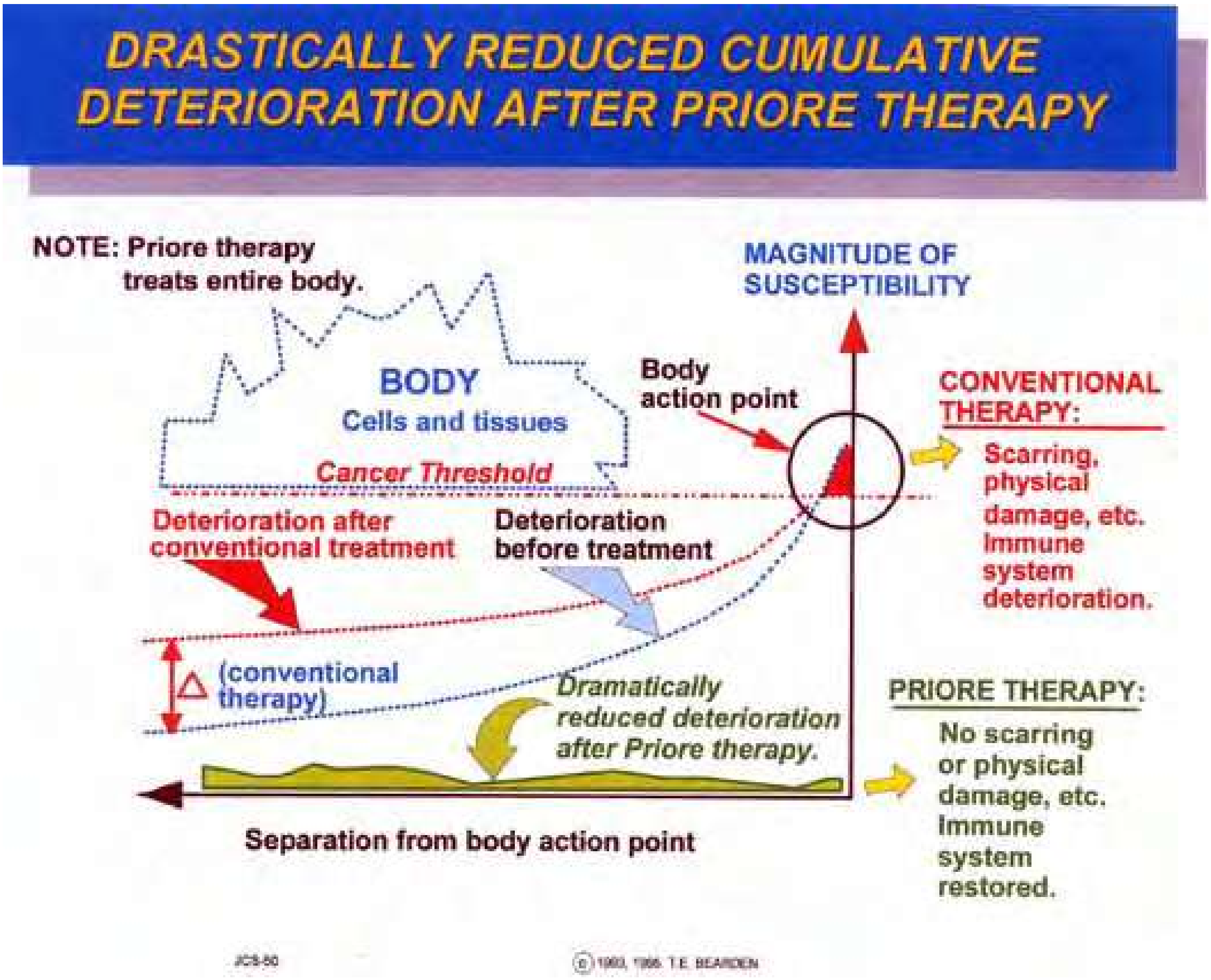

DRASTICALLY REDUCED CUMULATIVE DETERIORATION AFTER PRIORE THERAPY

NOTE: Priore therapy treats entire body.

BODY
Cells and tissues

Cancer Threshold

Body action point

MAGNITUDE OF SUSCEPTIBILITY

CONVENTIONAL THERAPY:

Scarring, physical damage, etc. Immune system deterioration.

Deterioration after conventional treatment

Deterioration before treatment

Δ (conventional therapy)

Dramatically reduced deterioration after Priore therapy.

PRIORE THERAPY:

No scarring or physical damage, etc. Immune system restored.

Separation from body action point

JCS-50

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

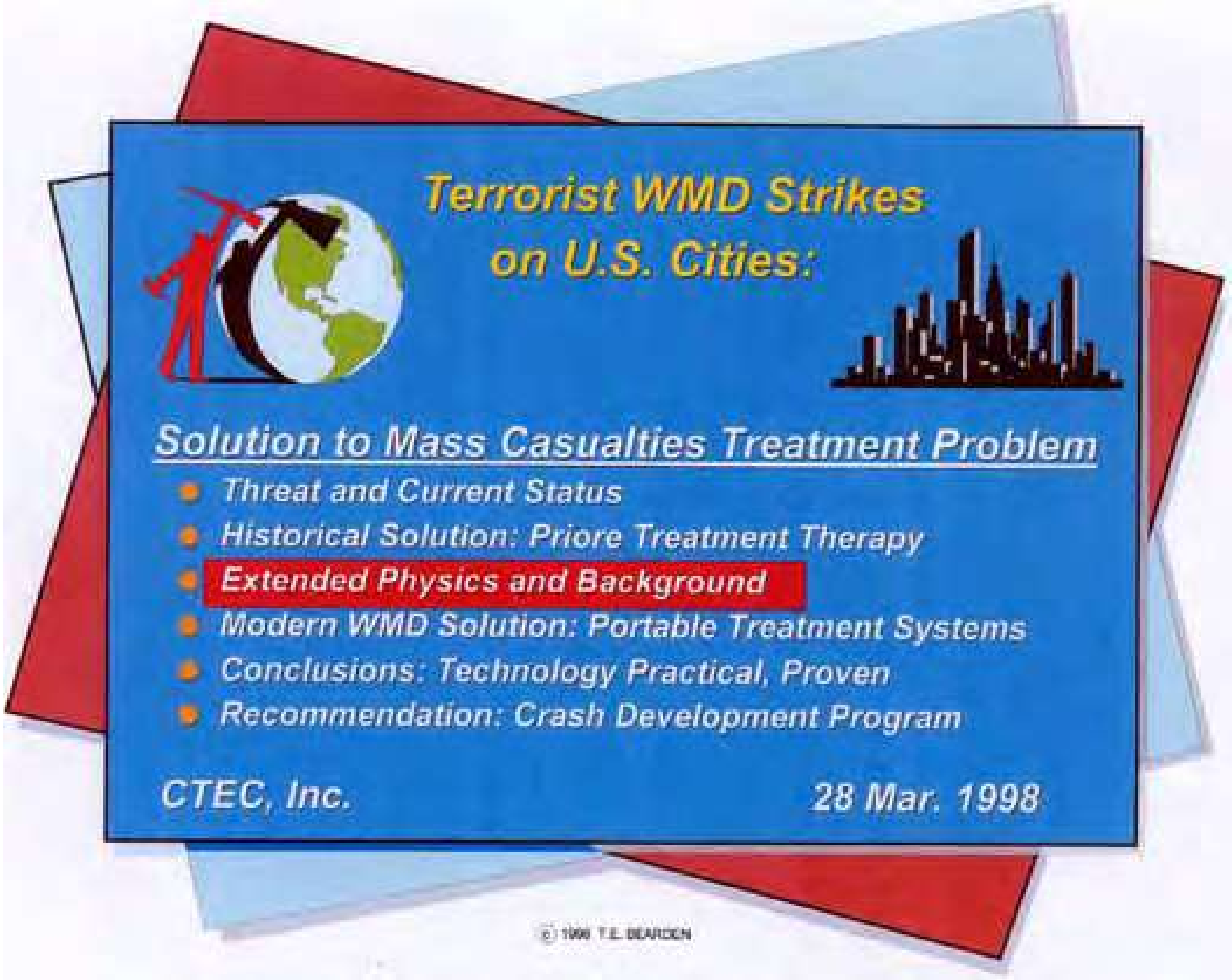

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

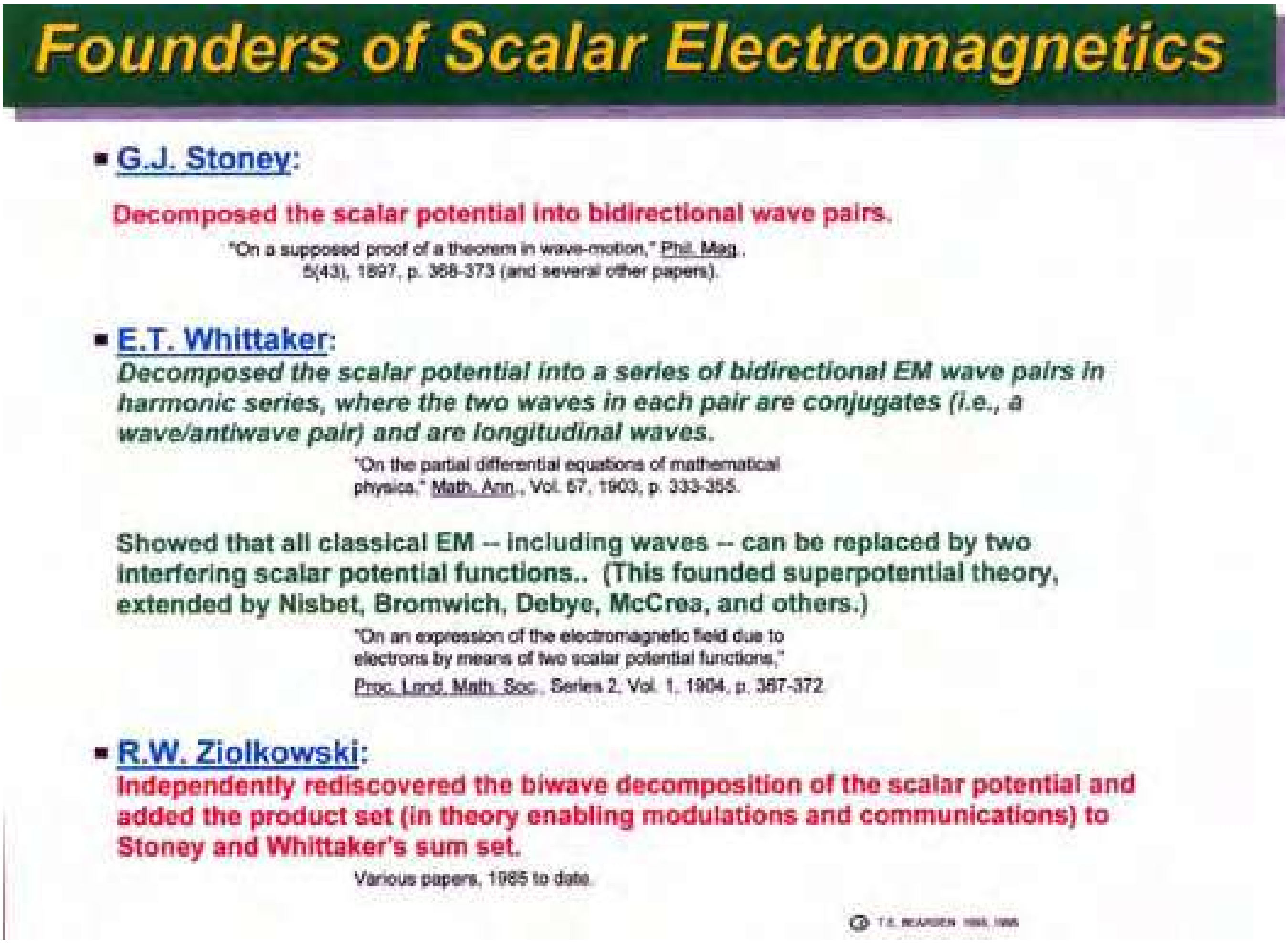

Founders of Scalar Electromagnetics

- **G.J. Stoney:**

Decomposed the scalar potential into bidirectional wave pairs.

"On a supposed proof of a theorem in wave-motion," Phil. Mag., 5(43), 1897, p. 368-373 (and several other papers).

- **E.T. Whittaker:**

Decomposed the scalar potential into a series of bidirectional EM wave pairs in harmonic series, where the two waves in each pair are conjugates (i.e., a wave/antiwave pair) and are longitudinal waves.

"On the partial differential equations of mathematical physics," Math. Ann., Vol. 57, 1903, p. 333-355.

Showed that all classical EM -- including waves -- can be replaced by two interfering scalar potential functions.. (This founded superpotential theory, extended by Nisbet, Bromwich, Debye, McCrea, and others.)

"On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372

- **R.W. Ziolkowski:**

Independently rediscovered the biwave decomposition of the scalar potential and added the product set (in theory enabling modulations and communications) to Stoney and Whittaker's sum set.

Various papers, 1985 to date.

© T.E. BEARDEN 1995, 1996

Return to Slide Index

Next Slide

Previous Slide

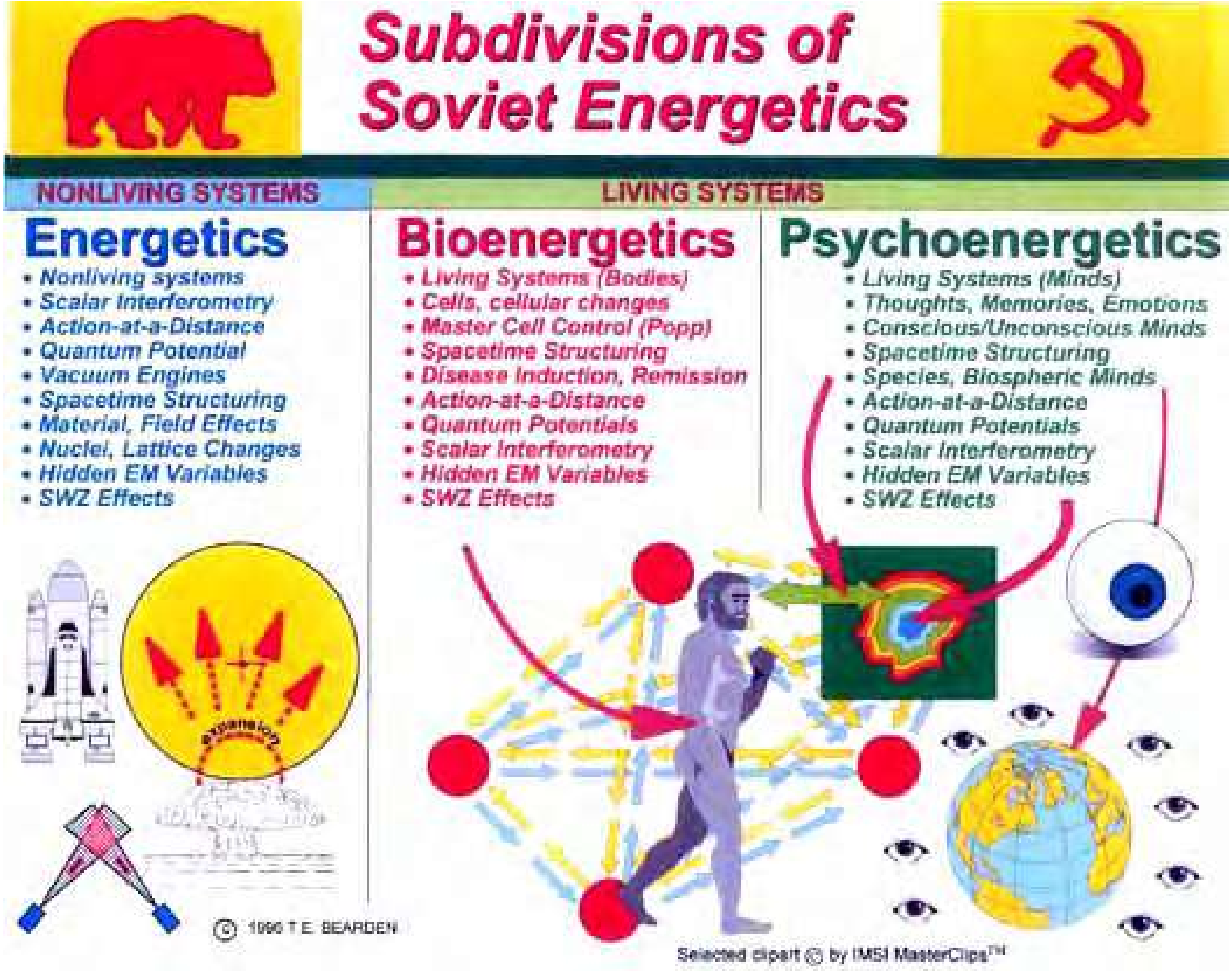

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

LIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1990 T.E. BEARDEN

Selected clipart © by IMSI MasterClips™

Return to Slide Index

Next Slide

Previous Slide

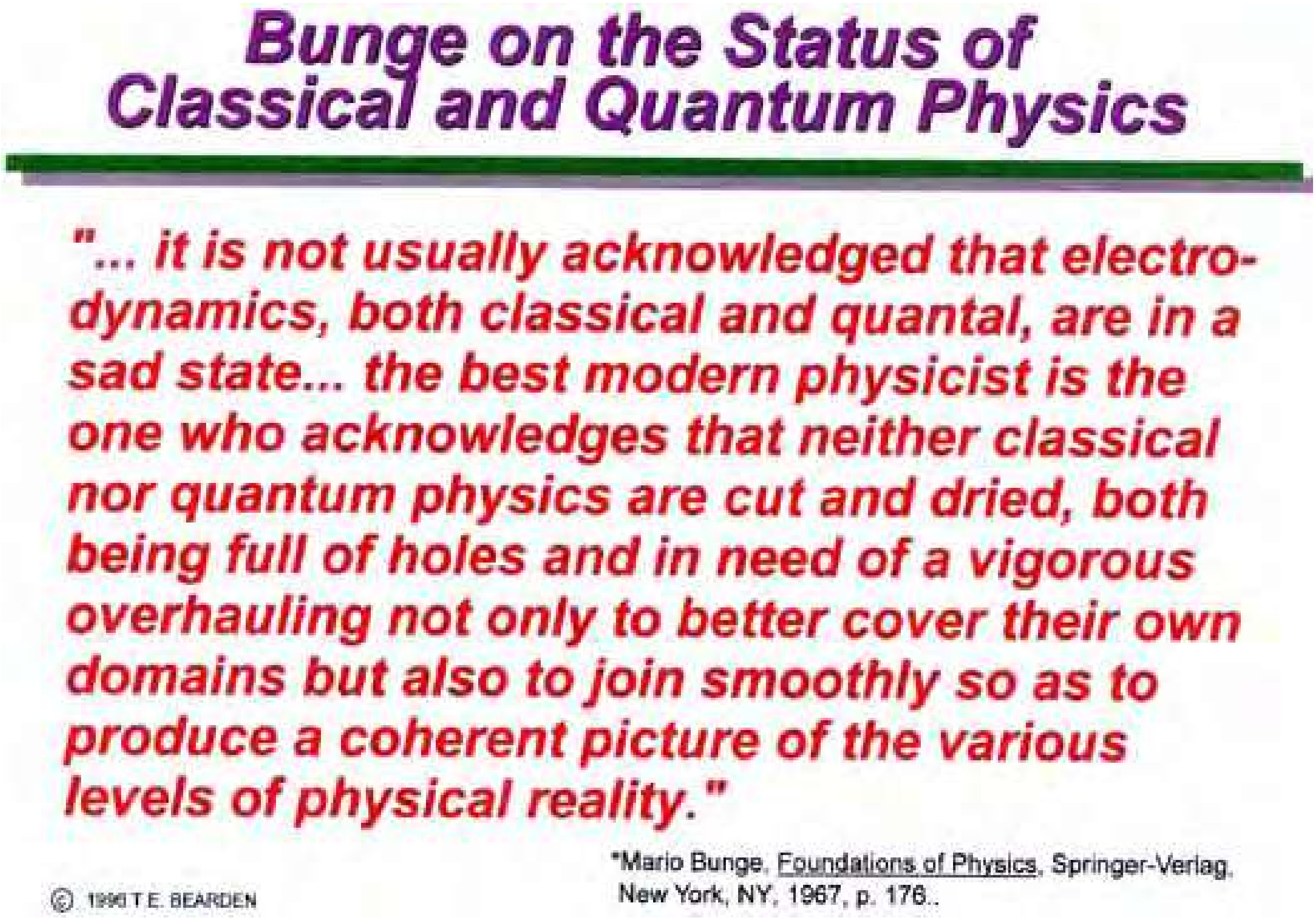

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

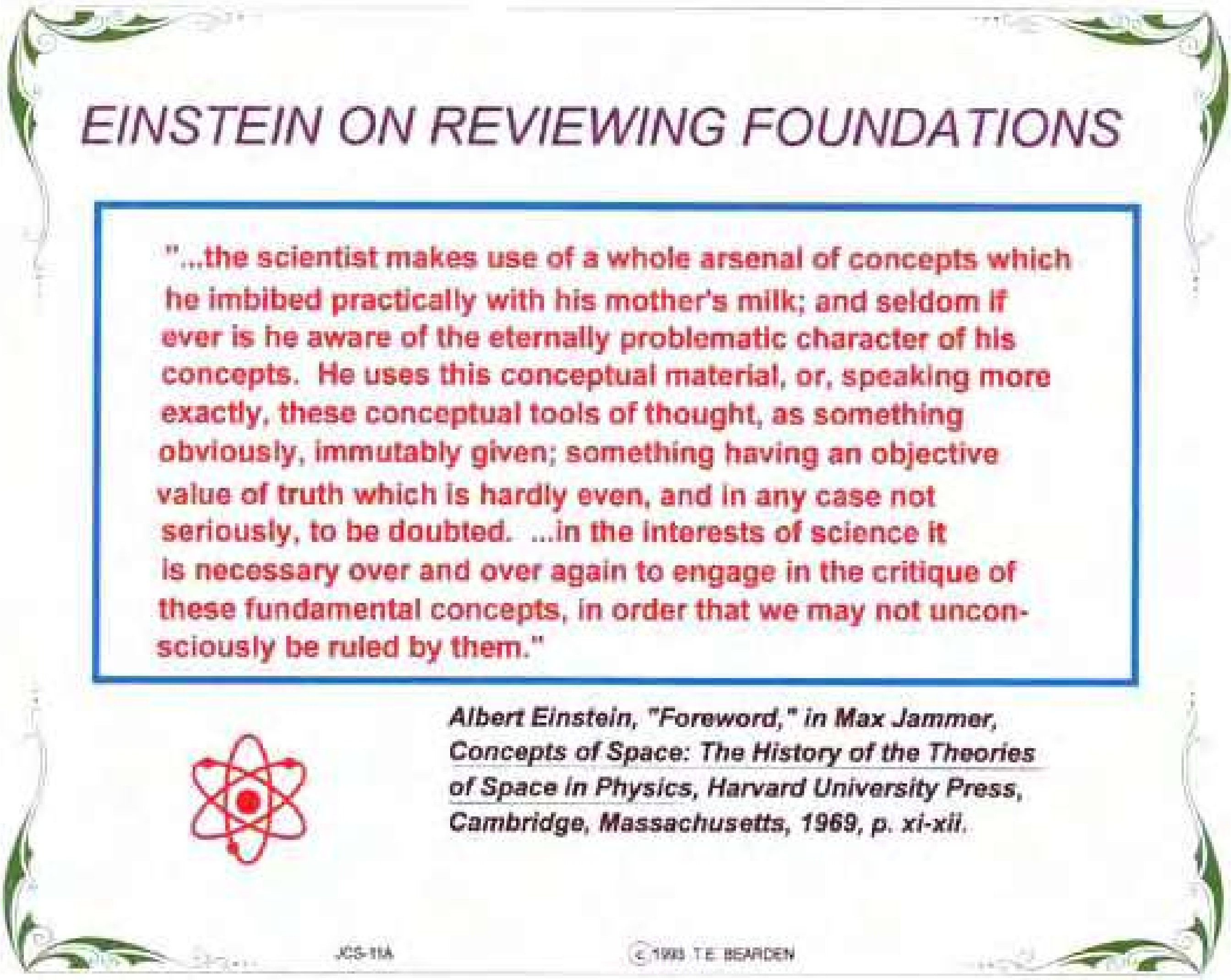

EINSTEIN ON REVIEWING FOUNDATIONS

> "...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

JCS-11A

© 1993 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

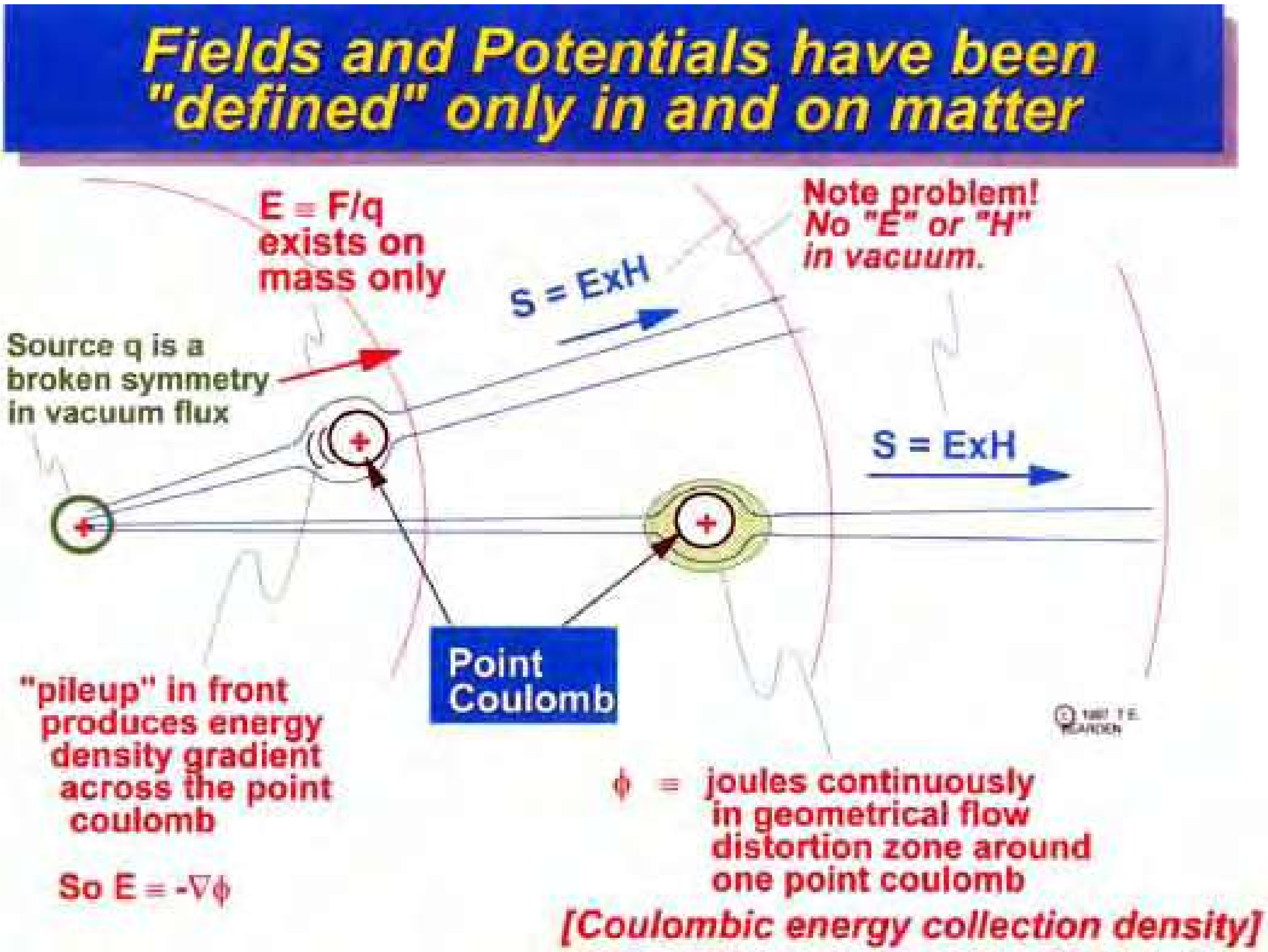

Fields and Potentials have been "defined" only in and on matter

$E \equiv F/q$ exists on mass only

Note problem! *No "E" or "H" in vacuum.*

$S = E \times H$

Source q is a broken symmetry in vacuum flux

$S = E \times H$

Point Coulomb

"pileup" in front produces energy density gradient across the point coulomb

So $E \equiv -\nabla\phi$

© 1987 T.E. BEARDEN

$\phi \equiv$ joules continuously in geometrical flow distortion zone around one point coulomb

[Coulombic energy collection density]

Return to Slide Index

Next Slide

Previous Slide

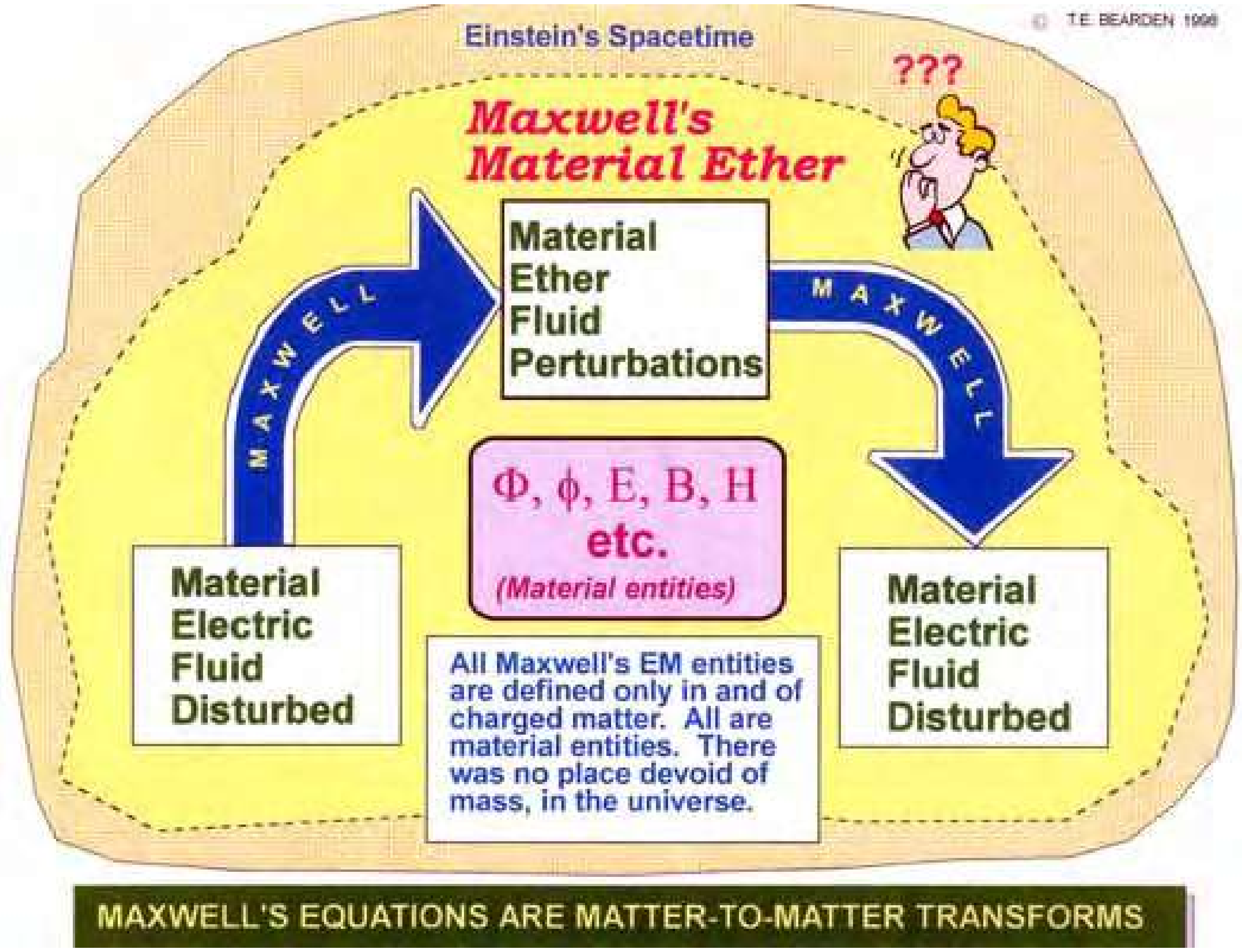

© T.E. BEARDEN 1996

Einstein's Spacetime

Maxwell's Material Ether

Material Ether Fluid Perturbations

Material Electric Fluid Disturbed

Φ, ϕ, E, B, H etc.

(Material entities)

All Maxwell's EM entities are defined only in and of charged matter. All are material entities. There was no place devoid of mass, in the universe.

Material Electric Fluid Disturbed

MAXWELL'S EQUATIONS ARE MATTER-TO-MATTER TRANSFORMS

Return to Slide Index

Next Slide

Previous Slide

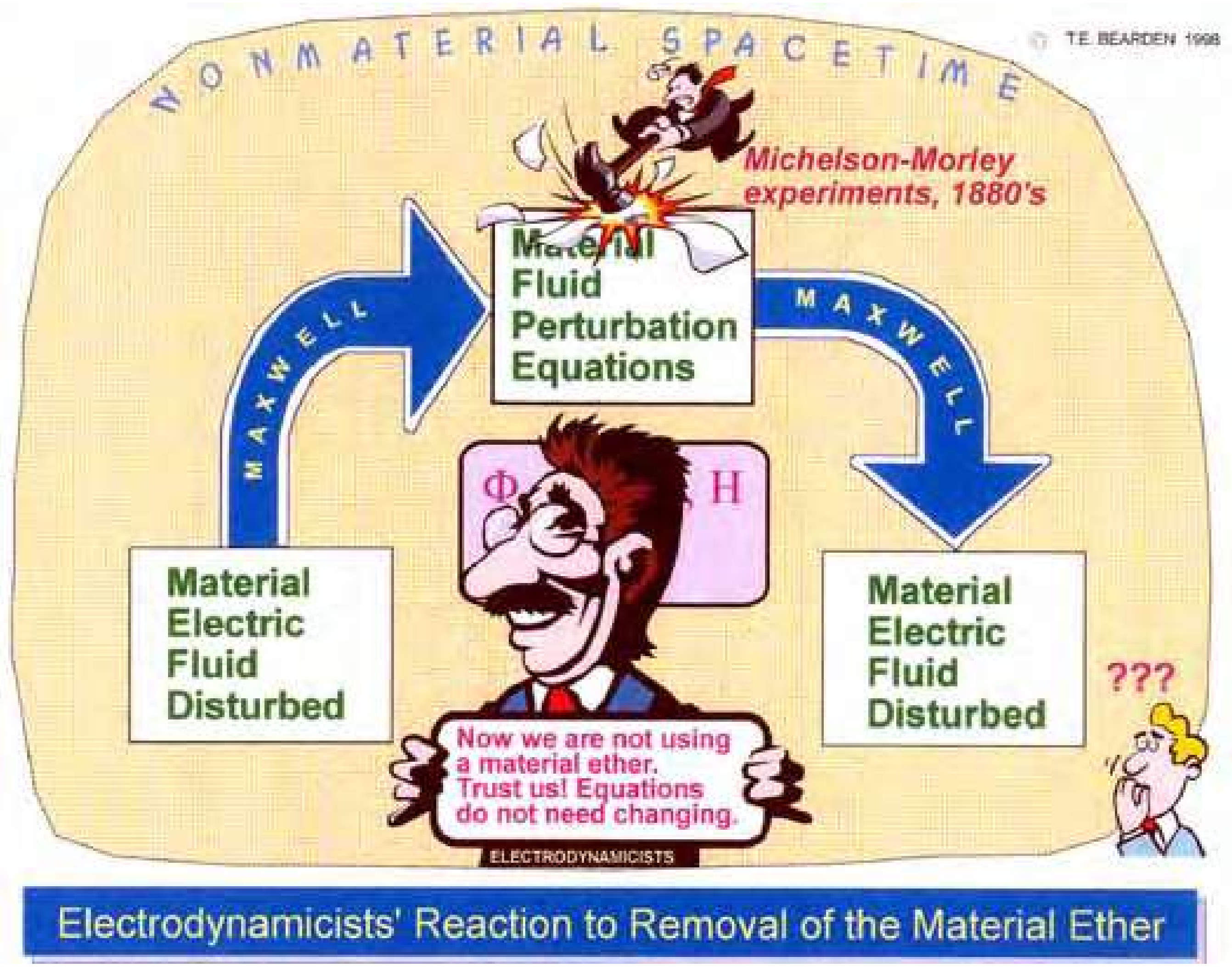
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Michelson-Morley experiments, 1880's
Material Fluid Perturbation Equations
MAXWELL
MAXWELL
Φ
H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! Equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

Return to Slide Index

Next Slide

Previous Slide

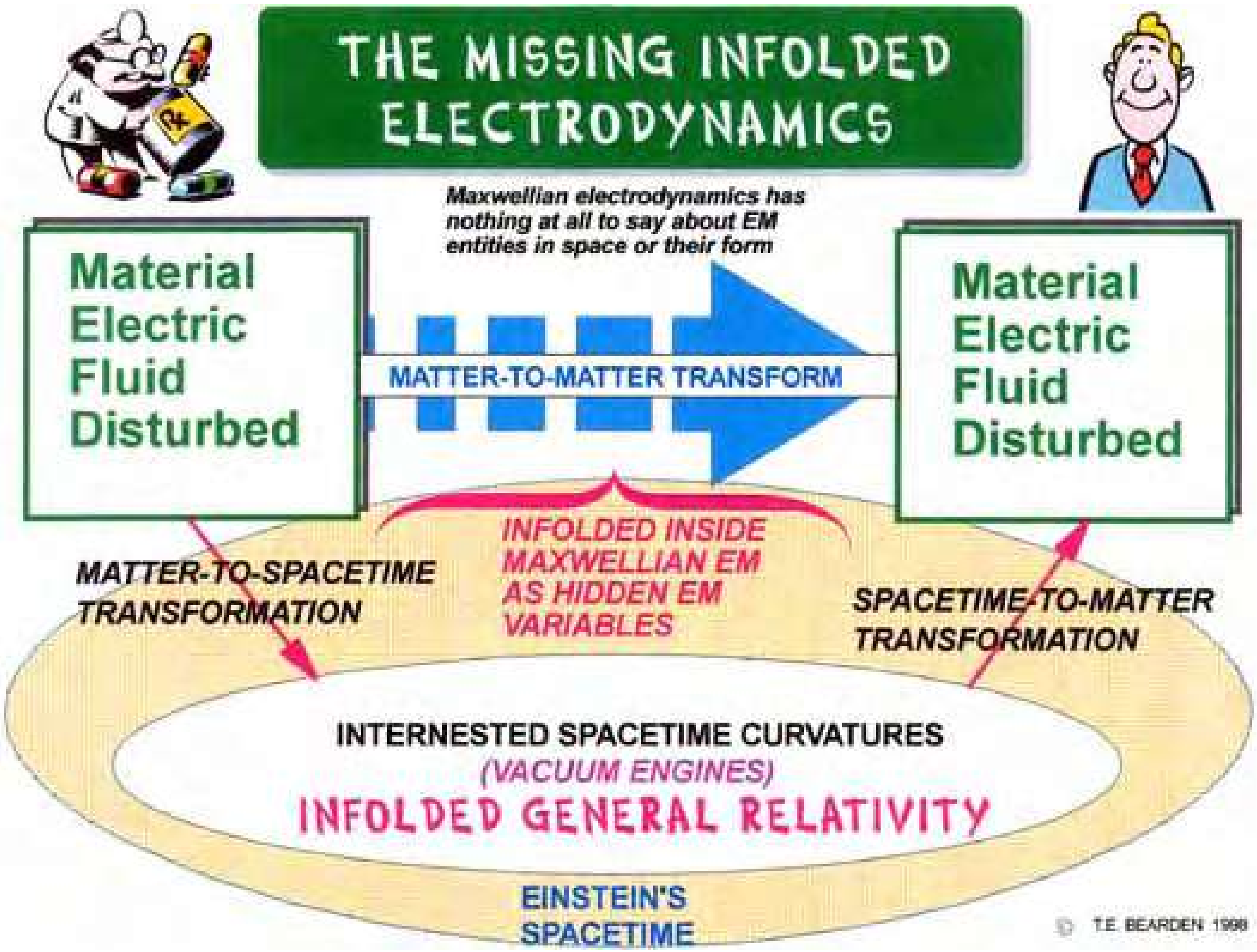

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

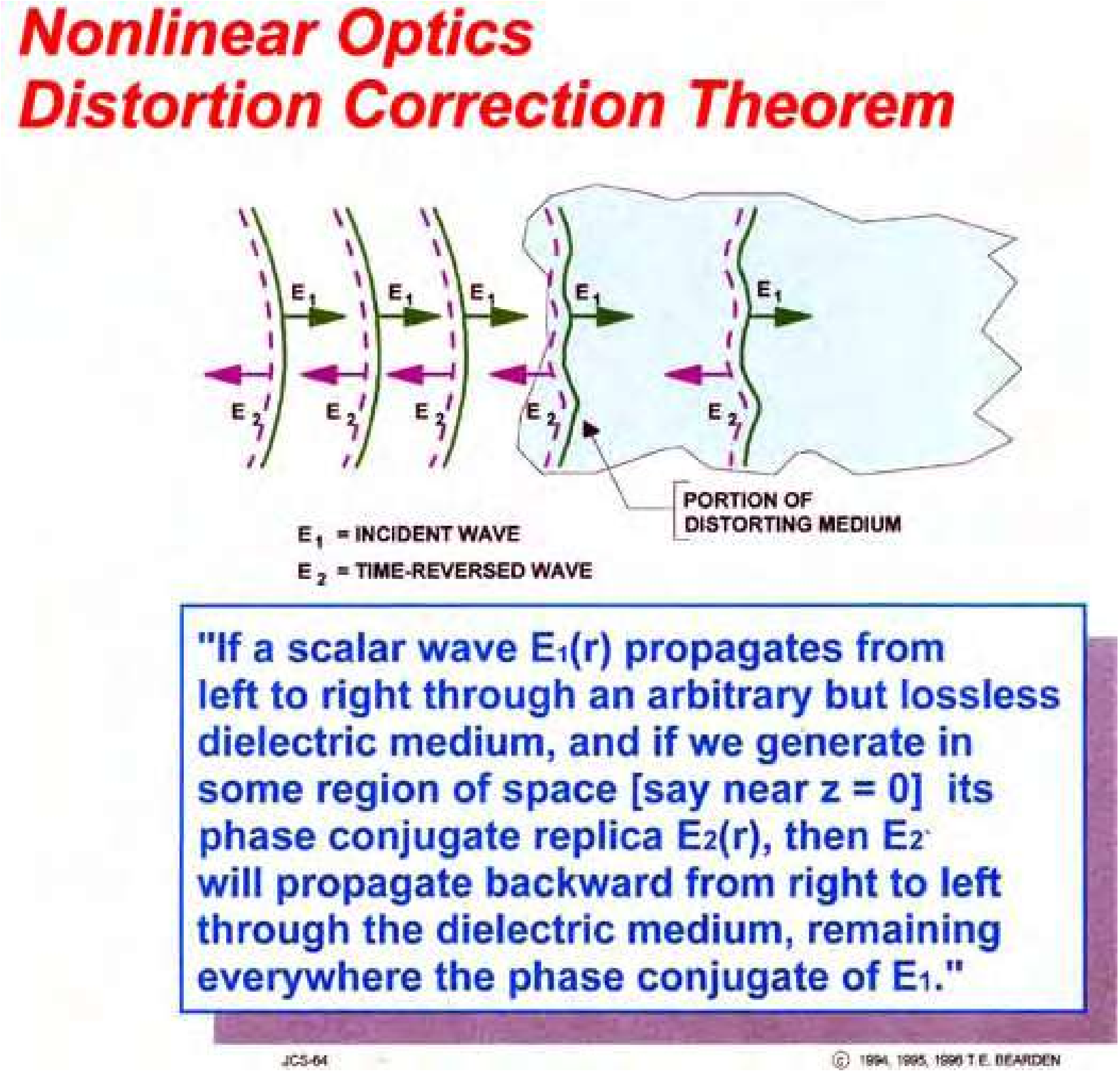

Nonlinear Optics
Distortion Correction Theorem

PORTION OF DISTORTING MEDIUM

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near $z = 0$] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

JCS-64

© 1994, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

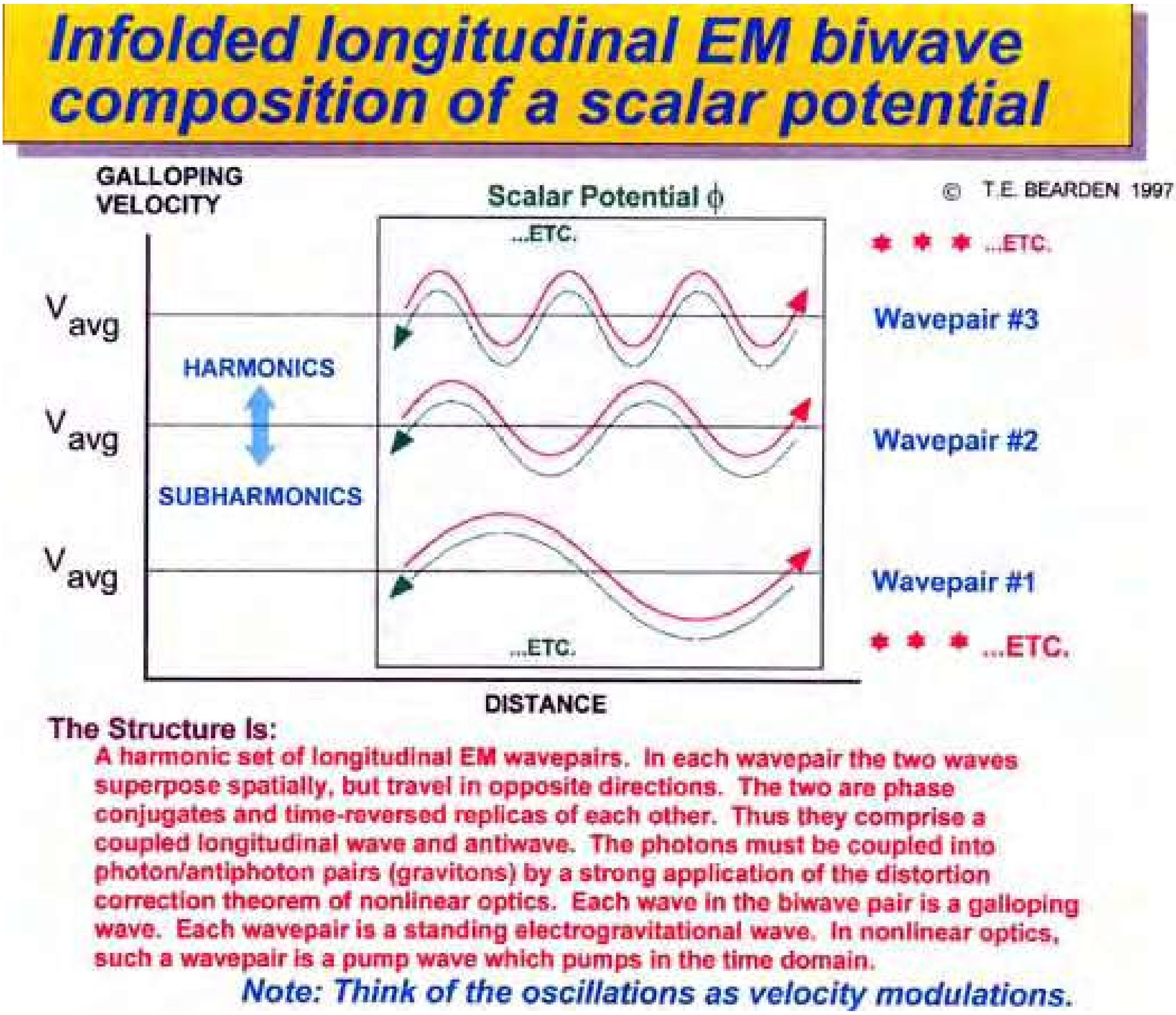

Infolded longitudinal EM biwave composition of a scalar potential

GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC.
* * * ...ETC.
V_{avg}
Wavepair #3
HARMONICS
V_{avg}
Wavepair #2
SUBHARMONICS
V_{avg}
Wavepair #1
...ETC.
* * * ...ETC.
DISTANCE

The Structure Is:

A harmonic set of longitudinal EM wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave. In nonlinear optics, such a wavepair is a pump wave which pumps in the time domain.

Note: Think of the oscillations as velocity modulations.

Return to Slide Index

Next Slide

Previous Slide

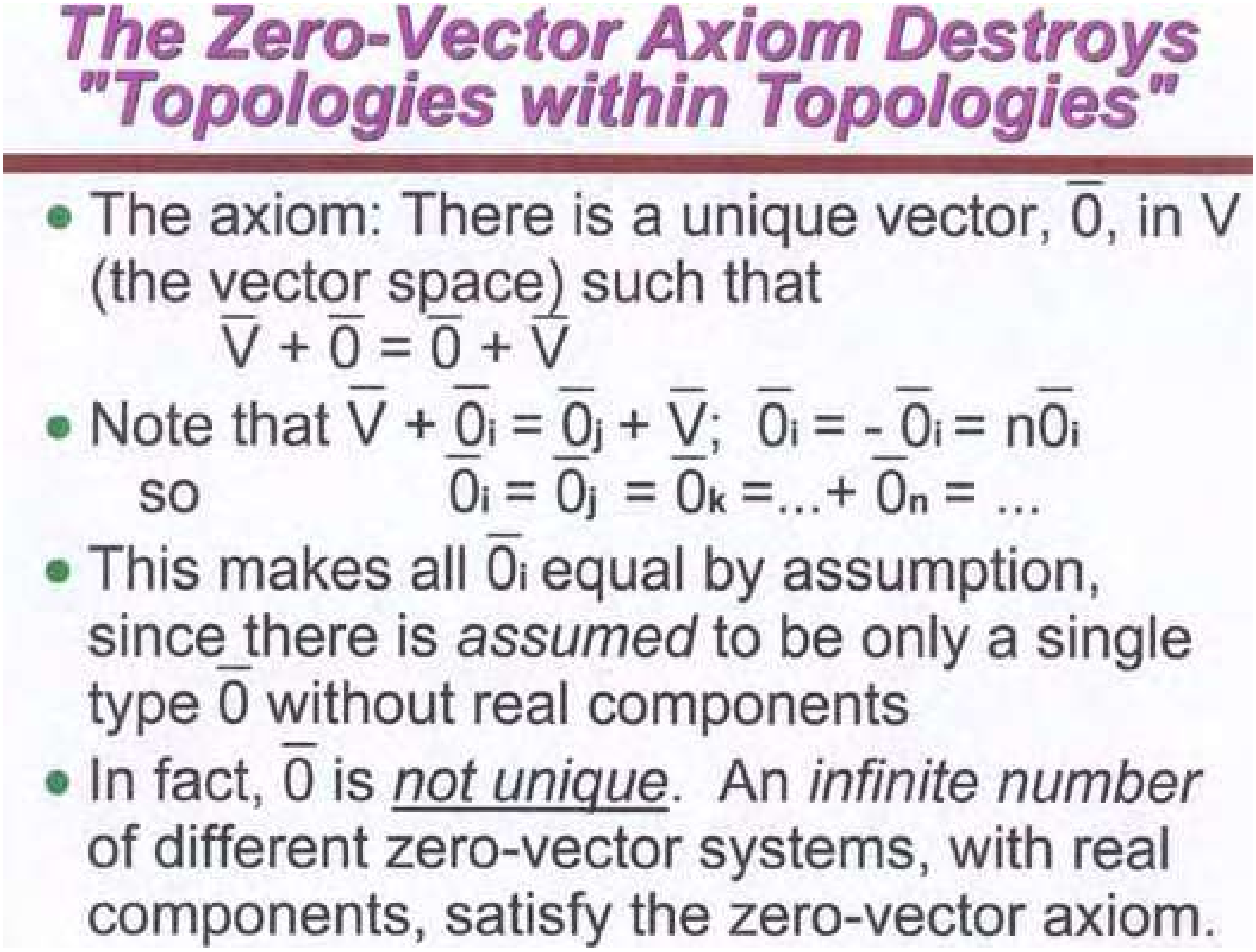

The Zero-Vector Axiom Destroys "Topologies within Topologies"

- The axiom: There is a unique vector, $\overline{0}$, in V (the vector space) such that
  $$\overline{V} + \overline{0} = \overline{0} + \overline{V}$$
- Note that $\overline{V} + \overline{0}_i = \overline{0}_j + \overline{V}$; $\overline{0}_i = -\overline{0}_i = n\overline{0}_i$
  so $\overline{0}_i = \overline{0}_j = \overline{0}_k = \ldots + \overline{0}_n = \ldots$
- This makes all $\overline{0}_i$ equal by assumption, since there is *assumed* to be only a single type $\overline{0}$ without real components
- In fact, $\overline{0}$ is *not unique*. An *infinite number* of different zero-vector systems, with real components, satisfy the zero-vector axiom.

Return to Slide Index

Next Slide

Previous Slide

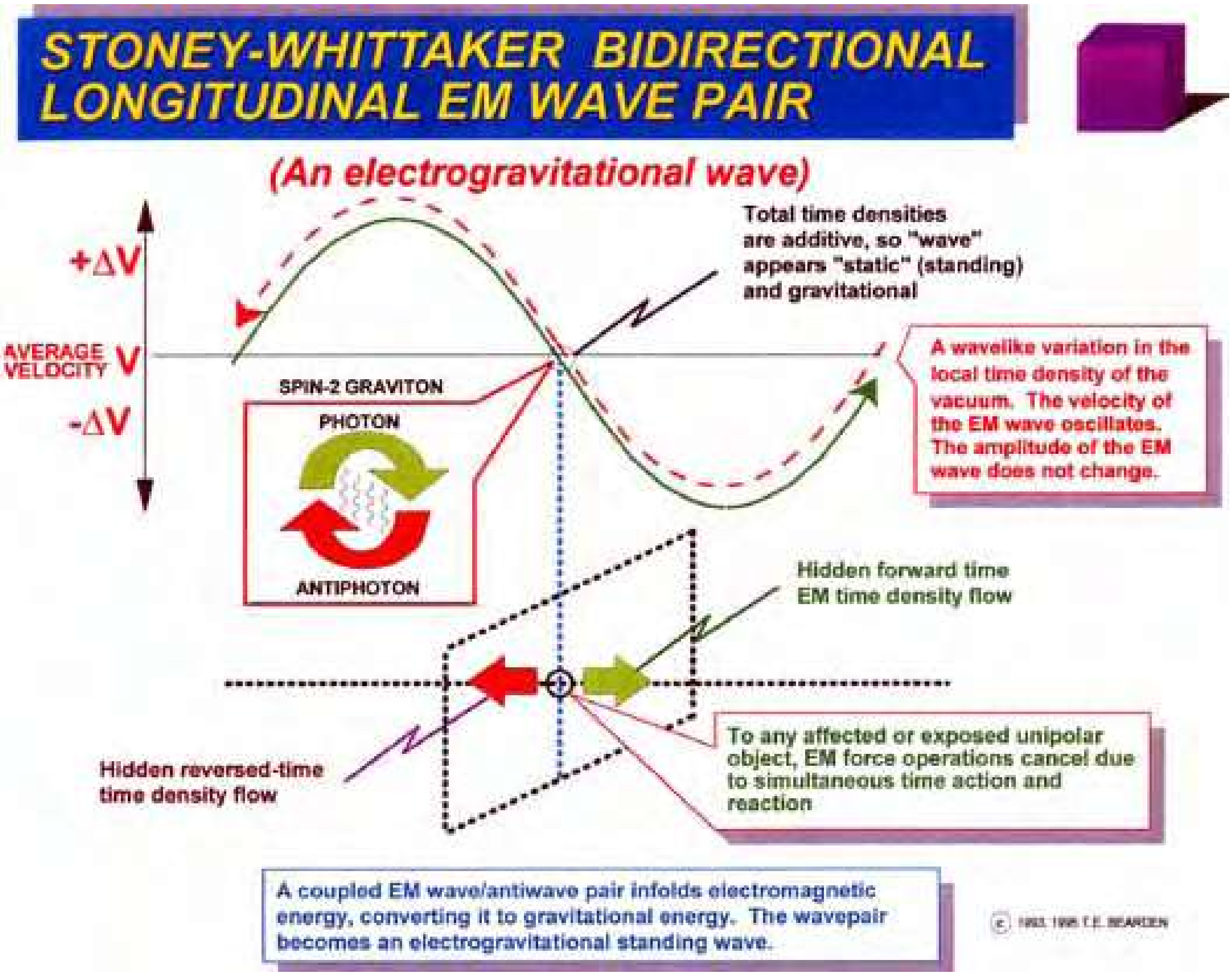

STONEY-WHITTAKER BIDIRECTIONAL LONGITUDINAL EM WAVE PAIR

(An electrogravitational wave)

Total time densities are additive, so "wave" appears "static" (standing) and gravitational

A wavelike variation in the local time density of the vacuum. The velocity of the EM wave oscillates. The amplitude of the EM wave does not change.

Hidden forward time EM time density flow

Hidden reversed-time time density flow

To any affected or exposed unipolar object, EM force operations cancel due to simultaneous time action and reaction

A coupled EM wave/antiwave pair infolds electromagnetic energy, converting it to gravitational energy. The wavepair becomes an electrogravitational standing wave.

Return to Slide Index

Next Slide

Previous Slide

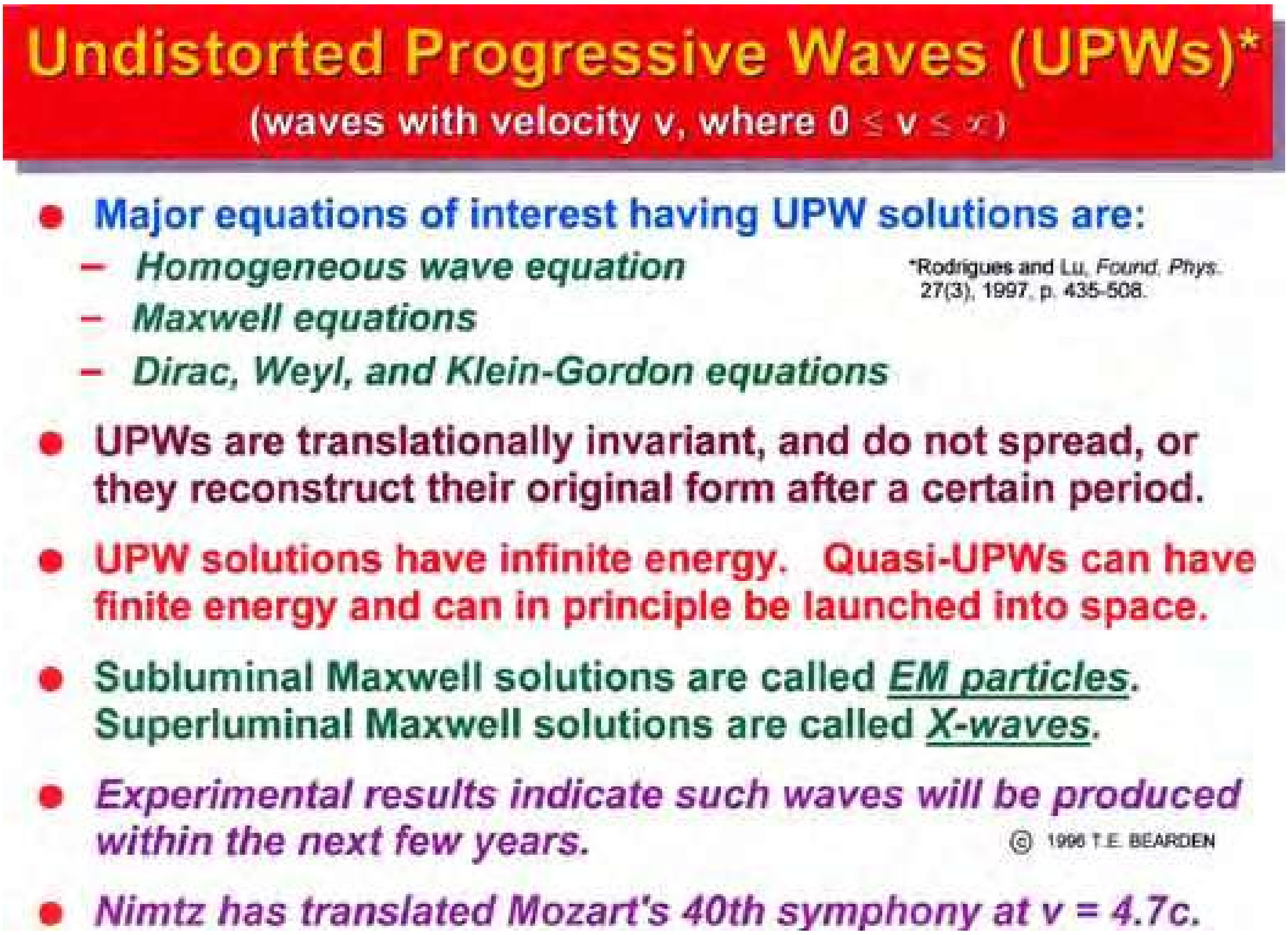

Undistorted Progressive Waves (UPWs)*

(waves with velocity v, where $0 \leq v \leq \infty$)

- Major equations of interest having UPW solutions are:
  - *Homogeneous wave equation*
  - *Maxwell equations*
  - *Dirac, Weyl, and Klein-Gordon equations*
- UPWs are translationally invariant, and do not spread, or they reconstruct their original form after a certain period.
- UPW solutions have infinite energy. Quasi-UPWs can have finite energy and can in principle be launched into space.
- Subluminal Maxwell solutions are called *EM particles*. Superluminal Maxwell solutions are called *X-waves*.
- *Experimental results indicate such waves will be produced within the next few years.*
- *Nimtz has translated Mozart's 40th symphony at v = 4.7c.*

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

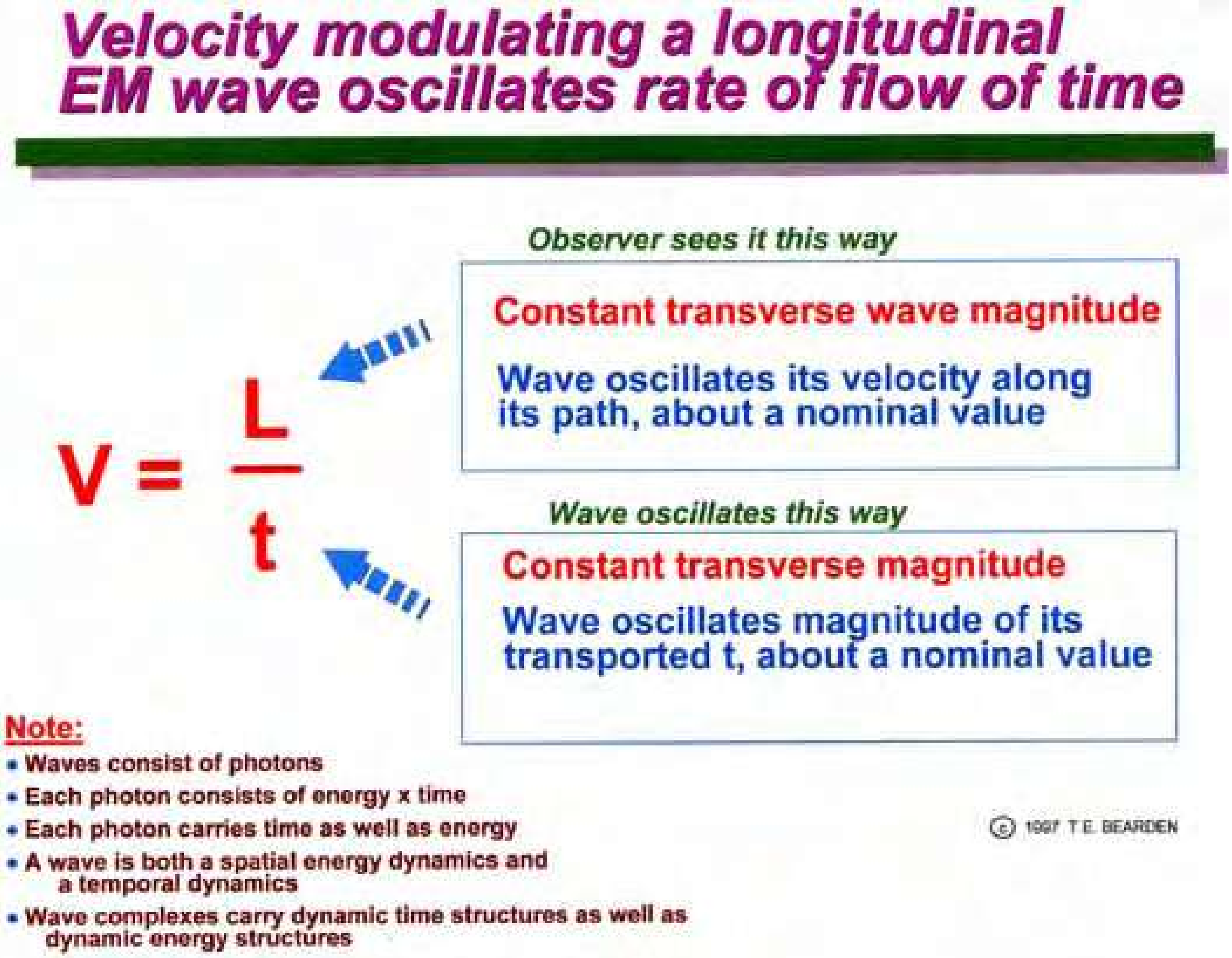

Velocity modulating a longitudinal EM wave oscillates rate of flow of time

$$V = \frac{L}{t}$$

Observer sees it this way

Constant transverse wave magnitude

Wave oscillates its velocity along its path, about a nominal value

Wave oscillates this way

Constant transverse magnitude

Wave oscillates magnitude of its transported t, about a nominal value

Note:

- Waves consist of photons
- Each photon consists of energy x time
- Each photon carries time as well as energy
- A wave is both a spatial energy dynamics and a temporal dynamics
- Wave complexes carry dynamic time structures as well as dynamic energy structures

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

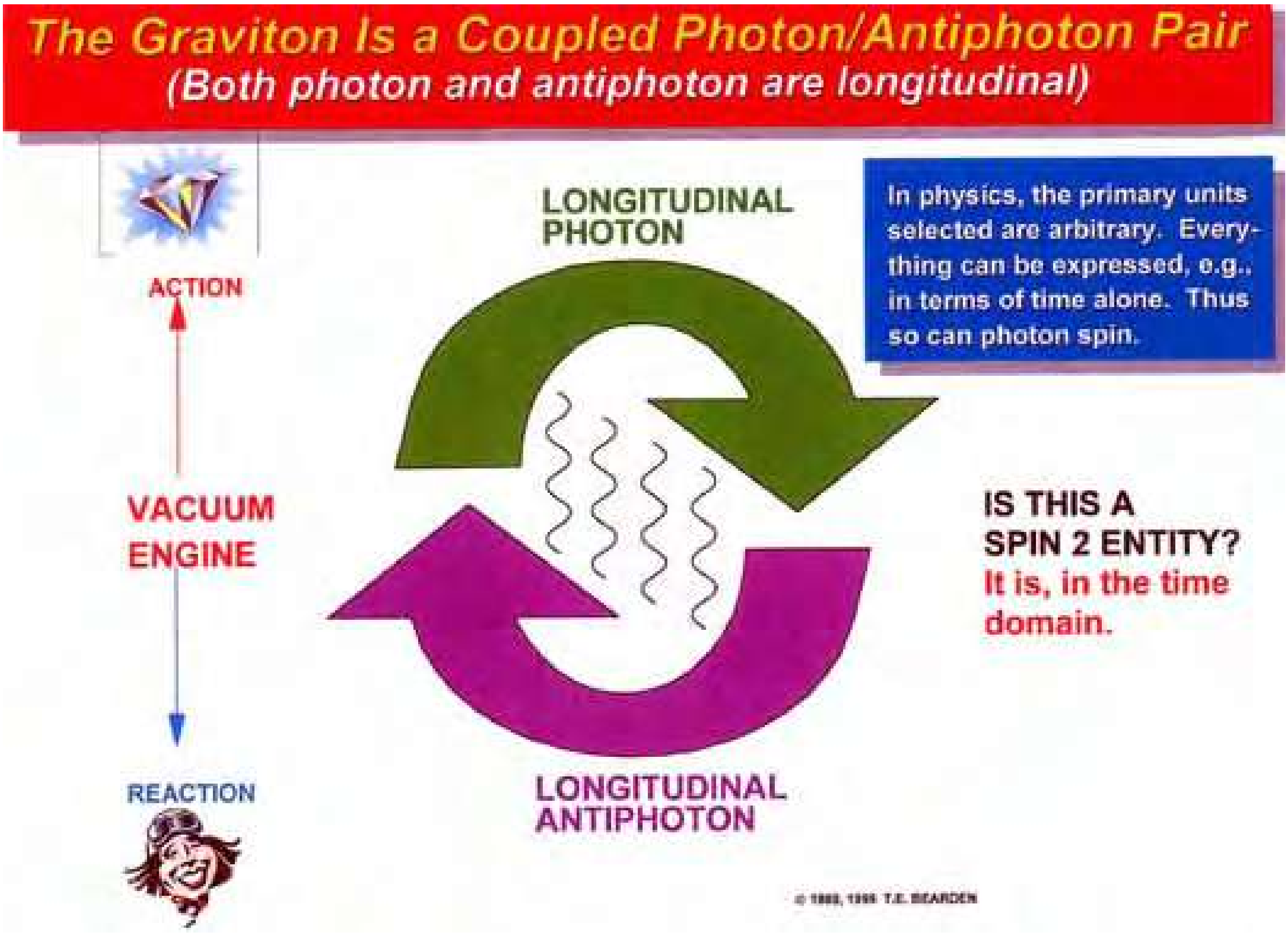

The Graviton Is a Coupled Photon/Antiphoton Pair

(Both photon and antiphoton are longitudinal)

In physics, the primary units selected are arbitrary. Everything can be expressed, e.g., in terms of time alone. Thus so can photon spin.

IS THIS A SPIN 2 ENTITY?

It is, in the time domain.

ACTION
VACUUM ENGINE
REACTION
LONGITUDINAL PHOTON
LONGITUDINAL ANTIPHOTON

© [illegible], [illegible] T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

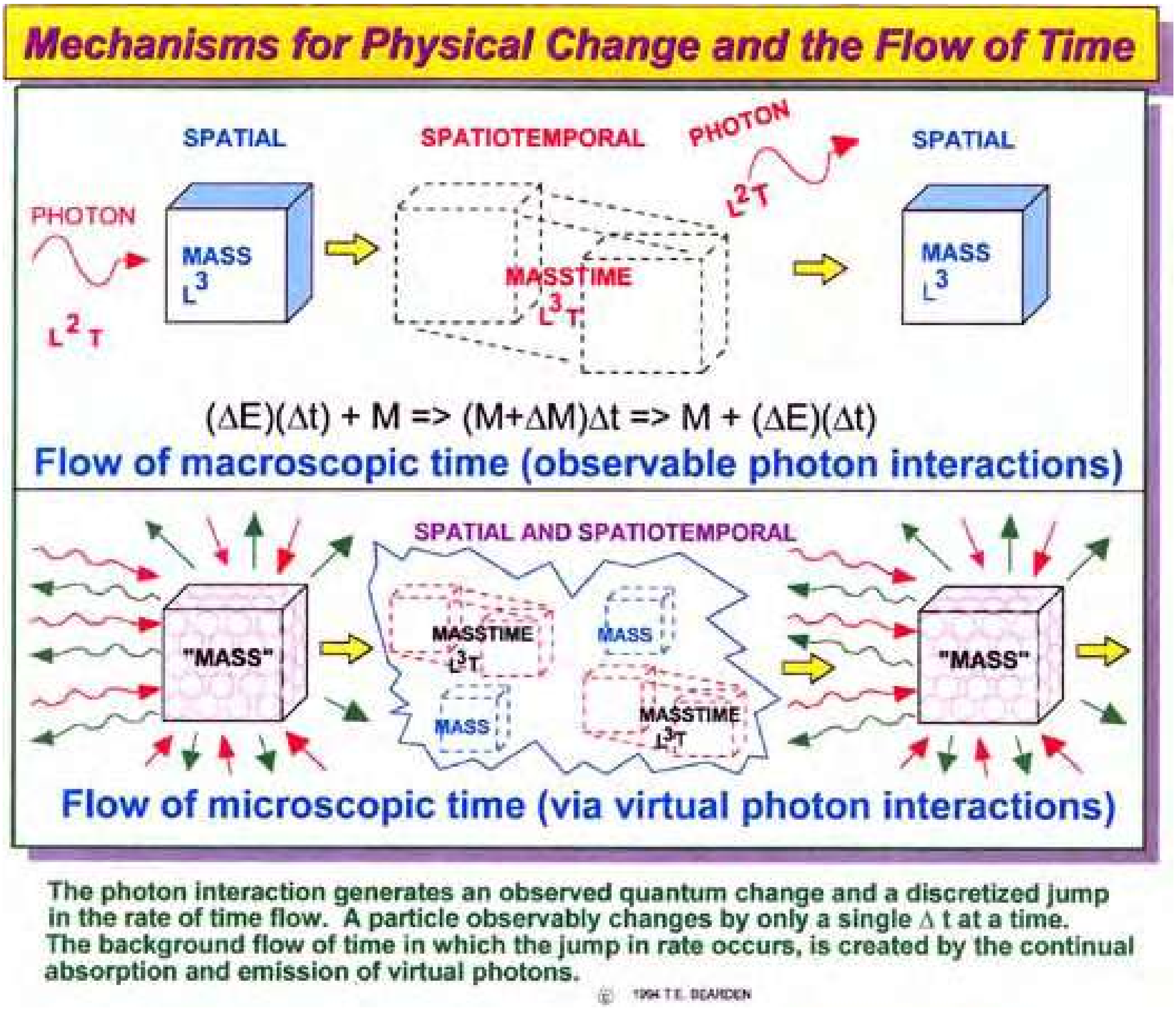

Mechanisms for Physical Change and the Flow of Time

SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS
L^3
L^2T
MASSTIME
L^3T
L^2T
MASS
L^3

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"

Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δ t at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

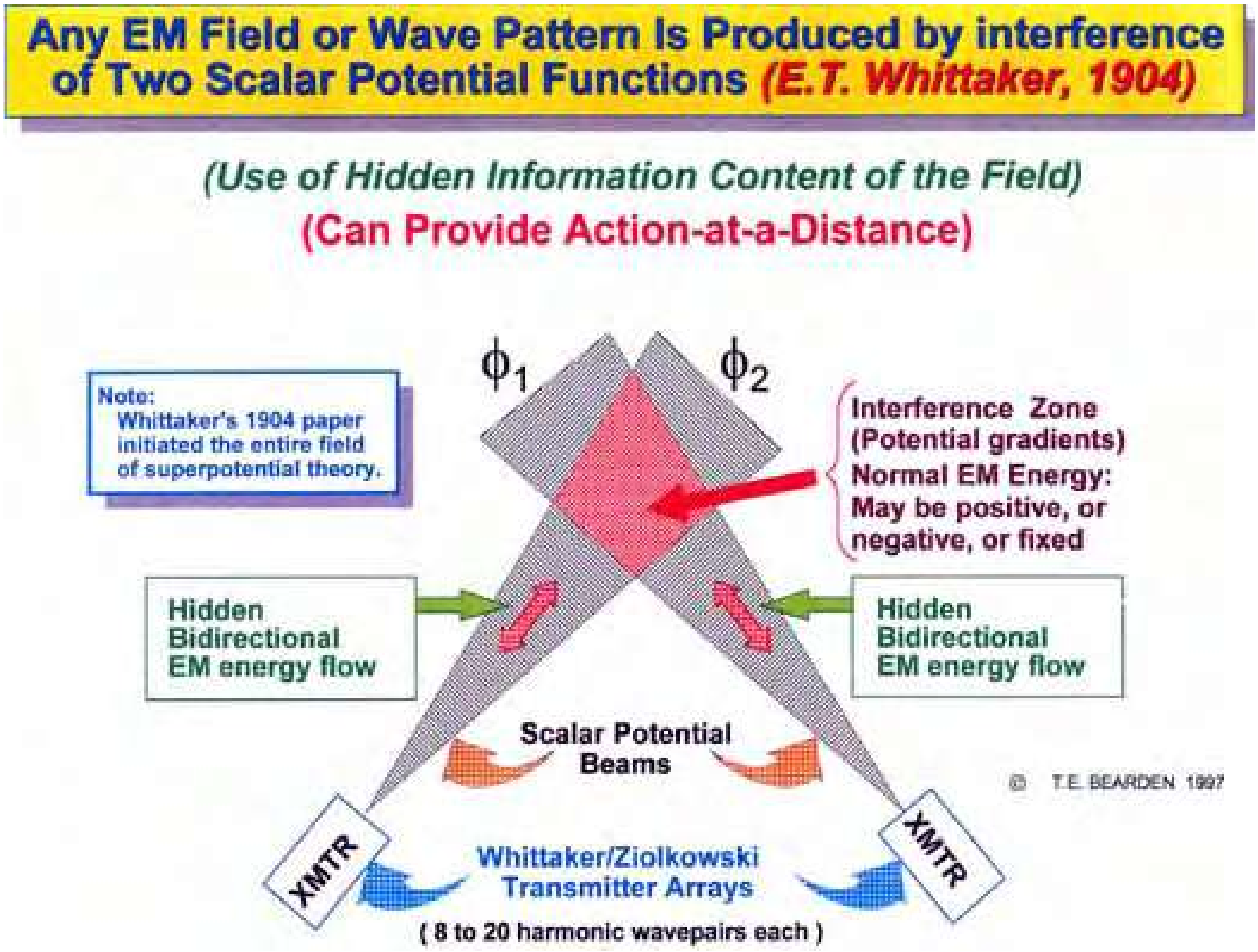

Any EM Field or Wave Pattern Is Produced by interference of Two Scalar Potential Functions *(E.T. Whittaker, 1904)*

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.

Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or negative, or fixed

Hidden Bidirectional EM energy flow

Hidden Bidirectional EM energy flow

Scalar Potential Beams

Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

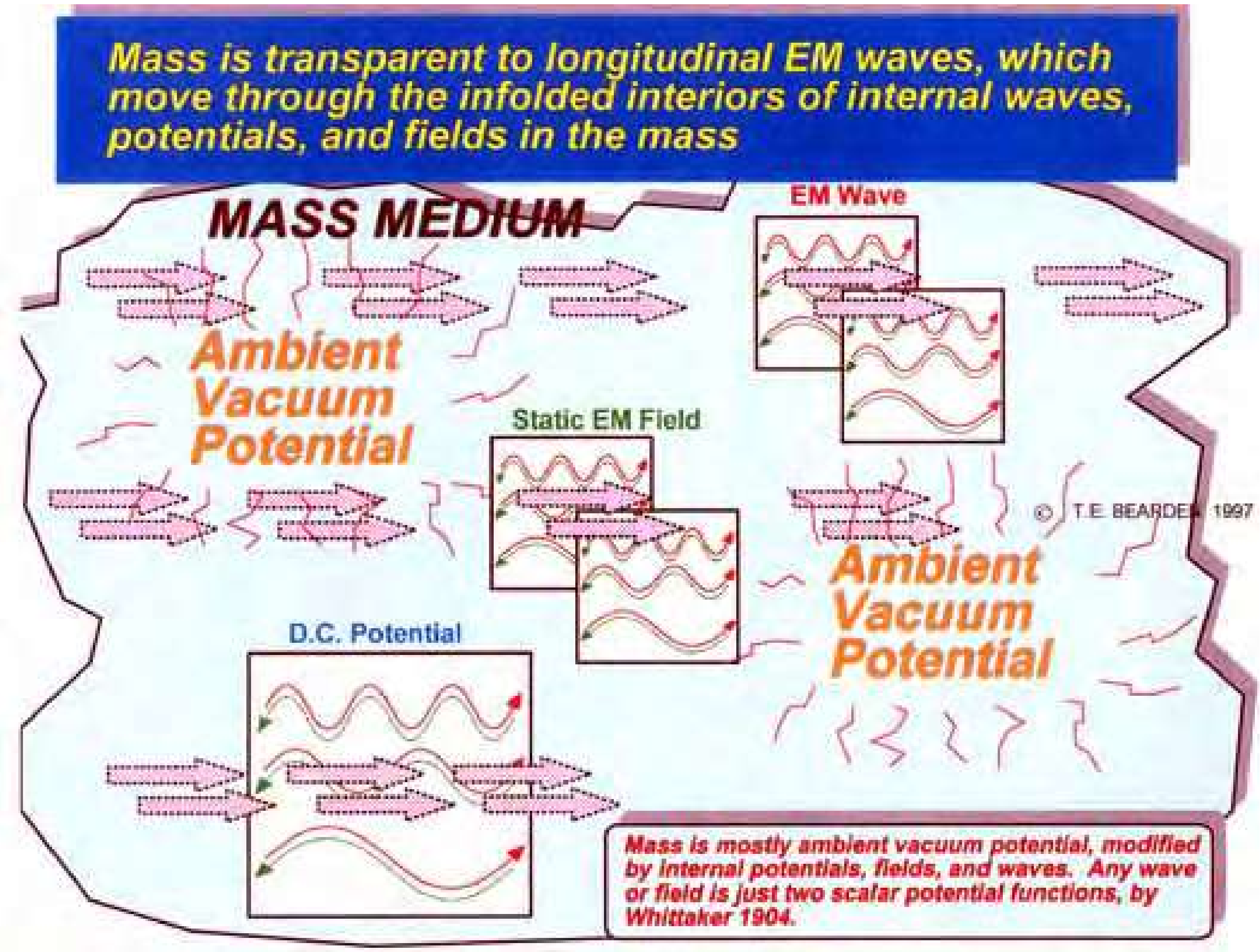

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
Ambient Vacuum Potential
EM Wave
Static EM Field
Ambient Vacuum Potential
D.C. Potential

© T.E. BEARDEN 1997

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

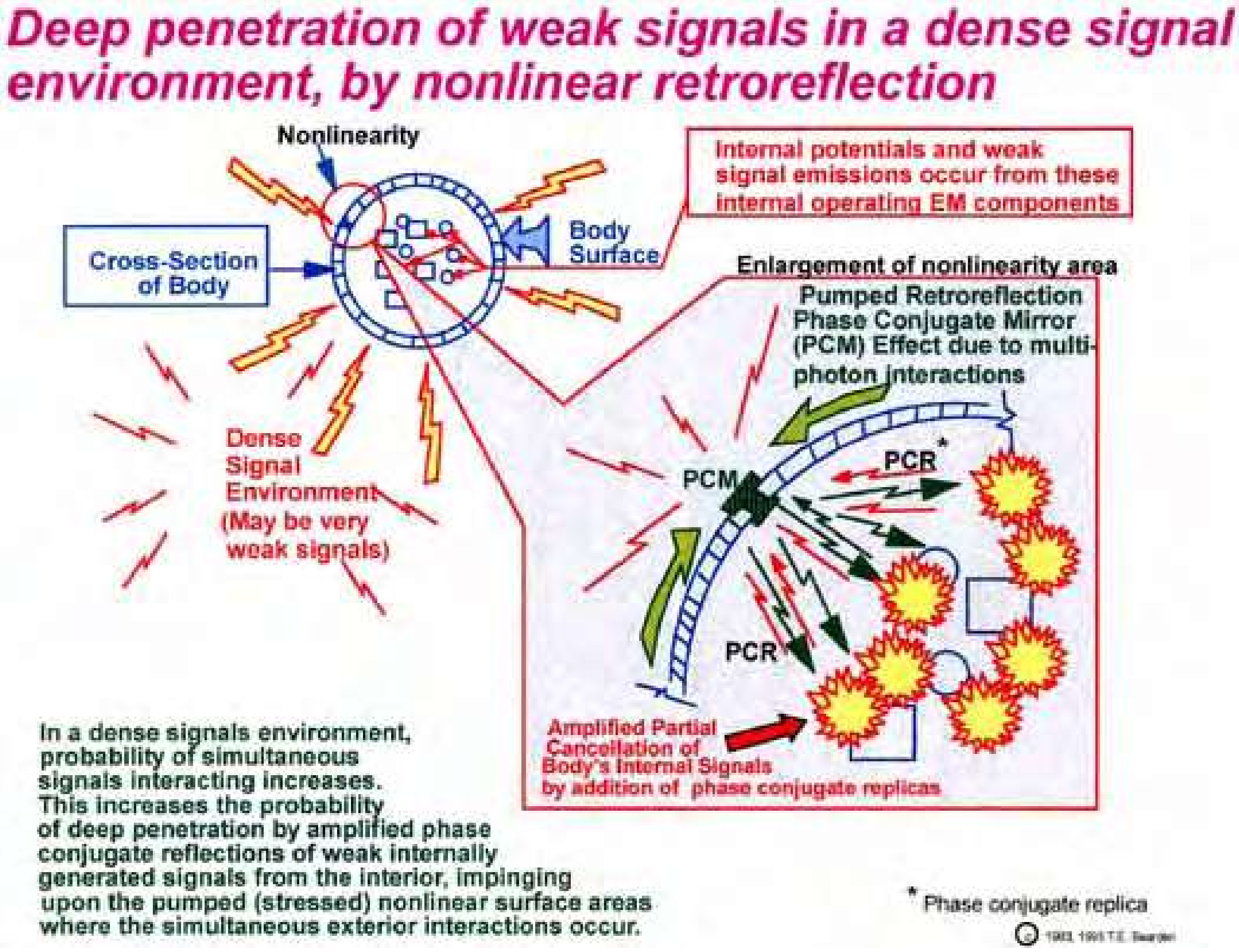

Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection

Internal potentials and weak signal emissions occur from these internal operating EM components

Dense Signal Environment (May be very weak signals)

Pumped Retroreflection Phase Conjugate Mirror (PCM) Effect due to multi-photon interactions

Amplified Partial Cancellation of Body's Internal Signals by addition of phase conjugate replicas

In a dense signals environment, probability of simultaneous signals interacting increases. This increases the probability of deep penetration by amplified phase conjugate reflections of weak internally generated signals from the interior, impinging upon the pumped (stressed) nonlinear surface areas where the simultaneous exterior interactions occur.

* Phase conjugate replica

© 1993, 1995 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

THE IMMUNE SYSTEM

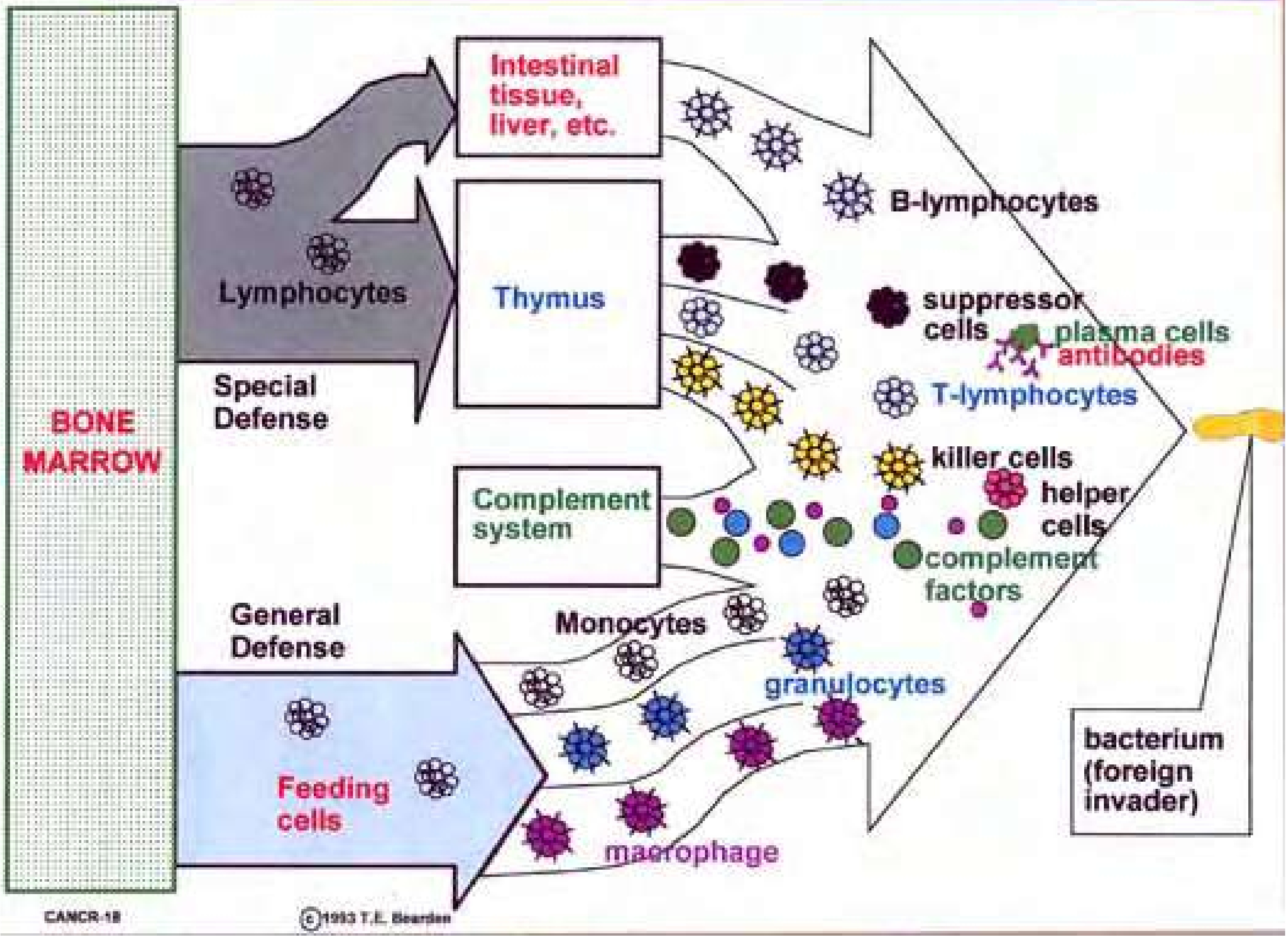

BONE MARROW

Lymphocytes

Special Defense

Intestinal tissue, liver, etc.

Thymus

Complement system

General Defense

Feeding cells

Monocytes

B-lymphocytes

suppressor cells

plasma cells

antibodies

T-lymphocytes

killer cells

helper cells

complement factors

granulocytes

macrophage

bacterium (foreign invader)

CANCR-18

©1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

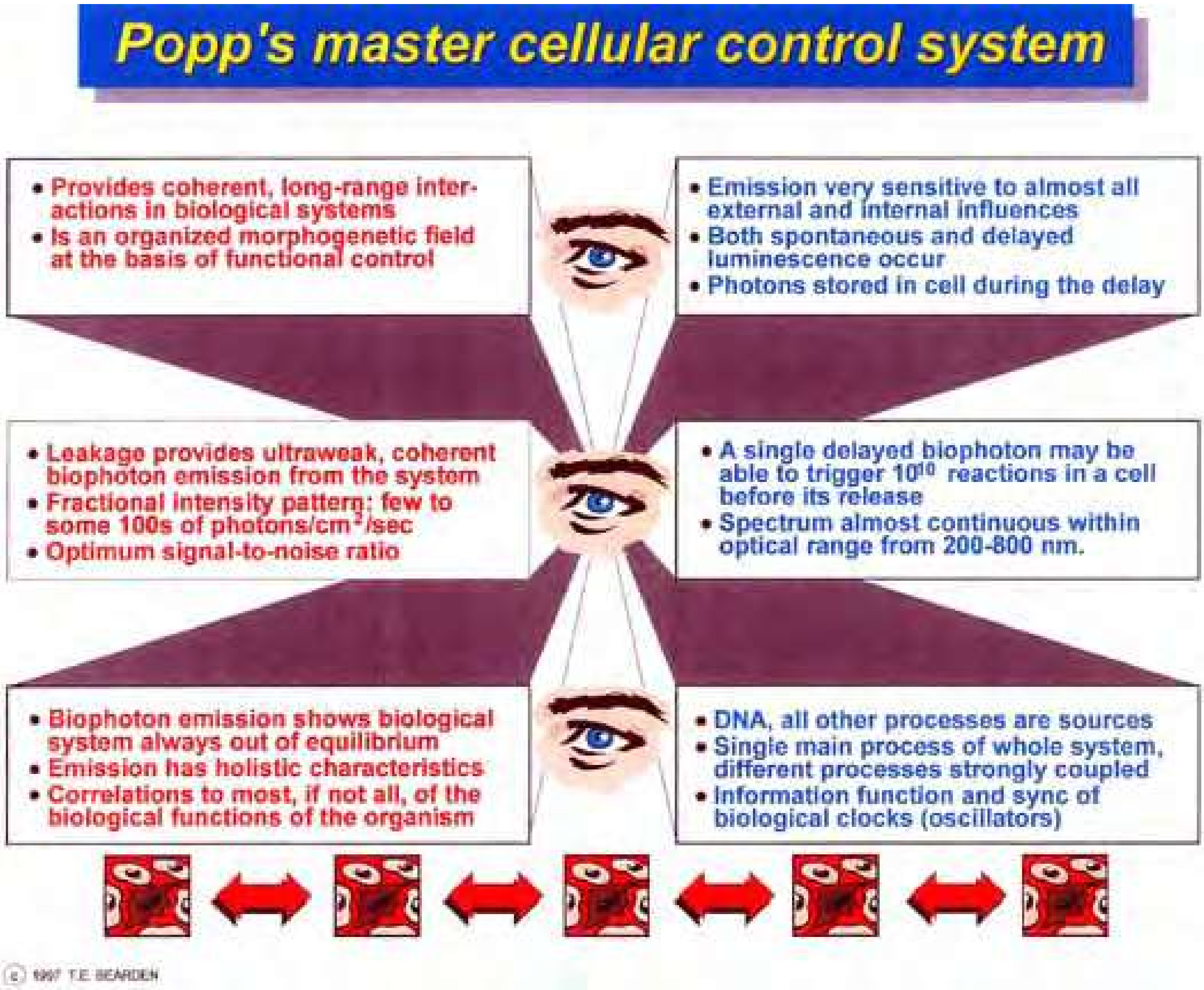

Popp's master cellular control system

- Provides coherent, long-range interactions in biological systems
- Is an organized morphogenetic field at the basis of functional control

- Emission very sensitive to almost all external and internal influences
- Both spontaneous and delayed luminescence occur
- Photons stored in cell during the delay

- Leakage provides ultraweak, coherent biophoton emission from the system
- Fractional intensity pattern: few to some 100s of photons/cm^2/sec
- Optimum signal-to-noise ratio

- A single delayed biophoton may be able to trigger 10^{10} reactions in a cell before its release
- Spectrum almost continuous within optical range from 200-800 nm.

- Biophoton emission shows biological system always out of equilibrium
- Emission has holistic characteristics
- Correlations to most, if not all, of the biological functions of the organism

- DNA, all other processes are sources
- Single main process of whole system, different processes strongly coupled
- Information function and sync of biological clocks (oscillators)

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

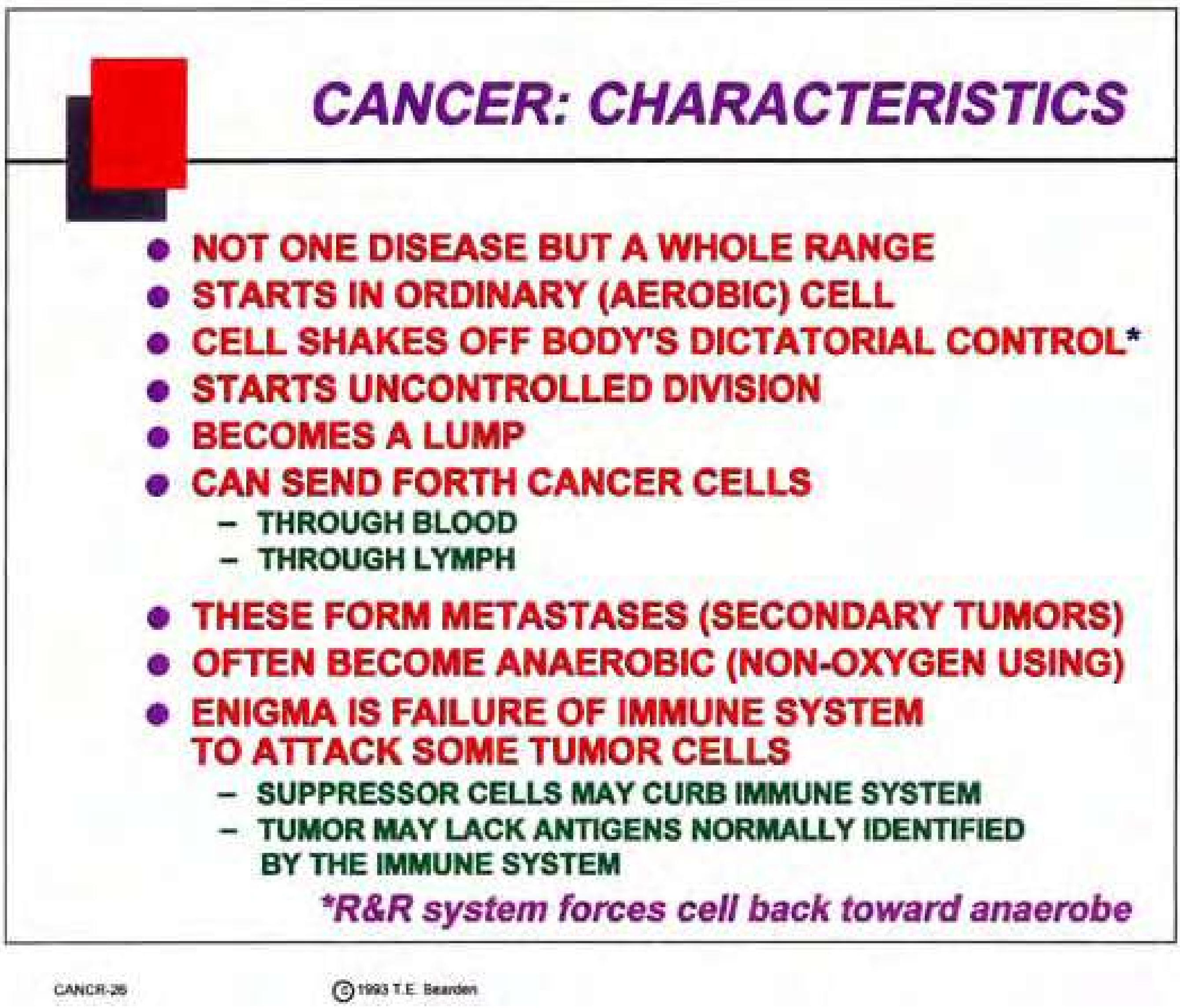

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

**R&R system forces cell back toward anaerobe*

CANCR-26

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

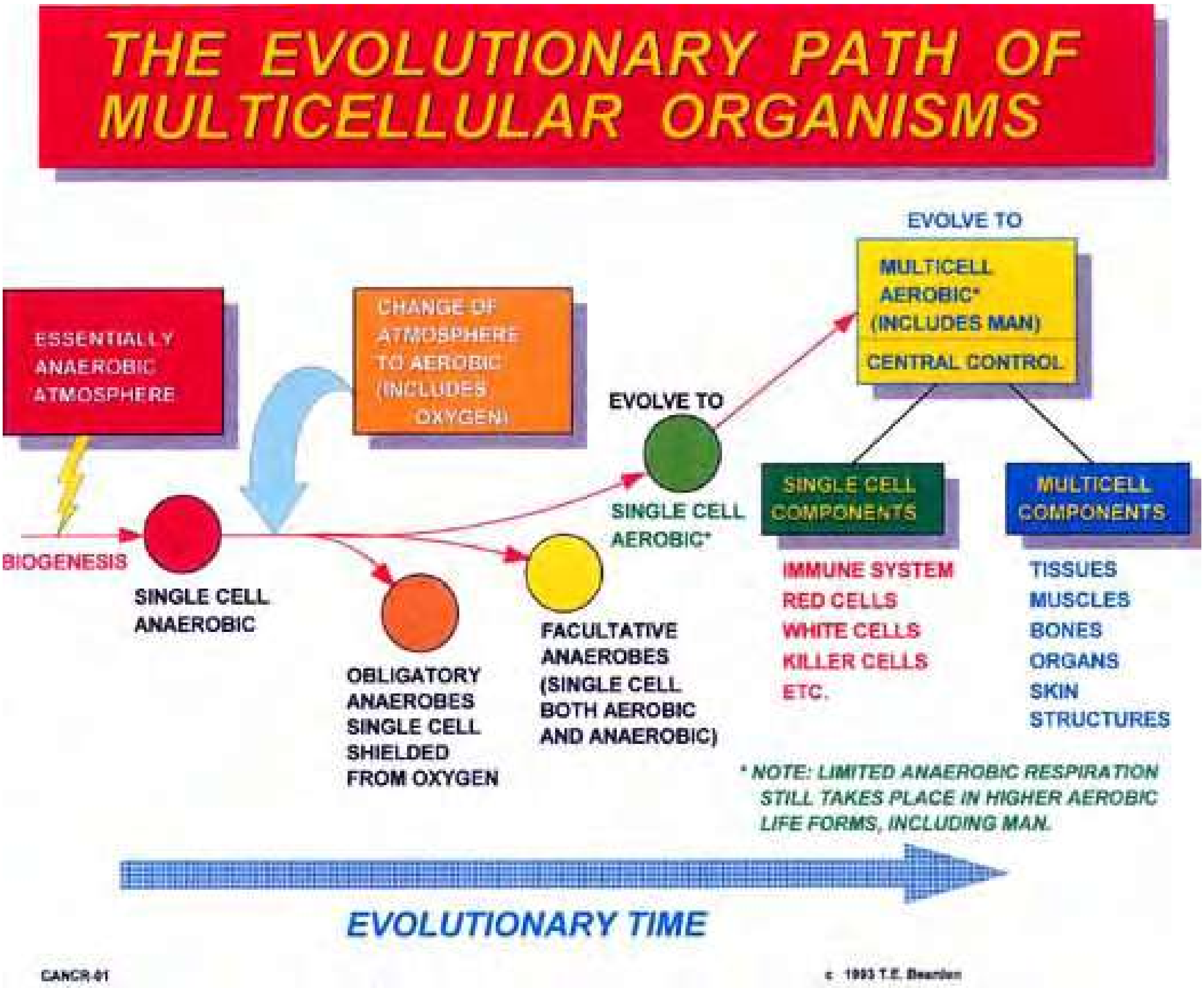

THE EVOLUTIONARY PATH OF MULTICELLULAR ORGANISMS

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

CANCR-01

c 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

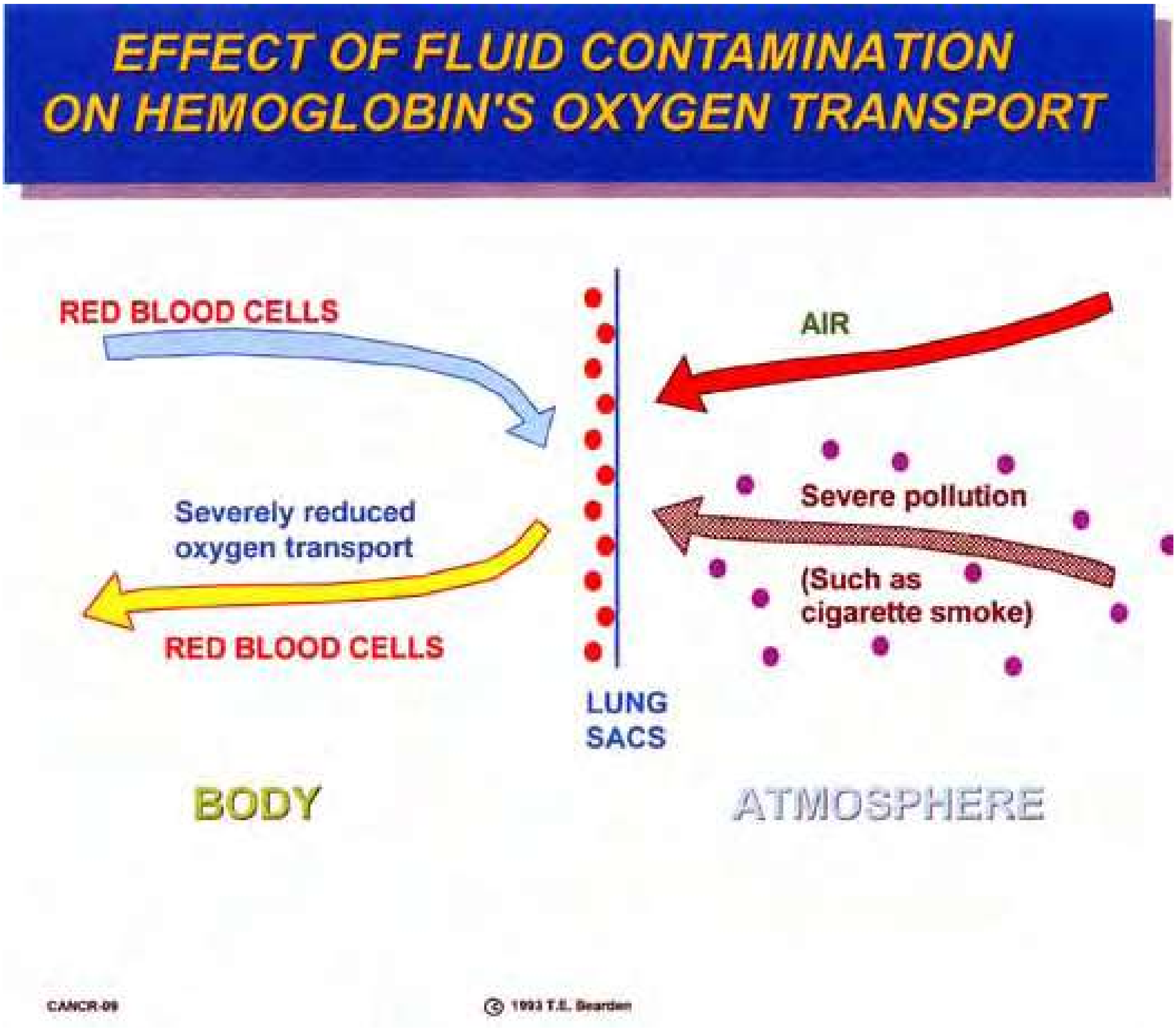

EFFECT OF FLUID CONTAMINATION ON HEMOGLOBIN'S OXYGEN TRANSPORT

RED BLOOD CELLS

AIR

Severe pollution

(Such as cigarette smoke)

Severely reduced oxygen transport

RED BLOOD CELLS

LUNG SACS

BODY

ATMOSPHERE

CANCR-09

© 1993 T.E. Bearden

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

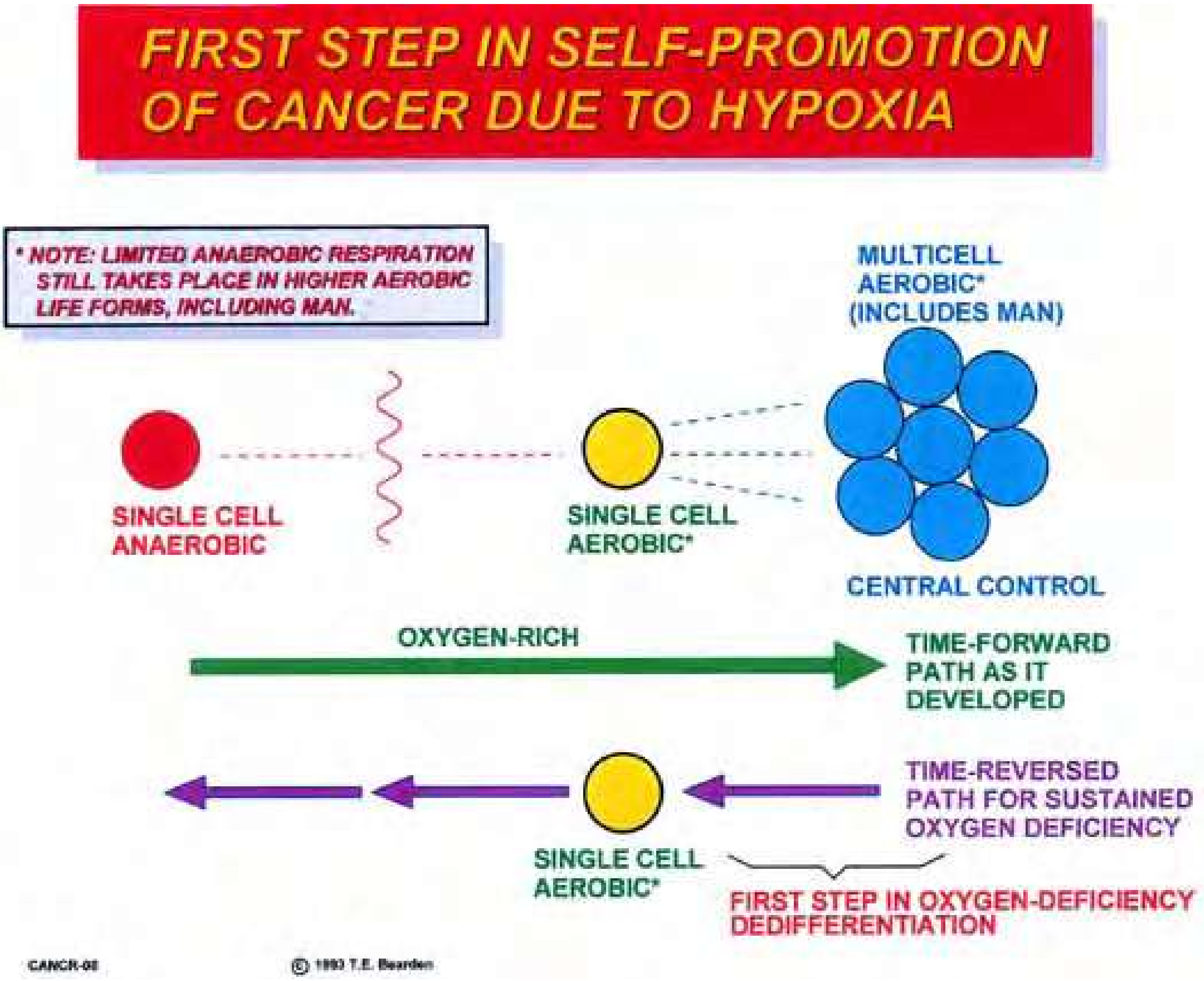

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

SINGLE CELL
ANAEROBIC
SINGLE CELL
AEROBIC*
MULTICELL
AEROBIC*
(INCLUDES MAN)
CENTRAL CONTROL
OXYGEN-RICH
TIME-FORWARD
PATH AS IT
DEVELOPED
TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY
SINGLE CELL
AEROBIC*

FIRST STEP IN OXYGEN-DEFICIENCY DEDIFFERENTIATION

CANCR-08

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

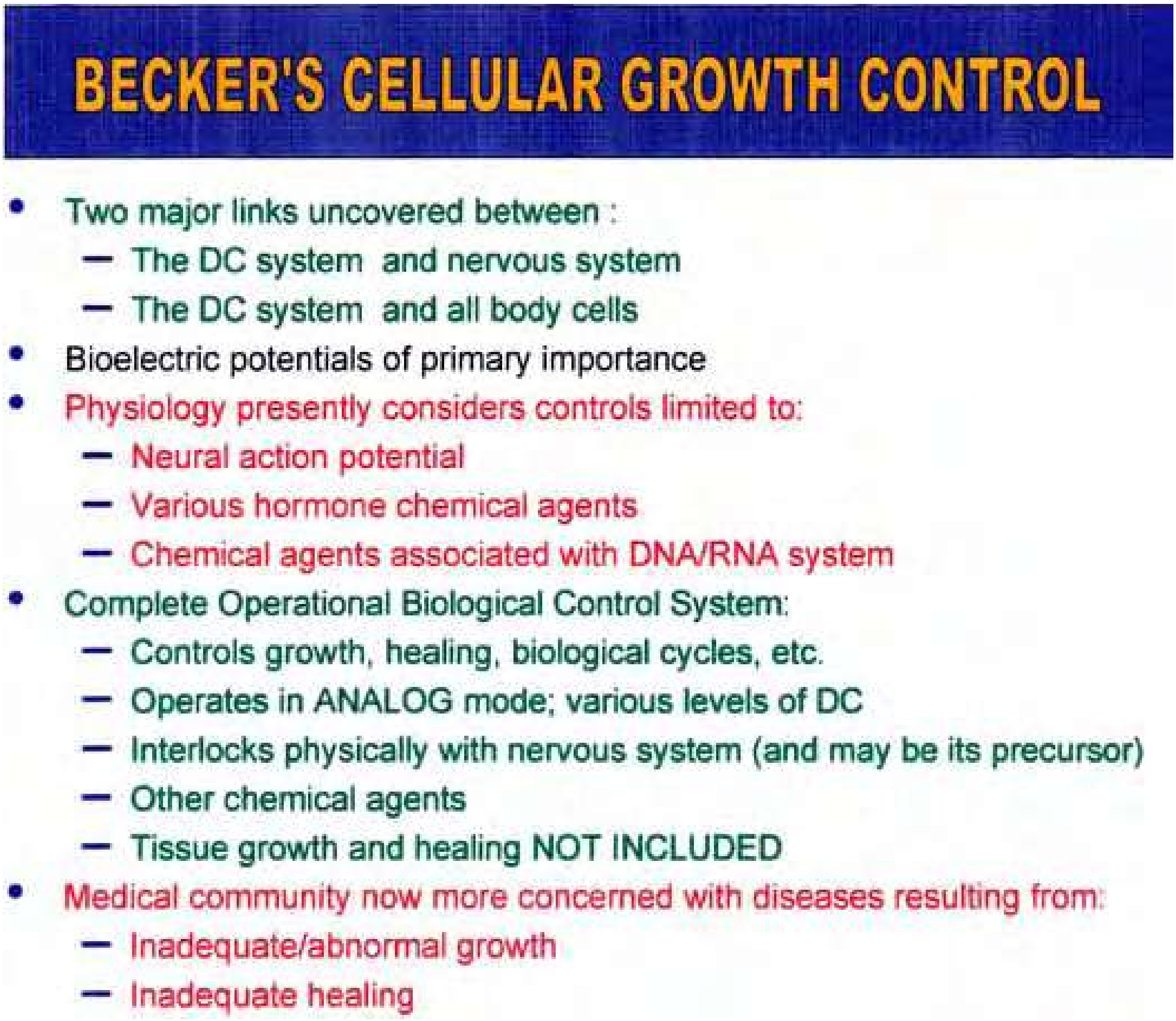

BECKER'S CELLULAR GROWTH CONTROL

- Two major links uncovered between :
  - The DC system and nervous system
  - The DC system and all body cells
- Bioelectric potentials of primary importance
- Physiology presently considers controls limited to:
  - Neural action potential
  - Various hormone chemical agents
  - Chemical agents associated with DNA/RNA system
- Complete Operational Biological Control System:
  - Controls growth, healing, biological cycles, etc.
  - Operates in ANALOG mode; various levels of DC
  - Interlocks physically with nervous system (and may be its precursor)
  - Other chemical agents
  - Tissue growth and healing NOT INCLUDED
- Medical community now more concerned with diseases resulting from:
  - Inadequate/abnormal growth
  - Inadequate healing

Return to Slide Index

Next Slide

Previous Slide

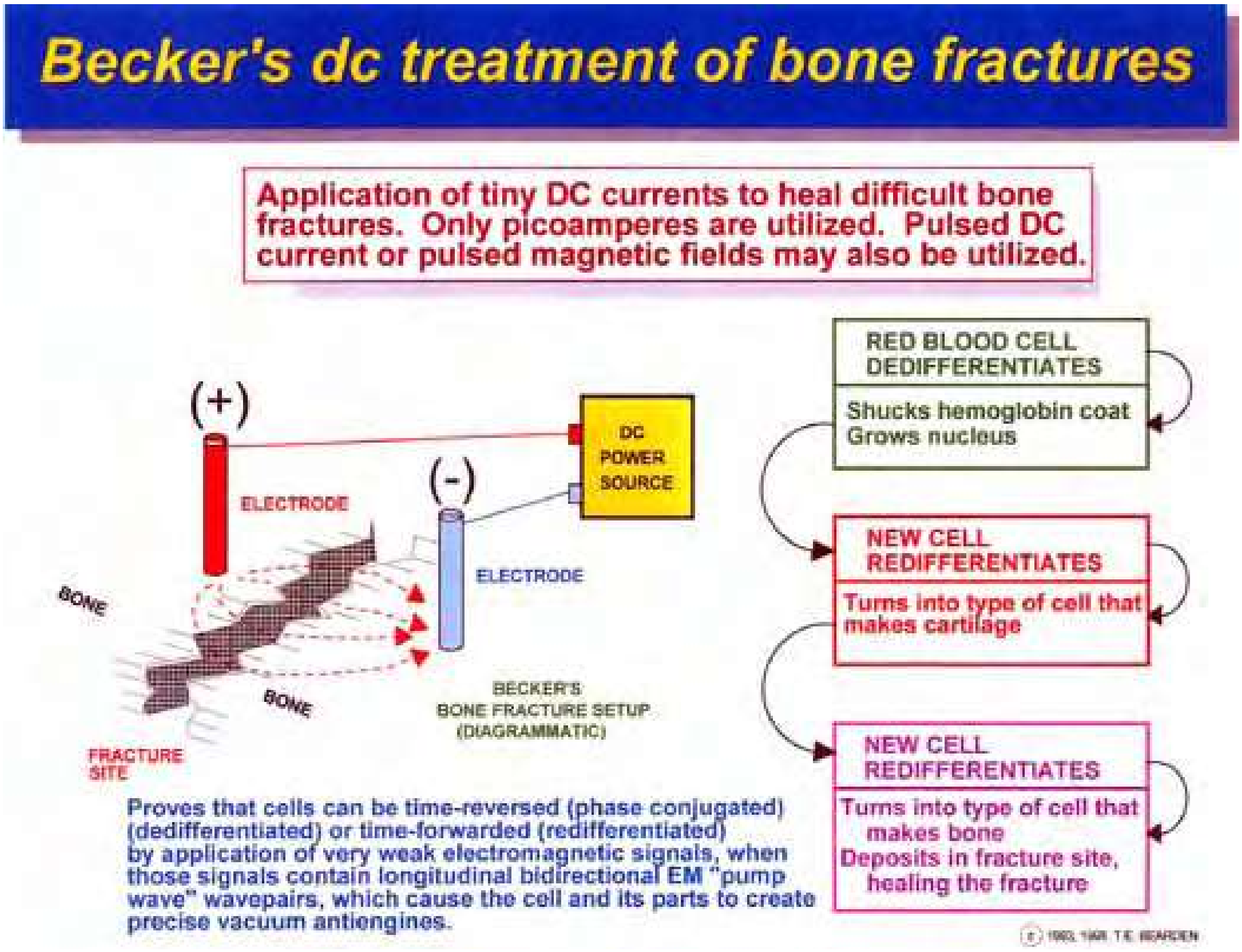

Becker's dc treatment of bone fractures

Application of tiny DC currents to heal difficult bone fractures. Only picoamperes are utilized. Pulsed DC current or pulsed magnetic fields may also be utilized.

(+)
ELECTRODE
(-)
DC POWER SOURCE
ELECTRODE
BONE
BONE
FRACTURE SITE

BECKER'S
BONE FRACTURE SETUP
(DIAGRAMMATIC)

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Proves that cells can be time-reversed (phase conjugated) (dedifferentiated) or time-forwarded (redifferentiated) by application of very weak electromagnetic signals, when those signals contain longitudinal bidirectional EM "pump wave" wavepairs, which cause the cell and its parts to create precise vacuum antiengines.

© [illegible] T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

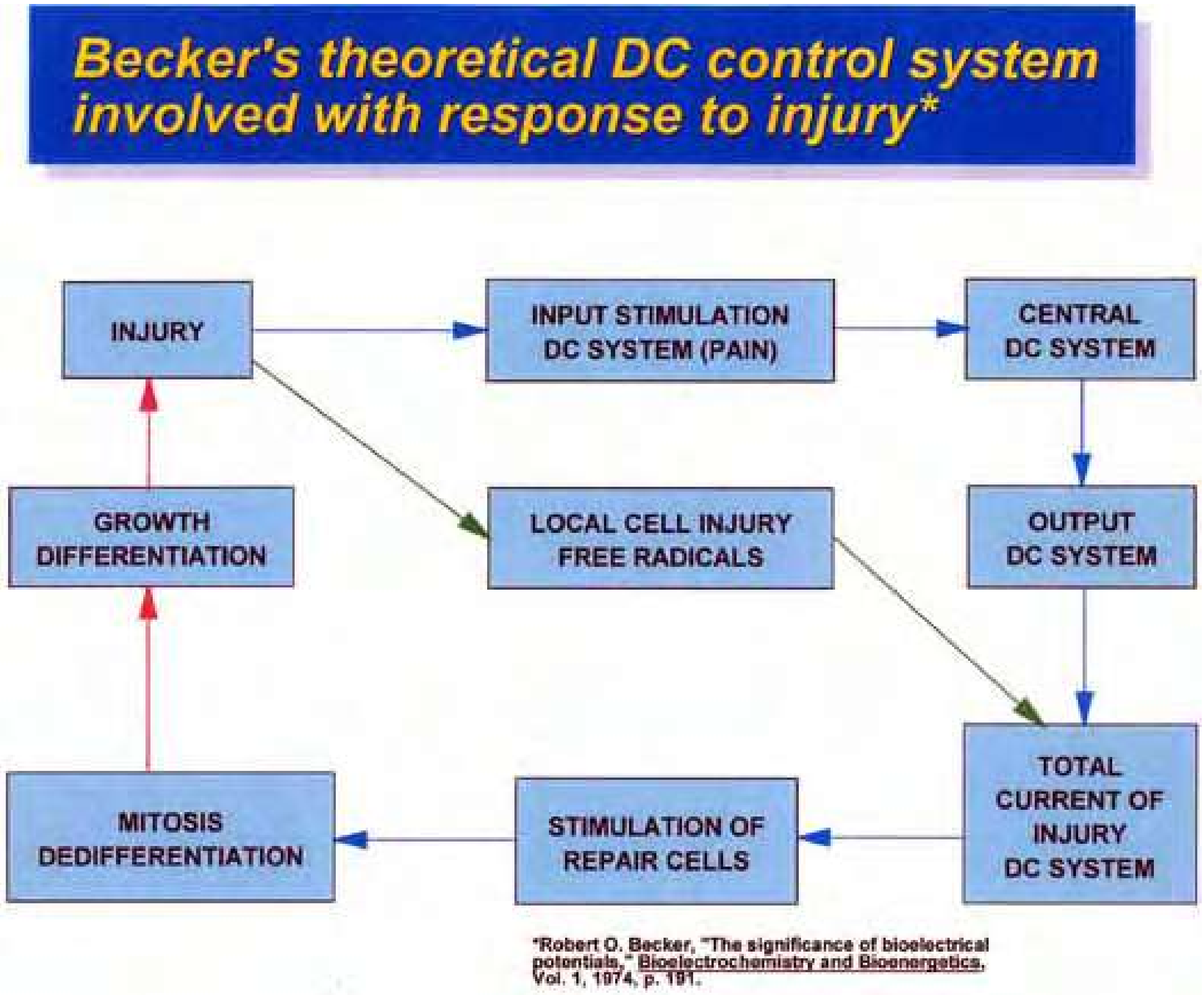

*Becker's theoretical DC control system involved with response to injury**

INJURY
INPUT STIMULATION DC SYSTEM (PAIN)
CENTRAL DC SYSTEM
GROWTH DIFFERENTIATION
LOCAL CELL INJURY FREE RADICALS
OUTPUT DC SYSTEM
TOTAL CURRENT OF INJURY DC SYSTEM
MITOSIS DEDIFFERENTIATION
STIMULATION OF REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Return to Slide Index

Next Slide

Previous Slide

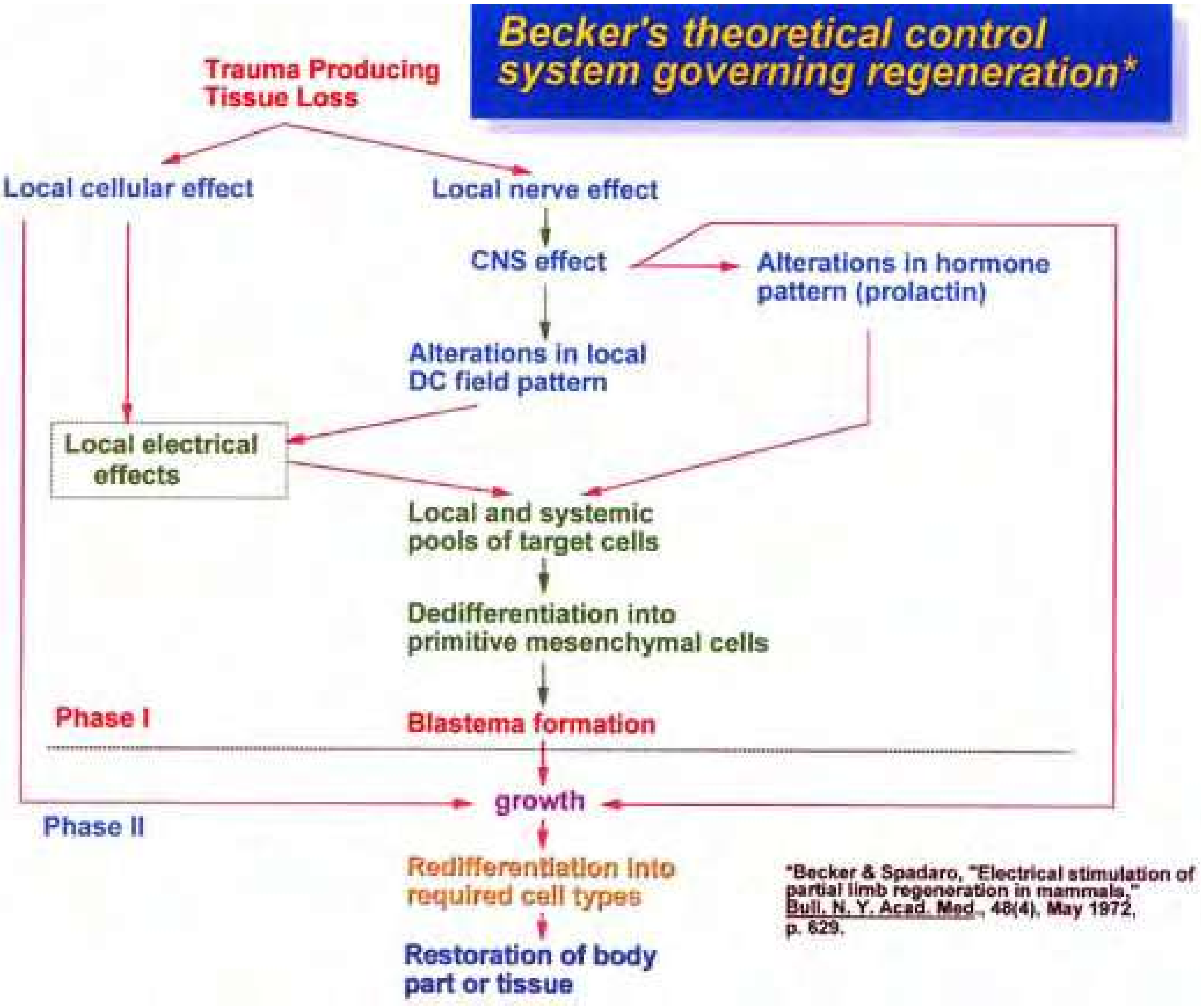

*Becker's theoretical control system governing regeneration**

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

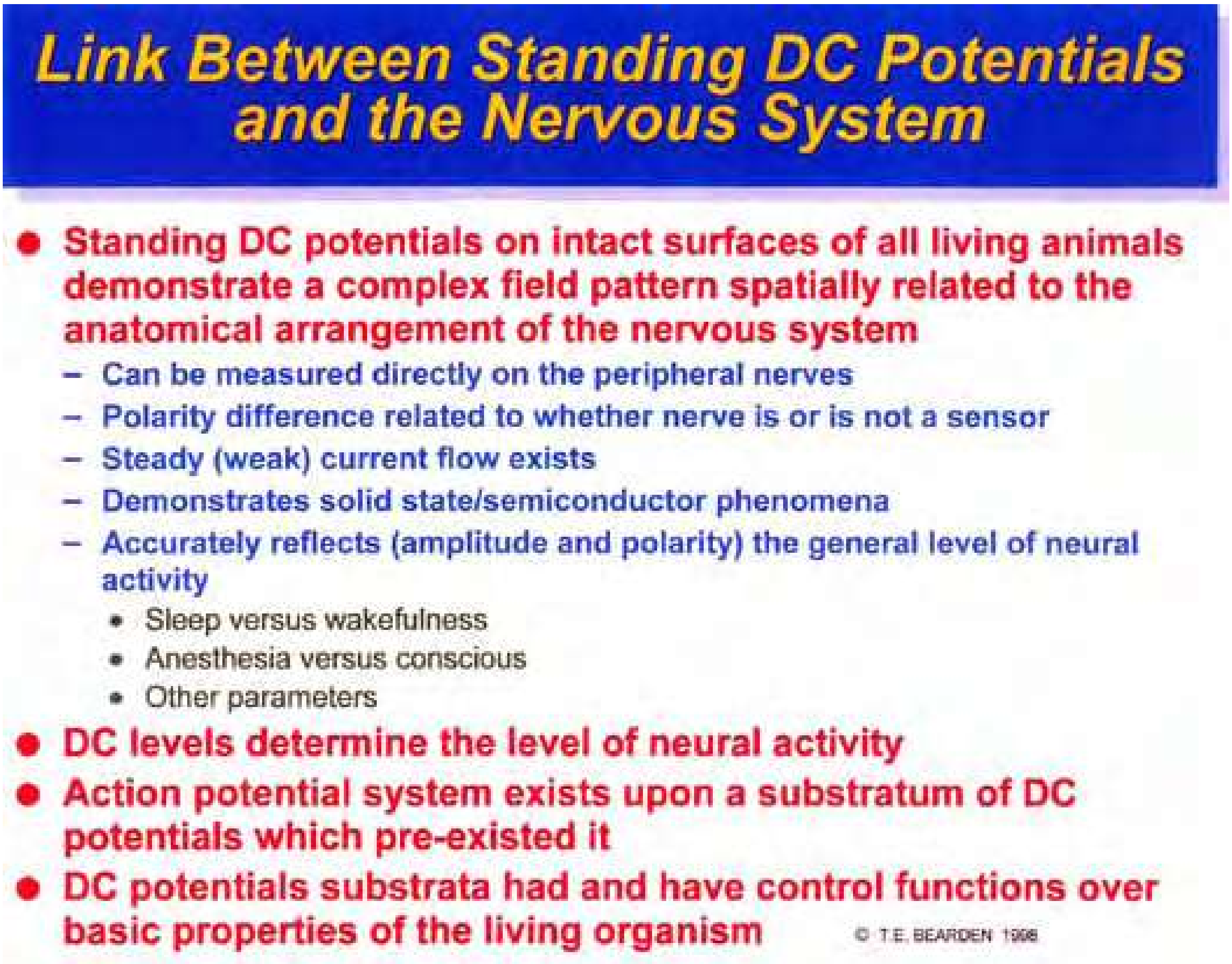

Link Between Standing DC Potentials and the Nervous System

- Standing DC potentials on intact surfaces of all living animals demonstrate a complex field pattern spatially related to the anatomical arrangement of the nervous system
  - Can be measured directly on the peripheral nerves
  - Polarity difference related to whether nerve is or is not a sensor
  - Steady (weak) current flow exists
  - Demonstrates solid state/semiconductor phenomena
  - Accurately reflects (amplitude and polarity) the general level of neural activity
    - Sleep versus wakefulness
    - Anesthesia versus conscious
    - Other parameters
- DC levels determine the level of neural activity
- Action potential system exists upon a substratum of DC potentials which pre-existed it
- DC potentials substrata had and have control functions over basic properties of the living organism

© T.E. BEARDEN 1998

[Return to Slide Index]

[Next Slide]

[Previous Slide]

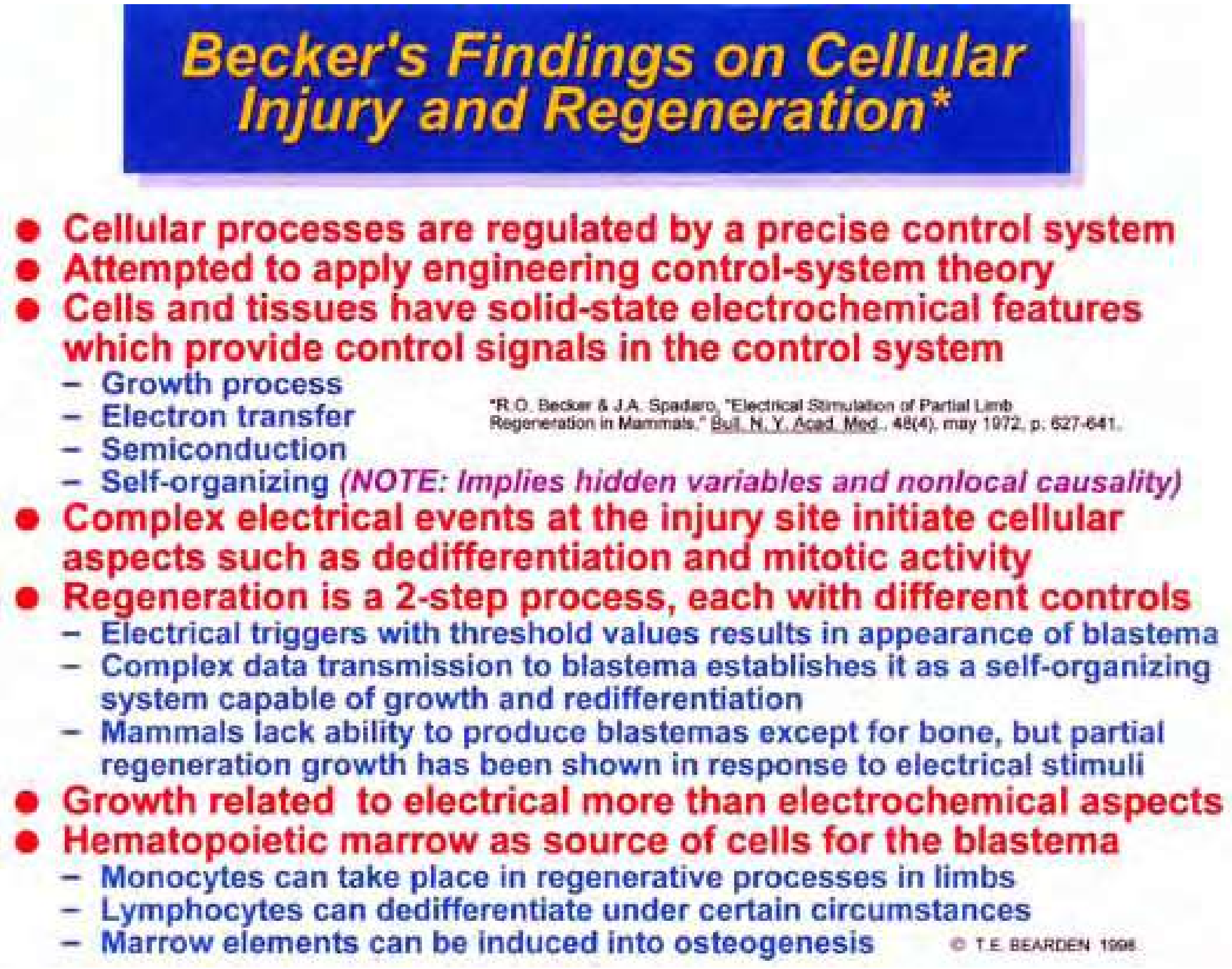

Becker's Findings on Cellular Injury and Regeneration*

- Cellular processes are regulated by a precise control system
- Attempted to apply engineering control-system theory
- Cells and tissues have solid-state electrochemical features which provide control signals in the control system
  - Growth process
  - Electron transfer
  - Semiconduction
  - Self-organizing *(NOTE: Implies hidden variables and nonlocal causality)*
- Complex electrical events at the injury site initiate cellular aspects such as dedifferentiation and mitotic activity
- Regeneration is a 2-step process, each with different controls
  - Electrical triggers with threshold values results in appearance of blastema
  - Complex data transmission to blastema establishes it as a self-organizing system capable of growth and redifferentiation
  - Mammals lack ability to produce blastemas except for bone, but partial regeneration growth has been shown in response to electrical stimuli
- Growth related to electrical more than electrochemical aspects
- Hematopoietic marrow as source of cells for the blastema
  - Monocytes can take place in regenerative processes in limbs
  - Lymphocytes can dedifferentiate under certain circumstances
  - Marrow elements can be induced into osteogenesis

*R.O. Becker & J.A. Spadaro, "Electrical Stimulation of Partial Limb Regeneration in Mammals," Bull. N. Y. Acad. Med., 48(4), may 1972, p. 627-641.

© T.E. BEARDEN 1998

[Return to Slide Index]

[Next Slide]

[Previous Slide]

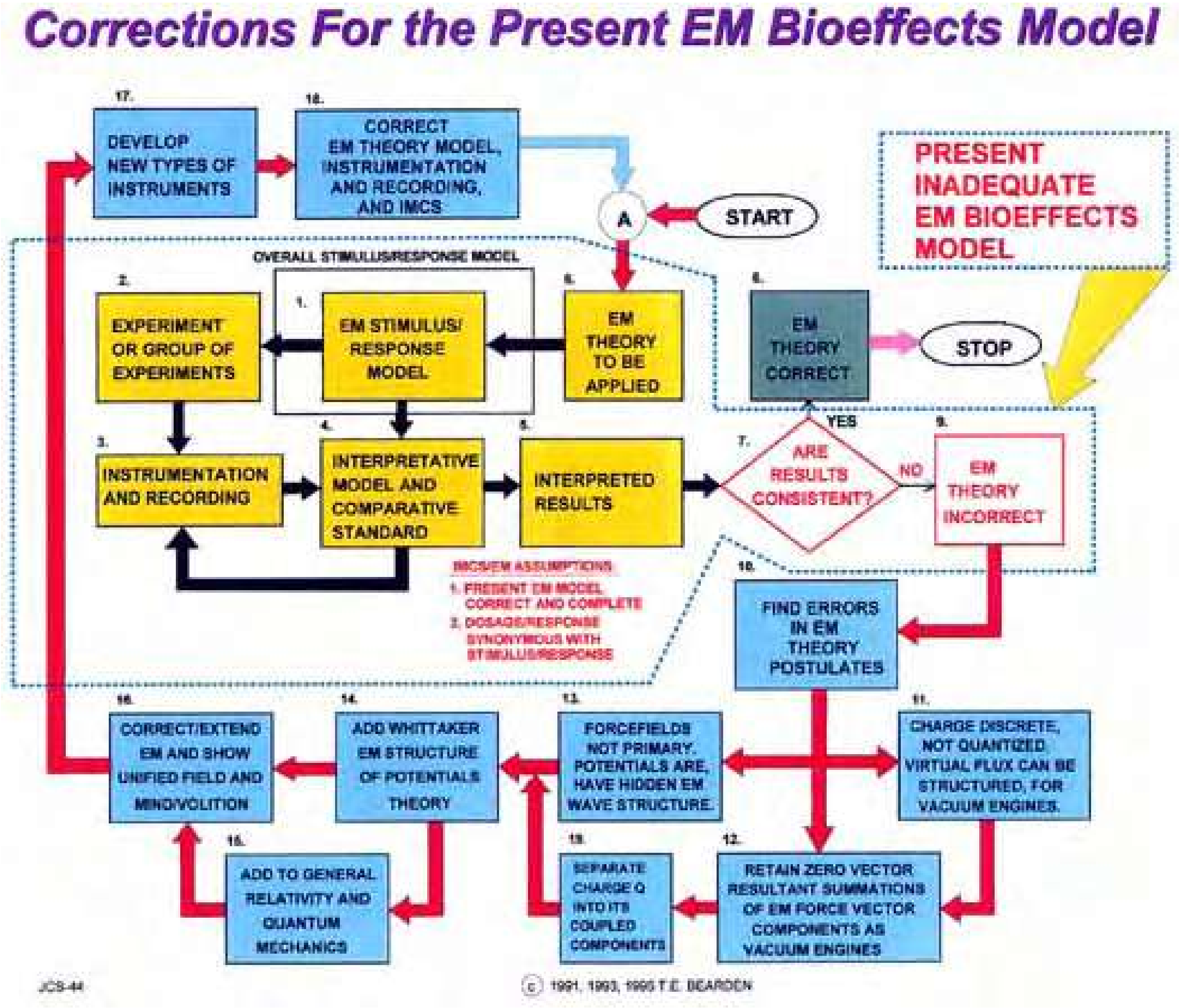

Corrections For the Present EM Bioeffects Model

17. DEVELOP NEW TYPES OF INSTRUMENTS

18. CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS

A

START

PRESENT INADEQUATE EM BIOEFFECTS MODEL

OVERALL STIMULUS/RESPONSE MODEL

2. EXPERIMENT OR GROUP OF EXPERIMENTS

1. EM STIMULUS/ RESPONSE MODEL

6. EM THEORY TO BE APPLIED

8. EM THEORY CORRECT

STOP

YES

3. INSTRUMENTATION AND RECORDING

4. INTERPRETATIVE MODEL AND COMPARATIVE STANDARD

5. INTERPRETED RESULTS

7. ARE RESULTS CONSISTENT?

NO

9. EM THEORY INCORRECT

IMCS/EM ASSUMPTIONS
1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE

10. FIND ERRORS IN EM THEORY POSTULATES

16. CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION

14. ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY

13. FORCEFIELDS NOT PRIMARY. POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.

11. CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.

15. ADD TO GENERAL RELATIVITY AND QUANTUM MECHANICS

19. SEPARATE CHARGE Q INTO ITS COUPLED COMPONENTS

12. RETAIN ZERO VECTOR RESULTANT SUMMATIONS OF EM FORCE VECTOR COMPONENTS AS VACUUM ENGINES

JCS-44

© 1991, 1993, 1995 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

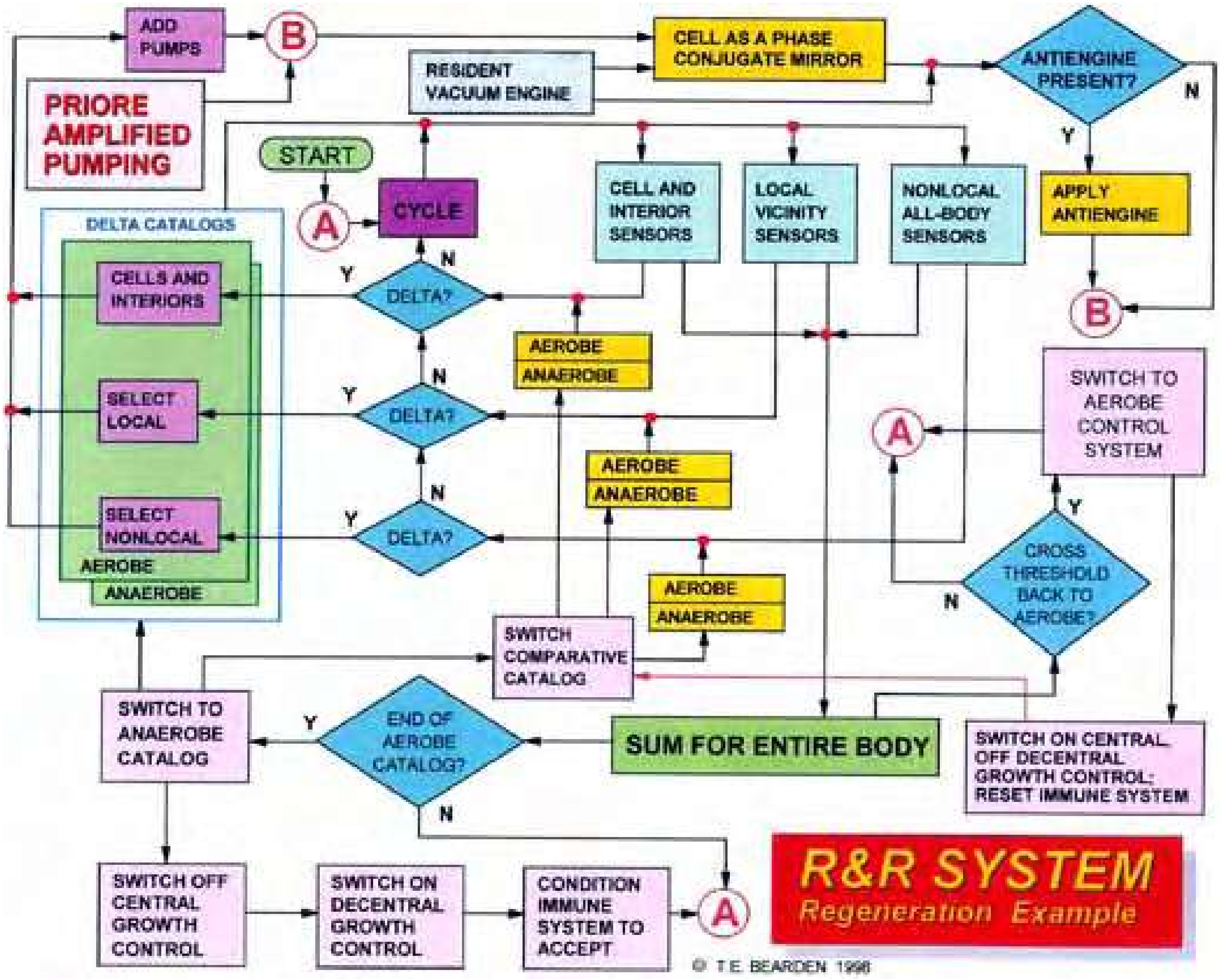
ADD PUMPS
B
CELL AS A PHASE CONJUGATE MIRROR
RESIDENT VACUUM ENGINE
ANTIENGINE PRESENT?
N
Y
PRIORE AMPLIFIED PUMPING
START
A
CYCLE
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NONLOCAL ALL-BODY SENSORS
APPLY ANTIENGINE
DELTA CATALOGS
CELLS AND INTERIORS
Y
DELTA?
N
B
AEROBE
ANAEROBE
SELECT LOCAL
Y
DELTA?
N
SWITCH TO AEROBE CONTROL SYSTEM
A
AEROBE
ANAEROBE
SELECT NONLOCAL
Y
DELTA?
AEROBE
ANAEROBE
Y
CROSS THRESHOLD BACK TO AEROBE?
N
AEROBE
ANAEROBE
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
Y
END OF AEROBE CATALOG?
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM
N
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT
A

R&R SYSTEM
Regeneration Example

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

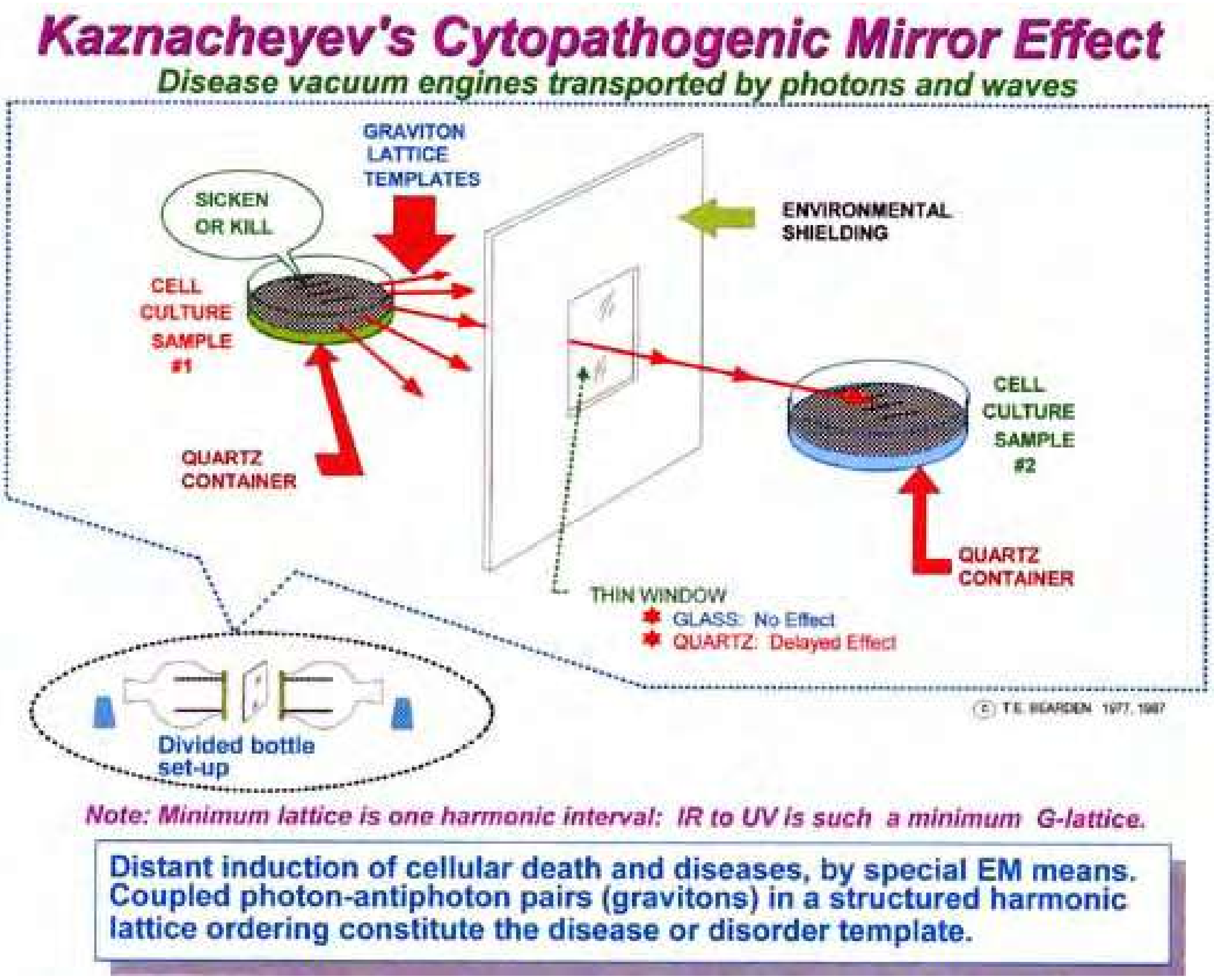

Kaznacheyev's Cytopathogenic Mirror Effect

Disease vacuum engines transported by photons and waves

GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
CELL
CULTURE
SAMPLE
#1
QUARTZ
CONTAINER
ENVIRONMENTAL
SHIELDING
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
✱ GLASS: No Effect
✱ QUARTZ: Delayed Effect
Divided bottle
set-up

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Distant induction of cellular death and diseases, by special EM means. Coupled photon-antiphoton pairs (gravitons) in a structured harmonic lattice ordering constitute the disease or disorder template.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

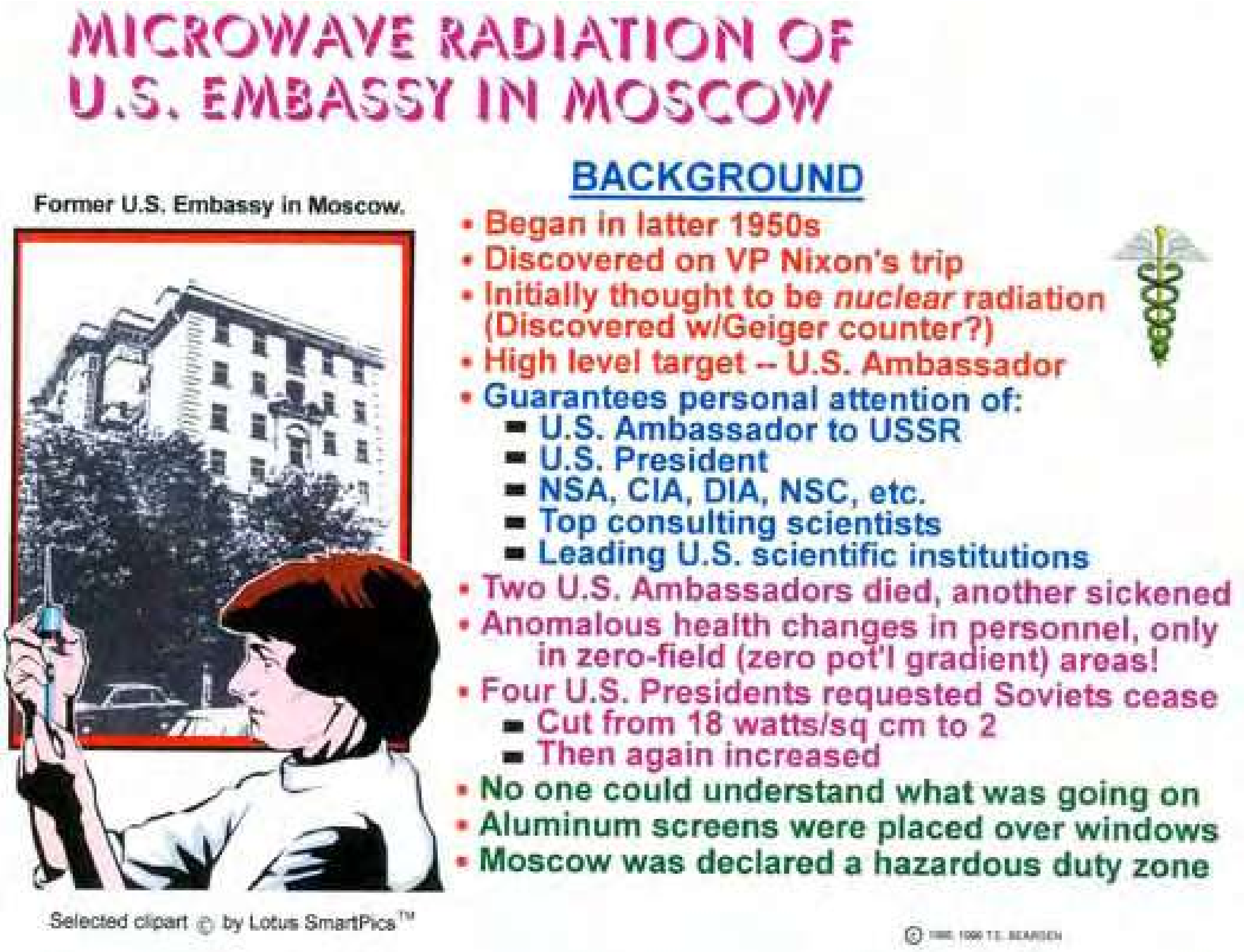

MICROWAVE RADIATION OF U.S. EMBASSY IN MOSCOW

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Former U.S. Embassy in Moscow.

Selected clipart © by Lotus SmartPics™

© [illegible] 1990 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

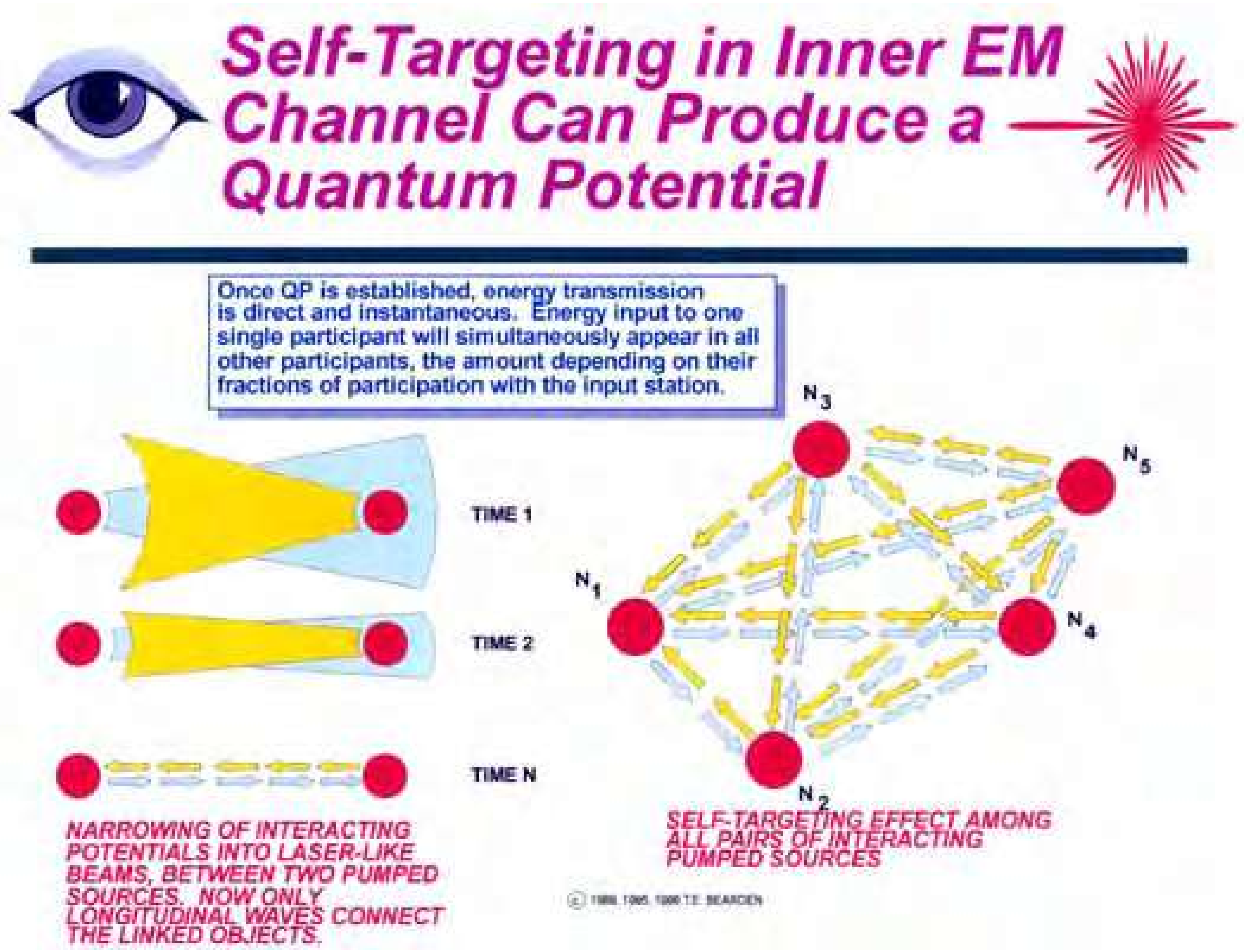

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

TIME 1
TIME 2
TIME N
N_3
N_5
N_1
N_4
N_2

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1986, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

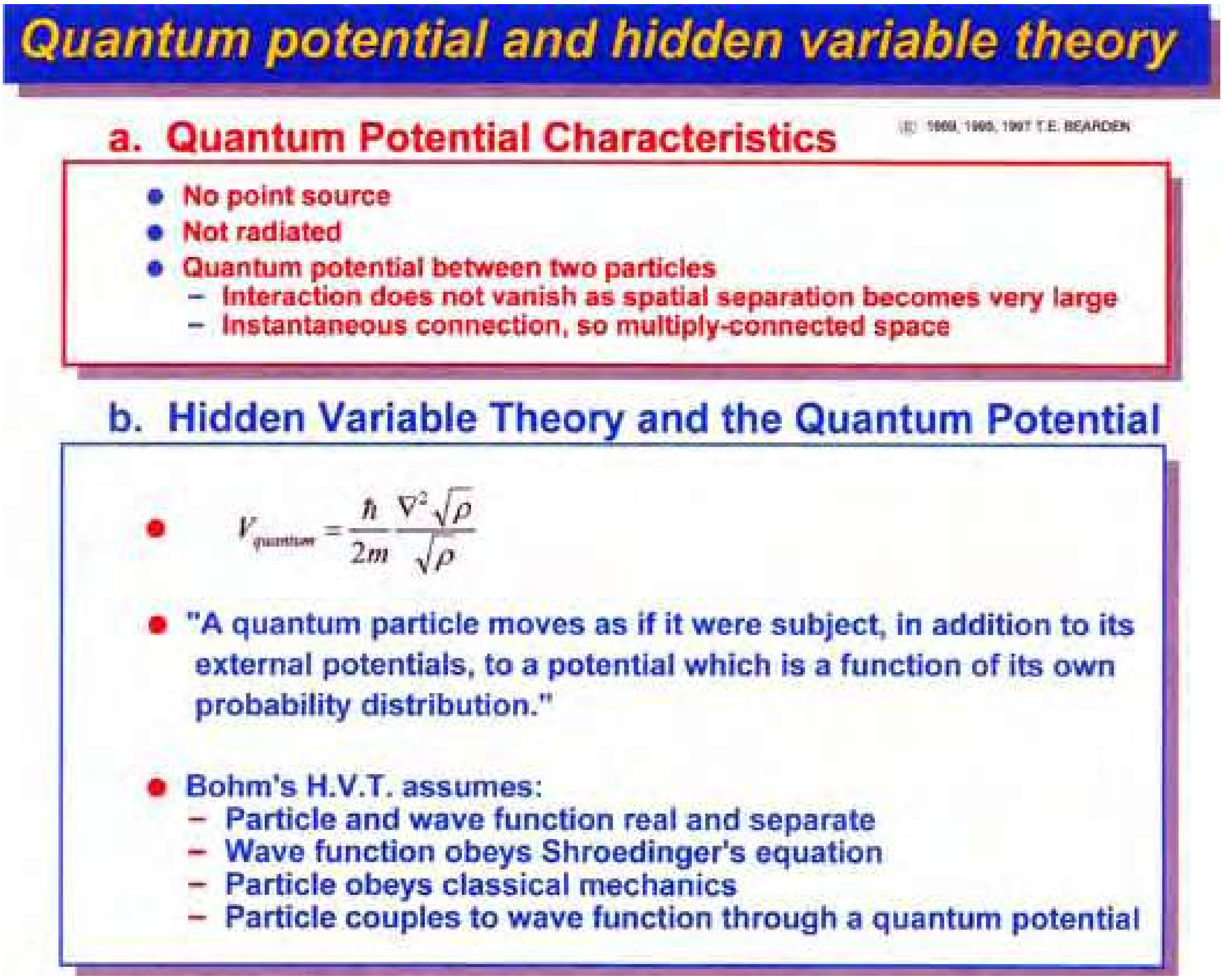

Quantum potential and hidden variable theory

a. Quantum Potential Characteristics

© 1988, 1989, 1997 T.E. BEARDEN

- No point source
- Not radiated
- Quantum potential between two particles
  - Interaction does not vanish as spatial separation becomes very large
  - Instantaneous connection, so multiply-connected space

b. Hidden Variable Theory and the Quantum Potential

- $V_{quantum} = \frac{\hbar}{2m} \frac{\nabla^2 \sqrt{\rho}}{\sqrt{\rho}}$
- "A quantum particle moves as if it were subject, in addition to its external potentials, to a potential which is a function of its own probability distribution."
- Bohm's H.V.T. assumes:
  - Particle and wave function real and separate
  - Wave function obeys Shroedinger's equation
  - Particle obeys classical mechanics
  - Particle couples to wave function through a quantum potential

Return to Slide Index

Next Slide

Previous Slide

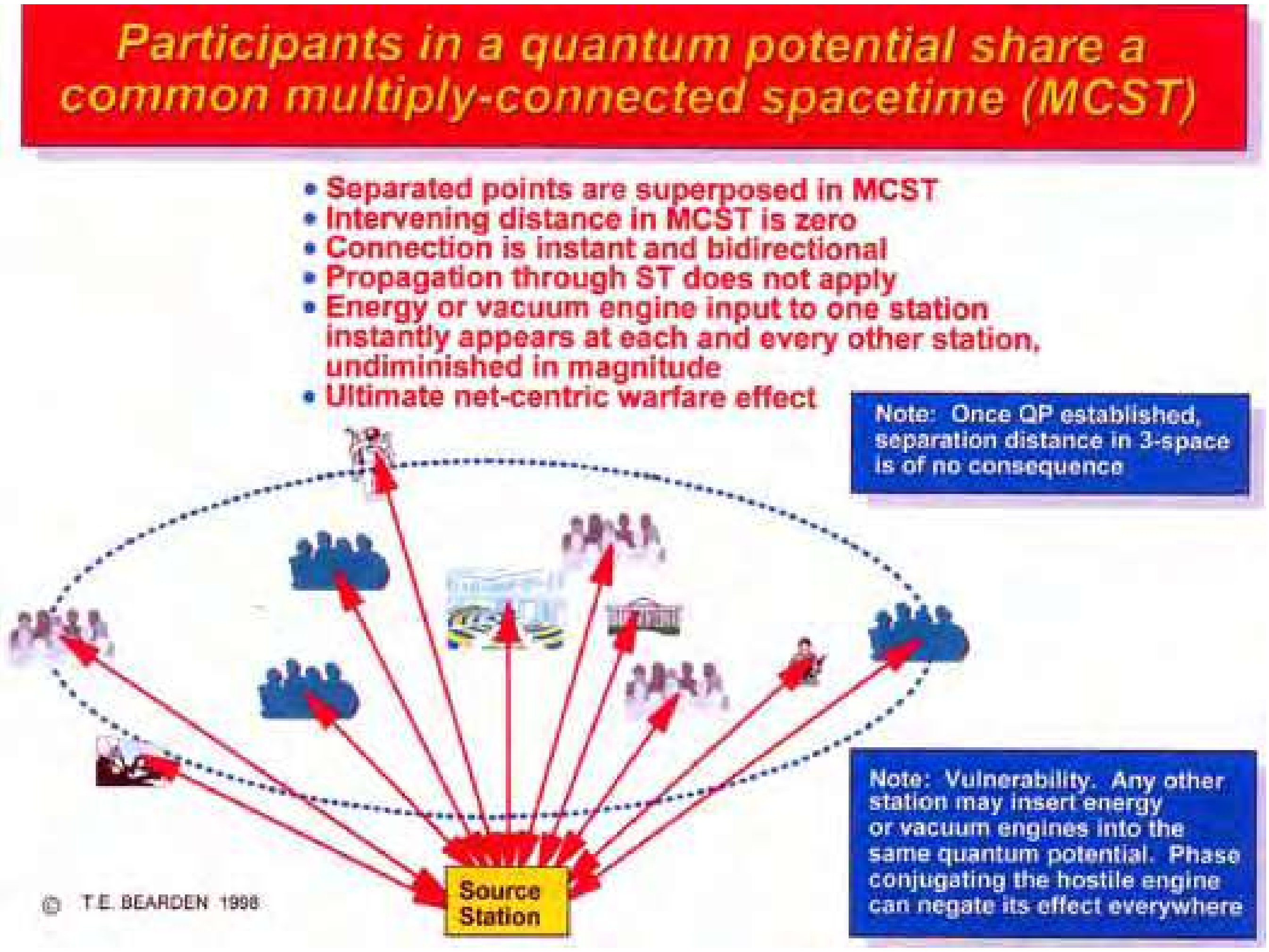

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Intervening distance in MCST is zero
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude
- Ultimate net-centric warfare effect

Note: Once QP established, separation distance in 3-space is of no consequence

Note: Vulnerability. Any other station may insert energy or vacuum engines into the same quantum potential. Phase conjugating the hostile engine can negate its effect everywhere

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

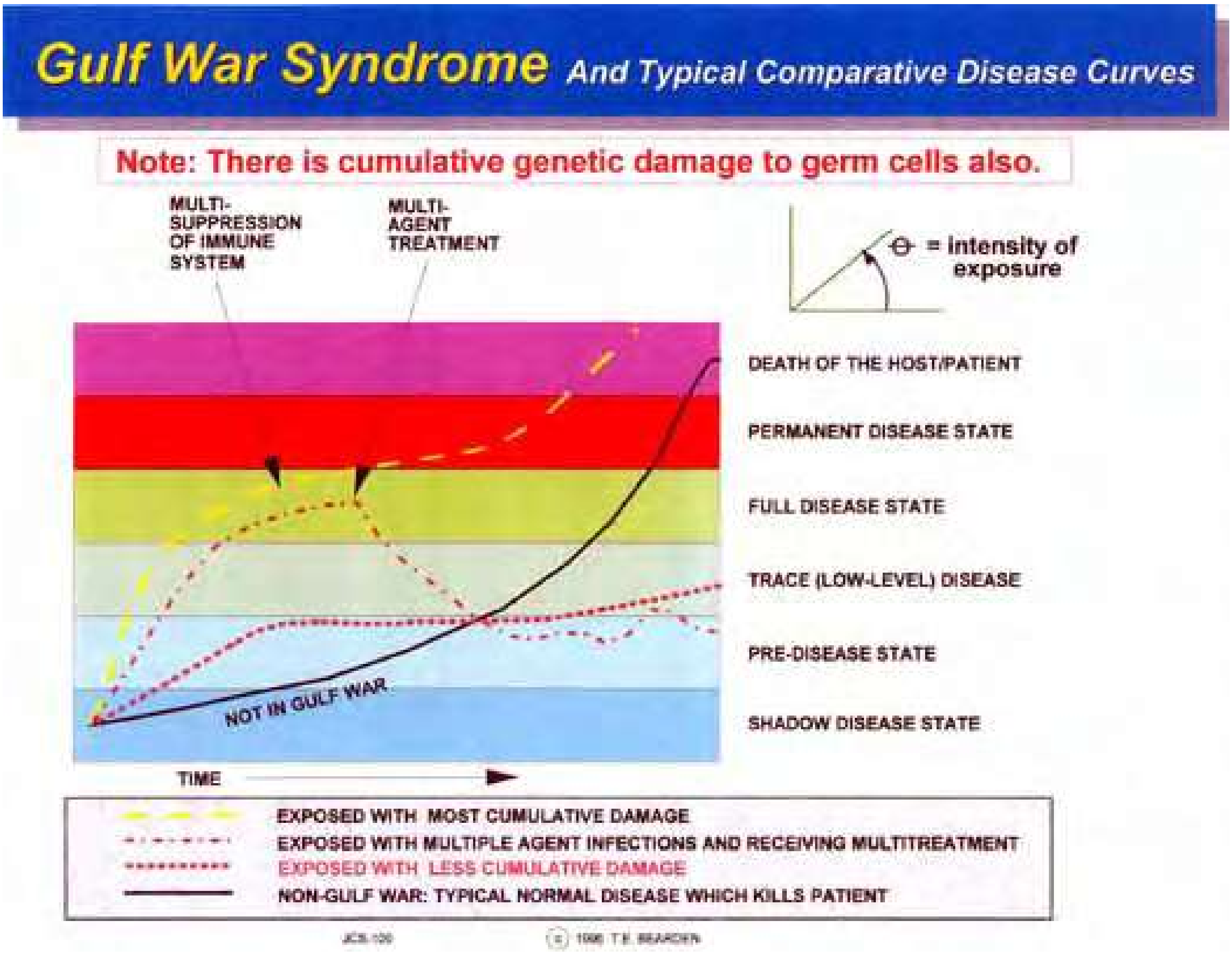

Gulf War Syndrome And Typical Comparative Disease Curves

Note: There is cumulative genetic damage to germ cells also.

- - - EXPOSED WITH MOST CUMULATIVE DAMAGE
-·-·-·- EXPOSED WITH MULTIPLE AGENT INFECTIONS AND RECEIVING MULTITREATMENT
*********** EXPOSED WITH LESS CUMULATIVE DAMAGE
—— NON-GULF WAR: TYPICAL NORMAL DISEASE WHICH KILLS PATIENT

JCS-100

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

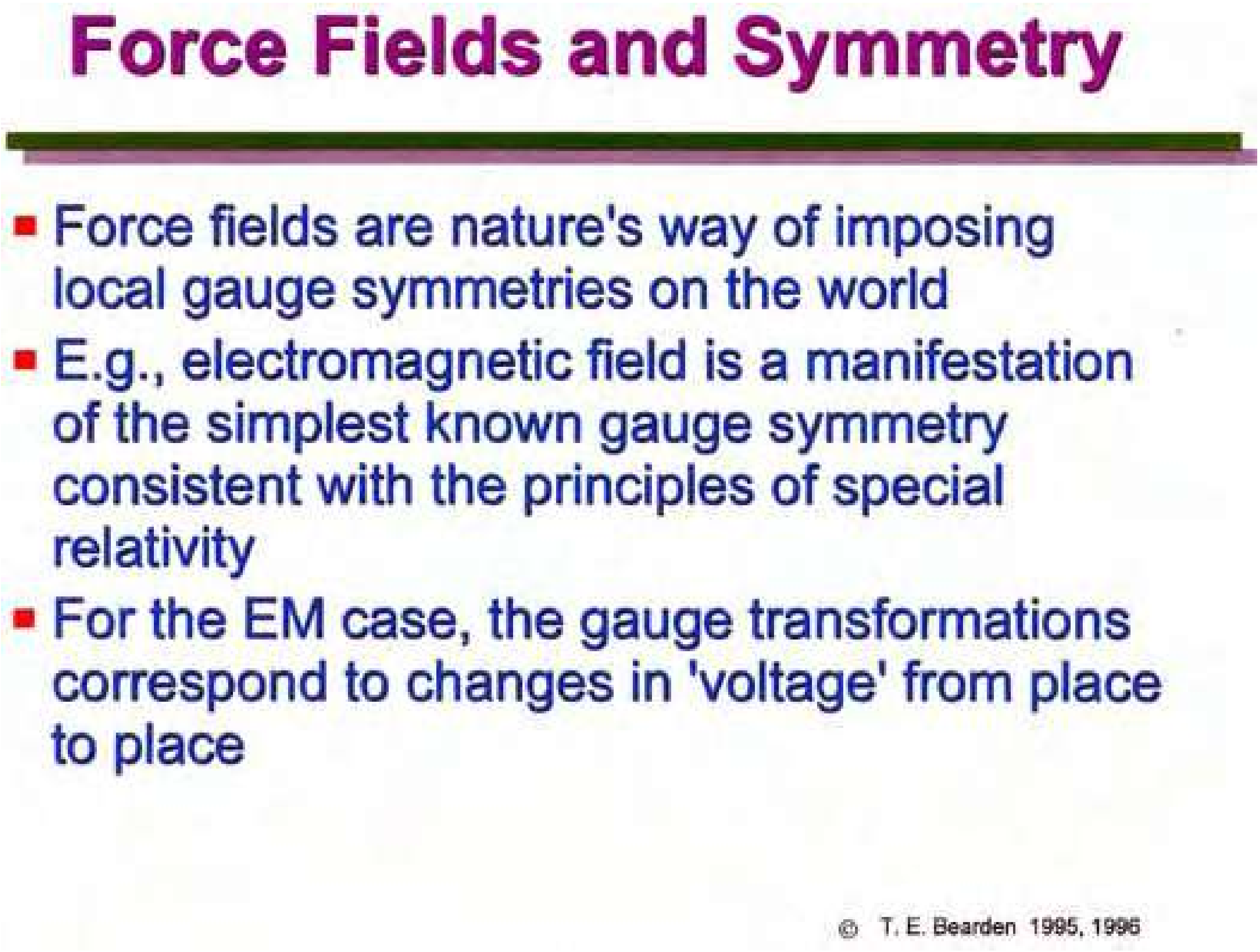

Force Fields and Symmetry

- Force fields are nature's way of imposing local gauge symmetries on the world
- E.g., electromagnetic field is a manifestation of the simplest known gauge symmetry consistent with the principles of special relativity
- For the EM case, the gauge transformations correspond to changes in 'voltage' from place to place

© T. E. Bearden 1995, 1996

Return to Slide Index

Next Slide

Previous Slide

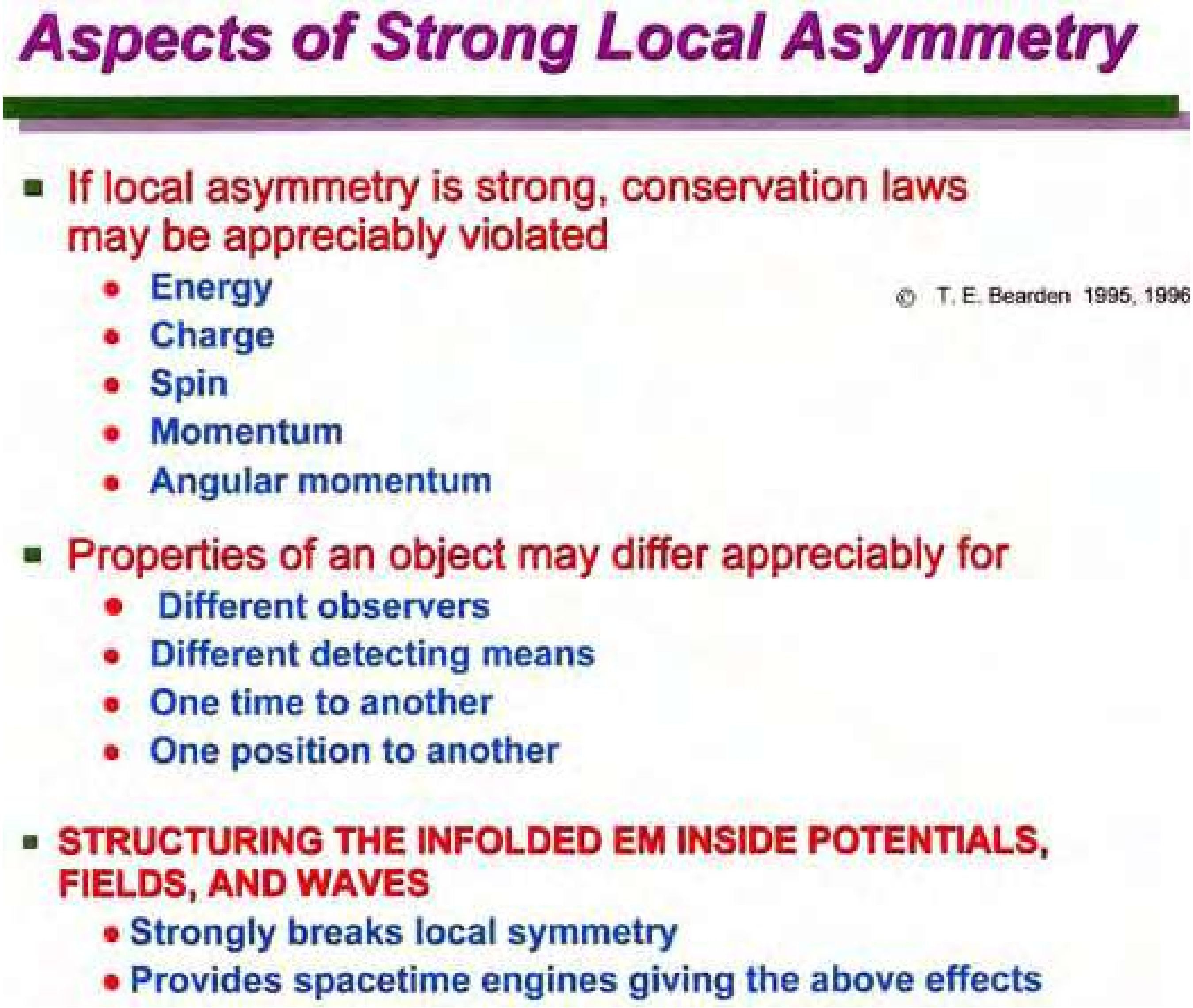

Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

Return to Slide Index

Next Slide

Previous Slide

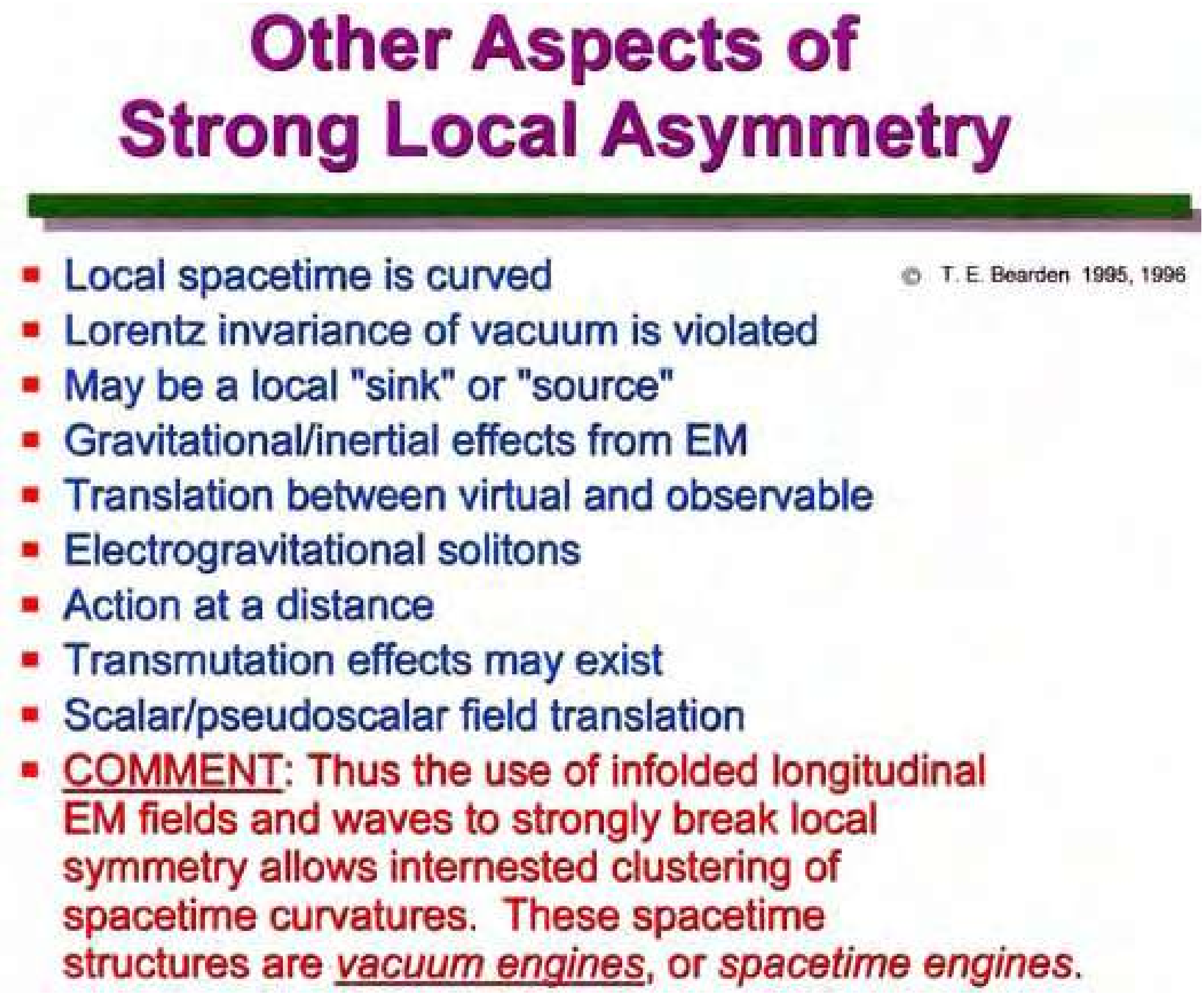

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational/inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

Return to Slide Index

Next Slide

Previous Slide

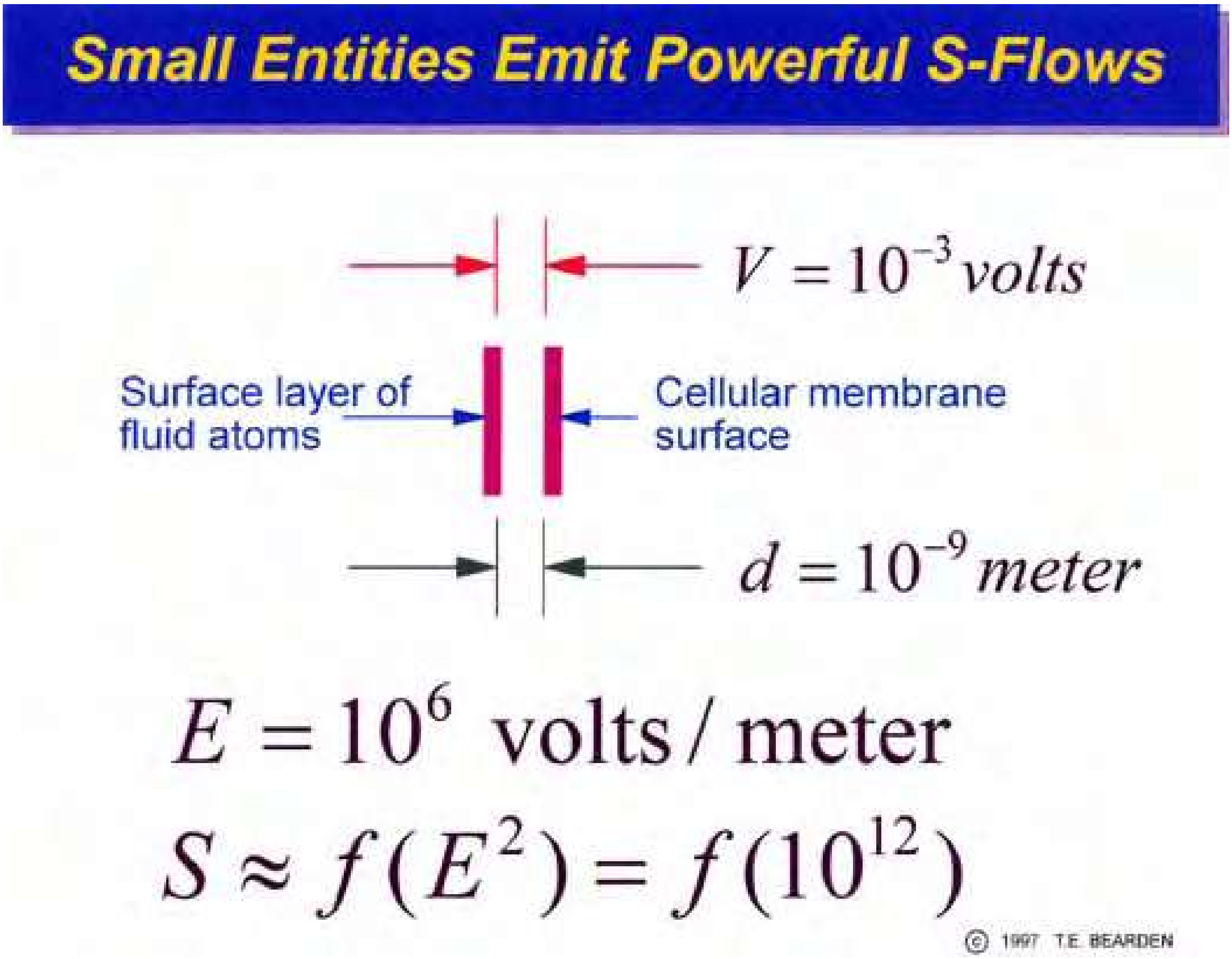

Small Entities Emit Powerful S-Flows

$V = 10^{-3}\, volts$

Surface layer of fluid atoms

Cellular membrane surface

$d = 10^{-9}\, meter$

$$E = 10^{6} \text{ volts / meter}$$

$$S \approx f(E^{2}) = f(10^{12})$$

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

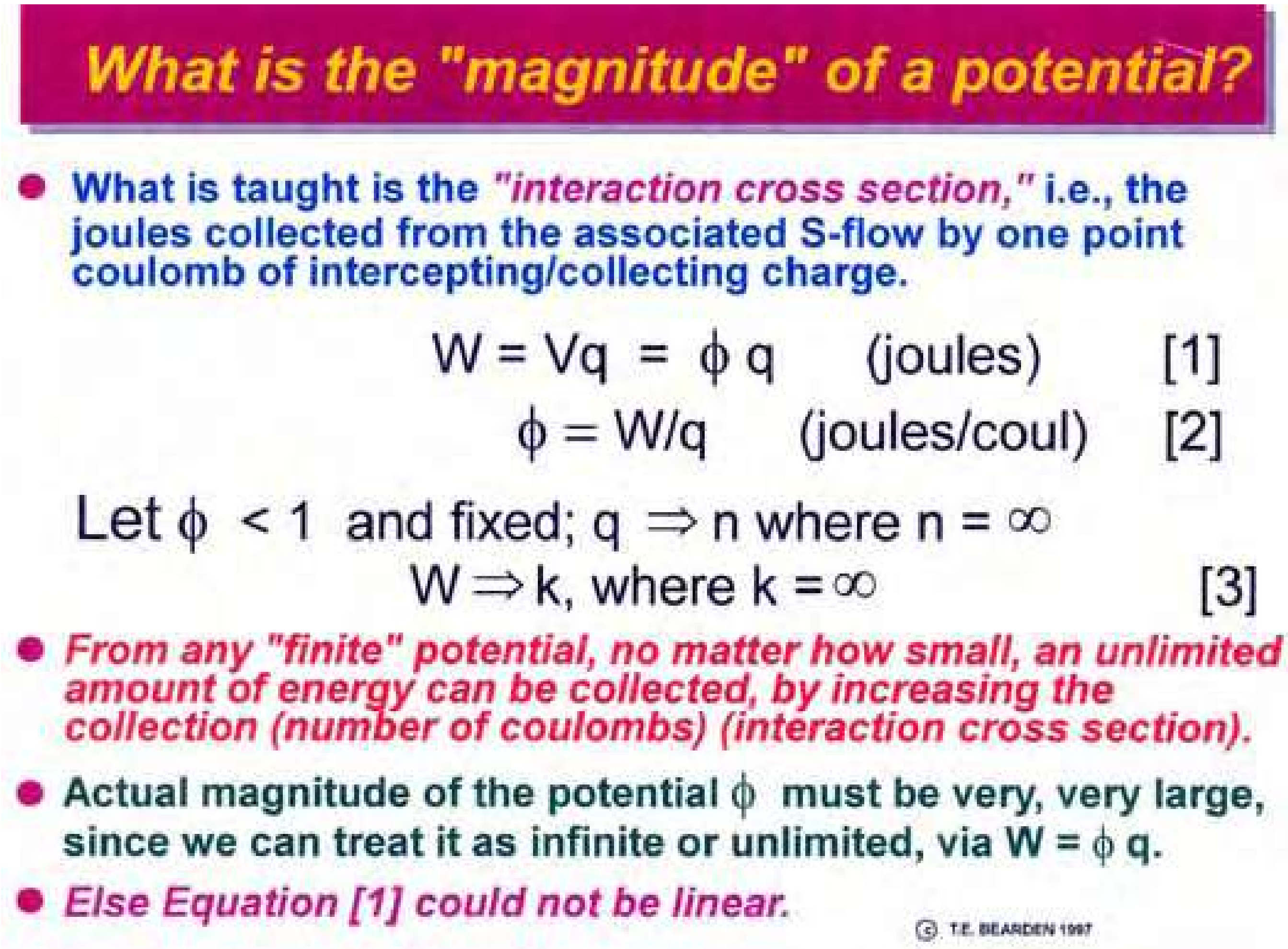

What is the "magnitude" of a potential?

- What is taught is the *"interaction cross section,"* i.e., the joules collected from the associated S-flow by one point coulomb of intercepting/collecting charge.

$$W = Vq = \phi q \quad \text{(joules)} \qquad [1]$$

$$\phi = W/q \quad \text{(joules/coul)} \qquad [2]$$

$$\text{Let } \phi < 1 \text{ and fixed; } q \Rightarrow n \text{ where } n = \infty$$

$$W \Rightarrow k, \text{ where } k = \infty \qquad [3]$$

- *From any "finite" potential, no matter how small, an unlimited amount of energy can be collected, by increasing the collection (number of coulombs) (interaction cross section).*
- Actual magnitude of the potential ϕ must be very, very large, since we can treat it as infinite or unlimited, via $W = \phi q$.
- *Else Equation [1] could not be linear.*

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

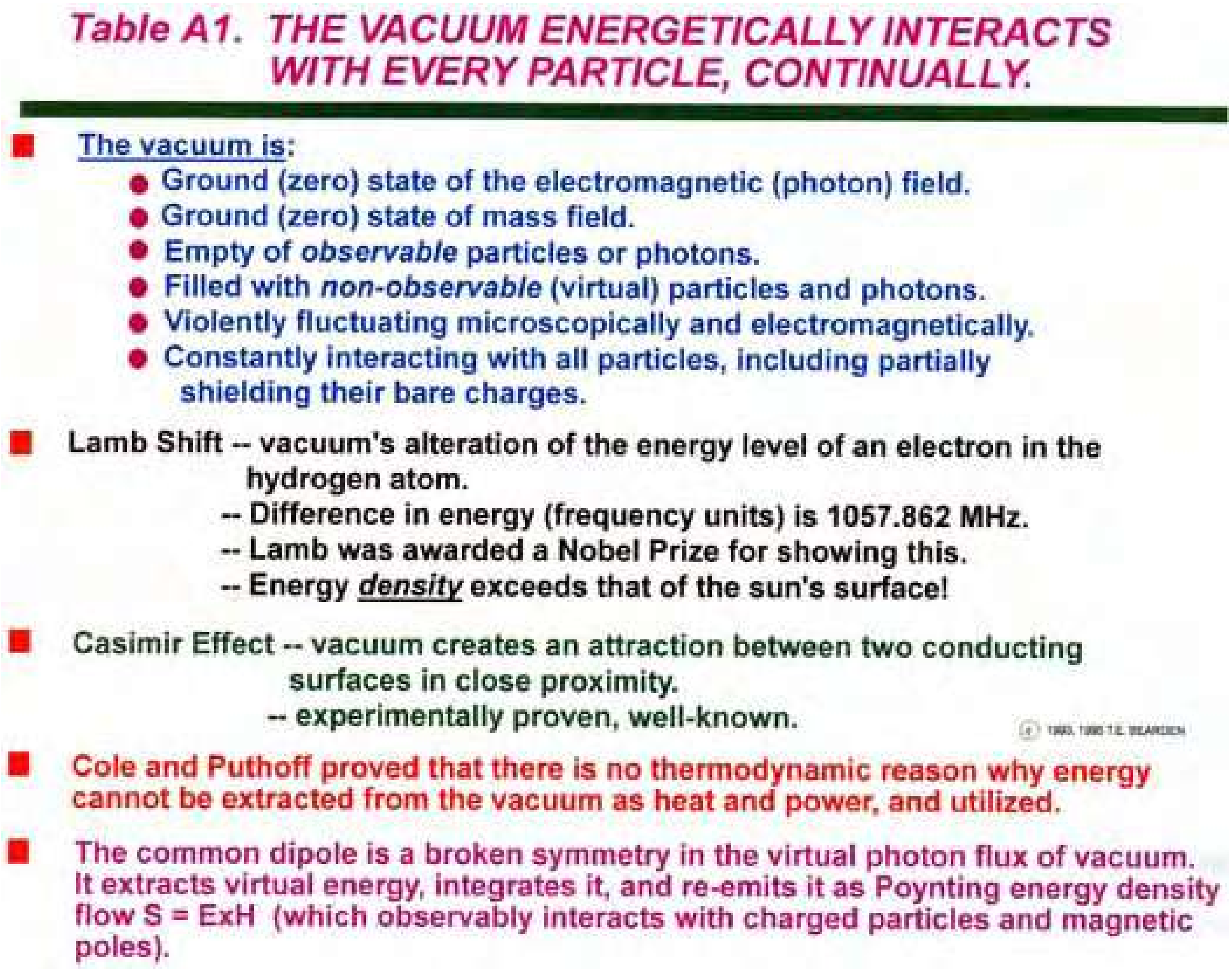

Table A1. THE VACUUM ENERGETICALLY INTERACTS WITH EVERY PARTICLE, CONTINUALLY.

- The vacuum is:
  - Ground (zero) state of the electromagnetic (photon) field.
  - Ground (zero) state of mass field.
  - Empty of *observable* particles or photons.
  - Filled with *non-observable* (virtual) particles and photons.
  - Violently fluctuating microscopically and electromagnetically.
  - Constantly interacting with all particles, including partially shielding their bare charges.
- Lamb Shift -- vacuum's alteration of the energy level of an electron in the hydrogen atom.
  - -- Difference in energy (frequency units) is 1057.862 MHz.
  - -- Lamb was awarded a Nobel Prize for showing this.
  - -- Energy ***density*** exceeds that of the sun's surface!
- Casimir Effect -- vacuum creates an attraction between two conducting surfaces in close proximity.
  - -- experimentally proven, well-known.

© [illegible] T.E. BEARDEN

- Cole and Puthoff proved that there is no thermodynamic reason why energy cannot be extracted from the vacuum as heat and power, and utilized.
- The common dipole is a broken symmetry in the virtual photon flux of vacuum. It extracts virtual energy, integrates it, and re-emits it as Poynting energy density flow S = ExH (which observably interacts with charged particles and magnetic poles).

Return to Slide Index

Next Slide

Previous Slide

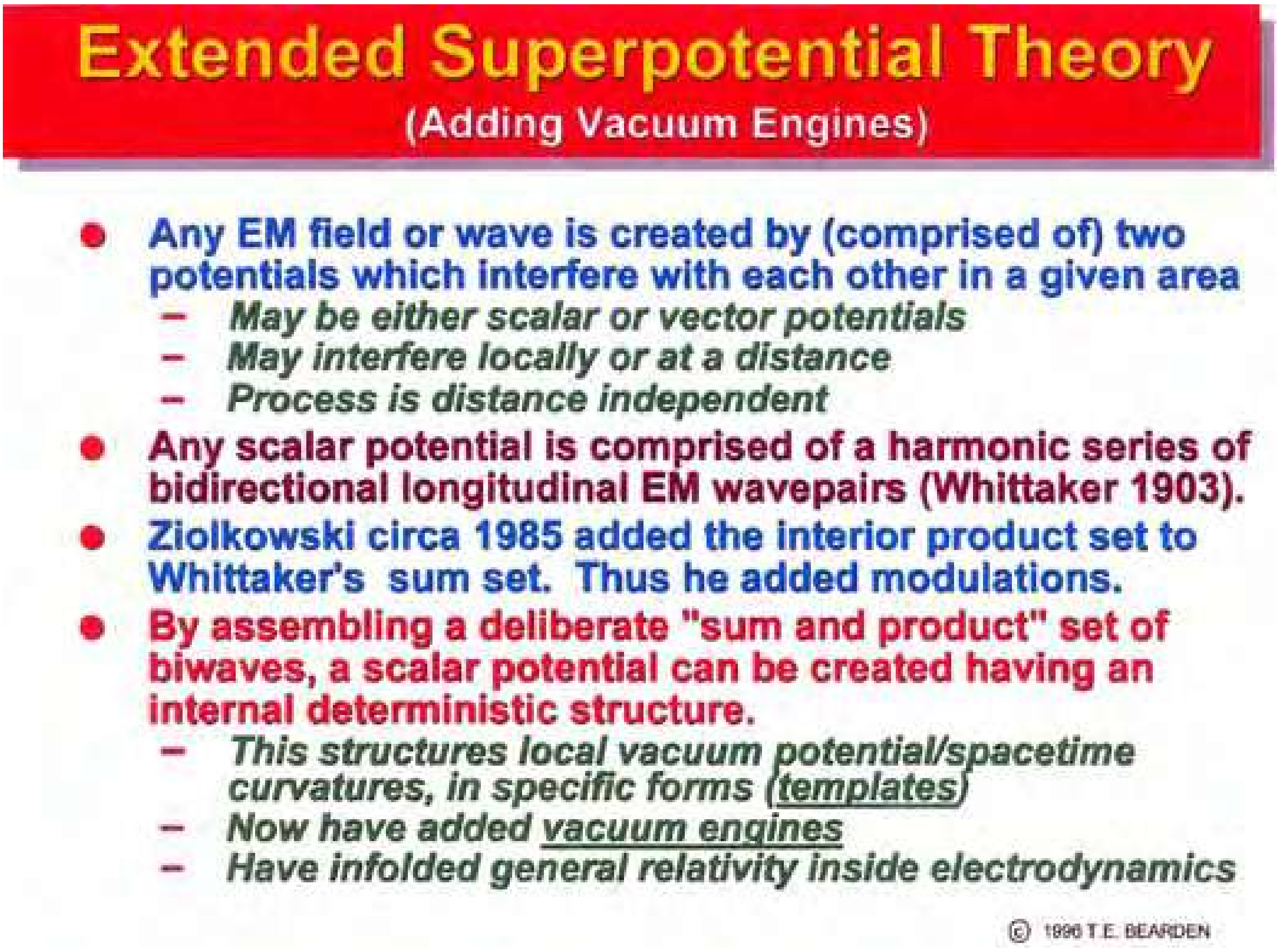

Extended Superpotential Theory

(Adding Vacuum Engines)

- Any EM field or wave is created by (comprised of) two potentials which interfere with each other in a given area
  - *May be either scalar or vector potentials*
  - *May interfere locally or at a distance*
  - *Process is distance independent*
- Any scalar potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs (Whittaker 1903).
- Ziolkowski circa 1985 added the interior product set to Whittaker's sum set. Thus he added modulations.
- By assembling a deliberate "sum and product" set of biwaves, a scalar potential can be created having an internal deterministic structure.
  - *This structures local vacuum potential/spacetime curvatures, in specific forms (templates)*
  - *Now have added vacuum engines*
  - *Have infolded general relativity inside electrodynamics*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

Vacuum Engine: Jillions of Working ST Demons

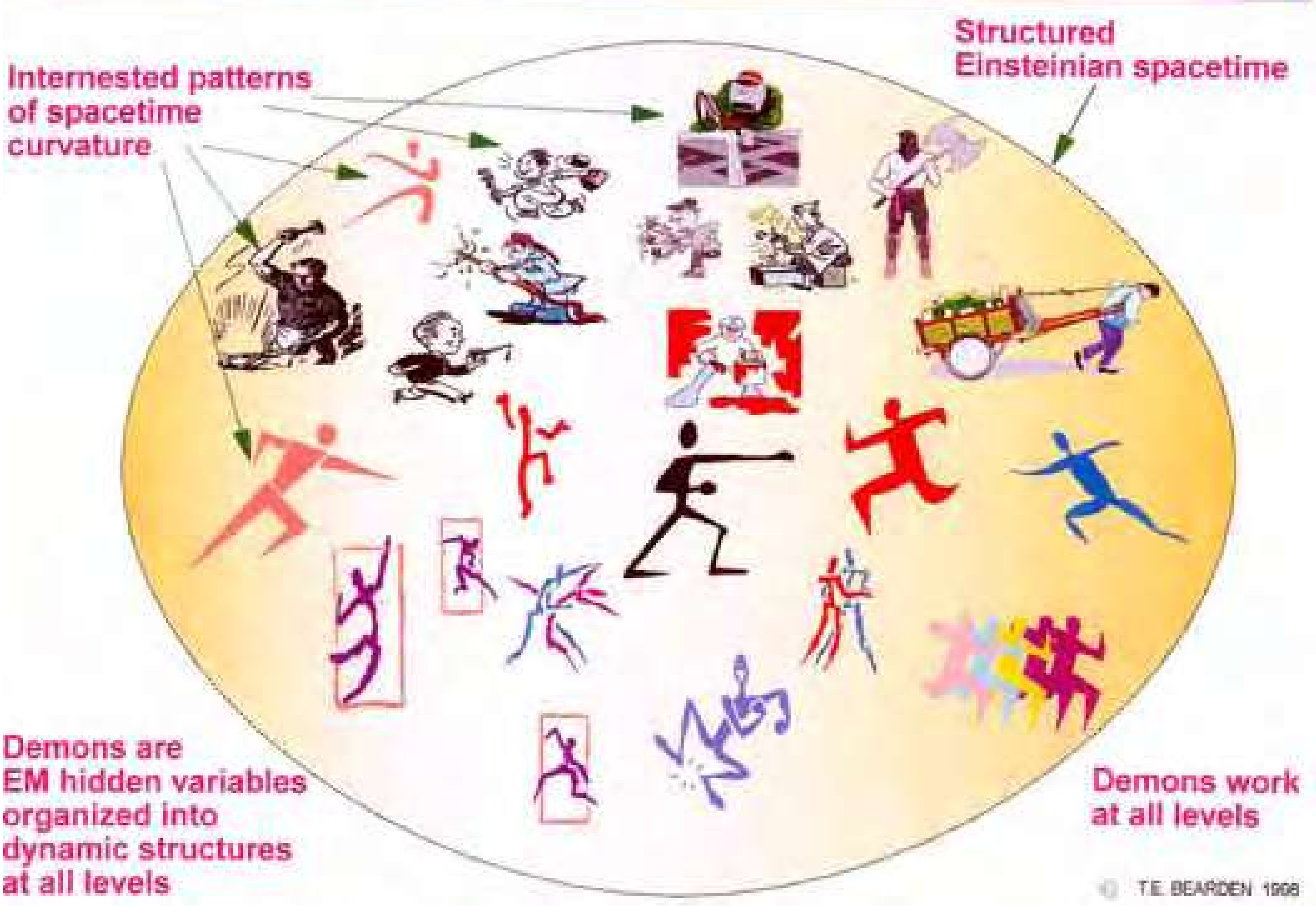

Internested patterns of spacetime curvature

Structured Einsteinian spacetime

Demons are EM hidden variables organized into dynamic structures at all levels

Demons work at all levels

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

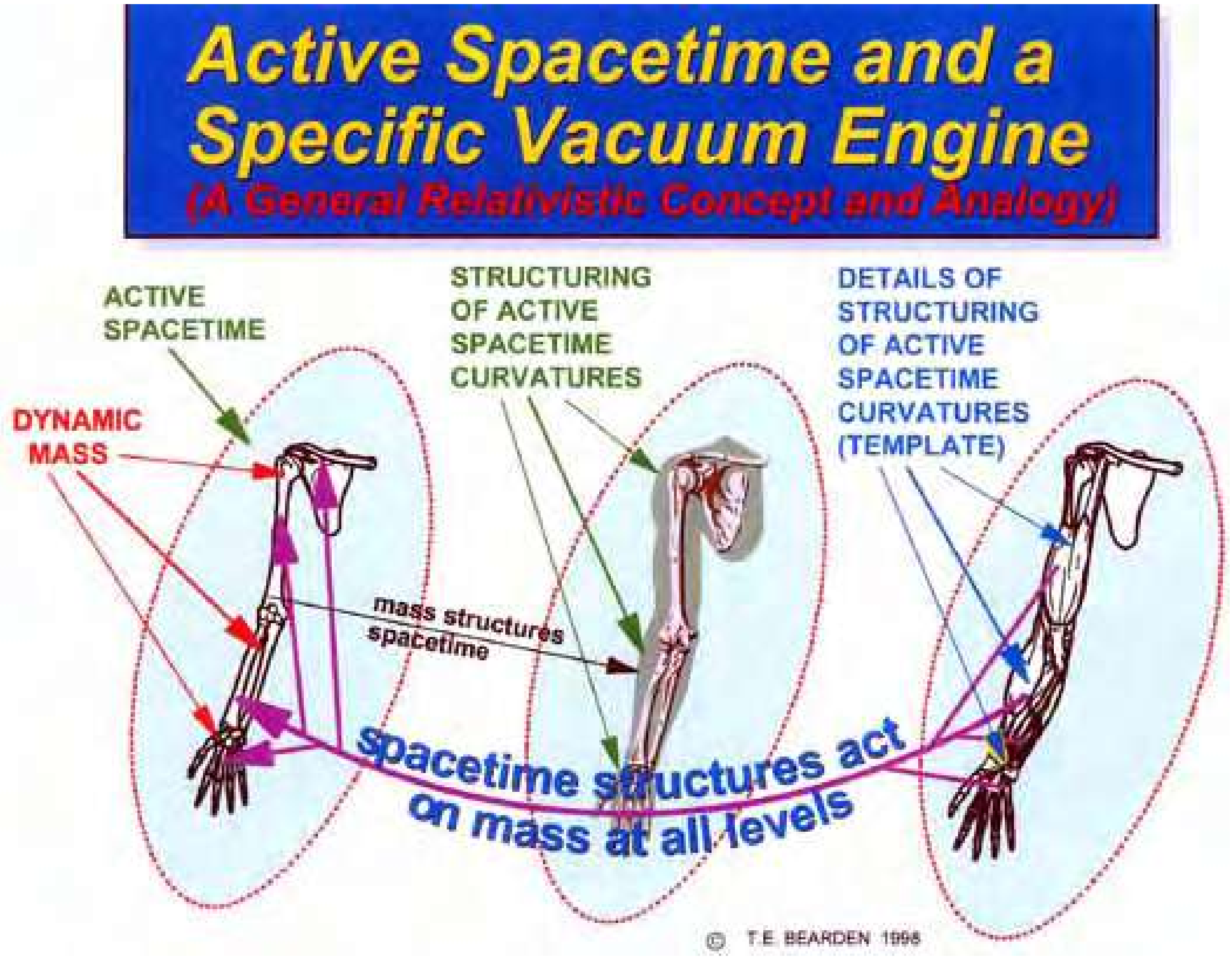

Active Spacetime and a Specific Vacuum Engine

(A General Relativistic Concept and Analogy)

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)
mass structures spacetime
spacetime structures act on mass at all levels

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

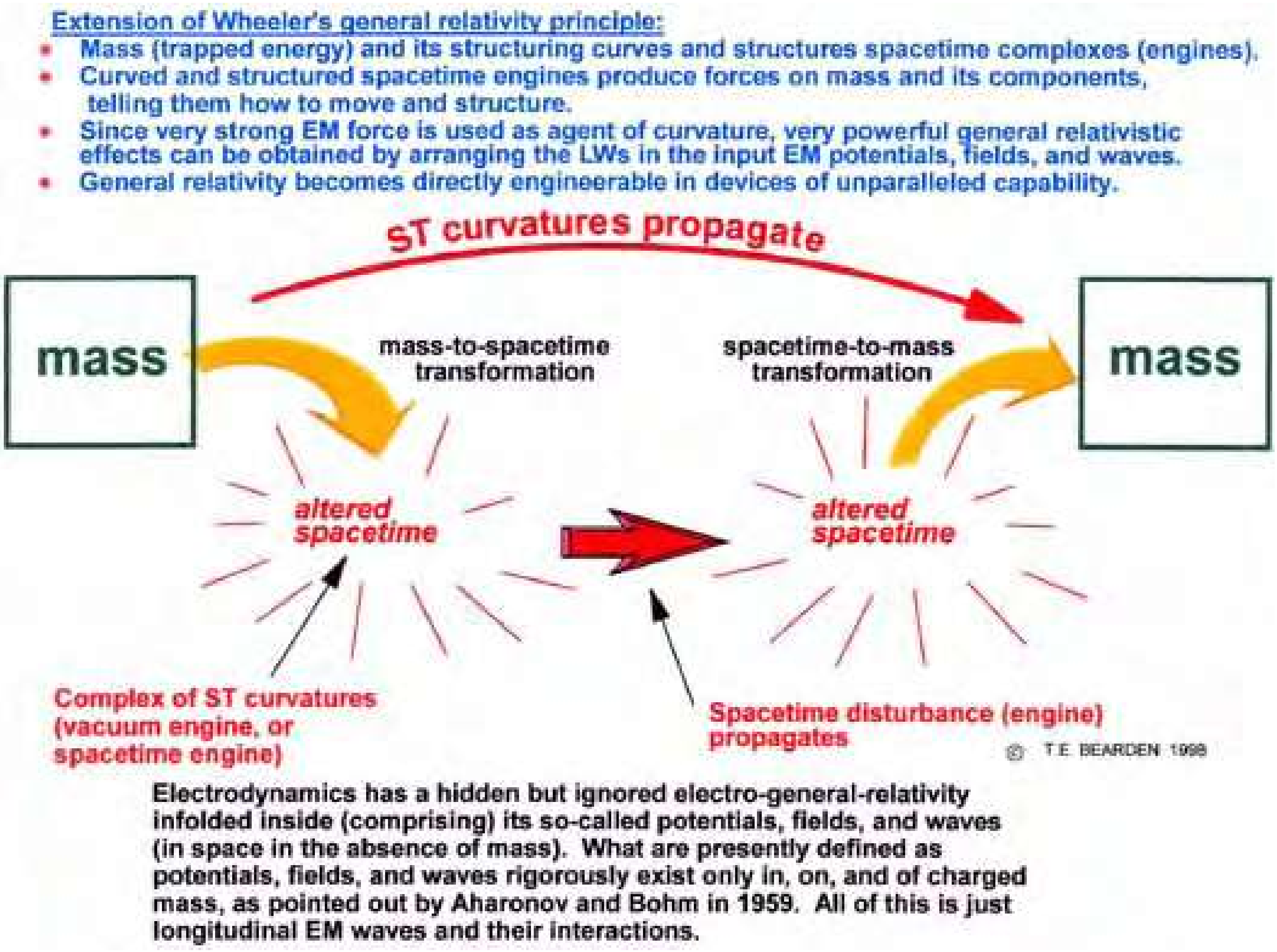

Extension of Wheeler's general relativity principle:

- Mass (trapped energy) and its structuring curves and structures spacetime complexes (engines).
- Curved and structured spacetime engines produce forces on mass and its components, telling them how to move and structure.
- Since very strong EM force is used as agent of curvature, very powerful general relativistic effects can be obtained by arranging the LWs in the input EM potentials, fields, and waves.
- General relativity becomes directly engineerable in devices of unparalleled capability.

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

© T.E. BEARDEN 1998

Electrodynamics has a hidden but ignored electro-general-relativity infolded inside (comprising) its so-called potentials, fields, and waves (in space in the absence of mass). What are presently defined as potentials, fields, and waves rigorously exist only in, on, and of charged mass, as pointed out by Aharonov and Bohm in 1959. All of this is just longitudinal EM waves and their interactions.

Return to Slide Index

Next Slide

Previous Slide

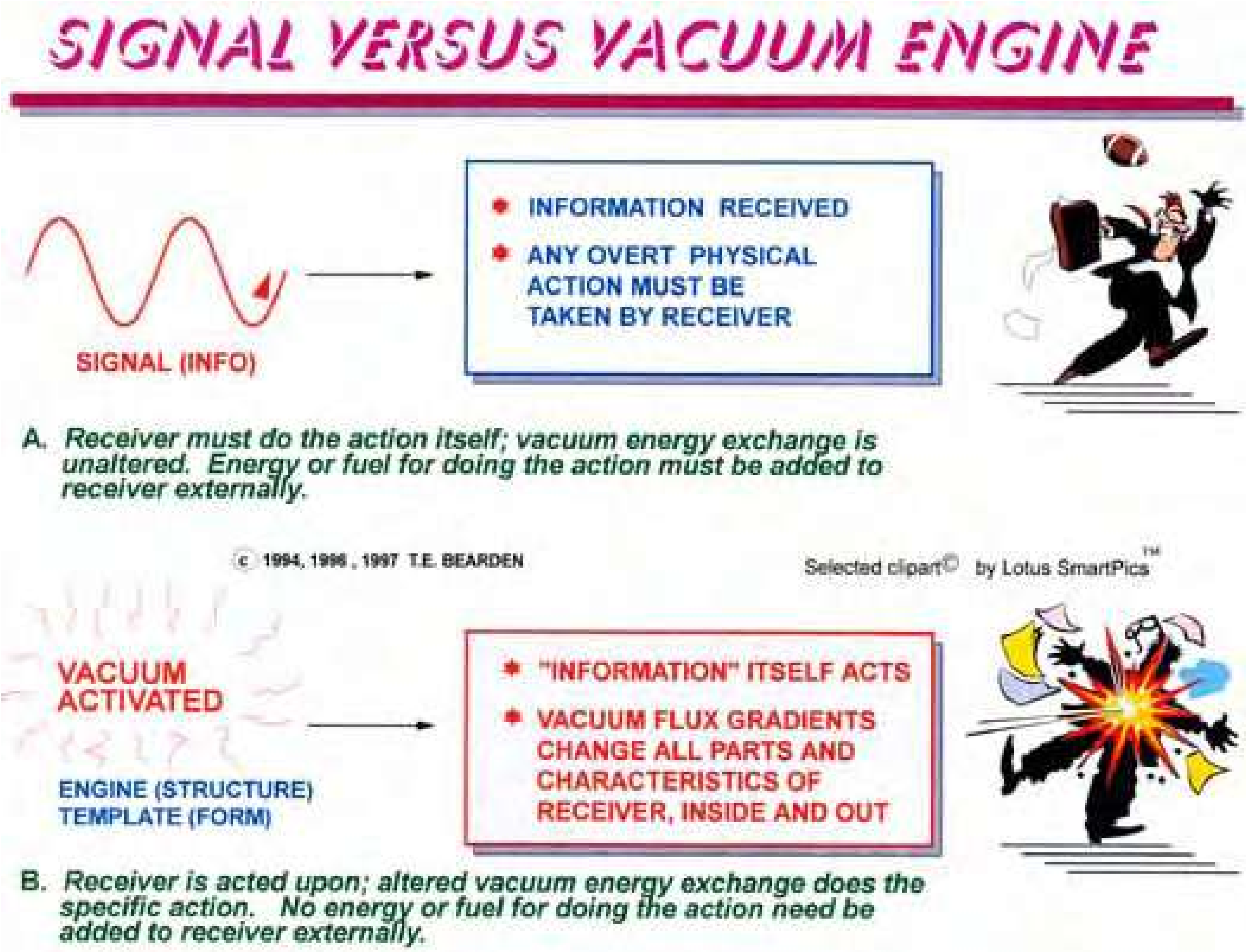

SIGNAL VERSUS VACUUM ENGINE

SIGNAL (INFO)

- INFORMATION RECEIVED
- ANY OVERT PHYSICAL ACTION MUST BE TAKEN BY RECEIVER

A. *Receiver must do the action itself; vacuum energy exchange is unaltered. Energy or fuel for doing the action must be added to receiver externally.*

© 1994, 1996, 1997 T.E. BEARDEN

Selected clipart© by Lotus SmartPics™

VACUUM ACTIVATED
ENGINE (STRUCTURE) TEMPLATE (FORM)

- "INFORMATION" ITSELF ACTS
- VACUUM FLUX GRADIENTS CHANGE ALL PARTS AND CHARACTERISTICS OF RECEIVER, INSIDE AND OUT

B. *Receiver is acted upon; altered vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to receiver externally.*

Return to Slide Index

Next Slide

Previous Slide

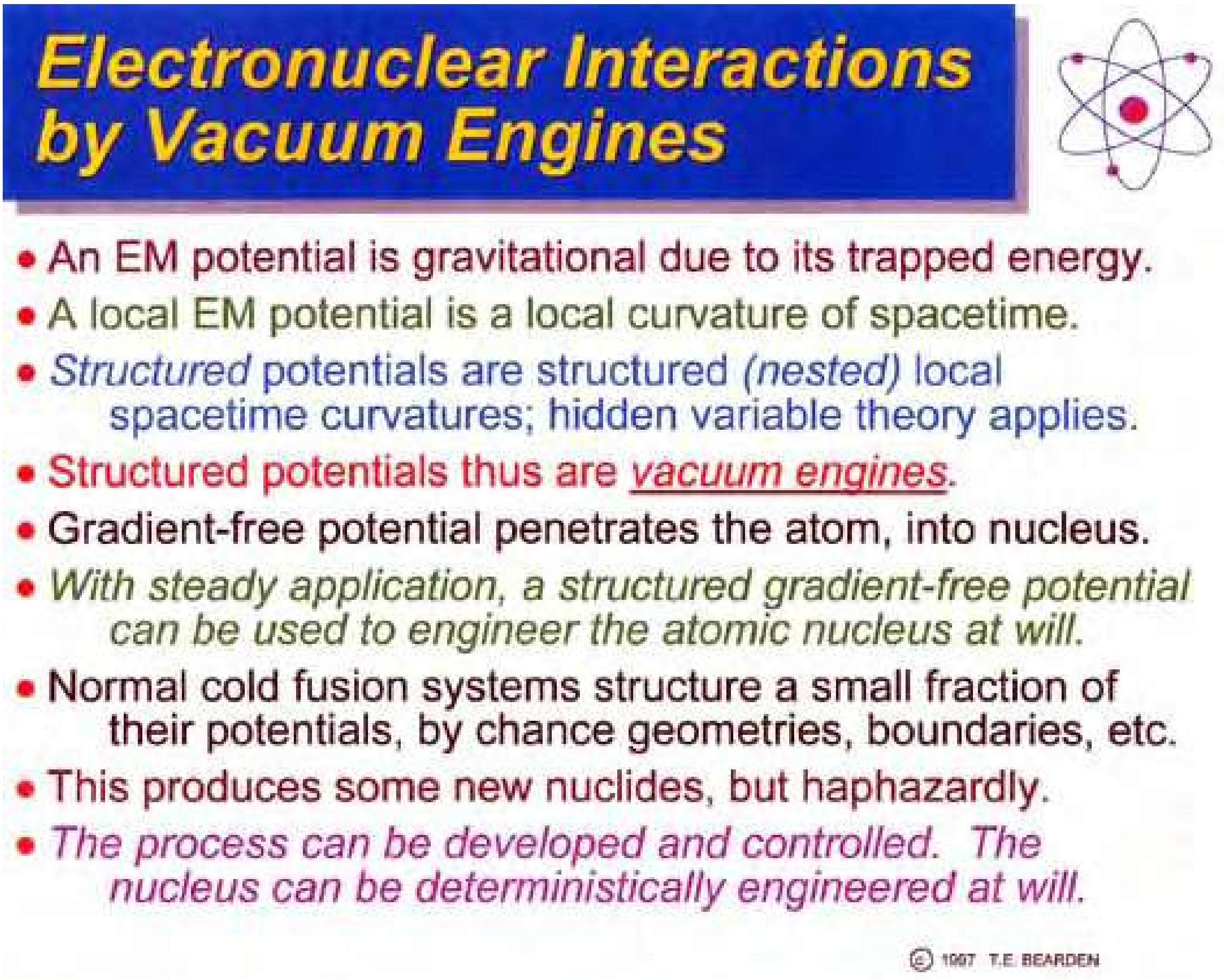

Electronuclear Interactions by Vacuum Engines

- An EM potential is gravitational due to its trapped energy.
- A local EM potential is a local curvature of spacetime.
- *Structured* potentials are structured *(nested)* local spacetime curvatures; hidden variable theory applies.
- Structured potentials thus are *vacuum engines*.
- Gradient-free potential penetrates the atom, into nucleus.
- *With steady application, a structured gradient-free potential can be used to engineer the atomic nucleus at will.*
- Normal cold fusion systems structure a small fraction of their potentials, by chance geometries, boundaries, etc.
- This produces some new nuclides, but haphazardly.
- *The process can be developed and controlled. The nucleus can be deterministically engineered at will.*

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

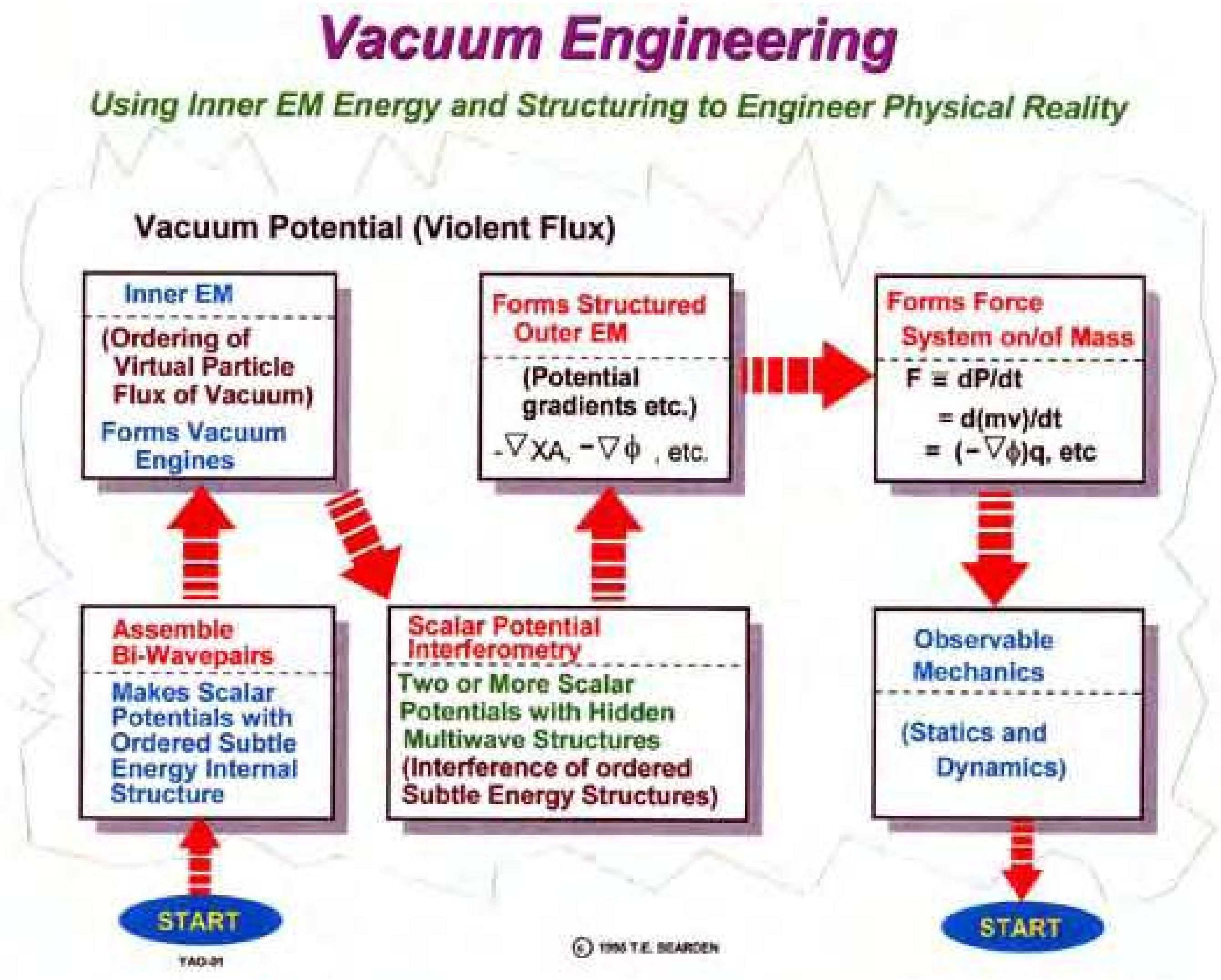

Vacuum Engineering

Using Inner EM Energy and Structuring to Engineer Physical Reality

Vacuum Potential (Violent Flux)

Inner EM
(Ordering of Virtual Particle Flux of Vacuum)
Forms Vacuum Engines

Forms Structured Outer EM
(Potential gradients etc.)
$-\nabla \times A$, $-\nabla \phi$, etc.

Forms Force System on/of Mass
$F \equiv dP/dt$
$= d(mv)/dt$
$\equiv (-\nabla \phi)q$, etc

Assemble Bi-Wavepairs
Makes Scalar Potentials with Ordered Subtle Energy Internal Structure

Scalar Potential Interferometry
Two or More Scalar Potentials with Hidden Multiwave Structures (Interference of ordered Subtle Energy Structures)

Observable Mechanics
(Statics and Dynamics)

START

START

TAO-91

© 1996 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

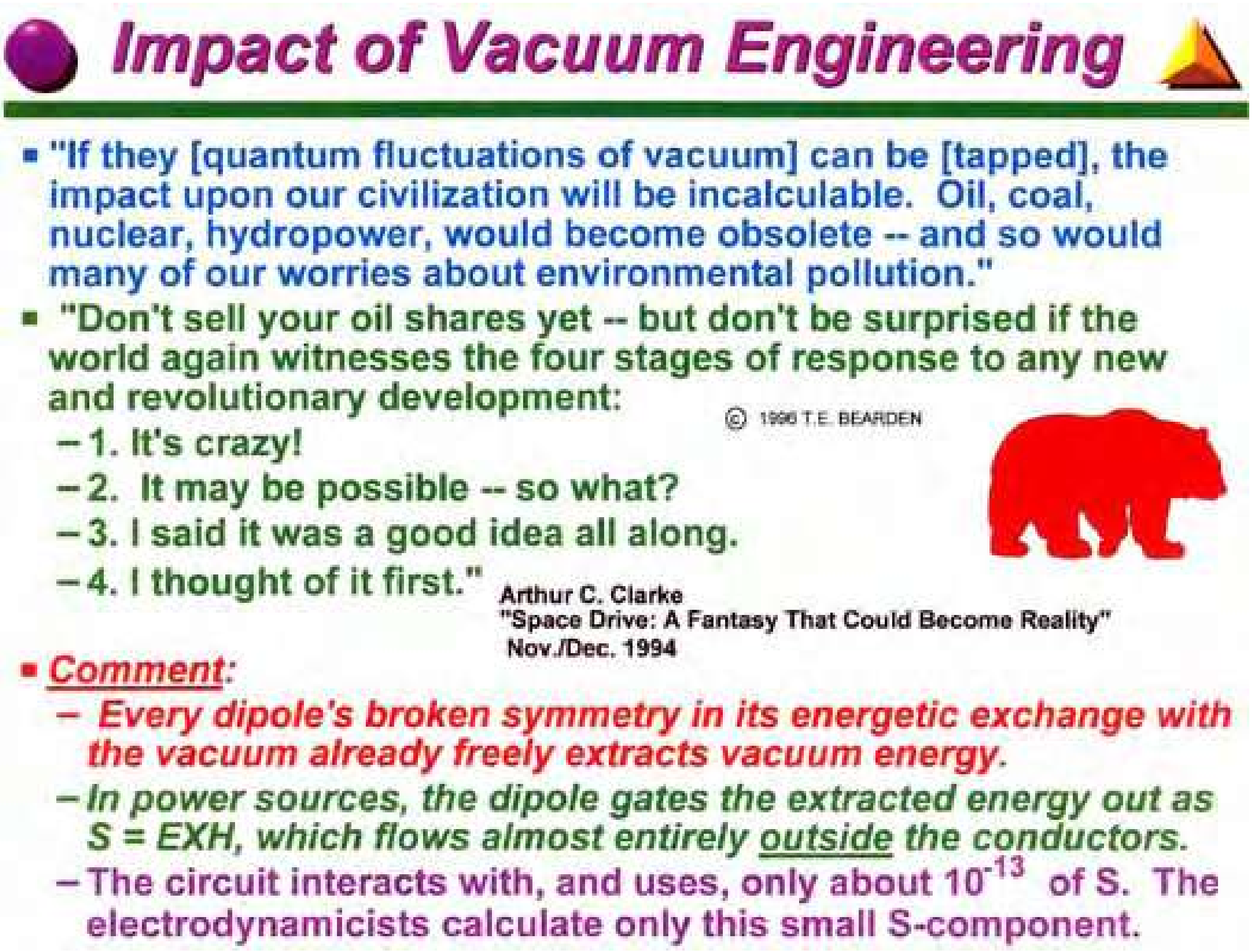

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete -- and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet -- but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

© 1996 T.E. BEARDEN

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

- Comment:
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

Return to Slide Index

Next Slide

Previous Slide

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

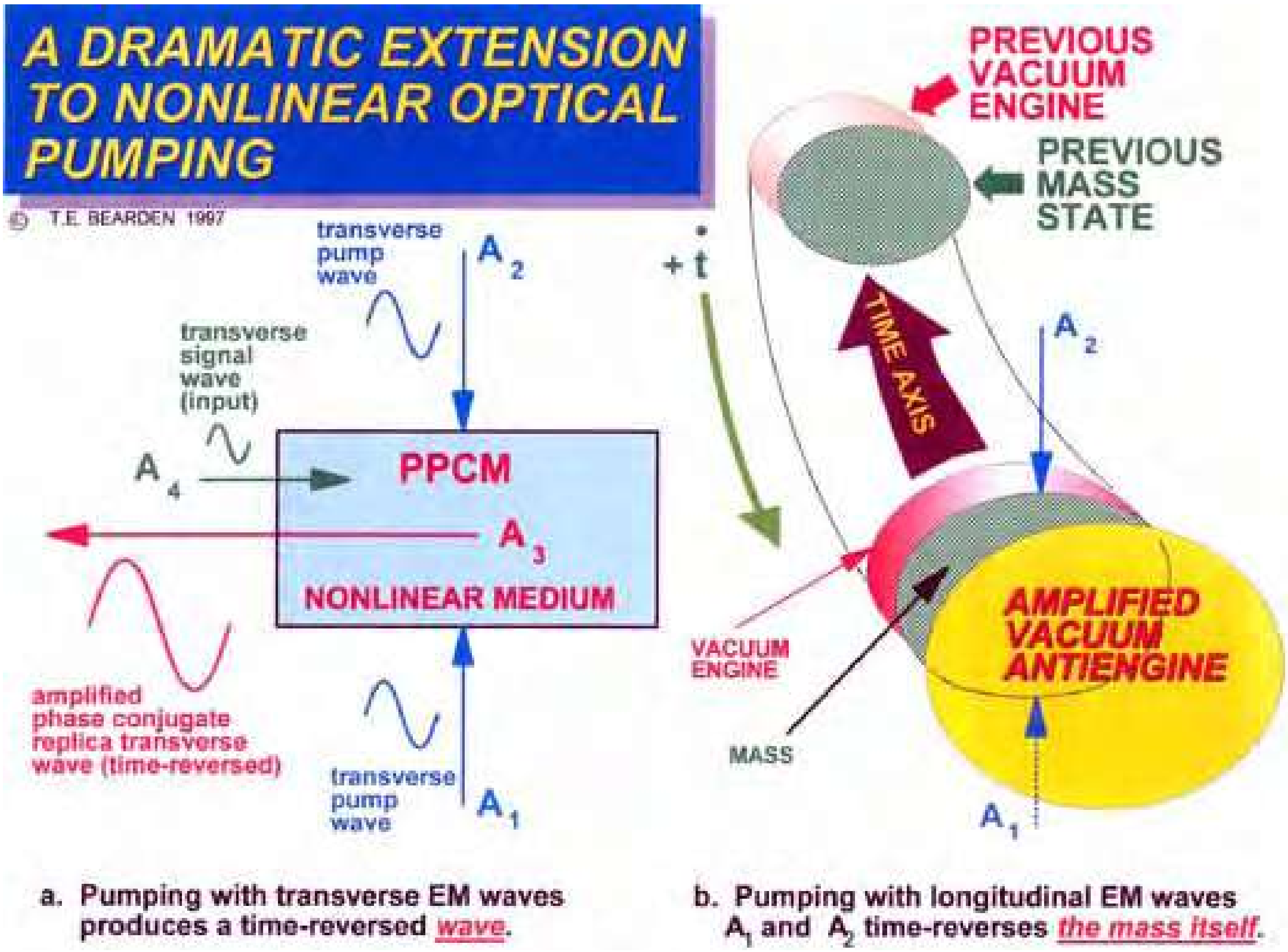
transverse pump wave
A_2
transverse signal wave (input)
A_4
PPCM
A_3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A_1
+ t
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
TIME AXIS
A_2
VACUUM ENGINE
AMPLIFIED VACUUM ANTIENGINE
MASS
A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Return to Slide Index

Next Slide

Previous Slide

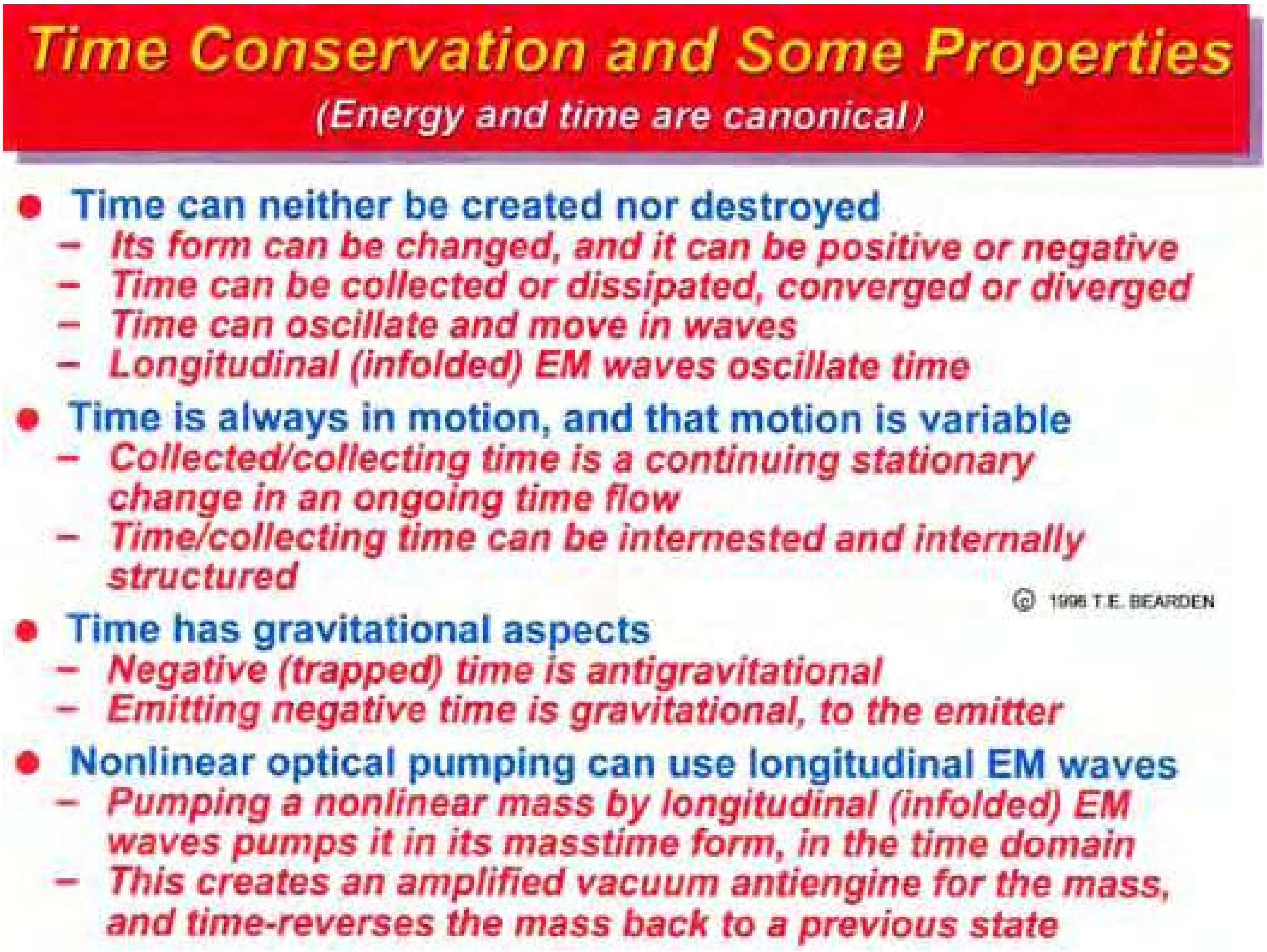

Time Conservation and Some Properties

(Energy and time are canonical)

- Time can neither be created nor destroyed
  - *Its form can be changed, and it can be positive or negative*
  - *Time can be collected or dissipated, converged or diverged*
  - *Time can oscillate and move in waves*
  - *Longitudinal (infolded) EM waves oscillate time*
- Time is always in motion, and that motion is variable
  - *Collected/collecting time is a continuing stationary change in an ongoing time flow*
  - *Time/collecting time can be internested and internally structured*
- Time has gravitational aspects
  - *Negative (trapped) time is antigravitational*
  - *Emitting negative time is gravitational, to the emitter*
- Nonlinear optical pumping can use longitudinal EM waves
  - *Pumping a nonlinear mass by longitudinal (infolded) EM waves pumps it in its masstime form, in the time domain*
  - *This creates an amplified vacuum antiengine for the mass, and time-reverses the mass back to a previous state*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

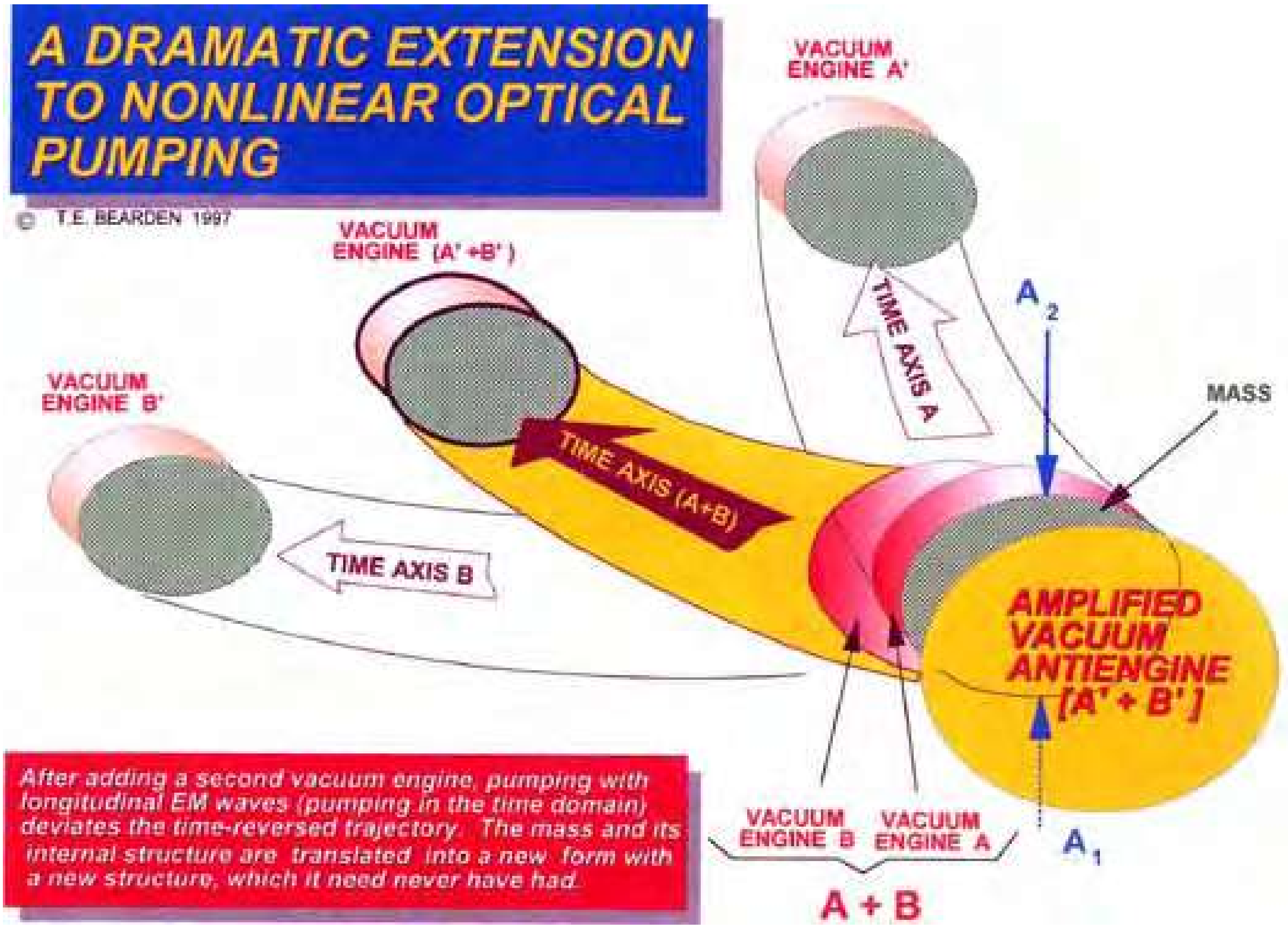

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
A_2
MASS
TIME AXIS (A+B)
TIME AXIS B
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A_1
A + B

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

Return to Slide Index

Next Slide

Previous Slide

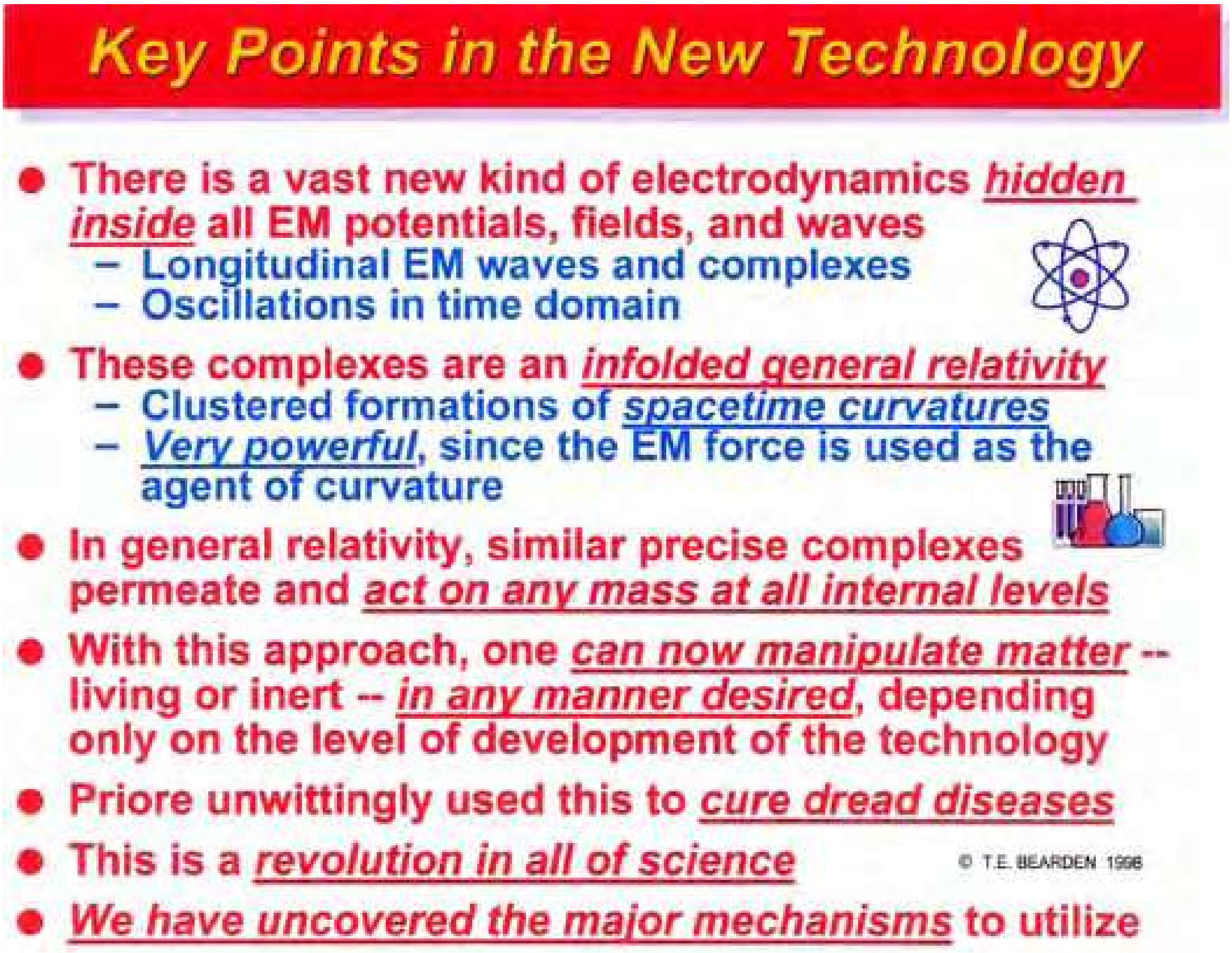

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

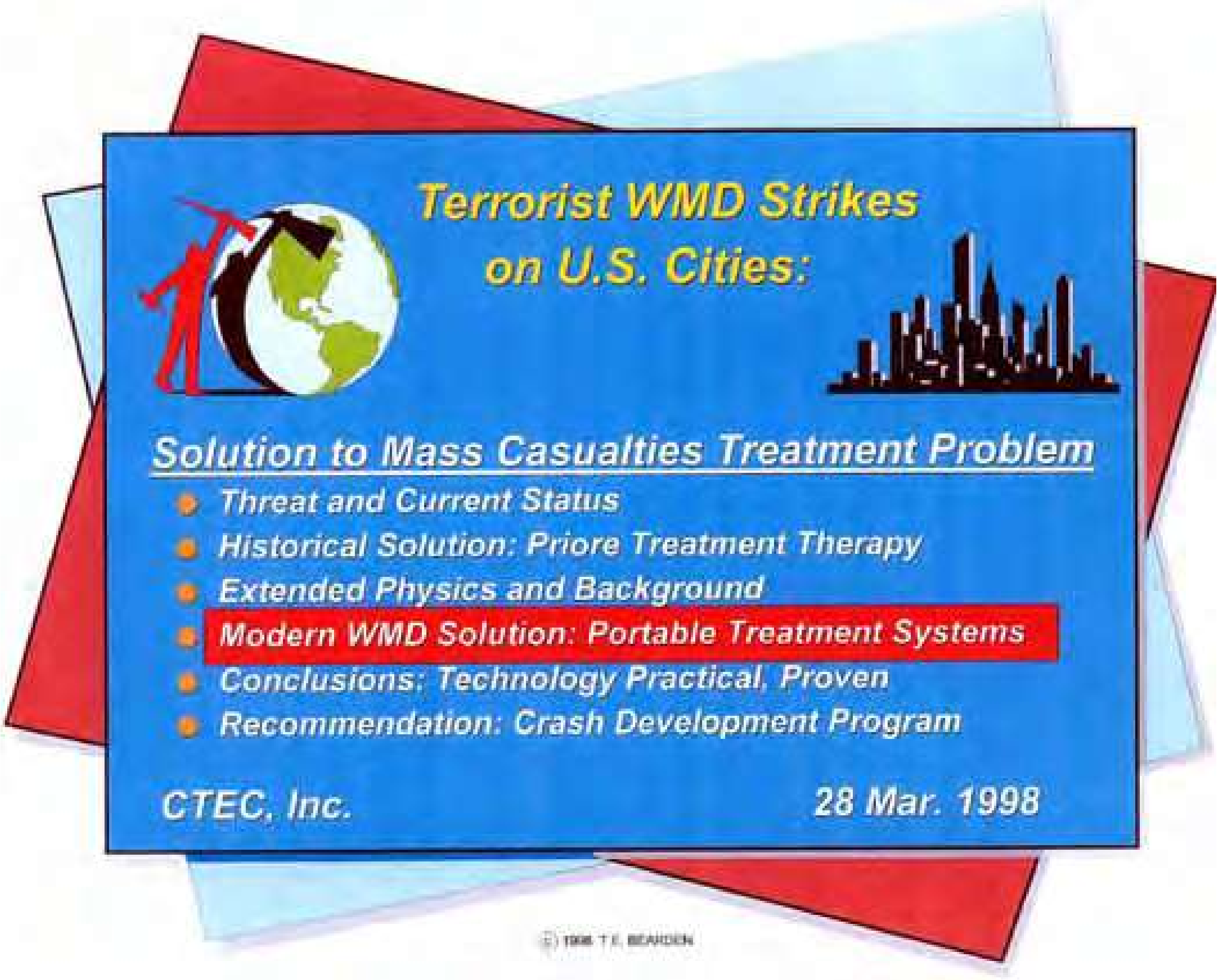

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

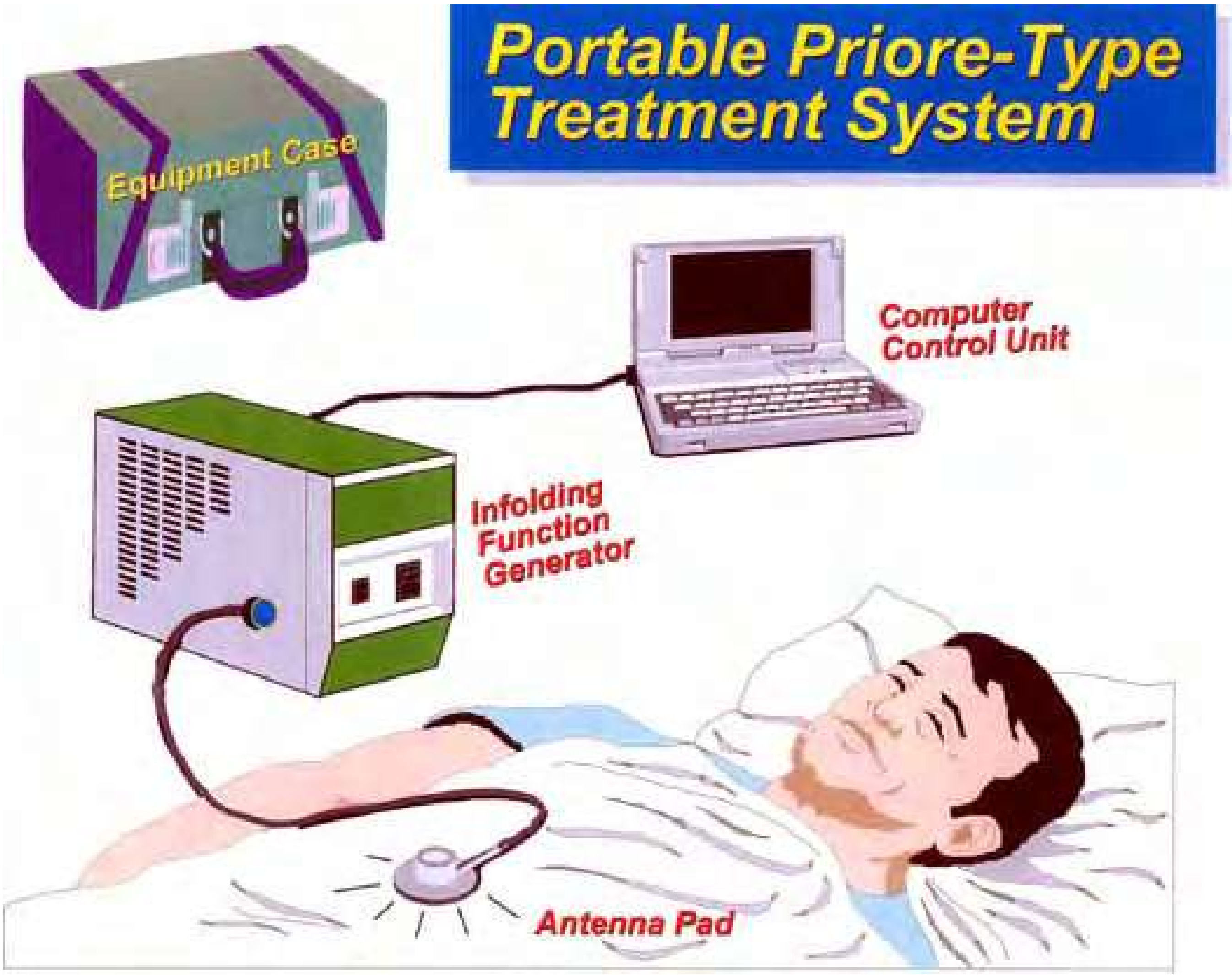

Portable Priore-Type Treatment System

Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Return to Slide Index

Next Slide

Previous Slide

Portable System Block Diagram

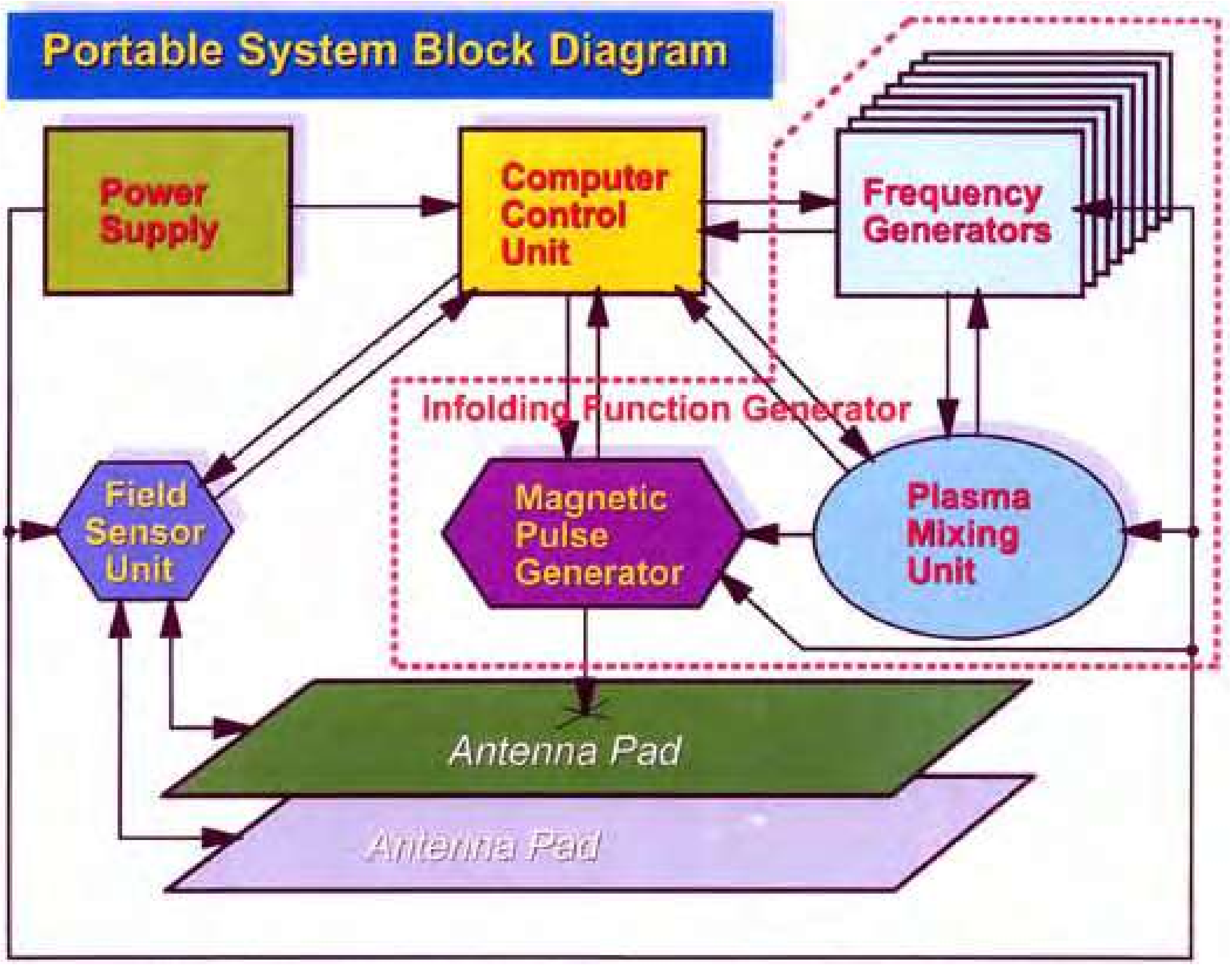
Power Supply
Computer Control Unit
Frequency Generators
Infolding Function Generator
Field Sensor Unit
Magnetic Pulse Generator
Plasma Mixing Unit
Antenna Pad
Antenna Pad

[Return to Slide Index]

[Next Slide]

[Previous Slide]

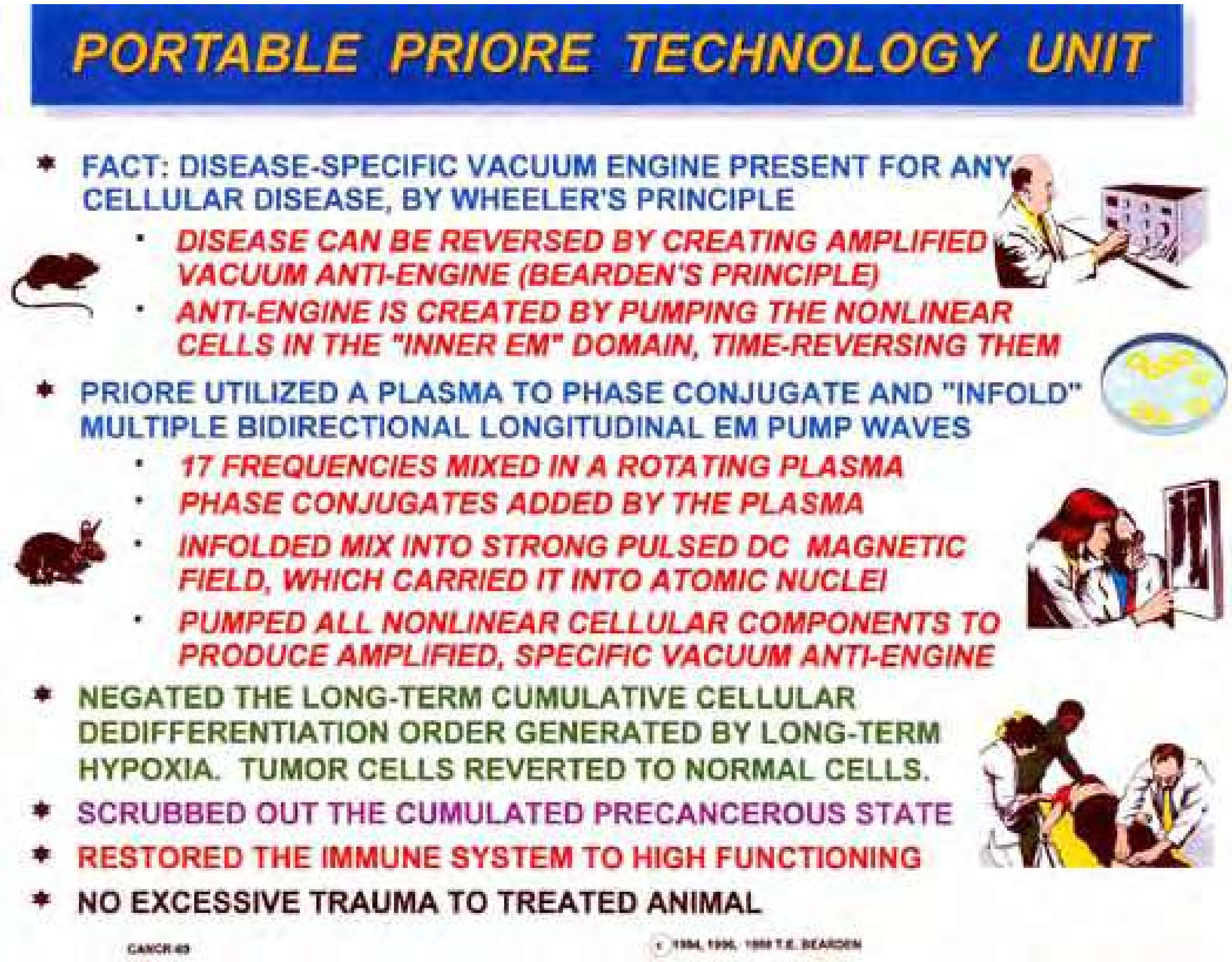

PORTABLE PRIORE TECHNOLOGY UNIT

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

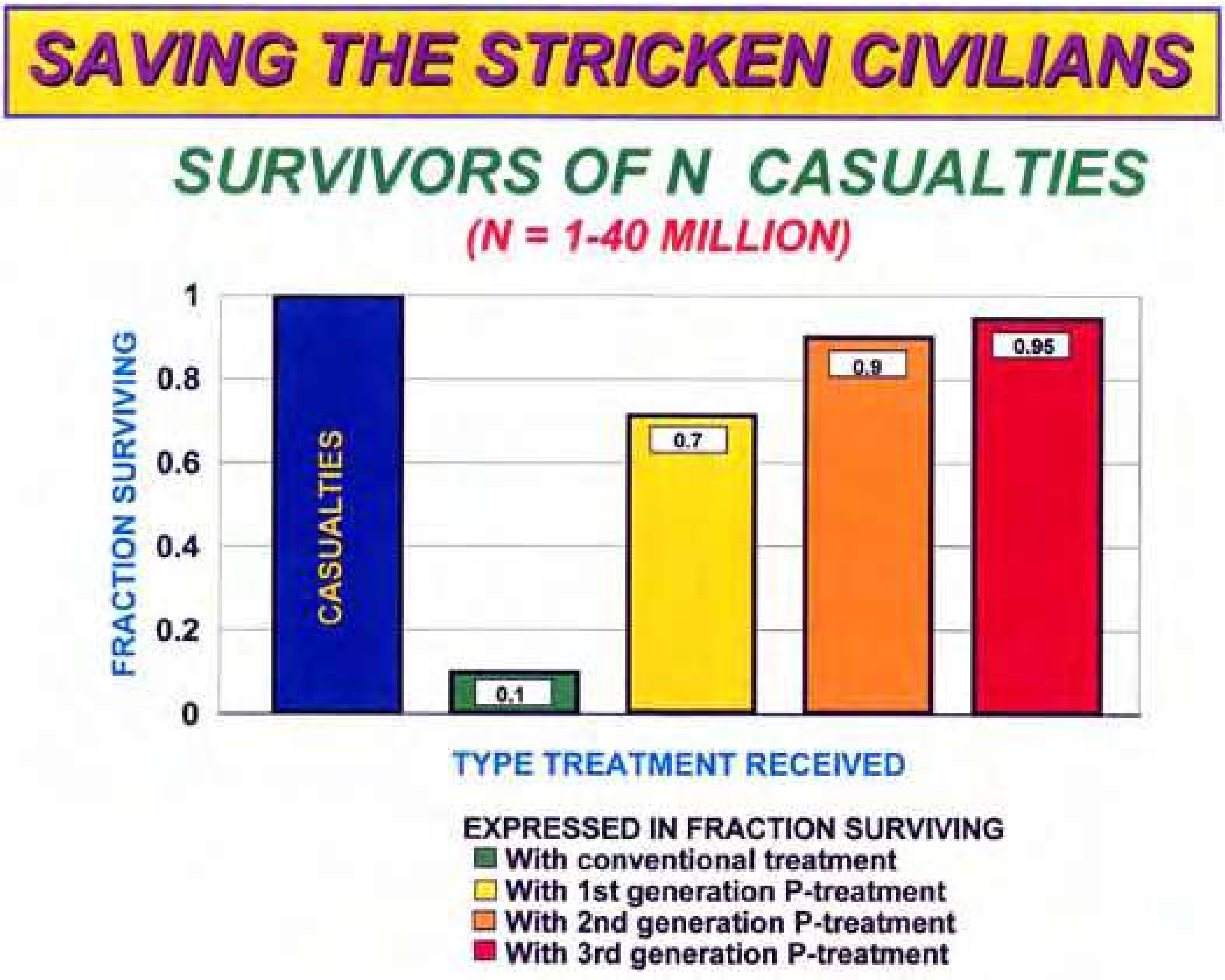

SAVING THE STRICKEN CIVILIANS

SURVIVORS OF N CASUALTIES

(N = 1-40 MILLION)

EXPRESSED IN FRACTION SURVIVING

- 🟩 With conventional treatment
- 🟨 With 1st generation P-treatment
- 🟧 With 2nd generation P-treatment
- 🟥 With 3rd generation P-treatment

Return to Slide Index

Next Slide

Previous Slide

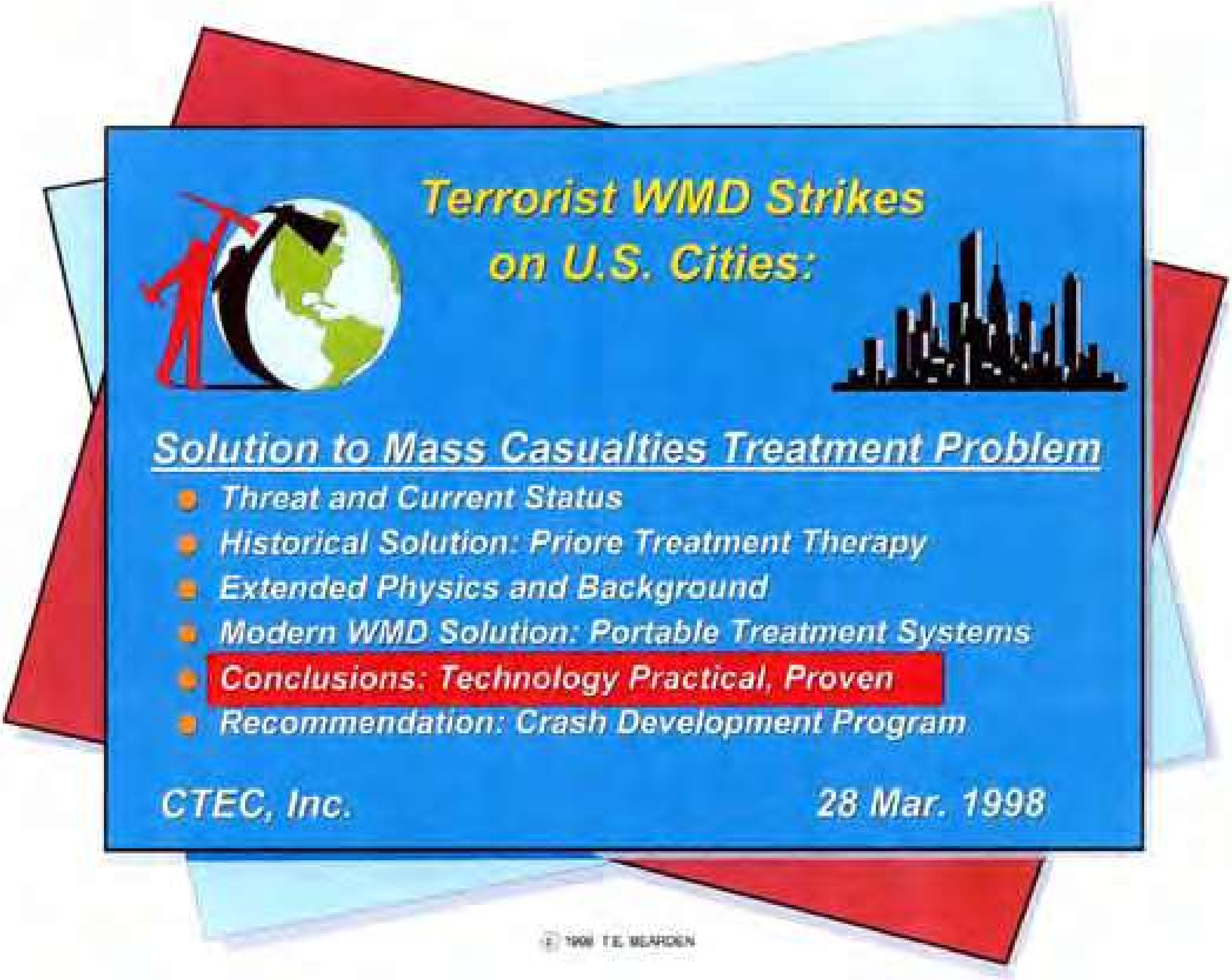

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc. 28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

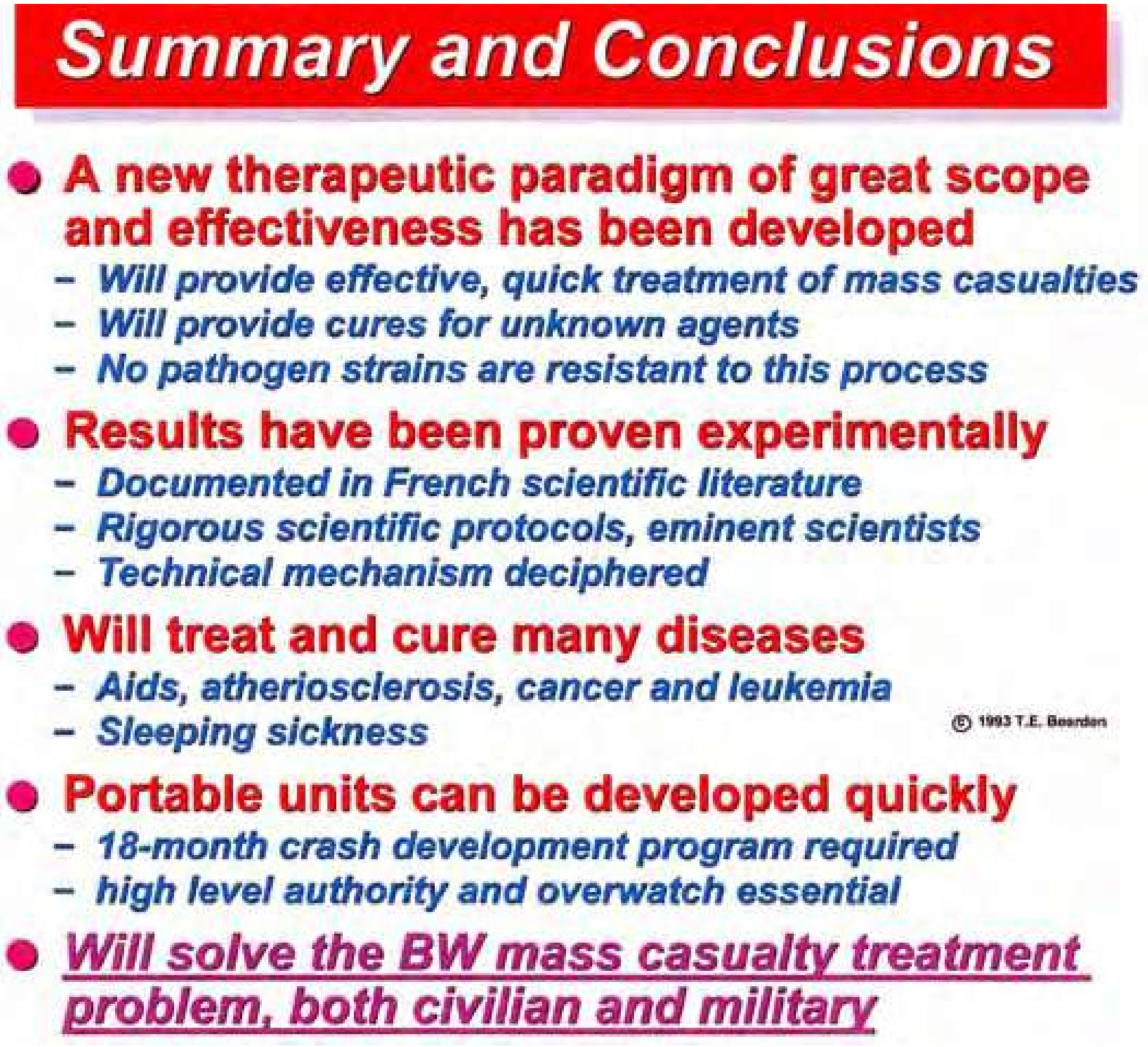

Summary and Conclusions

- A new therapeutic paradigm of great scope and effectiveness has been developed
  - *Will provide effective, quick treatment of mass casualties*
  - *Will provide cures for unknown agents*
  - *No pathogen strains are resistant to this process*
- Results have been proven experimentally
  - *Documented in French scientific literature*
  - *Rigorous scientific protocols, eminent scientists*
  - *Technical mechanism deciphered*
- Will treat and cure many diseases
  - *Aids, atheriosclerosis, cancer and leukemia*
  - *Sleeping sickness*
- Portable units can be developed quickly
  - *18-month crash development program required*
  - *high level authority and overwatch essential*
- *Will solve the BW mass casualty treatment problem, both civilian and military*

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

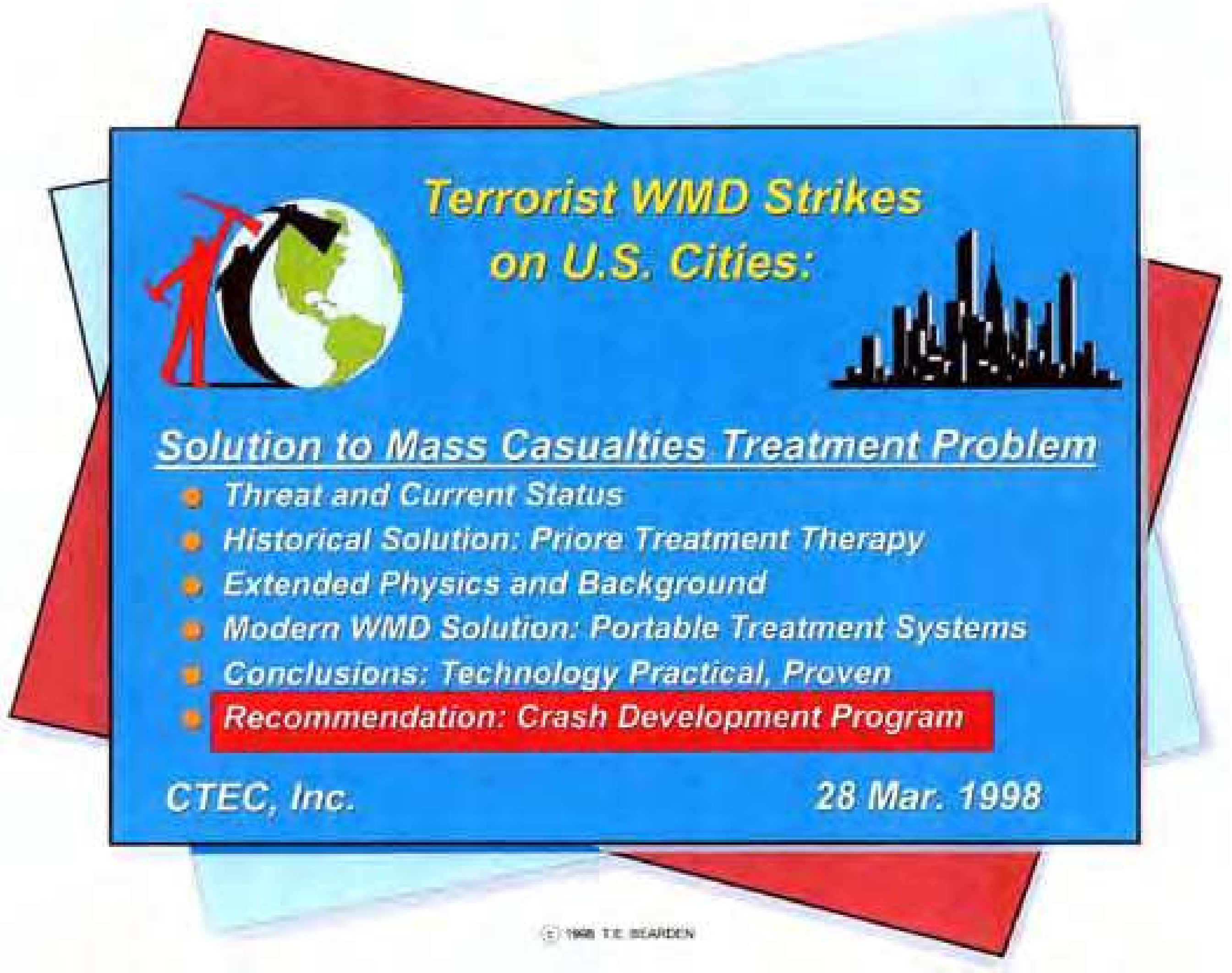

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

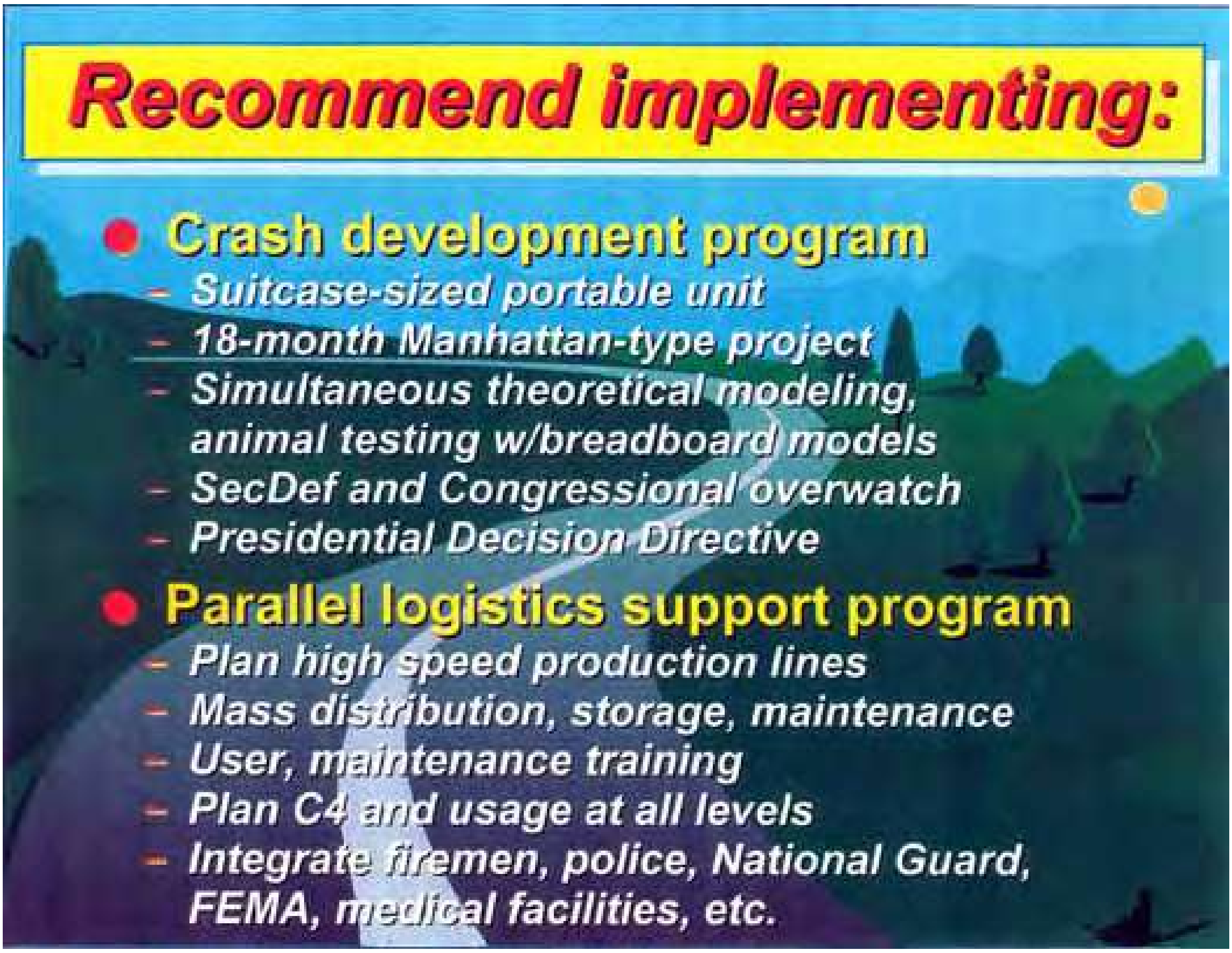

Recommend implementing:

- Crash development program
  - Suitcase-sized portable unit
  - 18-month Manhattan-type project
  - Simultaneous theoretical modeling, animal testing w/breadboard models
  - SecDef and Congressional overwatch
  - Presidential Decision Directive
- Parallel logistics support program
  - Plan high speed production lines
  - Mass distribution, storage, maintenance
  - User, maintenance training
  - Plan C4 and usage at all levels
  - Integrate firemen, police, National Guard, FEMA, medical facilities, etc.

Return to Slide Index

Next Slide

Previous Slide

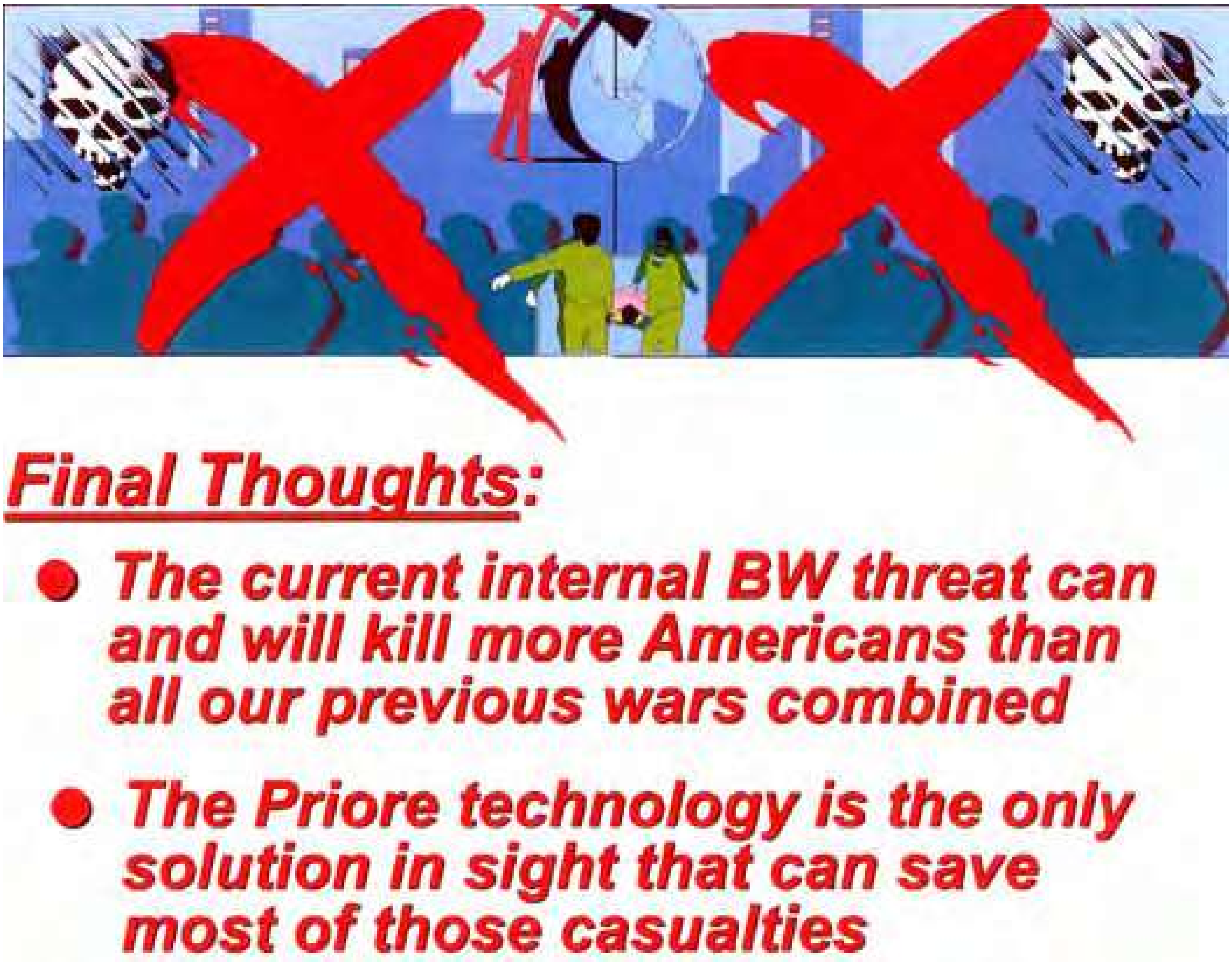

Final Thoughts:

- The current internal BW threat can and will kill more Americans than all our previous wars combined
- The Priore technology is the only solution in sight that can save most of those casualties

Return to Slide Index

Previous Slide

CTEC, Inc.
2311 Big Cove Road
Huntsville, AL 35801
(256) 533-3682 Ph/
(256) 539-5951 Fax

CTEC, Inc.

June 24, 1998

Dr. Jane F. Kinsel, Director
Office of Policy Analysis
Department of Health & Human Services
Bathesda, MD 20892

Dear Dr. Kinsel:

Your letter of June 8 is appreciated. However, the entire action package had already gone to the Assistant SecDef and also to Major General Neary in the USAF.

The proof of a revolution in medical treatment was in the package. The basis for it is already demonstrated experimentally in the scientific literature, both in the U.S. and in France. We are speaking of something already proven, but missed by the entire U.S. medical community. Strong words, but true. Let me be specific.

Any cell in the body can be time-reversed (physics term) or dedifferentiated (biology term) back to an earlier state, by pumping (i.e., "squeezing" in simple terms) that cell with longitudinal EM waves as pump waves. There are considerable papers in the French scientific literature showing the experimental results.

Apply that to cancer. We are stating bluntly that all cancerous cells in the body can be "depromoted" back to normal cells easily. Further, the damage previously done to them (such as by sustained hypoxia due to contaminants reducing the oxygen-carrying ability normally given to it by some 60-to-80 surrounding water molecules per red cell and the consequent H-bonding interactions) will also be reversed. The cells can all be returned to healthy, normal cells again, without "killing" or "burning" or "cutting" anything.

With some development of the technology, you can quickly and easily cure every cancer patient in every hospital in the United States.

Now apply it to AIDS. Since you can time-reverse every cell in the body, you can time-reverse the HIV-infected cells – genetics and all – right back to normal cells with normal genetics. Those cells not infected and healthy, will just get a "little younger" again.

So with that small development, you can quickly and easily cure every AIDS patient in every hospital in the United States.

I would have thought that NIH would certainly be interested in such a revolutionary therapy which has already been demonstrated in laboratory experiments. World-renowned French scientists – such as the eminent parasitologist Pautrizel and research PhDs assigned directly by Robert Courrier, head of the Biology Section of the French Academy of Sciences, worked with Prioré in performing those startling experiments that proved what we are saying.

The method was rigorously shown to cure infectious diseases (such as trypanosomiasis), atherosclerosis (which now afflicts millions of Americans), and terminal tumors. It was shown to reverse suppressed immune systems back to normal vigorous functioning.

Please permit me to disagree with you that the examination of such is not the mission of the NIH. To the contrary, it is precisely the NIH which should have some of its very best scientists looking into this.

The action package to DoD was to propose a crash program to develop this methodology in the form of a small, portable unit for treatment of mass casualties resulting from a terrorist BW attack on our cities (now expected in the next few weeks, as you must be aware). The leading Arab terrorist of the world has already promised just this.

We are speaking of several million Americans who are going to die. They are going to die in spite of everything that NIH and its labs can do, even though I know you will make a heroic effort. With NIH present methods, you cannot do very much for those stricken Americans at all. Everything we have in the "normal" kit bag is almost useless against a knowledgeable spray attack of modified smallpox (as you know, the Russian secret BW labs have made and sold tons of that already, to guess who!). Or a spray of tularemia (you are well aware of the extreme lethality of that), or anthrax (easily obtained). It is also common knowledge that these terrorist teams with their BW agents are already on site in this country, waiting for the word to attack. This has now been officially raised to our primary Strategic Threat.

In the face of such looming strikes on the U.S., I simply cannot believe that the NIH does not feel it within their mission to rigorously check out a proposed method for saving perhaps 70% of those coming deaths (first generation equipment), and up to 90% with second generation equipment.

Perhaps you just did not read the package carefully. I urge you to do so.

Else NIH is going to look very, very bad when (1) the strikes do occur, (2) they can't handle them (and they cannot, as you well know), (3) they had within their hands in advance, a proposed methodology that could have been investigated intensely, to save millions of those stricken Americans.

As a staunch supporter of NIH and especially CDC, I have to believe that there is a different attitude there. If NIH is not interested in a previously experimentally demonstrated method of saving millions of Americans, then the only recourse left is the political channel.

I assure you that my critique of electrodynamics is well-founded, as supported by various leading physicists such as Nobelist Feynman (who stated bluntly that the field concept is wrong), Wheeler, Bunge, and so on. It is inexplicable why all our universities are still teaching a 130 year old abridged (Heaviside's) subset of Maxwell's theory, and one that still contains a material ether even though that was falsified 110 years ago!

You already have seen the misuse of this technology by the Russians, in the decades-long radiation of the U.S. Embassy in Moscow. Is it not NIH's mission to understand how they generated all those diseases and health changes? Have you ever looked into Kaznacheyev's experiments showing that ANY cellular disease or disorder can be induced in cells at a distance by purely EM means? Have you looked into what got eliminated from electrodynamics by ignoring the hidden longitudinal EM inside every potential, field, and wave? Do you realize what ignoring that has done to American chemistry, biochemistry, and medical science?"

I would hope that the NIH would at least be interested in its own mission area!

I challenge you bluntly to admit that no one presently understands the primary operation of the human cellular *regenerative* system. That's because its technical principle is not presently in our recognized physics, biochemistry, and medical science – even though it has been in the hard physics literature since 1903-4. Becker's epochal work came closest; since he only had access to the conventional EM model, it failed him. But he at least showed you rigorously that cells can be changed (both differentiated and redifferentiated) all over the map by laughably weak EM – picoamperes of current, which means persistent DC potentials. In the microwave radiation of the U.S. Embassy, all the health changes occurred in regions absent of fields – i.e., absent of bleed-off of the potentials. Hence in regions of stable DC potentials! With 100% correlation of the health changes to persistent, gradient free potentials, and 100% anticorrelation to changing potentials (i.e., fields), our own scientists concluded totally erroneously that it could not have been the EM radiation causing the changes, when their own results had just proven that it was.

Now I ask you: Can NIH explain how it was that the DC potentials caused those health changes? I think not. If it could, it would understand the action package I wrote, and its importance.

This is not a "political" letter. I'm just an ordinary American citizen who has done about 30 years very hard work on this particular area. The work is good enough now to stand on its own merits. The scientific community will eventually accept it, about 50 years from now. But we cannot wait 50 years! Those millions of Americans are going to die, because of the bureaucracy and inaction of our own scientific organizations. Whether they like it or not, the U.S. scientific community is now the forefront of the struggle for this nation to survive. And they are failing us dramatically.

So as a parent who wants his children to live, and his neighbors' children to live, and those millions of soon-to-be stricken Americans to live also, I urge you to at least have someone competent in physics (and knowledgeable that the foundations of physics itself are in woeful shape!) check the work. Has anybody there *read* any of the references to the Prioré work I cited? Does anyone there realize that *there are no EM forcefields in space?* Probably not.

Dr. Kinsel, there is a time for "business as usual" and there is a time to do some serious new work. Those stricken Americans are going to be lying there. Nothing NIH can do is going to stop that. That's a given, and it is not my own estimate. It's the official estimate. And you *know* you presently do not have the tools to do anything really effective about it.

Frankly, I interpret your letter to me as just the normal "spin control" for easy disposition. I would hope that NIH would seriously rethink the situation, get off the "spin control" and "business as usual" posturing, and do something dramatically different for a change.

What have you got to lose? A little time from perhaps two or three of your best scientists.

What have you got to gain? Saving untold millions of American lives. A new, *already partially demonstrated* cure for most dread diseases such as AIDS and cancer – diseases which you presently cannot do very much about in spite of decades of heroic struggle.

Sincerely,

cc: Senator Shelby, Congr. Cramer

T.E. Bearden
President and CEO

CTEC, Inc.
T.E. Bearden, President & CEO
2311 Big Cove Road
Huntsville, AL 35801-1351
(256)533-3682 ph/(256)536-0411 Fax

CTEC, Inc.

May 12, 1998

Dr. Harold Varmus, Director
National Institute of Health
Bathesda, MD 20892
(301) 496-1766

Dear Dr Varmus:

We have recently sent action documents to the Director of Nuclear and Counterproliferation, Office of the DCS, Air and Space Operations, HQ USAF and to the DoD (General Busbee) Assistant Secretary of Defense, as well as several Senate and House committees, which may be of interest to you. For one thing, we give the technical mechanism used to generate the Gulf War Syndrome. The same weaponry is now being used to rapidly induce new strains of certain pathogens worldwide. I urge you to closely read that section of the letter to MG Neary which explains the GWS, why the French did not get it, and why Southern Iraqis got an increase in cancer and leukemia but not GWS. Also explained is how the "microwave radiation" of the U.S. Embassy in Moscow induced diseases for decades, and how we failed to comprehend it because the standard EM model deceived us.

There is a way to treat and quickly cure most diseases, including the mass casualties resulting from a terrorist BW strike on our civilian population centers, our military bases, or our forces in the field. The method is little known, but was scientifically demonstrated in France in the 1960s and early 1970s before cancellation of the Prioré Project when the French Government changed. Eminent French scientists worked on the project, and the results are fully documented in leading French scientific journals. References are listed in a separate attachment.

Previously the Western scientific community has been unable to understand the technical mechanism responsible for such remarkable cures, because of serious foundations flaws in Western physics and electrodynamics. *Foundations* physicists have long been aware our science is flawed, and have pointed out many of these fundamental errors. However, the operational science establishment has not paid attention, even though physicists such as Feynman and Wheeler pointed out that the notion of "force fields in space" was totally wrong.

As a primary example, there is an "infolded" electrodynamics inside all potentials, fields, and waves that is far more primary than the coarse EM that is in our textbooks. It is actually an infolded general relativity, and this is what the Russians have weaponized. This infolded EM-GR has been ignored in the West (but not in Russia) since 1903.

As another example the potentials, fields, and waves of present electrodynamics are rigorously defined – and exist – only in the presence of mass. They do not exist in such form (force fields and oscillating force fields) at all in space, in the absence of mass. Rigorously, Maxwell's equations are mass-to-mass transforms, since Maxwell and everyone at the time assumed the ubiquitous presence of the material ether. In other words, to the founding electrodynamicists, there was no place in all the universe that was devoid of mass. Hence they defined all EM entities as entities containing mass. Even though Michelson-Morley experiments of more than a century ago destroyed that material ether, not a single Maxwellian equation has ever been changed! They still assume the material ether.

Specifically missing from our electrodynamics are two essential transforms: (1) the mass-to-spacetime transform, and (2) the spacetime-to-mass transform. In other words, these two transforms are "infolded" or hidden inside (comprise) the end point extremum case – the mass-to-mass transform actually written by Maxwell.

But these two missing transforms are just general relativity (they correspond precisely to Wheeler's principle of general relativity). So general relativity has always been infolded inside electrodynamics as we know it, but ignored. Further, this is a very powerful general relativity, since it uses the very strong EM force as the agent of spacetime curvature. Hence now one can engineer a powerful general relativity in the laboratory and in practical devices, and even action-at-a-distance, which one cannot do using merely the staid and weak G-force.

It turns out that longitudinal EM waves as shown by E.T. Whittaker in 1903 and 1904 are far more primary than our present EM theory and even our present GR theory. Both EM and GR are directly engineerable using longitudinal EM waves, including subluminal and superluminal EM waves in vacuum, and including action at a distance. One can engineer complexes of spacetime curvatures as desired, which in turn are *spacetime engines* (Wheeler's term). These engines can act upon mass – either living or inert – in any fashion desired, and at any distance desired.

The little-known and poorly understood Regeneration & Recovery (R&R) system – as contrasted to the immune system – utilizes this exact "spacetime engine" methodology to restore damaged or diseased cells, within its limitations – "dedifferentiating" or "time-reversing" them back to normal. The mechanism is fully explained in the attached package. The proposed program applies this fundamental method to greatly amplify the effect. Thus cancer cells can be reversed back to normal cells, HIV-infected cells (HIV-factories) can be reversed back to normal cells (genetics and all), etc. This represents a dramatic and unparalleled new medical therapeutic methodology of extreme power. The same techniques can simply be conjugated one additional time, and used to produce diseases and cellular changes at a distance, including in entire mass populations, as the KGB is doing right now. Specifically, by impressing an extremely weak "cocktail mix" of disease spacetime engines on the U.S. populace, the immune systems are reacting and being "spread thin" across several "shadow state" infectious pathogenic conditions. Thus when an actual terrorist BW attack occurs, the lethality of the strike will be remarkably enhanced because the immune systems will be "thinned" in their ability to fend off this new pathogen. By doing it this way, BW warfare can actually be conducted upon the U.S. without anyone being the wiser.

Our very survival as a nation is threatened by remarkable KGB weapons using this fully documented, extended electrodynamics. Bluntly, unless we take off our scientific blinders and understand Russian energetics weapon science, GWS and the Embassy health changes in Moscow were just a picnic before what is upon us. After two decades of work on the KGB energetics weapon problem, we have fully deciphered the technical mechanisms and what energetics is, as well as citing many of the Russian weapon tests and how they were done. Again, you can see that for yourself in the package.

Now that we understand the technical mechanisms, we are proposing rapid redevelopment of the Prioré process, in portable suitcase-sized units to be mass produced and filtered down into the entire emergency community for mass treatment of lethal infectious diseases resulting from BW strikes. And we are strongly urging the U.S. scientific community to quickly investigate and come to grips with this vast new *biological engineering* mechanism. Our very survival depends on it.

The inclosed documents provide an overview of the process. There is nothing else on the scientific horizon that can save the majority of all those stricken Americans, once the first foreign-sponsored terrorist teams strike our population centers with BW agents.

Sincerely,

Incls: Full package with table of contents

T.E. Bearden
President & CEO

Weapons Slides

The Tom Bearden Website

Help support the research

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts." *

Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Slides
(covers period up to 1999)

- The Coming Strategic Attack on the United States
- Outline
- Loss of two separate A-10 Warthogs in 1997
- Typical ABM Defenses (Widely Deployed by mid-70s)

- Launch Phase Anti-Missile System
- Launch Phase Anti-Bomber System
- HC-130s Engines Quit, Aircraft Crashes
  Nov. 22, 1996
- 1996 Actions by this analyst
- Additional Information Exists: Not Analyzed, Not
  Integrated in Briefing
- Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985
- Satellite Photos of Plumes in Soviet Arctic Since 1974
  See "Fer de Lance" for additional information
- Aum Shinrikyo ("Supreme Truth" Cult), Shoko Asahara, Leader

  Aum Shinrikyo - Anatomy of a Dangerous Cult
- Ashkhabad Problem and Its Solution
- Strange Characteristics of Iran Quake, Sep. 1978
- Iran Earthquake, September 1978
- Situation Now (Assessment)
- Situation Now (Assessment 2)
- Aspects of Strong Local Asymmetry
- Mass Casualty Problem: Aerial Anthrax

Spray on Washington, D.C.

- Present Status of the WMD Threat to the United States
- Some Present Problems in the U.S. WMD Defense Capabilities
- Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future
- Attempt to Ban Energetics Weapons Worldwide
- Brezhnev's Schedule
- Large glowing Tesla globe witnessed in Red China by hundreds
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  Sequence of Events
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  What They Were
- Cold Explosion -- 9 April 1984 Near Kurils (Plus additional weapons effects)
- Countering Insane MAD Systems:
  Find Loophole, Exploit It
- Next Generation Counter to Both Sane and Insane MAD systems
- Simultaneous Dudding of Nuclear Weapons Worldwide

- Simultaneous Dudding of Nuclear Weapons Worldwide
- Operation Divine Wind II
- Advanced EMBW Applications
- EM Biological Warfare (EMBW)
- Microwave Radiation of U.S. Embassy in Moscow
- Pentagon EMI Studies, 1989
- EM Missile Practice (6 slides)
- Subdivisions of Soviet Energetics Program
- Communism's Goal
- Manuilsky's Goal vs. Status Today (Dec. 1996)
- Professor Hellman, German Scientist Released from Work in Soviet Union
- Hemispheres and Globes of Light (4 slides)
- How Could Such Superweapons Ever Be Concealed?
- Soviets Easily Concealed These Superweapons for Four Decades
- Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes
- Longitudinal Wave Interferometry: Endothermic

- Longitudinal Wave Interferometry: Exothermic
- U.S. Air Attack on Libya, April 1986
- Lightning Wall
  circa 1959-1960
- Lisitsyn's Report: Brain Code Broken
- Large Longitudinal Wave Interferometer (LWI) Accident
- Large Longitudinal Wave Interferometer (LWI) Accident
- Use of Hidden Information Content of the Field Can Provide Action-At-A-Distance
- Two Types of Mutual Assured Destruction (MAD) System
- Psychoenergetics Weapons Teams Mentally Disable All Personnel
- Strikes Against Special Ship and Energetics Weapons Sites
- "Mindsnapper" Attack on Special Ships and Energetics Weapons Sites
- How Much Off Guard Are We?
  (7 Slides)
- Possible Targets: 1995-9 Escalation Phase
- Propagation of Weapon Effects
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15th, 1979

- Countering an Insane QP Weapon
- Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases
- Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
- Mechanism for Producing a Quantum Potential
- Major Principles for Use of Quantum Potential Weapons
- Major Principles for Use of Quantum Potential Weapons (2)
- Instantaneous Communication by a Quantum Potential
- Participants in a quantum potential share a common multiply-connected spacetime (MCST)
- Quake in Tangshan, China
  28 July, 1976
- Exchanging Earthquakes
- Instant Communication by a Quantum Potential
- Westerners Do Not React to Slow Threats
- Characterizing a Given Risk or Risk Factor
- Technical Risk Management
- Aum Shinrikyo's Sarin Facility

- In Scalar Beam Interference Zone
- Physicist Schappeller Auralaz Munster, Austria
- Initial Soviet Plan for Cold War, After WWII
- Josef Stalin: Absolute Dictator
- Exploding any Resistance by Operational Scientists
- Significant Soviet Statements
- Captain Svoboda Dives Headlong to Her Death
- T-polarized Beam Projector
- Possible Tests (1)
- Possible Tests (2)
- Possible Tests (3)
- Possible Tests (4)
- Mushroom Cloud Rising From Sea
- Launches from Cape Canaveral and Vandenberg AFB Provided Practice Targets of Opportunity
- The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth
- Embodiment of an instrument array sufficient to discriminate when time-density charging has occurred, and that transduction

of time-charge decay has occurred.

- Time-Polarized EM Wave Interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone
- Transmutation of Radioactive Nuclei at a Distance
- Transmutation of Radioactive Nuclei
- Typical Tesla Shield, Northern Route of Japan
- A Curious Tide in Human Affairs
- Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958
- The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
- The KGB/Yakuza-Aum War Plan
- Why Western Science is Ignorant of KGB Energetics
- Transmutation of Radioactive Nuclei at a Distance
- TheYakuza
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15, 1979

---

* Note that the Secretary did not speak the words in square blocks; these were added by the writers later,

probably to "soften the tone" of the Secretary's information release. So the Secretary confirmed that some nations of the world -- not just some ragged terrorists, as the writers tried to imply for spin control -- do possess novel electromagnetic weapons and are using them to induce earthquakes, engineer the weather and climate, and trigger volcanic eruptions.

THE COMING STRATEGIC ATTACK ON THE UNITED STATES

By an AUM/Yakuza/KGB/Communist Coalition Using Energetics SuperWeapons

T.E. Bearden, LTC
U.S. Army (Retired)
M.S. Nuclear Eng.
Fellow Emeritus, AIAS

With Technical Background and History of Energetics Weapons

Partial support by the Association of Distinguished American Scientists is gratefully acknowledged.d

© [illegible] T.E. BEARDEN

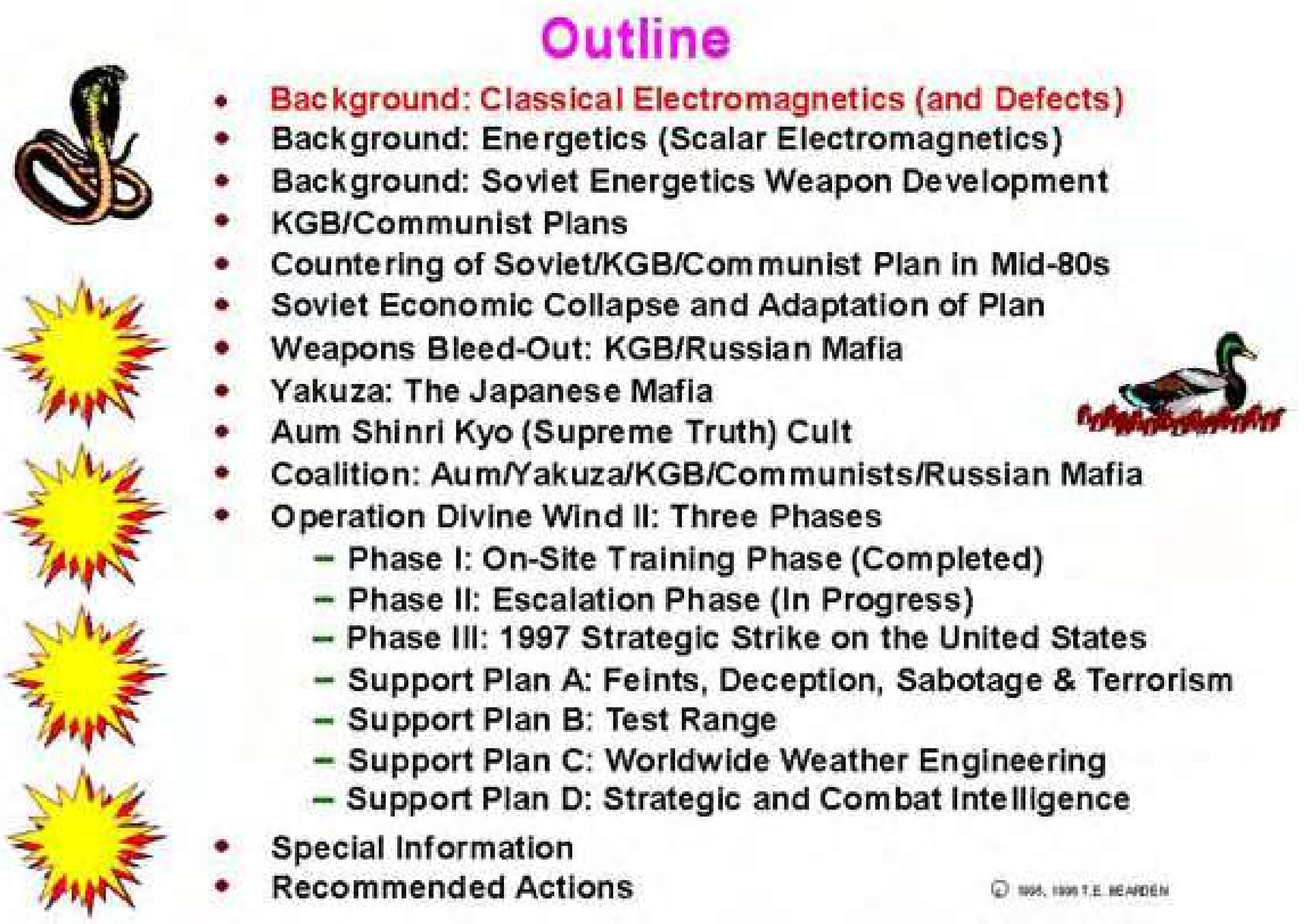

Outline

- Background: Classical Electromagnetics (and Defects)
- Background: Energetics (Scalar Electromagnetics)
- Background: Soviet Energetics Weapon Development
- KGB/Communist Plans
- Countering of Soviet/KGB/Communist Plan in Mid-80s
- Soviet Economic Collapse and Adaptation of Plan
- Weapons Bleed-Out: KGB/Russian Mafia
- Yakuza: The Japanese Mafia
- Aum Shinri Kyo (Supreme Truth) Cult
- Coalition: Aum/Yakuza/KGB/Communists/Russian Mafia
- Operation Divine Wind II: Three Phases
  - Phase I: On-Site Training Phase (Completed)
  - Phase II: Escalation Phase (In Progress)
  - Phase III: 1997 Strategic Strike on the United States
  - Support Plan A: Feints, Deception, Sabotage & Terrorism
  - Support Plan B: Test Range
  - Support Plan C: Worldwide Weather Engineering
  - Support Plan D: Strategic and Combat Intelligence
- Special Information
- Recommended Actions

© 1996, 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

Loss of two separate A-10 Warthogs in 1997

- A-10 Mysterious Flyaway and Crash - April 2, 1997
- Death of Second A-10 Pilot - May 27, 1997
- Captain Button's Mysterious Flyaway and Eventual Crash
- Death of Second A-10 Pilot - May 27, 1997

A-10 Mysterious Flyaway and Crash

April 2, 1997

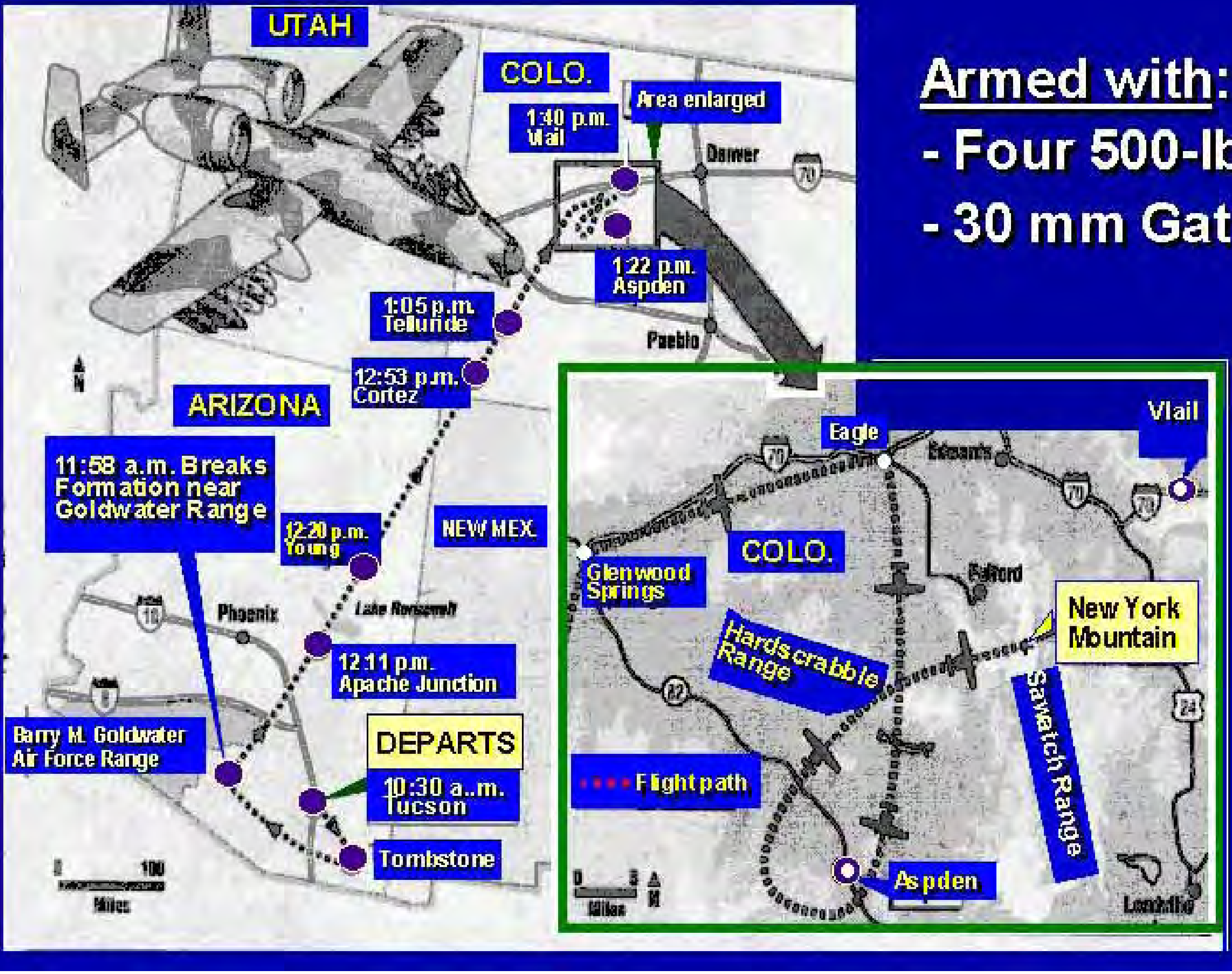
UTAH
COLO.
Area enlarged
1:40 p.m.
Vail
Denver
70
1:22 p.m.
Aspden
1:05 p.m.
Telluride
Pueblo
12:53 p.m.
Cortez
ARIZONA
N
11:58 a.m. Breaks
Formation near
Goldwater Range
12:20 p.m.
Young
NEW MEX.
10
Phoenix
Lake Roosevelt
12:11 p.m.
Apache Junction
8
Barry M. Goldwater
Air Force Range
DEPARTS
10:30 a..m.
Tucson
Tombstone
0
100
Miles
Vlail
Eagle
Edwards
70
70
COLO.
Glenwood
Springs
Gilford
New York
Mountain
Hardscrabble
Range
82
Sawatch Range
24
Flight path
0
5
Miles
N
Aspden
Leadville

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

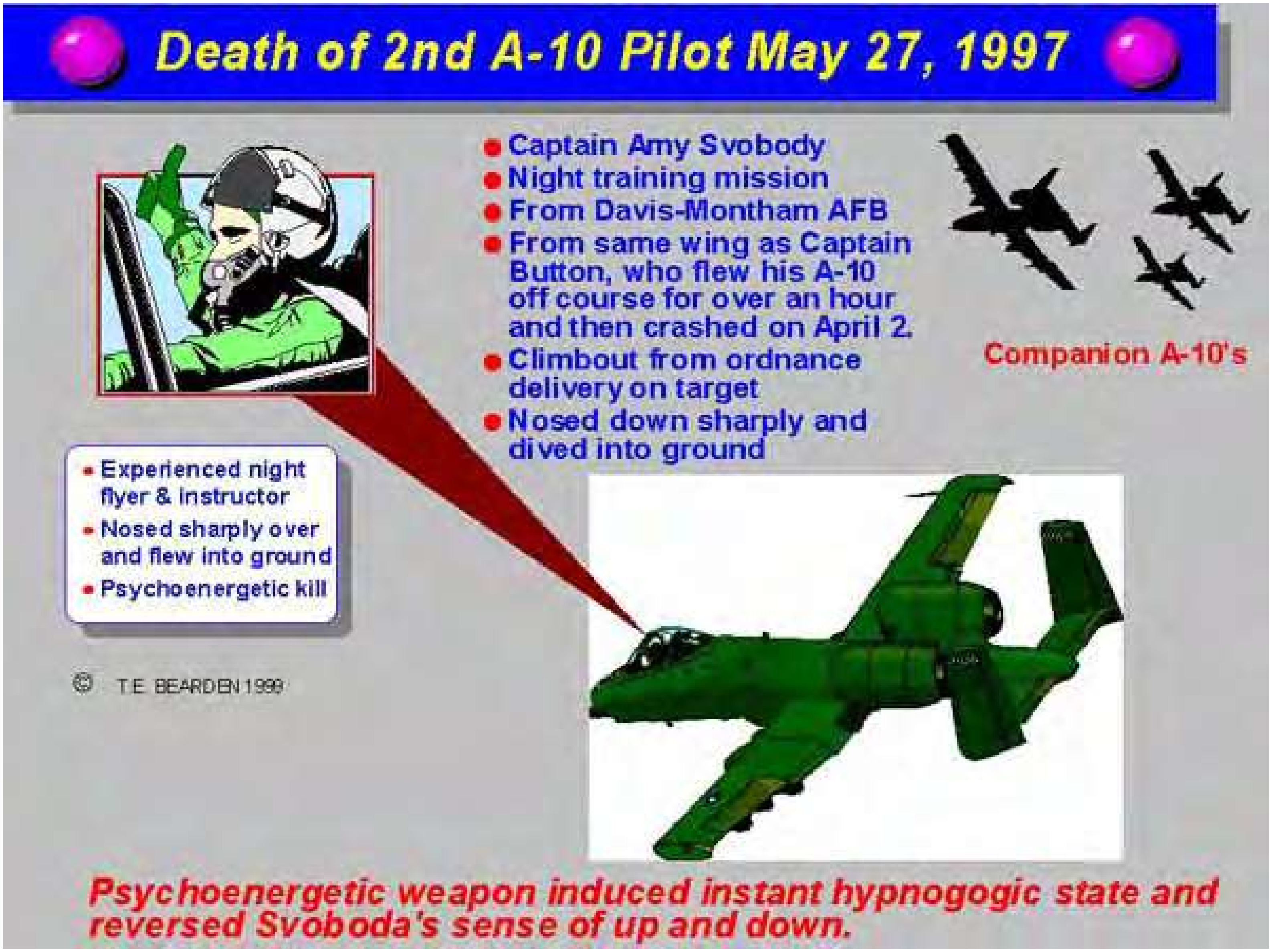

Death of 2nd A-10 Pilot May 27, 1997

- Captain Amy Svobody
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Climbout from ordnance delivery on target
- Nosed down sharply and dived into ground

Companion A-10's

- Experienced night flyer & instructor
- Nosed sharply over and flew into ground
- Psychoenergetic kill

© T.E. BEARDEN 1999

Psychoenergetic weapon induced instant hypnogogic state and reversed Svoboda's sense of up and down.

A-10 Mysterious Flyaway and Crash

April 2, 1997

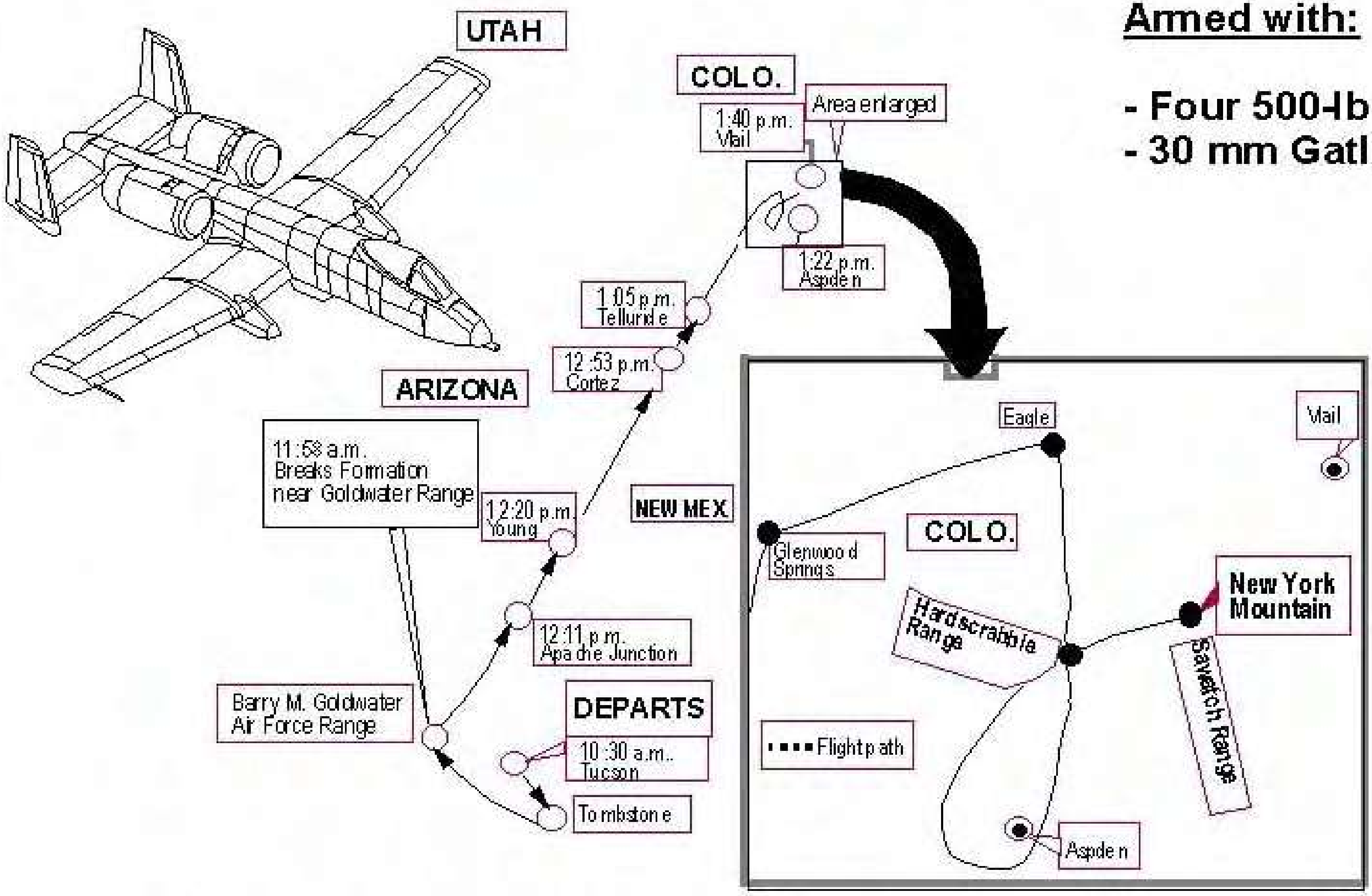
UTAH
COLO.
1:40 p.m. Vail
Area enlarged
1:22 p.m. Aspden
1:05 p.m. Telluride
12:53 p.m. Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
NEW MEX
12:11 p.m. Apache Junction
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a.m. Tucson
Tombstone
Eagle
Vail
Glenwood Springs
COLO.
Hardscrabble Range
New York Mountain
Sawatch Range
Flight path
Aspden

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

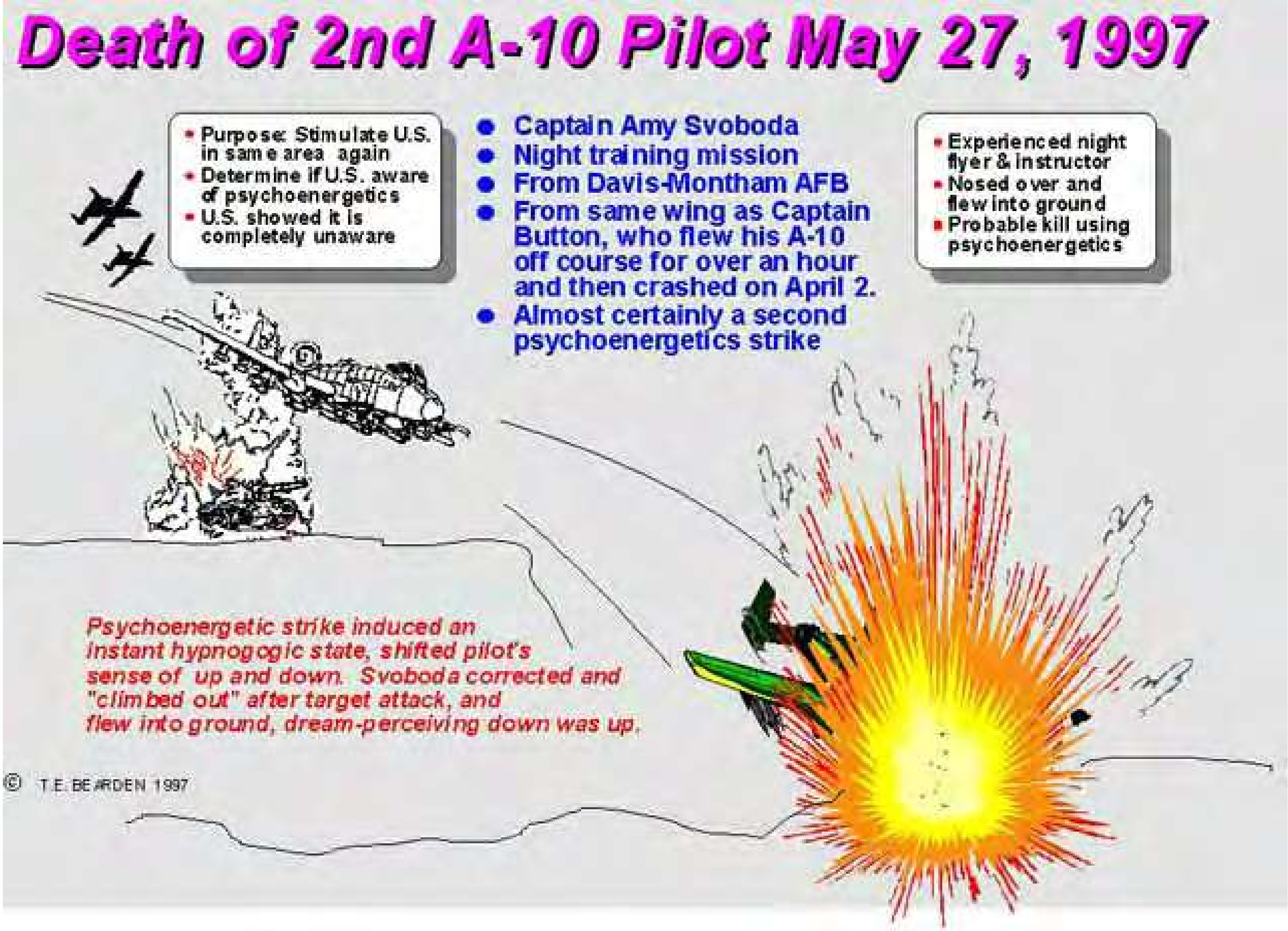

Death of 2nd A-10 Pilot May 27, 1997

- Purpose: Stimulate U.S. in same area again
- Determine if U.S. aware of psychoenergetics
- U.S. showed it is completely unaware

- Captain Amy Svoboda
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Almost certainly a second psychoenergetics strike

- Experienced night flyer & instructor
- Nosed over and flew into ground
- Probable kill using psychoenergetics

Psychoenergetic strike induced an instant hypnogogic state, shifted pilot's sense of up and down. Svoboda corrected and "climbed out" after target attack, and flew into ground, dream-perceiving down was up.

© T.E. BEARDEN 1997

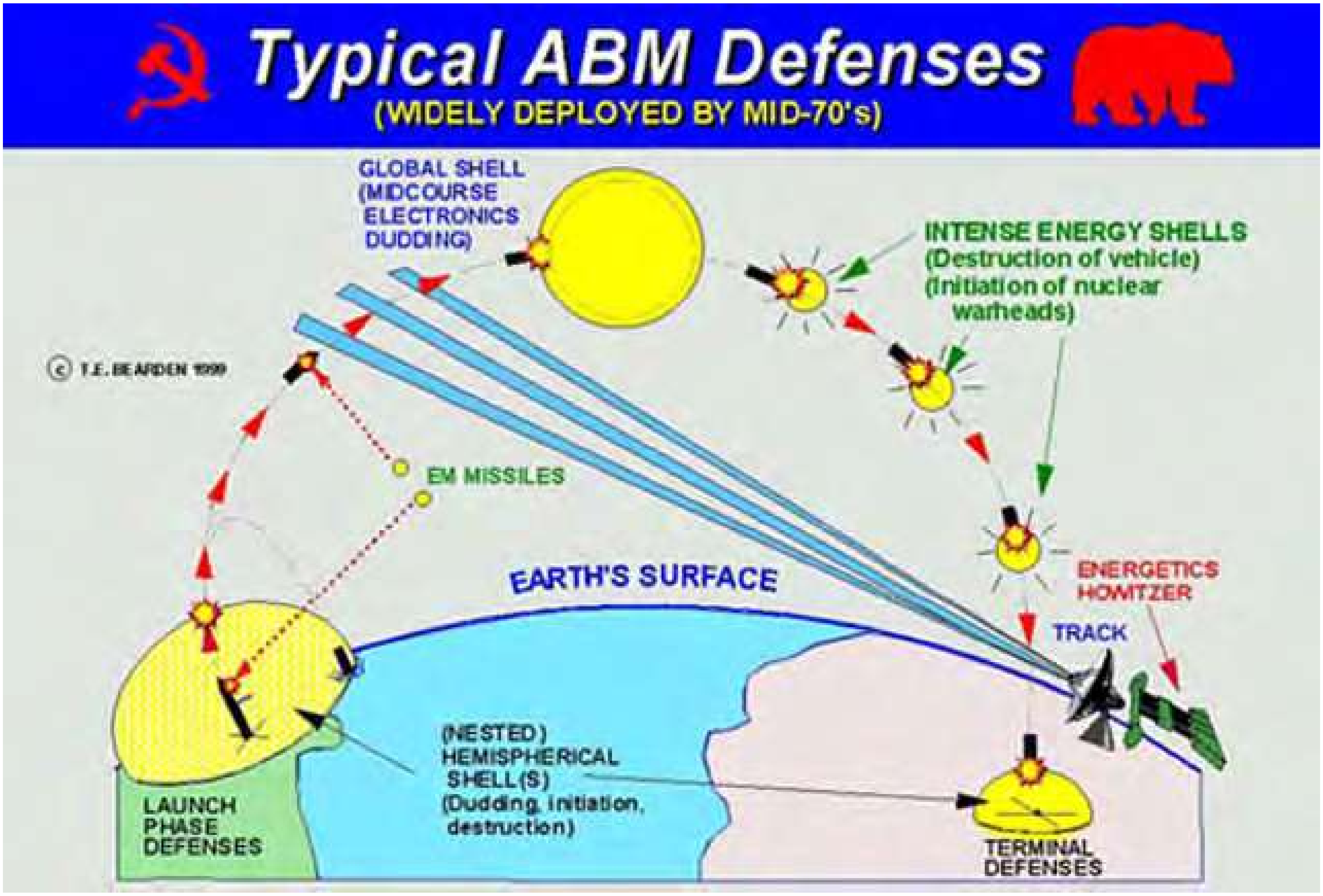
Typical ABM Defenses
(WIDELY DEPLOYED BY MID-70's)
GLOBAL SHELL (MIDCOURSE ELECTRONICS DUDDING)
INTENSE ENERGY SHELLS (Destruction of vehicle) (Initiation of nuclear warheads)
© T.E. BEARDEN 1999
EM MISSILES
EARTH'S SURFACE
ENERGETICS HOWITZER
TRACK
(NESTED) HEMISPHERICAL SHELL(S) (Dudding, initiation, destruction)
LAUNCH PHASE DEFENSES
TERMINAL DEFENSES

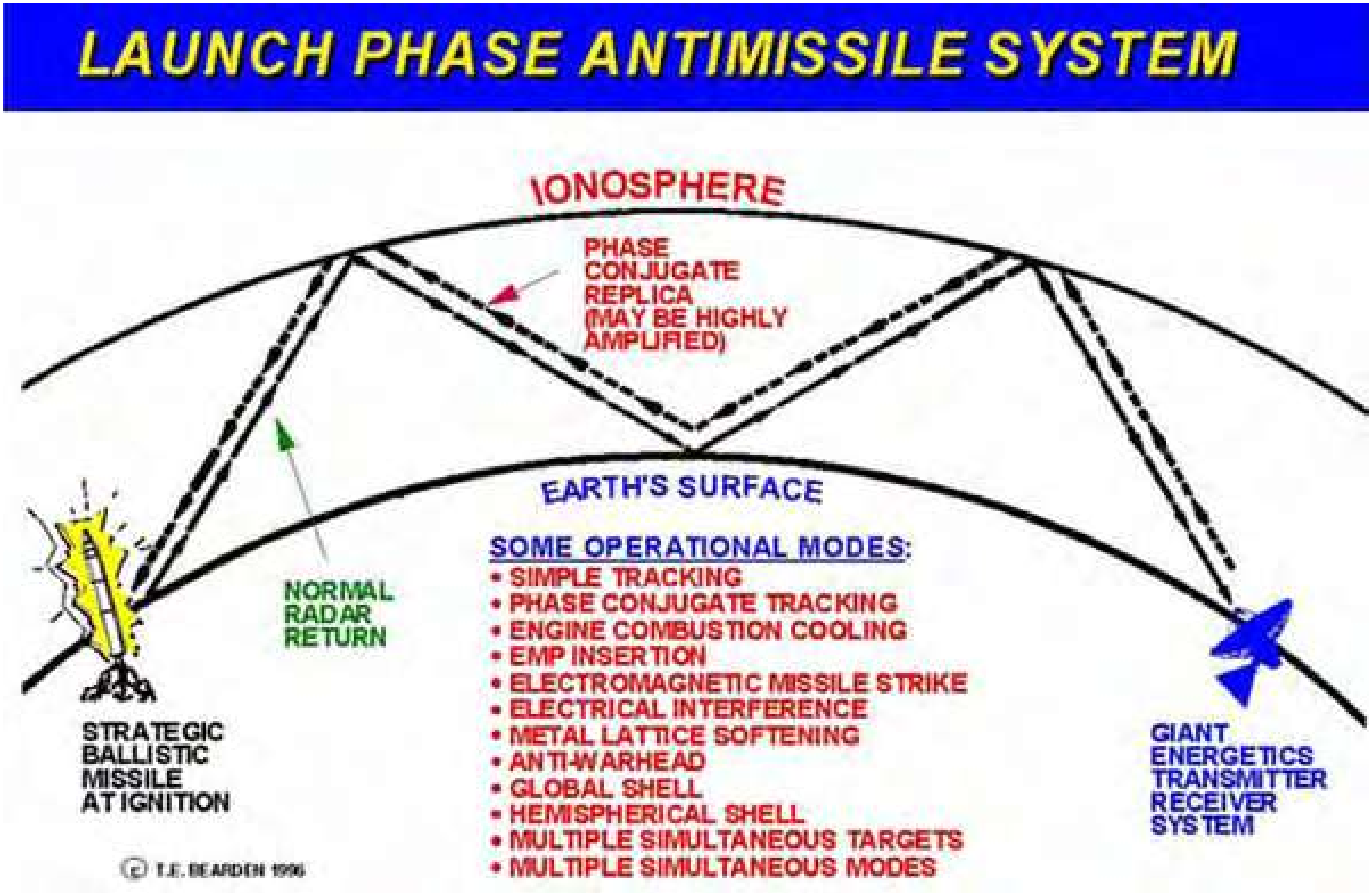

LAUNCH PHASE ANTIMISSILE SYSTEM

IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC BALLISTIC MISSILE AT IGNITION
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM
© T.E. BEARDEN 1996

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-WARHEAD
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

LAUNCH PHASE ANTI-BOMBER SYSTEM

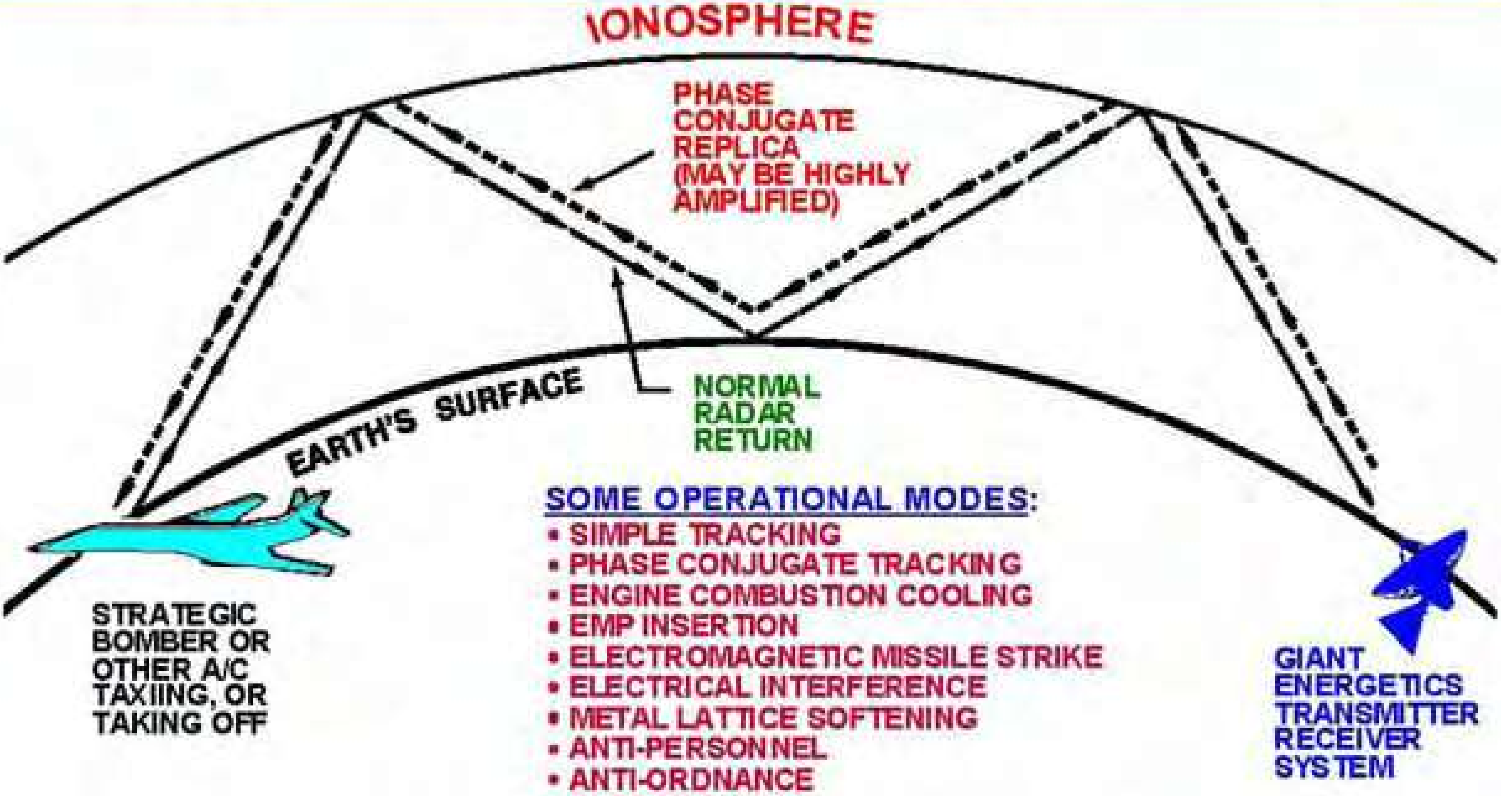
IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
NORMAL RADAR RETURN
EARTH'S SURFACE
STRATEGIC BOMBER OR OTHER A/C TAXIING, OR TAKING OFF
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-PERSONNEL
- ANTI-ORDNANCE
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

© T.E. BEARDEN 1996

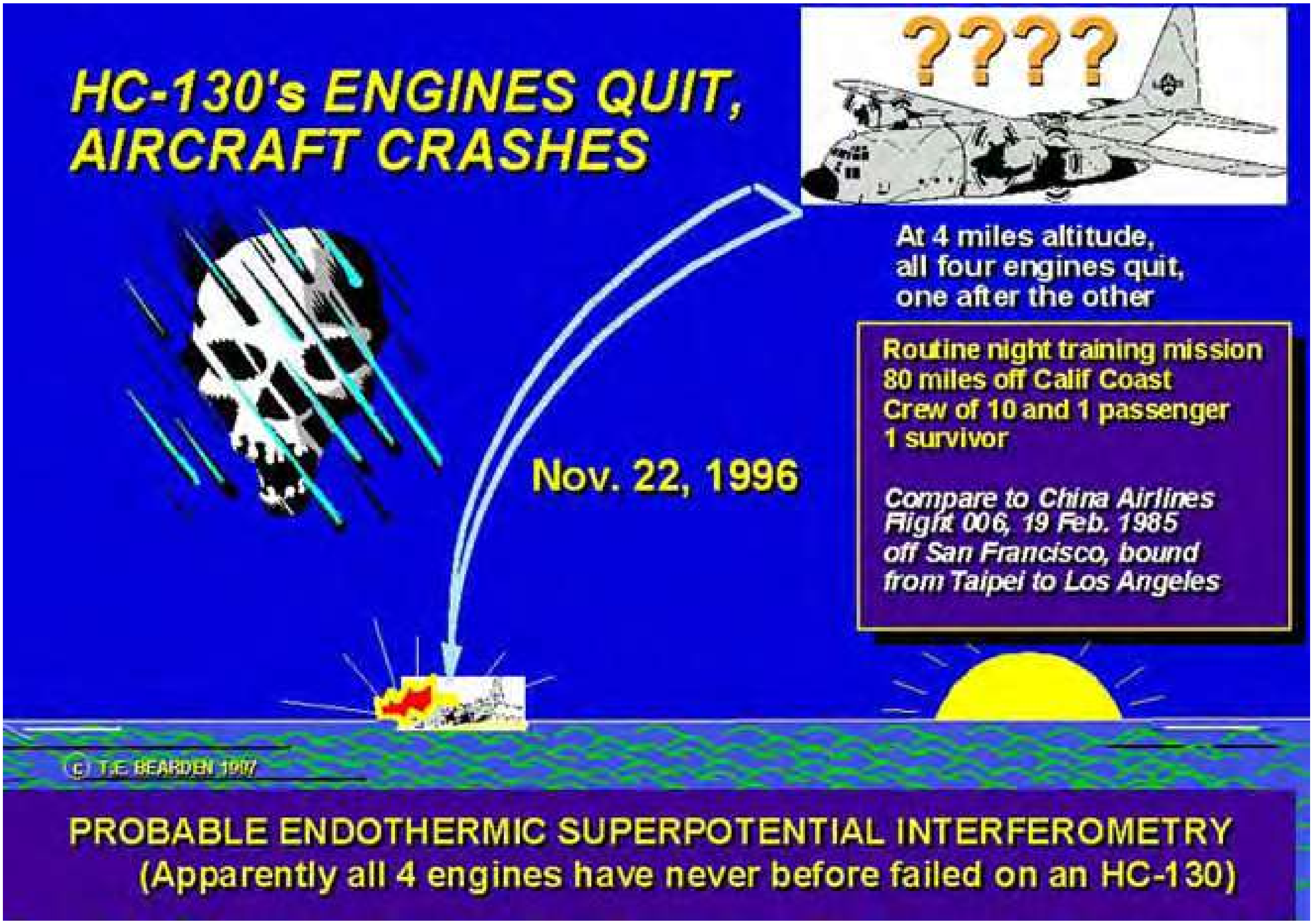

HC-130's ENGINES QUIT, AIRCRAFT CRASHES

At 4 miles altitude,
all four engines quit,
one after the other

Routine night training mission
80 miles off Calif Coast
Crew of 10 and 1 passenger
1 survivor

Compare to China Airlines Flight 006, 19 Feb. 1985 off San Francisco, bound from Taipei to Los Angeles

Nov. 22, 1996

© T.E. BEARDEN 1997

PROBABLE ENDOTHERMIC SUPERPOTENTIAL INTERFEROMETRY
(Apparently all 4 engines have never before failed on an HC-130)

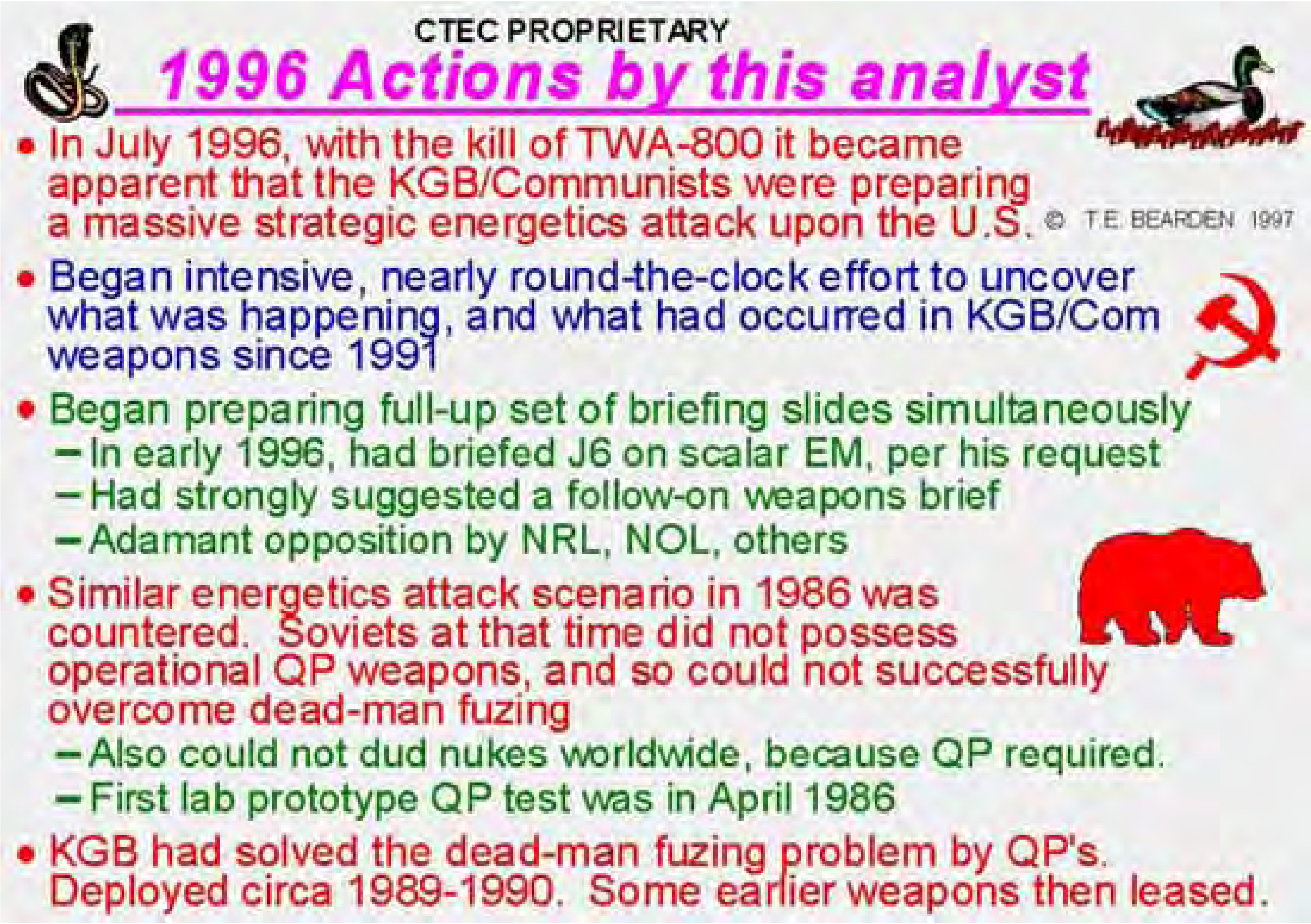

CTEC PROPRIETARY

1996 Actions by this analyst

- In July 1996, with the kill of TWA-800 it became apparent that the KGB/Communists were preparing a massive strategic energetics attack upon the U.S.

© T.E. BEARDEN 1997

- Began intensive, nearly round-the-clock effort to uncover what was happening, and what had occurred in KGB/Com weapons since 1991
- Began preparing full-up set of briefing slides simultaneously
  - In early 1996, had briefed J6 on scalar EM, per his request
  - Had strongly suggested a follow-on weapons brief
  - Adamant opposition by NRL, NOL, others
- Similar energetics attack scenario in 1986 was countered. Soviets at that time did not possess operational QP weapons, and so could not successfully overcome dead-man fuzing
  - Also could not dud nukes worldwide, because QP required.
  - First lab prototype QP test was in April 1986
- KGB had solved the dead-man fuzing problem by QP's. Deployed circa 1989-1990. Some earlier weapons then leased.

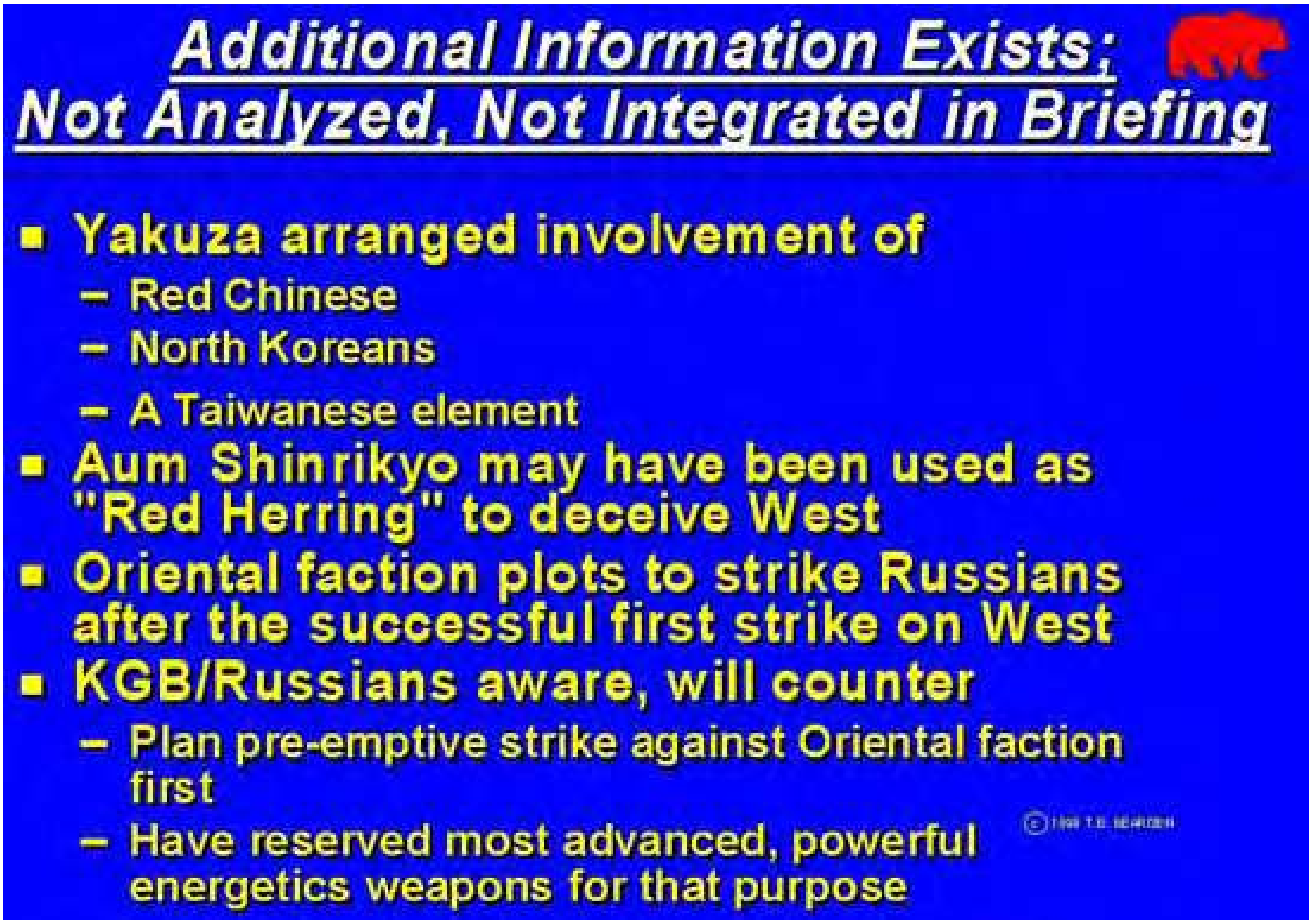

Additional Information Exists; Not Analyzed, Not Integrated in Briefing

- Yakuza arranged involvement of
  - Red Chinese
  - North Koreans
  - A Taiwanese element
- Aum Shinrikyo may have been used as "Red Herring" to deceive West
- Oriental faction plots to strike Russians after the successful first strike on West
- KGB/Russians aware, will counter
  - Plan pre-emptive strike against Oriental faction first
  - Have reserved most advanced, powerful energetics weapons for that purpose

The Tom Bearden Website

Help support the research

Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6

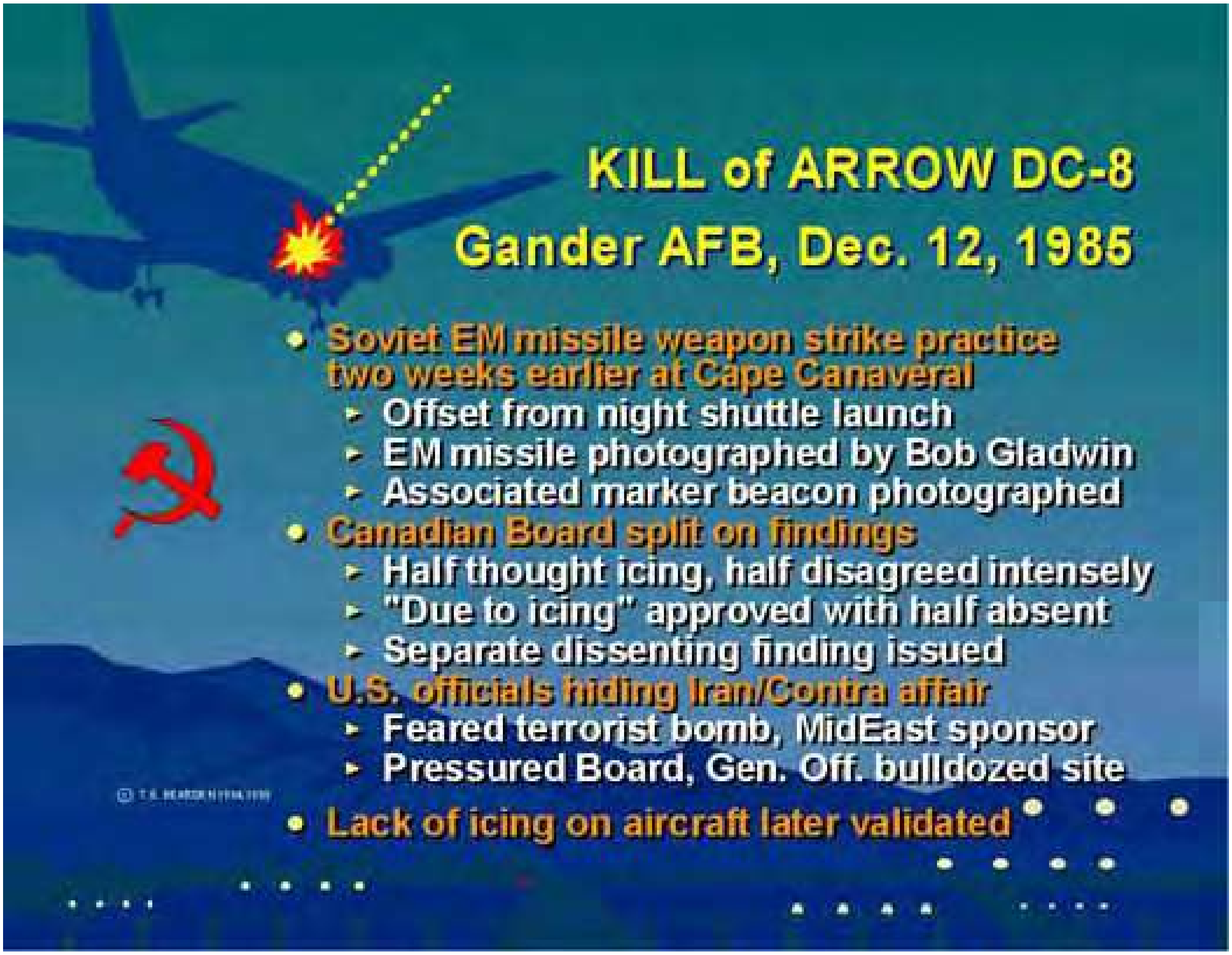

KILL of ARROW DC-8

Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. bulldozed site
- Lack of icing on aircraft later validated

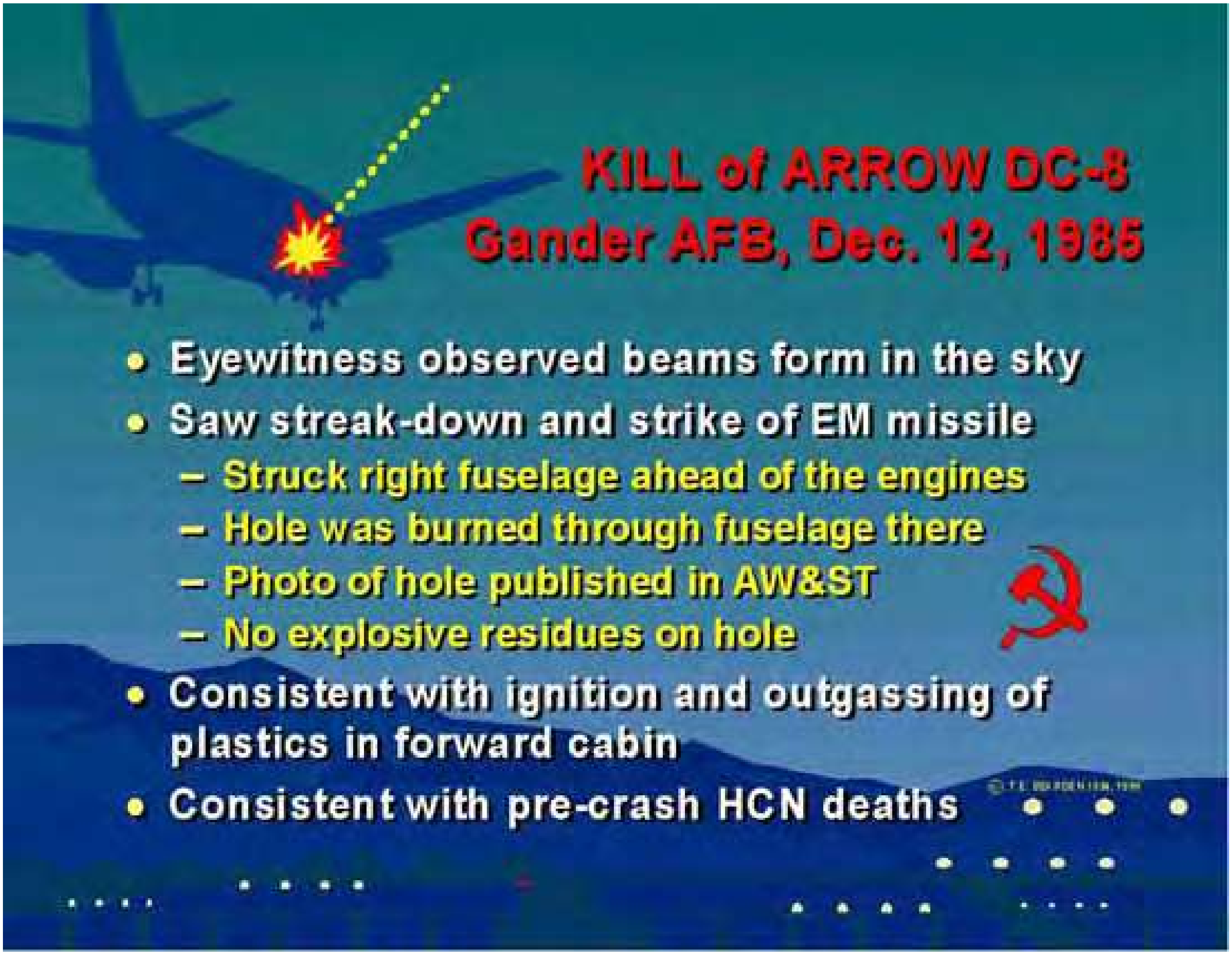

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Eyewitness observed beams form in the sky
- Saw streak-down and strike of EM missile
  - Struck right fuselage ahead of the engines
  - Hole was burned through fuselage there
  - Photo of hole published in AW&ST
  - No explosive residues on hole
- Consistent with ignition and outgassing of plastics in forward cabin
- Consistent with pre-crash HCN deaths

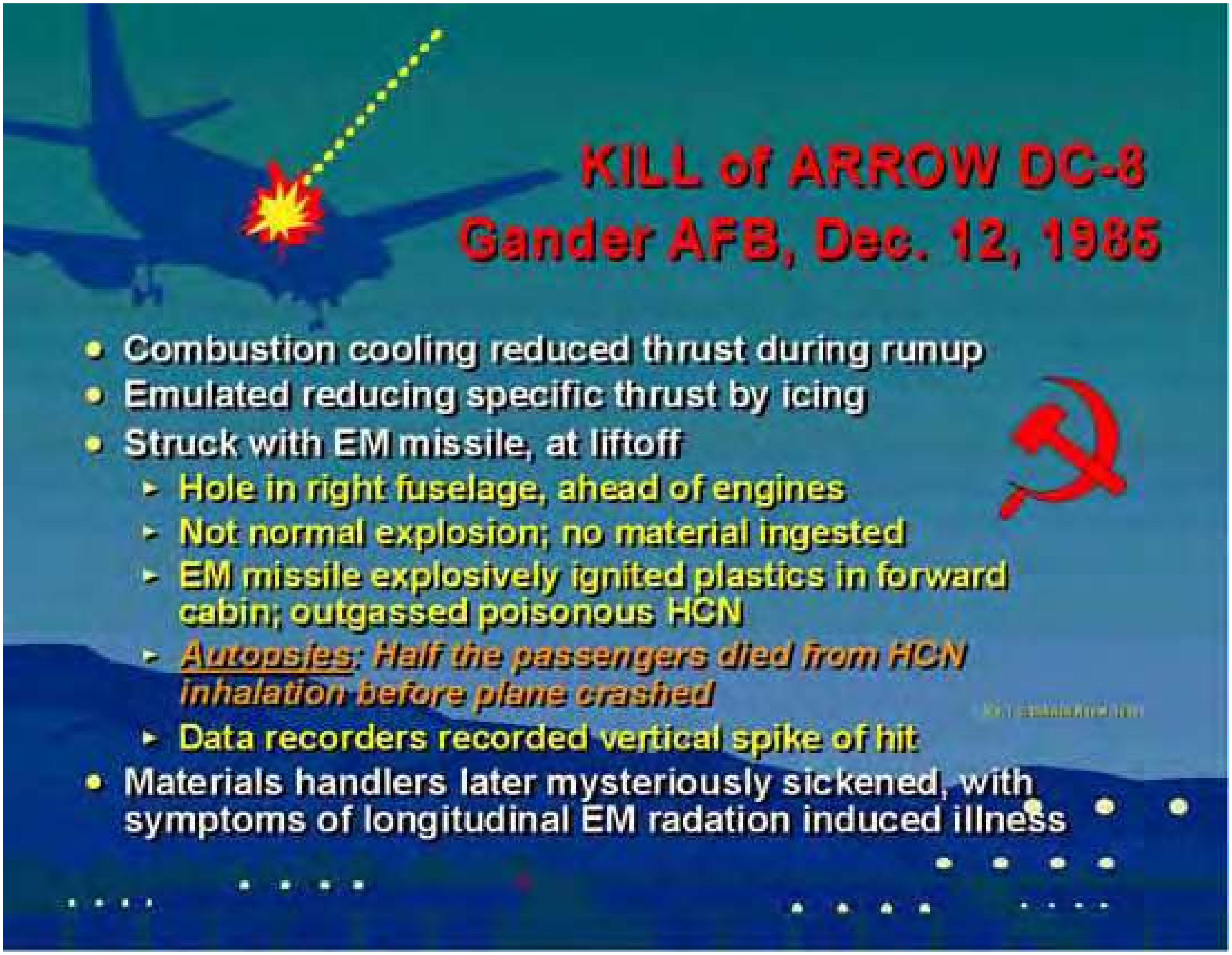

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation before plane crashed*
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of longitudinal EM radation induced illness

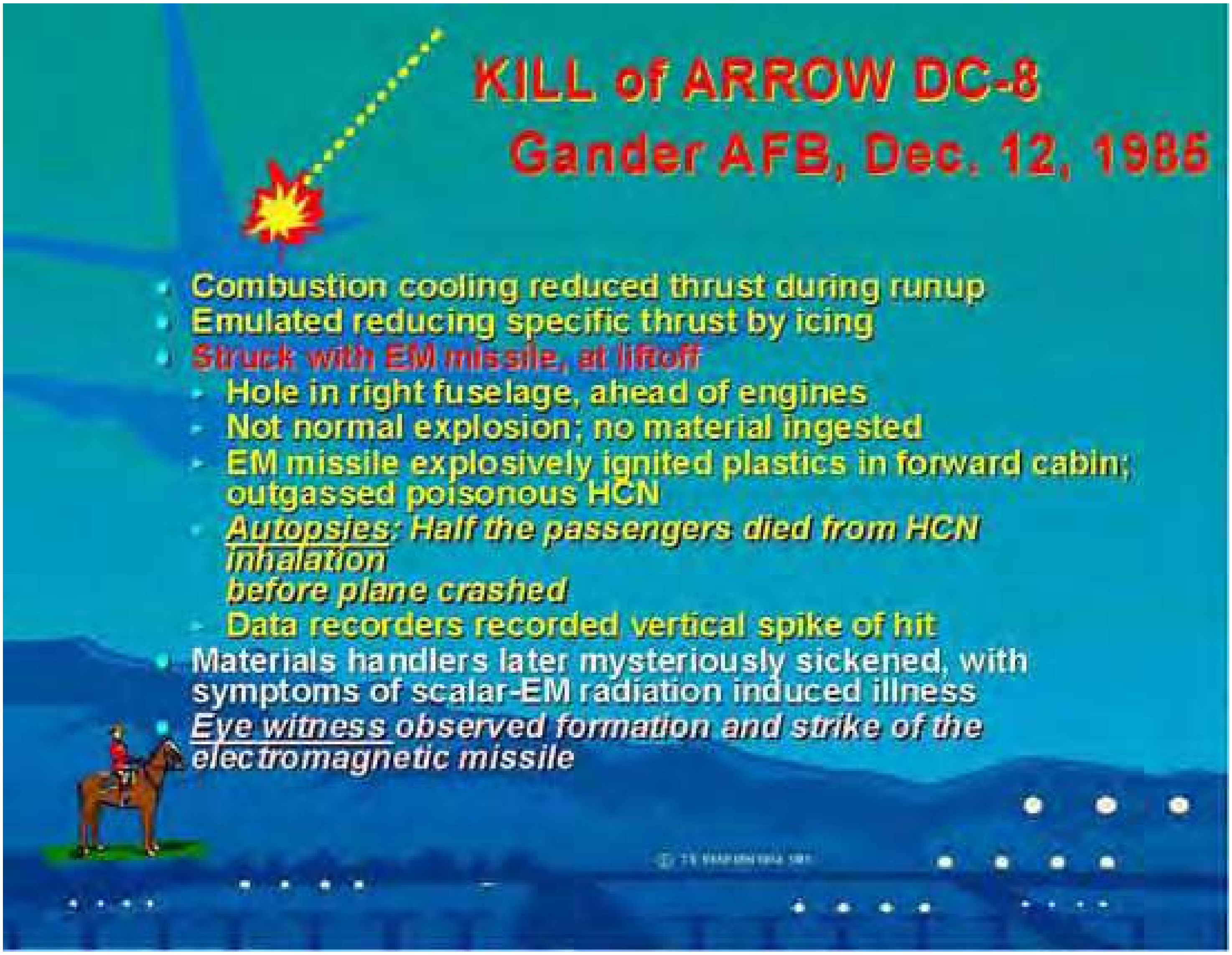

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation before plane crashed*
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of scalar-EM radiation induced illness
- *Eye witness observed formation and strike of the electromagnetic missile*

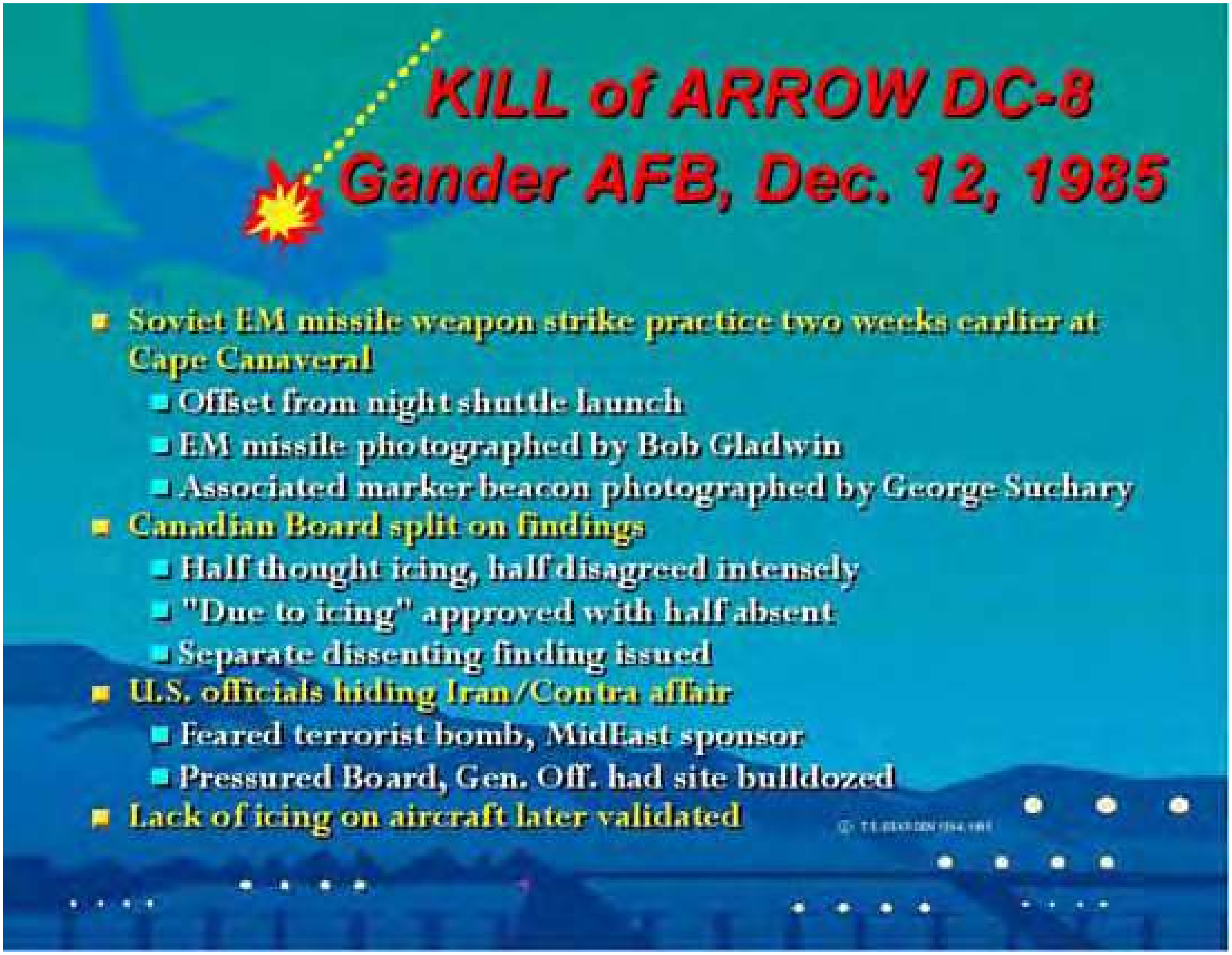

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed by George Suchary
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. had site bulldozed
- Lack of icing on aircraft later validated

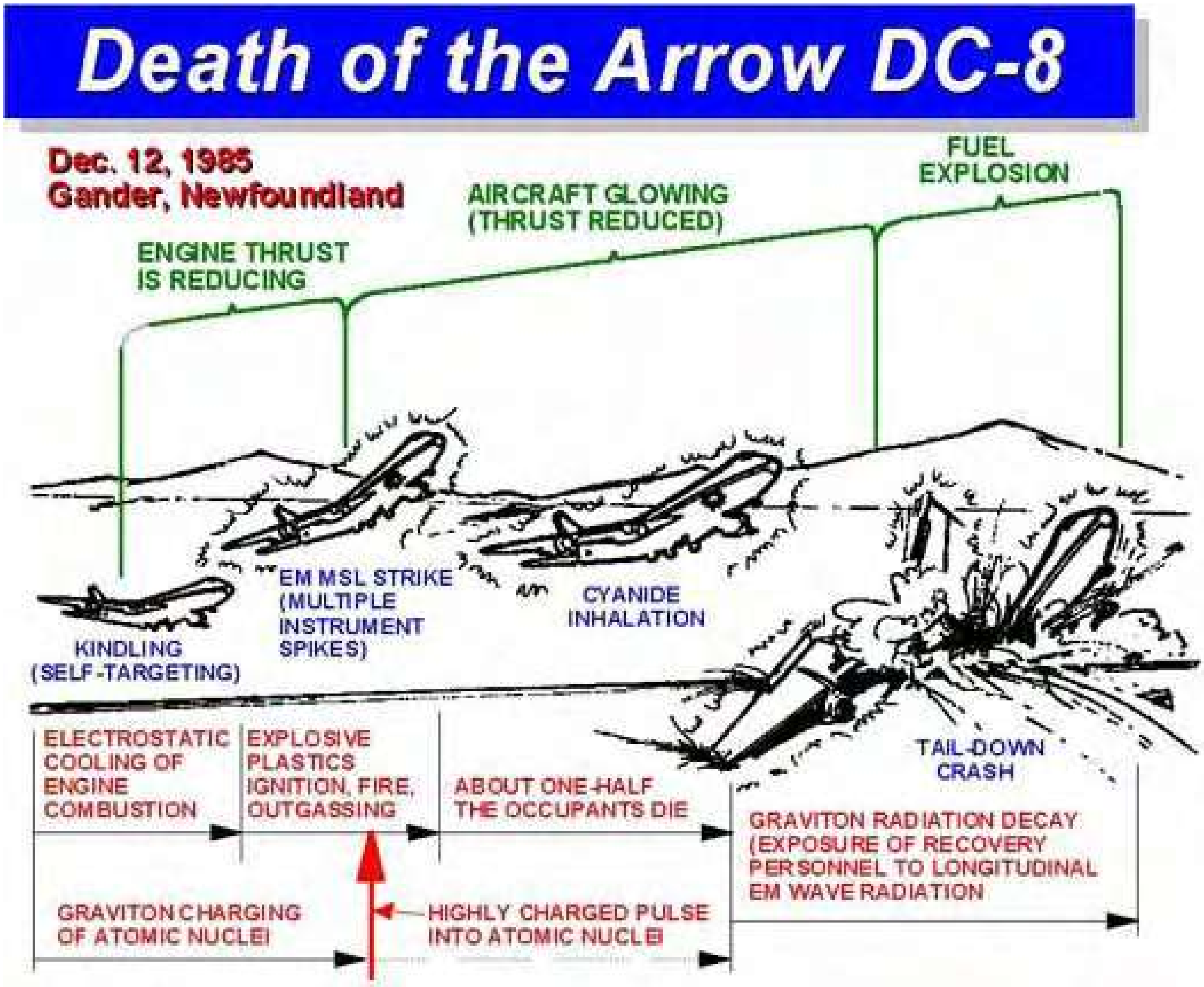
Death of the Arrow DC-8
Dec. 12, 1985
Gander, Newfoundland
ENGINE THRUST
IS REDUCING
AIRCRAFT GLOWING
(THRUST REDUCED)
FUEL
EXPLOSION
KINDLING
(SELF-TARGETING)
EM MSL STRIKE
(MULTIPLE
INSTRUMENT
SPIKES)
CYANIDE
INHALATION
TAIL-DOWN
CRASH
ELECTROSTATIC
COOLING OF
ENGINE
COMBUSTION
EXPLOSIVE
PLASTICS
IGNITION, FIRE,
OUTGASSING
ABOUT ONE-HALF
THE OCCUPANTS DIE
GRAVITON RADIATION DECAY
(EXPOSURE OF RECOVERY
PERSONNEL TO LONGITUDINAL
EM WAVE RADIATION
GRAVITON CHARGING
OF ATOMIC NUCLEI
HIGHLY CHARGED PULSE
INTO ATOMIC NUCLEI

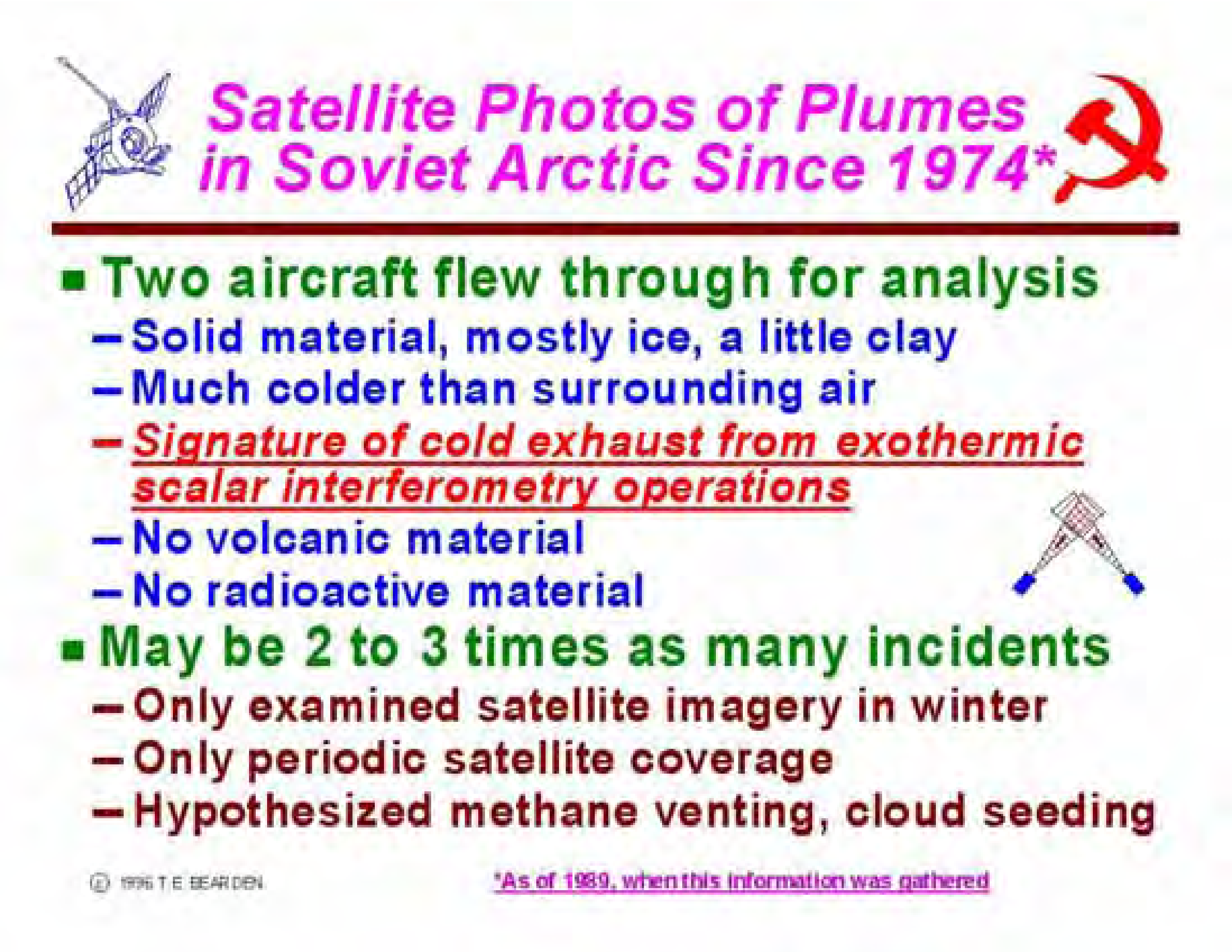

Satellite Photos of Plumes in Soviet Arctic Since 1974*

- Two aircraft flew through for analysis
  - Solid material, mostly ice, a little clay
  - Much colder than surrounding air
  - *Signature of cold exhaust from exothermic scalar interferometry operations*
  - No volcanic material
  - No radioactive material
- May be 2 to 3 times as many incidents
  - Only examined satellite imagery in winter
  - Only periodic satellite coverage
  - Hypothesized methane venting, cloud seeding

© 1996 T.E. BEARDEN

*As of 1989, when this information was gathered

AUM SHINRIKYO

("SUPREME TRUTH" CULT)

Shoko Asahara, Leader

© 1995, 1998 T.E. BEARDEN

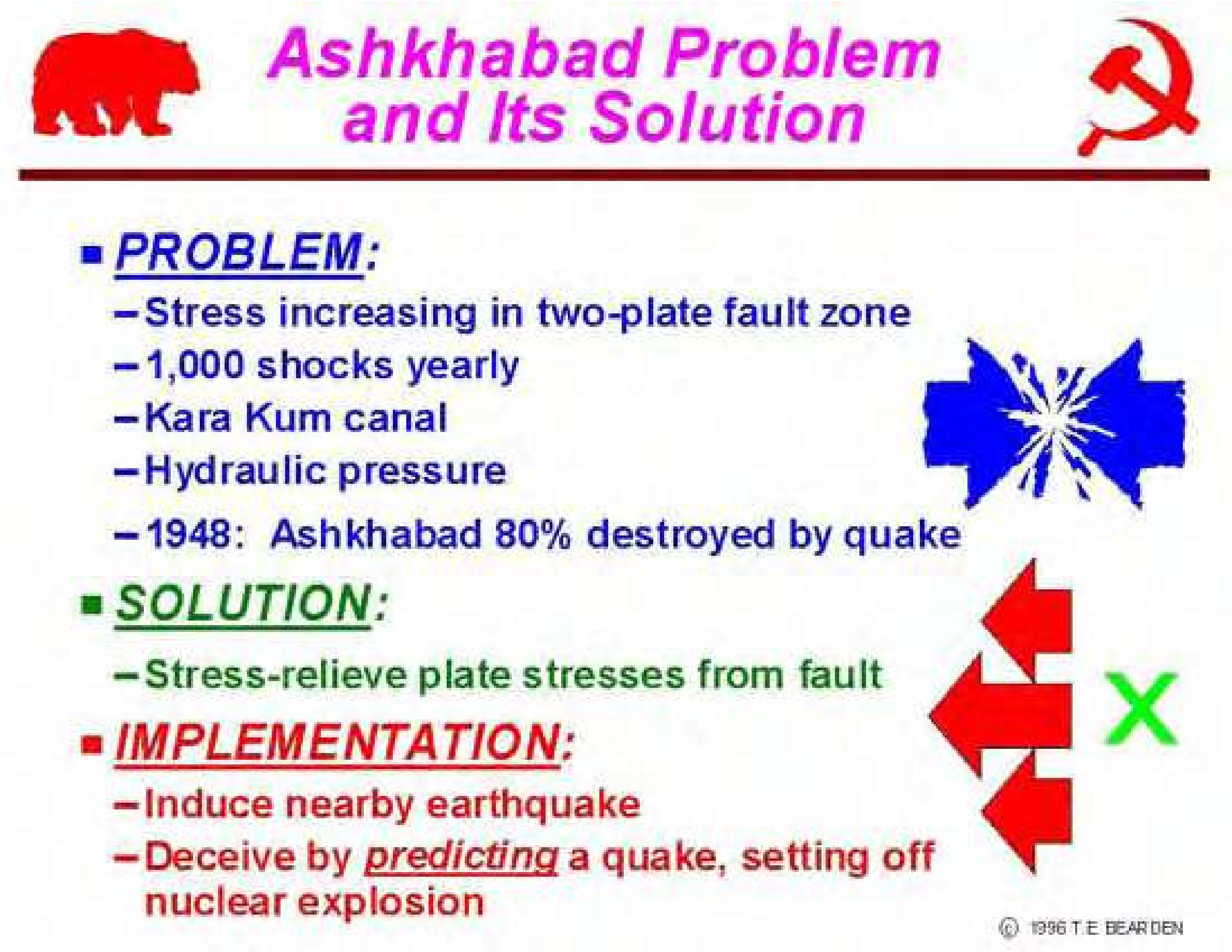

Ashkhabad Problem and Its Solution

- ***PROBLEM:***
  - Stress increasing in two-plate fault zone
  - 1,000 shocks yearly
  - Kara Kum canal
  - Hydraulic pressure
  - 1948: Ashkhabad 80% destroyed by quake
- ***SOLUTION:***
  - Stress-relieve plate stresses from fault
- ***IMPLEMENTATION:***
  - Induce nearby earthquake
  - Deceive by *predicting* a quake, setting off nuclear explosion

© 1996 T.E. BEARDEN

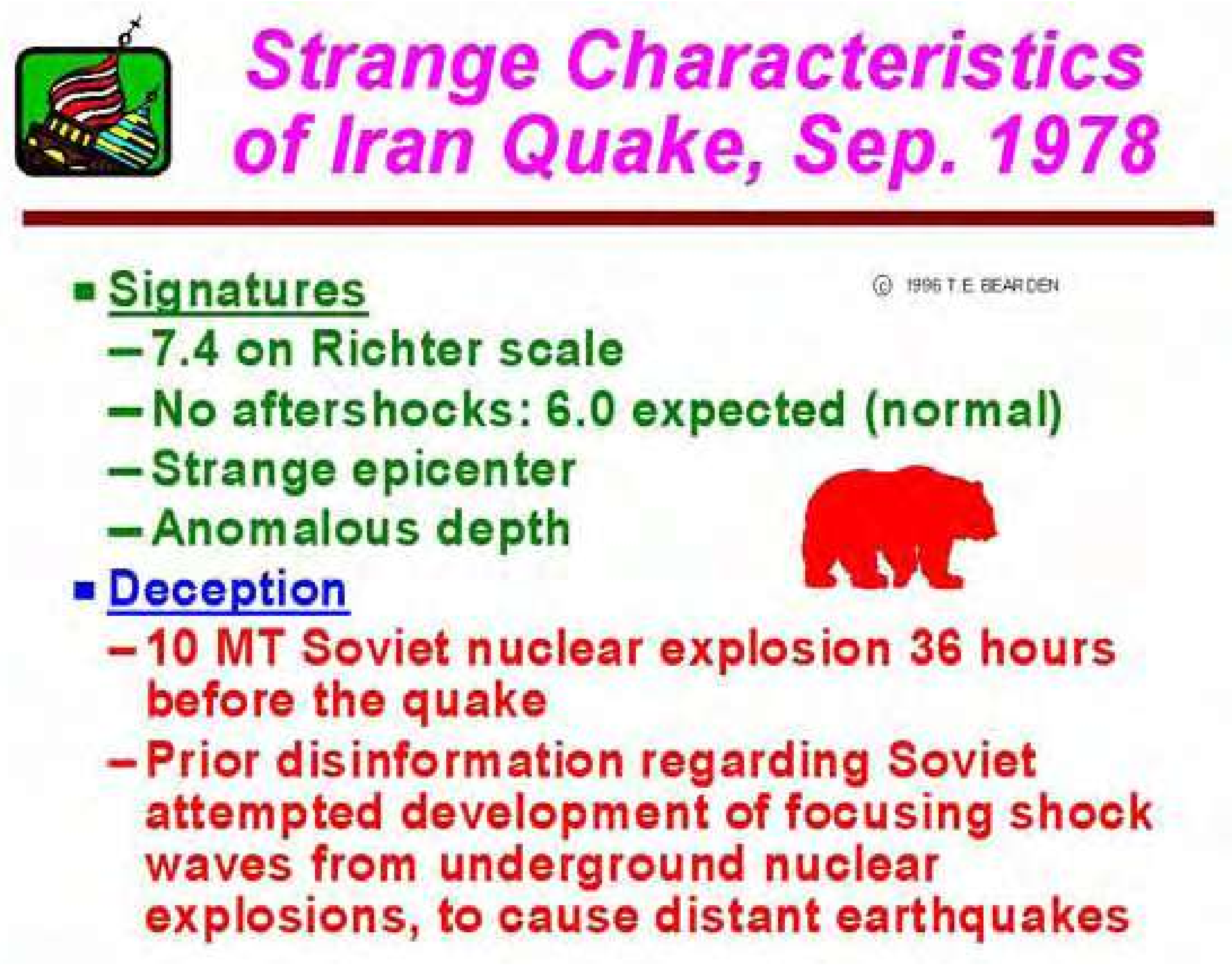

Strange Characteristics of Iran Quake, Sep. 1978

© 1996 T.E. BEARDEN

- Signatures
  - 7.4 on Richter scale
  - No aftershocks: 6.0 expected (normal)
  - Strange epicenter
  - Anomalous depth
- Deception
  - 10 MT Soviet nuclear explosion 36 hours before the quake
  - Prior disinformation regarding Soviet attempted development of focusing shock waves from underground nuclear explosions, to cause distant earthquakes

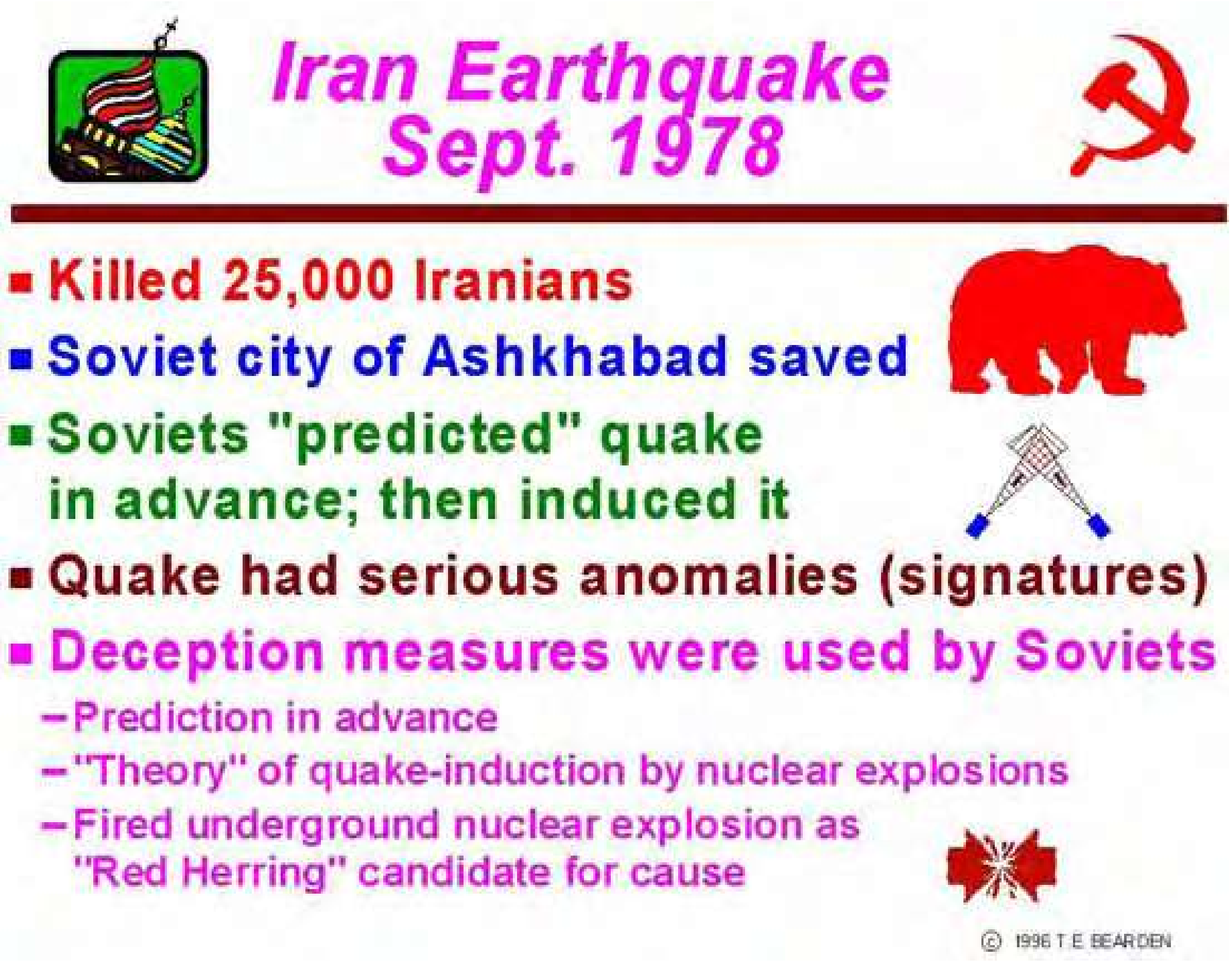

Iran Earthquake Sept. 1978

- Killed 25,000 Iranians
- Soviet city of Ashkhabad saved
- Soviets "predicted" quake in advance; then induced it
- Quake had serious anomalies (signatures)
- Deception measures were used by Soviets
  - Prediction in advance
  - "Theory" of quake-induction by nuclear explosions
  - Fired underground nuclear explosion as "Red Herring" candidate for cause

© 1996 T E BEARDEN

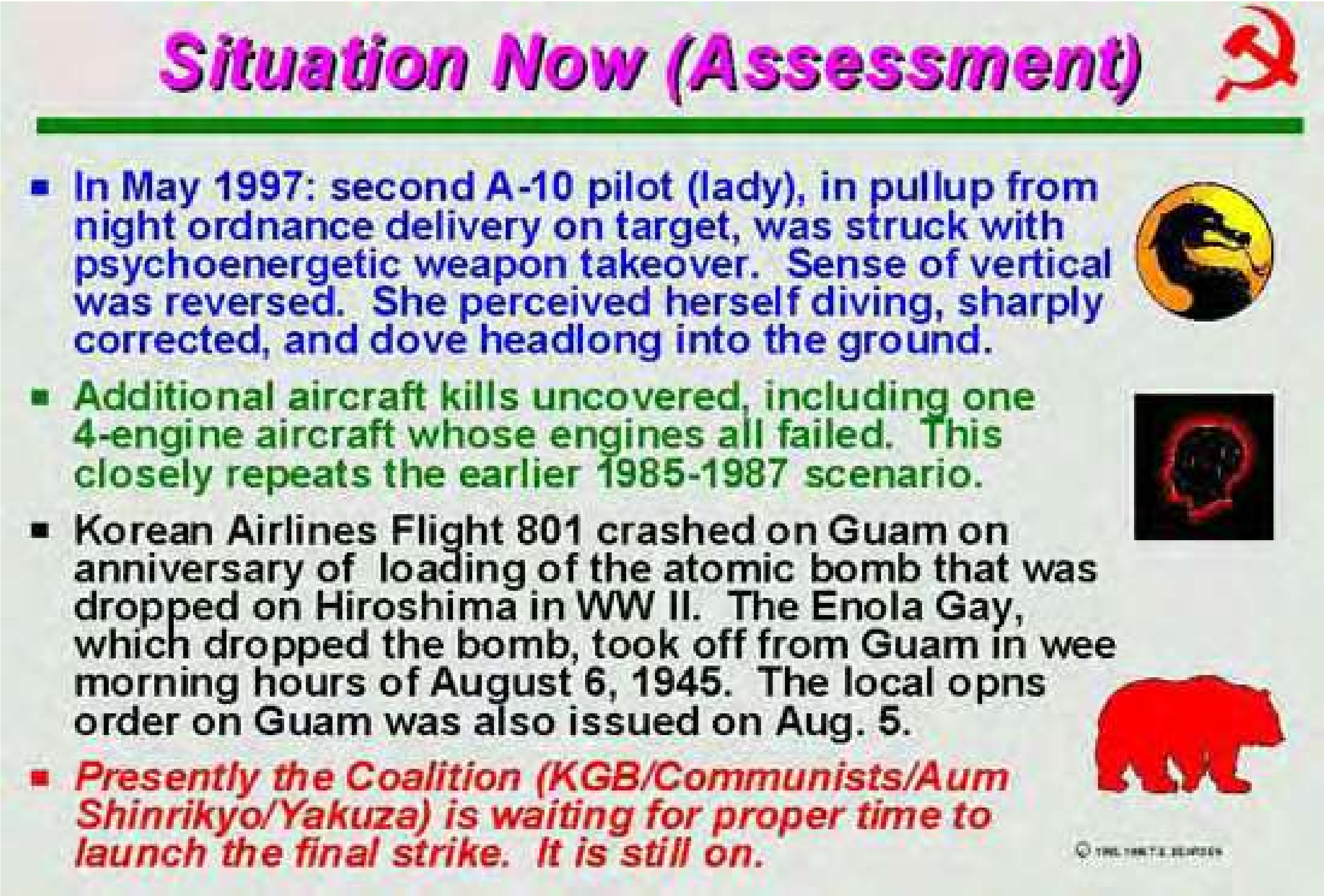

Situation Now (Assessment)

- In May 1997: second A-10 pilot (lady), in pullup from night ordnance delivery on target, was struck with psychoenergetic weapon takeover. Sense of vertical was reversed. She perceived herself diving, sharply corrected, and dove headlong into the ground.
- Additional aircraft kills uncovered, including one 4-engine aircraft whose engines all failed. This closely repeats the earlier 1985-1987 scenario.
- Korean Airlines Flight 801 crashed on Guam on anniversary of loading of the atomic bomb that was dropped on Hiroshima in WW II. The Enola Gay, which dropped the bomb, took off from Guam in wee morning hours of August 6, 1945. The local opns order on Guam was also issued on Aug. 5.
- *Presently the Coalition (KGB/Communists/Aum Shinrikyo/Yakuza) is waiting for proper time to launch the final strike. It is still on.*

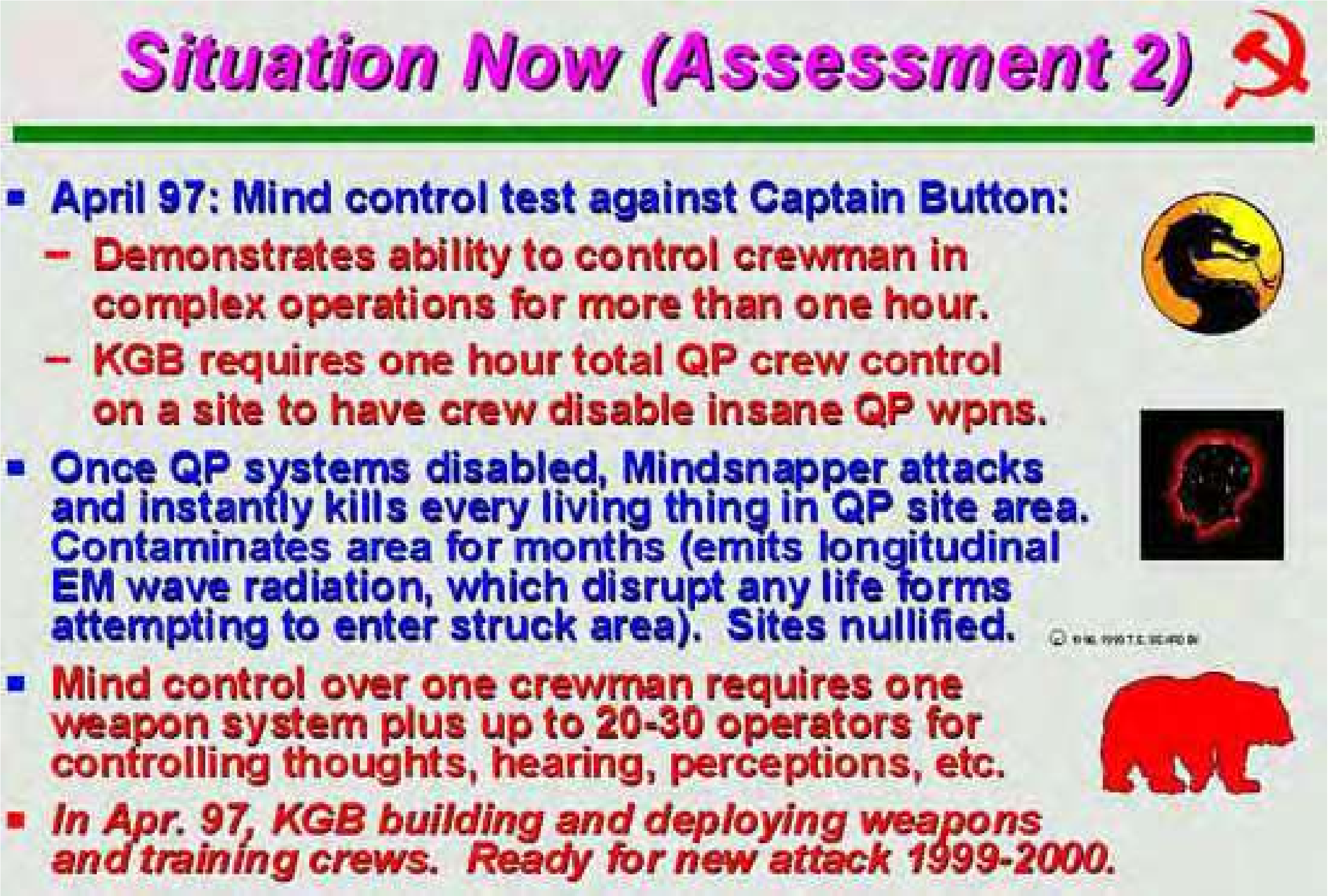

Situation Now (Assessment 2)

- **April 97: Mind control test against Captain Button:**
  - **Demonstrates ability to control crewman in complex operations for more than one hour.**
  - **KGB requires one hour total QP crew control on a site to have crew disable insane QP wpns.**
- **Once QP systems disabled, Mindsnapper attacks and instantly kills every living thing in QP site area. Contaminates area for months (emits longitudinal EM wave radiation, which disrupt any life forms attempting to enter struck area). Sites nullified.**
- **Mind control over one crewman requires one weapon system plus up to 20-30 operators for controlling thoughts, hearing, perceptions, etc.**
- ***In Apr. 97, KGB building and deploying weapons and training crews. Ready for new attack 1999-2000.***

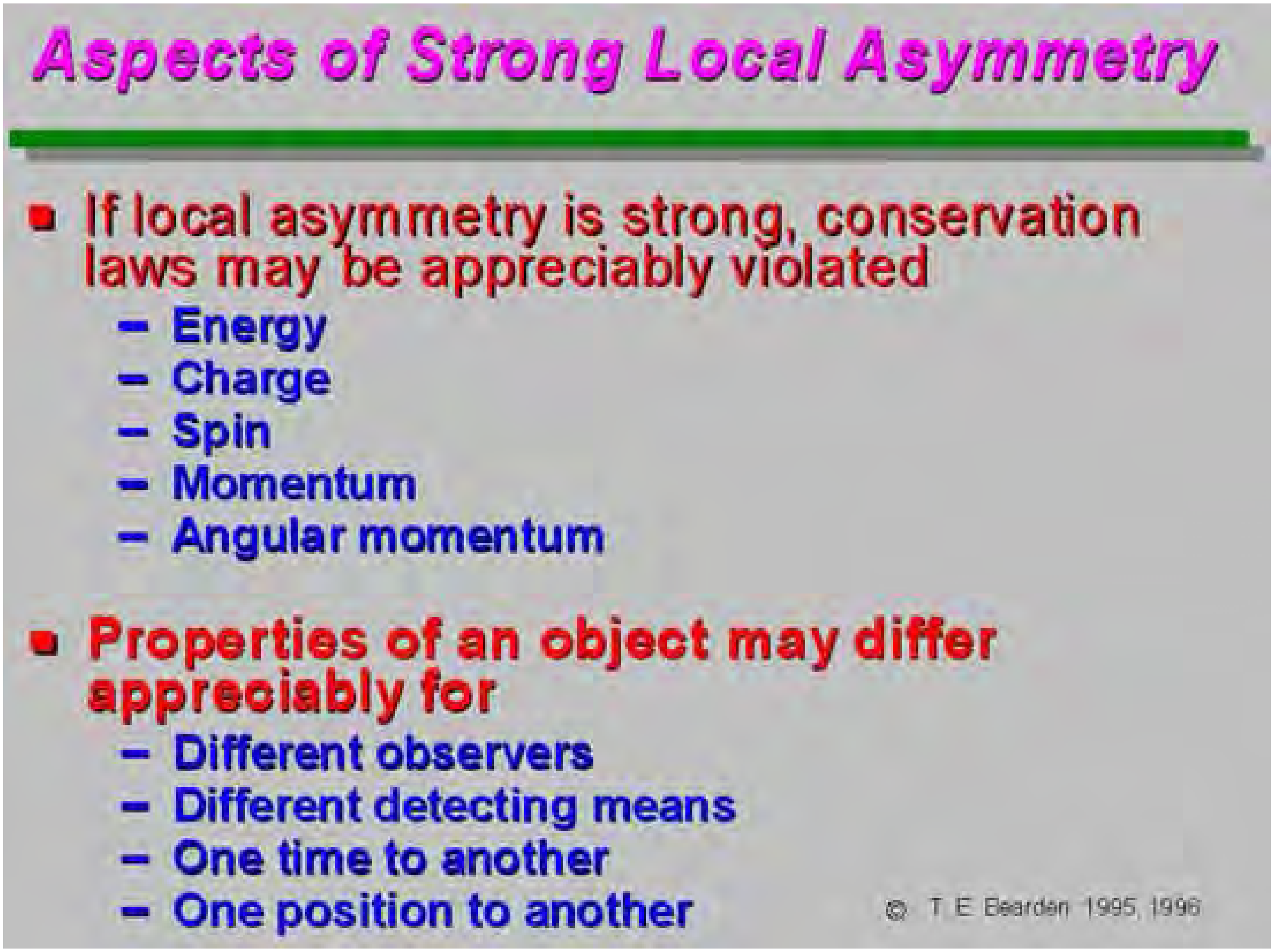

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- **Properties of an object may differ appreciably for**
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

© T. E. Bearden 1995, 1996

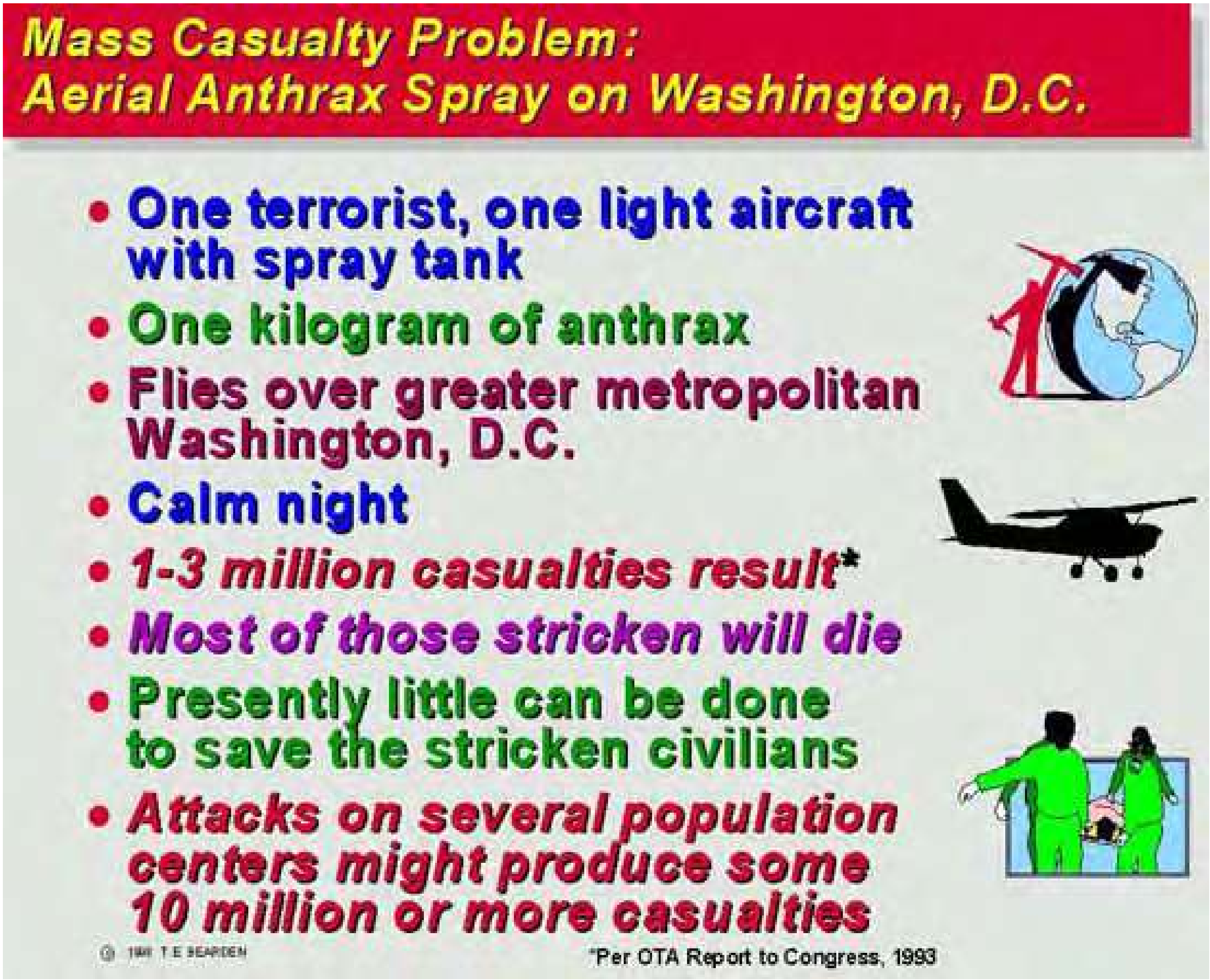

Mass Casualty Problem:
Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- One kilogram of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

© [illegible] T. E. BEARDEN

*Per OTA Report to Congress, 1993

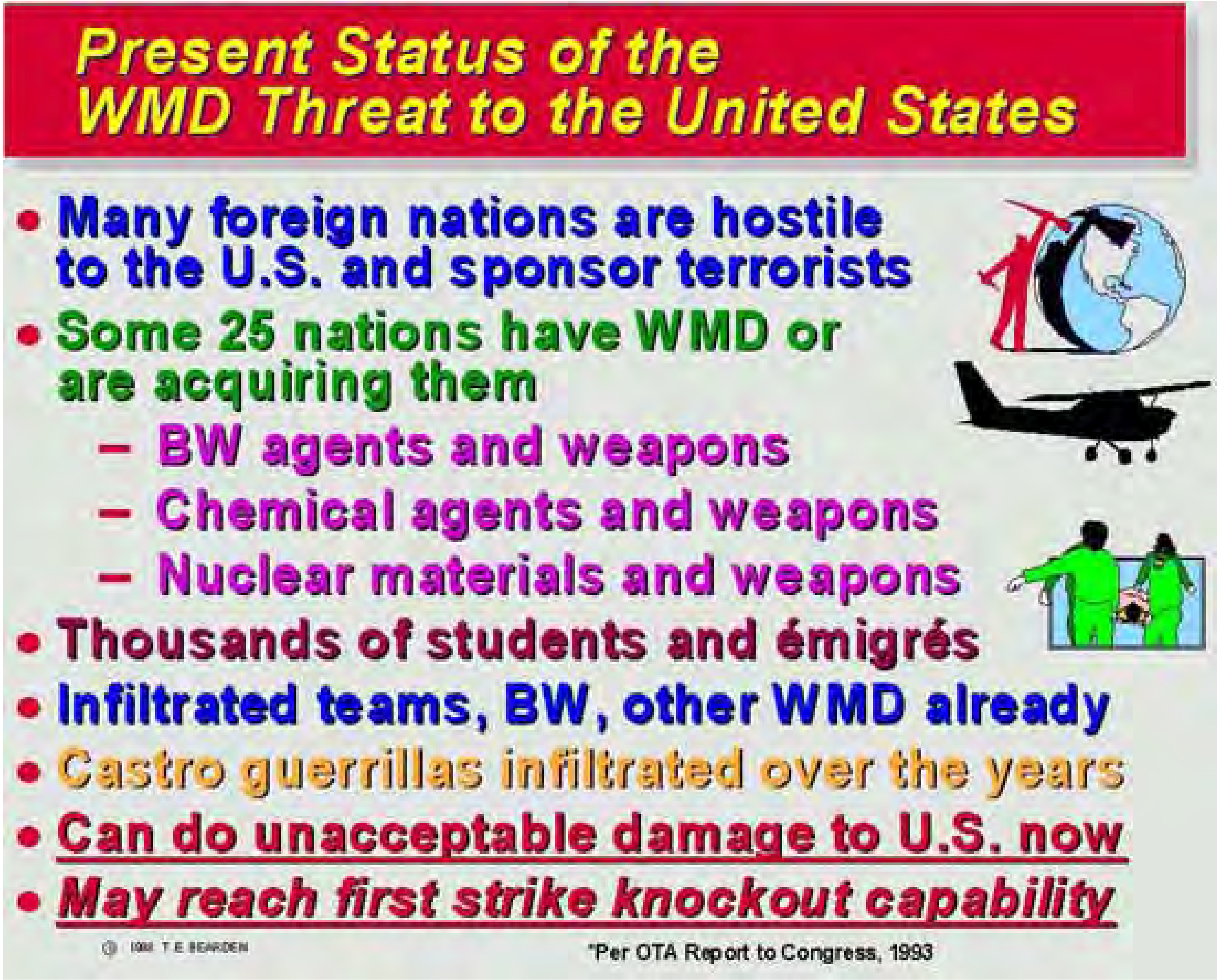

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
  - BW agents and weapons
  - Chemical agents and weapons
  - Nuclear materials and weapons
- Thousands of students and émigrés
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© [illegible] T.E. BEARDEN

*Per OTA Report to Congress, 1993

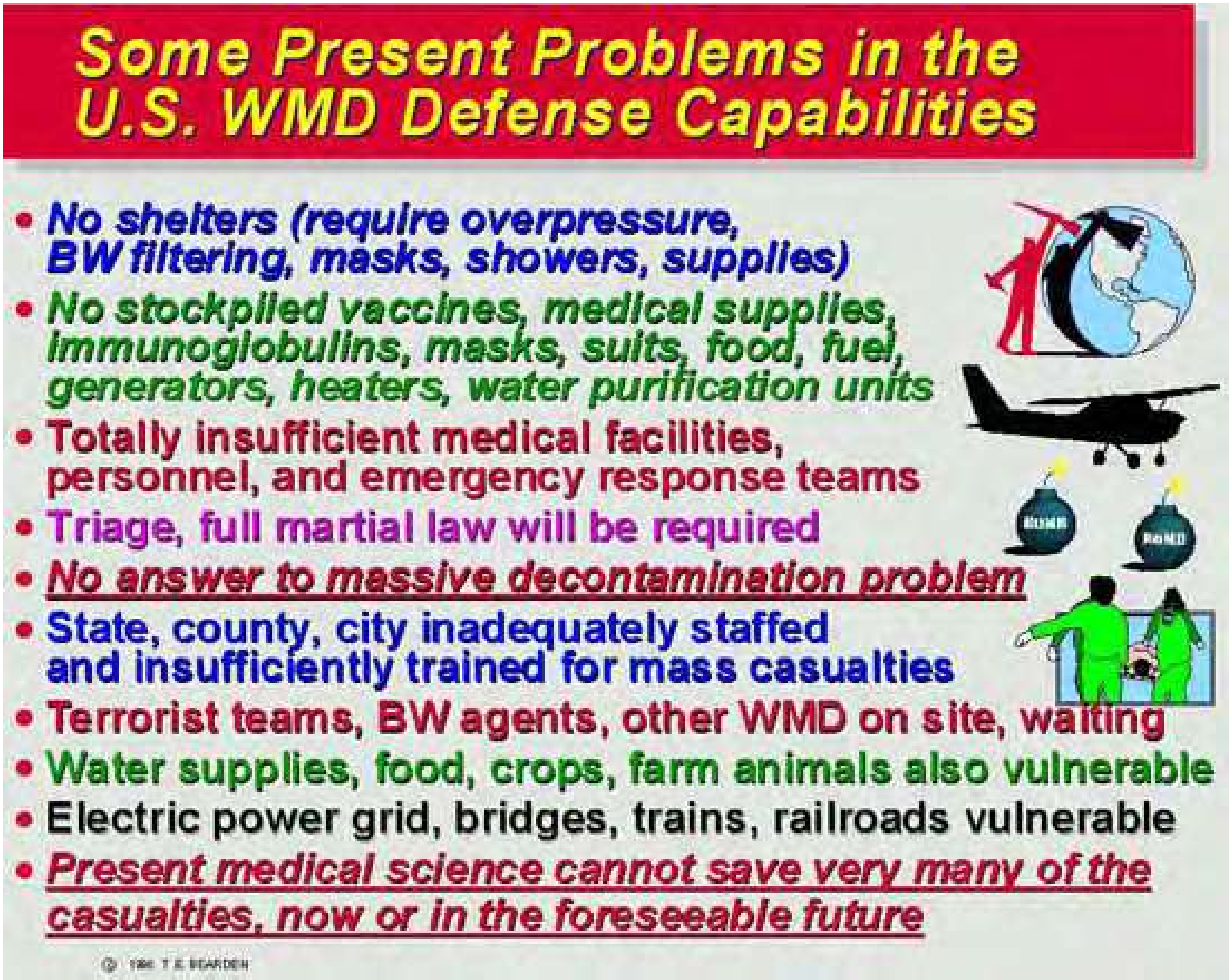

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© 1996 T. E. BEARDEN

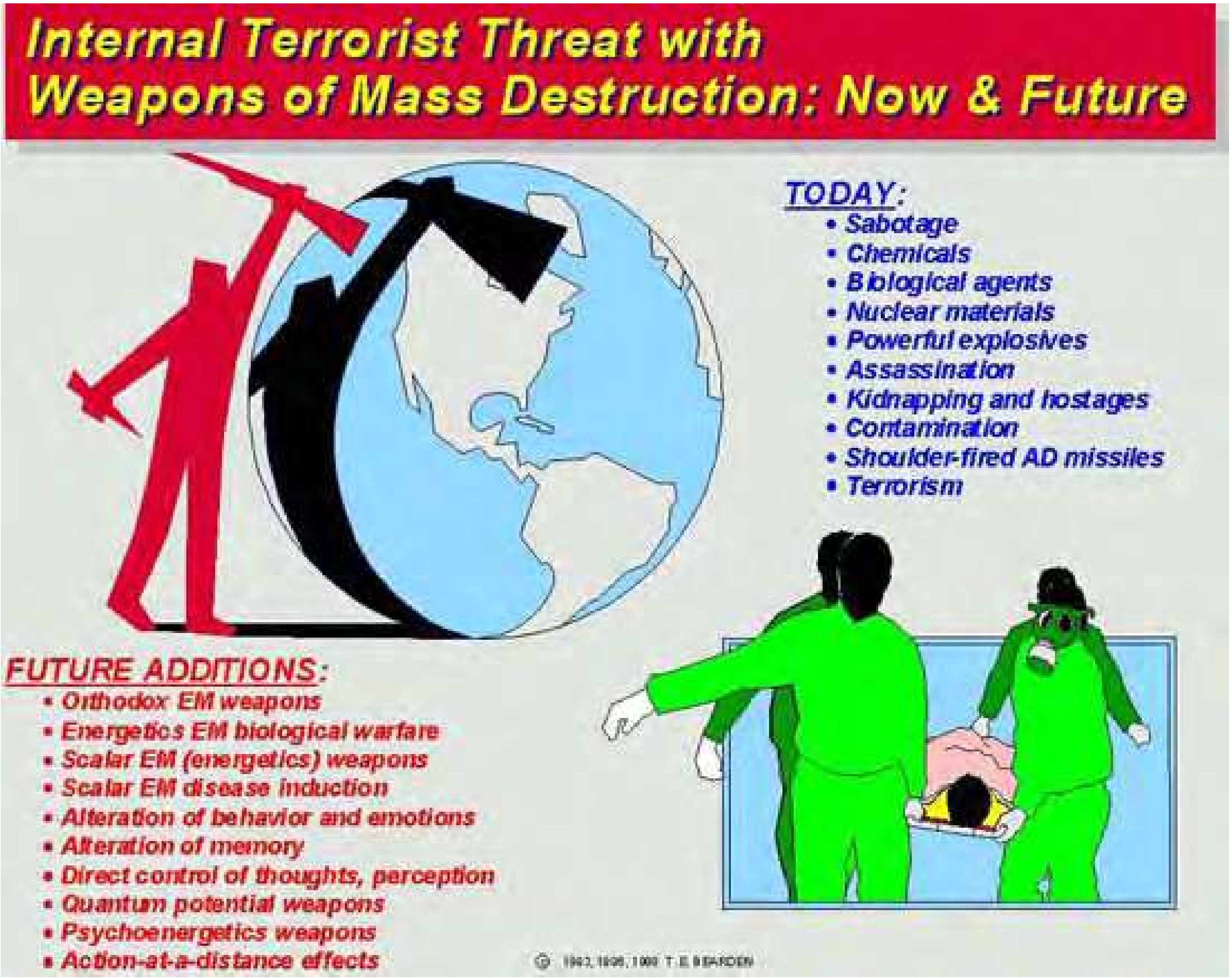

Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- Orthodox EM weapons
- Energetics EM biological warfare
- Scalar EM (energetics) weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct control of thoughts, perception
- Quantum potential weapons
- Psychoenergetics weapons
- Action-at-a-distance effects

© 1993, 1996, 1999 T. E. BEARDEN

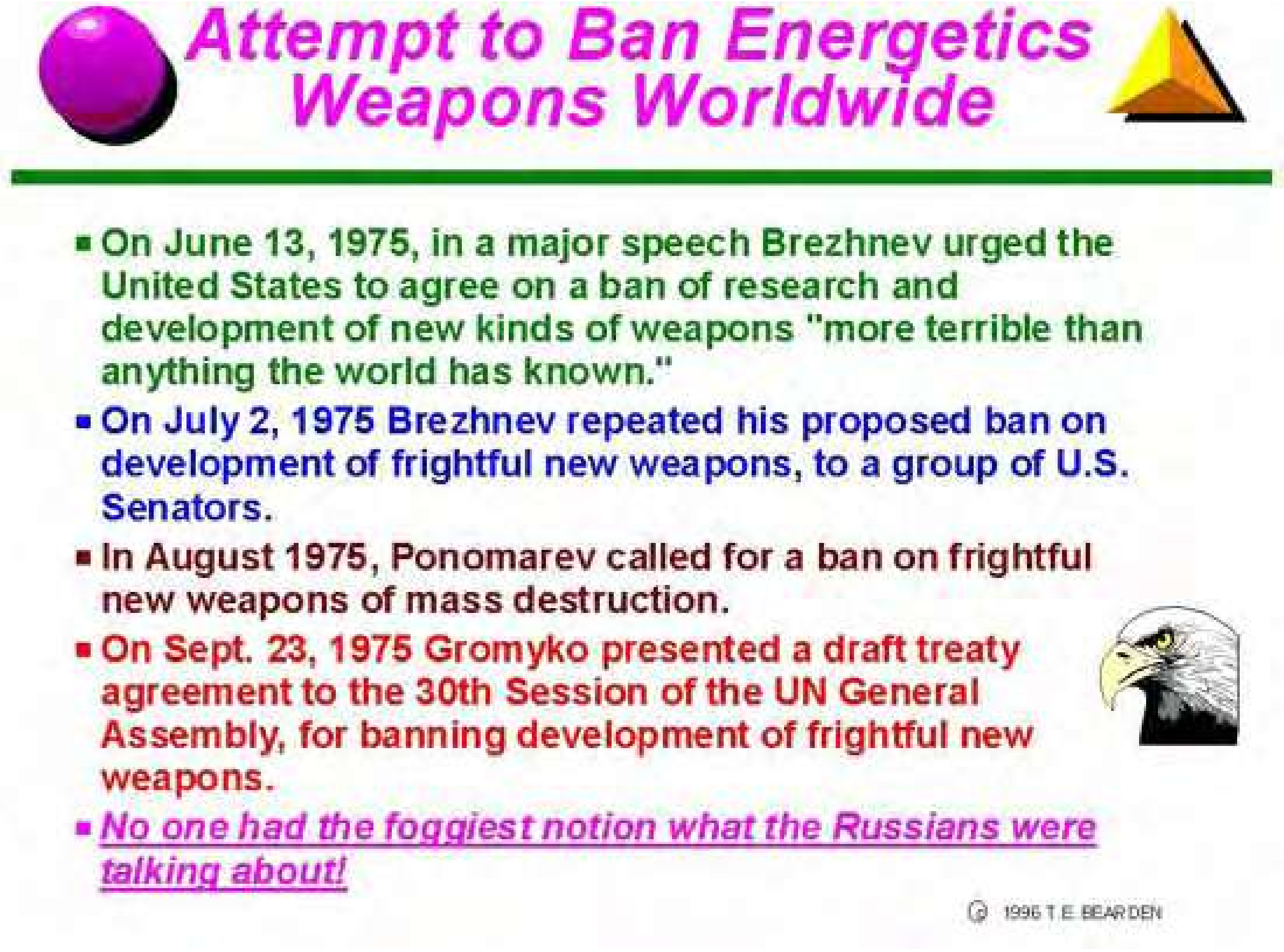

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T.E. BEARDEN

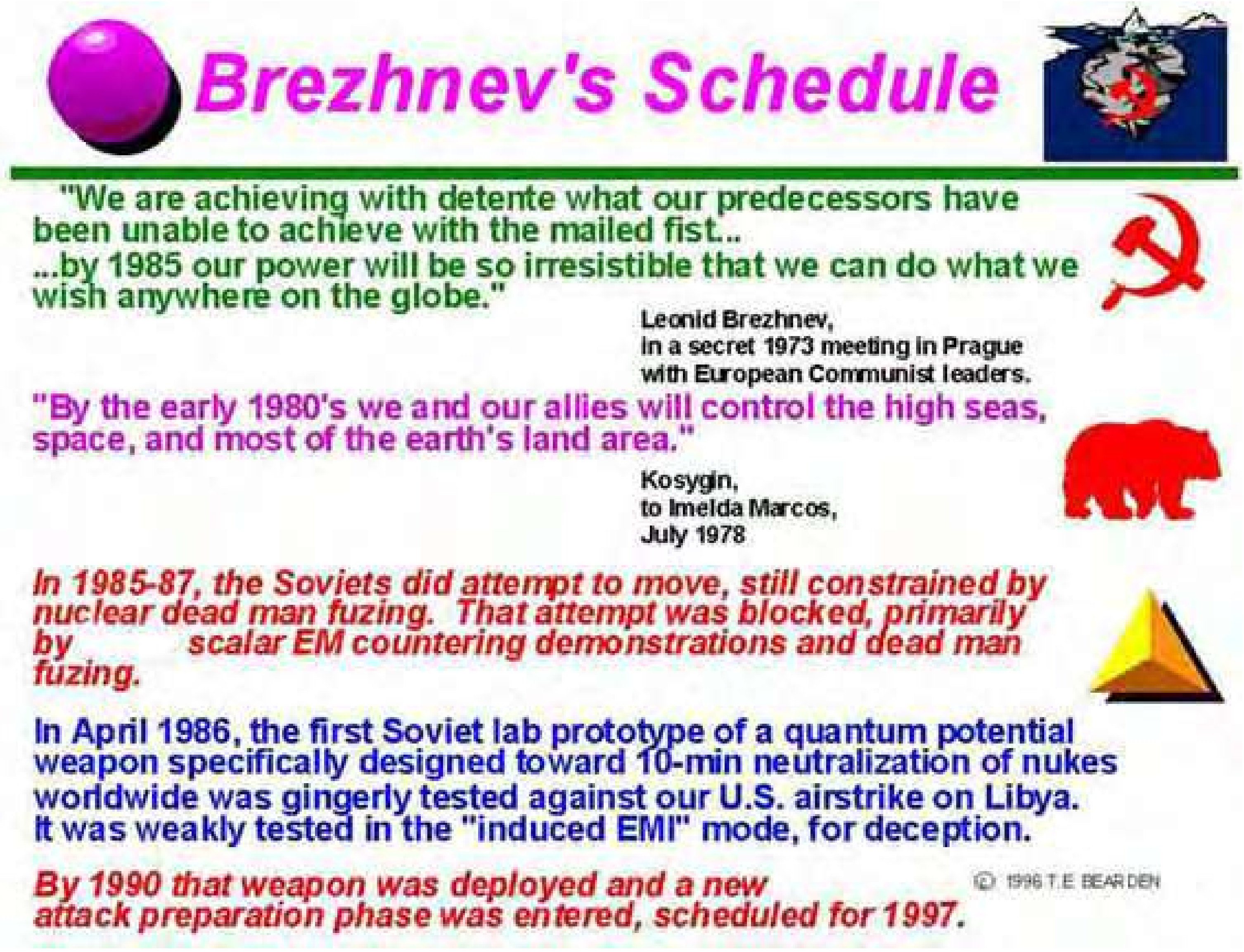

Brezhnev's Schedule

"We are achieving with detente what our predecessors have been unable to achieve with the mailed fist..
...by 1985 our power will be so irresistible that we can do what we wish anywhere on the globe."

Leonid Brezhnev,
in a secret 1973 meeting in Prague
with European Communist leaders.

"By the early 1980's we and our allies will control the high seas, space, and most of the earth's land area."

Kosygin,
to Imelda Marcos,
July 1978

In 1985-87, the Soviets did attempt to move, still constrained by nuclear dead man fuzing. That attempt was blocked, primarily by scalar EM countering demonstrations and dead man fuzing.

In April 1986, the first Soviet lab prototype of a quantum potential weapon specifically designed toward 10-min neutralization of nukes worldwide was gingerly tested against our U.S. airstrike on Libya. It was weakly tested in the "induced EMI" mode, for deception.

By 1990 that weapon was deployed and a new attack preparation phase was entered, scheduled for 1997.

© 1996 T.E. BEARDEN

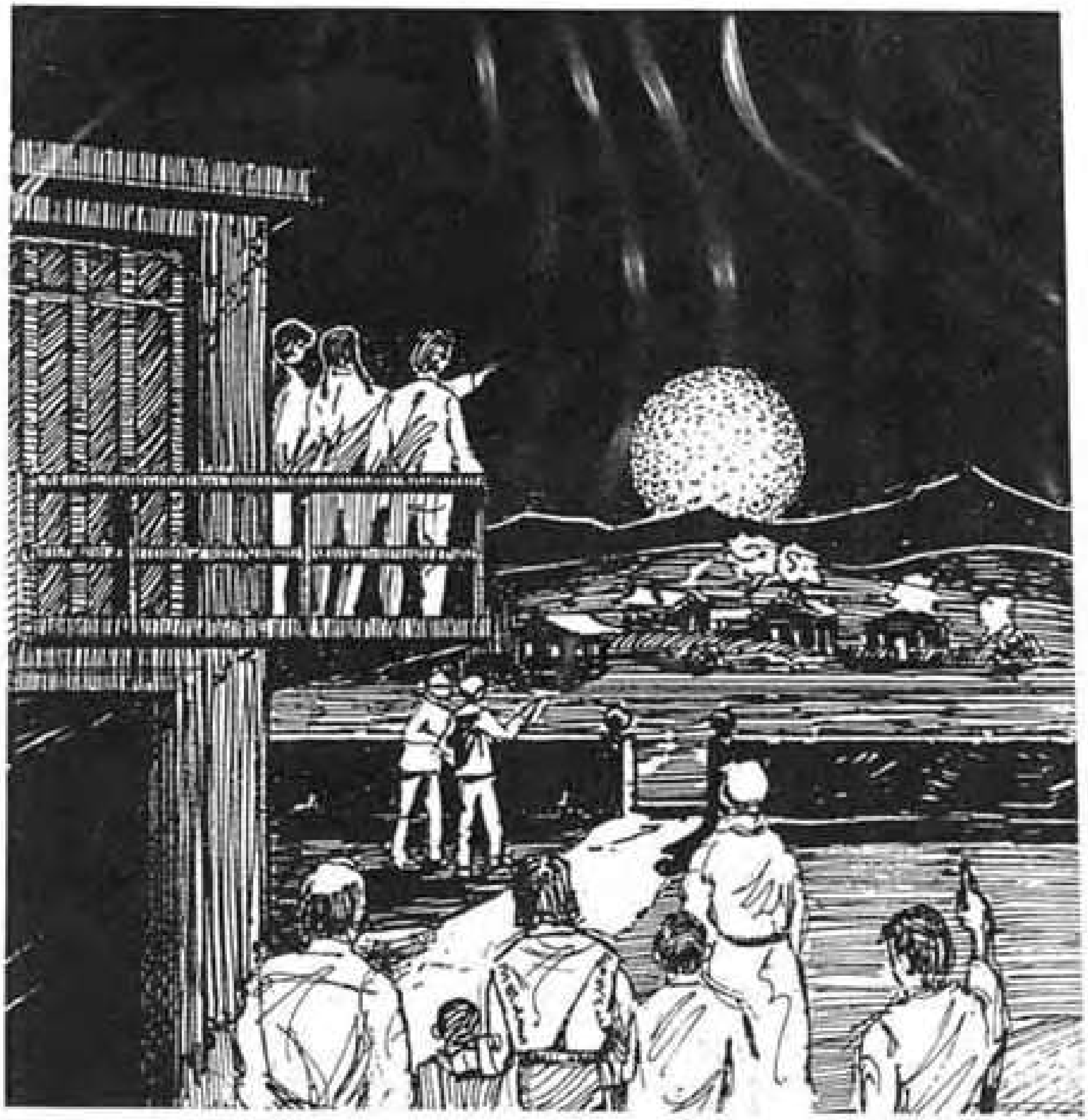

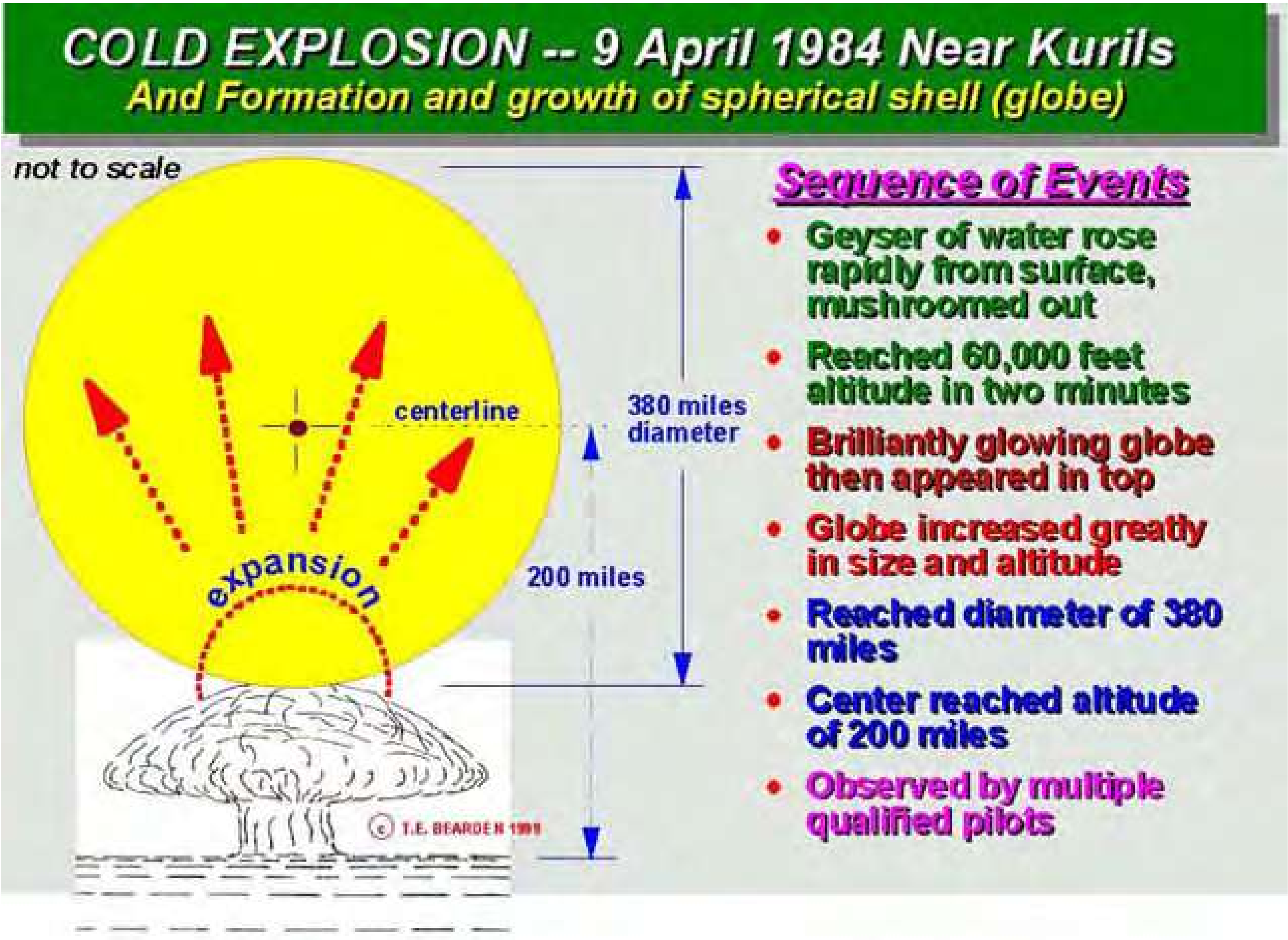

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
380 miles diameter
200 miles
expansion
© T.E. BEARDEN

Sequence of Events

- Geyser of water rose rapidly from surface, mushroomed out
- Reached 60,000 feet altitude in two minutes
- Brilliantly glowing globe then appeared in top
- Globe increased greatly in size and altitude
- Reached diameter of 380 miles
- Center reached altitude of 200 miles
- Observed by multiple qualified pilots

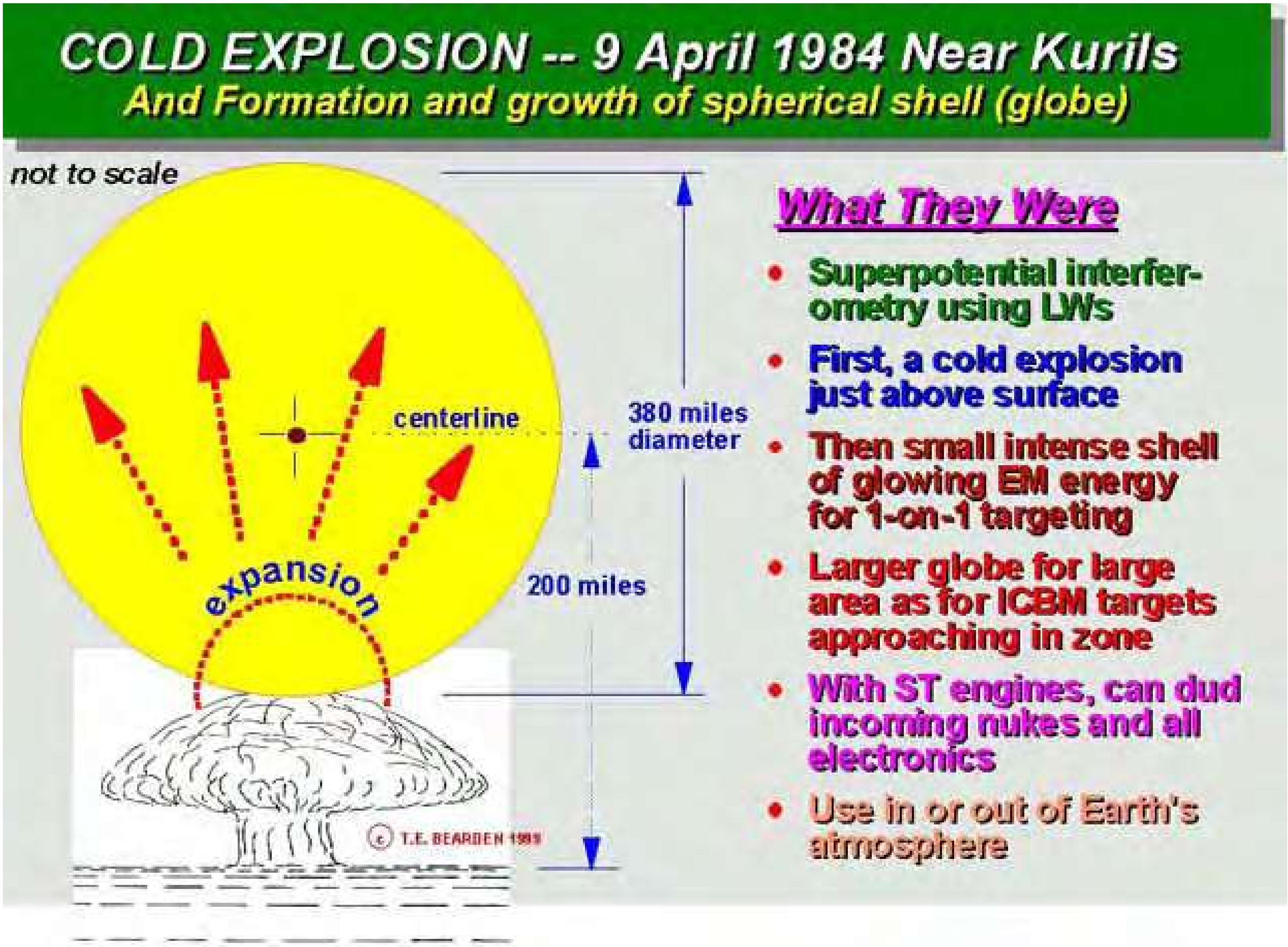

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
380 miles
diameter
200 miles
expansion
© T.E. BEARDEN 1999

What They Were

- Superpotential interferometry using LWs
- First, a cold explosion just above surface
- Then small intense shell of glowing EM energy for 1-on-1 targeting
- Larger globe for large area as for ICBM targets approaching in zone
- With ST engines, can dud incoming nukes and all electronics
- Use in or out of Earth's atmosphere

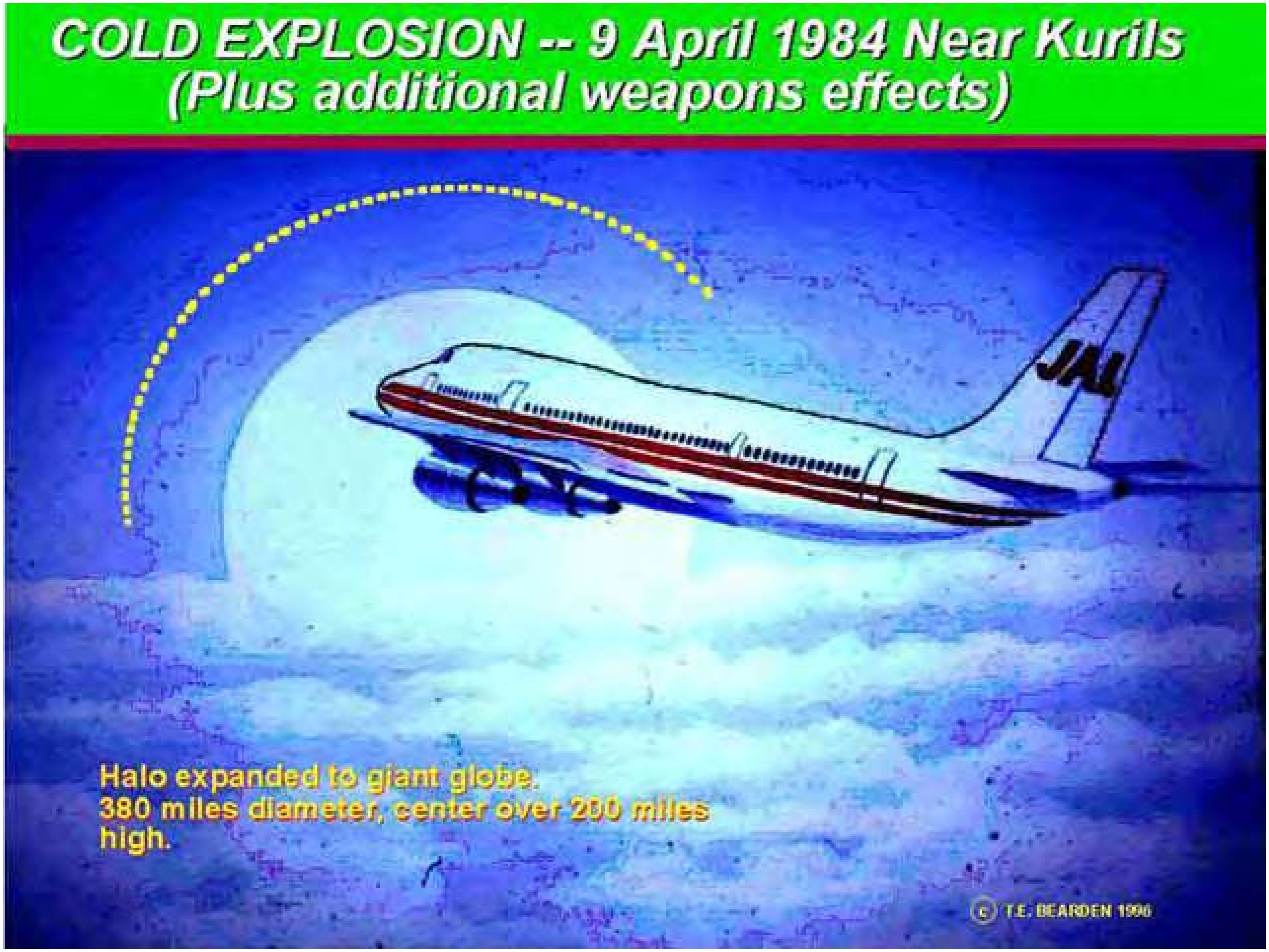

COLD EXPLOSION -- 9 April 1984 Near Kurils
(Plus additional weapons effects)

Halo expanded to giant globe.
380 miles diameter, center over 200 miles high.

© T.E. BEARDEN 1996

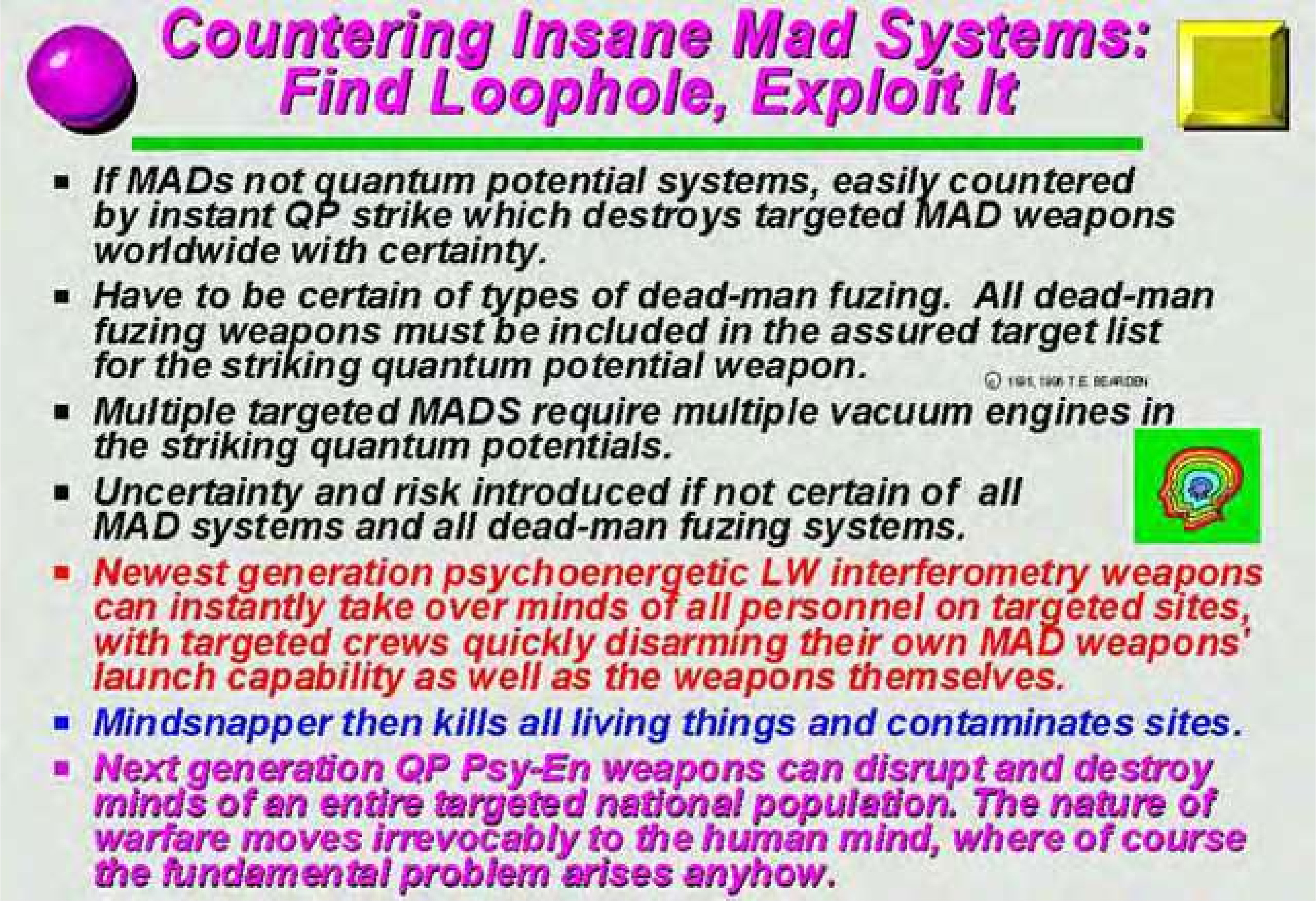

Countering Insane Mad Systems: Find Loophole, Exploit It

- *If MADs not quantum potential systems, easily countered by instant QP strike which destroys targeted MAD weapons worldwide with certainty.*
- *Have to be certain of types of dead-man fuzing. All dead-man fuzing weapons must be included in the assured target list for the striking quantum potential weapon.*

© [illegible] T.E. BEARDEN

- *Multiple targeted MADS require multiple vacuum engines in the striking quantum potentials.*
- *Uncertainty and risk introduced if not certain of all MAD systems and all dead-man fuzing systems.*
- *Newest generation psychoenergetic LW interferometry weapons can instantly take over minds of all personnel on targeted sites, with targeted crews quickly disarming their own MAD weapons' launch capability as well as the weapons themselves.*
- *Mindsnapper then kills all living things and contaminates sites.*
- *Next generation QP Psy-En weapons can disrupt and destroy minds of an entire targeted national population. The nature of warfare moves irrevocably to the human mind, where of course the fundamental problem arises anyhow.*

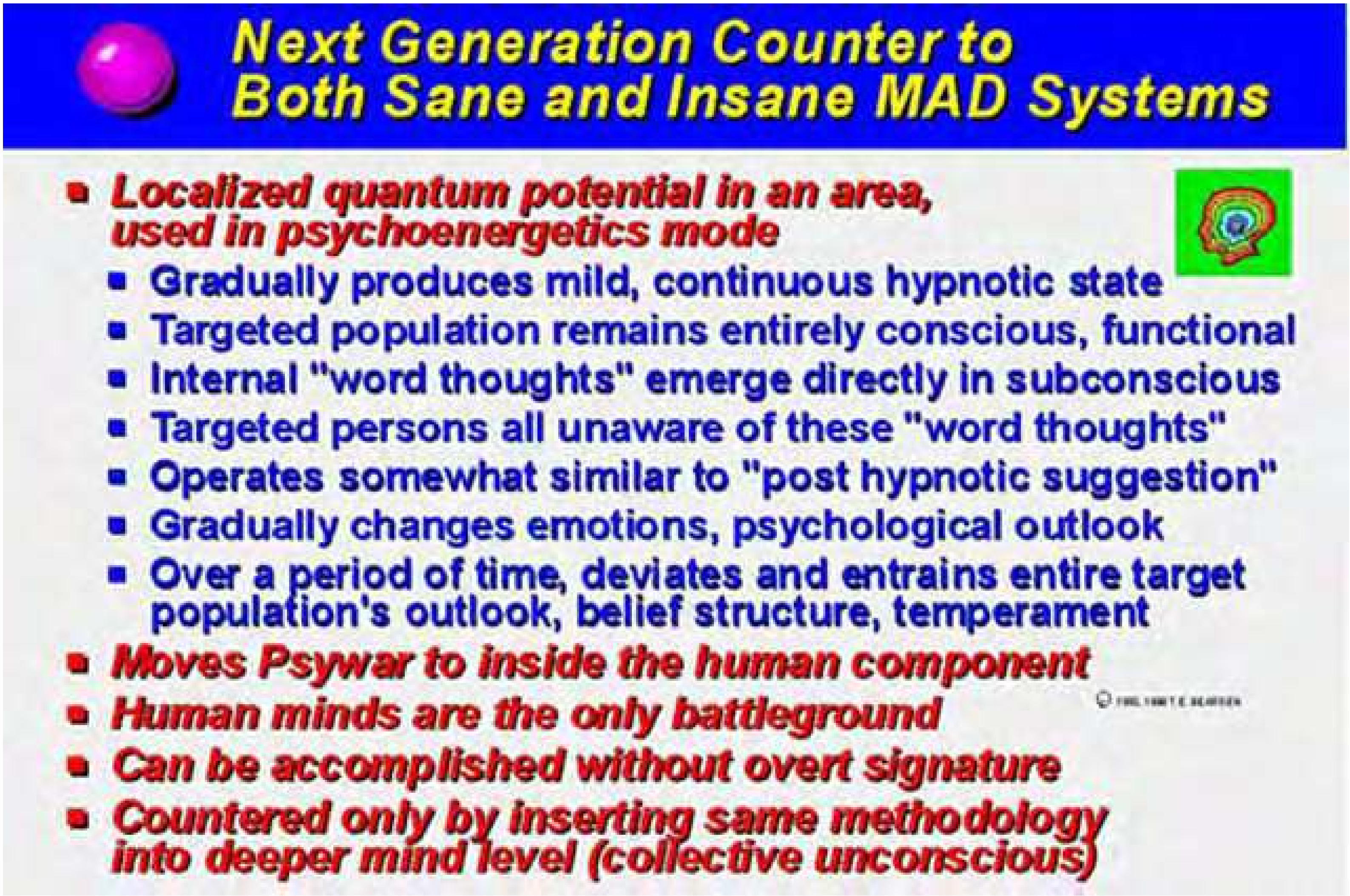

Next Generation Counter to Both Sane and Insane MAD Systems

- *Localized quantum potential in an area, used in psychoenergetics mode*
  - Gradually produces mild, continuous hypnotic state
  - Targeted population remains entirely conscious, functional
  - Internal "word thoughts" emerge directly in subconscious
  - Targeted persons all unaware of these "word thoughts"
  - Operates somewhat similar to "post hypnotic suggestion"
  - Gradually changes emotions, psychological outlook
  - Over a period of time, deviates and entrains entire target population's outlook, belief structure, temperament
- *Moves Psywar to inside the human component*
- *Human minds are the only battleground*
- *Can be accomplished without overt signature*
- *Countered only by inserting same methodology into deeper mind level (collective unconscious)*

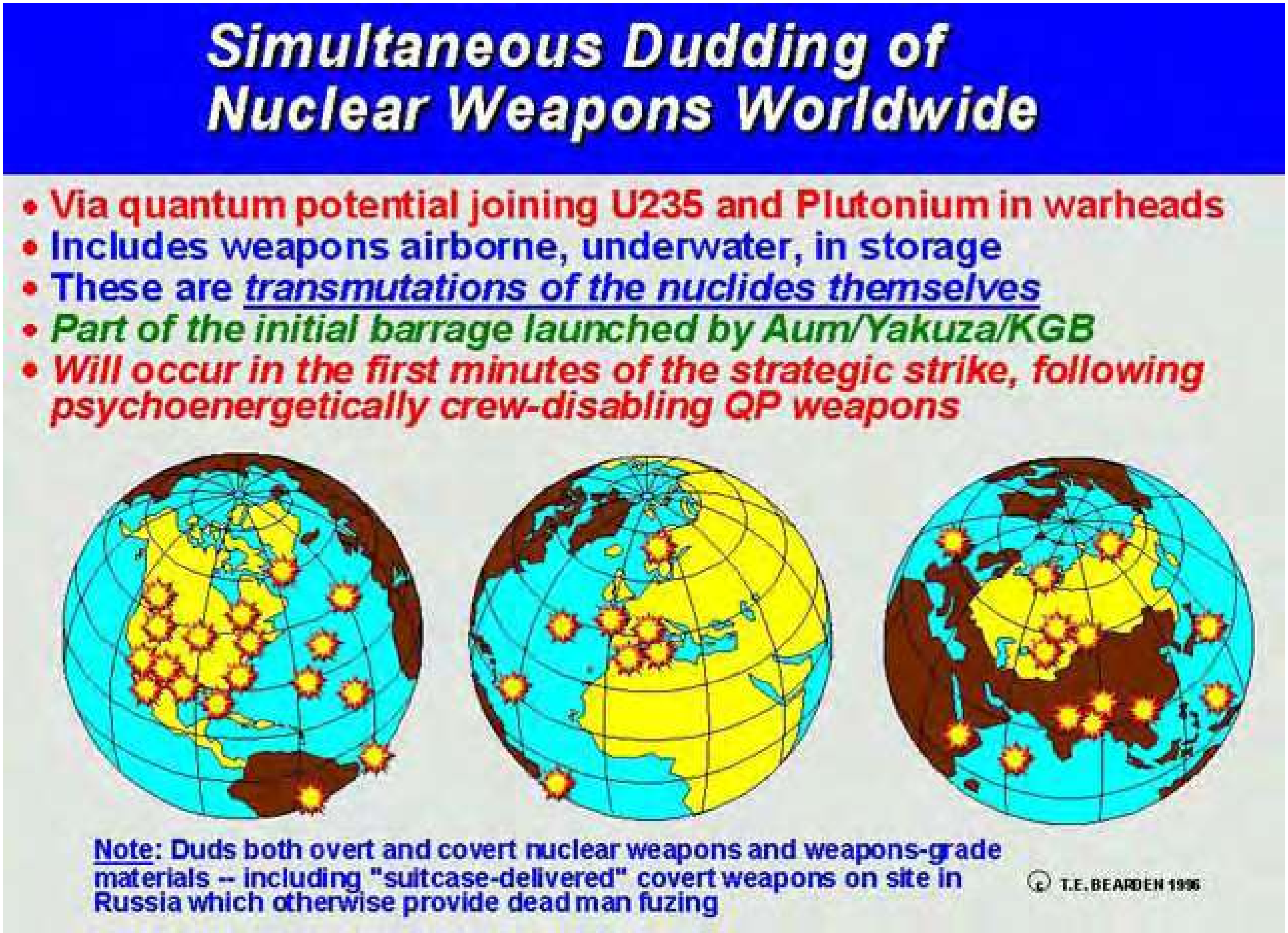

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium in warheads
- Includes weapons airborne, underwater, in storage
- These are *transmutations of the nuclides themselves*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike, following psychoenergetically crew-disabling QP weapons*

Note: Duds both overt and covert nuclear weapons and weapons-grade materials -- including "suitcase-delivered" covert weapons on site in Russia which otherwise provide dead man fuzing

© T.E. BEARDEN 1996

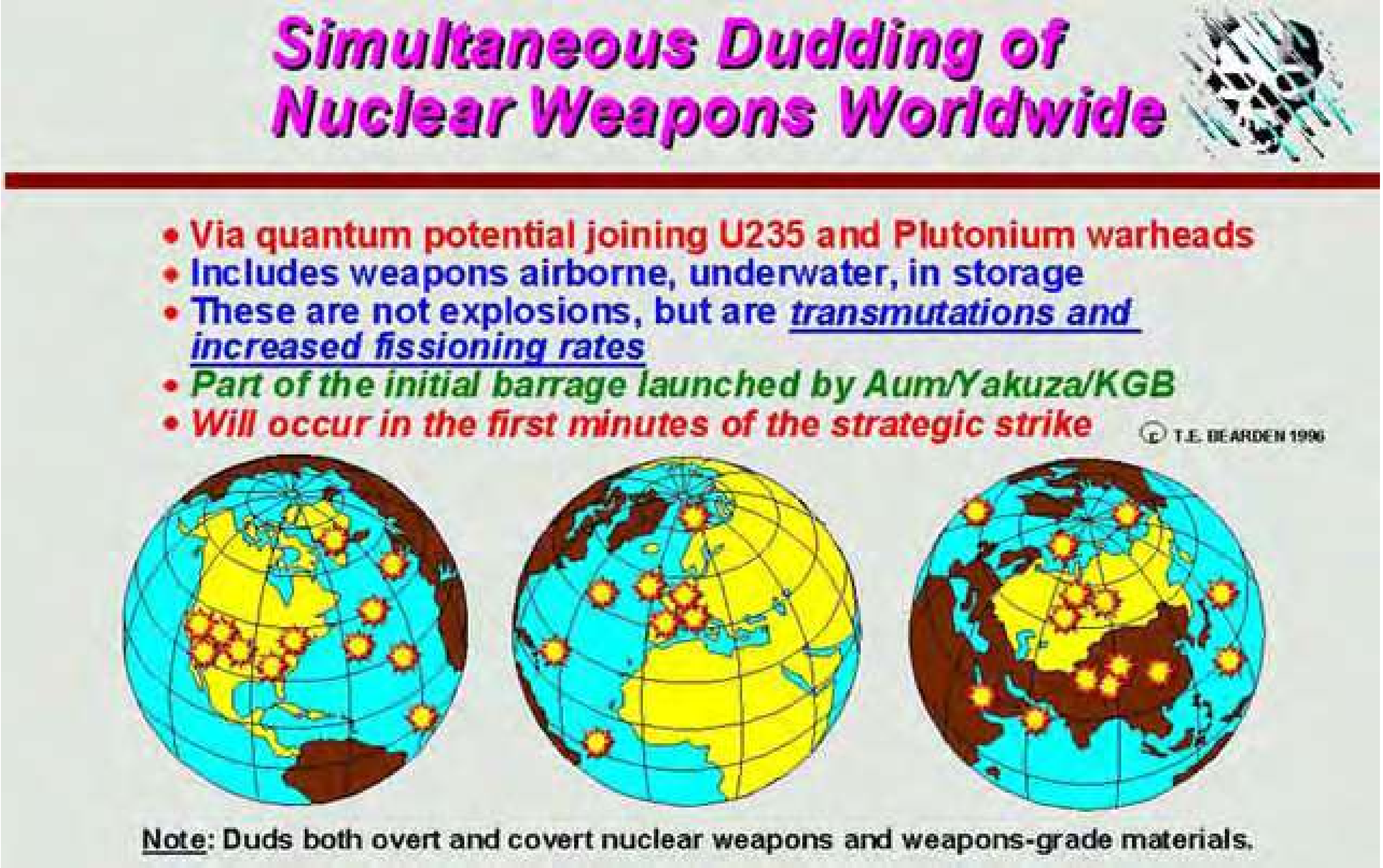

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium warheads
- Includes weapons airborne, underwater, in storage
- These are not explosions, but are *transmutations and increased fissioning rates*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike*

© T.E. BEARDEN 1996

Note: Duds both overt and covert nuclear weapons and weapons-grade materials.

The Tom Bearden Website

Help support the research

Operation Divine Wind II

Slide 1

Slide 2

Slide 3

Slide 4

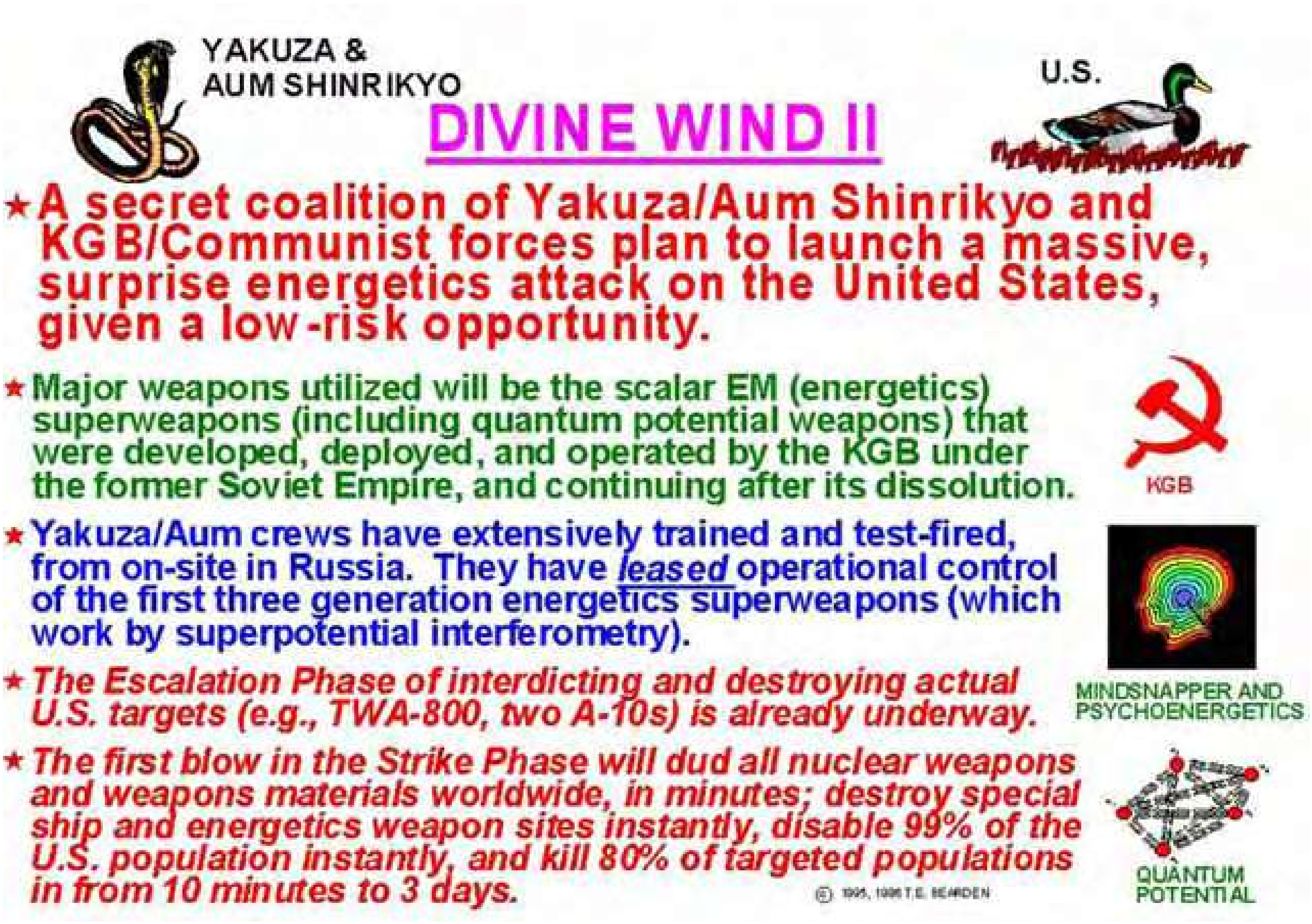
YAKUZA &
AUM SHINRIKYO
U.S.
KGB
MINDSNAPPER AND
PSYCHOENERGETICS
QUANTUM
POTENTIAL

DIVINE WIND II

- ★ **A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plan to launch a massive, surprise energetics attack on the United States, given a low-risk opportunity.**
- ★ **Major weapons utilized will be the scalar EM (energetics) superweapons (including quantum potential weapons) that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.**
- ★ **Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the first three generation energetics superweapons (which work by superpotential interferometry).**
- ★ ***The Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800, two A-10s) is already underway.***
- ★ ***The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, disable 99% of the U.S. population instantly, and kill 80% of targeted populations in from 10 minutes to 3 days.***

© 1995, 1996 T.E. BEARDEN

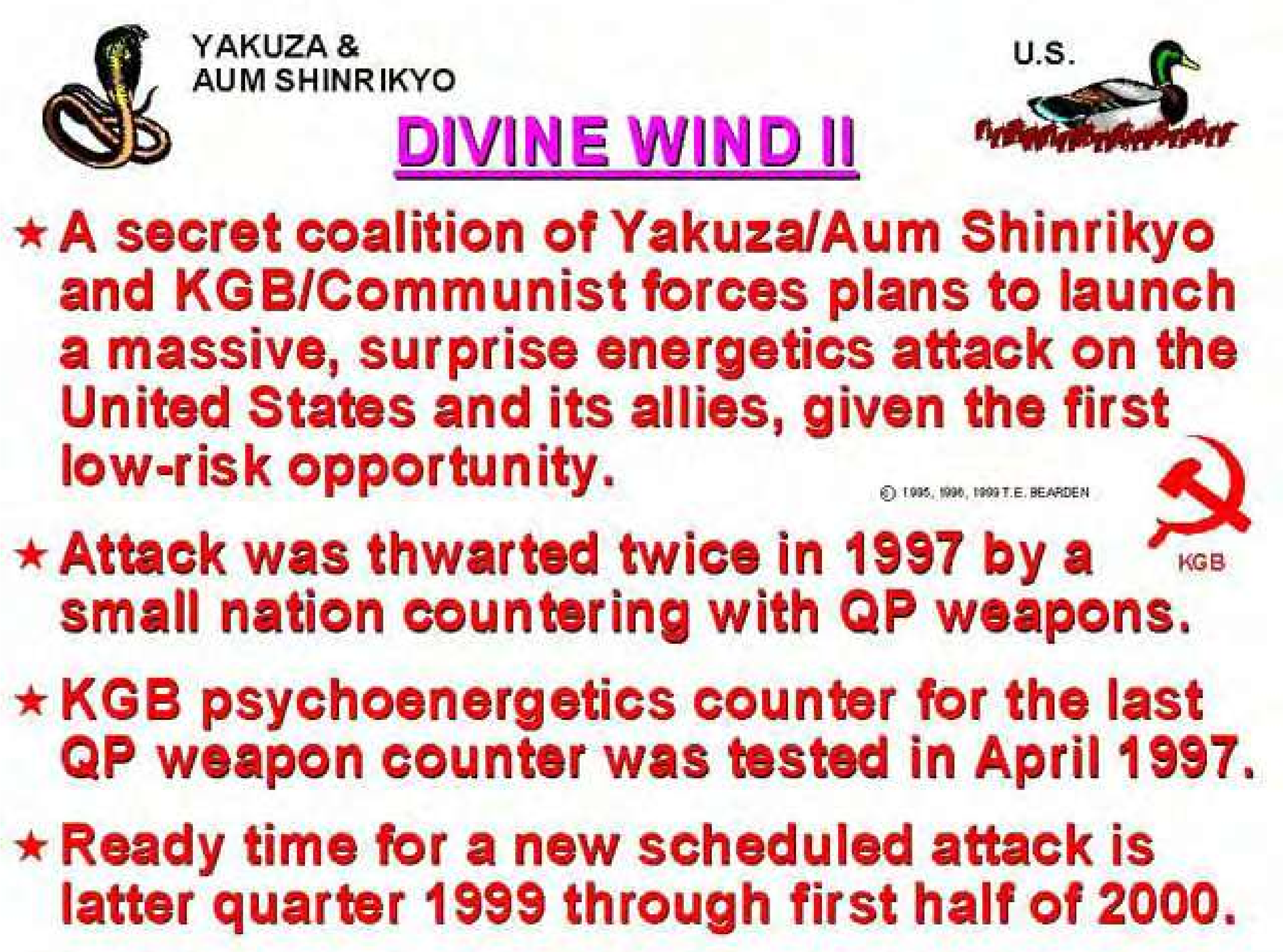

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

- ★ A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plans to launch a massive, surprise energetics attack on the United States and its allies, given the first low-risk opportunity.

© 1995, 1996, 1999 T.E. BEARDEN

KGB

- ★ Attack was thwarted twice in 1997 by a small nation countering with QP weapons.
- ★ KGB psychoenergetics counter for the last QP weapon counter was tested in April 1997.
- ★ Ready time for a new scheduled attack is latter quarter 1999 through first half of 2000.

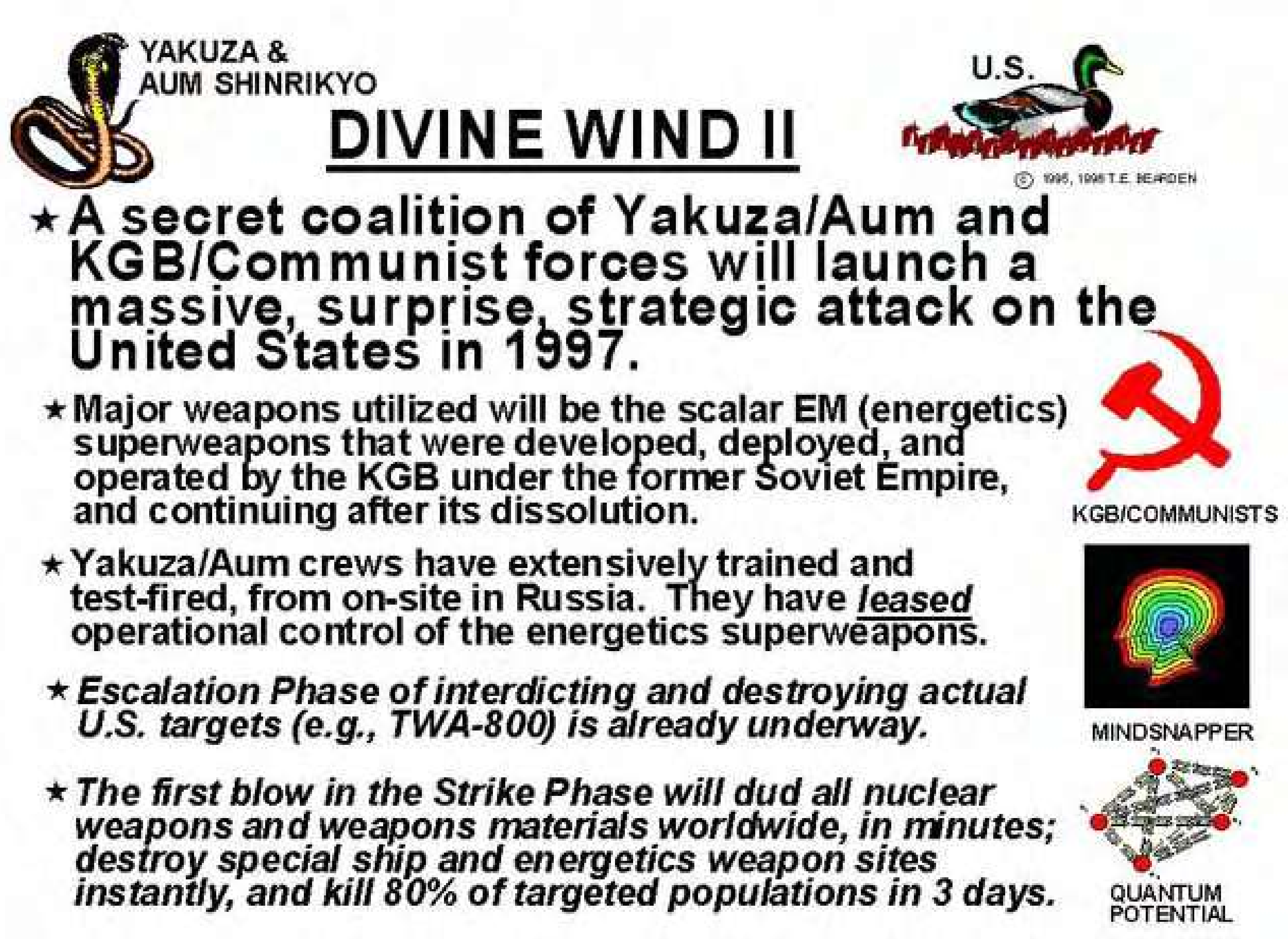

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

© 1995, 1996 T.E. BEARDEN

★ **A secret coalition of Yakuza/Aum and KGB/Communist forces will launch a massive, surprise, strategic attack on the United States in 1997.**

★ Major weapons utilized will be the scalar EM (energetics) superweapons that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.

★ Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the energetics superweapons.

★ *Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800) is already underway.*

★ *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, and kill 80% of targeted populations in 3 days.*

KGB/COMMUNISTS

MINDSNAPPER

QUANTUM
POTENTIAL

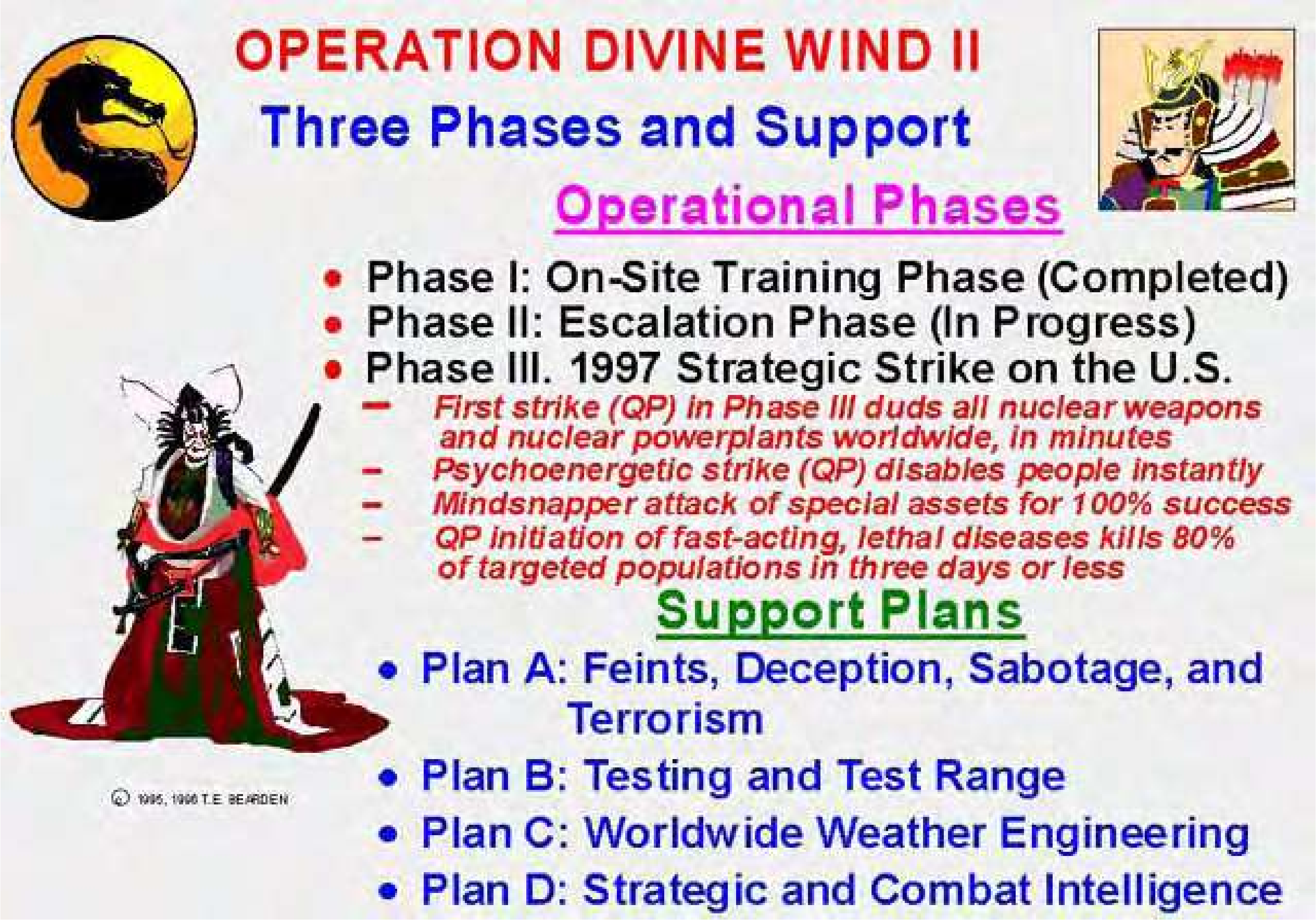

OPERATION DIVINE WIND II

Three Phases and Support

Operational Phases

- Phase I: On-Site Training Phase (Completed)
- Phase II: Escalation Phase (In Progress)
- Phase III. 1997 Strategic Strike on the U.S.
  - *First strike (QP) in Phase III duds all nuclear weapons and nuclear powerplants worldwide, in minutes*
  - *Psychoenergetic strike (QP) disables people instantly*
  - *Mindsnapper attack of special assets for 100% success*
  - *QP initiation of fast-acting, lethal diseases kills 80% of targeted populations in three days or less*

Support Plans

- Plan A: Feints, Deception, Sabotage, and Terrorism
- Plan B: Testing and Test Range
- Plan C: Worldwide Weather Engineering
- Plan D: Strategic and Combat Intelligence

© 1995, 1996 T.E. BEARDEN

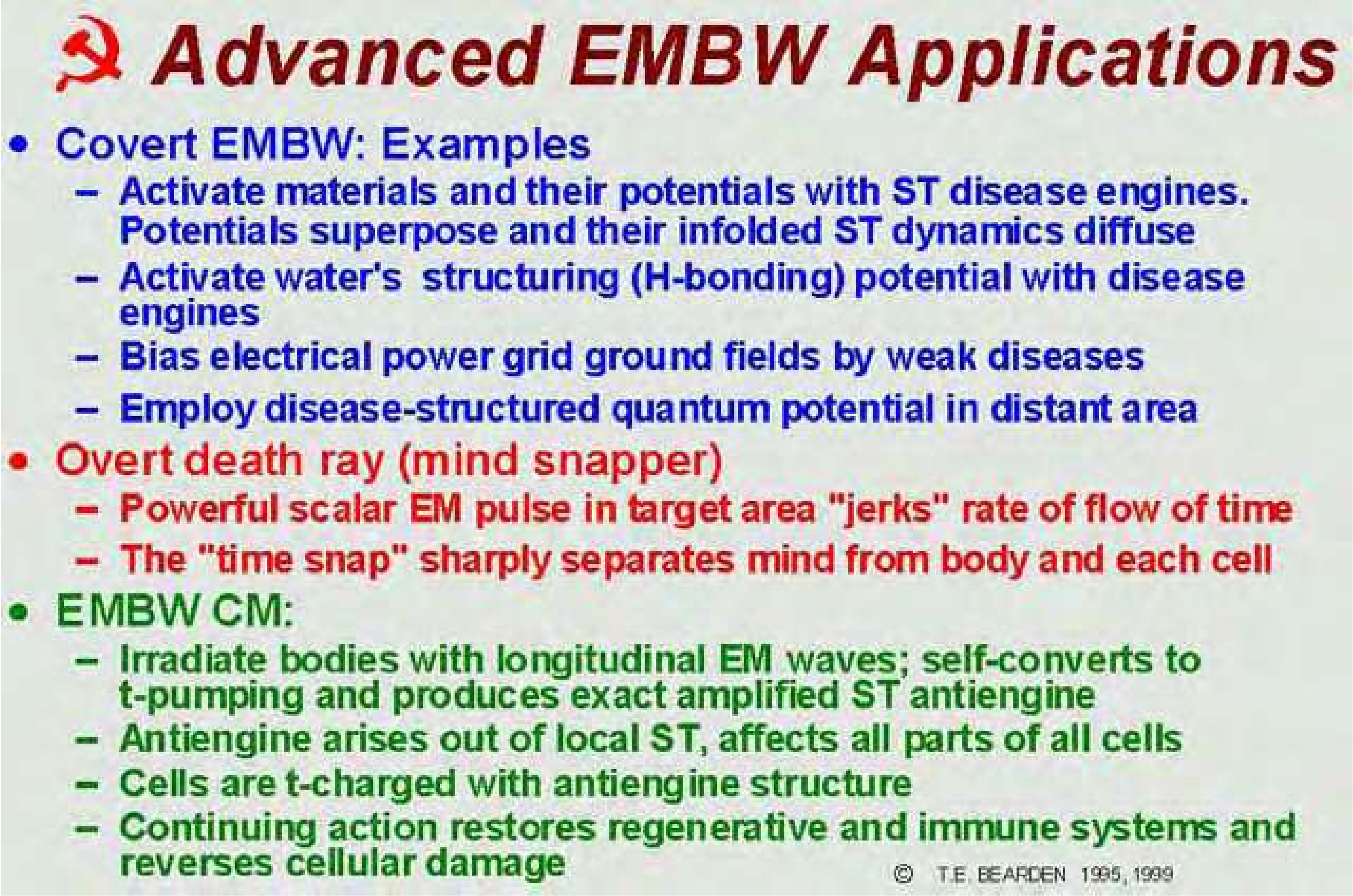

Advanced EMBW Applications

- Covert EMBW: Examples
  - Activate materials and their potentials with ST disease engines. Potentials superpose and their infolded ST dynamics diffuse
  - Activate water's structuring (H-bonding) potential with disease engines
  - Bias electrical power grid ground fields by weak diseases
  - Employ disease-structured quantum potential in distant area
- Overt death ray (mind snapper)
  - Powerful scalar EM pulse in target area "jerks" rate of flow of time
  - The "time snap" sharply separates mind from body and each cell
- EMBW CM:
  - Irradiate bodies with longitudinal EM waves; self-converts to t-pumping and produces exact amplified ST antiengine
  - Antiengine arises out of local ST, affects all parts of all cells
  - Cells are t-charged with antiengine structure
  - Continuing action restores regenerative and immune systems and reverses cellular damage

© T.E. BEARDEN 1995, 1999

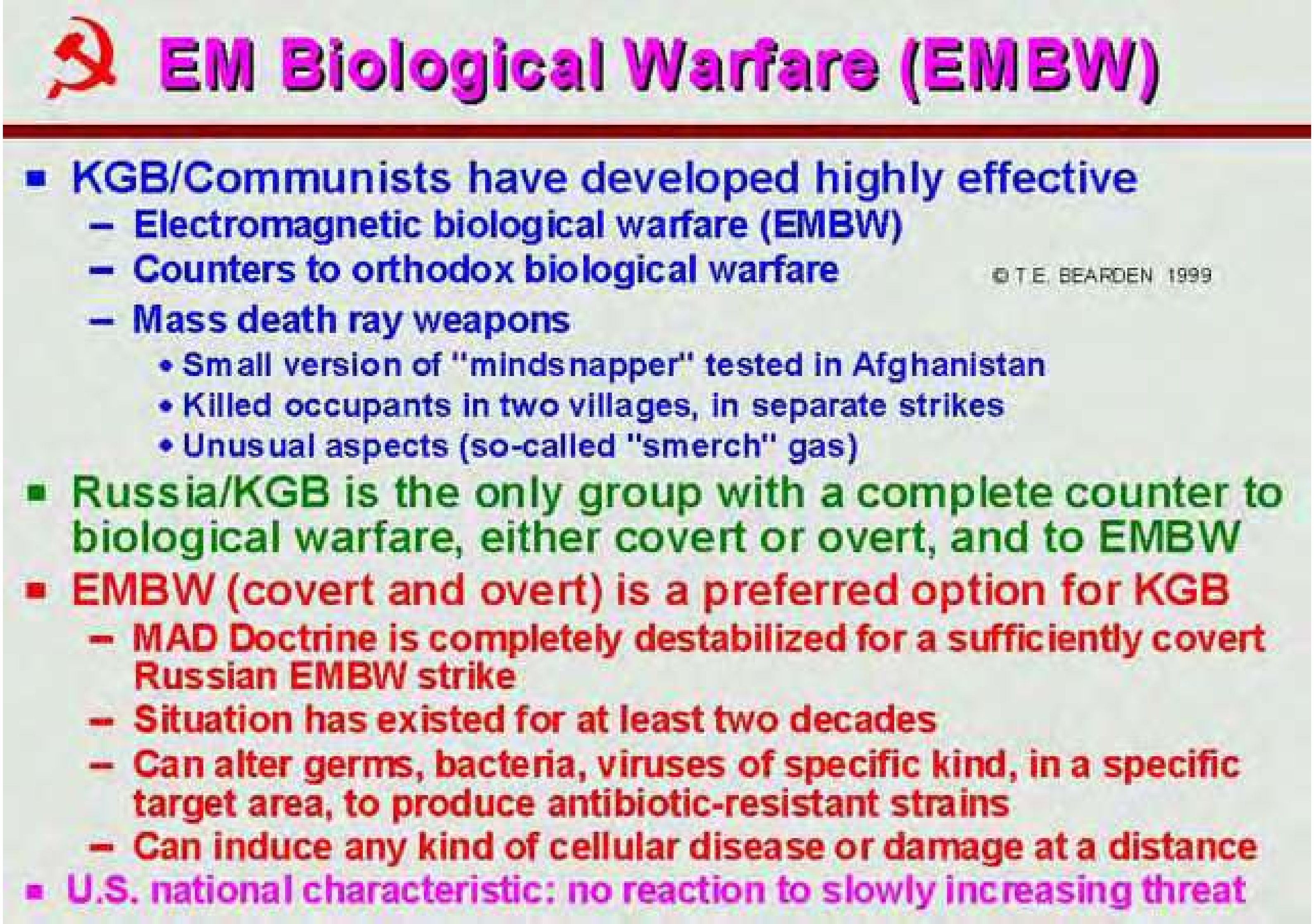

EM Biological Warfare (EMBW)

- KGB/Communists have developed highly effective
  - Electromagnetic biological warfare (EMBW)
  - Counters to orthodox biological warfare
  - Mass death ray weapons
    - Small version of "mindsnapper" tested in Afghanistan
    - Killed occupants in two villages, in separate strikes
    - Unusual aspects (so-called "smerch" gas)
- Russia/KGB is the only group with a complete counter to biological warfare, either covert or overt, and to EMBW
- EMBW (covert and overt) is a preferred option for KGB
  - MAD Doctrine is completely destabilized for a sufficiently covert Russian EMBW strike
  - Situation has existed for at least two decades
  - Can alter germs, bacteria, viruses of specific kind, in a specific target area, to produce antibiotic-resistant strains
  - Can induce any kind of cellular disease or damage at a distance
- U.S. national characteristic: no reaction to slowly increasing threat

© T.E. BEARDEN 1999

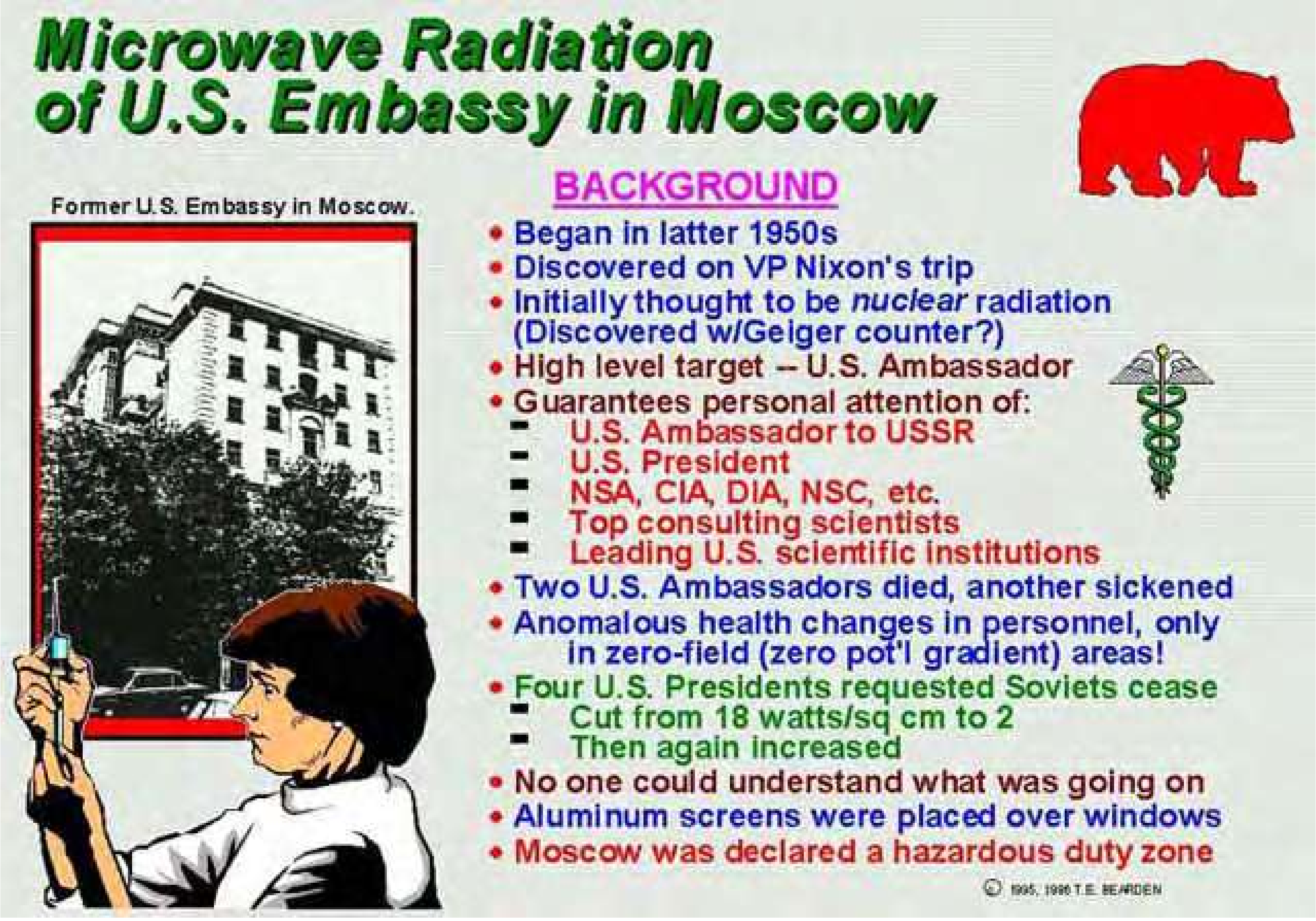

Microwave Radiation of U.S. Embassy in Moscow

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

© 1995, 1996 T.E. BEARDEN

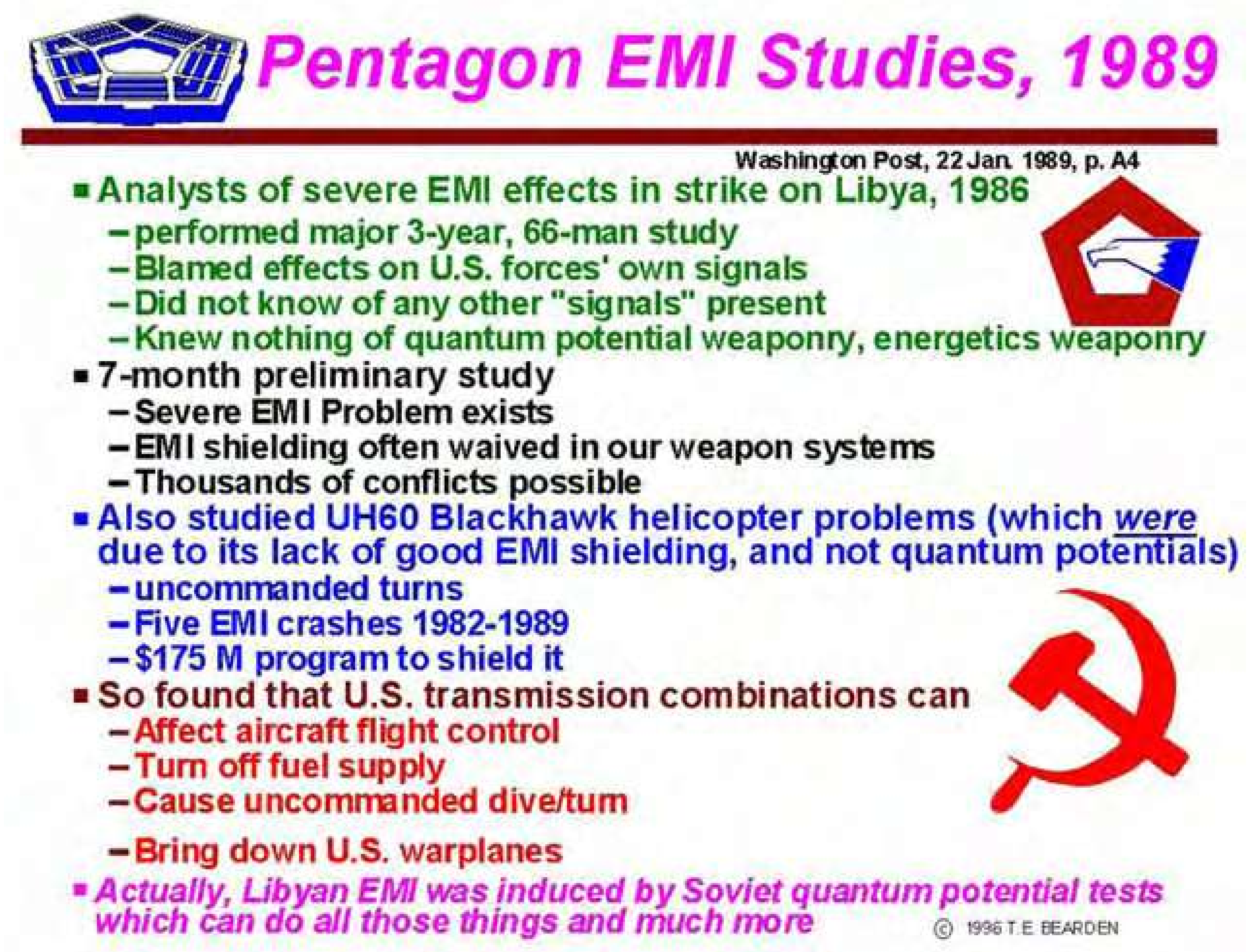

Pentagon EMI Studies, 1989

Washington Post, 22 Jan. 1989, p. A4

- Analysts of severe EMI effects in strike on Libya, 1986
  - performed major 3-year, 66-man study
  - Blamed effects on U.S. forces' own signals
  - Did not know of any other "signals" present
  - Knew nothing of quantum potential weaponry, energetics weaponry
- 7-month preliminary study
  - Severe EMI Problem exists
  - EMI shielding often waived in our weapon systems
  - Thousands of conflicts possible
- Also studied UH60 Blackhawk helicopter problems (which *were* due to its lack of good EMI shielding, and not quantum potentials)
  - uncommanded turns
  - Five EMI crashes 1982-1989
  - $175 M program to shield it
- So found that U.S. transmission combinations can
  - Affect aircraft flight control
  - Turn off fuel supply
  - Cause uncommanded dive/turn
  - Bring down U.S. warplanes
- *Actually, Libyan EMI was induced by Soviet quantum potential tests which can do all those things and much more*

© 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

EM Missile Practice

- EM Missile Practice
  Latter Sept. - Early Nov. 1996
  Melbourne, Australia
- Deception: Using EM Missile to Reinforce Notion of SAM Attack
  Saturday evening, Nov. 16, 1996
- Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry
  Thursday, December 12th, 1996
- EM Missile Strike, Offset from Night Shuttle Launch, Cape Canaveral, Florida
  The Smoking Gun in November 1985
- Distant Operator Onsite in Russia Slew Away Registration Point
  Marker Beacon over Night Shuttle Launch, 26 November, 1985
- EM Missile Practice and Deception
  November 17, 1996

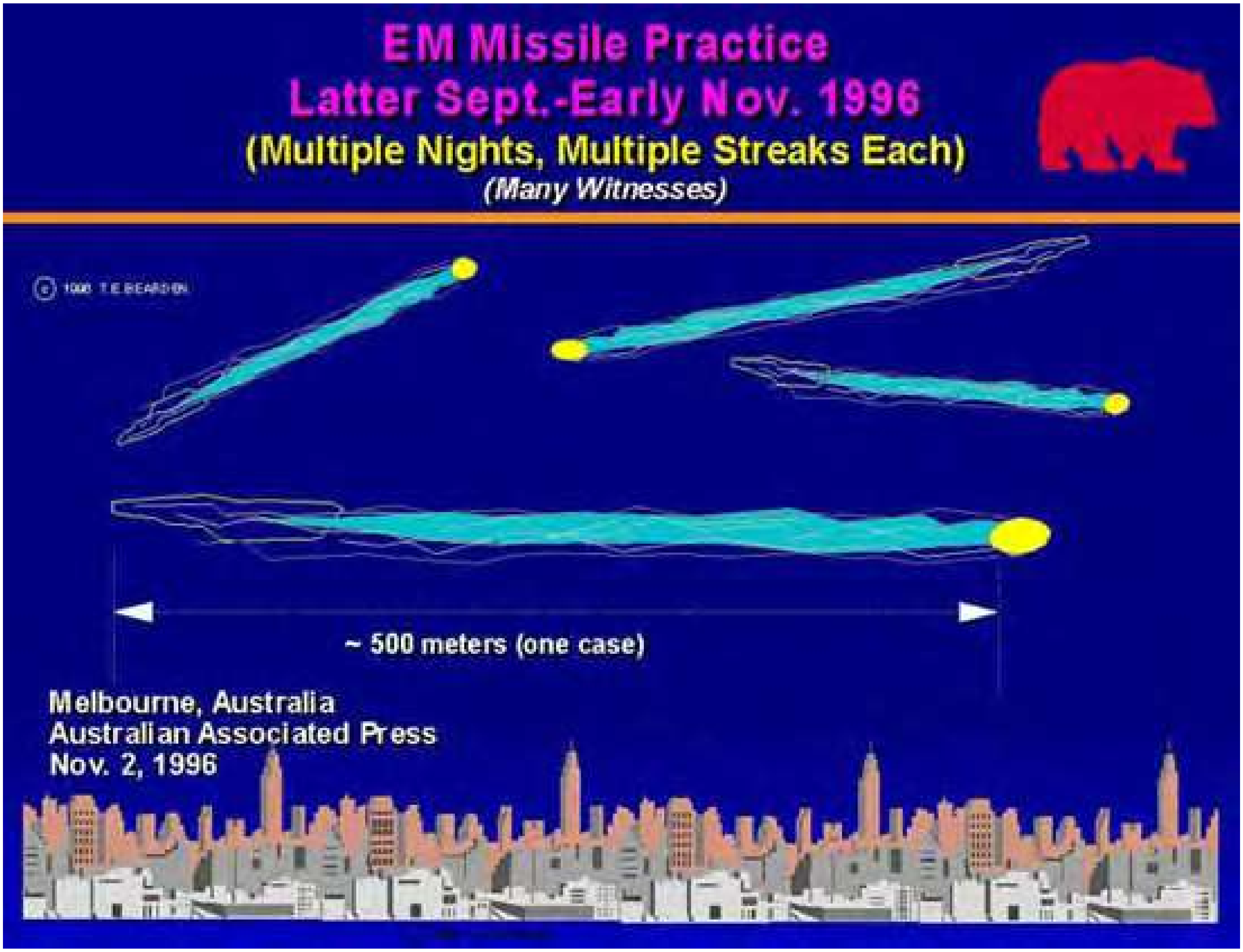

EM Missile Practice
Latter Sept.-Early Nov. 1996

(Multiple Nights, Multiple Streaks Each)

(Many Witnesses)

© 1996 T.E. BEARDEN

~ 500 meters (one case)

Melbourne, Australia
Australian Associated Press
Nov. 2, 1996

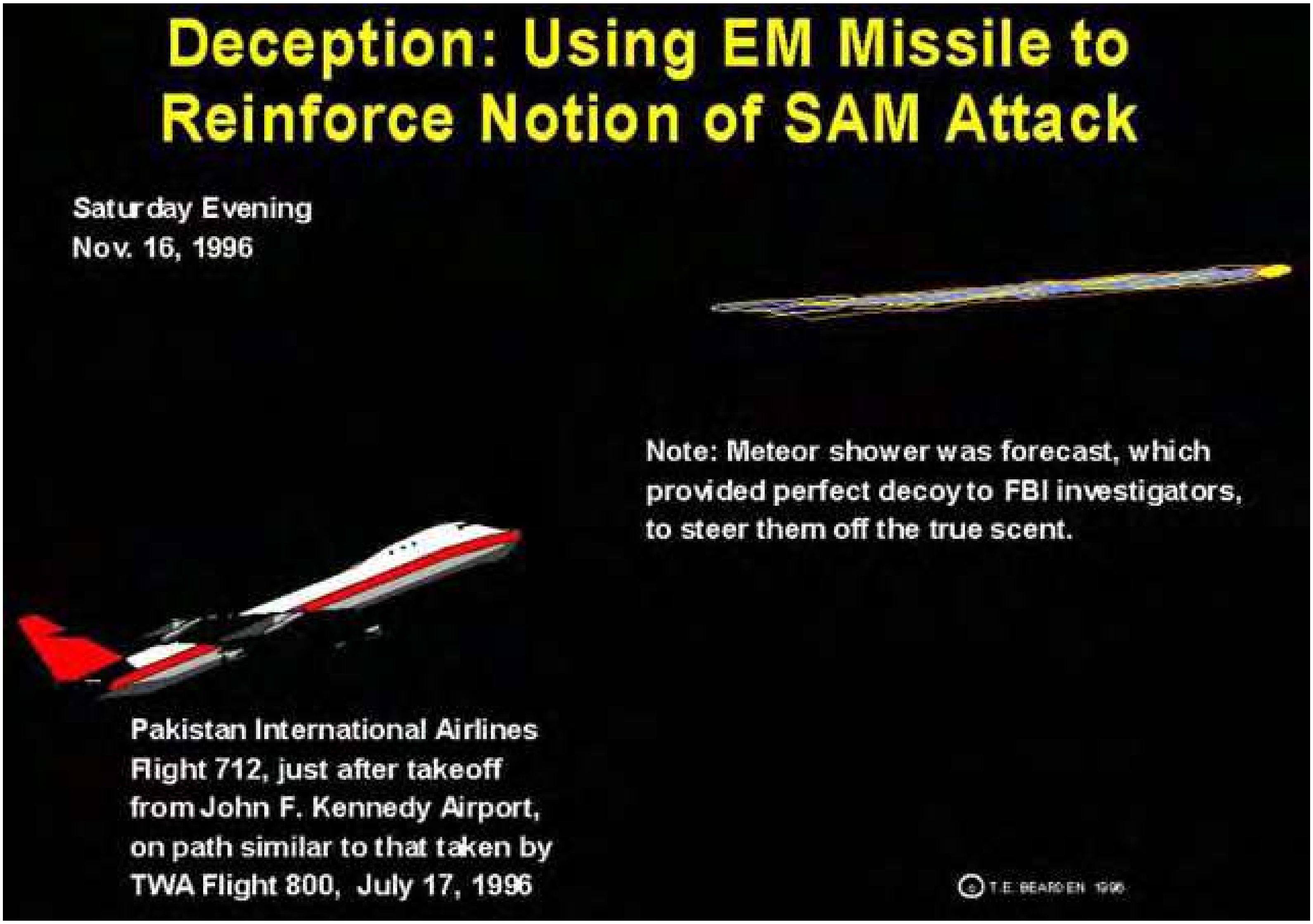

Deception: Using EM Missile to Reinforce Notion of SAM Attack

Saturday Evening
Nov. 16, 1996

Note: Meteor shower was forecast, which provided perfect decoy to FBI investigators, to steer them off the true scent.

Pakistan International Airlines Flight 712, just after takeoff from John F. Kennedy Airport, on path similar to that taken by TWA Flight 800, July 17, 1996

© T.E. BEARDEN 1996

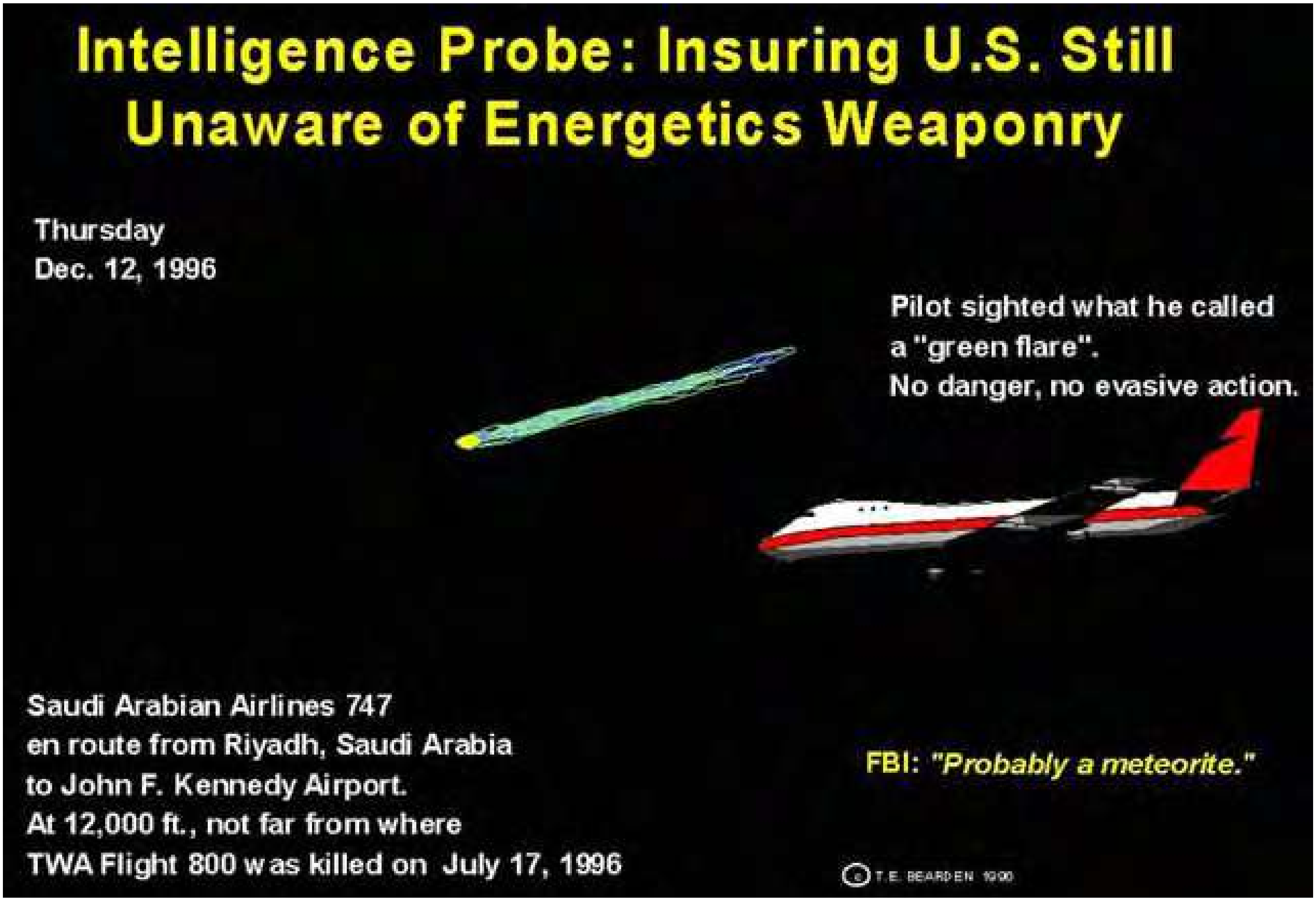

Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry

Thursday
Dec. 12, 1996

Pilot sighted what he called
a "green flare".
No danger, no evasive action.

Saudi Arabian Airlines 747
en route from Riyadh, Saudi Arabia
to John F. Kennedy Airport.
At 12,000 ft., not far from where
TWA Flight 800 was killed on July 17, 1996

FBI: *"Probably a meteorite."*

© T.E. BEARDEN 1990

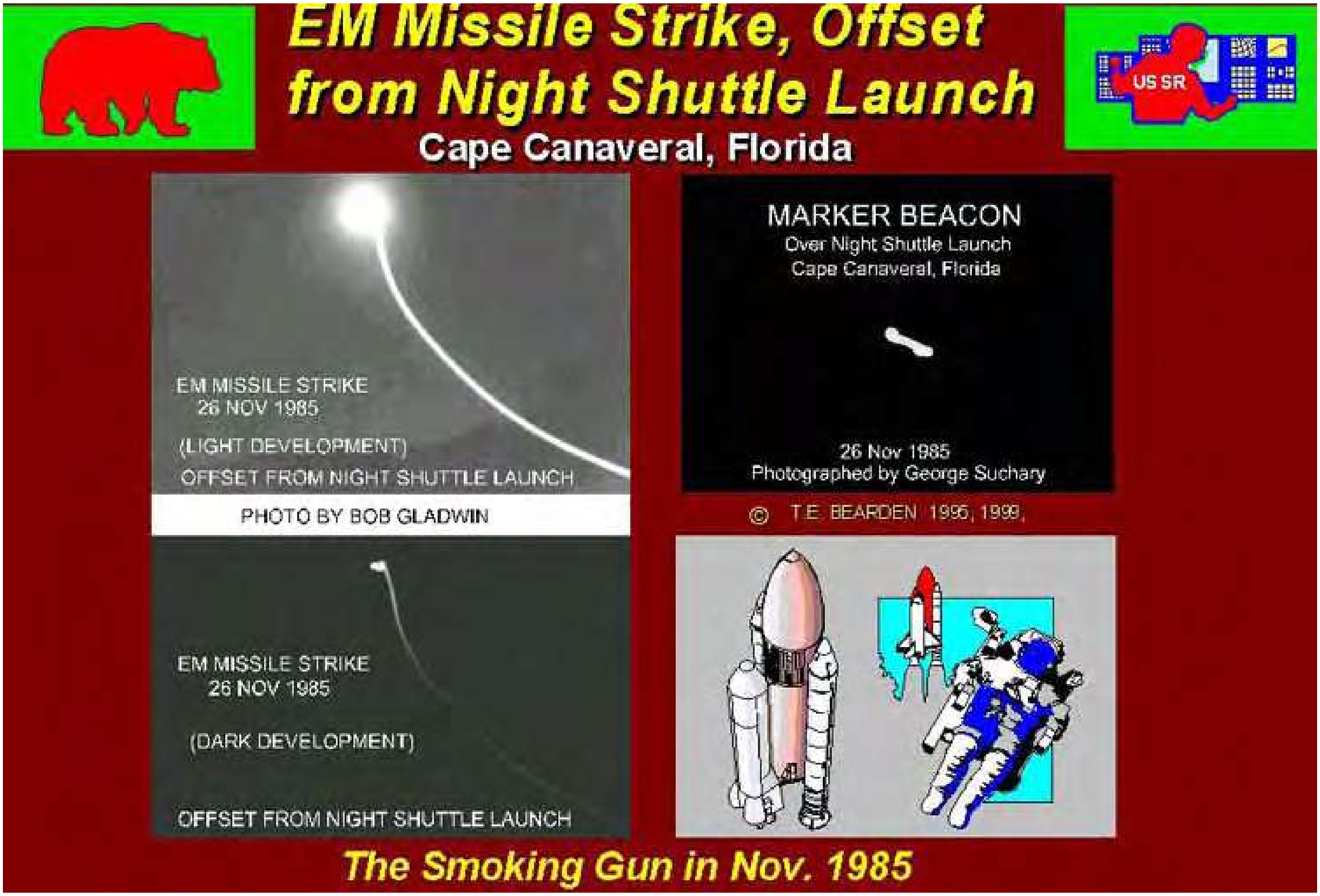

EM Missile Strike, Offset from Night Shuttle Launch

Cape Canaveral, Florida

EM MISSILE STRIKE
26 NOV 1985

(LIGHT DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

PHOTO BY BOB GLADWIN

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

© T.E. BEARDEN 1995, 1999,

EM MISSILE STRIKE
26 NOV 1985

(DARK DEVELOPMENT)

OFFSET FROM NIGHT SHUTTLE LAUNCH

The Smoking Gun in Nov. 1985

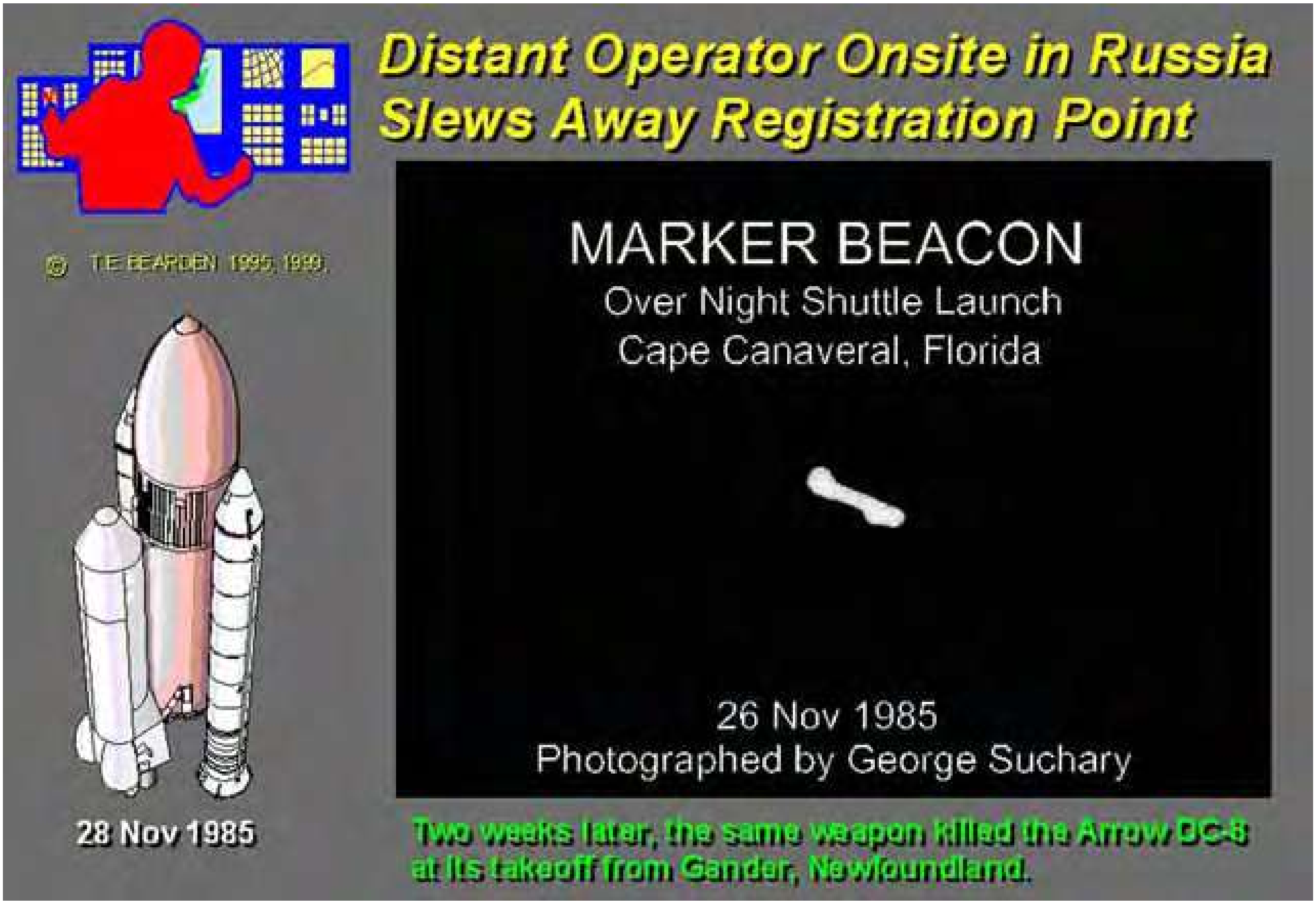

Distant Operator Onsite in Russia Slews Away Registration Point

© T.E. BEARDEN 1995, 1999.

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

28 Nov 1985

Two weeks later, the same weapon killed the Arrow DC-8 at its takeoff from Gander, Newfoundland.

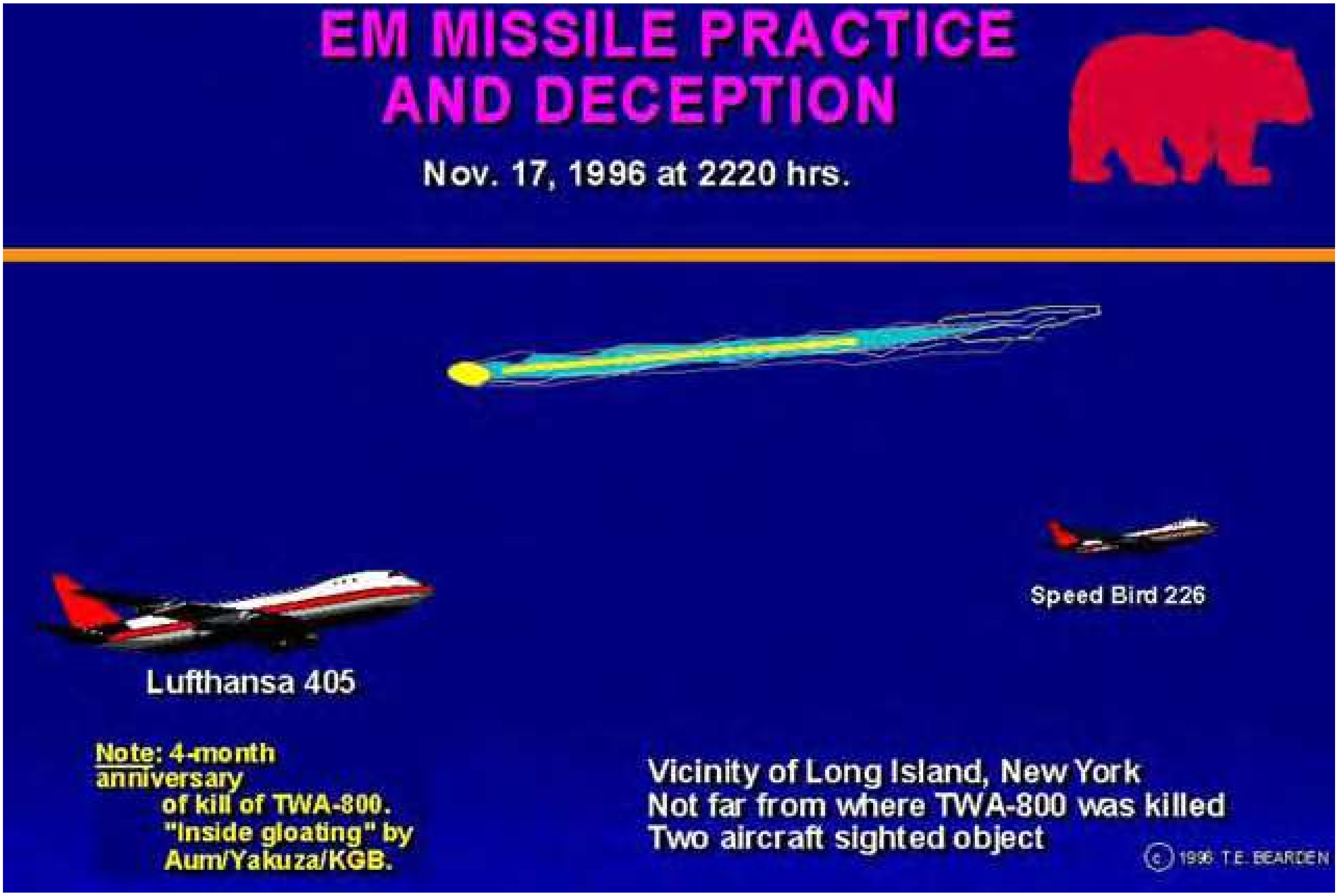

EM MISSILE PRACTICE AND DECEPTION

Nov. 17, 1996 at 2220 hrs.

Speed Bird 226

Lufthansa 405

Note: 4-month anniversary
of kill of TWA-800.
"Inside gloating" by
Aum/Yakuza/KGB.

Vicinity of Long Island, New York
Not far from where TWA-800 was killed
Two aircraft sighted object

(c) 1996 T.E. BEARDEN

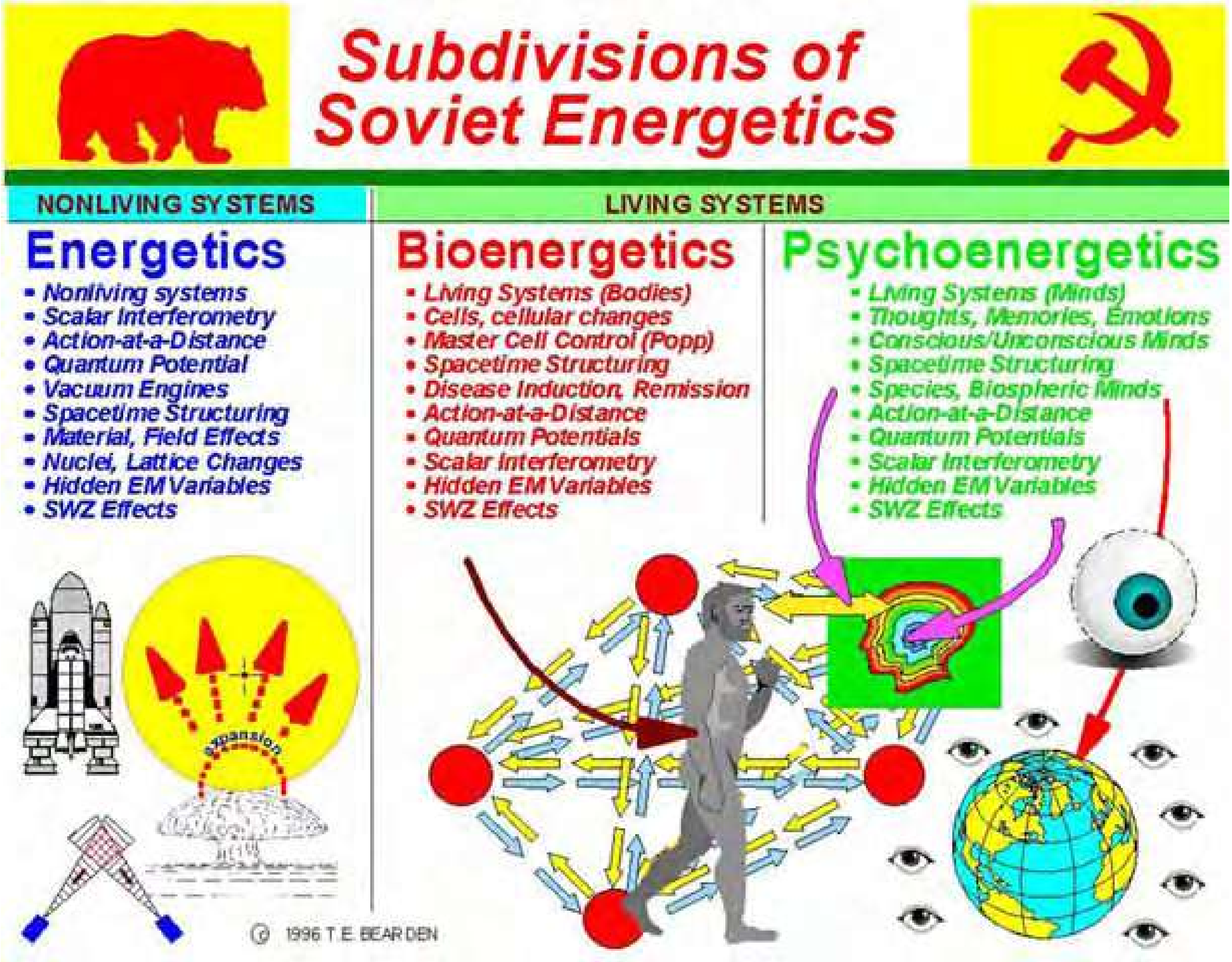

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

LIVING SYSTEMS

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T.E. BEARDEN

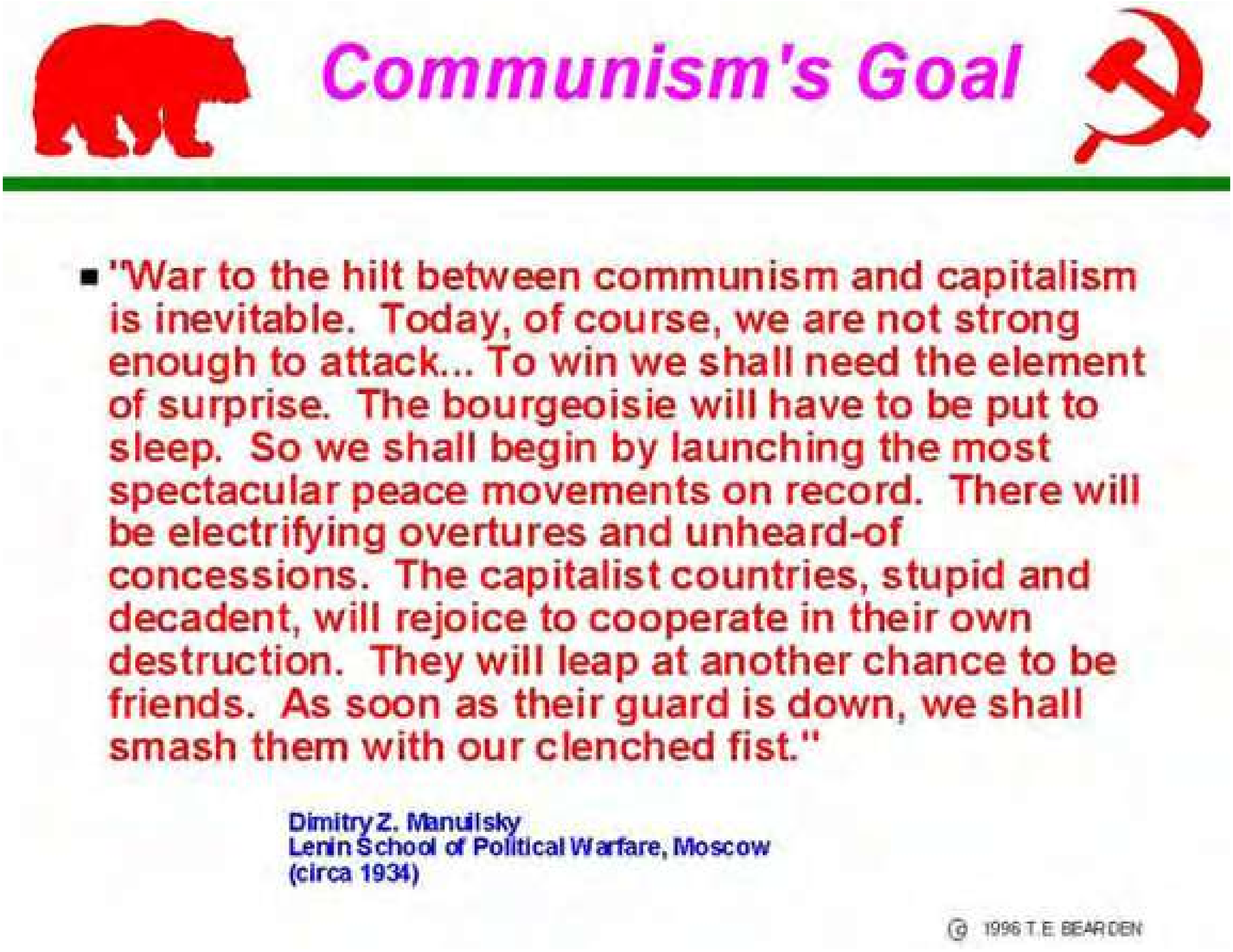

Communism's Goal

- "War to the hilt between communism and capitalism is inevitable. Today, of course, we are not strong enough to attack... To win we shall need the element of surprise. The bourgeoisie will have to be put to sleep. So we shall begin by launching the most spectacular peace movements on record. There will be electrifying overtures and unheard-of concessions. The capitalist countries, stupid and decadent, will rejoice to cooperate in their own destruction. They will leap at another chance to be friends. As soon as their guard is down, we shall smash them with our clenched fist."

Dimitry Z. Manuilsky
Lenin School of Political Warfare, Moscow
(circa 1934)

© 1996 T.E. BEARDEN

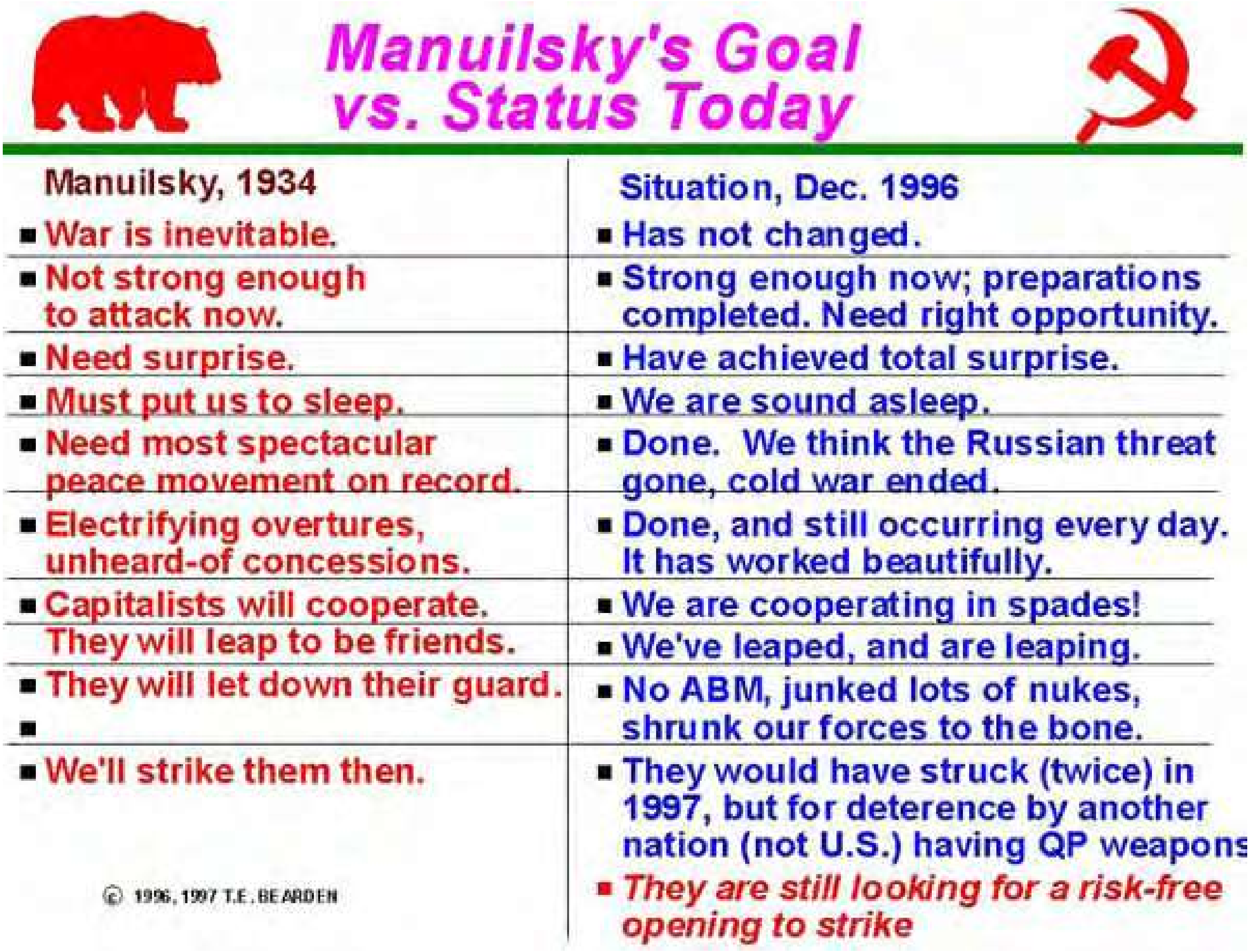

Manuilsky's Goal vs. Status Today

Manuilsky, 1934	Situation, Dec. 1996
■ War is inevitable.	■ Has not changed.
■ Not strong enough to attack now.	■ Strong enough now; preparations completed. Need right opportunity.
■ Need surprise.	■ Have achieved total surprise.
■ Must put us to sleep.	■ We are sound asleep.
■ Need most spectacular peace movement on record.	■ Done. We think the Russian threat gone, cold war ended.
■ Electrifying overtures, unheard-of concessions.	■ Done, and still occurring every day. It has worked beautifully.
■ Capitalists will cooperate.	■ We are cooperating in spades!
They will leap to be friends.	■ We've leaped, and are leaping.
■ They will let down their guard. ■	■ No ABM, junked lots of nukes, shrunk our forces to the bone.
■ We'll strike them then.	■ They would have struck (twice) in 1997, but for deterence by another nation (not U.S.) having QP weapons ■ *They are still looking for a risk-free opening to strike*

© 1996, 1997 T.E. BEARDEN

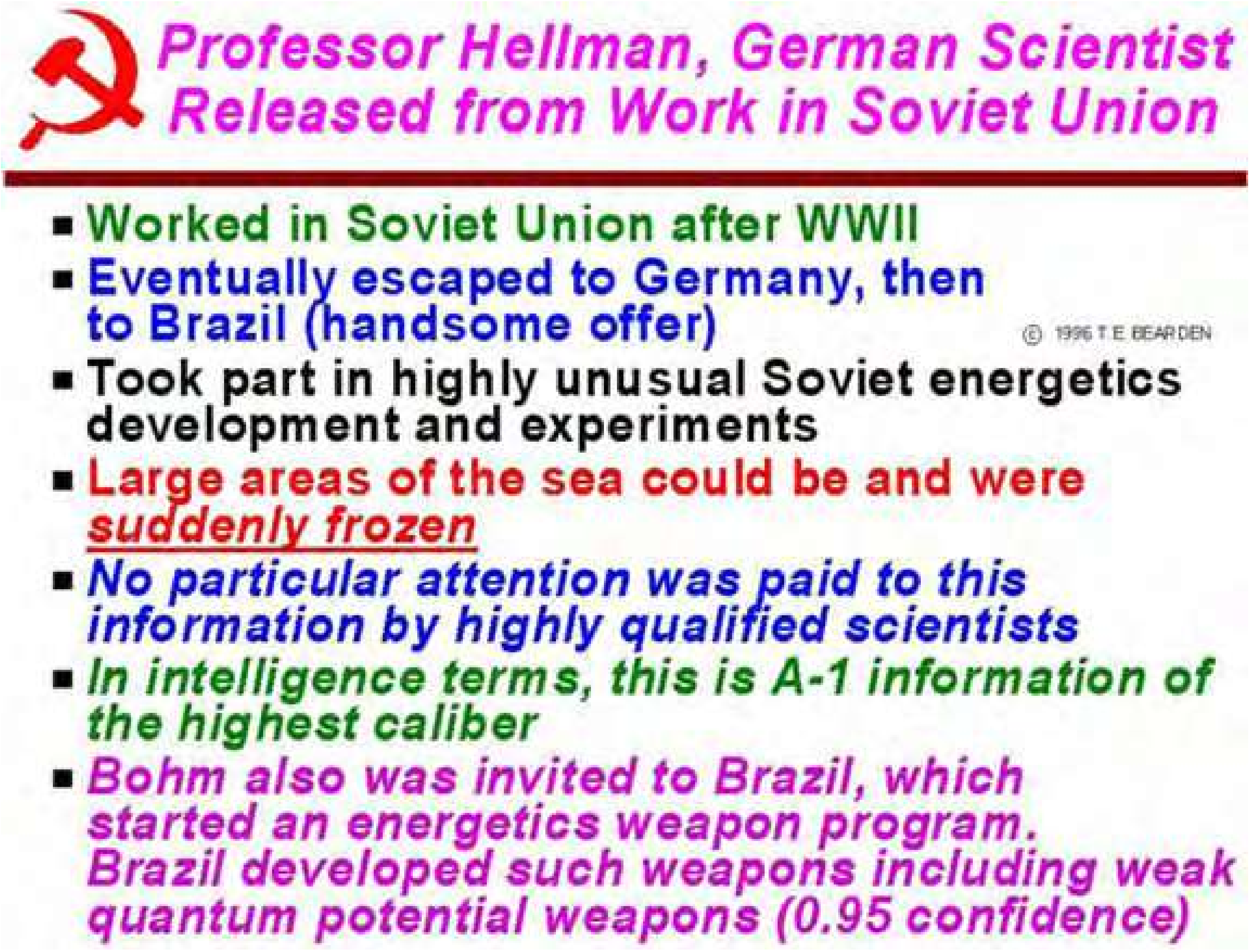

Professor Hellman, German Scientist Released from Work in Soviet Union

- Worked in Soviet Union after WWII
- Eventually escaped to Germany, then to Brazil (handsome offer)

© 1996 T. E. BEARDEN

- Took part in highly unusual Soviet energetics development and experiments
- Large areas of the sea could be and were ***suddenly frozen***
- *No particular attention was paid to this information by highly qualified scientists*
- *In intelligence terms, this is A-1 information of the highest caliber*
- *Bohm also was invited to Brazil, which started an energetics weapon program. Brazil developed such weapons including weak quantum potential weapons (0.95 confidence)*

The Tom Bearden Website

Help support the research

Hemispheres of Light

- March 20th, 1969
  Caribbean and Western North Atlantic
- Hemisphere and Globes, 1977
  24 March, 1977
- Virgin Islands Incident
  August 1969
- Typical Shield Test, Northern Route to Japan
  June, 1982

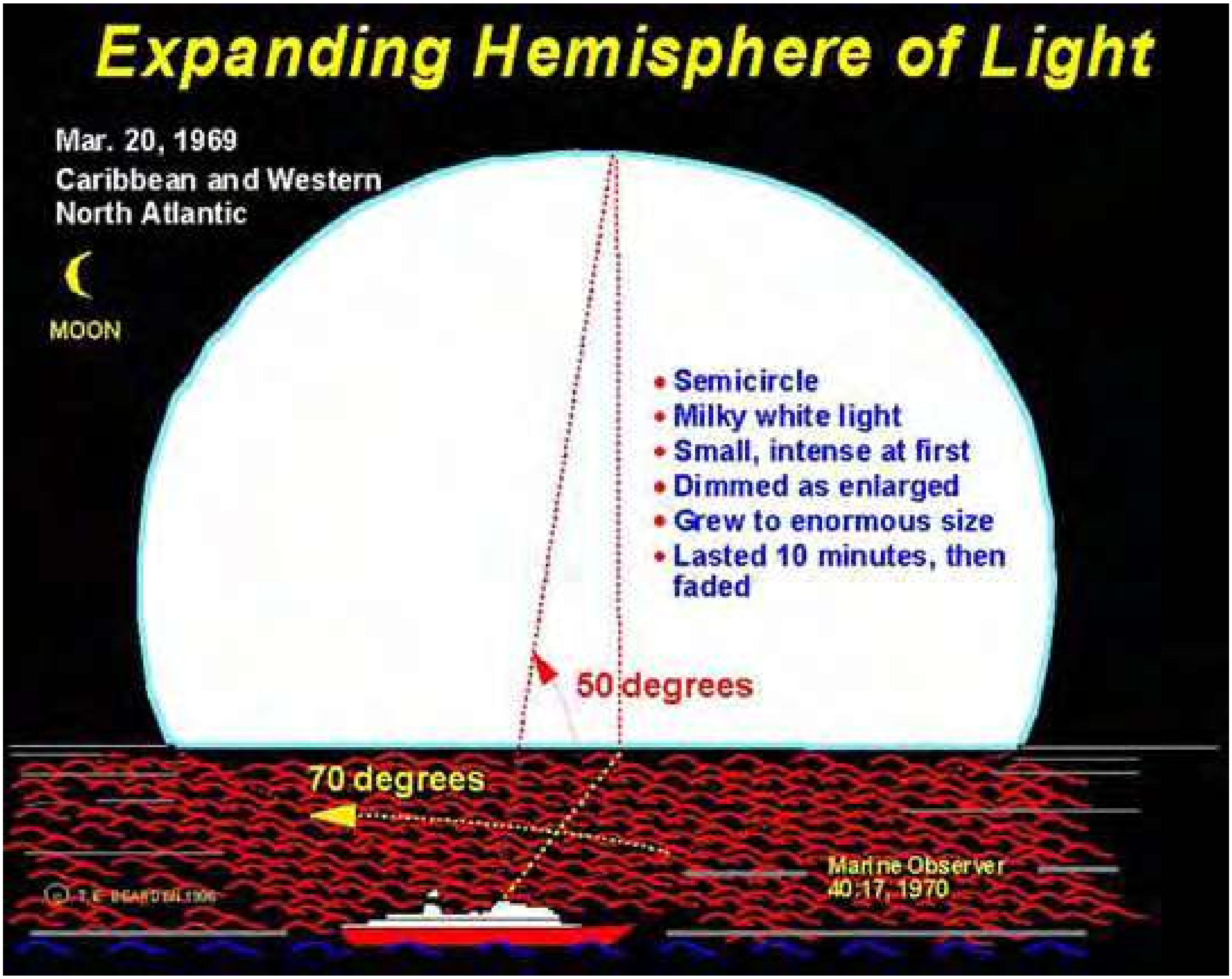

Expanding Hemisphere of Light

Mar. 20, 1969
Caribbean and Western
North Atlantic

- Semicircle
- Milky white light
- Small, intense at first
- Dimmed as enlarged
- Grew to enormous size
- Lasted 10 minutes, then faded

MOON
50 degrees
70 degrees
Marine Observer
40:17, 1970

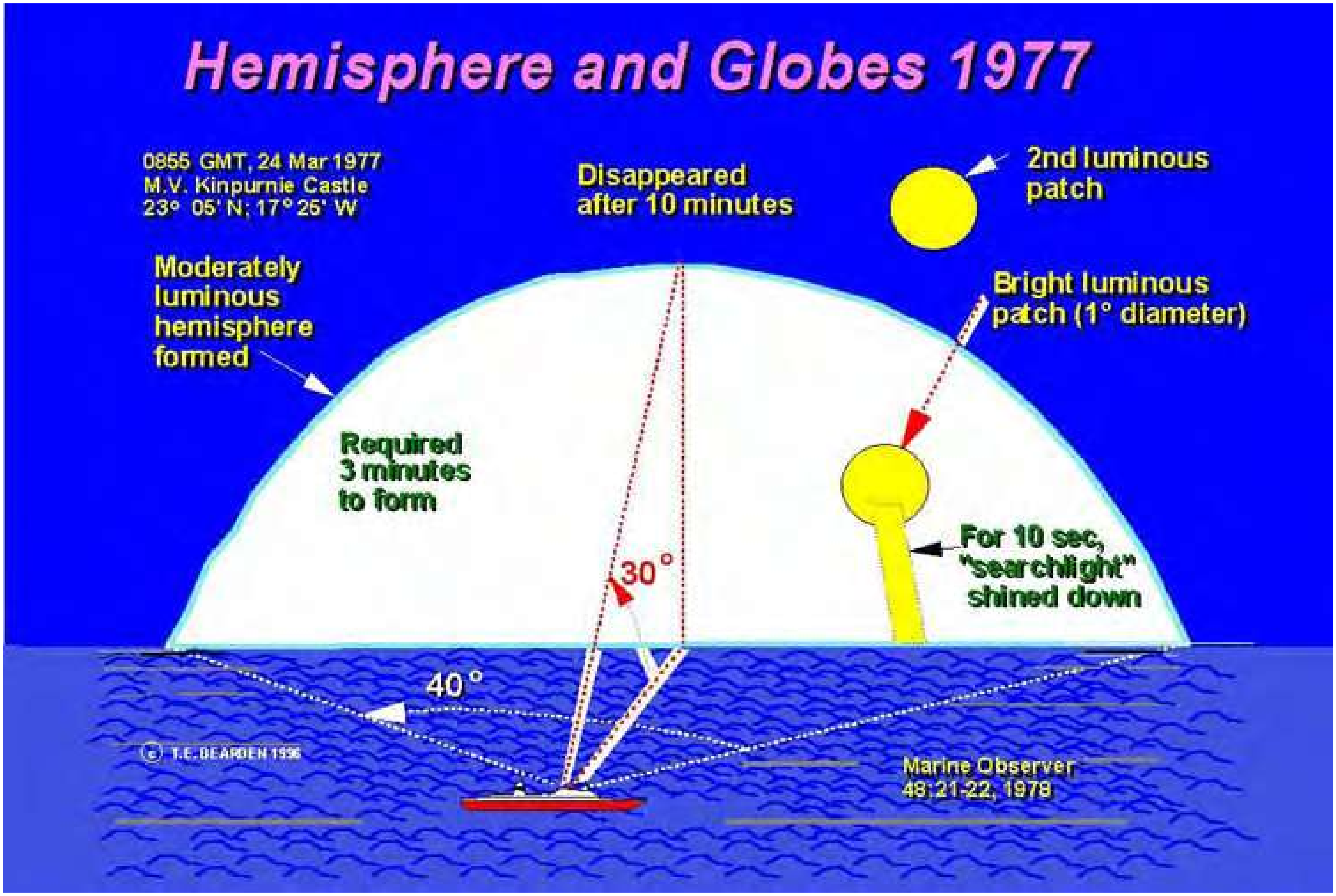

Hemisphere and Globes 1977

0855 GMT, 24 Mar 1977
M.V. Kinpurnie Castle
23° 05' N; 17° 25' W

Disappeared after 10 minutes

2nd luminous patch

Moderately luminous hemisphere formed

Bright luminous patch (1° diameter)

Required 3 minutes to form

For 10 sec, "searchlight" shined down

© T.E. BEARDEN 1996

Marine Observer
48:21-22, 1978

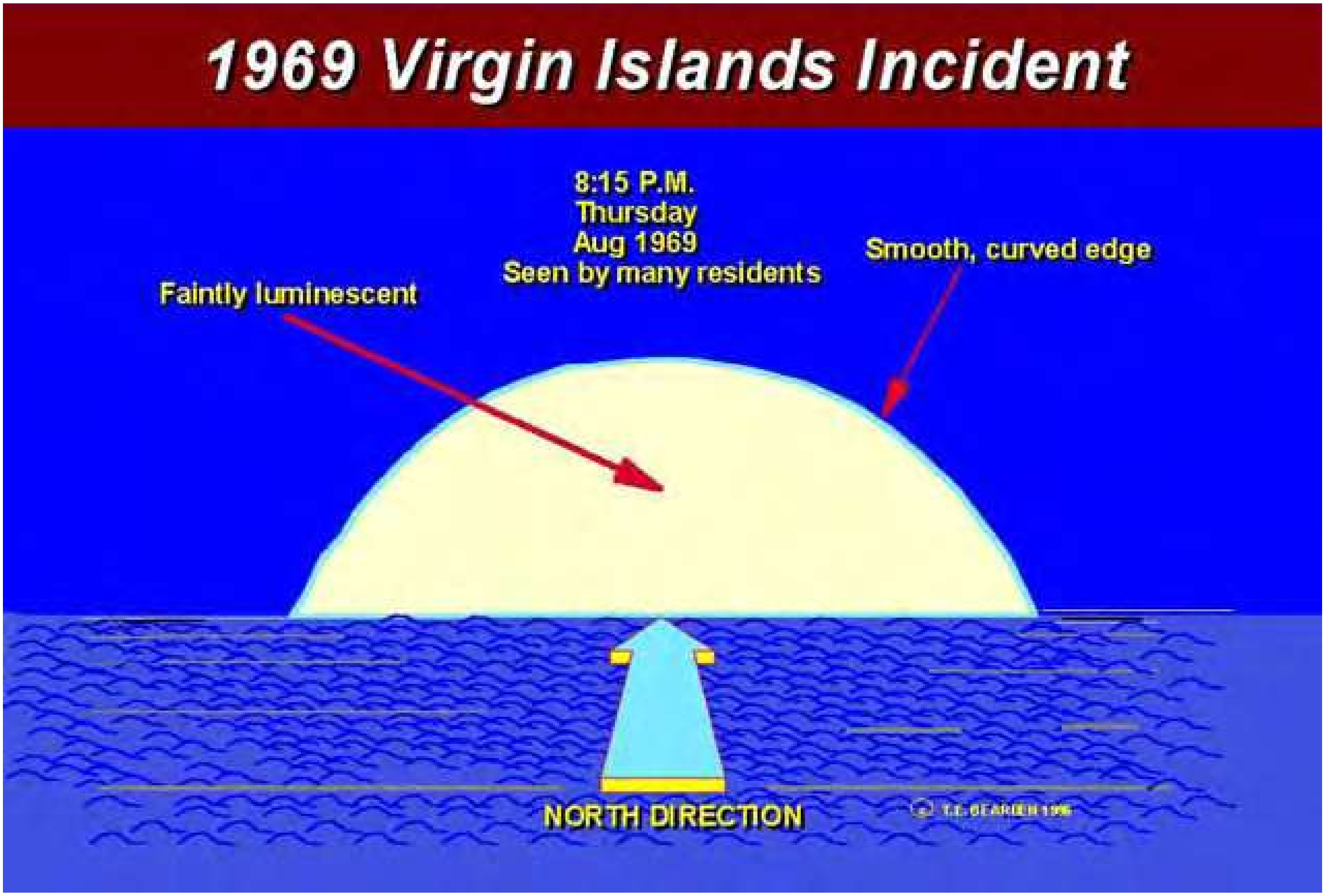

1969 Virgin Islands Incident

8:15 P.M.
Thursday
Aug 1969
Seen by many residents

Faintly luminescent

Smooth, curved edge

NORTH DIRECTION

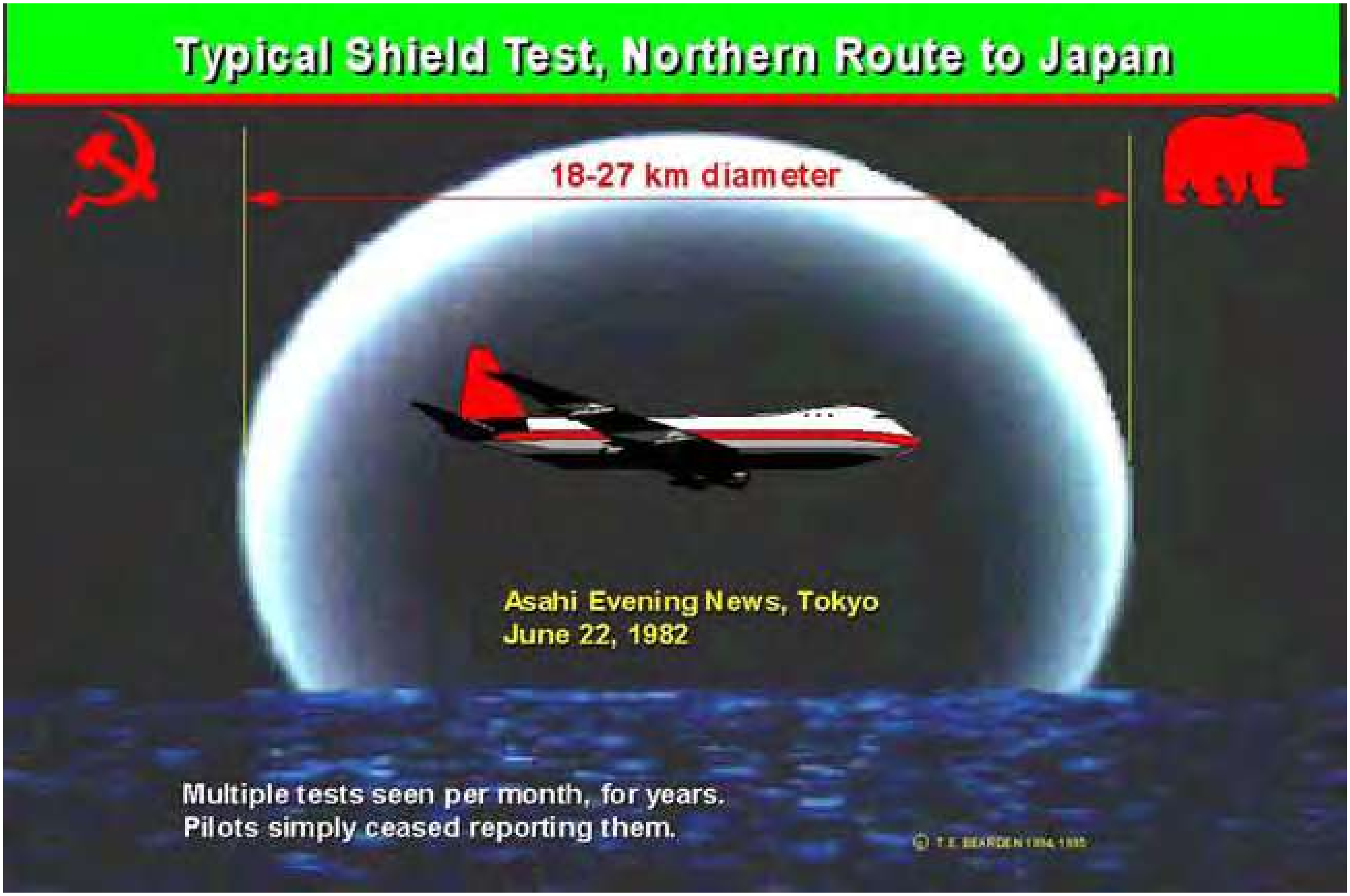

Typical Shield Test, Northern Route to Japan

18-27 km diameter

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T.E. BEARDEN 1994, 1995

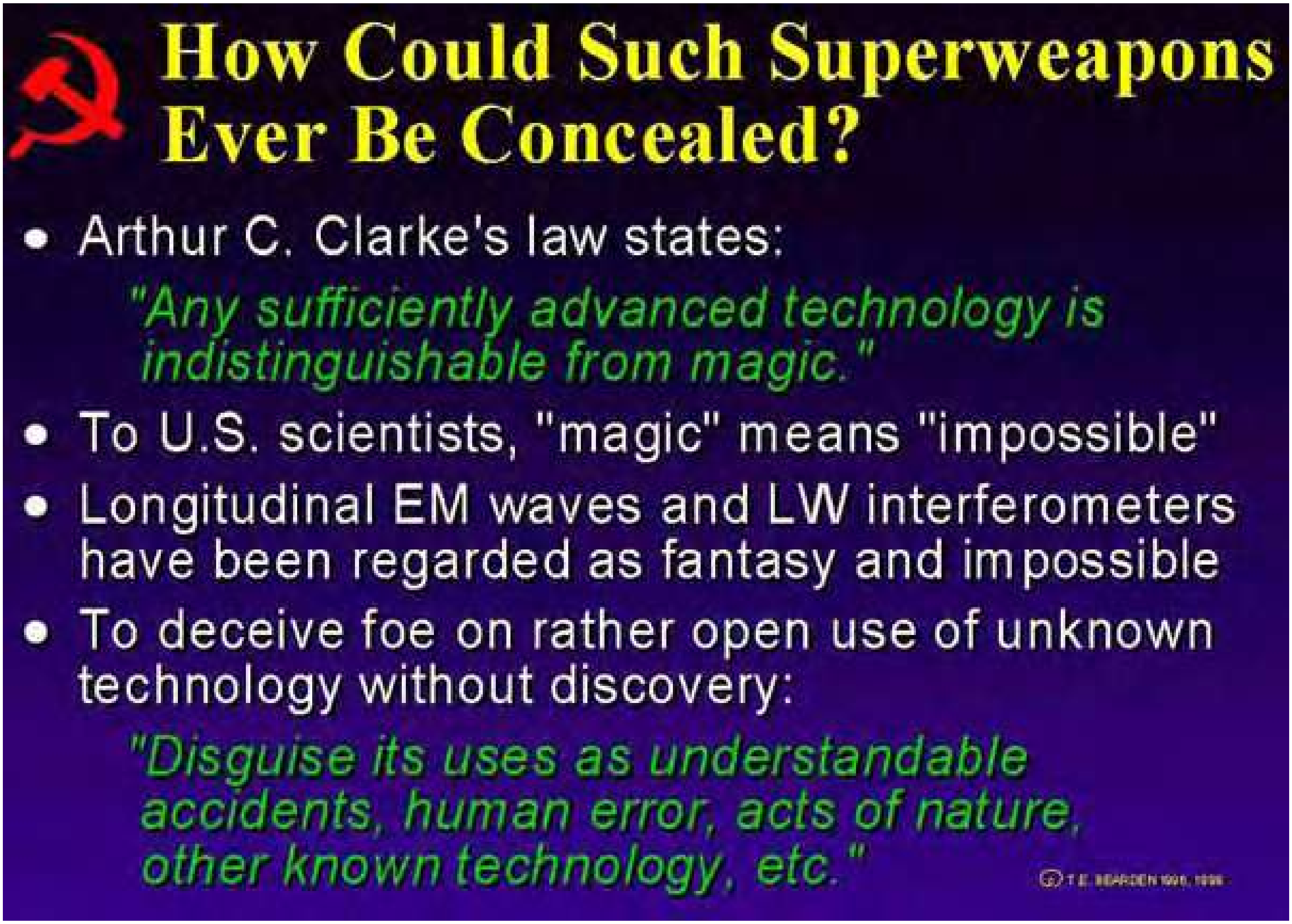

How Could Such Superweapons Ever Be Concealed?

- Arthur C. Clarke's law states:

  "Any sufficiently advanced technology is indistinguishable from magic."
- To U.S. scientists, "magic" means "impossible"
- Longitudinal EM waves and LW interferometers have been regarded as fantasy and impossible
- To deceive foe on rather open use of unknown technology without discovery:

  "Disguise its uses as understandable accidents, human error, acts of nature, other known technology, etc."

© T.E. BEARDEN [illegible]

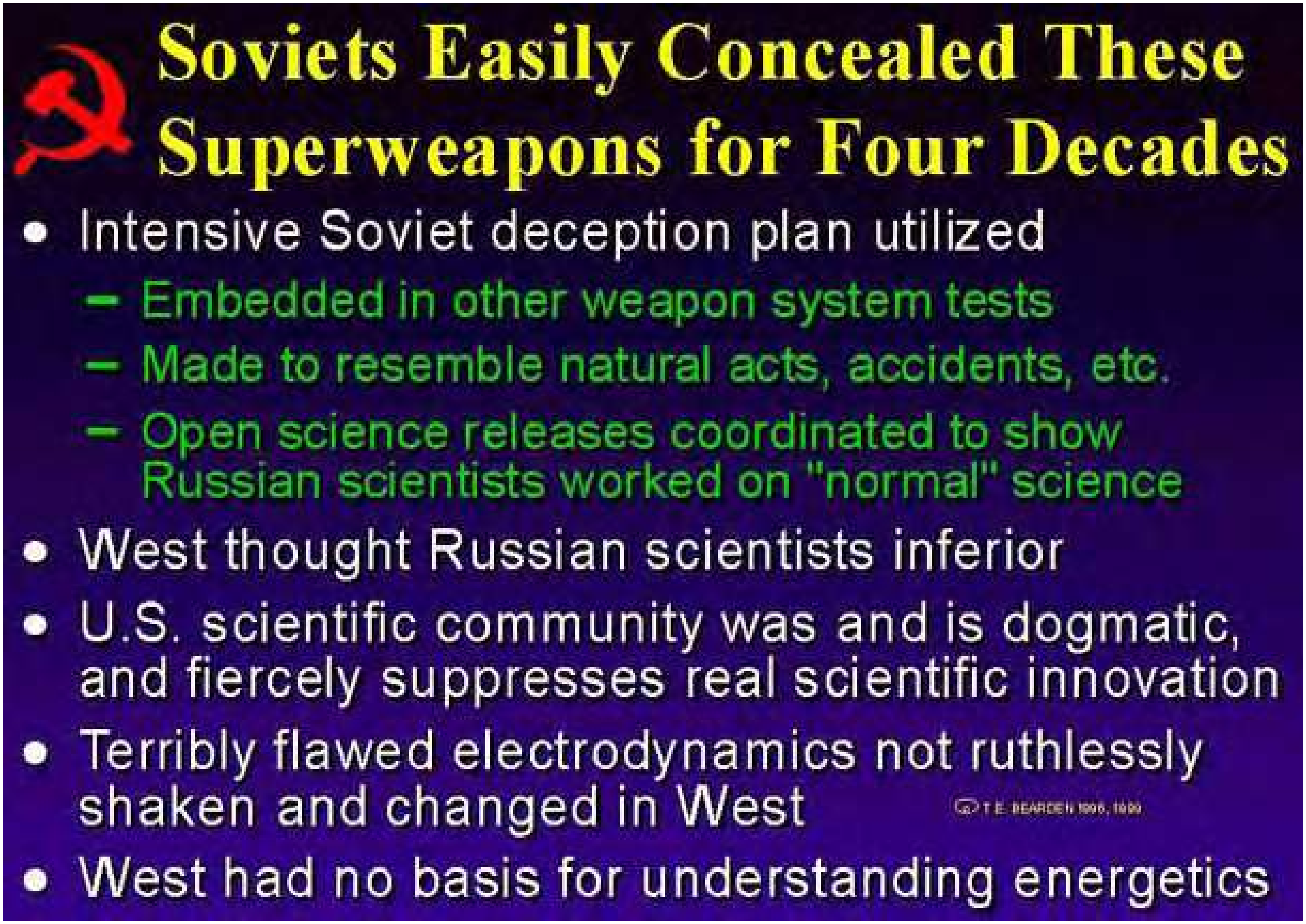

Soviets Easily Concealed These Superweapons for Four Decades

- Intensive Soviet deception plan utilized
  - Embedded in other weapon system tests
  - Made to resemble natural acts, accidents, etc.
  - Open science releases coordinated to show Russian scientists worked on "normal" science
- West thought Russian scientists inferior
- U.S. scientific community was and is dogmatic, and fiercely suppresses real scientific innovation
- Terribly flawed electrodynamics not ruthlessly shaken and changed in West
- West had no basis for understanding energetics

© T.E. BEARDEN 1996, 1999

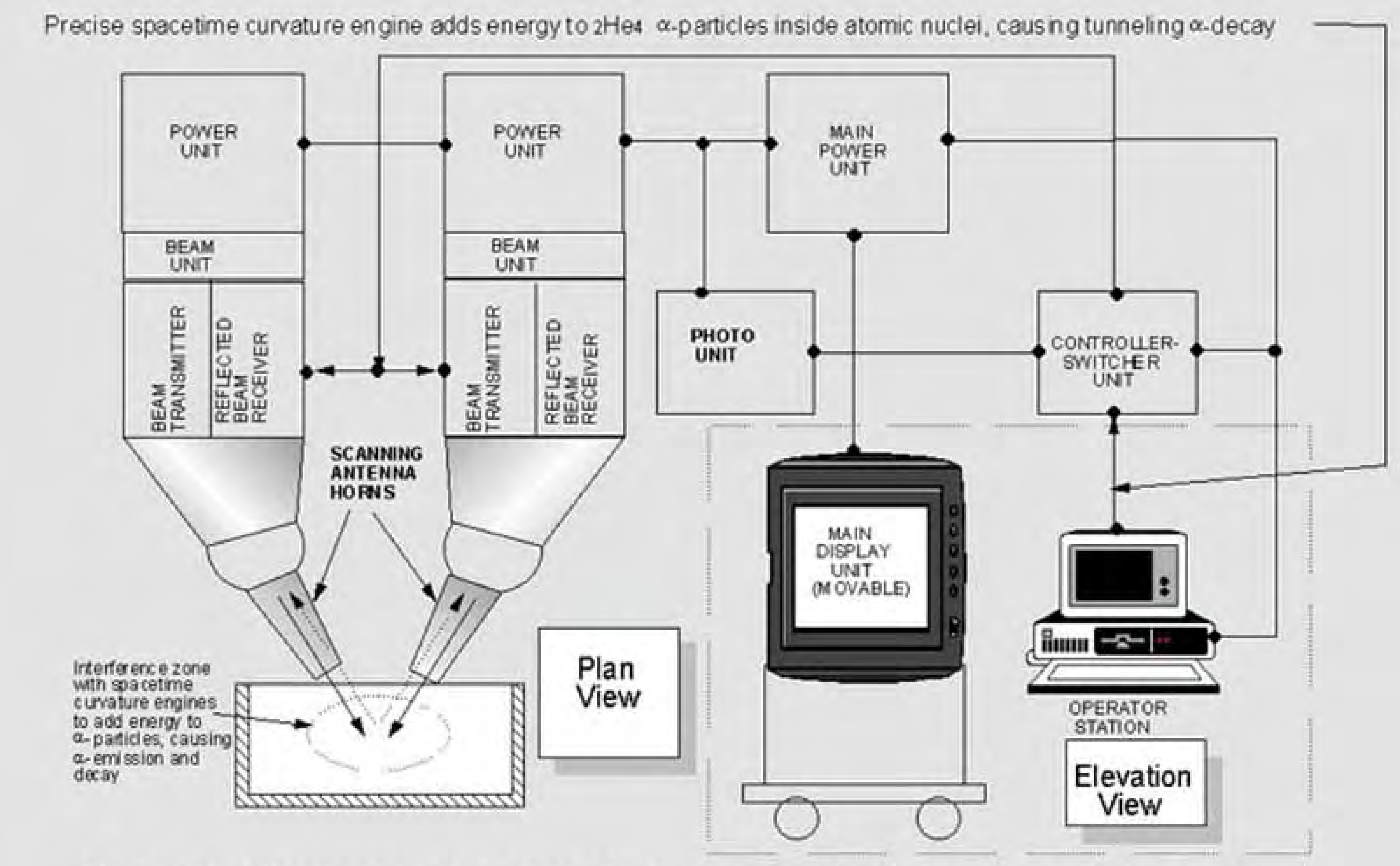
Precise spacetime curvature engine adds energy to 2He4 α-particles inside atomic nuclei, causing tunneling α-decay
POWER UNIT
POWER UNIT
MAIN POWER UNIT
BEAM UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
PHOTO UNIT
CONTROLLER-SWITCHER UNIT
SCANNING ANTENNA HORNS
MAIN DISPLAY UNIT (MOVABLE)
Interference zone with spacetime curvature engines to add energy to α-particles, causing α-emission and decay
Plan View
OPERATOR STATION
Elevation View

Figure 26. Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes.

© T.E. BEARDEN 1996

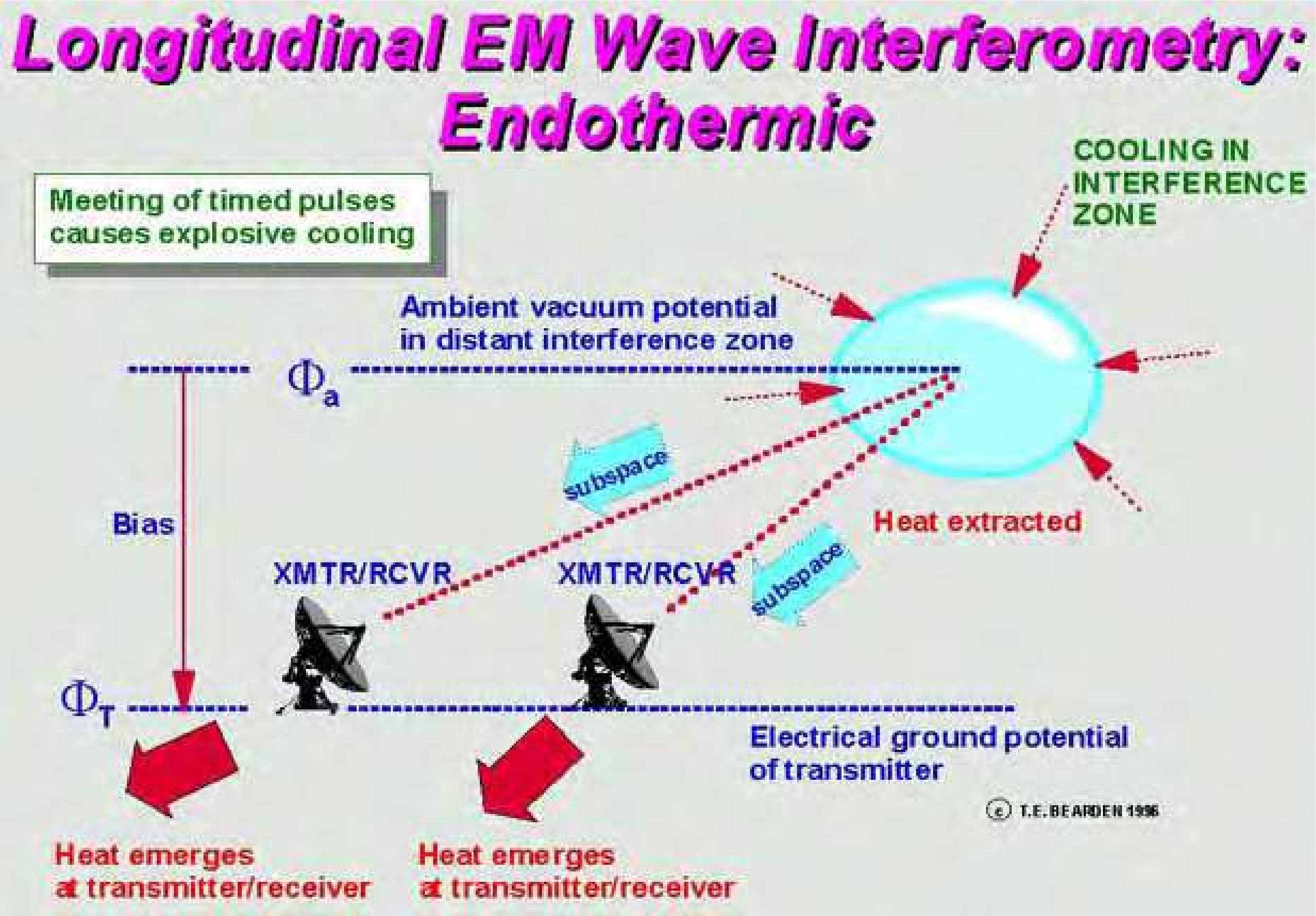

Longitudinal EM Wave Interferometry: Endothermic

Meeting of timed pulses
causes explosive cooling

COOLING IN
INTERFERENCE
ZONE

Ambient vacuum potential
in distant interference zone

Φ_a

subspace

subspace

Bias

Heat extracted

XMTR/RCVR

XMTR/RCVR

Φ_T

Electrical ground potential
of transmitter

© T.E. BEARDEN 1996

Heat emerges
at transmitter/receiver

Heat emerges
at transmitter/receiver

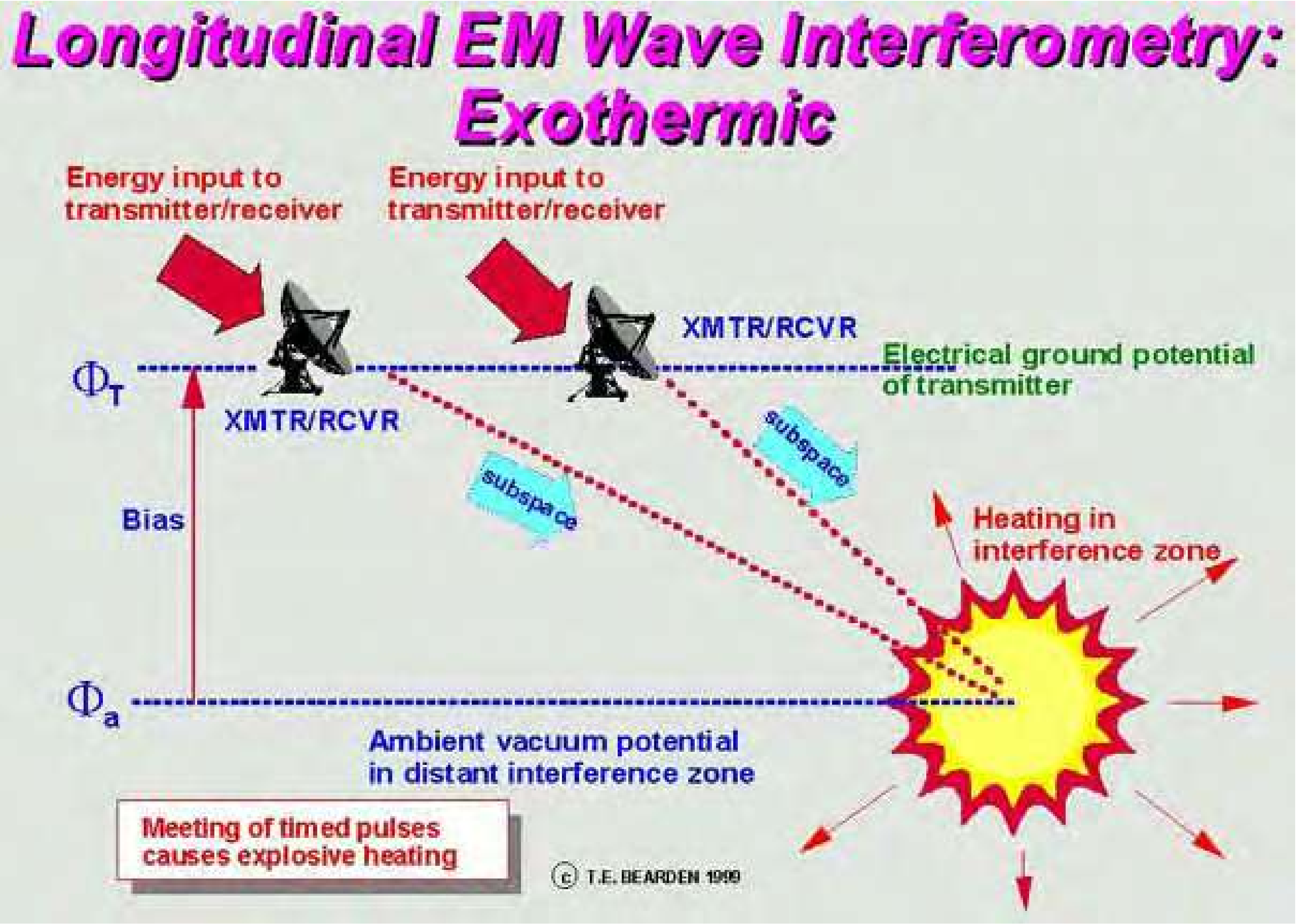

Longitudinal EM Wave Interferometry: Exothermic

Energy input to transmitter/receiver

Energy input to transmitter/receiver

XMTR/RCVR

XMTR/RCVR

Φ_T

Electrical ground potential of transmitter

subspace

subspace

Bias

Heating in interference zone

Φ_a

Ambient vacuum potential in distant interference zone

Meeting of timed pulses causes explosive heating

Ⓒ T.E. BEARDEN 1999

The Tom Bearden Website

Help support the research

U.S. Air Attack on Libya
April 1986

- Slide 1
- Slide 2
- "Washington Post" Article
  January 22, 1989

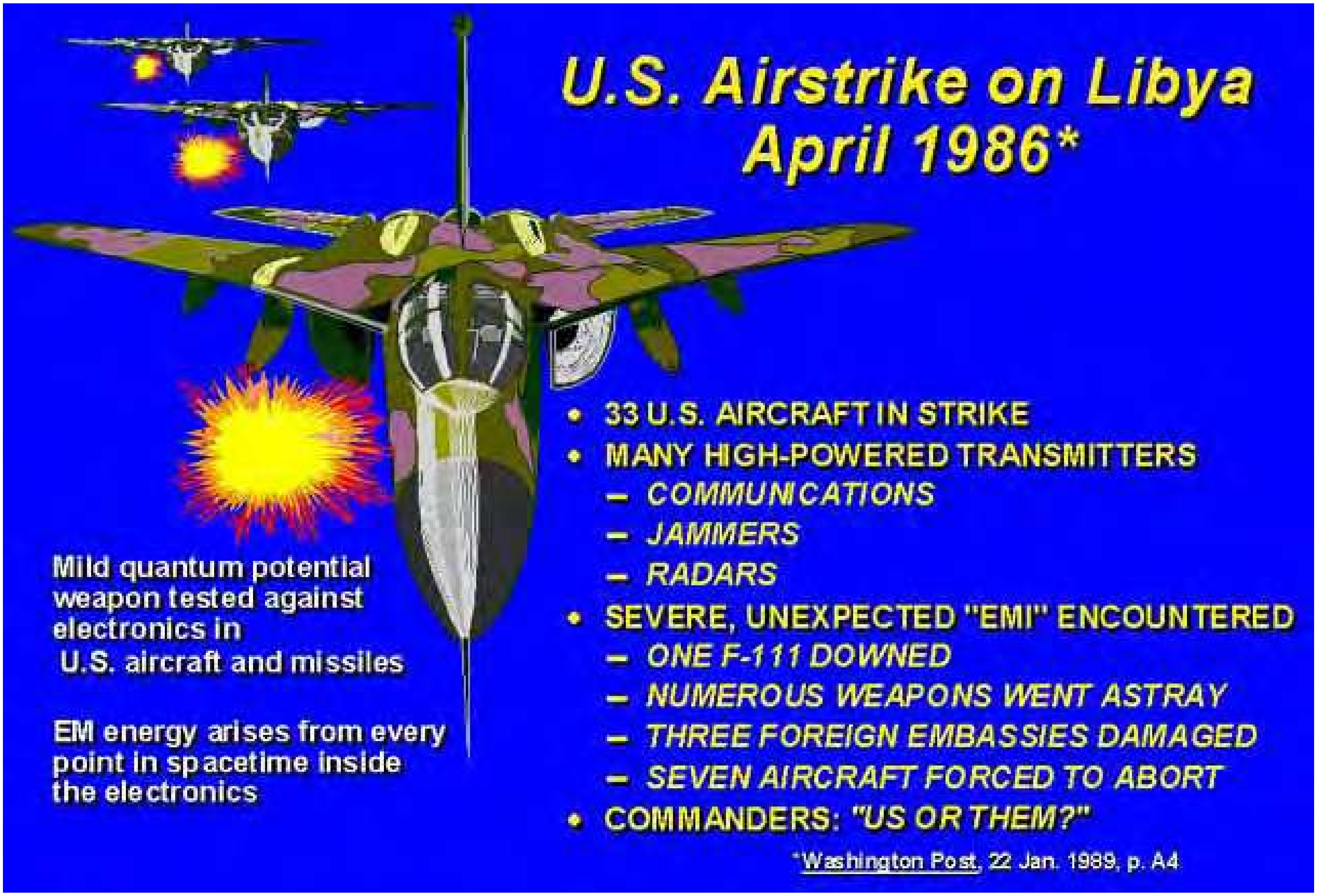

U.S. Airstrike on Libya April 1986*

- 33 U.S. AIRCRAFT IN STRIKE
- MANY HIGH-POWERED TRANSMITTERS
  - *COMMUNICATIONS*
  - *JAMMERS*
  - *RADARS*
- SEVERE, UNEXPECTED "EMI" ENCOUNTERED
  - *ONE F-111 DOWNED*
  - *NUMEROUS WEAPONS WENT ASTRAY*
  - *THREE FOREIGN EMBASSIES DAMAGED*
  - *SEVEN AIRCRAFT FORCED TO ABORT*
- COMMANDERS: *"US OR THEM?"*

Mild quantum potential weapon tested against electronics in U.S. aircraft and missiles

EM energy arises from every point in spacetime inside the electronics

*Washington Post, 22 Jan. 1989, p. A4

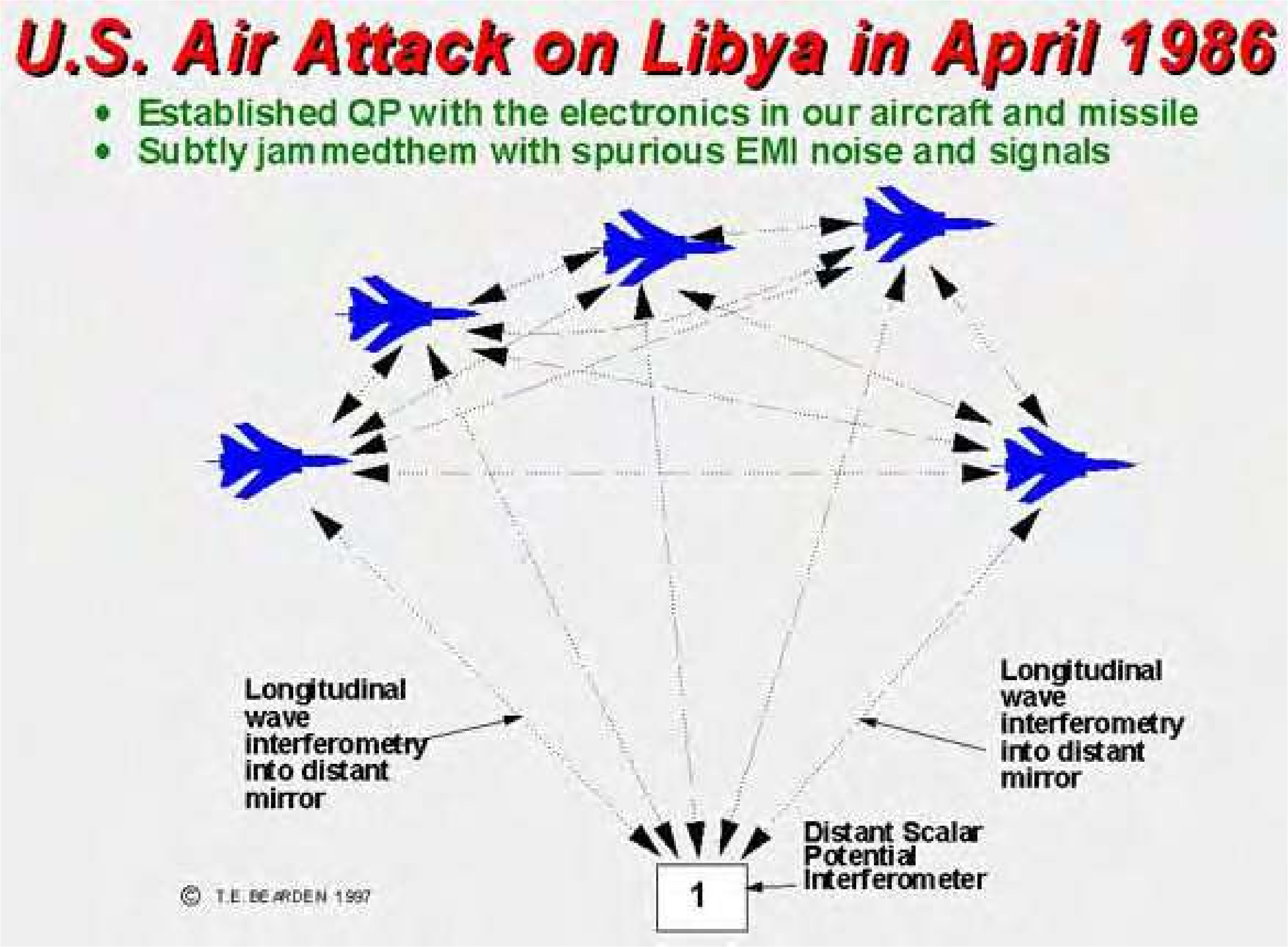

U.S. Air Attack on Libya in April 1986

- Established QP with the electronics in our aircraft and missile
- Subtly jammedthem with spurious EMI noise and signals

Longitudinal wave interferometry into distant mirror
Longitudinal wave interferometry into distant mirror
Distant Scalar Potential Interferometer
1
© T.E. BEARDEN 1997

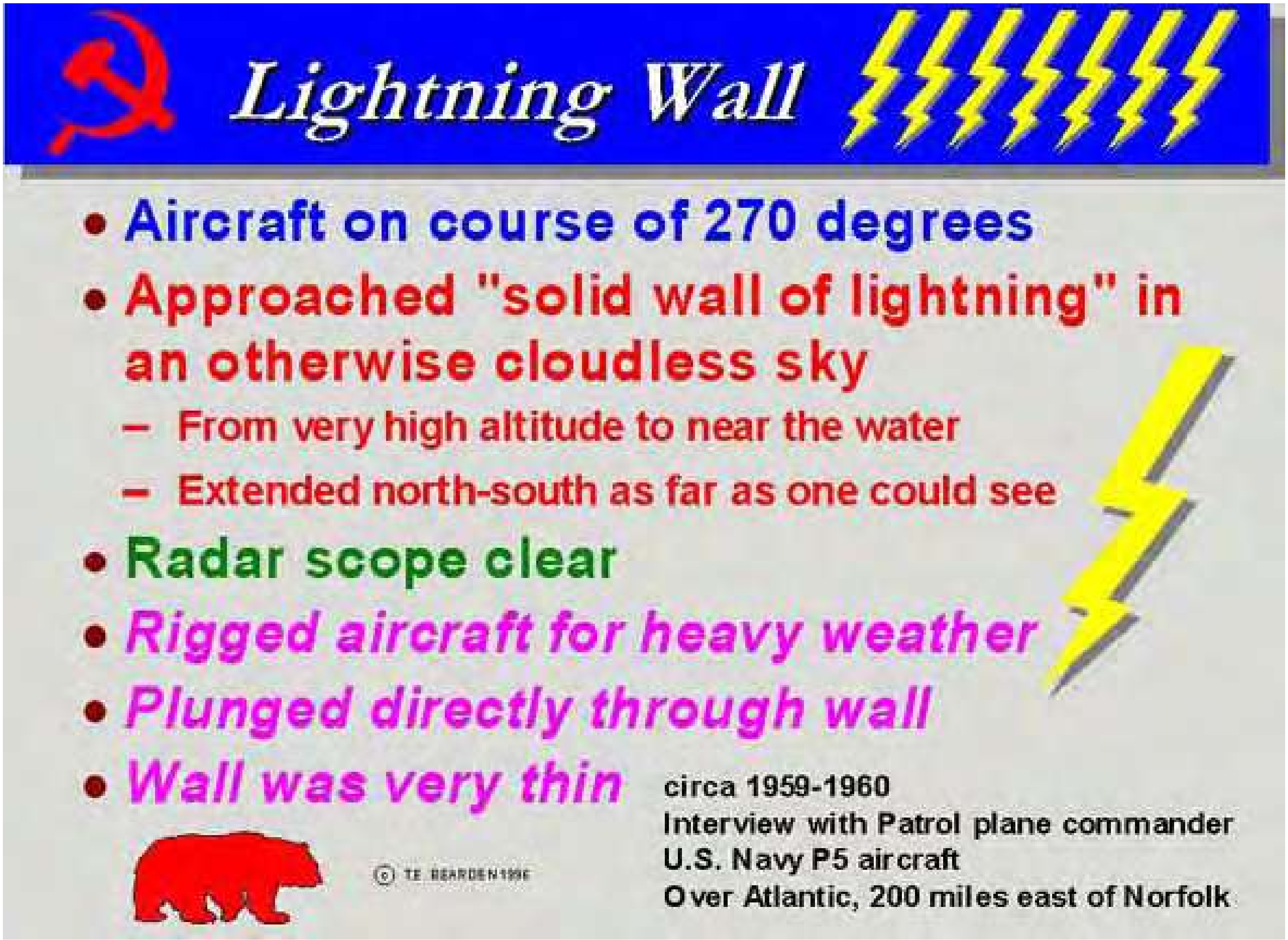

Lightning Wall

- Aircraft on course of 270 degrees
- Approached "solid wall of lightning" in an otherwise cloudless sky
  - From very high altitude to near the water
  - Extended north-south as far as one could see
- Radar scope clear
- *Rigged aircraft for heavy weather*
- *Plunged directly through wall*
- *Wall was very thin*

circa 1959-1960
Interview with Patrol plane commander
U.S. Navy P5 aircraft
Over Atlantic, 200 miles east of Norfolk

© T.E. BEARDEN 1996

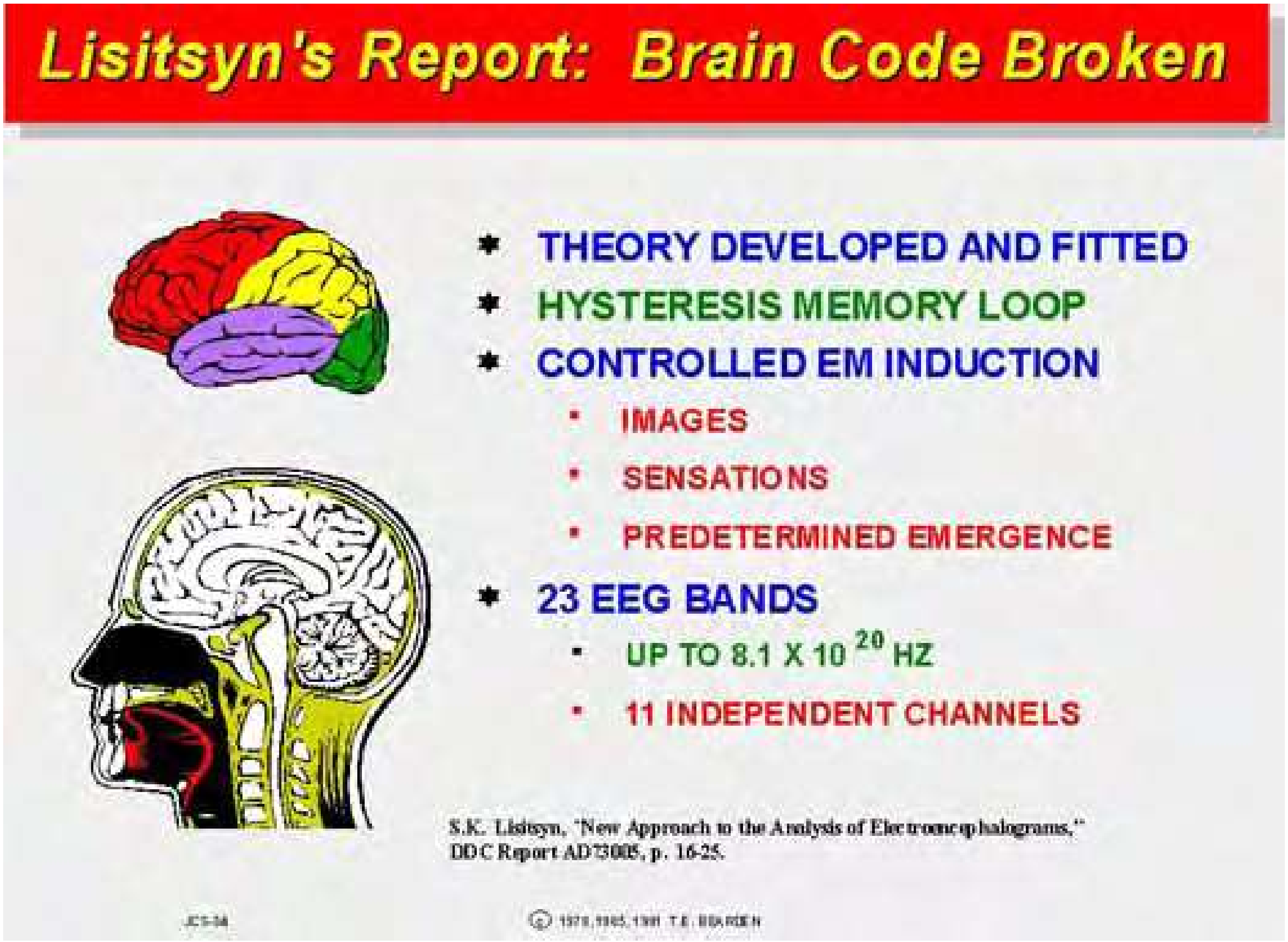

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms,"
DDC Report AD73005, p. 16-25.

JCS-34

© 1978, 1985, 1991 T.E. BEARDEN

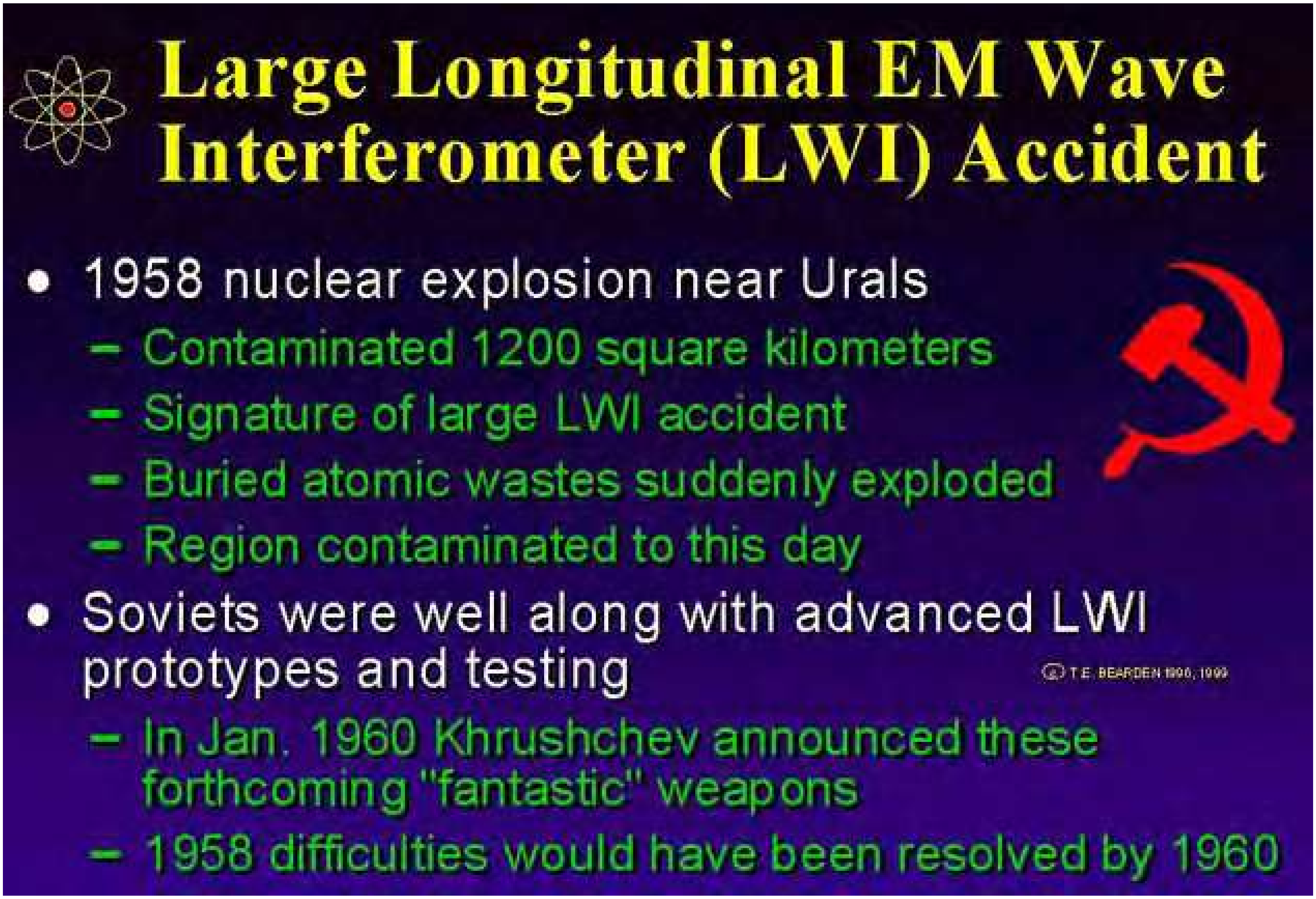

Large Longitudinal EM Wave Interferometer (LWI) Accident

- 1958 nuclear explosion near Urals
  - Contaminated 1200 square kilometers
  - Signature of large LWI accident
  - Buried atomic wastes suddenly exploded
  - Region contaminated to this day
- Soviets were well along with advanced LWI prototypes and testing
  - In Jan. 1960 Khrushchev announced these forthcoming "fantastic" weapons
  - 1958 difficulties would have been resolved by 1960

© T.E. BEARDEN 1996, 1999

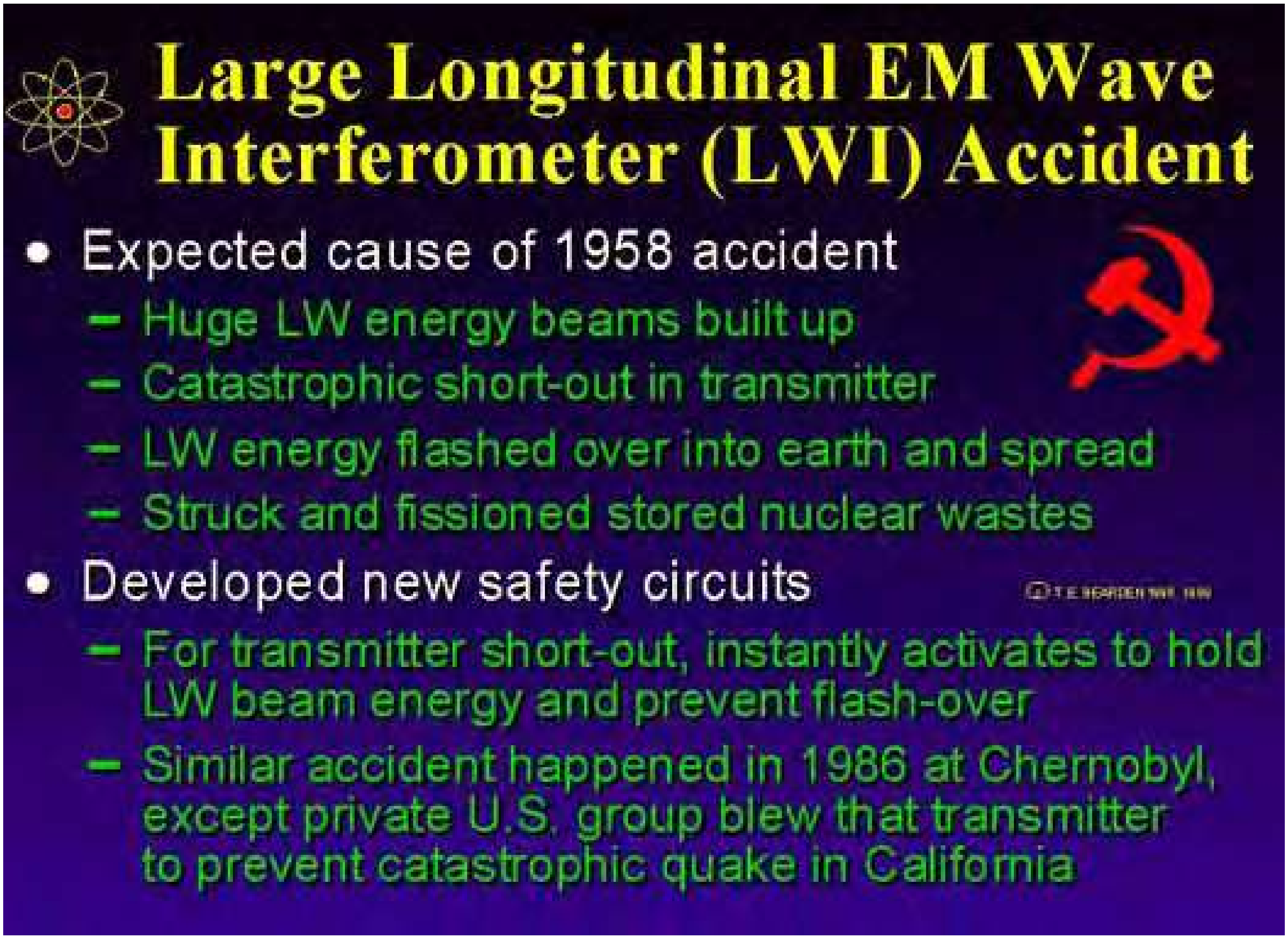

Large Longitudinal EM Wave Interferometer (LWI) Accident

- Expected cause of 1958 accident
  - Huge LW energy beams built up
  - Catastrophic short-out in transmitter
  - LW energy flashed over into earth and spread
  - Struck and fissioned stored nuclear wastes
- Developed new safety circuits
  - For transmitter short-out, instantly activates to hold LW beam energy and prevent flash-over
  - Similar accident happened in 1986 at Chernobyl, except private U.S. group blew that transmitter to prevent catastrophic quake in California

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

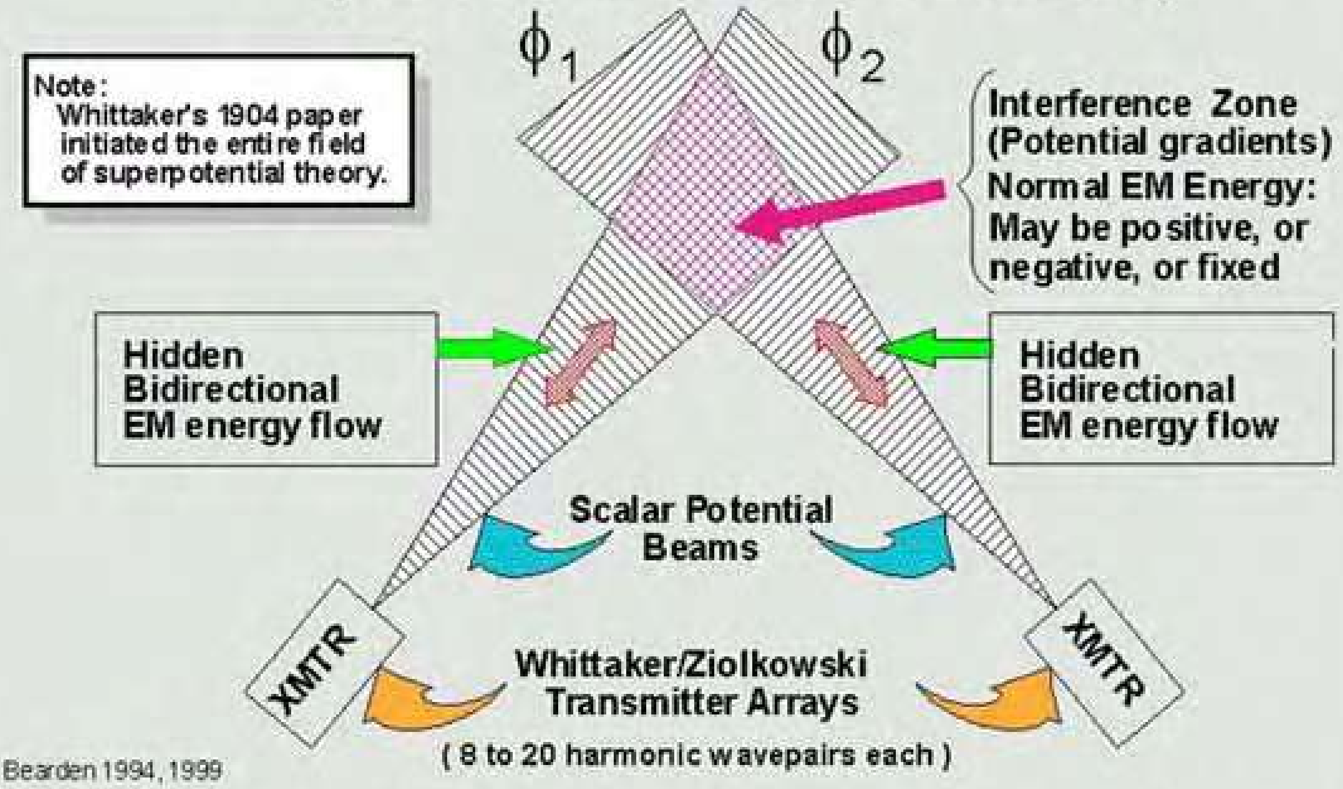
ϕ_1
ϕ_2
Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

© T. E. Bearden 1994, 1999

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepair functions) produces all EM force fields and waves.

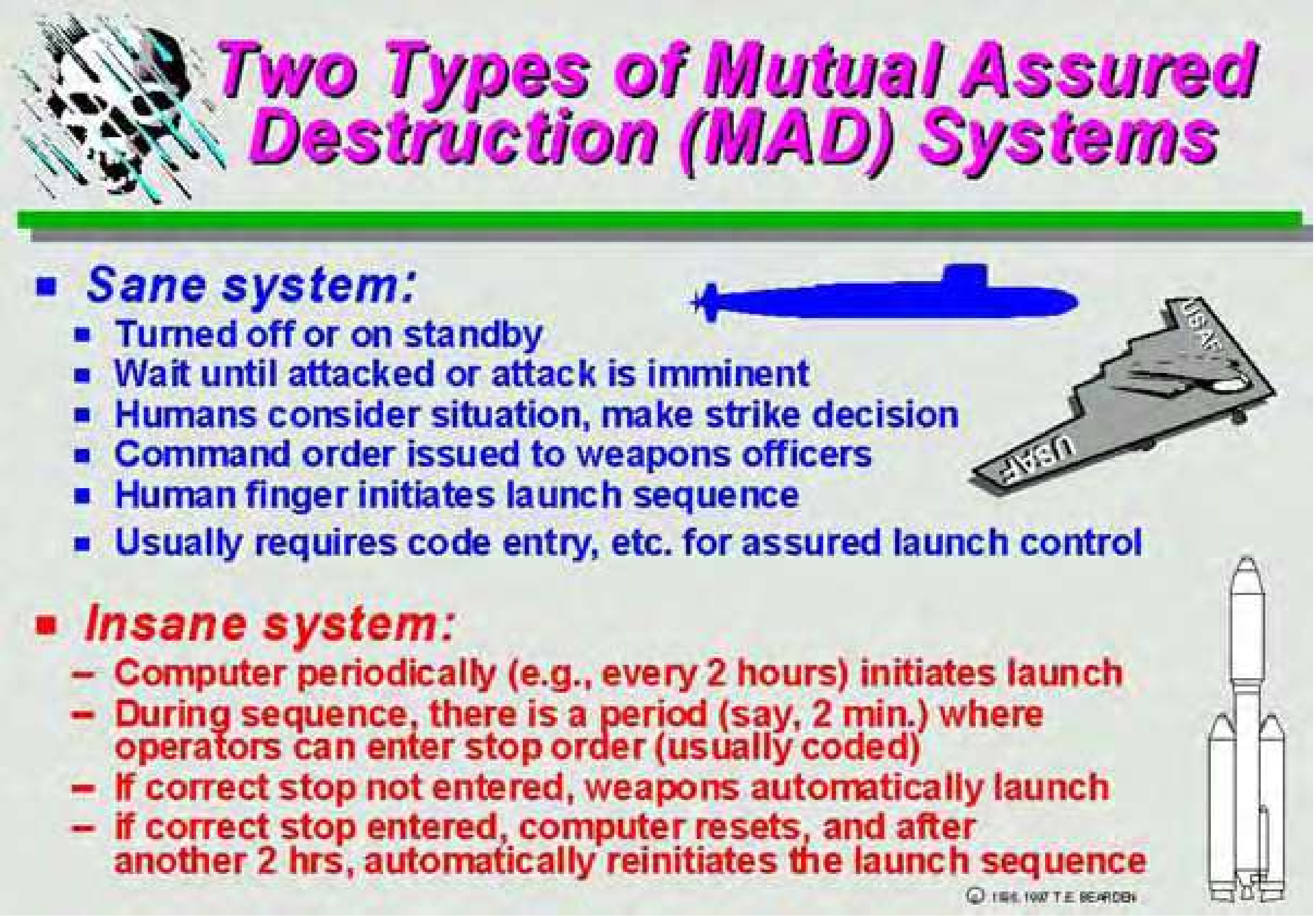

Two Types of Mutual Assured Destruction (MAD) Systems

- ***Sane system:***
  - Turned off or on standby
  - Wait until attacked or attack is imminent
  - Humans consider situation, make strike decision
  - Command order issued to weapons officers
  - Human finger initiates launch sequence
  - Usually requires code entry, etc. for assured launch control
- ***Insane system:***
  - Computer periodically (e.g., every 2 hours) initiates launch
  - During sequence, there is a period (say, 2 min.) where operators can enter stop order (usually coded)
  - If correct stop not entered, weapons automatically launch
  - If correct stop entered, computer resets, and after another 2 hrs, automatically reinitiates the launch sequence

© 1996, 1997 T.E. BEARDEN

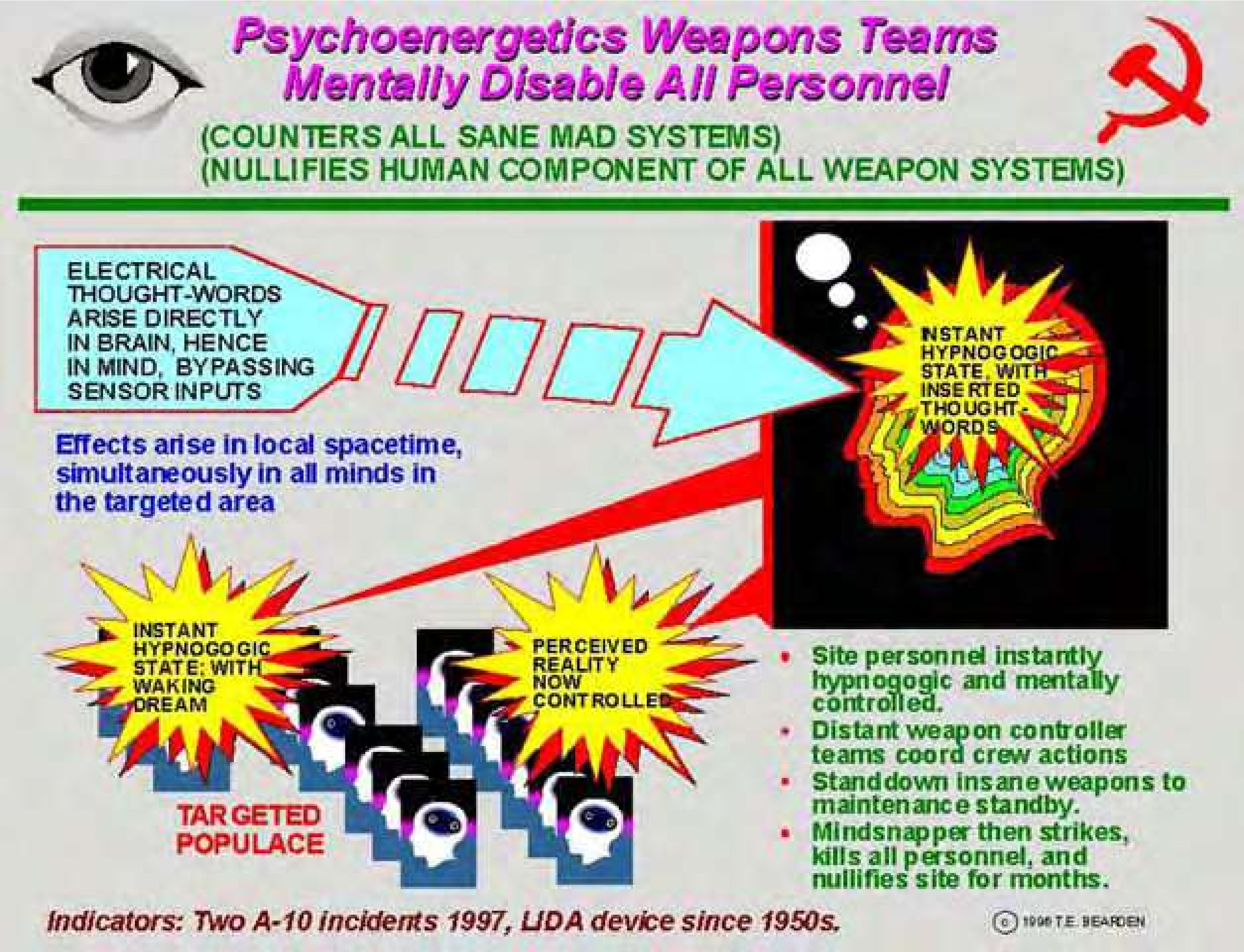

Psychoenergetics Weapons Teams Mentally Disable All Personnel

(COUNTERS ALL SANE MAD SYSTEMS)
(NULLIFIES HUMAN COMPONENT OF ALL WEAPON SYSTEMS)

ELECTRICAL THOUGHT-WORDS ARISE DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED THOUGHT-WORDS

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED POPULACE

- Site personnel instantly hypnogogic and mentally controlled.
- Distant weapon controller teams coord crew actions
- Standdown insane weapons to maintenance standby.
- Mindsnapper then strikes, kills all personnel, and nullifies site for months.

Indicators: Two A-10 incidents 1997, LIDA device since 1950s.

© 1998 T.E. BEARDEN

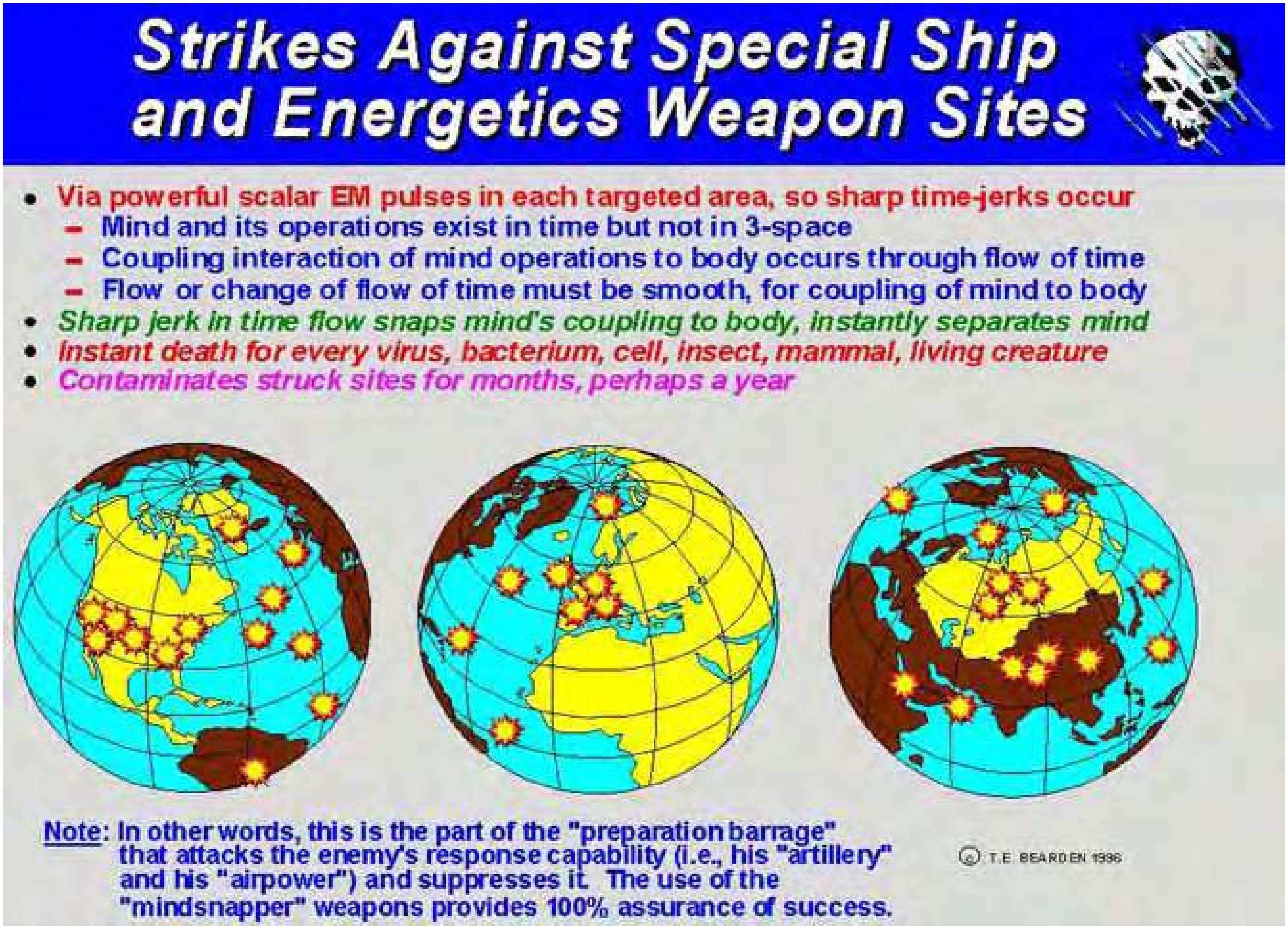

Strikes Against Special Ship and Energetics Weapon Sites

- Via powerful scalar EM pulses in each targeted area, so sharp time-jerks occur
  - Mind and its operations exist in time but not in 3-space
  - Coupling interaction of mind operations to body occurs through flow of time
  - Flow or change of flow of time must be smooth, for coupling of mind to body
- *Sharp jerk in time flow snaps mind's coupling to body, instantly separates mind*
- *Instant death for every virus, bacterium, cell, insect, mammal, living creature*
- *Contaminates struck sites for months, perhaps a year*

Note: In other words, this is the part of the "preparation barrage" that attacks the enemy's response capability (i.e., his "artillery" and his "airpower") and suppresses it. The use of the "mindsnapper" weapons provides 100% assurance of success.

© T.E. BEARDEN 1996

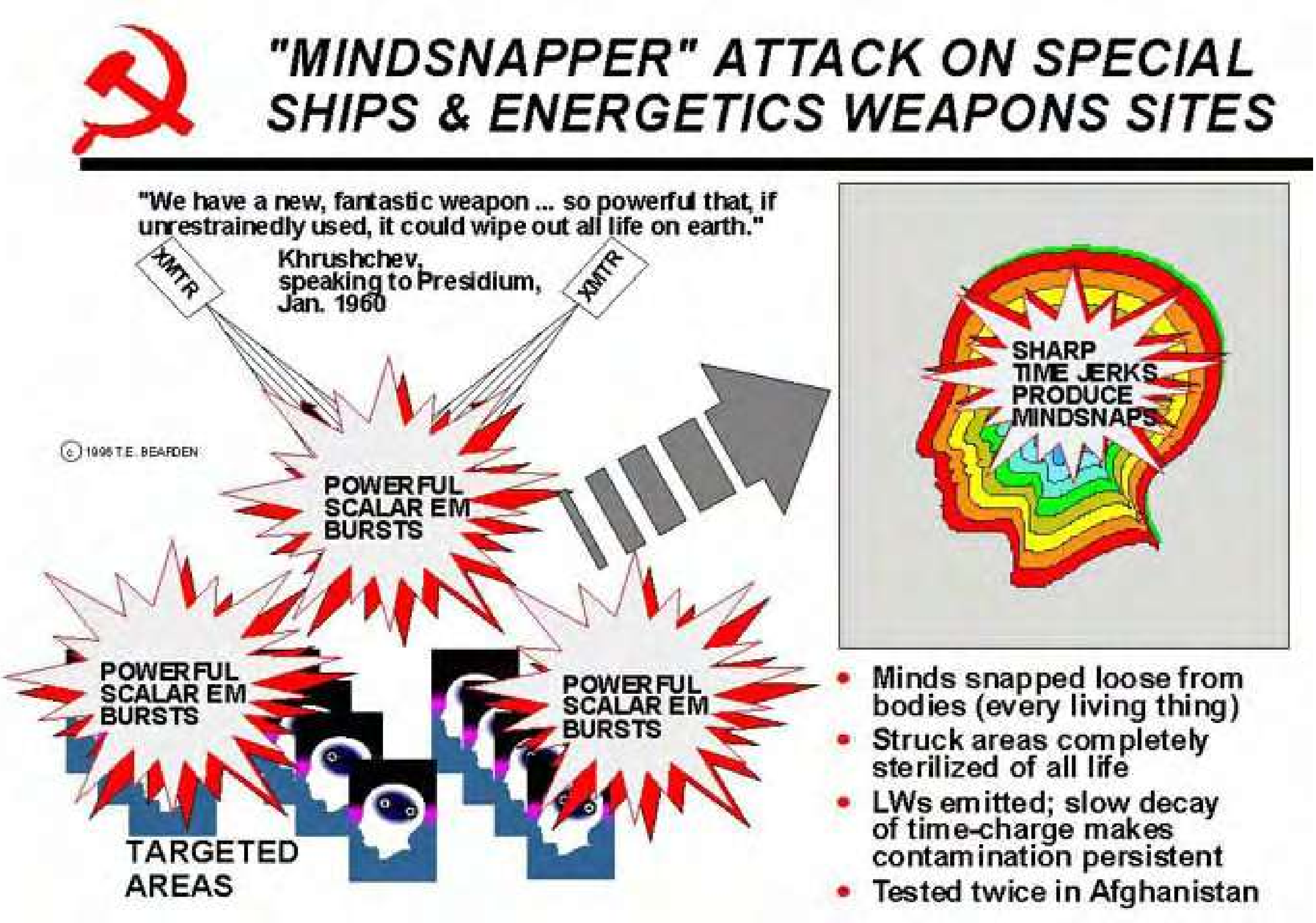

"MINDSNAPPER" ATTACK ON SPECIAL SHIPS & ENERGETICS WEAPONS SITES

"We have a new, fantastic weapon ... so powerful that, if unrestrainedly used, it could wipe out all life on earth."

Khrushchev,
speaking to Presidium,
Jan. 1960

XMTR
XMTR
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
TARGETED AREAS
SHARP TIME JERKS PRODUCE MINDSNAPS

(c) 1996 T.E. BEARDEN

- Minds snapped loose from bodies (every living thing)
- Struck areas completely sterilized of all life
- LWs emitted; slow decay of time-charge makes contamination persistent
- Tested twice in Afghanistan

The Tom Bearden Website

Help support the research

How Much Off Guard Are We?

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6
- Slide 7

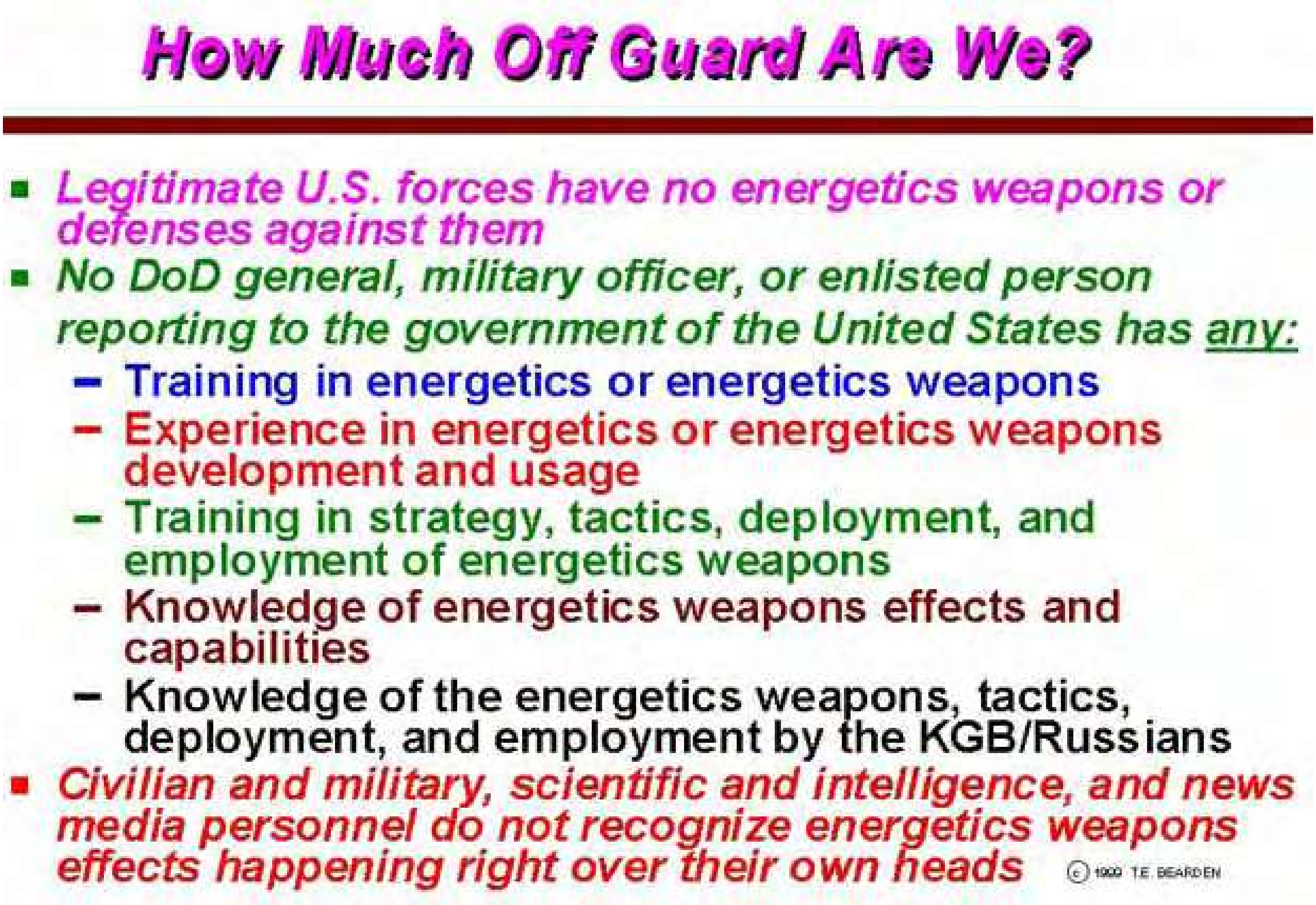

How Much Off Guard Are We?

- *Legitimate U.S. forces have no energetics weapons or defenses against them*
- *No DoD general, military officer, or enlisted person reporting to the government of the United States has any:*
  - Training in energetics or energetics weapons
  - Experience in energetics or energetics weapons development and usage
  - Training in strategy, tactics, deployment, and employment of energetics weapons
  - Knowledge of energetics weapons effects and capabilities
  - Knowledge of the energetics weapons, tactics, deployment, and employment by the KGB/Russians
- *Civilian and military, scientific and intelligence, and news media personnel do not recognize energetics weapons effects happening right over their own heads*

© 1999 T.E. BEARDEN

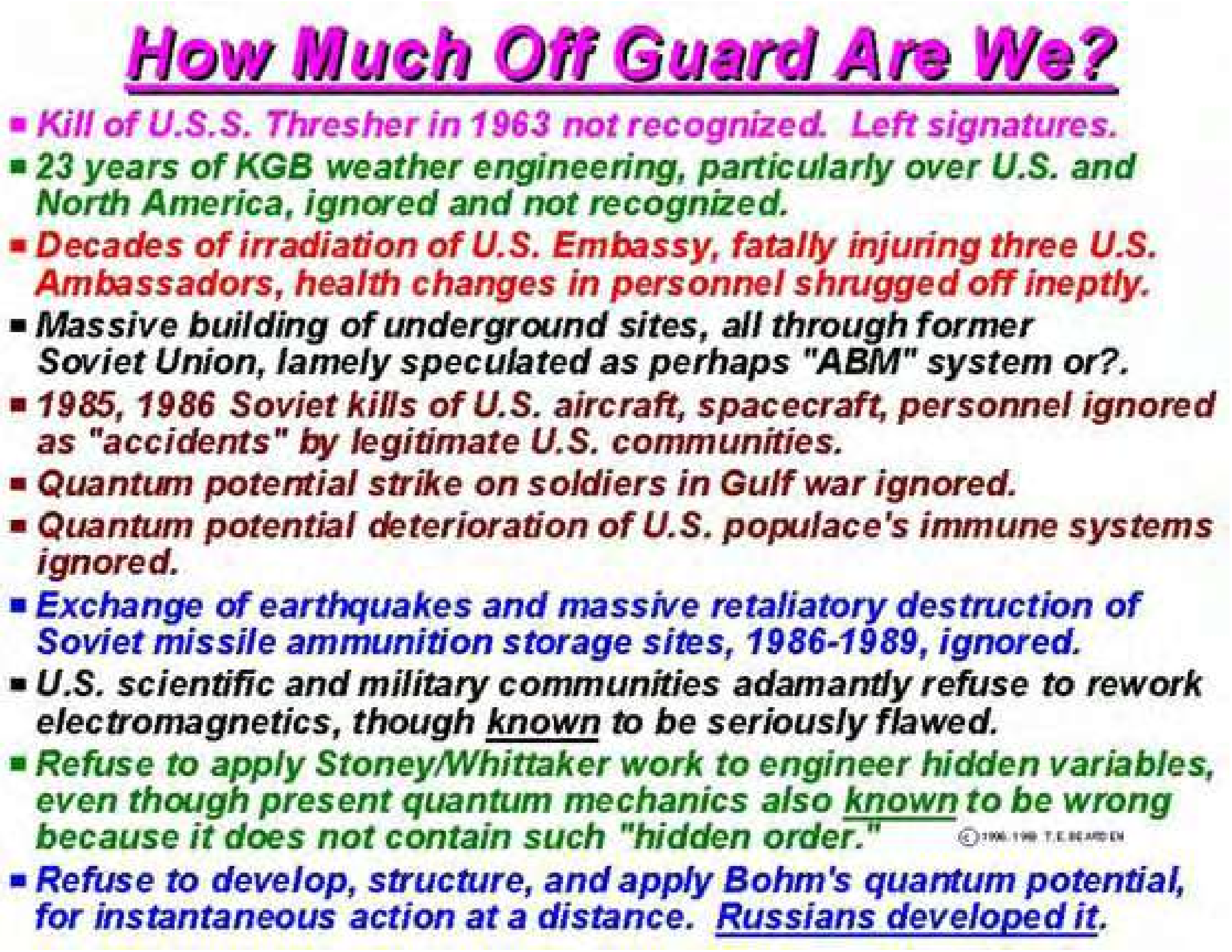

How Much Off Guard Are We?

- *Kill of U.S.S. Thresher in 1963 not recognized. Left signatures.*
- *23 years of KGB weather engineering, particularly over U.S. and North America, ignored and not recognized.*
- *Decades of irradiation of U.S. Embassy, fatally injuring three U.S. Ambassadors, health changes in personnel shrugged off ineptly.*
- *Massive building of underground sites, all through former Soviet Union, lamely speculated as perhaps "ABM" system or?.*
- *1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.*
- *Quantum potential strike on soldiers in Gulf war ignored.*
- *Quantum potential deterioration of U.S. populace's immune systems ignored.*
- *Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.*
- *U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.*
- *Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."*
- *Refuse to develop, structure, and apply Bohm's quantum potential, for instantaneous action at a distance. Russians developed it.*

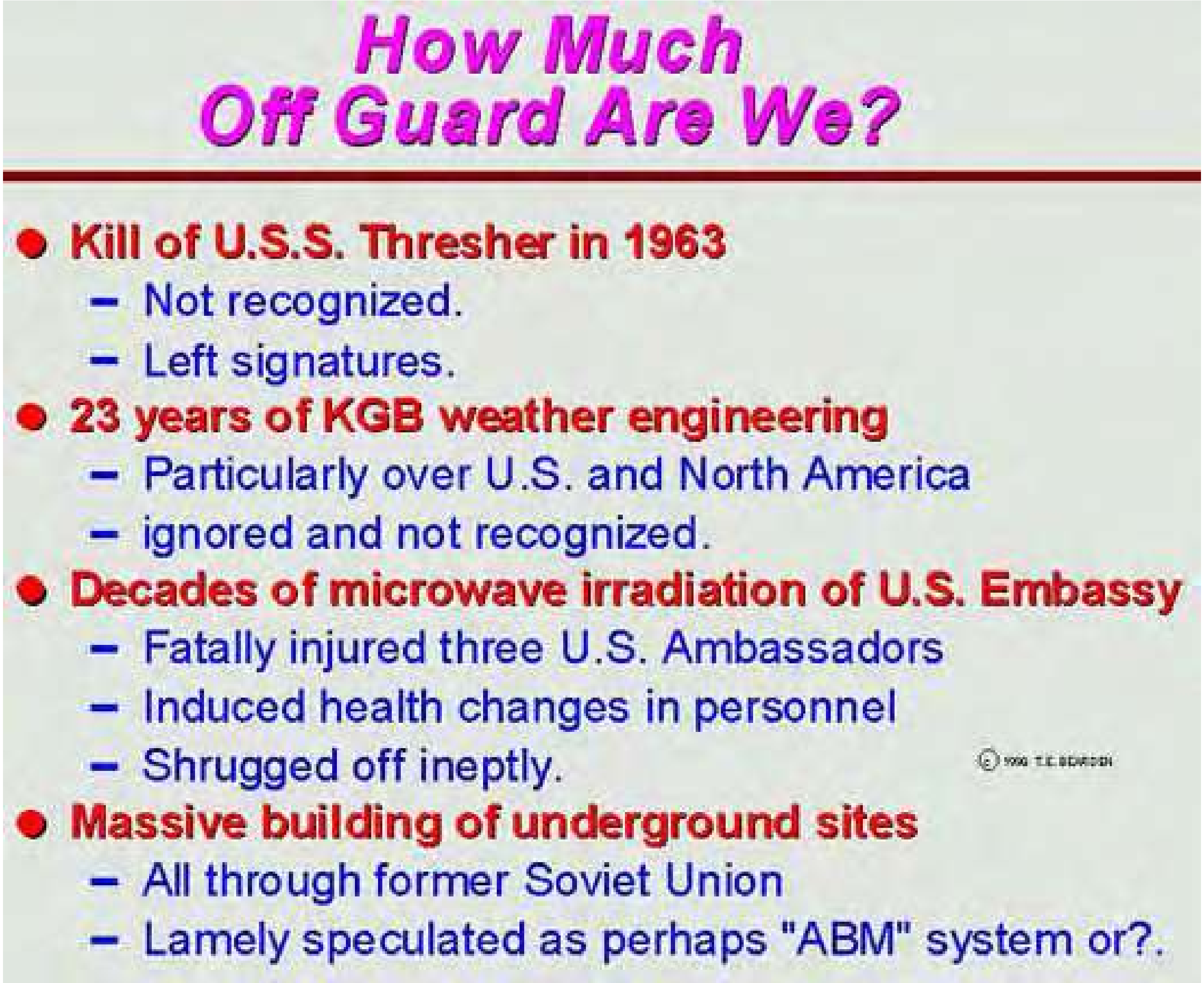

How Much Off Guard Are We?

- **Kill of U.S.S. Thresher in 1963**
  - Not recognized.
  - Left signatures.
- **23 years of KGB weather engineering**
  - Particularly over U.S. and North America
  - ignored and not recognized.
- **Decades of microwave irradiation of U.S. Embassy**
  - Fatally injured three U.S. Ambassadors
  - Induced health changes in personnel
  - Shrugged off ineptly.
- **Massive building of underground sites**
  - All through former Soviet Union
  - Lamely speculated as perhaps "ABM" system or?.

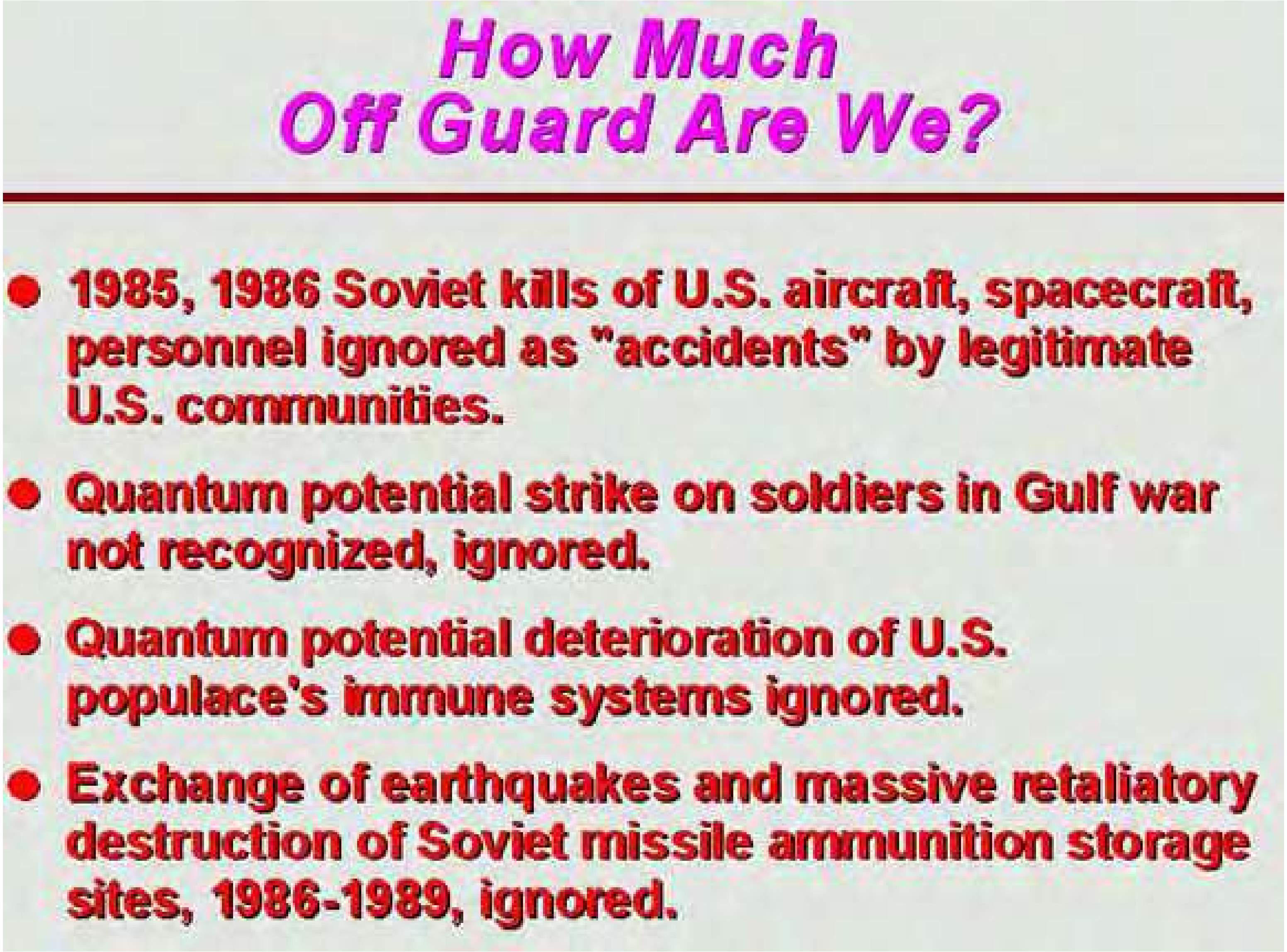

How Much Off Guard Are We?

- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war not recognized, ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.

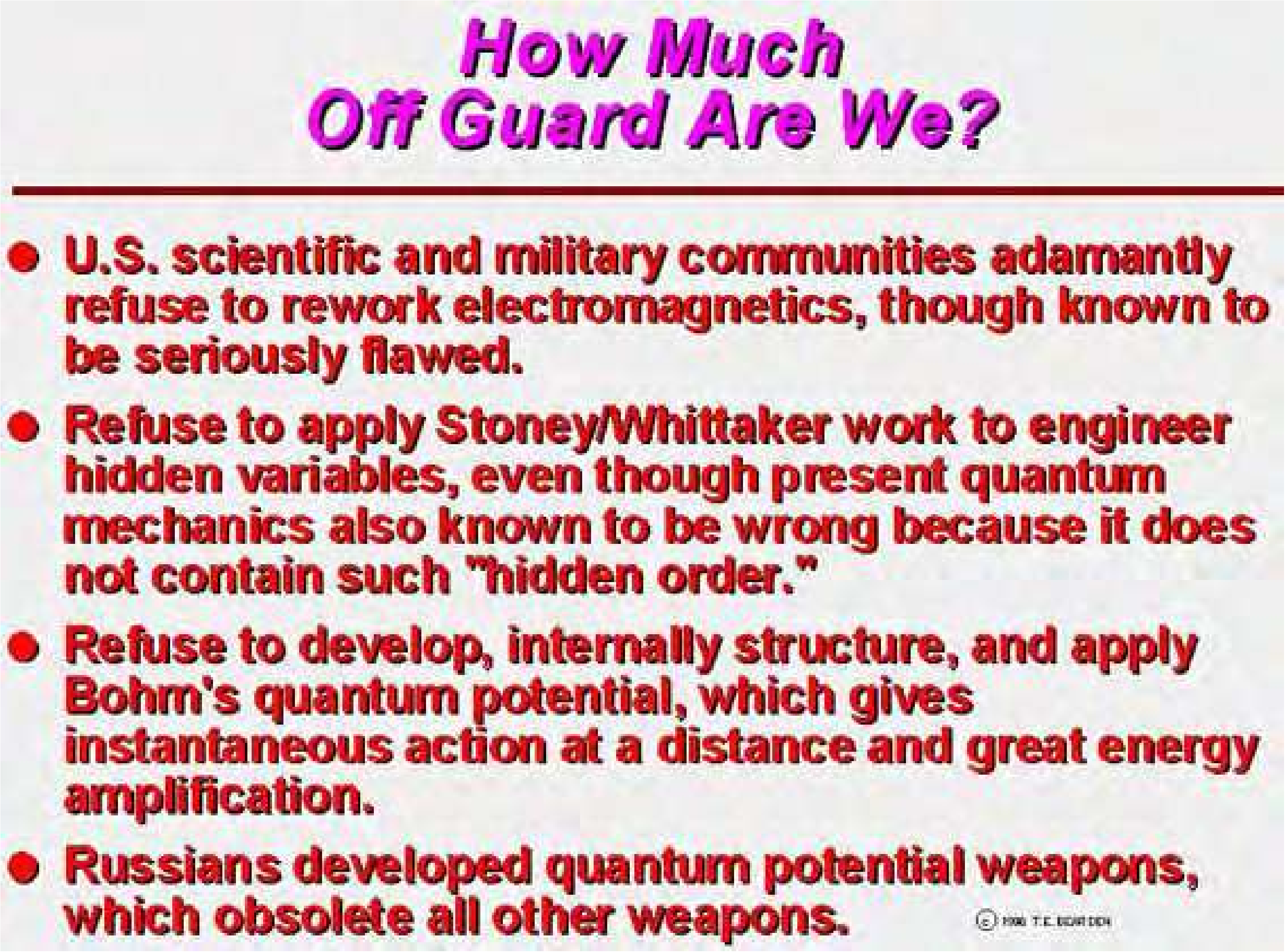

How Much Off Guard Are We?

- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."
- Refuse to develop, internally structure, and apply Bohm's quantum potential, which gives instantaneous action at a distance and great energy amplification.
- Russians developed quantum potential weapons, which obsolete all other weapons.

© [illegible] T.E. BEARDEN

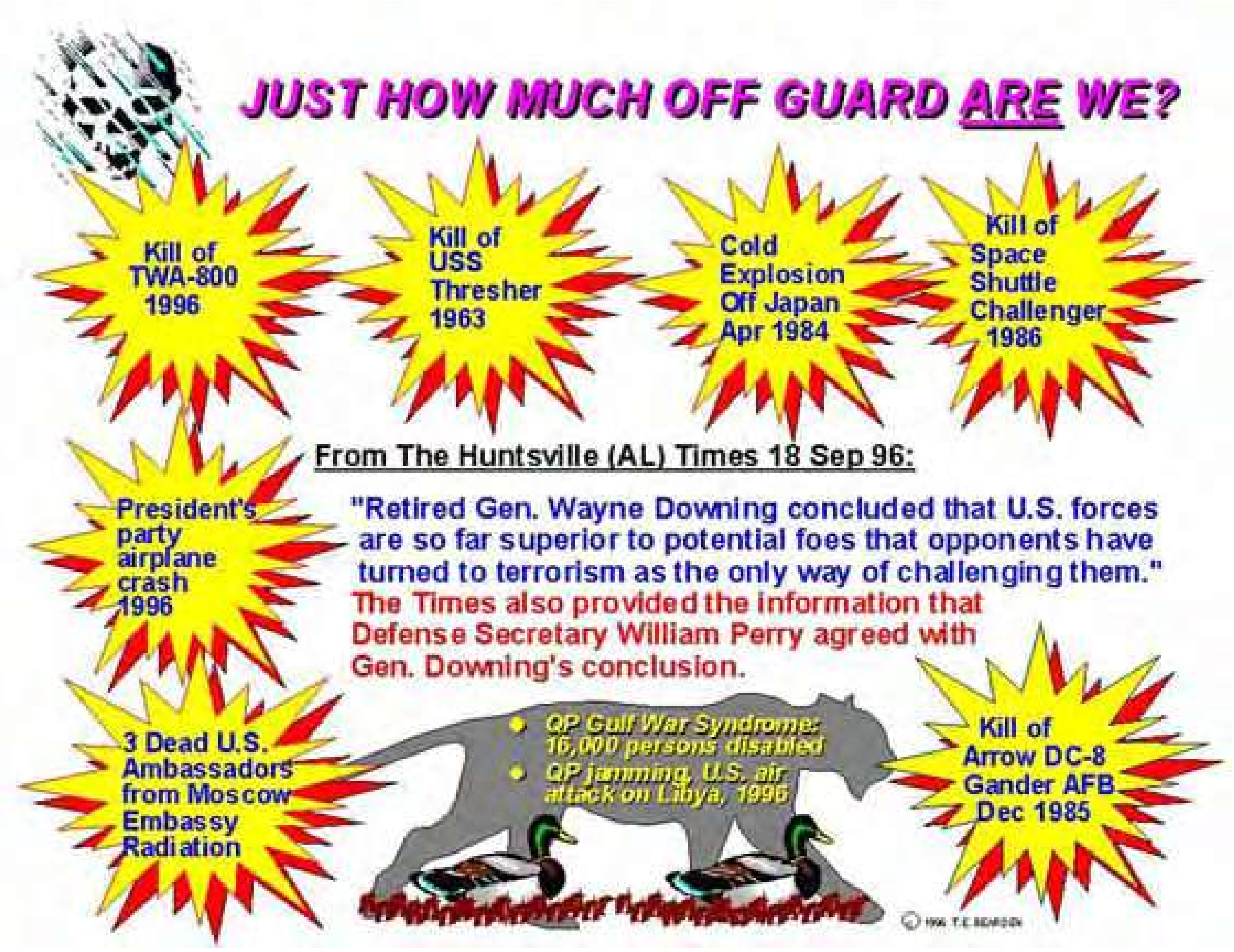

JUST HOW MUCH OFF GUARD ARE WE?

Kill of
TWA-800
1996

Kill of
USS
Thresher
1963

Cold
Explosion
Off Japan
Apr 1984

Kill of
Space
Shuttle
Challenger
1986

President's
party
airplane
crash
1996

From The Huntsville (AL) Times 18 Sep 96:

"Retired Gen. Wayne Downing concluded that U.S. forces are so far superior to potential foes that opponents have turned to terrorism as the only way of challenging them."

The Times also provided the information that Defense Secretary William Perry agreed with Gen. Downing's conclusion.

3 Dead U.S.
Ambassadors
from Moscow
Embassy
Radiation

- *QP Gulf War Syndrome: 16,000 persons disabled*
- *QP jamming, U.S. air attack on Libya, 1996*

Kill of
Arrow DC-8
Gander AFB
Dec 1985

© 1996 T.E. BEARDEN

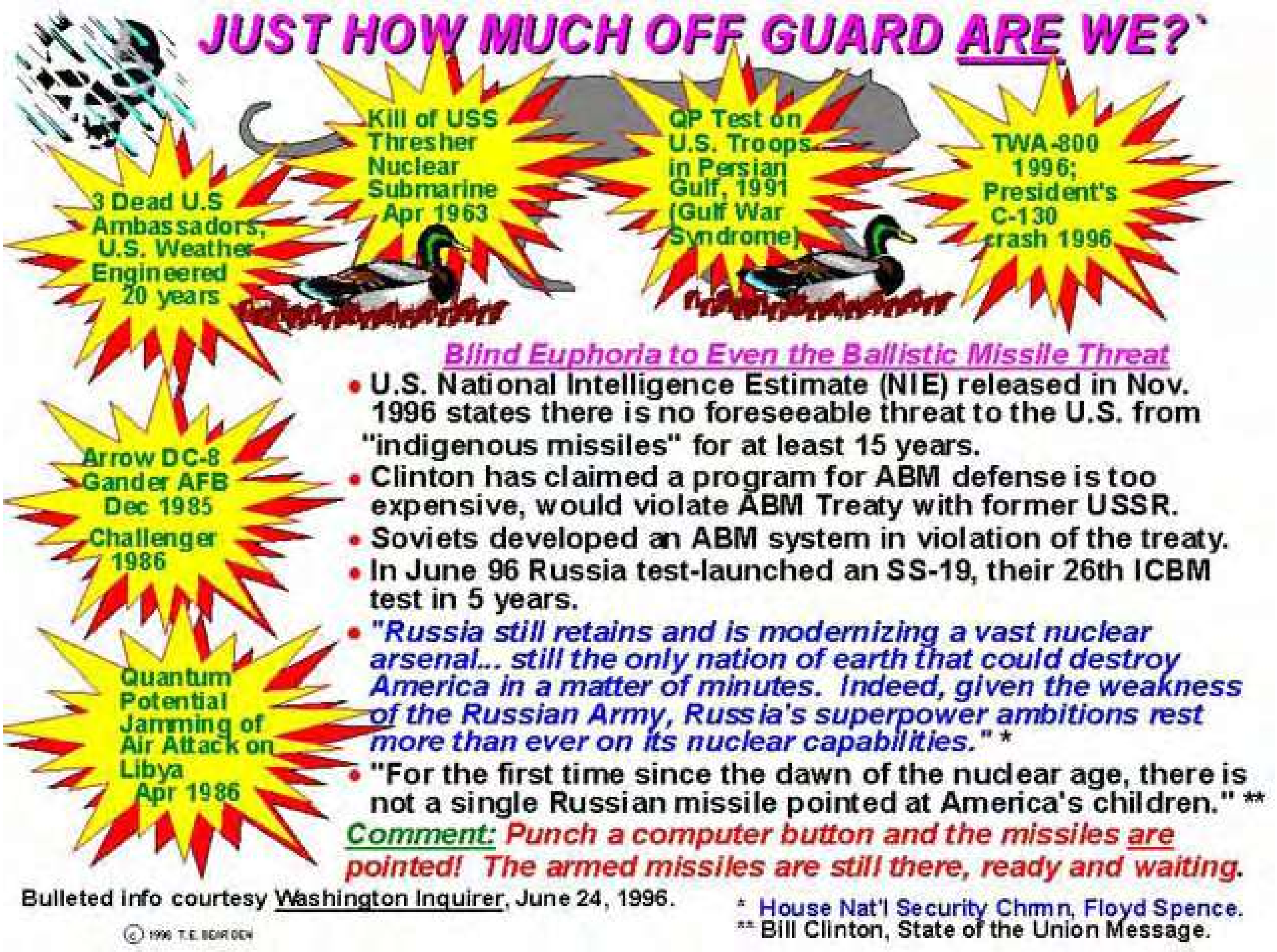

JUST HOW MUCH OFF GUARD ARE WE?

3 Dead U.S Ambassadors, U.S. Weather Engineered 20 years

Kill of USS Thresher Nuclear Submarine Apr 1963

QP Test on U.S. Troops in Persian Gulf, 1991 (Gulf War Syndrome)

TWA-800 1996; President's C-130 crash 1996

Arrow DC-8 Gander AFB Dec 1985

Challenger 1986

Quantum Potential Jamming of Air Attack on Libya Apr 1986

Blind Euphoria to Even the Ballistic Missile Threat

- U.S. National Intelligence Estimate (NIE) released in Nov. 1996 states there is no foreseeable threat to the U.S. from "indigenous missiles" for at least 15 years.
- Clinton has claimed a program for ABM defense is too expensive, would violate ABM Treaty with former USSR.
- Soviets developed an ABM system in violation of the treaty.
- In June 96 Russia test-launched an SS-19, their 26th ICBM test in 5 years.
- *"Russia still retains and is modernizing a vast nuclear arsenal... still the only nation of earth that could destroy America in a matter of minutes. Indeed, given the weakness of the Russian Army, Russia's superpower ambitions rest more than ever on its nuclear capabilities."* *
- "For the first time since the dawn of the nuclear age, there is not a single Russian missile pointed at America's children." **

Comment: Punch a computer button and the missiles are pointed! The armed missiles are still there, ready and waiting.

Bulleted info courtesy Washington Inquirer, June 24, 1996.

* House Nat'l Security Chmn, Floyd Spence.
** Bill Clinton, State of the Union Message.

© 1996 T.E. BEARDEN

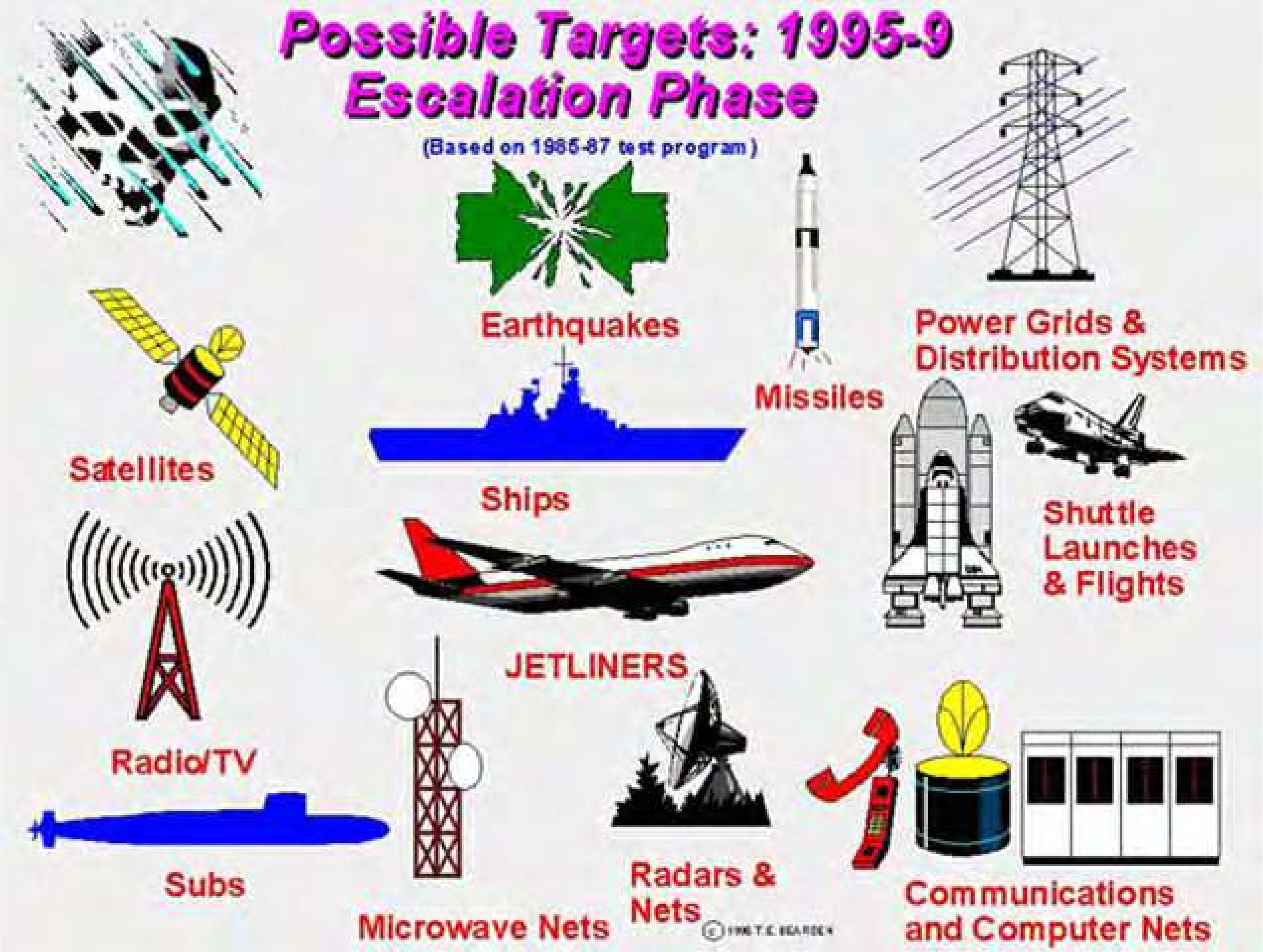

Possible Targets: 1995-9 Escalation Phase

(Based on 1985-87 test program)

Earthquakes
Power Grids & Distribution Systems
Missiles
Satellites
Ships
Shuttle Launches & Flights
JETLINERS
Radio/TV
Subs
Microwave Nets
Radars & Nets
Communications and Computer Nets

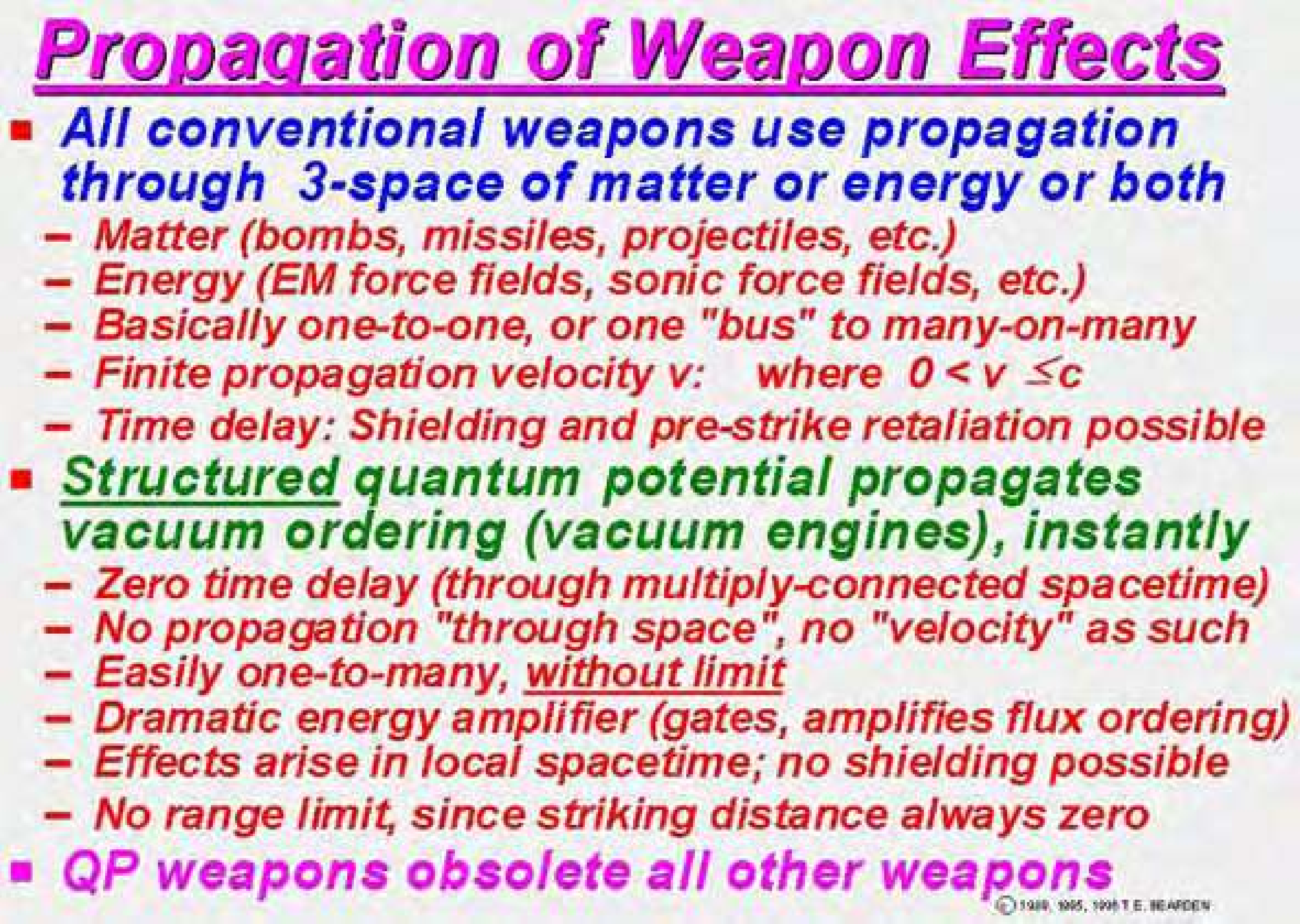

Propagation of Weapon Effects

- All conventional weapons use propagation through 3-space of matter or energy or both
  - Matter (bombs, missiles, projectiles, etc.)
  - Energy (EM force fields, sonic force fields, etc.)
  - Basically one-to-one, or one "bus" to many-on-many
  - Finite propagation velocity v: where $0 < v \leq c$
  - Time delay: Shielding and pre-strike retaliation possible
- Structured quantum potential propagates vacuum ordering (vacuum engines), instantly
  - Zero time delay (through multiply-connected spacetime)
  - No propagation "through space", no "velocity" as such
  - Easily one-to-many, without limit
  - Dramatic energy amplifier (gates, amplifies flux ordering)
  - Effects arise in local spacetime; no shielding possible
  - No range limit, since striking distance always zero
- QP weapons obsolete all other weapons

© 1988, 1995, 1998 T.E. BEARDEN

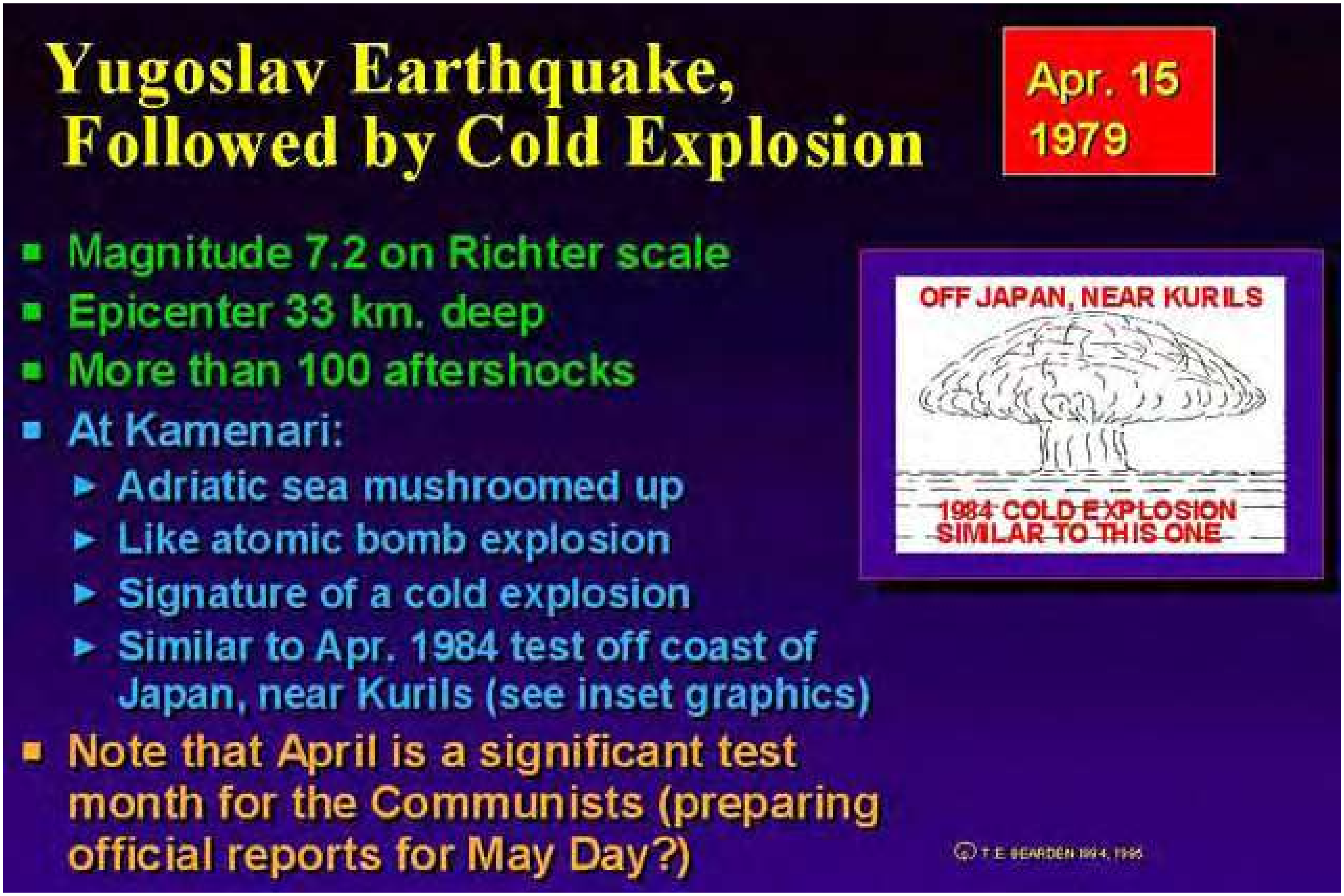

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- Note that April is a significant test month for the Communists (preparing official reports for May Day?)

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

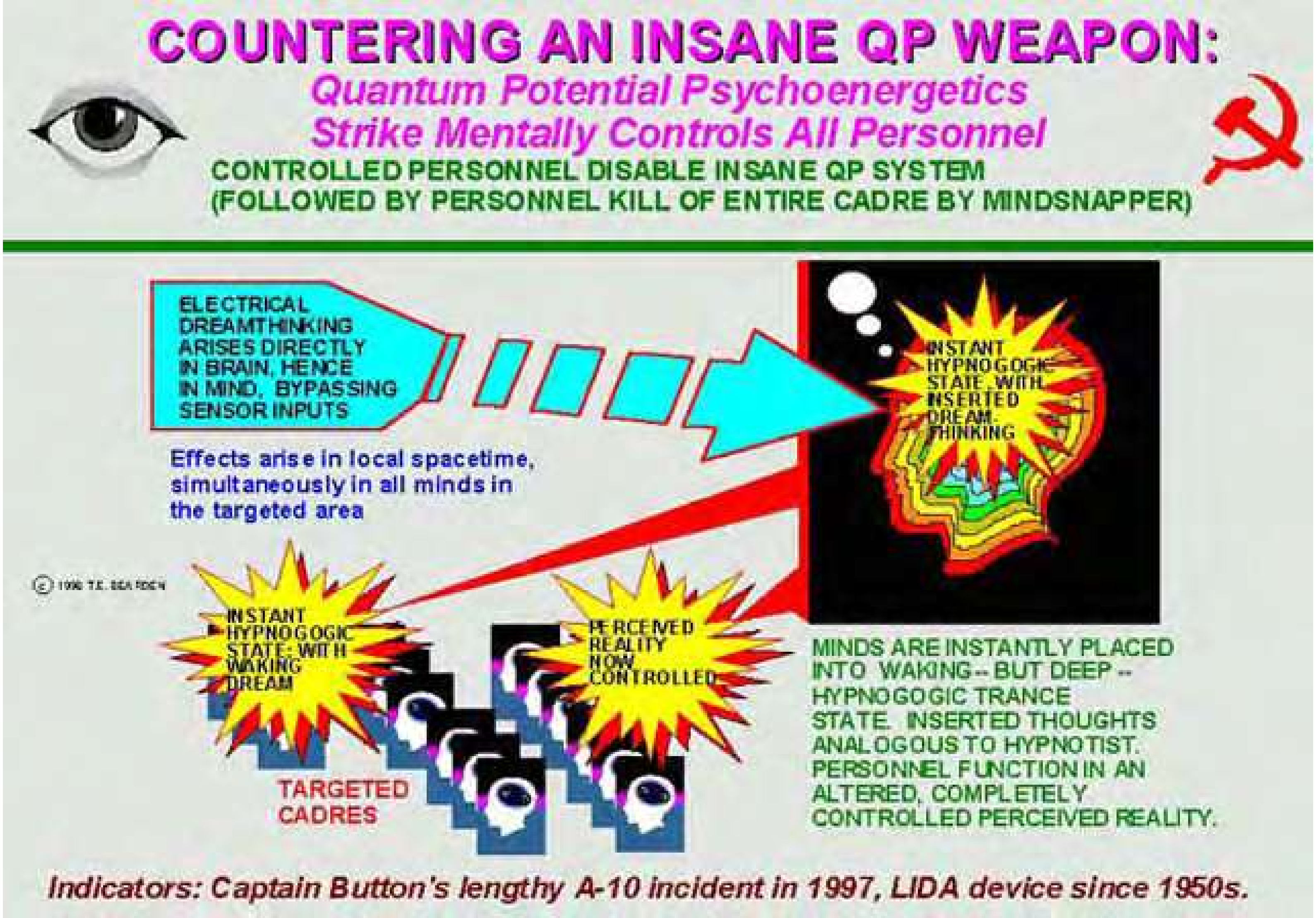

COUNTERING AN INSANE QP WEAPON:

Quantum Potential Psychoenergetics Strike Mentally Controls All Personnel

CONTROLLED PERSONNEL DISABLE INSANE QP SYSTEM
(FOLLOWED BY PERSONNEL KILL OF ENTIRE CADRE BY MINDSNAPPER)

ELECTRICAL DREAMTHINKING ARISES DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED DREAM THINKING
INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM
PERCEIVED REALITY NOW CONTROLLED
TARGETED CADRES

© [illegible] T.E. BEARDEN

MINDS ARE INSTANTLY PLACED INTO WAKING -- BUT DEEP -- HYPNOGOGIC TRANCE STATE. INSERTED THOUGHTS ANALOGOUS TO HYPNOTIST. PERSONNEL FUNCTION IN AN ALTERED, COMPLETELY CONTROLLED PERCEIVED REALITY.

Indicators: Captain Button's lengthy A-10 incident in 1997, LIDA device since 1950s.

Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases

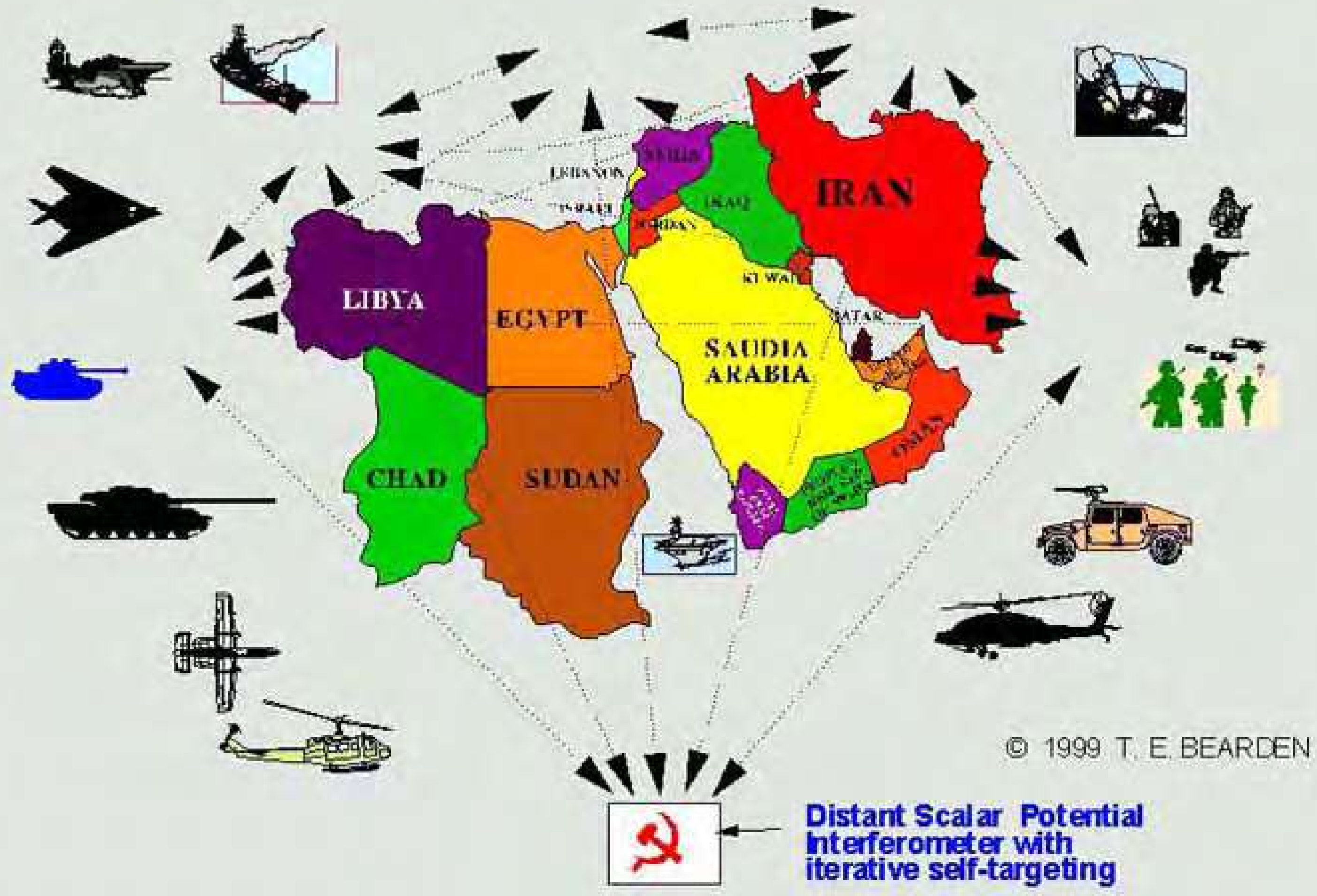
LEBANON
ISRAEL
SYRIA
IRAQ
JORDAN
IRAN
KUWAIT
QATAR
LIBYA
EGYPT
SAUDIA ARABIA
OMAN
CHAD
SUDAN

© 1999 T. E. BEARDEN

Distant Scalar Potential Interferometer with iterative self-targeting

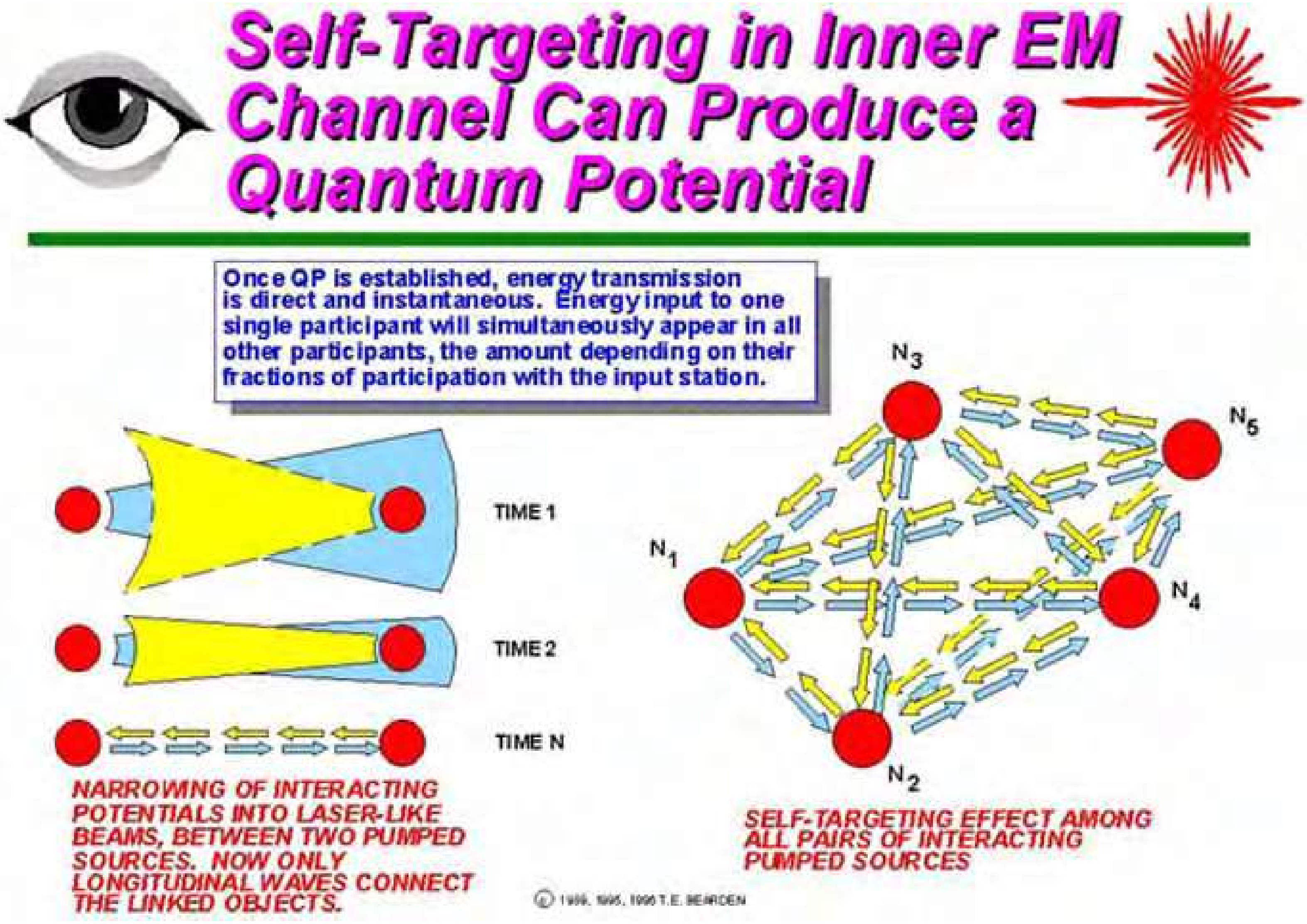

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1989, 1995, 1996 T.E. BEARDEN

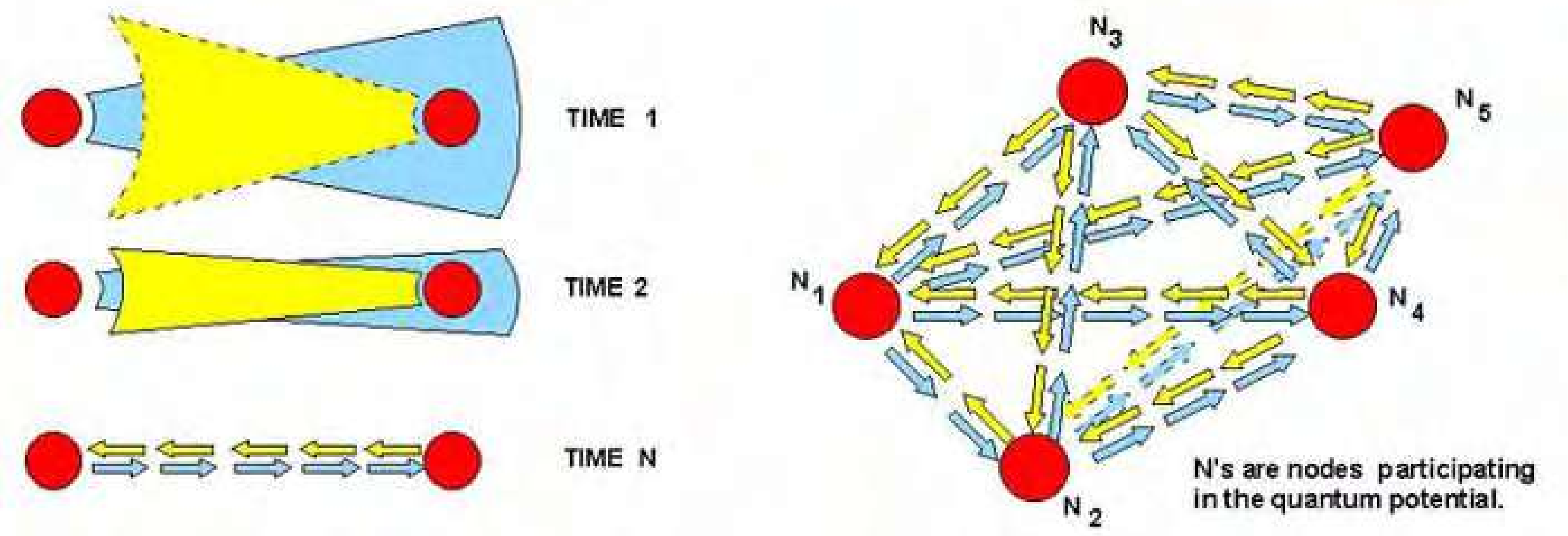

a. Iterative retroreflection between any pair of nodes narrows the interconnecting potentials to a laser-like, bidirectional beam. Iterative phase conjugate reflection provides self-targeting.

b. The quantum potential exists as narrow, laser-like beams between participating nodes, even when the nodes are widely separated. The beam space is multiply-connected; transmission is instantaneous.

© 1989, 1995 T.E. BEARDEN

Consider the hidden Stoney/Whittaker bidirectional EM waves that comprise the potential from each source node. If these nodes have sufficient intersymmetry of dynamic material form, then self-targeting occurs between the two nodes of any nodal pair. By iterative phase conjugate reflection, the scattering is narrowed into interconnecting wavepairs, in laser-like beams, between each two nodes. Thus the interconnecting potentials narrow into laser-like, bidirectional beams. The Coulomb gauge may apply, and a quantum potential may be formed.

Figure 30. Mechanism for generating a quantum potential, by iterative self-targeting induced into a system of multiple scattering nodes. Once a QP is established, spatial separation has little or no effect.

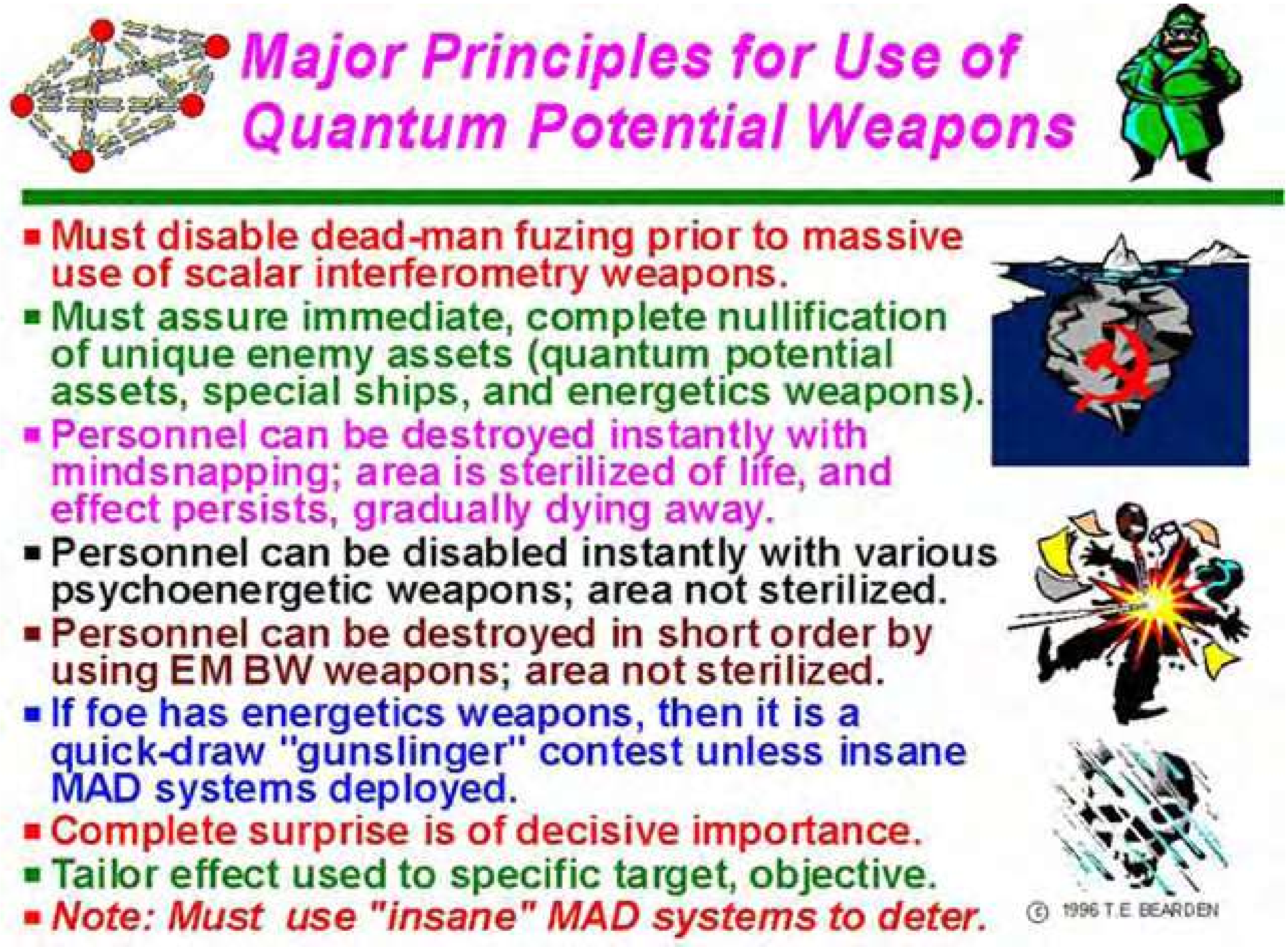

Major Principles for Use of Quantum Potential Weapons

- Must disable dead-man fuzing prior to massive use of scalar interferometry weapons.
- Must assure immediate, complete nullification of unique enemy assets (quantum potential assets, special ships, and energetics weapons).
- Personnel can be destroyed instantly with mindsnapping; area is sterilized of life, and effect persists, gradually dying away.
- Personnel can be disabled instantly with various psychoenergetic weapons; area not sterilized.
- Personnel can be destroyed in short order by using EM BW weapons; area not sterilized.
- If foe has energetics weapons, then it is a quick-draw "gunslinger" contest unless insane MAD systems deployed.
- Complete surprise is of decisive importance.
- Tailor effect used to specific target, objective.
- *Note: Must use "insane" MAD systems to deter.*

© 1996 T.E. BEARDEN

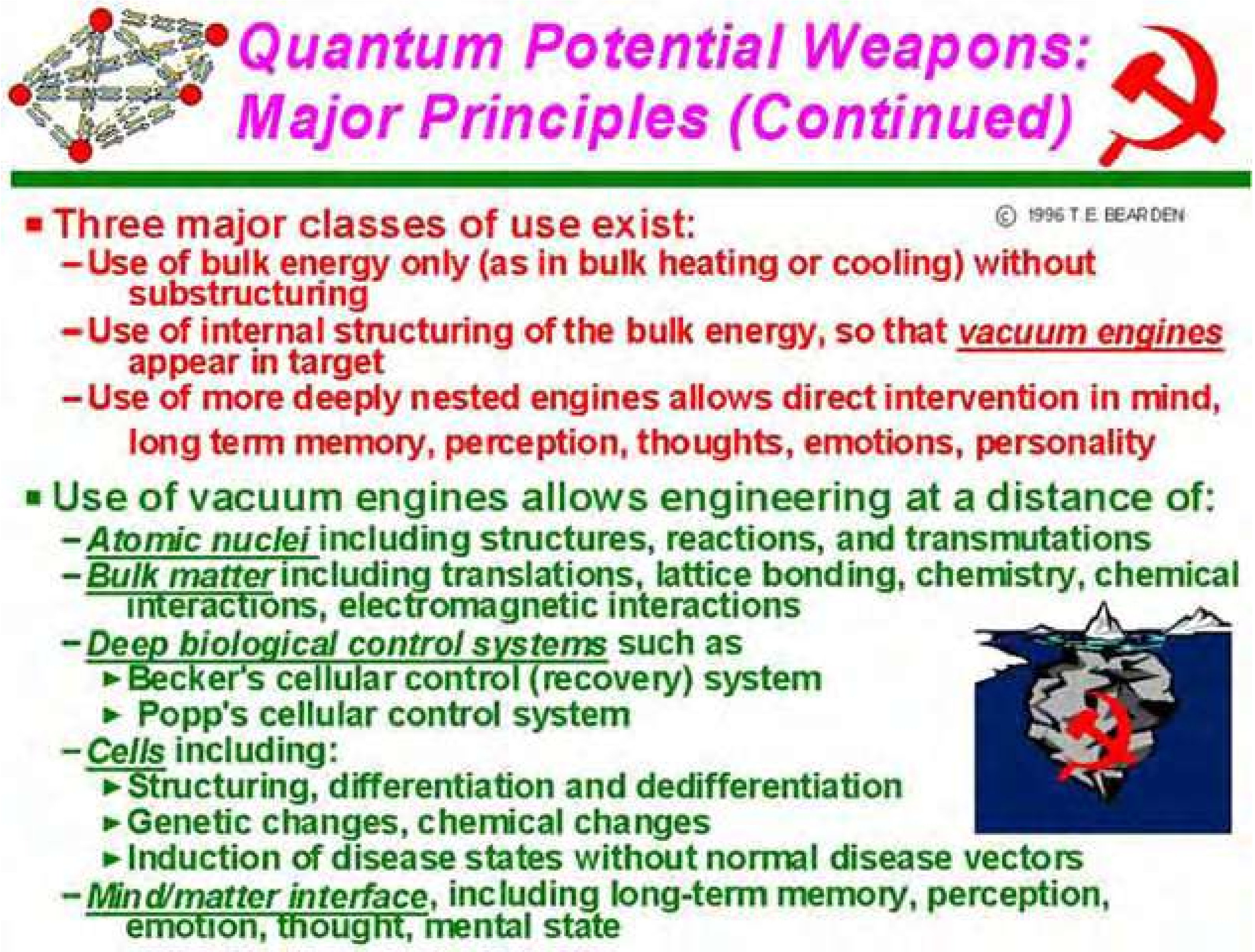

Quantum Potential Weapons: Major Principles (Continued)

© 1996 T. E. BEARDEN

- Three major classes of use exist:
  - Use of bulk energy only (as in bulk heating or cooling) without substructuring
  - Use of internal structuring of the bulk energy, so that *vacuum engines* appear in target
  - Use of more deeply nested engines allows direct intervention in mind, long term memory, perception, thoughts, emotions, personality
- Use of vacuum engines allows engineering at a distance of:
  - *Atomic nuclei* including structures, reactions, and transmutations
  - *Bulk matter* including translations, lattice bonding, chemistry, chemical interactions, electromagnetic interactions
  - *Deep biological control systems* such as
    - Becker's cellular control (recovery) system
    - Popp's cellular control system
  - *Cells* including:
    - Structuring, differentiation and dedifferentiation
    - Genetic changes, chemical changes
    - Induction of disease states without normal disease vectors
  - *Mind/matter interface*, including long-term memory, perception, emotion, thought, mental state

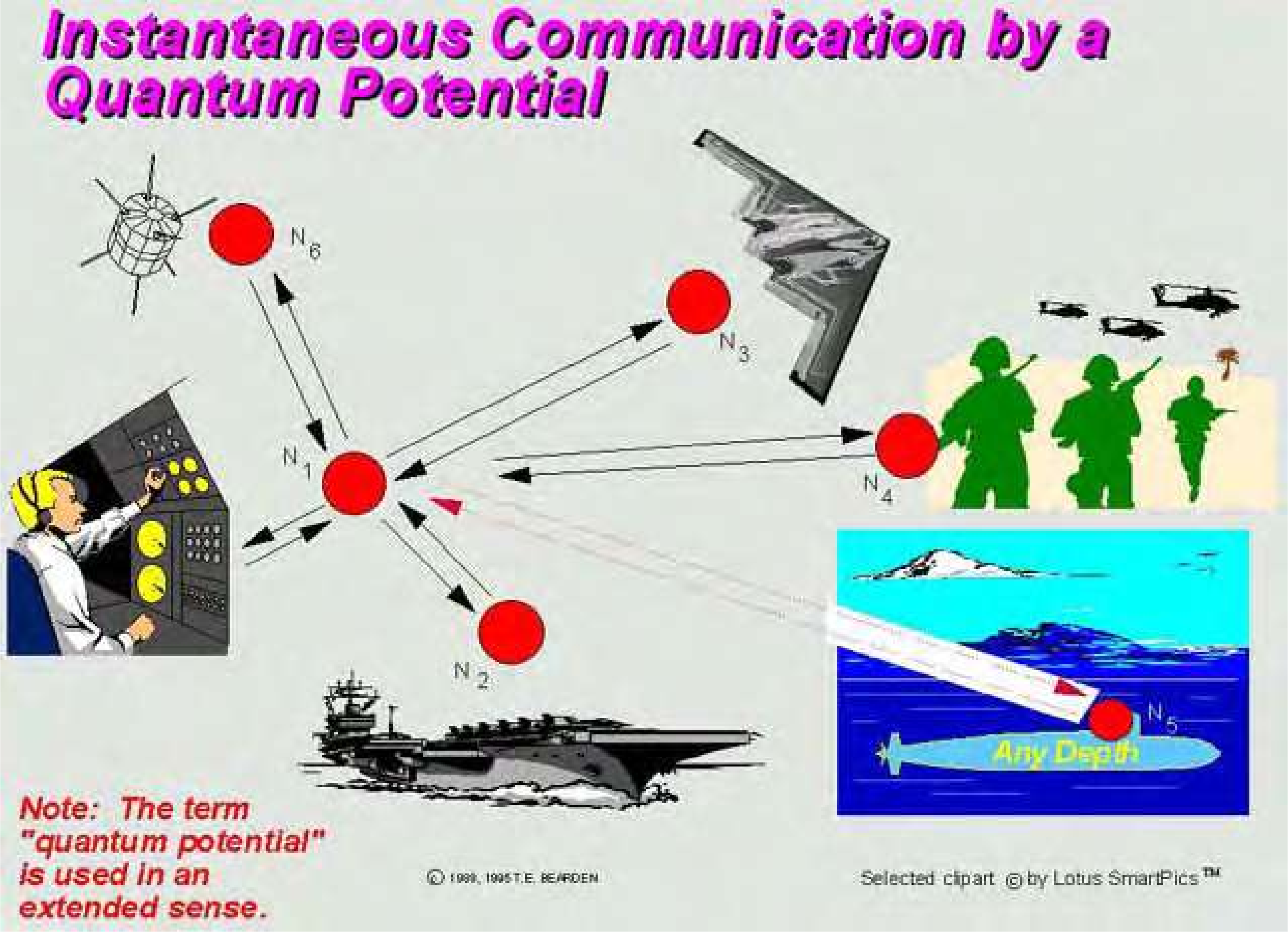

Instantaneous Communication by a Quantum Potential

Note: The term "quantum potential" is used in an extended sense.

© 1989, 1995 T.E. BEARDEN

Selected clipart © by Lotus SmartPics™

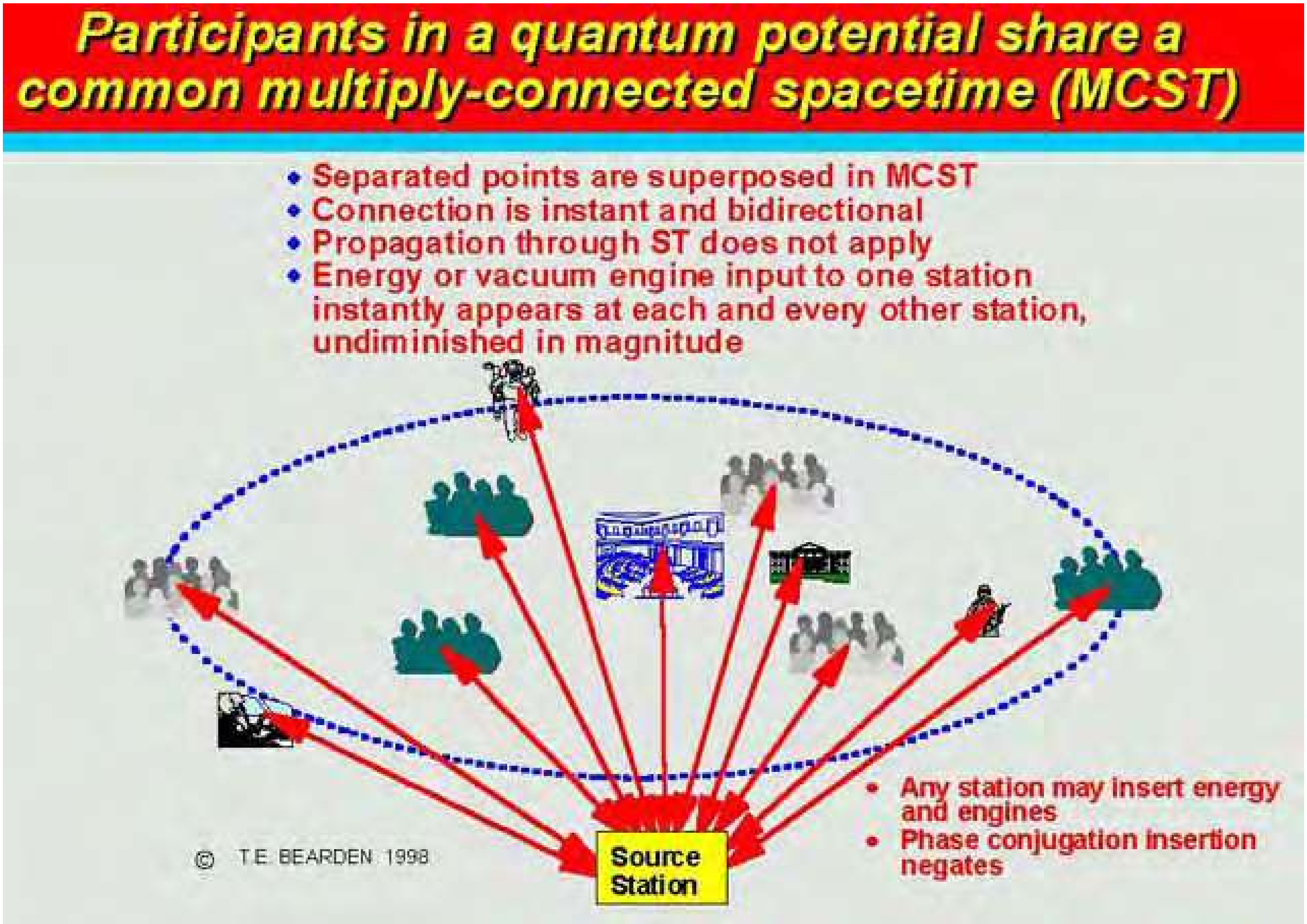

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude

Source Station

© T.E. BEARDEN 1998

- Any station may insert energy and engines
- Phase conjugation insertion negates

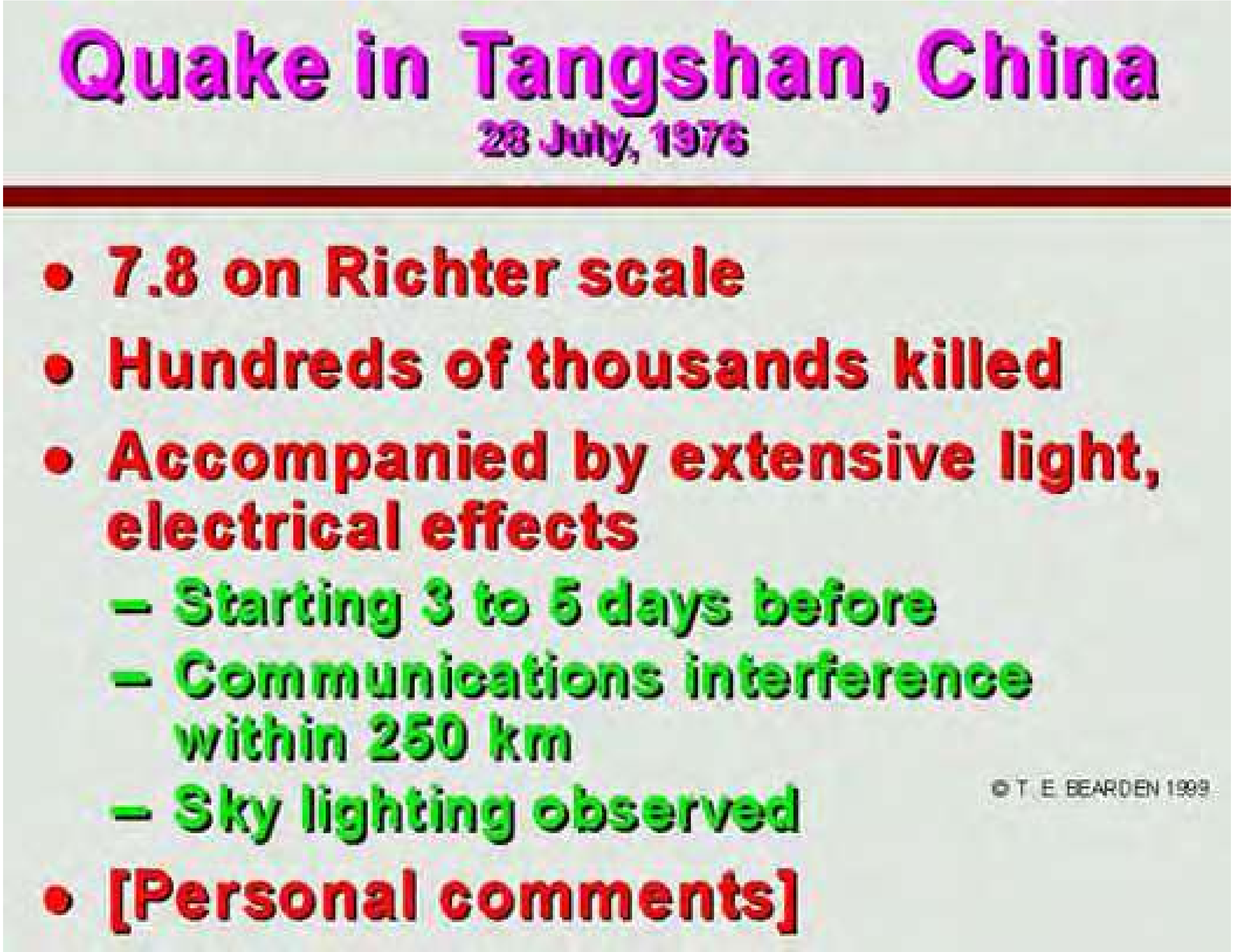

Quake in Tangshan, China

28 July, 1976

- 7.8 on Richter scale
- Hundreds of thousands killed
- Accompanied by extensive light, electrical effects
  - Starting 3 to 5 days before
  - Communications interference within 250 km
  - Sky lighting observed
- [Personal comments]

© T. E. BEARDEN 1999

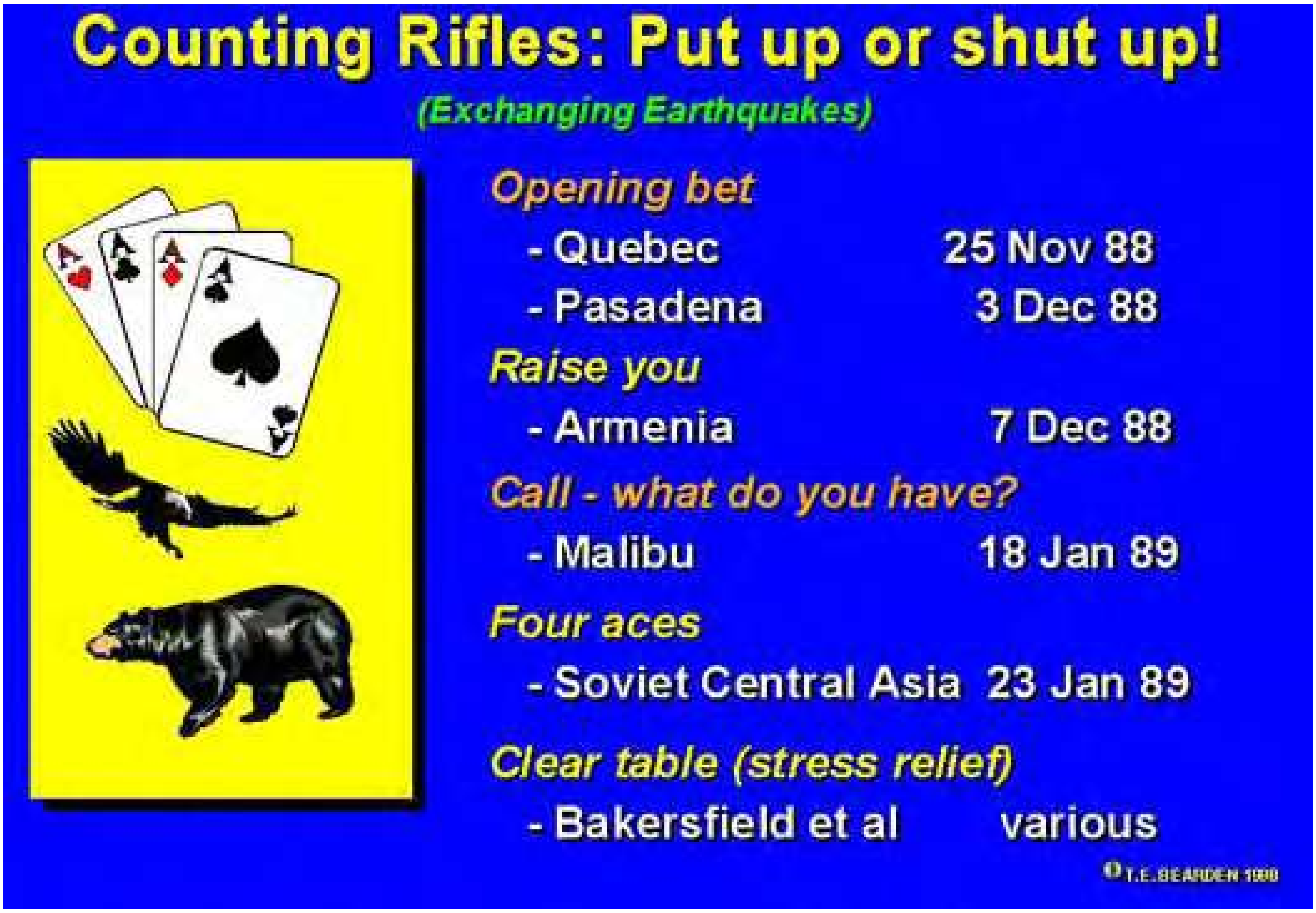

Counting Rifles: Put up or shut up!

(Exchanging Earthquakes)

Opening bet	
- Quebec	25 Nov 88
- Pasadena	3 Dec 88
Raise you	
- Armenia	7 Dec 88
Call - what do you have?	
- Malibu	18 Jan 89
Four aces	
- Soviet Central Asia	23 Jan 89
Clear table (stress relief)	
- Bakersfield et al	various

© T.E. BEARDEN 1990

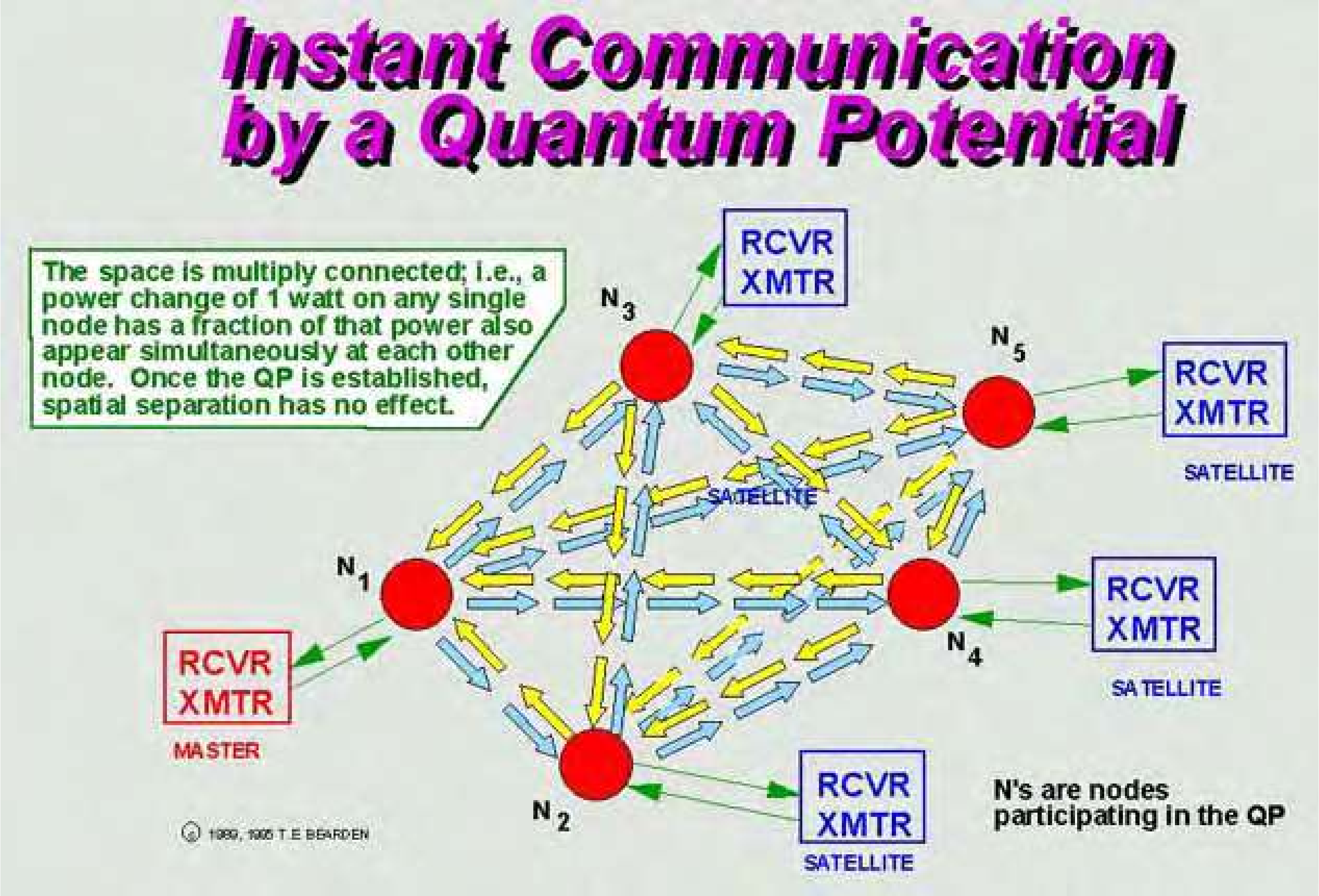

Instant Communication by a Quantum Potential

The space is multiply connected; i.e., a power change of 1 watt on any single node has a fraction of that power also appear simultaneously at each other node. Once the QP is established, spatial separation has no effect.

N's are nodes participating in the QP

N3
RCVR
XMTR
SATELLITE
N5
RCVR
XMTR
SATELLITE
N1
RCVR
XMTR
MASTER
N4
RCVR
XMTR
SATELLITE
N2
RCVR
XMTR
SATELLITE

© 1989, 1995 T.E. BEARDEN

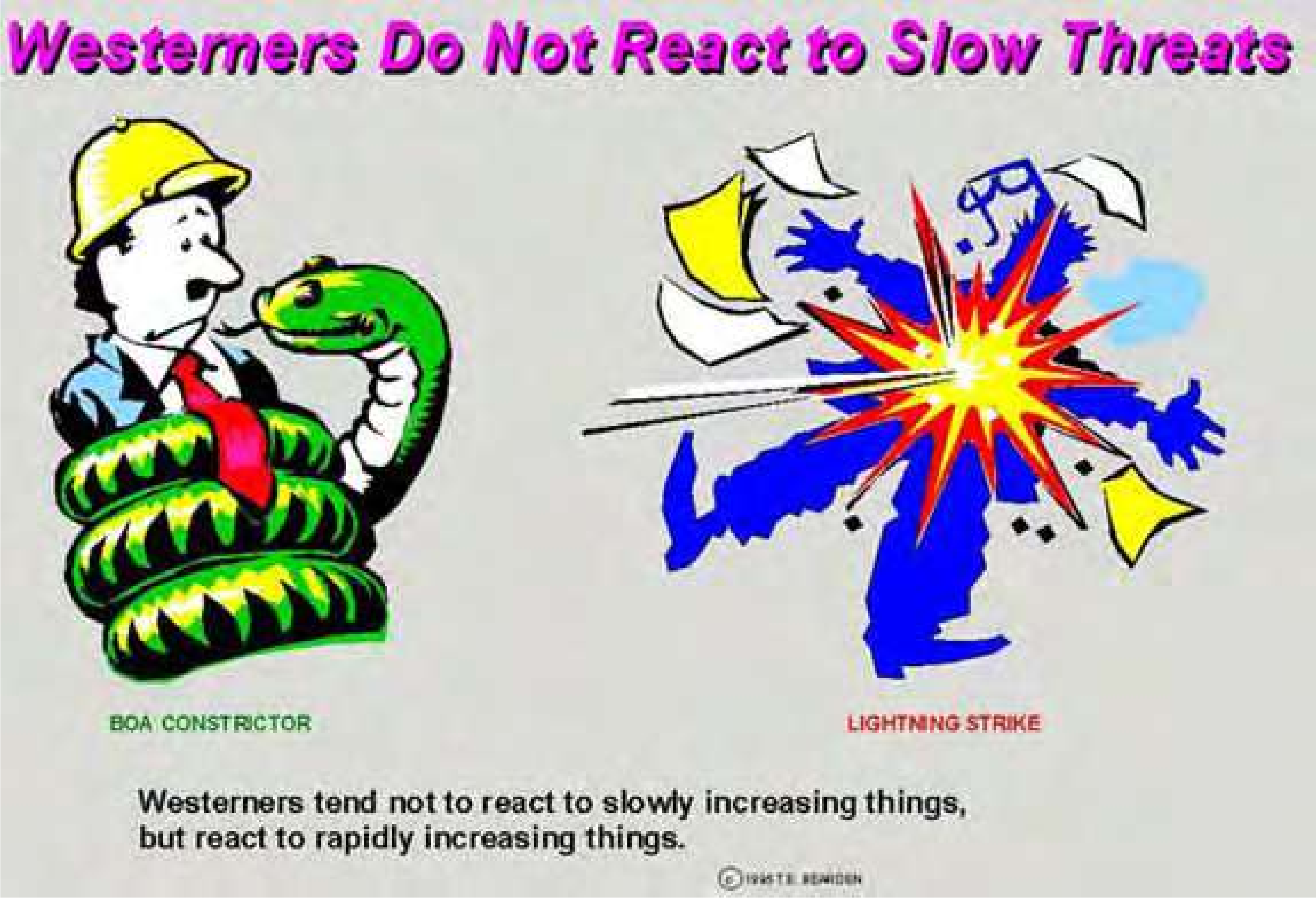

Westerners Do Not React to Slow Threats

BOA CONSTRICTOR
LIGHTNING STRIKE

Westerners tend not to react to slowly increasing things, but react to rapidly increasing things.

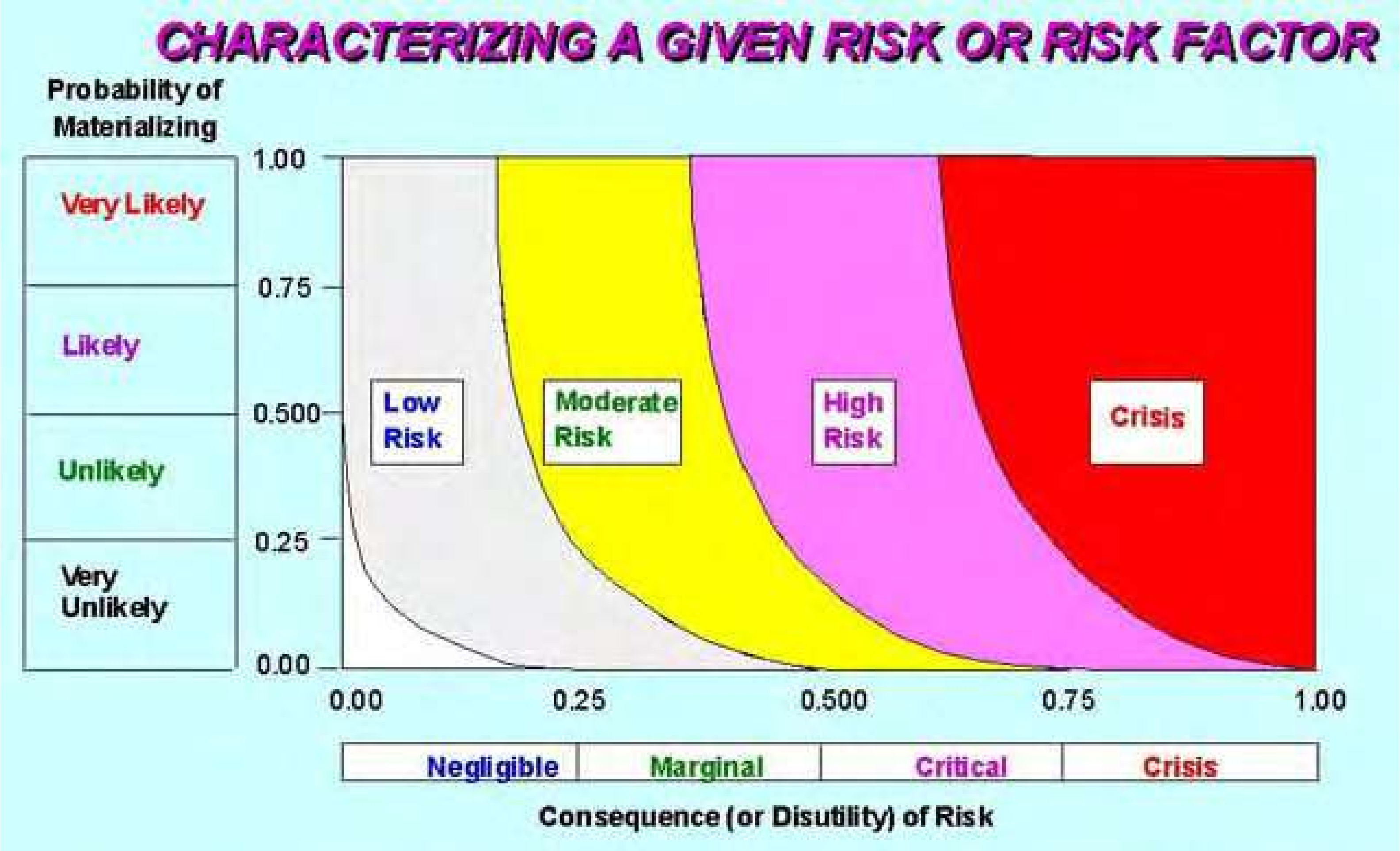
CHARACTERIZING A GIVEN RISK OR RISK FACTOR
Probability of Materializing
Very Likely
Likely
Unlikely
Very Unlikely
1.00
0.75
0.500
0.25
0.00
Low Risk
Moderate Risk
High Risk
Crisis
0.00
0.25
0.500
0.75
1.00
Negligible
Marginal
Critical
Crisis
Consequence (or Disutility) of Risk

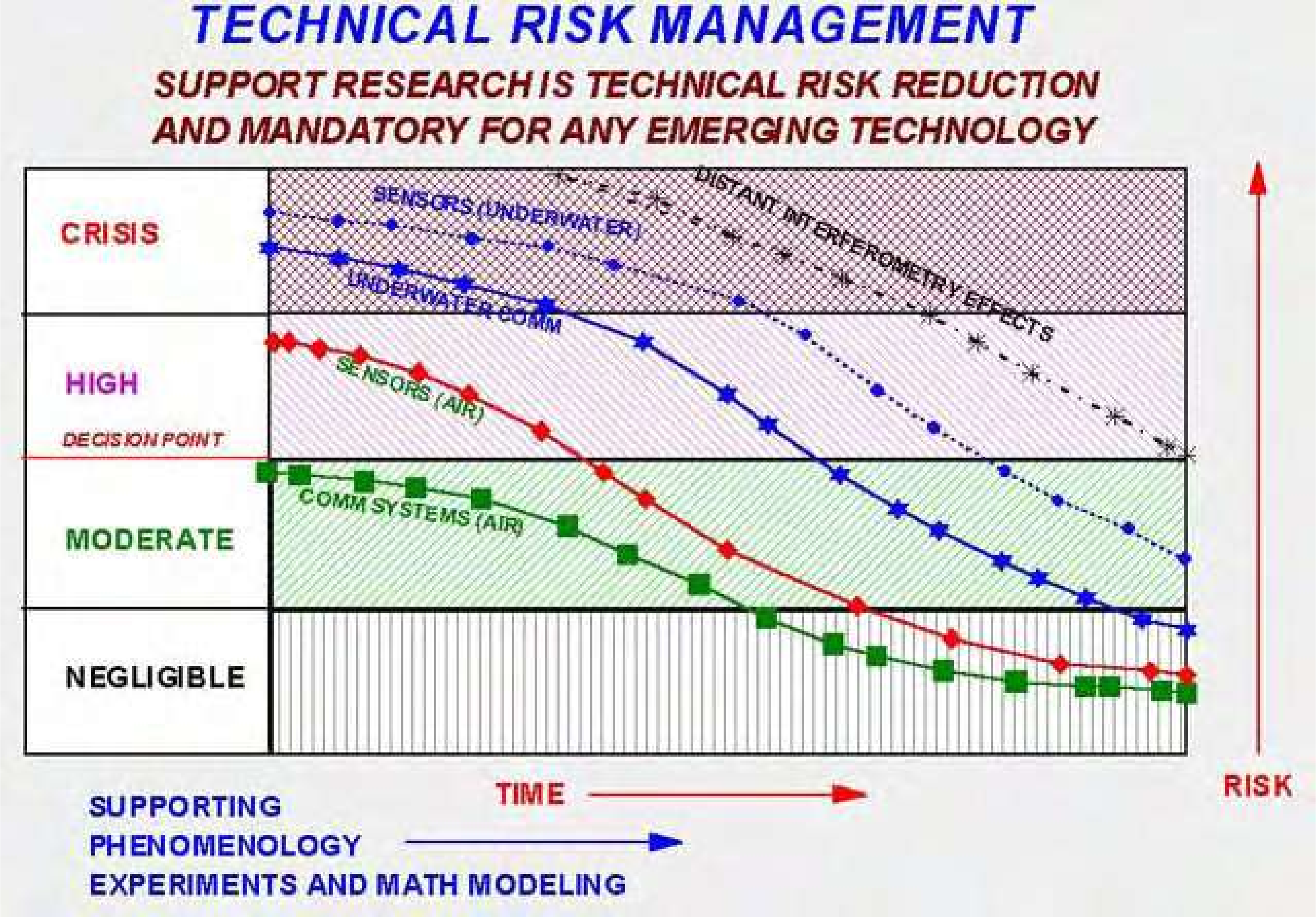

TECHNICAL RISK MANAGEMENT

SUPPORT RESEARCH IS TECHNICAL RISK REDUCTION AND MANDATORY FOR ANY EMERGING TECHNOLOGY

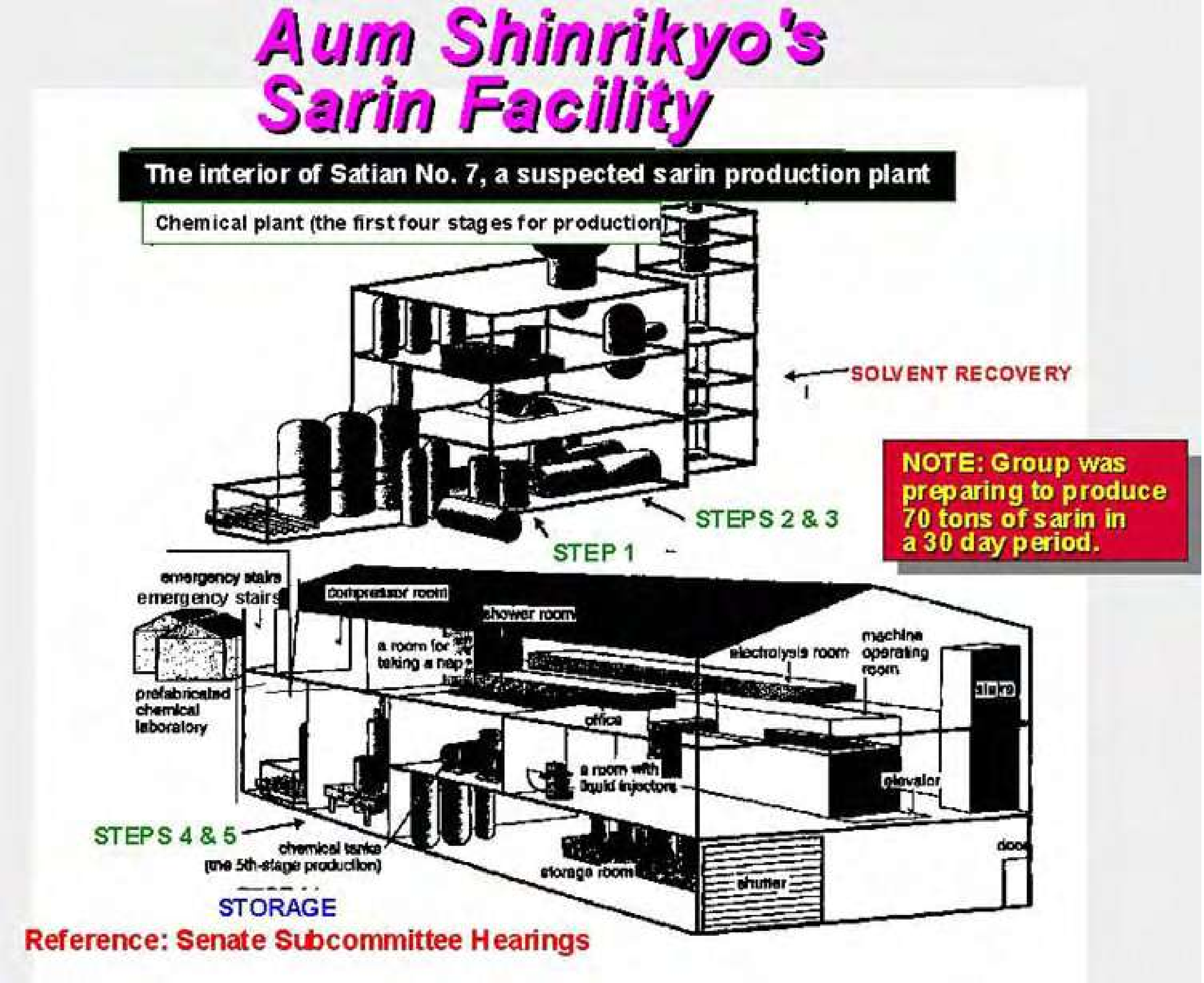

Aum Shinrikyo's Sarin Facility

The interior of Satian No. 7, a suspected sarin production plant

Chemical plant (the first four stages for production)

NOTE: Group was preparing to produce 70 tons of sarin in a 30 day period.

SOLVENT RECOVERY
STEPS 2 & 3
STEP 1
emergency stairs
emergency stairs
compressor room
shower room
a room for taking a nap
electrolysis room
machine operating room
stairs
prefabricated chemical laboratory
office
a room with liquid injectors
elevator
STEPS 4 & 5
chemical tanks (the 5th-stage production)
storage room
shutter
door
STORAGE

Reference: Senate Subcommittee Hearings

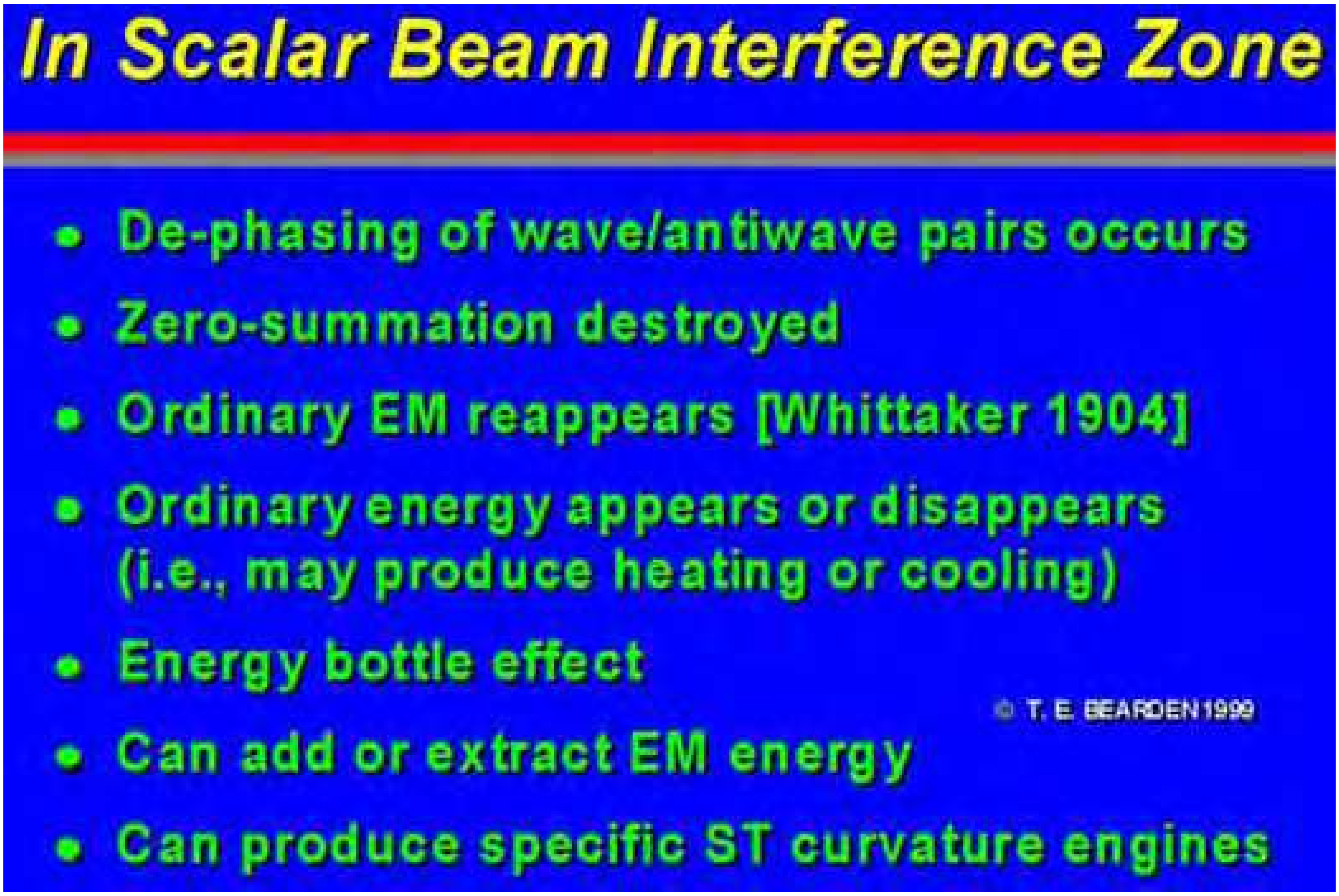

In Scalar Beam Interference Zone

- De-phasing of wave/antiwave pairs occurs
- Zero-summation destroyed
- Ordinary EM reappears [Whittaker 1904]
- Ordinary energy appears or disappears (i.e., may produce heating or cooling)
- Energy bottle effect
- Can add or extract EM energy
- Can produce specific ST curvature engines

© T. E. BEARDEN 1999

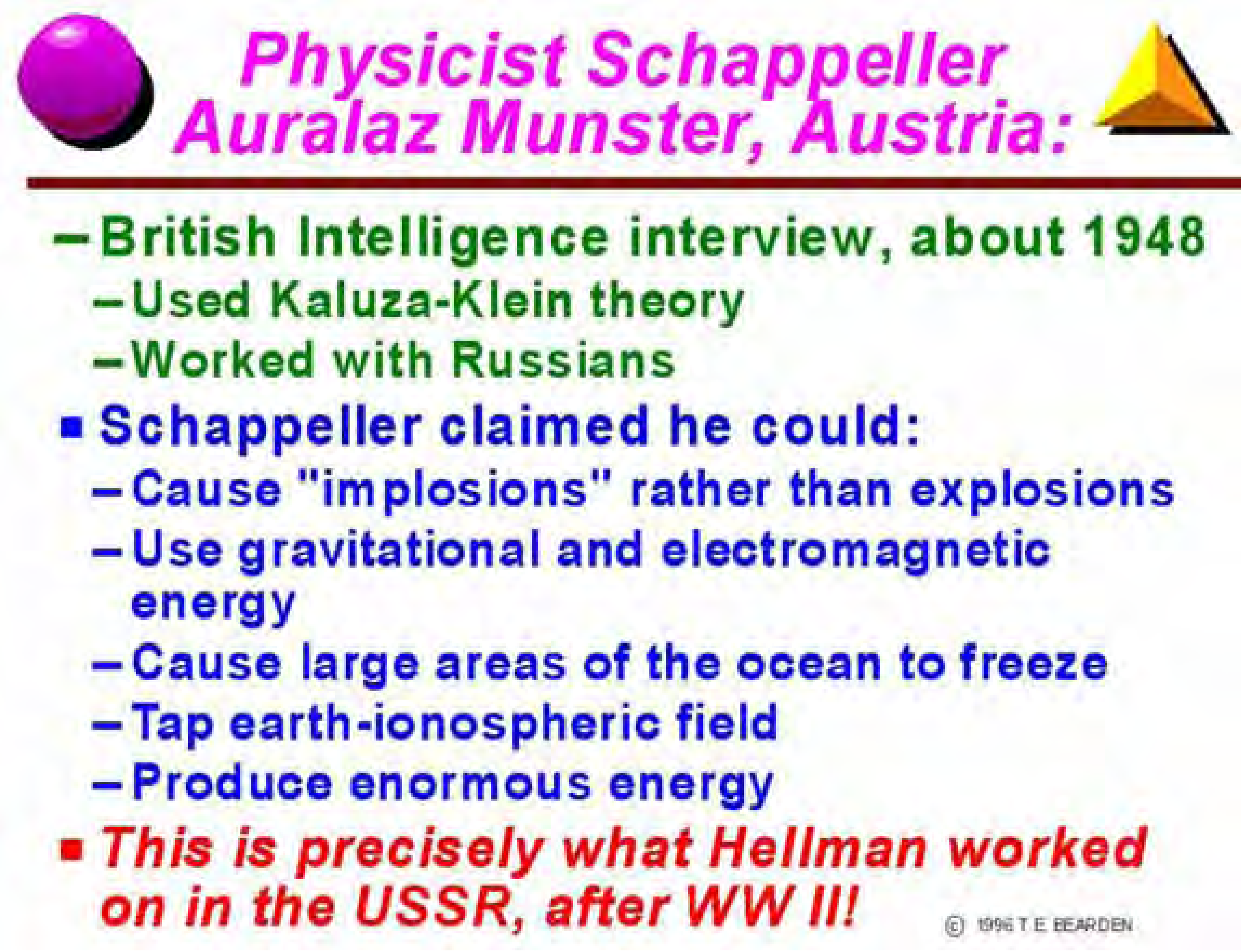

Physicist Schappeller Auralaz Munster, Austria:

- British Intelligence interview, about 1948
  - Used Kaluza-Klein theory
  - Worked with Russians
- Schappeller claimed he could:
  - Cause "implosions" rather than explosions
  - Use gravitational and electromagnetic energy
  - Cause large areas of the ocean to freeze
  - Tap earth-ionospheric field
  - Produce enormous energy
- *This is precisely what Hellman worked on in the USSR, after WW II!*

© 1996 T.E. BEARDEN

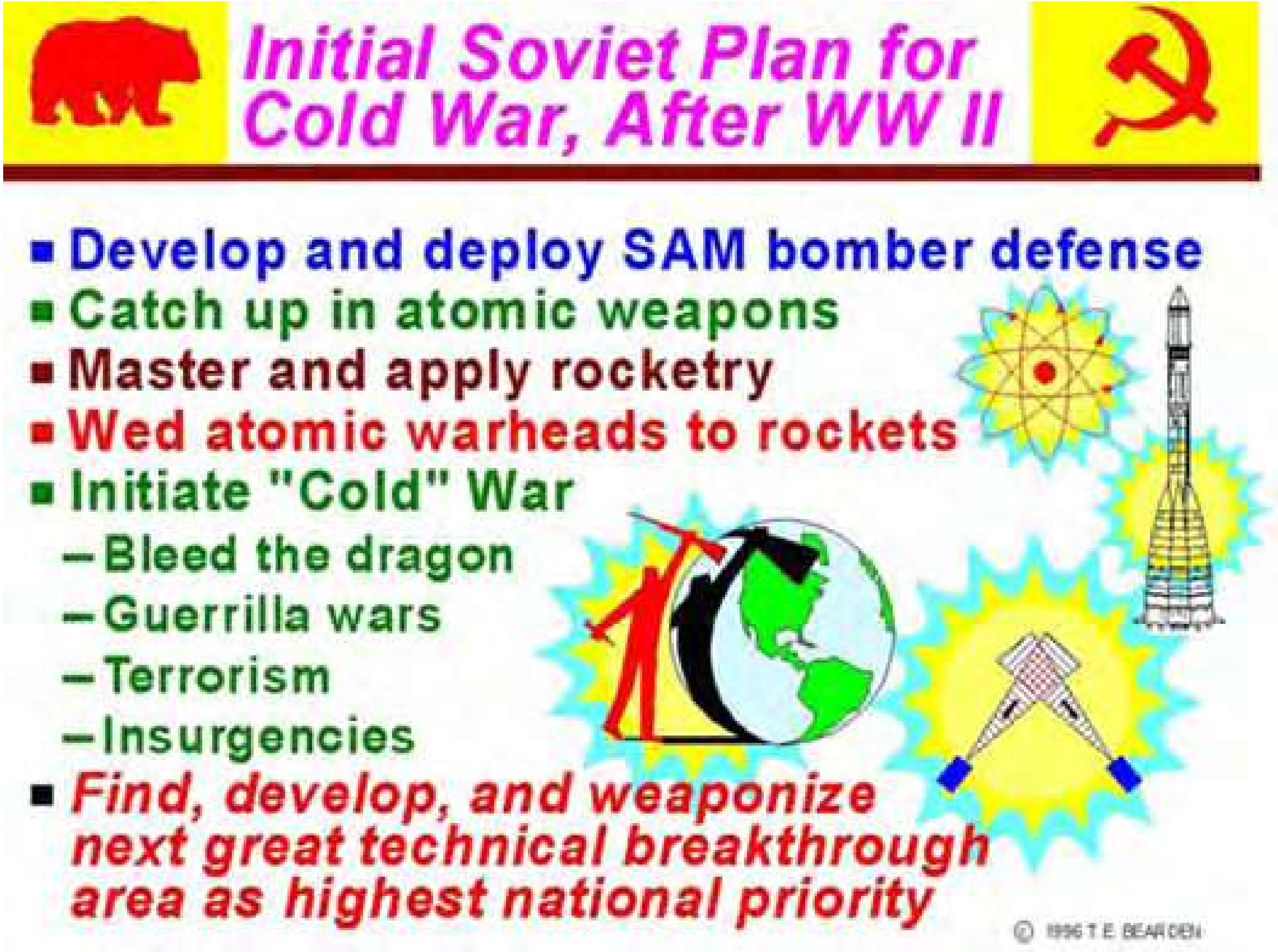

Initial Soviet Plan for Cold War, After WW II

- Develop and deploy SAM bomber defense
- Catch up in atomic weapons
- Master and apply rocketry
- Wed atomic warheads to rockets
- Initiate "Cold" War
  - Bleed the dragon
  - Guerrilla wars
  - Terrorism
  - Insurgencies
- *Find, develop, and weaponize next great technical breakthrough area as highest national priority*

© 1996 T.E. BEARDEN

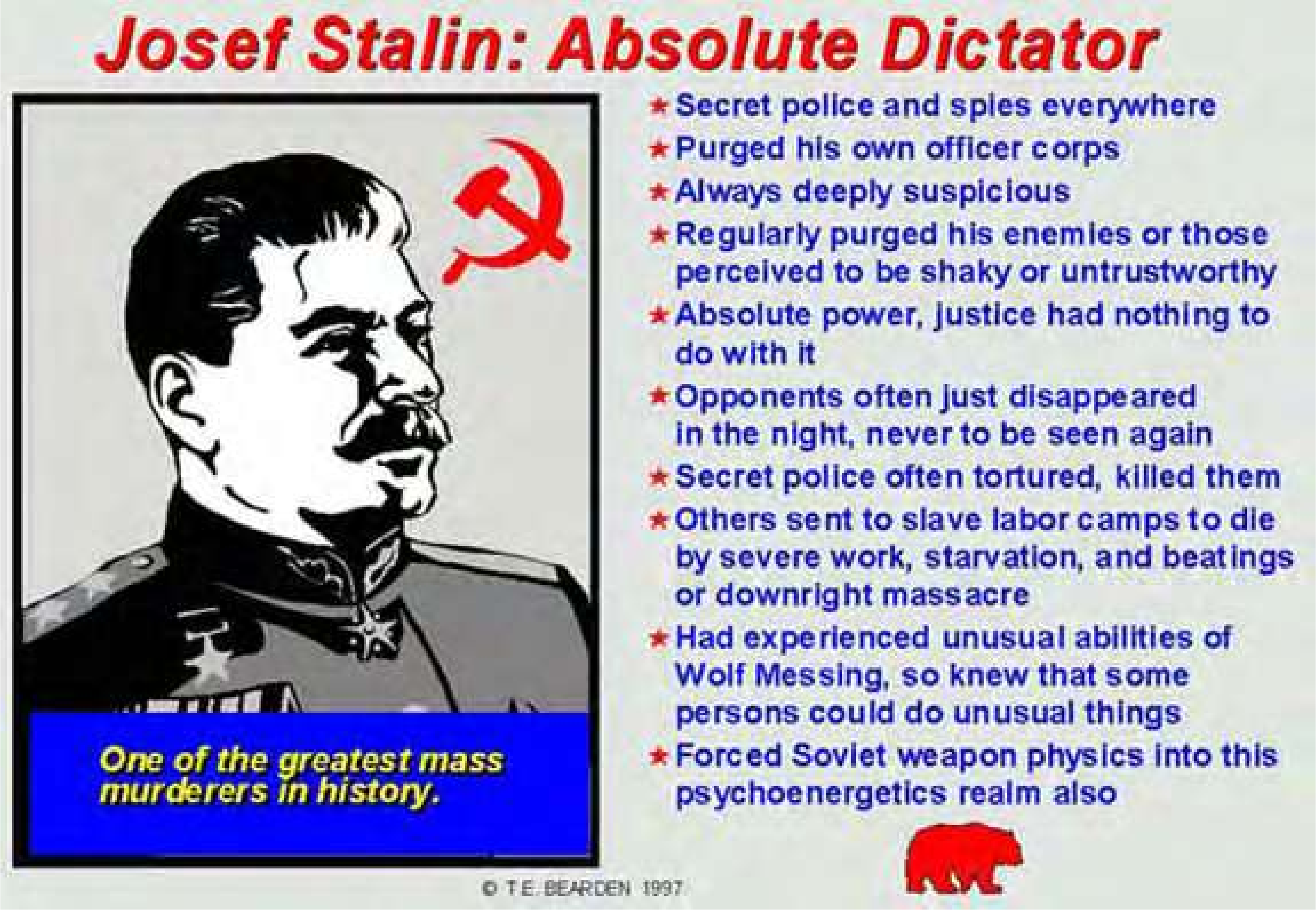

Josef Stalin: Absolute Dictator

- Secret police and spies everywhere
- Purged his own officer corps
- Always deeply suspicious
- Regularly purged his enemies or those perceived to be shaky or untrustworthy
- Absolute power, justice had nothing to do with it
- Opponents often just disappeared in the night, never to be seen again
- Secret police often tortured, killed them
- Others sent to slave labor camps to die by severe work, starvation, and beatings or downright massacre
- Had experienced unusual abilities of Wolf Messing, so knew that some persons could do unusual things
- Forced Soviet weapon physics into this psychoenergetics realm also

One of the greatest mass murderers in history.

© T.E. BEARDEN 1997

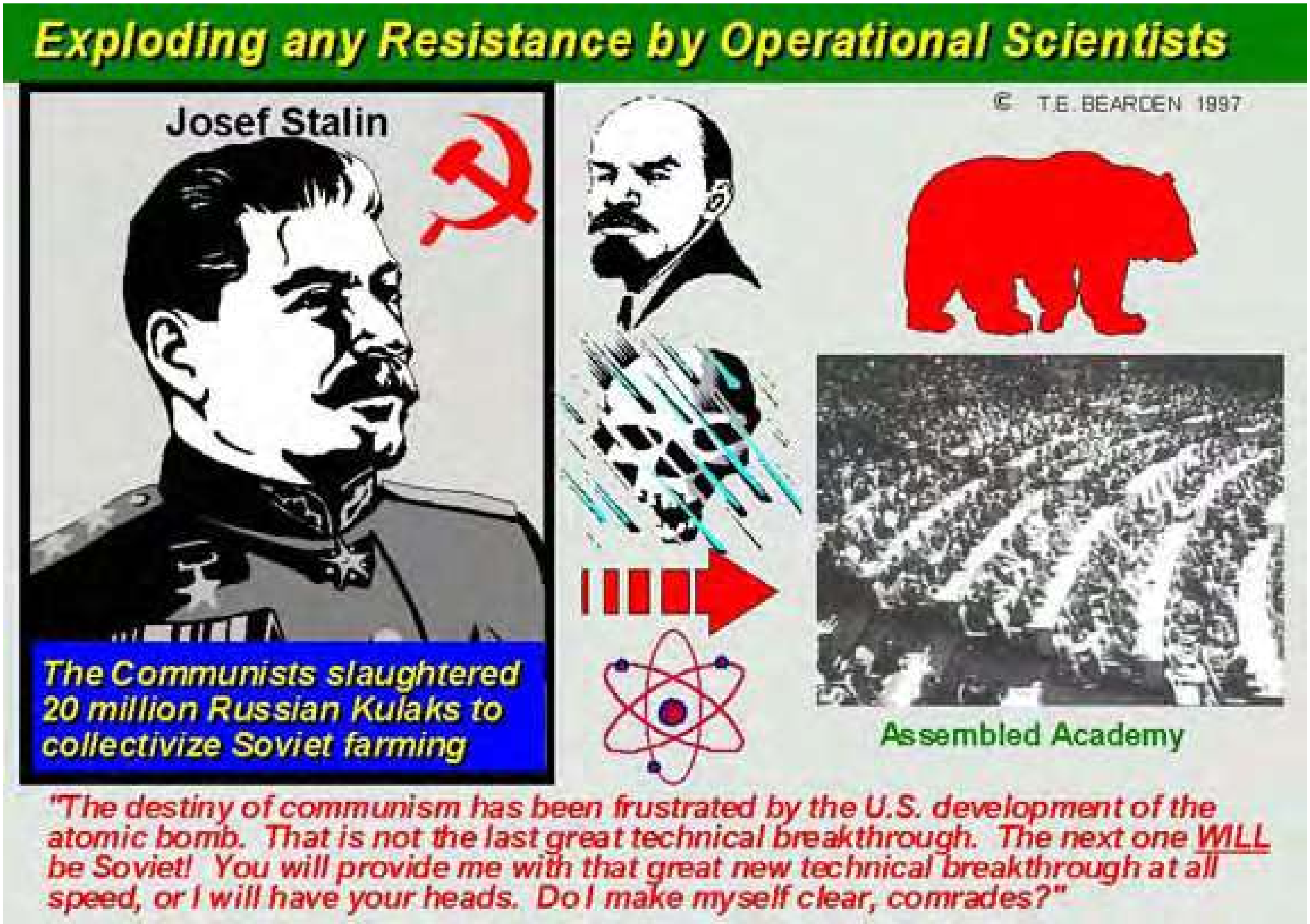

Exploding any Resistance by Operational Scientists

© T.E. BEARDEN 1997

The Communists slaughtered 20 million Russian Kulaks to collectivize Soviet farming

"The destiny of communism has been frustrated by the U.S. development of the atomic bomb. That is not the last great technical breakthrough. The next one WILL be Soviet! You will provide me with that great new technical breakthrough at all speed, or I will have your heads. Do I make myself clear, comrades?"

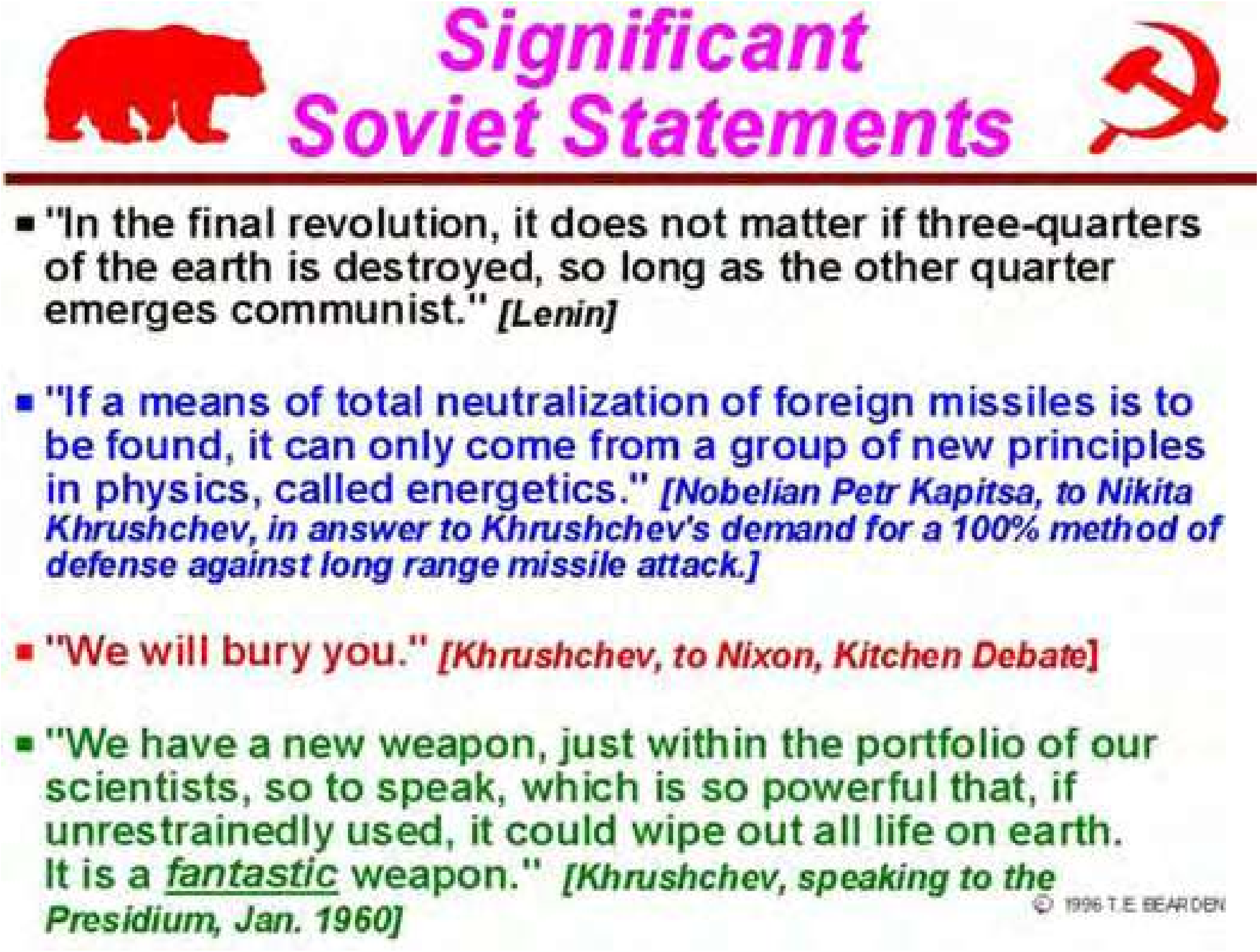

Significant Soviet Statements

- "In the final revolution, it does not matter if three-quarters of the earth is destroyed, so long as the other quarter emerges communist." *[Lenin]*

- "If a means of total neutralization of foreign missiles is to be found, it can only come from a group of new principles in physics, called energetics." ***[Nobelian Petr Kapitsa, to Nikita Khrushchev, in answer to Khrushchev's demand for a 100% method of defense against long range missile attack.]***

- "We will bury you." ***[Khrushchev, to Nixon, Kitchen Debate]***

- "We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that, if unrestrainedly used, it could wipe out all life on earth. It is a *fantastic* weapon." ***[Khrushchev, speaking to the Presidium, Jan. 1960]***

© 1996 T.E. BEARDEN

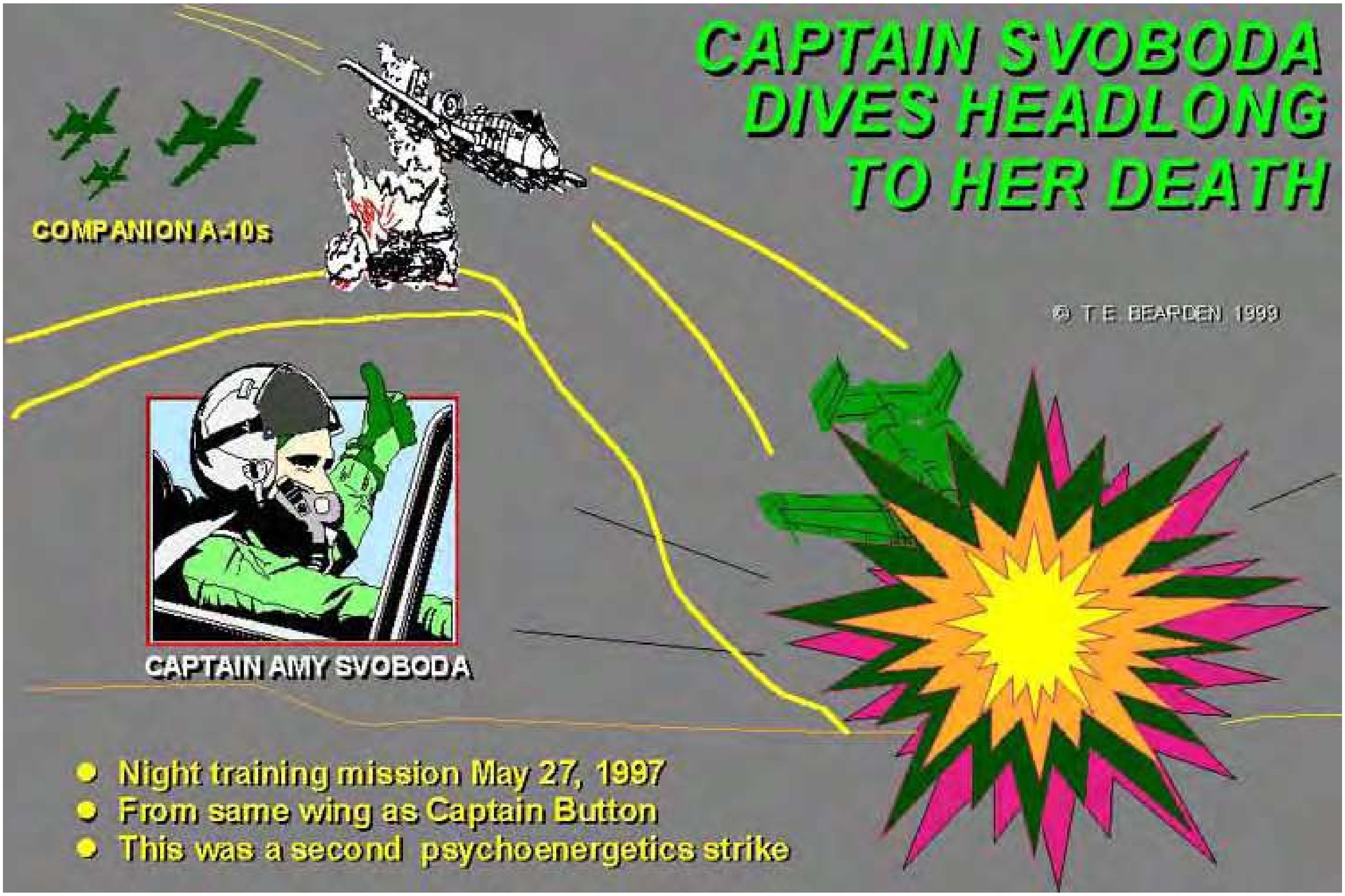

CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH

COMPANION A-10s
CAPTAIN AMY SVOBODA

© T.E. BEARDEN 1999

- Night training mission May 27, 1997
- From same wing as Captain Button
- This was a second psychoenergetics strike

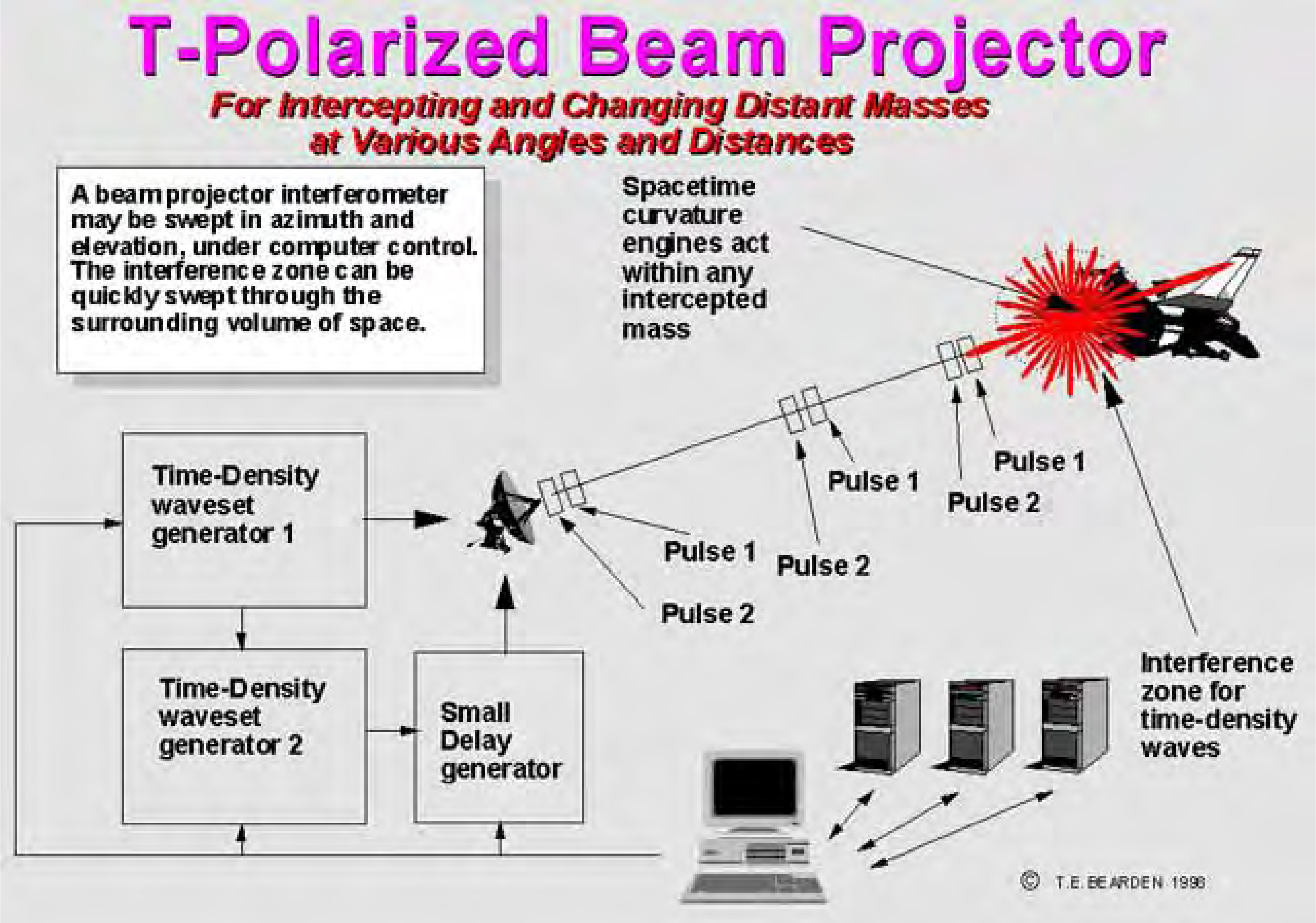

T-Polarized Beam Projector

For Intercepting and Changing Distant Masses at Various Angles and Distances

A beam projector interferometer may be swept in azimuth and elevation, under computer control. The interference zone can be quickly swept through the surrounding volume of space.

Spacetime curvature engines act within any intercepted mass
Time-Density waveset generator 1
Time-Density waveset generator 2
Small Delay generator
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Interference zone for time-density waves

© T.E. BEARDEN 1996

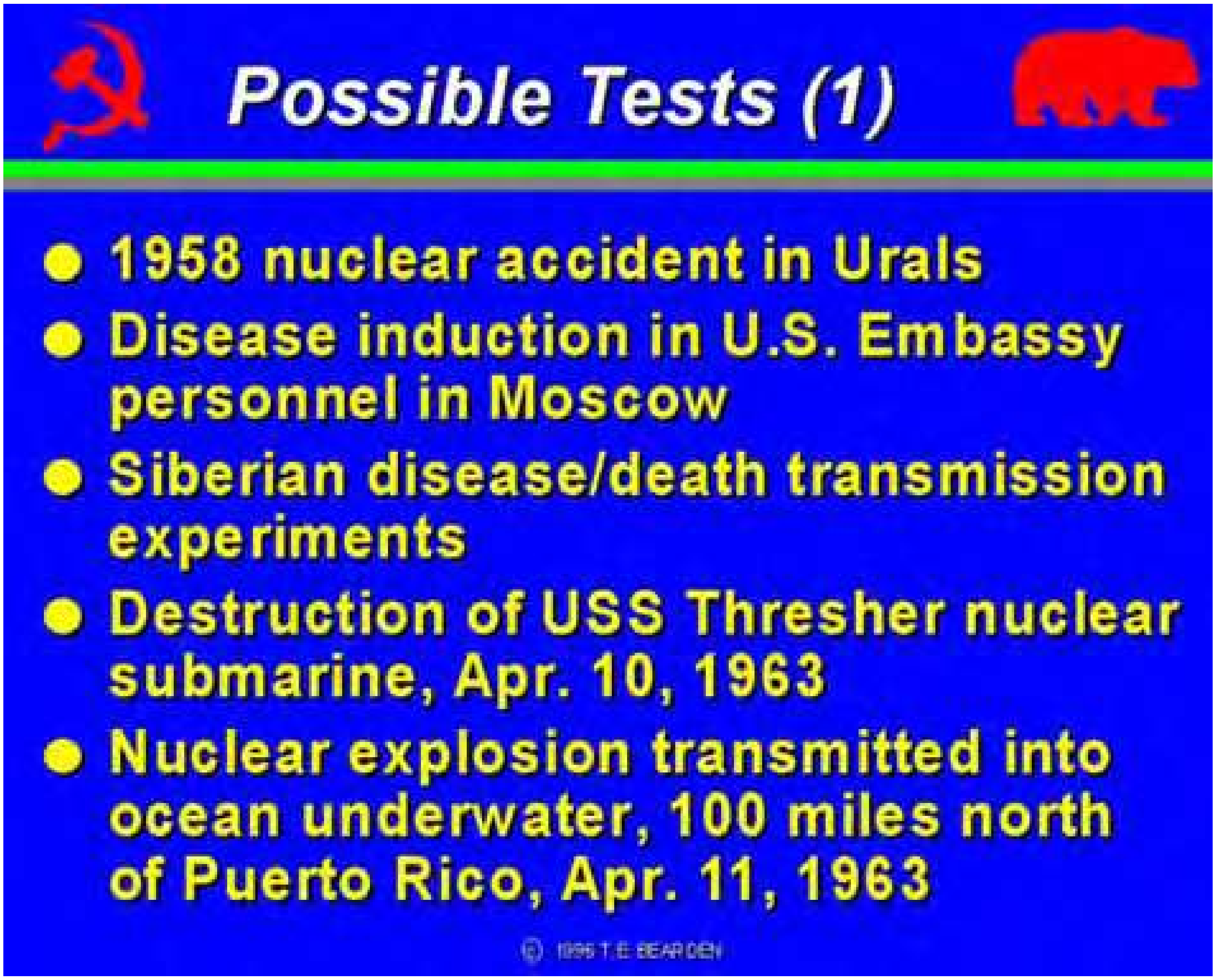

Possible Tests (1)

- 1958 nuclear accident in Urals
- Disease induction in U.S. Embassy personnel in Moscow
- Siberian disease/death transmission experiments
- Destruction of USS Thresher nuclear submarine, Apr. 10, 1963
- Nuclear explosion transmitted into ocean underwater, 100 miles north of Puerto Rico, Apr. 11, 1963

© 1996 T.E. BEARDEN

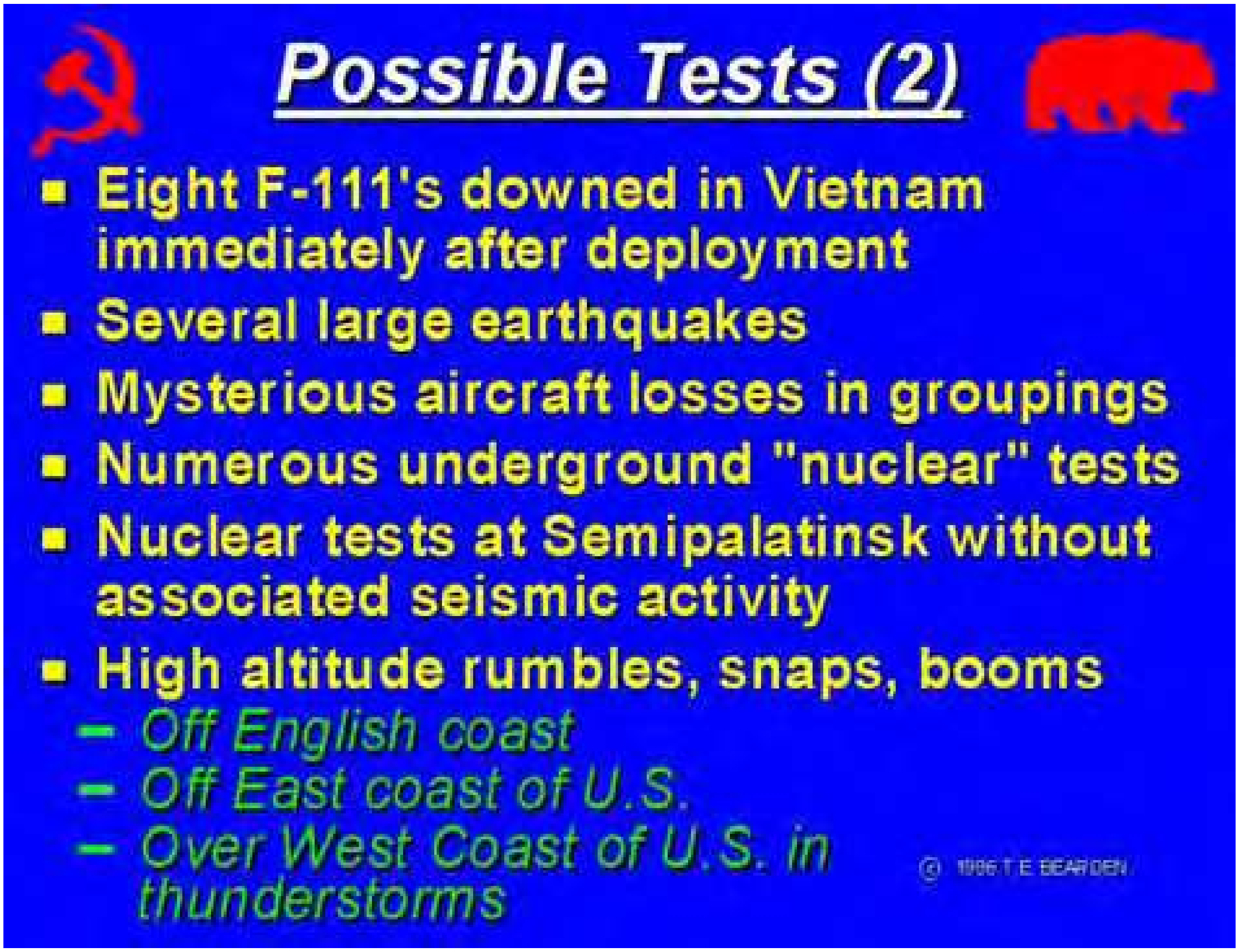

Possible Tests (2)

- Eight F-111's downed in Vietnam immediately after deployment
- Several large earthquakes
- Mysterious aircraft losses in groupings
- Numerous underground "nuclear" tests
- Nuclear tests at Semipalatinsk without associated seismic activity
- High altitude rumbles, snaps, booms
  - *Off English coast*
  - *Off East coast of U.S.*
  - *Over West Coast of U.S. in thunderstorms*

© 1996 T E BEARDEN

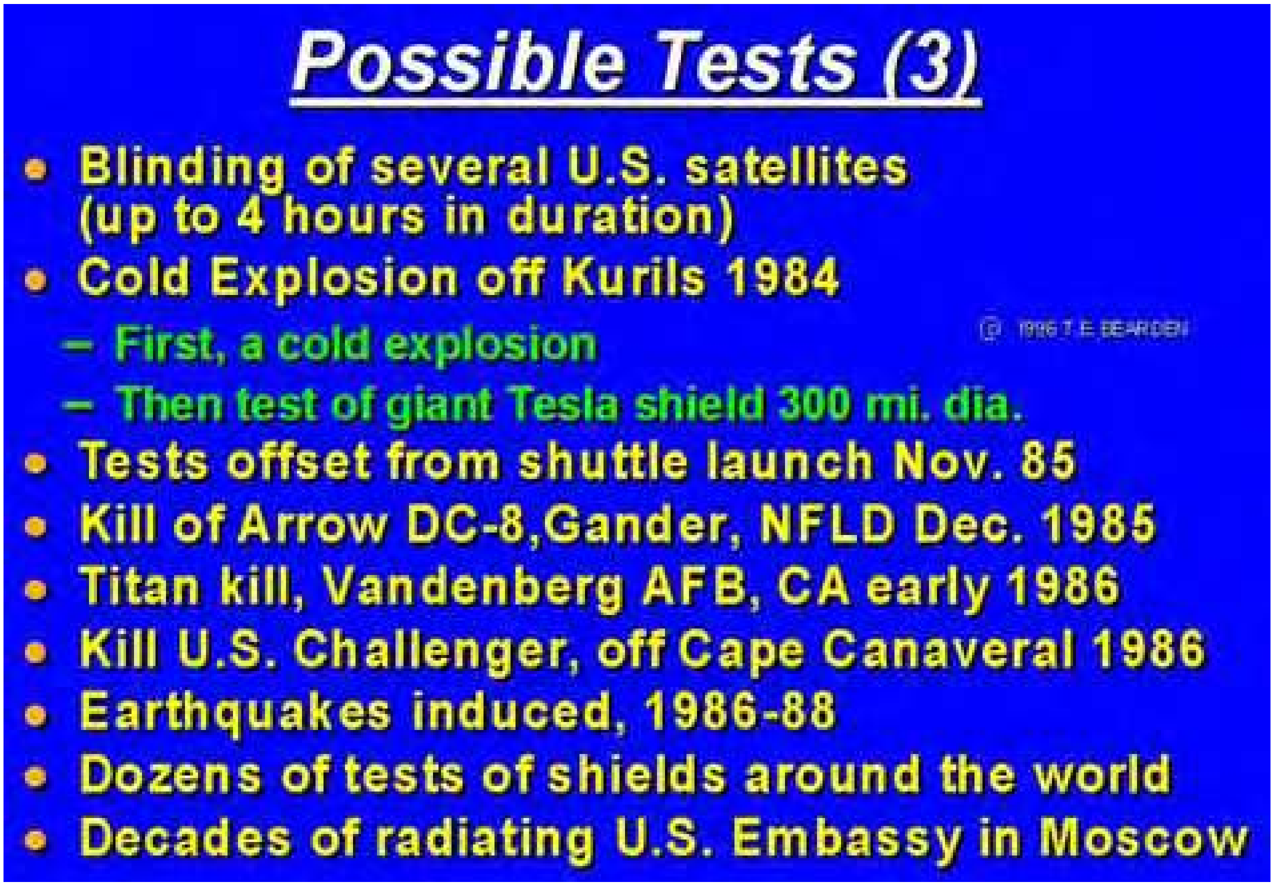

Possible Tests (3)

- Blinding of several U.S. satellites (up to 4 hours in duration)
- Cold Explosion off Kurils 1984
  - First, a cold explosion
  - Then test of giant Tesla shield 300 mi. dia.
- Tests offset from shuttle launch Nov. 85
- Kill of Arrow DC-8,Gander, NFLD Dec. 1985
- Titan kill, Vandenberg AFB, CA early 1986
- Kill U.S. Challenger, off Cape Canaveral 1986
- Earthquakes induced, 1986-88
- Dozens of tests of shields around the world
- Decades of radiating U.S. Embassy in Moscow

© 1996 T.E. BEARDEN

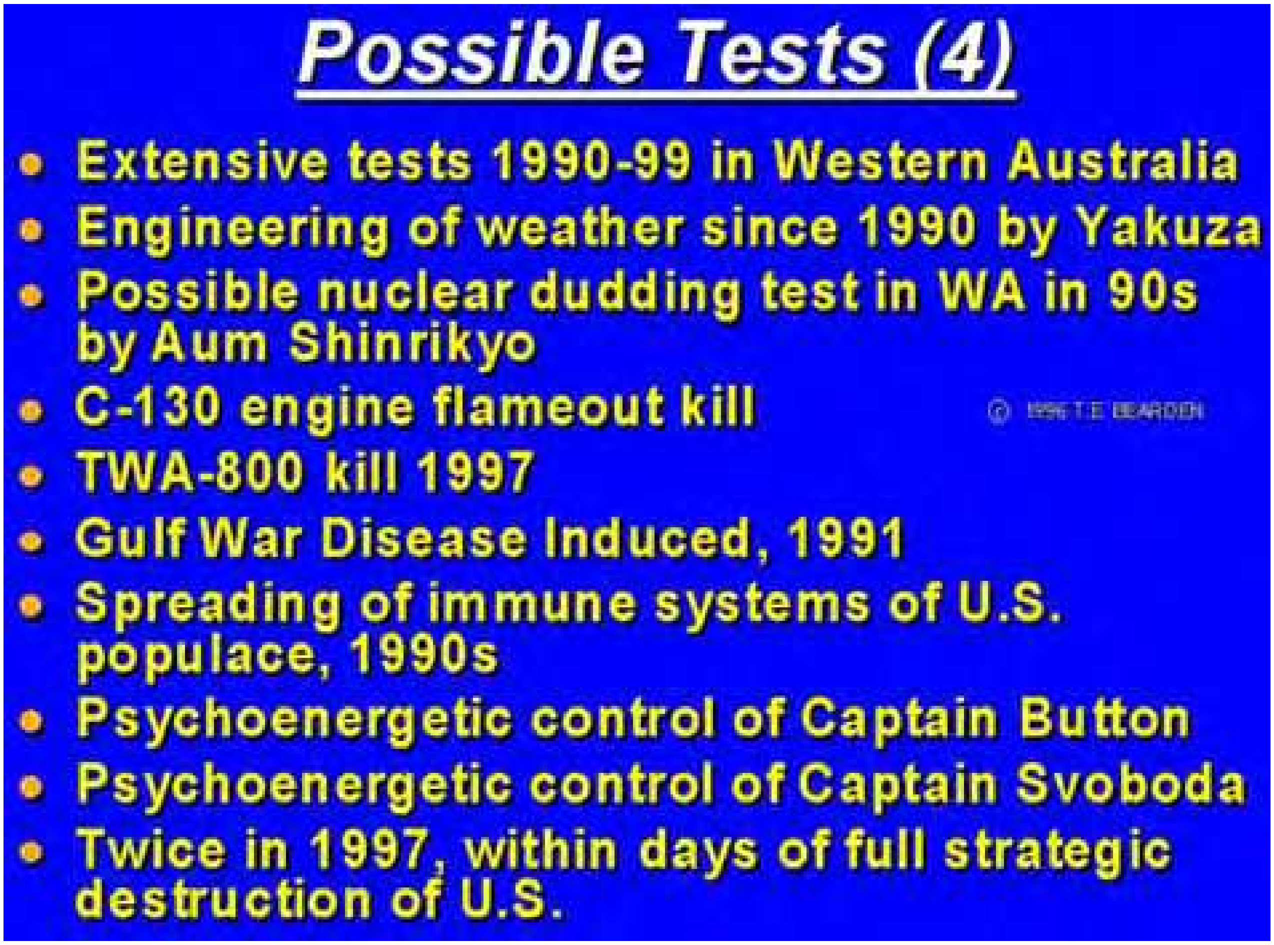

Possible Tests (4)

- Extensive tests 1990-99 in Western Australia
- Engineering of weather since 1990 by Yakuza
- Possible nuclear dudding test in WA in 90s by Aum Shinrikyo
- C-130 engine flameout kill
- TWA-800 kill 1997
- Gulf War Disease Induced, 1991
- Spreading of immune systems of U.S. populace, 1990s
- Psychoenergetic control of Captain Button
- Psychoenergetic control of Captain Svoboda
- Twice in 1997, within days of full strategic destruction of U.S.

© 1996 T.E. BEARDEN

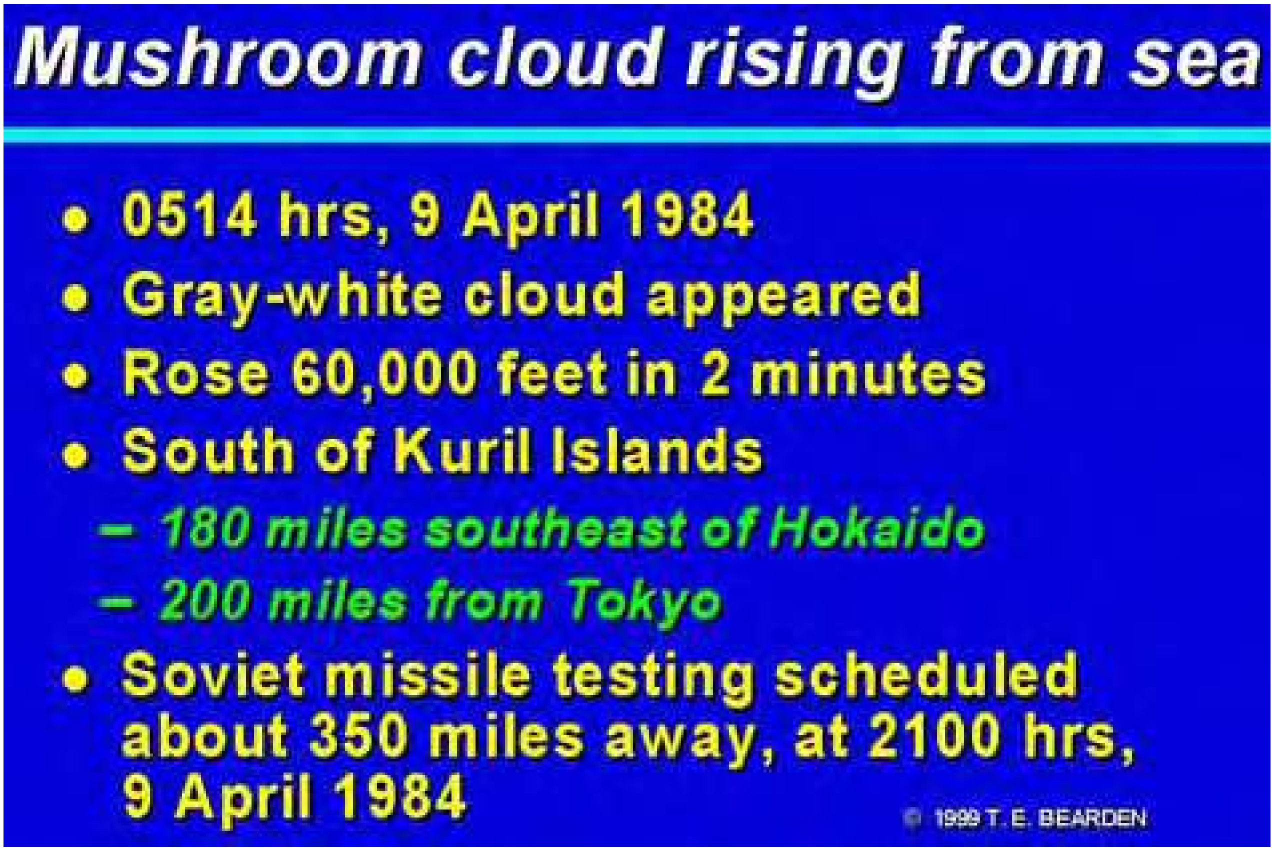

Mushroom cloud rising from sea

- 0514 hrs, 9 April 1984
- Gray-white cloud appeared
- Rose 60,000 feet in 2 minutes
- South of Kuril Islands
  - *180 miles southeast of Hokaido*
  - *200 miles from Tokyo*
- Soviet missile testing scheduled about 350 miles away, at 2100 hrs, 9 April 1984

© 1999 T. E. BEARDEN

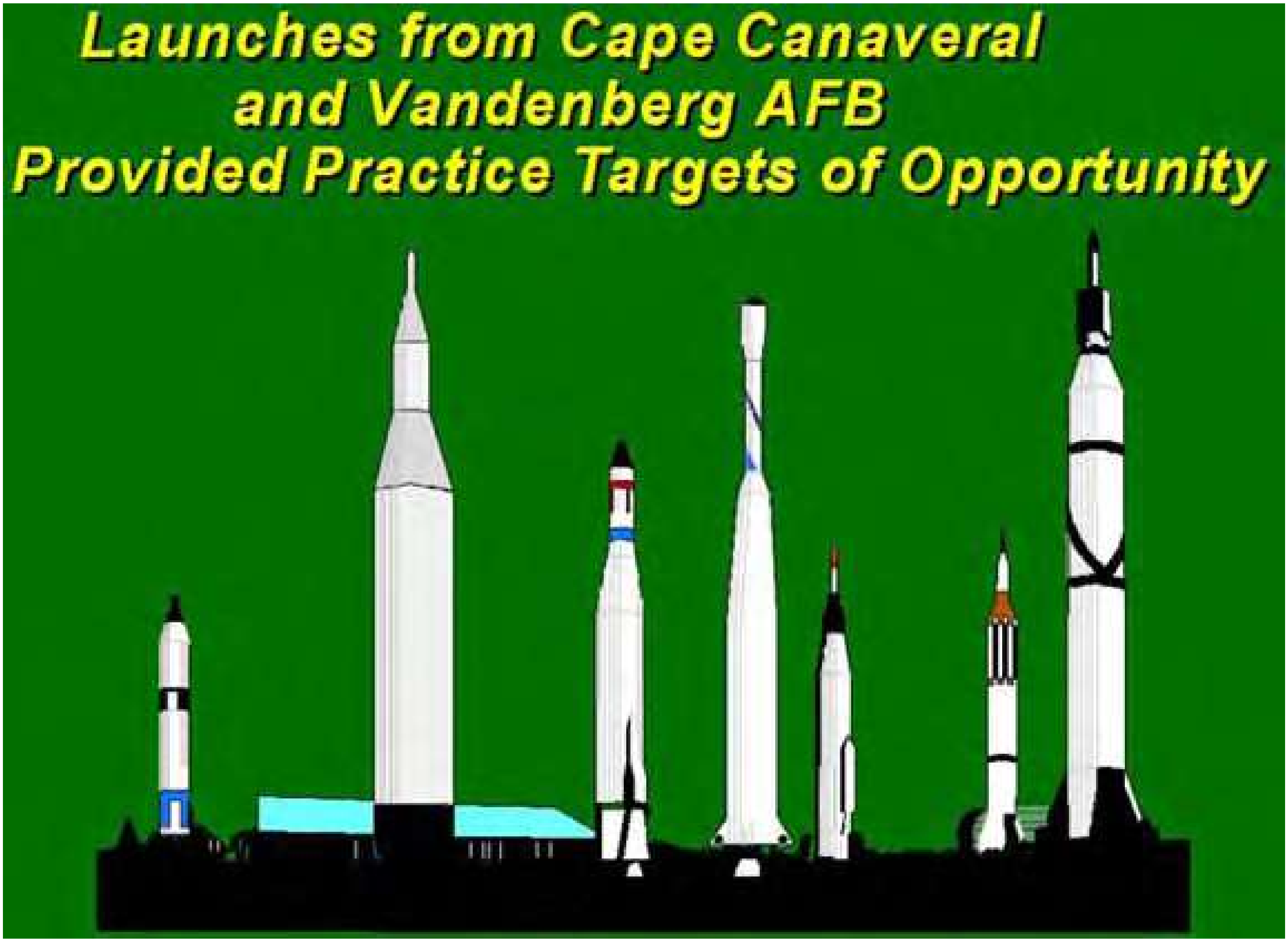
Launches from Cape Canaveral
and Vandenberg AFB
Provided Practice Targets of Opportunity

The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth

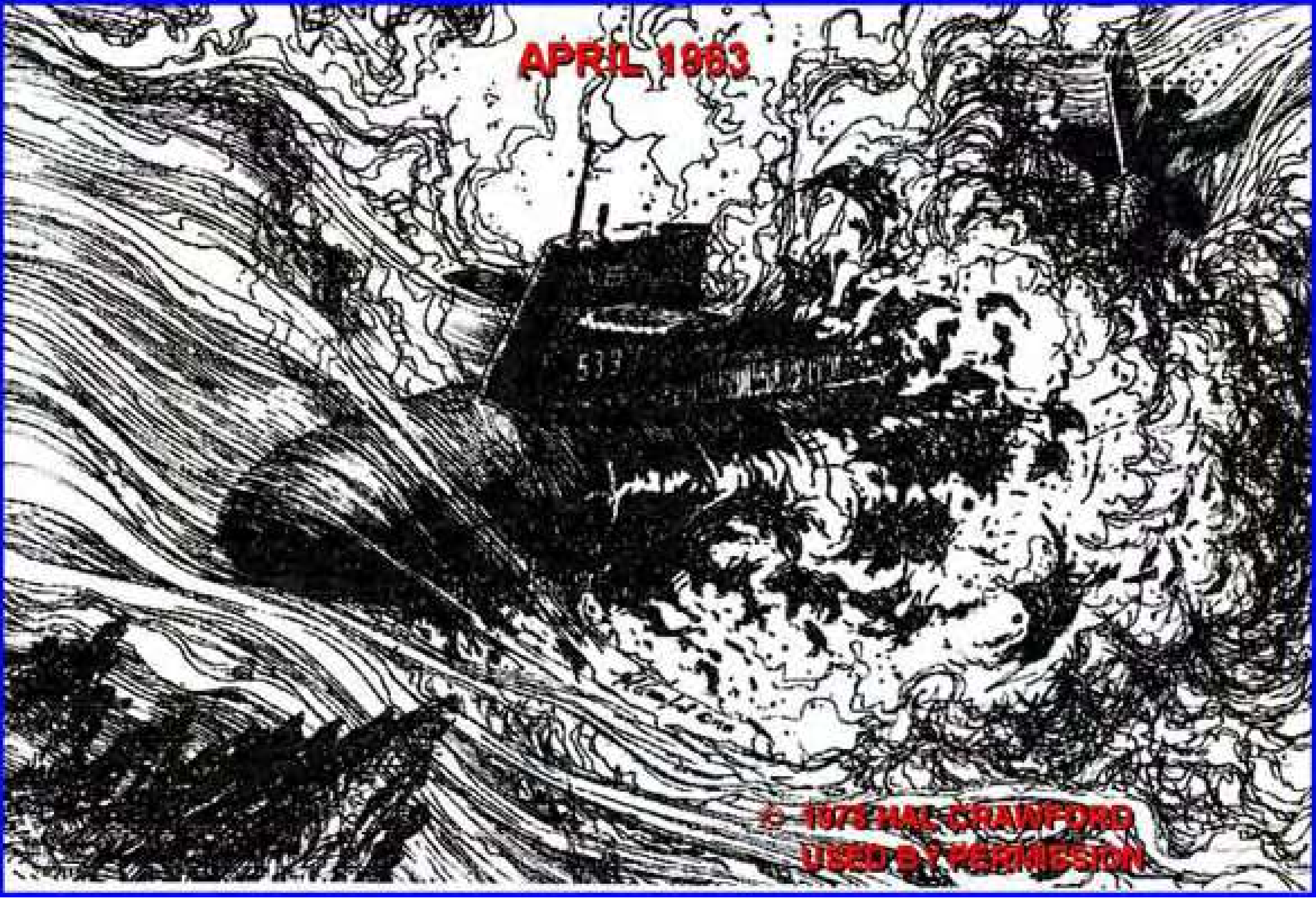
APRIL 1963
593
© 1973 HAL CRAWFORD
USED BY PERMISSION

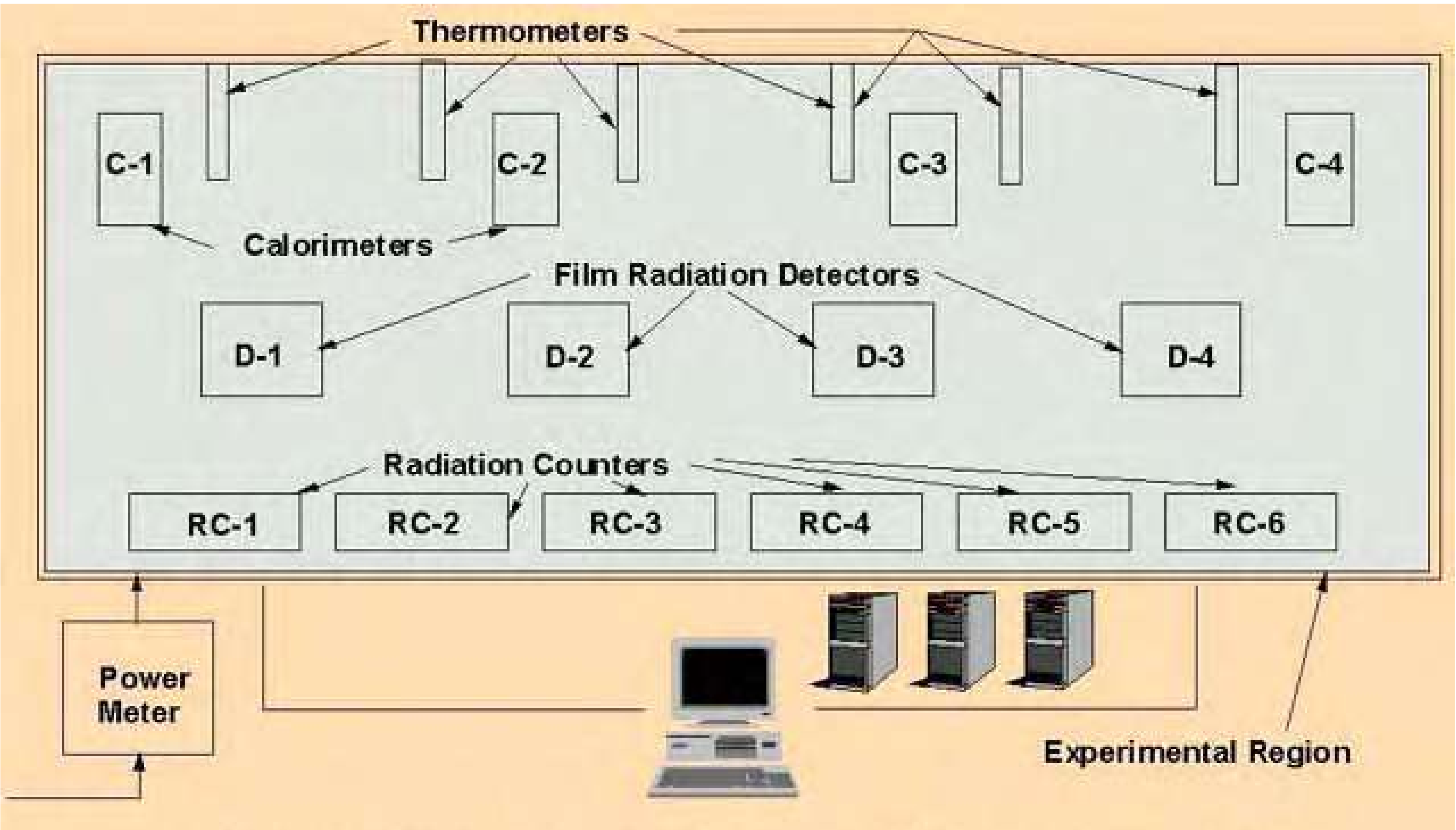
Thermometers
C-1
C-2
C-3
C-4
Calorimeters
Film Radiation Detectors
D-1
D-2
D-3
D-4
Radiation Counters
RC-1
RC-2
RC-3
RC-4
RC-5
RC-6
Power Meter
Experimental Region

Figure 32. Embodiment of an instrumental array sufficient to discriminate when time-density charging has occurred, and that transduction of time-charge decay is occurring.

© T.E. BEARDEN 1998

Time-Polarized EM Wave interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone

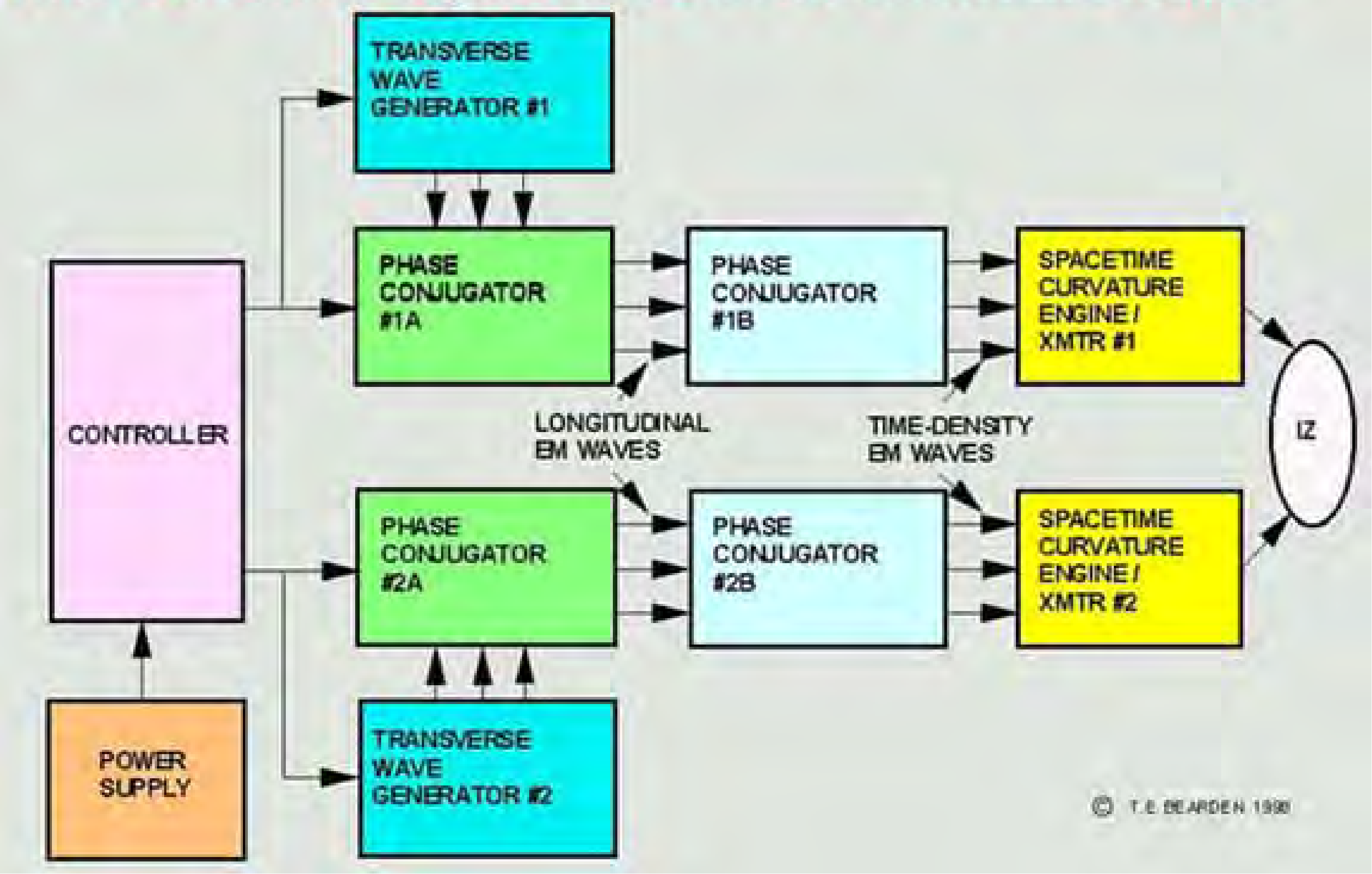
TRANSVERSE WAVE GENERATOR #1
CONTROLLER
PHASE CONJUGATOR #1A
PHASE CONJUGATOR #1B
SPACETIME CURVATURE ENGINE / XMTR #1
LONGITUDINAL EM WAVES
TIME-DENSITY EM WAVES
IZ
PHASE CONJUGATOR #2A
PHASE CONJUGATOR #2B
SPACETIME CURVATURE ENGINE / XMTR #2
POWER SUPPLY
TRANSVERSE WAVE GENERATOR #2
© T.E. BEARDEN 1990

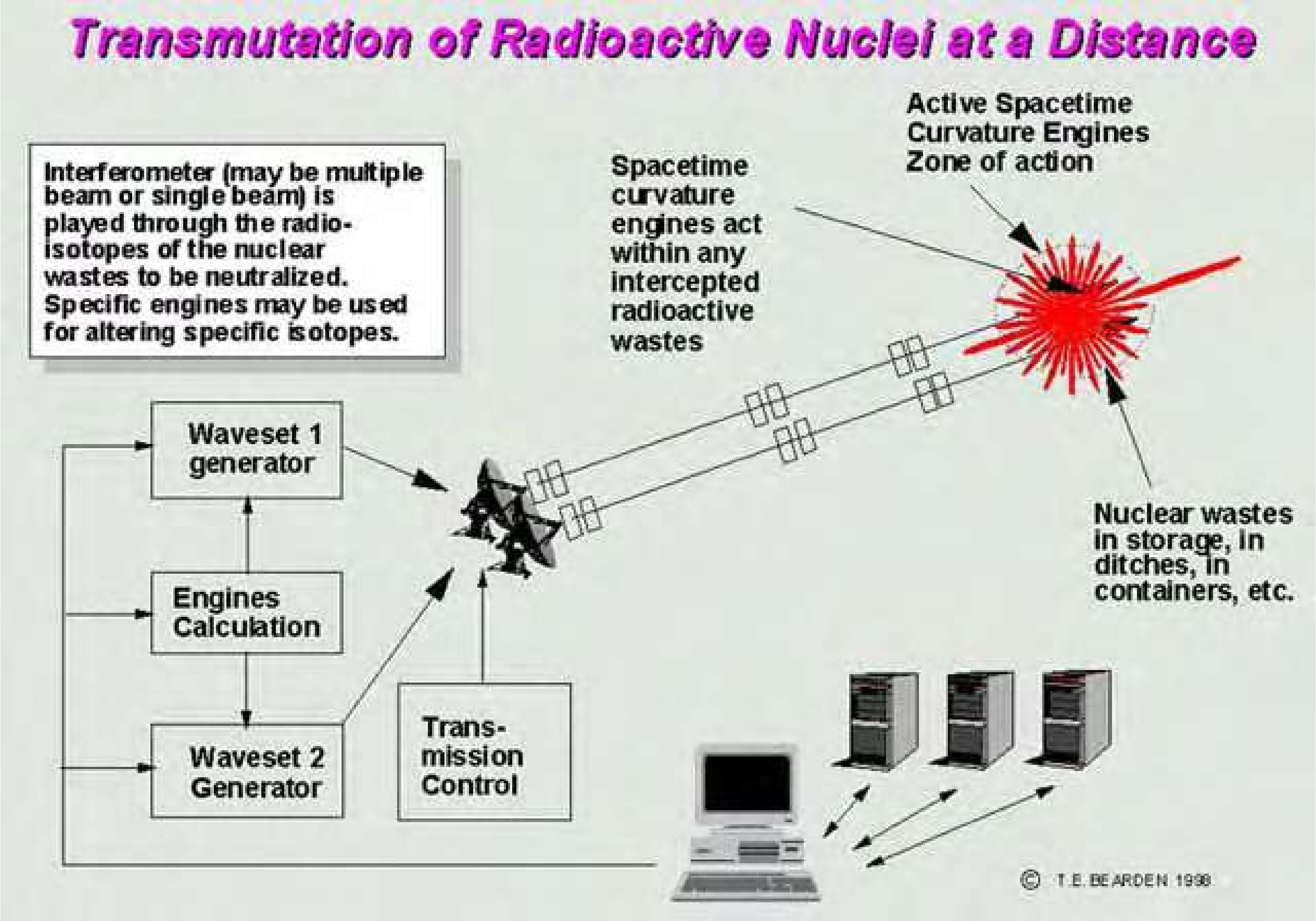

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radioisotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes

Active Spacetime Curvature Engines Zone of action
Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control
Nuclear wastes in storage, in ditches, in containers, etc.

© T.E. BEARDEN 1998

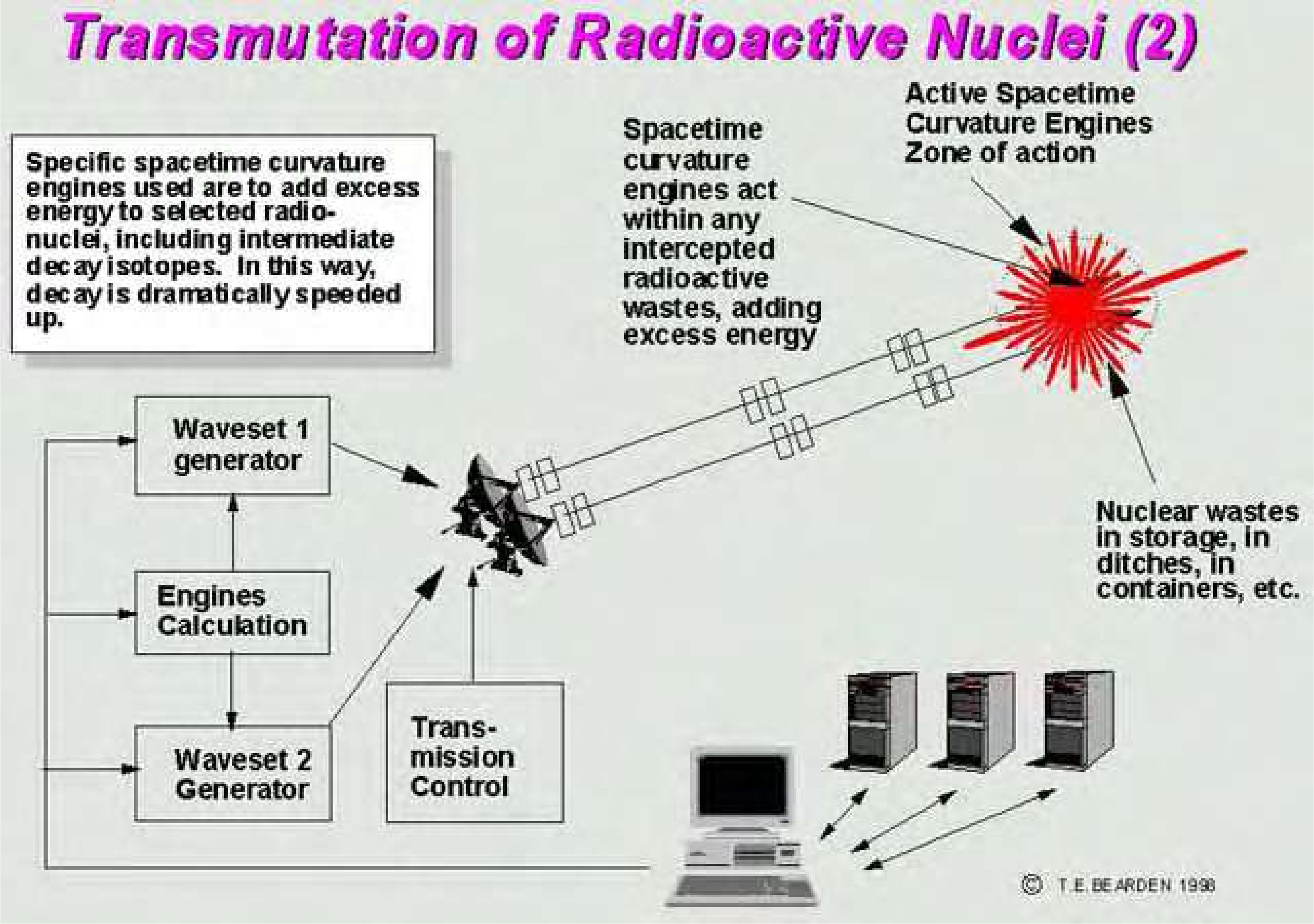

Transmutation of Radioactive Nuclei (2)

Specific spacetime curvature engines used are to add excess energy to selected radionuclei, including intermediate decay isotopes. In this way, decay is dramatically speeded up.

Spacetime curvature engines act within any intercepted radioactive wastes, adding excess energy

Active Spacetime Curvature Engines Zone of action
Nuclear wastes in storage, in ditches, in containers, etc.
Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control
© T.E. BEARDEN 1998

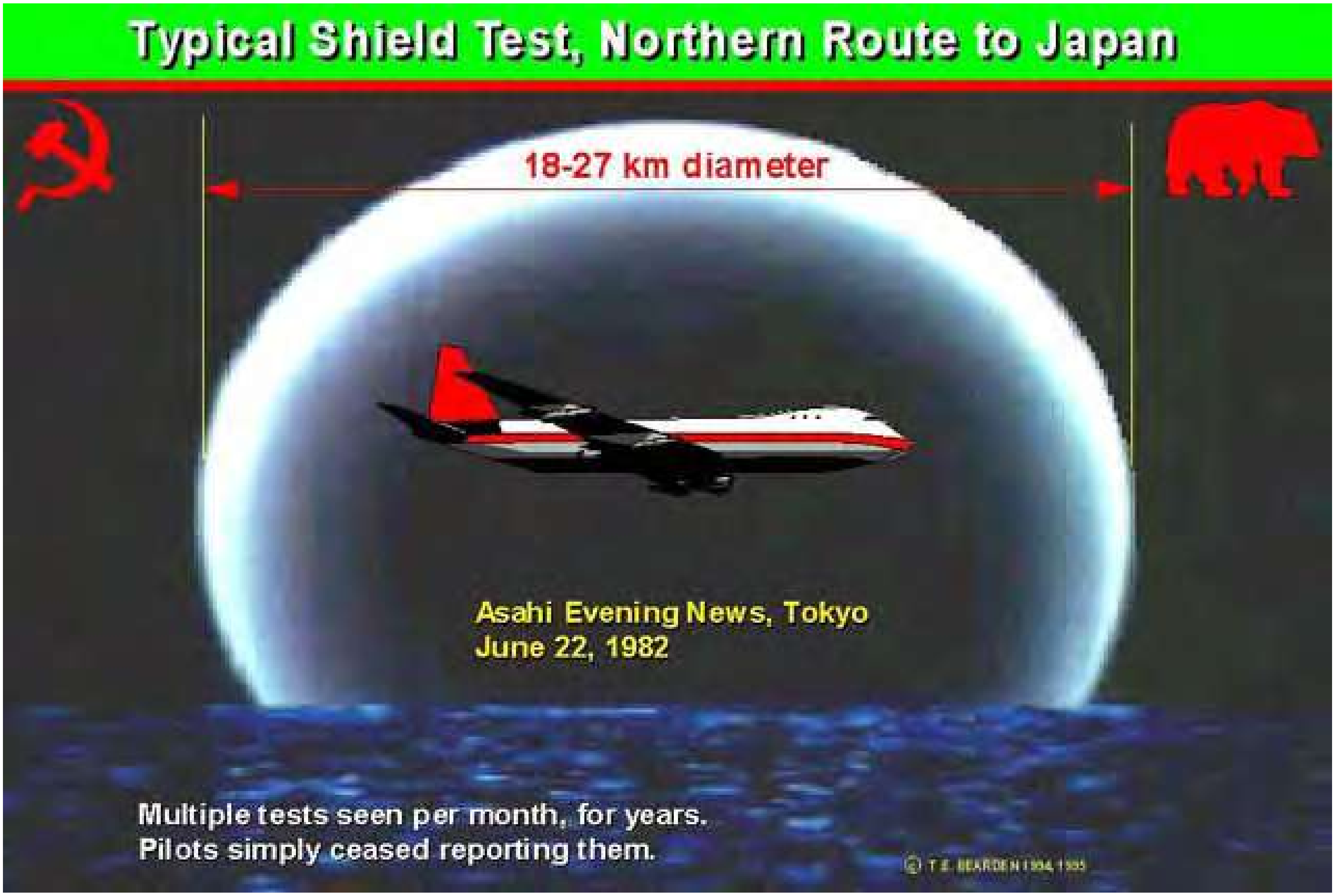
Typical Shield Test, Northern Route to Japan
18-27 km diameter
Asahi Evening News, Tokyo
June 22, 1982
Multiple tests seen per month, for years.
Pilots simply ceased reporting them.
© T.E. BEARDEN 1994, 1995

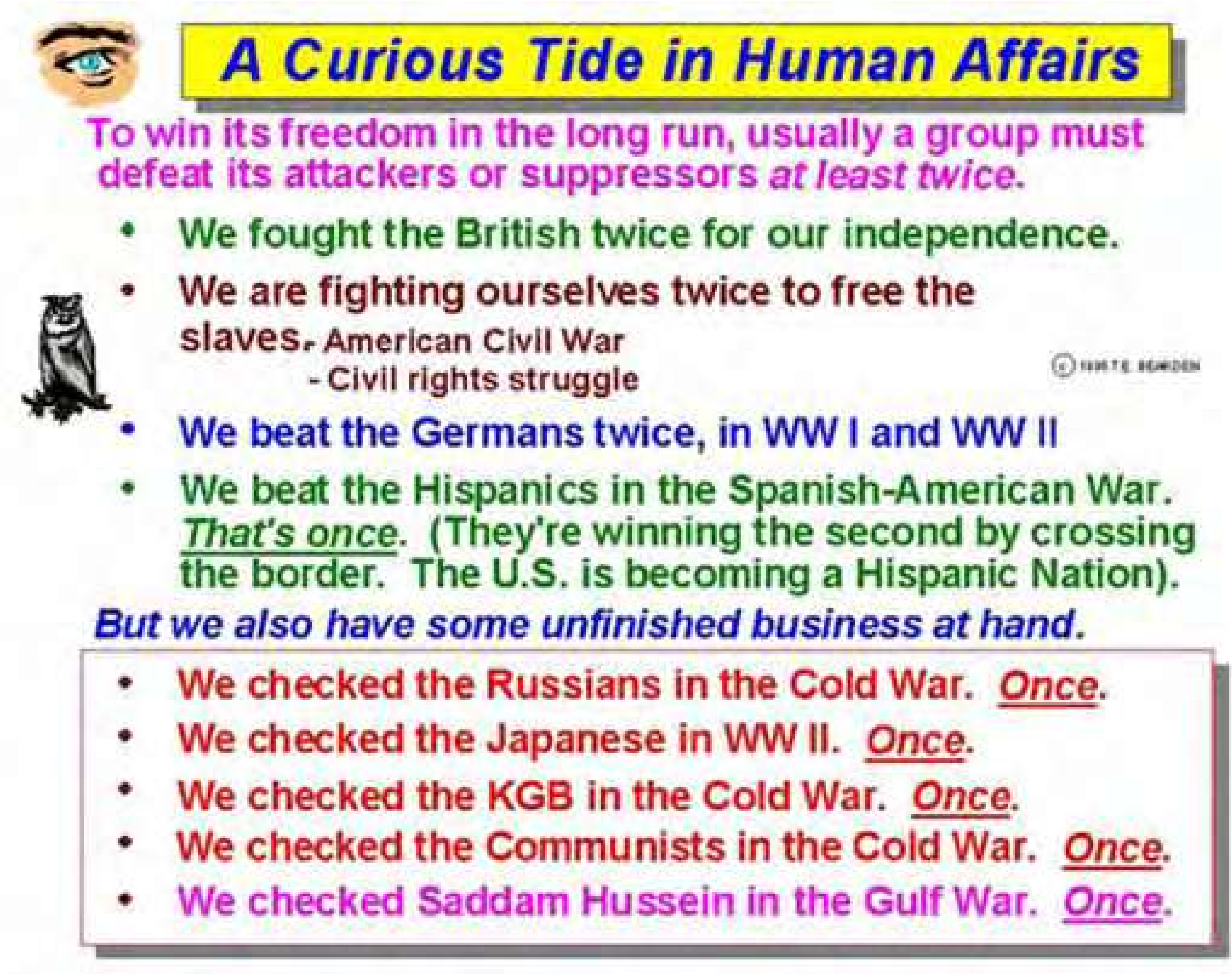

A Curious Tide in Human Affairs

To win its freedom in the long run, usually a group must defeat its attackers or suppressors *at least twice*.

- We fought the British twice for our independence.
- We are fighting ourselves twice to free the slaves. - American Civil War
  - Civil rights struggle
- We beat the Germans twice, in WW I and WW II
- We beat the Hispanics in the Spanish-American War. *That's once*. (They're winning the second by crossing the border. The U.S. is becoming a Hispanic Nation).

But we also have some unfinished business at hand.

- We checked the Russians in the Cold War. *Once*.
- We checked the Japanese in WW II. *Once*.
- We checked the KGB in the Cold War. *Once*.
- We checked the Communists in the Cold War. *Once*.
- We checked Saddam Hussein in the Gulf War. *Once*.

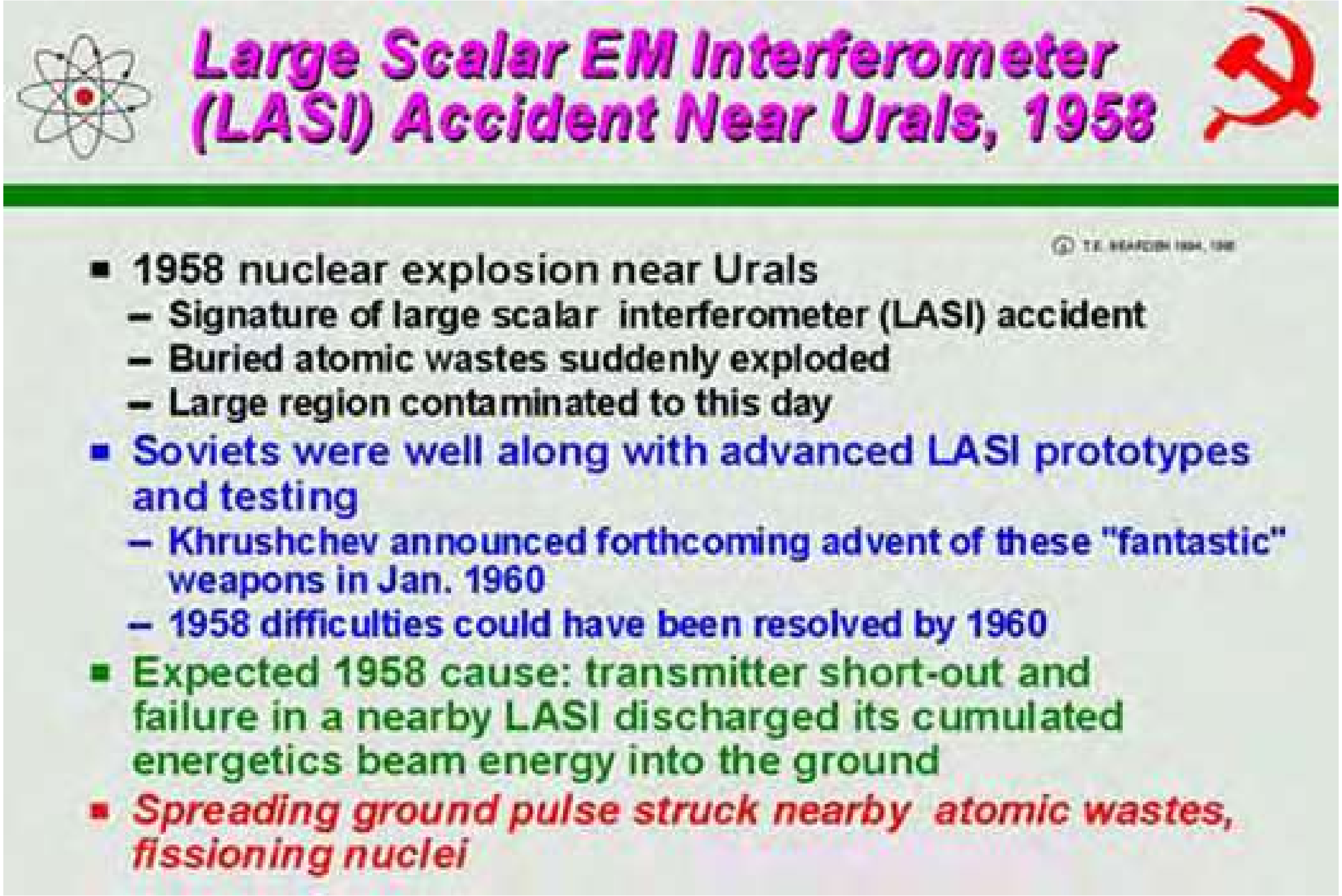

Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958

- 1958 nuclear explosion near Urals
  - Signature of large scalar interferometer (LASI) accident
  - Buried atomic wastes suddenly exploded
  - Large region contaminated to this day
- Soviets were well along with advanced LASI prototypes and testing
  - Khrushchev announced forthcoming advent of these "fantastic" weapons in Jan. 1960
  - 1958 difficulties could have been resolved by 1960
- Expected 1958 cause: transmitter short-out and failure in a nearby LASI discharged its cumulated energetics beam energy into the ground
- *Spreading ground pulse struck nearby atomic wastes, fissioning nuclei*

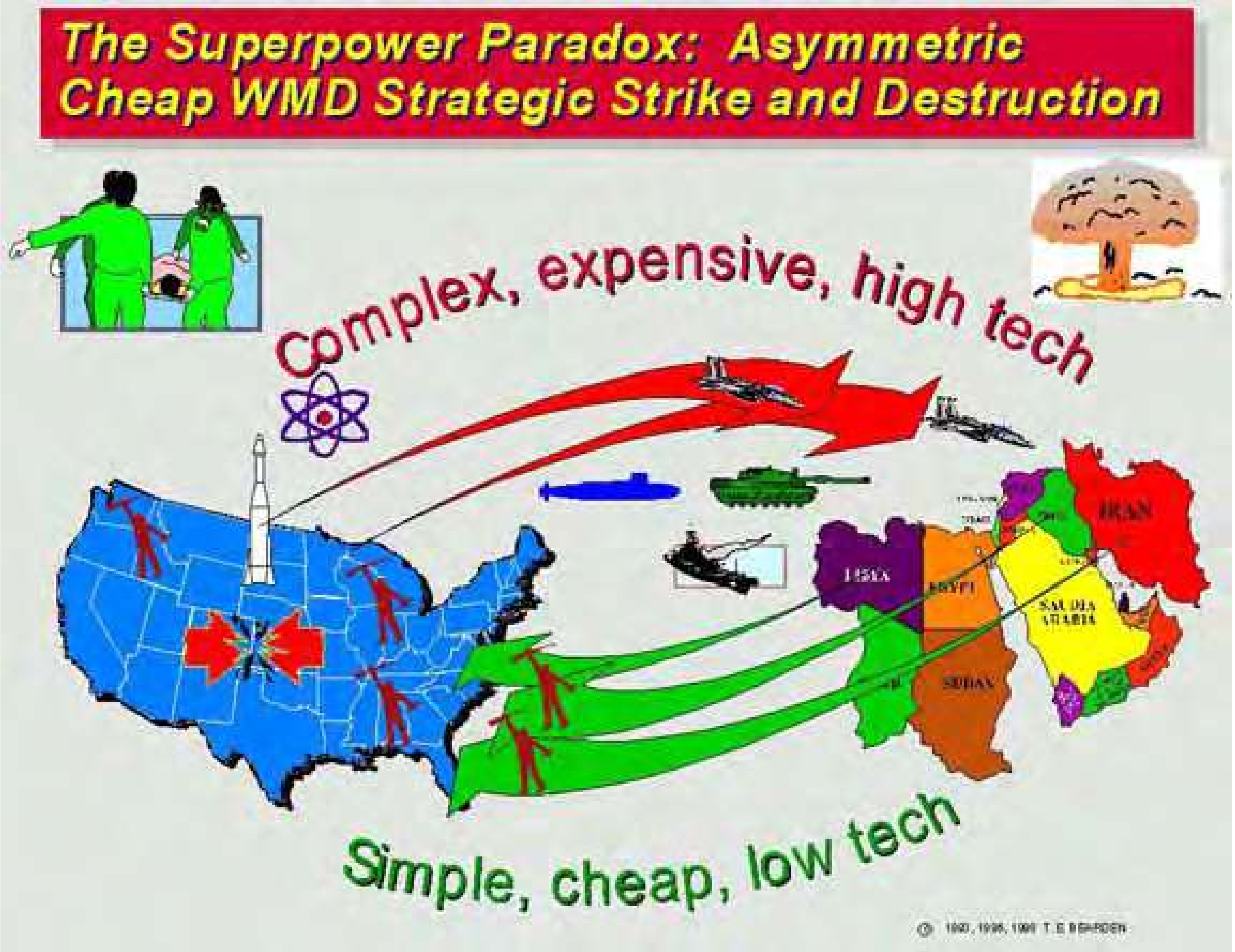

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

Simple, cheap, low tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN
© 1993, 1996, 1997 T. E. BEARDEN

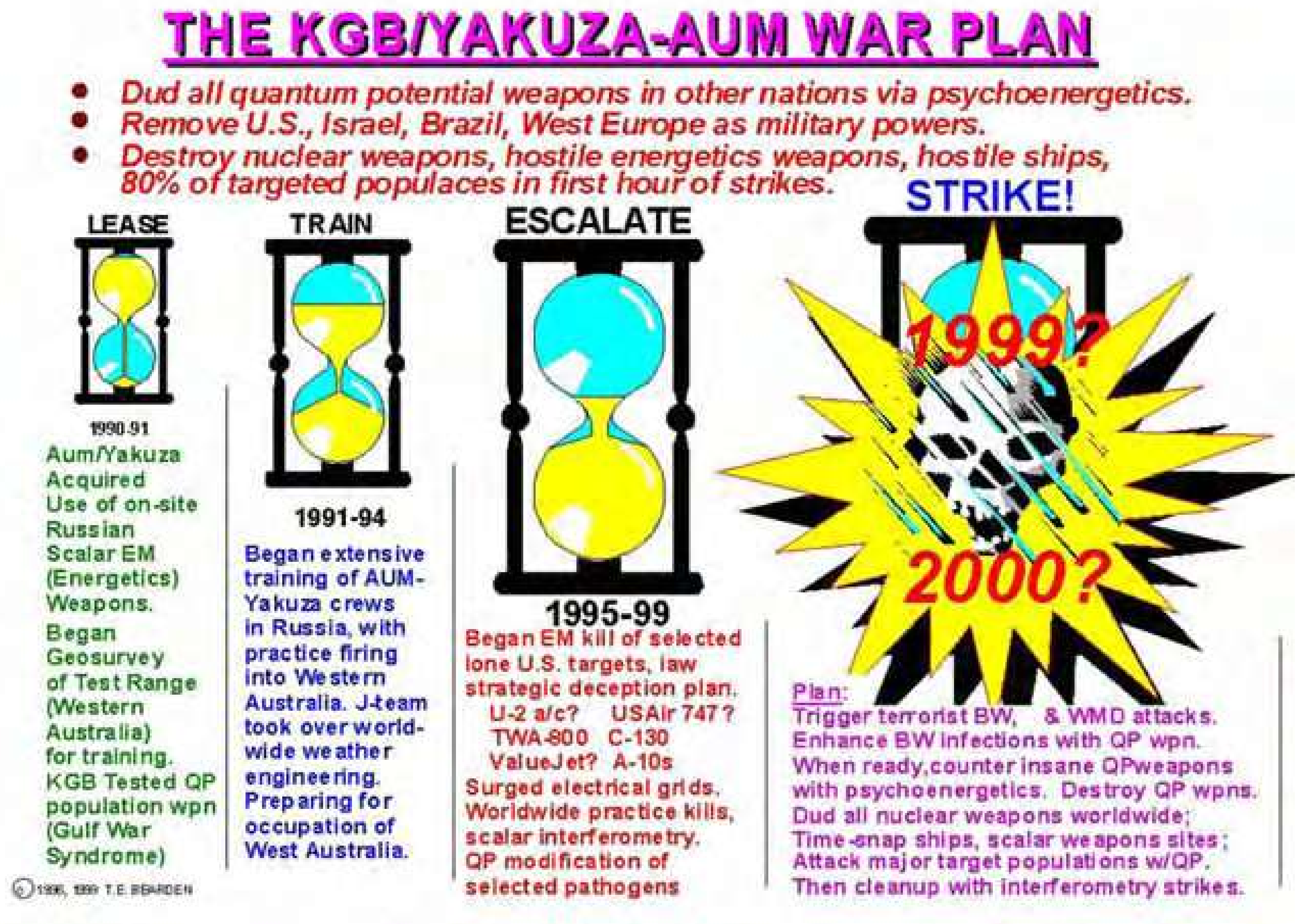

THE KGB/YAKUZA-AUM WAR PLAN

- *Dud all quantum potential weapons in other nations via psychoenergetics.*
- *Remove U.S., Israel, Brazil, West Europe as military powers.*
- *Destroy nuclear weapons, hostile energetics weapons, hostile ships, 80% of targeted populaces in first hour of strikes.*

LEASE

1990-91

Aum/Yakuza Acquired Use of on-site Russian Scalar EM (Energetics) Weapons.

Began Geosurvey of Test Range (Western Australia) for training. KGB Tested QP population wpn (Gulf War Syndrome)

TRAIN

1991-94

Began extensive training of AUM-Yakuza crews in Russia, with practice firing into Western Australia. J-team took over world-wide weather engineering. Preparing for occupation of West Australia.

ESCALATE

1995-99

Began EM kill of selected lone U.S. targets, law strategic deception plan.
U-2 a/c? USAir 747?
TWA-800 C-130
ValueJet? A-10s
Surged electrical grids. Worldwide practice kills, scalar interferometry. QP modification of selected pathogens

STRIKE!

Plan:
Trigger terrorist BW, & WMD attacks.
Enhance BW infections with QP wpn.
When ready, counter insane QPweapons with psychoenergetics. Destroy QP wpns.
Dud all nuclear weapons worldwide;
Time-snap ships, scalar weapons sites;
Attack major target populations w/QP.
Then cleanup with interferometry strikes.

© 1996, 1999 T.E. BEARDEN

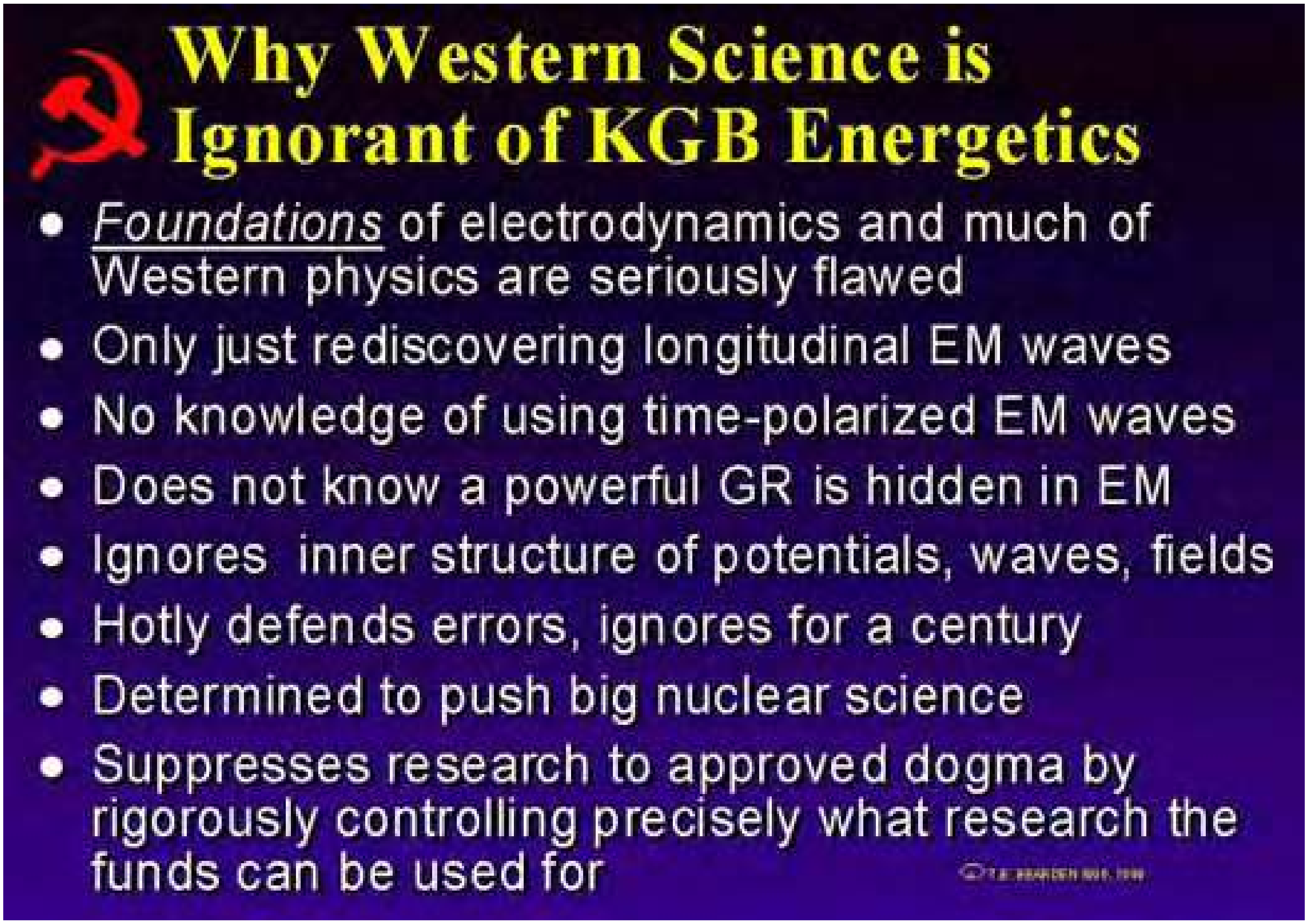

Why Western Science is Ignorant of KGB Energetics

- *Foundations* of electrodynamics and much of Western physics are seriously flawed
- Only just rediscovering longitudinal EM waves
- No knowledge of using time-polarized EM waves
- Does not know a powerful GR is hidden in EM
- Ignores inner structure of potentials, waves, fields
- Hotly defends errors, ignores for a century
- Determined to push big nuclear science
- Suppresses research to approved dogma by rigorously controlling precisely what research the funds can be used for

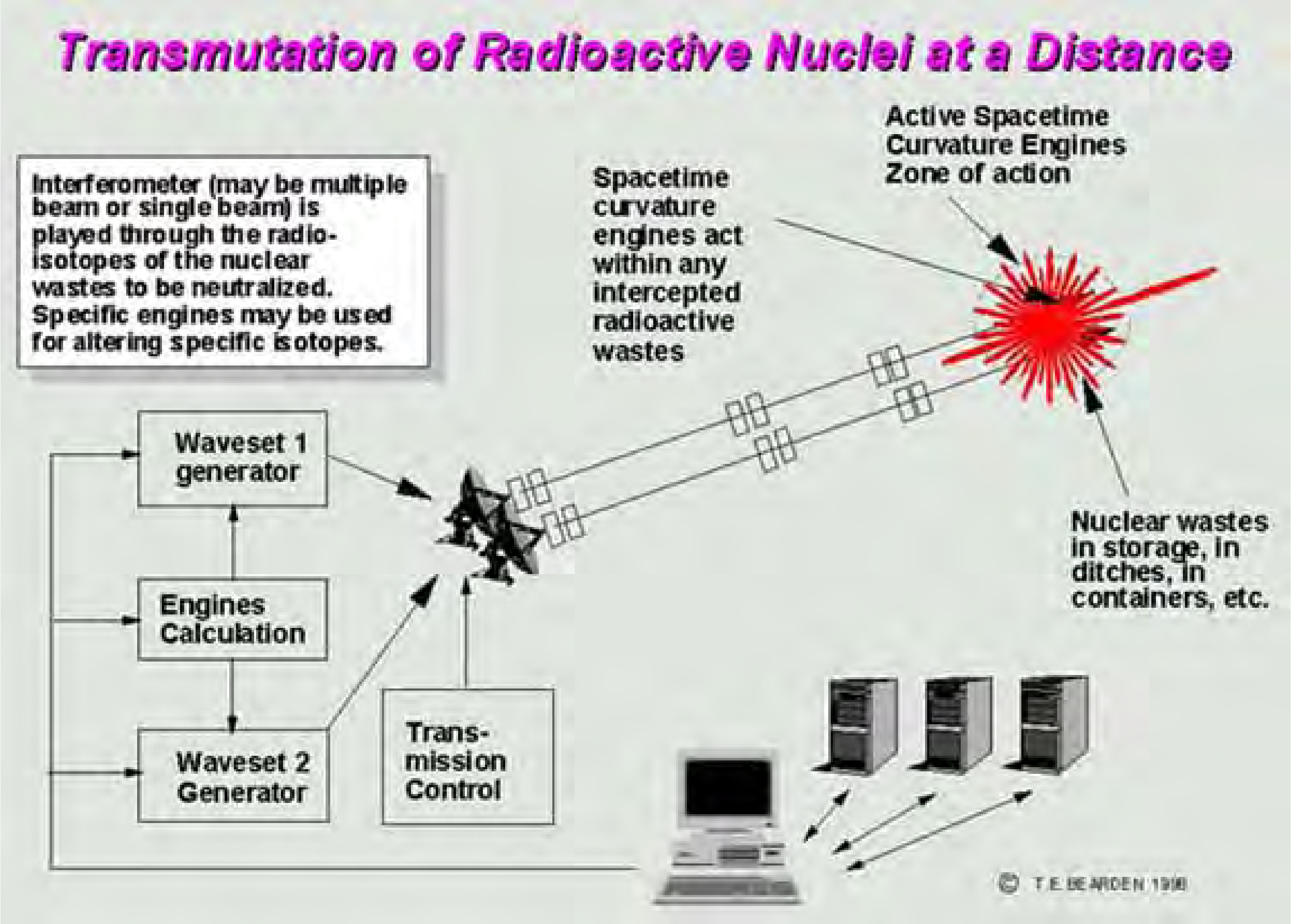

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radioisotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes

Active Spacetime Curvature Engines Zone of action

Nuclear wastes in storage, in ditches, in containers, etc.

Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control
© T.E. BEARDEN 1998

THE YAKUZA

(JAPANESE MAFIA)

To do business with Japan is to do business with the Yakuza

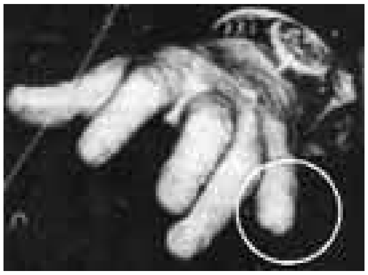

© 1995, 1996, 2001 T.E. BEARDEN

The Last Samurai, Defending Bushido

- **3 Major Yakuza Groups, In Coalition.**
- **Penetrate Japanese Government.**
- **Open Offices, Signs on Street.**
- **Public Romanticizes Yakuza as Last Japanese Samurai**
- **In Large High-Tech Corporations.**
- **In Control of Japanese Banks.**
- **Japanese Real-Estate Rip-off.**
- **$500 - 1,000 Billion per Year Income.**
- **Rapidly Spreading in U.S.**
- **Uniting Mafia Groups Worldwide.**
- **Coalition with Aum Shinrikyo.**
- **Leased Russian Energetics Weapons.**
- **Crews On-site in Russia, Trained.**
- **Practice Firings into Western Australia for Three Years. (Dozens of Shots per Month)**
- **Now in Escalation Phase, Killing U.S. Targets.**
- **All-out Strike on U.S. ready 1999-2000.**
- **Now held in check by KGB, dead-man fuzed by small nation friendly to U.S.**

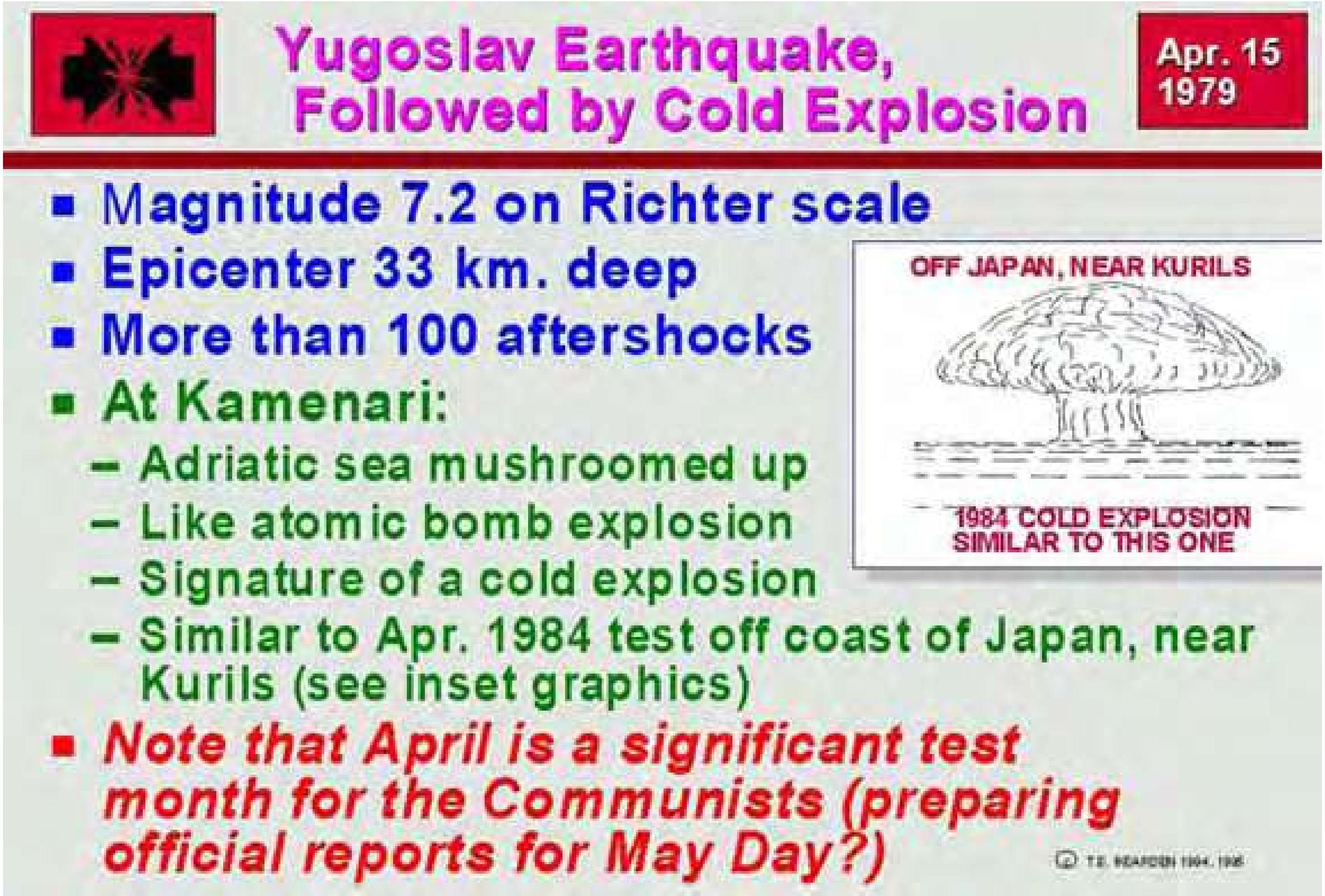

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- *Note that April is a significant test month for the Communists (preparing official reports for May Day?)*

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

Briefing to
Request for Information Symposium

14 February 2000

Air Force Research Laboratory

Directed Energy Directorate
AFRL/DE

Kirtland AFB, New Mexico

Colonel Doug Beason
Deputy Director

“VICTORY SMILES UPON THOSE WHO ANTICIPATE THE CHANGES IN THE CHARACTER OF WAR”

Guilio Douhet

Directed Energy Directorate

A paradigm shift in weapon system technology for the 21st Century.

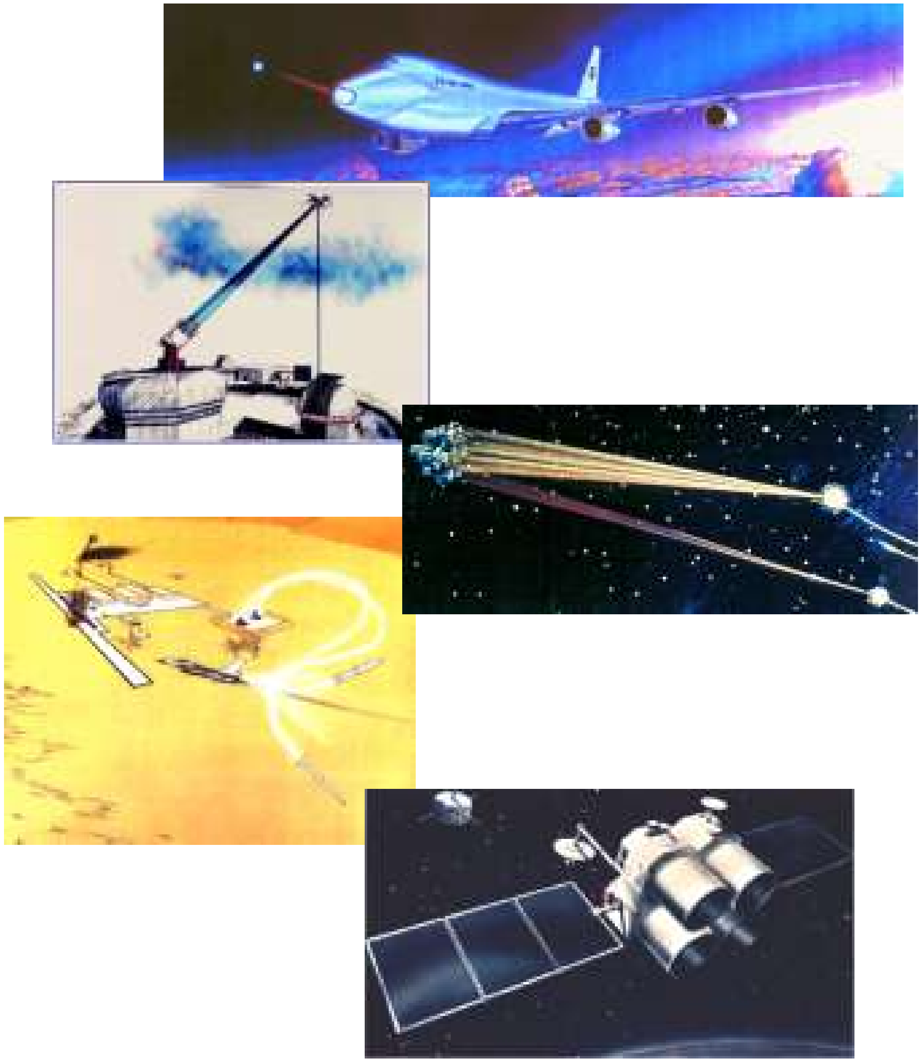

Our Mission

- ***To support*** user needs for DEW applications and to address mission area deficiencies
- ***To exploit*** the relevance of Directed Energy technology to Air Force and DoD needs
- ***To foster*** user awareness of the potential of Directed Energy for Air Force and DoD applications
- ***To avoid*** technological surprise
- ***To explore*** DE technology avenues that offer high payoff to directed energy capabilities and applications

To do what has never been done before!

Directed Energy Directorate

Our Vision of the Future:
Directed Energy will Dominate the Battlespace of the 21st Century

Strike

We will strike deep in the enemy's territory at the speed of light, with little or no collateral damage or loss of life, crippling his ability to wage aggression.

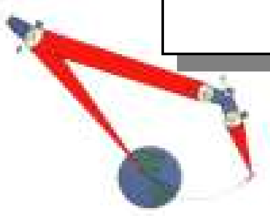

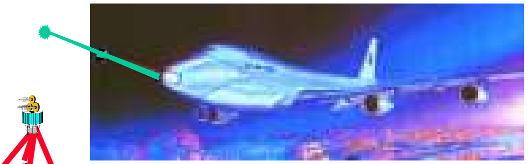

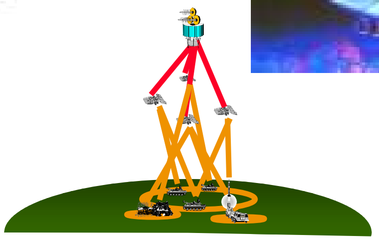

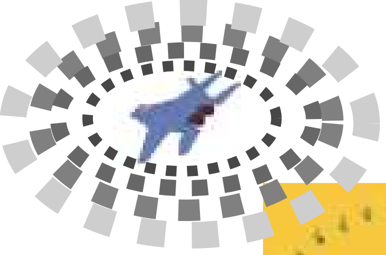

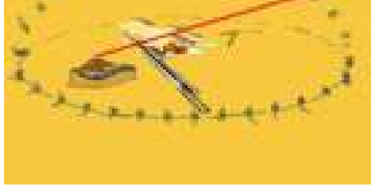

Our high value and limited military assets will be protected by invisible shields of directed energy.

Protect

Graduated Deterrence

We will provide our warfighters with weapons that provide a wide range of graduated force for every military contingency.

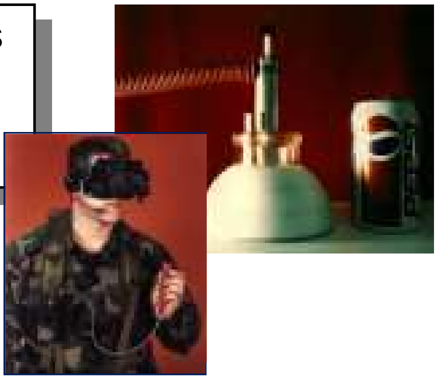

Awareness

Our battleflield awareness will be greatly enhanced with directed energy tools capable of detection and identification

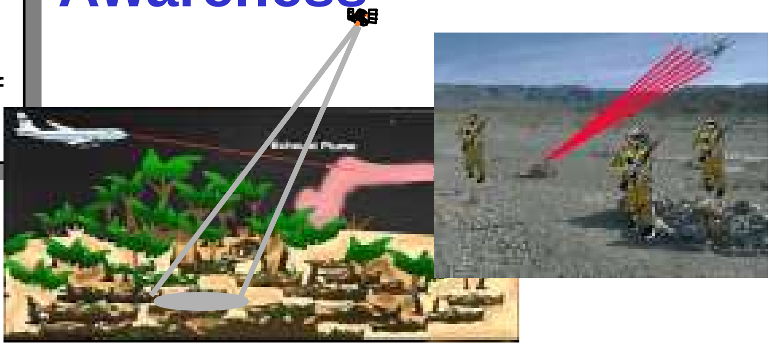

Security & safety in
isolated canyons

Access to 50,000
acres for testing

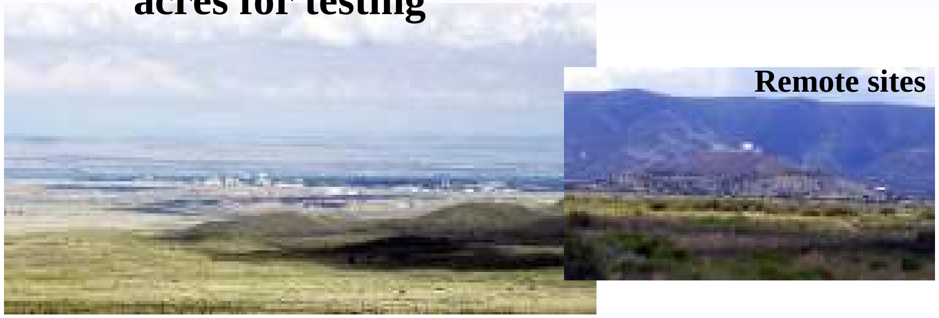
Remote sites

Ideal weather
“thin” atmosphere

Kirtland AFB

Unique & ideally suited for the nation’s

directed energy technology development

Dark skies
No light
“pollution”

Sandia National Laboratories

Many diverse, high tech tenants

International airport & metro area

Directed Energy Directorate
Remote Sites

- Maui, HI (leased land)
  - Relay Mirror Experiment Facility
  - Maui High Performance Computer Center
  - Maui Space Surveillance Complex

- North Oscura Peak, NM
  - Telescope site at White Sands Missile Range
  - Supports Space Imaging R&D (ABL Technology)

Lasers and Microwaves

Synergistic, Complementary and Capable

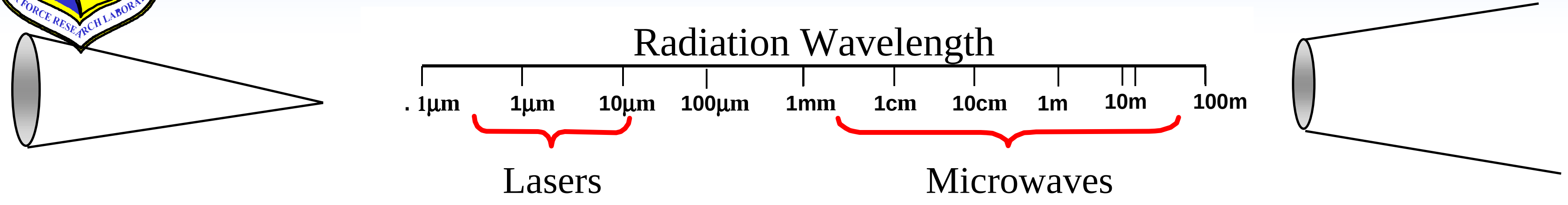
Radiation Wavelength
. 1µm
1µm
10µm
100µm
1mm
1cm
10cm
1m
10m
100m
Lasers
Microwaves

High Energy Lasers	High Power Microwave s
Small spot: surgical attack	**Flood target area: multiple simultaneous targets**
Affects targets from outside: structural destruction	**Affects components inside: electronic disruption**
Weather sensitive: clouds, dust, molecular absorption	**Weather insensitive: sees through clouds and dust.**
Modest apertures: glass mirrors, coatings	**Big apertures: metal antennas, ground planes**

But on the other hand,... they both

- **Travel at the speed of light to the target**
- **Capable of graduated effects from deny, disrupt, degrade, and destroy.**
- **Minimum collateral damage**

Directed Energy Directorate
Broad Technology Investments

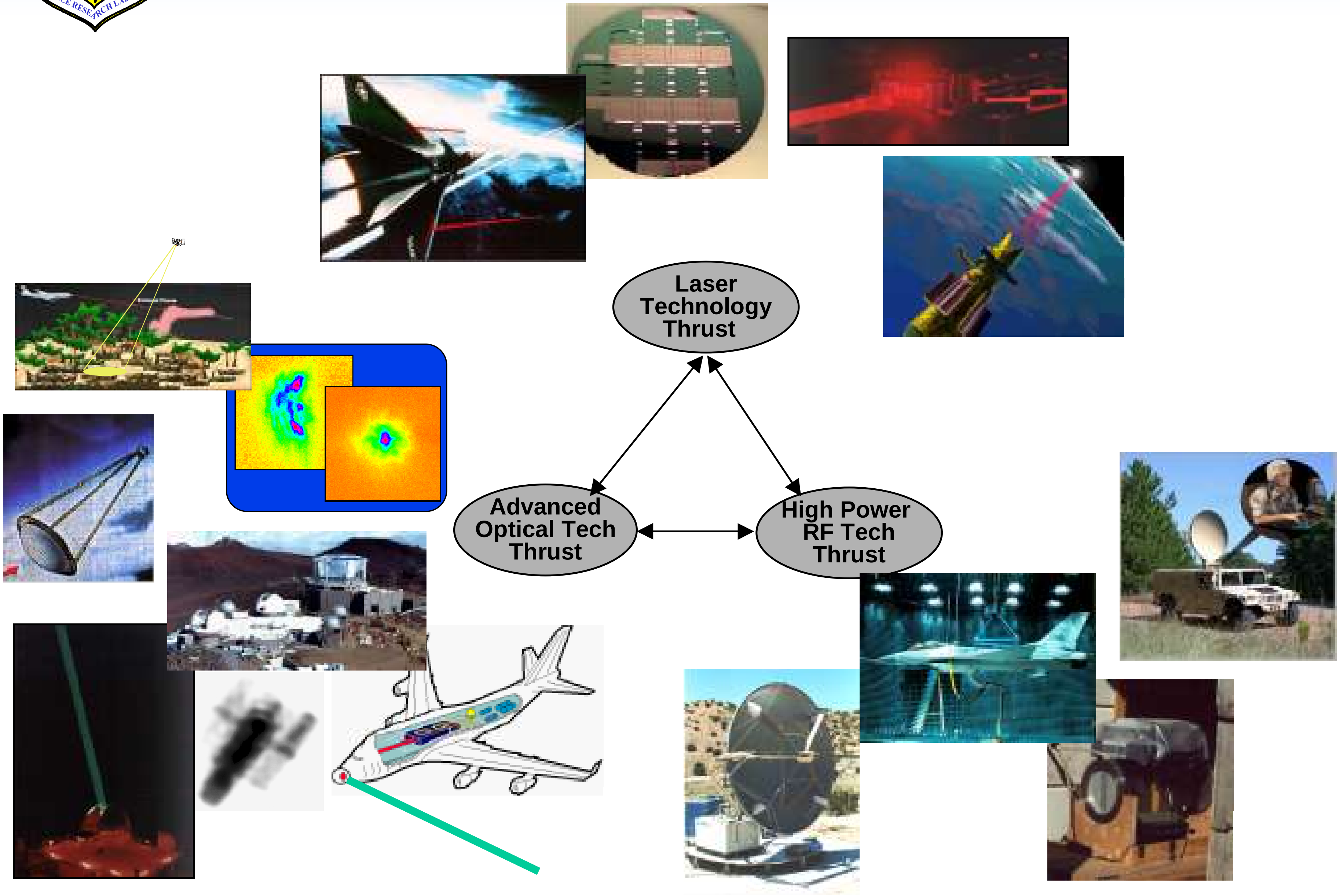
Laser Technology Thrust
Advanced Optical Tech Thrust
High Power RF Tech Thrust

Directed Energy Directorate

12 Major DE Technology Programs

Ground Based Laser Technology Program

Airborne Laser Technology Program

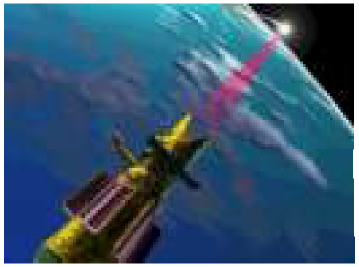
High Power Laser Program

Scalable Multi-Wavelength Laser Program

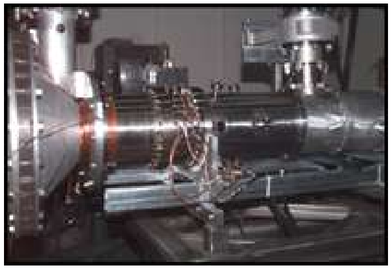
HPM Sources Program

HPM Pulsed Power Program

HPM Effects Program

HPM Applications Program

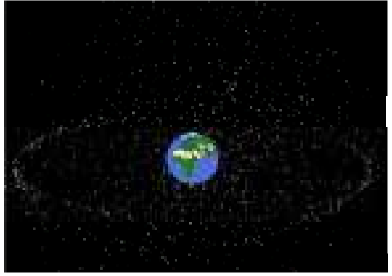
Space Awareness Program

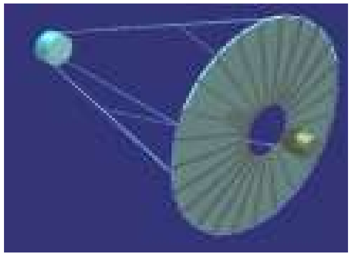
Large Optics

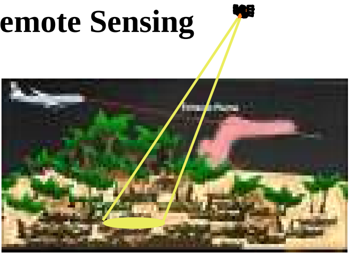
Remote Sensing

High Efficiency Electric Laser Program

AFRL
Directed Energy Directorate

Mission

Develop, integrate and transition science & technology for Directed Energy to include high power microwaves, lasers, adaptive optics, imaging and its effects to assure the preeminence of US in air & space

Major Technology Programs

Ground Based Laser
Airborne Laser
High Power Lasers
Multi-Wavelength Laser
Electric Lasers
Large Optics

Space Awareness
HPM Applications
HPM Effects
HPM Pulsed Power
HPM Sources
Remote Sensing

AFRL/DE - factoids
- **952 people (Civ, Mil, Contr)**
- **$109M FY00 budget**
- **4,325 acres**
- **757,000 sf of space**
- **Located throughout Kirtland**
- **Sites at Maui & WSMR**

Air Force Research Laboratory
Directed Energy Directorate

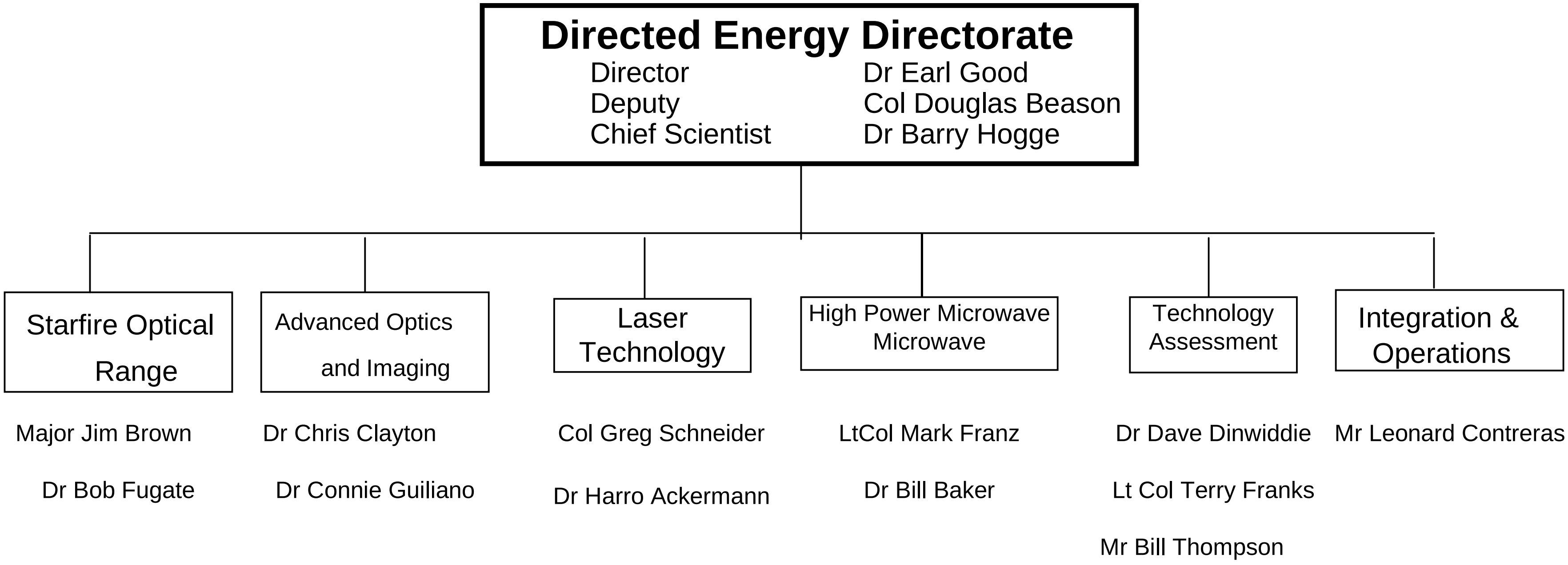

Directed Energy Directorate

Director	Dr Earl Good
Deputy	Col Douglas Beason
Chief Scientist	Dr Barry Hogge

Starfire Optical Range
- Major Jim Brown
- Dr Bob Fugate

Advanced Optics and Imaging
- Dr Chris Clayton
- Dr Connie Guiliano

Laser Technology
- Col Greg Schneider
- Dr Harro Ackermann

High Power Microwave Microwave
- LtCol Mark Franz
- Dr Bill Baker

Technology Assessment
- Dr Dave Dinwiddie
- Lt Col Terry Franks
- Mr Bill Thompson

Integration & Operations
- Mr Leonard Contreras

Directed Energy Directorate

Major Kirtland Facilities

Lasers, Optics, Imaging

Starfire Optical Range

Advanced Laser Facility

Argus Facility

Optics Development
& Beam Control Laboratory

Chemical-Oxygen-Iodine-
Laser Facility

Optics Coating & Components Evaluation Laboratory

Laser Effects Test Facility

Directed Energy Directorate

Major Facilities - High Power Microwave

High Energy Microwave Laboratory

High Energy Research & Technology Facility

Plasma Research Laboratory

Electric Research Laboratory

Electromagnetic Energy/RF Effects Laboratory

Directed Energy Directorate

Manning

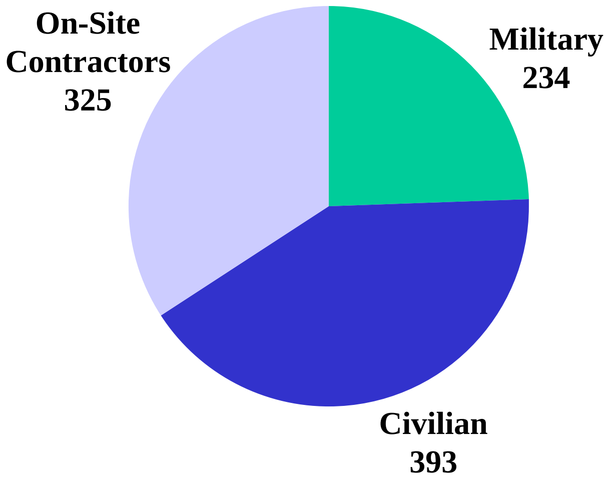

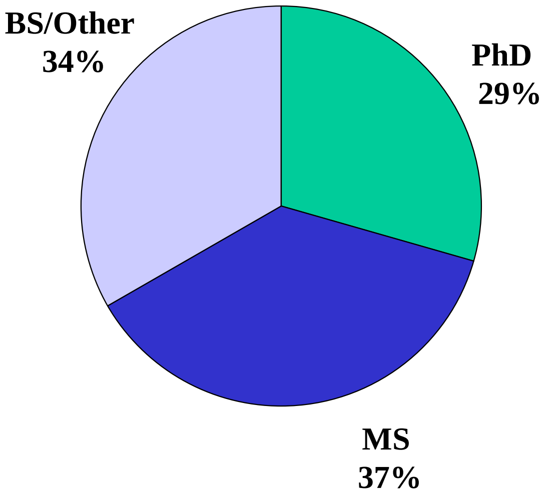

Assigned Personnel
952

Government S&E Education Levels

Directed Energy Directorate

FY00 AFRL PB Resources

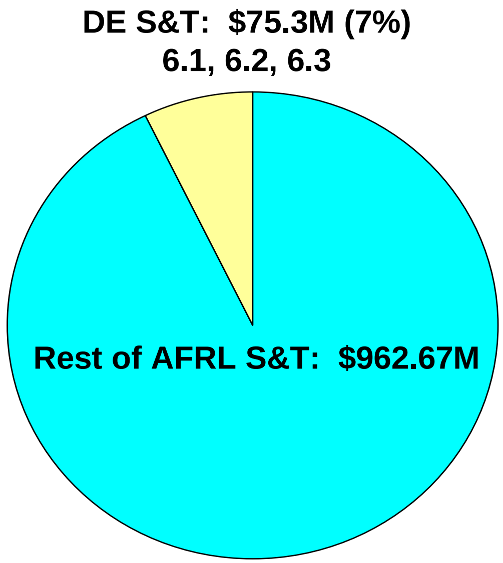

AFRL TOTAL S&T: $1,037.98M

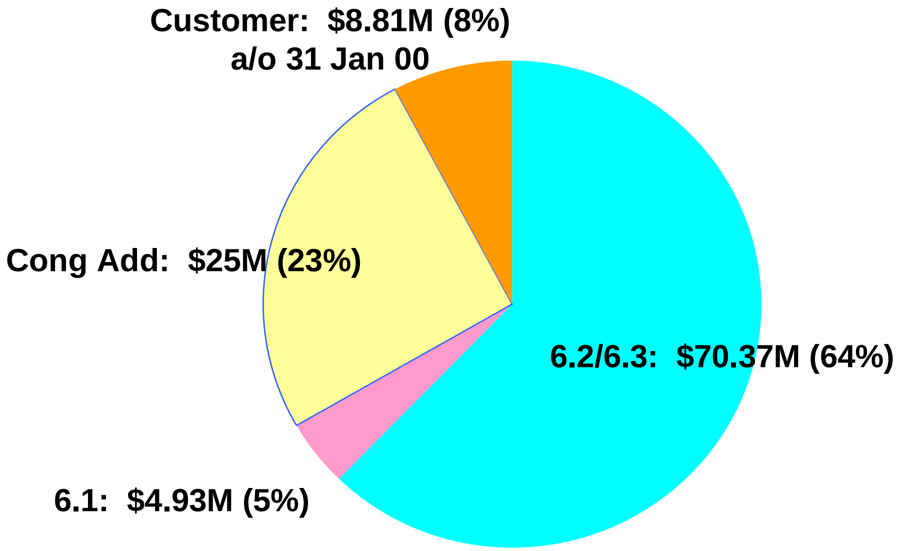

Directed Energy Total: $109.11M

Directed Energy Directorate

Potential Areas for Collaboration

- Electric Lasers
- Beam Control
- Large, Lightweight Optics
- Optical Coatings
- High Power RF Sources
- High Power RF Antennae
- Atmospheric Compensation
- Remote Sensing

For Information contact Mr Rich Garcia, AFRL/DEOB-PA (505)846-4583

AiMT Advances in Military Technology
Vol. 4, No. 2, December 2009

Directed Energy Weapons in Modern Battlefield

L. Palíšek[*]

Division VTÚPV Vyškov, VOP-026 Šternberk, s.p., Czech Republic

The manuscript was received on 25 June 2009 and was accepted after revision for publication on 21 September 2009.

Abstract:

The paper deals with electromagnetic threats caused due to Directed Energy Weapons (DEW). The area of interest is focused on HPM (High Power Microwaves) and UWB (Ultra Wide Bandwidth) as an important part of DEW area. Possible impact on electronic structures which can be part of military systems is considered in this paper. Some practical results achieved during experiments related to assessment of electronic equipment vulnerability is provided with overview of immunity of some equipment to HPM and UWB signals at the end of this study.

Keywords:

DEW, HPM, UWB, electromagnetic threats, NEC, vulnerability

1. Introduction

Directed Energy Weapons (DEW) with various principles seem to be more important in the future battlefield. There are several principles which are considered for use in the DEW area. Special optical technologies (lasers), sound technologies as well as microwave technologies are developed and can be used in special weapon systems to disturb military system behaviour or can be used for counter personnel weapons (for example Active Denial System working on frequency 94 GHz, see Fig. 1). DEW has been mainly considered as technologies suitable for less-lethal or even non-lethal technologies which have been subjects of interest within the past years.

Recent developed military systems consist of equipment which is more and more sophisticated and relying on electronics. Very high importance of electronics within the military systems is obvious from development of Network Enabled Capabilities (NEC) which is a key task of the Czech Armed Forces as well as within the NATO (NNEC – NATO Network Enabled Capability). The stage related to "Communication and information sharing" has been topical in recent years and characterised by development of communication networks. Common network infrastructure can be

---

[*] *Corresponding author: Division VTÚPV Vyškov, VOP-026 Šternberk, s.p., Olomoucká 1841/175, 785 01 Šternberk, Czech Republic, phone: +420 517 303 638, E-mail: l.palisek@vtupv.cz*

considered as a fundament of integrated NEC environment. Communication network and especially its reliability will have significant influence on mission success. Disturbance of these networks could be obviously critical.

Power electromagnetic fields can be generated by relevant power technologies and focused (directed) towards the target with electronics. Shortcuts like HPM (High Power Microwave) and UWB (Ultra Wide Bandwidth) are widely used in this area as well as HPEM (High Power Electromagnetics) and IEMI (Intentional Electromagnetic Interference). Examples of developed technologies which could be easily used like DEW against electronics are shown in Figs 2-6 (these technologies were offered or presented mainly during relevant conferences in past few years, radiated pulse power was hundreds of MW or higher in frequency range of microwaves).

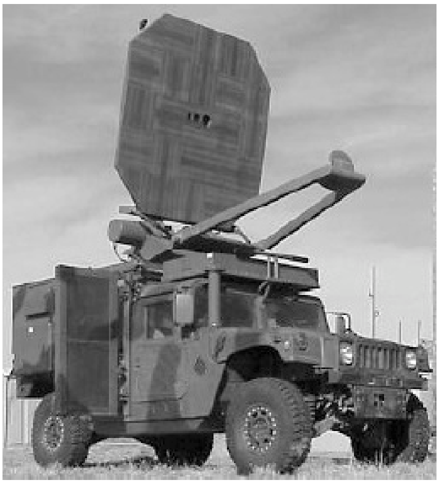

Fig. 1 Humvee with ADS (Active Denial System) mounted [1]

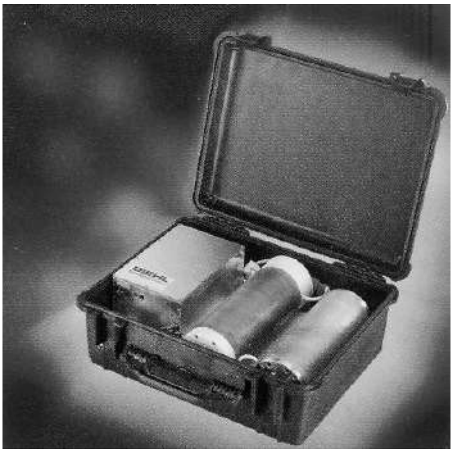

Fig. 2 HPM system DS-110 (built in suitcase, pulse power 250 MW [2]

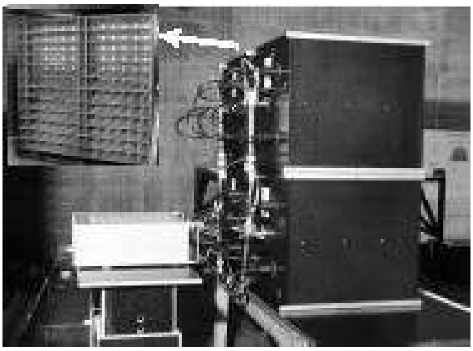

Fig. 3 HPM system with antenna array [2]

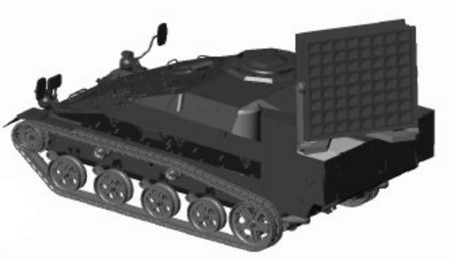

Fig. 4 HPM system with antenna array installed on a tank [2]

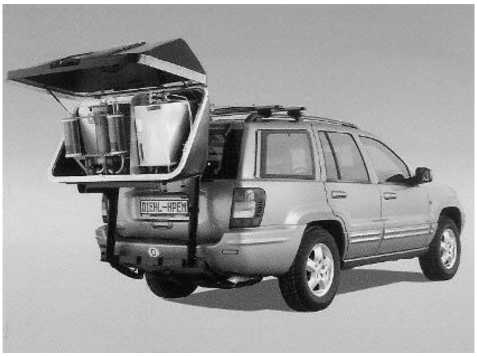

Fig. 5 Electromagnetic System for Selective Car Stopping [2]

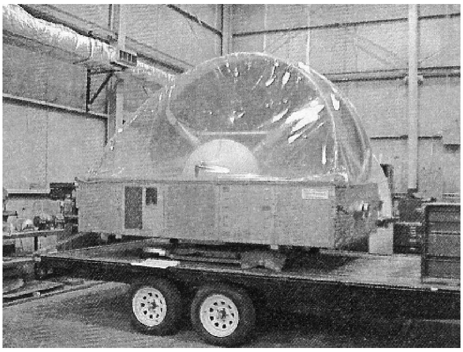

Fig. 6 JOLT hyperband system, UWB system [3]

2. Electromagnetic Threats

As it was mentioned in the introduction an important group of DEW technologies, non-lethal technologies, was represented by weapons with use of power electromagnetic field destructive effects.

Next electromagnetic environments are typically considered related to DEW (see Fig. 7):

- HPM: High Power Microwaves, narrowband signals in frequency range from hundreds of MHz up to several GHz, electromagnetic strength at the place of target typically up to tens of kV/m (strong dependence on range).
- UWB: Ultra Wide Bandwidth, pulses with rise time typically in hundreds of picoseconds and with duration of several nanoseconds, electromagnetic strength at the place of target typically up to tens of kV/m (strong dependence on range).

Because HPM and UWB environment is strongly dependent on range and moreover these environments may have a wide variety of wave shapes, their standardization is very difficult. Now there are only few standards which describe HPM and UWB environment generally (e.g. IEC 61000-2-13 [4]). A possible approach is to consider various types of HPM and UWB environments that have been produced and to consider their possible use in the future against sensitive targets. It is always necessary to make relevant analysis of electromagnetic threat where the shortest possible distance from HPM or UWB source is a very important parameter.

2.1. Electromagnetic Strength Assessment

As a typical possible scenario can be considered HPM (UWB) source with real pulse power $P = 1$ GW and as a worst case special antenna with extremely narrow radiation pattern ($G = 40$ dB, respectively $D = 8\,750$). Now the expected electromagnetic strength at the place of threatened equipment (system) can be calculated according to next Eq. (1)

$$E = \frac{\sqrt{30PD}}{r} \tag{1}$$

where P is radiated power, D is directivity of antenna, r is a distance of threatened equipment (system) from radiating antenna.
Note: Eq. (1) is valid only for direct propagation where reflections are neglected. This simplified equation is usually used for the microwave area and short distances.

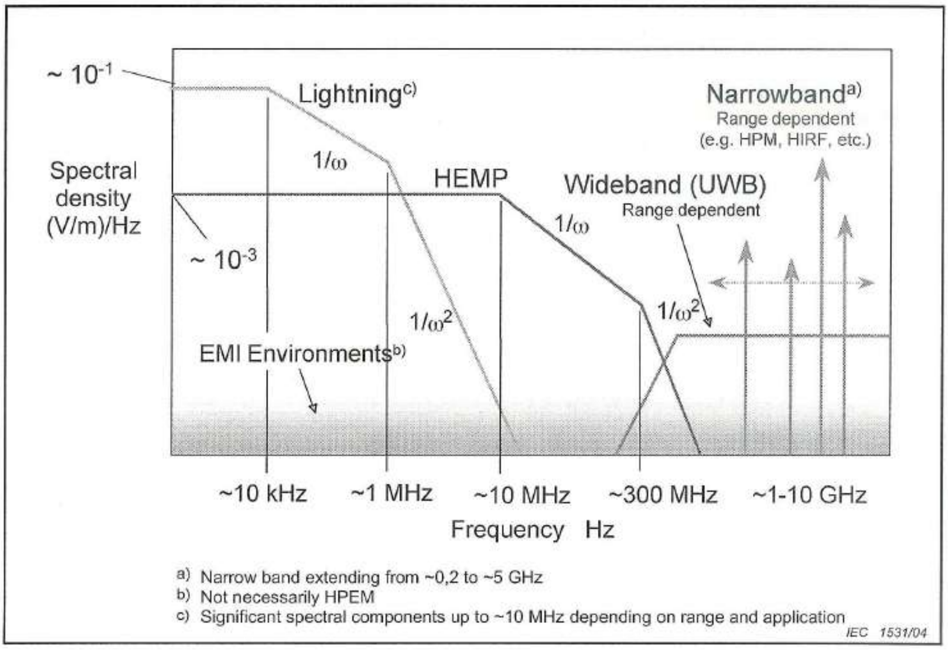

a) Narrow band extending from ~0,2 to ~5 GHz
b) Not necessarily HPEM
c) Significant spectral components up to ~10 MHz depending on range and application
IEC 1531/04

Fig. 7 HPM and UWB comparison with other electromagnetic threats [4]

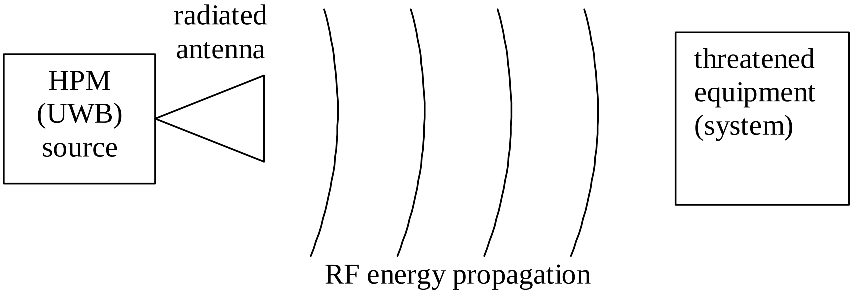

Fig. 8 RF energy propagation from source (HPM or UWB) to equipment (system)

It is obvious from Eq. (1) that for most cases (distance of directed energy weapon from equipment (system) is generally tens of meters), it is possible to consider electromagnetic strength up to tens of kV/m.

2.2. *Critical Frequencies Assessment*

It is very useful to consider transfer function (*T*) of investigated equipment (system) for critical frequencies assessment. Transfer function T_O (outer) is combined with T_I (internal) to form a composite transfer function from incident fields to response box inputs in Fig. 9 [5]. Note: T_O typically corresponds to relationship between incident electromagnetic field and surface current and charge densities on the outer surface. T_I corresponds to transfer function from the surface current and charge densities on the outer surface to some interior port of interest, producing voltage and current waveforms there.

From Fig. 9 it is obvious the most important part related to critical frequencies of considered system (equipment) is band 2 – resonance region. It is possible to use first of all analysis where sizes of equipment (system) as well as used cable lengths are considered and next it is important to make relevant measurement of transfer functions (shielding effectiveness of used boxes, induced currents and voltages etc.) for transfer function (*T*) assessment.

Considered frequency spectrum is divided as follows [5] (see Fig. 9):

Band 1: $f \le f_l \approx \dfrac{c}{4l}$ (aperture and small antenna coupling region),

Band 2: $f_1 \le f \le f_h \approx \dfrac{5c}{l}$ (resonance region, external and internal),

Band 3: $f_h \le f$ (integration region),

where *l* is characteristic dimension of the object, *c* is speed of the light, *f* is considered frequency.

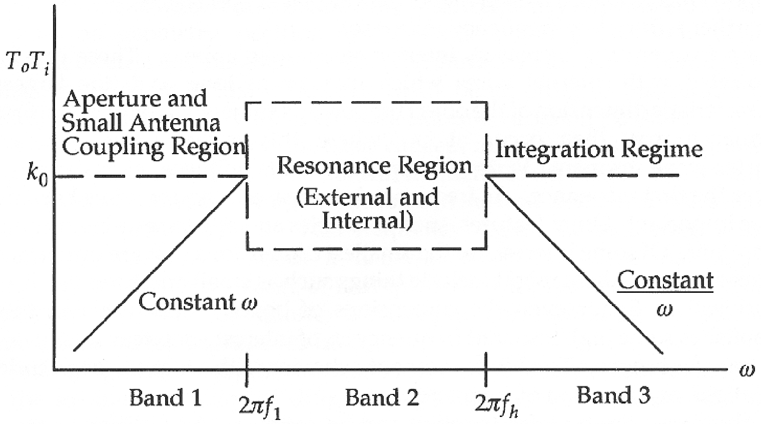

Fig. 9 System response as a function of frequency (logarithmic scales) [5]

3. HPM and UWB Impact on Electronics

Electronic systems can be affected by high intense electromagnetic fields causing the disturbance or even the permanent damage of the system. According to relevant EMC (Electromagnetic Compatibility) standards next performance criteria are usually used:

A - normal performance within the specification limits

B - temporary loss or degradation of function (self-recoverable)
C - temporary loss of degradation of function, which requires operator intervention or resetting
D - degradation or loss of function which is not recoverable due to damage of equipment (components) or software damage

On the one hand performance criteria are a very important parameter on the other hand consequence of IEMI (Intentional Electromagnetic Interference) attack has to be considered too. For consequence scaling it is possible to use for example such terms like "very limited", "limited", "severe", "very severe" and "catastrophic" [6].

4. Vulnerability Assessment

First of all it is necessary to make analysis where electromagnetic threat assessment is done (see chapter 2). Moreover it is necessary to consider criticality of relevant system (consequences – see chapter 3). Finally testing plan and testing can be prepared. The best way is to start with transfer function measurements (shielding effectiveness measurement, induced currents and voltages measurement) which is possible to carry out in wide frequency range up to few GHz or higher. After that it is necessary to choose possible critical frequencies according to transfer function measurements and according to analysis. Chosen frequencies can be used for final high power electromagnetic field measurement with HPM and UWB simulator use. Possibilities of these simulators are restricted to parameters which cannot be changed or changing these parameters takes too much time. Due to this fact it is necessary to choose suitable parameters according to analysis and transfer function measurement results for this testing. The data gained from analyses and testing has to be considered together with possible consequences which can occur. Such data can be used for vulnerability assessment.

5. Practical Results

A lot of experiments related to electronic equipment vulnerability have been carried out in previous projects within the last few years.

Transfer function measurements were carried out for better understanding of electromagnetic field coupling into relevant cables. Tested cables were irradiated with electromagnetic field with vertical polarisation and induced voltages were measured in frequency range 30 MHz to 1 GHz. It was low power measurement, electromagnetic strength of incident electromagnetic field was 1 V/m and measured results were recalculated for expected high power electromagnetic field 10 kV/m. Mainly cables which are widely installed in computer technologies were used for the purpose of these experimental measurements. Few examples are shown in Figs. 10-12. It is possible to expect induced voltage levels on regular widely used cables approximately up to 1 kV during electromagnetic field incidence with electromagnetic strength 10 kV/m on the basis of presented results. The highest induced voltages occur when the length of cables is comparable with wavelength of used signal (resonant region, see Figs. 10-12). For lower frequencies as well as for higher frequencies induced voltages are becoming lower.

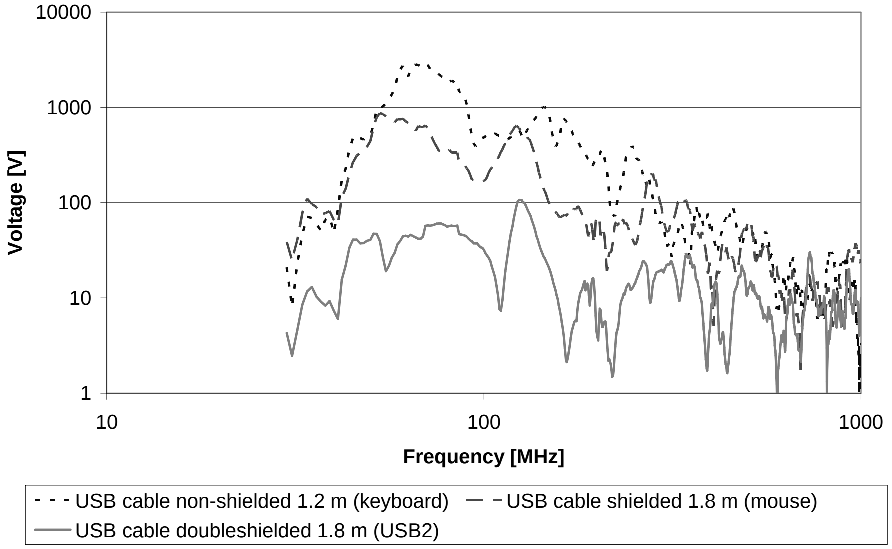

Fig. 10 Induced voltages on USB cables for electromagnetic strength E = 10 kV/m

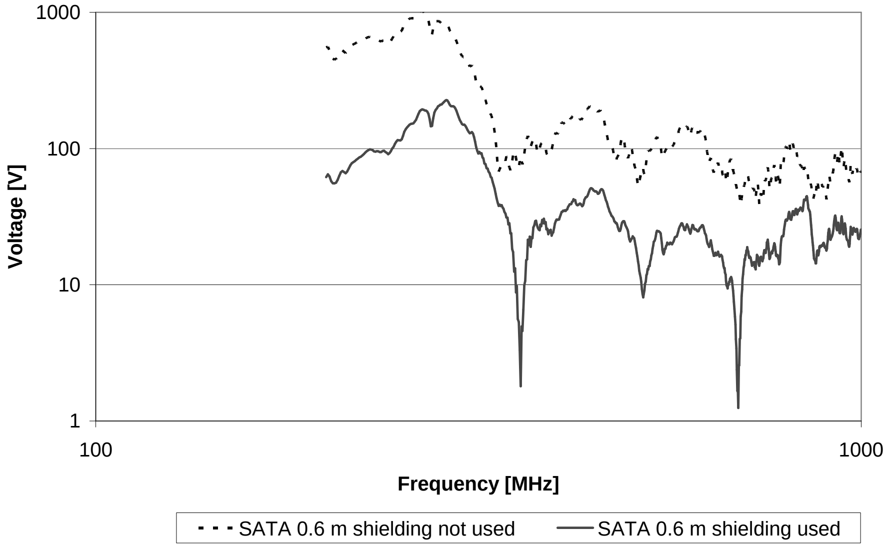

Fig. 11 Induced voltages on SATA cables for electromagnetic strength E = 10 kV/m

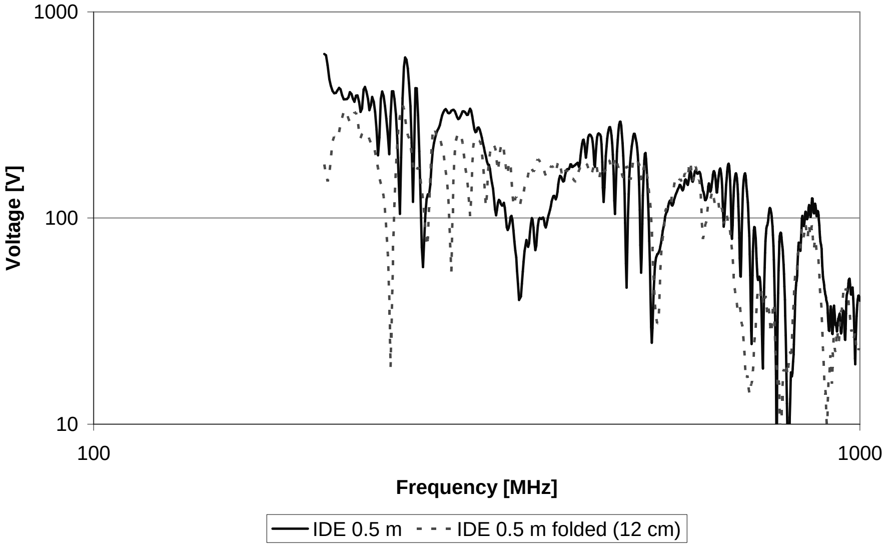

Fig. 12 Induced voltages on IDE cables for electromagnetic strength E = 10 kV/m

Other kinds of experiments in this area are testing with high power electromagnetic field irradiation of possible "targets" and evaluation of effects which can occur.

Some examples of tested setups are shown in Figs 13-15. It was verified that HPM signals can be used for stopping the car engine (see Fig. 13, HPM generator working on frequency 3 GHz with pulse power 500 kW was used for this experiment in shielded semi-anechoic chamber) as well as for effective disrupting of computer technologies and data transfers (see Figs. 14-16). Standard performance criteria (see chapter 3) are usually used for the evaluation of effects. From experiments which were carried out it was obvious that electromagnetic strength of few kV/m can be enough to achieve effective disturbance on electronics.

Repetition rate as a parameter of HPM and UWB signals can have an influence on some kind of temporary failures (see Fig. 16). Overview of immunity of some equipment to HPM and UWB signals is shown in Fig. 17. Typical electronic equipment was tested against effects of HPM 3 GHz, 6 GHz, and 9 GHz signals with pulse widths in hundreds of ns, repetition rates in hundreds of Hz and against effects of UWB signal with rise time of 0.5 ns, pulse duration of 3 ns in single shot regime. The most effective signal to disturb electronics from used ones was HPM 3 GHz. Very often the sensitivity level was very close to damage level (see Fig. 17). For UWB signal it was more typical to achieve temporary failures (sensitivity level) rather than some kind of damage. The results from testing showed the most vulnerable parts of recent systems are computer technologies including data transfers.

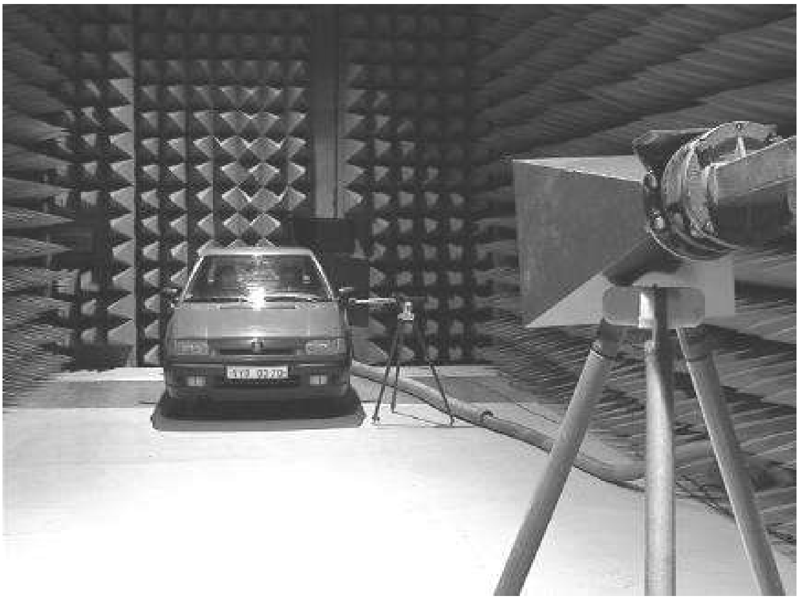

Fig. 13 Stopping the car engine with HPM irradiation use (HPM 3 GHz)

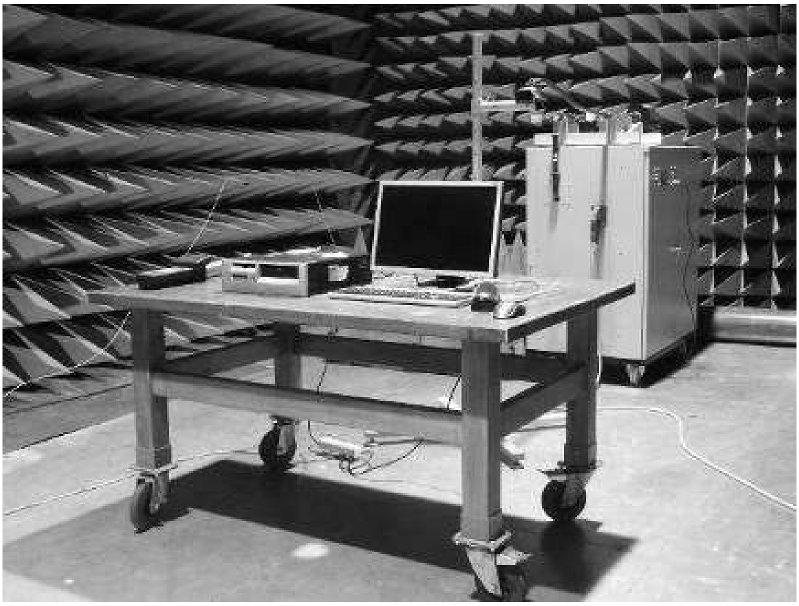

Fig. 14 PC setup irradiated with HPM 9 GHz

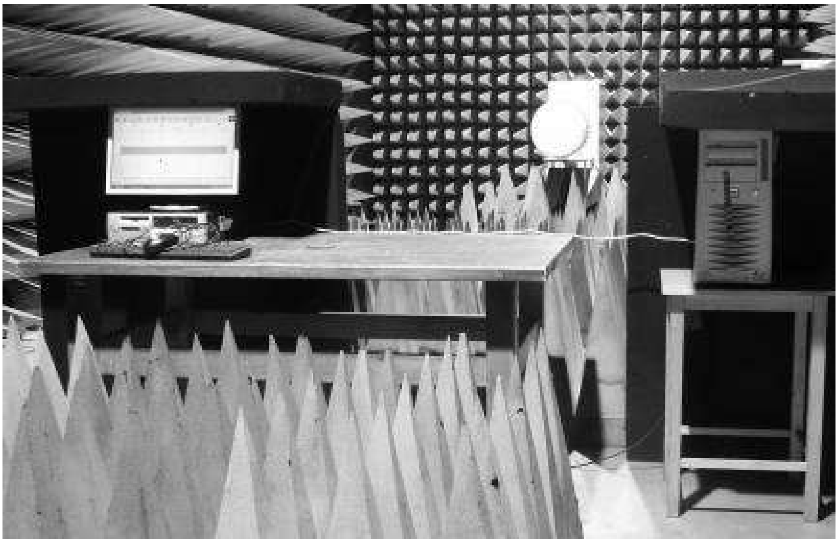

Fig. 15 Data transfer through UTP cable irradiated with UWB signal

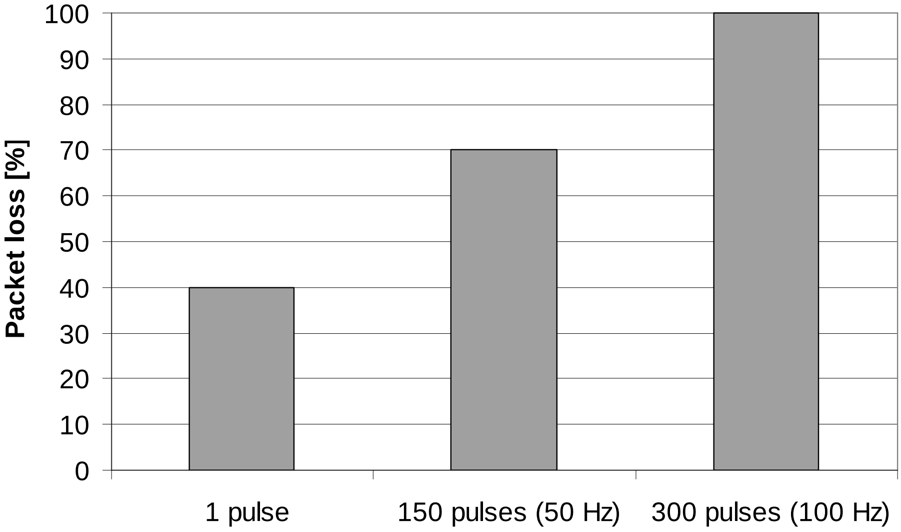

Fig. 16 UTP 2 m irradiated with UWB

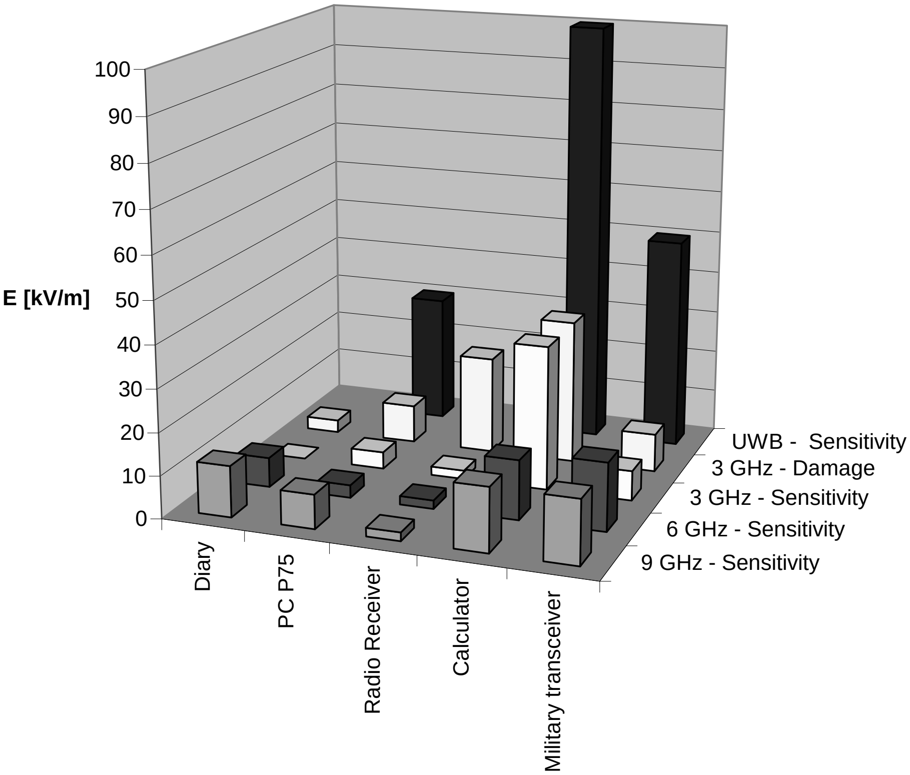

Fig. 17 Sensitivity and damage levels comparison for tested equipment against HPM and UWB effects

6. Conclusion

Electromagnetic threats HPM and UWB as a part of DEW technologies were briefly defined in this paper. Electromagnetic threats assessment as well as vulnerability assessment was introduced. Vulnerability of chosen tested equipment was presented on relevant examples. It was demonstrated that electronic equipment was vulnerable due to electromagnetic field irradiation. In case sensitive electronic equipment is used in crucial applications it is necessary to evaluate vulnerability of the system to such threats.

Gained results in the past have lead to starting a new project for the Ministry of Defence of the Czech Republic (OSPROZ-DEW1) focused on the assessment of vulnerability of weapon systems and parts of communication and information systems integrated in NEC which are installed to ordnance of the Army of the Czech Republic now. Assessment will be carried out on the basis of relevant analyses as well as relevant experimental measurements. The project started in 2009.

References

[1] *Wikipedia* [online]. Wikipedia, the Free Encyclopaedia. [cited 2009-09-15]. Available from: <http://en.wikipedia.org/wiki/Active_Denial_System>.

[2] *DIEHL Defence* [online]. DIEHL Defence. [cited 2009-09-15]. Available from: <http://www.diehl-bgt-defence.de>.

[3] GIRI, DV. *High – Power Electromagnetic Radiators: Nonlethal Weapons and Other Applications.* Cambridge (Massachusetts) and London : Harvard University Press, 2004. 194 p. ISBN 0-674-01569-X.

[4] International Standard IEC 61000-2-13 *Electromagnetic compatibility (EMC) – Part 2-13: Environment – High-power electromagnetic (HPEM) environments – Radiated and conducted*. 1st ed., 2005.

[5] TAYLOR, CD. and GIRI, DV. *High-Power Microwave Systems and Effects*. Washington : Taylor and Francis, 1994. 199 p. ISBN 1-56032-302-7.

[6] MANSSON, D., THOTTAPPILLIL, R. and BACKSTROM, M. Methodology for Classifying Facilities With Respect to Intentional EMI. *IEEE Trans. on Electromagnetic Compatibility*, 2009, vol. 51, p. 46-52. ISSN 0018-9375.

Acknowledgement

This paper is a part of study of the project OSPROZ-DEW1 "Defence capabilities against DEW – Vulnerability assessment of weapon systems and infrastructure C2" supported by the Ministry of Defence of the Czech Republic.

will be available to treat and possibly cure diabetes, hemophilia and other life-threatening and chronic illnesses.

Space manufacturing will develop the new metal alloys needed for high-temperature processing using advanced nuclear and fusion energy sources. Larger and more perfect crystals for the electronics and other industries will be produced.

By themselves, these new industries created by exploiting the unique environment of space, "pay" for the NASA programs. But the indirect impact to the economy has been even greater, and more important. These are the so-called "spin-offs" from space exploration which have been developed by industry using technological breakthroughs funded by NASA.

New materials, such as refractory ceramics and new alloys which withstand the temperature extremes of space flight, are used in everything from nuclear power plants to everyday kitchen implements. Developments in electronics, including miniaturization and sophisticated automation techniques, have streamlined the functioning of industry and shortened the time (thereby increasing the productivity) of almost everything we do.

The machines that monitor all of the vital bodily functions of the astronauts who walked on the Moon, now monitor premature infants in incubators, and have provided heart monitors and miniaturized pacemakers for thousands of people.

The same qualitative impact, magnified many times, will result from the development of beam weapons combined with the recommitment of the civilian space program to establish a permanently manned station in space toward the goal of colonizing the Moon and other planets.

However, studies done by Chase Econometrics and confirmed by independent studies using the Fusion Energy Foundation's econometric model show that there is a much greater impact that the development of a qualitatively new technology has on an economy. This impact is, strictly speaking, not measurable by adding up all the new products and new techniques that a new technology introduces; it is the *increase in productivity* throughout the economy as the result of the combination of higher manpower skill levels and new scientific knowledge entering industrial production. One study of this induced productivity effect estimated that U.S. productivity increased 0.1 percent for every billion dollars spent on the space program. This change in productivity alone represented an additional $3 billion to the GNP every year it was present. An interesting comparison is possible between the expenditure of a $1 billion aliquot of the federal budget on a high-technology R&D oriented program (like the Apollo program or the development of a beam weapon) and its expenditure on transfer payments, bureaucratic services, or the like. The Chase Econometrics study showed that the expenditure of this money on high-technology R&D actually *lowered* inflation, while the other expenditure had the opposite effect, raising inflation by 0.2 percent.

Four types of directed-energy weapons

by Mary McCourt

In his *EIR* multi-client report, *Beam Weapons: The Science to Prevent Nuclear War,* Dr. Steven Bardwell describes the types of beam weapons on line for development. Each type, laser beams, particle beams, microwave beams, and plasma beams, is, Bardwell states, "in principle capable of generating the required power and energy [to reach and disarm its target] in a form efficiently absorbed by the missile." A beam weapon effectively disarms a nuclear warhead. A hydrogen bomb can be detonated only by an initial powerful atomic-bomb explosion capable of setting off a chain reaction in the lithium-deuterium fuel. A beam weapon, by pumping energy into the very delicately balanced triggering mechanism, prevents the initial explosion and essentially turns the warhead into a "dud." The missile, like a satellite, might fall to the earth, but it can no longer be detonated.

Scientists agree that laser-defense battle stations, even with the lowest level of laser-beam technology, can be defended from other beam weapons themselves. But a missile cannot be effectively defended from the beam without such massive protection that it would lose both the necessary speed and distance.

Laser-beam weapons

Laser beams, particularly the chemical laser, will likely be the first deployable beam weapons developed. A laser is a beam of very intense, single wavelength electromagnetic waves, either of light or high energy X-rays. Such a weapon can be focused very precisely because either the light or X-ray wavelengths all have the same frequency and phase. The five different types of lasers, which can be applied to fusion energy as well as beam weapons, are all being researched at U.S. laboratories.

The **chemical laser,** which could be developed for military deployment within five years, uses a gaseous medium in which a chemical reaction is induced. The product of the reaction emits laser light. The Soviet Union used such a laser last year in tests that downed a ballistic missile.

In a **gas laser,** a burning gas such as a hydrogen and fluorine mixture is suddenly compressed, and the energy distribution that results from the compression is then stimulated to emit single-frequency light waves at very high energy. Both the United States and Japan are currently using huge gas lasers for nuclear fusion development.

An **electron discharge laser** uses replaceable energy

© 1982 EIR News Service Inc. All Rights Reserved. Reproduction in whole or in part without permission strictly prohibited.

from an electron beam to create the source of laser light. Such a laser would be very efficient for use in space because its energy source is electricity, not an exhaustible chemical fuel.

Two other types of lasers, **X-ray lasers** and **free-electron lasers,** yet to be perfected technologically, have greater advantages of energy density and flexibility than those listed above. The **X-ray laser** is widely recognized as the most promising long-range laser for ballistic missile defense, based in space. The X-ray laser, which is just a single pulse, is by far the most energy-dense, delivering thousands of times more energy per pulse than conventional lasers. In addition, the target absorbs the X-rays very efficiently, making this weapon capable of very efficient destruction of missiles.

Particle-beam weapons

Particle beams also deliver energy in a highly controlled pulse traveling at near the speed of light. But instead of a pulse of intense electro-magnetic radiation, the particle beam consists of **subatomic particles, (electrons** or **protons), neutral atoms** (usually hydrogen), or (usually magnetized) **macroscopic particles** accelerated to high speeds. A particle beam destroys its target, the triggering mechanism, by creating a very intense shock wave within the mechanism, like a very small, but extremely heavy and powerful hammer striking down on the target.

Electron beams can be generated in the range of millions of volts. Scientists researching the electron beam for military or civilian energy use have discovered that the electron beam becomes a complex structure of electrons and a magnetic field. Such structured beams are capable of carrying higher currents and more energy for much longer distances at much great levels of power output.

Proton beams, which have been researched intensively for the past 30 years, use an electron beam as a seed and then are accelerated in their own right. As protons are 2,000 times heavier than electrons, a proton beam of the same velocity has 2,000 times the energy of an electron beam.

Neutral particles eliminate many of the problems of charged-particle beams, which can degrade both the efficiency and controllability of the beam. By spring of 1983, U.S. researchers will have produced a beam of protons at an energy of 2.5 million electron volts, capable of traveling at 99 percent of the speed of light.

Control and targeting of a **macroscopic particle beam** would be more difficult, but the unparalleled power density of the beam—due to the large mass of the particles—gives it great potential as a beam weapon.

Microwave and **plasma beams** have only been discussed in the United States in the past two years, although it is estimated that the Soviet Union is two or three years ahead in the production of microwaves. Intense, directional microwaves are generated when electron beams are propagated at or near the speed of light through a plasma. Such beams, if focused, could destroy delicate electronic equipment in a target.

A plasma beam consists of the highest energy-dense form of matter, a gas so hot that the electrons and neutrons have separated. The plasma forms itself into a complex structure of particles and magnetic fields. The unique feature of a plasma beam is that it actually requires the atmosphere, which hinders the guidance and propagation of other beams, to hold in the plasma and sustain the structure.

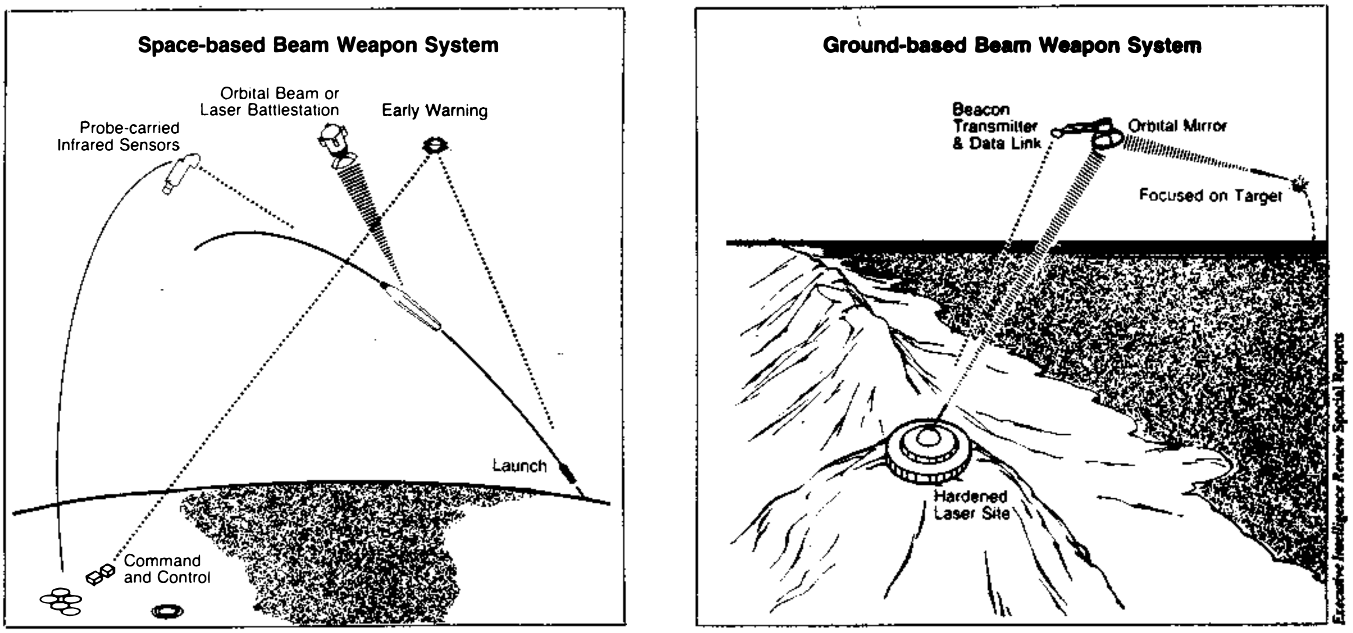
Space-based Beam Weapon System
Orbital Beam or Laser Battlestation
Early Warning
Probe-carried Infrared Sensors
Launch
Command and Control
Ground-based Beam Weapon System
Beacon Transmitter & Data Link
Orbital Mirror
Focused on Target
Hardened Laser Site
Executive Intelligence Review Special Reports

Space-based beam weapons, (left) provide the only feasible means for area defense against nuclear-armed ballistic missiles. The deployment of a weapon capable of generating an intense beam of laser light, atomic particles, or plasma, in an orbit around the earth, would protect the entire United States from incoming ballistic missiles. **Ground-based beam weapons** (right) can provide both area and point defense. Shown here is a conceptual design of a laser beam-weapon system built on a mountaintop, which uses a relay mirror to provide aiming and tracking for the weapon.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/242532678

ELECTROMAGNETIC DIRECTED ENERGY WEAPONS FOR ELIMINATING ELECTRONIC SYSTEMS

Article

CITATIONS
2

READS
40

1 author:

Jan Valouch
Tomas Bata University in Zlín

28 PUBLICATIONS 20 CITATIONS

SEE PROFILE

All content following this page was uploaded by Jan Valouch on 06 March 2015.

The user has requested enhancement of the downloaded file.

ELECTROMAGNETIC DIRECTED ENERGY WEAPONS FOR ELIMINATING ELECTRONIC SYSTEMS

Capt. Dipl. Eng. Jan VALOUCH

Annotation:

The article presents information on the current situation in the development and application of electromagnetic weapons and indicates possible trends of their future employment. Under the term electromagnetic weapons we understand means and systems that use technology generating directed power electromagnetic pulses degrading the activities or destructing the electronic circuits of the enemy equipment. This will also enable to neutralize or completely destroy the operation of information, communication, control, firing and other electronic devices. Electromagnetic weapons are currently in use and greater number of countries realizes the perspectives of their future development.

✶✶✶

1. CURRENT SITUATION OF DEVELOPMENT AND USE OF ELECTROMAGNETIC WEAPONS IN FOREIGN COUNTRIES

Principle of electromagnetic weapon activities

Typical representative of electromagnetic (EMC) weapons are so called **Directed Energy Weapons** (DEW) employing the technology of **High Power Microwave** (HPM). These weapons represent highly sophisticated technology of 21st century. Their destructive effect results from the electromagnetic field and they threaten the operation of all devices equipped with electronic circuits, especially by the effect of arcing, overload or by discharge of single electronic component parts. These weapons have strong physical and psychological impact.

In general, DEW consists of impulse source of energy, source of microwave radiation and antenna. The impulse source transforms accumulated energy to high power electric pulse with the length of duration in units of nanoseconds. Accumulated energy can be in kilo-Joule with the power of gigawatt. In the source of microwave radiation this impulse then acts upon diode and generates a beam of electrons with energy – approx, 400 kV, 10—60 kA. This energy is dissipated by the directional antenna.

Electromagnetic weapons can fundamentally differ in terms of its design and application e.g. according to the used frequency band. EMC wide band weapons emit in wide frequency range but with low density of energy. These equipments are suitable where it is not possible to exactly identify the characteristics of the target – especially its working frequencies. Contrary to the narrow-band EMC weapons emit pulses on individual frequencies with enormously high power. Their action upon the target is very effective since impulse resonate with the known frequency of the attacked device.

Development and application of electromagnetic weapons

Concerning a development and application of electromagnetic weapons we can say that after the EMC weapons emerged from "the unknown" world of classified projects, now they are tuned and EMC weapon capability is enhanced. And they are gradually mounted on suitable carriers – vehicles, aircraft, ships, bombs, missiles and even space ships. The interest is focused not only on single pieces of equipment but also on complete systems incl. sensors and equipment to control the combat functions.

Within the armed forces the employment of EMC weapons have the following advantages:

- very rapid effect against the enemy targets,
- usage irrespective of the weather conditions,
- coverage of a great amount of various targets with minimum need to be informed about their characteristics,
- threat to less available target – under ground,
- operational attack (neutralization, destruction or denial of activities of electronic assets) at selected levels of warfare,
- minimum of collateral destruction in politically sensitive environment and use of this environment after the end of conflict,
- reduction of minimum time for tracking and guidance to target.

USA, Russia, France, China and UK are the countries that achieved the greatest progress in the development and application of electromagnetic weapons. The interest of other countries (Germany, Belgium, the Netherlands, Denmark, Norway …) is for a long time focused on High Power Microwave (HPM) especially on protection against the effects of HPM. As to my opinion this is the beginning of the path for research in the Czech Republic.

Currently, discussions are held on their employment in Iraqi by the U.S. armed forces. They could facilitate quickly eliminate the command and control systems of Iraq and eliminate communication of their forces without any loss of lives and any collateral damage. There were some intentions to employ these weapons of electromagnetic warfare by the U.S. force in Kosovo and Afghanistan at least in trial operation. It is also interesting the assumption that Russian forces used DEW prior the action against terrorists in Moscow theatre in 2002. This weapon was designed to disable the electronic detonating primers. The fact is that no Chechen managed to initiate her primer.

Misuse of electromagnetic weapons

More frequent are the attacks of terrorists and criminal underworld using the electromagnetic means. Among the targets of terrorists can be financial institutions, medical facilities, aircraft, automobiles, computer network and other daily used civilian and military equipment. The EMC radiation generators can be for example in the briefcase and that is why it is quite easy to prepare such an attack. First known terrorist application of electromagnetic weapons is from 1995 when Chechen rebels used this technology against the security system of Russian facility.

Nobody knows about the use of DEW for criminal activities and terrorist attacks with exception of the offenders themselves and their victims. The German experts even officially recommended application of these means to the German police units since there were recorded several cases of use of DEW by the German underground. With DEW it is possible to commit perfect crime, as they leave no evidence. No doubt, in future a wide use of these weapons can be expected which can means an increased risk of their misuse.

Future of electromagnetic weapons

Future of DEW is often discussed in many articles and literature. A part of study called Air Force 2025 also deals with the future of DEW employment that was developed by the Air University Maxwell Air Force Base, Alabama. This study discusses a possibility to employ these means in the spaceships or satellites in order to destroy the hostile satellite information and communication channels.

Development of DEW in terms of their employment is mainly connected with mutual co-operation of the army, air force and navy forces. These will be the components that are the potential users of DEW. Protection of the aircraft and ships against the effects of enemy DEW is being solved. Further, e.g. active use of DEW against the air defense and employment of DEW in the aerospace. Within 3—5 years it is expected that DEW will be installed into the drones (USA). In respect of technologies for DEW, the research will be especially focused on increase of input power of DEW, reduction of size and weight and improvement of antenna systems. A very interesting is also a research of biological impact of DEW on human being and hazard for the personnel.

2. CURRENT SITUATION IN THE AREA OF ELECTROMAGNETIC WEAPONS IN THE CZECH REPUBLIC

The following part of this article briefly discusses NATO RTO – Research and Technology Organization (http://www.rta.nato.int), or the activities of selected panels with representation of the Czech Republic and which deal with the mentioned issue.

The panels dealing with this topic are:

- SCI–119 Tactical Implications of High Power Microwaves,
- SCI–132 High Power Microwave Threat to Infrastructure and Military Equipment.

Since most of activities of these technical panels is classified as NATO SECRET, only general character information not subject to classification are provided.

Panel SCI–119 Tactical Implications of High Power Microwaves started its activities in 1998 and terminated it in this year. The objective of the work of its members have been mainly solution of issues of use of HPM in the military sector both in terms of defense as well as of potential active employment in attack. The most important areas solved:

- design of resources generating HPM,
- vulnerability of information technology of military infrastructure,
- risk of back effect of electromagnetic weapons on friendly troops,
- testing of HPM effects on the off-the-shelf equipment equipped with electronic components (personal computers, cellular phones, vehicles, aircraft …),
- content of HPM national programs from the viewpoint of:
  - earmarking of specialized workplaces,
  - refinement of the content of solution in the given country,
  - allocation of funds,
  - planning of the number of students studying HPM at the universities,
  - gathering of results.

The issue of HPM is solved by NATO systematically since the beginning of 80-ies of the last century. In the short time, it is assumed that new advanced HPM weapons will be developed and within NATO this process must be systematically controlled. In this area – development of electromagnetic weapons – USA, Russian, France and the UK have the greatest success.

Panel SCI–132 High Power Microwave Threat to Infrastructure and Military Equipment commenced its activity by the introductory session in October 2002, held in Munster in Germany. Its orientation is linked to the preceding activity of the Panel SCI–119. The main activities of the Panel are planned for the period of 2003—2005 and single members will deal with the following issues:

- identification of potential threat of HPM on the military and civilian infrastructure and equipment,

- research of the penetration and dissemination of HPM in specific devices,
- research of DEW (Directed Energy Weapons),
- possibilities how to protect the military equipment against the effect of weapon assets on the basis of HPM.

Other areas considered important by the member of the Panel SCI–132 and that are to be in the area of interest are:

- resistance of military equipment against HPM,
- implication of HPM in the military field test and system of their evaluation,
- EMC terrorism.

The Panel consists of representatives of the following countries: Canada, Czech Republic, Denmark, Germany, France, UK, USA, the Netherlands, Norway and Italy.

CONCLUSION

Currently, the use of electromagnetic weapons plays the more important role. Though, the wider employment of these weapons is expected within 5—10 years, there already exist and are used the devices/equipment capable to reliably effect on the enemy electronic assets (control, information, communication…) and deny the enemy any further operations. The issue of electromagnetic weapons in the Czech Republic is in the stage of research and that is why the representation and involvement of the Czech representatives in the special Panels of NATO RTO or any other international organizations and boards is desirable.

REFERENCES

[1] MUNZERT, Reinhard: Targeting the Human with Directed Energy Weapons, 2002, Erlangen, Germany

[2] MERKLE, Laurence D.: Virtual Prototyping of RF Weapons, 2002, The Air Force Research Laboratory, Kirtland.

[3] COOP, Carlo: The Electro-magnetic Bomb: A Weapon of Electronic Mass Destruction, 2000, Melbourne Australia.

[4] VALOUCH, Jan: Activities of VTÚ PV Vyškov in the area of EMC and DEW, presentation of NATO RTO / SCI–132, 2002, Munster, Germany.

Selection of unclassified documents from the session of Panel NATO RTO/SCI:

[5] SCI–119 Workshop Tactical Implications of High Power Microwaves, Copenhagen, Denmark, 11—13.6.2002.

[6] Open Meeting of NATO RTO/SCI–132 on High Power Microwave to Infrastructure and Military Equipment, Munster, Germany, 22—24.10.2002.

View publication stats

James Madison University

From the SelectedWorks of George H Baker

August, 2011

High Power Electromagnetic Weapons: A Brief Tutorial

George H Baker, III, *James Madison University*

Available at: https://works.bepress.com/george_h_baker/33/

High Power Electromagnetic Weapons: A Brief Tutorial

George H. Baker

Background. High power electromagnetic weapons, also referred to as high power radiofrequency (HPRF) weapons, are a type of directed energy weapons. The system effects of high power electromagnetic environments are well recognized by world scientific and military communities. Former CIA Director John Deutch has said that, "the electron is the ultimate precision-guided weapon." In the course of the investigation ofnuclear EMP effects on electronics during the Cold War period, it became evident that garden variety, unprotected electronics would malfunction, in some cases burn out, in the presence of electromagnetic fields in the hundreds to thousands of volts per meter. The EMP experience has led to the development of non-nuclear high power electromagnetic sources to create fields that equal or exceed EMP levels, albeit over relatively small ranges. Achievable electronic effects could have serious consequences in terms of interruption or termination of critical system operation. The effects are of particular interest to the military in the context of information warfare and missile defense. Because most critical infrastructures are controlled by electronics, HPRF weapons are a concern for civilian systems as well. The weapons could be used to disrupt computer electronics controlling electric power grids, telecommunications networks, financial institution databases, security systems, and aircraft.

Military forces in many countries are pursuing the development of HPRF weapons. These programs are normally classified. HPRF weapons represent a revolutionary concept because they operate at the speed of light, can be used covertly, and harm equipment rather than humans (non-lethal). HPRF weapons may be enclosed in briefcases, packing containers, truck beds, or aircraft and maneuvered to expose critical military or infrastructure systems.

Electromagnetic Weapon Characteristics. There are three elements to be considered in discussing HPRF weapons systems: the weapon itself, the propagation of the weapon output to the target, and the target response (see Figure 1).

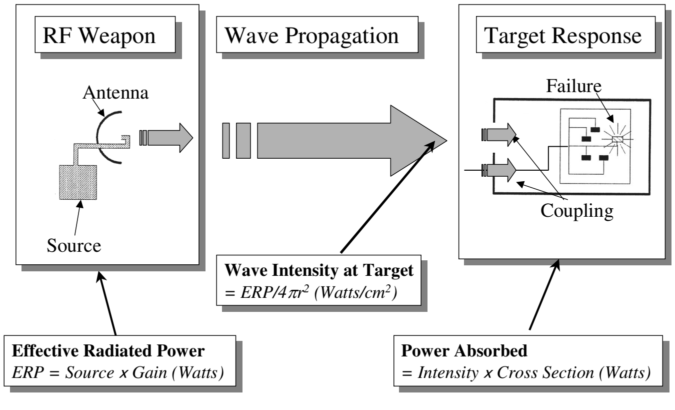

Figure 1. High Power RF Weapon Operation Elements

©2011 George H. Baker

Weapon Design. An HPRF weapon consists of a power source, or driver and a radiating antenna. The source is normally pulsed. Outputs may be sine waves (providing a narrow frequency band signal) or pulses (providing a wide frequency band signal). High power electromagnetic source technology includes electron beam devices (magnetrons, vircators, gyrotrons, backward wave oscillators), solid-state devices (bulk avalanche, optical switches, silicon carbide circuits) and explosive generators (magnetic flux compression using high explosives). Output power levels in the gigawatts are feasible over a wide range of frequencies. Design details will determine output characteristics. Power, frequency, bandwidth, repetition rate, and duty cycle are the important HPRF weapon output parameters.

The HPRF weapon source must be connected to an antenna with sufficient gain to "beam" the RF energy to a useful range. Portability requirements impose major limits on antenna area which in turn largely governs the maximum intensity (power per area) that can be delivered to a target system at a given range. Compact sources that may be moved covertly in briefcases, packing containers, truck beds, or aircraft and maneuvered for close-up exposures of critical military or infrastructure systems are of most interest. Antenna size limits largely determine the maximum intensity (power/area) that can be delivered to a target system at a given range. A table of approximate relative portable platform sizes is provided below.

Table 1. Portable Platform Size Comparisons (Approximate)

Capacity: Platform:	Volume	RF system Weight	Basis
Briefcase	0.02 m^3	5 kg	Typical hard side
Footlocker	0.15	40 kg	1 person portable
Pickup truck	4 m^3	1000 kg	1 ton capacity
Econoline van	8 m^3	2000 kg	Bed volume
Tractor trailer	80 m^3	20,000 kg	20 ton capacity

High Power Electromagnetic Wave Propagation. Electromagnetic waves propagate at the speed of light through the atmosphere. Under most conditions, the atmosphere will not attenuate HPRF waves as they travel from source to target. However, if the electromagnetic wave's peak field level exceeds the air ionization threshold a cascade process will occur in which an atmospheric plasma is created that will absorb most or all of the energy in the wave. This phenomenon is referred to as air breakdown.

A typical threshold for breakdown at sea level is approximately 1 megawatt/cm^2 . As an example, a 1 Gigawatt source radiating through a 1-foot radius circular antenna will produce an average intensity over the antenna of 0.35 megawatts/cm^2, a factor of 3 below air breakdown. Once away from the antenna, EM intensity falls off as the beam spreads (typically as $1/4\pi r^2$ where r is the distance to the target).

©2011 George H. Baker

The HPRF signal intensity on target depends on source transmission power (P_t), antenna gain, and range according to the following equation:

$$S_t = \frac{ERP}{4\pi r^2} = \frac{P_t \cdot G}{4\pi r^2} \frac{watts}{m^2}$$

The antenna gain may be approximated by $G \approx \frac{4\pi A_e}{\lambda^2}$ where A_e is the effective antenna area and λ is the wavelength of the radiated HPRF signal. This yields a simple formula for the intensity on target:

$$S_t = \frac{P_t A_e}{r^2 \lambda^2}$$

So, for example, a 1 Gigawatt source radiating at 1 GigaHz ($\lambda = 0.3$m) with a 1-m^2 antenna will produce a beam intensity of 11 kW/m^2 (or 1.1 watts/cm^2) on a target 1000 meters away. If the pulse duration were 1 microsecond, the energy fluence on target would be power x time or, 1.1 microjoules/cm^2.

Target Response. Once the electromagnetic wave signal reaches the target, its energy couples to the system in a very complex manner through various paths associated with the topology of the target system. The HPRF signal will induce currents on any external conductors (antennas, wires, etc) that penetrate to the system interior. The signal will also couple to any external metal shielding and then may reradiate to internal conductors. Signal waves will penetrate through any holes in external shielding to interior portions of the system. The main concern is the amount of energy that finds its way to critical electronic circuits, causing the system to malfunction.

To simplify coupling calculations, it is possible to determine an effective "coupling cross section" for critical internal circuits. In effect, this treats each internal circuit as a receiver. The power received by the circuit may be expressed as a simple function in incident wave power and an effective coupling cross section:

$$P_i = S_t \mathrm{A}_\sigma$$

P_i is the power induced in the internal circuit, S_t is the incident power at the system's exterior, and A_σ is the coupling cross section of the internal circuit. A_σ is a function of frequency and incorporates all the complexity of coupling including multiple paths, layers and mechanisms. There are two major coupling modes, front door and back door, as shown in figure 2.

©2011 George H. Baker

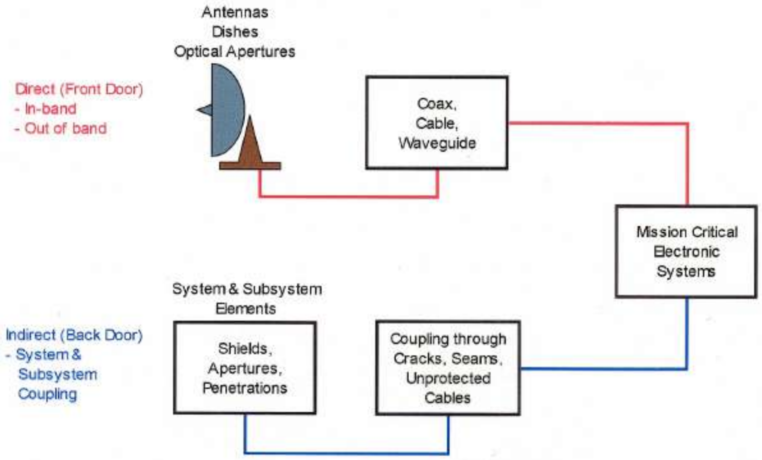
Antennas
Dishes
Optical Apertures
Direct (Front Door)
- In-band
- Out of band
Coax,
Cable,
Waveguide
Mission Critical
Electronic
Systems
System & Subsystem
Elements
Indirect (Back Door)
- System &
Subsystem
Coupling
Shields,
Apertures,
Penetrations
Coupling through
Cracks, Seams,
Unprotected
Cables

Figure 2. Front Door and Back Door Coupling Modes

The physics of front and back door coupling are the same. In each case, energy resident in the incident wave induces currents in the system that flow to a sensitive system circuit. Each can be characterized parametrically by equation $P_i = S_t A_\sigma$. The coupling cross-section, A_σ, is typically large for front door coupling and small for back door coupling. The coupling cross sections for front door coupling are of the order of the physical area of the antenna or aperture. Effective coupling cross section values for back door coupling of unhardened systems range from $10^{-4} – 10^2$ cm^2. Coupling cross-sections are hard to predict analytically and can be determined confidently only by direct measurement. This creates a problem for an attacker's confidence since HPRF effects are subject to much higher uncertainties than conventional weapon effects.

HPRF Weapon Effects on Systems. EMP and HPRF affect systems by disrupting the operation of electronic components either temporarily (upset or latchup) or permanently (component damage). Damage may be "direct" where energy inherent in the EMP/HPRF field is sufficient to cause malfunctions or "indirect" where EMP/HPRF energy triggers effects involving a system internal power supply. Even though the power delivered to a circuit by the HPRF field is small, the much larger energy available in the system's power supply (or fuel and ordnance) can be improperly diverted by an HPRF induced overvoltage arc or malfunction of system digital control circuit, causing major system damage.

Upset refers to an induced change of state in a digital circuit in which the system continues to operate, although possibly with erroneous data bit streams. Latchup refers to changes of state in digital circuits where the affected portion of the system ceases to operate until the system resets itself or (worse) a manual reset is required. In either case, system components are not directly damaged and recovery is often possible (depending on time criticality of affected function).

Exploitation of upset effects should not be discounted. In some cases they are tantamount to permanent damage. Upset may result in major system damage, e.g., a missile plunges into ocean due to a guidance system upset, or computer equipment is destroyed by an upset sprinkler system. Small, upset-level transient pulses can also act to trigger the release of energy from a system's own power supply causing components to burn out at fluences much lower than would normally cause permanent damage.

To give an indication of HPRF wave energies required to cause system effects, ITT industries (formerly Kaman Sciences) openly exposed circuit cards to 2.9 GHz microwave pulses with duration 1 microsecond. Onset of upset occurred at 1 watt/ cm^2 (equivalent to 1 microjoule/cm^2 for this pulse duration). All components were upset at wave intensities of 1000W/cm^2. Onset of damage occurred at 100W/cm^2 (for a 1 microsecond pulse, this is equivalent to 100 microjoules/cm^2). All components were damaged at wave intensities of 1000 W/ cm^2 (equivalent to 1000 microjoules/cm^2).

System Protection. System hardening involves a combination of operational and hardware techniques. Operational techniques may include the provision of spares for soft critical subsystems or boxes, disconnecting susceptible circuits upon warning, and/ or establishment of a physical keep-out perimeter (with barriers and/or security force) or zones around critical equipment to prevent positioning of HPRF weapons at close ranges. Operational controls may also be built into software to provide circumvention and reset, error-correcting codes, voting logic, and status detection. For some non time-sensitive systems, provisions for rapid system repair may be an option.

Conceptually, hardware approaches involve placing a conducting material between the incident HPRF wave and susceptible internal circuits. Hardening techniques have been successfully demonstrated and codified for EMP (ref. article on nuclear EMP). The EMP community has placed a heavy reliance on exterior shielding while limiting the number of penetrations that have to be individually protected. Such protection applied at the system exterior allows interior boxes to go untreated. This approach works well when designed in from the start. For retrofit protection, however, it is often prohibitively expensive.

Hardware approaches for HPRF protection, while conceptually similar, have some differences in emphasis from EMP. Because EMP is extremely broadband, typically only a small fraction of the energy comes through the front door in band. HPRF weapons can be tuned to the front door center frequency such that all the beam energy flows into the system. Also since HPRF weapons operate at higher frequencies, attention to smaller apertures (including cracks and seams) is required and dimensions of waveguide-beyond-cutoff penetration treatments will require changes (longer and more narrow waveguides are needed).

Front door in-band protection is one of the more challenging (but not insurmountable) HPRF protection problems. The high gains associated with most front door paths make these potentially the most susceptible portion of the system. However these well-characterized front door receive paths have received much attention in terms of protection engineering. Radar systems are often protected from their own or neighboring transmitters by a receiver protector or RP. Similar protection can be applied to communication receivers against in band HPRF environments. Table 2 summarizes HPRF hardening techniques.

©2011 George H. Baker

Table 2. HPRF Hardening Methods

Operational Techniques	Hardware Techniques	
	Front Door	Back Door
Disconnect on warning	Tuned metallic radomes	Shielding, topology control
Spares	Bandstop reflectors	Internal cable shields
Physical keep-out perimeters or zones	Antenna gain pattern control	Non corrosive mating surfaces
Software Techniques: Circumvention/reset Error correcting codes Voting logic Status detection/ alarm	Filters	Closely spaced fasteners
	TR devices/ limiters	RF gaskets and seals
	Terminal protection devices	Aperture screens and conductive films
		Selection of hard circuits and components
		Pin protection
		Fiber Optics
		Filtering, limiting at penetration points

Future Directions. HPRF generation techniques have matured to the point where practical devices have become technically feasible. Miniaturization of pulse-power source components, more efficient power supplies, and advances in electronic pulse forming, energy conversion, and antennas enable reduced size and increased efficiency of HPRF generation hardware. Simple weapons can also be built using inexpensive magnetrons from common microwave ovens.

Military trends to computerized battle management, weapons tracking, and landline/wireless network communications are a double-edged sword, introducing serious HPRF vulnerabilities. Military use of commercial off-the-shelf equipment and dependence on civilian infrastructure exacerbate the problem. Senior military officials have dropped hints about pursuing offense technology but there are no officially published details concerning weapon availability or capabilities.

Several other countries also have extensive background in the development of RF weaponry. The former Soviet Union pioneered the development of HPRF weapon technology and this technology is now being offered to other countries. According to a recent report from the Office of the US Secretary of Defense on the military power of China, “Captain Shen Zhongchang from the Chinese Navy Research Institute…envisions a weaker military defeating a superior one by attacking its space-based communications and surveillance systems…in future wars, Shen highlights radar, radio stations, communications facilities, and command ships as priority targets vulnerable to smart weapons, electronic attack, and electromagnetic pulse weapons.” It is expected that in future conflicts the United States will encounter adversaries using HPRF weapons as part of asymmetric tactics to disrupt information systems.

©2011 George H. Baker

See also: Nuclear Electromagnetic Pulse (EMP)

References and Further Reading

Abrams, M., "The Dawn of the E-bomb," IEEE Spectrum, October, 2003.

Backstrom, M. et al; *Preliminary Study Regarding the Resistance of Critical Societal Systems to High Intensity Electromagnetic Radiation*; Royal Swedish Defense Research Agency; Report FOA-R-97-00538-612-SE; August, 1997.

Barker, R., Schamiloglu, E., *High-Power Microwave Sources and Technologies,*Wiley-IEEE Press, 2001.

Benford, J., Segle, J., *High Power Microwaves*, 2nd Edition, Institute of Physics, 2004.

Guyatt, D. G., "Some Aspects of Antipersonnel Electromagnetic Weapons," ICRC Symposium on the Medical Profession and the Effects of Weapons, February, 1996.

Kopp, C., "The E-Bomb – A Weapon of Electrical Mass Destruction," INFOWAR Conference, Washington, D.C. September 1996.

Lucier, J. P., "Backyard Terrorism - E-Zapper Could Break the Bank,"
Insight Magazine, April 1, 2002.

McKenna, T., "Flash in the Plan," Journal of Electronic Defence, May 2003.

Price, D. et al, "Compact Pulse Power for Directed Energy Weapons," Journal of Directed Energy, Volume 1, Fall 2003.

Saxton, J., Chairman, *Congressional Hearing on Electromagnetic Weapons*, Joint Economic Committee, Intelligence and Security, Source: http://www.house.gov/jec/hearings/02-25-8h.htm, June 17, 1997.

National Research Council, Technology for the United States Navy and Marine Corps, 2000-2035, Volume 2: Technology, National Academy Press, Washington, DC, 1997.

About the author: George Baker is Professor of Integrated Science and Technology at James Madison University. He also serves as Technical Director of the University's Institute for Infrastructure and Information Assurance (IIIA). He is a consultant in the areas of critical infrastructure assurance, high power electromagnetics, nuclear and directed energy weapon effects, and risk assessment. He served as senior staff on the Congressional EMP Commission. Baker is former director of the Defense Threat Reduction Agency's Springfield Research Facility, a national center for critical system vulnerability assessment. Much of his career was spent at the Defense Nuclear Agency (DNA) as the Integrated Electromagnetics Team Leader directing the development of

©2011 George H. Baker

electromagnetic protection, underground testing, and standards including MIL-STD-188-125, MIL-STD-2169B, and MIL-HDBK-423. He also headed the Agency's Innovative Concepts Division overseeing the joint US-Russian space nuclear power technology, electro-thermal chemical (ETC) gun development, radiofrequency directed energy source concept development and testing, and DNA's university grants programs. Baker received the Agency Legacy Award for his leadership and innovation. He is a member of the NDIA Homeland Security Executive Board, the Institute of Electrical and Electronic Engineers, the Directed Energy Professional Society (Charter Member), and the Association of Old Crows. He is a Summa Foundation EMP Fellow and holds a Ph.D. from the U.S. Air Force Institute of Technology.

©2011 George H. Baker

archived as **http://www.stealthskater.com/Documents/Hamilton_04.pdf**

more of Dan Burisch is at ***http://www.stealthskater.com/Burisch.htm***

note: because important websites are frequently "here today but gone tomorrow", the following was archived from http://www.skywatch-research.org/QuantumHolographic.htm on April 8, 2004 . This is NOT an attempt to divert readers from the aforementioned web-site. Indeed, the reader should only read this back-up copy if it cannot be found at the original author's site.

The Biophotonic Quantum Holographic Matrix

by **William F. Hamilton III**

ABSTRACT: *This paper is an attempt to integrate studies of the quantum potential, quantum holography, biophotonics and the enveloping matrix of biointegration, biocommunication, and bioinformation that composes the web of light and life in living organisms and their possible genesis in an electromagnetic infoton.*

Introduction:

Theoretical physicist David Bohm proposed a new hidden variables theory of Quantum Mechanics. He assumes the wave function does not represent just a set of probabilities but an actual field. This field exists and acts upon particles the same way a classical potential does. Thus it is a quantum potential which is associated to this field and is a function of the wave function. It is postulated here that the quantum potential is active -- as a field -- in living cells and organisms and acts upon them.

The equation Bohm used to represent this quantum potential is:

$$U = -\sum_k \left(\hbar^2 / 2m_k\right)\left(\partial_k^2 R / R\right) \qquad \textbf{(1)}$$

Research reveals that Bohmian quantum potential is still used, but that it may only represent a means to extend Quantum Theory into a realistic theory of particles and fields.

It resolves the dilemma of the appearance -- in one and the same phenomenon -- of both particle and wave properties in a rather straightforward manner. Bohmian mechanics is a theory of motion describing a particle (or particles) guided by a wave. This guiding wave is usually attributed to atomic particles. But it might be more insightfully seen as a traveling **soliton** or vortex.

Bohmian mechanics is manifestly nonlocal, and the velocity may depend upon the positions of other -- even distant -- particles whenever the wave function of the system is entangled.

The behavior of the particle is determined by the particle's position and momentum, the wave field, and the sub-quantum fluctuations.

Bohm's quantum potential binds the entire Universe together into what he liked to call a seamless "unbroken wholeness". Every particle in the Universe is connected by the quantum potential to every other particle. He likened the cosmos to a **hologram** in which each point on the film carries information about the entire picture. Bohm's GWT -- far more sophisticated than de Broglie's crude version -- is a "holistic" vision in which all parts of the Universe are joined to every other part. "Interconnectedness" was one of Bohm's favorite words. He saw the Universe as resembling the unity of a living organism. A kind of pantheism not unlike Spinoza's -- a pantheism Einstein himself favored.” **[2]**

Bohm referred the holomovement as the basic reality. The pattern formed on a photographic plate by light waves that interfere and create a hologram. This pattern is the interaction or interference pattern of two parts of a laser light (coherent light). One beam reflects off an object, the other off a mirror. In addition, any portion of the holographic plate contains information on the whole object.

Biohologram:

Every cell in the human body contains nuclear DNA. Information on the whole body -- much like a hologram -- only the cells can go through a process of division and reproduction (mitosis).

In quantum holography, things are spookier still. While holograms are typically constructed with interfering beams of light, in quantum holography the researchers measure the simultaneous arrivals of an illuminating photon that is sent into a chamber and a companion photon in the other entangled beam.**[2]**

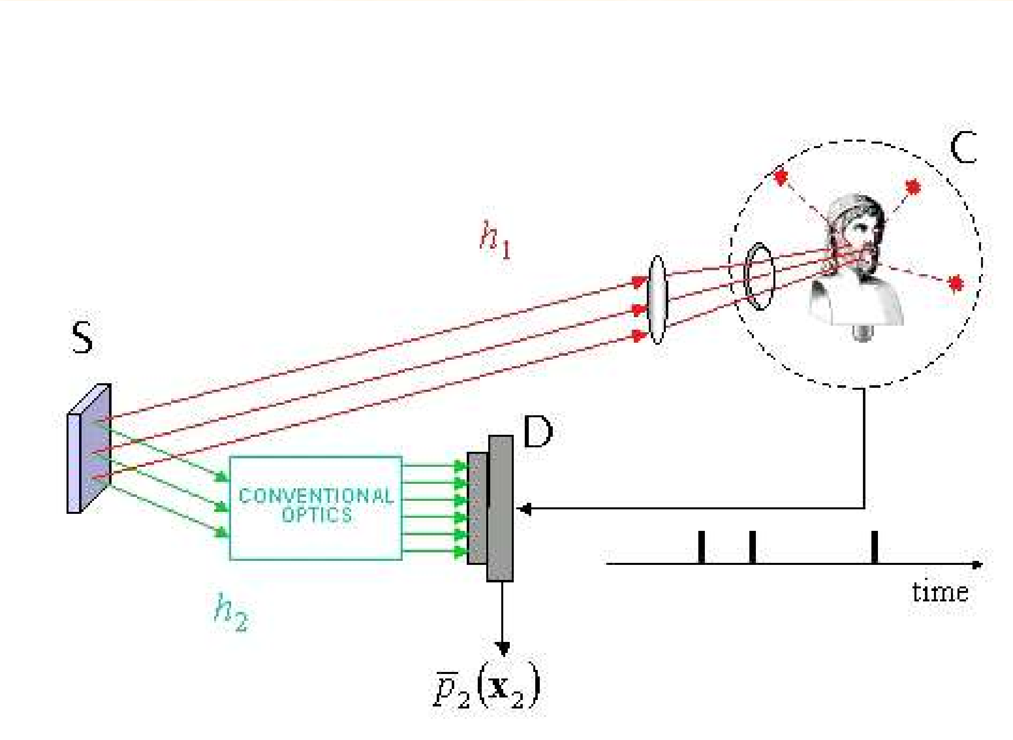

Quantum holography -- as depicted in this cartoon -- is a way that physicists could produce 3-dimensional images of objects that would be invisible to systems relying on classical physics. A source **S** produces 2 photon beams (**h_1** and **h_2**). The photons in **h_1** are quantum mechanically **entangled** with the photons in **h_2**. When a photon in **h_1** enters chamber **C**, it encodes 3-dimension information about the enclosed object (in this case a Grecian bust) in the interference of the 2 paths it could take -- a path that intercepts the object and subsequently is reflected to the chamber wall, and a path that misses the object and strikes the chamber wall directly. The second beam **h_2** passes through conventional optics and strikes a detector array **D**. A coincidence counter extracts the holographic information by monitoring the relative time between a photon in **h_1** striking the chamber wall and its entangled mate in **h_2** arriving at detector **D**.

Wave Genetics Inc. and the Moscow Institute of Control Sciences are exploring a new concept of the genetic code that asserts:

(1) that the evolution of biosystems has created genetic "texts" -- similar to natural context dependent texts in human languages -- shaping the text of these speech-like patterns.

2) that the chromosome apparatus acts simultaneously both as a source and receiver of these genetic texts, respectively decoding-and-encoding them.

3) that the chromosome continuum of multicellular organisms is analogous to a static-dynamical multiplex time-space holographic grating, which comprises the space-time of an organism in a convoluted form.

That is to say, the DNA action -- theory predicts and which experiment confirms -- is …

(i) that of a "gene-sign" laser and its solitonic electro-acoustic fields, such that the gene-biocomputer "reads and understands" these texts in a manner similar to human thinking but at its own genomic level of "reasoning". It asserts that natural human texts (irrespectively of the language used) and genetic "texts" have similar mathematical-linguistic and entropic-statistic characteristics, where these concern the fractality of the distribution of the character frequency density in the natural and genetic texts and where -- in case of genetic "texts" -- the characters are identified with the nucleotides, and

(ii) that DNA molecules -- conceived as a gene-sign continuum of any biosystem -- are able to form holographic pre-images of biostructures and of the organism as a whole as a registry of dynamical "wave copies" (or "**matrixes**") succeeding each other. This continuum is the measuring, calibrating field for constructing its biosystem.

“Complex information can be encoded in EM fields, as we all know from coding-and-decoding of television and radio signals. Even more complex information can be encoded in holographic images. DNA acts as a holographic projector of acoustic and EM information which contains the informational quintessence of the biohologram. Only 3% of human DNA encodes the physical body. The remaining 97% of the 3 billion base pair genome contains over a million genetic structures -- called transposons -- that have the capacity to jump from one chromosomal location to another (Kelleher, 1999). We are 99.9% alike in our genetic legacy. Our individuality is expressed in 3 million small variations in our cells called single nucleotide polymorphisms.

"**Gene-expression** is the mechanism by which new patterns are called into being (Rossi, 2000). There is also a strong correlation between modulation of the brain’s EM field and consciousness (Persinger, 1987; McFadden, 2002). The Gariaev group has discovered a wave-based genome and DNA phantom effect which strongly supports the holographic concept of reality (Miller, Webb, Dickson, 1975). This main information channel of DNA is the same for both photons and radio waves. Superposed coherent waves of different types in the cells interact to form diffraction patterns, first in the acoustic domain and second in the electromagnetic domain -- a **quantum hologram** -- a translation process between acoustical and optical holograms.” **[4]**

The distribution of the character frequency in genetic texts is fractal, so the nucleotides of DNA molecules are able to form holographic pre-images of biostructures. This process of "reading and writing" the very matter of our being manifests from the genome's associative holographic memory in conjunction with its quantum nonlocality. Rapid transmission of genetic information and gene-expression unite the organism as holistic entity embedded in the larger Whole. The system works as a biocomputer -- a **wave biocomputer**.

The quantum nonlocality of the genetic information is fundamental. Experimental work of the Gariaev group shows how quantum nonlocality is directly related to laser radiation from chromosomes (coherent light), which jitterbugs its polarization plane to radiate or occlude photons. DNA and the genome have now been identified as active "laser-like" environments. Roughly speaking, DNA can be considered as a liquid crystal gel-like state that acts on the incoming light in the manner of a solitonic lattice. **[5]**

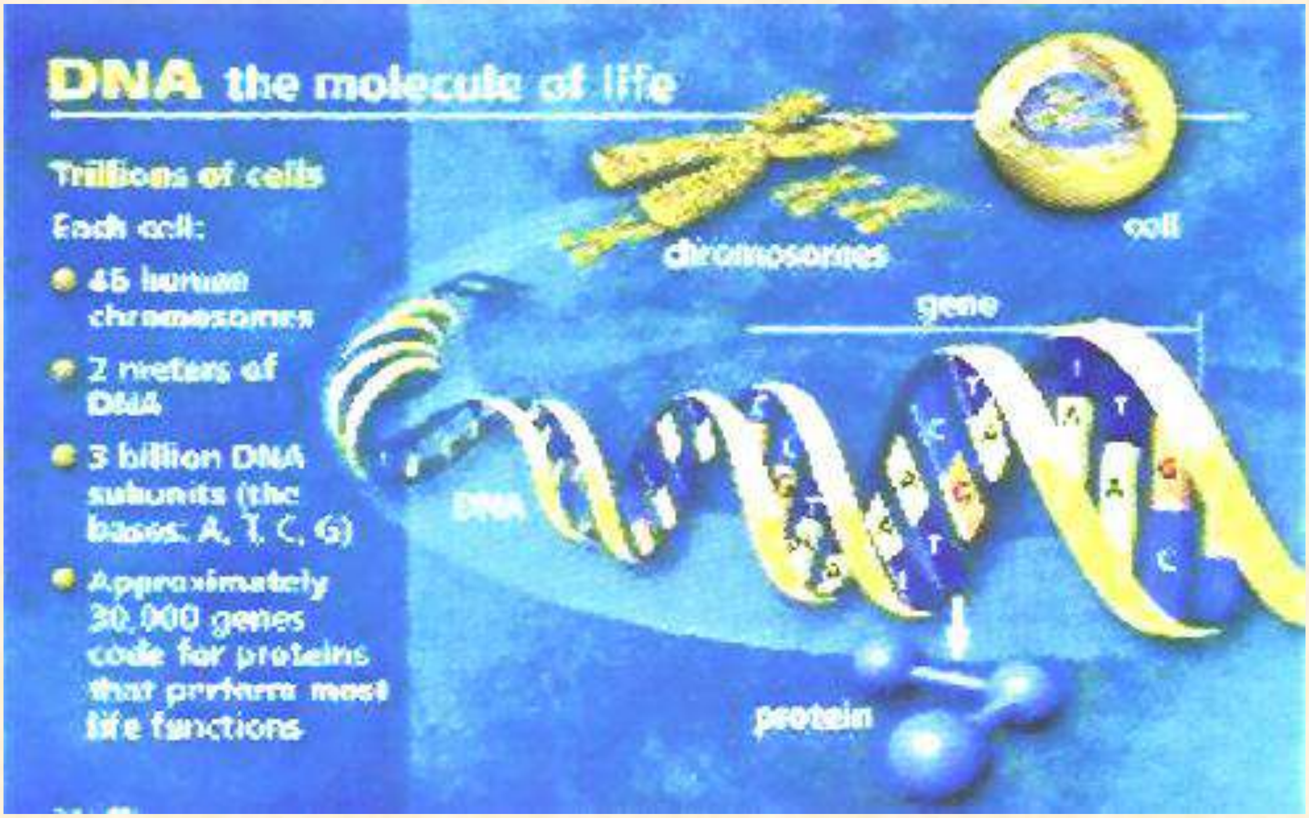
DNA the molecule of life
Trillions of cells
Each cell:
46 human chromosomes
2 meters of DNA
3 billion DNA subunits (the bases: A, T, C, G)
Approximately 30,000 genes code for proteins that perform most life functions
chromosomes
cell
gene
DNA
protein

Biophotonics

This entire section is extracted from the History of Biophotonics website:

Around 1923, Alexander Gurwitsch discovers an "ultra-weak" photon emission from living systems (onions, yeast, ...), since he suggested connections between photon emission and cell division rate. He calls this **photonemission "mitogenetic radiation**". His experiments indicate that the wavelength is in the range around 260 nm (Bibliography under Gurwisch and also Ruth (1977, 1979)).

Around 1950, Russian scientists rediscover "ultra-weak photon emission" from living organisms. Most results are published in "Biophysics" (Engl.) and originally in "Biofizika"). (Bibliography under Ruth, 1979).

Italian nuclear physicists discover by chance "bioluminescence" of seedlings. They do not think that this finding is significant, but they publish the results (Colli et al. 1954, 1955, Ruth 1979).

The Russian biophysicist and the American chemist enunciate the first theory of ultra-weak photonemission (UWPE) from biological systems -- the so called "**imperfection theory**". UWPE shall be an expression of the deviation from equilibrium -- some kind of distortion of metabolic processes (Zhuravlev 1972, Seliger 1975, Ruth 1979).

Independently from each other and by different motivations, scientific groups in Australia (Quickenden), Germany (Fritz-Albert Popp), Japan (Inaba), and Poland (Slawinski) show evidence of ultra-weak photon emission from biological systems by use of modern single-photon counting systems. Bibliography (Quickenden, Inaba, Popp &Ruth, Slawinski).

While Quickenden, Slawinski, and Inaba prefer the imperfection theory, Popp and his group enunciate just the opposite theory:

1. The radiation originates from an almost perfect coherent photon field.

2. Essential sources are the DNA and corresponding resonators in the cells.

3. The mechanism describes photon storage in cavities and information channels, tuned by Casimir forces.

4. There is a close connection to delayed luminescence which corresponds to excited states of the coherent photon field.

5. The radiation is not the product but essentially the initiator of chemical reactions in the cells. The radiation submits the information within and between cells.

6. The radiation is not limited to the optical range but follows an **f = const** rule (the occupation probability of the phase space is equal for all wavelengths) and extends to longer wavelengths including the so-called heat radiation of the body.

7. This radiation is the proper regulator and information carrier of life.

The Marburg group of Fritz-Albert Popp calls this phenomenon "**biophotons**" in order to stress the difference to "bioluminescence". Biophotons are single quanta which are permanently-and-continuously emitted by all living systems. They are subjects of quantum physics, and they display a universal phenomenon attributed to all living systems. Worldwide, all scientists who agree with these statements call the radiation "biophotons" and the scientific field "**biophotonics**".

From 1972 to 1980, the Marburg group of the leader -- the physicist and Dr. Habil Fritz-Albert Popp -- evaluated experimentally all the essential physical properties of biophotons.

1. The intensity ranges from a few up to some hundreds photons/(sec-cm^2).

2. The spectral distribution follows in the time average an **f = const** rule.

3. The modes are strongly coupled.

4. The delayed luminescence that approaches continuously the biophoton emission follows a hyperbolic rather than an exponential relaxation function.

5. The biophotons origin from an almost fully coherent field.

6. Cells are able to establish cavity resonators which contribute to biophoton regulation.

7. The essential source of non-equilibrium biophoton emission is the DNA.

This group introduces the first time photocount statistics (PCS) into biophotonics. They show evidence that biophotons are emitted according to a Poissonian PCS. Furthermore, they show evidence that (1) the delayed luminescence follows a hyperbolic relaxation function rather than an exponential one; (2) the modes are strongly coupled; and (3) there are hyperbolic oscillations around the continuous hyperbolic relaxation function. The group finds the first time intercellular communication by means of biophotons. Later this was confirmed by Albrecht-Bühler (Bacteria), Popp and Chang (dinoflagellates), Galle (daphnia), Shen (blood), and Vogel (bacteria).

(Bibliography under Popp and coworkers, i.e., Bahr, Böhm, Grass, Grolig, Herrmann, Kramer, Rattemeyer, Ruth, Schmidt, Wulle, Albrecht-Bühler, Chang, Galle, Shen, Vogel).

The papers of Popp and his group are examined mainly by the group of B. Chwirot (Kopernikus University, Torun). They confirm the essential results (Bibliography Chwirot et al.).

Herbert Klima (Atom Institute Vienna) performs his dissertation in Popp`s group at the University in Marburg. He transfers "Biophotonics" to the University in Vienna -- in particular investigations on laser excitation of living systems.

(Bibliography Klima or Atominstitut Wien).

J.Slawinski cooperates with the groups in Japan, USA and the Popp-group in Germany. He follows mainly the links between biophotons and biochemical reactions. There arises a branch biochemical biophotonics that becomes an essential part mainly in Japan and USA.” **[6]**

Bioinfotons

This term “**bioinfoton**” is one that I am adopting to specify any particle (such as the **Ganesh Particle***), wave, or biomolecule that transfers information from an emitter to a receptor. It is possible that if a biophoton is emitted from a coherent source that it will carry holographic information as a form of modulation (possibly frequency-modulated) and produce action or results in biological systems.

It is possible that biophotons may even be carriers of ***psi*** information and that a coherent coupling can be established between 2 conscious lifeforms resulting in a transference of information from a higher potential field to a field at lower potential. Even though this is conjecture at this point, it suggests other means by which disparate lifeforms can communicate. We know there is a process by which trees communicate and even signal each other in the face of danger. We may be exchanging information with pets and other animals through biophotonic communication. This may be how some people who have a "green thumb" affect plants they care for.

Even communication with other intelligent life forms in the Cosmos may be effected by the amplification and vectoring of biophotons or -- via the quantum potential -- produce information at a distance.

A virtual photon flux in the quantum potential may be integrating the coordinated wholeness of the organism while the action of bioinfotons may keep every part of the organism in communication with every other part.

Conclusions

Life and living organisms are more than the sum of their parts and are composed of material and non-material parts. The spiral form of DNA evidently produces an electromagnetic wave field that informs the organism’s growth and structure.

Scientist Richard Alan Miller is a pioneer in this field and has written:

“In a hologram, wave fields interfere with one another to lay the foundations for the reconstruction of the image of an object. But how are the wave fields produced?

The term "holography" comes from the Greek roots meaning "entire" and "to write". In holography, the image is projected by a coherent light source split into both the object wave and the reference wave background. This dichotomous nature is reflected in the particle/wave nature of the DNA molecule, which can be "read out" with biophotons from chromosomes to set up a holographically-produced wave field. This superposition of wave fields (object wave and reference wave) creates a wave guide for the formation of biological structure. The image is constructed according to the reference information contained in the genes. The reconstructed object wave is identical with the object wave field. The reconstructed wave fields reproduce exactly the recorded ones (the DNA with genetic code)." **[7]**

Mr. Miller also proposes this hypothesis:

> **“QUANTUM BIOHOLOGRAPHY**
>
> **Hypothesis: The organization of any biological system is established by a complex electrodynamical field that is, in part, determined by its atomic physiochemical components. These, in part, determine the behavior and orientation of these components. This dynamic is mediated through wave-based genomes wherein DNA functions as the holographic projector of the psychophysical system -- a quantum biohologram."**

I propose that the biohologram is embedded as an integral part of the biocosmic hologram we call the"Universe".

Dr. Dan Burisch has seen evidence that a **Ganesh Particle*** can transmute itself into a fully functional atypical eukaryotic cell in a matter of seconds -- not eons. Such evidence -- when it becomes available -- will change our entire perspective on the existence of a universal holonomic mind that informs its creation.

**Ganesh Particles are charged clusters that extracted from ancient bedrock and have a seemingly unknown origin, but carry information which is transferred to cells via a crossbridge. (discovered by Dr. Dan B.C. Burisch)*

References:

1. **http://plato.stanford.edu/entries/qm-bohm/#qp**
2. **http://www.ksharpe.com/Word/BM05.htm**

3. **http://www.aip.org/enews/physnews/2001/split/566-1.html**
4. **http://www.rialian.com/rnboyd/dna-wave.doc**
5. **http://www.geocities.com/iona_m/Chaosophy4/bioholography/bioholography.htm**
6. **http://www.lifescientists.de/history.htm**
7. **http://www.nwbotanicals.org/oak/newphysics/Helix%20to%20Hologram.pdf**

if on the Internet, Press <BACK> on your browser to return to the previous page (or go to www.stealthskater.com)

else if accessing these files from the CD in a MS-Word session, simply <CLOSE> this file's window-session; the previous window-session should still remain 'active'

The Tom Bearden Website

Help support the research

Foundations of Electro-Magnetism

Slide Series

- Major Points of Presentation
- Chung's Carbon Filament Negative Resistor
- Aharanov-Bohm Effect
- Absorption and emission reactions (1)
- Absorption and emission reactions (2)
- Anti-engine for cell's deviation reverses cell back to normal
- Asymmetric regauging produces excess force, which can be used to do work on the system
- Asymmetric regauging produces excess force, which can be used to do work on the regauging system
- Becker's bone fracture healing experiments
- Becker's theoretical control system governing regeneration
- Becker's theoretical DC control system involved with response to injury
- Geometrical Distortion of Poynting Energy Flow
- Cosmic Bootstrap: Before the Big Bang
- Bridging Enables COP > 1

- Aspects of Strong Local Symmetry
- Other Aspects of Strong Local Symmetry
- Bunge on the Status of Electrodynamics
- Bunge on the Status of Electrodynamics and Physics
- Bunge on the Status of Classical and Quantum Physics
- Fractions of BW Casualties Surviving
- First Step in Self-Promotion of Cancer Due to Hypoxia
- Differentiation and Dedifferentiation Timelines
- Evolution of Cells on Earth
- A Charged Particle q is a Coupled System of m and Ø
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (1)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (2)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (3)
- A Charged Particle is a Coupled System
- The Anomalies in Navy Electrolyte Experiments at China Lake
- Chung's Carbon Filament Negative Resistor

- Typical circuit has about 10^{-13} energy collection efficiency
- Classical View of EM
- Curved Spacetime acts as source or sink
- Chronic Disease Puzzling
- Cancer Characteristics
- Solving a Cancer Enigma
- Cancer: Characteristics
- Nonlinear Optics Distortion Correction Theorem
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)
- In Time-Reversed Zone: Nuclear reactions are biased
- In Time-Reversed Zone: Dynamics may be reversed
- Time-Reversed Zone: Significant only after time-density charging
- A New Conservation of Energy Law
- Questions on Carcinogens and EM

Radiation

- AS DEFINED, fields and potentials only exist in and on charged matter
- Serious flaws and errors in classical EM theory
- Electrodynamicists' Reaction to Removal of the Material Ether
- Reaction to Loss of Material Ether (1)
- Reaction to Loss of Material Ether (2)
- Some EM waves in Spacetime
- Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String
- Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna
- Phase Conjugate Wavepairs Produce New Waves
- Scalar Electromagnetics (Energetics) View of EM (1)
- Scalar Electromagnetics (Energetics) View of EM (2)

- Scalar Electromagnetics (Energetics) View of EM (3)
- Extended General Relativity Principle
- Extension of Work-Energy Theorem
- 34 Flaws in Classical EM Theory
- MTW's General Relativity Principle
- Use of General Relativity (GR) in Particle Physics
- The Missing Infolded Electrodynamics
- Infolded Engines: Excluded by Present Physics
- Mass curves and structures spacetime; ST curvatures change and structure mass (1)
- Mass curves and structures spacetime; ST curvatures change and structure mass (2)
- At Infinite Velocity: Each Is Every Other
- Lisitsyn's Report: Brain Code Broken
- Lorentz Closed Surface Integration
- Lorentz's physically insignificant energy flow can be collected and utilized
- Lorentz surface integration of the Poynting vector around a closed surface
- Longitudinal EM Photon Interaction with Charge
- Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass.

- Pumping with Longitudinal EM waves Time-reverses the Mass
- Maxwell's Quaternion Theory
- Maxwellian Systems before and after Regauging
- Maxwell's four equations reduce to: Page 1
- Maxwell's four equations reduce to: Page 2
- More to come

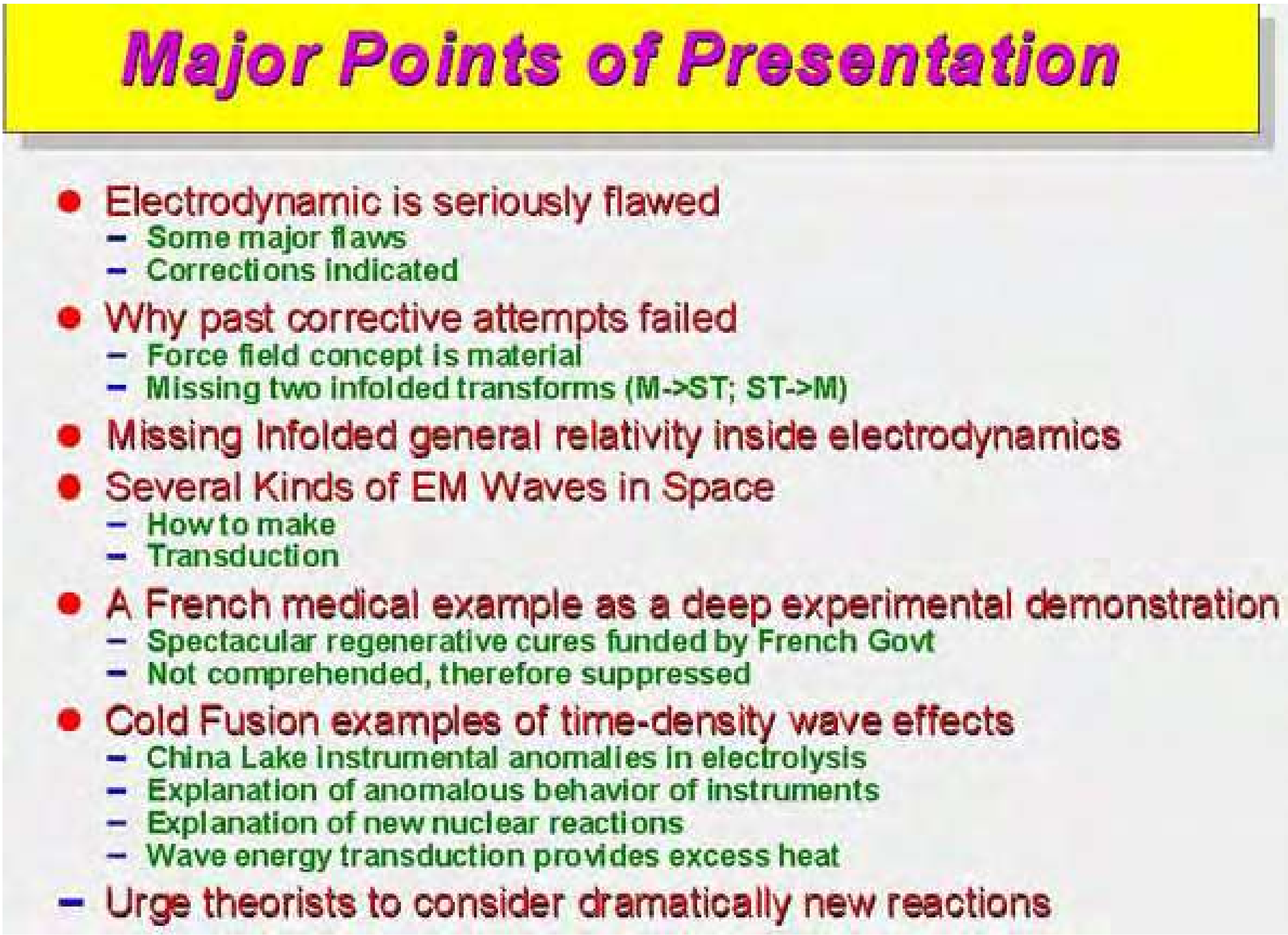

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

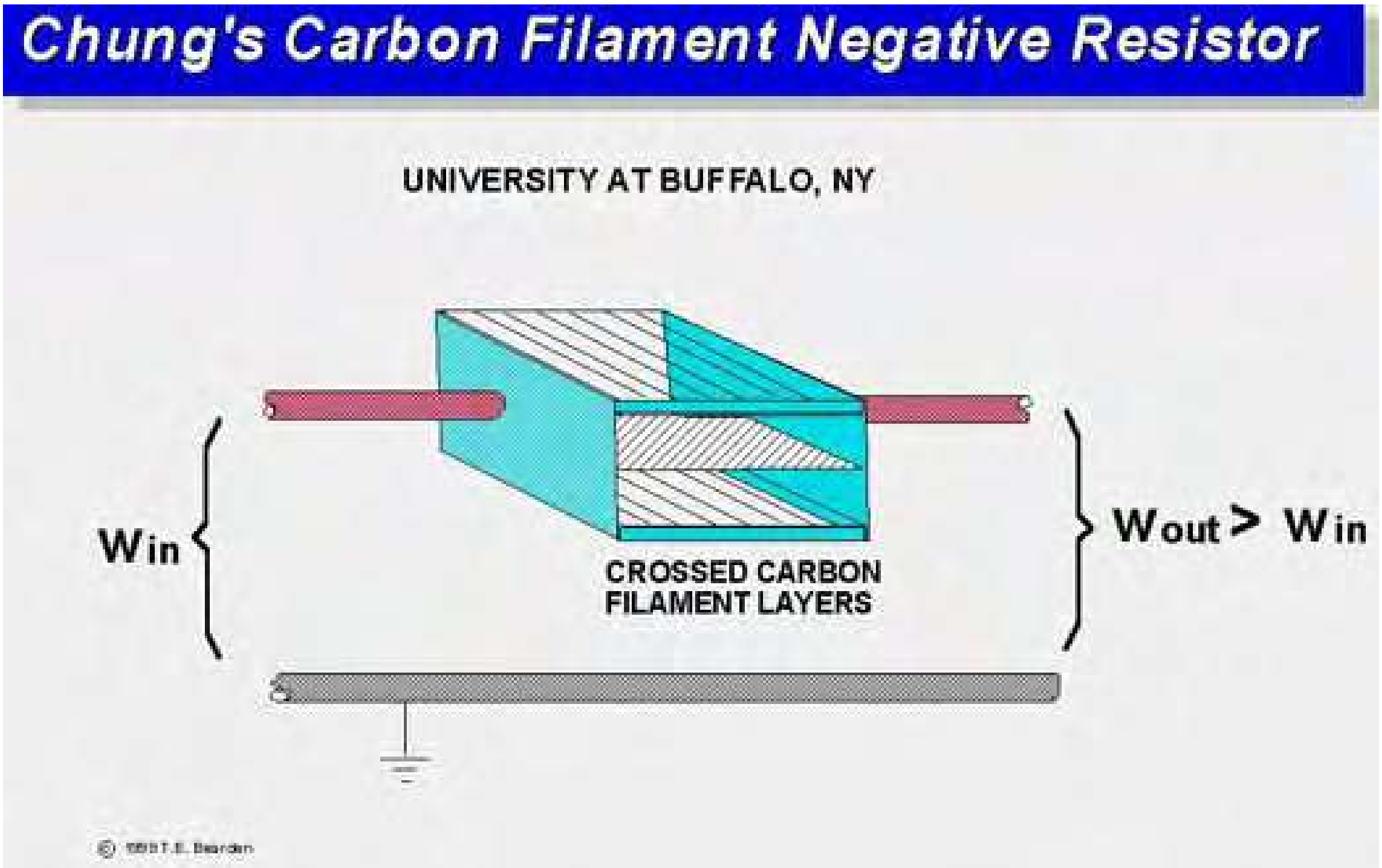

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

W_{in}

CROSSED CARBON FILAMENT LAYERS

$W_{out} > W_{in}$

© 1998 T.E. Bearden

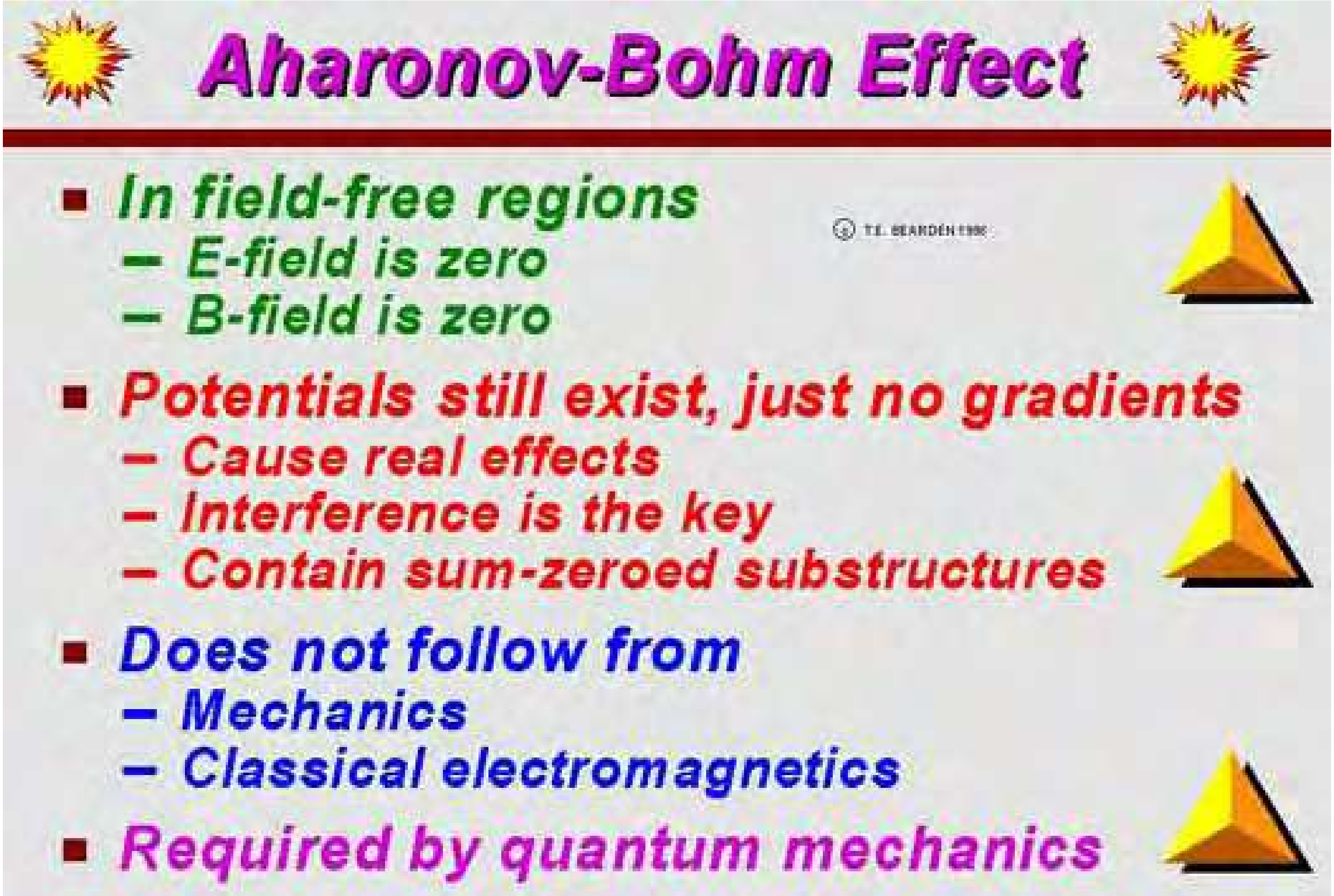

Aharonov-Bohm Effect

© T.E. BEARDEN 1998

- In field-free regions
  - E-field is zero
  - B-field is zero
- Potentials still exist, just no gradients
  - Cause real effects
  - Interference is the key
  - Contain sum-zeroed substructures
- Does not follow from
  - Mechanics
  - Classical electromagnetics
- Required by quantum mechanics

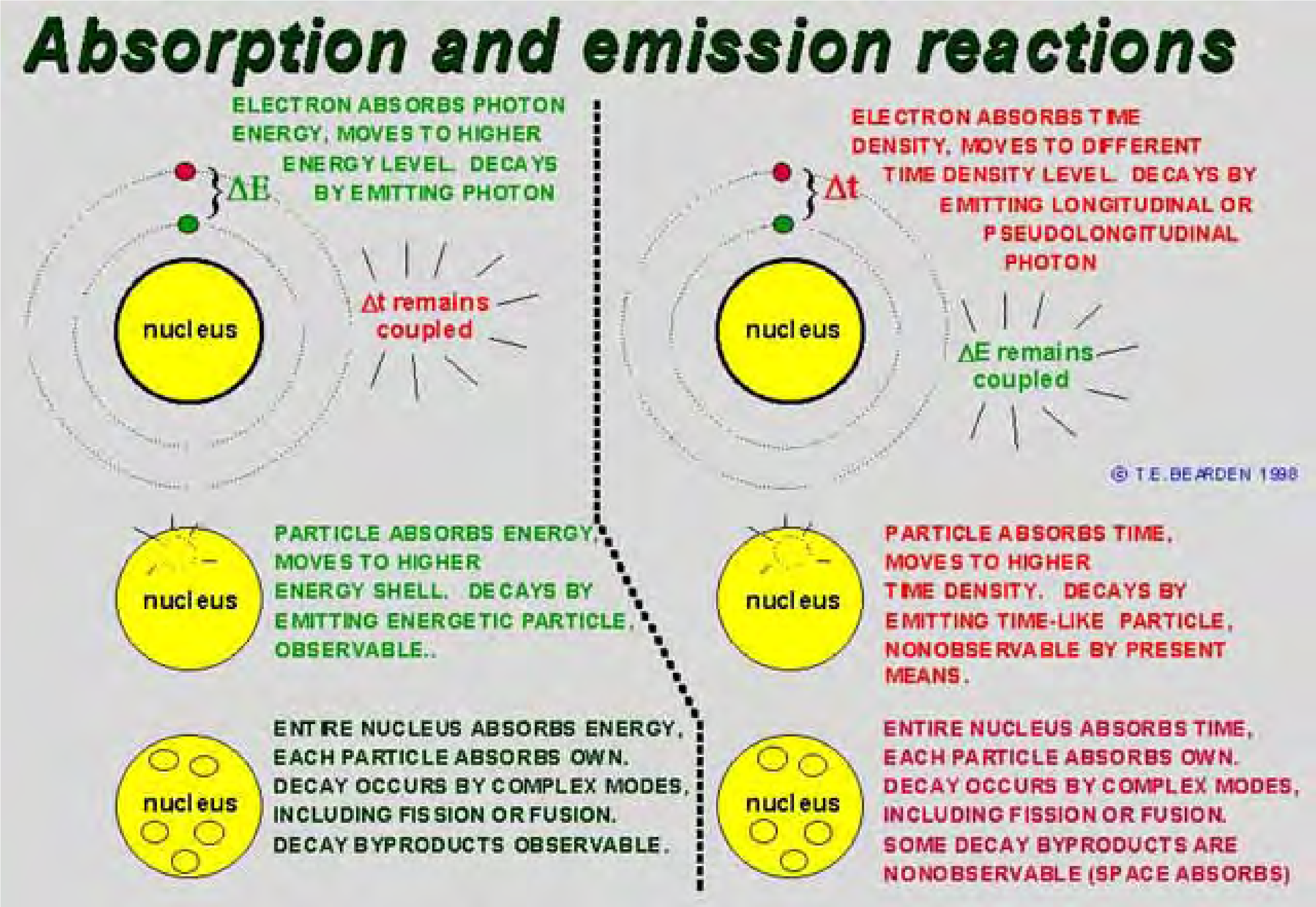

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

© T.E. BEARDEN 1998

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE, OBSERVABLE..

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

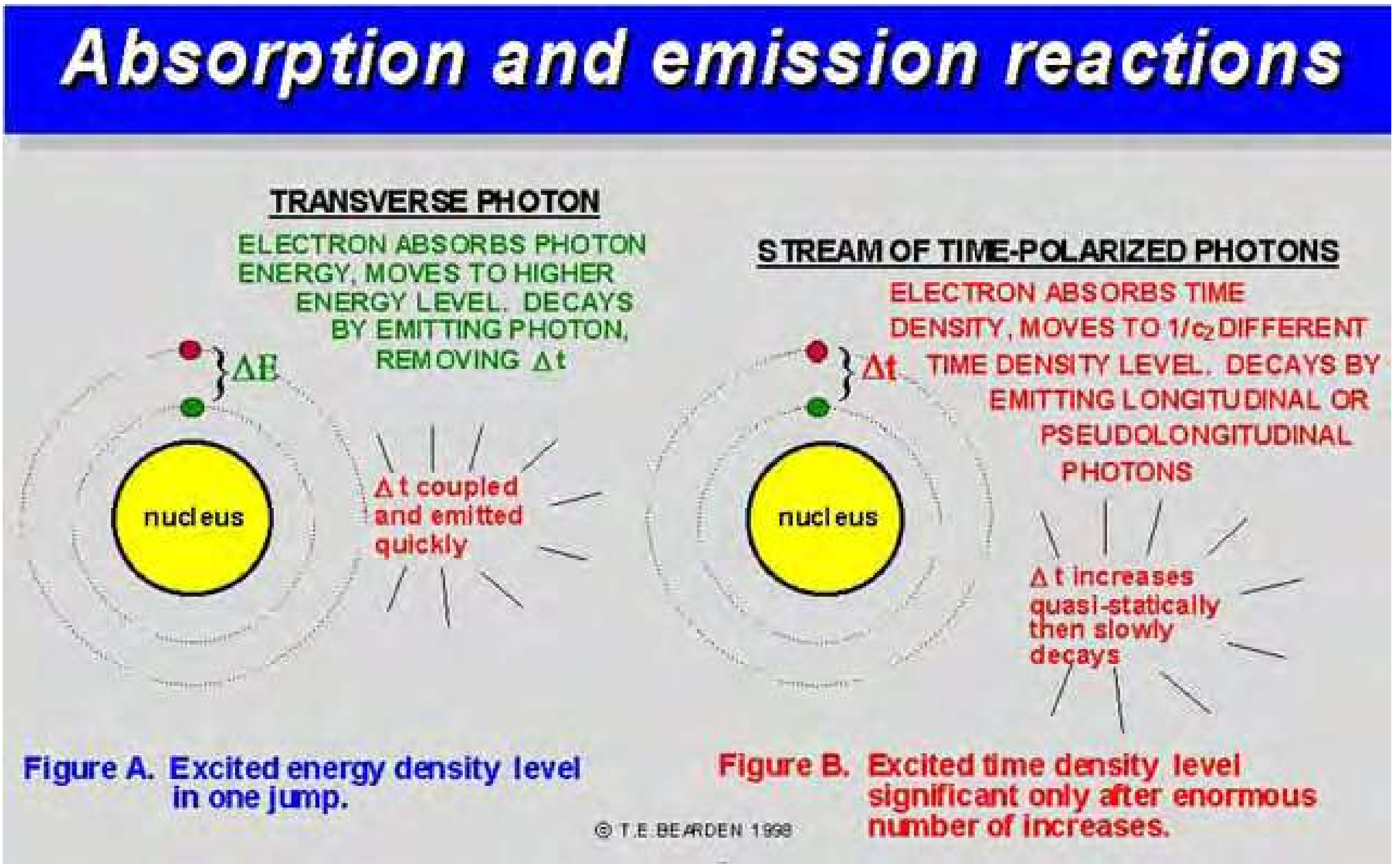

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

Δt coupled and emitted quickly

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c_2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt increases quasi-statically then slowly decays

Figure A. Excited energy density level in one jump.

Figure B. Excited time density level significant only after enormous number of increases.

© T.E.BEARDEN 1998

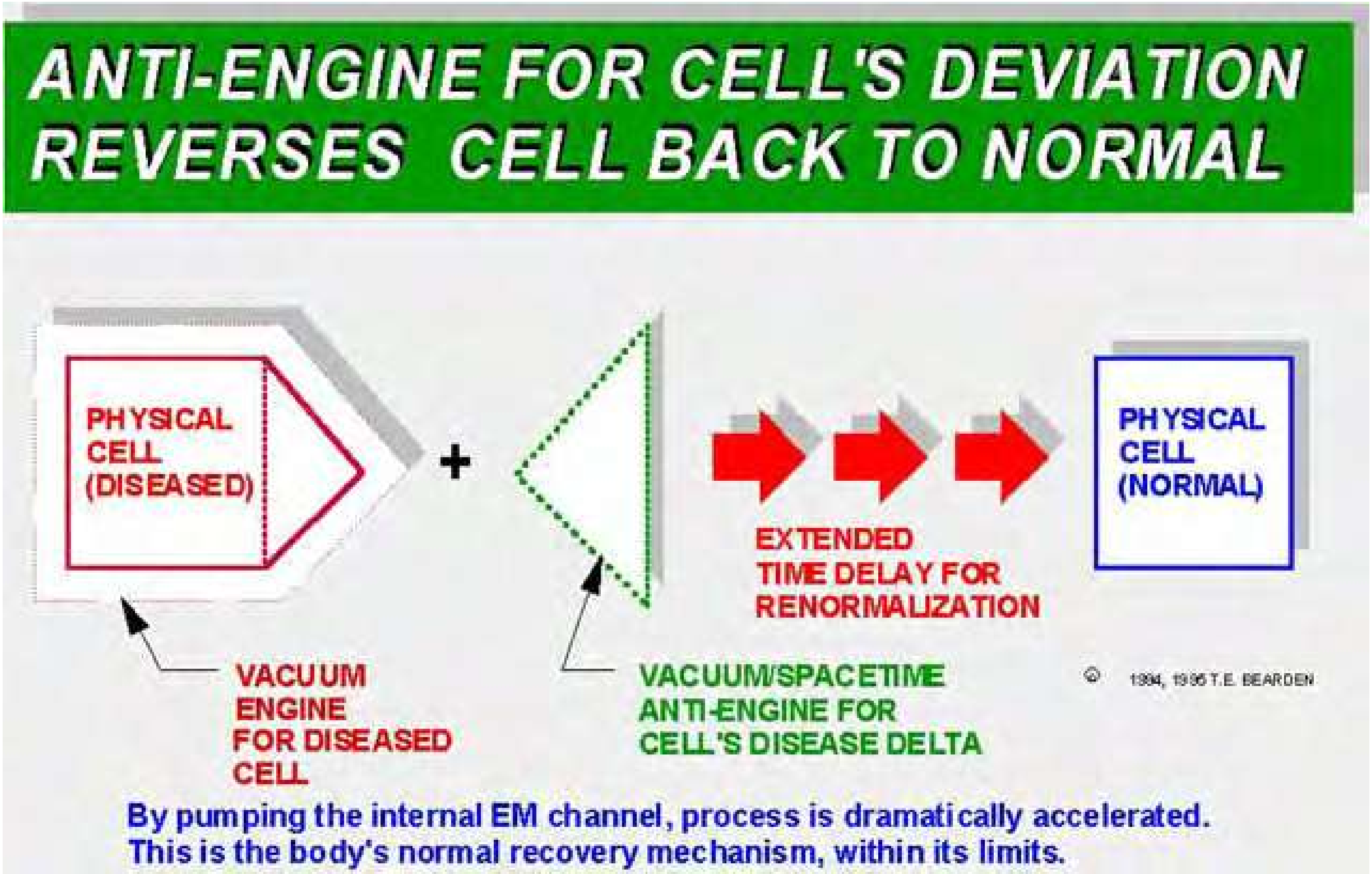

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL
CELL
(DISEASED)
+
EXTENDED
TIME DELAY FOR
RENORMALIZATION
PHYSICAL
CELL
(NORMAL)
VACUUM
ENGINE
FOR DISEASED
CELL
VACUUM/SPACETIME
ANTI-ENGINE FOR
CELL'S DISEASE DELTA

© 1994, 1995 T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated.
This is the body's normal recovery mechanism, within its limits.

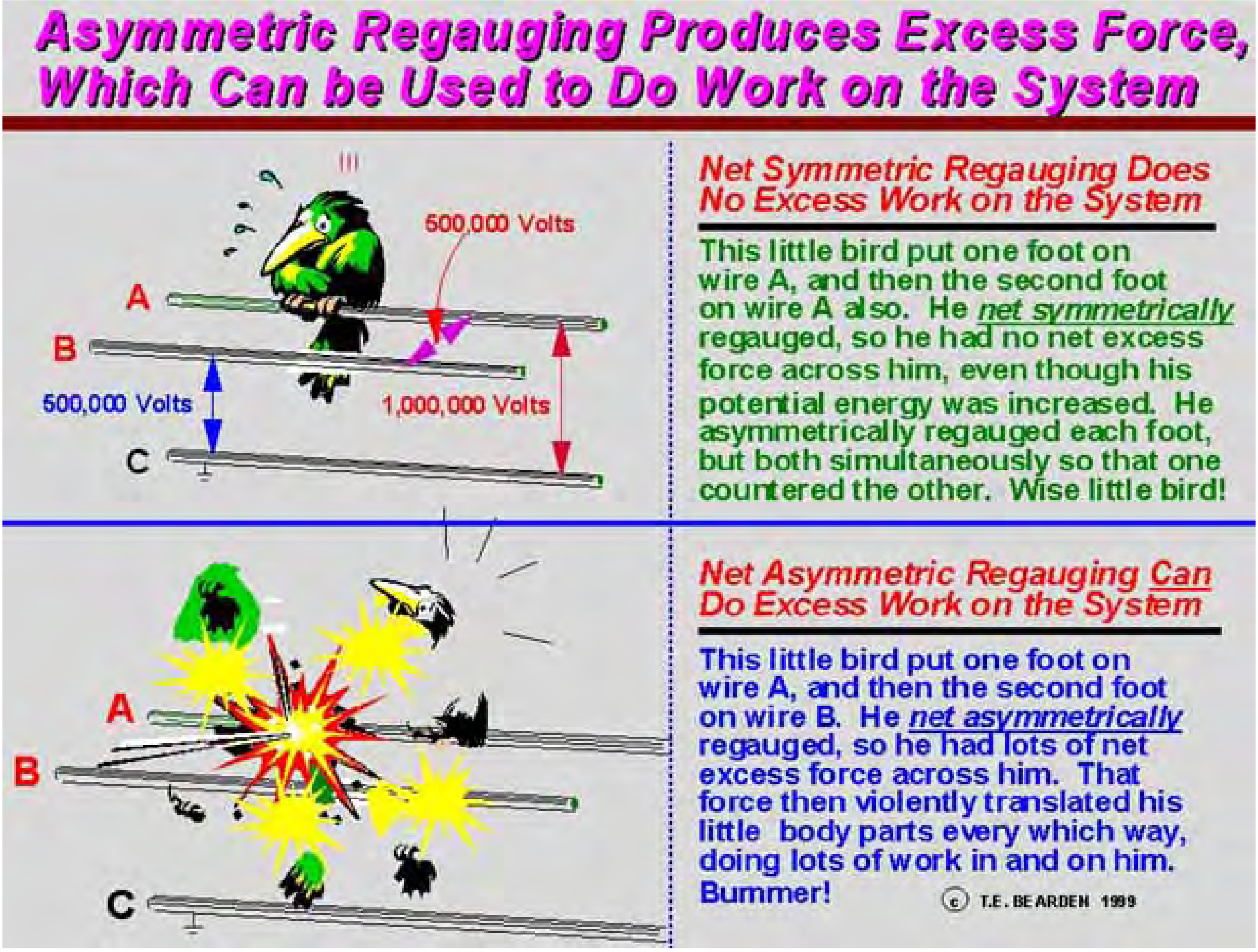

Asymmetric Regauging Produces Excess Force, Which Can be Used to Do Work on the System

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging *Can* Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1999

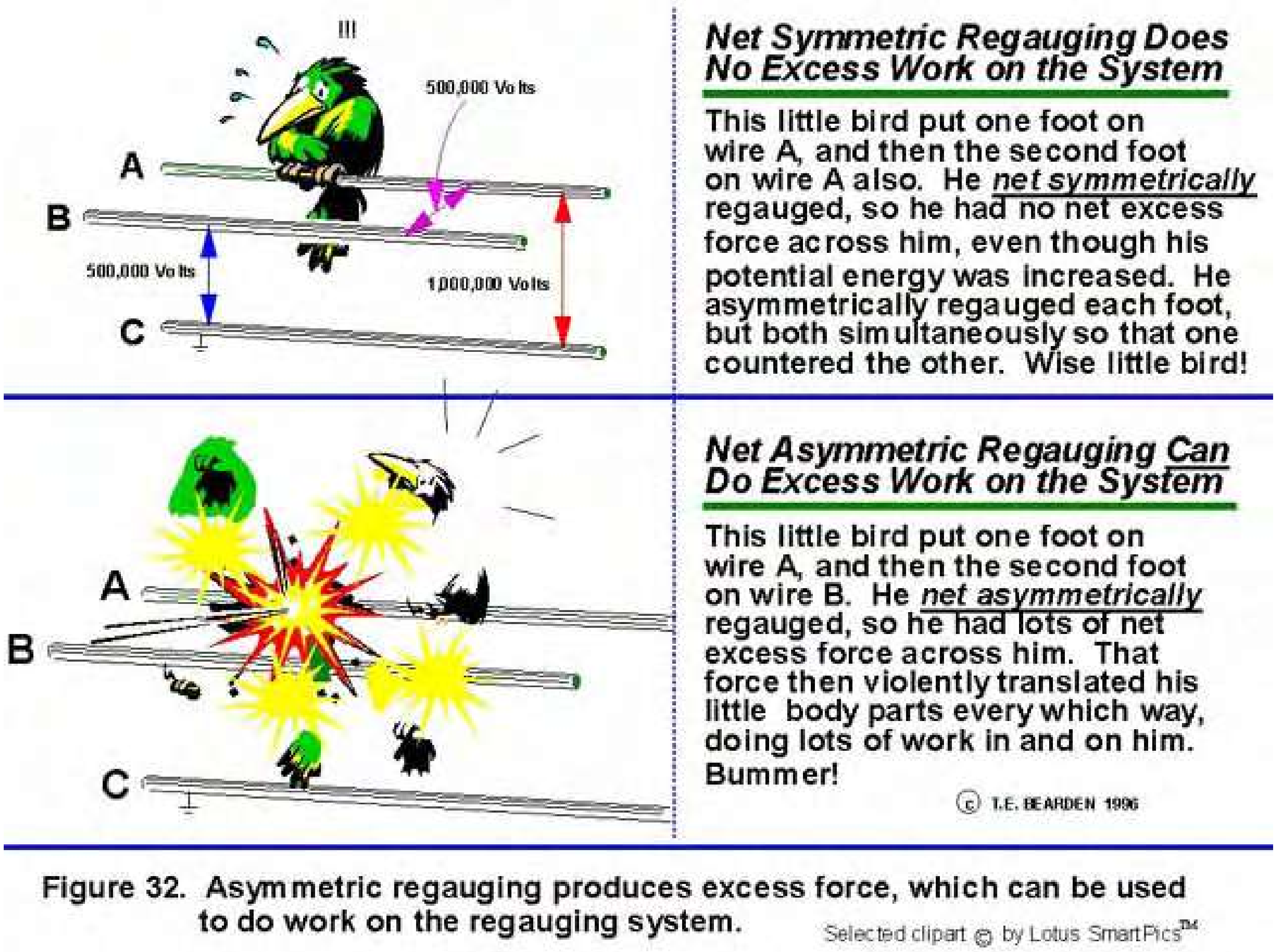

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging Can Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1996

Figure 32. Asymmetric regauging produces excess force, which can be used to do work on the regauging system.

Selected clipart © by Lotus SmartPics™

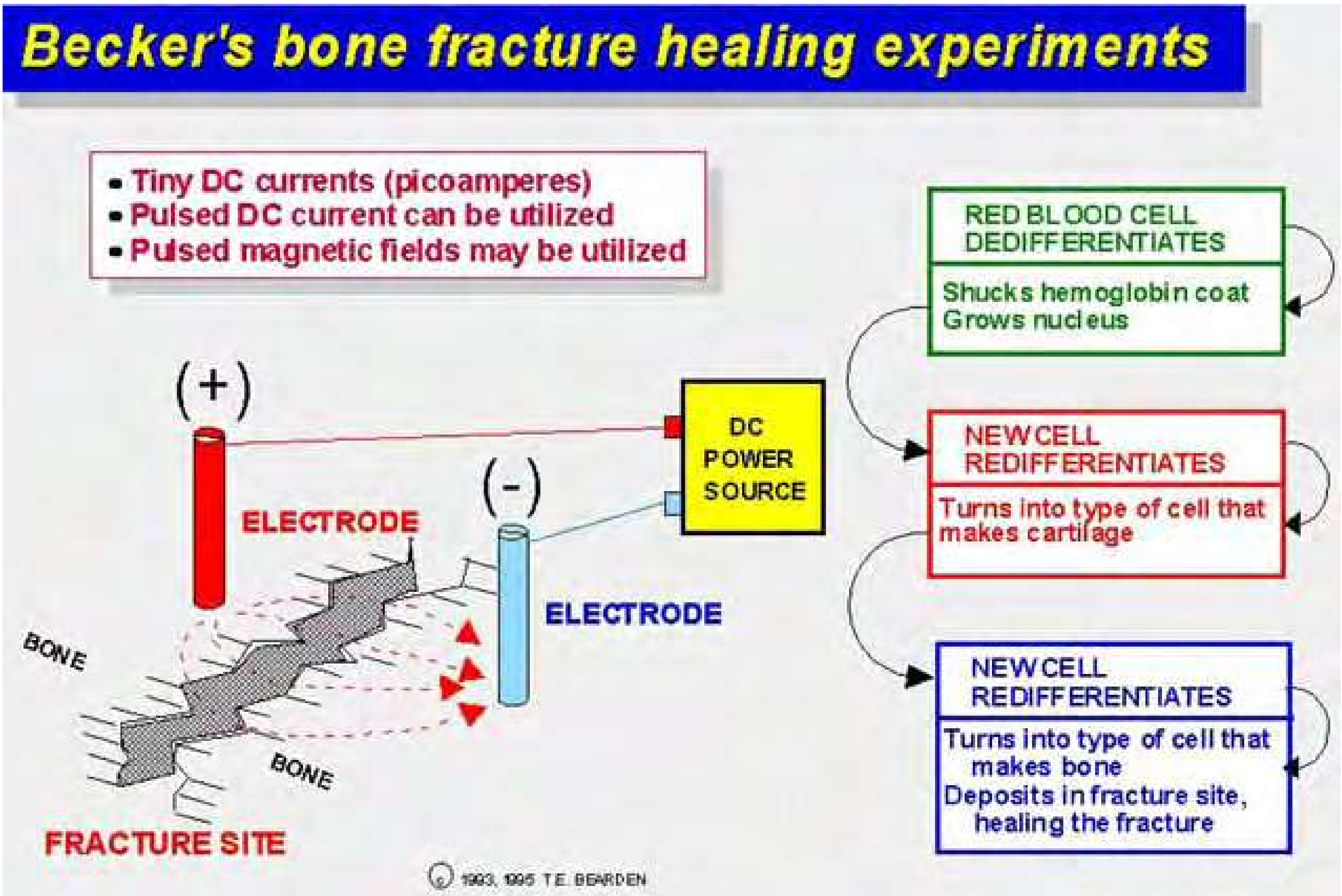

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

(+)
ELECTRODE
(-)
ELECTRODE
DC
POWER
SOURCE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1995 T.E. BEARDEN

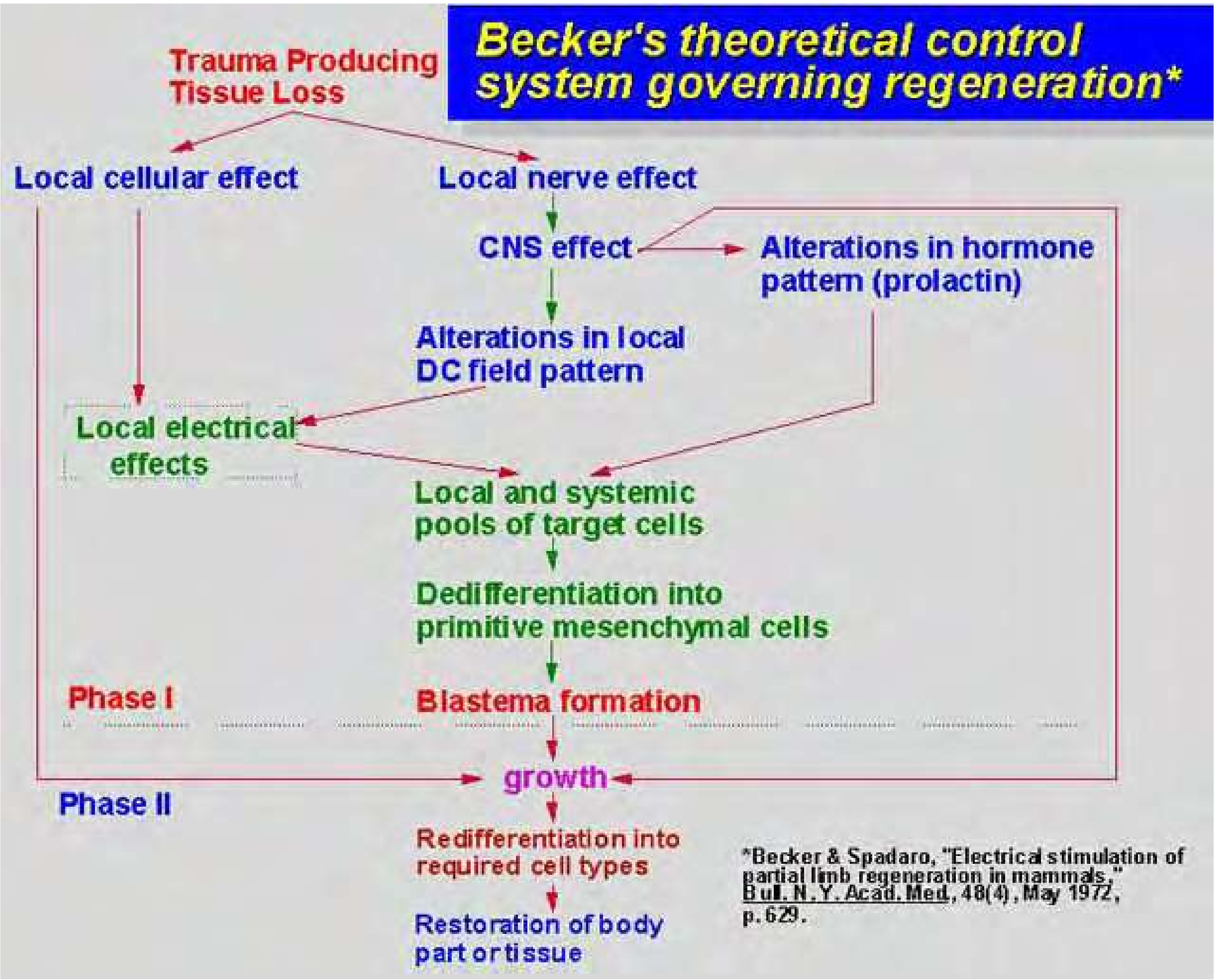

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss
Local cellular effect
Local nerve effect
CNS effect
Alterations in hormone pattern (prolactin)
Alterations in local DC field pattern
Local electrical effects
Local and systemic pools of target cells
Dedifferentiation into primitive mesenchymal cells
Phase I
Blastema formation
growth
Phase II
Redifferentiation into required cell types
Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N. Y. Acad. Med., 48(4), May 1972, p. 629.

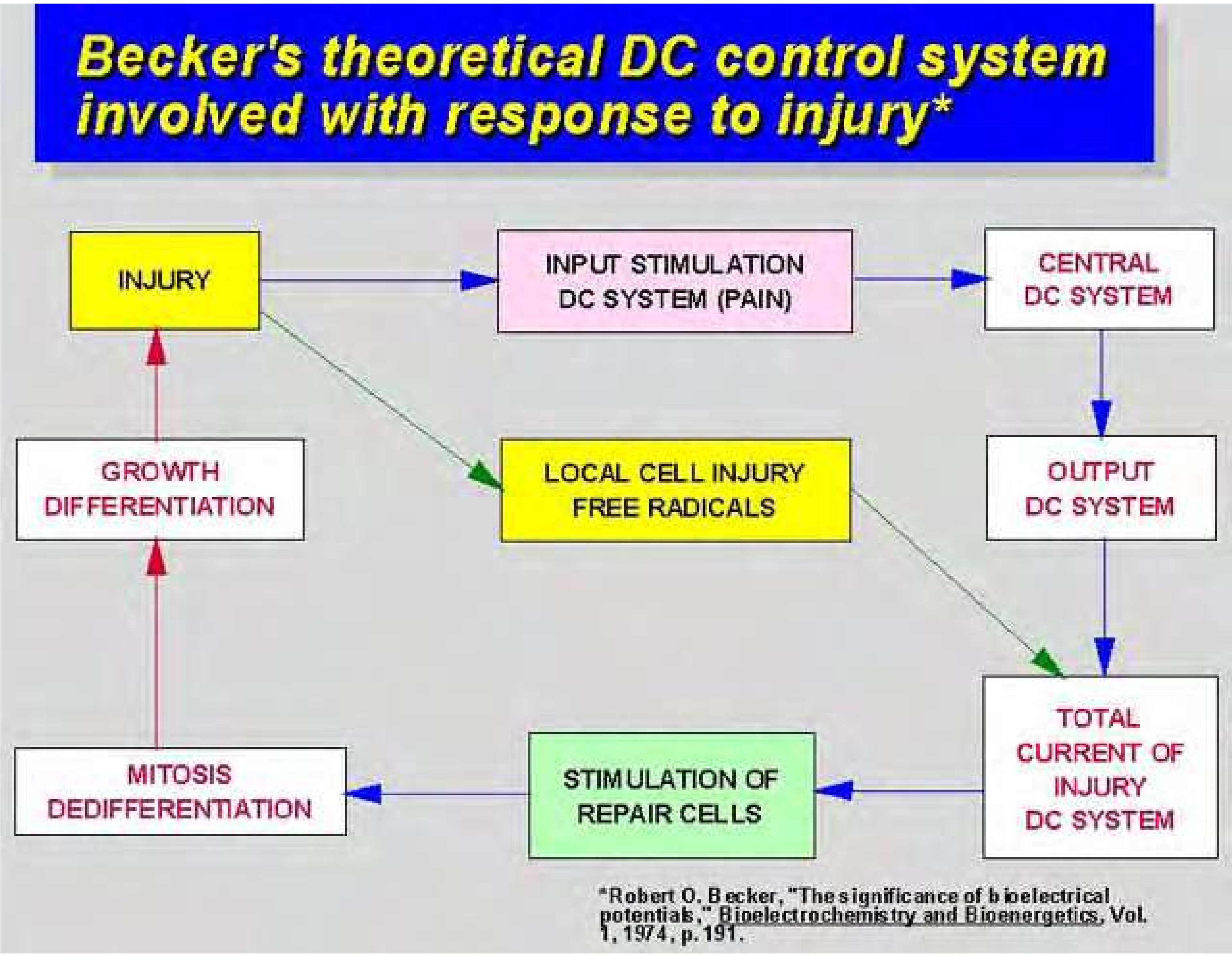

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

GROWTH DIFFERENTIATION

TOTAL CURRENT OF INJURY DC SYSTEM

STIMULATION OF REPAIR CELLS

MITOSIS DEDIFFERENTIATION

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p.191.

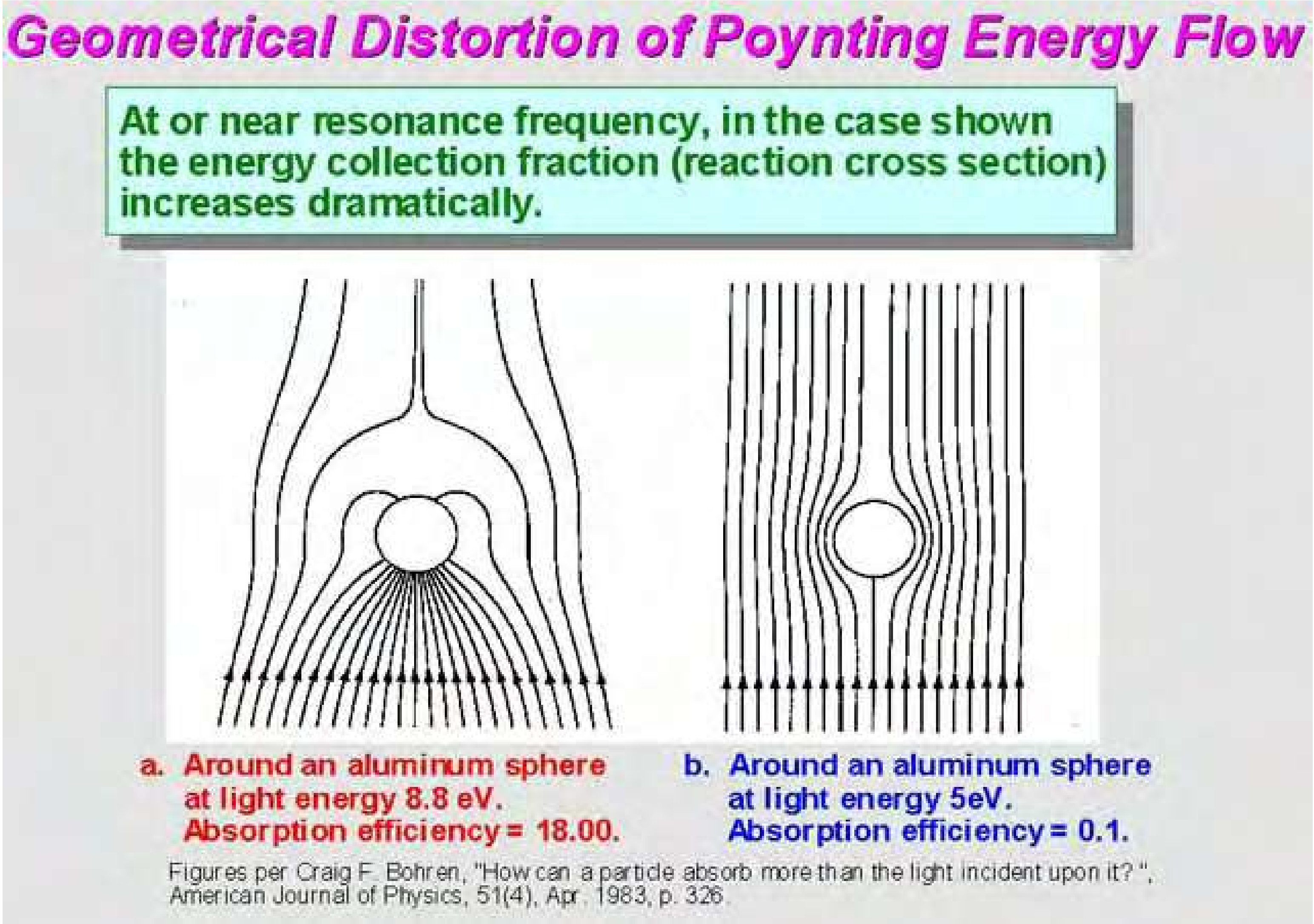

Geometrical Distortion of Poynting Energy Flow

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

a. Around an aluminum sphere at light energy 8.8 eV. Absorption efficiency = 18.00.

b. Around an aluminum sphere at light energy 5eV. Absorption efficiency = 0.1.

Figures per Craig F. Bohren, "How can a particle absorb more than the light incident upon it? ", American Journal of Physics, 51(4), Apr. 1983, p. 326.

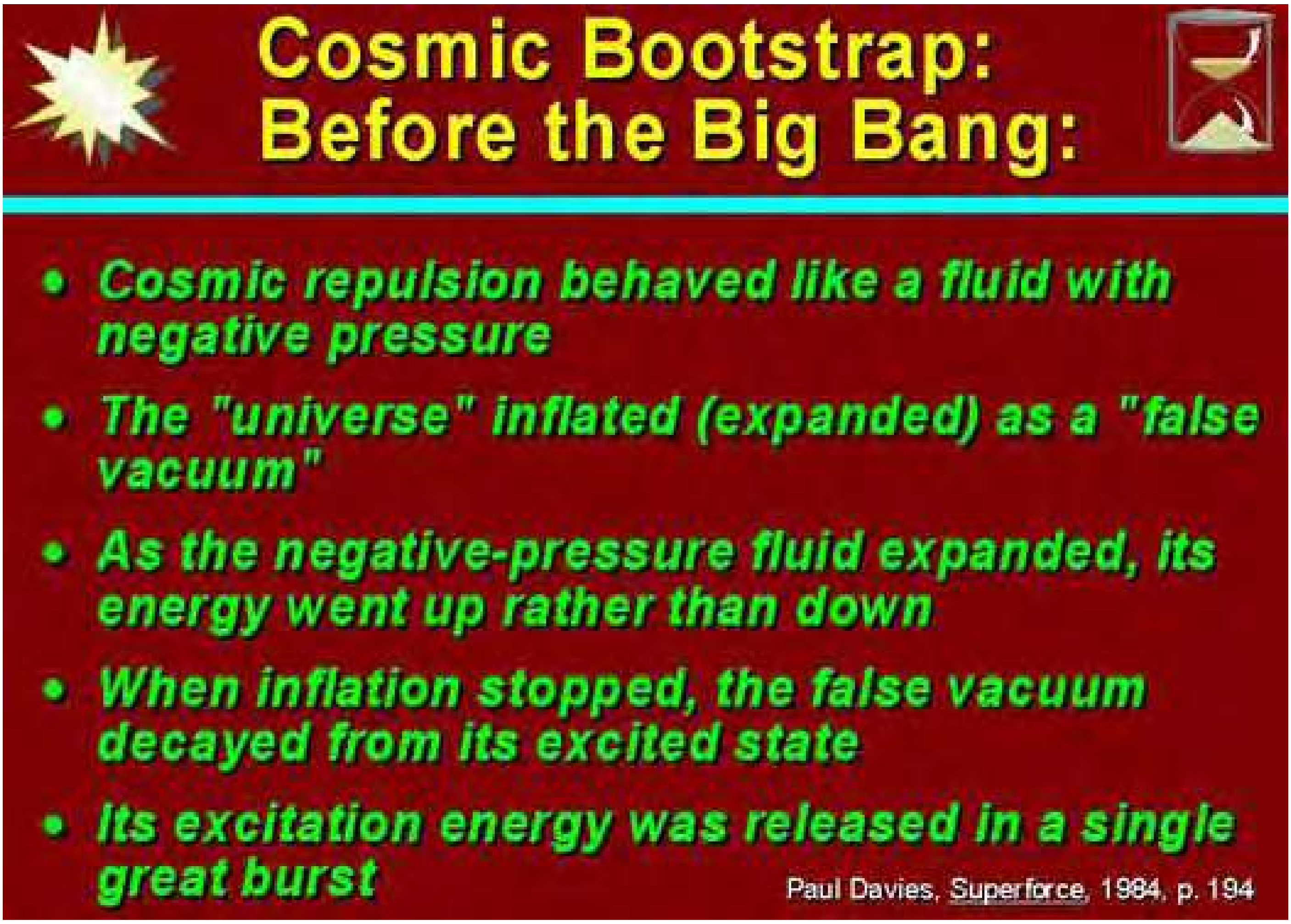

Cosmic Bootstrap: Before the Big Bang:

- *Cosmic repulsion behaved like a fluid with negative pressure*
- *The "universe" inflated (expanded) as a "false vacuum"*
- *As the negative-pressure fluid expanded, its energy went up rather than down*
- *When inflation stopped, the false vacuum decayed from its excited state*
- *Its excitation energy was released in a single great burst*

Paul Davies, Superforce, 1984, p. 194

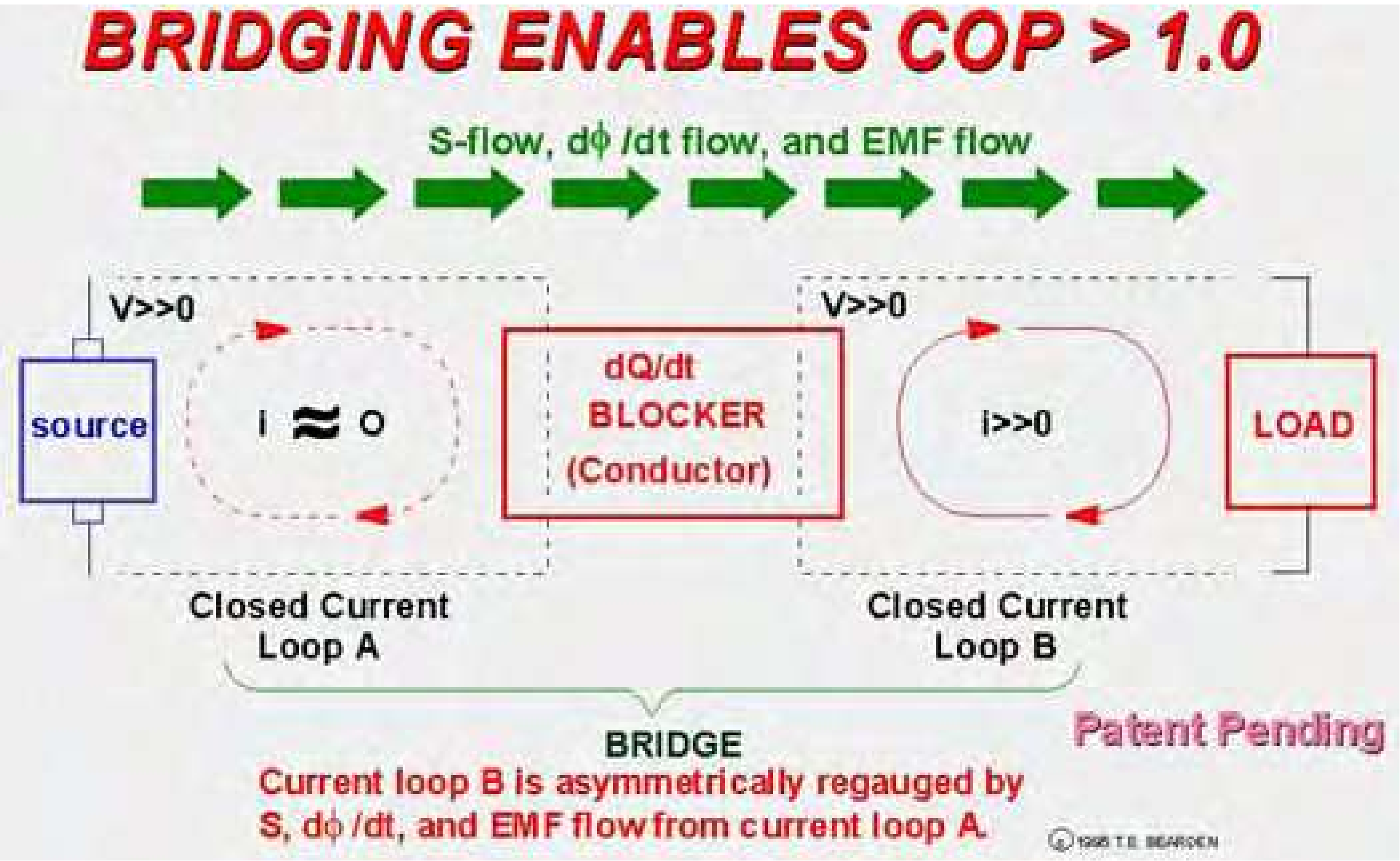

BRIDGING ENABLES COP > 1.0

S-flow, dφ /dt flow, and EMF flow

V>>0
source
I ≈ 0
dQ/dt
BLOCKER
(Conductor)
V>>0
I>>0
LOAD

Closed Current Loop A

Closed Current Loop B

BRIDGE

Current loop B is asymmetrically regauged by S, dφ /dt, and EMF flow from current loop A.

Patent Pending

©1998 T.E. BEARDEN

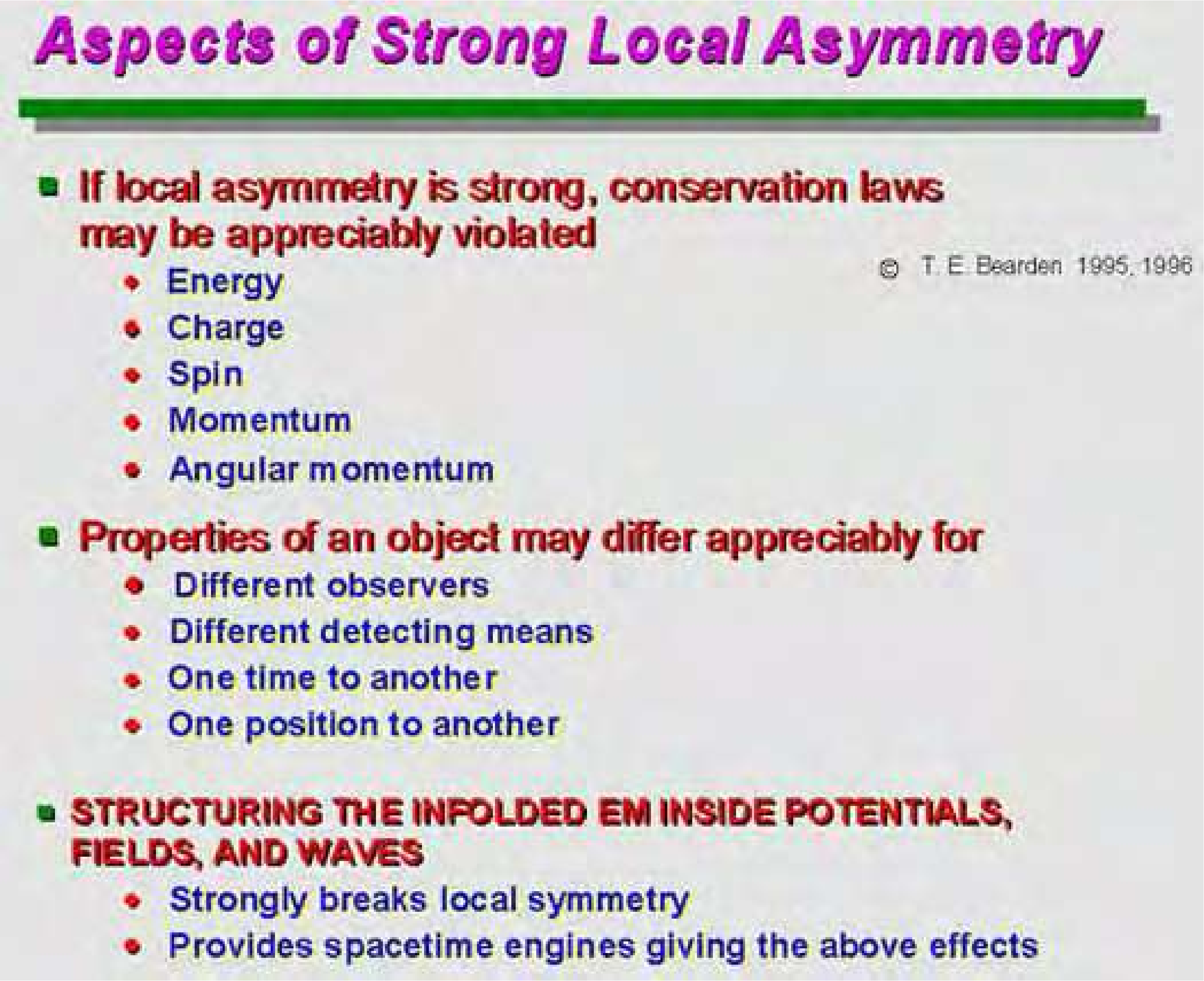

Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

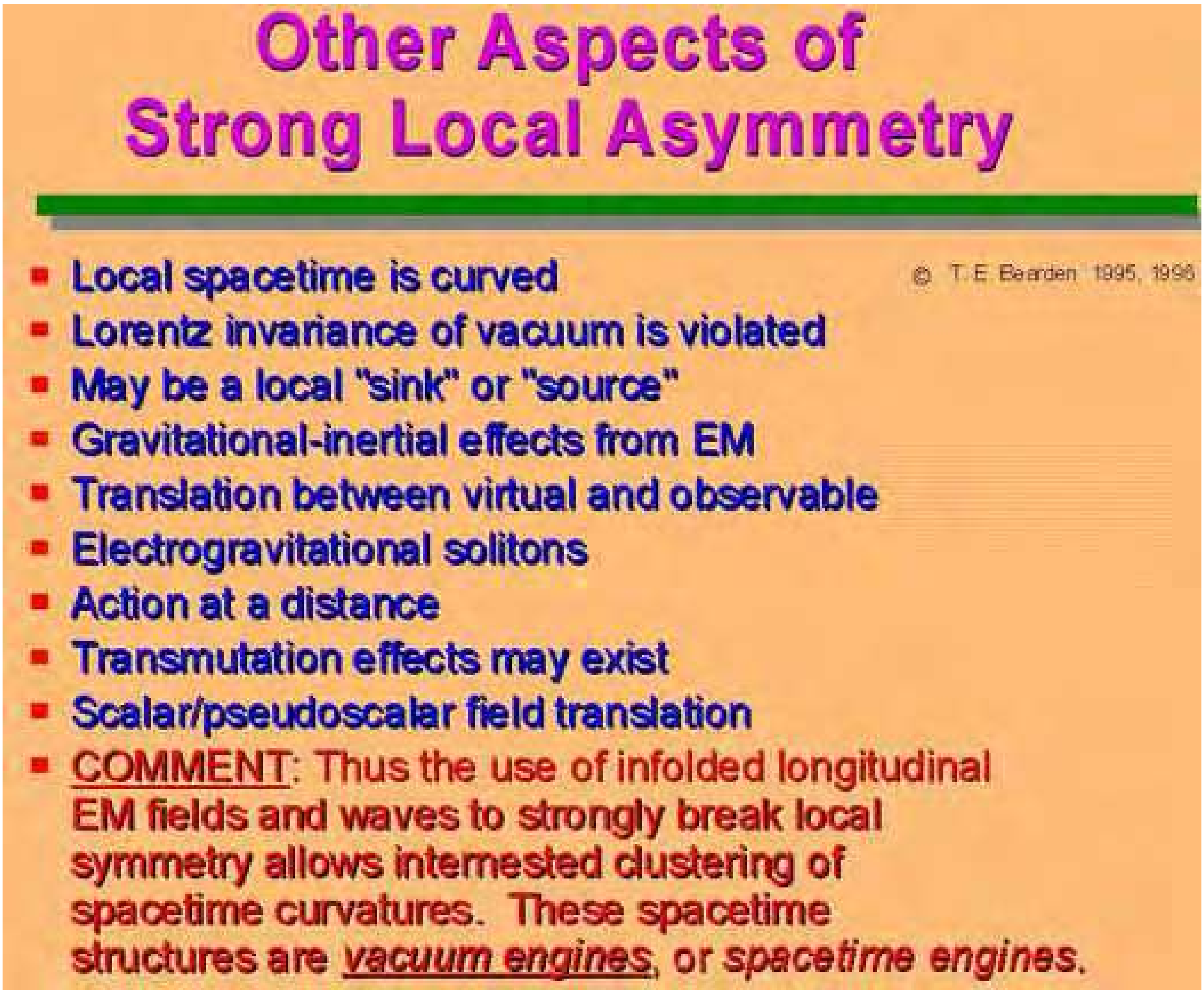

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational-inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

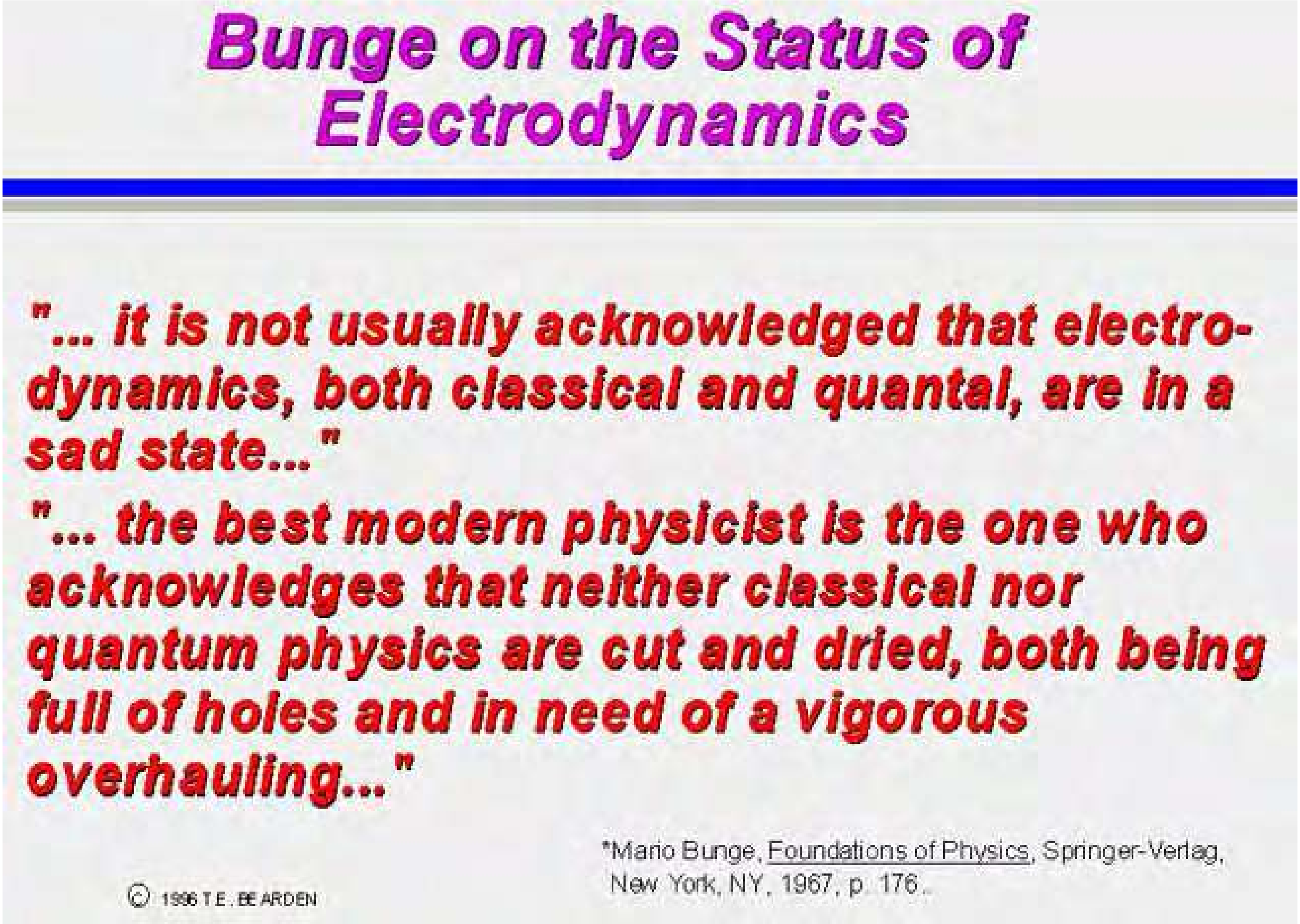

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176.

© 1996 T.E. BEARDEN

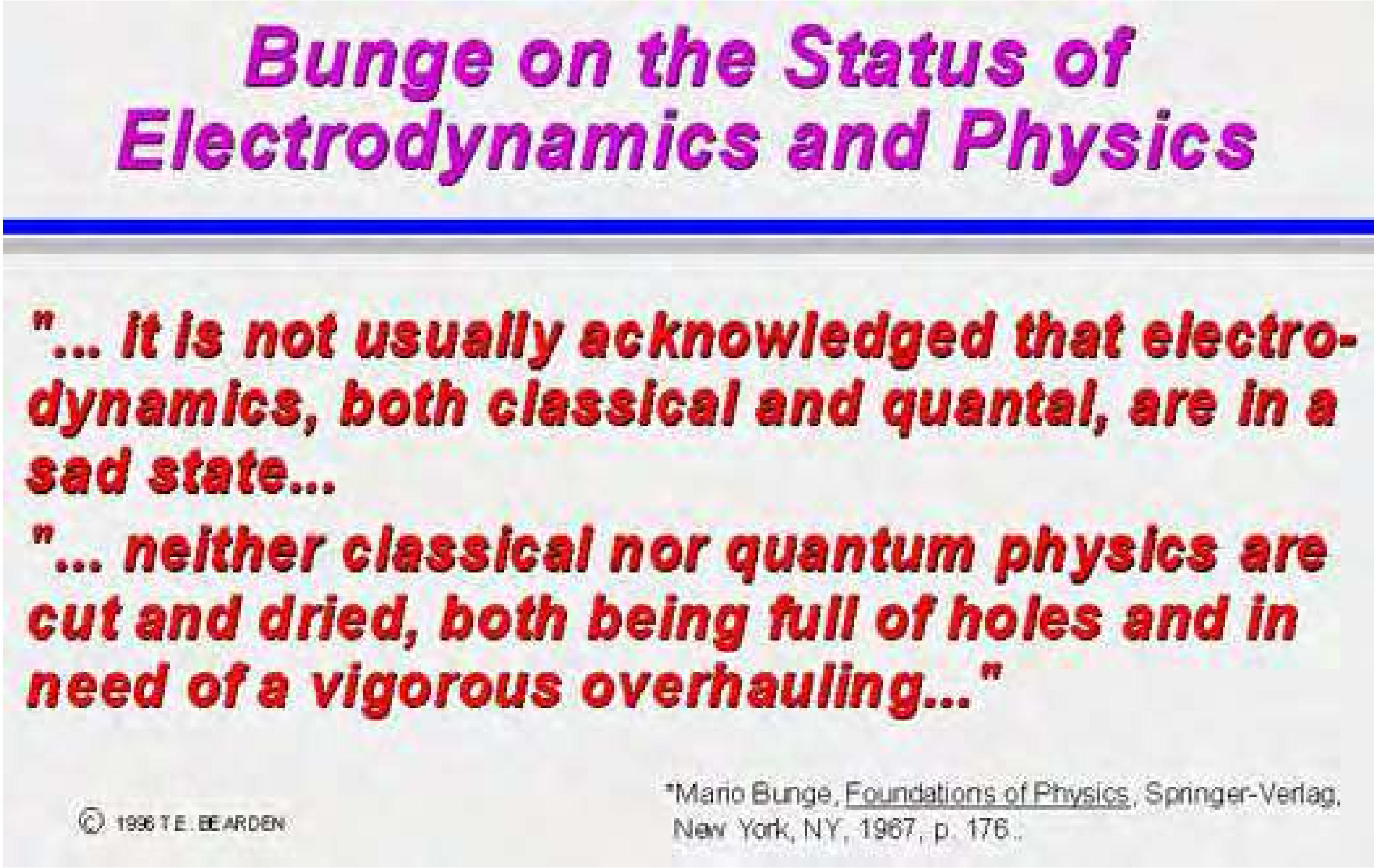

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

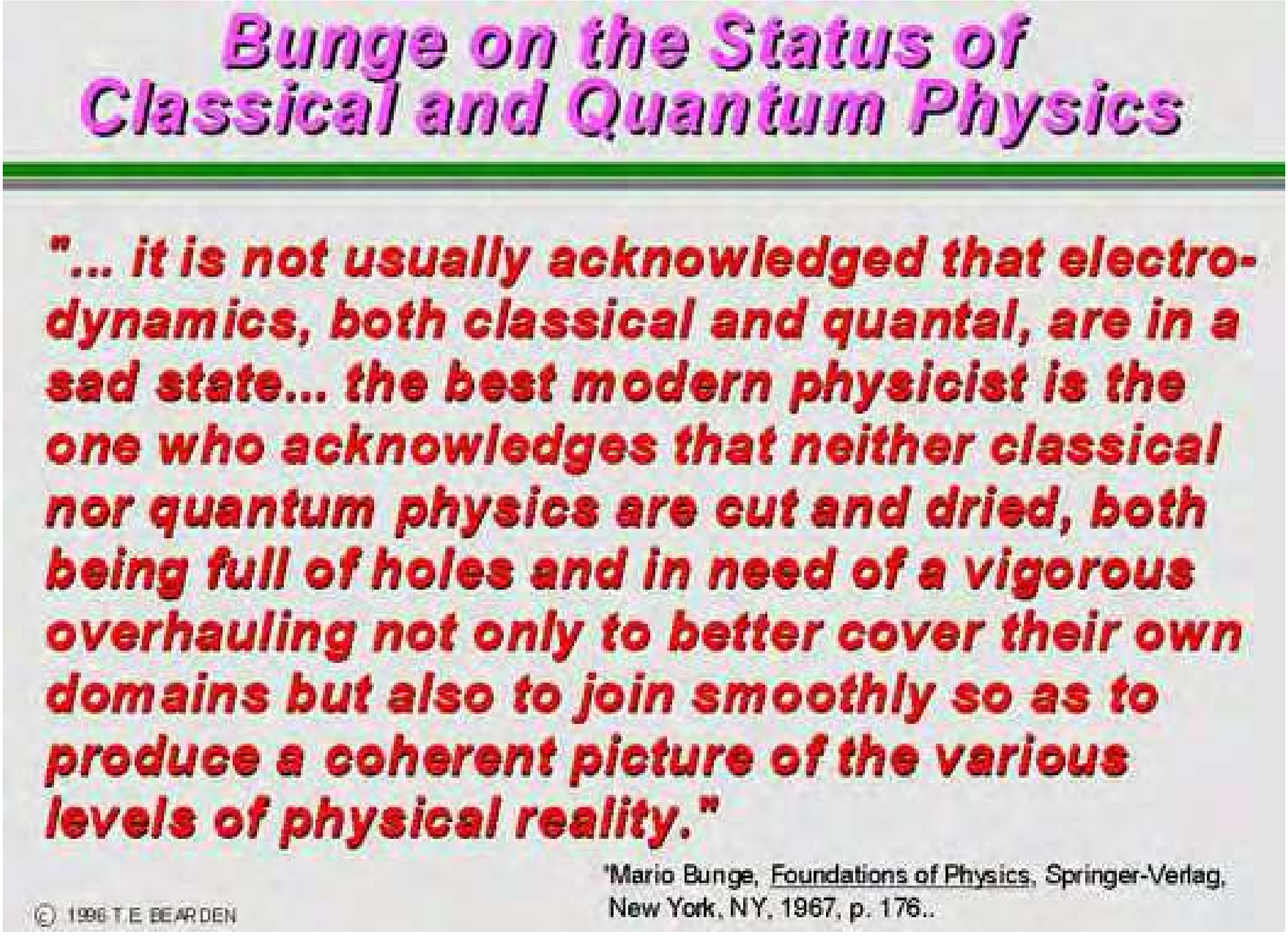

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

'Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

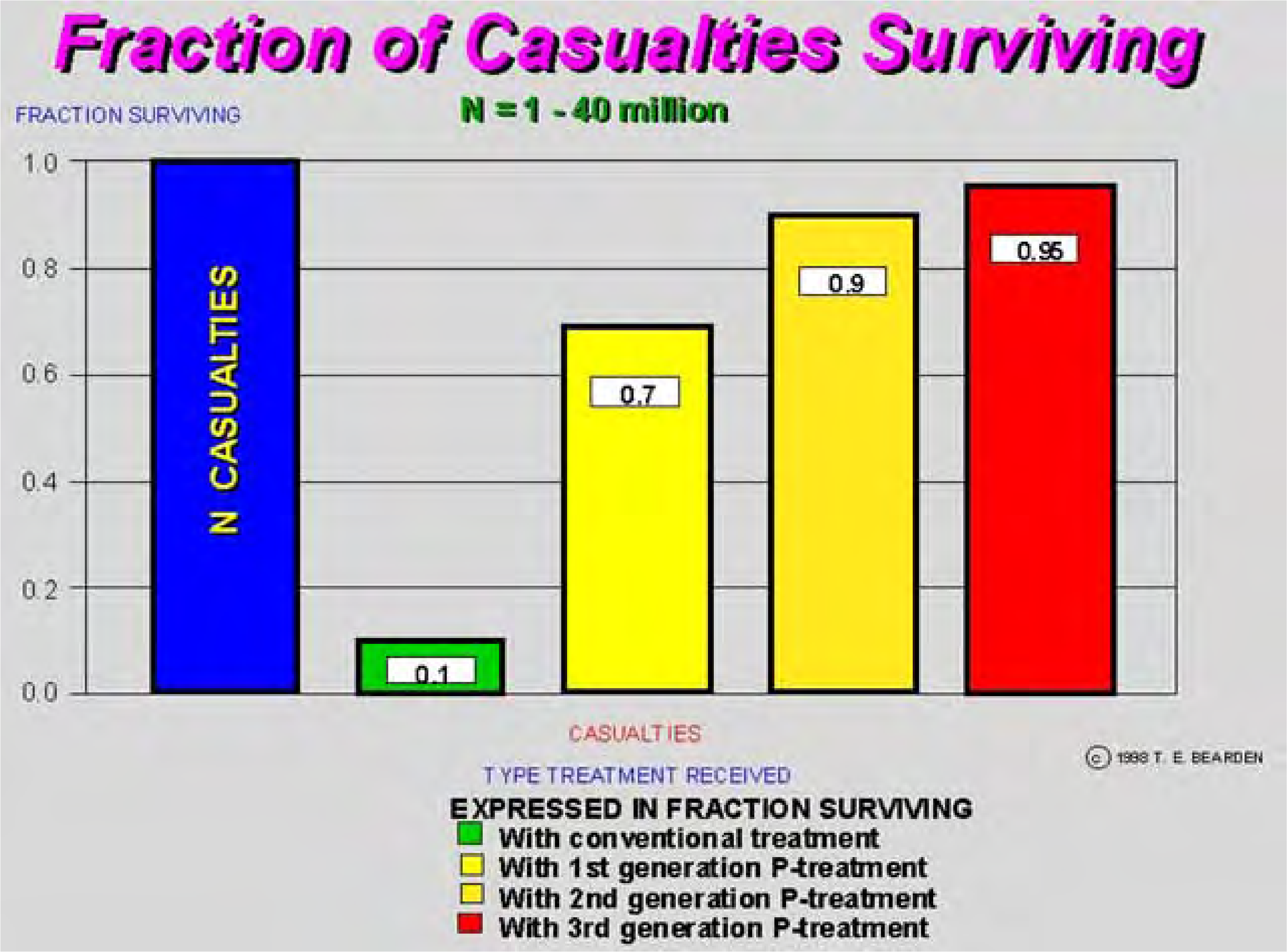
Fraction of Casualties Surviving
FRACTION SURVIVING
N = 1 - 40 million
1.0
0.8
0.6
0.4
0.2
0.0
N CASUALTIES
0.1
0.7
0.9
0.95
CASUALTIES
TYPE TREATMENT RECEIVED
© 1998 T. E. BEARDEN
EXPRESSED IN FRACTION SURVIVING
🟩 With conventional treatment
🟨 With 1st generation P-treatment
🟨 With 2nd generation P-treatment
🟥 With 3rd generation P-treatment

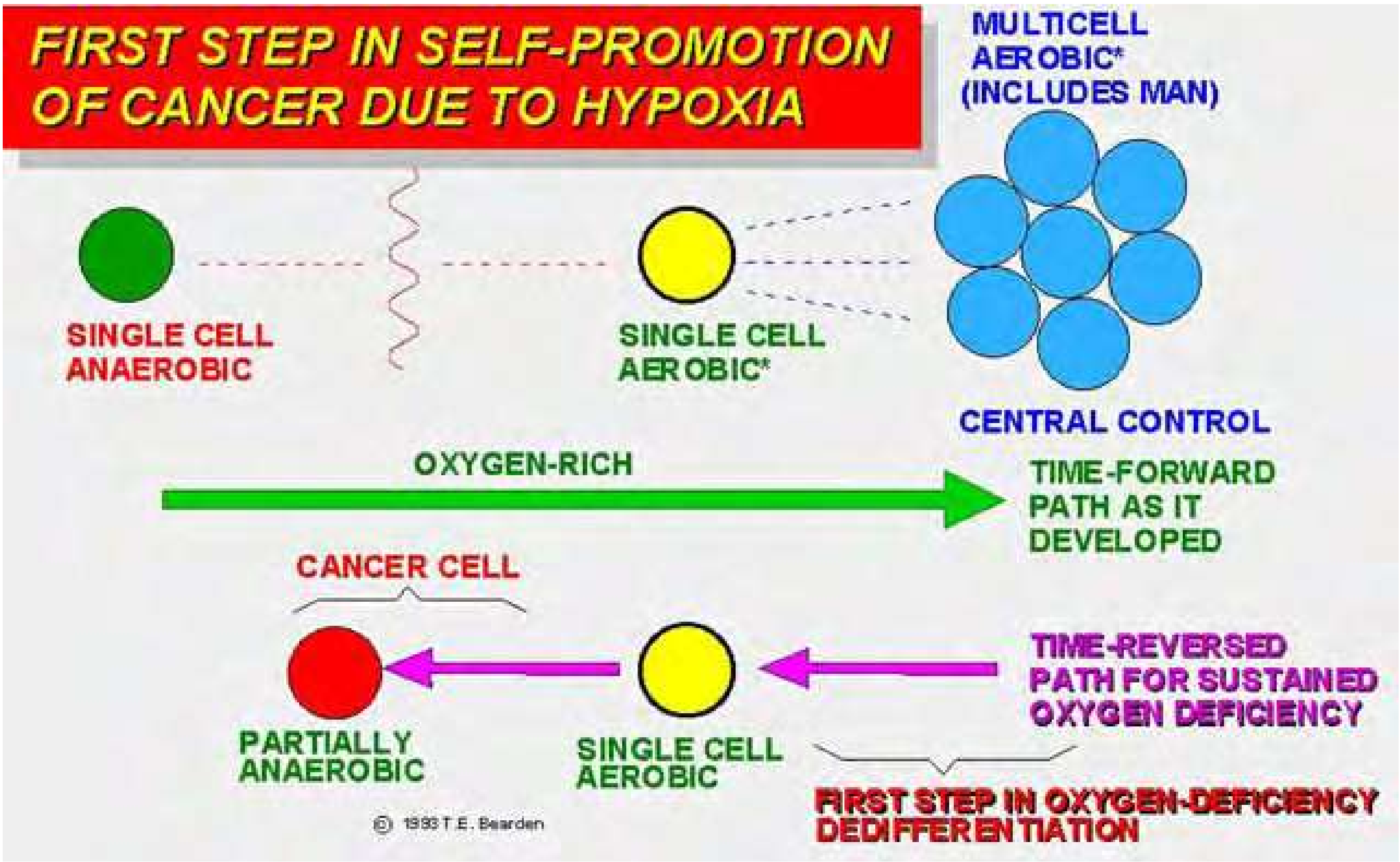

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

MULTICELL
AEROBIC*
(INCLUDES MAN)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

© 1993 T.E. Bearden

Differentiation and Dedifferentiation Timelines

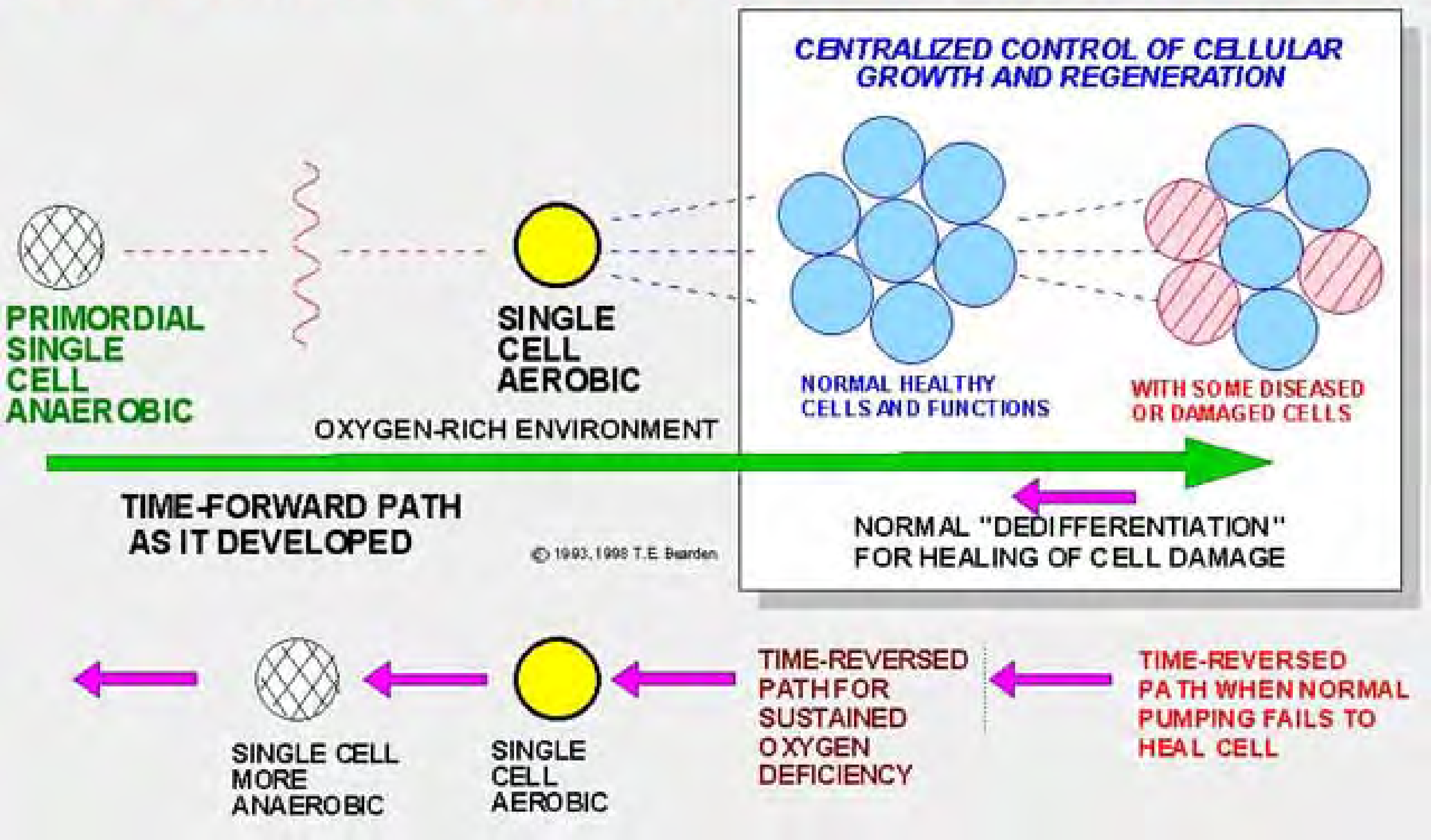
CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION
PRIMORDIAL SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC
OXYGEN-RICH ENVIRONMENT
NORMAL HEALTHY CELLS AND FUNCTIONS
WITH SOME DISEASED OR DAMAGED CELLS
TIME-FORWARD PATH AS IT DEVELOPED
NORMAL "DEDIFFERENTIATION" FOR HEALING OF CELL DAMAGE
SINGLE CELL MORE ANAEROBIC
SINGLE CELL AEROBIC
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
TIME-REVERSED PATH WHEN NORMAL PUMPING FAILS TO HEAL CELL

© 1993, 1998 T.E. Bearden

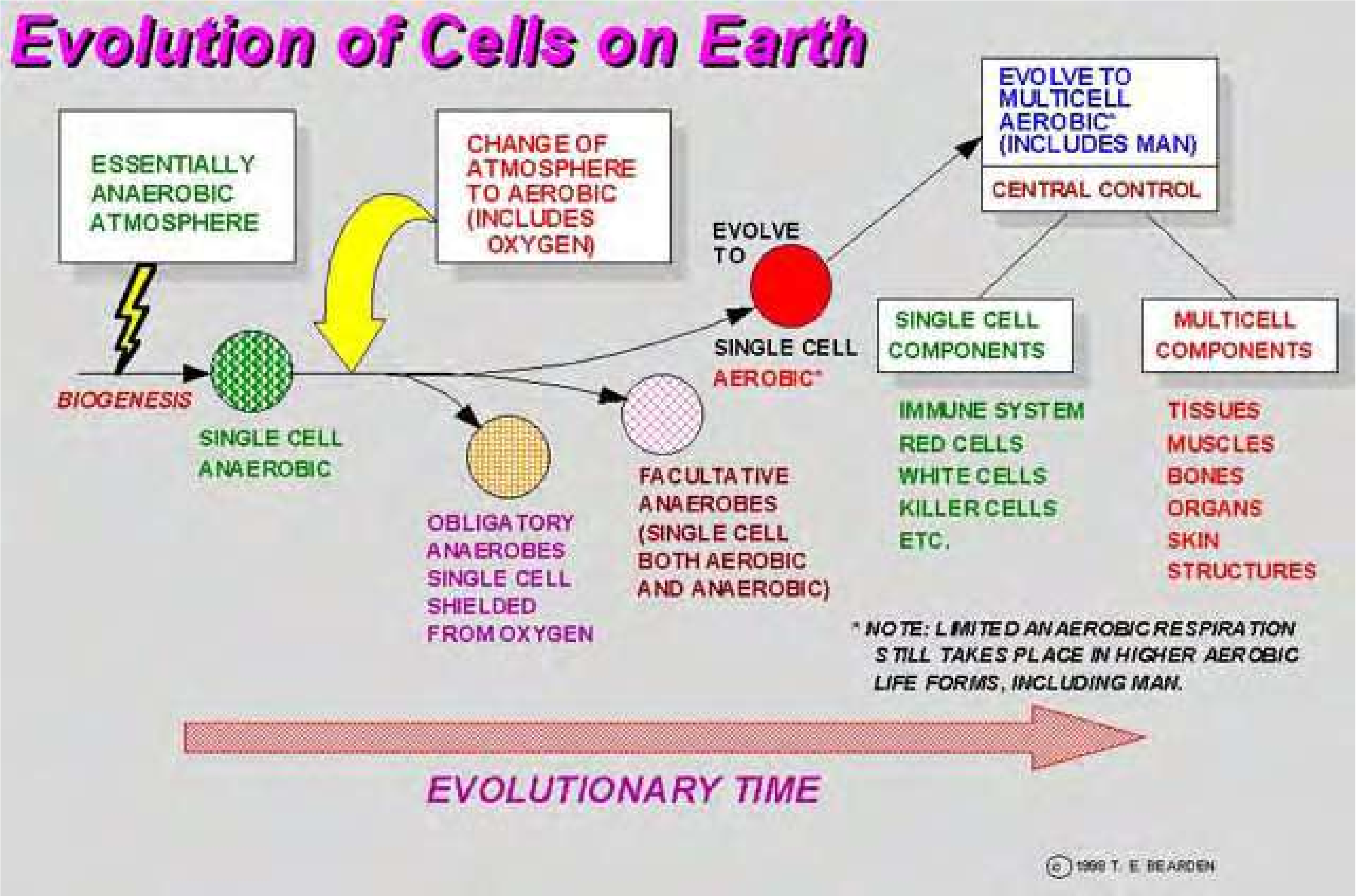

Evolution of Cells on Earth

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

*NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

(c) 1999 T. E. BEARDEN

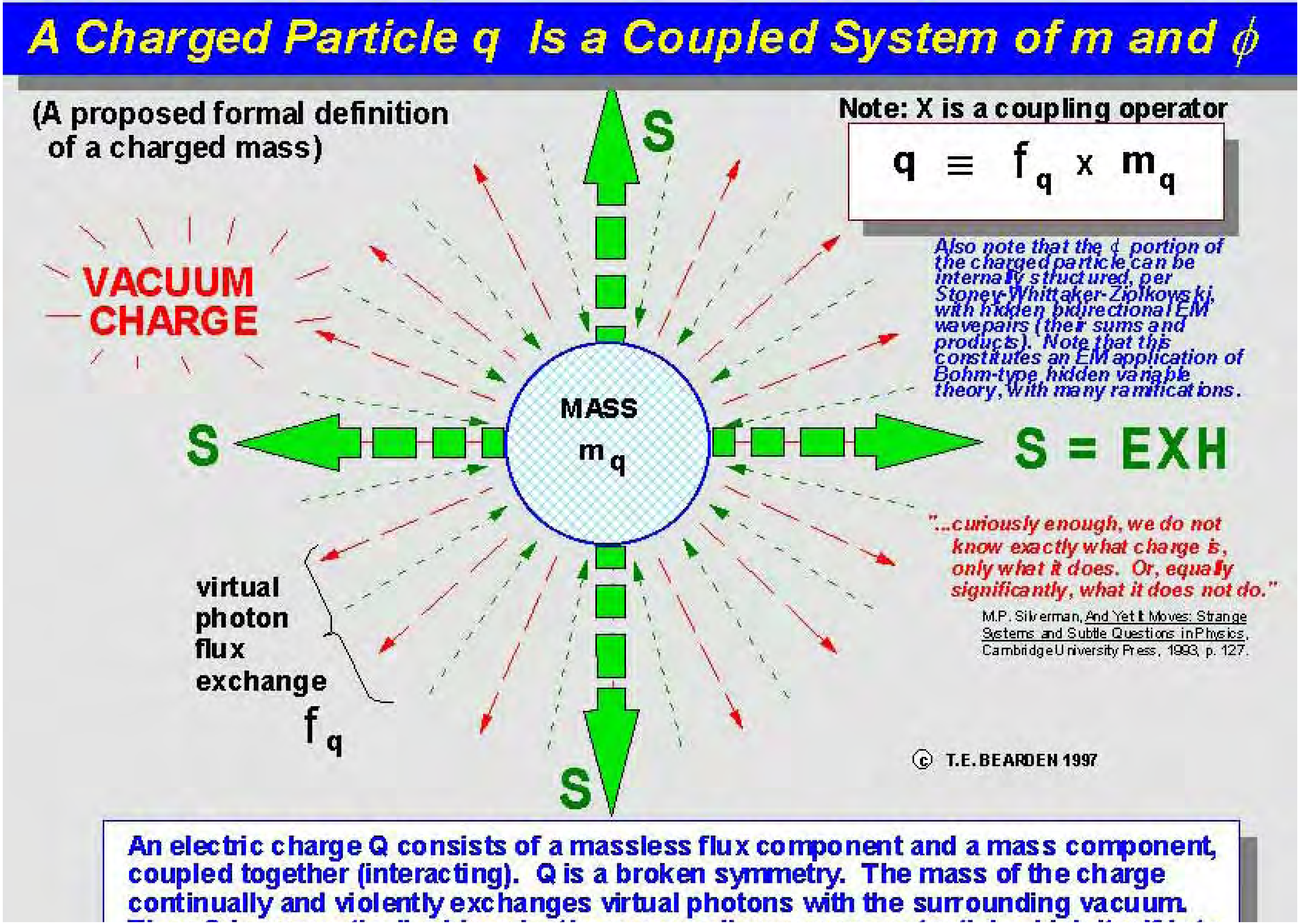

A Charged Particle q Is a Coupled System of m and ϕ

(A proposed formal definition
of a charged mass)

Note: X is a coupling operator

$$q \equiv f_q \times m_q$$

Also note that the ϕ portion of the charged particle can be internally structured, per Stoney-Whittaker-Ziolkowski, with hidden bidirectional EM wavepairs (their sums and products). Note that this constitutes an EM application of Bohm-type hidden variable theory, with many ramifications.

S = EXH

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

© T.E. BEARDEN 1997

An electric charge Q consists of a massless flux component and a mass component, coupled together (interacting). Q is a broken symmetry. The mass of the charge continually and violently exchanges virtual photons with the surrounding vacuum.

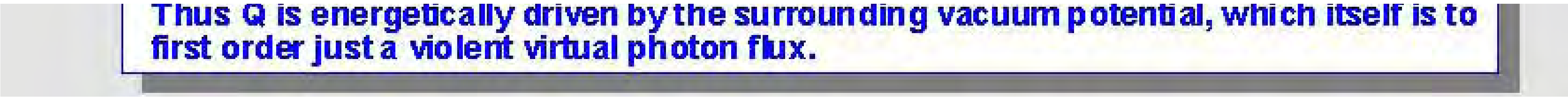

Thus Q is energetically driven by the surrounding vacuum potential, which itself is to first order just a violent virtual photon flux.

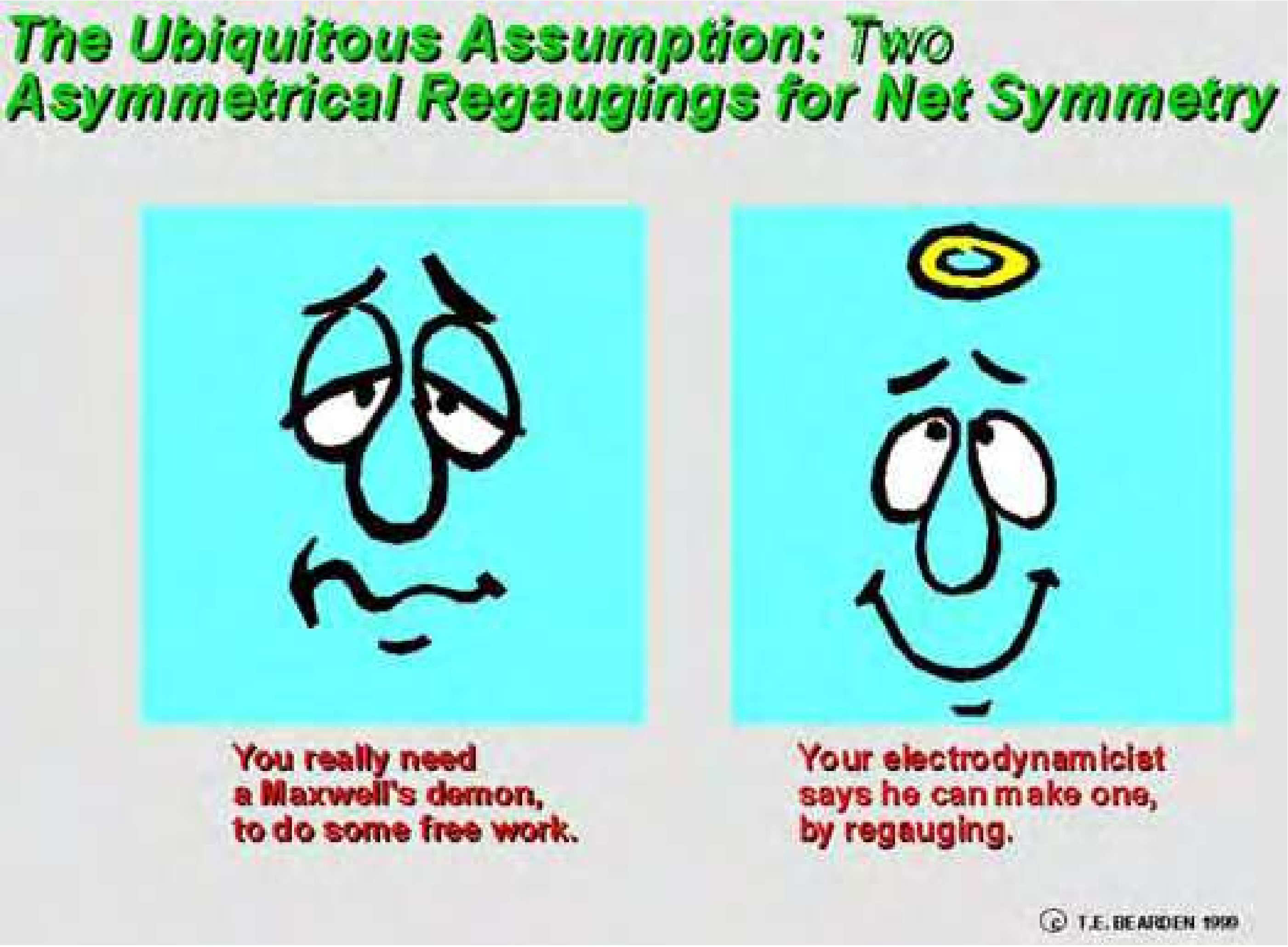

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

You really need
a Maxwell's demon,
to do some free work.

Your electrodynamicist
says he can make one,
by regauging.

© T.E. BEARDEN 1999

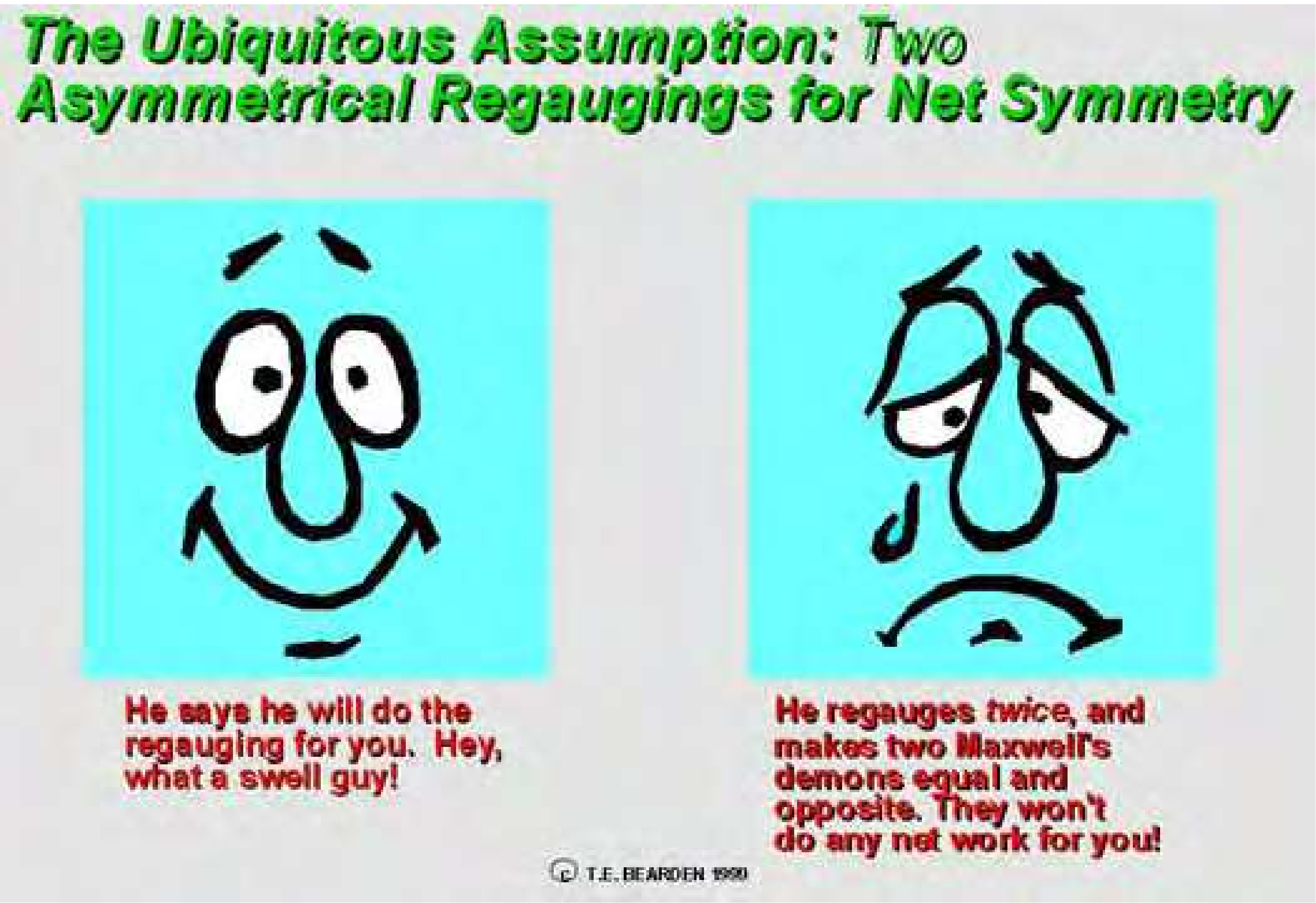

The Ubiquitous Assumption: *Two* Asymmetrical Regaugings for Net Symmetry

He says he will do the regauging for you. Hey, what a swell guy!

He regauges *twice*, and makes two Maxwell's demons equal and opposite. They won't do any net work for you!

© T.E. BEARDEN 1999

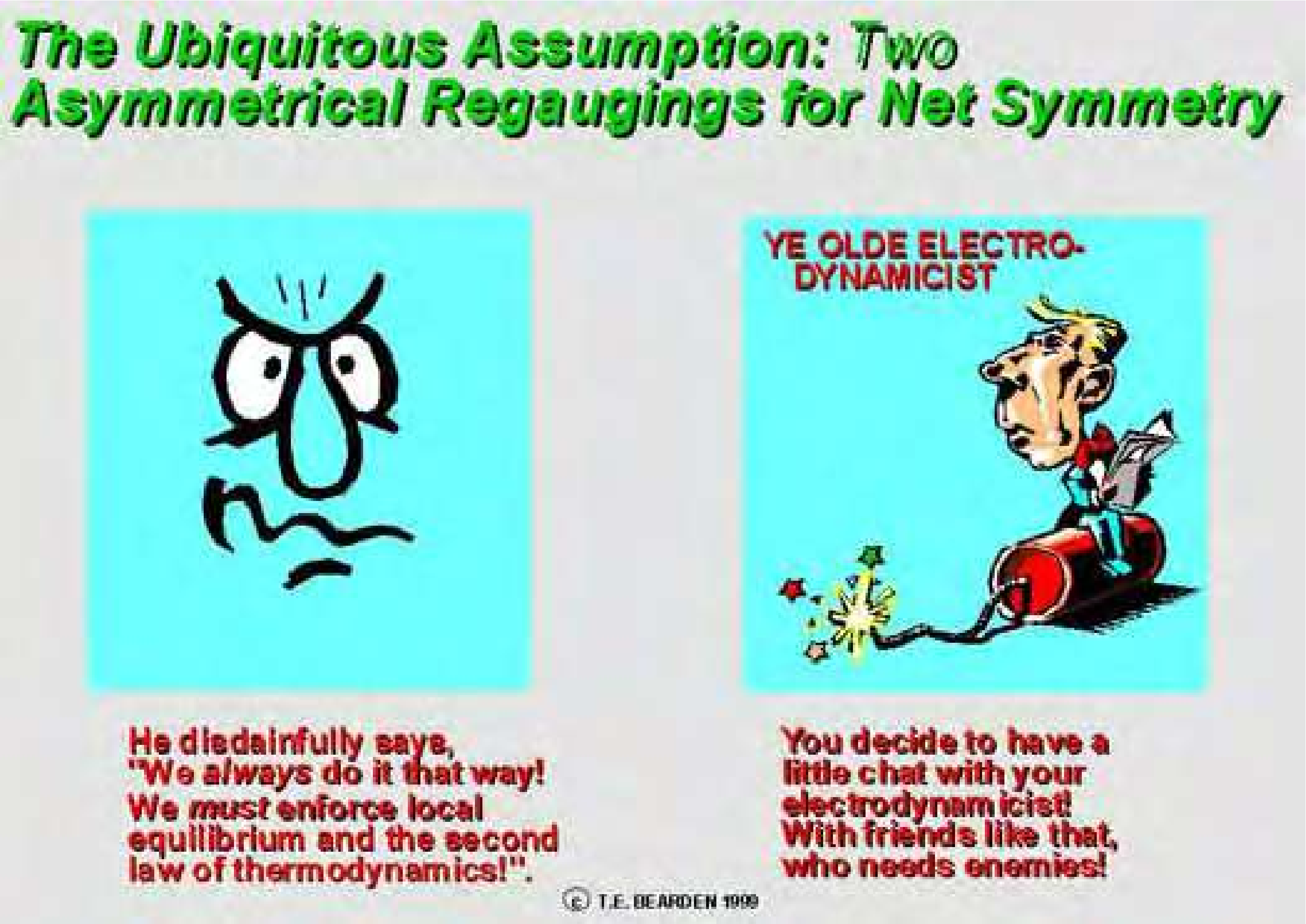

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

YE OLDE ELECTRO-
DYNAMICIST

He disdainfully says,
"We *always* do it that way!
We *must* enforce local
equilibrium and the second
law of thermodynamics!".

You decide to have a
little chat with your
electrodynamicist!
With friends like that,
who needs enemies!

© T.E. BEARDEN 1999

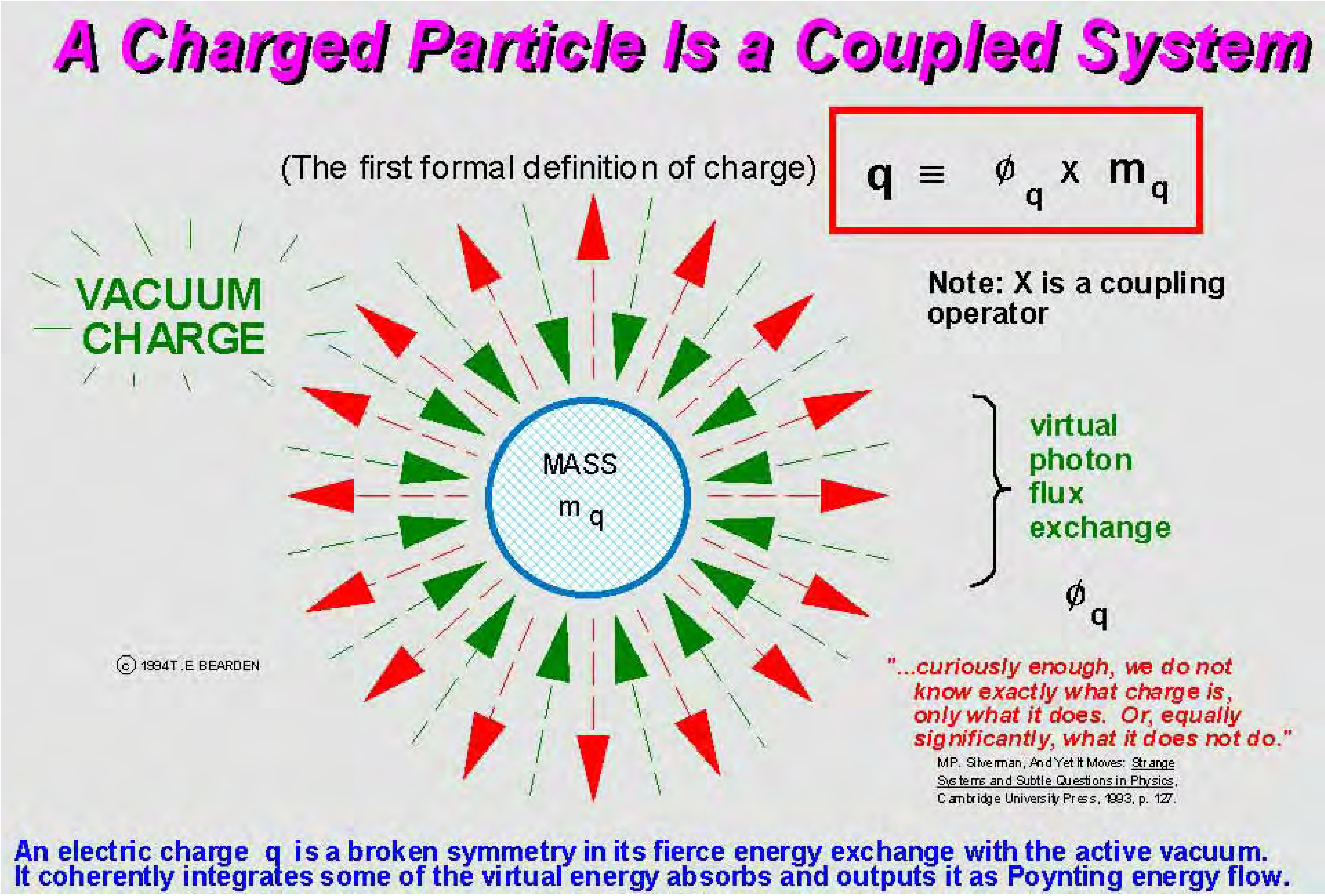

A Charged Particle Is a Coupled System

(The first formal definition of charge)

$$\boxed{q \equiv \phi_q \; x \; m_q}$$

Note: X is a coupling operator

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge q is a broken symmetry in its fierce energy exchange with the active vacuum. It coherently integrates some of the virtual energy absorbs and outputs it as Poynting energy flow.

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

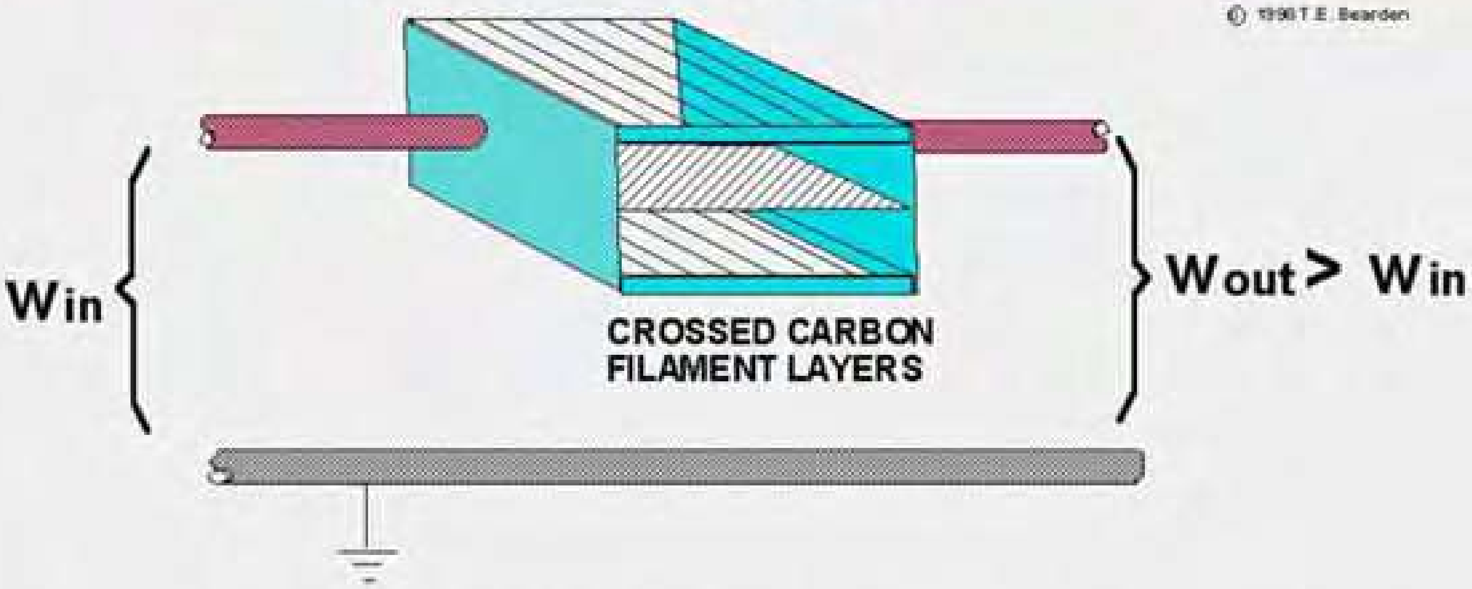
© 1998 T.E. Bearden
Win
Wout > Win
CROSSED CARBON
FILAMENT LAYERS

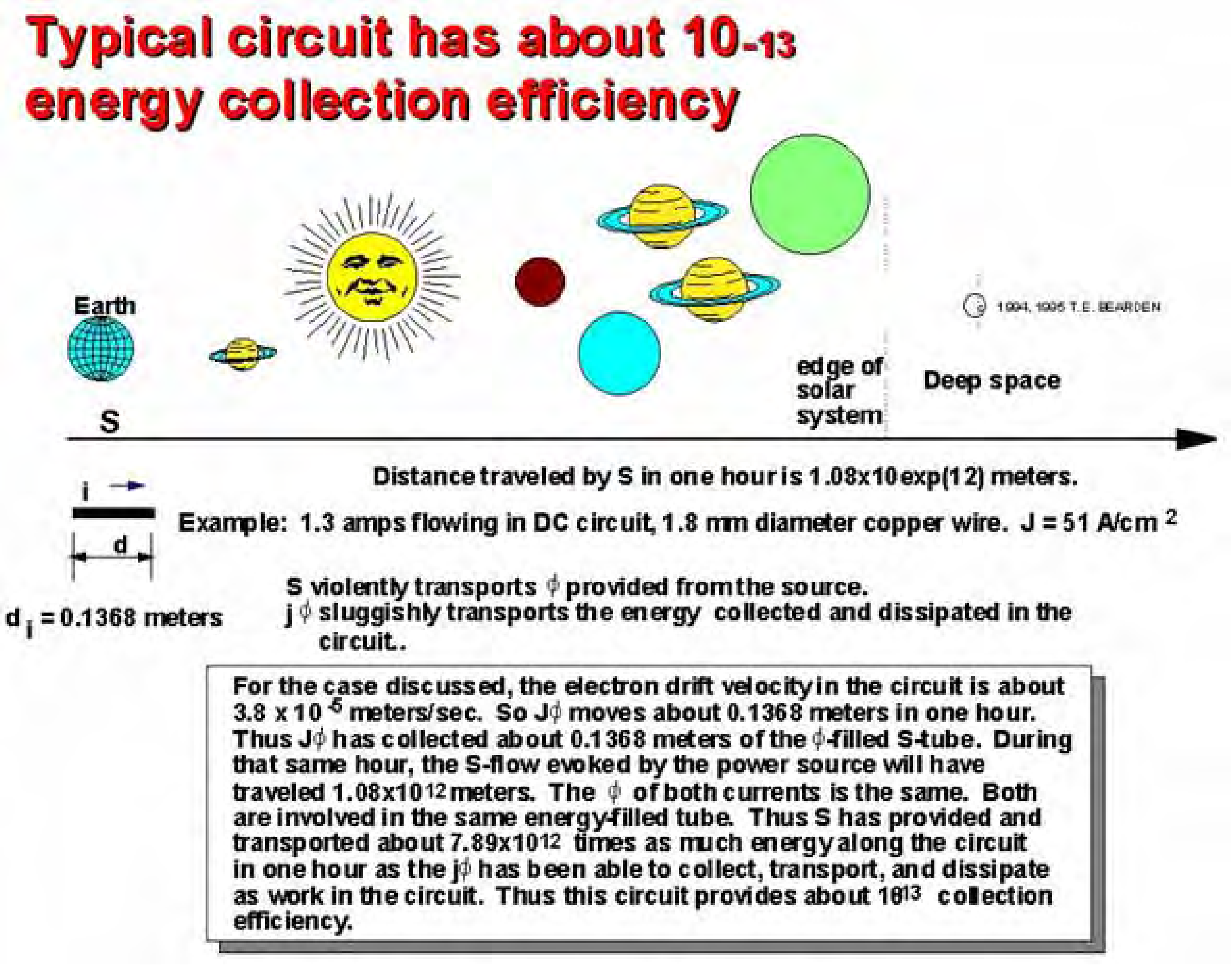

Typical circuit has about 10_{-13} energy collection efficiency

Earth
edge of solar system
Deep space
S
© 1994, 1995 T.E. BEARDEN
i
d

Distance traveled by S in one hour is 1.08x10exp(12) meters.

Example: 1.3 amps flowing in DC circuit, 1.8 mm diameter copper wire. $J = 51$ A/cm ^{2}

$d_i = 0.1368$ meters

S violently transports ϕ provided from the source.
j ϕ sluggishly transports the energy collected and dissipated in the circuit..

For the case discussed, the electron drift velocity in the circuit is about 3.8×10^{-5} meters/sec. So Jϕ moves about 0.1368 meters in one hour. Thus Jϕ has collected about 0.1368 meters of the ϕ-filled S-tube. During that same hour, the S-flow evoked by the power source will have traveled $1.08\text{x}10^{12}$ meters. The ϕ of both currents is the same. Both are involved in the same energy-filled tube. Thus S has provided and transported about $7.89\text{x}10^{12}$ times as much energy along the circuit in one hour as the jϕ has been able to collect, transport, and dissipate as work in the circuit. Thus this circuit provides about 10^{13} collection efficiency.

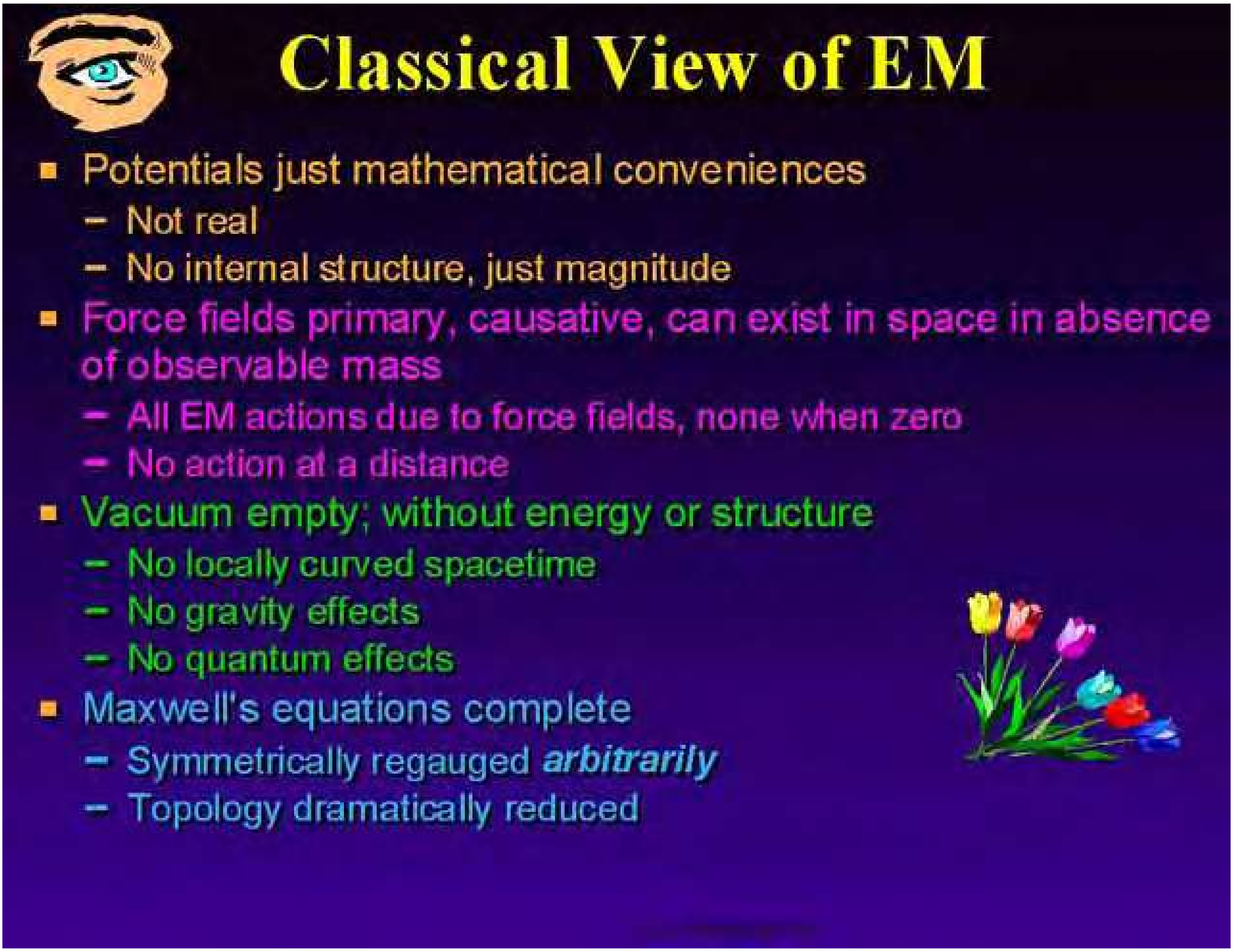

Classical View of EM

- Potentials just mathematical conveniences
  - Not real
  - No internal structure, just magnitude
- Force fields primary, causative, can exist in space in absence of observable mass
  - All EM actions due to force fields, none when zero
  - No action at a distance
- Vacuum empty; without energy or structure
  - No locally curved spacetime
  - No gravity effects
  - No quantum effects
- Maxwell's equations complete
  - Symmetrically regauged ***arbitrarily***
  - Topology dramatically reduced

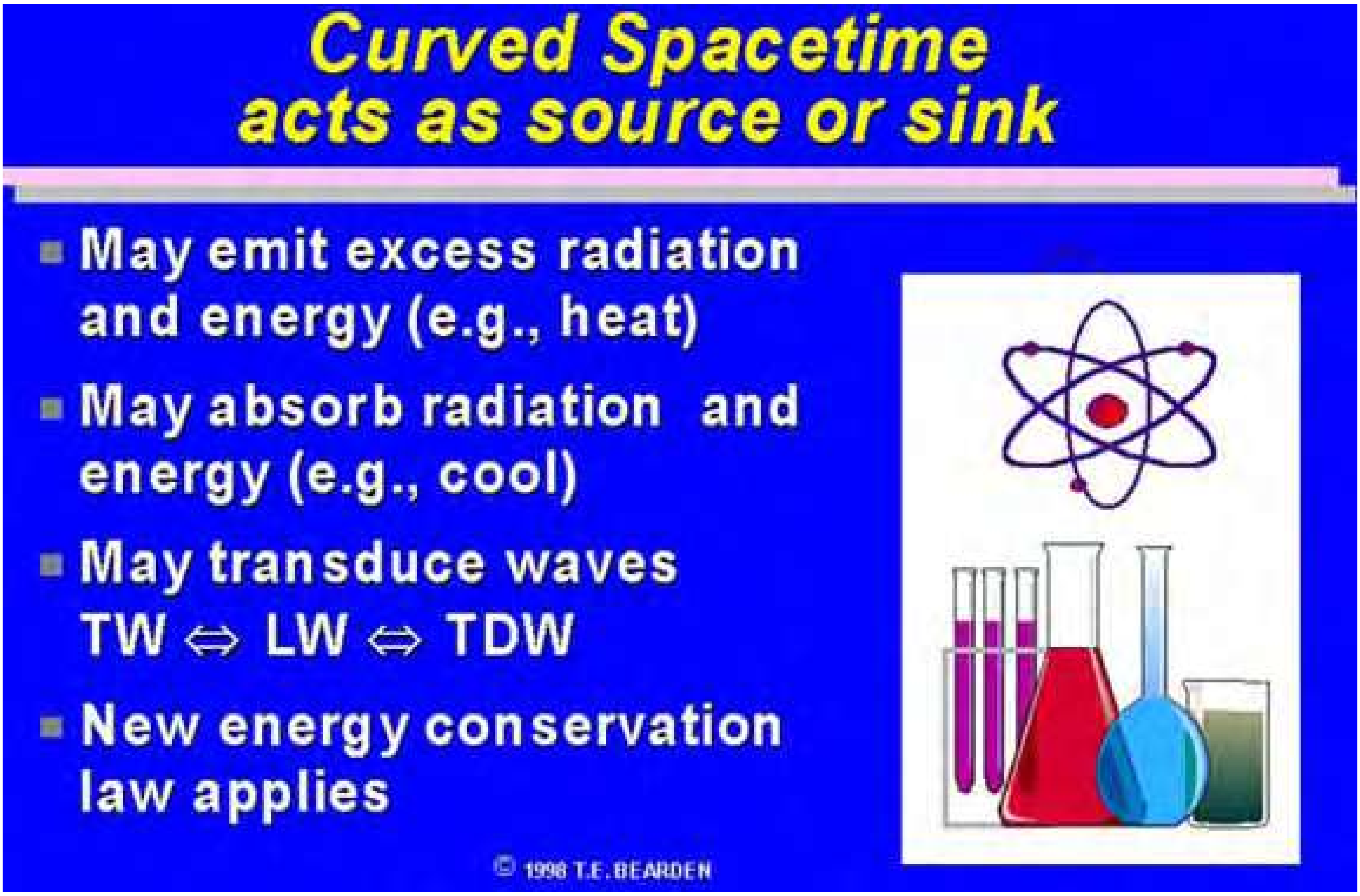

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E. BEARDEN

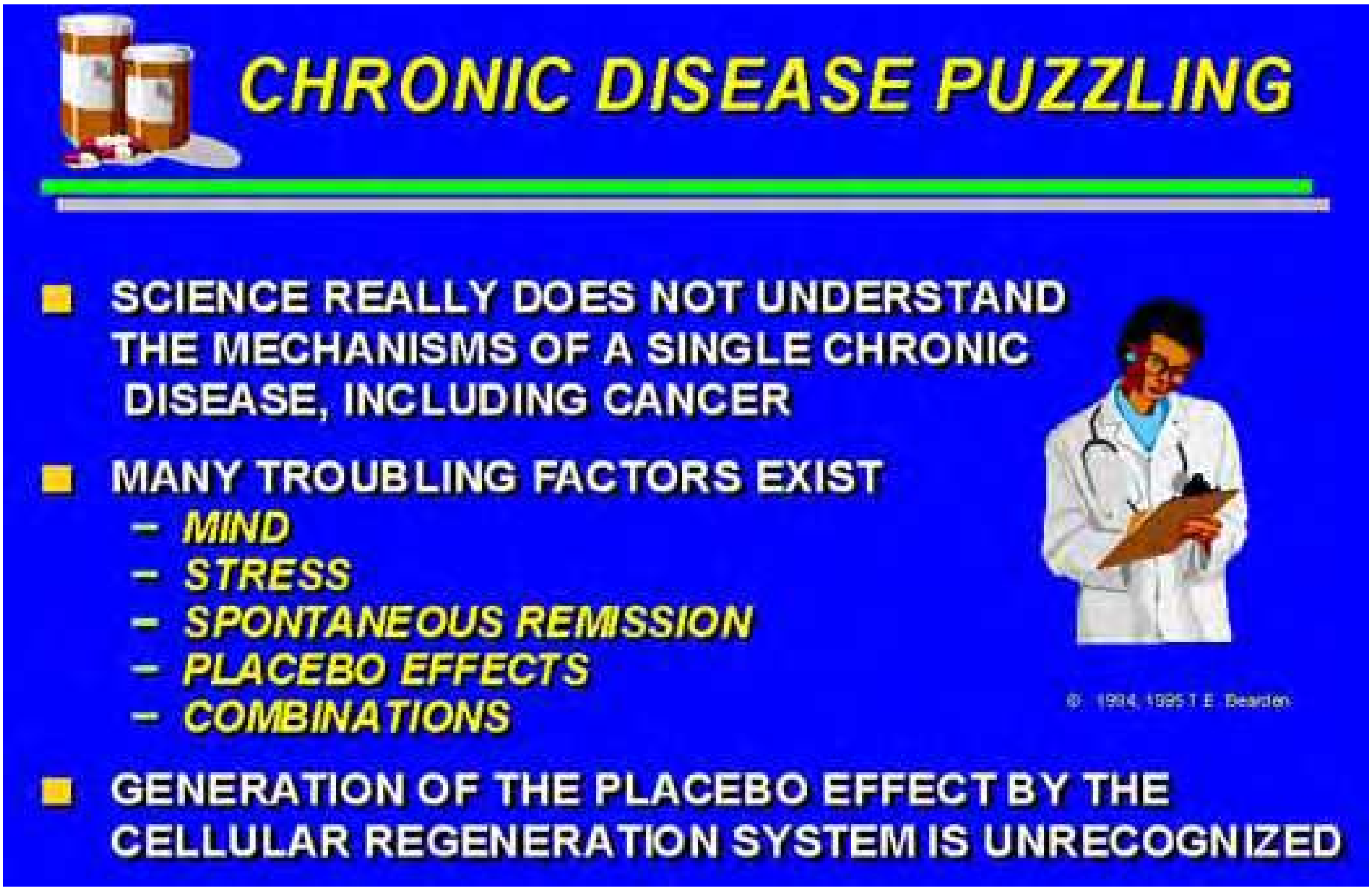

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

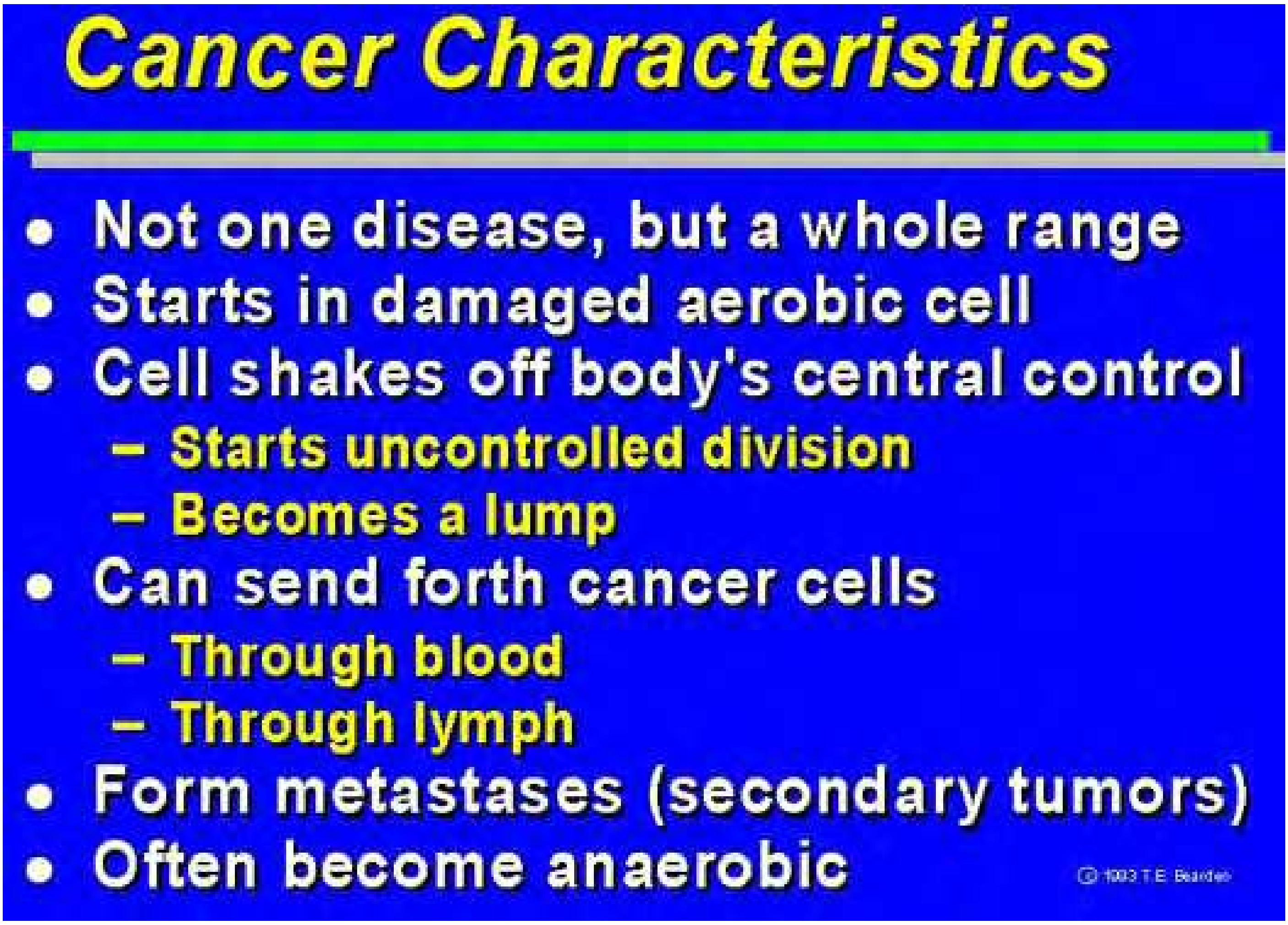

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

© 1993 T.E. Bearden

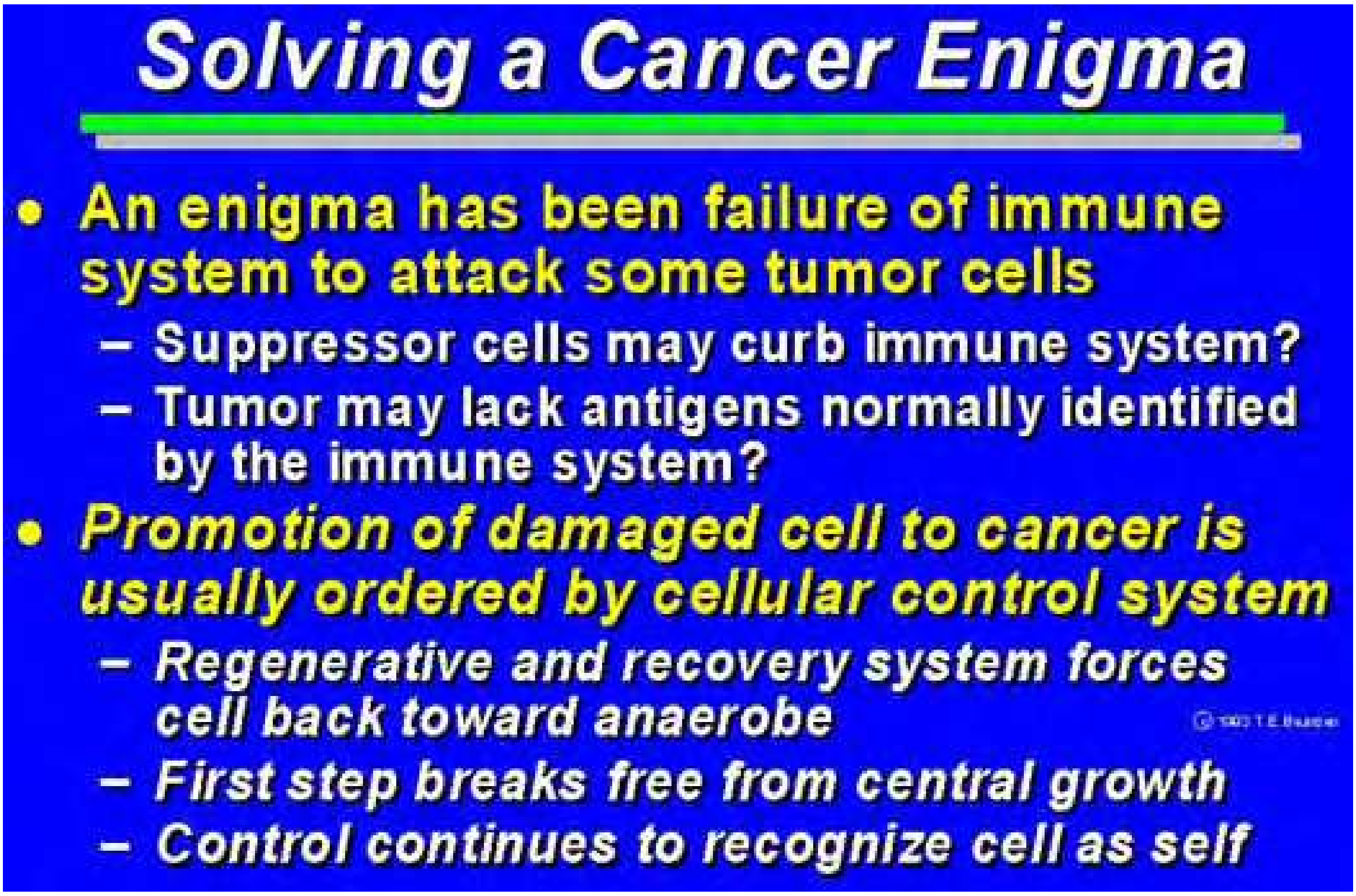

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

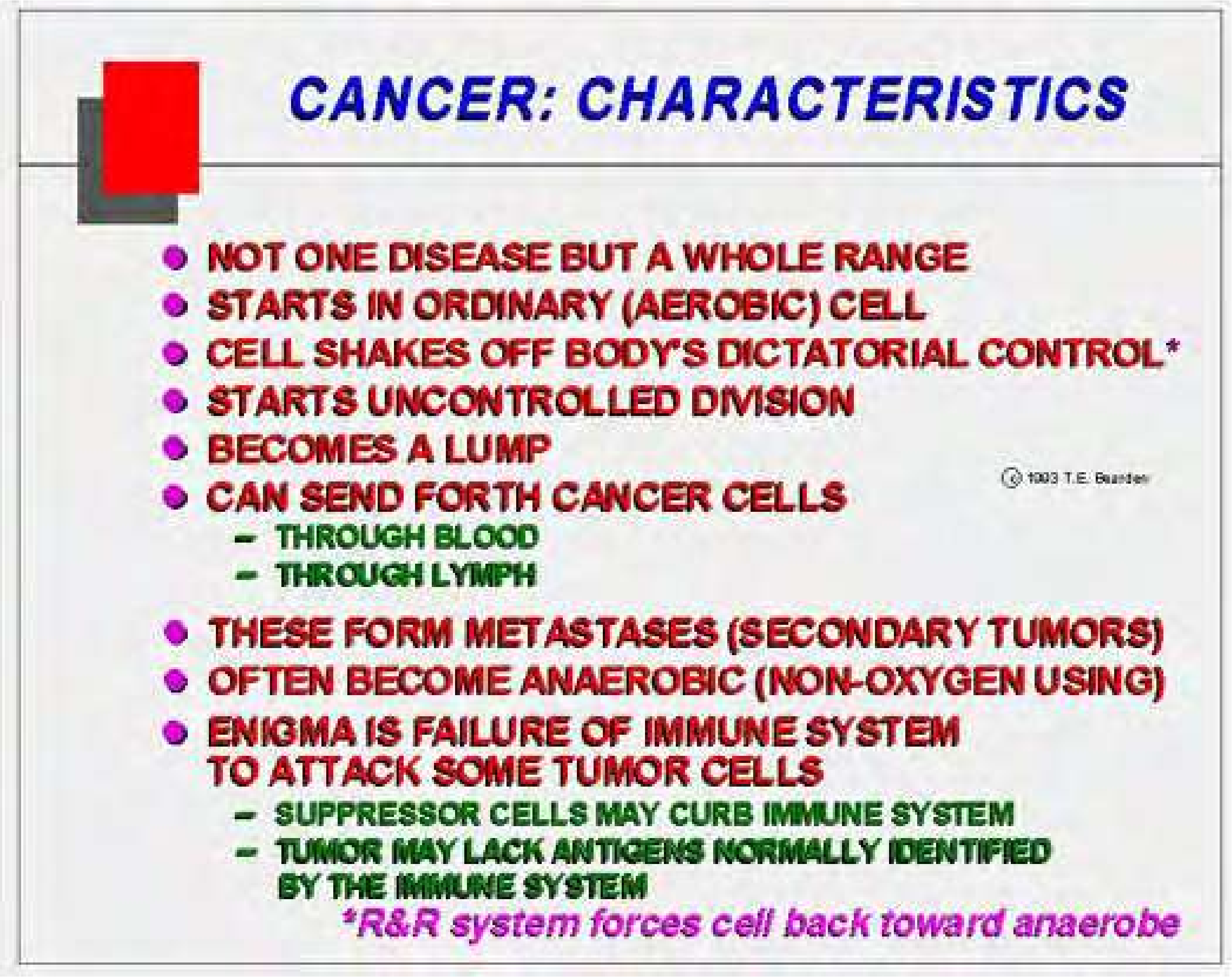

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

*R&R system forces cell back toward anaerobe

© 1993 T.E. Bearden

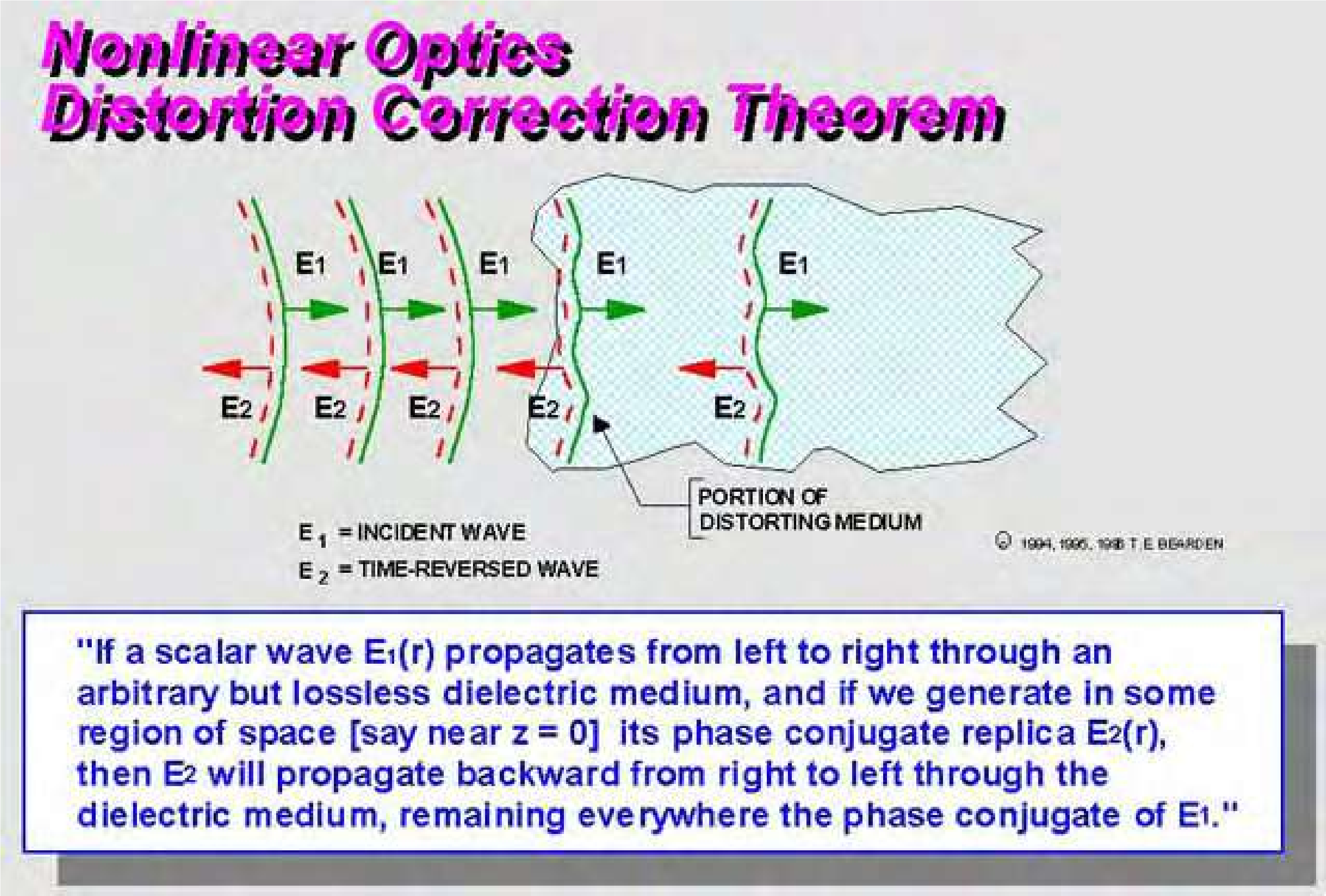

Nonlinear Optics
Distortion Correction Theorem

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

© 1994, 1995, 1998 T E BEARDEN

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

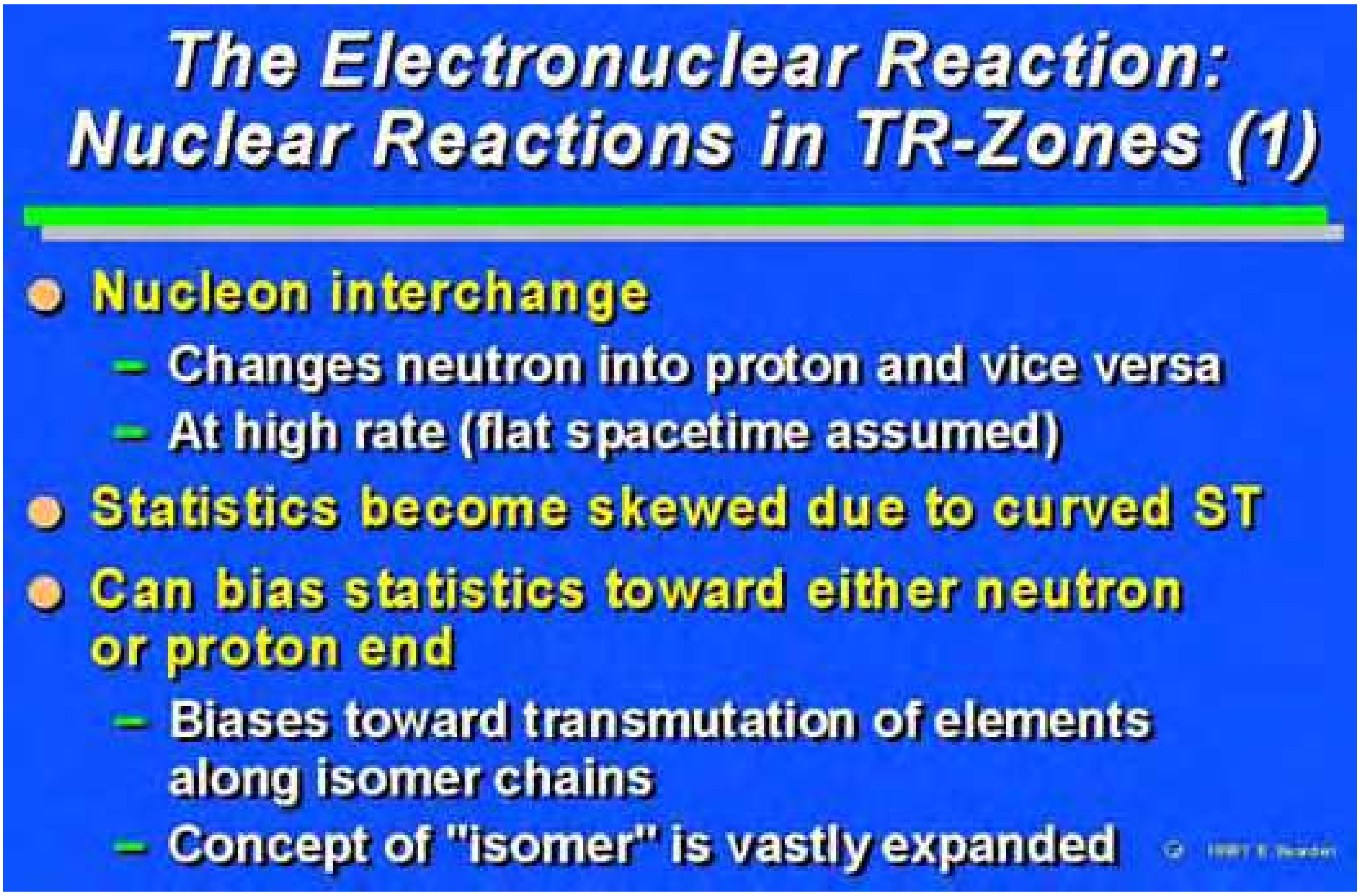

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- Nucleon interchange
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- Statistics become skewed due to curved ST
- Can bias statistics toward either neutron or proton end
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

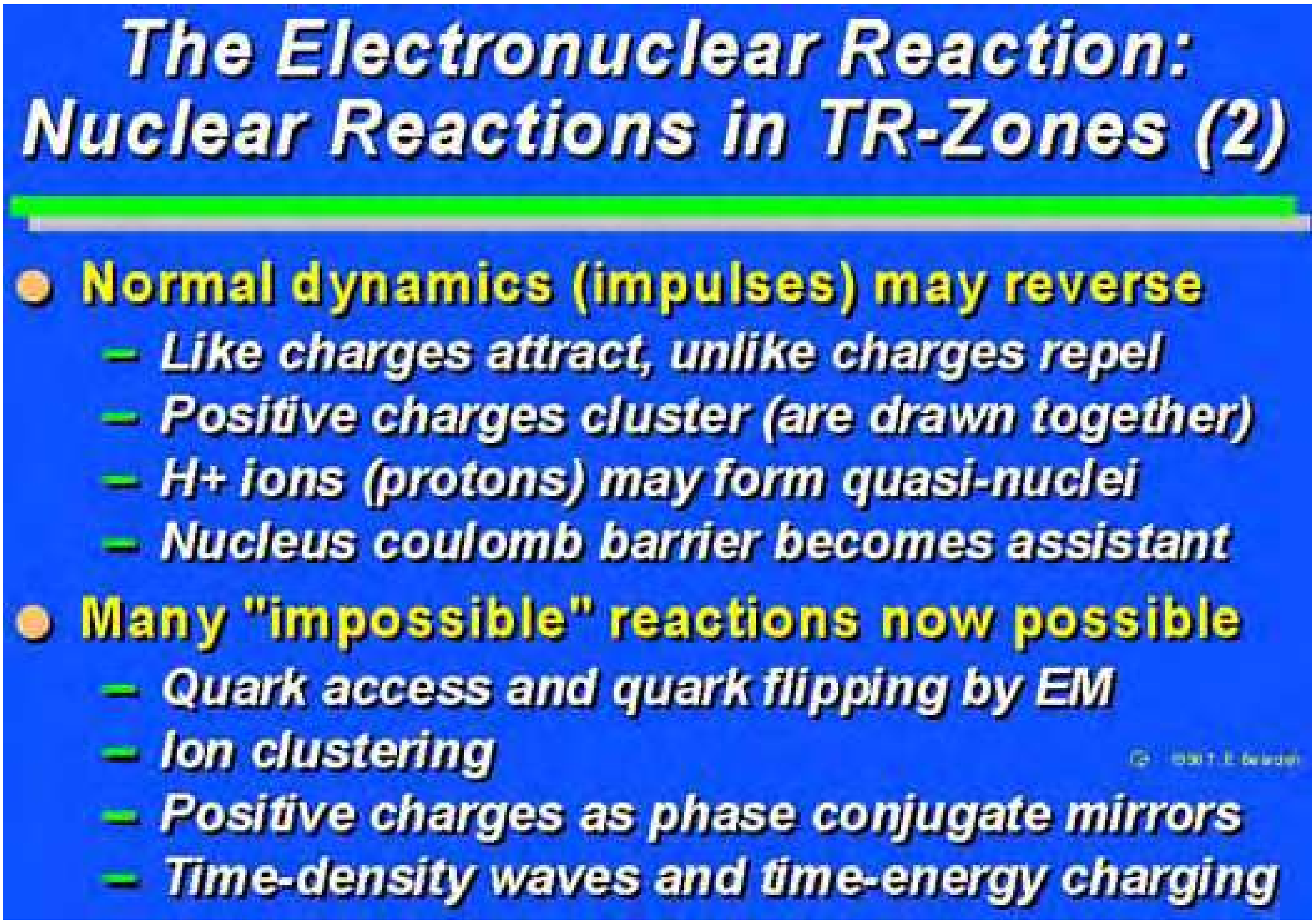

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- **Normal dynamics (impulses) may reverse**
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- **Many "impossible" reactions now possible**
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

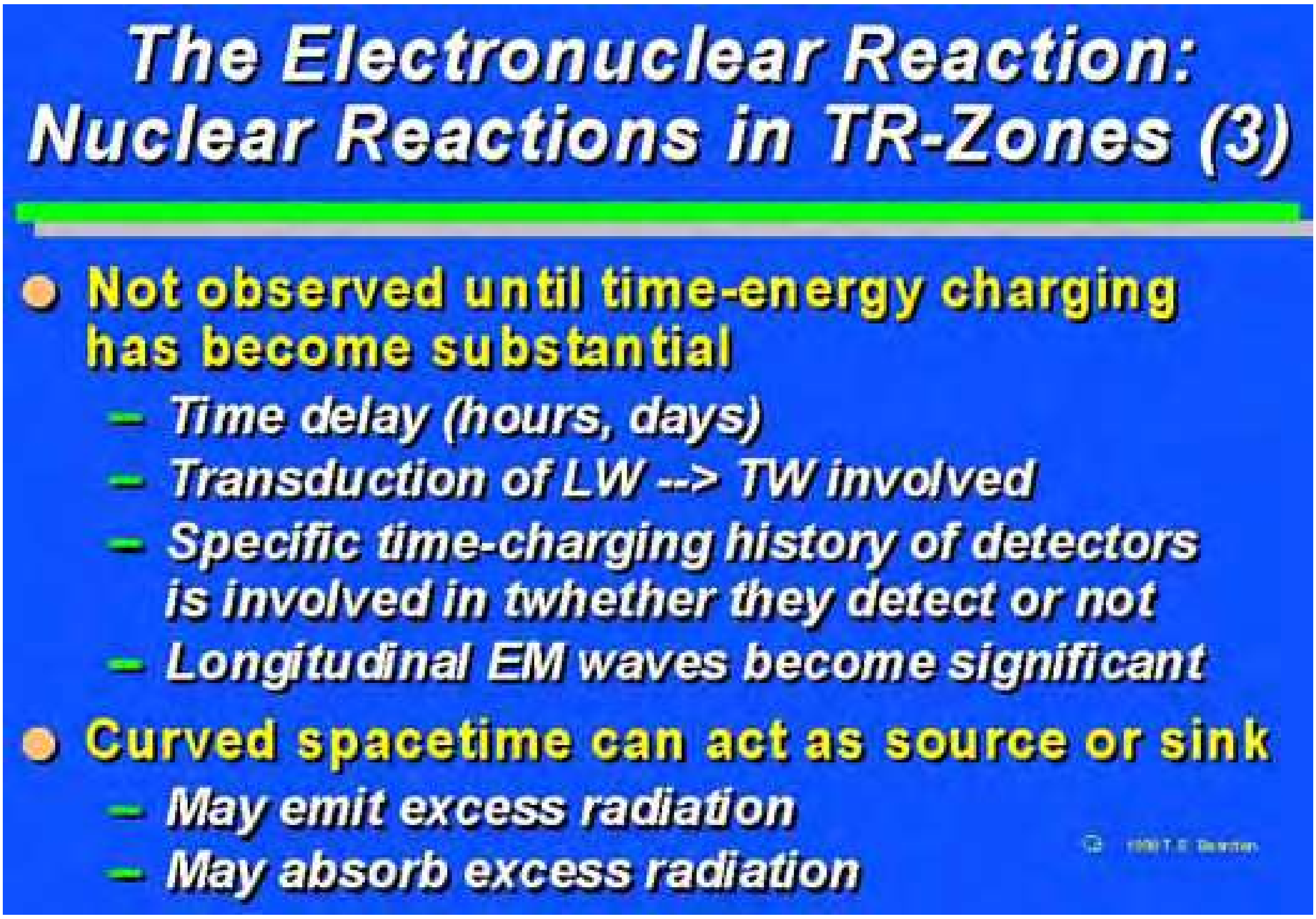

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- **Not observed until time-energy charging has become substantial**
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- **Curved spacetime can act as source or sink**
  - *May emit excess radiation*
  - *May absorb excess radiation*

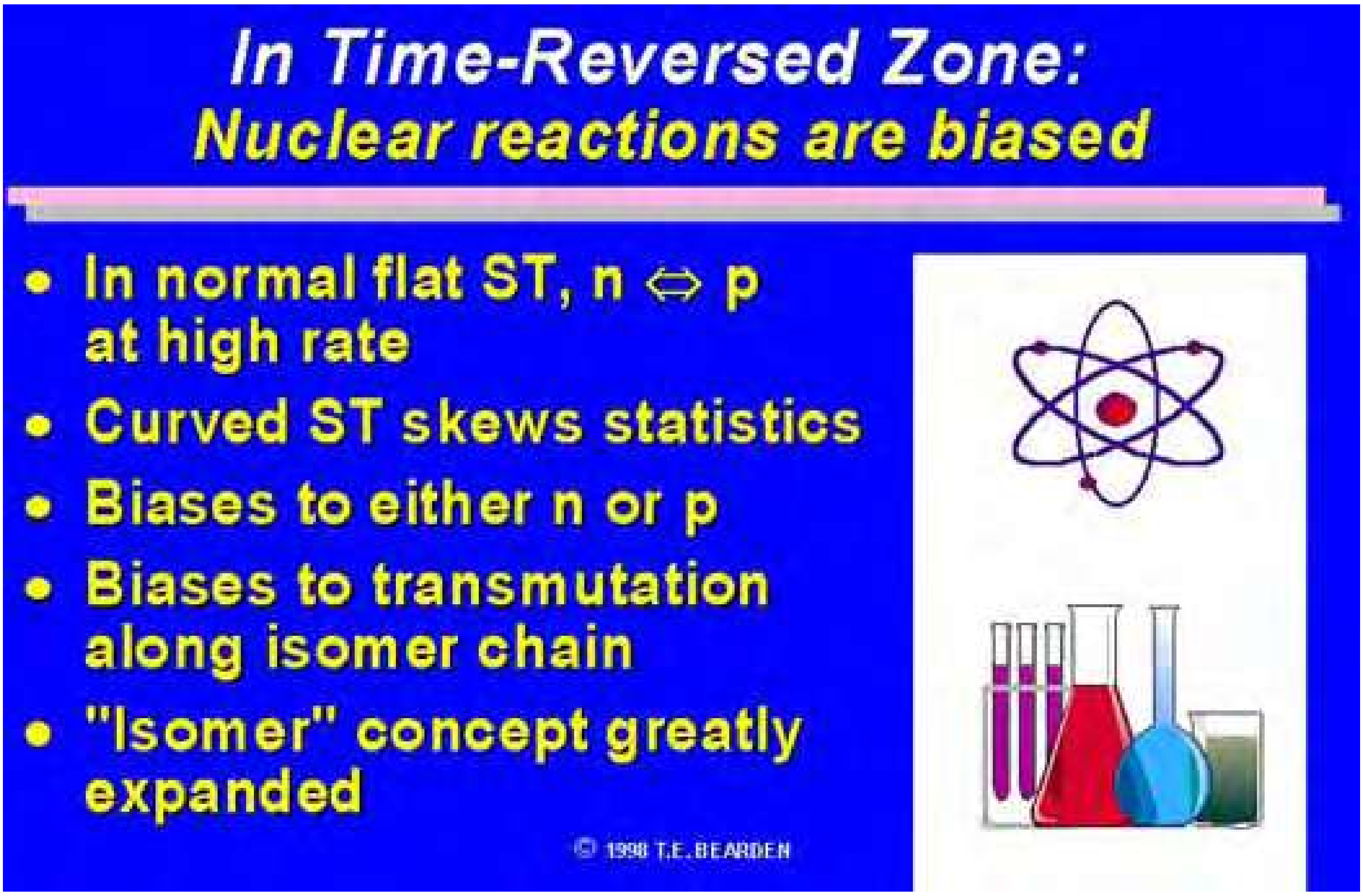

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T.E.BEARDEN

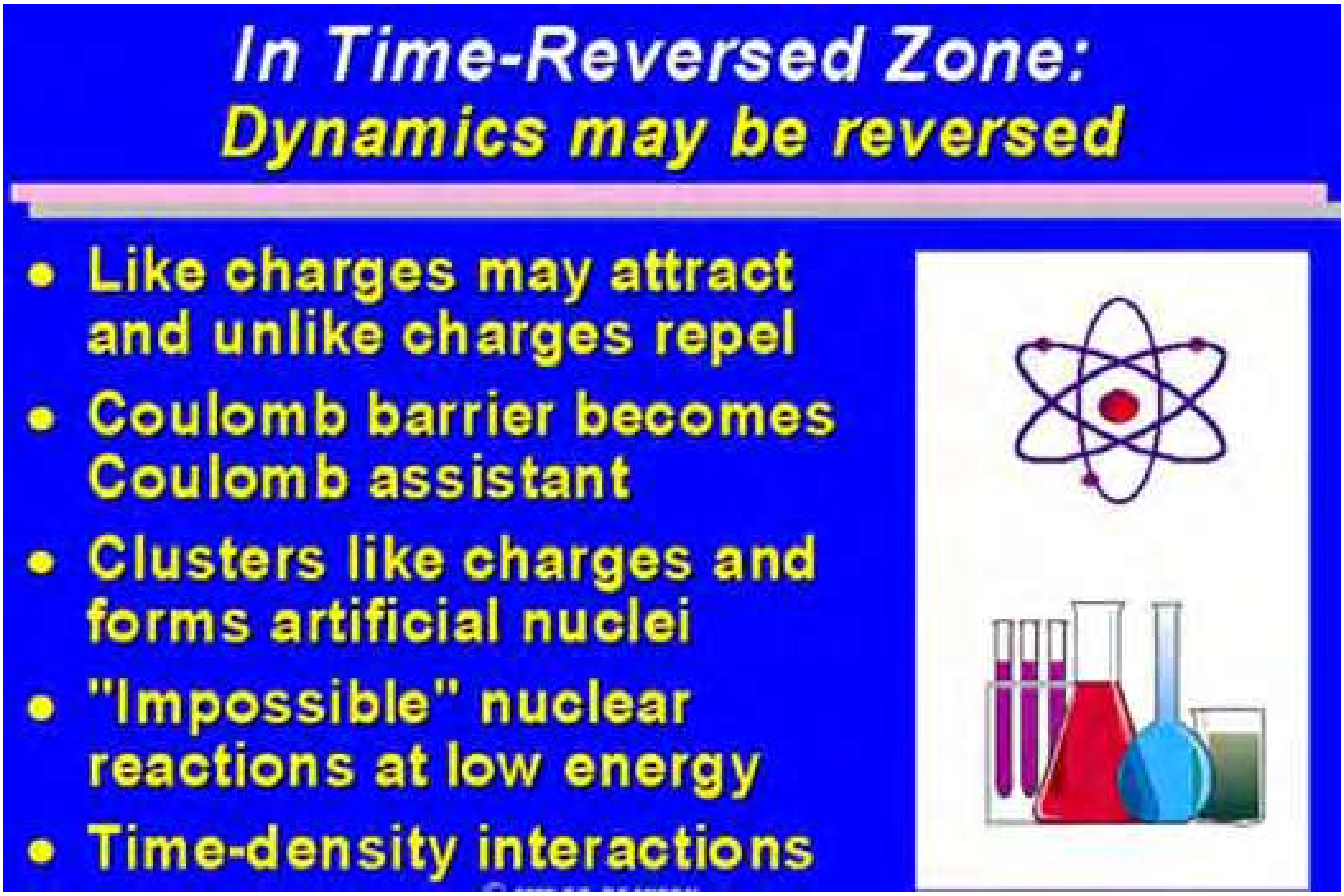

In Time-Reversed Zone:
Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

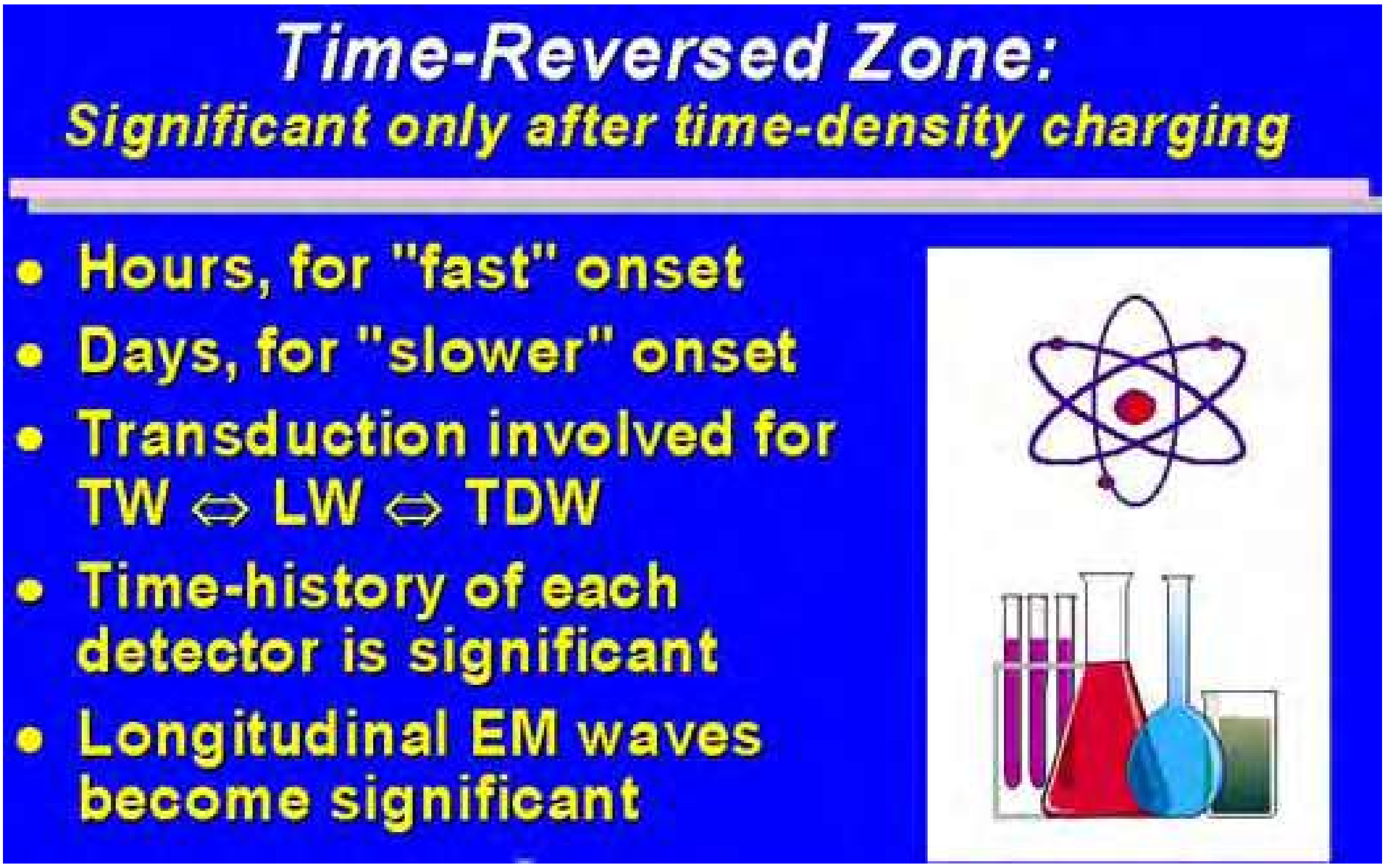

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

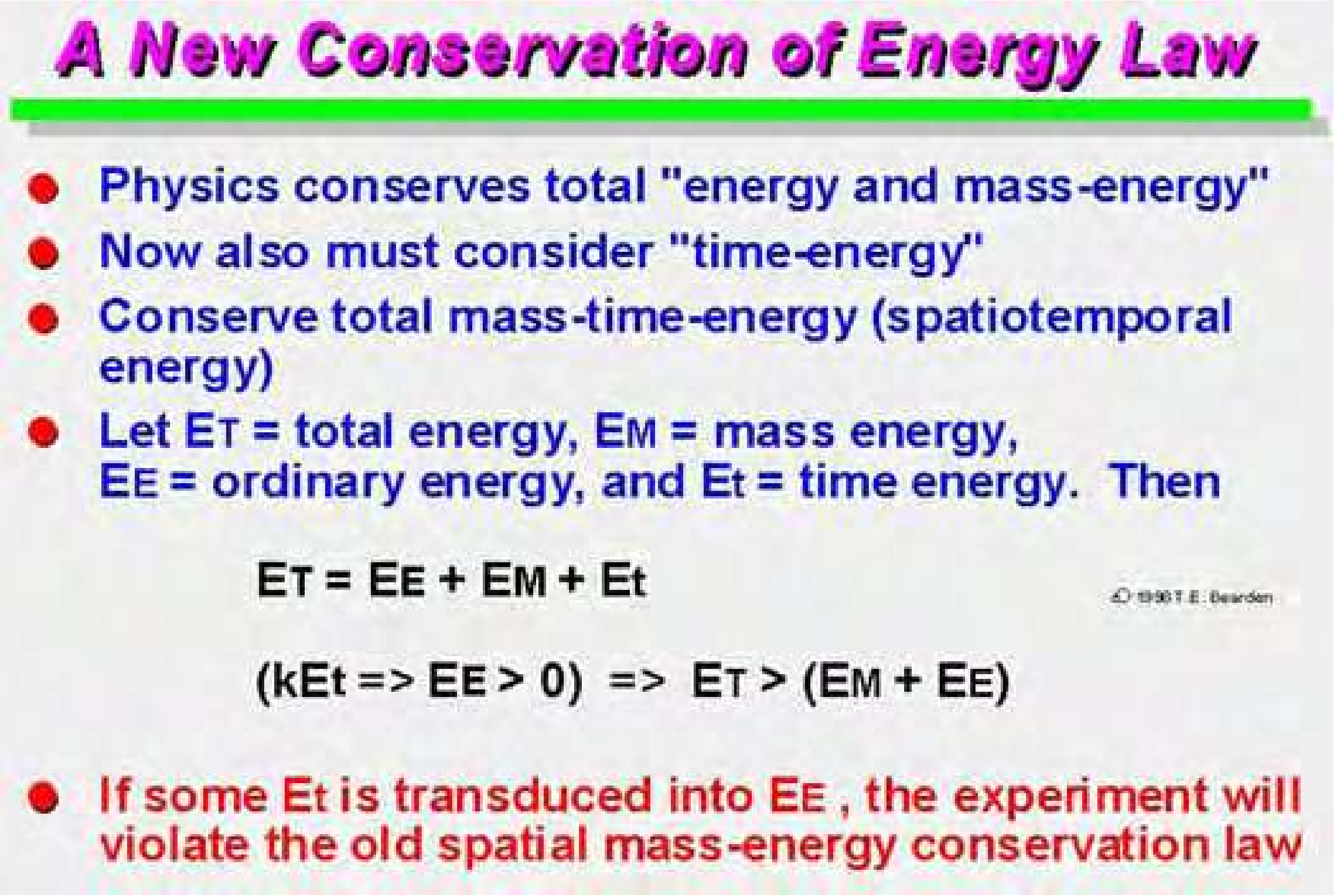

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

© 1998 T. E. Bearden

$$(kE_t \Rightarrow E_E > 0) \;\; \Rightarrow \;\; E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

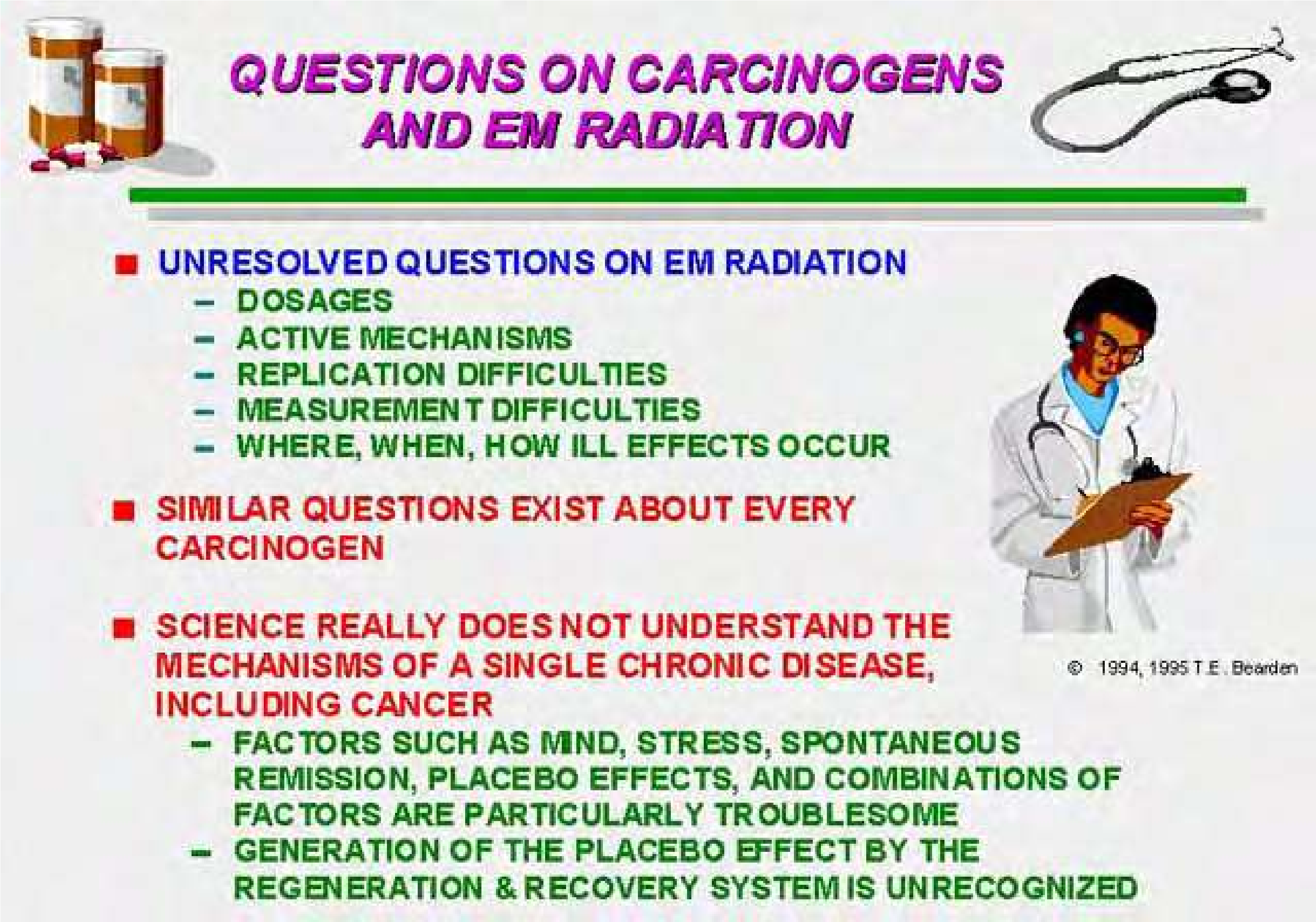

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

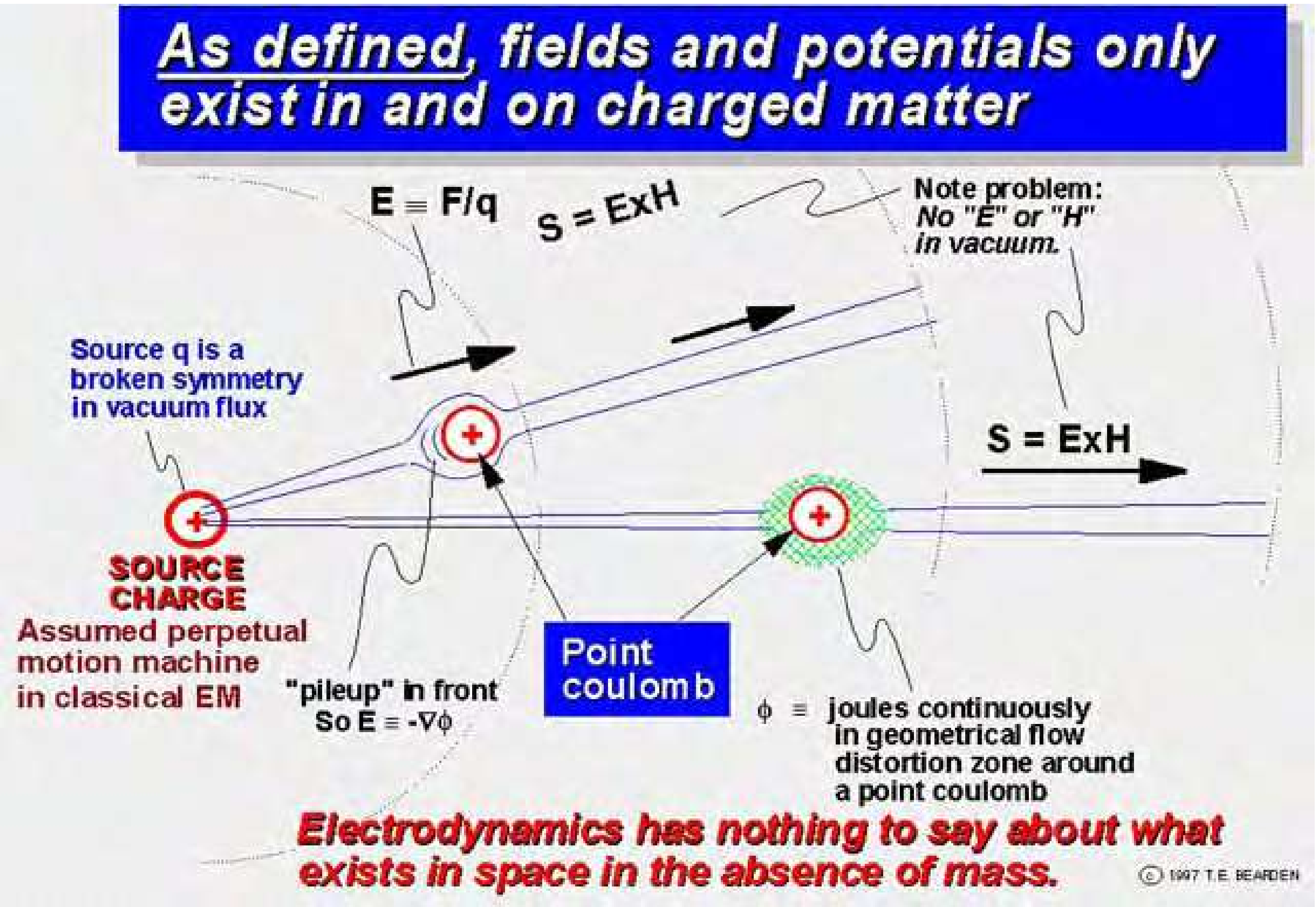

As defined, fields and potentials only exist in and on charged matter

$E \equiv F/q$

$S = E \times H$

Note problem:
No "E" or "H" in vacuum.

Source q is a broken symmetry in vacuum flux

$S = E \times H$

SOURCE CHARGE

Assumed perpetual motion machine in classical EM

"pileup" in front
So $E \equiv -\nabla\phi$

Point coulomb

$\phi \equiv$ joules continuously in geometrical flow distortion zone around a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

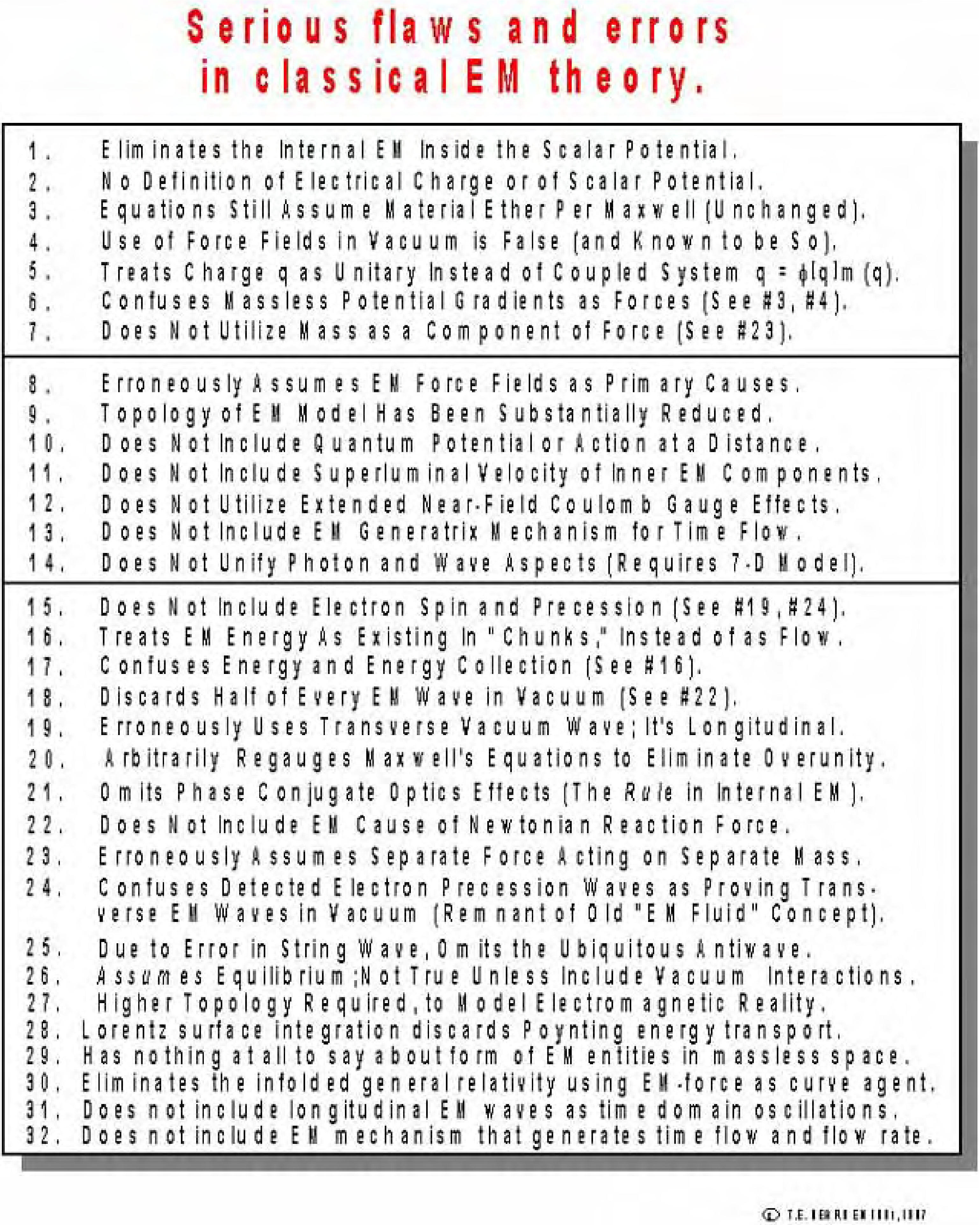

Serious flaws and errors in classical EM theory.

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi[q]m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

© T.E. BEARDEN 1991, 1997

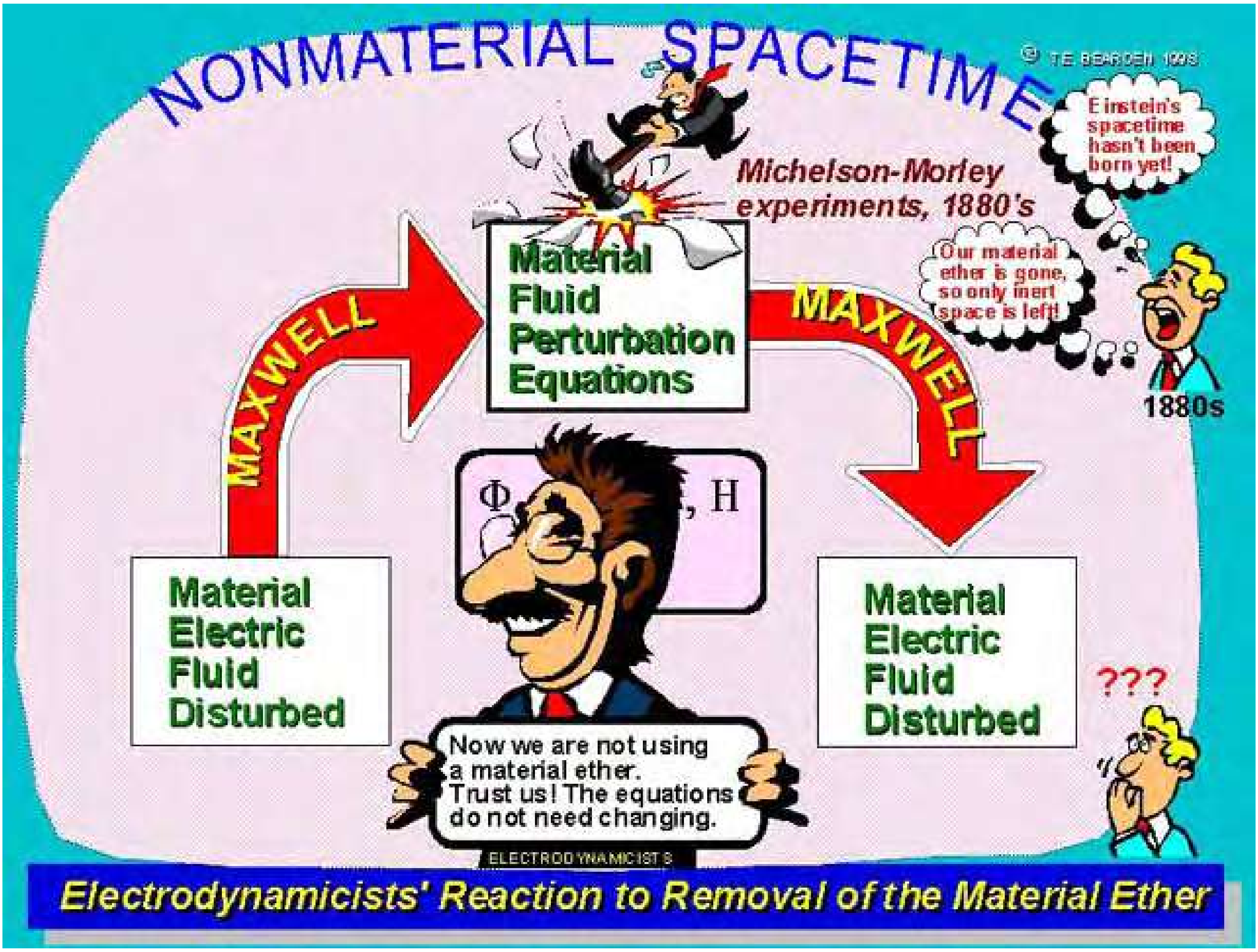
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Einstein's spacetime hasn't been born yet!
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left!
1880s
Material Fluid Perturbation Equations
MAXWELL
MAXWELL
Φ
, H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

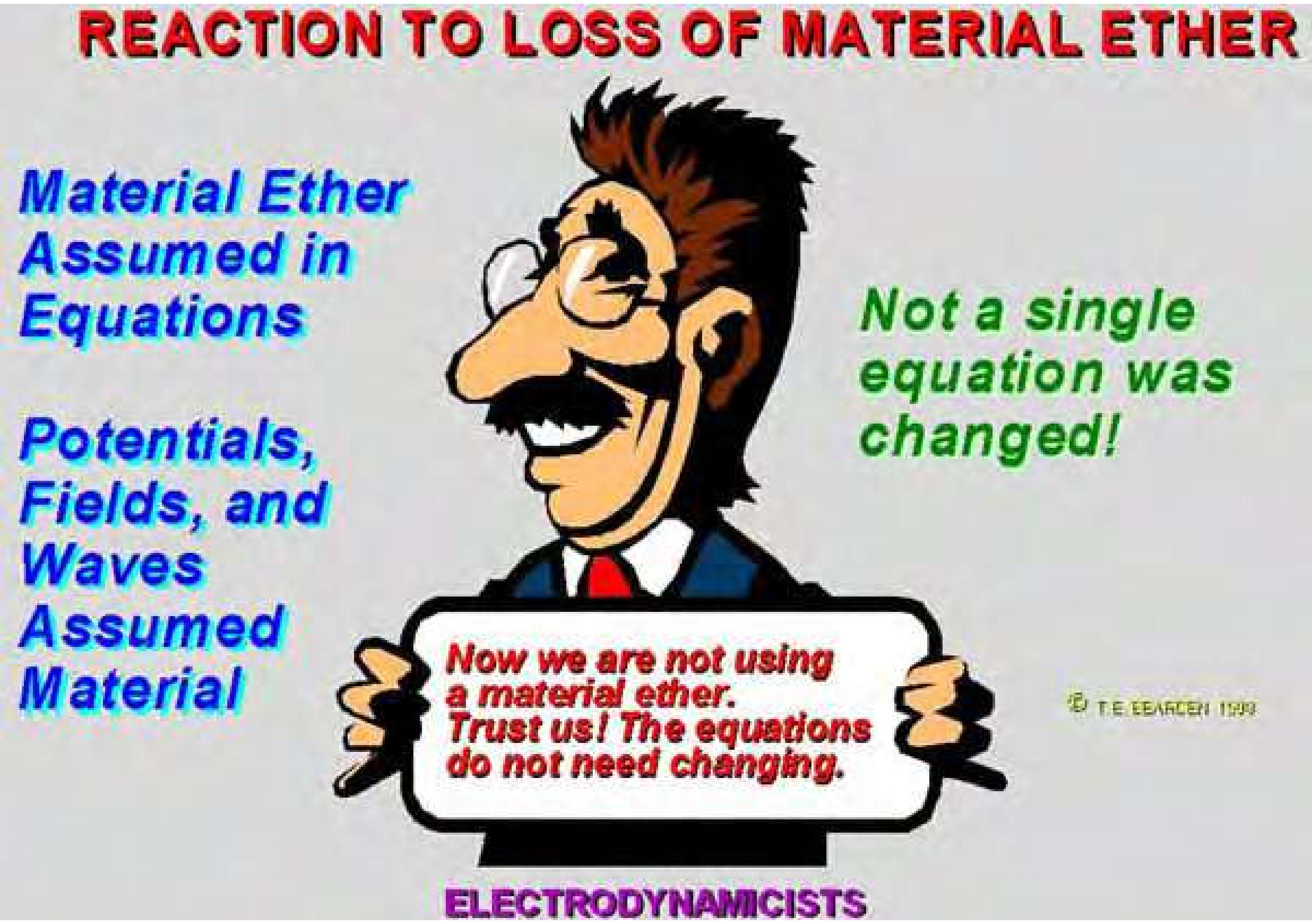

REACTION TO LOSS OF MATERIAL ETHER

Material Ether
Assumed in
Equations

Potentials,
Fields, and
Waves
Assumed
Material

Not a single
equation was
changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.

© T.E. BEARDEN 1999

ELECTRODYNAMICISTS

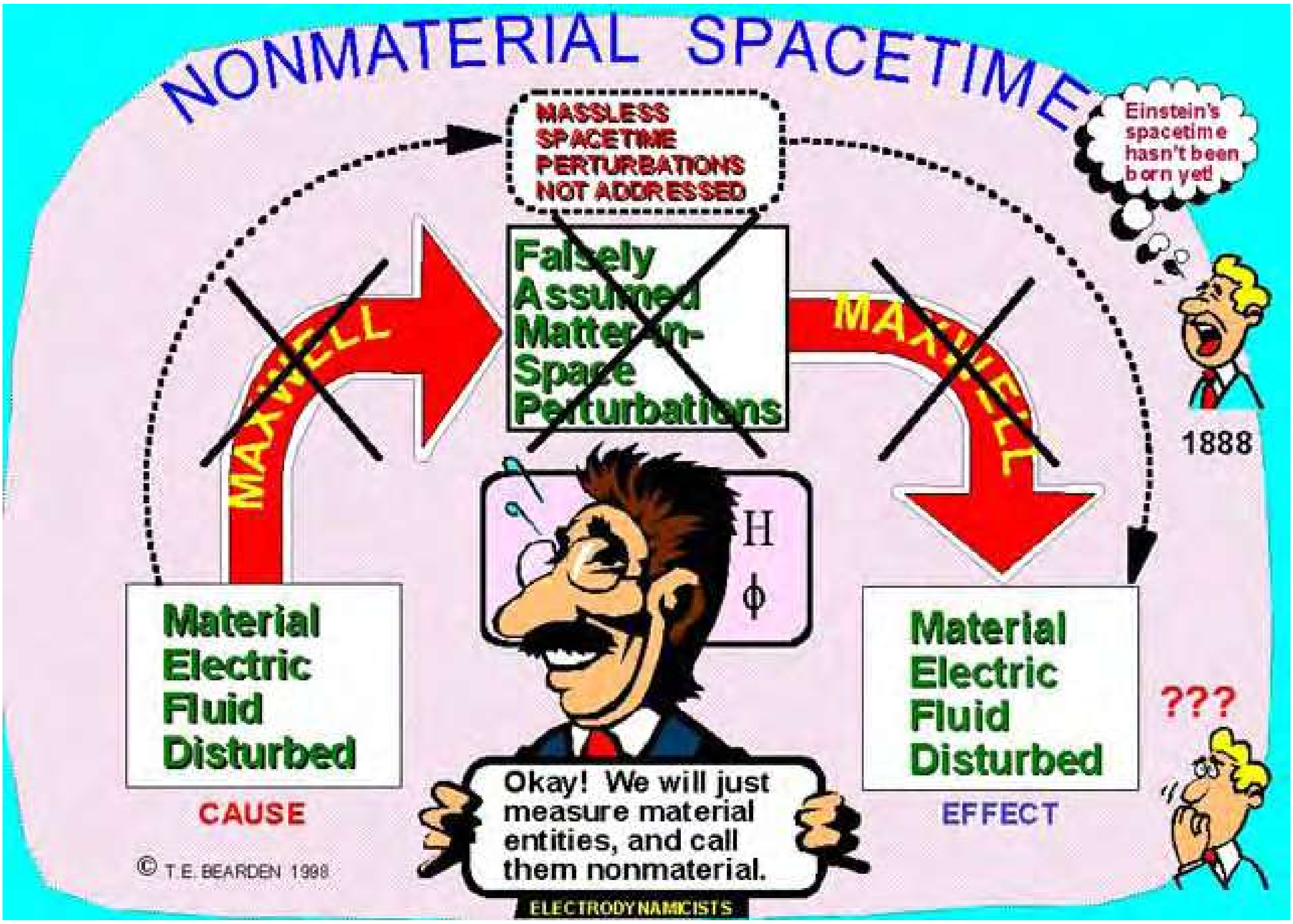
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
1888
MAXWELL
Falsely Assumed Matter-in-Space Perturbations
MAXWELL
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E. BEARDEN 1998

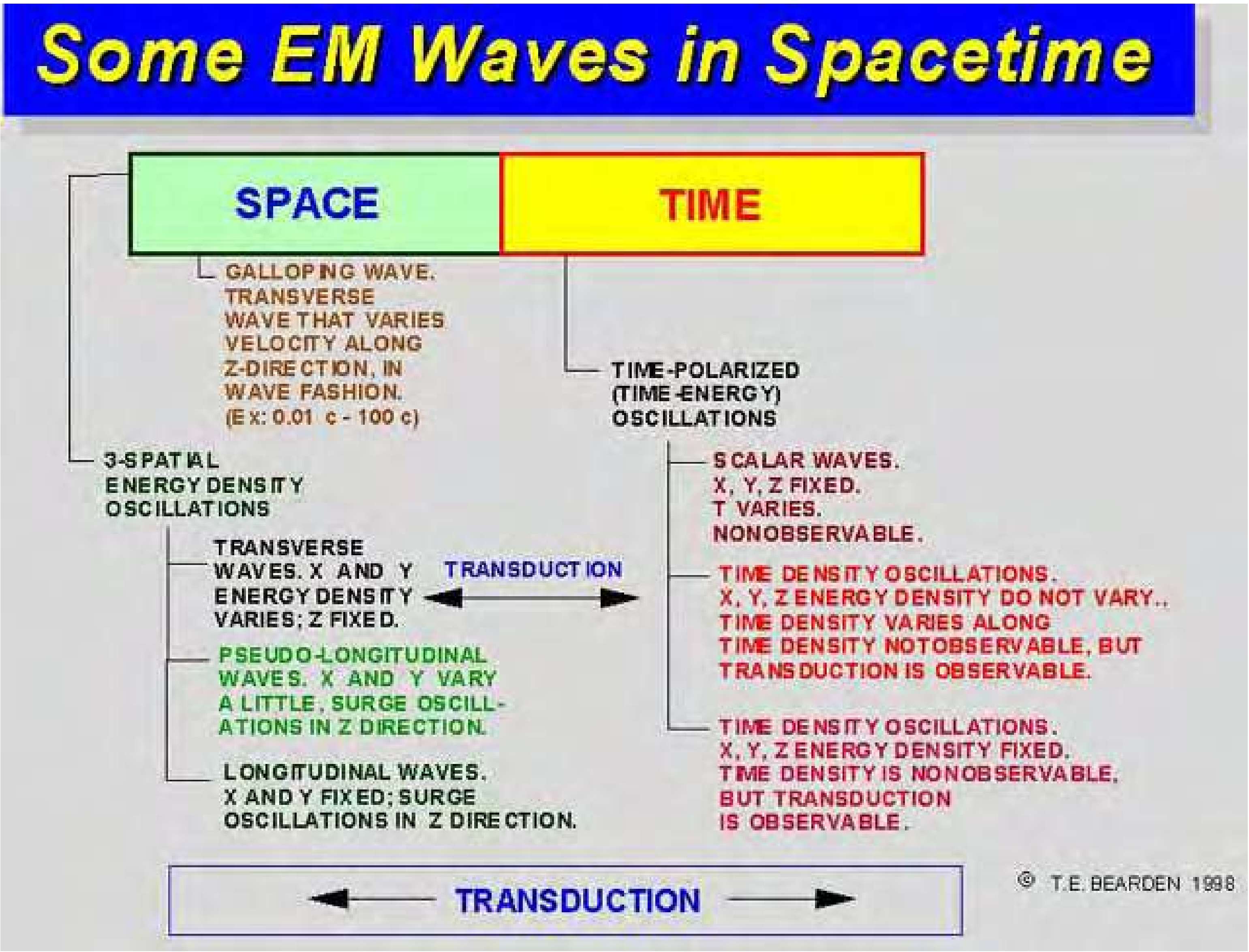

Some EM Waves in Spacetime

SPACE

TIME

GALLOPING WAVE.
TRANSVERSE
WAVE THAT VARIES
VELOCITY ALONG
Z-DIRECTION, IN
WAVE FASHION.
(Ex: 0.01 c - 100 c)

TIME-POLARIZED
(TIME-ENERGY)
OSCILLATIONS

3-SPATIAL
ENERGY DENSITY
OSCILLATIONS

SCALAR WAVES.
X, Y, Z FIXED.
T VARIES.
NONOBSERVABLE.

TRANSVERSE
WAVES. X AND Y
ENERGY DENSITY
VARIES; Z FIXED.

TRANSDUCTION

TIME DENSITY OSCILLATIONS.
X, Y, Z ENERGY DENSITY DO NOT VARY..
TIME DENSITY VARIES ALONG
TIME DENSITY NOTOBSERVABLE, BUT
TRANSDUCTION IS OBSERVABLE.

PSEUDO-LONGITUDINAL
WAVES. X AND Y VARY
A LITTLE, SURGE OSCILL-
ATIONS IN Z DIRECTION.

TIME DENSITY OSCILLATIONS.
X, Y, Z ENERGY DENSITY FIXED.
TIME DENSITY IS NONOBSERVABLE,
BUT TRANSDUCTION
IS OBSERVABLE.

LONGITUDINAL WAVES.
X AND Y FIXED; SURGE
OSCILLATIONS IN Z DIRECTION.

TRANSDUCTION

© T.E. BEARDEN 1998

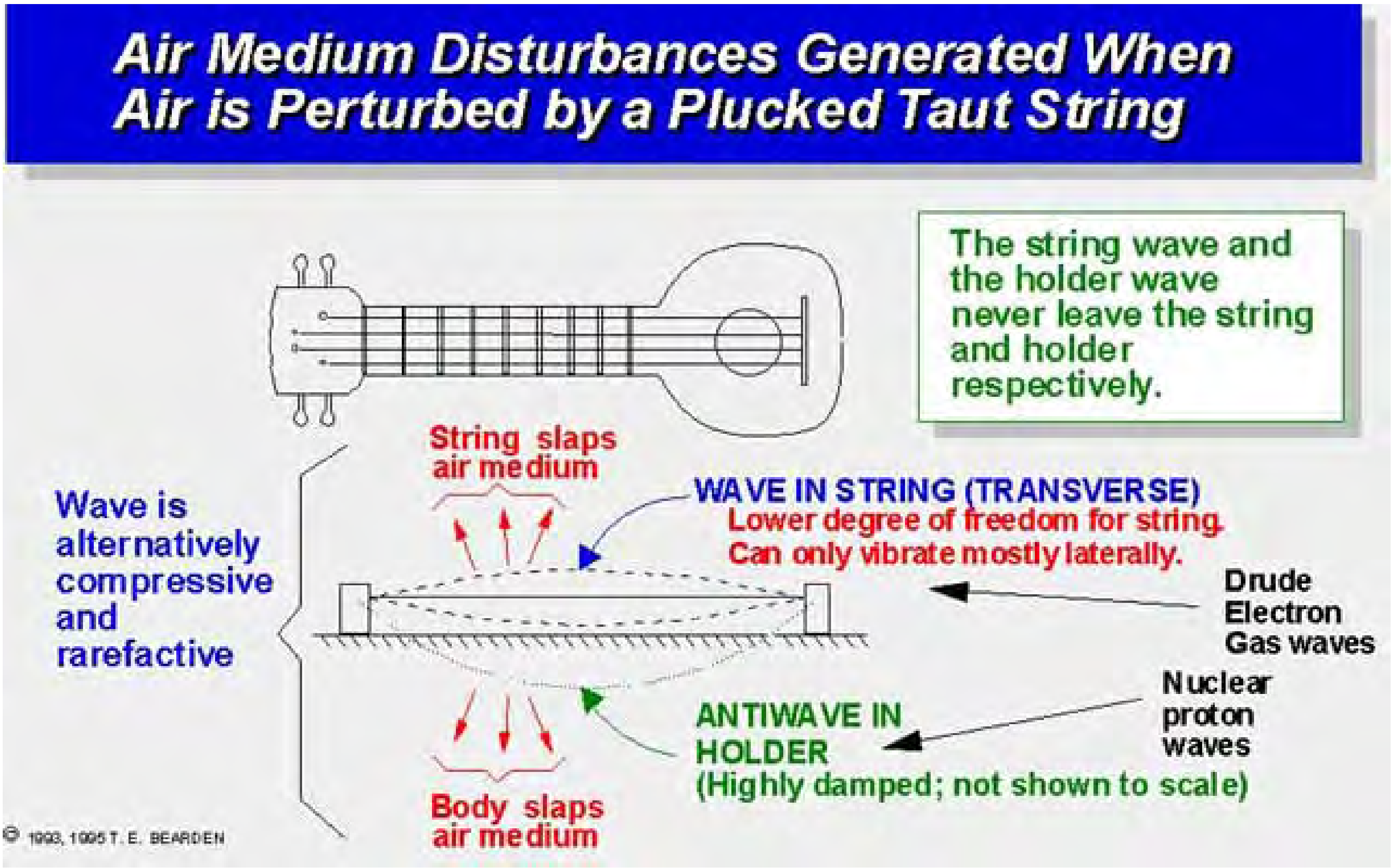

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

Wave is alternatively compressive and rarefactive

String slaps air medium

WAVE IN STRING (TRANSVERSE)

Lower degree of freedom for string.
Can only vibrate mostly laterally.

Drude Electron Gas waves

Nuclear proton waves

ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)

Body slaps air medium

© 1993, 1995 T. E. BEARDEN

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

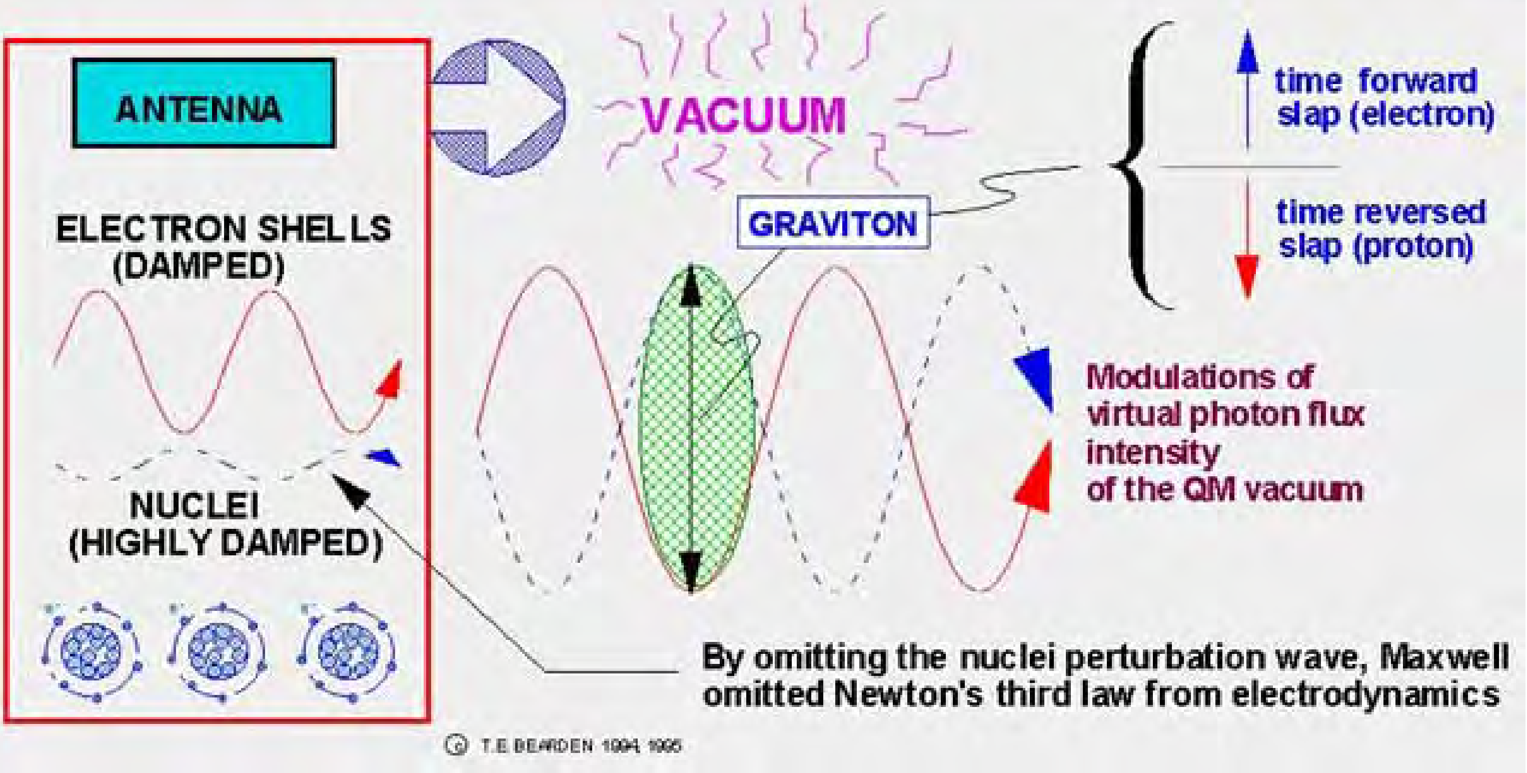
ANTENNA
ELECTRON SHELLS (DAMPED)
NUCLEI (HIGHLY DAMPED)
VACUUM
GRAVITON
time forward slap (electron)
time reversed slap (proton)
Modulations of virtual photon flux intensity of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

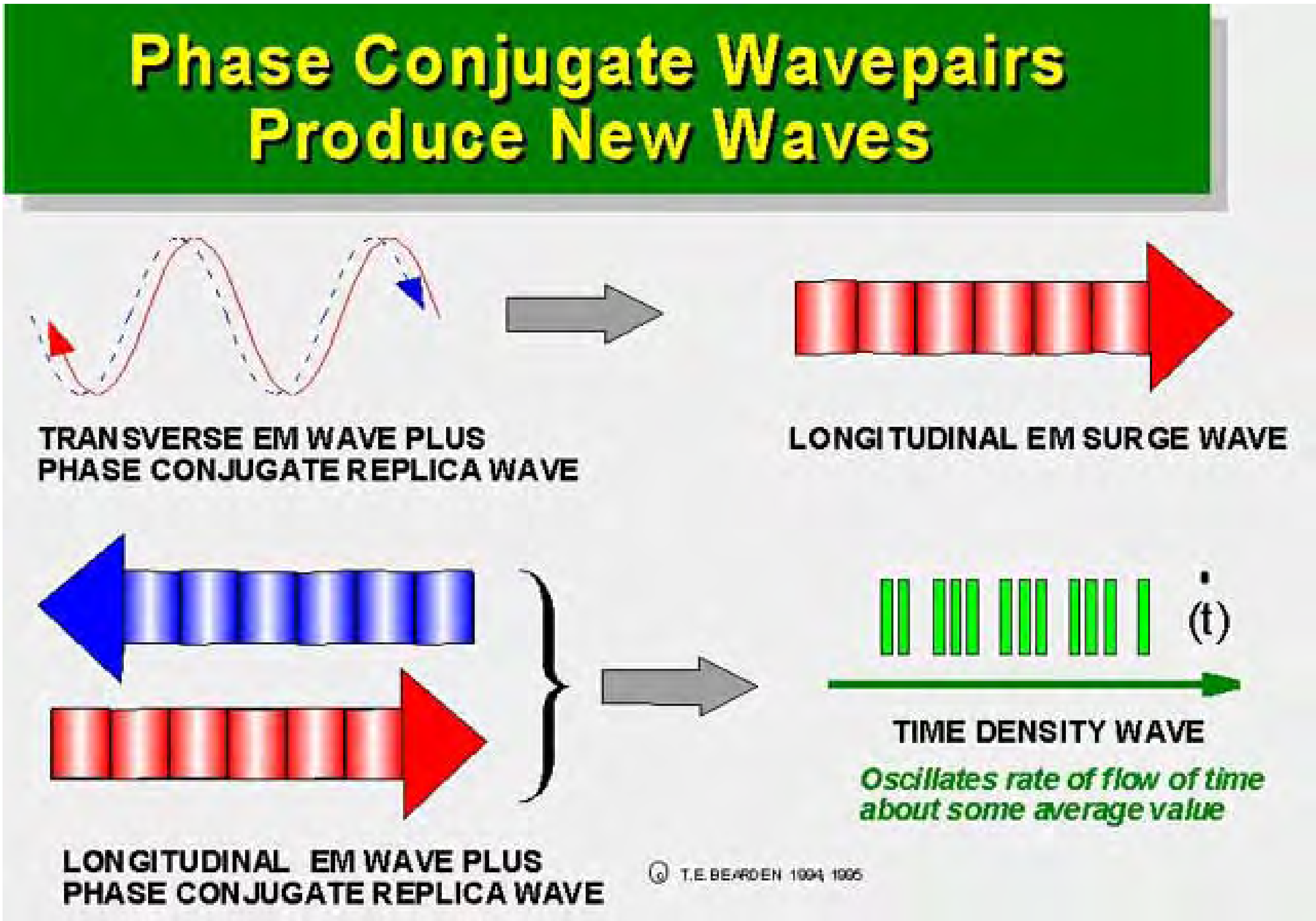

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM SURGE WAVE

$(\dot{t})$

TIME DENSITY WAVE

Oscillates rate of flow of time about some average value

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

© T.E. BEARDEN 1994, 1995

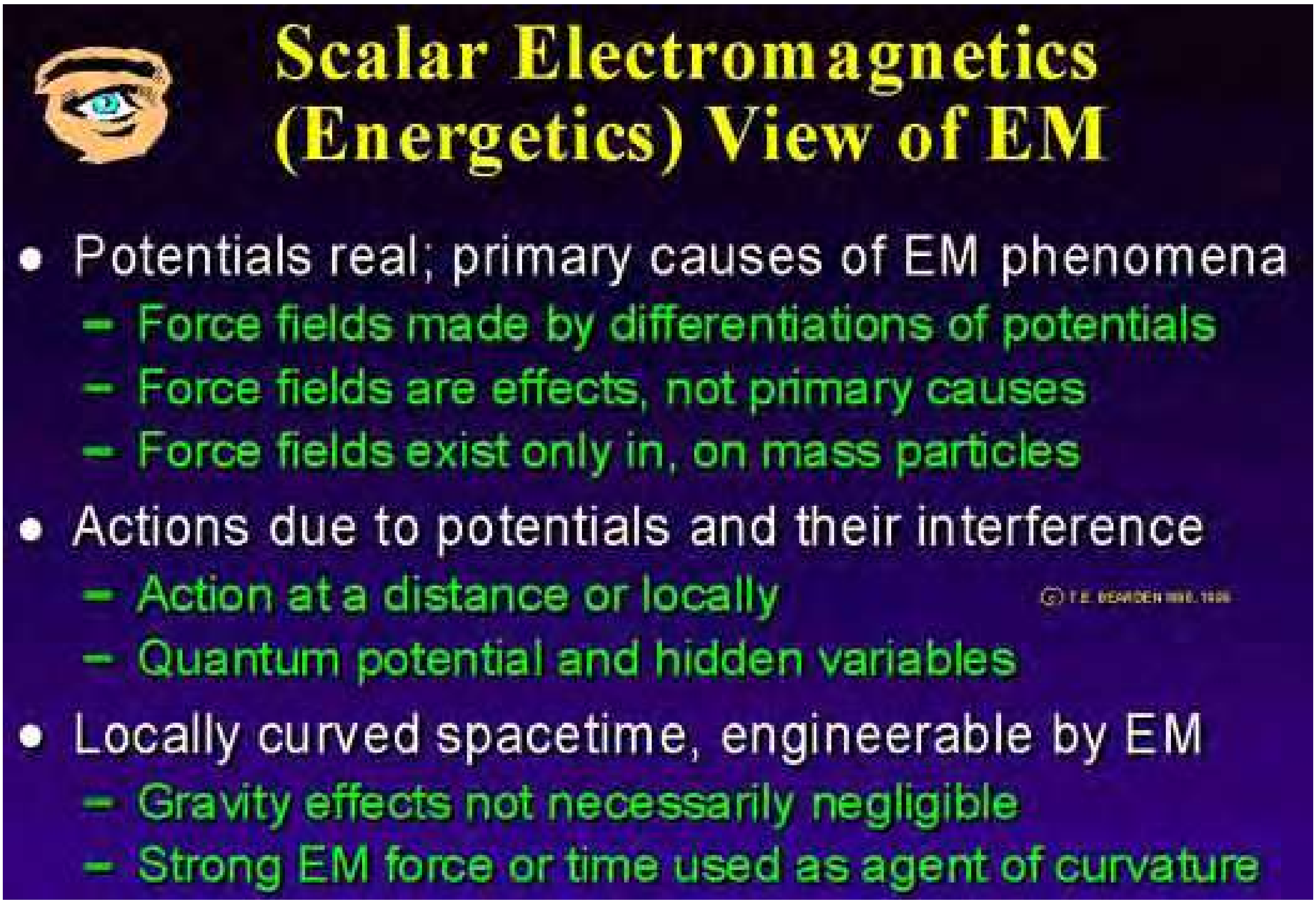

Scalar Electromagnetics (Energetics) View of EM

- Potentials real; primary causes of EM phenomena
  - Force fields made by differentiations of potentials
  - Force fields are effects, not primary causes
  - Force fields exist only in, on mass particles
- Actions due to potentials and their interference
  - Action at a distance or locally
  - Quantum potential and hidden variables
- Locally curved spacetime, engineerable by EM
  - Gravity effects not necessarily negligible
  - Strong EM force or time used as agent of curvature

© T.E. BEARDEN [illegible]

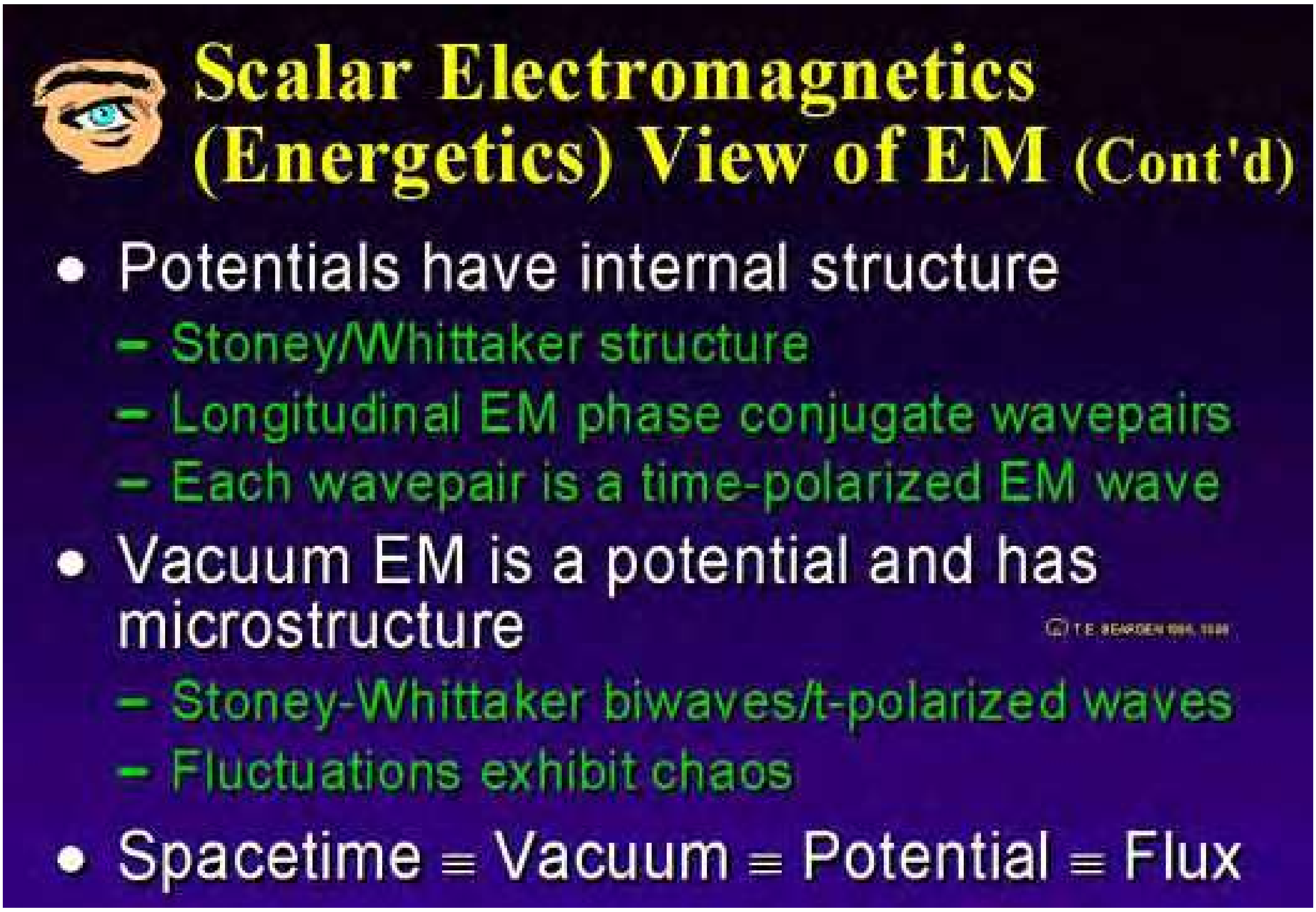

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Potentials have internal structure
  - Stoney/Whittaker structure
  - Longitudinal EM phase conjugate wavepairs
  - Each wavepair is a time-polarized EM wave
- Vacuum EM is a potential and has microstructure
  - Stoney-Whittaker biwaves/t-polarized waves
  - Fluctuations exhibit chaos
- Spacetime ≡ Vacuum ≡ Potential ≡ Flux

© T.E. BEARDEN [illegible]

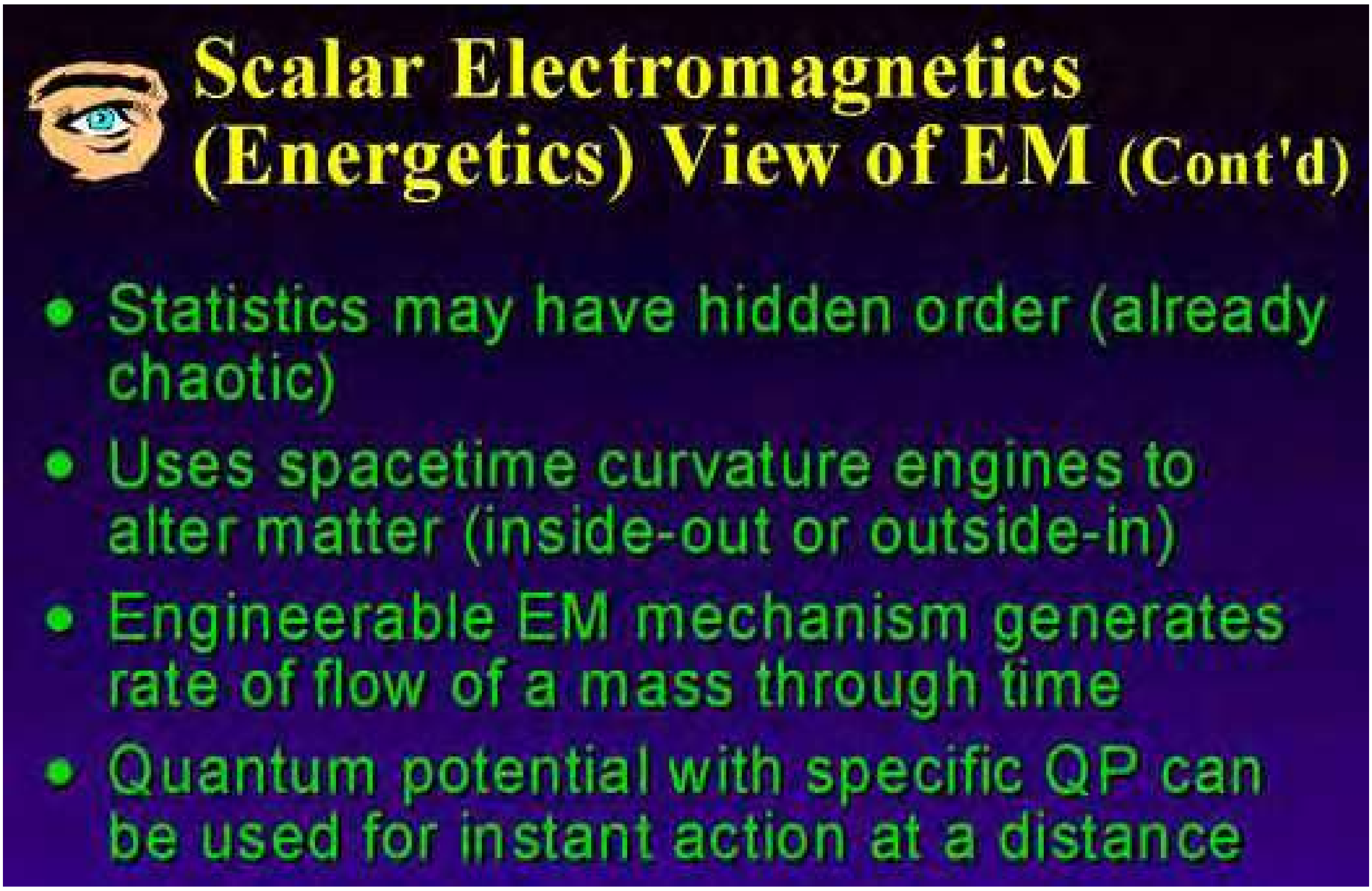

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Statistics may have hidden order (already chaotic)
- Uses spacetime curvature engines to alter matter (inside-out or outside-in)
- Engineerable EM mechanism generates rate of flow of a mass through time
- Quantum potential with specific QP can be used for instant action at a distance

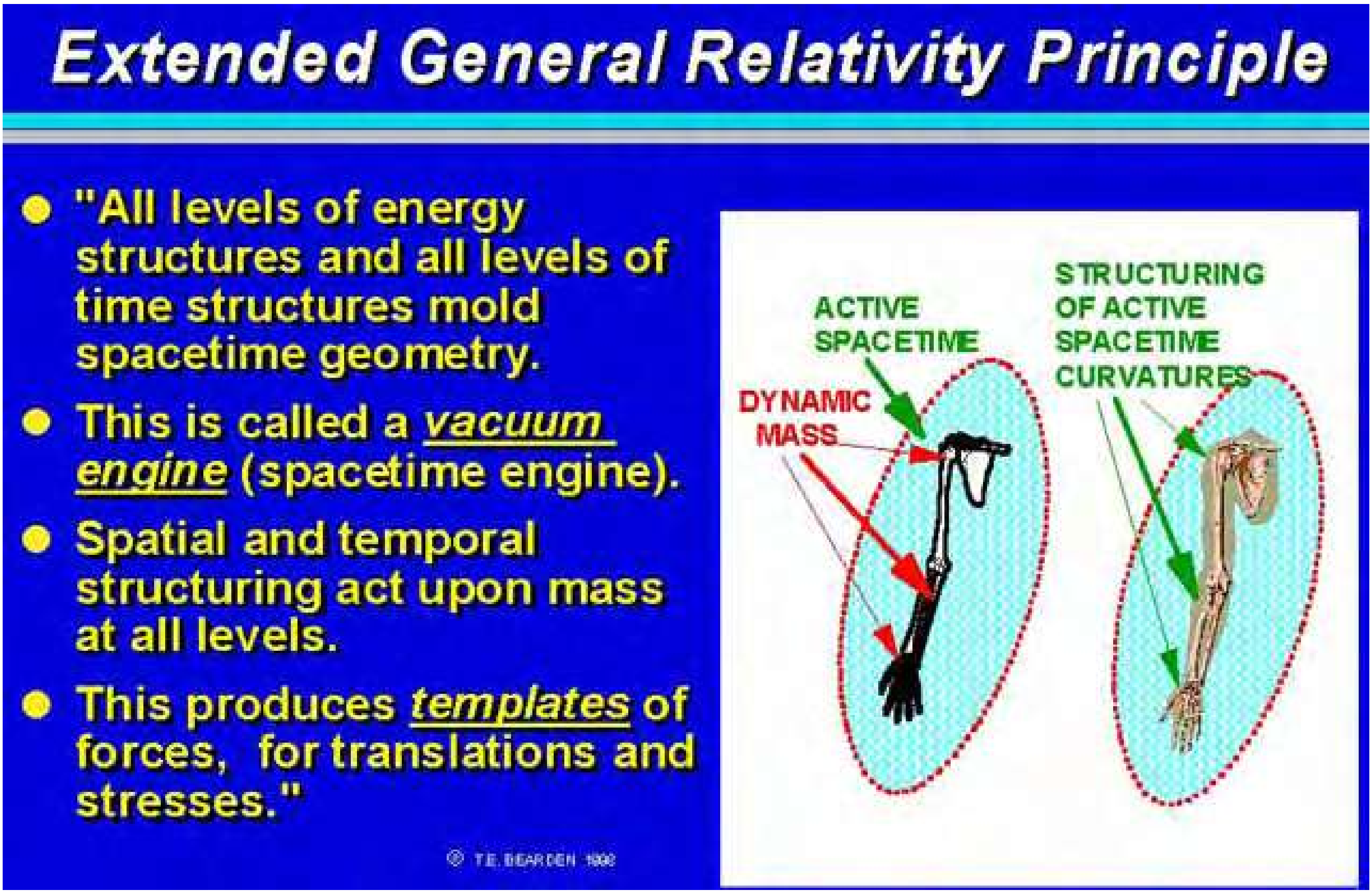

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a *vacuum engine* (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces *templates* of forces, for translations and stresses."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E. BEARDEN 1990

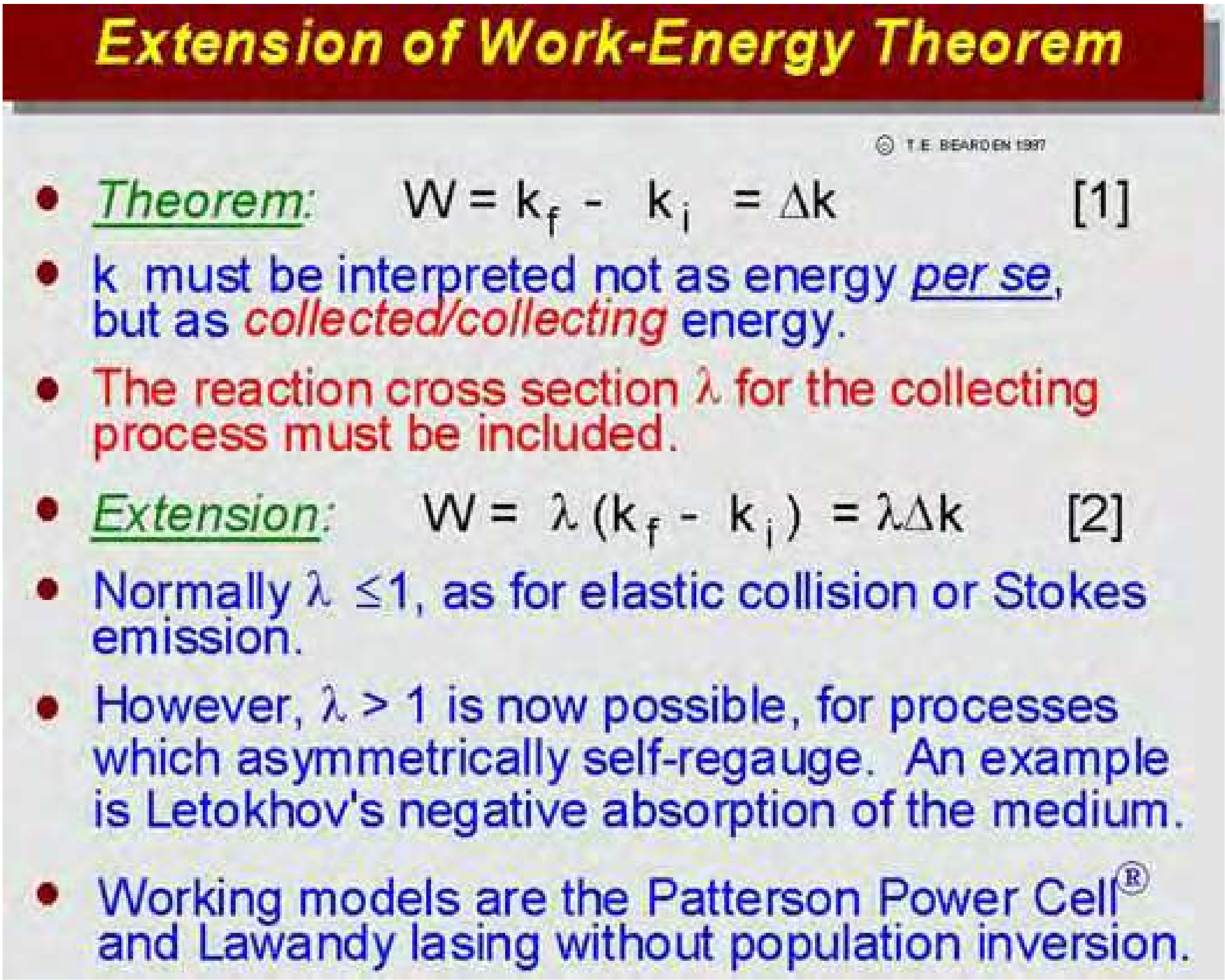

Extension of Work-Energy Theorem

© T.E. BEARDEN 1997

- *Theorem:* $W = k_f - k_i = \Delta k$ [1]
- k must be interpreted not as energy *per se*, but as *collected/collecting* energy.
- The reaction cross section λ for the collecting process must be included.
- *Extension:* $W = \lambda (k_f - k_i) = \lambda \Delta k$ [2]
- Normally $\lambda \leq 1$, as for elastic collision or Stokes emission.
- However, $\lambda > 1$ is now possible, for processes which asymmetrically self-regauge. An example is Letokhov's negative absorption of the medium.
- Working models are the Patterson Power Cell® and Lawandy lasing without population inversion.

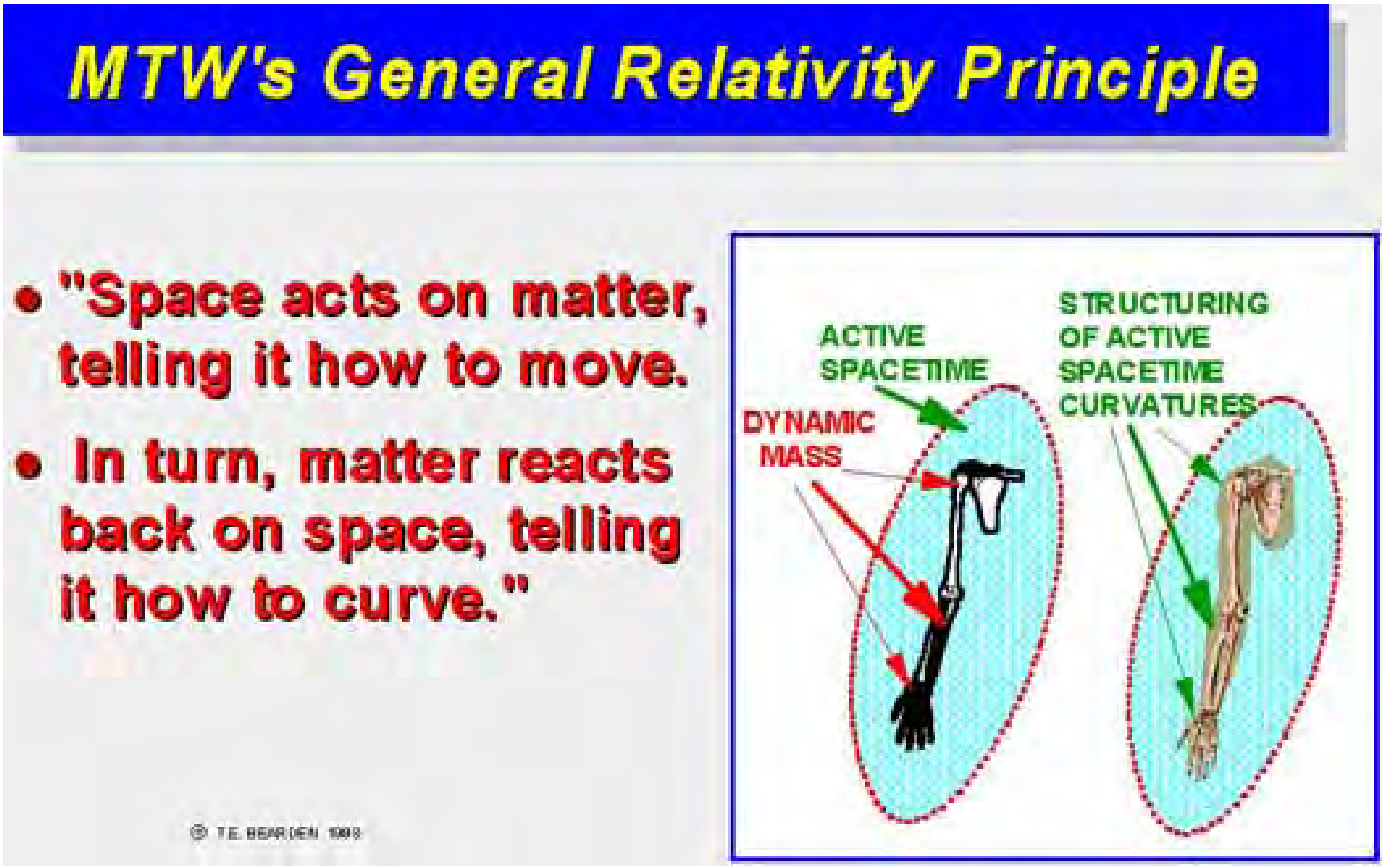

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE SPACETIME
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DYNAMIC MASS

© T.E. BEARDEN 1998

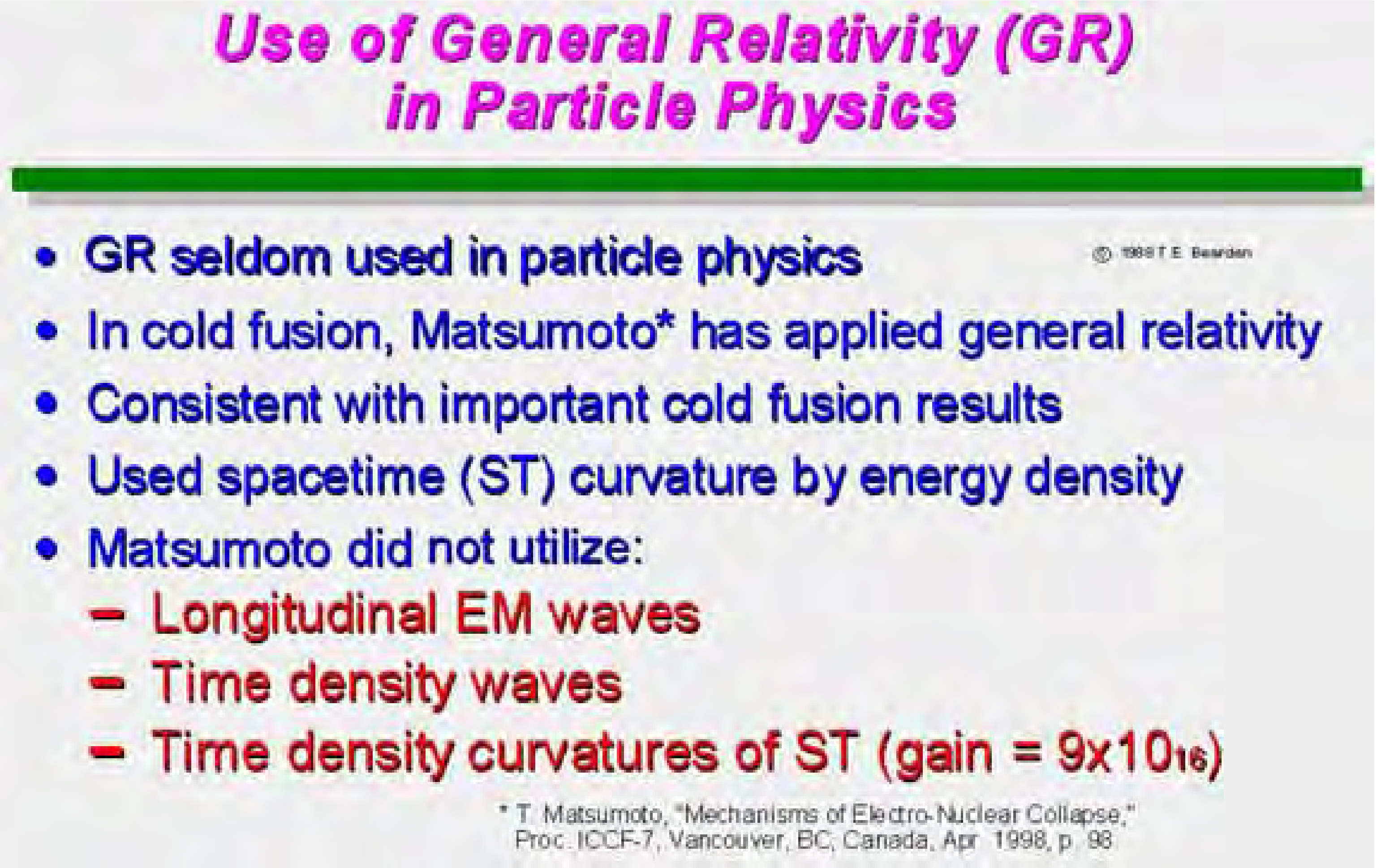

Use of General Relativity (GR) in Particle Physics

© 1998 T.E. Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density waves
  - Time density curvatures of ST (gain = 9x10$_{16}$)

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98

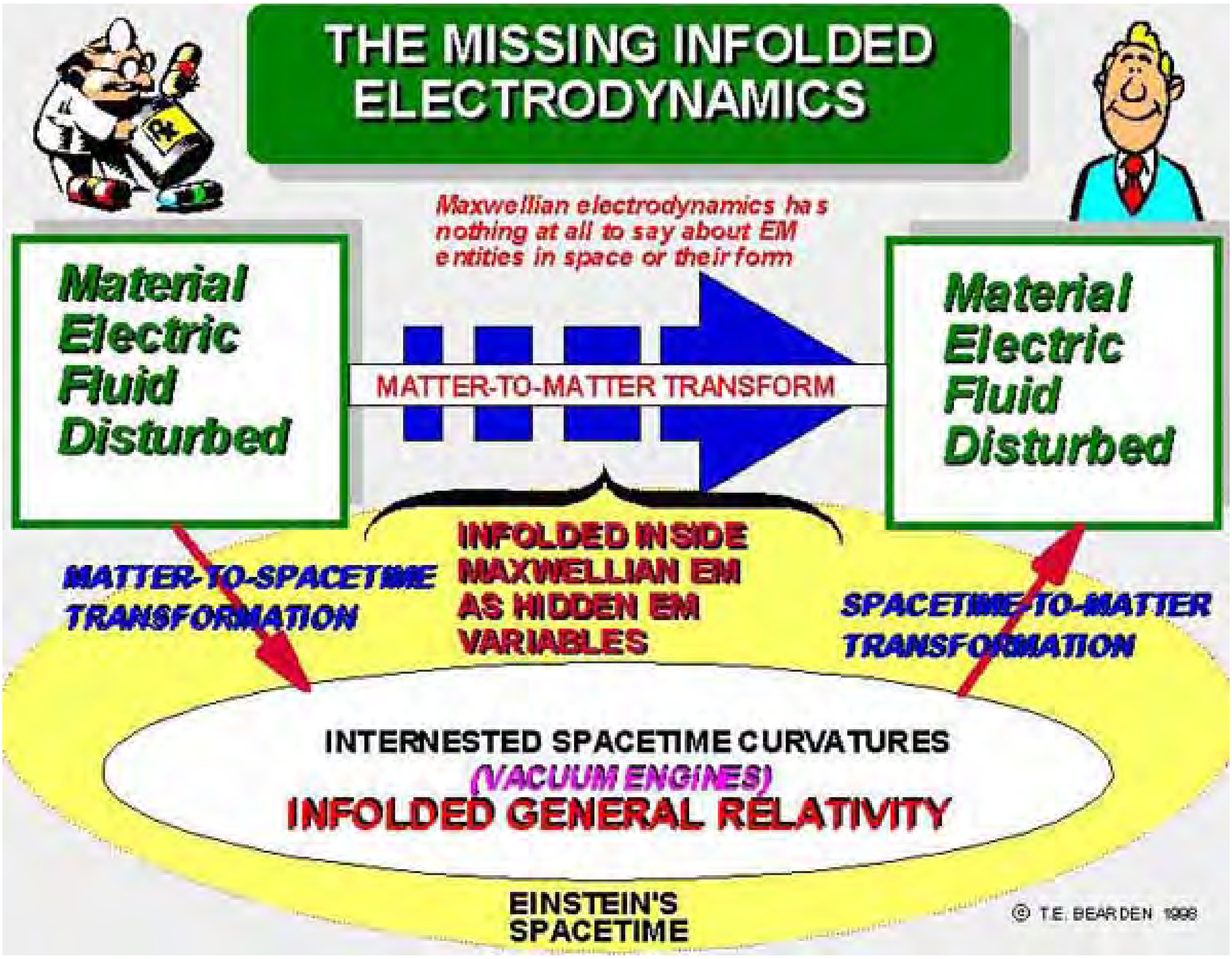

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

MATTER-TO-SPACETIME TRANSFORMATION

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

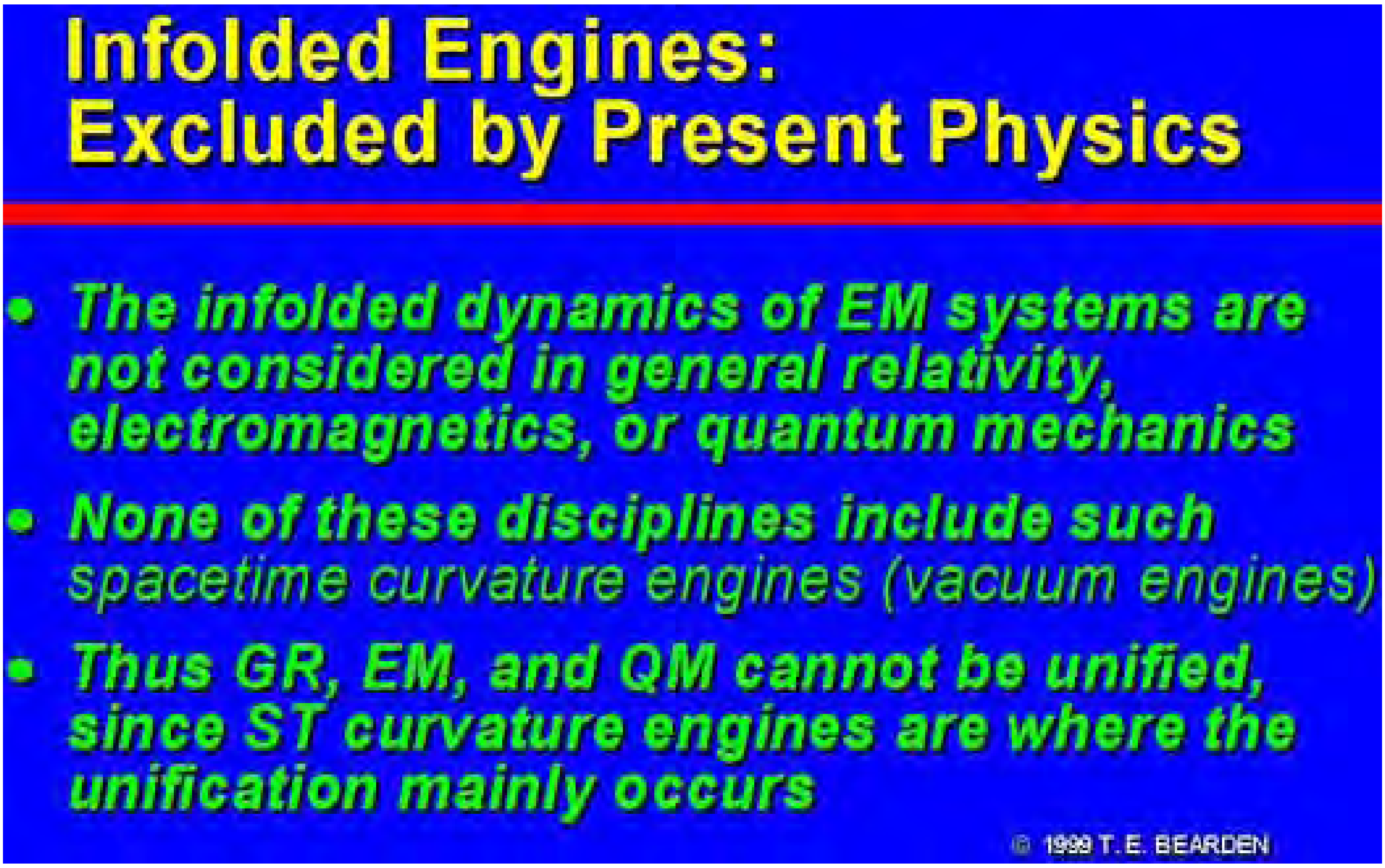

Infolded Engines: Excluded by Present Physics

- *The infolded dynamics of EM systems are not considered in general relativity, electromagnetics, or quantum mechanics*
- *None of these disciplines include such spacetime curvature engines (vacuum engines)*
- *Thus GR, EM, and QM cannot be unified, since ST curvature engines are where the unification mainly occurs*

© 1999 T. E. BEARDEN

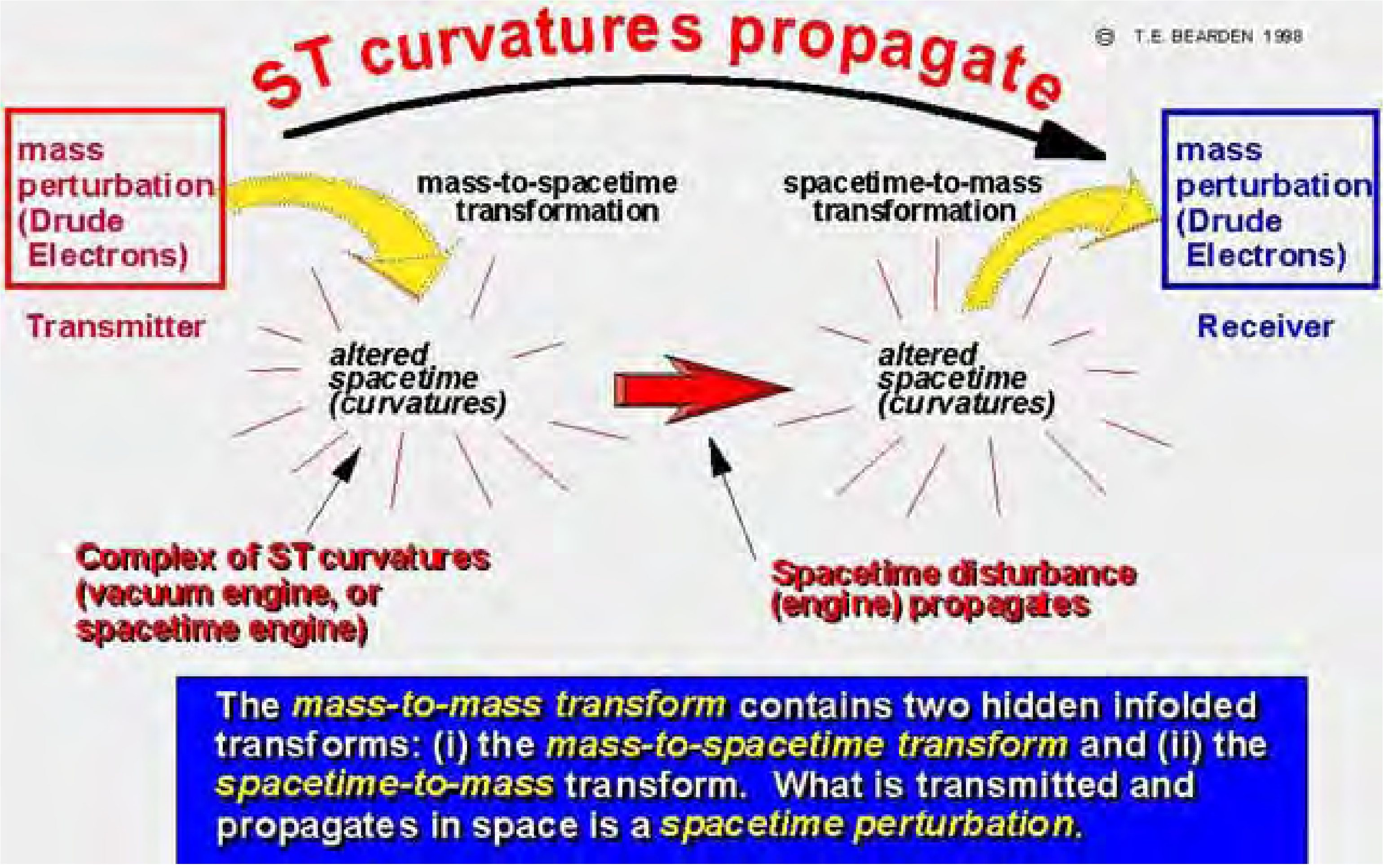

ST curvatures propagate

© T.E. BEARDEN 1998

mass
perturbation
(Drude
Electrons)

Transmitter

mass-to-spacetime
transformation

altered
spacetime
(curvatures)

spacetime-to-mass
transformation

altered
spacetime
(curvatures)

mass
perturbation
(Drude
Electrons)

Receiver

Complex of ST curvatures
(vacuum engine, or
spacetime engine)

Spacetime disturbance
(engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation.*

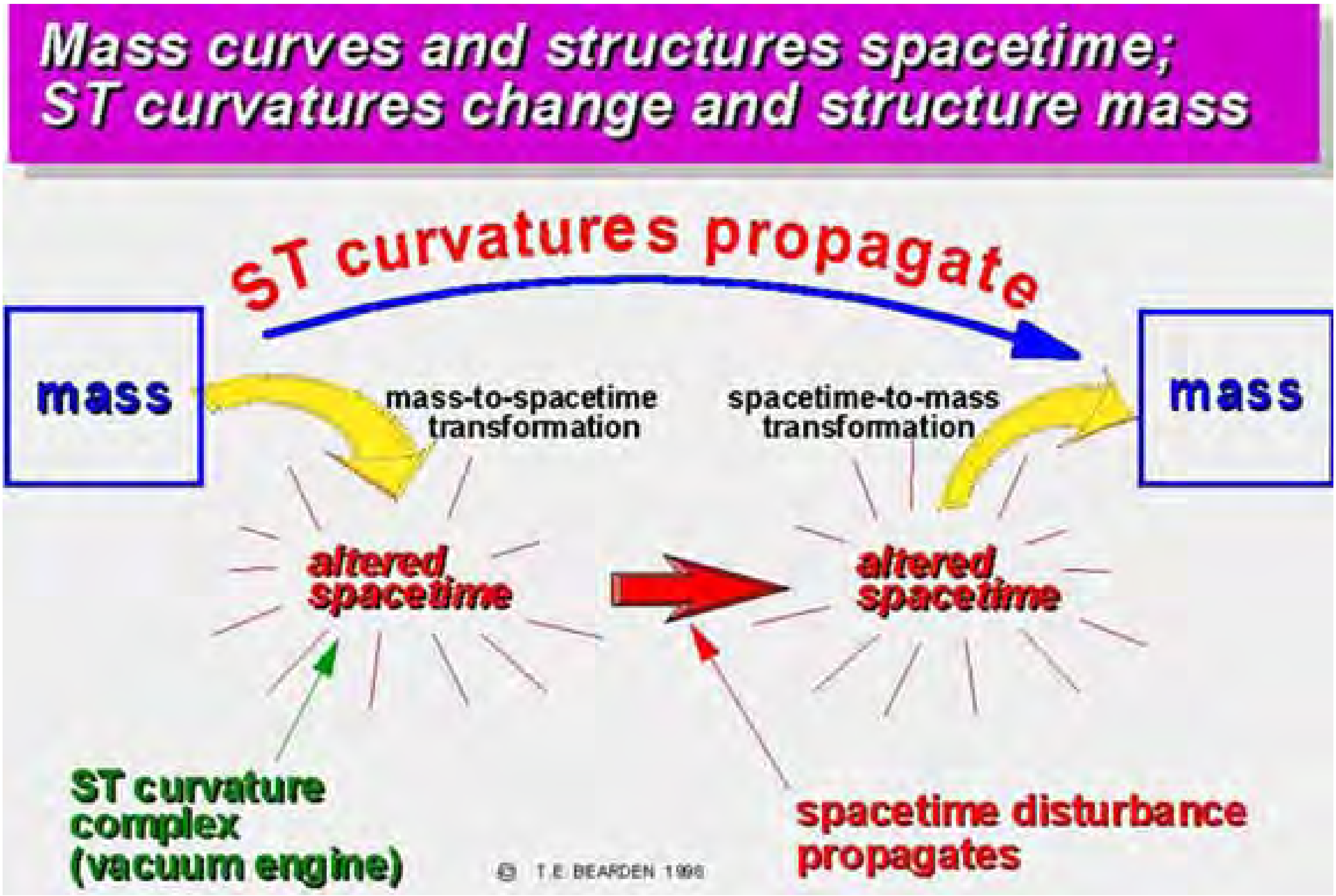

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate

mass

mass-to-spacetime transformation

altered spacetime

spacetime-to-mass transformation

altered spacetime

mass

ST curvature complex (vacuum engine)

spacetime disturbance propagates

© T.E. BEARDEN 1996

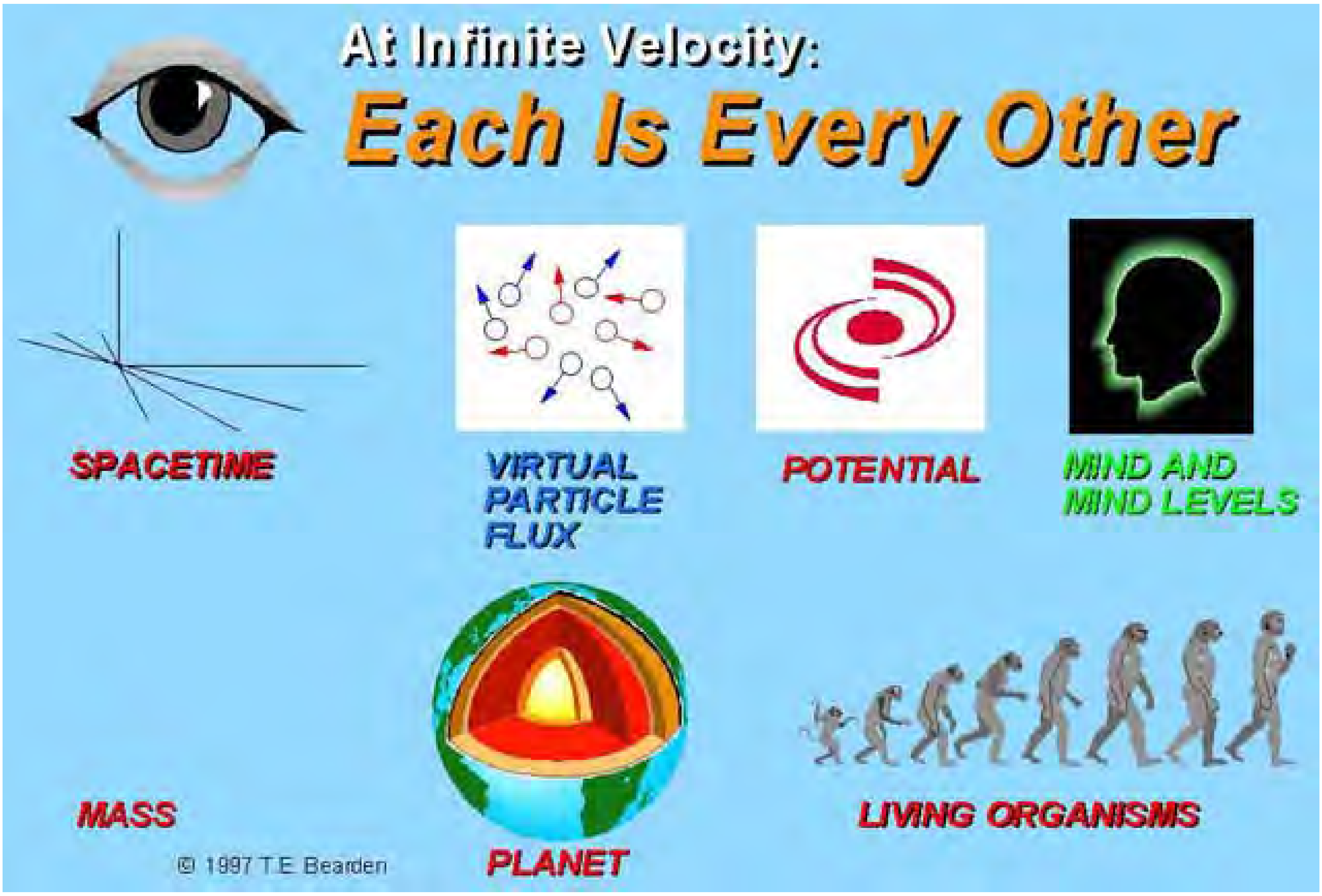

At Infinite Velocity:

Each Is Every Other

SPACETIME

VIRTUAL PARTICLE FLUX

POTENTIAL

MIND AND MIND LEVELS

MASS

PLANET

LIVING ORGANISMS

© 1997 T.E. Bearden

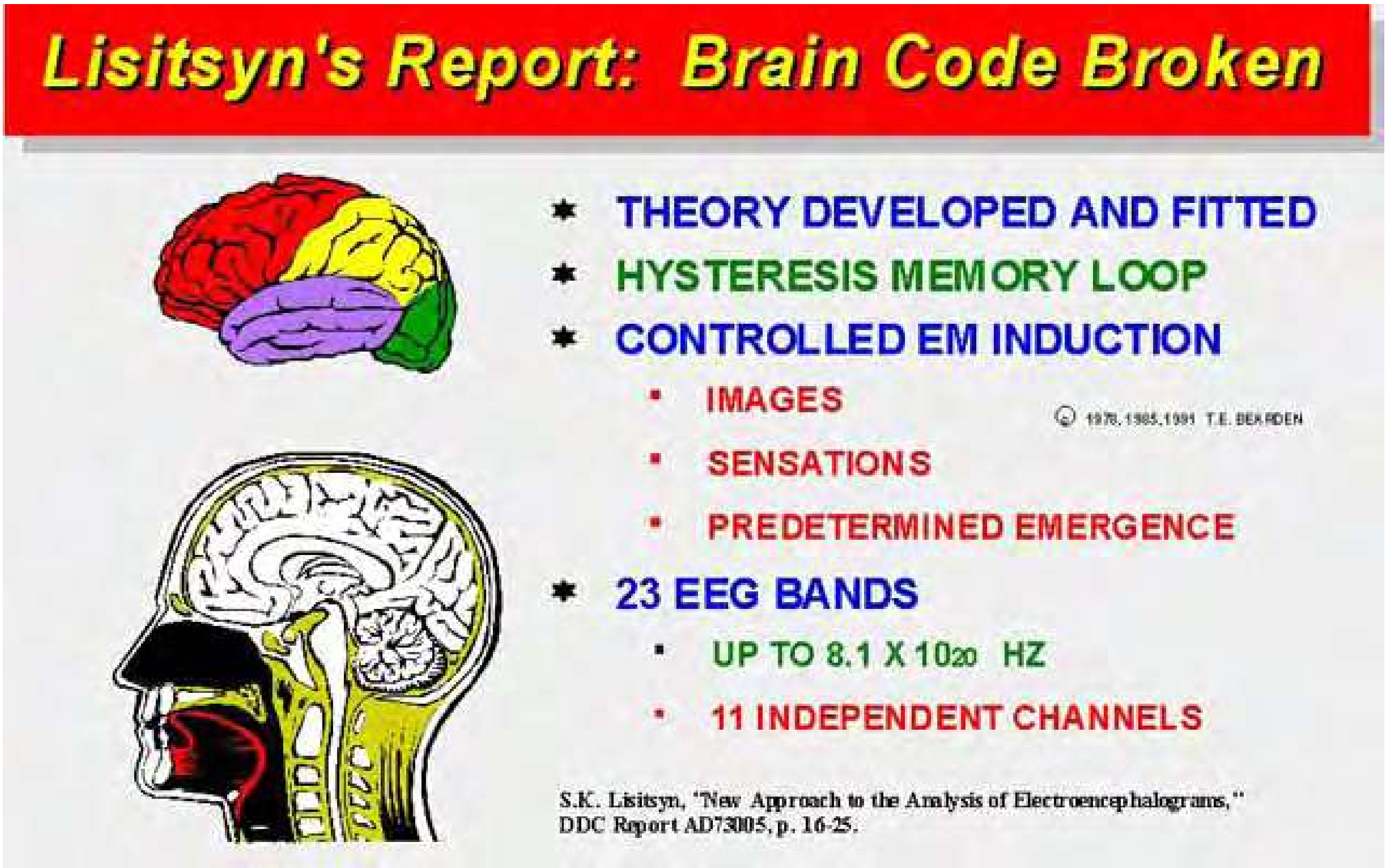

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO $8.1 \times 10_{20}$ HZ
  - 11 INDEPENDENT CHANNELS

© 1978,1985,1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

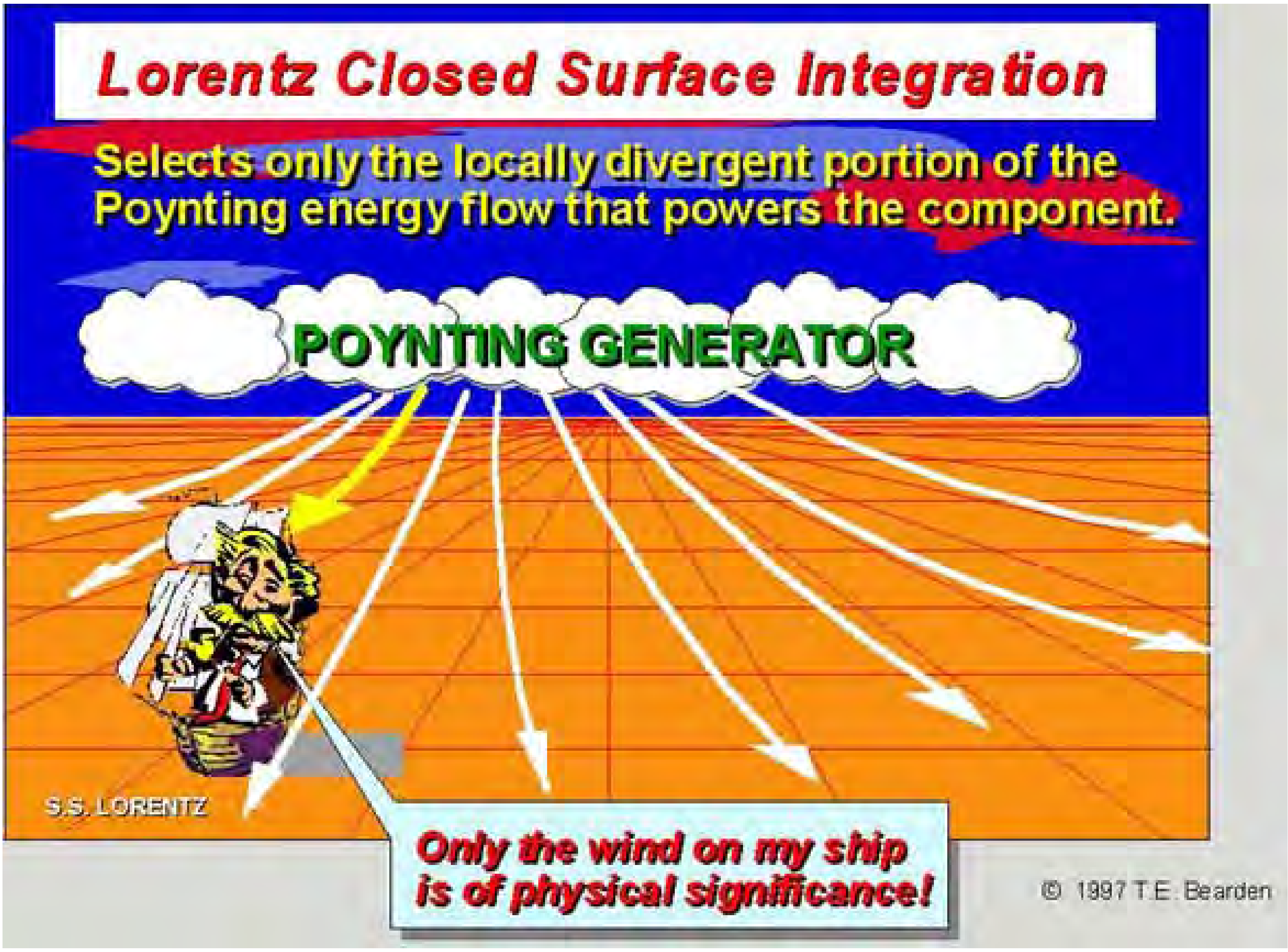

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

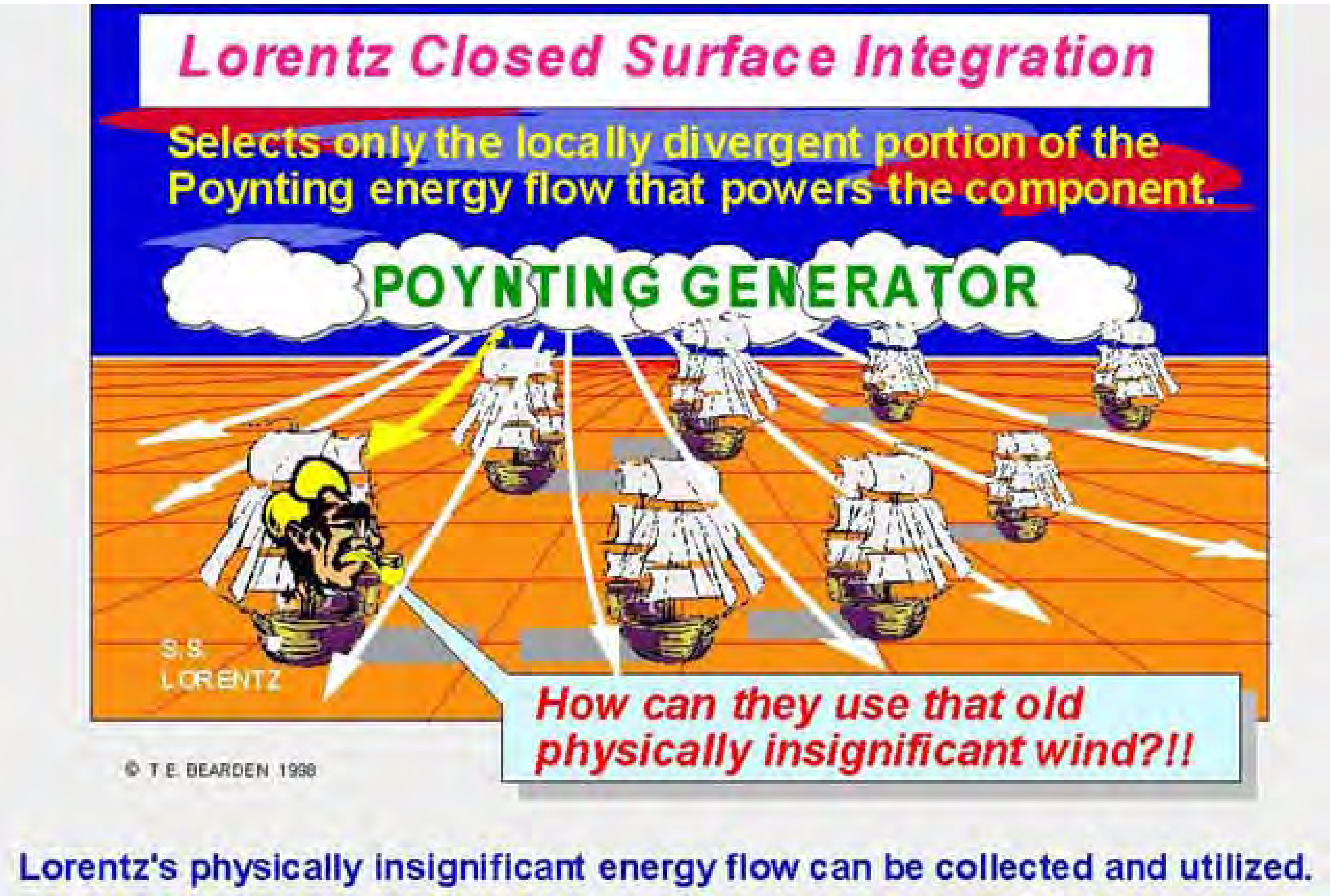

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

How can they use that old physically insignificant wind?!!

© T E BEARDEN 1998

Lorentz's physically insignificant energy flow can be collected and utilized.

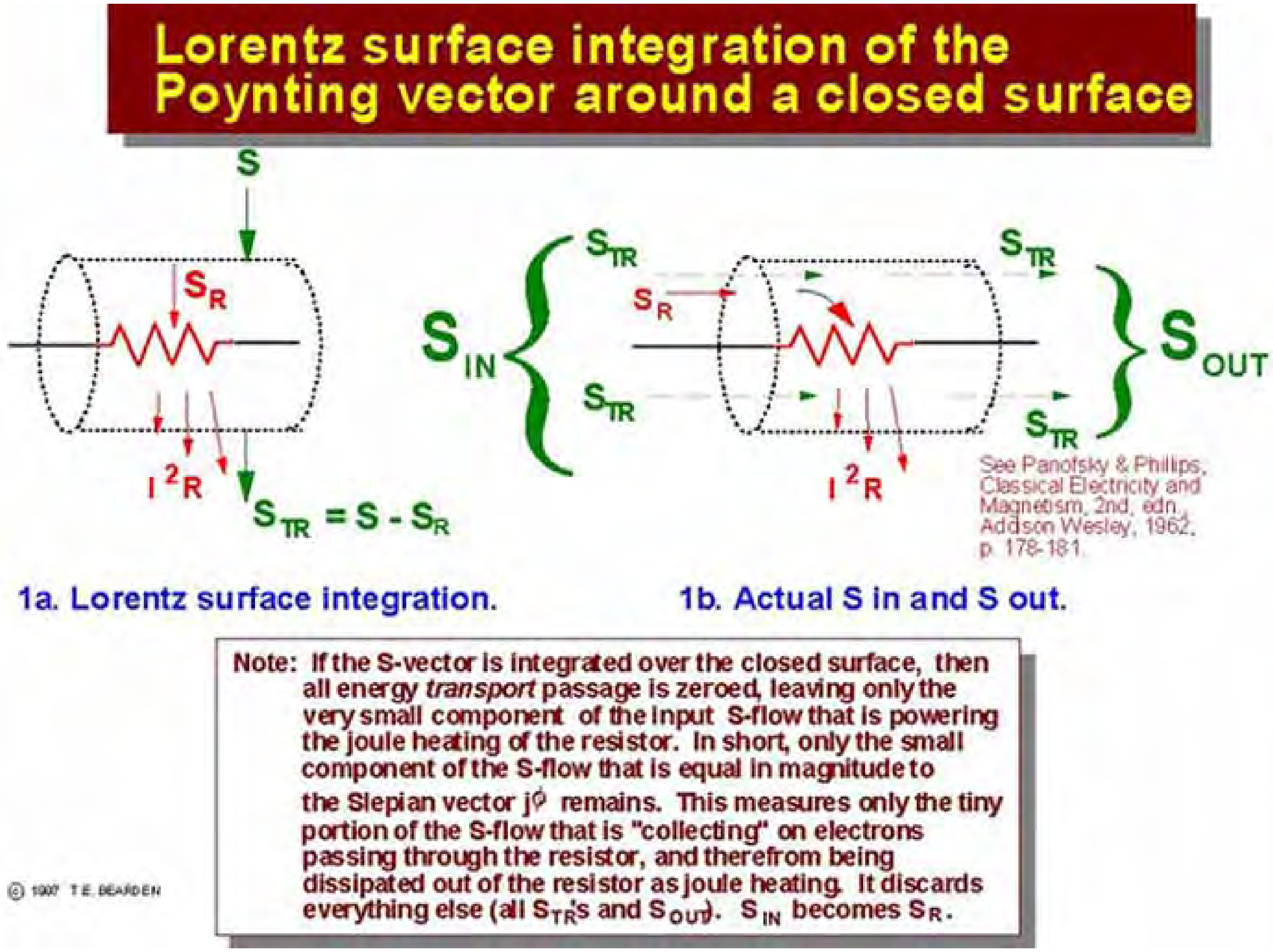

Lorentz surface integration of the Poynting vector around a closed surface

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd, edn., Addison Wesley, 1962, p. 178-181.

1a. Lorentz surface integration.

1b. Actual S in and S out.

Note: If the S-vector is integrated over the closed surface, then all energy *transport* passage is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Slepian vector $j\phi$ remains. This measures only the tiny portion of the S-flow that is "collecting" on electrons passing through the resistor, and therefrom being dissipated out of the resistor as joule heating. It discards everything else (all S_{TR}'s and S_{OUT}). S_{IN} becomes S_R.

© 1997 T.E. BEARDEN

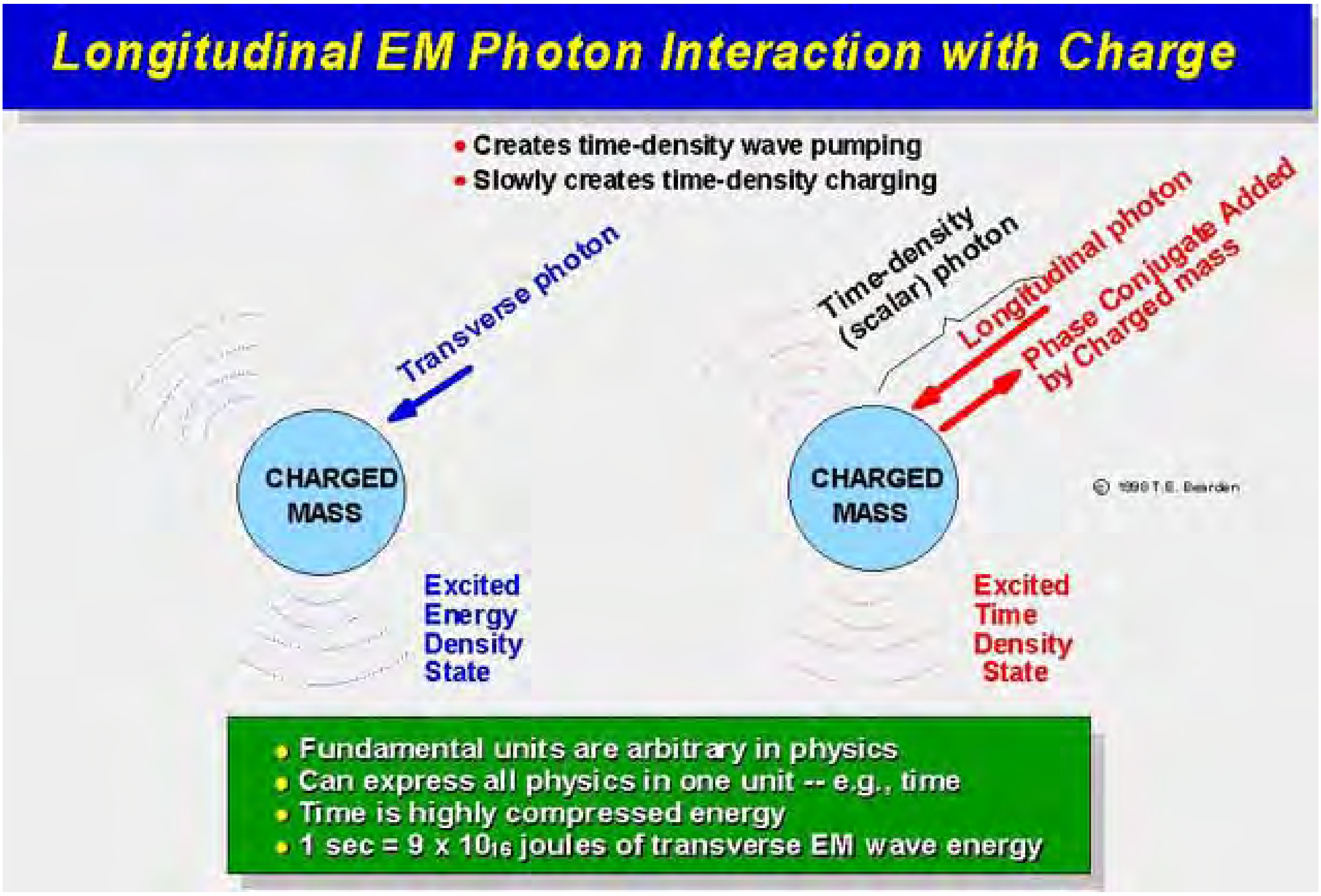

Longitudinal EM Photon Interaction with Charge

- Creates time-density wave pumping
- Slowly creates time-density charging

Transverse photon

CHARGED MASS

Excited Energy Density State

Time-density (scalar) photon

Longitudinal photon

Phase Conjugate Added by Charged mass

CHARGED MASS

Excited Time Density State

© 1990 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9 x 10_{16} joules of transverse EM wave energy

Figure 1-8. Infolded biwave composition of a scalar potential.

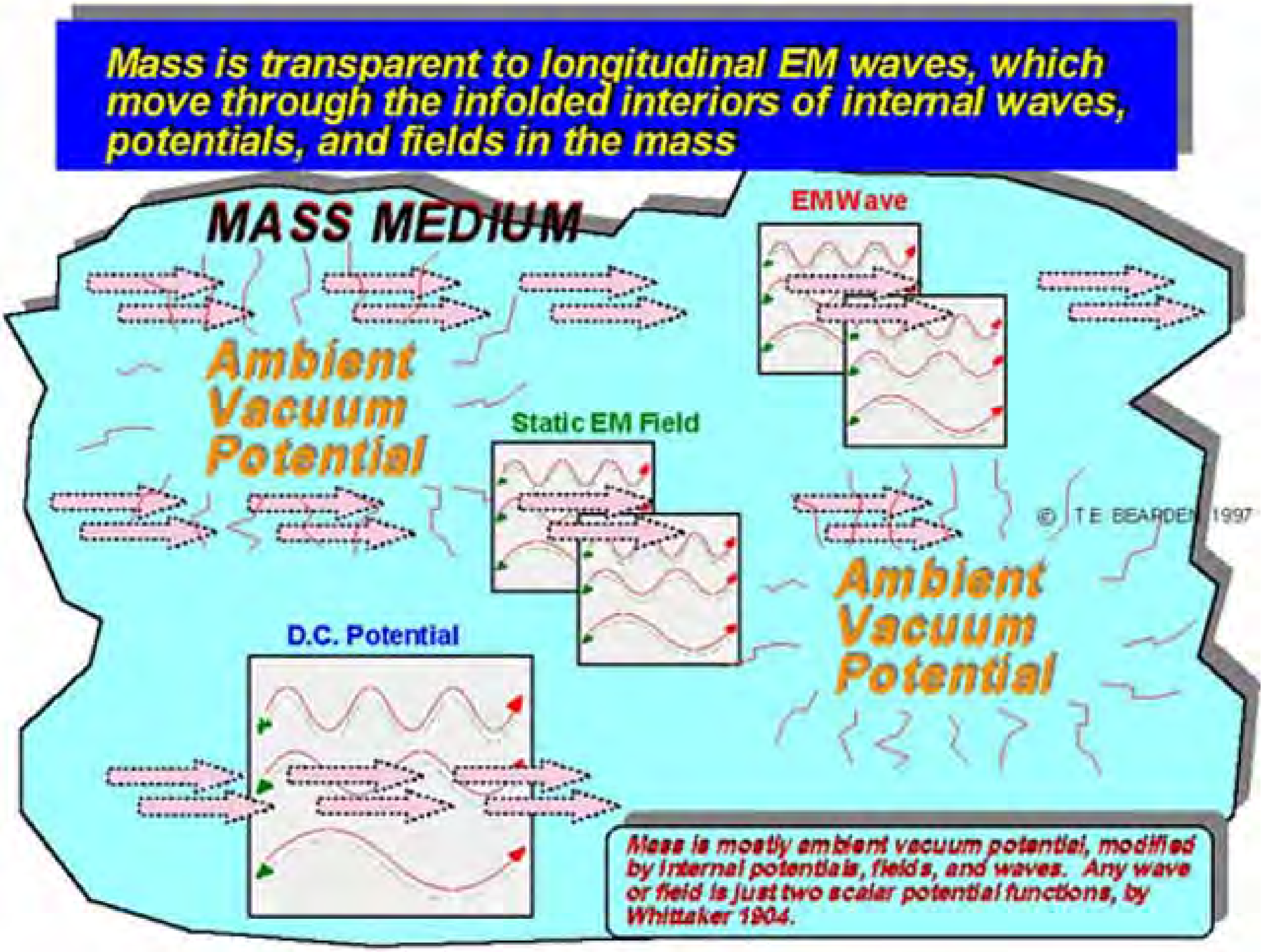

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
EMWave
Ambient
Vacuum
Potential
Static EM Field
© T E BEARDEN 1997
Ambient
Vacuum
Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

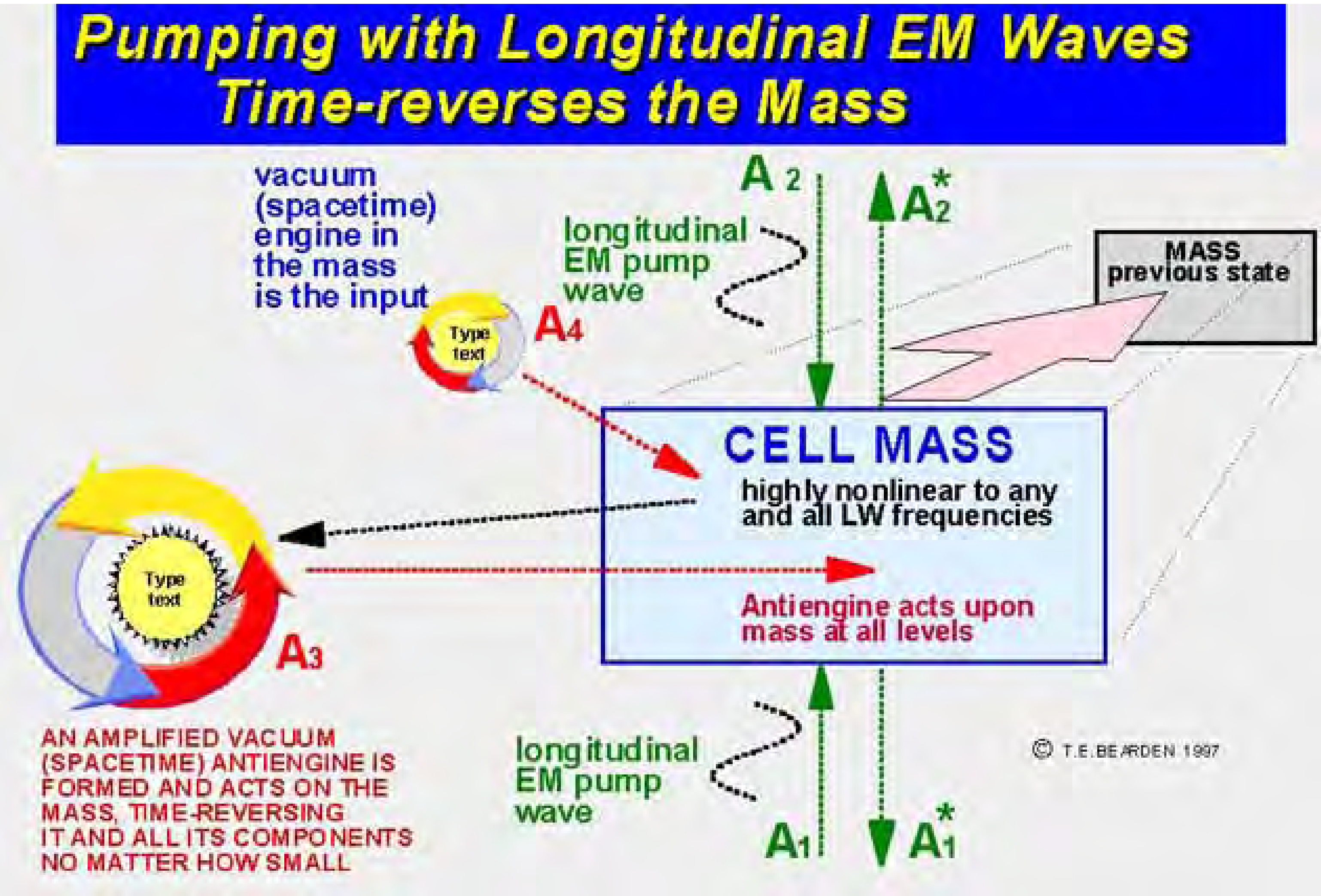

Pumping with Longitudinal EM Waves Time-reverses the Mass

vacuum
(spacetime)
engine in
the mass
is the input

longitudinal
EM pump
wave

MASS
previous state

CELL MASS

highly nonlinear to any
and all LW frequencies

Antiengine acts upon
mass at all levels

AN AMPLIFIED VACUUM
(SPACETIME) ANTIENGINE IS
FORMED AND ACTS ON THE
MASS, TIME-REVERSING
IT AND ALL ITS COMPONENTS
NO MATTER HOW SMALL

longitudinal
EM pump
wave

© T.E.BEARDEN 1997

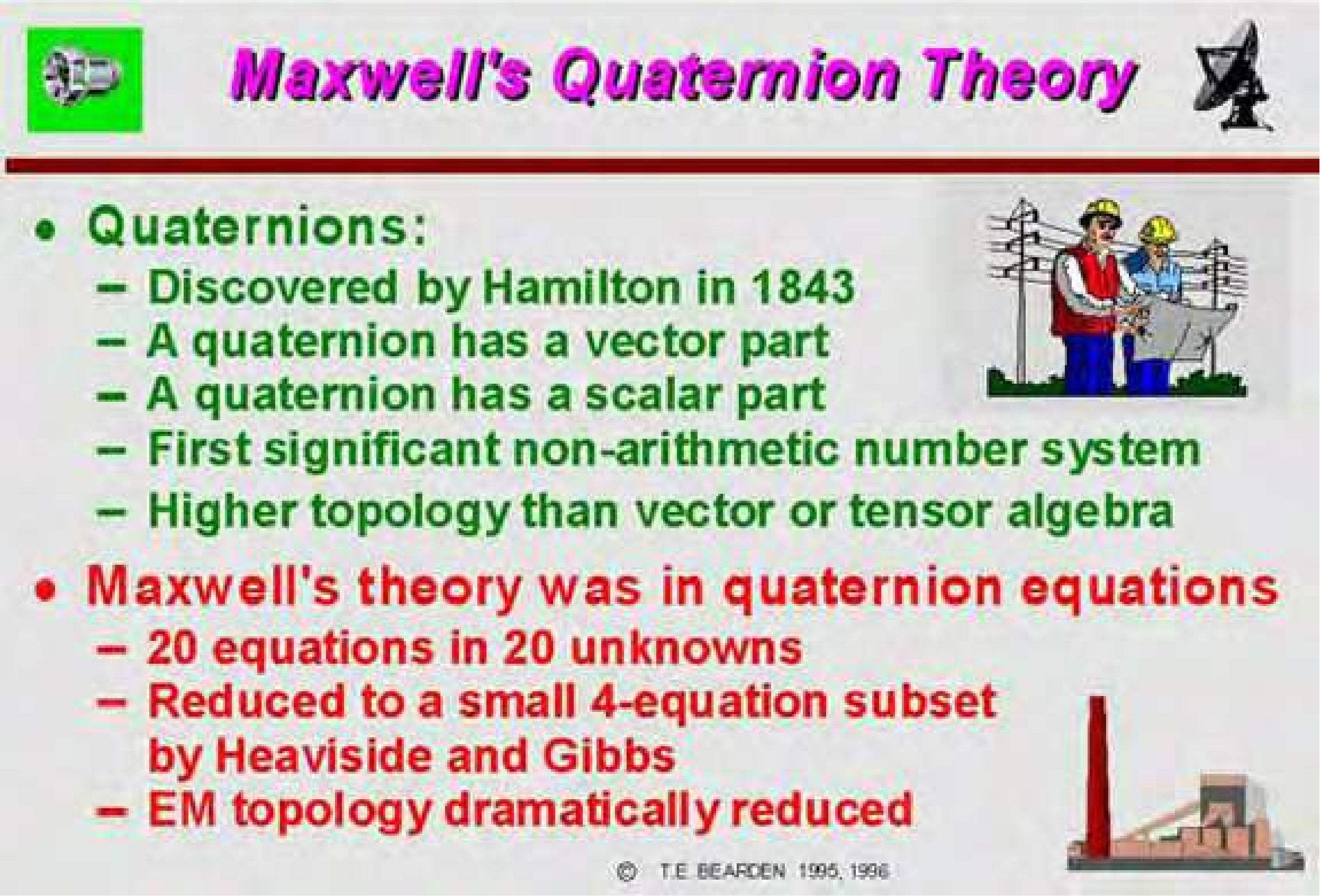

Maxwell's Quaternion Theory

- Quaternions:
  - Discovered by Hamilton in 1843
  - A quaternion has a vector part
  - A quaternion has a scalar part
  - First significant non-arithmetic number system
  - Higher topology than vector or tensor algebra
- Maxwell's theory was in quaternion equations
  - 20 equations in 20 unknowns
  - Reduced to a small 4-equation subset by Heaviside and Gibbs
  - EM topology dramatically reduced

© T.E. BEARDEN 1995, 1996

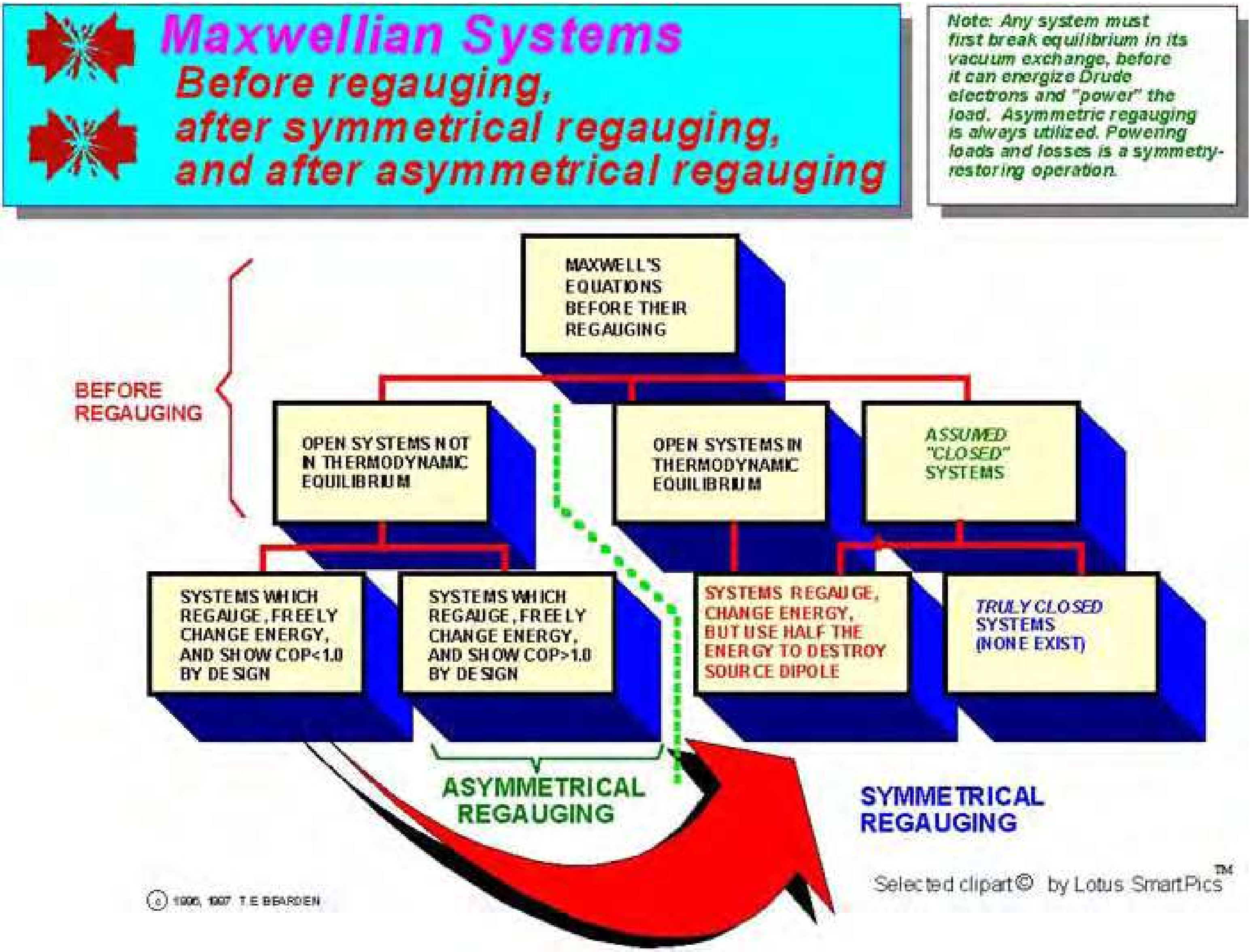

Maxwellian Systems
Before regauging, after symmetrical regauging, and after asymmetrical regauging

Note: Any system must first break equilibrium in its vacuum exchange, before it can energize Drude electrons and "power" the load. Asymmetric regauging is always utilized. Powering loads and losses is a symmetry-restoring operation.

MAXWELL'S EQUATIONS BEFORE THEIR REGAUGING

BEFORE REGAUGING

OPEN SYSTEMS NOT IN THERMODYNAMIC EQUILIBRIUM

OPEN SYSTEMS IN THERMODYNAMIC EQUILIBRIUM

ASSUMED "CLOSED" SYSTEMS

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP<1.0 BY DESIGN

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP>1.0 BY DESIGN

SYSTEMS REGAUGE, CHANGE ENERGY, BUT USE HALF THE ENERGY TO DESTROY SOURCE DIPOLE

TRULY CLOSED SYSTEMS (NONE EXIST)

ASYMMETRICAL REGAUGING

SYMMETRICAL REGAUGING

© 1996, 1997 T.E. BEARDEN

Selected clipart© by Lotus SmartPics™

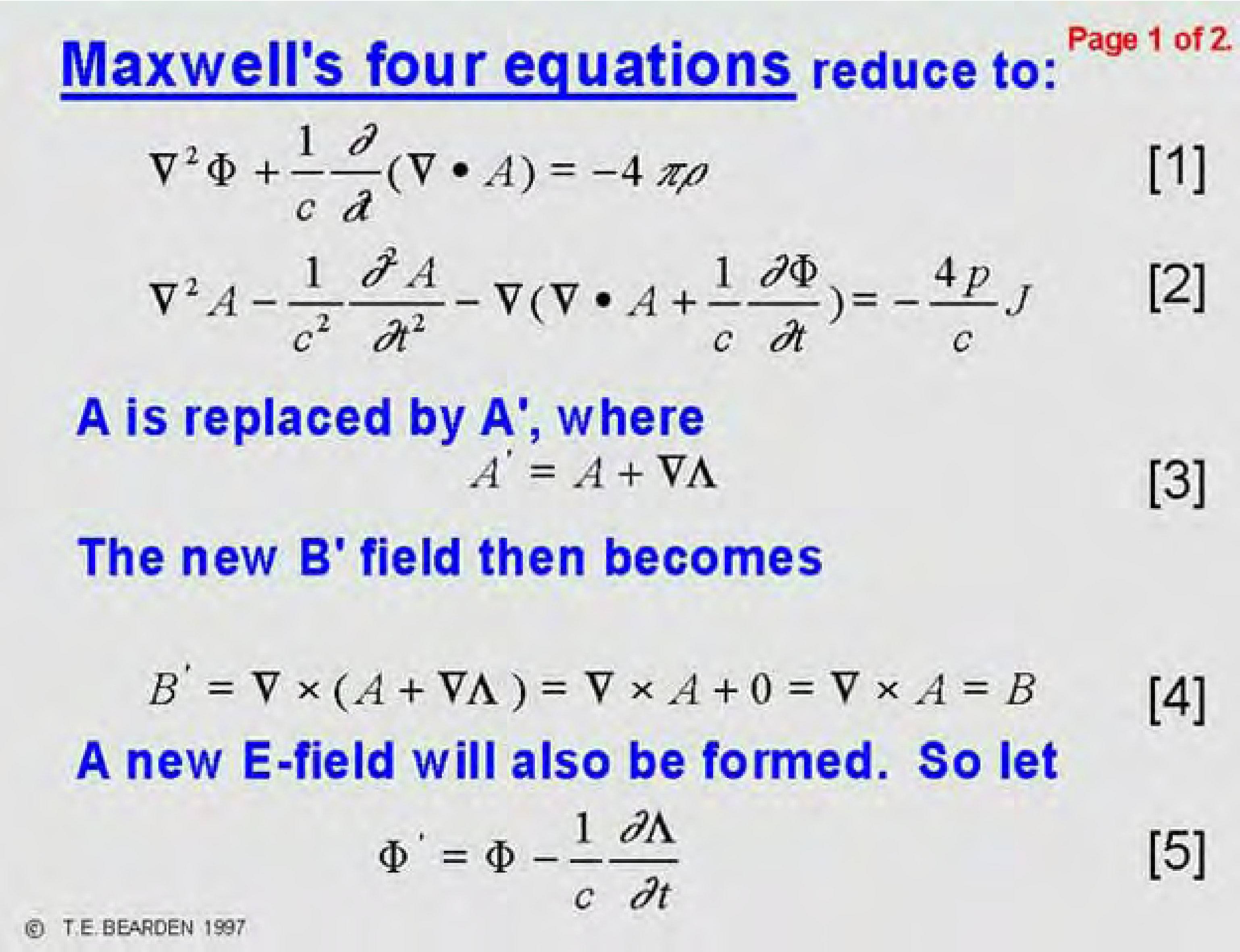

Page 1 of 2.

Maxwell's four equations reduce to:

$$\nabla^2\Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet A) = -4\pi\rho \qquad [1]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} - \nabla\left(\nabla \bullet A + \frac{1}{c}\frac{\partial\Phi}{\partial t}\right) = -\frac{4p}{c}J \qquad [2]$$

A is replaced by A', where

$$A' = A + \nabla\Lambda \qquad [3]$$

The new B' field then becomes

$$B' = \nabla \times (A + \nabla\Lambda) = \nabla \times A + 0 = \nabla \times A = B \qquad [4]$$

A new E-field will also be formed. So let

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t} \qquad [5]$$

© T.E. BEARDEN 1997

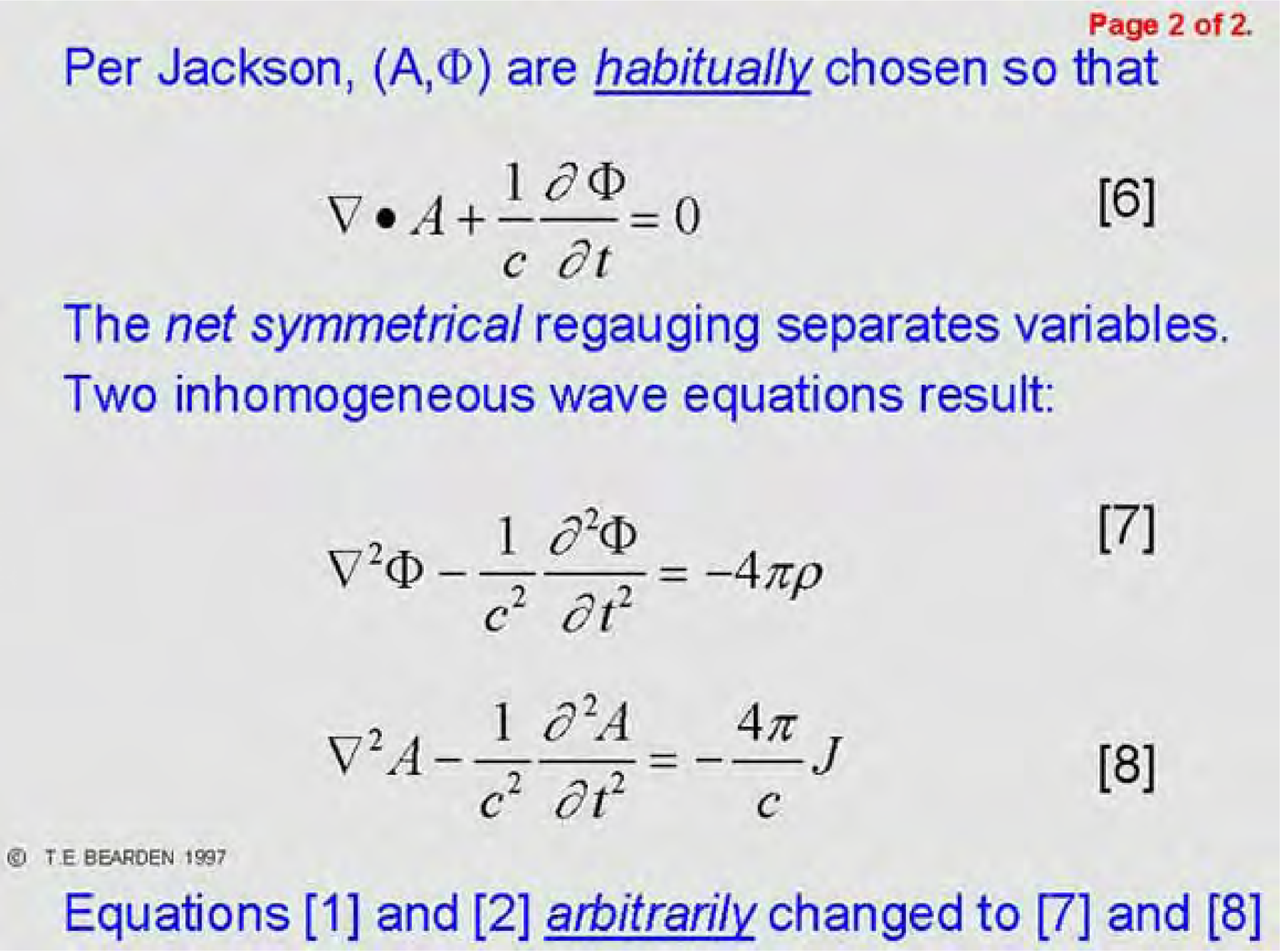

Per Jackson, (A,Φ) are *habitually* chosen so that

$$\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [6]$$

The *net symmetrical* regauging separates variables.
Two inhomogeneous wave equations result:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [7]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{4\pi}{c} J \qquad [8]$$

© T.E BEARDEN 1997

Equations [1] and [2] *arbitrarily* changed to [7] and [8]

The Tom Bearden Website

Mind Control and EM Wave Polarization Transductions - Part I, II & III

Reprinted from "Explore" magazine, 1999

Part I

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9
- Page 10

Part II

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8

Part III

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9

- The U.S. approach to mind control

- "Voices in my head"
  Flanagan's neurophone - used to give instructions to diplomats during negotiations, etc.

The Tom Bearden Website

ORDER BOOKS & DVDs

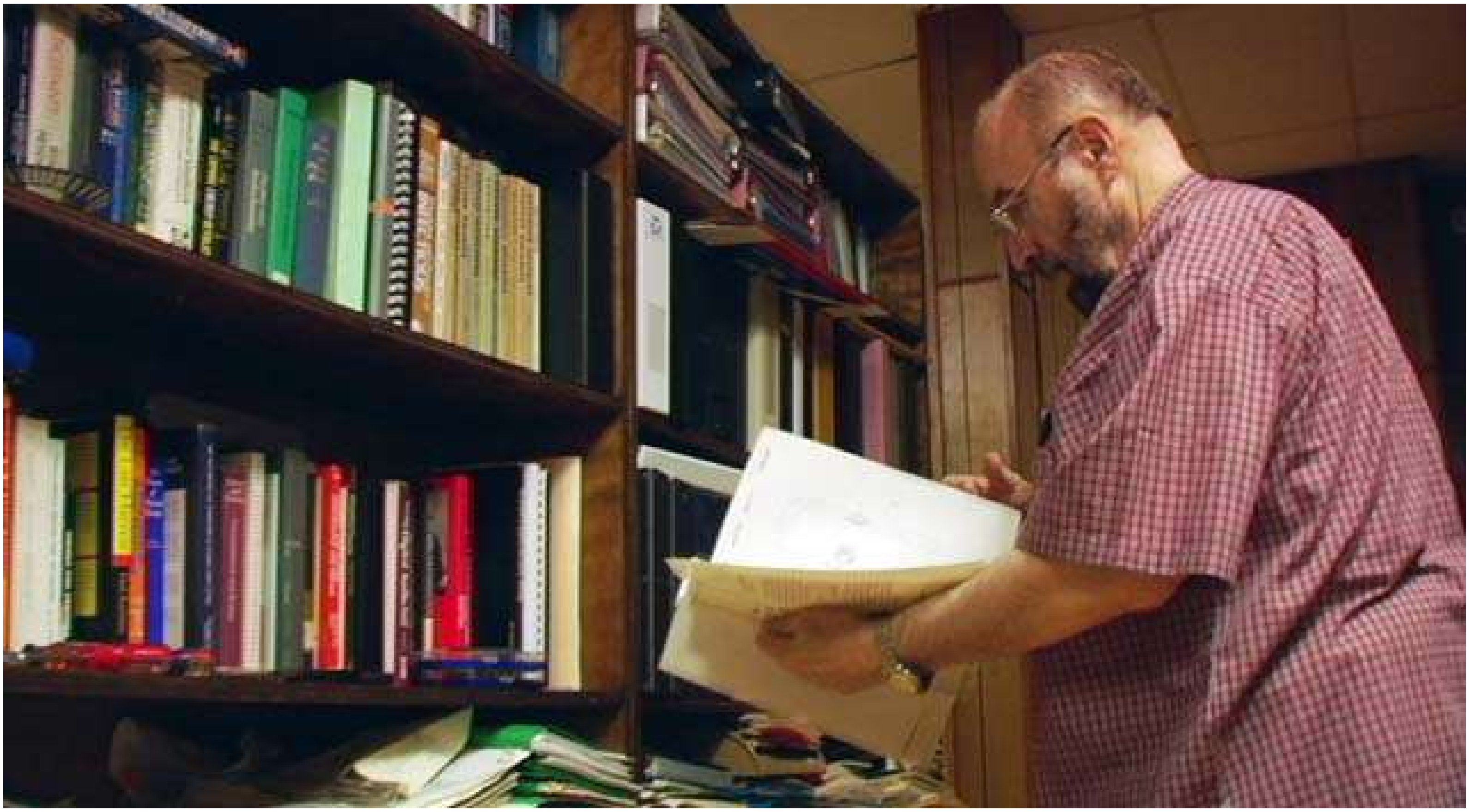

Articles

- **Job Opportunities in Energy**
  25 August, 2010

- **THE WORLD ENERGY CRISIS IN PERSPECTIVE:**
  The Breakthrough Technology that Solves it Quickly
  22 March, 2010

- **Direct Engineering of Physical Reality Itself**
  3 June, 2009

- **Solving The Problem of Freely Using EM Energy from the Vacuum**
  18 September, 2008

- **How an EM circuit powers its load - and solving the world energy crisis**
  16 September, 2008

- **Rapid and Decisive Solution of the World Energy Crisis and Global Warming**
  Updated
  9 September, 2008

- **On Negative Energy EMP Weapons**
  26 July, 2008

- **In Memoriam - Howard Johnson**
  31 March, 2008

- **Suppression of the Self-Powering Battery?**
  1st December, 2007

- **Never Mind - Junk Science**

and All That
11 September, 2007

- **The Deliberate Discard of Asymmetric Maxwellian Systems, Thus Preventing COP>1.0**
  and Self-Powering Energy-from-the-Vacuum Systems
  21 June, 2007

- **What We Have to Get Across to Our Electrical Engineering Profession:**
  To the EE Departments, Professors, Students, Curricula, and Textbooks
  May 24, 2007

- **Phases of Research and Development and Some Problems**
  18 February, 2007

- **Scalar Electromagnetic Weapons and their Terrorist Use: Immediate Strategic Aspects of the Asymmetric War on the U.S.**
  11 September, 2004

- **Reducing our Dependence on Fossil Fuels**
  28 April, 2003

- **Flash News: Second Law of Thermodynamics Curtailed**
  4th August, 2002

- **Maxwell's demons are alive and thriving**
  4 July, 2002

- **Clean Electrical Energy from the Active Vacuum**
  January, 2002

Mind Control and EM Wave Polarization Transductions

Part I
Vol. 9, No. 2 1999

Part II
Vol 9, No. 3 1999

Part III
Vol 9, No. 4,5 1999

Vacuum Engines and Priore's Methodology - The True Science of Energy-Medicine

Part I
Vol. 6 No. 1 1995

Part II
Vol. 6 No. 2 1995

Paranormal Pictures
Psychic Observer 39(3) ~ 224, 255-258

Many of Tom Bearden's older articles on a wide range of topics can be found in the back issues of Explore! magazine.

See http://www.explorepub.com

Mind Control and EM Wave Polarization Transductions, Part 1

©Copyright 1999 by T.E. Bearden, USA

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental pro-cedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

Abstract

For some time we have been repeatedly queried about the technical mechanisms and unusual electrodynamics of advanced mind control research, both in the West and abroad. Calling full attention to the special note above, in this paper we present a high-level over-view of the novel electromagnetic nature of mind operations, mind and body coupling, and *intent*•the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transduction mechanisms whereby differing EM wave polarizations can be transformed one-into-the-other are presented. We give two specific examples of lethal foreign military tests in 1997 of advanced mind control weapons against two military pilots over the central U.S. Some dangers and potential benefits of the emerging mind control technology are pointed out

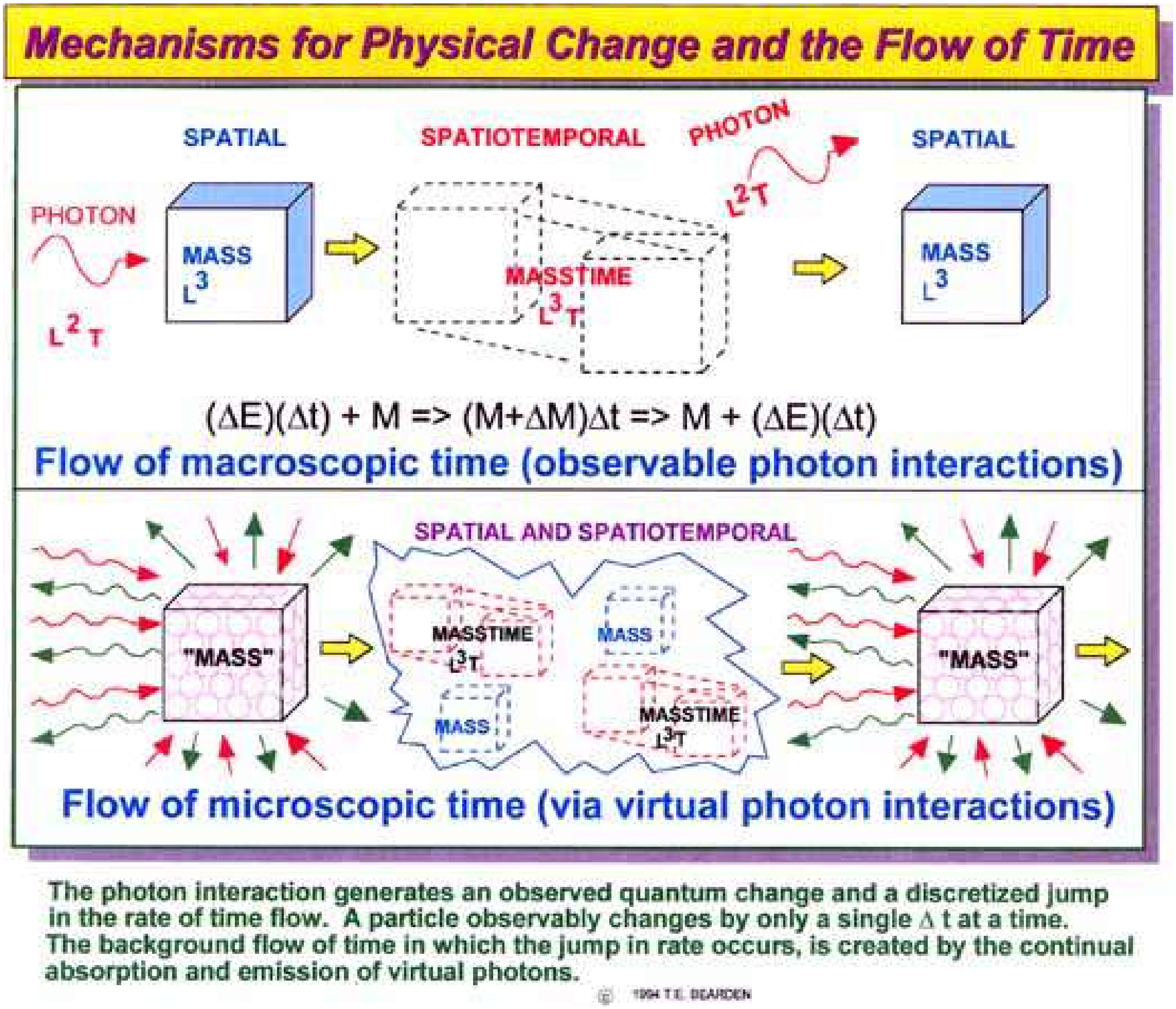
Mechanisms for Physical Change and the Flow of Time
SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS
L^3
L^2T
MASSTIME
L^3T
L^2T
MASS
L^3
$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$
Flow of macroscopic time (observable photon interactions)
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"
Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δ t at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Figure 1. Mechanism that generates a mass's flow through time.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Introduction and Background

In quantum field theory, there are four polarizations of photons.[1] Using 4-space and the z-direction as the direction of propagation, we have x- and y- polarizations where the 3-spatial energy of the photon is oscillating laterally, in the x- or y- direction. These are *transverse polarized* photons, as is any combination of the two. The third polarization is along the z-direction, which is a longitudinal polarization. In other words, the 3-spatial energy of the longitudinal photon cannot oscillate in the x- or y- direction, and so it is oscillating to-and-fro along the line of motion, z-. The fourth polarization occurs when the energy is frozen in all three spatial dimensions x-, y-, and z- and it cannot oscillate in any of those directions. In that case, the photon oscillates its energy in the t- direction, providing the t-polarized photon. We will later discuss how time is actually highly condensed energy.

In Minkowski 4-space, body operations are space-like, and are so treated in conventional materialistic physics. Mind and mind-operations are time-like, not space-like, even though they are totally electromagnetic in nature.[2]

As is well-known, all observation in physics is considered 3-spatial..[3] Mass is a 3-spatial concept, and we detect changes to mass (as, e.g., in the shift of electrons in the circuits of electrical instruments). So one may model the common physical observation mechanism as a *time-differentiating process* which *a priori* discards time-change and retains 3-spatial energy change.[4] That is, Minkowski reality is modeled in the fundamental units of L^3t. Physical observation (via the transverse photon interaction) is the process given by applying the operator $\partial/\partial t$ to L^3t, yielding an L^3 output. Hence mind and mind operations are excluded by the usual physics instruments and observation, which simply exclude the time domain in their outputs and do not "measure" it.

For this reason, physicists have erroneously considered mind to be "metaphysical" and nonreal. Indeed, most physicists to one extent or the other are materialists, and consider the "mind" to be nothing but the operations ongoing in a "meat computer." Nonetheless, in the time-domain the time is absolutely real, and it is completely electrodynamic in nature. There is no metaphysics involved, and the temporal domain—along with mind and mind operations—is simply an erroneously neglected area of physics. In present physics, the notion of mind is comfortably disposed of by imposing the use of the "observer" concept, with out ever specifying that the observer has a consciousness and a mind. Indeed, "observation" is only about what that "observer" perceives. Obviously, one has a dramatically crippled physics if one eliminates that nonobservable called "time." Similarly, one also has a dramatically crippled physics when one eliminates the mechanisms and physics ongoing in those time-like and dynamic "things" such as mind, that occupy time and function in it.

From this viewpoint, Western physics adheres to its 3-spatial measurement foundation only by ignoring the transduction of time-polarized EM changes into detectable longitudinal and transverse EM wave changes. In this aspect, present Western physics is severely self-crippled.

Time is totally electromagnetic and energetic in nature. The flow of time is not a separate external river on which a mass floats along like a boat drifting down the current of a great flowing river. Instead, the flow of time is generated directly on every mass by its total set of photon interactions, both virtual and ob servable. We have previously presented the exact mechanism for the flow of time. *[Figure 1]*

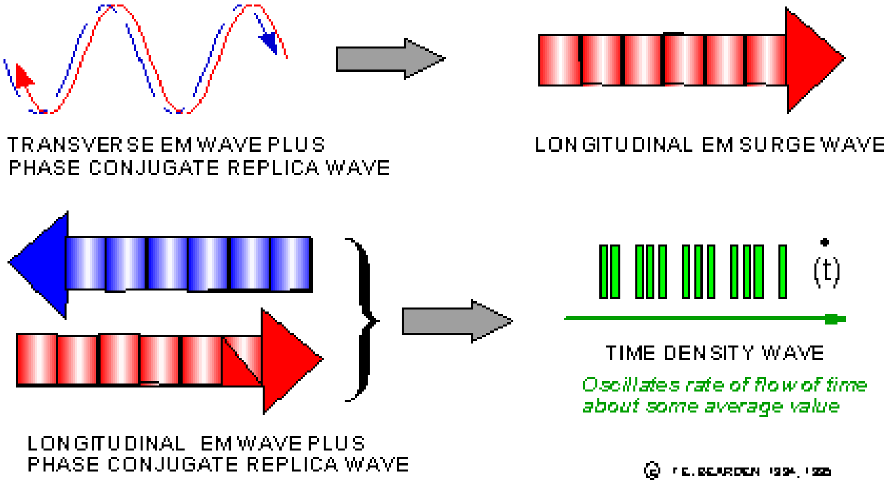
TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
LONGITUDINAL EM SURGE WAVE
LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
$\dot{(t)}$
TIME DENSITY WAVE
Oscillates rate of flow of time
about some average value
© T.E. BEARDEN 1994, 1995

Figure 2. Transduction of EM wave type by successive phase conjugate pairing.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Further, a photon is comprised of angular momentum, therefore of (energy)x(time). It transports both energy and time, not just energy alone. When absorbed by a mass, not only does it "energy-excitation" charge the mass, but also it "time-excitation" charges it, converting the former "mass" to "mass time". Rigorously it is *masstime* that emits a photon, not *mass*. So a mass moves through time in little spurts, by the continual macroscopic addition and subtraction of little Dt components. Further, in being driven through time, mass is continually altered to masstime (a quite different critter from mass), to mass, to masstime, to mass, etc.

In any masstime state, a myriad of tiny virtual photon interactions made of very tiny (DE)(Dt)'s interact with that same mass during that particular larger Dt of the masstime state (m+Dm)(Dt). Hence mass in its alternate masstime state has that state internally structured by its surrounding environment's interactions with it. The masstime state is internally structured energetically in its (DE) component, and also internally structured temporally in its (Dt) component.

A standard charged fundamental particle such as an electron, e.g., is not necessarily identical with another, when the internal structuring of its masstime state is considered. Further, by two papers by Whittaker,[5] interferometry of masstime states with either energy reactions or temporal reactions can yield observable effects and changes due to these neglected "hidden variables" in the electron's masstime alternative states. To engineer the mind and its operations directly, one must perform electrodynamic engineering in the *time* domain, not in the 3-space EM energy density domain. The direct engineering of time-like mind and mind operations—in all levels and all aspects—requires the use of *time-polarized photons and time-polarized EM waves*. This is the rarest form of electrodynamics, almost untouched by Western physicists.

One can either painfully produce such time-polarized photons and EM waves and directly irradiate a target mass with specific assemblies of them, or one can force the mass itself to iteratively transduce ordinary transverse EM waves first into longitudinally-polarized EM waves and then into time-polarized EM waves. Both will be discussed briefly.

In the West, it appears that the present author's discovery of mechanisms for producing time-polarized (scalar) EM waves [see *Figure 2*] and for transducing between wave polarizations has no precedent. [See *Figure 3*]. While scalar (time-polarized) photons are known in the literature,[1] the creation and use of time-polarized EM waves does not seem to appear in the Western physics literature.[6]

Wave transduction or *polarization transductions* are terms used by the author for the process of transforming an EM wave's particular polarization into another polarization type. Such transduction to other than transverse polarization forms is little known in the West. It appears essentially haphazardly in some experiments, usually without any recognition by the experimenters themselves. Significant transduction in experiments will also produce "strange" and unexplained anomalies in the instruments being used for experimental measurements.[7]

So most probably Western mind control researchers have not recognized the methodology and mechanisms[8] for transducing one type of EM wave polarization directly into another.[9] They continue to seek the "mind" in 3-space and hence in the brain, rather than in the time domain. On the other hand, it appears

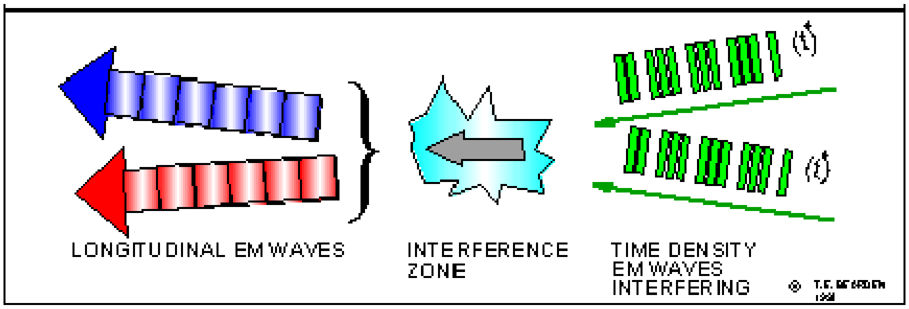

Figure 3. Interference processes for transducing higher EM wave polarizations into lower polarization states.

EXPLORE! VOLUME 9, NUMBER 2, 1999

that the Russian KGB *energetics*[10] weapons scientists—particularly those in *psychoenergetics*[11]—have known and used methods of transducing one wave polarization into another, for at least two decades. It follows that those same scientists have very probably developed *mind engineering* and *mind control* via novel time-polarized EM wave means and a dramatically extended electrodynamics of the time-like mind operations.

In the present paper we briefly develop the wave transductions and the basis for mind engineering, including the specialized use of ordinary transverse wave (TW) EM spectra to force internal EM wave transductions and time-domain operations inside irradiated bodies, cells, and tissues. Through the mind-body coupling mechanism, these transductions of transverse EM waves can operate upon the mind and its deepest operations as well as upon the body, every cell, and every part of every cell.

Russian Mind Control Uses Higher EM Wave Polarizations

It is apparent that the KGB psychoenergetics weapons scientists know and use the full extended EM wave polarization range. As a postulation, one would expect our own weapons scientists to know and utilize the orthodox transverse wave (TW) EM for similar research and experimentation. However, in the West scientists are *just beginning* to realize the importance of a very general (and weaker) type of longitudinal EM wave (LW) polarization, as evidenced by the appearance of dozens of papers on "undistorted progressive waves" (UPWs, which are essentially imperfect LW s with TW residues remaining). The Los Alamos National Laboratory web site has a number of such papers—particularly by Rodrigues[12] and Lu—available for free downloading. UPWs have very interesting characteristics. If they were perfect longitudinal EM waves, they would have infinite energy and infinite speed. Since only imperfect UPWs can be physically made, their speed can vary from slower than the speed of light to faster than the speed of light. Their energy can also vary over a great range.

Another characteristic of UPWs that are reasonably good longitudinal EM waves, is that they pass readily through a large depth of water and mass, including through the ocean and the earth with only small interactions and losses. Yet by interfering two such beams of "high quality" UPWs at a great distance, then in the interference zone ordinary EM energy will rise directly out of local spacetime potential, as essentially shown by Whittaker[13] nearly a century ago. Russian weapon research facilities have weaponized these effects[14] for nearly 50 years, under rigid KGB control and operation. Nonetheless, even if using only ordinary TW waves, Western mind control researchers may get some fairly good results, brute-force-like, by using gross correlates between just the input irradiating TWs and the exhibited behavioral responses of the individual. The net input-output correlations can be determined, even though not taking into account the actual wave polarization transduction mechanisms ongoing inside the irradiated dielectric (or brain, or mind, etc.). However, Western researchers appear to have no knowledge of the exact mechanisms by means of which coherent time-like mind operations of a biological organism couple to the organism's 3-spatial body to provide the mental control loop. They also appear unaware of how the coherent 3-spatial behavior responses of the body couple back to the time-like mind to provide it with a sensory feedback of the body's responses.

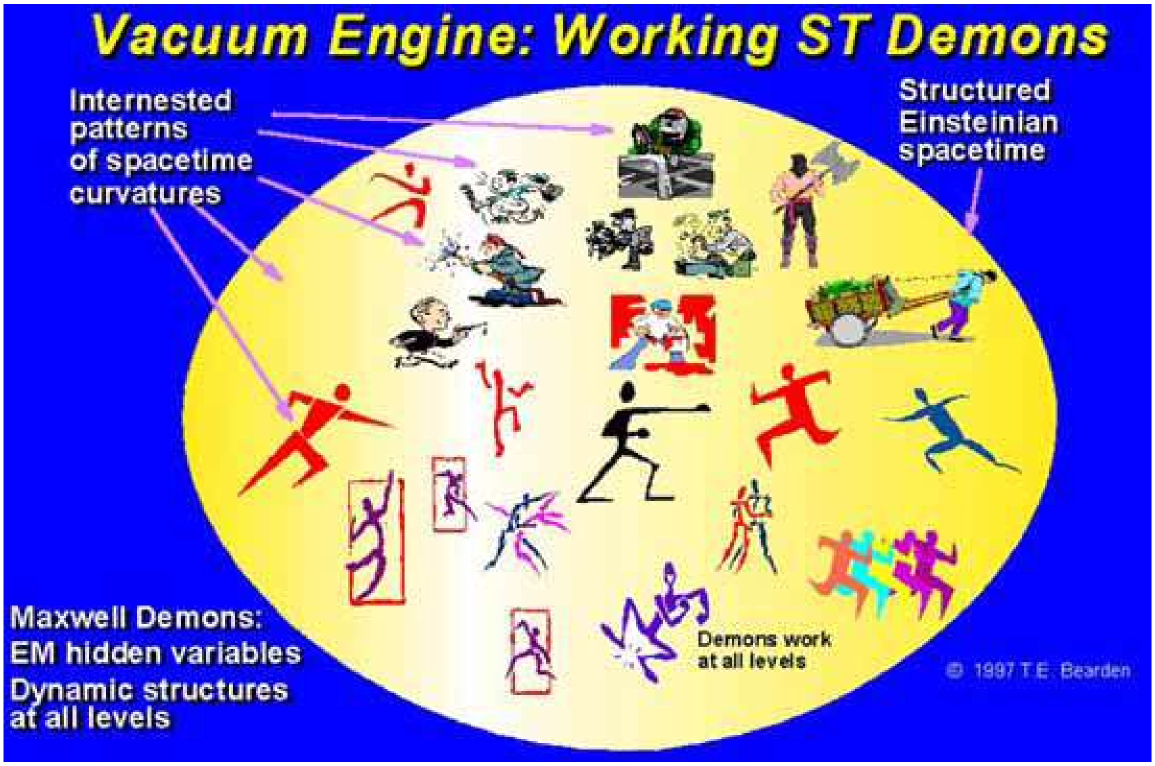
Vacuum Engine: Working ST Demons
Internested patterns of spacetime curvatures
Structured Einsteinian spacetime
Maxwell Demons: EM hidden variables
Dynamic structures at all levels
Demons work at all levels
© 1997 T.E. Bearden

Figure 4. A spacetime curvature engine has myriads of small ST curvature components which act upon mass at all levels.

EXPLORE! VOLUME 9, NUMBER 2, 1999

General Relativity Aspects

Both these coupling mechanisms can be taken directly from general relativity, if one puts one's mind to it. However, unless the wave polarization transformations are known, it is difficult or impossible to apply the ordinary GR *directly*, since relativists usually concentrate upon spacetime curvature by altering 3-space energy density rather than by altering time-polarized "time-energy" density.

In applied general relativity, spacetime itself is an active medium. One speaks of a change in spacetime as a "spacetime curvature". Any curvature of spacetime in a local region directly performs continuous work upon any mass embedded in that region. Vice versa, any mass (or other concentration of energy) in a local spacetime continuously acts upon that spacetime to "curve" it.

So if one wishes to perform actions upon mass in a distant region, one may generate local curvatures of spacetime in that region, and these curvature "engines" will directly act upon the mass at all levels. *[Figure 4]* This is very different from energy propagation through space from one point to another. Now the "ordinary EM energy and actions" arise from every spatiotemporal point within the mass, at every level, and move upwards (from inside to outside). So we speak of specific forms of "templates" of nested spacetime curvatures formed and utilized to engineer mass "from inside out" as *spacetime curvature engines or vacuum engines*.

Use of vacuum (spacetime curvature) engineering[15] is a far more powerful form of engineering than is provided by energy propagation through space. As an example, it is easy to alter the quarks in a nucleon, using spacetime curvature engines and time-charging decay. Indeed, we have developed the mechanisms for cold fusion and the electronuclear interaction (formation of new nuclides) at feeble energy. An Invention has been filed with the U.S. Patent Office, and formal patent applications are in preparation. Some information on these mechanisms and principles has been released[17].

Time as Dense EM Energy and a Strong Spacetime Curvature Agent

The advantage of using the time-polarized "time-energy" for spacetime curvature is that time is ordinary energy compressed by a factor of at least c^2—which, in the MKS system of units, is some 9¥10^{16}. Thus use of time-polarized EM photons and waves as ST curvature agents gives an amplification of 9 ¥10^{16} over the use of transverse-polarized EM waves for that purpose. In turn, the use of the strong EM force in ordinary TW waves as an agent of ST curvature is already a nominal 10^{40} times as strong as is the weak G force used as the agent of ST curvature.

The bottom line is this: For spacetime curvature effects, the use of the t-polarized domain provides amplification of some 9¥10^{56} greater than the weak G-force ST curvature agent usually considered in general relativity[18].

Western Science Remains Largely Materialistic

Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system.[8] This serious self-limitation exists because in the body we *measure* only weak TW EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to

Disclosure[16] on this work

measure "mind operations" directly.

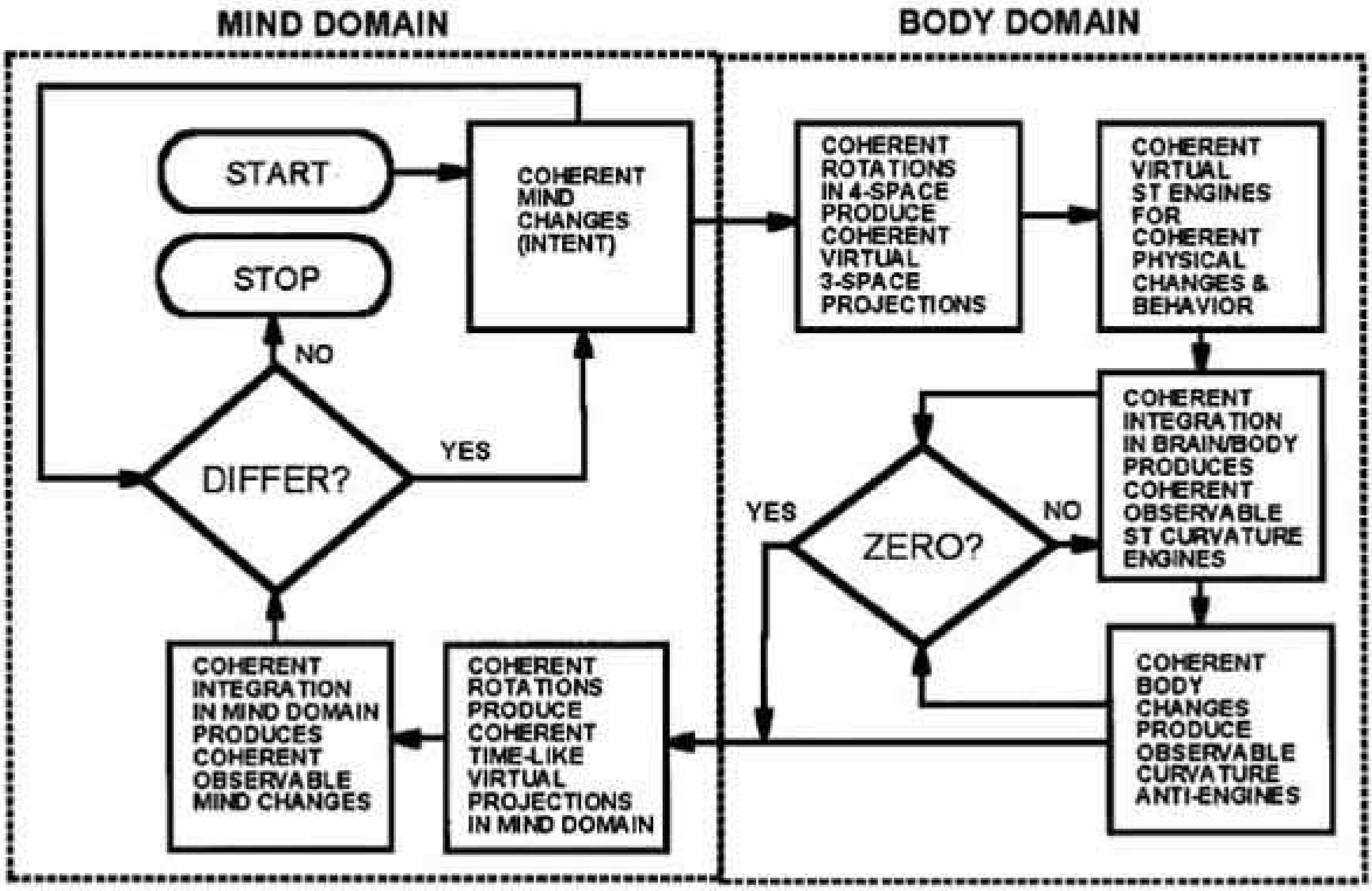
MIND DOMAIN
BODY DOMAIN
START
STOP
COHERENT MIND CHANGES (INTENT)
NO
YES
DIFFER?
COHERENT INTEGRATION IN MIND DOMAIN PRODUCES COHERENT OBSERVABLE MIND CHANGES
COHERENT ROTATIONS PRODUCE COHERENT TIME-LIKE VIRTUAL PROJECTIONS IN MIND DOMAIN
COHERENT ROTATIONS IN 4-SPACE PRODUCE COHERENT VIRTUAL 3-SPACE PROJECTIONS
COHERENT VIRTUAL ST ENGINES FOR COHERENT PHYSICAL CHANGES & BEHAVIOR
COHERENT INTEGRATION IN BRAIN/BODY PRODUCES COHERENT OBSERVABLE ST CURVATURE ENGINES
YES
NO
ZERO?
COHERENT BODY CHANGES PRODUCE OBSERVABLE CURVATURE ANTI-ENGINES

Figure 5. Mind-to body coupling and body-to-mind coupling.

EXPLORE! VOLUME , NUMBER 2, 1999

With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.

Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW functions exist in the body and in all of nature in general. Presently only highly theoretical quantum field theorists seem to even be aware of the existence of scalar (t-polarized) photons, and even they are unaware of t-polarized EM waves. Certainly our scientists do not measure t-polarized EM waves in and around the body and in nature, nor do they make them in the laboratory. Apparently they have not studied such waves and their interactions with matter—living and inert—at all.

But it's even worse. In ignoring the time component transported by photons and EM waves, science has also erroneously omitted half of the excitation charging and excitation decay processes whenever a mass interacts with photons and EM waves. More on that later.

Brute Forcing Time Functions versus Fine Control Methods

Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.

While KGB scientists also use TW EM "brute force" TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired. They also do not hesitate to use LWs and t-polarized waves overtly. So in blunt terms, Western compared to Russian mind control research is probably like comparing an automobile body shop with a fine surgical ward. Having long ago worked out all those "exact correlates" to the internal "information content of the field" is where the KGB scientists are at least 20 years ahead of the West.

The reason one can get mind and behavior results with the TW electromagnetics, while bypassing the real mechanism which uses t-polarized waves and photons, is due to (1) the very peculiar nature of the EM emission from a dielectric, and (2) what can be done by re-radiating that dielectric with its emitted spectrum, deliberately and very carefully altered in selected parts. We will return to that important feature later.

Polarization and Observability

As we stated, there are four photon polarizations and therefore there must be correspondingly four EM wave polarizations[1] The first three polarizations are the x-, y-, and z- spatial polarizations. The x- and y- polarizations are transverse polarizations and the z-polarization is a longitudinal polarization along the direction of propagation (along the z-axis, by standard notation). Simply put, we may visualize the transverse polarizations as rather like the wiggling of a fish's tail from side to side as the fish moves forward, or a whale's flukes up and down as the whale moves forward, or some combination thereof. We may visualize the longitudinal polarization as a sort of "repetitive accordion effect" along the line of motion of the wave. Usually the z- polarization is neglected in EM wave theory, although in recent years physicists have "rediscovered" longitudinal EM waves and are now intensely researching the use and characteristics of such waves.[12]

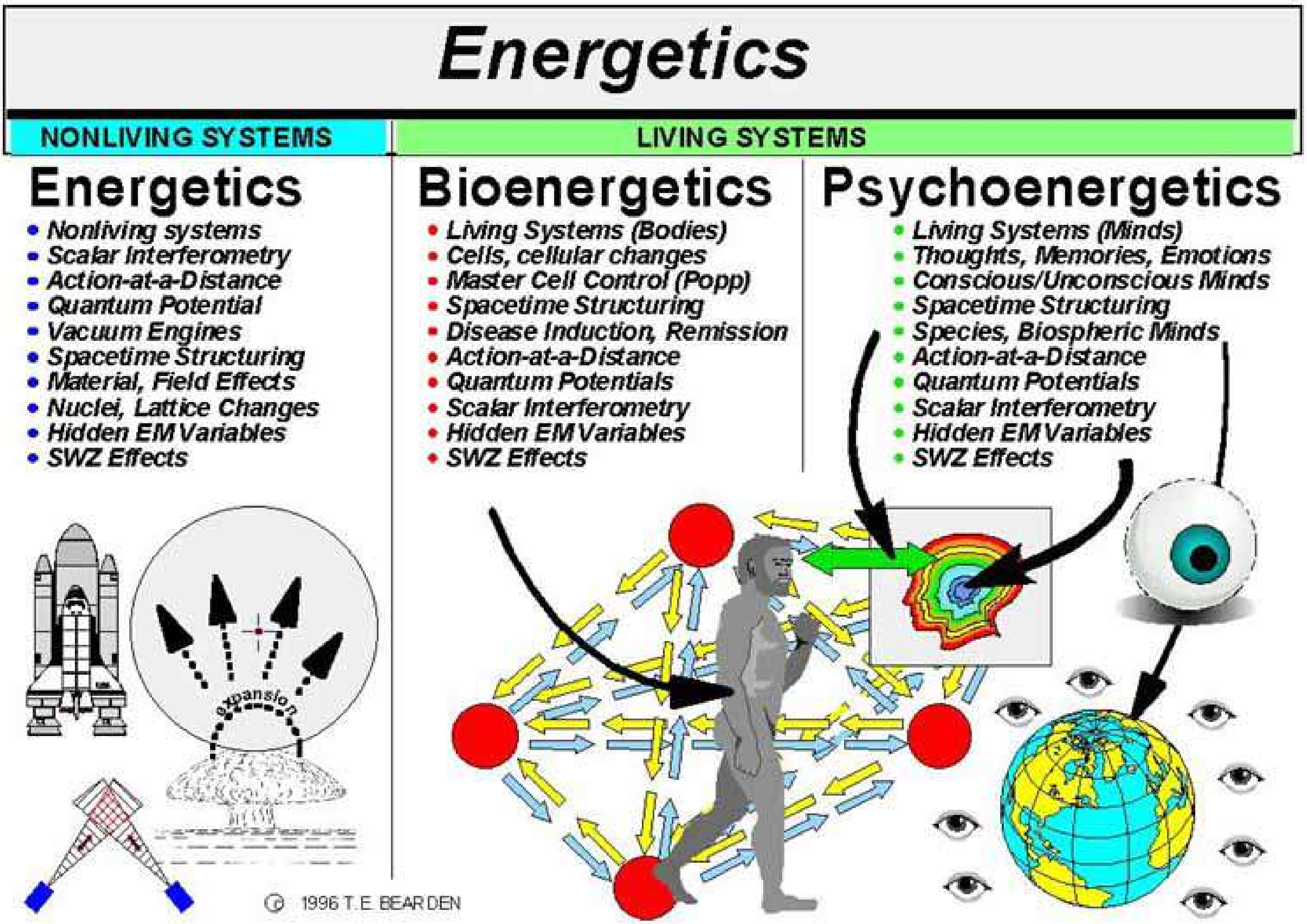

Energetics

NONLIVING SYSTEMS | **LIVING SYSTEMS**

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T.E. BEARDEN

Figure 6. The three divisions of Soviet energetics and their characteristics.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Unfortunately, in quantum field theory there has been a tendency to regard the t-polarized or "scalar" photon—where the local rate of time is oscillated—as unobservable. Oscillation of the local rate of time produces powerful oscillations of local space-time curvatures, due to the extreme energy density of time. Individually, the t-polarized photon tends to be unobservable. However, a *coherent sequential group* of such t-polarized photons, each individually in the virtual state with respect to the external observer, may simply integrate in its interaction with mass over a short time period into observable change because of the coherent integration of virtual spacetime curvatures into a larger, observable spacetime curvature.

Rotations and Projections in 4-Space

Relativistically, any velocity change in 4-space is a rotation. Any such rotation in the time-domain also creates a tiny projection component into 3-space. Any such rotation in the 3-spatial domain also creates a tiny projection component into the time domain.

Hence a sufficient series of coherent time-domain (mental) changes produces a coherent series of virtual changes in 3-space (the body), thereby coherently integrating into an observable energy change in the body. A sufficient series of coherent 3-spatial energy changes produces a coherent series of virtual changes in the time domain which coherent integrate into an observable time-domain change.

Solution to the Age-Old Philosophical Problem of Intent

This is in fact the solution to the age-old problem of *intent*, or how the nonmaterial (i.e., non-3-spatial and non-observable) mind induces a 3-spatial, observable energy change upon the 3-spatial body. It is also the solution to the problem of *awareness*; i.e., how the mind is aware of itself[19] and of the responses of the body. Together the two form a closed loop coupling of the mind and body. The time delay in the loop together with memory recall for comparison, creates the sense of "persistence" of self in time. This also creates the sense of being a "separate, closed being" (i.e., of separate persistence in time—in the living entity). The sense of "separation of self from an external world" is created by comparing those body-to-mind sensory feedbacks which are not correlated to the mind's previous feed-forward intent.

Time As Energy and Why It Is Very Dense Energy

In addition to the three spatial polarizations of photons and EM waves, there is a very, very useful t-polarization along the time axis. In this polarization, the 3-spatial energy is not oscillating at all. Instead, the time or time-energy is oscillating. Time can be taken to be energy compressed by at least c^2, so it has at least the same energy density as mass. In other words, one second is 9×10^{16} joules of time-energy (energy compressed into time). The t-polarized photon or EM wave is called the scalar photon or scalar EM wave, respectively.

To demonstrate why time can be regarded as energy, we need only point out that the choice of fundamental units in any physics model is totally arbitrary. We usually choose these units to ease the mathematical manipulation, ease our understanding, and simplify the ability to visualize or "grasp" the physics. However, if one wishes, one can build all of physics from a single fundamental unit—e.g., energy. In that case, one readily sees that time is a function of energy and only energy. Hence it is perfectly proper to regard time as energy, and to seek out what form the function takes in our normal system of units (say, the MKS system).

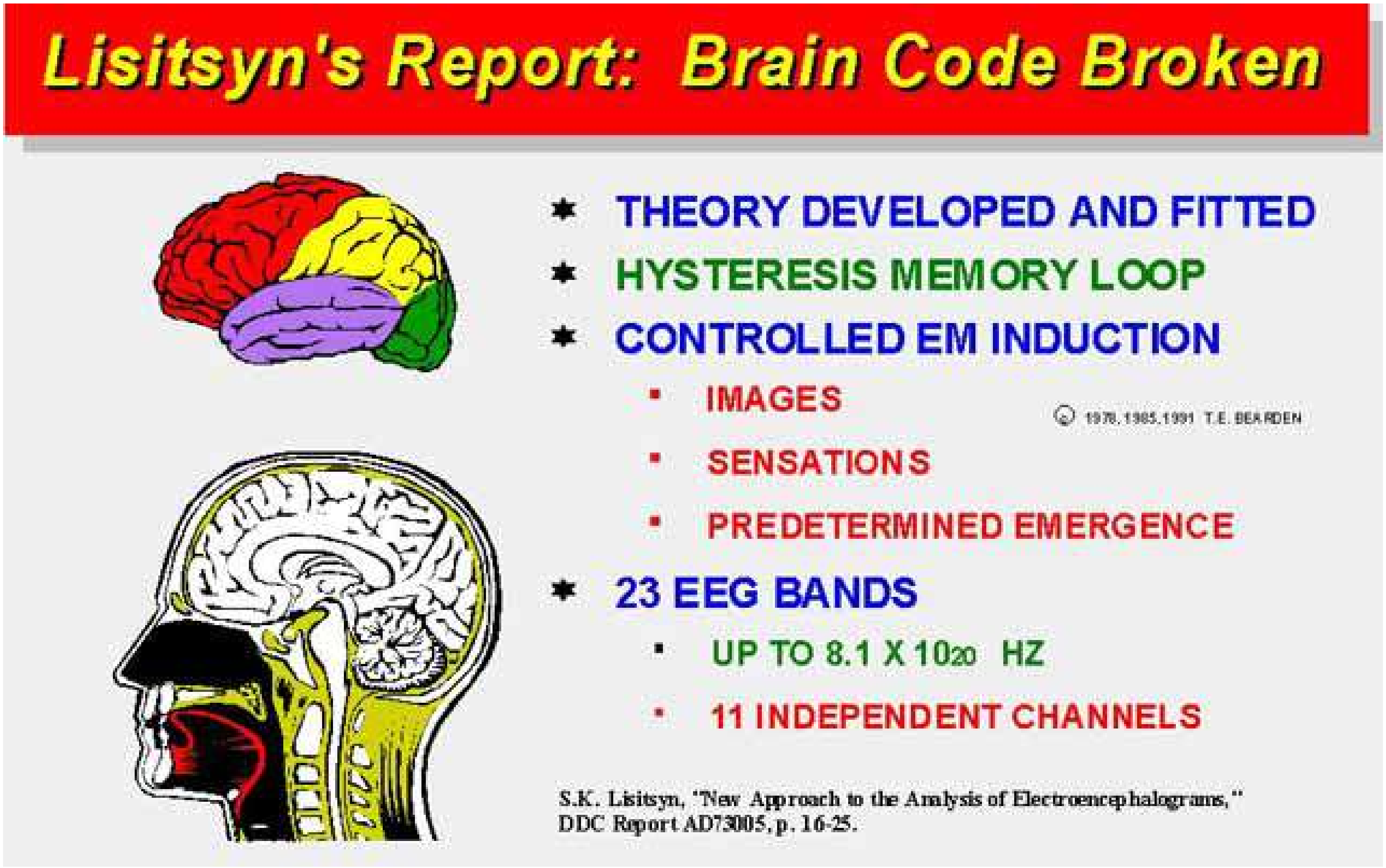

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

Figure 7. In the 1960s, Lisitsyn revealed Russian scientists had deciphered the human brain code.

EXPLORE! VOLUME 9, NUMBER 2, 1999

In that system, it appears that any change in time Δt can be taken as always a function of a corresponding change in energy ΔE, where $\Delta t \leq (\Delta E) \div c^2$ [1] from which it follows that $\Delta E \geq (\Delta t) \times c^2$ [2] For convenience, we take the special case where $\Delta t = (\Delta E) \div c^2$ [3] $\Delta E = (\Delta t) \times c^2$ [4].

Observation as Used in Physics Is Spatial

We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain, and thus "lost" by the stripping away of time in the ordinary observation process. Physical observation is essentially a time-differentiating mechanism applied to a 4-spatial change, or in terms of MKS fundamental units L and t, observation θ is $\theta = \partial/\partial t(L^3 t) = L^3$ [5].

Thus, as is well-known in quantum mechanics, physical observation is 3-spatial, and time is not a physical observable, even in theory. Since mind is time-like, it follows that mind is not a physical observable either, since discarding the time dimension also discards the mind. In short, one may also take physical operation as the mechanism that separates mind and body. To observe is to separate. We point out, however, that merely "separating and discarding" the time domain (as in physical observation of a single change) does not eliminate it, nor does it eliminate the single time-like mind change that may be involved in an intentional volition. The time domain certainly remains, even though only the 3-spatial intersection of the 4-spatial mechanism and the resulting "outputs". So a series of coherently integrated "mental intent" changes introduced into the human body's overall servo-mechanism provides the continuing input. From there, ordinary physics will generate the resulting actions induced in the body by that servomechanism and its amplifying mechanisms.

Two Coupling Mechanisms Make a Closed Loop

This is the "mind-to-body" coupling mechanism. [See *Figure 5*] It is the mechanism whereby the mind is coupled to the body, and whereby mental intent is able to induce a series of physical inputs into the body servomechanism.

So the body's servomechanism then generates the responses of the body (including everything from chemistry to electrical changes to muscular movements, etc.). These responses are changes in 3-space.

The conscious mind is a *serial processor*, though extremely rapid. It produces the series of coherent intent inputs for volitional behavior.

The unconscious mind is a *massively parallel processor*.[20] It continually produces the vast series of coherent "unconscious intent inputs" for control of all the deeper processes in the body, beyond usual conscious awareness.

However, as the body moves or changes in 3-space, each resulting quantum change in body 3-space is also a slight rotation out of 3-space and toward the time axis. Hence it induces—in the "virtual state" in the time domain—a precisely correlated projection. The body's responses are in general

Minkowski change is given by physical observation. So we may say *that the time-change remains in the virtual state, with respect to attempted physical observation of a single Minkowski 4-space change.*

Rotation, Coherent Integration, and Intent

Previously we pointed out that any change in a 4-space entity may be regarded as a rotation away from the "trajectory" of the entity. Hence each and every t-polarized change creates a very small projection into 3-space by its rotation slightly away from the time axis. All mind changes in the time-domain actually produce virtual 3-spatial projections in 3-space (in the physical domain). We define "intent" as the continued production of successively coherent mental changes, producing coherent virtual changes in the 3-space body system, with coherent integration of those 3-space virtual changes into observable 3-space energy inputs into the 3-space body system. Successively coherent mind-changes will produce successively coherent 3-space virtual changes. In short, coherent mind-changes will produce coherent integration of those 3-space virtual projections into an observable 3-space change. This is the creation of an ordinary 3-spatial energy change induced upon the 3-space body. In short, this is the mechanism whereby time-like mental intent is able to produce a series of coherent observable quantum changes in the physical body (as in the brain and nervous system).

The Body as a Servomechanism

From the standpoint of control theory, we may consider the body system to be a coherent, so a coherent series of virtual state changes in the time-domain (in the mind realm) are created successively. In short, again we have coherent integration, this time in the mind or time domain. This produces "observable" changes in the time-mind domain, which are coherent with the body's 3-spatial changes actually performed. Thus the mind receives feedback directly from the physical movements and changes of the body. This is the manner in which the body is coupled back to the mind.[21]

By comparing the "intent" behavioral move that was "fed forward" into the body, with the return "response" move analog that was "fed back" from the body to the mind, the mind is able to determine errors and differences, and originate additional correctional commands.

Thus the entire mind-body loop is a closed-circuit system of feedforward and feedback, together with corrections. It also has multiple levels of such, infolded in the larger volitional levels. See again *Figure 4*.

This solves the age-old problem of the mechanisms for the mind-body coupling, intent, volition, conscious and unconscious functioning, sense of the external world, sense of the internal world, sense of "being in" the external world, etc.[22]

All This Is Included in Russian Psychoenergetics

This is the highly summarized basis for *psychoenergetics*, the KGB's division of energetics that deals with the mind and body coupling and functions, and direct engineering of (1) the mind-body coupling and (2) the mind operations directly. [See *Figure 6*] The KGB intent, of course, has always been to exploit this science for the complex servomechanism system comprised of many subordinate servo systems with feedback and feedforward looping.

Given the input from intent, ordinary servo theory will take it from there. Once a servo has an input, servomechanism theory describes the response of the

degradation, killing, and control of human beings, including all humanity.

The Russians know full well that, if you produce and utilize scalar EM photons and t-polarized EM waves, you can directly affect and engineer mind and mind operations at any and all levels. Western researchers, who know nothing of how to make

EXPLORE! VOLUME 9, NUMBER 2, 1999

time-polarized photons and time-polarized EM waves, do not yet know that. Consequently the Russians have developed a highly secret science of *directly* engineering the mind and its operations, including thought, images, perceptions, feelings, emotions, memory, and mind processing. Indeed, Lisitsyn[23] wrote quite specifically of this capability in the 1960s. [See *Figure 7*]

Western clandestine mind control researchers are apparently still slowly and painfully fitting TW EM irradiation correlates to induced or resulting mental and physical behavior responses. They seem unaware of the actual wave transductions occurring inside the body and mind, but are unwittingly inducing those transductions in hidden fashion anyway and in "brute-force input-response fitting" models.

These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design—for example—of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gaia's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.

This fine research article will be continued in Part II with Russian Methodology, Waves and Wave

spacetime that is highly active. The mind is rooted in the time-domain and projects from that domain into 3-space. The body is rooted in the 3-space domain and projects into the time-domain. To "change" or "function" in one domain is automatically to function in the other. EM and gravitational phenomena are still modeled separately by human scientists, but not by nature.

3. In quantum mechanics, time is not an observable, but merely a parameter.

4. It is convenient to consider mass as a 3-spatial form of condensed energy.

5. E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics, " Mathematische Annalen, Vol. 57, 1903, p. 333-355;— "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions, "Proc. Lond. Math. Soc., Series 2, Vol. I, 1904, p. 367-372.

6. A NERAC document search yielded some 16 good references on t-polarized photons, but not a single reference on t-polarized EM waves.

7. E.g., such previously inexplicable instrument anomalies have accompanied some excellent and rigorous electrolyte experiments at U. S. Navy research facilities at China Lake. For a description, see Melvin H Miles and Benjamin F. Bush, "Radiation measurements at China Lake: Real or Artifacts?", Proc. ICCF - 7 (International Conference on Cold Fusion—7, Vancouver, BC, Canada, Apr. 1998, p. 101. For a brief explanation of the anomalies, see T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves, " Explore!, 8(6), 1998, p. 7-16;— "Toward a Practical Unified Field Theory and a Deep Experimental Example," Proc. INE Symposium, Univ. Utah, Aug. 14-15, 1998.

8. Many U. S. researchers and journalists—and even many scientists—have wrestled with the problems of mind, intent, and mind-body

Transduction, The Cellular Control System, and other matters of interest. ♦

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel US. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, US. Army. Graduate of Guided Missile Staff Officers Course, U. S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses.

Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor: He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuums virtual particle flux is organized and used to energetically shape and manipulate matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations

coupling. Enumeration of those efforts would itself require an entire book. For an introduction into that domain, the reader is referred to the very determined exposé by Cheryl Welsh in her very timely book, The 1950's Discovery of the Code of the Brain, May 1988, published on the Internet. Simply employ any Net search engine and search on the name Cheryl Welsh. A connection to the site and to Welsh's book will immediately be found Welsh's important compendium is highly recommended, giving the reader a "birds eye view", so to speak, of how Western scientists think and proceed regarding the subject of mind control by electromagnetic means. Unfortunately all journalists so far researching the area have not been aware that Western electrodynamics itself is seriously flawed, and that the great Russian advances in mind engineering and mind control are a result of their secret but complete revision and correction of Western electrodynamics to provide the basis for energetics. Until Western scientists revise their own decrepit old electrodynamics fouling the textbooks, they will never catch up to the Russian mind control developments springing from intense development programs that did that revision in the late 1940s and early 1950s.

9. An Invention Disclosure on these processes and enhancement embodiments has been filed with the U. S. Patent Office. Formal patent applications are also in preparation.

10. Energetics is a unified science of an extended electrodynamics possessing a hidden infolded general relativity inside the potentials, fields, and waves. It deliberately employs higher polarizations (longitudinal and tempic) of photons and EM waves to engineer action at a distance, use of specific patterns of spacetime curvature created at a distance, and direct engineering of either inert or living bodies. It also encompasses the direct engineering of time-like mind and mind functions. It is divided into three divisions, depending upon the nature of the target. Targeted against inert materials, structures, fields, and waves it is called "energetics. "Against living bodies, their fields,

and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

---

1. E.g., see Lewis H Ryder, Quantum Field Theory, Cambridge University Press, 2nd Edition, 1996, p. 147+. For an even deeper discussion of photon polarization, see F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, under the heading "Covariant Quantization of the Photon Propagator" in Chapter 5. Mandl and Shaw argue that the longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e. electrostatic) potential.
2. Nature is unified and not separated into the "separate" disciplines our science has prescribed. Thus the mind is also totally electrogravitational. The mind works not just with EM forms in an inert space, but also with spacetime curvatures in a

and biochemistry, it is called "bioenergetics." Against the mind and mind functions, it is called "psychoenergetics".
11. Psychoenergetics is that branch of energetics used to directly affect and engineer the mind, memory, perception, thought, emotion, sensation, personality, and other mind functions of a living biological organism. It is usable against a single personal conscious (serial processor) mind and a single personal unconscious (massively parallel) processor mind In theory it is also usable against lungs collective unconscious mind and its functions, as well as the deeper unconscious for all species on Earth (i.e., against Gaia). We are speaking of physics, not metaphysics.
12. W A. Rodrigues Jr.. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 \leq v < \infty$ in nature," Foundations of Physics, 27(3), 1997, p. 435-5•08 A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory website.
13. E. T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions, "Proc. Lond. Math. Soc., Series 2, Vol. I, 1904, p. 367-372. The paper was published in 1904 and orally delivered in 1903.
14. E.g., see T. E. Bearden, Fer-de-Lance, Tesla Book Co., 1986;—AIDS: Biological Warfare, Tesla Book Co., 1988;—Soviet Weather Engineering Over North America, 1-hour videotape, 1985,—Gravitobiology; A New Biophysics, Tesla Book Co., P.O. Box 121873, Chula Vista, CA 91912, 1991,—Energetics of Free Energy Systems and Vacuum Engine

EXPLORE! VOLUME 9, NUMBER 2, 1999

Therapies, Tara Publishing, Internet node www. tarapublishing.com/books, July 1997; — Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ pages furnished to selected government agencies and key Congressional committee leaders.

15. E.g., see T.D. Lee, Particle Physics and Introduction to Field Theory, Harwood; New York, 1981. On P.383, Nobelist Lee points out that the microstructure of the scalar vacuum field (i.e., of vacuum charge) is not utilized. Particularly see Lee's own attempt to indicate the possibility of using vacuum engineering, in his "Chapter 25; Outlook; Possibility of Vacuum Engineering, " p. 824-828. With wave transductions and use of higher polarization EM waves, there are far better and more powerful ways to do vacuum (spacetime curvature) engineering.

16 T.E. Bearden, "Formation and Use of Time-Reversal Zones, EM Wave Transduction, Time-Density (Scalar) EM Excitation and Decay, and Spacetime curvature Engines to Alter Matter and Convert Time Into Energy," Invention Disclosure Document #446522, Oct. 26, 1998.

17. T.E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves, "Explore!, 8(6), 1998, p. 7-16; — "Toward a Practical Unified Field Theory and a Deep Experimental Example," Proc. INE Symposium, Univ. Utah, Aug. 14-15, 1998.

18. We have used the enormous energy density of time in our approach to a totally new kind of electronuclear interaction at feeble spatial energy. By use of a time-reversal zone (TRZ) in a fluid; the free H+ ions—which are free protons—can be induced to attract each other so tightly that the edge of the nuclear force region is entered. In that case the protons form a quasi-nucleus. When the TRZ decays back to a time-forward zone (TRZ), this new quasi-nucleus decays by changing quarks in one or more

means "sense of existing in time".

20. The "unconscious" is totally conscious, just multiply so. The conscious mind can see only "a single slide in the slide projector at a time, " so to speak. When it "looks" at the ongoings and outputs of the unconscious, it "sees" millions of slides in the slide projector simultaneously. Hence it cannot distinguish "one and one only". So it just sees nothing, where by "nothing" we mean "no single distinguishable thing." It therefore is unable to consciously perceive the actions of the unconscious. When the unconscious wishes to communicate with the conscious (because of unresolved conflicts, etc.), it must produce large groupings integrated into "symbolic scenarios". Hence the symbolism of the unconscious, and the basis for therapeutic psychology and psychiatry. A "symbol", after all, is just a thing which can have many different meanings, all folded into one.

21. The mind continually has activities of its own, each of which produces coherent changes in the body to a much smaller degree. The interweaving of these "smaller self-changes" infolded inside the 3-space major intent changes also produces slight body changes, which in turn "project back" into the mind. This huge group of "infolded self-loops" produces the sense of "my body" being "volitionally directed or occupied" simultaneously with the "greater intent" loops which represent "my intentional actions with my body". Further; the body servo continually receives a myriad of inputs from its sensory systems, both from changes inside the body and from incoming changes from the external universe. This interweaving produces the awareness of "being in the external world" by having (i) changes whose feedback do not match mind-intent changes, and (ii) changes whose feedback does match mind-intent changes. This of course applies to both the conscious and unconscious levels. The purpose here is to demonstrate that there exists an actual physics of the living being, the mind as well as

nucleons, thereby producing new nuclides. We point out that this is actually a very high energy reaction; however; since the time-energy is used and is very, very dense, the amount of 3-spatial energy that must be used is very small. If this new mechanism holds, it represents a revolution in particle physics.

19. "Self-awareness" refers to be continually "refreshed" by time-delayed coherent feed-back inputs from the body to the mind; where these inputs compare to the recalled memory of the minds inputs to the body. In short, the time-delay coherence creates the sense of "existing in time". Awareness usually

the body, and the coupling and interfiinctioning of mind and body.

22. The methodology also easily extends to include deeper interlays for the collective human unconscious mind and for Gaia—the collective unconscious mind level for all the species on earth.

23. S. K Lisitsyn, "New Approach to the Analysis of Electroencephalograms, " in Problems of Bionics— (Selected Articles), Defense Documentation Center Report AD 73005, date unknown, p. 16-25. Available through the National Technical Information System (NTIS).

Mind Control and EM Wave Polarization Transductions, Part 2

©Copyright 1999 by T.E. Bearden, USA

This article continues on with T.E. Bearden's article, Part 1 of which was published in Volume 9, #2 of Explore! for the Professional. Because we feel that Col. Bearden's caution below is so very important we are repeating its publication in all parts of this article.

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for non-approved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

A Strategic Threat Completely Unrecognized by the West

The Russian methodology leads directly to rather mind-bending, extended capabilities. KGB psychoenergetics scientists seem fixed on an eventual goal of directly engineering Jung's collective human unconscious, thereby converting the entire human species into a sort of "ant" species. In short, they seek to engineer what they regard as "perfect communism and perfect order" directly in the entire species. *Figure 8* indicates the direct connection of every human's conscious and unconscious mind to the collective species unconscious and to Gaia, the collective unconscious for all species on Earth. Since time is internally structured and layered, it follows that time-like mind levels are also internested and layered.

This appears to be a significant strategic threat, *well along in its development*, that our scientific and intelligence communities have completely failed to recognize.

The direct engineering of the entire human collective species is a doable.[1] It is going to be done, whether we like it or not. In that race (which may eventually determine whether or not we retain any personal freedom at all), the "democratic" nations are woefully behind. In fact, the West has not even realized there is such a new kind of "arms race" for dictatorship and control of the "collective species mind" and we have already almost lost it. The West has not yet even developed the overall theory of psychoenergetics.

Waves and Wave Transductions

A pure longitudinal EM wave (LW) would have infinite energy and infinite speed. In practice, one can only approach this perfect LW wave, and one always has transverse residues remaining. Hence the "imperfect" longitudinal EM waves we can make in suitable equipment and techniques, are known as "undistorted progressive waves" (UPWs). The UPWs may have velocity far greater than the speed of light, or less than the speed of light. So the speed v of UPWs can vary as

$$0 < v < \infty \qquad [6]$$

The Russians also know that, when a nearly-pure longitudinal EM wave interacts with matter (it usually doesn't interact very much, and has a low reaction cross section), the mass immediately adds the phase conjugate replica to the interacting LW portion, regardless of frequency, thereby transducing the input LW interactions into scalar (time-polarized) EM interactions.

One transduces from TW to LW to Scalar EM via iterative phase conjugation. One transduces from scalar EM to LW to TW via iterative interferometry. It took the present author 20 years to discover this simple thing; we now have filed an invention disclosure on the process and embodiments for performing it, and will be following with formal patent applications. Let TW = transverse EM wave, LW = longitudinal EM wave, and TDW = time-density EM wave (another name I use for t-polarized or scalar EM waves, since the density of time-energy is what is varying). Let (I) = interfering with, PCR = phase conjugate replica wave with a coupled set of parentheses enclosing the type of wave. The rules for wave type transduction are:

$$\text{TW} + \text{PCR(TW)} \Rightarrow \text{LW} \qquad [7]$$

$$\text{LW} + \text{PCR(LW)} \Rightarrow \text{TDW} \qquad [8]$$

Going the other way, the rules are:

$$\text{TDW (I) TDW} \Rightarrow \text{LW} \qquad [9]$$

$$\text{LW (I) LW} \Rightarrow \text{TW} \qquad [10]$$

In EM smog (dense EM signals, even though very weak), there is increased nonlinear multiwave interactions, both interferometric and phase conjugative. So in such EM smog there exists an increased, non-negligible component of LW and TDW irradiation, formed by the above transduction interactions in the irradiated living cells and bodies. This latter part accounts for many of the long term effects which show up in "correlative" studies, but which do not show up in the normal limited-frequencies TW experiments produced by so-called "physical experiments."

The standard EM bioeffects manner of experimenting uses a theoretical model of simple, direct energy deposition by means of absorption in cells and tissues. Now we must utilize an "open system" throughout the body dielectric in which higher polarized EM energy (either LW or t-polarized) travels readily throughout all areas of the body.

EXPLORE! VOLUME 9, NUMBER 3, 1999

EM Bioeffects and Wave Transductions

In the EM bioeffects experiments to date, there has been absolutely zero control of the internal transductions, and there has been no consideration of them. Almost all transductions of TWs into LWs will immediately result into transduction into TDW and time-excitation charging of the cells and all their internal components. These time-excitation charges (excited states of time-charging) will then slowly decay, emitting longitudinal EM waves in the process which travel throughout the body. These emitted LWs interfere with each other and the LW processes in the body, producing low level TW "noise jamming" of the cellular processes and—most especially—of the deep cellular control system of the body. The latter system functions by means of LWs and TDWs. Long-term effects of such "noise jamming" of biochemistry reactions and the control systems of the cells and body, can lead to degenerative diseases such as arthritis, atherosclerosis, Alzheimer's disease, heart disease, cataracts, tumors, leukemia, somewhat accelerated aging, immune system weakening and depressing, etc. As the body's natural defense and regeneration ability weakens, opportunistic infections further add to the degeneration.

It was shown by Whittaker in 1903 that any scalar EM potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs. [See *Figure 9*] Each longitudinal EM wave pair in the harmonic series is a phase conjugate pair. Unknown to Whittaker, that "LW phase conjugate pair" is actually a time-polarized EM "photon" structure, containing a phase conjugate LW photon substructure. So staid old "voltage" (potential) is comprised of a harmonic series of time-polarized EM waves!

The corresponding "t-polarized photons" are actually spin-4 *supergravitons* comprised of coupled antipairs of LW "photons". Each LW photon is a spin-2 *graviton* comprised of a coupled TW photon-antiphoton pair. Each TW photon and antiphoton in the coupled pair has spin-1. The couplet, being the longitudinal "photon" or graviton, thus has spin-2. And so on.

In 1904, Whittaker showed that any EM field or wave pattern can be decomposed into two scalar potential functions. By applying Whittaker 1903 decomposition to each scalar potential function, any EM field or wave has an enormously rich internal structure, comprised of scalar potential functions in interference. Each scalar potential function is comprised of time-polarized EM waves (made of spin-4 supergravitons). Each time-polarized EM wave is further comprised of bidirectional longitudinal EM wavepairs. Each longitudinal EM wave is further comprised of phase-conjugate pairs of transverse EM waves. Each transverse EM wave is further comprised of two scalar potential functions, each of which is in turn comprised of an infinite harmonic series of time-polarized EM waves, and so on *ad infinitum* and *ad nauseum*. Obviously all EM potentials, fields, and waves have an enormous "internal structure" or internal "information content of the field".

The Cellular Control System Uses t-Polarized EM Photons and Waves

The master cellular control system (MCCS) in the body utilizes these photon-gravitons, their corresponding EM wave polarizations, and transductions between them. By the use of very powerfully condensed gravitons and supergravitons, it is able to utilize small but highly significant direct curvatures of spacetime. It does not merely send a "signal" and hope the cell itself acts upon it. Instead, it sends an "engine" which alters the local spacetime in which the cell and all its components are embedded. This "local ST curvature engine" then physically works directly upon the cell and all its parts, to alter the cell according to the specific "template" of the engine.

This is the direct application of the Wheeler's simple statement of general relativity: *Mass (trapped energy) acts upon spacetime to curve it, and curved spacetime acts upon mass to change it.*

Popp's "ordinary photon" study of this cellular control system[2] therefore will detect the coherent ordinary TW photon residues which result from the photon polarization transductions occurring from the actions of the ST curvature engines upon the cells.

Cellular division, biochemical processes, growth and growth control, cell differentiation and dedifferentiation, etc. are all controlled by the MCCS or one of its subsystems, using spacetime curvature engines. The ordinary TW electronic activity our biophysicists detect and correlate are simply the intermediate components of the actual ongoing polarization transductions.

The Cellular Control System Uses EM Time-Energy

One important subsystem of the MCCS controls the immune system's functions, including the deepest and most subtle functioning performed by all elements of the immune system. This system has been rather intensively studied, but since our biophysicists have not possessed the full electrodynamics utilized, only the correlated residue of normal TW EM waves and photons have been recognized and studied. Hence Western biophysics is in its very crude infancy.

Incorporation of the total wave and photon polarizations, together with their transductions, will eventually present a revolution in both biomedicine and biophysics. Unfortunately, presently our entire medical community is very firmly entrenched in only ordinary TW EM polarizations and ordinary biochemistry. Our scientists thus are studying less than 1% of the functioning of the living organism.

As an example, since Western scientists have no knowledge or research into the time-polarized mind domain and mind-body coupling, the mystery of the *placebo effect*—in which the mind has been verified to produce actual physical effects – will indefinitely remain a mystery.[3] As one result of that "crippling" of our medical theory, medical experimenters have to deliberately drive experiments on drug effects, etc. beyond the "mind's own threshold" of physical effect. In short, they have to drive the experimental results over and above the results being accomplished by the placebo effect. They must perform valid placebo control experiments, and subtract the placebo component results from their own experimental results, to determine the "delta" induced purely by the drugs.

The Regenerative Subsystem

Another important subsystem of the MCCS controls the *regenerative system* of the body. The regenerative system has various mechanisms, including growth control, differentiation and dedifferentiation of cells, etc. Becker's portrayal of the regenerative system is shown in *Figure 10*. One of its most important functions is healing and restoring damaged or diseased cells. It does that by a most remarkable process.

This system has been very poorly studied, mostly by Becker[4] powerfully condensed gravitons and supergravitons, it is able to utilize small but highly significant direct curvatures of spacetime in precise patterns. In short, it utilizes general relativistic spacetime curvature engines, rather than "EM photon signals", to accomplish its control of the cells and their processes. It does

does that by a most remarkable process.

This system has been very poorly studied, mostly by Becker[4] *et al.*, although there is still sporadic, ill-funded, and ill-planned research on regenerative effects from time to time.[5] Becker and his colleagues came closest, and—considering the severe limitations of the orthodox U(1) EM model available to them—Becker's work is particularly revolutionary. He demonstrated beyond question that simple DC potentials could generate cellular differentiation and redifferentiation, and produce results at global sites in the body and not just at the local injury sites where the DC potentials were applied.

Becker was nominated for a Nobel Prize for his Herculean efforts and remarkable results. For his trouble he was hounded, his funds were withdrawn, and he was forced to retire at an early age in his 50s. His biggest "heresy" was to bravely point out that power line radiations had deleterious effects upon biological systems, and to testify to it in a series of hearings and court cases where a very powerful electrical power industry essentially ran rough shod over science and the public.

Organized Science Is Highly Politicized and Imperfect

It is not only politicians that are "bought and paid for" by powerful vested financial interests. In organized science, precisely the same arrangement holds true. Scientists are no better or no worse than any other segment of humanity. A surprising percentage will produce the results desired by their benefactors, if provided with cushy jobs, good income, prestige, and secure positions.

It has been said that money is power. It has also been said that power corrupts, and absolute power corrupts absolutely. In the scientific community, as well as in the political community, this "great corruption" by great financial power has been elevated to an art form. Simply examine the funds made available to scientists. There is not enough for all the scientists, so it is fiercely competitive. If the scientist does not get grants (e.g., a would-be professor at a university), he does not bring in extra money to the university. He cannot get his graduate students funded, so his own innovative (read: heretical) research is thwarted. His papers are rejected by the journals, and before long he has no funded position in science. If lucky, he will be employed in the grocery store or the butcher shop.

Every bit of funding available for the scientists to compete for (and apply for grants) already has firmly fixed controls of exactly what research can be done with that money. The "science" that will be applied is already formulated. That and only that can be done. It is a tribute to the dedication and resourcefulness of our working scientists themselves, that anything innovative at all gets done.

One can have a very fine scientific career, tenure as a professor, get one's graduate students funded, get one's papers published in prestigious journals, win awards and achieve stature and prestige, if one just plays the desired game prescribed by the "old boys schools" in control of almost all the scientific funding. Yes, one can permissibly move the decimal point in the present models a bit, and be rewarded handsomely for doing so. But one is damned if one attempts any serious change or overthrow of obsolete scientific models and practices favored by the "system".

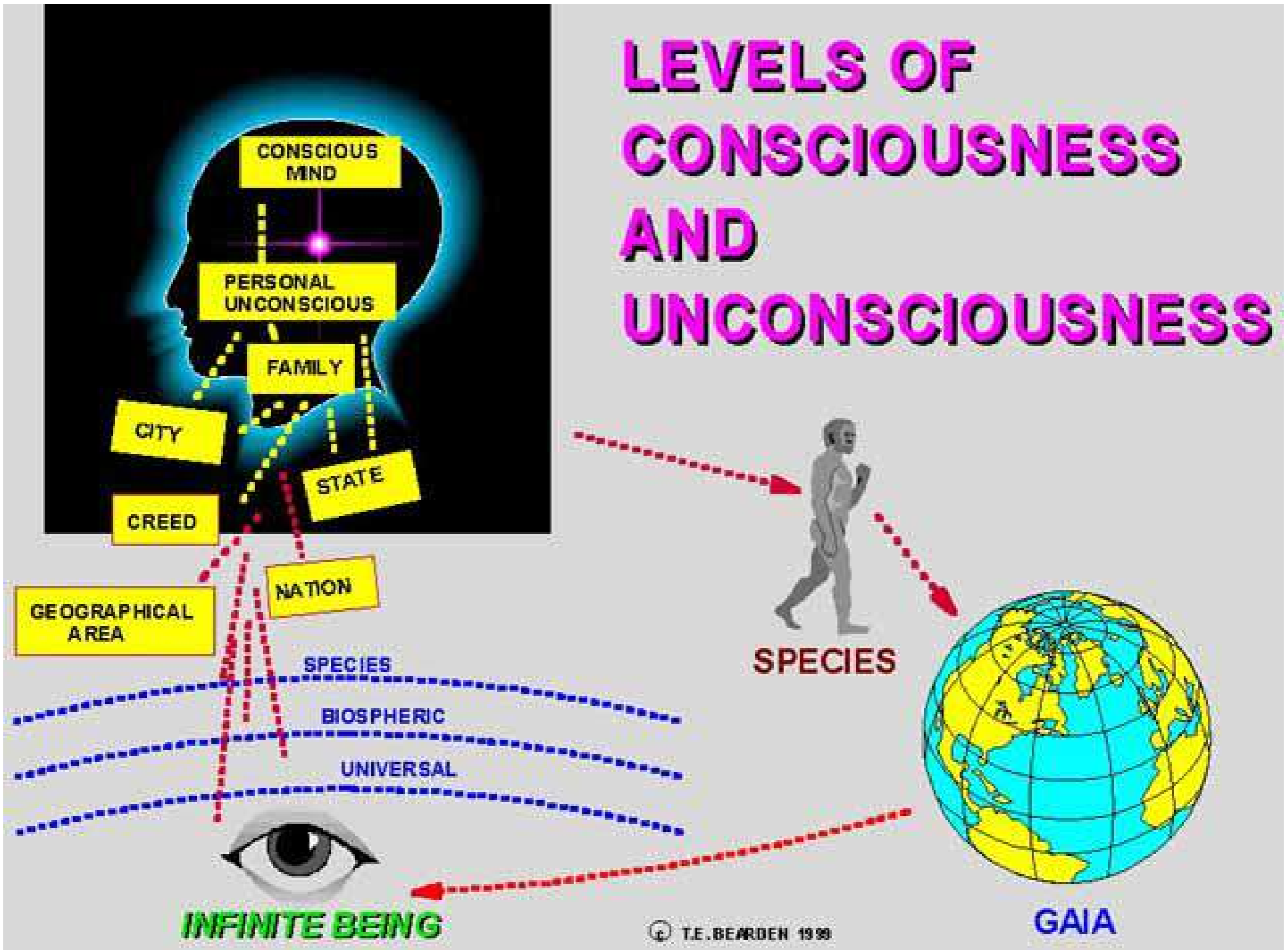

Figure 8. Nested levels of consciousness and unconsciousness.
The mind levels contain the "life" and "mind" dynamics.

EXPLORE! VOLUME 9, NUMBER 3, 1999

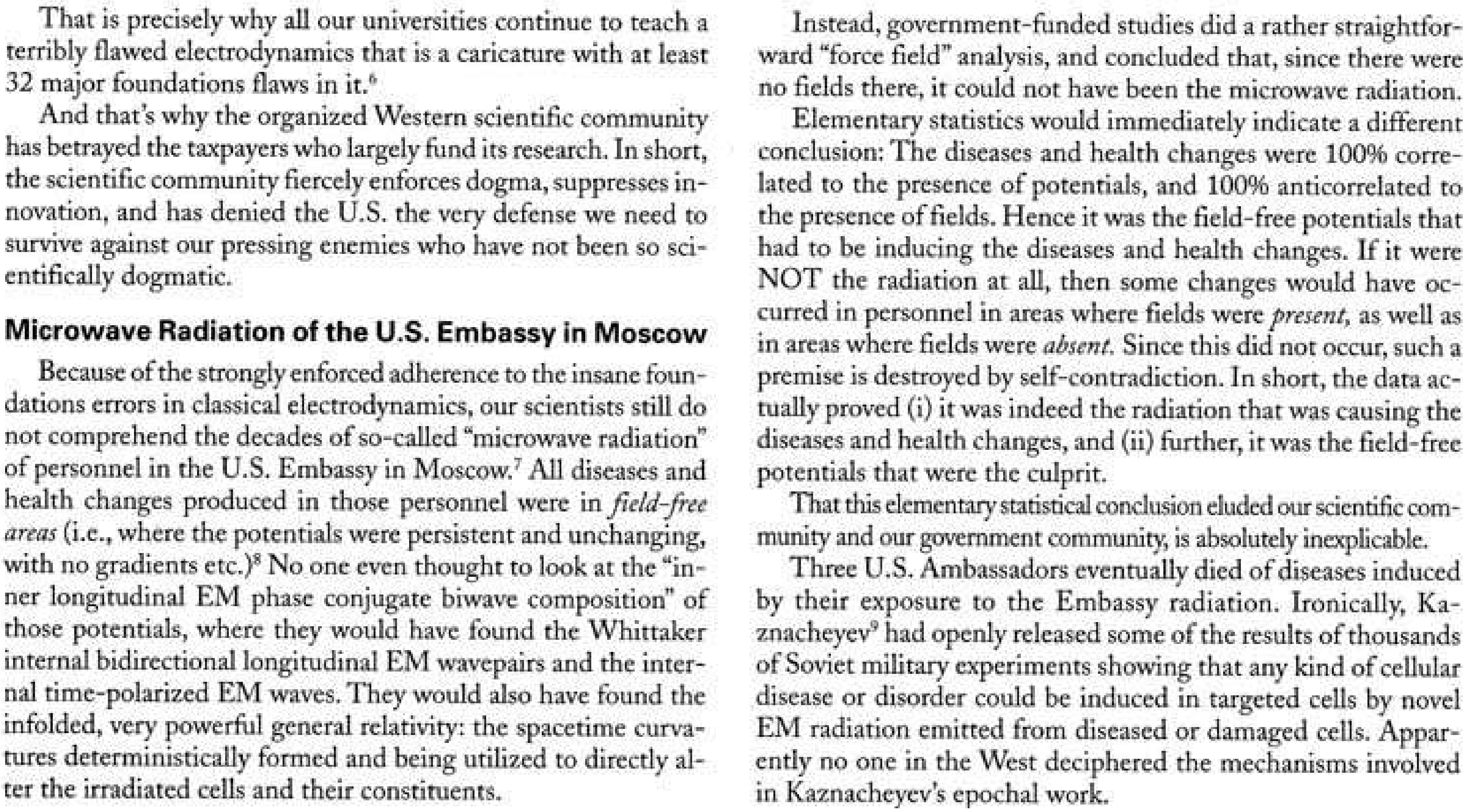

That is precisely why all our universities continue to teach a terribly flawed electrodynamics that is a caricature with at least 32 major foundations flaws in it.[6]

And that's why the organized Western scientific community has betrayed the taxpayers who largely fund its research. In short, the scientific community fiercely enforces dogma, suppresses innovation, and has denied the U.S. the very defense we need to survive against our pressing enemies who have not been so scientifically dogmatic.

Microwave Radiation of the U.S. Embassy in Moscow

Because of the strongly enforced adherence to the insane foundations errors in classical electrodynamics, our scientists still do not comprehend the decades of so-called "microwave radiation" of personnel in the U.S. Embassy in Moscow.[7] All diseases and health changes produced in those personnel were in *field-free areas* (i.e., where the potentials were persistent and unchanging, with no gradients etc.)[8] No one even thought to look at the "inner longitudinal EM phase conjugate biwave composition" of those potentials, where they would have found the Whittaker internal bidirectional longitudinal EM wavepairs and the internal time-polarized EM waves. They would also have found the infolded, very powerful general relativity: the spacetime curvatures deterministically formed and being utilized to directly alter the irradiated cells and their constituents.

Instead, government-funded studies did a rather straightforward "force field" analysis, and concluded that, since there were no fields there, it could not have been the microwave radiation.

Elementary statistics would immediately indicate a different conclusion: The diseases and health changes were 100% correlated to the presence of potentials, and 100% anticorrelated to the presence of fields. Hence it was the field-free potentials that had to be inducing the diseases and health changes. If it were NOT the radiation at all, then some changes would have occurred in personnel in areas where fields were *present*, as well as in areas where fields were *absent*. Since this did not occur, such a premise is destroyed by self-contradiction. In short, the data actually proved (i) it was indeed the radiation that was causing the diseases and health changes, and (ii) further, it was the field-free potentials that were the culprit.

That this elementary statistical conclusion eluded our scientific community and our government community, is absolutely inexplicable.

Three U.S. Ambassadors eventually died of diseases induced by their exposure to the Embassy radiation. Ironically, Kaznacheyev[9] had openly released some of the results of thousands of Soviet military experiments showing that any kind of cellular disease or disorder could be induced in targeted cells by novel EM radiation emitted from diseased or damaged cells. Apparently no one in the West deciphered the mechanisms involved in Kaznacheyev's epochal work.

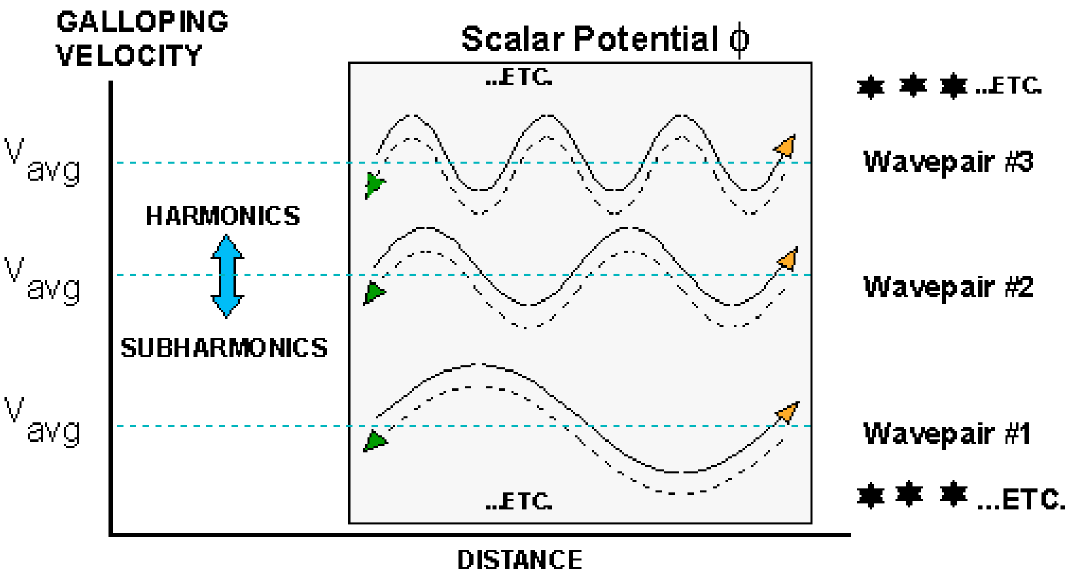

The Structure Is:

- A harmonic set of longitudinal phase conjugate wavepairs.
- In each wavepair the two waves superpose spatially after detection, but travel in opposite directions.
- The convergent wave set is in the imaginary plane, and hence is not observable. It is EM energy incoming to the potential (dipolarity) from the time domain.
- The charge's spin is 720 degrees, 360 in the real plane and 360 in the imaginary plane.
- Hence the charge receives the complex convergent EM energy, transduces it into real EM energy, and emits enormous energy at the speed of light in all directions -- which includes bidirectional pairs in 3-space (after the reaction, being after "observation").
- This produces the fields and potentials from the "source charge or dipole."
- Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus their quantum field

- **Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus their quantum field theory work strongly supports the "negative resistor" interpretation of the scalar potential and our solution to the source charge and source dipole problem.**

Figure 9. Infolded longitudinal biwaves (time-polarized EM waves) composition of a scalar potential.

EXPLORE! VOLUME 9, NUMBER 3, 1999

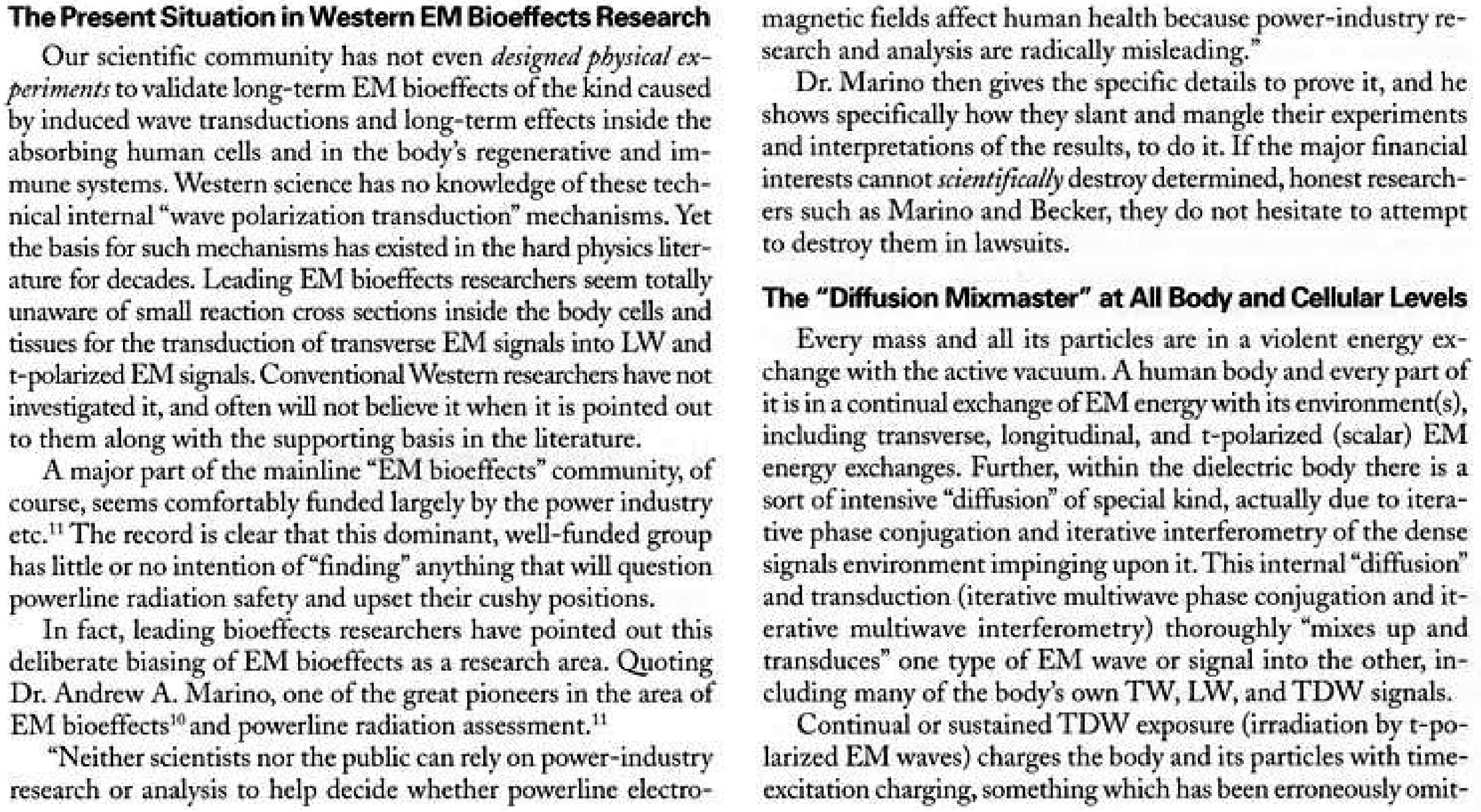

The Present Situation in Western EM Bioeffects Research

Our scientific community has not even *designed physical experiments* to validate long-term EM bioeffects of the kind caused by induced wave transductions and long-term effects inside the absorbing human cells and in the body's regenerative and immune systems. Western science has no knowledge of these technical internal "wave polarization transduction" mechanisms. Yet the basis for such mechanisms has existed in the hard physics literature for decades. Leading EM bioeffects researchers seem totally unaware of small reaction cross sections inside the body cells and tissues for the transduction of transverse EM signals into LW and t-polarized EM signals. Conventional Western researchers have not investigated it, and often will not believe it when it is pointed out to them along with the supporting basis in the literature.

A major part of the mainline "EM bioeffects" community, of course, seems comfortably funded largely by the power industry etc.[11] The record is clear that this dominant, well-funded group has little or no intention of "finding" anything that will question powerline radiation safety and upset their cushy positions.

In fact, leading bioeffects researchers have pointed out this deliberate biasing of EM bioeffects as a research area. Quoting Dr. Andrew A. Marino, one of the great pioneers in the area of EM bioeffects[10] and powerline radiation assessment.[11]

"Neither scientists nor the public can rely on power-industry research or analysis to help decide whether powerline electromagnetic fields affect human health because power-industry research and analysis are radically misleading."

Dr. Marino then gives the specific details to prove it, and he shows specifically how they slant and mangle their experiments and interpretations of the results, to do it. If the major financial interests cannot *scientifically* destroy determined, honest researchers such as Marino and Becker, they do not hesitate to attempt to destroy them in lawsuits.

The "Diffusion Mixmaster" at All Body and Cellular Levels

Every mass and all its particles are in a violent energy exchange with the active vacuum. A human body and every part of it is in a continual exchange of EM energy with its environment(s), including transverse, longitudinal, and t-polarized (scalar) EM energy exchanges. Further, within the dielectric body there is a sort of intensive "diffusion" of special kind, actually due to iterative phase conjugation and iterative interferometry of the dense signals environment impinging upon it. This internal "diffusion" and transduction (iterative multiwave phase conjugation and iterative multiwave interferometry) thoroughly "mixes up and transduces" one type of EM wave or signal into the other, including many of the body's own TW, LW, and TDW signals.

Continual or sustained TDW exposure (irradiation by t-polarized EM waves) charges the body and its particles with time-excitation charging, something which has been erroneously omit-

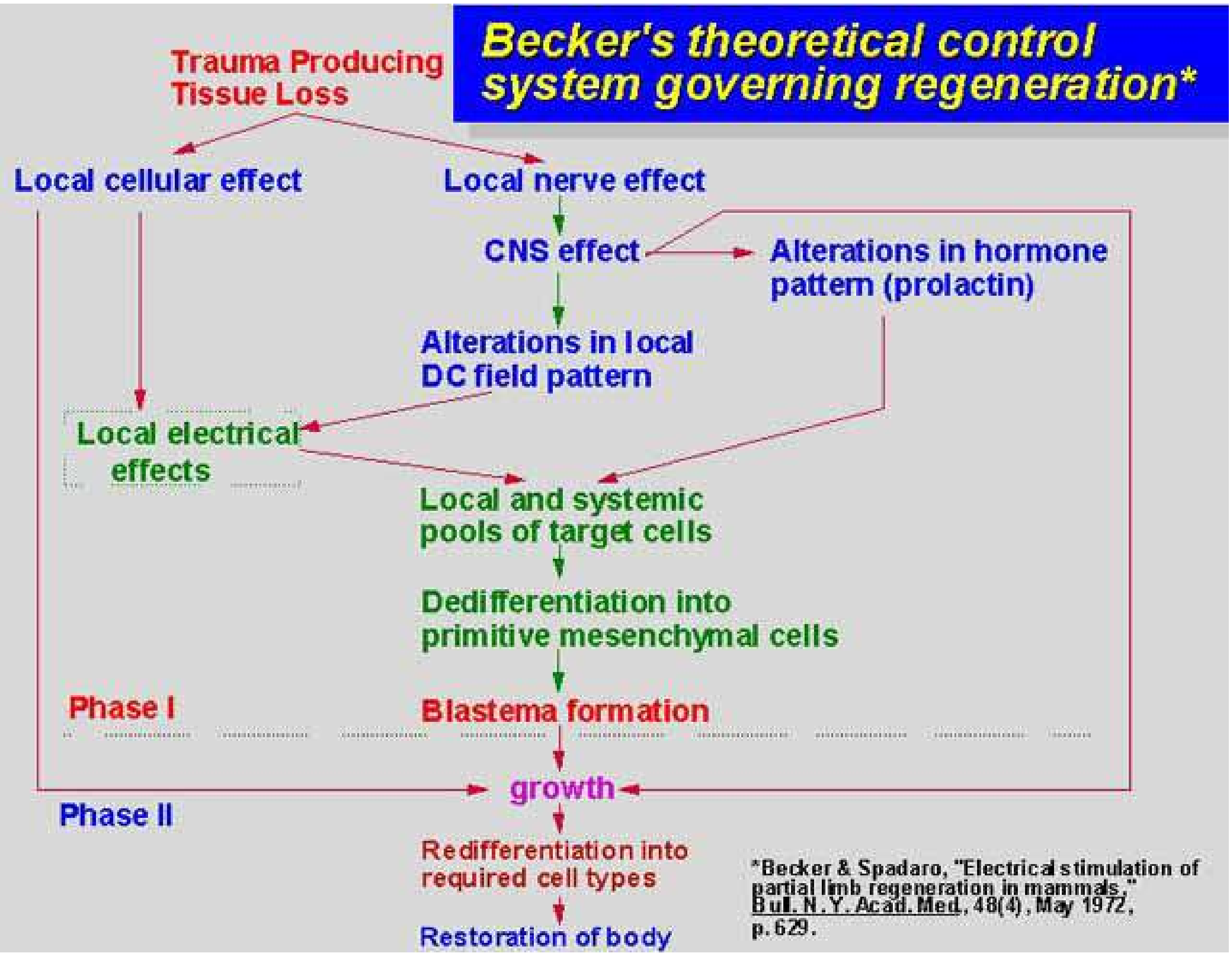

Becker's theoretical control system governing regeneration*
Trauma Producing Tissue Loss
Local cellular effect
Local nerve effect
CNS effect
Alterations in hormone pattern (prolactin)
Alterations in local DC field pattern
Local electrical effects
Local and systemic pools of target cells
Dedifferentiation into primitive mesenchymal cells
Phase I
Blastema formation
growth
Phase II
Redifferentiation into required cell types
Restoration of body

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Restoration of body part or tissue
Bull. N. Y. Acad. Med., 48(4), May 1972, p.629.

Figure 10. Becker's model of the Regenerative system of the body.

EXPLORE! VOLUME 9, NUMBER 3, 1999

ted in physics, but is true nonetheless. Inexplicably, physicists just ignore what happens time-wise when a mass absorbs a photon. We have covered that elsewhere.

So inside the body dielectric, there is a "diffusion mix" correlate of any and every type of the waves, to any or all of the exchanges, plus the additional internal interactions of the body in its TW, LW, and scalar EM wave transductions. There is a one-to-one correspondence to any one of the three types of polarization waves (transverse, longitudinal, and scalar) and everything going on anywhere in the body.

Note that the type of LW we are using is actually made of phase conjugate pairs of transverse EM waves, so its "photons" are made of photon/antiphoton pairs. That is, our longitudinal EM wave has an internal TW structure, and the LW wave is comprised of spin-2 *gravitons*. Our longitudinal EM wave has an internal structure, and it dramatically differs from the LW used by the U.S. scientific community which ignores wave, field, and potential infolded substructures. Our LW (which seems to be what the Russians use) is also a gravitational wave because it is comprised of oscillations in its distributions of spin-2 photon/antiphoton pairs (i.e., *gravitons*).

The type of t-polarized (scalar) EM wave we are using is a phase conjugate coupled pair of the specialized electrogravitational (EG) LWs. The "t-polarized photons" comprising this t-polarized "EM" wave are comprised of spin-4 *supergravitons*. These are, I believe, what the KGB energetics weapons scientists use also. By iteration, we can go into as deep an internal multilevel structuring of EM and EG waves, of any type, via methods shown by Whittaker.[12]

Now let us examine a body as a dielectric, sitting there in its exchange, in relative equilibrium. It turns that every tiniest piece of a dielectric participates in any emission from it[13]—even the emission of a single photon, graviton, or supergraviton. Well, it *would* so participate, due to the special sort of correlated diffusion and interferometry and phase conjugation that we pointed out. Anything that escapes (is re-emitted) from the body, has been through this "diffusion wringer". It has escaped from each and every internal part of the body.

Novel Signatures in Body-Emitted EM Radiation of Any Polarization

A single photon or wave emitted from a dielectric carries in its internally structured time component a *template* of the exact internal time structuring of every part of the dielectric emitting that photon. It carries in its internally structured energy component a *template* of the internal energy structuring of every part of the dielectric emitting it. An emitted EM wave, being a collection of emitted photons and photon structures, thus carries the same internal structuring of both its transported energy and its transported time.[14] Western scientists do not use the internal

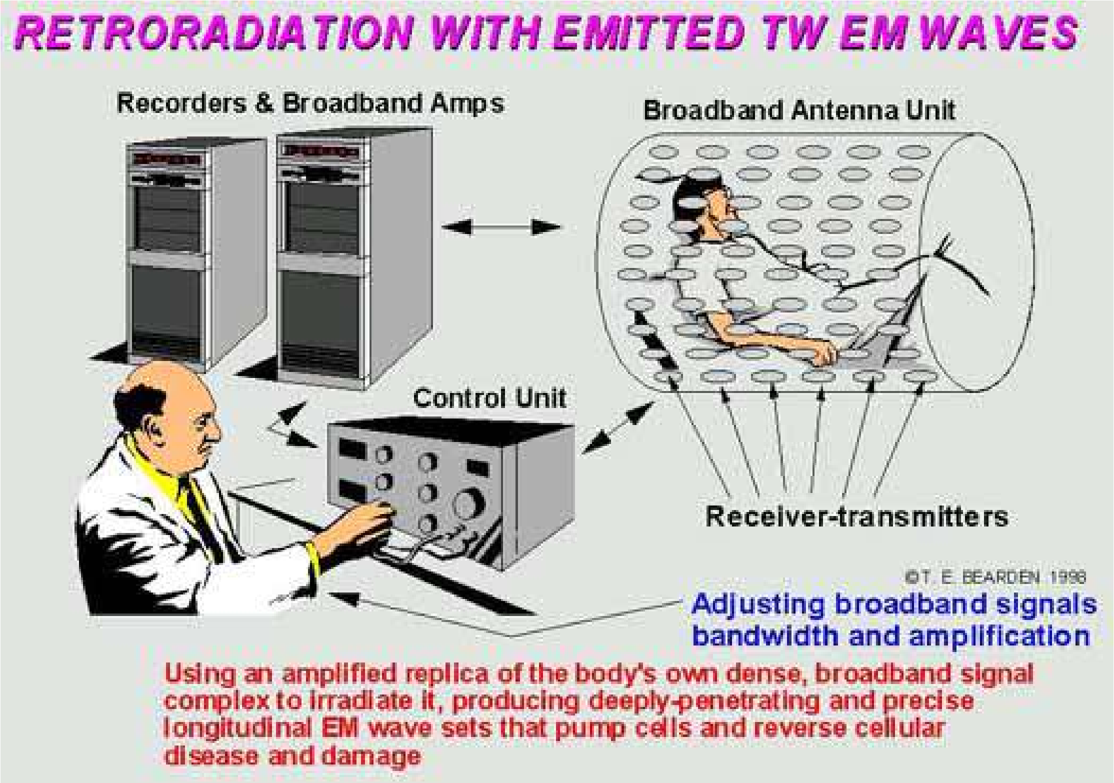
RETRORADIATION WITH EMITTED TW EM WAVES
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
©T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification
Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

Figure 11.

EXPLORE! VOLUME 9, NUMBER 3, 1999

structuring of photons and EM sine waves, have no instruments for that, and so do not know it. The Russians know it,[15,16] use it, and have developed sophisticated instruments for it for the last 30 years or more.

Shortly we will postulate how conventional scientists experimenting with mind and behavior control electromagnetically might utilize an immediate consequence of the fact that the specific emission from a body at any time—whatever type of wave component we look at—is a total correlate to all that body's internal functions, its energy exchanges with its environment, its states, its mental states and conditions, etc.

And we accent that the coherently integrated intent-changes in that body from its coupled mind's operations are also in the substructure of photons and EM waves emitted from the living body into its environment.

How Western Mind Control Scientists Might Proceed

Conventional scientists would ignore scalar photons and waves, determine the TW spectrum, and then slowly and painfully correlate that to the various physical, mental, and emotional things. That's the way they would *have* to do it, since apparently they use only standard electrodynamics. They can in fact do this via brute force, because that emission TW spectrum is of waves escaping from the entire "diffusion machinery" mixmaster in the body dielectric. It does contain all those integrated *mind-intent* 3-space correlates as well as the ordinary body 3-space energy correlates.

Anyhow, in envisioning Western procedures, the test body would be sitting there, in its normal (let's say fairly normal and quiet) EM environment. [See *Figure 11*] Let's assume this body is not sick, but is healthy and normal. Its emissions are a part of its equilibrium condition. It follows that the totality of its TW emissions is correlated to the sum total of everything that interacted with the body, plus its own internal reactions, including from mind-body coupling. Indeed, everything that has ever happened to that entity—physically, mentally, emotionally, etc.—has internal correlates infolded inside the emitted transverse EM wave spectrum.[17]

So first the scientists would simply measure the full environmental TW spectrum down to very tiny levels in all frequency bands possible—everything that is coming in to the body from its environment. They would probably perform the experiments in a triply-shielded Faraday cage, so the external environment outside the cage contributes very little transverse EM wave noise (perhaps some Schumann resonance). Magnetic wave shielding and ELF shielding is something else again, and longitudinal EM wave shielding is in general not possible unless longitudinal EM waves are employed in the shielding. But a "quiet external TW environment" enhances the isolation and measurement of the specific TW signal transforms the Western experimenters would be seeking.

Then they would measure everything likewise that the body is *emitting* when in that environment. That is the "zero reference" or "normal reference" level. They would measure every aspect: frequency, amplitude, phase, wave shape, modulation, etc.

Probable Experimental and Measurement Techniques

Next they would selectively irradiate the body with specific signals, frequencies, waveforms, etc. They would begin by using

They would consider the body dielectric as a "ship with portholes". Light (EM transverse wave energy) coming out of the portholes has originated and adapted from everywhere inside the ship and its operations. When the experimenters reinsert energy through one or more portholes, they are interested in determining precisely which operations inside the ship are affected. So the body physiological functions are also heavily instrumented.

By feeding back in the exact EM pattern and magnitude escaping the body dielectric through a single porthole, one "nullifies" that part of the emission by producing a net equilibrium in that component. This will in turn produce changes elsewhere in the spectrum emission from other portholes, and this can be measured. It will also produce "stress changes" in specific portions of the body dielectric, and in the heavily instrumented body these physiological, chemical, electrical, etc. changes can also be detected. Monitoring of the skin conductance, e.g., indicates the level of stress, and EEG monitoring with sophisticated instrumentation can indicate something of the brain wave states.

In short, the necessary measurements from which to construct specific correlates can be performed, recorded, and later analyzed.

This article will be concluded in Volume 9, #4 with Improving the Methodology, Further Phenomenology Experiments, Rejuvenation and other matters. ◆

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel U.S. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, U.S. Army. Graduate of Guided Missile Staff Officer's Course, U.S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses.

Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor. He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuum's virtual particle flux is organized and used to energetically shape and manipulate matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

1. *In fact, it simply applies in the laboratory that method that is already used by species for the so-called "directed evolution" component of evolution. The exact mechanisms for directed evolution also arise naturally in the new approach. Directed evolution is required; e.g., no reptilian species could have evolved slowly into a bird. As it started slow changes, it would have not been able to fly or to run, and would have been simply destroyed immediately by*

Next they would selectively irradiate the body with specific signals, frequencies, waveforms, etc. They would begin by using portions of the same frequencies that are being emitted from the body in the "zero reference" environment. They would insure that the returned signal was of higher amplitude than the usual signal being emitted by the body at that same frequency and type.

reptilian species could have evolved slowly into a bird. As it started slow changes, it would have not been able to fly or to run, and would have been simply destroyed immediately by predators. Instead, such evolution occurs in a single jump in a species hard-pressed for a lengthy period. In the evolution of birds, one day every reptile egg that hatched, now hatched into a complete and functional bird with light bones, feathers, etc. A continuing debate (with perhaps more emotion than science) continues to this day on directed evolution. But it is real, and the mechanism can readily be generated (we've already done the gist of it).

EXPLORE! VOLUME 9, NUMBER 3, 1999

2. *Fritz Albert Popp, "Photon Storage in Biological Systems," in Fritz Albert* Popp et al, *eds.*, Electromagnetic Bio-Information: Proceedings of the Symposium, Marburg, Federal Republic of Germany, Sep. 5, 1977; *Urban & Schwarzenberg, Baltimore, 1979, p. 123-149;—"On the coherence of ultraweak photoemission from living tissues,"* in Disequilibrium and Self-Organization, *ed. by C. W. Kilmister, D. Reidel, Hingham, MA, 1986;*—et al., *"Physical aspects of biophotons,"* Experientia, *Vol. 44, 1988, p. 576-585;—"Biophoton emission: New evidence for coherence and DNA as source,"* Cell Biophys., *Vol. 6, 1984, p. 33-52. See particularly Fritz Albert Popp, Ke-hsueh Li, and Qiao Gu, Eds.,* Recent Advances in Biophoton Research and its Applications, *World Scientific Publishing Co., New York, NY, 1992.*
3. *With novel use of EM time-energy to curve spacetime, the physical actions of the "placebo effect" are due to the use of such "spacetime curvature engines" to produce actual physical effects—at least temporarily —in the cells and throughout the body's chemistry, physiology, etc. Experimental detection of these physical effects are well-established in the orthodox medical literature.*
4. *For example, see particularly R. O. Becker, "The direct current field: A primitive control and communication system related to growth processes,"* Proceedings of the. XVI International. Congress of Zoology, *Washington, D.C., Vol. 3, 1963, p. 179-183;—"The direct current control system: A link between environment and organism,"* New York State Journal of Medicine, *Vol. 62, 1962, p. 1169-1176;—and D. G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents,"* Trans. N.Y. Acad. Sci., *Vol. 29, 1967, p. 606-615;—and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals,"* Bulletin of the New York Academy of Medicine, *Second Series, 48(4), May 1972, p. 627-64.*
5. *E.g., a popular discussion is given by John Travis, "Starting Over: Some Animals can regenerate limbs or even most of their bodies. How?",* Science News, *152(18), Nov. 1, 1997, p. 280-282. For a typical technical paper, see Jeremy P. Brockes, "Amphibian limb regeneration: Rebuilding a complex structure,"* Science, *276(5309), Apr. 4, 1997, p. 81-87.*
6. *I have one viewgraph which lists those 32 major flaws in electrodynamics. Further, leading theorists (Evans, Crowell, and others) in the Alpha Foundation's Institute for Advanced Study (AIAS)—of which I am a Fellow Emeritus by AIAS's kind tolerance—have added about another very technical and very fundamental dozen or so errors to my own list. A leading foundations scientist with over 500 papers in the literature commented dryly to the AIAS that "I broadly agree. that the transverse plane wave view of vacuum electrodynamics is the biggest blunder of twentieth century science. Faraday visualized that "EM disturbances" surrounding wires, magnets, etc. were simply the transverse plucking of taut strings. As he states in his book, Maxwell—who also firmly believed in the material ether—simply captured Faraday's thinking and assumed that transverse "plucked string" wave disturbance in the material ether. At the time all this was expressed, the electron had not been discovered, the atom had not been discovered, the molecule was a featureless blob, and no one had the slightest notion of an atomic nucleus filled with protons and neutrons. Charge was simply a "certain volume of electric fluid" and everything electromagnetic was expressed in fluid dynamics. As an example, that "material fluid" and "material ether" theory was destroyed in the late 1880s by the Michelson-Morley experiment. All the Maxwell equations assume that material ether to this day; not a single equation was ever changed to eliminate the material ether assumptions during the last 110 years.*
7. *For full explanations of the "Moscow Signals," see T. E. Bearden, .,* Gravitobiology: A New Biophysics, *Tesla Book Co., P.O. Box 121873, Chula Vista, CA 91912, 1991;*—AIDS: Biological Warfare, *Tesla Book Co., 1988. For an excellent Government study of the Embassy radiation by excellent but orthodox Johns Hopkins scientists, see A. M.* Lilienfeld et al., Foreign Service Health Status Study: Evaluation of Health Status of Foreign Service and Other Employees from Selected Eastern European Posts, *Department of Epidemiology, of Hygiene and Public Health, The Johns Hopkins University, Baltimore, Maryland, Final Report, July 31, 1978. On p. 245, it is noted that a marked excess of symptoms were noted in the Moscow employee group as compared to other East European employees in other East European nations.*
8. *Johns Hopkins researchers established the exact field patterns present in the Embassy from the microwave carrier radiation; see Robert C. Mallaliew,* A Model of the Microwave Intensity Distribution within the U.S. Embassy, *Moscow 1966 to 1977, Report FS-80-166, Applied Physics Laboratory, Johns Hopkins University, Aug. 1980. Nowhere in any of the Government-funded studies is there any consideration of field-free scalar potentials and their internal time-polarized EM wave structures (the bidirectional phase conjugate wavepairs shown by Whittaker to comprise the scalar potentials). Yet the field radiation patterns surprisingly showed that in just such field-free, persistent potential zones was where 100% of the health changes occurred.*

issue, see "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144. See also Vlail Kaznacheyev, "Apparent information transfer between two groups of cells," Psychoenergetic Systems, *1(1), Dec. 1974. The Soviet Union awarded Kaznacheyev a very high medal for his achievements.*

10. *For a very nice summary of EM bioeffects from the more conventional view, see Robert O. Becker and Andrew A. Marino,* Electromagnetism and Life, *State University of New York Press, Albany, 1982. The book is also available at http://www.ortho.lsumc.edu/Faculty/Marino/El/ElPub.html.*
11. *Andrew A. Marino,* Powerline Electromagnetic Fields and Human Health, *http://www.ortho.lsumc.edu/Faculty/Marino/Marino.html. This book is highly recommended. Particularly see "Chapter 5, Blue-Ribbon Committees and Powerline EMF Health Hazards," and "Chapter 6: Power-Industry Science and Powerline EMF Health Hazards." Biophysicist Marino is one of the leaders in the field and has been personally involved in many skirmishes with powerline-dominated studies and findings.*
12. *E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics,"* Mathematische Annalen, *Vol. 57, 1903, p. 333-355; — "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* Proc. Lond. Math. Soc., *Series 2, Vol. 1, 1904, p. 367-372.*
13. *Reali, G. C., "Reflection from dielectric materials,"* American Journal of Physics, *50(12), Dec. 1982, p. 1133-1136. The reflected field from a dielectric material is not generated just at its surface but comes from everywhere in the interior of it.*
14. *Present physics models the EM wave as transporting energy, but erroneously omits that it also transports time. A photon is made of angular momentum, or "energy x time". Since EM waves are made of photons, it follows that they transport both the energy components and the time components making up those photons. An EM wave is a wave of angular momentum, not just a wave of energy.*
15. *KGB energetics weapons scientists call this internal structuring the "information content of the field." However, the Russians use a good deception plan, which implies that the "information content of the field" is just the normal transverse EM spectral content of EM signals. Appreciable "pertinent" work using that phrase is produced by the open Russian scientific community, thus confusing the issue. Our electrodynamicists seem to buy it hook, line, and sinker. The deception plan also tricks Western scientists into thinking that psychoenergetics is nothing but a Russian version of Western parapsychology. Again, our scientists comfortably accept this with smug satisfaction.*
16. *Presently the Russians are in a great new version of this deception plan, with the goal of deceiving Western scientists into thinking that the "secret" Russian energetics is based on torsion fields. Consequently they are releasing substantial material on torsion wave theory, including replicable experimental results. The released version of torsion fields has one glaring omission: the time components and structuring. This deception plan is excellent, because the torsion field theory that is released, is indeed "real" and testable. It is also revolutionary, with respect to present Western science. It almost positively assures diverting the attention of much of the Western science community into this highly limited subset of the full Russian torsion theory—effectively diverting the West away from psychoenergetics and the direct use of t-polarized EM waves and fields, torsion or otherwise.*
17. *We point out that E. T. Whittaker, in "On the Partial Differential Equations of Mathematical Physics,"* Mathematische Annalen, *Vol. 57, 1903, p. 333-355, showed that every scalar potential is comprised of an internal structure of bidirectional longitudinal EM phase conjugate wavepairs in an infinite set of harmonics. Each of those phase conjugate longitudinal EM wavepairs comprises a time-polarized EM wave. So the common potential contains an internal longitudinal EM wave structure that is further organized into an internal time-polarized EM wave structure. The mind-intent correlates and the total mind and emotional experience correlates from fertilization of the egg that led to development of that person, is in the internal substructure of every potential in the living body. In his "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* Proc. Lond. Math. Soc., *Series 2, Vol. 1, 1904, p. 367-372, Whittaker also showed that any EM field or wave is comprised of two scalar potential functions. It follows from Whittaker and Reali that the EM emissions from the living body carry the internal correlates of everything that has been experienced by that person from conception in the womb. Sometimes this deep embedded "hidden information content" can even be accessed in deep hypnosis. I once observed the deep hypnotic regression of a subject back to her prenatal experiences prior to birth, where the subject relived hearing words of rejection by the father, before the subject was born. The occurrence of the incident and the father's words were directly confirmed on the spot by the mother, who was closely observing the session as it occurred. It turned out that this particular rejection of the baby still in the womb had played a major role in subsequent deep psychological*

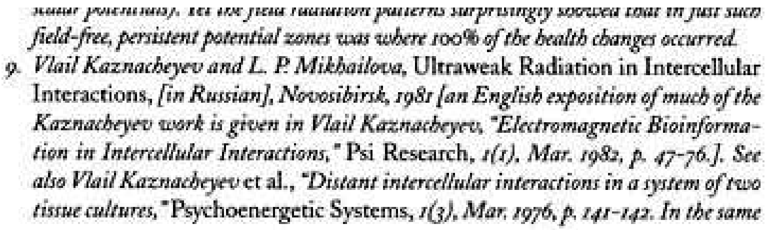

scalar potentials). Yet the field radiation patterns surprisingly showed that in just such field-free, persistent potential zones was where 100% of the health changes occurred.

9. *Vlail Kaznacheyev and L. P. Mikhailova,* Ultraweak Radiation in Intercellular Interactions, *[in Russian], Novosibirsk, 1981 [an English exposition of much of the Kaznacheyev work is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions,"* Psi Research, *1(1), Mar. 1982, p. 47-76.]. See also Vlail Kaznacheyev* et al., *"Distant intercellular interactions in a system of two tissue cultures,"* Psychoenergetic Systems, *1(3), Mar. 1976, p. 141-142. In the same*

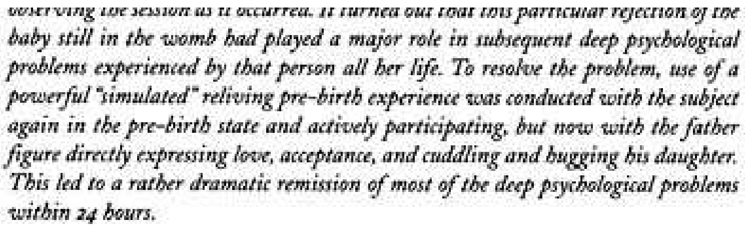

observing the session as it occurred. It turned out that this particular rejection of the baby still in the womb had played a major role in subsequent deep psychological problems experienced by that person all her life. To resolve the problem, use of a powerful "simulated" reliving pre-birth experience was conducted with the subject again in the pre-birth state and actively participating, but now with the father figure directly expressing love, acceptance, and cuddling and hugging his daughter. This led to a rather dramatic remission of most of the deep psychological problems within 24 hours.

EXPLORE! VOLUME 9, NUMBER 3, 1999

Mind Control and EM Wave Polarization Transductions, Part 3 of 3

©Copyright 1999 by T.E. Bearden, USA

This outstanding article is the last, Part 3, Part I and 2 of which were published in Volume 9, #2 and #3 of Explore! for the Professional. Because we feel that Col. Bearden’s Special Note below is so very important we are repeating its publication in all 3 parts of this article.

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation

achieving many more effects in the targeted bio-organism or groups of them. With Hunt's method, the system as an *open dissipative system far from thermodynamic equilibrium in its environment* is being measured as such. This more directly allows for determination of the environmental inputs on a much broader range of subsystems, both space-like and time-like. Particularly for mental and emotional correlates to environmental stimuli, such measurements and correlations are essential.

More Comprehensive Phenomenology Experiments

To return: Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.

In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion.

of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

Two Ways to Improve the Methodology and Experimental Correlates

We diverge momentarily: If they wish to improve this methodology, they will consider the body dielectric as a nonlinear isotropic medium. Then they will pretend that the spectrum of *difference frequencies* between each two adjacent waves is the spectrum of "emissions" from the body. They would also use those, as if they were actual emitted spectral signals, instead of the actual signals emitted. That procedure will give superior correlations, but it is more complicated. Part of that is known for underwater sonar,[1] and so our own scientists may or may not have realized it for use in mind and behavior correlate determination with respect to the body dielectric.

Another dramatic improvement in the depth to which the correlates can be constructed, is provided by detecting each porthole output in a two-channel device, where one channel has an adjustable time-delay as utilized by Dr. Valerie Hunt.[2] The two channels—one real-time and one slightly delayed—are then mixed, as by Hunt's method, producing an instrumental measurement of the chaotic part of the functions. Again, this can be measured, recorded, and later analyzed. It adds (i) a completely different kind of thermodynamics in the system measurements, (ii) a dramatically extended set of correlates, (iii) more direct measurement of the time-polarization causative signals, and (iv) an applications Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions.

A particularly vulnerable aspect of every mammal is the pleasure center in the brain. Experiments have shown that, when this center is stimulated, it is the most addictive experience possible. Rats will forego food, endure electrical shocks, starve, and even die to obtain stimulation of this center. Obviously a device capable of generating signals that evoke direct stimulation of the pleasure center in humans would be a powerful weapon.[3]

Extending the Data Base

See Figure 12. After the researchers have made an "individual porthole signal insertion" correlate database, then they would make a "two, three, four," porthole signal insertion correlate database by inducing selected multiple changes at once. Again they would analyze everything; the program would need one or more supercomputers. It would also require quite a highly qualified multidisciplinary team, extensive facilities, substantial funding, and would be many years in duration. We are describing a decades-long program. But it's a doable.

With sufficient development and performance of such a program, the researchers would have produced (with sufficient testing and analysis, and sufficient retest validation and verification) a database of "specific correlates for a given overall technology eventually having many additional degrees of freedom and therefore capable of

desired physical, mental, or physical and

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

mental delta."

A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment.

The Dielectric Body as a Ship With Portholes

Again let us review what the EM transverse wave irradiation from the body dielectric represents. Think of the body dielectric as a sort of "ship", with lots of portholes. When it gets irradiated, certain operations happen internally to that radiation, i.e., certain "processes" and "interactions" occur, which change both the mental and physical states. The output emissions back from the body dielectric, out through the filtering "portholes", will directly correlate to what has happened inside the ship.

It is terribly important to realize that one can use the same frequencies and spectra coming back out of those "portholes" to insert EM signals and waves etc. back in there, and into the deepest processes going on in the body and in the mind (at all levels, times. So it will reach directly into and affect any and every part of the dielectric, as you wish, and in whatever manner you have correlates for.

We filed an invention disclosure[4] on the above correlate and "back-porthole" process for use in directly inducing full and efficient conversion directly to time-density (scalar, or time-polarized) EM waves inside the body/ cells, *therefore immediately and directly time-excitation charging the body cells and all their internal components.*

Subtracting the Normal Correlates to Obtain the Deltas

A specifically diseased body, e.g., has a specifically tailored change in its "normal" emission spectra, because of the disease condition alteration of its resident spacetime curvature engine. By having a database for the normal emission spectra of the healthy body, one can subtract it from the actual emitted spectra, to obtain the "delta spectrum".

If the correlates have been previously determined by the above mentioned research program, then one may simply introduce an amplified "delta spectrum" into the body, i.e., through the portholes. This will immediately time-excite the cells and their components throughout the body, particularly where that disease does exist. It turns out that the time-excitation charging by that method will "pump" those diseased including the deep unconscious).

Remember, whatever you put back in there, is going back through the "mixmaster" hopper. It will be iteratively phase conjugated and interfered with by everything, countless cells in the time domain. The cells themselves (and all their parts) are highly nonlinear—to time-excitation charging, they are *totally* nonlinear. So they will act as rather perfect pumped phase conjugate mirrors, but pumped in the time-energy domain rather than in the spatial

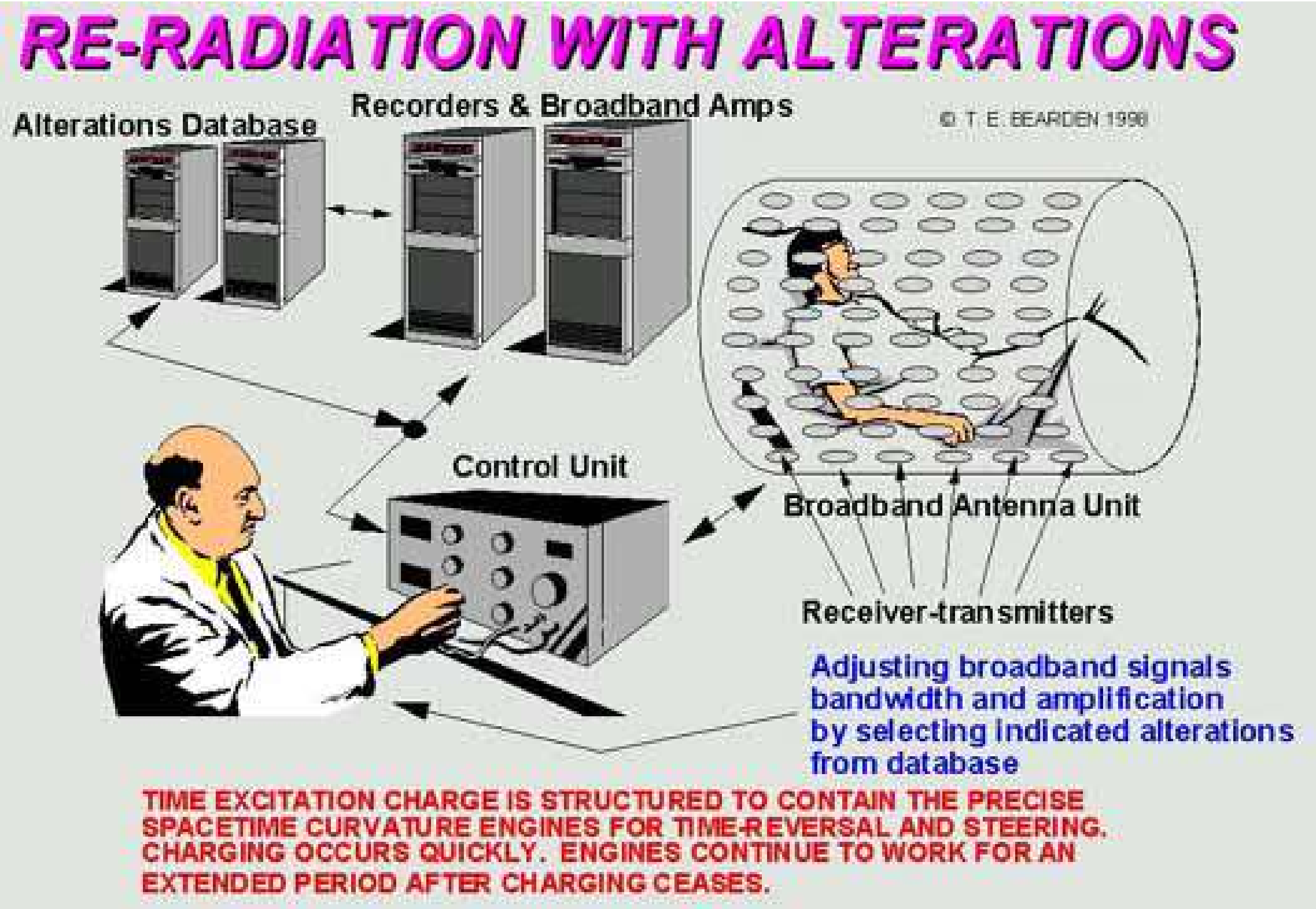
RE-RADIATION WITH ALTERATIONS
Alterations Database
Recorders & Broadband Amps
© T. E. BEARDEN 1998
Control Unit
Broadband Antenna Unit
Receiver-transmitters
Adjusting broadband signals bandwidth and amplification by selecting indicated alterations from database
TIME EXCITATION CHARGE IS STRUCTURED TO CONTAIN THE PRECISE SPACETIME CURVATURE ENGINES FOR TIME-REVERSAL AND STEERING. CHARGING OCCURS QUICKLY. ENGINES CONTINUE TO WORK FOR AN EXTENDED PERIOD AFTER CHARGING CEASES.

Figure 12. Using an amplified replica of the body's own dense, broadband signal complex, containing added alterations, to irradiate the body and produce deeply-penetrating precise time-density charging for cellular time-reversal while steering the time-reversal trajectory to also correct a congenital condition.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

energy domain. This is a dramatic extension to nonlinear phase conjugate optics theory.[5]

General relativity tells us that each specific detail of the mass and energy structure of each cell and all its parts—down to the finest details—is accompanied by a precise set of spacetime curvatures. By reintroducing the same energy as the disease delta TW waves, the affected cells add the phase conjugates, converting the inserted TW signals into longitudinal EM wave signals. The longitudinally-pumped cells again phase conjugate, converting these longitudinal EM waves to time-polarized EM waves, and "charging" the cells and all their parts in the time domain.

Suddenly, at that point the *resident spacetime curvature engine*—rather than TW or LW EM waves-serves as the "input signal" to the cell-and-its-parts as pumped phase conjugate mirrors (PCMs), pumped in the time domain. An extension to the distortion correction theorem of nonlinear optics results. An exact, amplified antiengine for the diseased cells—i.e., a phase conjugate replica of the resident spacetime curvature engine—is formed. This amplified antiengine overpowers the resident disease engine, powerfully time-reversing the cells and their components back to a previous healthy state. The time-reversal actually occurs over a period of time after the "time excitation charge-up" ceases, say over the next week or two. Usually two to three quick irradiations are required, one week apart, to give a time-reversing period of three to four weeks. force upon the cells or any part. E.g., as Pepper[6] states,

"On a more fundamental level, the ideal lossless PCM reverses all the quantum numbers of the incident photon (i. e., linear momentum, angular momentum, etc.). It can be shown that the PCM therefore experiences no linear or angular momentum transfer from the incident photons; hence, the PCM is free from photon radiation pressure and torques."

Note that, in ordinary EM theory, it is precisely translation forces that are engendered upon charged mass by EM fields. This process is quite different. In biological terms, this process dedifferentiates (in physics: time-reverses) the diseased cells and all their components—including the genes—back to an earlier healthy state.

By taking a diseased body and feeding back into it an exact amplified replica of its emitted TW spectrum, one can force the body to "time-excitation charge" all cells and their components immediately, with high efficiency. The human body and every cell and every part of every cell does the necessary "calculation" for the exact antiengine for each part. Since time is extremely condensed spatial EM energy (by a factor of almost 10^{17}), a startling contraction occurs in the radiation time required to time-charge the cells and their components and charged particles. This is a method for instant initiation of time-charging, hence it is extremely rapid. What took Priore several hours of LW irradiation to accomplish in the body,[7] can now be done

Even though appreciable forces are operating internally upon the cells and all their parts, there is no net translation

in seconds with this "thwarting" and spectrum transduction method. In fact, it is an intense stress on the body, and—say—should only

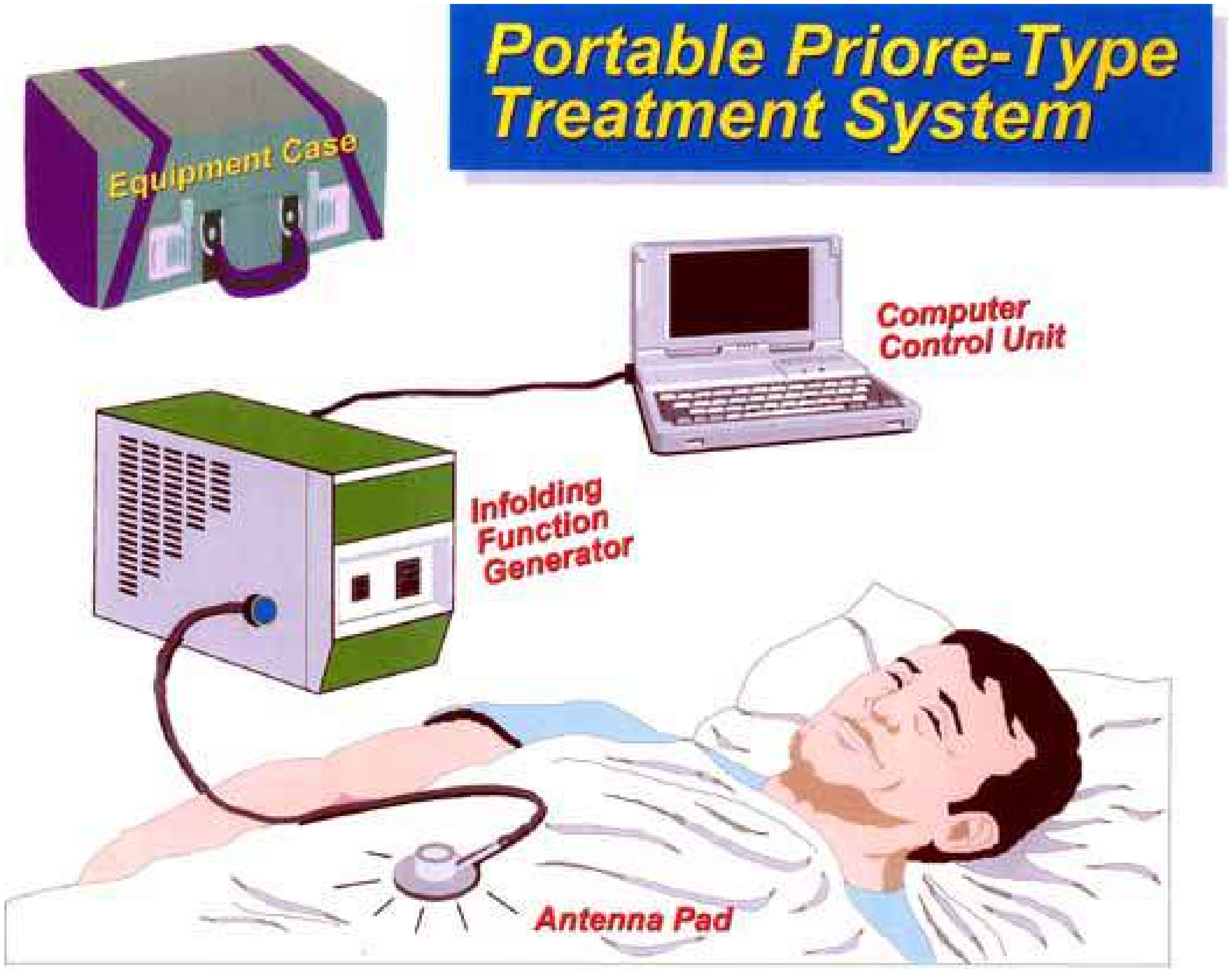
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Figure 13. Portable first-method treatment unit proposed to the Department of Defense. For treating and curing mass casualties from terrorist BW attacks on U.S. population centers.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

be continued for 30 seconds or so. We highly stress that overcharging in this time domain can be detrimental or even lethal if prolonged.

At the end of the very short time-excitation charging time, the body is "time-excitation charged in all its minutest particles". So everything is therefore pumped in the time domain for the next several days after the irradiation is removed, since the time-charge decays very gradually over a period of time. During that decay period, every part of the cell and body acts as a pumped phase conjugate mirror material, pumped in the time domain rather than in the 3-space energy domain. When time-pumped, the mass itself reverses to a previous physical state. This can be used for both nonliving and living masses. Because there is no forcible translation manipulation of the cells and their parts, it can be utilized on living cells in vivo. In the case of diseased cells—e.g., AIDS-it appears that three precisely determined 30-second "AIDS correlate" irradiations a week apart, should be sufficient to completely cure AIDS, including removal of all HIV-genetics changes from every previously infected cell of the body. Also, it would accomplish considerable rejuvenation of the body (making the body younger).

Rejuvenation: Restore the Telomeres and Produce Immortal Cells

Saving Most of the Casualties from Mass Terrorist Attacks

To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S. Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines could be developed cheaply and quickly, and present government facilities could be used to massively attack the correlate database production problem. The portable treatment units could be cranked out by the hundreds of thousands and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals, emergency response teams, etc.

Each device could be used in "assembly line" fashion upon hundreds of sickened patients, in case of a WMD (weapons of mass destruction) attack upon one or more of our cities. The first generation equipment would save probably 70% of the stricken (as from anthrax, etc.). With second generation equipment, one would expect 90% of the stricken casualties could be saved.

Just now, there is a very dim prognosis for all those casualties, using present available treatment capabilities. Time and lack of vaccines defeats just about everything that can be done. Brutally speaking, most of those unfortunate enough to breath in

In fact, it is now known that the gradual loss of the telomeres—the natural ends of chromosomes—in a cell's chromosomes is what causes the cell to age. The gradual loss of the telomeres and the resulting shortening of the chromosomes represents a cellular "disease state" or "disorder" state" accompanied by a precise delta in the spacetime curvature engine accompanying every affected (aging) cell.

By determining the correlate for that precise telomere loss in the aging body's emitted radiation spectrum, an effective telomere regeneration process immediately emerges. One simply amplifies that aging "precise telomere reduction" correlate and reradiates the body with it, passing an amplified correlate back into every cell of the body. The cells and all their parts—including the chromosomes—add *the phase conjugate replicas*, converting the incoming TW EM correlates to longitudinal EM wave correlates. The cells again phase conjugate the LWs and add *the phase conjugate replicas*, converting the LW correlates to time-polarized EM wave pumping correlates. This latter process *time-excitation-charges* the aging cells and all their parts, including the genes and chromosomes. The cellular matter itself produces the amplified "aging antiengine" for each and every cell and telomere ending. The result is to again add telomeres, converting the cell back to a younger, more vigorous cell, and reversing the process of aging.

We point out without further elaboration that these methods also introduce direct engineering signals into the body's master cellular control system, including the regenerative subsystem. Hence the body's regenerative subsystem is involved in helping produce the necessary engines and anthrax spores from an attack, are going to die. At least 80% of them will expire, in spite of everything the present system can do for them.

Note each device could rapidly treat a whole series patients, one right after the other, merely laying them between "blankets" containing wire antennas, irradiating with the proper correlate spectrum quickly, moving on to the next patient, irradiating quickly (seconds), etc. The entire apparatus could be made to fit in a large suitcase-sized container, and it would be highly portable and controlled by a laptop-sized computer.

Unfortunately, so far the DOD doesn't even understand what the heck we're talking about! The National Institute of Health (NIH) and the National Science Foundation (NSF) are not interested at all. After all, it isn't drugs, vaccines, antibiotics, etc. —which are simply not going to do the job. However, those are what the government and scientific communities are going to finance to the tune of hundreds of millions of dollars. They are then going to lose almost all those mass casualties when the terrorist attacks on our population centers inevitably occur.

Those are Americans. They are going to die. Most of them could be saved. We ought to save them.

For any change in state (physical, chemical, emotional, mental, perceptual), the body dielectric "portholes" change and the output spectrum changes. The delta from zero reference now represents that exact change.

Where Western Mind Control Research Has Probably Arrived

Western mind control researchers

operate upon the short telomere chains to add back telomeres. More specifically, the regenerative subsystem is changed into perfect consonance with the external introduction of the amplified antiengines.

That exact process can be used to rejuvenate the aged population and remedy most of the debilities of growing old. That is how we eventually intend to do it, if we can convince the U.S. government and the U.S. scientific community to perform the necessary research and validation.

probably first established a correlation database of porthole emission correlates to changes in internal states, both physical and mental. They almost certainly recognize the peculiar participation of every part of the body dielectric in the TW emission spectrum from the body. They will almost certainly have opted for this "porthole" notion of some similar expression of it, and the "thwarting" of these "relaxation and release" emission processes to engender internal physical changes in the body and in the brain-meat-computer, as they see it.

EXPLORE! VOLUME 9, NUMBERS 4 & 5, 1999

They probably experimented with different TW EM radiation mixes, to see what the body emission change to, and what states are induced. Undoubtedly they found that the induced states began to correlate with the former emission delta data base. That is, they found out which delta emissions spectra, reversed and amplified and fed back thorough the portholes, produce what internal changes.

Use for healing is the simplest of all. Use for killing or debilitating is more complicated. Of course weapons researchers wish to look only at the killing aspects. That is something they understand. They do not understand healing, for they do not even have the "regenerative" or "healing" mechanism in any of the scientific papers or texts. Becker[8] came closest in this country, and Priore did it in France. Both of those scientists were suppressed and paid a bitter price for their revolutionary and innovative research.

But look at what has been done, by "input-output" correlation research. The scientists will have a database now, that tells what state emerges in the body and in the mind, for what irradiation spectra (frequency, relationships, phases, amplitudes, wave shapes, etc.) one uses. And they did not work out all the wave-to-wave transductions ongoing in the mind and body connection and in the body and in the mind, individually, between TWÛLWÛTDW.

We suspect that this is the course Western mind control researchers will have taken, without even realizing the internal wave transductions occurring. And they would not be concerned with that. They would have what they want, for weapons, mind control, and behavior control. And they of the internal "mix" of the dielectric and all its tiniest portions. So Western mind control researchers and rogue groups probably do get good results.

We are not just picking on the U.S. government here! Internal divisive groups exist in all large and powerful groups, and joust for power. In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old "Power corrupts, and absolute power corrupts absolutely" routine.

The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.

In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, *it could account for what seems to be actually occurring*. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the would of course ignore the healing aspects. And they would get results of course, revolutionary results. So they would undoubtedly classify the program very highly, thinking that they had the best of the best.

Rogue Groups May Also Have Arisen to Divert the Purpose of the Projects

And that is where real problems may arise. Every human group has within it competing subgroups, jousting for power. The more highly classified the group and its activities, and the less that is known about it outside its confines, the more strongly and secretly the subgroups can play. So they can emerge as **rogue groups** pursuing their own agendas, not that of their duly elected governments. In the extreme, such a deep black program can even become a "captured" program, which is totally in the hands of rogues and no longer reports to or is bound by the dictates of the parent government.

It has long been suspected in some quarters that even the US's highly classified research community may be riddled by such rogue groups, and so may be certain highly sensitive parts of the Research and Development community.

Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A world." They may actually believe they are ahead of the Russians.

Where the Russians Probably Have Arrived

On the other hand, the Russians will have developed instruments for detecting and isolating the LWs and the scalar waves, in addition to the ordinary mundane instruments for measuring the TWs. So they will have not only the TW correlates, but also the much finer TWÛLWÛTDW (scalar EM wave) correlates. They will understand and employ the use of vacuum engines (spacetime curvature engines) which Western researchers will often invoke but not understand or even be aware that this is what is being invoked. In other words, Western researchers probably just use the old method of "input fi set of functions fi output". That is the standard basis for almost all Western system analysis. With it, one doesn't have to know the exact functions and processes going on in the middle. One just determines a transfer function from a given input to the output that results. Then one catalogs that transfer function. One does that for lots and lots of transfer functions for particularly desirable results, and one has the system analysis. One also has the immediate basis for a rapidly-developed applications technology.

Western scientists are good system analysts. But they have little or no comprehension of what is really going on in that "inside the ship functions" portion, even though they have determined the transform function which acts upon the input to give the "from the portholes" output.

The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced certain percentage of a highly secret rogue group will wind up using all these things and more. It's a **human** characteristic, the old primate dominance game. Only now disguised and hidden under deep classification.

Nonetheless, the "simplified" TW EM "porthole correlate" approach will yield very positive results, due to the correlation existing with the LW s and scalar portions than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U(l) electrodynamics. Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW interferometers, through intervening mass

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

including the earth and ocean. Western mind-control scientists and their rogue groups—assuming they exist as postulated—largely will not be able to do that, and probably won't even recognize Russian testing of psychoenergetics weapons directly in or over Western nations.

Russian Lethal Psychoenergetics Mind Control Tests Over the U. S.

In fact, the KGB tested just such highly advanced mind control weapons over the middle of the United States in 1997 on two occasions. The tests were conducted against Captain Button in his A-10 "Warthog" aircraft on April 2, 1997 [See Figure 14], and upon Captain Svoboda in her A-10 "Warthog" on May 27, 1997. [See Figure 15].

Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona. At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—dream-thinking until he died that everything was normal. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour.

Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks. A secondary purpose was to stimulate the U.S. government, watch its investigation and findings, and ascertain if the U.S. knew of the extent to which Russian KGB mind control weaponry had advanced. Obligingly, we proved to me KGB that we did not understand what had actually happened to Captain Button.

Captain Svoboda was from the same Wing as Captain Button and was its most experienced night fighter A-10 pilot. In fact, she taught other pilots the fine points of night flying combat missions in the A-10. About two months after Captain Button's bizarre flight and death, Svoboda was engaged in night target bombing on the range. Rising from her ordnance drop on me target, Svoboda was suddenly struck and rendered hypnogogic and in a waking-dream state. Her sense of the vertical was instantly reversed. Superb pilot that she was, she immediately dream-perceived she was diving, not climbing out of the delivery. She instantly corrected to "climb out". So in the real world she dived sharply into the ground and her aircraft exploded.

This test demonstrated the instant alteration of a skilled pilot's perception, causing that pilot to inadvertently react promptly in an unintentionally suicidal manner. A secondary purpose of the test was to stimulate the U.S. government system again, to see if the system recognized what had

A-10 Mysterious Flyaway and Crash

April 2, 1997

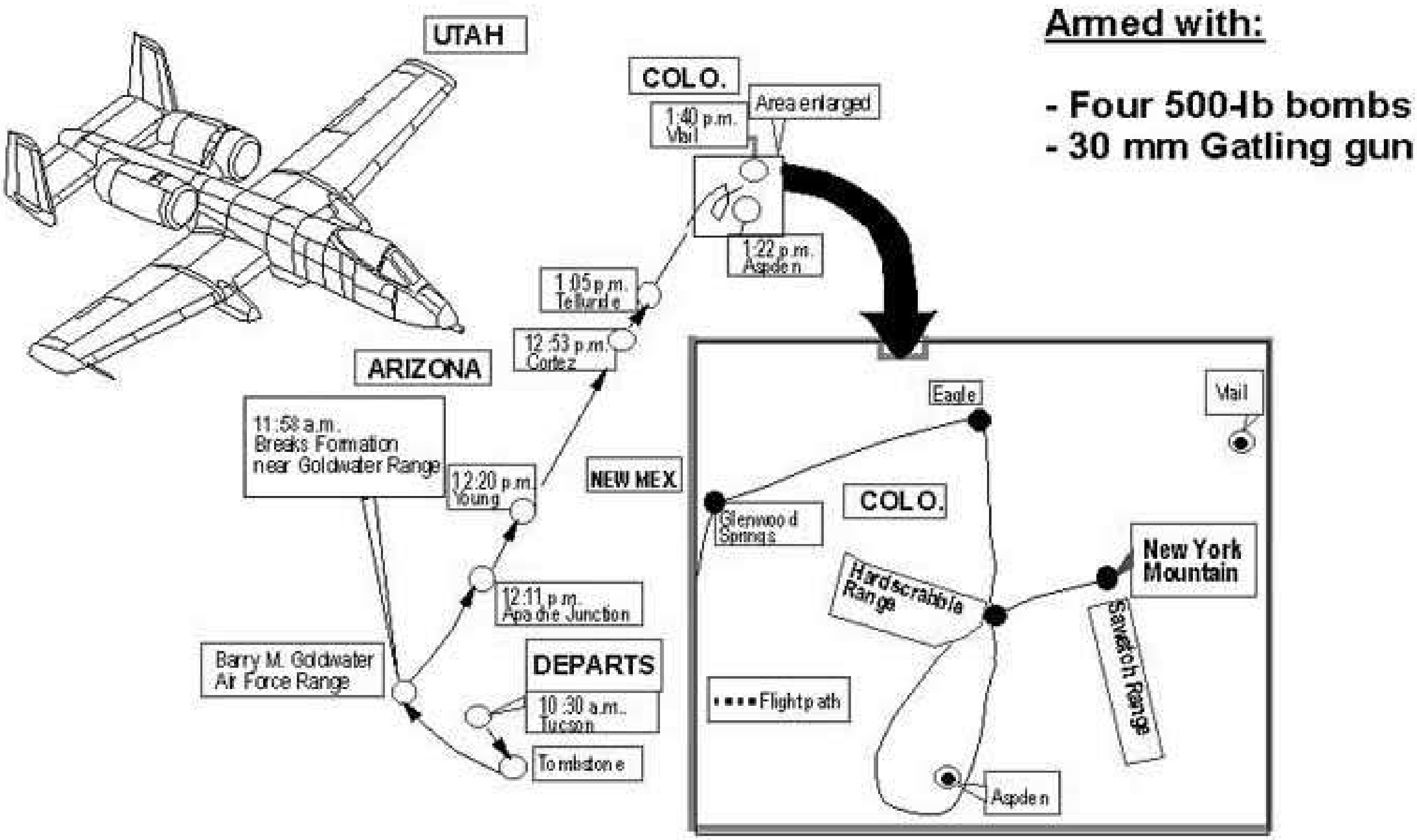
Armed with:
- Four 500-lb bombs
- 30 mm Gatling gun
UTAH
COLO.
Area enlarged
1:40 p.m. Vail
1:22 p.m. Aspden
1:05 p.m. Telluride
12:53 p.m. Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
NEW MEX.
12:11 p.m. Apache Junction
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a.m. Tucson
Tombstone
Eagle
Vail
Glenwood Springs
COLO.
New York Mountain
Hardscrabble Range
Sawatch Range
Flightpath
Aspden

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

occurred. Sad to say, it did not and it has not. It was lamely speculated that Svoboda probably looked over her shoulder like a rank greenhorn, and inadvertently dived her aircraft into the ground while looking back at the target. If there was one pilot in the USAF who would have not done that, it was Captain Svoboda. At any rate, the KGB determined that the U.S. scientific, intelligence, and governmental communities did not have the foggiest notion of what had occurred.

In private proprietary communications[9] we have explained further, to proper parties, exactly the reasons for these two tests, and what exactly was being tested, and what is its planned strategic use.

The Final Race for Direct Mind Control of the Entire Human Species

The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials.

The KGB scientists also know that something very like Gaia— a collective unconscious mind for all species on earth—also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials. If one reads some later books by Kaznacheyev,[10] e.g , this human species unconscious mind aspect and the biosphere unconscious mind aspect emerges loud and clear, if one interpolates between the lines, so to speak. Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.

I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society"—communist style—is still a variation of feudalism (all systems tend to feudalism or some form thereof).

The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now. It is simply a matter of developing sufficiently complex equipment to be able to directly engineer the time-domain substructuring at sufficiently deep levels. It is a doable, although it is very, very complex and very, very difficult. On the other hand, they have been at it for decades. Even in the 50s the Russians already had the old LIDA device,[ll] capable of inducing a catatonic trance-like state in a mammal—either a human or a cat,[12] for example. Some versions of Russian mind control devices were used on Russian soldiers in the

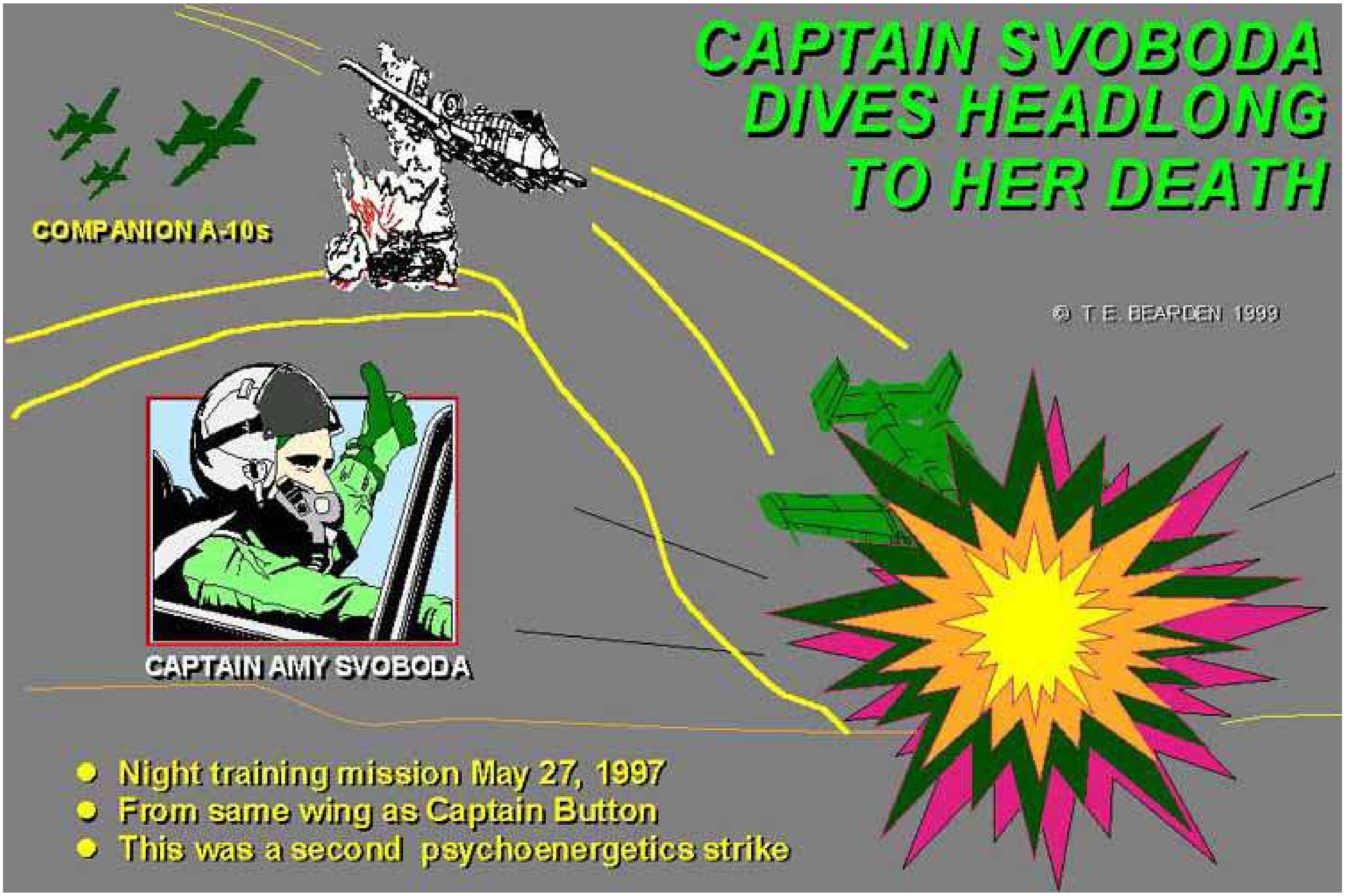
CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH
COMPANION A-10s
© T. E. BEARDEN 1999
CAPTAIN AMY SVOBODA
● Night training mission May 27, 1997
● From same wing as Captain Button
● This was a second psychoenergetics strike

Figure 15. Captain Svoboda dives headlong to her death.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

Afghan War to condition them for performance of merciless acts.[13]

Potential U.S. Rogue Groups and Two Recent Examples

Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.

Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed.

We Certainly Have Proof of Such Rogue Activity. For Example, Here are Two Prominent Cases:

as a result of those ghoulish experiments. So what happened when this gruesome thing was revealed? Again President Clinton publicly apologized! The head of the Executive Branch did not turn the macabre matter over to his Attorney General to prosecute. How many of these scientists were indicted? Not one. How many are going to be indicted? Not one. Suppose again this had been done by a private doctor in his own private clinic. You get the point.

There is Little or No Punishment of Rogue Scientific Groups

Shockingly, the U.S. government at the highest level has directly shown (and these are not all such cases by any means!) that mass crimes against U.S. civilians, *perpetrated by portions of the U.S. scientific community in direct conspiracy and in secret*, will likely be condoned. The perpetrators will not be indicted, tried, or convicted.

Note the connection of the "rogue groups" thesis to the above incidents. Note how rogue **scientific** groups got away with it in both these cases. There may be other rogue

1. For three decades, scientists from the U.S. government, universities, and civilian contractors secretly conspired to treat unsuspecting human patients with whole body nuclear radiation, including some retarded children and cancer patients. Some of those patients—including some of the retarded children—died as a result. These illegal experiments were conducted in great secrecy, and the results were highly classified. Eventually these actions were revealed, and a Presidential investigation committee investigated. These experiments would seem to be little different from the WWII Nazi and Japanese experiments on human prisoners. Are those U.S. scientists who were responsible for those retarded children's' deaths any different from the Nazi scientists and criminals we executed at Nuremberg? So what happened when this terrible thing finally came out into the open? President Clinton publicly apologized! How many of the responsible scientists were indicted for murder? Not one. None is ever going to be.
   Suppose a doctor here in South Alabama, with his own private clinic, had secretly irradiated those human patients and retarded children with whole body nuclear radiation, over a period of 30 years, resulting in some agonizing deaths. How many state and federal agencies would have come after him with arrest warrants? Probably about 50 or so. Would he have been indicted, tried, convicted, and executed or sent to prison for scientific groups getting away with such things today. They should not be able to get away with it. But they can.

Perhaps a most startling additional part of those two incidents is that there was no great public outcry from the scientific community, deploring these murders and demanding that the criminal scientists be indicted and tried.

As the old saying goes, "By their silence they have convicted themselves." The organized scientific community, **as a community**, has shown that it has little or no ethics, and—while deploring any scientific murdering that "gets revealed,"—is not really interested in justice. In short, much of the U.S. scientific community may now have very little ethics left.

We are **not** talking about normal individual scientists, but the Big Science community. There is a whale of a lot of difference between the two. In the Big Science community, there can be and there are rogue groups. Lots of them. There is deep cover, deep classification. And there is very probably advanced mind control research and testing, be it legal or illegal. Hopefully most of it is legal and constrained. However, some of it is almost certain to be illegal and ill constrained.

Human beings are still human beings. All the good and evil is still there, regardless of the group. The stage settings change, but the cast of characters and the play never change.

It's sad, but 'twas ever thus. Hidden parts of our own governments—and other governments throughout the world—are no different from the old medieval groups, where nobles etc. were always plotting against the king, or using the king's power for their own nefarious end. The rogue groups today are no different from all the scheming and conniving groups that life? Absolutely! Would the President of the United States have apologized? Not on your life; he would have turned it over to the U.S. Attorney General with a strong directive to investigate and prosecute on criminal charges.

2. Again for three decades, scientists from (i) the U.S. Government, (ii) universities, and (iii) civilian contractors conspired to give poor syphilitic blacks in Tuskegee, Alabama a placebo while ostensibly treating them for syphilis. The purpose of the program was to deliberately observe and record the ravages of the disease to its fruition in their wracked bodies. So the scientists deliberately sat there and watched their brains rot and their bodies rot. They kept meticulous notes, of course. It was very scientific. And it was macabre. They "treated" some 400 blacks in the "program." Over 100 blacks died

destroyed the Roman Empire. *Great empires fall from within, not usually from without!*

Conclusion

This concludes our brief portrayal of the probable directions that mind control research has taken, in Russia and in the West. It is a novel and powerful area, and it poses both severe dangers and great benefits to all the peoples of the Earth.

Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction of helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth. Yet it can revolutionize medical science, education, communication, and psychology. Eventually it can engineer the mind and memory directly, beneficially. We foresee the day—perhaps 30 years hence—when education will be

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

accomplished *by directly loading the software into the mind.* Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks one will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an M. D. And so on. When that happens, then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.

We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on Earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to

446955, Oct. 23, 1998, forwarded to US. Patent Office on Nov. 4, 1998.

5. A very good introduction to nonlinear phase conjugate optics is given by David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(2), March/April 1982, p. 156-183. Also see David M. Pepper, "Applications of Optical Phase Conjugation," Scientific American, 254(1), Jan. 1986, p. 74-83.

6. Pepper, Optical Engineering, ibid., p. 158.

7. See T.E. Bearden, Energetics of Free Energy Systems and Vacuum Engine Therapies, Tara Publishing, Internet node www.tarapublishing.com/books,July 1997. See also T.E. Bearden, "Energetics Update and Summary, " Part I, Explore, 7(6), 1997, p. 60-67, Part II, Explore!, 7(7), 1997, p. 53-56; Part III Explore!, 8(1), I997, p. 53-56; Part IV; Explore!, 8(3), 1997, p. 56-63. See also T.E. Bearden, "Vacuum Engines and Priore’s Methodology: The True Science of Energy-Medicine: Parts I and II " Explore!, 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62.

8. E.g., see R.O. Becker; Charles H. Bachman, and Howard Friedman, The direct current system: A link between the environment and the organism, " New York State Journal of Medicine, Vol. 62,

be the Sword of Damocles and turn it into the golden Millennium.

We are reminded of a poignant quote from Teilhard de Chardin:

"Someday, after we have mastered the winds, the waves, the tides and gravity, we shall harness for God the energies of love. Then for the second time in the history of the world man will have discovered fire."

May the nations of the world utilize the new mind engineering science to truly free all minds into a common love of, and respect for, their fellow humans of every color, creed, and circumstance. As de Chardin envisioned, may we use the new science to discover a new fire, burning clearer and brighter than anything that has gone before. ✦

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel US. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, US. Army. Graduate of Guided Missile Staff Officers Course, U. S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses.

Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor: He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuums virtual particle flux is organized and used to energetically shape and manipulate

April 15, 1962, p. 1169-1176. See also R. 0. Becker; "The direct current field: A primitive control and communication system related to growth processes, "Proceedings of the XVI International Congress of Zoology, Washington, D.C., VOL.3, 1963, p. 179-183; —"The neutral semiconduction control system and its interaction with applied electrical current and magnetic fields," Proc. 1lth International. Cong. Radiol., Amsterdam, Excerpta Medica Foundation, 1966, Series 105, p. 1753-1759. See particularly Robert O. Becker and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, Vol. TBD, No. 73, Nov.-Dec. 1970, p. 169-198.

9. Bearden, T.E., Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ page enclosure to CTEC Letter, "Saving the Lives of mass BW Casualties from Terrorist BW Strikes on US. Population Centers," distributed to selected high U.S. Government agencies and Legislative subcommittees, May. 4, 1998.

10. E. g. , Vlail Petrovich Kaznacheyev, V. I. Vernadsky's Study on Biosphere and NooSphere, 1989 [in Russian]. Vernadskys concepts included living matter, biosphere, and human-transformed biosphere (noosphere). Kaznacheyev writes of evolution and

matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

---

1. Owen Flynn, "Parametric arrays: A new concept for sonar, "Electronic Warfare Magazine, June 1977, p. 107-112.
2. Valerie Hunt, presentation to the U.S. Psychotronics Associations Annual Symposium, Bergamo Center, Ohio, 1988. Hunt presented epochal work showing chaos—and thus hidden order—in recorded ELF electromagnetic radiation from human bodies. Hunt split the received EM signals into two identical channels, time-delayed one channel by a very slight amount, then recombined (mixed) the nondelayed and delayed signals. The result was that attractors precisely indicative of chaos appeared. This positively proves that human systems produce ordered time-

organization of the living matter as the biosphere of the earth. Without specifying the exact technical details, Kaznacheyev discusses the gist of much of what is in this present article.
The present author has an autographed copy of Kaznacheyevs book referenced above.

11. Rabichev, Lev Yakovlevich; Vladislav Fedorovich Vasiliev; Alexandr Sergeevich Putilin; Tatyana Grigorievna Ilina; Petr Vasilievich Raku; and Leopold Pavlovich Kernitsky, "apparatus for the treatment of neuropsychic and somatic diseases with heat, light, sound; and VHF electromagnetic radiation," U.S. Patent. No. 3,773,049, Nov. 20, 1973. This is the U. S. patent for the LIDA machine, developed by the Russians prior to the Korean War. Additional patents or applications are as follows: Italy, #619,129, March 1961. Germany, #1,183,607, Dec. 1964. Great Britain, #826,766, Jan. 1960.

12. Stefan T. Possony, "Psy-War: Soviet Device Experiment," Defense & Foreign Affairs Daily, 12(104), June 7, 1983, p. 1-2. Reports on Dr. Ross Adey's investigation of the Soviet device called LIDA which is used to bombard human brains with radio waves in the 40-MHz region. Also reports on work by A.S. Davydov of the Ukrainian Academy of Sciences, who discovered how the blood-brain barrier can be penetrated by low-frequency radio waves so that brain cells are affected. See also "Soviet device uses radio

structuring of EM.

3. Years ago, I read an obscure reference to Russian research directly along that line, with successful development of just such a portable device. Unfortunately I lost the reference in a disastrous flood in my basement and have been unable to recover it.
4. T. E. Bearden, "Alteration of Materials by Wideband Multiwave Irradiation Combined with Selective Emission Manipulation," Invention Disclosure Document Number

waves as tranquilizer, " Associated Press release, Huntsville [Alabama] Times, May 19, 1983. Dr. Ross Adey tested the Soviet LIDA machine on a cat, and in two or three minutes the cat was sitting quietly, almost as though transfixed, and uninterested in its surroundings. The animal remained uninterested in its surroundings for some 20 to 30 minutes after the machine was turned off.

13. "KGB Used PIS to Program Soldiers, "Resonance, No. 30, Mar. 1996, p.29. Quoting: "During the 1970s, the Soviet KGB developed a Psychotronic Influence System (PIS) that was used to turn people into programmable "human weapons." Soldiers fighting in the Afghan conflict were trained using the system which employs a combination of high frequency radio waves and hypnosis. Commands are activated by code words or cyphers. The purpose of PIS was to render regular soldiers and members of the special forces incapable of feeling sympathy.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

The Tom Bearden Website

Latest cloning research findings by Bedini/Bearden

The huge dangers of current cloning technology

Mystery of the death of Australia's cloned sheep explained

"Spark of Life"

From "Stranger than Science"

by Frank Edwards © 1959, 1987

Did an obscure amateur scientist discover the spark of life? Did Andrew Crosse accidentally stumble upon a mixture of chemicals and electricity which produced living things from a lifeless compound? If he did not, then just what did he accomplish? The record of his work is clear, only the enigma of the results remains to plague science.

The neighbors of Andrew Crosse regarded him as more devil than man. They did not understand the bright flashes that lighted his laboratory windows at night when he was tinkering with his crude electrical devices. Not only was he dreaded and shunned as "the thunder and lightning man," but he was denounced as an atheist, a blasphemer, and a Frankenstein who had best be put in chains for the common safety.

Andrew Crosse minded his own business, it is true, but his was a very strange business for the early 1800's. And one of his experiments remains a very strange business to this day.

Andrew was an enthusiastic amateur experimenter in the new field of electricity. Out of touch with others in the same work, he labored under the double handicap of not knowing what had already been done - and of not understanding what he was doing himself. Yet it may have been this very lack of knowledge which led him to undertake the experiments which were to inscribe his name, however faintly, in the annals of science.

He decided to induce the development of artificial crystals by subjecting chemicals to prolonged exposure to weak electrical currents. Andrew mixed up some silicate of potash and hydrochloric acid and into this he dropped a fistsized chunk of oxide of iron. By inducing the current from a small battery to trickle through the solution to the oxide of iron, he hoped to bring about the growth of artificial crystals of silica.

Having arranged this combination of chemicals and current, Andrew set it aside and went back to his favorite pastime of studying the spark-gap - the flashes which alarmed his neighbors.

Did he accidentally stumble upon an arrangement which created life from inorganic matter?

In a paper which he wrote for the London Electrical Society in that same year of 1837, Andrew set down this account of his experience.

He wrote "On the fourteenth day after the commencement of this experiment, I observed through a small magnifying lens a few small whitish specks clustered around the middle of the electrified stone. Four days later these specks had doubled in size and had struck out six or eight fine filaments around each speck . . . the filaments longer than the hemisphere from which they projected.

"On the 26th day of the experiment, the objects assumed the form of perfect insects, standing erect on the bristles which they were growing. Although I regarded this as most unusual I attached no singular significance to it until two days later, the 28th day of the experiment, when the magnifying lens showed that these things were moving their legs. I must say now that I was quite astonished. After a few more days they detached themselves from the stone and moved about through the caustic acid solution.

"In the course of a few weeks more than a hundred of them made their appearance on the oxide of iron. Under a microscope I examined them and found that the smaller ones had six legs, the larger ones had eight. Others who have examined them pronounced them to be of the genus acari, but some say they are an entirely new species.

"I have never ventured an opinion on the cause of their birth for the reason that I have never been able to form one. I thought they might have been airborne creatures that had drifted into the liquid and prospered, but later experiments with closed vessels, in which the ingredients had been purified by baking in the oven, produced identical creatures; therefore, I suggest that they must originate in the electrified liquid by some process unknown to me."

Andrew Crosse realized that he was walking a tightrope before the top scientists of his day. He was describing an experience foreign to their accepted understanding and therefore he was inviting ridicule. He did not have long to wait. Cries of fraud and hoax engulfed him. He and his alleged insects were denounced as nothing more than humbugs.

Amid all the furor that his announcement had created, Crosse stood alone and helpless. Even other scientists who had duplicated his tests with similar results kept their silence.

All but one. That voice was raised in his defense and it was such a powerful voice that none dared challenge it; for it belonged to the great Michael Faraday. At last Crosse had found his champion.

Faraday reported to the Royal Institution that he too had experienced development of these little creatures in the course of his experiments. But he added that he could not decide whether they had been created in the sterile solutions or brought back to life by the electricity! Either development would have constituted a milestone in scientific advance, as Faraday realized, but he left it to his fellow scientists to make the decision, if any.

Crosse never claimed to have discovered anything. He was merely reporting what had happened. After the attacks upon him subsided, he retired to his home in the Quontock hills for many more happy years among his test tubes and batteries.

Although he spent his long life as a humble searcher after scientific truths, the contribution for which he is remembered is the controversy over his acari - unwanted then, and unexplained to this day.

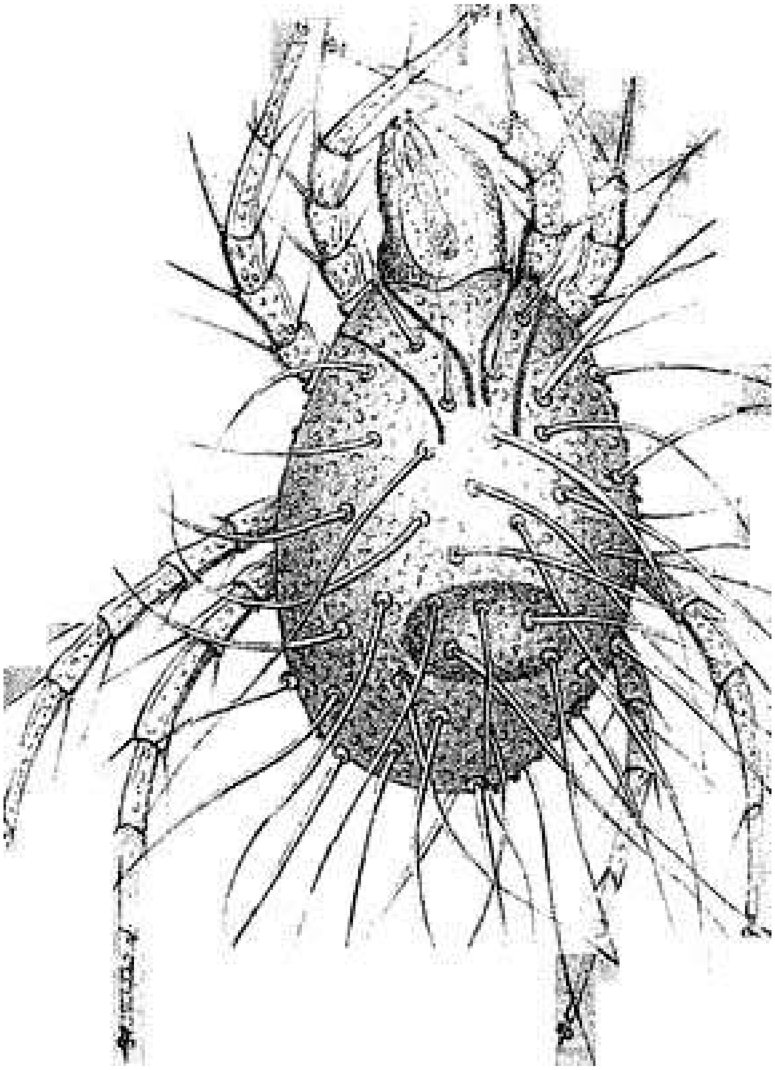

Genus Acari

Tom Bearden comments (finalized and slightly updated in March 2001).

It is difficult to make an understandable comment on the business of life, because presently scientists have not the foggiest notion what mind, being, life, death, time, etc. are. For that matter, they don't really know what matter and energy and charge are either. In the scheme of things, our modern "science" is only about three hundred years old, and that's a snap of the finger in even just the existence of this world. Much of modern science still considers the mind as a "meat computer", or just the physical electrodynamic activity of the brain, etc. Materialism has not died, even though quantum mechanics essentially wiped it out decades ago, and the old concept of "mass" as something hard, permanent, and "material" has also long since vanished. One is referred to the works of Max Jammer to see what the present view of "mass" is. Also, without looking up the references again, recent experiments with quarks and gluons wound up with experiments in which it mattered not a whit whether one considered the mass as zero or nonzero. Nature, it seems, does not have to play by our usual "either-or" rules. In physics, of course, this is the duality problem, which to this day has not been resolved because attempts to resolve it have applied Aristotelian logic -- and there is no solution to the problem of "opposites being identical" in that logic, which forbids that very thing by authoritarianism.

The modern biogenesis experiments, e.g., never took place in the absence of life. There was no "statistical generation of living or quasi-living forms" in the absence of engines for such forms. The great rush of Poynting energy flow around the earth, generally parallel to the equator, alone carries the "engines" (internal sets of spacetime curvatures) for every form of life that ever existed upon the earth. So the experiments were bathed in "living forms and dynamics", if one accounts for the internal Whittaker-structuring of electrodynamics -- noted but ignored by our scientists for nearly 100 years.

Quite some time ago, we worked out the nature of mind, and the mind-matter coupling mechanism (which is engineerable). We did that in pursuit of the mechanisms in the KGB/Russian psychoenergetics weapon work. I had the last of a 3-part article published in "Explore", dealing briefly with mind control technology both the way it is probably done clandestinely in this country and how it is done by the KGB/Russians. I'm still trying to get the proper parties to believe it; not much progress. However, some very neat information has been released on the Internet that very probably reveals why officially there is no interest. It's because three decades ago some of our "spooky folks" found (in deep hypnosis experiments) that humans are inordinately sensitive to extremely low-level EM signals and fields. Even though no detection or sensation may rise to the conscious level, the unconscious (which is totally conscious, just massively parallel rather than serial) does sense these signals. Voila! It would appear that "very deep black" parts of the community probably developed quite good mind influence and mind and behavior control or interference using techniques adapted from that work. The reader must always remember that just because the *overt* government does not know something or have something, does not at all mean that the *covert* government doesn't have it. And even in the covert government, there are layers within layers. So there, even when the upper levels do not have something, it still does not mean that the deeper levels do not have it. The problem with the best knowledge and technology, when unknown to the upper covert levels and the overt level, is that it is then quite likely to be subverted into the deepest levels, which may have their own agendas and may not report to the duly elected U.S. government at all. Indeed, such levels often assume that the President himself does not have a "need to know".

Present physics has erroneously omitted the time increment carried by the photon. The overemphasis on the so-called transverse wave (which does not exist as such in vacuum, fairly readily shown) has led to our missing the greatest and most important part of electrodynamics: the longitudinally polarized EM wave and the time-polarized EM wave. For example, mind and mind operations are totally electromagnetic, but use time-polarized EM waves rather than the standard stuff. Overtly we do not even have detectors for LPWs and TPWs yet; the Russians do. However, every major weapons lab on earth now is aware of longitudinal EM waves and their powerful characteristics. Simply read some of the summary material on Undistorted Progressive Waves (UPWs) on the Los Alamos National Laboratory website, e.g. by Rodrigues et al. A UPW is an "imperfect" longitudinal EM wave with some remaining transverse wave residue (that is the way it is modeled).

Let me give you one example of how fouled we are. We still use Aristotelian logic primarily. Yet it is an incomplete logic. Let me show you instantly. Take the simple Venn diagram of a rectangle lying on its side, divided in half by a vertical line in the middle. Label the right rectangular area A. Label the left rectangular area "not-A". What is the line in the middle and the border? Both A and not-A, so that A is identical to not-A. Yet the third law (law of the excluded middle) forbids A identical to not-A. Apply the third law to the Venn diagram. The border disappears, and the middle dividing line disappears. The logic self-destructs, because now there is no separate A and no separate not-A. Yet so far as I can determine, no one seems to have seen that very simple thing and questioned it. Aristotelian logic cannot be applied to the very Venn diagrams used to establish and "prove by demonstration" its propositions. We corrected that nearly 20 years ago, and got called every kind of lunatic imaginable -- the usual reaction when something fundamentally wrong is pointed out in a comfortable part of the prevailing paradigm. These days I simply point to Morris Kline's "Mathematics: The Loss of Certainty" to the Aristotelians, and tell skeptics to go and read it and then we will discuss the foundations of logic. With a 5-law logic (or 4-law with an additional applications rule), one solves the "dividing line" problem nicely. Then suddenly things like spacetime and being and mind and nothing etc. can be solved -- in the 5-law logic, but never in the 3-law logic. The problem of life itself -- whether terrestrial or extraterrestrial -- has no solution in 3-law logic.

The entire method of Zen is to try to break up the iron Aristotelian bent of the conscious (serial processor) mind, and get into the far broader unconscious (totally conscious, massively parallel processor) mind. Few have done that. But one cannot present a 5-law solution to a 3-law processor; it just "does not compute". It's like putting a thousand slides at once on the normal "single slide at a time" projector. The "conscious mind" can see only a single slide at a time. So it just sees "black" or "nothing at all". That's why we "see" the "massively parallel conscious" mind as "unconscious"! It isn't. But it is a quite different *form* of consciousness which we all have.

Time is never observed, and no observable is a *cause*, a priori, because an observable is only a frozen, nonchanging 3-spatial "instant snapshot" and therefore the *output* of the observation process, while the *causal input* is what exists in 4-space prior to its interaction with the observation process. When one adds time and time dynamics to one's science, one moves far in front of much of present Western physics. E.g., physics has nearly hopelessly confused the *causal* side of the observation mechanism/process with the *effect* side, everywhere, throughout physics -- simply by assuming that observables persist in time -- a non sequitur since an observation is a frozen 3-space snapshot at one single instant.

And sadly, the physicists do not seem to have even realized it!. Not a single paper or text in the West, e.g., illustrates an EM wave in spacetime prior to its interaction with a unit point charge. E.g., simply check a fine editorial by Romer. Romer is the editor of American Journal of Physics, and we are referring to Robert H. Romer, "Heat is not a noun," American Journal of Physics, 69(2), Feb. 2001, p. 107-109; and specifically to his end note #24, p. 109. Romer takes to task *"…that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. …it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "…perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

But in general physicists continue to charge on, confusing effect as cause. Electrodynamics, e.g., is particularly guilty of greatly confusing the two, which is really what is preventing a successful and engineerable unified field theory. Further, contemporary classical electrodynamicists do not calculate either the field or the potential; instead they calculate the *reaction cross section* of each, interacting with a unit point charge assumed out of nowhere. In short, they calculate how much is diverged from the actual field or potential, and then call that "little extracted part" the field or potential! Again, they mistake the causal entity existing before the interaction with the effect entity existing after the interaction. Well, after the interaction has occurred, that result is the *effect*, not the *cause prior to interaction*. Physics has this so hopelessly snarled that it will require 50 to 100 years before the journals and physics societies are likely to even realize it and admit it, and then it will require them another 50 years to straighten it out and purge this huge non sequitur from physics theory and modeling.

Anyway, we did model how mind and mind operations operate, including their precise nature, how they fit into all this, etc. We thus were able to resolve many previously unsolved riddles such as the nature of intent, how a space-less mind operation can generate a spatial input to a spatial body and vice versa, etc. We also worked out how to use "effects" systems to generate "causal" systems, completely independent of mass -- because the KGB has recently completed a completely different kind of weaponry based on this, and different from anything that has ever existed in the history of mankind (at least on this planet!). The Russians of course were already there; I got there much more crudely by intensely working on how to explain what they were doing.

To see that the effect can generate the cause, just consider the normal "forward time" observation process having to operate in a "time-reversed" process. You see, not only must we shift observers from frame to frame, but we must also be able to shift them from forward-time (of some conveniently chosen outside reference observer) to reversed time (of that same reference observer).

At any rate, it turns out that the entire universe is already "alive" in a very special sense. There is no death; there is mind (I use the term "mind" in the most general sense, not just the conscious mind which is merely a "periscope" put up and taken down from the unconscious). And mind is not a part of the brain, and it is indeed separated from body but intimately linked to it by the mind-body coupling dynamics. Further, any life that has "ever existed" still dynamically exists in a special kind of more fundamental electrodynamics infolded inside all normal EM fields, waves, and potentials. The Tibetans knew much of this very well, but of course in terms that our materialistic science rejects or completely fails to understand (as we stated, most of our scientists believe the human is a robot and the mind is simply the electrochemical reactions of a "meat computer"!).

Since ambient spacetime may be modeled as a special kind of "scalar" potential with intense energy density, it then decomposes into a series of bidirectional EM longitudinal wavepairs, a la Whittaker 1903, except one must use impulse waves so as to include the time dynamics part of spacetime, rather than forcefield 3-space type waves. The entire thing -- mind, life, matter, fields, waves, spacetime, etc. -- then becomes structurable and directly engineerable by novel EM means (such as the O(3) higher symmetry electrodynamics pioneered by Evans and Vigier). Longitudinal EM wave technology is the key to developing such mind-body coupling mechanisms and engineering of them, at all. One only gets the time-polarized EM wave technology *after* the LPW technology has "done its thing" -- the time-polarized technology is far and away the most powerful and primary.

Also, low frequency photons have far more energy than high frequency photons, if one includes the time component, since time is spatial energy compressed by c-squared and has the same energy density as mass. Time-energy (highly compressed spatial energy, compressed by c-squared but in the time domain rather than in 3-space) and spatial energy in the photon are canonical. As the frequency of the photon decreases, its spatial energy decreases also with the decreasing frequency. However, the time-energy component increases, and its compacted energy increases as the square of the time-component expressed in seconds. Hence at lower frequencies the total energy (sum of spatial energy and time energy) in the photon increases. Indeed, if one wishes to engineer gravitational effects, it is the low frequency photons (low ELF) that are important, because they have far and away more "spacetime curving energy" than high frequency photons such as gamma rays.

At least eight nations on earth now have longitudinal EM wave technology weapons to one extent or another. Several of them -- hostile to one another -- are using Western Australia as a convenient nearly uninhabited target range.

Anyway, when you come at the mind-body-life problem with that background (which unfortunately required more than 20 years of very hard work!) there appears to be no great difficulty in explaining the Faraday, Crosse, Fox, Miller, and Urey biogenesis experiments. And yes, there appears to be no difficulty in explaining the "appearance" of "life on earth" in the prebiotic soup. It did not really appear out of nowhere spontaneously at all, but simply was a sort of sympathetic resonant coupling of physical forms to mind-forms already there. None of the experimenters realized that the great Poynting energy flow (and one must add back in the missing Heaviside component which is some 10exp13 times larger than the accounted Poynting flow) around the earth parallel to the equator and from West to East, is already internally structured with every possible form of life ever in the universe anywhere. If we further multiply that Heaviside component by c-squared, so that we can account for the time-energy portion of that great energy flow, we can see immediately that life and its environment are both living and intercommuting all the while. The creator in fact created a living universe, not an inert one (as best we can speak it in the inadequate 3-law logic). If one believes in big bangs, then life structuring emerged right along in all the fields and waves and energy, right in and out of the big bang. There is not a single point in the universe anywhere that is absent of energy and therefore absent of life, in the view of the model I am using.

Understand, one does not have to agree with this model! That is one's own decision. We are just pointing out that, if we correct many of the errors known to be in the presently used models, this is the kind of reality that readily emerges. We get a physics that is not just an "inert" physics, but a living physics where literally the entire universe is "alive" in very special manner.

In deciphering (after 14 years!) the Priore mechanism, I found that incredibly the cellular regenerative system does retain the form and "memory" of all the precessor cellular forms. Cancer, in fact, is very often due to a direct "order" by that system to a cell (aerobic) to start back down the "chain" to an anaerobic cell, because of severe hypoxia and the sustained inability of the regenerative system to maintain the cell in aerobic form. It doesn't just send an "order", but sends a "precise set of spacetime curvature engines" that go to work on that cell and all its parts. The first major step back is to break it loose from centralized control of its growth -- whereupon we recognize it as a tumor cell. There are two requirements: (1) cellular damage, and (2) "promotion". Biologists do not really understand "promotion", because they do not understand the cellular regenerative system—which used time-polarized EM waves, and nonlinear optical pumping of the cell and all its parts in the TIME-energy domain rather than the SPATIAL energy domain. That sort of pumping creates an "antiengine" to its forward time engine resident in the sick cell, thereby time-reversing the cell and all its parts back to a previous state on the trail from the early anaerobic single cells to the present aerobic cells.

By the way, recent experiments in cloning are beginning to bear out the fallacy that exact electrochemical processes produce the same end physical results, particularly when the persistence of the created forms is examined. E.g., cloned rats may live half their lives in apparently normalcy—then suddenly grow greatly obese and die. Many other anomalies are being uncovered in this fashion. The reason is quite simple: We have some diagrams showing that one may introduce an additional "engine" to the pumping of a diseased or damaged cell (or a healthy one for that matter), and in the time-reversal process then induced, the "time-reversal" process will be diverted from the original trajectory. In short, the original full state will not be recovered.

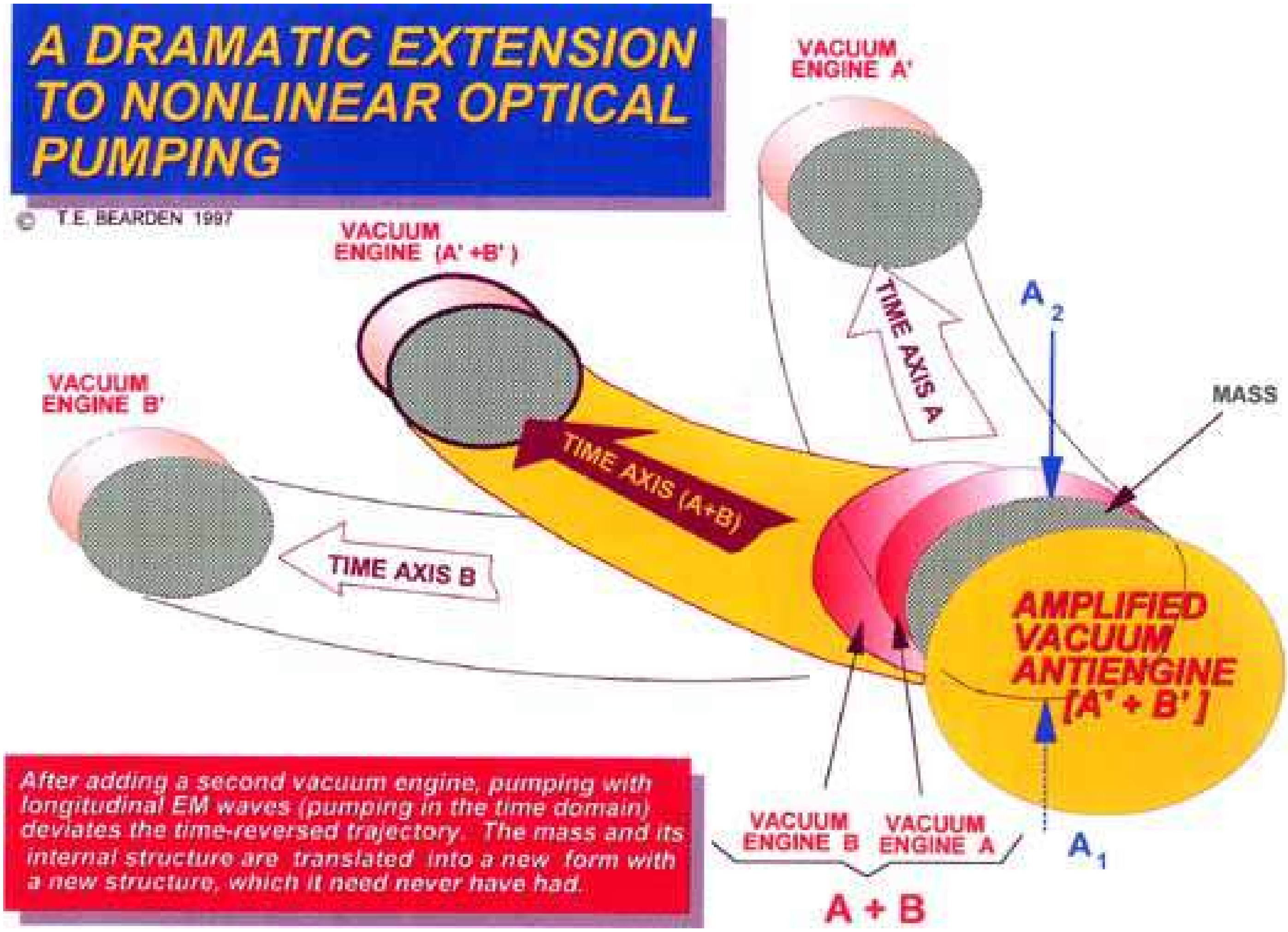

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

This, we pointed out, is a second generation technology to be developed, where a disease or disorder must be treated and where also there was never a normal past state anyway. A case in point is birth defects, etc. Eventually, if the technology ever is allowed to be funded and developed, even those kinds of medical conditions will be treatable and curable. The point is that in cloning, one does not use the normal sperm cell and egg, without abnormal disruption. There is indeed an abnormal disruption -- and every such physical action or change and dynamics certainly involves an additional set of engines. In short, with present methods the cloned entity also has introduced into the equation an additional "engine". In the fertilization process, there is quite a jump back to a "time-reversed" situation (fertilization is the beginning of a physical embodiment, hence from two parts of an embodiment already in being, there is a rather dramatic "jump backwards"). The "normal" backwards jump of that kind is a normal fertilization of the egg by the sperm. Any change to that procedure, represents the insertion of an abnormal engine to be processed in that same "jump back". For the methods being used, it appears that the abnormal engines are sufficiently weak that, in the cloned animal, some time must expire in its lifetime before the impact of that little constant deviation from normal grows to become suddenly significant.

Indeed, all "living forms" are materialized from energy, but from both matter-energy and time-energy and the dynamics of the two. Natural processes include fertilization, birth, etc., but there is no technical reason whatsoever that the process cannot be direct and even directly engineerable. In this "expanded energy" sense, mind energy and dynamics can be engineered as well as body matter energy and dynamics.

To cinch the matter of "engineering living forms", I once participated in a accidental and unplanned experiment where Golden materialized living things. Not little bitty things, but big things. In my view we actually materialized thought forms, since much later we were able to piece together exactly what apparently engendered each part of the phenomena. There were multiple (scared) witnesses who prefer never to discuss it. The experiment also altered all the clocks in the area for four days before the "time-energy charge" we had inadvertently created dissipated. Let me just say that it really got my attention! We never repeated that experiment again. But the changes were substantial and the phenomena were substantial, and to this day I prefer not to discuss that experiment. It simply cannot be understood without the outline sketch we are speaking of, where one accounts for time-energy and dynamics as well as material energy and dynamics, and also for the coupling dynamics between the two.

Not only can one have a "material entity" or thing communicating with another "material entity" or thing, but one can also have a "mental thing" (time-domain thing with no spatial existence) communicating with a "material entity or thing" with only spatial existence, and vice versa. There is indeed a coupling two-way mechanism

Look at the so-called "physical form" -- which almost always one is speaking of as purely 3-space. Most scientists have not yet understood that a priori a piece of 3-dimensional mass *as observed* cannot even persist in time! Masstime exists in time, but "mass" is a frozen 3-spatial intersection of the spacetime entity. It is exactly comparable to a frozen frame in a movie film. The individual frame never has any existence except at a frozen moment in the slide projector. "Motion" and "movement" etc. involve the time domain a priori -- hence something without any time-component at all, cannot persist or even exist except at one frozen moment in time. Hence any observable existed in that single "observation slice" and at that "point in time" only. After all, a "point in time" is where time symmetry is broken or ceases. In short, it is a time-discontinuity. All observations are such time-discontinuities.

The SETI problem will not be solved, in my personal opinion (understand, I do not object at all if others do not see it this way) until a far greater and deeper understanding is reached by our modern physics. Forget communicating with "transverse" waves; longitudinal EM waves can travel at any velocity, limited only by how much "residue" of the normal EM wave junk is still hanging in there (how much "transverse EM wave noise" is still present. A pure LW moves at infinite velocity. As we suggested, simply download the Rodrigues and Lu papers carried on the Los Alamos National Laboratory Web site. It appears that now most major weapon labs have "discovered" the fact that there can be longitudinal EM waves, and some eight or so nations have developed weapons technology from such. If one has such technology, all the rest is obsolete.

For the skeptic, here is a statement made in 1997 by Dr. William Cohen, then Secretary of Defense:

> *"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts."*
>
> [Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.].

Any technological civilization that has not solved the fundamental "self-killer" problem of the technological species, must and will self-destruct. This follows from simple servo-mechanism theory, as we have previously pointed out. The KGB is working on a "solution" that is far along: engineer the entire human species quantum potential so that the human species can be turned into an "antlike" species, which of course would be the "ideal" communist state, in the minds of many Communist party theoreticians. Our guys haven't even got started on engineering the quantum potential yet. They have not yet even discovered the big ball park, let alone the big ball game.

Anyway, that is just some comments for your personal information. What really is happening in the "hidden weapons" area and "hidden weapons science area" is already so advanced in some quarters that science fiction pales by comparison. The real question is whether we shall survive very much longer, and whether we shall mature rapidly enough to begin to use these technologies for the benefit of humanity, rather than for its destruction.

Sincerely,
Tom Bearden

Additional Comments from Tom dated 9 July, 2001

CHENIERE PRESS

The Tom Bearden Website

Return to Order *Excalibur Briefing*

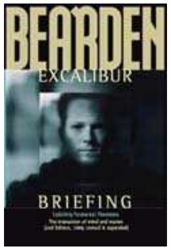

Help support the research

CHECK ONLINE STORE FOR UPDATES

EXCALIBUR BRIEFING

THE NEUROPHONE

Another device that uses the new hyperspatial, virtual state, nested modulation technology (and has done so for seventeen or eighteen years) is Dr. Pat Flanagan's neurophone. With brilliant insight and intuition far beyond that of science at the time, Pat invented and patented the instrument by the time he was seventeen years old. The neurophone is a device that, contrary to all present theory and knowledge, will directly “pump the brain” and reproduce sound and information directly in the brain and mind system, without going through the auditory system at all. A simplified diagram of Pat's improved neurophone is shown in Figure 11.

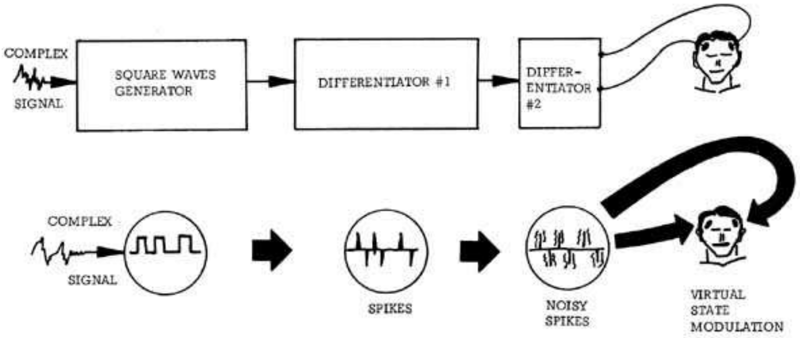

Figure 11. A simplified diagram of the improved Flanagan neurophone

Briefly, the device takes a complex signal, such as the sound of an orchestra playing a musical interlude, and electrically processes it as shown in the figure. First the signal is passed into a section that clips everything into a series of square waves, remarkably analogous to the sort of clipped waves Lisitsyn confirms are the information carriers of the human brain. Next the square waves are differentiated, yielding a series of sharp spikes (note that these spikes retain the pulse-time content of the clipped signal). These spikes are again differentiated, and since these are finite spikes with real nonzero rise times and decay times rather than theoretical constructs, a series of noisy spikes results from the second differentiator section. From here, the noisy spikes are introduced to special contact electrodes, one of which is normally placed on the forehead, while the other may be placed almost anywhere, including on the foot. Nowhere are any sound waves introduced to the head.

The square-wave clipper section reduces the complex signals, their overtones, and their complex modulations to square waves, retaining the temporal content of the wave mix but not the waves themselves. The first and second differentiators heighten or filter through the temporal content of the higher-order differentiations, that is, they serve as a band pass filter unit to accent the time keying of the neutrinic and mindfield portions or aspects of the signal. When these time spikes are then introduced across the body as pulsed voltages, they are modulated directly on the dendrite firings of the brain and nervous system, providing direct and pulsed modulation of the neutrinic and mindfield component channels of the mind-brain-consciousness-life loop itself. Thus the neurophone directly inputs information into the brain and nervous system, bypassing all the normal sensory systems that lie between the mind-brain loop and the outside environment.

Indeed, a similar development has been reported by two University of Missouri researchers. Dr. Donald York, a neurophysiologist, and Dr. Thomas Jensen, a speech pathologist, have recently reported identifying and decoding twenty-seven words and syllables in specific brain wave patterns and correlating these electroencephalographic patterns with both the spoken word and the silently thought word in about forty subjects. At present, Dr. York and Dr. Jensen are programming a computer with a brain wave vocabulary, to monitor and read the EEG of a stroke victim's brain and help stroke victims who have lost their powers of speech to communicate.

While all such developments can obviously be misused, their potential for assistance to mankind is enormous. Indeed, my own premise that brain linkage is possible and can be used to advance all mankind to the sixth stage of species evolution is slowly being proven. It is within our reach now to develop a multi-channel communications system that will directly link brains, minds, and consciousness into a single functional being; and I have already pointed out that the psychokinetic power of the emergent entity increases exponentially with the number of linked stages. The normal corpus callosum linkage of the two cerebral halves in each human head proves that brains, minds, and beings can be linked and integrated into one.

The Excalibur Briefing details the basic concepts of the theory involved in hyperspace-virtual state engineering, and thus in the engineering of mind-life-consciousness links to multicellular organisms here on earth. Pat Flanagan's neurophone proves that complex information can be implanted directly in the brain and mind, bypassing the normal isolating barriers. Drs. York and Jensen have shown that the brain wave does indeed contain recognizable, decodable information analogues. Lisitsyn has already reported measuring the number of independent channels involved and has documented the importance of the clipped brain waves as opposed to the sine-wave content. If we have the vision, we can still develop a brainlink and rapidly apply it to link sixty to one hundred persons, thereby opening a conscious, multidimensional "eye" for the collective human unconsciousness, gently rousing ZARG into wakefulness, and fulfilling human destiny without undergoing the psychotronic Armageddon planned by the Soviets. The time is short, but it can still be done if the funding, the vision, and the team can be established on a crash priority basis.

SOVIET "WOODPECKER" SIGNALS

But, as I have said, all this can be used for the detriment of mankind as well as for its benefit. Unfortunately, it appears that the Soviet Union has chosen to weaponize the effects on a global scale. A prime example is the "woodpecker" signals emanating from the USSR in the 5-30 megahertz region and interfering with communications around the earth.

These complex woodpecker signals appear to originate from two or three dozen powerful Soviet transmitters, each with a power estimated as high as 40 megawatts (Figure 12).

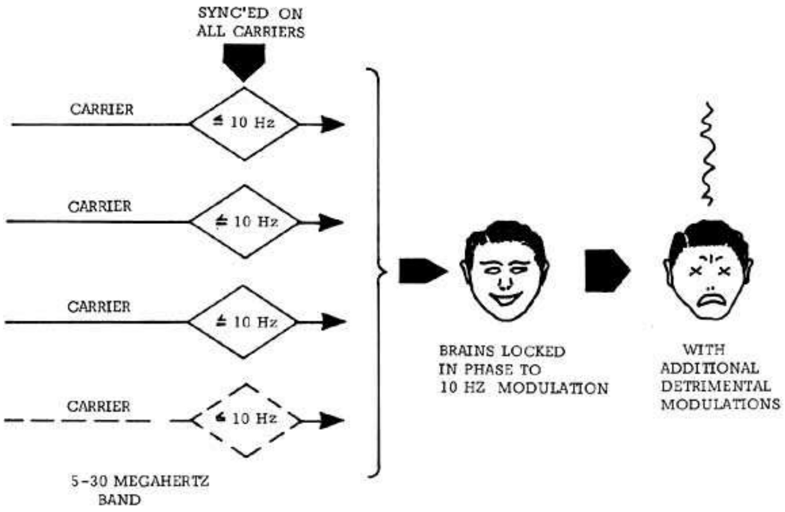

Figure 12. The Soviet "woodpecker" signals

The pioneering experimental measurements of these signals by Dr. Robert Beck and William Bise have shown just how deadly a potential may be possessed by the signals. These measurements have been performed in Eugene and Portland, Oregon; Los Angeles, California; Huntsville, Alabama; and several other locations. They have been particularly significant on the West Coast in and around Eugene, Oregon because of the presence of a direct current transmission line several hundred miles long, which has acted as a long wire antenna, picking up the signals and rebroadcasting them with appreciable gain in the vicinity.

Typically the signals may be found on, say, sixteen different carriers between 10 and 20 megahertz. Twelve of the carriers may appear normal, with normal side-bands, and the other four may have the carrier and both sidebands suppressed but still show the biologically significant modulation (for example, 10 hertz). On all sixteen channels a strong 10 hertz modulation may appear, all perfectly time synchronized and in phase channel-to-channel. The received signal from one of these carriers may be twenty-five to thirty times as strong as the earth's back-ground magnetic field, which is oscillating at approximately 7.0 to 7.5 hertz. Other complex modulation frequencies, many of them changing, are present on the various channels.

Normally, the brainwaves of mammals in an area are gently entrained by the normal oscillations of the earth's magnetic field and are oscillating along at, say, 7.5 hertz. Under continuous radiation from the Soviet woodpecker signals, a percentage (say 30 percent) of the mammalian brains may be captured by the 10 hertz modulation on the overpowering Soviet signals. These captured brains are now in forced oscillation and are riding along in phase with the 10 hertz Soviet modulation. In other words, a certain percentage of the human brains in an area will phase lock to the 10 hertz modulation, and that will effectively lock into those brains the sixteen carrier frequencies with their concomitant frequency mixes and complex signal modulations. If the signals are made much more powerful (for example, by focusing them onto one particular area or increasing the power of the transmitters), then a much higher percentage of brainwave entrainment can be accomplished in the targeted area. By simply adding on the desired modulation patterns, the Soviets can now pump material directly into the mind-brain-consciousness-life loops of the entrained brains. The weapon implications are enormous: raw emotion such as sheer terror or panic can be transmitted; death or disease patterns of all kinds can be transmitted; informational content (thoughts and ideas) can be impressed directly into the captured brains and minds and processed as if originating inside each brain itself. Indeed, as long as the Soviets keep the effects rather gentle, an entire population can be subtly influenced without governmental notice. Certain specific ELF frequencies can rapidly disable or even kill, and for wartime use those frequencies can be directly implanted in the captured brains themselves by the woodpecker signals. It appears that 1984 came a little early, and one can perhaps now understand Brezhnev's strange 1975 proposal to the SALT negotiations that we should also consider outlawing the development of new weapons more frightful than the mind of man has ever conceived.

The signals also have direct application against strategic targets other than personnel. For example, by incorporating the correct nested modulation pattern, the positron kindling (electron squelching) pattern can be impressed. By beaming through the earth (with a circularly polarized signal or with other complex modulation forms), a beam with a strong neutrinic component, which will squelch electrons and which will interact with great preference for strong voltages or strong magnetic fields, can be directed at a target complex. In the targeted area aircraft and communications systems can be disabled; automobile, truck, and other ignition systems will fail; power transformers and large motors and generators will fail; power transmission systems with their associated high voltages will fail; and so on. Even independent, self-contained emergency power equipment, such as at military installations and strategic missile sites, will fail to operate. Further, to some of these signals, the earth and the sea are as transparent as glass, and so underwater nuclear submarines can be electrically disabled on station before their missiles can be fired, while strategic missile complexes can be electrically disabled before their deadly nuclear missiles can be hurled toward their distant targets. If some submarines and land complexes succeed in firing missiles before being disabled, these missiles can be electrically dudded and scrubbed from the sky over Soviet territory while still far from their intended targets.

So, as stated at the end of *Excalibur*, the players are on stage and in place. The final act begins. One thing is sure—mankind will never be the same again after seeing the play!

Return to top of page

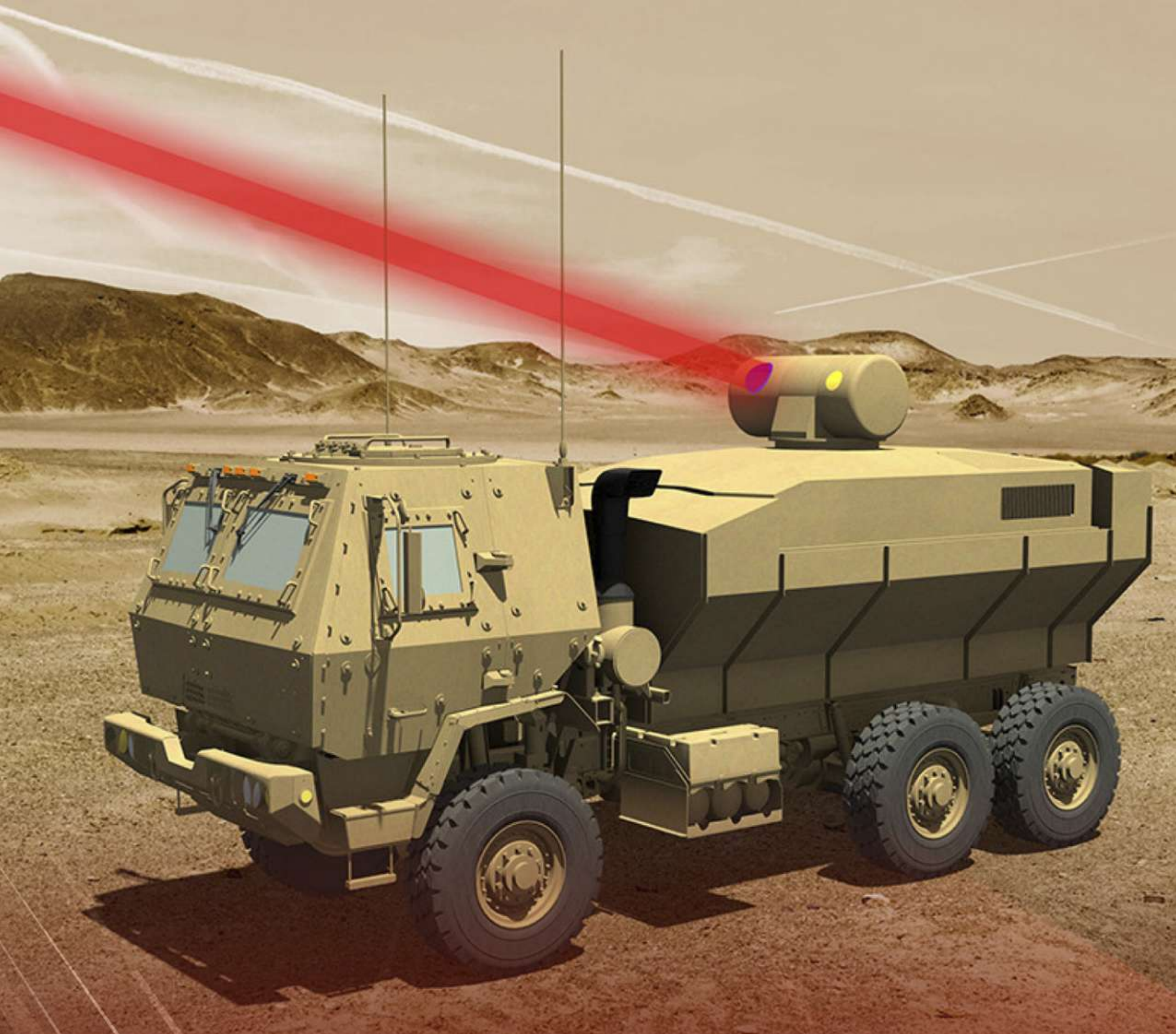

LASER WEAPON SYSTEMS

TURNING UP THE HEAT WITH SPEED-OF-LIGHT RESPONSE

LOCKHEED MARTIN

We never forget who we're working for®

LASER WEAPON SYSTEMS

Lockheed Martin has specialized in laser weapon system development for 40 years, with advancements in areas such as precision pointing and control, line-of-sight stabilization and adaptive optics – essential functions in harnessing and directing the power of a laser beam – and in compact, robust, spectrally beam-combined fiber laser devices that provide unmatched performance.

Lockheed Martin's directed energy (DE) laser program draws upon our proven expertise across a broad range of capabilities, as well as our spirit of innovation. Our experience includes:

- The beam-control/fire-control system for a megawatt-class laser that destroyed a ballistic missile in flight
- High-power spectrally beam-combined fiber lasers that provide the most efficient conversion of platform prime power into lethal power on target
- Unique SWIR component capabilities that significantly extend the ISR (intelligence, surveillance, and reconnaissance) range of the DE laser system
- Widespread weapons-system and platform integration experience across all DoD services

REVOLUTIONARY TECHNOLOGY

Laser weapons are a revolutionary technology because of the advantages of speed, flexibility, precision and low cost per engagement that are only possible with lasers.

These advantages apply to stand-alone DE laser systems as well as to weapon systems that combine DE and kinetic energy capabilities. In these cases, DE operates as a force multiplier, enabling the warfighter to counter a growing range of emerging threats.

DE weapons are characterized by:

- Very deep magazine
- Extremely low cost per engagement
- Speed of light delivery

These strengths mean that they are well-suited to countering large numbers of inexpensive, highly maneuverable threats that might otherwise exhaust the magazines of our current defensive kinetic energy (KE) weapons. By using DE and KE weapons together, the warfighter will be able to neutralize emerging swarming threats while reserving our most capable kinetic weapons to defend against our adversaries' largest and most hardened threats. DE weapon systems support multi-mission scenarios, and can be readily augmented by communications functions, as well as unprecedented intelligence, surveillance and reconnaissance range, and precision.

In addition, Lockheed Martin's directed energy weapons bring some important and unique advantages to the warfighter:

- A robust and highly-maintainable laser architecture that is designed to minimize life-cycle cost and to maximize up-time
- Output laser power that can be rapidly adjusted between low and maximum power to support disrupt, disable, and destroy capabilities
- Highly parallel laser architecture that supports graceful degradation by eliminating almost all single points of failure

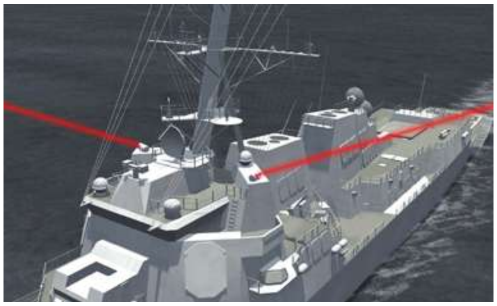

- Highest system efficiency demonstrated in any DE laser system; minimizing size weight and power requirements for DE platforms such as Army and Marine Corps ground vehicles, Navy DDG and LCS ships, and SOCOM AC-130
- Highest lethality laser to minimize engagement time

LOCKHEED MARTIN IS ADVANCING AND DEMONSTRATING A RANGE OF LASER WEAPON SYSTEM TECHNOLOGIES:

ADAM (Area Defense Anti-Munitions)

- Autonomous operations for rocket threats; accepts external sensor cue when required; capable of continuous operation
- Successfully engaged constrained and free-flying rockets, an unmanned aircraft systems (UAS) target in flight, and a small boat
- Affordable commercial-off-the-shelf (COTS) based system with very low cost-per-kill; deep magazine; scalable precision effects
- Capable of close-in defense (1-4 km)

ALADIN (Accelerated Laser Demonstration Initiative)

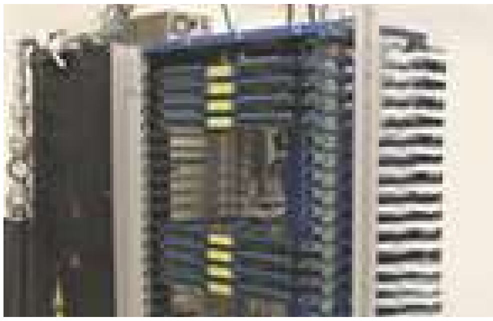

- 30-kilowatt laser made by combining many fiber lasers into a single, near-perfect-quality beam of light
- Uses approximately 50 percent less electricity than alternative solid-state laser technologies
- Spectral Beam Combining sends beams from multiple fiber laser modules, each with a unique wavelength, into a combiner that forms a single, powerful, high-quality beam

ATHENA (Advanced Test High Energy Asset)

- Represents highest power level documented by a laser weapon system of this type, while retaining excellent beam quality and electrical efficiency
- First field testing of an integrated 30-kilowatt single-mode fiber laser weapon system prototype
- Uses the proven high-energy laser weapon system architecture from our ADAM system, and incorporates the 30-kilowatt ALADIN laser

ABC Turret (Aero-adaptive Aero-optic Beam Control)

- Prototype turret with the ability to fire in any direction mounted on tactical aircraft
- First turret to demonstrate a 360-degree field of regard for laser weapon systems on an aircraft flying near the speed of sound
- Validated performance with nearly 60 flight tests conducted in 2014 and 2015 using a business jet as a low-cost flying test bed

RELI (Robust Electric Laser Initiative) for Army HELMTT (High Energy Laser Mobile Test Truck) Demonstrator

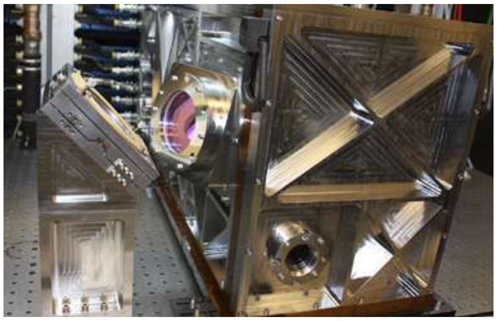

- Modular 60-kW laser in development for the U.S. Army
- Proven affordable weapon architecture that supports size, weight and power constraints for air, sea and land platforms
- Scalable design combines multiple kilowatt lasers to attain various weapon power levels
- Modularity results in a reliable and low maintenance laser system that minimizes single points of failure

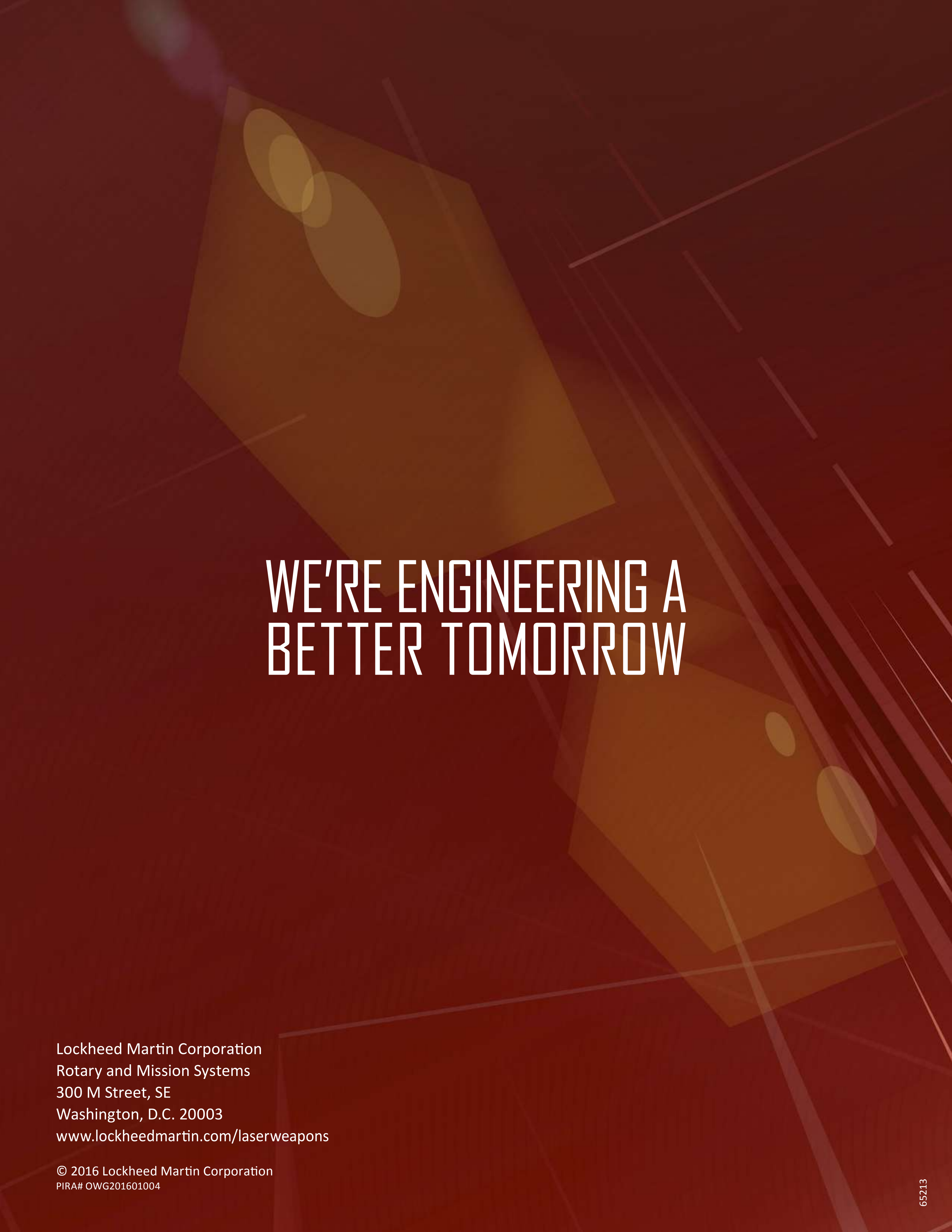

WE’RE ENGINEERING A BETTER TOMORROW

Lockheed Martin Corporation
Rotary and Mission Systems
300 M Street, SE
Washington, D.C. 20003
www.lockheedmartin.com/laserweapons

© 2016 Lockheed Martin Corporation
PIRA# OWG201601004

Axial Vircator for Electronic Warfare Applications

Libor DRAŽAN, Roman VRÁNA

Dept. of Radar, University of Defence, Kounicova 65, 662 10 Brno, Czech Republic

libor.drazan@unob.cz, roman.vrana@unob.cz

Abstract. *This paper deals with a high power microwave generator with virtual cathode – vircator in axial release for electronic warfare applications. The classification of directed energy weapons microwave (DEWM) is introduced together with basic block diagrams of a particular class of DEWM. In the paper, methods for designing vircator pulsed power supply, axial vircator structure, measurement methods and experimental results are presented. The vircator in electromagnetic ammunition is powered by magneto-cumulative generator and in weapons for defense of objects (WDO), it is powered by Marx generator. The possible applications of a vircator in the DEWM area are discussed.*

Keywords

Electronic warfare, directed energy weapons microwave, high power microwave, magneto-cumulative generator, Marx generator, pulse forming line, vircator.

1. Introduction

Electronic warfare (EW) is defined as the art and science of preserving the use of the electromagnetic spectrum for friendly use while denying its use to the enemy. Electromagnetic spectrum is, of course, reaching from DC to light (and beyond). Thus, electronic warfare covers the full radio frequency spectrum, the infrared spectrum, the optical spectrum, and the ultraviolet spectrum.

Electronic warfare has classically been divided into:

- Electronic warfare support (ES), which includes the receiving part of EW.
- Electronic attack (EA), which includes jamming, chaff, flares used to interfere with the operation of radars, military communication, and head-seeking weapons. Over the last years, EA also includes antiradiation weapons (ARW) and directed energy weapons (DEW).
- Electronic protection (EP), which is directed into design or operation of radars or communication systems to counter the effect EA.

A detailed classification of EW can be found in [1].

Today the directed energy weapons (DEW) are significant in terms of the electronic attack (EA). Considerable financial resources are directed especially to the research of DEW working in microwave band (directed energy weapons microwave − DEWM), which derive their benefits from incidence of electromagnetic pulses (EMP) on electronic devices. The EMP effect was first observed during the early testing of high altitude airburst nuclear weapons [2]. The effect is characterized by production of a very short (hundreds of nanoseconds) but intense electromagnetic pulse, which propagates away from its source with ever diminishing intensity, governed by the theory of electromagnetism. The EMP is in effect an electromagnetic shock wave. This pulse of energy produces a powerful electromagnetic field, particularly within the vicinity of the weapon burst. The field can be sufficiently strong to produce short lived transient voltages of hundreds of volts to kilovolts on exposed electrical conductors, such as wires or conductive tracks on printed circuit boards.

At present, a high intensity EMP can be successfully generate without using nuclear weapons. This fact is very significant especially for applications in the EW area. It is this aspect of the EMP effect which is of military significance, as it can result in irreversible damage to a wide range of electrical and electronic equipment, particularly computers and radio or radar receivers. Depending on the electromagnetic hardness of the electronics (a measure of the equipment's resilience to this effect) and the intensity of the field produced by the weapon, the equipment can be irreversibly damaged or even electrically destroyed. The damage inflicted is not unlike the damage experienced by exposure to close proximity lightning strikes, and may require complete replacement of the equipment, or at least substantial portions of it [3], [4].

2. Directed Energy Weapons Microwave

Directed energy weapons microwave (DEWM) generate very high-powered electromagnetic impulse (EMP) in the microwave band. From the generated signal bandwidth point of view, there are two classes of generators for the design and construction of DEWM in order to produce electromagnetic field with sufficient power:

- Narrowband – very similar to the radar transmitters; usually generate a modulated harmonic signal. This class is called the high power microwave (HPM).
- Ultra-wideband (UWB) – generates high power videosignal illuminated by a broadband antenna.

The application of directed energy weapons microwave (DEWM) determines their design. There are two groups of directed energy weapons microwave:

- Single shot DEWM,
- Repetitive pulsed power systems.

The electromagnetic weapon modes closely relate to the mission of these weapons. Electromagnetic ammunition usually uses single shot DEWM while weapons for defense of objects (WDO) usually use repetitive pulsed power systems. The typical pulse repetition frequency of DEWM is nearly one thousand pulses per seconds. Recently, there has been a trend of increase in the pulse repetition frequency beyond tens of thousands pulses per seconds. The actual DEWM construction depends on the concrete supposed combat use, on the required measure, range and so on. The following text will be dedicated to experimental verification of some technology usable in narrowband DEWM or HPM. Questions regarding classification, construction and possible use of generators for DEWM are answered in more detail in [4], [5] and [6].

2.1 Basic Block Diagram of DEWM

HPM design depends on the specific assumed application of DEWM. There are difference approaches to the design and realization of electromagnetic ammunition and to the design and realization of weapons for defense of objects (WDO). These differences mostly affect the choice of the power sources and the construction of modulators. In electromagnetic ammunition, a highly limiting factor is the demand for low weight and small proportions. This fact also sharply limits the choice of the power microwave tube.

Electromagnetic ammunition usually uses single shot DEWM. The single shot high-energy high-voltage pulsed source has to fulfill hard requirements on the high amount of the delivered energy with respect to the source volume [MJ/m^3]. The best solution of this requirement is a magneto-cumulative generator that reaches up to 8000 MJ/m^3. A block diagram of a HPM generator in electromagnetic ammunition is shown in Fig. 1.

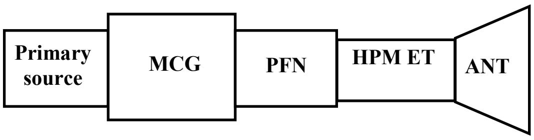

Fig. 1. Block diagram of a HPM generator in electromagnetic ammunition.

In the block diagram, the primary source feeds the stator coil of a magneto-cumulative generator (MCG), which creates a powerful magnetic field in the surrounding fittings filled with an explosive. At the moment of achieving the primary current peak value, the explosive is shooting. The explosion causes expansion of fitting that performs compression of the magnetic field inside the stator coil of MCG. Compression of magnetic field causes dramatic growth of the current flowing through MCG (hundreds of kiloamperes). The MCG output is connected to a pulse forming network that provides transformation of heavy current to an impulse high voltage (hundreds of kilovolts), which feeds a high power microwave electron tube (HPM ET). The generated microwave is emitted through antenna (ANT) to the space.

A block diagram of a HPM generator in weapons for defense of objects (WDO) is shown in Fig. 2.

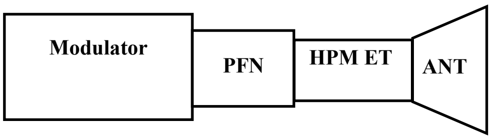

Fig. 2. The block diagram of HPM generator in Weapons for Defense of Objects (WDO).

In weapons for defense of objects (WDO), the HPM generator works similarly as the HPM generator in electromagnetic ammunition. Since WDOs usually generate power microwave pulses repetitively, a modulator with this capability has to be used instead of MCG. Modulators with Marx capacitor banks, Tesla transformers or transductors with fast opening switches are used most often. In this class of DEWM, other types of high power microwave electron tubes (HPM ET) are also used. They may generate microwave pulses with high pulse repetition frequency.

3. Energy Sources for Directed Energy Weapons Microwave

The electromagnetic pulse sources for DEWM are entirely based on energy acquired from the chemical energy of explosion that compresses the magnetic flux made by coil (MCG) or by discharging a capacitor bank. These two possibilities are the best solutions for generation of pulses with sufficient amount of energy in a short time. Other methods were used to generate pulses in the beginning of DEWM research or in laboratory.

3.1 Helical Magneto-Cumulative Generator

MCGs are entirely based on the energy acquired from chemical energy of explosion that compresses magnetic flux made by coil. MCG physical principles, classification and applications are described in more detail in [7]. The main part of research was focused on the possibility of using a magneto-cumulative generator for electromagnetic ammunition power supply. For that reason, 4 versions of low power MCGs were designed, realized and tested in order to verify their characteristics, construction principle, explosive initialization method and verifying a measurement method.

The tested magneto-cumulative generator in Fig. 3 consists of an aluminous cylindrical stator, inside is a 16 screw helical coil from 2 mm copper wire, a copper armature tube with 38 mm outer diameter and 4 mm wall thickness filed by octogen (explosive). The explosive is electrically initiated. The load coil represented by a copper strip in shape of a loop connects the stator with the armature. In the right side of stator, there is a crowbar switch made of a brass plate to perform disconnecting of the primary current source. An insulator circle is responsible for separation of the stator and the armature.

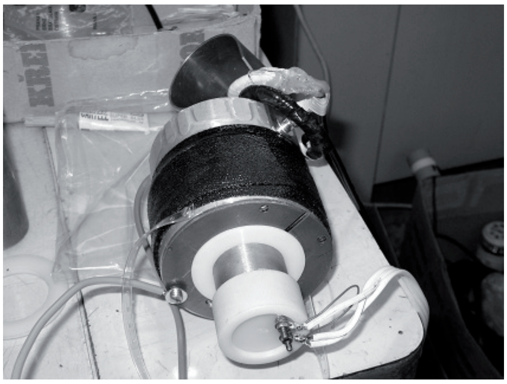

Fig. 3. Low power MCG.

Low power magneto-cumulative generator characteristics:

C_1 = 190 µF − primary capacitor capacity

U_0 = 800 V − voltage of capacitor

L_G = 13.1 µH − stator coil inductance

R = 0.0175 Ω − stator coil resistance

T_0 = 3.13×10^{-4} s − primary current period

I_m = 2750 A − maximum primary current

T_d = 73 µs − explosive initiation delay

Fig. 4 presents the low power magneto-cumulative generator output current time history calculated from the voltage measured by Rogowski coil. The primary current $I_m = 2750$ A gives the generator peak output current $I_{out} = 27$ kA. The detailed test description and measured results can be found in [8], [9], [10] and [11]. The purpose of the mentioned tests was:

- Verify the method of generating a high energy pulse by magnetic flux compression using explosive;
- Solve the primary current magneto-cumulative generator power supply;
- Find the peak primary current and an explosive initiation synchronization method;
- Find a high current measurement method.

All the requirements were met.

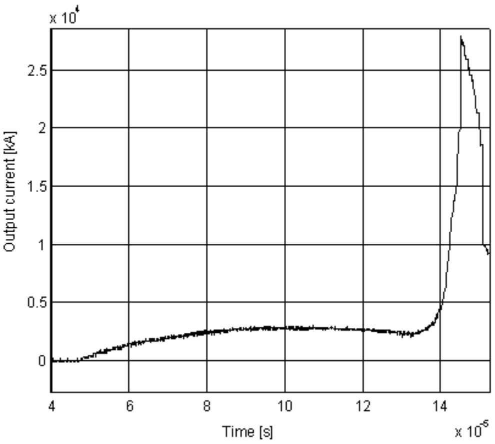

Fig. 4. Low power magneto-cumulative generator output current time history.

To use a magneto-cumulative generator for single shot high power high voltage pulse generator, it is necessary to obtain a high-level output current. Because of that reason, a middle power MCG was designed and realized. It is a two-stage helical magneto-cumulative generator using a dynamic transformer to increase the output voltage. In Fig. 5, there is a schematic plot of the generator's first stage. The magneto-cumulative generator power supply primary current, the initiation of the explosive and the output high current measure was analogical to the low power magneto-cumulative generator described above.

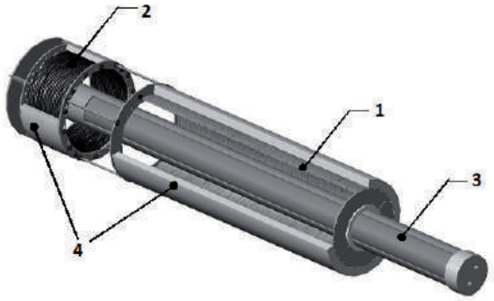

1 – First stage coil, 2 – Load coil, 3 – Armature with explosive, 4 – Insulator

Fig. 5. Schematic drawing of first stage of middle power MCG.

Fig. 6 shows a photo of the middle power magneto-cumulative generator's first stage prepared to the test.

Magneto-cumulative generator characteristics:

C_1 = 125 µF − primary capacitor capacity

U_0 = 5000 V − voltage of capacitor

L_G = 323 µH − stator coil inductance

T_0 = 3×10^{-3} s − primary current period

I_m = 2650 A − maximum primary current value

L_L = 300 nH − load inductance

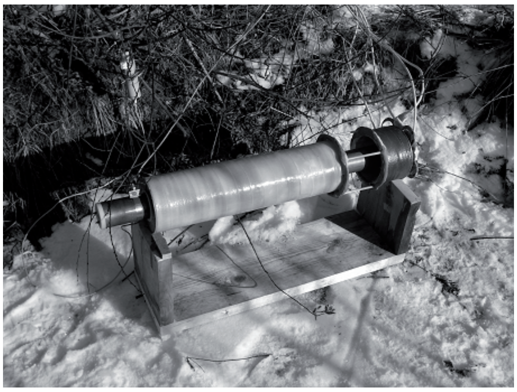

Fig. 6. The picture of middle power MCG prepared to testing.

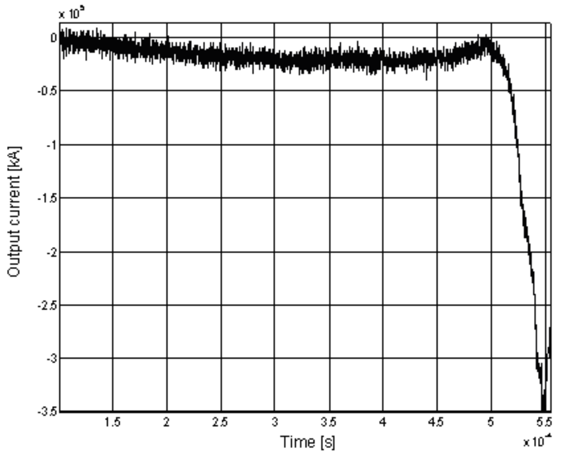

Fig. 7. Middle power magneto-cumulative generator output current time history.

In Fig. 7, the generator output current time history measured during test is shown. The primary current $I_m = 2650$ A gave the generator peak output current $I_{out} = 350$ kA.

The next step in the magneto-cumulative generator research is the implementation and verification of a dynamic transformer. The dynamic transformer's role is an increase of the output voltage to about 70 kV. Magneto-cumulative generator's second stage supplies a load coil through an electric breaker. The beaker interrupts the current in the load circuit in a very short time to enable transfer of the energy from the magnetic field to the high voltage pulse. These activities will be done in the next step of the research and development. The further trends are described in [12].

3.2 Pulsed Power Supply With Capacitive Storage of Energy

For some pulse applications in weapons for defense of objects (WDO), it is desirable to couple the Marx generator directly to the vacuum diode; however, the pulse rise time is then limited by the Marx inductance and capacitance, and the impedance of the generator is greater than typically several tens of ohms. In order to produce short, fast-rising, low-impedance beam outputs, it is customary to use the Marx to charge a pulse-forming line (PFL). Although the PFLs may be constructed in a variety of shapes (strip, coaxial, radial, etc.), they are typically used in only two types of circuits— the simple transmission line, and the double, or Blumlein line. In contrast to the simple transmission line, an alternate circuit invented by A. D. Blumlein is capable of producing an output pulse into a matched load that equals the charge voltage. A cylindrical version of the Blumlein circuit fabricated by our team is represented in Fig. 8. It consists of three coaxial cylinders with the intermediate cylinder being charged by the Marx generator. The Marx generator is shown in Fig. 9. The center cylinder is connected to the outer grounded cylinder by an inductor. Ideally the inductor acts as a short during the charge cycle, and then as an open for the short duration of the output pulse [13]. Pulsed power supply parameters are shown in Tab. 1.

Marx generator		Pulse forming line	
Number of stage	n=18	Impedance	9,1 Ω
Capacity of 1stage	100 nF	Length	1 m
Input voltage	U_{in}=25 kV	Pulse duration	60 ns
Output voltage	U_{out}=450 kV	Working medium	water

Tab. 1. Pulsed power supply parameters.

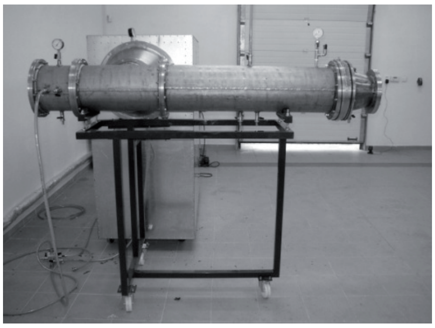

Fig. 8. Cylindrical version of the Blumlein circuit.

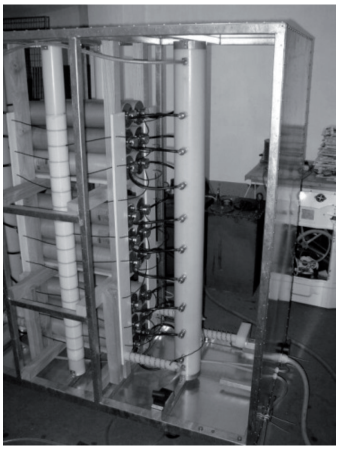

Fig. 9. Marx generator.

4. High Power Microwave Generator

A wide range of HPM devices exists. Relativistic klystrons, magnetrons, slow wave devices, reflex triodes and vircators are all examples of the available technology base [5]. From the perspective of an electromagnetic weapon or warhead designer, the device of choice will be the vircator at this time. The vircator is mainly a one shot device capable of producing a very powerful single pulse of radiation, yet it is mechanically simple, small and robust, and can operate over a relatively broad band of microwave frequencies. The physics of the vircator tube are substantially more complex than those of the preceding devices. The fundamental idea behind the vircator is that of accelerating a high current electron beam against a foil or a grid anode. Many electrons will pass through the anode, forming a bubble of space charge behind the anode. Under the proper conditions, this space charge region will oscillate at microwave frequencies. If the space charge region is placed in a resonant cavity which is appropriately tuned, very high peak powers may be achieved. Conventional microwave engineering techniques may then be used to extract the microwave power from the resonant cavity. Because the frequency of oscillation is dependent upon the electron beam parameters, vircators may be tuned or chirped in frequency, where the microwave cavity will support appropriate modes. Power levels achieved in vircator experiments range from 170 kW to 40 GW over frequencies spanning the decimeter and centimeter bands [5].

4.1 Vircator Operation Fundamentals

The basic idea of the vircator is to accelerate a dense flush of an electron beam against a grid or a foil anode. Plenty of electrons pass through the anode and form a region of a space charge behind the anode called ”virtual cathode“. This region of a space charge at corresponding conditions can oscillate in a region of microwave frequencies. It is possible to tune the vircator in a broad band of frequencies using only a change of a space charge density. There is no necessity to have an external magnetic field for a correct vircator function. In Fig. 10, an axial vircator is shown. Electron beam passes through the foil or the grid anode. Microwave power is brought out axially, too. Frequency changes appear in vircator with standard geometry when the distance between the anode and the cathode gets smaller due to filling the working space by plasma. Efficiency of a standard geometry vircator is ordinarily about ones percent.

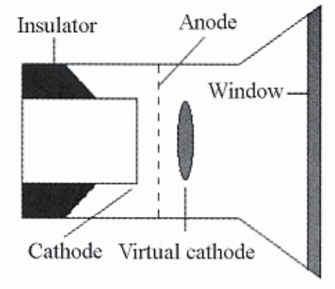

Fig. 10. Schematic drawing of axial vircator.

Despite its low efficiency, vircator is very attractive for army applications because it is very simple to be made, it is a compact device and there it is no need to have an external magnetic field. For more information about construction and properties of various types of vircators you can see e. g. [5], [13] and [14].

4.2 Mathematical Description of Vircator Operation

For microwave frequency generation, it is necessary to meet a number of conditions relevant to the power supply and geometric proportions of a vircator electrode. Determination of these conditions results from [15] and we use simplified geometry displayed in Fig. 11 to derive the mathematical characterization.

The most common case is that a pulse duration τ is much longer than a beam transit time across cavity $\tau >> L/C$.

We assume that an external axial magnetic field obstructs transverse electron motion. The space charge of the beam makes a negative potential energy $e\Phi$ in drift space which breaks the electrons. If the space charge potential reaches the value of accelerating voltage (electrons are stopped in beam), the beam with bigger current cannot expand. There is a value of current which causes stopping of electrons. This value is called the vacuum critical current. There has to be the above mentioned critical current to make the vircator generate microwave oscillations.

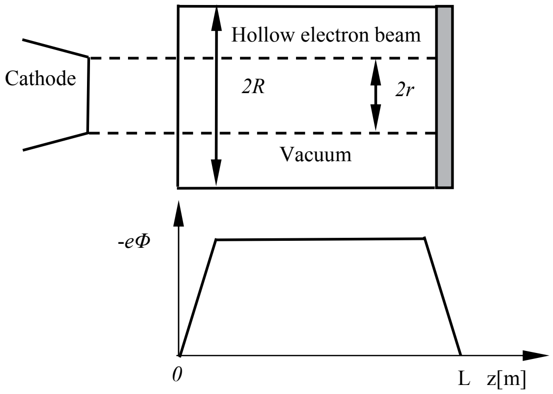

Fig. 11. Simplified geometry of axial vircator.

We suppose a hollow beam; whose charge is concentrated in a thin layer with radius r. Behind the transition space (which is comparable to the radius of chamber – backward conductor), there is nearly homogenous potential. It is possible to determine the potential quantity with consideration that it is a coaxial capacitor and its inner cylinder is saturated by beam charge.

Capacity of the coaxial capacitor is given by

$$C = \frac{2 \cdot \pi \cdot \varepsilon_0}{\lg\left(\frac{R}{r}\right)} \qquad (1)$$

where ε_0 is the vacuum electric permittivity, R is the chamber radius, r is the beam radius.

The potential is given by the equation

$$\Phi = \frac{Q}{C} \tag{2}$$

where Q is the charge linear density, C is the capacity per length unit. We can define the charge linear density as

$$Q = \frac{I}{v} \tag{3}$$

where v is an electron module velocity and the current is defined by the equation

$$I = e \cdot n \cdot v \cdot S \tag{4}$$

where e is the charge of an electron, n is the linear electron density, S is the profile of the electron beam.

When (1) and (3) are substituted into (2), we can obtain expression for the potential in the form

$$\Phi = \frac{Q}{C} = \frac{I \cdot \lg\left(\frac{R}{r}\right)}{2 \cdot \pi \cdot \varepsilon_0 \cdot v} . \tag{5}$$

The electron velocity v is associated with its initial energy and potential by the energy preservation law

$$m_e \cdot \gamma_0 \cdot c^2 + e\Phi = m_e \cdot \gamma \cdot c^2 \tag{6}$$

where γ_0 is the initial relativistic factor of the beam, γ is the relativistic factor inside a system reduced due to potential.

The relativistic factor inside a system can be expressed by the equation

$$\gamma = \gamma_0 + \frac{e\Phi}{m_e \cdot c^2} \tag{7}$$

where m_e is the electron mass and c is the speed of light. The relativistic factor can also be expressed by

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} . \tag{8}$$

The electron velocity in drift space v can be specified from following equations:

$$\gamma^2 = \left(\gamma_0 + \frac{e\Phi}{m_e \cdot c^2}\right)^2 , \tag{9}$$

$$v = c \cdot \sqrt{1 - \frac{1}{\left(\gamma_0 - \frac{e\Phi}{m_e \cdot c^2}\right)^2}} . \tag{10}$$

Substituting (10) into (3) and multiplying by $(e \cdot v) / (m \cdot c^3)$, we can get

$$F(\Phi) = \frac{e \cdot \Phi}{m \cdot c^2} \cdot \sqrt{1 - \frac{1}{\left(\gamma_0 - \frac{e\Phi}{m_e \cdot c^2}\right)^2}} = \frac{e \cdot I \cdot \lg\left(\frac{R}{r}\right)}{2 \cdot \pi \cdot \varepsilon_0 \cdot m \cdot c^3} . \tag{11}$$

The function $F(\Phi)$ has its peak value under the conditions describes by

$$\frac{e \cdot \Phi}{m \cdot c^2} = \gamma_0 - \gamma^{\frac{1}{3}} . \tag{12}$$

The evaluation is written in the form

$$F(\Phi) = \sqrt{\left(\gamma^{\frac{2}{3}} - 1\right)^3} = \frac{e \cdot I \cdot \lg\left(\frac{R}{r}\right)}{2 \cdot \pi \cdot \varepsilon_0 \cdot m \cdot c^3} . \tag{13}$$

There is no possibility to find the solution of (11) for bigger values of function $F(\Phi)$. According to (13), the critical vacuum current is

$$I_{cr} = \frac{2 \cdot \pi \cdot \varepsilon_0 \cdot m \cdot c^3}{e \cdot I \cdot \lg\left(\frac{R}{r}\right)} \sqrt{\left(\gamma^{\frac{2}{3}} - 1\right)^3} . \tag{14}$$

When we substitute the known constants (like $\pi = 3.1415$), we obtain the equation

$$I_{cr} = \frac{8.5}{\lg\left(\frac{R}{r}\right)} \sqrt{\left(\gamma^{\frac{2}{3}} - 1\right)^3} \tag{15}$$

where I_{cr} is the critical vacuum current [kA].

For vircator implementation, energy of electrons of about 500 keV is considered, which corresponds to the relativistic factor of $\gamma_0 = 2$. Thus planar diode current limited by space charge follows the three-half Child-Langmuir law

$$I \approx \frac{U^{\frac{3}{2}} \cdot S_k}{d_{ka}^{\ 2}} \tag{16}$$

where U is the voltage between anode and cathode, S_k is the surface of cathode, d_{ka} is the distance between anode and cathode.

An important characteristic of a vacuum planar diode is its impedance, which is given by

$$Z = 1.36 \cdot 10^5 \cdot U^{-\frac{1}{2}} \cdot \frac{d_{ka}^2}{r_k^2} \tag{17}$$

where r_k is the radius of the cathode.

The vacuum critical current is the maximal current that can distribute itself through vacuum. It means it is possible to measure the same current until the distance is bigger than the diameter of beam. It is just the same value when the potential of space charge of electron beam is equal to the accelerating voltage. That is the limit where the virtual cathode is generated and the vircator starts working. The generated frequency depends on the plasma frequency of the electron beam. The plasma frequency is determined by the electron beam current density. The plasma frequency is the frequency of electrons in space charge field oscillations. The electrons can oscillate due to the influence of repulsive force which takes effect between particles with the same charge. The plasma frequency depends on the concentration of electrons in the beam. The current density is given by the number of charges which are able to pass through the surface of 1 m^2 per one second. Thus we can calculate the current density i [Am^{-2}]

$$i = e \cdot n_V \cdot c_e \tag{18}$$

where e is the electron charge, n_V is the electron volume density and c_e is the electron velocity.

Plasma frequency of the electron beam is approximately given by

$$f_{pe} = 9 \cdot 10^7 \cdot \sqrt{i}\ . \tag{19}$$

4.3 Axial Vircator Design

Following the theoretical assumptions given above, the axial vircator with grid anode was fabricated. The structure of the grid anode enables repeated operation in contradistinction to vircator with a foil anode. The cathode diameter $r_k = 17.5$ mm and the working chamber radius $R = 40$ mm were chosen in order to decrease the value of the vacuum current. The critical vacuum current is $I_{cr} = 4.63$ kA in this configuration. The vircator impedance matching to the forming line is a very important requirement for vircator's correct operation. Because of that, a cathode with a possibility to change the distance d_{ka} between cathode and anode is fabricated. The cathode is made as a cylindrical carbon block. The anode grid is made from pyrolitic carbon.

In Fig. 12, there is section picture of the built-up axial vircator with the possibility to the change distance between anode and cathode. In Fig. 13, there is a view of the fabricated vircator. The vircator is fed by a pulse generator formed by Marx generator and pulse forming line described in Section 3.2. The connection between the axial vircator and the output of pulse forming line is shown in Fig 14. Vircator voltage is measured by a cylindrical capacitive probe, which is a part of pulse forming line output circuit. Vircator current is measured by Rogowski coil placed in the outer cylinder of the pulse forming line. Vircator voltage time history is displayed in Fig. 15 and vircator current time history is displayed in Fig. 16.

Compared to other types of microwave tubes, the microwaves generated by an axial vircator cover a relatively wide frequency band. In Fig. 17, the frequency spectrum of the signals generated by the realized axial vircator is shown. The dominant frequencies of the generated microwave signals are placed in the band of 0.5 to 1.5 GHz. The frequency spectrum was acquired on the basis of information measured by a digital oscilloscope Tektronix TDS 7704B connected to broadband antenna EMCO 3115. For the measurement the microwave output power of the vircator, a measuring device was designed and made, using calorimetric method for the power measurement. A detailed power sensor description can be found in [18]. A more detailed description the of axial vircator research results is given in [6], [14], [16] and [17].

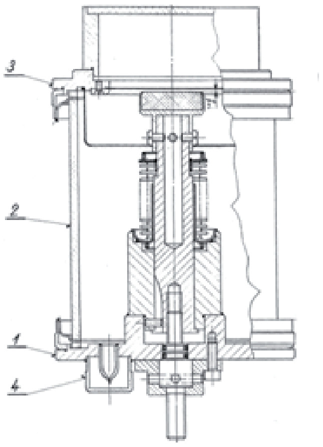

1 cathode contact ring, *2* ceramic insulator, *3* anode contact ring, *4* pipe for vacuum pumping shelter

Fig. 12. Section picture of built-up axial vircator with possibility to change distance between anode and cathode.

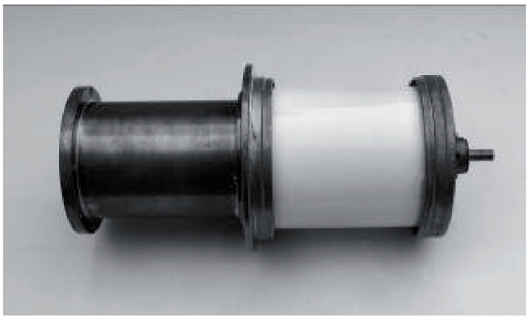

Fig. 13. View of fabricated axial vircator with adjustable distance between anode and cathode.

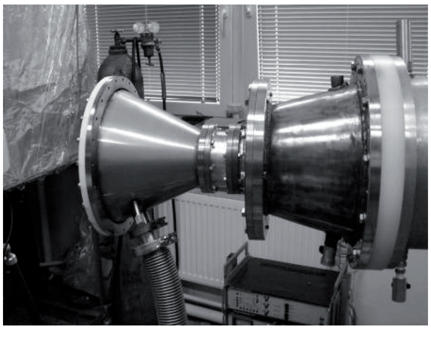

Fig. 14. View of fabricated axial vircator connected to the pulse forming line.

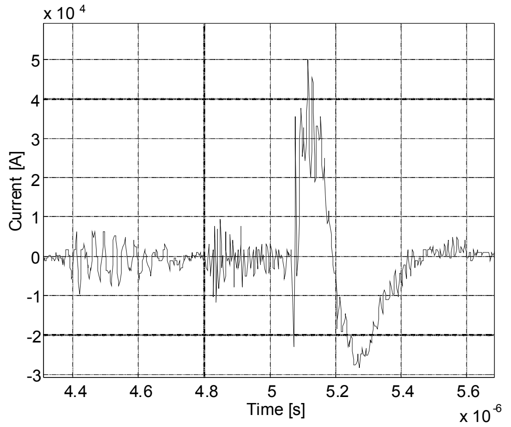

Fig. 15. Vircator current time history.

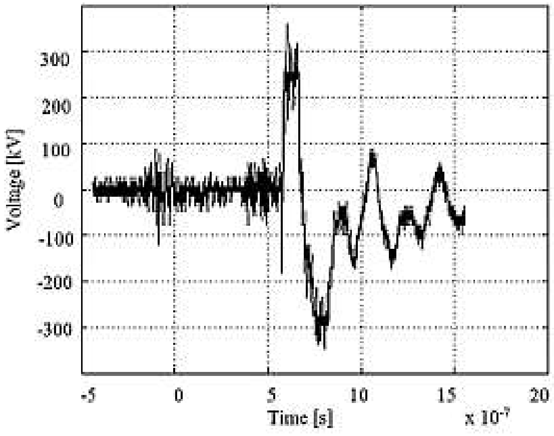

Fig. 16. Vircator voltage time history.

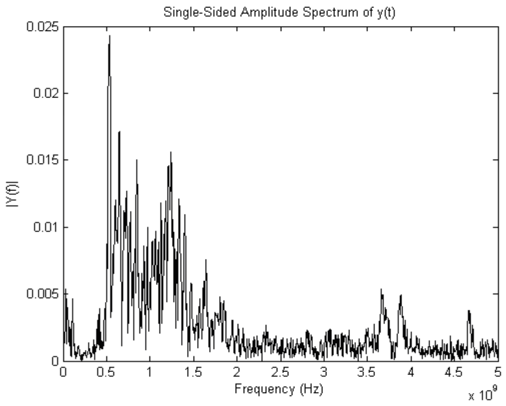

Fig. 17. Frequency spectrum of signals generated by realized axial vircator.

5. Conclusion

The research of directed energy weapons microwave validated substantial technology applicable in electromagnetic ammunition and in weapons for defense of objects (WDO). The axial vircator is the most important component part applicable in both categories of DEWM. Dominant frequencies of the generated microwave signal are placed in band of 0.5 to 1.5 GHz. To fulfill the vircator minimal proportion requirements to be able to use the vircator as a DEWM, it is necessary to shift the vircator dominant frequencies to the 3 GHz band. This requirement can be reached by arranging the cathode size efficiently and consequentially optimizing the distance between anode and cathode. Great attention also has to be paid to increase the specific dielectric strength of the output window of the axial vircator.

Acknowledgements

Research of the directed energy weapons microwave substantial technology was realized within the scope of ŽELETON, APPM and GENVLN research projects supported by the Ministry of Defense of Czech Republic.

References

[1] ADAMY, D. *EW101: A First Course in Electronic Warfare*. Norwood: Artech House, 2001.

[2] GLASSTONE, S., DOLAN, P. *The Effect of Nuclear Weapons.* 3rd ed. USA: United States Department of Defence, 1977.

[3] COOP, C. *An Introduction to the Technical and Operational Aspects of the Electromagnetic Bomb.* Canberra: RAAF Air Power Studies Centre Technical Reports, 1996.

[4] TAYLOR, C. D., GIRI, D. V. *High-Power Microwave Systems and Effects*. Washington: Taylor and Francis, 1994.

[5] BARKER, R. J., SCHAMILOGLU, E. *High-Power Microwave Sources and Technologies.* IEEE Press, 2002.

[6] DRAŽAN, L. *The Chosen Problems of Directed Energy Weapons Radio Frequency and Microwave Design.* Inaugural dissertation. Brno: University of Defence, 2008. 144 p. [in Czech].

[7] ALGILBERS, L. L., GRISHNAEV, I., TKACH, Y. *Magnetocumulative Generators.* New York: Springer-Verlag, 2000.

[8] DRAŽAN, L., ZAJONC, J. *Magneto-cumulative generator test results - version G3 and G4 (22.10.2002).* APPM project technical report. Brno: Military Academy, 2002. [in Czech].

[9] DRAŽAN, L., ZAJONC, J. *Magneto-Cumulative Generator Test Results, ver. G3 and G4 (26.11.2002).* APPM project technical report. Brno: Military Academy, 2002. [in Czech].

[10] DRAŽAN, L., ZAJONC, J. *Magneto-Cumulative Generator Test Results – ver. G3 (9.12.2002).* APPM project technical report. Brno: Military Academy, 2002. [in Czech].

[11] DRAŽAN, L., ZAJONC, J. *Magneto-Cumulative Generator Test Results – ver. G3 (7.1.2003).* APPM project technical report. Brno: Military Academy, 2003. [in Czech].

[12] DRAŽAN, L., ZAJONC, J. *Magneto-Cumulative Generator Optimisation Possibility.* APPM project technical report. Brno: Military Academy, 2003. [in Czech].

[13] MILLER, R., B. *An Introduction to the Physics Intense Particle Beams.* New York: Plenum Press, 1982.

[14] BEZDĚK, M., TECL, S., DRAŽAN, L. *Report Concerning the Vircator Research Results.* Research report. Vyškov: VOP-026 Šternberk, 2005. [in Czech].

[15] ŠUNKA, P. *High Power Pulse Generation*. Lectures collection. Prague: Academy of Sciences of the Czech Republic, Institute of Plasma Physics, 2004. [in Czech].

[16] DRAŽAN, L.,ZAJONC, J. *Test of Pulse Forming Line and Design of Microwave Generator with Virtual Cathode.* APPM project technical report. Brno: Military Academy, 2003. [in Czech].

[17] DOSTÁL, F., PLCH, Z. *GENVLN - Measurement.* Technical report. Vyškov: VOP 026 Šternberk, 2008. [in Czech].

[18] DREXLER, P., FIALA, P. Methods for high power EM pulse measurement. *Radioengineering*, 2006, vol. 15, no. 4, p. 43-49.

About Authors...

Libor DRAŽAN was born in Strakonice, Czech Republic, in 1960. He received his Ing. (M.Sc.) and CSc. (Ph.D) degrees from Brno Military Academy in 1984 and 1998, respectively. Since 1990, he has been with the Dept. of Radar at Military Academy and University of Defence in Brno, as an assistant professor (1990 to 2002), head of Dept. (since 2003) and associate professor (since 2008). His research interests include active and passive radar theory, electronic warfare and directed energy weapons microwave.

Roman VRÁNA was born in Havířov, Czech Republic, in 1972. He received his Ing. (M.Sc.) degree from Brno Military Academy in 1995. Since 1999, he has been with the Dept. of Radar, Military Academy and University of Defence in Brno, as an assistant professor (since 2008). His research interests include active radar theory and electronic warfare.

DARPA NESD
INDUSTRY GROUP

The *Biological Technologies Office (BTO)* of the *Defense Advanced Research Projects Agency (DARPA)* is sponsoring an Industry Group to support the ***Neural Engineering System Design (NESD) Program***.

NESD is a highly interdisciplinary program requiring extensive integration of new research and technology. To facilitate the development and adoption of NESD technologies, DARPA has organizedan industry group for potential proposers. Members of this group have agreed to provide access to rapid prototyping and manufacturing of advanced neuro-engineering components, including state-of-the-art electronics, photonics, computing, assembly and packaging.

This document contains all responses to **Special Notice 16-17** as of March 10, 2016. Any updates will be posted at **http://www.darpa.mil/work-with-us/opportunities**.

ALLEN INSTITUTE *for*
BRAIN SCIENCE

Allen Institute for Brain Science

Primary Contact
Christof Koch, Ph.D. / christofk@alleninstitute.org

About

The Allen Institute for Brain Science is an independent, non-profit medical research organization that has embarked on a concerted, ten-year effort, involving more than 300 scientists, engineers, technologists and other specialists, to understand the cerebral cortex in the mouse and in the human at the cellular level. We are built upon the notion of Big Science, Team Science and Open Science in the service of Basic Science with an explicit mission to accelerate understanding of the human brain in health and disease. Our data are freely and publicly available at www.brain-map.org.

Capabilities

- We seek to experimentally and computationally characterize the components (cell types) of the rodent visual thalamo-cortical system at the molecular, transcriptional, electron-microscopic, electrophysiological, morphological and functional levels to study how these components integrate information to generate high-level cognitive functioning such as object recognition.
- To study such high-level functioning, we use various in vivo perturbation and monitoring techniques (e.g. optical stimulation, extracellular depth recordings, wide and narrow-field Ca-imaging, etc.), very high density silicon probes (364 probes per shaft) as well as state-of-the art computer models that capture cortical information processing at varying levels of granularity.
- To understand the human brain, the Institute is also engaged in cell typing from human brain tissue harvested from patients undergoing brain surgery.
- Thus, we are routinely recording intracellularly from single and multiple cells from ex vivo human lateral temporal cortex and as well as performing morphological characterization on these same cells).

Capabilities

- We also have extensive experience with the genomic and transcriptional architecture of the human brain.
- We are interested in applying this knowledge for medical and translational purposes and to help facilitate advancement of CNS disorder therapies.

Facilities II

The Allen Institute headquarters, a 270,000 square foot life sciences building, is located in Seattle's South Lake Union neighborhood, the city's biotechnology hub. We are a high-throughput facility, home to custom-designed robotic systems, automated technologies, premier bench laboratory space and all the high-powered computing hardware and software necessary for processing petabytes of data.

Additional Contacts

Erica Sessle, MPH, MPhil / ericas@alleninstitute.org
Jennifer Pawlosky / jenniferp@alleninstitute.org

Bionics Institute

Primary Contact

Rob Shepherd / rshepherd@bionicsinstitute.org

Capabilities

Medical device safety and efficacy testing (in vitro and in vivo)

- Design, rapid prototyping & validation testing
- Accelerated aging & flexion testing
- Electrochemical impedance spectroscopy
- Acute feasibility studies
- Chronic safety and biocompatibility studies
- Drug delivery techniques
- Histology & data analysis
- Scanning electron microscopy of implanted materials
- Preclinical animal model development

Custom device prototyping & fabrication

- Design and fabrication of custom electrode arrays and surgical tools
- Use of FDA approved materials

'First in human' clinical trials

Clinical and Scientific Track Record

- 35 years of internationally recognised experience & scientific excellence
- Extensive clinical, academic & commercial partnerships
- Proven research & clinical development pathways

Staff and Facilities

- Experienced and qualified scientists, engineers and clinicians
- Co-located in a major Australian hospital
- Purpose built equipment and facilities

BOSTON
MICROMACHINES
CORPORATION

Boston Micromachines Corp

Primary Contact

Paul Bierden / pab@bostonmicromachines.com

About

Boston Micromachines Corporation (BMC) is the leading provider of microelectromechanical systems (MEMS) deformable mirror products and has expertise in the design of adaptive optics systems. The company's suite of award-winning compact DM products is the most cost-effective, highest performance mirrors in the market today. By applying wavefront correction, BMC devices can be used in a variety of applications which include laser beam shaping, microscopy, astronomy, and vision science. Located in Cambridge, MA, BMC is privately held and offers custom-designed manufacturing services in addition to its portfolio of standard DM products and adaptive optics systems.

Capabilities

The specific capabilities/services the company provides in relation to the NESD falls into two categories:

Deformable mirrors systems for use in neurological imaging instruments
- MEMS deformable mirrors ranging in size from 140 – 1000 actuators
- Mirror surface geometries of continuous, segmented, and hexagonal tip-tilt
- High speed drive electronics

These devices are currently being used in in vivo neural imaging systems at various premier research institutions around the country. Boston Micromachines also provides research and development services for advancing the deformable mirror and drive electronics technology and design to better fit specific applications.

- Advanced imaging systems utilizing adaptive optics to improve resolution and contrast in in vivo subsurface imaging application (i.e. deep penetration brain imaging)

Boston Scientific

Primary Contact

Rafael Carbunaru / Rafael.Carbunaru@bsci.com

Capabilities

- Lead prototyping
- Firmware modifications to the implantable pulse generators (IPG)
- Software modifications to the computer programmers (CP)
- Use of advanced computational models and predictive algorithms for parameters and programming guidance for neurostimulation

Additional Contacts

- Stephen Carcieri / Stephen.Carcieri@bsci.com
- Hemant Bokil / Hemant.Bokil@bsci.com
- Michael Moffitt / Michael.Moffitt@bsci.com

Chronocam

Primary Contact

Luca Verre / lverre@chronocam.com

Capabilities

- Extensive know-how in designing, developing and industrialising asynchronous, event-based, bio-inspired vision sensors
- Expertise in developing and customising event-based, bio-inspired computer vision algorithms for a large number of applications spanning from autonomous navigation, security and surveillance, aerospace and defence to biomedical devices and smart objects
- System integration capabilities: development of camera systems relying on Chronocam's proprietary vision sensors and development of embedded software porting Chronocam's proprietary computer vision algorithms on a large number of processing platforms (such as: CPUs, GPUs, FPGA)

Additional Contacts

Christoph Posch / cposch@chronocam.com
Xavier LaGorce / xlagorce@chronocam.com
Thomas Finateau / tfinateu@chronocam.com

cirtec
MEDICAL SYSTEMS

Cirtec

Primary Contact

Heather Dunn / Heather.Dunn@cirtecmed.com

About Cirtec

Cirtec Medical has been providing contract design, development, and manufacturing services to the medical device industry since our founding in Massachusetts in 1986. Today, Cirtec has over 150 employees in our two facilities on the east and west coasts.

Cirtec has supported active implantable developers throughout our history. Early projects focused on hermetic welding and testing services for implantable neurostimulation devices, with service offerings expanding to provide full design and development. In all, Cirtec has worked with over 60 different companies and institutions developing active implantable devices, with over 30 in the neuromodulation space. We work with companies and institutions of all sizes, from early stage starts-ups to large, well-established medical device manufacturers. We also work frequently with research laboratories and academic groups; our current projects include development and manufacturing of an implantable neurostimulator used for research and recently approved for clinical study under a feasibility IDE.

Capabilities

Cirtec is ISO 13485 certified and FDA registered.
Cirtec supports the full product development cycle from conceptual development and requirements definition through clinical and commercial manufacturing.

Our engineering teams including mechanical, electrical, and software design resources as well as process development, manufacturing, and quality engineering. Cirtec has a robust and flexible quality system; start-up companies and research groups often choose to work under our quality system rather than developing their own. We provide design history file management and document control services. We can also perform or contribute to requirements documentation, risk management activities, and design verification activities. Our manufacturing capabilities relevant to active implantable devices include:

- Laser welding
- Resistance welding
- Electrical assembly
- Laser marking
- Laser ablation
- Surface treatment
- Ultrasonic cleaning
- Epoxy molding and adhesion
- Silicone molding and adhesion
- Packaging and labelin

Manufacturing capabilities are supported by in-house test capabilities. These include mechanical testing, hermetic leak testing, bench level and automated electrical testing, environmental conditioning, and custom test setups such as soak or pressure testing.

Our manufacturing infrastructure includes multiple ISO class 7 cleanrooms and ESD safe handling procedures for all devices. Cirtec provides complete supply chain management services, leveraging our extensive Approved Supplier List and robust inventory management and material traceability procedures.

VALUE TO PROPOSING TEAMS: There are six primary areas where Cirtec capabilities provide value to proposing teams.

Hermetic barrier development	Hermetic package and barrier design
	Leak specification development
	Internal moisture management: Vacuum bake, desiccant handling, residual gas analysis testing
	Hermetic sealing (welding, brazing) and testing
Electrical assembly	Soldering, wire bonding, and die bonding
	Coil winding
	ESD-safe workspaces
	Lithium battery safe handling procedures
	Bench and automated testing
Interconnect development	Feedthrough design and sourcing
	Hardwired and disconnectable design
	Fabrication and testing
Encapsulation	Epoxy, silicone, and parylene encapsulation
	Layered encapsulation design to manage fluid ingress
	In-house design, tooling, molding, testing
Sterile packaging	Package design
	Pouch and tray packaging and labeling
	Off-the-shelf double sterile tray system
	Sterilization service provider management
Quality system support	Design control support and design history file maintenance
	GMP manufacturing
	Supply chain management
	Procedures for and support of requirements definition, risk management, design verification, and process validation

Dragon ID

Primary Contact

Eugene Malinskiy / eugene@dragonid.com

Capabilities

- Our staff is very experienced working within the complexities of the medical system from observing physicians in the operating room to bringing devices through regulatory approval.
- We have engineers with process development and industrial experience, including interfacing directly with manufacturers.
- We have software engineers experienced in both hardware and software engineering and experience in microcode and embedded systems, end user applications, and high performance computing.
- In addition to engineering talent, we staff researchers, designers, and SMEs to bring a truly interdisciplinary approach to innovation and design for a project.
- We have experience recruiting, establishing, and coordinating relationships between global stakeholders for complex projects to maximize accountability and efficiency for a product while minimizing bureaucracy required.
- We work with our world--class partners in industries such as industrial design, rapid multi--material prototyping, materials selection, and manufacturing in order to round out our capabilities for a product.

Freedom Photonics LLC

Primary Contact

Leif Johansson / leif@freedomphotonics.com

About

Founded in 2005 we are privately held Limited Liability Company based in Santa Barbara, California with ISO Qualification in process, certification in Q2 2016. We develop, manufacture and market complete photonic integrated circuit based products (Indium Phosphide, PLC, Si Photonics).

Our main markets in components, modules and subsystems are for fiber-optic communication, free-Space-optical (FSO) communication and infrared sensors and illuminators.

Capabilities

We offer electronics design, assembly and testing with expertise in photonic design, packaging, control electronics, manufacturing and qualification.

We offer services across development and have a premier R&D compound semiconductor facility with lithography (steppers, DUV stepper, EBL), wet and dry etching, deposition and characterization capabilities.

Our Business Model

Conceptualization, prototyping and development

- Fast R&D turn-around and low cost device fabrication in Nanotech cleanroom facility (Freedom Photonics employees)
- Internal multi-project wafer runs

Volume device production with commercial foundries worldwide

- Fabrication processes designed for manufacturability
- Access to manufacturability and production capacity, qualified processes
- Experience in outsourcing volume device manufacturing

Significant internal capability for assembly, packaging, test

Capabilities and Facilities I – Photonics Design and Fabrication

Design

- Commercial and proprietary tools for passive and active PIC modeling, (Lumerical, native FDTD/BPM code, mode solvers, mask layout, SOA/active region models, Simwindows, SRIM)

Mask layout

- Proprietary scripting tool, with design rule checker and automated mask plate layout

Fabrication

- Full wafer process, die pick-and-place, wedge bonder, ball- bonder, auto-cleaver, pull-tester, etc.

Capabilities and Facilities II – Photonics Production Test

- Photonic device bar level probe
- Chip-on-carrier level (probe card)
- Package; DC, spectral, and data testing
- Burn in – monitored, environmental chambers

Capabilities & Facilities III – Packaging

Several in-house systems with laser welding and UV curing capability.

- Custom tooling, TOSA packages, 14 pin butterflies etc.
- ZEMAX commercial optical train design tool

Capabilities and Facilities IV – Electronics

- Digital and Analog Electronics
- FPGA design and verification tools in house
- Commercial tools for board layout
- Outside foundry for PCB fabrication and stuffing
- Electronics workstations for assembly and test
- Oscilloscopes, sources, multi-meters

We perform contract fabrication and engage with a number of domestic and international fabrication providers, both full process or selected fabrication steps are available.

IMEC

Primary Contact

Barun Dutta / Barundeb.Dutta@imec.be

Capabilities

IMEC is interested in making its comprehensive capabilities available for joint-research programs for solutions in Life Sciences Technologies, Healthcare and Bio-medical engineering, including:

- Semiconductor/nanotechnology
- Process technology
- Design technology and systems/software technologies

Additional Contacts

Maarten Willems / Maarten.Willems@imec.be

Infinite Arthroscopy

Infinite Arthroscopy

Primary Contact

Daniel Dudley / d.dudley@infinitearthro.com

About

We are exclusively focused on developing new medical technologies with an emphasis on translating innovations from other industries to the operating theatre. We have experience in the full life cycle of medical device development from initial conception to regulatory submission. We have large number of rapid prototyping and fabrication techniques available in-house including but not limited to software-based design, stereolithography, fused deposition modelling, PCB milling, and machining. Our staff includes designers, engineers, health care professionals, and subject matter experts from a diversity of industries allowing us to leverage a far wider knowledge base than the standard medical device company. We maintain a nationwide network of leading physicians and institutional partners to develop, refine, and validate our technologies according to contemporary healthcare practice and needs.

Capabilities

Infinite Arthroscopy also maintains a portfolio of intellectual property derived from our medical device technologies that would be available to the NESD Industry Group. Two relevant examples are listed below.

Infinite Arthroscopy, Inc. Proprietary Technologies:
Novel Light Source

- Miniaturized / Ruggedized
- Low power (sub 1 watt)
- Low heat (to LEDs of comparable lumen-output)
- High lumen output
- Long Lifetime (50,000+ hours)
- Emits only target wavelength(s) (Capacity to customize to project needs)

Command/Control Architecture for Medical Devices

- HIPAA/FTC Approved for use in Operating Rooms
- Uninterrupted High-bandwidth capacity
- Transmission below standard RF noise floor
- Encrypted protocol

Additional Contacts
Howard Fein / h.fein@infinitearthro.com

Inscopix

Primary Contact

Pushkar Joshi / pushkar@inscopix.com

Capabilities

- Neuroscience application development and validation (surgical procedures for rodents and non-human primates, application of biological reagents within our own life science lab)
- Optical system development (lens design and simulation, LED light source development, prototyping)
- Hardware system (FPGA-based and/or microcontroller based mixed-signal systems especially involving imaging, mechanical packaging) development, testing, and prototyping
- Software (application, image processing, computer vision, analytical) development and testing
- Mechanical accessories development (design, prototyping via 3D printing)
- With these capabilities, we can customize Inscopix's existing and emerging systems and applications, or develop new systems and applications for or with other proposers in support of this program.

Additional Contacts

Shung Chieh / schieh@inscopix.com

Nanomedical Diagnostics

Primary Contact

Brett Goldsmith / brgoldsmith@nanomedicaldiagnostics.com

Capabilities

- Graphene FET sensor array chips appropriate for integration with nearly any biochemical or electrochemical interaction
- Current sensor hardware capable of measuring biochemical and bioelectric interactions over time with 10 millisecond time resolution, and spatial resolution less than 10 microns
- Capability to produce sensor hardware with time resolution approaching one microsecond with a spatial resolution less than 2 microns
- Expertise in combining electronic materials, packaging and assembly with active biological materials
- Projected FDA clearance for an in vitro clinical product, and associated manufacturing chain, using graphene sensors by mid 2017
- Graphene appropriate for use in medical sensors integrated into an electronic or implantable platform of your choosing
- Chemical and biochemical expertise in biochemical attachment and blocking techniques specialized for boosting specificity of electronic readout of biological interactions
- A complete "turnkey" R&D system for using graphene sensor chips
- Hardware and software appropriate for reading graphene biosensors

Additional Contacts

Francie Barron / fbarron@nanomedicaldiagnostics.com

Naval Research Laboratory

Primary Contact

Dr. Jas. S. Sanghera / sanghera@nrl.navy.mil

About

NRL is the corporate research laboratory for the Navy and Marine Corps and conducts a broad program of scientific research, technology and advanced development.

The Optical Materials and Devices Branch has five primary areas: high purity chemicals, specialty optical materials, silica fiber technology, optical devices and advanced concepts.

Capabilities

We have expertise in the research, development and fabrication of glasses, glass-ceramics, ceramics, crystals, bulk optics, films, waveguides and optical fibers.

We offer a full suite of capabilities to make transparent ceramics to solve the ceramic packaging needs of NESD, some key properties:

- UV to RF transparency tailoring
- Tougher, stronger, harder than glass
- Electrically inert
- Demonstrated conformal shapes
- Demonstrated bonding technologies
- Vertical integration

Our facilities include class 100 clean rooms, powder processing equipment, sintering systems, and characterization capabilities (XRD, SEM, TEM, optics labs, mechanical testing).

Some of our prior applications include: Conformal Optics, IR simulators, photonic mast design, flexible PV's, IRCM, and Directed Energy Weapons.

Additional Contacts

Shyam Bayya / shyam.bayya@nrl.navy.mil

Star Lab

Primary Contact

Adam Fraser / adam.fraser@starlab.io

About

Star Lab is a high-tech small business focused on complex systems security research and development, with significant expertise and a pedigree of success developing security technologies for safety-critical and mission-critical systems. The company specializes in the areas of high-assurance operating systems, technology protection, applied cryptography, and anti-tamper, and is also a leader in kernel/hypervisor/embedded software development. Star Lab has offices in Washington D.C. and Huntsville, AL.

Additional Contacts

Irby Thompson / irby@starlab.io

Qualcomm

Primary Contact

Karim Arabi / karabi@qti.qualcom.com

Capabilities

- Experience in building an end-to-end platform as a reference design for interface with brain
- Experience in driving an industry wide standard effort to standardize a communication medium and protocol for brain interface
- Advanced low power wireless solutions, RF interface, MODEM system solutions and related electronics for brain interface
- Advanced wireless charging and powering solutions for brain implants
- Ultra low power electronic design solutions at advanced technology nodes suitable for brain implants to enable always on monitoring and processing
- Ultra low power DSP solutions, stream data processing and data compression and decompression solutions

Additional Contacts

Rudy Beraha / rberaha@qti.qualcomm.com

New World War Revolutionary Methods for Political Control

Mind-Reading

History

Probably no more “intrusive and persistent” method of obtaining information about a person exists than reading their mind.[1] Research on mind-reading has been vigorously pursued by US government agencies and various academic centers since the 1970s, and continues to this day.

Since 1973 DARPA has been studying mind-reading with EEG hooked to computers, using scientists at the University of Illinois, UCLA, Stanford Research Institute, Massachusetts Institute of Technology, and the University of Rochester.

They developed a system that could determine how a person perceived colors or shapes and were working on methods to detect daydreaming, fatigue, and other brain states. Although the device had to be calibrated for each person’s brain by having them think a series of specific thoughts, the calibration was quick.

In 1974 another very basic mind-reading machine was created by researchers at Stanford Research Institute. It used an EEG hooked to a computer which allowed a dot to be moved across a computer screen using thought alone. When interpreting people’s brainwaves, it was right about 60% of the time. During these tests scientists discovered that brain patterns are like fingerprints, each person has their own. So, each computer would have to be calibrated for a specific person.

Another method to address this issue was to store a large amount of generic patterns on the computer, so when it encountered a brain pattern it didn’t recognized, it used one that most resembled it. Since then, DARPA has sponsored Brain-Computer Interface (BCI) and mind-reading programs at Duke University, MIT, University of Florida, and New York State University, Brooklyn.

The Human Computer Interaction group at Tufts University has studied mind-reading funded by grants from a government research and education agency known as the National Science Foundation (NSF). Carnegie Mellon University, Stanford University, and the MIT Sloan School of Management have studied mind-reading. The Computer Laboratory at the University of Cambridge in England has developed mind-reading machines based on facial expressions.

Other academic institutions that have participated in mind-reading projects include the University of California, Berkeley, University of Maryland, and Princeton University in New Jersey. Microsoft has studied mind-reading using EEG to better accommodate its users. Emotiv Systems built a mind-reading gaming device which uses EEG to infer the mental states of video game

Main Menu

- Home
- Read *New World War* Online
- Articles
- Solutions
- Biography
- Download my books for free

Updates

Update: January 20, 2018
A freemasonic death threat....more

Important Updates

All Previous Updates

Donations

Your Donations Are Appreciated

Available Books

PDF, Kindle, & ePub editions of my books can be downloaded for free, or purchased on Amazon & LuLu.

players. Honda Motors and Advanced Telecommunications Research Institute International (ATR) have studied mind-reading.

Neuroimaging Devices

Scientists discovered that the neural code of the human brain is similar to the digital code of a computer. To some extent, they have deciphered this code. Prior to this, they assumed that it was necessary to identify the neurons associated with specific acts, which would have made mind-reading much more difficult.[2]

They now understand that it's not necessary to monitor billions of neurons to determine which are connected to a particular thought or act. Only a small number of them need to be monitored to accomplish this. To monitor these neurons researchers use neuroimaging devices. They include event-related optical signal (EROS), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), functional near-infrared imaging (fNIR), magnetoencephalography (MEG), and positron emission tomography (PET).[3] These devices may be combined for a more accurate reading.

There are basically two types of measurements, *direct* methods and *indirect* methods. Direct methods measure changes in electromagnetic fields and currents around the brain which are emitted from the surface of the scalp, or they monitor the neurons themselves. Indirect methods measure hemodynamic (blood movement) changes of hemoglobin in specific tissue compartments.

Both of these methods are almost simultaneous with neuronal activity. Regarding sensors, there are invasive ones which must be implanted, and non-invasive ones which can be worn on the scalp, in the form of a headband.

Electroencephalography (EEG) provides a direct method for determining brain states and processes by measuring the electrical activity on the scalp produced by the firing of neurons in the brain. EEG has been around for over 100 years. EEG is commonly used in neuroscience, cognitive science, and cognitive psychology. It is inexpensive, silent, non-invasive, portable, and tolerates movement.

Wireless EEG which uses non-invasive sensors that have physical contact with the scalp can transmit the signals to a remote machine for deciphering. Although, in 1976 the *Los Angeles Times* reported that DARPA was working on an EEG to detect brain activity several feet from a person's head, which was to be completed in the 1980s. EEG normally produces only a general indicator of brain activity.

However, in 2008 *Discovery News* reported that a company called Emotiv Systems developed an algorithm that decodes the cortex, providing a more accurate measurement. "We can calibrate the algorithm across a wide range of technologies with the same resolution you would get from placing an invasive chip inside the head," said Tan Le, president of Emotiv Systems.

Functional magnetic resonance imaging (fMRI) measures the blood flow in the brain in response to neural activity. Active neurons use oxygen, which is brought to them by blood. The more active a region of the brain is the more

GEOENGINEERING
PLANETARY
SABOTAGE
&
HUMAN
EXTERMINATION
Mark M. Rich

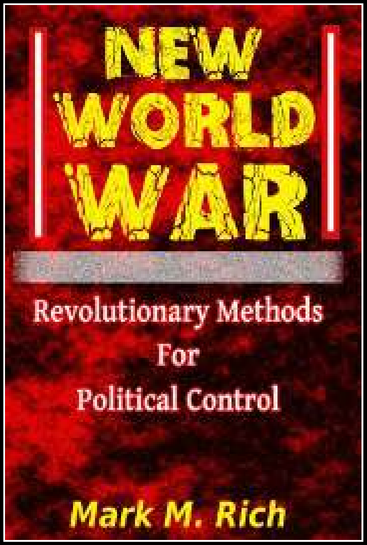

blood flows in the area. This movement of blood is referred to as *hemodynamic activity*. FMRI can detect which areas are receiving blood, which indicates that they're processing information.

The fMRI provides an indirect measurement of brain processes. It is the most common method of neuroimaging, and can produce 2 and 3-dimensional images. It is non-invasive, and can record signals from all brain regions, unlike EEG which focuses on the surface only.

Functional near-infrared imaging (fNIR) provides an indirect measurement of brain activity by detecting hemodynamic changes in the cortex. Although it is based on different principles, in that it uses light, it functions in the same manner as fMRI. FNIR can provide an almost continuous display of these changes in the cortex. It is inexpensive, non-invasive, and portable. A wireless headband with sensors exists for this device.

Event-related optical signal (EROS) is a brain-scanning device that focuses near-infrared light into the cerebral cortex to detect the density of neurons indicated by the transparency of brain tissue. Because it can only detect these changes a few centimeters deep, it can only image the cerebral cortex. Unlike fNIR, which is an optical method for measuring blood flow, EROS detects the intensity of neurons themselves and provides a direct measurement of brain activity. It is very accurate, portable, inexpensive, and non-invasive. A wireless headband with sensors exists for this device.

Capabilities

Mind-reading can be accomplished by first having a computer learn which brain patterns are associated with specific thoughts, then store the decoded information in a database. This machine learning is accomplished using a type of *artificial intelligence* (AI) called an *algorithm*. A very basic algorithm is a spell checker, which uses a database of common mistakes associated with a particular sequence of letters to present suggestions to a user.

“The new realization is that every thought is associated with a pattern of brain activity,” proclaimed neuroscientist John Dylan Haynes, in *Newsweek International* on February 4, 2008. “And,” says Haynes, “you can train a computer to recognize the pattern associated with a particular thought.”

In a January 2000 issue of *US News and World Report*, Lockheed Martin neuroengineer Dr. John Norseen announced, “Just like you can find one person in a million through fingerprints ... you can find one thought in a million.” This can be accomplished using AI and HCI, or what Dr. Norseen calls biofusion.

The decoded brain signals can be stored in a database. Then when someone is scanned, the computer detects the pattern and matches the signals to the database of known meanings. But it's not necessary to scan a brain to decode its signals for every single thought, such as a picture.

Instead, after the machine has learned how to decipher patterns associated with specific thoughts such as images, more images can be added to the program and the computer can use the process it used for the other images as a model to somewhat accurately detect additional thoughts.

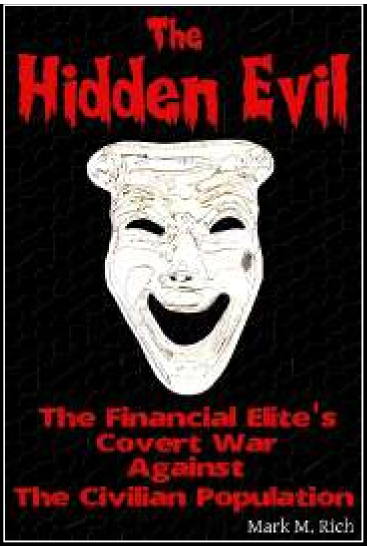
The Hidden Evil
The Financial Elite's Covert War Against The Civilian Population
Mark M. Rich

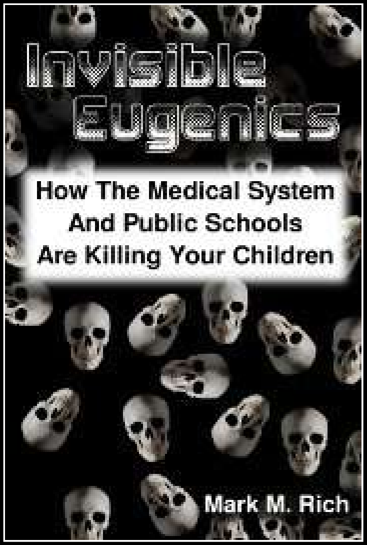
Invisible Eugenics
How The Medical System And Public Schools Are Killing Your Children
Mark M. Rich

New World War: Revolutionary Methods for Political Control

Dedication & Thanks

[Peer Review & Thanks]
[Dedication]

Volume I: Current Political Situation

[Overview]
[Introduction]
[Revolution in Warfare]
[The Other World]
[Dictatorship Creation]
[The Groups Facilitating the Revolution]
[Their Goal is Neo-Feudalism]
[Problem-Reaction-Solution]
[Volume I Commentary]

Both words and images can be detected using mind-reading devices with varying degrees of accuracy. This can occur for words and images being viewed by a person on an external display, such as a book, or words and images just being thought of with no external stimuli.

“It is possible to read someone’s mind by remotely measuring their brain activity,” announced *New Scientist* in their *Mind-Reading Machine Knows What You See* article of April of 2005. The Computational Neuroscience Laboratories at the Advanced Telecommunications Research Institute International (ATR) in Kyoto Japan, and Princeton University in New Jersey, proved that by monitoring the visual cortex with fMRI they could determine which basic objects (sets of lines) a person was looking at.

When the objects were combined, they could even determine which one was being focused on. According to the scientists, it may be possible not only to view but also to record and replay these images. They announced that the technology could be used to figure out dreams and other secrets in people’s minds.

Vanderbilt University in Nashville has conducted simple mind-reading tests using an fMRI/Computer, which learned what basic images a group of test subjects was looking at. They were able to predict with 50% accuracy which objects the test subjects were thinking of when they were asked only to remember what they had seen, without being shown the images.

On March 6, 2008 *ABC News* reported that neuroscientists at the University of California at Berkeley accomplished mind-reading by monitoring the visual cortex with an fMRI connected to a self-learning (artificial intelligence) computer program.

First, they used 1750 pictures to build a computational database for the computer to learn with by flashing the pictures in front of test subjects connected to an fMRI. This allowed the algorithm to decipher the brain patterns and associate them with the images.

In addition to deciphering these brain patterns, the computer recorded the process that it used to accomplish this, and built a model based upon it. Then, without scanning the test subjects, they added 120 new pictures to the program and allowed it to create its interpretation of what the new brain signals would be, based on the previous model.

Then they had the test subjects look at these pictures which they had never seen while being scanned. The computer predicted what they were looking at 72% of the time. The scientists announced that the model could be used as a basis to predict the brain activity associated with any image.

What this means is that it’s not necessary to scan a brain to obtain the meaning of each signal. Once the model had been developed, they could simply add new pictures to the database/dictionary. The scientists suggested that out of 1 billion pictures, the computer would be accurate about 20% of the time.

Images which are not consciously seen by a person can even be detected by mind-reading machines. Researchers at University College London flashed

Volume II: The New War

The New War
The New Enemy
Initiatives to Remove Civil Liberties
The Investigation
Surveillance Technology and Methods
Mind-Reading
Volume II Commentary

Volume III: Weapons of The New War

Introduction to Nonlethal Weapons
Psychological Operations
Introduction to Directed-Energy Weapons
High-Powered Microwaves
High-Powered Lasers
Sonic Weapons
Computer Network Operations
Microwave Hearing
Silent Subliminals
Use of Citizen Informants
Chemical and Biological
Weather Warfare
Miscellaneous Weapons and Tactics
Volume III Commentary

Volume IV: The Coverup

Volume IV Introduction
Schizophrenia Spectrum Disorders
Control of the Medical Industry
Another Look at Schizophrenia
Political Considerations
Punitive Psychiatry in Communist Russia
Coverup Initiatives
Volume IV Commentary
Conclusion

Appendix

A Brief History of PsyOp
Small-Scale Wars
Nongovernmental Organizations
Human-Computer Intelligence Network

pictures in quick succession to test subjects connected to an fMRI. Although some of these pictures were invisible to the subjects, they were accurately recorded 80% of the time by the computer.

Like a fingerprint, each person has their own *brainprint*. Therefore, calibration for each brain is necessary. This is accomplished by having the person think a series of specific thoughts. In the case of EEG, this calibration can take less than a minute. However, because the signals which represent thoughts are similar from one person to the next, a universal mind-reading database has been suggested.

Using fMRI, scientists at Carnegie Mellon University (CMU) discovered that the brain patterns associated with specific thoughts are quite similar among multiple people. This, they stated, would provide the opportunity to create a universal mind-reading dictionary.

Scientists at the University of California at Berkeley mentioned that a “general visual decoder” would have great scientific use. Likewise, the brain patterns associated with specific words that occur when people are reading are also basically the same. This similarity of brain functions associated with words seems to have been an evolutionary development which allowed for an advantage in communication.

A mind-reading machine capable of determining the brain pattern associated with a specific word was developed by scientists at CMU. Brain scans using fMRI were taken of test subjects who were given a variety of words to think of in order to train the computer. An important consideration here is that they were not viewing these words on an external display, only thinking about them. After the computer identified the brain patterns associated with those words, the subjects were given two new words to think about, which the computer accurately determined.

Although, in this particular study only a couple of words were tested, it proves that after a model of how to decipher brain signals was created, AI could accurately determine new words that subjects were thinking about. “These building blocks could be used to predict patterns for any concrete noun,” proclaimed Tom Mitchell of the Machine Learning Department.

In February of 2004 *Popular Science* announced that a mind-reading computer could, in theory, translate a person’s working verbal memory onto a computer screen. “You could imagine thinking about talking and having it projected into a room 2,000 miles away,” says Professor Craig Henriquez at Duke University’s Center for Neuroengineering, who has studied mind-reading for DARPA. He added, “It’s very, very possible.”

FMRI can be used to determine if someone is reading or writing. Neuroscientists can determine when a person is reading by monitoring their brainwaves. They can almost determine exactly what they’re reading. And because these patterns are similar from one person to the next, a universal device for determining what people are reading is possible.

In March of 2008 both Technology Review and *ABC News* revealed that an fMRI could in theory be used to display a person’s dreams. Then in December of 2008, scientists at ATR in Kyoto Japan announced that they developed a

Electronic Tyranny
Other Devices Connected to the GIG
My Experience

Sources

Other Sites

www.TheHiddenEvil.com
www.RichEssence.com

technology that would eventually allow them to record and replay a person's dreams.

Emotions from love to hate can be recognized by neuralimaging. The level of stress a person is experiencing can also be measured. Brain states such as honesty, deception, and even self deception, can be measured.

Patterns associated with decisions can also be read. Scientists from CMU, Stanford University, and the MIT Sloan School of Management were able to accurately predict the purchasing decisions of test subjects in a virtual shopping center. They monitored the subject's level of interest in a product as well as their decision to purchase it.

Neuroimaging can also detect decisions about how someone will later do a high-level mental activity. Neuroimaging can be used to determine if someone is speaking or reading. It can be used to detect areas of the brain that are active when someone is hearing a sound, or touching an object.

Brain patterns associated with specific physical movements, such as a finger, can be deciphered with neural imaging. The mere intention to make a physical movement can be detected before the actual movement is made.

Cameras

A type of mind-reading is possible with cameras connected to computers. One such device, called the Emotional Social Intelligence Prosthetic (ESP), was developed at the MIT Media Laboratory in 2006. It consists of a tiny camera that can be worn on a hat, an earphone and a small computer that is worn on a belt. It infers a person's emotional state by analyzing combinations of subtle facial movements and gestures.

When an emotional state is detected, the wearer is signaled through the earphone to adjust their behavior in order to gain the attention of the target. The computer can detect 6 emotional states. It can also be adjusted for cultural differences and configured specifically for the wearer.

Around this time, the Computer Laboratory at the University of Cambridge, UK, developed a similar camera-based mind-reading machine. It uses a computer to monitor, in real-time, combinations of head movement, shape, color, smiles, and eyebrow activity to infer a person's emotional state.

It detects basic emotional states such as happiness, sadness, anger, fear, surprise, and disgust, as well as more complex states. It's accurate between 65 and 90 percent of the time. "The mind-reading computer system presents information about your mental state as easily as a keyboard and mouse present text and commands," they announced.

Used for Surveillance

Mind reading exists. The DOD and various institutions have vigorously researched this subject since at least the mid 1970s. "Mapping human brain functions is now routine," declared *US News and World Report*, in an article entitled *Reading Your Mind—And Injecting Smart Thoughts*, of January of 2000.

Both words and images, being viewed or thought, can be mind-read. Various emotional states as well as mental processes such decisionmaking, reading, writing, movement and the intention to make a movement, can be detected with mind-reading devices. Perceptions such as touch, sound, and light can also be detected.

The proposed uses for mind-reading technology are positive. Some include determining if people in comas can communicate, helping stroke patients and those who suffered brain injuries, aiding those with learning disorders, assisting with online shopping, and improving people's communications skills.

However, other uses that have been suggested include the monitoring of unconscious mental processes, and interrogation of criminal suspects and potential terrorists. Dr. John Alexander mentioned that the recent developments in mind-reading technology would take surveillance to new levels by allowing investigators to "peer into the inner sanctum of the mind," in order to determine if a suspect has caused, or will likely cause a crime.

Dr. Norseen has sent R&D plans to the pentagon to have tiny mind-reading devices installed at airports to profile potential terrorists. He suggested that these devices could be functional by 2005. In August of 2008, *CNN* stated that the US military's knowledge obtained from mind-reading research could be used to interrogate the enemy.

Law Enforcement Technology announced in September of 2005 the existence of a new forensic technology known as *Brain Fingerprinting*, which has already been used in hundreds of investigations as a lie detector by the CIA, FBI and law enforcement agencies in the United States.

Brain Fingerprinting is admissible in court, because unlike a polygraph, which relies on emotional responses, it uses EEG to see how the brain reacts to words and pictures related to a crime scene. Dr. Larry Farwell, its inventor, says it is completely accurate. According to the report it will be used to help speed-up investigations.

Sources Next

Endnotes

[1] Another possible method to obtain information is Remote Viewing. RV is the ability to produce correct information about people, events, objects, or concepts that are somewhere else in space and time, and are completely blind to the viewer collecting the information. It can be used to describe people or events, produce leads, reconstruct events, make decisions, and make predictions about the future. See *Remote Viewing Secrets* by Joseph McMoneagle. RV tests were conducted by the US government over a 20-year period during Project Stargate, a classified initiative by the CIA which began in 1972 and lasted until about 1994. Most of the 154 tests and 26,000 trials took place at the Cognitive Sciences Laboratory at Forte Meade, Maryland. A majority of the results of the project are still classified. See the *Journal of Parapsychology* articles, *Remote Viewing by Committee*, September 22, 2003, by Lance Storm and *Experiment One of the SAIC Remote Viewing Program*, of December 1, 1998, by Richard Wiseman and Julie Milton. The success of the project varies depending on the source. Allegedly the original tests were conducted under rigid scientific conditions, which had impressive results. However, the same sources describe RV

in general as ineffective. See *Discover Magazine's* article, *CIA ESP*, on April 1, 1996, by Jeffrey Kluger, and the *Washington Post's* report, *Many Find Remote Viewing a Far Fetch from Science*, on December 2, 1995, by Curt Suplee. According to author McMoneagle, an original viewer during Project Stargate, it is accurate about 50 or 60 percent of the time. Nevertheless, RV will be used to obtain intelligence, according to John B. Alexander. See *The New Mental Battlefield*, which appeared in the December 1980 issue of *Military Review*. Also see the June 1998 *Research Report Number 2* of the University of Bradford's Non-Lethal Weapons Research Project (BNLWRP), for how RV has been added to the NLW arsenal. According to multiple sources, US government agencies are now using the consulting services of RV professionals. This was reported on January 9, 2002 in the *University Wire's (Colorado Daily)* article, *Clairvoyant Discusses Reveals Details of Remote Viewing*, by Wendy Kale, and in the *Bulletin of the Atomic Scientists* on September 1, 1994, in its report, *The Soft Kill Fallacy* by Steven Aftergood. In his book *Winning The War: Advanced Weapons, Strategies, and Concepts for the Post-911 World*, Alexander had this to say regarding RV: "Since the beginning of history, humans have made anecdotal references to innate abilities to foretell the future, to know what was occurring at distant locations or the status of people separated from them, and to find resources they need without any traditional means of accessing that information." He continued: "Studies have demonstrated beyond any doubt that these nontraditional capabilities exist. ... [RV can] radically change our means of gathering intelligence. It holds the promise of providing information about inaccessible redoubts and advances in technology. More importantly, once these skills are understood, those possessing them will be able to determine an adversary's intent and be predictive about the events."

[2] Because neuroimaging technology decodes brain patterns to thoughts, some argue that it technically doesn't read a person's mind. However, because specific thoughts and brain states can be deciphered, here it is referred to as mind-reading. Additionally, most mainstream documents refer to this as mind-reading, despite the fact that it is actually brainwave-reading.

[3] Magnetoencephalography (MEG) and positron emission tomography (PET) can also be used to infer a person's neurophysiological state. But because they are impractical for field use due to their large size and harmful radiation, MEG and PET won't be considered here. However, DARPA is in the process of developing a small helmet-sized MEG device which would be connected to a portable computer. See the article *Mind over Machine* in the February 1, 2004 issue of *Popular Science*, by Carl Zimmer.

Copyright © 2011 Mark Rich. All rights reserved.

Mark (at) NewWorldWar.org

Congressional Research Service
Informing the legislative debate since 1914

Navy Lasers, Railgun, and Hypervelocity Projectile: Background and Issues for Congress

Ronald O'Rourke
Specialist in Naval Affairs

May 25, 2016

Congressional Research Service
7-5700
www.crs.gov
R44175

Summary

The Navy is currently developing three potential new weapons that could improve the ability of its surface ships to defend themselves against enemy missiles—solid state lasers (SSLs), the electromagnetic railgun (EMRG), and the hypervelocity projectile (HVP).

Any one of these new weapon technologies, if successfully developed and deployed, might be regarded as a "game changer" for defending Navy surface ships against enemy missiles. If two or three of them are successfully developed and deployed, the result might be considered not just a game changer, but a revolution. Rarely has the Navy had so many potential new types of surface-ship missile-defense weapons simultaneously available for development and potential deployment.

Although the Navy in recent years has made considerable progress in developing SSLs, EMRG, and HVP, a number of significant development challenges remain. Overcoming these challenges will likely require years of additional development work, and ultimate success in overcoming them is not guaranteed.

The issue for Congress is whether to approve, reject, or modify the Navy's funding requests and proposed acquisition strategies for these three potential new weapons. Potential oversight questions for Congress include the following:

- Using currently available approaches for countering anti-ship cruise missiles (ASCMs) and anti-ship ballistic missiles (ASBMs), how well could Navy surface ships defend themselves in a combat scenario against an adversary such as China that has large numbers of ASCMs (including advanced models) and ASBMs? How would this change if Navy surface ships in coming years were equipped with SSLs, EMRG, HVP, or some combination of these systems?
- How significant are the remaining development challenges for SSLs, EMRG, and HVP?
- Are current schedules for developing SSLs, EMRG, and HVP appropriate in relation to remaining development challenges and projected improvements in enemy ASCMs and ASBMs? To what degree are current schedules for developing SSLs, EMRG, or HVP sensitive to annual funding levels?
- When does the Navy anticipate issuing roadmaps detailing its plans for procuring and installing production versions of SSLs, EMRGs, and HVP on specific Navy ships by specific dates?
- Will the kinds of surface ships that the Navy plans to procure in coming years have sufficient space, weight, electrical power, and cooling capability to take full advantage of SSLs (particularly those with beam powers above 200 kW) and EMRG? What changes, if any, would need to be made in Navy plans for procuring large surface combatants (i.e., destroyers and cruisers) or other Navy ships to take full advantage of SSLs and EMRG?
- Are the funding sources for SSLs, EMRG, and HVP in Navy and Defense-Wide research and development accounts sufficiently visible for supporting congressional oversight?

Contents

Introduction 1
  Issue for Congress 1
  Scope of Report 1
Background 2
  Strategic and Budgetary Context 2
    Concern about Survivability of Navy Surface Ships 2
    Depth of Magazine and Cost Exchange Ratio 3
  SSLs, EMRG, and HVP in Brief 4
    SSLs 4
    EMRG 13
    HVP 16
  Indirectly Improving Ability to Counter ASCMs and ASBMs 22
  Remaining Development Challenges 22
    SSLs 23
    EMRG and HVP 24
Issues for Congress 26
  Potential Oversight Questions 26
Legislative Activity for FY2017 27
  Summary of Congressional Action on FY2017 Funding 27
  FY2017 National Defense Authorization Act (H.R. 4909/S. 2943) 27
    House 27
    Senate 30
  FY2017 DOD Appropriations Act (H.R. 5293) 33
    House 33
  Directed Energy Weapon Systems Acquisition Act of 2016 (H.R. 4964/S. 2778) 34
    House 34
    Senate 35

Figures

Figure 1. Laser Weapon System (LaWS) on USS Ponce 7
Figure 2. Laser Weapon System (LaWS) on USS Ponce 7
Figure 3. Laser Weapon System (LaWS) on USS Ponce 8
Figure 4. Laser Weapon System Demonstrator (LWSD) on Self Defense Test Ship 10
Figure 5. Laser Weapon System Demonstrator (LWSD) on Self Defense Test Ship 11
Figure 6. ONR Graphic of LWSD Components 12
Figure 7. Industry-Built EMRG Prototype Demonstrator 14
Figure 8. Industry-Built EMRG Prototype Demonstrator 15
Figure 9. EMRG Prototype Demonstrator Installed on a JHSV 16
Figure 10. Photograph Showing HVP 17
Figure 11. HVP 18
Figure 12. HVP Launch Packages 19
Figure 13. HVP Application to Various Launchers 20

Figure 14. Navy Slide Depicting Operations Against Various Target Types 21
Figure 15. Development Challenges for SSLs 24
Figure 16. Development Challenges for EMRG 25

Tables

Table 1. Summary of Congressional Action on FY17 Funding.. 27

Appendixes

Appendix. Potential Advantages and Limitations of Shipboard Lasers 38

Contacts

Author Contact Information 40

Introduction

Issue for Congress

This report provides background information and issues for Congress on three potential new weapons that could improve the ability of Navy surface ships to defend themselves against enemy missiles—solid state lasers (SSLs), the electromagnetic railgun (EMRG), and the hypervelocity projectile (HVP).[1]

Any one of these new weapon technologies, if successfully developed and deployed, might be regarded as a "game changer" for defending Navy surface ships against enemy missiles. If two or three of them are successfully developed and deployed, the result might be considered not just a game changer, but a revolution. Rarely has the Navy had so many potential new types of surface-ship missile-defense weapons simultaneously available for development and potential deployment. Although the Navy in recent years has made considerable progress in developing SSLs, EMRG, and HVP, a number of significant development challenges remain.

The issue for Congress is whether to approve, reject, or modify the Navy's funding requests and proposed acquisition strategies for these three potential new weapons. Congress' decisions on this issue could affect future Navy capabilities and funding requirements and the defense industrial base.

Scope of Report

SSLs are being developed by multiple parts of the Department of Defense (DOD), not just the Navy. SSLs, EMRG, and HVP, moreover, have potential application to military aircraft and ground forces equipment, not just surface ships. And SSLs, EMRG, and HVP can be used for missions other than defending against ASCMs and ASBMs.[2] This report focuses on Navy efforts to develop SSLs, EMRG, and HVP for potential use in defending Navy surface ships against ASCMs and ASBMs. It supersedes an earlier CRS report that provided an introduction to potential Navy shipboard lasers.[3]

Note that while fictional depictions of laser weapons in popular media often show them being used to attack targets at long ranges, the SSLs currently being developed by the Navy for potential shipboard use would be used to counter targets at short ranges of about a mile to perhaps a few miles.

---

[1] Railgun is also spelled as rail gun; EMRG is also abbreviated as EM railgun; hypervelocity is also spelled as hyper-velocity or hyper velocity.

[2] As discussed later in the report, the Navy is exploring the potential for using shipboard lasers to counter small boats and unmanned aerial vehicles (UAVs), and EMRG can be used to attack land targets.

[3] CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke. This earlier CRS report has been archived and remains available as a supplementary reference source on potential Navy shipboard lasers.

Background

Strategic and Budgetary Context

Concern about Survivability of Navy Surface Ships

Although Navy surface ships have a number of means for defending themselves against anti-ship cruise missiles (ASCMs) and anti-ship ballistic missiles (ASBMs),[4] some observers are concerned about the survivability of Navy surface ships in potential combat situations against adversaries, such as China, that are armed with advanced ASCMs and with ASBMs.[5] Concern about this issue has led some observers to conclude that the Navy's surface fleet in coming years might need to avoid operating in waters that are within range of these weapons, or that the Navy might need to move toward a different fleet architecture that relies less on larger surface ships and more on smaller surface ships and submarines.[6] Such changes in Navy operating areas and fleet architecture could substantially affect U.S. military strategy and the composition of the Navy's shipbuilding expenditures.

Navy surface fleet leaders in early 2015 announced a new organizing concept for the Navy's surface fleet called distributed lethality. Under distributed lethality, offensive weapons such as ASCMs are to be distributed more widely across all types of Navy surface ships, and new operational concepts for Navy surface ship formations are to be implemented. The aim of distributed lethality is to boost the surface fleet's capability for attacking enemy ships and make it less possible for an enemy to cripple the U.S. fleet by concentrating its attacks on a few very-high-value Navy surface ships (particularly the Navy's aircraft carriers).[7] Perspectives on whether it would be cost effective to spend money spreading offensive weapons across a wider array of Navy surface ships might be influenced by views on whether those surface ships can adequately defend themselves against enemy missiles.

---

[4] These include the following: operating ships in ways that make it hard for others to detect and accurately track Navy ships; jamming or destroying enemy targeting sensors; interfering with the transmission of targeting data from sensors to weapon launchers; attacking weapon launchers (which can land-based launchers or launchers on surface ships, submarines, or aircraft); and countering ASCMs and ASBMs headed toward Navy ships. Navy measures for countering ASCMs and ASBMs headed toward Navy ships include the following: jamming a missile's guidance system; using decoys of various kinds to lure enemy missiles away from Navy ships; and shooting down enemy missiles with surface-to-air missiles and the Phalanx Close-In Weapon System (CIWS), which is essentially a radar-controlled Gatling gun. Employing all these measures reflects a long-standing Navy approach of creating a multi-layered defense against enemy missiles, and of attacking the enemy's "kill chain" at multiple points so as to increase the chances of breaking the chain. (The kill chain is the sequence of steps that an enemy must complete to conduct a successful missile attack on a Navy ship. This sequence includes, at a basic level of description, detecting and tracking the Navy ship, passing that information from sensors to the weapon launcher, launching the weapon, and guiding the weapon all the way to the Navy ship. Interfering with any one of these actions can break the kill chain and thereby prevent or defeat the attack.)

[5] See, for example, Andrew F. Krepinevich, Maritime Warfare in a Mature Precision-Strike Regime, Washington, Center for Strategic and Budgetary Assessments, 2014, 128 pp. For more on China's ASCMs and ASBMs, see CRS Report RL33153, *China Naval Modernization: Implications for U.S. Navy Capabilities—Background and Issues for Congress*, by Ronald O'Rourke.

ASCMs and ASBMs are not the only reasons that some observers are concerned about the future survivability of U.S. Navy surface ships in combat situations; observers are also concerned about threats to U.S. Navy surface ships posed by small boats, mines, and torpedoes.

[6] See, for example, Phillip E. Pournelle, "The Deadly Future of Sea Control," *U.S. Naval Institute Proceedings*, July 2015: 26-31.

[7] See, for example, Thomas Rowden, Peter Gumataotao, and Peter Fanta, "Distributed Lethality," *U.S. Naval Institute Proceedings*, January 2015: 18-23; Sam LaGrone, "SNA: Navy Surface Leaders Pitch More Lethal Ships, Surface Action Groups," *USNI News*, January 14, 2015; Kris Osborn, "Navy Unveils New Surface Warfare Strategy," *Military.com*, January 14, 2015; Sydney J. Freedberg Jr., "'If It Floats, It Fights,': Navy Seeks 'Distributed Lethality,'" *Breaking Defense*, January 14, 2015; Mike McCarthy and Megan Eckstein, "Navy Eyeing A 'Hunter Killer' Surface (continued...)

Depth of Magazine and Cost Exchange Ratio

Two key limitations that Navy surface ships currently have in defending themselves against ASCMs and ASBMs are limited depth of magazine and unfavorable cost exchange ratios. Limited depth of magazine refers to the fact that Navy surface ships can use surface-to-air missiles (SAMs) and their Close-in Weapon System (CIWS) Gatling guns to shoot down only a certain number of enemy unmanned aerial vehicles (UAVs) and anti-ship missiles before running out of SAMs and CIWS ammunition[8]—a situation (sometimes called "going Winchester"), that can require a ship to withdraw from battle, spend time travelling to a safe reloading location (which can be hundreds of miles away),[9] and then spend more time traveling back to the battle area.

Unfavorable cost exchange ratios refer to the fact that a SAM used to shoot down a UAV or anti-ship missile can cost the Navy more (perhaps much more) to procure than it cost the adversary to build or acquire the UAV or anti-ship missile. In the FY2016 defense budget, procurement costs for Navy SAMs range from about $900,000 per missile to several million dollars per missile, depending on the type.[10]

In combat scenarios against an adversary with a limited number of UAVs and anti-ship missiles, an unfavorable cost exchange ratio can be acceptable because it saves the lives of Navy sailors and prevents very expensive damage to Navy ships. But in combat scenarios (or an ongoing military capabilities competition) against a country such as China that has many UAVs and anti-ship missiles and a capacity for building or acquiring many more, an unfavorable cost exchange ratio can become a very expensive—and potentially unaffordable—approach to defending Navy surface ships against UAVs and anti-ship missiles, particularly in a context of constraints on U.S. defense spending and competing demands for finite U.S. defense funds.

(...continued)

Fleet, Would Require Upgunning Existing Ship Fleets," *Defense Daily*, January 15, 2015: 1-3; Richard Scott, "Offensive Language: USN Sets Out Surface Firepower Strategy," *Jane's International Defence Review*, May 2015: 42-47; Megan Eckstein, "Navy Studying Implications of Distributed Lethality in Wargames Series," *USNI News*, July 9, 2015; Lara Seligman, "Navy Establishes Task Force To Study Impact of Distributed lethality," *Inside the Navy*, July 10, 2015.

[8] Navy cruisers have 122 missile cells; Navy destroyers have 90 or 96 missile cells. Some of these cells are used for storing and launching Tomahawk land attack cruise missiles or anti-submarine rockets. The remainder are available for storing and launching SAMs. A Navy cruiser or destroyer might thus be armed with a few dozen or several dozen SAMs for countering ASCMs and ASBMs. Countering ASCMs or ASBMs with SAMs might sometimes require shooting two SAMs at each ASCM or ASBM.

[9] The missile cells on a Navy cruiser or destroyers are clustered together in an installation called a Vertical Launch System (VLS). VLS cells cannot be reloaded while the ship is underway; a ship needs to return to a port or a calm anchorage to reload its VLS.

[10] Unit procurement costs for ship-launched SAMs in the FY2016 are as follows: about $900,000 for the Rolling Airframe Missile (RAM), about $1.1 million to about $1.5 million for the Evolved Sea Sparrow Missile (ESSM), about $3.9 million for the SM-6 Block 1 missile, about $14 million for the SM-3 Block 1B missile, and more than $20 million for theSM-3 Block IIA missiles. RAM and ESSM are short-range missiles for defense against aircraft and ASCMs. The SM-6 Block 1 is a medium-range missile used for both defense against aircraft and ASCMs, and terminal (i.e., endo-atmospheric) defense against theater-range ballistic missiles. The SM-3 Block 1B and SM-3 Block IIA are used for mid-course (i.e., exo-atmospheric) defense against theater-range ballistic missiles.

SSLs, EMRG, and HVP offer a potential for dramatically improving depth of magazine and the cost exchange ratio:

- **Depth of magazine.** SSLs are electrically powered, drawing their power from the ship's overall electrical supply, and can be fired over and over, indefinitely, as long as the SSL continues to work and the ship has fuel to generate electricity. The EMRG's projectile and the HVP (which are one and the same—see next section) can be stored by the hundreds in a Navy surface ship's weapon magazine.[11]
- **Cost exchange ratio.** An SSL can be fired for a marginal cost of less than one dollar per shot (which is the cost of the fuel needed to generate the electricity used in the shot), while the EMRG's projectile/HVP has an estimated unit procurement cost of about $25,000.[12]

For additional discussion of the strategic and budgetary context in which the programs discussed in this report and other Navy programs may be considered, see CRS Report RL32665, *Navy Force Structure and Shipbuilding Plans: Background and Issues for Congress*, by Ronald O'Rourke.

SSLs, EMRG, and HVP in Brief

SSLs

Overview

The Navy in recent years has leveraged both significant advancements in industrial SSLs and decades of research and development work on military lasers done by other parts of DOD to make substantial progress toward deploying high-energy SSLs[13] on Navy surface ships. Navy surface ships would use high-energy SSLs initially for countering small boats UAVs, and potentially in the future for countering ASCMs and ASBMs as well.[14] High-energy SSLs on Navy ships would be short-range defensive weapons—they would counter targets at ranges of about one mile to perhaps eventually a few miles.[15]

---

[11] In July 2015, the Navy issued a request for information (RFI) to industry for the fabrication of a prototype EMRG mount that would store a minimum of 650 rounds. (RFI for Fabrication of Prototype Mount for Naval Railgun, Solicitation Number: N00024-15-R-4132, FedBizOpps.gov, July 29, 2015. See also Justin Doubleday, "Navy Developing Integrated Mount For Electromagnetic Railgun," *Inside the Navy*, July 31, 2015.)

[12] Sources for cost of HVP: David Martin, "Navy's Newest Weapon Kills at Seven Times the Speed of Sound," CBS News (cbssnews.com), April 7, 2014; Kris Osborn, "Navy Will Test its Electromagnetic Rail Gun aboard DDG 1000," *DefenseTech*, April 15, 2015.

[13] In discussions of potential Navy shipboard lasers, a high-energy laser is generally considered to be a laser with a beam power of at least 10 kilowatts (kW).

[14] In general, lasers would counter small boats and missiles by heating and burning holes in their skins, and causing thermal damage to their interiors. Lasers can also be used to "dazzle" (i.e., interfere with) electro-optical sensors on a boat or missile.

[15] The Navy has also performed research and development work on a different kind of laser, called the free electron laser (FEL). In recent years, Navy research and development work on potential shipboard lasers has shifted more to SSLs. For background information on the FEL, see CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke.

In addition to a low marginal cost per shot and deep magazine, potential advantages of shipboard lasers include fast engagement times, an ability to counter radically maneuvering missiles, an ability to conduct precision engagements, and an ability to use lasers for graduated responses ranging from detecting and monitoring targets to causing disabling damage. Potential limitations of shipboard lasers relate to line of sight; atmospheric absorption, scattering, and turbulence (which prevent shipboard lasers from being all-weather weapons); an effect known as thermal blooming that can reduce laser effectiveness; countering saturation attacks; possible adversary use of hardened targets and countermeasures; and risk of collateral damage, including damage to aircraft and satellites and permanent damage to human eyesight, including blinding. These potential advantages and limitations are discussed in greater detail in the **Appendix**.

Selected Key Developments

Key developments in the Navy's high-energy SSL development effort include the following:

- Between 2009 and 2012, the Navy successfully tested a prototype SSL called the Laser Weapon System (LaWS) against UAVs in a series of engagements that took place initially on land and subsequently on a Navy ship at sea.
- Between 2010 and 2011, the Navy tested another prototype SSL called the Maritime Laser Demonstration (MLD) in a series of tests that culminated with an MLD installed on a Navy ship successfully engaging a small boat.
- In April 2013, the Navy announced that it planned to install LaWS on the USS *Ponce* (pronounced pon-SAY)—a converted amphibious ship that is operating in the Persian Gulf as an interim Afloat Forward Staging Base (AFSB[I]) [16]—to conduct evaluation of shipboard lasers in an operational setting against swarming boats and swarming UAVs.[17] The system was installed in August 2014 (see **Figure 1**, **Figure 2**, and **Figure 3**).
- In March 2014, it was reported that the Navy anticipated moving to a shipboard laser program of record in "the FY2018 time frame" and achieving an initial operational capability (IOC) with a shipboard laser in FY2020 or FY2021.[18]

---

[16] An AFSB operates as a "mother ship" for Navy helicopter and small boat operations. The Ponce is serving as an interim AFSB pending the arrival of a new AFSB that is currently being built.

[17] "Navy Leaders Announce Plans for Deploying Cost-Saving Laser Technology," *Navy News Service*, April 8, 2013; Thom Shanker, "Navy Deploying Laser Weapon Prototype Near Iran," *New York Times*, April 9, 2013: 4; Mike McCarthy, "Navy Deploying Laser For Taking Out Drones," *Defense Daily*, April 9, 2013; Graham Warwick, "U.S. Navy Planning Gulf Deployment For Laser Weapon," *Aerospace Daily & Defense Report*, April 9, 2013: 6; Megan Eckstein, "Navy-Built Laser Weapon System Will Begin Demo On Ponce In Early 2014," *Inside the Navy*, April 15, 2013. See also Lara Seligman, "Navy-built LaWS To Begin Demo This Summer, IOC Slated For FY-20-21," *Inside the Navy*, March 24, 2014; Office of Naval Research, "All Systems Go: Navy's Laser Weapon Ready for Summer Deployment," *Navy News Service*, April 7, 2014.

Swarming refers to the use of boats and UAVs in large numbers, or swarms, in an attempt to confuse and overwhelm a target ship's defensive systems.

[18] Lara Seligman, "Navy-built LaWS To Begin Demo This Summer, IOC Slated For FY-20-21," *Inside the Navy*, March 24, 2014. A program of record, or POR, is a term sometimes used by DOD officials that means, in general, a program in the Future Years Defense Plan (FYDP) that is intended to provide a new, improved, or continuing materiel, weapon, or information system or service capability in response to an approved need. The term is sometimes used to refer to a program in a service's budget for procuring and deploying an operational weapon system, as opposed to a research and development effort that might or might not eventually lead to procurement and deployment of an operational weapon system.

- In December 2014, the Navy declared LaWS on the Ponce to be an "operational" system.[19]
- In January 2016, the Navy stated that it anticipated releasing a directed energy weapon roadmap in February 2016.[20]

---

[19] A December, 11, 2014, press report stated

> The Navy's first-of-a-kind laser deployed on a vessel sailing in the Persian Gulf has been declared operational and can be used by the crew to defend itself against potential threats, the service's head of the Office of Naval Research said on Wednesday [December 10, 2014].
>
> Rear Adm. Matthew Klunder told reporters on a conference call that Central Command has been green lighted to use the laser in the event of a threat, approval that has been passed along to the ship's commanding officer. The 30-kilowat laser, known as the Laser Weapon System, or LaWS, was installed on the USS Ponce in August [2014].
>
> The ship later departed for the Persian Gulf and the LaWS successfully carried out operational testing recently by striking a fast attack boat and drone, Klunder said, adding that this marks the "historic" first ever operational deployment of a directed energy weapon.
>
> (Mike McCarthy, "Navy Authorized To Use Ship-Based Laser In Battle," *Defense Daily*, December 11, 2014: 3. See also Sam LaGrone, "U.S. Navy Allowed to Use Persian Gulf Laser for Defense," *USNI News*, December 10, 2014; Philip Ewing, "Navy Declares Laser Weapon 'Operational,'" *Politico Pro (Pro Defense Report)*, December 10, 2014.)

The Navy testified on February 24, 2016, that

> the Solid State Laser Quick Reaction Capability (SSL-QRC) was fielded as a science and technology demonstration aboard the USS PONCE. It was successfully demonstrated as an effective weapon system and was subsequently transitioned to the fleet in the Central Command area of responsibility and is now an operational system.
>
> (Statement of Rear Admiral Mathias W. Winter, United States Navy, Chief of Naval Research, Before the Emerging Threats and Capabilities Subcommittee of the House Armed Services Committee on The Fiscal Year 2017 Budget Request, February 24, 2016, p. 15.)

[20] Justin Doubleday, "Winter: Navy Directed-Energy Strategy To Be Released This Month," *Inside the Navy*, February 1, 2016.

Figure 1. Laser Weapon System (LaWS) on USS Ponce

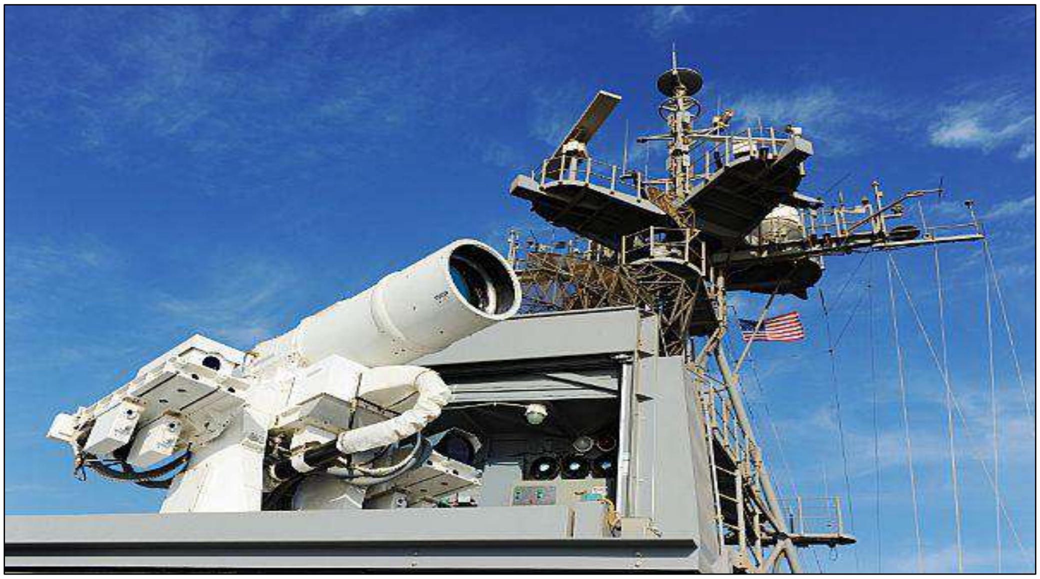

Source: Navy photograph dated November 16, 2014, accompanying David Smalley, "Historic Leap: Navy Shipboard Laser Operates in Arabian Gulf," *Navy News Service*, December 10, 2014, accessed August 12, 2015, at http://www.navy.mil/list_all.asp?id=84805.

Figure 2. Laser Weapon System (LaWS) on USS Ponce

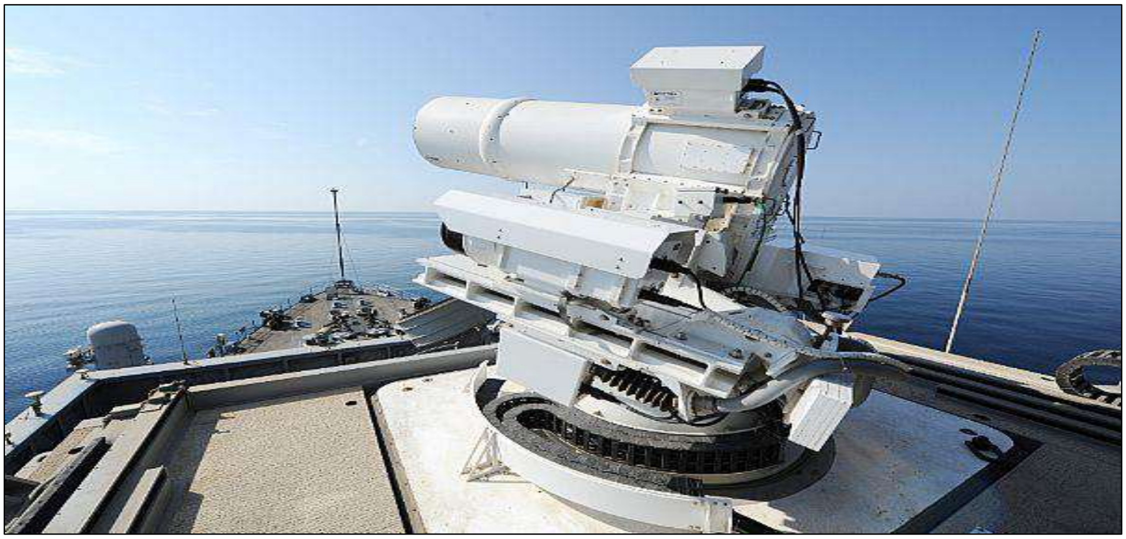

Source: Navy photograph dated November 17, 2014, accompanying David Smalley, "Historic Leap: Navy Shipboard Laser Operates in Arabian Gulf," *Navy News Service*, December 10, 2014, accessed August 12, 2015, at http://www.navy.mil/list_all.asp?id=84805.

Figure 3. Laser Weapon System (LaWS) on USS Ponce

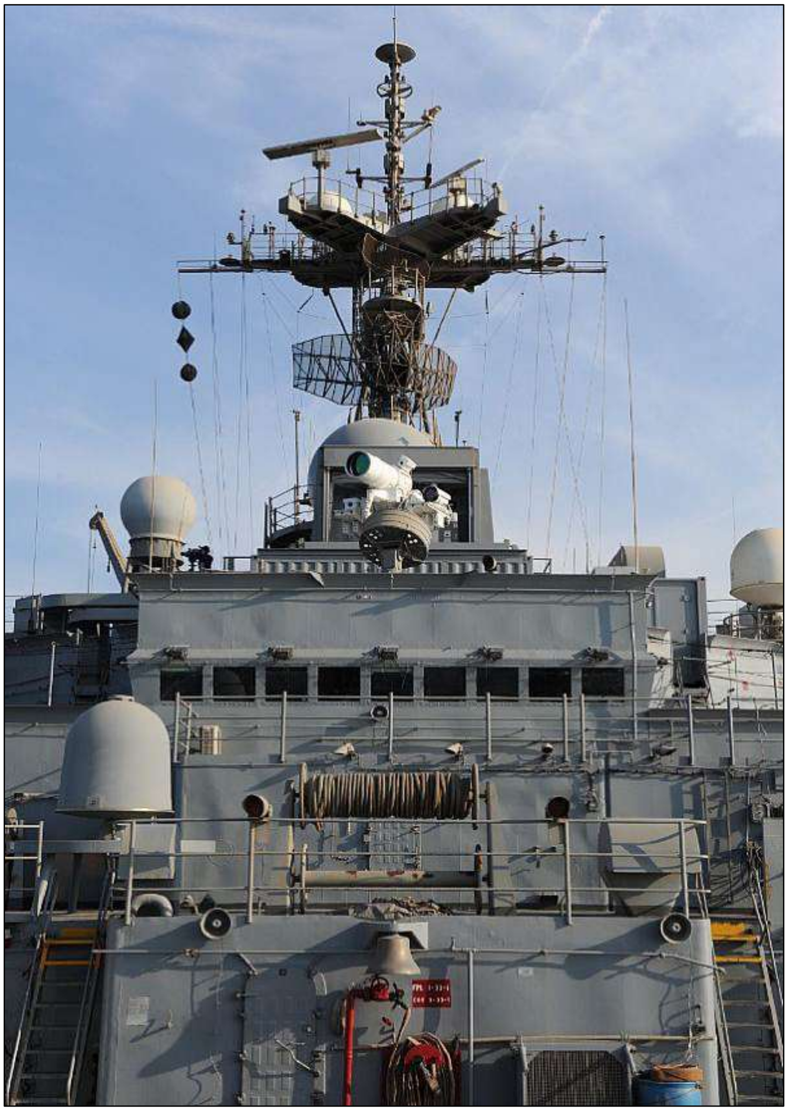

Source: Navy photograph dated November 16, 2014, accompanying David Smalley, "Historic Leap: Navy Shipboard Laser Operates in Arabian Gulf," *Navy News Service*, December 10, 2014, accessed August 12, 2015, at http://www.navy.mil/list_all.asp?id=84805.

SSL Technology Maturation (SSL-TM) Program

LaWS has a reported beam power of 30 kilowatts (kW),[21] which is strong enough to counter small boats and UAVs. As a follow-on effort to LaWS and MLD, the Navy initiated the SSL Technology Maturation (SSL-TM) program, in which industry teams led by BAE Systems, Northrop Grumman, and Raytheon, among others, competed to develop a shipboard laser with a beam power of 100 kW to 150 kW, which would provide increased effectiveness against small boats and UAVs.[22] Boosting beam power further—to something like 200 kW or 300 kW—could permit a laser to counter at least some ASCMs. Even stronger beam powers—on the order of at least several hundred kW, if not one megawatt (MW) or more—could improve a laser's effectiveness against ASCMs and enable it to counter ASBMs.[23]

---

[21] See, for example, Mike McCarthy, "Navy Authorized To Use Ship-Based Laser In Battle," *Defense Daily*, December 11, 2014: 3.

On October 22, 2015, DOD announced that it had selected Northrop Grumman as the winner of the SSL-TM competition. DOD's contract-award announcement stated:

> Northrop Grumman Space and Mission Systems Corp., Redondo Beach, California, is being awarded a $53,151,809 cost-plus-fixed-fee contract for the Solid State High Power Laser Weapon System Demonstrator (LWSD) program.... The Office of Naval Research seeks to continue the advancement of SSL weapon system designs, architectures, and component technologies. The government believes that improvements in lethality may be achieved through maturation and optimization of a variety of system characteristics, including laser power, beam quality, beam director architecture, and other physical and optical aspects of the laser, beam director, and system design. Leveraging our experience and internal investments, the Northrop Grumman team is ready to fully support the three phases of the LWSD program. This contract contains options, which if exercised, will bring the contract value to $91,057,597. Work will be performed in Redondo Beach, California, and is expected to be completed Oct. 21, 2016. If options are exercised, work will continue through July 7, 2018.... This contract was competitively procured under the Office of Naval Research broad agency announcement 15-0005 entitled "Solid State, High Power Laser Weapon System Demonstrator (LWSD) Design, Development and Demonstration for Surface Navy, USN." Six proposals were received in response to this solicitation.[24]

A December 22, 2016, Northrop Grumman news release about the October 22, 2016, contract award stated:

> During Phase 1 of the LWSD contract, Northrop Grumman will develop a detailed design for the new system. Phase 2 will include assembly and ground test of the system, while Phase 3 will comprise at-sea testing of the system aboard the Navy's Self Defense Test Ship (SDTS). The Navy will lead this testing with Northrop Grumman providing technical support. The SDTS is the former USS Paul F. Foster (DD-964).
>
> According to Renard, Northrop Grumman's LWSD is well suited to support the Navy's planned initial testing on the SDTS. The company has designed its system to be installed, however, with minimal modification or additional costs, for demonstration on the Navy's DDG-51 FLT II class destroyers.[25]

---

[22] For more on the SSL-TM program, see Office of Naval Research, "Solid-State Laser Technology Maturation Program," accessed August 11, 2015, at http://www.onr.navy.mil/Media-Center/Fact-Sheets/Solid-State-Laser-Technology-Maturation-Program.aspx; Office of Naval Research, "Solid State Laser Technology Maturation Program," September 2012, accessed August 11, 2015, at http://www.onr.navy.mil/~/media/Files/Fact-Sheets/35/Solid-State-Laser-Technology-Maturation-Program-2012-a.ashx; Office of Naval Research, "Research and Development/Technology Maturation of Solid State High Power Laser Weapon Systems, Subsystems, and/or Components for Surface Navy, USN, Broad Agency Announcement (BAA)," ONR BAA # 12-019, 2012, accessed August 11, 2015, at http://www.onr.navy.mil/~/media/files/funding-announcements/baa/2012/12-019.ashx; Future Force, "Developing a High-Energy Laser for the Navy," January 23, 2015, accessed August 11, 2015, at http://futureforce.navylive.dodlive.mil/2015/01/high-energy-laser/.

[23] For additional discussion, see CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke, particularly the section entitled "Required Laser Power Levels for Countering Targets" and Appendix A on "Laser Power Levels Required to Counter Targets."

[24] DOD contract award announcements for October 22, 2015, accessed December 18, 2015, at: http://www.defense.gov/News/Contracts/Contract-View/Article/625630.

[25] "US Navy Selects Northrop Grumman to Design and Produce Shipboard Laser Weapon System Demonstrator," (continued...)

Figure 4 shows an artist's rendering of LWSD installed on the Navy's Self Defense Test Ship (the USS Paul F. Foster [DD-964], an old Spruance [DD-963] class destroyer).

Figure 4. Laser Weapon System Demonstrator (LWSD) on Self Defense Test Ship

Artist's rendering

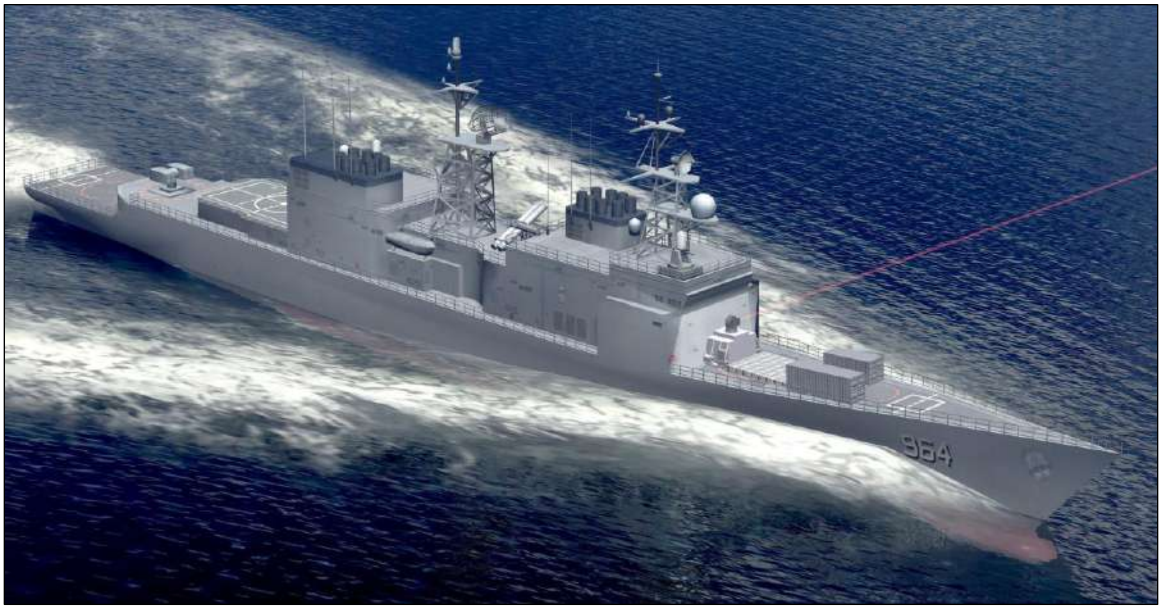

Source: Cropped version of image accessed on March 18, 2016, at http://media.globenewswire.com/cache/189/hires/39412.jpg.

Figure 5 is a detail from the above photo.

(...continued)

December 22, 2016, accessed March 18, 2016, at: http://www.globenewswire.com/newsarchive/noc/press/pages/news_releases.html?d=10158731 . See also Richard Scott, "Northrop Grumman To Build on MLD for SSL Demonstrator," *IHS Jane's International Defence Review*, February 2016: 5.

Figure 5. Laser Weapon System Demonstrator (LWSD) on Self Defense Test Ship

Artist's rendering

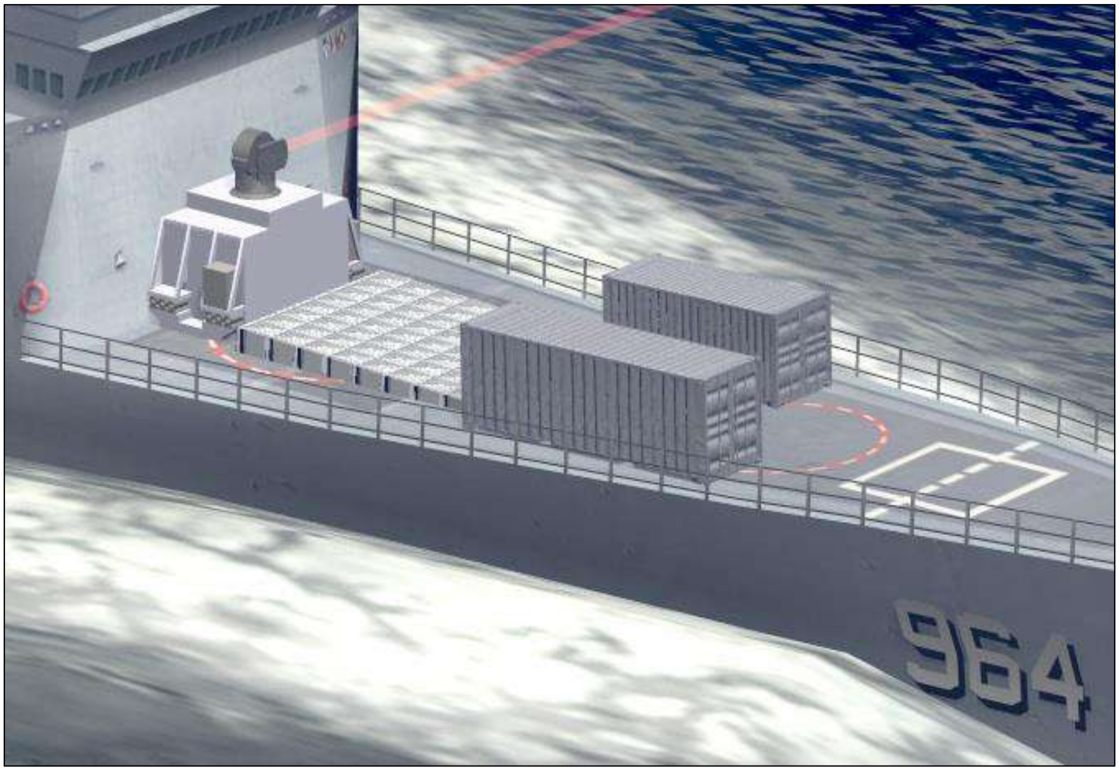

Source: Cropped version of image accessed on March 18, 2016, at http://media.globenewswire.com/cache/189/hires/39412.jpg.

Figure 6 Is an Office of Naval Research (ONR) graphic illustration of the LWSD's major components.

Figure 6. ONR Graphic of LWSD Components

Artist's rendering

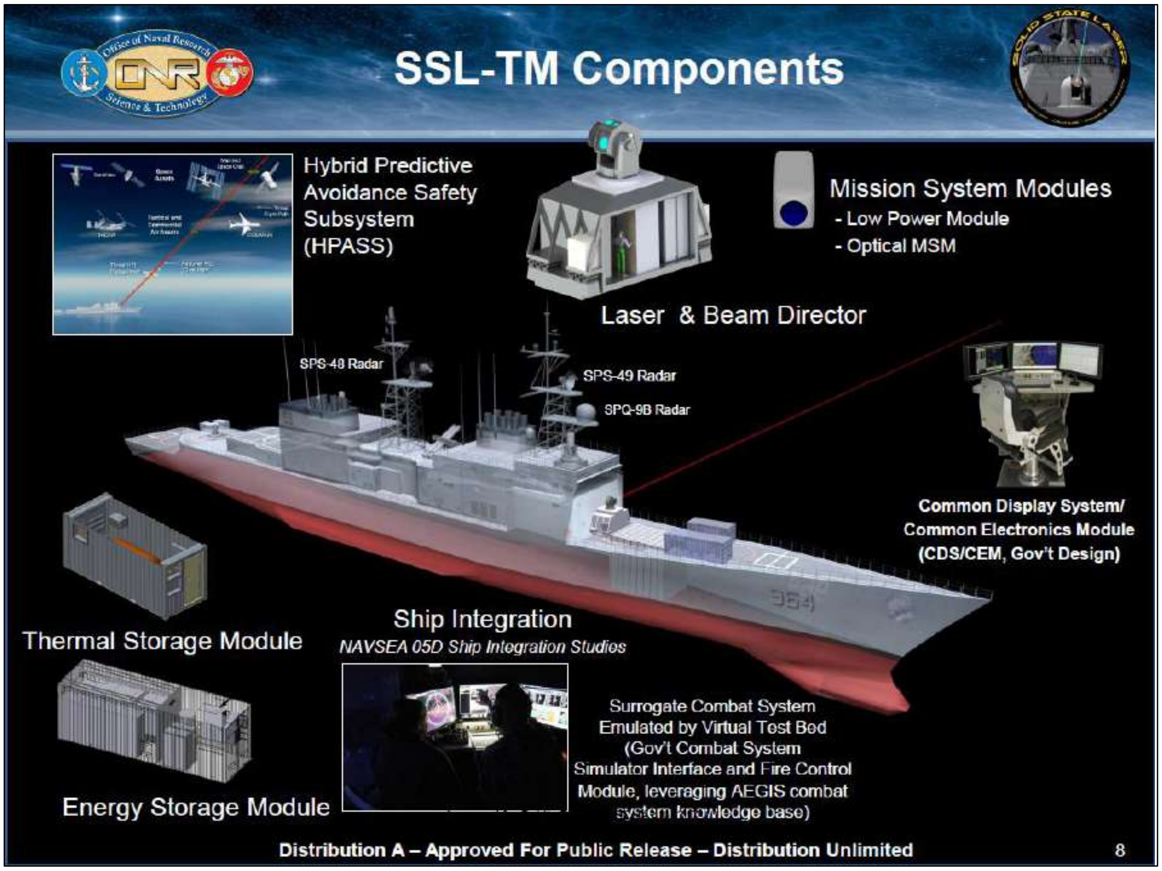

SSL-TM Components

Hybrid Predictive Avoidance Safety Subsystem (HPASS)

Mission System Modules
- Low Power Module
- Optical MSM

Laser & Beam Director

SPS-48 Radar

SPS-49 Radar

SPQ-9B Radar

Common Display System/ Common Electronics Module (CDS/CEM, Gov't Design)

Thermal Storage Module

Ship Integration

NAVSEA 05D Ship Integration Studies

Surrogate Combat System Emulated by Virtual Test Bed (Gov't Combat System Simulator Interface and Fire Control Module, leveraging AEGIS combat system knowledge base)

Energy Storage Module

Distribution A – Approved For Public Release – Distribution Unlimited

8

Source: Slide from February 2016 ONR briefing to CRS on SSL-TM program, received from Navy Office of Legislative Affairs February 26, 2016.

Directed Energy Roadmap

A July 28, 2015, press report stated:

> [Secretary of the Navy Ray] Mabus said he would release a DE [directed energy][26] roadmap this fall that "charts our course for research, development, and fielding of high power radio frequency weapons, lasers, and directed energy countermeasures. And I will follow it up with my guidance to the Program Objective Memorandum for [Fiscal Year 2018],[27] which, importantly, establishes a resource sponsor and a program of record."...
>
> Also meant to help quicken the pace of progress, the Office of Naval Research will take lessons learned from the [USS] Ponce to inform the Solid State Laser Technology Maturation program that aims to produce a 100-150 kilowatt laser prototype for at-sea testing in 2018, or sooner if possible. Rear Adm. Bryant Fuller, Naval Sea Systems Command (NAVSEA) chief engineer, said... that everything the Navy learned about rules

---

[26] Lasers and another class of weapons called high-power microwave (HPM) weapons are referred to collectively as directed-energy weapons because they achieve their effects by directing electromagnetic energy at their targets.

[27] The Program Objective Memorandum (POM) is an internal DOD document that guides the preparation of a budget for a particular fiscal year.

> of engagement and how to use LaWS in an operational environment would apply to larger laser weapons as well. Leveraging the operational knowledge Ponce gained will help the Navy field whatever comes out of the SSL-TM effort much more rapidly.
>
> In the meantime, Mabus said the Laser Weapon System (LaWS) will continue its work in the Middle East after early success led officials to extend its deployment.[28]

EMRG

In addition to SSLs, the Navy since 2005 has been developing EMRG, a cannon that uses electricity rather than chemical propellants (i.e., gunpowder charges) to fire a projectile.[29] In EMRG, "magnetic fields created by high electrical currents accelerate a sliding metal conductor, or armature, between two rails to launch projectiles at [speeds of] 4,500 mph to 5,600 mph,"[30] or roughly Mach 5.9 to Mach 7.4 at sea level.[31] Like SSLs, EMRG draws its power from the ship's overall electrical supply.[32] The Navy originally began developing EMRG as a naval surface fire support (NSFS) weapon for supporting U.S. Marines operating ashore, but subsequently determined that the weapon also has potential for defending against ASCMs and ASBMs.[33] In response to Section 243 of the FY2012 National Defense Authorization Act (H.R. 1540/P.L. 112-81 of December 31, 2011), the Navy in September 2012 submitted to the congressional defense committees a report on the EMRG development effort.[34]

Following tests with early Navy-built EMRG prototypes, the Navy funded the development of two industry-built EMRG prototype demonstrators, one by BAE Systems and the other by General Atomics (see **Figure 7** and **Figure 8**).

---

[28] Megan Eckstein, "Mabus: Adversaries Showing Interest in Directed Energy; Navy Needs to Move Faster," *USNI News*, July 28, 2015.

[29] Because it uses electricity rather than a powder charge to accelerate the projectile, Navy officials sometimes refer to EMRG as a launcher rather than a gun or cannon.

[30] Grace Jean, "With a Bang, Navy Begins Tests on EM Railgun Prototype Launcher," Navy News Service, February 28, 2012, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=65577.

[31] The speed of sound in air (i.e., Mach 1), varies with altitude; at sea level, it is approximately 761 miles an hour. (See for example, the table entitled "Speed of Sound at Different Altitudes," accessed August 12, 2015, at http://www.fighter-planes.com/jetmach1.htm.

[32] Unlike SSLs, however, EMRG is not a directed energy weapon, because it achieves its effects by firing a physical projectile at the target, not by directing electromagnetic energy at the target. See also footnote 26.

[33] For a recent article discussing the use of EMRG in countering ASCMs and ASBMs, see Sam LaGrone, "Navy Wants Rail Guns to Fight Ballistic and Supersonic Missiles Says RFI," *USNI News*, January 5, 2015.

[34] U.S. Navy, *Electromagnetic Railgun System: Final Report to the Congressional Defense Committees*, August 2012, with cover letters dated September 18, 2012. For a press report discussing the Navy's report to Congress, see Dan Taylor, "Stackley: Navy Identifies Four Technical Hurdles To Railgun Development," *Inside the Navy*, November 19, 2012.

Figure 7. Industry-Built EMRG Prototype Demonstrator

BAE prototype

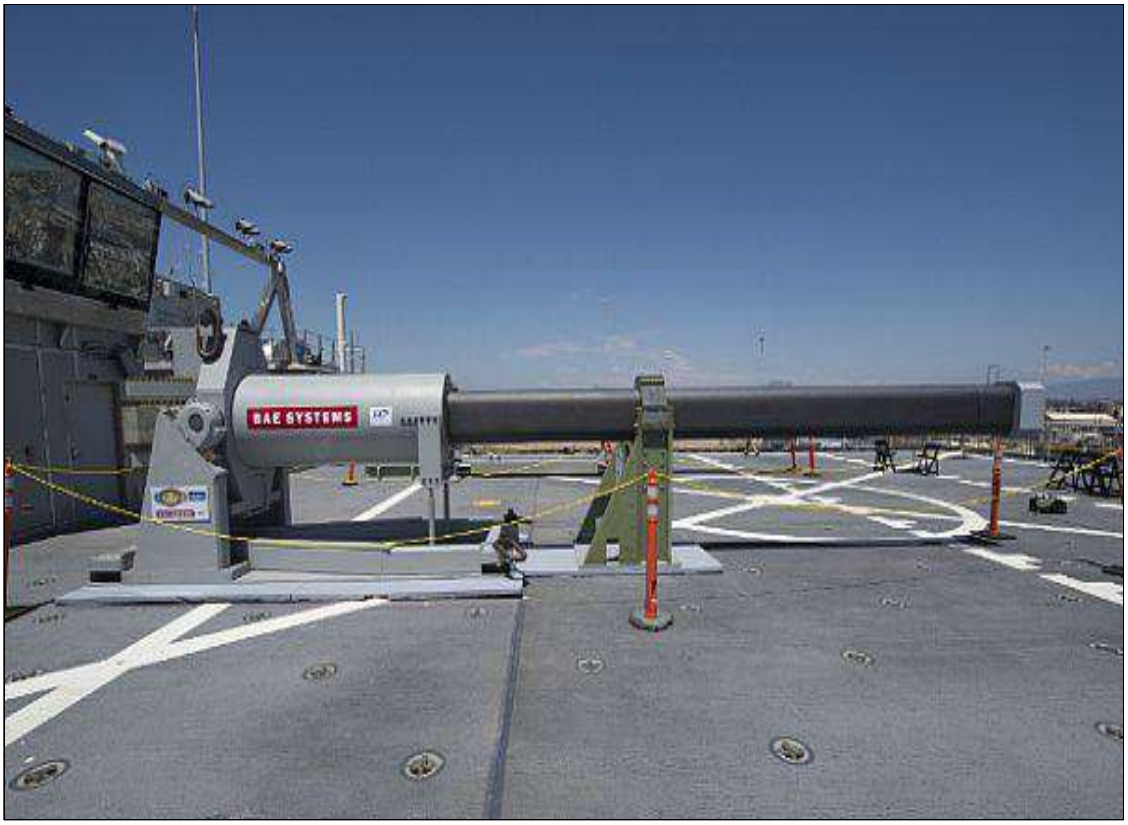

Source: Navy photograph dated July 8, 2014, associated with Office of Naval Research Public Affairs, "From Research to Railgun: Revolutionary Weapon at Future Force EXPO," *Navy News Service*, January 13, 2015, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=85166.

Figure 8. Industry-Built EMRG Prototype Demonstrator

General Atomics prototype

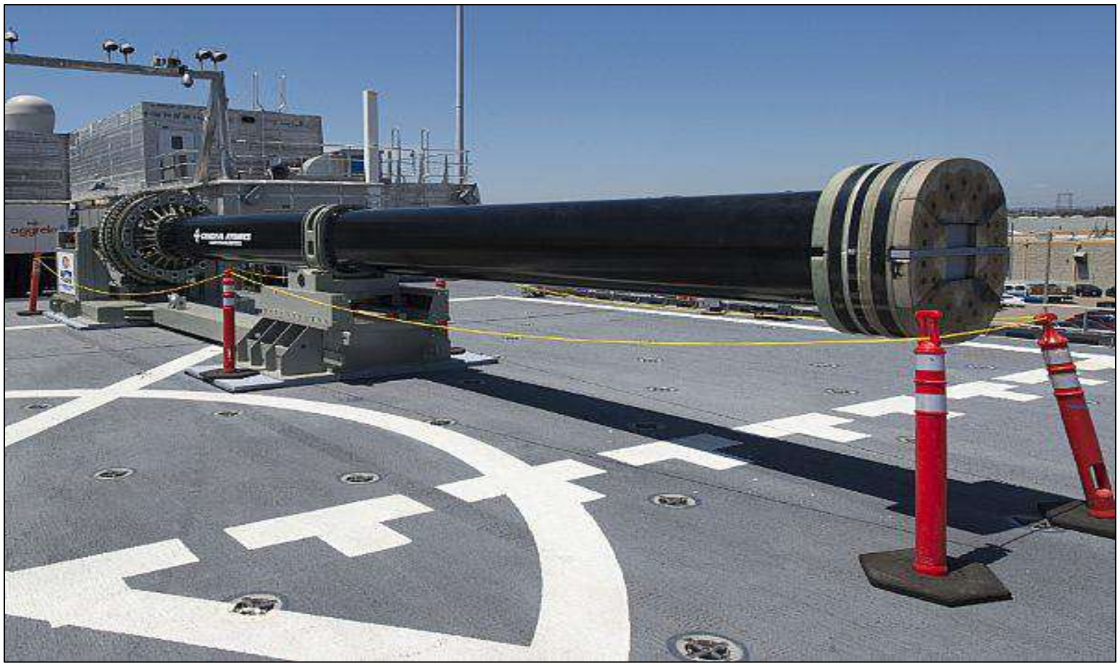

Source: navy photograph dated July 8, 2014, accessed August 12, 2015, at http://www.navy.mil/view_image.asp?id=180994.

The two industry-built prototypes are designed to fire projectiles at energy levels of 20 to 32 megajoules,[35] which is enough to propel a projectile 50 to 100 nautical miles.[36] (Such ranges might refer to using the EMRG for NSFS missions. Intercepts of ASCMs and ASBMs might take place at much shorter ranges.) The Navy began evaluating the two industry-built prototypes in 2012.

In April 2014, the Navy announced that it plans to temporarily install a prototype EMRG aboard a Navy Joint High Speed Vessel (JHSV) in FY2016, for use in at-sea tests.[37] **Figure 9** is an artist's rendering of that installation.

In January 2015, it was reported that the Navy is projecting that EMRG could become operational on a Navy ship between 2020 and 2025.[38] In April 2015, it was reported that the Navy is considering installing an EMRG on a Zumwalt (DDG-1000) class destroyer by the mid-2020s.[39]

---

[35] The Navy states that "a megajoule is a measurement of energy associated with a mass traveling at a certain velocity. In simple terms, a one-ton vehicle moving at 100 mph equals a magajoule of energy." (Office of Naval Research Public Affairs, "Navy Sets New World Record with Electromagnetic Railgun Demonstration," *Navy News Service*, December 10, 2010, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=57690.)

[36] Grace Jean, "With a Bang, Navy Begins Tests on EM Railgun Prototype Launcher," Navy News Service, February 28, 2012, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=65577.

[37] Naval Sea Systems Command Office of Corporate Communication, "Navy to Deploy Electromagnetic Railgun Aboard JHSV," *Navy News Service*, April 7, 2014, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=80055.

[38] Sam LaGrone, "Navy Wants Rail Guns to Fight Ballistic and Supersonic Missiles Says RFI," *USNI News*, January 5, 2015.

[39] Sam LaGrone, "Navy Considering Railgun for Third Zumwalt Destroyer," *USNI News*, February 5, 2015 (updated (continued...)

Figure 9. EMRG Prototype Demonstrator Installed on a JHSV

Artist's rendering

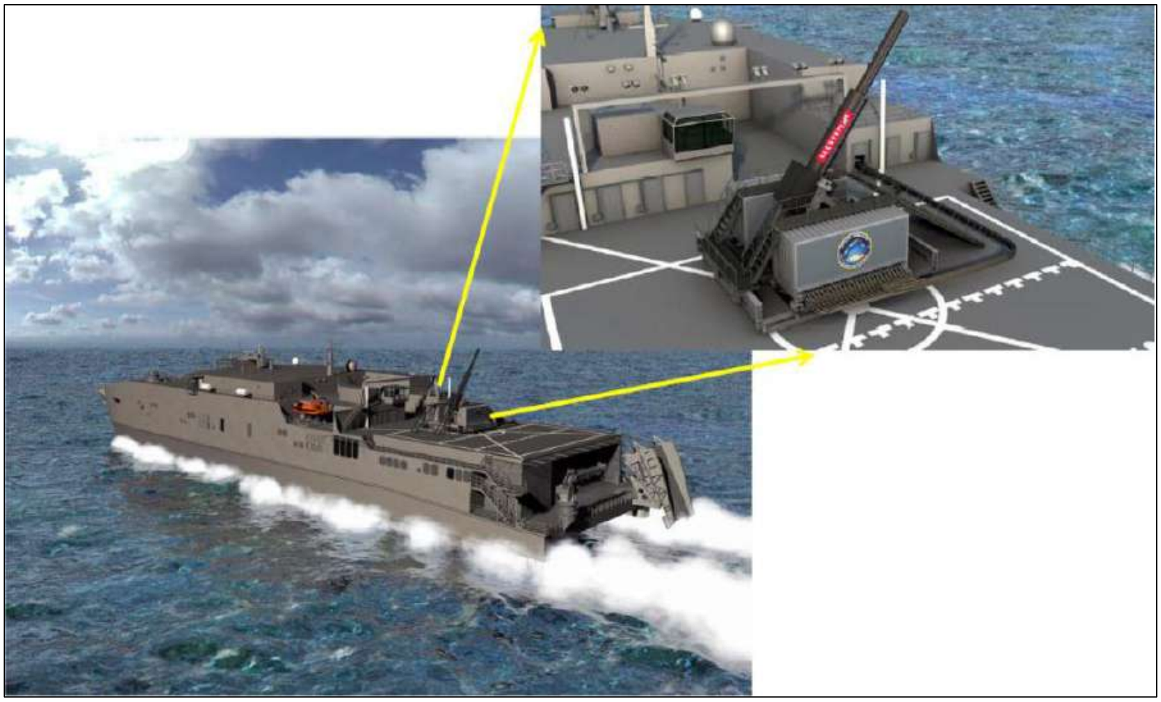

Source: Briefing slide entitled "FY16 At-Sea Test of Railgun" in Navy briefing entitled "Railgun Program Overview," undated but posted at InsideDefense.com on April 14, 2015. (InsideDefense.com states that the briefing was presented at a public conference on April 14, 2015.)

HVP

As the Navy was developing EMRG, it realized that the guided projectile being developed for EMRG could also be fired from 5-inch and 155mm powder guns.[40] Navy cruisers each have two

---

(...continued)

February 11, 2015); Mike McCarthy, "Navy Aiming To Put Railgun On Third Zumwalt Destroyer," *Defense Daily*, February 6, 2015; Kris Osborn, "Navy Will Test its Electromagnetic Rail Gun aboard DDG 1000," *DefenseTech*, April 15, 2015. For more on Zumwalt-class destroyers, see CRS Report RL32109, *Navy DDG-51 and DDG-1000 Destroyer Programs: Background and Issues for Congress*, by Ronald O'Rourke.

[40] The Navy describes the HVP as "a next generation, common, low drag, guided projectile capable of completing multiple missions for gun systems such as the Navy 5-Inch, 155-mm, and future railguns.... HVP's low drag aerodynamic design enables high velocity, maneuverability, and decreased time-to-target. These attributes coupled with accurate guidance electronics provide low cost mission effectiveness against current threats and the ability to adapt to air and surface threats of the future." (Office of Naval Research, Hypervelocity Projectile," September 2012, accessed August 14, 2015, at http://www.onr.navy.mil/~/media/Files/Fact-Sheets/35/Hypervelocity-Projectile-2012B.ashx.) The Navy states that HVP weighs 23 pounds. (Source: David Martin, "Navy's Newest Weapon Kills at Seven Times the Speed of Sound," *CBS News (cbssnews.com)*, April 7, 2014.)

BAE Systems states that HVP is 24 inches long and weighs 28 pounds, including a 15-pound payload. The total length and weight of an HVP launch package, BAE Systems states, is 26 inches and 40 pounds. BAE states that the maximum rate of fire for HVP is 20 rounds per minute from a Mk 45 5-inch gun, 10 rounds per minute from the 155mm gun on DDG-1000 class destroyers (called the Advanced Gun System, or AGS), and 6 rounds per minute from EMRG. HVP's firing range, BAE Systems states, is more than 40 nautical miles (when fired from a Mk 45 Mod 2 5-inch gun), more than 50 nautical miles (Mk 45 Mod 4 5-inch gun), more than 70 nautical miles (155mm gun on DDG-1000 class destroyers), and more than 100 nautical miles (EMRG). (BAE Systems, "Hypervelocity Projectile (HVP)," 2014, (continued...)

5-inch guns, and most Navy destroyers each have one 5-inch gun. The Navy's three new Zumwalt class (DDG-1000) destroyers, which are under construction, each have two 155mm guns.

The projectile is a hypervelocity projectile when fired from either EMRG or a powder gun, but the term HVP tends to be used more frequently in connection with the concept of firing it from a powder gun. **Figure 10** and **Figure 11** show the HVP.

Figure 10. Photograph Showing HVP

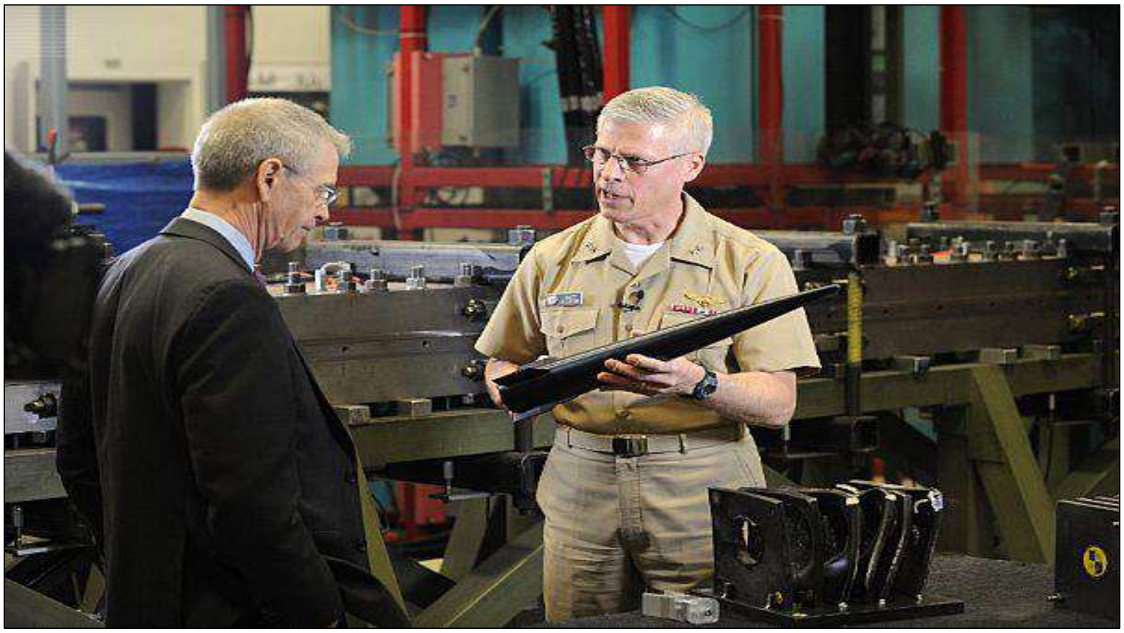

Source: Navy photograph dated April 4, 2014, with a caption that reads: "Rear Adm. Matthew Klunder, chief of naval research, shows off a Hypervelocity Projectile (HVP) to CBS News reporter David Martin during an interview held at the Naval Research Laboratory's materials testing facility. The HVP is a next-generation, common, low drag, guided projectile capable of completing multiple missions for gun systems such as the Navy 5-inch, 155-mm, and future railguns," accessed August 12, 2015, at http://www.navy.mil/view_image.asp?id=174517.

(...continued)

accessed August 14, 2015, at http://www.baesystems.com/download/BAES_178505/hyper-velocity-projectile-hvp-datasheet.)

In July 2015, the Navy issued a request for information (RFI) to industry for the fabrication of a prototype EMRG mount capable of handling an integrated launch weight package of 22 kg, or about 48.5 pounds. (RFI for Fabrication of Prototype Mount for Naval Railgun, Solicitation Number: N00024-15-R-4132, FedBizOpps.gov, July 29, 2015. See also Justin Doubleday, "Navy Developing Integrated Mount For Electromagnetic Railgun," *Inside the Navy*, July 31, 2015.)

Figure 11. HVP

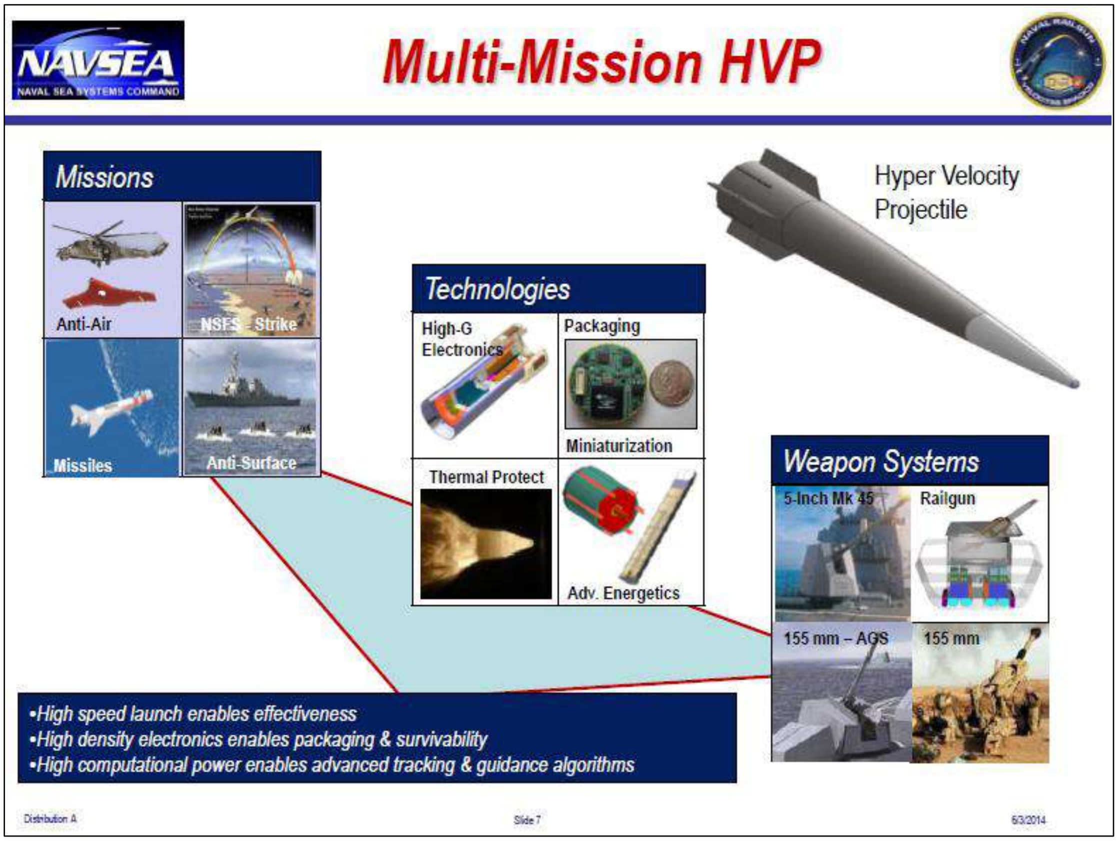

Multi-Mission HVP

Missions

Anti-Air

NSFS - Strike

Missiles

Anti-Surface

Technologies

High-G Electronics

Packaging

Miniaturization

Thermal Protect

Adv. Energetics

Hyper Velocity Projectile

Weapon Systems

5-Inch Mk 45

Railgun

155 mm – AGS

155 mm

- High speed launch enables effectiveness
- High density electronics enables packaging & survivability
- High computational power enables advanced tracking & guidance algorithms

Distribution A

Slide 7

6/3/2014

Source: Slide 7 from Navy briefing entitled "Electromagnetic Railgun," NDIA Joint Armaments Forum, Exhibition & Technology Demonstration, May 14, 2014, LCDR Jason Fox, USN, Assistant PM [Program Manager], Railgun Ship Integration, Distribution A, Approved for Public Release, accessed August 13, 2015, at http://www.dtic.mil/ndia/2014armaments/WedFox.pdf.

When fired from 5-inch powder guns, the projectile achieves a speed of roughly Mach 3, which is roughly half the speed it achieves when fired from EMRG, but more than twice the speed of a conventional 5-inch shell fired from a 5-inch gun.[41] This is apparently fast enough for countering at least some ASCMs. The Navy states that "The HVP—combined with the MK 45 [5-inch gun][42]—will support various mission areas including naval surface fire support, and has the capacity to expand to a variety of anti-air threats, [and] anti-surface [missions], and could expand the Navy's engagement options against current and emerging threats."[43]

One advantage of the HVP/5-inch gun concept is that the 5-inch guns are already installed on Navy cruisers and destroyers, creating a potential for rapidly proliferating HVP through the cruiser-destroyer force, once development of HVP is complete and the weapon has been integrated into cruiser and destroyer combat systems. **Figure 12** shows HVP launch packages configured for 5-inch guns, 155mm guns, and EMRG.

---

[41] Source: Sam LaGrone, "Updated: Navy Researching Firing Mach 3 Guided Round from Standard Deck Guns," *USNI News*, June 1, 2015 (updated June 2, 2015).

[42] The type of 5-inch gun on Navy cruisers and destroyers is called the Mark 45.

[43] Naval Surface Warfare Center Dahlgren Division Corporate Communications, "DEPSECDEF Loads HVP on Test Range, Observes Repetitive Rate Electromagnetic Railgun's Commissioning Series," *Navy News Service*, May 8, 2015, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=86987.

Figure 12. HVP Launch Packages

Launch packages for 5-inch gun, 155mm gun, and EMRG

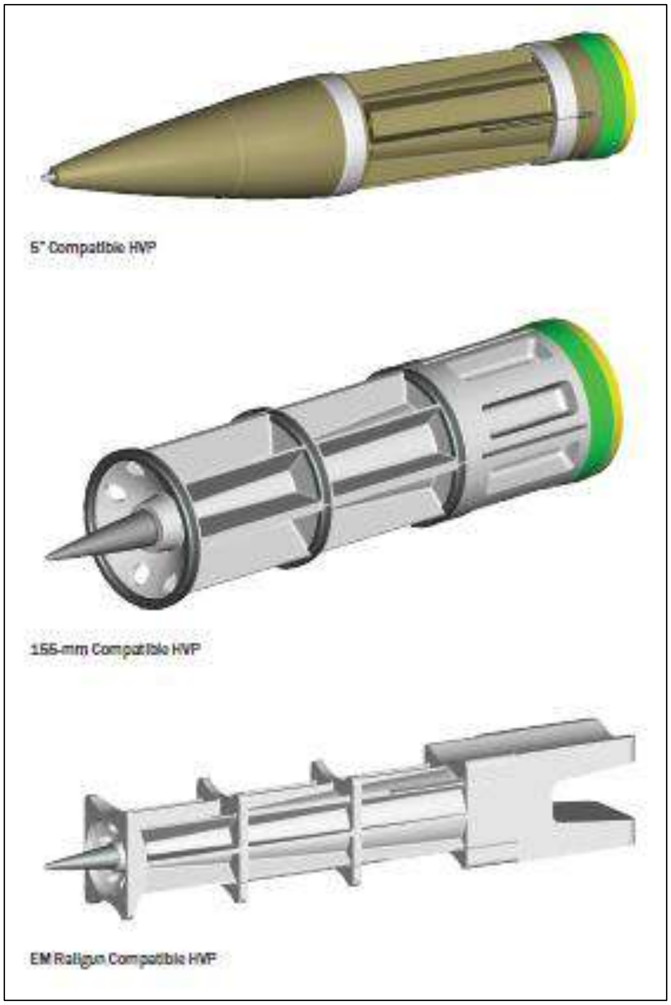

Source: BAE Systems, "Hypervelocity Projectile (HVP)," 2014, accessed August 14, 2015, at http://www.baesystems.com/download/BAES_178505/hyper-velocity-projectile--datasheet.

Figure 13 is a slide showing the potential application of HVP to 5-inch power guns, 155mm powder guns, and EMRG. The first line of the slide, for example, discusses HVP's use with 5-inch powder guns, stating that it uses a high-explosive (HE) warhead for the NSFS mission;[44] that a total of 113 5-inch gun barrels are available in the fleet (which could be a reference to 22 cruisers with two guns each, and 69 destroyers with one gun each); and that as a game-changing capability, it is guided and can be used at ranges of up to 26 nautical miles to 41 nautical miles for NSFS operations, for countering ASCMs, and for anti-surface warfare (ASuW) operations (i.e., attacking surface ships and craft).

---

[44] The "KE" in the next line down means that when fired from EMRG, the projectile can alternatively attack targets using its own kinetic energy (i.e., by simply impacting the target at hypersonic speed).

Figure 13. HVP Application to Various Launchers

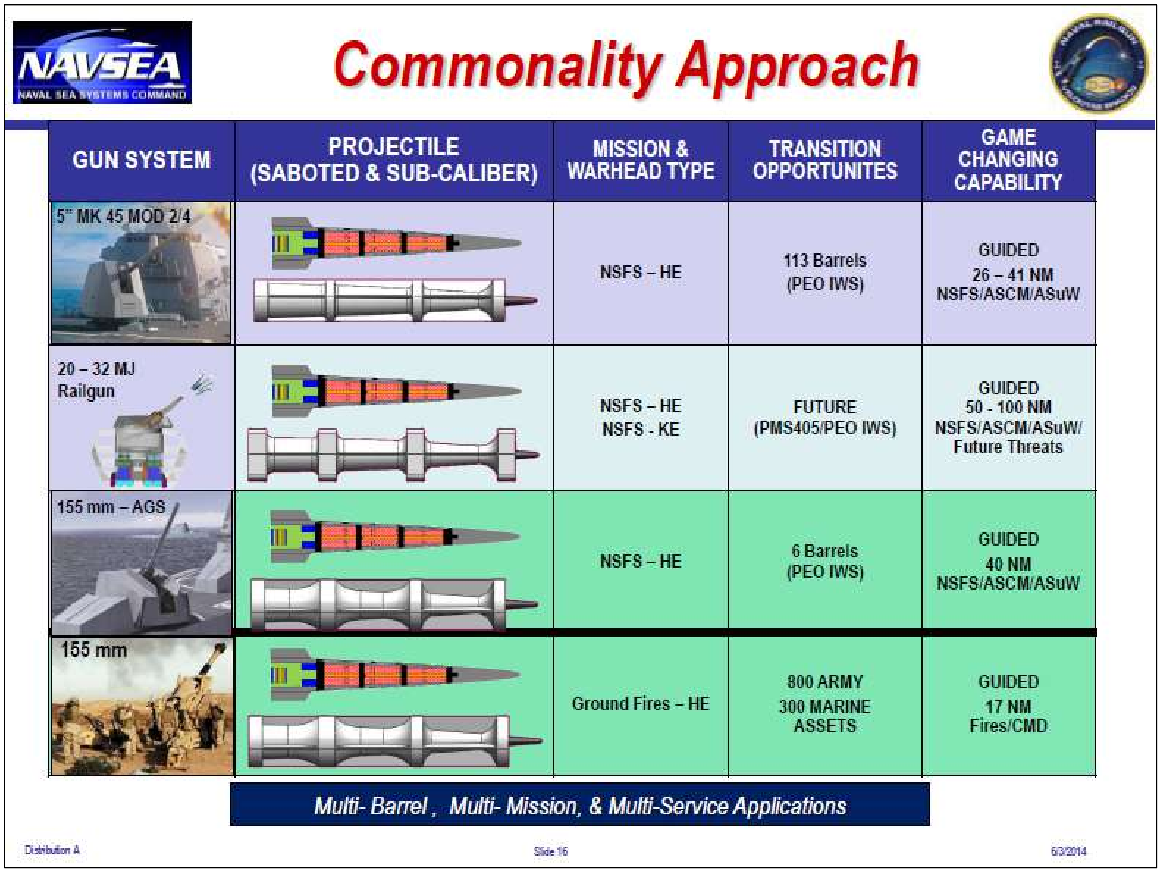

NAVSEA
NAVAL SEA SYSTEMS COMMAND

Commonality Approach

GUN SYSTEM	PROJECTILE (SABOTED & SUB-CALIBER)	MISSION & WARHEAD TYPE	TRANSITION OPPORTUNITES	GAME CHANGING CAPABILITY
5" MK 45 MOD 2/4		NSFS – HE	113 Barrels (PEO IWS)	GUIDED 26 – 41 NM NSFS/ASCM/ASuW
20 – 32 MJ Railgun		NSFS – HE NSFS - KE	FUTURE (PMS405/PEO IWS)	GUIDED 50 - 100 NM NSFS/ASCM/ASuW/ Future Threats
155 mm – AGS		NSFS – HE	6 Barrels (PEO IWS)	GUIDED 40 NM NSFS/ASCM/ASuW
155 mm		Ground Fires – HE	800 ARMY 300 MARINE ASSETS	GUIDED 17 NM Fires/CMD

Multi- Barrel , Multi- Mission, & Multi-Service Applications

Distribution A

Slide 16

6/3/2014

Source: Slide 16 from Navy briefing entitled "Electromagnetic Railgun," NDIA Joint Armaments Forum, Exhibition & Technology Demonstration, May 14, 2014, LCDR Jason Fox, USN, Assistant PM [Program Manager], Railgun Ship Integration, Distribution A, Approved for Public Release, accessed August 13, 2015, at http://www.dtic.mil/ndia/2014armaments/WedFox.pdf.

Figure 14 is a not-to-scale illustration of how HVPs fired from EMRGs and 5-inch guns can be used to counter various targets, including ASCMs and ASBMs.

Figure 14. Navy Slide Depicting Operations Against Various Target Types

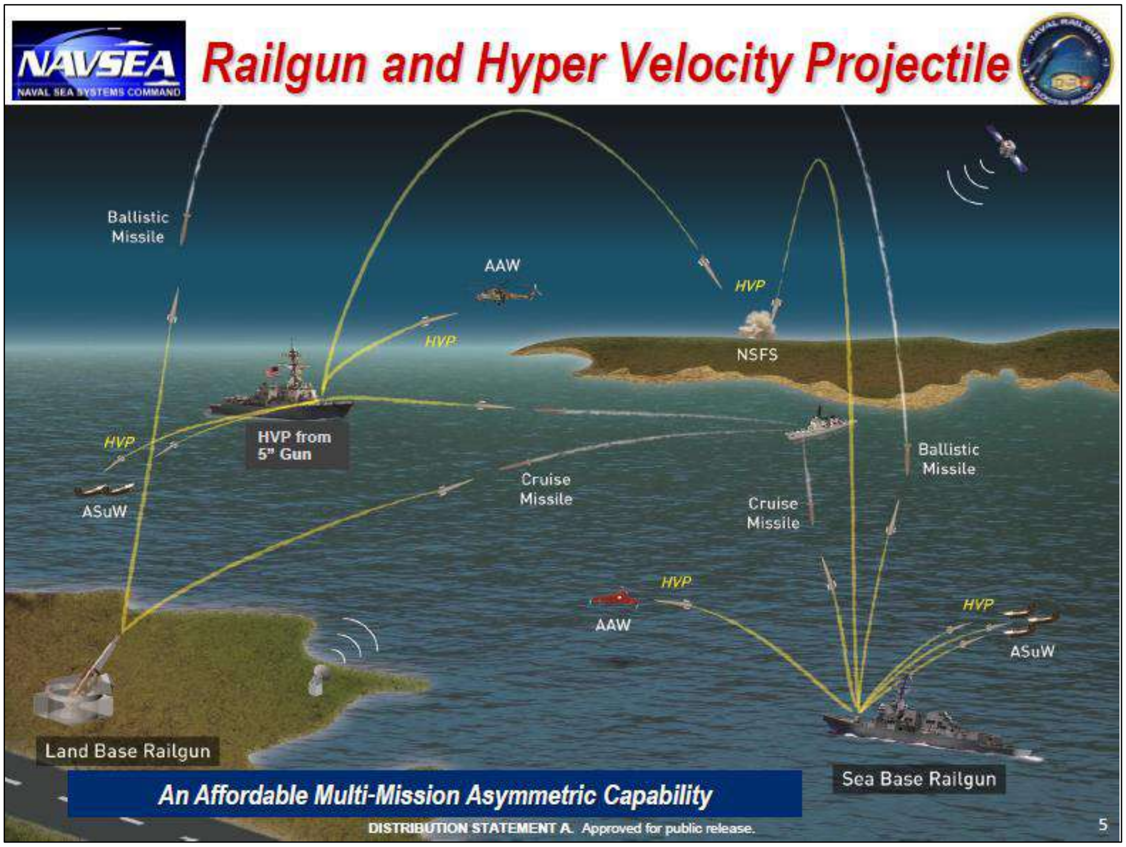

NAVSEA NAVAL SEA SYSTEMS COMMAND

Railgun and Hyper Velocity Projectile

An Affordable Multi-Mission Asymmetric Capability

DISTRIBUTION STATEMENT A. Approved for public release.

5

Source: Slide 5 from Navy briefing entitled "Electromagnetic Railgun," NDIA Joint Armaments Forum, Exhibition & Technology Demonstration, May 14, 2014, LCDR Jason Fox, USN, Assistant PM [Program Manager], Railgun Ship Integration, Distribution A, Approved for Public Release, accessed August 13, 2015, at http://www.dtic.mil/ndia/2014armaments/WedFox.pdf.

An April 11, 2016, press report states:

> The Pentagon wants to take a weapon originally designed for offense, flip its punch for defense and demonstrate by 2018 the potential for the Army and Navy to conduct missile defense of bases, ports and ships using traditional field guns to fire a new hypervelocity round guided by a mobile, ground variant of an Air Force fighter aircraft radar.
>
> The Strategic Capabilities Office is working with the Army, Navy and Air Force to craft a Hypervelocity Gun Weapon System that aims, in part, to provide China and Russia an example of a secret collection of new U.S. military capabilities the Defense Department is bringing online in an effort to strengthen conventional deterrence.
>
> "It is a fantastic program," Will Roper, Strategic Capabilities Office director, said in a March 28 interview with reporters, who said the project aims "to completely lower the cost of doing missile defense" by defeating missile raids at a lower cost per round and, as a consequence, imposing higher costs on attackers.[45]

A May 5, 2016, press report states:

---

[45] Jason Sherman, "SCO Aims To Flip The Script on Missile Defense With Hypervelocity Gun," *Inside the Navy*, April 11, 2016.

> Come January [2017], the Pentagon will almost assuredly have new leadership, complete with a new vision for how the Department of Defense should operate, organize and plan for the future.
>
> It's a reality facing down Defense Secretary Ash Carter and Deputy Secretary Bob Work as they try to complete a transformation at the Pentagon, one which both men have said is vital to making sure the US is able to maintain its technological edge against great powers like Russia and China in the future.
>
> While Carter has made his reach-out to nontraditional communities the centerpiece of his initiative, Work's focus has been the "Third Offset," a series of technological bets, largely focused on man-unmanned teaming, that he believes will pay off in the future.
>
> So how do you make sure that a long-term project like the Third Offset keeps going under a new administration?
>
> "We have three really complimentary ways to go about this," Work said Monday at an event hosted by the Atlantic Council.
>
> The first, he explained, is giving many options to the next administration.
>
> "First of all, I'm going to be central to the transition, so I am going to be able to personally talk with the transition team and explain to them what we have pursued and why we have pursued it, and let them make their own decisions," Work said.
>
> "One of the things we have done in our program is build in a lot of different options that they can pull levers on," Work explained.
>
> As an example, he pointed to the idea of an electromagnetic railgun. Initially, Work and his team thought that was an area that would be a major focus of development, but as they experimented they realized that a powder gun with a hypervelocity round could have almost the same impact — but at a fraction of the cost, because it did not require the development, testing and adaptation of a new gun.
>
> "We're going to say 'look, this is the place where [we think] you want to put your money,' but we're going to have enough money in both the electromagnetic railgun and the powder gun that if the new administration says 'I really want the electromagnetic railgun, this is the way I want to go,' knock yourself out," Work said. "We've set you up for success."[46]

Indirectly Improving Ability to Counter ASCMs and ASBMs

As discussed earlier, SSLs currently under development have enough beam power to counter small boats and UAVs, but not enough to ASCMs or ASBMs. Even so, such SSLs could indirectly improve a ship's ability to counter ASCMs and ASBMs by permitting the ship to use fewer of its SAMs for countering UAVs, and more of them for countering ASCMs and ASBMs. Similarly, even though HVPs fired from 5-inch powder guns would not be able to counter ASBMs, they could indirectly improve a ship's ability to counter ASBMs by permitting the ship to use fewer of its SAMs for countering ASCMs and more of its SAMs for countering ASBMs.

Remaining Development Challenges

Although the Navy in recent years has made considerable progress in developing SSLs, EMRG, and HVP, a number of significant development challenges remain. Overcoming these challenges will likely require years of additional development work, and ultimate success in overcoming them is not guaranteed.[47]

---

[46] Aaron Mehta, "Pentagon No. 2: How to Keep Third Offset Going in the Next Administration," *Defense News*, May 5, 2016.

SSLs

As shown in **Figure 15**, remaining development challenges for SSLs include, among other things, making the system rugged enough for extended shipboard use, making the beam director (the telescope-like part of the laser that sends the beam toward the target) suitable for use in a marine environment (where moisture and salt in the air can be harsh on equipment), and integrating the system into the ship's electrical power system and combat system.

A January 23, 2015, blog post co-authored by the Office of Naval Research's program officer for the Navy's SSL program states:

> In the near term, many challenges remain to develop and operate high-energy laser systems in the maritime environment that are unique to the Navy and Marine Corps. Among these challenges is dealing with the heat generated as power levels increase. A second issue is packing sufficient power on the platform, which will require advanced battery, generator, power conditioning, and hybrid energy technologies. Current laser technologies are approximately 30 percent electrically efficient. Corrosion and contamination of optical windows by shipboard salt spray, dirt, and grime also are technical challenges. In addition, atmospheric turbulence resulting from shifting weather conditions, moisture, and dust is problematic. Turbulence can cause the air over long distances to act like a lens, resulting in the laser beam's diffusing and distorting, which degrades its performance.
>
> Much progress has been made in demonstrating high-energy laser weapon systems in the maritime environment, but there is still much to be done. Additional advances will be required to scale power levels to the hundreds of kilowatts that will make high[-]energy lasers systems robust, reliable, and affordable. Higher power levels are important for the ability to engage more challenging threats and improve the rate and range at which targets can be engaged.
>
> The programs managed by ONR are addressing these remaining issues while positioning this important warfighting capability toward an acquisition program and eventual deployment with the fleet and force.[48]

---

[47] Laser skeptics sometimes note that laser proponents over the years have made numerous predictions about when lasers might enter service with DOD, and that these predictions repeatedly have not come to pass. Viewing this record of unfulfilled predictions, skeptics might argue that "lasers are X years in the future—and always will be." Laser proponents acknowledge the record of past unfulfilled predictions, but argue that the situation has now changed because of rapid advancements in SSL technology and a shift from earlier ambitious goals (such as developing megawatt-power lasers for countering targets at tens or hundreds of miles) to more realistic goals (such as developing kilowatt-power lasers for countering targets at no more than a few miles). Laser proponents might argue that laser skeptics are vulnerable to what might be called cold plate syndrome (i.e., a cat that sits on a hot plate will not sit on a hot plate again—but it will not sit on a cold plate, either).

[48] Peter Morrison and Dennis Sorenson, "Developing a High-Energy Laser for the Navy," Future Force, January 23, 2015, accessed August 13, 2015, at http://futureforce.navylive.dodlive.mil/2015/01/high-energy-laser/. The authors are identified at the end of the post as follows: "Peter Morrison is the Office of Naval Research's program officer for the Navy's Solid-State Laser program. Dennis Sorenson is a contractor with the Office of Naval Research."

Figure 15. Development Challenges for SSLs

As of February 2013

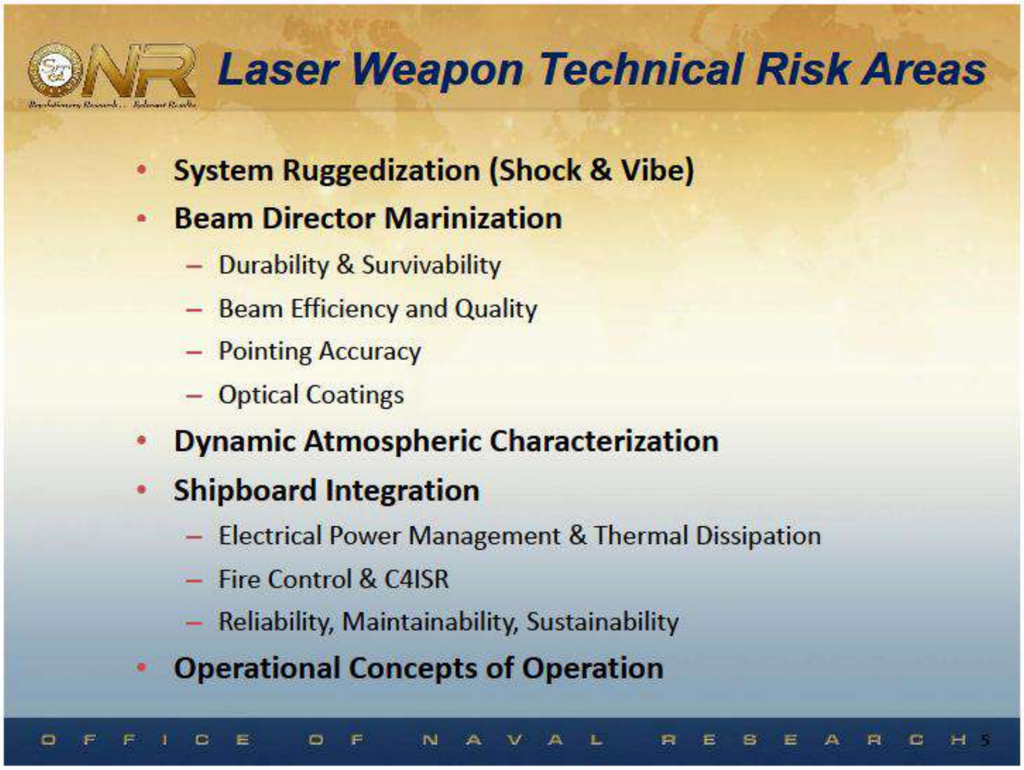

Laser Weapon Technical Risk Areas

- **System Ruggedization (Shock & Vibe)**
- **Beam Director Marinization**
  - Durability & Survivability
  - Beam Efficiency and Quality
  - Pointing Accuracy
  - Optical Coatings
- **Dynamic Atmospheric Characterization**
- **Shipboard Integration**
  - Electrical Power Management & Thermal Dissipation
  - Fire Control & C4ISR
  - Reliability, Maintainability, Sustainability
- **Operational Concepts of Operation**

OFFICE OF NAVAL RESEARCH

Source: Slide from Navy briefing entitled "Navy Solid State Laser Program Overview," ASNE Day 2013, Mr. Peter "Rollie" Morrison, ONR 35 S&T Program Office, February 22, 2013, accessed August 13, 2015, at https://www.navalengineers.org/ProceedingsDocs/ASNEDay2013/Morrison_Pres.pdf.

EMRG and HVP

As shown in **Figure 16**, remaining development challenges for EMRG involve items relating to the gun itself (including increasing barrel life to desired levels), the projectile, the weapon's electrical power system, and the weapon's integration with the ship. Fielding HVP on cruisers and destroyers ships equipped with 5-inch and 155mm powder guns would additionally require HVP to be integrated with the combat systems of those ships.

Figure 16. Development Challenges for EMRG

As of May 2014

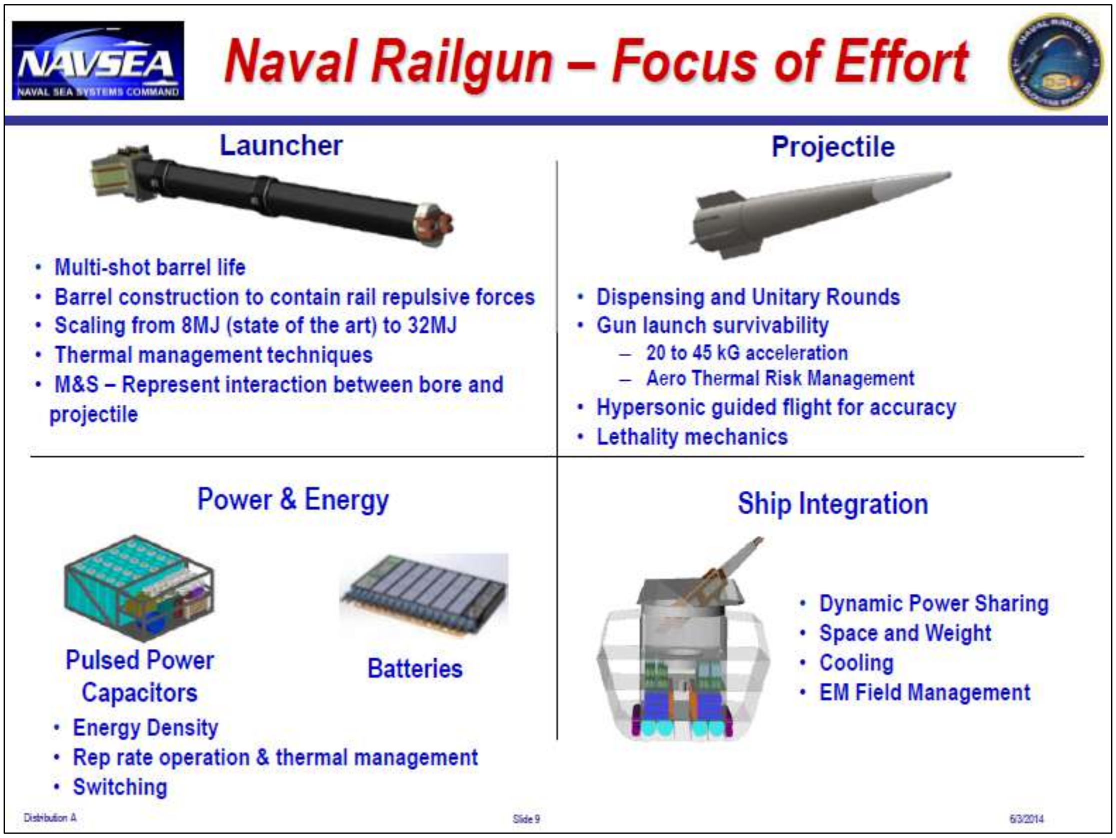

NAVSEA
NAVAL SEA SYSTEMS COMMAND

Naval Railgun – Focus of Effort

Launcher

- Multi-shot barrel life
- Barrel construction to contain rail repulsive forces
- Scaling from 8MJ (state of the art) to 32MJ
- Thermal management techniques
- M&S – Represent interaction between bore and projectile

Projectile

- Dispensing and Unitary Rounds
- Gun launch survivability
  - 20 to 45 kG acceleration
  - Aero Thermal Risk Management
- Hypersonic guided flight for accuracy
- Lethality mechanics

Power & Energy

Pulsed Power Capacitors

Batteries

- Energy Density
- Rep rate operation & thermal management
- Switching

Ship Integration

- Dynamic Power Sharing
- Space and Weight
- Cooling
- EM Field Management

Distribution A

Slide 9

6/3/2014

Source: Slide 9 from Navy briefing entitled "Electromagnetic Railgun," NDIA Joint Armaments Forum, Exhibition & Technology Demonstration, May 14, 2014, LCDR Jason Fox, USN, Assistant PM [Program Manager], Railgun Ship Integration, Distribution A, Approved for Public Release, accessed August 13, 2015, at http://www.dtic.mil/ndia/2014armaments/WedFox.pdf.

The Navy states:

> The EMRG effort began in FY 2005 with a focus on the barrel, power storage, and rail technology. In 2015, the Navy is testing full-scale industry advanced composite launchers for structure strength and manufacturability, and has advanced the pulsed-power system design from single-shot to actively cooled repeated rate operations. Building on the success of the first phase, the second phase started in 2012 with a focus on developing equipment and techniques to fire ten rounds per minute. Thermal-management techniques required for sustained firing rates are in development for both the launcher system and the pulsed-power system. The Office of Naval Research will develop a tactical prototype EMRG launcher and pulsed-power architecture suitable for advanced testing both afloat and ashore. Railgun demonstration has been funded to occur in FY 2016.[49]

A June 2015 press report states:

> As the Navy prepares to test its electromagnetic railgun at sea for the first time in 2016, service leaders said one of the biggest challenges will be integrating the new technology onto existing platforms.....

---

[49] U.S. Navy, *U.S. Navy Program Guide 2015*, p. 169.

> [Vice Adm. William Hilarides, commander of Naval Sea Systems Command] said he is positive the Navy will successfully demonstrate the weapon's ability to fire from the Trenton, but one of the biggest challenges will be configuring the railgun so that it fits within the power structure of other existing platforms.
>
> "Those are not 600-ton margin ships," he said [meaning ships with 600 tons of growth margin available to accommodate EMRG]. "If they have 60 tons, if they have 16 tons, then we'll be talking about what do we take off our existing destroyers, cruisers and other ships in order to get this incredible capability [on them]."
>
> These types of discussions are influencing ship designs as program managers look at what systems are indispensable and what can be exchanged, Hilarides said.
>
> Integrating the railgun into the fleet won't be a swift process.
>
> It will be at least 10 years until the railgun is fielded on new ships and potentially 30 years past that before the Navy considers removing powder guns from the fleet entirely and transitioning to energy weapons alone, according to Hilarides.[50]

Issues for Congress

Potential Oversight Questions

Potential oversight questions for Congress regarding Navy programs for SSLs, EMRG, and HVP include the following:

- Using currently available approaches for countering ASCMs and ASBMs, how well could Navy surface ships defend themselves in a combat scenario against an adversary such as China that has large numbers of ASCMs (including advanced models) and ASBMs? How would this change if Navy surface ships in coming years were equipped with SSLs, EMRG, HVP, or some combination of these systems?
- How significant are the remaining development challenges for SSLs, EMRG, and HVP?
- Are current schedules for developing SSLs, EMRG, and HVP appropriate in relation to remaining development challenges and projected improvements in enemy ASCMs and ASBMs? To what degree are current schedules for developing SSLs, EMRG, or HVP sensitive to annual funding levels?
- When does the Navy anticipate issuing roadmaps detailing its plans for procuring and installing production versions of SSLs, EMRGs, and HVP on specific Navy ships by specific dates?
- Will the kinds of surface ships that the Navy plans to procure in coming years have sufficient space, weight, electrical power, and cooling capability to take full advantage of SSLs (particularly those with beam powers above 200 kW) and EMRG? What changes, if any, would need to be made in Navy plans for procuring large surface combatants (i.e., destroyers and cruisers) or other Navy ships to take full advantage of SSLs and EMRG?

---

[50] Allyson Versprille, "Integration Biggest Challenge for Railgun," *National Defense*, June 2015. See also Lance M. Bacon, "3-Star: 'Lot of Work' Before Railgun Arrives in Fleet," *Navy Times*, February 5, 2015.

- Are the funding sources for SSLs, EMRG, and HVP in Navy and Defense-Wide research and development accounts (see "Summary of Congressional Action on FY2017 Funding" below) sufficiently visible for supporting congressional oversight?

Legislative Activity for FY2017

Summary of Congressional Action on FY2017 Funding

Funding in the defense budget for research and development work on Navy SSLs, EMRG, and HVP is spread across several research and development account line items (which are known as program elements, or PEs). The PEs shown in **Table 1** capture much but not necessarily all of the funding for developing Navy SSLs, EMRG, and HVP. The PEs shown in the table, moreover, include funding for efforts other than Navy SSLs, EMRG, and HVP, so congressional changes from requested amounts might or might not relate to SSLs, EMRG, or HVP.

Table 1. Summary of Congressional Action on FY17 Funding

In millions of dollars, rounded to nearest tenth

Program Element (PE) number, PE name, FY16 budget line number		Authorization			Appropriation		
	Req.	**HASC**	**SASC**	**Conf.**	**HAC**	**SAC**	**Conf.**
0602114N, Power Projection Applied Research, line 4	41.4	41.4	41.4		41.4		
0602750N, Future Naval Capabilities Applied Research, line 13	165.1	165.1	165.1		157.1		
0603114N, Power Projection Advanced Technology, line 16	96.4	106.4	81.4		76.6		
0603673N, Future Naval Capabilities Advanced Technology Development, line 21	249.1	249.1	239.1		252.1		
0603925N, Directed Energy and Electric Weapon System, line 75	32.7	32.7	32.7		32.7		
0604250D8Z, Advanced Innovative Technology, line 95	844.9	804.9	844.9		844.9		

Source: Table prepared by CRS based on Navy FY17 budget submission and committee and conference reports.

Notes: HASC is House Armed Services Committee; **SASC** is Senate Armed Services Committee; **HAC** is House Appropriations Committee; **SAC** is Senate Appropriations Committee; **Conf.** is conference agreement. The PEs shown in the table below capture much but not necessarily all of the funding for work on Navy SSLs, EMRG, and HVP. The PEs shown in the table, moreover, include funding for efforts other than Navy SSLs, EMRG, and HVP.

FY2017 National Defense Authorization Act (H.R. 4909/S. 2943)

House

The House Armed Services Committee, in its report (H.Rept. 114-537 of May 4, 2016) on H.R. 4909, recommended the funding levels shown in the HASC column of **Table 1**. The recommended increase of $10 million for PE 0603114N, Power Projection Advanced Technology (line 16) is for "Program increase for common mount." (Page 498) The recommended reduction of $40 million for PE 0604250D8Z, Advanced Innovative Technology (line 95) is for "SCO" (the Strategic Capabilities Office). (Page 518) H.Rept. 114-537 states:

> *Common mount for electromagnetic railgun*
>
> The budget request contained $96.4 million in PE 63114N for power projection advanced technology. Of this amount, $15.4 million was included for the Navy's electromagnetic railgun prototype.
>
> The committee remains supportive of the Navy's program for developing and deploying an electromagnetic railgun. The committee recognizes the growing imperative for the Navy to field this type of weapon, not only to increase capabilities for naval surface fire support and ballistic missile defense, but to also decrease the cost exchange model when comparing the railgun to conventional missiles or guns. However, the committee is increasingly concerned that the shift in emphasis to the hypervelocity projectile by the Strategic Capabilities Office has left the Navy with a funding gap in developing the requirements and design for a common mount, which is a necessary prerequisite to getting this capability into operational use. Therefore, the committee directs the Secretary of the Navy to provide a briefing to the House Committee on Armed Services by February 15, 2017, on the plan and milestone schedule for demonstrating and deploying a common railgun mount for sea- and land-based applications.
>
> The committee recommends $106.4 million, an increase of $10.0 million, in PE 63114N to support the development of a common mount for the sea-based and land-based electromagnetic railgun. (Page 61)

H.Rept. 114-537 also states:

> *Strategic Capabilities Office*
>
> The budget request contained $844.9 million in PE 64250D8Z for development activities of the Strategic Capabilities Office (SCO).
>
> Created in 2012 by the Deputy Secretary of Defense, SCO has the mission to identify, analyze, demonstrate, and transition game-changing applications of existing and near-term technology to shape and counter emerging threats. SCO is comprised of a relatively small number of personnel and relies on other program office personnel and resources to execute its mission. The committee appreciates the nature of SCO's mission and sustained leanness of the organization; however, the committee notes the budget for SCO has grown exponentially each fiscal year. For example, the fiscal year 2017 budget request is nearly double the request for fiscal year 2016.
>
> The committee is concerned that such rapid budget growth may bring with it some risks, including the demands on SCO's small staff, demands on other Department of Defense personnel, and impact of SCO decisions on existing programs. For example, the committee is aware of SCO's inclusion on the electromagnetic railgun development, and subsequent reprioritizing of its planned investment in that program for fiscal year 2017, resulting in a funding gap that could not be covered by the program office.
>
> Additionally, the committee remains concerned that the transition of technologies from SCO has not been adequately captured and conveyed to the oversight committees. The report required by the committee report (H. Rept. 114–102) accompanying the National Defense Authorization Act for Fiscal Year 2016 has not been delivered and is now almost 6 months late. In order to support prudent use of taxpayer resources, and to ensure proper oversight of these activities, the committee believes this report should be provided and concerns addressed before supporting full funding of planned activities.
>
> Therefore, the committee recommends $804.9 million, a decrease of $40.0 million, in PE 64250D8Z for development activities of the Strategic Capabilities Office. (Page 92)

Section 220 of H.R. 4909 as reported states:

> SEC. 220. Designation of Department of Defense senior official with principal responsibility for directed energy weapons.
>
> Not later than 180 days after the date of the enactment of this Act, the Secretary of Defense shall—
>
> (1) designate a senior official already serving within the Department of Defense as the official with principal responsibility for the development and demonstration of directed energy weapons for the Department; and
>
> (2) set forth the responsibilities of that senior official with respect to such programs.

H.Rept. 114-537 also states:

> *Five-inch precision guided projectile development for naval surface fire support*
>
> In the committee report (H. Rept. 114–102) accompanying the National Defense Authorization Act for Fiscal Year 2016, the committee noted "that current surface Navy gunnery requirements are outdated and that new technologies such as railgun and directed energy weapons are nearing readiness for technology transition." The committee referenced the Advanced Naval Surface Fires (ANSF) initiative and noted the ANSF was assessing options for providing a near-term 5-inch guided munition capability. The committee understands this capability would provide for improved and extended-range naval surface fire support. The committee continues to support the need for this precision guided capability and is also aware of the Hypervelocity Gun Weapon System (HGWS) program that is currently under consideration by the Strategic Capabilities Office (SCO). The committee notes the HGWS program would "flip the cost equation using conventional guns to defend forward bases against raids of advanced cruise and ballistic missiles" and believes there could be applications for use in 5-inch gun systems for naval surface fires support. The committee is encouraged by the development of both of these initiatives and expects the Navy and SCO to coordinate on these capabilities. The committee also expects the Navy to proceed forward with an accelerated development and acquisition strategy for this needed capability that is consistent with acquisition reform principles. (Pages 62-63)

H.Rept. 114-537 also states:

> *Technology enablers for directed energy weapon systems*
>
> The committee is aware that the Department of Defense has made significant advances in the development and operational demonstration of directed energy weapons systems. Each military department has demonstrated a marquee program in this area, such as the Navy's Laser Weapon System deployed on the USS Ponce, the Army High Energy Laser Mobile Demonstrator, and the Marine Corps' Ground Based Air Defense System. Along with technology demonstration activities like the Robust Electric Laser Initiative and the High Energy Liquid Laser Area Defense System, each of these programs demonstrated the increased power output and power on target necessary to develop a militarily useful directed energy weapon.
>
> However, as the Department has made progress in raising the power levels of these systems, it has also demonstrated the need for emphasis on development in other technology areas necessary to realize the full potential of laser weapons. For example, higher power output requires improved beam control to engage targets at greater distances, as well as better thermal management systems to dissipate the increased heat load. As the Department has been overcoming foundational technical challenges, new challenges have emerged that will impact the operational uses for directed energy weapons.

> Therefore, the committee directs the Assistant Secretary of Defense for Research and Engineering, in coordination with the research components of the military departments and the High Energy Laser Joint Technology Office, to provide a briefing to the House Committee on Armed Services by January 20, 2017. This briefing should provide a roadmap for enabling technologies, including:
>
> (1) Beam directors and adaptive optics, including deformable mirrors;
>
> (2) Thermal management needs and capabilities;
>
> (3) Integration challenges with fire control systems, including potential
>
> future needs for fire control for laser systems;
>
> (4) Power architectures and power electronics needs;
>
> (5) Facilities and test range capabilities; and
>
> (6) Other areas as deemed by the Secretary. (Page 93)

Senate

The Senate Armed Services Committee, in its report (S.Rept. 114-255 of May 18, 2016) on S. 2943, recommended the funding levels shown in the SASC column of **Table 1**. The recommended reduction of $15 million to PE 0603114N, Power Projection Advanced Technology (line 16) is for "General decrease." (Page 481) The recommended reduction of $10 million to PE 0603673N, Future Naval Capabilities Advanced Technology Development (line 21) is for "Capable manpower, and power and energy." (Page 481).[51] S.Rept. 114-255 states:

> **Power projection advanced technology**
>
> The budget request included $96.4 million in PE 63114N for power projection advanced technology. The committee notes that the Navy, Air Force, Defense Advanced Research Projects Agency, Strategic Capabilities Office, and other elements within the Department of Defense are all pursuing advanced power projection technologies and systems. The committee is concerned that these efforts are not well-coordinated and have uncertain pathways for transition to programs of record. In addition, the committee notes that the budget request represents an almost 200 percent increase over the amount enacted for fiscal year 2016. The committee believes that such a large increase in budget is not warranted and is concerned about the ability of the programs to absorb the additional funding. Consequently, the committee recommends a decrease of $15.0 million in PE 63114N, but directs that this reduction not be assessed against solid state laser maturation efforts. (Page 52)

Section 216 of S. 2943 as reported states:

---

[51] This recommended reduction of $10 million does not appear to relate (at least not directly) to lasers, the electromagnetic railgun, or the HVP. S.Rept. 114-255 states:

> **Capable manpower and power and energy**
>
> The budget request included $249.1 million in PE 63673N for future naval capabilities advanced technology developments. The activities listed under this program element include capable manpower and power and energy. The committee believes that the work plans for fiscal year 2017 on these activities does not warrant the level of funding included in the budget request. For example, the committee notes that the research included in these two projects include development of new personnel and management methodologies, and capabilities in energy security. Both of these efforts could be better coordinated with other organizations performing similar research. Consequently, the committee recommends a decrease of $10.0 million in PE 63673N to be distributed appropriately from capable manpower and power and energy. (Page 53)

SEC. 216. Directed energy weapon system programs.

(a) Inclusion of Directed Energy Weapon System programs in the rapid acquisition authority program.—

(1) IN GENERAL.—Section 806(c)(1) of the Bob Stump National Defense Authorization Act for Fiscal Year 2003 (Public Law 107–314; 10 U.S.C. 2302 note) is amended by adding at the end the following new subparagraph:

"(D) (i) In the case of any supplies and associated support services that, as determined in writing by the Secretary of Defense without delegation, are urgently needed to eliminate a deficiency in directed energy weapon systems, the Secretary may use the procedures developed under this section in order to accomplish the rapid acquisition and deployment of needed offensive or defensive directed energy weapon systems capabilities, supplies, and associated support services.

"(ii) For the purposes of directed energy weapon systems acquisition, the Secretary of Defense shall consider use of the following procedures:

"(I) The rapid acquisition authority provided under this section.

"(II) Use of other transactions authority provided under section 2371 of title 10, United States Code.

"(III) The acquisition of commercial items using simplified acquisition procedures.

"(IV) The authority for procurement for experimental purposes provided under section 2373 of title 10, United States Code.

"(iii) In this subparagraph, the term 'directed energy weapon systems' means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.".

(2) CONFORMING AMENDMENTS.—Section 2373 of title 10, United States Code, is amended—

(A) in subsection (a), by striking "and aeronautical supplies" and inserting ", aeronautical supplies, and directed energy weapon systems"; and

(B) by adding at the end of the following new subsection:

"(c) Directed energy weapon systems defined.—In this section, the term 'directed energy weapon systems' means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.".

(b) Joint Directed Energy Program Office.—

(1) REDESIGNATION.—The High Energy Laser Joint Technology Office of the Department of Defense is hereby redesignated as the "Joint Directed Energy Program Office" (in this subsection referred to as the "Office").

(2) STRATEGIC PLAN FOR DEVELOPMENT AND FIELDING OF DIRECTED ENERGY WEAPONS CAPABILITIES.—In addition to the functions and duties of the Office in effect on the day before the date of the enactment of this Act, the Office shall develop a strategic plan for development and fielding of directed energy weapons capabilities for the Department, in which the Office may define requirements for directed energy capabilities that address the highest priority warfighting capability gaps of the Department.

(3) ACCELERATION OF DEVELOPMENT AND FIELDING OF DIRECTED ENERGY WEAPONS CAPABILITIES.—

(A) IN GENERAL.—To the degree practicable, the Office shall use the policies of the Department that are revised pursuant to this section and new acquisition and management

> practices established pursuant to this section to accelerate the development and fielding of directed energy capabilities.
>
> (B) ENGAGEMENT.—The Secretary shall ensure that use of policies and practices described in subparagraph (A) include engagement with defense and private industries, research universities, and unaffiliated, nonprofit research institutions.

Regarding Section 216, S.Rept. 114-255 states:

> **Directed energy weapon system programs (sec. 216)**
>
> The committee remains concerned about the Department of Defense's inability to field an operational directed energy system. The committee is aware that the military services and industry partners have developed sufficient directed energy weapon capabilities for specific scenarios—like the High Energy Laser Mobile Demonstrator (HEL–MD) to counter rocket, artillery and mortar for base protection purposes and the Counter Electronics High Powered Microwave Advanced Missile Project (CHAMP) for disabling an adversary's electronics while avoiding collateral damage. These programs, as well as other high energy laser weapon systems, have been tested and demonstrated, but have failed to transition to acquisition programs of record.
>
> The committee notes that directed energy capabilities have the potential to support many operational missions in cost effective and efficient manners. In response to these factors, the committee recommends a provision that would amend section 806 of the Bob Stump National Defense Authorization Act for Fiscal Year 2003 (Public Law 107–314) to grant rapid acquisition authorities for directed energy weapon systems to accelerate the development and fielding of this technology and to help offset the gains of potential adversaries.
>
> The committee notes that since 1960, the Department of Defense has invested more than $6.0 billion in directed energy science and technology initiatives. However, the committee remains concerned that, despite this significant investment, the Department's directed energy initiatives are not resourced at levels necessary to transition them to full-scale acquisition programs. The committee notes with concern that years of investment have not to date resulted in any operational systems with high energy laser capability.
>
> The committee highlights that the Defense Science Board Task Force on Directed Energy Weapon Systems and Technology Applications found that "directed energy offers promise as a transformational 'game changer' in military operations, able to augment and improve operational capabilities in many areas." The task force further concluded that the range of potential applications is sufficient to warrant significantly increased attention to the scope and direction of efforts to assess, develop, and field appropriate laser, microwave, and millimeter wave weapons. Consistent with the findings of the task force, the committee believes that directed energy weapons systems offer significant benefits in terms of cost effectiveness, sustainability, magazine capabilities, and precision targeting. (Pages 46-47)

S.Rept. 114-255 also states:

> **High energy laser joint technology office**
>
> The budget request included $42.3 million in PE 62890F for high energy laser research. The committee notes that this program element funds defense high energy laser applied research through the High Energy Laser Joint Technology Office. However, the committee is concerned that the Joint Technology Office has not received sufficient funding in recent years to drive the maturation of high energy laser technology. As an example, the committee notes with concern that no laser technologies have yet been fielded or deployed, despite promising development and field tests. Given the importance of directed energy weapons systems in general as noted elsewhere in this Act, and of high energy laser systems in particular, the committee is concerned that budget request for this

> program element will be insufficient for supporting the joint technology office. Accordingly, the committee recommends an increase of $5.0 million in PE 62890F for the high energy laser joint technology office. (Pages 54-55)

S.Rept. 114-255 also states:

> **Directed energy systems prototyping**
>
> The budget request included no money in PE 64342D8Z for defense technology offsets. The committee notes with disappointment that the administration did not view it as a priority to request funds through this program element. Particularly with the high-profile emphasis placed on the Department of Defense's Third Offset Strategy, the committee is disappointed to see this program be unfunded. In addition, as noted elsewhere in this report, the committee is deeply disappointed with how the technology offset funding enacted in fiscal year 2016 was allocated. As noted, none of the money was put towards directed energy, in contradiction to the clear intent of Congress that half of the money be used to bolster directed energy technologies. While the committee does not recommend additional unrestricted funds for the technology offsets program, the committee underscores that directed energy systems are still critical areas of work in need of greater support and attention. The committee believes that the Department needs to focus in particular on the transition from lab development to deployment and fielding. Consequently, the committee recommends a general increase of $25.0 million in PE 64342D8Z to be used only for the purposes of directed energy systems prototyping. (Pages 59-60)

S.Rept. 114-255 also states:

> **Laser weapon system demonstrator**
>
> The Committee commends the Navy for initiating and funding the Laser Weapon Systems Demonstrator (LWSD) and believes that this is an important step toward maturing technologies that could ultimately enable the deployment of a shipboard maritime laser weapons system. While the Committee understands that the Navy envisions transitioning laser weapons to a formal Program of Record in the 2020s, it appears that the Navy has not programmed funding beyond the LWSD sea-based tests to support the installation of LWSD on a DDG or for the design and procurement of a formal maritime laser program.
>
> The committee expects that the Secretary of the Navy will keep the congressional defense committees updated on its plan to seamlessly transition the LWSD to a shipboard weapons system following sea-based testing and to a formal maritime laser Program of Record, technical progress toward developing the capability, and programmatic steps being taken to move to demonstration and deployment of advanced laser systems. (Page 70)

FY2017 DOD Appropriations Act (H.R. 5293)

House

The House Appropriations Committee, in its report (H.Rept. 114-577 of May 19, 2016) on H.R. 5293, recommended the funding levels shown in the HAC column of **Table 1**. The recommended reduction of $8 million to PE 0602750N, Future Naval Capabilities Applied Research (line 13) is for "FORCENET excessive growth" ($5 million), "Power and energy previously funded efforts" ($2 million), and "Sea shield previously funded efforts" ($1 million). (Page 228) The recommended reduction of $19.8 million to PE 0603114N, Power Projection Advanced Technology (line 16) is for "Precision strike technology excess growth." (Page 228) The recommended net reduction of $3 million to PE 0603673N, Future Naval Capabilities Advanced

Technology Development (line 21) consists of a reduction of $2 million for “Power and energy previously funded efforts,” a reduction of $2 million for “Sea shield previously funded efforts,” and an increase of $7 million for “Program increase—ASW [antisubmarine warfare] research.” (Page 228)

H.Rept. 114-577 states:

> HIGH ENERGY LASERS
>
> The Committee is aware of efforts within the High Energy Laser Joint Technology Office to develop advanced, directed-energy, high energy laser weapons that have the potential to perform a wide variety of military missions. The Committee encourages the Secretary of Defense to explore further development and evaluation of this important technology. (Page 264)

Directed Energy Weapon Systems Acquisition Act of 2016 (H.R. 4964/S. 2778)

House

H.R. 4964 was introduced in the House on May 5, 2016. The text of H.R. 4964 as introduced states:

> SECTION 1. Short title.
>
> This Act may be cited as the “Directed Energy Weapon Systems Acquisition Act of 2016”.
>
> SEC. 2. Findings.
>
> Congress makes the following findings:
>
> (1) The Committee on Armed Services of the Senate noted in the report accompanying S. 1356 (S. Rept. 114–49; 114th Congress) that since 1960, the Department of Defense has invested more than $6,000,000,000 in directed energy science and technology initiatives, and that the Committee is concerned that, despite this significant investment, the Department's directed energy initiatives are not resourced at levels necessary to transition them to full-scale acquisition programs.
>
> (2) The Defense Science Board Task Force on Directed Energy Weapon Systems and Technology Applications (the “Task Force”) found that “directed energy offers promise as a transformational ‘game changer’ in military operations, able to augment and improve operational capabilities in many areas”.
>
> (3) Despite this potential, years of investment have not resulted in any operational systems with high energy laser capability.
>
> (4) The Task Force believes that the range of potential application is sufficient to warrant significantly increased attention to the scope and direction of efforts to assess, develop, and field appropriate laser, microwave, and millimeter wave weapons.
>
> SEC. 3. Inclusion of directed energy weapon system programs in the rapid acquisition authority program.
>
> (a) In general.—Section 806(c)(1) of the Bob Stump National Defense Authorization Act for Fiscal Year 2003 (Public Law 107–314; 10 U.S.C. 2302 note) is amended by adding at the end the following new subparagraph:
>
> “(D) (i) In the case of any supplies and associated support services that, as determined in writing by the Secretary of Defense without delegation, are urgently needed to eliminate

> a deficiency in directed energy weapon systems, the Secretary may use the procedures developed under this section in order to accomplish the rapid acquisition and deployment of needed offensive or defensive directed energy weapon systems capabilities, supplies, and associated support services.
>
> "(ii) For the purposes of directed energy weapon systems acquisition, the Secretary of Defense shall consider use of the following procedures:
>
> "(I) The rapid acquisition authority provided under this section.
>
> "(II) Use of other transactions authority provided under section 2371 of title 10, United States Code.
>
> "(III) The acquisition of commercial items using simplified acquisition procedures.
>
> "(IV) The authority for procurement for experimental purposes provided under section 2373 of title 10, United States Code.
>
> "(iii) In this subparagraph, the term 'directed energy weapon system' means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.".
>
> (b) Conforming amendments.—Section 2373 of title 10, United States Code, is amended—
>
> (1) in subsection (a), by striking "and aeronautical supplies" and inserting ", aeronautical supplies, and directed energy weapon systems"; and
>
> (2) by adding at the end of the following new subsection:
>
> "(c) Directed energy weapon system defined.—In this section, the term 'directed energy weapon system' means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.".
>
> SEC. 4. Joint Directed Energy Program Office.
>
> (a) Redesignation.—The High Energy Laser Joint Technology Office of the Department of Defense is hereby redesignated as the "Joint Directed Energy Program Office" (in this section referred to as the "Office").
>
> (b) Strategic plan for development and transition of directed energy weapons capabilities toward fielding.—In addition to the functions and duties of the Office in effect on the day before the date of the enactment of this Act, the Office shall develop a strategic plan for development and transition of directed energy weapons capabilities toward fielding for the Department, in which the Office may define requirements for directed energy capabilities that address the highest priority warfighting capability gaps of the Department.
>
> (c) Acceleration of development and transition of directed energy weapons capabilities toward fielding.—
>
> (1) IN GENERAL.—To the degree practicable, the Office shall use the policies of the Department that are revised pursuant to this Act and new acquisition and management practices established pursuant to this Act to accelerate the development and transition of directed energy capabilities toward fielding.
>
> (2) ENGAGEMENT.—The Secretary shall ensure that use of policies and practices described in paragraph (1) include engagement with defense and private industries, research universities, and unaffiliated, nonprofit research institutions.

Senate

S. 2778 was introduced in the Senate on April 12, 2016. The text of S. 2778 as introduced states:

SECTION 1. Short title.

This Act may be cited as the “Directed Energy Weapon Systems Acquisition Act of 2016”.

SEC. 2. Findings.

Congress makes the following findings:

(1) The Committee on Armed Services of the Senate noted in the report accompanying S. 1356 (S. Rept. 114–49; 114th Congress) that since 1960, the Department of Defense has invested more than $6,000,000,000 in directed energy science and technology initiatives, and that the Committee is concerned that, despite this significant investment, the Department's directed energy initiatives are not resourced at levels necessary to transition them to full-scale acquisition programs.

(2) The Defense Science Board Task Force on Directed Energy Weapon Systems and Technology Applications (the “Task Force”) found that “directed energy offers promise as a transformational ‘game changer’ in military operations, able to augment and improve operational capabilities in many areas”.

(3) Despite this potential, years of investment have not resulted in any operational systems with high energy laser capability.

(4) The Task Force believes that the range of potential application is sufficient to warrant significantly increased attention to the scope and direction of efforts to assess, develop, and field appropriate laser, microwave, and millimeter wave weapons.

SEC. 3. Inclusion of directed energy weapon system programs in the rapid acquisition authority program.

(a) In general.—Section 806(c)(1) of the Bob Stump National Defense Authorization Act for Fiscal Year 2003 (Public Law 107–314; 10 U.S.C. 2302 note) is amended by adding at the end the following new subparagraph:

“(D) (i) In the case of any supplies and associated support services that, as determined in writing by the Secretary of Defense without delegation, are urgently needed to eliminate a deficiency in directed energy weapon systems, the Secretary may use the procedures developed under this section in order to accomplish the rapid acquisition and deployment of needed offensive or defensive directed energy weapon systems capabilities, supplies, and associated support services.

“(ii) For the purposes of directed energy weapon systems acquisition, the Secretary of Defense shall consider use of the following procedures:

“(I) The rapid acquisition authority provided under this section.

“(II) Use of other transactions authority provided under section 2371 of title 10, United States Code.

“(III) The acquisition of commercial items using simplified acquisition procedures.

“(IV) The authority for procurement for experimental purposes provided under section 2373 of title 10, United States Code.

“(iii) In this subparagraph, the term ‘directed energy weapon system’ means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.”.

(b) Conforming amendments.—Section 2373 of title 10, United States Code, is amended—

(1) in subsection (a), by striking “and aeronautical supplies” and inserting “, aeronautical supplies, and directed energy weapon systems”; and

(2) by adding at the end the following new subsection:

"(c) Directed energy weapon system defined.—In this section, the term 'directed energy weapon system' means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.".

SEC. 4. Joint Directed Energy Program Office.

(a) Redesignation.—The High Energy Laser Joint Technology Office of the Department of Defense is hereby redesignated as the "Joint Directed Energy Program Office" (in this section referred to as the "Office").

(b) Strategic plan for development and transition of directed energy weapons capabilities toward fielding.—In addition to the functions and duties of the Office in effect on the day before the date of the enactment of this Act, the Office shall develop a strategic plan for development and transition of directed energy weapons capabilities toward fielding for the Department, in which the Office may define requirements for directed energy capabilities that address the highest priority warfighting capability gaps of the Department.

(c) Acceleration of development and transition of directed energy weapons capabilities toward fielding.—

(1) IN GENERAL.—To the degree practicable, the Office shall use the policies of the Department that are revised pursuant to this Act and new acquisition and management practices established pursuant to this Act to accelerate the development and transition of directed energy capabilities toward fielding.

(2) ENGAGEMENT.—The Secretary shall ensure that use of policies and practices described in paragraph (1) include engagement with defense and private industries, research universities, and unaffiliated, nonprofit research institutions.

Appendix. Potential Advantages and Limitations of Shipboard Lasers

This appendix presents additional information on potential advantages and limitations of shipboard lasers.

Potential Advantages

In addition to a low marginal cost per shot and deep magazine, potential advantages of shipboard lasers include the following:

- **Fast engagement times.** Light from a laser beam can reach a target almost instantly (eliminating the need to calculate an intercept course, as there is with interceptor missiles) and, by remaining focused on a particular spot on the target, cause disabling damage to the target within seconds. After disabling one target, a laser can be redirected in several seconds to another target.
- **Ability to counter radically maneuvering missiles.** Lasers can follow and maintain their beam on radically maneuvering missiles that might stress the maneuvering capabilities of Navy SAMs.
- **Precision engagements.** Lasers are precision-engagement weapons—the light spot from a laser, which might be several inches in diameter, affects what it hits, while generally not affecting (at least not directly) separate nearby objects.
- **Graduated responses.** Lasers can perform functions other than destroying targets, including detecting and monitoring targets and producing nonlethal effects, including reversible jamming of electro-optic (EO) sensors. Lasers offer the potential for graduated responses that range from warning targets to reversibly jamming their systems, to causing limited but not disabling damage (as a further warning), and then finally causing disabling damage.

Potential Limitations

Potential limitations of shipboard lasers include the following:

- **Line of sight.** Since laser light tends to fly through the atmosphere on an essentially straight path, shipboard lasers would be limited to line-of-sight engagements, and consequently could not counter over-the-horizon targets or targets that are obscured by intervening objects. This limits in particular potential engagement ranges against small boats, which can be obscured by higher waves, or low-flying targets. Even so, lasers can rapidly reacquire boats obscured by periodic swells.
- **Atmospheric absorption, scattering, and turbulence.** Substances in the atmosphere—particularly water vapor, but also things such as sand, dust, salt particles, smoke, and other air pollution—absorb and scatter light from a shipboard laser, and atmospheric turbulence can defocus a laser beam. These effects can reduce the effective range of a laser. Absorption by water vapor is a particular consideration for shipboard lasers because marine environments feature substantial amounts of water vapor in the air. There are certain wavelengths of light (i.e., “sweet spots” in the electromagnetic spectrum) where atmospheric absorption by water vapor is markedly reduced. Lasers can be

designed to emit light at or near those sweet spots, so as to maximize their potential effectiveness. Absorption generally grows with distance to target, making it in general less of a potential problem for short-range operations than for longer-range operations. Adaptive optics, which make rapid, fine adjustments to a laser beam on a continuous basis in response to observed turbulence, can counteract the effects of atmospheric turbulence. Even so, lasers might not work well, or at all, in rain or fog, preventing lasers from being an all-weather solution.

- **Thermal blooming.** A laser that continues firing in the same exact direction for a certain amount of time can heat up the air it is passing through, which in turn can defocus the laser beam, reducing its ability to disable the intended target. This effect, called thermal blooming, can make lasers less effective for countering targets that are coming straight at the ship, on a constant bearing (i.e., "down-the-throat" shots). Other ship self-defense systems, such as interceptor missiles or a CIWS, might be more suitable for countering such targets. Most tests of laser systems have been against crossing targets rather than "down-the-throat" shots. In general, thermal blooming becomes more of a concern as the power of the laser beam increases.
- **Saturation attacks.** Since a laser can attack only one target at a time, requires several seconds to disable it, and several more seconds to be redirected to the next target, a laser can disable only so many targets within a given period of time. This places an upper limit on the ability of an individual laser to deal with saturation attacks—attacks by multiple weapons that approach the ship simultaneously or within a few seconds of one another. This limitation can be mitigated by installing more than one laser on the ship, similar to how the Navy installs multiple CIWS systems on certain ships.
- **Hardened targets and countermeasures.** Less-powerful lasers—that is, lasers with beam powers measured in kilowatts (kW) rather than megawatts (MW)—can have less effectiveness against targets that incorporate shielding, ablative material, or highly reflective surfaces, or that rotate rapidly (so that the laser spot does not remain continuously on a single location on the target's surface) or tumble. Small boats (or other units) could employ smoke or other obscurants to reduce their susceptibility to laser attack.[52] Measures such as these, however, can increase the cost and/or weight of a weapon, and obscurants could make it more difficult for small boat operators to see what is around them, reducing their ability to use their boats effectively.
- **Risk of collateral damage to aircraft, satellites, and human eyesight.** Since light from an upward-pointing laser that does not hit the target would continue flying upward in a straight line, it could pose a risk of causing unwanted collateral damage to aircraft and satellites. The light emitted by SSLs being developed by the Navy is of a frequency that can cause permanent damage to human eyesight, including blinding. Blinding can occur at ranges much greater than ranges for damaging targeted objects. Scattering of laser light off the target or off fog or particulates in the air can pose a risk to exposed eyes.[53]

---

[52] See, for example, "Kelsey D. Atherton, "China Plans To Defeat American Lasers With Smoke," *Popular Science*, May 3, 2016.

[53] The United States in 1995 ratified the 1980 Convention on Prohibitions or Restriction on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects. An (continued...)

For additional background information on potential Navy shipboard SSLs, see CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke.

Author Contact Information

Ronald O'Rourke
Specialist in Naval Affairs
rorourke@crs.loc.gov, 7-7610

(...continued)

international review of the convention began in 1994 and concluded in May 1996 with the adoption of, among other things, a new Protocol IV on blinding laser weapons. The protocol prohibits the employment of lasers that are specifically designed to cause permanent blindness to the naked eye or to the eye with corrective eyesight devices. The United States ratified Protocol IV on December 23, 2008, and it entered into force for the United States on July 21, 2009. DOD views the protocol as fully consistent with DOD policy. DOD believes the lasers discussed in this report are consistent with DOD policy of prohibiting the use of lasers specifically designed to cause permanent blindness to the naked eye or to the eye with corrective eyesight devices. For further discussion, see Appendix I ("Protocol on Blinding Lasers") in CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke.

Torpedo & Directed Energy Weapons Indian Navy Perspective

RAdm Om Prakash Singh Rana, VSM,
Director General of Naval Armament Inspection

IN supporting Indigenisation

IN Indigenisation Plan 2015-30

Maritime Cap Perspective Plan

Self-Reliance

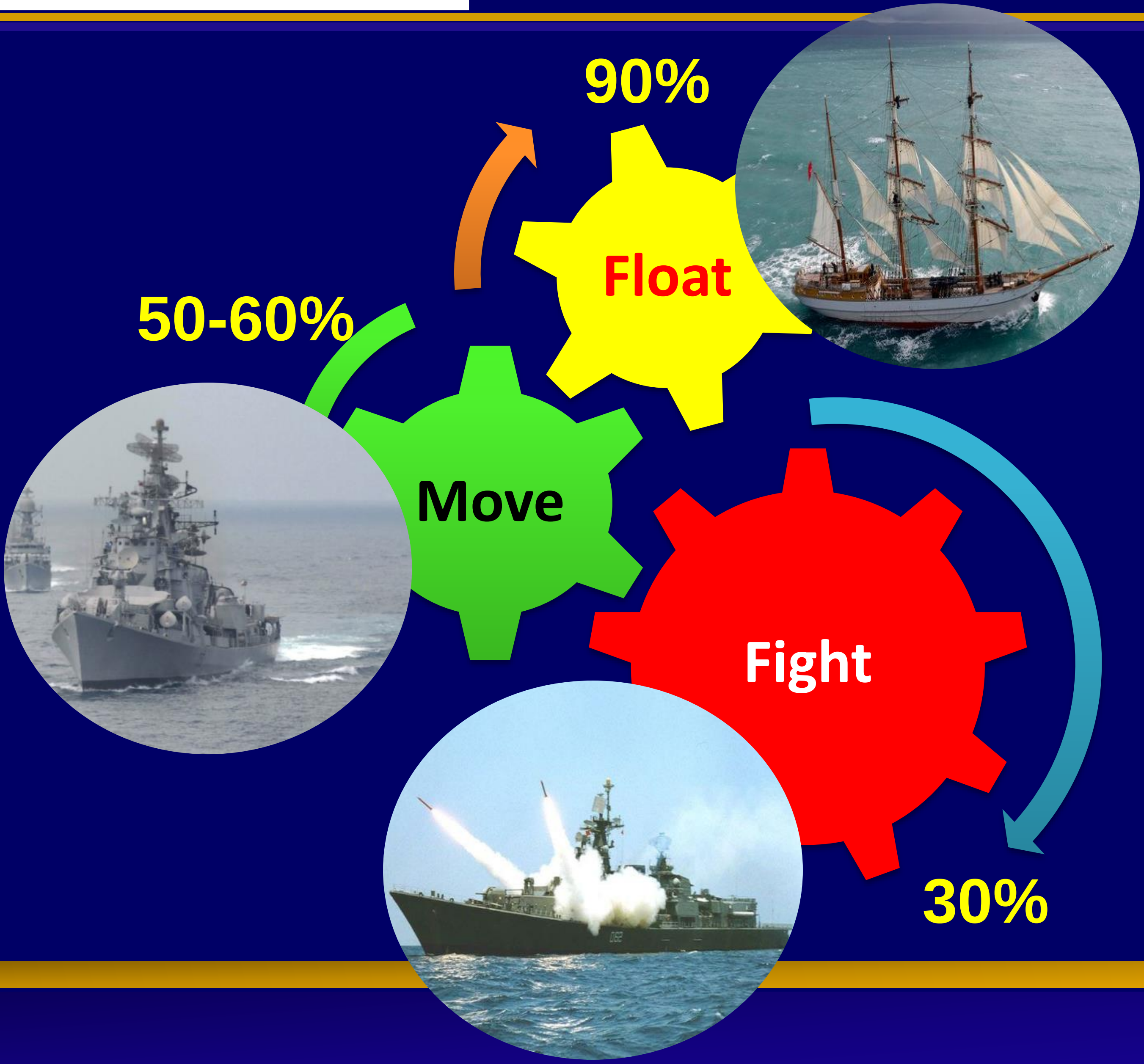

90%

Float

50-60%

Move

Fight

30%

Scope

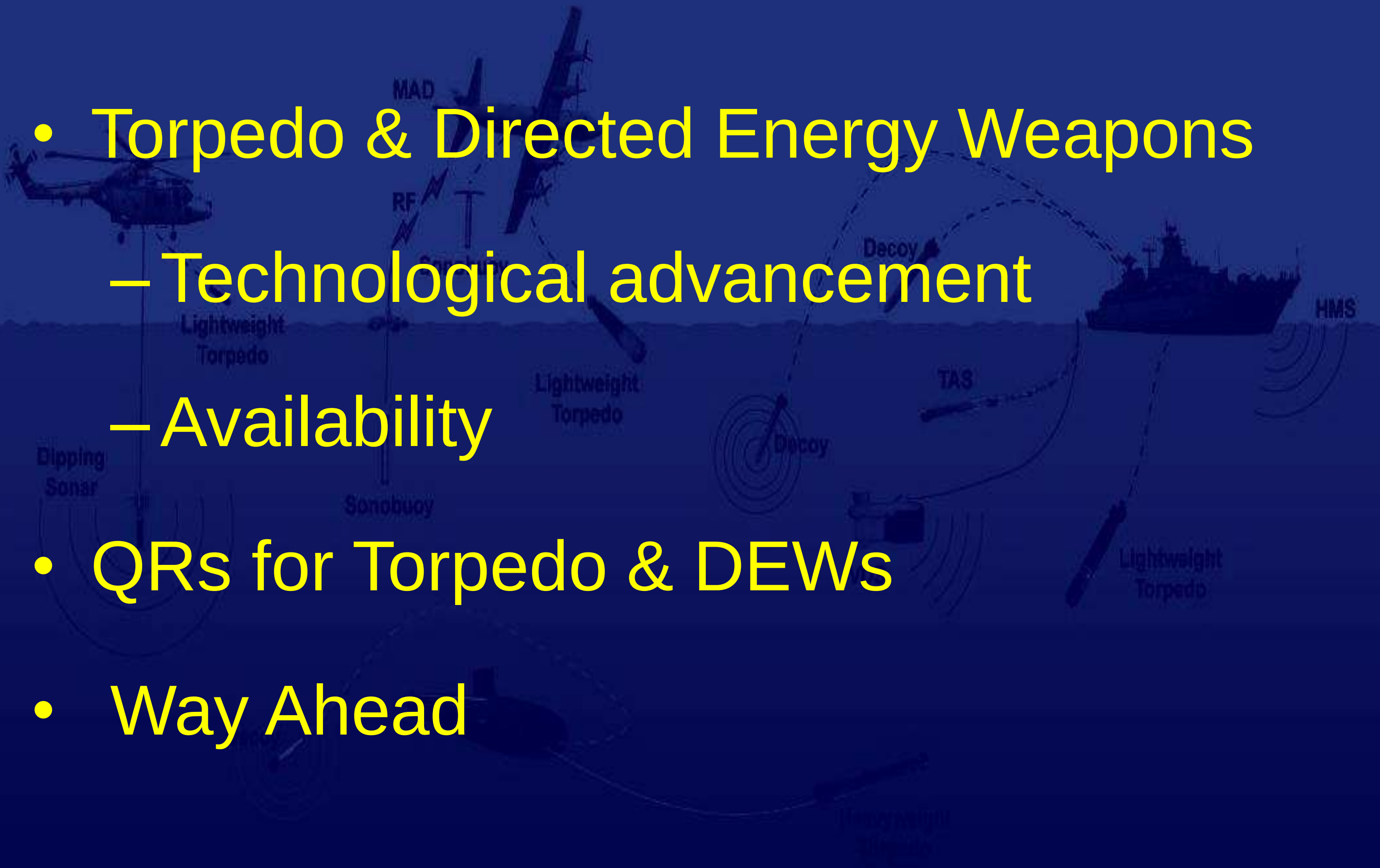
MAD
RF
Decoy
HMS
Lightweight Torpedo
Lightweight Torpedo
TAS
Decoy
Dipping Sonar
Sonobuoy
Lightweight Torpedo
Heavyweight Torpedo

- Torpedo & Directed Energy Weapons
  - Technological advancement
  - Availability
- QRs for Torpedo & DEWs
- Way Ahead

Advancement
in
Torpedo Technology

Advancement in Torpedo Technology Contd...

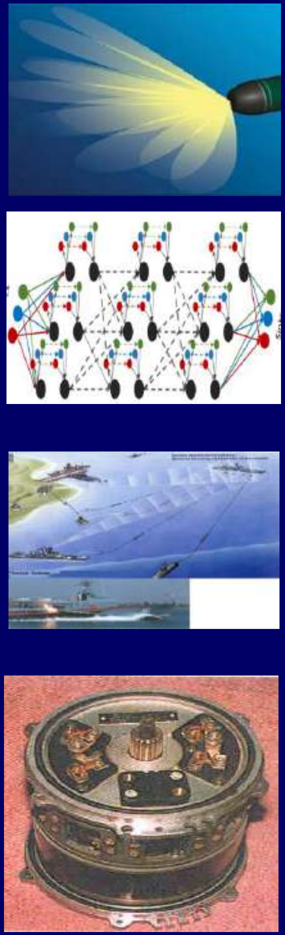

- Broadband sonar, multi-beam & INS
- Multi target tracking capability
- S/W based & high speed sig processing
- Re-programmable S/W & search pattern
- ACCM features
- Digital servo and retractable controls

Advancement in Torpedo Technology Contd...

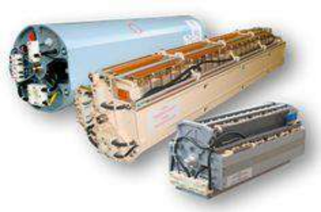

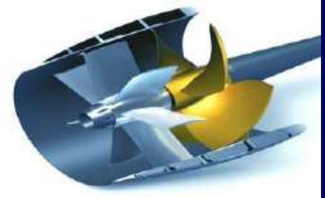

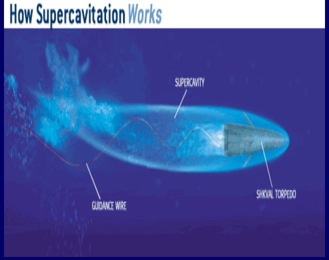
How Supercavitation Works
SUPERCAVITY
GUIDANCE WIRE
SHKVAL TORPEDO

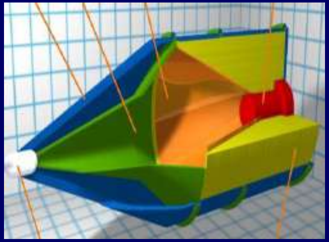

- Al-Ag_2O, Li-ion & Cu-Ni batteries
- Brushless motor & pump jet propulsion
- Semi open/close loop thermal propulsion
- Super-cavitation propulsion
- Tandem shaped charge insensitive W/H

Advance Torpedoes

Torpedo	Origin	Speed (Kts)	Range (Kms)
Black Shark Adv	Italy	18- 50	>50
DM2 A4 Sea Hake	Ger	50	> 50
A244S MOD-3	Italy	36	13.5
Flash Black	Italy	50	20
F21 HWT	France	25-50	> 50
Spearfish HWT	UK	40	48
Torpedo (2000)	Swed	40	> 40
MU-90	Eutorp	50	23
MK 48 ADCAP	USA	55	38
MK 54	USA	28 -43	12
Shkval E	Russia	> 200	7 - 10
UGST	Russia	50	40

Torpedoes in Indian Navy

Torpedo		Origin/ Country
A244S		Italy
AT1ME		Russian
MK 54		USA
TAL MK-I		India
53-65 KE	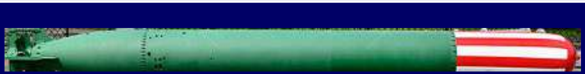	Russian
CET 65 E		Russian
TEST 71ME		Russian
SUT		Germany
Varunastra		India

Directed Energy Weapons

Directed Energy Weapons

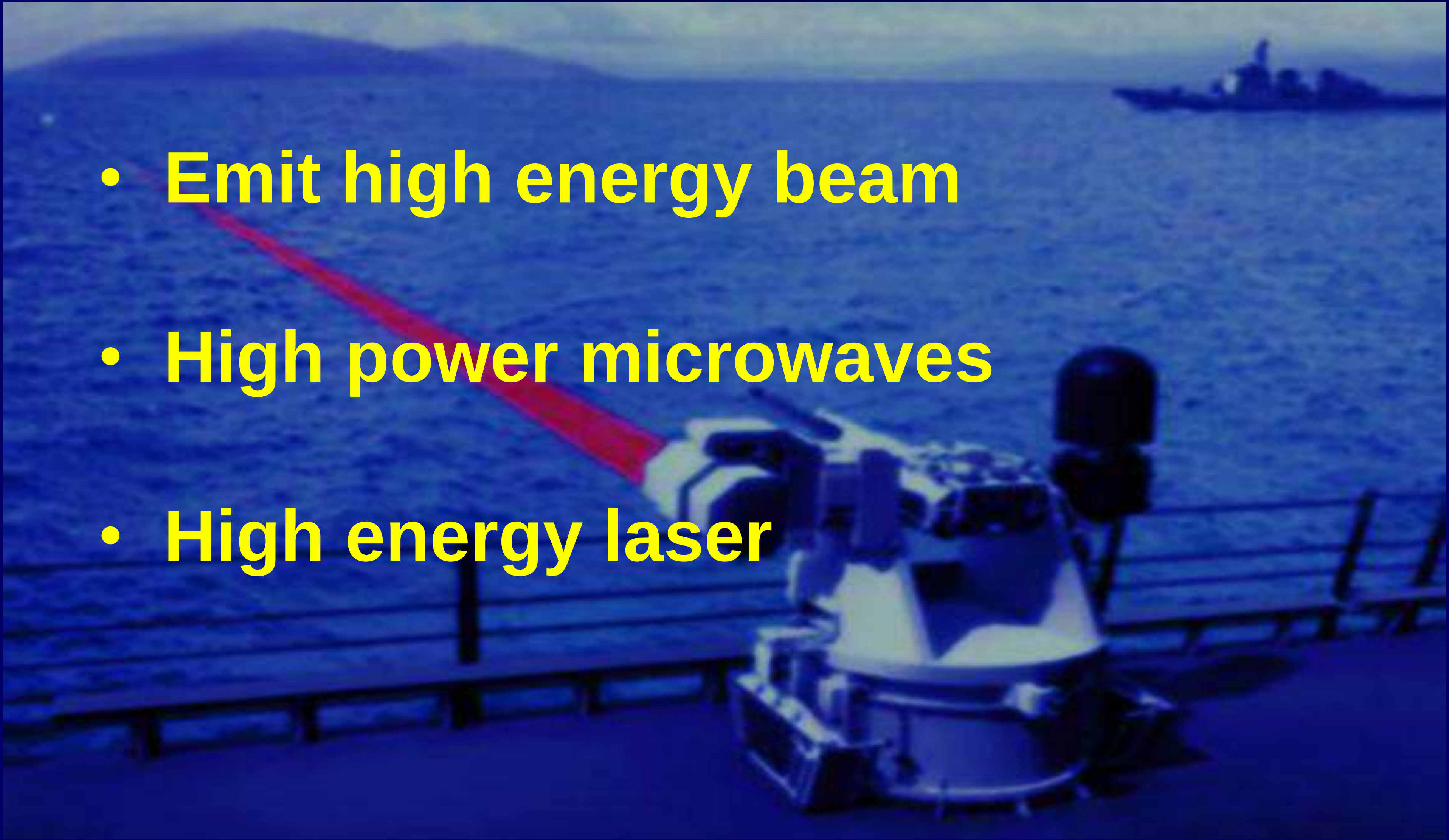

- Emit high energy beam
- High power microwaves
- High energy laser

DEW Advantages

- High Speed of Engagement
- Effective for Defensive/Offensive Role
- Immunity from Gravitation Constraints
- Space Requirement – Limited
- Effective against Moving Air Targets
- Minimal Collateral Damage
- Low Marginal Cost per Shot

High Energy Laser (HEL)
on Rocket from 1.6km away

DEWs from Sea, Air & Land Platforms

HE Laser Mobile Demonstrator
(USA)

Electromag Rail Gun

(USA)

Active Denial System
(USA/ Israel)

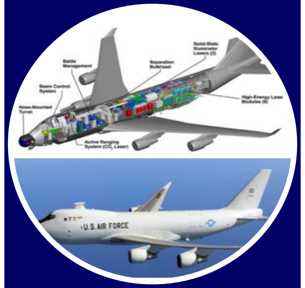

Airborne Ballistic Laser
(USA)

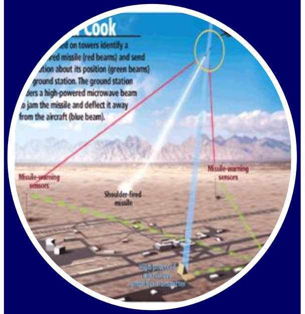

Vigilant Eagle
(USA)

Laser Weapon System

(USA)

DEWs in Indian Navy

DEWs not as potential destroyer

10 KW HEL demonstrated by DRDO

Work in progress for 25 KW HEL

Develop 100-500 KW HEL

Qualitative Requirements (QRs)

QRs....

QRs for Torpedoes

Requirement	**Remark**
Speed	Upto 55 Knot with variable speed
Range	Upto 40 Kms
Battery	$Mg\text{-}Cu_2Cl_2$, $Al\text{-}Ag_2O$ & Li-ion batteries
Propulsion	Gas turbine, brushless motor & super-cavitation
Warhead	Tandem insensitive munition W/H & prox EBW fuze
Guidance	CBASS & improved signal processing algorithm
Launch I/ P	Tactical S/W algorithm with target data & upgrades
Control	Intelligent digital controllers & actuators
Camouflage	ACCM features

QRs….

QRs for DEWs

- 10-25 KW for non-lethal engagement
- 100-500 KW against:-
  - Missile
  - Rocket
  - Aircraft
  - UAVs

Way Ahead

Way Ahead….

Joint Venture

Requirement	**Scope for Designers & Industry**
Torpedo & Sub-systems	
Mech Parts, W/H & Prop	Al-Mg alloy, FRP & nano materials, **IM**, Mg-Cu, $Al\text{-}Ag_2O$ & Li-ion batteries, brushless motor, GT, pump jet & supercavitation prop
Guidance & Control	Sonar imaging & det, CABSS, PCBs, tgt wake seeking, mag anomaly det & super quantum induction device homing, adv accl & sensors.
Process Control	Adv sig processing, tact S/W algorithm for tgt data, friendly forces positions, sound profile & mid-course upgrade, dig ctrl & actuator
Miscellaneous	Fibre optic cable, gene, invertor, shaft, bearing, gear box, bottle, gyro, pump, valve, fastener, cable, pipe, dynamic seal, loc & rec aids
Directed Energy Weapons	
DEW Systems	Ship borne 25-500 KW using High Power Microwaves or Lasers for non-lethal & lethal engagement

Way Ahead….

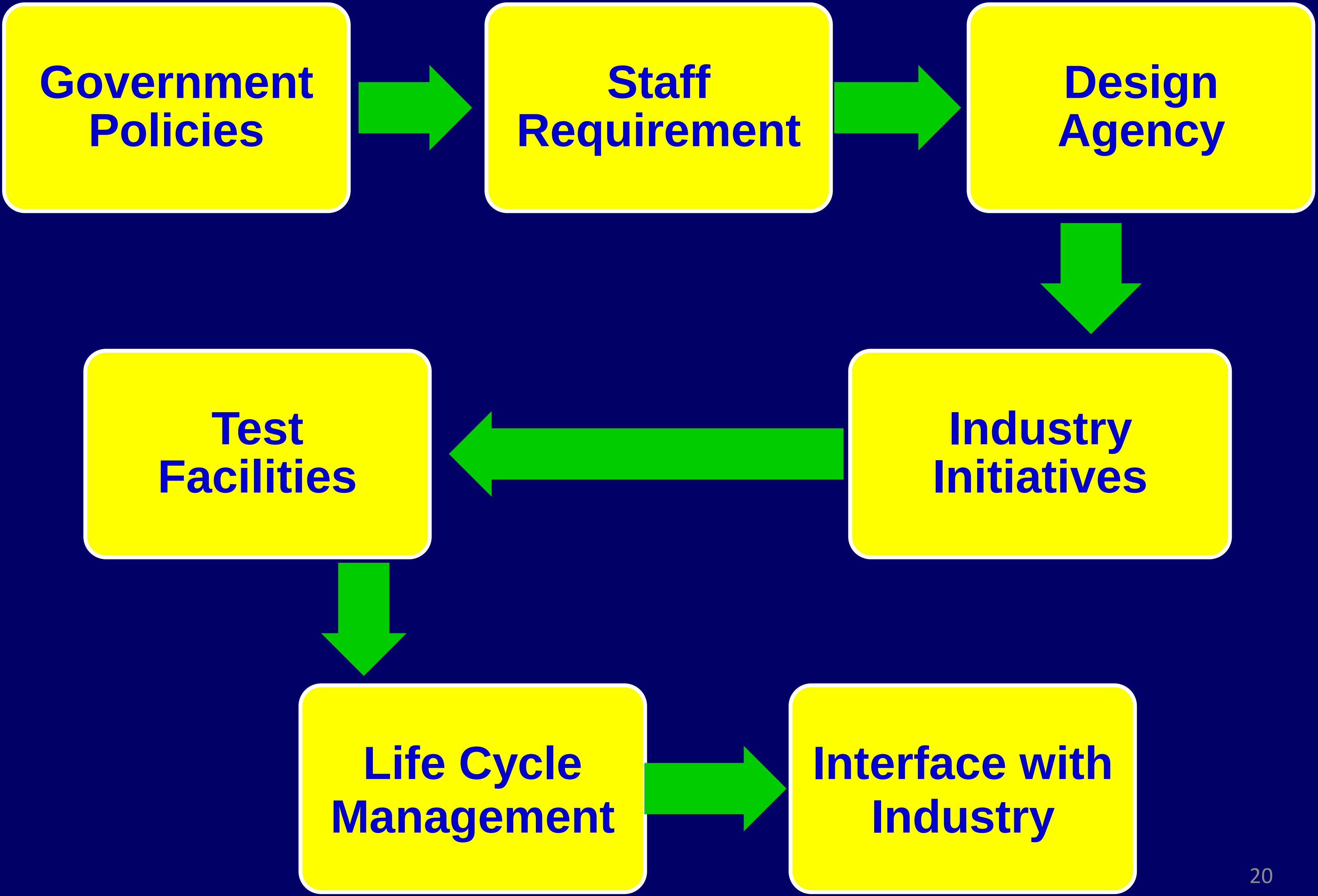

Government Policies

Staff Requirement

Design Agency

Industry Initiatives

Test Facilities

Life Cycle Management

Interface with Industry

Conclusion

- Govt promulgated various major policy guidelines
- ***IN*** "Indigenisation & MCPP"
- Harnessing knowledge base, infrastructures for "**Make in India**" prg under "**FIGHT**" category
- DEWs identified as core mil technology requirement
- Opportunities for strategic partnership

Thank You

5. Psychoenergetics

Mental entrainment, inducing hypnogogic trance, mind control, broadcasting thoughts and feelings, the Ant World, education via EM [LW] mental download, the "Feelies," revolution in spirituality, "enlightenment machine?" Scalar churches?

Psychoenergetics is simply *using the scalar interferometers to manipulate and engineer the human psyche*. To me it is the most mind-boggling and frightening aspect of longitudinal wave engineering.

The mind is electromagnetic in nature, and itself of the time-domain. It is not observable in 3-space. There is nothing you can point to that is the mind. As Bearden says, "The mind is time-like."

The longitudinal scalar EM waves are also of the time-domain. By overlaying certain infolded patterns (oscillations in time) on the longitudinal waves, a thought or feeling may be made to arise in a person or persons who are in the interference zone (where the wave beams cross).

The person will not notice anything, feeling that this thought or idea is his or her own idea. The longitudinal wave pattern might be a wave of panic or fear that spontaneously sweeps over you, perhaps inexplicably. It could even be in the form of a sudden intense patriotism. It could be explicit words and phrases, which everyone would think they had thought of by themselves. Or it could be rage and hatred, angry and violent. On the other hand It could also be a continuous sense of docility and placidity. This is the brave new world of *psychoenergetics and the coming "Mind Wars."*

Scalar Psychoenergetics in its most primitive form simply "entrains" all minds in the target area into a deep hypnogogic trance. In this state of mind people would be suddenly extremely suggestible, and would likely believe anything they are told, and would obey any orders given. Talk about winning the hearts and minds of the American People!

Tom Bearden makes me think of Paul Revere, who heroically conveyed important warnings for the sake of the new nation. But the words of the famous phrase have now strangely and curiously mutated into the once-inconceivable: "One if by land, Two if by mind."

Indeed one strategy in a Mind War would be to simply take over the minds of the "leaders" of the enemy nation. The targeted leader would not be aware that anything was amiss, although he might begin to make unexpected changes in policy.

And are, in fact, any of our "leaders" already having secret thoughts that are not their own? Who knows? Frighteningly, we can no longer be sure.

The new sciences of scalar electromagnetics and psychoenergetics are even now plunging humanity into a sudden science-fiction-like world beyond anything in humanity's previous imaginings.

> "The Russians reached the point in the mid 1990s that they could take over control of a person's mind, with modified longitudinal EM waves including some time-polarized EM waves, and with a team of specialists (estimate 25-30 per transmitter, and one transmitter and team per controlled person)."
> **Bearden: http://www.cheniere.org/misc/time.htm**

In a slide (below image) **Bearden** illustrates how a Scalar War would involve a psychoenergetics attack on the operators of the enemy scalar installation, entraining their minds into hypnogogic trance and getting them to shut down their systems.

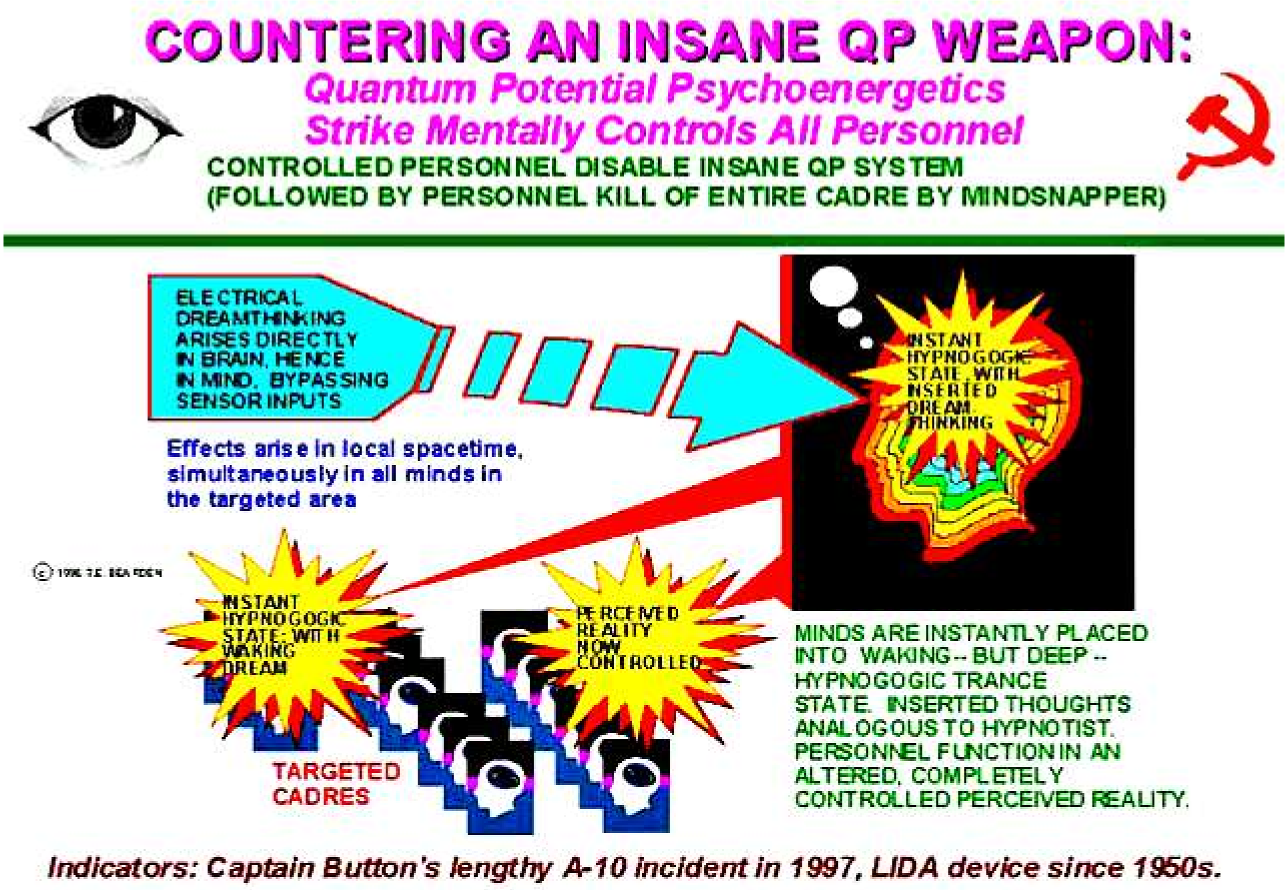
COUNTERING AN INSANE QP WEAPON:
Quantum Potential Psychoenergetics
Strike Mentally Controls All Personnel
CONTROLLED PERSONNEL DISABLE INSANE QP SYSTEM
(FOLLOWED BY PERSONNEL KILL OF ENTIRE CADRE BY MINDSNAPPER)
ELECTRICAL DREAMTHINKING ARISES DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS
Effects arise in local spacetime, simultaneously in all minds in the targeted area
INSTANT HYPNOGOGIC STATE, WITH INSERTED DREAM THINKING
© 1996 T.E. BEARDEN
INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM
PERCEIVED REALITY NOW CONTROLLED
TARGETED CADRES
MINDS ARE INSTANTLY PLACED INTO WAKING-- BUT DEEP -- HYPNOGOGIC TRANCE STATE. INSERTED THOUGHTS ANALOGOUS TO HYPNOTIST. PERSONNEL FUNCTION IN AN ALTERED, COMPLETELY CONTROLLED PERCEIVED REALITY.
Indicators: Captain Button's lengthy A-10 incident in 1997, LIDA device since 1950s.

And ultimately psychoenergetic warfare goes to the very heart of human identity itself.

For if my thoughts might no longer be "mine,"

- Then who and what am I?
- Can my very sense of being "me" be hijacked by some nefarious psychoenergetic scheme?
- Will the "secret government" eventually dictate directly into your mind how you feel about yourself?
- Or what you think you are?
- Or what you should do?

How Does Longitudinal EM Mind Control Work?

The physics of **scalar psychoenergetics** as expounded by **Col. Bearden** seems even more arcane and over-my-head than that of extracting energy from the vacuum. I can't understand it, but I look through it anyway. Each time another tiny understanding might occur, or a term might fall into place. I know the mathematics is forever beyond me.

One paper to begin with is **Mind Control** and **EM Wave Polarization Transductions.** This is such serious stuff that **Bearden** includes a strong warning about misuse of this knowledge.

Psychoenergetics weapons can mentally maim and physically kill.

Warning!

(From "Mind Control and EM Wave Polarization Transductions")

> "This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories.
>
> We do not propose or condone any use of these procedures for non-approved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results.
>
> We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral."
> **Bearden: http://www.cheniere.org/explore articles/mind control1/p01.htm**

After the serious warning comes the description of the paper. This is heady stuff for the layman.

> "Calling full attention to the special note above, in this paper we present a high-level overview of the novel electromagnetic nature of mind operations, mind and body coupling, and intent - the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations.
>
> We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transaction mechanisms whereby differing EM wave polarization s can be transformed one-into-the-other are presented."
> **Bearden: http://www.cheniere.org/explore articles/mind control1/p01.htm**

I will leave it to the technically astute to try to understand the mechanisms of operation which are described in scientific language. But there are tidbits for the layman.

Here he speaks of the difficulty Western scientists have in opening up to the "immaterial" *(longitudinal wave)* nature of the human mind.

Western Science Remains Largely Materialistic

> "Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system. This serious self-limitation exists because in the body we measure only weak TW [transverse wave] EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.
>
> "With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.
>
> "Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW [longitudinal wave] functions exist in the body and in all of nature in general."
> Bearden: http://www.cheniere.org/explore articles/mind control1/p05.htm

Western methods of influencing the mind with EM waves have only used transverse waves, not the longitudinal waves of the vacuum.

If only transverse waves are used one pretty well has to hit the target mind with a sledgehammer of waves.

Western Science knows only transverse EM waves

Brute Forcing Time Functions versus Fine Control Methods

> "Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.
>
> "While KGB scientists also use TW EM "brute force", TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired...
> Bearden
>
> "These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design for example of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gaia's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.
> **Bearden:** http://www.cheniere.org/explore articles/mind control1/p09.htm

The seemingly magical operations of psychoenergetics, sometimes called "psychotronics," cannot be achieved with ordinary transverse EM wave energy. But they are possible with longitudinal scalar waves because the mind itself is scalar in nature.

> "We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain..."
> **Bearden:** http://www.cheniere.org/explore articles/mind control1/p08.htm

The engineering of the mind can only be accomplished with the waves of time.

The course of psychoenergetics development necessitates the building of a vast database of settings and patterns which bring about given mental and/or emotional states.

A kind of "gnome" of the psyche. Bearden describes the process of the decoding of this "gnome."

> "Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.
>
> "In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions."
> **Bearden:** http://www.cheniere.org/explore articles/mind control3/p01.htm
>
> "A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment."
> **Bearden:** http://www.cheniere.org/explore articles/mind control3/p02.htm

Summing up his paper Bearden speculates on what the actual situation the science of psychoenergetics probably is.

The intense secrecy with which this knowledge has been kept from the public has led to the current farce of the "government" pretending it all does not exist, and this secrecy may have now put us at great risk.

> "The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time.
>
> Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.
>
> "In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring.
>
> As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians..."
>
> "Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are to rogue groups often just idle words. They have their own agendas.
>
> And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification."
>
> "The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U (I) electrodynamics.
>
> Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance.
>
> So they will also be able to do the same things here in "mind control" with LW [longitudinal wave] interferometers, through intervening mass."
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p05.htm**

Bearden cites two examples where he believes the Russian KGB tested the psychoenergetics capabilities of their scalar interferometers.

One was the case of Captain Button in his A-10 "Warthog" aircraft on April 2, 1997.

> "Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona."
>
> "At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion dream-thinking until he died that everything was normal.
>
> All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour.
>
> > "Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks."
>
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p06.htm**

The paper Mind Control and EM Wave Polarization Transductions ends with the most chilling and frightening possibility of all: the psychoenergetic engineering of the human species as a whole.

The Ant World

The Final Race for Direct Mind Control of the Entire Human Species

> "The KGB psychoenergetics weapons scientists because of their direct measurements and detections (and use) of t-polarized waves and LWs [longitudinal waves] for decades also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials.
>
> "The KGB scientists also know that something very like Gaia a collective unconscious mind for all species on earth also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials."
>
> "Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.
>
> "I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society" communist style is still a variation of feudalism (all systems tend to feudalism or some form thereof).
>
> "The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now...
>
> "Some versions of Russian mind control devices were used on Russian soldiers in the Afghan War to condition them for performance of merciless acts." 13
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p07.htm**
> **More about "The Ant World"**

Whether it is the Russians planning an "ant world" society or rogue U.S. black-ops groups withholding a great boon from human race, the secrecy on all sides around has allowed largely criminal elements to take over this technology while leaving the civilized elements in the dark.

Openness and the spreading of this knowledge among good and decent people is essential, and for that we will have to overcome our frightened denial.

> "Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.
>
> "Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed."
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p08.htm**

Like the huge energy of the Tesla howitzers, the ability to engineer the mind with longitudinal waveforms is a seeming "magical" thing which can be used for good or ill. Bearden envisions that further developments will allow something like the "downloading" of knowledge via scalar interferometry. Everyone could be highly educated, and quickly and easily.

In the conclusion of the paper "Mind Control and EM Wave Polarization Transductions" Bearden makes a plea to the nations of the world regarding the use of psychoenergetic engineering.

> "Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth.
>
> "Yet it can revolutionize medical science, education, communication, and psychology. We foresee the day - perhaps 30 years hence - when education will be accomplished by directly loading the software into the mind. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks on will also load a doctorate in chemistry. Another three weeks, in electrical engineering.
>
> Another three weeks, an MD And so on. When that happens then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized.
>
> There will be no impoverished large groups lacking the education to find decent, productive jobs.
>
> "We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium."
>
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p08.htm**

(Perhaps after so many frightening topics a slight break is in order here to consider the lighter side of psychoenergetics. For example, what will be the ultimate impact on the entertainment industry?)

The Feelies at Last!

Psychoenergetics will be able to bring about at last, for better or for worse, Aldous Huxley's fanciful idea in his novel "Brave New World," an entertainment format which he called "The Feelies."

Those were something you went to, like going to the movies, but you would experience all the actual feelings themselves, as if the depicted events were actually happening to you.

Although our "leaders" seem too dim to grasp what is going on in the world of scalar electromagnetics, it should not take some young Hollywood genius too long to understand that The Feelies are now possible, and that the immense megabucks of potential profit more than justify heavy investment in bringing it to market.

One can imagine that first "Feelie" theater, "The Huxley" perhaps, a large comfortable area surrounded by nine giant screens (not just one screen!), which become 3D if you put on your headset. And The Huxley is fitted with the finest sound, and oh, by the way, a modest sized scalar interferometer with computer. And there are one or more new tracks now recorded on the film next to the sound track.

These are the digital patterns of thoughts, feelings, and ideas to be fed into the theater's interferometer. And these thoughts and feelings and ideas will be felt by the audience as if they were their own thoughts and ideas. It is an entirely new art form. Psychevision.

As important (and no doubt famous) as the actors would be the "feelers," those from whose minds the feelings were lifted in the original programming of the Feelie tracks. Because, for example, Brad Pitt might be a good and handsome actor, but can he really feel it? Like the dubbing of foreign language films, it might be necessary to dub the "emotional track" using a deeper and more profound person's feelings, thoughts and ideas.

It might even be engineered so that if you sat on one side of the theater you would feel like the hero, and on the other side you would feel exactly like the bad guy! What would that lead to?

We have all cried at a movie some time, most people have, though they might not want to admit it. But this would be deep sobbing and a flood of tears, if the director so desired it. You would feel it at your core. Or, take a big chase scene, you would be terrified. They are breathing down your neck! You would be gripped by panic.

For that matter, the whole genre of Horror-Feelie would be too much for many people to take. Do I really want to feel that an actual vampire is tearing at my throat with his fangs?

A Religious Feelie would make you feel the exaltation of a saint at his prayers, or give you the definite feeling that you are Christ suffering on the cross. All the pain could be there too. The stickiness of the blood. The very feeling itself: "Why hast thou forsaken me?" Who among us could come up that despair for scanning?

Had this been developed a bit earlier we might watch a documentary of, say, Mother Theresa, containing her own actual deep sense of compassion, which we would feel arising in us as our own.

Then again, a Psychedelic Feelie could induce an LSD-like state and simply blow the audience’s minds.

Even in its simplest most primitive mode, the hypnogogic trance induction with its increased suggestibility, would completely alter the experience of even ordinary current movie going.

Go Back to Scalar War

Go Back to Psychoenergetics

<<< Scalar Wars

The Brave New World of Scalar Electromagnetics

by Bill Morgan

(See also **NEW 2007**: Book Review of Bearden's "Oblivion - America at the Brink")

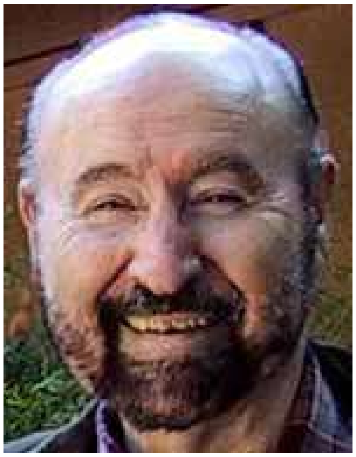
Tom Beardon (2001)
Lieutenant Colonel, U.S. Army (Retired)

For the past six months I have been undergoing the greatest **paradigm shift** I have ever had to go through. It has rattled my nerves and shaken my bones. This intense adjustment of my "world"has come about by studying the information given by Col. Tom Bearden at his website **Cheniere**. The new knowledge there has necessitated a total revision of my ideas about physical reality, the world we live in, and the future of humanity.

This paradigm shifting even actually made me dizzy on certain days as I tried to absorb and digest Bearden's vast amount of information. I am not a scientist at all, just a layman, and I have little comprehension of the math and high physics of this new science called "Scalar Electromagnetics."But there is a great deal of information at **Cheniere** which **needs to become common knowledge** as fast as possible, for the sake of the survival of life on earth.

To that end I have put together this small primer of Bearden's ideas, as a kind of "beginner's guide" to his website. (I have tried to underline certain words and phrases which are part of the new "jargon" of talking about these "new" phenomena. Also I have taken the liberty of **bolding** certain words and phrases to help them stick in the mind).

(Note: Throughout "**EM**" means "**electromagnetic,**" and "**LW**" means "**longitudinal wave**")
You can download this article in Word format

This article has six sections:

1. New Waves 2. Tapping the Waves 3. Weaponization 4. Healing 5. Psychoenergetics 6. As It Stands

2006 NOTE: Click here for a current status history of the MEG from Tom Bearden

1. New waves discovered

Longitudinal EM energy fills vacuum of space, the time domain of spacetime, time as compressed energy, E=tc2, waves of time, phase conjugate wave pairs.

> **"Electric power is everywhere present in unlimited quantities and can drive the world's machinery without the need of coal, oil, gas, or any other of the common fuels."** Nikola Tesla
>
> **"At any point and at any time, one can freely and inexpensively extract enormous EM energy flows directly from the active vacuum itself."** Tom Bearden

I guess the first thing to try to comprehend is that a **"new" kind of electromagnetic (EM) wave has been discovered** in the empty vacuum of space which, when engineered, can be an inexhaustible supply of energy in great magnitude at any place in the universe. The word "new" is in quotes because the

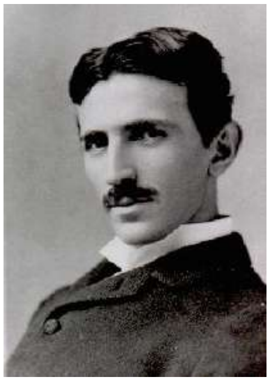

discovery really goes back to Nikola Tesla and his discovery of what he called "radiant energy." It is also not "new" because the Russians (KGB) have been working on this technology **for over 30 years** and have **weaponized** these "new" **longitudinal scalar waves** to a great degree.

These are the weapons Nikita Khrushchev spoke of in January, 1960.

> "By 1957-8 the Soviets had progressed to the point of a giant scalar EM accident in the Urals which exploded nearby atomic wastes, devastating the area. They had also progressed to development of great new superweapons using their new energetics - weapons to which Khrushchev referred in 1960 when he informed the Soviet Presidium of a new, fantastic weapon in development, a weapon so powerful that it could **wipe out all life on earth** if unrestrainedly employed.'" Tom Bearden

After over 30 years of development, and extensive testing around the globe, these new scalar electromagnetic weapons are up and running and ready to go. Tom Bearden, at his website Cheniere, dicusses the history of these new scalar electromagnetic weapons in his paper "Historical Background of Scalar EM Weapons."

Some Immediate Implications

> The implications of successful engineering of the longitudinal waves are enormous, and will change the world as we know it, one way or another. Among other things, these discoveries mean that:
>
> **1. The solutions to the energy crisis and the "oil problem" are in hand. These oil wars are unnecessary. There is endless energy available freely from the domain of time.**
>
> **2. Unbelievably powerful weapons are not only possible, but are already operating in several nations. The many powers of these weapons are unprecedented and mind-boggling.**
>
> **3. The cure of diseases such as cancer and AIDS, in fact nearly any disease, has become possible within a few years of sufficient funding. Everyone can be made healthy and stay healthy.**
>
> **4. Mind control on a mass scale has now become possible, and the machines to do it are already in place in certain nations. It has become possible to mentally enslave whole populations with the twist of a few dials.**

So the layman will need to understand that **there is a new kind of electromagnetic energy** that is altogether different from what he knows, e.g. radio, TV, cell phones, etc. The ordinary EM waves that we have known about are called **transverse** EM waves, to distinguish them from the new **longitudinal** EM waves. These scalar waves do not actually exist in our "material" world, but exist only in the vacuum of empty space, or the time domain. And we must keep in mind that this vacuum of space we speak of exists all through everything. Even our bodies are mostly empty space between atoms and molecules. So the gateway to this seething ocean of energy can be there at every point in the universe. This seething ocean of energy is all around us and all through us.

Emptiness is Full

This amazing discovery announces that **the "emptiness" of empty space is in fact not empty**, but a **great ocean** of seething energy!

Col. Bearden refers to this ocean of energy as being of the "time domain." Energy out of time? It seems like something from Star Trek but this is the point which the new science of scalar electromagnetics has reached. And where it is going may be beyond anything Star Trek could have dreamed of.

We live in a 3-dimensional world, which physics calls "3-space." But there is also **spacetime**, or 4-space, or the "4th dimension." Then suddenly comes this amazing new knowledge that **time itself is actually compressed energy**. And it is energy which is compressed by exactly the same factor by which **matter** is considered compressed energy: the speed-of-light-squared!

So we have a new companion to the famous **E=mc2**. It is now paired with **E=tc2** (where t is actually "delta-t," or change in time). Or (amaze your family and friends and) say:

"E equals em-cee-squared"
and
"E equals delta-tee-cee-squared"

It has a nice ring to it, like some freedom bell announcing a new era.

And as the atomic bomb released the compressed energy in matter, so can we now unleash **the tremendous energy that is compressed into time itself.** It gives a completely new meaning to the term "time bomb."

Some New Terminology of Scalar Electromagnetics

A new science, **several** new names

"Those [Russian] weapon scientists resurrected an old term from the history of electrodynamics, called "**energetics**". That is their approach to a **unified field theory**, where everything is based on "energetics". This model as its foundations uses a very similar approach to that "single fundamental unit" model, where energy is the unit. If one makes the energy EM in nature, then one has the Russian energetics approach. This unified approach gathers everything in, including all energy actions and relations in inert matter (the first branch of

> energetics, called by the same name), all field and matter interactions in living matter (the second branch of energetics, called "**bioenergetics**"), and all mind operations and mind-matter interactions (the third branch of energetics, called "**psychoenergetics**"). [ed: sometimes also called "psychotronics" by the Russians]
>
> "Note that the weaponeers sliced energetics into three branches, with each branch depending on the nature of the "targeting".
>
> "So I took a hint from them, since the stuff and the approach worked. They had already weaponized it highly."
>
> Bearden : http://www.cheniere.org/correspondence/021502a.htm
> See also "Subdivisions of Soviet Energetics"

Rick Andersen in an article called "What is Scalar Electromagnetics?" describes the **new science** this way:

"Scalar EM is the brainchild of Lt. Col. (retired) Thomas E. Bearden, a systems analyst and wargames specialist who has been advocating a view of electromagnetics which is based on the notion of a vast, unseen background of scalar energies (as opposed to vector energies) **which underlie all physical reality.**

"If Bearden is correct in his Scalar EM theory, then we can build devices which would enable us to **alter gravity**, time, inertia, and the apparent mass of an object. This of course has ENORMOUS implications for military applications, **space-vehicle drives, time-travel, teleportation, paranormal phenomena**, and just about every other area one can think of."
http://twm.co.nz/Beard_scalem.html

> **"Time-energy, time-currents, and time-structuring play the dominant role in electromagnetics. *Time-as-energy* eventually becomes engineerable, as easily as is spatial energy now. We are always dealing with *spacetime* and with *spacetime curvature*."**
>
> **Bearden:http://www.cheniere.org/techpapers/Vision%202000%20paper.doc**

The scalar energy discovered in the vacuum of space is sometimes called "**zero-point**" energy. Thomas Valone explains how this term came about.

What does "Zero Point" mean?

"Boyer traces the historical “creation of the vacuum” as proceeding in stages in parallel with the historical development of ideas about the vacuum. To paraphrase, he says that in the 17th century, it was thought that a totally empty volume of space could be created by simply removing all matter and, in particular, all gases. That was our first concept of the vacuum. Just get rid of all the gas.

Late in the 19th century, it became apparent that the region still contained thermal radiation. But it seemed that the radiation might be eliminated by cooling. So the second concept of getting a real

> vacuum is to cool it down to zero temperature. Just go all the way to absolute zero. Then we've got a real vacuum. Right? Well, since then, both theory and experiment have shown that **there is a non-thermal radiation in the vacuum** and that **it persists** even if the temperature could be **lowered to absolute zero**. Therefore, it was simply called the "**zero point**" radiation."
>
> Thomas Valone http://www.seaspower.com/InsideZeroPoint.htm

"For our very survival, it is absolutely imperative that informed citizens be aware of this dramatic change, which is just now starting. The powerful new science and engineering must be controlled and used for humanity's benefit, not its detriment. Else it will eventually be let loose unrestrainedly, to destroy all life on earth - a possibility indicated by Nikita Khrushchev in 1960." - Bearden

http://www.cheniere.org/explore%20articles/priore1/p4.jpg

2. Tapping the waves

Endless free energy, Tesla's radiant energy, the dipole as the gateway to the energy of the void, MEG in one year, Bedini, Yakuza, etc. End of Oil, geopolitical chaos, rogue groups

Move Over G.E. & Big Oil, Energy is Everywhere!

Third-world nations set to rise

> "Certain powerful interests did not wish the extended electromagnetics to be discovered or taught. That, after all, would lead to **free energy** and loss of economic control of the citizens. That was precisely why Nikola Tesla had already been suppressed ... Today it is still being suppressed by the orthodox establishment."
> http://www.cheniere.org/books/ferdelance/s25.htm

The vast seething ocean of energy of the vacuum, the longitudinal scalar EM waves of the time domain, can now be tapped and "transduced" into ordinary (transverse) EM energy in our 3-Space world. The process can be compared to putting a paddlewheel into a river. The energy acquired is free, since the river is there flowing whether we tap it or not. And it is a mighty river, and is not diminished by our paddlewheel.

This means that **a permanent solution to the "energy crisis" is at hand**, if only those who hold it **so secret** would give it up to humankind.

It means our species' **insane use of oil** is no longer necessary. We can make automobiles that need no fuel at all. We can build motors that power themselves. It means that unlimited electrical power can be available for free anywhere in the world. Every house and even shack can have its own power supply.

Of course there is the cost of the equipment at first, but the energy flow is free for the presumably very long life (no moving parts) of the generator. It will probably take decades to shift over to scalar power.

But the ability to tap the scalar (longitudinal) EM waves of the vacuum (which exists everywhere) also means that **unbelievably powerful weapons** can be, and already have been developed. In fact these weapons are so dangerous that the **fear engendered** becomes a psychological obstacle to the grasping of the facts.

I struggled with this myself for months, and even though Tom Bearden had completely convinced me already, it is hard to accept the fact that for everyone on earth the "worst that might happen" has suddenly gotten much worse. It takes time to deal with that, to realize it and get over it.

Indeed, the possibilities of these new weapons, which are tapping the **huge energies of time**, are so terrifying it is natural for the mind to simply say: "I don't want to hear about that."

And the government, likewise, pretends these weapons do not exist and tries to keep it secret from the people. Because **most people would indeed be truly alarmed** to know the actual situation. Col. Tom Bearden has decided that the people need to know the truth anyway. As a People on a Planet we must face this now, and **ban the use of scalar weapons in war or in mass mind control**. The current deployment of these weapons is a **world emergency** which we must all face.

The physics and mathematics of the longitudinal waves and the time domain are clearly beyond the layman, but I recommend scanning through them anyway just to get a sense of the wonder and the elegance of it all and some of the exotic terms which are used. Two of the most important papers by Col. Bearden are:

Giant Negentropy from the Common Dipole
The Final Secret of Free Energy.

Despite the complexity of the science involved, Bearden sums up the **essential secret** of it all in a few words.

> "There is no problem at all in extracting all the energy one wishes from the active vacuum, anywhere in the universe, at any time. **Just make a dipole.**
>
> The problem is in (1) catching some of that freely gushing EM energy in a circuit containing a load, and (2) dissipating the caught and collected EM energy in that load to power it, without using half the caught energy to destroy the source dipole's).
>
> That is the ONLY real energy problem on the planet, and always has been."
>
> http://www.cheniere.org/correspondence/022502.htm

A dipole could be a battery, which is polarized into negative and positive. Or the terminals of a generator. Or any magnet with its two poles. Or an atom with its positive and negative charges. Or the earth's magnetic field, the sun, and all heavenly bodies. Even a galaxy must be some kind of mega-dipole. Wherever there is a dipole there is already immense scalar EM energy ushering out of and back into the vacuum. And there have already been demonstrated a number of diverse methods for tapping that energy as it flows. (John Bedini, Edwin Gray, Bruce de Palma, and others). Bearden provides what he calls a "partial list" on his website.

> "In other words, **nature then gladly gives us** as much EM energy flow as we need, indefinitely — just for paying a tiny little bit initially to "make the little dipole." After that, we never have to pay anything again, and nature will happily keep on pouring out that 3-flow of EM energy for us. This is the **giant negentropy** mechanism I uncovered, performed in the simplest way imaginable: just make an ordinary little dipole." Bearden
> http://www.cheniere.org/techpapers/Unnecessary%20Energy%20Crisis.doc

Important note 2006: Since the enthusiastic early report below, the hopeful timetable envision has gone quite off track. Bearden had said that the MEG could appear as early as 1-2 years ***after getting funding*** for the final research and scale-up. They have **still** NOT gotten the funding as of 2006.

Click here for a current status history of the MEG from Tom Bearden

The original text of article continues here...

How to Buy a Free Energy Machine

Answer: Just wait one more year! 2003. Then get off the Grid!
[note 2006: The above line was unfortunately naive and overly optimistic! - author]

Perhaps the most eagerly awaited event now is the marketing of Bearden's **M**otionless **E**lectromagnetic **G**enerator (MEG) scheduled to begin in about a year's time. This is when the proof will hit the fan. Many people have said to me, when talking about free energy, "I'll believe it when I can go to the store and buy one." Hopefully in a year's time they will be able to order their own MEG. And then they will believe, because it works and is working.

Says Bearden: "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat."
Bearden : http://www.cheniere.org/correspondence/011202.htm

Soon the cat will be fully out of the bag. The first MEG units will supply 2.5 kilowatts of free electricity. Forever. Units can be connected together to double or triple the wattage. After that gets going, 10 kilowatt generators are planned. Energy will flow freely from the time domain indefinitely, and there are no moving parts to wear out. This manufacturing project is a heroic effort and will change the world as we know it. For once people are in possession of such devices there can be no further denial from anyone, including the government and/or "secret-government."

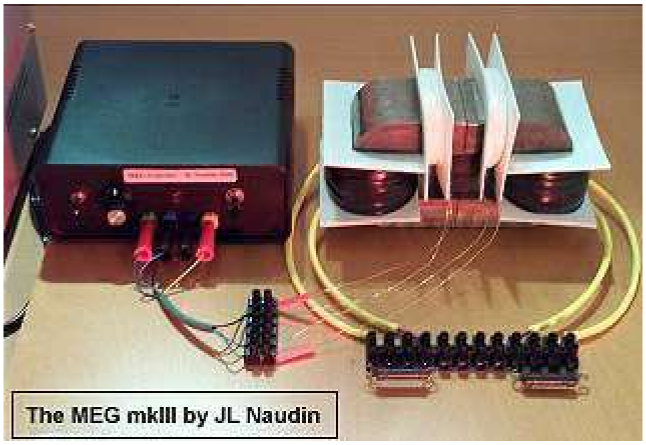

The Motionless Electromagnetic Generator
(Successfully replicated here by J. L. Naudin)

The physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum."

The device was successfully replicated by J.L. Naudin in France, and others, like Steve Utne, are also working on their own models. But **Magnetic Energy Limited's MEG** is actually scheduled for production soon and looks to be the first commercially available free-energy generator in history!

I imagine it will not take long for tinkerers to put them into their electric cars and make the first fuelless automobiles since Tesla's unheralded (and unheeded) experiments. Or for builders to begin building houses powered by MEG's. Houses which will never see electric bills. How can they ever make enough of these to meet the demand that will suddenly be there? Bearden is right, it needs to be an all-out crash program of the government **on the order of the Manhattan Project**.

I must include here that Bearden is quick to point out that he is not himself the inventor of the MEG, but is a partner and advisor in bringing it to the public market through **Magnetic Energy Limited, Inc**. I must also admit that I am praying for the success of this venture, and the ending of the absurd and criminal withholding of this technology from the people by the black-ops military-industrial people. This secrecy has cost the planet over 30 years of unnecessary pollution.

The first units off the assembly line, which is being set up in a "friendly" unnamed nation, will usher the public at last into the new era of scalar electromagnetics.

As of December, 2001, the situation with the MEG was described by Bearden in a letter.

> "What we presently have with the MEG is a successful laboratory experiment. At least a year's very hard research will have to be done before we will be ready to put a commercial power supply into production.
>
> "Consequently, we have made an agreement with a foreign partner (the National Materials Science Lab of the National Academy of Sciences of a friendly foreign nation) to do that year's research. At the same time, we are trying to make an agreement with one or more large financial partners here in the U.S." http://www.cheniere.org/correspondence/121901.htm

> "So we are hopeful that we will be able to start introducing units on the market about a year from now. Quite simply, we will either succeed or we will fail. But we will give it our very best effort."
> http://www.cheniere.org/correspondence/012202a.htm

Bearden has given another technical explanation of how the MEG works here in a letter.

Update : Motionless Electromagnetic Generator patented, 3-26-2002

Click here to read the patent

"In Magnetic Energy Limited, all business discussions and arrangements are taken care of by Dr. Lee Kenny, Managing Partner. There are five of us who are co-inventors of the MEG, and two of the others are really the principal inventors." Bearden
http://www.cheniere.org/megstatus.htm

Lee Kenny can be reached by E-mail at: jlkenny@ingr.com

Free Energy has been Suppressed

Secret Government Aware of Free Energy

"In this world of contemporary times, all the agencies, CIA, FBI, KGB, NSA, Electric Power Research Institute, DARPA, Brookings Institution, Henry Kissinger, Edward Teller, et al., all are aware of [free-energy researcher Bruce] DePalma and his ideas.

"Because these individuals and institutions are employed by the ruling elite to forecast the future and satisfy present needs and demands, it is clear that **free energy is a threat** to the world order constructed by **business** and the **mindset of those who want to own the world**.

"Free energy represents Man's aspirations and dreams of freedom and equality, uniform division of resources and the ability to choose one's own future.

"The fact that free energy is suppressed speaks to the greed and self-interest of a ruling elite which, even in the face of an emergency of global starvation, resource depletion and environmental pollution, will not give one inch if it means loss of control. This is an attitude of **paranoid delusion** and fantasy which can only arise from the alienation of a **class of elitists** who, through their money, are completely **insulated** from the reality of the day to day fight for existence of the common man."

http://depalma.pair.com/Absurdity/Absurdity07/ProblemOfFreeEnergy.html

"Before the turn of the century, **Nikola Tesla** had discovered and was utilizing a **new type of electric wave**. Tesla repeatedly stated his

> waves were non-Hertzian, and his wireless transmissions did not fall off as the square of the distance. His discovery was apparently so fundamental (and his intent to provide free energy to all humankind was so clear) that it was responsible for the withdrawal of his financial backing, his deliberate isolation, and the gradual removal of his name from the history books."
>
> Bearden http://www.cheniere.org/books/part1/starting%20pages.htm

Bearden describes one clear case of the suppression of free energy when the Yakuza killed the **Kawai engine** in his very presence in Huntsville, Alabama.

> "That night a jet arrived from Los Angeles, with a Yakuza on board. The next morning Kawai no longer controlled his company, his invention, or his own fate. The Kawai party was in fear and trembling. The Yakuza coldly dissolved the agreement, they packed up the two Kawai engines we had, and left. And that was that."

So it is not only governments who are keeping this revolutionary technology from seeing the light of day. The Yakuza? KGB? The Aum Shinrikyo cult? Our "leaders" need to be honest about what is going on, otherwise they look foolish, acting out a kind of "Alice In Wonderland" script before the public, a script that has become completely unbelievable.

Concerning the criminal withholding of this scalar technology by the military black-budget people, and the cover-up of the actual situation regarding scalar weapons throughout the world, Tom has this to say:

> "The high government officials of the United States of America grossly violated our constitution and their sworn oaths of office. Our constitution specifically charges the government of the United States to **defend** the people of the United States. That means our officials are specifically charged with defending our people. Yet they deliberately agreed not to do so, and they deliberately sacrificed us in advance.
>
> **"They committed treason of the highest kind, by deliberately surrendering us to total destruction anytime the big balloon goes up."**
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Energy Crisis Solved

One of the "must read" documents at Cheniere is The Unnecessary Energy Crisis: How to Solve It Quickly, especially since the "**oil wars**," which Bearden has foreseen for decades, have now already begun.

Here are some clips from that Document:

> "To avoid the impending collapse of the world economy and/or the destruction of civilization and the biosphere, we must quickly replace much of the "electrical energy from oil" heart of the crisis at great

> speed, and simultaneously replace a significant part of the "transportation using oil products" factor also.
>
> "The technical basis for that solution and part of the prototype technology required, are now at hand. We discuss that solution in this paper." Bearden

And although the solution to the energy problem and the oil wars may be at hand, Bearden suggests there is still an urgent rush that needs to be made to save human civilization from collapse.

> "Once the technology hardware solutions are ready for mass production, even with a massive worldwide deployment effort some five years are required to deploy the new systems sufficiently to contain the problem of world economic collapse. This means that, by the end of 2003, those hardware technology solutions must have been completed, and the production replacement power systems must be ready to roll off the assembly lines en masse.
>
> "The 2003 date appears to be the critical "point of no return" for the survival of civilization as we have known it.
>
> "Reaching that point, say, in 2005 will not solve the crisis in time, and the collapse of the world economy as well as the destruction of civilization and the biosphere will still almost certainly occur, even with the solutions in hand."
>
> "The **good news** is that we now know how to easily initiate continuous and powerful "electromagnetic winds" from the vacuum at will. Once initiated, each free EM energy wind flows continuously so long as the simple initiator is not deliberately destroyed."

On page 9 he explains the concept of the **dipole**, and how there is already huge scalar energy flowing in the time domain around any dipole. Hold a magnet or a battery in your hand and you hold a stress in time and a beautiful flowing loop of longitudinal waves!

Bearden decries the scientific community in the West for being unable to let go of their dogma long enough to see that the **greatest discovery in human history** lies at their feet, but for their stubborn ignorance.

> "The present energy crisis then is due totally to that "**conspiracy of ignorance**" we referred to, that is maintained by the scientific community, and that has been maintained by it for more than 100 years.
>
> "This is the real situation that the environmentalists must become aware of, if they are to see the correct path into which their energies and efforts should be directed — to solve both the energy crisis and the problem of gigantic pollution of the biosphere."

As far as anyone in the legislative branch of government ever understanding the profundity of the discovery and engineering of scalar waves, Tom Bearden says they are at a great disadvantage in doing so.

> "The nontechnical person — e.g., a Senator or a Congressperson — is operating under a distinct disadvantage. If he or she takes the stance that much better electrical power systems can readily be achieved, **he or she is in fact opposing almost the entire set of University, Government Technical, Power Company, Battery Company, and Organized Science communities**. Further, in most cases his technical advisors are themselves from one or the other of those communities, and likely to go back into that community or those communities when the Senator or Congressperson leaves office, or even before. So the Congress and the non-technical government community at large operate at a great disadvantage...Bearden

National Emergency Declaration is Needed

"The Government *Non-Technical* community (the Senate and the Congress, in particular) are in far better shape than the Government *Technical* community, to appreciate the **world implications of the pending economic disaster**. I am hopeful that both the environmentalists and the Government Non-Technical community will rapidly unite in a common goal to **get this vacuum energy program launched**, under a **National Emergency** declaration. If so, then they can solve the energy crisis and the pending economic crisis, in fairly short order, and permanently."

unnecessary_energy_crisis.htm#congress_disadvantage

In the conclusion of his paper The Unnecessary Energy Crisis: How to Solve It Quickly Bearden again states the urgency of replacing our centralized power grids with highly dispersed and local taps on the freely flowing energy of the active vacuum. Given the abilities of the Tesla howitzers, the power grids of all nations are sitting ducks.

"Personally, the present author regards the increasing **energy crisis as the greatest strategic threat** to the United States in its entire history. I will do anything within my power to help prevent what I perceive to be the looming economic collapse of the Western world, preceded or accompanied by a sudden, explosive, all-out and continuing exchange of the WMD arsenals of most of the world.

...*In the name of all humanity, let us begin!* Else by the time this first decade of the new millennium ends, much of humanity may not remain to see the second decade."

3. Weaponization

Scalar interferometry, Tesla howitzer, Tesla dome, heat/cold, dudding, 10 nations have them, dead-man fusing & Insane Systems, scalar strategy, earthquakes and weather wars, solar storms, mind attacks, leased to Yakuza, many examples of testing, quantum potential weapons.

> **"Indeed, Soviet energetics weapons are now capable of destroying both our triad shield, our homeland, our armed forces in the field, and our population, quickly and efficiently. We have a new "gap" of monumental proportions: not a missile gap, not a submarine or bomber gap, and not even a particle beam or laser gap. We have a scalar electromagnetics or electrogravitation gap."** Bearden
> http://www.cheniere.org/books/ferdelance/intro.htm
>
> "**We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that, if unrestrainedly used, it could wipe out all life on earth. It is a fantastic weapon."** Khrushchev, to the Presidium, Jan. 1960
>
> "We've totally missed the most incredible "secret superweapons" development program of all times, and the worldwide testing of the weapons themselves.
>
> "Hiroshima and Nagasaki should serve as object lessons in the price of ignorance of such technological breakthroughs.
>
> "History does repeat itself. But this time it's not we Americans who got the great new superweapon first."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Warfare has been changed forever by the development of these scalar energy longitudinal wave howitzers. To get a basic understanding of scalar waves is to have the imagination suddenly run wild as all the implications and possibilities regarding warfare fall into place. One realizes with a certain horror that the world has totally changed, and that there are some very fearsome possibilities. Remember, the power for these weapons comes from the time domain, longitudinal EM waves in the vacuum of empty space, and the power is tremendous and mind-boggling. Being able to blast away at any target from a distant control booth is something that has never happened before. This is incredible power to be in control of and it divides the history of weaponry into "before" and "after." And the destructive power of these weapons is delivered instantaneously to the target **from the local vacuum at the place of the target.**

From an interview:

Questioner: "Are you saying that real energy can be transmitted faster than the speed of light?"

Bearden: "Yes. But hyperspatially. Not through 3-space, but "around" it. Again, this means that it's possible to produce energetic changes in a distant system at a distant place, without transmitting energy "through space" in the normal sense. You transmit "around" space, so to speak, and directly in time or through a higher dimension, depending on the model you're using to understand this. Further, the speed of the internal EM energy transmission is not limited to the speed of light."

http://twm.co.nz/beard_interview.htm

Here are some initial considerations about scalar weapons and scalar wars

1. Tesla Howitzer 2. Tesla Domes 3. Rogue groups 4. WW III 5. Russian "UFOs?"

1. The Tesla Howitzer and its modes of action

Source: Tom Bearden's weapons slides and Fer-de-Lance Briefing

For one thing there is no delivery of any thing to any place. Bombs are obsolete. So are the planes to get them there. No helicopters necessary. All current nuclear devices are rendered nearly useless by electronic "dudding." All distant-destruction can be done from a control room, and to any point on earth. The destructive power of "men" has just increased by orders of magnitude. The new weapons could even cause storms on the sun! Thus there is great urgency to make these **new facts public** and known, with the hope that all nations would come together to ban the use of scalar weapons.

Two scalar antennae together along with the computers to control them, make up a scalar interferometer, and according to Col. Bearden the Russians have hundreds of such installations already. These interferometers are called "**Tesla Howitzers**." They can deliver a giant blast of energy to the distant target site, true "action-at-a-distance".

The **first howitzer mode** is called the "**exothermic**" mode of operation because immense EM energy blasts outward at the target site. The blast of a scalar howitzer can be of near nuclear level in destructiveness, and can be repeated easily, at that place, or nearby, or anywhere.

The howitzer can use a lesser exothermic power setting and simply **destroy all electronics** in the target area. Thus they can render our nuclear missiles inoperable as they sit in their silos by "frying" the electronic circuits that guide them. They can bring down any airplane, anywhere in the world, at any time. Any person anywhere, if their exact position is known, can be assassinated without a shot being fired.

The howitzers can also bring down power grids anywhere in the world, at any time. Thus they could bring Wall Street to an instant halt by disabling all its electronics. Even this one little easy tweak of the howitzer could cripple the U.S. economy in a single day. No wonder the "**leaders**" don't want you to know! This is a new era we are in. The discovery and development of scalar interferometry has put us in a **new world**, whether we like it or not. Anyone could be expected to go into denial, the **befuddled** senators and congressmen included. But we have **to wake up**.

In a **second howitzer mode** called the "**endothermic**" mode, the howitzer sucks energy out of the target area, essentially creating a blast of cold at the distant target. It is even capable of freezing parts of the ocean. Bearden gives a number of cases where these "**cold explosions**" have been witnessed, mostly by airline pilots. April 9, 1984: 1 2 3) A huge mushroom cloud of mostly water is seen rising miles into the sky from out of the ocean, an awesome sight.

In the endothermic mode the sucked-out energy must go somewhere, so it is vented out at some other chosen spot on the earth. These endothermic plumes have been photographed by satellite.

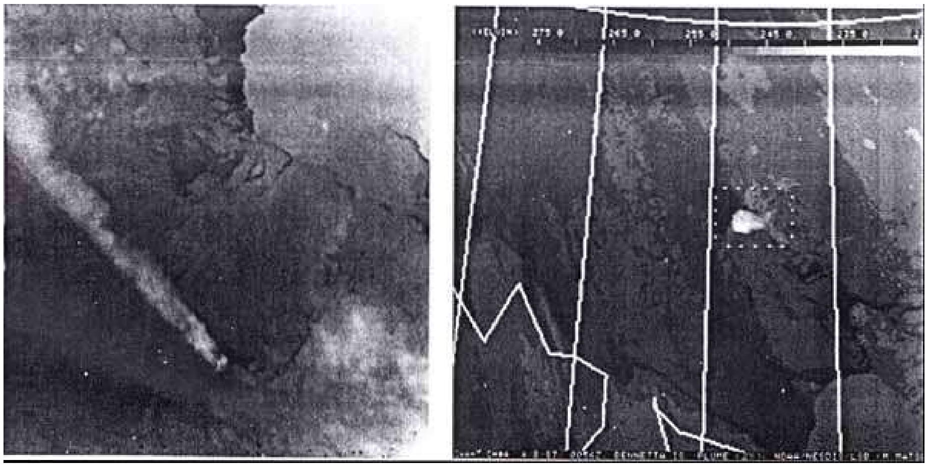

"The exhaust in the left picture is about 150 miles long, and is nearly horizontal (about 1.5 degrees above the horizontal). It is entirely consistent with the continuous exhaust from a "dumping transfer" scalar EM howitzer in the continuous exothermic mode. (The primary howitzer, of course, was activated in the endothermic mode.)" Bearden More Plumes

By using both of these **heat** (exothermic) and **cold** (endothermic) modes together the weather can be altered anywhere. Warm the air over here, cool it down over there, put a curl in the jet stream, dissipate clouds, create clouds, whip up a tornado. Bearden speaks about using the **endothermic howitzer mode** as a weapon of war:

> "Cold explosions can be used to freeze tanks, personnel, and equipment. The equipment and tanks thaw out. The personnel thaw out too, but they are dead when they do." Bearden, Fer-de-Lance

(This endothermic mode of the howitzers makes me think that there might one day be a cure for the problem of global warming after all. Could not a big device, situated in space or on the moon, be able to vent a great deal of excess heat outside the atmosphere into space, thus functioning as a kind of planetary thermostat?)

While the United States has experimented with electomagnetic weapons using ordinary **transverse** EM waves, the **longitudinal** [**LW**] weaponized waves are truly of another order altogether.

For one thing they do not have to travel through space, for they come from the time domain, which is everywhere. The energy of the blast, the heat or light or whatever, comes from the vacuum of space **at the location of the target itself**. This is astonishing. The weapons actually just trigger the release of immense energy from the vacuum at the target location.

> "... it is possible to focus the potential for the effects of a weapon **through spacetime itself**, in a manner so that mass and energy do not "travel through space" from the transmitter to the target at all. Instead, ripples and patterns in the fabric of spacetime itself are manipulated to meet and interfere in and at the local spacetime of some distant target. There **interference of these ripple patterns** creates the desired energetic effect (hence the term energetics) directly in and through the target itself, emerging from the very spacetime (vacuum) in which the target is imbedded at its distant location." Bearden Fer-de-Lance

The big Tesla howitzers are aimed at their targets by using a worldwide electromagnetic pattern called the "Woodpecker Grid," begun by the Russians in 1976.

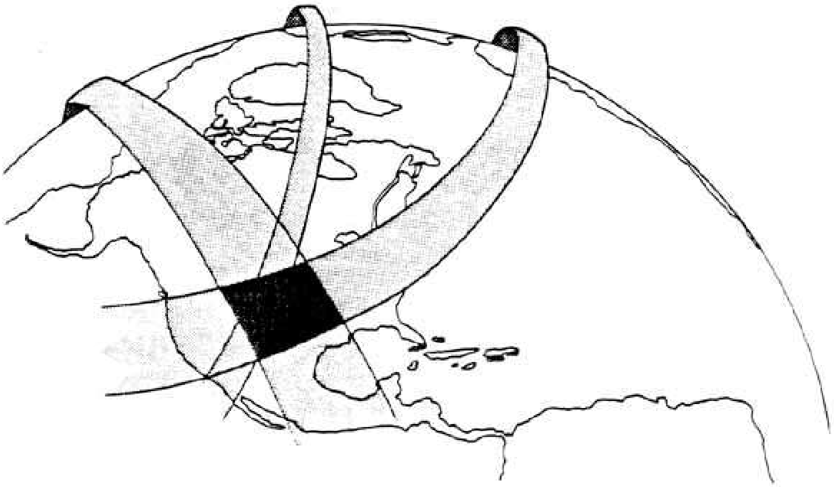

You can hear the Woodpecker signal here:
http://www.cheniere.org/misc/woodpeckersound.htm

The woodpecker grid is just that, a grid of EM waves which provide a channel for the **LW** waves to any point on earth. Computers use the woodpecker grid information to hone down the aiming of the howitzers **to a pinpoint location on earth**. (For that matter, it could be under the earth, in space, or under the sea).

A third howitzer mode is what Col. Bearden calls "**Mindsnapper**" mode. This is a truly frightening mode which affects the electromagnetic mind-body connection. Being mind-snapped at a low level would cause you to lose consciousness. Being mind-snapped at a slightly higher power would "entrain" all minds in the target area into a kind of hypnogogic trance, a state in which they would all be highly susceptible to suggestions and orders.

Being mind-snapped at high level would so disrupt the mind-body connection that everyone in the target area would fall down, instantly dead. Not only that, everything around you and inside you would be killed, so that your sterilized body would **not rot** for quite a long time.

The **Mindsnapper** can be used for a small area, or to **kill an entire population**. In fact, using the scalar waves to affect the human mind has become a new field in its own right, and has been dubbed "psychoenergetics." (This is discussed below in section 5).

Bearden describes the use of the howitzer in **mind-snapping mode** against people, to cause unconsciousness or death:

> "From a military viewpoint, these scalar EM weapons are really quite all-around weapons.
>
> "For example, they are very lethal against personnel.
>
> "Set in the "high intensity pulse mode," one blast and a person's nervous system is destroyed instantly. Total, instantaneous death results. One can blast away at entire groups of infantrymen, for

> example. It's the peak power that's important, not the average power. So the weapon doesn't require too much battery power.
>
> "Set on a lower intensity pulse mode, one zap can simply knock out a person, rendering him instantly unconscious. Quite useful if one needs prisoners to interrogate. Also very silent, which is useful in operations employing stealth and surprise." Bearden http://www.cheniere.org/books/ferdelance/s80.htm
>
> "Those hit by the scalar EM weapon, however, have a most peculiar death mode.
>
> "Death comes-instantly and totally. There is no convulsion, no response. The entire nervous system is destroyed instantly. Every living cell in the body is killed instantly, including all bacteria, germs, etc.
>
> "A body hit with this thing falls like a limp rag and lies where it falls. It doesn't decay in even 30-45 days. In a macabre fashion, it's been reduced to something like food irradiated with nuclear radiation; everything is killed, so the material is preserved for an extended period before any decay can set in."
>
> Bearden http://www.cheniere.org/books/ferdelance/s81.htm

He goes on to say that the same "bazooka-sized scalar EM pulse weapon" can also be used to knock out a tank with one shot (or pulse). So tank warfare is obsolete. Helicopters? Bearden points out: "... if he's attacked by a helicopter or a fixed-wing aircraft, he can bring the same weapon to bear on it. He can knock down a chopper or a jet. With one shot."

Obituary for Tank Warfare

Tanks useless against portable LW weapons

"For example, suppose one attacks a tank. [with an LW weapon]
With scalar EM pulse, the personnel die instantly. Total personnel kill is achieved.
All electronic systems of the tank are dudded. Total systems kill is achieved.
The ammunition in the tank explodes. Total kill is achieved.
The fuel explodes. This is another total kill mechanism.
That tank has been killed totally, by a variety of mechanisms, all simultaneously.
That's a k-kill in any analyst's book.
And don't worry about retrieval and repair. That tank is finished permanently.
Even the most modern tank is just as vulnerable as the most obsolete."

Bearden http://www.cheniere.org/books/ferdelance/s88.htm

One begins to get the idea how profoundly the conventional means of warfare have been obsoleted. All the usual weapons of war have been rendered nearly useless! And **Star Trek's "Phaser**," with its "stun" and "kill" modes is here at last.

Although it is the weapons aspect of scalar electromagnetic which has received the most developmental effort, Bearden points out that a great many other amazing things that may be accomplished in this new field of science. Beam me up Scotty!

> "Superluminal communications systems, hyperspace drive, and **materialization** and **dematerialization** are all hypothetically possible, using scalar electromagnetics. As the technology develops, we should see the development of many of the systems long thought impossible except in science- fiction."
>
> Bearden http://www.cheniere.org/books/ferdelance/s15.htm

Bearden describes a huge test of the Russian scalar weapons systems which occurred in 1985, a test that went completely unnoticed by western scientists:

> "Just before May 1, 1985 the Soviet Union performed a "full-up" operational test of their entire strategic scalar EM weapons complex. This test was detected and monitored periodically by Frank Golden. Some 27 of these giant Soviet "power taps" were locked into the molten core of the earth, producing forced (entrained) scalar resonance of the entire planet on 54 controlled frequencies under our very feet. The remainder of the scalar frequency spectrum was ablaze with literally hundreds of Soviet scalar transmitters: probably the entire strategic scalar command and control system to underwater submarines, higher command centers, distant commands, etc. was activated in the giant exercise.
>
> "For several days, the system was exercised on a mind-boggling scale, apparently as part of the Soviet Union's highly accentuated 40th anniversary celebration of the end of World War II.
>
> "Ironically, not a single U.S. intelligence agency, laboratory, or scientist detected this monstrous exercise for the new Soviet leader, Gorbachev. Not a one of them had a detector for scalar EM radiation, and not a one of them officially believes the exercise ever happened."
>
> Bearden http://www.cheniere.org/books/ferdelance/s38.htm
>
> "The **Woodpecker grid/howitzer** weapon system can be placed over the ocean and used against cruise missiles, naval surface-to-air and surface-to-surface missiles, submarine-launched ballistic and cruise missiles, etc. Placed over a carrier task force, it can also take care of the aircraft launched by the carrier as fast as they are launched. It can also handily take care of the missiles launched by guided-missile cruisers of the accompanying task force."
>
> (Bearden http://www.cheniere.org/books/ferdelance/s65.htm)

The whole series of the Fer-de-Lance briefing slides and comments is a "must read" for anyone wanting to be aware of the actual state of things in this world right now. It is a new world and it takes some getting used to. Fer-de-Lance speaks clearly, for the layman, of the way things stand.

What is it like to operate these Tesla Howitzers? Col. Bearden gives us an idea in his "Fer-de-Lance" briefing, Slides 66-71.

Who needs armies anymore? A entire war can be fought and won by a handful of men sitting in a comfortable control booth! The scalar electromagnetics revolution has made all previous forms of warfare nearly obsolete.

Why haven't the Russians used the scalar weapons yet?

Bearden gives his thoughts on this in "Fer-de-Lance." The power of these weapons is so great that an accident or war could wipe out all life on earth, and even **affect the sun**.

> "As can be appreciated, this entire concept [MAD] is now totally obsoleted by the Soviet possession of scalar EM weapons.
>
> 'The "mutual assured destruction" capability -- or MAD concept as it was so aptly named! -- is **no longer mutual.** And for two reasons: (1) the Soviet surprise attack could destroy essentially all our Triad forces wherever they are, in the first attack. (2) Even if a very small number of our strategic forces survive, the puny strike we could attempt to launch could be repeatedly decimated from launch to terminal phase.
>
> 'The restraint on the Soviets is not because of our Triad forces. The restraint is twofold:
>
> > (1) a second nation has scalar EM weapons, and would undoubtedly unleash them upon the Soviet Union with great violence if such a massive world-wide destruction of U.S. forces occurred or was threatened (is it really "accidental" that six major Soviet missile ammunition storage sites mysteriously exploded within seven months?);
> >
> > (2) the closed-loop **scalar radiation exchange system of Earth-Sun could be inadvertently "tweaked"** in the feedback loop from Earth to Sun, so that a large solar scalar resonance was stimulated. In that case the Sun could emit a mighty resonant "burp" that would simply **wipe out all life on Earth in a fiery shower**.
>
> 'The "terror" in the "balance of terror" between nations has just increased by many, many orders of magnitude.
>
> 'The biblical prophecy that the Earth will be wiped out by "fire and brimstone from heaven" is very near fulfillment now.
>
> 'Put another way, the pucker factor -- even on the Russians who may contemplate unleashing Armageddon -- is higher than any Western strategic analyst has ever imagined, even **in his wildest nightmares**.
>
> 'Truly these weapons are "more frightening than the mind of man has imagined," as Brezhnev put it in 1975.
>
> 'The frightening scalar EM weapons can be used, but only very, very gingerly indeed. If a slight mistake is made, everybody loses everything."
>
> Bearden http://www.cheniere.org/books/ferdelance/s74.htm

But the Russians (KGB) are not the worst of it. Bearden claims that some of the scalar weapons have been leased to the Yakuza for about a billion a year, and that only the KGB is holding the Yakuza in check from making a **full-scale all out scalar attack on the west.**

U.S. Defense Secretary William Cohen

Warns about eco-terrorism using scalar electromagnetic weapons

> "Others [terrorists] are engaging even in an eco-type of terrorism whereby they can **alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves...** So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...**It's real**, and that's the reason why we have to intensifour [counter terrorism] efforts." - Defense Secretary William Cohen, 1997

In a letter to a writer named "Russell" Bearden says,

"In short, Russell, the Secretary of Defense of the United States confirmed that there are indeed novel kinds of EM weapons, right now and have been for some time, which have been and are being used to (1) initiate earthquakes, (2) engineer the weather and climate, and (3) initiate the eruption of volcanoes. We wrote about those exact uses of the weaponry decades ago. Several nations now have such weapons. Three of them (two on one side and the other on a hostile side) are even firing practice shots into Western Australia, as a convenient test range."

http://www.earthchangestv.com/ufo/0209gandor.htm

See Also:
Examples of scalar weapons testing around the globe

See Also: Powers of the Longitudinal Interferometers

Tesla Howitzers in Weather mode

How to Make Weather

"Engineering the weather is **duck soup**; they tested that over the U.S. in 1967, and entered upon continuing operations over North America on July 4, 1976 as a grim kind of KGB "Bicentennial Gift" to the United States. Here's how it is done with several interferometers.

"First, the interferometers can deliberately make "high pressure areas" (cool the air so it shrinks and its "footprint" pressure on the ground increases because its density increases) and "low pressure areas" (heat the air so that expands and its "footprint" pressure on the ground decreases because its density decreases). Well, if one makes the highs and lows where one wishes them, and judiciously and somewhat slowly moves them along a given path, these highs and lows will entrain the jet streams and thus "steer" the weather.

"So if you wish a very cold snap or attack, go up into Canada and start these actions to steer down some streams that bring some very frigid air. Establish other highs and lows judiciously to "block" or "slow" other jet streams and flows as desired. In this way, one can pull the "large cold air masses coming down from Canada" deep into the southern U.S. at will. They do it regularly. If you wish an ice storm, add an additional current of warm moist air you bring up (steer up by using artificial highs and lows) from the ocean. Where they meet, you will get freezing rain, then sleet, then severe icing accumulating. This type of attack does rather large damage to the struck area."

http://www.earthchangestv.com/ufo/0209gandor.htm

Whole websites have now sprung up showing weather **radar anomalies** which may well be indications of scalar weather engineering. One list of over 500 images is here. Cheniere has its own collection here.

Here is an animated anomaly with cloud formation.

In another section at Cheniere are many photos of **cloud anomalies** which reveal the underlying grid along which the clouds are forming or un-forming. Bearden relates a startling incident in these **ongoing "weather wars**."

> "During the spring of 1986, abnormally strong Soviet weather engineering occurred over the U. S., causing a drastic drought in the southeastern U.S. **This drought was broken by a colleague who used an extremely powerful scalar EM device** to redirect jetstreams. A most unusual and unique signature of the "blocking" against the Soviet scalar EM actions resulted: Two huge circulations developed in the atmosphere, clearly showed as two adjacent giant "holes" in the swirling cloud cover over the middle and eastern U.S.
>
> "Between these two giant holes, the cloud circulations formed a stream of clouds, moving to the south, looking very: much like a giant vertical "bar" of a huge "Y-shaped" cloud flow.* Several national weathermen commented on this unusual pattern, which had not been observed before. The pattern continued, day after day, as the blocking continued."
>
> * On the weather maps.
>
> http://www.cheniere.org/books/analysis/history.htm

Tesla Howitzers in Earthquake mode

How to Make an Earthquake

"Here's how you initiate a very large earthquake with such weapons. Take a convenient fault zone of set of them. Focus the interferometry on the fault zone, in the "diverging" mode, and deposit EM energy there in the rocks on both sides, increasing (slowly) the stress in the rocks by the reverse piezoelectric effect (deposit excess energy, get crystal mechanical movements).

"Do it slowly, and the stress will build up to large pressures well-above a plate slip minimum energy required. At some point, the rocks yield and one or both sides "slip" and move rather sharply, giving a very large earthquake in that zone.

"Do the same thing down in the earth (remember, LWs easily penetrate right through the earth and ocean at will, and so the "interference zone" focus can be inside the earth or beneath the ocean, at will.

"Anyway, focus this thing down to where the active part of the volcano is still slumbering, down where the hole in the plate has been made. Keep increasing the deposition of energy in the magma itself, and eventually the increasing pressure from deep within that volcano, underground, will cause an eruption. Build the energy slow, and the eruption will likely be much larger."

http://www.earthchangestv.com/ufo/0209gandor.htm

Yugoslav Earthquake, Followed by Cold Explosion
Strange Characteristics of Iran Quake, Sep. 1978

Tesla Howitzers in Tesla's "Big Eye" mode

How to make a Big Eye in the Sky

By using longitudinal wave interferometry in a weakly endothermic mode a new kind of "radar" is now possible, one which can look right through the earth and oceans. Because the endothermic mode extracts energy from the distant target, a pattern of the target area can be scanned for in the returned energy.

"Thus the receiver produces a representation of the energy extracted from various locations within the distant endothermic zone. By displaying the received signals on an appropriately scanned screen, a representation of the distant scene in the intersection zone can be created. Indeed this is a special kind of "microwave interferometry," and -- with modern techniques -- the imagery obtained might be surprisingly good. With development, it might even become as good as the image presently obtained by side-looking acquisition radars.

"Interestingly enough, since scalar EM beams will easily penetrate the earth or the ocean, one can also look beneath the earth or beneath the ocean with this type of scanning scalar interferometer.

"The importance of this capability to strategic and tactical reconnaissance is obvious. Camouflage, cover, and concealment have no effect on such a system. One can easily look inside buildings and into underground facilities. With a small system such as this, the U.S. Marines at Khe Sahn would have had little difficulty locating the tunnels continually dug under the perimeter by the Viet Cong. And targets under jungle canopies are directly visible."

Bearden http://www.cheniere.org/books/ferdelance/s42.htm

Tesla Howitzers can destroy all life on earth

How to Make a Storm on the Sun!!

(Just make a big mistake with your Tesla howitzers)

The **deepest dangers** of the use of scalar weapons are frighteningly pointed out by Bearden here where he considers that the sun, earth and moon (actually the whole solar system) is a delicately balanced arrangement of scalar (longitudinal) EM radiation. There is a scalar connection between the earth and the sun, which are both giant dipoles gushing energy in the time domain. Big scalar events on the earth could alter this balance causing **true catastrophe**.

"Indeed, a solar response could be stimulated so that the Sun would violently belch and destroy our biosphere, among other effects."

Unrestrained scalar warfare could cause huge solar storms? To realize that **human activity** could actually **affect the sun itself** is truly mind boggling.

Bearden goes on to say:

> "Accordingly, use of huge scalar EM weapons is a double-edged sword. Unless carefully employed, use of the weapons could cause a terrible backlash to the user as well as the victim, and even accidentally cause the **destruction of the earth itself**. It is not accidental that in 1960 Khrushchev stated that his new fantastic weapon could -- if unrestrainedly used -- destroy all life on earth.
>
> "This appalling backlash potential is apparently what prompted Brezhnev in 1975 to make so strong an effort to get agreement to outlaw the development of "new weapons of mass destruction" more frightful than the mind of man had even dreamed of. Gromyko even introduced such a draft agreement into the proceedings of the United Nations. The first article of the draft agreement indicated that the nature of the weapons referred to would be negotiated. Sadly, the entire Western

> world did not even know what the Russians were talking about."
>
> http://www.cheniere.org/books/ferdelance/s24.htm

2. Tesla domes provide near-perfect defense shield.

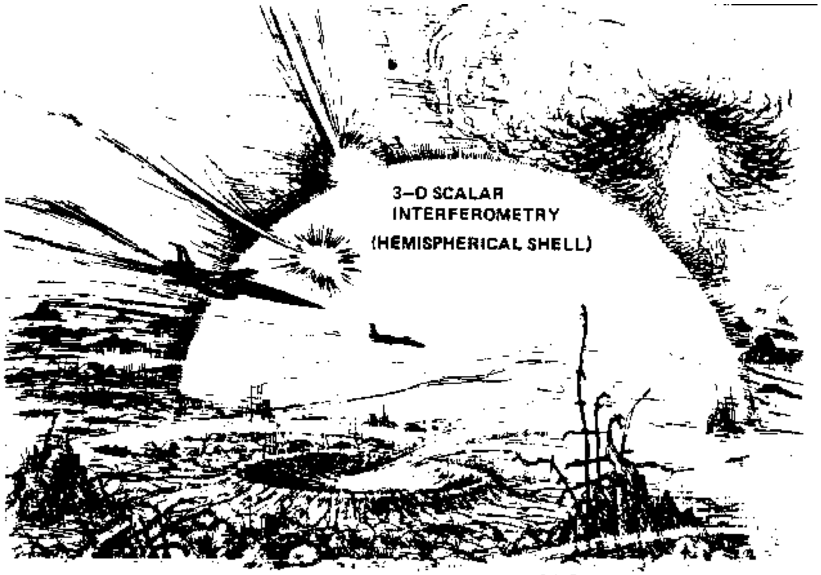

(Slide from Cheniere.org)

The Tesla howitzers can be used in a nearly **impregnable defensive mode** whereby they throw up a dome (**Tesla Dome**) or a sphere (**Tesla Globe**) of highly powerful electromagnetic energy, enough to "dud" or destroy missiles which try to penetrate them. Once again, this energy is not going through space from the howitzer, but being made to emerge from the local vacuum **at the location of the shell**. KGB tests of these domes have been witnessed by airline pilots around the world. Bearden gives many examples of the Tesla domes being sighted around the world in his briefing paper "Fer-de-Lance." (Russian dome test, another globe incident, yet another globe test). These giant electromagnetic domes can be hundreds of miles across, or narrowed down more to total impermeability.

> "Such a shell may be several hundred miles in diameter at the base. The enormous energy required to form such a defense shell is obtained by a "scalar power tap" into the **molten core of the earth itself**, as previously explained. In late April/early May of 1985, 27 such "**power taps**" were placed in the earth by the Soviets. If each tap is capable of powering four to six large scalar EM weapons, then the **Soviet strategic scalar EM arsenal contains over 100 monstrous superweapons** capable of generating exothermic explosions, endothermic explosions, engineering the weather, locating and destroying underwater submarines, detecting and destroying ballistic missiles shortly after launch, detecting and

> destroying long range strategic bombers as soon as they are airborne, etc."
> http://www.cheniere.org/books/ferdelance/s52.htm

By using nested domes one can protect the domed area **even against nuclear radiation itself**. Tesla globes can be used to hit airborne targets by simply placing a globe of any chosen size in the flight path of the incoming missile. It does double duty because the missile hits it going in, and then the rubble hits it again going out the other side.

While the U.S. persists in its useless boondoggle "missile defense system" by shooting a missile with a missile (missiles which would already have been "dudded" by the Tesla howitzers), between the domes and the globes the Russians (KGB) have in hand (**operational**) a nearly **perfect missile defense system**. Someone should ask the Secretary of Defense if the U.S. Tesla domes are in place. And over which cities? How powerful are our American Tesla domes?

Bearden cites a number of examples of these domes being tested around the world:
March 20, 1969 | March 24, 1977 | August 17, 1980 | June 17, 1966 | June 18, 1982 June 22, 1976

In a slide from the Fer-de-Lance briefing it is shown how the "woodpecker grid" itself also acts as a sensor, detecting any missile launch anywhere, and how the howitzers can respond instantly by delivering a blast to exactly the right coordinates, right along the channels of that same grid which is feeding the location information about the missile launch. This certainly beats "trying to hit a bullet with a bullet." Especially since YOUR bullet would probably be instantly inoperable in a real scalar war situation.

3. Rogue Groups in the government

Bearden fears that these scalar electromagnetic secrets may have fallen into the hands of what he called secret "**rogue groups**" within the labyrinthine depths of the black-ops special-access domain of the secret government. "In the extreme, such a deep black program can even become a "captured" program," he says, "which is totally in the hands of rogues and **no longer reports to or is bound by the dictates of the parent government**."

> "They have their own agendas. And being rogue groups, they may well bring in **unethical, immoral tricks**: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. **It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification. ...**
>
> "In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old 'Power corrupts, and absolute power corrupts absolutely' routine. ... Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, **may not operate under very much legitimate government control at all**, and may operate specifically for the purposes of the rogue group or groups that have gained control."

> "The involvement of at least some rogue groups, some being "**cowboys**" who operate well **outside all laws and ethics**, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, ***it could account for what seems to be actually occurring***. "
>
> http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

4. World War III has already begun

Because it is now possible to cripple the power grids and electronics of any nation, that nation could be reduced to economic ruin without ever firing an actual "shot." Without electricity Wall Street cannot operate. Without electricity you quickly need martial law. One no longer needs to wage a conventional war anymore at all. There are now **lots of new ways to bring any nation to its knees**.

> "We are at war, whether we like it or not. And we are going to have strategic strikes on an unparalleled scale for us. Pearl Harbor was nothing compared to what is coming down, once we consider the "augmentation" and such.
>
> "When that factor is considered, then it's a totally new ball game on replacing the power system. There isn't going to be much choice; the old one is going to be destroyed or so badly crippled that **it will collapse the U.S. economy** in a way we have not seen since **1929**. Of course, in such a case hitting us, we will be under martial law, the works, everywhere. For one thing, it's the only way you can keep the cities from imploding from within, by the criminal element, looting, robbing, killing, etc. Simply wait and see; it's on the way."

Although Bearden has for years decried the "scalar-gap," and chided the military for falling so far behind the Russian KGB in the field of scalar electromagnetics, recent statements indicate that things may at last have changed somewhat.

> "In the new war, the first phase of WW III has already been completed. Hence President Bush's recognition that this is a total war, and has to be taken to the enemy no matter where he may be hiding or in what sanctuary or in what foreign nation. This is an entirely different war than anything the U.S. has ever fought in its history. **And before it is over, it is going to kill more Americans than have died in all the previous wars in our history. The news media and the public have not yet recognized that awful situation and predicament.** Most are already thinking that, well, we won in Afghanistan so it's just about over. Instead, it has only just begun.
>
> "Anyway, that's a brief uptake on some of the weapon situation today. Some 10 nations or more now have LWIs, and five have the QPs. I'm still working on how many have the appalling negative EMP weapons, but would estimate 6 or 7 nations have them in one stage or the other, either deployed or in at least advanced development. And in my estimation, the Yakuza will have them within three to five years if they have to build them themselves, and quicker if they can buy them from the Russians, Chinese, or???" Bearden

When he was asked the question "**Does the government listen to Tom Bearden?**" Bearden said:

> "Let me speak carefully. I am very satisfied with how I've been listened to in the quarters where the rubber meets the road. **The West is no longer defenseless, not by any means.** Cannot amplify it further than that."
> http://www.cheniere.org/correspondence/021102a.htm

Bearden has briefed the government a number of times and the Cheniere website contains a great number of his graphic military briefing slides.

5. Are the Russians making "UFOs?"

Scalar electromagnetics engineering opens the possibility of antigravity vehicles and "hyperspace travel." In other words we can now make our own UFO's! Of such **scalar-powered antigravity vehicles**, which have now become possible, Bearden has this to say:

> "We might expect to see such vehicles glowing. Their various surface features and mechanisms might appear to be glowing or revolving lights, etc.
>
> "They could exhibit incredible "aerodynamic performance," seemingly in the atmosphere. **Actually they would not be moving "through" the atmosphere at all, but through a higher space outside each particle of atmospher**e.
>
> "They could seem to **materialize** and **dematerialize**.
>
> "They could seem to plunge into the ocean or rise out of it.
>
> "They could even seem to operate under the ocean or inside the earth itself.
>
> "**Such anomalous vehicle performances have been seen all over the world**, particularly since a few years after WWII.
>
> "It appears that one or more nations of Earth are operating such vehicles now **in great secrecy**. Our own "government" may even be one of these.
>
> "If so, it **isn't** our true "government/government." Instead, it's the "control group/government." It's government at the operational level, but at higher level **it belongs to certain control groups who have penetrated our government's vitals and taken over all such projects**."
>
> Bearden http://www.cheniere.org/books/ferdelance/slide32.htm

In an excellent article on her Earthfiles website, Linda Moulton Howe reports on the utter bafflement of Norwegian scientists over mysterious lights appearing in the valley of Hessdalen, Norway.

She writes of the research there: "The results can be broken down into two groups: 95% are thermal plasmas and 5% are unidentified solid objects. The plasmas emit long wave radio frequencies and strangely, their temperatures do not vary with change in size or brightness." She quotes the scientists' research summary:

> "1) most of the luminous phenomenon is a thermal plasma;
> 2) the light-balls are not single objects but are constituted of many small components which are vibrating around a common barycenter;
> 3) the light-balls are able to eject smaller light-balls;
> 4) the light-balls change shape all the time;
> 5) the luminosity increase of the light balls is due to the increase of the radiating area. **But the cause, and the physical mechanism with which radiation is emitted, is currently unknown**."

This would almost certainly seem to be another example of the testing of the scalar howitzers. It is baffling because, as Bearden points out repeatedly, Western scientists have limited knowledge of the scalar electromagnetic principles by which the howitzers operate. The Hessdalen scientists were completely **baffled** as to where the energy of these plasma balls **was coming from**.

Ms. Howe interviews Massimo Teodorani, Ph.D., Astrophysicist, who released the report on the 2001 Hessdalen sightings. The scientist speaks of the baffling phenomon observed.

> "I don't know how it is possible that Nature is spontaneously able to do that. Anyway, we deduce that the plasma is trapped inside a sort of magnetic cage and the magnetic cage closes around the plasma and keeps it fixed in some way, prevents it from expanding. **But where does it come from? We don't know**."

By what we have been considering here, we can recognize the signatures of scalar electromagnetic engineering. If it is a Tesla howitzer, then the energy for the plasma balls **is coming from the vacuum of spacetime** at the very **location of the balls themselves**, triggered by scalar interferometers aimed through the woodpecker grid.

These kind of balls can be used as **marker beacons** giving feedback for precision aiming of the howitzers. The energy of the marker beacon can be read back into the computers giving precise location information for pinpoint aiming. The target area can be very small, or widened out.

There is a little movie of one of the Hessdalen plasma balls here. It would seem to be a clear documentation of a scalar howitzer marker-beacon. How many other current "UFO" sightings are actually the witnessing of Tesla howitzer tests?

In fact there are a great many **anomalous events** of the past few decades which might very well be explained by the Russian testing of their longitudinal electromagnetic weapons. It would seem that the mysterious appearance of crop circles around the world could be accomplished by feeding a precise mathematical graphic pattern into the computerized aiming software and change the very molecular structure of the crops themselves along the lines of that pattern. The stalks fall over from the localized effect of the longitudinal waves. A plasma ball tracing patterns? Are they slyly sending a message to see if we are "getting it" yet?

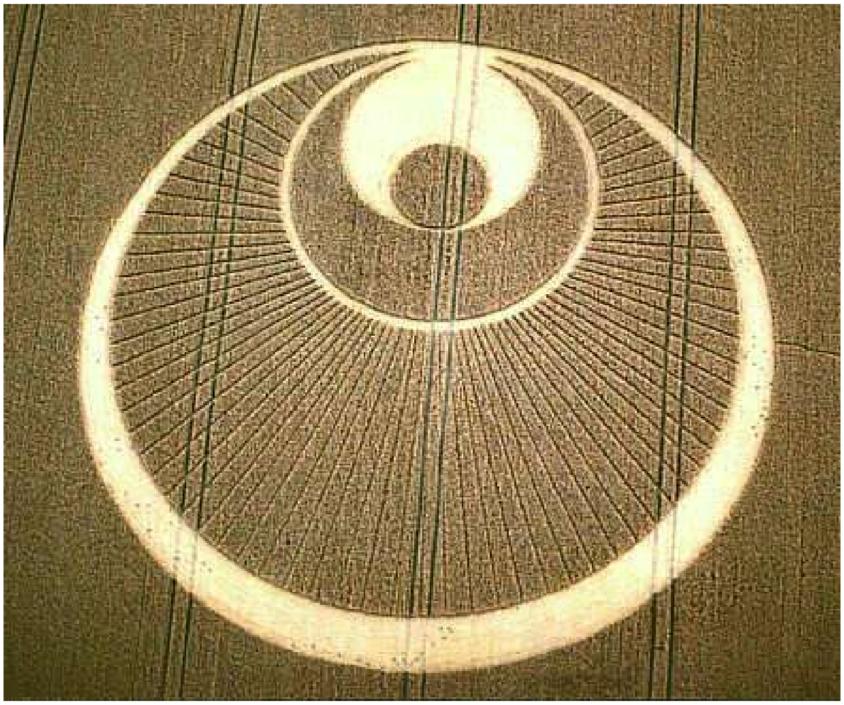

The "Cambridge Angel" - Made by Russian scalar plasma balls?

Also see: Interview with Tom Bearden

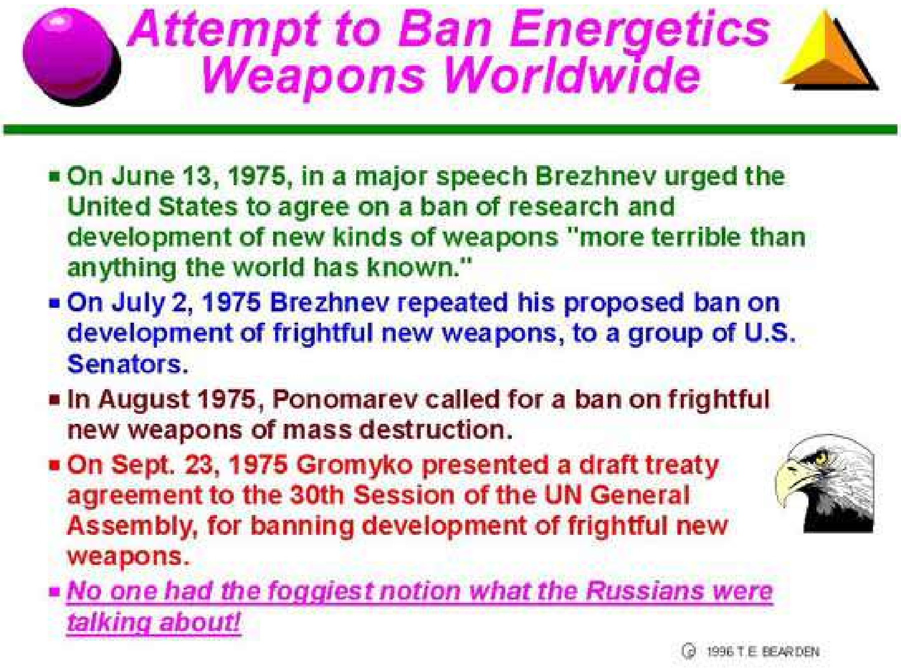

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T.E. BEARDEN

(Slide from Cheniere.org)

4. Healing powers of longitudinal waves

Healing via longitudinal waves, time reversal of cells (de-differentiation), Priore's machine, Bearden's proposed "blanket" version, using the electric power grid for healing waves or for broadcasting disease patterns. Scalar Biowar.

The good news!

"**Any sort of disease whatsoever - physical or mental** - will be rapidly curable, simply and cheaply. It will be possible to reverse aging and rejuvenate the person. It will be possible to regrow limbs and straighten misshapen spines - and directly remove the causes of mental diseases and cure them as well. Lifespan will be staggeringly increased, without 'old age's debilitation.' AIDS, cancer, leukemia, and genetic diseases will be completely conquered. The potential for a true golden era is upon us, for all humanity."

Bearden- http://twm.co.nz/beard_interview.htm

Proof of the healing powers of longitudinal waves from the vacuum is seen in the work of Antoine Priore who produced longitudinal scalar waves using a **gigantic glass tube** filled with plasma. He himself did not understand why it worked or how it cured cancer in animals.

> "The Priore team cured terminal tumors, trypanosomias, and other dread diseases in laboratory animals in France, in the 1960s and 1970s, under rigorous scientific protocols by eminent scientists. The Priore

> project was funded by the French Government to the tune of several millions of dollars (equivalent). http://www.explorepub.com/articles/energetics.html
>
> ". . . Prioré's machines concretely demonstrated a nearly 100% cure of all kinds of terminal cancers and leukemias, in thousands of rigorous laboratory tests with animals. These results were shown to medical scientists as early as 1960." http://www.cheniere.org/priore/index.html

Bearden has stated that the longitudinal EM waves can have a healing effect because they always come in pairs called "**phase conjugate pairs**" and one of the pair is **time-reversed**. This enables engineering of the waves to "pump" the cells of the body (with waves from the time domain) and actually **time-reverse the cells** back to a previous healthy condition. The scalar cure of cancer does not involve killing the cancer cells but rather pumping them with time energy into a previous condition. I know, this seems miraculous, and is it not?

Priore and his giant plasma tube

This healing power in the longitudinal waves of the time domain is a mind-boggling thing to face. It took me months to get through it, and to **suffer the fact** that the **cure** for cancer and AIDS was right there, **in this new scalar electromagnetics technology**, and that it has been kept back from the people for so many decades now.

Bearden has a great deal to say about the work of Antoine Prioré, who himself did not understand why his machine worked. Priore did not know about the longitudinal scalar waves of the time domain. In fact, Tom Bearden has supplied, at last, the missing explanation as to how that **big giant two-story glass tube** consistently cured cancer in animals.

> "**The medical implications are enormous**. By just reversing the damaged cells back to a previous physical state, this gives a physics mechanism for *controlled cellular dedifferentiation*, in biology terms. Physicians will **time-reverse diseased, damaged or aged cells** back to an earlier healthy condition, including all the cellular genetics. Normal cells are just reversed to a slightly younger condition."
>
> "We believe this mechanism explains the revolutionary but baffling electromagnetic cures achieved by a team of outstanding French scientists in the 1960s and early 70s, working with inventor **Antoine Prioré**. In lab animals, cures were achieved for terminal tumors, infectious diseases such as trypanosomiasis, and atheriosclerosis. In addition, deliberately suppressed immune systems were dramatically restored.

> "We hypothesize that this is the long-sought mechanism of how the living body heals itself. In short, the body's cellular control system induces **cellular reversal by weak time-domain pumping** of the damaged cells and the affected area. The cellular regeneration system performs all healing of cellular damage in the body; the immune system heals nothing, not even its own damaged cells) .
>
> "For the first time medical science can develop as a science of *healing* rather than a science of *intervention*. And it will develop as a science of *unlimited* healing, since **no pathogen can resist action by curved spacetime engines**." Bearden

http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

See Priore's machine See Slides of how it works.

Bearden has extended the Priore work, with the previously missing knowledge now in place, and designed a revised system (for irradiating the body with longitudinal waves) which gets rid of Priore's giant glass plasma tube and uses instead an **antenna unit which surrounds the body**. He has also provided the missing explanation of how it works. There is an article from "Explore" magazine, Part I and Part II.

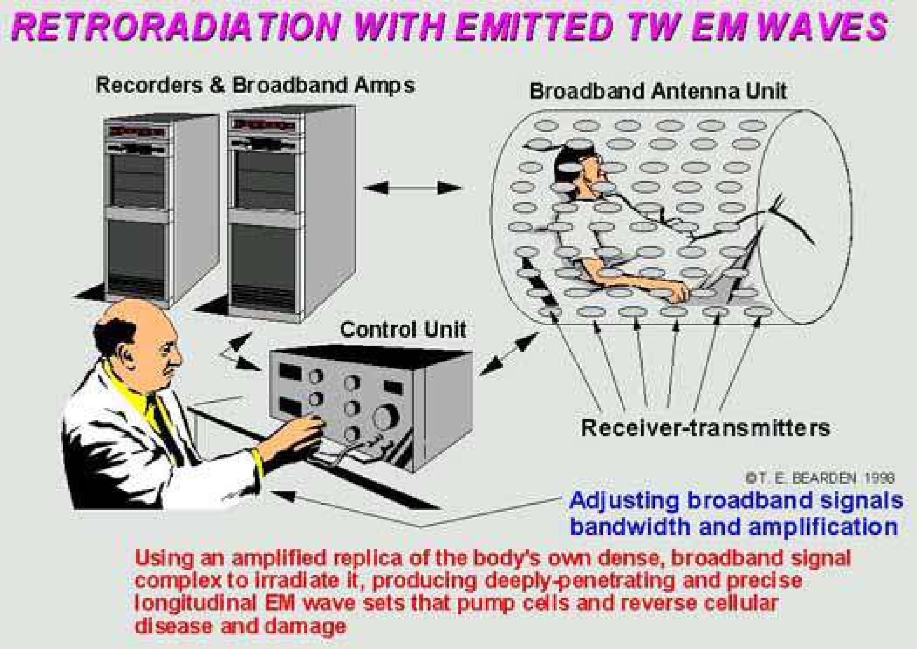
RETRORADIATION WITH EMITTED TW EM WAVES

Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
©T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification

Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

(Slide from Cheniere.org)

In fact, Bearden has presented to the U.S. government a plan to use a **portable version** of a **Scalar Healing Blanket** where wire channels in the blanket act as a scalar antenna, which can both send and receive scalar electromagnetic waves. A **two minute treatment** of the waves is sufficient to restore the body to its earlier healthy condition. Thus even **thousands could be treated easily and quickly in a biowar mass attack** situation.

> "To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S.

> Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines **could be developed cheaply and quickly**, and present government facilities could be used to massively attack the correlate database production problem.
>
> "The portable treatment **units could be cranked out by the hundreds of thousands** and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals, emergency response teams, etc." - Bearden

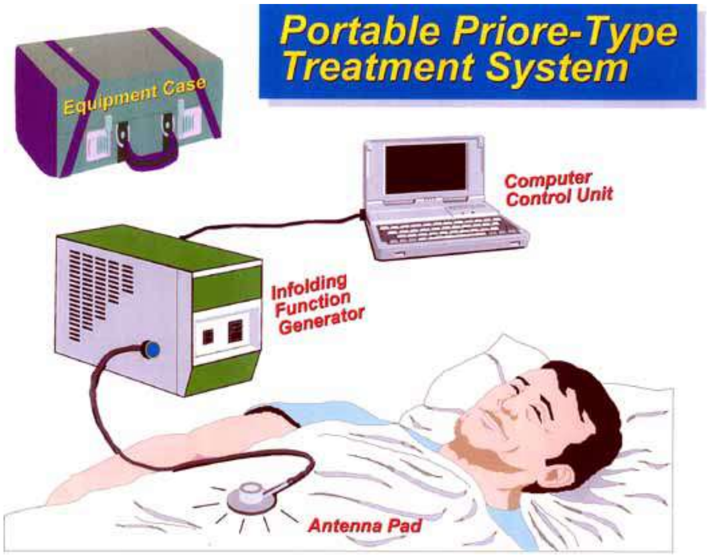
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

(Slide from Cheniere.org)

The **Scalar Healing Blanket** would fit into large suitcase. It has three basic parts, the longitudinal wave generator, a laptop computer, and the antenna-blanket. The plan was a stroke of Bearden's genius, but unfortunately it fell on **befuddled government ears**. "Unfortunately, so far the DOD [Dept. of Defense] doesn't even understand what the heck we're talking about!"

Article "Healing with Energy": http://www.explorepub.com/articles/energetics.html

> "We tried very hard in 1998 to call attention to the extreme national need for the most urgent possible development of a portable computer-controlled treatment device based on the experimentally proven Prioré methodology, but extended to a much simpler and far more rapid method of treatment.
>
> "Quick development of this technology could lead to **a suitcase size device, capable of treating a patient in less than one minute, and stopping the symptoms and disease progress cold, then reversing it**. Three "less than a minute" treatments one week apart would be required for complete cure. As an added advantage, partial reversal of aging in older patients would also be accomplished as a bonus, in addition to time-reversing the damaged and diseased cells...
>
> "The method proposed to the DoD in 1998 used normal EM in a peculiar way to force the body itself to make its own time-polarized

> EM waves, and pump every part of its cells -- including the genetics -- in the time domain.
>
> "Quite simply, we are going to be struck this way (QP spreading the immune systems of the populace plus professional smallpox, anthrax, etc. BW attacks on our population centers) and by other means we've spoken of in other papers and briefings."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

The Dark Side of the Force:
Creating and broadcasting disease

Speaking of the scalar wave induction of disease by longitudinal wave patterns Bearden refers to work by French scientist Kervran and Russian scientist Kaznacheyev, who proved that any disease could be transmitted electromagnetically.

> "It works this way: the Kaznacheyev experiments (and others) demonstrated that any cellular disease or disorder can be initiated into cells at a distance, by electromagnetic means (albeit unusual EM in nature). The decades of microwave radiation of the U.S. Embassy in Moscow, was an ongoing set of "stimuli" to ping the U.S. system and, by its response, ascertain how far along (or if it was "along") in its knowledge of such weapons. The radiation was responsible for many health changes in personnel and for the **eventual deaths of three U.S. Ambassadors**." Bearden http://www.cheniere.org/correspondence/021702a.htm

In a letter he writes,

"Yes, the Kaznacheyev experiments were quite real; 15,000 or so in military research institutes in Siberia. As with so many other novel areas investigated and developed by the Soviets, most U.S. investigators just assumed normal EM, which is not what the Kaznacheyev experiments involved, and not what was carrying disease induction patterns from one cell culture to another in rigorous lab tests."

How to "Broadcast" Diseases
Longitudinal EM Biowars

Bearden explains how the new quantum potential weapons can be used to induce disease-at-a-distance in a population, or to "spread" the immune system so thin that a **conventional bio-attack would be greatly potentiated**. By "broadcasting" virtual disease patterns over a population, the immune system is so overloaded by the presence of the **"shadow" diseases** that its effectiveness is thinned out.

> "In short, alter the internal wavestructure, and one creates a curved spacetime 'engine' that acts on mass in any fashion one chooses to design -- including initiate diseases . . . Now visualize one of these '**vacuum engines**' or 'spacetime curvature engines' that **acts on mass** to **generate the exact effects produced by anthrax**. (Or any other disease one wishes) . . .

Call the virtual state of a disease engine the disease pattern in the "shadow" state, just prior to becoming observable. . . .

"And the master cellular control system responds to "coming events that cast their shadows before". In short, it responds to the upper level of the **"shadow state" of a disease**, still in the virtual state! (Hey, virtual particles are real, virtual photons are real, virtual ST curvatures are real --- just very fleeting, only for another to immediately arise). The exchange of virtual particles is known to generate all forces in physics anyway! . . .

"So the cellular control system responds to "shadow state" disease patterns. Enter a QP weapon application. Place a quantum potential weapon "area" on the United States, so that the entire populace is in it. The slowly introduce and bring up in the shadow state, the **necrotizing fasciitis disease engine**. At some point, the cellular control systems will react, and order the immune system into action. In other words, immune system resources will be committed against this "immediately coming" enemy. Bring it on up a little, but still just in the shadow state. Voila! Humans form a bell-shaped distribution curve. And even if the structure of the QP fluctuates a little, fluctuating a bit stronger and a bit weaker, etc. **A few of those targeted bodies will now actually develop necrotizing fasciitis**. So far, just like the U.S. Embassy targeting, except a different "carrier".

"But now there is an exact signature that this is what is doing it, and not normal vector carriers of disease and contact or exposure infection. **The few cases of necrotizing fasciitis that break out will be most puzzling.** They will be randomly scattered across the entire nation, and so few that the vector carrier and contact or exposure explanation completely fails."

"And that one happened, right here in the U.S., not so long ago. Check it out.

"So carefully adjust the **patterns** for, say, 12 major terrible diseases, to that "adjusted and desired shadow state level". Now add all 12 patterns into the internal structure of the quantum potential. Now it has **12 shadow state disease engines** in it.

"**Place that one on a populace**. The cellular control systems order the immune systems into action, and they now have to commit their finite resources to 12 different areas. This greatly thins the response an immune system can make to any one of those 12 disease patterns! In short, it "**spreads**" the capabilities of the immune system, which can be galvanized to great effort, but only can perform a small bit against each area.

"Suppose one of those "galvanized" areas in the targeted populace's immune systems is anthrax. And suppose the bad guys now hit one or more of those cities with a professional anthrax attack. A former OTA study showed that a professional anthrax aerial spray attack against Washington D.C., dispensing

> 100 kilograms of anthrax spores in the spray, will account for 1 to 3 million casualties. And that's in an "unthinned" set of immune systems. **In the "thinned" set, one can at least double or triple, and even up to five times, that estimate.** So conservatively, now there would be from 2 to 6 million casualties. And even mild anthrax spray attacks in other "thinned" cities would also be amplified.
>
> "**That one is being set up, or already set up, right now.** There are other such "augmentation" setups also under way.
>
> **"So the weapons alluded to by the SecDef in 1997 are already being employed. WW III has already begun, and our populace (and our scientific community) is blissfully unaware of it."**
>
> http://www.cheniere.org/correspondence/021702a.htm

Most of us can remember the strange outbreaks of the so-called "**flesh eating disease**." It seemed to break out here and there randomly with no known cause. Bearden believes this was a test of the newer quantum potential weapons which go even beyond the capabilities of the Tesla howitzers.

> "The Kaznacheyev "disease induction by novel electrodynamics" work is also the basis for the spacetime curvature engines (not EM signals!) carried by quantum potential EM biological weapons for distant induction of diseases etc. into the population of an entire nation or area, such as the U.S.
>
> "This **quantum potential disease induction weapon** -- capable of attacking an entire population of a nation or area -- has been tested several times in the U.S. at very low levels, in the shadow state only (we explain shortly) and it has a most peculiar signature. E.g., **one test used necrotizing fasciitis (flesh-rotting disease) engines**.
>
> "The disease 'engine' was deliberately kept down in the 'shadow state' (virtual state, just below the observable state) so that only a tiny bit of the populace with depressed thresholds would "breach the observable threshold' due to the population forming a bell-shaped curve as to the threshold level. **The major signature of the test was that the cases of the disease broke out totally statistically and randoml**y, spread all over, without any 'disease vector' in between.
>
> "In short, it was not spread in any manner required by normal deterministic disease vectoring, **but was a totally random set of occurrences**. Several other diseases were also tested this way in the American populace, with precisely the same signature."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm
>
> "Apparently one planned use of such an augmented clandestine BW weapon will be smallpox, e.g. As you know, the Russians

> did develop smallpox for biological warfare, and even developed a new strain against which previous vaccination is ineffective.
>
> "The vast numbers of expected casualties, coupled with immune system spreading, are absolutely irresistible to the Russian mind. You can count on it that they have seen that the smallpox is already on site here, in the U.S. and in the hands of professional terrorist teams. Of course, the Russians/KGB will protest they have nothing to do with that! Yet likely the KGB will have flatly arranged it or at least assisted it. The capability for mass destruction of the nation is so lucrative and easy to achieve in advance, that **they would never be able to resist doing it**."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

Bearden believes the **Gulf War Syndrome** involved the use of scalar technology to induce a disease state, as he stated in a 1997 interview:

> **"I'll say this flatly, and I won't answer any further questions about it; the Gulf War Syndrome was induced...** It was induced with this stuff. All the rest... was contributing factors that everybody's talking about... There are some deeper signatures, if you look into it very deeply, that show you exactly how it was done and the fact that it was induced. It was a test. It was a test of a very special kind of weapon I have not talked about yet. Anyway you can use it to create diseases."
>
> Bearden http://twm.co.nz/beard_weapon.htm

Besides using the woodpecker grid to target a population, the entire **electrical grids** of nations can be used as "channels" for disease patterns to be broadcast to all people near that grid. Your own house wiring can become an antenna, broadcasting these "shadow" patterns of disease, weakening your immune system.

And here one faces the frustrating fact that the same power grids could be set to channel patterns which would keep the entire population healthy, rejuvenated, continuously bathing in the time-reversing old-age-defying energy of the spacetime vacuum. We have **no idea yet** of how long the human life span could be increased with this new technology. People might be able to live two centuries given full humane development of this technology.

5. Psychoenergetics

Mental entrainment, inducing hypnogogic trance, mind control, broadcasting thoughts and feelings, the Ant World, education via EM [LW] mental download, the "Feelies," revolution in spirituality, "enlightenment machine?" Scalar churches?

Psychoenergetics is simply using the scalar interferometers to manipulate and **engineer the human psyche**. To me it is the most **mind-boggling and frightening** aspect of longitudinal wave engineering.

The mind is electromagnetic in nature, and itself of the time-domain. It is not observable in 3-space. There is no **thing** you can point to that is the mind. As Bearden says, "**The mind is time-like**."

The longitudinal scalar EM waves are also of the time-domain. By overlaying certain infolded patterns (oscillations in time) on the longitudinal waves, **a thought or feeling may be made to arise in a person** or persons who are in the interference zone (where the wave beams cross).

The person will not notice anything, feeling that this thought or idea is his or her own idea. The longitudinal wave pattern might be a wave of panic or fear that spontaneously sweeps over you, perhaps inexplicably. It could even be in the form of a sudden intense patriotism. It could be explicit words and phrases, which everyone would think they had thought of by themselves. Or it could be rage and hatred, angry and violent. On the other hand It could also be a continuous sense of docility and placidity. This is the brave new world of psychoenergetics and the coming "**Mind Wars**."

Scalar Psychoenergetics in its most **primitive form** simply "entrains" all minds in the target area into a **deep hypnogogic trance**. In this state of mind people would be suddenly extremely suggestible, and would likely believe anything they are told, and would obey any orders given. Talk about winning the hearts and minds of the American People!

Tom Bearden makes me think of **Paul Revere**, who heroically conveyed important **warnings** for the sake of the new nation. But the words of the famous phrase have now strangely and curiously mutated into the once-inconceivable: "**One if by land, Two if by mind.**"

Indeed one strategy in a **Mind War** would be to simply **take over the minds of the "leaders"** of the enemy nation. The targeted leader would not be aware that anything was amiss, although he might begin to make unexpected changes in policy.

And are, in fact, any of our "leaders" already having secret thoughts that are not their own? Who knows? Frighteningly, we can no longer be sure. The new sciences of scalar electromagnetics and psychoenergetics are even now plunging humanity into a sudden science-fiction-like world beyond anything in humanity's previous imaginings.

> "The Russians reached the point in the mid 1990s that they could **take over control of a person's mind**, with modified longitudinal EM waves including some time-polarized EM waves, and with a team of specialists (estimate 25-30 per transmitter, and one transmitter and team per controlled person)." Bearden

In a slide Bearden illustrates how a Scalar War would involve a **psychoenergetics attack** on the operators of the enemy scalar installation, entraining their minds into hypnogogic trance and **getting them to shut down their systems**.

And ultimately psychoenergetic warfare goes to the very heart of **human identity itself**. For if my thoughts might no longer be "mine," then **who and what am I?** Can my very sense of being "me" be hijacked by some nefarious psychoenergetic scheme? Will the "secret government" eventually dictate directly into your mind how you feel about yourself? Or what you think you are? Or what you should do?

How Does Longitudinal EM Mind Control Work?

The physics of scalar psychoenergetics as expounded by Col. Bearden seems even more arcane and over-my-head than that of extracting energy from the vacuum. I can't understand it, but I look through it anyway. Each time another tiny understanding might occur, or a term might fall into place. I know the mathematics is forever beyond me.

One paper to begin with is Mind Control and EM Wave Polarization Transductions This is such serious stuff that Bearden includes a **strong warning about misuse of this knowledge**. Psychoenergetics weapons can mentally maim and physically kill.

Warning!

(From "Mind Control and EM Wave Polarization Transductions")

> "This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. **We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral**." Bearden

After the **serious warning** comes the description of the paper. This is heady stuff for the layman.

> "Calling full attention to the special note above, in this paper we present a high-level overview of the novel electromagnetic nature of mind operations, mind and body coupling, and intent - the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transduction mechanisms whereby differing EM wave polarizations can be transformed one-into-the-other are presented." Bearden

I will leave it to the technically astute to try to understand the mechanisms of operation which are described in scientific language. But there are tidbits for the layman. Here he speaks of the difficulty Western scientists have in opening up to the "immaterial" (longitudinal wave) nature of the human mind.

Western Science Remains Largely Materialistic

"Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system. This serious self-limitation exists because in the body we measure only weak TW [transverse wave] EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.

"With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.

"Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW [longitudinal wave] functions exist in the body and in all of nature in general." Bearden

Western methods of influencing the mind with EM waves have only used transverse waves, not the longitudinal waves of the vacuum. If only transverse waves are used one pretty well has to hit the target mind with a **sledgehammer of waves**.

Western Science knows only transverse EM waves

Brute Forcing Time Functions versus Fine Control Methods

"Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.

"While KGB scientists also use TW EM "brute force" TW waves when necessary, they do **"imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures** required to directly perform the mind engineering desired... Bearden

"These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design for example of the necessary time-polarized wave assemblies for **engineering the entire human collective unconscious simultaneously**, or for engineering the entire collective unconscious of all species on Earth (i.e., Gaia's collective unconscious), or even for precisely **engineering the memory and knowledge base of an individual**. Bearden

The seemingly magical operations of psychoenergetics, sometimes called "psychotronics," cannot be achieved with ordinary transverse EM wave energy. But they are possible with longitudinal scalar waves because the mind itself is scalar in nature.

> "We note that **all mind operations are time-like**, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain..." Bearden

The engineering of the mind can only be accomplished with the waves of time.

The course of psychoenergetics development necessitates the building of a vast database of settings and patterns which bring about given mental and/or emotional states. A kind of "**genome**" of the psyche. Bearden describes the process of the decoding of this "genome."

> "Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.
>
> "In these experiments, the experimenters will eventually be able to **provoke any body or mind change they wish**. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc.changes and interactions." Bearden
>
> "A second stage in the research would be to **test the correlates** and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment." Bearden

Summing up his paper Bearden speculates on what the actual situation the science of psychoenergetics probably is. The intense secrecy with which this knowledge has been kept from the public has led to the **current farce of the "government" pretending** it all does not exist, and this secrecy may have now put us at great risk.

> "The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, **may not operate under very much legitimate government control at all**, and may operate specifically for the purposes of the rogue group or groups that have gained control.

> "In a nutshell, that's what may be going on in the **clandestine mind control projects** in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well **outside all laws and ethics**, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, ***it could account for what seems to be actually occurring***. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians..."
>
> "Remember that, in every large and powerful human organization, the basis for rogue groups is **power** and **secrecy**. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a **human** characteristic, the **old primate dominance game**. Only now disguised and hidden under deep classification."
>
> "The **Russians**, with their additional knowledge of the actual mechanisms in the transforms, will be **much more advanced than the West,** because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U(l) electrodynamics. Further, the Russians have decades of use of **longitudinal interferometry beams to reach right through the earth and ocean** and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW [longitudinal wave] interferometers, through intervening mass." Bearden

Bearden cites two examples where he believes the Russian KGB tested the psychoenergetics capabilities of their scalar interferometers. One was the case of Captain Button in his A-10 "Warthog" aircraft on April 2, 1997.

> "Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona."
>
> "At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—**dream-thinking until he died that everything was normal**. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. **So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour**.

> "Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks."

The paper Mind Control and EM Wave Polarization Transductions ends with the most chilling and frightening possibility of all: the **psychoenergetic engineering of the human species** as a whole.

The Ant World

The Final Race for Direct Mind Control of the Entire Human Species

"The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs [longitudinal waves] for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. **The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials**.

"The KGB scientists also know that something very like **Gaia** — a collective unconscious mind for all species on earth — **also exists**, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials."

"Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.

"I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can then **convert the human species easily into a sort of "ant" society**, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society" — communist style — is still a variation of feudalism (**all systems tend to feudalism or some form thereof**).

"The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now...

"Some versions of Russian mind control devices were used on Russian soldiers in the Afghan War to **condition them for performance of merciless acts**." [13]

http://www.cheniere.org/explore%20articles/mind%20control3/p07.htm

More about "The Ant World"

Whether it is the Russians planning an "ant world" society or rogue U.S. black-ops groups withholding a great boon from human race, the secrecy on all sides around has allowed **largely criminal elements** to take over this technology while leaving the civilized elements in the dark. Openness and the spreading of this knowledge among good and decent people is essential, and for that **we will have to overcome our frightened denial**.

> "Meanwhile, **rogue groups** amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, **unless great care is exercised by the oversight committees in the House and the Senate**.
>
> "Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed." Bearden

Like the huge energy of the Tesla howitzers, the ability to engineer the mind with longitudinal waveforms is a seeming "magical" thing which can be used for good or ill. Bearden envisions that further developments will allow something like the "downloading" of knowledge via scalar interferometry. Everyone could be highly educated, and quickly and easily.

In the conclusion of the paper "Mind Control and EM Wave Polarization Transductions" Bearden makes **a plea to the nations of the world** regarding the use of psychoenergetic engineering.

> "Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction helping and healing people, rather than killing or abusing them. **The excesses in its bad use are a potent threat to all nations on earth**.
>
> "Yet it can revolutionize medical science, education, communication, and psychology. We foresee the day -- perhaps 30 years hence -- when **education will be accomplished by directly loading the software into the mind**. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks on will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an M.D. And so on. When that happens then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.
>
> "We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium." Bearden

(Perhaps after so many frightening topics a slight break is in order here to consider the **lighter side of psychoenergetics**. For example, what will be the ultimate impact on the entertainment industry?)

The Feelies at Last!

Psychoenergetics will be able to bring about at last, for better or for worse, Aldous Huxley's fanciful idea in his novel "**Brave New World**," an entertainment format which he called "**The Feelies**." Those were something you went to, like going to the movies, but you would **experience** all the actual feelings themselves, as if the depicted events were actually happening to you. Although our "leaders" seem too dim to grasp what is going on in the world of scalar electromagnetics, it should not take some young Hollywood genius too long to understand that **The Feelies are now possible**, and that the immense megabucks of potential profit more than justify heavy investment in bringing it to market.

One can imagine that first "Feelie" theater, "The Huxley" perhaps, a large comfortable area surrounded by nine giant screens (not just one screen!), which become 3D if you put on your headset. And The Huxley is fitted with the finest sound, and oh, by the way, a modest sized scalar interferometer with computer. And there are one or more **new tracks** now recorded on the film next to the sound track. These are the digital patterns of thoughts, feelings, and ideas to be fed into the theater's interferometer. And these thoughts and feelings and ideas will be felt by the audience as if they were their own thoughts and ideas. It is an entirely new art form. Psychevision.

As important (and no doubt famous) as the actors would be the "feelers," those from whose minds the feelings were lifted in the original programming of the Feelie tracks. Because, for example, Brad Pitt might be a good and handsome actor, but **can he really feel it?** Like the dubbing of foreign language films, it might be necessary to dub the "emotional track" using a deeper and more profound person's feelings, thoughts and ideas.

It might even be engineered so that if you sat on one side of the theater you would feel like the hero, and on the other side you would feel exactly like the bad guy! What would that lead to?

We have all cried at a movie some time, most people have, though they might not want to admit it. But this would be **deep sobbing** and a **flood of tears**, if the director so desired it. You would feel it at your core. Or, take a big chase scene, **you would be terrified**. They are breathing down your neck! You would be gripped by panic.

For that matter, the whole genre of **Horror-Feelie** would be too much for many people to take. Do I **really** want to feel that an **actual vampire** is tearing at my throat with his fangs?

A Religious Feelie would make you feel the exaltation of a saint at his prayers, or give you the **definite** feeling that you are Christ suffering on the cross. All the pain could be there too. The stickiness of the blood. The very feeling itself: "Why hast thou forsaken me?" Who among us could come up **that** despair for scanning? Had this been developed a bit earlier we might watch a documentary of, say, Mother Theresa, containing her own actual deep sense of compassion, which we would feel arising in us as our own.

Then again, a Psychedelic Feelie could induce an LSD-like state and simply blow the audience's minds.

Even in its simplest most primitive mode, the hypnogogic trance induction with its increased suggestibility, would completely alter the experience of even ordinary current movie going.

6. As it Stands - 2002

[Note 2006: Click here for a current status history of the MEG from Tom Bearden]

> "**Anyway, it's presently a "Mexican stand-off**" with the various parties maneuvering with check and countercheck. That is the real "balance of terror" that is held by such a fragile balance."
>
> "You can see the implications if such weapons fall into the hands of the radical terrorists. Ordinary positive energy EMP shooters **can be made fairly readily and cheaply** from nearly just **Radio Shack parts** or surplus parts. It just takes some know-how... When and if the Yakuza [Japanese mafia] gets those (and they will, it's just a question of time), you can begin to appreciate some of the turns this present terrorism and our war on it will take." Bearden

The very thought of the building of one of these howitzers by "some lone-nut gunman" from Radio Shack parts makes one realize how serious it is for humanity to face the crisis it has come to.

Once again, as at the dawn of the nuclear era, humanity finds itself at the crossroads of Heaven and Hell. As a species, we must stop and consider. And there is at least one real difference between this visit to the crossroads and the last: *this* time there is a very real possibility of a true "earthly heaven" on the Heaven side.

The poor nations could come to a good life. Disease could be largely eliminated. Fossil fuel pollution could end. Global warming could be solved, and the weather could be managed in a humane way, bringing water where there is drought, and sunny skies where there is flood. Scalar interferometry would be there to avert any possible hits by asteroids or comets. Mars could be colonized and terraformed. Everyone would be highly educated through psychoenergetic downloading of knowledge. And the **wisdom of the wise** could be realized directly through scalar connection to their minds and hearts.

These stunning sudden possibilities give me pause to think: if our species has reached this crossroads in such a short span of centuries, then surely other species in this vast universe have passed this way before and long ago. And of those who chose the path to "planetary heaven," what is their life like now? It might be good if some of them were with us now, to guide us through this last struggle over this pesky old good-and-evil problem. **It is our last chance to choose the path which benefits the whole species, and not just some elite group or faction**.

The Disclosure Project

One person working tirelessly to bring to bring these technologies into the open is Dr. Stephen Greer of the Disclosure Project, which is calling for congressional hearings into the issues of UFOs and the free energy technology that was, perhaps in part, reverse-engineered from them.

> "Once abundant and nearly free energy is available in impoverished areas for agriculture, transportation, construction, manufacturing and electrification, **there is no limit to what humanity can achieve**. It is ridiculous — obscene even — that mind-boggling poverty and famine exists in the world while we sit on classified technologies that could completely reverse this situation.
>
> "So why not release these technologies? **Because the social, economic and geo-political order of the world would be greatly altered**. Every deep insider with whom I have met has emphasized that this would be the greatest change in known human history. The matter is so highly classified not because it is so silly, but because its implications are so profound and far reaching. By nature, those who control such projects do not like change. And here we are talking about the biggest economic, technological, social and geo-political change in known human history. Hence, the status quo is maintained, even as our civilization hurtles towards oblivion...
>
> "With the types of weapons currently in the **covert arsenal** — weapons more fearsome even than thermonuclear devices — there is no possibility of a survivable conflict. Yet in the darkness of secrecy, actions have been taken on behalf of every human that may endanger our future. Only a full, honest disclosure will correct this situation. **It is not possible for me to convey in words the urgency of this**."
>
> Dr. Stephen Greer http://www.disclosureproject.org/ES-DisclosureImplications-2.htm

Greer and his team have assembled hundreds of witnesses, many of whom are military or ex-military, who are ready to testify to congressional hearings what they know about the covert black projects which have kept free-energy and antigravity propulsion technology secret from the people.

Greer points out the need to regulate these technologies so they are limited to peaceful uses, but this is of course difficult when the government simply chooses to pretend they do not exist. In the meantime, as the President and the Defense Department continue the charade of war-as-usual, those scalar (longitudinal wave) installations which DO exist continue to fall into ever more sinister hands. And there is no reporter in the press brave enough to ask, "**Mr. Secretary of Defense, what steps are you taking to defend our MINDS from attack by the big Russian longitudinal interferometers?**"

Longitudinal Interferometers are Proliferating

In a later statement Bearden reveals that in fact the Yakuza [Japanese mafia] has **already leased** some of the interferometers with a nearly $1 billion "down payment:"

> "After the collapse of the Soviet Union's economy, lots of things got available for lease or purchase that would never have been under the old system. That happened in the resulting economic chaos. The Russian Mafia rose to prominence, as did some wheeling and dealing entrepreneurs who played both ends against the middle.
>
> "**In that atmosphere, and in the need for money, the KGB allowed leasing of the earlier longitudinal EM wave interferometers to a consortium of the Yakuza and Aum Shinrikyo.** They did this only after they had tested and deployed their great new quantum potential weapons. That class of weapon is the dominant weapon on earth. Five nations now have it: Brazil, Russia (KGB), China, and two nations friendly to the U.S. The deal to lease many of the earlier LWIs [longitudinal wave interferometers] to the Japanese Yakuza and Aum Shinrikyo was consummated at the end of 1989, with a down payment of $900 million in gold bullion as "up front" money. I don't know what the lease per year was and is, but probably something like a billion dollars a year. **So the rogue Japanese acquired longitudinal EM weapons** and that technical know-how. In fact, the Yakuza is producing certain kinds of those in its own facilities in Japan now."
>
> http://www.cheniere.org/correspondence/030202a.htm

Personally, I am a little uncomfortable with the knowledge that the Japanese mafia may now be able to insert things into my mind. I'd like to sleep at night knowing that my dreams are my own, such as they may be, and not some weird fantastic visions cooked up by a criminal organization. **This is a world emergency**.

And in Bearden's analysis, it all comes back to the issue of **energy**. It is the energy crisis which is now provoking the current "oil wars," since oil production has hit the "Hubbard Peak" and will decline. And the electrical power grids must be replaced with scalar power as fast as possible, for these grids are now totally vulnerable to any possessor of longitudinal interferometers. They can all be wiped out in minutes. By anyone who possesses or controls a Tesla howitzer.

> "Anyway, in my personal view, we should have a national **Manhattan Project** right now, to develop decentralized electrical power systems taking their energy from the vacuum . . . The MEG is real and so are several other systems invented by private inventors and colleagues. All these efforts should be massively funded and massively and quickly developed. The survival of this nation is going to depend upon it. . . . There are many other facets, but the major point is that **our scientific apparatus and the public are still unaware of what a desperate struggle we have entered, since it was thrust upon us. 9-11 was a wake-up call**. It was certainly bad enough, and one grieves for those lost American lives. But I fear the **worst is yet to come**, and it is coming at us like a runaway train hurtling right down the track at us."
>
> http://www.cheniere.org/correspondence/021702a.htm
>
> "The electromagnetic weapons mentioned by the Secretary of Defense in April 1997 **are in terrorist hands also**, including the **Yakuza and Aum Shinrikyo**. That rogue group leased those earlier weapons on site in Russia, from the KGB, at the end of 1989. The SecDef stated that such weapons were being used to initiate earthquakes, initiate volcanoes into eruption, and control and engineer the weather. Quite true. And presently

there is a truly massive set of weather engineering operations going on over North America, from that rogue Japanese group manning the weapons in Russia. Some 10 nations of the world now have the type of longitudinal EM wave interferometer weapons (which are what the SecDef was most probably referring to). There are other even more fearsome weapons, possessed by five nations. **So a great deal of the state of the world is not covered in the news at all, and will not be . . .**

"We have been in an undeclared war of an eerie kind for some decades. That war seems now to have started to heat up also. The destructive capability of some of these weapons is awesome, and far greater than nuclear."

http://www.cheniere.org/correspondence/021702a.htm

"In that sense, the first phase of WW III is already completed. Now we are only waiting for Phase II.

"It is against that backdrop that I think one must "analyze" such things as how fast we will be replacing the power grid (**it's going to go anyway, and catastrophically, the terrorists will see to that**). And everything we have is critically dependent upon energy, from the economy to jobs to mobility, to the millions of trucks that transport our goods every day, etc.

"As far as I am concerned, we are already late in declaring a great national emergency in energy, because of the coming destruction of the present energy infrastructure. The scientific community should lead, follow, or get out of the way. The emergency is not in seeking "alternative but conventional" systems. It is in developing "energy from the vacuum" on an emergency, crash basis.

"Unfortunately, everywhere I look I just see business as usual. Most Americans (and news media) seem to think, hey, Afghanistan is finished except for a bit of cleanup, and it's just about over).

"It isn't. It has only just begun."

http://www.cheniere.org/correspondence/021702a.htm

Beyond the Howitzers: Quantum Potential Weapons

Do not use EM propagation through space

As more and more nations acquire the scalar interferometers, Bearden speculates on the next generation of longitudinal weapons called "quantum potential" [QP] weapons.

> "The **quantum potential weapons** are presently only held by three nations: Russia (in the hands of the KGB, which is now know by its new name), Brazil, and the "friendly little nation." Red China may be working on them, but is just at the beginning...

> "The QP weapons have been highly reserved by the Russians. However, the complete extent of their capabilities is still being figured out...
>
> "The quantum potential weaponry (possessed by only three nations) is supreme, because **it uses multiply connected spacetime**, and DOES NOT involve propagation of EM energy through space as does the longitudinal EM wave interferometers.
>
> http://www.cheniere.org/misc/qp.htm
>
> "The operational deployment of full-scale strategic QP weapons on site in Russia occurred at the end of 1989, which is when the formal weapons first went operational. These weapons are not in the inventory of the regular Russian armed forces, but all research and development, manufacturing, deployment, an manning and employment is under the ruthless control of the KGB. Shortly thereafter (within weeks), the KGB leased many of their earlier longitudinal EM wave interferometer (LWI) weapons to a rogue Japanese group comprised of the Aum Shinrikyo and Yakuza."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

In another paper, "Time Polarization - Significance and Weaponization," Bearden tells the story of how a KGB longitudinal wave attack was averted by the longitudinal weapons of a "friendly small nation."

So as it stands, it is precarious. Not until the people demand an end of what author Jim Marrs has dubbed the "**Rule by Secrecy**," can the people's representatives tackle the big problems of a worldwide ban on Scalar Wars and the shifting over to an essentially oil-free economy.

Latest Update from Tom Bearden - March 2002

The whole situation is changing so fast that when I asked Tom to look over this article he mentioned the seeming contradiction when some of the above statements are taken together. At different times the information he has is updated so one has to keep a kind of running track. He was kind enough to clear up the confusion here with the situation as it stands as of March, 2002.

> "We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy." Tom Bearden

Emptiness is Full of Energy

"The Tao is an empty vessel;
it is used, but never filled.
Oh, unfathomable source
of ten thousand things!

"Oh, hidden deep but ever present!
I do not know from whence it comes.
It is the forefather of the gods."

Lao Tzu -Tao te Ching

Afterword

Because of the extreme seriousness of the topics brought forward in this article, I asked Tom Bearden to look it over to make sure I had not gotten anything too terribly wrong or mixed up. I did not want to pass on any misunderstanding of my own. I felt a little nervous writing to an eminent scientist, and hoped I had not presumed upon his time.

But Tom was so kind as to answer immediately, and to my relief I had not gotten anything too terribly wrong. Here is his letter, in which he clears up a misunderstanding, and reports on his current health treatment, and gives a call to the young grad students to turn themselves loose in the whole new field of scalar electromagnetics and free energy.

Subject: RE: Article on Tom Bearden and "Scalar Wars"
Date: Fri, 22 Mar 2002 22:54:01 -0600

Bill,

Well, I must say I'm astounded at your persistence and hard work in congealing all that together. My compliments on a hard job quite well done.

Couple tiny things: In the area mentioning Kervran and Kaznacheyev: Kervran was French; Kaznacheyev was and is Russian. Little typo of "spredding" the immune system. Should be "spreading".

We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy.

Otherwise, it's pretty accurate.

Just now I'm letting the weapons take care of themselves; have to, as I will be recovering now for nearly a year from now after the heart attack last year and discovering that I had contracted mycoplasia (the BW modified kind) in Canada in 1968. Finally got a test which confirmed the chronic mycoplasia (which I had for 33 years!). The stuff burrows inside one's red corpuscles, then draws nutrients from the hemoglobin, hardening it and reducing its ability to take on oxygen (by as much as 50% or more). The resulting symptom is loss of endurance (which I had for that 33 years) -- the so-called "chronic fatigue syndrome".

Many of the Gulf War Veterans who are sick also have it, though they apparently have other complicating factors as well. A very high percentage of persons with chronic fatigue syndrome also have mycoplasia. The treatment for the long-standing mycoplasma infection of that nature is to stay on antibiotics for a year. The only time the mycoplasma is vulnerable is when one wears out a red cell in normal usage, and the body makes a new red cell to replace it. The mycoplasma comes out of the dying cell to infect the new replacement cell, so it is exposed and vulnerable at that specific time and that time only. The antibiotics kills it at that time. Hence the necessity to stay on antibiotics for such a long time so it is continuously in your bloodstream (I've already been on antibiotics for over two months, and have at least 10 more months to go). We also take a little medical oxygen every day.

Another symptom of the mycoplasma infection is the association of runaway fibrillation of the heart. To stop that, there are pills one takes, and they stop it. But they also further reduce the volume of blood pumped by the heart, so that this projects one right back into the hypoxia, dizzy, loss of balance, etc. Catch 22 situation; the fibrillation will build and cause heart attack, stroke, etc. if not stopped, and if stopped it puts you back in the hypoxia which can cause many things! With the conventional treatment, you're damned if you do and damned if you don't. Hence the need for a little extra medical oxygen, so you can take the "heart clamp" pill and stop the fibrillation, but also augment the oxygen enough to continue to survive without a stroke or some such. Interestingly, it's very difficult to even obtain a mycoplasma test, and Medicare fights you tooth and nail over paying for the oxygen. Either they will pay for it or I'll pay for it myself; it is absolutely necessary.

After a heart attack, one is treated in the heart clinic by specialists who are very caring and excellent practitioners, but know nothing of mycoplasma infection, so not consider it even a part of "heart disease", and who are rather puzzled when one has no clogging of the arteries, has the right count of red corpuscles (largely nonfunctional, of course, but there), pumps the right amount of blood, etc. So they identify and treat only the runaway fibrillation, do not prescribe the oxygen -- and essentially place you on a regime as best they know, but one that is guaranteed to kill or maim you or turn you into a totally disabled person or human vegetable via stroke, paralysis, etc. Fortunately my family doctor will prescribe the oxygen anyway, though I may have to pay for it myself. Whatever works!

However, I'm slowly improving little by little, working as much as I'm able to and trying to finish my book: Energy from the Vacuum: Concepts and Principles, and get it to the publisher for publication later this year. Hopefully the book will turn all the interested young grad students and post-docs loose in the free-energy area. They will not have to spend 30 agonizing years to get to where I am, and where a few others are, but can simply start here and go forward. Since speed to develop energy from the vacuum is of the essence because of the world situation, I'm moving on it as fast as I am physically able to.

Meanwhile, working with the AIAS (Alpha Foundation's Institute for Advanced Study), the AIAS (15 authors) has now gotten about 20 scientific papers published in leading journals (Foundations of Physics, Foundations of Physics Letters, Physical Scripta, Optik, etc.) dealing with electromagnetic energy from the vacuum. Several more are either approved for publication or in the review process also. So slowly we are getting it into the mainstream scientific literature. Hopefully it will be in time, but it's cutting it close.

Very best wishes,

Tom Bearden

Late update

Massive display of longitudinal interferometers?

UFO Fleet Taped
Over Bedhampton, England

Is this sighting a case of the testing of all the Russian (KGB) longitudinal interferometers at once? And coordinating their targeting through the woodpecker grid to a common place, as an exercise in making a massive scalar attack upon a given region? Was this a confluence of Russian Tesla howitzer marker-beacons?

Bedhampton, England - March 22, 2002 - Video Frames by: Anthony Woods

Certainly for a massive all-out scalar attack you would want coordinated movements of the target areas of many howitzers at once, for you could then just cut a bloody swath across a wide area, all the howitzer blasts marching across the land in a kind of scalar version of the "scorched-earth policy." All in a line, blasting, and heat, and mind waves. Or you might want to put the blasts all in the same place for a truly horrendous destruction. And on top of that biowar on citizens whose immune systems have been weakened with the quantum potenial weapons. This is the latest version of Armageddon.

And should any madmen gain control of these weapons, any rogue group at all, and should they have some aggenda to radically depopulate the world, to simply kill by the billion, the means will be in their hands to do so.

Update: MEG patent granted

Motionless Electromagnetic Generator Patent Granted

This message is from **Jean-Louis Naudin**, the French scientist who successfully replicated the Motionless Electromagnetic Generator (MEG) in France. He announces the granting of the patent to the MEG Builder's group at Yahoo.

----- Original Message -----
From: jnaudin509@aol.com
To: jlnlabs@yahoogroups.com
Sent: Tuesday, March 26, 2002 2:20 PM
Subject: [jlnlabs] (Info) GOOD NEWS : The Tom Bearden's MEG IS PATENTED

Dear ALL,

GOOD NEWS !!!!

The **Motionless Electromagnetic Generator**
from Tom Bearden is now PATENTED
US **6362718** granted on March 26, 2002

US Patent 6,362,718 :
Motionless Electromagnetic Generator (MEG)

Abstract: An electromagnetic generator without moving parts includes a permanent magnet and a magnetic core including first and second magnetic paths. A first input coil and a first output coil extend around portions of the first magnetic path, while a second input coil and a second output coil extend around portions of the second magnetic path. The input coils are alternatively pulsed to provide induced current pulses in the output coils. Driving electrical current through each of the input coils reduces a level of flux from the permanent magnet within the magnet path around which the input coil extends. In an alternative embodiment of an electromagnetic generator, the agnetic core includes annular spaced-apart plates, with posts and permanent magnets extending in an alternating fashion between the plates. An output coil xtends around each of these posts. nput coils extending around portions of the plates are pulsed to cause the induction of current within the output coils.

Inventors: Patrick; Stephen L. (2511 Woodview Dr. SE., Huntsville, AL 35801); Bearden; Thomas E. (2211 Cove Rd., Huntsville, AL 35801); Hayes; James C. (16026 Deaton Dr. SE., Huntsville, AL 35803); Moore; Kenneth D. (1704 Montdale Rd., Huntsville, FL 35801); Kenny; James L. (925 Tascosa Dr., Huntsville, AL 35802)

Appl. No.: 656313
Filed: September 6, 2000

For some technical info see my web site at : http://jnaudin.free.fr/html/meg.htm

Best Regards
Jean-Louis Naudin
Email: JNaudin509@aol.com
Main Web site : http://jlnlabs.org
Site France : http://jlnlabs.multimania.com

Click here to view patent # **6,362,718**

Patent Story on Rense.com:

Important Note 2006! MEG delayed: **Click here for a current status history of the MEG from Tom Bearden**

Rense.com

MEG Scalar Energy Device Patented - Production Starts Next Year

From Bill Morgan
3-28-2

Patent was granted on March 26, 2002 for "The Motionless Magnetic Generator," (MEG) US Patent 6,362,718, which is likely to become the first commercially available free energy device in history in about one year from now. The machine will provide free electricity from the vacuum, for the life of the device, which should be a very long life since it has no moving parts. You can see a picture of scientist Jean-

Louis Naudin's MEG replication model at: http://jnaudin.free.fr/html/meg.htm.

It has strong magnets, coils, and a controller unit with the electronics. Naudin made the announcement to the MEG-builder's Yahoo group. The announcement has significance since the patent office has always been skeptical of devices which seem to "get-something-for-nothing." But according to the new science of scalar electromagnetics, the MEG does not break the law of conservation of energy. It's just that the energy is conserved in the fourth dimension, time, and not our 3-space world.

The MEG provides electrical energy by tapping the longitudinal electromagnetic (EM) waves which exist in almost infinite abundance in the vacuum of space. This ocean of energy which permeates everything is sometimes called the "zero point" energy, since it remains there even at absolute zero temperature.

Four inventors are listed: Stephen L. Patrick, Thomas E. Bearden, James C. Hayes, Kenneth D. Moore.

Tom Bearden has explained the operation of the MEG on his website Cheniere.org, and also speaks about the new fearful weapons that can and have been made using the same "longitudinal waves" of the vacuum.

The complicated physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum," which can be found at Tom Bearden's website:

cheniere.org.(http://www.help4all.de/energy/MEGpaper.pdf)

The first MEG units to be produced for sale will output 2.5 kilowatts of free electricity. Forever. They should be in production about a year from now. Facilities for manufacturing the device are being set up in an unnamed "friendly nation."

This free electricity will flow indefinitely, without much, or any maintenance. The units may be hooked together to provide more wattage, so four of them would provide 10 kilowatts. After some production experience units will be made which output 10 kilowatts each. With a couple of those units a house could get off the electrical grid.

Tom Bearden, one of the inventors has said "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat." This cat, it would seem, is out of the bag now.

The ordinary EM waves we are familiar with are called "transverse" waves, to distinguish them from the new "longitudinal" EM waves of the vacuum. Bearden has explained in depth on his website cheniere.org that wherever there is a dipole (battery, generator, magnet) there is an unseen flow of longitudinal EM waves in that local vacuum, the only problem is in tapping that energy and "transducing" it to electricity. Bearden says that the problem with all the electrical circuits we have is that they are two-wire circuits, a loop by which half the energy goes back to destroy the dipole. In the MEG that closed loop is never made. So the dipole is not being destroyed.

Patents are not granted on devices which do not work, so in a sense this announcement proclaims a new era. The MEG device itself proclaims and proves that energy is abundantly available everywhere, for free. We only have to build the devices to tap it. MEG type devices could be put into a car with an electric engine to make a truly fuel-less automobile. Electricity can be available in very remote places. People will be able to get off our very terrorist-vulnerable power grids. In an era of terrorism a highly dispersed power system would be most desirable rather than our centralized systems.

What is somewhat astonishing to me is the discovery by Bearden et. al. that time itself is actually compressed energy, and that this free energy is actually coming from the time domain, the ocean of longitudinal EM waves which fill the empty vacuum of spacetime. In fact, time is energy compressed by the same factor that matter is compressed energy: the speed-of-light squared. Thus there is a new companion to Einstein's E=mc2. Can you say "E equals delta-tee-cee-squared?" The "tee" is time and delta-tee is change in time.

A company has been set up to manufacture the MEG called Magnetic Energy Limited.

The ramifications of free energy are enormous. The oil wars are not necessary. If we threw as much money at this technology as we are spending on the oil wars we would be free of the need for oil in less than a decade. With fuel-less cars air pollution will be greatly lessened. Third world nations can raise their standard of living eventually. And the energy is free. And it never runs out.

I have put together a kind of "Bearden for Beginners" article which explains some of the basic concepts of the current state of scalar technology.

http://www.prahlad.org/pub/bearden/scalar_wars.htm

The granting of the MEG patent is the herald of the new era of scalar electromagnetics, and the free energy which flows forever and never runs out.

Bill Morgan

Status History of the MEG

updated Oct. 26, 2006

Important notes on the status of the MEG (Motionless Electromagnetic Generator)

Since the time this "**Scalar Wars**" article was written the timetable of the **MEG** has been thrown off considerably. Bearden had originally said the MEG could be marketed within 1-2 years ***of receiving funding for final research and scale-up costs***. I thought certainly it would not take very long, but over 5 years later they still have not gotten the $10 million or so necessary. There are various reasons for this, some of which Tom Bearden has written about.

The place for updates on the MEG status is Tom Bearden's correspondence section at his website Cheniere.org:

http://www.cheniere.org/correspondence/index.html

Here are a few of his reports which detail the history of the MEG status :

Mar. 6, 2006
http://www.cheniere.org/correspondence/030206.htm

May 16, 2005
http://www.cheniere.org/correspondence/051605.htm

Jan. 19, 2005
http://www.cheniere.org/correspondence/011905.htm

Dec. 5, 2004
http://www.cheniere.org/correspondence/120504.htm

Do the magnets run down after a while?

Further developments

Tom Bearden clears up a question

When the MEG patent was released many questioned the patent language which seemed to imply that permanent magnets of the MEG would "wear out" after awhile. I myself wondered about the same thing so I put the question to Tom, and here is his reply.

> Dear Bill,
>
> Thanks for the kind words; much appreciated.
>
> I'm not up to going on radio or TV, and won't be for quite some time. Still suffering from hypoxia, though very slowly improving, and will still be on antibiotics for 10 more months at least (maybe longer).
>
> Obviously an extensive dialog developed between the Patent office and our patent attorneys. As best I can understand it, the Patent office accepts the fact that a permanent magnet produces and emits energy, or has a history of accepting it. They do not appear to be at all cognizant of the broken symmetry of opposite charges, and hence of a dipole. So in their opinion (which seems to be vague), the magnet has to use itself up very gradually, or some such. So some rewrites were essentially required by the Patent office.
>
> All I can say is that a magnet is a very durable thing, so long as you don't heat it up too much or shock it to much. Certainly one will make a good 20 years, and of course there are plenty of magnets that have been around awhile longer.
>
> That said, in our earliest experiments we did "measure" some such effect, or so we thought, until we found that a particular instrument was bad and giving erroneous readings. So with a new instrument, we have not observed since then any such readings.
>
> Out of all that fuzziness came the inclusion of "using up the magnets gradually". So far as we are aware (after correcting those early readings on a defective instrument), we have seen no such measurements or effects since then at all. Nonetheless, just in case there is some very small effect in that respect, at least it's in there.
>
> Actually, as you are aware, capacitors and solid state components are the major concerns for failure, and any system does require maintenance, both periodic and whenever malfunction or failure occurs. The MEG is no different from any other device in that respect, so of course it will have a component failure rate and maintenance required from time to time, just as any other electromagnetic device. But without moving parts, it can be ruggedized and made very durable indeed.
>
> Please give Jeff Rense my regrets, and I do admire his work and his show. He's laying it in there.
>
> The depth of the scientific mindset against COP>1.0 is inexplicable, particularly when physics itself already has COP>1.0 validated

experiments (such as Bohren's experiment, which outputs 18 times as much as one inputs by Poynting calculations), and it works every time. Any competent university optics lab can perform the experiment. Also, even the conventional texts admit that a charged capacitor or electret laid on a permanent magnet, so that the E of the cap or electret and the H of the magnet are at right angles, is a "free energy machine". That silly thing will just sit there and pour out Poynting energy S = EXH indefinitely. Years and years and years. If you wait just a single year, that silly thing has changed the energy density of a volume of space a light year in radius, reaching out well beyond the solar system. Now that's a WHALE of a lot of EM energy that beast poured out in that year, and it will still be going strong and pouring it out at a steady rate. The charges and dipoles in original matter in the universe have been pouring out EM energy in that fashion for some 14 billion years, give or take a billion or so. It's also easy to see that, prior to Lorentz's ARBITRARY symmetrical regauging of the Maxwell-Heaviside equations, those equations prescribe both EM systems in equilibrium with their active environment, and EM systems not in equilibrium with it. The latter systems are permitted by the thermodynamics of open disequilibrium systems to perform five magic functions: (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator inputs (the active environment inputs the rest of it), (4) power itself and its load (the active environment inputs all the energy, like a windmill), and (5) exhibit negentropy. That is a well-established thermodynamics. Lorentz arbitrarily discarded all such open disequilibrium EM systems in order to get simpler equations having analytical solutions and not requiring numerical methods. So he simply scrapped all the really interesting and challenging EM systems in Maxwell's theory. ARBITRARILY!

That our scientific establishment continues to teach only that "half" of the theory, and not point out the rest to the young students, and not fund research into developing such systems that take energy from the vacuum, is simply inexplicable. It's a matter of total dogma, not science at all.

But hopefully it is changing a little bit. At least now we cite chapter and verse when these things were done, and what they are, and the young grad students and post docs can go check the papers and make up their own minds.

That is the real hope of the future: that we get a generation of young fellows, with excellent skills, who now know where all the skeletons got hidden in the closet, and understand that the electrical engineering model arbitrarily assumes an inert vacuum environment (falsified for more than a half century in particle physics) and a local flat spacetime (falsified for nearly a century by general relativity).

Once they know that past history, and go read into what broken symmetry of opposite charges means for a dipole and for all dipolar EM circuits, they will set to and change that more than a century of error. In that case, we shall have numerous solutions to the energy crisis rather quickly, popping up all over.

Good thing. It will certainly displace some great energy barons, but it will also free the populace. One can extract all the EM energy from the vacuum one wishes, anywhere, anytime, with ridiculous ease. One can produce that "electrical wind of energy" at will, by making a simple dipole and then letting it alone. The only energy problem is in how to then intercept some of that energy flow and "catch" it in a circuit, discharge it in a load to power it, and not use half the "caught" energy to kill that dipole that is gushing out the extracted EM energy from the vacuum.

Very best wishes,

Tom Bearden

END

Revised: May 16, 2007

Download this article in Ms Word format
(ScalarWars.doc)
Send a comment to Bill Morgan

See also:

Scalar Electromagnetics

Scalar Wars:
The Brave New World of Scalar Electromagnetics
(Read this one first!)

Horrifying US Secret
Weapon Unleashed In Baghdad

Scalar Heaven or Hell:
Psychoenergetics

Engineering the Mind

Sacred_Scientist thread at Rumor Mill News

Are any of these Earthquakes Being caused by Scalar Weapons?

Safe Places, Earth Changes, and Gordon Scallion

Examples of probable Scalar Weapons Tests

Iraq: World's First Scalar War?

Space Energy Access Inc. Announces
New Free Energy Device

SINGLE MOST IMPORTANT BREAKTHROUGH IN HISTORY

Comparison: Ebomb and Scalar Waves

IT COULD COME AS SIMPLY AS THIS
Scalar destruction of U.S. power grid

WHAT A SCALAR "SMOKING GUN" WOULD BE

BEARDEN REPONDS TO SCALAR ATTACK POSSIBILITY (Columbia)

THE SCALAR ATTACK POSSIBILITY
Columbia Disaster

TOM BEARDEN ON THE CHALLENGER DISASTER

ANOTHER BEARDEN ACCOUNT OF '86 CHALLENGER SCALAR ATTACK

Broadcasting Thoughts & Disease

Thoughts on Diane Harvey's Appeal to the U.N.

Scalar War Has Already Begun

Musings on Scalar Electromagnetic Phenomena

The New Scalar Paradigm

ELECTROMAGNETIC WEAPONS TO BE USED

SCALAR: BEARDEN SAYS U.S. NOW DEFENDED

American Karma & The River of Tears

Open Letter to Donald Rumsfeld

My Rumor Mill Posts on All Topics

Your comments are welcome: prema@prahlad.org

Journal of Directed Energy, **1**, Fall 2003, 48–72

Compact Pulsed Power for Directed Energy Weapons

David Price,[*] Carl Bloemker, Edward Goldman, David Nett, Sidney Putnam, Doug Weidenheimer, Roger White, and Tony Wynn

Titan Corporation, 2700 Merced Street, San Leandro, California 94577

The Titan Corporation has been developing directed energy technologies since the early 1980s. Over this time Titan has provided advanced power supplies for electromagnetic gun and laser research while making broader contributions in the field of high-power microwaves (HPM). Titan's HPM work expanded from initial efforts fielding and operating the first gigawatt-level oscillators for susceptibility test applications, to research and development maximizing the peak and average output powers and overall efficiency realized from such systems. As interest in HPM technology has spread globally, Titan has leveraged this core competency and provided the HPM equipment for nearly all of the major European effects test facilities. Titan's current interests focus on compact, efficient and reliable directed energy weapon systems and the advanced subsystems and components that enable the same. Specific subsystems of interest include pulse-forming networks and intermediate energy storage and power conditioning elements (electronically reconfigurable batteries and power ride-thru subsystems). These subsystems are designed for reduced size and weight while still meeting severe service, platform integration, lifetime, and thermal management constraints. Specific components under development at Titan include laser-gated solid-state switches and both high peak and high average power, frequency-agile, HPM oscillators. Status and recent results from this research are presented.

KEYWORDS: Electronically reconfigurable batteries, High-power microwaves, Laser-gated solid-state switches, Magnetron, Power-ride thru, Pulsed power, Reltron

1. Introduction and Previous Work

Directed energy weapons (DEWs) are emerging as a key defense technology of the early 21st century. The first laser antimissile and antiartillery systems are within perhaps five years of deployment, and newly envisioned information warfare and nonlethal point defense missions are hastening the deployment of high-power microwave (HPM) systems. Department of Defense (DoD) programs in "more electric" and "all-electric" platforms and electrically driven weapon and self-defense systems are promoting the development of laser, high-power radio frequency (RF) and both electromagnetic (EM) and electrothermal/chemical (ETC) gun technologies. In fact, the DoD's current fundamental "transformation" efforts are based on stressing the development of capabilities to deal with threats emerging in the

---

Received February 13, 2003; revision received May 28, 2003.

[*] Corresponding author; e-mail: dprice@titan.com.

© 2003 Directed Energy Professional Society

Table 1. Triservice directed energy applications

U.S. Army [Future Combat System (FCS)—hybrid electric drive vehicles]
- Advanced armaments—electrothermal/chemical guns
- Enhanced area air defense—solid-state heat capacity laser system
- Counter-HARM
- Counter-ATGM
- Countermine (electronically fuzed and wide-area threats)

U.S. Navy [All-Electric Ship Programs (Integrated Propulsion System and DDX)]
- Advanced stand-off shore bombardment—electromagnetic guns
- Advanced anti-ship missile defense (free-electron laser and HPM)
- Nonlethal area denial

U.S. Air Force (more electric aircraft program and advanced UAV programs)
- Enhanced self-defense—solid-state heat capacity laser system
- Space-based laser program
- Airborne laser
- Suppression of enemy air defense
- Aircraft self-defense
- Close air support
- Defensive and offensive counter air
- Strategic attack/strike warfare, attack operation/air interdiction
- Combat search and rescue
- Area denial
- Cruise missile defense

Table 2. Countries now developing high-power RF technology

Programs	Countries
Major	United States, China
Medium-scale	United Kingdom, France, Germany, Russia
Emerging	Sweden, Japan, India, Taiwan, Australia, Israel, South Korea

Information Age instead of emphasizing countermeasures to specific threats. DEWs hold the promise of engaging multiple threats and multiple threat types, thereby fitting very well into this conceptual paradigm. Some specific examples of defense applications of DEW technologies are shown in Table 1.

The initiatory HPM programs in the USSR and United States have evolved and influenced foreign allies and threats to take an interest (Table 2). The collapse of the USSR led to the dispersal of Russian and Ukrainian HPM workers and the proliferation of the Soviet technology to the Third World. This global emergence is driven by at least two factors: 1) The dependence of both military systems and civilian support infrastructure on sensitive electronics is increasing, and 2) after a more-than-20-year effort, an understanding of the susceptibilities of both military and commercial systems to HPM threats is maturing.

For these reasons the Titan Corporation has taken a keen interest in the development of systems, subsystems, and advanced components across several different technologies to support HPM and laser directed energy programs.

In 1997, Maxwell Technologies purchased Physics International (PI) from Primex Technologies and merged two of the three most prominent, commercial, pulsed power entities in the United States. PI and Maxwell pulsed power capabilities were combined into a new entity, Maxwell Physics International, which became a component of Maxwell's Systems Division. Maxwell subsequently decided to leave the DoD and Department of Energy (DOE) pulsed power business altogether and sold the combined Maxwell/PI pulsed power divisions to the Titan Corporation, the parent company of Pulse Sciences Incorporated (PSI). Now, the three most prominent names in commercial pulsed power are combined as the Pulse Sciences Division (Titan-PSD) of the Titan Corporation.

In March 2002, Titan completed the acquisition of Jaycor and its subsidiary California Tube Laboratory (CTL). This acquisition established Titan as one of the world's primary industrial authorities on electromagnetic effects and suppliers of RF and narrow-band microwave DEW technologies and associated electrical power systems. High average power, high efficiency, magnetrons; high peak power, pulsed, relativistic magnetrons and super-reltrons; short pulsed, ultra-broad-bandwidth systems and megawatt-class power supplies and power conditioning subsystems are all designed and manufactured by Titan. This paper will chronicle some of the past and report on some of the current contributions that Titan has made to the development of pulsed power technology for DEW applications.

1.1. Previous Titan work in support of HPM programs

Titan-PSD has been developing directed energy technologies since the early 1980s. In this time, Titan has delivered capacitor banks for electromagnetic gun research with a total stored energy exceeding 200 MJ and advanced pulsed power systems for several large laser systems (Fig. 1). Over 40 years, Titan-PSD has delivered more than 200 pulsed power systems for electromagnetic pulse (EMP), x-ray, and lightning simulation; flash radiography; and other applications worldwide. Although somewhat peripheral to DEW research proper, these projects have nonetheless formed much of the experience base supporting Titan's directed energy technologies contributions.

Titan's primary contributions have been in the field of HPM (Fig. 2). Work started in 1984 when Titan set up and began operating a susceptibility effects test facility for the Defense Threat Reduction Agency (DTRA). A 30 × 20 × 16 ft anechoic chamber to support this work was originally designed to operate in X-band and above. At the outset of the U.S. HPM directed energy program the community felt that the higher gains possible from fixed antenna sizes at higher frequencies would drive technology development into these bands. Over the next 10-plus years, the 25 effects test programs conducted on strategic and tactical systems in the Titan facility and tests conducted at other sites both confirmed that the lowest susceptibility thresholds are observed in the S-band and below. HPM source development history has mirrored these findings.

1.1.1. HPM source developments at Titan. In 1983 Titan initiated HPM source research in its test facility by developing first C-X band overmoded vircators[17] and later an L-band vircator.[13] These systems produced gigawatt outputs through the oscillating virtual cathode mechanism. Their main shortcoming was their broad bandwidth due to chirping induced by diode gap closure. To address this shortcoming Titan developed the first cavity vircator[3] that reduced the output bandwidth from several hundred megahertz to a few tens of megahertz. Versions of these devices were soon after sold to Nucletudes in France

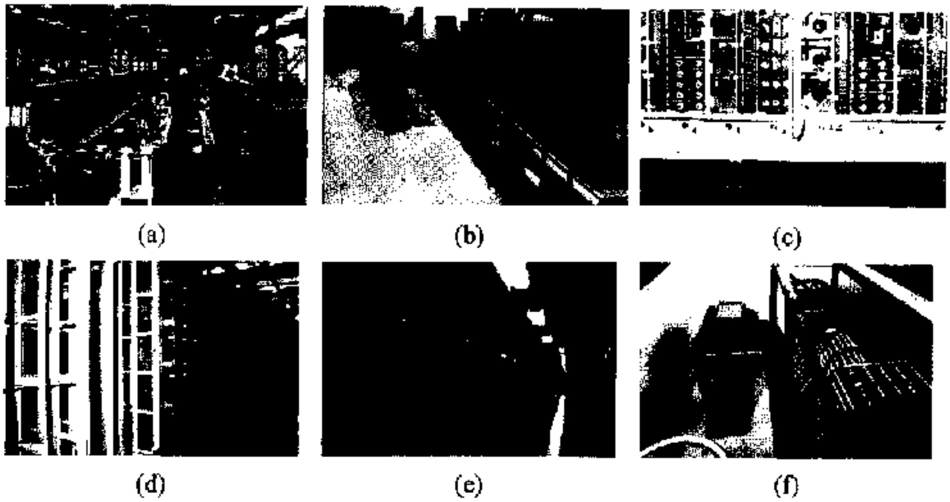

Fig. 1. Titan-PSD has designed, fabricated, and delivered the pulse power subsystems for most of the major electric gun facilities worldwide: (a) the decommissioned 32-MJ Thunderbolt System built for a Strategic Defence Initiative Office (SDIO) project; (b) the existing 32-MJ Kirkcudbright system currently being operated by the United Kingdom; (c) the first transportable pulse power subsystem designed for an electric weapon system. [This 8.5-MJ stored energy subsystem, consisting of four separate PFNs, was designed by Titan-PSD under the Army/Army Research Development and Engineering Center (ARDEC) Balanced Technology Initiative and was capable of delivering up to 5 shots at 3 rounds per minute.]; (d) the 52-MJ bank delivered to the ARDEC for electromagnetic and coil gun research; (e) the pulsed power driver for the 40-MJ OMEGA laser at the University of Rochester; and (f) a portion of the 100-Hz, 750-kV, 1.25-μs modulator for NRL's EMRLD laser.

and Culham Laboratory in the United Kingdom to help kick-start HPM effects testing and technology development programs at each location.

Titan initiated work on magnetrons by acquiring S- and X-band, A-6 versions from Bekifi at the Massachusetts Institute of Technology.[2] These devices were integrated into the Titan test facility, and within two years an L-band version of this same design was developed.[16] An S-band magnetron was sold commercially to Thales in France in 1987 to help kick-start HPM effects testing and technology development programs there.

This HPM source development work soon expanded from the initial efforts to field gigawatt-level vircators and magnetrons for the effects testing applications to the development of other high-power sources. These include narrow-band klystrons (both high- and low-perveance variants) and reltrons, as well as ultra-wide-band systems. In parallel, the peak and average power and spectral characteristics of these HPM sources were being explored in research involving phase locking, repetitive pulsing, and frequency agility. Other related Titan work included the development of pulsed power systems to drive these HPM sources involving several different architectures [pulse-forming networks (PFNs) with capacitive energy storage and voltage adders with magnetic compression power conditioning]. These advancements are described briefly in the following.

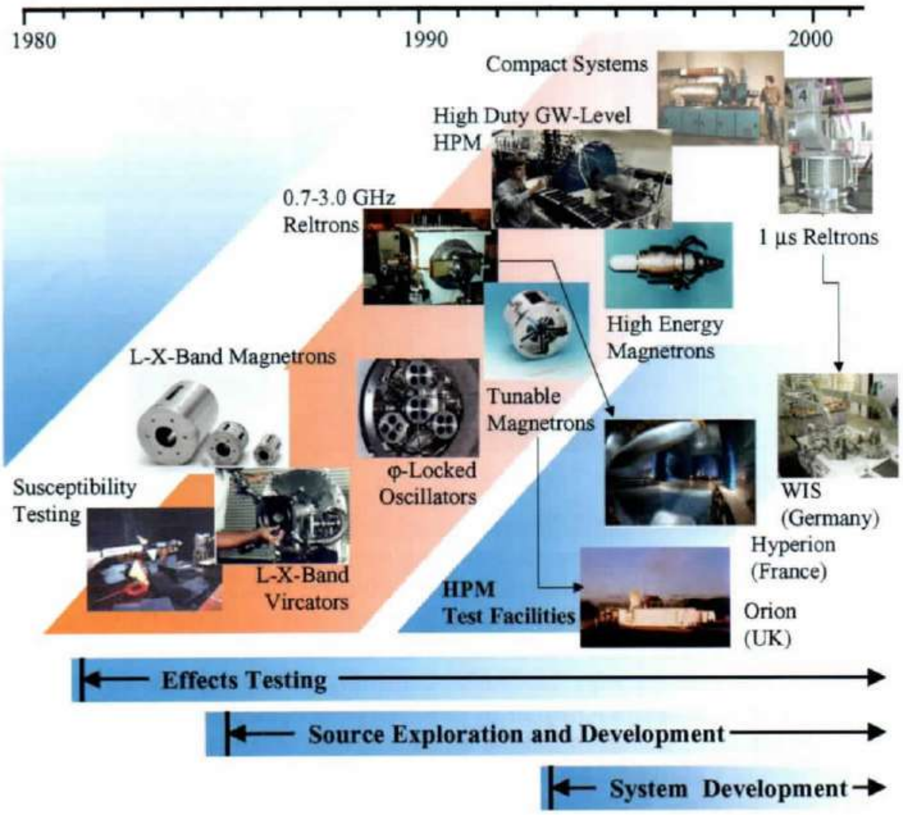
1980
1990
2000
Compact Systems
High Duty GW-Level HPM
0.7-3.0 GHz Reltrons
1 μs Reltrons
High Energy Magnetrons
L-X-Band Magnetrons
Tunable Magnetrons
φ-Locked Oscillators
Susceptibility Testing
L-X-Band Vircators
HPM Test Facilities
WIS (Germany)
Hyperion (France)
Orion (UK)
Effects Testing
Source Exploration and Development
System Development

Fig. 2. Titan contributed to the field of HPMs in three distinct eras: 1) susceptibility effects testing, 2) HPM source development, and (currently) 3) DEW system development. The effects testing work started in 1984. Then, Titan operated the DTRA susceptibility effects test facility supporting tests for external DoD users. Soon thereafter Titan began using the facility for its own end-to-end DEW research programs. In parallel, Titan HPM source development efforts initially focused narrowly on providing the facility with minimum essential capability but soon expanded to meet specific weaponization and testing requirements. During this period, Titan performed some of the seminal narrowband HPM work associated with phase-locking, frequency agility, and repetitive operation. Source development work continues today, focusing on improving energy output and tunable bandwidth. These efforts are combined with the development of other advanced pulsed power subsystems and components needed to realize compact, efficient, and reliable DEW systems.

In the late 1980s, Titan S-band magnetrons were used in a three-year effort to understand how to achieve output powers exceeding 10 GW from a phase-locked array.[9] Many mission scenarios for DEW systems that could produce burnout or upset in military targets using only a single HPM pulse drove output requirements to this 10-GW level. The combination of gigawatt-level sources locked in phase appeared an effective way to realize this output level without increasing the RF fields and beam energies within individual diodes (and dealing with the RF breakdown and thermal issues that would certainly ensue). In this effort both master/slave and peer configurations were investigated to understand the optimal architecture for such an array. In the end, a module of seven "peer-coupled," 2.8-GHz, magnetrons was built that produced an output power of 2.9 GW. Extrapolation to even higher output levels appeared feasible.

At this time, the HPM community was beginning to recognize that there must exist a middle ground between HPM and electronic warfare (EW) DEW concepts. The HPM approaches were characterized by superpower, single-shot, single-frequency, pulse formats, while the EW approaches made use of very low powers but incorporated modulations in the output pulse formats to effectively exploit susceptibilities within their intended targets. To access this middle ground the HPM sources had to first be shown capable of operation in repetitive or burst modes. This burst capability would be surely required of any DEW system, if only to engage multiple threats deployed in sorties or clusters. In response to these weaponization considerations Titan-PSD developed relativistic magnetrons capable of operation in 1-kHz bursts.[4] A 1.1-GHz, L-band, magnetron was driven by a magnetic pulse compression modulator (see discussion below) and achieved 250-pps repetitive operation at 1.2-GW peak power, in a burst of 100 shots. The average power of 12.6 kW during the burst remains today the state of the art for gigawatt peak power sources. To demonstrate that the magnetron's diode could recover in a time short enough to support higher repetition rates, the device was run for ~5 pulses (limited by the modulator) with 1 ms between individual pulses.

In the early 1990s both Jaycor and PI were actively extolling the use of frequency-agile waveforms to enhance lethality in backdoor attacks. Frequency agility exploits the characteristics of typical coupling cross sections that show pronounced dependence on microwave frequency. An HPM source that can sweep or hop frequency across a band has a higher probability of matching to a coupling resonance than a source fixed at a single frequency. Because of this strong dependence of backdoor coupling and susceptibility levels on microwave frequency, it was apparent that test facilities needed the capability to vary frequency continuously in order to obtain a comprehensive and accurate assessment of any given test asset's susceptibilities. In response to these testing and weaponization considerations Titan-PSD developed frequency-tunable magnetrons. These oscillators can be tuned ±17% about a central frequency,[10] which today still represents the state of the art. In the mid-1990s these magnetrons were integrated into advanced concept demonstrations for the Army and were also exported to the United Kingdom as part of an HPM test facility.

In the late 1980s and early 1990s Titan-PSD developed several ultra-broad-bandwidth (UWB) systems for Army and Air Force DEW research programs. The first system was based on a spark gap-switched Marx that drove a variety of different broadband antennas. It could launch 200-MW total RF power in 3-ns pulses with rise times of ~300 ps at a repetition rate of a few pulses per second. This compact, self-contained system was used in both indoor susceptibility tests and outdoor technology demonstrations. A second project produced a system that could launch impulse waveforms at 200 pps in burst mode.[15] A 3.66-m-diameter paraboloidal reflector was driven at its feed by a hydrogen switch that was integrally coupled to a novel electromagnetic lens. The radiated waveform had a rise time of ~100 ps and a 10–90 rise-to-fall pulse width of 45 ns. The peak electric field on bore-sight measured at 305 m was 4.2 kV/m.

Titan invented the super-reltron in 1992. A suite of tubes that operate between 0.7 and 4.5 GHz has been developed since that time.[12] These novel tubes represent a marriage between pulse power and conventional high-power klystron techniques. An energetic electron beam is modulated by periodic virtual cathode formation, and the bunched electron beam is postaccelerated to energies approaching 1 MeV. Microwave power is extracted with ~50%

Table 3. Operating parameters of Titan-built modulators for DEW applications

Modulator	Voltage, kV	Current	Pulse width	Pulse repetition frequency	Pulse energy	Average power	Switch type
LS-15	50	10 kA	100 ns	1.9 kHz	50 J	95 kW	Spark gap
NED laser driver	±40	260 A	2 μs	10 Hz	3 kJ	30 kW	Thyratron
EMRLD PFN	750	38 kA	1.3 μs	125 Hz	35 kJ	4.4 MW	Spark gap
EMRLD (trigger)	600	260 A	400 ns	125 Hz	100 J	12.5 kW	Spark gap
OMEGA[a] laser driver	15	128 kA	200 μs	Single shot	40 MJ	—	Ignitrons
CLIA HPM driver	750	10 kA	100 ns	250 Hz	750 J	188 kW	Magnetic
ORION HPM driver	500	10 kA	50–500 ns	100 Hz	250 J	25 kW	Thyratron
Compact HPM driver	450	9 kA	450 ns	10 Hz	1.8 kJ	18 kW	Spark gap

[a] 120 modules, each storing 0.33 MJ.

conversion efficiency downstream in a dual-cavity output section that is tuned to the bunch frequency. The reltrons can produce microwave pulse widths approaching 1 μs and pulse energies of a few 100 J, outstripping the relativistic magnetrons that to date are limited to pulse widths of about 100 cycles and pulse energies below 100 J.

In response to the HPM effects testing considerations discussed above, Titan developed frequency-tunable reltrons. These oscillators can be tuned ±10% about a central frequency.[11] These devices are still in development today (see Sec. 2) and have been provided to both France and Germany as HPM threat simulators.

1.1.2. Development of pulse power modulators for DEW applications at Titan. Titan has developed a large number of modulators for a variety of laser and HPM DEW applications, with voltage ratings from tens of kilovolts to several megavolts and average power levels from tens of kilowatts to multimegawatts (Table 3). They also range from relatively simple units to complex, state-of-the-art devices with various different architectures and subsystems. These include PFNs for generating the required pulse shapes; thyratron switching for precise timing of the pulse switching into the load; low inductance layouts for fast output pulse rise times and high-turns-ratio pulse transformers for generating the high voltage outputs. The switching technologies include thyratrons, spark gaps, magnetic switches, and solid-state devices, such as metaloxide semiconductor field effect transistors (MOSFETs), insulated gate bipolar transistors (IGBTs), and thyristors (Fig. 3).

Several of these modulators have been designed for long-lifetime, reliable operations. Such features are essential if these systems are to find their way into military applications. As an example, the Defense Advanced Research Projects Agency (DARPA) excimer Raman-shifted laser device (EMRLD) modulator (a 100-Hz, 750-kV, 1.25-μs, lumped-element PFN to pump an excimer laser) was specifically designed for and demonstrated a lifetime of 10^{10} shots.

The ORION pulser (Fig. 3b) design has command resonant charge and intermediate energy storage sections each switched by thyratrons that drive a step-up transformer and

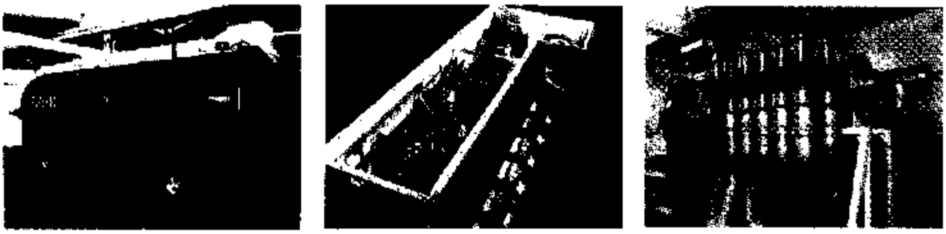

(a) (b) (c)

Fig. 3. Titan-PSD pulsed power systems based on (a) spark gap, (b) thyratron, and (c) magnetic switching.

Table 4. Key operating parameters in European HPM test facilities

HPM facility	Power density, W/cm^2	Range, m	Frequency, MHz	Maximum pulse width, ns	Maximum repetition rate, Hz	Shots per burst	Source type
UK—Orion	>120	100	1,070–3,300	100	100	100	Magnetrons
France—Hyperion	>1,000	~100	700–3,300	300	1	—	Reltrons and magnetrons
Germany	>170	15	700–1,450	300	10	100	Reltrons
Sweden[a]	~120	15–25	L-Ku	500–5,000	1,000	Continuous	Reltrons

[a] This facility uses conventional high power klystrons in L-, S-, C-, X-, and Ku-bands, with maximum power 25–0.25 MW.

PFN. The modulator can fire 1,000 pulses in a burst at repetition rates up to 100 pps and produces 500 kV into a 50Ω load. It has been used to drive relativistic magnetrons[6] for HPM effects measurements.

The Compact Linear Induction Accelerator (CLIA) (Fig. 3c) was developed for moderate-repetition-rate HPM source research and development.[1] CLIA operates at 250 pps with 750-kV, 10-kA, 100-ns output pulses. The system uses thyratron and magnetic switching throughout and produces 188-kW average power into the load. CLIA is not actively cooled and so is restricted to burst operation with 5,000 pulses in a burst. CLIA has been used to drive both magnetron and klystron loads. This technology is scalable to kilohertz repetitive operation.

1.1.3. Titan pulse power systems for HPM simulation. In addition to work done supporting domestic RF and microwave DEW research and development, Titan has provided HPM simulation systems for all of the world's major HPM test facilities in Sweden, France, the United Kingdom,[14] and, most recently, Germany. The microwave test parameters and operational capabilities achievable in these international facilities are summarized in Table 4. The types of Titan HPM sources that produce the intense microwave environments are also given. Work continues today upgrading those systems and providing similar hardware for several non-European countries.

2. Most Recent Contributions

Titan-PSD has taken an interest in the full spectrum of developments of advanced subsystems and components that the three U.S. services will require to realize their DEW and electric platform goals. These developments enable many different military missions and cut across many different technologies: advanced capacitors, switches, transmission systems, RF, and optical systems. Some of the pulse power and power electronic subsystem and component developments that Titan is pursuing are summarized in Table 5.

The following subsections describe five electrical subsystems and three components that Titan-PSD is currently developing for DEW applications.

2.1. Titan-PSD development of subsystems for DEW applications

2.1.1. HPM modulator. Titan-PSD has a conceptual design of a high-voltage-pulse power subsystem for a short-pulse HPM system. The given RF pulse output parameters led us to choose a specific tube that is compatible with the constraints of an airborne system, namely the magnetron with a 20Ω impedance and 33% efficiency in a short-pulse mode. Our pulse power design concept is driven by the tube input requirements (20Ω load, 40-ns pulse width, 10 pps) and comprises a low-impedance, water-filled, stepped-pulse forming line (PFL) charged by a fast Marx and discharged through a self-breaking, spark-gap output switch. An advantage to using a liquid dielectric PFL is that its impedance can be easily varied along its length to accommodate variations in load impedance. Input power to the Marx is from two 400-V series strings of lithium ion batteries with solid-state switched intermediate voltage multipliers providing ±40 kV. Batteries are recharged from aircraft power through a small voltage converter.

Figure 4 is a simplified circuit diagram of the pulse power modulator coupled to a 20Ω constant impedance load though a transit-time isolator. The transit-time isolation is integral to the two parallel cables that couple the output of the pulsed power modulator to the microwave source. The Marx stores 1,140 J at a voltage of ±40 kV. It erects to a 1,280-kV open circuit voltage after a trigger pulse is applied to several of its switches from the trigger generator. Near peak PFL voltage, the self-firing spark gap closes, and energy is transferred to the tube at a peak voltage of 640 kV.

Figure 5 shows the voltage and current in the load computed from a more detailed version of the circuit depicted in Fig. 4. Note that the current is multiplied by a factor of 10. Voltage rises in about 11 ns (10–90%), and the width at the peak is approximately 40 ns. We recognize that the impedance of the magnetron can vary during RF generation, but the constant impedance model is sufficient here. An advantage to using a liquid dielectric PFL

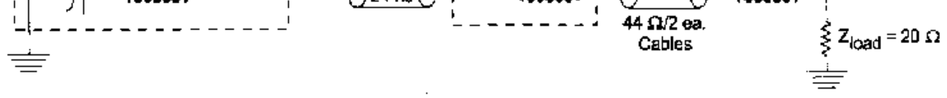
SF6 Insulated 16-stage Marx Generator
PFL 20 Ω w/pkr
24 ns
Self-Break SF6 Gas Switch
Transit Time Isolator/Coupler
44 Ω/2 ea. Cables
Vacuum Bushing Inductance
$Z_{load} = 20\ \Omega$

Fig. 4. Simplified circuit diagram of the pulse power modulator coupled to a 20Ω constant impedance load though a transit time isolator. This impedance represents a magnetron electron tube load.

Table 5. Titan advanced DEW power technology development rationales

What need does the development of pulsed power technology fulfill?		What need does the development of high power/high current technology fulfill?	
It is an enabler for advanced electric weapons and self-defense systems.		It is an enabler for advanced electric weapons and self-defense systems.	
System	User	System	User
Dynamic protection	Army (FCS), Navy (carrier)	Solid-state lasers	Army (Anti-Aircraft and Anti-Missile), Navy (ASMD)
ETC	Army (FCS), Navy (shore bombardment)	HPM	Army (FCS), Navy (non-lethal area denial)
HPM/electromagnetic pulse	Army (FCS), Navy (non-lethal area denial and ASMD), Air Force (SEAD)	EMALS	Navy (carriers)
EM gun	Navy (shore bombardment)	Electric arresting gear	Navy
UWB	All services (counter C4I)	Ride-thru	All services
		It is an enabler for electrically driven platforms: • FCS platforms for the Army • Combatants such as DD21 for the Navy • More electric aircraft for the Air Force	
It is an enabler for commercial spin-offs: • Environmental remediation systems • Material surface treatments • Pulsed thrusters for satellite station keeping		It is an enabler for commercial spin-offs: • Electric vehicles • Commercial power distribution and control • Advanced semiconductor lithography	
What must we accomplish at the subsystem level? • Develop compact, reliable, affordable PFNs • Develop compact, reliable, affordable modulators • Develop compact, reliable, affordable intermediate energy stores		What must we accomplish at the subsystem level? • Develop compact, reliable, affordable prime power, power conversion and controls • Develop compact, reliable, affordable intermediate energy stores • Develop compact, reliable, affordable ride-thru systems	
What must we accomplish at the component level? • Develop high-energy-density, fast-discharge capacitors • Develop long-life, reliable, high-current switches • Develop compact inductors for pulse control • Develop affordable ultracapacitors • Develop compact, reliable, efficient, high-average-power HPM sources		What must we accomplish at the component level? • Develop high-current-density, high-voltage stand-off solid-state switches • Develop affordable ultracapacitors • Develop compact, reliable, efficient, high-average-power HPM sources	

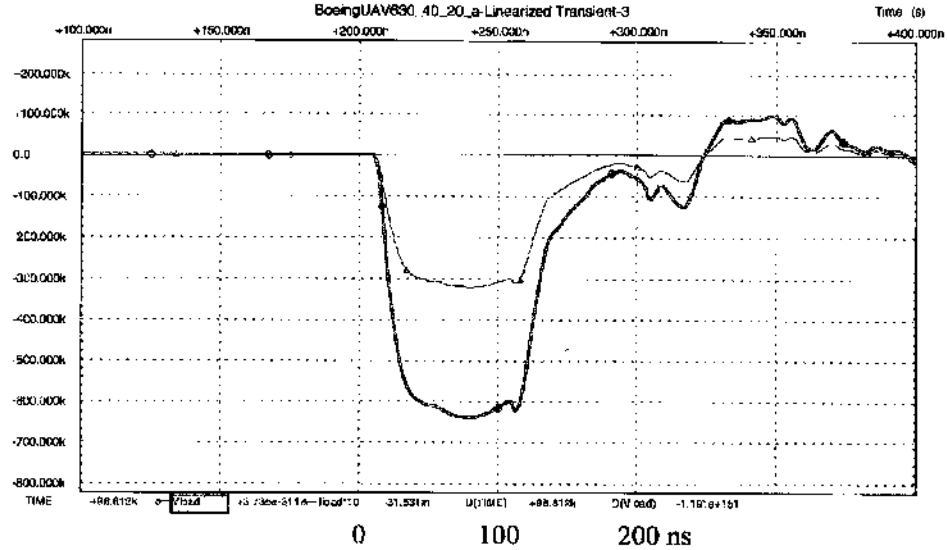

Fig. 5. Voltage and current (multiplied 10×) in the magnetron load computed using detailed version of the simplified circuit diagram in Fig. 4.

is that its characteristic impedance can easily be varied along its length to accommodate variations in load impedance.

Our compact pulse power system conceptual design mounts all of the subsystems and components in a frame that is 24 in. wide × 24 in. high × 144 in. long. The total weight of the system, including the frame, sealed enclosures, and internal mountings, is 1,420 lb. Figure 6 is a dimensioned diagram of the subsystem that is designed to fit into the port bay of an unmanned aerial vehicle (UAV) with the source in the starboard bay. Frames and supports have been omitted for clarity.

2.1.2. Compact PFN. Titan-PSD has designed, fabricated, and delivered the pulse power subsystems for most of the major electric gun facilities worldwide. These facilities were dedicated to electric gun research and for this reason were designed with a premium on reliability and ease of use of the pulse power subsystem. Compactness was a secondary consideration given the large amount of real estate available for the subsystem.

The 8.5-MJ stored energy subsystem, consisting of four separate PFNs (Fig. 1c), was the first design to use the high-energy-density, 2.5-J/cc (2.5-MJ/m^3) polyvinylidene fluoride (PVDF) Aerovox capacitors; the overall energy density of the PFNs is less than 0.5 MJ/m^3. This low energy density is partially due to the use of "jelly-roll" inductors whose large external fields forced the separation of components and resulted in a low component packing fraction.

Figure 7 is a photograph of a 250-kJ PFN that was constructed under a DTRA/Navy-sponsored ETC gun development project and that was incorporated for a time into the Army/Tank-Automotive and Armaments Command (TACOM) Combat Hybrid Power System (CHPS) system integration laboratory.[5] Titan-PSD designed this PFN module for compactness by developing innovative bus work and a closed-field inductor for use with the 2.5-J/cc PVDF capacitors. These advances in component design allowed us to achieve a

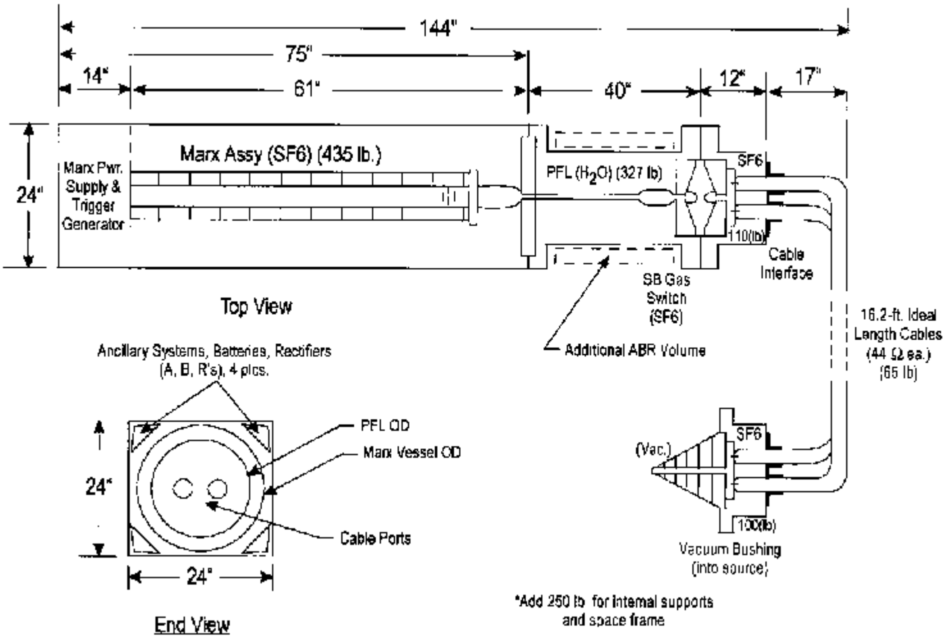
144"
75"
14"
61"
40"
12"
17"
24"
Marx Pwr. Supply & Trigger Generator
Marx Assy (SF6) (435 lb.)
PFL (H_2O) (327 lb)
SF6
110(lb)
Cable Interface
SB Gas Switch (SF6)
Additional ABR Volume
Top View
16.2-ft. Ideal Length Cables (44 Ω ea.) (65 lb)
Ancillary Systems, Batteries, Rectifiers (A, B, R's), 4 plcs.
PFL OD
Marx Vessel OD
Cable Ports
24"
24"
End View
(Vac.)
SF6
100(lb)
Vacuum Bushing (into source)
*Add 250 lb for internal supports and space frame

Fig. 6. Top and end views of the pulse power subsystem concept design. The upper portion will be mounted in a frame and placed in one bay of a UAV. The vacuum bushing and the source will be placed in a second bay.

very high packing factor and an overall energy density of 1.25 MJ/m^3. This module designed in 1994 still represents the state of the art in PFN design. Minor variants to the design are under consideration today for EM gun applications.

Titan-PSD has estimated the increases in the overall energy density of small (250–500-kJ) PFN modules that could be realized by achieving specific advances in component technologies.[†] This includes the use of an existing, more compact vacuum output switch (developed by Titan-PSD) and existing, more compact diodes for the crowbar switches. We would design a higher-energy-density, closed-field inductor and eliminate the dump resistors and associated hardware. Rather than having to dump energy into a resistor, advanced pulsed power systems would be charged by four-quadrant converters capable of returning energy from the PFN to an intermediate energy store. When 5-J/cc film or ceramic, high-energy, reversal tolerant capacitors become available, the utilization of advanced components would lead to a PFN design with an energy density of 3.25 MJ/m^3.

2.1.3. Power supplies for laser applications. Titan-PSD's San Diego operations are actively engaged in building innovative high-average-power systems. Titan is currently delivering power conditioning and control elements for the high-power RF klystrons that will power the DOE Spallation Neutron Source that is now under construction at Oak Ridge, Tennessee, and a series of multimegawatt rectifiers for the U.S. Navy's Electro-Magnetic

---

[†] Higher energy densities could be achieved in the design of larger modules due to economics of scale.

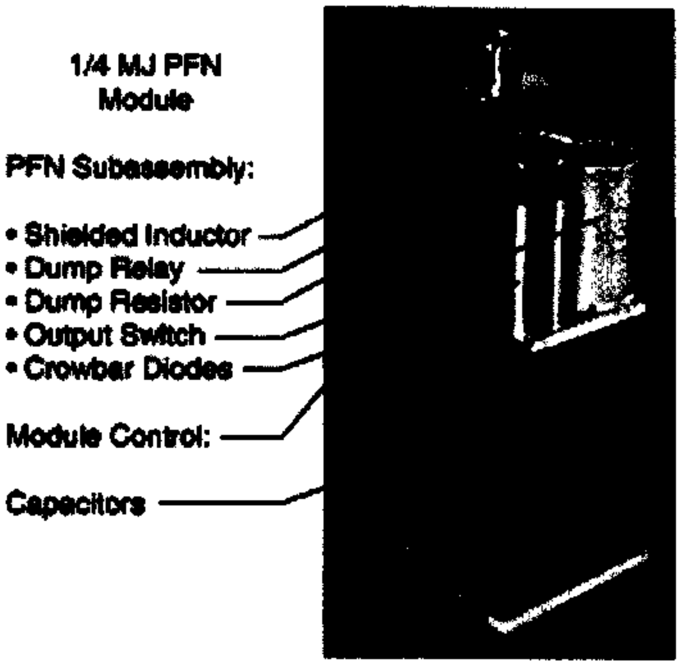

Fig. 7. Photograph of the 250-kJ PFN with the state-of-the-art, 1.25-MJ/m^3 overall energy density. Titan-PSD has operated this PFN into resistive loads at a rate of 1/3 Hz for extended periods. Operational limits were imposed by the heating of the load, not by the performance of any component of the PFN.

Aircraft Launch System (EMALS). Most recently, Titan has developed conceptual compact power supply designs for deployable pulse power systems for the Strategic Illumination Laser (SIL). The SIL is a diode-pumped, solid-state laser for use as a designator as a part of the U.S. Air Force's Missile Defense Agency's Airborne Laser (ABL). These power supplies will drive compact arrays of light-emitting diodes (LEDs), which will in turn pump the solid-state lasing medium.

Titan's proposed SIL power supply system will consist of a group of innovative ballast-type isolated power converter modules, each fed by its own active-power-factor-corrected rectifier module. By this method the input current harmonic distortion will be minimized, resulting in an overall system power factor of 0.995, which will be acceptable for the Boeing-747 type aircraft (ABL platform) power system. The whole power system will be protected against faults by appropriate fuses and circuit breakers. Initial size and weight estimates for the total power system is 65 liters of total volume and 140 kg of mass. The driving factor in the weight and volume is the need to meet the power factor requirements. Given a dedicated power source, the size and weight could be halved.

Two basic topologies have been examined for power factor correction. They are the "Vienna rectifier,"[8] first proposed by researchers in the Vienna Technical University, and the hex-bridge bidirectional inverter. Both topologies do not contain any magnetic components, operating at line frequency, and both can provide a power factor over 0.99. All merits and drawbacks of both topologies have been initially analyzed, and the Vienna approach is favored for this application.

For the ballast inverter Titan-PSD selected the variable-frequency, zero-power switching full bridge quasi-resonant inverter topology. It is new and until recently not well known but is very promising for multikilowatt power supplies. It shows excellent average to peak ratio for the semiconductor switches, close to the theoretical limit of 50%, and a good power transformer copper utilization factor, typically 90%. It is intrinsically output current limiting and therefore operates well in a current regulated mode and is tolerant to output short circuits. Using this converter topology allows the design to reach unprecedented power conversion efficiency (real value 95%) with excellent reliability.

2.1.4. Electronically reconfigurable battery. We consider a DEW power system in which a 10-kV capacitive store is required to operate repetitively. If the system operates at 3 pps, even for a short period of time, the recharge time would be about 300 ms and the recharge power would be about 500 kW_{avg} (1 MW_{pk}). An important system issue is the size and weight of the 500-kW power conditioning unit.

Titan-PSD has developed a means to eliminate the need for intermediate power conditioning for many DEW systems mounted on hybrid electric vehicles by using the on-board batteries to charge the capacitive store directly. Our approach uses electronic switching to convert the battery modules in the mobility store from a parallel configuration providing vehicle load leveling and/or silent mobility capability, to a series configuration that delivers 500 kW at 10 kV to recharge the capacitor bank and back again. We refer to this as an "electronically reconfigurable battery" (ERB). Note that all of the batteries are recharged in parallel. A similar approach is used in capacitive Marx generators to achieve very high voltage pulsed output.

Figure 8 is a schematic drawing of the circuit of an ERB in a hybrid electric drive vehicle. In this schematic circuit, some battery modules are always in parallel and connected to the vehicle bus. Others, along the left-hand side of the drawing, can be switched between a parallel configuration, which supports the vehicle bus, and series operation for charging the DEW store. Electronic reconfiguration requires three switches per battery module. All switches, with the exception of the output switch, need only block the voltage of a single battery module and open at near zero current (characteristic of capacitor charging) in normal operation. Passive snubbing controls the transient conditions during erection and de-erection. Reconfiguration can be accomplished in less than 1 ms using off-the-shelf, solid-state switches such as IGBTs or MOSFETs.

Figure 8 shows an erectable battery module charging a Guillemin E-type network. Switches 2 and 3 are battery isolation switches, and switch 4 is a series switch. These switches are rated only for the module voltage (1 kV in this case). Mechanical contactors parallel the isolation switches for better efficiency during periods when the pulse power system is not in use. Switch 5 is the output switch, which is rated for the full output of the ERB (10 kV and 50 A) and is most likely a series stack of the same switches used for 2–4. For fault protection and charge interrupt at current, a vacuum contactor and fuse are placed in series with the output switch. All switches are optically isolated with gate power drawn from their adjacent modules. Switch 6 is a high-voltage and current-closing switch, either vacuum or solid state.

Assuming a 20-ton-class vehicle and extrapolating from CHPS requirements, we have developed a conceptual design of an ERB for a hybrid electric vehicle with a system meeting the following requirements: 1) Deliver up to 400 kW to the vehicle bus at 1 kV in parallel operation; 2) recharge a 150-kJ capacitive store to 10 kV in 300 ms; and 3) support 30–45-min silent operation at 80 kW. These requirements mandate the use of very high energy

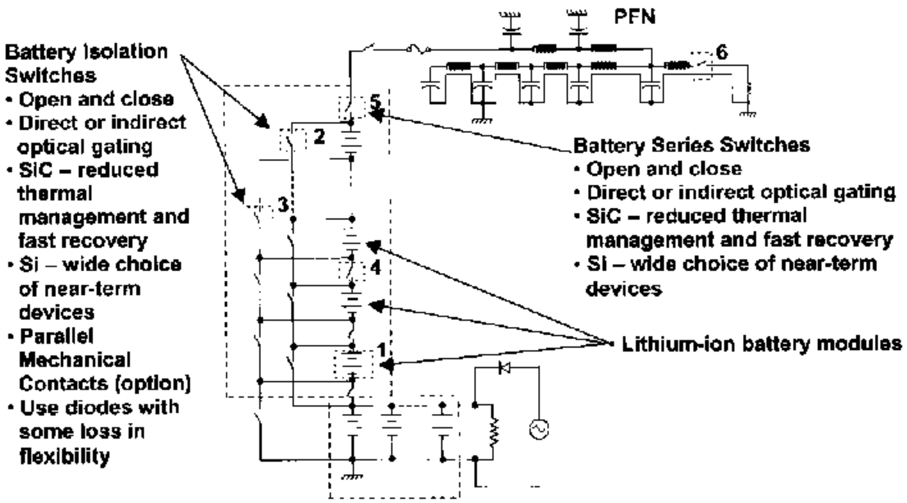
PFN
6
Battery Isolation Switches
• Open and close
• Direct or indirect optical gating
• SiC – reduced thermal management and fast recovery
• Si – wide choice of near-term devices
• Parallel Mechanical Contacts (option)
• Use diodes with some loss in flexibility
5
2
3
4
1
Battery Series Switches
• Open and close
• Direct or indirect optical gating
• SiC – reduced thermal management and fast recovery
• Si – wide choice of near-term devices
Lithium-ion battery modules

Fig. 8. Schematic circuit of the ERB. Shown is a generic circuit topology for charging a Guillemin E-type network. For the present application, the 10 modules in the dynamic store are all erected in series to produce a 10-kV output. A feature of the ERB is that the dynamic portion of the store may be erected in combinations of series/parallel modules to provide any output voltages that are an integer multiple of the DC bus voltage. This voltage agility, in theory, would allow a single dynamic store to power several loads with different input voltage requirements on the same platform such as ETC guns and HPM systems.

and power density batteries such as the HP series of lithium ion batteries being developed by SAFT and the lithium polymer batteries produced by Ultralife Batteries, Inc., for use in cell phones. The SAFT batteries have a slight advantage in usable power density and packaging for military use, whereas the Ultralife batteries have an advantage in cost due to volume production and a potential for more compact packaging (thin, rectangular) in ERB service. A design and trade study will select the most appropriate cell for the application. We chose the Ultralife cells for the conceptual design used for purposes of discussion in this paper because of our experience with them in our laboratory tabletop ERB and because of the availability of these cells over a range of capacities from 120 mA-h to 3.5 A-h.

The ERB system in this case constitutes only 1/3rd of the total battery for erected (dynamic) operation. The remaining 2/3rds of the store (static) is dedicated to load leveling and silent mobility. The total capacity of ~290 MJ (80 kW-h) accommodates silent mobility requirements. With only 1/3rd (90 MJ, 25 kW-h) of the total capacity configured for on-command electronic erection and de-erection, the vehicle energy storage system maintains its load leveling and silent operation capability even when the pulse power system is active.

For our conceptual design, we assemble 53 Ultralife polymer batteries into individual 200-V stacks. Five of these stacks are placed in series to obtain an output voltage equal to that of the vehicle bus. In the static portion of the store, 20 such series stacks of the UBC34106102 cells (5,300 total cells) are needed to provide the 200-MJ nominal capacity. In the dynamic portion of the store, the UBC383562 cell is used because of its higher current rating. Due to the smaller size of these cells, there can be as many as 4 in parallel by 53 in series per 200-V unit, and five such units will form an erectable module. The module will incorporate all necessary switches and isolation and thermal management hardware.

Ten modules will constitute the entire dynamic store (~90 MJ, 10,600 total cells). The total battery volume in the static and dynamic stores is ~0.282 m^3, and the accessories are expected to add another 0.4–0.6 m^3 for a total volume of less than 1 m^3. The weight for just the batteries would be ~572 kg, with accessories of much lower density adding another 100–200 kg, for a total weight of less than 700 kg. Charging of the ERB is always performed at bus voltage with the dynamic store in parallel configuration.

In load leveling or silent mobility operation, the charge state of the ERB is continuously monitored and controlled by the vehicle's systems. When the pulse power mode is activated, the mechanical contactors bypassing the isolation switches are opened. Next, all isolation switches are opened, followed by the closing of all series switches, and the ERB begins charging the capacitive store. Initially, the load looks like a short circuit, and the charging current is limited by the internal impedance of the ERB cells. As the charge on the bank increases, the charging current diminishes, and the cell voltage begins to rise. Near full charge, the current approaches zero and the ERB output voltage rises to its maximum of ~10 kV. Voltage on the store is monitored and is regulated by chopping the ERB output (i.e., opening and closing the series and output switches as necessary to maintain charge against leakage; this has been demonstrated in the laboratory at rates of up to 1 kHz). When the store discharges, the output switch opens immediately and remains open long enough for the closing switch to recover (1–10 ms). The ERB can remain erected for as long as the pulse power system is active. Recharging of the store is then initiated by reclosing the ERB output switch. When the pulse power system is deactivated, the dynamic store is returned to its parallel configuration by opening the series switches (de-erecting). The dynamic store can then be recharged from the vehicle power bus.

We have created an ERB laboratory demonstration consisting of five erectable modules of 40 V. The Ultralife UBC383562 cells were used and arranged on PC boards that were in turn stacked via standoffs. Isolation and series switching was via optically coupled IGBTs. Some figures of merit (FOM) from the ERB testing include a demonstrated current rise time to 80 A of 4 μs at 200 V and an erection/de-erection frequency of 1 kHz for use in voltage regulation. A FOM from the four series cell tests is a demonstrated current rise time to 120 A of 1 μs, in a maximum power transfer configuration.

2.1.5. Intermediate energy storage. In the preceding section Titan-PSD identified a need for and established the technical feasibility of an advanced technology power conditioning element to charge capacitive stores in DEW systems. The ERB is the solution with the highest efficiency and energy and power density possible. This is critical in order to realize a manageably sized DEW system. A second similar, critical issue in the design and use of pulsed electrical weapon systems involves the level of available platform power. If this power is insufficient to support the required operational burst length and repetitive fire rate of the system, then it may be necessary to incorporate intermediate energy storage into the system design. Intermediate storage could be practically achieved through the use of secondary batteries, flywheels, and ultracapacitors (also referred to as electric double-layer capacitors or EDLCs). Figure 9 is a schematic diagram of intermediate energy stores in two DEW systems.

Figure 10 shows a present application of an EDLC energy storage system in providing ride-through of voltage sags and momentary outages for an induction motor. The output of the store is connected across the direct current (DC) link of the asynchronous drive and supports its voltage. Titan-PSD has manufactured a ride-through system that has been demonstrated to respond to sags and outages within 10 ms. It can provide

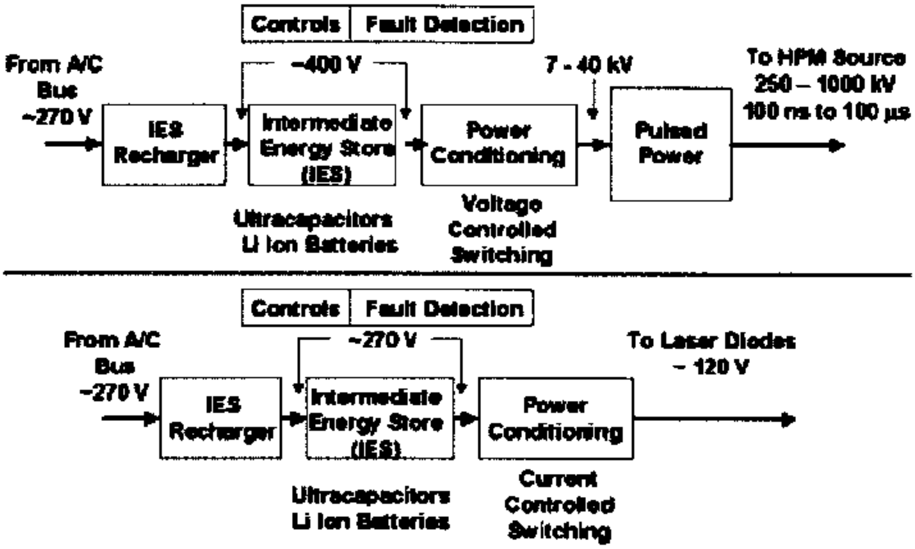
Controls
Fault Detection
From A/C Bus ~270 V
~400 V
7 - 40 kV
To HPM Source 250 – 1000 kV 100 ns to 100 μs
IES Recharger
Intermediate Energy Store (IES)
Power Conditioning
Pulsed Power
Ultracapacitors Li Ion Batteries
Voltage Controlled Switching
Controls
Fault Detection
From A/C Bus ~270 V
~270 V
To Laser Diodes ~ 120 V
IES Recharger
Intermediate Energy Store (IES)
Power Conditioning
Ultracapacitors Li Ion Batteries
Current Controlled Switching

Fig. 9. Two DEW block system schematics with intermediate energy storage.

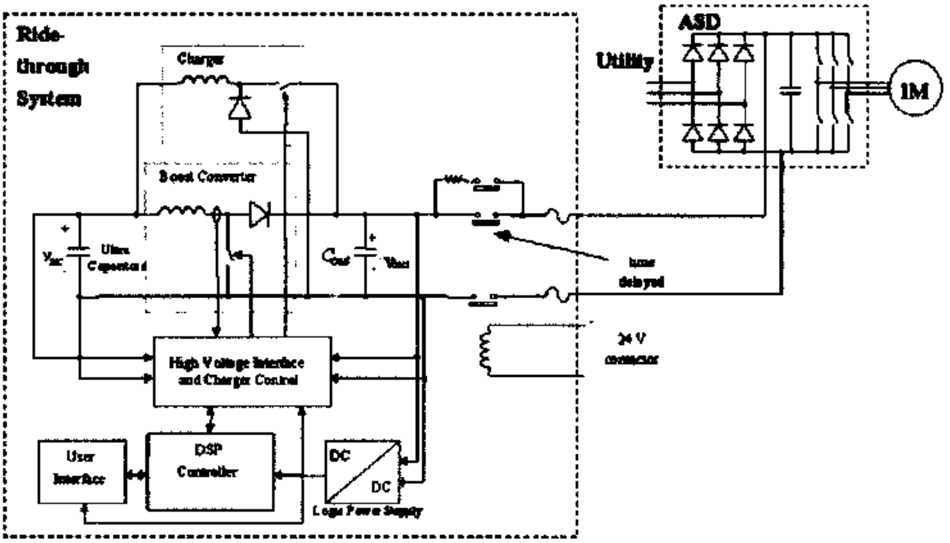
Ride-through System
Charger
Boost Converter
V_{uc}
Ultra Capacitor
C_{out}
V_{out}
fuse delayed
24 V contactor
High Voltage Interface and Charger Control
User Interface
DSP Controller
DC
DC
Logic Power Supply
Utility
ASD
IM

Fig. 10. Circuit diagram of an EDLC ride-through system supporting the DC link of an asynchronous drive (ASD).

100 kW for up to 5 s, which is sufficient to allow for orderly shutdown in a prolonged outage.

Titan-PSD has also developed several concept designs of EDLC energy storage systems for military applications. For example, one EDLC system design would be capable of supporting the DC link in a Navy shipboard system at 480 V for 12.5 s at 250 kW. This system is estimated to be 50% smaller and lighter than a corresponding flywheel system (i.e., the Caterpillar UPS-250 system). With minor modifications, the same system (or several systems in parallel) could be used as an intermediate energy store for DEW systems.

2.2. Titan advanced component developments to enable DEWs

2.2.1. Solid-state switches for severe service. The key enabling technology for nearly all pulse power systems for DEW applications is the switch that initiates and delivers the pulsed electrical output from the electrical energy store to the load. This switch must hold off high voltage (typically ≥ 10 kV), close rapidly (< 1 μs), carry high peak current (100–500 kA) with high initial rate of rise (50–100 kA/μs), sustain a high total action ($\geq 10^7$ A^2-s), and recover rapidly (< 10 ms). In addition, the switch must support high current density, operate reliably under mechanically stressing conditions, have minimum weight and volume claims, and not require ancillaries such as vacuum pumps or gas flow subsystems. At present, the hydrogen thyratron is most frequently selected to meet these simultaneous requirements, but thyratrons are large and heavy devices, with auxiliary equipment that consumes several kilowatts and often require immersion in oil. The Army Research Laboratory has been actively seeking solid-state devices for DEW utility[7] and Titan-PSD is weighing in by developing a high fluence, laser gated and pumped (silicon) thyristor (LGPT) that will satisfy these requirements.

High fluence optical gating of thyristor switches has been chosen for numerous pulsed power switching applications. Light activation with high fluence laser sources enables the use of solid-state switching devices in performance parameter spaces previously serviced only by vacuum and gas spark gaps. If the output from the high fluence source is properly distributed within the switching medium, the di/dt capability of the switch does not depend on the rate of plasma spreading. Rather, di/dt becomes a function of the rate of rise of the optical fluence. In addition, optical gating eliminates the extensive and intrusive physical gate structure that limits the conduction area of the highest di/dt conventional solid-state devices. Properly designed laser-gated devices are typically capable of higher di/dt, peak current, and charge transfer for a given silicon area than conventional solid-state devices.

The band edge for intrinsic silicon is ~1.12 eV, corresponding to ~1,111 nm. However, silicon is an indirect band-gap material so that the optical phonon energy (0.063 eV) adds to the photon energy to create e–h pairs out to 1,170 nm. In most previous efforts involving direct laser-gating of silicon devices, a solid-state laser was used; typically either Nd:YAG or Nd:YVO at 1,064 nm. However, over the past 10 years, the development of compact diode-pumped solid-state lasers has significantly reduced the size of the requisite optical sources and improved their lifetime and reliability. More important, the InGaAs laser diode bars that were developed for pumping these solid-state lasers have been pushed to longer wavelengths, approaching and exceeding that of the silicon band edge. In principle, the center wavelength of InGaAs laser diode bars is selectable by design over a continuum from 850 nm through at least 1,150 nm. This allows direct optical gating and pumping of silicon by the diodes without any of the intermediary options of the preceding paragraph. Figure 11 is an illustration of an on-board laser diode gated and pumped silicon thyristor, which is one of the candidate approaches in development now.

Titan-PSD is currently performing on two programs and has finished a third that provide technical leverage for further switch development. Our Electra Advanced Pulsed Power Program is a five-year, Naval Research Laboratory (NRL)/DOE-sponsored effort to develop advanced pulsed power components and systems for KrF laser IFE (krypton fluoride laser inertial fusion energy). The most critical component is the primary switch, and we are

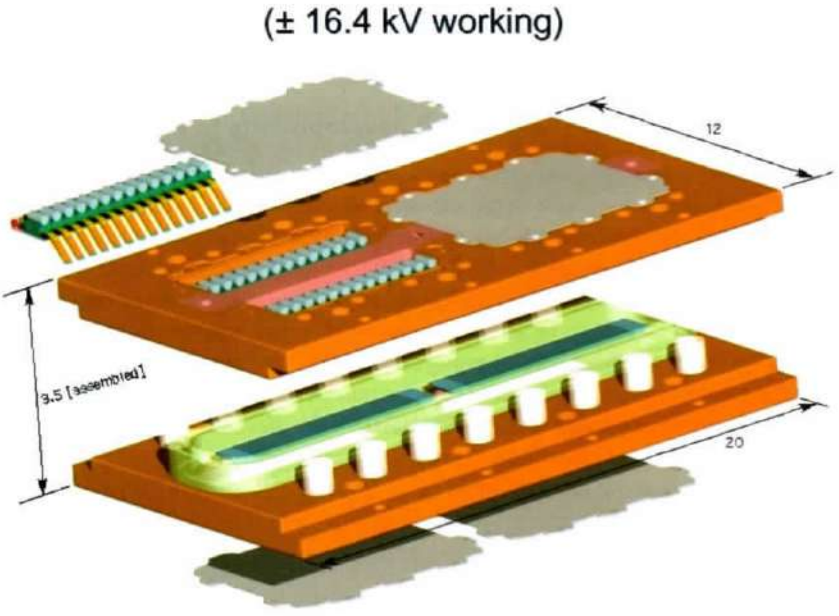

Fig. 11. Illustration of a laser gated and pumped thyristor.

developing an on-board LGPT for this role. The switching requirements are repetitive, high peak current and di/dt, as well as high efficiency (the switch dissipates $<1.5\%$ of stored energy) and long lifetime (10^9 shots). The LGPT is to be used in a Marx generator, and the single device operating parameters are 16.4-kV working voltage, current density 2.25 kA/cm^2, peak current 225 kA, max di/dt 900 kA/μs, pulse width 800 ns, and repetition rate 5 pps continuous. The lifetime requirement limits the thermal excursion per pulse to a few degrees, and therefore the action per square centimeter is low ($\sim$2 A^2-s).

Another current Titan-PSD program is to develop compact pulsed power components for the Air Force Research Laboratory (AFRL) at Kirtland Air Force Base. Components under development enable a compact, long-lifetime Marx-PFN driver for a narrowband HPM source. The di/dt requirement is a few hundred kA/μs/cm^2(Si) or 2 MA/μs total, with a peak current of 30 kA. To satisfy the switching requirements for this application, we are considering the LGPT, a solid-state laser activated switch, and a hybrid combination of the two.

Although the action and charge transfer in these Electra and AFRL applications are substantially less than that found in some pulse power applications, they serve to illustrate our confidence in the high power laser diode bars for high di/dt service. In both cases, the pulse width of the energy transfer is short enough that thyristor action has barely begun by the time the pulse is over. Because of this, a large fraction of the total charge conducted during the pulse must be supplied optically. The diode laser bars and their drive circuitry are capable of supplying an illumination fluence of up to 6 kW/cm^2(Si) within 20 ns and sustaining that level for at least several microseconds. Fluence of several kilowatts per square centimeter is desirable for initial gating in many DEW applications. The required fluence drops to 250–500 W/cm^2 for continuous pumping.

Another effort was recently undertaken in a conceptual design study for an optically gated and pumped switch for rectification of compulsator (pulsed alternator) output for driving an electromagnetic gun. This completed study was performed by Titan-PSD and the Optiswitch Technology Corporation for the Institute for Advanced Technology (IAT). The on-board LGPT design was chosen because of the customer's desire to reduce the overall device count and silicon area compared with conventional devices. Specifically, we designed a 16.7 kV, 100–115-cm^2 asymmetric (blocks voltage forward direction only) thyristor and series diode. Depending on the allowable temperature excursion per shot, our approach replaces ~80 conventional devices with 3–5 series pairs (6–10 devices total; 340–500 cm^2 active) for the same forward losses. This is possible because continuous pumping of thick (2.5-mm), high-voltage (15–20-kV) devices with photo-carriers mitigates the carrier population deficit in the n-base. Electrically gated devices of the same voltage capability typically exhibit huge commutation and conduction losses and are therefore impractical for this type of service. Reverse blocking was handled by a series optically pumped diode. Separating the forward and reverse blocking functions allowed more freedom in designing for reverse recovery, a feature deemed necessary due to repeated failures of symmetric (blocks voltage both directions) devices in this service. Laser diode pumping served to reduce the forward losses in the series diode element as in the thyristor.

The switch was designed to handle a peak current of 1.6 MA (3.3–4.7 kA/cm^2) for 600 μs when installed in a half-wave, four-phase rectifier. Nine total cycles are rectified per shot of the electromagnetic gun. The initial design (500 cm^2 active) limited the temperature excursion per shot to ~100°C, commensurate with a service life of ~10^4 shots. In the extreme, the current density could be pushed to 4.7 kA/cm^2 and optical pump power doubled, resulting in a ΔT approaching 180°C and ~200-shot lifetime. The laser diode bars in the initial case provided a constant 500 W/cm^2 of silicon for each 600-μs conduction phase. This goes to 1 kW/cm^2 in the extreme case.

2.2.2. Next generation reltron microwave generator development. Reltron HPM sources were invented and developed by R. Bruce Miller of Titan Advanced Innovative Technologies from the mid-1980s through 1998. The PSD of Titan assumed responsibility for reltron contracts in 1998. Projects included one complete HPM simulator for a government research laboratory in Germany and eight reltron tubes of various specifications. In developing these eight tubes, enough mechanical and electrical modifications have been included to justify their designation as "second-generation" designs.

Reltrons are both compact and efficient. The electron beam is highly modulated over ~10 cm and postaccelerated prior to drifting to the output cavities and beam collector. This beam transport requires only ~75-cm overall length for L-band tubes (Fig. 12). The output cavities are incorporated into a standard rectangular waveguide. Microwave power is efficiently extracted (~80%) directly in TE_{10} mode. The overall efficiency (peak electrical to microwave power) is 30–40%.

In general these new tubes move the reltron technology base in the directions of longer pulsewidths (~1–2 μs), increased frequency tunability (~±13%), lower driver voltages (500–600 kV), and higher repetition rates (~10 Hz for demountable versions, ~300 Hz for sealed tubes). Second-generation improvements include a new high voltage insulator design, new movable current contacts for the modulation and output cavity tuners, reduced outgassing, and monolithic grids. The frequency range for these eight tubes covers 700–1,450 MHz. Programs are in place now to extend this range.

In addition to the objective to extend the frequency coverage, efforts are underway to improve beam optics to reduce current losses and thereby improve efficiency and microwave output power. A new beam current diagnostic package has been recently implemented to

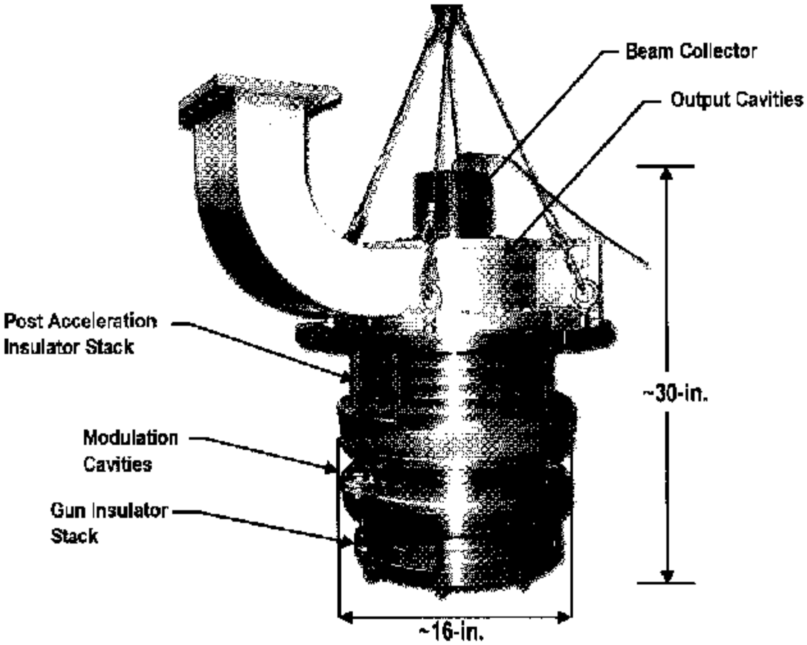

Fig. 12. L-band reltron capable of generating flat 200-MW, 900-ns pulses with frequency stability $\delta f/f < 0.3$. This generation of reltrons is capable of RF pulse widths of ~300 RF cycles, repetitive pulse operation from single-shot to 10 Hz, and ±10% continuous-frequency tunability.

better infer basic performance parameters and to assist in comparing benchmarking simulations. These calculations simulate performance in the actual geometry and are generated with a three dimensional particle-in-cell code (the LSP code from Mission Research Corporation). Measured waveforms generated by a recently delivered L-band tube are shown in Fig. 13. This tube produces ~185 MW in flat, 900-ns pulses with excellent frequency stability. The LSP code was essential during this tube's development. A new era in microwave tube design has emerged with the development of accurate three-dimensional models.

2.2.3. HEM magnetron. Backdoor microwave effects on electronics systems in modern weapons fall into three broad categories. Two categories are characterized by high peak power density and microwave formats that overwhelm the target system with an electric field that is high enough to cause circuit upset or burnout. In the third category, effects are caused by coupling spurious signals into the target system, akin to classical smart jamming, interference or spoofing[‡] techniques used by the EW countermeasures community. The DEW microwave pulse formats appropriate to the latter class of effects are generally distinguished from upset and damage pulse formats by lower peak power density thresholds but more specific requirements on frequency and amplitude modulations and

---

[‡] Low incident microwave levels can cause "interference" with the operation of threat weapon electronic systems, producing high levels of noise in sensitive circuitry to "mask" true signals. The incident microwave signal is said to "spoof" if it is tailored to "mimic" the true signal and to thereby produce false commands.

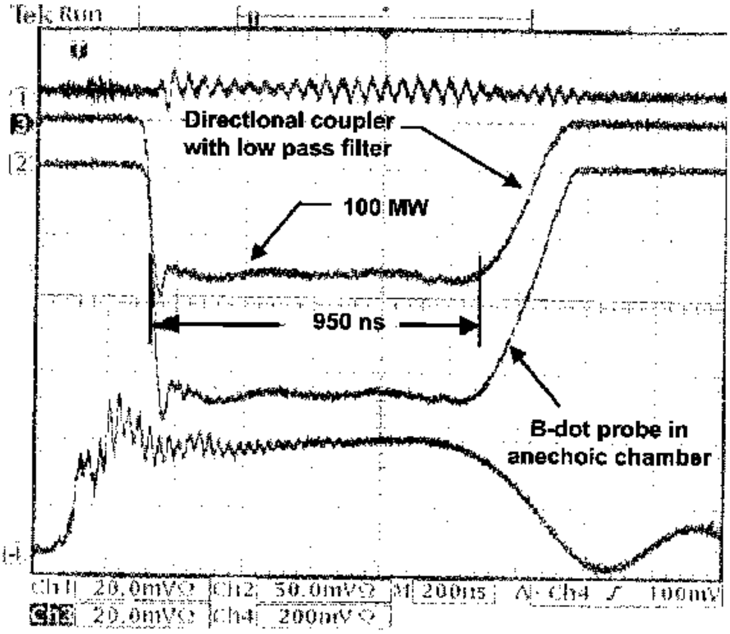

Fig. 13. Waveforms from the 185-MW L-band reltron: the top trace is the heterodyne (l.o. = 1.3 GHz) showing excellent frequency stability, and the second trace is the steady, flattop output power measured with a calibrated directional coupler. (The data are conditioned with a low-pass filter; higher harmonic content is know to be more than 13 dB below the primary L-band signal.) The third trace is the radiated power monitored with an uncalibrated, free-field B-dot, and the last trace is an uncalibrated profile of the Marx current.

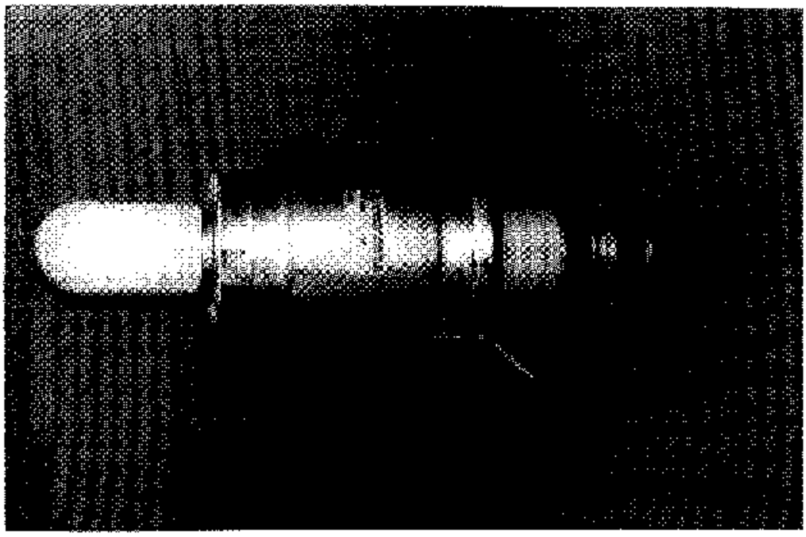

Fig. 14. Titan's L-band HEM magnetron. This is the highest-average-power magnetron that is commercially available.

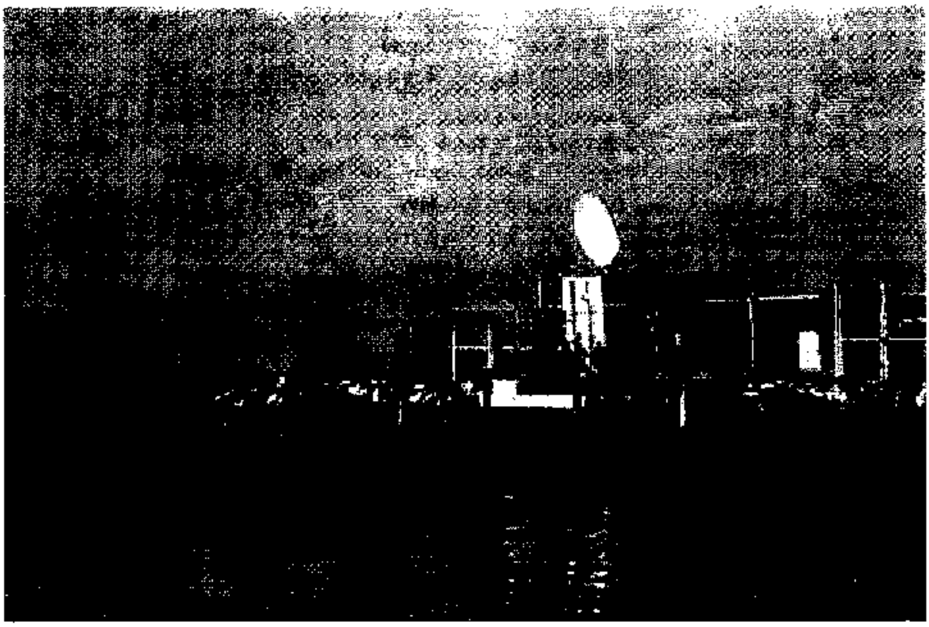

Fig. 15. The principal system elements of the ATD system are a power supply, a modulator, an HEM source, and an antenna. These components are supported by a control system and a cooling system. For the DTRA ATD ground-based system the power supply and operators' console(s) are housed in a shelter mounted on one flatbed trailer while the remaining components are mounted on another flatbed trailer (out of the field of view to the right). The system was designed to broadcast 2-MW peak power pulses with an average power of 300 kW for 30-s missions. This yields a peak fluence of 1 mW/cm^2 at a range of 2,400 m.

repetition rate. All three categories of effects can be induced by DEW systems based on high-power RF, narrow-band, pulsed HPM or long-pulse, high-energy microwaves (HEMs).

Titan has developed an HEM laboratory source and an HEM advanced technology demonstration (ATD) source for DEW purposes at its California Tube Laboratory. These HEM sources have efficiencies of approximately 85%. This high efficiency minimizes prime power and cooling requirements, thus significantly minimizing overall DEW system size and weight. The HEM magnetron, shown in Fig. 14, has produced 900-kW peak and 150-kW average powers and can easily be upgraded to 300-kW average power. The unique modulation characteristics of this magnetron can produce modulation effects of 30 MHz or greater in the target systems. More operational details are given in Table 6.

The ATD system shown in Fig. 15 combines the outputs of two 150-kW-average-output power magnetrons. The maximum peak output power obtained to date is 1.2 MW. The magnetrons with their respective magnet yokes each weigh about 200 lb, and each takes up about 4 ft^3. When combined their efficiency is over 76%, therefore together they absorb only 95 kW and are actively cooled. Variants of this RF system are suitable for Army, Navy, and Air Force missions.

Table 6. California tube laboratory HEM magnetron operational specifications

Parameter	Value
Frequency, fixed	915 MHz (890–920 MHz possible as well)
Power output	
Continuous wave	300-kW minimum
Pulse	600-kW peak (900 actually achieved) 300-kW avg (nominal 50% duty)
Pulse width	10s of microseconds to 5 ms
Peak anode voltage	42 kV
Peak anode current	17 A
Beam efficiency	88% typical
Water-cooled anode	20 gpm typical
Filament	12 VAC, 200-A typical standby

3. Conclusions and Significance

The first laser and HPM DEWs are just now being deployed, and new threats emerging in the information age will sustain the need for continued DEW development and acceptance. The deployment of even more advanced DEW systems and in fact the DoD's intent to field more (or all) electric fighting platforms depend critically on the availability of compact, lightweight, efficient pulsed power subsystems and long-lived, reliable advanced power conditioning and front-end components.

Titan has been supporting U.S. directed energy programs by providing services and analysis and developing advanced technologies since the early 1980s. Today, Titan is the only U.S. industrial supplier of commercial HPM generators. We have advanced pulsed power and power electronics technology development programs involving intermediate and primary electrical energy storage, solid-state switching, power conditioning, electrical-to-microwave conversion, and novel microwave antennas underway now. Titan is focused on providing pulsed power and C4ISR (command, control, communications, computers, intelligence, surveillance, and reconaissance) solutions for national defense and is intent on maintaining its industrial leadership in military information technologies.

Acknowledgments

The authors recognize and acknowledge contributions made by Jim Benford, Jerry Levine, R. Bruce Miller, and David Blank to the development of the primary narrow-band microwave sources that were the precursors of much of the HPM and HEM technology that Titan offers today.

References

[1] Ashby, S., D. Drury, G. James, P. Sincerny, and L. Thompson, "CLIA—A Compact Linear Induction Accelerator System," *Proceedings 8th International IEEE Pulsed Power Conference*, San Diego, CA, p. 940 (1991).

[2] Bekifi, G., and T.J. Orzechowski, Phys. Rev. Lett. **37**(6) 379 (1976).

[3] Benford, J., D. Price, H. Sze, and D. Bromley, J. Appl. Phys. **61**(5), 2098 (1987).

[4] Benford, J.N., N.J. Cooksey, J.S. Levine, and R.R. Smith, IEEE Trans. Plasma Sci. **21**(4), 388 (1993).

[5] Bhasavanich, D., C.M. Gilman, H.G. Hammon, and K. Nielsen, "80 kW, 1/3 Hz Repetitive-Fire PFN for Electrothermal Launchers," *Proceedings 10th International IEEE Pulsed Power Conference*, Albuquerque, NM, p. 155 (1995).

[6] Hammon, J., S.K. Lam, and S. Pomeroy, "A Transportable 500 kV, High Average Power Modulator with Pulse Length Adjustable from 100 ns to 500 ns," *Proceedings 10th International IEEE Pulsed Power Conference*, Albuquerque, NM, p. 429 (1995).

[7] Kingsley, L.E., R. Pastore, and H. Singh, "Solid-State Power Switches for HPM Modulators," *Proceedings 10th International IEEE Pulsed Power Conference*, Albuquerque, NM, p. 65 (1995).

[8] Kolar, J.W., and F.C. Zach, "A Novel Three-Phase Interface Minimizing Line Current Harmonics of High-Power Telecommunications Rectifier Modules," *Record of the 16th IEEE International Telecommunications Energy Conference*, Vancouver, Canada, p. 367 (1994).

[9] Levine, J.S., N. Aiello, J. Benford, and B. Harteneck, J. Appl. Phys. **70**(5), 2838 (1991).

[10] Levine, J.S., B. Harteneck, and H.D. Price, *Intense Microwave Pulses III*, SPIE Vol. 2557, p. 74 (1995).

[11] Miller, R.B., K.W. Habiger, W.R. Beggs Jr., and J.R. Clifford, "Advances in Super-Reltron Source Development," *Intense Microwave Pulses III*, SPIE Vol. 2557, p. 2 (1995).

[12] Miller, R.B., C.A. Muehlenweg, K.W. Habiger, J.R. Smith, and D.A. Shiffler, *Intense Microwave Pulses II*, SPIE Vol. 2154, p. 99 (1994).

[13] Price, D., D. Fittinghoff, J. Benford, H. Sze, and W. Woo, IEEE Trans. Plasma Sci. **16**(2), 177 (1988).

[14] Price, D., J.S. Levine, and J. Benford, "ORION-A Frequency-Agile HPM Field Test System," *Seventh National Conference on High Power Microwave Technology*, Laurel, MD (1997).

[15] Smith, I.D., D.W. Morton, D.V. Giri, H. Lackner, C.E. Baum, and J.R. Marek, "Design, Fabrication and Testing of a Paraboloidal Reflector Antenna and Pulser System for Impulse-Like Waveforms," *Proceedings 10th International IEEE Pulsed Power Conference*, Albuquerque, NM, p. 56 (1995).

[16] Smith, R.R., J. Benford, D. Harteneck, and H.M. Sze, IEEE Trans. Plasma Sci. **19**(4), 628 (1991).

[17] Sze, H., J. Benford, W. Woo, and B. Harteneck, Phys. Fluids **29**(11), 3873 (1986).

The Tom Bearden Website

TOWARDS A NEW ELECTROMAGNETICS
PART I: SOLUTION TO TESLA'S SECRETS AND THE SOVIET TESLA WEAPONS

Table of Contents

Foreword **NEW**

Starting Pages **NEW**

Soviet Tesla Weapons

References

The Tom Bearden Website

TABLE OF CONTENTS

Acknowledgements

Foreword

Part I

Solutions to Tesla's Secrets and the Soviet Tesla Weapons

References

Part II

Reference Articles for Solutions to Tesla's Secrets

Bottone, A., "Nikola Tesla's New Wireless", *Electrical Engineer*, London, Dec. 24, 1909.

"Tesla's New Device Like Bolts of Thor", *N. Y. Times*, Dec. 8, 1915.

Secor, H. Winfield, "Tesla's views on Electricity and the War", *Electrical Experimenter*, Aug., 1917.

Tesla, Nikola, "The Effect of Statics on Wireless Transmission", *Electrical Experimenter*, Jan., 1919.

Gernsback, H., "Underground Wireless", *Electrical Experimenter*, March, 1919

Secor, H. Winfield, "America's Greatest War Invention, The Rogers Underground Wireless", *Electrical Experimenter*, March, 1919.

Tesla, Nikola, "Tesla on High Frequency Generators", *Electrical Experimenter*, April, 1919.

Tesla, Nikola, "The True Wireless", *Electrical E:x:perimenter*, May, 1919.

Tesla, Nikola, "My Inventions, v. The Magnifying Transmitter", *Electrical Experimenter*, June, 1919.

"The Rogers Underground Aerial For Amateurs", *Electrical Experimenter,* June, 1919.

"A Giant Eye to See Round the World", *Albany Telegram,* Feb. 25, 1923.

Tesla, Nikola, "'World System' of Wireless Transmission of Energy", *Telegraph and Telephone Age,* Oct. 16, 1927.

Secor, H. Winfield, "Tesla Maps Out Our Electrical Future", *Science and Invention,* April, 1930.

Tesla, Nikola, "Man's Greatest Achievement, *N.* Y. *American,* July 6, 1930.

"Tesla at 75", *Time,* July 20, 1931.

Gernsback, Hugo, "Faster Than Light:", *Everyday Science and Mechanics,* Nov., 1931.

"No High Speed Limit, Says Tesla", *The Literary Digest,* Nov. 7, 1931.

O'Neill, John J., "Tesla Cosmic Ray Motor May Transmit Power 'Round Earth", *Brooklyn Eagle,* July 10, 1932.

"Tesla, 76, Reports His Talents At Peak", *N. Y. Times,* July 10, 1932.

Tesla, Nikola, "Pioneer Radio Engineer Gives Views on Power", *N. Y. Herald Tribune,* Sept., 11, 1932.

Bird, Carol, "Tremendous New Power Soon to be Unleashed", *Kansas City Journal-Post,* Sept. 10, 1933

"Tesla 'Harnesses' Cosmic Energy", *Philadelphia Public Ledger,* Nov. 2, 1933.

"Tesla Sees Evidence Radio and Light are Sound", *N.* Y. *Times,* April 8, 1934.

"Invents Peace Ray", *N. Y. Sun,* July 10, 1934.

"Death-Ray Machine Described", *N. Y. Sun,* July 11, 1934.

"Tesla, at 78, Bares New 'Death-Beam'", *N. Y. Times,* July 11, 1934.

Alsop, Joseph W., Jr., "Beam to Kill Army at 200 Miles, Tesla's Claim",
N. Y. Herald Tribune, July 11, 1934

Tesla, Nikola, "Tesla on Power Development and Future Marvels", *N. Y. World*
Telegram, July 24, 1934.

Wilshimer, Helen, "Dr. Tesla Visions the End of Aircraft in War", *Every*
Week Magazine, Oct. 21, 1934.

"A Machine to End War", by Nikola Tesla as told to George Sylvester Viereck,
Liberty, Feb., 1935.

"Tesla Predicts Ships Powered by Shore Beam", *N. Y. Herald Tribune,*
June 5, 1935.

"3 Tesla Inventions", *N. Y. Sun,* July 9, 1935.

Sparling, Earl, "Nikola Tesla, at 79, Uses Earth to Transmit Signals;
Expects to Have $100,000,000 Within Two Years", *N. Y. World Telegram,*
July 11, 1935.

"Tesla's Controlled Earthquake", *N. Y. American*, July 11, 1935.

"Tesla, 79, Promises to Transmit Force", *N. Y. Times,* July 11, 1935.

Tesla, Nikola, "Expanding Sun will Explode Someday Tesla Predicts",
N. Y. Herald Tribune, Aug. 18, 1935.

"Tesla, 80, Reveals New Power Device", *N. Y. Times*, July 11, 1936.

"Sending of Messages to Planets Predicted by Dr. Tesla on Birthday",
N. Y. Times, July 11, 1937.

"Tesla Has Plan to Signal Mars", *N. Y. Sun,* July 12, 1937.

"In the Realm of Science: Tesla, Who Predicted Radio, Now Looks Forward
to Sending Waves to the Moon", *N.* Y. *Herald Tribune,* Aug. 22, 1937.

"Aerial Defense 'Death-Beam' Offered to U. S. By Tesla", *Baltimore Sun,* July 12, 1940.

"'Death Ray' for Planes", *N. Y. Times,* Sept. 22, 1940.

"Proposing the 'Death Ray' for Defense", *Philadelphia Inquirer,* Oct. 20, 1940.

Swezey, Kenneth M., "Nikola Tesla", *Science,* May 16, 1958.

"Cableless Power", *Electronics* (Digest), July 22, 1960.

"Fireballs for Defense?", *Christian Science Monitor,* Feb. 7, 1962.

Lagus, George W., "The Use of Ground in High-Frequency Circuits", *Radio-TV and Science and Electronics,* June-July, 1969.

Jamison, S. L., "Life Energy", *Probe, The Unknown,* June, 1973.

Jueneman, Frederic B., "The First Law of Thermodynamics", *Industrial Research,* Feb., 1974.

Wade, Nicholas, "Charged Debate Erupts over Russian Beam Weapons", *Science,* May 27, 1977.

Benford, Gregory, "Zeus in Orbit", *Omni,* Sept., 1981.

JAMES HARRIS ROGERS' PATENTS

(In order of patent application date)

958,829, Method and Apparatus for Producing High Frequency Oscillating Currents. James H. Rogers. Filed Jan. 20, 1910.

1,220,005, Wireless Signaling System. James Harris Rogers and Henry H. Lyon. Filed Nov. 10, 1916.

1,322,622, wireless Signaling System, James H. Rogers and Henry H. Lyon, Filed Nov. 10, 1916.

1,349,103, Radiosignaling System. James Harris Rogers. Filed May 2, 1917.

1,315,862, Radiosignaling System. James Harris Rogers. Filed Jan. 17,

1918.

1,349,104, Radiosignaling System. James Harris Rogers. Filed Jan. 17, 1918.

1,303,729, Wireless Signaling System. James Harris Rogers. Filed Jan. 10,
1919.

1,303,730, Radiosignaling System. James Harris Rogers. Filed Jan. 11, 1919.

1,316,188, Radiosignaling System. James H. Rogers. Filed Mar. 28, 1919.

ACKNOWLEDGEMENTS

PART I

I am much indebted to Mr. Carey Briggs for furnishing references which strongly support the existence of scalar potential waves and scalar interferometry. Among those papers are Whittaker (1903), Nisbet (1955) , and other papers dealing with Hertzian potential theory. As ever, my profound gratitude is due Hal Crawford for his splendid illustrations which accompany this paper. I am also indebted to John Ratzlaff and the Tesla Book Company for graciously furnishing important Tesla references for me to analyze. My sincere thanks go to Messrs. Jim Beal and Williard Van de Bogart for furnishing important references on exoelectron emission and tectonic fault zones, and to Dr. Robert Beck for furnishing important material on the Rogers underground and undersea transmission system. Additional assistance was also received from Mr. Peter Kelly, Joe Siegeldorf, Joe Gambill, and Bob Beutlich. I am also very appreciative of the information furnished by Bob Jones and Vestigia on the beautiful Vestigia mystery light experiments.

PART II

A. L. Lez, San Anselmo, Calif.
Thorn L. Mayes, Saratoga, Calif.
Bob Nelson, Berkeley, Calif.
Mildred Ratzlaff, Millbrae, Calif.

The Tom Bearden Website

FOREWORD

The long-searched-for Tesla secrets have been solved by Tom Bearden. Part I contains the specific description of the solutions to Tesla's secrets and the Soviet Tesla weapons.

The present electromagnetic theory has several flaws, and it is these errors that have hidden the long-sought unified field theory. The explanations presented by the author provide clarity and understanding to related Tesla writings and patents.

Before the turn of the century, Nikola Tesla had discovered and was utilizing a new type of electric wave. Tesla repeatedly stated that his waves were non-Hertzian, and that his wireless transmissions did not fall off as the square of the distance. The author has found that the wave which Tesla was using is a longitudinal scalar wave, capable of many times the speed of light. When coupled with another scalar wave (Tesla wave), the resulting vector wave can be the source of electromagnetic energy for electrical power, a protective force field, or a destructive weapon.

The booklet is well illustrated with special drawings by Hal Crawford. The bibliography includes corroborating information on all major points. Part II presents reprints or the entire subject matter of selected articles from the bibliography, plus additional references of importance. Articles are arranged in chronological order. Also included are nine patents by James Harris Rogers which were the basis of successful wireless communication during World War I. These principles of underground and sub-sea wireless were based on Tesla's discoveries.

Bearden's solutions are truly a major breakthrough, and an absolute necessity to the understanding of existing Tesla writings. It is also important that more should be known of this extraordinary individual, Thomas E. Bearden.

BIOSKETCH:

A nuclear engineer, wargames analyst, and military tactician, Lieutenant Colonel (Retired) Tom Bearden has over 26 years experience in air defense systems, tactics and operations; technical intelligence, anti-radiation missile countermeasures; nuclear weapons employment, computerized wargames; and military systems requirements.

He is currently with the Alabama division of a large aerospace company where he is involved in determining the future requirements for laser weapons.

Lieutenant Colonel Bearden obtained a Master of Science degree in nuclear engineering from the Georgia Institute of Technology and a Bachelor of Science degree in mathematics from Northeast Louisiana University. He is also a graduate of the U.S. Army Command and General Staff College and several U.S. Army artillery and missile schools.

He has had direct experience with tube artillery and with the Ajax, Hercules, Hawk, and Patriot missile systems and the production of technical intelligence on Soviet surface-to-air missile systems for the U.S. Army/Department of Defense.

Tom holds a Sandan (third degree black belt) in Yoseikan aikido; is a singer, guitarist and songwriter; member of Mensa, Society for the Investigation of the Unexplained, Vestigia, the American Association of Meta-Science: and is on the board of governors of the U.S. Psychotronics Association. In addition, he is Alabama Director of a small nonprofit foundation, the Association of Distinguished American Scientists and is a MUFON adviser on nuclear engineering.

He also edits *Specula,* Journal of the American Association of Meta-Science.

Tom is primarily active as a theorist on UFOs, parapsychological and paranormal phenomena, and psychotronics. He is the author of a book, *The Excalibur Briefing,* Strawberry Hill Press, 2594 15th Avenue, San Francisco, California 94127 and has numerous technical papers dealing with new paradigm physics, paranormal phenomena, and Soviet psychotronic weapons.

Finally, Bearden's *Solutions to Tesla's Secrets and the Soviet Tesla Weapons* promises to be recognized in the future as a major contribution to our present technology.

JTR

The Tom Bearden Website

TESLA'S SECRET AND THE SOVIET TESLA WEAPONS

© T.E. Bearden 1981

With Special Drawings by Hal Crawford

Before the turn of the century, Nikola Tesla had discovered and was utilizing a new type of electric wave. Tesla repeatedly stated his waves were non-Hertzian, and his wireless transmissions did not fall off as the square of the distance. His discovery was apparently so fundamental (and his intent to provide free energy to all humankind was so clear) that it was responsible for the withdrawal of his financial backing, his deliberate isolation, and the gradual removal of his name from the history books.

By 1914 or so, Tesla had been successfully isolated and was already nearly a "nonperson." Thereafter Tesla lived in nearly total seclusion, occasionally surfacing (at his annual birthday party for members of the press) to announce the discovery of an enormous new source of free energy, the perfection of wireless transmission of energy without losses, fireball weapons to destroy whole armies and thousands of airplanes at hundreds of miles distance, and a weapon (the "Tesla Shield," I've dubbed it) that could provide an impenetrable defense and thus render war obsolete.

In my pursuit of Tesla's secret, it gradually became apparent to me that present orthodox electromagnetic theory is seriously flawed in some fundamental respects. One of these is in the definition and use of ϕ, the scalar electrostatic potential. It is this error which has hidden the long-sought unified field theory from the theorists.

In the theory of the scalar electrostatic potential (SEP), the idea is introduced of work accomplished on a charge brought in from a distance against the scalar field. The SEP is not a vector field, but is a scalar field. Indeed, scalar potential cannot of itself perform work on a charged mass; if it could do so, then tremendous force would exist on every mass due to the extremely high SEP of the vacuum itself. Only a differential of SEP between two spatial points can produce force or accomplish work. (Rigorously, a differential of scalar potential between two spatial points constitutes a vector. Only a vector can produce force and do work.)

Also, work can only be done on a mass. Further, it takes time* to move an electron or other charged mass between two spatial points, and so the work performed by a spatial differential of the ϕ-field requires time. Rigorously, the delta SEP is voltage, not SEP per se, and is directly related to

E field. The entire voltage concept depends on the work performed in moving a mass, after that mass has moved. The idea of "voltage" always implies the existence of a steady differential of ϕ between two spatial points for a finite length of time, and it also involves the assumption of a flow of actual mass having occurred. SEP, on the one hand, is always a single-point function; on the other hand, difference in potential (i.e., V) is always a two point function, as is any vector.

Yet many graduate level physics and electromagnetics papers and texts erroneously confuse ϕ and V in the static case! Such an interpretation is of course quite incorrect.

---

* *Two spatial points involve at least $\Delta t = \Delta L/c$ in time. All vectors and gradients involve 2 separated spatial points, and thus present timelines in 4-space. ϕ_4 is a point, not a line, in 4-space.*

Another common assumption in present EM theory -- that the electrostatic potential (ϕ_0) of the normal vacuum is zero -- has no legitimate basis. In fact, we know ϕ_0 is nonzero because the vacuum is filled with enormous amounts of fluctuating virtual state activity, including incredible charge fluctuations. And by virtue of its point definition, ϕ_0 must be the "instantaneous intensity" of these fluctuations -- but both in space and time. The scalar electrostatic potential is therefore the "instantaneous stress" on spacetime itself, and a measure of the intensity of the virtual state flux through a 4-dimensional spacetime point.

Potential theory was largely developed in the 1800's, before the theory of relativity. Time flowrate was then regarded as immutable. Accordingly, electrostatic "intensity" was chosen as "spatial intensity," with the connotation of "spatial flux density." This assumes a constant, immutable rate of flow of time, which need not be true at all if we believe relativity. Such a spatial "point" intensity is actually a "line" in 4-space, and not a 4-dimensional "point" at all. Thus the spatial potential -- ϕ_3 -- is a very special case of the real spacetime potential -- ϕ_4, or charge and electromagnetic theory today is accordingly a special case of the real 4-space electromagnetism that actually exists! Note also that charge is a 4-dimensional concept.

Now mass is a spatial, 3-dimensional concept. Rigorously, mass does not exist in time -- masstime exists in time. Mass and charge are thus of differing dimensionalities!

Also, according to quantum mechanics, the charge of a particle -- e.g., of an electron -- is due to the continual flux of virtual particles given off and absorbed by the observable particle of mass. Thus charge also is conceptually a measure of the virtual flux density, and directly related to ϕ. Further, since the charge exists in time, it is the charge of a particle of spatial mass that gives it the property of masstime, or existing in time.

Here a great confusion and fundamental error has been thrown into the present EM theory by the equating of "charge" and "charged mass." As we have seen, the two things are really very different indeed.

To speak of a spatial "amount" of charge erroneously limits the basic EM theory to a fixed time flowrate condition (which of course it was considered to be, prior to Einstein's development of relativity). Thus when the limited present theory encounters a "relativistic" case (where the time flowrate changes), all sorts of extraordinary corrections must be introduced. The real problem, of course, is with the fundamental definitions of electrostatic potential and charge. The spatial "amount" of charge (i.e. the coulomb) as we presently erroneously use the term, is actually the spatial amount of observable "charged mass." To correct the theory, one must introduce the true 4-space SEP and separate the definitions of charge and charged mass.

Only when a mass is moved does one have work -- and voltage or vector fields. (The reason one has voltage and E field connected to a normal electrostatically charged object in the laboratory is because an excess of charged-particle masses are assembled on the object, and these masses are in violent motion! A true static charge would have no E field at all.)

Table 1. Some Present Theoretical Facts

- $Ø_3$-FIELD CONCEIVED BEFORE RELATIVITY
    - POINT FUNCTION IN SPACE
    - LINE FUNCTION IN SPACETIME
    - TIME FLOWRATE CONSIDERED IMMUTABLE
    - SPECIAL CASE OF $Ø_4, \ldots\ldots, Ø_N$
- GRADIENT BETWEEN TWO POINTS IN $Ø_3$ FIELD
    - PRODUCES VOLTAGE
    - PRODUCES E FIELD
    - MOVES ELECTRON
    - IS AN AREA IN 4-SPACE
- GRADIENT MAY BE BETWEEN SEPARATED POINTS IN TWO SUPERPOSED Ø-FIELDS
- Ø-FIELD CAN "MOVE" AT INFINITE VELOCITY*

* JACKSON, CLASSICAL ELECTRODYNAMICS, 2ND EDITION, P. 223

Table 2. Some Proposed $\varnothing_4$ Characteristics

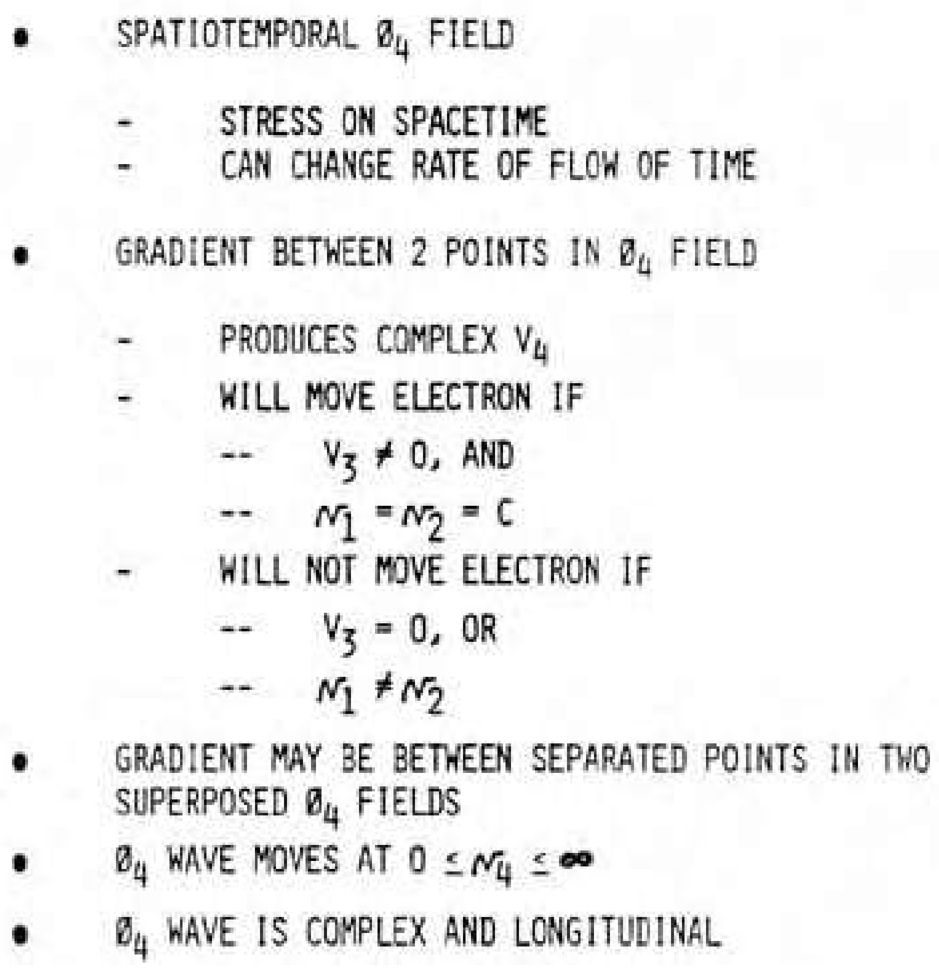

- SPATIOTEMPORAL $\varnothing_4$ FIELD
  - STRESS ON SPACETIME
  - CAN CHANGE RATE OF FLOW OF TIME
- GRADIENT BETWEEN 2 POINTS IN $\varnothing_4$ FIELD
  - PRODUCES COMPLEX V_4
  - WILL MOVE ELECTRON IF
    - $V_3 \neq 0$, AND
    - $\nu_1 = \nu_2 = c$
  - WILL NOT MOVE ELECTRON IF
    - $V_3 = 0$, OR
    - $\nu_1 \neq \nu_2$
- GRADIENT MAY BE BETWEEN SEPARATED POINTS IN TWO SUPERPOSED $\varnothing_4$ FIELDS
- $\varnothing_4$ WAVE MOVES AT $0 \leq \nu_4 \leq \infty$
- $\varnothing_4$ WAVE IS COMPLEX AND LONGITUDINAL

ϕ-field need not involve observable mass accumulation, but only charge (virtual flowrate intensity) accumulation. Accumulated masses are like so many gallons of water; accumulated charge is like so much pressure on both the water (space) and the time in which the water is existing.

Now, if one varies the SEP solely as a point function, one would have a purely scalar complex longitudinal wave, and not a vector wave at all. This is the fundamentally new electrical wave that Tesla discovered in 1899.

Rigorously, all vector fields are two-point functions and thus decomposable into two scalar fields, as Whittaker showed in 1903. It follows that any vector wave can be decomposed into two scalar waves. By implication, therefore, a normal transverse EM vector wave, e.g., must simply be two coupled scalar (Tesla) waves -- and these scalars independently would be longitudinal if uncoupled. An ordinary transverse EM vector wave is thus two pair-coupled Tesla scalar longitudinal waves, and only a single special case of the much more fundamental electromagnetics discovered by Nikola Tesla.

A Tesla (scalar potential) wave -- i.e., a massless wave in pure ϕ_0, the stress of the spacetime medium would have very strange characteristics indeed. For one thing, since it moves in a complex 4-space, it has many more modes of movement than does a simple wave in 3-space. And for another thing, it need not be bound at all by the speed of (vector) light. In current theory, one ϕ_3-field does not directly interact or couple with other existing ϕ_3-fields except by simple superposition. Therefore presently the ϕ-field is considered to have no drag limitation at all, hence infinite velocity. (E.g., as stated in Jackson, Classical Electrodynamics, 2nd edition, page 223.)

Actually, a ϕ_4-wave can and will interact with some of the other existing ϕ_4-waves in the medium transversed, and this interaction can involve pair-coupling into EM vector fields and waves, an interaction not presently in the electrodynamics theory. The result of scalar pair-coupling creates a finite amount of vector "drag" on the ϕ_4-wave, so it then has less than infinite velocity. However, if this drag is small due to limited pair-coupling, the scalar wave's velocity through the slightly dragging medium still may be far greater than the speed of vector EM waves (light) in vacuum. On the other hand, if the pair-coupling is made severe, the ϕ-wave may move at a speed considerably below the speed of vector light waves in vacuum. The velocity of the ϕ_4-wave is thus both variable and controllable or adjustable (e.g., simply by varying its initial amplitude, which through a given medium changes the percentage of pair-coupling and hence the degree of drag on the scalar wave). The Tesla scalar wave thus can have either subluminal or superluminal velocity, in contradiction to present theory.

Note that the scalar wave also violates one of Einstein's fundamental postulates -- for the speed of our "new kind of light" wave is not limited to c, and need not be the same to every observer. Thus Tesla scalar waves lead to a new "superrelativity" of which the present Einstein relativity is only a highly special case!

Table 3. Tesla Waves Can:

- **ESTABLISH STANDING WAVES**
    - **in the earth**
    - **in the ionosphere**
- **TAP ENERGY FROM THE EARTH'S CORE**
- **TRAVEL FASTER OR SLOWER THAN LIGHT**
- **CHANGE RATE OF TIME FLOW**
- **AFFECT ALL FIELDS, INCLUDING GRAVITY**
- **COMMUTE BETWEEN VIRTUAL AND OBSERVABLE**

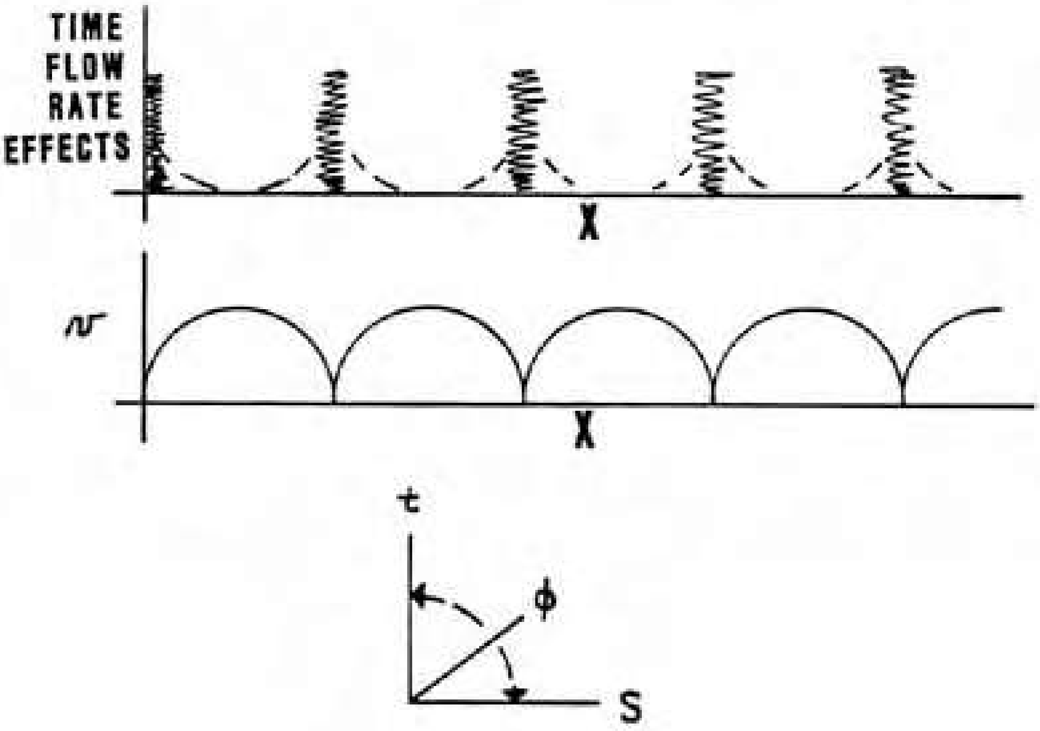

Figure 1. Tesla's Standing Columnar Wave

But let us now look for some subtle but real examples of scalar waves and scalar pair-coupling in nature. As is well known, a tectonic fault zone can provide anomalous lights, sounds, etc. from stresses, piezoelectrical activity, and telluric currents in the earth and through the fault zone. In examining the fault zone phenomena, I finally realized that a fault zone was literally a scalar interferometer i.e., if one can have scalar ϕ_4-waves, they can interfere, either constructively or destructively. Their interference, however, produces scalar pair-coupling into vector EM waves. This coupling may be at a distance from the interferometer itself, and thus the interferometer can produce energy directly at a distance, without vector transmission through the intervening space. Coupling of ϕ waves with the paired scalars comprising ordinary EM vector waves can also occur. If this triplex coupling forms additional EM vector waves 180 degrees out of phase, the ordinary EM wave is diminished or extinguished. If the scalar triplex coupling occurs so as to create vector EM waves in phase with the interacting vector EM wave, the amplitude of the ordinary vector wave is increased.

Scalar potential waves can thus augment or diminish, or create or destroy, ordinary EM waves at a distance by pair-coupling interference under appropriate conditions, and this is in consonance with the implications of Whittaker's fundamental 1903 work.

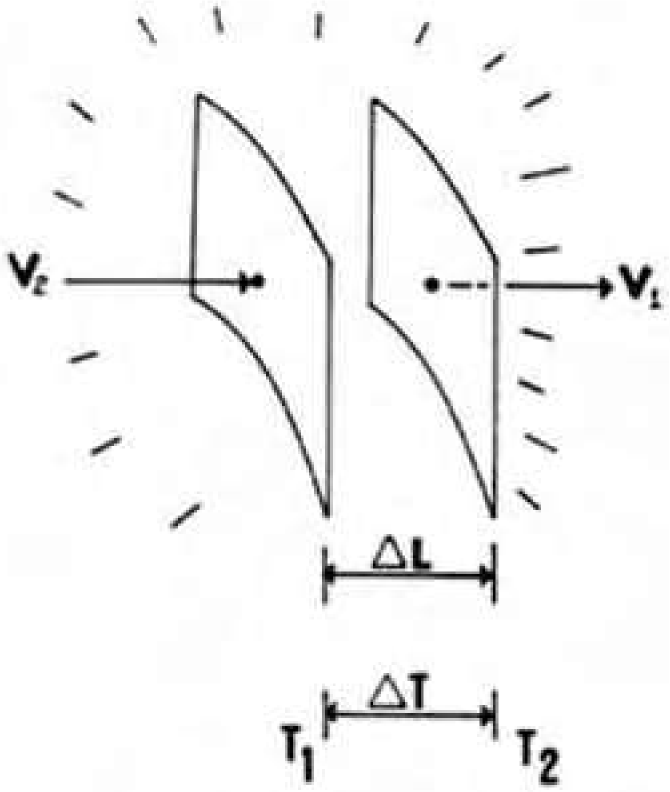
V_2
V_1
ΔL
ΔT
T_1
T_2

ALL VECTORS AND VECTOR FIELDS ARE CREATED BY SCALAR INTERFEROMETRY AND PAIR COUPLING

EM COUPLING CONDITION

$$\Delta L/c = \Delta T$$

Figure 2. Pair Coupling: Producing EM at a Distance

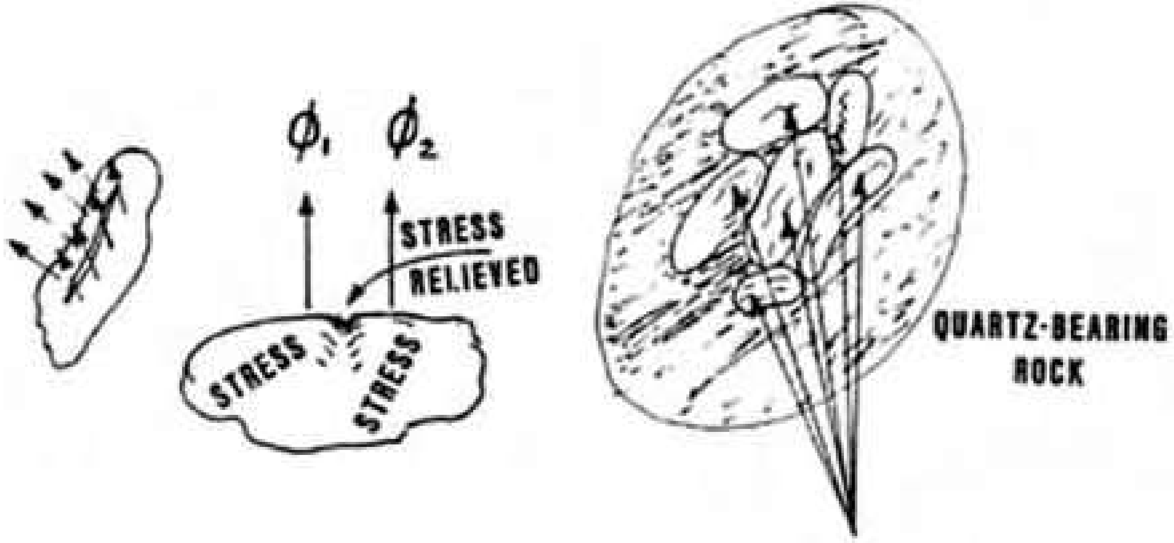
ϕ_1
ϕ_2
STRESS RELIEVED
STRESS
STRESS
QUARTZ-BEARING ROCK

LOCKED-IN STRESSED CRACKS

EACH TINY STRESSED CRACK ACTS AS A SCALAR INTERFEROMETER

BECOMES A PRIGOGINE CRYSTAL

Figure 3. Natural Scalar Interferometers

An earthquake fault zone is such a scalar interferometer. Stresses and charge pileups exist in the plates on each side adjacent to the fault, with stress relief existing in the middle in the fault fracture itself. Since the rock is locally nonlinear, the mechanical stresses and electrical currents in it are also locally nonlinear. This results in the generation of multiple frequencies of ϕ_4-waves from each side of the fault interferometer, yielding two complex Fourier expansion patterns of scalar potential waves. On occasion these two Fourier-transformed scalar wave patterns couple at a distance to produce stable ordinary electromagnetic field in a 3-dimensional spatial pattern -- e.g., a stress light such as the Vestigia light covered in Part I of The Excalibur Briefing. Driven by the erratic two scalar Fourier expansion patterns of the scalar interferometer (whose input stresses normally slowly change), an erratic, darting, hovering "spooklight" of the variety studied by Vestigia is produced.

As the stresses change in each side of the interferometer, the distant scalar coupling zone is affected. Thus the stresslight moves and its form changes, but it may be relatively stable in form for seconds or minutes. Since the stresses in the rock may be intense, the stress light may involve an intense pair of ϕ-patterns coupling into the sphere or ball of vector EM energy. The atoms and molecules of the air in the region of the coupled stresslight ball thus become highly excited, giving off radiant energy as the excited states decay.

Since much of the piezoelectric material in the stressed rocks is quartz grains, the features of quartz are of particular interest. Each little quartz grain is itself highly stressed, and has stress cracks. It is therefore a little scalar interferometer. Further, quartz is transparent to infrared and ultraviolet; and the random orientation of all the quartz scalar interferometers may also form a Prigogine system far from thermodynamic equilibrium. If so, this system can tap into highly energetic microscopic electromagnetic fluctuations to produce large-scale, ordered, relatively stable patterns of electromagnetic energy at a distance.

In short, all of this lends support to the formation of relatively stable but somewhat erratic patterns of electromagnetic energy at a distance from the fault itself. In the atmosphere, such scalar interferometers could form in clouds or even in the air or between clouds and earth. If so, such rare but occasional "weather" scalar interferometers could account for the rare phenomenon of ball lightning. The intense energy of the ball of lightning, as compared to the lesser energy of an earthstress light, could well be due to the enormous electrical charges between clouds or between cloud and earth, available to fuel the scalar interferometer. Very probably it is this phenomenon which gave Tesla the clue to scalar wave interferometry.

Thus such phenomena as earthstress lights, ball lightning, and the Tesla system of wireless transmission of energy at a distance with negligible losses and at speeds exceeding the speed of light may be explained. They are complex, however, and involve fundamental changes to present electromagnetic theory. These changes include utilizing 4-space scalar electrostatic potentials, scalar waves, pair coupling, ordinary 3-dimensional Fourier expansion, the Prigogine effect, and the properties of piezoelectric materials in rocks.

Since the scalar potential also stresses time, it can change the rate of flow of time itself. Thus it affects anything which exists in time -- including the mind, both of the individual and at various levels of unconsciousness. Therefore the same functions that result in earthstress lights also affect mind and thought, and are in turn affected by mind and thought. This is the missing ingredient in Persinger's theory that UFO's are correlated with, and a result of, fault zones and earth stresses. While Persinger seems to feel this is a "normal physics" explanation, it indeed involves a paranormal explanation.

The time-stressing ability of the true ϕ scalar wave also explains the interaction of such earthstress lights with humans and human intent, as noted by other researchers. (E.g., the lights that repeatedly seemed to react to the observers, as detailed by Dr. Harley Rutledge in his epoch-making Project Identification, Prentice-Hall, 1981.)

These ideas in condensed form comprise the concepts required to violate the speed of light and produce an ordinary electromagnetic field at a distance, using scalar interferometry, without losses -- as Tesla had done in his wireless transmission system which he had tested prior to 1900 and had perfected by the 1930's. Scalar interferometry can give stable regions of EM or "light energy" at a distance without losses, particularly as detailed in the beautiful Vestigia experiments, and it is within our grasp to utilize the new effects. Indeed, any stress crack in a material can result in the scalar potential interferometer effect. Exophoton and exoelectron emission -- poorly understood but already known in fatiguing of materials -- must be at least partly due to the scalar interferometer effect.

However, one additional caution should be advanced. Normal movement of electrons allows so much "sideplay" movement of the electrons -- and there is so much such sideplay electron motion in the surrounding vicinity -- that pair coupling is almost instantaneous for small waves. Thus orbital electrons in atoms seem to absorb and emit vector EM photons. Actually they also emit some percentage of scalar waves as well. Since a scalar wave is comprised exclusively of disturbance in the virtual state, it need not obey the conservation of energy law. Further, a scalar wave of itself does not "push electrons" or other charges; hence it is nearly indetectable by present detectors. Ionization detectors such as a Geiger counter tube, e.g., are exceptions if the scalar wave encountered is fairly strong. In that case sufficient triplex coupling with the ionized gas occurs to produce additional ionization or charge, breaching the tube's cutoff threshold and producing a cascade discharge of electrons and voltage which is detected. But weak scalar waves are presently indetectable by ordinary instruments. However, these small scalar waves are detectable by sensitive interferometry techniques e.g., such as an electron interferometer. Since the use of such instruments is quite rare, then indeed we have been living immersed in a sea of scalar waves without knowing it.

Finally, the percentage of scalar waves produced by changes in charged mass pileups can be increased by utilizing charged mass streaming. Essentially the charged masses must be moved suddenly, as quickly as possible, at or near the complete breakdown of the medium. For this reason, Tesla utilized sparkgaps in his early transmission systems, but also found that he could induce ionized media to "breakdown" in such fashion by a slow growth process. One of his early patented atmospheric wireless transmission systems is based on this fact. However, it was necessary to use a very high voltage, insuring extreme stress on the medium and hence some spillover stress onto time itself. In other words, ϕ_3 is always an approximation; at sufficiently high spatial stress, sufficient spillover ϕ_4 exists to give Tesla scalar waves. For this reason, Tesla used very high voltages and extremely sharp discharges to give "streaming" of the charged masses and thus high percentages of ϕ_4 waves. This suggests that the breakdown of dielectrics is a much richer phenomenon than is presently allowed for in the conventional theory.

To summarize, electrostatic potential -- ϕ-field is stress on the spacetime medium at a four-dimensional point. I. e. , it is a sort of pressure on the medium, but pressure on all four dimensions, not just on the three spatial dimensions. Thus in the new standard theory, ϕ_4 may have complex values. In addition, a ϕ-wave is to be interpreted as a scalar longitudinal wave in complex spacetime -- directly in ϕ_0, the normal average 4-space stress itself. And charge and charged mass must be recognized as two separate concepts. This is the gist of what I finally recognized about Nikola Tesla's work and fundamental discovery.

This is exciting, for it means that Tesla stress waves can affect either space or time individually, or both space and time simultaneously, or even oscillate back and forth between primarily affecting time and primarily affecting space, Tesla's waves were actually these ϕ-field scalar waves. As such, they were fundamentally different from ordinary electromagnetic waves, and had entirely different characteristics, just as Tesla often stated. E. g., a Tesla wave can either move spatially, with time flowing linearly; move temporally only (sitting at a point and waxing and waning in magnitude -- but changing the rate of flow of time itself in doing so, and affecting gravitational field, fundamental constants of nature, etc,), or move in a combination of the two modes. In the latter case, the Tesla wave moves in space with a very strange motion -- it oscillates between (1) spatially standing still and flexing time, and (2) moving smoothly in space while time flows smoothly and evenly. I.e., It stands at one point (or at one columnar region), flexing for a moment; then slowly picks up spatial velocity until it is moving smoothly through space; then slows down again to a "standing column," etc. This is Tesla's fabulous "standing columnar wave."

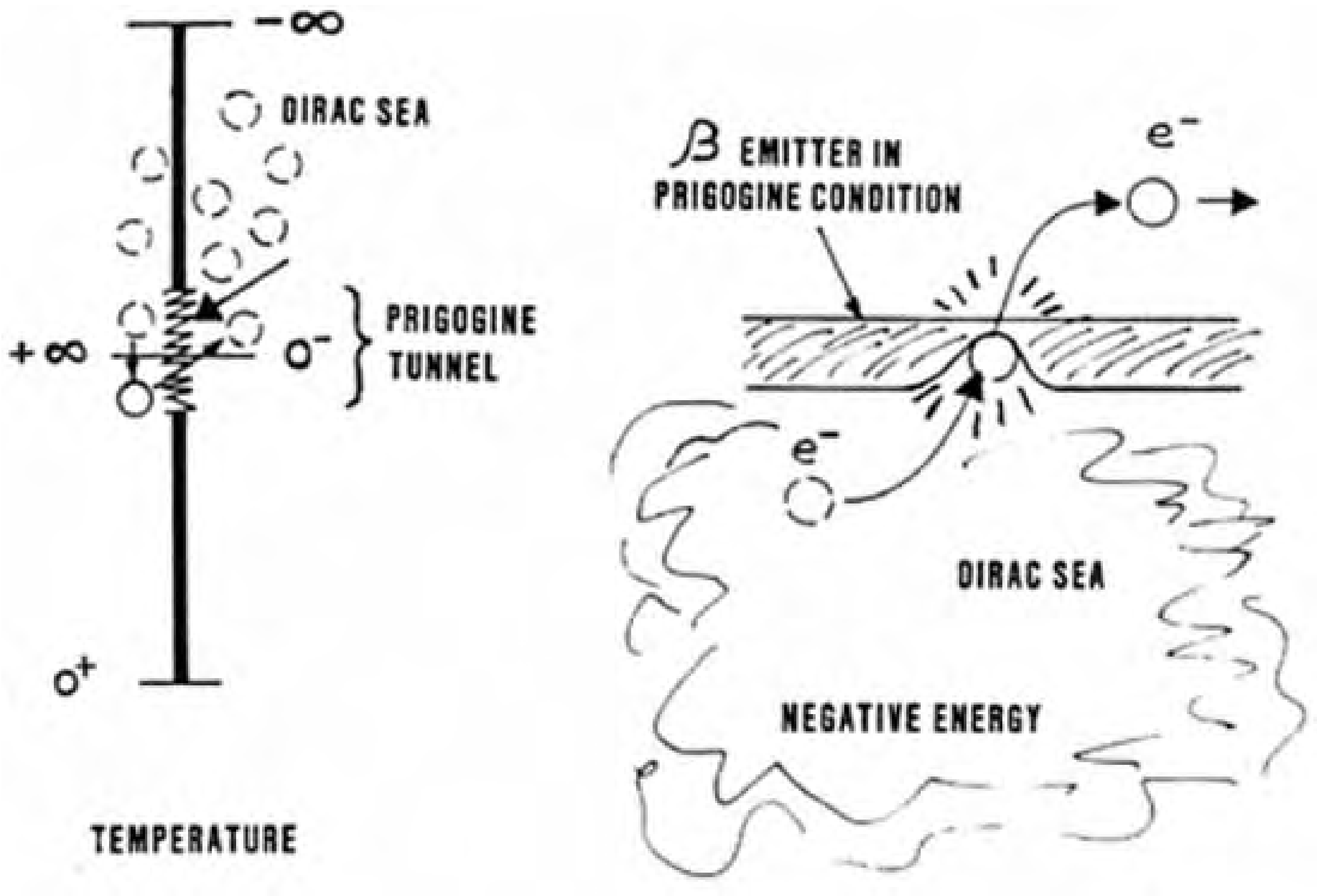
−∞
DIRAC SEA
+∞
0⁻
PRIGOGINE
TUNNEL
0⁺
TEMPERATURE
β EMITTER IN
PRIGOGINE CONDITION
e⁻
e⁻
DIRAC SEA
NEGATIVE ENERGY

Figure 4. Exoelectric Emission

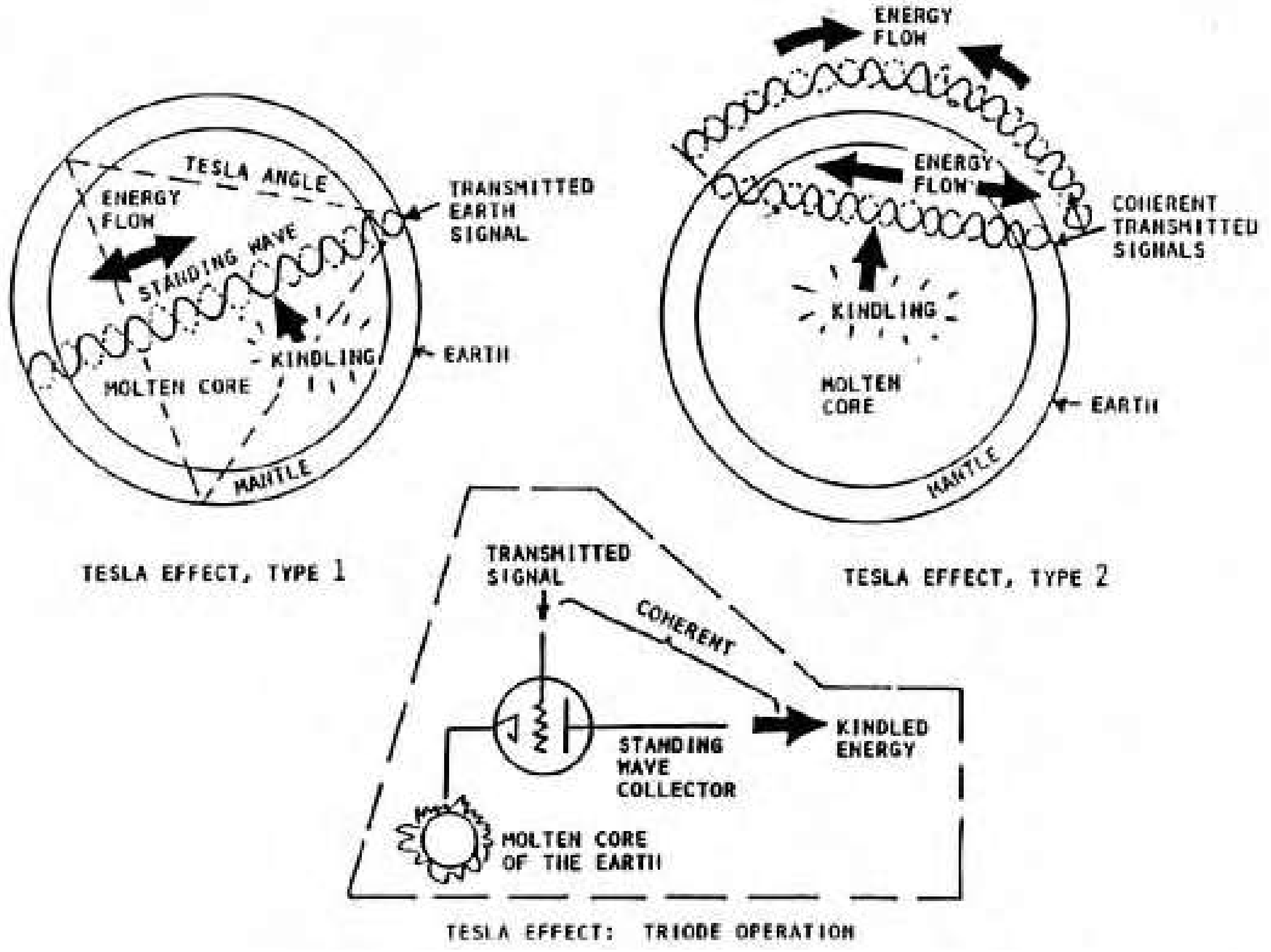
TESLA ANGLE
ENERGY FLOW
STANDING WAVE
TRANSMITTED EARTH SIGNAL
KINDLING
EARTH
MOLTEN CORE
MANTLE
TESLA EFFECT, TYPE 1
ENERGY FLOW
ENERGY FLOW
COHERENT TRANSMITTED SIGNALS
KINDLING
MOLTEN CORE
EARTH
MANTLE
TESLA EFFECT, TYPE 2
TRANSMITTED SIGNAL
COHERENT
KINDLED ENERGY
STANDING WAVE COLLECTOR
MOLTEN CORE OF THE EARTH
TESLA EFFECT: TRIODE OPERATION

Figure 5. The Tesla Effect

Another wild characteristic of the Tesla wave is that it can affect the rate of flow of time itself; hence it can affect or change every other field including the gravitational field -- that exists in time flow. It can also affect all universal constants, the mass of an object, the inertia of a body, and the mind and thoughts as well! All of these exist in the flow of time, and they are affected if the time stream in which they exist is affected. This was the awful secret that Tesla partially discovered by 1900, and which he came more and more to fully realize as he pursued its nature and its ramifications into the 1920's and 1930's.

Tesla also found he could set up standing ϕ-field waves through the earth. He in fact intended to do so, for he had also discovered that all charges in the highly stressed earth regions in which such a standing wave existed produced ϕ-fields which would feed (kindle) energy into the standing ϕ-field wave by pair coupling. I.e., normal vector field energy would "assemble" onto the scalar matrix wave by means of pair-coupling. Thus by transmitting a scalar standing wave into the earth, he could easily tap the fiery scalar fields produced in the molten core of the planet itself, turning them into ordinary electromagnetic energy. In such case, a single generator would enable anyone to put up a simple antenna and extract all the free energy desired.

When Tesla's alarmed financial backers discovered this was his real intent, they considered him a dangerous madman and found it necessary to ruthlessly stop him at all costs. And so his financial support was withdrawn, he was harassed in his more subtle patent efforts (and the patents themselves were adulterated), and his name gradually was removed from all the electrical textbooks. By 1914 Tesla, who had been the greatest inventor and scientist in the world, had become essentially a nonperson.

A few other persons in the early 1900's also were aware that potential and voltage are different. And some of them even learned to utilize Tesla's ϕ-field, even though they only vaguely understood they were utilizing a fundamentally different kind of electromagnetic wave. For example, James Harris Rogers patented an undersea and underground communications system which Tesla later confirmed utilized Tesla waves. The U.S. secretly used the Rogers communications system in World War I to communicate with U.S. submarines underwater, and to communicate through the earth to the American Expeditionary Force Headquarters in Europe. The Rogers system was declassified after the War and very shortly after that, it had mysteriously been scrubbed off the face of the earth. Again, potential stress waves -- Tesla waves -- were eliminated and "buried."

Probably the most brilliant inventor and researcher into Tesla's electromagnetics was T. Henry Moray of Salt Lake City, Utah. Dr. Moray actually succeeded in tapping the limitless zero-point energy of vacuum (spacetime) itself. By 1939, Dr. Moray's amplifier contained 29 stages and its output stage produced 50 kilowatts of power from vacuum. Interestingly, another 50 kilowatts could be tapped off any other stage in the device -- which consequently could have produced almost 1.5 megawatts of electrical power! Dr. Moray's epoch-making work was suppressed also. His device -- which represented over 20 years of heartbreaking accumulation of 29 working tubes from thousands made -- was destroyed by a Soviet agent in 1939, but not before the agent had obtained the drawings for building the tubes and the device itself. Today the Moray amplifier is a standard component of many of the Soviet secret superweapons and Tesla weapons.

In the 20's and 30's, Tesla announced the final perfection of his wireless transmission of energy without losses even to interplanetary distances. In several articles (e.g., H. Winfield Secor, "Tesla Maps Our Electrical Future, " Science and Invention, Vol. XVII, No.12, pp. 1077, 1124-1126), Tesla even revealed he used longitudinal stress waves in his wireless power transmission. Quoting from the article, "Tesla upholds the startling theory formulated by him long ago, that the radio transmitters as now used, do not emit Hertz waves, as commonly believed, but waves of sound." "He says that a Hertz wave would only be possible in a solid ether, but he has demonstrated already in 1897 that the ether is a gas, which can only transmit waves of sound; that is such as are propagated by alternate compressions and rarefactions of the medium in which transverse waves are absolutely impossible." The wily Tesla did not reveal, of course, that such scalar waves nearly always immediately pair-coupled into vector waves when produced by normal means. Tesla himself was working with longitudinal scalar waves.

Table 4. Rogers' Undersea and Underground Communication System

- **SECRETLY USED BY U.S. IN WWI**
- **COMMUNICATED TO SUBS UNDERWATER**
    - **- ANY DEPTH, UP TO 30 KHZ**
    - **- REDUCED DEPTH, >30 KHZ**
- **COMMUNICATED TO EXPEDITIONARY HQ OVERSEAS**
    - **- 5,000 TIMES STRONGER THAN AERIAL**
    - **- STATIC-FREE**
- **DECLASSIFIED IN 1919**
    - **- TO REVOLUTIONIZE COMMUNICATIONS**
    - **- TO END ALL SURFACE ANTENNAS**
- **MYSTERIOUSLY "LOST"**
- **REDISCOVERED AND "LOST" THREE TIMES AFTER WWII**

continued

BEARDEN
FER DE LANCE
12.14 15:24
BRIEFING ON SOVIET SCALAR
ELECTROMAGNETIC WEAPONS

ORDER THE ALL NEW 2ND EDITION OF "FER DE LANCE"

Updated to the year 2002

Stunning new revelations about the secret scalar EM wars the media ignores

THIS BOOK IS THE ONLY SOURCE FOR THIS INFORMATION

The Tom Bearden Website

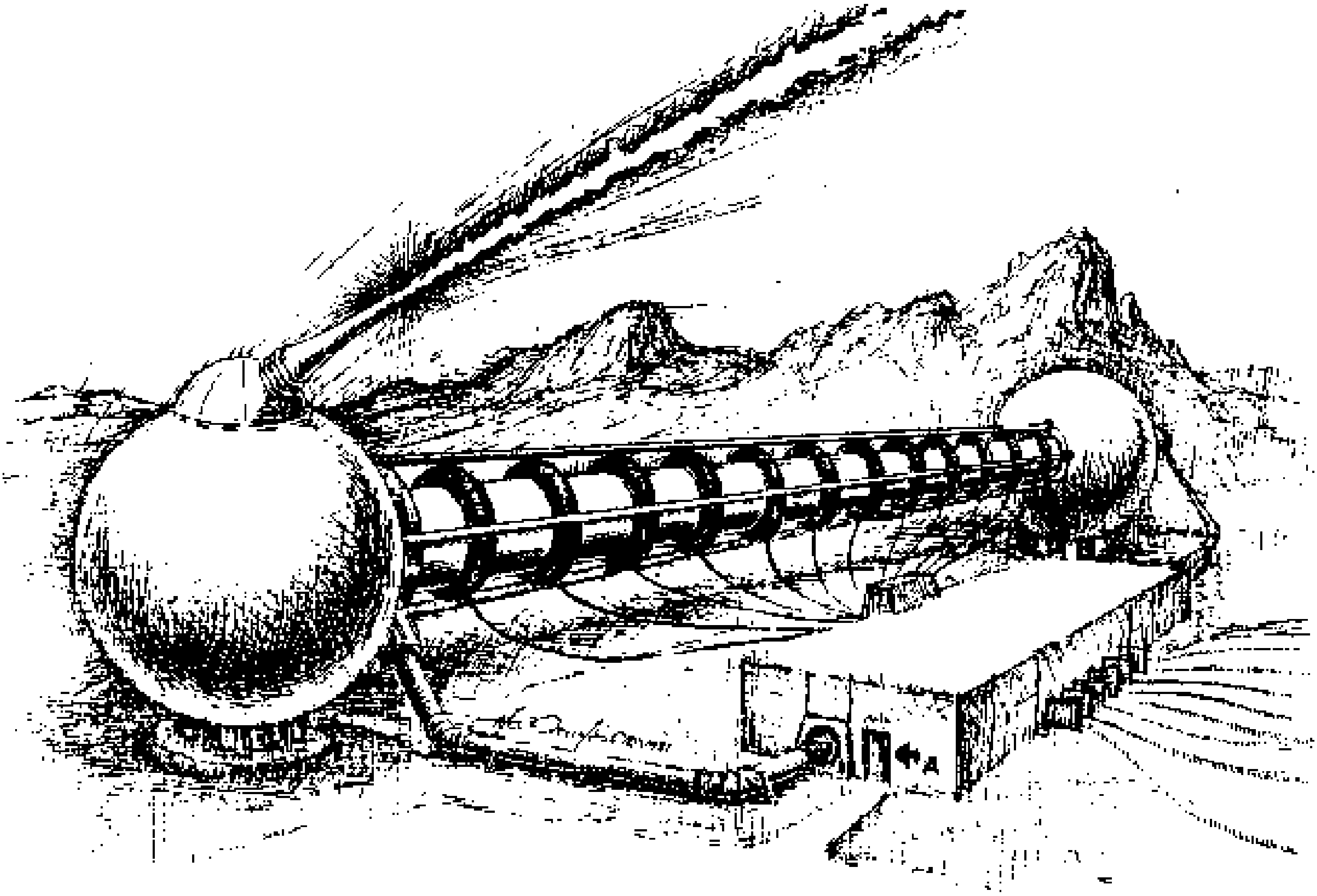

Scalar Potential Interferometer

Figure 6. Multimode Tesla Weapon

In the 1930's Tesla announced other bizarre and terrible weapons: a death ray, a weapon to destroy hundreds or even thousands of aircraft at hundreds of miles range, and his ultimate weapon to end all war -- the Tesla shield, which nothing could penetrate. However, by this time no one any longer paid any real attention to the forgotten great genius. Tesla died in 1943 without ever revealing the secret of these great weapons and inventions.

Unfortunately, today in 1981 the Soviet Union has long since discovered and weaponized the Tesla scalar wave effects. Here we only have time to detail the most powerful of these frightening Tesla weapons -- which Brezhnev undoubtedly was referring to in 1975 when the Soviet side at the SALT talks suddenly suggested limiting the development of new weapons "more frightening than the mind of man had imagined." One of these weapons is the Tesla howitzer recently completed at the Saryshagan missile range and presently considered to be either a high-energy laser or a particle beam weapon, (See Aviation Week & Space Technology, July 28, 1980, p. 48 for an artist's conception.)

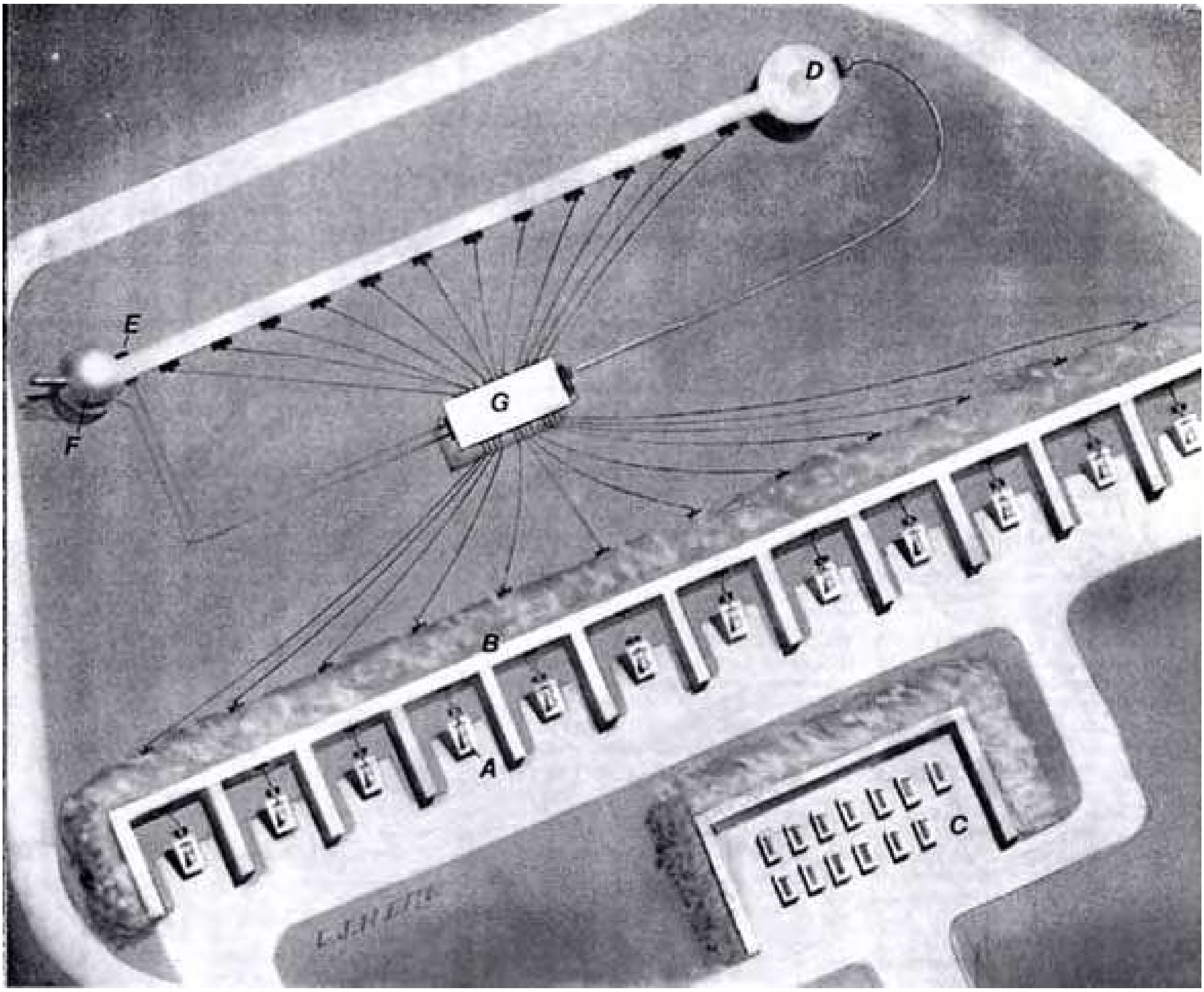

Aviation Week & Space Technology July 28, 1980

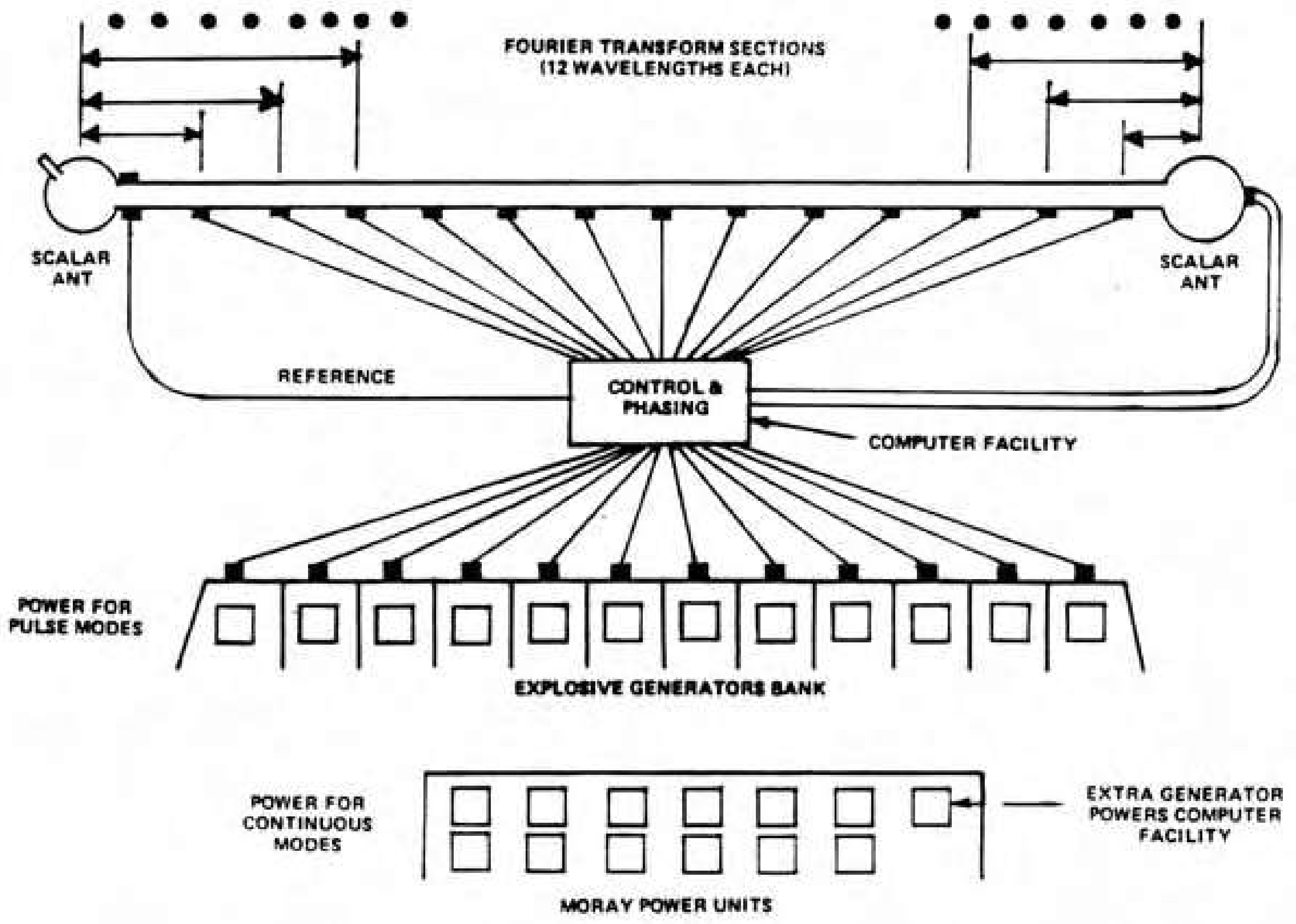
FOURIER TRANSFORM SECTIONS
(12 WAVELENGTHS EACH)
SCALAR ANT
SCALAR ANT
REFERENCE
CONTROL & PHASING
COMPUTER FACILITY
POWER FOR PULSE MODES
EXPLOSIVE GENERATORS BANK
POWER FOR CONTINUOUS MODES
EXTRA GENERATOR POWERS COMPUTER FACILITY
MORAY POWER UNITS

Figure 7. Tesla Weapons at Saryshagan

The Saryshagan howitzer actually is a huge Tesla scalar interferometer with four modes of operation. One continuous mode is the Tesla shield, which places a thin, impenetrable hemispherical shell of energy over a large defended area. The 3-dimensional shell is created by interfering two Fourier-expansion, 3-dimensional scalar hemispherical patterns in space so they pair-couple into a dome-like shell of intense, ordinary electromagnetic energy. The air molecules and atoms in the shell are totally ionized and thus highly excited, giving off intense, glowing light. Anything physical which hits the shell receives an enormous discharge of electrical energy and is instantly vaporized -- it goes pfft! like a bug hitting one of the electrical bug killers now so much in vogue.

If several of these hemispherical shells are concentrically stacked, even the gamma radiation and EMP from a high altitude nuclear explosion above the stack cannot penetrate all the shells due to repetitive absorption and reradiation, and scattering in the layered plasmas.

In the continuous shield mode, the Tesla interferometer is fed by a bank of Moray free energy generators, so that enormous energy is available in the shield. A diagram of the Saryshagan-type Tesla howitzer is shown in figure 7. Hal Crawford's fine drawing of the interferometer end of the Tesla howitzer is shown in figure 6. Hal's exceptional rendition of the Tesla shield produced by the howitzer is shown in figure 8.

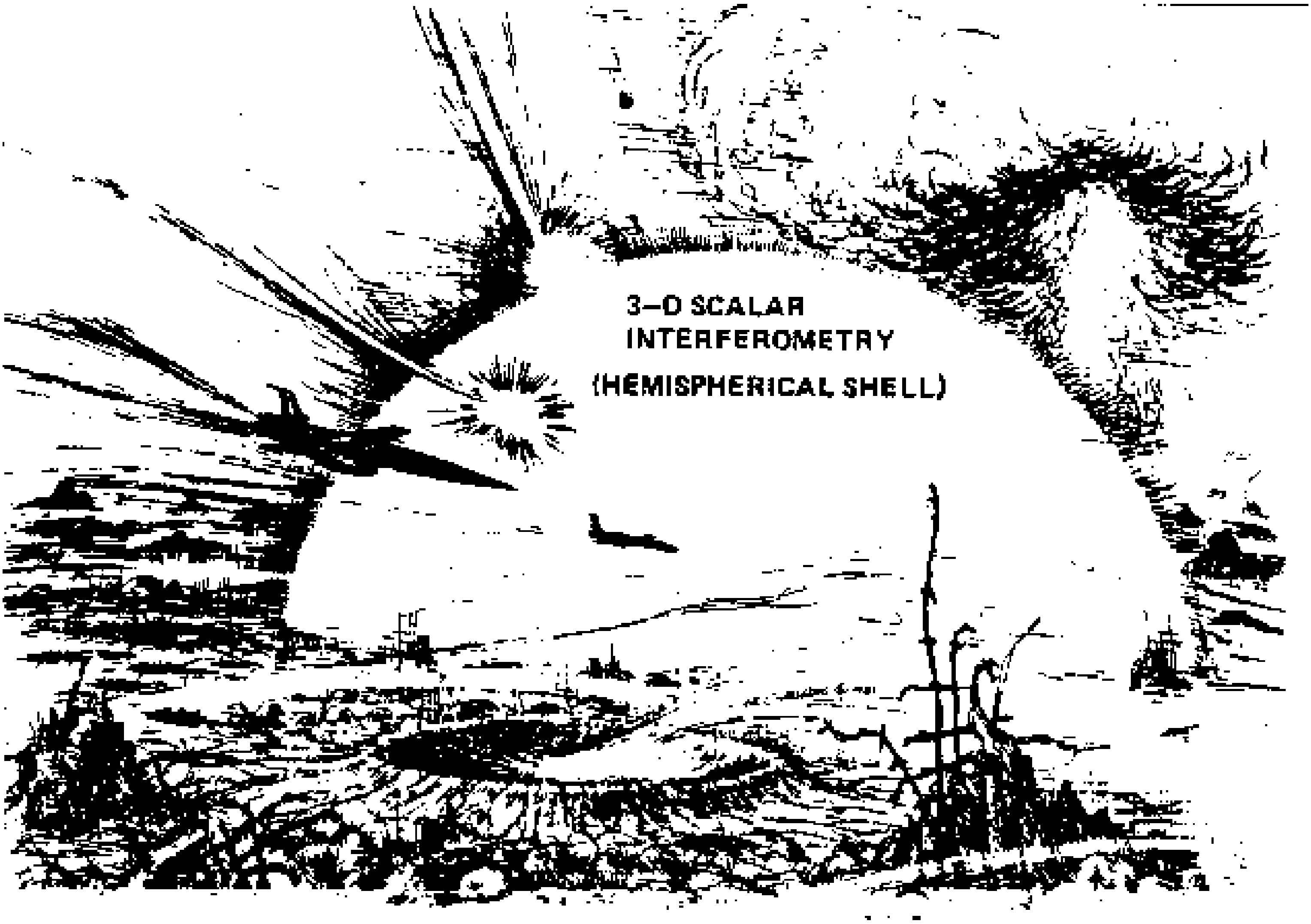
3–D SCALAR
INTERFEROMETRY
(HEMISPHERICAL SHELL)

Figure 8. The Tesla Shield

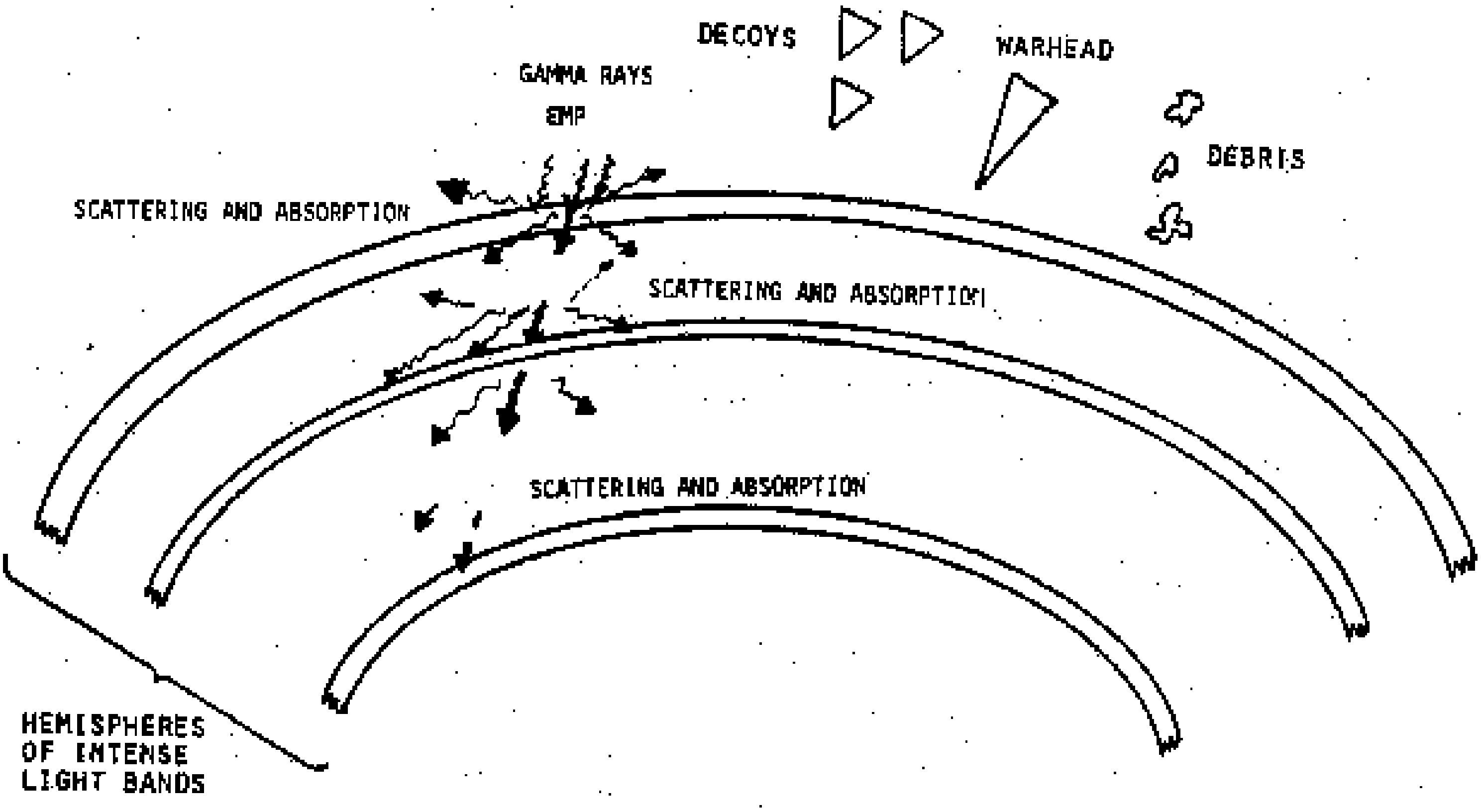
DECOYS
WARHEAD
GAMMA RAYS
EMP
DEBRIS
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
HEMISPHERES
OF INTENSE
LIGHT BANDS

Figure 9. Tesla Terminal Area Defense System

In the pulse mode, a single intense 3-dimensional scalar phi-field pulse form is fired, using two truncated Fourier transforms, each involving several frequencies, to provide the proper 3-dimensional shape (Figure 10). This is why two scalar antennas separated by a baseline are required. After a time delay calculated for the particular target, a second and faster pulse form of the same shape is fired from the interferometer antennas. The second pulse overtakes the first, catching it over the target zone and pair-coupling with it to instantly form a violent EMP of ordinary vector (Hertzian) electromagnetic energy. There is thus no vector transmission loss between the howitzer and the burst. Further, the coupling time is extremely short, and the energy will appear sharply in an "electromagnetic pulse (EMP)" strikingly similar to the 2-pulsed EMP of a nuclear weapon.

This type weapon is what actually caused the mysterious flashes off the southwest coast of Africa, picked up in 1979 and 1980 by Vela satellites. The second flash, e.g., was in the infrared only, with no visible spectrum. Nuclear flashes do not do that, and neither does superlightning, meteorite strikes, meteors, etc. In addition, one of the scientists at the Arecibo Ionospheric Observatory observed a gravitational wave disturbance -- signature of the truncated Fourier pattern and the time-squeezing effect of the Tesla potential wave -- traveling toward the vicinity of the explosion.

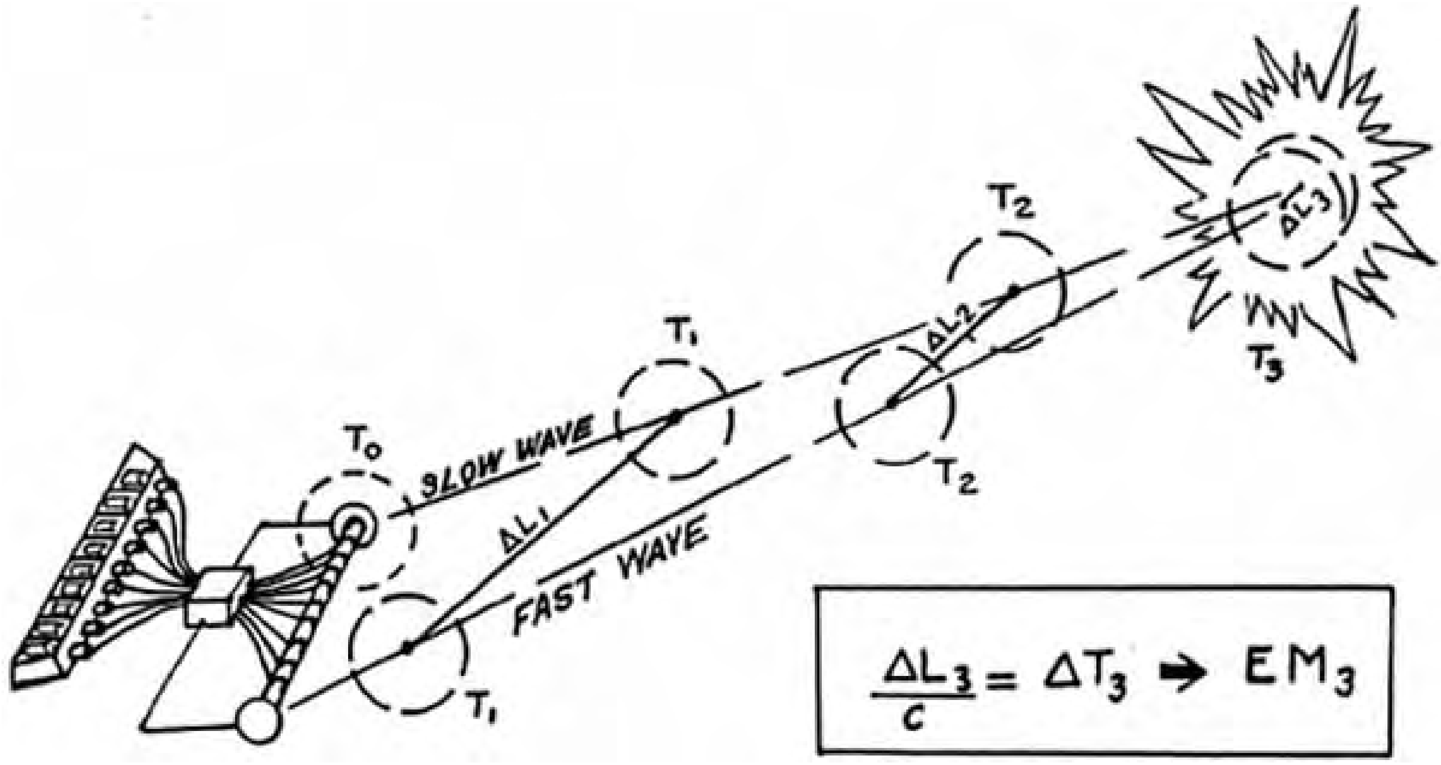
T₀
SLOW WAVE
T₁
T₂
T₃
ΔL1
ΔL2
ΔL3
FAST WAVE
T₁
T₂
ΔL₃/c = ΔT₃ → EM₃

TESLA HOWITZER
(SCALAR INTERFEROMETER)

Figure 10. "Nuclear" Flashes off the Coast of Africa

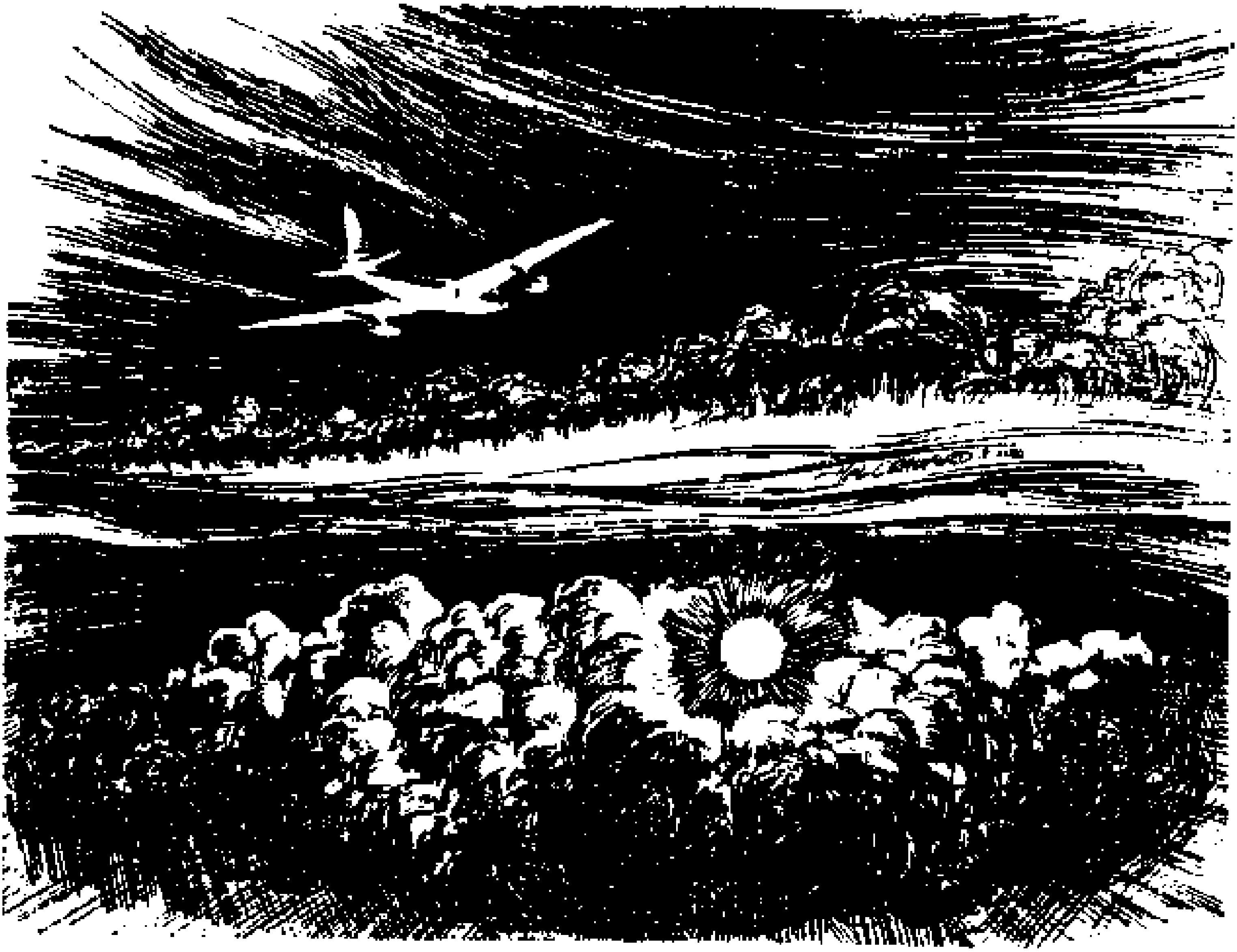

Lithuania - 10 Sep 1976 - British European Airways Flight #831 between Moscow and London

CIA Report Released under FOIA

Figure 11. Continuous Tesla Fireball

The pulse mode may be fed from either Moray generators or -- if the Moray generators have suffered their anomalous "all fail" malfunction -- ordinary explosive generators. Thus the Tesla howitzer can always function in the pulse mode, but it will be limited in power if the Moray generators fail.

In the continuous mode, two continuous scalar waves are emitted -- one faster than the other -- and they pair-couple into vector energy at the region where they approach an in-phase condition. In this mode, the energy in the distant "ball" or geometric region would appear continuously and be sustained -- and this is Tesla's secret of wireless transmission of energy at a distance without any losses. It is also the secret of a "continuous fireball" weapon capable of destroying hundreds of aircraft or missiles at a distance. An example of a Soviet test of this mode of operation is shown in figure 11.

Witness to a super weapon?

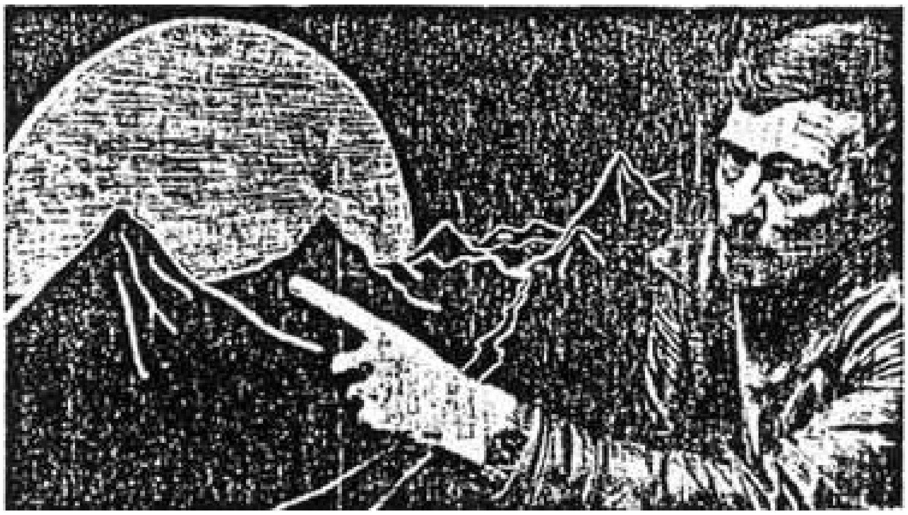

Nick Downie describes the strange lurid glow that flared silently over the Hindu Kush

THE SUNDAY TIMES, 17 AUGUST 1980

(Multiple incidents in Sept., 1979)

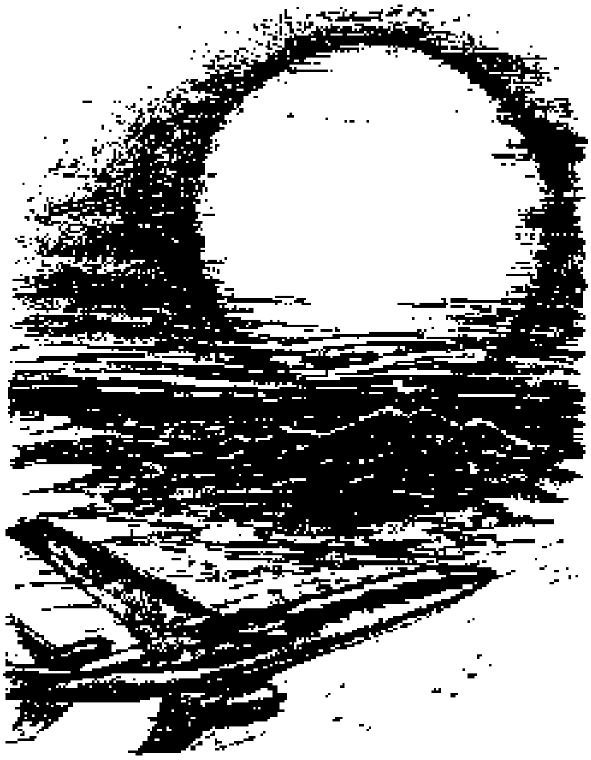

TEHERAN, IRAN 17 JUNE 1966

SEEN FROM NEAR MEHRABAD AIRPORT

OBSERVED 4-5 MINUTES

SEEN BY 2 AIRCRAFT

CIA REPORT RELEASED UNDER FOIA

Figure 12. Tesla EMP Globe

The volume of the Tesla fireball can be vastly expanded to yield a globe which will not vaporize physical vehicles but will deliver an EMP to them to dud their electronics. A test of this mode is shown in figure 12. (See also Gwynne Roberts, "Witness to a Super Weapon?", the London Sunday Times, 17 August 1980 for several other tests of this mode at Saryshagan, seen from Afghanistan by British TV cameraman and former War Correspondent Nick Downie.)

If the Moray generators fail anomalously, then a continuous mode limited in power and range could conceivably be sustained by powering the interferometer from more conventional power-sources such as advanced magnetohydrodynamic generators.

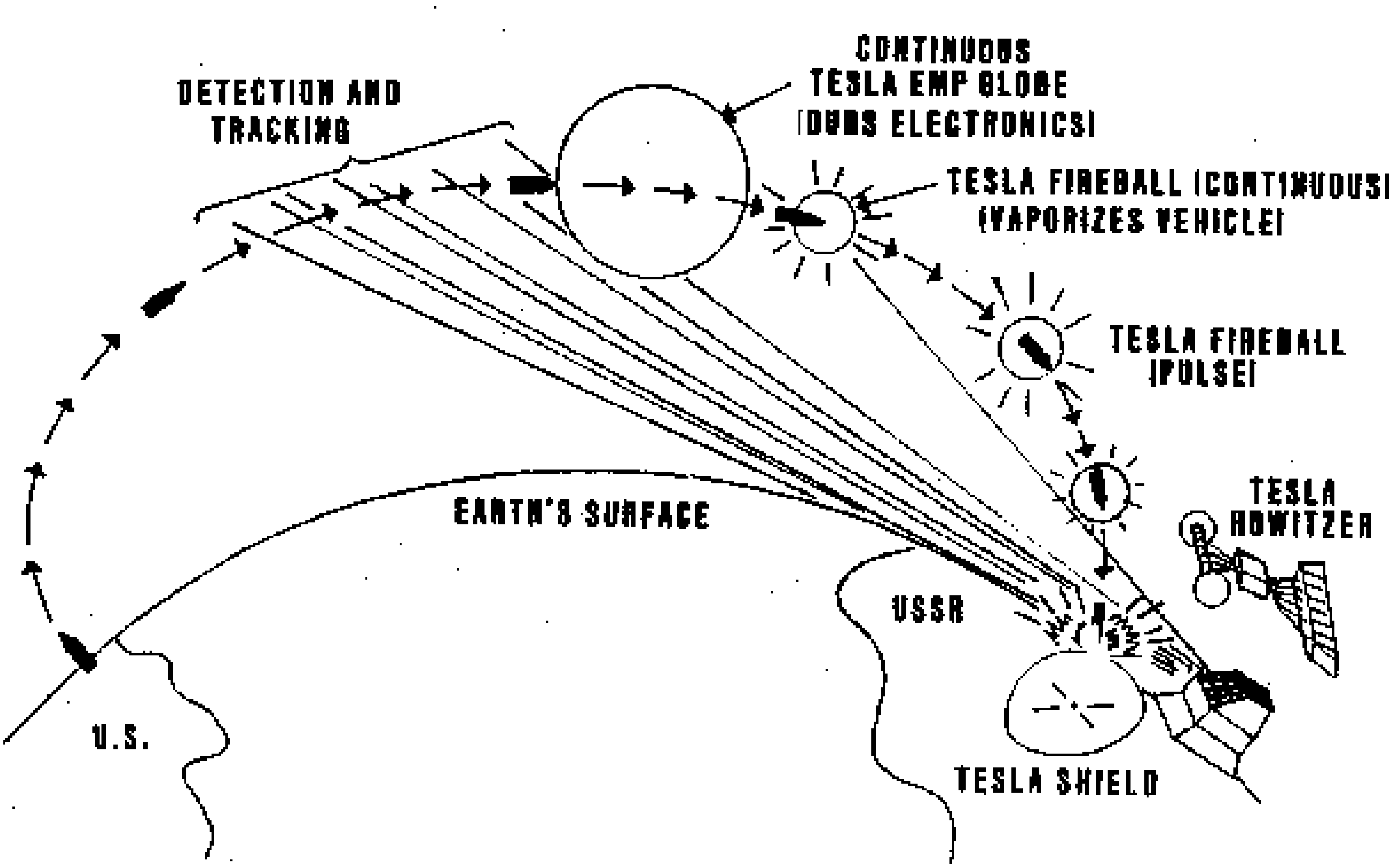
DETECTION AND TRACKING
CONTINUOUS TESLA EMP GLOBE (DUDS ELECTRONICS)
TESLA FIREBALL (CONTINUOUS) (VAPORIZES VEHICLE)
TESLA FIREBALL (PULSE)
TESLA HOWITZER
EARTH'S SURFACE
USSR
U.S.
TESLA SHIELD

Figure 13. Tesla ABM Defenses

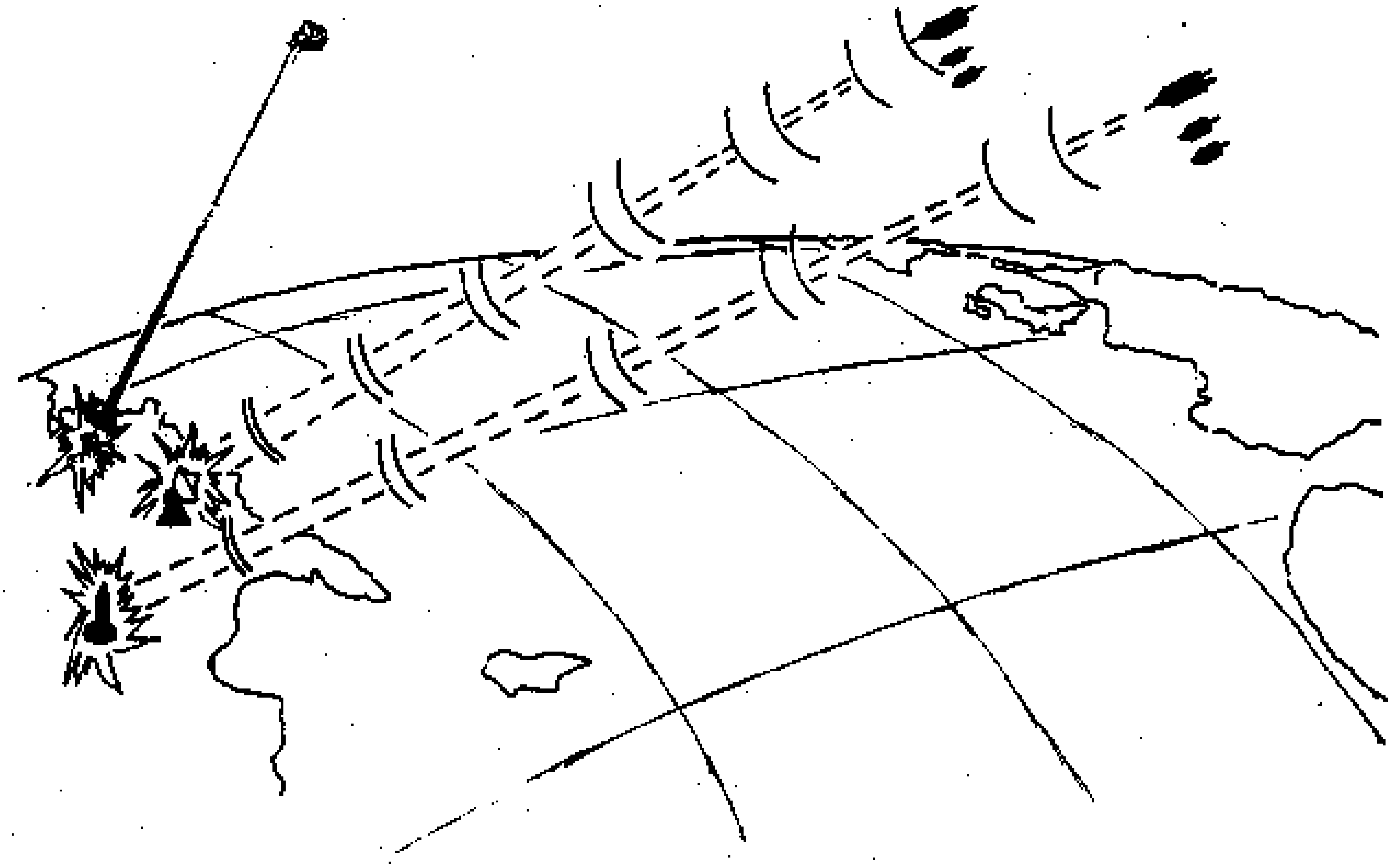

Figure 14. Moray/Tesla Technology: Star Wars Now

Typical strategic ABM uses of Tesla weapons are shown in figure 13. In addition, of course, smaller Tesla howitzer systems for anti-tactical ballistic missile defense of tactical troops and installations could be constituted of more conventional field missile systems using paired or triplet radars, of conventional external appearance, in a scalar interferometer mode.

With Moray generators as power sources and multiply deployed reentry vehicles with scalar antennas and transmitters, ICBM reentry systems now can become long range "blasters" of the target areas, from thousands of kilometers distance (figure 14). Literally, "Star Wars" is liberated by the Tesla technology. And in air attack, jammers and ECM aircraft now become "Tesla blasters." With the Tesla technology, emitters become primary fighting components of stunning power.

The potential peaceful implications of Tesla waves are also enormous. By utilizing the "time squeeze" effect, one can get antigravity, materialization and dematerialization, transmutation, and mindboggling medical benefits. One can also get subluminal and superluminal communication, see through the earth and through the ocean, etc. The new view of phi-field also provides a unified field theory, higher orders of reality, and a new super-relativity, but detailing these possibilities must wait for another book.

With two cerebral brain halves, the human being also has a Tesla scalar interferometer between his ears. And since the brain and nervous system processes avalanche discharges, it can produce (and detect) scalar Tesla waves to at least a limited degree. Thus a human can sometimes produce anomalous spatiotemporal effects at a distance and through time. This provides an exact mechanism for psychokinesis, levitation, psychic healing, telepathy, precognition, postcognition, remote viewing, etc. It also provides a reason why an individual can detect a "stick" on a radionics or Hieronymus machine (which processes scalar waves), when ordinary detectors detect nothing. Unfortunately there is not room to develop the implications of this human Tesla interferometry in detail, for that must wait for yet another book, presently in its initial stages, that Hal Crawford and I are writing.

Table 5. Orders of Reality

- **PHOTONS ARE**
  - **> PAIR-COUPLED SCALARS**
  - **> VELOCITY-LIMITED TO C**
  - **> CARRIERS OF T**

- **PHOTON INTERACTION**
  - **> IS UBIQUITOUS**
  - **> PRODUCES $\dot{T}$ = C**
  - **> YIELDS 1st ORDER REALITY**

- **SCALAR O WAVES**
  - **> NOT VELOCITY-LIMITED**
  - **> YIELD HIGHER ORDER REALITIES**

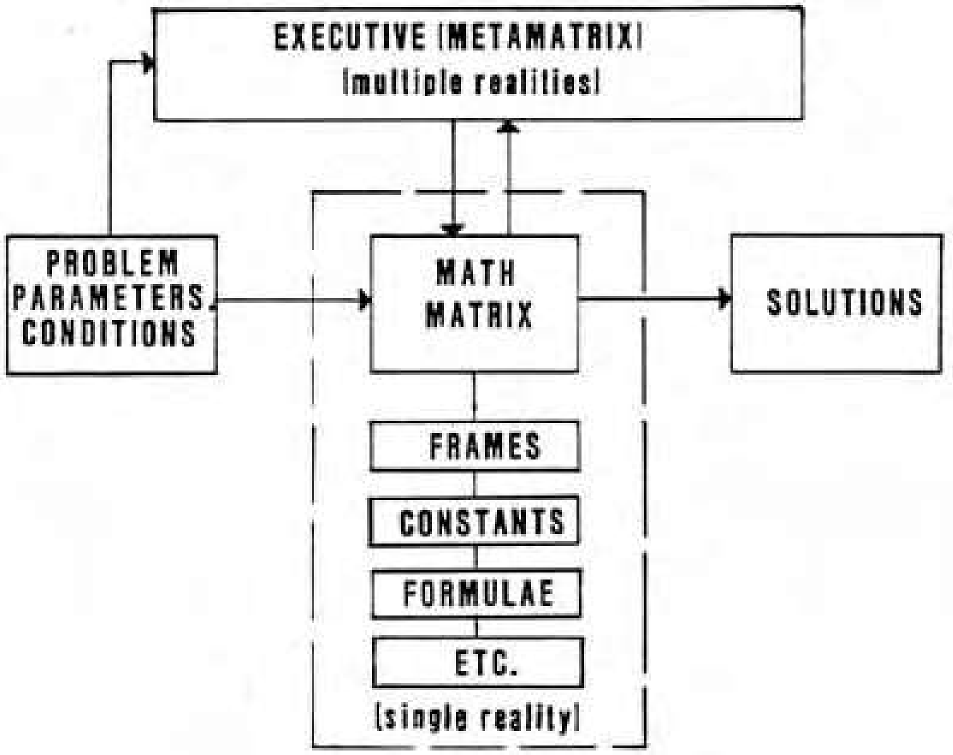
EXECUTIVE (METAMATRIX)
(multiple realities)
PROBLEM PARAMETERS, CONDITIONS
MATH MATRIX
SOLUTIONS
FRAMES
CONSTANTS
FORMULAE
ETC.
(single reality)

Figure 15. Implications of Tesla Potential

At the July 1981 U.S. Psychotronics Association's Annual Conference in Dayton, Ohio, I presented the first rough paper on the Tesla secret and scalar interferometry. A videotape of the presentation was made and will shortly be available. I am also scheduled to make a special presentation at the Alternate Energy Conference in Toronto, Canada in latter October, 1981. A professional, videotaped two-hour presentation on this subject is also being prepared. Wide distribution of the material through the international underground physics and technology network has already been made. This time, God willing, Tesla's secret will not be suppressed for another 80 years!

And perhaps it is not yet too late. The material has cost me (now) some 16 years of agonizing labor and nearly $100,000 of my own personal funds. No orthodox university, scientific group, foundation, or governmental agency would support such an effort, either financially or otherwise. Indeed, most ordinary journals will not even accept material on such matters. Nonetheless, the area is of overwhelming importance and I truly believe Tesla's lost secret will shortly affect the lives of every human being on earth.

Perhaps with the free and open release of Tesla's secret, the scientific and governmental bureaucracies will be. shocked awake from their slumber, and we can develop defenses before Armageddon occurs. Perhaps there is hope after all -- for even Brezhnev, in his strange July, 1975 proposal to the SALT talks, seemed to reveal a perception that a turning point in war and weaponry may have been reached, and that human imagination is incapable of dealing with the ability to totally engineer reality itself. Having tested the weapons, the Soviets must be aware that the ill-provoked oscillation of timeflow affects the minds and thoughts -- and the very lifestreams and even the collective species unconsciousnesses -- of all lifeforms on earth. They must know that these weapons are two-edged swords, and that the backlash from their use can be far more terrible to the user than was the original effect to his victim.

If we can avoid the Apocalypse, the fantastic secret of Nikola Tesla can be employed to cure and elevate man, not kill him. Tesla's discovery can eventually remove every conceivable external human limitation. If we humans ourselves can elevate our consciousness to properly utilize the Tesla electromagnetics, then Nikola Tesla -- who gave us the electrical twentieth century in the first place -- may yet give us a fantastic new future more shining and glorious than all the great scientists and sages have imagined.

References

The Tom Bearden Website

REFERENCES

1. Whittaker, E. T., Proc. Lond. Math. Soc. 1, 367 (1903).

2. Debye, P., Ann. Phvs. (Leipz.) 30, 57 (1909).

3. Bromwich, T. J. I'A. Phil. Trans. A, 220, 175 (1920); also Phil. Mag., 38, 143 (1919).

4. Laporte, O. & Uhlenbeck, G. E., Phys. Rev. 37, 1380 (1931).

5. Nisbet, A., Proc. R. Soc. London A 231, 250 (1955).

6. Essex, E. A., Am. Jour. Phys. 45, 1099 (1977).

7. Braunlich, P., Ed., Thermally Stimulated Relaxation in Solids, Springer-Verlag, New York, 1979.

8. Ratzlaff, John. T., Dr. Nikola Tesla: Selected Patent Wrappers, Volumes I, II, III, & IV, Tesla Book Company, 1580 Magnolia Avenue, Millbrae, CA 94030, 1980.

9. Ratzlaff, John T. & Anderson, Leland I., Dr. Nikola Tesla Bibliography, Ragusan Press, 936 Industrial Ave., Palo Alto, CA 94303, 1979.

10. "Tesla- 85th Birthday," N. Y. Sun, July 11, 1941. (Tesla says he could build, within three months, a plant at a cost of $2 million that would melt the engines of an approaching aircraft at a distance of thousands of miles.)

11. "Tesla Promises to Light Dark Spot on Moon. It's Part of a Scheme of His for Interplanetary Radio; Distance Means Nothing. Has 4 New Inventions. Tells of Them on 81st Birthday; 2 Nations Honor Him," N. Y. Herald Tribune, July 11, 1937.

12. O'Neill, John J., "In the Realm of Science: Tesla, who predicted radio, now looks forward to sending waves to the Moon," N. Y. Herald Tribune, Aug. 22, 1937. (Inventor hopes to use energy-transmitting device to make spot glow on the lunar surface. Theory is traced to 1897 experiments. His mechanism is to use vast natural forces, possibly cosmic rays.)

13. "Tesla, 80, Reveals New Power Device," N. Y. Times, July 11, 1936, p. 13, col. 2. (Says his wireless system of power transmission will supply the earth with energy for industry.)

14. Sparling, Earl. "Nikola Tesla, at 79, Uses Earth to Transmit Signals; Expects to have $100,000,000 Within Two Years," N. Y. World-Telegram, July 11, 1935. (Inventor tells of 'quake' in his laboratory that brought police and ambulances during experiments with mechanical oscillator.)

15. "Tesla's Controlled Earth Quakes Power Through the Earth, A Startling Discovery," N. Y. American, July 11, 1935, Section 2. (Announces the successful passage of an induction current with a varying flux through a circuit without the use of a commutator. Cosmic ray studies indicate many tenets of theory of relativity to be fallacious. Possible to convey mechanical effects to any distance.)

16. "Tesla, 79, Promises to Transmit Force -- Transmission of Energy Over World," N. Y. Times, July 11, 1935, p. 23, col. 8. (Tesla has plans to send energy over the entire world. Measurement of cosmic rays said to be 50 times greater than the speed of light, demolishing the theory of relativity.)

17. Tesla, Nikola, "Expanding Sun Will Explode Some Day, Tesla Predicts," N. Y. Herald-Tribune, Aug. 18, 1935. (Present literature on cosmic rays is erroneous. Some cosmic rays reach speeds 50 times that of light. Sun will increase in mass and energy and will ultimately explode. Tesla's view is that the condensation of primary substance is going on continuously. Finds secret of cosmic rays in the positive electrical charge of the sun. Discusses radioactive emanations.)

18. "3 Tesla Inventions -- Famous Scientist Will Tell Them. Tomorrow," N. Y. Sun, July 9, 1935. (One of the discoveries is a new way of transmitting energy, an entirely new principle, nothing like wireless. Also method of harnessing cosmic rays.)

19. Welshimer, Helen, "Dr. Tesla Visions the End of Aircraft in War," Every Week Magazine, Oct. 21, 1934, p. 3. (Claims to have created a new agent, which kills without a trace and yet pierces the thickest armor. Can destroy armies or aircraft.)

20. Tesla, Nikola. "Tesla on Power Development and Future Marvels," N. Y. World Telegram, July 24, 1934. (Source of reference is Prodigal Genius by J. J. O'Neill, p. 241. Reply to articles of June 29, July 12, and July 13, 1934. Praises Westinghouse and Insul to create the power system he had barely suggested in 1893. With regard to death ray effect, Tesla employs an agent

in which intensity does not diminish with the square of the distance.)

21. Dunlap, Orrin E., Jr., "Tesla Sees Evidence That Radio and Light Are Sound, " N. Y. Times, Apr. 8, 1934, X, p. 9, Col. 1. (Tesla points to errors of the past, explains radio as he sees it at age 77. He expects television.). (Note: Sound is a longitudinal wave -- as is the Tesla wave. Hertzian waves are transverse waves, not longitudinal.)

22. Bird, Carol, "Tremendous New Power Soon to be Unleashed," Philadelphia Public Ledger, Sep. 10, 1933, Magazine Section, p. 6. (Revolutionary power project by Tesla, who is also completing process for thought photography.) (Note: Here we have a possible clue that the same principle may be used both in Tesla's energy device and in Tesla's approach to thought photography.)

23. "Tesla 'Harnesses' Cosmic Energy," Phila. Public Ledger, Nov. 2, 1933. (A principle has been discovered to derive cosmic energy which operates the universe. Power is "everywhere present in unlimited quantities." Will eliminate the need for coal, oil, gas, or any of the common fuels.) (Note: Was Tesla referring to what we today call "zero-point energy of the vacuum"? His principle will tap it.)

24. Blakeslee, Howard W., "Discovery of Force to Surround Nations & Smash Attacker Claims of Aged Inventor - Nikola Tesla Makes Announcement on 75th Birthday -- Will Turn Plans Over to Geneva," Minneapolis Tribune, July 11, 1934. (Note: Here the inventor is referring to the "Tesla Shield.")

25. "Tesla, at 78, Bares New 'Death Beam'", N. Y. Times; July 11, 1934, p. 18, col. 1. (Invention powerful enough to destroy 10,000 airplanes at 250 miles away. Defense weapon only.)

26. Alsop, Joseph W., Jr., "Beam to Kill Army at 200 Miles Tesla's Claim on 78th Birthday," N. Y. Herald Tribune. July 11, 1934, pp. 1, 15. (Beam of force similar to death ray, involves four electrical devices. Can also be used in peacetime to transmit power over distances limited only by the curvature of the earth.) (Note: In this reference, we note that the Tesla death ray and the Tesla wireless transmission devices apparently use the same effect or basic principle. There appear to be four devices involved in an operational system.)

27. "A Giant Eye to See 'Round the World' ", Albany Telegram, Feb. 25, 1933. (Based on the mechanism of the human eye. The first two parts of Tesla's invention have been completed. Will allow man to see any part of the earth.) (Note: Similarity of binocular vision to interferometry.)

28. Tesla, Nikola. "Pioneer Radio Engineer Gives Views on Power," N. Y. Herald Tribune, Sept. 11, 1932. (Tesla says wireless waves are not electromagnetic, but sound in nature. Holds space is not curved.) (Note again that sound waves are longitudinal, as are Tesla's waves. Electromagnetic waves are transverse.)

29. "No High-Speed Limit, Says Tesla, " Literary Digest, Nov. 7, 1931, p. 28. (Speeds greater than light, deemed impossible by the Einstein theory, have been accomplished. As early as 1900 Tesla showed that power from his transmitter passed over the earth at a speed of 292,830 miles per second.) (Note: As is well known, velocity represents the rotation of an object out of the normal 3-dimensional space toward a 4th dimensional direction. The speed of light, c, represents a full orthogonal turn. A normal electromagnetic wave, being transverse oscillatory, is a priori limited to a single orthogonal turn, for that is all it possesses. It thus moves at the speed c. A longitudinal scalar wave, on the other hand, need not be so limited at all.)

30. O' Neill , J. J. "Sun Emits Super Ray, Nikola Tesla Asserts, " Brooklyn Eagle, Feb. 8, 1932, p. 4, col. 1. (Note: Logically, the sun should also emit scalar waves as well as transverse vector waves.)

31. O'Neill, J. J., "Tesla Cosmic Ray Motor May Transmit Power 'Round Earth'", Brooklyn Eagle, July 10, 1932, A, 1:4, pp. 1, 17. (Efforts by Tesla to harness cosmic rays started 25 years ago, and he now announces success in operating a motive device by means of these rays. Hopes to build a motor on a large scale.)(Note: possibly implies that all these Tesla devices for the last 25 years have been dealing with the same basic principle.)

32. "Tesla, 76, Reports His Talents at Peak," N. Y. Times, July 10, 1932, p. 19, col. 1. (New invention in tapping tremendous and unused source of energy. One invention to permit generation of all kinds of rays of almost unlimited intensity.) (Note: By scalar interferometry, electromagnetic waves of any frequency -- and hence of any "type" can be assembled by coupling the appropriate scalars together. A Hertzian wave is just two coupled Tesla scalar waves.)

33. Tesla, Nikola, "Man's Greatest Achievement," N. Y. American, July 6, 1930, p. 10. (Editorial Section). ("To create and annihilate material substance, cause it to aggregate in forms according to his desire...would place him beside his Creator and fulfill his ultimate destiny.") (Note: Here Tesla seems to be referring to controlled materialization and dematerialization of matter.)

34. Tesla, Nikola. "World System of Wireless Transmission of Energy," Telegraph & Telephone Age -- N. Y., Oct. 16, 1927, pp. 457-460.

Transmission of power without wires is not a theory, but "a fact demonstrated by Tesla." Mode of propagation of currents from transmitter has mean speed 57% greater than Hertz waves.) (Note: a faster than light, non-Hertzian type of wave is involved. The Tesla wave is not a normal Hertzian wave.)

35. Secor, H. Winfield, "The Rogers Underground Wireless, Electrical Experimenter, Mar., 1919, pp. 787-789, 832-835, 839. (Rogers system does away with aerial wires.)

36. Gernsback, Hugo, "Underground Wireless," Electrical Experimenter, Mar. 1919, p. 762. (Development of James H. Rogers. Receives messages from Europe during thunderstorm. Tesla agrees that messages are not Hertzian waves.)

37. Tesla, Nikola, "The True Wireless," Electrical Experimenter, May 1919, pp. 28-30, 61-63, 87. (Tesla says Hertz wave theory is a delusion. Signals must be from earth currents. Note again that Tesla does not like Hertzian waves, but himself has something more fundamental.)

38. Tesla, Nikola. "The Effects of Statics on Electrical Transmission," Electrical Experimenter, Jan., 1919, pp. 627, 658. (Tells of defects in Hertz waves. System devised by Tesla releases energy at infinite velocity.) (Note: electrostatic potential is already known to be able to travel at infinite velocity. See Jackson, Classical Electrodynamics, 2nd Edition, 1975, p.223. This is strong evidence that Tesla used scalar potential waves or Ø-fields.)

39. Bot tone, A., Nikola Tesla's New Wireless," Electrical Engineer, London, Dec. 24, 1909, p. 893. (Experiments conducted on Long Island have resulted in principles of transmission which are direct opposite of Hertzian wave transmission.) (Note: In other words, using uncoupled scalar waves -- Tesla waves instead of pair-coupled scalar waves -- Hertzian waves?)

40. Tesla, Nikola. "Electrical Control of the Weather Will Soon Be an Accomplished Fact," St. Louis Republic, Nov. 15, 1908, V, p. 3. (Tesla concurs in concept of electrical control of weather.)

41. Swezy, Kenneth M. , "Nikola Tesla, " Science, May 16, 1958 , pp. 1147-1158. (Electricity today is generated, transmitted, and converted to mechanical power by means of his inventions.)

42. Ramsay, Jack, "Cableless Power, " Electronics (Digest) , July 22, 1960, pp. 6, 8. (Soviet attempt at development of wireless power transmission, but a failure according to Swedish observers, and the idea having been bandied about in the U. S. since Tesla.) (Note: Establishes that the Soviets were at least trying to work with the Tesla techniques.)

43. Anderson, Leland I., "Correspondence: Sub-Surface Communications Systems," Proceedings of the I.R.E., Mar., 1961, p. 645. (Reference to subsurface or "earth current" communication systems. Pioneering work by Nikola Tesla and James H. Rogers.)

44. Rebert, Edwina, "Fireballs For Defense?", Christian Science Monitor, Feb. 7, 1962, p. 9c. (Feature article on possibility of using ball lightning as defense weapon -- may be inspired by early experiments of Tesla at Colorado Springs.)

45. Matsch, Lee and Rice, Warren, "Potential Flow Between Two Parallel Circular Disks with Partial Transmission," Journal of Applied Mechanics, Trans. ASME, Mar., 1967, Vol. 34, Series E., No.1, pp. 239-240. (Reference article -- bladeless turbines.)

46. Lagus, George W., "The Use of Ground in High-Frequency Circuits," Radio-TV Experimenter and Science and Electronics, June-July, 1969, PP. 71, 72. (Tesla gave several demonstrations of feasibility of electrifying large areas of land by ground propagation of high frequency currents.)

47. Jamison, S. L. , "Life Energy, " Probe. The Unknown, June , 1973, pp. 3,4. (Letter to editor -- report that vitamins and minerals could be impressed on the body using a Tesla coil.)

48. Jueneman, F. B. J. , "The First Law of Thermodynamics, " Industrial Research, Feb., 1974, pp. 17-18. (Nikola Tesla, at Colorado Springs in 1899, found that the earth was a large electron sink, and electrical storms generated standing wave potentials.)

49. Puharich, Andrija, "The Work of Nikola Tesla Ca. 1900 and its Relationship to Physics, Bioenergy and Healing," paper read at the International Inter-disciplinary Conference on Consciousness and Healing, Oct. 13, 1976, University of Toronto.) ,

50. Golka, Robert K. and Bass, Robert W., "Tesla's Ball Lightning Theory, a BKG-Wave, The Ransworth Effect, and the 'Hydrotron' Electrostatic-Inertial Self-Confined Plasmoid Concept." (Paper presented at the Annual Controlled Fusion Theory Conference, San Diego, California, May 4-6, 1977.)

51. Popovic, Prof. Vojin, "Research of Nikola Tesla in the Light of His Diary From Colorado Springs," 10 pp. (Diary contains details of investigations for determining nodal points in the earth as a conductor of waves. Comments on "fire globes.")

52. Curtis, George D., PhD, "An Electromagnetic Radiation Pattern Over the Ocean," Undersea Technology, Vol. 5, no.8, August 1964. Curtis reports the presence of a previously unknown weak electromagnetic radiation pattern over the ocean. When all known effects are accounted for, the anomalous pattern still remains.

53. Santilli, R. M., "Partons and Gravitation: Some Puzzling Questions," Annals of Physics, Vol. 83, No.1, March 1974, pp. 108-157. In this paper Santilli proved that one of the cornerstone assumptions of physics -- that electric field and gravitation are separate things -- is false. One is therefore left with only two alternatives: either they are totally the same thing, or they are partially the same thing. (Comment: Note that the ~4 Tesla potential provides the unifying connection.)

54. Science News, Vol. 113, No.1, January 7, 1978, p. 3. A photograph by T. Kuribayashi of the earthquake lights involved in the Matsushiro earthquake swarm (1965-1967). This is the only known photo of earthquake lights.

55. Wiedemann, C. Louis, "Results of the N.J. 'Spook Light' Study," Vestigia Newsletter, Vestigia, RD 2, Brookwood Rd., Stanhope, NJ 07874, May 1977, pp. 1-3. This article describes an ongoing Vestigia experiment in which a nocturnal light or "mystery light" was scientifically photographed in conjunction with visual sightings and instrumental recordings of data. Various later issues of the Vestigia Newsletter contain other update articles on the phenomenon.

56. Moray, T. Henry, The Sea of Energy, fifth edition, History and Biography by John E. Moray, Foreword by Tom Bearden; Cosray Research Institute, 2505 South 4th East, Salt Lake City, Utah 84115, 1978. In addition to his free-energy device, Moray also built a special radio which he used to clearly listen to Admiral Byrd at the South Pole. No static was experienced on Moray's radio, showing he was' not using ordinary electromagnetics. In addition, he built a device which he could tune to listen to persons several miles distant, so that their voices sounded as if they were immediately nearby. Both these devices were demonstrated to a visiting Russian Colonel/ PhD.

57. I. Procaccia, J. Ross, Science, 198, 716 (18 Nov 1977). Describes Prigogine's Nobel Prize-winning work on thermodynamics of nonlinear systems far from thermodynamic equilibrium. Prigogine's work shows that such a system can indeed demonstrate negentropy. Out of unrestrained disorder can and does arise order, contrary to the old thermodynamics.

58. "Persinger's 'Earthquake Lights' ...Ho-Hum," Frontiers of Science, Vol. III, No.3, March-April 1981, p. 15-16. Gives cogent and "fatal" objections to

Persinger's "earthquake light" scheme for explaining UFO's. Mentions Dr. Brian Brady's experiments obtaining fast, short electrical lights or sparks from crushing quartz-bearing granite cylinders. (Comment: Brady's experiments are certainly repeatable, so it is true that "earthstress lights" can be produced at a distance, although the piezoelectric effect alone in no way explains the production of the phenomena outside the piezoelectric crystal. The fact that UFO's occurring near a fault zone are not necessarily repeatable simply means that at least one other major factor must be involved in the production of UFOs than is involved in the production of earthstress lights. UFOs may sometimes involve scalar interferometry from a fault zone, but other causative factors must also be involved. Earthstress lights, on the other hand, need have no additional causative factors than what lies in the earth itself.)

59. Jackson, John David, Classical Electrodynamics, Second Edition, John Wiley & Sons, New York, 1975, p. 223.

60. Roberts, Gwynne, "Witness to a Super Weapon ?", The Sunday Times, London, 17 August, 1980. Article deals with a strange semicircular phenomenon seen inside the Soviet Union from Afghanistan in September, 1979 by Nick Downie, a former member of the SAS who has established a considerable reputation as a war cameraman. Downie saw the phenomenon twice, and later met an Afghan who had seen it several months before. A possibly related phenomenon is reported in Brackenbridge, M., "Unidentified Phenomenon," Marine Observer, 48: 21-22, 1978. Note the first "nuclear flash" was detected by Vela satellite on September 22, 1979 - close in time to Downie's sightings.

61. Bearden, Tom and Crawford, Hal, "Possible Soviet Test of a Tesla Weapon," Specula, Journal of the A.A.M.S., Vol. 3, No.2, April-June 1980, p. 29. See also Specula 3, 2, 30-32 by Bearden and Crawford. These reports are from CIA reports released under the Freedom of Information Act.

62. Bearden, Thomas E., The Excalibur Briefing, Strawberry Hill Press, San Francisco, 1980.

63. Muldrew, D. B., "Generation of Long-Delay Echoes," Journal of Geophysical Research, 84: 5199-5215, 1979. For anomalous long delays of electromagnetic signals in the atmosphere, Muldrew favors a rather complex interaction between signals from separate transmitters that (theoretically at least) can create a long-lived electrostatic wave that travels in the ionosphere -- a sort of natural memory device. The coded signals could then be read out much later when the proper natural conditions developed. Delays of up to 40 seconds might be possible with this "ionospheric memory."

64. Zhugzhda, Yu. D., Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation, "Magnetogravitational Waves in a Conducting Isothermic Atmosphere," Moscow ASTRONOMICHESKIY ZHUGZHDA, Vol. 56, No.1, 1979, pp. 74-83. Among other things, the paper indicates the possibility of transforming longitudinal waves into transverse waves in the region of a strong magnetic field.

65. Ranada, Antonio and Vazquez, Luis, "Kinks and the Heisenberg Uncertainty Principle," Physical Review D, Vol. 19, No.2, 15 January 1979, pp. 493-495. Shows that the velocity and position of the center of a kink, soliton, or solitary wave can be known with arbitrary precision, in violation of the Heisenberg uncertainty principle. (Comment: A Fourier-transformed scalar Tesla wave pattern is either a soliton or closely related to it.)

66. Don Moser, Photographs by Blair Pittman, "Big Thicket of Texas, National Geographic, Vol. 146, No.4, October 1974, pp. 504-529. The author photographed a "ghost light" or nocturnal spooklite -- i.e., an earths tress light. The photo is shown in the article.

67. Gehring, Gillian, "Actinide magnetism: an extraordinary tale," Nature, Vol. 279, 3 May 1979, pp. 16-18. Anomalous magnetic properties -- including spin orbit coupling as a large effect -- are exhibited by actinide intermetallics containing uranium. Present theory cannot explain these effects.

68. Gintsburg, M. A., Institute of Terrestrial Magnetism, the Ionosphere, and Radio-Wave Propagation, Academy of Sciences of the USSR, Astron. Zh. 51: 218-221, Jan-Feb 1974, English translation in Sov. Astron., Vol. 18, No.1, July-August 1974, pp. 128-130. Among other things, the article points out that thus far a Fourier analysis in mass electrodynamics has not been made. One property characteristic of mass electrodynamics is the existence of longitudinal electromagnetic waves in vacuo, i.e., longitudinal photons. They have not yet been discovered but theoretically they are entirely possible.

69. Burman, R., "A Photon rest mass and the propagation of longitudinal electric waves in interstellar and intergalactic space," J. Phvs. A: Math. Nucl. Gen., Vol. 6, March 1973, pp. 434-444. This paper on longitudinal electric waves in plasma may be very important. Weber may be detecting longitudinal electric waves. Proca waves can go faster or slower than c.

70. Klass, Philip J., "Anti-Satellite Laser Use Suspected," Aviation Week & Space Technology, December 8, 1975, pp. 12-13. Some details on the "laser blinding" of U. S. satellites by the Soviets. On one occasion, the satellite was blinded for 4 hours. (Comment: difficult to explain with current laser

technology, but simple to explain with Tesla scalar interferometry.)

71. "Soviets Build Directed-Energy Weapon," Aviation Week & Space Technology, July 28, 1980, pp. 57-50. Presents a background summary in Soviet particle beam weapon work, particularly the device being constructed at Saryshagan.

72. "Scientists Fail to Solve Vela Mystery," Science, Vol. 207, 1 February 1980, pp. 504-506. See also article in Science, 30 November, 1979.

73. "Satellite Evidence Shows 'Possibility' of Nuclear Test, DOD Says," Aerospace Daily, October 29, 1979, p. 286.

74. "A Flash of Light," Newsweek, November 5, 1979, pp. 64-65.

75. "Was It a Nuclear Device?", Newsweek, July 21, 1980, p. 19.

76. "A Nuclear Blast -- or 'Zoo Animals'?", Newsweek, April 7, 1980, p. 21.

77. "Diverging Views," Washington Roundup, Aviation Week & Space Technology, July 21, 1980, p. 15.

78. Klass, Philip J. , "Clandestine Nuclear Test Doubted," Aviation Week & Space Technology, August 11, 1980, pp. 67, 69, 71-72.

79. "Debate Continues on the Bomb That Wasn't," Science, Vol. 209, 1 August 1980, pp. 572-573.

80. "Navy Lab Concludes the Vela Saw a Bomb," Science, Vol. 209, 29 August, 1980, pp. 996-997.

81. Bloch, Ingram and Crater, Horace; "Lorentz-invariant potentials and the nonrelativistic limit," Am. J. Phys. 49(1), Jan. 1981, pp. 67-75. Among other things, this interesting paper points out some decidedly unusual and unfamiliar influences of the size of the scalar potential upon what is considered nonrelativistic behavior. For scalar potential energyof appreciable size relative to a particle's rest energy, ordinary Newtonian mechanics and the Schrodinger equation may be inadequate, even if v/c is small. This reference by itself is sufficient to establish that Ø-field alone can alter the flow rate of time.

82. Semon, Mark D. and Schmieg, Glenn M., "Note on the analogy between inertial and electromagnetic forces," Am. J. Phys. 49(7), July 1981, pp. 689-690. Compares physical forces in a rotating frame with electromagnetic forces in a Newtonian frame. The transverse mechanical force in a rotating

frame is associated with an induced EMF in a Newtonian frame.

83. Harlacher, Von Wolfgang M., "Bomben aus dem Hyperraun," Esotera, 4 April 1979, pp. 359-365.

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

TABLE OF CONTENTS

Cover

List of Figures

Implications

Electrical Physics Presently Has a Mindset

It Started With Geometry and Grew

Points and Motion

Four Types of Vectors Actually Emerged

Quantum Mechanics Compounds the Problem

To Summarize Briefly

Uncharged Spatial (Massless) System Vector

Chargeless Mass System Vector

What Force Is

Force and Hertzian Waves Cannot Exist in Vacuum

Charged-Mass-System Vector

Charged Spatial (Massless) Vector

The Shadow Vector

A Scalar is a Zero Vector

Virtual and Observable Aspects

Scalars and Vectors Have Substructures

Substructures, Virtual Levels, and Hyperspaces

Notes and References

Additional Notes and References

TOWARD A NEW ELECTROMAGNETICS

PART III: CLARIFYING THE VECTOR CONCEPT

by

T. E. Bearden

© Copyright 1983

by T. E. Bearden

All rights reserved

LIST OF FIGURES

1. Uncharged spatial (massless) system vector
2. Chargeless mass system vector
3. Detection of "transverse" and longitudinal waves
4. Charged mass system vector
5. The "charge" on an electron mass consists of a flux of virtual particles on and off the mass
6. A test charge (charged mass) brought near a fixed charge (charged mass) experiences an acceleration
7. A charged mass system vector
8. Repeating the "test charge" experiment
9. Charged spatial (massless) vector
10. Removing the bare particle (mass) from a charged particle leaves the charge. The vacuum is DEFINED AS the charge
11. Assigning a spatial vector to the charged vacuum
12. A "shadow vector" ε_s
13. Substructures of a "zero resultant" spatial vector. (The substructures of two zero vectors may be vastly different)

The Tom Bearden Website

TOWARD A NEW ELECTROMAGNETICS

Part III: Clarifying the Vector Concept

© 1983 T.E. Bearden

-- IMPLICATIONS --

Some of the fundamental concepts of the new Tesla electromagnetics are presented. The new concepts have startling implications:

(1) No force or force field exists as such in vacuum.

(2) Hertzian (transverse) electromagnetic waves do not travel through the vacuum, just as Tesla stated.

(3) Forceless, massless Tesla (scalar) longitudinal waves actually transit the vacuum. Tesla called them "electrical sound waves."

(4) At present there are actually four different FUNDAMENTAL TYPES of vectorial entities in physics, erroneously confused as one and the same.

(5) Tesla longitudinal scalar waves are also "time" waves and can affect anything and everything that exists in time.

(6) The fundamental constants of nature (which exist in time) can be altered by Tesla scalar waves, which oscillate the values of the constants.

(7) Every vector and scalar has an internal substructure, which can be independently affected and changed. This allows the direct engineering of the virtual state and the vacuum itself.

(8) All observable forces (electrical, mechanical, gravitational, etc.) arise in, on, and OF the actual substructure of the "accelerating mass particle" itself, not as an "external" massless force or force field applied "to" a mass .

(9) Physical reality itself -- and the "physical laws of nature"

can be deliberately changed and engineered.

(10) All "physical reality" is totally internal to the physical changes of the mass particles of the detector system of the observer.

(11) Relativity's speed of light limitation applies only to the changes of the basic mass particles of the detecting instrument.

(12) Detection of superluminal effects cannot be accomplished by a "single stage" or "single shift " (single interaction) detector .

(13) Detection of superluminal effects is permitted by "multiple stage" or "multiple shift" interactions where the last interaction is a conventional interaction of photon vs. detector particle. (The two-slit apparatus for detection of electron diffraction is an example. First, the superluminal DeBroglie waves interact with the slits, which are "tuned" toward the electron's DeBroglie wavelength. The interaction with TWO slits produces subluminal interference effects, which then interact back upon the physical electron. The apparatus is thus an electron interferometer capable of detecting superluminal waves by a two-stage interaction).

(14) Interference is the most common first-stage superluminal interaction to accomplish "downshifting" superluminal entities to luminal or subluminal velocities. Superposition of superluminal "phase" waves (such as deBroglie waves, which individually always move faster than the speed of light) interferes the waves to create a subluminal group velocity, which may then interact with an ordinary mass particle in the detection system.

(15) Any otherwise physical vector must exist as an unzipped (segmented) or "shadow" vector in vacuum. "Radiation" of a vector EM wave from the electron gas in an antenna into vacuum results in the "choking off" of the mass of the transversely oscillating electrons in the antenna. Since the spinning electron mass is the "zipper" that makes or comprises the physical vector in the first place, this throttling of the mass flow unzips the E and B vectors, leaving whirling (massless magnetic scalar potential) segments of massless charge flux (massless electrostatic scalar potential). This unzipped whirling pattern of charge flux (scalar massless A/Ø) is what radiates into vacuum and propagates through it. This is a special kind of

scalar wave pattern, not a physical or vector wave.

(16) The spin of a charged particle is the mechanism for integrating or "zipping together" the individual virtual fragments of a shadow vector into a real (observable) vector. For "uncharged particles" such as neutrons, it is the spins of its virtual charged components. that accomplish the integration or zipping.

(17) All fundamental charged particles are constantly accelerated. There is no such thing as an "unaccelerated" particle, except as a gross average over time or length. Further, all of them are spinning.

(18) All changes to and from a physical vector or scalar system must arise in and come from its own internal substructure, which is zipped to its spinning particle of mass.

(19) All fundamental particles are charged internally. That is, they are dynamic assemblages of smaller charged particles. If the average sum of the total internal charge is essentially zero over some finite, small increment of time, the particle is externally uncharged. If the sum is not essentially zero, the particle is also externally charged.

(20) There are no static physical things in existence. In physical reality, something appears "static" only at a particular level. Upon sufficiently fine examination, it is composed of accelerating parts, and thus comprised of "fluctuations."

(21) Since (a) the basic physical (mass) vector consists of a "smeared particle," where particle and smear are inseparable, (b) the conceptual particle also is accelerated, and (c) the "smearing" is for a small increment of time and a small increment of length; then the basic constituency of "physical reality" is inseparable "force x time x length," or action. The basic "quantum" of physical change is thus comprised of action.

(22) Since to "detect" we must "stop" the action, separate or split the quantum into two pieces ("canonical" pieces) , and compare (measure) one piece by throwing away the other, then each physical observable must have a differential operator (the "separating agent") corresponding to it. This accounts for the fundamental postulate of quantum mechanics whereby every observable has a corresponding operator. Further, since what remains is totally relative to what was split out and thrown away, physical change

is totally "relative." This accounts for the fact that observed reality is relative, each part to each other.

(23) As a special case, we may assume that we can evaluate a physical change at a point (without length). If so, when we discard length, the remaining basic vector is momentum. This approximation holds only so long as the system to which it is applied essentially does not change over the quantal fragment of length discarded -- i.e., it holds for the linear case. Conservation of momentum, then, is violated when sufficient nonlinearity in length is present.

(24) As a second special case, we may assume that we can evaluate a physical change in a spatial manner (without time). If so, when we discard time, the remaining basic vector is energy (has the units of energy or work). This approximation holds only so long as the system to which it is applied essentially does not change over the quantal fragment of time discarded -- i.e., it holds for the linear case. Conservation of energy, then, is violated when sufficient nonlinearity in time is present. Since a "virtual change" a priori is defined as a total nonlinearity in the observer's quantal time increment but not outside it, then virtual interactions can and do violate conservation of energy within that time increment, but not out of it -- so long as the time interval itself is considered linear. If the time interval is sufficiently nonlinear, then the virtual change may result in violation of the conservation of energy externally to the time increment. In that case, an "observable change" results .

(25) As a third special case, we may assume that we can evaluate the "instantaneous value" of a physical change at a static point in space. To do so, we must discard both time (to be instantaneous) and length (to be at a spatial point), and the remaining basic vector is force. This approximation holds only so long as the system to which it is applied essentially does not change over the quantal fragment of time or the quantal fragment of length discarded. Conservation of force, then, is violated when sufficient nonlinearity in time or length is present.

(26) A new conservation of energy law is required, one which unites the present conservation of energy law with an altered form of the conservation of charge law. Briefly, the total equivalent of mass,

observable energy, and massless charge (anenergy) is conserved.

(27) All AL and At fragments are produced and destroyed one at a time, in the action fissioning of a single quantum of action (detection process). Each ∇L and ∇t is discretized but not quantized. Since quanta do not superpose, the "external universe" is continually created and destroyed in the detector's mass system, one quantum at a time, at a very high rate. This interpretation gives physical meaning to the creation and annihilation operators of quantum mechanics.

(28) Since the detecting mass system is itself continually created and destroyed one quantum at a time, ultimately all is mind changes, and only mind changes. The observer's life, mind, and being transcend all materialistic interpretations of reality -- as indeed does the very fact of the "existence" of a perceived external universe.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

-- Electrical Physics Presently Has a Mindset --

In examining the foundations of geometry, mechanics, and electromagnetics, it becomes strikingly clear that substantial -- even grave theoretical errors were made early on and perpetuated into the existing theory. These errors are now so firmly entrenched that they form a part of the "mindset" of almost all physicists, engineers, and scientists.

So ingrained are these errors and inconsistencies that the orthodox scientist/theoretician finds it almost impossible to break out of them.

The present mindset is analogous to the Newtonian mindset which so fiercely resisted the new ideas of relativity, shortly after the turn of the century. However, after a few scientists formulated the rules and theory of the "relativity mathematics game," a new generation of students, not yet so firmly engrained in the Newtonian mindset, could grasp the new relativity when their teachers expounded it.

In this short series of papers[1],[2] I will roughly outline where the founders of mechanics and electromagnetics went wrong, and indicate the way to correct the fundamental errors. In addition, I will briefly point out some of the implications, and speak of some direct experimental proof.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

NOTES AND REFERENCES

1. Bearden, T. E., "Comments on the New Tesla Electromagnetics: Part I: Discrepancies in the Present EM Theory;" "Part II: The Secret of Electrical Free Energy," Tesla Book Company, 1580 Magnolia Ave., Millbrae, CA 94030, 1982.

2. Bearden, T. E., "Solutions to Tesla's Secrets and the Soviet Tesla Weapons," Tesla Book Co., 1981. Also Ratzlaff , John T., "Reference Articles for Solutions to Tesla's Secrets," Tesla Book Co., 1982.

3. Note we are applying the rule, "A thing is that which it does, and it does that which it is." Actually this is one statement of a fourth fundamental law of logic not incorporated by Aristotle. See Bearden, "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," DTIS report, available through the National Technical Information System, Port Royal Road, Springfield, VA 22161.

4. Specifically, the resulting theory becomes a curtailed, special case of the much more fundamental electrodynamics and electromagnetics that actually exist .

5. We point out here that measuring a field of force existing in the electron gas in a probe of the measuring instrument is not at all the same thing as measuring a force in the vacuum, nor does it establish that a force exists in vacuum. Indeed, it is already well known that the FIELD concept itself will not withstand rigorous logical examination. For a discussion rather clearly showing the present difficulty in defining a field, see Robert Bruce Lindsay and Henry Margenau, Foundations of Physics, Dover Publications, New York, 1963, pp. 283-287. Note particularly on p. 283 that a "field of force" at any point is actually DEFINED only for the case when a unit mass is present at that point. It is then illogically ASSUMED that the force

continues to exist at the point in the ABSENCE of the mass, which of course need not follow at all. On p. 284, note the similar logical paradox connected with the idea of a scalar gravitational potential field. The potential (field) is only defined at a point when mass is present at that point, .and it is specifically defined as the potential energy per unit mass for a particle present at that point. IF THERE IS NO MASS PRESENT, NEITHER A FORCE VECTOR FIELD NOR A SCALAR POTENTIAL FIELD IS DEFINED THERE. ASSIGNMENT OF THESE FIELDS TO THE POINT IN THE ABSENCE OF THE MASS IS AN ASSUMPTION, NOT AT ALL A DEFINITION. SINCE A TRUE DEFINITION IS AN IDENTITY, THEN THE ENTITY IDENTIFIED (DEFINED) TO INCLUDE THE PRESENCE OF MASS IS NOT IDENTICAL TO THE ENTITY RESULTING WHEN THAT MASS IS ABSENTED.

To see just how arbitrary and postulational are present. "definitions" of mass and force, see Lindsay and Margenau, op. cit., pp. 84-101. Also see Richard P. Feynman, Robert B. Leighton, and Matthew Sands, The Feynman Lectures on Physics, Addison-Wesley, New York, Vol.1,. 1963, Fourth Printing July 1966, p.2-4 for a definition of the electric field in the context of its POTENTIALITY for producing a force. Again, the force only exists when a particle of mass is present. From these examples, one can see the implication that A PHYSICAL FIELD IS SOMETHING SUCH THAT, WHEN A MASS IS INTRODUCED INTO IT, THE MASS EXHIBITS AN EFFECT. For a "force field," this is tantamount to stating that there exists some mechanism connected with a field which, in the presence of a mass causes a force to be exhibited. In that case the force is an EFFECT, not a cause, and there is a more fundamental mechanism that GENERATES FORCE ITSELF. See also field discussions in Feynman, Richard, The Character of Physical Law, M.I.T. Press, Cambridge, MA, March 1967, 2nd printing September 1967, passim.

6. While in Europe prior to 1881, Albert Abraham Michelson performed his first interferometer experiments to determine the velocity of the earth through the ether, obtaining essentially null results. At the Case School of Applied Science in Cleveland,

Ohio, he perfected his interferometer experiment from 1883 to 1887, assisted by a colleague, chemist Edward Williams Morley. By 1887 the results were ready and announced. Michelson himself thought his experiment had proven Stokes' theory of an ether dragged along by the earth in motion, and thus motionless with respect to the earth. This was at odds, however, with certain other experiments indicating a moving ether. The Michelson-Morley experiment was finally reconciled with these other experiments by Fitzgerald's suggestion in 1892 that the physical dimensions of material bodies are altered when they are in motion. In 1907 Michelson was awarded the Nobel prize, the first American to receive it in the sciences.

7. See Lindsay and Margenau, Foundations of Physics, 1963, pp. 324-326; D. C. Miller, "The Ether Drift Experiment and the Determination of the Absolute Motion of the Earth," Reviews of Modern Physics, Vol. 5, p. 203, 1933. Actually the experiments did not yield a conclusively null result, but rather showed large systematic trends. For a typical elimination of the systematic trends, see Handschy, M. A., "Re-examination of the 1887 Michelson-Morley experiment," American Journal of Physics, Vol. 50, No. 11, Nov. 1982, pp. 987-990. See Rho Sigma, Ether-Technology, CSA Printing & Bindery, Lakemont Georgia, 30552, 1977 for several enlightening points on the vacuum ether: See A. K. Lapkovskii, "Relativistic Kinematic Equations and the Theory of Continuous Media," Soviet Physics Journal, Vol. 21., No. 6, June 1978 for an abstract describing Soviet utilization of the concept of a small particle (called by Bearden a quiton, in Quiton/Perceptron Physics, DTIS, 1973) of the medium. See Belyaev, B. N. , "On Random Fluctuations of the Velocity of Light in Vacuum," Azvestiya Vysshikh Uchebnykh Zavedenii, Fizika, No. 11, Nov. 1980, pp. 37-42, translation by Plenum, for discussion of the proven variation of the speed of light in vacuum; the velocity of light in a vacuum on earth is measured to be higher than the velocity of light in the vacuum of deep space. See Graham, G. M. & Lahoz, D. G., "Observation of static electromagnetic angular momentum in vacuo," Nature, Vol. 285, 15 May 1980, pp. 154-155 for the first direct observation of free electromagnetic angular

momentum in vacuum. See Davies, Paul, "Something for nothing," New Scientist, 27 May 1982, pp. 580-582 for a discussion showing that modern theories of the vacuum reveal that even empty space is seething with activity; an ether of sorts emerges from vacuum fluctuations due to quantum mechanics considerations. See Hooper, William J., "All-Electric Motional Electric Field Generator," U. S. Patent No. 3,610,971, October 5, 1971 for a generator which produces a gravitational or inertial field. Einstein suggested that vacuum, complete with electromagnetic and gravitational fields, be called the ether. Dirac certainly did not abandon an ether, for in 1954 he stated "The aetherless basis of physical theory may have reached the end of its capabilities and we see in the aether a new hope for the future." James Clerk Maxwell derived his famous equations based on an ether theory. Sir Arthur Eddington also believed firmly in an ether. Sir Oliver Lodge actually pointed out the dilemma which yields the approach in this paper: writing of the ether in his book, The Ether of Space, Harper & Bros., New York, 1909, he stated: "We have no means of getting hold of the ether mechanically; we cannot grip it or move it in the ordinary way: we can only get it electrically. We are straining the ether when we charge a body with electricity; it tries to recover, it has the power of recoil.. . . "But when electrical theory was being founded, scientists thought of space as something rather fixed, and FILLED WITH a thin material ether. They did not realize that space itself does not exist except after an observation; before the observation, spacetime exists -- indefinite in both length and time. They did not know that electrostatic scalar potential in fact was spacetime, hence also the vacuum and the ether. In assuming that the charge of vacuum is zero and that charge and charged mass are identical, they hid the answer to the dilemma and placed the foundations of electromagnetics on its present unsound basis.

8. The field, of course, is indeed a highly useful concept and this author certainly does not recommend its abandonment. Instead, he recommends that it be placed on a sounder logical basis.

9. Specifically, they came to feel that the "electric

field" which was improperly defined -- was what was waving.

10. In fact Einstein once proposed that the vacuum, complete with its electromagnetic and gravitational fields, should be called the ether. His proposal was not adopted. (See Born, Max, Einstein's Theory of Relativity, Revised Edition, Dover Publications, New York, 1965, p. 224.)

11. Particularly from the work of Schrodinger, Born, Dirac and others.

12. For example, see Lindsay and Marge. nau, op. cit., pp. 287-288. A physical vector is thought to be defined by its magnitude, its direction in space, and its transformation characteristics. Actually that is a geometrical vector, not a physical vector. It does not tell us WHAT A VECTOR CONSISTS OF, but only tells us some of its important characteristics. Remember that a true definition must be an identity.

13. Here a reading of Lindsay and Margenau, op. cit., pp. 79-81 may prove enlightening. Also note that velocity, or L/T considered "at an instant" (stopped), represents an idea of "motionless motion" and is an application of the fourth law of logic. For a discussion of the fourth law of logic and its usage, see Bearden, Thomas E., "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," Specula, Journal of the A.A.M.S., P.O. Box 1182, Huntsville, AL 35807, combined Vol. 3, No. 4/Vol. 4, No. 3, Oct 1980 - Mar 1981, pp. 50-57 (also available from Defense Technical Information Service).

14. Note this is an identity of opposites, which explicitly violates the three Aristotlean laws of logic. See Bearden, "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," loc. cit., 1981.

15. Again note the fourth law of logic: zero motion (the absence of motion) being recognized as a special case of the presence of motion. Also, physical reality consists of internested levels, and any physical object has an internal substructure of nested levels of finer structure, extending down into the virtual (nonobserved) state. For a vector to model (apply to) a

physical object, it itself must be modeled in such fashion as to reflect this kind of substructure. Thus the use of geometrical vectors as models of physical objects in motion is presently flawed in a fundamental fashion.

16. The reader is most strongly urged to read Morris Kline, Mathematics: The Loss of Certainty, Oxford University Press, New York, 1980 as a prelude to understanding what mathematics is and is not, and what it does and does not.

17. Refer to Lindsay & Margenau, op. cit., pp. 79-81 to see how the ideas of motion and vector are inextricably entangled with the idea of a particle.

18. Call it uncertainty or call it constituency; a quantum change is composed of two canonical entities inextricably welded together into a single entity.

19. Time is an unavoidable, nonexclusive constituency of the welded quantum.

20. Simply from the definition of force as CONSISTING OF a time- and length-smeared mass motion change.

21. The force is an effect, not a cause. It IS the smeared charged particle. It is CAUSED by a more fundamental mechanism. It is the result of the combination of (1) a nonzero del phi, and (2) the presence of a spinning charged particle. IN A DEL PHI, THE SPINNING CHARGED PARTICLE ACCELERATES ITSELF! This is the fundamental secret of free energy that was suppressed, to bury the fundamental work of Nikola Tesla, shortly after Tesla was forced to abort his Wardenclyffe attempt to provide the world with free energy.

22. To quote: "The Hertz wave theory of wireless transmission may be kept up for a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history. " Nikola Tesla, "The True Wireless, " Electrical Experimenter, May 1919, p. 87.

23. De Beauregard, O. Costa, "Running backwards the Mermin device: Causality in EPR correlations," American Journal of Physics, Vol. 51, No. 6, June 1983, p. 515.

24. Note the Soviet scientist Kozyrev's experiments with time waves. See Kozyrev, N. A., "Possibility of Experimental Study of the Properties of Time," September 1967, pp. 1-49, in JPRS 45238, May 1968. Kozyrev reports real physical effects from the oscillation of time. Also, note that scalar potential energy of appreciable size with respect to a particle's rest energy can force the situation to be relativistic, even though the velocity of the particle with respect to the velocity of light is small. That is, electrostatic scalar potential alone can cause variation in the rate of flow of time and hence vary physical characteristics. See Bloch & Crater, "Lorentz-invariant potentials and the non-relativistic limit," American Journal of Physics, Vol. 49, No.1, 1981, pp. 67-75. By inference, oscillating the electrostatic scalar potential can produce time waves and lead to direct physical effects.

25. It is already shown in the literature that the electrostatic scalar potential (ESP) can affect spacetime (ST) in the same manner as velocity. Cf Bloch & Crater, op. cit., 1981. Now note that, to any quantal or macroscopic observer, the existence of the 4-space volume of ST implicit in $(\Delta t)(\Delta v)$, where v is volume, cannot be separated from the existence of the subquantal entities that exist therein. We therefore DEFINE the magnitude of the ESP as the summation of the absolute values of all the internal virtual vectors in the $(\Delta t)(\Delta v)$ quantum of ST, divided by the absolute value (magnitude) of $(\Delta t)(\Delta v)$. We take the view that no such thing as "unstressed" ST physically exists, and that "spacetime" and "stressed spacetime" are identical. Hence ESP and ST are one and the same thing. Note that this implies that the virtual density of ST is variable, and is nothing but the magnitude of the ESP. In EM theory, the assumption that the ESP of vacuum ($Ø_O$) is equal to zero is in serious error. In fact, $Ø_O$ IS "spacetime of the laboratory observer," in the new view.

26. For example, the definition of the electrostatic

potential (ESP) is usually taken as "the work which must be done against electric forces to bring a unit charge from a reference point to the point in question; the reference point is located at an infinite distance, or, for practical purposes, at the surface of the earth or some other large conductor." (McGraw-Hill Dictionary of. Scientific and Technical Terms, ed. Daniel N. Lapedes, second edition, 1978, p. 518.) Note that this is NOT a definition at all, for it is not an identity. Instead, it is the statement that, if an ESP exists at a point and a unit charged mass (assumed to be at a point) is brought in from infinity toward the ESP location point, the amount of work it is necessary to expend upon the mass of the particle is equal to the magnitude of the ESP. The ESP exists whether or not any work at all is expended, and whether or not a charged unit mass is brought in. To adequately define ESP, we must define its identity, or what it consists of, in the absence of mass, since we have conceived the ESP to exist at a vacuum point. Further, the definition usually taken is completely a 3-space definition. Instead, in our new view the ESP is to be taken at a point in n-space, where n is equal to or greater than 4.

27. See Bearden, Quiton/Perceptron Physics, 1973, available through the DTIS. See also Bearden, The Excalibur Briefing, Strawberry Hill Press, San Francisco, CA, 1980. Ultimately all physical phenomena are mindchanges in the minds of all the observers.

28. And then assumes this summation value is zero.

29. See note 25 above.

30. Bearden, The Excalibur Briefing, Strawberry Hill Press, San
Francisco, CA, 1980.

31. Cf Rauscher, E. A., "Electromagnetic and Non-Linear Phenomena in Complex Minkowski Spaces," Tecnic Research Laboratories, 64 Santa Margarita, San Leandro, CA 94579. Presented at the 1983 March Meeting of The American Physical Society in Los Angeles, CA 21-25 March, 1983. This is a truly remarkable paper of great significance. Rauscher, a world-class physicist, has presented a new theoretical

model for some rather extraordinary possible extensions of present electromagnetics.

32. Cf Muses, Charles, Introduction to Jerome Rothstein's Communication, Organization, and Science, The Falcon's Wing Press, Indian Hills, Colorado, 1958. The entire foreword by Muses is a remarkable document, which analyzes the structure of time itself. See also his profound summary paper, "Hypernumbers II" in the January 1978 issue of Applied Mathematics and Computation, published by Elsevier.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

ADDITIONAL NOTES AND REFERENCES

Although quotes and direct utilization of material from these references were not incorporated in this paper, the following references were also consulted. In addition,' several notes are added for further clarity.

33. Rupert Sheldrake, A New Science of Life: The Hypothesis of Formative Causation, J. P. Tarcher, Inc., Los Angeles, CA, 1981. Ø, the electrostatic scalar potential field, in my opinion is actually the morphogenetic field that Sheldrake proposes.

34. Briefly, by a "particle" we mean an entity so constructed that, if any part of it changes all of it changes. From the viewpoint of the particle, this implies that to change is to detect, and to detect is to change. Also, internal and external become synonymous, in the "detected" sense. The idea of a "fundamental particle" in physics actually invokes the fourth law of logic implicitly.

35. Only if a thing dimensionally contains time, can it "occupy time." This point is so obvious that one wonders how so many of the scientists and mathematicians seem to have missed it. By this criterion, e.g., mass does not exist in time, a priori. To "observe" or detect, in fact, means to stop time, thus collapsing the wave function. However, it reduces the observable or detectable to a spatial quantity, not a spatiotemporal quantity. In other words, the ordinary scientific method destroys a part of reality in each detection or measurement, yielding only a partial truth. not fundamental truth.

36. Note that electrostatic scalar potential is actually infinite-dimensional and hyperspatial. The coverage of

this paper is still only a special case. By Tesla technology, it is possible to do direct engineering in hyperspace -- beyond our present space and time, with all that that statement implies.

37. Bob Sloan, "Nikola Tesla: The Greatest Inventor of all Time?". IEEE Antennas and Propagation Society Newsletter, June 1983, pp. 9-11. A very succinct summary of the importance Tesla played in ushering in the modern age.

38. Gerald E. Brown and Mannque Rho, "The structure of the nucleon," Physics Today, Vol. 36, No.2, February 1983, pp. 24-32. Recommended as a summary of the new thinking as to the structure of the nucleon: a bag containing three quarks, surrounded by a cloud of mesons which squeeze the bag.

39. John J. O'Neill. Prodigal Genius: The Life of Nikola Tesla. Angriff Press, P.O. Box 2726, Hollywood, CA 90028, new printing 1981.

40. Margaret Cheney, Tesla: Man Out of Time, Prentice-Hall, Englewood Cliffs, NJ, 1981.

41. John T. Ratzlaff and Leland I. Anderson. Dr. Nikola Tesla
Bibliography. Palo Alto. CA, 1979. Indispensable.

42. Dr. Nikola Tesla: Selected Patent Wrappers, compiled by John T. Ratzlaff, multiple volumes. 1980. Available from The Tesla Book Company, 1580 Magnolia, Millbrae, CA 94030. Tesla's correspondence with the U.S. Patent Office, when patiently trying to obtain patents. He spent a great deal of time trying to convince the Patent Office that his inventions would indeed work. Some of them required 12 years to obtain, and then were "watered down" in the process.

43. Thomas Commerford Martin, The Inventions, Researches and Writings of Nikola Tesla, Originally published in 1894 by The Electrical Engineer, New York; republished in 1977 by Omni Publications, Hawthorne, CA 90250.

44. Ernest Nagel and James R. Newman, Godel's

Proof, New York University Press, 1958.

45. Yakov P. Terletskii, Paradoxes.in the Theory of Relativity, With a Foreword by Banesh Hoffman, translated from the Russian, Plenum Press, New York, 1968. Of particular interest is the discussion on particles with imaginary masses, moving faster than the speed of light, contained in pp. 104-107. Such particles can in principle be experimentally detected. In fact, it would appear that the well known exchange of virtual particles between two other particles, such that each turns into the other, is such a case. (Note that protons and neutrons in the nuclei of atoms do precisely this.)

46. Robert M. Besancon, Ed., The Encyclopedia of Physics, Second Edition, Van Nostran Reinhold, New York, 1974. Particularly see the discussion on the electron, pp. 272-274. Note this discussion predates Stanford University's experiments yielding fractional charge, though it does point out that several physicists had also reported measuring fractional charges on the electron. See also the discussions of ionization, Michelson-Morley experiment, the photon, and propagation of electromagnetic waves.

47. Robert Eisberg and Robert Resnick, Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles, John Wiley & Sons, New York, 1974.

48. R. K. Bullough and P. J. Caudrey, eds., Solitons, Springer-Verlag, New York, 1980.

49. James Dale Barry, Ball Lightning and Bead Lightning, Plenum Press, New York, 1980. Note particularly p.196, for a short discussion on flashless discharges. An extensive bibliography is also included.

50. Harley D. Rutledge, Project Identification, The First Scientific Field Study of UFO Phenomena, Prentice-Hall Inc., Englewood Cliffs, NJ, 1981.

51. John J. Reitz, Frederick J. Milford, and Robert W. Christy, Foundations of Electromagnetic Theory, Third Edition, Addison-Wesley, Reading, MA, 1979.

52. The entire series of handbooks by William Corliss, dealing with anomalies and unusual natural phenomena

of all kinds. Corliss is a national treasure, and his handbooks are absolutely indispensable.
See particularly his Handbook of Unusual Natural Phenomena, The Sourcebook Project, Box 107, Glen Arm, MD 21057, 1977 and his Lightning, Auroras, Nocturnal Lights, and Related Luminous Phenomena, 1982.

53. Bernard d'Espagnat, Conceptual Foundations of Quantum Mechanics, W. A. Benjamin, Menlo Park, CA, 1971.

54. D.W.G. Ballentyne and D.R. Lovett, A Dictionary of Named Effects and Laws in Chemistry, Physics and Mathematics, Fourth Edition, Chapman and Hall, New York, 1980. Check this neat little book to discover some very odd effects in materials.

55. David Bohm, The Special Theory of Relativity, W. A. Benjamin, New York, 1965.

56. Albert Einstein, Relativity: The Special and the General Theory, Crown Publishers, New York, 1961. See particularly the discussion of relativity and the problem of space, in Appendix V.

57. Edwin F. Taylor and John Archibald Wheeler, Spacetime Physics, W. H. Freeman and Co., San Francisco, 1966. Note particularly the discussion on observers and frames in the first two dozen pages. On p.19, note that the notion of a frame requires an infinite observer distributed through each and every "point" that was clock-synchronized in a frame. Since all observers are localized, a better idea is to realize that all the "external" modeling just represents the relative changes inside the physical detection system of the observer himself/herself. All detection/observation is totally inside the physical observer.

58. Paul Edwards, Ed. in Chief, The Encyclopedia of Philosophy, Vols. 1-8, Macmillan Publishing Co., New York, 1967.

59. Michael A. Persinger and Gyslaine F. Lafreniere, Space-Time Transients and Unusual Events, Nelson-Hall, Chicago, IL, 1977.

60. John David Jackson, Classical Electrodynamics, Second Edition, John Wiley & Sons, New York, 1975.

61. James Clerk Maxwell, A Treatise On Electricity & Magnetism, Vols. 1 & 2, Third Edition, Dover Publications, New York, 1954. An unabridged, slightly altered, republication of the third edition, originally published by the Clarendon Press in 1891.

62. Jack S. Greenberg and Walter Greiner, "Search for the sparking of the vacuum," Physics Today, August 1982, pp. 24-32. A beautiful summary article on the present concept of the vacuum, from the standpoint of quantum mechanics and quantum field theory. I specifically urge anyone interested in tapping the vacuum energy to read this article.

63. Max Jammer, Concepts of Mass, Harvard University Press, Cambridge, MA, 1961. Recipient of the monograph prize of the American Academy of Arts and Sciences for the year 1960 in the field of physical and biological sciences. This book, strange as it may seem, was the first monograph to subject the notion of mass to an integrated and coherent historical investigation, something which had never before been done. The reading of this book is an absolute must for anyone seriously concerned about whether or not science speaks fundamental truth, or to what degree it does so. Most scientists assume foundations concepts such as "field," "force," "mass," "time," etc. are well-defined and well-understood in physics, since they are ubiquitous and so familiar. In fact, none of the absolute fundamentals in the foundations of physics is unambiguously understood!

The last paragraph by Jammer is illuminating: "The modern physicist may rightfully be proud of his spectacular achievements in science and technology. However, he should always be aware that the foundations of his imposing edifice, the basic notions of his discipline, such as the concept of mass, are entangled with serious uncertainties and perplexing difficulties that have as yet not been resolved."

64. G. Burniston Brown, "Gravitational and inertial mass," American Journal of Physics, Vol. 28, p. 475, 1960: "...one of the most astonishing features of the

history of physics is the confusion which surrounds the definition of the key term in dynamics, mass."

65. Whittaker, Sir Edmund, A History of the Theories of Aether and Electricity, Vol. 1: The Classical Theories, and Vol. 2: The Modern Theories 1900-1926, Harper Torchbooks, Harper & Brothers, New York, 1960.

66. Weyl, Hermann, Space -- Time -- Matter, Fourth Edition, translated from the German by Henry L. Brose, Dover Publications, New York, 1922.

67. Jammer, Max, Concepts of Space: The History of Theories of space in Physics, Second Edition, Foreword by Albert Einstein, Harvard Univ. Press, Cambridge, MA, 1969.

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- It Started With Geometry and Grew --

At the very beginning of what we call the "scientific period," mathematics was both king and queen, and Euclidean geometry was its handmaiden. So we ask, "What precisely is geometry?" Here we are not interested in a "textbook" answer, but in an answer indicating what geometry really does.[3] In other words, with what does geometry concern itself, and what is the fundamental nature of those things with which it concerns itself?

Briefly, geometry -- at its foundation -- is totally spatial. It is fitted to, and expressed in terms of, the TOTAL ABSENCE OF MASS. Thus the geometer deals in abstract, massless entities called "points," "lines," "planes" etc. When the geometer speaks of "motion," he speaks of a time-smeared, length-smeared point. Geometry at heart is massless, and a "geometer's vector" is a highly specific type of "system." In fact, it represents the "time-smearing" and "length-smearing" of a point. A priori, the fundamental concept of the geometrical vector has taken a "spatial" entity and introduced a hidden involvement with "time."

Modern mathematics and physics have followed an intertwined development for several hundred years. And both sprang as offshoots of the original work of the geometers. Let us briefly sketch the overall path of interest taken by these two developing disciplines.

With the advent of Descartes's fundamental work, algebra was combined with geometry to yield analytic geometry, a new and powerful mathematical tool. With the invention of calculus by Leibniz and Newton, both mathematics and physics received a giant impetus. Differential geometry and vector mathematics arose in full splendor and, in physics, mechanics leaped to the forefront with Newton's profound work.

But the mechanics made a most fundamental error when they simply applied the geometer's vector to a mass, to produce -- so they thought -- a mass vector. That which rigorously applies only to the absence of mass cannot be so lightly applied to the presence of mass without the risk of serious limitations in the resulting theory. The precise difference between a geometer's massless vector and a mechanic's mass-vector is one of the issues to be developed in this thesis.

As rapid development continued in mechanics and mathematics, certain physicists were involved in intense experimental work on charged matter, becoming the first electricians. Both the preceding mathematical ideas and constructs as well as the preceding (partially erroneous) mechanics constructs and ideas were applied by the electricians, struggling with their pith balls, cat fur, and glass rods to understand, quantify, and model electrical forces and the phenomena of charged matter. In other words, the electricians strove to formulate the physics and dynamics of charged matter and its interactions by simply "adding to" the work of the geometers and mechanics. Here again, a fundamental logical error was made. That (geometry) which a priori applies only to the absence of mass, and that (mechanics) which a priori applies only to the absence of charge, cannot be lightly applied to the presence of charged mass (both mass and charge)[4] without risking the incorporation of grave limitations in the resulting theory.

After the profound work of Maxwell, the idea of FIELDS OF FORCE became more prominent, until the field concept ruled the day[5]. The electricians continued, pushing the idea of fields into space and vacuum itself, along the way inventing the idea of "charge effects" existing even in the massless vacuum, with concomitant fields. Meanwhile, they had thoroughly confused chargeless point-smeared, chargeless mass-smeared, length-smeared and time-smeared vectors.

After a set of fundamental experiments designed to detect motion of the material ether yielded essentially null results[6], Michelson and Morley were regarded as having completely disposed of the ether -- even though the experiments only disposed of material ethers, and not Lorentz-invariant non-material ethers[7]. Maxwell's equations and the field concept were elevated to profound importance.[8] Then, after Einstein's fundamental relativity work shortly after the turn of the century, the ether concept faded away and the field concept reigned supreme. Indeed, in their enthusiasm the interpreters of relativity went so far as to affirm that one can have a wave without any medium; that is. that something can be moving (waving) without anything there to move![9] And with great glee they pronounced the final end to the idea of "ether" as a medium, even though Einstein himself never did any such thing.[10] With the advent of Einstein's General Theory of Relativity, even matter came to be regarded as just a special "kink" or curvature in spacetime or "vacuum nothing."

Quantum mechanics arose and even certainty and determination fell. Chaos, probability, and randomness now assumed the ruling position. Probability waves (and probability fields) arose,[11] as did quantum fields of various kinds. The intermingling of these concepts with the concepts of electrodynamics pushed the idea of the field even farther into esoteric realms.

The point is, each of these developing disciplines incorporated and built on the foregoing disciplines. From the beginning of geometry, there was no rigorous definition of a vector, and there is none today.[12] From the beginning of mechanics, in their foundations the theorists made grave logical errors by incorporating the geometer's vector; errors so great that today mechanics and electromagnetics are severely flawed, as is everything that came after them and built upon their illogical foundations.

Next Page

Help support the work

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- Points and Motion --

It is my purpose in this paper to expose in a very simple fashion the most basic errors that were made. One basic error involves the idea of motion itself.[13]

In formulating concepts of motion, the geometers used a "point in motion" to determine or specify, for example, velocity. Now a "point" is a static concept a priori. To determine (or even to think and perceive) motion, one must determine that it occupies two different points (positions or locations) at two different times, yet consider both points at the same time. Indeed, that is precisely what the arrow means that is used to represent a vector. A "point in motion" therefore represents a contradiction of opposites. That is, it represents the idea that "that which is motionless has motion.".[14] Even with this, there is a difference in a spatial point and a spatiotemporal (spacetime) point. To exist at all, a spatial point must be moving in time; in other words, it is a spatiotemporal line, even if it is a static spatial point.

Vector analysis was constructed in the abstract -- again, a massless point in motion possessed or constituted a velocity vector, etc. In massless (and timeless) space, FIELDS were defined: "scalar" fields constituted the assignment of a simple motionless number (magnitude) to each spatial point, while "vector" fields constituted the assignment of a "simple vector" (magnitude and velocity) to each spatial point. But the MATHEMATICAL vector system consisted of massless (point) motional relationships, recognizing zero motion as a special case of motion.[15]

Of course mathematics development was also always intertwined with practical problems. With the sustained application of mathematics to gross physical material problems, mechanics slowly arose.

These developments required decades and even centuries to occur completely. All along the way, innovations and changes -- and additions to the mathematical formulism were being derived and taught to students as the "natural" system of reality. A permanent mindset was being forged.

Indeed, mathematics was regarded as THE single human expression of fundamental truth. Not until Godel's work in the twentieth century did it become evident that MATHEMATICS IS SIMPLY A GAME PLAYED ACCORDING TO ASSIGNED RULES, AND THERE IS NO ULTIMATE TRUTH IN MATHEMATICS ALONE.[16] It is a most useful game, of course, since it is the game fitted to perception. Thus it applies, essentially, to whatever can be perceived. But to be applied to physical systems, it must be changed, altered, updated, and fitted as the perceiving/detecting instruments become ever more subtle.

Next Page

Help support the work

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

-- Four Types of Vectors Actually Emerged --

As the physical sciences slowly developed and incorporated abstract geometry and mathematics, in actuality four major types of vectors and two major states of observation evolved, although this fact did not become apparent to the scientists. Specifically, the mathematicians and scientists failed to recognize the differences in the four types of vectors, hopelessly intermingling them and confusing them as a single class of vector. Further, they did not appreciate that a fundamental vector conceptually is a UNITARY SYSTEM, and the system represented by one of these four types of vectors utilizes and is comprised of different components, "welded together with no seam in the middle."

Conceptually (and from a systems viewpoint), the four types of vectors are (1) the chargeless, massless spatial system vector (geometer's vector), (2) the uncharged mass system vector (mechanic's vector), (3) the charged mass system vector (electrician's vector), and (4) the charged space system vector (advanced electrician's vector). These four fundamentally different vectors are shown in Figures 1, 2, 4, and 9 in a simplified manner.

The major problem was that, beginning with the geometer's vector, these four major types of vectors were not treated as systems. Instead, their "vector" aspects were hopelessly confused and intermeshed, and no distinctions were made between them. And in the foundations of the mathematical constructs, time-smearing was not recognized at all.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

-- Quantum Mechanics Compounds the Problem --

In addition, the two presently recognized observation states -- observable and nonobservable (virtual) -- were of course unknown to the early geometers and electricians, and these ideas were not incorporated directly into the theoretical foundation.

From particle physics and quantum mechanics, we now understand that physical reality is structured of an observable state, underlaid with an infinite number of ever finer, successive levels of virtual (unobservable) states. At least reality is most accurately modeled in that fashion, according to particle physics today.

It is also well known, for example, that at the most fundamental level, one cannot actually separate nonmotion from motion (which implies, for example, that one cannot separate mass and velocity). In other words, a "mass in motion" idea is actually incorrect, at the most basic level. What actually exists is a sort of "smeared mass". That is, "mass-motion" is fundamentally what exists, not mass IN motion .

Actually, all that the Heisenberg Uncertainty Principle implies is this fact: If one examines the concept of "static (non-smeared) thing in non-static (smeared) motion", in ever finer detail, one reaches a degree of fineness where the "smearing" is paramount and one cannot have an un-smeared or "separate static thing" to be in motion. Instead, one only has the smeared, 4-dimensional spacetime entity, without 3-dimensional spatial separations.

This means, for example, that at the most basic level, it is actually incorrect to represent a momentum with a little static particle of mass connected to a spatial velocity vector. It is incorrect to think of the system as comprised of TWO SEPARATE ENTITIES, (1) a mass, and (2) a massless spatial system velocity vector (a geometer's vector).

We mention in passing that, presently, we understand that every particle is continually accelerating. First, the particle has spin, which involves rotation, which means that every "part" of the periphery of the particle is accelerated toward the center axis of spin. Second, every particle is continuously "fluctuating," and these fluctuations are accelerations. Further, we must consider any change such as an acceleration -- as existing in a small time increment, and occurring in a small length increment. Thus mass particles actually exist as (mass x acceleration x time x length). This of course has the dimensions of ACTION or angular momentum. The "real" world of physical matter, then, is composed of building blocks of action, called "quanta." Any other physical "quantity" must be obtained by fissioning (differentiating) the action quantum. For this reason, quantum mechanics presently must postulate that to every observable there corresponds an operator.

Next Page

-- To Summarize Briefly --

Let us now summarize these concepts and further examine their impact.

In physics, there must actually exist four major KINDS of vectors, rather than just one as prescribed by present interpretation. These vectors are "built" starting with four major kinds of particles.[17] These are: (1) the spatial point, (2) an uncharged mass particle, (3) a charged mass particle, and (4) a charged spatial point. Further, each of the four vectors at its most fundamental level (that is, at the quantum level)[18] is actually an inseparable, unitary SYSTEM welded into a single undivided entity containing time and existing nonexclusively in time.[19] When we look at or represent the so-called "parts" of the system, we are looking at them before they are welded together into the physical vector. That is, whenever we speak of "parts, " we imply that a "cutting" or "differentiating" action has been implied to separate the system into such "parts."

From such considerations, four different kinds of system vectors result.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

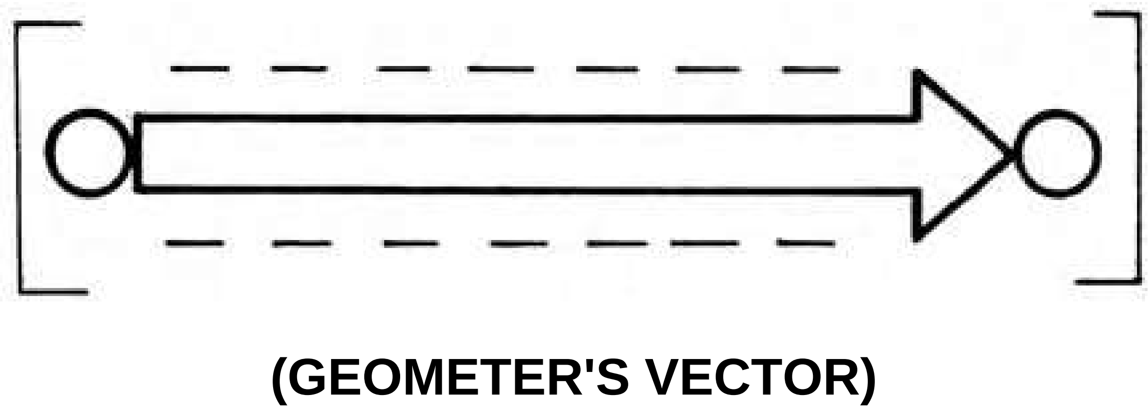

Figure 1. Uncharged spatial (massless) system vector.

UNCHARGED SPATIAL (MASSLESS) SYSTEM VECTOR
-- (Figure 1) --

This is the geometer's abstract vector, consisting of a "point in motion." (Actually, it is a "smeared point," for example.) However, so ingrained is the concept of a "point in motion" velocity vector that we now consider it to be "natural" because of its total familiarity. But simply ask, "WHAT is in motion?"[17] and you immediately see the difficulty. To have a WHAT, one must "stop the action" (detect or measure), separating "static" from "non-static." Acceleration and other vectors, etc. have also been derived by the geometer and utilized in similar fashion. All are massless.

Next Page

TOWARD A NEW ELECTROMAGNETICS

PART III: CLARIFYING THE VECTOR CONCEPT

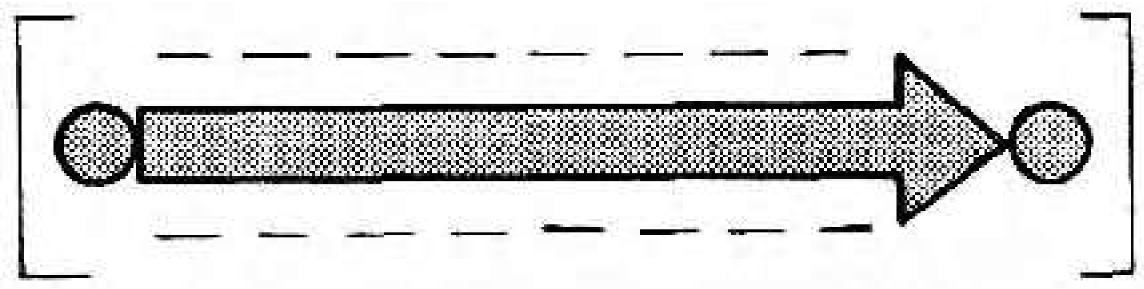

(MECHANIC'S VECTOR)

Figure 2. Chargeless mass system vector.

CHARGELESS MASS SYSTEM VECTOR

-- (Figure 2) --

This type of vector actually is the essential vector of mechanics, involving mass-motion (momentum), force (mass-motion change), etc. The fundamental difference between this type of vector and the geometer's vector is the presence of smeared mass existing in time (i.e., mass-time), welded together with a geometer's vector, but with no seam in the middle, into a new kind of "vector" AS A SYSTEM. The vector is the SYSTEM EXISTING IN TIME .

To illustrate:

"Momentum" is more properly referred to as "mass-motion" rather than "mass in motion." That is, at the fundamental level, the mass is NOT "separate" from its motion. It is NOT separated from the time in which the smearing occurs. The idea of momentum, however, is really to express the time-density of the mass-smeared-through-length. It is, in other words, the time rate of length-smearing of mass. Now in our minds we have conceived that

$$\bar{p} = m \odot \bar{v} \qquad (1)$$

where, by the symbol $\odot$ we mean "coupled to"; but we actually have

$$\bar{p} \equiv [\overline{m\bar{v}}] \qquad (2)$$

where we are not allowed to separate (even in thought) m and $\bar{v}$. Quantum mechanics agrees with this essentially, because $\bar{p}$ is a canonical variable linked to length, in any observable physical change. (The REASON p is canonically linked to ΔL is because $\bar{p}$ is the time-rate of length-smearing of m. If there is no length, there is no length-smearing of m to have a time rate OF in the first place.)

Note that, not only is

$$\bar{p} = [\overline{m\bar{v}}] \qquad (3)$$

but also

$$\bar{p} \equiv [\overline{m\bar{v}}] \qquad (4)$$

which is a much stronger and quite different statement. That is, $\bar{p}$ is IDENTICALLY $(\overline{mv})$, not just calculably EQUAL TO (m)($\bar{v}$). This means that $\bar{p}$ is a SYSTEM that is COMPRISED of mass-motion.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

-- What Force Is --

We now note that force, for example, is -- and may be defined as -- the time-rate of change of momentum, or

$$\bar{F} \equiv d/dt[\bar{p}] \qquad (5)$$

and this identity states that a force -- any force is COMPRISED OF time-changing "mass-motion." As such, the force vector is a mass-system vector, not just a massless spatial vector. Fundamentally, this mass-system vector is a totally different creature from a massless spatial vector. Our present manner of considering force as a geometer's vector "separately applied to" a mass particle is completely erroneous at the quantum level. Instead, fundamentally force is always a mechanic's vector. Force is an EFFECT, not a CAUSE.

And here mechanics made a most fundamental error, in not recognizing the difference between its kind of vector and that of the geometer.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

--Force and Hertzian Waves Cannot Exist in Vacuum --

Note that one cannot have an observable "force vector" existing' in vacuum a priori.

For example, we have the definition of force as

$$\bar{F} \equiv \frac{d\bar{p}}{dt} \qquad (6)$$

or

$$\bar{F} \equiv \overline{[m\bar{a} + \dot{m}\bar{v}]} \qquad (7)$$

and we see that, rigorously, a force vector CONSISTS OF (not, "is equal to") a time-changing mass-motion vector system. IF THERE IS NO OBSERVABLE ACCELERATING MASS PRESENT, THEN THERE CAN BE NO OBSERVABLE FORCE PRESENT. The mass can accelerate in time (increase or decrease of mass) or space (increase or decrease of velocity) or both.

Observable force CANNOT exist in vacuum (in the absence of mass), a priori.[20]

However, assume for a moment that one could have a massless force vector, as assumed in present electrical theory. Let this force vector appear at a point in the vacuum. Since the vacuum has zero observable mass, it would have zero inertial resistance to this hypothetical observable force -- hence the observable force would instantly produce an "infinite" acceleration of its point of application, vanishing with it into the infinite distance. Therefore our fictitious force would disappear the instant it appeared! In any case, it could not be retained at a point in the vacuum for any finite length of time, however small.

The direct implications are that (1) something other than an observable electrical force field exists in the vacuum, and (2) there must exist a more fundamental mechanism by which this "something else" generates or CREATES a force on/of a moving electrically charged mass. (Note again that, at the most basic level, any particle of mass is always quivering, in motion, and accelerating. from quantal considerations alone.)

Thus immediately we have discovered something unique about so-called "force-fields" in vacuum: for example, about gravitational field, electrical field, and magnetic field (and the strong force and the weak force as well). These fields do not exist at all as ordinary force vectors -- and real force fields -- in vacuum! $\bar{E}$ and $\bar{B}$ fields, e.g., are defined in terms of force per unit electrical charged mass and magnetically charged mass, respectively. In the absence of mass, they cannot exist.

And this in turn means that transverse $\bar{E}-\bar{H}$ field waves (Hertzian waves) cannot exist in a vacuum. Indeed, they appear on, and ARE CONSTITUTED of, the charged-mass-motion that changes, and they appear where such change occurs, as a result of an introduced mass.[21] But in the absence of the spinning charged particle of mass, they do not exist as force fields at all.

Hertzian waves exist in a transmitting whip antenna, for example, in the oscillating electron gas along its length. Something else entirely different exists in vacuum between the transmitting antenna and the receiving antenna. Then in the receiving antenna, Hertzian waves again exist in the oscillating electron gas along its length. (See Figure 3.)

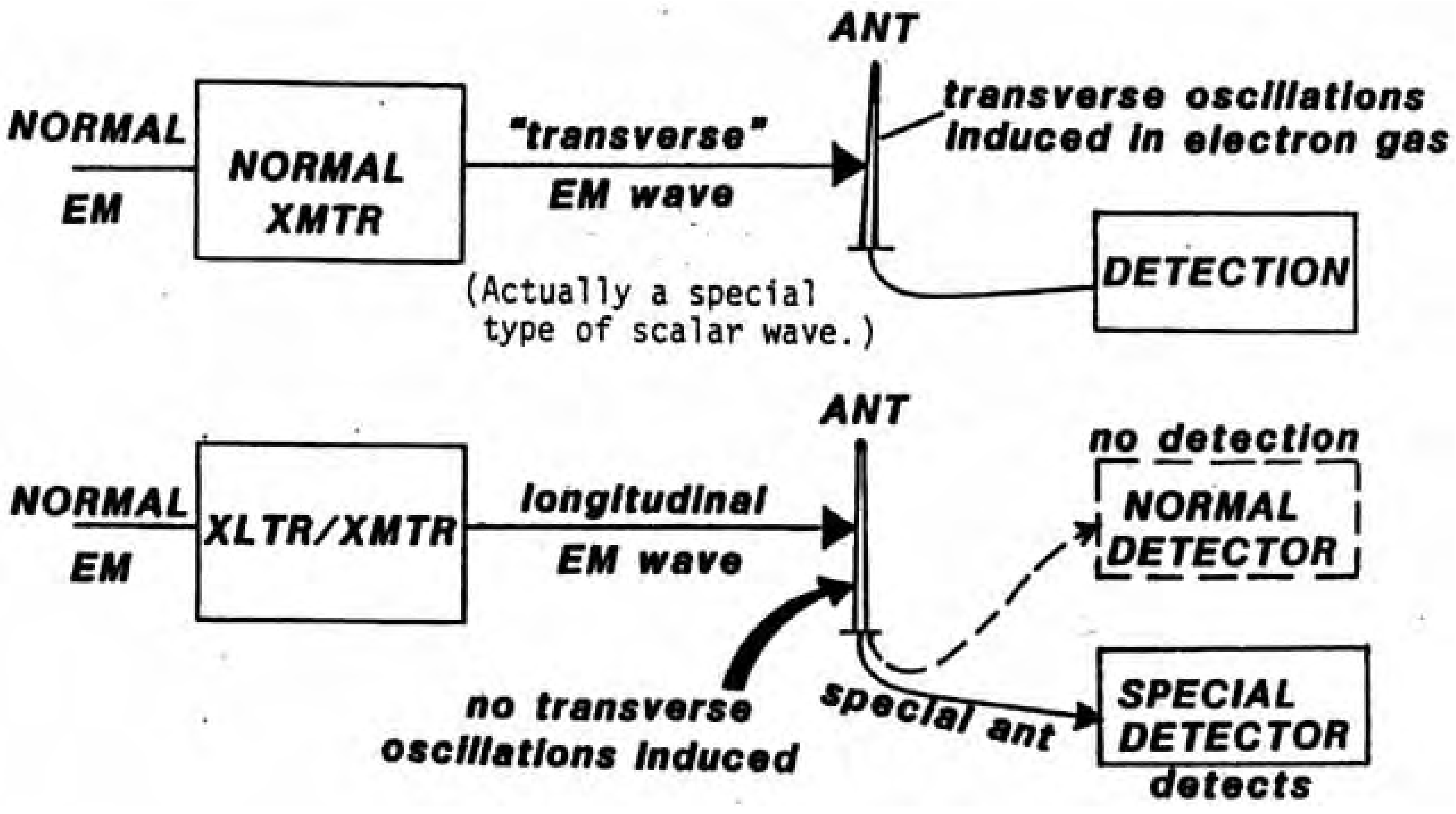
ANT
NORMAL
EM
NORMAL XMTR
"transverse"
EM wave
transverse oscillations induced in electron gas
DETECTION
(Actually a special type of scalar wave.)
ANT
NORMAL
EM
XLTR/XMTR
longitudinal
EM wave
no detection
NORMAL DETECTOR
no transverse oscillations induced
special ant
SPECIAL DETECTOR
detects

Figure 3. Detection of "transverse" and longitudinal waves

This is interesting, for Nikola Tesla stated several times that HERTZIAN WAVES CANNOT BE PRODUCED IN A VACUUM, NOR CAN THEY TRAVEL IN A VACUUM.[22]

Tesla was correct, as we are beginning to see.

We shall later return to show in what form so-called "force-fields" actually exist in vacuum.

For now, I point out that I am stating a fundamental change to all of physics, including both mechanics and electromagnetics.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support this work

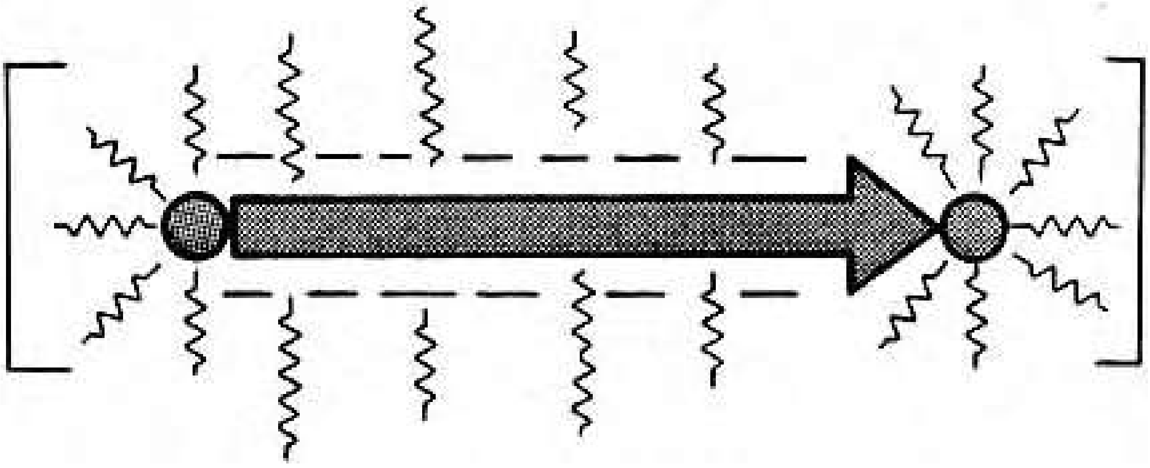

(ELECTRICIAN'S VECTOR)

Figure 4. Charged mass system vector.

-- CHARGED-MASS-SYSTEM VECTOR --
-- (Figure 4) --

The third type of vector we meet is the vector mass system where the mass is charged. First, we point out a serious error in present electromagnetic (EM) theory. That is, in present theory it is implicitly assumed that

$$\mathbf{q} \equiv \mathbf{q_m} \qquad (8)$$

In other words, "charge" and "charged mass" are erroneously assumed to be identically the same thing.

In the days when electricians were playing with pith balls and striving to uncover the secrets of electricity, they knew nothing at all about the virtual state, and consequently nothing about a "virtual particle flux" on a particle of mass causing (and comprising) the "charge" of that mass.

Today, of course, we know from particle physics and quantum mechanics that the "charge" on an observable particle of mass IS due to a flux of virtual (nonobservable) particles on and off the mass of the observable particle (see figure 5). A charged mass is thus presently known to be a SYSTEM: a massless charge flux, coupled to a bare particle (chargeless mass) constitutes a "charged particle."

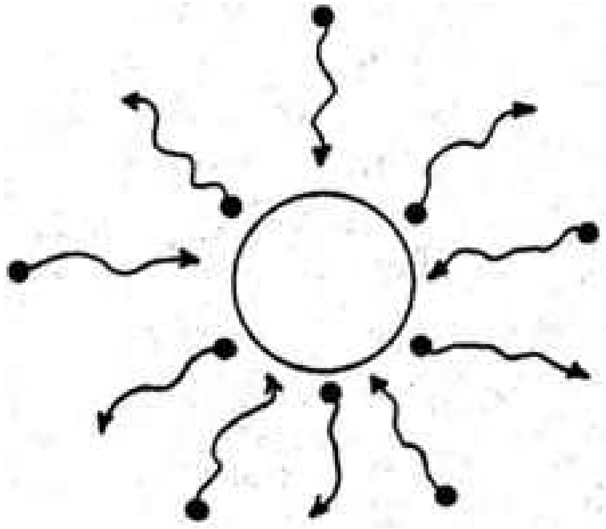

Figure 5. The "charge" on an electron mass consists of a flux of virtual particles on and off the mass.

Thus, actually the "charge" is the virtual (unobservable, or SPATIO-TEMPORAL) flux to and from the observable SPATIAL particle of mass. So, rigorously,

$$\mathbf{q \not\equiv q_m} \qquad (9)$$

But instead,

$$\mathbf{q \equiv [d/dm(q_m)]} \qquad (10)$$

and this is a definition and therefore an identity. This definition alone affects all present electromagnetics theory.

To illustrate: In founding electrical theory, early scientists dealt with forces generated by charged masses (for example, charged pith balls). They later extrapolated the experimental results they obtained (or thought they obtained) with the smallest charged mass, a charged particle. In Figure 6, I show the classic situation for derivation of the idea of E-field (except we have used an electron for our test charge, rather than a pith ball).

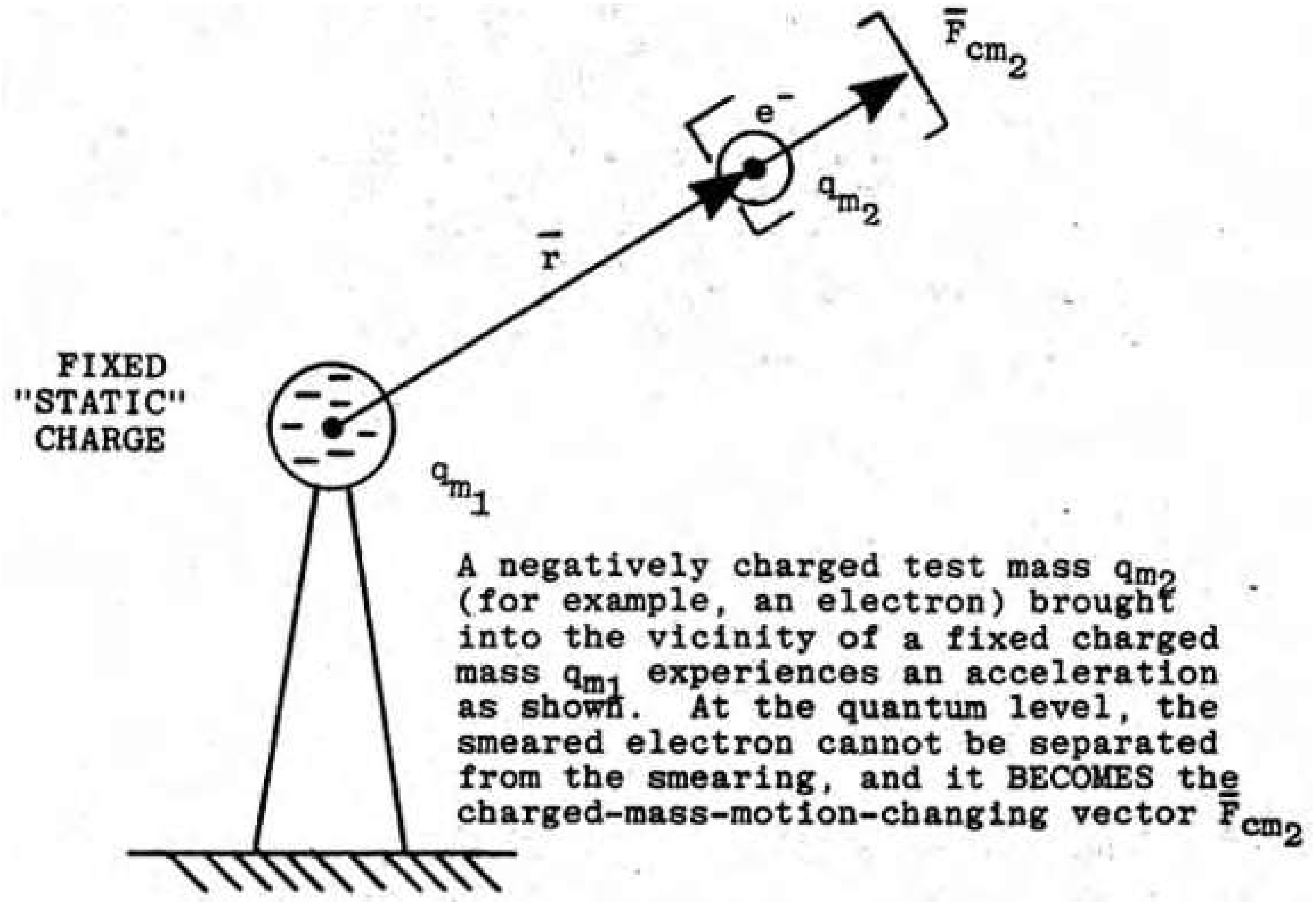

A negatively charged test mass q_{m_2} (for example, an electron) brought into the vicinity of a fixed charged mass q_{m_1} experiences an acceleration as shown. At the quantum level, the smeared electron cannot be separated from the smearing, and it BECOMES the charged-mass-motion-changing vector $\bar{F}_{cm_2}$

Figure 6. A test charge (charged mass) brought near a fixed charge (charged mass) experiences an acceleration.

Now note that what actually happens is that the unrestrained test charge becomes a CHARGED MASS SYSTEM VECTOR (a "smeared charged mass-motion changing"). The "test charge" BECOMES a charged mass force vector; it does not have a separate geometer's vector "appear on it." What actually happens is shown in Figure 7.

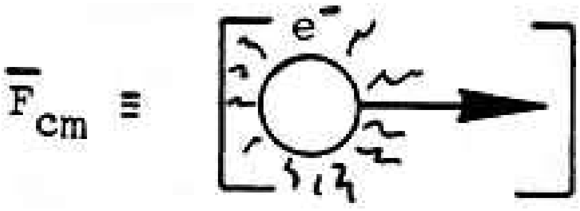

Figure 7. A charged-mass-system vector.

That is, in the simplest (nonrelativistic) case, for an electron what happens is

$$\bar{F}_{cm} \equiv [\overline{q_e \odot m_e \odot \bar{a}_s}] \qquad (11)$$

and this is a DEFINITION. That is, considered instantly, the electron exists as a charged-mass electrical force CONSISTING OF/COMPRISED OF a charge flux q_e canonically coupled to a mass, with that subsystem then canonically coupled to a spatial acceleration vector, ALL AS A SINGLE ENTITY, WITHOUT ANY "SEAMS" BETWEEN ITS "PARTS." The $\bar{F}_{cm}$ IS THE ELECTRON SYSTEM ITSELF; it is NOT a "spatial vector." Rigorously, it does not exist in the absence of the smeared electron mass, a priori.

Again, in assuming this force exists in the absence of the smeared mass of the moving particle, electromagnetics theory is in serious logical error.

Referring back to Figure 6, we see that, if we repeat the experiment many times and with the test charge in many locations, we have the situation shown in Figure 8.

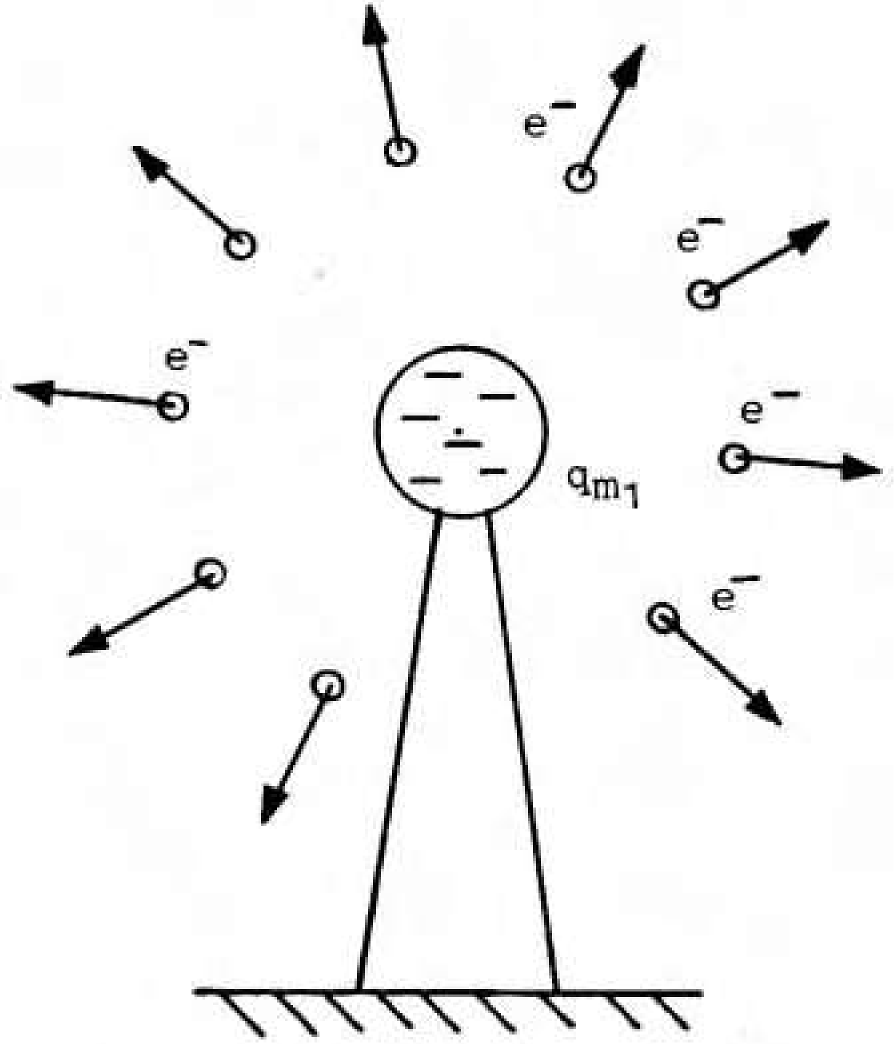

Figure 8. Repeating the "test charge" experiment.

It is found that, rigorously,

$$|\overline{F}_{cm}| \propto \left|\frac{q_{m_1} q_{m_2}}{r^2}\right| \qquad (12)$$

where $\overline{F}_{cm}$ is a charged mass system vector. Erroneously, this has been stated one way or another as

$$|\overline{F}| \propto \left|\frac{q_1 q_2}{r^2}\right| \qquad (13)$$

where $\overline{F}$ is assumed to be a spatial system vector. Further, this confusion has been carried over into the definition of the $\overline{E}$-field as:

$$\bar{E}_{cm} \equiv \frac{\bar{F}_{cm}}{m_{cm}} \qquad (14)$$

In this definition, $\bar{E}$ -- which is a charged mass system vector -- has been confused as a charged spatial system vector, where $\bar{F}$ is regarded simply a spatial system vector! Actually, the definition of the $\bar{E}$-field should be

$$\bar{E}_{cm} \equiv \frac{\bar{F}_{cm}}{m_{cm}} \qquad (15)$$

where $\bar{F}_{cm}$ is a charged mass system vector. Failure to properly define the $\bar{E}$-field has caused the conception of the $\bar{E}$-field to be falsely perpetuated as existing in vacuum.

The $\bar{E}$-field is TREATED this way in present EM theory. Hence present theory falsely assumes that the observable $\bar{E}$-field can exist in vacuum.

What actually exists in space, $\bar{E}$-field-wise, is a special kind of ordered virtual state pattern in a series of spinning "scalar" fields. This virtual state pattern or "shadow vector" field will be explained later.

Note again that one cannot have a "force vector" existing in vacuum - a priori.

However, assume for a moment that one could have a massless force vector, as presently assumed. Let this force vector appear at a point in the vacuum. Since the vacuum has zero observable mass, it would have zero inertial resistance to this hypothetical observable force hence the observable force would instantly produce an "infinite" acceleration of its point of application, vanishing with it into the distance. Therefore our fictitious force would disappear the instant it appeared! In any case, it could not be retained at a point in the vacuum for any finite length of time, however small.

The direct implications are that (1) something other than an observable electrical force field exists in the vacuum, and (2) there must exist a more fundamental mechanism by which this "something else" generates or CREATES a change on/of an accelerating electrically charged mass particle. (Note again that at the basic level, any particle of mass is ALWAYS quivering and accelerating, from quantal fluctuation considerations alone.) Causality has no arrow microscopically.[23]

Next Page

Help support this work

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

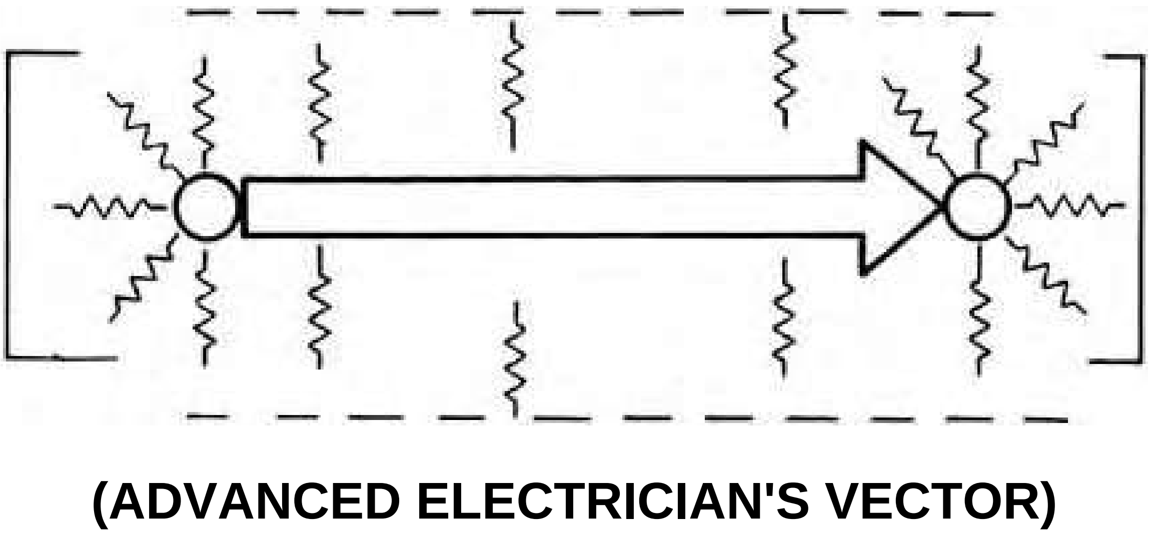

Figure 9. Charged spatial (massless) vector.

-- CHARGED SPATIAL (MASSLESS) VECTOR --

We recognize now that

$$q_c \not\equiv q_m \qquad (16)$$

and that q_c is simply the virtual-particle flux that constitutes charge -- and indeed constitutes vacuum itself!

We DEFINE vacuum, based on Figure 5, as shown in Figure 10.

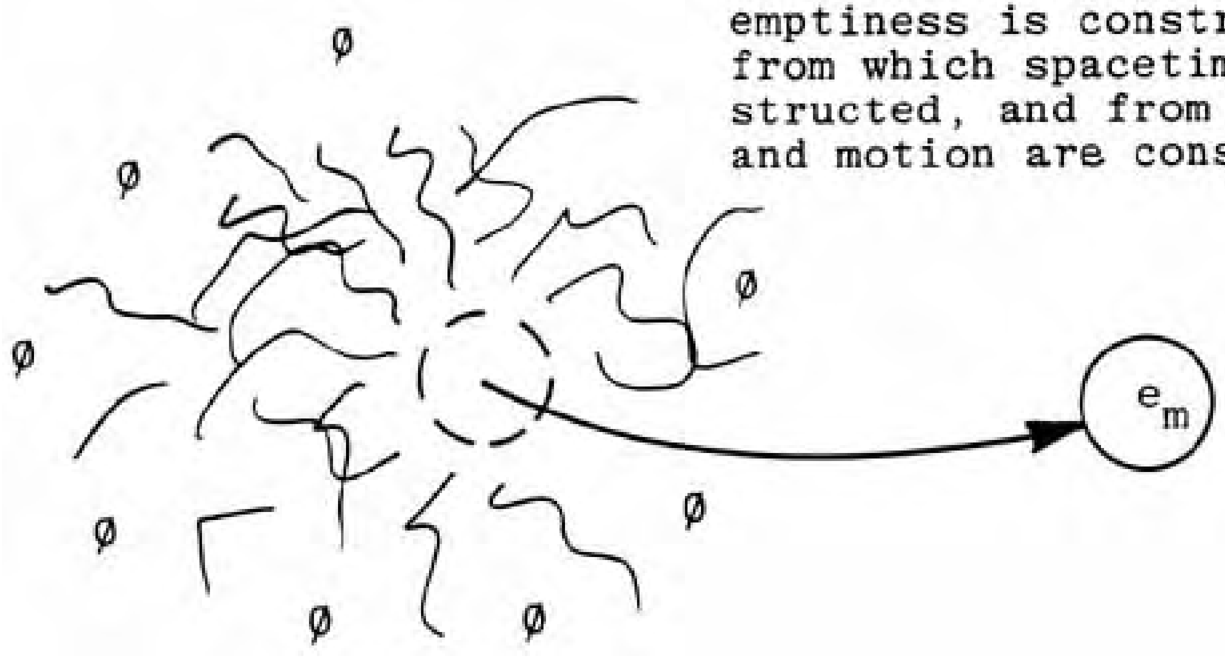

There is no "emptiness filled with charge." Rather, there is massless charge, from which emptiness is constructed, and from which spacetime is constructed, and from which mass and motion are constructed.

Figure 10. Removing the bare particle (mass) from
a charged particle leaves the charge.
The vacuum is DEFINED AS the charge.

That is, vacuum may be defined as pure massless charge flux. This flux IS identically "spacetime" as well. Vacuum is pure Ø-field (electrostatic scalar potential). Here again, in present theory it is assumed that

$$\text{Ø}_O \equiv 0 \tag{17}$$

which, by our new definition of vacuum, is quite false.

We now note that, if we insist on assigning a spatial vector to the vacuum, we have the situation shown in Figure 11.

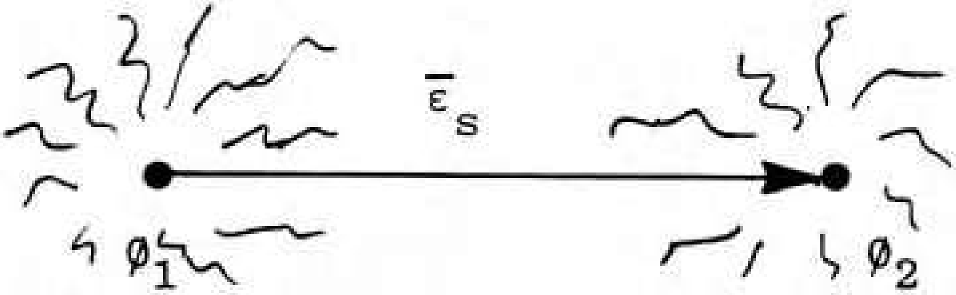

Figure 11. Assigning a spatial vector to the
charged vacuum

Next Page

Help support the work

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- THE SHADOW VECTOR --

Note that this spatial vector $\bar{\epsilon}_s$ represents $-\overline{\nabla\phi_s}$, that is,

$$\bar{\epsilon}_s = -\overline{\nabla\phi}_s \qquad (18)$$

but $\bar{\epsilon}_s$ cannot be a force (mass system) vector. It can only exist as an ordered pattern in the virtual flux between two separated points of the vacuum; that is, as an ordered pattern in the virtual state. Literally, $\bar{\epsilon}_s$ exists only as a tiny bit of order existing in great disorder.

In other words, the present EM theory is incorrect in stating that

$$\bar{E}_m = -\overline{\nabla\phi}_s \qquad (19)$$

in vacuum in the absence of an observable spinning charged particle, since

$$\bar{\epsilon}_s \not\equiv \bar{E}_m \qquad (20)$$

The actual existence of $\bar{\epsilon}_s$ may be visualized in terms of successive differentials of $\bar{E}_m$, broken into differentials $\bar{e}_m$ so small that, observably, each little differential's mass component m has become virtual, so that

$$\bar{e}_m \equiv \bar{e}_{vm} \qquad (21)$$

where subscript m stands for mass, subscript v for virtual, and observably

$$\bar{e}_{vm} \equiv \overline{dE}_m \qquad (22)$$

but, in the absence of a spinning charged particle,

$$\bar{E}_m \neq \int \overline{dE}_m \qquad (23)$$

since the $\bar{e}_{vm}$ components remain individually separated. That is, in macro-time a SHADOW force vector exists, made of microscopically ordered BUT UNJOINED (unintegrated) "virtual state" vector differentials of what would be an observable mass system force vector $\bar{E}_m$ if integrated.

Thus, the "$\bar{E}$-vector" $\bar{\epsilon}_s$ that exists in vacuum is a "shadow" vector as shown in figure 12.

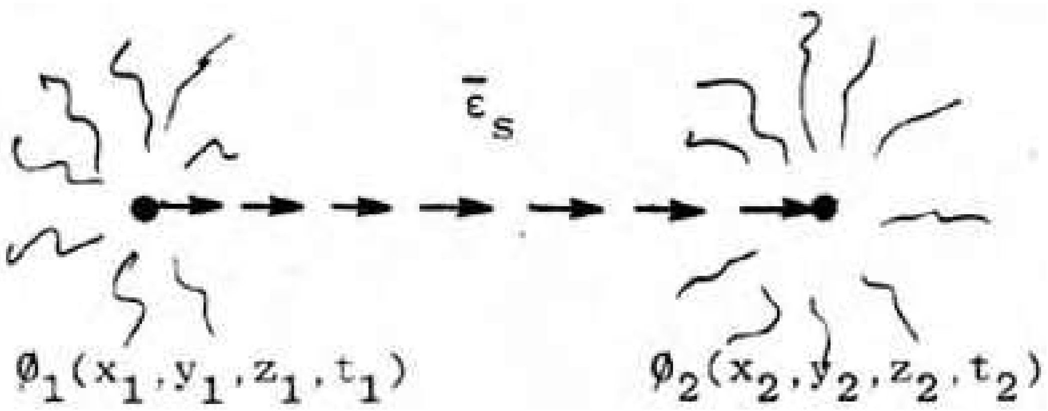

Figure 12. A "shadow vector" $\bar{\epsilon}_s$

We say that such a previous mass system vector, broken into ordered but unjoined virtual vectors by the absenting of all mass, is a SHADOW VECTOR, and we label it with a subscript vm, to represent "virtual mass" system. To the macro observer, this is the kind of "vector" that exists in vacuuo.

Note that, observably, the shadow vector merely represents a special ordering in $\nabla\phi$. It is NOT an OBSERVABLE (mass system) vector, but it IS an ordered series of consecutive virtual vectors.

With each virtual bit vector, a virtual time exists as well, and these "virtual time bits" are also ordered consecutively (in macrotime).

I point out that any observable vector must be finite, and so it must have a finite magnitude (finite length). In the simplest case, this length ΔL is related to a Δt by

$$\Delta L = c\Delta t \qquad (24)$$

What I am saying is that ANY observable spatial vector is actually a spatiotemporal vector, and the MAGNITUDE of any vector is related to TIME (to the existence of that vector in time) at the most fundamental level. Suffice it to say that, if the fundamental quantum level (Δt) aspect of a vector is interfered with, then the MAGNITUDE of the vector is interfered with.[24] That is, if we can make a time wave, we can change or affect ANY vector's magnitude, including the magnitude of mass system vectors and charged mass system vectors. Such a "time wave" can be made easily, and it has been.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- A Scalar is a Zero Vector --

Now let us look at the idea of a scalar.

A "scalar" may in a general sense be considered as the sum of the "absolute values" of the individual vector components of a system of vectors whose observable resultant is zero. That is, it represents the magnitude of the internal stress of a vector system, with the absence of a single observable directionality of the system. It also follows that every scalar is actually a stressed zero vector, and every zero vector is a scalar.

Thus we have four major types of scalars related to the four types of vectors:

(a) $$S_1 = |\overline{v}_s| \tag{25}$$

(b) $$S_2 = |\overline{v}_m| \tag{26}$$

(c) $$S_3 = |\overline{v}_{cm}| \tag{27}$$

(d) $$S_4 = |\overline{v}_{cs}| \tag{28}$$

where S stands for scalar, $\overline{V}$ for vector, and subscript s for spatial, m for mass, and c for charged.

For example, comparing equations (25) and (26), it can easily be seen that twice as many "point-motions" is not at all the same thing as twice as many "gram-mass-motions." The two resulting vector systems are quite different.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- Virtual and Observable Aspects --

We must also examine some aspects of "virtual" and "observable."

For example, we construct several spatial vector summations in Figure 13. The "resultants" of these spatial vectors are all equal. However, the actual sums, even though equal, are quite different, because their internal "stresses" (substructure forms) are quite different.

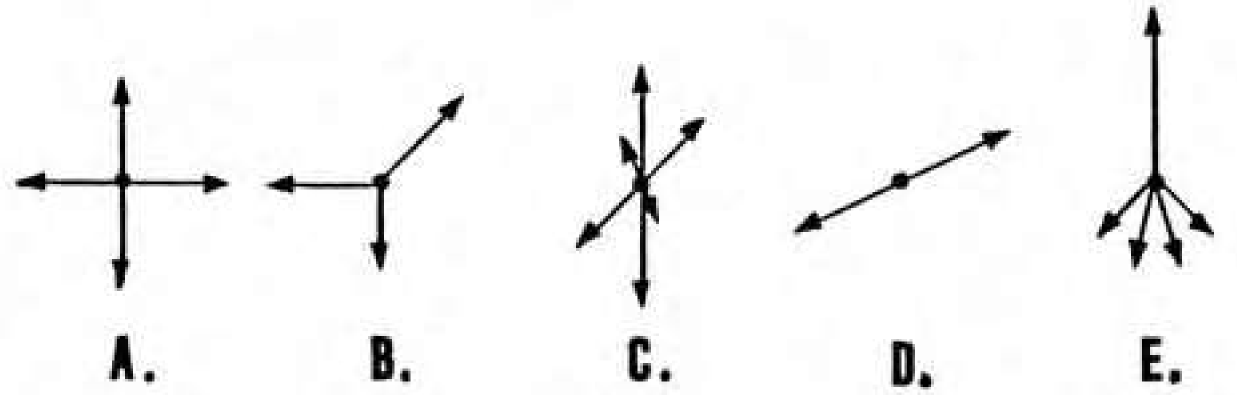

SUPERPOSITION DOES NOT ELIMINATE THE VIRTUAL SUBSTRUCTURE.

When the time aspects of the vector systems of Figure 13 are considered, one can easily understand the problem. That is, the resultant of each of these "systems" is zero, and so one can say that the vectorial "magnitude" of the system is zero since the magnitude of the resultant vector is zero. However, in each case the "action" represented by each vector element actually occurs in a finite tiny Δt. So: (1) The zero resultant must exist for a finite Δt, and (2) all the actions indicated by the system component vectors actually occur in that Δt. The absolute value of the "activity per unit time per unit volume" of such a zero-resultant system thus has physical meaning, and one may refer to this notion as the "stress" on spacetime[25], or the "electrostatic scalar potential" of the system. Note that this differs from the present definition of electrostatic scalar potential, which becomes just a special case of the more fundamental potential defined here.[26]

The derivatives of this spatiotemporal stress also have physical meaning. The time derivative is indicative of the stress on the flow of macroscopic time at a fixed spatial point, and the spatial derivative is indicative of the stress on space. Here one is confronted with the fact that what we call "space" and "time" are continually being created, directly in the physical observing/detecting apparatus itself.[27] That is, rigorously, "detected physical reality" exists totally in and of the mass-changes of the observer's mass or his detecting instruments. In the fundamental detection process itself, there is a flow of the rate of creation of spatial lengths and a flow of the rate of creation of time lengths. Indeed, to a linear observer the stress on the creation of the flow of time controls the flow of the creation of space, and the stress on the creation of the flow of space controls the flow of the creation of time. The change in the stress on 4-space (ordinary Minkowskian space-time) controls the "curvature of that spacetime" in the fifth dimension. The change in the stress on 5-space controls the "curvature of that 5-space spacetime" in 6-space, and so on. Development of these facets of the new concepts is beyond the scope of this paper.)

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- SCALARS AND VECTORS HAVE SUBSTRUCTURES --

As can now be seen, the sum of each structure in figure 13 is observably zero. Therefore we might define the sum as a "zero spatial vector." We note, however, that it actually exists for a time Δt and is thus a spatiotemporal entity, rigorously.

If we define the internal stress action A in a region $\Delta s^3\Delta t$ of spacetime as

$$A_{(s^3t)} \equiv \sum_{k=1}^{n} |\bar{i}_s| \qquad (29)$$

and the 4-space internal stress intensity or potential as

$$S \equiv \lim_{(\Delta s^3\Delta t)\to 0} \left[\frac{A(s^3t)}{\Delta s^3\Delta t} \right] \qquad (30)$$

where $\bar{i}_s$ is any internal vector in the substructure, Δs^3 is the spatial volume (about a point) containing vector $\bar{i}_s$, and Δt is the inseparable time during which these component actions occurred, then we see that, stress-wise, all the "zero-vectors" in figure 13 are quite different in their internal stresses, 4-space potentials, and internal substructures. For the five "zero sum" vectors, OBSERVABLY we have

$$\bar{0}_1 = \bar{0}_2 = \bar{0}_3 = \bar{0}_4 = \bar{0}_5 \qquad (31)$$

whether or not

$$\Delta s^3\Delta t_m = \Delta s^3\Delta t_n \ (m \neq n;\ 1 \leq m \leq 5;\ 1 \leq n \leq 5) \qquad (32)$$

But considering the substructures,

$$\bar{0}_m \neq \bar{0}_n \quad (1 \leq m \leq 5;\ 1 \leq n \leq 5;\ m \neq n) \qquad (33)$$

I now point out that a scalar can be regarded as a stressed zero-sum vector, where the magnitude S of the scalar represents the internal stress intensity caused by the substructure of the zero-vector.

Thus, generally,

$$S(s^n t) \equiv \lim_{\Delta s^n \Delta t \to 0} \left[\frac{\sum_{n=1}^{k} |\bar{i}_s|}{\Delta s^n \Delta t} \right] \qquad (34)$$

That is, in general any observable scalar has, consists of, and is comprised of a VIRTUAL (unobservable) substructure that is very real indeed. One must also consider the scalar as existing for some finite time Δt, (at least for the time of one quantum change), and the intensity of the virtual actions occurring in the spatiotemporal substructure of the scalar during that time Δt is proportional to the magnitude of the scalar.

Normally, the concept of a scalar -- as presently used -- makes no allowance for the scalar to exist in time, or for a virtual vector substructure, or for any patterning inside the substructure. This is equivalent to assuming that

$$A \equiv 0 \qquad (35)$$

and that all $\bar{i}_s$'s are evenly distributed. That is, from this new viewpoint, presently the mathematical theory assumes all scalars to have an equal density of virtual activity per spatiotemporal volume in its virtual substructure, and an isotropic virtual pattern distribution of an infinite number of equal virtual vectors in its 4-space substructure.[28]

In the new approach, neither of these two assumptions need hold -- though in special cases they can hold. Thus present orthodox theory is just a single special case of a more fundamental approach indicated here.

Note that, by directly affecting and changing the virtual substructures of scalars and vectors, we can directly perform virtual. state engineering, and this allows us to directly "engineer" the so-called "laws of nature" of the normal observable laboratory state and thus ENGINEER AND CHANGE PHYSICAL REALITY ITSELF.[29]

In the new approach, we can (observably) have

$$2 + 2 \neq 4 \qquad (36)$$

or

$$2 + 2 = 4 \qquad (37)$$

by the following means: In the first case (equation 36), we assume that the virtual substructures are patterned, and interact nonlinearly in such a way as to produce an extra observable. Thus we have a delta added to the normal observable scalar results of the interaction, as follows:

$$2_o + 2_o = 4_o + \Delta_{v\text{->}o} \qquad (38)$$

where subscript "o" means observable and "v" means virtual. Note that

$$\Delta_{v\text{->}o} \qquad (39)$$

indicates a delta due to virtual substructure interactions yielding an extra observable delta. This extra delta may be either scalar or vector in nature, depending on the circumstances and the particular interactions.

Note also that any vector or scalar must now be considered to HAVE, CONTAIN, and CONSIST OF an infinite substructure. And note that, similar to the scalar case, from the new viewpoint the present theory assumes each scalar (point) of the vector to have a structure similar to that of equation (34), except that now the scalars are ordered, with a linearly decreasing internal stress density per unit scalar along the line of the vector.

In the new approach, vector interaction (superposition, for example) can now violate present theory, if the two virtual substructures interact nonlinearly to produce a nonzero, observable delta. Observably (macroscopically) , this delta, again, may be either "scalar" or "vector."

This approach now becomes consistent with quantum mechanics at the foundation level.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- Substructures, Virtual Levels, and Hyperspaces --

In the new approach, our definitions and assumptions immediately drive us to a picture of an infinite set of nested levels of substructures in the virtual state. That is, anyone component (scalar or vector) in one level of virtual state has an infinite number of even finer virtual components, one level more subtle.

AND THAT IS WHAT PARTICLE PHYSICS AND QUANTUM MECHANICS ALREADY REVEAL ABOUT THE STRUCTURE OF PHYSICAL REALITY.

So these definitions and assumptions now provide the basis for a new vector mathematics that is in accord with, and fitted to, modern physical observations.

We have a picture such that any observable scalar or vector contains a virtual substructure (virtual level 1). Any scalar or vector in virtual level 1 also contains a finer virtual substructure, in virtual level 2. And so on ad infinitum.

Each succeedingly finer level of virtual state can be modeled as a hyperdimension (higher spatial dimension) as I pointed out in Appendix 1 to my book, The Excalibur Briefing.[30]

Thus this approach immediately ties into hyperdimensional or hyperspatial theory -- such as Elizabeth Rauscher's 8-dimensional theory[31] and C. Muses's hypernumber theory.[32]

The new definitions and assumptions are far richer than what is allowed by tensors, though there are many similarities. Muses's work, however, essentially can encompass most of these definitions and concepts, except the distinct types of vectors are not so clearly delineated in his theory (at least to my comprehension of it.) His theory does provide a nested, hyperdimensional structure of time, however, and thus allows "scalar" waves in the hyperspatial structure of time -- in other words, observably "scalar" waves in the virtual state structure of spacetime, or pure Tesla waves, or simply "time" waves.

These are the bare notes; from this approach, already new (proprietary) mechanisms and exact specifications to make scalar waves -- in essentially whatever quantity and degree desired -- appear to have been successfully accomplished by my close colleagues.

The new approach is real and it leads to a new physics. And I believe that the very beginnings of the new physics are already working on the laboratory bench.

Nikola Tesla discovered the most essential features of the new electromagnetics over eighty years ago and was simply suppressed for his efforts. Now, although it has been eighty years in the reborning, Tesla electromagnetics is once again loose in the Western world.

This time, let us hope that it fares better at the hands of orthodox science and large financial control groups than it did for Nikola Tesla.

Next Page

The Tom Bearden
Website

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

by Lt. Col. T.E. Bearden (retd.), 1983

Copyright

Title Page

Foreword

1. TOWARDS A NEW ELECTROMAGNETICS - ACKNOWLEDGEMENTS
2. OUTLINE
3. CHARACTERISTICS OF SELECTED CONCEPTS
4. QUOTE FROM NIKOLA TESLA
5. FORCE BETWEEN TWO LIKE CHARGES
6. REPETITIONS LEAD TO THE VECTOR "FIELD" CONCEPT
7. A SCALAR HAS A HIDDEN (VIRTUAL) VECTOR SUBSTRUCTURE
8. SCALAR AND VECTOR SUBSTRUCTURES
9. SUPERPOSITION DOES NOT ELIMINATE THE VIRTUAL SUBSTRUCTURE
10. PHYSICAL VECTORS
11. THE FOUR VECTORS OF PHYSICS
12. FORCE DEFINITION
13. FORCE IS AN EFFECT, NOT A CAUSE
14. FORCE AND Ë-FIELD DO NOT EXIST IN VACUUM
15. COROLLARIES
16. ENERGY IS DERIVED, NOT FUNDAMENTAL
17. EXISTENCE IN, AND PASSAGE THROUGH, TIME
18. SOME IMPLICATIONS
19. CONSERVATION OF ANENERGY (AE): (AN EXPANDED CONSERVATION LAW)
20. FREE ENERGY IS ALLOWED

21. SCALAR POTENTIAL CAN FORCE THE RELATIVISTIC CASE
22. SCALAR STRESS WAVE (Ö-WAVE) (TESLA WAVE)
23. SCALAR Ö-WAVE PRODUCTION
24. A SIMPLIFIED CHARGED PARTICLE
25. A CHARGED PARTICLE IS A SPECIAL KIND OF "SPRAY NOZZLE"
26. AN OBSERVABLE CHARGED PARTICLE IS DIRECTLY CONNECTED TO VIRTUAL PARTICLE FLUX VACUUM
27. SPIN COUPLES VACUUM TO PARTICLE
28. ELECTRON PRECESSION (SIMPLIFIED)
29. VACUUM/SPACETIME IS PURE VIRTUAL PARTICLE FLUX ($Ø_0$)
30. VACUUM IS:
31. GRADIENT OF ELECTROSTATIC POTENTIAL
32. IN A ∇Ø, A CHARGED PARTICLE MOVES ITSELF
33. ELECTROMAGNETIC WAVES IN ELECTRON GAS IN A CONDUCTOR
34. ELECTROMAGNETIC WAVES IN VACUUM
35. DETECTION OF TRANSVERSE AND LONGITUDINAL WAVES
36. CREATING A "RIVER" OF ANENERGY
37. A 2-WIRE SYSTEM IS CONSERVATIVE
38. A HOLLOW CHARGED SPHERE
39. ONE-WAY GATE VALVE
40. SIMPLE FREE ENERGY MOTOR
41. THE TRANSLATOR
42. TRANSVERSE WAVE INTERFERENCE
43. LONGITUDINAL WAVE INTERFERENCE (SCALAR WAVES)
44. CREATING ENERGY AT A DISTANCE
45. NONLINEARITIES CAUSE SHIFTING (PARTIAL TRANSLATION)
46. THE SCALAR AURA.
47. VIRTUAL FLUX AND MAGNETIC FIELD (LINEAR CASE)
48. PRODUCING MAGNETIC MONOPOLES
49. BREAKING METAL WITH MAGNETIC MONOPOLES
50. CREATING AN INERTIAL FIELD (ANTIGRAVITY)
51. EFFECTS ON HUMANS
52. INITIATING SPECIFIC BIOLOGICAL EFFECTS
53. GROUND-TO-AIR USE

54. TACTICAL AIR-TO-GROUND USAGE
55. INTEGRATED WEAPONS DISPLAY (MODIFIED JAMMER)
56. RADAR INVISIBILITY
57. NETWORK OF VIRTUAL TRANSMITTERS
58. CONCLUSION

Return to Homepage

TOWARD A NEW ELECTROMAGNETICS

PART 4: VECTORS AND MECHANISMS CLARIFIED

(©) 1983 T. E. Bearden

All rights reserved.

Delivered to the 1983 Annual Conference

of the

U. S. Psychotronics Association

July 20-24, 1983

at the

University of Portland

Portland, Oregon

The Tom Bearden Website

--FOREWORD --

This paper is an adaptation of the oral/visual presentation given by the author at the 1983 Annual Conference of the U. S. Psychotronics Association on Saturday, July 23, 1983.

The presentation was delivered directly from the vugraphs alone, and not from a formal, written text.

The title of the presentation at the conference was "New Electromagnetics: Where the 'Old Math' Took the Wrong Turn."

As originally conceived, the lecture was to be given in a single period of some 45 minutes and cover only the first part of this paper, the clarification of the vector concept in physical theory. The Symposium director so graciously scheduled two periods instead, allowing time to present additional material further clarifying vector and scalar waves, and to present a variety of unusual physical mechanisms that apply the concepts.

After the conference, the author prepared this paper directly from the same slides, in the same order. However, no attempt was made to match the words in this paper to the words actually delivered in the formal presentation. The gist of the two, however, are the same -- except that, in the conference lecture, the author covered (as an aside) some important additional material: (1) the nature of what actually occurs in an "electrical ground," and (2) the author's explanation of the theory of Tesla's magnifying transmitter.

An audio tape and a videotape of the author's USPA presentation were made by the conference officials. Copies are available at nominal cost from Mr. Robert Beutlich, Secretary/ Treasurer, U. S. Psychotronics Association, 3459 Montrose Avenue, Chicago, Illinois 60618, U.S.A.,
phone (312) 478-7715.

The Tom Bearden Website

SLIDE 1.

TOWARD A NEW ELECTROMAGNETICS

© 1983 T.E. BEARDEN

ALL RIGHTS RESERVED

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to two of my fellow colleagues, Dr. William Tiller and Mr. Frank Golden, for many fruitful hours of discourse and interchange on these and related subjects.

Also, I wish to thank the Association of Distinguished American Scientists (A.D.A.S.) for support and assistance in the long and arduous struggle to revise the basic foundations of electromagnetics and mechanics.

Next Slide

SLIDE 2.

OUTLINE

- **MAIN BRIEFING**
  - **> TRANSLATOR EFFECTS**
  - **> POTENTIAL WEAPONS**
  - **> INCIDENTS**
  - **> CONCLUSION**
- **HISTORICAL BACKGROUND**
  - **> TESLA/MORAY**
  - **> WWII AND AFTER**
- **ELEMENTS OF THE EMERGING THEORY**
  - **> LONGITUDINAL/ TRANSVERSE WAVES**
  - **> VACUUM/MASS EFFECTS**
  - **> FLAWS IN VECTOR THEORY**
  - **> IMPLICATIONS**

This presentation is one part of a three part presentation, consisting of the main briefing on physical effects, the historical background, and the elements of the emerging new electromagnetics.

In this present briefing I will not cover the historical background. Instead, first I will address the concepts of the new theory, and then I will give an abbreviated main briefing, including some weapon aspects. I will present one simple free energy motor and show how it works, and where it gets its energy.

Much of this material I have known for some time, but have hesitated to release it because it enables one to develop -- in a fairly straightforward manner -- weapons and beams (ray) systems that can disable electronics, computers, communications, and people. Anyone handy in electronics, who has a few thousand dollars and access to bench equipment and an electronics parts supply house, can develop these weapons in his or her basement workshop. But because certain very radical elements in this country already know some of this information and may plan to utilize such devices in criminal activities, I am openly releasing the information as a warning, to be prepared.

Of course there are many good things that can be developed with this technology, including machines to heal a remarkable variety of ailments and illnesses. I strongly urge everyone to use this information wisely and only to benefit other human beings. There exists a cosmic or karmic law for those who misuse it against their fellow persons; if one does that, one will assuredly suffer the consequences.

Also, I strongly urge the development of Western defenses against the Soviet strategic and tactical weapons; already developed utilizing these principles. Unless the West wakes up to the extreme urgency of this task, in the near future we may be overwhelmed in short order by a much harsher, dictatorial system. If that should happen, it will plunge the human race into a new dark ages far more terrible than that dark period in history following the fall of Rome.

These are the final times, and this is the beginning of the final technology. The ability to directly engineer physical reality itself has been laid in our laps. Pandora's box is already spilled open, and there is no longer any stopping the new technology.

Let us prepare, then, for that which is to come.

With these introductory remarks, let us now turn to the main presentation.

Next Slide

The Tom Bearden Website

SLIDE 3.

CHARACTERISTICS OF SELECTED CONCEPTS

CONCEPT	CHARACTERISTIC
• ENERGY • ANENERGY • PRESENT "CHARGE" • TRUE CHARGE	FLOW OF OBSERVABLE MASS PARTICLES FLOW OF VIRTUAL PARTICLES VIRTUAL PARTICLE FLUX TIED TO OBSERVABLE MASS PARTICLES VIRTUAL PARTICLE FLUX WITHOUT MASS
• FLUX • FLOW • CURRENT	MULTI-DIRECTIONAL FLOW THROUGH A POINT UNI-DIRECTIONAL MOVEMENT OF A FLUX [GRADIENT] UNI-DIRECTIONAL GRADIENT OF A [USUALLY CONSTRAINED] FLUX

On this slide I show some fundamental defining characteristics of energy, "charge" or charged mass as presently in the theory, flux, flow, and current. These fundamental concepts are often somewhat confused in the literature.

Particularly note that the concept of energy, being tied to "the capacity to do work," is specifically tied to the concept of mass. Eventually, energy must be expressed in terms of moving masses -- specifically, in accelerating or decelerating masses.

I introduce the term "anenergy" here, specifically in relation to

the concept of a virtual particle. Anenergy is roughly equivalent to the present idea of "virtual energy,"
except it more precisely implies the flow or flux of virtual particles.

We also redefine charge as virtual particle flux, in order to remove the present error in EM theory where "charge" always implies a charged mass. The electrostatic scalar potential, phi (Ø) , then merely becomes the intensity of the massless charge **--** that is, the intensity of the virtual particle flux -- at a point. Let me point out, however, that in the new view this "point" is always in n-dimensional space, where n is equal to or greater than four. The present 3-space formulation of Ø is thus a highly special case.

Next Slide

The Tom Bearden Website

SLIDE. 4

QUOTE FROM NIKOLA TESLA:

> **"THE HERTZ WAVE THEORY OF WIRELESS TRANSMISSION MAY BE KEPT UP FOR A WHILE, BUT I DO NOT HESITATE TO SAY THAT IN A SHORT TIME IT WILL BE RECOGNIZED AS ONE OF THE MOST REMARKABLE AND INEXPLICABLE ABERRATIONS OF THE SCIENTIFIC MIND WHICH HAS EVER BEEN RECORDED IN HISTORY."**

Nikola Tesla
"The True Wireless,"
The Electrical Experimenter

One of the things we will find is that Hertz or transverse electromagnetic waves are strictly mass waves in a charged mass particle medium.

For example, transverse EM waves do not, and CAN not, exist in vacuum. Vacuum EM waves are longitudinal waves of compression and rarefaction. What is compressed and rarified is the virtual particle flux that comprises vacuum itself. Vacuum EM waves in this virtual flux are just like sound waves in air -- in the flux of molecules and atoms that comprise the atmospheric gas.

Nikola Tesla, of course, knew this and he frequently pointed out the falsity of the accepted transverse wave theory. Here is one Tesla quote, for example, taken from The Electrical Experimenter. (PAUSE)

Almost every scientist -- even the "Tesla bugs" -- have thought Tesla wrong in his adamant opposition to the idea of Hertzian waves in the vacuum.

As we shall see, Tesla was right and all the present scientists and textbooks are wrong .

Next Slide

The Tom Bearden Website

SLIDE 5.

FORCE BETWEEN TWO LIKE CHARGES

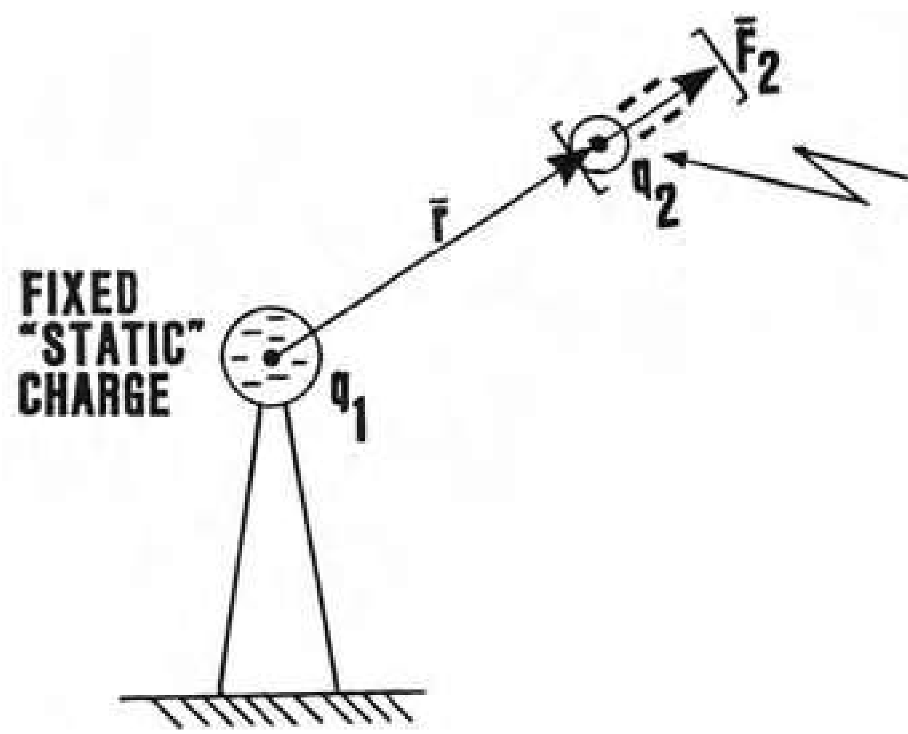

A NEGATIVELY CHARGED TEST MASS [e.g., AN ELECTRON] BROUGHT INTO THE VICINITY OF q_1 EXPERIENCES A FORCE $\bar{F}_2$ AS SHOWN. NOTE THE ELECTRON ACTUALLY IS SPINNING. THIS HAS NEVER BEEN DONE WITH A TRUE "STATIC CHARGE", FOR THERE IS NO SUCH THING IN NATURE.

In my presentation last year, I noted some 22 fundamental flaws in present electromagnetics theory. One of these flaws was in the actual vector theory itself. As we shall discover, one cannot simply plug geometrical vectors into physical systems without making fundamental errors. The vector concept itself must be adapted, for there are now four different kinds of vectors in physics, all confused as the same thing. We will come to that shortly.

To begin, let us look at the fundamental kind of experiment, shown on this slide, from which the idea of electrical force and electric field was taken.

We have a fixed static charged object q(l) in the laboratory, and we bring in a test charged mass q(2). The test charged mass may even be a single electron, as we illustrate here. We show in our example the fixed charge to be negative. When we release the test charge electron, it accelerates away from the fixed charged mass, as shown on the diagram. Note that we have no force at all until we bring in the test charged mass. Indeed, we shall find that the force is actually COMPRISED OF the accelerating test electron. Specifically, it is not something mysterious which appears in vacuum and PUSHES on the electron. We shall find that FORCE IS AN EFFECT, NOT A CAUSE. It CONSISTS OF the "smeared mass" that is accelerating -- in this case, it consists of the smeared, accelerating electron.

We summarize the experiment: in the absence of the charged test particle, no force exists in the vacuum. When we bring in the test charged mass, it accelerates away, and THE FORCE CONSISTS OF THAT ACCELERATING CHARGED MASS PARTICLE. The experiment does not at all address or deal with WHAT EXISTS IN THE VACUUM AROUND CHARGED MASS Q(l) IN THE ABSENCE OF TEST CHARGED MASS Q(2). Any assumption of a force existing in the vacuum is sheer speculation and not based on this experiment.

Next Slide

The Tom Bearden
Website

SLIDE 6.

REPETITIONS LEAD TO THE VECTOR "FIELD" CONCEPT

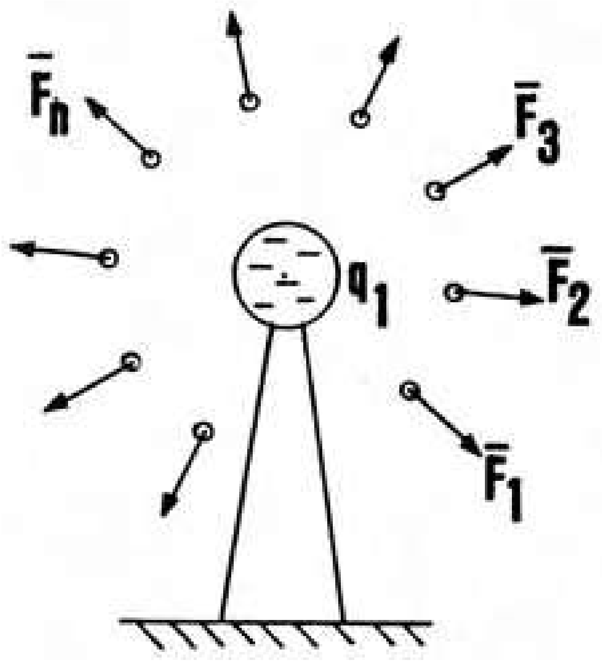

REPEATING FIGURE 4 A NUMBER OF TIMES LEADS TO SETS OF FORCE ARROWS, WHICH WE THEN VISUALIZE IN TOTAL AS A FIELD, Ë, OF FORCE(S) ABOUT q_1

RIGOROUSLY, THIS "FORCE FIELD" WOULD REQUIRE THAT A SPINNING CHARGED PARTICLE OF MASS EXIST AT EACH AND EVERY POINT IN SPACE ABOUT q_1

On this slide, we show that, as we repeat the experiment by bringing in the test charged mass from differing directions, the same thing happens around the fixed charged mass q(l). Test charged mass q (2) -- our electron -- accelerates radially away from q(l).

As we repeat the experiment and draw many of the little arrows radially outward, gradually the idea of a "vector field" emerges. That is, we form the idea that, at every point in space around fixed charge q (l), there exists one of our little vector arrows. Thus we assign a vector field to the space: to every point of it, there is assigned both a magnitude and a direction.

However, note that this does not describe the situation as it exists, at any time, in the "present." That is, rigorously any one arrow only exists whenever a little test charged mass is present and accelerating radially away.

We may take the view that the arrow WILL exist WHEN a test charged mass is brought to that point and released. In that case, the arrow WILL exist in the FUTURE, not the present. And we may refer to that as, it POTENTIALLY exists, when and if.. ..etc.

So already we see a fundamental logical problem in the idea of such a field. It WILL exist in the future, but only POTENTIALLY exists in the present.

Of course I am not the first one to point out such difficulties; the fact that the field concept is erroneous is already well-known to foundations scientists. But it is such a USEFUL concept that it is retained.

Now the Ë-field, or electric field, is defined as "the accelerating force on a charged mass per unit charged mass accelerated by it." If there is no charged mass present to be accelerated, then there can be no Ë-field present.

So Ë-field cannot exist in vacuum, for vacuum presupposes the absence of all observable mass.

So we can rigorously say that no Ë-field of force exists around the charged mass q(l) IN THE PRESENT. At best, it can only POTENTIALLY exist IN THE FUTURE.

Next Slide

The Tom Bearden Website

SLIDE 7.

A SCALAR HAS A HIDDEN [VIRTUAL] VECTOR SUBSTRUCTURE

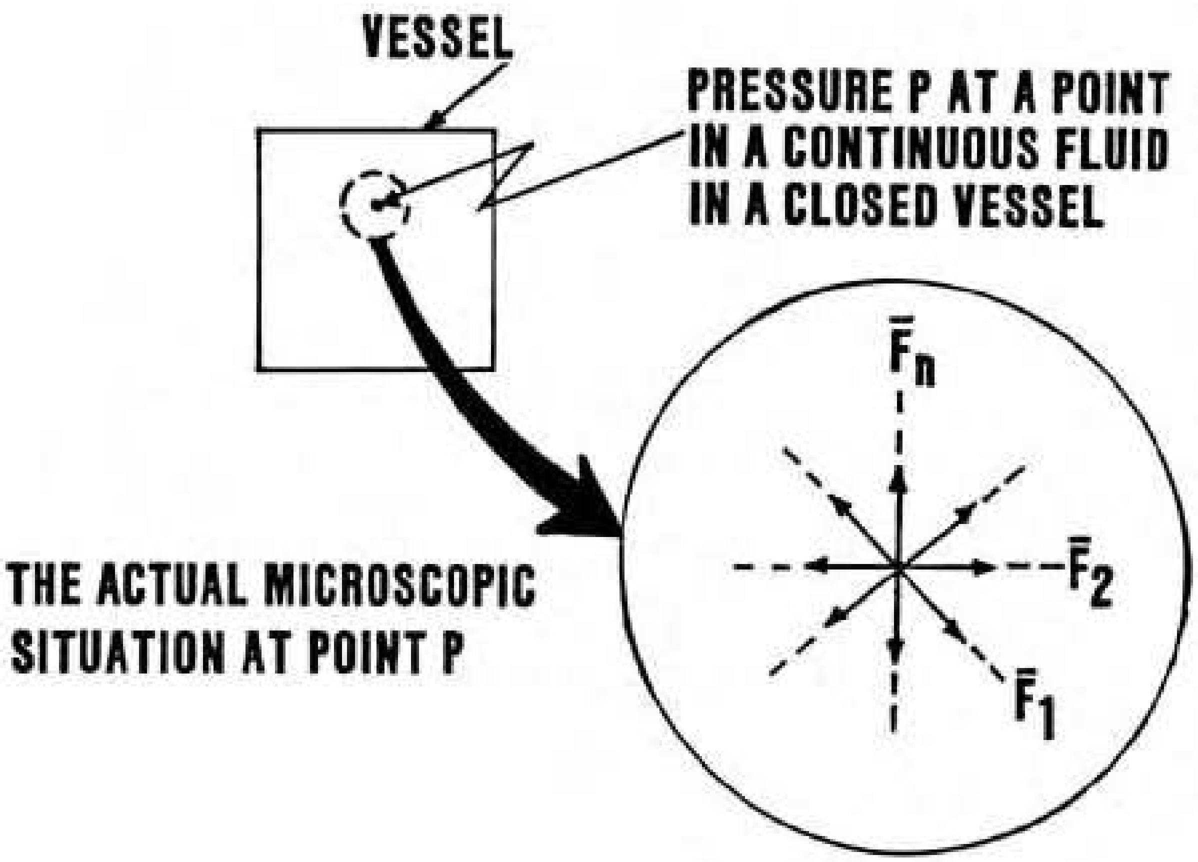
VESSEL
PRESSURE P AT A POINT IN A CONTINUOUS FLUID IN A CLOSED VESSEL
THE ACTUAL MICROSCOPIC SITUATION AT POINT P
$\bar{F}_n$
$\bar{F}_2$
$\bar{F}_1$

Our next slide shows that a physical scalar quantity has a hidden vector substructure.

For example, we show on the left a closed vessel containing a pressurized gas. The pressure at a point in the gas is a scalar quantity, since it has a magnitude but not a direction. However, physically what we have in the gas is a flux of individual molecules continually through any small volume, which we may let approach a point in the limit. Thus we have a flux in all directions through the point at which we specify pressure.

That is, the scalar "pressure" at a point may be said to contain a substructure of individual molecule flow vectors.

All physical process scalars have such hidden substructures. In the equilibrium case, the summation of all the individual flux vectors yields a resultant vector zero.

Thus we take the view that ANY PHYSICAL PROCESS SCALAR IS ALSO A ZERO VECTOR, AND ANY ZERO VECTOR IS A SCALAR AND CONTAINS A SUM-ZERO HIDDEN VECTOR SUBSTRUCTURE.

Next Slide

The Tom Bearden Website

SLIDE 8.

SCALAR AND VECTOR SUBSTRUCTURES

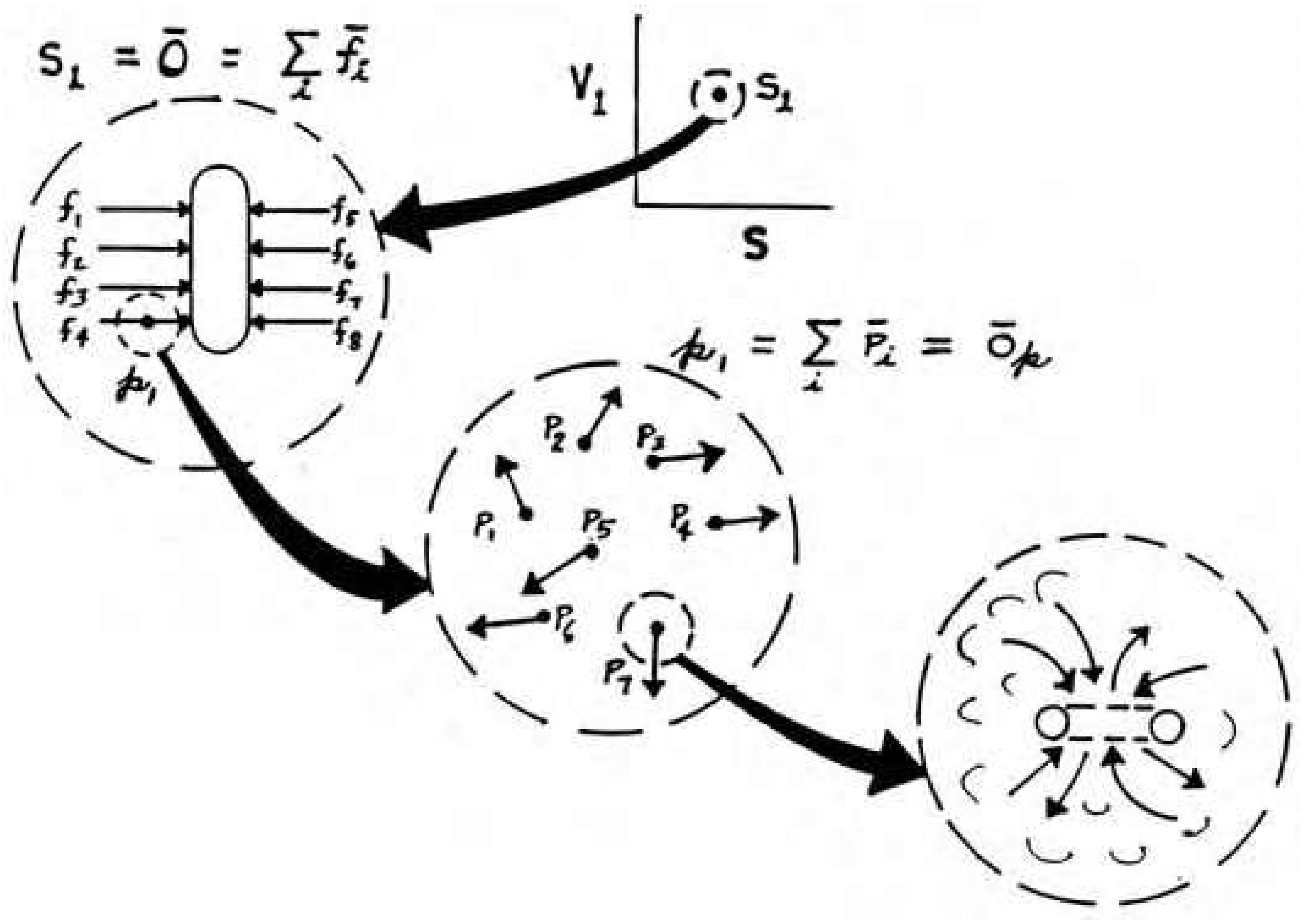
$S_1 = \bar{0} = \sum_i \bar{f}_i$
V_1
S_1
S
f_1
f_2
f_3
f_4
f_5
f_6
f_7
f_8
p_1
$p_1 = \sum_i \bar{P}_i = \bar{0}_p$
P_2
P_3
P_1
P_5
P_4
P_6
P_7

Now let us see if we can do some unusual thinking about the hidden vector substructure of a scalar or zero-vector.

On the leftmost figure in this slide, we show a plate with balanced opposing forces. The vector summation of all the forces' is zero, so the system has no resultant vector. The plate therefore does not move. However, the plate -- that is, the medium in which the forces act -- is stressed, much like the previous example of pressure in a gas. The difference is that now the stress is patterned in its substructure, and not random. Now suppose we rhythmically vary the forces, each the same, in a regular manner in time. Yet we maintain a continuous balance, so there is never a system resultant greater than zero. In this case, we maintain our vector zero, but we are producing regular stress waves in the plate medium..

I think you now can see that I can vary the individual opposing vector pairs so that rhythmic waves flow from top to bottom, or from bottom to top. That, of course, will describe a vector gradient traveling through the medium, and that will be an ordinary "vector wave." Is there any way I can make waves through the medium, and not have such vector waves ever appear?

Indeed there is, if I look further into the substructure. Take one point on one of the vectors shown acting on the plate in the left diagram. That point is a scalar, and it also has a substructure of even finer vectors, as shown in the lower middle figure. And any point in one of these second-order hidden vectors is a scalar and composed of a still finer hidden vector substructure, as shown in the lower right figure. This process repeats ad infinitum, and all is structures within structures. At least it can be modeled that way, for it already is treated that way in particle physics.

Now go back to the second order vector substructure. Here I can establish whatever patterns I wish, and not affect the first order pattern shown in the figure on the left. I can even rhythmically vary things in the second order substructure, and leave a blissfully unchanged first order substructure, to the first order observer.

So I can make as many waves; as I wish through the medium, as long as I stay second order or higher, and the external observer will never see any change at all unless he has a special detector of some sort to detect the lower substructure changes.

As I showed in the Appendix to my book, The Excalibur Briefing, these substructures are in reality the same thing as hyperspaces. And movements in them can proceed much faster than the speed of light.

So I can have higher order signals and waves through the stressed medium without changing the overall first order stress in the medium.

This is a sort of super-relativity; the present relativity is just a special case.

Superluminal communication systems are perfectly possible, but not with the present first order transmitters and receiver/detectors.

Note also that higher order stresses -- stresses in the higher order substructures -- represent special stresses in time, and time has an internal stress-pattern substructure. The higher order stress waves are phase waves or phase fluctuations in that respect.

Also note that Charles Musés has already created the hypernumber mathematics of the structure of time, and it is largely in his mathematics that these concepts will have to be mathematically modeled.

Next Slide

[The Tom Bearden Website](#)

SLIDE 9.

SUPERPOSITION DOES NOT ELIMINATE THE VIRTUAL SUBSTRUCTURE

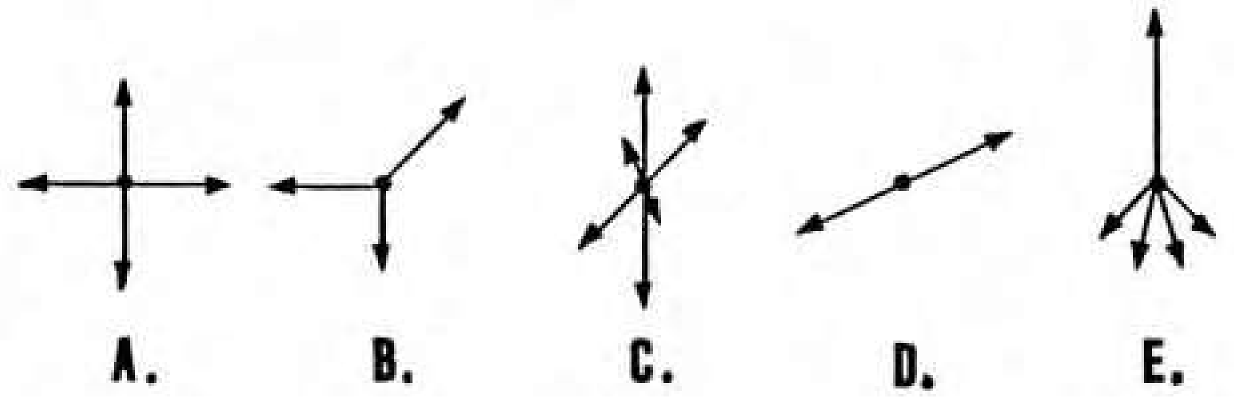

TWO OR MORE MACROSCOPICALLY "EQUAL" VECTORS OR SCALARS CAN DIFFER DRASTICALLY IN THEIR VIRTUAL SUBSTRUCTURE DISTRIBUTIONS. EVEN TWO "ZEROS" MAY BE DYNAMICALLY DIFFERENT DUE TO THEIR DIFFERING SUBSTRUCTURES.

Next we show some simple substructures of a zero vector. All of these vector systems sum to zero vectorially. Note that if we superpose two of these structures, all substructure vectors are still present., and the vector summation is still zero.

Also notice that, from the standpoint of substructuring, zeros can be and are drastically different.

We can even make rhythmic waves in the substructure by superposing patterned zero vectors in a regular fashion.

Next Slide

The Tom Bearden Website

SLIDE 10.

PHYSICAL VECTORS

- **VECTOR: A "SMEARED PARTICLE"**
- **FOUR MAJOR PARTICLES IN PHYSICS:**
  - POINT [MASSLESS, CHARGELESS]
  - MASS [CHARGELESS]
  - CHARGED MASS
  - CHARGED POINT [MASSLESS}
- **LEADS TO FOUR MAJOR VECTORS:**
  - GEOMETER'S
  - MECHANIC'S
  - ELECTRICIAN'S
  - ADVANCED ELECTRICIAN'S

In quantum mechanics, as we approach the quantum level, we find that the notions of a particle (static concept) and a wave become inseparable. That is, any particle becomes a "smear" in length and time, and it becomes inseparable from its smear. At the quantum level, there is no separate energy, length, mass, momentum, energy, or time -- all are welded together inextricably into a quantity called "action," having the units of angular momentum.

So at the fundamental quantum level, a moving point becomes a smear, and a vector represents a physical, undifferentiated entity.

To go quickly, we now define a vector as consisting of a smeared particle. That is, the particle is smeared in both time and length, and it is not differentiated from time or length.

There are four kinds of particles in physics which can be smeared to form vectors. First, there is a point in space or spacetime. This point may be approximated as massless and chargeless. At least that is the way the geometers of old conceived it in the abstract, and so the idea of a "point" in space assumes it to be chargeless and massless. Second, there is a tiny particle of mass, often abstracted to be a "point-mass," but visualized as having no charge. This is the mechanic's particle. Third, there is a tiny charged mass, such as an electron, and this is the electrician's particle. Again, it is often abstracted as a charged point-mass. Fourth, there is a charged point, having no mass, and this is the advanced electrician's particle, which he uses to form the idea of potentials and fields.

Smearing these four particles at the fundamental quantum level produces four different kinds of physical vectors: the geometer's, the mechanic's, the electrician's and the advanced electrician's vectors. They are all different.

Next Slide

The Tom Bearden Website

SLIDE 11.

THE FOUR VECTORS OF PHYSICS

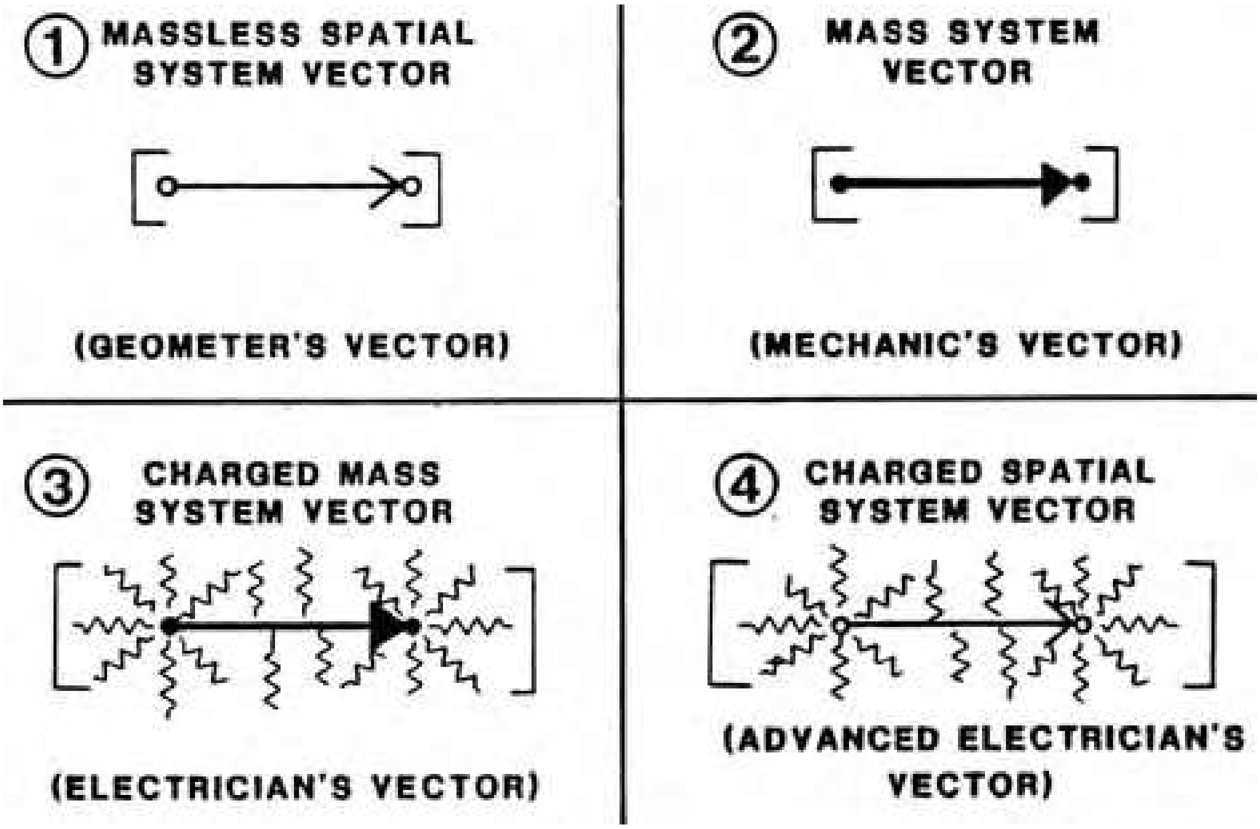
① MASSLESS SPATIAL SYSTEM VECTOR
(GEOMETER'S VECTOR)
② MASS SYSTEM VECTOR
(MECHANIC'S VECTOR)
③ CHARGED MASS SYSTEM VECTOR
(ELECTRICIAN'S VECTOR)
④ CHARGED SPATIAL SYSTEM VECTOR
(ADVANCED ELECTRICIAN'S VECTOR)

On this slide we show a graphic illustration of each of these vectors.

Note that they may be visualized as smeared particles, but the particle is not separate from the smear in time and length.

Each vector must be considered as a unitary entity -- it is a special kind of quantum, if you will.

Each is a special quantum of motion.

And Dewey Larson has it right when he says, "All is motion."

Next Slide

The Tom Bearden Website

SLIDE 12.

FORCE DEFINITION

- ***A TRUE DEFINITION IS AN IDENTITY.***

  $$F \equiv d/dt\,(mv)$$

- **FORCE CONSISTS OF**
  THE TIME RATE OF CHANGE OF MOMENTUM

- **MOMENTUM IS A MASS SMEARED THROUGH**
  A LENGTH IN A CERTAIN TIME

- **FORCE CONSISTS OF MASS-MOTION-CHANGE**

- ***FORCE IS AN EFFECT, NOT A CAUSE !***

We now look a little closer at the concept of a force.

I point out that a true definition is an identity. If a supposed definition of an entity is not an identity for that entity, then it is not a definition at all, but only a statement ABOUT the entity.

It may come as somewhat of a shock to the normal engineer and scientist that the foundations concepts of physics -- such as force, mass, energy, etc. -- are ALL in serious difficulty and contain many unresolved logical conflicts.

Force is in that kind of difficulty, and it has that kind of logical conflict.

In fact, in its force equations, physics has not even applied its own present definition of force.

As we show here, force is generally defined as the time rate of change of momentum.

If that is true, then as an identity the definition states that a force CONSISTS OF the time rate of change of momentum.

And of course, momentum is mass times velocity. Hence momentum is a mass smeared through a length and through a certain time.

Force then consists of the time rate of change of that smearing. That is, we smear the mass through time and length, faster or slower.

But if force CONSISTS OF that smearing change, it does not CAUSE it.

Force is thus an effect or result, not a cause.

This implies that there is a more fundamental mechanism that causes or generates force itself -- ANY force. If we can find that hidden mechanism, we may well find the long-sought unified field theory, since it will unify all forces on a common mechanism.

Next Slide

The Tom Bearden Website

SLIDE 13.

FORCE IS AN EFFECT NOT A CAUSE

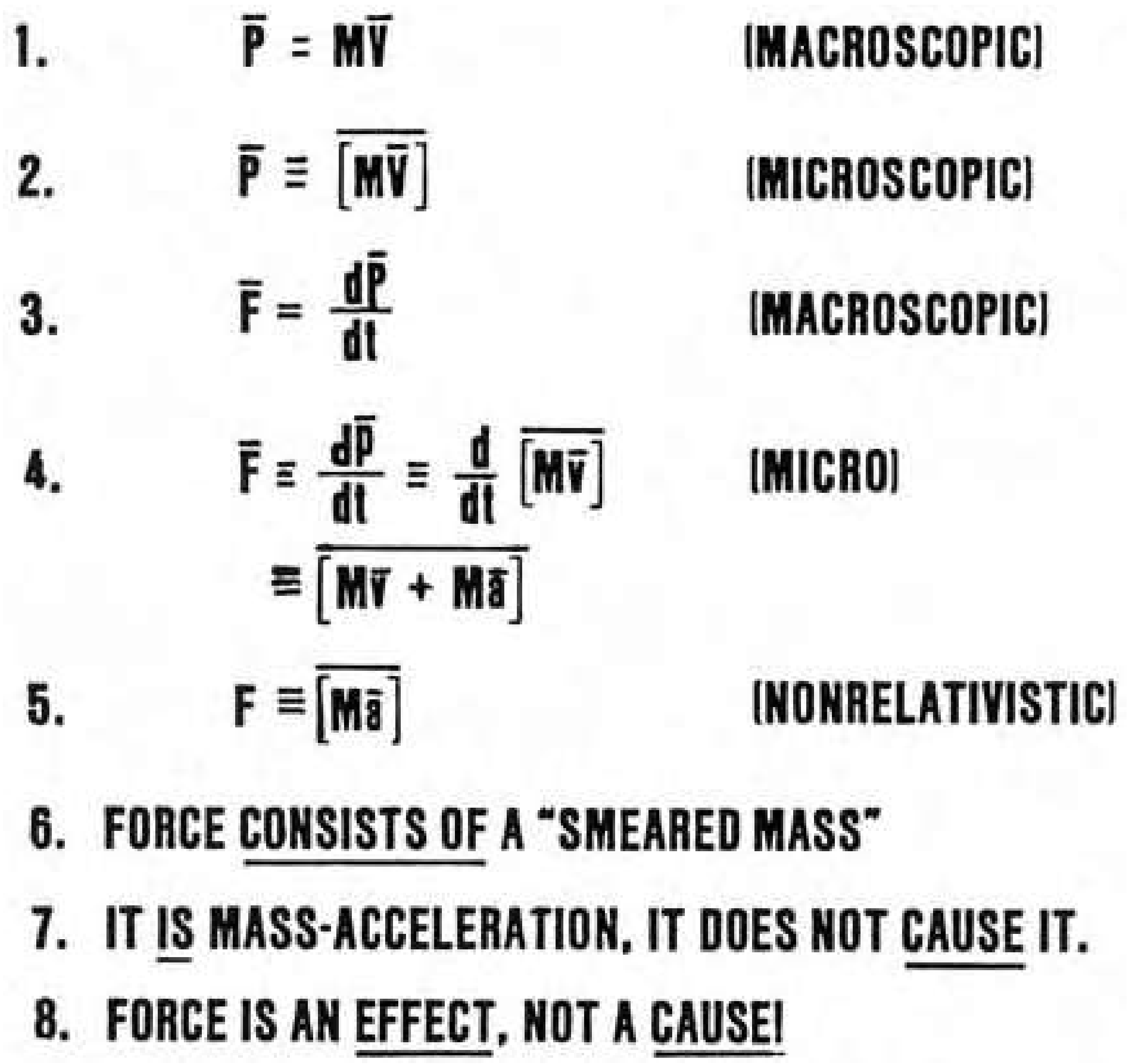

1. $\bar{P} = M\bar{V}$ (MACROSCOPIC)
2. $\bar{P} \equiv \overline{[M\bar{V}]}$ (MICROSCOPIC)
3. $\bar{F} = \frac{d\bar{P}}{dt}$ (MACROSCOPIC)
4. $\bar{F} \equiv \frac{d\bar{P}}{dt} \equiv \frac{d}{dt}\overline{[M\bar{v}]} \equiv \overline{[M\bar{v} + M\bar{a}]}$ (MICRO)
5. $F \equiv \overline{[M\bar{a}]}$ (NONRELATIVISTIC)
6. FORCE CONSISTS OF A "SMEARED MASS"
7. IT IS MASS-ACCELERATION, IT DOES NOT CAUSE IT.
8. FORCE IS AN EFFECT, NOT A CAUSE!

On this slide, we show briefly how one might express this idea in mathematical symbology.

Macroscopically, momentum consists of mass times velocity. Microscopically, the mass and the velocity are inextricable, and it is not proper to speak of a "mass in motion," but only of "mass-motion." We show this in equation 2.

We take the macro definition of momentum, as shown in equation 3, and express it for the microscopic, unseparated case, as shown in equation 4.

In the nonrelativistic case, we may consider m-dot to be zero, since the time rate of change of mass is almost zero. In that case, force may be defined for the microscopic case as a smeared accelerating mass, as shown in equation 5.

Force thus consists of the welded mass and acceleration, with no seam in the middle.

It IS mass-acceleration; it does not CAUSE it. Therefore force is an effect, not a cause.

Next Slide

The Tom Bearden Website

SLIDE. 14

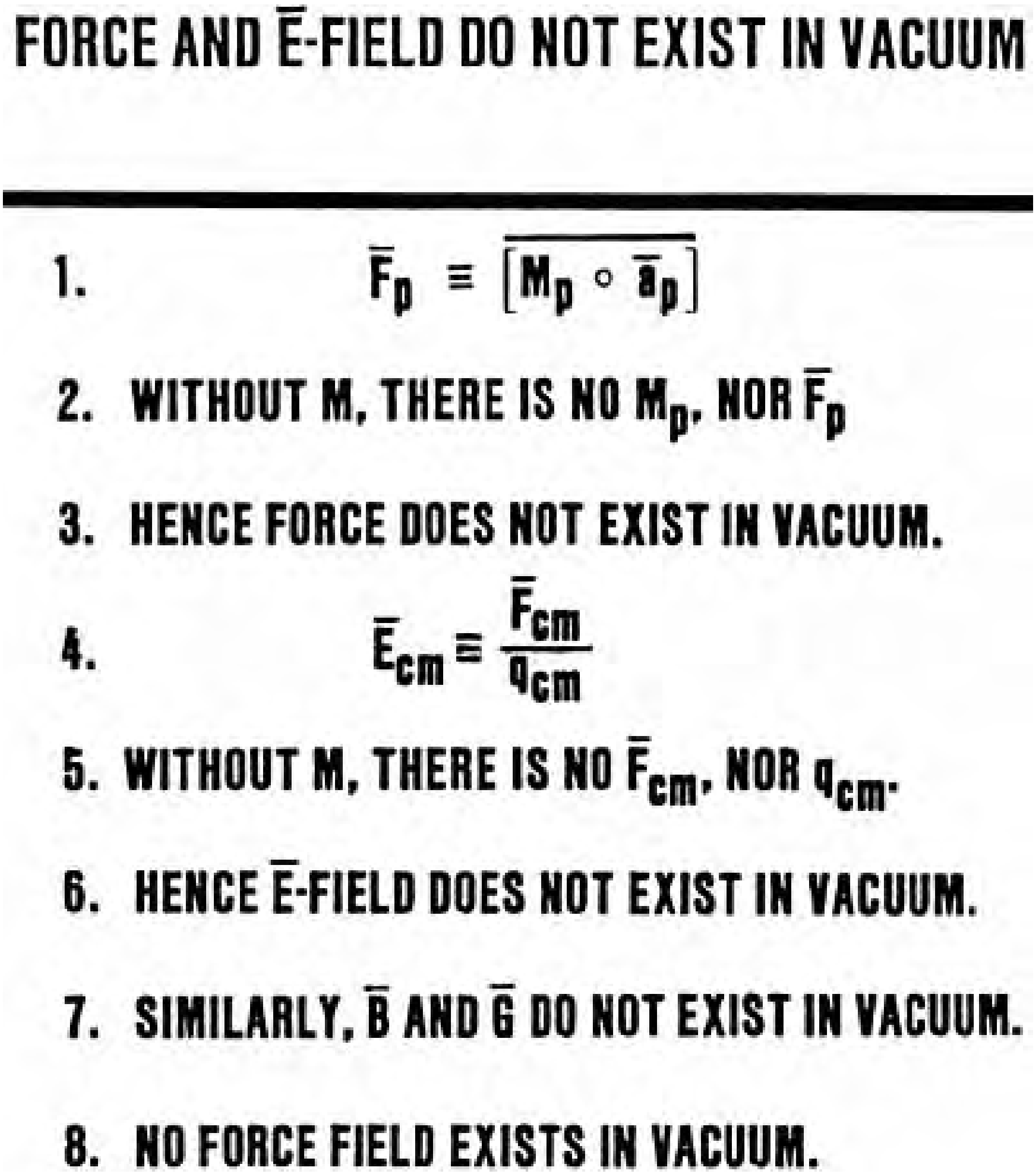

FORCE AND $\bar{E}$-FIELD DO NOT EXIST IN VACUUM

1. $\bar{F}_p \equiv \overline{\left[M_p \circ \bar{a}_p\right]}$
2. WITHOUT M, THERE IS NO M_p, NOR $\bar{F}_p$
3. HENCE FORCE DOES NOT EXIST IN VACUUM.
4. $\bar{E}_{cm} \equiv \dfrac{\bar{F}_{cm}}{q_{cm}}$
5. WITHOUT M, THERE IS NO $\bar{F}_{cm}$, NOR q_{cm}.
6. HENCE $\bar{E}$-FIELD DOES NOT EXIST IN VACUUM.
7. SIMILARLY, $\bar{B}$ AND $\bar{G}$ DO NOT EXIST IN VACUUM.
8. NO FORCE FIELD EXISTS IN VACUUM.

But if force CONSISTS OF mass welded to acceleration, it cannot exist in the absence of mass.

Hence force does not exist in vacuum.

Further, the Ë-field (Electric field) of a charged particle is defined as the force on a charged mass, divided by the mass. That is, the Ë-field is thought to represent the force per unit charged mass. And then it is assumed to exist in vacuum.

But vacuum is without mass or charged mass, observably. Without mass, neither force nor charged mass exists.

Therefore the Ë-field does not exist in the vacuum.

Magnetic field (B) and gravitational field (G) are similarly defined, and they have the same logical difficulties.

Therefore neither magnetic field nor gravitational field exist in vacuum.

A priori, NO force field exists in vacuum.

The greatest error in physics has been the assignment of a force as a cause, and thus assigning it to the vacuum. The force is an effect, and it never exists in vacuum, a priori.

Next Slide

SLIDE 15.

COROLLARIES

- **MORE PRIMARY MECHANISM CAUSES/GENERATES FORCE -- ANY FORCE**
- **FORCE DOES NOT EXIST IN VACUUM, SINCE VACUUM IS THE ABSENCE OF MASS**
- **FORCE FIELDS DO NOT EXIST IN VACUUM**
  - **E-FIELD**
  - **B-FIELD**
- **TRANSVERSE [HERTZIAN] WAVES DO NOT EXIST IN VACUUM**

Without belaboring the issue, I now point out and summarize some of these conclusions.

First, a more primary mechanism causes or generates force -- ANY force.

Second, force implies the presence of mass a priori. Force does not exist in vacuum, since vacuum is the absence of mass.

_

Third, no force fields of any kind exist in vacuum, including Ë-field, B-field, and gravitational field.

_

Fourth, since Hertzian waves are defined in terms of Ë and H force fields, and these fields do not exist in vacuum, then Hertzian waves do not exist in vacuum.

Tesla pointed out long ago that vacuum EM waves are longitudinal waves, not Hertz waves. Vacuum EM waves are similar to sound waves in a gas.

The Hertz wave exists as a matter wave in the electron gas in a transmitting antenna, and in the electron gas in a receiving antenna, but does not exist in the vacuum in between the two antennas. Between them, EM waves are longitudinal waves.

Next Slide

The Tom Bearden Website

SLIDE 16.

ENERGY IS DERIVED, NOT FUNDAMENTAL

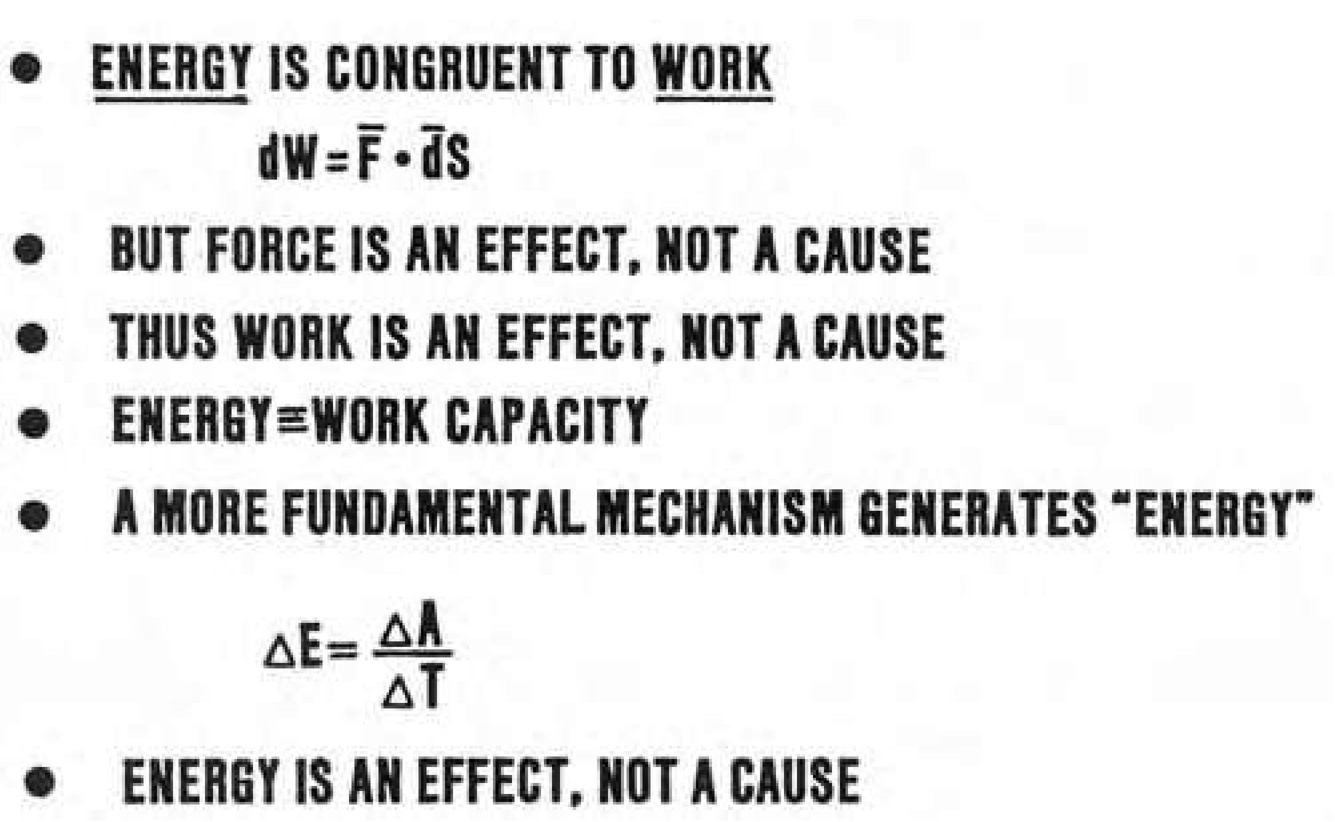

- ENERGY IS CONGRUENT TO WORK

$$dW = \bar{F} \cdot \bar{d}S$$

- BUT FORCE IS AN EFFECT, NOT A CAUSE
- THUS WORK IS AN EFFECT, NOT A CAUSE
- ENERGY≅WORK CAPACITY
- A MORE FUNDAMENTAL MECHANISM GENERATES "ENERGY"

$$\Delta E = \frac{\Delta A}{\Delta T}$$

- ENERGY IS AN EFFECT, NOT A CAUSE

Energy is usually taken as "the capacity to do work" at some future time. In other words, when we refer to the energy of something, we refer to HOW MUCH WORK IT CAN DO IN THE FUTURE, IF AND WHEN WE ARRANGE FOR IT TO DO IT. Rigorously, a body has no "work" at a particular instant in time, so we use "energy" to refer to potential work that can, or will, be done in the future. Thus energy is always congruent to work in the future.

And for that reason, the units of energy are the units of work; the only difference is in the time the work is done, relative to the "now" or "present" we are speaking of.

Energy is the potential for work to be done in the future.

But what is work?

Well, work is defined in terms of the movement of a force through a distance. And it is understood that only the projection of the force along the line of movement is "working." Thus work is defined differentially as shown on the slide. The differential of work is equal to the dot product of the moving force times the differential of displacement.

But as we have seen, force is an effect, not a cause.

That is, when we calculate the work done, it's always in the past, never in the present or future. It's always been DONE, and has been expended.

Being in the past, it's gone forever, so it certainly isn't the CAUSE of anything which happens in the future.

Thus work is an effect, not a cause.

That's interesting, because then it certainly does not directly generate "energy," since energy is something in the present which speaks of future work to be done.

Rigorously, the fundamental quantum -- which is what exists in the present, in unseparated and unquantized form -- consists of action. NOT energy, NOT time, NOT length, NOT momentum exclusively, but all of these INclusively.

And a change in energy occurs when a quantum is separated at a certain rate in time. Thus delta energy equals the time rate of change of action, or delta action divided by the delta time used in separating the quantum and fissioning it into parts (collapsing the wave function).

Thus energy itself is an effect or result of a more fundamental mechanism, and it is not a cause.

In quantum mechanics all this has been neatly "buried" in the mysterious "collapse of the wave function," which is supposedly what "causes" everything to happen. And then we swear that what results after the collapse is a totally statistical selection from the possible states of the system we are examining. We thus say that the microworld is totally statistical at base. Of course we also build a wave equation or wave function which propagates forward in time with rigorous causality, and that wave or function absolutely controls the basket of "possible states" allowed .

All that says is that the "wave function times time" is what exists in the present, when observation has not been made by "stopping time." And then we "stop time" in a rather blunt manner, somewhat like firing a billiard ball blindly into a group of billiard balls on a table, and we live with what gets hit and pops out.

If we examine this "time" that is flowing, look into its substructure, and arrange that substructure as we will, we can control which "billiard ball" is hit and which pocket it goes into. By working on the vector substructures of time, we can thus deliberately control the microworld and make it causal, not statistical. Quantum mechanics is in error in stopping where it has; it is presently only a first order approximation of the real quantum mechanics that awaits.

Einstein was correct after all; God does NOT play dice with the universe when one considers all levels.

One finds oneself in a statistical situation at one level only when the constituent sublevels of it are uncontrolled.

Next Chapter

The Tom Bearden Website

SLIDE 17.

EXISTENCE IN, AND PASSAGE THROUGH, TIME

- **IN RELATIVITY, TRANSLATION OF A BODY AND PASSAGE THROUGH TIME ARE RELATED**

  $$\Delta L = C \Delta T$$

- **"MASS" IS SPATIAL**
  - DOES NOT EXIST IN TIME
  - "MASS-TIME" EXISTS IN TIME

- **VIRTUAL FLUX TO AND FROM A BARE PARTICLE DETERMINES ITS "PASSAGE THRU TIME"**

- **SPHERICAL SPIN OF A MASS PARTICLE [L^3] COUPLES IT TO THE SPACETIME VACUUM [L^3T] VIRTUAL PARTICLE FLUX**

- **PASSAGE THROUGH ORDINARY TIME IS CONTROLLED BY CREATION AND ANNIHILATION OF QUANTA**

On this slide, we simply point out that translation of a body through space and its passage through time are related, according to relativity.

Now "mass" is spatial.

It does not exist in time.

Only things which contain the unit of time can "exist in time."

Otherwise, the thing only exists as such when the passage of time has been momentarily stopped -- usually by photon emission and the consequent carrying off of the time unit by the photon.

Since every mass that has temperature -- and every mass does -- is continually absorbing and emitting photons, then the mass is continually grabbing a time tail (absorption) and losing it (emission) shortly thereafter.

Masstime is what exists in time, not mass.

And every mass particle is continually turning from mass to masstime to mass, etc.

Indeed, it is the summation of the absorption and emission of virtual subquanta by a mass particle -- i.e., its "charge" -- that determines its rate of flow through "time."

And the spherical spin of a mass particle couples the particle of mass to the spacetime vacuum; that is, to the virtual particle flux.

Passage through ordinary time is controlled by creation and annihilation of quanta, and these processes are controlled by substructure (subquantal) actions and interactions.

Next Slide

The Tom Bearden Website

SLIDE 18.

SOME IMPLICATIONS

- **MODIFIES CONSERVATION OF CHARGE**
- **MODIFIES CONSERVATION OF ENERGY**
  - ***MASS***
  - ***ENERGY***
  - ***ANENERGY (MASSLESS CHARGE)***
- **MODIFIES NEWTON'S LAW**
  - ***FOR EVERY ACTION, THERE IS AN OPPOSITE AND EQUAL REACTION***
  - ***REACTION NEED NOT BE LOCAL***
  - ***REACTION NEED NOT BE ANTIPARALLEL***
- **UNIFIED FIELD THEORY**

Without belaboring the issue, I now simply point out some implications of the new approach.

First, we modify two laws -- the conservation of charge and the conservation of energy -- by combining them into one law, a new conservation law in which anenergy is conserved.

Neither charge per se nor energy per se need be conserved; but the overall anenergy (massless charge) equivalency is what is conserved. This allows us to derive mechanisms for free energy from the vacuum, for example. These mechanisms readily violate the present conservation of energy law, but do not violate the overall conservation of anenergy law.

It is the total anenergy equivalency of mass, energy, and massless charge that is now to be conserved in general. The old laws are now just special cases.

Also, Newton's third law -- for every action there is an equal and opposite reaction -- is modified. It still applies, but in changed form.

The reaction still occurs, but it need not be local, and it need not be antiparallel to the original action.

Thus it is now possible to design and build a "locally reactionless" inertial space drive for a vehicle.

The final implication is that the common mechanism producing force -- any force -- provides a unified field theory. Specifically, it provides one which easily lends itself to direct engineering. It is not esoteric nor mysterious. It is practical and physical, and it can be done with some modifications to our present electrical circuits and devices.

Next Slide

The Tom Bearden
Website

SLIDE 19.

CONSERVATION OF ANENERGY (AE):
(AN EXPANDED ENERGY CONSERVATION LAW)

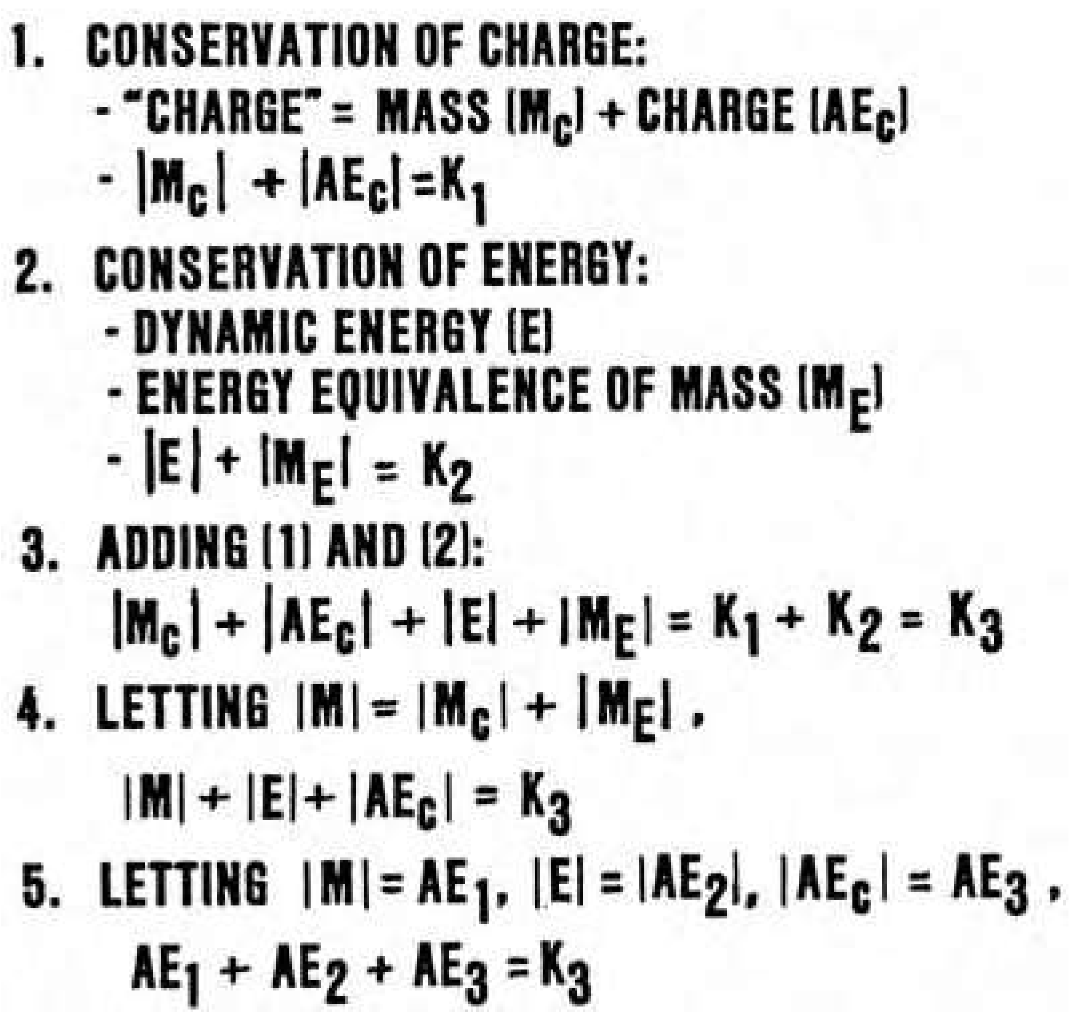

1. CONSERVATION OF CHARGE:
   - "CHARGE" = MASS (M_C) + CHARGE (AE_C)
   - $|M_C| + |AE_C| = K_1$
2. CONSERVATION OF ENERGY:
   - DYNAMIC ENERGY (E)
   - ENERGY EQUIVALENCE OF MASS (M_E)
   - $|E| + |M_E| = K_2$
3. ADDING (1) AND (2):

   $$|M_C| + |AE_C| + |E| + |M_E| = K_1 + K_2 = K_3$$

4. LETTING $|M| = |M_C| + |M_E|$,

   $$|M| + |E| + |AE_C| = K_3$$

5. LETTING $|M| = AE_1$, $|E| = |AE_2|$, $|AE_C| = AE_3$,

   $$AE_1 + AE_2 + AE_3 = K_3$$

We briefly show a simplified derivation of the new conservation of anenergy law.

We express the conservation of charge law by correcting the term "charge" to show its true meaning: charged mass. We use the absolute value symbol to mean the "anenergy equivalency."

As shown on the slide, we account separately :for the mass (m-sub-c) of the conserved charges and the virtual particle charge flux (the anenergy -- AE-sub-c) of the charges conserved. That law now says that the absolute values of the masses and the massless charge fluxes have a total constant summation.

For a dynamic mass, we do a similar thing, in this simplified case ignoring potential energy of position, and taking the energy equivalence of the dynamic energy (M-sub-E) and the rest mass energy of the particles, M-sub-E. The conservation of energy now states that these two terms have a constant absolute value summation.

We add equations 1 and 2, getting equation 3. We consolidate like terms to get equation 4.

Expressing all terms in their anenergy equivalencies (absolute values), we get equation 5, which is the overall conservation of anenergy equivalency.

Note that we now can change mass into energy or anenergy (massless charge), energy into mass or anenergy, and anenergy into mass or energy.

Thus we can turn everything into pure vacuum ether (pure massless charge, pure anenergy, pure electrostatic scalar potential, pure stress on spacetime) or we can turn part of the vacuum into energy or particles of mass.

Vacuum now becomes a very real thing indeed.

It is pure anenergy.

Pure massless charge, pure electrostatic scalar potential, pure "broken bits" (subquantal bits) of (quantized) energy.

The vacuum is identically the virtual state.

It is not an emptiness filled with virtual things; rather, it IS the sum total of those virtual things, for they are totally unseparated until some sort of "conceptual observation" is invoked.

Physics should have adopted Einstein's suggestion that the vacuum, complete with its EM and gravitational fields, be called the ether.

Next Slide

The Tom Bearden Website

SLIDE 20.

FREE ENERGY IS ALLOWED

- **ANY ONE-WAY GATING PROCESS FROM ANENERGY TO ENERGY YIELDS "FREE ENERGY" SINCE THE ANENERGY IS AUTOMATICALLY REPLENISHED BY THE UNIVERSE.**

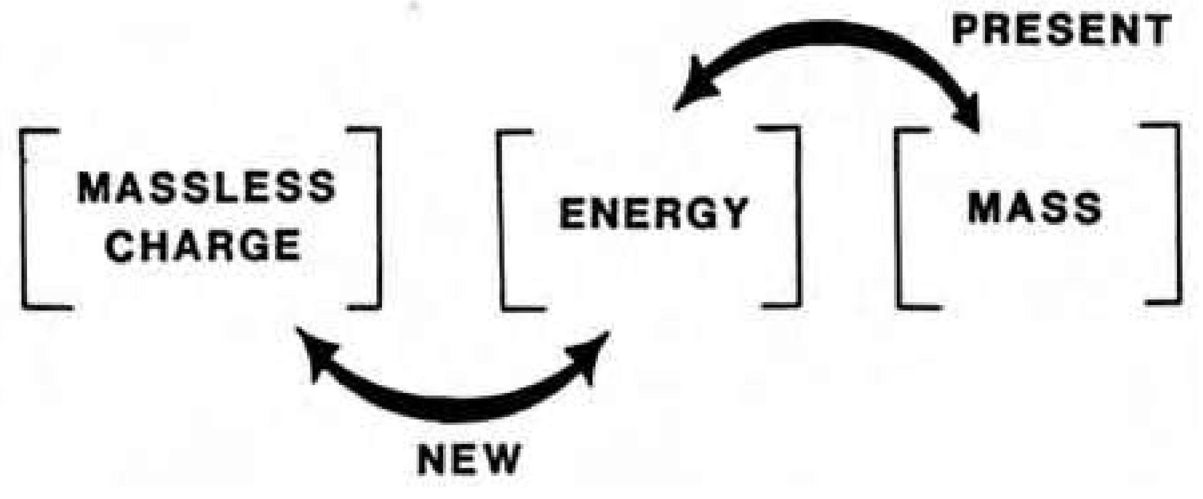
PRESENT
MASSLESS CHARGE
ENERGY
MASS
NEW

Now we can put the search for free energy devices on a rigorous scientific basis.

As shown here, mass and energy are translatable one into the other. So are massless charge and energy.

And the universe is filled with fantastic amounts of anenergy; ultimately, that is what the universal vacuum IS.

Any one-way gating process from anenergy to energy yields "free energy" since the anenergy is automatically replenished by the universe.

A sort of Maxwell's demon, after all, is what we need. Only we need one which works with the virtual state, not observable molecules.

Fortunately nature has provided several Maxwell's demons of this sort, if we but recognize them and learn to use them.

Next Slide

The Tom Bearden Website

SLIDE 21.

SCALAR POTENTIAL CAN FORCE THE RELATIVISTIC CASE

- **FOR SCALAR POTENTIAL ENERGY OF APPRECIABLE SIZE RELATIVE TO A PARTICLE'S REST ENERGY.**
    - **NEWTONIAN MECHANICS AND**
    - **THE SCHRODINGER EQUATION**
- **MAY BE INADEQUATE**
    - **EVEN IF V/C IS SMALL**

Bloch & Crater, "Lorentz-invariant potentials and the non-relativistic limit," American Journal of Physics, Vol. 49, No. 1, Jan. 1981, pp. 67-75

First, let us now digress to point out that one does not have to have relativistic velocity to obtain relativistic effects.

One can use the common electrostatic scalar potential to drive the situation relativistic. On this slide we show just one of several references in the standard literature that address this fact.

Let's understand what we are saying.

Anything you get from a relativistic situation, you can get directly by cleverly applying electrostatic scalar potential.

You can get a change in the passage through time, you can get energy changes, mass changes, inertial resistance changes, etc.

You can bend, warp, and twist spacetime like a pretzel.

If you "wave" the scalar potential by simply varying it , you can create pure time waves. You can also produce pure inertial field waves, pure gravitational waves, etc.

The ability to engineer the curving of spacetime allows the direct ability to engineer physical reality itself.

Next Slide

The Tom Bearden Website

SLIDE 22.

SCALAR STRESS WAVE
[O-WAVE] [TESLA WAVE]

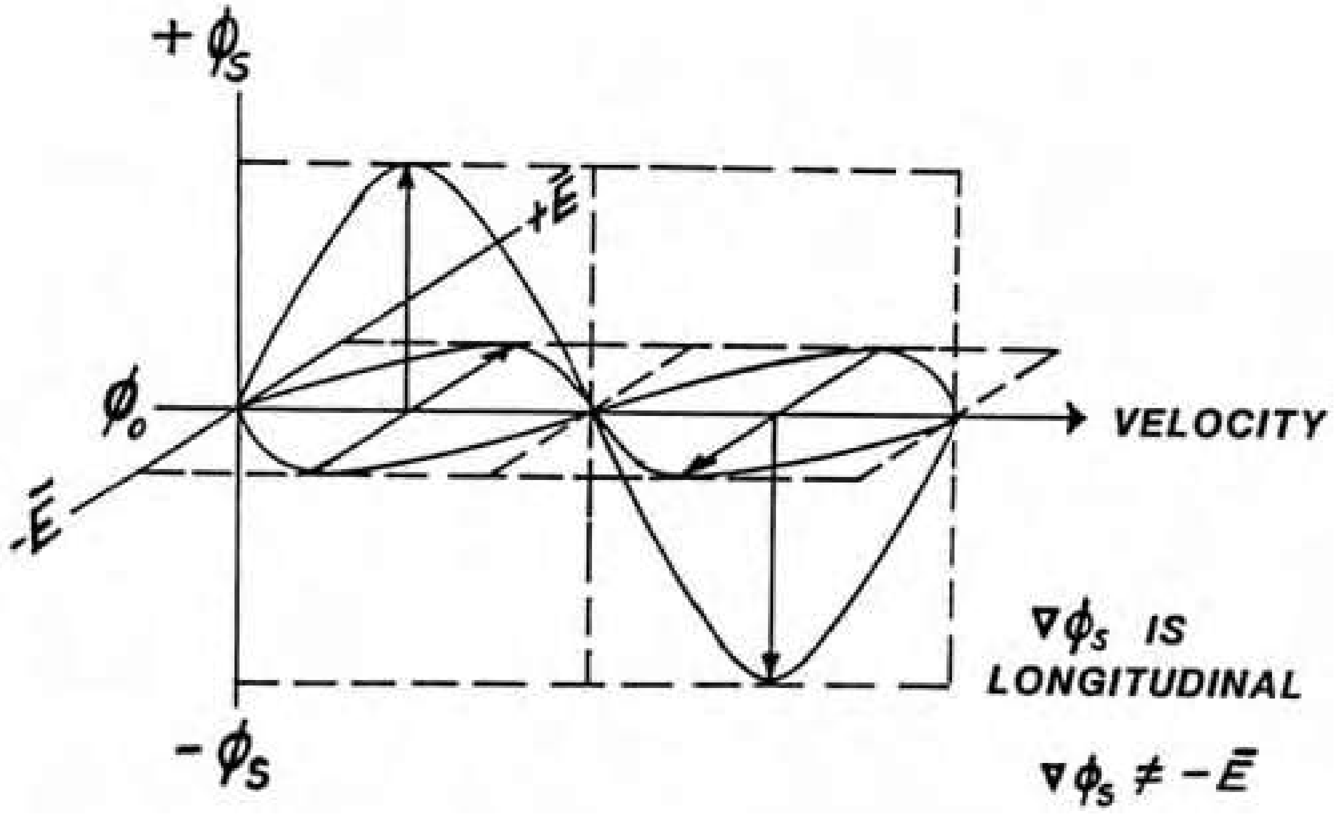
$+\phi_S$
$+\vec{E}$
ϕ_o
VELOCITY
$-\vec{E}$
$\nabla\phi_S$ IS LONGITUDINAL
$-\phi_S$
$\nabla\phi_S \neq -\vec{E}$

On this slide we now show a simple way to make a scalar, zero-vector wave -- the kind of wave originally discovered by Nikola Tesla.

It's simple. We just believe that a sum-zero vector substructure makes a scalar quantity, and we MAKE some scalars that way.

We also understand that a zero-gradient of a scalar is a zero vector, so that the scalar itself may be taken to be a zero vector.

The simplest explanation of this wave is as follows:

First, in physics we have two competing, mutually exclusive theories as to the nature of electromagnetic energy: the wave theory and the particle theory. Physicists argued for decades over these theories, for some experiments support one and some support the other. They never solved the problem; they just agreed to quit arguing. They formulated the "duality" principle to allow the saving of face to both sides.

Briefly, the duality principle implies that, whatever the nature of electromagnetic energy is before an interaction, in the interaction you can get it to act as a wave or as a particle. In other words, AS IT EXISTS, BEFORE THE INTERACTION, it is implicitly both particle and wave, joined together in some fashion, without being explicitly either one.

With the fourth law of logic, this becomes perfectly clear. With three-law Aristotlean logic, the problem is unresolvable.

Let us use this idea of "explicit duality without implicit duality" to analyze the wave shown on the slide.

First, from a wave aspect, the E-fields and the B-fields of the two waves do superpose and vectorially add. Since the waves are 180 degrees out of phase, the exterior resultant wave has a zero electric field and a zero magnetic field. Therefore it is a "zero-vector" wave, or "scalar" wave. It's a wave of pure stress in spacetime.

However, this scalar wave has a precisely determined substructure, consisting_of two ordinary sine waves, each of which comprises an ordinary E-H vector EM wave.

Now we apply the photon consideration (remember, before we interact with the wave, it must implicitly possess BOTH wave and particle natures combined, and we have so far only examined the implication of the wave nature.

The theory of photons' states that they are monocular critters. Photons pass right through other photons without interaction, in a linear situation. Therefore they can coexist without interaction, which is what we show here.

One photon, by the way, is one wavelength .

The photon theory requires that both substructure waves continue to exist as independent photons. Therefore we are assured that our substructure is intact.

However, notice that the totality of the two waves stresses spacetime. In other words, we have twice the stress on spacetime now as we would have from either wave separately.

This wave is therefore just a pure stress wave in spacetime itself.

This thing oscillates time, oscillates the relativity of the situation, and can affect energy, time flow rate, inertia, gravity, etc. aspects of an absorbing system.

Note that we have a rhythmic oscillation in phi (Ø), and we have a longitudinal stress wave, very similar to a sound wave. The MEDIUM for this wave is the virtual particle flux that identically comprises vacuum spacetime itself.

Next Slide

The Tom Bearden Website

SLIDE 23.

SCALAR Ö-WAVE PRODUCTION

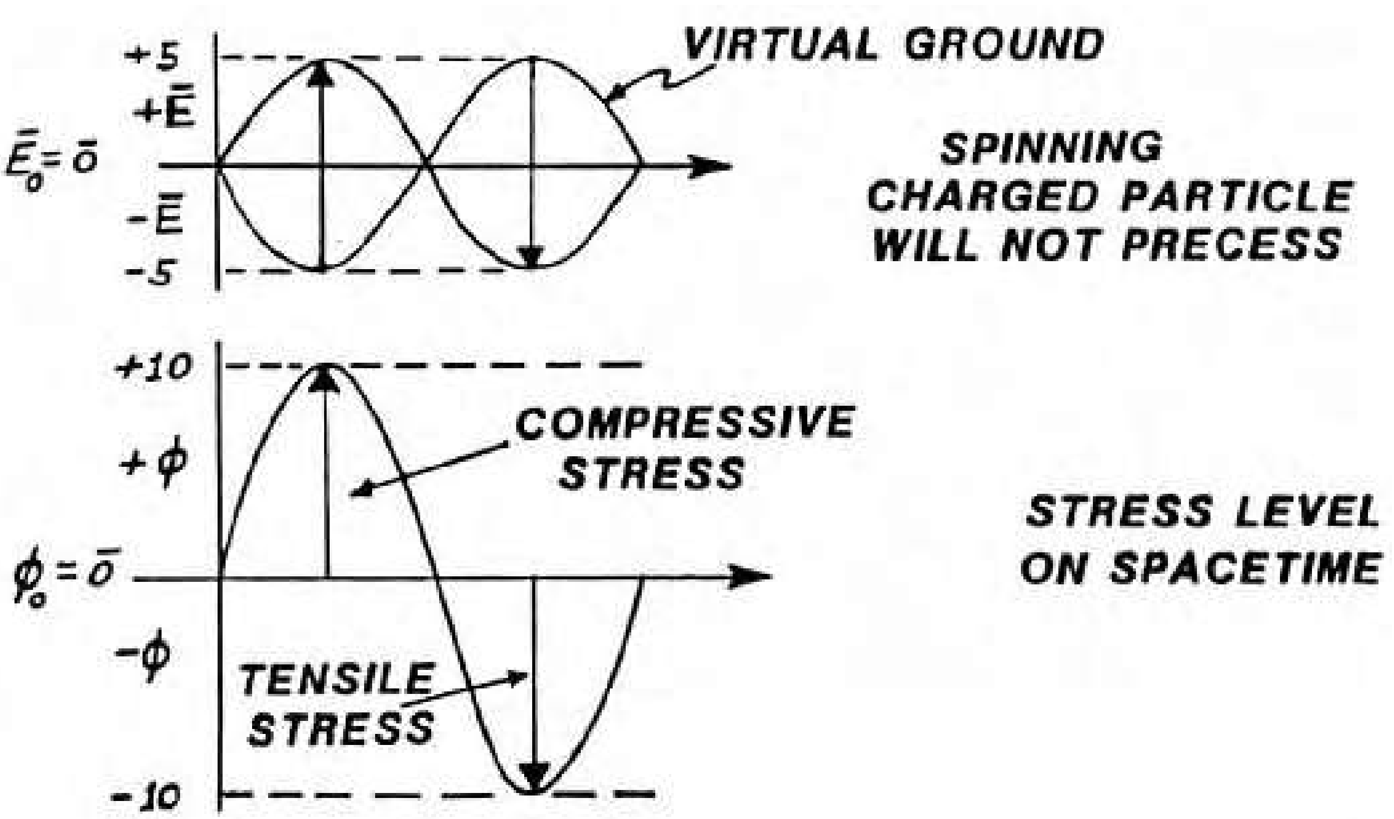
VIRTUAL GROUND
+5
+Ē
$\bar{E}_0 = \bar{0}$
-Ē
-5
SPINNING CHARGED PARTICLE WILL NOT PRECESS
+10
+φ
COMPRESSIVE STRESS
$\phi_0 = \bar{0}$
-φ
TENSILE STRESS
-10
STRESS LEVEL ON SPACETIME

On this slide we show how to regard the magnitude of the stress, by using one wave envelope as "virtual ground." In the top diagram, the original E-field stress of each wave has a magnitude of 5, and the composite stress wave now has a magnitude of 10, in whatever units we choose to express them.

We show in the bottom diagram that during one half cycle we have compressive stress in the virtual particle vacuum flux, and during the second half cycle we have tensile stress.

This shows the wave is like a sound wave in the gaseous molecules of the air.

However, this wave has one difference. It also oscillates time, and thus has at least one additional degree of freedom, compared to ordinary EM waves.

In fact, this wave can be made n-dimensional and hyperspatial.

As a first order approximation, we can treat such a wave in
a spatial fashion, if the wave is not too great in magnitude and the relativistic oscillation of time and inertia is not too large.

Next Slide

[The Tom Bearden Website](http://www.cheniere.org)

SLIDE 24.

A SIMPLIFIED CHARGED PARTICLE

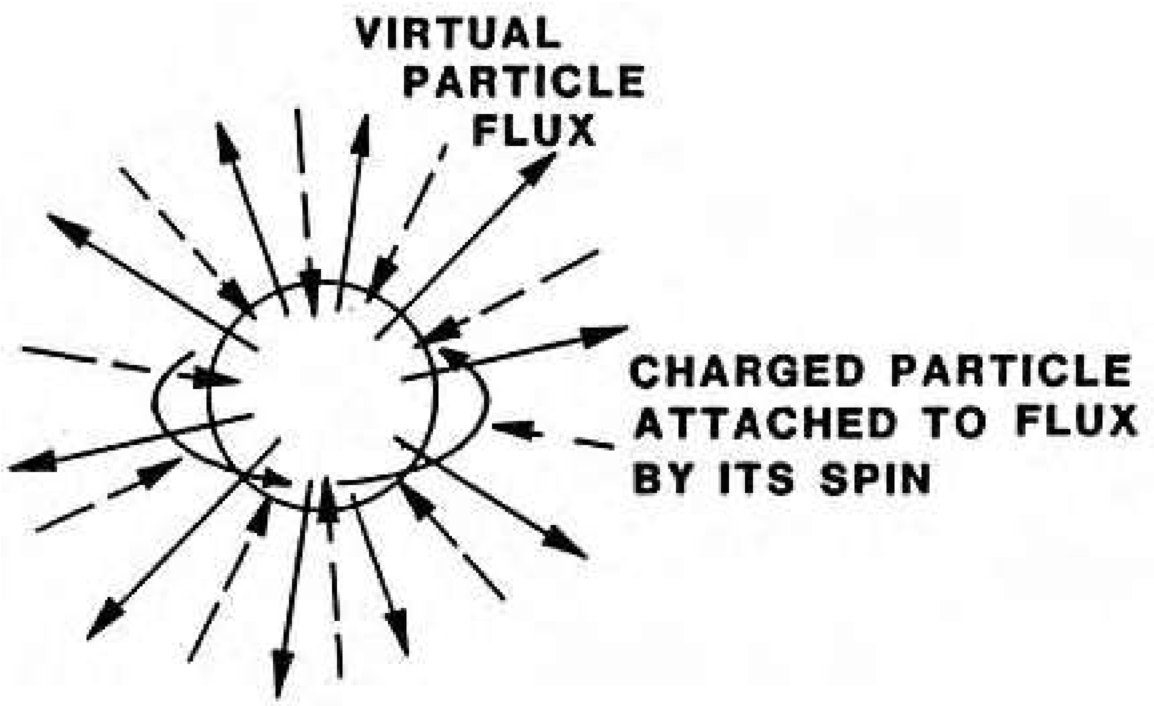

(THINK OF THE FLUX IN AND OUT, EXCEPT GRADIENTS ARE SPIRAL-LIKE, DUE TO PARTICLE'S SPIN)

We now digress slightly to pick up some more concepts.

On this slide, we show a simplified physical model of a charged particle. According to present particle physics, we model it as an observable, spinning bare mass in the middle, connected to a spray of virtual particles on and off the mass in all directions.

It is the spin of the mass that "attaches" the spray of virtual particles to and from itself.

Note that, because of the spin, the gradient lines of flux will actually be spiral or vortex in nature.

Also note that, when EM theory was founded, no one knew that an electron was spinning. Electron spin was therefore omitted from the foundations of present electromagnetic theory, and compensated for later by strictly ad hoc measures.

Next Slide

The Tom Bearden Website

SLIDE 25.

A CHARGED PARTICLE IS A SPECIAL KIND OF "SPRAY NOZZLE"

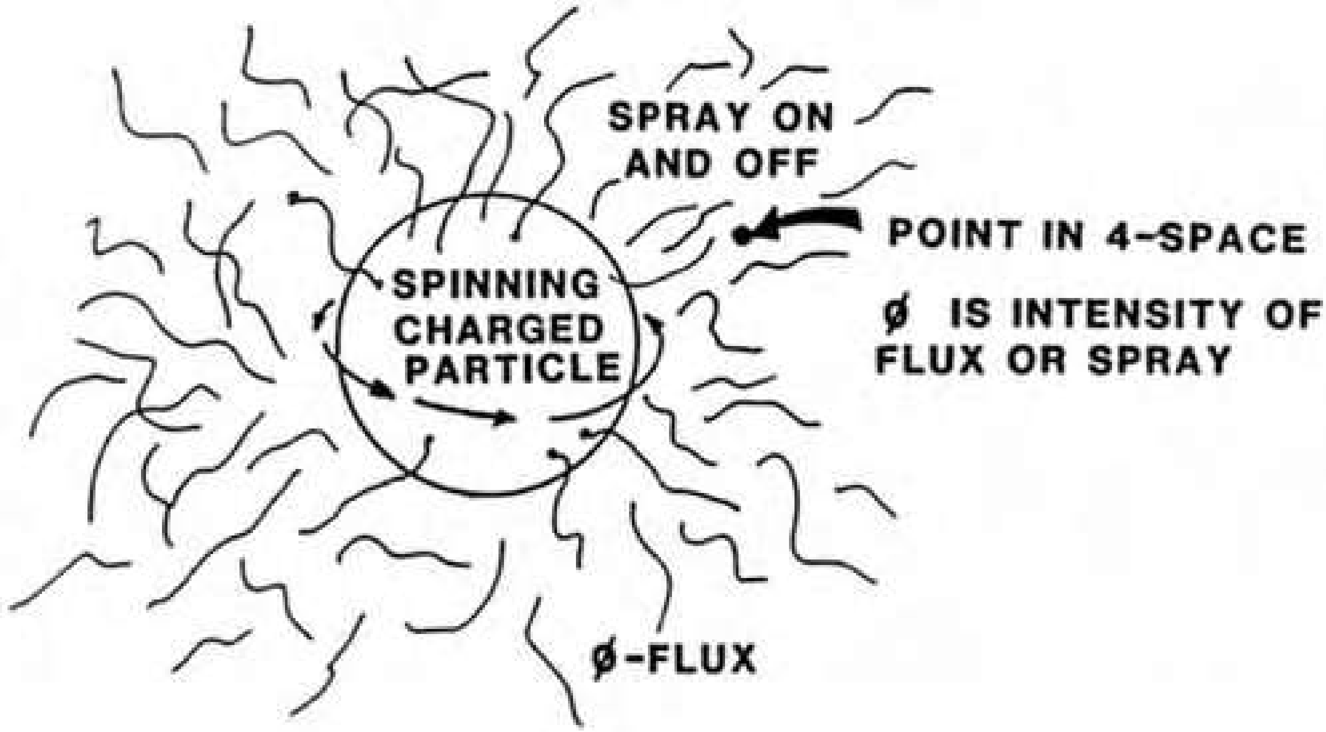
SPRAY ON
AND OFF
POINT IN 4-SPACE
Ø IS INTENSITY OF
FLUX OR SPRAY
SPINNING
CHARGED
PARTICLE
Ø-FLUX

(SPRAY IS THROUGH 4 DIMENSIONS, NOT JUST 3)

Here we show a very important concept, and a very important fact of nature.

A charged particle may be idealized as a special kind of spray nozzle.

The 3-dimensional particle of mass is spraying out virtual particle fluid, and sucking in virtual particle fluid, in four or more dimensions.

We never have to furnish or replenish the spray. Nature provides that. And it's inexhaustible. It comes from the entire virtual particle flux vacuum of the universe. It's already connected firmly to matter, by its connection to the spinning charged particle of mass.

I repeat and strongly stress: This tremendous, continuous spray of virtual particles is absolutely free, and absolutely furnished by the universe.

We don't have to create the basic pumps for the fluid. We just move the spray nozzles around, by moving the spinning charged masses around.

We never furnish any ultimate energy to or in our present EM circuits. We just pump around the spray nozzles, gating the spray furnished by the universe.

Next Slide

The Tom Bearden Website

SLIDE 26.

AN OBSERVABLE CHARGED PARTICLE IS DIRECTLY CONNECTED TO VIRTUAL PARTICLE FLUX VACUUM

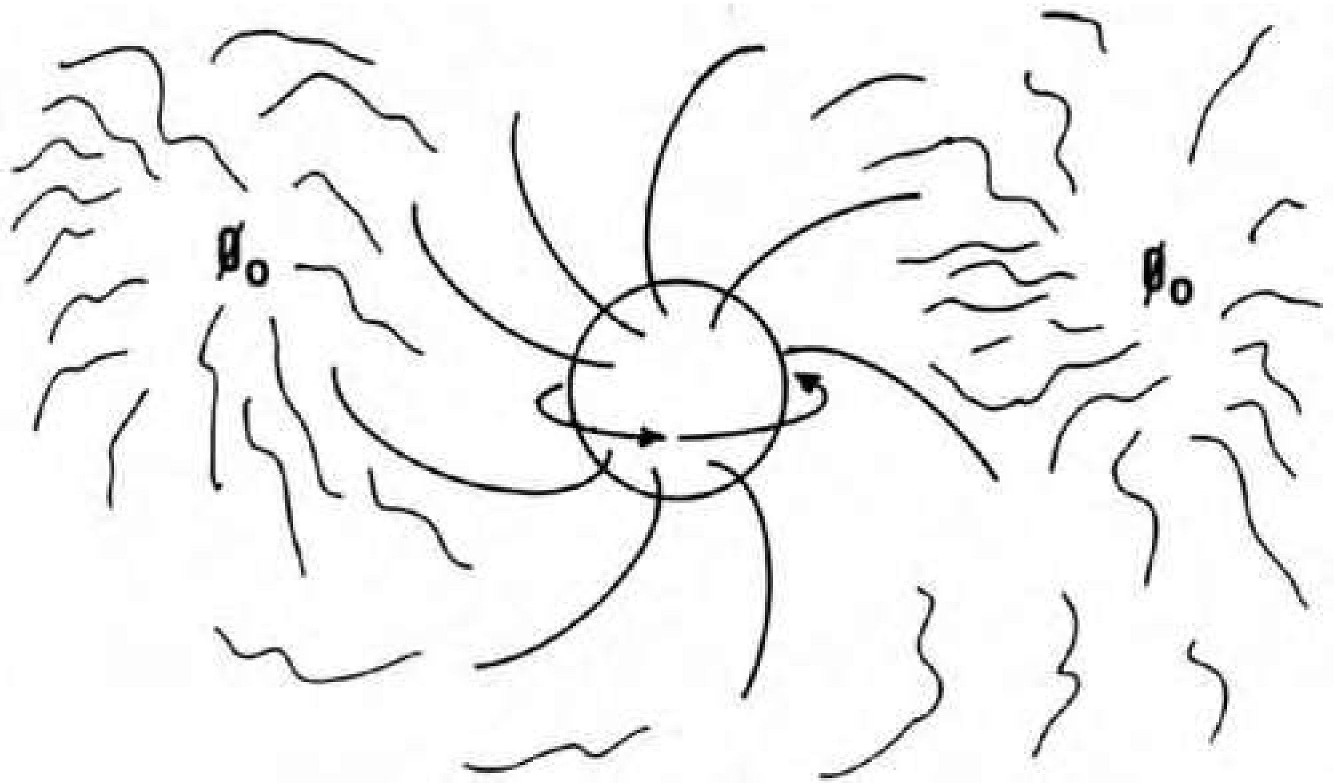

On this slide we accent that our 3-dimensional particle of mass is directly connected by its spin to the vacuum ether itself. To the virtual particle flux that identically IS the vacuum spacetime.

Interestingly, relativity states that the mass itself is nothing but a "kink" or curvature in spacetime .

The fact that the surface of the particle is spinning at right angles to the virtual particle vacuum flux "drags" part of the flux along with it, spiraling it, and certainly "kinking" or "bending" the vacuum flux .

That's the MECHANISM for the bending of spacetime. And that's what MAKES the particle of mass in the first place, and reduces its dimensionality to three.

Also note that the vacuum IS phi-sub-zero, and that is never a zero value. This is contrary to present EM theory, which erroneously assumes phi-sub-zero -- the charge of vacuum -- to be zero or non-existent.

Next Slide

The Tom Bearden Website

SLIDE 27.

SPIN COUPLES VACUUM TO PARTICLE

- **SPHERICAL ROTATION IS THE KEY**
- **A PARTICLE CAN BE MODELED AS A SPHERICALLY ROTATING VORTEX OF SPACETIME**
- **ITS MASS IS DUE TO ITS SPIN**
- **THE SPINNING OBJECT IS CONTINUALLY CONNECTED TO ITS ENVIRONMENT**
- **VERY HIGH VALUES OF ELECTROSTATIC POTENTIAL CAN INDUCE RELATIVISTIC CONDITIONS**
    - **CHANGE RATE OF FLOW OF TIME**
    - **EVEN THOUGH VELOCITY IS NONRELATIVISTIC**

E.P.BATTEY-PRATT AND T.J.RACEY
"GEOMETRIC MODEL FOR FUNDAMENTAL PARTICLES,"
INTL. J. OF PHYS. 19, NO. 6, 437-475, 1980

In fact, the bits and pieces of many of the new concepts are strewed all through the scientific literature. Here we show one example, establishing a good model that spin of a particle couples the particle to the vacuum .

This paper is by E.P. Battey-Pratt and T.J. Racey, and was . published in the International Journal of Physics in 1980.

Note again that very high values of electrostatic potential
can induce relativistic conditions, even though relativistic velocity is not present.

Next Slide

The Tom Bearden Website

SLIDE 28.

ELECTRON PRECESSION (SIMPLIFIED)

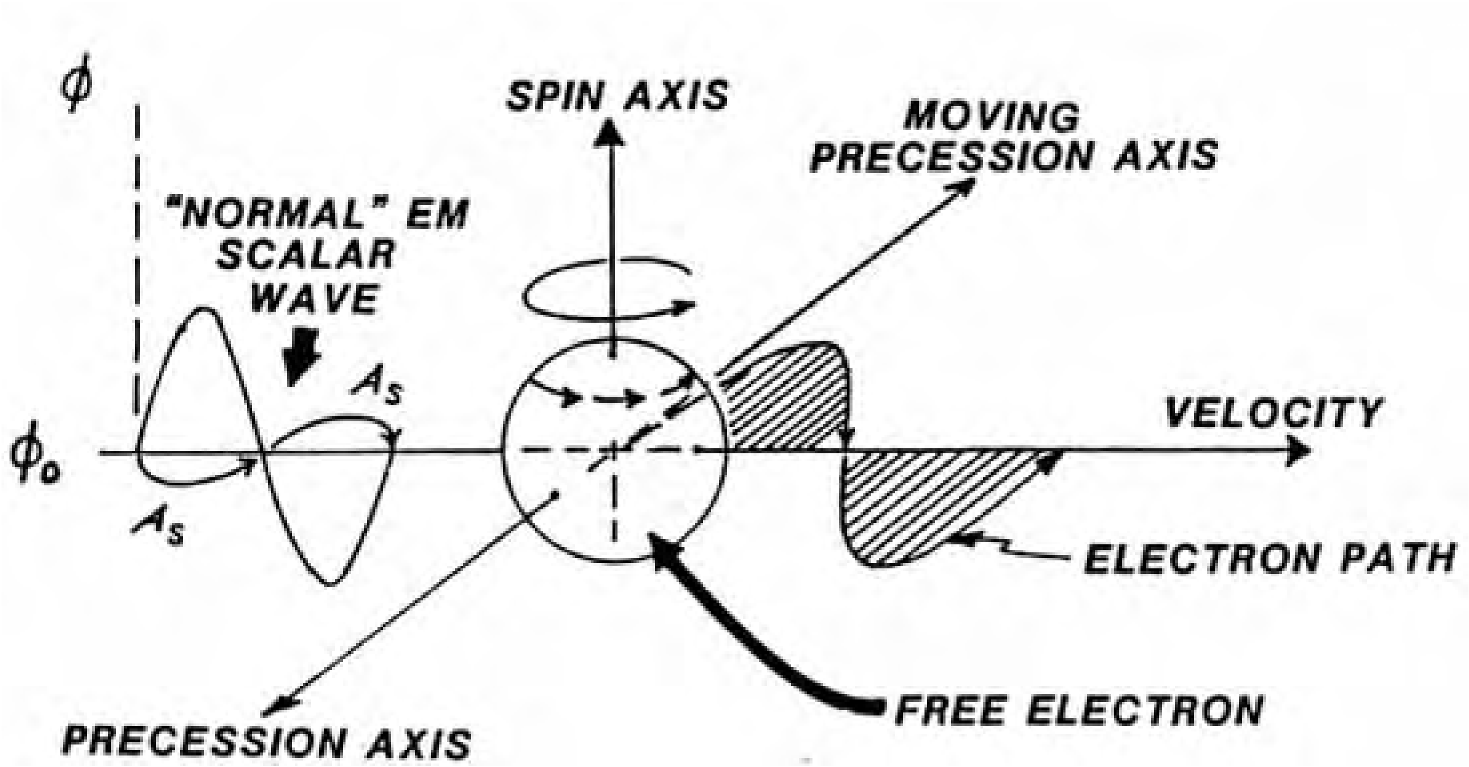

ϕ
SPIN AXIS
MOVING PRECESSION AXIS
"NORMAL" EM SCALAR WAVE
A_s
VELOCITY
ϕ_o
A_s
ELECTRON PATH
FREE ELECTRON
PRECESSION AXIS

On this slide I show the effects of precession of an electron when it encounters a longitudinal scalar wave that contains "spin vortex holes" for the electrons to fall into and mesh its spin with.

The basic idea here -- that electron precession accounts for the Hertzian waves in the electron gas in a transmitting antenna and in a receiving antenna, came from my close colleague and friend, Frank Golden, and I am most happy to give him full credit for this important insight.

As we have previously stated, transverse (force) waves cannot exist in vacuum in the absence of mass. Hertz waves therefore cannot exist in vacuum, just as Tesla stated.

Yet we know that Hertz waves exist in the electron gas in our transmitting antennas and in the electron gas in our receiving antennas. How then do we get Hertz waves here, if only longitudinal waves can exist in the vacuum in between?

Electron precession is the key.

We never measure what's happening in vacuum with our instruments. Instead, almost always we measure what is happening to and in the electron gas in our antenna or probe and feeding current to the instrument.

Here we show a "normal" EM vacuum wave -- which is a longitudinal wave containing spin vortexes from the electrons that generated it -- approaching and striking a spinning charged electron. As the peak and trough of the wave passes, it is as if we had a force pressing against the spinning electron, first along the line of wave travel, and then antiparallel to the wave travel. *(that is, "explaining" it in present concepts.)*

The electron acts as a little gyroscope, and precesses laterally, first in one direction and then the other.

Therefore the wave recovered in the electron gas in our receiving antenna or instrument probe is a transverse Hertzian matter wave.

Hertz waves are always matter vector waves.

Vacuum EM waves are always nonmaterial longitudinal scalar waves.

Longitudinal scalar waves in vacuum normally contain many spinning vortex "holes" of flux, created from the spinning electrons which launched the wave and stayed behind in the transmitting antenna. This kind of longitudinal wave is directly detectable by a normal free electron charged gas in a receiving antenna or probe. It also directly interacts with free electrons in a conducting metal shield, and so is shielded by Faraday cages.

On the other hand, our zero-vector longitudinal wave, made
by opposing waves, contains opposing spin holes which annihilate or cancel each other.

In the absence of spin holes, the longitudinal wave will not mesh with spinning electrons in a conductor, and so it is not detectable in the normal fashion. It also does not interact with free electrons in a conducting metal shield, so it readily penetrates Faraday cages.

An easy way to see that electrons do not interact with that substructured longitudinal zero-vector wave is to visualize both substructure component waves interacting on the electron simultaneously, pushing in opposite directions equally. In that case the electron tries to precess in both directions, equally, and so it does not precess in either direction. Therefore it does not "detect" the passing wave.

The wave without "golf ball holes", however, is detectable by any circuit having high nonlinearity actions occurring in it. Such highly nonlinear dynamic areas act to provide a phase shifting between the composite substructure waves. This phase shift results in violation of the sum-zero condition, producing a "normal" EM wave which deposits energy in the out-of-phase area.

From the spin vortex "golf ball hole" concept, the out-of-phase condition means that now we have an alternating preponderance of spin holes, spinning first in one direction and then in the other. Thus the electrons in the nonlinear, phase shift area are hooked and oscillated (precessed) to and fro, producing energy.

Solid state, highly doped transistors are particularly vulnerable to this effect, as are gas discharge tubes, spark and cascade ion discharges, plasmas, etc.

Next Slide

The Tom Bearden Website

SLIDE 29.

VACUUM/SPACETIME IS PURE VIRTUAL PARTICLE FLUX ($Ø_o$)

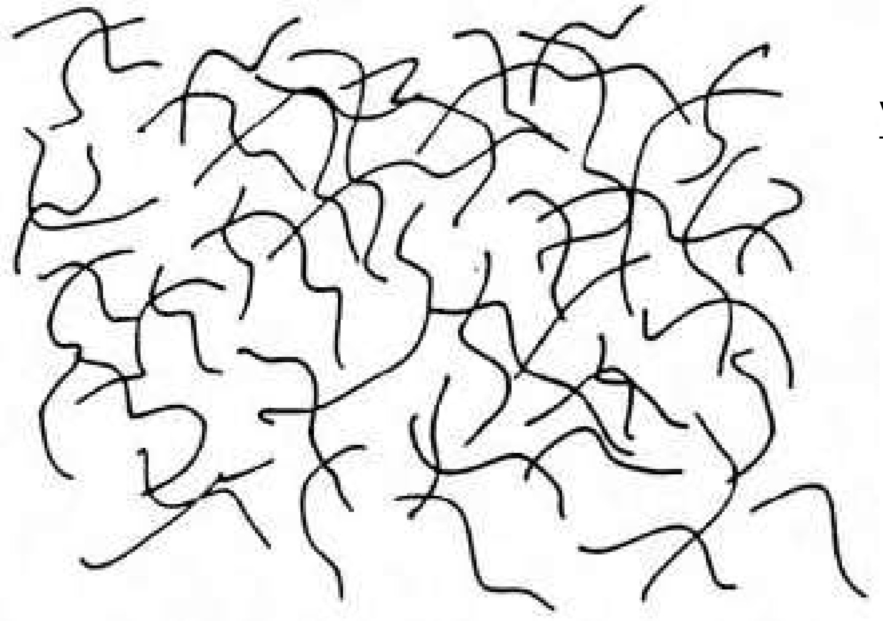

VACUUM IS IDENTICAL TO:

- **VIRTUAL PARTICLE FLUX**
- **SPACETIME**
- **ANENERGY**
- **$Ø_0$**
- **MASSLESS CHARGE**

VACUUM IS:

- **MADE OF UNQUANTIZED ACTION**
- **WITHOUT DEFINITE LENGTH INTERVALS**
- **WITHOUT DEFINITE TIME INTERVALS**
- **N-DIMENSIONAL (UNFIXED)**

We pause here to again summarize some important things we now know about vacuum.

Vacuum IS IDENTICAL TO virtual particle flux, spacetime, anenergy, Phi-subzero, massless charge, unquantized action, and infolded energy, momentum, length, time, power, velocity, acceleration, etc.

Vacuum is made of unquantized action, dimensionally.

It is without definite length or time, as Einstein stated.

It is n-dimensional, where n is always four or greater.

It is not an emptiness filled with things. Rather, it is identically a plenum filled with things which are totally unseparated, unfixed, and undetermined.

From vacuum: mass, energy, time, frames, and all the rest of physical reality are dynamically made.

Vacuum is the ether. The ether has never been eliminated from physics. Ether just has become much more subtle, and physicists have used other names for it.

Next Slide

The Tom Bearden Website

SLIDE 30.

VACUUM IS:

- **SPACETIME (L^nT, WHERE $n \geq 3$)**
- **CHARGE (MASSLESS)**
- **ELECTROSTATIC SCALAR POTENTIAL $\text{Ø}_0 \neq 0$ ERROR IN PRESENT THEORY**
- **BROKEN BITS OF ENERGY (SUBQUANTAL)**
- **PURE VIRTUAL PARTICLE FLUX**
- **O-WAVE FLUX**
- **MULTILEVEL, STRUCTURED, PATTERNED**
- **A VIRTUAL PLENUM**
- **AN OBSERVABLE EMPTINESS**

Again, we list some of the things that vacuum identically is.

It is spacetime, massless charge, electrostatic scalar potential, broken bits (subquantal) of energy, pure virtual particle flux, zero-vector wave flux, multilevel, structured, patterned, a virtual plenum, and an observable emptiness.

It is all things and contains all things in potential state.

It is not, in that it is not observable. But from it comes all observables.

It is both ordered and disordered, simultaneously.

It involves four-law logic, not just three-law Aristotlean logic.

Next Slide

The Tom Bearden Website

SLIDE 31.

GRADIENT OF ELECTROSTATIC POTENTIAL

IN VACUUM

$$\nabla \text{Ø} \neq \ddot{E}$$

ON AND OF A MOVING CHARGED PARTICLE

$$\nabla \text{Ø} = \ddot{E}$$

Again I accent that, in vacuum in the absence of a charged spinning particle of observable mass, del-phi does not yield an Ë-field, and the conventional equations of EM are wrong in that respect.

In the presence of a spinning charged particle, in a del-phi which contains electron-vortex-holes to mesh with, the charged particle attaches itself to the moving del-phi flux gradient, moving itself with the river. This produces an Ë field. The Ë-field CONSISTS OF the smeared electron, it does not CAUSE THE MOVEMENT of the electron. It is an effect, not a cause. The conventional equation for del-phi equals E is correct for matter waves in electron gases; it is not correct in vacuum itself.

Next Slide

The Tom Bearden Website

SLIDE 32.

IN A ∇Ø, A CHARGED PARTICLE MOVES ITSELF

∇Ø>0 IMPLIES A MOVING RIVER, AND THE CHARGED PARTICLE IS "HOOKED TO" THE RIVER

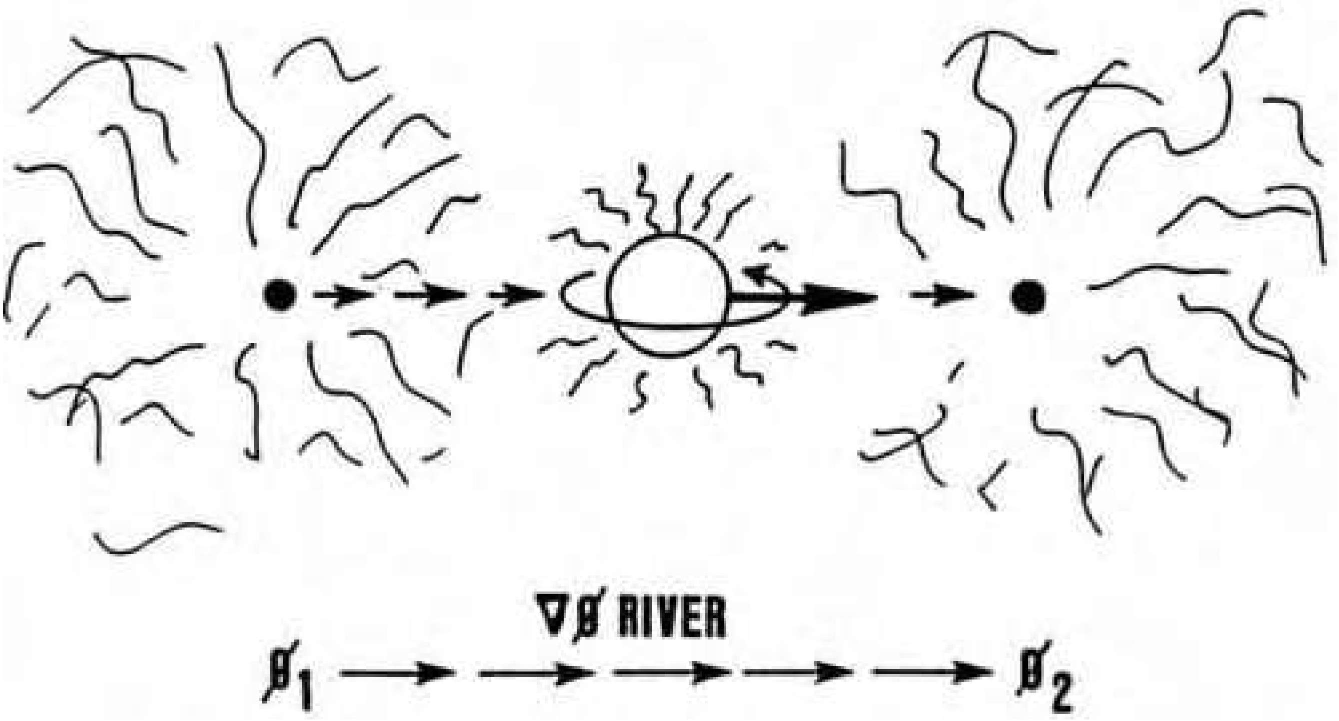

We show clearly on this slide that a spinning charged particle, when it hooks to a spin-hole in a del-phi river, MOVES ITSELF.

Note that a del-phi implies a "pressure difference" in phi-flux between two points. Therefore there is a moving river -- a gradient -- of phi-dot, moving from the higher pressure to the lower.

If a charged particle is attached to this moving river, it flows along with the river. It is just like putting a boat into a moving stream of water. If the boat "latches" to the moving water, it moves with the river.

Electron spin and spin vortexes in the del-phi mesh, to accomplish the "hooking" or "latching."

Once hooked into a moving river, the electron moves itself. One does not have to furnish work to move it.

If the river is accelerating, the electron will accelerate itself.

By cleverly gating such rivers into and onto sources of free electrons, one can directly produce free energy.

The extra energy comes from converting anenergy (massless charge, vacuum flux) to energy of a moving mass.

Next Slide

The Tom Bearden Website

SLIDE 33.

ELECTROMAGNETIC WAVES IN ELECTRON GAS IN A CONDUCTOR

- **VECTOR WAVES**
- **CONTAIN E, B FIELDS**
- **GENERATED BY SPIN COUPLING OF ELECTRONS WITH SCALAR WAVE "SPIN-HOLES"**
- **WAVES OF FORCE**
- **HERTZIAN WAVES**
- **TRANSVERSE WAVES**
- **WAVES OF ELECTRON PRECESSION**

Here we pause to summarize the characteristics of electromagnetic waves in the electron gas of a conductor.

First, these are vector EM waves in the accepted sense.

They contain E and B fields.

They are generated by spin coupling of electrons with spin-holes in a scalar wave delta-phi river. The electrons move themselves under that condition, producing work and forces.

They are waves of force or force field variations.

They are Hertzian waves.

They are transverse oscillatory, not longitudinal.

They are matter waves.

They are waves of electron precession.

Next Slide

The Tom Bearden Website

SLIDE 34.

ELECTROMAGNETIC WAVES IN VACUUM

- **SCALAR LONGITUDINAL WAVES**
- **VECTOR ZERO [Ö] WAVES**
- **WITHOUT Ë, B FIELDS**
- **INTERNALLY STRUCTURED**
- **PATTERNED**
- **MAY CONTAIN ELECTRON "SPIN-HOLES"**

Here we summarize the characteristics of electromagnetic waves in vacuum.

They are scalar longitudinal waves of alternate compression and rarefaction of the vacuum virtual particle flux.

That is, they are waves of electrostatic potential.

They are zero-vector waves.

They are internally structured and patterned.

They usually contain electron "spin holes" unless made in a fashion so as to make opposing spin holes that cancel each other.

Since they are pure phi-waves, they need not be limited in velocity to the speed of light.

They are hyperspatial waves.

They are waves in virtual state itself.

Next Slide

The Tom Bearden Website

SLIDE 35.

DETECTION OF TRANSVERSE AND LONGITUDINAL WAVES

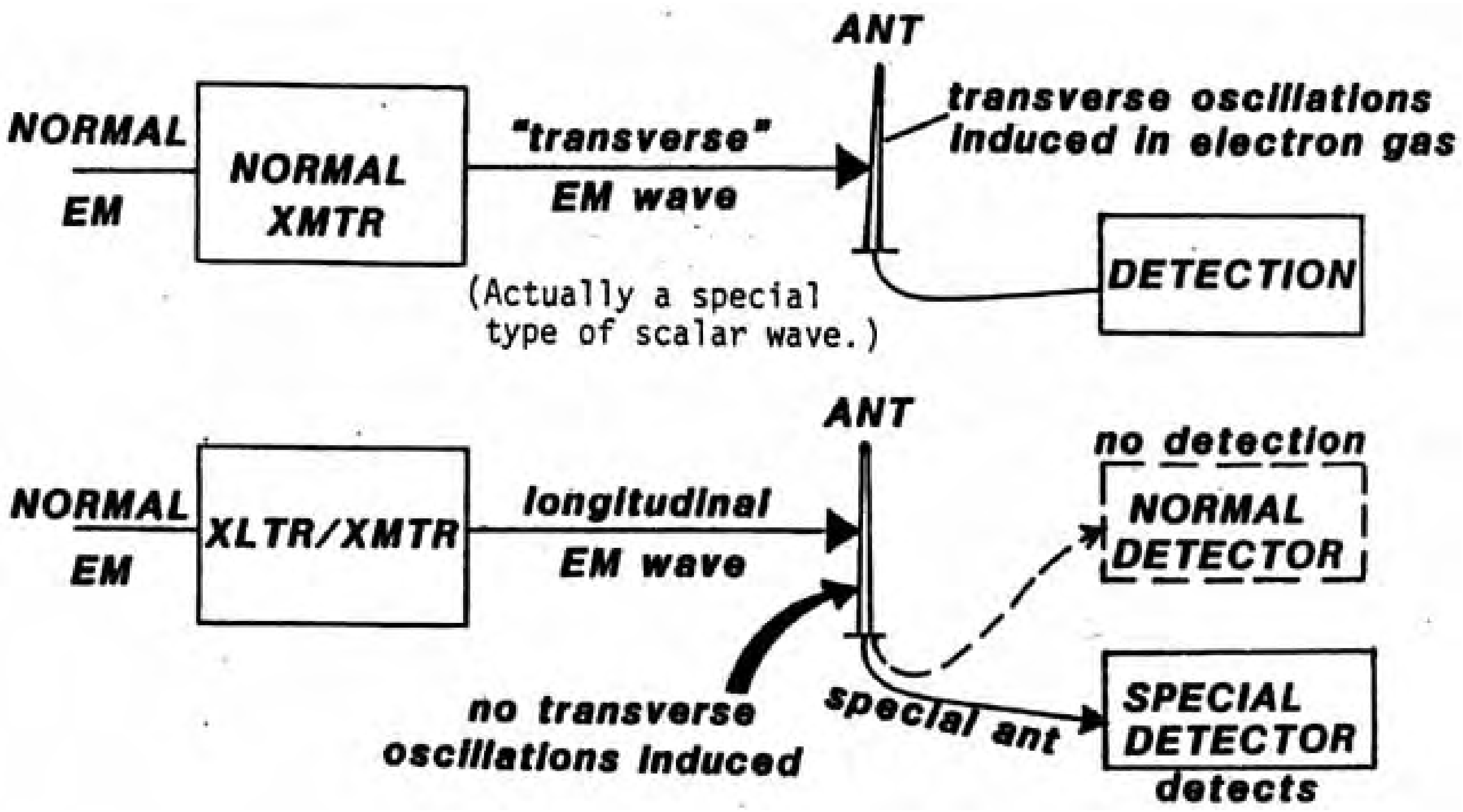
ANT
NORMAL
EM
NORMAL
XMTR
"transverse"
EM wave
transverse oscillations
induced in electron gas
DETECTION
(Actually a special
type of scalar wave.)
ANT
NORMAL
EM
XLTR/XMTR
longitudinal
EM wave
no detection
NORMAL
DETECTOR
no transverse
oscillations induced
special ant
SPECIAL
DETECTOR
detects

In the top drawing on this slide, we show a normal transmitter putting out a normal EM wave, which received in a normal antenna/receiver. What we call a "transverse" wave rigorously exists only in the electron gas in the transmitting antenna and in the electron gas in the receiving antenna. Specifically, a longitudinal wave exists in the vacuum in between the two antennas. However, due to the method of production, the longitudinal wave contains spin-holes for electrons, so the electrons in the receiving antenna readily couple with the wave by falling into and meshing their spins with the spins of the spin-vortex holes. Electron precession produces electron gas waves that are transverse oscillatory, hence Hertzian waves in nature.

In the bottom drawing, we show a translator/

transmitter. In other words, we oppose ordinary EM waves in a sum-zero substructure, deliberately producing longitudinal EM waves in vacuum with the spin-hole vortexes canceling each other. This type of wave does not "hook" spinning electrons in the normal receiving antenna, and thus it is not detected. The normal antenna/receiver system never sees it at all.

However, by means of a special antenna which generates nonlinear phase shifts in the composite substructure waves, "hooking" holes are restored in an oscillatory nature. Electrons then hook with this output and detect the wave.

Note that it requires a two-stage detector, operated in an interferometer fashion, to detect this pure scalar wave.

Next Slide

The Tom Bearden Website

SLIDE 36.

CREATING A "RIVER" OF ANENERGY

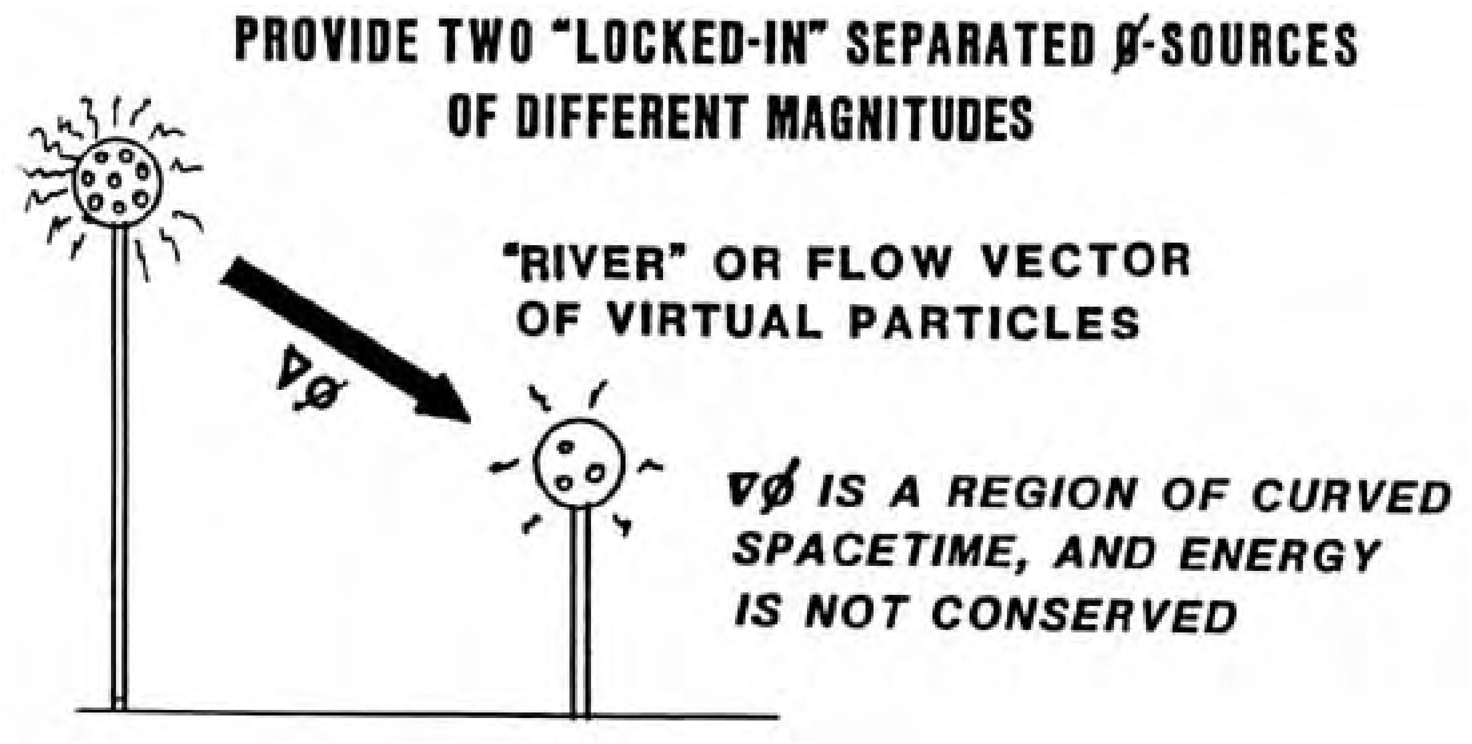

PROVIDE TWO "LOCKED-IN" SEPARATED Ø-SOURCES OF DIFFERENT MAGNITUDES

"RIVER" OR FLOW VECTOR OF VIRTUAL PARTICLES

∇Ø IS A REGION OF CURVED SPACETIME, AND ENERGY IS NOT CONSERVED

IN VACUUM, THIS IS NOT YET A VOLTAGE DROP (NO MASS FLOW), NOR IS THERE AN Ë-FIELD

Now if we wish to get free energy, we are going to have to provide a continuous anenergy river, and a means of tapping it to produce mass movement.

This slide shows one easy way to make an anenergy river.

We simply pump some electrons (spray nozzles) onto an elevated charged ball, and we LEAVE them there. A second ball is connected to ground. A higher phi -- that is, a denser spray -- is now in the vicinity of the elevated ball on the left. A lower phi -- that is, a less dense spray -- is in the vicinity of the lower ball on the right. Between the two balls, now, there is a gradient in phi, and a virtual flux flowing from the "higher virtual pressure" to the "lower virtual pressure."

This del-phi river does NOT constitute an Ë-field, as we
have previously pointed out.

The del-phi region, however, is definitely a region of
curved spacetime. As is well known from general relativity, in such a region energy need not be conserved.

Therefore it is entirely possible -- consistent with ordinary physics -- to violate conservation of energy in this del-phi river, if one believes general relativity. If one argues adamantly that conservation of energy cannot be violated under any circumstances, then one must throw out general relativity. Also, one must throw out most of particle physics, whose explanations presently involve virtual interactions, each of which violates the conservation of energy.

Next Slide

The Tom Bearden Website

SLIDE 37

A 2-WIRE SYSTEM IS CONSERVATIVE

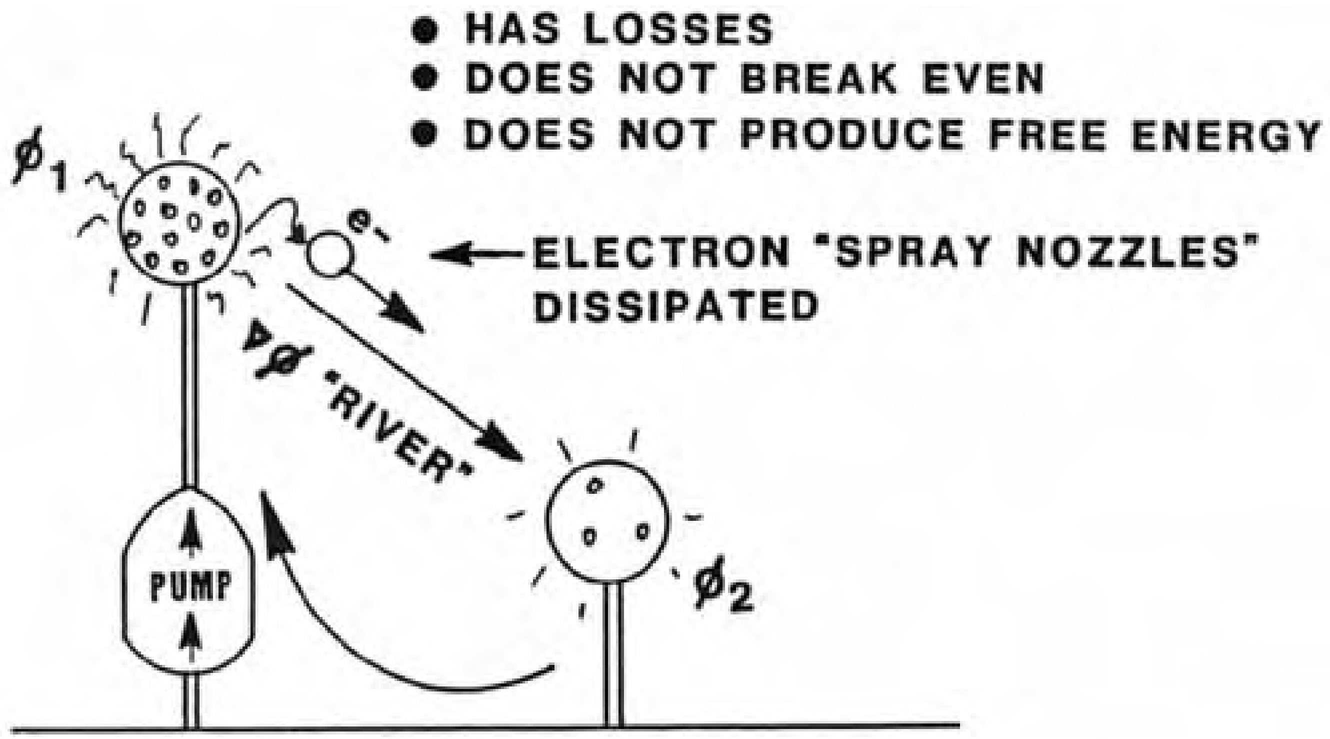

- HAS LOSSES
- DOES NOT BREAK EVEN
- DOES NOT PRODUCE FREE ENERGY

REPLENISH SPRAY NOZZLES BY PUMPING MORE ELECTRONS. (MOVING ELECTRON MASSES AGAINST A "SPRAY PILEUP" REQUIRES WORK.

Here we show that, as engineers, we've all been tricked into releasing our virtual rivers so that we would not discover free energy.

While I cannot prove this as yet, I tentatively suspect that J. P. Morgan and Thomas Edison are directly responsible for this occurring. Tesla utilized a single-wire system; we have all been taught to utilize only two-wire systems.

What we in effect do is connect a wire or circuit to our charged ball, allowing our hose nozzles to drain away.

In other words, we first do some work to pump hose nozzles (electrons) into and onto our elevated reservoir, so we will obtain a denser flux, and a flux river to utilize. Then we release the nozzles themselves, and let THEM move on the flux river, deplenishing our river. Of course we can get back as much energy as we pumped up there in the first place. But this two-wire system is at best theoretically even. Since real circuits have losses, we can never even quite break even. Thus we have to continue to pump hose nozzles, and let them bleed off to provide energy, in a less-than-100% fashion.

By foolishly releasing our hose nozzles, we guarantee that we will never achieve free energy. And that's nice, because the people who control things, and for their own wealth and power depend on selling us energy at high prices, are guaranteed to stay in power and control.

One is never truly free unless one is energy independent, in a self-contained fashion.

Tesla nearly succeeded in giving us free or very cheap, inexhaustible energy -- and for that reason was quickly and dramatically suppressed.

Only three graduate schools of electrical engineering existed in the U.S. The only one of any real importance was at Columbia University.

Only one or two professors were necessary to "reach" to modify and subtly direct the entire foundations of the emerging electrical theory.

Professors were in general poor, and had no research funds unless some rich industrialist or philanthropist gave them a few dollars.

The setup then was certainly very ripe for Morgan and Edison to stop free energy, and suppress Tesla at the same time. They could have done so by influencing only one -- or at most two -- theoreticians at Columbia.

At any rate, the theory was developed in such fashion as to rule out the single wire circuit and the longitudinal wave. Free energy was suppressed, and the financial empires of Morgan and Edison stayed secure.

Next Slide

The Tom Bearden Website

SLIDE 38.

A HOLLOW CHARGED SPHERE

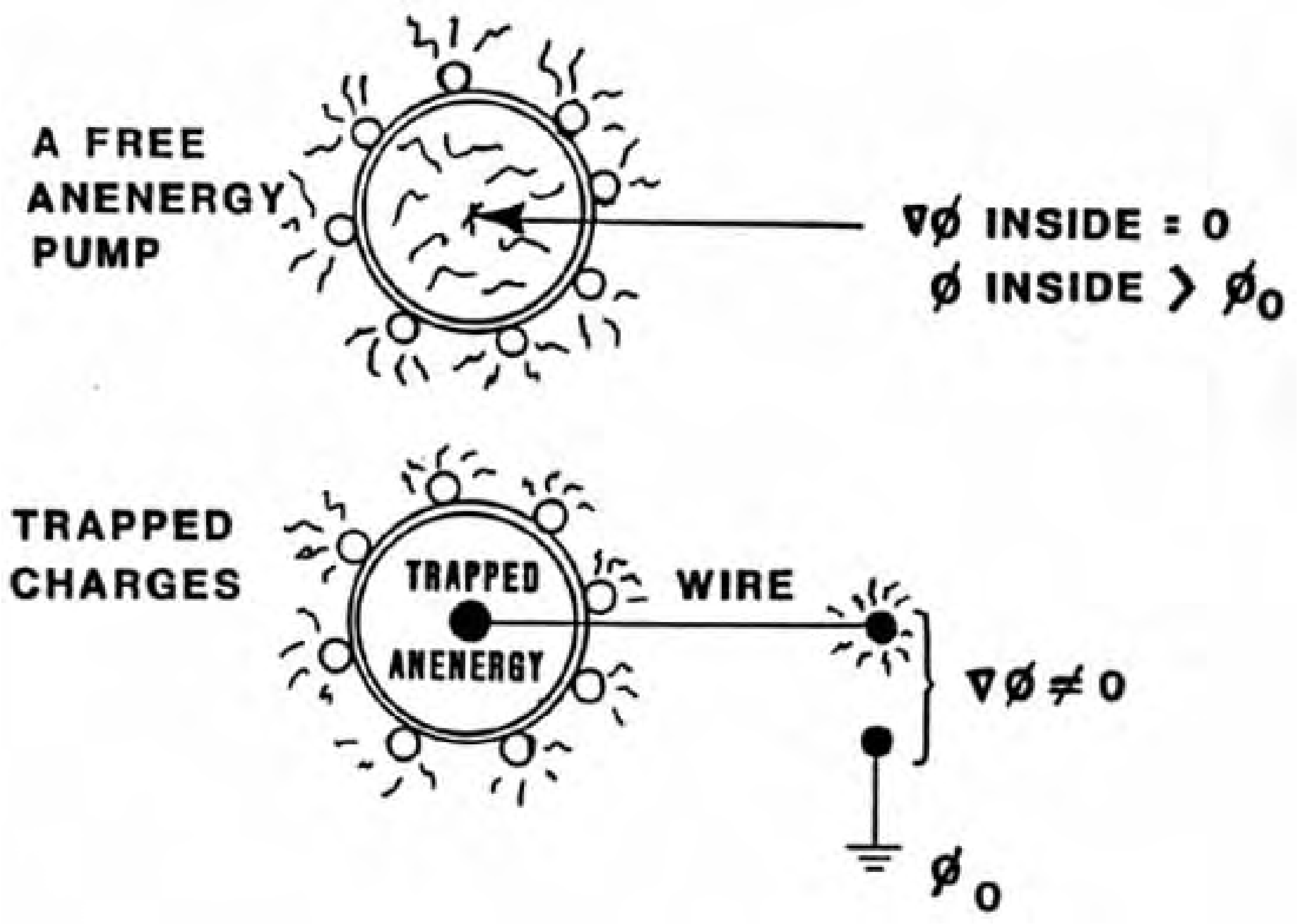
A FREE ANENERGY PUMP
∇ø INSIDE = 0
ø INSIDE > ø₀
TRAPPED CHARGES
TRAPPED ANENERGY
WIRE
∇ø ≠ 0
ø₀

Here we show another easy source of a higher vacuum charge, hence an easy source for a free river.

Simply charge up a sphere. As is well known, inside that sphere one has no del-phi. So one can introduce an electron therein and not produce an Ë-field .

However, since we've piled up additional spray nozzles, the density of the spray inside the sphere (as well as outside it) is certainly higher.

In other words, the inside of the sphere is also a source
for an increased del-phi river. In the bottom drawing, we drill
a tiny hole in the sphere, insulate a thin wire running into the inside of the sphere, and run the wire out to a terminal opposing another grounded terminal. Between the two terminals now there must exist a del-phi river.

And between those terminals there now exists a curvature of spacetime, hence the theoretical possibility to obtain free energy without violation of principles.

Next Slide

The Tom Bearden Website

SLIDE 39.

ONE-WAY GATE VALVE

(ANENERGY TO ENERGY)

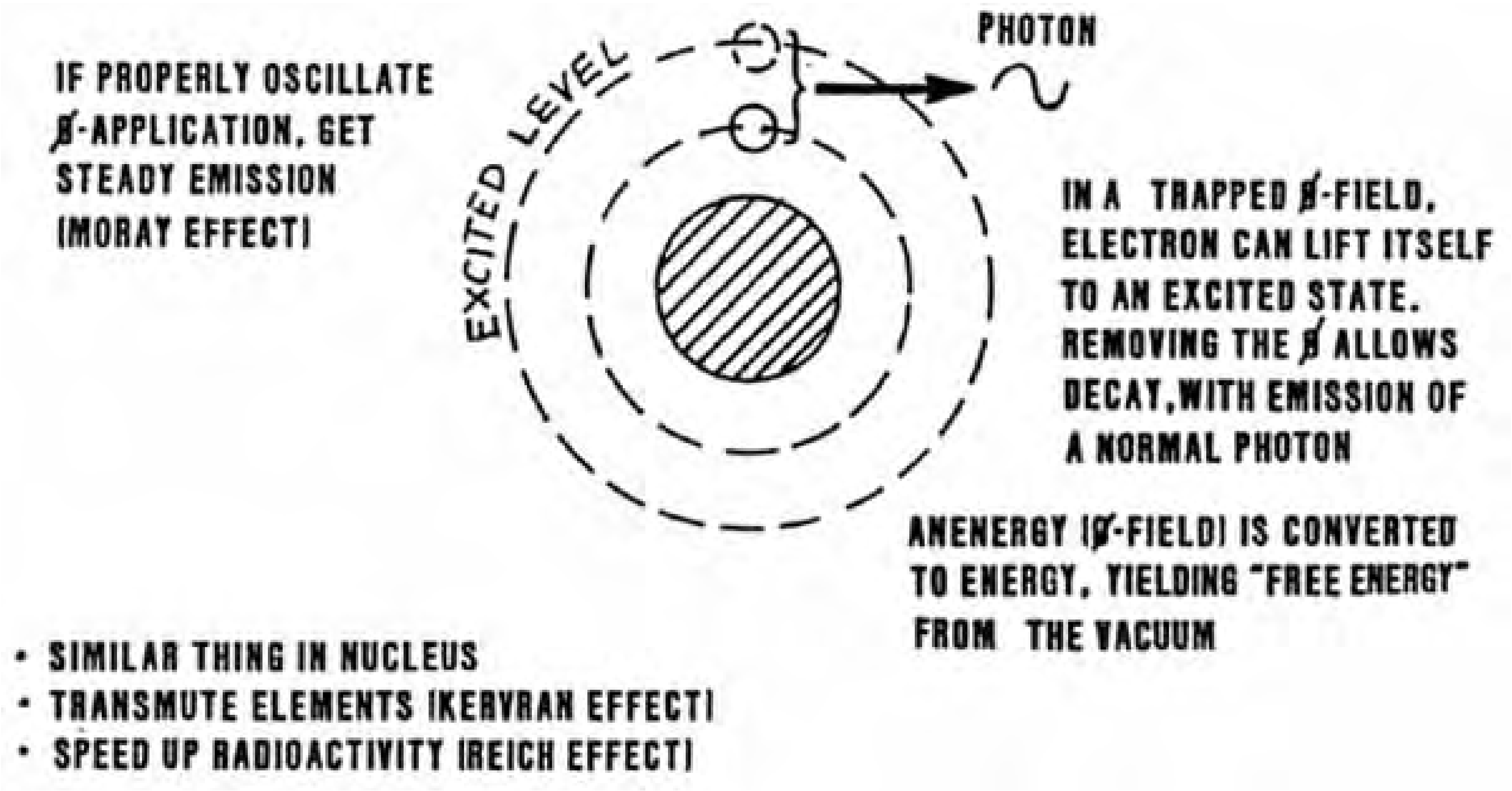

IF PROPERLY OSCILLATE
Ø-APPLICATION, GET
STEADY EMISSION
(MORAY EFFECT)

IN A TRAPPED Ø-FIELD,
ELECTRON CAN LIFT ITSELF
TO AN EXCITED STATE.
REMOVING THE Ø ALLOWS
DECAY, WITH EMISSION OF
A NORMAL PHOTON

ANENERGY (Ø-FIELD) IS CONVERTED
TO ENERGY, YIELDING "FREE ENERGY"
FROM THE VACUUM

- SIMILAR THING IN NUCLEUS
- TRANSMUTE ELEMENTS (KERVRAN EFFECT)
- SPEED UP RADIOACTIVITY (REICH EFFECT)

Here we show one way to provide a one-way gate valve from a del-phi river to energy released into the outside world.

By providing a properly tuned, oscillating phi wave, one can cause an orbital electron in an atom to raise itself to an excited state. Remember, in a del-phi, an electron that hooks to it raises itself.

We then switch out the phi application, and the excited electron decays in normal fashion, emitting a normal photon of energy.

By phase locking the processes, one produces a steady stream of photons.

By absorbing or directing the photons, one can produce heat, light, energy beams, etc.

By this means, one produces free energy by gating anenergy into energy.

And if one does not release spray nozzles in the phi-oscillations which feed the mechanism, one has an inexhaustible energy production.

A very similar thing can be done with the excited states of the nucleus. For example, by using a radioactive material with a predisposition to emit some particle -- say an electron -- the nucleus can be made to continually emit the particles WITHOUT ITSELF DECAYING INTO BYPRODUCTS. In other words, an electron is lifted from the Dirac sea of vacuum to instantly replace the electron emitted from the nucleus.

This provides a one-way gate valve from the Dirac sea to the external universe.

Theoretically, these mechanisms are no more mysterious than a hydraulic ram, which raises water higher than the stream that feeds the ram.

In the 20's and 30's, T. Henry Moray successfully built such gating devices for converting vacuum anenergy to external energy.

In the Kervran effect, living systems are able to directly transmute elements, even though they only possess feeble energies, by influencing the virtual particle interactions in the nucleus that furnish its binding energy.

Reich used an adaptation of a cloudbuster -- which emits scalar waves -- to drastically speed up radioactive decay of radioelements.

Both plants and animals transmute elements by similar processes.

Next Slide

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

SLIDE 40,

SIMPLE FREE ENERGY MOTOR

BATTERY RECHARGES ITSELF

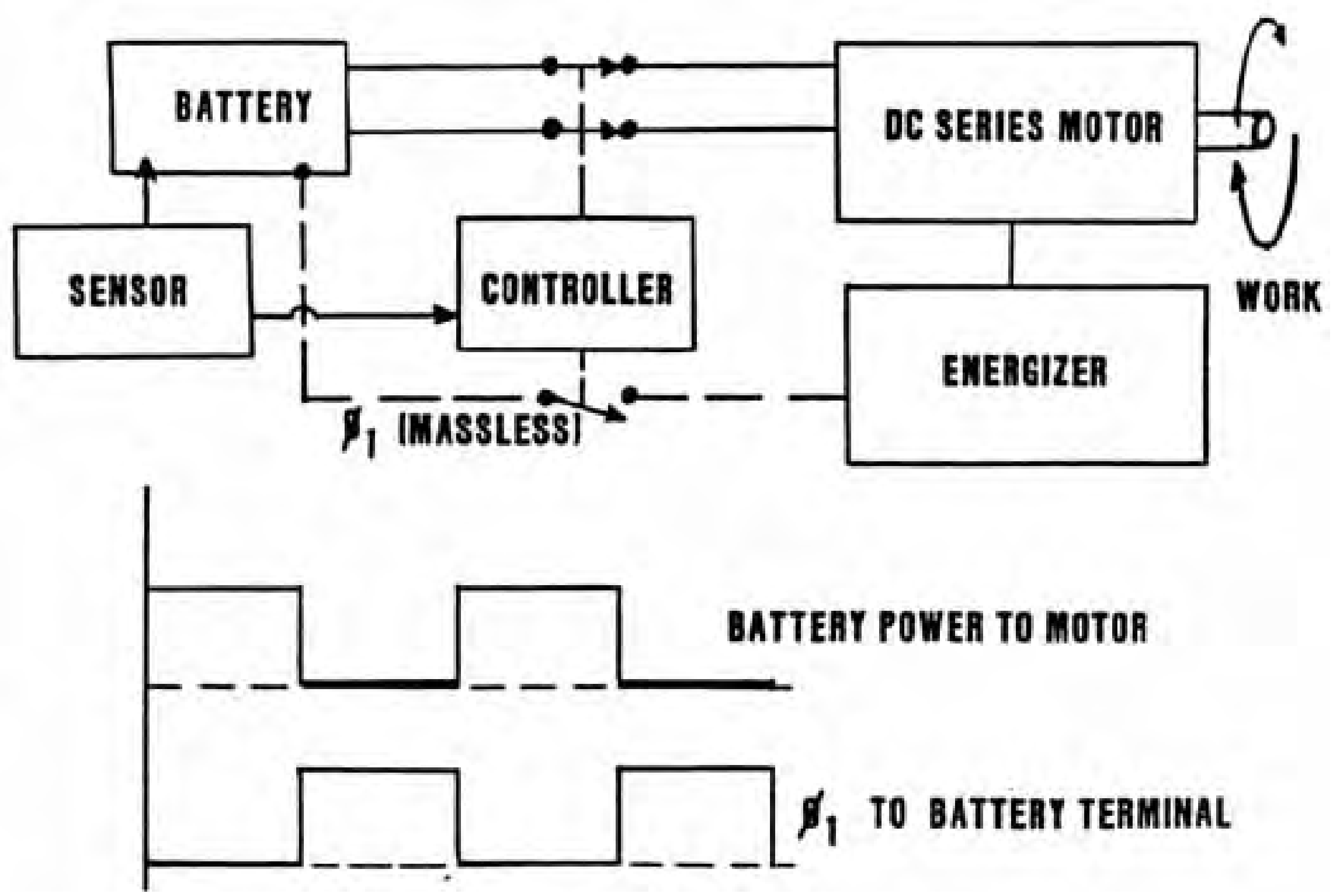
BATTERY
DC SERIES MOTOR
SENSOR
CONTROLLER
WORK
ENERGIZER
Ø1 (MASSLESS)
BATTERY POWER TO MOTOR
Ø1 TO BATTERY TERMINAL

On this slide, we show a theoretical scheme which several researchers have discovered and used to build simple free energy motors.

In this scheme, we drive an ordinary d.c. series motor by a two wire system from an ordinary battery. The motor produces shaft horsepower, at -- say -- some 30 or 40 percent efficiency, compared to the power drained from the battery. This much of the circuit is perfectly ordinary.

The trick here is to get the battery to recharge itself, without furnishing normal power to it, or expending work from the external circuit in the process.

To do this, recall that a charged particle in a "hooking" del-phi river moves itself. This is true for an ion, as well as for an electron. We need only make the del-phi in correct fashion and synchronize it; specifically, we must not release the hose nozzles we utilize to produce our del-phi river or waves.

The inventors who have discovered this have used various variations, but here we show a common one.

First, we add an "energizer" (often referred to by various other names) to the circuit. This device makes the del-phi waves we will utilize, but does NOT make currents of electron masses. In other words, it makes pure Ø-dot. It takes a little work to do this, for the energizer circuit must pump a few charges now and then. So the energizer draws a little bit of power from the motor, but not very much.

Now we add a switching device, called a controller, which breaks up power to the motor in pulses. During one pulse, the battery is connected and furnishes power to the motor; during the succeeding pulse, the battery is disconnected completely from the motor and the output from the energizer is applied across the terminals of the battery.

If frequency content, spin-hole content, etc. are properly constructed by the energizer, then the ion movements in the battery reverse themselves, recharging the battery. Again, remember that these ions MOVE THEMSELVES during this recharge phase. Specifically, we are NOT furnishing ordinary current to the battery, and we are not doing work on it from the energizer.

If things are built properly, the battery can be made to more than recover its charge during this pulse cycle.

To prevent excess charge of the battery and overheating and destroying it, a sensor is added which senses the state of charge of the battery, and furnishes a feedback signal to the controller to regulate the length of recharge time per "power off" pulse. In other words, the system is now self-regulating.

The relation between power pulses and recharge pulses is shown on the graphs at the bottom. Note that regulation may decrease the time of recharge application of the del-phi river.

This system, if properly built and tuned, will furnish "free shaft energy" continually, without violating conservation of anenergy. Remember that the del-phi condition across the battery terminals means that spacetime is suddenly curved there, and conservation of energy need no longer apply.

Again, this system is consistent with general relativity and with the fact that Ø-field alone can drive a situation relativistic. We have deliberately used these facts to do direct engineering. Our "extra energy" comes from shifting phi-flux -- the energy of the universal vacuum spacetime -- directly into ordinary energy for our use. Thus we draw on an inexhaustible source, and our device is no more esoteric than a paddlewheel in a river. The only difference is that, in this case, we have to be clever enough to make and divert the river in the right timing sequency.

This is a free energy device which an ordinary person, who knows a little electronics, can experiment with in the basement. To develop it, one is talking several thousands of dollars and a lot of persistence and tinkering; one is not talking millions.

WARNING: THE PRECEDING EXPERIMENTS ARE HAZARDOUS. DO NOT ATTEMPT THESE EXPERIMENTS UNDER ANY CIRCUMSTANCES UNLESS YOU ARE AN EXPERIENCED ELECTRICAL RESEARCHER, EXPERIENCED IN PERFORMING EXPERIMENTS WITH LEAD-ACID BATTERIES AND PULSE CHARGE AND DISCHARGE OF SAME, AND UNLESS YOU ALSO USE ALL SAFETY PRECAUTIONS SUCH AS GOGGLES

AND PROTECTIVE GLOVES, SLEEVES, AND APRON. YOU MUST NOT HAVE OTHER INFLAMMABLE LIQUIDS OR OTHER SUBSTANCES PRESENT WHICH COULD BE IGNITED AND BURN OR EXPLODE. SURGED LEAD-ACID BATTERIES PRODUCE HYDROGEN GAS, WHICH CAN EASILY EXPLODE SINCE SPARKING ALSO CAN OCCUR. THE ACID FROM SUCH AN EXPLOSION CAN EASILY BLIND YOU IF IT GETS IN YOUR EYES, AND IT CAN BURN YOUR SKIN. IN ADDITION, LEAD AND LEAD COMPOUNDS ARE POISONS, AND ARE TO BE HANDLED ONLY BY EXPERIENCED RESEARCHERS. THESE EXPERIMENTS ARE NOT FOR AMATEURS UNDER ANY CIRCUMSTANCE, BUT ONLY FOR EXPERIENCED PROFESSIONALS WITH PROPER KNOWLEDGE AND TRAINING, AND USING PROPER PRECAUTIONS. NEITHER THE AUTHOR NOR THE PUBLISHER IS REPONSIBLE OR LIABLE FOR ANY ACCIDENTS OR DAMAGE YOU MAY ENCOUNTER, AND ALL EXPERIMENTATION WITH THESE DEVICES AND PROCEDURES IS AT YOUR OWN ASSUMED PERSONAL RISK.

Next Slide

The Tom Bearden Website

SLIDE 41.

THE TRANSLATOR

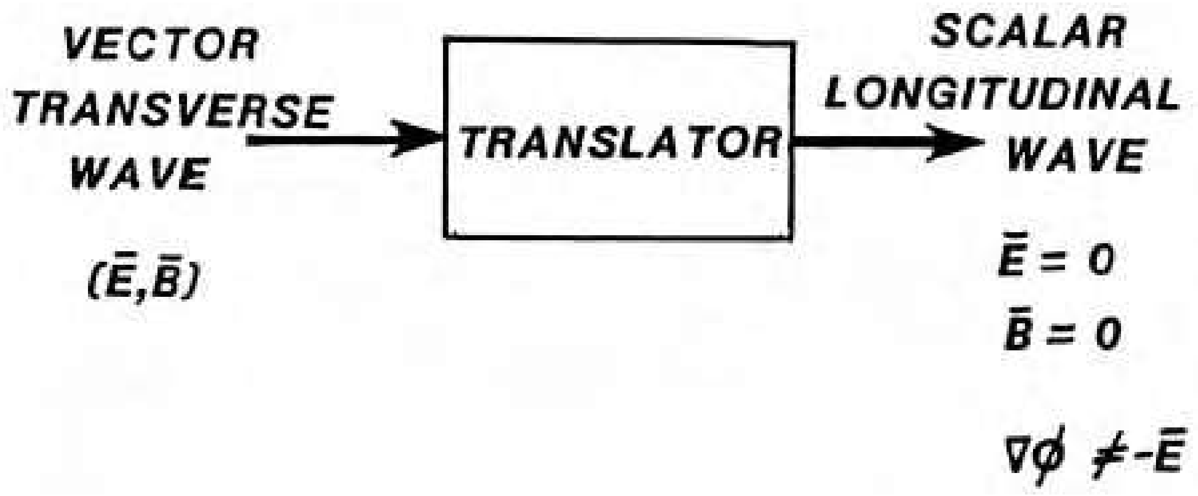

We now turn to a fundamental concept we need in order to understand some of the effects used by the Soviets in their Tesla weapons.

Note that one other country -- one not hostile to the U.S. -- also has such weaponry.

This slide shows the concept of a TRANSLATOR. Briefly, if we input a transverse matter wave, the translator outputs a scalar longitudinal wave without spin vortex hooks. If we input a scalar longitudinal wave -- even one without hooks -- the translator puts out a normal vector transverse matter wave.

We define a translator as anything which will accomplish one or both of those functions (usually both).

Basically, any device which is TOTALLY nonlinear to transverse waves will produce scalar waves from a transverse wave input. Any device which is totally nonlinear to longitudinal waves will produce transverse waves from a longitudinal wave input.

It is comparable to certain plasmas, which, as is well
known, produce such a translator effect. If you input longitudinal waves, you get transverse waves. If you input transverse waves, you get longitudinal waves. And the plasma is rather totally nonlinear.

Again, if our translator outputs non-hooking waves, it means that the component substructure waves oppose each other in sum zero fashion, and "kill" or nullify the vortex spin-holes.

Next Slide

The Tom Bearden Website

SLIDE 42.

TRANSVERSE WAVE INTERFERENCE

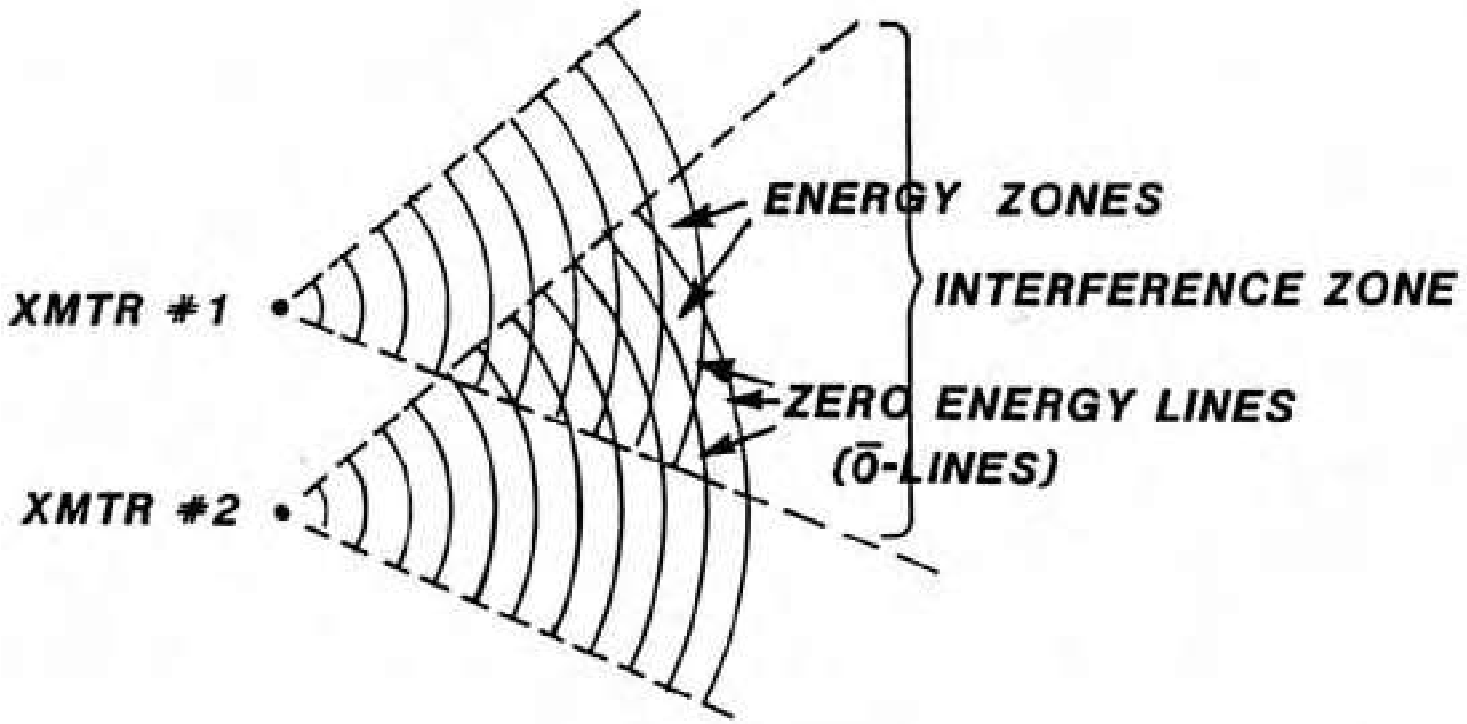

On this slide we show "normal " transverse wave interference, as we presently teach it in our textbooks.

We visualize two identical transmitters, each producing a beam in a pattern of about 55 degrees of so.

Where these beams overlap -- in the ideal, perfect case, we have constructive and destructive interference, with absolute zero-vector linear regions being formed. We put in the energy, and out-of-phase interference creates the zero zones, which contain no energy as such.

We point out, however, that these zero-lines are scalars,
and contain high stresses on spacetime.

Next Slide

The Tom Bearden Website

SLIDE 43.

LONGITUDINAL WAVE INTERFERENCE [SCALAR WAVES]

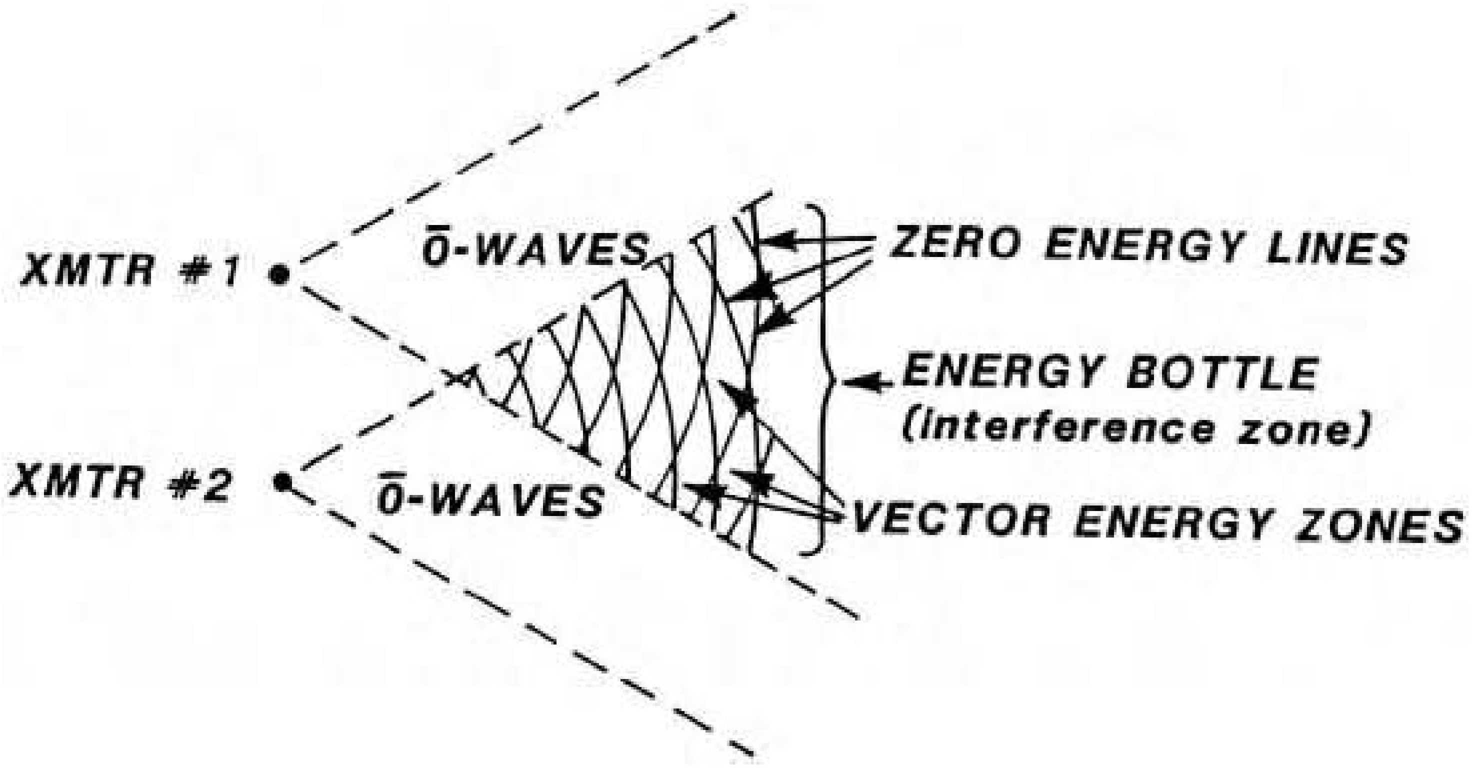
XMTR #1
XMTR #2
0̄-WAVES
0̄-WAVES
ZERO ENERGY LINES
ENERGY BOTTLE
(Interference zone)
VECTOR ENERGY ZONES

On this slide we show the interference pattern that emerges from the interference of two identical scalar wave beams from translator/transmitters.

Now note we are interfering two zero-vector waves, each containing a substructure. In this case we assume similar substructures.

What happens now is that we get a similar interference pattern, but with some startling differences.

First, we put in the zero-lines. These represent in-phase conditions for the substructures.

. Second, the substructures themselves interfere, and form energy in the grid zones between intersecting zero-lines.

In the perfect case, however, this energy cannot radiate away. It is physically created and trapped as if in an "energy bottle." We are CREATING this energy at a distance, and storing it in a bottle. If the bottle were perfect, we would continue to accumulate energy in the grid zones -- notice we are continuing to pour energy into both transmitters, and there is zero ordinary energy anywhere else except in the grid zone "bottles."

If we turn off the transmitters smoothly, we extinguish the energy in the bottle completely. It is sucked right out of there, and the vacuum returns to its normal condition.

If we erratically and nonlinearly turn off the transmitters, we destroy the bottle and "dump" the energy suddenly. This is exactly like a sudden EMP from a nuclear weapon. A sudden pulse of energy is freed to radiate and interact in the region where the bottle was.

The high altitude booms off the East coast of the U.S. a few years ago were made in precisely this fashion, from such a weapon being adjusted and calibrated in the Soviet Union.

Next Slide

SLIDE 44.

CREATING ENERGY AT A DISTANCE

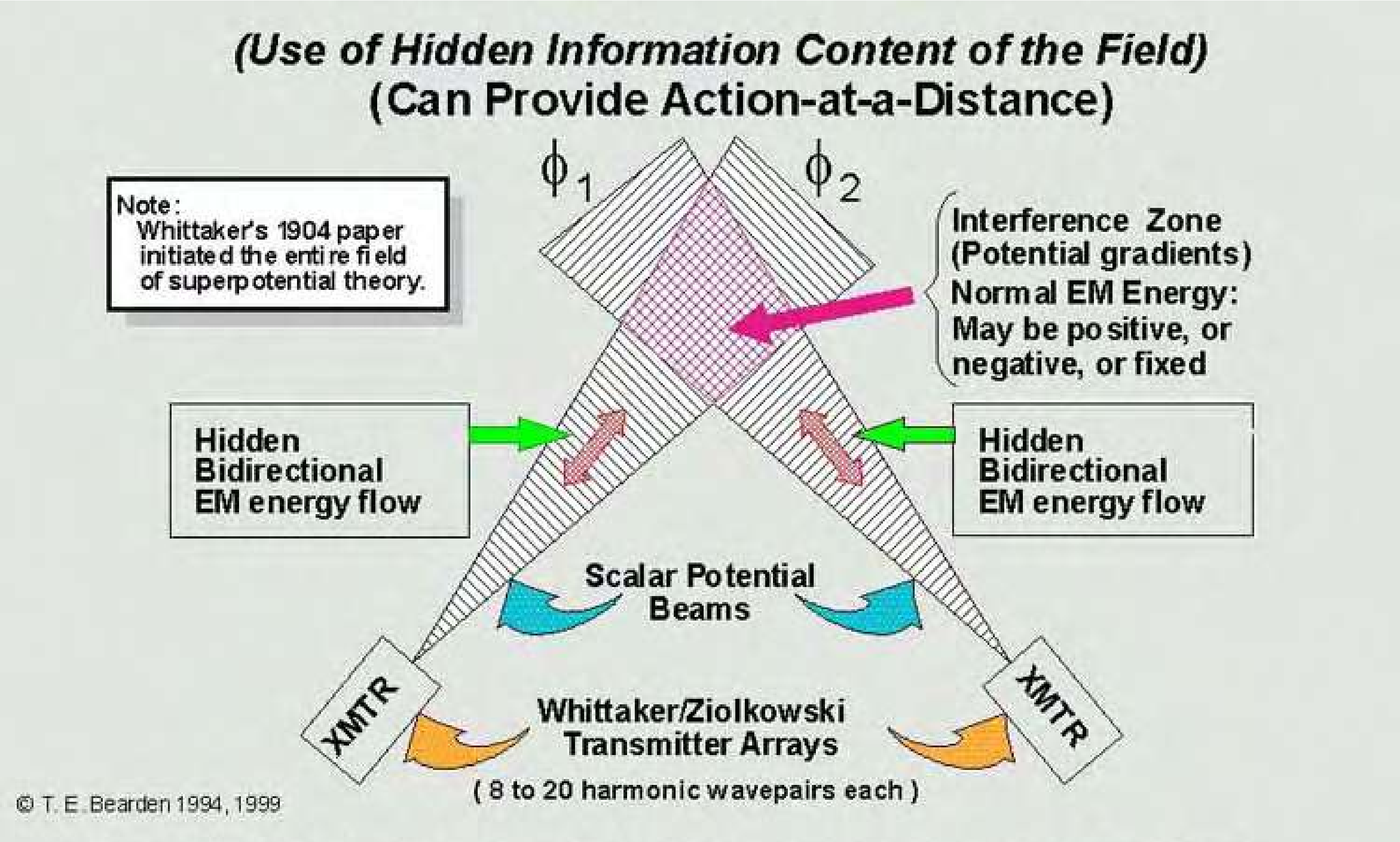
(Use of Hidden Information Content of the Field)
(Can Provide Action-at-a-Distance)
ϕ1
ϕ2
Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.
Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)
© T. E. Bearden 1994, 1999

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepair functions) produces all EM force fields and waves.

But let us look at one way to do this sort of thing more efficiently.

Here we show two scalar transmitters (that is, two translator /transmitters) which form narrow beams, and which cross those beams at a distance.

In the crossing zone, scalar interference is established, making an energy bottle and producing trapped or locked in energy.

If I have a physical target in that crossed zone, I can literally "fry" the target, for I get all the power inputted to the transmitters contained in the bottle zone. I don't have any square law losses.

I can heat a metal object white hot at a distance, for example. Because all real bottles are slightly imperfect, the real bottle is leaky and some photons escape. I can observe the white hot steel in the bottle, and I can record and measure the temperature from the escaping photons, by simple radiometric methods.

Now if I smoothly turn off both transmitters linearly, all the energy in the bottle disappears immediately. I can immediately reach in and pick up a black, cool piece of steel. And that cannot be done with ordinary inductive or radiowave heating.

That has been done in a laboratory on the North American continent.

Next Slide

The Tom Bearden Website

SLIDE 45.

NONLINEARITIES CAUSE SHIFTING [PARTIAL TRANSLATION]

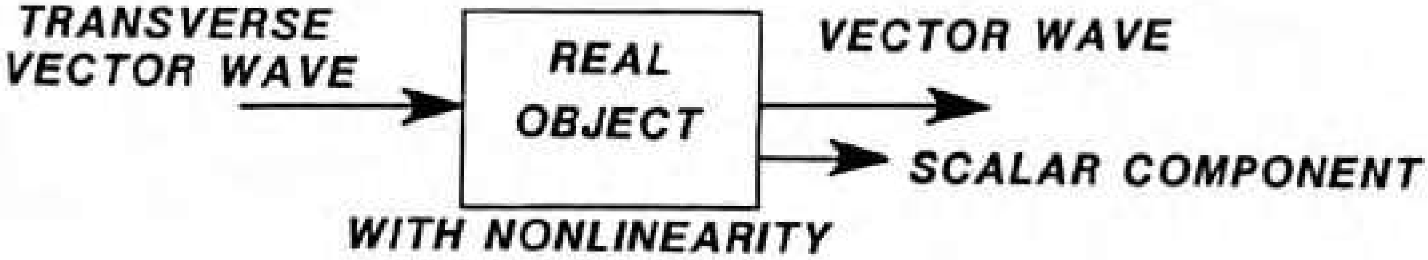
TRANSVERSE VECTOR WAVE
REAL OBJECT
WITH NONLINEARITY
VECTOR WAVE
SCALAR COMPONENT

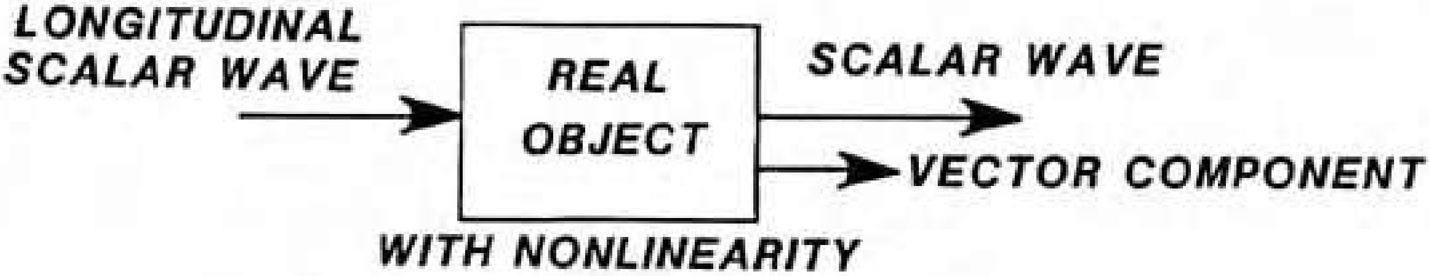
LONGITUDINAL SCALAR WAVE
REAL OBJECT
WITH NONLINEARITY
SCALAR WAVE
VECTOR COMPONENT

Here we show the results of electrical nonlinearities in an object which encounters a true translated (non-hooking) scalar wave. All objects have some electrical nonlinearities in them, and so all act partially as translators.

This means that any real object creates a small scalar aura around itself from its continual temperature interactions -- from absorbing and radiating ordinary EM waves. In addition, the object normally is subject to undetected scalar wave radiations, and so it also produces a small aura of transverse vector waves around itself by translation.

All real objects accomplish at least a little bit of translation.

Next Slide

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

SLIDE 46.

THE SCALAR AURA

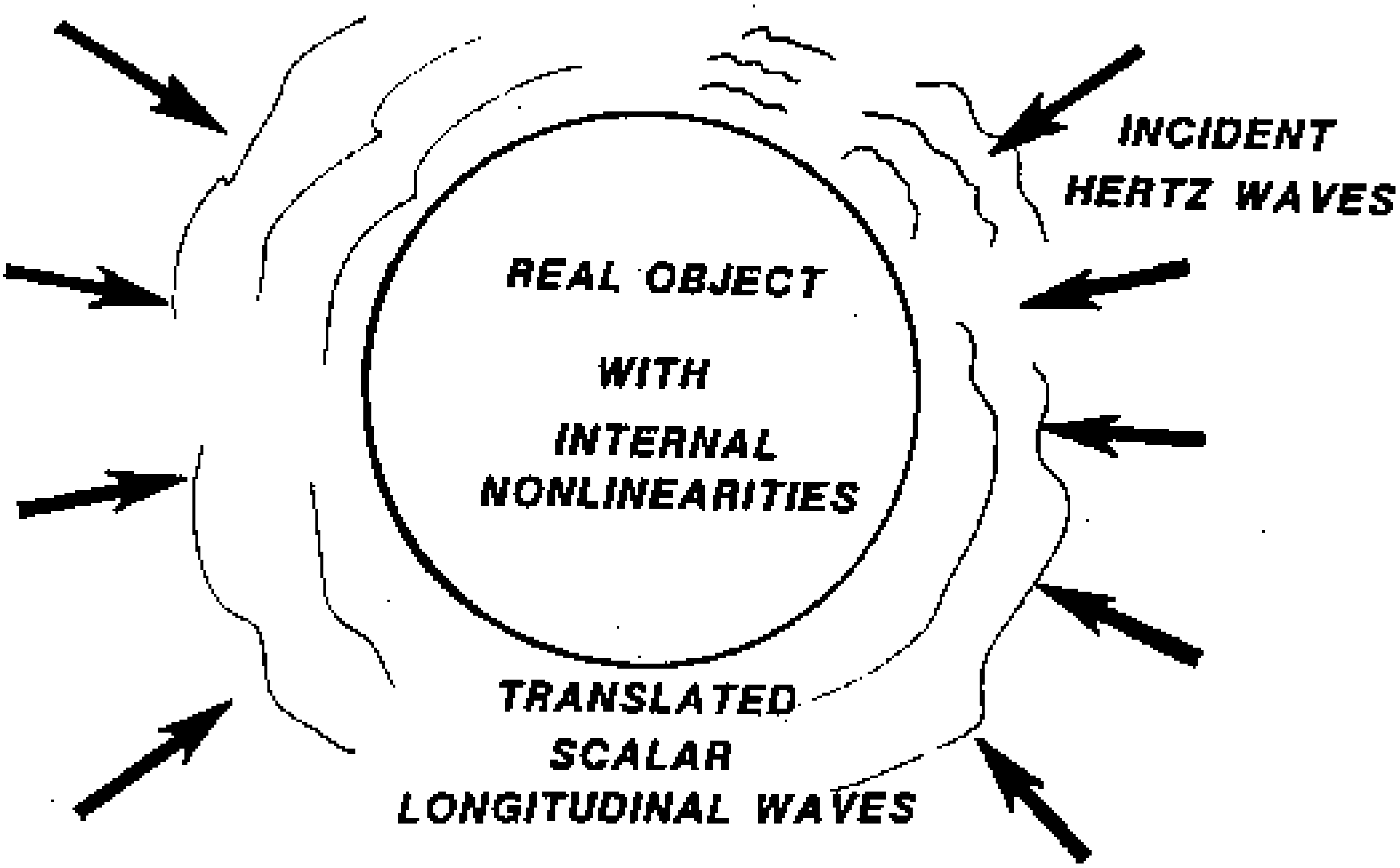
INCIDENT
HERTZ WAVES
REAL OBJECT
WITH
INTERNAL
NONLINEARITIES
TRANSLATED
SCALAR
LONGITUDINAL WAVES

We examine now an object with temperature, that is, a normal object continually absorbing ordinary EM radiation and continually emitting EM radiation.

The object has a small fraction of its incident energy expressed as a scalar aura around it, because of translation. Note that normal detectors will not detect this aura, but special two-stage interferometric detectors will detect it. Also note that the object replenishes- the ordinary radiation flow by its translation of some of its incident scalar radiation into the normal or "transverse" (so-called) radiation.

The entire internal pattern -- all interactions and functions of the body -- are precisely present in its scalar aura, down to the finest detail. This makes an x-ray pale by comparison.

True psychics actually use the nonlinearity of their nervous systems (note the avalanche ion discharges across dendrite gaps) and the interferometric aspects (note two cerebral hemispheres, connected by a corpus callosum) to detect the scalar aura. This is a physical fact and not esoteric delusion. To control this, however, is normally as difficult as learning to walk a high wire and do acrobatics on it. There are not many high wire walkers in the world, but it certainly is a skill within human control capability and purview.

T. Galen Hieronymus was absolutely correct when he stated that the entire internal working pattern of an object is radiated by that object in terms of an extraordinary energy, which he called "eloptic" energy. It is actually translated scalar wave energy, and entirely consistent with present EM theory when the glaring errors in its foundations are corrected.

Shortly I hope we will see instrumentation which can directly examine the scalar aura in the finest detail, seeing every illness, and simply setting dials to radiate the body with gentle, harmless scalar waxes to directly correct illnesses.

Next Slide

The Tom Bearden Website

SLIDE 47.

VIRTUAL FLUX AND MAGNETIC FIELD [LINEAR CASE]

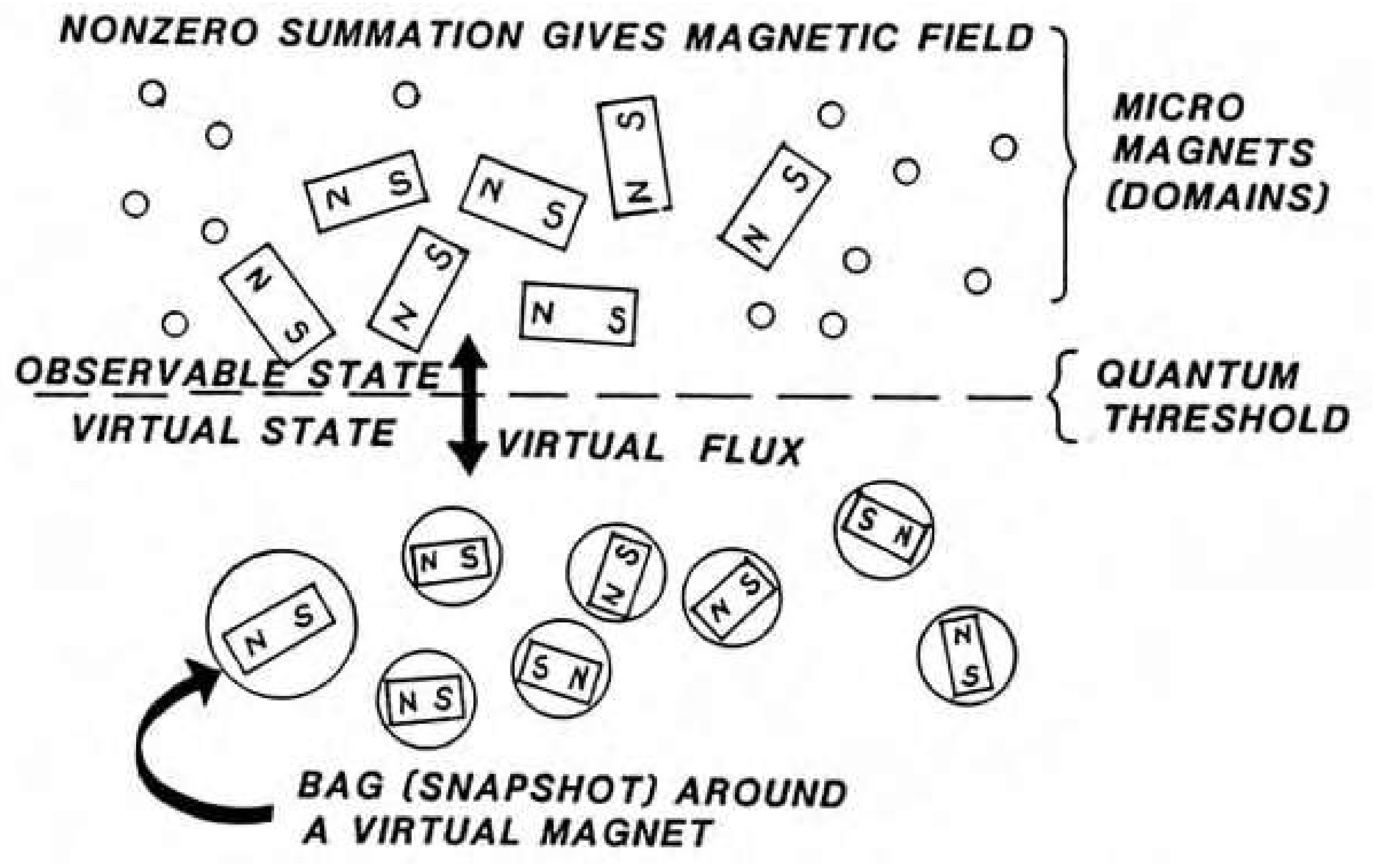
NONZERO SUMMATION GIVES MAGNETIC FIELD
MICRO MAGNETS (DOMAINS)
N
S
OBSERVABLE STATE
VIRTUAL STATE
VIRTUAL FLUX
QUANTUM THRESHOLD
BAG (SNAPSHOT) AROUND A VIRTUAL MAGNET

Briefly, let's look at some nice things we can do with magnetism. Specifically, let's find out how to make magnetic monopoles and use them to do things in the laboratory. This again has been done in a laboratory on the North American continent.

First, about all the present science can tell us about a magnet is that "a big magnet is made of littler magnets." And if we examine one of the smaller magnets, it's made of even smaller magnets. This is like saying big dirt piles are made of smaller dirt piles, and that explains what dirt is.

However, let's use that with our knowledge of the virtual flux vacuum.

In this slide, we show diagrammatically the situation for a common magnet. Note that domains in the magnetic material are themselves little magnets, and their alignment and vector summation determine whether or not there's said to be a magnetic field present. Nonzero summation states that there's a resultant magnetic vector, and hence an external magnetic field. Zero summation states there's no external nonzero resultant, hence no external magnetic field.

Actually, there's an external scalar magnetic potential field, even when the external vector magnetic field is zero. The substructure is still there.

If we pursue this "big magnets are composed of smaller magnets," eventually we reach the quantum threshold, and we have a substructure of virtual, subquantal magnets in vacuum, in the virtual particle flux. At least we conceive each little virtual particle as if it were spinning, and hence a little magnet.

We put a conceptual bag around each little virtual magnet.

In a linear situation, the north pole is as strong as the south pole, and so just outside the bag, the poles cancel or appear zero. Since the virtual bag appears to be zero length to an external macro observer, the poles seem to be directly superposed on top of each other, yielding no pole at all to the observer.

However, you and I now know that both poles are still in there, in the virtual substructure, and we certainly have remaining with us a translated scalar magnetic field.

There's a virtual flux to and from each observable particle of charged mass in the observable state as shown, but this flux is now scalar in virtual magnets, except for nonlinearities in the structures above the quantum level. Thus accelerated portions -- atoms with electrons in whirling orbit, spinning electrons, protons, etc. -- possess nonzero ordinary vector magnetic fields by translation. Again, notice the successive interlocking levels of reality, all the way from deep in the virtual state into interlocking levels in observable state.

Next Slide

SLIDE 48.

PRODUCING MAGNETIC MONOPOLES

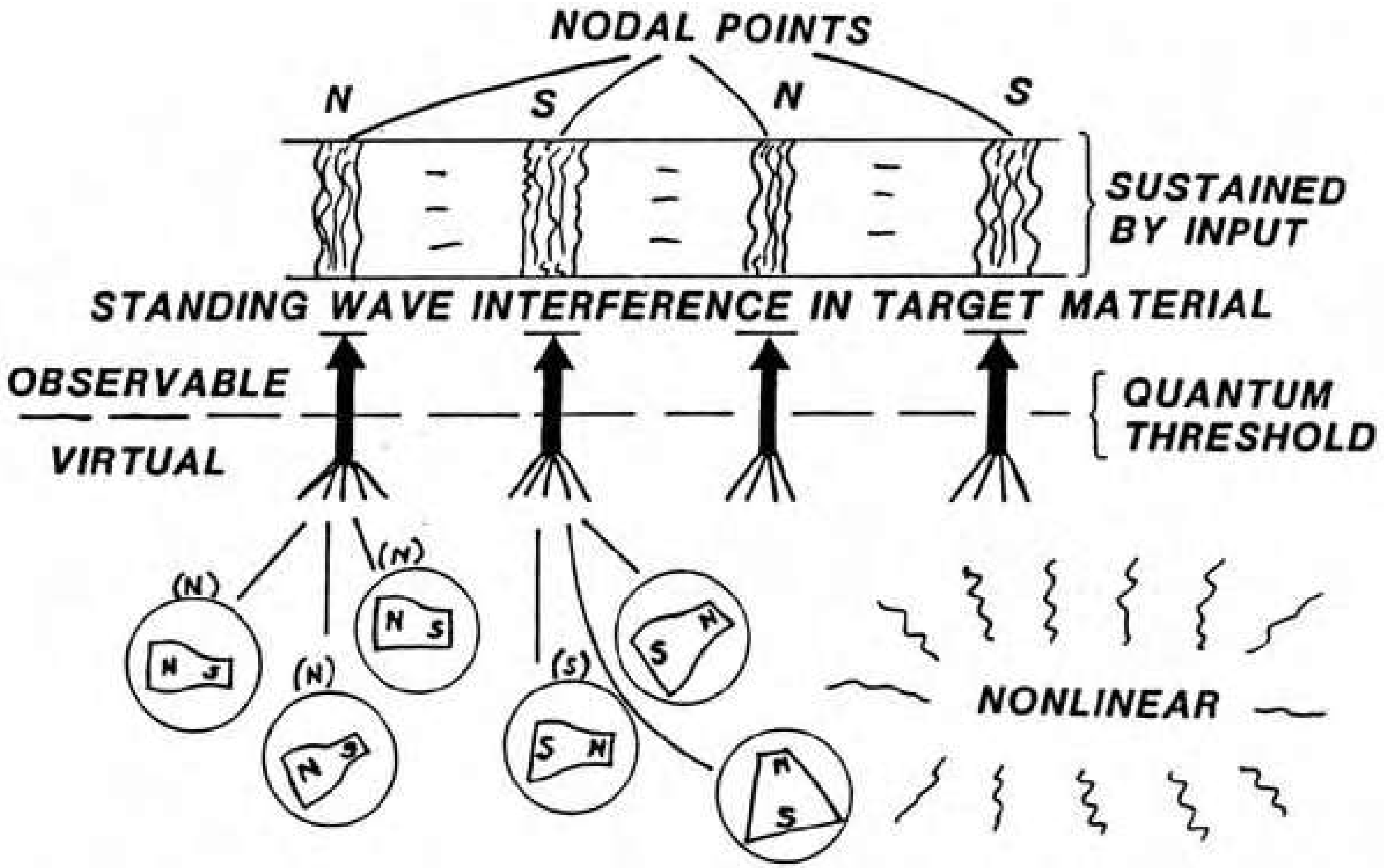
NODAL POINTS
N
S
N
S
SUSTAINED BY INPUT
STANDING WAVE INTERFERENCE IN TARGET MATERIAL
OBSERVABLE
VIRTUAL
QUANTUM THRESHOLD
(N)
(N)
(N)
(S)
N S
N S
N S
S N
S N
N S
NONLINEAR

Now let us look at a non-linear case of the same sort of thing. Now the bags are still around the little virtual magnets, but they are nonlinear. This means that, to an observer outside the bag, one pole seems bigger or more powerful than the other. In other words, to the quantum observer, this bag appears to be a magnetic monopole at a point.

In the flux still on and off each observable spinning charged mass, we now have a steady component of "monopoles". If we have a standing scalar wave present in a physical material in which the nonlinearity exists -- and the scalar waves can even be PRODUCING that nonlinearity -- we will have nodes available at which the monopoles will congregate and emerge and interact.

North monopoles will congregate at one node, while south monopoles will congregate at the next, and so on in alternating fashion.

That means that, at anyone node, monopoles of the same kind are steadily being "deposited" in the material. These monopoles strongly repel each other, and so the material at that node is increasingly stressed in a tensile fashion.

Eventually the material will be torn apart at the node,
stress relieving the situation. Movement of the material will release the nonlinear condition, stopping monopole production and deposit. However, at a break node, the same kind of magnetic pole will appear on each side of the break.

That is, the breaks will be N-N, S-S, N-N, S-S, etc.

An ordinary magnetic does not do that when it breaks. Instead, it breaks N-S, N-S, etc.

Next Slide

The Tom Bearden Website

SLIDE 49.

BREAKING METAL WITH MAGNETIC MONOPOLES

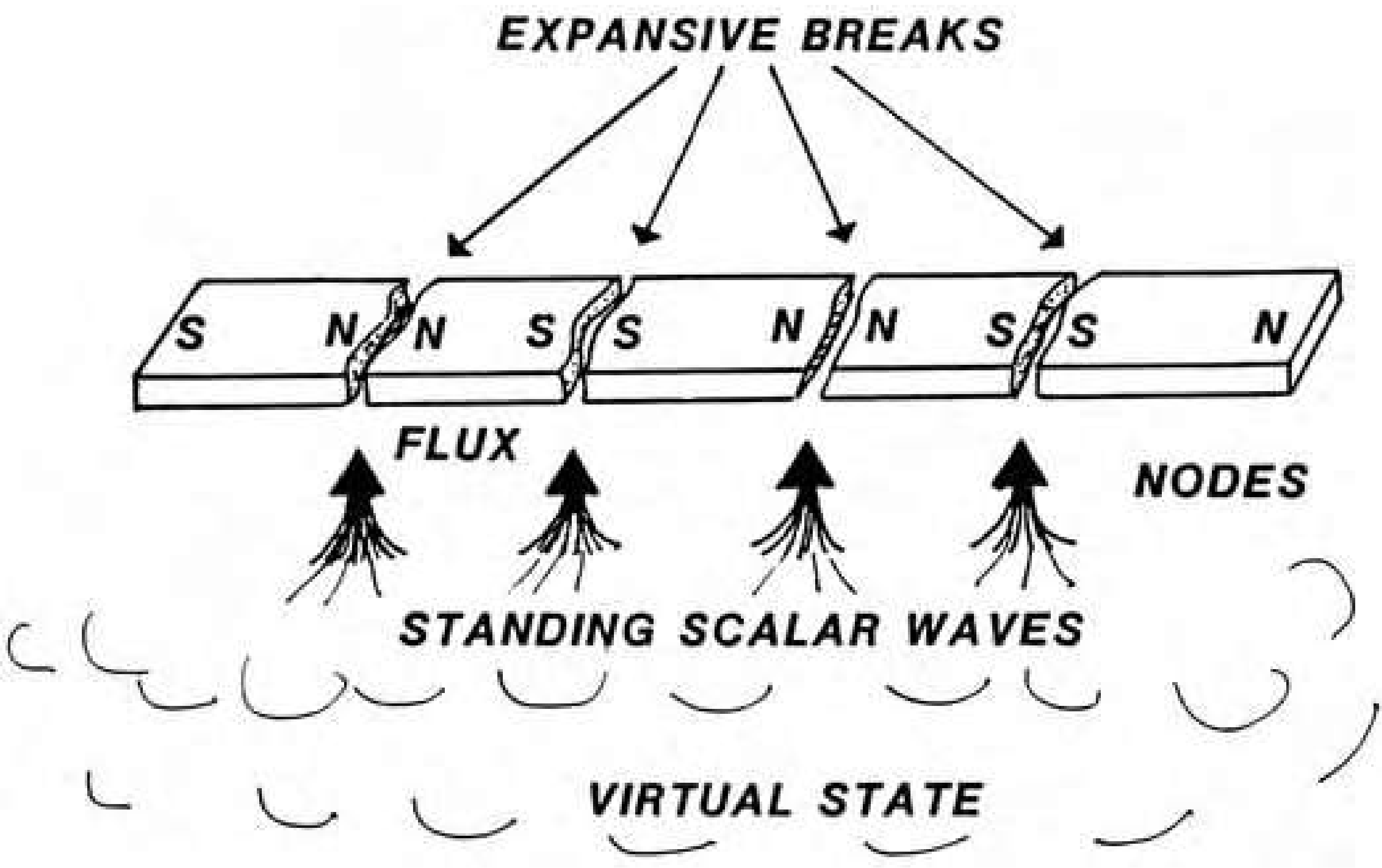

Again, metal has been broken in a North American laboratory utilizing the production and deposit of magnetic monopoles in this fashion.

On this slide we show what was done.

An ordinary bar of steel was exposed to standing scalar
waves in a nonlinear situation, depositing magnetic monopoles at the wave nodal points in the metal bar. The breaks occurred as shown on this slide.

Much of what I am presenting is based on certain experiments accomplished in experimental laboratories here and abroad. Most of the mechanisms have actually been produced in the lab. However, these are proprietary labs and I cannot violate trust and reveal the precise construction details utilized. In some cases I have not seen the precise constructions, but only the direct results.

But ironically, the production of magnetic monopoles -- which has fiercely resisted the normal scientific approach -- yields readily to the new approach.

Next Slide

[The Tom Bearden Website](http://www.cheniere.org)

SLIDE 50.

CREATING AN INERTIAL FIELD (ANTIGRAVITY)

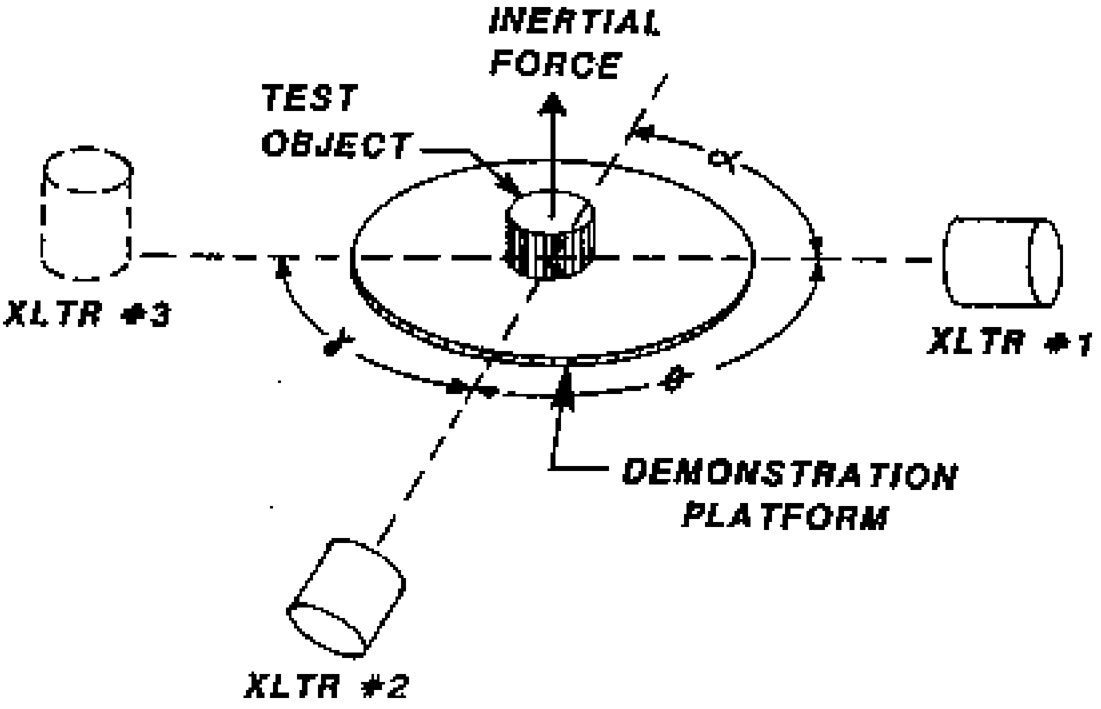

We now show schematically how to produce an inertial field, or antigravity. Also an inertial space drive can be produced in this manner.

This again has been done in a North American laboratory. In fact, inertial fields have been produced in more than one lab here in North America.

In the setup shown here, three translator/ transmitter projectors are used and oriented so as to give three-dimensional interference, or interference along all three spatial axes.

By controlling the substructures of the three projectors all together, one can control the precise phasing in all substructures, with respect to a common time reference.

By this means, a unilateral force can be created of each and every charged mass particle (that is, on each proton) in the nucleus.

Since we know that the neutrons and protons in a nucleus of an atom are continually interchanging, one into the other, back and forth, all of the nucleons can be treated as "diffuse protons" for mass purposes, and scalar wave interaction purposes.

Thus if we have the correct pattern to affect protons, we can control the phases of the substructure waves of the projectors so as to produce unilateral, parallel forces on and of each spinning proton in the nucleus.

This means we can produce unilateral thrust in an object exposed to the triad interference of the projectors.

In short, this produces antigravity. If sufficient force is produced, the test object in the interference zone simply lifts itself off against earth gravity. This has been done in North America.

If we phase the projectors so that counterbalancing forces are produced on each proton, we can either make the proton "lighter" inertially or "heavier."

In fact, we could "freeze" a common pistol so that it could not be moved and its trigger could not be pulled, simply by sufficiently increasing its inertia. This was done in experimental demonstrations in Toronto, Canada in 1969, and witnessed by the Chief of the Homicide Squad of the Toronto police. It was his pistol that was frozen for the demonstration.

If the projectors themselves are on board a space capsule, for example, they can produce such a unilateral thrust in and of a central mass or disc attached to the of the craft. By this means, an inertial spacedrive can be built.

Next Slide

The Tom Bearden Website

SLIDE 51.

EFFECTS ON HUMANS

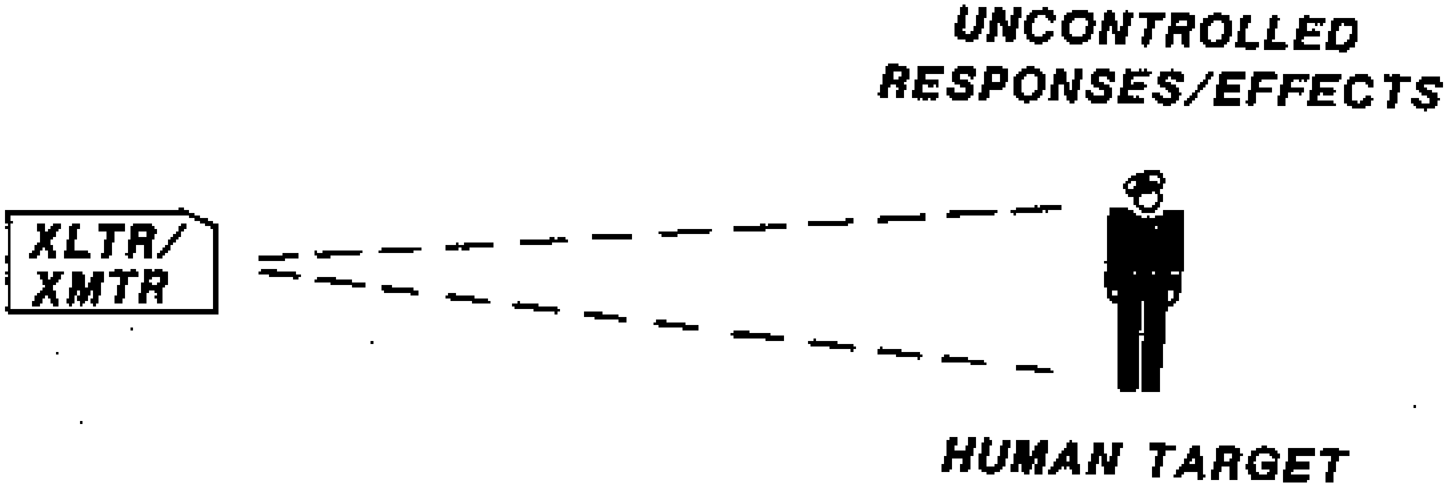

Now let us look at some of the effects on humans, when radiated with a true scalar wave beam from a translator/transmitter.

First, the human nervous system is highly nonlinear. Essentially, across billions of synapses, sudden ion discharges -- analogous to spark discharges -- are continually or periodically occurring. Brain waves, for example, are not waves along wires; instead, they are waves of avalanche discharges across tremendous numbers of nerve cell gaps.

The central nervous system is thus highly nonlinear, and the gap firings of the nerve endings act as translators.

These translators partially translate incident scalar waves, so that additional transverse matter waves -- or ordinary energy inputs -- occur in the ion discharges.

Thus additional energy is being added to the nervous system by incident scalar waves.

Many long-term effects of the environment on the body are gradually kindled in this fashion, even over a period of years.

For a stronger dosage of scalar waves, the nervous system experiences "jamming" of its signals. This leads to four basic "noise jamming" effects.

If a small dose is absorbed, the interference is slight. One feels as if one had taken a whiff of anaesthesia -- an unreal, dreamy feeling.

If a slightly larger dose is absorbed, the interference is moderate. Now the single channel system -- the motor system -- goes, and the individual is paralyzed. Mentally he is barely conscious, in a sort of hypnogogic, waking dream state. Any tendency toward epileptic seizure or other nervous disorder may surface here, as may heart palpitations or even heart failure.

For a heavier dosage yet, the interference is now strong. Now the multiple systems start failing. Consciousness is lost and the individual is in a deep coma. In addition, seizures and convulsions, and loss of control of some body functions, may occur. The autonomic nervous system is keeping the heart beating with difficulty and keeping the body gasping for breath.

For a very strong dosage, the entire nervous system fails and death results almost immediately. This is quite similar to the effects of nerve gas.

These are the four major symptoms of scalar wave radiation of the body and translating absorption by the nervous system in an uncoordinated fashion.

Next Slide

The Tom Bearden
Website

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

SLIDE 52.

INITIATING SPECIFIC BIOLOGICAL EFFECTS

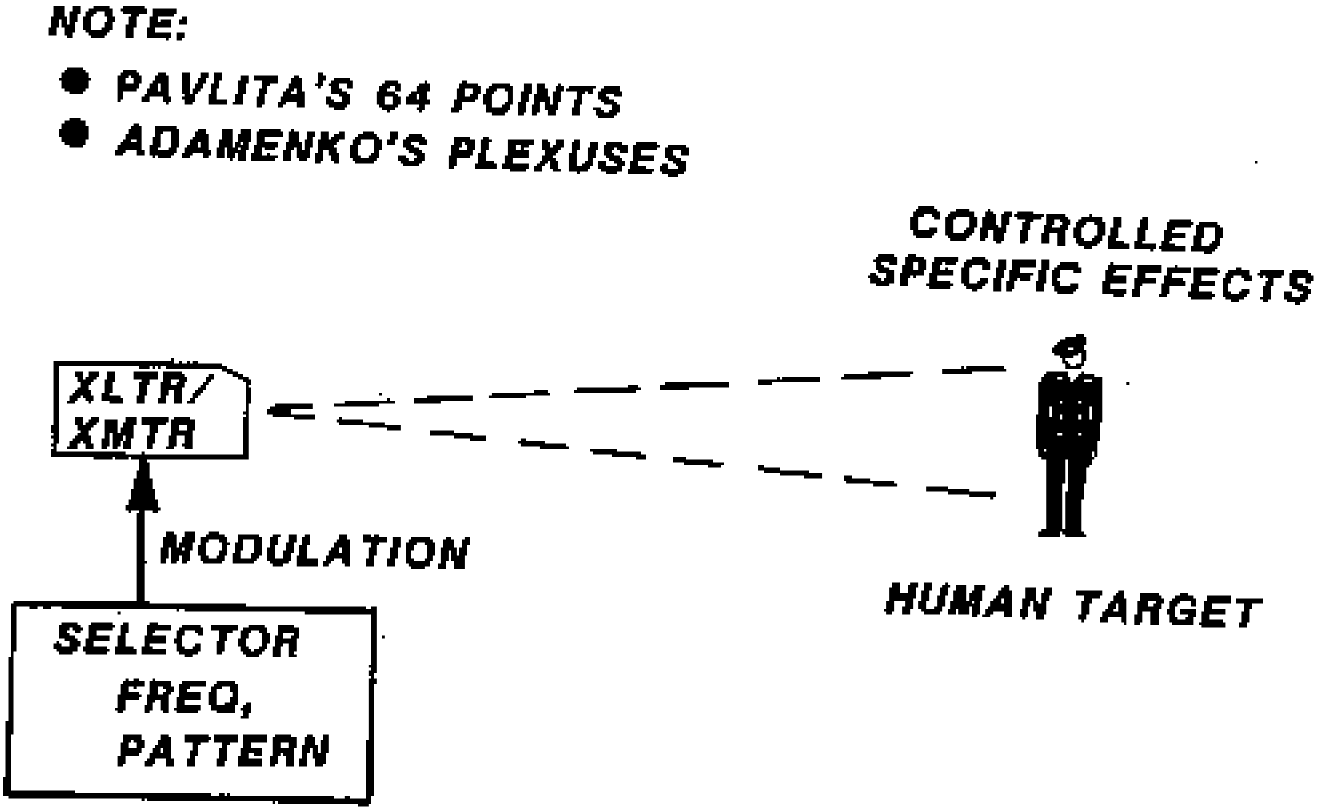
NOTE:
• PAVLITA'S 64 POINTS
• ADAMENKO'S PLEXUSES
CONTROLLED SPECIFIC EFFECTS
XLTR/ XMTR
MODULATION
HUMAN TARGET
SELECTOR FREQ, PATTERN

If we now add precisely patterned modulations to the substructure waves of the scalar zero vector beam, we can induce deliberate patterns in the dosage absorbed by a biological system.

As Kaznacheyev's experiments showed, any death or disease pattern can be induced upon an electromagnetic carrier -- specifically, upon the near-ultraviolet as an example. Some of his experiments were duplicated at the University of Marburg in West Germany, utilizing infrared carriers. These patterns can be dumped into cell cultures, eventually kindling the pattern of death or disease in the target cells.

The same can be done to a human from an external source. The major entry point is the acupuncture system, which reacts to scalar waves and transfers them to the nervous system for translation.

Pavlita, Czechoslovakian engineer, has built "psychotronic generators" for over 30 years that affect the human body. Adamenko, Soviet physicist, found that the acupuncture points formed groupings or plexuses that were frequency sensitive. In other words, certain frequency bands affect one plexus more than others. Further, these acupuncture points are known to be connected to effects produced at remote locations in the body.

Pavlita stated he had found 64 different sensitive "points" on the body, and had succeeded in building a generator for each. If by his points he meant Adamenko's plexuses, then frequency patterns can be created that affect certain acupuncture groupings -- and thereby certain specific body locations and effects -- more than others.

It thus appears reasonable that, by placing the proper frequency pattern in the modulations of the substructure waves feeding into the translator/ transmitter, the scalar ray or beam can be made highly selective as to body part affected and the effect generated or kindled in the body.

By this means, specifically tailored attacks on the body can be created.

Strong emotions such as sheer terror, overpowering rage, blind panic, etc. can be induced. Or physical effects such as epileptic seizure, loss of control of internal functions, loss of muscular control, severe weakness, specific failure of an individual organ such as the heart or liver, or blackouts and comas can be induced. In addition, patterns for specific toxins and diseases can be induced. With a strong absorption, these changes can be effected quickly. Imagine bubonic plague developing in 30 seconds, for example, or the effects of a cobra bite developing in 10 seconds.

If the individual is placed in crossed interference beams (that is, in the interference zone of a scalar interferometer), strong effects can be produced essentially instantly, including instantaneous death.

Of course BENEFICIAL effects can also be accomplished. As instrumentation is developed, one should be able to scan the aura, do a computerized readout to show the different medical problems either developed or developing in the body, have the computer compose a "cancellation or offsetting pattern," and radiate the body with a totally beneficial pattern, customized for that particular condition at that particular time. One should be able to quickly cure cancers and leukemias, for example, even of the brain and of the bone marrow. One may even be able to quickly reverse the effects of aging; at least there is no foreseeable reason why it should not be possible.

The new technology is a very powerful tool, and it can be used for either good or bad, depending upon the intent of the developer and the user.

Next Slide

[The Tom Bearden Website](http://www.cheniere.org)

SLIDE 53.

GROUND-TO-AIR USE

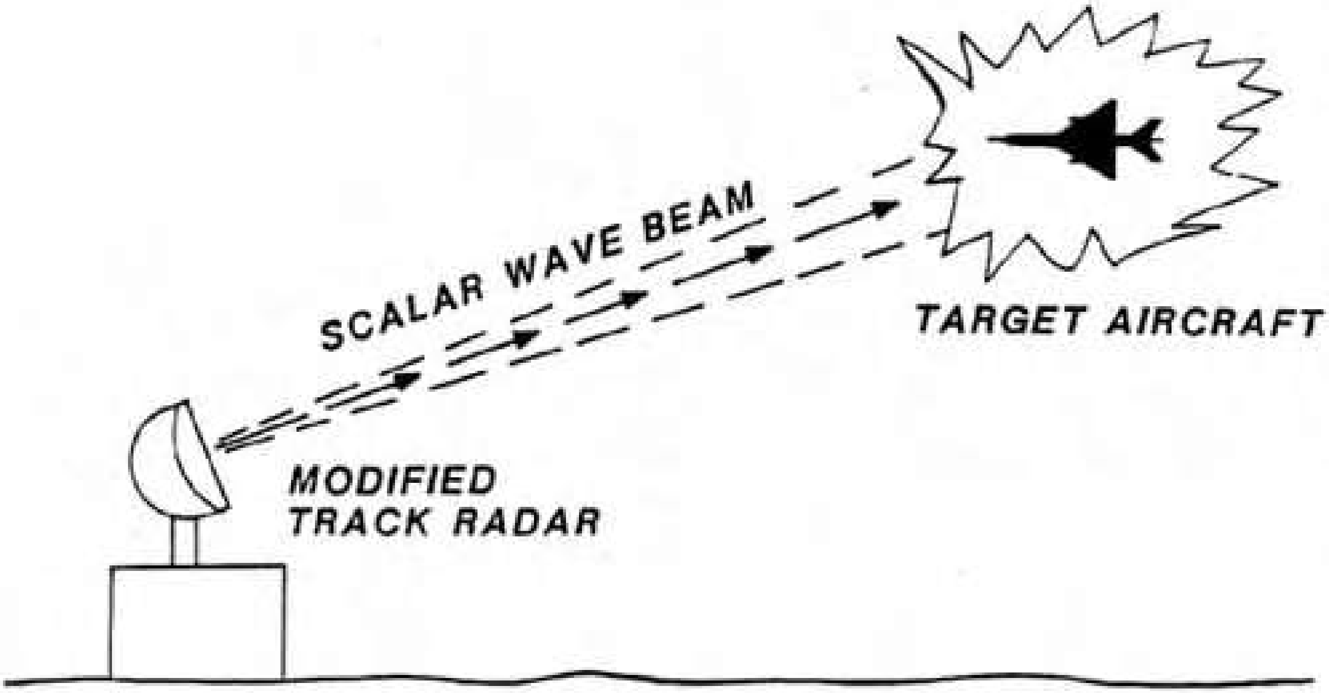

Here we show another use of the scalar beam technology already developed by foreign powers.

A modified track radar -- either continuous wave or pulsed -- transmits a scalar wave beam. It radiates an incoming hostile aircraft, keeping the beam on it, and kindling energy in the nonlinear electronic circuits of the aircraft by translation. Communications, navigation systems, avionics systems, electronic fuze systems, ignition and control systems, airborne radar, missile armament electronics, etc. are all vulnerable to this translation kindling of internally jamming energy from the incident scalar beam.

The electronics can be knocked out and the aircraft knocked down.

The pilot himself is also vulnerable because of his highly nonlinear central nervous system.

Some or all of the eight F-111's mysteriously

lost in
Vietnam may have been bagged by modified SA-2
FANSONG radars, using this effect.

Next Slide

The Tom Bearden Website

SLIDE 54.

TACTICAL AIR-TO-GROUND USAGE

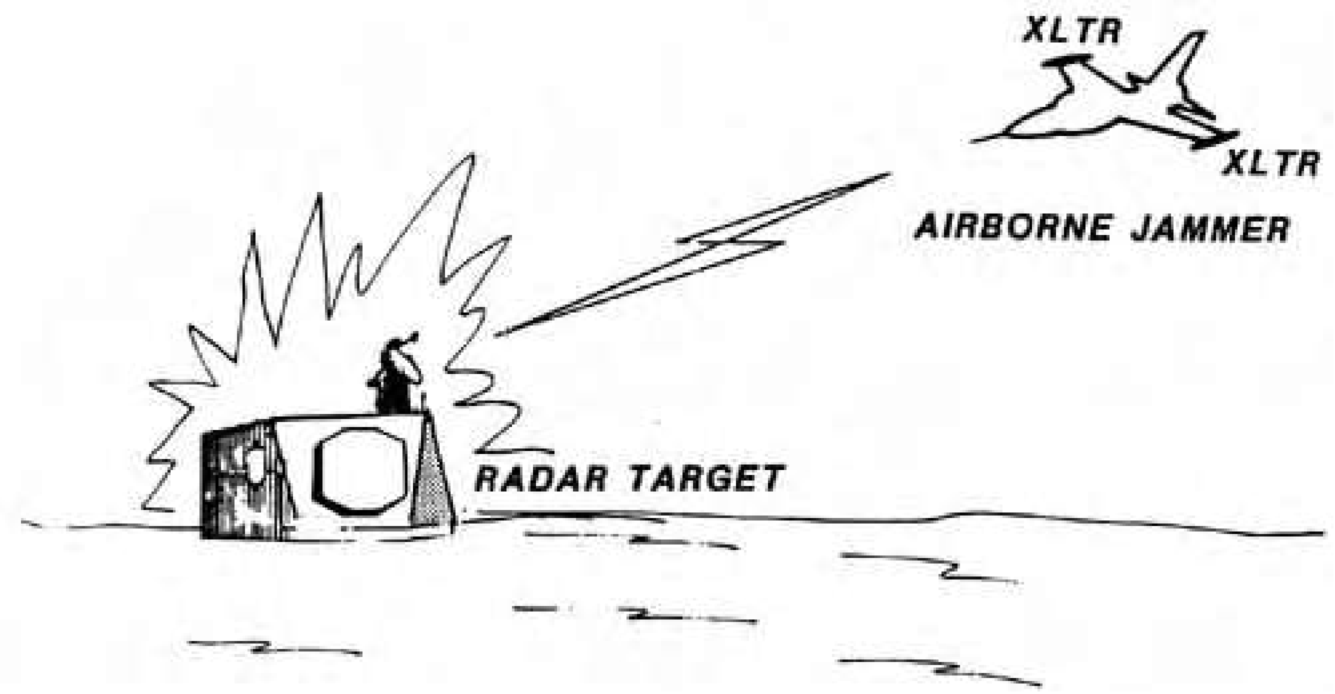

As shown on this slide, the technology is also usable from aircraft against ground targets, particularly ground-based radars and communications systems.

As shown, the aircraft has a modified jammer, able to emit a beam of scalar energy from one or more translator/transmitters.

By simply training a beam on the radar, eventually the radar can be knocked out. That is the simplest use.

A more sophisticated usage is provided by more than one translator/transmitter, acting as a scalar interferometer.

In this mode, each translator/transmitter emits multiple scalar carrier frequencies. Two Fourier expansions are established, so that a three dimensional "ball" or sphere of interference-zone energy is established. By pulsing or phasing, this ball is moved to strike the location of the radar, delivering a large burst of energy to it and knocking it out. Alternatively, the energy bottle is "dumped" when it reaches the targeted radar or its near vicinity.

A still better way is to fire one Fourier scalar pattern directly at the target radar, then fire a second pulse faster so that it overtakes the first directly upon the target. (The ensemble of such a pattern acts as a soliton, and its velocity is controllable by controlling the amplitude of the pattern generated).

When the second scalar overtakes the first, a sharp interference is formed, with a fast energy bottle, which is then nonlinearly dumped in microseconds or milliseconds. In this fashion, all the energy put out by the jammer is delivered directly upon and into the target in one sharp EMP pulse, much like that from a small nuclear explosion. The target is destroyed.

Note that such a weapon can be utilized against tanks, vehicles, explosives, missiles, ammunition, bridges (the airblast), bunkers (the blast can occur INSIDE the enclosure), ships, planes, personnel, and targets of varied types.

The Soviets have had such secret weapons for years.

One additional country has them also.

Next Slide

The Tom Bearden Website

SLIDE 55.

INTEGRATED WEAPONS DISPLAY (MODIFIED JAMMER)

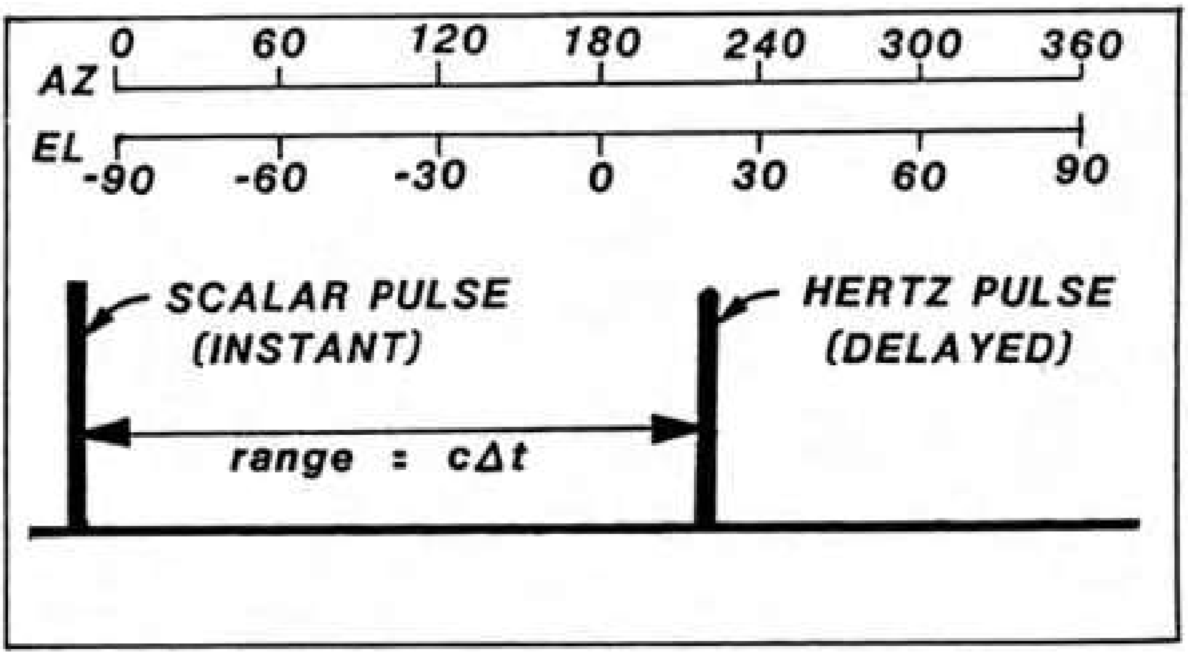

Here we show diagrammatically what a modified Soviet integrated weapons display for a modified jammer might be, to employ scalar interferometer weaponry.

As the "main bang" is emerging from the antenna of the
target radar on the ground, the phi-field from it reaches to infinity, including through the aircraft with a finite magnitude. In the a1rcraft, two channels act in interferometer fashion as a scalar detector for phi-dot. The main bang creation (phi-dot) is detected, and displayed as the leftmost pulse.

At that time, a computer time-base starts clocking off microseconds.

At some time later, the third channel, acting as a normal receiver, receives the ordinary "Hertz" pulse from the radar, which has traveled to the aircraft at the speed of light.

The computer converts elapsed time between the scalar pulse and the Hertz pulse to range.

Azimuth and elevation angle to the radar are obtained by normal angle tracking.

At this time, the computer has the exact coordinates of the radar, allowing for motion of the aircraft during the lapse time.

The computer switches the interferometer into transmit mode, and arranges amplitude patterns to target the radar.

It fires the first scalar pulse, then fires the second so as to create the full EMP directly upon the target.

The EMP flash occurs, destroying the radar.

At that time, another target can be processed.

In a more advanced weapon, scanning may be used to detect, locate, target, and engage multiple targets, including incoming air-to-air and surface-to-air missiles, one after another.

The same weapon is usable against a variety of aerial and ground based targets.

The major weapon is the electronic emitter, not a missile or ordinary ordnance.

Hordes of such special "jammer aircraft" now become a potent force on the battlefield, destroying all before them with annihilating ray weapons.

Star war is a reality now.

The High Frontier is already seriously obsolete.

Our entire defense establishment is completely vulnerable to these weapons at present, for we have not developed countermeasures or defenses.

Next Slide

The Tom Bearden Website

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

SLIDE 56

RADAR INVISIBILITY

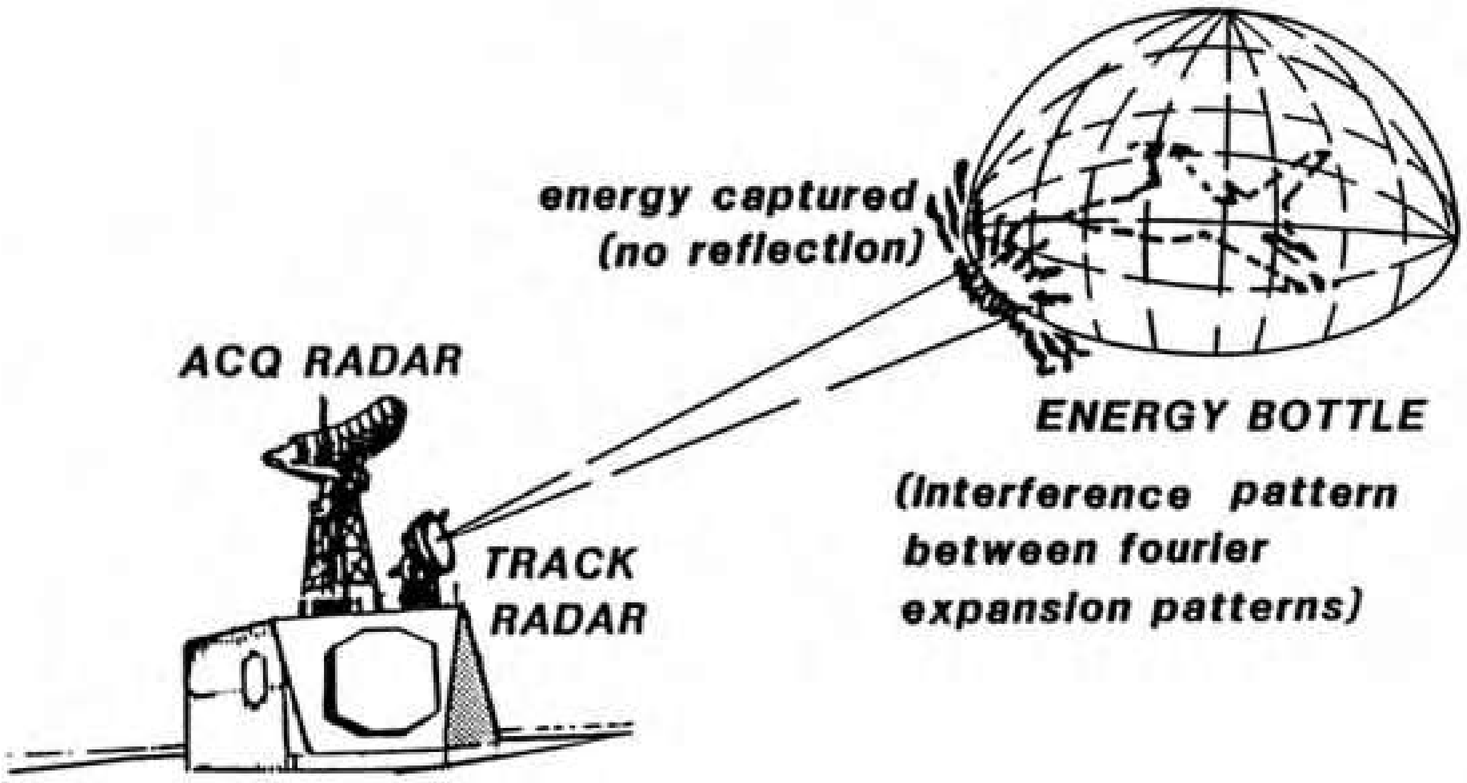
energy captured
(no reflection)
ACQ RADAR
ENERGY BOTTLE
(interference pattern
between fourier
expansion patterns)
TRACK
RADAR

On this slide we show another effect which can be obtained. We can make an airplane invisible to ordinary radar.

Imagine now the same "jammer" type aircraft with three channels, one for use in scalar interferometry and one as an ordinary channel.

This time, a Fourier interference pattern is established around the aircraft, so this establishes an "energy bottle" around the plane. We specifically arrange this bottle in the frequency band of the hostile radar we wish to avoid. For multiple bands, we establish multiple bottles -- this just requires more frequencies or more translator/transmitters.

The beam from the ground radar now strikes the energy bottle, and its energy enters and is trapped in the bottle grids. Since the bottle is leaky, the incident energy gradually disperses and "leaks" out all around, with very little emerging energy.

The ground radar receives no return pulse or reflection.

Therefore it sees no target.

The same interferometer system is usable both in this defensive screen fashion and in the attack fashion previously discussed.

Use of this system has been demonstrated internationally.

An energy bottle has been created in a North American laboratory as well.

Next Slide

The Tom Bearden Website

SLIDE 57.

NETWORK OF VIRTUAL TRANSMITTERS

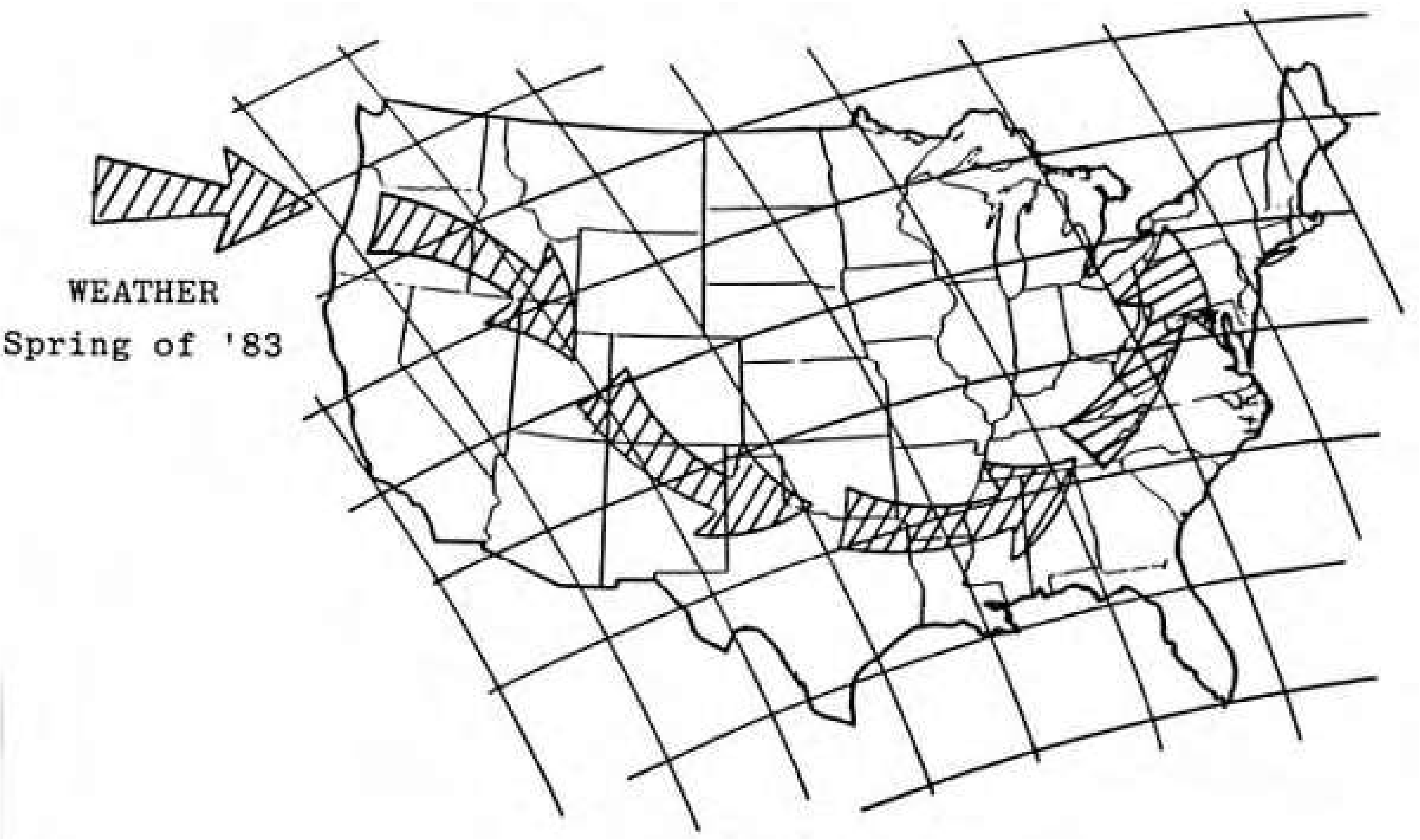

As is well-known, for years the Soviets have blanketed the world with large, powerful transmitter signals in the communications band, referred to as the "Woodpecker" signals because of the staccato noise the chirped signals make to the ear. The carriers have been modulated with known harmful bioactive ELF frequencies. At least some effects have been noted on humans from these signals, and other human effects are suspected but not proven.

In addition, evidence exists that such transmitters, utilized in the scalar transmission mode, may have been used to cause earthquakes and to control the weather. High altitude bursts off the coast of the Eastern United States and Canada, over

Timmons, Ontario, Canada, and off the coast from St. Petersburg, Florida have occurred also.

Mysterious "blackouts" of U. S .satellites -- and British satellites -- have occurred. These would be difficult to explain as the result of illumination from a ground laser, but easily explained as the result of illumination from a ground-based scalar beam or scalar interferometer.

Strange patterns of micro sky quakes and micro earthquakes occurred at various locations throughout the U.S. just before and shortly after the death of Leonid Brezhnev. These strongly appear to be connected with the adjustment of a scalar interference pattern -- a gigantic "energy bottle" across North America, as shown on this slide.

Shortly after this suspected adjustment and alignment, the underground net in the U.S. detected the emergence of numerous ELF-modulated signal carriers throughout the U.S., generally utilizing carriers in the vicinity of 40 MHz. Locations of the apparent transmitters were such places as Atlanta, North Georgia, St. Petersburg, Florida; Mississippi; Alabama; in the vicinity of Washington, D.C.; near Boston; around the Great Lakes; Michigan; in the Pacific Northwest (Washington and Oregon); and in both northern and southern California.

Researchers knowledgeable of the Soviet woodpecker signals and ELF modulations detected and reported these signals.

Physical effects were reported and correlated to the signals, including wakefulness at night, restlessness, fatigue, headaches, nausea, ringing in the ears, anxiety, disturbed dreams, etc.

What appears to have occurred is the adjustment and

alignment of a vast scalar interference pattern across the U.S. Recall that energy emerges in each grid, by phase interference between the composite substructure waves of the scalar beams.

Thus the Woodpecker "over-the-horizon radars," as accepted

by Western intelligence analysts, may have acquired a diabolical new strategic capability.

It is as if the Soviets had been permitted to build a vast network of transmitters directly on U.S. soil, for hostile employment against U.S. citizens.

Shortly after the net was aligned, a curious pattern of weather assaults on the U.S. began.

Storms assaulted California unceasingly. Then they roared across the Rockies, dipping sharply down to hit the Texas Panhandle, and bending northeast again to assault the South and the East.

The mighty jetstreams themselves were diverted and "kinked" into this looped, distorted pattern.

And the storms and floods assaulted the U.S. one after the other in this highly unusual fashion.

I maintain that the virtual transmitter net, just established by the Soviets and adjusted in on the U.S., was tested in the weather control mode. If so, it proved highly effective. Imagine what would have happened to the U.S. if it had been unleashed full force, and not just "gently tested."

And imagine what happens when disease, death, and nervous system disturbances are modulated upon these blanketing signals. The "transmitters" for the patterns are in each grid, all across the U.S.! And there are no physical transmitters there to destroy.

A counter is needed quickly. A simple one would consist of recording the signals in each grid zone, analyzing the substructure, reversing the signals and their structured patterns 180 degrees, and feeding the "cancellation signals" out in each area.

A great strategic Russian weapon of devastating power and potential is now operating directly on U.S. soil. The need for a crash countermeasures program is urgent.

In 1975 Brezhnev had introduced to the SALT talks the
strange proposal that we should also consider outlawing the development of new weapons more frightful than the mind of man
had ever imagined. We didn't know what the Russians were talking about. When we asked them later what weapons they meant, they referred vaguely to weapons of an electromagnetic nature.

The presentation I have just given you is part of what Brezhnev meant.

Next Slide

The Tom Bearden Website

SLIDE 58.

CONCLUSION

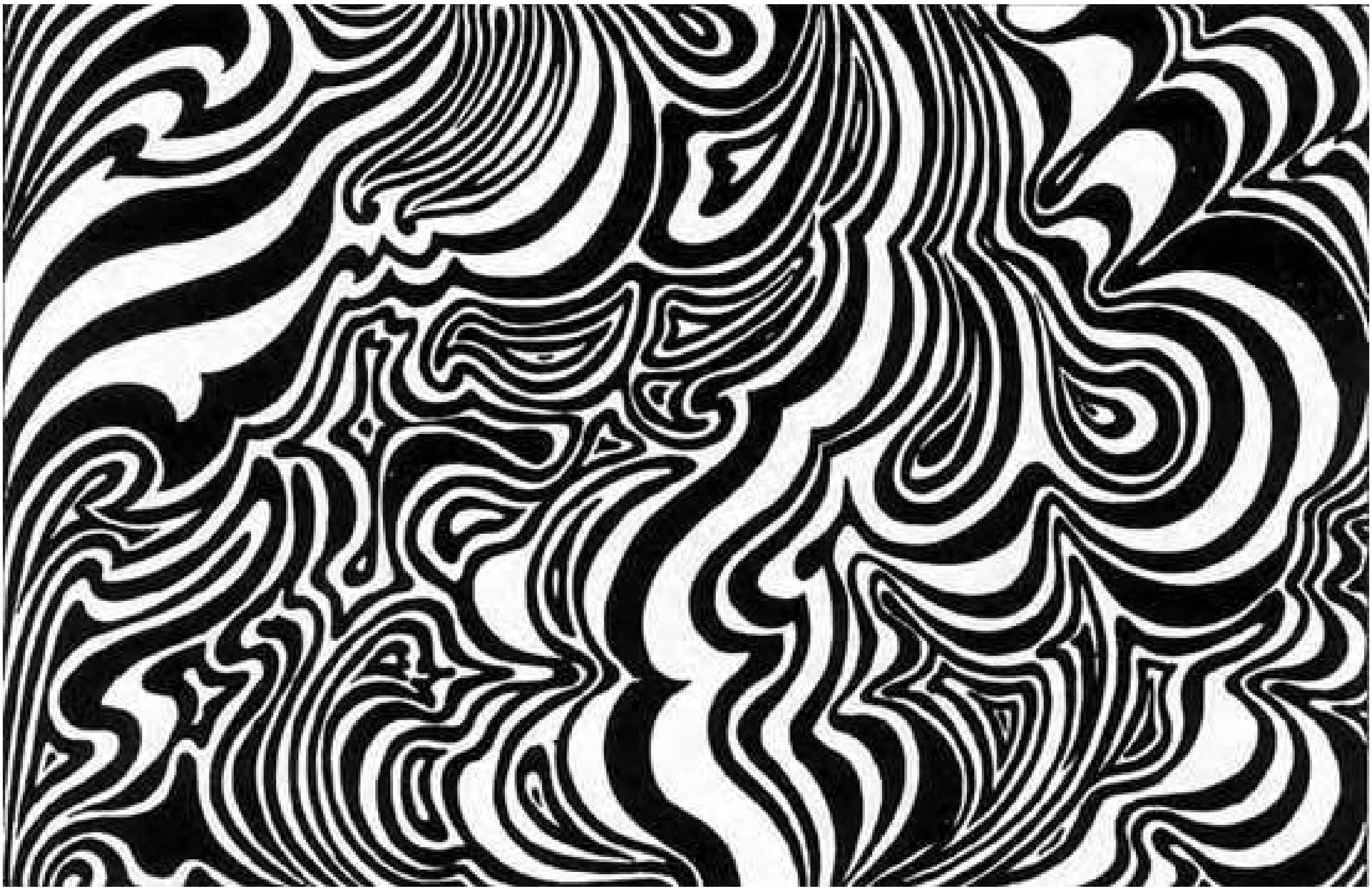

In conclusion, today I have released a number of powerful mechanisms that result from the new technological concepts. Many of these were originally discovered -- at least in rudimentary form -- by Nikola Tesla. In the West, the suppression of Tesla led to the suppression of these effects. And our knowledge of them went to the grave with Tesla.

After World War II, the Soviet Union mounted a massive campaign to obtain all the scientific literature of the West and
thoroughly digest it. Great centers were set up, and thousands

of PhD's devoted to this purpose. Gradually they synthesized
this technology from the errors in EM foundations they uncovered, and from obscure, ignored papers published in the orthodox
Western scientific literature.

They rediscovered the technology of Tesla, Moray, and
Rogers. And they have highly weaponized the effects, and adjusted and aligned the weapons upon us.

Now the radical movements are onto part of this technology. As late as about six months ago, one such group discovered how to build a translator and a scalar beam projector. Excitedly, they planned to build them the size of a large pistol, powered by a portable battery pack, and sell them to street criminals to kill or stop the police. They planned to get rich by this illicit means. At that time, a personal friend of mine, who moves in underground circles, was attempting to purchase one of the prototypes for my analysis. Fortunately, the inventor stepped into the beam of one of his devices and was almost killed. This frightened the group, and at least for now they have desisted. But be forewarned: this technology is coming down the pipe, and the police establishments in the West are in for the shock of their lives if the scientific community does not wake up and prepare defenses.

Again I urge my fellow colleagues to experiment vigorously with these new concepts, but build equipment for the benefit of man. Insect control, disease elimination and control -- both are noble purposes. Defense against the hostile use of such devices is also a fitting purpose.

We are in the final days, and the final technology is upon us. It is a stupendous tool for either good or bad, and assuredly it will be utilized for both means. I call attention once again to the fact that my ultimate purpose is to use this technology benevolently to achieve mind linkage, and elevate man from his present brutish state to the lofty state that is his destiny and his true God-nature.

For better or worse, Tesla technology is once again loose in the world, this time for good, and on a massive scale.

Let us prepare ourselves for a wild canoe ride, for the real white water lies just ahead.

Thank you for your attention, and good evening.

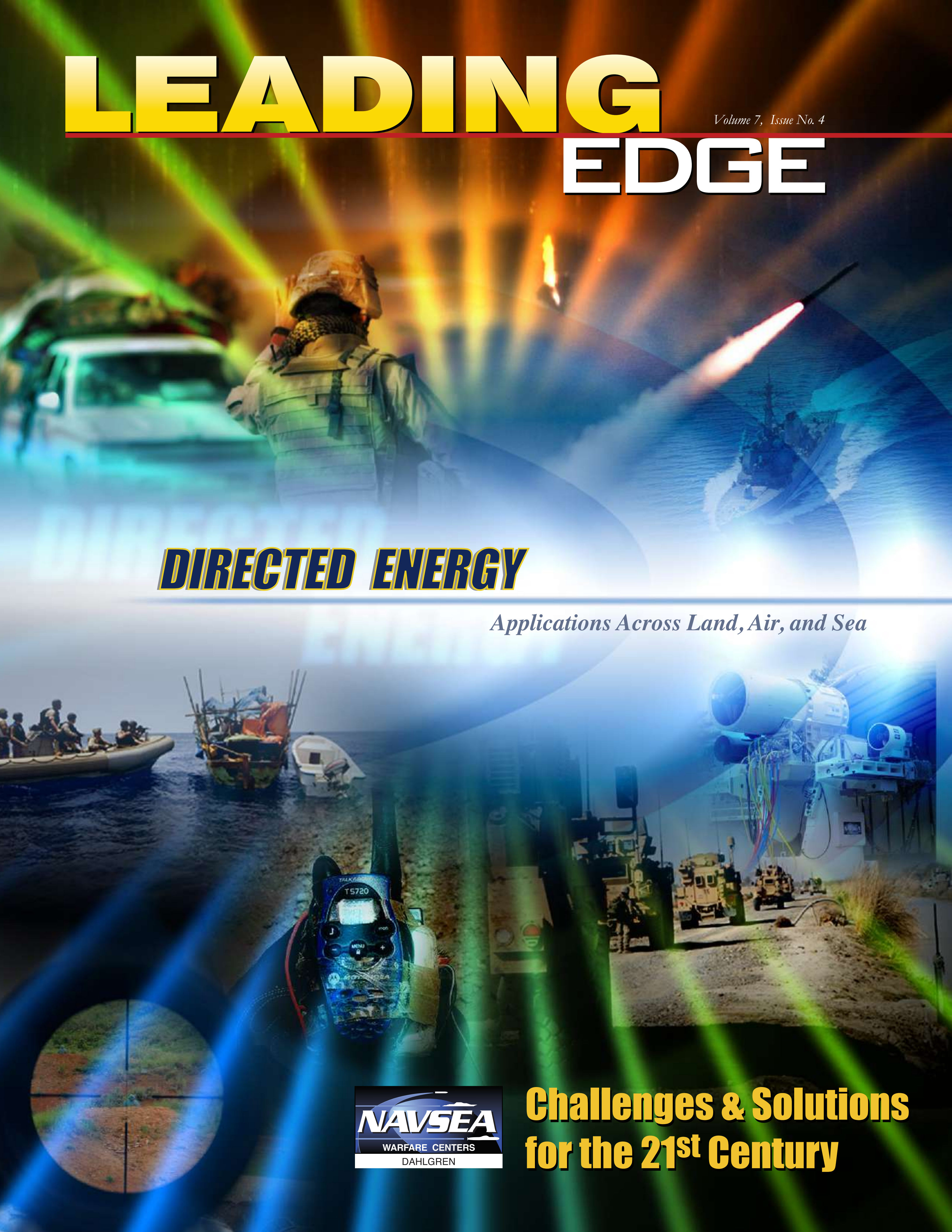
TALKABOUT
T5720
MENU
MOTOROLA

LEADING EDGE

Volume 7, Issue No. 4

DIRECTED ENERGY

Applications Across Land, Air, and Sea

NAVSEA
WARFARE CENTERS
DAHLGREN

Challenges & Solutions for the 21st Century

LEADING EDGE
Volume 7, Issue No. 4

TABLE OF CONTENTS

Directed Energy

3 LEADING EDGE DIRECTED-ENERGY RDT&E
Captain Michael H. Smith

4 DIRECTED-ENERGY TOPICS IN THIS ISSUE
Mr. Dale Sisson

PAST, PRESENT, AND FUTURE

6 NAVAL DIRECTED-ENERGY WEAPONS – NO LONGER A FUTURE WEAPON CONCEPT
David C. Stoudt

12 HISTORICAL OVERVIEW OF DIRECTED-ENERGY WORK AT DAHLGREN
Stuart Moran

26 HISTORY OF LASER WEAPON RESEARCH
Melissa Olson

36 LASER WEAPON SYSTEM (LAWS) ADJUNCT TO THE CLOSE-IN WEAPON SYSTEM (CIWS)
Robin Staton and Robert Pawlak

RDT&E, ACQUISITION, AND WARFARE MANAGEMENT

44 THE ACQUISITION CHALLENGE ASSOCIATED WITH DIRECTED-ENERGY RDT&E
Mike Kotzian

TECHNOLOGY, MODELING, AND ASSESSMENT

50 THE BASICS OF ELECTRIC WEAPONS AND PULSED-POWER TECHNOLOGIES
Stuart Moran

58 SOLID MODELING OF DIRECTED-ENERGY SYSTEMS
Joseph F. Sharrow

64 A FUNDAMENTAL KEY TO NEXT-GENERATION DIRECTED-ENERGY SYSTEMS
Directed Energy Division, Electromagnetic and Sensor Systems Department

70 ACTIVE DENIAL ARRAY
Randy Woods and Matthew Ketner

74 DIRECTED ENERGY IN THE MILITARY ENVIRONMENT
LT Leedjia Svec, Jeremy Beer, and Dave Freeman

HIGH-POWER MICROWAVE

78 DIRECTED ENERGY USING HIGH-POWER MICROWAVE TECHNOLOGY
Jacob Walker and Matthew McQuage

HIGH-ENERGY LASER

82 LASER COUNTER ROCKET, ARTILLERY, AND MORTAR (C-RAM) EFFORTS
Michael Libeau

NONLETHAL CAPABILITIES

86 MULTIFREQUENCY RADIO-FREQUENCY (RF) VEHICLE STOPPER
Stephen A. Merryman

92 HIGH-POWER ELECTRICAL VEHICLE-STOPPING SYSTEMS
Jordan Chaparro and Melanie Everton

96 NONLETHAL SMALL-VESSEL STOPPING WITH HIGH-POWER MICROWAVE TECHNOLOGY
Jacob Walker

LEADING EDGE

Naval Surface Warfare Center, Dahlgren Division (NSWCDD)
Captain Michael H. Smith, *Commander*
Carl R. Siel, Jr., *Technical Director*
David C. Stoudt, *Distinguished Engineer for Directed Energy (ST) and NAVSEA Technical Warrant for Directed Energy and Electric Weapon Systems*
Janice Miller, *Corporate Communications Director (Acting)*
Steve Zehring, *Managing Editor*
Margie Stevens, *Production Coordinator*
Patrice Waits, *Editor & Layout*
Clement Bryant, *Layout Design & Graphic Artist*
Kellie Yeatman, *Graphic Artist*
Trey Hamlet, *Graphic Artist/3-D Modeling*

Electromagnetic & Sensor Systems
Dale Sisson, *Head, Q Department*
Brandy Anderson, *Graphic Artist/Cover Design*

NSWCDD
Jordan Chaparro
Melanie Everton
Matthew Ketner
Michael Libeau
Matthew McQuage
Stephen A. Merryman
Stuart Moran
Melissa Olson
Robert Pawlak
Joseph F. Sharrow
Robin Staton
Jacob Walker
Randy Woods

Defense Acquisition University
Mike Kotzian

Naval Medical Research Unit – San Antonio
Jeremy Beer
Dave Freeman
LT Leedjia Svec

The Leading Edge *magazine is produced by the Naval Surface Warfare Center, Dahlgren, Virginia. The purpose of the publication is to showcase technical excellence across the Warfare Centers and promote a broader awareness of the breadth and depth of knowledge and support available to the Navy and DoD.*

Address all correspondence to Corporate Communications, C6
Email: dlgr_nswc_c6@navy.mil; or write to
Commander
Naval Surface Warfare Center, Dahlgren Division
Corporate Communications, C6
6149 Welsh Road, Suite 239
Dahlgren, VA 22448-5130

NSWCDD/MP-09/34
Approved for public release; distribution is unlimited.

NAVSEA
WARFARE CENTERS
DAHLGREN

Introduction

Leading Edge Directed-Energy RDT&E

Captain Michael H. Smith, USN
Commander, NSWCDD

Dahlgren first launched research and development efforts devoted to harnessing the power of electromagnetic energy over 40 years ago. From early work with voltage multipliers and pulse-powered technology, to today's high-energy lasers and high-power microwave technologies, the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) has led, and continues to lead, cutting-edge directed-energy research, development, testing, and evaluation. Our commitment in this area only grows stronger—evidenced by our chartering of the Directed Energy Warfare Office (DEWO)—in order to provide increased focus on warfighting applications of these technologies.

Today's military forces face a wide array of challenges in diverse operating environments around the world. Directed energy offers unique and flexible options to address today's challenges, as traditional kinetic weapons are often of limited value in peacekeeping missions and in urban environments, where restricted rules of engagement typify the norm. Kinetic weapons can also be more costly or ineffective to employ against asymmetric threats.

The Chief of Naval Operations (CNO) recently placed added emphasis on directed energy and on expanding the range of directed-energy capabilities. In response, scientists and engineers at NSWCDD are actively developing prototype systems in a number of areas that you will read about in this issue—areas that have been successfully demonstrated and tested in our Navy laboratories and ranges.

In this issue of *The Leading Edge* magazine, you will trace the rich history of directed-energy work at Dahlgren, gain insight into directed-energy weapons already fielded or being readied for the field, and learn about prototypes that show real promise for providing incredibly effective offensive and defensive directed-energy solutions. For example, scientists and engineers at NSWCDD are leading the way toward realizing small, lightweight radio frequency (RF) transmitters using high-power, solid-state switching amplifiers for the development of counter-improvised explosive device detection and neutralization systems. You will also learn about diverse applications of directed-energy technology—such as research and testing of laser glare devices and laser eye protection—and have the opportunity to gain a better understanding of the Department of Defense (DoD) acquisition framework and the challenge of maintaining cost and schedule estimates while delivering weapons systems that are critical to the warfighter.

From lasers to high-power electrical vehicle-stopping systems, I am sure you will be fascinated and, along with me, be impressed with the advancements our scientists, engineers, and technical staff are achieving in the directed-energy arena to support of our men and women in uniform.

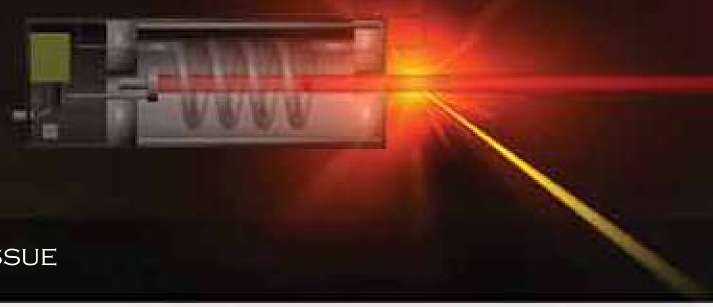

DIRECTED-ENERGY TOPICS IN THIS ISSUE

Dale Sisson
Head, Electromagnetic and
Sensor Systems Department
NSWCDD Dahlgren, Virginia

Welcome to our Directed Energy issue of the *Leading Edge* magazine. This issue represents the third in a trilogy of issues covering the truly fascinating and incredibly challenging area of naval warfare in the operational electromagnetic environment. In our first issue, we covered the full range of operational and readiness implications when operating in the electromagnetic environment. Then, in our second issue, we highlighted the complexities and dynamics of providing relevant and effective sensors and radars to our warfighters. Now, we focus on directed energy and relate how the Naval Sea Systems Command (NAVSEA) Warfare Centers, and the Naval Surface Warfare Center, Dahlgren Division's (NSWCDD's), in particular, are working on state-of-the-art directed-energy weapons capabilities for the warfighter.

In this issue, we first look back to the early years, decades ago, when directed-energy weapons research began. We examine the history of directed energy, and we cover significant discoveries and achievements made by NAVSEA Warfare Center scientists and engineers, and others in the scientific community. We then relate information about several of our current directed-energy initiatives, and about how we're working hard to solve some of the most complex technical challenges associated with directed-energy weapons. We highlight how others in the Navy, such as the Naval Medical Research Unit in San Antonio, Texas, are also conducting research into directed energy and how our forces can better protect themselves from the effects of directed energy. We show how directed energy can be employed in a variety of offensive and defensive, lethal and nonlethal situations. We explain how directed-energy weapons work and how they can be employed in various environments against a wide range of situations. Lastly, we look forward as we provide technical and strategic leadership for the efficient and effective development, acquisition, and fielding of directed-energy systems for the warfighter.

So, if you want to learn about what the NAVSEA Warfare Centers and others in the Navy are doing in the area of directed-energy weapons, look no further than this issue of the *Leading Edge* magazine. I'm confident that you will be impressed by the progress made in this most important technology field.

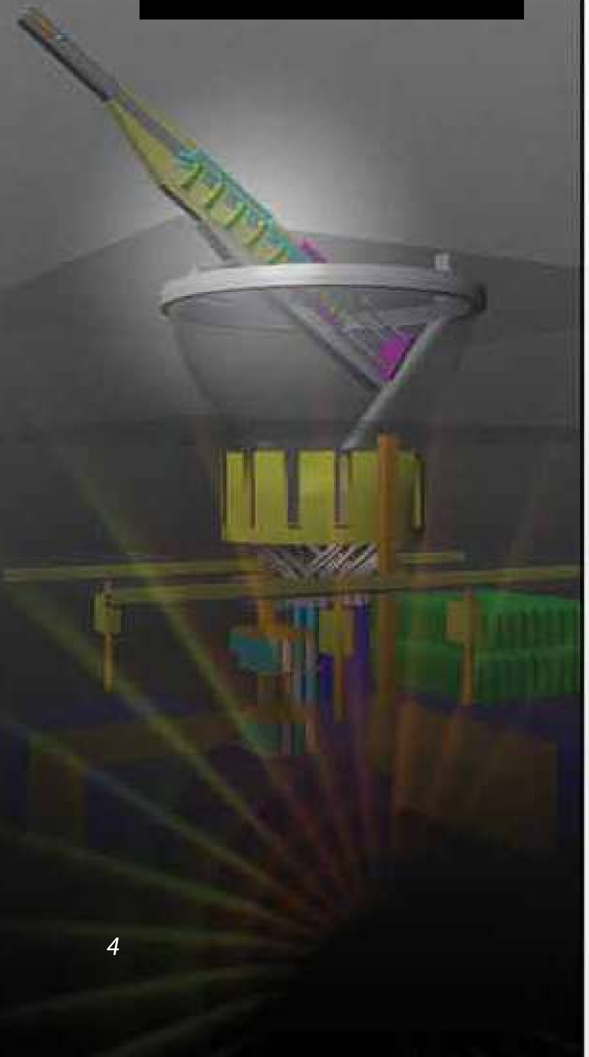

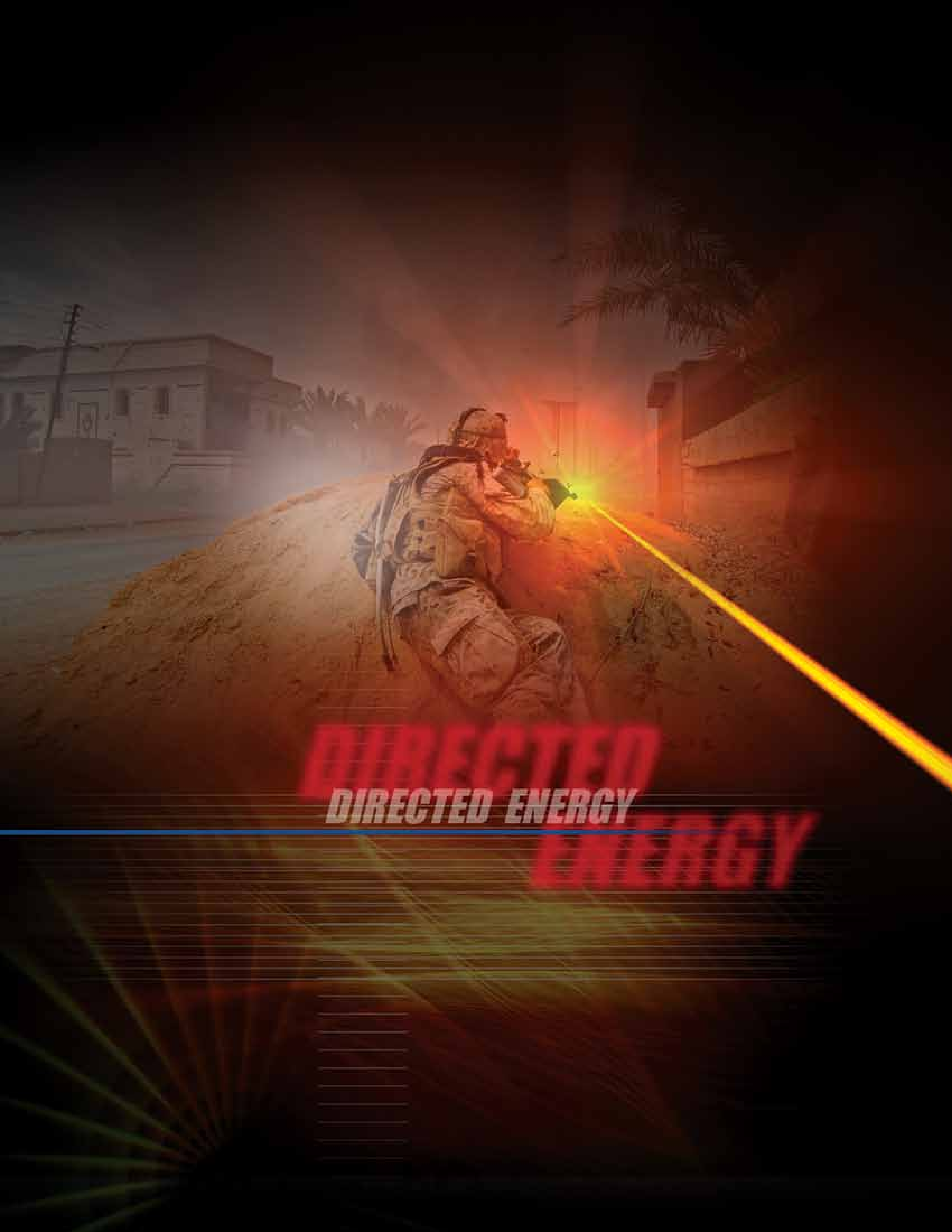

DIRECTED ENERGY

NAVAL DIRECTED-ENERGY WEAPONS – NO LONGER A FUTURE WEAPON CONCEPT

By David C. Stoudt

Dr. Stoudt is the Distinguished Engineer for Directed Energy (ST) and the NAVSEA Technical Warrant for Directed Energy and Electric Weapon Systems.

Directed-energy weapon (DEW) technologies typically take the form of high-energy lasers (HELs), high-power microwaves (HPMs), and charged-particle beams. This article focuses on the first two technology areas, as they have reached the point of being ready for operational testing and evaluation, and in some cases, operational use on the battlefield. DEWs have been popularized in science-fiction writings for over a hundred years. The Department of Defense (DoD) has been investing in their development since the 1970s. This article will not go into technical depth regarding the various directed-energy (DE)-related efforts currently underway in the Navy, but rather, it will overview DE areas under development and relate recent Navy leadership activity. Other articles in this issue of *The Leading Edge* magazine will provide the reader with much greater technical and programmatic details on various DE efforts.

HIGH-ENERGY LASER WEAPONS

HEL weapon systems have been envisioned for a great many years, to include being referred to as Martian "Heat Ray" weapons in H.G. Wells' epic novel *The War of the Worlds*, originally published in 1898. In reality, a high-average-power laser weapon system is very similar to a "heat ray", or even a blow torch. During the early years of DoD investments in DE technology, the Navy led the development of HEL with the creation of the world's first megawatt-class, continuous-wave, Mid-Infrared Advanced Chemical Laser (MIRACL), located at White Sands Missile Range (WSMR). Roughly 80 years after the work of H.G. Wells, the U.S. Navy tested the MIRACL laser and ultimately used that laser system to engage static and aerial targets in the desert of WSMR in the following years. While that laser proved to be the wrong choice for the Surface Navy's self-defense mission, it did spawn work by the Air Force on the Airborne Laser (ABL), and the Army on the Tactical High-Energy Laser (THEL). In 2000 and 2001, the THEL successfully shot down 28 supersonic Katyusha artillery rockets and 5 artillery shells.

In 2010, the ABL successfully engaged and destroyed tactical ballistic missiles during the boost phase of their flight. All three of these laser systems—the MIRACL, the ABL, and the THEL—are chemical lasers that utilize toxic chemicals and operate in less than optimal wavelengths that make them a poor choice for most naval applications. The MIRACL is shown in Figure 1.

Recent advances in solid-state lasers, to include fiber lasers, have moved these electric lasers to the forefront of the Department's research and development (R&D) for near-term HEL applications in the services. The Navy has particular interest in electric lasers, to include the free-electron laser (FEL), for shipboard self-defense and force protection applications. The speed-of-light delivery of HEL energy can defeat the high-g maneuvers of newly developed foreign antiship cruise missiles (ASCMs). Thus, the Office of Naval Research (ONR) started an FEL Innovative Naval Prototype (INP) program in FY10, with a goal of reaching the output power of 100 kW. The eventual goal of the FEL program is to reach the multi-megawatt power level with wavelength selectivity. The Naval Sea Systems Command (NAVSEA) Directed Energy and Electric Weapons Program Office (PMS 405) has been actively developing a fiber laser-based Laser Weapon System (LaWS) that could be a retrofit to augment the current capabilities of the Close-In Weapon System (CIWS) currently deployed on many surface combatants. The Naval Surface Warfare Center, Dahlgren Division (NSWCDD), is the Technical Direction Agent and lead system integrator for PMS 405 on the LaWS program. The Naval Air Systems Command (NAVAIR) has interest in compact, solid-state HEL systems for aircraft self-protect and air-to-ground engagements, and will be starting a

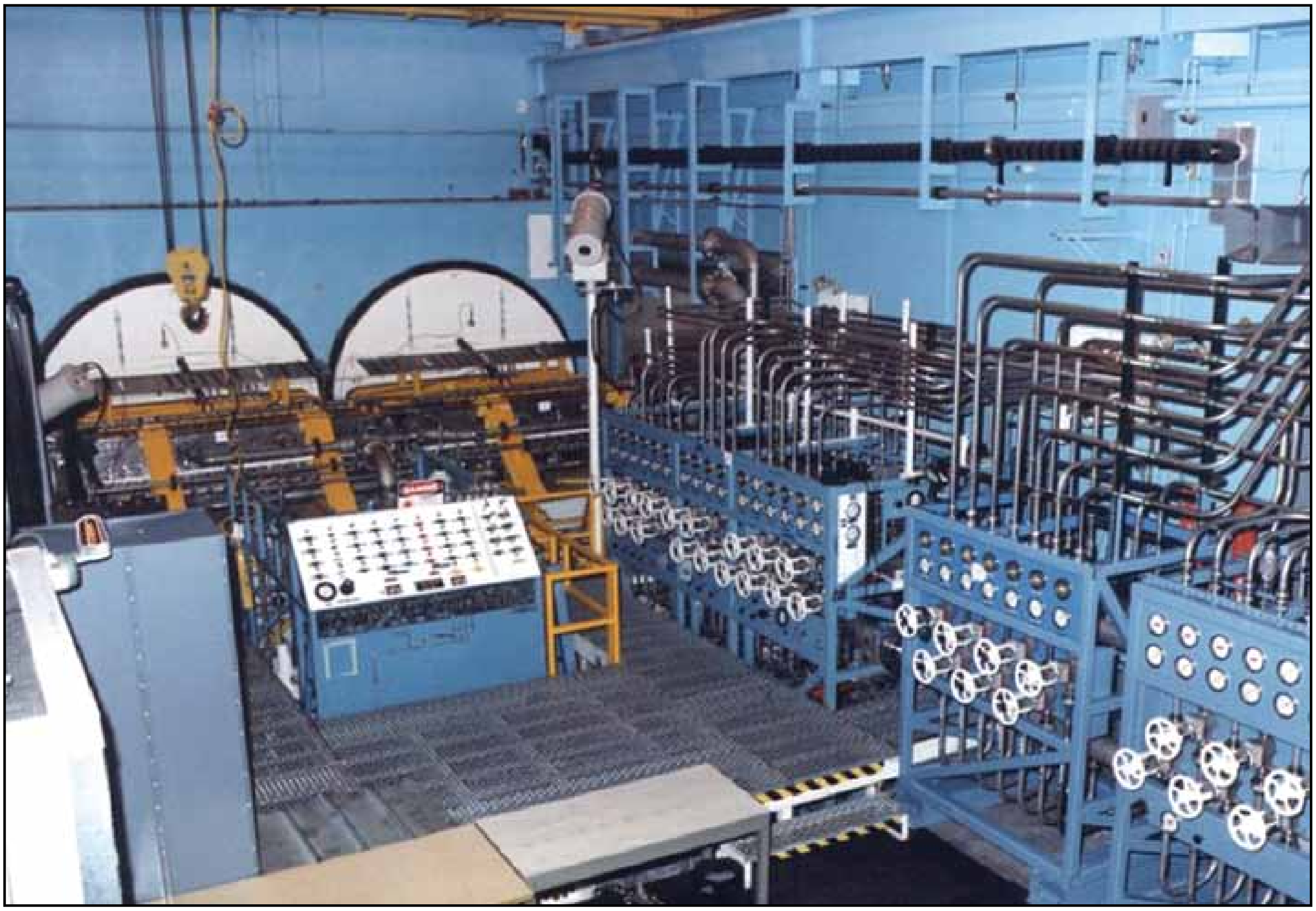

Figure 1. Mid-Infrared Advanced Chemical Laser (MIRACL)

fiber laser-based ONR Future Naval Capability effort in FY12. LaWS is shown in Figure 2.

High-Power Microwave Weapons

Like lasers, microwave weapons have been fantasized about ever since the invention of microwave power generators. In fact, in 1932 it was generally recognized by the British government that bombers, ostensibly German bombers, would be able to penetrate British air space and bomb its civilian population and infrastructures. In 1934, the Air Ministry initially asked Robert Watson-Watt, of the National Physical Laboratory, if he could build a "death ray" that could kill enemy pilots or detonate bombs while they are still on the planes of enemy aircraft. Such a "death ray" had been proposed to the Air Ministry by Harry Gindell-Mathews 10 years earlier in 1924. Watson-Watt, a former meteorologist who had become an expert on radio signals, suggested that energy reflected from an aircraft could be used to locate it. His experiments were successful and RADAR (radio detection and ranging), a name coined by the U.S. Navy in 1940, was born. While RADAR is not a DEW in the way they are thought of today, its roots can clearly be traced to the military's desire for such capabilities.

The Navy's HPM, or high-power radio-frequency (RF) systems, have been progressively increasing in power density to the point where it is now feasible to integrate the technology into weapon systems for deployment. While initial HPM applications suffered from their inability to obtain militarily useful outcomes, either due to technology limitations, difficult concept of operations (CONOPS), or inherent robustness of potential target systems, many feasible military applications for using HPM

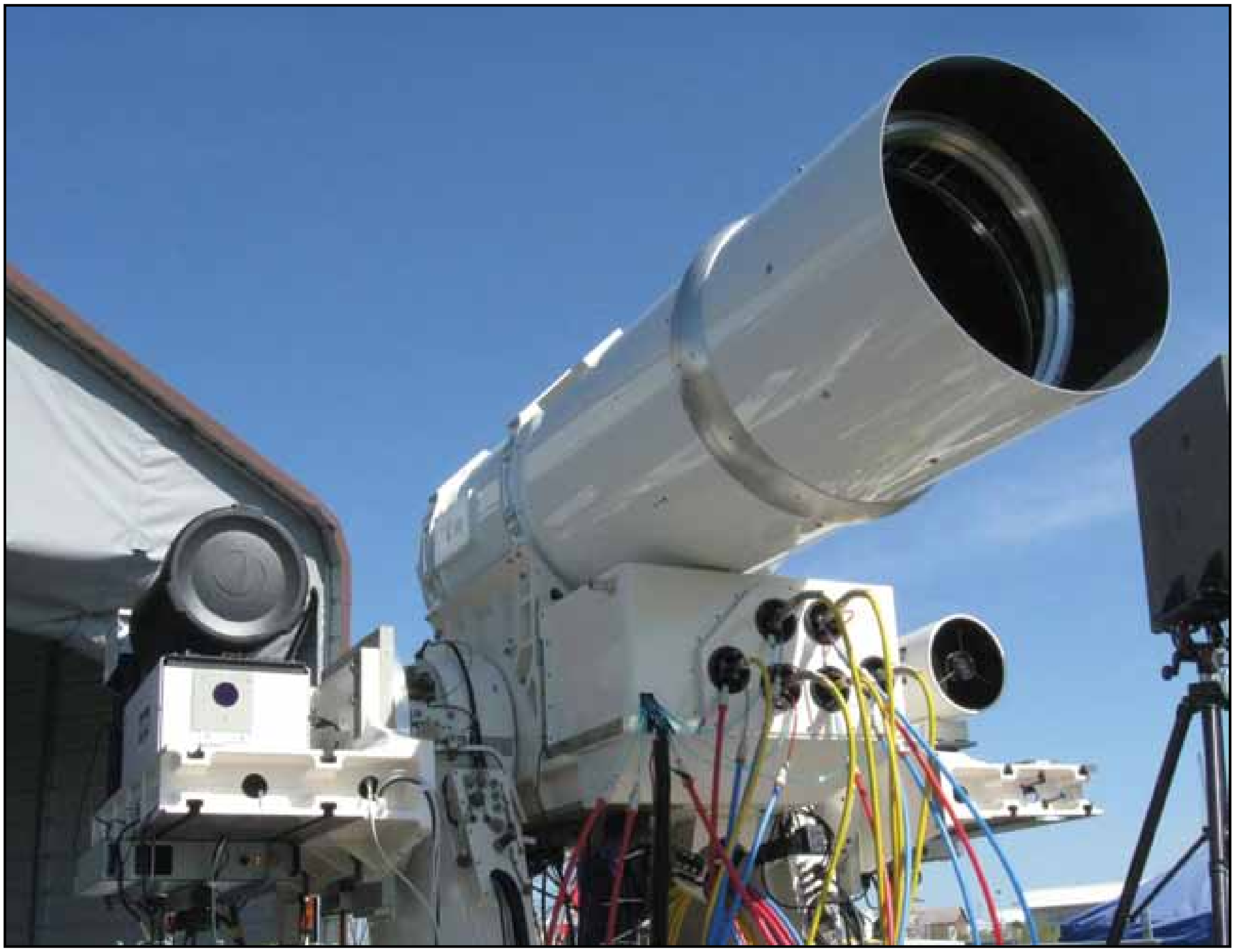

Figure 2. Laser Weapon System (LaWS)

devices have surfaced over recent years to include nonlethal, antipersonnel weapons and nonkinetic, antimateriel weapons. While these concepts offer unique capabilities to the warfighter due to the nonkinetic effects they generate, other warfighting concepts—such as stopping vehicles, or countering hidden roadside bombs or improvised explosive devices (IEDs)—are difficult to achieve by any other means. The multifrequency Radio-Frequency Vehicle Stopper (RFVS) system is shown in Figure 3.

In addition, the difficulty in overcoming the propagation losses associated with HPM has driven some concepts into platforms such as unmanned aerial vehicles (UAVs) or cruise missiles that deliver the HPM device to the target for a close-in engagement. Over the past 10 years, field-testable prototypes have been developed to demonstrate the operational utility of these concepts, and in some cases, those prototypes have or will be deployed operationally to support our troops in theater. It is only through the hard work and perseverance of the Naval Research Enterprise (NRE), as well as other DoD laboratories, that concepts that were once only laboratory curiosities are now making their way onto the battlefield and contributing to the fight.

FOREIGN DIRECTED-ENERGY WEAPON (DEW) DEVELOPMENT

While the United States has been very active in this warfighting area, significant foreign DEW development also has elevated the need for the Navy to afford these threats a higher priority. This can be done either by incorporating the necessary DEW countermeasures into weapon systems, platforms, and critical infrastructures, or by adapting the CONOPS and tactics, techniques, and procedures (TTPs) employed by our armed forces to properly account for those foreign DEW systems. Materiel developers need to understand how this threat

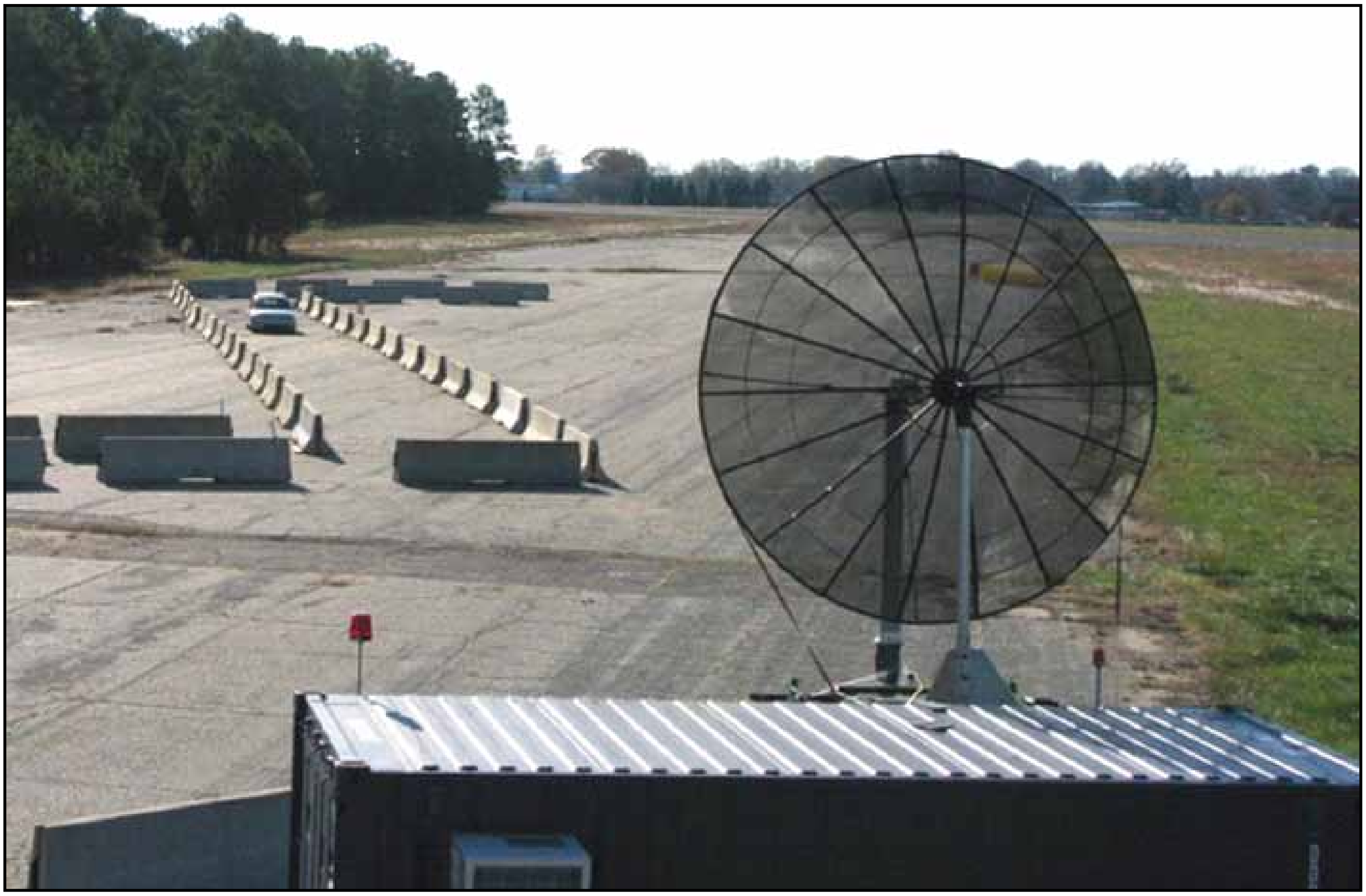

Figure 3. Multifrequency RF Vehicle-Stopper (RFVS) System

is evolving and properly address it during the design of their systems. They also need to address DE in the development of their system threat assessments. There has been movement on the HPM side to modify existing military standards, such as MIL STD 464[1] and others, to now include information on potential HPM threats. For example, in the HEL arena, work has been accomplished in the development of protective measures for eyes; however, this threat needs to be considered during the system development process. It is well known that building in countermeasures is much cheaper during the initial development of a system, vice trying to retrofit systems with countermeasures once a new threat is on the battlefield. As analysts evaluate the foreign development of DE technologies, and the trends become clearer, it is the responsibility of the acquisition community to take this threat into consideration and ensure that weapon systems, platforms, and infrastructures will be available and at full capability when needed. By accounting for foreign threat developments, assessing blue force susceptibilities and vulnerabilities, and adopting appropriate measures to negate or counter these threats, naval forces will avoid technological surprise on the battlefield in the future.

REQUIREMENTS

The DE programs briefly mentioned in this article, and covered more deeply in this and other publications, offer warfighters unique capabilities not currently found in their arsenal. The continuing problem, however, is matching those unique capabilities to vetted operational requirements. The DE technical community has made great strides in helping the operational community understand the capabilities of DE weapons and their potential military effects on targets. The lack of formal requirements, however, has yielded more of a technology push—rather than an operational pull—of various DE capabilities. Progress has been made, but more effort is required if DE capabilities are to be developed and transitioned between science and technology (S&T), and formal programs of record. Notwithstanding, the current outlook and trends are positive.

A RESURGENCE OF NAVY INTEREST IN DIRECTED ENERGY

The Navy's interest in DEWs for future maritime operations has increased in recent years due to a number of weapons development successes. Recognizing the importance and value of DEWs, NAVSEA reestablished the Navy Directed Energy Weapons Program Office (PMS 405) in 2004. Accordingly, PMS 405 was designated as the point of contact for matters related to DE and electric weapon systems (EWS) development and acquisition initiation for NAVSEA, and for matters being coordinated with other federal agencies and military services. PMS 405's mission is to transition technology from the laboratory to prototype/advanced development/testing for operational development and use.[2,3]

The Navy also established its first formal executive position for DE (ST-level), the Navy's Distinguished Engineer/Scientist for Directed Energy, at NSWCDD in August 2004. Following the establishment of this position, NAVSEA then formally established a Technical Authority Warrant for Directed Energy and Electric Weapon Systems (DE&EWS)—Surface Ships in July 2008. The scope of the warrant includes the transition of S&T development to weapon system development of lethal and nonlethal capabilities associated with the DE&EWS for Surface Ships.[4] This included, but was not limited to, the following:

- Laser Weapon Systems
  - High-Energy Lasers
  - Solid-State Lasers
  - Free-Electron Lasers
  - Femtosecond Ultrashort Pulse Lasers
  - Laser-Induced Plasma Channel
  - Lethality/Vulnerability
- Electromagnetic Rail Gun Weapon System
- High-Power Microwave
  - Active Denial System
  - Laser-Guided Energy
- Maritime Directed Energy Test Center
- Electromagnetic Launch of Weapons (excluding the Electromagnetic Aircraft Launch System (EMALS))

Then, within the NAVSEA Warfare Center Enterprise, Warfare Center leadership established two technical capabilities (TCs): an NSWCDD TC for DE systems research, development, test, and evaluation (RDT&E); and a Naval Surface Warfare Center, Port Hueneme Division (NSWCPHD) TC for in-service engineering, test and evaluation (T&E), and integrated logistics support to DE systems. NSWCDD leads all S&T and RDT&E for the development and weaponization of DE systems for surface, air, and ground environments. It also leads the development of offensive and defensive DE technologies needed to characterize and exploit vulnerabilities, provide weapons, and protect against attack. NSWCDD provides the technologies, devices, and systems designed to create or control electromagnetic energy used to cause persistent disruption or permanent damage by attacking target materials, electronics, optics, antennas, sensors, arrays, and personnel, including nonlethal applications. NSWCPHD provides in-service engineering, T&E, and integrated logistics support to DE systems throughout the system life cycle.

The Navy further demonstrated increased interest in DE when Assistant Secretary of the Navy (Research, Development & Acquisition (ASN(RDA)) designated NAVAIR offensive and defensive leads for naval aviation DE activities:

- Program Executive Officer for Unmanned Aviation and Strike Weapons (PEO(U&W)), assigned as the offensive DE lead for naval aviation
- PEO for Tactical Aircraft Programs (T), assigned as the defensive lead for naval aviation

Concerning future initiatives, the Chief of Naval Operations (CNO) tasked the Strategic Studies Group (SSG) to examine a topic entitled "Maritime Operations in the Age of Hypersonic and Directed-Energy Weapons."[5] The intent of the study was to provide Navy leadership with an understanding of where DE technologies and weapons are today and how they might influence future maritime operations. The theme of the study was completed during FY10, the results of which discuss many DE concepts, as well as tactics for the employment of DE capabilities. The study's findings are currently under review and consideration by senior Navy leadership.

CONCLUSION

While H.G. Wells' *The War of the Worlds* novel and television programs like *Star Trek* popularized the notion of using DE for weapons in years past, today— through persistent DEW RDT&E—Navy leadership is realizing the great potential that DEWs offer naval warfighters and homeland defenders. The scientific and technical advances the Navy has made in HEL and HPM in recent years have been nothing short of extraordinary. Moreover, future technological and engineering advances undoubtedly will result in profound differences in our nation's future warfighting capabilities. Naval DEWs, therefore, are no longer just a future weapon concept…they are here today.

REFERENCES

1. *Department of Defense Interface Standard, Electromagnetic Environmental Effects—Requirements for Systems*, MIL-STD-464, 18 March 1997.
2. NAVSEA Notice 5400, Ser 09B/240, Subject: Establishment of the Navy Directed Energy Weapons Program Office (PMS 405), dated 4 January 2002.
3. NAVSEA Instruction 5400.101, Ser SEA 06/058, Subject: Directed Energy and Electric Weapons Program Office (PMS 405) Charter, dated 21 July 2004.
4. NAVSEANOTE 5400 Ser 09B/240 of 4 January 2002 and NAVSEAINST 5400.101 Ser SEA 06/058 of 21 July 2004.
5. Letter from the CNO tasking the SSG XXIX Theme—"Maritime Operations in the Age of Hypersonic and Directed-Energy Weapons," dated 23 September 2009.

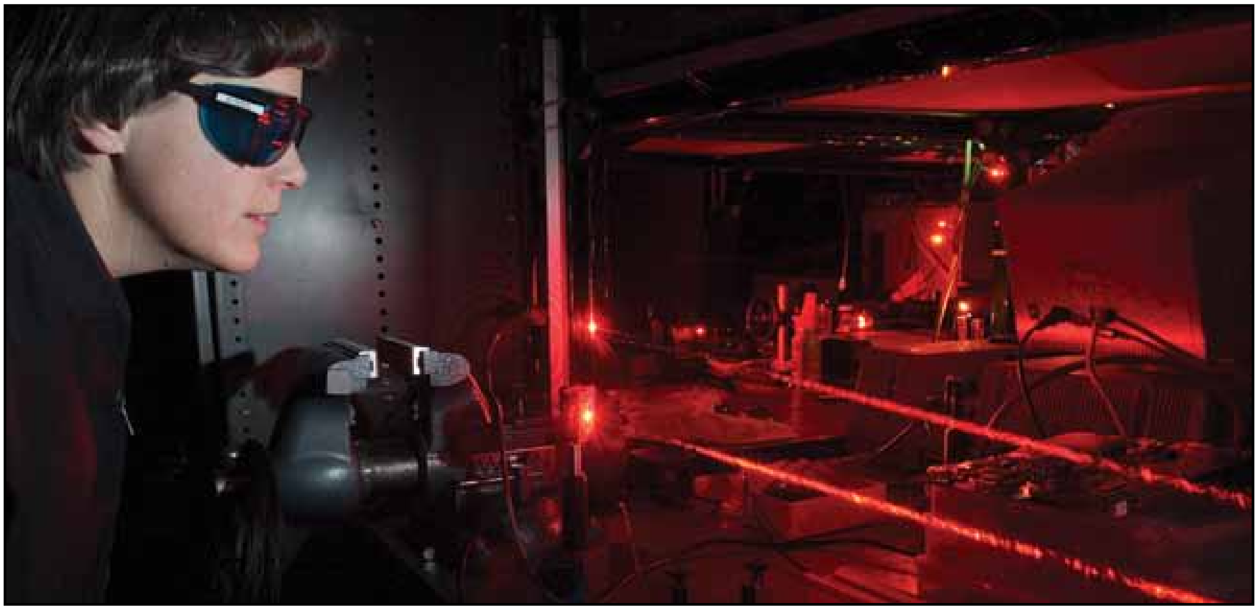

HISTORICAL OVERVIEW OF DIRECTED-ENERGY WORK AT DAHLGREN

By Stuart Moran

In 1962, the United States set off a megaton nuclear weapon 250 miles above the Pacific. The blast caused a large imbalance of electrons in the upper atmosphere that interacted with the Earth's magnetic field to create oscillating electric fields over a large area of the Pacific. These fields were strong enough to damage electronics in Hawaii, a thousand miles away, and clearly demonstrated the effects of an electromagnetic pulse (EMP). It didn't take long for the military to begin considering ways to create such pulses without using nuclear weapons.

In the late 1960s, the Special Applications Branch at the Naval Weapons Laboratory at Dahlgren began studying ways to generate high-power oscillating electric fields that could be used as a weapon to damage enemy electronics. These devices were basically high-power versions of the old spark-gap transmitters used in the early days of radio. To construct a device that could produce nuclear EMP-like fields, stored electrical energy was converted to radio-frequency (RF) energy that could be radiated from an antenna through the atmosphere to a target. These devices typically would store energy in a high-voltage capacitor and release the energy quickly using a spark-gap switch. This would then drive oscillating currents on an antenna, causing it to radiate. To achieve field strengths of thousands of volts per meter, typical of a nuclear EMP, devices operating at hundreds of thousands of volts or more were needed.

A number of radiating devices were studied in the early 1970s. Most belonged to a class of devices called Hertzian oscillators. A capacitor is charged to high voltage, the switch is closed, and current flows in the circuit, causing the stored energy to oscillate between the electric field of the capacitor and the magnetic field of the inductor. To charge the capacitor to extremely high voltages, a step-up transformer of some type must be used. One of the fastest voltage multipliers, the Marx generator, was frequently used. The losses from internal resistance and external radiation damp the oscillating waveform, typically after a few cycles. The radiated pulses are, therefore, short in time and broad in frequency content.[1] A simple diagram of the inductance-capacitance oscillator (L-C oscillator) is shown in Figure 1.

SINGLE-PULSE BURNOUT DEVICES

Many types of Hertzian devices were designed, constructed, and tested at Dahlgren during the 1970s. The transmission-line oscillator, or cavity oscillator, used a quarter-wavelength

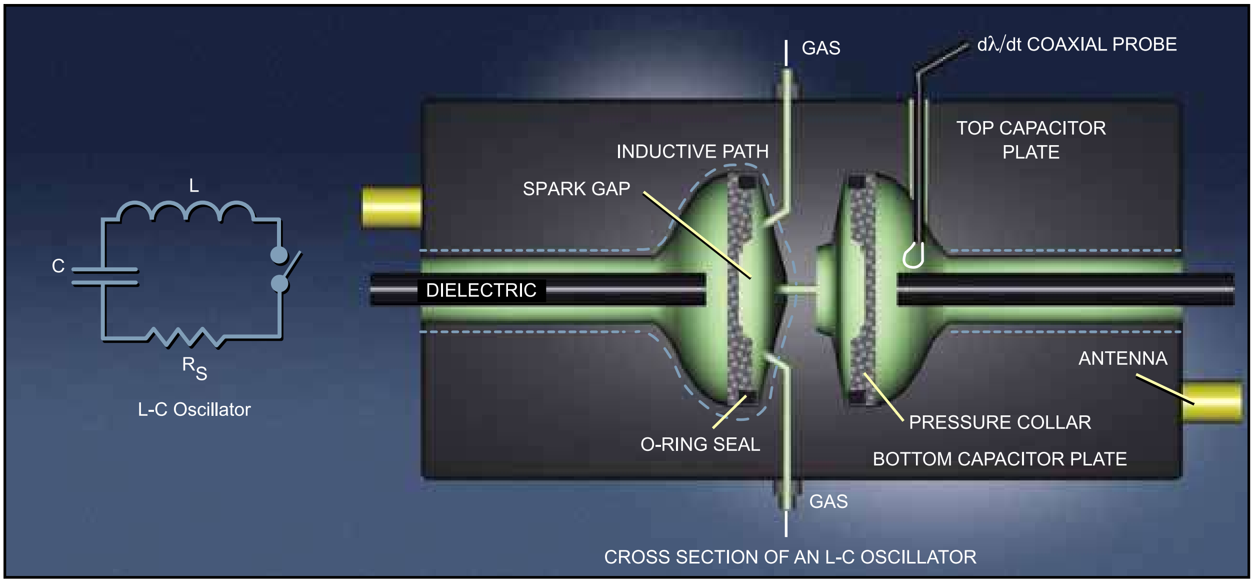

Figure 1. Inductance-Capacitance Oscillator (L-C Oscillator) Diagram

coaxial pipe, which was switched at one end, to create the oscillating waveform. A frozen wave generator, a different type, had quarter-wave sections of cable that were charged plus and minus to create a two-cycle waveform "frozen" in the cable. All sections were simultaneously switched, causing the wave to travel to an antenna. A special folded design was developed so one switch could be used, eliminating the multiswitch synchronization problem. A Ross circuit used a square wave pulse, which traveled down cable "tees," creating reflections, which were timed to create several RF cycles. In the Travetron, the turn-on time of a series of spark-gap switches was incorporated as a designed delay, creating reflections through a series of gaps to produce the waveform. This design allowed higher frequencies. All of these devices were designed, built, and tested to determine power and frequency capabilities, as well as efficiency.

Scientists and engineers at Dahlgren built and tested versions of Hertzian oscillators operating up to half a million volts. These devices powered relatively simple monopole or dipole antennas that could produce very high electric fields at hundreds of meters. In the early 1970s, a special outdoor field-measurement range was constructed. It housed high-voltage systems in underground trailers that fed antennas above ground on a specially-built, 100-m-long ground plane that was constructed for testing and field measurements. A picture of the ground place in a fielded measurement range is shown in Figure 2. Field probes were even carried aboard helicopters to make measurements above ground effects, as shown in Figure 3.

Other types of devices to produce pulses were constructed, too. Vector inversion generators used spiral-wound capacitive plates to generate high voltages without transformers.[2,3] The Landecker ring used a paddle-wheel arrangement of capacitors and inductors charged in parallel and discharged in series. The circular arrangement was designed so the entire system would radiate as a magnetic dipole, thus forming its own antenna.[4] Switch timing was critical, and Dahlgren engineers attempted to verify reports that Landecker developed a specific type that brought all capacitor leads into a single-center spark gap.

Scientists and engineers also looked at devices that used explosives to generate the electrical energy needed. These included explosive flux compressors of several types, which generated fields and then explosively squeezed the fields between conductors to amplify the peak power. In the early 1970s, a large (70-ft clear zone) anechoic chamber was constructed at Dahlgren with an explosive chamber in one end. Explosives would be set off in the chamber to drive various types of flux compressor schemes that would generate electrical pulses fed into an oscillator and antenna in the anechoic chamber. Pulse parameters and field strengths could be measured. Impedance-matching networks, matching transformers, and methods of improving efficiency were studied. Tests were performed at Dahlgren and at Los Alamos using large antennas suspended from balloons.[5] In other schemes, piezoelectric devices were developed, which could be compressed hydraulically and then quickly released to produce high voltages. The concept was to use explosives to generate the high

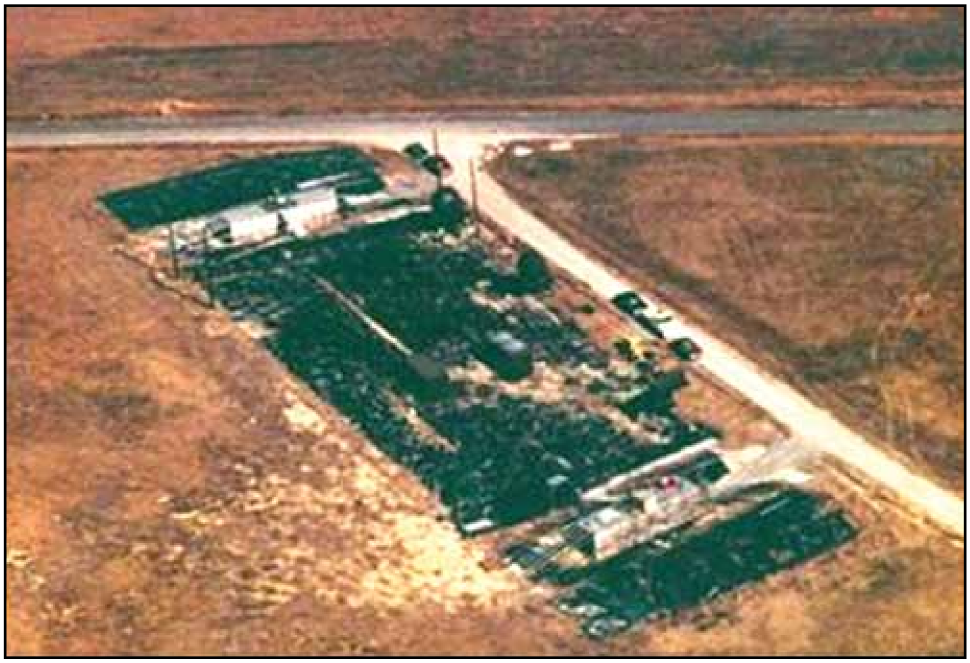

Figure 2. Field Measurement Range

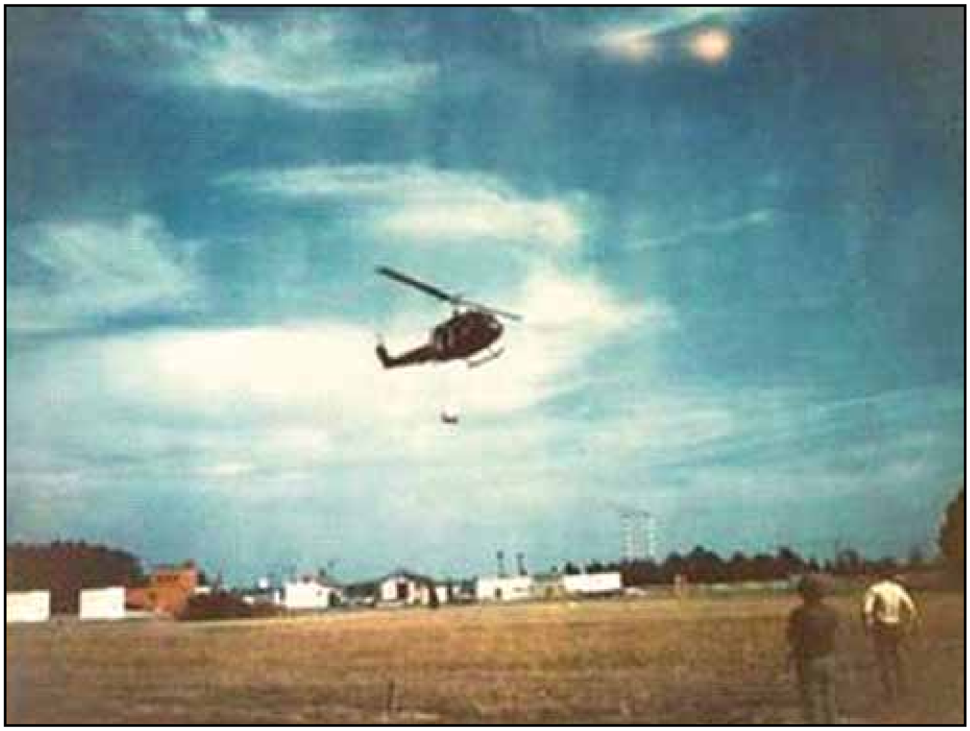

Figure 3. Airborne Electric Field Measurements

pressures. Ferroelectric and ferromagnetic transducers driven by explosives were also tested.[6]

Special Effects Warhead (SEW) Program

In 1973, Dahlgren began the SEW Program to look at the feasibility of "burning out" enemy radar and missile systems using single-shot, very high-peak-power EMPs. The program looked at the feasibility of constructing an electromagnetic warhead that could disable electronics beyond a normal hard-kill explosive range as far as a mile away. The program was funded at several million dollars a year through most of the 1970s.

A major thrust of the SEW Program was to better understand the effects of high fields on military electronics. Little information was available on the vulnerability of foreign or U.S. electronics, particularly entire systems. A trailer-based RF impulse system, employing a Marx-driven L-C oscillator charged at two million volts, was constructed at Dahlgren. This Transportable Oscillating Pulser System (TOPS) was connected to a large bounded-wave structure that produced uniform fields over a region large enough to place an entire radar or missile system. The electric field emitted from the throat of this system was so high that a special bag of high-voltage gas was needed until the radiating structure became large enough to transition to the normal atmosphere. A picture of TOPS is shown in Figure 4.

Since many important target systems were not available for testing, much of the vulnerability information was obtained from U.S. electronics, and estimates were then made for foreign systems. In addition to the tests done at Dahlgren, pulsers were also constructed in mobile trailers that could be transported to other sites for testing against simulated or actual targets. The Mobile Oscillating Pulser System (MOPS) was an example that was carried to test sites, such as China Lake, to perform tests against radars and simulated foreign systems.

A key requirement for the SEW Program was to demonstrate enforceable target vulnerability, which means that a high percentage of the time a large percentage of the targets are affected. One important finding was the broad difference between an electromagnetic safety concern—where a 1 percent vulnerability was far too great—and a weapon concern—where a 10 percent vulnerability was not good enough. The field strengths between the safety requirements and weapon requirements often were many orders of magnitude apart.

The SEW Program looked at many types of electronic component vulnerability, subsystem vulnerability, and complete system vulnerability. As a result, energy tables for burnout effects were developed. Subsequently, Dahlgren performed numerous field tests against radar and communications systems between 1973 and 1978, and funded component and subsystem testing on missiles.

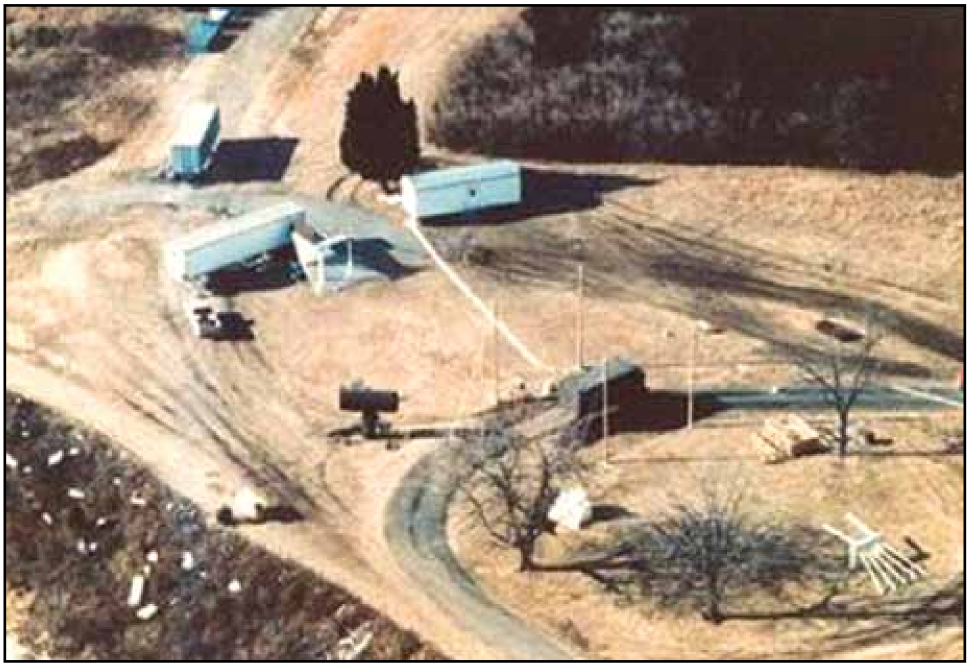

Figure 4. Transportable Oscillating Pulser System (TOPS)

REPETITIVE SYSTEMS FOR ELECTRONIC WARFARE

The electric fields required to damage military electronics in the 1970s often were very high, and ranges typically were limited. As a spinoff of programs trying to damage targets with a single pulse, some of these devices were reduced in size and power, and operated in a repetitive mode to generate noise pulses for the purpose of electronically jamming target systems. In 1976, the Naval Air Systems Command (NAVAIR) began the Electromagnetic Countermeasures Program to study the application of high-repetition-rate Hertzian devices for use as noise jammers. The initial targets were low-frequency radars.

In late 1976, Dahlgren performed effectiveness tests against various radars using helicopter-mounted Hertzian jammers. These devices were able to screen incoming target aircraft at useful ranges. The concept of a forward-launched rocket to deliver a parachute-suspended Hertzian jammer also was investigated. Dahlgren teamed with engineers at China Lake to study packaging concepts of utilizing an extended 5-inch Zuni rocket as a forward-fired delivery vehicle. A prototype is shown in Figure 5.

Similar Hertzian devices were considered for use as communications and data-link jammers. Several antenna deployment schemes were developed, and by fall 1978, successful ground launches had been performed in which the deployment sequence and jammer operation were demonstrated. The name Zuni Expendable Pulsed-Power Oscillator (ZEPPO) was given to the project. Dahlgren teamed with the Naval Avionics Center (NAC) to build the systems. By 1980, China Lake fired the first air-launched prototypes at both low and high altitudes. Devices, batteries, spark gaps, and antennas continued to be developed, and new targets—such as spread-spectrum systems—were tested. Other delivery systems besides rockets were also considered.

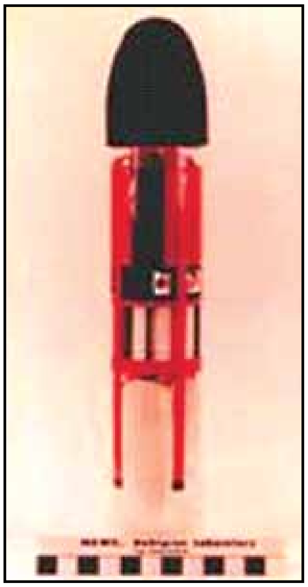

Figure 5. ZEPPO Payload

THE PULSED POWER TECHNOLOGY PROGRAM

Large directed-energy weapons (DEWs) often required megawatts or gigawatts of peak power, so methods of supplying and modifying this power were needed. As Dahlgren became involved in a broad range of DEW systems, one attribute became more and more obvious: the size, weight, and cost of a directed-energy (DE) system were dominated by the pulsed-power technologies needed to drive the system, not by the source device itself. Consequently, more effort began to be devoted to the power-delivery technologies needed for many of the weapon concepts. Pulsed-power components enabled energy to be stored over long periods of time (seconds) and released very quickly (nanoseconds) to obtain a billion times increase in peak power.

Dahlgren hosted a pulsed-power systems symposium and workshop in 1976 and helped initiate the International Pulsed Power Conferences, which began in 1977 and continues today under the Institute of Electrical and Electronics Engineers (IEEE). As Dahlgren's involvement with systems design increased, it became apparent that new technologies were needed in the prime-power and pulsed-power area to support a variety of new concepts. Dahlgren urged the Navy to initiate a Pulsed Power Technology Program to develop power sources, energy storage systems, high-power switches, and power conditioning systems needed for a variety of future weapons. This program was initiated in 1978 and was originally funded by NAVAIR and then by the Directed Energy Program Office (PMS 405) in the early 1980s. In addition to the Pulsed Power Technology Program, PMS 405 also began funding free-electron lasers (FELs), chemical lasers, high-power microwaves (HPMs), and charged-particle beams (CPBs). The Pulsed Power Technology Program at Dahlgren, in turn, funded many areas of research, both internal and external, over the next 10 years. Dahlgren served as the focal point for the Navy's science and technology (S&T) in pulsed power and funded many universities, government laboratories, and commercial companies under the Pulsed Power Technology Program.

To provide large amounts of electrical prime power, new types of rotating machines were studied, including flywheels, conventional alternators, homopolar generators, rotary flux compressors, and compensated pulsed alternators. These machines attempted to produce fast, high-power pulses using special materials to reduce losses, eddy currents, and mechanical stresses. MHD generators were developed using rocket-motor propellant that could be started and stopped. In the mid-1980s, a full-scale hybrid (solid fuel/liquid oxidizer) combustor was fabricated and tested at 10 MW, achieving world records for power-to-weight ratio and conductivity. By 1980, new types of energy storage systems were studied, including inductive storage and advanced capacitors using new types of insulating materials and geometries. During the late 1980s, programs such as the Mile-Run Capacitor Program reduced the capacitor size by a factor of 10 through better synthesis of polymer films.

Beginning with internal independent research funds, Dahlgren developed liquid dielectric materials based on water/glycol mixtures at low temperatures. These water-capacitor devices could hold energy for orders-of-magnitude longer time periods than ever before, allowing pulseforming lines to be constructed that could be charged directly from rotating machines. Dahlgren scientists developed a world-record high-voltage water capacitor that could hold pulses for milliseconds and became internationally recognized experts in water breakdown.[7, 8]

High-power fast switching was another important area of research. Dahlgren funded companies to develop new types of multistage thyratrons that could operate at very high voltages. By the early 1980s, multistage thyratrons capable of operating at over 200 kV, 40 kA with 20 nsec risetimes were demonstrated. Vacuum switches, ignitrons, plasma pinch switches, pseudospark switches, back-lighted thyratrons, and e-beam switches all were studied, as well as a variety of spark-gap switches. Higher power solid-state switches were developed, too, using new geometries and substrate material. Superconducting coils were considered, both for energy storage and as opening switches. Dahlgren engineers developed exploding-wire opening switches, and several types of plasma pinch switches were funded. They also worked on stacked cable pulsers. Additionally, concepts for electromagnetic armor were developed. These systems used high-density capacitors to blunt penetrators. Inductive energy storage—which could be far denser than capacitors—was studied, including methods of generating the seed current and the problematic high-voltage opening switch. Opening switches—which were needed for inductive energy store systems—were studied, as well as magnetic switches, which used saturating magnetic material to sharpen pulses. Magnetic switches operating at 10 kHz were demonstrated by 1983.[9]

In 1985, Dahlgren used internal funds to upgrade a facility to provide controls, diagnostics, and 200 kW of average power at 50 kV to accommodate testing of new switches and water-based capacitors. This facility could control the power with a vacuum-tube pulser and could generate over a million volts with a rep-rated Marx generator. The facility was used to:

- Develop water-dielectric energy storage, rep-rated spark gaps, and pseudospark switches.
- Test a variety of switches developed by contractors, such as back-lighted thyratrons.[10, 11]

A picture of one system being tested—a water pulse-forming line and spark-gap switch—is shown in Figure 6.

Dahlgren concentrated in-house switching efforts in spark gaps. New types of gases were studied, as well as electrode materials, gas-flows, switch geometries, and triggering techniques to produce high-repetition-rate switches for electronic warfare, as well as particle-beam weapons.[12] Dahlgren scientists and engineers demonstrated 100-μs recovery of spark-gap switches after handling kilojoules of energy at hundreds of kilovolts, a world record.[13] The High Energy 2-Pulse System for fast recovery experiment is shown in Figure 7.

In 1986, Dahlgren ran a workshop on high-power switching for Navy tactical and Department of Defense (DoD) strategic applications and became involved with numerous DoD working groups on electromagnetic propulsion, high-power diagnostics, advanced energy conversion, power modulators, and pulsed power. Spark gaps were investigated to create underwater noise for submarines. Dahlgren also led four North Atlantic Treaty Organization (NATO) Advanced Study Institutes in Europe and the UK on various pulsed-power topics. International assessments of key pulsed-power technologies were also performed.

PARTICLE-BEAM WEAPONS

Particle-beam weapons were a major focus of DE work during the 1970s and 1980s. A CPB weapon takes subatomic particles, generally electrons, and accelerates them to near the speed of light before sending them toward a target. These fast electrons penetrate deeply into most materials, so they are difficult to counter. The high-current electron beam was to be accelerated by an induction-type

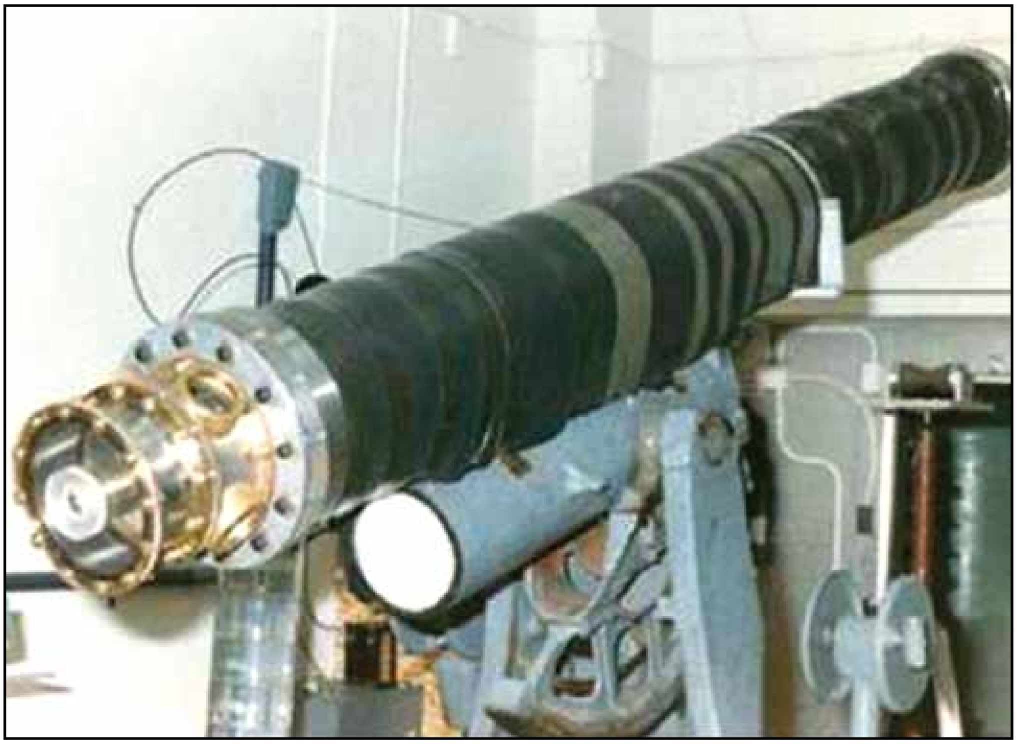

Figure 6. A Water Pulse-Forming Line and Spark-Gap Switch Test

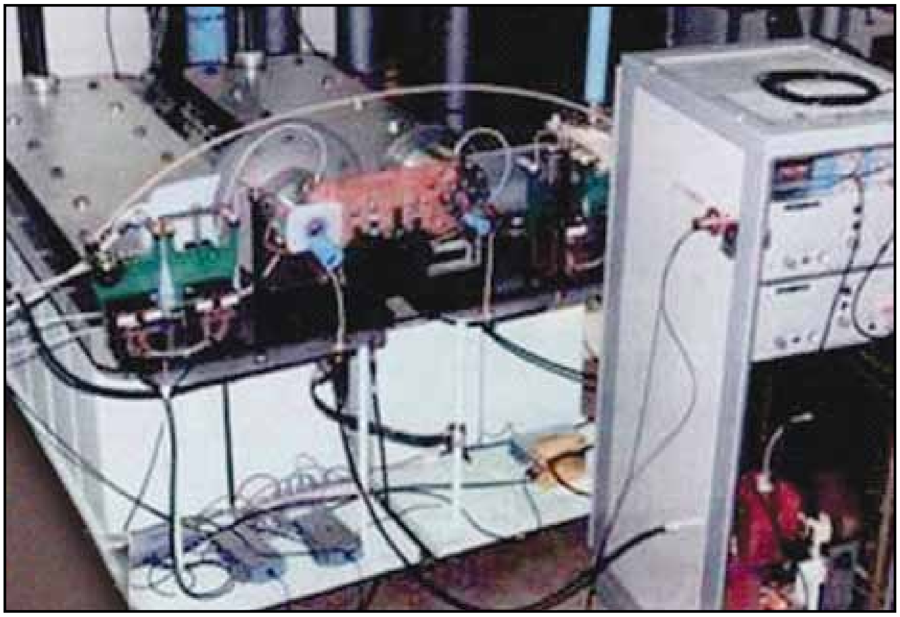

Figure 7. High Energy 2-Pulse System

accelerator, repetitively pulsed. High electron-beam currents (kiloamps) and a hole-boring series of pulses were anticipated to create a stable, long-range beam. Since the beam was capable of penetrating quickly and deeply into any target material, it had the potential to damage electronics and set off explosives before salvage fuzing could occur. The beam was predicted to be all-weather and essentially countermeasure-proof. Even a near miss could cause substantial damage from high fields and X-rays produced by the deceleration of electrons as they hit air molecules near the target. The CPB concept is shown in Figure 8.

Scientists and engineers from Dahlgren worked on the pulsed-power technologies needed to drive these machines beginning in 1980, and it became a major focus of the Pulsed Power Technology Program.[14] The White Oak Laboratory developed beam-steering concepts and looked at material interactions. By 1989, the program investigated:

- Propagation
- Compact Recirculating Accelerators
- Pointing and Tracking
- Prime Power
- Material Interaction
- Fratricide

For a compact shipboard system, recirculating accelerators were needed to make multiple passes of the electron beam past the accelerating cavities. This required a high-power, fast recovery switch, which Dahlgren began working on in 1988. Using patented hydrogen switches and special triggering techniques—efforts that had begun with internal research funds—Dahlgren demonstrated spark-gap switches, the only technology that could meet the current, voltage, and recovery requirements at that time.[15] The High-Voltage 5-Pulse System experiment is shown in Figure 9.

During these technology efforts, significant advances were achieved in all aspects of the program. These included:

- Generating high-current, high-energy beams (although still below weapons parameters)
- Demonstrating a 360° turn in a high-current beam
- Propagating a single pulse through the air
- Demonstrating beam steering on a small scale
- Performing target interaction measurements

Multipulse, long-range propagation was never demonstrated. A comprehensive tri-service summary called the Net Technical Assessment for CPB was sponsored by the Defense Advanced Research Projects Agency (DARPA) in 1987 to describe the accomplishments of the program. The report said compact accelerators were the most pressing technology need. As a result, most funding was directed toward this topic. Funding was stopped in the early 1990s, however, due to the high expense, stretched timelines, and changes in the threat.

PULSED POWER AND ELECTROMAGNETIC LAUNCHERS

During the 1980s, the Army and Air Force looked at short-range electromagnetic weapons to penetrate stronger armor with higher velocities. The Navy worked on concepts for a weapon that could be mounted on ships to intercept missile systems at line-of-sight distances. The Navy—then the biggest user of space systems—was also interested in studies showing that small satellites could

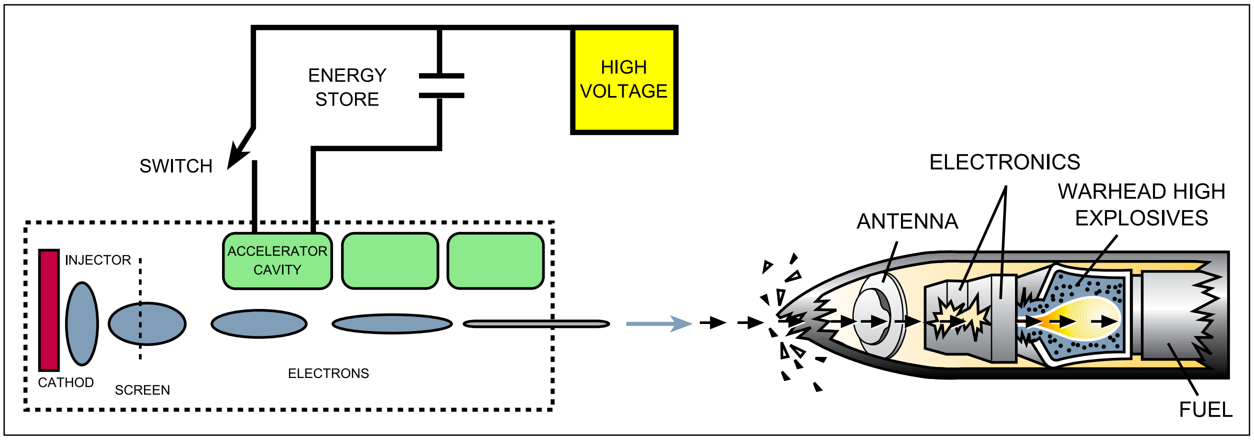

Figure 8. Charged-Particle Beam (CPB) Concept

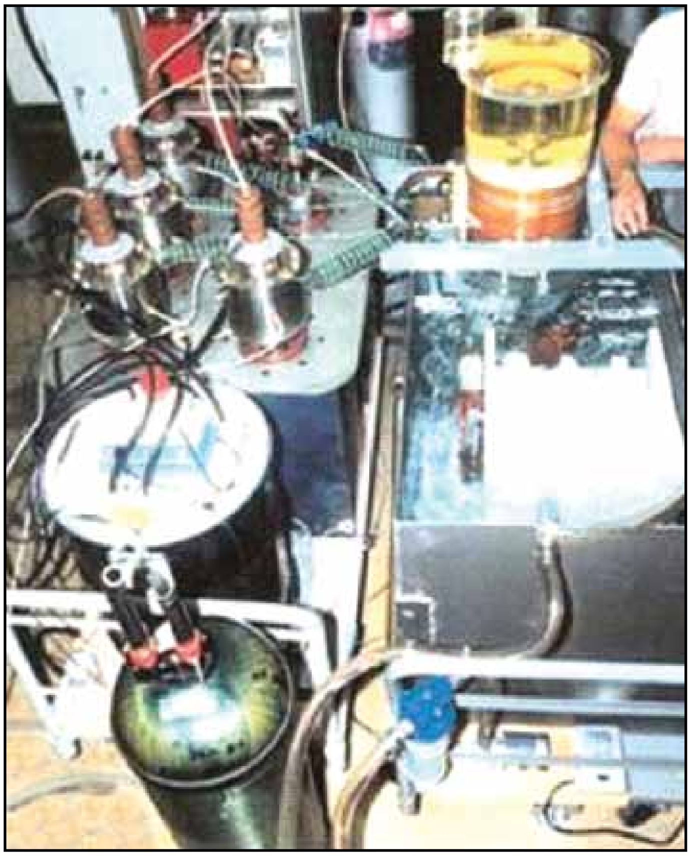

Figure 9. High-Voltage 5-Pulse System Experiment

be electromagnetically launched into low Earth orbit for the fraction of the cost for a normal launch.

Through the 1980s, electric guns were funded by independent research and independent exploratory development programs at Dahlgren, studying electric gun concepts for both rail guns and electrothermal (ET) guns. Kinetic energy weapons were also investigated as part of the Pulsed Power Technology Program. Under these programs, pure electric launchers were developed and tested at Dahlgren, including ones that self-formed projectiles.[16–18] Also studied were ET guns that used the discharge of electrical energy at the gun breech to generate a plasma jet. This plasma jet heated a low-molecular-weight working fluid, such as water, to produce a heated gas that accelerated the projectile to higher velocities than conventional explosives. The Electrothermal-Chemical (ETC) Gun concept augmented the electrical energy generating the plasma jet with a chemical reaction. A 127mm ETC gun was investigated, and a 60mm ETC gun was tested at Dahlgren, with the ability to fire short bursts at a rate of 100 rounds per minute.[19]

Early Dahlgren work on electromagnetic launchers—along with capacitor development and switch advances from the Pulsed Power Technology Program—allowed Dahlgren to provide the Navy with detailed conceptual designs in the late 1990s for near-term, long-range rail guns based on capacitor energy store. These efforts helped support the decision to begin a long-range rail-gun program at Dahlgren that continues today, resulting in world-record achievements. Capital investment funds were used to construct a high-energy facility in 2005 to test pulsed-power components and module designs for use in electromagnetic launcher programs. An early electromagnetic launcher is shown in Figure 10.

High-Energy Lasers (HELs)

In general, megawatts of continuous laser power are required to kill hard targets at long ranges. Laser technologies that can produce this much power are very limited. The Navy was a leader in developing powerful chemical lasers in the 1970s and 80s. These lasers burned chemical reactants to

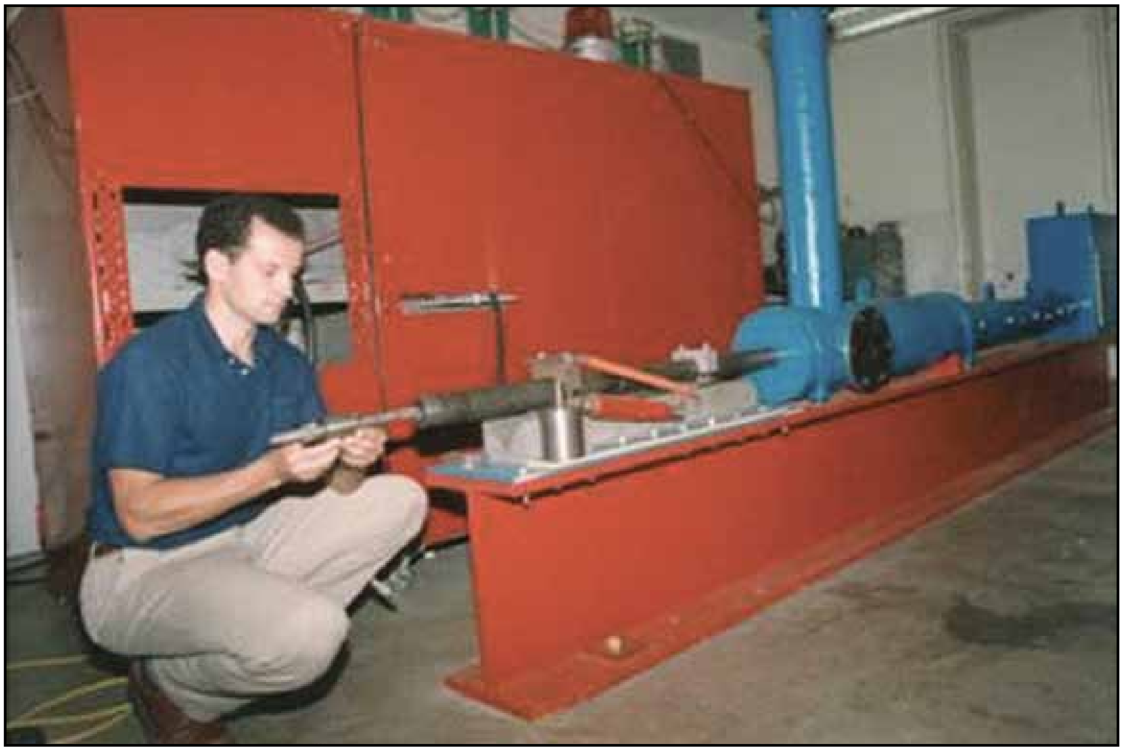

Figure 10. Early Electromagnetic Launcher at Dahlgren

generate the excited states for lasing, thus reducing the need for large amounts of electrical power. The Navy built an entire HEL system, including the Mid-Infrared Advanced Chemical Laser (MIRACL) and the Sea-Lite beam director. By 1990, this building-sized system demonstrated shooting boosters, missiles in flight, and supersonic vehicles. However, the system had drawbacks because it:

- Used hazardous, expensive chemicals
- Had propagation problems at the mid-infrared wavelength
- Was large in size and high in cost

FELs require electron accelerators similar to CPB weapons, so they also are large and complex. However, they can be designed to operate at optimum wavelengths and scale nicely to higher powers. The Strategic Defense Initiative began working on FELs in the late 1980s, funding the advanced test accelerator at LLNL, originally developed for CPBs. FELs were also studied under the Strategic Defense Initiative Organization (SDIO) to be used as an antisatellite weapon. These lasers went from milliwatts to watts under SDIO, and then to kilowatts more recently with work at the Thomas Jefferson National Accelerator Facility in Virginia. Space-based lasers and relay mirror systems were studied under SDIO funding, too, including the development of the Advanced Beam Control System for beam steering, beam control, rapid optical retargeting, and self-alignment.

Dahlgren engineers concentrated its internal laser efforts on medium-power soft-kill weapons. They performed tests against sensors and cameras, and investigated damage thresholds. In the late 1980s, Dahlgren engineers worked with optical augmentation to locate enemy optics for targeting and on green laser dazzlers for defense against small-boat attack. There were efforts to harden electro-optical equipment, including sights and night-vision systems for the Marines, and laser eye-protection filters for goggles and binoculars. Laser systems were also investigated for remotely cutting holes and wires to disable electronics. Lethality work continued under funding from the Joint Technology Office for High-Energy Lasers to look at alternative wavelengths and pulse shapes in addition to modern target materials.[20]

Dahlgren scientists continued to investigate laser-damage thresholds for materials, components, and subsystems for a variety of laser technologies. Near the start of the 21st century, commercial lasers based on pumping optical fibers with semiconductor lasers became common and more powerful. Dahlgren purchased the Navy's largest collection of fiber lasers in 2004 and began investigating ways to combine multiple beams into a laser weapon. These lasers have very high efficiencies, above 20 percent, and the fiber-optic output reduces the requirement for complex optical paths. In 2008, Dahlgren engineers demonstrated a laser capability to ignite spinning mortar rounds, and in 2009, engineers demonstrated the capability of fiber lasers in a shoot down of soft targets at China Lake, California.

RESURGENCE OF DIRECTED ENERGY

With the fall of the Soviet Union and a greatly altered threat, DoD funding (particularly technology funding) experienced an overall decline in the late 1980s and early 1990s. This caused Navy managers to emphasize near-term, lower risk, evolutionary concepts. The Pulsed Power Technology Program and the Navy's Charged Particle Beam Program both came to an end. Investigations into HPM weapons declined as the difficulty of burnout of military electronics—particularly analog components—became apparent. Problems with propagation and cost caused the Navy to greatly reduce efforts on chemical lasers. With the cancellation of major programs, Dahlgren used internal funding in 1990 to keep a core technical capability together, which was necessary for the Center to remain in the mainstream of tactical DE and its associated technologies. Efforts continued in water breakdown, testing of contractor-developed pulsed-power components, and electric guns. New talent and technologies from universities were brought in to jump-start new projects. Tunable waveform generators using unique semiconductor materials were developed. These used bulk semiconductor material, fabricated in-house, that could be used as a fast switch controlled by laser light for both on and off operation. This allowed faster repetition rates and better triggering than could be done with small spark gaps, as well as the ability to create specific waveforms.[21] "Green" technologies were also investigated using nonthermal plasmas and spark-gap shock waves for cleaning and pollution reduction.[22] New types of particle detectors and magnetic field sensors were developed, and new methods of infrastructure protection were investigated.[23] Soft-kill weapons, both optical and HPM, continued to be studied. Short-pulse jamming of spread-spectrum systems was investigated, as well as beat-wave coupling and special waveforms.[24]

A number of trends led to a resurgence of DEWs by the end of the 20th century. The DoD trend in using digital electronics and off-the-shelf commercial technologies increased dramatically. The pace of change in electronics and computers changed rapidly, too. Most of these new electronic systems had never been tested for vulnerability, and there was a question of how much they would increase military vulnerability to RF or HPM attack. The reduced emphasis on nuclear EMP shielding meant more military electronics were not as well protected from RF attack. Consequently, interest in protecting U.S. military and civilian infrastructure increased, including systems in foreign countries. Moreover, with the increasing reliance on civilian infrastructure, such as power, communications, and emergency and industrial systems—all of which were controlled by digital electronics—the potential that an adversary could attack infrastructure systems to affect or divert military operations became an increasing concern. Following several major terrorist attacks during this time period, there was also concern about the impact of an RF attack on airport towers, financial systems, alarm systems, and industrial plants. Human factors—such as a state of confusion experienced by humans—also played an important part in determining the overall effects of an RF attack.

The asymmetric threat—where large numbers of cheap weapons in a swarm attack could overrun a few sophisticated weapons—caused more concern. As the asymmetric threat to the surface Navy pushed the limits of conventional defensive systems, DE—with it speed-of-light propagation, soft-kill potential, and cheap rounds—offered tactical advantages, either as an adjunct to conventional systems or as stand-alone systems. Additionally, there was an increased emphasis on nonlethal, precise accuracy and graduated effects that could be used. Moreover, the idea that future battles would be fought together with civilians and friendly forces on the battlefield increased the importance of low collateral damage and antimateriel attacks.

The Joint Program Office for Special Technology Countermeasures (JPO/STC), located at Dahlgren, began efforts concerning the vulnerability of new digital systems to RF attack. The program also established a DoD-wide database of vulnerability data, source designs, and RF-effects information—bringing together much of the information collected by the services over the years. The program looked at the protection of modern digital infrastructure systems and funded a facility constructed in 1992 to test large-scale electromagnetic vulnerabilities to various methods of attack.

In the late 1990s and early 2000s, Dahlgren initiated programs regarding the potential for RF attack using nonkinetic disruption, with minimal collateral damage. Capital investment funds were used to construct a test facility for this effort in 1998. Dahlgren developed RF payloads for remotely piloted vehicles and demonstrated their effectiveness in field tests in 1999, and in similar tests in 2007. The successful completion of Project Guillotine was DoD's first demonstration of this type of HPM technology. As the need for statistical vulnerability to commercial digital systems became apparent, Dahlgren constructed instrumented test facilities in 1999 and 2002. Two multistory buildings could be reconfigured to reflect different types of building construction and electromagnetic shielding. Large complexes of electronics, computer networks, server systems, telephone systems, security systems, and various types of digital industrial controls could be assembled, instrumented and exposed to attack from an external device or technique. This program-funded complex—called the Maginot Open Air Test Site (MOATS) facility—continues to be used to test target systems, as well as a variety of RF weapon technologies developed internally and by external and international organizations. A picture of the MOATS facility is shown in Figure 11.

As the need for additional DE laboratory space and testing capabilities became apparent, Dahlgren applied for military construction funds, and in 2008, constructed the Naval Directed Energy Center (NDEC), with access to Dahlgren's over-water test range. Other construction funds were used to construct a remote facility at the Pumpkin Neck Explosive Test Range to serve as a laser backstop and measurement facility, as well as an explosive-test staging area. These facilities already have been used to develop and test fiber lasers against modern threat targets. Construction is currently underway to build an expansion of the NDEC and a 120-m laser test laboratory building using an existing tunnel structure. This collection of facilities represents very important capabilities to develop and test future DE systems.

CONCLUSION

For over 40 years, the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) has been a leader in developing DE devices, pulsed-power systems, and electric weapons. Its people have contributed many publications and patents, and set world records. DEWs tend to be complex and technically challenging to build. Regardless, these weapons offer important, powerful advantages, such as:

- Deep Magazines
- Cheap Rounds
- Fast Targeting
- Variable Lethality
- Pinpoint Targeting

As a result of NSWCDD's leadership, persistent scientific initiatives, and leading-edge engineering

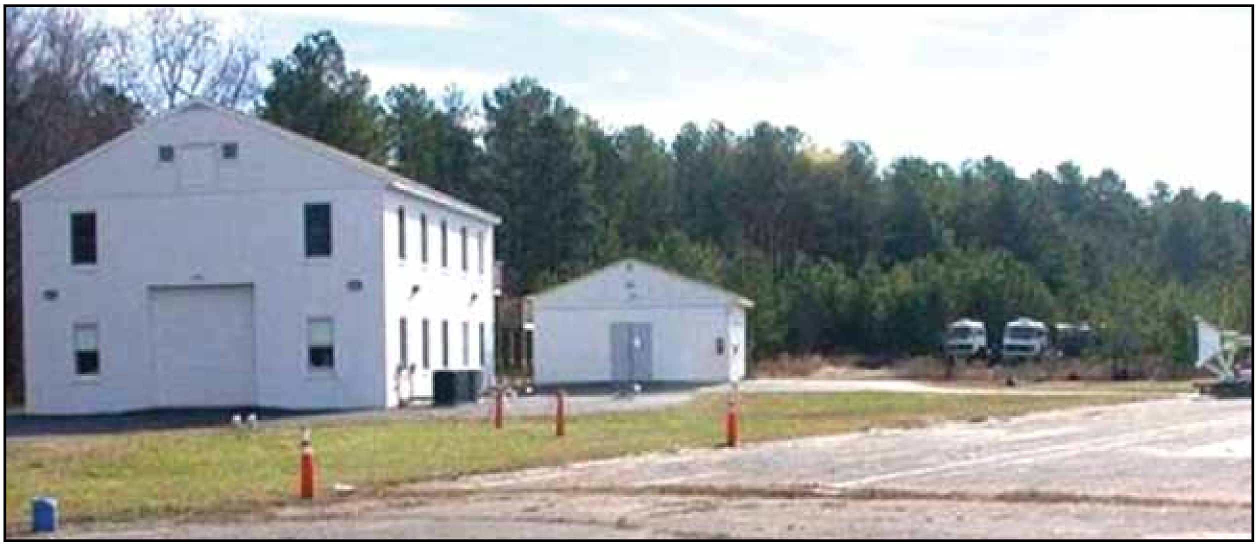

Figure 11. MOATS Facility Undergoing Testing with an RF Weapon (on right)

over the years, naval warfighters will increasingly find themselves turning to DEWs when dealing with situations spanning the spectrum of conflict.

References

1. Moran, S.L., "High Repetition Rate L-C Oscillator," *IEEE Transactions on Electron Devices*, Vol. ED-26 No. 10, October 1979.
2. Fitch, R.A. and Howell, V.T.S., "Novel Principle of Transient High-Voltage Generation," *Proceedings IEEE*, Vol. III, April 1964.
3. Ramrus, A. and Rose F., "High-Voltage Spiral Generators," *1st International Pulsed Power Conference*, Lubbock Texas, pIIIC9, November 1976.
4. Landecker, K.; Skatterbol, L.V. and Gowdie, D.R., "Single Spark Ring Transmitter," *Proc. IEEE*, Vol. 59, No. 7, July 1971, pp. 1082–1090.
5. Erickson, D.J.; Fowler, C.M. and Caird, R., *EM Radiation from an Explosive Generator System Coupled to a Large Antenna*, Los Alamos Scientific Laboratory Report M-6-258, March 1978.
6. Fowler C.M. and Brooks, M.L., *Explosive Flux Compression Devices Project*, ARPA Order No. 2161, Semiannual Technical Report, 1 March 1972–31 August 1972.
7. Zahn, M.; Ohki, Y.; Fenneman, D.; Gripshover, R. and Gehman, V., "Dielectric Properties of Water and Water/Glycol Mixtures for Use in Pulsed Power System Design," *Proceedings of the IEEE*, September 1986, pp. 1182–1221.
8. Gehman, Jr., V.H.; Gripshover, R.J.; Berger, T.L.; Bowen, S.P. and Zia, R.K.P., "Effects of Filtration on the Impulse Breakdown Strength of High-Purity Water," *IEEE 1995 Tenth International Pulsed Power Conference*, Albuquerque, New Mexico, 10–13 July 1995.
9. Moran, S.L., *Pulsed-Power Technology Program: Accomplishments in Switching*, NAVSWC Tech Note TN-89-259, 1990.
10. Berger, T.L.; Gehman, Jr., V.H.; Lindberg, D.D. and Gripshover, R.J., "A 2.5 Gigawatt Liquid Dielectric Coaxial Pulse Forming Line," *Proceedings of the 6th IEEE International Pulsed Power Conference*, Arlington, Virginia, 29–30 June and 1 July 1987.
11. Bernardes, J.S., "A High Power Fast Response Switch Testing System," *7th IEEE Pulsed Power Conference*, June 1989.
12. Moran, S.L. and Hutcherson, R.K., *High-PRF High-Current Switch*, U.S. Patent No. 4,912,396, 27 March 1990.
13. Moran, S.L. and Hardesty, L.H., "High-Repetition-Rate Hydrogen Spark Gap," *IEEE Transactions on Electron Devices*, Vol. 38, No.4, April 1991.
14. Rose, M.F. and Huddleston, C.M., *Pulsed Power and Directed Energy Technology*, 10th Anniversary Technical Symposium, NSWC MP 84-560, October 1984.
15. Moran, S.L., "Hydrogen Spark Gap for High-Repetition-Rate Compact Accelerators," *NSWCDD Technical Digest*, NSWCDD/MP-92/104, September 1992, pp. 126–139.
16. Bernardes, J.S. and Merryman, S.,"Parameter Analysis of a Single Stage Induction Mass Driver," *Proceedings of the Fifth IEEE Pulsed Power Conference*, Arlington, Virginia, June 1985, pp. 552–555.
17. Bernardes, J.S., "Muzzle Velocity Control with Electrothermal Guns," *Proceedings First Navy Independent Research/Independent Exploratory Development Symposium*, Laurel, Maryland, June 1988, Vol. 2, pp. 123–131.
18. Bernardes, J. S., "Focusing Induction Accelerator," *Proceedings of the Sixth IEEE Pulsed Power Conference*, Arlington, Virginia, June 1987, pp. 735–738.
19. Bernardes, J.S., *Sliding Breech-Block for Repetitive Electronic Ignition*, Patent Number: 5,233,902, 10 August 1993.
20. Kim, H.S. et al., "Effects of Chopping Laser Penetration of Metal Targets," *Applied Physics Letters*, 55 (8), 21 August 1989.
21. Stoudt, D.C. and Schoenbach, K.H., "The Electrical Characteristics of Semi-Insulating GaAs for High-Power Switches," *7th IEEE Pulsed Power Conference Proceedings*, June 1989.
22. Grothaus, M.G. et. al., "Gaseous Effluent Treatment using a Pulsed Corona Discharge," *Proceedings of 10th IEEE Pulsed Power Conference*, Albuquerque, New Mexico, July 1995.
23. Gripshover, R.J. et. al., *Power Frequency Magnetic Field Detection Systems*, Patent No. 6,483,309, 19 November 2002.
24. *Spread Spectrum Countermeasures Final Report*, E-Systems, Inc., Melpar Division under NSWC contract # N68786-82-C-0109.

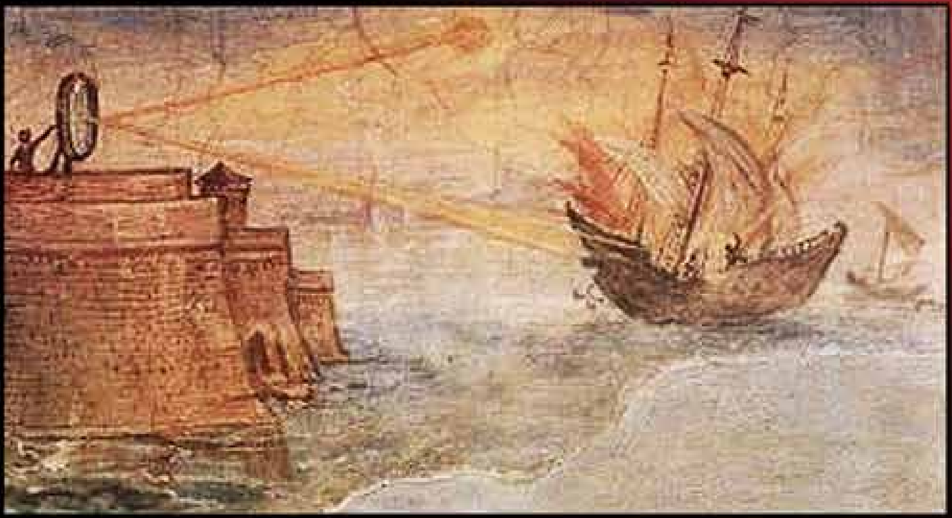

HISTORY OF LASER WEAPON RESEARCH

By Melissa Olson

The idea of using light as a weapon can be traced back to Hippocrates, commander of the Greek forces in 212 B.C. His forces supposedly set fire to the sails of the Roman fleet by focusing sunlight with mirrors. Weapons systems based on lasers and "ray guns," long a staple of science fiction, have captured the imagination of people everywhere. But with steady progress toward the development of lasers in the last 40 years, viable, state-of-the-art laser weapon systems have now become a reality.

The production of lasers in the modern scientific world is fairly new. The first laser was developed in the 1960s and represented the beginning of a drastic change in how the military viewed warfare. The late 1970s and 1980s, too, marked a busy time period for developing lasers into possible weapon systems. All branches of the military and industry were striving to master high power levels, beam control, and adaptive optics. In 1999, the Department of Defense (DoD) formally recognized lasers as future weapons and began research and development (R&D). In 2000, the Joint Technology Office for High Energy Lasers was formed to bring all laser technologies together to develop a complete laser weapon system that could be used by the warfighter.

ELECTROMAGNETIC SPECTRUM

The electromagnetic spectrum contains all the types of electromagnetic energy, including radio waves, microwaves, infrared, visible light, ultraviolet, and gamma rays. ***Laser*** is an acronym for "light amplification by stimulated emission of radiation." Light, therefore, is a type of electromagnetic radiation. Light is made up of tiny packets of energy called photons. The amount of energy is what determines the wavelength. Lasers are usually infrared (1 mm to 750 nm) and visible light (750- to 400-nm wavelength). Microwaves are mostly high-frequency radio waves (millimeters to centimeters), with wavelengths 10,000 times longer than lasers. Diffraction of any electromagnetic radiation beam is based on the wavelength and aperture size. For the same aperture size, lasers diffract 10,000 times less than microwaves. This allows the beam to reach farther ranges while maintaining a small spot size of concentrated energy on the target. Lasers are preferred in specific scenarios because of minimal diffraction. The electromagnetic spectrum is shown in Figure 1.

LASER FUNDAMENTALS

The quantum mechanical idea of stimulated emission of light was discovered by Albert Einstein in 1917 and is one of the fundamental ideas behind the laser. Einstein

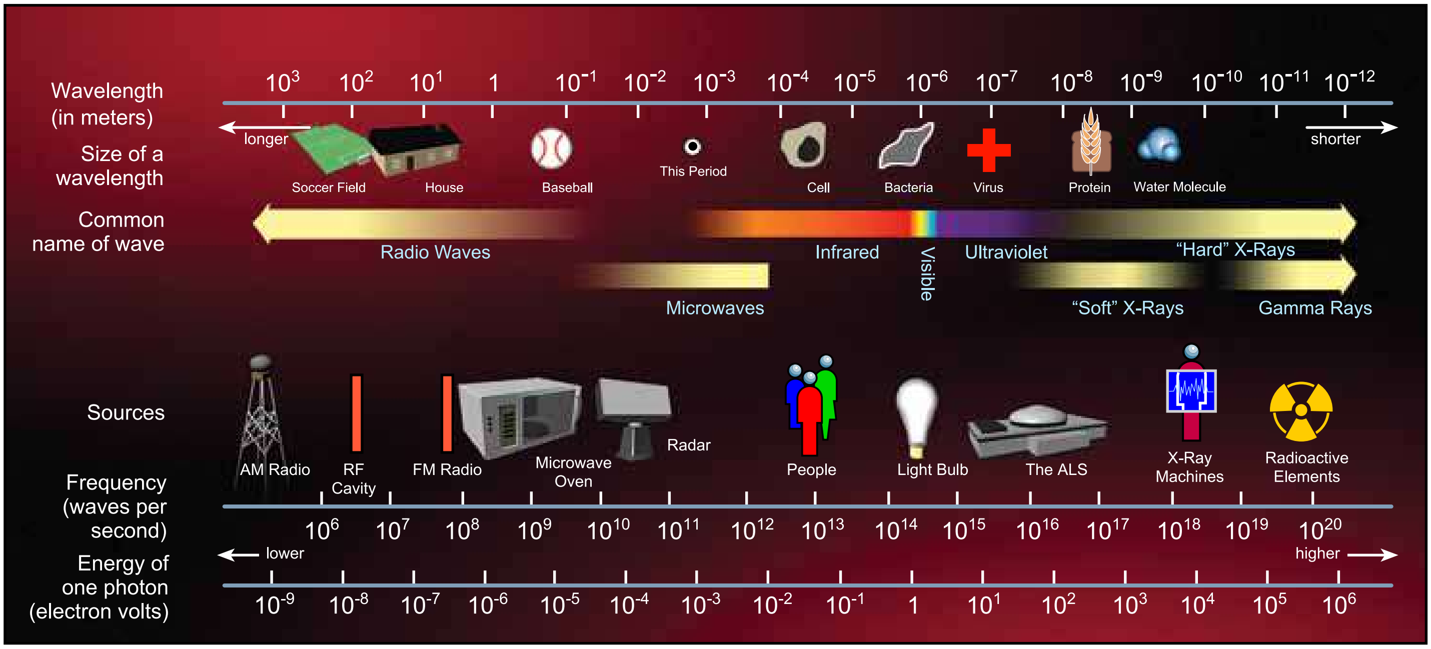
Wavelength (in meters)
10^3 10^2 10^1 1 10^-1 10^-2 10^-3 10^-4 10^-5 10^-6 10^-7 10^-8 10^-9 10^-10 10^-11 10^-12
longer
shorter
Size of a wavelength
Soccer Field
House
Baseball
This Period
Cell
Bacteria
Virus
Protein
Water Molecule
Common name of wave
Radio Waves
Infrared
Visible
Ultraviolet
"Hard" X-Rays
Microwaves
"Soft" X-Rays
Gamma Rays
Sources
AM Radio
RF Cavity
FM Radio
Microwave Oven
Radar
People
Light Bulb
The ALS
X-Ray Machines
Radioactive Elements
Frequency (waves per second)
10^6 10^7 10^8 10^9 10^10 10^11 10^12 10^13 10^14 10^15 10^16 10^17 10^18 10^19 10^20
lower
higher
Energy of one photon (electron volts)
10^-9 10^-8 10^-7 10^-6 10^-5 10^-4 10^-3 10^-2 10^-1 1 10^1 10^2 10^3 10^4 10^5 10^6

Figure 1. Electromagnetic Spectrum

theorized that when a photon interacts with an atom or molecule in an excited state, two photons are produced when the atom or molecule leaves the excited state. Population inversion occurs when the atoms or molecules are in the excited state. In order for molecules to come out of the normal "ground" state, a source of power must be introduced to the system energizing the atoms to the excited state. When many photons are passed through many excited atoms, more and more photons are produced. The photons are contained and reflected back and forth in a cavity, with mirrors usually on each end. The mirror on the output end is only partially reflective, allowing some photons to leak through, creating the laser beam.

The difference between an everyday light bulb and the light of a laser is temporal and spatial coherence. In a light bulb, the light emits photons equally in all directions. The light is random, out of phase, and multiwavelength. A laser emits coherent light, so photons travel in identical direction and phase. A laser is also monochromatic, i.e., light of one wavelength. Another significant difference is that laser light is highly collimated, which means the laser beam can travel long distances with minimum spreading.

The laser gain medium through which the photons travel to become amplified or magnified can vary. The source of power used to excite the medium, achieving population inversion, can be the result of a chemical reaction, an electric discharge, a flash lamp, another laser, or some other excitation mechanism. The type of the lasing medium determines the type of laser. The three categories in which lasers are usually classified are chemical, gas, and solid state. A laser can also be continuous wave (CW) or pulsed. Each type of laser produces a specific wavelength of radiation. It is important to note that different wavelengths of radiation interact with the atmosphere differently. A laser beam is either scattered or absorbed by air molecules, water vapor, or dust. Longer wavelengths scatter less and are absorbed more than shorter wavelengths; our sky is blue because the shorter blue wavelengths of light are scattered more than the longer wavelengths.[1] Gamma rays are so highly absorbed that they cannot propagate more than a few feet in the air. Thus, some laser wavelengths are scattered or absorbed more than others. This makes laser wavelengths with minimum absorption better for use as directed-energy weapons since they propagate through the atmosphere better than others. For example, the carbon-dioxide (CO_2) laser is strongly absorbed by water vapor, so any use near the ocean will be negatively affected. Near-infrared and infrared lasers have shorter wavelengths with negligible absorbance. The optimal laser choice, therefore, would be a wavelength-tunable laser that could vary depending on the atmospheric conditions, such as the free-electron laser (FEL).

Lasers have affected almost every type of modern technology. Most laser technologies use low powers and were mastered very quickly. They are used in many everyday appliances, such as scanning/inventory devices, surgery/medicine, hair removal, presentation pointers, law enforcement, ranging and sighting devices, welding applications, and much more. Using a laser as a weapon has many advantages. For example, a laser:

- Is unaffected by gravity
- Causes minimal collateral damage
- Travels at the speed of light
- Can precisely reach far distances
- Is capable of causing a specific, predetermined amount of damage to targets

The theory behind these capabilities makes the laser weapon a prime choice in multiple engagement scenarios. However, developing lasers with higher powers to use as a weapon has proven more difficult than first considered.

Military Laser History and Laser Types

Generally, a laser weapon is any laser used against the enemy with more than 50 kW to megawatts of power. This is much greater power than commercial lasers. Accordingly, they have greater support needs, including:

- Environmental and personnel safety
- Mirror coatings
- Chilling requirements
- Power requirements
- Laser fuel storage
- Alignment and tracking requirements

In 1960, the very first laser (a ruby laser) was built, producing minimal power. This event was followed by many other laser technology developments. The first chemical laser, hydrogen fluoride (HF), was built in 1965, producing 1 kW. It was then that DoD became interested in researching and developing a more powerful laser for weapon applications. Subsequently, in 1968, the Defense Advanced Research Projects Agency (DARPA) Baseline Demonstration Laser produced 100 kW, and the Navy-ARPA Chemical Laser (NACL) produced 250 kW in 1975. The very first laser is depicted in Figure 2.

Solid-State Lasers (SSLs)

An SSL uses a solid lasing medium, such as a rod made up of glass or crystal, or a gem, like the ruby laser. Along with the rod or host material is an active material, such as chromium, neodymium, erbium, holmium, or titanium. Chromium is the active material used in ruby lasers. Neodymium is the active material in the most widespread applications. A flash lamp, arc lamp, or another laser carries out the optical cavity pumping to achieve population inversion and stimulate the laser beam. The Neodymium Yttrium-aluminum garnet (Nd:YAG) laser is a popular SSL. It operates at a 1064.5-nm wavelength and can be pulsed wave or CW. A great advantage of these lasers is that the wavelength and pulse duration can be varied considerably.[1] The power level can reach up to megawatts when using Q-switching to achieve

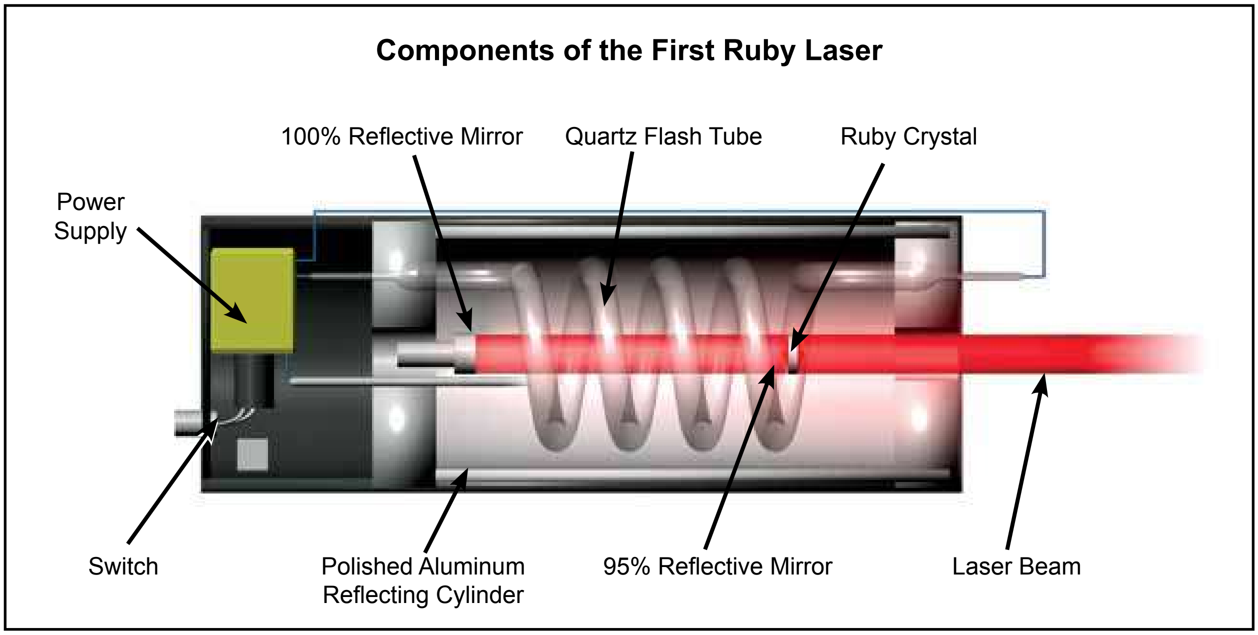

Figure 2. First Ruby Laser Developed in 1960 by Research Physicist Theodore H. Maiman

short pulse lengths. The various interactions with the laser and different crystalline materials can double the electromagnetic frequency, which will halve the wavelength, bringing the laser beam into the visible range, 532 nm (green). The wavelength can be further divided down three or four times, making this laser range from the near-infrared to the ultraviolet wavelength. These lasers are commonly used for rangefinders and target designators. Other advantages of these lasers are that they can be made very small, rugged, cheap, and battery-powered. Characteristics of SSLs are shown in Table 1.

Chemical Lasers

A chemical laser uses chemical reaction to create population inversion in the lasing medium. One example is the Mid-Infrared Advanced Chemical Laser (MIRACL) developed in the mid-1980s. The MIRACL is a continuous-wave, mid-infrared (3.8-μ) laser. Its operation is similar to a rocket engine in which a fuel (ethylene, C2H4) is burned with an oxidizer (nitrogen trifluoride, NF3).[2] Free, excited fluorine atoms are among the combustion products. Just downstream from the combustor, deuterium and helium are injected into the exhaust. Deuterium (U) combines with the excited fluorine to create excited deuterium fluoride (DF) molecules, while the helium stabilizes the reaction and controls the temperature.[2] The laser's resonator mirrors are wrapped around the excited exhaust gas, and optical energy is extracted. The cavity is actively cooled and can be run until the fuel supply is exhausted. The laser's megawatt-class output power can be varied over a wide range by altering the fuel flow rates and mixture. The laser beam in the resonator is approximately 21-cm high and 3-cm wide. Beam-shaping optics are used to produce a 14- × 14-cm (5.5- × 5.5-inch) square, which is then propagated through the rest of the beam train. Diagnostics for evaluating the beam shape, absolute power, and intensity profile are used on each firing of the laser. The beam can be directed to a number of different test areas or to the SEA LITE beam director.[2] The DF Chemical Laser (MIRACL) and the Sea Lite Beam are shown in Figure 3.

The laser and beam director were integrated in the mid-1980s at the Army's High Energy Laser Systems Test Facility (HELSTF) at White Sands Missile Range, New Mexico. Following integration, extensive tests were conducted in the areas of:

- High-power optical components and beam-path conditioning
- Beam-control techniques
- High-power propagation
- Target damage and vulnerability
- Target lethality[3]

Tests supported by the MIRACL included:

- The high-power dynamic with flying drone (BQM-34)
- Conventional defense initiative with flying drone
- High-velocity target test with Vandal Missile
- High-altitude target tests with flying drone
- Missile and plume tests using the 1.5-m aperture
- Radiometrically calibrated images and spectral radiometry

These successful tests are what made many believe that MIRACL was the first and only successful laser weapon system developed by the Navy prior to the Navy Laser Weapon System (LaWS).[3]

Gas Lasers

Gas lasers are a type of chemical laser that uses a pure gas or gas mixture to produce a beam. The typical gas laser contains a tube with mirrors on each end. One end transmits the beam out of the cavity. Most gas lasers use electron-collision pumping, with electric current passing through the gas. Some use optical pumping with flash lamps. The helium

Table 1. Characteristics of Solid-State Lasers

Name	Wavelength (nm)	Typical Power	Typical Operation
Alexandrite	700-830	5 watts	Pulsed/CW tunable
Erbium	850/1230/1540/1730/2900	8 watts	Pulsed
Holmium: glass	1950	milliwatts	Pulsed
Neodymium	1064/1123/1318/1370	megawatts	Pulsed
Neodymium: glass	1060	megawatts	Pulsed
Neodymium: YAG	1064.5	megawatts	Pulsed/CW
Ruby	694.3	10–15 watts	Pulsed
Titanium-sapphire	660–1060	15 watts	Pulsed/CW tunable

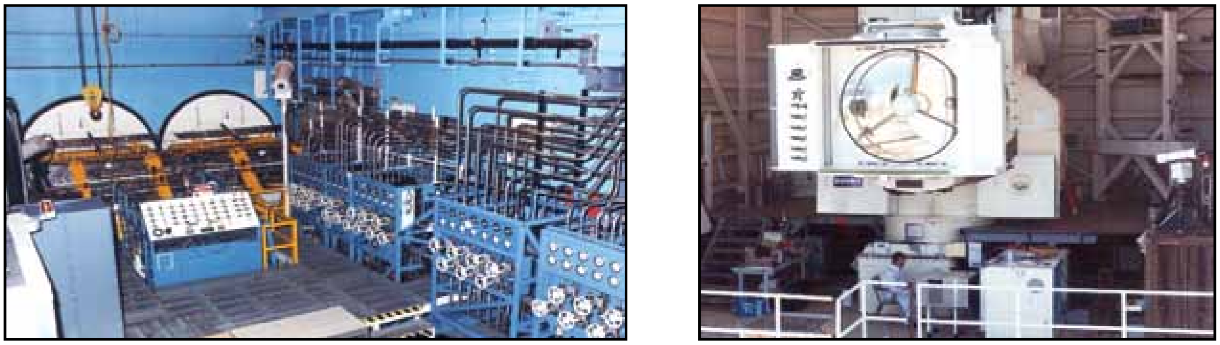

DF Chemical Laser (MIRACL)

Sea Lite Beam

Figure 3. DF Chemical Laser (MIRACL) and Sea Lite Beam

neon (HeNe) laser is a very well-known gas laser. It produces a bright red, continuous beam of low power. It is used for many applications such as scanning, alignment, measurement, and stabilization devices. University students use them in optical training laboratories. Many larger lasers contain a HeNe inside the beam path, as well to verify beam alignment. HeNe lasers are fairly cheap and very rugged. They can work continuously for thousands of hours.

CO_2 lasers are in the gas family. These lasers were the earliest, truly high-power lasers and have been among the most crucial lasers used in R&D for high-energy laser (HEL) weapons. In industry, the more powerful CO_2 lasers are used for welding, drilling, and cutting. There are many different types of CO_2 lasers that vary in pumping design. CO_2 lasers work by burning a hydrocarbon fuel (like kerosene or methane) in oxygen or nitrous oxide. The hot gas flows through a comb of nozzles, expands quickly, and achieves population inversion. The gas then flows through an optical resonator at supersonic speeds, resulting in stimulated emission and a laser beam.[4]

CO_2 lasers have been researched for use as nonlethal weapons. The wavelength produced by a CO_2 laser is also absorbed by glass. For example, the beam does not penetrate a windshield. Thus, shooting a CO_2 laser at a vehicle's windshield could deter a threat by damaging the windshield or by causing a dazzling effect to reduce the visibility of the driver, while not reaching the driver at all.

The gas dynamic laser (GDL) is a CO_2 laser based on differences in relaxation velocities of molecular vibrational states. The laser medium's gas has properties such that an energetically lower vibrational state relaxes faster than a higher vibrational state; thus, a population inversion is achieved in a particular time. A GDL is shown in Figure 4. Characteristics of chemical and gas lasers are identified in Table 2.

Fiber Lasers

Modern fiber lasers are considered SSLs. They are powered by electricity, making them highly mobile and supportable on the battlefield. Fiber lasers use optical fibers as the gain media. In most cases, the gain medium is a fiber doped with rare earth elements—such as erbium (Er3+), neodymium (Nd3+), ytterbium (Yb3+), thulium (Tm3+), or praseodymium (Pr3+)—and one or several laser diodes are used for pumping. Optical fibers have been used in industry, specifically for telecommunications to transport information via light. With developing technology, optical fibers have become high-energy, powerful laser energy sources. Fiber lasers have proven to have much benefit over traditional SSLs. They are rugged and do not require a clean room to operate or maintain, as most other laser systems do. They also are extremely efficient; however, they cannot operate well in all weather conditions. One example is the IPG CW fiber lasers, which produce moderate beam quality, causing damage to materials and components through thermal heating and burn-through. The Naval Surface Warfare Center, Dahlgren Division (NSWCDD) purchased eight commercially available 5.5-kW IPG lasers, where two multimode (seven fibers) lasers are housed per cabinet. This type of laser is easy to mount due to the flexible fibers. The IPG CW Fiber Laser is shown in Figure 5.

Miscellaneous Lasers

There are other types of lasers that do not necessarily fit into the chemical or solid-state categories. These include semiconductor lasers, used in:

- Television
- CD Players
- Dye Lasers
- Spectroscopy
- Radios
- Telecommunications
- Medicine
- Astronomy

There also are the FELs mentioned previously. The FEL is a completely different breed of laser.

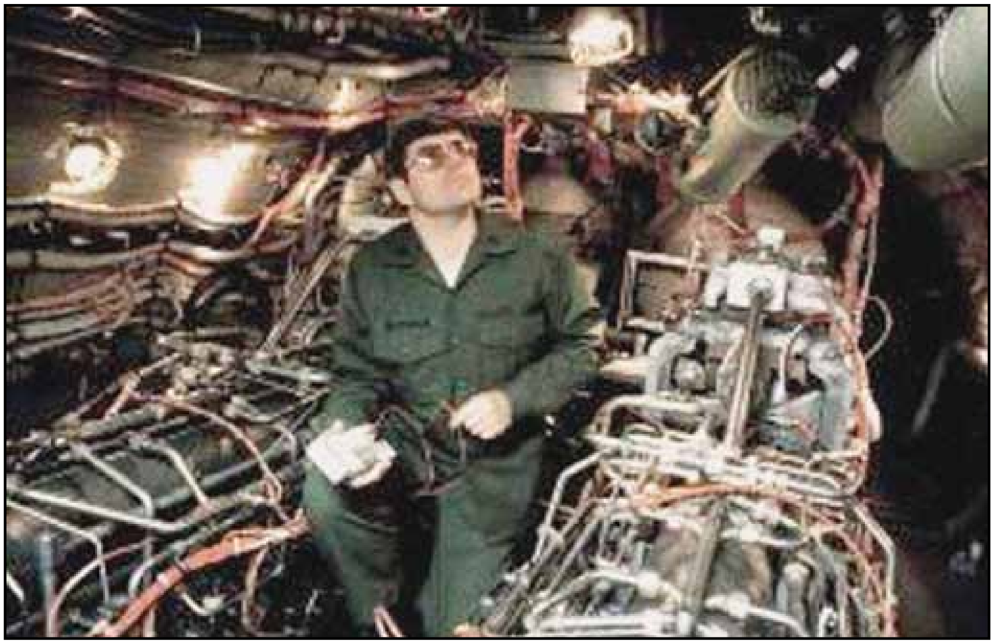

Figure 4. A laser engineer inspects a gas dynamic laser after installation aboard an NKC-135 airborne laser laboratory.

Table 2. Characteristics of Chemical/Gas Lasers

Name	Wavelength (nm)	Typical Power	Typical Operation
Helium-Neon	543/632.8	.0001–.001 watts	CW
Krypton	350–647	.0001–.05 watts	CW
Argon	350–514.5	.001–6.0 watts	CW
Xenon fluoride (excimer)	351	.001–20 watts	CW
Argon fluoride (excimer)	193	.05–30 watts	Pulsed
Krypton fluoride (excimer)	249	7–100 watts	Pulsed
Deuterium fluoride (chemical)	3,000–4,200	.01–100 megawatts	Pulsed/CW
Hydrogen fluoride (chemical)	2,600–3,000	.01–150 megawatts	Pulsed/CW
Carbon dioxide	9,000–12,000	.1–15,000 megawatts	Pulsed/CW
GaAlAs (semiconductor)	750–900	10–4,000 milliwatts	Pulsed

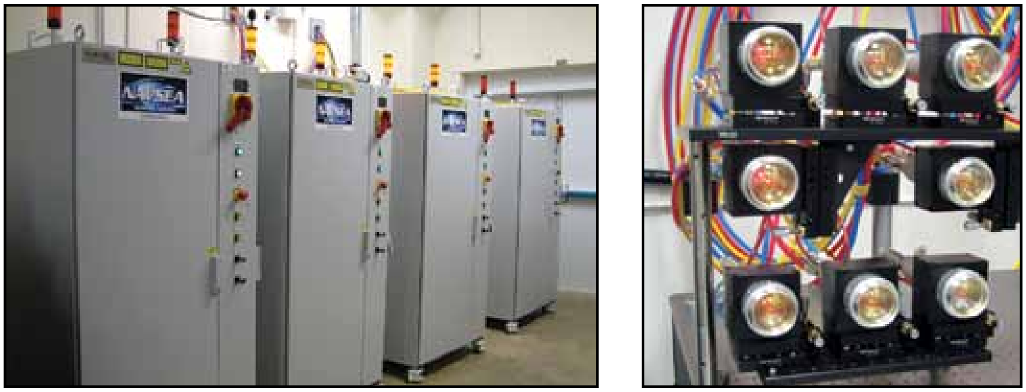

Figure 5. IPG CW Fiber Laser System

It uses electrons to create photons instead of some type of matter. The electrons are produced, collected, and directed to flow at very high speeds. To excite the electrons, they are passed through a "wiggler," i.e., a series of magnets positioned in such a way that electromagnetic radiation (light) is produced when the electrons release photons. The significant feature of the FEL is that the wavelength can be controlled, depending on the magnet positions and the speed of electrons. This versatility makes the FEL particularly appealing. However, the footprint of the FEL system is too large to transform into any ideal defense weapon. The Jefferson Laboratory in Newport News, Virginia, has an FEL and continues to maintain and test its capabilities and effects. This laser was new to the military in the late 1990s and received funding to optimize its capabilities and integrate as a defense weapon. Although great progress has been made, the required footprint could be much larger than desired. Consequently, some interest in the FEL has shifted to other HEL sources.

Many scientists foresee the probability of using the laser as a global weapon. This possibility is proven through basic laws of physics. Actually implementing such a system, however, can be more difficult. The global weapon concept uses a base laser with optics and is strategically positioned in space to be able to direct its beam multiple places on Earth at the speed of light with maximum power levels. This idea faces significant problems, including appropriate power levels, optics to handle such levels, propagation issues, and the ethical measures behind any global weapon. Still, the idea presents interesting possibilities.

LASER WEAPON DEVELOPMENT

The following paragraphs highlight some of the laser weapons that have been successfully developed over the last 40 years.

Baseline Demonstrator Laser (BDL) Hydrogen Fluoride (HF)

In 1973, TRW Inc. produced the world's first high-energy chemical laser, the Baseline Demonstration Laser, for DoD. After that, TRW Inc. produced and demonstrated six more HELs, including the MIRACL (1985) and Alpha (2000), the nation's only megawatt-class chemical lasers.

Navy-ARPA Chemical Laser (NACL) HF

The NACL was mated with the Navy Pointer Tracker at TRW Inc.'s San Juan Capistrano, California, facilities in the 1975–1978 time frame. This was the Navy's initial, integrated HEL system test bed and was used to provide the first demonstrated kill of an operational missile in 1978.

Alpha HF—Built for Strategic Defense Initiative (SDI) Space-Based Laser (SBL)

Alpha, an HF laser, was the baseline technology for the SBL readiness demonstration (SBLRD). In 1991, the Alpha laser demonstrated megawatt-class power levels similar to MIRACL, but in a low-pressure, space operation environment. Alpha demonstrated that multimegawatt, space-compatible lasers can be built and operated.

Tactical High-Energy Laser (THEL)

The THEL is a DF chemical laser developed by the Army. In 2000 and 2001, THEL shot down 28 Katyusha artillery rockets and 5 artillery shells. On 4 November 2002, THEL shot down an incoming artillery shell and a mobile version successfully completed testing. Subsequently, during a test conducted on 24 August 2004, the system successfully shot down multiple mortar rounds. These tests represented actual mortar threat scenarios in which both single mortar rounds and salvo were tested and intercepted. A photograph of THEL is shown in Figure 6.

Advanced Tactical Laser (ATL)

The ATL uses a closed-cycle, chemical oxygen-iodine laser (COIL) with beam control, which lases at a 1.315-µ wavelength. The ATL was developed to engage tactical targets from a moving platform at ranges of approximately 10 km. It can spot a 10-cm-wide beam on a distant target for up to 100 shots. This beam has enough power to slice through metal at a distance of 9 miles. The aircraft equipped with the ATL weapon system is shown in Figure 7.

A specially modified 46th Test Wing NC-130H aircraft equipped with the ATL weapon system fired its laser while flying over White Sands Missile Range, New Mexico, successfully hitting a target board located on the ground. Equipped with a chemical laser, a beam control system, sensors, and weapon-system consoles, the ATL is designed to damage, disable, or destroy targets with little or no collateral damage.

Airborne Laser (ABL) (CO_2) Chemical Oxygen

The ABL C-130H aircraft contains three laser beam systems: the powerful killing primary laser beam (ATL), a set of illuminating laser beams for infrared surveillance and high-speed target acquisition, and a beacon laser for a high-precision laser target tracking beam control system. The primary

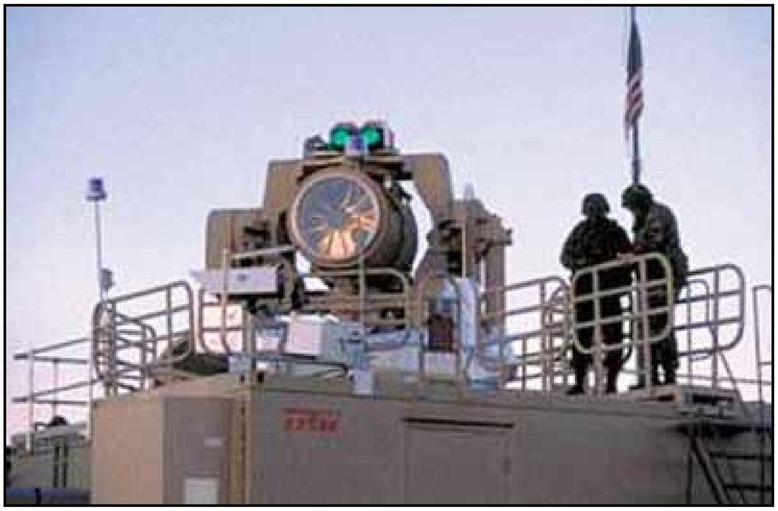

Figure 6. Tactical High-Energy Laser (THEL)

Figure 7. 46th Test Wing NC-130H Aircraft Equipped with the ATL Weapon System

laser beam is generated by a megawatt COIL located at the rear of the fuselage. The high-power laser beam travels towards the front of the aircraft through a pipe. The pipe passes through a Station 1000 bulkhead/airlock, which separates the rear fuselage from the forward cabins. The high-power beam passes through the fine beam control system mounted on a vibration-isolated optical bench. Beam pointing is achieved with very fast, lightweight steering mirrors, which are tilted to follow the target missile. The ABL finally destroyed a target while in flight at White Sands Missile Range in August 2009. The 12,000-lb ABL locked onto an unspecified ground target and fired the laser, making the target disappear. Although it was successful at this demonstration, using the ABL in the fleet has fallen out of favor due to affordability and technology problems. The ABL is shown on an aircraft in Figure 8.

Joint High-Power Solid-State Laser (JHPSSL)

In hopes of accelerating SSL technology for military uses, work is being performed by the U.S. Army Space and Missile Defense Command (SMDC) and the Army Test and Engineering Center at White Sands Missile Range. The technology uses an electric laser diode to shoot light into 32 garnet crystal modules that combine to create “laser amplifier chains” producing 15 kW. By using seven chains and by combining multiple beams, they have reached 105 kW in the laboratory operating in a clean room. The program’s ultimate goal is for a laser system to reach high powers outside a laboratory environment. Fielding such a delicate optical structure can present significant barriers for this laser system. Nonetheless, it will be a great accomplishment for a variety of force protection missions, such as shipboard defense against cruise missiles. The JHPSSL system is shown in Figure 9.

Navy Laser Weapon System (LaWS)

The Navy LaWS is the most recent, successful laser weapon. It uses an electric-fiber laser design, avoiding the problems that chemical lasers present. In the summer of 2009, the Naval Sea Systems Command (NAVSEA)—with support from NSWCDD—successfully tracked, engaged, and destroyed unmanned aerial vehicles (UAV)

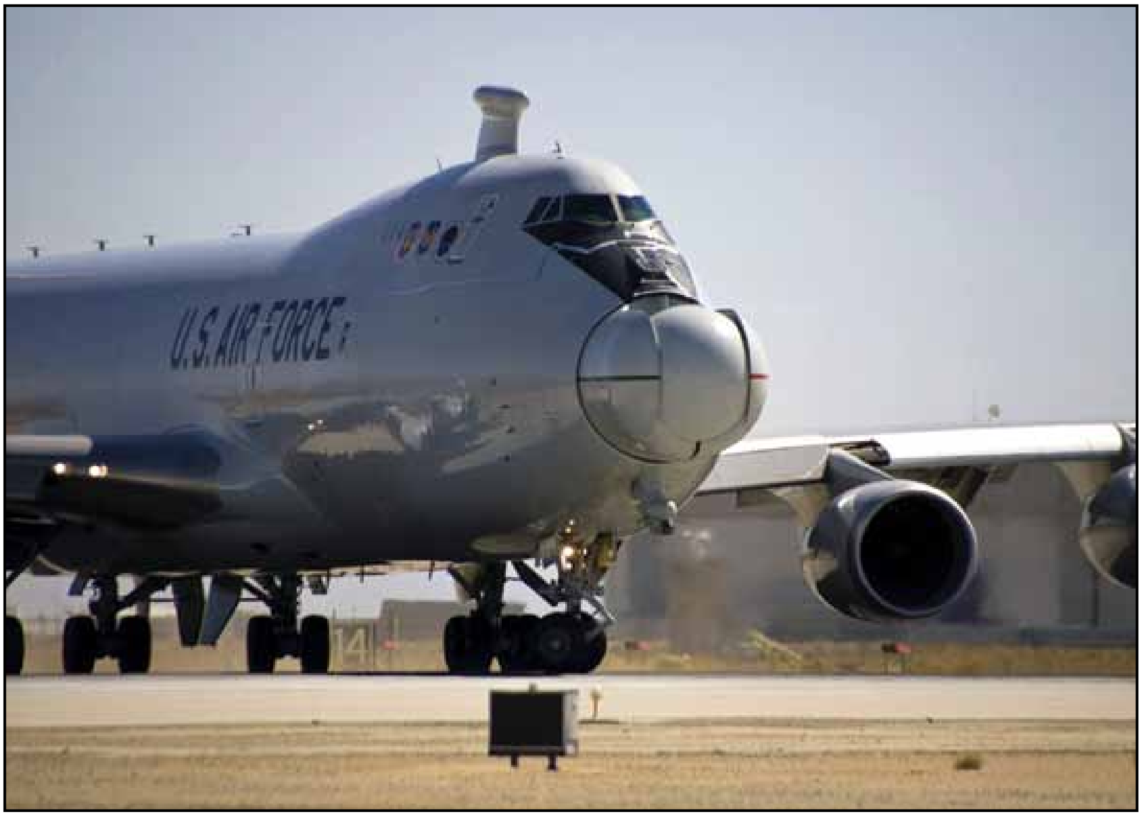

Figure 8. Airborne Laser (ABL)

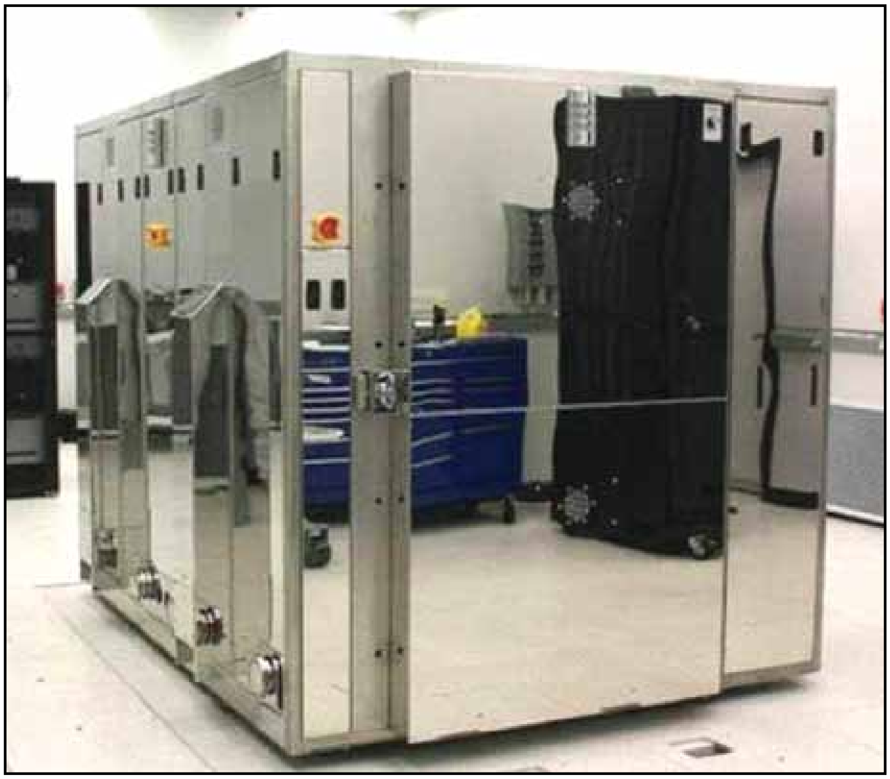

Figure 9. Joint High-Power Solid-State Laser (JHPSSL) System

in flight at the Naval Air Warfare Center, China Lake, California. A total of five targets were engaged and destroyed during the testing, which represented a first for the U.S. Navy. The laser was fired through a beam director on a Kineto Tracking Mount similar to the Sea Lite beam director. The system used fiber lasers in the configuration and has proven to be a rugged and dependable choice for the warfighter's needs. A photograph of LaWS is shown in Figure 10.

Laser weapon systems development in recent years has taken giant steps forward. Dedicated R&D has advanced the state of the art considerably. What was unimaginable only a few short years ago, today has become reality. Accordingly, given continued R&D, warfighters in the near term will have additional weapon options to choose from for dealing with a spectrum of threats and contingencies.

REFERENCES

1. Beason, Doug Ph.D., *The E-Bomb: How America's New Directed Energy Weapons Will Change The Way Future Wars Will Be Fought*, Da Capo Press, 2005.
2. Pike John, *Mid-Infrared Advanced Chemical Laser (MIRACL)*, FAS Space Policy Project, Military Space Programs, March 1998, http://www.fas.org/spp/military/program/asat/miracl.htm
3. Albertine, John R., "History of Navy HEL Technology Development and Systems Testing," (Proceedings Paper), *Proc. SPIE*, Vol. 4632, Laser and Beam Control Technologies, Santanu Basu and James F. Riker, Eds., pp. 32–37.
4. Anderberg, Major General Bengt and Dr. Myron L. Wolbarsht, *Laser Weapons: The Dawn of a New Military Age*, Plenum Press, New York, 1992.

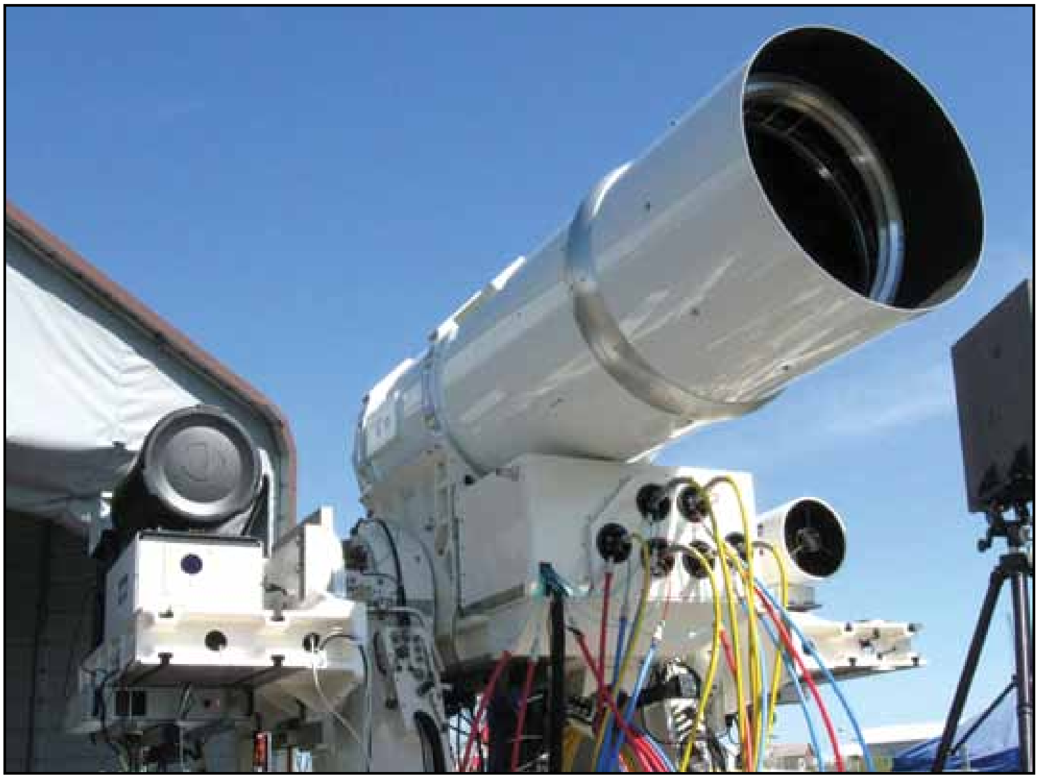

Figure 10. Navy Laser Weapon System (LaWS)

LASER WEAPON SYSTEM (LAWS) ADJUNCT TO THE CLOSE-IN WEAPON SYSTEM (CIWS)

By Robin Staton and Robert Pawlak

The Naval Sea Systems Command (NAVSEA) established the Navy Directed Energy Weapons Program Office in January 2002 and subsequently chartered the Directed Energy and Electric Weapon Systems Program Office (PMS 405) in July 2004.[1, 2] Its mission is to change the way the Navy fights in the 21st century by transitioning directed-energy and electric weapon technology, providing the warfighter with additional tools to fight today's and tomorrow's wars. In support of this mission, the Laser Weapon System (LaWS) was developed, which potentially adds a suite of tools for offensive and defensive operations.

The LaWS program is managed by PMS 405 in cooperation with the Program Executive Office Integrated Warfare Systems (PEO IWS), the Navy's Close-In Weapon System (CIWS) manager. A multilaboratory/multicontractor organization led by the Naval Surface Warfare Center, Dahlgren Division (NSWCDD), has been executing the program since March 2007. The potential advantages of a lethal, precise, speed-of-light weapon are numerous and have been recognized for many years. However, even in light of these advantages, there are realities that need to be considered for any program to succeed to the point that an actual system is placed in the hands of the warfighters.

The LaWS system offers viable solutions for an important subset of threats while fitting into acceptable size and weight constraints. In addition, since LaWS is a fully electric laser, the operation of the system does not require the handling and storage of hazardous chemicals, such as hydrogen fluoride. As will be discussed later, due to the incorporation of high levels of commercial off-the-shelf (COTS) technology, the LaWS system also has advantages for topside design, logistic supportability, and cost. Thus, LaWS could enable the Navy to address adverse cost-exchange situations, which can occur when engaging proliferated inexpensive threats such as unmanned aerial vehicles (UAVs).

BACKGROUND

Based on mission analysis work conducted prior to the LaWS program and additional work done as part of the program, it became clear that a number of factors require careful consideration. First, a high-power laser is not likely to replace anything on a ship in the next 5 years. For a new system to be added to a ship, a high-power laser must supplement current capabilities or provide new capabilities that clearly justify its addition. Second, because a laser provides such a diverse set of capabilities, conventional air-to-air warfare (AAW) models—such as the Fleet AAW Model for Comparison of Tactical Systems (FACTS), Antiair Warfare Simulation (AAWSIM), and Extended Air Defense Simulation (EADSIM), as well as other existing AAW analysis approaches—are not well suited for showcasing current or near-term laser-weapon capabilities. While they can (and have) been used for laser-weapon analysis, their application to a megawatt-class laser that could "instantly" destroy boats or cruise missiles (akin to missile engagements) is a more straightforward application of the existing models and techniques.

In November 1995, the Chief of Naval Operations requested that the National Research Council initiate, through its Naval Studies Board, a thorough examination of the impact of advancing technology on the form and capability of the naval forces to the year 2035. A major observation of the report is quoted below:

> Numerous laboratory and field-test versions of laser weapons have been developed and demonstrated. They have worked as expected and demonstrated suitable lethality against their intended targets. The primary factors that have inhibited the transition of the technology into deployed systems are size and weight. Generally, the conceptual designs of laser weapons that are scaled for combat effectiveness are too large to be appealing to users; conversely, weapons that are sized for platform convenience generally lack convincing lethality.[3]

Subsequently, an August 2006 U.S. Air Force (USAF) Scientific Advisory Board Study examined the increasing threat posed by UAVs in some detail. Key conclusions included:

> No single system can completely address the UAV threat. A single sensor solution is inadequate because of the size and speed challenges presented by small UAVs. A single-weapon-layer solution fails to provide for adaptability to multiple scenarios or adequate probability of kill.

Key recommendations of the USAF Advisory Scientific Board Study included:

> Develop and field longer-term upgrades to counter increased UAV threats. They include:…a small, multimission air/air and air/ground weapon; and directed-energy air defense weaponry.[4]

In addition to the USAF Scientific Advisory Board study, a 2007 OPNAV Deep Blue Study noted the potential advantage of nonkinetic defeat options and recommended that the Navy accelerate development of nonkinetic systems to include high-energy lasers (HELs).[5]

The laser power levels likely to be available in the near term, within reasonable size and cost, are in the neighborhood of 100 kW of radiated power. While this power level is not adequate to engage certain threats, such as cruise missiles or tactical ballistic missiles at tactically useful ranges, there is still a wide spectrum of threats that could be engaged at ranges that are comparable to many current ship-defense weapons, including minor-caliber guns and small missiles. The spectrum of threats includes:

- UAVs
- Missile Seekers
- Intelligence, Surveillance, and Reconnaissance Systems
- Rockets
- Man-Portable Air-Defense Systems (MANPADS)
- Mortar Rounds
- Floating Mines
- Artillery Rounds

LaWS on CIWS

The Mk 15 Phalanx CIWS can often detect, track, and (sometimes) identify potential threats at ranges well outside the effective range of the 20mm gun. These functions are accomplished using the search/track radar system and the Phalanx Thermal Imager (PTI). When added to the Phalanx mount and pointed in the same direction as the gun (see Figure 1), a laser weapon could potentially add a number of useful functions and capabilities to the mount, but technical challenges must be overcome. Preliminary analyses of the mechanical characteristics of the mount suggest that the additional weight that could be added to the mount must be kept under approximately 1200–1500 lb. Additionally, it is highly desirable that the addition of the laser weapon not substantially affect the train/elevation operation of the mount in angle, peak velocity, or acceleration. Consequently, use of rapidly evolving fiber laser technology appears to be the only currently foreseeable path to adding significant laser energy directly to the mount within these constraints.

One major driver in the genesis of the LaWS system was the availability of relatively low-cost COTS fiber-optic lasers. Because these fibers are flexible, they obviate the need for an expensive coudé path system (an optical mirror/lens assembly that turns radiation 90° and may also support rotation of the beam director), thus allowing the use of low-cost mount technology, as well as the retrofitting of the system on existing mounts. The last factor is extremely important because of the scarcity of topside real estate on today's ships. These fiber-optic lasers do have limitations in terms of power, although power levels are growing with advancing technology. The reality today is that, in order to get adequate lethality from a system based on this technology, the use of a beam-combining apparatus utilizing several individual fibers is necessary. (Figure 2 depicts combining multiple fibers in the same beam director.) Furthermore, a smaller beam size is desirable since this drives power density up—increasing the performance required for the tracking and pointing elements of the system. Thus, a high-resolution fine track sensor is needed, as well as an appropriately robust line-of-sight control.

A Potential Suite of LaWS-Related Capabilities

Potential added capabilities that an adjunct LaWS could contribute to the total ship combat system are briefly outlined in the following subsections.

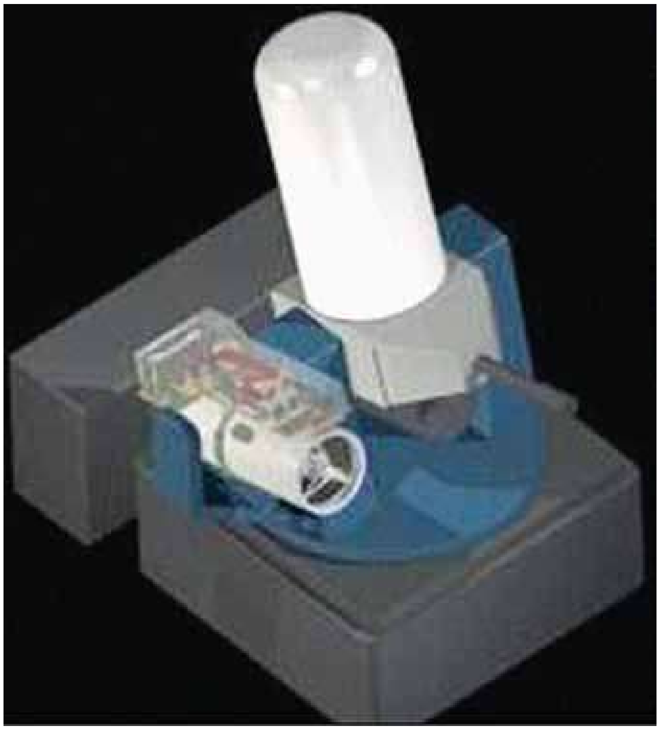

Figure 1. LaWS Mounted on CIWS

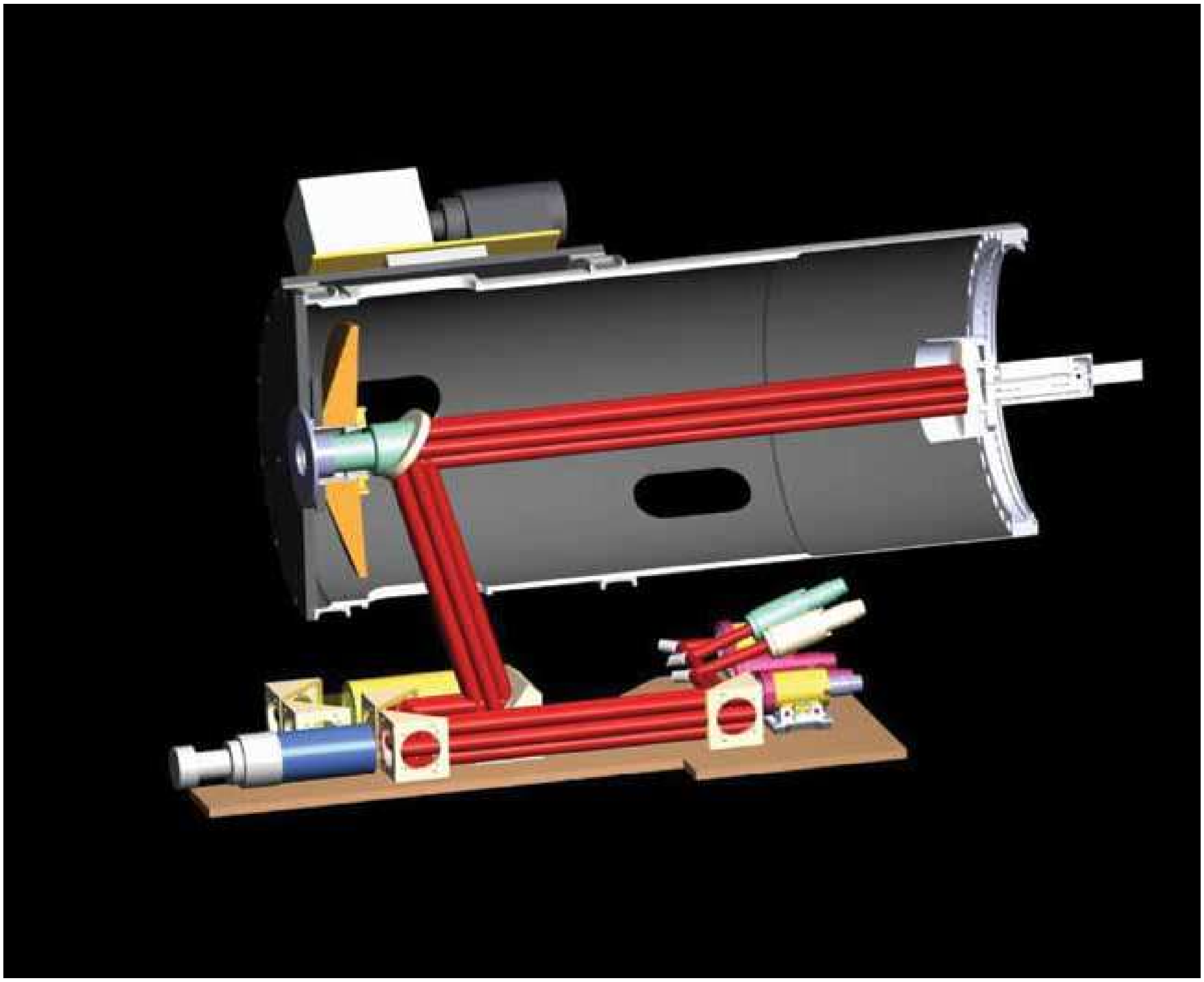

Figure 2. Cutaway View of the LaWS Beam Director

Target Identification, Tracking, and Intent Determination at Range

The optics that would be added for the laser to detect and track targets in support of a laser engagement would immediately contribute additional capabilities to the entire ship combat system even without operating the laser. A laser-gated illuminator, part of the tracking system, significantly increases the signal to the background level of tracked targets and provides good range resolution as well. The additional sensitivity and angle resolution provided by the LaWS optics would allow the identification, precision tracking, and "monitoring" (at high resolution) of potential threats or vehicles of interest at substantially greater ranges than could be achieved by the PTI alone. The Phalanx radar, or another source, would have to provide an initial, accurate cue to facilitate initial acquisition. Once acquired, the target could be examined and monitored with high resolution at range. This capability could make a substantial contribution to identification efforts—efforts to determine intent and potentially even to documenting target behavior to resolve issues with rules-of-engagement doctrine. It is widely recognized that rules-of-engagement issues, such as threat identification and intent determination, are among the most difficult problems faced by ship commanding officers.

Unambiguous Warning at Range

If a fraction of the laser energy is routed through a frequency-doubling crystal, an intense, visible beam can be projected to significant ranges to provide a clear, unambiguous warning that a potential target is about to be engaged unless an immediate change in behavior is observed. This feature also would have utility for dazzling aircraft, surface vehicle, or submarine sensors, and would provide exceptional long-range, unambiguous warning to boats or aircraft at night.

Sensor Destruction at Range

Many electro-optical (EO) sensors are quite susceptible to damage by laser energy in the fiber-laser band as is the case with infrared (IR) missile seekers with germanium optics. The frequency-doubling feature described in the previous paragraph also would be useful to ensure that a band-pass filter at a single frequency could not be applied as an effective countermeasure. The intent here would be to destroy the seeker or imager at ranges well beyond those achievable by the Phalanx 20mm gun. Other examples include intelligence, surveillance, reconnaissance, and targeting sensors on UAVs or unmanned surface vehicles (USVs).

IR Missile Assist at Range

Many targets of interest—including UAVs, USVs, and small boats—are somewhat "marginal" from a target-signature standpoint, particularly at the maximum range of existing IR guided missiles such as the FIM-92 Stinger, the FGM-148 Javelin, the RIM-116 RAM, and the AIM-9X Sidewinder. The CIWS laser adjunct could potentially "correct" this situation by laser heating target vehicles to enhance their signature to existing IR guided missiles. Note that this is NOT "conventional" semiactive-laser (SAL) guidance—the LaWS is not a coded illuminator, nor do the seekers in question rely on this coding. The IR missiles would be unmodified weapons taken from inventory. The LaWS adjunct would simply contribute laser energy that heats the target and enhances its signature for the missile. While, at the ranges envisioned, this laser heating alone would not be sufficient to "kill" the target, it could definitely heat the target. It should also be noted that the laser "illumination" could potentially be used to preferentially select a specific target from among a group of targets for engagement by a missile. It is expected that these engagements could occur at ranges of two to four times the effective Phalanx gun engagement ranges. Use of LaWS in this manner would be exactly analogous to the use of a SAL designator for a SAL guided missile, such as the AGM-114 Hellfire. It is expected that similar rules of engagement would apply.

Direct Target Destruction by Laser Heating

Some threats are known to be vulnerable to direct destruction by the application of laser energy for an appropriate period of time. The currently envisioned system would be able to destroy a subset of naval threats at ranges comparable to, and in some cases greater than, the ranges achieved with modern, stabilized guns using EO fire control systems and modern ammunition. In the case of a LaWS adjunct, the addition of the laser would open new options for a firing/engagement doctrine and would be expected to conserve CIWS rounds for use on threats that are not appropriate for this laser power level. While the laser is often quoted as having an "unlimited magazine," the true number of threats that can be engaged by the laser in any period of time is limited by the required illumination time and by the time required to evaluate a kill and transition to the next target. Thus, for particular target velocities and numbers, the "effective laser magazine" might be added to the CIWS magazine to increase the total number of targets engaged by the combined system.

LAWS ACCOMPLISHMENTS

A government/industry team, led by government technical personnel, have achieved significant accomplishments since the start of the LaWS program in 2007; specifically, the team:

- Conducted mission analyses
- Developed threat lethality estimates
- Performed industry surveys for critical components and subsystems
- Performed extensive trade-off analyses
- Designed a prototype system
- Constructed the system—the prototype director and mount (see Figure 3)
- Performed numerous laboratory-based tests of subsystems and the complete prototype
- Validated system operation with a full-up field test at high power using BQM-147A UAV target drones

Additionally, the team was able to minimize the cost of the prototype by leveraging hardware that had already been developed or procured for other applications, including an L3-Brashear tracking mount, a 50-cm telescope, and high-performance IR sensors. Some components were commercially procured, such as the 5.4-kW fiber lasers. Figure 4 shows three laser cabinets, containing two lasers apiece, resulting in a total power output of 32.4 kW. Other components, such as the beam combiner and much of the system software required for operation and target tracking, had to be specifically designed, fabricated, and tested.

The LaWS program achieved a highly successful field test/demonstration in June 2009 when the prototype successfully engaged and destroyed five drone targets at tactically significant ranges at the China Lake, California, test range (see Figure 5).

ADDITIONAL WORK TO BE DONE

Since the LaWS prototype sits on a dedicated gimbal, much additional work needs to be done

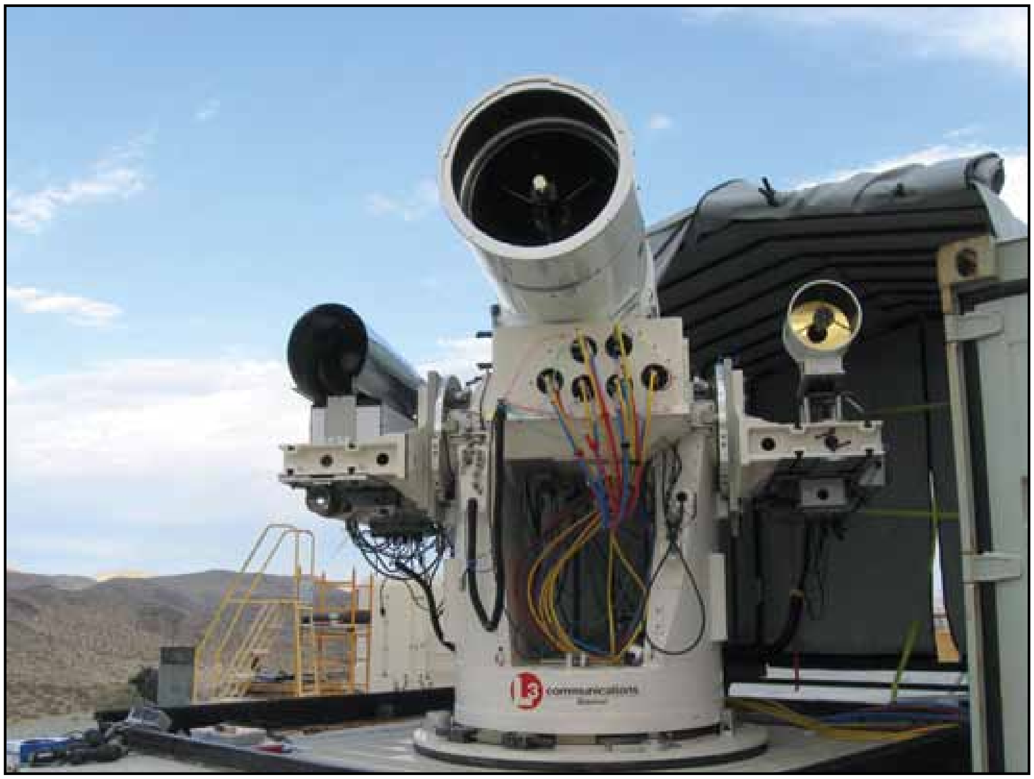
L3 communications

Figure 3. Photo of LaWS During Testing at the Naval Weapons Center, China Lake

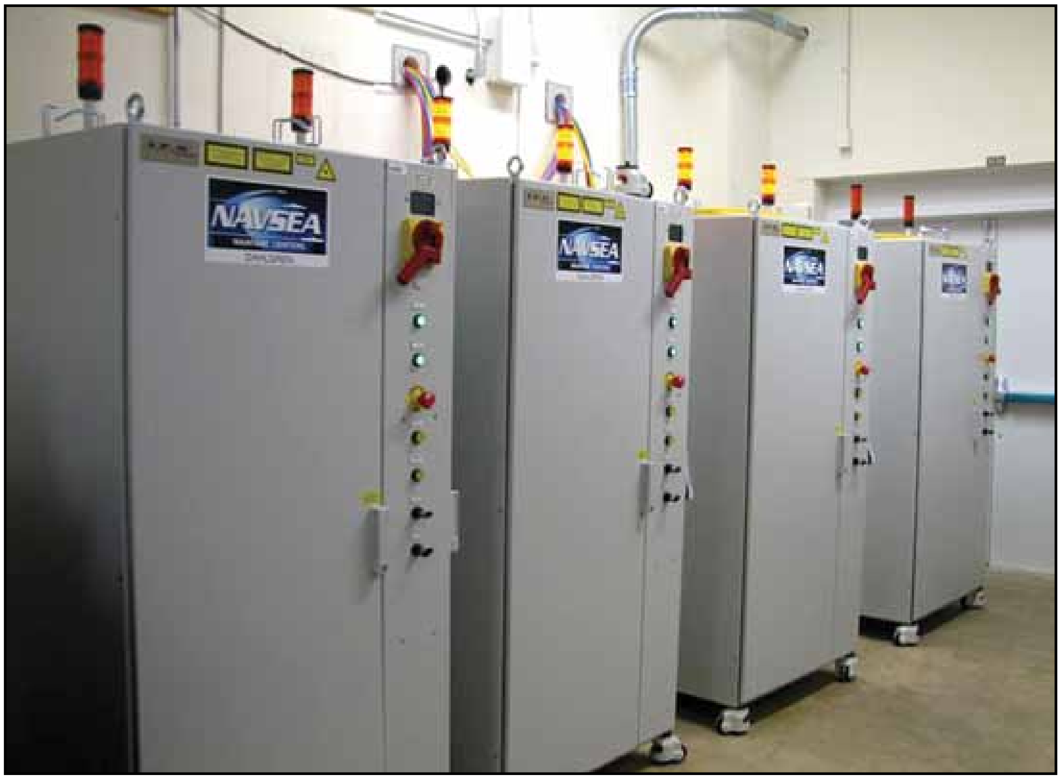

Figure 4. IPG Laser Cabinets

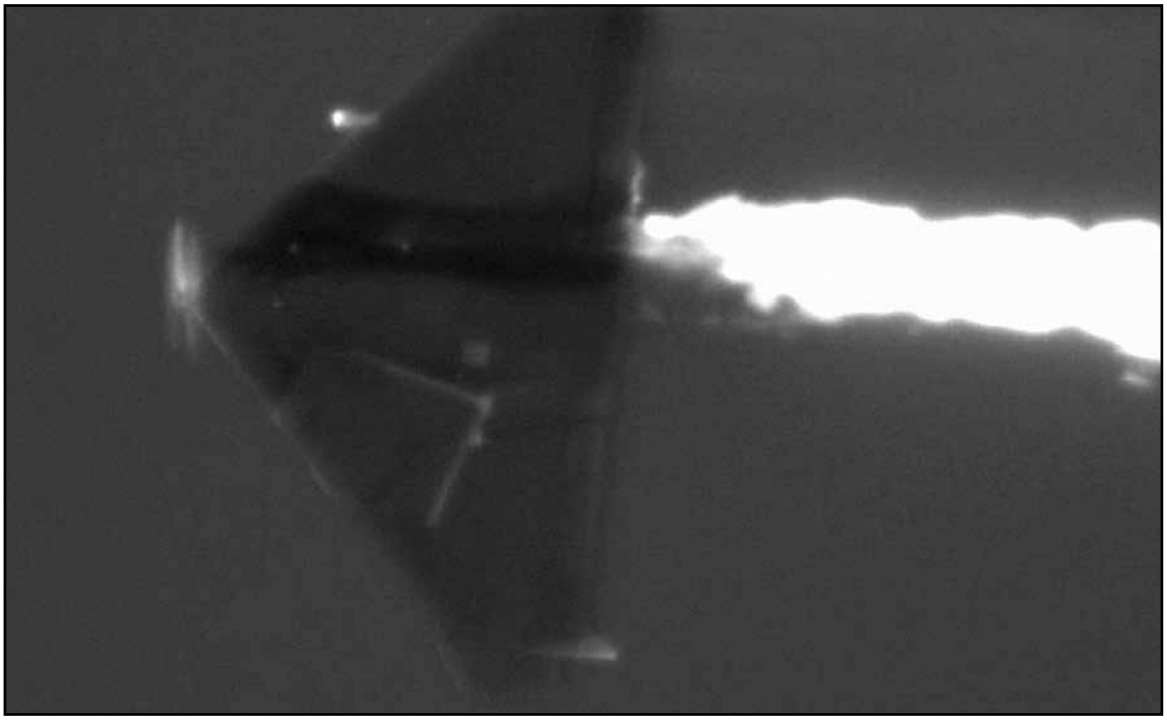

Figure 5. BQM-147A During LaWS Engagement

to place the weapon on the CIWS mount. The latter would require new control systems and optomechanical hardware for line-of-sight stabilization. Other aspects of the shipboard environment are also more stressful, and future mission areas may require an increasingly robust capability to deal with optical turbulence and the high-clutter environment of the ocean surface. Additional laser power might also be required. These modifications, depending on the level of capability desired, will require engineering modifications to the system. Engineering analysis and design to address these issues is currently underway at NSWCDD.

While the aforementioned engineering issues are important to address, there are additional technical issues that have yet to be analyzed. These issues are concerned with the potential utility of the system. Indeed, most of the detailed technical analyses and experiments performed thus far have focused on target destruction, with some effort expended on the issue of seeker damage/destruction. Developing credible lethality estimates for various potential threat targets is clearly very important, but one consequence of the lethality focus is that necessary, detailed, defendable technical analysis, analytic model development, and experiments have not been performed to explore the other functions/features that a CIWS Adjunct LaWS might provide to the overall ship combat system. Some of these contributions might become "routine" if the LaWS were available.

For example, a hard-kill engagement of a target by a Navy shipboard weapon is a relatively rare event, even during wartime conditions. On the other hand, ships in combat zones—and elsewhere—constantly have the problem of detecting potential threats, tracking them, identifying them, determining their intent, and providing warning. Thus, use of the LaWS system, at less than its full lethal potential, could become a daily, standard practice. It is still not clear how these potential benefits and capabilities could be measured or quantified to the satisfaction of key decision makers.

Likewise, other potential advantages of laser weapons—such as the potential for precision engagement, covert engagement, fire starting, graduated lethality, low cost per shot, and "unlimited" magazine—have not been subjected to rigorous technical analysis for feasibility, utility, and practicality. These investigations need to be performed and are gradually being addressed within the LaWS program.

Although the Phalanx CIWS system is currently installed on a number of Navy surface warships—either a single mount or a double mount—there are still significant numbers of ships that do not have a Phalanx system. It is highly desirable to make LaWS potentially available to virtually any ship that could benefit from the enhanced capabilities.

While the technical issues associated with the addition of LaWS to the Phalanx CIWS will be somewhat different from those associated with adding a LaWS system to other weapon systems—or the provision of a "stand-alone" LaWS—they do not appear to be insurmountable. For example, a LaWS beam director might be added to the stabilized Mk 38 Mod 2 25mm gun or the Mk 46 Mod 2 30mm gun. A LaWS beam director might be added to (or even substituted for) the Mk 46 EO Sight on DDGs or added to the trainable RAM launcher. Other options may exist as well.

The issue of defending combat logistics force ships, joint sealift ships, and certain support vessels from attacks from small boats or UAVs is also relevant. These ships often have little or no installed defensive capabilities for potential terrorist or pirate threats, and expeditionary security detachments do not have decisive warning or engagement capability. In addition, there are severe limitations placed on concept of operations (CONOPS) and rules of engagement due to the limited objectives/limited means of the various missions.

A system such as LaWS could provide graduated lethality from warning to destruction. It also could provide additional applications to minimize risk to sea base platforms and enhance sea shield capabilities against nonstate threats. If acceptable rules of engagement can be established, the advantages of graduated lethality might be extended to ships in port or entering/exiting harbors.

While considerable additional work needs to be done to produce a tactical system, the LaWS program's recent demonstration of capability provides strong evidence that a useful, tactical system could be produced within reasonable cost, volume, weight, and power constraints to provide the warfighter with a suite of additional tools to fight today's and tomorrow's wars.

REFERENCES

1. NAVSEA Notice 5400, Ser 09B/240, "Establishment of the Navy Directed Energy Weapons Program Office (PMS 405)," 4 January 2002.
2. NAVSEA Instruction 5400.101, Ser SEA 06/058, "Directed Energy and Electric Weapon Systems Program Office (PMS 405) Charter," 21 July 2004.
3. *Technology for the United States Navy and Marine Corps 2000–2035: Becoming a 21st-Century Force*, Naval Studies Board, National Research Council, National Academy Press, 1997, Volume 5: Weapons, Chapter 6, "Navy and Marine Corps Applications for Laser Weapons."
4. *Report on the Air Defense Against Unmanned Vehicles*, USAF Scientific Advisory Board, SAB-TR-06-01, August 2006.
5. "Counter Unmanned Aerial Systems (UAS)," *OPNAV Deep Blue Report 08-01*, August 2007.

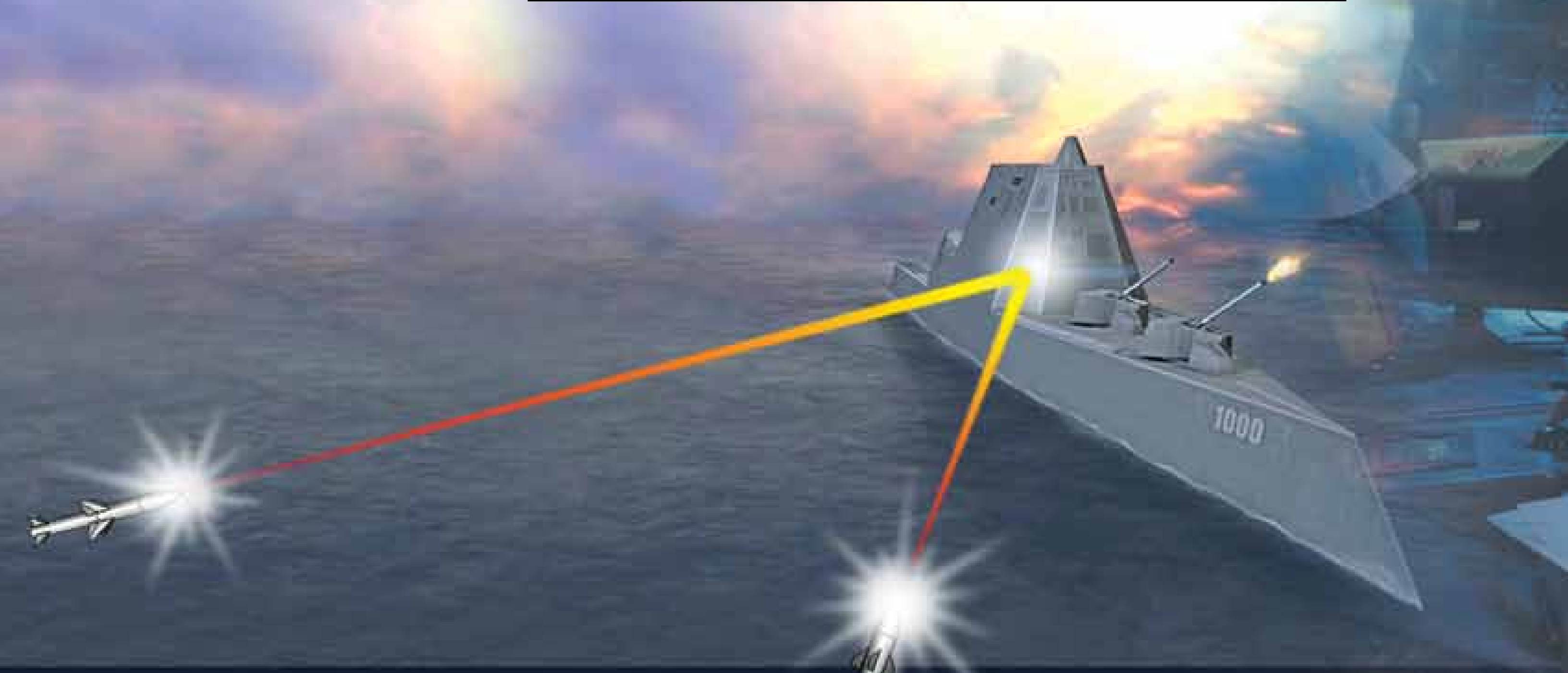

THE ACQUISITION CHALLENGE ASSOCIATED WITH DIRECTED-ENERGY RDT&E

By Mike Kotzian

An already tense situation quickly escalated. Everyone within the combat information center of the Navy's newest all-electric ship suddenly realized that two surface-skimming, antiship missiles were bearing down on their destroyer. With less than 30 seconds to impact, the tactical warfare officer gave the order to fire. Seconds later, the first surface-skimming missile vanished from all tracking consoles. Another order to fire closely followed, and the second missile threat was also destroyed. Consequently, within a matter of 10 seconds from threat recognition to threat elimination, the Navy's newest all-electric ship was able to destroy two incoming threats by using one of the Navy's newest weapon systems—the free-electron laser.

Does this scenario of a Navy all-electric ship, employing a high-energy laser to shoot down enemy surface-skimming antiship missiles, sound like inevitable reality or unattainable science fiction? For scientists and engineers working on directed-energy systems for the Navy, the answer does not lie solely in the advanced technical challenges associated with developing directed-energy weapons. Rather, the answer also lies in how well scientists and engineers understand and adhere to the Department of Defense's (DoD's) Defense Acquisition Management System (DAMS) framework governing the development of new weapon systems.

EVOLUTION OF DEFENSE ACQUISITION

The way in which DoD identifies needs and subsequently develops, tests, procures, and sustains weapon systems has evolved over time. Today's acquisition foundation can be traced back to the Packard Commission report in 1986, where many of this report's recommendations became part of the Goldwater-Nichols DoD Reorganization Act of 1986. This evolution continued along three tracks:

1. Requirements moving from threat-based to capability-based
2. The resource allocation system adding execution reviews with concurrent program and budget reviews

3. The acquisition process attempting to incorporate a more flexible and tailored process

These three tracks form the Defense Support System organizational structure: the Joint Capabilities Integration and Development System (JCIDS) process; the Planning, Programming, Budgeting, and Execution (PPBE) process; and the DAMS process, respectively. These three processes operate as "systems of systems" and are referred to as the "Big A" acquisition process shown in Figure 1.[1]

While all three of these phases hold their own level of importance, the major focus for scientists and engineers at research and development (R&D) facilities is the "Little a" acquisition process. It is this "Little a" acquisition process, where the rules and processes are found, that governs how DoD goes about developing a new materiel solution to a validated warfighter requirement. These rules and processes are codified within DoD Instruction 5000.02, *Operation of the Defense Acquisition System*, which was issued in December 2008.

The acquisition framework associated with DoD Instruction 5000.02 is the DAMS structure. This framework, shown in Figure 2, consists of numerous strategically placed milestones and major program reviews to ensure proper programmatic oversight.[2] Each of the milestones has specific

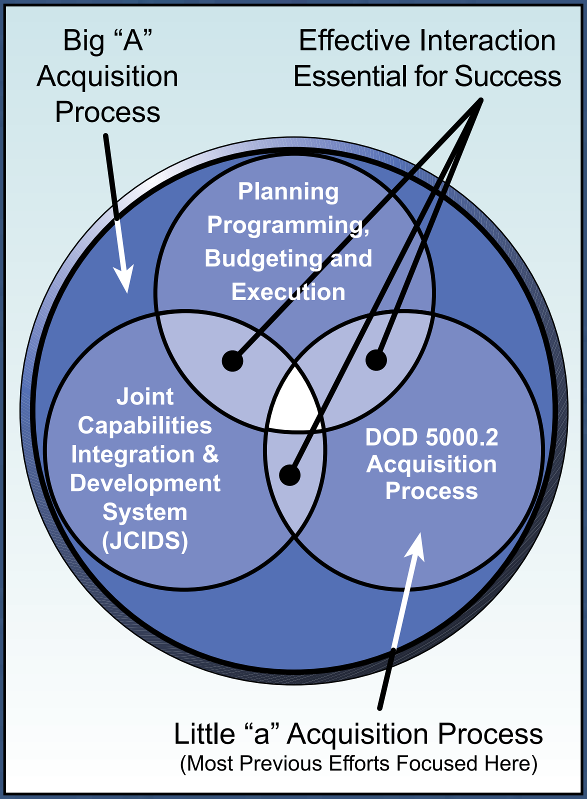
Big "A" Acquisition Process
Effective Interaction Essential for Success
Planning Programming, Budgeting and Execution
Joint Capabilities Integration & Development System (JCIDS)
DOD 5000.2 Acquisition Process
Little "a" Acquisition Process
(Most Previous Efforts Focused Here)

Figure 1. Defense Support System Organizational Structure

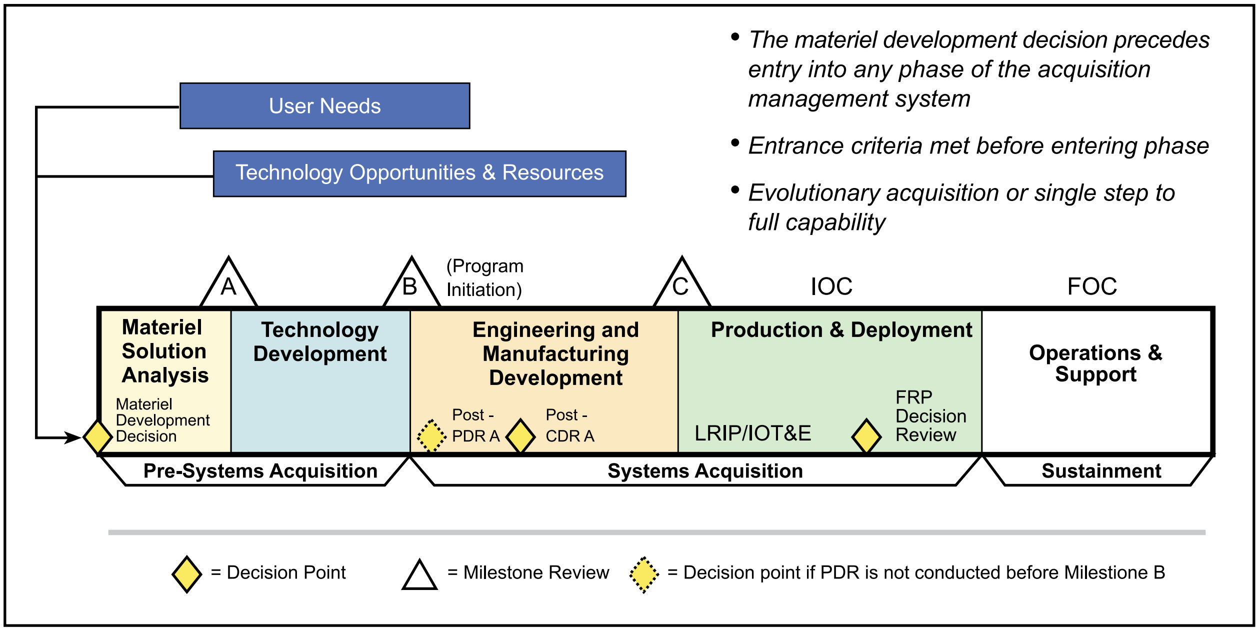
The materiel development decision precedes entry into any phase of the acquisition management system
Entrance criteria met before entering phase
Evolutionary acquisition or single step to full capability
User Needs
Technology Opportunities & Resources
A
B
(Program Initiation)
C
IOC
FOC
Materiel Solution Analysis
Materiel Development Decision
Technology Development
Engineering and Manufacturing Development
Post - PDR A
Post - CDR A
Production & Deployment
LRIP/IOT&E
FRP Decision Review
Operations & Support
Pre-Systems Acquisition
Systems Acquisition
Sustainment
= Decision Point
= Milestone Review
= Decision point if PDR is not conducted before Milestone B

Figure 2. DoD Acquisition Framework

criteria that must be satisfied before a program is allowed to further proceed along the DAMS. The program's Milestone Decision Authority (MDA) rests with the individual responsible for deciding if the milestone criteria have been met and, if so, for allowing the program to proceed to the next phase of the acquisition process. Designation of a program's MDA depends on a program's level of research, development, test, and evaluation (RDT&E) and procurement funding. For example, an Acquisition Category (ACAT) I program is defined as an eventual total expenditure for RDT&E of more than $365 million in fiscal year (FY) 2000 constant dollars or, for procurement, of more than $2.19 billion in FY 2000 constant dollars. In this case, for an ACAT ID ("D" refers to the Defense Acquisition Board (DAB)) the Under Secretary of Defense for Acquisition, Technology and Logistics (USD(AT&L)) is the MDA; for an ACAT IC ("C" refers to Component or Service), the MDA is the Head of the DoD Component or, if delegated, the Component Acquisition Executive.[3]

In addition, civilian and military workforce members within the DoD whose job responsibilities are deemed acquisition-related find themselves with a training requirement necessary to carry out their acquisition-related job responsibilities. Specifically, these workforce members are required to gain acquisition training and education with the passage of the Defense Acquisition Workforce Improvement Act (DAWIA) signed into law in 1990. The current certification process comprises three levels covering 16 different career fields. Each of these 16 career fields has a set of specific training, education, and experience requirements that must be met in order for an individual to achieve Level 1, Level 2, or Level 3 certification. The Defense Acquisition University (DAU) provides the necessary training classes required for the certification. DAU identifies "core-plus" training classes and continuous learning modules for each level of certification. The core-plus classes and modules are not required for certification but are identified as additional sources of information to assist individuals in becoming more knowledgeable about their career field beyond the minimum standards required for certification. The most up-to-date certification frameworks for all 16 career fields can be found at the following DAU website: http://icatalog.dau.mil/onlinecatalog/CareerLvl.aspx

DEFENSE ACQUISITION REFORM

The DoD acquisition environment is undergoing continuous change. The issuance of DoD Instruction 5000.02 marked the opening salvo of what has become seemingly constant updates, modifications, and guidance impacting how DoD procures weapon systems to meet warfighter requirements. In addition to DoD's issuance of DoD Instruction 5000.02, the Government Accountability Office published a stream of reports and findings that indicate significant cost growth and schedule delays in major defense acquisition programs. In 2009, Secretary of Defense Robert M. Gates proclaimed

a new way of doing business within DoD when it comes to weapon systems acquisition. Pressures are building for every program to maintain cost and schedule estimates while delivering the technical requirements originally developed to support the warfighter.

Moreover, there have been two major policy issuances. As previously mentioned, the first was DoD Instruction 5000.02 in December 2008. This update of the rules and processes governing DoD weapon systems acquisition primarily impacted the early part of the DAMS framework. The problem was that weapon system programs were failing their initial operational test and evaluation phases at alarming rates—many times traced to program offices attempting to design weapon systems with immature technology. Such failures were preventing those programs from proceeding to a full-rate production decision review and, more importantly, causing a repeat of some of the DAMS framework, which translated to increased costs and delayed initial operational capability timelines.

DoD Instruction 5000.02 attempted to solve this problem with three main emphases. First, a mandatory requirement was inserted for competitive prototyping prior to program initiation at Milestone B. The intent was to ensure a competition among contractors competing for a contract award. The theory was that such a competition would reduce technical risk, validate designs, improve cost estimates, evaluate manufacturing processes, and refine requirements. Reducing technical risks was especially important because weapon system programs were expected to demonstrate a technology readiness level (TRL) of six—where the system/subsystem model or prototype is demonstrated in a relevant environment—by the time a program reached Milestone B. TRLs are categorized on a scale of 1 to 9. A TRL of 1 is the lowest level of technology readiness, where scientific research begins to be translated into applied R&D. A TRL of 9 is the highest level of technology readiness, where the actual system is proven through successful mission operations. A TRL of 6 represents a major step up in a technology's demonstrated readiness. Using TRLs enables consistent comparisons of technical maturity across different types of technologies, giving program decision makers a common benchmark to consider when assessing program risk. Note that TRLs are meant to capture a level of technical maturity, not the probability of occurrence (i.e., the likelihood of attaining a required maturity level) or the impact of not achieving a level of technical maturity.[4]

The second emphasis was on a stricter adherence to systems engineering processes and technical reviews. Too often weapon system programs were not closely following systems engineering processes or avoiding due diligence when it came to the definition of successful exit criteria for a technical review. Consequently, all technical efforts must be outlined in a program's systems engineering plan. The program manager will use the eight technical management processes—decision analysis, technical planning, technical assessment, requirements management, risk management, configuration management, technical data management, and interface management—to manage the technical development of the system increments, including the supporting or enabling systems.[5] The program manager will use the eight technical processes—stakeholders requirements definition, requirements analysis, architectural design, implementation, integration, verification, validation, and transition—to design the system, subsystems, and components, including the supporting or enabling systems required to produce, support, operate, or dispose of a system.[6] Figure 3 provides an overlay of the new DoD Instruction 5000.02 and Secretary of the Navy (SECNAV Instruction) 5000.2D (Implementation and Operation of the Defense Acquisition System and the JCIDS), and shows the timing of specific systems engineering technical reviews as a program matures through the DAMS.

The third emphasis was a more prominent role of the MDA, starting with a mandatory requirement that all weapon system programs seeking a full or partial materiel solution must hold a Materiel Development Decision chaired by the MDA. Thus, the old Design Readiness Review was replaced with the Post-Critical Design Review Assessment chaired by the MDA. In short, the MDA was to become a more prominent figure in the oversight of a weapon system program's progress.

The second relatively recent major policy issuance was the Weapon Systems Acquisition Reform Act (WSARA) of 2009, implemented by Directive-Type Memorandum (DTM) 09-027 in December 2009. This DTM amended DoD Instruction 5000.02, the *Defense Federal Acquisition Regulation Supplement (DFARS)*, and associated business practices within the *Defense Acquisition Guidebook (DAG)*. The WSARA implementation brought about changes to policies and procedures across 13 categories. Some of the WSARA changes most relevant to the Navy directed-energy community include:

- Analysis of alternatives study guidance
- Acquisition strategies to ensure competition
- Competitive prototyping
- Developmental test and evaluation
- Systems engineering
- Preliminary design reviews
- Critical cost growth

THE ACQUISITION IMPACT

So why should the directed-energy community care about these acquisition policy changes? Because these policy changes impact the community's

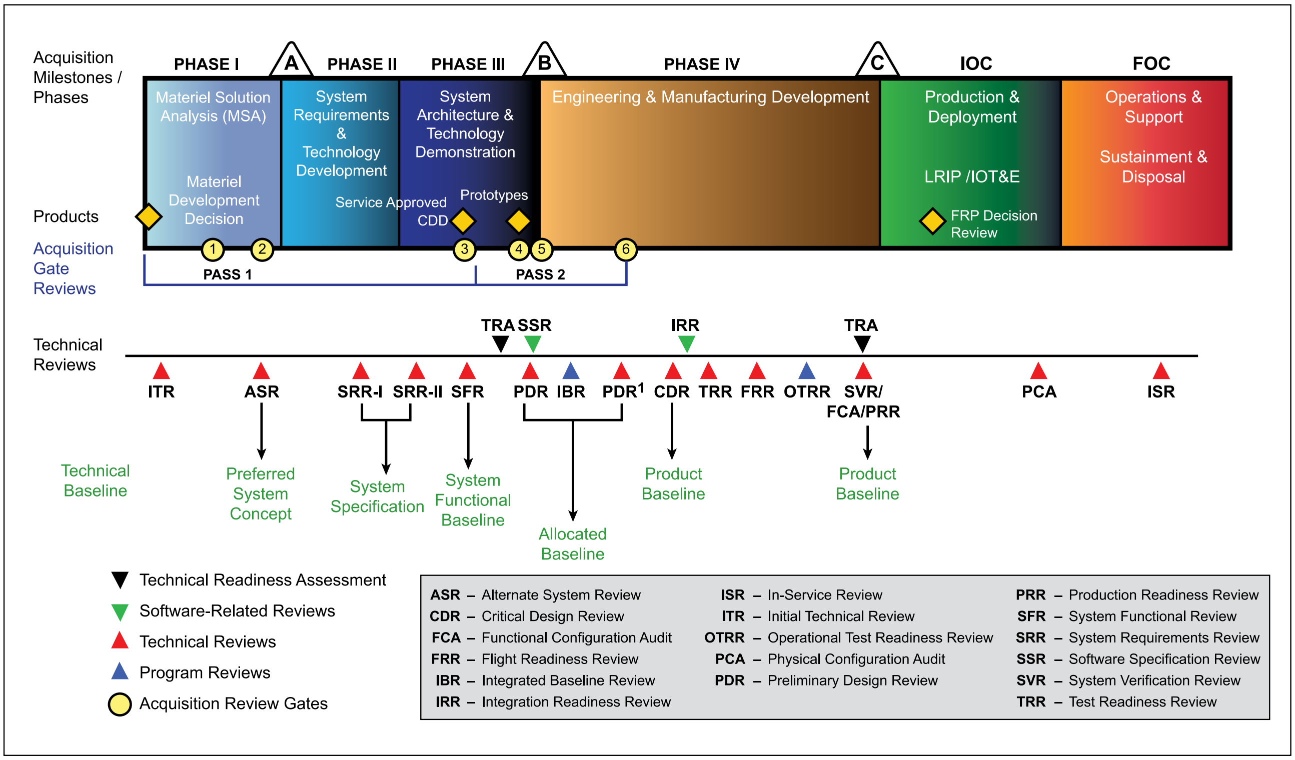
Acquisition Milestones / Phases
PHASE I
A
PHASE II
PHASE III
B
PHASE IV
C
IOC
FOC
Materiel Solution Analysis (MSA)
System Requirements & Technology Development
System Architecture & Technology Demonstration
Engineering & Manufacturing Development
Production & Deployment
Operations & Support
Materiel Development Decision
Service Approved CDD
Prototypes
LRIP /IOT&E
FRP Decision Review
Sustainment & Disposal
Products
Acquisition Gate Reviews
1
2
3
4
5
6
PASS 1
PASS 2
Technical Reviews
TRA
SSR
IRR
TRA
ITR
ASR
SRR-I
SRR-II
SFR
PDR
IBR
PDR1
CDR
TRR
FRR
OTRR
SVR/ FCA/PRR
PCA
ISR
Technical Baseline
Preferred System Concept
System Specification
System Functional Baseline
Allocated Baseline
Product Baseline
Product Baseline
Technical Readiness Assessment
Software-Related Reviews
Technical Reviews
Program Reviews
Acquisition Review Gates
ASR – Alternate System Review
CDR – Critical Design Review
FCA – Functional Configuration Audit
FRR – Flight Readiness Review
IBR – Integrated Baseline Review
IRR – Integration Readiness Review
ISR – In-Service Review
ITR – Initial Technical Review
OTRR – Operational Test Readiness Review
PCA – Physical Configuration Audit
PDR – Preliminary Design Review
PRR – Production Readiness Review
SFR – System Functional Review
SRR – System Requirements Review
SSR – Software Specification Review
SVR – System Verification Review
TRR – Test Readiness Review

Figure 3. Systems Engineering Technical Review Timing

ability to develop, produce, and/or sustain directed-energy weapon systems. The ultimate goal of the directed-energy community is to deploy directed-energy weapons to the fleet. Accordingly, regardless of which phase or phases an organization in the community supports, its actions are impacted by the language in DoD Instruction 5000.02 and the WSARA of 2009. The more scientists and engineers in the organization are aware of governing policy documents like DoD Instruction 5000.02, the better their chances are of meeting DoD leadership's expectations in terms of cost, schedule, and technical effectiveness.

Actions have shown that DoD senior leadership has come to expect all weapon system programs to adhere to the current acquisition-related policy and guidance changes. As mentioned earlier, major weapon system programs have recently been canceled or restructured for not meeting DoD senior leadership expectations—something that rarely occurred previously. In today's environment, technology alone will not carry the argument for a program's survivability. Directed-energy weapons definitely carry the allure of a "Star Wars-like" capability, but these same weapon systems will need to show sustainable cost and schedule compliance if they are to come to fruition. Resources are too limited, and the warfighter has too many needs to allow unsustainable weapon system programs to continue. Therefore, everyone involved with the development, procurement, and/or sustainment of a directed-energy weapon system needs to have an adequate understanding of the acquisition underpinnings now governing DoD.

SUMMARY

The proverbial "winds of change" are blowing across the DoD acquisition landscape. The management of major weapon systems dependent upon cutting-edge technologies—such as those of directed energy—cannot afford to conduct business in a manner reminiscent of bygone days. Everyone involved with the development, production, or sustainment of a directed-energy weapon system needs to understand the "rules of engagement" laid down by the most recent DoD acquisition policy guidance. Highly skilled scientists and engineers typically already understand the need for a structured systems engineering approach to problem solving. Today, though, more than ever, cost and schedule must be factored in as potential tradespace to deliver the ultimate goal: a cost-effective, directed-energy weapon system delivered in a timely manner while meeting the warfighter's requirements. Scientists and engineers who adhere to these recent acquisition changes will help their organizations achieve this goal, thereby ensuring that warfighters will be armed with the most technologically superior weapons possible.

REFERENCES

1. Congressional Research Service Report 7-5700, 10 July 2009, p. 3.
2. Ibid., p. 8.
3. DoD Instruction 5000.02, 8 December 2008, p. 33.
4. *Defense Acquisition Guidebook*, 5 May 2010, paragraph 10.5.2.2.
5. Ibid., paragraph 4.2.3.1.
6. Ibid., paragraph 4.2.3.2.

THE BASICS OF ELECTRIC WEAPONS AND PULSED-POWER TECHNOLOGIES

By Stuart Moran

WHAT ARE ELECTRIC WEAPONS?

Most conventional weapons rely on chemical energy (explosives) as their destruction mechanism, either to explode on target, like bombs, or to create kinetic energy, like a bullet. Electric weapons are different. Electric weapons use stored electrical energy, rather than explosives, to attack or destroy the target. Electric weapons generally fall into two categories: directed-energy weapons (DEWs) and electromagnetic (EM) launchers. DEWs send energy, instead of matter, toward a target, and can be separated into three types: laser weapons, particle-beam weapons, and high-power microwave (HPM) or radio-frequency (RF) weapons. EM launchers use electrical energy to throw a mass at a target, thus making them distinct from directed energy. There are also three types of EM launchers: rail guns, coil guns, and induction drivers. All involve the use of strong magnetic fields to push against projectiles. While electric guns are an electric weapon, they are not a DEW.

High electrical powers and large energies are needed for all these weapons. Technologies for storing and controlling electric power are needed and are commonly called pulsed-power technologies. Electric guns are often associated with DEWs due to their common reliance on pulsed-power technology. The types of electric weapons are shown in Figure 1.

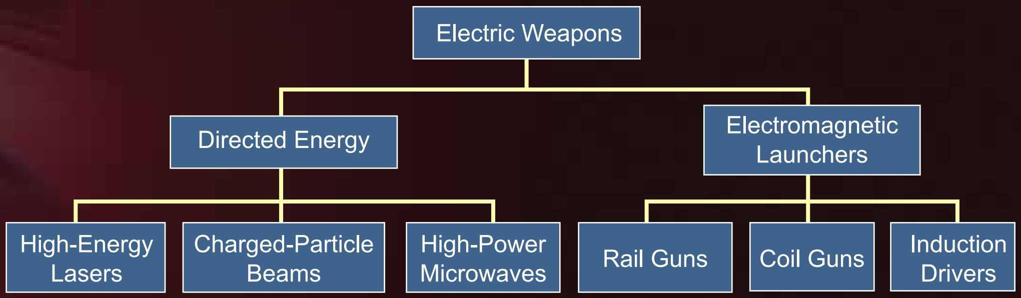
Electric Weapons
Directed Energy
Electromagnetic Launchers
High-Energy Lasers
Charged-Particle Beams
High-Power Microwaves
Rail Guns
Coil Guns
Induction Drivers

Figure 1. Types of Electric Weapons

There are a number of powerful advantages of electric weapons over conventional explosives:

- DEWs have a near-zero time of flight compared to conventional ordnance, allowing longer decision times and quicker reaction times.
- Electric weapons have a large "magazine" capacity, often limited only by the ability of the power source to recharge the system. The firing rate depends on how fast the system can be recharged, which in turn, depends on the available power source.
- The cost of engagement is greatly reduced. With increasingly sophisticated conventional weapons, the cost of practice rounds, such as a missile, can be millions. For an electric weapon, the cost per engagement is greatly reduced, making the attack of small targets (the asymmetric threat) less costly and training much more affordable.
- There is the potential for variable lethality, where the weapon effects can be controlled or attenuated to provide a warning or nonlethal effect. Otherwise, a full-power setting can be used to destroy the target.
- Electric weapons have the benefit of increased safety since less ordnance needs to be stored. Logistics costs less, and underway replenishment is easier since explosives are reduced or eliminated.
- Electric weapons can be used in conjunction with conventional weapons to heighten overall combat system effectiveness, such as knocking out electronics before engaging with a kinetic weapon.

Historically, the key Navy scenario for using directed-energy technologies has been close-in protection of naval vessels from antiship cruise missiles, particularly in a littoral environment. The ability of a DEW's speed-of-light engagement is particularly attractive under conditions of short warning times from supersonic stealthy missiles. However, increasingly difficult and problematic threats from nonmilitary aircraft and surface ships, countersurveillance platforms, fast patrol boats, unmanned aerial vehicles (UAVs), and terrorist inflatable boats or jet skis present different challenges. The threat has shifted from small numbers of expensive targets in open water to large numbers of small and cheap targets among neutral forces. The unique characteristics offered by DEWs, when compared to traditional weapon systems, allow them to be applied across a spectrum of threat roles, particularly in friendly or neutral-rich regions where precision pointing or less-than-lethal capability is paramount. The potential for HPM to counter electronics at levels below human effects makes them ideal nonlethal weapons. Electromagnetically launched projectiles allow longer range, shorter flight times, reduced reliance on air strikes and missiles, and safer storage and replenishment. With military budgets being squeezed, the low cost of directed-energy engagements, which often require just a few gallons of fuel, cannot be overemphasized. Instead of million-dollar missile shots, electric weapons allow new tactics, warning shots, and continual fire against large and small targets. They also allow inexpensive practice and training for improved readiness.

Pulsed Power for Electric Weapons

A useful rule of thumb is that a stick of TNT contains about a megajoule (MJ) of chemical energy, and this amount is often needed to destroy a military target. To destroy a target with an electric weapon, the electrical energy must also be deposited quickly. Surprisingly, a candy bar also has a megajoule of chemical energy, but it is released very slowly when we eat it. Many electric weapons require peak powers of more than a gigawatt (GW) or energies more than a megajoule. The time scales for delivery range from milliseconds to nanoseconds. As an example, delivering 1 MJ of energy in 10 µs requires 100 GW of power, which is more than a commercial power plant can produce. It is not practical to build continuous power supplies to directly drive most electric weapons. Consequently, pulsed-power technologies are needed to store energy at low power rates and release it quickly for weapon use. A pulsed-power system takes electrical power from a prime source (like a motor), stores it, and transforms the power to meet specific user requirements. The importance of a pulsed-power system is often underappreciated. For most electric weapon systems, the system size, weight, volume, and reliability are dominated by the pulsed-power chain. Pulsed-power components must be improved along with the weapon technology to make electric weapons systems practical. A block diagram of a pulsed-power system is shown in Figure 2.

Electrical energy can be stored in many ways, such as a battery (actually a chemical storage). A car battery has about a megajoule of energy, but it takes many seconds to drain it. A much faster method of storing electrical energy is in a capacitor, which can be discharged in milliseconds or faster. Inductive methods store the energy in the magnetic fields of a coil. This has the potential of achieving higher energy density than capacitors, but

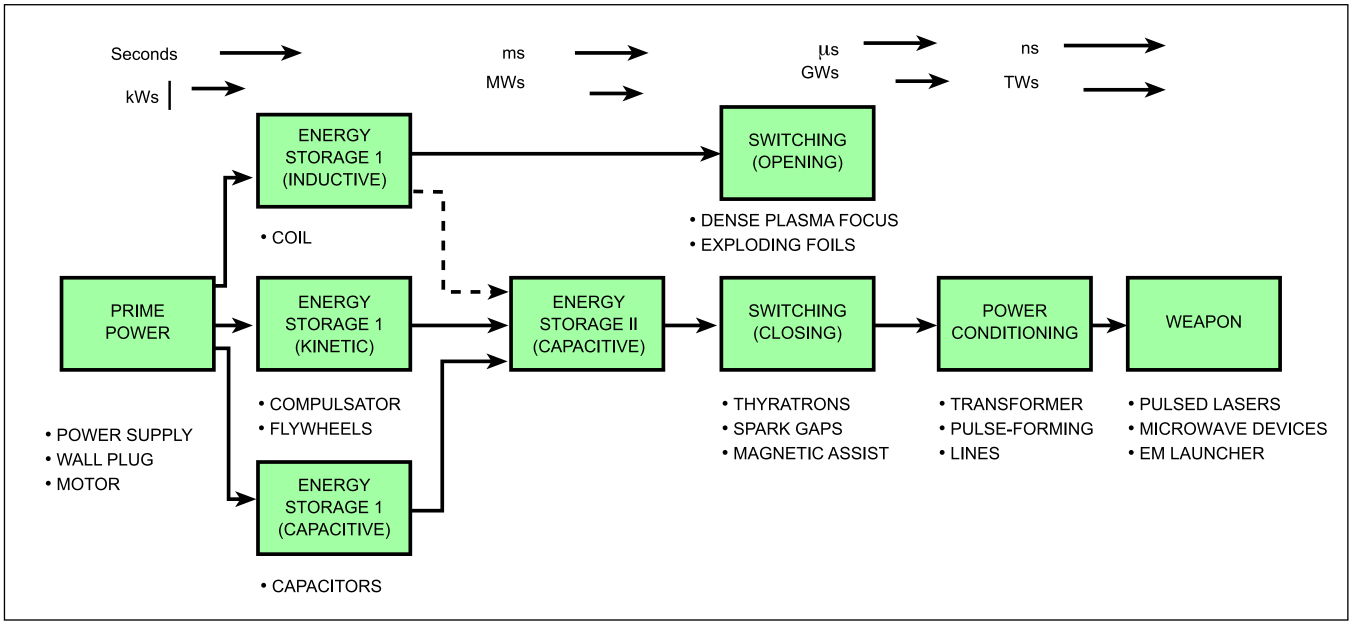
Seconds
kWs
ms
MWs
µs
GWs
ns
TWs
ENERGY STORAGE 1 (INDUCTIVE)
• COIL
SWITCHING (OPENING)
• DENSE PLASMA FOCUS
• EXPLODING FOILS
PRIME POWER
• POWER SUPPLY
• WALL PLUG
• MOTOR
ENERGY STORAGE 1 (KINETIC)
• COMPULSATOR
• FLYWHEELS
ENERGY STORAGE 1 (CAPACITIVE)
• CAPACITORS
ENERGY STORAGE II (CAPACITIVE)
SWITCHING (CLOSING)
• THYRATRONS
• SPARK GAPS
• MAGNETIC ASSIST
POWER CONDITIONING
• TRANSFORMER
• PULSE-FORMING
• LINES
WEAPON
• PULSED LASERS
• MICROWAVE DEVICES
• EM LAUNCHER

Figure 2. Block Diagram of a Pulsed-Power System

when the supporting systems are considered, the technology becomes less attractive. Energy storage for electric weapons can also be done with chemical explosive energy, where an explosive force is converted into electrical energy using techniques such as flux compression. Energy can be stored in the inertia of rotating machines and flywheels, but the energy can be released only as fast as the flywheel or motor can be stopped, usually in seconds. In many cases, several stages of energy store are used where each stage is faster than the last. Once the energy is stored, it must be released quickly using a high-power switch. There are many types of switches. Perhaps the most common type for electric-weapon applications has been the spark gap. Many types of controlled spark gaps exist, including pin-triggered, laser-triggered, field distortion, and simple overvolted. To achieve high repetition rates, flowing oil or gas can be used to flush the hot spark products, or sealed gaps using special fast-recovery gases, such as hydrogen, can be employed. Other switches, such as vacuum tubes and solid-state switches, can be used if they can handle the voltages and currents needed. Solid-state technologies, such as thyristers, have become very capable in recent years. Once the energy is switched out, there is usually some additional power conditioning, where transformers or pulse-forming networks are used to provide the desired pulse shape, voltage, and current required for the weapon. For rapid firing rates or continuous use, high average input powers are needed.

ALL-ELECTRIC SHIP

One of the major impediments to the development of electric weapons systems for Navy ships has been a lack of electrical prime power. Current surface combatant designs employ up to 90 percent of engine power mechanically dedicated solely to propulsion. These designs are unable to provide the tens to hundreds of megawatts (MW) of electrical power capacity required for many electric weapons. The solution is an electric-drive ship that uses all the engine power to generate electricity, enabling it to allocate power to weapons or propulsion as needed. In recent years, the Navy has been investigating cost-effective power-system options to meet future platform requirements.

HIGH-POWER MICROWAVE (HPM) AND RF WEAPONS

Microwave weapons are generally considered to use frequencies above a gigahertz, whereas lower frequencies are generally called RF weapons. These weapons are more powerful than electronic warfare systems and are designed to create extended disruption or permanent damage. An HPM weapon is considered to have a peak power of more than 100 MW, or energies above 1 J. The energy can enter a target through intended RF paths, such as target antennas (front door), or unintended paths, such as housing joints, cavities, and circuit wires (back door). Pulses ranging from a few nanoseconds to microseconds in duration can be sufficient to reset computers, cause loss of stored data, or cause microprocessors to switch operating modes. Nonlinear circuits and components can rectify signals and absorb energy outside of their normal operating parameters. Figure 3 illustrates some of the vulnerability areas on a missile body.

RF or HPM devices can be divided into narrowband or wideband systems, dependent upon the employed pulse length. Narrowband systems are similar to high-power radar pulses and produce RF radiation with a very narrow bandwidth (frequency coverage). The damage concept is to create enough energy in a target to overheat or overload electronic components. Wideband systems generally produce very short pulses (nanoseconds) and typically operate in lower frequency ranges. Wideband systems produce much lower average powers and rely on high-peak electric fields to produce reset or arcing of digital components. Creating short pulses—often only a few RF cycles long—generates a very broad frequency output to take advantage of a target's weak point. But, it also means that the energy is spread over many frequencies, so there may be very little energy at a specific vulnerable frequency. Vulnerability data is critical to estimate the effectiveness of HPM weapons. Ultimately, air breakdown will limit the amount of energy out of an antenna to around 1 MW/cm^2.

HPM devices can produce effects that range from denying the use of electronic-based equipment to disrupting, damaging, or destroying such equipment. HPM weapon advantages include all-weather capability, low precision pointing requirements, and effects persistence after the radiated EM energy "beam" has been turned off. One major advantage of HPM is that electronics are generally more vulnerable to high fields and high energies than humans. This provides the ability to attack electronics without harming people, which makes HPM an ideal choice for nonlethal applications.

Two major challenges of implementing HPM technologies into an operational weapon systems platform are:

1. Fratricide, or self-destruction, can be a problem because of the large areas affected by the sidelobes and near field of any meaningful HPM weapon system. Therefore, when attacking a target of interest with an HPM weapon, there is a greater risk of disruption to systems that were not intended to be targeted but fell within the sphere of influence. Host platforms, therefore, may need to undergo interference hardening.
2. With regard to battle damage assessment, kinetic weapons have the advantage of typically leaving visual evidence. HPM weapon systems do not leave large holes in a target but create more subtle influences as a result of attacking critical electronic components. Consequently, it can be more difficult to ascertain whether a target's capabilities have been sufficiently degraded or destroyed—and for how long—in determining whether a mission was successful.

For HPM system development, a fundamental challenge is the understanding of what it takes to affect the target. Coupling mechanisms, where EM energy enters and affects the target system, are extremely complex. The vulnerability of components is often vastly different if it is outside or inside a circuit board or enclosure. Effects depend upon the interactions with other components, connectors, and nearby conductors. The effects on a component can vary many orders of magnitude depending on frequency, orientation, cracks and seams, protective circuits, pulse energy, and duration. Research regarding effects on missiles has shown large variations not only between designs, but also between different serial numbers due to assembly methods, cable routing, and component variations. With the increasing use of commercial equipment by the military, such as computers and radios, effects are difficult to predict due to constant design and component changes. In general, electronics are getting smaller and operating at lower voltages, making them more sensitive to high fields. But smaller components often have lower pickup areas, and the proliferation of interfering signals has increased the amount of shielding on modern electronics. When

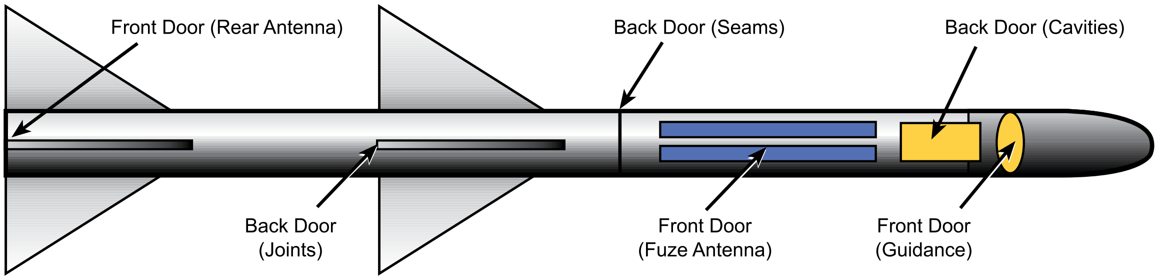

Figure 3. HPM Coupling Paths on Missile Body

target systems are located inside structures or buildings, it becomes even more difficult to predict. Efforts to predict reflections and interference inside complex structures become extremely complicated. Accordingly, generic electronics kill using universal waveforms is not likely. There continues to be a lot of hype about what RF weapons can do, but the idea that a backpack device can wipe out all electronics in a city is no more realistic than a hand-held laser cutting through a bank vault door.

HIGH-ENERGY LASERS (HELS)

A laser generally produces a beam of coherent light at a specific wavelength dependent on the atomic structure of the lasing substance. Only certain substances have the atomic properties appropriate for producing laser light, and these are often limited in power. Lasers are characterized by the substance being lased (gas, liquid, or solid) and the "pumping" process (light energy, electricity, or chemical reaction). A resonant optical cavity provides the means for aligning the energy in the beam and extracting that energy. A military laser system also includes beam processing or beam-path conditioning, beam pointing and control and—for long-range applications—adaptive optics to compensate for the atmosphere.

Until recently, HELs have been driven by chemical energy, so very little electrical power or pulsed power was needed. Chemical lasers use the reactions of gases or liquids to create the excited energy states necessary for laser emission. Large chemical lasers and beam directors have been developed by the Navy in recent decades and have successfully ruptured fuel tanks and downed supersonic missiles. However, these lasers required high-velocity, chemical-reaction chambers and emitted hazardous gaseous by-products. They often operated at wavelengths where the atmosphere absorbed much of the energy. Absorption creates thermal blooming, whereby absorbed energy in the air creates a negative lens that defocuses the beam. Increasing the power of the laser increases the energy absorbed and worsens the problem. The Army and Air Force are developing chemical lasers for airborne applications, where atmospheric absorption is less of a problem. Recent Navy interest in HELs has concentrated on lasers that are electrically powered, rather than chemically powered, and that operate at shorter wavelengths to allow smaller optics and more efficient propagation near the water.

Small semiconductor (or diode) lasers use current flow through an electrical junction to excite electrons and create laser light. These lasers are very limited in power, so research has focused on using large numbers of lasers assembled into a coherent array. Semiconductor lasers also create efficient light to excite or "pump" other types of lasers. Solid-state lasers (SSLs) use crystalline materials mixed (doped) with elements needed for proper lasing. SSLs show strong promise for compact, medium-power HEL weapon systems. Scaling these systems up to megawatt levels creates extreme heat in the crystal material, making it very difficult to prevent internal damage. Forced cooling and the heat capacity of large masses are under study.

Fiber lasers—which use semiconductor diode lasers to pump a flexible, doped crystalline fiber (similar to a fiber-optic line)—have demonstrated high efficiency and relatively high power. The technology is being used in the welding and cutting industries. Methods of pumping large numbers of fiber-optic lasers and combining them are being investigated. An example is shown in Figure 4.

The free-electron laser (FEL) operates differently from a conventional laser. An FEL uses a high-voltage electron accelerator to push electrons through a magnetic "wiggler" to create light radiation across a tunable band of frequencies. The FEL is extremely complex and large, but scaling to very high powers may be possible. Perhaps the biggest promise of the FEL is the ability to design the laser at an ideal atmospheric propagation wavelength. Significant technical hurdles remain in reaching the status of a deployable FEL, in scaling the beam to megawatt powers and in providing the necessary engineering to turn a laboratory device into a weapon system of reasonable size. For Navy application, FELs will require improvements in areas of radiation shielding, high vacuum, high-current photo-injectors, and probably cryogenic cooling—all of which must be integrated into a ship's basic design.

Fiber lasers and SSLs are the leading-candidate Navy lasers for medium power, as FELs are for high power. All are electrically driven and can meet the requirement for shorter wavelength, capable of transmitting at the "maritime window" of approximately 1 μ.

HEL weapons' advantages include a highly directional and narrowly focused beam, providing:

- Minimal collateral damage
- Speed-of-light delivery
- Rapid retargeting
- Low cost of engagement

Disadvantages center on:

- Limited range due to atmospheric attenuation
- Weather limitations
- Low efficiency (often less than 10 percent)
- Need for eye protection
- Relatively large size and weight requirements

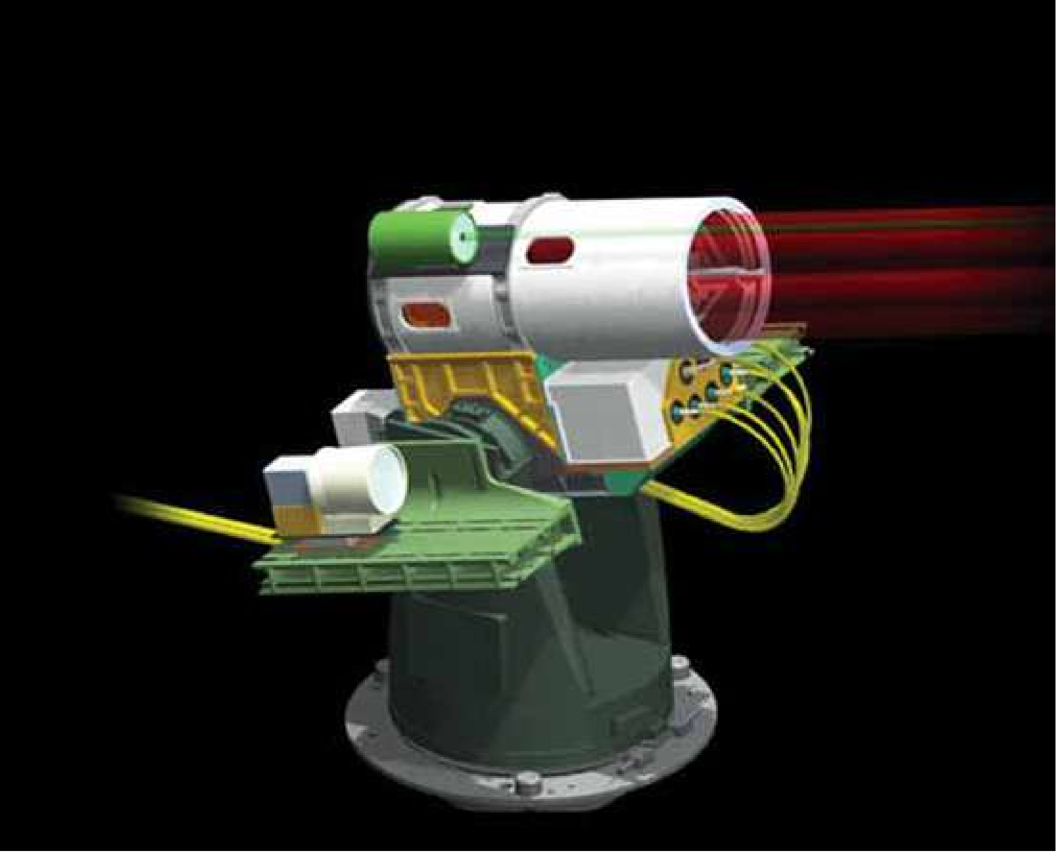

Figure 4. Drawing of Laser Weapon System (LaWS)

Long dwell times (seconds) will be needed for most targets. As with RF systems, there is a potential nonlethal or variable lethality capability since the energy can be easily defocused. A critical challenge is the understanding of a laser beam's propagation through a maritime boundary layer environment, where the sea and air interface creates turbulence and moisture gradients. Measuring the atmosphere and compensating for variations in real time may require adaptive optics or "rubber mirrors" that can be constantly adjusted to compensate for changes. Focusing a small spot at long range will require high beam quality and large optics, probably meter-size mirrors that are very highly reflective and very clean.

HELs in the future are expected to be able to focus energy to a spot size of much less than a meter at ranges of kilometers. This will necessitate very accurate target tracking systems, and precise stabilization and beam-pointing systems, both of which are difficult but should be feasible in the near term. Real-time atmospheric measuring systems will be needed for compensation techniques. Methods to protect the sensitive optical system from salt spray and corrosion will also be needed.

From a lethality perspective, three considerations need to be better understood before a HEL can be deemed a true weapon system:

1. Achievable spot size of beam on target at range
2. Amount of coupling into the target material
3. Subsequent effects of the damage inflicted

For the more severe threats, such as high-speed, antiship cruise missiles, HELs face the difficult task of engaging maneuverable, stealthy, inbound missiles. As such, a better quantitative understanding of the interactions among a laser beam's energy deposition, target material, and flight dynamics is needed.

PARTICLE BEAMS

A particle-beam weapon is a directed flow of atomic or subatomic particles. These particles can be neutral or electrically charged. Neutral beams need to be used outside the atmosphere (in space), where charged particles would repel and fly apart. Charged-particle beams (CPBs) are easier to make and are used within the atmosphere, where air molecules can constrain the beam. A CPB weapon transmits matter—not just EM waves—like lasers and microwave weapons. The particles are near the speed of light and deposit their kinetic energy deeply into any target material. They have the potential to be highly destructive weapons and are very difficult to shield against.

Charged particles are produced by applying a strong electric field near a material that emits electrons. These electrons then pass through accelerating stages with high voltage gradients (often megavolts), which increase the electron's velocity. As the electrons pass each stage, the velocity increases until they approach the speed of light (become relativistic), at which point they have substantial energy to penetrate a target. The accelerating systems can be linear, but a recirculating design is more compact and can reuse stages. These systems are basically high-current versions of scientific particle accelerators.

Once the electron beam is produced, it must propagate to the target. High-velocity electrons will not go far before they collide with air molecules and lose energy. The fact that air molecules struck by the beam are heated and moved out of the way for a short period of time creates a rarified "hole" in the atmosphere through which a second pulse can travel farther. In this manner, a fast series of pulses can "hole-bore" to the target, each pulse going farther than the last. The final pulse must have enough energy to damage the target. The deceleration of electrons in the atmosphere causes

Bremsstrahlung radiation in the forward direction toward the target, creating gamma rays that, in turn, create X-rays and RF radiation.[a] These effects can cause electronic upset and "soft-kill" mechanisms even if the beam slightly misses the target.

The beam of electrons is typically a few centimeters in diameter. When a beam strikes a target, the energy is deposited deep in the material (the collision cross section is small because of the relativistic speeds) in microseconds (much faster than a laser), creating thermal shock that is very difficult to shield against. For an explosive target, there is also the possibility of causing a deflagration or low-order burn, disrupting the normal warhead mechanism.

Scientists studying CPB weapons made significant technical advancements in the 1980s, but the weapons are still far from being practical. A CPB weapon is technically very challenging and expensive to build. Studies project that the volume requirements necessary for a CPB system could be on the order of a 5-inch gun system. Advantages of a CPB weapon include rapid penetration, a deep magazine, all-weather capability, and soft-kill mechanisms for a near miss. Problems include complexity, size, limited range, and the need to demonstrate compact accelerators and propagation mechanisms.

ELECTROMAGNETIC (EM) LAUNCHERS

A number of technology concepts to launch projectiles exist using electrical energy. These systems rely on large currents in conductors, creating strong magnetic fields that drive a projectile. The velocity of a normal powder gun projectile is limited by the expansion speed of the explosive powder, and present military guns are reaching that limit. With an electric gun, the fields can push projectiles much faster, providing longer ranges and increased kinetic energies. The simplest version is an EM rail gun, shown in Figure 5.

In any conducting loop, the generated magnetic field tries to expand the loop. If everything is held in position, the only movable item is the conducting projectile, which moves down the rails in an attempt to expand the loop. Since megajouoles of projectile energy are needed for EM rail guns, energy storage

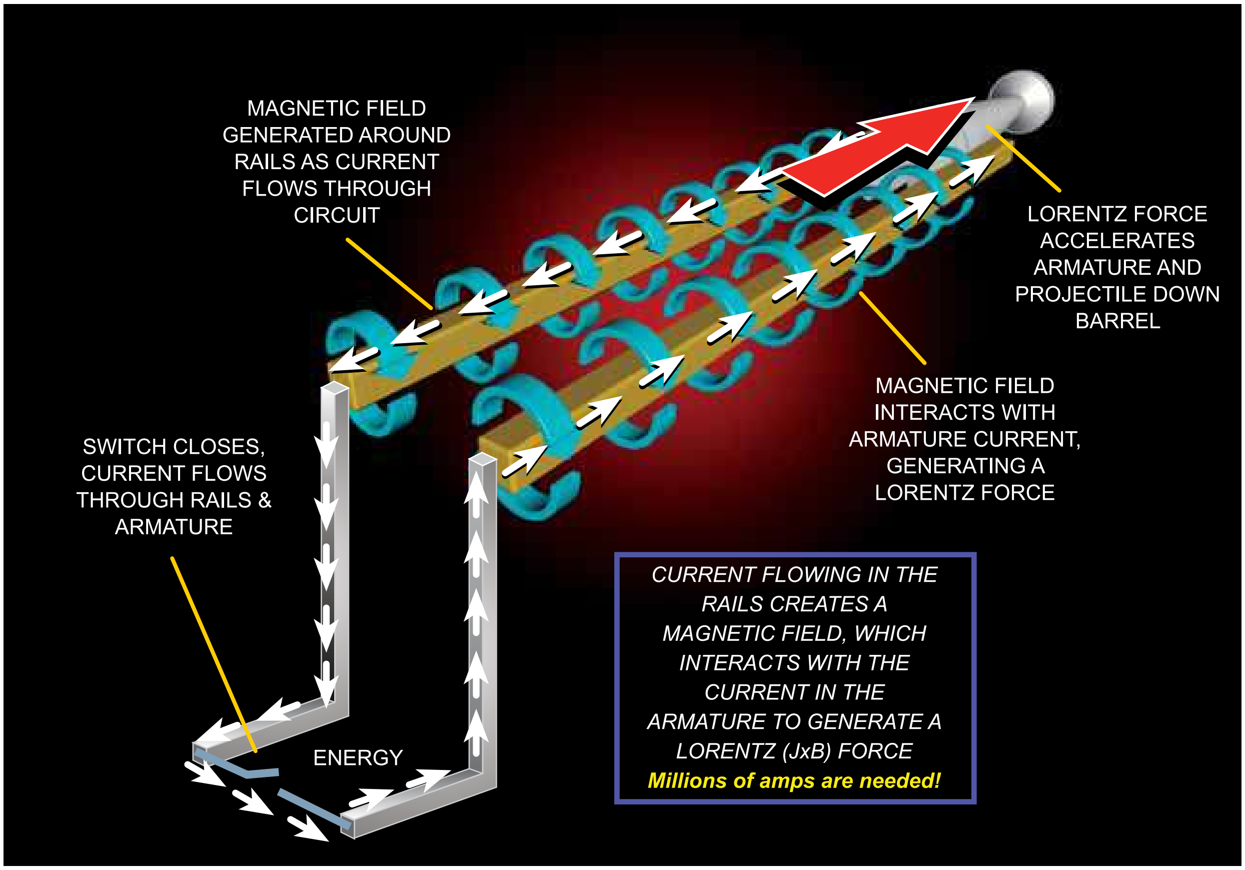
MAGNETIC FIELD GENERATED AROUND RAILS AS CURRENT FLOWS THROUGH CIRCUIT
LORENTZ FORCE ACCELERATES ARMATURE AND PROJECTILE DOWN BARREL
MAGNETIC FIELD INTERACTS WITH ARMATURE CURRENT, GENERATING A LORENTZ FORCE
SWITCH CLOSES, CURRENT FLOWS THROUGH RAILS & ARMATURE
ENERGY
CURRENT FLOWING IN THE RAILS CREATES A MAGNETIC FIELD, WHICH INTERACTS WITH THE CURRENT IN THE ARMATURE TO GENERATE A LORENTZ (JxB) FORCE
Millions of amps are needed!

Figure 5. Electromagnetic (EM) Rail Gun Concept

mechanisms that can store about 100 MJ are needed, along with the ability to discharge the energy in milliseconds. To generate useful forces, millions of amps of current are needed—a major challenge and significant loss mechanism. Large capacitor banks with very high-current switches are required. Spark gap switches have historically been the only option, but new high-current solid-state switches are now becoming available. Capacitor energy densities, too, have improved an order of magnitude in the last few decades. Rotating machines have also been considered because they are smaller than equivalent capacitor banks, but extracting the energy quickly, without tearing the machine apart, has been problematic. The launch energy of various projectiles is shown in Figure 6.

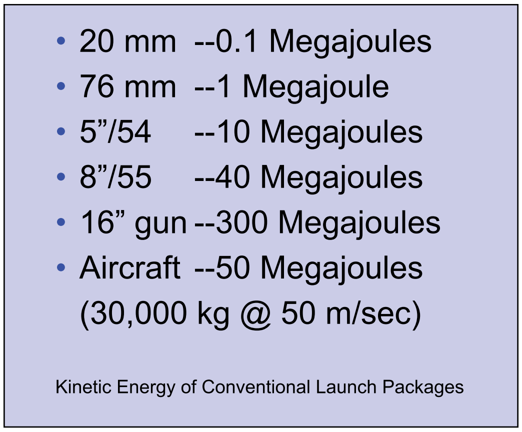

- 20 mm --0.1 Megajoules
- 76 mm --1 Megajoule
- 5”/54 --10 Megajoules
- 8”/55 --40 Megajoules
- 16” gun --300 Megajoules
- Aircraft --50 Megajoules (30,000 kg @ 50 m/sec)

Kinetic Energy of Conventional Launch Packages

Figure 6. Launch Energy of Various Projectiles

A rail gun is probably the most compact form of electric launcher. However, it requires direct electrical contact between the projectile and barrel rails, creating the potential for arcing, melting, and erosion. Coil guns use a series of sequentially fired coils around a “barrel” to push the projectile in stages. This does not require direct electrical contact, so it avoids rail erosion but requires a series of fast timed switches and more space. Linear induction motors are basically unrolled electric motors and have been used on electric trains and roller coasters, typically with magnetic levitating systems to avoid contact erosion. This concept is being developed by the Navy for launching aircraft. The energy to launch an aircraft is similar to a large-caliber projectile—more weight but less speed. The slower speeds are more suitable for rotating machines since the launch times are seconds rather than microseconds.[1] Electrothermal guns and electrothermal-chemical (ETC) guns use a combination of electricity and chemicals. Electrical energy is used to initiate chemical reactions that can produce lightweight driving gases, like steam, or allow more energetic propellants that are difficult to ignite in a conventional fashion.

Some advantages of electrically driven projectiles include:

- Higher projectile velocity (over conventional explosives)
- Very long range (>100 miles) with lower cost than missiles
- Time-critical delivery (because of shorter time of flight)
- Safer projectile stowage (minimal explosives)
- Potentially adjustable velocity levels, for better accuracy and controllable damage

The potential of having nonexplosive rounds and magazines is very attractive for the Navy. For long-range, large-caliber EM projectiles, the kinetic energy from the projectile velocity is greater than the chemical explosive energy in a conventional round traveling much slower. Therefore, damage can be equivalent even without explosives. System size and lifetime are still behind conventional systems, but getting close.

Outlook

Challenges remain for many electric weapon concepts. These weapon systems appear promising to meet the increasingly important asymmetric threats with low-cost precision rounds. They also can be employed across the energy spectrum for nonlethal targeting. Electric weapon systems will, in many cases, continue to supplement existing kinetic weapon systems in the near term. Despite technology challenges, directed-energy and electric weapons hold great promise in offering the future warfighter unique combat capabilities not currently available.

Endnote

a. Bremsstrahlung—a type of radiation emitted when high-energy electrons are decelerated. (German for ***braking radiation***)

Reference

1. Shope, S. et al., *Long-Range Naval Fire Support with a Coilgun*, Sandia Report 2001-3832, February 2002.

SOLID MODELING OF DIRECTED-ENERGY SYSTEMS

By Joseph F. Sharrow

Not long ago, around the mid-1980s, development of most new mechanical systems—such as automobiles, consumer products, and military devices—was performed manually on a drafting table or drawing board, much like the present-day version shown in Figure 1. These tables and boards performed a necessary function, but they offered little assistance other than for drawing lines. Engineers used them to prepare layouts, or two-dimensional sketches of what they were designing. They then would take these layouts to a draftsman, who would create drawings of each part in the device. The drawings would subsequently be sent to a manufacturing facility.

This layout and drawing preparation process typically would need to be repeated multiple times because mistakes would be made, or design issues would be discovered late in the process. Similarly, the manufacturing process would sometimes require multiple iterations as well because of the inherent limitations in designing three-dimensional (3-D) devices on two-dimensional boards. This less-than-ideal process made it difficult to design and manufacture even mundane products and frequently resulted in things that just didn't work. With the emergence of early computerization, numerical analyses of more complex systems began to be performed. These analyses were conducted to ensure that the systems worked in the real world. For example, engineers might conduct a structural analysis of the forces in a loaded dump-truck bed to make sure that the frame wouldn't bend and fail. Because of the difficulty in performing these analyses, they would often require a specially trained group of structural engineers, expensive software, and large mainframe computers, limiting their use to only the largest, most well-funded companies or organizations.

EMERGENCE OF COMPUTER-AIDED DESIGN (CAD) AND SOLID MODELING

With the availability of smaller scale computers and more economical software in the mid-to-late 1980s, CAD was born, initiating a period of rapid improvement in the design process. This was driven, in part, by the introduction of software packages such as AutoCAD. Initially, these software packages only attempted to automate drawing lines by making wireframe (stick-figure) versions on the computer of what previously had been made by hand on the drafting board. This reduced the difficulty in making changes in the development process, but it still limited the engineer's pallet to a two-dimensional space. What was really needed was a 3-D method of design. Solid modeling

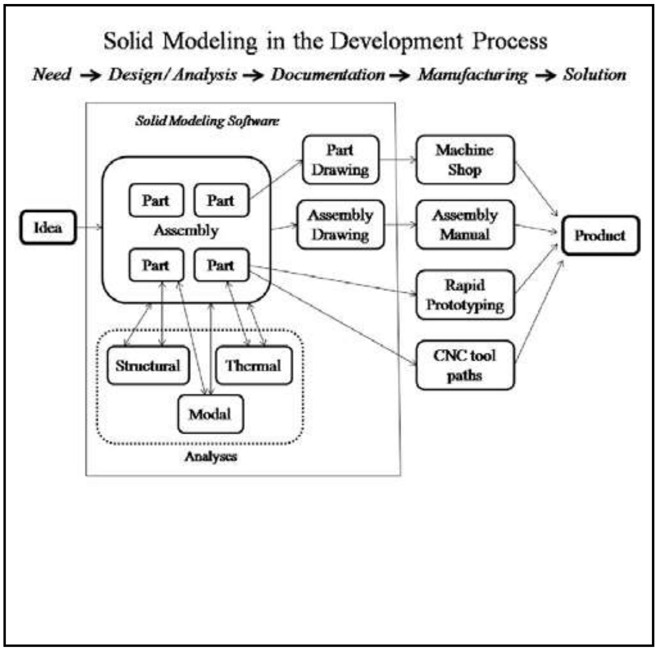
Solid Modeling in the Development Process
Need → Design/Analysis → Documentation → Manufacturing → Solution
Solid Modeling Software
Idea
Part
Part
Assembly
Part
Part
Part Drawing
Assembly Drawing
Machine Shop
Assembly Manual
Rapid Prototyping
CNC tool paths
Product
Structural
Thermal
Modal
Analyses

Figure 1. Drawing Board

addressed this need beginning in the late 1980s to early 1990s.

Solid modeling is analogous to taking blocks of clay and cutting and forming them into the shape of a solid part on a computer. These 3-D parts are then put together in an assembly, more accurately representing real-world devices. Though originally used only in a limited way for specialized applications in the aircraft and automobile industries, it wasn't until the 1990s that solid modeling experienced widespread availability and mainstream acceptance due to software packages such as Pro/ENGINEER. Figure 2 summarizes how Pro/ENGINEER and other similar packages fit into the development of new products. The general flow of the process moves from left to right.

Initially, nearly all 3-D solid modeling packages required significant computing and graphics display power, necessitating the use of large graphics workstations running the UNIX operating system. Rapid advances in computing and graphics power have since enabled nearly all packages to run efficiently on personal computers (PCs) and laptops, bringing solid modeling and analysis capability into the mainstream.

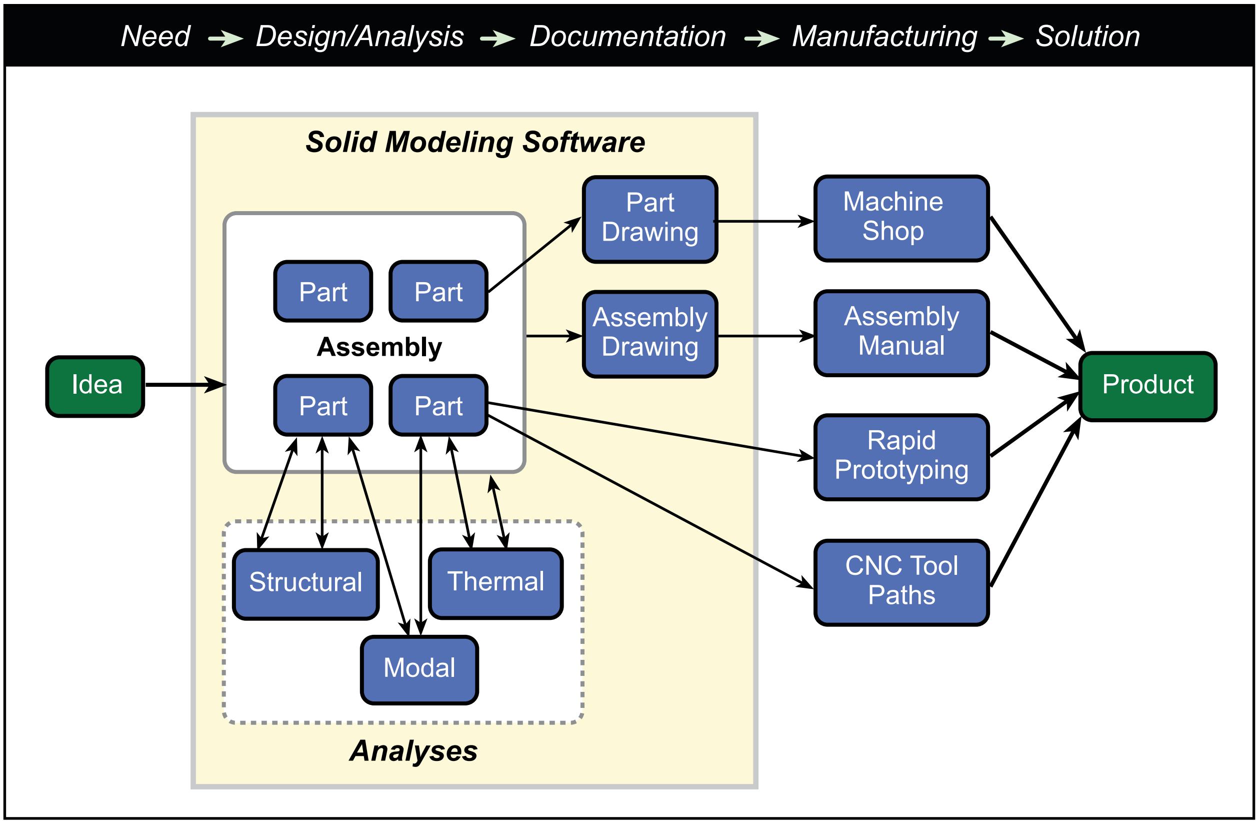
Need → Design/Analysis → Documentation → Manufacturing → Solution
Solid Modeling Software
Idea
Part
Part
Assembly
Part
Part
Part Drawing
Assembly Drawing
Machine Shop
Assembly Manual
Rapid Prototyping
CNC Tool Paths
Product
Structural
Thermal
Modal
Analyses

Figure 2. Solid Modeling in the Development Process

Solid Modeling of Directed-Energy Systems

Engineers working in the Directed Energy Division at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD), use solid modeling to develop hardware for nearly all of its programs. Both Pro/ENGINEER and SolidWorks are used extensively to develop new products in virtual 3-D space. Additionally, the structural simulation package within Pro/ENGINEER is used to determine stresses and natural frequencies of parts and assemblies. Consequently, these packages have enabled a single mechanical engineer in the Directed Energy Division and a draftsman in the Engagement Systems Department at Dahlgren to perform the design and analysis work that would have required an entire group of engineers and draftsmen just a few years ago. Today, collaboration among many organizations using similar packages has become commonplace. Insofar as solid modeling has become an indispensable tool for development and collaboration, its successful implementation requires proper training and experience before engineers can use it effectively, just as medical surgeons require training in the use of advanced robotic surgical devices before they can be used effectively. Thus, while these high-tech modeling systems not only have reduced the number of personnel needed for design and development, they have enabled the Navy to get significantly more bang for its buck while supporting warfighting needs. An example of how solid modeling is currently being used is discussed below.

Navy Laser Weapon System (LaWS) Beam Director

The Directed Energy Warfare Office (DEWO) and Directed Energy Division at Dahlgren are currently developing the Navy LaWS for the Naval Sea Systems Command's Directed Energy and Electric Weapon Systems (DE&EWS) Program Office (PMS 405). The program's goal is to take advantage of currently available industrial laser technology and incorporate it into a future naval weapon system. As part of the development process, major subsystems have been integrated with a Kineto Tracking Mount (KTM) into a LaWS beam director. The KTM/beam director was modeled and analyzed using Pro/ENGINEER. Ultimately, the resulting LaWS will be installed on Navy ships on the Close-In Weapon System (CIWS) gun mount. During field testing in June 2009 at the Naval Air Warfare Center, China Lake, California, the prototype KTM/beam director successfully destroyed five unmanned aerial vehicles (UAVs). The actual beam director used in the China Lake testing is shown in Figure 3; the Pro/ENGINEER assembly model used for development is shown in Figure 4.

The LaWS effort took advantage of many aspects of solid modeling including collaboration, structural and modal analysis, and manufacturing drawing creation. The project required development of new, unique hardware, as well as the integration of electronic models from commercial vendors. The KTM model was provided by L-3 Brashear and was originally designed using Pro/ENGINEER. The beam-directing telescope model was provided by RC Optical Systems, Incorporated, and was originally made in SolidWorks. These models were combined with many new optical and structural components developed by the Directed Energy Division into a single, comprehensive assembly model. This model was instrumental in understanding the interaction of the many components, and its use increased accuracy and precision that would have been impossible with old-fashioned two-dimensional development processes. Figure 5 shows a cross section through the main portion of the beam director, revealing the complexity of the many parts and subassemblies required for such a device. In addition to modeling the mechanical components, the actual laser beams were also included to better understand their path through the various mirrors and optical devices in the beam director, and to better highlight any interference they might have with structural components within the KTM or telescope.

Numerous analyses were performed to make sure that everything worked the way it was intended. One major analysis addressed the telescope mount. To ensure that the beams were stable at range, the mount had to be extremely stiff. The best way to ensure this was to perform a structural analysis using the structural simulation package within Pro/ENGINEER. Figure 6 shows the results of that analysis: a displacement plot in which different colors represent how much the telescope will move when the KTM rotates at its maximum speed. The large cylindrical object simulates the mass of the telescope. The minimum amount of displacement is indicated by blue, and the maximum is shown in red. This analysis verified that the movement of the telescope, relative to the optical components within the optics breadboard, was acceptable and should perform well at the range specified by the program office.

After modeling and analysis were completed, manufacturing drawings of custom parts were created by the Engagement Systems Department to be

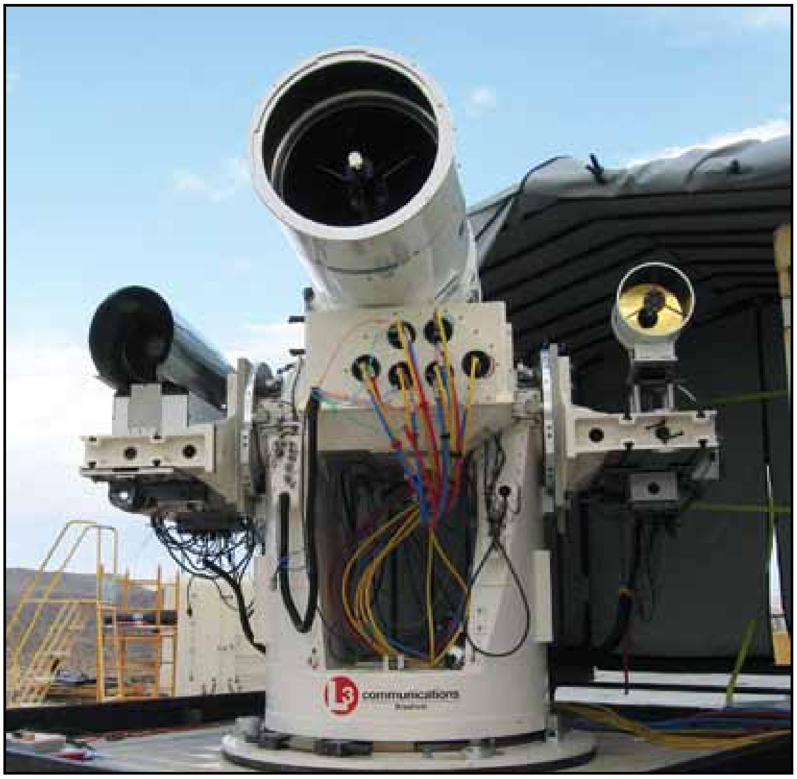
L3 communications

Figure 3. LaWS Beam Director

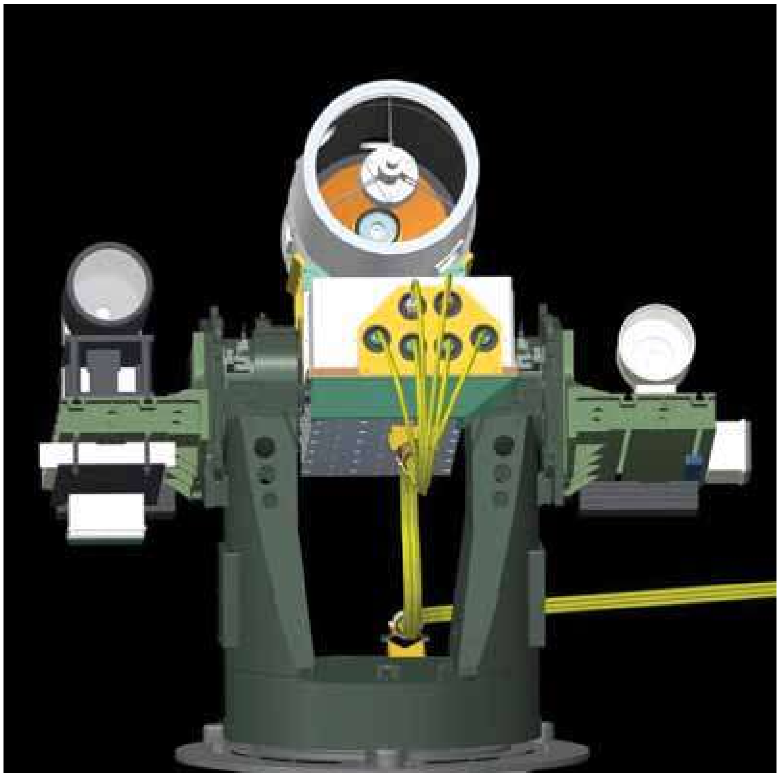

Figure 4. LaWS Beam Director Assembly Model

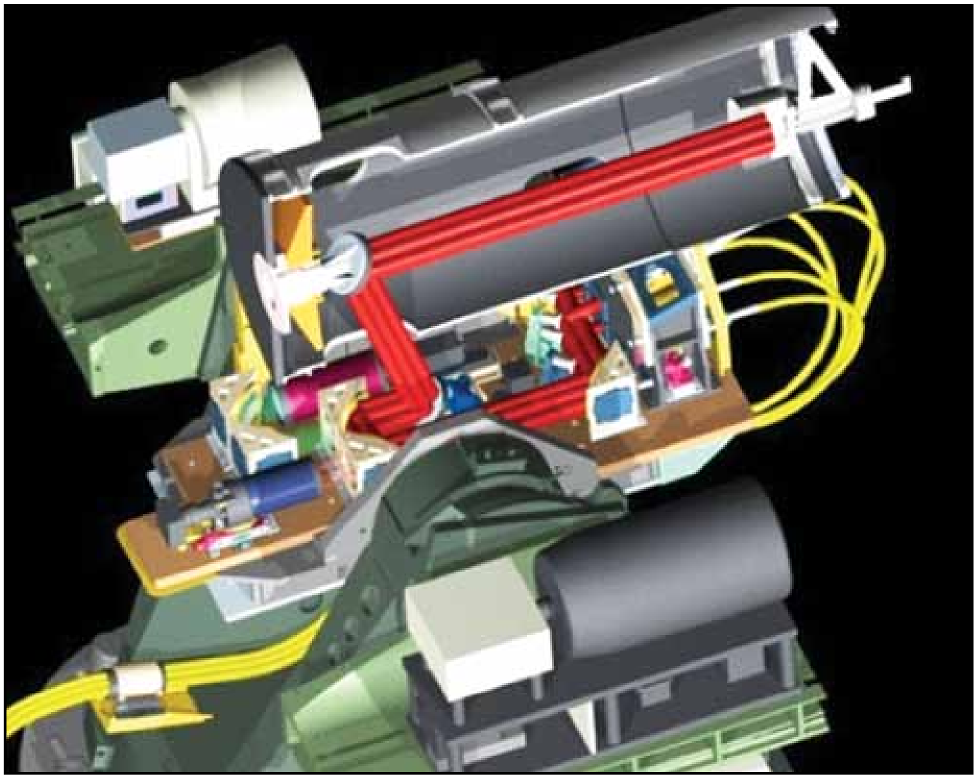

Figure 5. LaWS Beam Director Cutaway

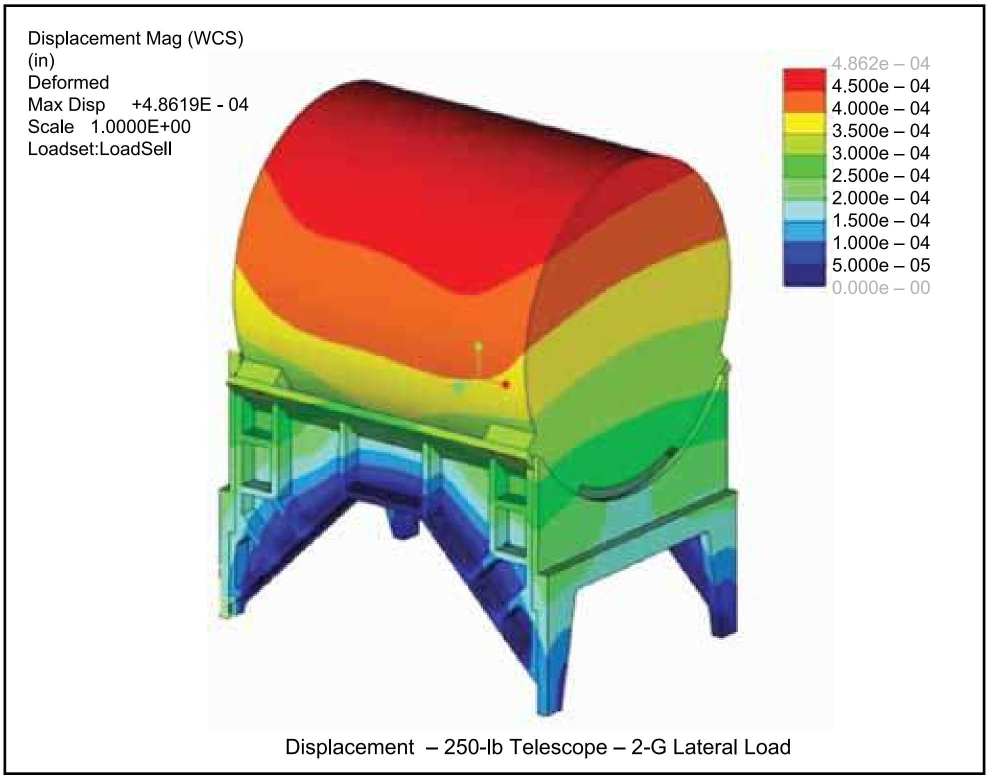

Displacement – 250-lb Telescope – 2-G Lateral Load

Figure 6. Displacement Plot

sent to manufacturing facilities, such as machine shops. One example is shown in Figure 7, which shows the first sheet of the multisheet drawing needed to manufacture the large plates that support the telescope from the center platform of the KTM. One of these plates is also shown in the displacement plot in Figure 6.

Even though it would be possible for one person to do all of the modeling, analyses, and drawings for a particular program, a more efficient process takes advantage of using the best skills available by collaborating with other experts. Collaboration enables assembly, part, and drawing files to be sent electronically, eliminating the need for collocating personnel. Drawings for the LaWS program, for instance, were made using noncollocated personnel across base at NSWCDD. They could just as easily have been made using personnel from across the country.

The LaWS program exemplifies how the Directed Energy Division uses solid modeling to enhance the quality and effectiveness of Navy directed-energy capabilities. As a result, warfighters will be better armed with more effective weapons and capabilities for future naval conflicts.

Acknowledgment

The author would like to acknowledge the valuable contributions of the Engagement Systems Department's drafting group in support of numerous DEWO and Directed Energy Division programs.

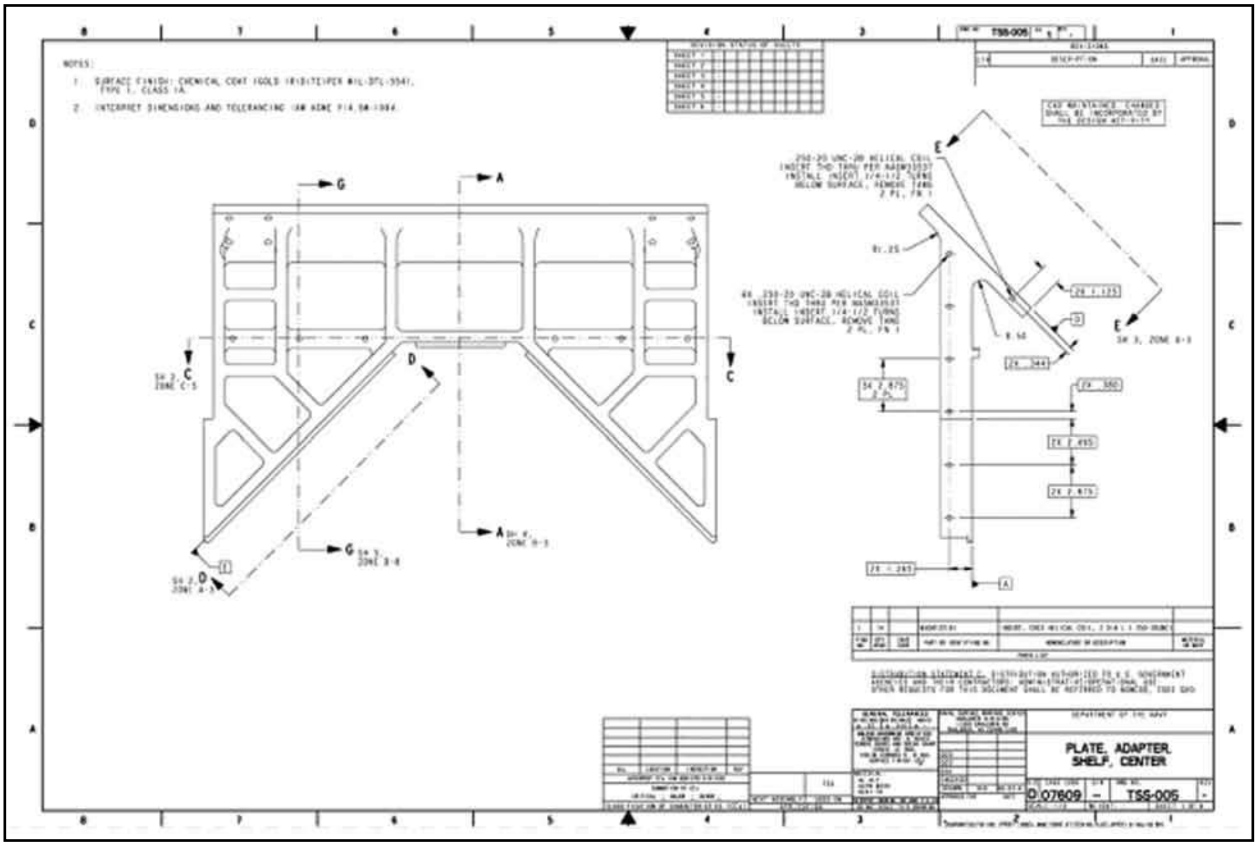

Figure 7. Manufacturing Drawing

A FUNDAMENTAL KEY TO NEXT-GENERATION DIRECTED-ENERGY SYSTEMS

By Directed Energy Division, Electromagnetic and Sensor Systems Department

Imagine an explosive ordnance disposal (EOD) unit on a routine scouting patrol deep in the notorious "Triangle of Death" south of Baghdad, where Marines, Sailors, and Soldiers frequently find themselves exposed to improvised explosive devices (IEDs). Fortunately, this newly outfitted unit is equipped with the latest unmanned, mobile, remote-controlled, radio frequency (RF) transmitter used as a directed-energy weapon (DEW). The integrated system provides comprehensive IED prediction, detection, prevention, and neutralization capabilities. Lightweight, pocket-sized transmitters carried by each warfighter constantly communicate sensor intelligence, key vital signs, critical conditions, and location telemetry to a geostationary satellite (GEOSAT). It intercepts, collects, and retransmits intelligence and situational awareness data simultaneously to any command post in the world and to each member of the unit on patrol. Highly efficient, miniature, switch-mode, RF amplifiers with high-power density (small size and weight with high-power output) enable these visions of future capabilities as their systems' transmitter backbone.

To civilians, the miniaturization of modern wireless (electromagnetic) devices is considered a mere convenience or luxury, i.e., Blackberries, mobile phones, and high-speed wireless network connections. To the next-generation warfighter, miniaturized, wireless, directed-energy (DE) systems open the door to the realization of a whole new set of effective and efficient wireless modalities. And while the capabilities mentioned in the above scenario are not yet available to warfighters, researchers believe they have uncovered the key to next-generation DE systems leading to the miniaturization of DE devices.

NEXT-GENERATION DE SYSTEM REQUIREMENTS

At the Naval Surface Warfare Center, Dahlgren Division (NSWCDD), key system requirements for effective next-generation DE systems are being researched and developed for applications to counter IEDs, to detect explosively formed penetrators (EFPs), to neutralize explosives, and to predict threat locations. Next-generation DE systems must yield a high probability of mission success and be inherently safe to operate. By design, they must minimize or eliminate the risk of hostile attack or collateral damage especially during screening missions. Considering the DEW example above, practical

080102-N-1132M-006 SHEIK SA'ID, Iraq (2 January 2008) U.S. Army Soldiers attached to 3rd Squadron, 2nd Cavalry Regiment patrol and search for weapons or Improvised Explosive Devices (IEDs) during a clearing mission. (U.S. Navy photo by Mass Communication Specialist 1st Class Sean Mulligan/Released)

next-generation DE systems must be physically characterized by:

- Low mass (weight)
- Small size (volume)
- High-power output with respect to size or high-power density
- High efficiency for extended mission use
- Minimized prime power and cooling support
- Portability
- Mobility
- Configurability

They must also ensure a high probability of mission effectiveness. The DEW must be easily transportable and agile, adapting to the immediate, local military mission requirements in various warfighting environments. Additionally, DE systems must be mechanically robust and able to withstand the shock and vibration of combat missions in rough and rugged environments. The key requirement—efficiency—fundamentally facilitates all required characteristics, including mass and size.

MOVING BEYOND REQUIREMENTS

Scientists at NSWCDD, sponsored by the Office of Naval Research (ONR), are researching and developing key system requirements for effective next-generation DE systems to counter IEDs, to detect EFPs, to neutralize explosives, and to predict threat locations.

Researchers at NSWCDD are leading the way toward realizing small, lightweight, RF transmitters using high-power, solid-state, switch-mode amplifiers, theoretically 100 percent efficient. These practical switch-mode amplifier realizations are at least 1/100 the volume and weight of any commercially available linear solid-state amplifier of comparable power output. The challenges included assessing what type of active amplifier device and operation would provide the greatest power density (power output per unit volume and mass) with its necessary auxiliary systems, such as prime power generation and cooling of waste heat. Such a device also needed to provide sufficient output power based on required standoff range and IED system-coupling efficiency while also maintaining a manageably-sized, easily transportable system. Researchers initially considered tube-based systems, but large, heavy, direct-current (DC) power supplies are required, and typically 40 percent of the input power is dissipated in heat, which negates any possibility of miniaturization.

Upon a practical review of amplifier-class operations and suitable active amplifier devices, however, research pointed to contemporary switch-mode amplifier schemes (e.g., Class-E and Class-F) using solid-state technology—such as the high-electron mobility transistor (HEMT)—as satisfying the high-power density and abusive mechanical requirements for expected worst-case transportation and operation in a rugged environment. To significantly impact reduction of size and weight, practical, high-efficiency thresholds were defined for next-generation DE systems at 90 percent and greater. The key technology enabler to realize amplifier high efficiency in high-power amplifiers up to 60 kW was found in exploiting contemporary switch-mode amplifier architecture with efficient power combining. Particularly, switch-mode schemes in Class-E and Class-F operation as solid-state, active-hybrid planar topology designs were found to be necessary and sufficient for DE applications. These analyses led to a novel, Class-E RF switch-mode amplifier design. A Class-E RF switch-mode amplifier can theoretically operate at 100-percent efficiency. For every input watt supplied, an RF output watt is produced. The conductors and dielectric substrate of the hybrid planar load network and the commercial off-the-shelf (COTS) transistor all exhibit some small degree of power loss, suggesting an estimated practically realized efficiency of 90 percent.

Moreover, the amplifier under research consisted of a novel microwave load network operating with high-power output at ultrahigh frequency (UHF). This research led to the state of the art in Class-E designs leading by hundreds of watts, several hundred megahertz in frequency, and roughly 10 percentage points in efficiency. A common, solid-state, high-power amplifier design technique sums the phase and amplitude of smaller amplifier units to the large values required for DE systems. A practical hardware limitation exists that limits the theoretically infinite number of fixed RF output power units to a finite number. Approximately 60-kW RF output power sets the boundary as the largest hardware realization. By applying spatial power combining in the propagating medium, phased-array antennas can be employed with constructive wave interference in air that would allow sufficient RF power densities on target, based on the number of elements in the array. This technique eliminates the traditional hardware necessary to power combine the smaller power-amplifier elements, realizing a much simplified DE system with enhanced power density in the transmitter, and reduced mass and volume.

The key to ultrahigh efficiency in a switch-mode amplifier, such as Class-E or Class-F, is found in zero-voltage switching (ZVS). Here, the load network is not only designed to be resonant at and around a particular desired switching frequency, it must simultaneously act to force the voltage across the switch to be zero when current flows and when it switches off; hence, theory suggests that no power is dissipated because the product of current through, and voltage across, the switch is zero. It is this aspect of the design that makes the job of switch-mode amplifier realization difficult. Of course, in practice, a small voltage exists for a very short time during the switching action, resulting in a small amount of input power being dissipated in heat. This theoretical description also assumes that all components are ideal (i.e., no impedance to current flow exists in the switch when turned on). All realistic switches exhibit finite impedance when turned on, which does dissipate some wasted energy, but again, this is very small in modern HEMT devices using the ZVS technique.

Class-E switch-mode amplifier theory development began in the United States during the 1960s, with details published in 1975, although some earlier reports were published in Russia. Lumped element electrical components (RF choke inductors and metal film capacitors) were initially used in lower frequency (3 to 30 MHz) prototypes. As engineers attempted higher frequency designs in the very high frequency (VHF) range, solid-state transistor switch parasitic intrinsic and packaging elements found inside the transistor began to be used as some of the key components necessary for ZVS. These parasitic elements included stray capacitance caused by differences of potential between parts inside the transistor and inductance caused by bond wire length that is used to connect the transistor to accessible terminals in its packaging. At microwave frequencies, these parasitic elements become sensitive, invoking unintended significant changes to load networks designed to operate with the transistors. Intrinsic elements include drain-to-source breakdown voltage capability and peak current capability. As the need for higher frequency operation and higher power increased, constraints of key transistor parameters became difficult to produce in traditional silicon technology:

- High instantaneous transient (peak) current capability through the transistor
- Moderate breakdown potential across the transistor
- Low output capacitance

Only within the past few years have transistor manufacturers produced COTS transistors that meet the required capabilities necessary to operate in switch mode for microwave frequencies and

081107-N-1120L-072 RAMADI, Iraq (7 November 2008) Joint EOD Rapid Response Vehicles (JERRVs) assigned to Naval Mobile Construction Battalion (NMCB) 7's convoy security element are secured following an escort mission from a forward operating base. The Cougar-type JERRVs are employed by coalition forces for escort and logistics missions, and to protect personnel from IEDs. NMCB 7 is deployed to U.S. Forces Central Command to provide contingency construction support to coalition forces in support of Operations Enduring Freedom and Iraqi Freedom. (U.S. Navy photo by Mass Communication Specialist 2nd Class Michael B. Lavender/Released)

high-power output. Selection is still somewhat limited for designers.

New transistor technology known as gallium nitride (GaN) HEMTs—using state-of-the-art manufacturing processes with GaN on silicon carbide materials—now facilitates Class-E high-power amplifier (100-W) designs at ultrahigh frequencies. The design process for switch-mode amplifiers is radically different than linear amplifiers, so engineers have tended to continue using linear amplifier design techniques due to familiarity, rather than advance to the switch-mode designs. Today, the Class-E and Class-F unit power output (greater than 100 W) capability and upper frequency limitation is based on a lack of available HEMTs with the necessary parameter capabilities.

Most recently, transistor manufacturers have limited their investment in the Class-E amplifier solid-state switch market due to no commercial market mandate. An assortment of presently available HEMTs provides a low-power capability in terms of 1- to 10-W output power for Class-E amplifiers in the cell phone market. The need remains to continue advancing in commercially manufactured HEMTs with key capabilities necessary to realize larger unit power output, hundreds of watts to a thousand watts, for practical implementation in DE systems.

POSSIBLE MULTIPLE APPLICATIONS

Directed-Energy Weapon Systems

Expanding on the vision of the next-generation DEW system mentioned at the beginning of this article, further imagine that EOD scouts detect a laser fluorescence signature of C4 high explosive and chlorine outgasses in the vicinity of an abandoned vehicle 2-km north of their current position. An electronic support measure (ESM) team on board an approaching clearing vehicle initiates RF jamming and electromagnetic surveillance procedures. Electronic specialists also scan the area with ground-surface differential thermography—particularly to detect possible buried IEDs and EFPs or their tiny command wires, crush wires, or pressure plates—while clearing a pathway to the abandoned roadside vehicle.

Upon arrival at a 500-m safe distance, the EOD specialists command the RF transmitter's robotic platform, also equipped with sensitive gamma-ray planar and computed tomography (CT) imaging to navigate toward and around the vehicle, interrogating every possible hiding place. It discloses an IED in the fuel tank. The specialist lifts the transmitter arming safety and commands the remote transmitter to radiate a prescribed dose of RF energy directed at a carefully chosen component of the vehicle-borne IED (VBIED) system. Without entering the vehicle, the advanced screening system detects and defuses the deadly IED buried within the rusty, metal vehicle chassis. Within minutes, the suspected VBIED threat is entirely neutralized, with absolutely no wounded warfighters or casualties.

Pictured here is the National Aeronautics and Space Administration/National Oceanic and Atmospheric Administration (NASA/NOAA) Geostationary Operational Environmental Satellite-P (GOES-P) launching from Cape Canaveral Air Force Station, Florida, aboard a Delta IV rocket procured by Boeing Launch Services on 4 March 2010. Built by Boeing Space and Intelligence Systems, GOES-P will provide NOAA and NASA scientists with data to support weather, solar, and space operations, and will enable future science improvements in weather prediction and remote sensing. Additionally, GOES-P will provide data on global climate changes and capability for search and rescue.

Mobile Ad-Hoc Wireless Network (MANET)

Beyond IED detection and neutralization, imagine an expeditionary unit on patrol, with each member equipped with an RF transceiver about the size and weight of a cigarette pack with an ultrahigh-efficient switch-mode amplifier. The miniature transceiver constantly communicates sensor intelligence, key vital signs, critical conditions, and location telemetry to a GEOSAT. This small switch-mode amplifier has the needed output power to reach an altitude of 35786 km, where the GEOSAT intercepts, collects, and retransmits this intelligence and situational awareness data to any command post in the world and to each member of the unit on patrol simultaneously. The expeditionary unit, spread out over a wide area with large interspacing, shares the situational awareness and intelligence data of each other at the speed of light. Thus, near real-time, worldwide communications with ubiquitous secure access from the battlefield is possible in a multiple-input, multiple-output (MIMO) architecture. The same system could provide a soldier-to-soldier MANET.

Next-generation switch-mode RF amplifier designs could also optimize payload weight and volume on board new communication satellites while supplying higher power density and making efficient use of the solar power supply budget. Improved switch-mode amplifier power output, when combined with enhanced antenna design, would minimize Earth-station antenna size requirements. The recently launched satellite shown at left demonstrates an example of the latest antenna technology.

Looking Forward

Miniaturizing next-generation DE systems opens up a whole new world of applications to support warfighters in ways unimaginable just a few years ago. Reduction of transmitter mass and volume, accompanied with high efficiency, creates a welcome trickle-down effect. Low profile, small, lightweight DE systems means:

- Less vulnerability to attack
- Greater mobility and maneuverability
- Simplified logistics with less fuel-supply demands
- Less impact on the environment

Clandestine operations, too, could be executed with greater ease and simplified logistics support. In the case of MIMO MANETs, miniaturized high-power density transmitters could further expand capabilities for the warfighter, enabling them to carry high-power transmitters to communicate with satellites or other supporting platforms. The

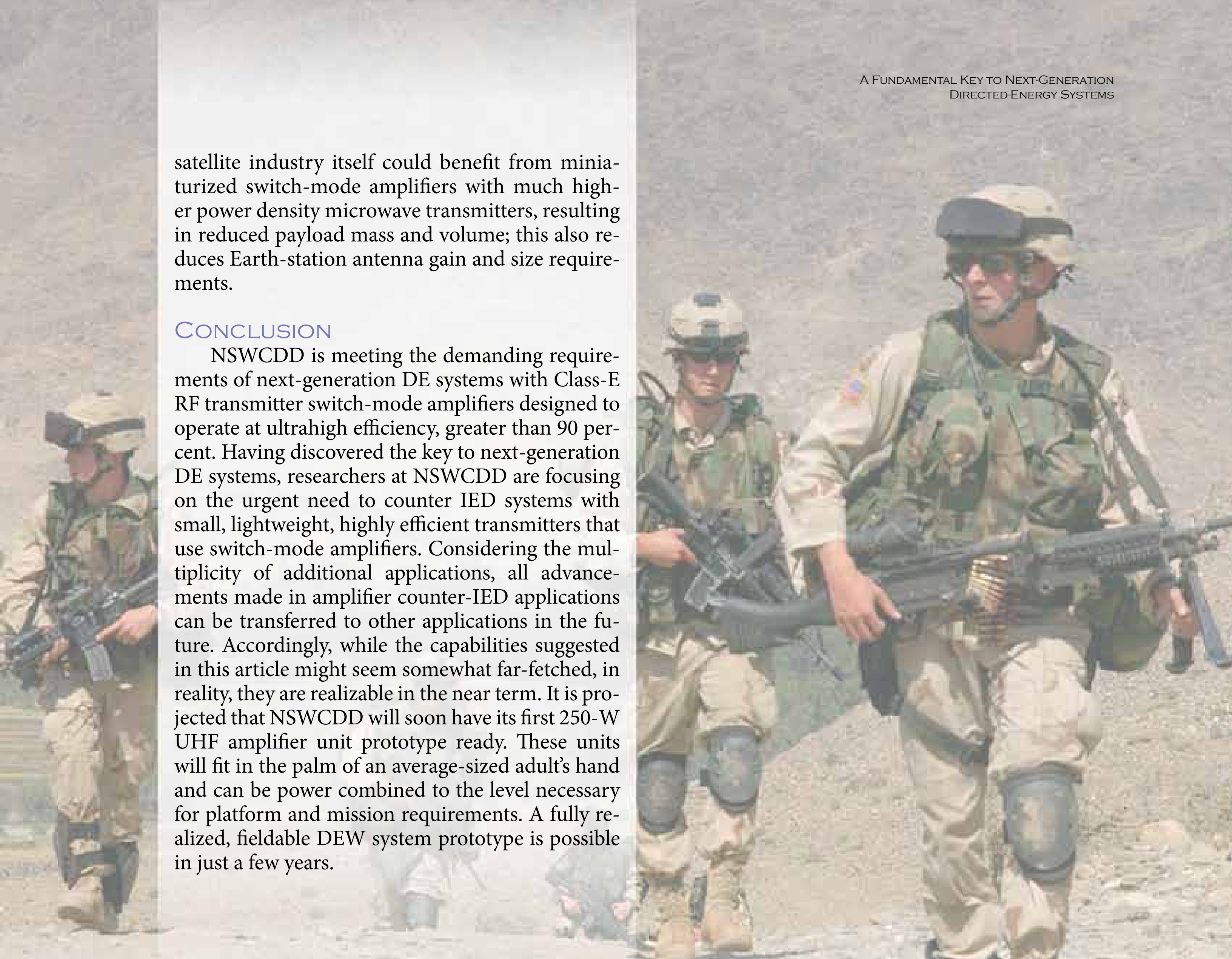

satellite industry itself could benefit from miniaturized switch-mode amplifiers with much higher power density microwave transmitters, resulting in reduced payload mass and volume; this also reduces Earth-station antenna gain and size requirements.

CONCLUSION

NSWCDD is meeting the demanding requirements of next-generation DE systems with Class-E RF transmitter switch-mode amplifiers designed to operate at ultrahigh efficiency, greater than 90 percent. Having discovered the key to next-generation DE systems, researchers at NSWCDD are focusing on the urgent need to counter IED systems with small, lightweight, highly efficient transmitters that use switch-mode amplifiers. Considering the multiplicity of additional applications, all advancements made in amplifier counter-IED applications can be transferred to other applications in the future. Accordingly, while the capabilities suggested in this article might seem somewhat far-fetched, in reality, they are realizable in the near term. It is projected that NSWCDD will soon have its first 250-W UHF amplifier unit prototype ready. These units will fit in the palm of an average-sized adult's hand and can be power combined to the level necessary for platform and mission requirements. A fully realized, fieldable DEW system prototype is possible in just a few years.

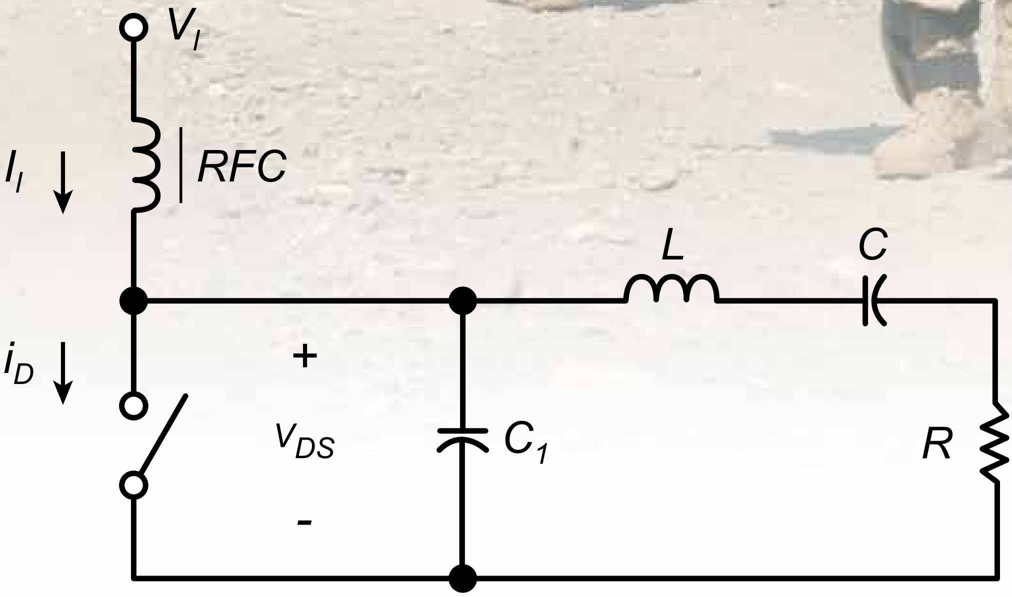

ACTIVE DENIAL ARRAY

By Randy Woods and Matthew Ketner

Active Denial Technology (ADT)—which encompasses the use of millimeter waves as a directed-energy, nonlethal, counterpersonnel weapon—has the potential to provide an important new escalation-of-force capability to U.S. operating forces. ADT projects a focused beam of 95-GHz millimeter waves to induce an intolerable heating sensation on an adversary's skin, repelling the individual with minimal risk of injury. More than a decade of research has established the biological and behavioral effects of ADT for large spot size systems, such as Active Denial System 1 (Figure 1). While the effects of this large spot size system have been successfully established, the technology that produces those effects has the potential to progress in a number of ways, particularly with the development of smaller, lighter, and lower-cost systems.

One research effort focuses on the development of smaller, lighter, and lower cost ADT demonstrators that produce commensurate "ADS-effects," with effective spot size and power densities on target. In support of this effort, the Joint Non-Lethal Weapons Program (JNLWP) sponsored the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) to develop a "smart target system," which measures the millimeter-wave beam using fast-response, 95-GHz diode detectors. NSWCDD subsequently developed and tested the W-Band Beam Diagnostic Array to characterize the system's beam with a temporal resolution of 30 Hz and a high spatial resolution of 1 inch.

The current method of measuring the 95-GHz beam is to use carbon-loaded Teflon (CLT) to produce an average power beam image. This method works as the CLT is exposed to the system's beam. The material heats, over a period of seconds, proportional to the magnitude of the radio frequency (RF) field, resulting in an image as shown in Figure 2. After the exposure, the specific heat capacity of the CLT can be used with the temperature increase in the CLT to provide an indication of the total energy deposited in the material. This method produces a good representation of the average RF field; however, any peak variations in the beam are averaged out.

To allow for high temporal-resolution measurements of the 95-GHz beam, a high-density, 95-GHz diode-detector array was commissioned by the Joint Non-Lethal Weapons Directorate (JNLWD), and was designed and built by NSWCDD, with support from Millitech, Inc. The array consists of a center 11 × 11 matrix (shown in Figure 3) with four removable arms that can be attached (shown in Figure 4), resulting in a measurement area of approximately 1 × 1 m.

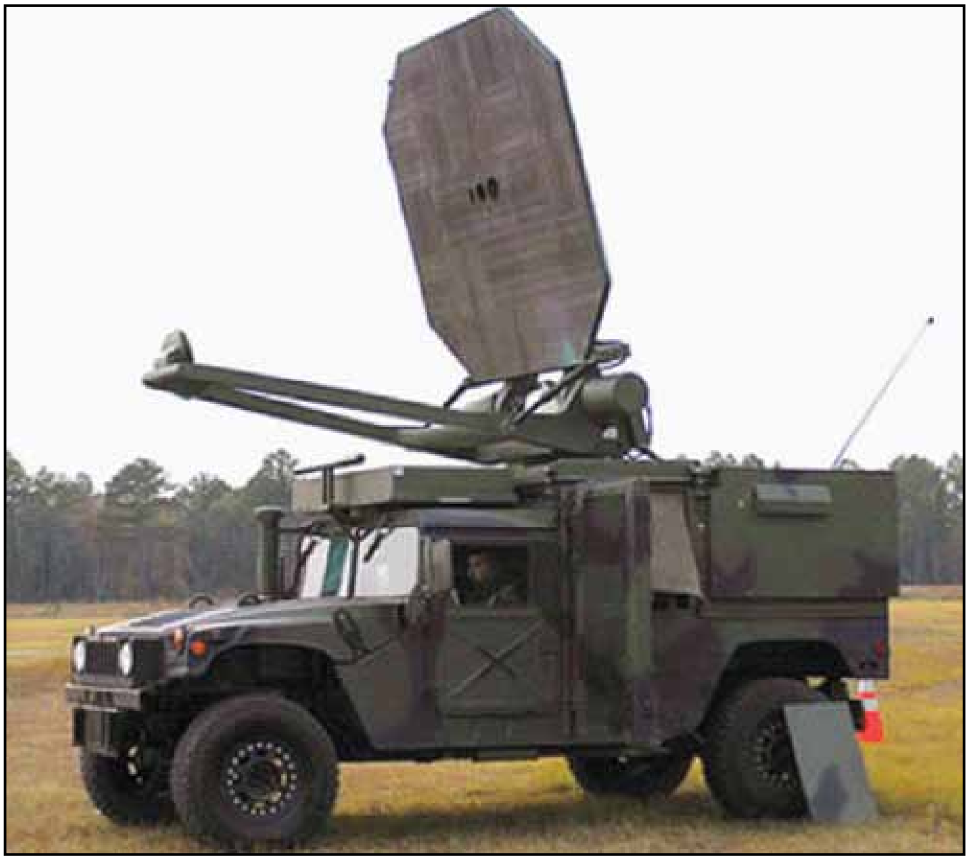

Figure 1. Active Denial System 1

Each element's profile consists of the individual horn antenna from the array, an attenuator, a detector, and a SubMiniature version A (SMA) connection to the digitizer circuitry. This configuration allows for the power received from the antenna to be attenuated and converted to a direct current (DC) output capable of being measured by an analog-to-digital converter. The machined antenna elements provide a uniform effective area for each element, allowing field strength (W/cm^2) to be converted into power received (W or dBm). The aperture antennas also provide an impedance match between free space and the waveguide system. A cross-sectional view of the array element is shown in Figure 5, followed by a signal flow diagram shown in Figure 6.

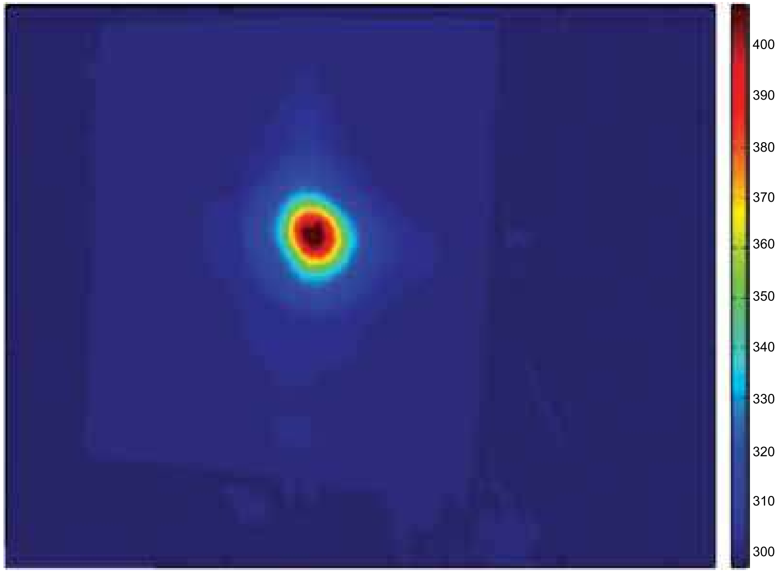

Figure 2. CLT Representation of Small, 95-GHz Spot

The basic principle of operation behind the array is that the derivative of the diode detector's power vs. output voltage curve is very repeatable between detector elements. Therefore, when the detector elements arrived at NSWCDD, each detector element was paired with a variable

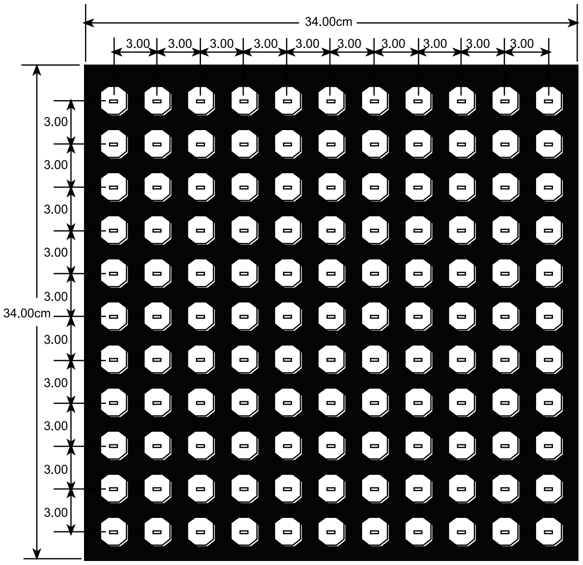
34.00cm
3.00
34.00cm

Figure 3. Main Array Face

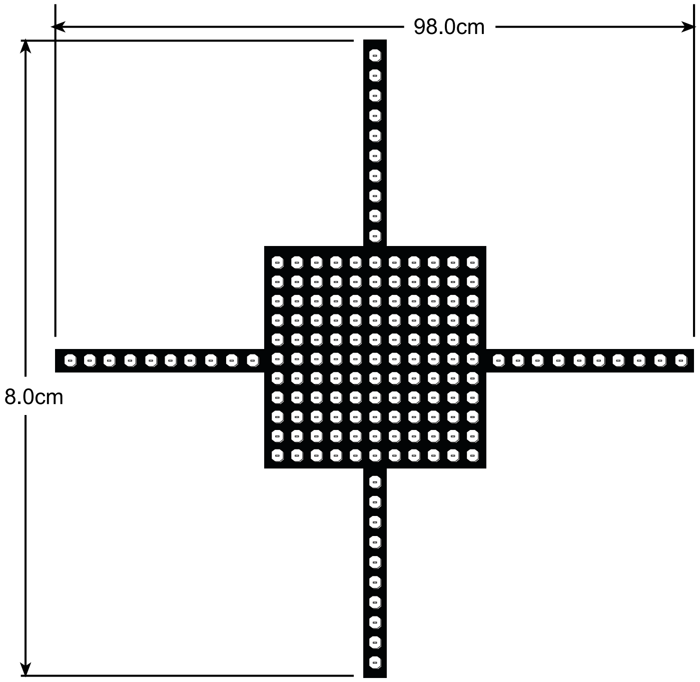

Figure 4. Full W-Band Array

attenuator and calibrated as a single unit. The calibration was accomplished by inserting a known input power of +5 dBm into the input of the attenuator and setting the DC output voltage at a predetermined millivolt (mV) output. This allowed the detector's individual offset voltages to be removed and caused the detectors to behave in a repeatable manner. The attenuator is able to be adjusted by varying the depth that the aluminum nickel card is inserted into the section of waveguide.

The final section of the electrical system converts the DC voltage output from the detectors to a digital signal to send back to the operator station. For this, it was determined that a 16-bit digitizer would be required to enable measuring the microvolts output by the detectors on the low end of their range, while still allowing the digitizer to measure the full output voltage of 1.8 V for high-input powers. Also, due to the proximity of the operator to the array and overall system flexibility, it was determined that Ethernet communications would provide a sufficient means of reading the system data.

To display the data to the operator, a two-dimensional array is populated and displayed for the user (shown in Figure 7). This allows values to be

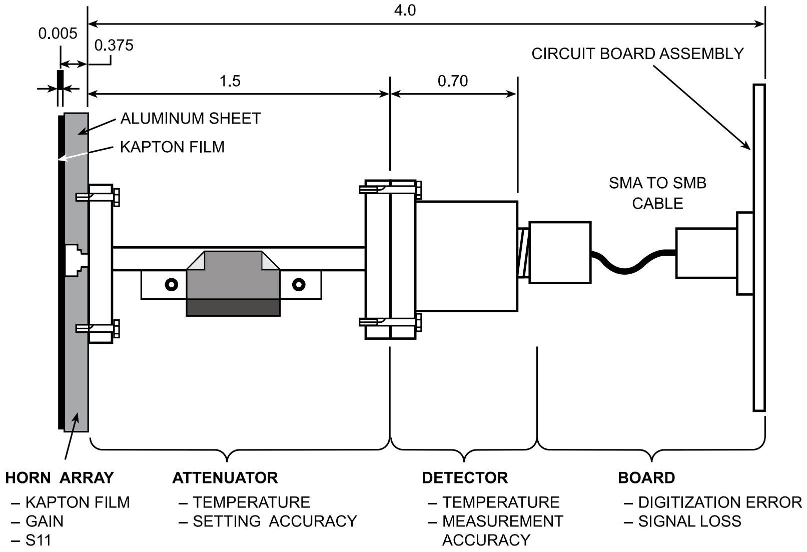
4.0
0.005
0.375
1.5
0.70
CIRCUIT BOARD ASSEMBLY
ALUMINUM SHEET
KAPTON FILM
SMA TO SMB
CABLE
HORN ARRAY
– KAPTON FILM
– GAIN
– S11
ATTENUATOR
– TEMPERATURE
– SETTING ACCURACY
DETECTOR
– TEMPERATURE
– MEASUREMENT ACCURACY
BOARD
– DIGITIZATION ERROR
– SIGNAL LOSS

Figure 5. Cross-Sectional View of Array Element

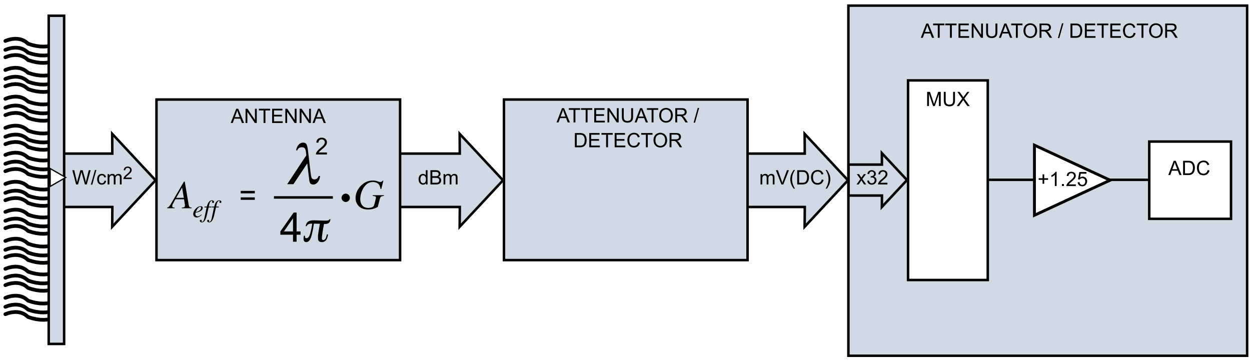
W/cm2
ANTENNA
$A_{eff} = \frac{\lambda^2}{4\pi} \cdot G$
dBm
ATTENUATOR / DETECTOR
mV(DC)
ATTENUATOR / DETECTOR
x32
MUX
+1.25
ADC

Figure 6. Signal Flow Diagram

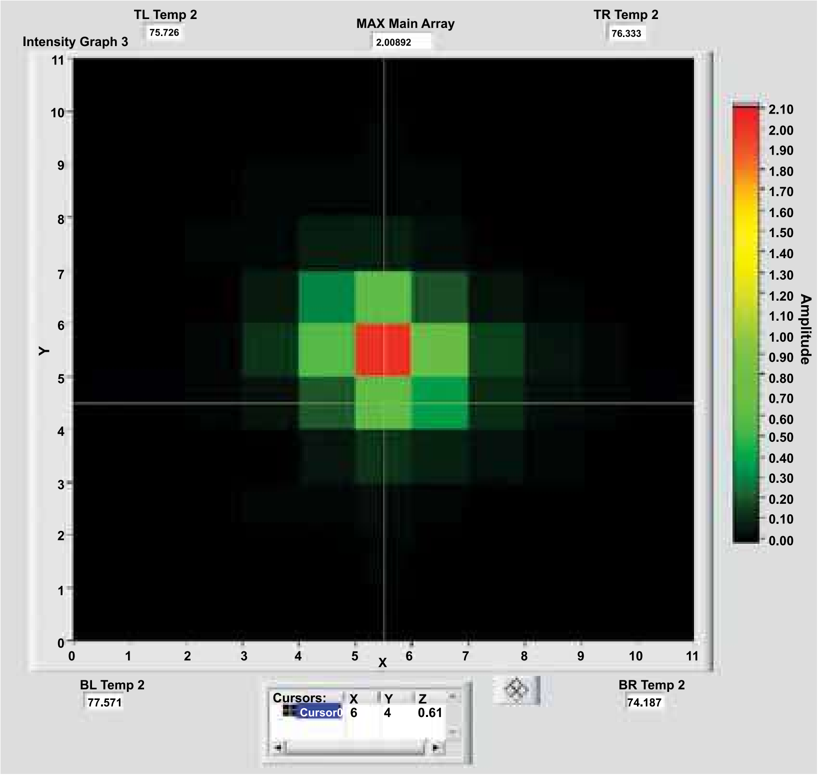
TL Temp 2 75.726

MAX Main Array 2.00892

TR Temp 2 76.333

BL Temp 2 77.571

BR Temp 2 74.187

Cursors:	X	Y	Z
Cursor0	6	4	0.61

Intensity Graph 3
Y
0
1
2
3
4
5
6
7
8
9
10
11
X
Amplitude
2.10
2.00
1.90
1.80
1.70
1.60
1.50
1.40
1.30
1.20
1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00

Figure 7. Array's Operator Interface Showing a Small Spot Source

read directly from the display corresponding to the watts per centimeter squared (W/cm^2) present at the array face. Data also is recorded so that it can be viewed later in a player application, such as a video file, or it can be viewed in a spreadsheet application, frame by frame. The data shown in Figure 7 is representative of small-source testing performed recently and very clearly shows the beam profile.

CONCLUSION

NSWCDD engineers successfully met the W-band array's design goals of providing a high temporal-resolution image of 95-GHz beams. The system has been tested against two active denial systems, providing good agreement with the currently accepted methods, as well as valuable information regarding the system's beam characteristics. These accomplishments will allow future system development to take advantage of this better understanding to possibly reduce system size and increase the effective range. A better understanding of the 95-GHz beam helps to facilitate future ADT development for this much-needed, nonlethal escalation-of-force capability for U.S. warfighters, homeland defenders, and law enforcement personnel.

DIRECTED ENERGY IN THE MILITARY ENVIRONMENT

By LT Leedjia Svec, Jeremy Beer, and Dave Freeman

The military operates in the land, air, and maritime environments. In each of these environments, lasers and laser devices are increasingly being seen and used in a variety of ways. Accordingly, the military must protect itself and civilians from the potentially dangerous effects of lasers and other directed-energy devices.

Lasers are being used on the ground to determine the intentions of people who approach checkpoints and to dissuade aircraft from entering restricted airspace. Laser weapons are also being developed for use in the maritime environment. With the use of lasers comes the requirement for eye protection. The eye is particularly sensitive to lasers and its anatomy includes optical components that amplify the power of incoming light. Consequently, the potential for injury or blinding is great.

Naval Medical Research Unit – San Antonio (NAMRU-SA) is poised to lead the way in researching and testing laser glare devices and laser eye protection. The mission of the NAMRU-SA is to conduct medical, dental, and directed-energy biomedical research, which focuses on ways to enhance the health, safety, performance, and operational readiness of Navy and Marine Corps personnel, and addresses their emergent medical and dental problems in routine and combat operations. NAMRU-SA was officially commissioned on 6 May 2009 and is a subordinate command under the Naval Medical Research Center (NMRC) in Silver Spring, Maryland, reporting to Navy Medicine Support Command (NMSC) in Jacksonville, Florida. NAMRU-SA consolidates the Naval Health Research Center Detachment Directed Energy Bioeffects Laboratory, the Naval Institute for Dental and Biomedical Research in Great Lakes, and the NMRC Combat Casualty Care research function. As part of the Base Realignment and Closure (BRAC) 2005, NAMRU-SA has moved to Fort Sam Houston. Two new buildings that have been constructed are the Battlefield Health and Trauma Research Institute and the Tri Service Research Laboratory. A conceptual drawing of the NAMRU-SA Tri-Service Research Laboratory (to house directed-energy research) is shown in Figure 1.

Many factors must be considered when lasers operate in military environments. On the ground, lasers offer a greater likelihood of close contact exposure. In aviation and maritime environments, the mobility of lasers is limited to permanent fixtures on aircraft or ships, so target acquisition can be much more complicated. Often ignored, but just as important and common to all environments, are the psychological factors that need to be explored. These factors include clarifying intentions, communications, and effectiveness. In certain situations,

Figure 1. Naval Medical Research Unit – San Antonio Tri-Service Research Laboratory at Fort Sam Houston, San Antonio, Texas (artist's concept)

sometimes lasers are coupled with other modalities, such as auditory instructions.

On the ground, laser exposure has been shown to interfere with driving vehicles, making color judgments, and target shooting. In aviation, lasers can interfere with pilot vision, causing afterimages, glare, or temporary ocular injury, with attendant effects on navigation and control. In the maritime environment, lights frequently are used to signal a variety of messages, from direction (left, right, etc.) to more complicated messages such as "man overboard." More prolific use of lasers underscores the need for laser eye protection, a dynamic area of research, which must respond to changing threat wavelengths and changing environments. Figure 2 shows NAMRU-SA personnel executing an operational field test at Kennedy Space Center, July 2009.

Recent studies undertaken by NAMRU-SA have investigated the use of laser dazzlers on sailors in small boats.[a] In these studies, participants were exposed to the laser glare at different angles and distances, in both day and night conditions. Study protocols were approved in accordance with the Institutional Review Board in compliance with all applicable federal regulations governing the protection of human subjects. Participants were given a survey assessing their subjective response to the laser, as well as a more objective visual eye chart. The

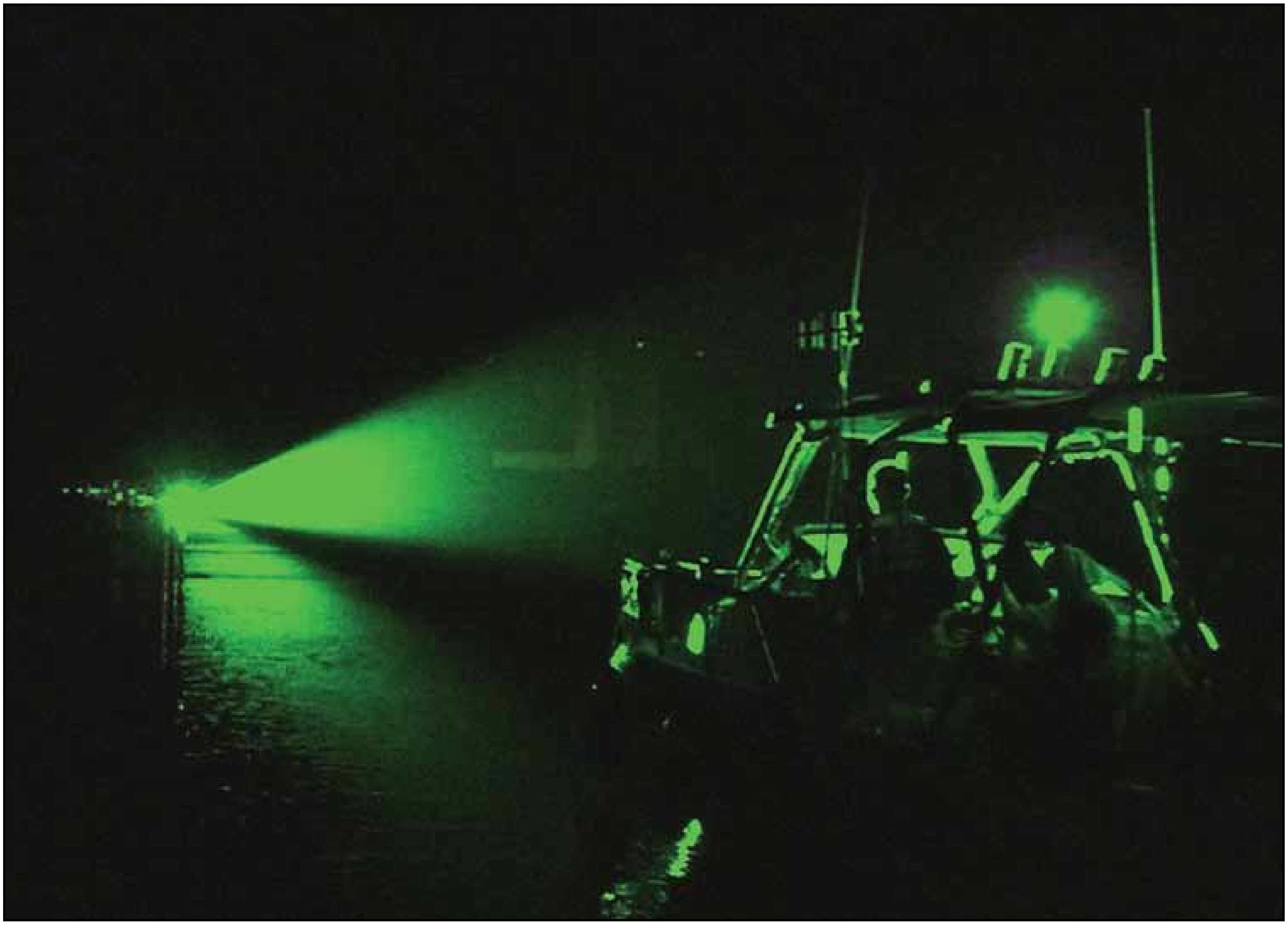

Figure 2. NAMRU-SA personnel execute operational field test at Kennedy Space Center, July 2009, in which a nonlethal laser prototype is evaluated for power delivery (stability and beam propagation) at range and human visual effectiveness aboard a maritime target.

results suggested that participants were most affected by the laser at night when they were looking straight at it (as opposed to many degrees away) and at the closest exposure distances. The most surprising finding, however, was that some participants reported being drawn to the laser rather than away from it, especially at farther distances. Participants remarked that they couldn't tell what the signal was, so they would want to go closer to find out. This illustrates that the assumption (by some)—that distant laser lights will deter and repel innocent mariners—might not always be true. Further research is needed to verify this finding, however, before employing laser glare devices in the maritime environment. Figure 3 shows NAMRU-SA personnel executing operational field tests, which were conducted at Cheatham Annex, Virginia, and Panama City, Florida, in 2008–2009.

These studies also brought the factor of communication to light. Participants remarked that "green is not a threatening color," and some thought "it could be a signal for help." Many felt curious about the "blinking light" used in the study and would go closer or try to contact the vessel to determine the intent of the message. Green lasers are used because they are more visually salient; however, they may not be as psychologically salient. Participants remarked that if the signal were paired with another signal, such as an auditory one, then the message of "warning" or "do not come closer" might be clearer.

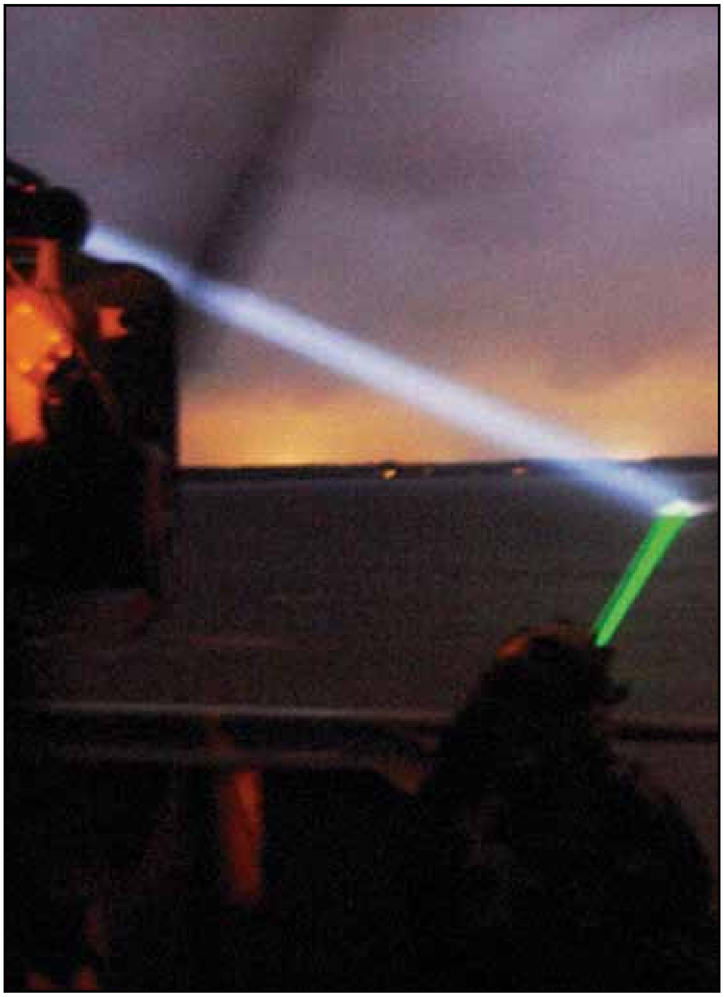

Figure 3. A compact hand-held laser is evaluated for effectiveness in maritime defense against small-boat attacks.

Lastly, these studies brought to light the matter of effectiveness. Laser glare devices are used to stop or alter the behavior of the recipient, but one study yielded mixed results. At close distances, participants noticed the signal, felt affected by it, and reported that their behavior changed in the manner desired by the person pointing the laser. But at greater distances, behavior might not change. Thus, these findings need to be replicated in different maritime scenarios in order to be truly useful in developing laser glare devices. This particular study was encouraging regarding the effectiveness and visual usefulness of glare devices, but it brought up new questions about their psychological impact on behavior. Resolving these questions must be an integral goal of technical research and development studies to determine the operational effectiveness of directed-energy devices, not just for the maritime environment, but for all military environments.

Endnote

a. Results and technical reports are available upon request from the corresponding author or from the NAMRU-SA Public Affairs Officer.

DIRECTED ENERGY USING HIGH-POWER MICROWAVE TECHNOLOGY

By Jacob Walker and Matthew McQuage

The Directed Energy Warfare Office (DEWO) and Directed Energy Division at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) merge past research and data with continuous innovation in the field of high-power microwave(s) (HPM) to address the critical need for nonlethal, nonkinetic weapons. HPM weapons can be described as nonkinetic devices that radiate electromagnetic energy in the radio frequency (RF) or microwave spectrum. They are designed to disrupt, deny, degrade, damage, or destroy targets. In essence, this is achieved when high-power electromagnetic waves propagate through air and interdict targets by traveling through the exterior layers of structures and coupling energy to critical electronic components. Since effectiveness against a wide range of targets is the goal, HPM has become a collective term for various technologies: wave shapes, source frequencies, and the distribution of varying signal bandwidths. It is the objective of HPM research and assessment, therefore, to address targets for which no engagement option currently exists. NSWCDD is working to identify optimal HPM mission platforms and move relevant technologies into the field.

HPM INITIATIVES

NSWCDD has actively pursued HPM research since the advent of the field in the 1970s. Since then, scientists and engineers have conducted HPM research and development in many areas, including hydrogen spark-gap switching, spiral generators, and related technologies. More currently, the Directed Energy Division developed a variety of high-power wideband RF systems based on pulsed power and Marx generators (Figure 1). In addition to the extensive work accomplished in HPM and RF source development, NSWCDD contributed substantially to the area of counter-HPM vulnerability assessments. Researchers developed site assessment guides and threat brochures, as well as a number of wideband RF sources, to determine the susceptibility of electronic equipment to high-power RF interference. This latter effort involved assessing and exploiting the weaknesses of specified electronic targets to various HPM and RF threats. Data gleaned from these efforts was then used to support optimized prototypes and system designs employing effects-based design methodology. NSWCDD utilized these wideband RF sources to determine the susceptibility of a multitude of military

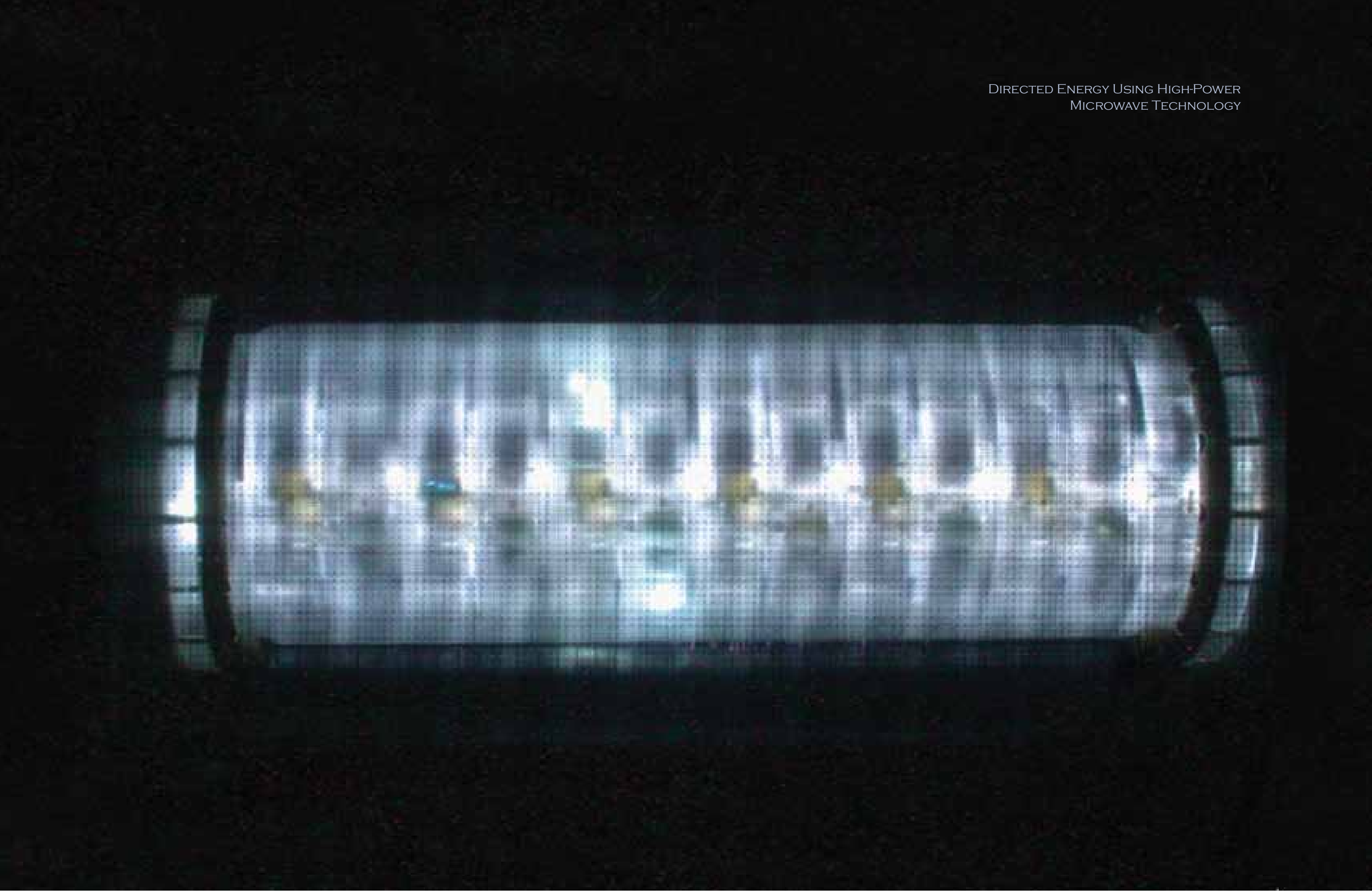

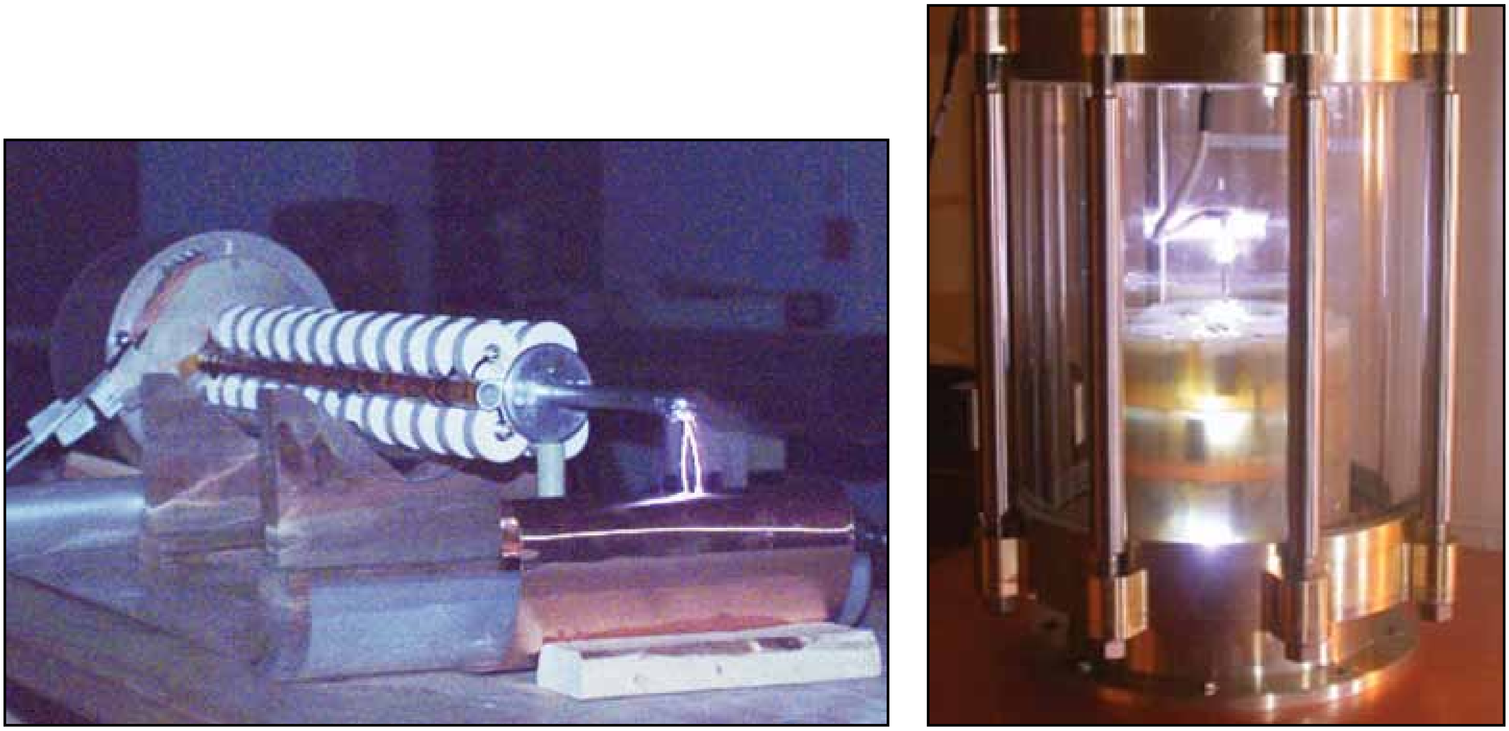

Figure 1. Examples of NSWCDD Marx Generators

and electronic infrastructure equipment to high-power RF interference.

HPM Counterattack Operational Overview

Research in support of HPM-driven electronic attack increased significantly as the demand for nontraditional warfare emerged. Traditional kinetic weapons often are of limited value in peace-keeping missions, for example, as today's enemies frequently are embedded within civilian populations and structures. This creates the need for novel HPM technologies that minimize the risk of collateral damage while effectively neutralizing threats. Dahlgren researchers conduct HPM system research and development—as well as lethality and weapon effectiveness assessments—to address this need while developing technologies against a wide variety of electronic targets. These projects leverage NSWCDD's assets, including the Maginot' Open Air Test Site (MOATS), state-of-the-art RF diagnostics, and modeling and simulation tools to identify applications and platforms in which HPM technologies can be employed. Figure 2 shows a computer model of the MOATS facility and a modeling and

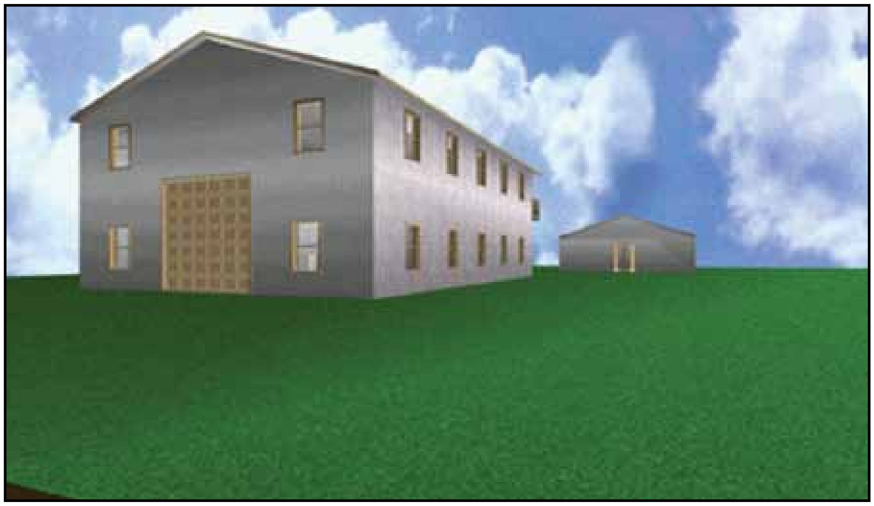

(a)

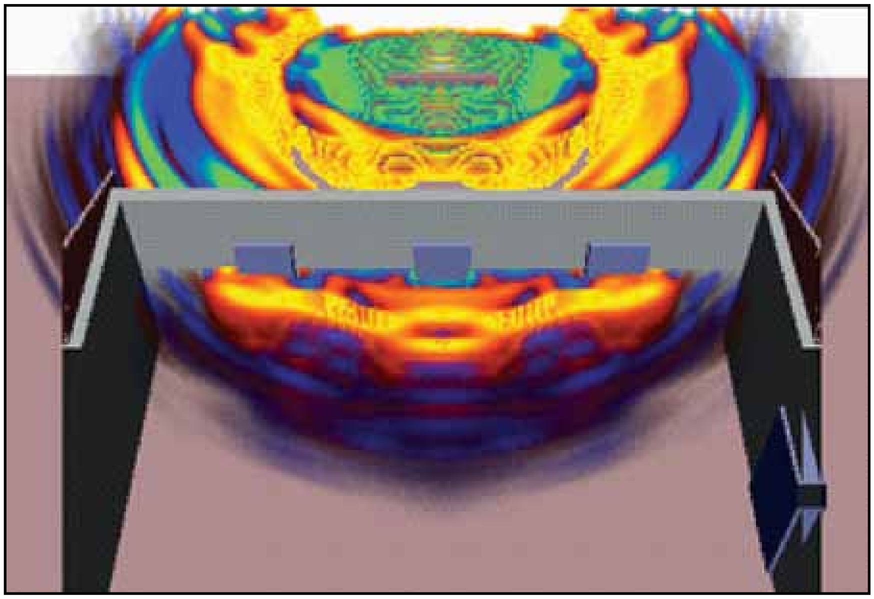

(b)

Figure 2. Modeling and Simulation Depicting (a) NSWCDD Test Facility and (b) Simulation of Radiated RF

simulation graphic depicting the RF emitted by an HPM dipole antenna.

Potential platforms for HPM integration include: man-portable, aerial, vehicle, and vessel-mounted systems. These platforms all provide unique methods for delivery of HPM sources. For example, aerial delivery—which, in many ways, is the most challenging due to size and weight constraints—can increase the effective range of these systems and can engage multiple targets at close range without endangering personnel. Likewise, vehicles and vessel-mounted HPM systems provide a way for law enforcement and the military to stop vehicles in chase scenarios almost as soon as they begin. The goal of all of these projects is to provide military forces with the ability to employ nonkinetic, electronic strike technologies against an adversary's electronics.

The DEWO and Directed Energy Division are uniquely positioned to provide numerous capabilities for in-house development while engaging with the private sector to test and provide feedback on HPM systems developed externally. In the past decade, NSWCDD has evaluated several HPM systems at Dahlgren to determine their effectiveness against various electronic targets while maintaining the Office of the Secretary of Defense's Tri-Service RF Directed Energy Weapon (DEW) Database. This database contains all effects data collected from directed-energy tests performed within the U.S. Air Force, Army, and Navy.

Conclusion

NSWCDD continues to pioneer HPM source development and lethality and integration studies, leading to the demonstration and delivery of prototype capabilities. It also is committed to researching and developing critical subsystems for HPM delivery. By leveraging numerous target assets and sophisticated diagnostic equipment—in conjunction with MOATS—NSWCDD has positioned itself at the forefront of HPM electronic attack, leading the way in the development and delivery of these capabilities to the warfighter.

Acknowledgment

Nancy Muncie, Bowhead, contributed to this article.

LASER COUNTER ROCKET, ARTILLERY, AND MORTAR (C-RAM) EFFORTS

By Michael Libeau

Mortars and rockets are common weapons confronting U.S. troops abroad. Insurgents fire the inexpensive projectiles into populated areas, intending to kill or injure service members and to inflict physical damage. While kinetic solutions like guns and missile interceptors are used to counter rockets and mortars, laser counter rocket, artillery, and mortar (C-RAM) systems present a promising solution to counter these challenging threats in the near future.

Scientists and engineers at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) have been researching, developing, testing, and evaluating laser C-RAM systems through collaboration, modeling and simulation, and experimentation. The Joint Technology Office (JTO) and the Directed Energy and Electric Weapons Program Office (PMS 405) sponsored the first year of these initiatives in 2007. Consecutive and current work has been sponsored by the Office of Naval Research (ONR) Expeditionary Maneuver Warfare and Combating Terrorism S&T Department.

BACKGROUND

In preparation for the development of a laser C-RAM system, an understanding of the vulnerability of rockets and mortars to laser energy was crucial. Engineers from NSWCDD and the U.S. Army Space and Missile Defense Command (SMDC) collaborated on laser C-RAM efforts. Engineers analyzed the RAM threat and examined a variety of targets, accessing RAM vulnerabilities to laser energy by utilizing theoretical, numerical, and experimental work. They then developed theoretical models that captured the physics of the laser-induced failures of targets containing high explosives (HE). Additionally, NSWCDD engineers enhanced lethality simulations using a tool called the Effectiveness Toolbox to model engagements of RAM targets with laser energy. Figure 1 shows a screen capture from the Effectiveness Toolbox.

The resulting simulations included results from a laser atmospheric propagation model and a thermal model to determine the effect of the laser energy on the target. The simulations also incorporated target trajectories necessary for modeling the changing laser conditions on the target resulting from the engagement of a ballistic target. Subsequent to modeling these effects, live testing was performed. Figure 2 shows the lasing and destruction of a RAM target during live testing.

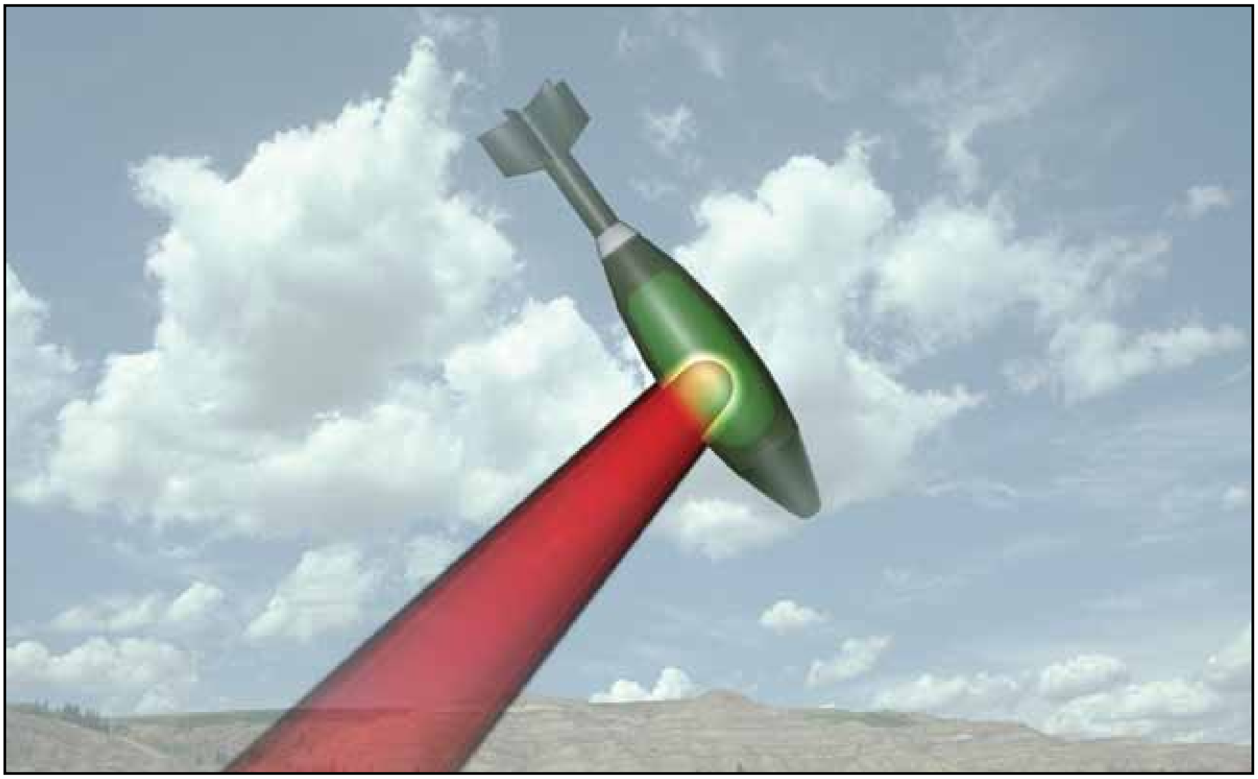

Figure 1. Screen Capture from the Effectiveness Toolbox Showing the Laser Engagement of a Mortar Target

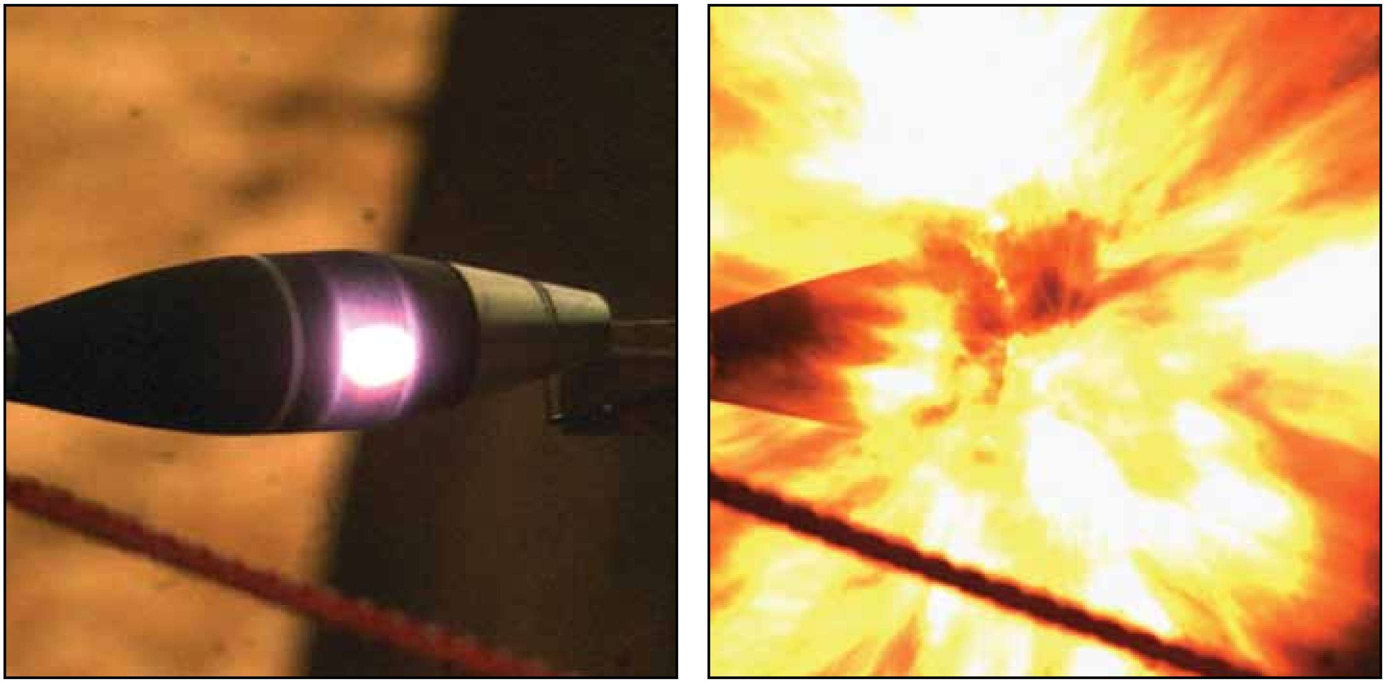

Figure 2. Explosive Target is Destroyed with NSWCDD’s Fiber Lasers

NSWCDD engineers conducted two large experimental tests to determine the vulnerability of HE targets to laser energy using NSWCDD's High-Energy Fiber lasers. The first test was conducted jointly with SMDC. During these two tests, over 40 RAM targets were destroyed under different laser conditions, producing significant information on laser lethality. Researchers measured the failure times of multiple targets for different laser powers, spot sizes, incidence angles, and aimpoints. The experimental data yielded by the tests increased engineers' understanding of the vulnerability of targets containing energetic materials. This data was then used to benchmark predictive models.

Future tests are planned with additional HE targets to further the knowledge of RAM vulnerability. These tests are controlled and conducted carefully to ensure that good data is obtained. Accurate measurements of laser power on the target and the resulting target failure times must be made during the tests. To that end, NSWCDD engineers leverage Division-wide expertise in lasers and optics with its long history of explosives testing to achieve meaningful test results. NSWCDD personnel have been instrumental in improving techniques to measure the spatial profile of laser power on a target. The spatial distribution of laser power on a target is critical to understanding the target's failure. Figure 3 shows a laser beam's spatial power distribution measured during a test.

ONGOING LASER C-RAM INITIATIVES

Recent advances in fiber lasers have increased the power outputs of these rugged, solid-state devices. Both government and contractor efforts are examining the application of commercial off-the-shelf (COTS) lasers and other electric lasers for application into advanced weapon systems. The Navy's Laser Weapon System (LaWS) Program, for example, is examining a laser system built around efficient fiber lasers. This is significant because a high-energy fiber laser system offers two critical advantages over gun and missile interceptor C-RAM systems. First, the laser has a great depth of magazine since it requires only electricity for operation. Consequently, unlike a gun system, which has a limited supply of ammunition, a laser system is limited only by its supply of electrical energy. Second, a laser system offers a cost per kill that is significantly lower than alternative systems because only electricity is being expended instead of gun ammunition or a costly missile interceptor. This low cost per kill also better matches the low cost of the RAM target being engaged.

High-Energy Fiber Laser C-RAM systems will provide significant advantages in defeating the RAM threat while augmenting existing C-RAM solutions. More importantly, laser C-RAM systems will help protect members of the armed forces from the inexpensive, yet often deadly threats posed by rockets and mortars.

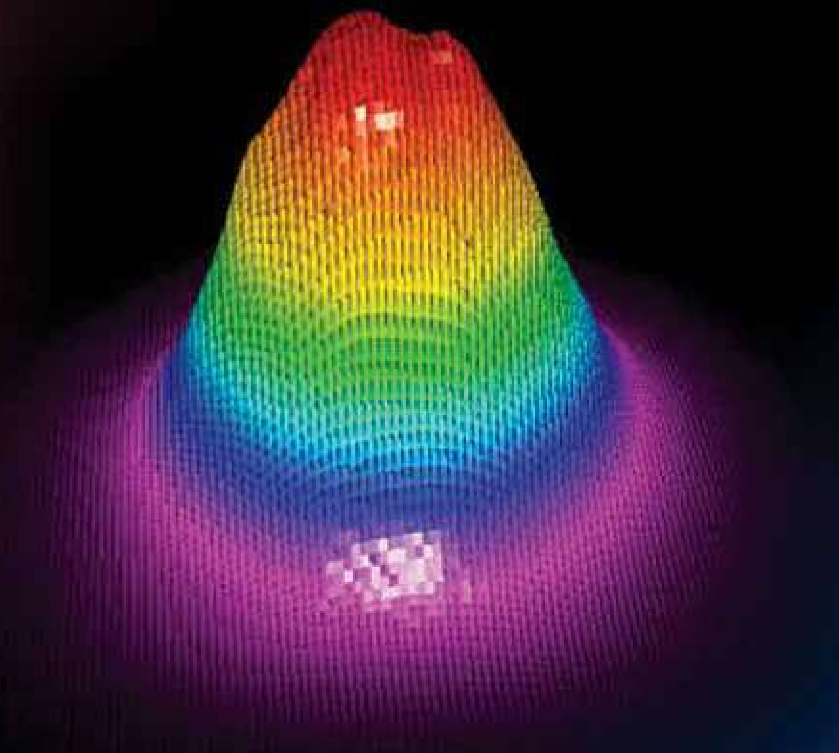

Figure 3. Laser Power Spatial Variation from a C-RAM Test

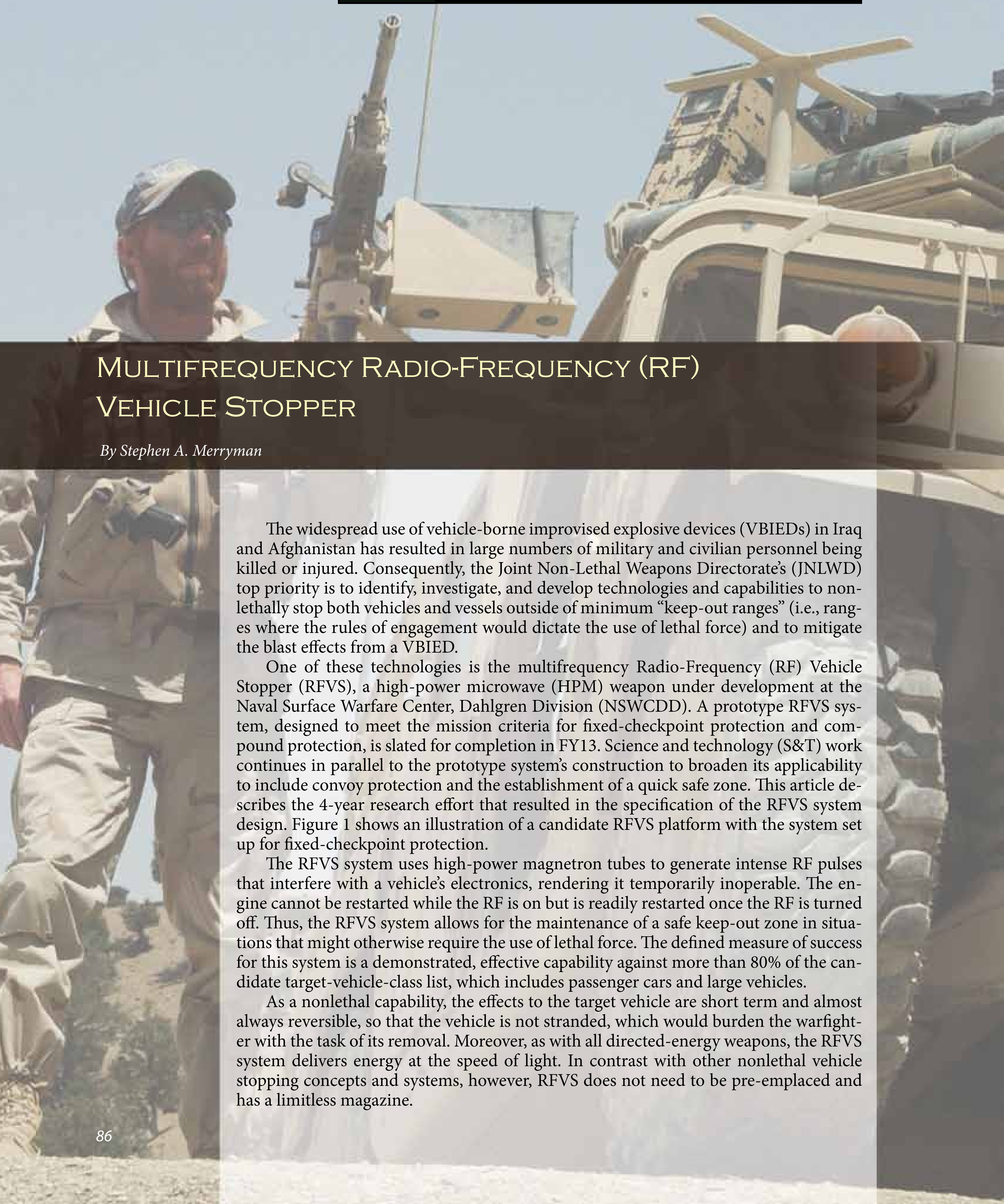

MULTIFREQUENCY RADIO-FREQUENCY (RF) VEHICLE STOPPER

By Stephen A. Merryman

The widespread use of vehicle-borne improvised explosive devices (VBIEDs) in Iraq and Afghanistan has resulted in large numbers of military and civilian personnel being killed or injured. Consequently, the Joint Non-Lethal Weapons Directorate's (JNLWD) top priority is to identify, investigate, and develop technologies and capabilities to non-lethally stop both vehicles and vessels outside of minimum "keep-out ranges" (i.e., ranges where the rules of engagement would dictate the use of lethal force) and to mitigate the blast effects from a VBIED.

One of these technologies is the multifrequency Radio-Frequency (RF) Vehicle Stopper (RFVS), a high-power microwave (HPM) weapon under development at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD). A prototype RFVS system, designed to meet the mission criteria for fixed-checkpoint protection and compound protection, is slated for completion in FY13. Science and technology (S&T) work continues in parallel to the prototype system's construction to broaden its applicability to include convoy protection and the establishment of a quick safe zone. This article describes the 4-year research effort that resulted in the specification of the RFVS system design. Figure 1 shows an illustration of a candidate RFVS platform with the system set up for fixed-checkpoint protection.

The RFVS system uses high-power magnetron tubes to generate intense RF pulses that interfere with a vehicle's electronics, rendering it temporarily inoperable. The engine cannot be restarted while the RF is on but is readily restarted once the RF is turned off. Thus, the RFVS system allows for the maintenance of a safe keep-out zone in situations that might otherwise require the use of lethal force. The defined measure of success for this system is a demonstrated, effective capability against more than 80% of the candidate target-vehicle-class list, which includes passenger cars and large vehicles.

As a nonlethal capability, the effects to the target vehicle are short term and almost always reversible, so that the vehicle is not stranded, which would burden the warfighter with the task of its removal. Moreover, as with all directed-energy weapons, the RFVS system delivers energy at the speed of light. In contrast with other nonlethal vehicle stopping concepts and systems, however, RFVS does not need to be pre-emplaced and has a limitless magazine.

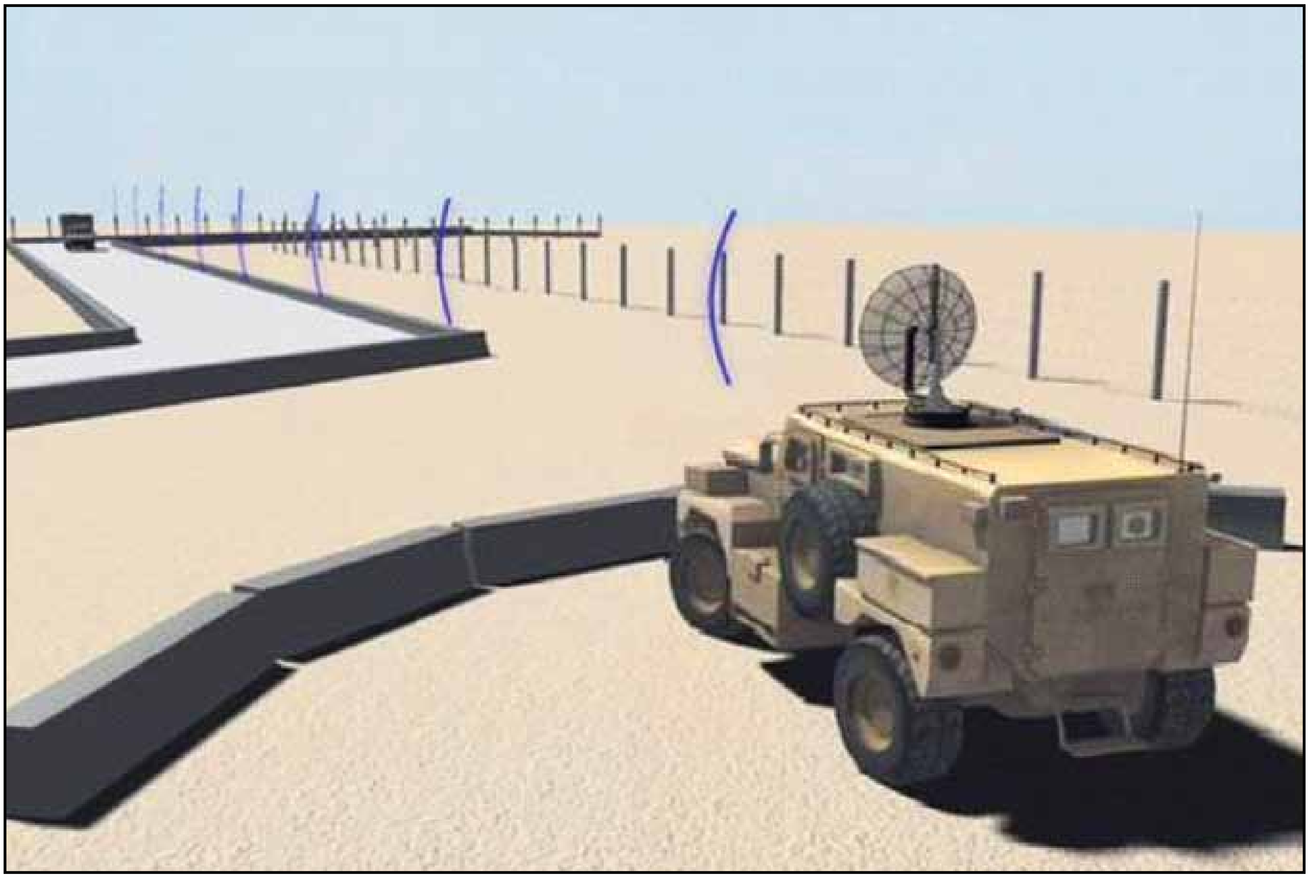

Figure 1. Illustration of Candidate RFVS System Setup for Checkpoint Protection

BACKGROUND

Using HPM or RF energy to stop an automobile engine is not a new concept; it has been under investigation for some time in private, academic, and military sectors. To that end, the RFVS program leveraged as much historic work as possible while collaborating with academic and military laboratories and while aggressively pursuing contacts in the automobile industry to gain knowledge of vehicle electronic design and function.

In 2005, the JNLWD funded the then-Directed Energy Technology Office (DETO) at NSWCDD to perform an extensive reverberation chamber test series to characterize the vulnerability of a representative cross section of automobiles to a wide range of HPM source frequencies.[a] The purposes of the tests were twofold. First, the applicability of the Army's Ground Vehicle Stopper (GVS) data set needed to be established for newer vehicles, and second, a thorough, source-technology independent assessment of vehicle vulnerabilities needed to be performed. The rationale behind the latter was to establish vehicle vulnerabilities without inadvertently biasing the process. Only after the full assessment was performed would factors such as concept of operations (CONOPS) and system requirements come into play. Figure 2 is a photograph of reverberation chamber testing.

Over the past decade, a significant number of private, academic, and military laboratories have investigated the susceptibility of automobiles to HPM energy. The range in approaches spans the gamut from isolated component testing, through direct injection and radiated testing of electronic control units (ECUs), and continuing through full vehicle radiated testing. Each of the different test methods has its strengths and weaknesses. Testing of isolated ECUs in controlled laboratory conditions is arguably the best way to determine exactly how a specific unit is responding to the RF. However, whether or not the identified susceptibilities continue to hold true when the unit is in place in a

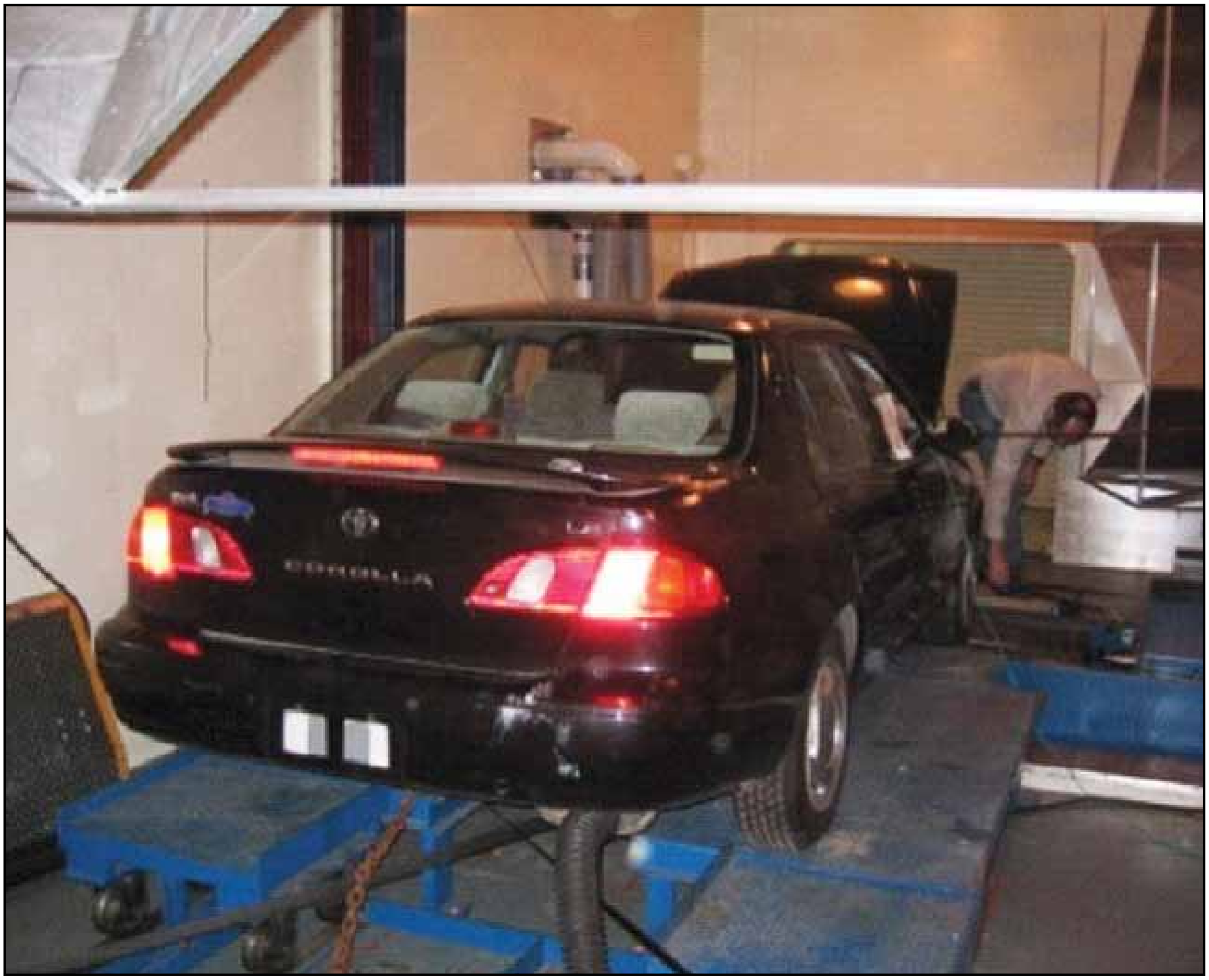

Figure 2. Vehicle on a Dynamometer in the NSWCDD Reverberation Chamber

vehicle, or whether the results apply to other vehicles' ECUs, remains a significant question that limits the applicability of the results.

Full vehicle testing and failure analysis of the ECU can be a daunting task. That said, full vehicle testing affords the advantage of ensuring that the response is commensurate with expectations of genuine engagements. The test approach one chooses to take depends upon resources, test facility availability, and most importantly, the objectives of the program. For the RFVS program, the objectives were to identify an HPM waveform that is effective against a broad class of the candidate target vehicles and to ensure that the identified waveform could be generated with a source that can be packaged in a footprint and cost amenable to military users. To meet the program's objectives, the RFVS program chose to invest the majority of its resources in full vehicle testing. While the focus of the effects testing portion of the RFVS program has remained on full vehicle testing, both time and resources have been devoted to fostering and maintaining connections with academia and the auto industry. There is concerted effort to keep abreast of the latest trends in automotive technology, to ensure that the current RFVS system design will continue to be effective against future vehicle designs, and to leverage all research that might aid in future RFVS designs.

SYSTEM OPERATION

The majority of current HPM system concepts employ a narrowband, single-frequency HPM source. In contrast, RFVS utilizes multiple HPM frequencies. The rationale for using multiple frequencies is associated with increased system effectiveness. Electromagnetic (EM) energy can be used to disrupt or damage an electronic target. In order for the energy to affect the electronics, however, it must be able to reach a critical component(s) inside the target. This involves a process referred to as coupling. Different EM waveforms are more or less effective against specific targets depending, in part, on their frequency, as different frequencies couple better or worse depending on varying target geometries. To be specific, each piece of electronics has specific resonance frequencies that most effectively facilitate coupling energy to the target. Unfortunately, these resonant frequencies can be unique to each piece of equipment. Consequently, a single-frequency waveform might be very effective against one target, but less effective against another target. Therefore, a system that utilizes either a sweep of frequencies or multiple frequencies will be more effective against a larger target set. This is not a novel idea, but rather one that has been readily acknowledged within the HPM community for some time and fervently embraced by the RFVS program. Current technology limitations prohibit high-power-swept frequency sources as viable options, leading to the idea of a multifrequency source. The more frequencies that are used, the more effective the system; however, a trade-off is made with system size and cost as the number of source frequencies is increased.

Brassboard System

After completion of the exhaustive vehicle effects characterization testing in 2006, the RFVS program identified the optimal number of frequencies needed to meet mission requirements. It then used this information in the design and construction of the Brassboard System. The purpose in constructing the Brassboard System was to demonstrate the benefit of the multifrequency approach and the ability to meet mission objectives with specified power on target requirements. Construction of the RFVS Brassboard System began in 2007 and was completed in 2008. The Brassboard System was not constructed with specific system footprints in mind. Thus, the antenna and conex used are significantly larger than those in the prototype design. The RFVS team collaborated with a Marine Corps service representative identified by the JNLWD to flesh out the specifics of the mock checkpoint to be used in the RFVS Brassboard System Demonstration. Figure 3 provides a diagram of the checkpoint setup used in the RFVS Brassboard Demonstration. Figure 4 provides photographs of the RFVS Brassboard System.

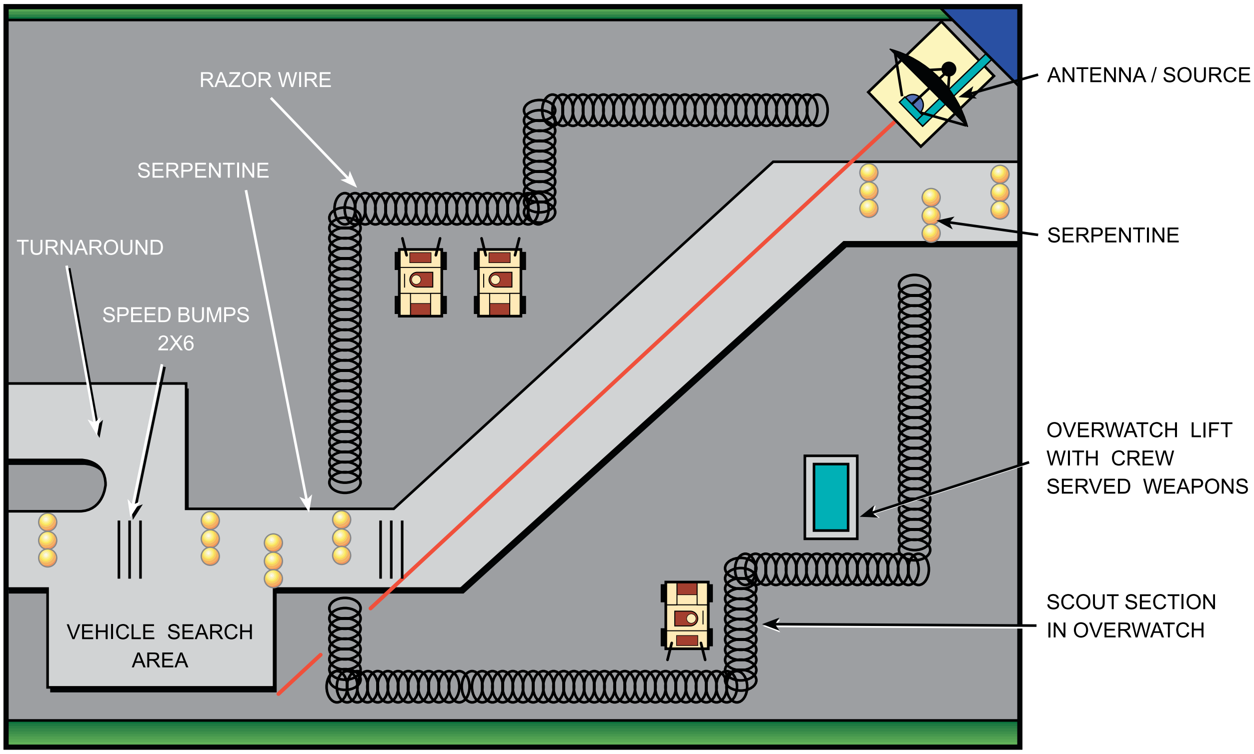

Figure 3. Schematic of Checkpoint Setup

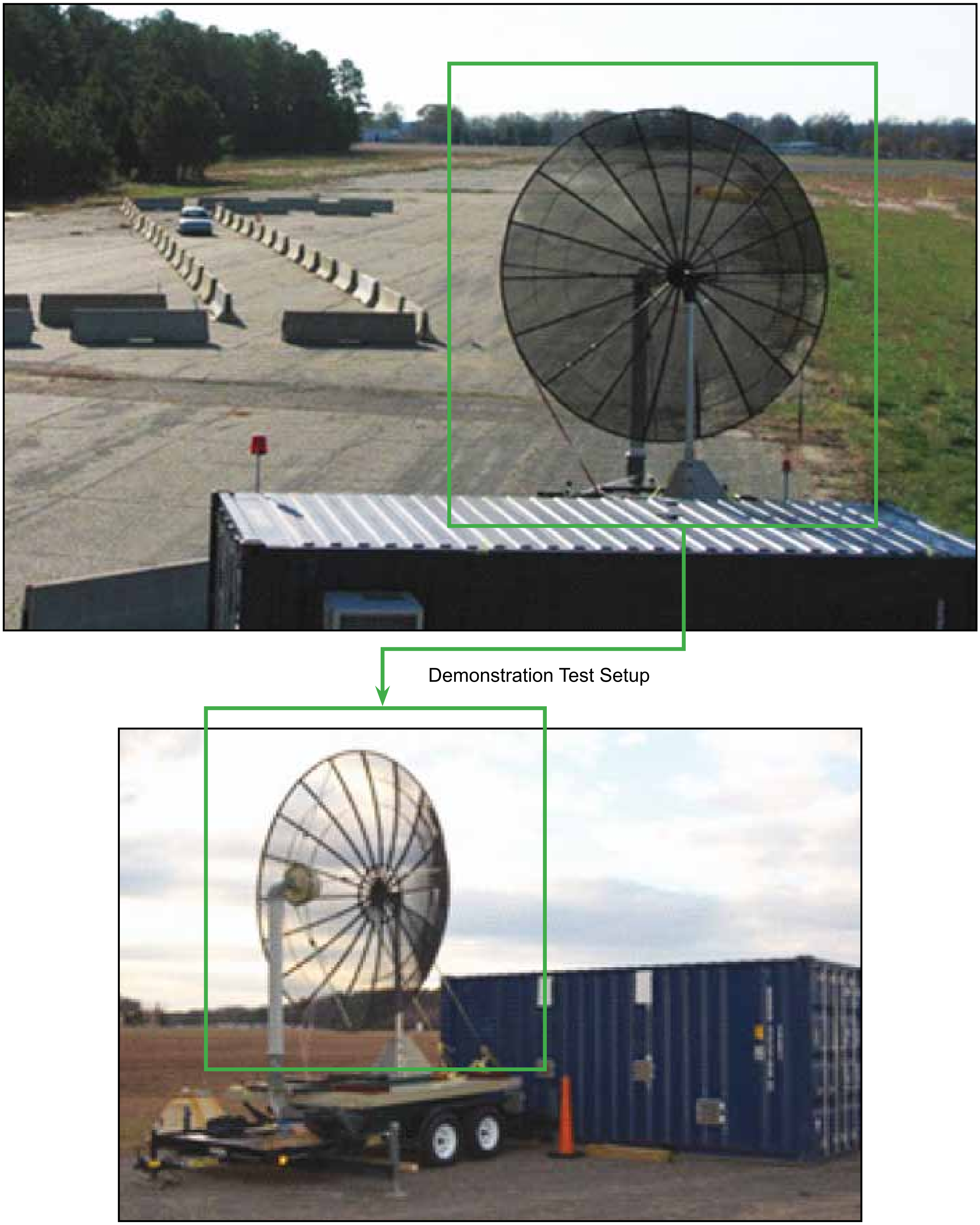
Demonstration Test Setup

Figure 4. Photograph of the RFVS Brassboard System and Demonstration Setup

The Brassboard System Demonstration was conducted in Spring 2008. The Demonstration was a success, and funding for the RFVS prototype was consequently approved. To date, 42 passenger vehicles (cars, pickup trucks, vans, and sport utility vehicles (SUVs)) and 3 large trucks (dump truck and tractors) have been tested as part of the RFVS program.

Way Ahead

The JNLWD continues to work with the Directed Energy Warfare Office (DEWO) toward the development of a fieldable multifrequency RFVS system. Once the capability is fully developed, tested, and certified ready for operational use, warfighters and civilians alike will benefit greatly. Lives will no doubt be saved using the ability to stop vehicles nonlethally and mitigate the blast effects from VBIEDs.

Acknowledgment

Dr. Cynthia Ropiak (SAQ Consulting) contributed to this article.

Endnote

a. The Directed Energy Technology Office (DETO) was renamed the Directed Energy Warfare Office (DEWO) in August 2009. For reference, see the charter for the DEWO, NSWCDD, 17 August 2009.

High-Power Electrical Vehicle-Stopping Systems

By Jordan Chaparro and Melanie Everton

The military needs devices that can safely and reliably stop or arrest vehicles. The primary concern is security at entry control points and vehicle check points similar to the one shown in Figure 1. In such scenarios, it is desirable to be able to stop unauthorized vehicles at predefined standoff ranges to protect personnel, equipment, and critical infrastructure.

Both the military and civilian law enforcement agencies face similar issues with chase scenarios, where concerns over bringing an offending vehicle to a stop without killing or injuring innocent civilians, or causing collateral damage, often prolongs high-speed pursuits. That said, currently employed nonlethal options for arresting vehicles have significant logistical limitations and carry a high cost per use.

The Naval Surface Warfare Center, Dahlgren Division's Directed Energy Warfare Office (DEWO), under the sponsorship of the Joint Non-Lethal Weapons Directorate (JNLWD), investigated compact systems designed to couple high-power electrical impulses to a target vehicle to stop its engine. Such systems are highly portable, can operate remotely, can be deployed quickly by a two-man team, and can engage hundreds of targets before requiring any significant maintenance.

System Overview

Conceptually, electrical vehicle-stopping systems are fairly simple devices. The systems use several stages of energy compression to take a low-peak power source—like a battery pack—and create very intense, short-duration, oscillating electrical impulses. The block diagram, shown in Figure 2, illustrates the principal components of such a system.

A high-energy density, 300-V lithium battery pack, similar to what might be found in a hybrid vehicle, serves as the prime power source for the device. These batteries are capable of driving the system for hundreds of engagements before requiring recharge.

The direct current bus from the batteries is stepped up to several kilovolts in order to charge a capacitive voltage multiplier, such as a Marx Generator, Spiral Line Generator, or Tesla Transformer. Once triggered, these generators charge a resonant circuit to hundreds of kilovolts which, when switched, generate the desired oscillating waveform. Coupling this electrical pulse to a target may be accomplished by direct electrode contact, by radiating the waveform from a broadband antenna structure, or by a combination of both methods.

Figure 1. An Azerbaijani Soldier Guarding Entry Control Point 1 at the Haditha Dam in Support of Operation Iraqi Freedom

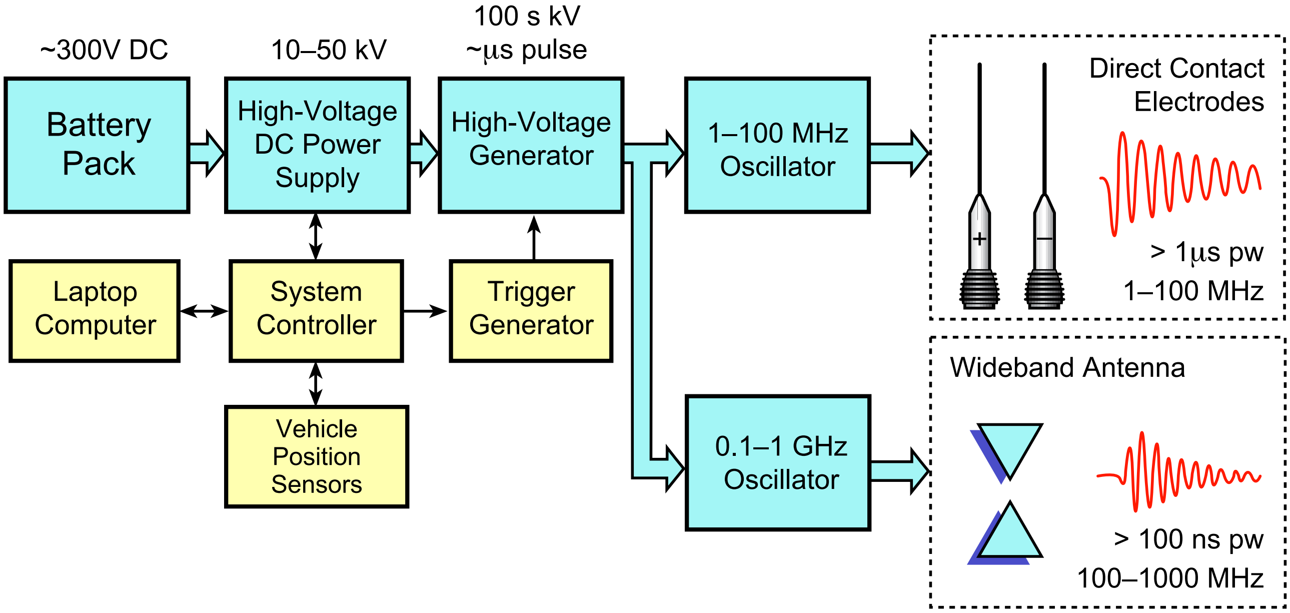
~300V DC
10–50 kV
100 s kV
~µs pulse
Battery Pack
High-Voltage DC Power Supply
High-Voltage Generator
1–100 MHz Oscillator
Direct Contact Electrodes
+
−
> 1µs pw
1–100 MHz
Laptop Computer
System Controller
Trigger Generator
Vehicle Position Sensors
0.1–1 GHz Oscillator
Wideband Antenna
> 100 ns pw
100–1000 MHz

Figure 2. System Block Diagram for Generic Electrical Vehicle-Stopping Systems

The system is monitored and controlled by an integrated system computer. A laptop computer, remotely connected to the system controller through either fiber optic or a wireless network, can be used to arm the system. At this point, motion detection sensors trigger the pulse train upon the targeted vehicle. The laptop can also be used to monitor the system's status, change system parameters, and receive data collected during the last engagement event.

A conceptual rendering of how such a system might look when in use is shown in Figure 3. Traffic would be funneled with barriers to a single lane. When not engaged, the system electrodes would sit flush with the roadway unit, with an exposed height of less than 3 inches. When required, the electrodes could be released to make contact with a vehicle's undercarriage and deliver the electrical impulses.

COMPARISON WITH EXISTING SYSTEMS

Tire spike systems are frequently employed but do not limit the momentum, drive, or control of a vehicle to an extent that could be useful in any type of control or checkpoint scenario. Consequently, while tire spike systems are primarily used in high-speed pursuit applications, they are limited, in that they cripple the target just enough to allow law enforcement to force the vehicle to a stop.

Restraining nets are most comparable to electrical vehicle stoppers with respect to their intended application and desired effect. Restraining net systems and electrical vehicle stoppers both completely arrest vehicles, although by different means. Restraining nets bind the front axle of the vehicle, causing it to forcibly lose momentum and skid to a stop. Thus, the vehicle operator loses the ability to steer the vehicle, further resulting in a lower potential for collateral damage. Electrical systems stop the engine of the vehicle, leaving the operator with control for the duration of the vehicle's momentum. Physical barrier structures can then be employed to force an affected vehicle to stop in a fairly short distance. Modern vehicles lose power steering when the engine is cut off, such that the maneuverability of the vehicle is limited enough to

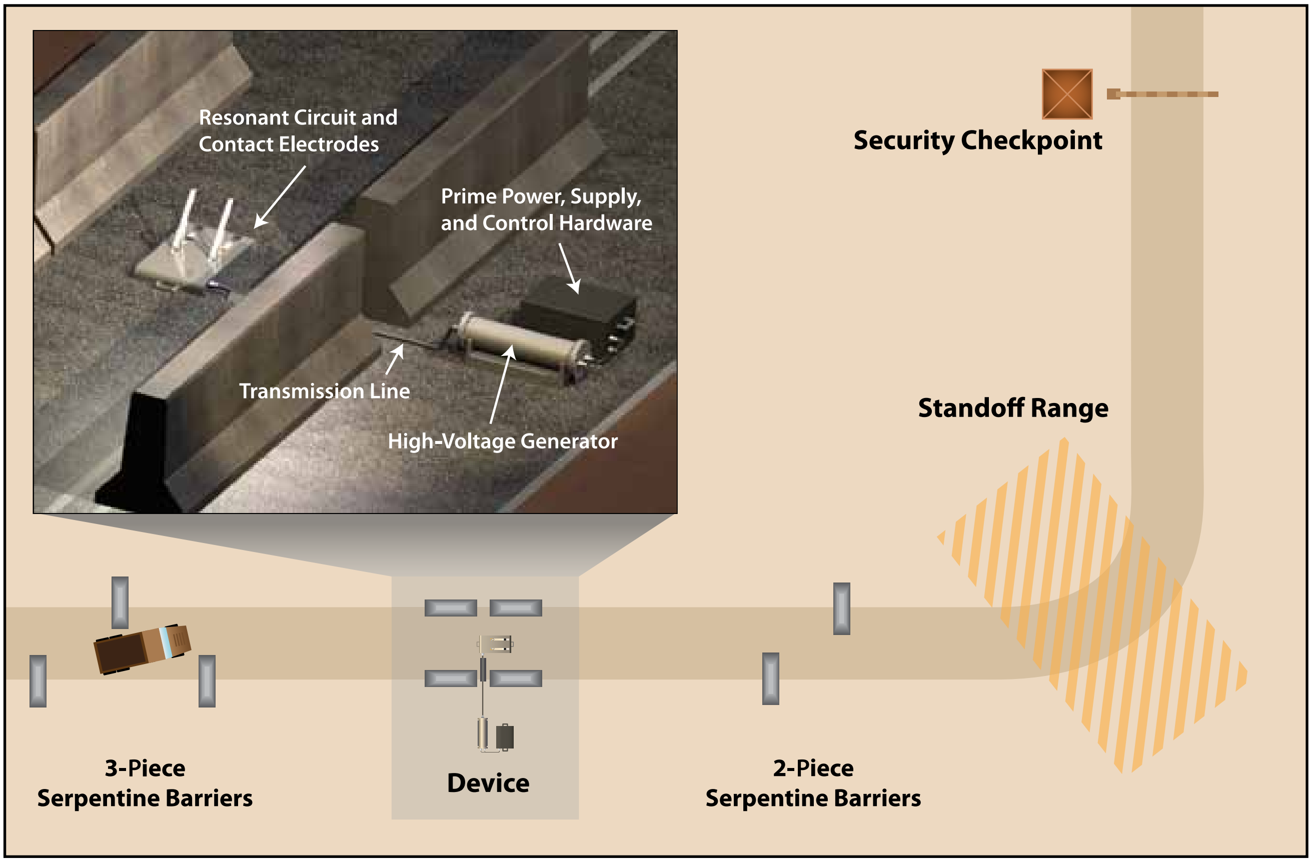
Resonant Circuit and Contact Electrodes
Prime Power, Supply, and Control Hardware
Transmission Line
High-Voltage Generator
Security Checkpoint
Standoff Range
3-Piece Serpentine Barriers
Device
2-Piece Serpentine Barriers

Figure 3. Conceptual Rendering of an Employed Electric Vehicle Stopping System

allow normally nonrestrictive serpentines to be effective at limiting roll-off distances.

One key logistical advantage of electrical vehicle stoppers, compared to restraining nets, is the average cost per engagement. Restraining net systems are one-time use devices that cost several thousand dollars each. Electrical systems initially cost tens of thousands of dollars but can perform thousands of stops within the expected lifetime of the device. Also, there is no requirement to physically reset or reload an electrical system, as with restraining nets. The maintenance required for electrical systems involves the occasional replacement of electrode arms and the inspection of the system connections and pressure levels.

Operationally, both systems have limitations on the types of targets that can be effectively stopped. Restraining nets are limited by vehicle momentum, which can be a product of high speeds or large vehicles. Electrical systems are not limited by vehicle size or speed, but they require additional support from structures—such as serpentines or speed bumps—to force the target to brake and dissipate its momentum once the engine has been stopped.

Both devices typically cause damage to targeted, stopped vehicles. Restraining nets almost always cause tire damage. Less commonly, brake lines, front axles, wheels, and transmissions also might be damaged. Electrical systems typically damage engine controllers, security modules, and engine sensors. In addition, noncritical parts—such as gauges, radios, and cabin fans—also might be damaged. Moreover, moving affected targets is much less of an issue with electrical stoppers than vehicles stopped by net systems, which must first have the net cut away and freed before the target is moved to the side of the roadway.

System Refinement and Look Forward

Previous attempts to field electrical vehicle-stopping systems have been hampered by limited success rates on a large population of vehicles. Many models of vehicles are easily affected by any type of large injected current, while others are fairly resistant. Through carefully designed and controlled experiments, and logistical regression modeling techniques, the DEWO team has been able to determine key waveform attributes that scale with stopping effectiveness rates on a representative population of vehicles. Successful stop rates exceeding 90 percent have been achieved on a diverse vehicle test set by engineering system resonators to enhance system performance The DEWO team, through continued research, testing, and evaluation, is continuing its work to increase the reliability and effectiveness of these systems to make them more compact and to improve their functionality for future military and law enforcement applications.

NONLETHAL SMALL-VESSEL STOPPING WITH HIGH-POWER MICROWAVE TECHNOLOGY

By Jacob Walker

The employment of small vessels to attack merchant ships and other seafaring units has emerged as a significant threat to international navigation and safe operations on the high seas. Along with swarm tactics, small vessels have been known to carry improvised explosive devices, help smuggle terrorists and weapons, and serve as attack platforms on the water for larger weapons. While kinetic solutions serve as the decisive option, alternative solutions that employ nonlethal means are being explored. A depiction of a swarm of small vessels ready to attack is shown in Figure 1.

The Naval Surface Warfare Center, Dahlgren Division's (NSWCDD's) Directed Energy Warfare Office (DEWO) is evaluating directed-energy (DE) concepts based on high-power microwave (HPM) technology for nonlethal vessel-stopping applications. Nonlethal weapons are defined by the Department of Defense (DoD) as weapons that are explicitly designed and primarily employed so as to incapacitate personnel or materiel while minimizing fatalities, permanent injury to personnel, and undesired damage to property and the environment.[1]

Several methodologies exist for using nonlethal means to stop small vessels. They include:

- Running-gear or prop entanglement systems
- Exhaust stack blockers
- A sea-anchor vessel-stopping system, which casts a net across the bow of a vessel to impart resistance
- Small-craft disablers, which insert a spear into the hull and deploy a fin that drags in the water, making steering impossible

Prop entanglement systems, exhaust stack blockers, and sea-anchor systems are useful and effective, but all are operationally difficult to deliver when deployment methods rely on positioning them in front of, or directly over, a vessel moving

Figure 1. Depiction of a Small-Vessel Swarm Ready to Attack

at high speeds. Small-craft disablers also are a formidable vessel-stopping solution and may be easier to deploy, but they cause permanent damage to the vessel in question.

Under the direction of the Joint Non-Lethal Weapons Directorate (JNLWD), the DEWO is in the initial stages of a multiyear effort to evaluate DE concepts for nonlethal vessel-stopping applications. It is currently focusing on HPM technology. This technology uses HPM sources to radiate radio frequency (RF) pulses downrange to interfere with motor-control electronics and significantly impede or stop small-vessel motors with minimal collateral damage. These RF pulses can be generated using different technologies ranging from wideband LC oscillators and microwave tubes (e.g., magnetrons, klystrons, and backward wave oscillators) to emerging solid-state technologies (e.g., nonlinear transmission line and photo-conductive switching). An outboard motor on a test stand is shown in Figure 2.

In comparison to kinetic weapons or other nonlethal systems, HPM avoids gross physical destruction to the vessel while, more importantly, providing zero-to-low risk of human injury. HPM accomplishes this at safe distances using speed of light delivery, therefore making evasion difficult, if not impossible, with the added benefit of scalable effects ranging from disruption to damage. Despite its numerous advantages, the use of HPM technology as a nonlethal weapon presents challenges as well, including a trade-off between system size and standoff range. This is particularly important when considering the use of HPM systems in different environments.

Upfront HPM source development costs represent one of the biggest challenges. However, long-term savings associated with HPM technology can offset this challenge. For example, prop entanglement systems might be deployed only once before they are rendered useless. HPM sources integrated onto a ship or other military vehicle can be employed in potentially thousands of missions, therefore resulting in a lower cost per single use, bringing overall associated costs of the system down significantly. Priorities for HPM nonlethal weapons include developing a system effective against different types of small vessels.

NSWCDD's Directed Energy Division began HPM susceptibility testing to determine the effectiveness of HPM weapons against relevant outboard engines. This involves testing small vessels in

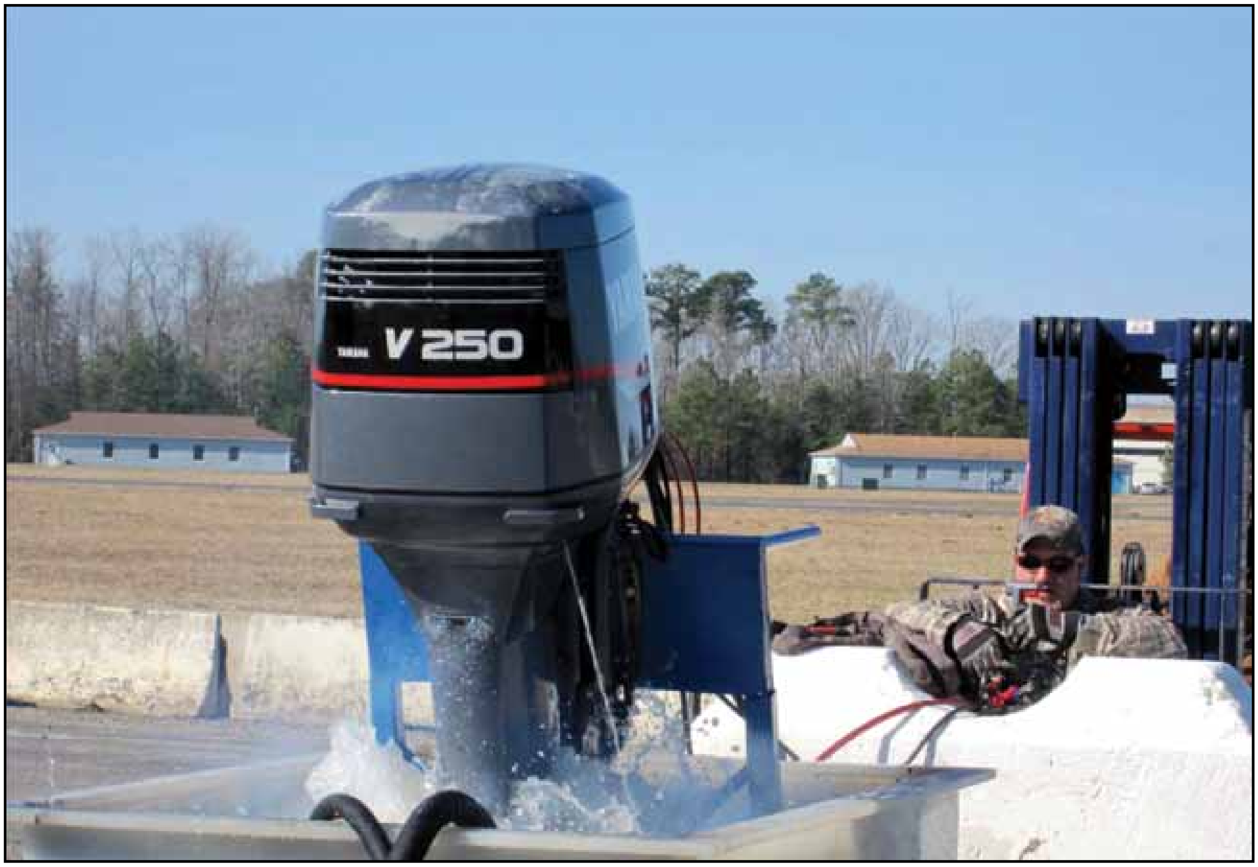

Figure 2. Outboard Motor Test Stand

a variety of environments, including reverberation and anechoic chambers, and open-air testing. All help identify different, effective waveform parameters such as frequency, pulse width, rise time, and required power or energy on target. They further facilitate the identification of design specifications necessary for an eventual HPM source. This source, once developed, will then be integrated into one of several potential platforms. Candidate concepts of deployment include U.S. Coast Guard and naval vessels in addition to unmanned surface or aerial vessels. Another potential application might be to supplement existing Coast Guard or Navy platforms used for fast-boat interdiction with an HPM vessel-stopping capability. A small-vessel test using an HPM source is shown in Figure 3.

Developing solutions for the growing threat that small vessels pose to navigation and safe operations in the world's oceans is one of JNLWD's top priorities. Using nonlethal HPM weapons to stop vessels will provide the warfighter with a viable option for swarm threat and fast-boat interdiction. DEWO is working diligently to accelerate this technology and provide a DE alternative to kinetic weapons and fulfill this long overdue capability gap.

Reference

1. *Department of Defense Dictionary of Military and Associated Terms*, Joint Publication 1-02, 12 April 2001 (as amended through 19 August 2009).

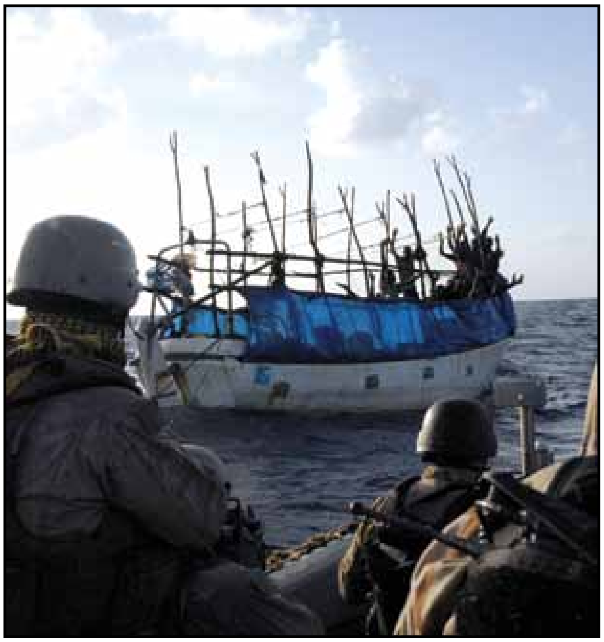

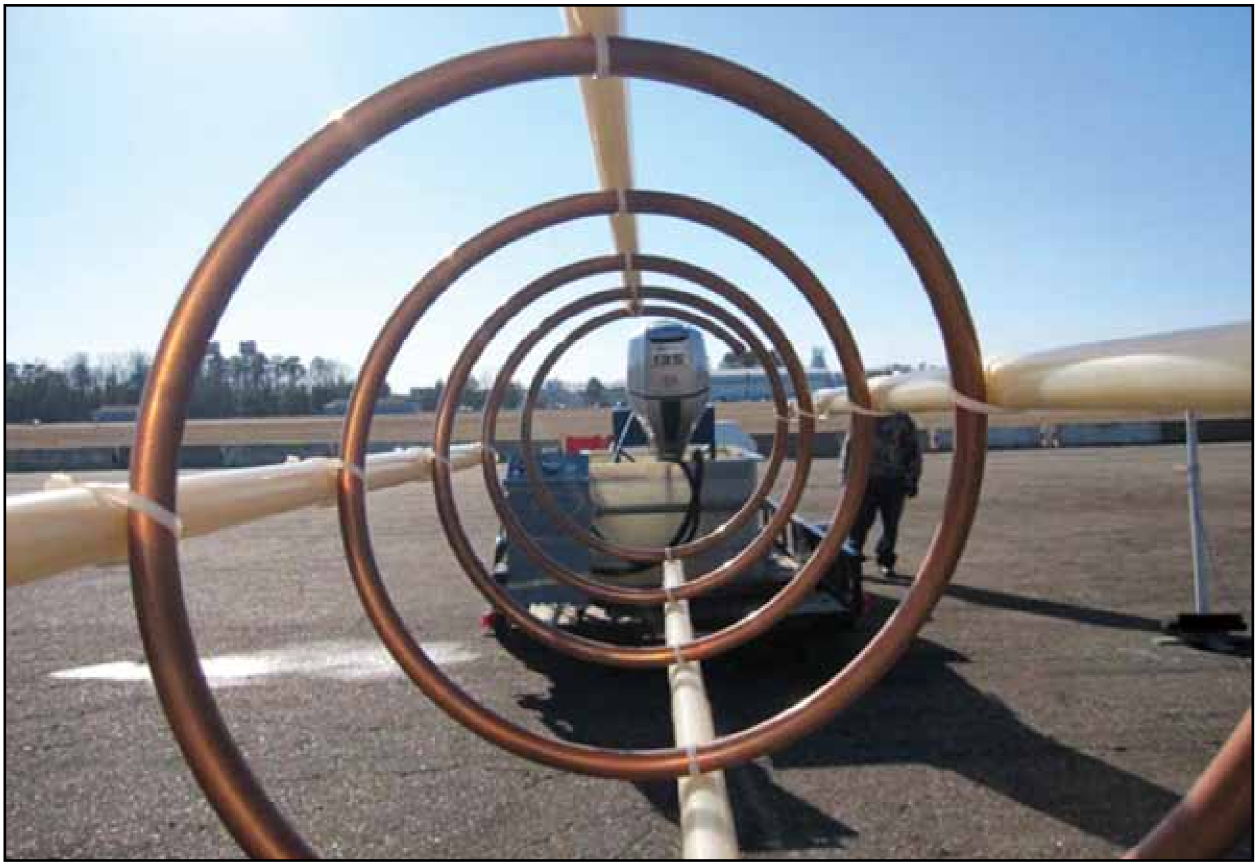

Figure 3. Small-Vessel Testing Using a High-Power Microwave Source

LEADING EDGE
Volume 7, Issue No. 4

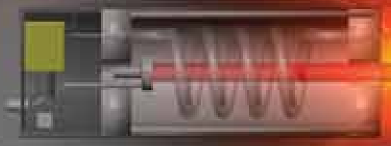

Directed Energy

We look forward as we provide enterprise-wide technical and strategic leadership for the efficient and effective development, acquisition, and fielding of directed-energy systems for the warfighter.

Dale Sisson
Head, Electromagnetic and
Sensor Systems Department
NSWCDD, Dahlgren, Virginia

LEADING EDGE

Volume 7, Issue No. 4

Fallen Warriors

Here we honor those who died while serving their country

NAVSEA
WARFARE CENTERS
DAHLGREN

NSWCDD/MP-09/34

Statement A: Approved for public release; distribution is unlimited

2013

NAWCWD TP 8347
Fourth Edition

Electronic Warfare and Radar Systems

Engineering Handbook

A Comprehensive Handbook for
Electronic Warfare and Radar Systems Engineers

Naval Air Warfare Center Weapons Division, Point Mugu, California

DISTRIBUTION STATEMENT A: Approved for public release.

NAWCWD TP 8347
Fourth Edition

Electronic Warfare and Radar Systems Engineering Handbook

by
Avionics Department

OCTOBER 2013

NAVAL AIR WARFARE CENTER WEAPONS DIVISION
POINT MUGU, CA 93042

DISTRIBUTION STATEMENT A: Approved for public release.

Naval Air Warfare Center Weapons Division

FOREWORD

This handbook is designed to aid electronic warfare and radar systems engineers in making general estimations regarding capabilities of systems. This handbook is sponsored by the NAVAIR Director of Electronic Warfare/Combat Systems Department. This fourth edition updates technical information in Sections 3-7 and 3-8 from previous editions.

This document was reviewed for technical accuracy by Dr. Andrew Chen.

Approved by
DR. R. E. SMILEY, *NAVAIR Director*
EW/Combat Systems Department
30 September 2013

Under authority of
M. MORAN
RDML, U.S. Navy
Commander

Released for publication by
S. O'NEIL
Director for Research and Engineering

NAWCWD Technical Publication 8347

Published by..Technical Communication Office
Collation..Cover, 226 leaves, + Tabs
Previously published as NAWCWPNS TS 92-78 (1st through 3rd Edition) .....4,040 paper
First Edition ....105 paper
Second Edition ....500 paper
Third Edition..1,000 paper
Fourth Edition ....100 paper, 500 electronic

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE ***(DD-MM-YYYY)*** 01-10-2013	**2. REPORT TYPE** Technical publication	**3. DATES COVERED (From - To)** 1 January 2012 – 30 September 2013

4. TITLE AND SUBTITLE

Electronic Warfare and Radar Systems Engineering Handbook (U)

5a. CONTRACT NUMBER N/A

5b. GRANT NUMBER N/A

5c. PROGRAM ELEMENT NUMBER N/A

6. AUTHOR(S)

Avionics Department

5d. PROJECT NUMBER N/A

5e. TASK NUMBER N/A

5f. WORK UNIT NUMBER N/A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Air Warfare Center Weapons Division
575 I Avenue, Suite 1 (Code 450000E)
Point Mugu, California 93042

8. PERFORMING ORGANIZATION REPORT NUMBER

NAWCWD TP 8347

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

N/A

10. SPONSOR/MONITOR'S ACRONYM(S)

N/A

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release.

13. SUPPLEMENTARY NOTES

None.

14. ABSTRACT

(U) This handbook is designed to aid electronic warfare and radar systems engineers in making general estimations regarding capabilities of systems. This handbook is sponsored by the NAVAIR Director of Electronic Warfare / Combat Systems Department.

15. SUBJECT TERMS

Aircraft Dynamics, Antenna, Data Transfer Busses, Electro-Optics and Infrared, Ethernet, Internet Protocol, Maneuverability, Microwave Components, Polarization, Radar Equations, Radiation Patterns, Receiver Characteristics, Receiver Sensitivity, Receiver Tests, Radio Frequency Components, RS-232 Interface, RS-422 Balanced Voltage Interface, RS-485 Interface, Transmission Control Protocol

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Timothy Stolsig
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	SAR	450	**19b. TELEPHONE NUMBER** *(include area code)*
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED			(805) 989-3452

CONTENTS

Section

ABBREVIATIONS AND ACRONYMS 1-1

FUNDAMENTALS

Constants, Conversions, and Characters 2-1
Mathematical Notation 2-2
Frequency Spectrum 2-3
Decibel (dB) 2-4
Duty Cycle 2-5
Doppler Shift 2-6
Electronic Formulas 2-7
Missile and Electronic Equipment Designations 2-8
Radar Horizon / Line of Sight 2-9
Propagation Time / Resolution 2-10
Modulation 2-11
Transforms / Wavelets 2-12

ANTENNAS

Antenna Introduction / Basics 3-1
Polarization 3-2
Radiation Patterns 3-3
Frequency / Phase Effects of Antennas 3-4
Antenna Near Field 3-5
Radiation Hazards 3-6
Active Electronically Scanned Arrays (AESA) 3-7
Fractal Antennas 3-8

RADAR EQUATIONS

Field Intensity and Power Density 4-1
Power Density 4-2
One-Way Radar Equation / RF Propagation 4-3
Two-Way Radar Equation (Monostatic) 4-4
Alternate Two-Way Radar Equation 4-5
Two-Way Radar Equation (Bistatic) 4-6
Jamming to Signal (J/S) Ratio - Constant Power [Saturated] Jamming 4-7
Burn-Through / Crossover Range 4-8
Support Jamming 4-9
Jamming to Signal (J/S) Ratio - Constant Gain (Linear) Jamming 4-10
Radar Cross Section (RCS) 4-11
Emission Control (EMCON) 4-12
EW Jamming Techniques 4-13

RADAR AND RECEIVER CHARACTERISTICS & TEST

RF Atmospheric Absorption / Ducting 5-1
Receiver Sensitivity / Noise 5-2
Receiver Types and Characteristics 5-3
Radar Modes 5-4
General Radar Display Types 5-5
IFF - Identification - Friend or Foe 5-6
Receiver Tests 5-7
Signal Sorting and Direction Finding 5-8

MICROWAVE / RF COMPONENTS

Microwave Waveguides and Coaxial Cable 6-1
Voltage Standing Wave Ratio (VSWR) / Reflection Coefficient
    Return Loss / Mismatch Loss 6-2
Microwave Coaxial Connectors 6-3
Power Dividers and Directional Couplers 6-4
Attenuators / Filters / DC Blocks 6-5
Terminations / Dummy Loads 6-6
Circulators and Diplexers 6-7
Mixers and Frequency Discriminators 6-8
Detectors 6-9
RF / Microwave Amplifiers 6-10
Signal Generation 6-11
Digital Processing Components 6-12
Microwave Measurements 6-13

ELECTRO-OPTICS AND IR 7-1
Introduction 7-1.1
Optical Spectrum 7-1.1
Radiometric Quantities and Terminology 7-1.2
Photometric Quantities 7-1.3
Basic Principles 7-1.4
Basic Radiant Power Relationships 7-1.5
Infrared Source Characteristics 7-1.10
Atmospheric Transmission 7-1.14
EO Components and Sensors 7-1.16
IR Threats to Aircraft and Their Countermeasures 7-1.30
IRCM 7-1.36
Lasers 7-1.38
Fiber Optics 7-1.41
Laser Safety 7-2

AIRCRAFT DYNAMICS CONSIDERATIONS

Free Fall / Aircraft Drag 8-1
Mach Number and Airspeed vs Altitude 8-2
Maneuverability 8-3
EMP / Aircraft Dimensions 8-4

DATA TRANSFER BUSSES

Data Busses 9-1
RS-232 Interface 9-2
RS-422 Balanced Voltage Interface 9-3
RS-485 Interface 9-4
IEEE-488 Interface Bus (HP-IB/GP-IB) 9-5
MIL-STD-1553 & 1773 Data Bus 9-6
Ethernet 9-7
Transmission Control Protocol / Internet Protocol 9-8

GLOSSARY 10-1

ABBREVIATIONS AND ACRONYMS

a	Acceleration or atto (10^{-18} multiplier)
A	Ampere, Area, Altitude, Angstrom (Å), Antenna Aperture, or Aerial
A-799	No evidence of failure report
A/A, A-A, AA	Air-to-Air or Anti-Aircraft
AA-()	Air-to-Air missile number ()
AAA	Anti-Aircraft Artillery
AAAA	Army Aviation Association of America
AAED	Advanced Airborne Expendable Decoy
AAM	Air-to-Air Missile
AARGM	Advanced Anti-Radiation Guided Missile
AAW	Anti-Air Warfare
A-BIT	Automatic Built-in-Test
ABM	Air Breathing Missile or Anti-ballistic Missile
A/C	Aircraft (also acft.)
AC	Alternating Current
ACA	Associate Contractor Agreement or Airspace Coordination Area
ACAT	Acquisition Category
ACC	Air Combat Command
ACCB	Aircraft Configuration Control Board
Acft	Aircraft (also A/C)
ACLS	Aircraft Carrier Landing System
ACM	Advanced Cruise Missile or Air Combat Maneuvering
ACQ	Acquisition
ACS	Antenna Coupler Set
ACTD	Advanced Concept Technology Demonstration
A/D	Analog to Digital
ADM	Advanced Development Model
ADP	Automatic Data Processing or Advanced Development Program
AEA	Airborne Electronic Attack
AEC	Aviation Electronic Combat (Army)
AEGIS	Automatic Electronic Guided Intercept System
AEL	Accessible Emission Limit
AESA	Active Electronically Scanned Array
AEW	Airborne Early Warning
AF	Antenna Factor, Air Force, or Audio Frequency
AFB	Air Force Base or Airframe Bulletin
AFC	Automatic Frequency Control or Airframe Change
AFIPS	Automated Financial Information Processing System

AFOTEC	Air Force Operational T&E Center
A/G	Air-to-Ground
AGB	Autonomous Guided Bomb
AGC	Automatic Gain Control
AGI	Auxiliary General Intelligence (Intelligence-gathering Ship)
AGL	Above Ground Level
AGM	Air-to-Ground Missile
AGS	Angle Gate Stealer
AHWS	Advanced Helicopter Weapons System
AI	Artificial Intelligence, Air Intercept, or Airborne Interceptor
AIAA	American Institute of Aeronautics and Astronautics
AIC	Air Intercept Control
AIM	Air Intercept Missile
AIRLANT	Commander, U.S. Naval Air Forces, Atlantic Fleet
AIRPAC	Commander, U.S. Naval Air Forces, Pacific Fleet
AJ	Anti-jamming or Anti-Jam
A-Kit	Aircraft wiring kit for a system (includes cabling, racks, etc. excluding WRAs)
ALC	Air Logistics Center
AM	Amplitude Modulation
AMD	Aircraft Maintenance Department
AMES	Advanced Multiple Environment Simulator
AMLV	Advanced Memory Loader/Verifier
Amp	Amplifier
AMRAAM	Advanced, Medium-Range, Air-to-Air Missile
ANSI	American National Standards Institute
ANT	Antenna
Ao	Operational Availability
AO	Acousto-Optical
AOA	Angle of Arrival, Angle of Attack, or Analysis of Alternatives (similar to COEA)
AOC	Association of Old Crows (Professional EW Society) or Award of Contract
AOT	Angle Only Track, Angle Off Tail, or Acquisition-on-Target
APC	Amphenol Precision Connector or Armored Personnel Carrier
APN	Aircraft Procurement, Navy
APO	Armed Forces (or Army or Air) Post Office, Acquisition Program Office

APU	Auxiliary Power Unit
AR	Anti-reflection or Aspect Ratio
ARM	Anti-radiation Missile
ARO	After Receipt of Order
A/S, A-S, AS	Air-to-Surface
ASC	Air Systems Command
ASCM	Anti-ship Cruise Missile
ASE	Aircraft Survivability (or Survival) Equipment, Allowable Steering Error, or Automatic Support Equipment
ASIC	Application Specific Integrated Circuit
ASK	Amplitude Shift Keying
ASM	Air-to-Surface Missile
ASO	Aviation Supply Office
A-Spec	System Specification
ASPJ	Airborne Self-Protection Jammer
ASPO	Avionics Support (also Systems) Project Office (also Officer)
ASR	Advanced Special Receiver or Airport/Airborne Surveillance Radar
ASRAAM	Advanced Short Range Air-to-Air Missile
ASTE	Advanced Strategic and Tactical Expendables
ASW	Anti-submarine Warfare
ATA	Advanced Tactical Aircraft
ATARS	Advanced Tactical Air Reconnaissance System
ATC	Air Traffic Control
ATD	Advanced Technology Demonstration
ATE	Automatic Test Equipment
ATEDS	Advanced Technology Expendables and Dispenser Systems
ATF	Advanced Tactical Fighter (F-22)
ATIMS	Airborne Turret Infrared Measurement System or Airborne Tactical Information Management System
ATIRCM	Advanced Threat Infrared Countermeasures
ATP	Acceptance Test Procedure
ATR	Autonomous Target Recognition, Airborne Transportable Rack, Atlantic Test Range
ATRJ	Advanced Threat Radar Jammer
AUTODIN	Automatic Digital Network
AUX	Auxiliary
avdp.	Avoirdupois (system of measures)
Avg	Average
AWACS	Airborne Warning and Control System
AZ	Azimuth (also Az)

B	Bandwidth (also BW) or Magnetic inductance
BAFO	Best and Final Offer
BAU	Bus Adapter Unit
BC	Bus Controller
BDA	Battle Damage Assessment
BDI	Battle Damage Indication
BFO	Beat Frequency Oscillator
BI	Background Investigation
BIFF	Battlefield Identification, Friend, or Foe
BIT	Built-in-Test, Binary Digit or Battlefield Information Technology
BITE	Built-in-Test Equipment
BIU	Bus Interface Unit
B-Kit	Avionics "Black Box" WRAs
B/N	Bombardier/Navigator
BNC	Bayonet Navy Connector
BOA	Basic Ordering Agreement
BOL	Swedish chaff dispenser in a launcher
BPF	Band Pass Filter
BPS	Bits Per Second
BUMED	Bureau of Medicine (Navy)
BUNO	Bureau Number (aircraft)
BUR	Bottom Up Review
BVR	Beyond Visual Range
BW	Beamwidth (referring to an antenna) or sometimes Bandwidth
BWA	Backward Wave Amplifier
BWO	Backward Wave Oscillator

c	Speed of Light = $3\text{x}10^8$ meters/sec = $1.8\text{x}10^{12}$ furlongs per fortnight or 1.8 terafurlongs per fortnight, or centi (10^{-2}) multiplier
C	Electron Charge, Coulomb, Capacitance, Celsius, Centigrade, Confidential, Roman numeral for 100, or a programming language (also C+ and C++)
C^2	Command and Control
C^3 (C^4)	Command, Control, Communications (and Computers)
C^3I (C^4I)	Command, Control, Communications, (Computers) and Intelligence
CAD	Computer-Aided Design
CAE	Computer-Aided Engineering
CAG	Carrier Air Group
CAGE	Commercial and Government Entry
CAIV	Cost as an Independent Variable
CAL	Calibration

CAM	Computer-Aided Manufacturing or Constant Addressable Memory
CAO	Competency Aligned Organization or Contract Administrative Officer
CAP	Combat Air Patrol
CAS	Close Air Support or Calibrated Airspeed
CASS	Consolidated Automated Support System
CAT	Catapult or Cockpit Automation Technology
CB	Citizens Band (also see Seabee)
CBD	Commerce Business Daily
CBIT	Continuous Built-in-Test
CBO	Congressional Budget Office
CCA	Circuit Card Assembly
CCB	Configuration Control Board
CCD	Charge Coupled Device
CCM	Counter-Countermeasures
CCN	Contract Change Number or Configuration Change Notice
CCU	Cockpit Control Unit
cd	Candela (SI unit of luminous intensity)
CD	Compact Disk or Control and Display
CDC	Combat Direction Center
CDR	Critical Design Review
CDRL	Contract Data Requirements List
CE	Conducted Emission
CECOM	Communications and Electronics Command (Army)
CEESIM	Combat Electromagnetic Environment Simulator
CEP	Circular Error Probability
CFA	Cognizant Field Activity
CFAR	Constant False Alarm Rate
CFE	Contractor Furnished Equipment
CG	Center of Gravity, Commanding General, Command Guidance, or Cruiser
CI	Configuration Item
CIA	Central Intelligence Agency
CIC	Combat Information Center (now called CDC)
CID	Combat Identification or Charge Injection Device
CILOP	Conversion in Lieu of Procurement
CINC	Commander in Chief
CIP	Capital Improvement Program
CIS	Commonwealth of Independent States (11 of 15 former Soviet Union territories except Estonia, Georgia, Latvia, and Lithuania)
CIWS	Close-In Weapon System
CJ	Coherent Jamming
CLC	Command Launch Computer
cm	Centimeter
CM	Countermeasures or Configuration Management
CMC	Command Mission Computer or Commandant Marine Corps
CMDS	Countermeasure Dispensing System
CMOS	Complementary Metal-Oxide Semiconductor
CMP	Configuration Management Plan
CMWS	Common Missile Warning System
CNAF	Commander, Naval Air Forces
CNAL	Commander, Naval Air Forces Atlantic (also COMNAVAIRLANT)
CNAP	Commander, Naval Air Forces Pacific (also COMNAVAIRPAC)
CNI	Communications, Navigation, and Identification
CO	Commanding Officer, Contracting Officer, Change Order, or Carbon Monoxide
COB	Close of Business
COCOM	Combatant Command
COEA	Cost and Operational Effectiveness Analysis
COG	Center of Gravity or Cognizant
COMM	Communications
COMSEC	Communications Security
CONSCAN	Conical Scanning Radar
CONUS	Continental United States
CO-OP	Cooperative (countermeasures)
COR	Contracting Officers Representative
CORPORAL	Collaborative On-Line Reconnaissance Provider/Operationally Responsive Attack Link
Cos	Cosine
COSRO	Conical-Scan on Receive Only
COTS	Commercial Off-The-Shelf (hardware/software)
CP	Circularly Polarized (antenna), Central Processor, or Command Post
CPS	Computer or Control Power Supply (depends on application)
CPU	Central Processing Unit
CRC	Originally Chemical Rubber Company, now published reference books by CRC Press
CRFM	Coherent RF Memory
CRISD	Computer Resources Integrated Support Document
CRLCMP	Computer Resources Life Cycle Management Plan

CRO	Countermeasures Response Optimization
CRT	Cathode Ray Tube or Combat Rated Thrust (afterburner)
CSAR	Combat Search and Rescue
Crypto	Cryptographic
CS	Conducted Susceptibility
CSC	Commodity Software Change
CSCI	Computer Software Configuration Item
C-Spec	Product Specification
CSS	Contractor Support Services
CTR	Chesapeake Test Range
CV	Aircraft Carrier
CVN	Nuclear Powered Aircraft Carrier
CVR	Crystal Video Receiver
CW	Continuous Wave or Chemical Warfare
CWBS	Contract Work Breakdown Structure
CWI	Continuous Wave Illuminator
CY	Calendar Year
d	Distance, Diameter, or deci (10^{-1} multiplier)
D	Distance, Diameter, Electron displacement, Detectivity, Doppler, Density, or Roman numeral for 500
da	deca (10^0 multiplier)
D/A	Digital-to-Analog
DAB	Defense Acquisition Board
DAC	Digital to Analog Converter or Dept of Army Civilian
DAR	Defense Acquisition Regulation
DARPA	Defense Advanced Research Projects Agency
DB	Database
dB	Decibel
dBc	dB referenced to the Carrier Signal
dBi	Decibel antenna gain referenced to an isotropic antenna
dBm	Decibel referenced to the power of one milliwatt
DBOF	Defense Business Operations Fund
dBsm	Decibel value of radar cross section referenced to a square meter
dBW	Decibel referenced to the power of one watt
DC	Direct Current, Discrete Circuit, or District of Columbia
DCE	Data Communication Equipment
DCS	Direct Commercial Sales or Distributed Control System
DDI	Digital Display Indicator
DDS	Direct Digital Synthesizers
DECM	Deceptive Electronic Countermeasures (also Defensive ECM)
deg	Degree
DEMVAL	Demonstration Validation (also DEM/VAL)
DET	Detachment
DF	Direction Finding
DFT	Discrete Fourier Transform
DI	Data Item
DIA	Defense Intelligence Agency or Diameter
DID	Data Item Description
DIRCM	Directed Infrared Countermeasures
DJ	Deceptive Jamming
D-Level	Depot Level Maintenance
DM	Data Management (also manager)
DMA	Direct Memory Address or Defense Mapping Agency
DME	Distance Measuring Equipment
DNA	Defense Nuclear Agency, Does Not Apply, or Deoxyribonucleic Acid
DOA	Direction of Arrival
DOD or DoD	Department of Defense
DoDISS	DoD Index of Specifications and Standards
DOM	Depth of Modulation
DON	Department of the Navy
DOS	Disk Operating System
DOT&E	Director, Operational Test & Evaluation
DPRO	Defense Plant Representative Office
DRB	Defense Review Board
DRFM	Digital RF Memory
DSARC	Defense Systems Acquisition (and) Review Council
DSN	Defense Switching Network
DSO	Dielectrically Stabilized Oscillator
DSP	Digital Signal Processor
D-Spec	Process Specification
DT (&E)	Development or Developmental Test (and Evaluation)
DTC	Design to Cost
DTE	Data Terminal Equipment
DTO	Digitally Tuned Oscillator or Defense Technology Objectives
DTRMC	Defense (or DoD) Test Recourse Management Center

e	Electron charge or base of natural logarithms (2.71828...)
E	Electric Field Intensity or Strength, Energy, East, or Exa (10^{18} multiplier)
E^3	Electromagnetic Environmental Effects
EA	Electronic Attack (similar to older term of ECM)
EC	Electronic Combat
ECAC	Electromagnetic Compatibility Analysis Center (DOD), now Joint Spectrum Center
ECCM	Electronic Counter-Countermeasures (similar to newer term of EP)
ECL	Emitter Coupled Logic
ECM	Electronic Countermeasures (similar to newer term of EA)
ECN	Engineering Change Notice
ECO	Engineering Change Order
ECP	Engineering Change Proposal or Egress Control Point
ECR	Electronic Combat Range (China Lake) or Electronic Combat & Reconnaissance
ECS	Environmental Control System
ECSEL	Electronic Combat Simulation and Evaluation Laboratory
ECU	Electronic Control Unit
EDM	Engineering Development Model
EED	Electro-Explosive Device
EEPROM	Electrically Erasable/Programmable Read-only Memory
EHF	Extremely High Frequency [30 to 300 GHz]
EIA	Electronic Industries Associates
EID	Emitter Identification Data
EIRP	Effective Isotropic Radiated power
EL	Elevation (also El)
ELF	Extremely Low Frequency [3 Hz to 3 kHz]
ELINT	Electronics Intelligence
ELNOT	Emitter Library Notation
EM	Electromagnetic
E-Mail	Electronic Mail
EMC	Electromagnetic Compatibility
EMCAB	EMC Advisory Board
EMCON	Emission Control
EMD	Engineering and Manufacturing Development
EME	Electromagnetic Environment
EMI	Electromagnetic Interference
EMP	Electromagnetic Pulse
EMR	Electromagnetic Radiation
EMS	Electromagnetic Susceptibility, Electromagnetic Spectrum
EMV	Electromagnetic Vulnerability
EO	Electro-Optic, Electro-Optical, or Engineering Order
EOB	Electronic Order of Battle or Expense Operating Budget
EOCM	Electro-Optic Countermeasures
EOF	Electro-Optical Frequency (300 to 3 x 10^7 GHz)
EP	Electronic Protection (similar to older terms of ECCM and DECM)
EPA	Environmental Protection Agency
EPROM	Electrically Programmable Read-only Memory
ERAM	Electronic Counter-Countermeasures (also Protection) Requirements and Assessment Manual
ERP	Effective Radiated Power
ES	Electronic Surveillance (similar to older term of ESM)
ESD	Electrostatic Discharge
ESM	Electronic Support Measures (similar to newer term of ES)
ESSM	Evolved Sea Sparrow Missile
ET	Electronics Technician
ETI	Elapsed Time Indicator
ETIRMS	EW Tactical Information and Report Management System
ETR	Estimated Time to Repair
EVM	Earned Value Management
EW	Electronic Warfare , Early Warning, or Expeditionary Warfare
EWBM	EW Battle Management
EWDS	EW Data Systems
EWIA	EW Intelligence Analysis
EWIR	Electronic Warfare Integration & Reprogramming (USAF database)
EWMP	Electronic Warfare Master Plan
EWO	Electronic Warfare Officer
EWRL	Electronic Warfare Reprogrammable Library (USN)
EWSI	EW Systems Integration
EWSSA	EW Software Support Activity
EXP	Expendable Countermeasure

f	femto (10^{-15} multiplier), Frequency (also F), or lens *f* number
F	Frequency (also *f*), Force, Farad, Faraday Constant, Female, Fahrenheit, Noise Figure, Noise Factor or “Friendly” on RWR display

F2T2EA	Find, Fix, Track, Target, Engage, Assess (targeting of hostile forces)
F/A	Fighter/Attack
FAA	Federal Aviation Administration
FAC	Forward Air Controller
FAR	Federal Acquisition Regulations or False Alarm Rate
f_c	Footcandle (unit of illuminance)
FCA	Functional Configuration Audit
FCR	Fire Control Radar
FDR	Frequency Domain Reflectometry
FEBA	Forward Edge of the Battle Area
FET	Field-Effect Transistor
FEWC	Fleet EW Center
FFT	Fast Fourier Transform
FIFO	First In / First Out
FIPR	Federal Information Processing Resources
fl	fluid
FLAK	AAA Shrapnel, from the German “Flieger Abwher Kanone” (AAA gun that fires fast and furiously)
FLIR	Forward Looking Infrared
FLPS	Flightline Payload Simulator
FLT	Flight
FM	Frequency Modulation or Failure Mode
FME	Foreign Material Exploitation
FMEA	Failure Mode and Effects Analysis
FMS	Foreign Military Sale(s)
FOC	Full Operational Capability
FOD	Foreign Object Damage
FORCECAP	Force Combat Air Patrol
FOT&E	Follow-On Test and Evaluation
FOTD	Fiber Optic Towed Device
FOUO	For Official Use Only
FOV	Field of View
FPA	Focal Plane Array
fps	feet per second
FRACAS	Failure, Reporting, Analysis, and Corrective Actions System
FRB	Failure Review Board
FRD	Functional Requirements Document
FSD	Full Scale Development
FSED	Full Scale Engineering Development
FSK	Frequency Shift Keying
FSU	Former Soviet Union
ft	Feet or Foot
FTC	Fast Time Constant
FTD	Foreign Technology Division (USAF)
FWD	Forward
FY	Fiscal Year

g	Gravity (also G)
G	Universal Gravitational Constant (also K), Giga (10^9 multiplier), Conductance, or Gain
G&A	General and Administrative (expense)
GaAs	Gallium Arsenide
GACIAC	Guidance and Control Information Analysis Center (DoD)
gal	Gallon
GAO	General Accounting Office
GBU	Guided Bomb Unit
GCA	Ground Controlled Approach
GCI	Ground Control Intercept
GENSER	General Service
GEN-X	Generic Expendable
GFE	Government Furnished Equipment
GHz	GigaHertz
GI	Government Issue
GIDEP	Government Industry Data Exchange Program
GIG	Global Information Grid
GIGO	Garbage In / Garbage Out
GOCO	Government Owned Contract Operated
GOFO	General Officer / Flag Officer
GP	General Purpose
GPI	Ground Plane Interference
GPIB	General Purpose Interface Bus
GPS	Global Positioning System
GSE	Ground Support Equipment

h	hours, hecto (10^2 multiplier), Plank’s constant, or height (also H)
H	Height (also h), Henry (Inductance), or Irradiance
HARM	High-speed Anti-Radiation Missile
HAWK	Homing All the Way Killer
HDBK	Handbook
HDF	High Duty Factor
HE	High Explosive
HEF	High Energy Frequency ($3x10^7$ to $3x10^{14}$ GHz)
HEL	High Energy Laser
HELO	Helicopter
HERF	Hazards of Electromagnetic Radiation to Fuel
HERO	Hazards of Electromagnetic Radiation to Ordnance
HERP	Hazards of Electromagnetic Radiation to Personnel
hex	hexadecimal
HF	High Frequency [3 - 30 MHz]
HIL or HITL	Hardware-in-the-Loop

HOJ	Home-On-Jam
HOL	Higher Order Language
HPF	High-Pass Filter
HP-IB	Hewlett-Packard Interface Bus
HP-IL	Hewlett-Packard Interface Loop
HPM	High Powered Microwave
HPRF	High Pulse Repetition Frequency
hr	hour
HSDB	High Speed Data Bus
HUD	Heads-Up Display
HV	High Voltage
H/W	Hardware
HWCI	Hardware Configuration Item
HWIL	Hardware-in-the-loop
Hz	Hertz (Cycles per second)
i	current (also I)
I	Current (also i), Intensity, Irradiance, Intermediate, or Roman Numeral for One
IA	Information Assurance
IADS	Integrated Air Defense System
I&Q	In-Phase and Quadrature
IAS	Indicated Airspeed
IAW	In Accordance With
IBIT	Initiated Built-in-Test
IBU	Interference Blanker Unit
IC	Integrated Circuit
ICD	Interface Control Document Initial Capabilities Document
ICMD	Improved Countermeasure Dispenser
ICNIA	Integrated Communication, Navigation, Identification Avionics
ICS	Inverse Conical Scan or Intercommunications System (aircraft)
ICW	In Compliance With
ID	Identification
IDA	Institute For Defense Analysis
IDECM	Integrated Defensive Electronic Countermeasures
IEEE	Institute of Electrical and Electronic Engineers
IF	Intermediate Frequency
IFF	Identification Friend-or-Foe
IFM	Instantaneous Frequency Measurement
IFR	Instrument Flight Rules
IG	Inspector General
IIR	Imaging Infrared
I-Level	Intermediate Level of Repair (also "I" Level)
ILS	Integrated Logistic Support, Instrument Landing System, or Inertial Locator System
ILSMT	Integrated Logistic Support Management Team
IM	Intermodulation or Item Manager
IMA	Intermediate Maintenance Activity
in	Inch
INEWS	Integrated Electronic Warfare System
INS	Inertial Navigation System
INT	Intensity
IO	Information Operations
I/O	Input/Output
IOC	Initial Operational (also Operating) Capability
IOT&E	Initial Operational Test and Evaluation
IPO	International Projects (Program) Office
IPR	In-Progress/Process Review
IPT	Integrated Product (also Program) Team
IR	Infrared
IR&D	Independent Research and Development
IRCM	Infrared Countermeasures
IRDS	Infrared Detecting System
IREXP	IR Expendables
IRIG-B	Inter-range Instrumentation Group B
IRLS	Infrared Line Scanner
IRS	Interface Requirements Specification, IR Suppression or Internal Revenue Service
IRST	Infrared Search and Track
ISAR	Inverse Synthetic Aperture Radar
ISO	Derived from the Greek "isos" meaning "equal," the official title is International Organization for Standardization
ISP	Integrated Support Plan Intelligence Support Plan
ISR	Interference to Signal Ratio (also I/S)
ITU	International Telecommunications Union
IV&V	Independent Validation and Verification
IW	Information Warfare

J	Jamming, Radiance, Current Density, or Joules
JAAS	Joint Architecture for Aircraft Survivability
JAFF	Jammer (illuminating) Chaff
JAG	Judge Advocate General
JAMS	Jamming Analysis Measurement System
JARS	Jamming Aircraft & Radar Simulation
JASSM	Joint Air-to-Surface Standoff Missile
JAST	Joint Advanced Strike Technology
JATO	Jet Assisted Takeoff or JAmmer Technique Optimization
JC2WC	Joint Command and Control Warfare Center (now JIOWC)
JCTD	Joint Concept Technology Demonstration
JCS	Joint Chiefs of Staff or Joint Spectrum Center (formerly ECAC)
JDAM	Joint Direct Attack Munition
JEACO	Joint Electronic Attack and Compatibility Office
JED	Journal of Electronic Defense (Published by the Association of Old Crows)
JEM	Jet Engine Modulation
JETS	Joint Emitter Targeting System
JEWC	Joint EW Conference or Joint EW Center (then JC2WC & now JIOWC)
JEWEL	Joint Electronic Warfare Effects Laboratory
JIOWC	Joint Information Operations Warfare Command
JMEM	Joint Munitions Effectiveness Manual
JMR	Jammer
JOVIAL	Julius' Own Version of International Algorithmic Language (Air Force computer programming language)
JPATS	Joint Primary Aircraft Training System
J/S	Jamming to Signal Ratio
JSF	Joint Strike Fighter
JSGCC	Joint Services Guidance and Control Committee
JSIR	Joint Spectrum Interference Resolution (signal interference portion of MIJI)
JSOW	Joint Stand-Off Weapon (AGM-154A)
JSTARS	Joint Surveillance Target Attack Radar System
JTA	Jammer Threat Analysis
JTAT	JATO Techniques Analysis and Tactics
JTCG/AS	Joint Technical Coordinating Group for Aircraft Survivability
JTCG/ME	Joint Technical Coordinating Group for Munitions Effectiveness
JTIDS	Joint Tactical Information Distribution System
JV or J/V	Joint Venture

k	kilo (10^3 multiplier) or Boltzmann Constant
K	Kelvin, Cathode, Universal gravitational constant (also G), or Luminous efficacy
KCAS	Knots Calibrated Airspeed
kg	kilogram
kHz	KiloHertz
KIA	Killed in Action
KIAS	Knots Indicated Air Speed
km	Kilometer
KSLOC	Thousand Source Lines of Code (software)
kt	Knot (nautical miles per hour)
kW	Kilowatt

l	length (also L) or liter
L	Length (also l), Loss, inductance, Luminance, or Roman Numeral for fifty
LADAR	Laser Detection and Ranging (i.e., laser radar)
LAN	Local Area Network
LANTIRN	Low Altitude Navigation & Targeting Infrared for Night
LASER	Light Amplification by Stimulated Emission of Radiation
LAT	Latitude (0-90° N or S from equator)
lbs	pounds
LCC	Life Cycle Cost(s)
LCD	Liquid Crystal Display or Lowest Common Denominator
LCP or LHCP	Left-hand Circular Polarization
LDF	Low Duty Factor
LDS	Laser Detecting Set
LED	Light-Emitting Diode
LEX	Leading Edge Extension
LGB	Laser Guided Bomb
LF	Low Frequency [30 - 300 kHz]
LIC	Low Intensity Combat or Laser Intercept Capability

LISP	List Processing (A programming language used in artificial intelligence)
LLL	Low Light Level (as in LLL TV)
lm	lumen (SI unit of luminous flux)
ln	Natural Logarithm
LO	Local Oscillator or Low Observable
LOA	Letter of Agreement (or Acceptance)
LOB	Line of Bearing (see also AOA)
LOG	Logarithm to the base 10 (also log) or Logistician
LONG	Longitude (0-180° E or W from Greenwich, U.K.)
LOR	Level of Repair
LORA	Level of Repair Analysis
LORAN	Long Range Navigation
LORO	Lobe on Receive Only
LOS	Line-of-Sight
LPAR	Large Phased-Array Radar
LPD	Low Probability of Detection
LPF	Low Pass Filter
LPI or LPOI	Low Probability of Intercept
LPRF	Low Pulse Repetition Frequency
LR	Lethal Range
LRA	Line Replaceable Assembly
LRF	Laser Rangefinder
LRIP	Low Rate Initial Production
LRU	Line Replaceable Unit
LSA	Logistic Support Analysis
LSAR	Logistic Support Analysis Record
LSB	Least Significant Bit
LSI	Large Scale Integration
LSO	Landing Signal Officer
LSSO	Laser System Safety Officer
LTBB	Look Through Blanking Bus
LWIR	Long Wave Infrared
LWR	Laser Warning Receiver
lx	Lux (SI unit of illuminance)
LZ	Landing Zone

m	milli (10^{-3} multiplier), meter, or electron mass
M	Mega (10^6 multiplier), Male, Mach number, or Roman numeral for 1,000
MA	Missile Alert or Missile Active
MAD	Magnetic Anomaly Detection (also Detector)
MADD	Microwave Acoustic Delay Device
MAF	Maintenance Action Form
MAG	Marine Aircraft Group or Magnetic
MAGTF	Marine Air-Ground Task Force
MANPADS	Man-portable Air Defense System
M&S	Modeling and Simulation
MASER	Microwave Amplification by Simulated Emission of Radiation
MATE	Modular Automatic Test Equipment
MAW	Missile Approach Warning system (also MAWS) or Marine Aircraft Wing
MAX	Maximum or Maximum aircraft power (afterburner)
MBFN	Multiple Beam Forming Network
MC	Mission Computer
MCIOC	Marine Corps Information Operations Center
MCP	Micro-Channel Plate
MDF	Mission Data File
MDI	Multiple Display Indicator or Miss Distance Indicator
MDG	Mission Data Generator
MDS	Minimum Discernible Signal or Minimum Detectable Signal
MDU	Multipurpose Display Unit
MF	Medium Frequency (300 kHz to 3 MHz)
MFD	Multifunction (video) Display
MG	Missile Guidance
MHz	MegaHertz (10^6 Hz)
MIA	Missing in Action
MIC	Microwave Integrated Circuit or Management Information Center
MICRON	10^{-6} meter
MiG	Mikoyan-Gurevich (Soviet aircraft manufacturer)
MIGCAP	MiG Combat Air Patrol
MIJI	Meaconing, Intrusion, Jamming, & Interference (also see JSIR)
mil	One-thousandth of an inch
MIL	Military power (100%, no afterburner) or Military
MILCON	Military Construction
MILSPEC	Military Specification
MILSTRIP	Military Standard Requisitioning and Issue Procedure(s)
MIMIC	Microwave Monolithic Integrated Circuit (also MMIC)
MIN	Minimum
Mincon	Minimal Construction
MIPPLE	RWR display switching between ambiguous emitters
MIPS	Millions of (Mega) Instructions Per Second
ML	Missile Launch
MLC	Main Lobe Clutter
MLV	Memory Loader Verifier
MLVS	Memory Loader Verifier Set

mm	Millimeter
MM	Man Month
MMIC	Microwave Monolithic Integrated Circuit (also MIMIC)
MMW	Millimeter Wave (40 GHz or higher per IEEE, but commonly used down to 30 GHz)
MOA	Memorandum of Agreement
MOE	Measure of Effectiveness
MOM	Methods of Moments (also MoM) or Metal-Oxide-Metal
MOP	Modulation on Pulse or Measure of Performance
MOPS	Million Operations Per Second
MOS	Minimum Operational Sensitivity, Military Occupational Specialty, Metal-Oxide Semiconductor, or Measure of Suitability
MOSAIC	Modeling System for Advanced Investigation of Countermeasures
MOU	Memorandum of Understanding
MPD	Multi-Purpose Display or Microwave Power Device
MPE	Maximum Permissible Exposure
mph	Miles per Hour
MPLC	Multi-Platform Launch Controller
MPM	Microwave Power Module
MPPS	Million Pulses Per Second
MPRF	Medium Pulse Repetition Frequency
mr or mrad	Milliradian
MRC	Maintenance Requirement Card or Medium Range CAP
MRE’s	Meals Ready to Eat
ms	Milliseconds
MSB	Most Significant Bit
MSI	Multi-Sensor (also Source) Integration, Management Support Issues, or Medium Scale Integration
MSIC	Missile and Space Intelligence Center
MSL	Mean Sea Level (altitude) or Missile
MTBF	Mean Time Between Failures
MTI	Moving Target Indicator (or Indication)
MTTR	Mean Time To Repair
MUXBUS	Multiplex Bus
MVS	Minimum Visible Signal
mw	Microwave
mW	Milliwatt
MWIR	Mid Wave Infrared
MWS	Missile Warning Set
MY	Man Year

n	nano (10^{-9} multiplier) or number of elements
N	Noise, Newton (force), Radiance, North, or No
n/a	Not Applicable (also N/A)
NA	Numerical Aperture
NADEP	Naval Aviation Depot
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NATOPS	Naval Air Training and Operating Procedures Standardization
NAV	Navigation
NAVAIR	Naval Air Systems Command (also NAVAIRSYSCOM)
NavMPS	Naval Mission Planning System
NAVSEA	Naval Sea Systems Command (also NAVSEASYSCOM)
NAWCAD	Naval Air Warfare Center Aircraft Division
NAWCWD	Naval Air Warfare Center Weapons Division
NBC	Nuclear, Biological, Chemical
NCTR	Non-Cooperative Target Recognition
NDI	Non-Developmental Item or Non Destructive Inspection
NEI	Noise Equivalent Power
NEMP	Nuclear Electromagnetic Pulse
NEOF	No Evidence of Failure
NEP	Noise Equivalent Power
NF	Noise Figure or Noise Factor (also F)
NFO	Naval Flight Officer
NGJ	Next Generation Jammer
NIOC	Navy Information Operations Command
NIPO	Navy International Program Office
NIR	Near Infrared
nm	nanometer or Nautical Mile (also NM or NMI)
NM or NMI	Nautical Mile (also nm)
NMCI	Navy Marine Corps Intranet
NNWC	Naval Network Warfare Command
NOHD	Nominal Ocular Hazard Distance
NORAD	North American Air Defense Command
NPG or NPGS	Naval Post Graduate School
NRE	Non-Recurring Engineering
NRL	Naval Research Laboratory
NRZ	Non Return to Zero
NSA	National Security Agency
nsec or ns	Nanosecond
NSN	National Stock Number
NSWC	Naval Surface Weapons Center
nt	Nit (SI unit of luminance)

NUWC	Naval Undersea Warfare Center
NVG	Night Vision Goggles
NWIP	Naval Warfare Information Publication
NWP	Naval Warfare Publication
O	Optical
OADR	Originating Agency's Determination Required
OAG	Operational Advisory Group
O&MN	Operations and Maintenance, Navy (also O&M,N)
OBE	Overtaken (Overcome) By Events
OCA	Offensive Counter Air
OEWTPS	Organizational Electronic Warfare Test Program Set
OFP	Operational Flight Program
OJT	On-the-Job Training
O-Level	Organizational Level of Repair (also "O" Level)
OMA	Organizational Maintenance Activity
OMB	Office of Management and Budget
OMEGA	Optimized Method for Estimating Guidance Accuracy (VLF Navigation System)
ONR	Office of Naval Research
OOK	On-Off Keying
OPEVAL	Operational Evaluation
OPM	Office of Personnel Management
OPSEC	Operational Security
OPTEVFOR	Operational Test and Evaluation Force
OR	Operational Requirement or Operationally Ready
ORD	Operational Requirements Document
OSD	Office of the Secretary of Defense
OSHA	Occupational Safety and Health Act
OSIP	Operational Safety Improvement Program
OSM	Operating System Memory or SMA connector made by Omni-Spectra
OSRB	Operational Software Review Board
OT (&E)	Operational Test (and Evaluation)
OTD	Operational Test Director
OTH	Over the Horizon
OTH-B	Over-the-Horizon Backscatter
OTH-R	Over-the-Horizon Radar
OTH-T	Over-the-Horizon Targeting
OTRR	Operational Test Readiness Review
OUSD	Office of the Under Secretary of Defense
oz	ounce

p	pico (10^{-12} multiplier) or page
P	Power, Pressure, or Peta (10^{15} multiplier)
P^3I	Pre-Planned Product Improvement
Pa	Pascal (pressure)
PA	Public Address or Program Analyst
PBIT	Periodic Built-in-Test
PC	Pulse Compression, Personal Computer, or Photoconductive
PCA	Physical Configuration Audit
PCM	Pulse Code Modulation
P_d	Probability of Detection
PD	Pulse Doppler
PDI	PD Illuminator or Post Detection Integration
PDP	Plasma Display Panel
PDQ	Pretty Darn [sic] Quick
PDR	Preliminary Design Review
PDW	Pulse Descriptor Word
PEL	Personnel Exposure Limits
PEM	Photoelectromagnetic
PEO	Program Executive Officer
pf	Power Factor or Pico Farads
PFA	Probability of False Alarm
PGM	Precision Guided Munition
ph	Phot (unit of illuminance)
P_h	Probability of Hit
pi	Greek letter π
P_i	Probability of Intercept (also POI)
PID	Positive Identification
PIN	Personal Identification Number
PIP	Product Improvement Plan or Predicted Intercept Point
Pixel	Picture Element
P_k	Probability of Kill or Peak
PLSS	Precision Location Strike System
PM	Phase Modulation or Program Manager
PMA	Program (also Project) Manager, Air
PMAWS	Passive Missile Approach Warning System
PMS	Program Manager, Ship
PMT	Photomultiplier Tube
PMW	Program Manager, Warfare
P-N	Positive to Negative Junction (also p-n)
PN or P/N	Part Number
POC	Point of Contact
POET	Primed Oscillator Expendable Transponder
POI	Probability of Intercept (also PI)
POL	Polarization
POM	Program Objective Memorandum

POP	Pulse-on-Pulse or Product Optimization Program
POST	Passive Optical Seeker Technology (Stinger missile)
PPI	Plan Position Indicator
PPS	Pulses Per Second
PRF	Pulse Repetition Frequency
PRI	Priority or Pulse Repetition Interval
PROM	Programmable Read-only Memory
PRR	Production Readiness Review or Pulse Repetition Rate
PRT	Pulse Repetition Time
P_s	Probability of Survival
P’s & Q’s	Pints and Quarts (small details)
PSK	Phase-shift Keying
PUPS	Portable Universal Programming System
PV	Photovoltaic
pw or PW	Pulse Width
PWB	Printed Wiring Board

q	electron charge
Q	Quantity Factor (figure of merit), Quadrature, aerodynamic pressure, or Charge (coulomb)
QA	Quality Assurance
QC	Quality Control
QED	Quod Erat Demonstradum (end of proof)(Satirically “quite easily done”)
QML	Qualified Manufacturer Listing
QPL	Qualified Parts List
QRC	Quick-Reaction Capability
QRD	Quick Reaction Demonstration
QRT	Quick-Reaction test

r or R	Radius or Range
R	Resistance, Reliability, or Roentgen
rad	Radian
R&D	Research and Development
RADAR	Radio Detection and Ranging
RADHAZ	Radiation Hazard
RAM	Random Access Memory, Radar Absorbing Material, Rolling Airframe Missile, or Reliability, Availability, and Maintainability
R&M	Reliability and Maintainability
RAT	Ram Air Turbine
RBOC	Rapid Blooming Offboard Chaff
RCP or RHCP	Right-hand Circular Polarization
RCS	Radar Cross Section
RCVR	Receiver
RDT&E	Research, Development, Test, & Evaluation
RDY	Ready
RE	Radiated Emissions
REC	Receive
RET	Return
RF	Radio Frequency
RFEXP	RF Expendables
RFI	Radio Frequency Interference, Ready-For-Issue, or Request for Information
RFP	Request for Proposal
RFQ	Request for Quotation
RFSS	Radio Frequency Simulation System (Army)
RGPO	Range Gate Pull Off
RGS	Range Gate Stealer
RGWO	Range Gate Walk Off (see RGPO)
RHAW	Radar Homing and Warning Receiver or Radar Homing All the Way
RHAWS	Radar Homing and Warning System
RINT	Radiation Intelligence
RIO	Radar Intercept Officer
RM	Radar Mile
rms or RMS	Root Mean Square
RNG	Range
ROC	Required Operational Capability
ROE	Rules of Engagement
ROI	Return on Investment
ROM	Read-only Memory or Rough Order of Magnitude
ROR	Range Only Radar or Rate of Return (financial)
ROT	Rate of Turn
ROWG	Response Optimization Working Group
RPG	Receiver Processor Group
RPM	Revolutions per Minute
RPT	Repeat
RPV	Remotely Piloted Vehicle
RRT	Rapid Reprogramming Terminal (a type of MLVS)
RS	Radiated Susceptibility or Remote Station
RSDS	Radar Signal Detecting Set
RSO	Range Safety Officer or Receiver, Set-on
RST	Receiver Shadow Time
RT	Remote Terminal, Termination Resistance, or Receiver/Transmitter (also R/T)
RUG	Radar Upgrade
RWR	Radar Warning Receiver
Rx	Receive

s, S, or sec	seconds
S	Signal Power, Surface Area, Secret, Electrical conductance (siemens), South, Scattering (as in S-parameters), or Seconds
SA	Situational Awareness, Semi-Active, Spectrum Analyzer, or Surface-to-Air (also S/A or S-A)
SA-()	Surface-to-Air missile number ()
SAE	Society of Automotive Engineers
SAM	Surface-to-Air Missile
SA-N-()	Naval Surface-to-Air missile number
SAR	Synthetic Aperture Radar, Special Access Required, Semi-Active Radar, Search and Rescue, or Specific Absorption Rate
SATS	Semi-Active Test System
SAW	Surface Acoustic Wave
SBIR	Small Business Innovative Research
SCI	Sensitive Compartmented Information
SCIF	Sensitive Compartmented Information Facility
SCN	Specification Change Notice
SCR	Software Change Request
SCP	Software Change Proposal
SCRB	Software Configuration Review Board
SCUD	Soviet short-range surface-to-surface missile
SE	Support Equipment
SDLM	Standard Depot Level Maintenance
SDI	Strategic Defense Initiative
SEAD	Suppression of Enemy Air Defense (pronounced “seed” or “C add”)
SEAL	Sea-Air-Land (Navy special forces)
sec	seconds (also S or s)
SECDEF	Secretary of Defense
SEI	Specific Emitter Identification
SEMA	Special Electronic Mission Aircraft
SERD	Support Equipment Recommendation Data
SHAPE	Supreme Headquarters Allied Powers Europe (NATO military command)
SHF	Super High Frequency (3 to 30 GHz)
SI	Special Intelligence or System International (Units)
SIF	Selective Identification Feature
SIGINT	Signals Intelligence
SIJ	Stand-In Jamming (also S/J)
SIM	Simulation
sin	Sine
SINCGARS	Single Channel Ground and Airborne Radio System
SIRCM	Suite of IR Countermeasures
SIRFC	Suite of Integrated RF Countermeasures
SJ	Support Jamming
S/J	Stand-In Jamming or Signal to Jamming Ratio
SL	Side lobe or Sea Level (also S.L.)
SLAM	Standoff Land Attack Missile
SLAR	Side-Looking Airborne Radar
SLC	Side Lobe Clutter
SLOC	Source Lines of Code or Sea Lines of Communication
SM	Statute Mile (also sm) or Standard Missile
SMA	Scheduled Maintenance Action or Sub-Miniature A connector
SMC	Sub-Miniature C connector
SME	Subject Matter Expert
SML	Support Material List
SMS	Stores Management Set or Status Monitoring System
S/N or SNR	Signal-to-Noise Ratio
SNORT	Supersonic Naval Ordnance Research Track
SNTK	Special Need to Know
SOF	Safety of Flight
SOJ	Stand-off Jammer
SONAR	Sound Navigation and Ranging
SOO	Statement of Objectives (replacing SOW)
SOP	Standard Operating Procedures
SORO	Scan-on-Receive Only
SOS	“Save Our Ship” (distress call with easy Morse code, i.e. ● ● ● – – – ● ● ●)
SOW	Statement of Work (being replaced by SOO)
SPAWAR	Space and Naval Warfare Systems Command
SPEC	Specification
SPIRITS	Spectral Infrared Imaging of Targets and Scenes
SPO	System Program Office
SPY	Radar on an AEGIS ship
sq	Square
sr	Steradian
SRA	Shop Replaceable Assembly
SRAM	Static Random Access Memory
SRB	Software Review Board
SRBOC	Super Rapid Blooming Offboard Chaff
SRD	Systems Requirements Document
SRS	Software Requirements Specification
SRU	Shop Replaceable Unit

SSA	Software (also Special or System) Support Activity, Source Selection Activity, or Solid State Amplifier
SSB	Single Side Band
SSI	Small Scale Integration
SSJ	Self Screening Jamming
SSM	Surface-to-Surface Missile
SSRO	Sector Scan Receive Only
SSW	Swept Square Wave
S&T	Science and Technology
STANAG	Standardization Agreement (NATO)
STAR	System Threat Assessment Report
stat	Statute
STBY	Standby
STC	Sensitivity Time Control, Short Time Constant or SHAPE Technical Center
STD	Software Test Description, Standard, or Sexually Transmitted Disease
STE	Standard Test Equipment
STOVL	Short Takeoff and Vertical Landing
STP	Software Test Plan, or Standard Temperature and Pressure (0°C at 1 atmosphere)
STR	Software (also System) Trouble Report
STT	Single Target Track
STU	Secure Telephone Unit
SUBSAM	Subsurface-to-Air Missile
SUT	System Under Test
S/W	Software (also SW)
SWAP	Size, Weight, And Power
SWC	Scan With Compensation
SWM	Swept Wave Modulation
SYSCOM	Systems Command
t	Time (also T)
T	Time (also t), tera (10^{12} multiplier), Temperature, or Telsa
TA	Target Acquisition or Terrain Avoidance
TAAF	Test, Analyze, and Fix
TAC	Tactical Air Command (now ACC)
TACAIR	Tactical Aircraft
TACAMO	Take Charge and Move Out (airborne strategic VLF communications relay system)
TACAN	Tactical Air Navigation
TACDS	Threat Adaptive Countermeasures Dispensing System
TACTS	Tactical Aircrew Combat Training System
TAD	Threat Adaptive Dispensing, Temporary Additional (also Active) Duty, or Tactical Air Direction
T&E	Test & Evaluation
TALD	Tactical Air Launched Decoy
TAMPS	Tactical Automated Mission Planning System
TAR	Target Acquisition Radar or Training Administrative Reserve
TAS	True Airspeed
TBA	To Be Announced
TBD	To Be Determined
TBMD	Theater Ballistic Missile Defense
TD	Technical Directive (also Director)
TDD	Target Detection Device
TDM	Time Division Multiplexing
TE	Transverse Electric
TEA	Technology Exchange Agreement
TEAMS	Tactical EA-6B Mission Support
TECHEVAL	Technical Evaluation
TEL	Transporter Erector Launcher
TEM	Transverse Electromagnetic
TEMP	Test and Evaluation Master Plan
TEMPEST	Not an acronym. Certification of reduced electromagnetic radiation for security considerations
TENA	Training Enabling Architecture
TERPES	Tactical Electronic Reconnaissance Processing and Evaluation System
TGT	Target
TIM	Technical Interchange Meeting
TM	Telemetry, Transverse Magnetic, or Technical Manual
TMD	Theater Missile Defense
TNC	Threaded Navy Connector
TOA	Time of Arrival
TOJ	Track on Jam
TOO	Target of Opportunity (HARM operating mode)
TOR	Tentative (also Tactical) Operational Requirement or Time of Receipt
TOS	Time on Station
TOT	Time on Target
TOW	Tube-Launched, Optically-Tracked, Wire-guided
TPI	Test Program Instruction
TPS	Test Program Set or Test Pilot School
TPWG	Test Plan Working Group
TQM	Total Quality Management
T/R	Transmit / Receive
TRB	Technical Review Board
TRD	Test Requirements Document
TREE	Transient Radiation Effects on Electronics

TRF	Tuned Radio Frequency
TRR	Test Readiness Review
TS	Top Secret
TSS	Tangential Sensitivity
TSSAM	Tri-Service Standoff Attack Weapon
TT	Target Track
TTI	Time To Impact/Intercept
TTG	Time-to-Go
TTL	Transistor-Transistor Logic
TTR	Target Tracking Radar
TV	Television
TVC	Thrust Vector Control
TWS	Track While Scan or Tail Warning System
TWSRO	Track While Scan on Receive Only
TWT	Traveling Wave Tube
TWTA	Traveling Wave Tube Amplifier
Tx	Transmit
TYCOM	Type Commander

u or µ	micron / micro (10^{-6} multiplier)
U	Unclassified, Unit, or Unknown (on RWR display)
UAS	Unmanned Aerial System
UAV	Unmanned (also uninhabited) Air (or Aerial) Vehicle
UCAV	Uninhabited Combat Air Vehicle (new USAF term for UAV)
UDF	User Data File
UDFG	User Data File Generator
UDM	User Data Module
UHF	Ultra High Frequency (300 MHz to 3 GHz)
ULF	Ultra Low Frequency (3 to 30 Hz)
µm	Micrometer
UN	United Nations
UNK	Unknown (also U)
UPC	Unique Planning Component
UPS	Uninterruptable Power Supply
us or µs	Microseconds
U.S.	United States
USA	United States of America or United States Army
USAF	United States Air Force
USMC	United States Marine Corps
USN	United States Navy
UTA	Uninhabited Tactical Aircraft
UUT	Unit Under Test
UV	Ultraviolet

v	Volts (also V), Velocity (also V or v_t)
V	Volts (also v), Velocity (also v or v_t), Volume, or Roman Numeral for five
VA	Veterans Administration, Volt-Amperes
VAQ	Prefix for Navy tactical EW squadron
V&V	Validation and Verification
VCO	Voltage Controlled Oscillator
Vdc or VDC	Volts Direct Current
VDT	Video Display Terminal
VECP	Value Engineering Change Proposal
VF	Prefix for Navy fighter squadron
VFO	Variable Frequency Oscillator
VFR	Visual Flight Rules
VGPO	Velocity Gate Pull Off
VGS	Velocity Gate Stealer
VGWO	Velocity Gate Walk Off
VHF	Very High Frequency (30 - 300 MHz)
VHSIC	Very High Speed Integrated Circuit
VID	Visual Identification
VLF	Very Low Frequency (3 to 30 kHz)
VLSI	Very Large Scale Integration
VLSIC	Very Large Scale Integrated Circuit
VMAQ	Prefix for Marine Tactical EW Squadron
VP	Prefix for Navy patrol squadron
VQ	Prefix for Navy special mission (usually reconnaissance) squadron
VRAM	Video Random Access Memory
VS or vs	Velocity Search or Versus (also vs.)
V/STOL	Vertical/Short Take-off and Landing (also VSTOL)
vt	Velocity (also V or v)
VTOL	Vertical Takeoff and Landing
VSWR	Voltage Standing Wave Ratio
VVA	Voltage Variable Attenuator

W	Watts, Weight, or West
W&T	Warning & Targeting
WARM	Wartime Reserve Mode
wb	Weber (magnetic flux)
WBS	Work Breakdown Structure
WC	Waveguide, circular
WFT	Windowed Fourier Transform
WGIRB	Working Group on Infrared Background
WORM	Write Once Read Many [times] (Refers to optical disks)
WOW	Weight on/off Wheels (also WonW or WoffW)
WPN	Weapons Procurement, Navy or Weapon

WR	Waveguide, rectangular
WRA	Weapon Replaceable Assembly
WRD	Waveguide, rectangular double ridged
WSSA	Weapons System Support Activity
WVR	Within Visual Range

x	Multiplication symbol
X	Reactance, Experimental, Extraordinary, Roman Numeral for ten, or X axis
X-EYE	Cross Eye
X-POL	Cross Polarization
XMIT	Transmit
XMTR	Transmitter

Y	Yes or Y-Axis
YAG	Yttrium-Aluminum Garnet
yd	Yard
YIG	Yttrium-Iron Garnet

Z	Impedance, Zenith, or Z-Axis

1xLR, 2xLR	One (or two or three etc.) Times Lethal Range
1v1 or 1-v-1	One versus One (Aerial engagement)

2D	Two Dimension

3D	Three Dimension
3M	Navy Maintenance and Material Management System

FUNDAMENTALS

Constants, Conversions, and Characters 2-1
Mathematical Notation 2-2
Frequency Spectrum 2-3
Decibel (dB) 2-4
Duty Cycle 2-5
Doppler Shift 2-6
Electronic Formulas 2-7
Missile and Electronic Equipment Designations 2-8
Radar Horizon / Line of Sight 2-9
Propagation Time / Resolution 2-10
Modulation 2-11
Transforms / Wavelets 2-12

This page intentionally left blank.

CONSTANTS, CONVERSIONS, and CHARACTERS

DECIMAL MULTIPLIER PREFIXES

Prefix	Symbol	Multiplier
exa	E	10^{18}
peta	P	10^{15}
tera	T	10^{12}
giga	G	10^{9}
mega	M	10^{6}
kilo	k	10^{3}
hecto	h	10^{2}
deka	da	10^{1}
deci	d	10^{-1}
centi	c	10^{-2}
milli	m	10^{-3}
micro	µ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto	a	10^{-18}

EQUIVALENCY SYMBOLS

Symbol	Meaning
∝	Proportional
~	Roughly equivalent
≈	Approximately
≅	Nearly equal
=	Equal
≡	Identical to, defined as
≠	Not equal
>>	Much greater than
>	Greater than
≥	Greater than or equal to
<<	Much less than
<	Less than
≤	Less than or equal to
∴	Therefore
°	Degrees
′	Minutes or feet
“	Seconds or inches

UNITS OF LENGTH

1 inch (in) = 2.54 centimeters (cm)
1 foot (ft) = 30.48 cm = 0.3048 m
1 yard (yd) ≅ 0.9144 meter
1 meter (m) ≅ 39.37 inches

1 kilometer (km) ≅ 0.54 nautical mile
≅ 0.62 statute mile
≅ 1093.6 yards
≅ 3280.8 feet

1 statute mile ≅ 0.87 nautical mile
(sm or stat. mile) ≅ 1.61 kilometers
= 1760 yards
= 5280 feet

1 nautical mile ≅ 1.15 statute miles
(nm or naut. mile) ≅ 1.852 kilometers
≅ 2025 yards
≅ 6076 feet
1 furlong = 1/8 mi (220 yds)

UNITS OF SPEED

1 foot/sec (fps) ≅ 0.59 knot (kt)*
≅ 0.68 stat. mph
≅ 1.1 kilometers/hr

1000 fps ≈ 600 knots

1 kilometer/hr ≅ 0.54 knot
(km/hr) ≅ 0.62 stat. mph
≅ 0.91 ft/sec

1 mile/hr (stat.) ≅ 0.87 knot
(mph) ≅ 1.61 kilometers/hr
≅ 1.47 ft/sec

1 knot* ≅ 1.15 stat. mph
≅ 1.69 feet/sec
≅ 1.85 kilometer/hr
≅ 0.515 m/sec

*A knot is 1 nautical mile per hour.

UNITS OF VOLUME

1 gallon ≅ 3.78 liters
≅ 231 cubic inches
≅ 0.1335 cubic ft
≅ 4 quarts
≅ 8 pints

1 fl ounce ≅ 29.57 cubic centimeter (cc) or milliliters (ml)

1 in^3 ≅ 16.387 cc

UNITS OF AREA

1 sq meter ≅ 10.76 sq ft

1 sq in ≅ 645 sq millimeters (mm)
= 1,000,000 sq mil
1 mil = 0.001 inch

1 acre = 43,560 sq ft

UNITS OF WEIGHT

1 kilogram (kg) ≅ 2.2 pounds (lbs)

1 pound ≅ 0.45 Kg
= 16 ounce (oz)
1 oz = 437.5 grains
1 carat ≅ 200 mg
1 stone (U.K.) ≅ 6.36 kg

NOTE: These are the U.S. customary (avoirdupois) equivalents, the troy or apothecary system of equivalents, which differ markedly, was used long ago by pharmacists.

UNITS OF POWER / ENERGY

1 H.P. = 33,000 ft-lbs/min
= 550 ft-lbs/sec
≅ 746 Watts
≅ 2,545 BTU/hr
(BTU = British Thermal Unit)

1 BTU ≅ 1055 Joules
≅ 778 ft-lbs
≅ 0.293 Watt-hrs

SCALES

OCTAVES

"N" Octaves = Freq to Freq x 2^N
i.e., One octave would be 2 to 4 GHz
Two Octaves would be 2 to 8 GHz
Three octaves would be 2 to 16 GHz

DECADES

"N" Decades = Freq to Freq x 10^N
i.e., One decade would be 1 to 10 MHz
Two decades would be 1 to 100 MHz
Three decades would be 1 to 1,000 MHz

TEMPERATURE CONVERSIONS

°F = (9/5) °C + 32
°C = (5/9)(°F - 32)
°K = °C + 273.16
°F = (9/5)(°K - 273) + 32
°C = °K - 273.16
°K = (5/9)(°F - 32) + 273

UNITS OF TIME

1 year = 365.2 days
1 fortnight = 14 nights (2 weeks)
1 century = 100 years
1 millennium = 1,000 years

NUMBERS

1 decade = 10
1 Score = 20
1 Billion = 1 x 10^9 (U.S.) (thousand million)
= 1 x 10^{12} (U.K.) (million million)

RULE OF THUMB FOR ESTIMATING DISTANCE TO LIGHTNING / EXPLOSION:

km: Divide 3 into the number of seconds which have elapsed between seeing the flash and hearing the noise.
miles: Multiply 0.2 times the number of seconds which have elapsed between seeing the flash and hearing the noise.

Note: Sound vibrations cause a change of density and pressure within a media, while electromagnetic waves do not. An audio tone won't travel through a vacuum but can travel at 1,100 ft/sec through air. When picked up by a microphone and used to modulate an EM signal, the modulation will travel at the speed of light.

Physical Constant	Quoted Value	S*	SI unit	Symbol
Avogadro constant	6.0221367×10^{23}	36	mol^{-1}	N_A
Bohr magneton	$9.2740154 \times 10^{-24}$	31	$J \bullet T^{-1}$	μ_B
Boltzmann constant	1.380658×10^{-23}	12	$J \bullet K^{-1}$	$k(=R\ N_A)$
Electron charge	$1.602177\ 33 \times 10^{-19}$	49	C	-e
Electron specific charge	$-1.758819\ 62 \times 10^{11}$	53	$C \bullet kg^{-1}$	$-e/m_e$
Electron rest mass	$9.1093897 \times 10^{-31}$	54	kg	m_e
Faraday constant	9.6485309×10^{4}	29	$C \bullet mol^{-1}$	F
Gravity (Standard Acceleration)	9.80665 or 32.174	0	m/sec^2 ft/sec^2	g
Josephson frequency to voltage ratio	4.8359767×10^{14}	0	$Hz \bullet V^{-1}$	2e/hg
Magnetic flux quantum	$2.06783461 \times 10^{-15}$	61	Wb	φ_o
Molar gas constant	8.314510	70	$J \bullet mol^{-1} \bullet K^{-1}$	R
Natural logarithm base	$\cong 2.71828$	-	dimensionless	*e*
Newtonian gravitational constant	6.67259×10^{-11}	85	$m^3 \bullet kg^{-1} \bullet s^{-2}$	G or K
Permeability of vacuum	$4\pi \times 10^{-7}$	d	H/m	μ_o
Permittivity of vacuum	$\cong 8.8541878 \times 10^{-12}$	d	F/m	ε_o
Pi	$\cong 3.141592654$		dimensionless	π
Planck constant	6.62659×10^{-34}	40	$J \bullet s$	h
Planck constant/2π	$1.05457266 \times 10^{-34}$	63	$J \bullet s$	$h(=h2\pi)$
Quantum of circulation	$3.63694807 \times 10^{-4}$	33	$J \bullet s \bullet kg^{-1}$	$h/2m_e$
Radius of earth (Equatorial)	6.378×10^{6} or 3963		m miles	
Rydberg constant	$1.0973731534 \times 10^{7}$	13	m^{-1}	R_χ
Speed of light	2.9979246×10^{8}	1	$m \bullet s^{-1}$	c
Speed of sound (dry air @ std press & temp)	331.4	-	$m \bullet s^{-1}$	-
Standard volume of ideal gas	22.41410×10^{-3}	19	$m^3 \bullet mol^{-1}$	V_m
Stefan-Boltzmann constant	5.67051×10^{-8}	19	$W \bullet K^{-4} \bullet m^{-2}$	σ

* S is the one-standard-deviation uncertainty in the last units of the value, d is a defined value.
(A standard deviation is the square root of the mean of the sum of the squares of the possible deviations)

THE SPEED OF LIGHT			
ACTUAL	UNITS	RULE OF THUMB	UNITS
≅ 2.9979246×10^8	m/sec	≅ 3×10^8	m/sec
≅ 299.79	m/μsec	≅ 300	m/μsec
≅ 3.27857×10^8	yd/sec	≅ 3.28×10^8	yd/sec
≅ 5.8275×10^8	NM/hr	≅ 5.8×10^8	NM/hr
≅ 1.61875×10^5	NM/sec	≅ 1.62×10^5	NM/sec
≅ 9.8357105×10^8	ft/sec	≅ 1×10^9	ft/sec
≅ 9.8357105×10^2	ft/μsec	≅ 1×10^3	ft/μsec

SPEED OF LIGHT IN VARIOUS MEDIUMS

The speed of EM radiation through a substance such as cables is defined by the following formula:

$$V = c/(\mu_r \varepsilon_r)^{1/2}$$

Where: μ_r = relative permeability

ε_r = relative permittivity

The real component of ε_r = dielectric constant of medium.

EM propagation speed in a typical cable might be 65-90% of the speed of light in a vacuum.

APPROXIMATE SPEED OF SOUND (MACH 1)

Sea Level (CAS/TAS)		**36,000 ft*** (TAS)	(CAS)
1230 km/hr	Decreases	1062 km/hr	630 km/hr
765 mph	Linearly	660 mph	391 mph
665 kts	To ⇒	573 kts	340 kts

* The speed remains constant until 82,000 ft, when it increases linearly to 1215 km/hr (755 mph, 656 kts) at 154,000 ft. Also see Section 8-2 for discussion of Calibrated Air Speed (CAS) and True Airspeed (TAS) and a plot of the speed of sound vs. altitude.

SPEED OF SOUND IN VARIOUS MEDIUMS

Substance	Speed (ft/sec)
Vacuum	Zero
Air	1,100
Fresh Water	4,700
Salt Water	4,900
Glass	14,800
Steel	20,000

DECIMAL / BINARY / HEX CONVERSION TABLE

Decimal	Binary	Hex	Decimal	Binary	Hex	Decimal	Binary	Hex
1	00001	01h	11	01011	0Bh	21	10101	15h
2	00010	02h	12	01100	0Ch	22	10110	16h
3	00011	03h	13	01101	0Dh	23	10111	17h
4	00100	04h	14	01110	0Eh	24	11000	18h
5	00101	05h	15	01111	0Fh	25	11001	19h
6	00110	06h	16	10000	10h	26	11010	1Ah
7	00111	07h	17	10001	11h	27	11011	1Bh
8	01000	08h	18	10010	12h	28	11100	1Ch
9	01001	09h	19	10011	13h	29	11101	1Dh
10	01010	0Ah	20	10100	14h	30	11110	1Eh

When using hex numbers it is always a good idea to use "h" as a suffix to avoid confusion with decimal numbers.

DECIMAL TO HEX CONVERSION

Both the following methods must use long division. Method one computes the digits from right to left while method two works from left to right.

Method one: To convert a decimal number above 16 to hex, divide the number by 16, then record the integer resultant and the remainder. Convert the remainder to hex and write this down - this will become the far right digit of the final hex number. Divide the integer you obtained by 16, and again record the new integer result and new remainder. Convert the remainder to hex and write it just to the left of the first decoded number. Keep repeating this process until dividing results in only a remainder. This will become the left-most character in the hex number (i.e., to convert 60 (decimal) to hex we have 60/16 = 3 with 12 remainder). 12 is C (hex) - this becomes the right most character. Then 3/16=0 with 3 remainder. 3 is 3 (hex). This becomes the next (and final) character to the left in the hex number, so the answer is 3C.

Method two: Use table of powers to work the digits from left to right:

For example: Here is your Decimal Number - 9379

Step 1 - Set up your chart:

65536	4096	256	16	1
16^4	16^3	16^2	16^1	16^0

Step 2 - Look in the table for the highest divisible number in the chart.
9379 / 4096 = 2 (the left-most Hex digit)
Must use long division to calculate the remainder (1187)

Step 3 - Divide the remainder with its highest divisible number in the chart:
1187 / 256 = 4 (the next digit to the right)
Must use long division to calculate the remainder (163)

Step 4 - Divide the remainder with its highest divisible number in the chart:
163 / 16 = 10 (or “A” from table L-1) (the next digit to the right)
Must use long division to calculate the remainder (3)

Step 5 - The remainder will not divide: remainder = 3 (the right-most Hex digit)

65536	4096	256	16	1
16^4	16^3	16^2	16^1	16^0
	2	4	A	3

HEX TO DECIMAL CONVERSION

To convert a hex number to decimal, multiply each hex digit converted to decimal by the decimal equivalent of the hex power represented and add the results.

For example: Here is your Hex Number - 24A3

Step 1 - Set up your chart:

4096	256	16	1
16^3	16^2	16^1	16^0

Step 2 - Place the numbers in a table:

4096	256	16	1
16^3	16^2	16^1	16^0
2	4	A	3

Step 3 - Multiply the Hex number times the power value:

2 x 4096 = 8192
4 x 256 = 1024
A(10) x 16 = 160
3 x 1 = 3

Step 4 - Add up your values:

Decimal value is 9379

GREEK ALPHABET

Case		Greek Alphabet Name	English Equivalent
Upper	Lower		
Α	α	alpha	a
Β	β	beta	b
Γ	γ	gamma	g
Δ	δ	delta	d
Ε	ε	epsilon	e
Ζ	ζ	zeta	z
Η	η	eta	e
Θ	θ , ‘	theta	th
Ι	ι	iota	i
Κ	κ	kappa	k
Λ	λ	lambda	l
Μ	μ	mu	m

Case		Greek Alphabet Name	English Equivalent
Upper	Lower		
Ν	ν	nu	n
Ξ	ξ	xi	x
Ο	ο	omicron	o
Π	π	pi	p
Ρ	ρ	rho	r
Σ	σ	sigma	s
Τ	τ	tau	t
Υ	υ	upsilon	u
Φ	φ	phi	ph
Χ	χ	chi	ch
Ψ	ψ	psi	ps
Ω	ω	omega	o

LETTERS FROM THE GREEK ALPHABET COMMONLY USED AS SYMBOLS

Symbol	Name	Use
α	alpha	space loss, angular acceleration, or absorptance
β	beta	3 dB bandwidth or angular field of view [radians]
Γ	Gamma	reflection coefficient
γ	gamma	electric conductivity, surface tension, missile velocity vector angle, or gamma ray
Δ	Delta	small change or difference
δ	delta	delay, control forces and moments applied to missile, or phase angle
ε	epsilon	emissivity [dielectric constant] or permittivity [farads/meter]
η	eta	efficiency or antenna aperture efficiency
Θ	Theta	angle of lead or lag between current and voltage
θ or ‘	theta	azimuth angle, bank angle, or angular displacement
Λ	Lambda	acoustic wavelength or rate of energy loss from a thermocouple
λ	lambda	wavelength or Poisson Load Factor
μ	mu	micro 10^{-6} [micron], permeability [henrys/meter], or extinction coefficient [optical region]

ν	nu	frequency
π	pi	3.141592654+
ρ	rho	charge/mass density, resistivity [ohm-meter], VSWR, or reflectance
Σ	Sigma	algebraic sum
σ	sigma	radar cross section [RCS], Conductivity [1/ohm-meter], or Stefan-Boltzmann constant
Τ	Tau	VSWR reflection coefficient
τ	tau	pulse width, atmospheric transmission, or torque
Φ	Phi	magnetic/electrical flux, radiant power [optical], or Wavelet’s smooth function [low pass filter]
φ	phi	phase angle, angle of bank, or beam divergence [optical region]
Ψ	Psi	time-dependent wave function or Wavelet’s detail function [high pass filter]
ψ	psi	time-independent wave function, phase change, or flux linkage [weber]
Ω	Omega	Ohms [resistance] or solid angle [optical region]. Note: inverted symbol is conductance [mhos]
ω	omega	carrier frequency in radians per second

MORSE CODE and PHONETIC ALPHABET

A - alpha	• –	J - juliett	• – –	S - sierra	• • •	1	• – – – –
B - bravo	– • • •	K - kilo	– • –	T - tango	–	2	• • – – –
C - charlie	– • – •	L - lima	• – • •	U - uniform	• • –	3	• • • – –
D - delta	– • •	M - mike	– –	V - victor	• • • –	4	• • • • –
E - echo	•	N - november	– •	W - whiskey	• – –	5	• • • • •
F - foxtrot	• • – •	O - oscar	– – –	X - x-ray	– • • –	6	– • • • •
G - golf	– – •	P - papa	• – •	Y - yankee	– • – –	7	– – • • •
H - hotel	• • • •	Q - quebec	– – • –	Z - zulu	– – • •	8	– – – • •
I - india	• •	R - romeo	• – •	0	– – – – –	9	– – – – •

Note: The International Maritime Organization agreed to officially stop Morse code use by February 1999, however use may continue by ground based amateur radio operators. (The U.S. Coast Guard discontinued its use in 1995.)

BASIC MATH / GEOMETRY REVIEW

EXPONENTS

$a^x a^y = a^{x+y}$

$a^x / a^y = a^{x-y}$

$(a^x)^y = a^{xy}$

$a^0 = 1$

Example:

$$\frac{x}{\sqrt{x}} = x \bullet x^{-\frac{1}{2}} = x^{(1-\frac{1}{2})} = x^{\frac{1}{2}} = \sqrt{x}$$

LOGARITHMS

$\log(xy) = \log x + \log y$

$\log(x/y) = \log x - \log y$

$\log(x^N) = N \log x$

If $z = \log x$ then $x = 10^z$

Examples: $\log 1 = 0$
$\log 1.26 = 0.1$; $\log 10 = 1$

if $10 \log N = \text{dB\#}$,
then $10^{(\text{dB\#}/10)} = N$

TRIGONOMETRIC FUNCTIONS

$\sin x = \cos(x-90°)$

$\cos x = -\sin(x-90°)$

$\tan x = \sin x / \cos x = 1 / \cot x$

$\sin^2 x + \cos^2 x = 1$

A radian is the angular measurement of an arc which has an arc length equal to the radius of the given circle, therefore there are 2π radians in a circle. One radian = 360°/2π = 57.296....°

ELLIPSE

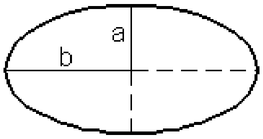

Area = π a b

Approx circumference

$$= 2\pi\sqrt{\frac{a^2 + b^2}{2}}$$

RECTANGLULAR SOLID

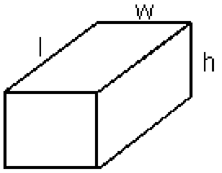

Area = lw

Volume = lwh

CYLINDER

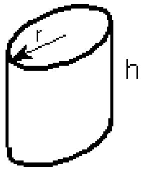

Volume − $\pi r^2 h$

Lateral surface area _ $2\pi rh$

ANGLES

$r^2 - x^2 + y^2$

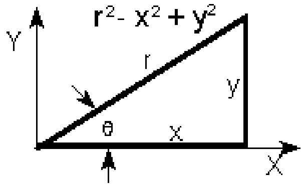

Sin θ = y/r

Cos θ = x/r

Tan θ = y/x

SPHERE

Surface area = $4\pi r^2$

Volume = $4/3\ \pi r^3$

Cross Section (circle) Area = πr^2

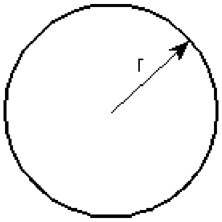

TRIANGLES

Angles: A + B + C = 180°

$c^2 = a^2 + b^2 - 2ab\cos C$

Area = 1/2 hc = 1/2 ab sin C

$b = \sqrt{e^2 + h^2}$

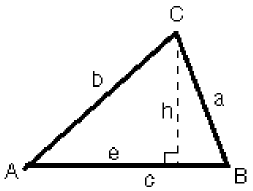

DERIVATIVES

Assume: a = fixed real #; u, v & w are functions of x

d(a)/dx = 0 ; d(sin u)/dx = du(cos u)/dx

d(x)/dx = 1 ; d(cos v)/dx = -dv(sin v)/dx

d(uvw)/dx = uvdw/dx + vwdu/dx + uwdv/dx +...etc

INTEGRALS

Note: All integrals should have an arbitrary constant of integration added, which is left off for clarity

Assume: a = fixed real #; u, & v are functions of x

∫adx = ax and ∫a f(x)dx = a∫f(x)dx

∫ (u +v)dx = ∫udx + ∫vdx ; $\int e^x dx = e^x$

∫(sin ax)dx = -(cos ax)/a ; ∫(cos ax)dx = (sin ax)/a

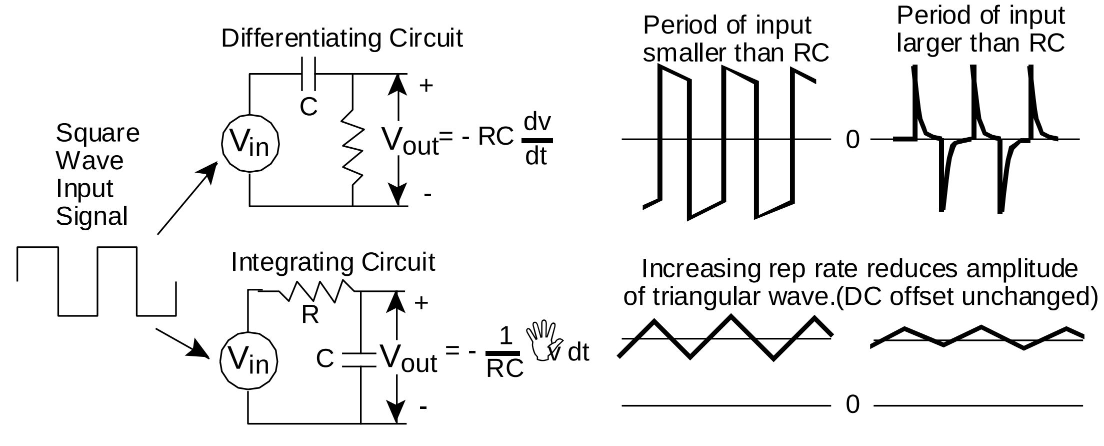
Differentiating Circuit
C
+
Vin
Vout = - RC dv/dt
-
Square Wave Input Signal
Integrating Circuit
R
+
Vin
C
Vout = - 1/RC ∫v dt
-
Period of input smaller than RC
Period of input larger than RC
0
Increasing rep rate reduces amplitude of triangular wave.(DC offset unchanged)
0

MATHEMATICAL NOTATION

The radar and EW communities generally accept some commonly used notation for the various parameters used in radar and EW calculations. For instance, “P” is almost always power and “G” is almost always gain. Textbooks and reference handbooks will usually use this notation in formulae and equations.

A significant exception is the use of “α” for space loss. Most textbooks do not develop the radar equation to its most usable form as does this reference handbook, therefore the concept of “α” just isn’t covered.

Subscripts are a different matter. Subscripts are often whatever seems to make sense in the context of the particular formula or equation. For instance, power may be “P”, “P_T”, “P_t”, or maybe “P_1”. In the following list, generally accepted notation is given in the left hand column with no subscripts. Subscripted notation in the indented columns is the notation used in this handbook and the notation often used in the EW community.

α = Space loss
α_1 = One way space loss, transmitter to receiver
α_2 = Two way space loss, xmtr to target (including radar cross section) and back to the rcvr
α_{1t} = One way space loss, radar transmitter to target, bistatic
α_{1r} = One way space loss, target to radar receiver, bistatic

Other notation such as α_{tm} may be used to clarify specific losses, in this case the space loss between a target and missile seeker, which could also be identified as α_{1r}.

A = Antenna aperture (capture area)
A_e = Effective antenna aperture
Å = Angstrom

B = Bandwidth (to 3dB points)
B_{IF} = 3 dB IF bandwidth of the receiver (pre-detection)
B_J = Bandwidth of the jamming spectrum
B_{MHz} = 3 dB bandwidth in MHz
B_N = Equivalent noise bandwidth, a.k.a. B
B_V = 3 dB video bandwidth of the receiver (post-detection) (Subscript V stands for video)
BF = Bandwidth reduction factor (jamming spectrum wider than the receiver bandwidth)
BW = Beamwidth (to 3 dB points)

c = Speed of Light

f = Frequency (radio frequency)
f_c = Footcandle (SI unit of illuminance)
f_D = Doppler frequency
f_R = Received frequency
f_T = Transmitted frequency

G = Gain
G_t = Gain of the transmitter antenna
G_r = Gain of the receiver antenna
G_{tr} = Gain of the transmitter/receiver antenna (monostatic radar)
G_J = Gain of the jammer
G_{JA} = Gain of the jammer antenna
G_{JT} = Gain of the jammer transmitter antenna
G_{JR} = Gain of the jammer receiver antenna
G_σ = Gain of reflected radar signal due to radar cross section

h = Height or Planks constant
h_{radar} = Height of radar
h_{target} = Height of target

J = Jamming signal (receiver input)
J_1 = Jamming signal (constant gain jammer)
J_2 = Jamming signal (constant power jammer)
J/S = Jamming to signal ratio (receiver input)

k = Boltzmann constant
$K_{1,2,3,4}$ = Proportionality constants, see Sections 4-3, 4-4, 4-5, and 4-1 respectively.

λ = Lambda, Wavelength or Poisson factor

L = Loss (due to transmission lines or circuit elements)

N = Receiver equivalent noise input (kT_oB)
NF = Noise figure

P = Power
P_d = Probability of detection
P_D = Power density
P_J = Power of a jammer transmitter
P_n = Probability of false alarm
P_r = Power received
P_t = Power of a transmitter

R = Range (straight line distance)
R_1 = Bistatic radar transmitter to target range
R_2 = Bistatic radar target to receiver range
R_J = Range of jammer to receiver (when separate from the target)
R_{NM} = Range in nautical miles

σ = Sigma, radar cross section (RCS)

S = Signal (receiver input)
S_R = Radar signal received by the jammer
S_{min} = Minimum receiver sensitivity

t = Time
t_{int} = Integration time
t_r = Pulse Rise Time
τ = Pulse Width

V = Velocity
V_r = Radial velocity

This page intentionally left blank.

FREQUENCY SPECTRUM

Figure 1, which follows, depicts the electromagnetic radiation spectrum and some of the commonly used or known areas. Figure 2 depicts the more common uses of the microwave spectrum. Figure 3 shows areas of the spectrum which are frequently referred to by band designations rather than by frequency.

Section 7-1 provides an additional breakdown of the EO/IR spectrum.

To convert from frequency (f) to wavelength (λ) and vice versa, recall that f = c/λ, or λ = c/f; where c = speed of light.

$$\lambda_{meter} = \frac{3x10^8}{f_{Hz}} = \frac{3x10^5}{f_{kHz}} = \frac{300}{f_{MHz}} = \frac{0.3}{f_{GHz}} \qquad \text{or} \qquad f_{Hz} = \frac{3x10^8}{\lambda_{meter}} \quad f_{kHz} = \frac{3x10^5}{\lambda_{meter}} \quad f_{MHz} = \frac{300}{\lambda_{meter}} \quad f_{GHz} = \frac{0.3}{\lambda_{meter}}$$

Some quick rules of thumb follow:

Metric:

Wavelength in cm = 30 / frequency in GHz
For example: at 10 GHz, the wavelength = 30/10 = 3 cm

English:

Wavelength in ft = 1 / frequency in GHz
For example: at 10 GHz, the wavelength = 1/10 = 0.1 ft

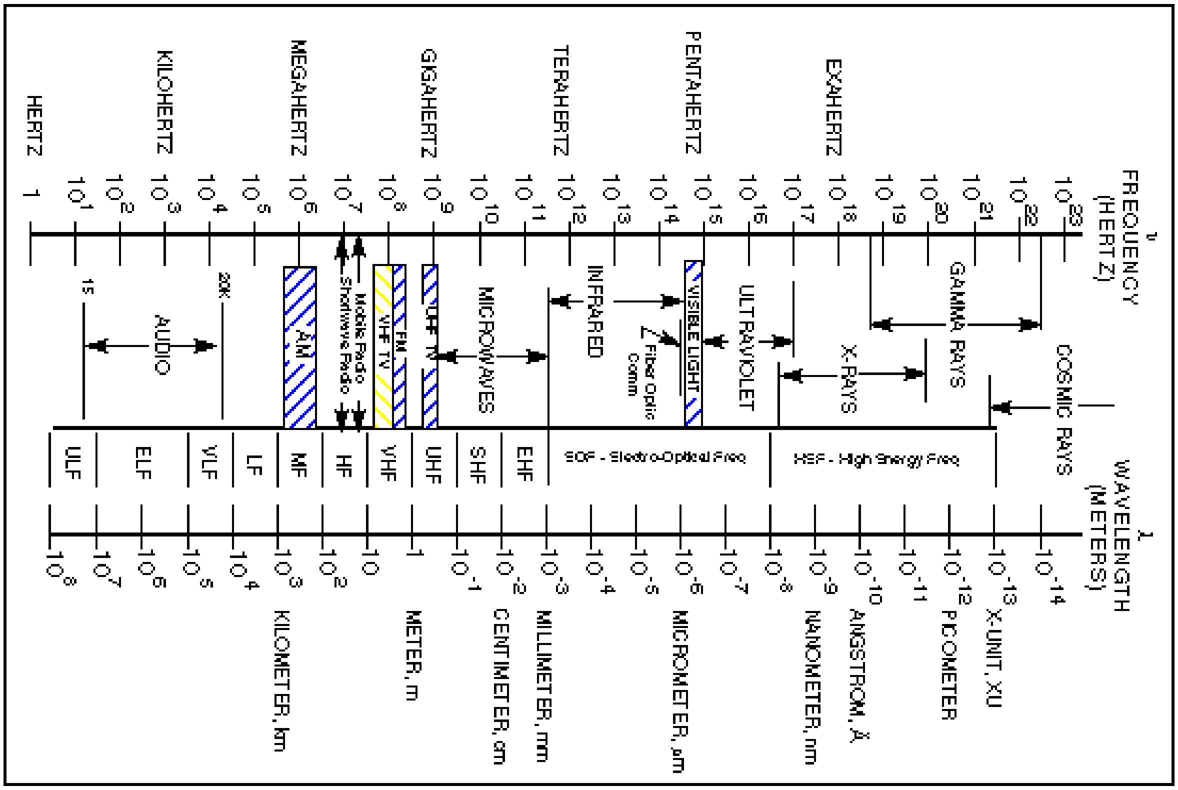
HERTZ
KILOHERTZ
MEGAHERTZ
GIGAHERTZ
TERAHERTZ
PENTAHERTZ
EXAHERTZ
1
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
10^{6}
10^{7}
10^{8}
10^{9}
10^{10}
10^{11}
10^{12}
10^{13}
10^{14}
10^{15}
10^{16}
10^{17}
10^{18}
10^{19}
10^{20}
10^{21}
10^{22}
10^{23}
ν
FREQUENCY (HERTZ)
15
20K
AUDIO
AM
Mobile Radio
Shortwave Radio
VHF TV
FM
UHF TV
MICROWAVES
INFRARED
Fiber Optic Comm
VISIBLE LIGHT
ULTRAVIOLET
X-RAYS
GAMMA RAYS
COSMIC RAYS
ULF
ELF
VLF
LF
MF
HF
VHF
UHF
SHF
EHF
EOF - Electro-Optical Freq
HEF - High Energy Freq
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
10^{-12}
10^{-13}
10^{-14}
λ
WAVELENGTH (METERS)
KILOMETER, km
METER, m
CENTIMETER, cm
MILLIMETER, mm
MICROMETER, μm
NANOMETER, nm
ANGSTROM, Å
PICOMETER
X-UNIT, XU

Figure 1. Electromagnetic Radiation Spectrum.

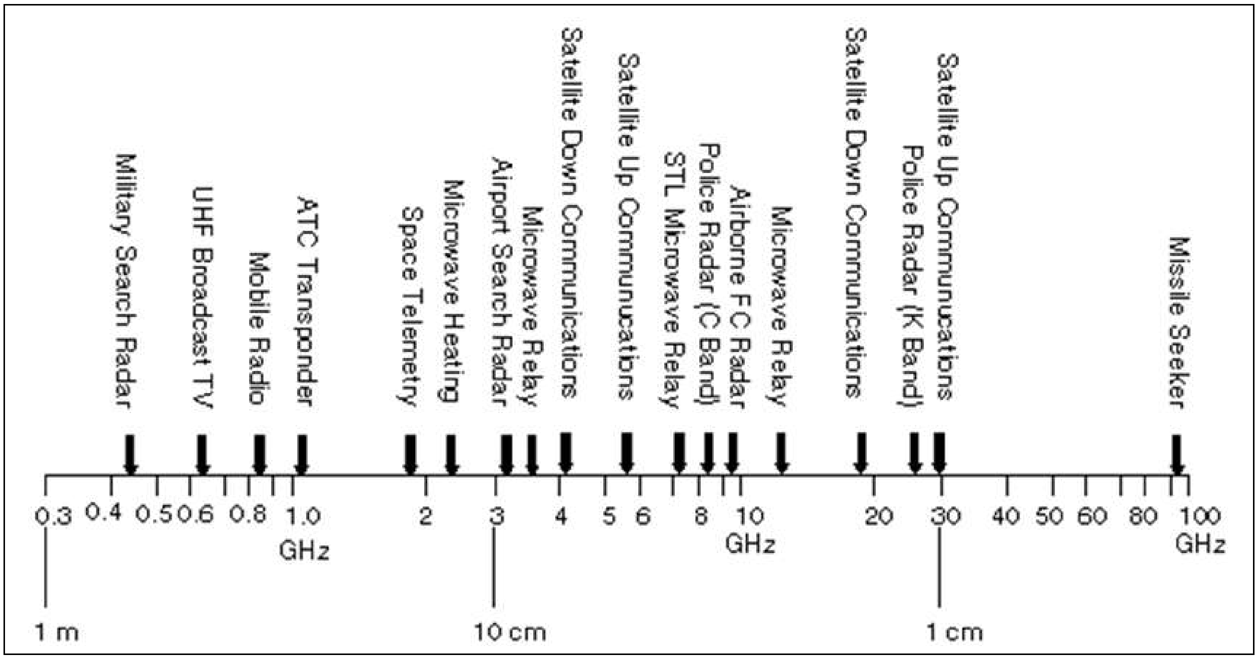

Figure 2. The Microwave Spectrum.

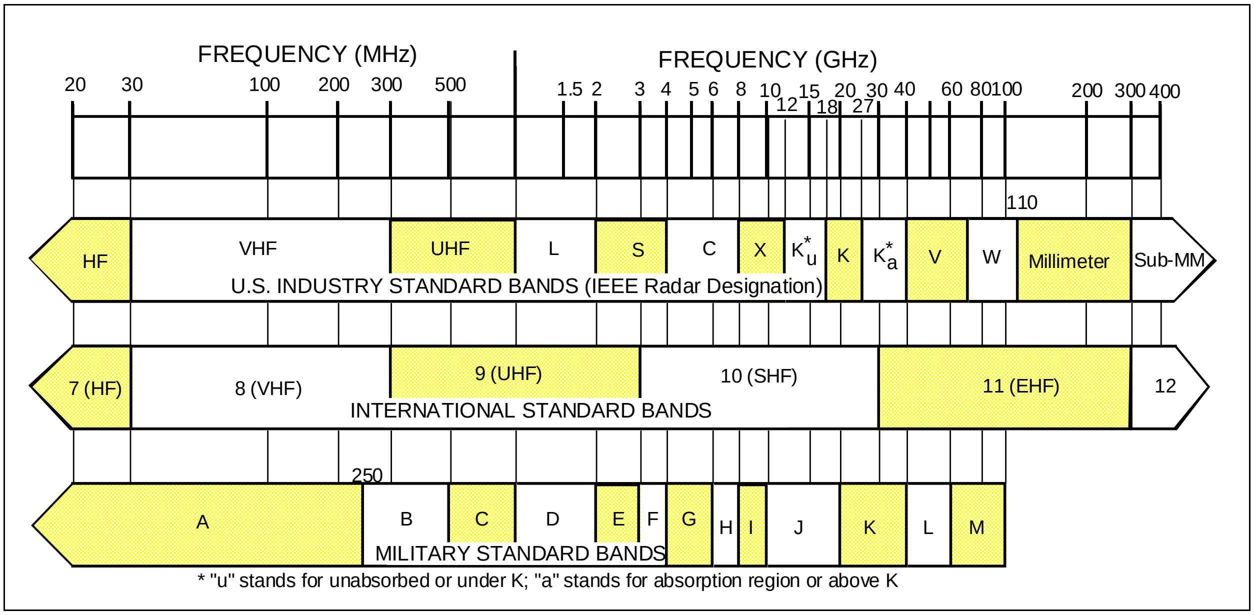
FREQUENCY (MHz)
20
30
100
200
300
500
FREQUENCY (GHz)
1.5
2
3
4
5
6
8
10
12
15
18
20
27
30
40
60
80
100
110
200
300
400
HF
VHF
UHF
L
S
C
X
K*u
K
K*a
V
W
Millimeter
Sub-MM
U.S. INDUSTRY STANDARD BANDS (IEEE Radar Designation)
7 (HF)
8 (VHF)
9 (UHF)
10 (SHF)
11 (EHF)
12
INTERNATIONAL STANDARD BANDS
250
A
B
C
D
E
F
G
H
I
J
K
L
M
MILITARY STANDARD BANDS
* "u" stands for unabsorbed or under K; "a" stands for absorption region or above K

Figure 3. Frequency Band Designations.

See Section 7, Figure 1 for a more detailed depiction of the UV and IR spectrum.

DECIBEL (dB)

The Decibel is a subunit of a larger unit called the bel. As originally used, the bel represented the power ratio of 10 to 1 between the strength or intensity i.e., power, of two sounds, and was named after Alexander Graham Bell. Thus a power ratio of 10:1 = 1 bel, 100:1 = 2 bels, and 1000:1 = 3 bels. It is readily seen that the concept of bels represents a logarithmic relationship since the logarithm of 100 to the base 10 is 2 (corresponding to 2 bels), the logarithm of 1000 to the base 10 is 3 (corresponding to 3 bels), etc. The exact relationship is given by the formula

$$\text{Bels} = \log(P_2/P_1) \qquad \textbf{[1]}$$

where P_2/P_1 represents the power ratio.

Since the bel is a rather large unit, its use may prove inconvenient. Usually a smaller unit, the Decibel or dB, is used. 10 decibels make one bel. A 10:1 power ratio, 1 bel, is 10 dB; a 100:1 ratio, 2 bels, is 20 dB. Thus the formula becomes

$$\text{Decibels (dB)} = 10 \log(P_2/P_1) \qquad \textbf{[2]}$$

The power ratio need not be greater than unity as shown in the previous examples. In equations [1] and [2], P_1 is usually the reference power. If P_2 is less than P_1, the ratio is less then 1.0 and the resultant bels or decibels are negative. For example, if P_2 is one-tenth P_1, we have

$$\text{bels} = \log(0.1/1) = -1.0 \text{ bels}$$

and $$\text{dB} = 10 \log(0.1/1) = -10 \text{ dB}.$$

It should be clearly understood that the term decibel does not in itself indicate power, but rather is a ratio or comparison between two power values. It is often desirable to express power levels in decibels by using a fixed power as a reference. The most common references in the world of electronics are the milliwatt (mW) and the watt. The abbreviation dBm indicates dB referenced to 1.0 milliwatt. One milliwatt is then zero dBm. Thus P_1 in equations [1] or [2] becomes 1.0 mW. Similarly, the abbreviation dBW indicates dB referenced to 1.0 watt, with P_2 being 1.0 watt, thus one watt in dBW is zero dBW or 30 dBm or 60 dBμW. For antenna gain, the reference is the linearly polarized isotropic radiator, dBLI. Usually the "L" and/or "I" is understood and left out.

dBc is the power of one signal referenced to a carrier signal, i.e., if a second harmonic signal at 10 GHz is 3 dB lower than a fundamental signal at 5 GHz, then the signal at 10 GHz is -3 dBc.

THE DECIBEL, ITS USE IN ELECTRONICS

The logarithmic characteristic of the dB makes it very convenient for expressing electrical power and power ratios. Consider an amplifier with an output of 100 watts when the input is 0.1 watts (100 milliwatts); it has an amplification factor of

$$P_2/P_1 = 100/0.1 = 1000$$

or a gain of:

$$10 \log(P_2/P_1) = 10 \log(100/0.1) = 30 \text{ dB}.$$

(notice the 3 in 30 dB corresponds to the number of zeros in the power ratio)

The ability of an antenna to intercept or transmit a signal is expressed in dB referenced to an isotropic antenna rather than as a ratio. Instead of saying an antenna has an effective gain ratio of 7.5, it has a gain of 8.8 dB (10 log 7.5).

A ratio of less than 1.0 is a loss, a negative gain, or attenuation. For instance, if 10 watts of power is fed into a cable but only 8.5 watts are measured at the output, the signal has been decreased by a factor of

$8.5/10 = .85$

or $10 \log(.85) = -0.7$ dB.

This piece of cable at the frequency of the measurement has a gain of -0.7 dB. This is generally referred to as a loss or attenuation of 0.7 dB, where the terms “loss” and “attenuation” imply the negative sign. An attenuator which reduces its input power by factor of 0.001 has an attenuation of 30 dB. The utility of the dB is very evident when speaking of signal loss due to radiation through the atmosphere. It is much easier to work with a loss of 137 dB rather than the equivalent factor of 2×10^{-14}.

Instead of multiplying gain or loss factors as ratios we can add them as positive or negative dB. Suppose we have a microwave system with a 10 watt transmitter, and a cable with 0.7 dB loss connected to a 13 dB gain transmit antenna. The signal loss through the atmosphere is 137 dB to a receive antenna with an 11 dB gain connected by a cable with 1.4 dB loss to a receiver. How much power is at the receiver? First, we must convert the 10 watts to milliwatts and then to dBm:

10 watts = 10,000 milliwatts

and

$10 \log (10{,}000/1) = 40$ dBm

Then

40 dBm - 0.7 dB + 13 dB - 137 dB + 11 dB - 1.4 dB = -75.1 dBm.

-75.1 dBm may be converted back to milliwatts by solving the formula:

$\text{mW} = 10^{(\text{dBm}/10)}$

giving: $10^{(-75.1/10)} = 0.00000003$ mW

Voltage and current ratios can also be expressed in terms of decibels, provided the resistance remains constant. First we substitute for P in terms of either voltage, V, or current, I. Since P=VI and V=IR we have:

$$P = I^2R = V^2/R$$

Thus for a voltage ratio we have:

$$\text{dB} = 10 \log[(V_2^2/R)/(V_1^2/R)] = 10 \log(V_2^2/V_1^2)$$
$$= 10 \log(V_2/V_1)^2 = 20 \log(V_2/V_1)$$

Like power, voltage can be expressed relative to fixed units, so one volt is equal to 0 dBV or 120 dBμV.

Similarly for current ratio: $\text{dB} = 20 \log(I_2/I_1)$

Like power, amperage can be expressed relative to fixed units, so one amp is equal to 0 dBA or 120 dBμA.

Decibel Formulas (where Z is the general form of R, including inductance and capacitance)

When impedances are equal: $dB = 10\log\frac{P2}{P1} = 20\log\frac{E2}{E1} = 20\log\frac{I2}{I1}$

When impedances are unequal: $dB = 10\log\frac{P2}{P1} = 20\log\frac{E2\sqrt{Z1}}{E1\sqrt{Z2}} = 20\log\frac{I2\sqrt{Z2}}{I1\sqrt{Z1}}$

SOLUTIONS WITHOUT A CALCULATOR

Solution of radar and EW problems requires the determination of logarithms (base 10) to calculate some of the formulae. Common "four function" calculators do not usually have a log capability (or exponential or fourth root functions either). Without a scientific calculator (or math tables or a Log-Log slide rule) it is difficult to calculate any of the radar equations, simplified or "textbook." The following gives some tips to calculate a close approximation without a calculator.

DECIBEL TABLE

DB	Power Ratio	Voltage or Current Ratio	DB	Power Ratio	Voltage or Current Ratio
0	1.00	1.00	10	10.0	3.16
0.5	1.12	1.06	15	31.6	5.62
1.0	1.26	1.12	20	100	10
1.5	1.41	1.19	25	316	17.78
2.0	1.58	1.26	30	1,000	31.6
3.0	2.00	1.41	40	10,000	100
4.0	2.51	1.58	50	10^5	316
5.0	3.16	1.78	60	10^6	1,000
6.0	3.98	2.00	70	10^7	3,162
7.0	5.01	2.24	80	10^8	10,000
8.0	6.31	2.51	90	10^9	31,620
9.0	7.94	2.82	100	10^{10}	10^5

For dB numbers which are a multiple of 10

An easy way to remember how to convert dB values that are a multiple of 10 to the absolute magnitude of the power ratio is to place a number of zeros equal to that multiple value to the right of the value 1.
40 dB = 10,000 : 1 (for power)

Minus dB moves the decimal point that many places to the left of 1.
-40 dB = 0.0001 : 1 (for power)

For voltage or current ratios, if the multiple of 10 is even, then divide the multiple by 2, and apply the above rules.
40 dB = 100 : 1 (for voltage)
-40 dB = 0.01 : 1

If the power in question is not a multiple of ten, then some estimation is required. The following tabulation lists some approximations. Some would be useful to memorize.

DB RULES OF THUMB

Multiply Current / Voltage By			Multiply Power By:	
if +dB	if -dB	**dB**	if +dB	if -dB
1	1	0	1	1
1.12	0.89	1	1.26	0.8
1.26	0.79	2	1.58	0.63
1.4	0.707	3	2	0.5
2.0	0.5	6	4	0.25
2.8	0.35	9	8	0.125
3.16	0.316	10	10	0.1
4.47	0.22	13	20	0.05
10	0.1	20	100	0.01
100	0.01	40	10,000	0.0001

You can see that the list has a repeating pattern, so by remembering just three basic values such as one, three, and 10 dB, the others can easily be obtained without a calculator by addition and subtraction of dB values and multiplication of corresponding ratios.

Example 1:

A 7 dB increase in power (3+3+1) dB is an increase of (2 x 2 x 1.26) = 5 times whereas
A 7 dB decrease in power (-3-3-1) dB is a decrease of (0.5 x 0.5 x 0.8) = 0.2.

Example 2: Assume you know that the ratio for 10 dB is 10, and that the ratio for 20 dB is 100 (doubling the dB increases the power ratio by a factor of ten), and that we want to find some intermediate value.

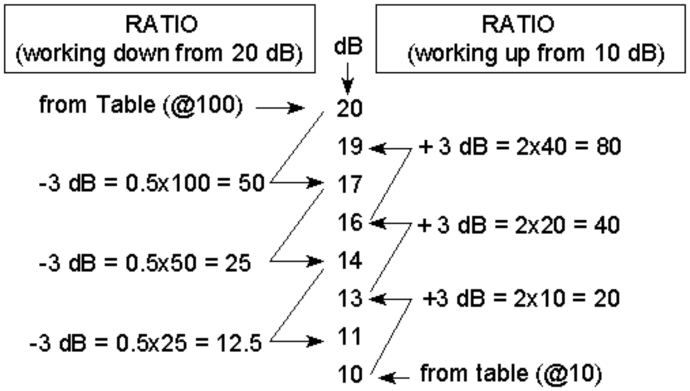

We can get more intermediate dB values by adding or subtracting one to the above, for example, to find the ratio at 12 dB we can:

work up from the bottom; 12 = 1+11 so we have 1.26 (from table) x 12.5 = 15.75
alternately, working down the top 12 = 13-1 so we have 20 x 0.8 (from table) = 16

The resultant numbers are not an exact match (as they should be) because the numbers in the table are rounded off. We can use the same practice to find any ratio at any other given value of dB (or the reverse).

dB AS ABSOLUTE UNITS

Power in absolute units can be expressed by using 1 Watt (or 1 milliwatt) as the reference power in the denominator of the equation for dB. We then call it dBW or dBm. We can then build a table such as the adjoining one.

From the above, any intermediate value can be found using the same dB rules and memorizing several dB values for determining the absolute power, given 48 dBm power output, we determine that 48 dBm = 50 dBm - 2 dB so we take the value at 50 dB which is 100W and divide by the value 1.58 (ratio of 2 dB) to get:

100 watts/1.58 = 63 W or 63,291 mW.

dB AS ABSOLUTE UNITS

dBμW	dBm	POWER	dBW
120	90	1 MW	60
90	60	1 kW	30
80	50	100 W	20
70	40	10 W	10
60	30	1 W (1000 mW)	0
50	20	100 mW	-10
40	10	10 mW	-20
33	3	2 mW	-27
32	2	1.58 mW	-28
31	1	1.26 mw	-29
30	0	1 mW	-30

Because dBW is referenced to one watt, the Log of the power in watts times 10 is dBW. The Logarithm of 10 raised by any exponent is simply that exponent. That is: $\text{Log}(10)^4 = 4$. Therefore, a power that can be expressed as any exponent of 10 can also be expressed in dBW as that exponent times 10. For example, 100 kW can be written 100,000 watts or 10^5 watts. 100 kW is then +50 dBW. Another way to remember this conversion is that dBW is the number of zeros in the power written in watts times 10. If the transmitter power in question is conveniently a multiple of ten (it often is) the conversion to dBW is easy and accurate.

This page intentionally left blank.

DUTY CYCLE

Duty cycle (or duty factor) is a measure of the fraction of the time a radar is transmitting. It is important because it relates to peak and average power in the determination of total energy output. This, in turn, ultimately affects the strength of the reflected signal as well as the required power supply capacity and cooling requirements of the transmitter.

Although there are exceptions, most radio frequency (RF) measurements are either continuous wave (CW) or pulsed RF. CW RF is uninterrupted RF such as from an oscillator. Amplitude modulated (AM), frequency modulated (FM), and phase modulated (PM) RF are considered CW since the RF is continuously present. The power may vary with time due to modulation, but RF is always present. Pulsed RF, on the other hand, is bursts (pulses) of RF with no RF present between bursts. The most general case of pulsed RF consists of pulses of a fixed pulse width (PW) which come at a fixed time interval, or period, (T). For clarity and ease of this discussion, it is assumed that all RF pulses in a pulse train have the same amplitude. Pulses at a fixed interval of time arrive at a rate or frequency referred to as the pulse repetition frequency (PRF) of so many pulse per second. Pulse repetition interval (PRI) and PRF are reciprocals of each other.

$$PRF = 1/T = 1/PRI \qquad \textbf{[1]}$$

Power measurements are classified as either peak pulse power, P_p, or average power, P_{ave}. The actual power in pulsed RF occurs during the pulses, but most power measurement methods measure the heating effects of the RF energy to obtain an average value of the power. It is correct to use either value for reference so long as one or the other is consistently used. Frequently it is necessary to convert from P_p to P_{ave}, or vice versa; therefore the relationship between the two must be understood. Figure 1 shows the comparison between P_p and P_{ave}.

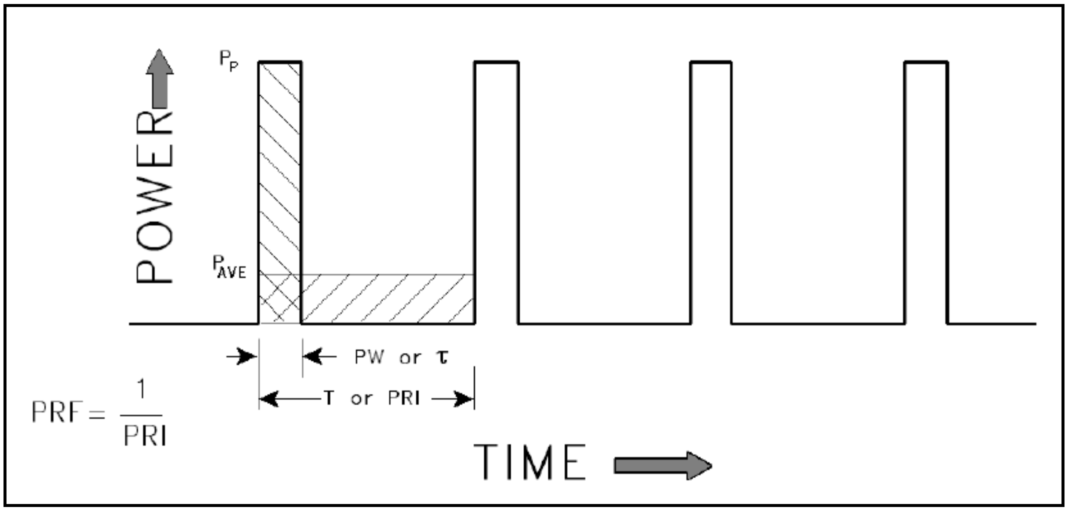

Figure 1. RF Pulse Train.

The average value is defined as that level where the pulse area above the average is equal to area below average between pulses. If the pulses are evened off in such a way as to fill in the area between pulses, the level obtained is the average value, as shown in Figure 1 where the shaded area of the pulse is used to fill in the area between pulses. The area of the pulse is the pulse width multiplied by the peak pulse power. The average area is equal to the average value of power multiplied by the pulse period.

Since the two values are equal:

$$P_{ave} \text{ x } T = P_p \text{ x } PW \qquad \textbf{[2]}$$

or

$$P_{ave}/P_p = PW/T \qquad \textbf{[3]}$$

Using [1]

$$P_{ave}/P_p = PW/T = PW \text{ x } PRF = PW/PRI = \text{duty cycle} \qquad \textbf{[4]}$$

(note that the symbol τ represents pulse width (PW) in most reference books)

The ratio of the average power to the peak pulse power is the duty cycle and represents the percentage of time the power is present. In the case of a square wave the duty cycle is 0.5 (50%) since the pulses are present 1/2 the time, the definition of a square wave.

For Figure 1, the pulse width is 1 unit of time and the period is 10 units. In this case the duty cycle is:

PW/T = 1/10 = 0.1 (10%).

A more typical case would be a PRF of 1,000 and a pulse width of 1.0 microseconds. Using [4], the duty cycle is 0.000001 x 1,000 = 0.001. The RF power is present one-thousandth of the time and the average power is 0.001 times the peak power. Conversely, if the power were measured with a power meter which responds to average power, the peak power would be 1,000 time the average reading.

Besides expressing duty cycle as a ratio as obtained in equation [4], it is commonly expressed as either a percentage or in decibels (dB). To express the duty cycle of equation [4] as a percentage, multiply the value obtained by 100 and add the percent symbol. Thus a duty cycle of 0.001 is also 0.1%.

The duty cycle can be expressed logarithmically (dB) so it can be added to or subtracted from power measured in dBm/dBW rather than converting to, and using absolute units.

$$\text{Duty cycle (dB)} = 10 \log(\text{duty cycle ratio}) \qquad \textbf{[5]}$$

For the example of the 0.001 duty cycle, this would be 10 log(0.001) = -30 dB. Thus the average power would be 30 dB less than the peak power. Conversely, the peak power is 30 dB higher than the average power.

For pulse radars operating in the PRF range of 0.25-10 kHz and PD radars operating in the PRF range of 10-500 kHz, typical duty cycles would be:

Pulse	~	0.1 - 3%	=	0.001 - .03	=	-30 to -15 dB
Pulse Doppler	~	5 - 50%	=	0.05 - .5	=	-13 to -3 dB
Continuous Wave	~	100%	=	1	=	0 dB

Intermediate Frequency Bandwidths of typical signals are:

Pulse	1 to 10 MHz
Chirp or Phase coded pulse	0.1 to 10 MHz
CW or PD	0.1 to 5 kHz

PRF is usually subdivided into the following categories: Low 0.25-4 kHz; Medium 8-40 kHz; High 50-300 kHz.

DOPPLER SHIFT

Doppler is the apparent change in wavelength (or frequency) of an electromagnetic or acoustic wave when there is relative movement between the transmitter (or frequency source) and the receiver.

Summary RF Equation for the Two-Way (radar) case

$$f_{Rec} = f_{Xmt} + f_D = f_{Xmt} + \frac{2(V_{Xmtr} + V_{Tgt})\, f_{Xmt}}{c}$$

Summary RF Equation for the One-Way (ESM) case

$$f_{Rec} = f_{Xmt} + f_D = f_{Xmt} + \frac{V_{Xmtr\ or\ Rec}\ f_{Xmt}}{c}$$

Rules of Thumb for two-way signal travel
(divide in half for one-way ESM signal measurements)
At 10 GHz, $f_D \cong$
35 Hz per Knot
19 Hz per km/Hr
67 Hz per m/sec
61 Hz per yd/sec
20 Hz per ft/sec

To estimate f_D at other frequencies, multiply these by:

$$\left[\frac{f_{Xmt}(GHz)}{10}\right]$$

The Doppler effect is shown in Figure 1. In everyday life this effect is commonly noticeable when a whistling train or police siren passes you. Audio Doppler is depicted, however Doppler can also affect the frequency of a radar carrier wave, the PRF of a pulse radar signal, or even light waves causing a shift of color to the observer.

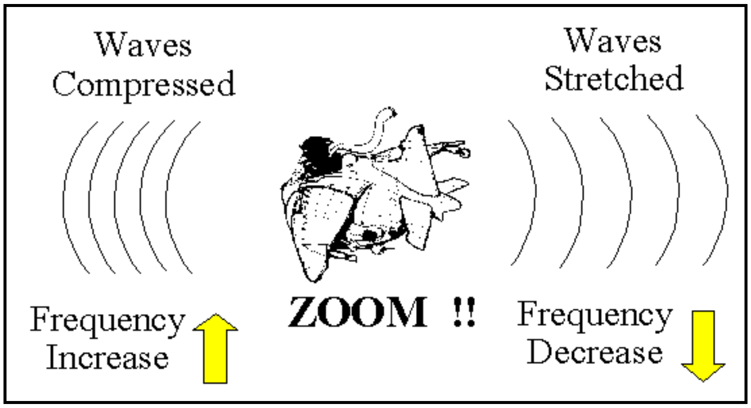

Figure 1. Doppler Frequency Creation From Aircraft Engine Noise.

How do we know the universe is expanding?

Answer: The color of light from distant stars is shifted to red (see Section 7-1: higher λ or lower frequency means Doppler shift is stretched, i.e., expanding).

A memory aid might be that the lights from a car (going away) at night are red (tail lights)!

Doppler frequency shift is directly proportional to velocity and a radar system can therefore be calibrated to measure velocity instead of (or along with) range. This is done by measuring the shift in frequency of a wave caused by an object in motion (Figure 2).

- * Transmitter in motion
- * Reflector in motion
- * Receiver in motion
- * All three

For a closing relative velocity:

- * Wave is compressed
- * Frequency is increased

For an opening relative velocity:

- * Wave is stretched
- * Frequency is decreased

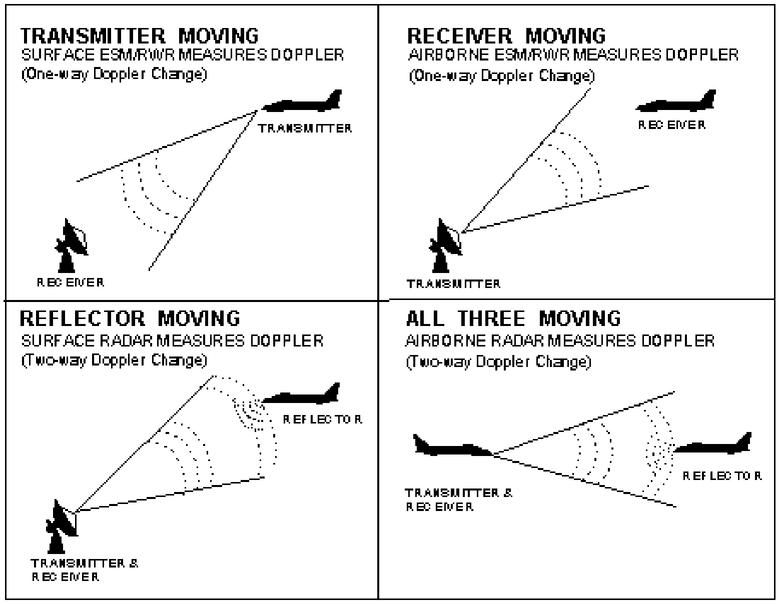
TRANSMITTER MOVING
SURFACE ESM/RWR MEASURES DOPPLER
(One-way Doppler Change)
TRANSMITTER
RECEIVER
RECEIVER MOVING
AIRBORNE ESM/RWR MEASURES DOPPLER
(One-way Doppler Change)
RECEIVER
TRANSMITTER
REFLECTOR MOVING
SURFACE RADAR MEASURES DOPPLER
(Two-way Doppler Change)
REFLECTOR
TRANSMITTER & RECEIVER
ALL THREE MOVING
AIRBORNE RADAR MEASURES DOPPLER
(Two-way Doppler Change)
REFLECTOR
TRANSMITTER & RECEIVER

Figure 2. Methods of Doppler Creation.

To compute Doppler frequency we note that velocity is range rate; V = dr/dt

For the reflector in motion case, you can see the wave compression effect in Figure 3 when the transmitted wave peaks are one wavelength apart. When the first peak reaches the target, they are still one wavelength apart (point a).

When the 2nd peak reaches the target, the target has advanced according to its velocity (vt) (point b), and the first reflected peak has traveled toward the radar by an amount that is less than the original wavelength by the same amount (vt) (point c).

As the 2nd peak is reflected, the wavelength of the reflected wave is 2(vt) less than the original wavelength (point d).

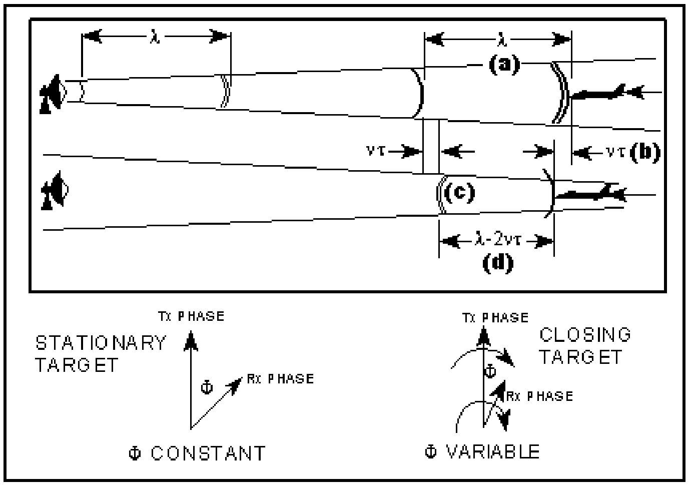
λ
λ
(a)
vτ
vτ (b)
(c)
λ-2vτ
(d)
Tx PHASE
STATIONARY TARGET
Φ
Rx PHASE
Φ CONSTANT
Tx PHASE
CLOSING TARGET
Φ
Rx PHASE
Φ VARIABLE

Figure 3. Doppler Compression Equivalent to Variable Phase Shift.

The distance the wave travels is twice the target range. The reflected phase lags transmitted phase by 2x the round trip time. For a fixed target the received phase will differ from the transmitted phase by a constant phase shift. For a moving target the received phase will differ by a changing phase shift.

For the closing target shown in Figure 3, the received phase is advancing with respect to the transmitted phase and appears as a higher frequency.

Doppler is dependent upon closing velocity, not actual radar or target velocity as shown in Figure 4.

For the following equations (except radar mapping), we assume the radar and target are moving directly toward one another to simplify calculations (if this is not the case, use the velocity component of one in the direction of the other in the formulas).

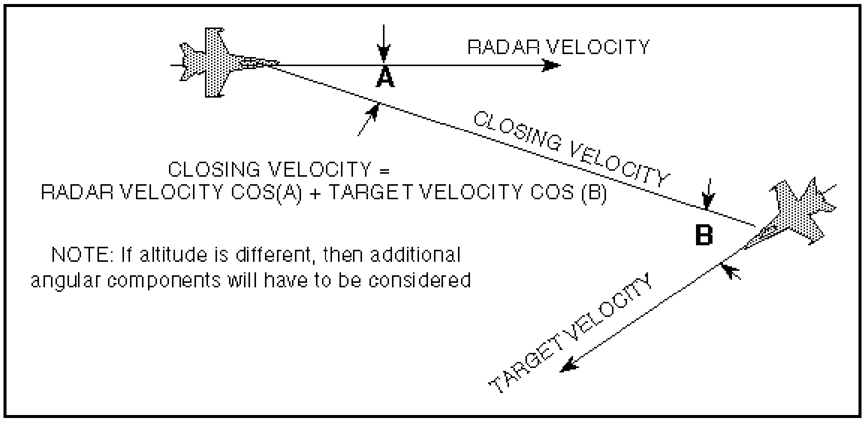
RADAR VELOCITY
A
CLOSING VELOCITY
CLOSING VELOCITY =
RADAR VELOCITY COS(A) + TARGET VELOCITY COS (B)
B
NOTE: If altitude is different, then additional
angular components will have to be considered
TARGET VELOCITY

Figure 4. Doppler Depends Upon Closing Velocity.

For the case of a moving reflector, doppler frequency is proportional to 2x the transmitted frequency:
Higher rf = higher doppler shift
$f_D = (2 \text{ x } V_{\text{Target}})(f/c)$

Speed of Light
Conversions
* * *
$c \cong 2.9979 \text{ x } 10^8$ m/sec
$c \cong 5.8275 \text{ x } 10^8$ nm/hr (knots)
$c \cong 9.8357 \text{ x } 10^8$ ft/sec

Likewise, it can be shown that for other cases, the following relationships hold:

For an airplane radar with an airplane target (The “all three moving” case, aircraft radar transmitter, target, and aircraft radar receiver)
$f_D = 2(V_{\text{Radar}} + V_{\text{Target}})(f/c)$

For the case of a semi-active missile receiving signals (Also “all three moving”)
$f_D = (V_{\text{Radar}} + 2V_{\text{Target}} + V_{\text{Missile}})(f/c)$

For the airplane radar with a ground target (radar mapping) or vice versa.
$f_D = 2(V_{\text{Radar}} \text{Cos}\theta \, \text{Cos}\varphi)(f/c)$, Where θ and φ are the radar scan azimuth and depression angles.

For a ground based radar with airborne target - same as previous using target track crossing angle and ground radar elevation angle.

For the ES/ESM/RWR case where only the target or receiver is moving (One-way Doppler measurements)
$f_D = V_{\text{Receiver or Target}} (f/c)$

Note: See Figure 4 if radar and target are not moving directly towards or away from one another.

Figure 5 depicts the results of a plot of the above equation for a moving reflector such as might be measured with a ground radar station illuminating a moving aircraft.

It can be used for the aircraft-to-aircraft case, if the total net closing rate of the two aircraft is used for the speed entry in the figure.

It can also be used for the ES/ESM case (one-way doppler measurements) if the speed of the aircraft is used and the results are divided by two.

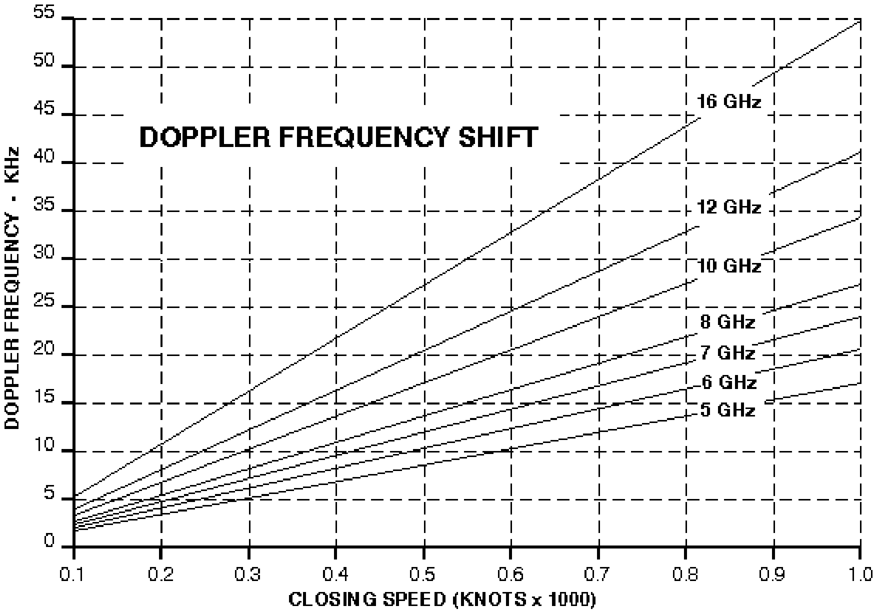

Figure 5. Two-Way Doppler Frequency Shift.

SAMPLE PROBLEMS:

(1) If a ground radar operating at 10 GHz is tracking an airplane flying at a speed of 500 km/hr tangential to it (crossing pattern) at a distance of 10 km, what is the Doppler shift of the returning signal?

Answer: Since the closing velocity is zero, the Doppler is also zero.

(2) If the same aircraft turns directly toward the ground radar, what is the Doppler shift of the returning signal?

Answer: 500 km/hr = 270 kts from Section 2-1. From Figure 4 we see that the Doppler frequency is about 9.2 KHz.

(3) Given that a ground radar operating at 7 GHz is Doppler tracking an aircraft 20 km away (slant range) which is flying directly toward it at an altitude of 20,000 ft and a speed of 800 ft/sec, what amount of VGPO switch would be required of the aircraft jammer to deceive (pull) the radar to a zero Doppler return?

Answer: We use the second equation from the bottom of page 2-6.3 which is essentially the same for this application except a ground radar is tracking an airplane target (vs. an airplane during ground mapping), so for our application we use a positive elevation angle instead of a negative (depression) angle.

$f_D = 2(V_r \text{ Cos } \theta \text{ Cos } \varphi)(f/c)$, where θ is the aircraft track crossing angle and φ is the radar elevation angle.

Since the aircraft is flying directly at the radar, $\theta = 0°$; the aircraft altitude = 20,000 ft = 6,096 meters.

Using the angle equation in Section 2-1, sin φ = x/r = altitude / slant range, so:
$\varphi = \sin^{-1}$ (altitude/slant range) = $\sin^{-1}$ (6,096 m / 20,000 m) = 17.7°

F_D = 2(800 ft/sec Cos 0° Cos 17.7°)(7x10^9 Hz / 9.8357 x 10^9 ft/sec) = 10,845 Hz

This page intentionally left blank.

ELECTRONIC FORMULAS

Ohm's Law Formulas for D-C Circuits. $E = IR = \frac{P}{I} = \sqrt{PR}$ $\quad P = I^2 R = EI = \frac{E^2}{R}$

Ohm's Law Formulas for A-C Circuits and Power Factor.

$$E = IZ = \frac{P}{I \cos \Theta} = \sqrt{\frac{P\,Z}{\cos \Theta}} \qquad P = I^2\, Z \cos \Theta = IE \cos \Theta = \frac{E^2 \cos \Theta}{Z}$$

In the above formulas Θ is the angle of lead or lag between current and voltage and cos Θ = P/EI = power factor or *pf*.

$$pf = \frac{Active\ \ power\ (in\ watts)}{Apparent\ \ power\ (in\ volt\text{-}amps)} = PoverE\ I \qquad pf = \frac{R}{Z}$$

Note: Active power is the "resistive" power and equals the equivalent heating effect on water.

Voltage/Current Phase Rule of Thumb Remember "ELI the ICE man"

ELI: Voltage (E) comes before (leads) current (I) in an inductor (L)
ICE: Current (I) comes before (leads) Voltage (E) in a capacitor (C)

Resistors in Series $R_{total} = R_1 + R_2 = R_3 + \ldots$

Two Resistors in Parallel $R_t = \frac{R_1\, R_2}{R_1 + R_2}$

Resistors in Parallel, General Formula $R_{total} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots}$

Resonant Frequency Formulas *Where in the second formula f is in kHz and L and C are in microunits.

$$f = \frac{1}{2\pi\sqrt{LC}}\ ,\ \ or\ \ f = \frac{159.2*}{\sqrt{LC}} \quad L = \frac{1}{4\pi^2 f^2 C}\ ,\ \ or\ \ L = \frac{25{,}330*}{f^2 C} \quad C = \frac{1}{4\pi^2 f^2 L}\ ,\ \ or\ \ C = \frac{25{,}330*}{f^2 L}$$

Conductance $G = \frac{1}{R}$ $(for\ D\text{-}C\ circuit)$ $\quad G = \frac{R}{R^2 + X^2}$ $(for\ A\text{-}C\ circuit)$

Reactance Formulas $X_C = \frac{1}{2\pi\, f\, C}$ $\quad C = \frac{1}{2\pi\, f\ X_C}$ $\quad X_L = 2\pi\, fL$ $\quad L = \frac{X_L}{2\pi\, f}$

Impedance Formulas $Z = \sqrt{R^2 + (X_L - X_C)^2}$ $(for\ series\ circuit)$

$Z = \frac{RX}{\sqrt{R^2 + X^2}}$ $(for\ R\ and\ X\ in\ parallel)$

Q or Figure of Merit $Q = \frac{X_L}{R}\ \ or\ \ \frac{X_C}{R}$

Frequency Response

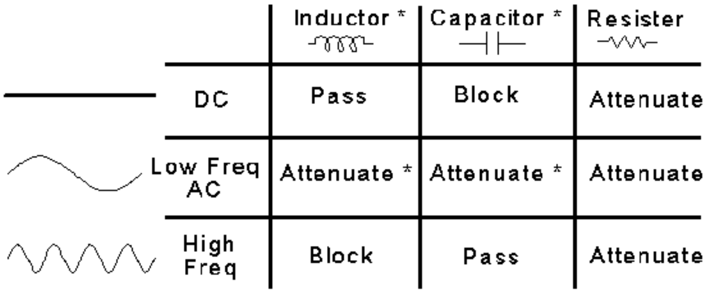

	Inductor *	Capacitor *	Resister
DC	Pass	Block	Attenuate
Low Freq AC	Attenuate *	Attenuate *	Attenuate
High Freq	Block	Pass	Attenuate

* Attenuation varies as a function of the value of the each device and the frequency

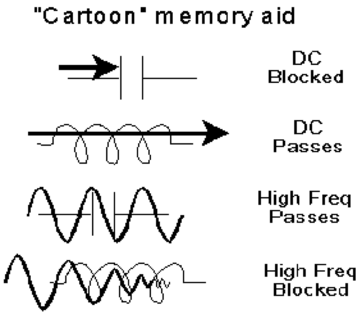

Sinusoidal Voltages and Currents

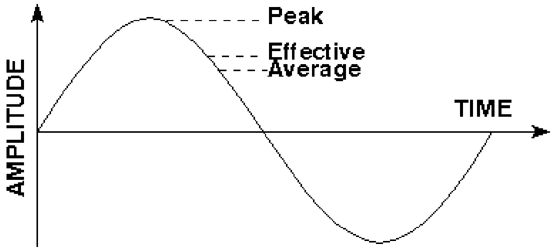

Effective value = 0.707 x peak value
[Also known as Root-Mean Square (RMS) value]

Half Cycle Average value = 0.637 x peak

Peak value = 1.414 x effective value

∴ Effective value = 1.11 x average value

Three-phase AC Configurations

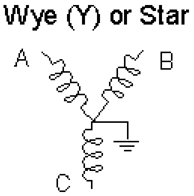

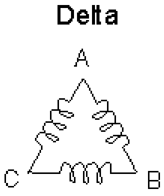

(120° phase difference between each voltage)
If the connection to a three phase AC configuration is miswired, switching any two of the phases will put it back in the proper sequence.
Electric power for ships commonly uses the delta configuration, while commercial electronic and aircraft applications commonly use the wye configuration.

Color Code for House Wiring:	**PURPOSE:**	**Color Code for Chassis Wiring:**
Black or red	HOT	Red
White	NEUTRAL (Return)	White
Green or bare	GROUND	Black

Color Code for Resistors: **First and second band:**
(and third band # of zeros if not gold/silver)

				Third band Multiplier		**Fourth band Tolerance**	
0	Black	5	Green	.1	Gold	5%	Gold
1	Brown	6	Blue	.01	Silver	10%	Silver
2	Red	7	Violet			20%	No color
3	Orange	8	Gray				
4	Yellow	9	White				

The third color band indicates number of zeros to be added after figures given by first two color bands. But if third color band is gold, multiply by 0.1 and if silver multiply by 0.01. Do not confuse with fourth color-band that indicates tolerance. Thus, a resistor marked blue-red-gold-gold has a resistance of 6.2 ohms and a 5% tolerance.

MISSILE AND ELECTRONIC EQUIPMENT DESIGNATIONS

Missiles are designated with three letters from the columns below plus a number (i.e., AIM-7M) Suffixes (M in this case) indicate a modification.

First Letter Launch Environment	**Second Letter** Mission Symbols	**Third Letter** Vehicle Type
A Air	D Decoy	M Guided Missile
B Multiple	E Special electronic	N Probe (non-orbital instruments)
C Coffin	G Surface attack	R Rocket (without installed or remote control guidance)
H Silo stored	I Intercept, aerial	
L Silo launched	Q Drone	
M Mobile	T Training	
P Soft Pad	U Underwater attack	
R Ship	W Weather	
U Underwater		

U.S. military electronic equipment is assigned an identifying alphanumeric designation that is used to uniquely identify it. This system is commonly called the "AN" designation system, although its formal name is the Joint Electronics Type Designation System (JETDS). The letters AN preceding the equipment indicators formerly meant "Army/Navy," but now are a letter set that can only be used to indicate formally designated DOD equipment. The first three letters following the "AN/" indicate Platform Installation, Equipment Type, and Equipment Function, respectively. The appropriate meaning is selected from the lists below. The letters following the AN designation numbers provide added information about equipment. Suffixes (A, B, C, etc.) indicate a modification. The letter (V) indicates that variable configurations are available. The letter (X) indicates a development status. A parenthesis ( ) without a number within it indicates a generic system that has not yet received a formal designation, e.g., AN/ALQ( ). Quite often the () is pronounced "bow legs" since they look like the shape of cowboy legs.

First Letter Platform Installation	**Second Letter** Equipment Type	**Third Letter** Function or Purpose
A Piloted aircraft	A Invisible light, heat radiation	B Bombing
B Underwater mobile, submarine	C Carrier	C Communications
D Pilotless carrier	D Radiac	D Direction finder, reconnaissance and/or surveillance
F Fixed ground	F Photographic	E Ejection and/or release
G General ground use	G Telegraph or teletype	G Fire control or searchlight directing
K Amphibious	I Interphone and public address	H Recording and/or reproducing
M Mobile (ground)	J Electromechanical or inertial wire covered	K Computing
P Portable	K Telemetering	M Maintenance and/or test assemblies
S Water	L Countermeasures	N Navigation aids
T Ground, transportable	M Meteorological	Q Special or combination of purposes
U General utility	N Sound in air	R Receiving, passive detecting
V Vehicular (ground)	P Radar	S Detecting and/or range and bearing, search
W Water surface and underwater combination	Q Sonar and underwater sound	T Transmitting
Z Piloted-pilotless airborne vehicle combination	R Radio	W Automatic flight or remote control
	S Special or combinations of types	X Identification and recognition
	T Telephone (wire)	Y Surveillance and control
	V Visual and visible light	
	W Armament	
	X Facsimile or television	
	Y Data Processing	

This page intentionally left blank.

RADAR HORIZON / LINE OF SIGHT

There are limits to the reach of radar signals. At the frequencies normally used for radar, radio waves usually travel in a straight line. The waves may be obstructed by weather or shadowing, and interference may come from other aircraft or from reflections from ground objects (Figure 1).

As also shown in Figure 1, an aircraft may not be detected because it is below the radar line which is tangent to the earth's surface.

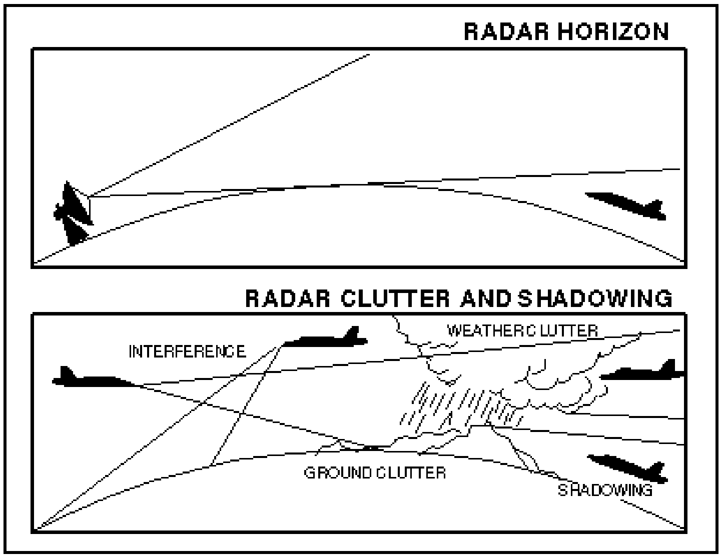

Figure 1. Radar Horizon and Shadowing.

Some rules of thumb are:

Range (to horizon):

$$R_{NM} = 1.23\sqrt{h_{radar}} \quad with\ h\ in\ ft$$

Range (beyond horizon / over earth curvature):

$$R_{NM} = 1.23\left(\sqrt{h_{radar}} + \sqrt{h_{target}}\right) \quad with\ h\ in\ ft$$

In obtaining the radar horizon equations, it is common practice to assume a value for the Earth's radius that is 4/3 times the actual radius. This is done to account for the effect of the atmosphere on radar propagation. For a true line of sight, such as used for optical search and rescue, the constant in the equations changes from 1.23 to 1.06.

A nomograph for determining maximum target range is depicted in Figure 2. Although an aircraft is shown to the left, it could just as well be a ship, with radars on a mast of height "h." Any target of height (or altitude) "H" is depicted on the right side.

See also Section 5-1 on ducting and refraction, which may increase range beyond these distances.

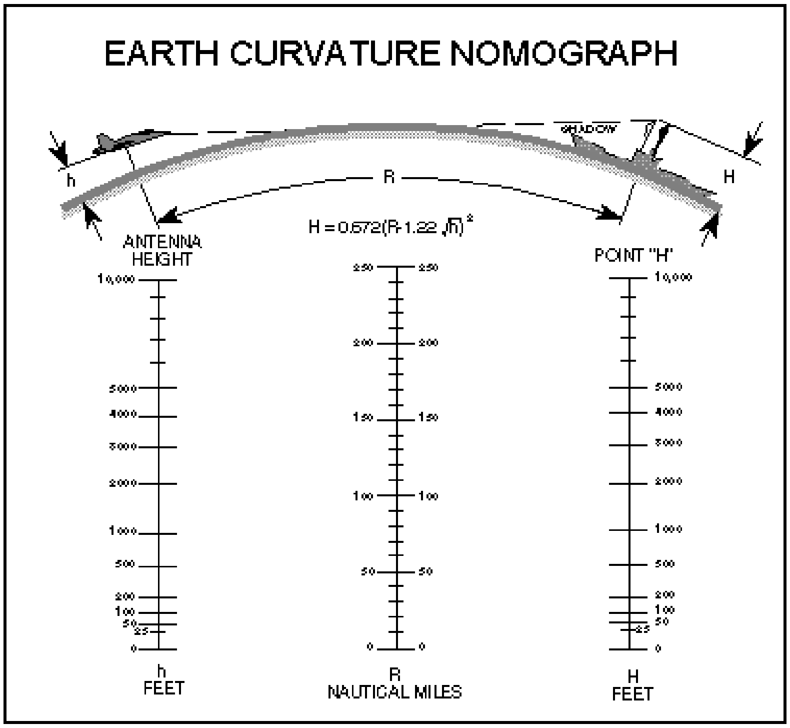

Figure 2. Earth Curvature Nomograph.

This data was expanded in Figure 3 to consider the maximum range one aircraft can detect another aircraft using:

$$R_{NM} = 1.23\left(\sqrt{h_{radar}} + \sqrt{h_{target}}\right)$$

(with h in feet)

It can be used for surface targets if $H_{target} = 0$. It should be noted that most aircraft radars are limited in power output, and would not detect small or surface objects at the listed ranges.

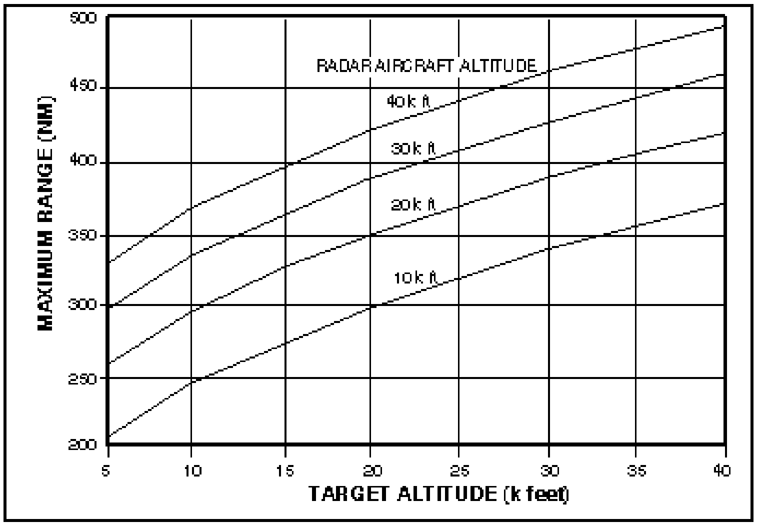

Figure 3. Aircraft Radar vs. Aircraft Target Maximum Range.

Figure 4 depicts the maximum range that a ship height antenna can detect a zero height object (i.e., rowboat).

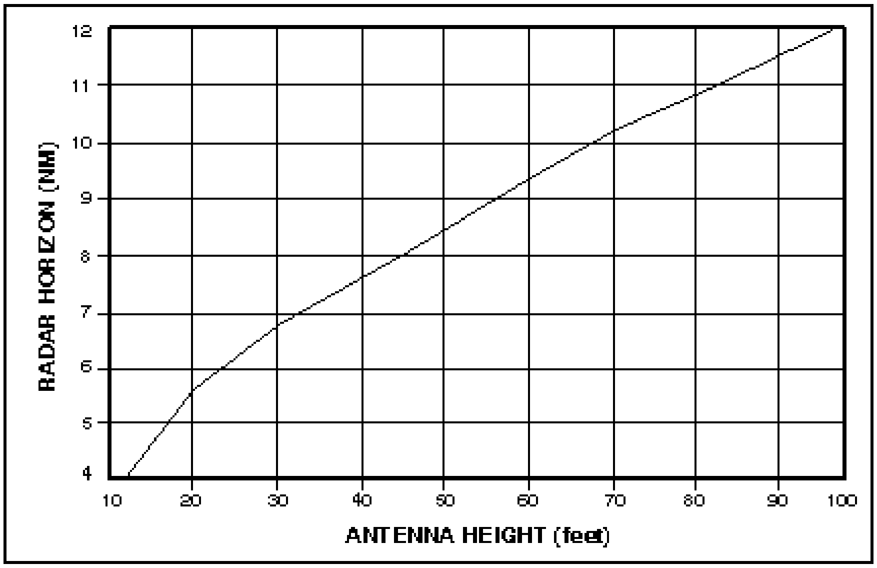

Figure 4. Ships Radar Horizon With Target on the Surface.

In this case "H" = 0, and the general equation becomes: $R_{max}\,(NM) = 1.23\sqrt{h_r}$

Where h_r is the height of the radar in feet.

Figure 5 depicts the same for aircraft radars. It should be noted that most aircraft radars are limited in power output, and would not detect small or surface objects at the listed ranges.

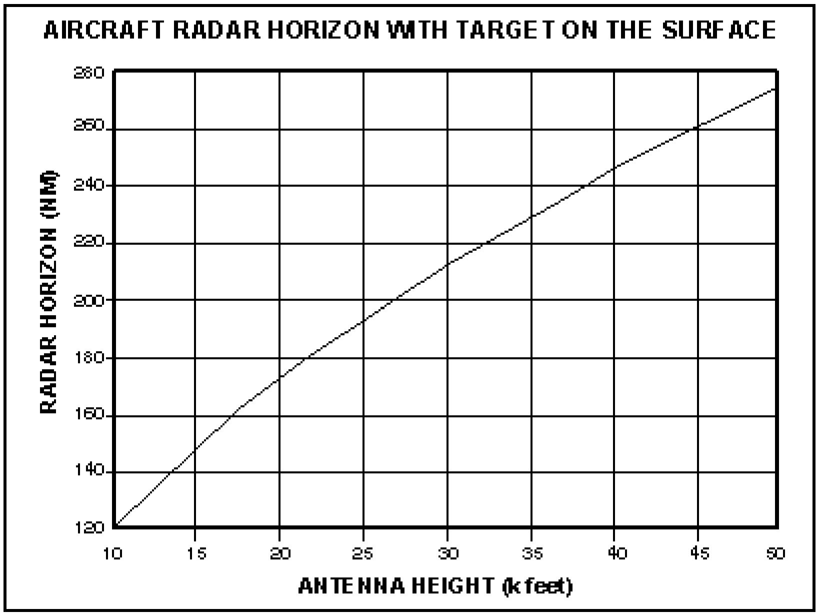

Figure 5. Aircraft Radar Horizon With Target on the Surface.

Other general rules of thumb for surface "targets/radars" are as follows:

For Visual SAR:

$$R_{Visual}(NM) = 1.05\sqrt{Acft\ Alt\ in\ ft}$$

For ESM:

$$R_{ESM}(NM) = 1.5\sqrt{Acft\ Alt\ in\ ft}$$

This page intentionally left blank.

PROPAGATION TIME / RESOLUTION

1. ROUND TRIP RANGE: $R = \frac{c\,t}{2}$ with t = time to reach target

Rules of Thumb

In one μsec round trip time, a wave travels to and from an object at a distance of:	The time it takes to travel to and from an object at a distance of:
≅ 150 m	1 m ≅ 0.0067 μsec
≅ 164 yd	1 yd ≅0.006 μsec
≅ 500 ft	1 ft ≅ 0.002 μsec
≅ 0.08 NM	1 NM ≅ 12.35 μsec
≅ 0.15 km	1 km ≅ 6.7 μsec

2. ONE WAY RANGE: ***R = ct*** with t = time to reach target

Time	Distance Traveled	Distance	Time it Takes
1 milli sec (ms)	165 NM	1 NM	6.18 μsec
1 micro sec (μs)	1000 ft	1 km	3.3 μsec
1 nano sec (ns)	1 ft	1 ft	1 nsec

3. UNAMBIGUOUS RANGE

(DISTANCE BETWEEN PULSES): $R = \frac{c \bullet PRI}{2}$

Normally a radar measures "distance" to the target by measuring time from the last transmitted pulse. If the inter-pulse period (T) is long enough that isn't a problem as shown in "A" to the right. When the period is shortened, the time to the last previous pulse is shorter than the actual time it took, giving a false (ambiguous) shorter range (figure "B").

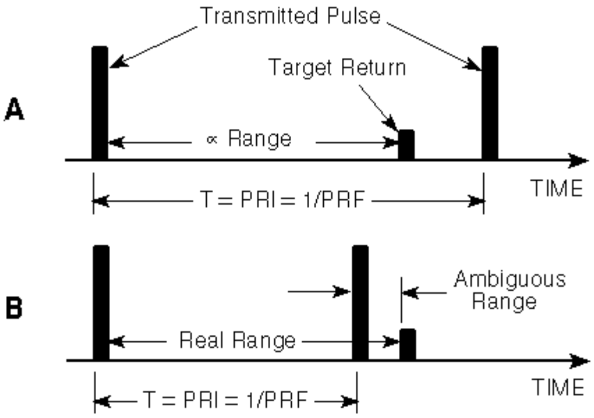

Rules of Thumb

RNM ≅ 81Pms

RKm ≅ 150Pms

Where Pms is PRI in milliseconds

4. RANGE RESOLUTION

Rules of Thumb

500 ft per microsecond of pulse width

500 MHz IF bandwidth provides 1 ft of resolution.

5. BEST CASE PERFORMANCE:

The atmosphere limits the accuracy to 0.1 ft

The natural limit for resolution is one RF cycle.

This page intentionally left blank.

MODULATION

Modulation is the process whereby some characteristic of one wave is varied in accordance with some characteristic of another wave. The basic types of modulation are angular modulation (including the special cases of phase and frequency modulation) and amplitude modulation. In missile radars, it is common practice to amplitude modulate the transmitted RF carrier wave of tracking and guidance transmitters by using a pulsed wave for modulating, and to frequency modulate the transmitted RF carrier wave of illuminator transmitters by using a sine wave.

Frequency Modulation (FM) - As shown in Figure 1, an unmodulated RF signal in the time domain has only a single spectral line at the carrier frequency (f_c) in the frequency domain. If the signal is frequency modulated, as shown in Figure 2 (simplified using only two changes), the spectral line will correspondingly shift in the frequency domain. The bandwidth can be approximated using Carson's rule: $BW = 2(\Delta f + fm)$, where Δf is the peak deviation of the instantaneous frequency from the carrier and fm is the highest frequency present in the modulating signal. There are usually many more "spikes" in the frequency domain than depicted. The number of spikes and shape of the frequency domain envelope (amplitude) are based on the modulation index β. The modulation index is related to the same two factors used in Carson's rule. A high Δf means a higher modulation index with many more "spikes" spread across a wider bandwidth.

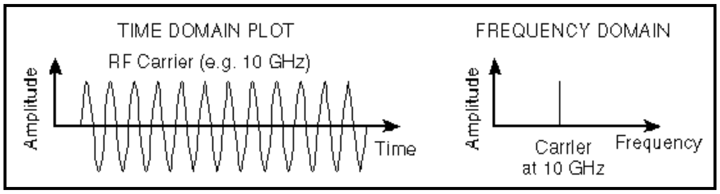
TIME DOMAIN PLOT
RF Carrier (e.g. 10 GHz)
Amplitude
Time
FREQUENCY DOMAIN
Amplitude
Carrier
at 10 GHz
Frequency

Figure 1. Unmodulated RF Signal.

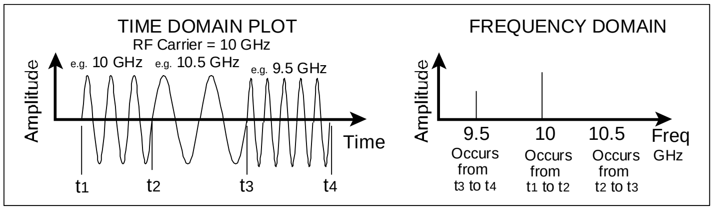
TIME DOMAIN PLOT
RF Carrier = 10 GHz
e.g. 10 GHz
e.g. 10.5 GHz
e.g. 9.5 GHz
Amplitude
Time
t1
t2
t3
t4
FREQUENCY DOMAIN
Amplitude
9.5
Occurs from t3 to t4
10
Occurs from t1 to t2
10.5
Occurs from t2 to t3
Freq GHz

Figure 2. RF Signal With Frequency Modulation.

Amplitude Modulation (AM) - If the signal in Figure 1 is amplitude modulated by a sinewave as shown in Figure 3, sidebands are produced in the frequency domain at $F_c \pm F_{AM}$. AM other than by a pure sine wave will cause additional sidebands normally at $F_c \pm nF_{AM}$, where n equals 1, 2, 3, 4, etc.

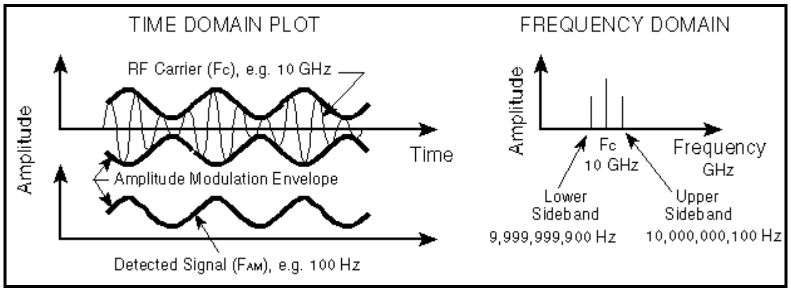
TIME DOMAIN PLOT
RF Carrier (Fc), e.g. 10 GHz
Amplitude
Time
Amplitude Modulation Envelope
Detected Signal (FAM), e.g. 100 Hz
FREQUENCY DOMAIN
Amplitude
Fc
10 GHz
Frequency
GHz
Lower
Sideband
9,999,999,900 Hz
Upper
Sideband
10,000,000,100 Hz

Figure 3. Sinewave Modulated RF Signal.

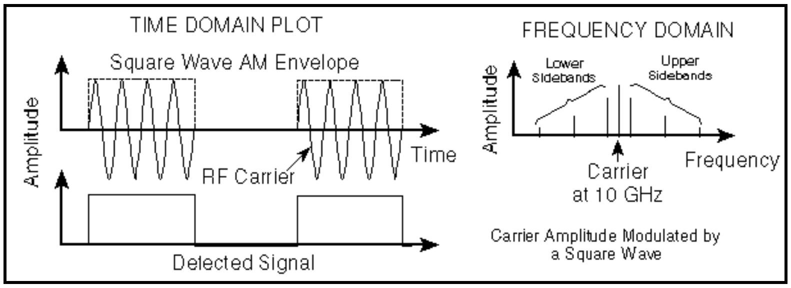
TIME DOMAIN PLOT
Square Wave AM Envelope
Amplitude
Time
RF Carrier
Detected Signal
FREQUENCY DOMAIN
Amplitude
Lower
Sidebands
Upper
Sidebands
Carrier
at 10 GHz
Frequency
Carrier Amplitude Modulated by
a Square Wave

Figure 4. Square Wave Modulated RF Signal (50% Duty Cycle).

Pulse modulation is a special case of AM wherein the carrier frequency is gated at a pulsed rate. When the reciprocal of the duty cycle of the AM is a whole number, harmonics corresponding to multiples of that whole number will be missing, e.g., in a 33.33% duty cycle, AM wave will miss the 3rd, 6th, 9th, etc. harmonics, while a square wave or

50% duty cycle triangular wave will miss the 2nd, 4th, 6th, etc. harmonic, as shown in Figure 4. It has sidebands in the frequency domain at $F_c \pm nF_{AM}$, where n = 1, 3, 5, etc. The amplitude of the power level follows a sine x / x type distribution.

Figure 5 shows the pulse width (PW) in the time domain which defines the lobe width in the frequency domain (Figure 6). The width of the main lobe is 2/PW, whereas the width of a side lobe is 1/PW. Figure 5 also shows the pulse repetition interval (PRI) or its reciprocal, pulse repetition frequency (PRF), in the time domain. In the frequency domain, the spectral lines inside the lobes are separated by the PRF or 1/PRI, as shown in Figures 7 and 8. Note that Figures 7 and 8 show actual magnitude of the side lobes, whereas in Figure 4 and 6, the absolute value is shown.

The magnitude of each spectral component for a rectangular pulse can be determined from the following formula:

$$a_n = 2A\frac{\tau}{T}\frac{\sin(n\pi\tau/T)}{n\pi\tau/T} \quad where: \begin{matrix} \tau = pulse\ width\ (PW) \\ T = period\ (PRI) \end{matrix} \quad and\ \ A = Amplitude\ of\ rectangular\ pulse \qquad [1]$$

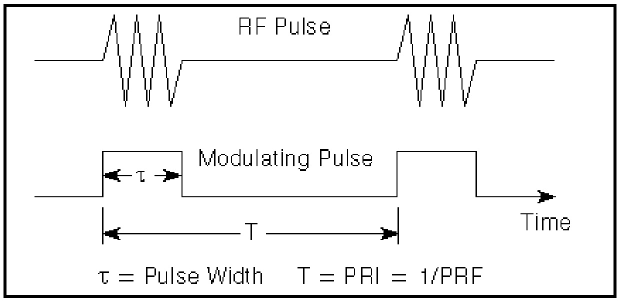

Figure 5. Pulse Width and PRI/PRF Waveforms.

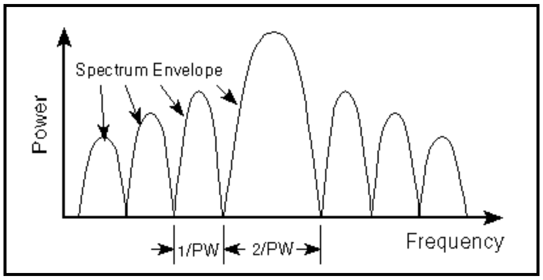

Figure 6. Sidelobes Generated by Pulse Modulation (Absolute Value).

Figure 7 shows the spectral lines for a square wave 50% duty cycle), while Figure 8 shows the spectral lines for a 33.33% duty cycle rectangular wave signal.

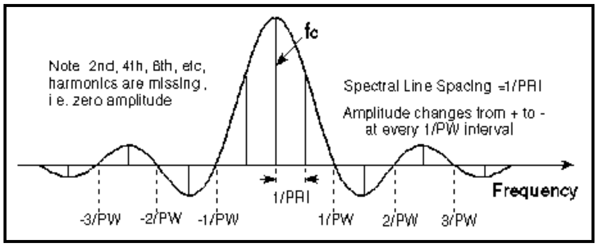

Figure 7. Spectral Times for a Square Wave Modulated Signal.

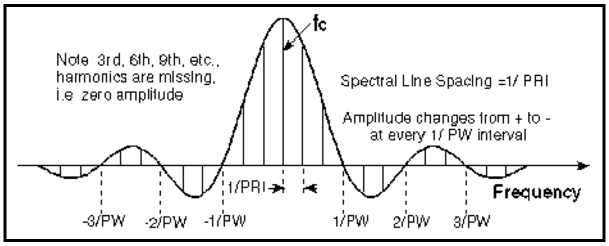

Figure 8. Spectral Lines for a 33.3% Duty Cycle.

Figure 9 shows that for square wave AM, a significant portion of the component modulation is contained in the first few harmonics which comprise the wave. There are twice as many sidebands or spectral lines as there are harmonics (one on the plus and one on the minus side of the carrier). Each sideband represents a sine wave at a frequency equal to the difference between the spectral line and f_c.

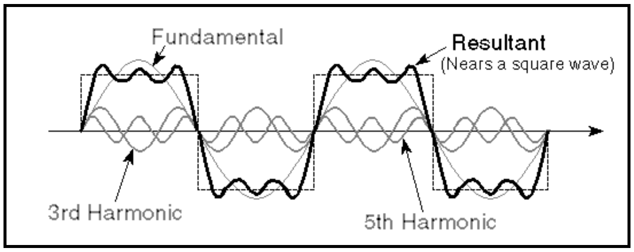

Figure 9. Square Wave Consisting of Sinewave Harmonics.

A figure similar to Figure 9 can be created for any rectangular wave. The relative amplitude of the time domain sine wave components are computed using equation [1]. Each is constructed such that at the midpoint of the pulse the sine wave passes through a maximum (or minimum if the coefficient is negative) at the same time. It should be noted that the "first" harmonic created using this formula is NOT the carrier frequency, f_c, of the modulated signal, but at $F_c \pm F_{AM}$.

While equation [1] is for rectangular waves only, similar equations can be constructed using Fourier coefficients for other waveforms, such as triangular, sawtooth, half sine, trapezoidal, and other repetitive geometric shapes.

PRI Effects - If the PW remains constant but PRI increases, the number of sidelobes remains the same, but the number of spectral lines gets denser (move closer together) and vice versa (compare Figures 7 and 8). The spacing between the spectral lines remains constant with constant PRI.

Pulse Width (PW) Effects - If the PRI remains constant, but the PW increases, then the lobe width decreases and vice versa. If the PW approaches PRI, the spectrum will approach "one lobe," i.e., a single spectral line. The spacing of the lobes remains constant with constant PW.

RF Measurements - If the receiver bandwidth is smaller than the PRF, the receiver will respond to one spectral line at a time. If the receiver bandwidth is wider than the PRF but narrower than the reciprocal of the PW, the receiver will respond to one spectral envelope at a time.

Jet Engine Modulation (JEM)

Section 2-6 addresses the Doppler shift in a transmitted radar signal caused by a moving target. The amount of Doppler shift is a function of radar carrier frequency and the speed of the radar and target. Moving or rotating surfaces on the target will have the same Doppler shift as the target, but will also impose AM on the Doppler shifted return (see Figure 10). Reflections off rotating jet engine compressor blades, aircraft propellers, ram air turbine (RAT) propellers used to power aircraft pods, helicopter rotor blades, and protruding surfaces of automobile hubcaps will all provide a chopped reflection of the impinging signal. The reflections are characterized by both positive and negative Doppler sidebands corresponding to the blades moving toward and away from the radar respectively.

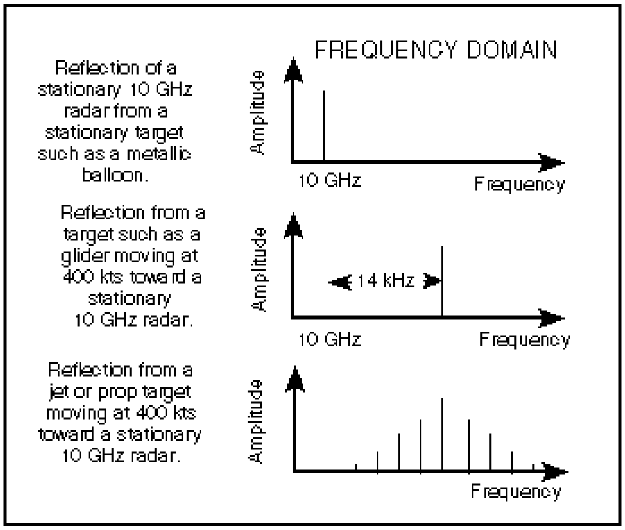
FREQUENCY DOMAIN
Reflection of a stationary 10 GHz radar from a stationary target such as a metallic balloon.
Amplitude
10 GHz
Frequency
Reflection from a target such as a glider moving at 400 kts toward a stationary 10 GHz radar.
Amplitude
14 kHz
10 GHz
Frequency
Reflection from a jet or prop target moving at 400 kts toward a stationary 10 GHz radar.
Amplitude
Frequency

Figure 10. Doppler Return and JEM.

Therefore, forward/aft JEM does not vary with radar carrier frequency, but the harmonics contained in the sidebands are a function of the PRF of the blade chopping action and its amplitude is target aspect dependent, i.e., blade angle, intake/exhaust internal reflection, and jet engine cowling all effect lateral return from the side. If the aspect angle is too far from head-on or tail-on and the engine cowling provides shielding for the jet engine, there may not be any JEM to detect. On the other hand, JEM increases when you are orthogonal (at a right angle) to the axis of blade rotation. Consequently for a fully exposed blade as in a propeller driven aircraft or helicopter, JEM increases with angle off the boresight axis of the prop/rotor.

This page intentionally left blank.

TRANSFORMS / WAVELETS

Transform Analysis

Signal processing using a transform analysis for calculations is a technique used to simplify or accelerate problem solution. For example, instead of dividing two large numbers, we might convert them to logarithms, subtract them, then look-up the anti-log to obtain the result. While this may seem a three-step process as opposed to a one-step division, consider that long-hand division of a four digit number by a three digit number, carried out to four places requires three divisions, 3-4 multiplication's, and three subtractions. Computers process additions or subtractions much faster than multiplications or divisions, so transforms are sought which provide the desired signal processing using these steps.

Fourier Transform

Other types of transforms include the Fourier transform, which is used to decompose or separate a waveform into a sum of sinusoids of different frequencies. It transforms our view of a signal from time based to frequency based. Figure 1 depicts how a square wave is formed by summing certain particular sine waves. The waveform must be continuous, periodic, and almost everywhere differentiable. The Fourier transform of a sequence of rectangular pulses is a series of sinusoids. The envelope of the amplitude of the coefficients of this series is a waveform with a Sin X/X shape. For the special case of a single pulse, the Fourier series has an infinite series of sinusoids that are present for the duration of the pulse.

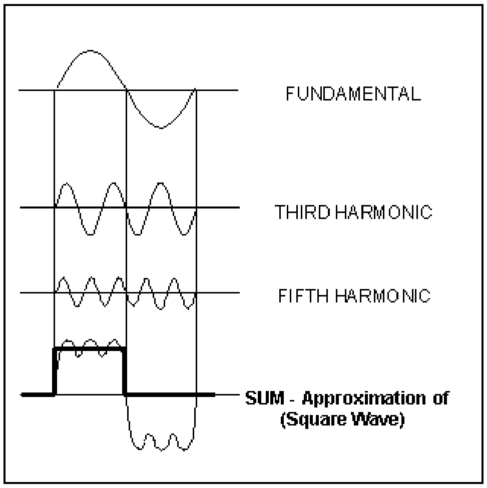

Figure 1. Harmonics.

Digital Sampling of Waveforms

To process a signal digitally, we need to sample the signal frequently enough to create a complete "picture" of the signal. The discrete Fourier transform (DFT) may be used in this regard. Samples are taken at uniform time intervals as shown in Figure 2 and processed.

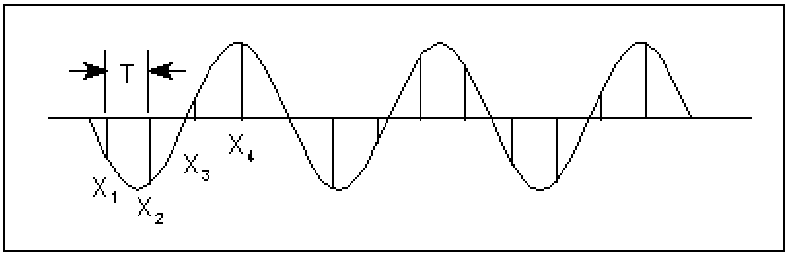

Figure 2. Waveform Sampling.

If the digital information is multiplied by the Fourier coefficients, a digital filter is created as shown Figure 3. If the sum of the resultant components is zero, the filter has ignored (notched out) that frequency sample. If the sum is a relatively large number, the filter has passed the signal. With the single sinusoid shown, there should be only one resultant. (Note that being "zero" and relatively large

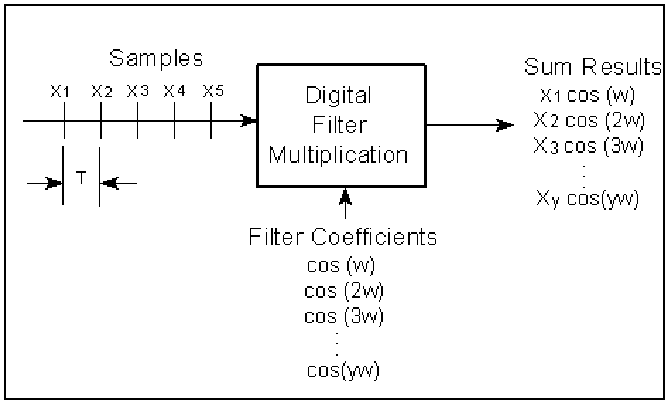

Figure 3. Digital Filtering.

may just mean below or above the filter's cutoff threshold.)

Figure 4 depicts the process pictorially: The vectors in the figure just happen to be pointing in a cardinal direction because the strobe frequencies are all multiples of the vector (phasor) rotation rate, but that is not normally the case. Usually the vectors will point in a number of different directions, with a resultant in some direction other than straight up.

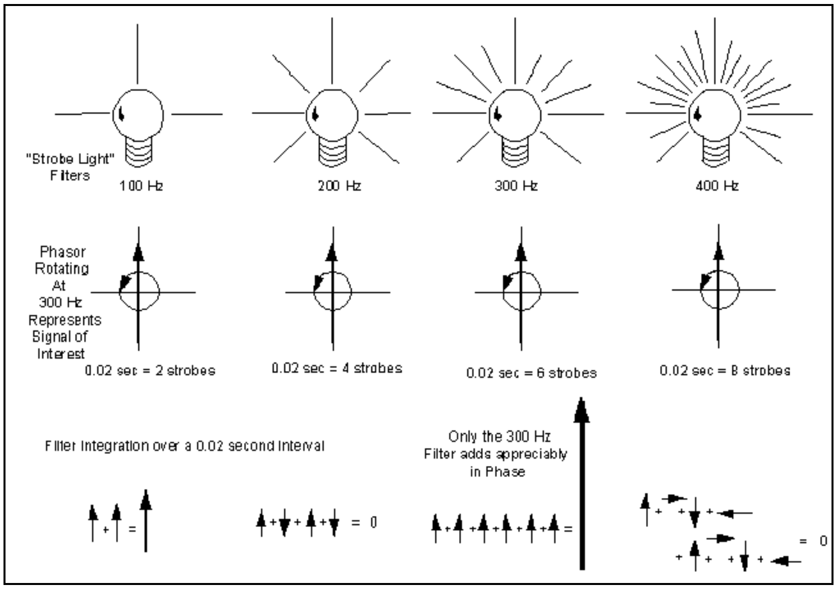
"Strobe Light" Filters
100 Hz
200 Hz
300 Hz
400 Hz
Phasor Rotating At 300 Hz Represents Signal of Interest
0.02 sec = 2 strobes
0.02 sec = 4 strobes
0.02 sec = 6 strobes
0.02 sec = 8 strobes
Filter Integration over a 0.02 second Interval
Only the 300 Hz Filter adds appreciably in Phase
+ =
+ + + = 0
+ + + + + =
+ + + + + + + = 0

Figure 4. Phasor Representation.

In addition, sampling normally has to taken at or above twice the rate of interest (also known as the Nyquist rate), otherwise ambiguous results may be obtained. Figure 4 is under-sampled (for clarity) and consequently does not depict typical filtering.

Fast Fourier Transforms

One problem with this type of processing is the large number of additions, subtractions, and multiplications which are required to reconstruct the output waveform. The Fast Fourier transform (FFT) was developed to reduce this problem. It recognizes that because the filter coefficients are sine and cosine waves, they are symmetrical about 90, 180, 270, and 360 degrees. They also have a number of coefficients equal either to one or zero, and duplicate coefficients from filter to filter in a multibank arrangement. By waiting for all of the inputs for the bank to be received, adding together those inputs for which coefficients are the same before performing multiplications, and separately summing those combinations of inputs and products which are common to more than one filter, the required amount of computing may be cut drastically.

- The number of computations for a DFT is on the order of N squared.

- The number of computations for a FFT when N is a power of two is on the order of N $\log_2$ N.

For example, in an eight filter bank, a DFT would require 512 computations, while an FFT would only require 56, significantly speeding up processing time.

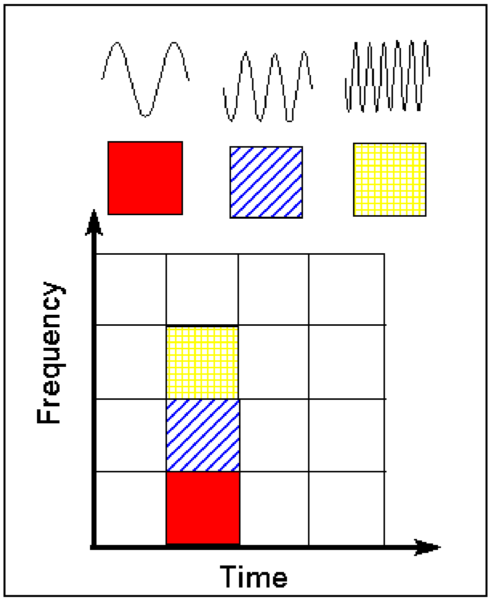
Frequency
Time

Figure 5. Windowed Fourier Transform.

Windowed Fourier Transform

The Fourier transform is continuous, so a windowed Fourier transform (WFT) is used to analyze non-periodic signals as shown in Figure 5. With the WFT, the signal is divided into sections (one such section is shown in Figure 5) and each section is analyzed for frequency content. If the signal has sharp transitions, the input data is windowed so that the sections converge to zero at the endpoints. Because a single window is used for all frequencies in the WFT, the resolution of the analysis is the same (equally spaced) at all locations in the time-frequency domain.

The FFT works well for signals with smooth or uniform frequencies, but it has been found that other transforms work better with signals having pulse type characteristics, time-varying (non-stationary) frequencies, or odd shapes.

The FFT also does not distinguish sequence or timing information. For example, if a signal has two frequencies (a high followed by a low or vice versa), the Fourier transform only reveals the frequencies and relative amplitude, not the order in which they occurred. So Fourier analysis works well with stationary, continuous, periodic, differentiable signals, but other methods are needed to deal with non-periodic or non-stationary signals.

Wavelet Transform

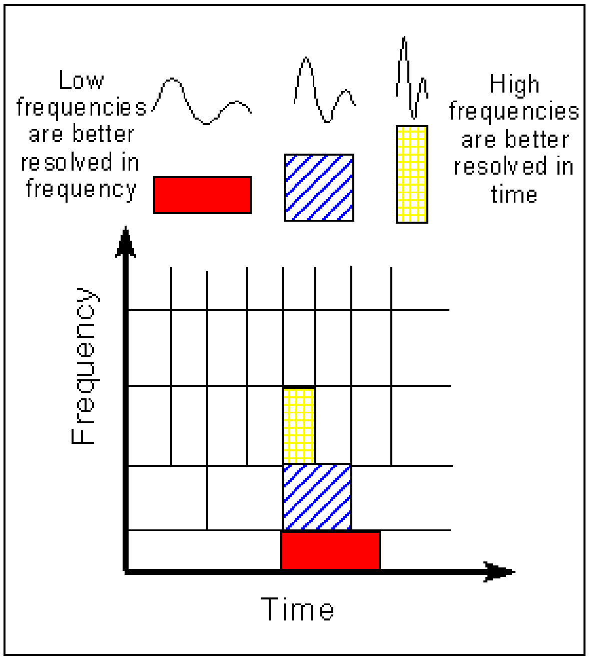
Low frequencies are better resolved in frequency
High frequencies are better resolved in time
Frequency
Time

Figure 6. Wavelet Transform.

The Wavelet transform has been evolving for some time. Mathematicians theorized its use in the early 1900's. While the Fourier transform deals with transforming the time domain components to frequency domain and frequency analysis, the wavelet transform deals with scale analysis, that is, by creating mathematical structures that provide varying time/frequency/amplitude slices for analysis. This transform is a portion (one or a few cycles) of a complete waveform, hence the term wavelet.

The wavelet transform has the ability to identify frequency (or scale) components, simultaneously with their location(s) in time. Additionally, computations are directly proportional to the length of the input signal. They require only N multiplications (times a small constant) to convert the waveform. For the previous eight filter bank example, this would be about twenty calculations, vice 56 for the FFT.

In wavelet analysis, the scale that one uses in looking at data plays a special role. Wavelet algorithms process data at different scales or resolutions. If we look at a signal with a large "window," we would notice gross features. Similarly, if we look at a signal with a small "window," we would notice small discontinuities as shown in Figure 6. The result in wavelet analysis is to "see the forest and the trees." A way to achieve this is to have short high-frequency fine scale functions and long low-frequency ones. This approach is known as multi-resolution analysis.

For many decades, scientists have wanted more appropriate functions than the sines and cosines (base functions) which comprise Fourier analysis, to approximate choppy signals. (Although Walsh transforms work if the waveform is periodic and stationary). By their definition, sine and cosine functions are non-local (and stretch out to infinity), and therefore do a very poor job in approximating sharp spikes. But with wavelet analysis, we can use approximating functions that are contained neatly in finite (time/frequency) domains. Wavelets are well-suited for approximating data with sharp discontinuities.

The wavelet analysis procedure is to adopt a wavelet prototype function, called an "analyzing wavelet" or "mother wavelet." Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, while frequency analysis is performed with a dilated, low-frequency version of the prototype wavelet. Because the original signal or function can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet functions), data operations can be performed using just the corresponding wavelet coefficients as shown in Figure 7.

If one further chooses the best wavelets adapted to the data, or truncates the coefficients below some given threshold, the data is sparsely represented. This "sparse coding" makes wavelets an excellent tool in the field of data compression. For instance, the FBI uses wavelet coding to store fingerprints. Hence, the concept of wavelets is to look at a signal at various scales and analyze it with various resolutions.

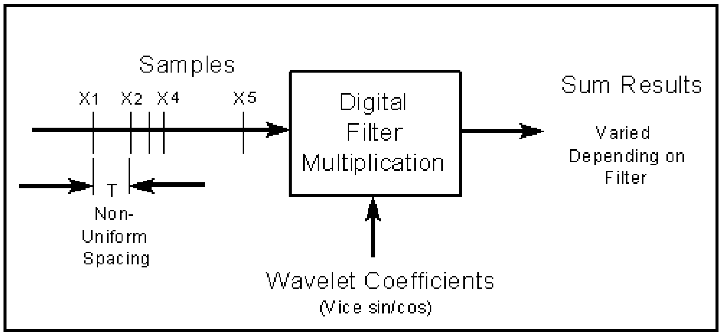

Figure 7. Wavelet Filtering.

Analyzing Wavelet Functions

Fourier transforms deal with just two basis functions (sine and cosine), while there are an infinite number of wavelet basis functions. The freedom of the analyzing wavelet is a major difference between the two types of analyses and is important in determining the results of the analysis. The "wrong" wavelet may be no better (or even far worse than) than the Fourier analysis. A successful application presupposes some expertise on the part of the user. Some prior knowledge about the signal must generally be known in to select the most suitable distribution and adapt the parameters to the signal. Some

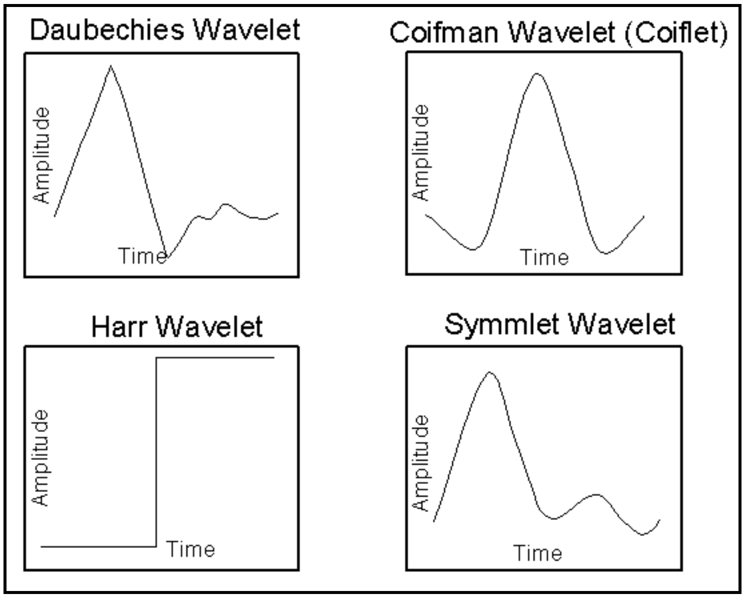

Figure 8. Sample Wavelet Functions.

of the more common ones are shown in Figure 8. There are several wavelets in each family, and they may look different than those shown. Somewhat longer in duration than these functions, but significantly shorter than infinite sinusoids is the cosine packet shown in Figure 9.

Wavelet Comparison With Fourier Analysis

While a typical Fourier transform provides frequency content information for samples within a given time interval, a perfect wavelet transform records the start of one frequency (or event), then the start of a second event, with amplitude added to or subtracted from, the base event.

Example 1.

Wavelets are especially useful in analyzing transients or time-varying signals. The input signal shown in Figure 9 consists of a sinusoid whose frequency changes in stepped increments over time. The power of the spectrum is also shown. Classical Fourier analysis will resolve the frequencies but cannot provide any information about the times at which each occurs. Wavelets provide an efficient means of analyzing the input signal so that frequencies and the times at which they occur can be resolved. Wavelets have finite duration and must also satisfy additional properties beyond those normally associated with standard windows used with Fourier analysis. The result after the wavelet transform is applied is the plot shown in the lower right. The wavelet analysis correctly resolves each of the frequencies and the time when it occurs. A series of wavelets is used in example 2.

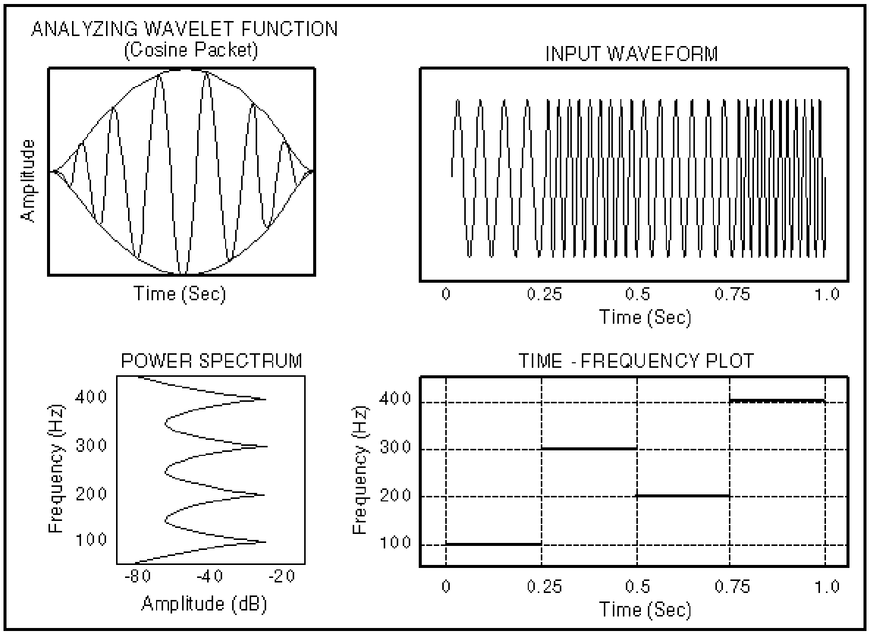

Figure 9. Sample Wavelet Analysis.

Example 2.

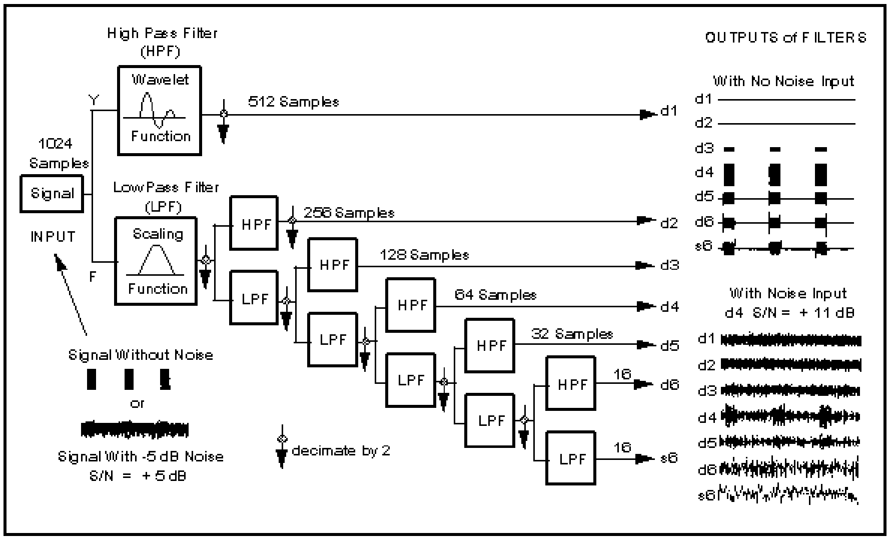
High Pass Filter
(HPF)
Wavelet
Function
Y
512 Samples
d1
1024
Samples
Signal
INPUT
Low Pass Filter
(LPF)
Scaling
Function
F
HPF
256 Samples
d2
LPF
HPF
128 Samples
d3
LPF
HPF
64 Samples
d4
LPF
HPF
32 Samples
d5
LPF
HPF
16
d6
LPF
16
s6
Signal Without Noise
or
Signal With -5 dB Noise
S/N = +5 dB
decimate by 2
OUTPUTS of FILTERS
With No Noise Input
d1
d2
d3
d4
d5
d6
s6
With Noise Input
d4 S/N = +11 dB
d1
d2
d3
d4
d5
d6
s6

Figure 10. Example 2 Analysis Wavelet.

Figure 10 shows the input of a clean signal, and one with noise. It also shows the output of a number of "filters" with each signal. A 6 dB S/N improvement can be seen from the d4 output. (Recall from Section 4.3 that 6 dB corresponds to doubling of detection range.) In the filter cascade, the HPFs and LPFs are the same at each level. The wavelet shape is related to the HPF and LPF in that it is the "impulse response" of an infinite cascade of the HPFs and LPFs. Different wavelets have different HPFs and LPFs. As a result of decimating by 2, the number of output samples equals the number of input samples.

Wavelet Applications

Some fields that are making use of wavelets are: astronomy, acoustics, nuclear engineering, signal and image processing (including fingerprinting), neurophysiology, music, magnetic resonance imaging, speech discrimination, optics, fractals, turbulence, earthquake-prediction, radar, human vision, and pure mathematics applications.

See Andrew Bruce, David Donoho, and Hong-Ye Gao, "Wavelet Analysis," *IEEE Spectrum*, Vol. 33 No. 10, October 1996.

ANTENNAS

Antenna Introduction / Basics 3-1
Polarization 3-2
Radiation Patterns 3-3
Frequency / Phase Effects of Antennas 3-4
Antenna Near Field 3-5
Radiation Hazards 3-6
Active Electronically Scanned Arrays (AESA) 3-7
Fractal Antennas 3-8

This page intentionally left blank.

ANTENNA INTRODUCTION / BASICS

Rules of Thumb:

1. The power gain of an antenna with losses, excluding input impedance mismatch, is given by:

$$G \approx \frac{4\pi\eta A}{\lambda^2} \quad \text{Where} \quad \begin{aligned} \eta &= \text{Efficiency} \\ A &= \text{Physical aperture area} \\ \lambda &= \text{wavelength} \end{aligned}$$

another is :

$$G = \frac{X\eta}{BW_\phi BW_\theta}$$

Where $BW_{\theta \text{ and } \phi}$ *are the elev & az beamwidths in degrees.*

For approximating an antenna pattern with :

(1) A rectangle; $X = 41253, \eta_{typical} = 0.7$

(2) An ellipsoid; $X = 52525, \eta_{typical} = 0.55$

2. Directive gain of rectangular X-Band Aperture
   $G = 1.4\ LW$ Where: Length (L) and Width (W) are in cm

3. Power gain of Circular X-Band Aperture
   $G = d^2\eta$ Where: d = antenna diameter in cm
   η = aperture efficiency

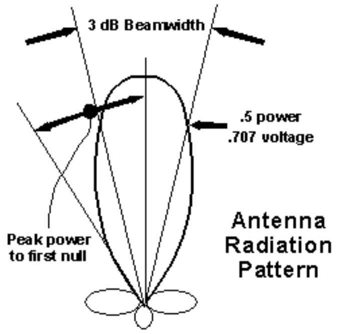

4. Directive gain of an imaginary isotropic antenna radiating in a uniform spherical pattern is one (0 dB).

5. Antenna with a 20 degree beamwidth has approximately a 20 dB directive gain.

6. 3 dB beamwidth is approximately equal to the angle from the peak of the power to the first null (see figure at right).

7. Parabolic Antenna Beamwidth: $$BW = \frac{70\lambda}{d}$$

   Where: BW = antenna beamwidth; λ = wavelength; d = antenna diameter.

The antenna equations which follow relate to Figure 1 as a typical antenna. In Figure 1, BW_φ is the azimuth beamwidth and BW_θ is the elevation beamwidth. Beamwidth is normally measured at the half-power or -3 dB point of the main lobe unless otherwise specified. See Glossary.

The gain or directivity of an antenna is the ratio of the radiation intensity in a given direction to the radiation intensity averaged over all directions.

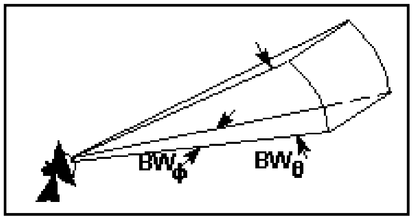

Figure 1. Antenna Aperture.

Quite often directivity and gain are used interchangeably and it sometimes leads to overly optimistic antenna performance estimations. The difference is that directivity is based solely on antenna pattern shape estimation where antenna losses such as dielectric, ohmic resistance, and polarization mismatch are neglected. If these losses are included in the antenna gain calculations, the antenna gain is then referred to as the power gain. Moreover, if additional impedance mismatch or VSWR losses are included in the antenna system gain estimation, the antenna gain calculations are then referred to as the realized gain. However, using directive gain (or directivity) calculations is very convenient in practice for a first order idealized antenna performance estimation.

Normalizing a radiation pattern by the integrated total power yields the directivity of the antenna. This concept in shown in equation form by:

$$D(\theta,\phi) = 10\ \mathrm{Log}\left[\frac{4\pi P\ (\theta,\phi)}{\int\int P_{in}(\theta,\phi)\,\mathrm{Sin}\ \theta\ d\theta\ d\phi}\right] \quad \begin{matrix} 0 < \phi \le 360° \\ 0 < \theta \le 180° \end{matrix} \qquad \textbf{[1]}$$

Where D(θ,φ) is the directivity in dB, and the radiation pattern power in a specific direction is $P_d(\theta,\varphi)$, which is normalized by the total integrated radiated power. Another important concept is that when the angle in which the radiation is constrained is reduced, the directive gain goes up. For example, using an isotropic radiating source, the gain would be 0 dB by definition (Figure 2(a)) and the power density (P_d) at any given point would be the power in (P_{in}) divided by the surface area of the imaginary sphere at a distance R from the source. If the spacial angle was decreased to one hemisphere (Figure 2(b)), the power radiated, P_{in}, would be the same but the area would be half as much, so the gain would double to 3 dB. Likewise if the angle is a quarter sphere, (Figure 2(c)), the gain would be 6 dB. Figure 2(d) shows a pencil beam. The gain is independent of actual power output and radius (distance) at which measurements are taken.

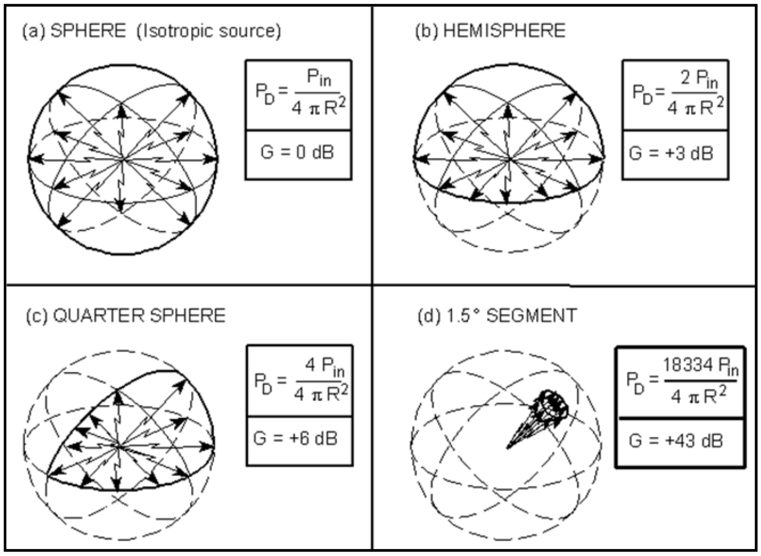
(a) SPHERE (Isotropic source)
$P_D = \frac{P_{in}}{4\pi R^2}$
G = 0 dB
(b) HEMISPHERE
$P_D = \frac{2 P_{in}}{4\pi R^2}$
G = +3 dB
(c) QUARTER SPHERE
$P_D = \frac{4 P_{in}}{4\pi R^2}$
G = +6 dB
(d) 1.5° SEGMENT
$P_D = \frac{18334 P_{in}}{4\pi R^2}$
G = +43 dB

Figure 2. Notional Representation of Directive Antenna Gain.

Real antennas are different, however, and do not have an ideal radiation distribution. Energy varies with angular displacement and losses occur due to sidelobes. However, if we can measure the pattern, and determine the beamwidth we can use two (or more) ideal antenna models to approximate a real antenna pattern as shown in Figure 3.

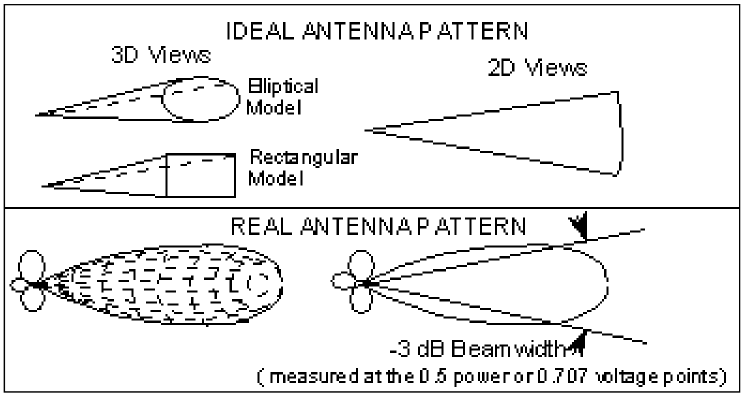
IDEAL ANTENNA PATTERN
3D Views
Elliptical Model
Rectangular Model
2D Views
REAL ANTENNA PATTERN
-3 dB Beamwidth
(measured at the 0.5 power or 0.707 voltage points)

Figure 3. Antenna Beamwidth.

Assuming the antenna pattern is uniform, the gain is equal to the area of the isotropic sphere ($4\pi r^2$) divided by the sector (cross section) area.

$$G = \frac{Area\ of\ Sphere}{Area\ of\ Antenna\ pattern} \qquad \textbf{[2]}$$

It can be shown that:

$$G \approx \frac{4\pi}{BW_{\phi\,az}\,BW_{\theta\,el}} \quad or \quad \frac{4\pi}{\phi\theta\ (radians)} \quad where: \begin{matrix} BW_{\phi\,az} = \text{Azmith beamwidth in radians} \\ BW_{\theta\,el} = \text{Elevation beamwidth in radians} \end{matrix} \qquad \textbf{[3]}$$

From this point, two different models are presented:

(1) Approximating an antenna pattern using an elliptical area, and
(2) Approximating an antenna pattern using a rectangular area.

Approximating the antenna pattern as an **elliptical** area:

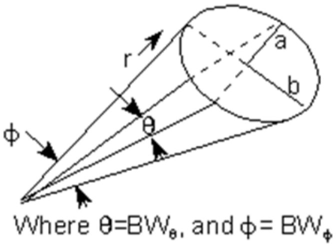

Area of ellipse = $\pi a b = \pi[(r \sin\theta)/2][(r \sin\varphi)/2] = (\pi r^2 \sin\theta \sin\varphi)/4$

$$G = \frac{Area\ of\ Sphere}{Area\ of\ Antenna\ pattern} = (4\pi r^2)\left(\frac{4}{\pi r^2 \sin\theta \sin\phi}\right) = \frac{16}{\sin\theta \sin\phi}$$

For small angles, $\sin\varphi = \varphi$ in radians, so:

$$G = \frac{16}{\sin\phi \sin\theta} \approx \frac{16}{\phi\,\theta\,(radians)} = \frac{16}{\phi\,\theta}\left(\frac{360°\,360°}{2\pi\,2\pi}\right) = \frac{52525}{\phi\,\theta\,(degrees)}\ or\ \frac{52525}{BW_\phi\,BW_\theta\,(degrees)} \quad \textbf{[4]}$$

The second term in the equation above is very close to equation [3].

For a very directional radar dish with a beamwidth of 1° and an average efficiency of 55%:

Ideally: G = 52525, or in dB form: 10 log G =10 log 52525 = 47.2 dB

With efficiency taken into account, G = 0.55(52525) = 28888, or in log form: 10 log G = 44.6 dB

Approximating the antenna pattern as a **rectangular** area:

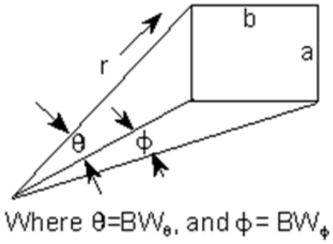

$a = r \sin\theta$, $b = r \sin\varphi$, area = $ab = r^2 \sin\theta \sin\varphi$

$$G = \frac{Area\ of\ Sphere}{Area\ of\ Antenna\ pattern} = \frac{4\pi r^2}{r^2 \sin\theta \sin\phi} = \frac{4\pi}{\sin\theta \sin\phi}$$

For small angles, $\sin\varphi = \varphi$ in radians, so:
The second term in the equation above is identical to equation [3].

$$G = \frac{4\pi}{\sin\phi \sin\theta} = \frac{4\pi}{\phi\,\theta\,(radians)} = \frac{4\pi}{\phi\,\theta}\left(\frac{360°\,360°}{2\pi\,2\pi}\right) = \frac{41253}{\phi\,\theta\,(degrees)}\ or\ \frac{41253}{BW_\phi\,BW_\theta\,(degrees)} \quad \textbf{[5]}$$

Converting to dB, $G_{\max}(dB) = 10\ \mathrm{Log}\left[\frac{41253}{BW_\phi\,BW_\theta}\right]$ with BW_ϕ and BW_θ in degrees **[6]**

For a very directional radar dish with a beamwidth of 1° and an average efficiency of 70%:
Ideally (in dB form): 10 log G =10 log 41253 = 46.2 dB.
With efficiency taken into account, G = 0.7(41253) = 28877, or in log form: 10 log G = 44.6 dB

Comparison between elliptical and rectangular areas for antenna pattern models:

By using the rectangular model there is a direct correlation between the development of gain in equation [5] and the ideal gain of equation [3]. The elliptical model has about one dB difference from the ideal calculation, but will yield the same real antenna gain when appropriate efficiencies are assumed.

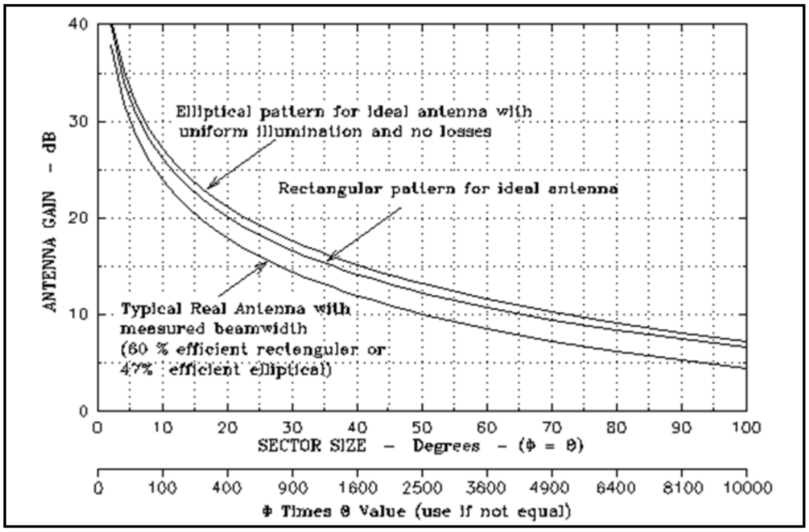

Figure 4. Antenna Sector Size vs. Gain.

The upper plot of Figure 4 shows the gain for an ideal antenna pattern using the elliptical model. The middle plot shows the gain for an ideal antenna using the rectangular model. The lower plot of Figure 4 shows the gain of a typical real antenna (rectangular model using an efficiency of 70% or elliptical model using an efficiency of 47%).

Gain as a function of λ:

When θ = 0, each wave source in Figure 5 is in phase with one another and a maximum is produced in that direction.

Conversely, nulls to either side of the main lobe will occur when the waves radiating from the antenna cancel each other. The first null occurs when there is a phase difference of λ/2 in the wave fronts emanating from the aperture. To aid in visualizing what happens, consider each point in the antenna aperture, from A to C in Figure 5, as a point source of a spherical wave front. If viewed from infinity, the electromagnetic waves from each point interfere with each other, and when, for a particular direction, θ in Figure 5, each wave source has a corresponding point that is one-half wavelength out of phase, a null is produced in that direction due to destructive interference.

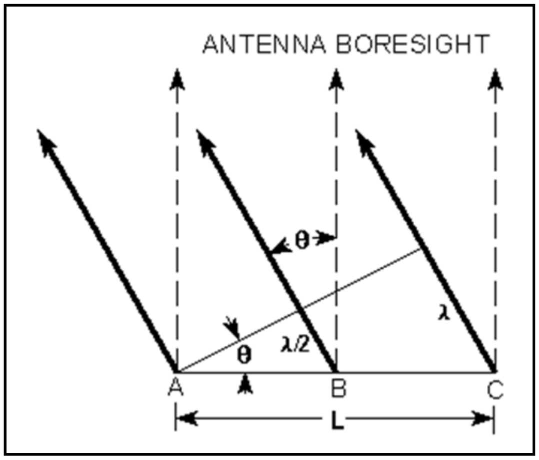

Figure 5. Directional Gain vs. Wavelength.

In Figure 5, the wave emanating from point A is out of phase with the wave from point B by one-half of a wavelength. Hence, they cancel. Similarly, a point just to the right of point A cancels with a point just to the right of point B, and so on across the entire aperture. Therefore, the first null in the radiation pattern is given by:

$$\operatorname{Sin}\theta = \lambda/L \text{ and, in radians, } \theta = \lambda/L \text{ (for small angles)} \qquad \textbf{[7]}$$

As the angle off boresight is increased beyond the first null, the intensity of the radiation pattern rises then falls, until the second null is reached. This corresponds to a phase difference of two wavelengths between the left and right edges of the aperture. In this case, the argument proceeds as before, except now the aperture is divided into four segments (point A canceling with a point halfway between A and B, and so on).

The angle θ is the angle from the center (maximum) of the radiation pattern to the first null. The null-to-null beam width is 2θ. Generally, we are interested in the half-power (3 dB) beamwidth. It turns out that this beamwidth is approximately one-half of the null-to-null beamwidth, so that:

$$BW_{3\,dB} \approx (½)(2\theta) = \lambda/L \qquad \textbf{[8]}$$

Therefore, beamwidth is a function of the antenna dimension "L" and the wavelength of the signal. It can be expressed as follows: Note: for circular antennas, L in the following equations = diameter

$$Bw_{\varphi(az)} = \lambda/L_{Az\ eff} \text{ and } BW_{\theta(el)} = \lambda/L_{El\ eff} \qquad \textbf{[9]}$$

Substituting the two variations of equation [9] into equation [3] and since $L_{Az\ eff}$ times $L_{El\ eff}$ = A_e (effective capture area of the antenna), we have:

$$G \approx \frac{4\pi}{BW_\phi\, BW_\theta\,(radians)} = \frac{4\pi\ L_{az}\ L_{el}}{\lambda^2} = \frac{4\pi\ A_e}{\lambda^2} \qquad \textbf{[10]}$$

Note: Equation is approximate since aperture efficiency isn't included as is done later in equation [12].

The efficiency (discussed later) will reduce the gain by a factor of 30-50%, i.e. real gain = .5 to .7 times theoretical gain.

Unity Gain Antenna.
If a square antenna is visualized and G=1, $A_e = \lambda^2 / 4\pi$. When a dimension is greater than 0.28 λ (~¼λ) it is known as an electrically large antenna, and the antenna will have a gain greater than one (positive gain when expressed in dB). Conversely, when the dimension is less than 0.28 λ (~¼λ)(an electrically small antenna), the gain will be less than one (negative gain when expressed in dB). Therefore, a unity gain antenna can be approximated by an aperture that is ¼λ by ¼λ.

Beamwidth as a Function of Aperture Length
It can be seen from Figure 5, that the wider the antenna aperture (L), the narrower the beamwidth will be for the same λ. Therefore, if you have a rectangular shaped horn antenna, the radiation pattern from the wider side will be narrower than the radiation pattern from the narrow side.

APERTURE EFFICIENCY, η

The Antenna Efficiency, η, is a factor which includes all reductions from the maximum gain. η can be expressed as a percentage, or in dB. Several types of "loss" must be accounted for in the efficiency, η:

(1) Illumination efficiency which is the ratio of the directivity of the antenna to the directivity of a uniformly illuminated antenna of the same aperture size,
(2) Phase error loss or loss due to the fact that the aperture is not a uniform phase surface,
(3) Spillover loss (Reflector Antennas) which reflects the energy spilling beyond the edge of the reflector into the back lobes of the antenna,

(4) Mismatch (VSWR) loss, derived from the reflection at the feed port due to impedance mismatch (especially important for low frequency antennas), and

(5) RF losses between the antenna and the antenna feed port or measurement point.

The aperture efficiency, η_a, is also known as the illumination factor, and includes items (1) and (2) above; it does not result in any loss of power radiated but affects the gain and pattern. It is nominally 0.6-0.8 for a planer array and 0.13 to 0.8 with a nominal value of 0.5 for a parabolic antenna, however η can vary significantly. Other antennas include the spiral (.002-.5), the horn (.002-.8), the double ridge horn (.005-.93), and the conical log spiral (.0017-1.0).

Items (3), (4), and (5) above represent RF or power losses which can be measured. The efficiency varies and generally gets lower with wider bandwidths. Also note that the gain equation is optimized for small angles - see derivation of wavelength portion of equation [7]. This explains why efficiency also gets lower for wider beamwidth antennas.

EFFECTIVE CAPTURE AREA

Effective capture area (A_e) is the product of the physical aperture area (A) and the aperture efficiency (η) or:

$$A_e = \eta A = \frac{\lambda^2 G}{4\pi} \qquad \textbf{[11]}$$

GAIN AS A FUNCTION OF APERTURE EFFICIENCY

The Gain of an antenna with losses is given by:

$$G = \frac{4\pi\eta A}{\lambda^2} \quad \begin{array}{c} \eta = \text{Aperture Efficiency} \\ \text{Where } A = \text{Physical aperture area} \\ \lambda = \text{wavelength} \end{array} \qquad \textbf{[12]}$$

Note that the gain is proportional to the aperture area and inversely proportional to the square of the wavelength. For example, if the frequency is doubled, (half the wavelength), the aperture could be decreased four times to maintain the same gain.

BEAM FACTOR

Antenna size and beamwidth are also related by the beam factor defined by:

Beam Factor = (D/λ)·(Beamwidth)   where D = antenna dimension in wavelengths.

The beam factor is approximately invariant with antenna size, but does vary with type of antenna aperture illumination or taper. The beam factor typically varies from 50-70°.

APERTURE ILLUMINATION (TAPER)

The aperture illumination or illumination taper is the variation in amplitude across the aperture. This variation can have several effects on the antenna performance:

(1) reduction in gain,
(2) reduced (lower) sidelobes in most cases, and
(3) increased antenna beamwidth and beam factor.

Tapered illumination occurs naturally in reflector antennas due to the feed radiation pattern and the variation in distance from the feed to different portions of the reflector. Phase can also vary across the aperture which also affects the gain, efficiency, and beamwidth.

CIRCULAR ANTENNA GAIN

Solving equation [12] in dB, for a circular antenna with area $\pi D^2/4$, we have:

$$10 \text{ Log } G = 20 \text{ Log } (D/\lambda) + 10 \text{ Log } (\eta) + 9.94 \text{ dB}; \quad \text{where } D = \text{diameter} \qquad \textbf{[13]}$$

This data is depicted in the nomograph of Figure 6. For example, a six foot diameter antenna operating at 9 GHz would have approximately 44.7 dB of gain as shown by the dashed line drawn on Figure 6. This gain is for an antenna 100% efficient, and would be 41.7 dB for a typical parabolic antenna (50% efficient).

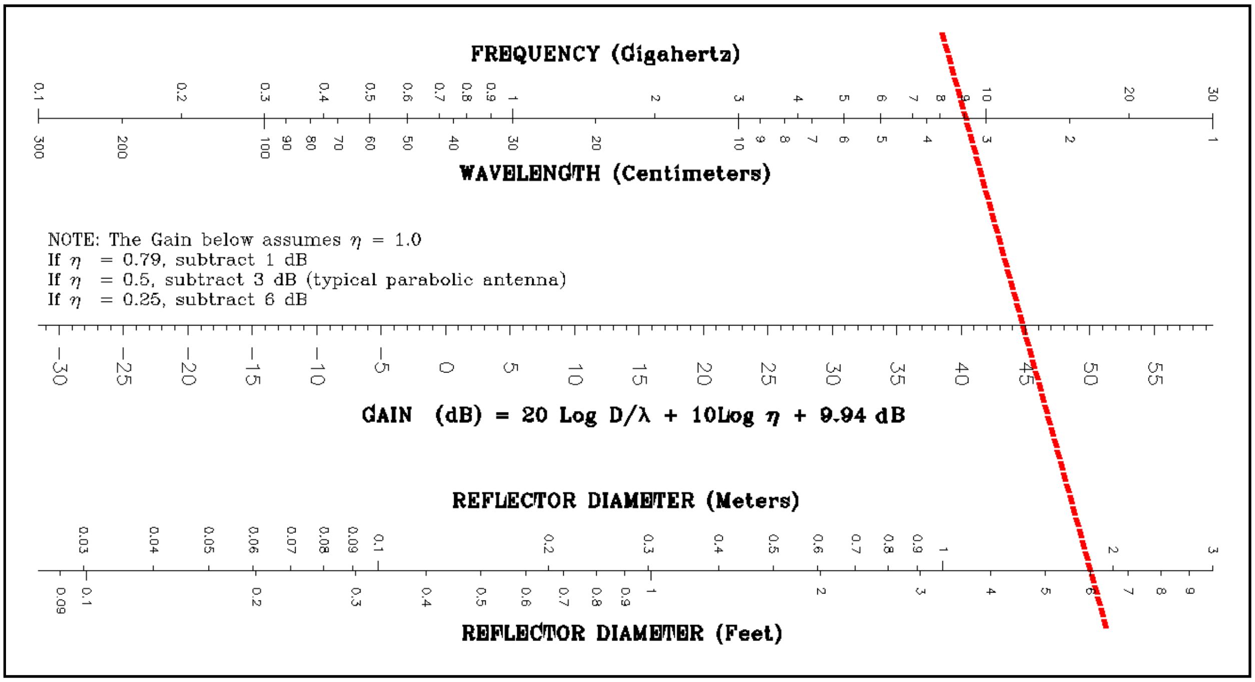

Figure 6. Antenna Gain Nomograph.

An example of a typical antenna (with losses) showing the variation of gain with frequency is depicted in Figure 7, and the variation of gain with antenna diameter in Figure 8. The circle on the curves in Figure 7 and 8 correspond to the Figure 6 example and yields 42 dB of gain for the 6 ft dish at 9 GHz.

Example Problem: If the two antennas in the drawing are "welded" together, how much power will be measured at point A? (Line loss $L_1 = L_2 = 0.5$, and 10log L_1 or $L_2 = 3$ dB)

Multiple choice:

A. 16 dBm b. 28 dBm c. 4 dBm d. 10 dBm e. < 4 dBm

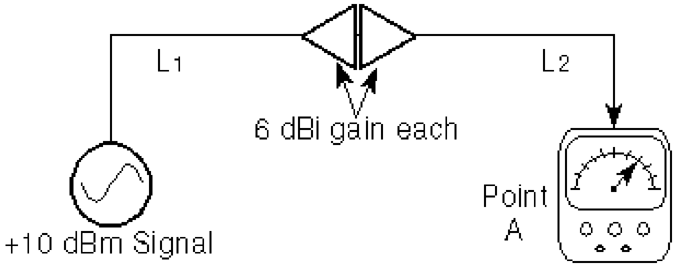

Answer:

The antennas do not act as they normally would since the antennas are operating in the near field. They act as inefficient coupling devices resulting in some loss of signal. In addition, since there are no active components, you cannot end up with more power than you started with. The correct answer is "e. < 4 dBm."

10 dBm - 3 dB - small loss -3 dB = 4 dBm - small loss

If the antennas were separated by 5 ft and were in the far field, the antenna gain could be used with space loss formulas to calculate (at 5 GHz): 10 dBm - 3 dB + 6 dB - 50 dB (space loss) + 6 dB -3 dB = -34 dBm (a much smaller signal).

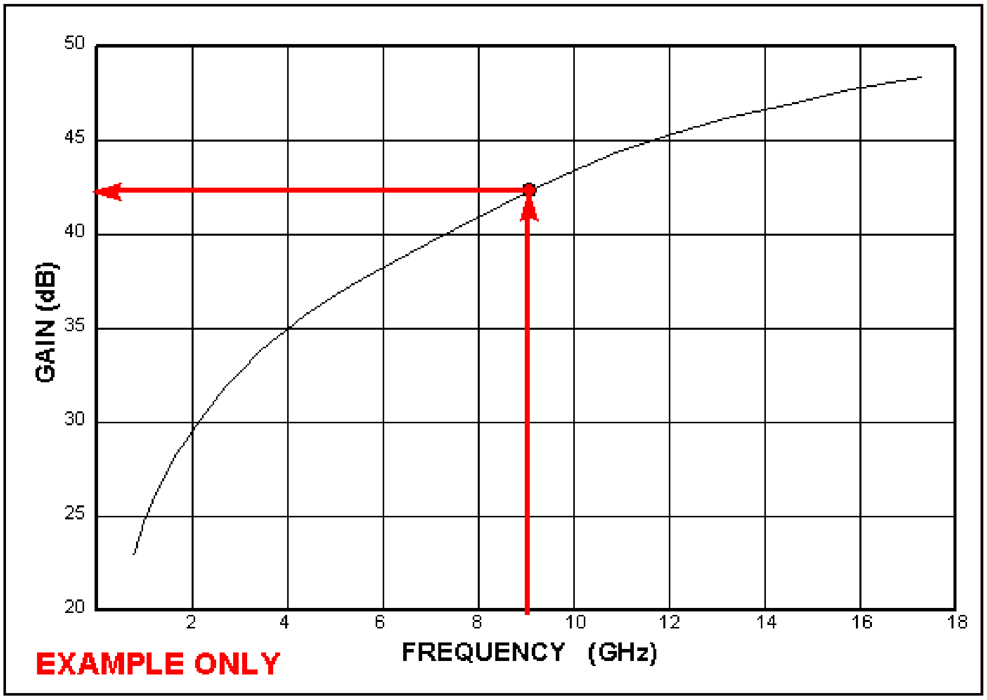

Figure 7. Gain of a Typical 6-Foot Dish Antenna (With Losses).

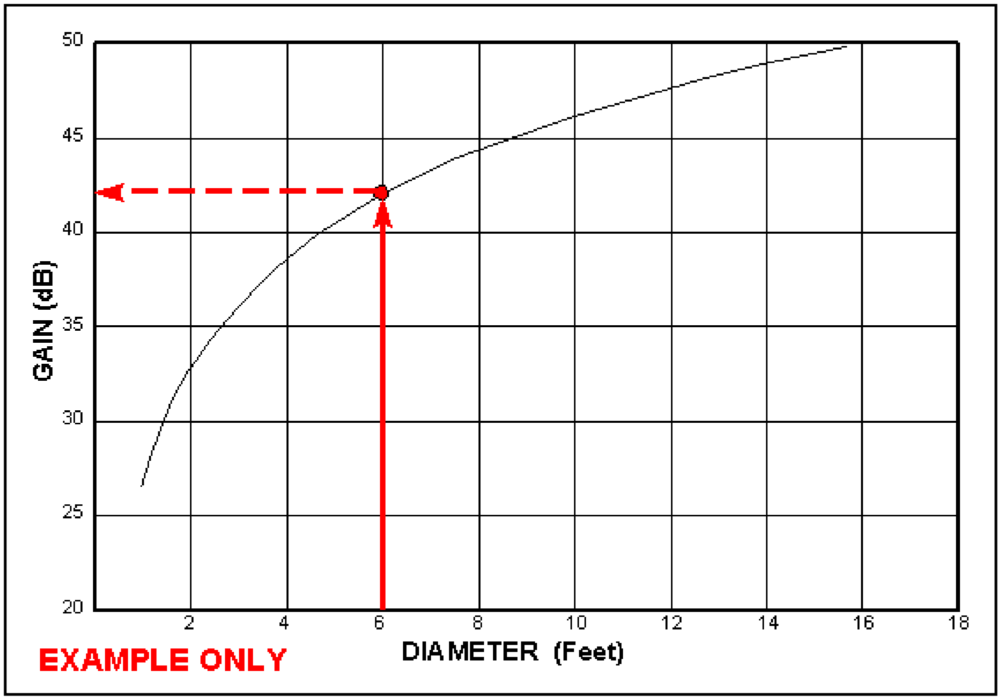

Figure 8. Gain of a Typical Dish at 9 GHz (With Losses).

This page intentionally left blank.

POLARIZATION

Table 1 shows the **theoretical** ratio of power transmitted between antennas of different polarization. These ratios are seldom fully achieved due to effects such as reflection, refraction, and other wave interactions, so some practical ratios are also included.

Table 1. Polarization Loss for Various Antenna Combinations.

Transmit Antenna Polarization	Receive Antenna Polarization	Ratio of Power Received to Maximum Power					
		Theoretical		Practical Horn		Practical Spiral	
		Ratio in dB	as Ratio	Ratio in dB	as Ratio	Ratio in dB	as Ratio
Vertical	Vertical	0 dB	1	*	*	N/A	N/A
Vertical	Slant (45° or 135°)	-3 dB	½	*	*	N/A	N/A
Vertical	Horizontal	- ∞ dB	0	-20 dB	1/100	N/A	N/A
Vertical	Circular (right-hand or left-hand)	-3 dB	½	*	*	*	*
Horizontal	Horizontal	0 dB	1	*	*	N/A	N/A
Horizontal	Slant (45° or 135°)	-3 dB	½	*	*	N/A	N/A
Horizontal	Circular (right-hand or left-hand)	-3 dB	½	*	*	*	*
Circular (right-hand)	Circular (right-hand)	0 dB	1	*	*	*	*
Circular (right-hand)	Circular (left-hand)	- ∞ dB	0	-20 dB	1/100	-10 dB	1/10
Circular (right or left)	Slant (45° or 135°)	-3 dB	½	*	*	*	*

* Approximately the same as theoretical.
Note: Switching transmit and receive antenna polarization will give the same results.

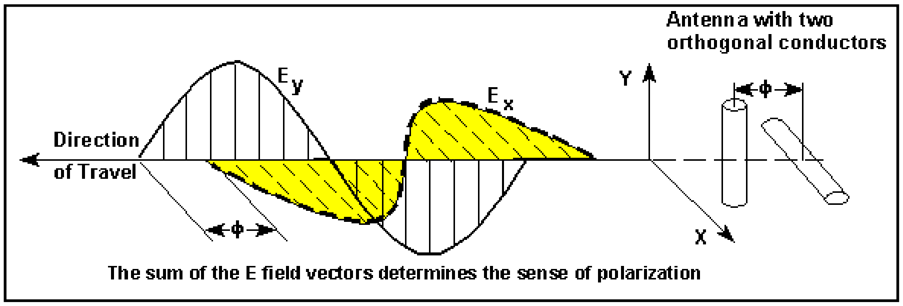
Antenna with two orthogonal conductors
E y
E x
Y
ϕ
Direction of Travel
ϕ
X
The sum of the E field vectors determines the sense of polarization

Figure 1. Polarization Coordinates.

The polarization of an electromagnetic wave is defined as the orientation of the electric field vector. Recall that the electric field vector is perpendicular to both the direction of travel and the magnetic field vector. The polarization is described by the geometric figure traced by the electric field vector upon a stationary plane perpendicular to the direction of propagation, as the wave travels through that plane. An electromagnetic wave is frequently composed of (or can be broken down into) two orthogonal components as shown in Figure 1. This may be due to the arrangement of power input leads to various points on a flat antenna, or due to an interaction of active elements in an array, or many other reasons.

The geometric figure traced by the sum of the electric field vectors over time is, in general, an ellipse as shown in Figure 2. Under certain conditions the ellipse may collapse into a straight line, in which case the polarization is called linear.

In the other extreme, when the two components are of equal magnitude and 90° out of phase, the ellipse will become circular as shown in Figure 3. Thus linear and circular polarization are the two special cases of elliptical polarization. Linear polarization may be further classified as being vertical, horizontal, or slant.

Figure 2 depicts plots of the E field vector while varying the relative amplitude and phase angle of its component parts.

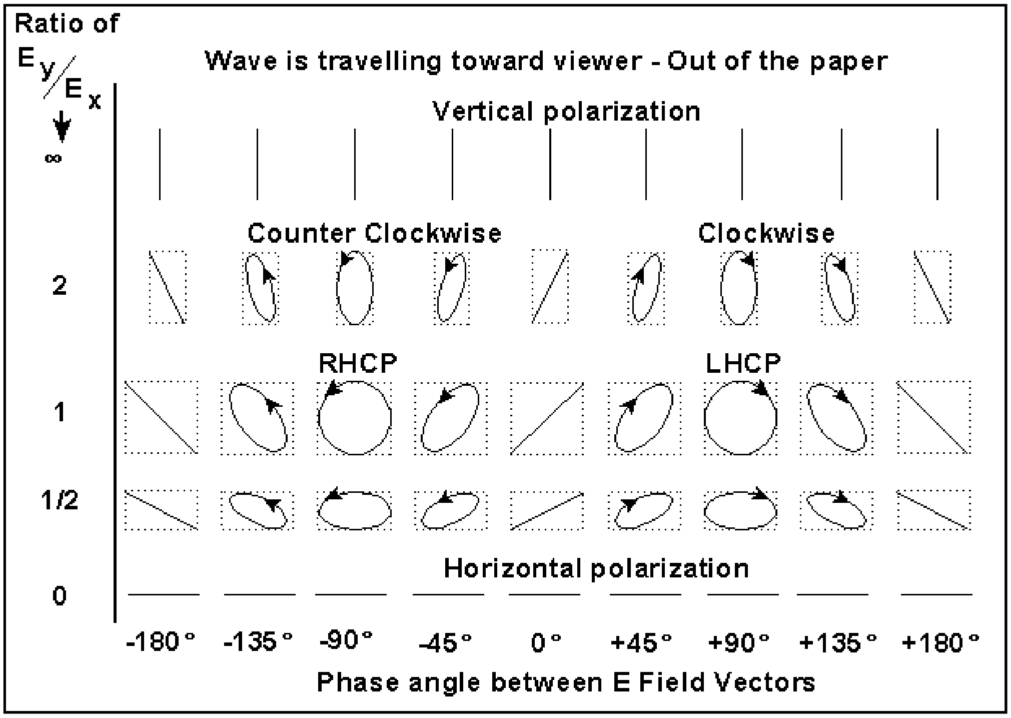
Ratio of E_y / E_x
∞
2
1
1/2
0
Wave is travelling toward viewer - Out of the paper
Vertical polarization
Counter Clockwise
Clockwise
RHCP
LHCP
Horizontal polarization
-180°
-135°
-90°
-45°
0°
+45°
+90°
+135°
+180°
Phase angle between E Field Vectors

Figure 2. Polarization as a Function of E_y / E_x Ratio and Phase Angle.
Adopted from J.D. Kraus, "Antennas," 2nd ed., Figure 2-37

For a linearly polarized antenna, the radiation pattern is taken both for a co-polarized and cross polarized response. The polarization quality is expressed by the ratio of these two responses. The ratio between the responses must typically be great (30 dB or greater) for an application such as cross-polarized jamming. For general applications, the ratio indicates system power loss due to polarization mismatch. For circularly polarized antennas, radiation patterns are usually taken with a rotating linearly polarized reference antenna. The reference antenna rotates many times while taking measurements around the azimuth of the antenna that is being tested. The resulting antenna pattern is the linear polarized gain with a cyclic ripple. The peak-to-peak value is the axial ratio, and represents the polarization quality for a circular polarized antenna. The typical RWR antenna has a maximum 3 dB axial ratio within 45° of boresight.

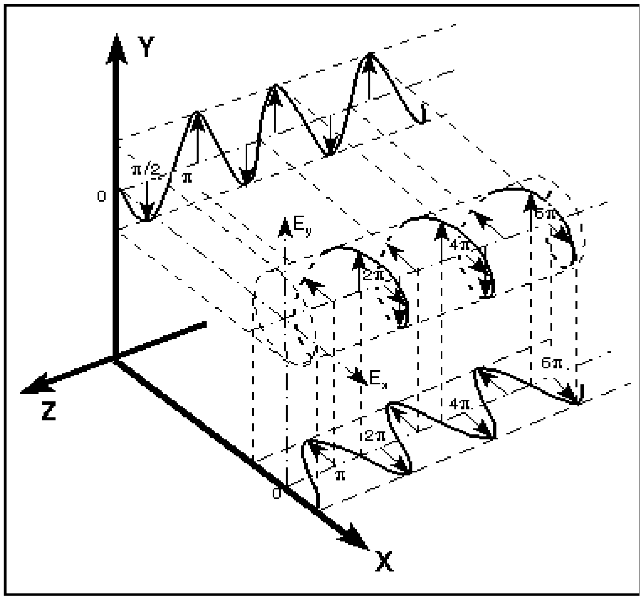
Y
Z
X
0
π/2
π
E_y
2π
4π
5π
E_x
6π
4π
2π
π
0

Figure 3. Circular Polarization – E Field.

For any antenna with an aperture area, as the aperture is rotated, the viewed dimension along the axis remains constant, while the other viewed dimension decreases to zero at 90° rotation. The axial ratio of an antenna will get worse as the antenna is rotated off boresight because the field contribution from the axial component will remain fairly constant and the other orthogonal component will decrease with rotation.

The sense of antenna polarization is defined from a viewer positioned behind an antenna looking in the direction of propagation. The polarization is specified as a transmitting, not receiving antenna regardless of intended use.

We frequently use "hand rules" to describe the sense of polarization. The sense is defined by which hand would be used in order to point that thumb in the direction of propagation and point the fingers of the same hand in the direction of rotation of the E field vector. For example, referring to Figure 4, if your thumb is pointed in the direction of propagation and the rotation is counterclockwise looking in the direction of travel, then you have left hand circular polarization.

Optics people view an aperture from the front and therefore use the opposite reference.

The polarization of a linearly polarized horn antenna can be directly determined by the orientation of the feed probe, which is in the direction of the E-field.

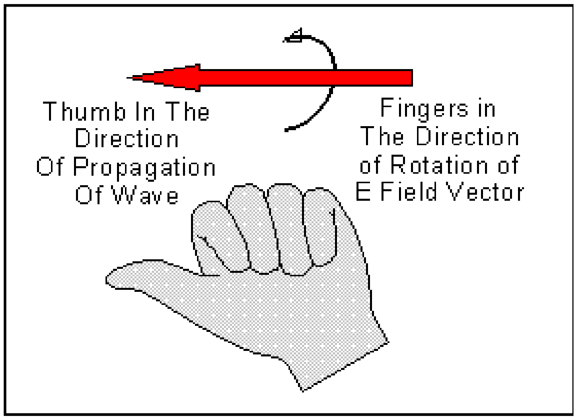
Thumb In The Direction Of Propagation Of Wave
Fingers in The Direction of Rotation of E Field Vector

Figure 4. Left Hand Polarization.

In general, a flat surface or sphere will reflect a linearly polarized wave with the same polarization as received. A horizontally polarized wave may get extended range because of water and land surface reflections, but signal cancellation will probably result in "holes" in coverage. Reflections will reverse the sense of circular polarization.

If the desired antenna is used for receiving a direct transmission as shown in Figure 5, the same polarization sense (specified if transmitting) is required for maximum signal reception in this situation. Buy two right-hand or two left-hand circularly polarized antennas for this case. When you procure antennas, remember that the polarization is specified as if transmitting, regardless of intended use.

Wave propagation between two identical antennas is analogous to being able to thread a nut from one bolt to an identical opposite facing bolt.

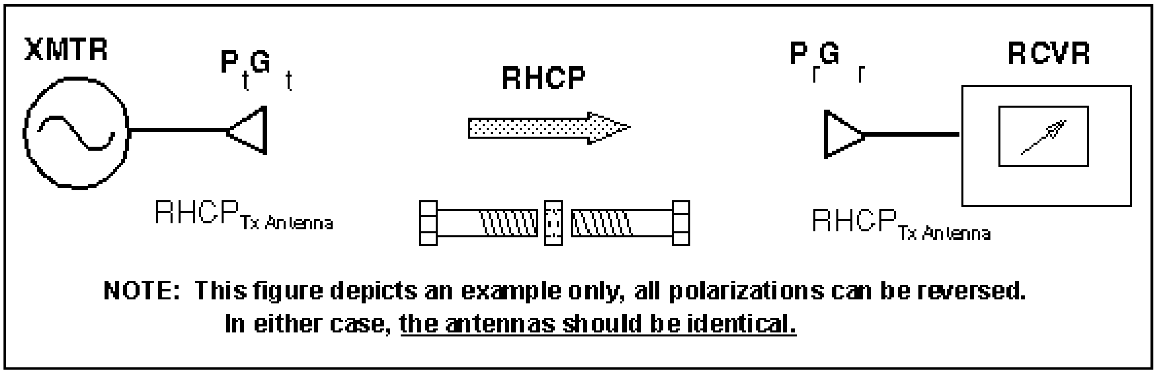
XMTR
P_tG_t
RHCP
P_rG_r
RCVR
RHCP Tx Antenna
RHCP Tx Antenna
NOTE: This figure depicts an example only, all polarizations can be reversed.
In either case, the antennas should be identical.

Figure 5. Same Circular Polarization.

If the desired antenna is used for a receiving a wave with a single or odd number of reflections, such as a bistatic radar where separate antennas are used for transmit and receive as shown in Figure 6, then opposite circularly polarized antennas would be used for maximum signal reception. In this case buy antennas of opposite polarization sense (one left hand and one right hand).

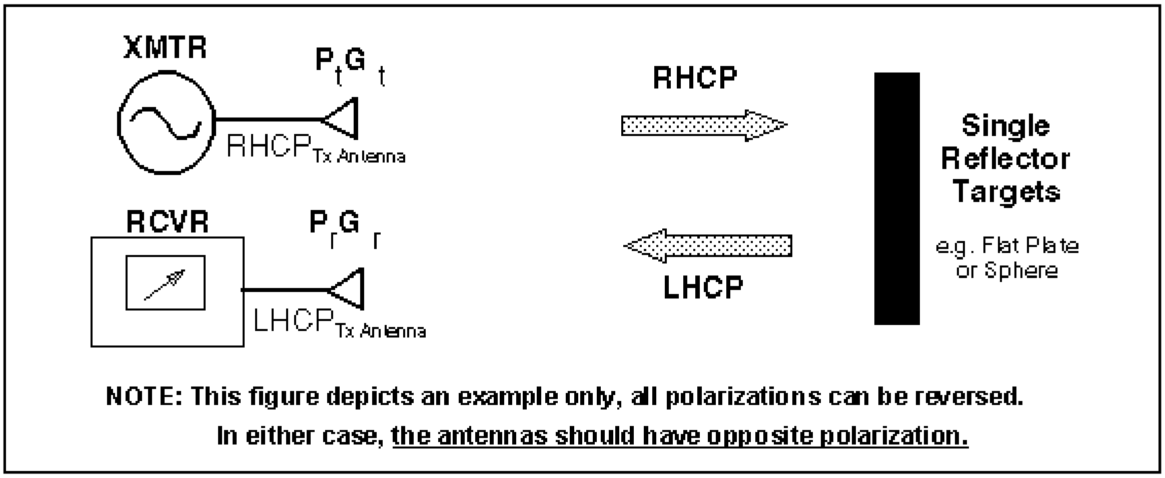
XMTR
PtGt
RHCP Tx Antenna
RCVR
PrGr
LHCP Tx Antenna
RHCP
LHCP
Single Reflector Targets
e.g. Flat Plate or Sphere
NOTE: This figure depicts an example only, all polarizations can be reversed.
In either case, the antennas should have opposite polarization.

Figure 6. Opposite Circular Polarization.

In a corner reflector, waves reflect twice before returning to the receiver as shown in Figure 7, consequently they return with the same sense as they were transmitted. In this case (or any even number of reflections) buy antennas of the same polarization sense.

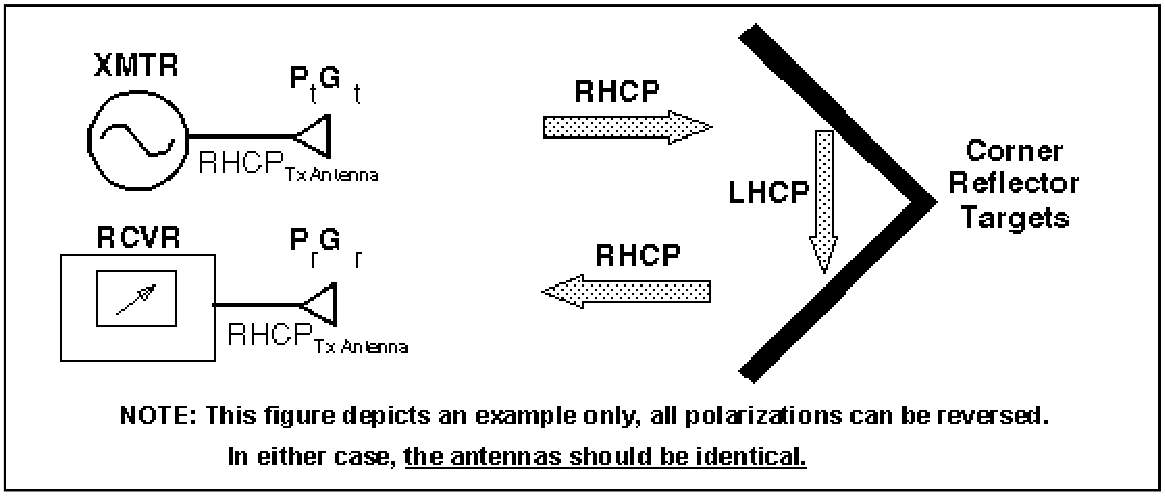
XMTR
PtGt
RHCP Tx Antenna
RCVR
PrGr
RHCP Tx Antenna
RHCP
LHCP
RHCP
Corner Reflector Targets
NOTE: This figure depicts an example only, all polarizations can be reversed.
In either case, the antennas should be identical.

Figure 7. Circular Polarization With Corner Reflector.

An aircraft acts as both a corner reflector and a "normal" reflector so the return has mixed polarization. Most airborne radars use the same antenna for transmitting and receiving in order to receive the corner reflections and help exclude receipt of reflections from rain (single polarization reversal), however in doing so there is about a 5-9 dB loss from the ideal receiver case. It should be noted that the return from raindrops is attenuated by approximately 20 dB.

RADIATION PATTERNS

The radiation pattern is a graphical depiction of the relative field strength transmitted from or received by the antenna. Antenna radiation patterns are taken at one frequency, one polarization, and one plane cut. The patterns are usually presented in polar or rectilinear form with a dB strength scale. Patterns are normalized to the maximum graph value, 0 dB, and a directivity is given for the antenna. This means that if the side lobe level from the radiation pattern were down -13 dB, and the directivity of the antenna was 4 dB, then the sidelobe gain would be -9 dB.

Figures 1 to 14 on the pages following depict various antenna types and their associated characteristics. The patterns depicted are those which most closely match the purpose for which the given shape was intended. In other words, the radiation pattern can change dramatically depending upon frequency, and the wavelength to antenna characteristic length ratio. See Section 3-4. Antennas are designed for a particular frequency. Usually the characteristic length is a multiple of $\lambda/2$ minus 2-15% depending on specific antenna characteristics.

The gain is assumed to mean directional gain of the antenna compared to an isotropic radiator transmitting to or receiving from all directions.

The half-power (-3 dB) beamwidth is a measure of the directivity of the antenna.

Polarization, which is the direction of the electric (not magnetic) field of an antenna is another important antenna characteristic. This may be a consideration for optimizing reception or jamming.

The bandwidth is a measure of how much the frequency can be varied while still obtaining an acceptable VSWR (2:1 or less) and minimizing losses in unwanted directions. See Glossary, Section 10.

A 2:1 VSWR corresponds to a 9.5 dB (or 10%) return loss - see Section 6-2.

Two methods for computing antenna bandwidth are used:

Narrowband by %, $B = \left(\frac{F_U - F_L}{F_C}\right)(100)$, where F_C = Center frequency

Broadband by ratio, $B = \frac{F_U}{F_L}$

An antenna is considered broadband if F_U / F_L > 2. The table at the right shows the equivalency of the two, however the shaded values are not normally used because of the aforementioned difference in broadband/narrowband.

Bandwidth	
%	Ratio
5	1.05 : 1
10	1.11 : 1
20	1.22 : 1
30	1.35 : 1
40	1.50 : 1
50	1.67 : 1
60	1.85 : 1
67	2 : 1
100	3 : 1
120	4 : 1
133	5 : 1
150	7 : 1
160	9 : 1
163	10 : 1

Should there be ever a need to express bandwidth of an antenna in one or the other alternative formats, a conversion between the two narrowband and broadband bandwidth quantities can be easily calculated using the following relationships:

Calculate broadband ratio B_{bb} given narrowband B%,

$$B_{bb} = (200 + B\%)/(200 - B\%) \qquad \textbf{[1]}$$

Calculate narrowband B% given broadband ratio B_{bb},

$$B\% = 200 * (Bbb - 1)/(Bbb + 1) \qquad \textbf{[2]}$$

For an object that experiences a plane wave, the resonant mode is achieved when the dimension of the object is nλ/2, where n is an integer. Therefore, one can treat the apertures shown in the following figure as half wave length dipole antennas for receiving and reflecting signals. More details are contained in Section 8-4.

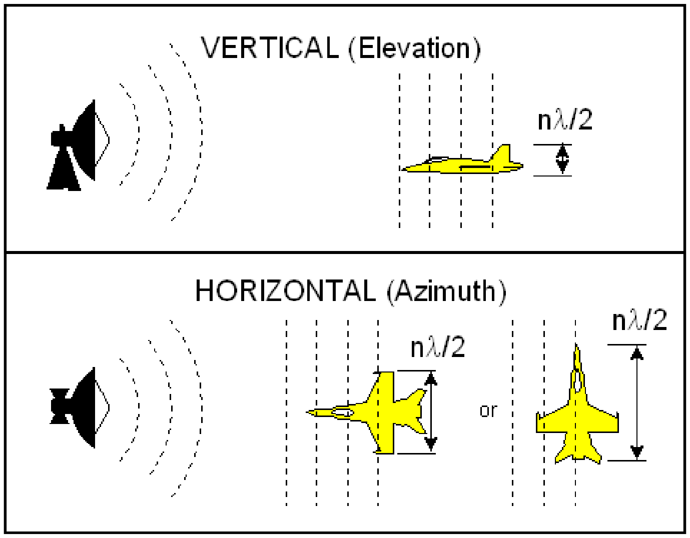

The following lists antenna types by page number. The referenced page shows frequency limits, polarizations, etc.

Type	Page	Type	Page
4 arm conical spiral	3-3.6	log periodic	3-3.8
alford loop	3-3.4	loop, circular	3-3.4
aperture synthesis	3-3.8	loop, alford	3-3.4
array	3-3.8	loop, square	3-3.4
axial mode helix	3-3.5	luneberg lens	3-3.9
biconical w/polarizer	3-3.6	microstrip patch	3-3.9
biconical	3-3.6	monopole	3-3.3
cavity backed circuit fed slot	3-3.5	normal mode helix	3-3.5
cavity backed spiral	3-3.9	parabolic	3-3.7
circular loop	3-3.4	patch	3-3.9
conical spiral	3-3.5	reflector	3-3.10
corner reflector	3-3.9	rhombic	3-3.3
dipole array, linear	3-3.8	sinuous, dual polarized	3-3.6
dipole	3-3.3	slot, guide fed	3-3.9
discone	3-3.4	slot, cavity backed	3-3.9
dual polarized sinuous	3-3.6	spiral, 4 arm conical	3-3.6
guide fed slot	3-3.9	spiral, conical	3-3.5
helix, normal mode	3-3.5	spiral, cavity backed	3-3.5
helix, axial mode	3-3.5	square loop	3-3.4
horn	3-3.7	vee	3-3.3
linear dipole array	3-3.8	yagi	3-3.8

Antenna Type	Radiation Pattern	Characteristics
MONOPOLE 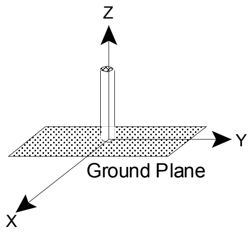	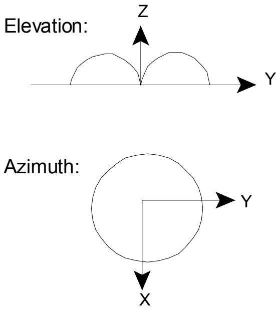	**Polarization: Linear** Vertical as shown **Typical Half-Power Beamwidth** 45 deg x 360 deg **Typical Gain:** 2-6 dB at best **Bandwidth:** 10% or 1.1:1 **Frequency Limit** **Lower:** None **Upper:** None **Remarks:** Polarization changes to horizontal if rotated to horizontal
λ/2 DIPOLE 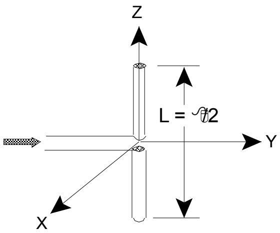	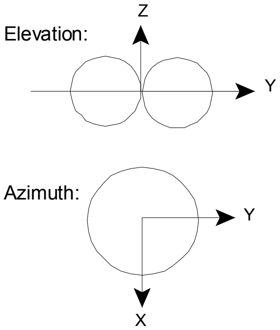	**Polarization**: Linear Vertical as shown **Typical Half-Power Beamwidth** 80 deg x 360 deg **Typical Gain:** 2 dB **Bandwidth:** 10% or 1.1:1 **Frequency Limit** **Lower:** None **Upper:** 8 GHz (practical limit) **Remarks:** Pattern and lobing changes significantly with L/f. Used as a gain reference < 2 GHz.

Figure 1. Monopole and Dipole Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
VEE 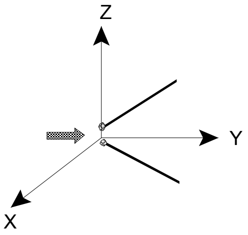	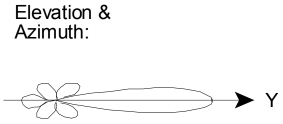	**Polarization**: Linear Vertical as shown **Typical Half-Power Beamwidth** 60 deg x 60 deg **Typical Gain:** 2 to 7 dB **Bandwidth:** "Broadband" **Frequency Limit** **Lower:** 3 MHz **Upper:** 500 MHz (practical limits) **Remarks:** 24KHz versions are known to exist. Terminations may be used to reduce backlobes.
RHOMBIC 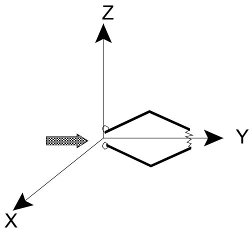	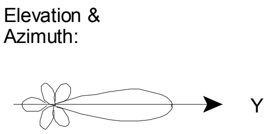	**Polarization**: Linear Vertical as shown **Typical Half-Power Beamwidth** 60 deg x 60 deg **Typical Gain:** 3 dB **Bandwidth:** "Broadband" **Frequency Limit** **Lower:** 3 MHz **Upper:** 500 MHz **Remarks:** Termination resistance used to reduce backlobes.

Figure 2. Vee and Rhombic Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
CIRCULAR LOOP (Small) 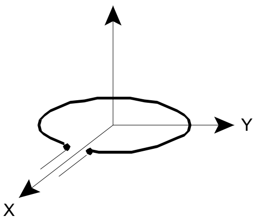	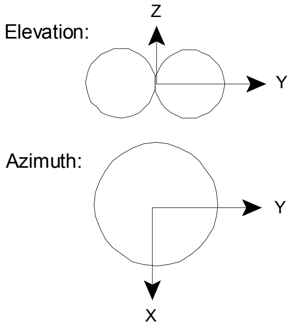 Elevation: Z Y Azimuth: Y X	**Polarization:** Linear Horizontal as shown **Typical Half-Power Beamwidth:** 80 deg x 360 deg **Typical Gain:** -2 to 2 dB **Bandwidth:** 10% or 1.1:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 1 GHz
SQUARE LOOP (Small) 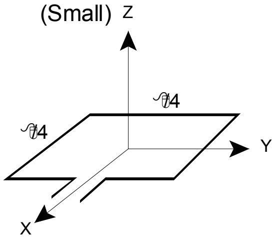	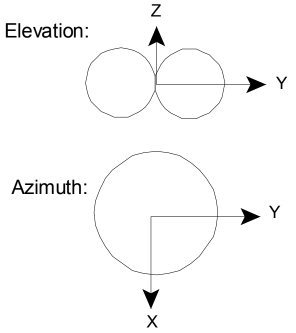 Elevation: Z Y Azimuth: Y X	**Polarization:** Linear Horizontal as shown **Typical Half-Power Beamwidth:** 100 deg x 360 deg **Typical Gain:** 1-3 dB **Bandwidth:** 10% or 1.1:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 1 GHz

Figure 3. Circular Loop and Square Loop Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
DISCONE 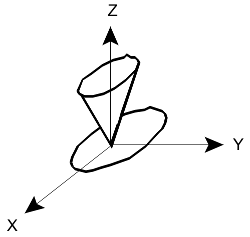	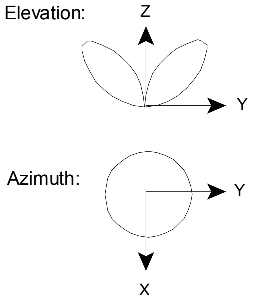 Elevation: Z Y Azimuth: Y X	**Polarization:** Linear Vertical as shown **Typical Half-Power Beamwidth:** 20-80 deg x 360 deg **Typical Gain:** 0-4 dB **Bandwidth:** 100% or 3:1 **Frequency Limit:** **Lower:** 30 MHz **Upper:** 3 GHz
ALFORD LOOP 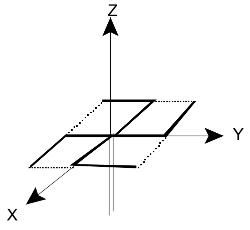	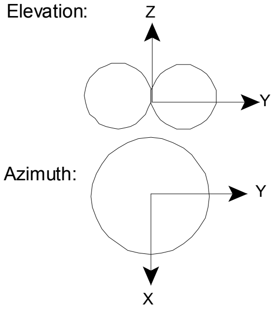 Elevation: Z Y Azimuth: Y X	**Polarization:** Linear Horizontal as shown **Typical Half-Power Beamwidth:** 80 deg x 360 deg **Typical Gain:** -1 dB **Bandwidth:** 67% or 2:1 **Frequency Limit:** **Lower:** 100 MHz **Upper:** 12 GHz

Figure 4. Discone and Alford Loop Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
AXIAL MODE HELIX 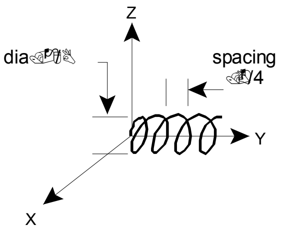 Z dia λ/π spacing λ/4 Y X	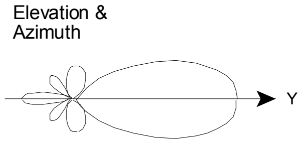 Elevation & Azimuth Y	**Polarization:** Circular Left hand as shown **Typical Half-Power Beamwidth:** 50 deg x 50 deg **Typical Gain:** 10 dB **Bandwidth:** 52% or 1.7:1 **Frequency Limit** **Lower:** 100 MHz **Upper:** 3 GHz **Remarks:** Number of loops >3
NORMAL MODE HELIX 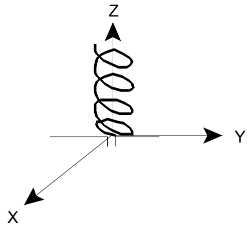	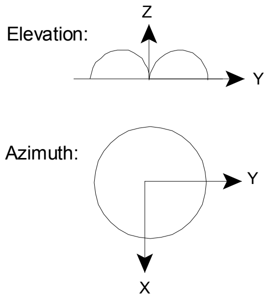 Elevation: Z Y Azimuth: Y X	**Polarization:** Circular - with an ideal pitch to diameter ratio. **Typical Half-Power Beamwidth:** 60 deg x 360 deg **Typical Gain:** 0 dB **Bandwidth:** 5% or 1.05:1 **Frequency Limit** **Lower:** 100 MHz **Upper:** 3 GHz

Figure 5. Axial Mode Helix and Normal Mode Helix Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
CAVITY BACKED SPIRAL (Flat Helix) 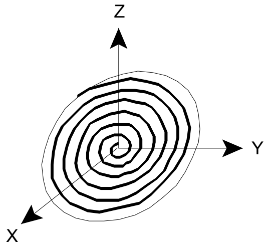	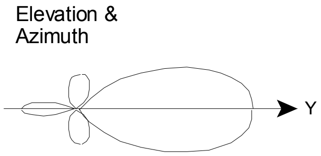 Elevation & Azimuth Y	**Polarization:** Circular Left hand as shown **Typical Half-Power Beamwidth:** 60 deg x 90 deg **Typical Gain:** 2-4 dB **Bandwidth:** 160% or 9:1 **Frequency Limit:** **Lower:** 500 MHz **Upper:** 18 GHz
CONICAL SPIRAL 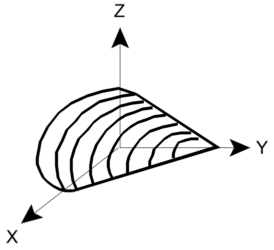	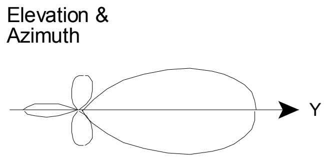 Elevation & Azimuth Y	**Polarization:** Circular Left hand as shown **Typical Half-Power Beamwidth:** 60 deg x 60 deg **Typical Gain:** 5-8 dB **Bandwidth:** 120% or 4:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 18 GHz

Figure 6. Cavity Backed Spiral and Conical Spiral Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
4 ARM CONICAL SPIRAL 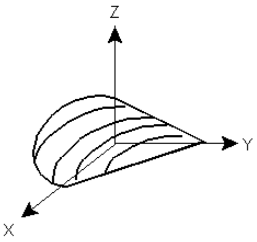 	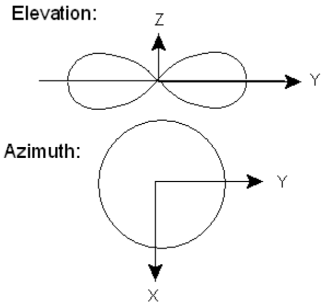 	**Polarization:** Circular Left hand as shown **Typical Half-Power Beamwidth:** 50 deg x 360 deg **Typical Gain:** 0 dB **Bandwidth:** 120% or 4:1 **Frequency Limit:** **Lower:** 500 MHz **Upper:** 18 GHz
DUAL POLARIZED SINUOUS 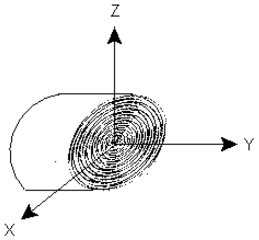 	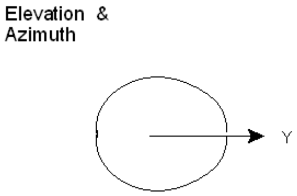 	**Polarization**: Dual vertical or horizontal or dual Circular right hand or left hand with hybrid **Typical Half-Power Beamwidth:** 75 deg x 75 deg **Typical Gain:** 2 dB **Bandwidth:** 163% or 10:1 **Frequency Limit:** **Lower:** 500 MHz **Upper:** 18 GHz

Figure 7. Dual Polarized Sinuous Antenna Characteristics and 4 Arm Conical Spiral.

Antenna Type	Radiation Pattern	Characteristics
BICONICAL 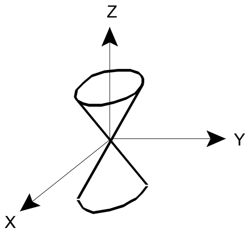 	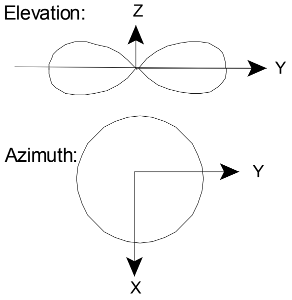 	**Polarization:** Linear, Vertical as shown **Typical Half-Power Beamwidth:** 20-100 deg x 360 deg **Typical Gain:** 0-4 dB **Bandwidth:** 120% or 4:1 **Frequency Limit:** **Lower:** 500 MHz **Upper:** 40 GHz
BICONICAL W/POLARIZER 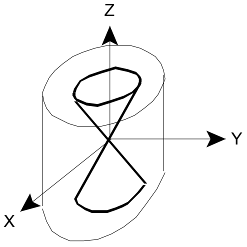 	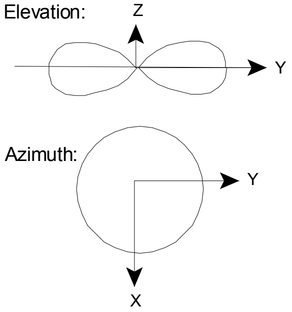 	**Polarization:** Circular, Direction depends on polarization **Typical Half-Power Beamwidth:** 20-100 deg x 360 deg **Typical Gain:** -3 to 1 dB **Bandwidth:** 100% or 3:1 **Frequency Limit:** **Lower:** 2 GHz **Upper:** 18 GHz

Figure 8. Biconical and Biconical With Polarizer Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
HORN 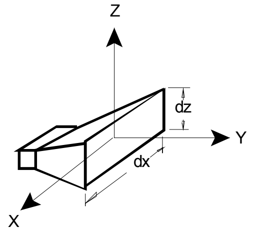	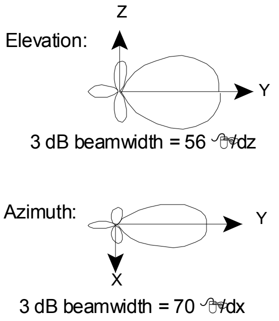 Z Elevation: Y 3 dB beamwidth = 56 λ/dz Azimuth: Y X 3 dB beamwidth = 70 λ/dx	**Polarization:** Linear **Typical Half-Power Beamwidth:** 40 deg x 40 deg **Typical Gain:** 5 to 20 dB **Bandwidth:** If ridged: 120% or 4:1 If not ridged: 67% or 2:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 40 GHz
HORN W / POLARIZER 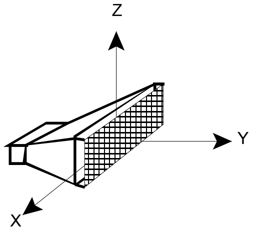	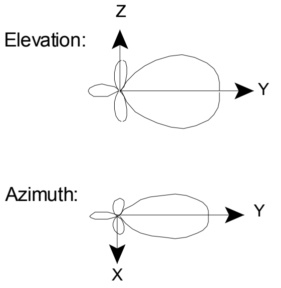 Z Elevation: Y Azimuth: Y X	**Polarization:** Circular, Depends on polarizer **Typical Half-Power Beamwidth:** 40 deg x 40 deg **Typical Gain:** 5 to 10 dB **Bandwidth:** 60% or 2:1 **Frequency Limit:** **Lower:** 2 GHz **Upper:** 18 GHz

Figure 9. Horn and Horn w/Polarizer Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
PARABOLIC (Prime) 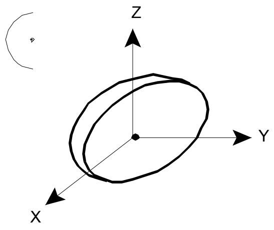	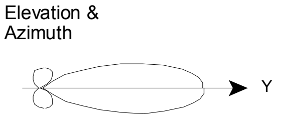 Elevation & Azimuth Y	**Polarization:** Takes polarization of feed **Typical Half-Power Beamwidth:** 1 to 10 deg **Typical Gain:** 20 to 30 dB **Bandwidth:** 33% or 1.4:1 limited mostly by feed **Frequency Limit:** **Lower:** 400 MHz **Upper:** 13+ GHz
PARABOLIC 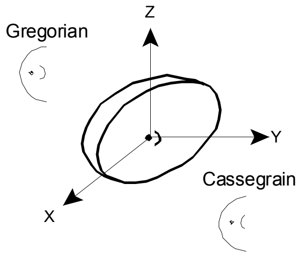 Z Gregorian Y Cassegrain X	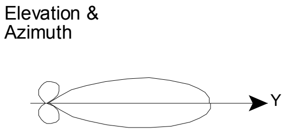 Elevation & Azimuth Y	**Polarization:** Takes polarization of feed **Typical Half-Power Beamwidth:** 1 to 10 deg **Typical Gain:** 20 to 30 dB **Bandwidth:** 33% or 1.4:1 **Frequency Limit:** **Lower:** 400 MHz **Upper:** 13+ GHz

Figure 10. Parabolic (Prime) and Parabolic Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
YAGI 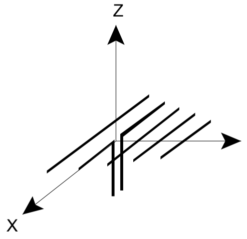	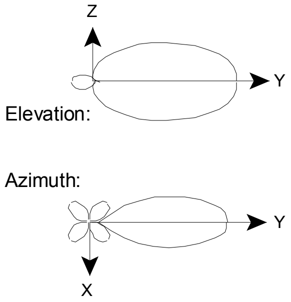	**Polarization:** Linear Horizontal as shown **Typical Half-Power Beamwidth** 50 deg X 50 deg **Typical Gain:** 5 to 15 dB **Bandwidth:** 5% or 1.05:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 2 GHz
LOG PERIODIC 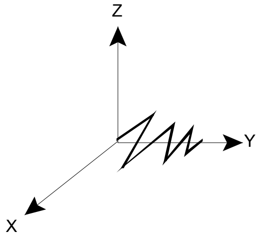	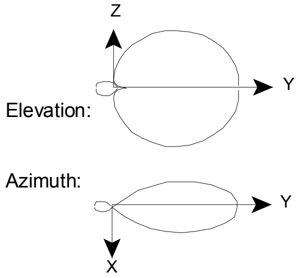	**Polarization:** Linear **Typical Half-Power Beamwidth:** 60 deg x 80 deg **Typical Gain:** 6 to 8 dB **Bandwidth:** 163% or 10:1 **Frequency Limit:** **Lower:** 3 MHz **Upper:** 18 GHz **Remarks:** This array may be formed with many shapes including dipoles or toothed arrays.

Figure 11. Yagi and Log Periodic Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
LINEAR DIPOLE ARRAY (Corporate Feed) 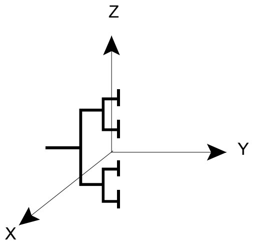	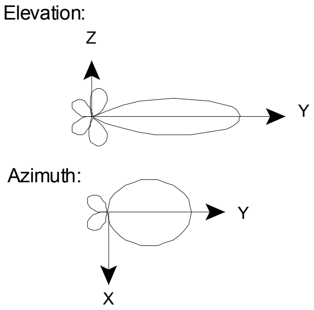	**Polarization:** Element dependent Vertical as shown **Typical Half-Power Beamwidth:** Related to gain **Typical Gain:** Dependent on number of elements **Bandwidth:** Narrow **Frequency Limit:** **Lower:** 10 MHz **Upper:** 10 GHz
APERTURE SYNTHESIS 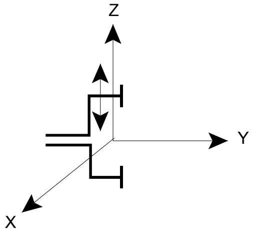	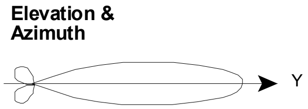	All characteristics dependent on elements **Remarks:** Excellent side-looking, ground mapping where the aircraft is a moving linear element.

Figure 12. Linear Dipole Array and Aperture Synthesis Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
CAVITY BACKED CIRCUIT FED SLOT (and Microstrip Patch) 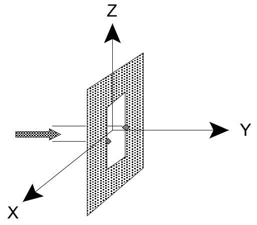	Elevation & Azimuth 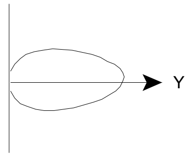	**Polarization:** Linear, vertical as shown **Typical Half-Power Beamwidth**: 80 deg x 80 deg **Typical Gain:** 6 dB **Bandwidth:** Narrow **Frequency Limit:** **Lower:** 50 MHz **Upper:** 18 GHz **Remarks:** The feed line is sometimes separated from the radiator by a dialetric & uses capacititive coupling. Large conformal phased arrays can be made this way.
GUIDE FED SLOT 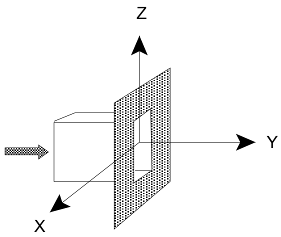	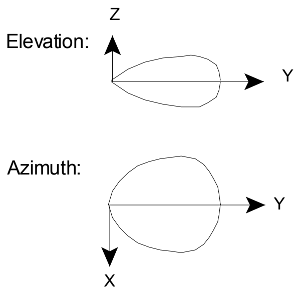	**Polarization:** Linear, **Typical Half-Power Beamwidth** Elevation: 45-50☜ Azimuth: 80☜ **Typical Gain:** 0 dB **Bandwidth:** Narrow **Frequency Limit:** **Lower:** 2 GHz **Upper:** 40 GHz **Remarks:** Open RF Waveguide

Figure 13. Cavity Backed Circuit Fed Slot and Guide Fed Slot Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
CORNER REFLECTOR 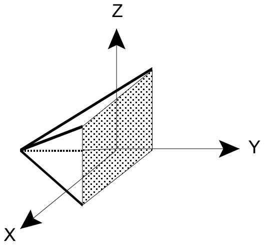	Elevation: (Z-Y) Azimuth: (X-Y) Dependent upon feed emitter	**Polarization**: Feed dependent **Typical Half-Power Beamwidth** 40 deg x variable **Typical Gain:** 10 dB above feed **Bandwidth:** Narrow **Frequency Limit** **Lower:** 1 GHz **Upper:** 40 GHz **Remarks:** Typically fed with a dipole or colinear array.
LUNEBURG LENS 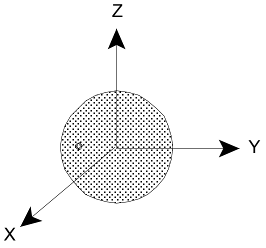	Elevation & Azimuth 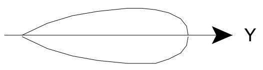	**Polarization:** Feed dependent **Typical Half-Power Beamwidth:** System dependent **Typical Gain:** System dependent **Bandwidth:** Narrow **Frequency Limit** **Lower:** 1 GHz **Upper:** 40 GHz **Remarks:** Variable index dielectric sphere.

Figure 14. Corner Reflector and Luneburg Lens Antenna Characteristics.

This page intentionally left blank.

FREQUENCY / PHASE EFFECTS OF ANTENNAS

The radiation patterns of the antennas presented in the previous section are for antenna geometries most commonly used. The antenna should be viewed as a matching network that takes the power from a transmission line (50 ohm, for example), and matches it to the free space "impedance" of 377 ohms. The most critical parameter is the change of VSWR with frequency. The pattern usually does not vary much from acceptable to the start of unacceptable VSWRs (> 2:1). For a given physical antenna geometric size, the actual radiation pattern varies with frequency.

The antenna pattern depicted in Figure 1 is for the dipole pictured in Section 3-3. The maximum gain is normalized to the outside of the polar plot and the major divisions correspond to 10 dB change. In this example, the dipole length (in wavelengths) is varied, but the same result can be obtained by changing frequency with a fixed dipole length. From the figure, it can be seen that side lobes start to form at 1.25λ and the side lobe actually has more gain than the main beam at 1.5λ. Since the radiation pattern changes with frequency, the gain also changes.

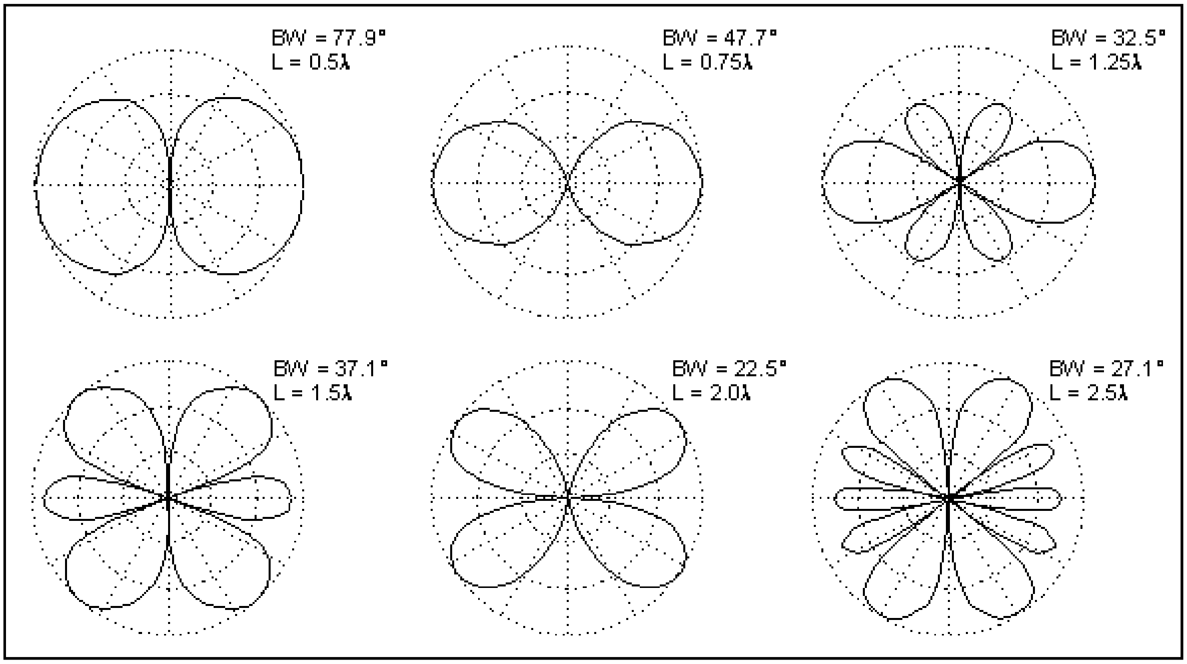

Figure 1. Frequency Effects.

Figure 2 depicts phase/array effects, which are yet another method for obtaining varied radiation patterns. In the figure, parallel dipoles are viewed from the end. It can be seen that varying the phase of the two transmissions can cause the direction of the radiation pattern to change. This is the concept behind phased array antennas. Instead of having a system mechanically sweeping the direction of the antenna through space, the phase of radiating components is varied electronically, producing a moving pattern with no moving parts. It can also be seen that increasing the number of elements further increases the directivity of the array. In an array, the pattern does vary considerably with frequency due to element spacing (measured in wavelengths) and the frequency sensitivity of the phase shifting networks.

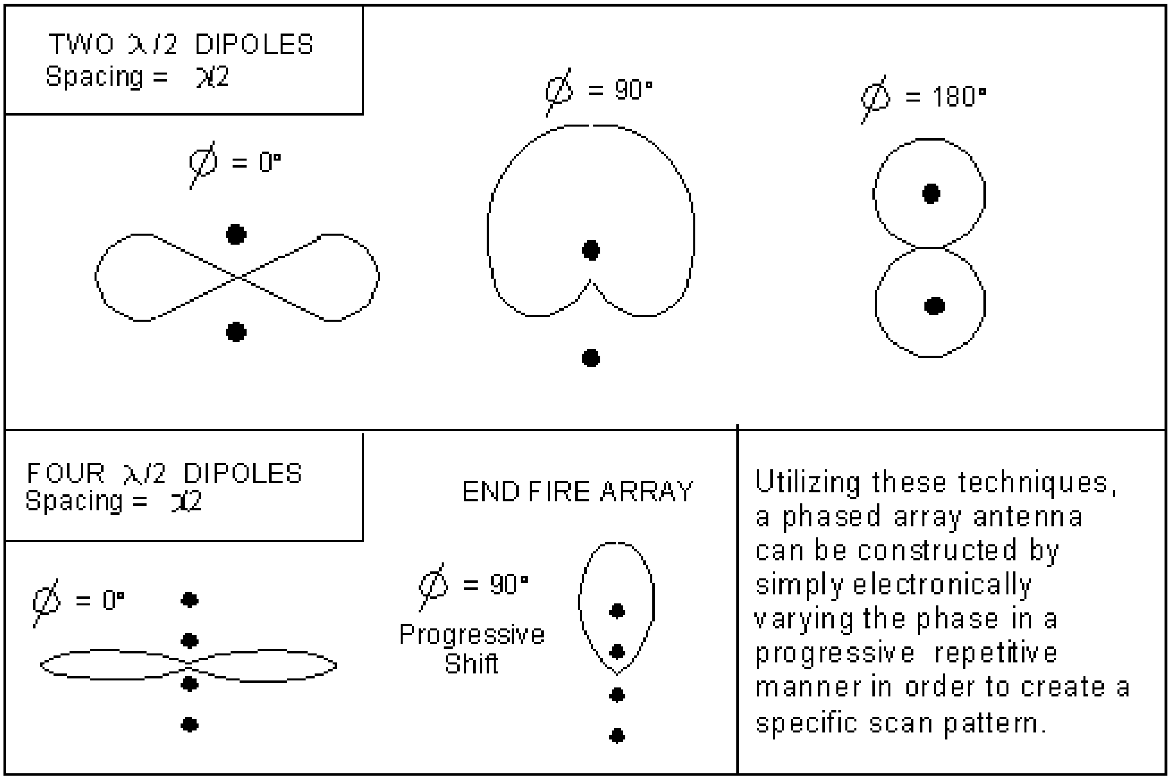

Utilizing these techniques, a phased array antenna can be constructed by simply electronically varying the phase in a progressive repetitive manner in order to create a specific scan pattern.

Figure 2. Phase / Array Effects*.

* Note: Assuming Figure 2 depicts x-y plane antenna pattern cross section, to achieve the indicated array patterns using dipole antennas, the dipole antenna elements must be aligned with the z-axis.

Two antennas that warrant special consideration are the phased array and the Rotman bootlace type lens. Both of these antennas find wide application in EW, RADAR, and Communications. The phased array will be described first.

LINEAR PHASED ARRAY

The linear phased array with equal spaced elements is easiest to analyze and forms the basis for most array designs. Figure 3 schematically illustrates a corporate feed linear array with element spacing d.

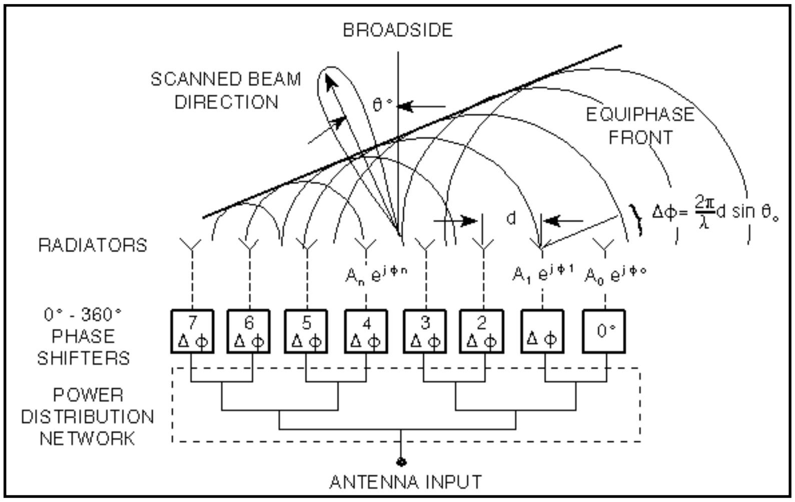

Figure 3. Corporate Fed Phased Array.

It is the simplest and is still widely used. By controlling the phase and amplitude of excitation to each element, as depicted, we can control the direction and shape of the beam radiated by the array. The phase excitation, $\varphi(n)$, controls the beam pointing angle, θ_o, in a phased array. To produce a broadside beam, $\theta_o=0$, requires phase excitation, $\varphi(n)=0$. Other scan angles require an excitation, $\varphi(n) = nkd \sin(\theta_o)$, for the nth element where k is the wave number ($2\pi/\lambda$). In this manner a linear phased array can radiate a beam in any scan direction, θ_o, provided the element pattern has sufficient beamwidth. The amplitude excitation, A_n, can be used to control beam shape and sidelobe levels. Often the amplitude excitation is tapered in a manner similar to that used for aperture antennas to reduce the sidelobe levels. One of the problems that can arise with a phased array is insufficient bandwidth, since the phase shift usually is not obtained through the introduction of additional path length. However, it should be noted that at broadside the corporate feed does have equal path length and would have good bandwidth for this scan angle.

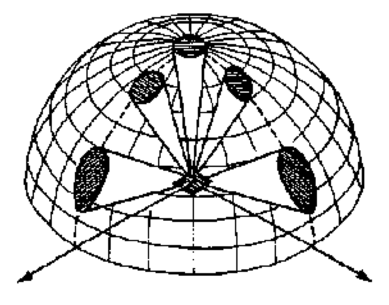

Figure 4. Beam Distortion.

The linear array described above would yield a narrow fan beam in the plane normal to the plane containing the array and scan direction, with the narrow beamwidth in the plane of the array. To obtain a pencil beam it would be necessary to array several of these linear arrays in such a manner resulting in a planar array of radiating elements. A problem associated with all electronic scanning is beam distortion with scan angle. Figure 4 illustrates this phenomenon. It results in spread of the beam shape with a concomitant reduction in gain. This effect is known as “scan loss.” For an ideal array element, scan loss is equal to the reduction in aperture size in the scan direction which varies as cos θ, where θ is the scan angle measured from the planar array normal.

When elements are spaced greater than $\lambda/2$ apart, grating lobes are possible when scanning. As the beam is scanned further from broadside, a point is reached at which a second symmetrical main lobe is developed at the negative scan angle from broadside. This condition is not wanted because antenna gain is immediately reduced by 3 dB due to the second lobe. Grating lobes are a significant problem in EW applications because the broad frequency bandwidth requirements mean that at the high end of the frequency band, the elements may be spaced greater than $\lambda/2$. Therefore in order to avoid grating lobes over a large frequency bandwidth, element spacing must be no greater than $\lambda/2$ at the highest frequency of operation.

There are many other factors to consider with a phased array such as coning, where the beam curves at large scan angles, and mutual coupling between elements that affect match and excitation. Excessive mutual coupling will invariably result in blind scan angles where radiation is greatly attenuated. These issues will not be covered in detail here.

Of interest is the gain of the array which is given by:

$$\text{Array Gain} = G_e(\theta) \bullet \sum_{n=1}^{N} A(n)\, e^{j\,\phi(n)}\, e^{j\,n\,k\,d\,\sin\theta}$$ Where each element is as described in Section 3-4.

$G_e(\theta)$ is the element gain which in this case has been taken the same for all elements. Note that if we set $A(n)=1$, and $\varphi(n)=0$, then at broadside where $\sin(\theta) = 0$, the gain would be ($N\ G_e$). This represents the maximum gain of the array, which typically will not exceed $n\pi$, and is a familiar figure. It should be noted that in practical array design, the element pattern characteristics are greatly influenced by mutual coupling and that the characteristic of elements at the edge of the array can deviate significantly from those near the center.

ROTMAN BOOTLACE LENS

Another method of feeding an array of elements is to use a lens such as the Rotman (rhymes with rotten) Bootlace type shown in Figure 5. The lens consists of a parallel plate region (nowadays microstrip or stripline construction) and cables of specified length connecting the array of elements to the parallel plate region. The geometry of the lens and the cable lengths are designed so that all ray paths traced from a beam port on the right side to its associated wavefront on the left array port side, are equal. This tailoring of the design is accomplished at three focus points (beam ports 1, 4, and 7 in Figure 5). Departure from perfect focus at intermediate beam ports is negligible in most designs.

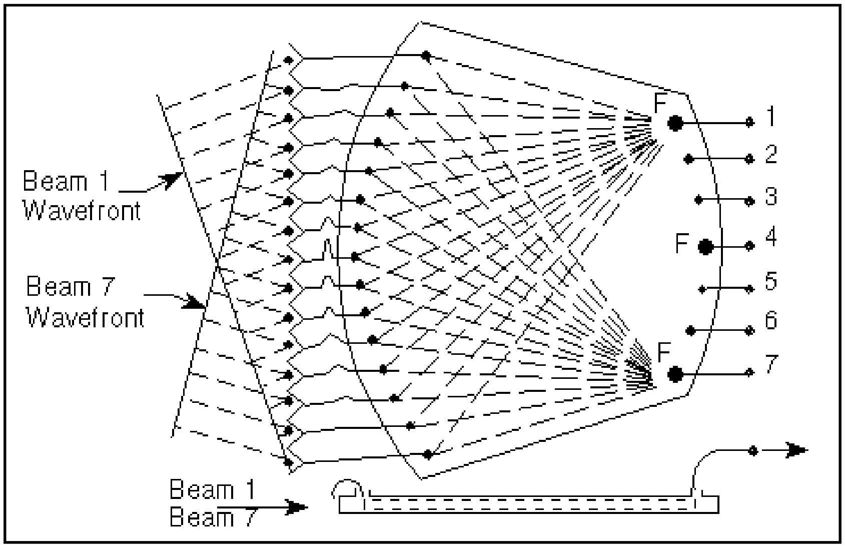

Figure 5. Rotman Bootlace Lens.

The Rotman lens provides both true time delay phase shift and amplitude taper in one lens component. The true time delay is one of the distinct advantages of the lens over the phase shifted array since that makes it independent of frequency. To understand how the taper is obtained requires knowledge of the parallel plate region. For a stripline design the unit would consist of a large flat plate-like center conductor sandwiched between two ground planes, and having a shape much like that of the plan view outline shown in Figure 5 with individual tapered launchers (connectors) attached to each beam port and array port. If the antenna is in the receive mode, the energy intercepted on the array port side can be controlled by the angle subtended by the tapered sections of the connector (launcher) much like a larger antenna would intercept a larger portion of energy from free space.

Unlike the phased array with its fine beam steering, the Rotman lens provides only a distinct set of beams. Fine steering is obtained by combining beams either equally or unequally to form intermediate beams. As can be seen in Figure 6, this results in a broader beam with less gain but lower side lobes than the primary beams.

High transmit power can be obtained using a Rotman lens by placing a low power amplifier between each lens output port and its antenna. In this case a separate Rotman lens would have to be used for receiving.

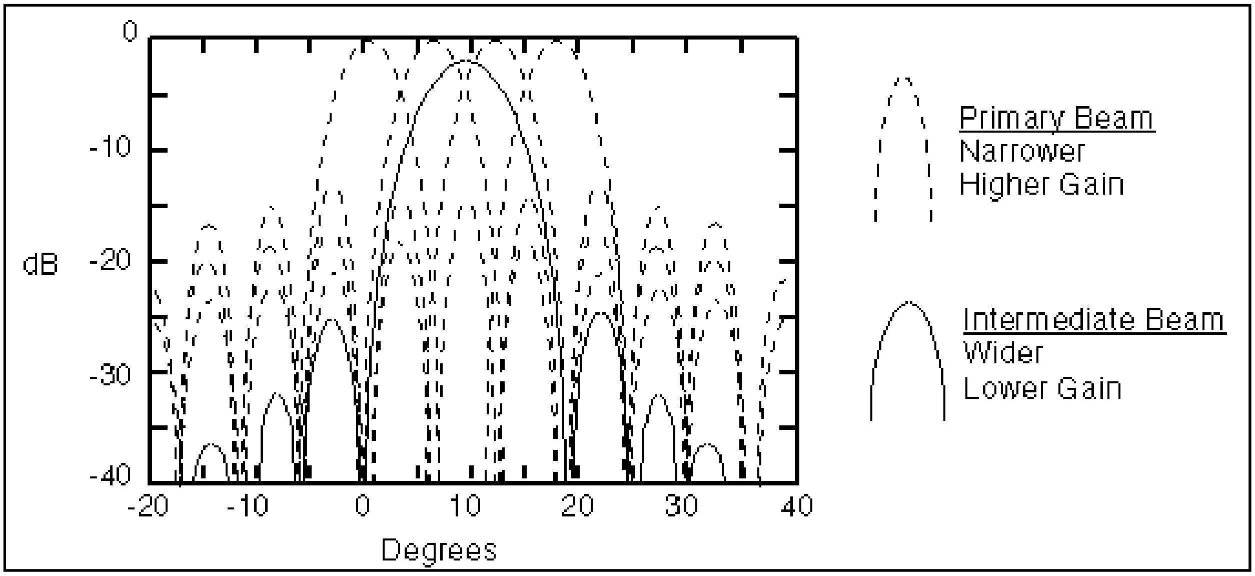

Figure 6. Primary and Intermediate Beam Formation in Lens Arrays.

This page intentionally left blank.

ANTENNA NEAR FIELD

As noted in the sections on RF propagation and the radar equation, electromagnetic radiation expands spherically (Figure 1) and the power density at a long range (R) from the transmitting antenna is:

$$P_D = \frac{P_t G_t}{4\pi R^2} \qquad \textbf{[1]}$$

When the range is large, the spherical surface of uniform power density appears flat to a receiving antenna which is very small compared to the surface of the sphere. This is why the far field wave front is considered planar and the rays approximately parallel. Also, it is apparent that at some shorter range, the spherical surface no longer appears flat, even to a very small receiving antenna.

The planer, parallel ray approximation is valid for distances greater than the distance where the phase error is 1/16 of a wavelength or 22.5 degrees. This distance is given by

$R_{ff} = \frac{2D^2}{\lambda}$ where λ is the wavelength and D is the largest dimension of the transmit antenna. **[2]**

Antenna measurements made at distances greater than R_{ff} generally result in negligible pattern error. Distances less than R_{ff} is termed the near-field.

If the same size antenna is used for multiple frequencies, R_{ff} will increase with increasing frequency. However, if various size antennas are used for different frequencies and each antenna is designed with D as a function of λ (λ/2 to 100λ), then R_{ff} will vary from $c/2f$ to $20000c/f$. In this case R_{ff} will decrease with increasing frequency. For example: a 10λ antenna at 3 GHZ has a D of 100 cm and corresponding R_{ff} of 20 m, while a 10λ antenna at 30 GHz has a D of 10 cm and corresponding R_{ff} of 2 m.

While the above analogy provides an image of the difference between the near and far fields, the relationship must be defined as a characteristic of the transmitting antenna.

Actual antennas, of course, are not ideal point source radiators but have physical dimensions. If the transmitting antenna placed at the origin of Figure 1 occupies distance D along the Z-axis and is boresighted along the Y-axis (φ = 90), then the geometry of point P on the sphere is represented in two dimensions by Figure 2. For convenience, the antenna is represented by a series of point sources in an array.

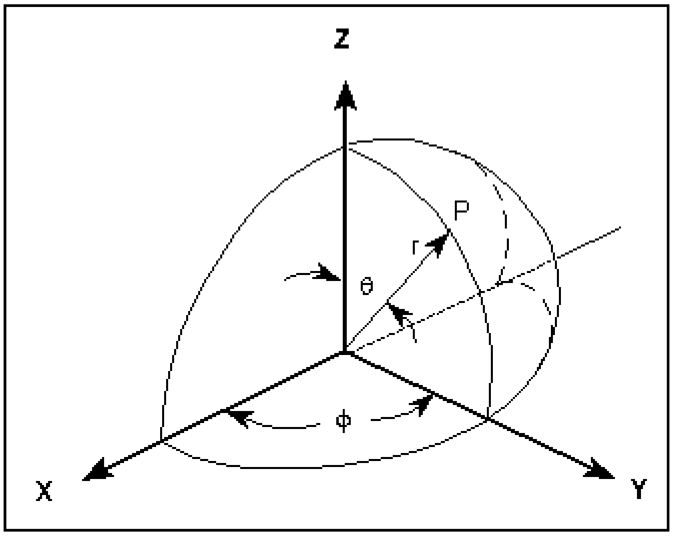

Figure 1. Spherical Radiation to Point "P" from an Ideal Point Source.

When point P is close to the antenna, as in Figure 2, then the difference in distance of the two rays r and R taken respectively from the center of the antenna and the outer edge of the antenna varies as point P changes.

Derivation of equation [2] is given as follows:

From Figure 2, the following applies:

$$r^2 = z^2 + y^2 \qquad [3]$$

$$z = r \cos \theta \qquad [4]$$

$$y = r \sin \theta \qquad \text{and} \qquad [5]$$

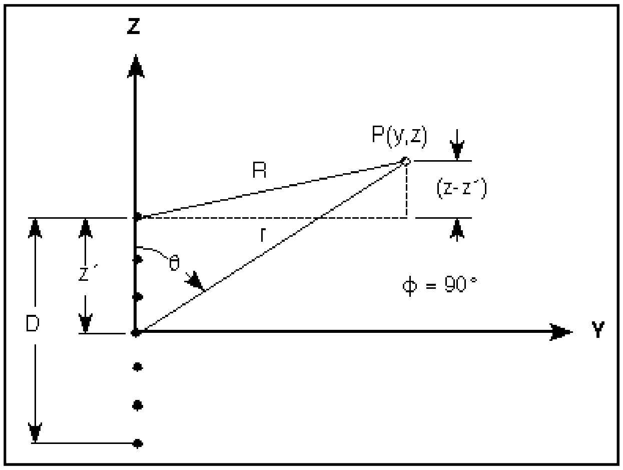

Figure 2. Near Field Geometry of Point "P" for a Non-Ideal Radiator With Dimension D.

$$R = \sqrt{y^2 + (z - z')^2} = \sqrt{y^2 + z^2 - 2z'z + (z')^2} \qquad [6]$$

Substituting [3] and [4] into [6] $R = \sqrt{r^2 + [-2(r \cos \theta) z' + (z')^2]}$ [7]
which puts point P into spherical coordinates.

Equation [7] can be expanded by the binomial theorem which for the first three terms, reduces to:

$$R = r - z' \cos \theta + \frac{(z')^2 \sin^2 \theta}{2r} + ... \qquad [8]$$

In the parallel ray approximation for far field calculations (Figure 3) the third term of [8] is neglected.

The distance where the far field begins (R_{ff}) (or where the near field ends) is the value of r when the error in R due to neglecting the third term of equation [8], equals 1/16 of a wavelength.

R_{ff} is usually calculated on boresight, so θ = 90° and the second term of equation [8] equals zero (Cos 90° = 0), therefore from Figure 3, where D is the antenna dimension, R_{ff} is found by equating the third term of [8] to 1/16 wavelength.

$$\frac{(z')^2 \sin^2 \theta}{2\, R_{ff}} = \frac{\lambda}{16}$$

$\text{Sin}\,\theta = \text{Sin}\, 90 = 1 \text{ and } z' = D/2$ so: $$\frac{\left(\frac{D}{2}\right)^2}{2\, R_{ff}} = \frac{\lambda}{16}$$

$$R_{ff} = \frac{16 (D/2)^2}{2\lambda} = \frac{2\, D^2}{\lambda} \qquad [9]$$

Equation [9] is the standard calculation of far field given in all references.

Besides [9] some general rules of thumb for far field conditions are:

$$r >> D \quad \text{or} \quad r >> \lambda$$

If the sphere and point P are a very great distance from the antenna, then the rays are very nearly parallel and this difference is small as in Figure 3.

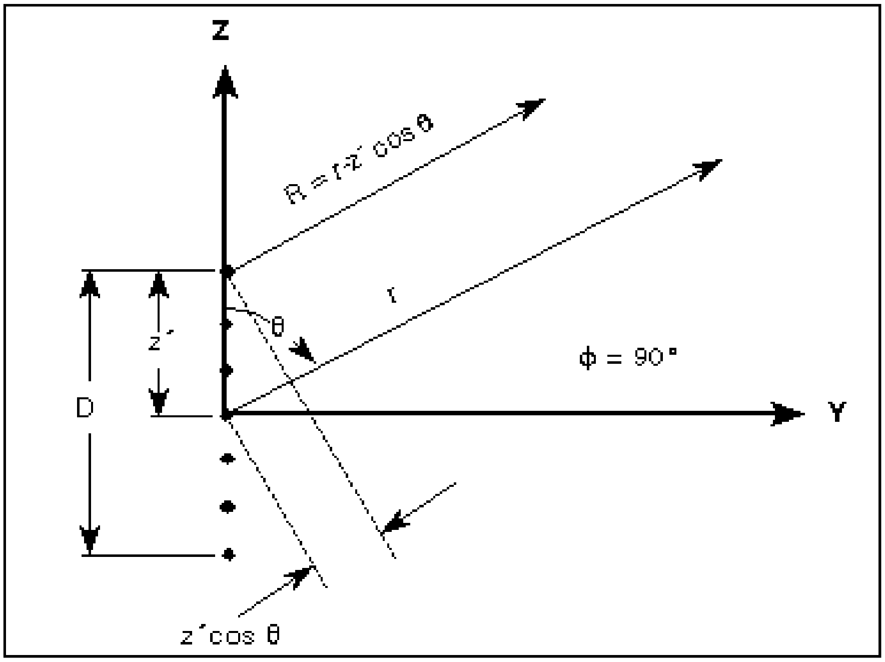

Figure 3. Far Field Parallel Ray Approximation for Calculations.

For reference purposes, a simplified alternative method to derive minimum far field distance approximation for an antenna or an antenna array with aperture size "d" is presented in Figure 4 without the need to resort to the spherical coordinate system and the Binomial Theorem. This approach illustrates a simple application of the Pythagorean Theorem. The P_{FF} symbol represents an arbitrary point in the far field at antenna boresight. Refer to relevant quantities shown in Figure 4.

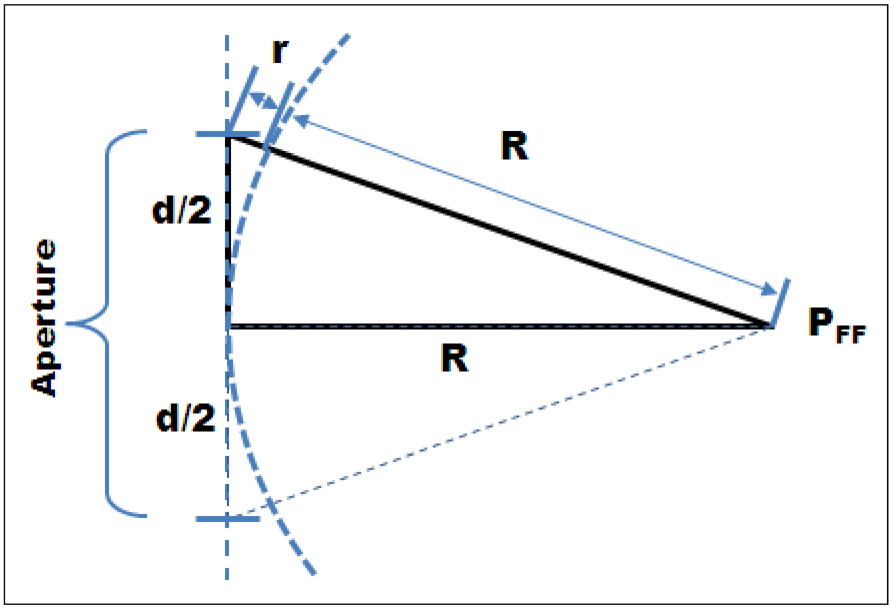

Figure 4. Alternative Geometry for Far Field Estimation.

From Figure 4, the distance "r" is the maximum propagation path difference across a given antenna aperture endpoint locations. Expressing this propagation path difference in terms of a wavelength provides means to derive an equation for generalized far field range relationship. Note a slight change of notation between Figures 3 and 4. A derivation of results already shown in Equations [2] & [9] is given as follows. From Figure 4, the following applies:

$$(R + r)^2 = (d/2)^2 + R^2 \qquad \textbf{[10]}$$

Collecting terms and cancelling,

$$2Rr + r^2 = d^2/4 \qquad \textbf{[11]}$$

Simplifying with a noted valid assumption,

$$2Rr \approx d^2/4, \text{ since } r << R \qquad \textbf{[12]}$$

Solving for "r" (see Figure 4)

$$r \approx d^2/(8R) \qquad \textbf{[13]}$$

For far field condition assume $r \leq \lambda/16$. Substituting "r" in terms of lambda,

$$\lambda/16 \geq r \approx d^2/(8R) \;\rightarrow\; P_{FF} \geq 2d^2/\lambda \qquad \textbf{[14]}$$

As shown, Equation [14] represents the same result as Equation [9] derived earlier.

The power density within the near field varies as a function of the type of aperture illumination and is less than would be calculated by equation [1]. Thus, in the antenna near field there is stored energy. (The complex radiation field equations have imaginary terms indicating reactive power.) Figure 5 shows normalized power density for three different illuminations.

Curve A is for reference only and shows how power density would vary if it were calculated using equation [1].

Curve B shows power density variations on axis for an antenna aperture with a cosine amplitude distribution. This is typical of a horn antenna in the H-plane.

Curve C shows power density variations on axis for a uniformly illuminated antenna aperture or for a line source. This is typical of a horn antenna in the E-plane.

Curve D shows power density variations on axis for an antenna aperture with a tapered illumination. Generally the edge illumination is approximately -10 dB from the center illumination and is typical of a parabolic dish antenna.

Point E For radiation safety purposes, a general rule of thumb for tapered illumination is that the maximum safe level of 10 mW/cm^2 (~200 V/m) is reached in the near field if the level at R_{ff} reaches 0.242 mW/cm^2 as can be verified by computing the power density at point E in Figure 5. (10 mW/cm^2 at point E extrapolates to 0.242 mW/cm^2 [16 dB lower] at $R=R_{ff}$, or Y axis value =1). Figure 1 in Section 3-6 depicts more precise values for radiation hazard exposure.

Point F Far Field Point. At distances closer to the source than this point (near field), the power density from any given antenna is less than that predicted using Curve A. At farther distances, (far field) power densities from all types of antennas are the same.

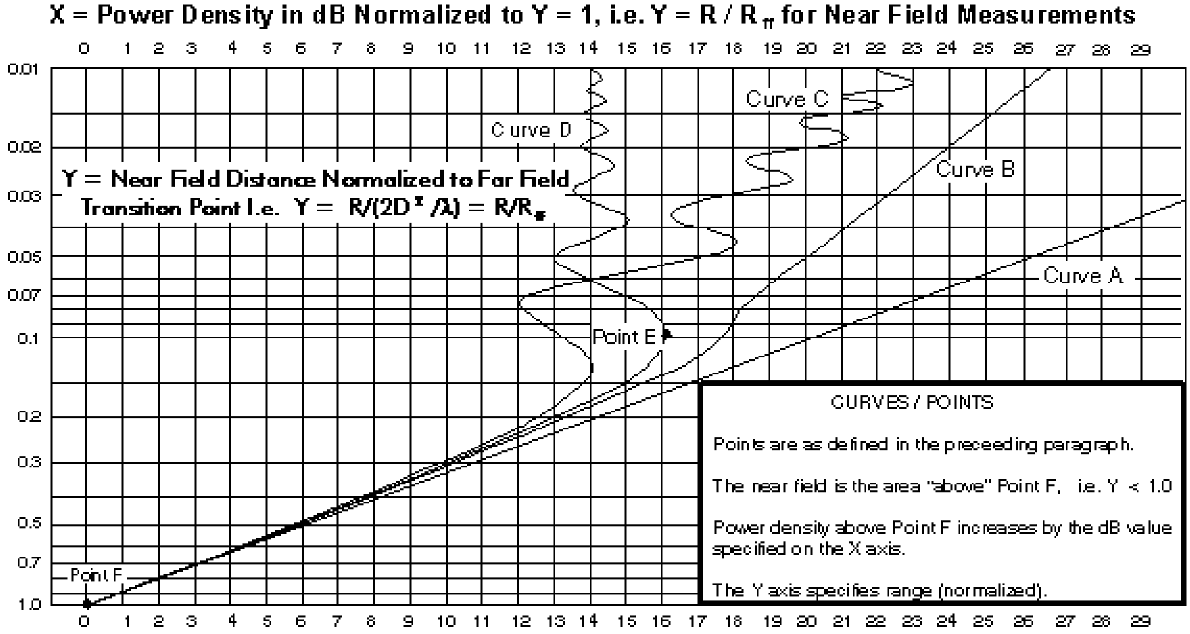

Figure 5. Antenna Near-Field On-Axis Power Density (Normalized) for Various Aperture Illuminations.

FOR FAR FIELD MEASUREMENTS:

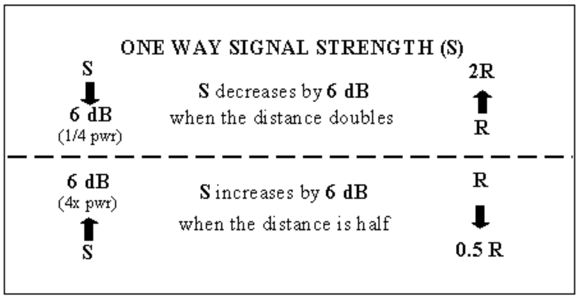
ONE WAY SIGNAL STRENGTH (S)
S
6 dB
(1/4 pwr)
S decreases by 6 dB
when the distance doubles
2R
R
6 dB
(4x pwr)
S
S increases by 6 dB
when the distance is half
R
0.5 R

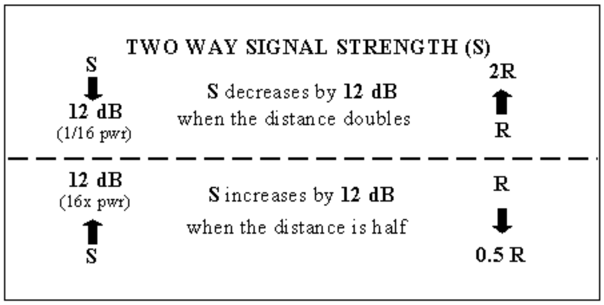
TWO WAY SIGNAL STRENGTH (S)
S
12 dB
(1/16 pwr)
S decreases by 12 dB
when the distance doubles
2R
R
12 dB
(16x pwr)
S
S increases by 12 dB
when the distance is half
R
0.5 R

When free space measurements are performed at a known distance from a source, it is often necessary to know if the measurements are being performed in the far field. As can be seen from Curve A on Figure 5, if the distance is halved (going from 1.0 to 0.5 on the Y axis), the power density will increase by 6 dB (going from 0 to 6 dB on the X axis). Each reduction in range by ½ results in further 6 dB increases. As previously mentioned, Curve A is drawn for reference only in the near field region, since at distances less than R_{ff} the power density increases less than 6 dB when the range is halved. In the far field, all curves converge and Equation [1] applies.

When a measurement is made in free space, a good check to ensure that is was performed in the far field is to repeat the measurement at twice the distance. The power should decrease by exactly 6 dB. A common error is to use 3 dB (the half power point) for comparison. Conversely, the power measurement can be repeated at half the distance, in which case you would look for a 6 dB increase, however the assumed extrapolation conclusion is not as certain, because the first measurement could have been made in the far field, and the second could have been made in the near field.

Care must be exercised in using the $\mathbf{2d^2/\lambda}$ far field measurement criterion. For antennas with moderate sidelobe levels (>-25 dB) pattern errors are negligible and the error in directivity is less than 0.1 dB and this measurement distance suffices in most cases. However low sidelobe antennas require longer measurement distances. For example, maintaining a 1 dB or less sidelobe error for a linear array with a -40 dB sidelobe level requires a measurement distance of $6d^2/\lambda$. (1). This corresponds to criterion for "r" in Equation [14] to be changed from $r \leq \lambda/16$ to $r \leq \lambda/48$ or 7.5 deg.

[1] R. C. Hansen. "Measurement Distance Effects on Low Sidelobe Patterns," *IEEE Transactions on Antennas and Propagation*, Vol. AP-32, No. 6, June 1984.

RADIATION HAZARDS

Radiation Hazard (RADHAZ) describes the hazards of electromagnetic radiation to fuels, electronic hardware, ordnance, and personnel. In the military these hazards are segregated as follows:

1) Hazards of Electromagnetic Radiation to Personnel (HERP)
2) Hazards of Electromagnetic Radiation to Ordnance (HERO)
3) Hazards of Electromagnetic Radiation to Fuel (HERF)

The current industrial specifications for RADHAZ are contained in ANSI/IEEE C95.1-1992 which was used as a reference to create the combined Navy regulation NAVSEA OP3565 / NAVAIR 16-1-529. Volume I contains HERP and HERF limits - its current version is REV 5. Volume II (REV 6) covers HERO. These limits are shown in Figure 1 although all values have been converted to average power density.

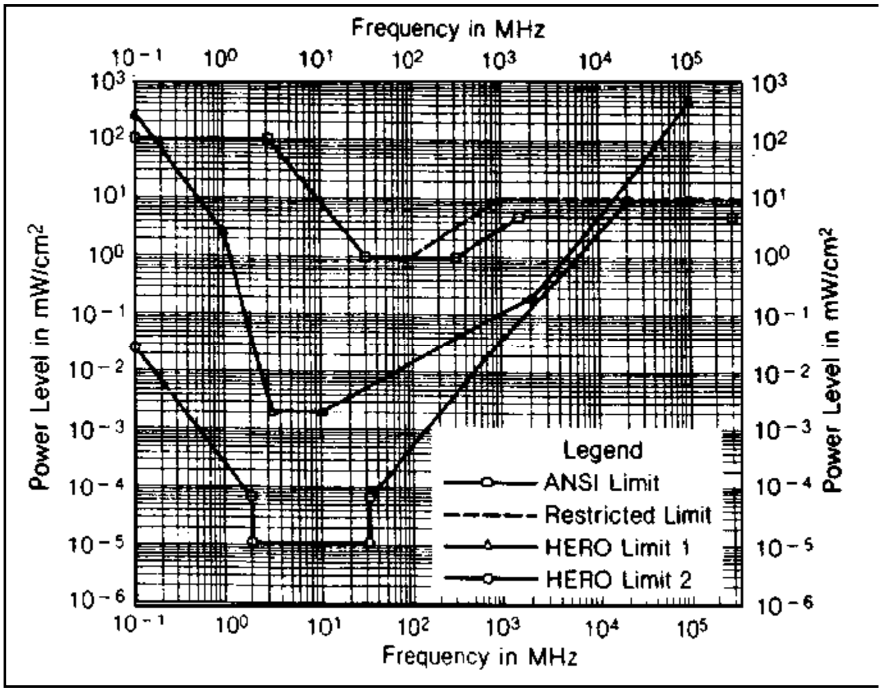

Figure 1. Radiation Hazards to Personnel and Ordnance.

OP 3565 specifies HERO RADHAZ levels at frequencies below 1 GHz in peak value of electric field strength (V/m), while levels above 200 MHz are specified in average power density (mW/cm^2) - note the overlapping frequencies. Since Figure 1 depicts power density as the limits, you must convert the average values to peak field strength for use at lower frequencies. Also many applications of EMC work such as MIL-STD-461 use limits based on the electric (E) field strength in volts/meter. Remember that $P=E^2/R$, and from Section 4-2, we note that $R=377\Omega$ for free space. It can also be shown that the magnetic field strength (H field in Amps/meter) = I/m where I=E/R. Don't forget that RMS = 0.707 Peak. With the units of P_D in mW/cm^2, E in V/m, and H in A/m, then P_D (mW/cm^2) = E^2 / 3770 = 37.7 H^2. It should thus be noted that a 100 times increase in power (mW/cm^2) is only a 10 times increase in V/m.

The potential dangers to ordnance and fuels are obvious because there could be an explosive "chain reaction" by exploding; consequently, these limits are generally lower than personnel limits. There are three **HERO** categories. The HERO limit 2 is for HERO "unsafe" or "unreliable" explosive devices with exposed wires arranged in optimum (most susceptible) receiving orientation. This usually occurs during the assembly/disassembly of ordnance, but also applies to new/untested ordnance until proven "safe" or "susceptible." The HERO limit 1 is for HERO susceptible ordnance fully assembled undergoing normal handling and loading operations. HERO safe ordnance requires no RF radiation precautions. A list of which specific ordnance (by NALC) falls into each category can be found in OP 3565 along with specific frequency restrictions for each piece of ordnance. For example, all missiles of one variety are susceptible

(HERO 1 limits), while another missile has both susceptible and safe variants (with no RADHAZ limits). Other ordnance may be HERO unsafe (HERO 2 limits).

The danger of **HERP** occurs because the body absorbs radiation and significant internal heating may occur without the individual's knowledge because the body does not have internal sensation of heat, and tissue damage may occur before the excess heat can be dissipated. As shown in Figure 1, the current "restricted" limit is for individuals more than 55" tall because they have more body mass. In other words, all people may be exposed to the lower limit, but only persons taller than 55" may be exposed to the higher limit of 10 mW/cm^2.

NAVSEA OP 3565 will be updated in the future to be compatible with DoD INST 6055.11 dated Feb 21, 1995 which supersedes it. The personnel radiation levels in Figures 2 and 3 were taken from the new release of DoD INST 6055.11.

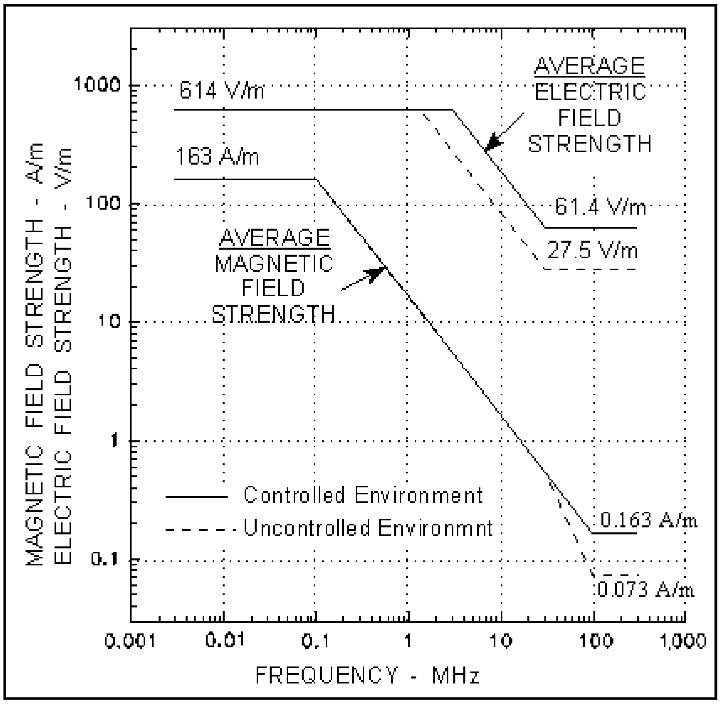

Figure 2. Lower Frequency HERP from DoD INST 6055.11.

Unlike the existing "restricted limit" of NAVSEA OP 3565 discussed above, in the revised DoD instruction for personnel radiation hazards, a different approach to exposure was taken.

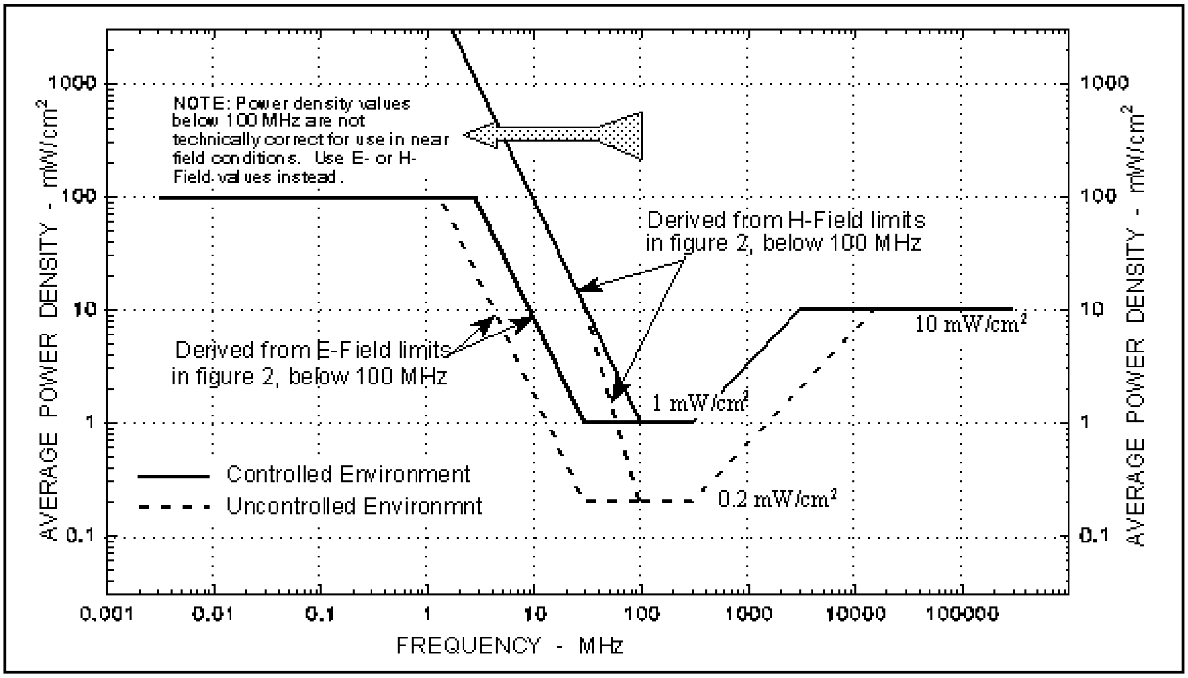

Figure 3. Radiation Hazards to Personnel from DoD INST 6055.11.

Two maximum hazard limits are defined;

1) Controlled Environments - where personnel are aware of the potential danger of RF exposure concurrently with employment, or exposure which may occur due to incidental transient passage through an area, and;
2) Uncontrolled Environments - A lower maximum level where there is no expectation that higher levels should be encountered, such as living quarters.

These Personnel Exposure Limits (PELs) are based on a safety factor of ten times the Specific Absorption Rate (SAR) which might cause bodily harm. The term PEL is equivalent to the terms "Maximum Permissible Exposure (MPE)" and "Radio Frequency Protection Guides (RFPG)" in other publications.

There are several exceptions to the max limits in Figs 2 and 3 (in some cases higher levels are permitted):

- High Power Microwave (HPM) system exposure in a controlled environment, which has a single pulse or multiple pulses lasting less than 10 seconds, has a higher peak E-Field limit of 200 kV/m.
- EMP Simulation Systems in a controlled environment for personnel who are exposed to broad-band (0.1 MHz to 300 GHz) RF are limited to a higher peak E-Field of 100 kV/m.
- The given limits are also increased for pulsed RF fields. In this case the peak power density per pulse for pulse durations < 100 msec and no more than 5 pulses in the period is increased to: PELPulse = PEL x TAVG / 5 x Pulse Width, and the peak E-field is increased to 100 kV/m. If there are more than 5 pulses or they are greater than 100 msec, a time averaged PD should not exceed that shown in Figure 3.
- A rotating or scanning beam likewise reduces the hazard, so although an on-axis hazard might exist, there may be none with a moving beam. The power density may be approximated with:
  PDscan = PDfixed (2 x Beam Width / scan angle)
- Many other special limitations also apply, such as higher limits for partial body exposure, so if in doubt, read the DoD Inst 6055.11 in detail. Field measurements may be measured in accordance with IEEE C95.3-1991.

The PELs listed in Figures 2 and 3 were selected for an average RF exposure time at various frequencies. In a controlled environment, this averaging time was selected as 6 minutes for 0.003 to 15,000 MHz. If the exposure time is less than 6 minutes, then the level may be increased accordingly. Similar time weighted averages apply to uncontrolled environments, but it varies enough with frequency such that DoD INST 6055.11 should be consulted.

NAVSEA OP 3565 contains a list of Navy avionics which transmit RF as well as radars along with their respective hazard patterns. Special training is required for individuals who work in areas which emit RF levels which exceed the uncontrolled levels. Warning signs are also required in areas which exceed either the controlled or uncontrolled limits.

Although E-Field, H-Field, and power density can be mathematically converted in a far-field plane wave environment, the relations provided earlier do not apply in the near field, consequently the E- or H-field strength must be measured independently below 100 MHz. It should be noted that the specifications in NAVSEA OP 3565 for lower frequency HERO limits are listed as peak E-field values, whereas lower RF limits in DoD INST 6055.11 on HERP are in average (RMS) E-field values. Upper frequency restrictions are based on average (RMS) values of power density in both regulations except for certain circumstances.

HERF precautions are of more general concern to fuel truck operators. However, some general guidelines include:

- Do not energize a transmitter (radar/comm) on an aircraft or motor vehicle being fueled or on an adjacent aircraft or vehicle.
- Do not make or break any electrical, ground wire, or tie down connector while fueling.
- Radars able to illuminate fueling areas with a peak power density of 5 W/cm2 should be shut off.
- For shore stations, antennas radiating 250 watts or less should be installed at least 50 ft from fueling areas (at sea 500 watts is the relaxed requirement).
- For antennas which radiate more than 250 watts, the power density at 50 ft from the fueling operation should not be greater than the equivalent power density of a 250 watt transmitter at 50 ft.

ACTIVE ELECTRONICALLY SCANNED ARRAYS (AESA)

Scanning phased arrays employing electronically controlled phase shifters (e.g., PiN diode, ferrite) have been used in high power radar applications since the early 1960s. These are planar arrays and generally use a corporate type feed structure to array rows of elements in one dimension of the array and use an analogous feed to array the rows in the orthogonal dimension of the array, resulting in a single feed point. A high power RF transmitter, usually employing some sort of liquid cooling, is used to excite the array. This architecture is cumbersome and difficult to package and the high power transmitter is a single point of failure for the system.

The maturation of solid state transmit/receive modules (T/R) using Gallium Arsenide (GaAs) technology or more recently Gallium Nitride (GaN) technology has made AESAs a practical reality. The T/R modules consist of a low noise amplifier (LNA) for the receive function and a solid state high power amplifier (SSPA) for realization of the transmit function. The T/R module may also contain the necessary phase control elements for beam scanning. A generic 8-element AESA for one dimension is shown in Figure 1.

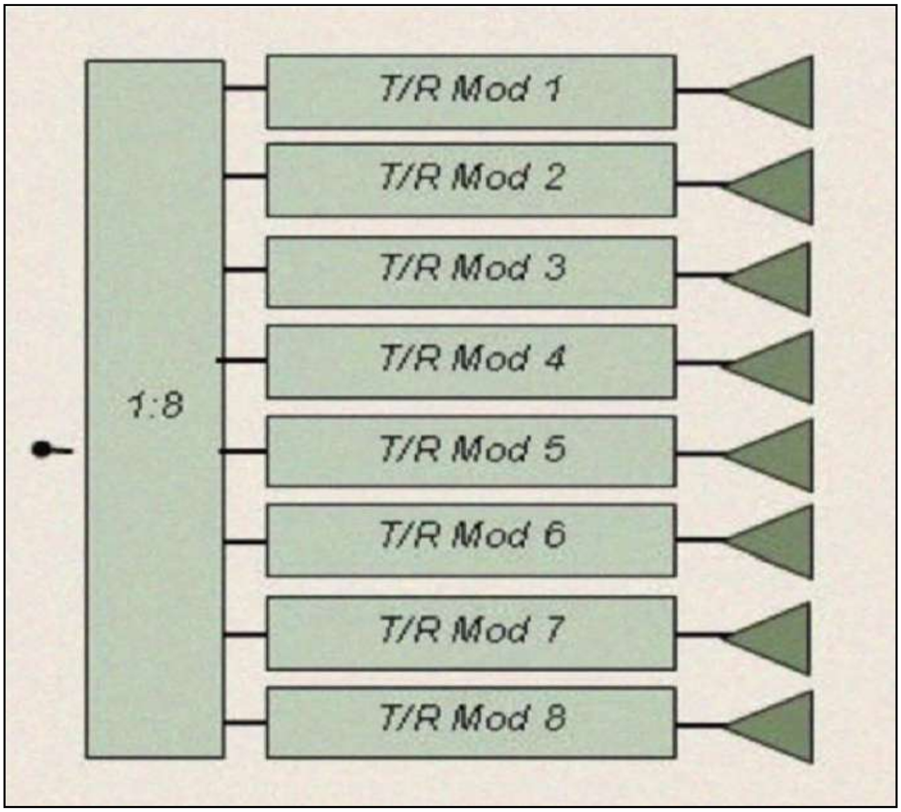

Figure 1. Eight Element AESA.

Electronically scanned phased array designs with decade bandwidths have been reported in the open literature for many years. However those reported are only lab versions and none have been implemented in a fielded system. In addition, the high average power requirement, particularly for EW systems, calls for the use of GaN SSPA technology and this, in conjunction with the dense array packaging requirement, poses severe heat dissipation/cooling issues.

AESAs for EW applications impose design considerations vastly different from an AESA for radar. In contradistinction with high **peak** power radar applications (e.g., AN/APG79), the high **average** power large duty factors (sometimes CW) requirements for EW impose severe design requirements and constraints for the wide bandwidth AESA. Incorporation of an AESA in airborne environment, with its concomitant restrictive volume, further exacerbates design constraints.

The heat dissipation problem is of major concern in modern high power AESAs. Due to the behavior of microwave transistor amplifiers, the power added efficiency (PAE) of a TR module transmitter is typically a relatively small fraction of the total prime power consumption. As a result, an AESA will dissipate a lot of heat, which must be extracted. The reliability of GaAs and GaN MMIC chips improves if the RF amplifier system operates at reduced temperatures. Traditional air cooling used in most established avionic hardware is ill suited to the high packaging density of an AESA. As a result modern AESAs are predominantly liquid cooled. A typical liquid cooling system will use pumps to drive the coolant through channels in the cooling plenum of the array, and then route it to a heat exchanger. In comparison, with a conventional air cooled fighter radar, the AESA will be more reliable but will require more electrical power and more cooling infrastructure, and typically can produce much higher average transmit power.

ADVANTAGES OF AESA RADIATORS

AESAs add many capabilities of their own to those of the Passive Electronically Steered Array (PESAs). Among these are: the ability to form multiple beams, to scan without mechanical steering, to use each transmit/receive module for different roles concurrently, like radar detection, and, more importantly, their multiple wave and scanning frequencies create multiple difficulties for traditional, correlation-type radar detectors.

LOW PROBABILITY OF INTERCEPT

Radar systems work by sending out a signal and then listening for its echo off distant objects. Each of these paths, to and from the target, is subject to the inverse square law of propagation. That means that a radar's received energy drops with the fourth power of distance, which is why radar systems require high powers, often in the megawatt range, to be effective at long range.

The radar signal being sent out is a simple radio signal, and can be received with a simple radio receiver. It is common to use such a receiver in the targets, normally aircraft, to detect radar broadcasts. Unlike the radar unit, which must send the pulse out and then receive its reflection, the target's receiver does not need the reflection and thus the signal drops off only as the square of distance. This means that the receiver is always at an advantage over the radar in terms of range. It will always be able to detect the signal long before the radar can see the target's echo. Since the position of the radar is extremely useful information in an attack on that platform, this means that radars generally must be turned off for lengthy periods if they are subject to attack; this is common on ships, for instance.

Turning that received signal into a useful display is the purpose of the RWR. Unlike the radar, which knows which direction it is sending its signal, the receiver simply gets a pulse of energy and has to interpret it. Since the radio spectrum is filled with noise, the receiver's signal is integrated over a short period of time, making periodic sources like a radar add up and stand out over the random background. Typically RWRs store the detected pulses for a short period of time, and compare their broadcast frequency and pulse repetition frequency against a database of known radars. The rough direction can be calculated using a rotating antenna, or similar passive array, and combined with symbology indicating the likely purpose of the radar - airborne early warning, surface to air missile, etc.

This technique is much less useful against AESA radars. Since the AESA can change its frequency with every pulse, and generally does so using a pseudo-random sequence, integrating over time does not help pull the signal out of the background noise. Nor does the AESA have any sort of fixed pulse repetition frequency, which can also be varied and thus hide any periodic brightening across the entire spectrum. Traditional RWRs suffered significantly decreased effectiveness against AESA radars.

HIGH JAMMING RESISTANCE

Jamming is likewise much more difficult against an AESA. Traditionally, jammers have operated by determining the operating frequency of the radar and then broadcasting a signal on it to confuse the receiver as to which is the "real" pulse and which is the jammer's. This technique works as long as the radar system cannot easily change its operating frequency. When the transmitters were based on klystron tubes this was generally true, and radars, especially airborne ones, had only a few frequencies to choose among. A jammer could listen to those possible frequencies and select the one to be used to jam.

Since an AESA could change its operating frequency with every pulse, and spread the frequencies across a wide band even in a single pulse, jammers are much less effective. Although it is possible to send out broadband white noise against all the possible frequencies, this means the amount of energy being sent at any one frequency is much lower, reducing its effectiveness. In fact, AESAs can then be switched to a receive-only mode, and use these powerful jamming signals instead to track its source, something that required a separate receiver in older platforms.

AESA radars can be much more difficult to detect, and so much more useful in receiving signals from the targets, that they can broadcast continually and still have a very low chance of being detected. This allows such radar systems to generate far more data than traditional radar systems, which can only receive data periodically, greatly improving overall system effectiveness.

OTHER ADVANTAGES

Since each element in an AESA is a powerful radio receiver, active arrays have many roles besides traditional radar. One use is to dedicate several of the elements to reception of common radar signals, eliminating the need for a separate radar warning receiver. The same basic concept can be used to provide traditional radio support, and with some elements also broadcasting, form a very high bandwidth data link.

AESAs are also much more reliable than either a PESA or older designs. Since each module operates independently of the others, single failures have little effect on the operation of the system as a whole. Additionally, the modules individually operate at lower powers and voltages so the need for a large high-voltage power supply is eliminated.

Replacing a mechanically scanned array with a fixed AESA mount can help reduce an aircraft's overall RCS, but some designs (such as the Eurofighter Typhoon) forgo this advantage in order to combine mechanical scanning with electronic scanning and provide a wider angle of total coverage.

This page intentionally left blank.

FRACTAL ANTENNAS

The term fractal, which means broken or irregular fragments, was originally coined by B.B. Mandelbrot **[1]** to describe a family of complex shapes that possess self-similarity in their geometrical structure. In-depth studies of the patterns of nature provided inspiration for the development of fractal geometry where it has been used to model numerous natural objects. Recent advances in the antenna art have led to the application of fractal geometry instead of convention Euclidean geometric concepts to the design of wide bandwidth, low profile, compact antennas.

A fractal antenna is an antenna that uses a fractal, self-similar design to maximize the length, or increase the perimeter (on inside sections or the outer structure), of material that can receive or transmit electromagnetic radiation within a given total surface area or volume. The earliest published reference to use of the term fractal radiators and fractal antennas to refer to fractal shaped antenna elements appeared May 1994 **[2]**.

Such fractal antennas are also referred to as multilevel and space filling curves, but the key aspect lies in their repetition of a motif over two or more scale sizes, or "iterations." For this reason, fractal antennas are very compact, multiband or wideband, and have useful applications in many commercial and military systems.

A good example of a fractal antenna as a space-filling curve is in the form of a Minkowski Island (Figure 1). Here, each line of copper is just a small fraction of a wavelength **[3]**.

A fractal antenna's response differs markedly from traditional antenna designs, in that it is capable of operating with good-to-excellent performance at many different frequencies simultaneously. Normally standard antennas have to be "cut" for the frequency for which they are to be used — and thus the standard antennas only work well at that frequency. This makes the fractal antenna an excellent design for wideband and multiband applications.

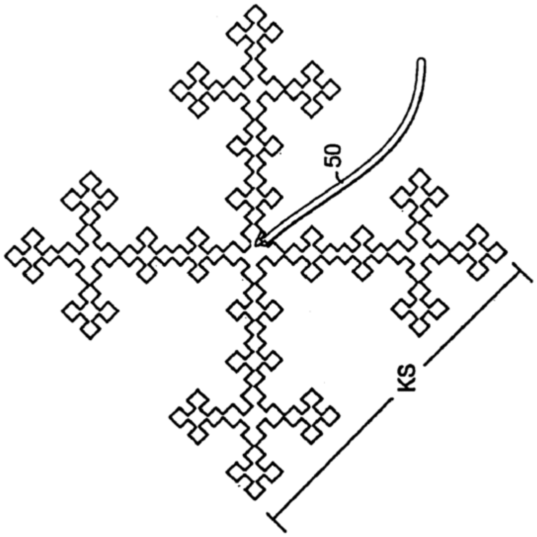

Figure 1. An Example of a Fractal Antenna: a Space-Filling Curve Called a Minkowski Island.

Antenna elements (as opposed to antenna arrays) made from self-similar shapes were first created by Nathan Cohen, then a professor at Boston University, starting in 1995 **[4]**. Cohen's efforts with a variety of fractal antenna designs were first published in 1995 (thus the first scientific publication on fractal antennas), and a number of patents have been issued from the 1995 filing priority of invention. Most allusions to fractal antennas make reference to these "fractal element antennas."

Many fractal element antennas use the fractal structure as a virtual combination of capacitors and inductors. This makes the antenna have so that it has many different resonances that can be selected and adjusted by choosing the proper fractal design. Electrical resonances may not be directly related to a particular scale size of the fractal antenna structure. The physical size of the antenna is unrelated to its resonant or broadband performance. The general rule of antenna length being near target frequency wavelength does not apply itself in the same way with fractal antennas.

This complexity arises because the current on the structure has a complex arrangement caused by the inductance and self capacitance. In general, although their effective electrical length is longer, the fractal element antennas are physically smaller.

Fractal element antennas are shrunken compared to conventional designs and do not need additional components. In general, the fractal dimension of a fractal antenna is a poor predictor of its performance and application. Not all fractal antennas work well for a given application or set of applications. Computer search methods, optimization algorithms, and antenna simulations are commonly used to identify which fractal antenna designs best meet the need of the application.

Although the first validation of the technology was published as early as 1995 **[4]** recent independent studies show advantages of the fractal element technology in real-life applications, such as radio frequency identification and cell phones.

A different and also useful attribute of some fractal element antennas is their self-scaling aspect. In 1999, it was discovered that self-similarity was one of the underlying requirements to make antennas "invariant" (same radiation properties) at a number or range of frequencies. Previously, under Rumsey's Frequency Independent Antenna Principle, it was believed that antennas had to be defined by angles for this to be true; the 1999 analysis **[5]**, based on Maxwell's equations, showed this to be a subset of the more general set of self-similar conditions. Hence fractal antennas offer a closed-form and unique insight into a key aspect of electromagnetic phenomena to wit the invariance property of Maxwell's equations.

Antenna tuning units are typically not required on fractal antennas due to their wide bandwidth and complex resonance. However, if a transmitting antenna has deep nulls in its response or has electromagnetic structural issues that require equalization then an antenna tuning unit should be used.

In addition to their use as antennas, fractals have also found application in other antenna system components including loads, counterpoises, and ground planes. Confusion by those who claim "grain of rice"-sized fractal antennas arises, because such fractal structures serve the purpose of loads and counterpoises, rather than bona fide antennas.

Fractal inductors and fractal tuned circuits (fractal resonators) were also discovered and invented simultaneously with fractal element antennas. An emerging example of such is in metamaterials. A recent report demonstrates using close-packed fractal resonators to make the first wideband metamaterial "invisibility cloak" at microwave frequencies **[6]**. Fractal filters (a type of tuned circuit) are another example of fractal geometry in microwave componentry.

As fractals can be used as counterpoises, loads, ground planes, and filters, all parts that can be integrated with antennas, they are considered parts of some antenna systems and thus are discussed in the context of fractal antennas.

REFERENCES

[1] B. B. Mandelbrot. *The Fractal Geometry of Nature.* New York, W. H. Freeman, 1983.

[2] D. H. Wemer. "Fractal Radiators," *Proceedings of the 4th Annual 1994 IEEE Mohawk Valley~Section Dual-Use Technologies* and *Applications Conference,* Volume I, SUNY Institute of Technology at UticalRomc, New York, 23-26 May 1994, pp. 478-482.

[3] J. P. Gianvittorio and Y. Rahmat-Samii. "Fractal Antennas: A Novel Antenna Miniaturization Technique, and Applications," *Antennas and Propagation Magazine*, *IEEE*, Vol. 44, No. 1, February 2002, pp. 20-36.

[4] N. Cohen. "Fractal Antennas," *Communications Quarterly*, Vol. 9, Summer 1995.

[5] R. Hohlfeld and N. Cohen. "Self-Similarity and the Geometric Requirements for Frequency Independence in Antenna," *Fractals,* Vol. 7, No. 1, 1999, pp. 79–84. doi:10.1142/S0218348X99000098.

[6] D. Schurig, J. J. Mock, B. J. Justice, S. A. Cummer, J. B. Pendry, A. F. Starr, and D. R. Smith. "Metamaterial Electromagnetic Cloak at Microwave Frequencies," *Science,* Vol. 314, No. 5801, 10 November 2006, pp. 977-980. Published online 19 October 2006 [DOI:10.1126/science.1133628].

This page intentionally left blank.

RADAR EQUATIONS

Field Intensity and Power Density....4-1
Power Density....4-2
One-Way Radar Equation / RF Propagation....4-3
Two-Way Radar Equation (Monostatic)....4-4
Alternate Two-Way Radar Equation....4-5
Two-Way Radar Equation (Bistatic)....4-6
Jamming to Signal (J/S) Ratio - Constant Power [Saturated] Jamming....4-7
Burn-Through / Crossover Range....4-8
Support Jamming....4-9
Jamming to Signal (J/S) Ratio - Constant Gain [Linear] Jamming....4-10
Radar Cross Section (RCS)....4-11
Emission Control (EMCON)....4-12
EW Jamming Techniques....4-13

This page intentionally left blank.

FIELD INTENSITY and POWER DENSITY

Sometimes it is necessary to know the actual field intensity or power density at a given distance from a transmitter instead of the signal strength received by an antenna. Field intensity or power density calculations are necessary when estimating electromagnetic interference (EMI) effects, when determining potential radiation hazards (personnel safety), or in determining or verifying specifications.

Field intensity (field strength) is a general term that usually means the magnitude of the electric field vector, commonly expressed in volts per meter. At frequencies above 100 MHZ, and particularly above one GHz, power density (P_D) terminology is more often used than field strength.
Power density and field intensity are related by equation [1]:

$$P_D = \frac{E^2}{Z_0} = \frac{E^2}{120\pi} = \frac{E^2}{377} \qquad \textbf{[1]}$$

where P_D is in W/m^2, E is the RMS value of the field in volts/meter and 377 ohms is the characteristic impedance of free space. When the units of P_D are in mW/cm^2, then P_D (mW/cm^2) = E^2/3,770.

Conversions between field strength and power density when the impedance is 377 ohms, can be obtained from Table 1. It should be noted that to convert dBm/m^2 to dBμV/m add 115.76 dB. Sample calculations for both field intensity and power density in the far field of a transmitting antenna are in Section 4-2 and Section 4-8. Refer to chapter 3 on antennas for the definitions of near field and far field.

Note that the "/" term before m, m^2, and cm^2 in Table 1 mean "per," i.e., dBm per m^2, not to be confused with the division sign which is valid for the Table 1 equation $P=E^2/Z_o$. Remember that in order to obtain dBm from dBm/m^2 given a certain area, you must add the logarithm of the area, not multiply. The values in the table are rounded to the nearest dBW, dBm, etc. per m^2 so the results are less precise than a typical handheld calculator and may be up to ½ dB off.

VOLTAGE MEASUREMENTS

Coaxial cabling typically has input impedances of 50, 75, and 93Ω, (±2) with 50Ω being the most common. Other types of cabling include the following: TV cable is 75Ω (coaxial) or 300Ω (twin-lead), audio public address (PA) is 600Ω, audio speakers are 3.2 (4), 8, or 16Ω.

In the 50Ω case, power and voltage are related by:

$$P = \frac{E^2}{Z_0} = \frac{E^2}{50} = 50\, I^2 \qquad \textbf{[2]}$$

Conversions between measured power, voltage, and current where the typical impedance is 50 ohms can be obtained from Table 2. The dBμA current values are given because frequently a current probe is used during laboratory tests to determine the powerline input current to the system.

MATCHING CABLING IMPEDANCE

In performing measurements, we must take into account an impedance mismatch between measurement devices (typically 50 ohms) and free space (377 ohms).

Table 1. Conversion Table - Field Intensity and Power Density.
$P_D = E^2/Z_0$ (Related by free space impedance = 377 ohms)

E (Volts/m)	20 log 10^6 (E) (dBµV/m)	P_D (watts/m^2)	10 Log P_D (dBW/m^2)	Watts/cm^2	dBW/cm^2	mW/cm^2	dBm/cm^2	dBm/m^2
7,000	197	130,000	+51	13	+11	13,000	+41	+81
5,000	194	66,300	+48	6.6	+8	6,630	+38	+78
3,000	190	23,900	+44	2.4	+4	2,390	+34	+74
4,000	186	10,600	+40	1.1	0	1,060	+30	+70
1,000	180	2,650	+34	.27	-6	265	+24	+64
700	177	1,300	+31	.13	-9	130	+21	+61
500	174	663	+28	.066	-12	66	+18	+58
300	170	239	+24	.024	-16	24	+14	+54
200	166	106	+20	.011	-20	11	+10	+50
100	160	27	+14	.0027	-26	2.7	+4	+44
70	157	13	+11	$1.3x10^{-3}$	-29	1.3	+1	+41
50	154	6.6	+8	$6.6x10^{-4}$	-32	.66	-2	+38
30	150	2.4	+4	$2.4x10^{-4}$	-36	.24	-6	+34
20	146	1.1	+0	$1.1x10^{-4}$	-40	.11	-10	+30
10	140	.27	-6	$2.7x10^{-5}$	-46	.027	-16	+24
7	137	.13	-9	$1.3x10^{-5}$	-49	.013	-19	+21
5	134	.066	-12	$6.6x10^{-6}$	-52	$66x10^{-4}$	-22	+18
3	130	.024	-16	$2.4x10^{-6}$	-56	$24x10^{-4}$	-26	+14
2	126	.011	-20	$1.1x10^{-6}$	-60	$11x10^{-4}$	-30	+10
1	120	.0027	-26	$2.7x10^{-7}$	-66	$2.7x10^{-4}$	-36	+4
0.7	117	$1.3x10^{-3}$	-29	$1.3x10^{-7}$	-69	$1.3x10^{-4}$	-39	+1
0.5	114	$6.6x10^{-4}$	-32	$6.6x10^{-8}$	-72	$66x10^{-4}$	-42	-2
0.3	110	$2.4x10^{-4}$	-36	$2.4x10^{-8}$	-76	$24x10^{-4}$	-46	-6
0.2	106	$1.1x10^{-4}$	-40	$1.1x10^{-8}$	-80	$11x10^{-4}$	-50	-10
0.1	100	$2.7x10^{-5}$	-46	$2.7x10^{-9}$	-86	$2.7x10^{-6}$	-56	-16
$70x10^{-3}$	97	$1.3x10^{-5}$	-49	$1.3x10^{-9}$	-89	$1.3x10^{-6}$	-59	-19
$50x10^{-3}$	94	$6.6x10^{-6}$	-52	$6.6x10^{-10}$	-92	$66x10^{-8}$	-62	-22
$30x10^{-3}$	90	$2.4x10^{-6}$	-56	$2.4x10^{-10}$	-96	$24x10^{-8}$	-66	-26
$20x10^{-3}$	86	$1.1x10^{-6}$	-60	$1.1x10^{-10}$	-100	$11x10^{-8}$	-70	-30
$10x10^{-3}$	80	$2.7x10^{-7}$	-66	$2.7x10^{-11}$	-106	$2.7x10^{-8}$	-76	-36
$7x10^{-3}$	77	$1.3x10^{-7}$	-69	$1.3x10^{-11}$	-109	$1.3x10^{-8}$	-79	-39
$5x10^{-3}$	74	$6.6x10^{-8}$	-72	$6.6x10^{-12}$	-112	$66x10^{-10}$	-82	-42
$3x10^{-3}$	70	$2.4x10^{-8}$	-76	$2.4x10^{-12}$	-116	$24x10^{-10}$	-86	-46
$2x10^{-3}$	66	$1.1x10^{-8}$	-80	$1.1x10^{-12}$	-120	$11x10^{-10}$	-90	-50
$1x10^{-3}$	60	$2.7x10^{-9}$	-86	$2.7x10^{-13}$	-126	$2.7x10^{-10}$	-96	-56
$7x10^{-4}$	57	$1.3x10^{-9}$	-89	$1.3x10^{-13}$	-129	$1.3x10^{-10}$	-99	-59
$5x10^{-4}$	54	$6.6x10^{-10}$	-92	$6.6x10^{-14}$	-132	$66x10^{-12}$	-102	-62
$3x10^{-4}$	50	$2.4x10^{-10}$	-96	$2.4x10^{-14}$	-136	$24x10^{-12}$	-106	-66
$2x10^{-4}$	46	$1.1x10^{-10}$	-100	$1.1x10^{-14}$	-140	$11x10^{-12}$	-110	-70
$1x10^{-4}$	40	$2.7x10^{-11}$	-106	$2.7x10^{-15}$	-146	$2.7x10^{-12}$	-116	-76
$7x10^{-5}$	37	$1.3x10^{-11}$	-109	$1.3x10^{-15}$	-149	$1.3x10^{-12}$	-119	-79
$5x10^{-5}$	34	$6.6x10^{-12}$	-112	$6.6x10^{-16}$	-152	$66x10^{-14}$	-122	-82
$3x10^{-5}$	30	$2.4x10^{-12}$	-116	$2.4x10^{-16}$	-156	$24x10^{-14}$	-126	-86
$2x10^{-5}$	26	$1.1x10^{-12}$	-120	$1.1x10^{-16}$	-160	$11x10^{-14}$	-130	-90
$1x10^{-5}$	20	$2.7x10^{-13}$	-126	$2.7x10^{-17}$	-166	$2.7x10^{-14}$	-136	-96
$7x10^{-6}$	17	$1.3x10^{-13}$	-129	$1.3x10^{-17}$	-169	$1.3x10^{-14}$	-139	-99
$5x10^{-6}$	14	$6.6x10^{-14}$	-132	$6.6x10^{-18}$	-172	$66x10^{-16}$	-142	-102
$3x10^{-6}$	10	$2.4x10^{-14}$	-136	$2.4x10^{-18}$	-176	$24x10^{-16}$	-146	-106
$2x10^{-6}$	6	$1.1x10^{-14}$	-140	$1.1x10^{-18}$	-180	$11x10^{-16}$	-150	-110
$1x10^{-6}$	0	$2.7x10^{-15}$	-146	$2.7x10^{-19}$	-186	$2.7x10^{-16}$	-156	-116

NOTE: Numbers in table rounded off.

FIELD STRENGTH APPROACH

To account for the impedance difference, the antenna factor (AF) is defined as: AF=E/V, where E is field intensity which can be expressed in terms taking 377 ohms into account and V is measured voltage which can be expressed in terms taking 50 ohms into account. Details are provided in Section 4-12.

POWER DENSITY APPROACH

To account for the impedance difference, the antenna's effective capture area term, A_e relates free space power density P_D with received power, P_r, i.e. $P_r = P_D A_e$. A_e is a function of frequency and antenna gain and is related to AF as shown in Section 4-12.

SAMPLE CALCULATIONS

Section 4-2 provides sample calculations using power density and power terms from Tables 1 and 2, whereas Section 4-12 uses these terms plus field intensity and voltage terms from Table 1 and Table 2. Refer the examples in Section 4-12 for usage of the conversions while converting free space values of power density to actual measurements with a spectrum analyzer attached by coaxial cable to a receiving antenna.

Conversion Between Field Intensity (Table 1) and Power Received (Table 2).

Power received (watts or milliwatts) can be expressed in terms of field intensity (volts/meter or μv/meter) using equation [3]:

$$\text{Power received}\,(P_r) = \frac{E^2}{480\pi^2}\,\frac{c^2}{f^2}\,G \qquad \textbf{[3]}$$

or in log form: $10 \log P_r = 20 \log E + 10 \log G - 20 \log f + 10 \log (c^2/480\pi^2)$ **[4]**

Then $10 \log Pr = 20 \log E_1 + 10 \log G - 20 \log f_1 + K_4$ **[5]**

$$\text{Where } K_4 = 10 \log \left[\frac{c^2}{480\pi^2} \bullet \left(\begin{matrix} \text{conversions} \\ \text{as required} \end{matrix} \; \frac{(\text{Watts to mW})}{(\text{volts to } \mu v)^2\,(\text{Hz to MHz or GHz})^2} \right) \right]$$

The derivation of equation [3] follows:

$P_D = E^2/120\pi$ Eq [1], Section 4-1, terms (v^2/Ω)

$A_e = \lambda^2 G/4\pi$ Eq [8], Section 3-1, terms (m^2)

$P_r = P_D A_e$ Eq [2], Section 4-3, terms $(W/m^2)(m^2)$

$\therefore\ P_r = (E2/120\pi)(\lambda^2 G/4\pi)$ terms $(v^2/m^2\Omega)(m^2)$

$\lambda = c/f$ Section 2-3, terms (m/sec)(sec)

$\therefore P_r = (E^2/480\pi^2)(c^2 G/f^2)$ which is equation [3]

terms $(v^2/m^2\Omega)(m^2/sec^2)(sec^2)$ or v^2/Ω = watts

Values of K_4 (dB)

P_r	E_1	f_1 (Hz)	f_1 (MHz)	f_1 (GHz)
Watts (dBW)	volts/meter	132.8	12.8	-47.2
	μv/meter	12.8	-107.2	-167.2
mW (dBm)	volts/meter	162.8	42.8	-17.2
	μv/meter	42.8	-77.2	-137.7

Table 2. Conversion Table - Volts to Watts and dBμA.
($P_x = V_x^2/Z$ - Related by line impedance of 50 Ω)

Volts	dBV	dBμV	Watts	dBW	dBm	dBμA
700	56.0	176.0	9800	39.9	69.9	142.9
500	53.9	173.9	5000	37.0	67.0	140.0
300	49.5	169.5	1800	32.5	62.5	135.5
200	46.0	166.0	800	29.0	59.0	132.0
100	40.0	160.0	200	23.0	53.0	126.0
70	36.9	156.9	98	19.9	49.9	122.9
50	34.0	154.0	50	17.0	47.0	120.0
30	29.5	149.5	18	12.5	42.5	115.5
20	26.0	146.0	8	9.0	39.0	112.0
10	20.0	140.0	2	3.0	33.0	106.0
7	16.9	136.9	0.8	0	29.9	102.9
5	14.0	134.0	0.5	-3.0	27.0	100.0
3	9.5	129.5	0.18	-7.4	22.5	95.6
2	6.0	126.0	0.08	-11.0	19.0	92.0
1	0	120.0	0.02	-17.0	13.0	86.0
0.7	-3.1	116.9	9.8×10^{-3}	-20.1	9.9	82.9
0.5	-6.0	114.0	5.0×10^{-3}	-23.0	7.0	80.0
0.3	-10.5	109.5	1.8×10^{-3}	-27.4	2.6	75.6
0.2	-14.0	106.0	8.0×10^{-4}	-31.0	-1.0	72.0
0.1	-20.0	100.0	2.0×10^{-4}	-37.0	-7.0	66.0
.07	-23.1	96.9	9.8×10^{-5}	-40.1	-10.1	62.9
.05	-26.0	94.0	5.0×10^{-5}	-43.0	-13.0	60.0
.03	-30.5	89.5	1.8×10^{-5}	-47.4	-17.7	55.6
.02	-34.0	86.0	8.0×10^{-6}	-51.0	-21.0	52.0
.01	-40.0	80.0	2.0×10^{-6}	-57.0	-27.0	46.0
7×10^{-3}	-43.1	76.9	9.8×10^{-7}	-60.1	-30.1	42.9
5×10^{-3}	-46.0	74.0	5.0×10^{-7}	-63.0	-33.0	40.0
3×10^{-3}	-50.5	69.5	1.8×10^{-7}	-67.4	-37.4	35.6
2×10^{-3}	-54.0	66.0	8.0×10^{-8}	-71.0	-41.0	32.0
1×10^{-3}	-60.0	60.0	2.0×10^{-8}	-77.0	-47.0	26.0
7×10^{-4}	-64.1	56.9	9.8×10^{-9}	-80.1	-50.1	22.9
5×10^{-4}	-66.0	54.0	5.0×10^{-9}	-83.0	-53.0	20.0
3×10^{-4}	-70.5	49.5	1.8×10^{-9}	-87.4	-57.4	15.6
2×10^{-4}	-74.0	46.0	8.0×10^{-10}	-91.0	-61.0	12.0
1×10^{-4}	-80.0	40.0	2.0×10^{-10}	-97.0	-67.0	6.0
7×10^{-5}	-84.1	36.9	9.8×10^{-11}	-100.1	-70.1	2.9
5×10^{-5}	-86.0	34.0	5.0×10^{-11}	-103.0	-73.0	0
3×10^{-5}	-90.5	29.5	1.8×10^{-11}	-107.4	-77.4	-4.4
2×10^{-5}	-94.0	26.0	8.0×10^{-12}	-111.0	-81.0	-8.0
1×10^{-5}	-100.0	20.0	2.0×10^{-12}	-117.0	-87.0	-14.0
7×10^{-6}	-104.1	16.9	9.8×10^{-13}	-120.1	-90.1	-17.1
5×10^{-6}	-106.0	14.0	5.0×10^{-13}	-123.0	-93.0	-20.0
3×10^{-6}	-110.5	9.5	1.8×10^{-13}	-127.4	-97.4	-24.4
2×10^{-6}	-114.0	6.0	8.0×10^{-14}	-131.0	-101.0	-28.0
1×10^{-6}	-120.0	0	2.0×10^{-14}	-137.0	-107.0	-34.0
7×10^{-7}	-124.1	-3.1	9.8×10^{-15}	-140.1	-110.1	-37.1
5×10^{-7}	-126.0	-6.0	5.0×10^{-15}	-143.0	-113.0	-40.0
3×10^{-7}	-130.5	-10.5	1.8×10^{-15}	-147.4	-117.4	-44.4
2×10^{-7}	-134.0	-14.0	8.0×10^{-16}	-151.0	-121.0	-48.0
1×10^{-7}	-140.0	-20.0	2.0×10^{-16}	-157.0	-127.0	-54.0

POWER DENSITY

Radio Frequency (RF) propagation is defined as the travel of electromagnetic waves through or along a medium. For RF propagation between approximately 100 MHz and 10 GHz, radio waves travel very much as they do in free space and travel in a direct line of sight. There is a very slight difference in the dielectric constants of space and air. The dielectric constant of space is one. The dielectric constant of air at sea level is 1.000536. In all but the highest precision calculations, the slight difference is neglected.

From chapter 3, Antennas, an isotropic radiator is a theoretical, lossless, omnidirectional (spherical) antenna. That is, it radiates uniformly in all directions. The power of a transmitter that is radiated from an isotropic antenna will have a uniform power density (power per unit area) in all directions. The power density at any distance from an isotropic antenna is simply the transmitter power divided by the surface area of a sphere ($4\pi R^2$) at that distance. The surface area of the sphere increases by the square of the radius, therefore the power density, P_D, (watts/square meter) decreases by the square of the radius.

$$\begin{matrix} \text{Power density from} \\ \text{an isotropic antenna} \end{matrix} = P_D = \frac{P_t}{4\pi R^2} \qquad \begin{matrix} \text{where : } P_t = \text{Transmitter Power} \\ R = \text{Range From Antenna (i.e., radius of sphere)} \end{matrix} \qquad \textbf{[1]}$$

P_t is either peak or average power depending on how P_D is to be specified.

Radars use directional antennas to channel most of the radiated power in a particular direction. The Gain (G) of an antenna is the ratio of power radiated in the desired direction as compared to the power radiated from an isotropic antenna, or:

$$G = \frac{\text{Maximum radiation intensity of actual antenna}}{\text{Radiation intensity of isotropic antenna with same power input}}$$

The power density at a distant point from a radar with an antenna gain of G_t is the power density from an isotropic antenna multiplied by the radar antenna gain.

Power density from radar, $$P_D = \frac{P_t G_t}{4\pi R^2} \qquad \textbf{[2]}$$

P_t is either peak or average power depending on how P_D is to be specified.

Another commonly used term is effective isotropic radiated power (EIRP), where **EIRP = P_t G_t.** ERP is also used but EIRP is preferred because it specifically defines the type of reference antenna as isotropic.

A receiving antenna captures a portion of this power determined by its effective capture Area (A_e). The received power available at the antenna terminals is the power density times the effective capture area (A_e) of the receiving antenna.

e.g., If the power density at a specified range is one microwatt per square meter and the antenna's effective capture area is one square meter then the power captured by the antenna is one microwatt.

For a given receiver antenna size the capture area is constant no matter how far it is from the transmitter, as illustrated in Figure 1. Also notice from Figure 1 that the received signal power decreases by 1/4 (6 dB) as the distance doubles. This is due to the R^2 term in the denominator of equation [2].

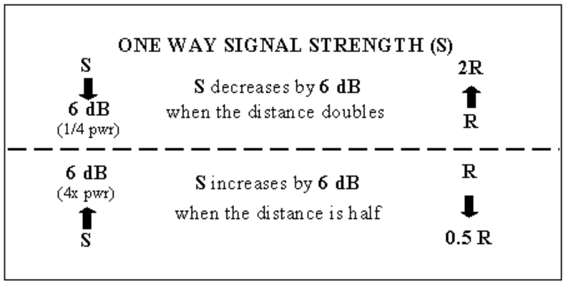
ONE WAY SIGNAL STRENGTH (S)
S
6 dB
(1/4 pwr)
S decreases by 6 dB
when the distance doubles
2R
R
6 dB
(4x pwr)
S
S increases by 6 dB
when the distance is half
R
0.5 R

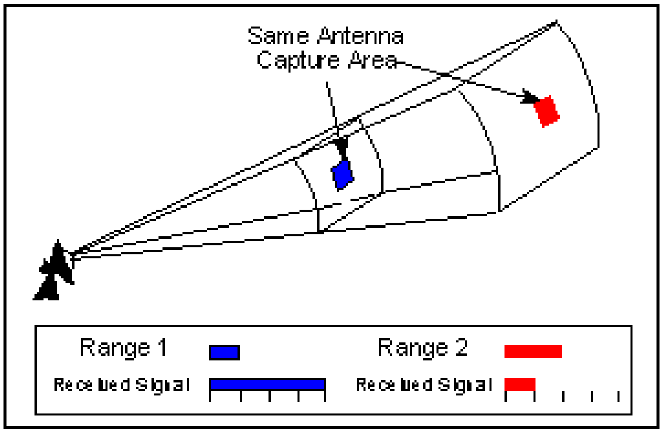
Same Antenna
Capture Area
Range 1
Received Signal
Range 2
Received Signal

Figure 1. Power Density vs. Range.

Sample Power Density Calculation - Far Field
(Refer to Section 3-5 for the definition of near field and far field)

Calculate the power density at 100 feet for 100 watts transmitted through an antenna with a gain of 10.

Given: P_t = 100 watts G_t = 10 (dimensionless ratio) R = 100 ft

This equation produces power density in watts per square range unit.

$$P_D = \frac{P_t G_t}{4\pi R^2} = \frac{(100 \text{ watts})(10)}{4\pi (100 \text{ ft})^2} = 0.0080 \text{ watts/ ft}^2$$

For safety (radiation hazard) and EMI calculations, power density is usually expressed in milliwatts per square cm. That's nothing more than converting the power and range to the proper units.

100 watts = 1 x 10^2 watts = 1 x 10^5 mW

100 feet = 30.4785 meters = 3047.85 cm.

$$P_D = \frac{P_t G_t}{4\pi R^2} = \frac{(10^5 \text{ mW}) \bullet (10)}{4\pi (3047.85 \text{cm})^2} = 0.0086 \text{ mW/ cm}^2$$

However, antenna gain is almost always given in dB, not as a ratio. It's then often easier to express EIRP in dBm.

$$P_t\,(dBm) = 10 \,\text{Log}\left[\frac{P_t \text{ watts}}{1 \text{ mW}}\right] = 10 \,\text{Log}\left[\frac{100}{.001}\right] = 50 \text{ dBm}$$

$$G_t\,(dB) = 10 \,\text{Log}\left[\frac{G_t}{1}\right] = 10 \,\text{Log}\,(10) = 10 \text{ dB}$$

EIRP (dBm) = P_t (dBm) + G_t (dB) = 50 + 10 = 60 dBm

To reduce calculations, the graph in Figure 2 can be used. It gives EIRP in dBm, range in feet and power density in mW/cm^2. Follow the scale A line for an EIRP of 60 dBm to the point where it intersects the 100 foot range scale. Read the power density directly from the A-scale x-axis as 0.0086 mW/cm^2 (confirming our earlier calculations).

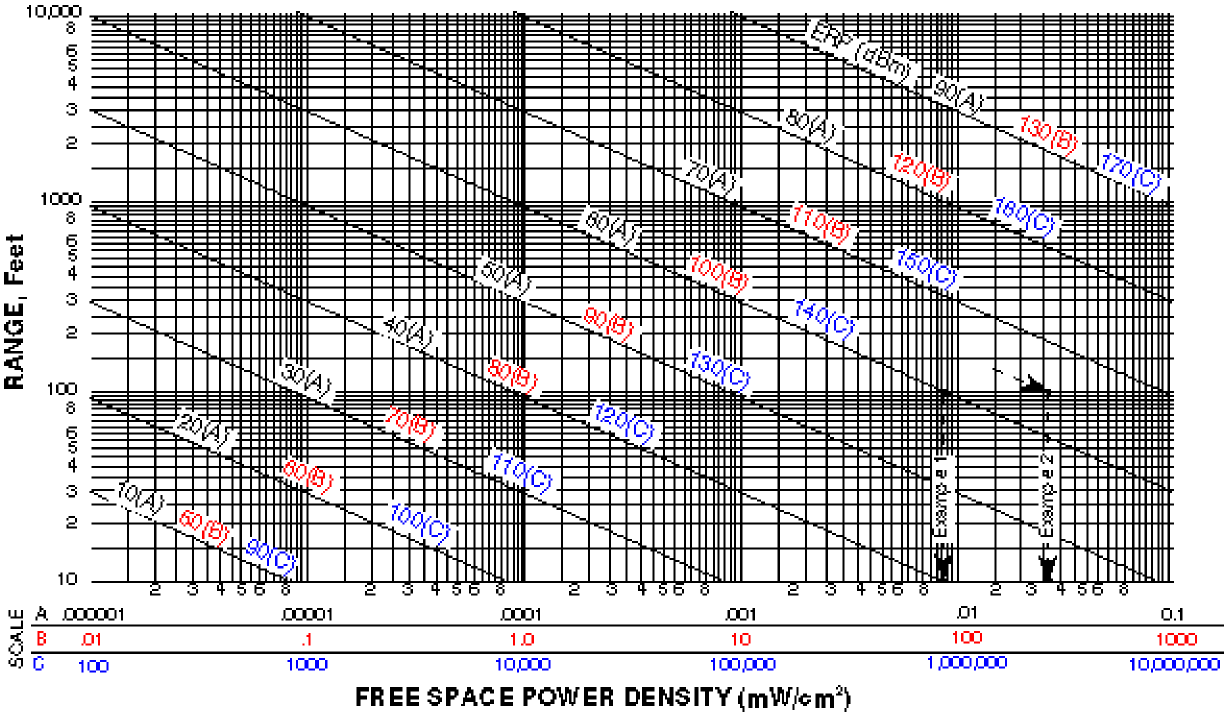

Figure 2. Power Density vs. Range and EIRP.

Example 2

When antenna gain and power (or EIRP) are given in dB and dBm, it is necessary to convert back to ratios in order to perform the calculation given in equation [2]. Use the same values as in example 1 except for antenna gain.

Suppose the antenna gain is given as 15 dB: **G_t (dB) = 10 Log (G_t)**

$$Therefore: \quad G_t = [10]^{\left(\frac{G_t(dB)}{10}\right)} = [10]^{\left(\frac{15}{10}\right)} = 31.6228$$

$$P_D = \frac{P_t G_t}{4\pi R^2} = \frac{(10^5\ mW)(31.6228)}{4\pi (3047.85)^2} = 0.0271\ mW/cm^2$$

Follow the 65 dBm (extrapolated) EIRP line and verify this result on the A-scale X-axis.

Example 3 - Sample Real Life Problem

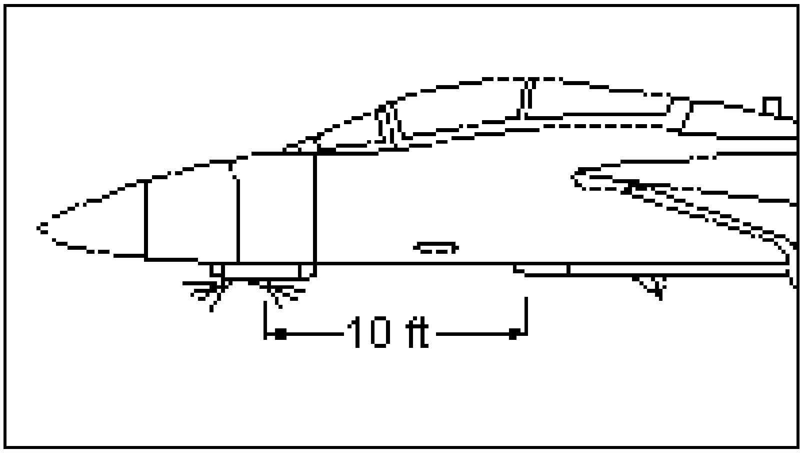

Assume we are trying to determine if a jammer will damage the circuitry of a missile carried onboard an aircraft and we cannot perform an actual measurement. Refer to the diagram at the right.

Given the following:

Jammer power: 500 W ($P_t = 500$)

Jammer line loss and antenna gain: 3 dB ($G_t = 2$)

Missile antenna diameter: 10 in

Missile antenna gain: Unknown

Missile limiter protection (maximum antenna power input): 20 dBm (100mW) average and peak.

The power density at the missile antenna caused by the jammer is computed as follows:

$$P_D = \frac{P_t G_t}{4\pi R^2} = \frac{500W\,(2)}{4\pi[(10ft)(.3048m/ft)]^2} = 8.56W/m^2$$

The maximum input power actually received by the missile is either:

$P_r = P_D A_e$ (if effective antenna area is known) or
$P_r = P_D G_m \lambda^2/4\pi$ (if missile antenna gain is known)

To cover the case where the missile antenna gain is not known, first assume an aperture efficiency of 0.7 for the missile antenna (typical). Then:

$P_r = P_D A \eta = 8.56\ W/m^2\ (\pi)[\ (10/2\ in)(.0254\ m/in)\ ]^2\ (0.7) = 0.3$ watts

Depending upon missile antenna efficiency, we can see that the power received will be about 3 times the maximum allowable and that either better limiter circuitry may be required in the missile or a new location is needed for the missile or jammer. Of course if the antenna efficiency is 0.23 or less, then the power will not damage the missile's receiver.

If the missile gain were known to be 25 dB, then a more accurate calculation could be performed. Using the given gain of the missile (25 dB= numeric gain of 316), and assuming operation at 10 GHz ($\lambda = .03$m)

$P_r = P_D G_m \lambda^2 / 4\pi = 8.56\ W/m^2\ (316)(.03)^2/ 4\pi = .19$ watts (still double the allowable tolerance)

ONE-WAY RADAR EQUATION / RF PROPAGATION

The one-way (transmitter to receiver) radar equation is derived in this section. This equation is most commonly used in RWR, communications, or ESM type of applications. The following is a summary of the important equations explored in this section:

ONE-WAY RADAR EQUATION

Peak Power at

Receiver Input, $P_r\,(or\ S) = P_D\,A_e = \frac{P_t G_t A_e}{4\pi R^2}$ and Antenna Gain, $G = \frac{4\pi A_e}{\lambda^2}$

or: Equivalent Area, $A_e = \frac{G\lambda^2}{4\pi}$

So the one-way radar equation is :

$$S\,(or\ P_r) = \frac{P_t G_t G_r \lambda^2}{(4\pi R)^2} = P_t G_t G_r \left[\frac{c^2}{(4\pi f R)^2}\right]^* \qquad (\ Note: \lambda = \frac{c}{f}\ )$$

* keep λ, c, and R in the same units

On reducing to log form this becomes:
$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r - 20\log fR + 20\log(c/4\pi)$

or in simplified terms:
$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r - \alpha_1$ (in dB)

Where: α_1 = one-way free space loss = $20\log(f_1 R) + K_1$ (in dB) and:
$K_1 = 20\log[(4\pi/c)$(Conversion factors if units if not in m/sec, m, and Hz)$]$

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain

Values of K_1 (in dB)

Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
NM	37.8	97.8
km	32.45	92.45
m	-27.55	32.45
yd	-28.33	31.67
ft	-37.87	22.13

Note: Losses due to antenna polarization and atmospheric absorption (Sections 3-2 & 5-1) are not included in any of these equations.

Recall from Section 4-2 that the power density at a distant point from a radar with an antenna gain of G_t is the power density from an isotropic antenna multiplied by the radar antenna gain.

Power density from radar, $P_D = \frac{P_t G_t}{4\pi R^2}$

If you could cover the entire spherical segment with your receiving antenna you would theoretically capture all of the transmitted energy. You can't do this because no antenna is large enough. (A two degree segment would be about a mile and three-quarters across at fifty miles from the transmitter.)

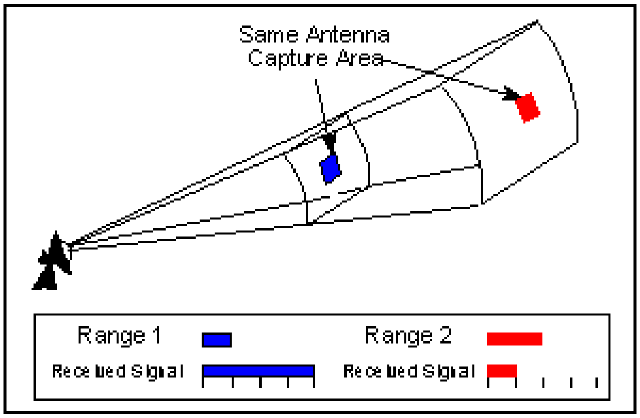

Figure 1. Power Density vs. Range.

A receiving antenna captures a portion of this power determined by its effective capture Area (A_e). The received power available at the antenna terminals is the power density times the effective capture area (A_e) of the receiving antenna.

For a given receiver antenna size the capture area is constant no matter how far it is from the transmitter, as illustrated in Figure 1. This concept is shown in the following equation:
Peak Power at Receiver input,

$$P_R \text{ (or S)} = P_D A_e = \frac{P_t G_t A_e}{4\pi R^2} \quad \text{which is known as the one-way (beacon) equation} \qquad [2]$$

In order to maximize energy transfer between an antenna and transmitter or receiver, the antenna size should correlate with frequency. For reasonable antenna efficiency, the size of an antenna will be greater than $\lambda/4$. Control of beamwidth shape may become a problem when the size of the active element exceeds several wavelengths.

The relation between an antenna's effective capture area (A_e) or effective aperture and its Gain (G) is:

$$Antenna\ Gain,\ G = \frac{4\pi A_e}{\lambda^2} \qquad [3]$$

$$or:\ Equivalent\ Area,\ A_e = \frac{G\lambda^2}{4\pi} \qquad [4]$$

Since the effective aperture is in units of length squared, from equation [3], it is seen that gain is proportional to the effective aperture normalized by the wavelength. This physically means that to maintain the same gain when doubling the frequency, the area is reduced by 1/4. This concept is illustrated in Figure 2.

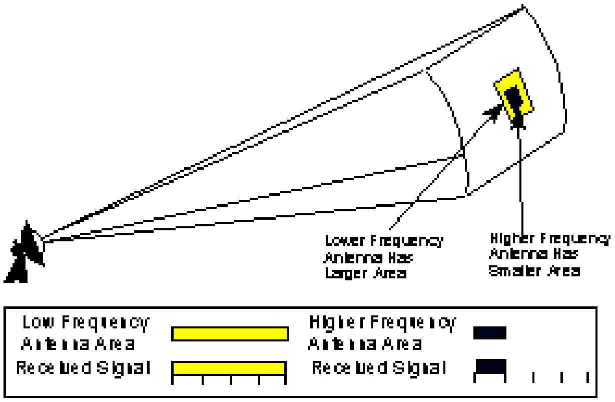
Lower Frequency Antenna Has Larger Area
Higher Frequency Antenna Has Smaller Area
Low Frequency Antenna Area
Received Signal
Higher Frequency Antenna Area
Received Signal

Figure 2. Capture Area vs. Frequency.

If equation [4] is substituted into equation [2], the following relationship results:

$$Peak\ Power\ at\ Receiver\ Input = S\ (or\ P_R) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 R^2} = \frac{P_t G_t G_r \lambda^2}{(4\pi R)^2} \qquad [5]$$

This is the signal calculated one-way from a transmitter to a receiver. For instance, a radar application might be to determine the signal received by a RWR, ESM, or an ELINT receiver. It is a general purpose equation and could be applied to almost any line-of-sight transmitter to receiver situation if the RF is higher than 100 MHZ.

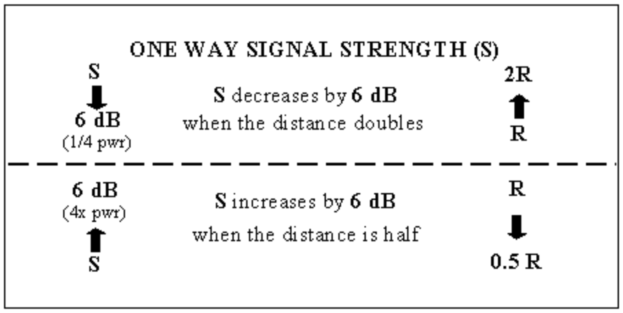
ONE WAY SIGNAL STRENGTH (S)
S
6 dB
(1/4 pwr)
S decreases by 6 dB when the distance doubles
2R
R
6 dB
(4x pwr)
S
S increases by 6 dB when the distance is half
R
0.5 R

Values of K_1 (dB)

Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
NM	37.8	97.8
km	32.45	92.45
m	-27.55	32.45
yd	-28.33	31.67
ft	-37.87	22.13

The free space travel of radio waves can, of course, be blocked, reflected, or distorted by objects in their path such as buildings, flocks of birds, chaff, and the earth itself.

As illustrated in Figure 1, as the distance is doubled the received signal power decreases by 1/4 (6 dB). This is due to the R^2 term in equation [5].

To illustrate this, blow up a round balloon and draw a square on the side of it. If you release air so that the diameter or radius is decreased by 1/2, the square shrinks to 1/4 the size. If you further blow up the balloon, so the diameter or radius is doubled, the square has quadrupled in area.

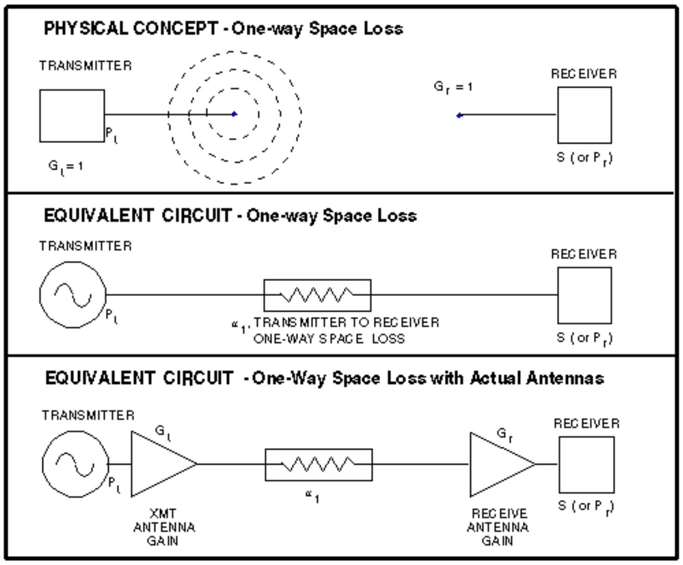
PHYSICAL CONCEPT - One-way Space Loss
TRANSMITTER
P_t
$G_t = 1$
$G_r = 1$
RECEIVER
S (or P_r)
EQUIVALENT CIRCUIT - One-way Space Loss
TRANSMITTER
P_t
α_1, TRANSMITTER TO RECEIVER ONE-WAY SPACE LOSS
RECEIVER
S (or P_r)
EQUIVALENT CIRCUIT - One-Way Space Loss with Actual Antennas
TRANSMITTER
G_t
P_t
XMT ANTENNA GAIN
α_1
G_r
RECEIVE ANTENNA GAIN
RECEIVER
S (or P_r)

Figure 3. Concept of One-Way Space Loss.

The one-way free space loss factor (α_1), (sometimes called the path loss factor) is given by the term $(4\pi R^2)(4\pi/\lambda^2)$ or $(4\pi R/\lambda)^2$. As shown in Figure 3, the loss is due to the ratio of two factors (1) the effective radiated area of the transmit antenna, which is the surface area of a sphere ($4\pi R^2$) at that distance (R), and (2) the effective capture area (A_e) of the receive antenna which has a gain of one. If a receiving antenna could capture the whole surface area of the sphere, there would be no spreading loss, but a practical antenna will capture only a small part of the spherical radiation. Space loss is calculated using isotropic antennas for both transmit and receive, so α_1 is independent of the actual antenna. Using Gr = 1 in equation [11] in Section 3-1, $A_e = \lambda^2/4\pi$. Since this term is in the denominator of α_1, the higher the frequency (lower λ) the more the space loss. Since G_t and G_r are part of the one-way radar equation, S (or P_r) is adjusted according to actual antennas as shown in the last portion of Figure 3. The value of the received signal (S) is:

$$S\,(or\ P_R) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 R^2} = P_t G_t G_r \left[\frac{\lambda^2}{(4\pi R)^2}\right] \qquad \textbf{[6]}$$

To convert this equation to dB form, it is rewritten as:

$$10\,\text{Log}\,(S\text{ or }P_r) = 10\,\text{Log}\,(P_t G_t G_r) + 20\,\text{Log}\left[\frac{\lambda}{4\pi f R}\right] \qquad \textbf{[7]}$$

Since $\lambda = c/f$, equation [7] can be rewritten as:

$$10\,\text{Log}\,(S\text{ or }P_r) = 10\,\text{Log}(P_t G_t G_r) - \alpha_1 \qquad \textbf{[8]}$$

Where the one-way free space loss, α_1, is defined as: $\alpha_1 = 20\,\text{Log}\left[\frac{4\pi f R}{c}\right]^*$ **[9]**

The signal received equation in dB form is: $10\log(P_r \text{ or } S) = 10\log P_t + 10\log G_t + 10\log G_r - \alpha_1$ **[10]**

The one-way free space loss, α_1, can be given in terms of a variable and constant term as follows:

$$\alpha_1 = 20\,\mathrm{Log}\left[\frac{4\pi f\,R}{c}\right]^* = 20\,\mathrm{Log}\, f_1 R + K_1 \quad (in\ dB) \qquad \textbf{[11]}$$

The value of f_1 can be either in MHz or GHz as shown with commonly used units of R in the adjoining table.

$$where\ K_1 = 20\,\mathrm{Log}\left[\frac{4\pi}{c}\bullet (Conversion\ units\ if\ not\ in\ m/\mathrm{sec}, m, and\ Hz)\right]$$

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain.

A value for the one-way free space loss (α_1) can be obtained from:

(a) The One-way Free Space Loss graph (Figure 4). Added accuracy can be obtained using the Frequency Extrapolation graph (Figure 5)

(b) The space loss nomograph (Figure 6 or 7)

(c) The formula for α_1, equation [11].

FOR EXAMPLE:

Find the value of the one-way free space loss, α_1, for an RF of 7.5 GHz at 100 NM.

(a) From Figure 4, find 100 NM on the X-axis and estimate where 7.5 GHz is located between the 1 and 10 GHz lines (note dot). Read α_1 as 155 dB. An alternate way would be to read the α_1 at 1 GHz (138 dB) and add the frequency extrapolation value (17.5 dB for 7.5:1, dot on Figure 5) to obtain the same 155 dB value.

(b) From the nomogram (Figure 6), the value of α_1 can be read as 155 dB (Note the dashed line).

(c) From the equation 11, the precise value of α_1 is 155.3 dB.

Remember, α_1 is a free space value. If there is atmospheric attenuation because of absorption of RF due to certain molecules in the atmosphere or weather conditions etc., the atmospheric attenuation is in addition to the space loss (refer to Section 5-1).

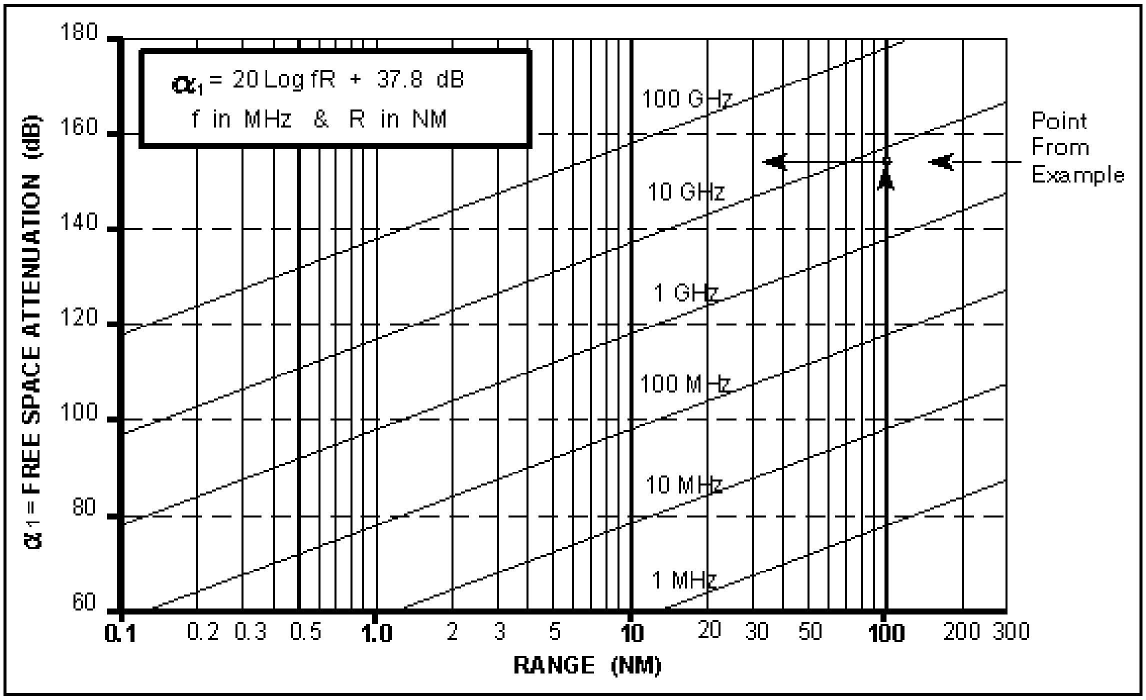

Figure 4. One-Way Free Space Loss.

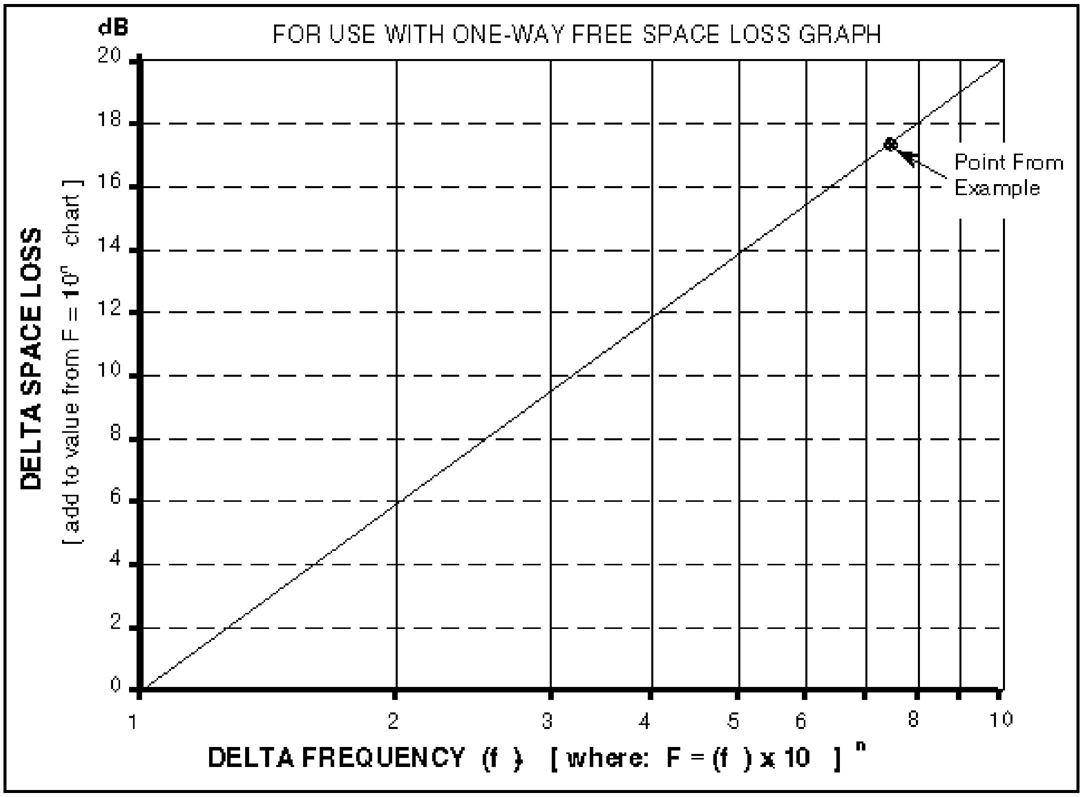

Figure 5. Frequency Extrapolation.

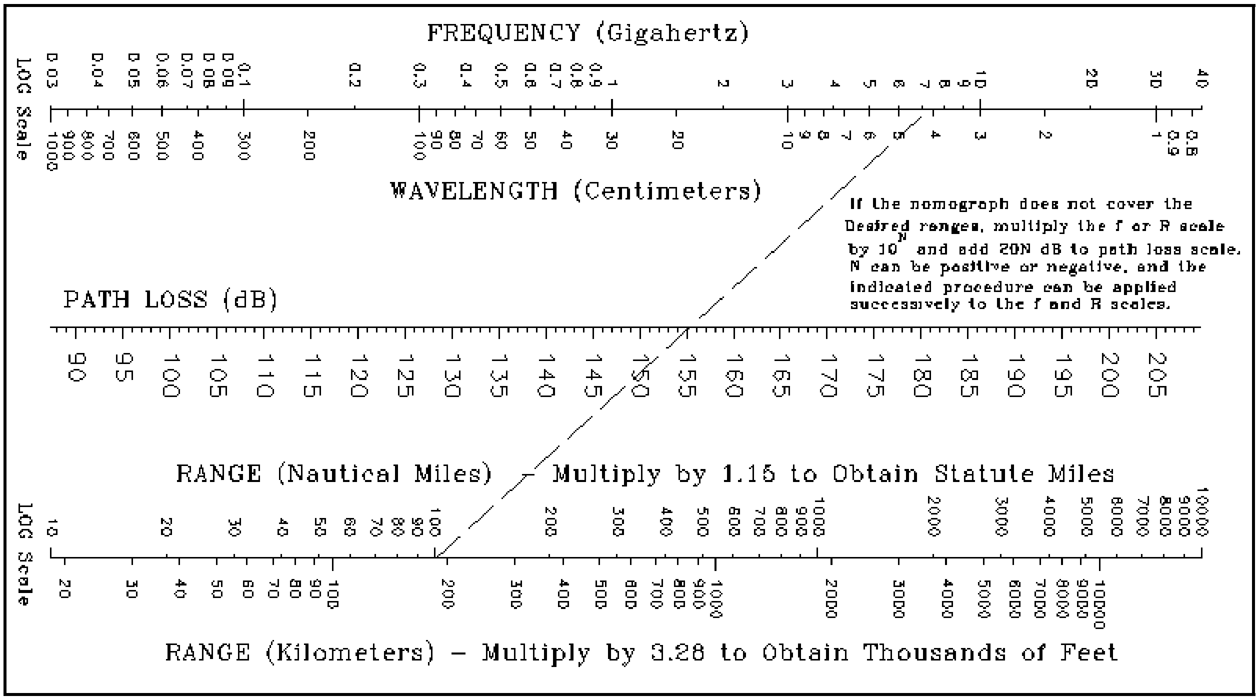

If the nomograph does not cover the Desired ranges, multiply the f or R scale by 10^N and add 20N dB to path loss scale. N can be positive or negative, and the indicated procedure can be applied successively to the f and R scales.

Figure 6. One-Way Space Loss Nomograph for Distances Greater Than 10 Nautical Miles.

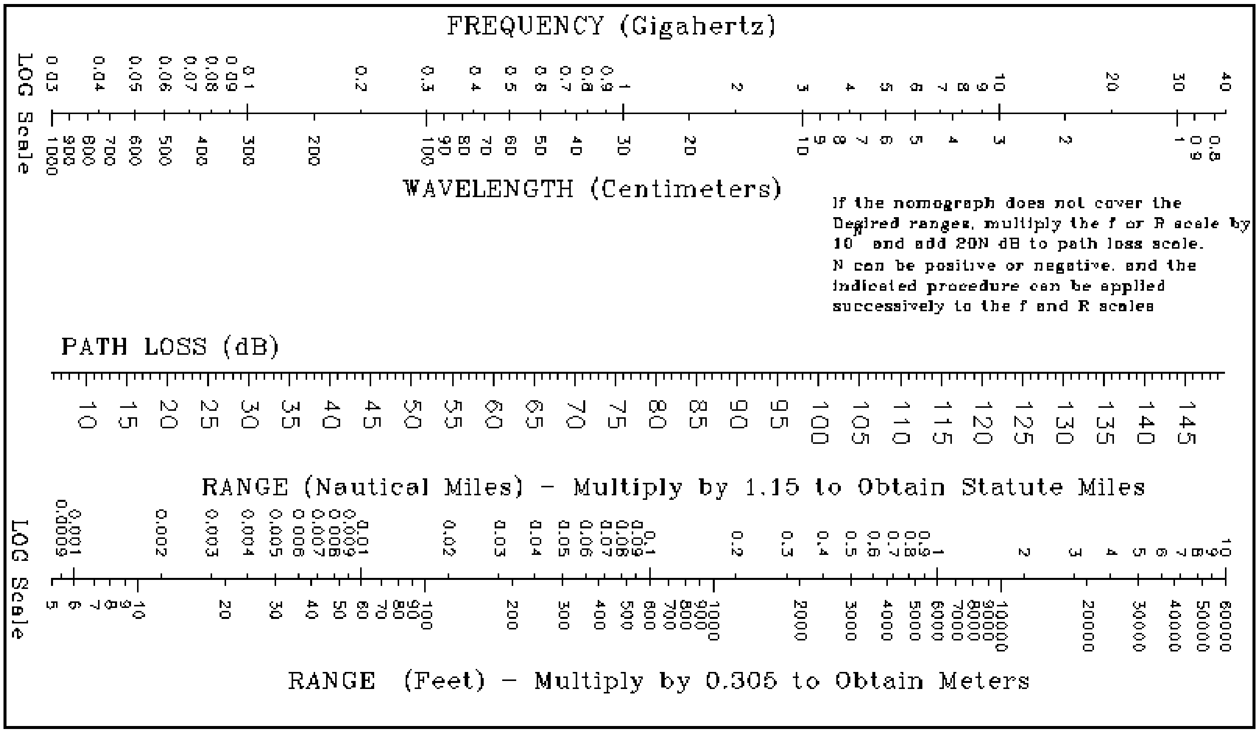

If the nomograph does not cover the Desired ranges, multiply the f or R scale by 10^N and add 20N dB to path loss scale. N can be positive or negative, and the indicated procedure can be applied successively to the f and R scales

Figure 7. One-Way Space Loss Nomograph for Distances Less Than 10 Nautical Miles.

Figure 8 is the visualization of the losses occurring in one-way radar equation. Note: To avoid having to include additional terms, always combine any transmission line loss with antenna gain. Losses due to antenna polarization and atmospheric absorption also need to be included.

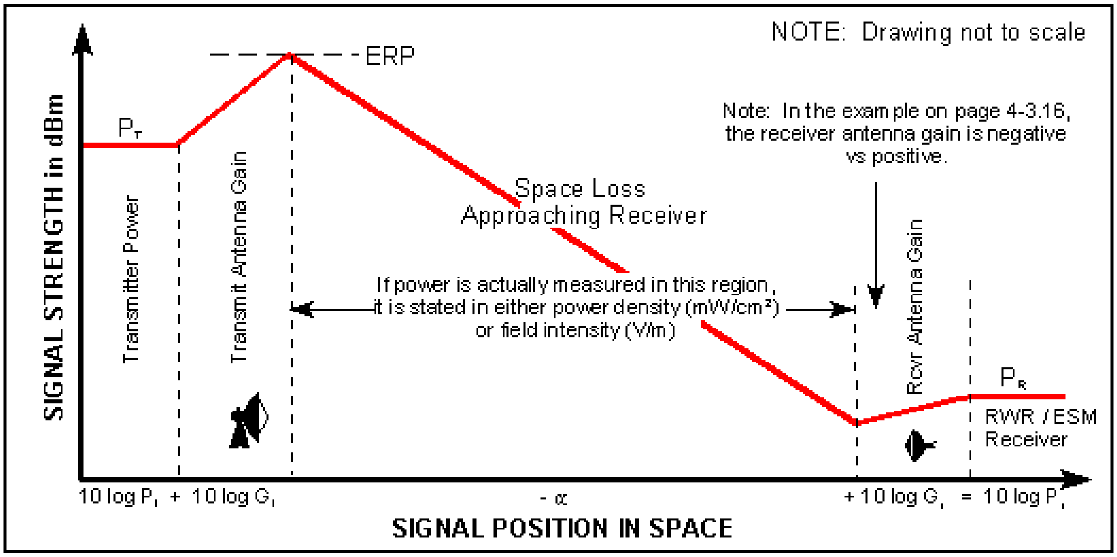
NOTE: Drawing not to scale
ERP
SIGNAL STRENGTH in dBm
P_T
Transmitter Power
Transmit Antenna Gain
Space Loss
Approaching Receiver
Note: In the example on page 4-3.16, the receiver antenna gain is negative vs positive.
If power is actually measured in this region, it is stated in either power density (mW/cm²) or field intensity (V/m)
Rcvr Antenna Gain
P_R
RWR / ESM Receiver
10 log P_t + 10 log G_t
- α
+ 10 log G_r = 10 log P_r
SIGNAL POSITION IN SPACE

Figure 8. Visualization of One-Way Radar Equation.

RWR/ESM RANGE EQUATION (One-Way)

The one-way radar (signal strength) equation [5] is rearranged to calculate the maximum range R_{max} of RWR/ESM receivers. It occurs when the received radar signal just equals S_{min} as follows:

$$R_{max} \cong \left[\frac{P_t G_t G_r \lambda^2}{(4\pi)^2 S_{min}}\right]^{\frac{1}{2}} \; or \; \left[\frac{P_t G_t G_r c^2}{(4\pi f)^2 S_{min}}\right]^{\frac{1}{2}} \; or \; \left[\frac{P_t G_t A_e}{4\pi S_{min}}\right]^{\frac{1}{2}} \qquad \textbf{[12]}$$

In log form:

$$20\log R_{max} = 10\log P_t + 10\log G_t - 10\log S_{min} - 20\log f + 20\log(c/4\pi) \qquad \textbf{[13]}$$

and since $K_1 = 20\log\{4\pi/c$ times conversion units if not in m/sec, m, and Hz$\}$
(Refer to Section 4-3 for values of K_1).

$$10\log R_{max} = \tfrac{1}{2}[\, 10\log P_t + 10\log G_t - 10\log S_{min} - 20\log f - K_1] \quad \text{(keep } P_t \text{ and } S_{min} \text{ in same units)} \qquad \textbf{[14]}$$

If you want to convert back from dB, then $R_{max} \cong \frac{10^{M\,dB}}{20}$, where M dB is the resulting number in the brackets of equation 14.

From Section 5-2, Receiver Sensitivity / Noise, S_{min} is related to the noise factor S:

$$S_{min} = (S/N)_{min}\,(NF)KT_oB \qquad \textbf{[15]}$$

The one-way RWR/ESM range equation becomes:

$$R_{max} \cong \left[\frac{P_t G_t G_r \lambda^2}{(4\pi)^2 (S/N)_{min} (NF) K T_o B}\right]^{\frac{1}{2}} \text{ or } \left[\frac{P_t G_t G_r c^2}{(4\pi f)^2 (S/N)_{min} (NF) K T_o B}\right]^{\frac{1}{2}} \text{ or } \left[\frac{P_t G_t A_e}{4\pi (S/N)_{min} (NF) K T_o B}\right]^{\frac{1}{2}} \qquad \textbf{[16]}$$

RWR/ESM RANGE INCREASE AS A RESULT OF A SENSITIVITY INCREASE

As shown in equation [12] $S_{min}^{-1} \propto R_{max}^2$. Therefore, $-10 \log S_{min} \propto 20 \log R_{max}$ and the following table results:

% Range Increase: Range + (% Range Increase) x Range = New Range
i.e., for a 6 dB sensitivity increase, 500 miles +100% x 500 miles = 1,000 miles
Range Multiplier: Range x Range Multiplier = New Range
i.e., for a 6 dB sensitivity increase 500 miles x 2 = 1,000 miles

dB Sensitivity Increase	% Range Increase	Range Multiplier	dB Sensitivity Increase	% Range Increase	Range Multiplier
+ 0.5	6	1.06	10	216	3.16
1.0	12	1.12	11	255	3.55
1.5	19	1.19	12	298	3.98
2	26	1.26	13	347	4.47
3	41	1.41	14	401	5.01
4	58	1.58	15	462	5.62
5	78	1.78	16	531	6.31
6	100	2.0	17	608	7.08
7	124	2.24	18	694	7.94
8	151	2.51	19	791	8.91
9	182	2.82	20	900	10.0

RWR/ESM RANGE DECREASE AS A RESULT OF A SENSITIVITY DECREASE

As shown in equation [12] $S_{min}^{-1} \propto R_{max}^2$. Therefore, $-10 \log S_{min} \propto 20 \log R_{max}$ and the following table results:

% Range Decrease: Range - (% Range decrease) x Range = New Range
i.e., for a 6 dB sensitivity decrease, 500 miles - 50% x 500 miles = 250 miles
Range Multiplier: Range x Range Multiplier = New Range
i.e., for a 6 dB sensitivity decrease 500 miles x .5 = 250 miles

dB Sensitivity Decrease	% Range Decrease	Range Multiplier	dB Sensitivity Decrease	% Range Decrease	Range Multiplier
- 0.5	6	0.94	-10	68	0.32
- 1.0	11	0.89	- 11	72	0.28
- 1.5	16	0.84	- 12	75	0.25
- 2	21	0.79	- 13	78	0.22
- 3	29	0.71	- 14	80	0.20
- 4	37	0.63	- 15	82	0.18
- 5	44	0.56	- 16	84	0.16
- 6	50	0.50	- 17	86	0.14
- 7	56	0.44	- 18	87	0.13
- 8	60	0.4	- 19	89	0.11
- 9	65	0.35	- 20	90	0.10

Example of One-Way Signal Strength: A 5 (or 7) GHz radar has a 70 dBm signal fed through a 5 dB loss transmission line to an antenna that has 45 dB gain. An aircraft that is flying 31 km from the radar has an aft EW antenna with -1 dB gain and a 5 dB line loss to the EW receiver (assume all antenna polarizations are the same).

Note: The respective transmission line losses will be combined with antenna gains, i.e.:

$$-5 + 45 = 40 \text{ dB}, \; -5 - 1 = -6 \text{ dB}, \; -10 + 5 = -5 \text{ dB}.$$

(1) What is the power level at the input of the EW receiver?

Answer (1): P_r at the input to the EW receiver = Transmitter power - xmt cable loss + xmt antenna gain - space loss + rcvr antenna gain - rcvr cable loss.

Space loss (from Section 4-3) @ 5 GHz = $20 \log f R + K_1 = 20 \log (5x31) + 92.44 = 136.25$ dB. Therefore:

$$P_r = 70 + 40 - 136.25 - 6 = -32.25 \text{ dBm @ 5 GHz } (P_r = -35.17 \text{ dBm @ 7 GHz since } \alpha_1 = 139.17 \text{ dB})$$

(2) If the received signal is fed to a jammer with a gain of 60 dB, feeding a 10 dB loss transmission line which is connected to an antenna with 5 dB gain, what is the power level from the jammer at the input to the receiver of the 5 (or 7) GHz radar?

Answer (2): P_r at the input to the radar receiver = Power at the input to the EW receiver+ Jammer gain - jammer cable loss + jammer antenna gain - space loss + radar rcvr antenna gain - radar rcvr cable loss. Therefore:

$$P_r = -32.25 + 60 - 5 - 136.25 + 40 = -73.5 \text{ dBm @ 5 GHz. } (P_r = -79.34 \text{ dBm @ 7 GHz since } \alpha_1 = 139.17 \text{ dB and } P_t = -35.17 \text{ dBm}).$$

This problem continues in Sections 4-4, 4-7, and 4-10.

This page intentionally left blank.

TWO-WAY RADAR EQUATION (MONOSTATIC)

In this section the radar equation is derived from the one-way equation (transmitter to receiver) which is then extended to the two-way radar equation. The following is a summary of the important equations to be derived here:

TWO-WAY RADAR EQUATION (MONOSTATIC)

Peak power at the radar receiver input is:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4} = P_t G_t G_r \left[\frac{\sigma c^2}{(4\pi)^3 f^2 R^4}\right]^*$$

Note : $\lambda = c/f$ *and* $\sigma = RCS$

* *keep* λ *or* $c, \sigma,$ *and* R *in the same units*

On reducing the above equation to log form we have:

$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 20\log f - 40\log R - 30\log 4\pi + 20\log c$

or in simplified terms: $10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1$ (in dB)

Note: Losses due to antenna polarization and atmospheric absorption (Sections 3-2 and 5-1) are not included in these equations.

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)

K_2 Values (dB)	RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

One-way free space loss, $\alpha_1 = 20\log(f_1 R) + K_1$ (in dB)

K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	Km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

Figure 1 illustrates the physical concept and equivalent circuit for a target being illuminated by a monostatic radar (transmitter and receiver co-located). Note the similarity of Figure 1 to Figure 3 in Section 4-3. Transmitted power, transmitting and receiving antenna gains, and the one-way free space loss are the same as those described in Section 4-3. The physical arrangement of the elements is different, of course, but otherwise the only difference is the addition of the equivalent gain of the target RCS factor.

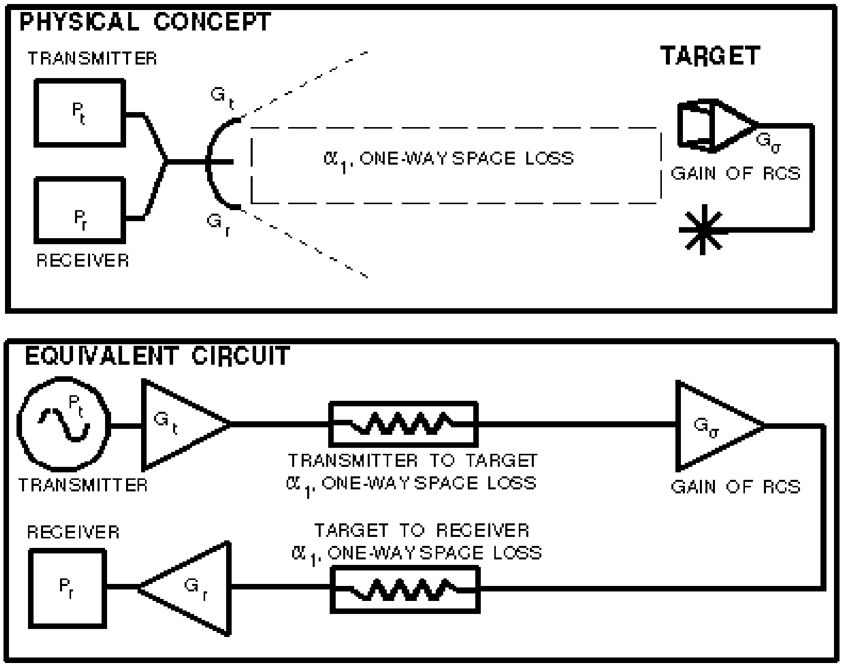
PHYSICAL CONCEPT
TRANSMITTER
P_t
P_r
RECEIVER
G_t
G_r
α_1, ONE-WAY SPACE LOSS
TARGET
G_σ
GAIN OF RCS
EQUIVALENT CIRCUIT
P_t
TRANSMITTER
G_t
TRANSMITTER TO TARGET
α_1, ONE-WAY SPACE LOSS
G_σ
GAIN OF RCS
RECEIVER
P_r
G_r
TARGET TO RECEIVER
α_1, ONE-WAY SPACE LOSS

Figure 1. The Two-Way Monostatic Radar Equation Visualized.

From Section 4-3, One-Way Radar Equation / RF Propagation, the power in the receiver is:

$$\begin{array}{c}\text{Received Signal} \\ \text{at Target}\end{array} = \frac{P_t G_t G_r \lambda^2}{(4\pi R)^2} \qquad \textbf{[1]}$$

From equation [3] in Section 4-3: $$\text{Antenna Gain}, G = \frac{4\pi A_e}{\lambda^2} \qquad \textbf{[2]}$$

Similar to a receiving antenna, a radar target also intercepts a portion of the power, but reflects (reradiates) it in the direction of the radar. The amount of power reflected toward the radar is determined by the Radar Cross Section (RCS) of the target. RCS is a characteristic of the target that represents its size as seen by the radar and has the dimensions of area (σ) as shown in Section 4-11. RCS area is not the same as physical area. But, for a radar target, the power reflected in the radar's direction is equivalent to re-radiation of the power captured by an antenna of area σ (the RCS). Therefore, the effective capture area (A_e) of the receiving antenna is replaced by the RCS (σ).

$$G_r = \frac{4\pi\sigma}{\lambda^2} \qquad \textbf{[3]}$$

so we now have:

$$\begin{array}{c}\text{Reflected Signal} \\ \text{from target}\end{array} = \frac{P_t G_t \lambda^2 4\pi\sigma}{(4\pi R)^2 \lambda^2} \qquad \textbf{[4]}$$

The equation for the power reflected in the radar's direction is the same as equation [1] except that P_t G_t, which was the original transmitted power, is replaced with the reflected signal power from the target, from equation [4]. This gives:

$$\begin{array}{c}\text{Reflected Signal Received Back} \\ \text{at Input to Radar Receiver}\end{array} = \frac{P_t G_t \lambda^2 4\pi\sigma}{(4\pi R)^2 \lambda^2} \; x \; \frac{G_r \lambda^2}{(4\pi R)^2} \qquad \textbf{[5]}$$

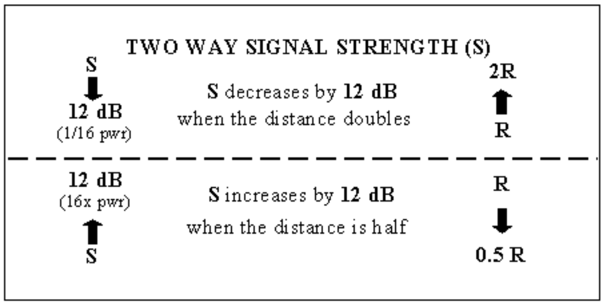

If like terms are cancelled, the two-way radar equation results. The peak power at the radar receiver input is:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4} = P_t G_t G_r \left[\frac{\sigma c^2}{(4\pi)^3 f^2 R^4} \right]^* \qquad \textbf{[6]}$$

* Note: $\lambda = c/f$ and σ = RCS. Keep λ or c, σ, and R in the same units.

On reducing equation [6] to log form we have:

$$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 20\log f - 40\log R - 30\log 4\pi + 20\log c \qquad \textbf{[7]}$$

Target Gain Factor

If Equation [5] terms are rearranged instead of cancelled, a recognizable form results:

$$S\,(or\ P_r) = (P_t G_t G_r) \bullet \left[\frac{\lambda^2}{(4\pi R)^2}\right] \bullet \left[\frac{4\pi\sigma}{\lambda^2}\right] \bullet \left[\frac{\lambda^2}{(4\pi R)^2}\right] \qquad \textbf{[8]}$$

In log form:

$$10\log[S\,(or\ P_r)] = 10\log P_t + 10\log G_t + 10\log G_r + 20\log\left[\frac{\lambda}{4\pi R}\right] + 10\log\left[\frac{4\pi\sigma}{\lambda^2}\right] + 20\log\left[\frac{\lambda}{4\pi R}\right] \qquad \textbf{[9]}$$

The fourth and sixth terms can each be recognized as -α, where α is the one-way free space loss factor defined in Section 4-3. The fifth term containing RCS (σ) is the only new factor, and it is the “Target Gain Factor.”

In simplified terms the equation becomes:

$$10\log\,[S\ (or\ P_r)] = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1 \qquad \text{(in dB)} \qquad \textbf{[10]}$$

Where α_1 and G_σ are as follows:

From Section 4-3, equation [11], the space loss in dB is given by:

$$\alpha_1 = 20\log\left[\frac{4\pi f R}{c}\right]^{*} = 20\log f_1 R + K_1 \quad where\ K_1 = 20\log\left[\frac{4\pi}{c} \bullet (Conversion\ units\ if\ not\ in\ m/s\text{ec}, m, and\ Hz)\right] \qquad \textbf{[11]}$$

* Keep c and R in the same units. The table of values for K_1 is again presented here for completeness. The constant, K_1, in the table includes a range and frequency unit conversion factor.

While it’s understood that RCS is the antenna aperture area equivalent to an isotropically radiated target return signal, the target gain factor represents a gain, as shown in the equivalent circuit of Figure 1. The Target Gain Factor expressed in dB is G_σ as shown in equation [12].

One-way free space loss, α_1 = 20log (f_1R) + K_1 (in dB)

K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	Km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

$$G_\sigma = 10\log\left[\frac{4\pi\sigma}{\lambda^2}\right] = 10\log\left[\frac{4\pi\sigma f^2}{c^2}\right] = 10\log\sigma + 20\log f_1 + K_2 \quad (in\ dB) \qquad \textbf{[12]}$$

$$where: K_2 = 10\log\left[\frac{4\pi}{c^2} \bullet \left(\begin{matrix} Frequency\ and\ RCS \\ conversions\ as\ required \end{matrix} \quad \frac{(Hz\ to\ MHz\ or\ GHz)^2}{(meters\ to\ feet)^2}\right)\right]$$

The “Target Gain Factor” (G_σ) is a composite of RCS, frequency, and dimension conversion factors and is called by various names: “Gain of RCS”, “Equivalent Gain of RCS”, “Gain of Target Cross Section”, and in dB form “Gain-sub-Sigma”.

If frequency is given in MHz and RCS (σ) is in m^2, the formula for G_σ is:

$$G_\sigma = 10\log\sigma + 20\log f_1 + 10\log\left[4\pi \bullet \left(\frac{\sec}{3\,x\,10^8\,m}\right)^2 \bullet m^2 \bullet \left(\frac{1x\,10^6}{\sec}\right)^2\right] \qquad \textbf{[13]}$$

or:

$$G_\sigma = 10\log\sigma + 20\log f_1 - 38.54 \quad (in\ dB) \qquad \textbf{[14]}$$

Target gain factor, G_σ = 10log σ + 20log f_1 + K_2 (in dB)			
K_2 Values			
(dB)	RCS (σ)	f_1 in MHz	f_1 in GHz
(units)	K_2 =	K_2 =	
m^2	-38.54	21.46	
ft^2	-48.86	11.14	

For this example, the constant K_2 is -38.54 dB. K_2 values for various area and frequency and frequency units are summarized in the adjoining table.

In the two-way radar equation, the one-way free space loss factor (α_1) is used twice, once for the radar transmitter to target path and once for the target to radar receiver path. The radar illustrated in Figure 1 is monostatic so the two path losses are the same and the values of the two α_1’s are the same.

If the transmission loss in Figure 1 from P_t to G_t equals the loss from G_r to P_r, and $G_r = G_t$, then equation [10] can be written as:

$$10\log\,[S \text{ or } P_r] = 10\log P_t + 20\log G_{tr} - 2\alpha_1 + G_\sigma \quad \text{(in dB)} \qquad \textbf{[15]}$$

The space loss factor (α_1) and the target gain factor (G_σ) include all the necessary unit conversions so that they can be used directly with the most common units. Because the factors are given in dB form, they are more convenient to use and allow calculation without a calculator when the factors are read from a chart or nomograph.

Most radars are monostatic. That is, the radar transmitting and receiving antennas are literally the same antenna. There are some radars that are considered “monostatic” but have separate transmitting and receiving antennas that are co-located. In that case, equation [10] could require two different antenna gain factors as originally derived:

$$10\log\,[S \text{ or } P_r] = 10\log P_t + 10\log G_t + 10\log G_r - 2\alpha_1 + G_\sigma \quad \text{(in dB)} \qquad \textbf{[16]}$$

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain.

Figure 2 is the visualization of the path losses occurring with the two-way radar equation. **Note:** to avoid having to include additional terms, always combine any transmission line loss with antenna gain. Losses due to antenna polarization and atmospheric absorption also need to be included.

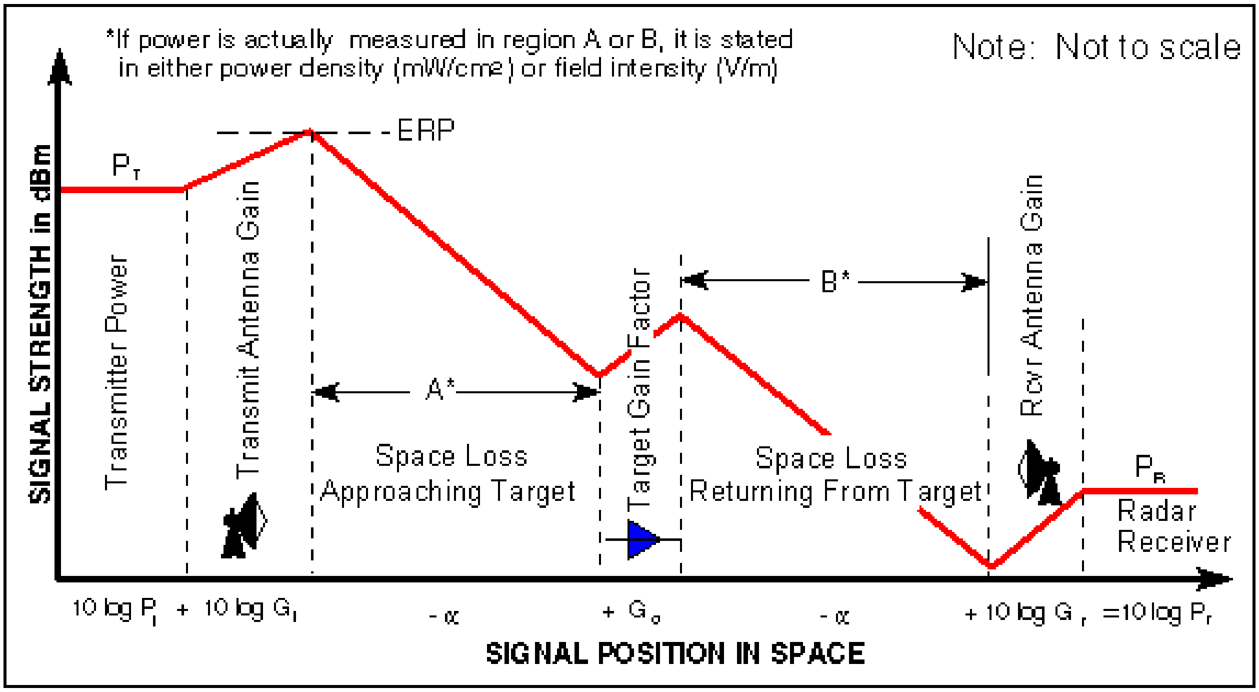
*If power is actually measured in region A or B, it is stated in either power density (mW/cm²) or field intensity (V/m)
Note: Not to scale
SIGNAL STRENGTH in dBm
ERP
P_T
Transmitter Power
Transmit Antenna Gain
A*
Space Loss Approaching Target
Target Gain Factor
B*
Space Loss Returning From Target
Rcvr Antenna Gain
P_R
Radar Receiver
10 log P_t
+ 10 log G_t
- α
+ G_σ
- α
+ 10 log G_r
=10 log P_r
SIGNAL POSITION IN SPACE

Figure 2. Visualization of Two-Way Radar Equation.

RADAR RANGE EQUATION (Two-Way Equation)

The Radar Equation is often called the "Radar Range Equation." The Radar Range Equation is simply the Radar Equation rewritten to solve for maximum Range. The maximum radar range (R_{max}) is the distance beyond which the target can no longer be detected and correctly processed. It occurs when the received echo signal just equals S_{min}.

The Radar Range Equation is then:

$$R_{\max} \cong \left[\frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 S_{\min}}\right]^{\frac{1}{4}} \text{ or } \left[\frac{P_t G_t G_r c^2 \sigma]}{(4\pi)^3 f^2 S_{\min}}\right]^{\frac{1}{4}} \text{ or } \left[\frac{P_t G_t A_e \sigma}{(4\pi)^2 S_{\min}}\right]^{\frac{1}{4}} \qquad \textbf{[17]}$$

The first equation, of the three above, is given in Log form by:

$$40\log R_{max} \cong 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 10\log S_{min} - 20\log f - 30\log 4\pi + 20\log c \qquad \textbf{[18]}$$

As shown previously, Since $K_1 = 20\log [(4\pi/c)$ times conversion units if not in m/sec, m, and Hz], we have:

$$10\log R_{\max} \cong \tfrac{1}{4}\left[10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 10\log S_{\min} - 20\log f_1 - K_1 - 10.99\text{ dB}\right] \qquad \textbf{[19]}$$

If you want to convert back from dB, then

$$R_{\max} \cong 10^{\frac{M\ dB}{40}}$$

Where M dB is the resulting number within the brackets of equation 19.

One-way free space loss, α_1 = 20log (f_1R) + K_1 (in dB)

K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	Km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

From Section 5-2, Receiver Sensitivity / Noise, $S_{\min}$ is related to the noise factors by:

$$S_{\min} = (S/N)_{\min}(NF)kT_0B \qquad \textbf{[20]}$$

The Radar Range Equation for a tracking radar (target continuously in the antenna beam) becomes:

$$R_{\max} \cong \left[\frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 (S/N)_{\min}(NF)kT_0B}\right]^{\frac{1}{4}} \text{ or } \left[\frac{P_t G_t G_r c^2 \sigma}{(4\pi)^3 f^2 (S/N)_{\min}(NF)kT_oB}\right]^{\frac{1}{4}} \text{ or } \left[\frac{P_t G_t A_e \sigma}{(4\pi)^2 (S/N)_{\min}(NF)kT_oB}\right]^{\frac{1}{4}} \qquad \textbf{[21]}$$

P_t in equations [17], [19], and [21] is the peak power of a CW or pulse signal. For pulse signals these equations assume the radar pulse is square. If not, there is less power since P_t is actually the average power within the pulse width of the radar signal. Equations [17] and [19] relate the maximum detection range to $S_{\min}$, the minimum signal which can be detected and processed (the receiver sensitivity). The bandwidth (B) in equations [20] and [21] is directly related to $S_{\min}$. B is approximately equal to 1/PW. Thus a wider pulse width means a narrower receiver bandwidth which lowers $S_{\min}$, assuming no integration.

One cannot arbitrarily change the receiver bandwidth, since it has to match the transmitted signal. The "widest pulse width" occurs when the signal approaches a CW signal (see Section 2-11). A CW signal requires a very narrow bandwidth (approximately 100 Hz). Therefore, receiver noise is very low and good sensitivity results (see Section 5-2). If the radar pulse is narrow, the receiver filter bandwidth must be increased for a match (see Section 5-2), i.e. a 1 μs pulse requires a bandwidth of approximately 1 MHz. This increases receiver noise and decreases sensitivity.

If the radar transmitter can increase its PRF (decreasing PRI) and its receiver performs integration over time, an increase in PRF can permit the receiver to "pull" coherent signals out of the noise thus reducing $S/N_{\min}$ thereby increasing the detection range. Note that a PRF increase may limit the maximum range due to the creation of overlapping return echoes (see Section 2-10).

There are also other factors that limit the maximum practical detection range. With a scanning radar, there is loss if the receiver integration time exceeds the radar's time on target. Many radars would be range limited by line-of-sight/radar horizon (see Section 2-9) well before a typical target faded below $S_{\min}$. Range can also be reduced by losses due to antenna polarization and atmospheric absorption (see Sections 3-2 and 5-1).

Two-Way Radar Equation (Example)

Assume that a 5 GHz radar has a 70 dBm (10 kilowatt) signal fed through a 5 dB loss transmission line to a transmit/receive antenna that has 45 dB gain. An aircraft that is flying 31 km from the radar has an RCS of 9 m^2. What is the signal level at the input to the radar receiver? (There is an additional loss due to any antenna polarization mismatch but that loss will not be addressed in this problem). This problem continues in Sections 4-3, 4-7, and 4-10.

Answer:

Starting with: $10\log S = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1$ (in dB)

We know that: $\alpha_1 = 20\log fR + K_1 = 20\log(5\text{x}31) + 92.44 = 136.25$ dB

and that: $G_\sigma = 10\log\sigma + 20\log f_1 + K_2 = 10\log 9 + 20\log 5 + 21.46 = 44.98$ dB (see Table 1)
(Note: The aircraft transmission line losses (-5 dB) will be combined with the antenna gain (45 dB) for both receive and transmit paths of the radar)

So, substituting in we have: $10\log S = 70 + 40 + 40 + 44.98 - 2(136.25) =$ **-77.52 dBm** @ 5 GHz

The answer changes to -80.44 dBm if the tracking radar operates at 7 GHz provided the antenna gains and the aircraft RCS are the same at both frequencies.

$\alpha_1 = 20\log(7\text{x}31) + 92.44 = 139.17$ dB, $G_\sigma = 10\log 9 + 20\log 7 + 21.46 = 47.9$ dB (see Table 1)

$10\log S = 70 + 40 + 40 + 47.9 - 2(139.17) =$ **-80.44 dBm** @ 7 GHz

Table 1. Values of the Target Gain Factor (G_σ) in dB for Various Values of Frequency and RCS.

Frequency (GHz)	RCS - Square meters						
	0.05	5	9	10	100	1,000	10,000
0.5 GHz	2.44	22.42	24.98	25.44	35.44	45.44	55.44
1 GHz	8.46	28.46	31.0	31.46	41.46	51.46	61.46
5 GHz	22.44	42.44	44.98	45.44	55.44	65.44	75.44
7 GHz	25.36	45.36	47.9	48.36	58.36	68.36	78.36
10 GHz	28.46	48.46	51.0	51.46	61.46	71.46	81.46
20 GHz	34.48	54.48	57.02	57.48	67.48	77.48	87.48
40 GHz	40.50	60.48	63.04	63.5	73.5	83.5	93.5

Note: Shaded values were used in the examples.

TWO-WAY RADAR RANGE INCREASE AS A RESULT OF A SENSITIVITY INCREASE

As shown in equation [17] $S_{min}^{-1} \propto R_{max}^4$. Therefore, $-10\log S_{min} \propto 40\log R_{max}$ and the table below results:

% Range Increase: Range + (% Range Increase) x Range = New Range
i.e., for a 12 dB sensitivity increase, 500 miles +100% x 500 miles = 1,000 miles
Range Multiplier: Range x Range Multiplier = New Range
i.e., for a 12 dB sensitivity increase 500 miles x 2 = 1,000 miles

Table 2. Effects of Sensitivity Increase.

dB Sensitivity Increase	% Range Increase	Range Multiplier	dB Sensitivity Increase	% Range Increase	Range Multiplier
+ 0.5	3	1.03	10	78	1.78
1.0	6	1.06	11	88	1.88
1.5	9	1.09	12	100	2.00
2	12	1.12	13	111	2.11
3	19	1.19	14	124	2.24
4	26	1.26	15	137	2.37
5	33	1.33	16	151	2.51
6	41	1.41	17	166	2.66
7	50	1.50	18	182	2.82
8	58	1.58	19	198	2.98
9	68	1.68	20	216	3.16

TWO-WAY RADAR RANGE DECREASE AS A RESULT OF A SENSITIVITY DECREASE

As shown in equation [17] $S_{min}^{-1} \propto R_{max}^{4}$ Therefore, $-10 \log S_{min} \propto 40 \log R_{max}$ and the table below results:

% Range Decrease: Range - (% Range Decrease) x Range = New Range
i.e., for a 12 dB sensitivity decrease, 500 miles - 50% x 500 miles = 250 miles

Range Multiplier: Range x Range Multiplier = New Range
i.e., for a 12 dB sensitivity decrease 500 miles x 0.5 = 250 miles

Table 3. Effects of Sensitivity Decrease.

dB Sensitivity Decrease	% Range Decrease	Range Multiplier	dB Sensitivity Decrease	% Range Decrease	Range Multiplier
- 0.5	3	0.97	-10	44	0.56
- 1.0	6	0.94	- 11	47	0.53
- 1.5	8	0.92	- 12	50	0.50
- 2	11	0.89	- 13	53	0.47
- 3	16	0.84	- 14	55	0.45
- 4	21	0.79	- 15	58	0.42
- 5	25	0.75	- 16	60	0.40
- 6	29	0.71	- 17	62	0.38
- 7	33	0.67	- 18	65	0.35
- 8	37	0.63	- 19	67	0.33
- 9	40	0.60	- 20	68	0.32

ALTERNATE TWO-WAY RADAR EQUATION

In this section the same radar equation factors are grouped differently to create different constants as is used by some authors.

TWO-WAY RADAR EQUATION (MONOSTATIC)

Peak power at the radar receiver input is:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4} = \frac{P_t G_t G_{rsigma} c^2}{(4\pi)^3 f^2 R^4} * \quad (\mathit{Note}: \lambda = \frac{c}{f} \text{ and } \sigma \text{ is RCS}) \qquad \textbf{[1]}$$

* Keep λ or c, σ, and R in the same units. On reducing the above equation to log form we have:

or: $\mathbf{10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r - \alpha_2}$ (in dB)

Where: $\alpha_2 = 20\log f_1 R^2 - 10\log \sigma + K_3$, and $K_3 = -10\log c^2/(4\pi)^3$

Note:
Losses due to antenna polarization and atmospheric absorption (Sections 3-2 and 5-1) are not included in these equations

K_3 Values:
(dB)

Range Units	f_1 in MHz σ in m^2	f_1 in GHz σ in m^2	f_1 in MHz σ in ft^2	f_1 in GHz σ in ft^2
NM	114.15	174.15	124.47	184.47
km	103.44	163.44	113.76	173.76
m	-16.56	43.44	-6.24	53.76
yd	-18.1	41.9	-7.78	52.22
ft	-37.2	22.8	-26.88	33.12

In the last section, we had the basic radar equation given as equation [6] and it is repeated as equation [1] in the table above.

In Section 4-4, in order to maintain the concept and use of the one-way space loss coefficient, α_1, we didn't cancel like terms which was done to form equation [6] there. Rather, we regrouped the factors of equation [5]. This resulted in two minus α_1 terms and we defined the remaining term as G_σ, which accounted for RCS (see equation [8] & [9]).

Some authors take a different approach, and instead develop an entirely new single factor α_2, which is used instead of the combination of α_1 and G_σ.

If equation [1] is reduced to log form, (and noting that $f = c/\lambda$) it becomes:

$$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r - 20\log (f R^2) + 10\log \sigma + 10\log (c^2/(4\pi)^3) \qquad \textbf{[2]}$$

We now call the last three terms on the right minus α_2 and use it as a single term instead of the two terms α_1 and G_σ. The concept of dealing with one variable factor may be easier although we still need to know the range, frequency, and radar cross section to evaluate α_2. Additionally, we can no longer use a nomograph like we did in computing α_1 and visualize a two-way space loss consisting of two times the one-way space loss, since there are now 3 variables vs. two.

Equation [2] reduces to:

$$\mathbf{10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r - \alpha_2} \quad \text{(in dB)} \qquad \textbf{[3]}$$

Where $\alpha_2 = 20\log(f_1R^2) - 10\log\sigma + K_3$ and where f_1 is the MHz or GHz value of frequency and $K_3 = -10\log(c^2/(4\pi)^3) + 20\log$ (conversion for Hz to MHz or GHz)+ 40log (range unit conversions if not in meters) - 20log (RCS conversions for meters to feet)

The values of K_3 are given in the table above.

Comparing equation [3] to equation [10] in Section 4-4, it can be seen that $\alpha_2 = 2\alpha_1 - G_\sigma$.

TWO-WAY RADAR EQUATION (BISTATIC)

The following table contains a summary of the equations developed in this section.

TWO-WAY RADAR EQUATION (BISTATIC)

Peak power at the radar receiver input is:

$$P_r = \frac{P_t\, G_t\, G_r\, \lambda^2\, \sigma}{(4\pi)^3\, R_{Tx}^2\, R_{Rx}^2} = P_t\, G_t\, G_r \left[\frac{\sigma\, c^2}{(4\pi)^3\, f^2\, R_{Tx}^2\, R_{Rx}^2} \right]^*$$

Note : $\lambda = c/f$ and $\sigma = RCS$

** keep λ or $c, \sigma,$ and R in the same units*

On reducing the above equation to log form we have:
$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 20\log f + 20\log c - 30\log 4\pi - 20\log R_{Tx} - 20\log R_{Rx}$

or in simplified terms: $10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - \alpha_{Tx} - \alpha_{Rx}$ (in dB)

Where α_{Tx} corresponds to transmitter to target loss and α_{Rx} corresponds to target to receiver loss.

Note: Losses due to antenna polarization and atmospheric absorption (Sections 3-2 and 5-1) are not included in these equations.

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)

K_2 Values (dB)	RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

One-way free space loss, $\alpha_{Tx\text{ or }Rx} = 20\log (f_1 R_{Tx\text{ or }Rx}) + K_1$ (in dB)

K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	Km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

BISTATIC RADAR

There are also true bistatic radars - radars where the transmitter and receiver are in different locations as is depicted in Figure 1. The most commonly encountered bistatic radar application is the semi-active missile. The transmitter is located on, or near, the launch platform (surface or airborne), and the receiver is in the missile which is somewhere between the launch platform and the target.

The transmitting and receiving antennas are not the same and are not in the same location. Because the target-to-radar range is different from the target-to-missile range, the target-to-radar and target-to-missile space losses are different.

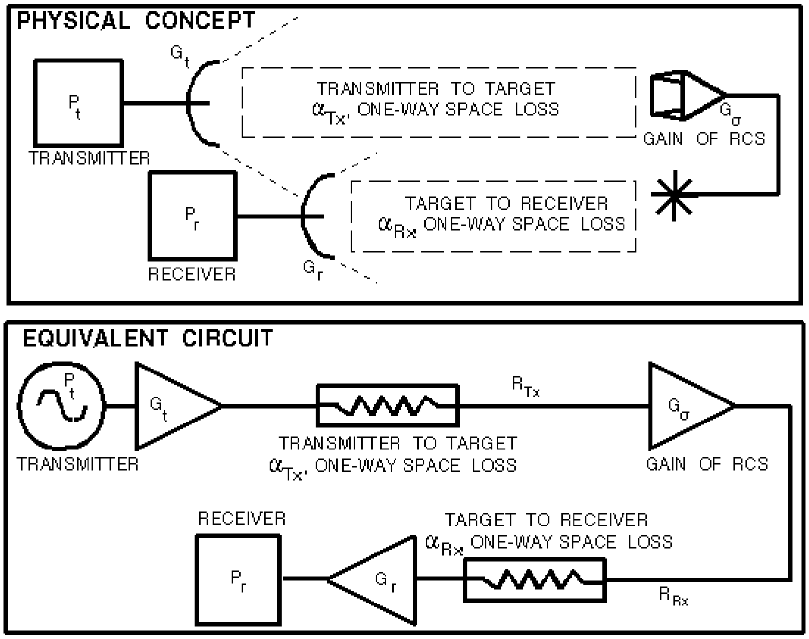
PHYSICAL CONCEPT
G_t
P_t
TRANSMITTER
TRANSMITTER TO TARGET
α_{Tx}, ONE-WAY SPACE LOSS
G_σ
GAIN OF RCS
P_r
RECEIVER
G_r
TARGET TO RECEIVER
α_{Rx} ONE-WAY SPACE LOSS
EQUIVALENT CIRCUIT
P_t
G_t
TRANSMITTER
R_{Tx}
TRANSMITTER TO TARGET
α_{Tx}, ONE-WAY SPACE LOSS
G_σ
GAIN OF RCS
RECEIVER
P_r
G_r
TARGET TO RECEIVER
α_{Rx} ONE-WAY SPACE LOSS
R_{Rx}

Figure 1. Bistatic Radar Visualized.

The peak power at the radar receiver input is:

$$P_r = \frac{P_t\, G_t\, G_r\, \lambda^2\, \sigma}{(4\pi)^3\, R_{Tx}^2\, R_{Rx}^2} = P_t\, G_t\, G_r \left[\frac{\sigma\, c^2}{(4\pi)^3\, f^2\, R_{Tx}^2\, R_{Rx}^2} \right] \qquad (\, Note : \lambda = \frac{c}{f} \text{ and } \sigma = RCS) \qquad \textbf{[1]}$$

Keep λ or c, σ, and R in the same units.

On reducing the above equation to log form we have:

$$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 20\log f + 20\log c - 30\log 4\pi - 20\log R_{Tx} - 20\log R_{Rx} \quad \textbf{[2]}$$

or in simplified terms:

$$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - \alpha_{Tx} - \alpha_{Rx} \quad \text{(in dB)} \qquad \textbf{[3]}$$

Where α_{Tx} corresponds to transmitter to target loss and α_{Rx} corresponds to target to receiver loss, or:

$\alpha_{Tx} = 20\log(f_1 T_{Tx}) + K_1$ (in dB) and $\alpha_{Rx} = 20\log(f_1 T_{Rx}) + K_1$ (in dB)

with K_1 values provided on page 4-6.1 and with f_1 being the MHz or GHz value of frequency.

Therefore, the difference between monostatic and bistatic calculations is that two α's are calculated for two different ranges and different gains may be required for transmit and receive antennas.

To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain.

As shown in Figure 2, it should also be noted that the bistatic RCS received by the missile is not always the same as the monostatic RCS. In general, the target's RCS varies with angle. Therefore, the bistatic RCS and monostatic RCS will be equal for receive and transmit antennas at the same angle to the target (but only if all three are in a line, as RCS also varies with elevation angle).

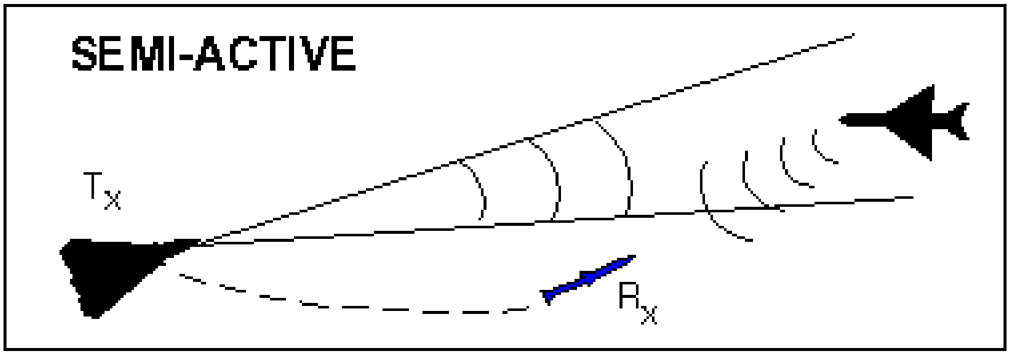

Figure 2. Bistatic RCS Varies.

JAMMING TO SIGNAL (J/S) RATIO - CONSTANT POWER [SATURATED] JAMMING

The following table contains a summary of the equations developed in this section.

<table>
<tr>
<td>JAMMING TO SIGNAL (J/S) RATIO (MONOSTATIC)
$J/S = (P_j G_{ja} 4\pi R^2) / (P_t G_t \sigma)$ (ratio form)* or:
$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log\sigma^* + 10.99 \text{ dB} + 20\log R^*$
Note (1): Neither f nor λ terms are part of these equations</td>
<td rowspan="2">* Keep R and σ in same units
Target gain factor, (in dB)
$G_\sigma = 10\log\sigma + 20\log f_1 + K_2$
K_2 Values (dB):
<table>
<tr><td>RCS (σ) (units)</td><td>f_1 in MHz K_2 =</td><td>f_1 in GHz K_2 =</td></tr>
<tr><td>m2</td><td>-38.54</td><td>21.46</td></tr>
<tr><td>ft2</td><td>-48.86</td><td>11.14</td></tr>
</table>
</td>
</tr>
<tr>
<td>If simplified radar equations developed in previous sections are used:
$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_1$ (in dB)
Note (2): the $20\log f_1$ term in $-G_\sigma$ cancels the $20\log f_1$ term in α_1</td>
</tr>
<tr>
<td>JAMMING TO SIGNAL (J/S) RATIO (BISTATIC)
R_{Tx} is the range from the radar transmitter to the target. See note (1).
$J/S = (P_j G_{ja} 4\pi R_{Tx}^2) / (P_t G_t \sigma)$ (ratio form) * or:
$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log\sigma^* + 10.99 \text{ dB} + 20\log R_{Tx}^*$</td>
<td rowspan="2">One-way free space loss (dB)
α_1 or α_{Tx} =
$20\log(f_1 R) + K_1$
K_1 Values (dB):
<table>
<tr><td>Range (units)</td><td>f_1 in MHz K_1 =</td><td>f_1 in GHz K_1 =</td></tr>
<tr><td>NM</td><td>37.8</td><td>97.8</td></tr>
<tr><td>km</td><td>32.45</td><td>92.45</td></tr>
<tr><td>m</td><td>-27.55</td><td>32.45</td></tr>
<tr><td>ft</td><td>-37.87</td><td>22.13</td></tr>
</table>
</td>
</tr>
<tr>
<td>If simplified radar equations developed in previous sections are used: see note (2).
$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_{Tx}$ (in dB)</td>
</tr>
</table>

This section derives the J/S ratio from the one-way range equation for J and the two-way range equation for S, and deals exclusively with active (transmitting) Electronic Attack (EA) devices or systems. Furthermore, the only purpose of EA is to prevent, delay, or confuse the radar processing of target information.

By official definition, EA can be either Jamming or Deception. This may be somewhat confusing because almost any type of active EA is commonly called "jamming," and the calculations of EA signal in the radar compared to the target signal in the radar commonly refer to the "jamming-to-signal" ratio ("J-to-S" ratio). Therefore this section uses the common jargon and the term "jammer" refers to any EA transmitter, and the term "jamming" refers to any EA transmission, whether Deception or Concealment.

Jamming: "Official" jamming should more aptly be called Concealment or Masking. Essentially, Concealment uses electronic transmissions to swamp the radar receiver and hide the targets. Concealment (Jamming) usually uses some form of noise as the transmitted signal. In this section, Concealment will be called "noise" or "noise jamming."

Deception: Deception might be better called Forgery. Deception uses electronic transmissions to forge false target signals that the radar receiver accepts and processes as real targets.

"J" designates the EA signal strength whether it originates from a noise jammer or from a deception system.

Basically, there are two different methods of employing active EA against hostile radars:

Self Protection EA
Support EA

For most practical purposes, Self Protection EEA is usually Deception and Support EA is usually noise jamming. As the name implies, Self Protection EA is EA that is used to protect the platform that it is on. Self Protection EA is often called “self screening jamming,” “Defensive EA” or historically “Deception ECM.” The top half of Figure 1 shows self-screening jamming.

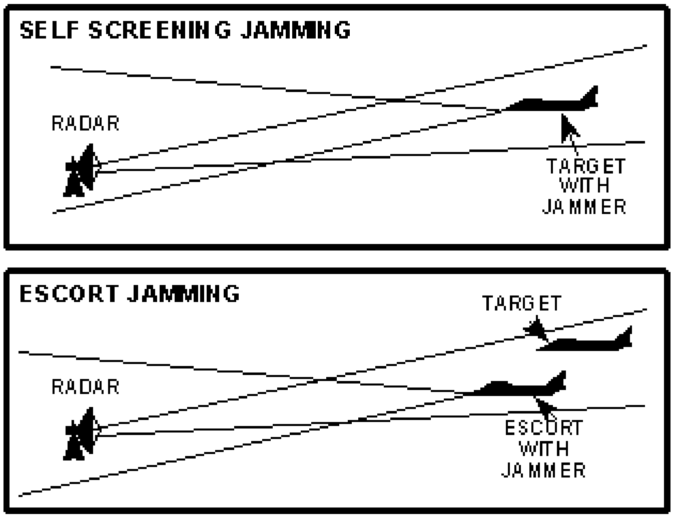

Figure 1. Self Protection and Escort Jamming.

The bottom half of Figure 1 illustrates escort jamming which is a special case of support jamming. If the escort platform is sufficiently close to the target, the J-to-S calculations are the same as for self protection EA.

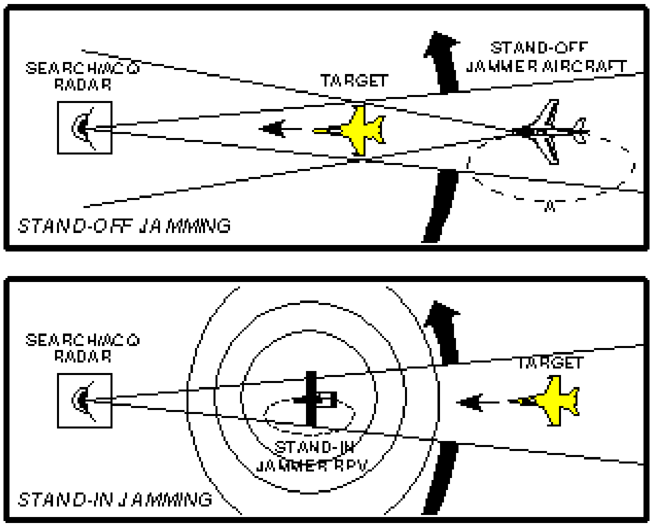

Figure 2. Support Jamming.

Support EA is electronic transmissions radiated from one platform and is used to protect other platforms or fulfill other mission requirements, like distraction or conditioning. Figure 2 illustrates two cases of support jamming protecting a striker - stand-off jamming (SOJ) and stand-in jamming (SIJ). For SOJ the support jamming platform is maintaining an orbit at a long range from the radar - usually beyond weapons range. For SIJ, a remotely piloted vehicle is orbiting very close to the victim radar. Obviously, the jamming power required for the SOJ to screen a target is much greater than the jamming power required for the SIJ to screen the same target.

When factoring EA into the radar equation, the quantities of greatest interest are “J-to-S” and Burn-Through Range.

“J-to-S” is the ratio of the signal strength of the jammer signal (J) to the signal strength of the target return signal (S). It is expressed as “J/S” and, in this section, is always in dB. J usually (but not always) must exceed S by some amount to be effective, therefore the desired result of a J/S calculation in dB is a positive number. Burn-through Range is the radar to target range where the target return signal can first be detected through the jamming and is usually slightly farther than crossover range where J=S. It is usually the range where the J/S just equals the minimum effective J/S (See Section 4-8).

The significance of "J-to-S" is sometimes misunderstood. The effectiveness of EA **is not** a direct mathematical function of "J-to-S." The magnitude of the "J-to-S" required for effectiveness is a function of the particular EA technique and of the radar it is being used against. Different EA techniques may very well require different "J-to-S" ratios against the same radar. When there is sufficient "J-to-S" for effectiveness, increasing it will rarely increase the effectiveness at a given range. Because modern radars can have sophisticated signal processing and/or EP capabilities, in certain radars too much "J-to-S" could cause the signal processor to ignore the jamming, or activate special anti-jamming modes. Increasing "J-to-S" (or the jammer power) does, however, allow the target aircraft to get much closer to the threat radar before burn-through occurs, which essentially means more power is better if it can be controlled when desired.

IMPORTANT NOTE: If the signal S is CW or PD and the Jamming J is amplitude modulated, then the J used in the formula has to be reduced from the peak value (due to sin x/x frequency distribution). The amount of reduction is dependent upon how much of the bandwidth is covered by the jamming signal. To get an exact value, integrals would have to be taken over the bandwidth. As a rule of thumb however:

- If the frequency of modulation is less than the BW of the tracking radar reduce J/S by 10 Log (duty cycle).
- If the frequency of modulation is greater than the BW of the tracking radar reduce J/S by 20 Log(duty cycle).

For example; if your jamming signal is square wave chopped (50% duty cycle) at a 100 Hz rate while jamming a 1 kHz bandwidth receiver, then the J/S is reduced by 3 dB from the maximum. If the duty cycle was 33%, then the reduction would be 4.8 dB. If the 50% and 33% duty cycle jamming signals were chopped at a 10 kHz (vice the 100 Hz) rate, the rule of thumb for jamming seen by the receiver would be down 6 dB and 9.6 dB, respectively, from the maximum since the 10 kHz chopping rate is greater than the 1 kHz receiver BW.

J/S for SELF PROTECTION EA vs. MONOSTATIC RADAR

Figure 3 is radar jamming visualized. The Physical concept of Figure 3 shows a monostatic radar that is the same as Figure 1, Section 4-4, and a jammer (transmitter) to radar (receiver) that is the same as Figure 3, Section 4-3. In other words, Figure 3 is simply the combination of the previous two visual concepts where there is only one receiver (the radar's).

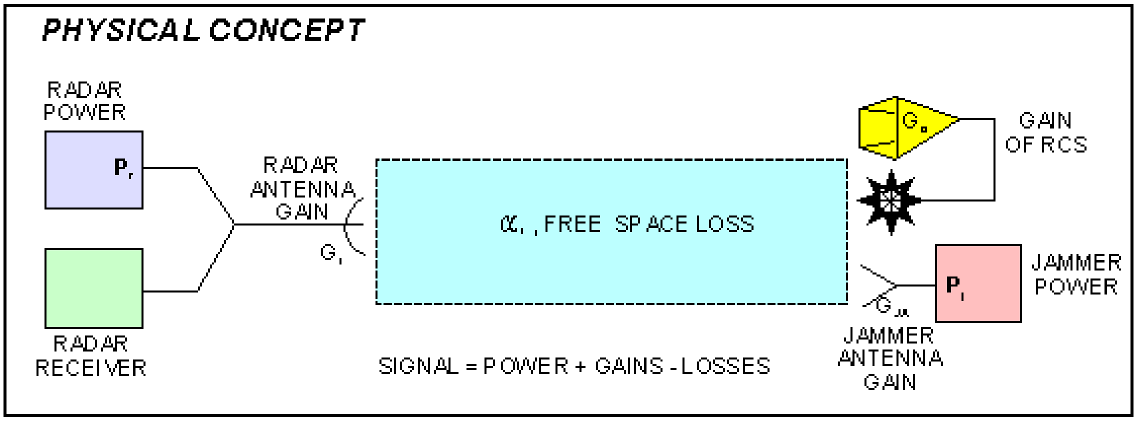

Figure 3. Radar Jamming Visualized.

The equivalent circuit shown in Figure 4 applies to jamming monostatic radars with either self protect EA or support EA. For self protect (or escort) vs. a monostatic radar, the jammer is on the target and the radar receive and transmit antennas are collocated so the three ranges and three space loss factors (α's) are the same.

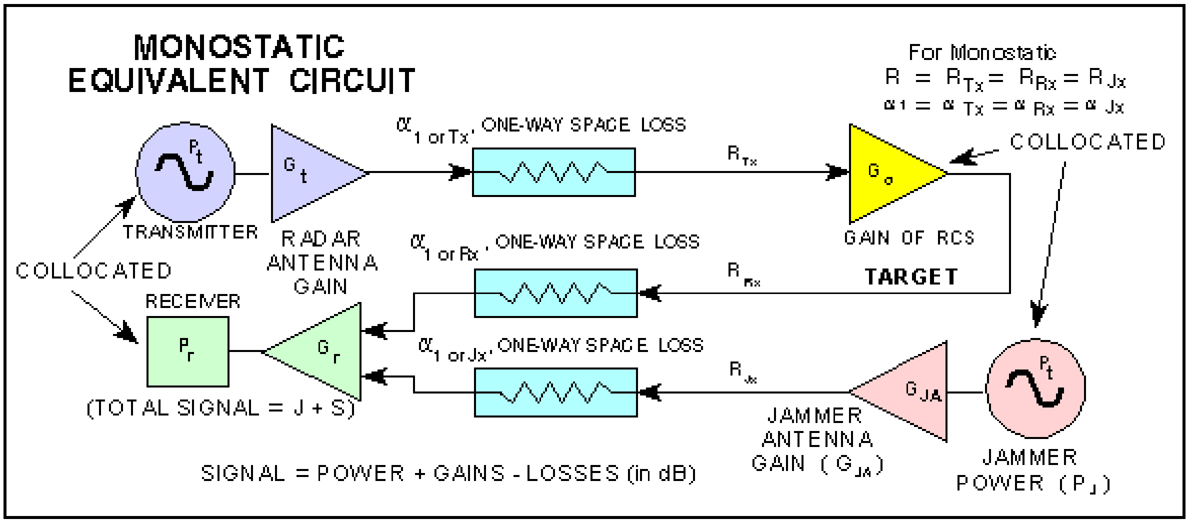
MONOSTATIC
EQUIVALENT CIRCUIT
For Monostatic
R = R_Tx = R_Rx = R_Jx
α1 = α Tx = α Rx = α Jx
P_t
TRANSMITTER
G_t
RADAR
ANTENNA
GAIN
α_1 or Tx, ONE-WAY SPACE LOSS
R_Tx
G_σ
GAIN OF RCS
TARGET
COLLOCATED
α_1 or Rx, ONE-WAY SPACE LOSS
R_Rx
COLLOCATED
RECEIVER
P_r
G_r
(TOTAL SIGNAL = J + S)
α_1 or Jx, ONE-WAY SPACE LOSS
R_Jx
G_JA
P_t
JAMMER
ANTENNA
GAIN (G_JA)
JAMMER
POWER (P_J)
SIGNAL = POWER + GAINS - LOSSES (in dB)

Figure 4. Monostatic Radar EA Equivalent Circuit.

J-S Ratio (Monostatic) - The ratio of the power received (P_{r1} or J) from the jamming signal transmitted from the target to the power received (P_{r2} or S) from the radar skin return from the target equals J/S.

From the one way range equation in Section 4-3: $$P_r 1 \text{ or } J = \frac{P_j\, G_{ja}\, G_r\, \lambda^2}{(4\pi R)^2} \qquad \textbf{[1]}$$

From the two way range equation in Section 4.4: $$P_r 2 \text{ or } S = \frac{P_t\, G_t\, G_r\, \lambda^2\, \sigma}{(4\pi)^3\, R^4} \qquad \textbf{[2]}$$

so $$\frac{J}{S} = \frac{P_j\, G_{ja}\, G_r\, \lambda^2 (4\pi)^3 R^4}{P_t\, G_t\, G_r\, \lambda^2\, \sigma\, (4\pi R)^2} = \frac{P_j\, G_{ja}\, 4\pi\, R^2}{P_t\, G_t\, \sigma}^{*} \qquad \textit{(ratio form)} \qquad \textbf{[3]}$$

* Keep R and σ in the same units.

On reducing the above equation to log form we have:

10log J/S = 10log P_j + 10log G_{ja} - 10log P_t - 10log G_t - 10log σ + 10log 4π + 20log R **[4]**

or 10log J/S = 10log P_j + 10log G_{ja} - 10log P_t - 10log G_t - 10log σ + 10.99 dB + 20log R **[5]**

Note: Neither *f* nor λ terms are part of the final form of equation [3] and equation [5].

J/S Calculations (Monostatic) Using a One Way Free Space Loss - The simplified radar equations developed in previous sections can be used to express J/S.

From the one way range equation Section 4-3:

$$10\log (P_{r1} \text{ or } J) = 10\log P_j + 10\log G_{ja} + 10\log G_r - \alpha_1 \quad \text{(in dB)} \qquad \textbf{[6]}$$

From the two way range equation in Section 4.4:

$$10\log (P_{r2} \text{ or } S) = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1 \quad \text{(in dB)} \qquad \textbf{[7]}$$

$$10\log (J/S) = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_1 \quad \text{(in dB)} \qquad \textbf{[8]}$$

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain. The $20\log f_1$ term in $-G_\sigma$ cancels the $20\log f_1$ term in α_1.

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)

K_2 Values (dB)	RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

One-way free space loss, $\alpha_1 = 20\log (f_1R) + K_1$ (in dB)

K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

J/S for SELF PROTECTION EA vs. BISTATIC RADAR

The semi-active missile illustrated in Figure 5 is the typical bistatic radar which would require the target to have self protection EA to survive. In this case, the jammer is on the target and the target to missile receiver range is the same as the jammer to receiver range, but the radar to target range is different. Therefore, only two of the ranges and two of the α's (Figure 6.) are the same.

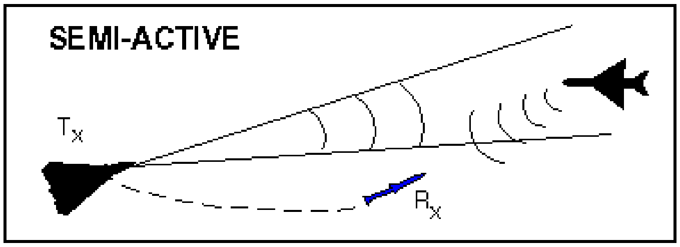

Figure 5. Bistatic Radar.

In the following equations:

α_{Tx} = The one-way space loss from the radar transmitter to the target for range R_{Tx}
α_{Rx} = The one-way space loss from the target to the missile receiver for range R_{Rx}

Like the monostatic radar, the bistatic jamming and reflected target signals travel the same path from the target and enter the receiver (missile in this case) via the same antenna. In both monostatic and bistatic J/S equations this common range cancels, so both J/S equations are left with an R_{Tx}^2 or $20 \log R_{Tx}$ term. Since in the monostatic case $R_{Tx} = R_{Rx}$ and $\alpha_{Tx} = \alpha_{Rx}$, only R or α_1 is used in the equations. Therefore, the bistatic J/S equations [11], [13], or [14] will work for monostatic J/S calculations, but the opposite is only true if bistatic R_{Tx} and α_{Tx} terms are used for R or α_1 terms in monostatic equations [3], [5], and [8].

The equivalent circuit shown in Figure 6 applies to jamming bistatic radar. For self protect (or escort) vs. a bistatic radar, the jammer is on the target and the radar receive and transmit antennas are at separate locations so only two of the three ranges and two of the three space loss factors (α's) are the same.

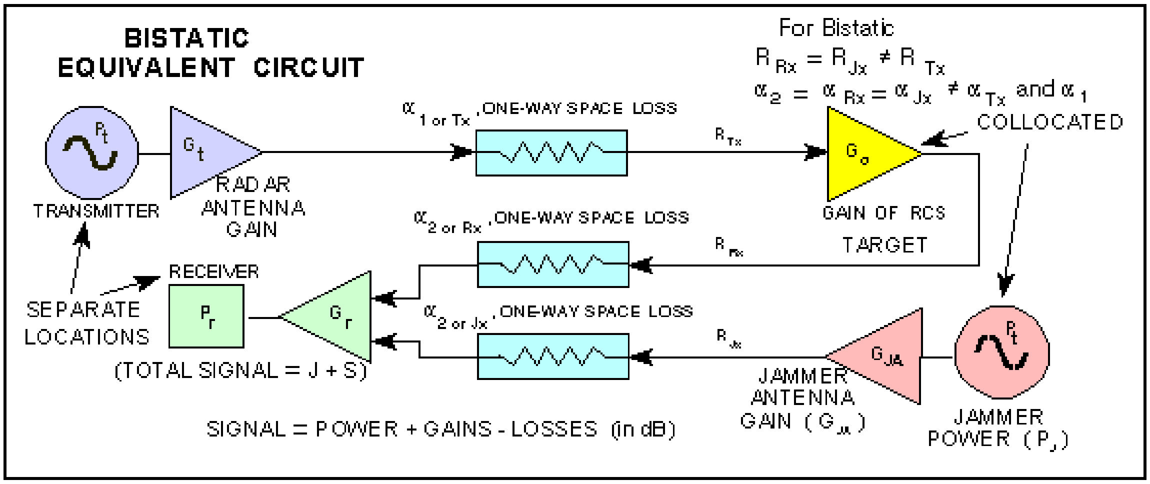
BISTATIC EQUIVALENT CIRCUIT
For Bistatic
$R_{Rx} = R_{Jx} \neq R_{Tx}$
$\alpha_2 = \alpha_{Rx} = \alpha_{Jx} \neq \alpha_{Tx}$ and α_1
COLLOCATED
P_t
TRANSMITTER
G_t
RADAR ANTENNA GAIN
$\alpha_{1 \text{ or } Tx}$, ONE-WAY SPACE LOSS
R_{Tx}
G_σ
GAIN OF RCS
TARGET
$\alpha_{2 \text{ or } Rx}$, ONE-WAY SPACE LOSS
R_{Rx}
RECEIVER
SEPARATE LOCATIONS
P_r
G_r
$\alpha_{2 \text{ or } Jx}$, ONE-WAY SPACE LOSS
R_{Jx}
G_{JA}
P_t
(TOTAL SIGNAL = J + S)
JAMMER ANTENNA GAIN (G_{JA})
JAMMER POWER (P_J)
SIGNAL = POWER + GAINS - LOSSES (in dB)

Figure 6. Bistatic Radar EA Equivalent Circuit.

J-to-S Ratio (Bistatic) When the radar's transmit antenna is located remotely from the receiving antenna (Figure 6), the ratio of the power received (P_{r1} or J) from the jamming signal transmitted from the target to the power received (P_{r2} or S) from the radar skin return from the target equals J/S. For jammer effectiveness J normally has to be greater than S.

From the one way range equation in Section 4-3:

$$P_r 1 \text{ or } J = \frac{P_j G_{ja} G_r \lambda^2}{(4\pi R_{Rx})^2} \qquad (R_{Jx} = R_{Rx})$$

[9]

From the two way range equation in Section 4.4:

$$P_r 2 \text{ or } S = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R_{Tx}^2 R_{Rx}^2}$$

[10]

so

$$\frac{J}{S} = \frac{P_j G_{ja} G_r \lambda^2 (4\pi)^3 R_{Tx}^2 R_{Rx}^2}{P_t G_t G_r \lambda^2 \sigma (4\pi R_{Rx})^2} = \frac{P_j G_{ja} 4\pi R_{Tx}^2}{P_t G_t \sigma}^{*} \qquad \textit{(ratio form)}$$

[11]

* Keep R and σ in the same units.

On reducing the above equation to log form we have:

$$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10\log 4\pi + 20\log R_{Tx}$$

[12]

or

$$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10.99\text{ dB} + 20\log R_{Tx}$$

[13]

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain. Neither *f* nor λ terms are part of the final form of equation [11] and equation [13].

Bistatic J/S Calculations (Bistatic) Using a One Way Free Space Loss - The simplified radar equations developed in previous sections can be used to express J/S.

From the one way range equation in Section 4-3:

$10\log (P_{r1} \text{ or } J) = 10\log P_j + 10\log G_{ja} + 10\log G_r - \alpha_{Rx}$ (all factors dB) **[14]**

From the two way range equation in Section 4-4:
$10\log (P_{r2} \text{ or } S) = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - \alpha_{Tx} - \alpha_{Rx}$ (all factors dB) **[15]**

$10\log (J/S) = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_{Tx}$ (all factors dB) **[16]**

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain. The $20\log f_1$ term in $-G_\sigma$ cancels the $20\log f_1$ term in α_1.

Target gain factor, $G_\sigma = 10\log\sigma + 20\log f_1 + K_2$ (in dB)			
K_2 Values (dB)	RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

One-way free space loss $\alpha_{Tx\text{ or }Rx} = 20\log f_1 R_{Tx\text{ or }Rx} + K_1$ (in dB)			
K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

Saturated J/S (Monostatic) Example (Constant Power Jamming)

Assume that a 5 GHz radar has a 70 dBm signal fed through a 5 dB loss transmission line to an antenna that has 45 dB gain. An aircraft is flying 31 km from the radar. The aft EW antenna has -1 dB gain and a 5 dB line loss to the EW receiver (there is an additional loss due to any antenna polarization mismatch but that loss will not be addressed in this problem). The aircraft has a jammer that provides 30 dBm saturated output if the received signal is above -35 dBm. The jammer feeds a 10 dB loss transmission line which is connected to an antenna with a 5 dB gain. If the RCS of the aircraft is 9 m^2, what is the J/S level received by the tracking radar?

Answer: The received signal at the jammer is the same as the example in Section 4-3, i.e. answer (1) = -32.3 dBm @ 5 GHz. Since the received signal is above -35 dBm, the jammer will operate in the saturated mode, and equation [5] can be used. (See Section 4-10 for an example of a jammer in the linear region.)

$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log\sigma + 10.99 \text{ dB} + 20\log R$

Note: the respective transmission line losses will be combined with antenna gains,
i.e. -5 + 45 = 40 dB & -10 +5 = -5 dB.

$10\log J/S = 30 - 5 - 70 - 40 - 9.54 + 10.99 + 89.8 = 6.25$ dB @ 5 GHz*

* The answer is still 6.25 dB if the tracking radar operates at 7 GHz provided the antenna gains and the aircraft RCS are the same at both frequencies.

In this example, there is inadequate jamming power at each frequency if the J/S needs to be 10 dB or greater to be effective. One solution would be to replace the jammer with one that has a greater power output. If the antenna of the aircraft and the radar are not the proper polarization, additional power will also be required (see Section 3-2).

BURN-THROUGH / CROSSOVER RANGE

The burn-through equations are derived in this section. These equations are most commonly used in jammer type of applications. The following is a summary of the important equations explored in this section:

J/S CROSSOVER RANGE (MONOSTATIC) (J = S)

$R_{J=S} = [\ (P_t\ G_t\ \sigma) / (P_j\ G_{ja}\ 4\pi)\ ]^{1/2}$ (dB Ratio)

or $20 \log R_{J=S} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} - 10.99$ dB

If simplified radar equations already converted to dB are used:

$20 \log R_{J=S} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 - 20\log f_1$ (dB)

* Keep P_t & P_j in same units
Keep R and σ in same units

K_1 Values (dB):

Range (units)	f_1 in MHz K_1 =	in GHz K_1 =
m	-27.55	32.45
ft	-37.87	22.13

BURN-THROUGH RANGE (MONOSTATIC)

The radar to target range where the target return signal (S) can first be detected through the EA (J).

$R_{BT} = [\ (P_t\ G_t\ \sigma\ J_{min\ eff}) / (P_j\ G_{ja}\ 4\pi\ S)\ ]^{1/2}$ (dB Ratio)

or $20\log R_{BT} = 10\log P_t + 10\log G_t + 10\log\sigma - 10\log P_j - 10\log G_{ja} + 10\log(J_{min\ eff}/S) - 10.99$

If simplified radar equations already converted to dB are used:

$20\log R_{BT} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 + 10\log(J_{min\ eff}/S) - 20\log f_1$(in dB)*

f_1 is MHz or GHz value of frequency

Target gain factor (dB)

$G_\sigma = 10\log \sigma + 20\log f_1 + K_2$

K_2 Values (dB):

RCS (σ) (units)	f_1 in MHz K_2 =	in GHz K_2 =
m^2	-38.54	21.46
ft^2	-48.86	11.14

BURN-THROUGH RANGE (BISTATIC)

R_{Tx} is the range from the radar transmitter to the target and is different from R_{Rx} which is the range from the target to the receiver. Use Monostatic equations and substitute R_{Tx} for R

CROSSOVER RANGE and BURN-THROUGH RANGE

To present the values of J and S, (or J/S) over a minimum to maximum radar to target range of interest, equation [1], Section 4-7. which has a slope of 20 log for J vs. range and equation [2], Section 4-7, which has a slope of 40 log for S vs. range are plotted. When plotted on semi-log graph paper, J and S (or J/S) vs. range are straight lines as illustrated in Figure 1.

Figure 1 is a sample graph - it cannot be used for data.

The crossing of the J and S lines (known as crossover) gives the range where J = S (about 1.29 NM), and shows that shorter ranges will produce target signals greater than the jamming signal.

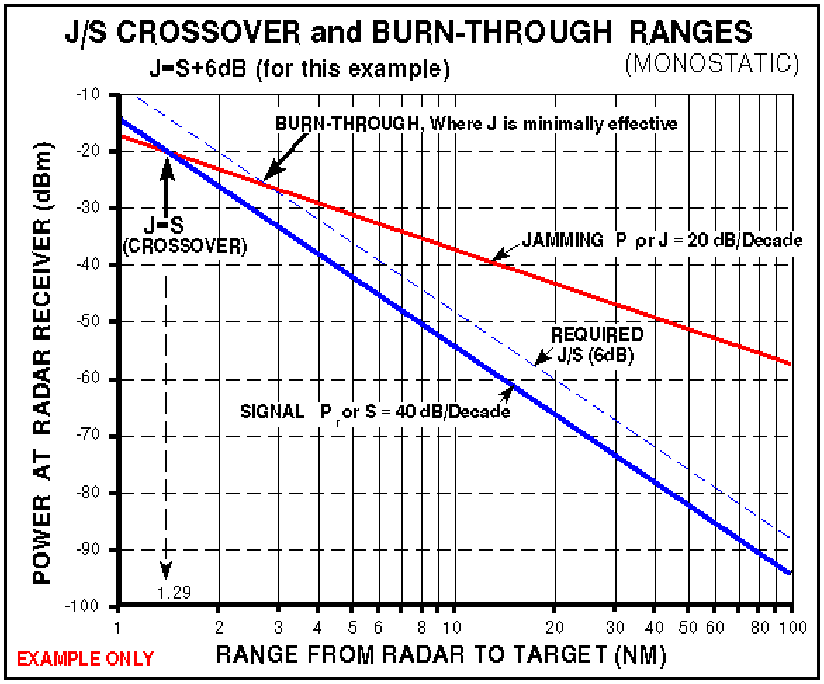

Figure 1. Sample J and S Graph.

The point where the radar power overcomes the jamming signal is known as burn-through. The crossover point where J = S could be the burn-through range, but it usually isn't because normally J/S > 0 dB to be effective due to the task of differentiating the signal from the jamming noise floor (see receiver sensitivity section). For this example, the J/S required for the EA to be effective is given as 6 dB, as shown by the dotted line. This required J/S line crosses the jamming line at about 2.8 NM which, in this example, is the burn-through range.

In this particular example, we have:

P_t = 80 dBm G_t = 42 dB
P_j = 50 dBm G_{ja} = 6 dB
$\sigma = 18\ m^2$ f = 5.9 GHz (not necessary for all calculations)

A radar can be designed with higher than necessary power for earlier burn-through on jamming targets. Naturally that would also have the added advantage of earlier detection of non-jamming targets as well.

Note: To avoid having to include additional terms for the following calculations, always combine any transmission line loss with antenna gain.

CROSSOVER AND BURN-THROUGH RANGE EQUATIONS (MONOSTATIC)

To calculate the crossover range or burn-through range the J/S equation must be solved for range. From equation [3], Section 4-7:

$$\frac{J}{S} = \frac{P_j\, G_{ja}\, 4\pi\, R^2}{P_t\, G_t\, \sigma} \quad \textit{(ratio form)} \qquad \text{Solving for R:} \quad R = \sqrt{\frac{P_t\, G_t\, \sigma\, J}{P_j\, G_{ja}\, 4\pi\, S}} \qquad \textbf{[1]}$$

BURN-THROUGH RANGE (MONOSTATIC) - Burn-through Range (Monostatic) is the radar to target range where the target return signal (S) can first be detected through the jamming (J). It is usually the range when the J/S just equals the minimum effective J/S.

$$R_{BT} = \sqrt{\frac{P_t\, G_t\, \sigma\, J_{\min eff}}{P_j\, G_{ja}\, 4\pi\, S}} \quad \text{(burn-through range)} \qquad \textbf{[2]}$$

or in dB form, (using 10log 4π = 10.99 dB):

$$20\log R_{BT} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} + 10\log (J_{\min eff}/S) - 10.99 \text{ dB} \qquad \textbf{[3]}$$

RANGE WHEN J/S CROSSOVER OCCURS (MONOSTATIC) - The crossover of the jammer's 20 dB/decade power line and the skin return signal's 40 dB/decade power line of Figure 1 occurs for the case where J = S in dB or J/S=1 in ratio. Substituting into equation [1] yields:

$$R_{(J=S)} = \sqrt{\frac{P_t\, G_t\, \sigma}{P_j\, G_{ja}\, 4\pi}} \quad \text{(Crossover range)} \qquad \textbf{[4]}$$

or in dB form:

$$20\log R_{J=S} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} - 10.99 \text{ dB} \qquad \textbf{[5]}$$

Note: keep R and σ in same units in all equations.

CROSSOVER AND BURN-THROUGH EQUATIONS (MONOSTATIC)
USING α - ONE WAY FREE SPACE LOSS

The other crossover burn-through range formulas can be confusing because a frequency term is subtracted (equations [6], [7] and [8]), but both ranges are independent of frequency. This subtraction is necessary because when J/S is calculated directly as previously shown, λ^2 or $(c/f)^2$ terms canceled, whereas in the simplified radar equations, a frequency term is part of the G_σ term and has to be cancelled if one solves for R. From equation [8], Section 4-7:

$$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_1 \quad \text{(factors in dB)}$$

or rearranging: $\alpha_1 = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} + 10\log (J/S)$

from Section 4-4: $\alpha_1 = 20\log f_1 R_1 + K_1$ or $20\log R_1 = \alpha_1 - K_1 - 20\log f_1$

then substituting for α_1:

$$20\log R_1 = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 + 10\log (J/S) - 20\log f_1 \quad \text{(in dB)} \qquad \textbf{[6]}$$

EQUATION FOR BURN-THROUGH RANGE (MONOSTATIC) - Burn-through occurs at the range when the J/S just equals the minimum effective J/S. G_σ and K_1 are as defined on page 4-8.1.

$$20\log R_{BT} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 + 10\log (J_{\text{min eff}}/S) - 20\log f_1 \quad \text{(in dB)} \qquad \textbf{[7]}$$

EQUATION FOR THE RANGE WHEN J/S CROSSOVER OCCURS (MONOSTATIC)

The J/S crossover range occurs for the case where J = S, substituting into equation [6] yields:

$$20\log R_{J=S} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 - 20\log f_1 \quad \text{(factors in dB)} \qquad \textbf{[8]}$$

BURN-THROUGH RANGE (BISTATIC)

Bistatic J/S crossover range is the radar-to-target range when the power received (S) from the radar skin return from the target equals the power received (J) from the jamming signal transmitted from the target. As shown in Figure 6, Section 4-7, the receive antenna that is receiving the same level of J and S is remotely located from the radar's transmit antenna. Bistatic equations [11], [13], and [14] in Section 4-7 show that J/S is only a function of radar to target range, therefore J/S is not a function of wherever the missile is in its flight path provided the missile is in the antenna beam of the target's jammer. The missile is closing on the target at a very much higher rate than the target is closing on the radar, so the radar to target range will change less during the missile flight.

It should be noted that for a very long range air-to-air missile shot, the radar to target range could typically decrease to 35% of the initial firing range during the missile time-of-flight, i.e. A missile shot at a target 36 NM away, may be only 12 NM away from the firing aircraft at missile impact.

Figure 2 shows both the jamming radiated from the target and the power reflected from the target as a function of radar-to-target range. In this particular example, the RCS is assumed to be smaller, 15 m^2 vice 18 m^2 in the monostatic case, since the missile will be approaching the target from a different angle. This will not, however, always be the case.

In this plot, the power reflected is:

$$P_{ref} = \frac{P_t\, G_t\, 4\,\pi\,\sigma}{(4\pi R\,)^2}$$

Substituting the values given previously in the example on page 4-8.1, we find that the crossover point is at 1.18 NM (due to the assumed reduction in RCS).

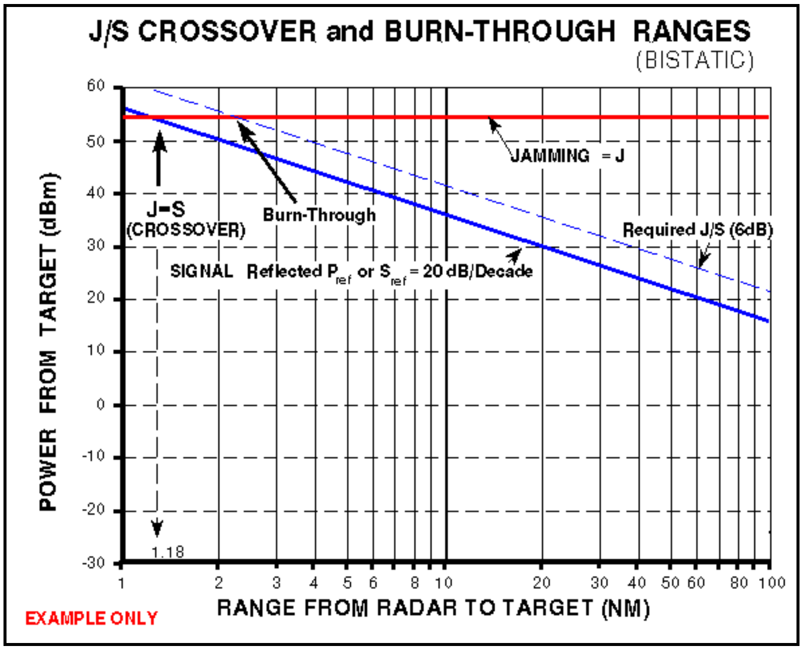

Figure 2. Bistatic Crossover and Burnthrough.

CROSSOVER AND BURN-THROUGH RANGE EQUATIONS (BISTATIC)

To calculate the radar transmitter-to-target range where J/S crossover or burn-through occurs, the J/S equation must be solved for range. From equation [11] in Section 4-7:

$$\frac{J}{S} = \frac{P_j\, G_{ja}\, 4\pi\, R_{Tx}^2}{P_t\, G_t\, \sigma} \qquad \textit{(ratio form)}$$

Solving for R_{Tx}:

$$R_{Tx} = \sqrt{\frac{P_t\, G_t\, \sigma\, J}{P_j\, G_{ja}\, 4\pi\, S}} \qquad [9]$$

Note:
Bistatic equation [10] is identical to monostatic equation [1] except R_{Tx} must be substituted for R and a bistatic RCS (σ) will have to be used since RCS varies with aspect angle. The common explanations will not be repeated in this section.

BURN-THROUGH RANGE (BISTATIC) - Burn-through Range (Bistatic) occurs when J/S just equals the minimum effective J/S. From equation [9]:

$$R_{Tx(BT)} = \sqrt{\frac{P_t\, G_t\, \sigma\, J_{\min\, eff}}{P_j\, G_{ja}\, 4\pi\, S}} \qquad \textit{(ratio form)} \qquad [10]$$

or in dB form:

$$20\log R_{Tx(BT)} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} + 10\log (J_{min\ eff}/S) - 10.99\text{ dB} \quad \textbf{[11]}$$

If using the simplified radar equations (factors in dB):

$$20\log R_{Tx(BT)} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 + 10\log (J_{min\ eff}/S) - 20\log f_1 \quad \textbf{[12]}$$

Where G_σ and K_1 are defined on page 4-8.1

RANGE WHEN J/S CROSSOVER OCCURS (BISTATIC)

The crossover occurs when J = S in dB or J/S = 1 in ratio.

$$R_{Tx(J=S)} = \sqrt{\frac{P_t G_t \sigma}{P_j G_{ja} 4\pi}} \qquad \textit{(ratio)} \quad \textbf{[13]}$$

or in log form:

$$20\log R_{Tx(J=S)} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} - 10.99\text{ dB} \quad \textbf{[14]}$$

If simplified equations are used (with G_σ and K_1 as defined on page 4-8.1) we have:

$$20\log R_{Tx(J=S)} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 - 20\log f_1 \qquad \text{(factors in dB)} \quad \textbf{[15]}$$

Note: keep R and σ in same units in all equations.

DETAILS OF SEMI-ACTIVE MISSILE J/S

Unless you are running a large scale computer simulation that includes maneuvering, antenna patterns, RCS, etc., you will seldom calculate the variation in J/S that occurs during a semi-active missile's flight. Missiles don't fly straight lines at constant velocity. Targets don't either - they maneuver. If the launch platform is an aircraft, it maneuvers too. A missile will accelerate to some maximum velocity above the velocity of the launch platform and then decelerate.

The calculation of the precise variation of J/S during a missile flight for it to be effective requires determination of all the appropriate velocity vectors and ranges at the time of launch, and the accelerations and changes in relative positions during the fly out. In other words, it’s too much work for too little return. The following are simplified examples for four types of intercepts.

	J/S (dB)	ΔJ/S (dB)
At Launch:	29	n/a
Intercept Type	At 2 sec. to Intercept:	
AAM Head-on:	23	-6
SAM Incoming Target:	25	-4
AAM Tail Chase:	29	0
SAM Outbound Target:	35	+6

In these examples, all velocities are constant, and are all along the same straight line. The missile velocity is 800 knots greater than the launch platform velocity which is assumed to be 400 kts. The missile launch occurs at 50 NM.

For the AAM tail chase, the range from the radar to the target remains constant and so does the J/S. **In these examples** the maximum variation from launch J/S is ± 6 dB. That represents the difference in the radar to target range closing at very high speed (AAM head on) and the radar to target range opening at moderate speed (SAM outbound target). The values shown above are examples, not rules of thumb, every intercept will be different.

Even for the simplified linear examples shown, graphs of the J and S will be curves - not straight lines. Graphs could be plotted showing J and S vs. radar to target range, or J and S vs. missile to target range, or even J/S vs. time of flight. If the J/S at launch is just barely the minimum required for effectiveness, and increasing it is difficult, then a detailed graph may be warranted, but in most cases this isn’t necessary.

This page intentionally left blank.

SUPPORT JAMMING

The following table contains a summary of equations developed in this section:

<table>
<tr>
<td>MAIN LOBE JAMMING TO SIGNAL (J/S) RATIO (For SOJ/SIJ)
$J/S = (P_j G_{ja} 4\pi R_{Tx}^4) / (P_t G_t \sigma [BW_J/BW_R] R_{Jx}^2)$ (ratio form)*
$10\log J/S = 10\log P_j - 10\log[BW_J/BW_R] + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma +$ ↓
$10.99\text{ dB} + 40\log R_{Tx} - 20\log R_{Jx}$ * ←←
or if simplified radar equations are used:
$10\log J/S = 10\log P_j - BF + 10\log G_{ja} - \alpha_{jx} - 10\log P_t - 10\log G_t - G_\sigma + 2\alpha_1$ (in dB)*</td>
<td>Target gain factor,
$G_\sigma = 10\log\sigma + 20\log f_1 + K_2$ (in dB)
K_2 Values (dB):
<table>
<tr><td>RCS (σ)
(units)</td><td>f_1 in MHz
K_2 =</td><td>f_1 in GHz
K_2 =</td></tr>
<tr><td>m^2</td><td>-38.54</td><td>21.46</td></tr>
<tr><td>ft^2</td><td>-48.86</td><td>11.14</td></tr>
</table></td>
</tr>
<tr>
<td>SIDE LOBE JAMMING TO SIGNAL (J/S) RATIO (For SOJ/SIJ)
$J/S = (P_j G_{ja} G_{r(SL)} 4\pi R_{Tx}^4) / (P_t G_t G_{r(ML)} \sigma [BW_J/BW_R] R_{Jx}^2)$ (ratio form)*
$10\log J/S = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - 10\log P_t - 10\log G_t - 10\log G_{r(ML)} +$ → ↓
$10.99\text{ dB} - 10\log \sigma + 40\log R_{Tx} - 20\log R_{Jx}$ * ←←
or if simplified radar equations are used (in dB)*:
$10\log J/S = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - \alpha_{jx} - 10\log P_t - 10\log G_t - 10\log G_{r(ML)} - G_\sigma + 2\alpha_1$</td>
<td rowspan="2">One-way free space loss,
α_1 or $\alpha_{Tx} = 20\log(f_1 R) + K_1$ (in dB)
K_1 Values (dB):
<table>
<tr><td>Range
(units)</td><td>f_1 in MHz
K_1 =</td><td>f_1 in GHz
K_1 =</td></tr>
<tr><td>NM</td><td>37.8</td><td>97.8</td></tr>
<tr><td>Km</td><td>32.45</td><td>92.45</td></tr>
<tr><td>m</td><td>-27.55</td><td>32.45</td></tr>
<tr><td>yd</td><td>-28.33</td><td>31.67</td></tr>
<tr><td>ft</td><td>-37.87</td><td>22.13</td></tr>
</table>
* Keep R and σ in same units</td>
</tr>
<tr>
<td><table>
<tr><td>R_{Jx}</td><td>Range from the support jammer transmitter to the radar receiver</td></tr>
<tr><td>R_{Tx}</td><td>Range between the radar and the target</td></tr>
<tr><td>BF</td><td>10 Log of the ratio of BW_J of the noise jammer to BW_R of the radar receiver</td></tr>
<tr><td>$G_{r(SL)}$</td><td>Side lobe antenna gain</td></tr>
<tr><td>$G_{r(ML)}$</td><td>Main lobe antenna gain</td></tr>
<tr><td>α_{JX}</td><td>One way free space loss between SOJ transmitter and radar receiver</td></tr>
<tr><td>α_1</td><td>One way space loss between the radar and the target</td></tr>
</table></td>
</tr>
</table>

Support jamming adds a few geometric complexities. A SOJ platform usually uses high gain, directional antennas. Therefore, the jamming antenna must not only be pointed at the victim radar, but there must be alignment of radar, targets, and SOJ platform for the jamming to be most effective. Two cases will be described, main lobe-jamming and side-lobe jamming.

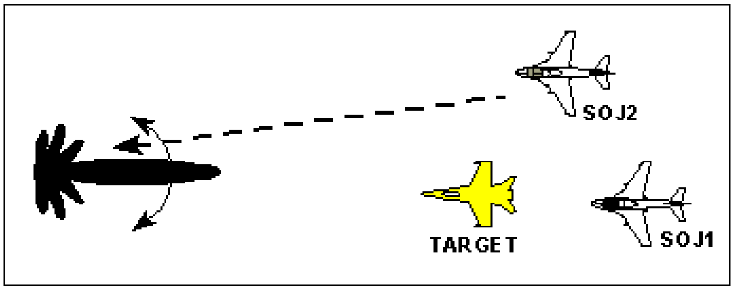

Figure 1. Radar Antenna Pattern.

Support jamming is usually applied against search and acquisition radars which continuously scan horizontally through a volume of space. The scan could cover a sector or a full 360°. The horizontal antenna pattern of the radar will exhibit a main lobe and side lobes as illustrated in Figure 1. The target is detected when the main lobe sweeps across it. For main lobe jamming, the SOJ platform and the target(s) must be aligned with the radar's main lobe as it sweeps the target(s).

For side lobe jamming, the SOJ platform may be aligned with one or more of the radar's side lobes when the main lobe sweeps the target. The gain of a radar's side lobes are many tens of dB less (usually more than 30 dB less) than the gain of the main lobe, so calculations of side lobe jamming must use the gain of the side lobe for the radar receive antenna gain, not the gain of the main lobe. Also, because many modern radars employ some form of side lobe blanking or side lobe cancellation, some knowledge of the victim radar is required to predict the effectiveness of side lobe jamming.

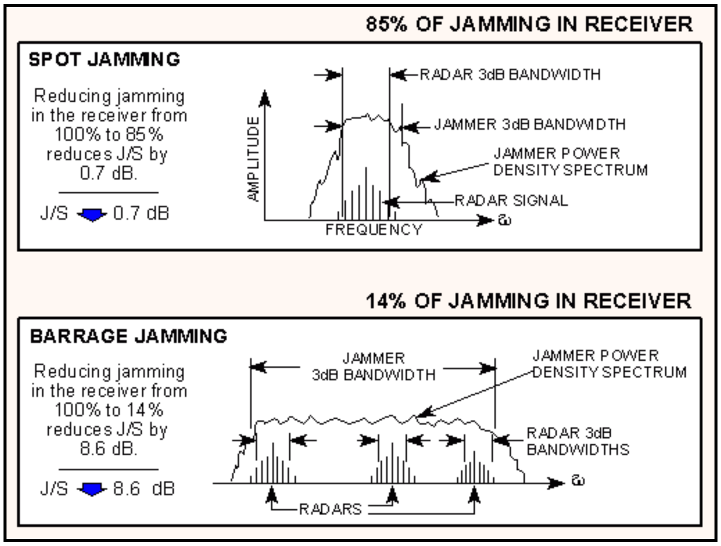
85% OF JAMMING IN RECEIVER
SPOT JAMMING
Reducing jamming in the receiver from 100% to 85% reduces J/S by 0.7 dB.
J/S ⬇ 0.7 dB
RADAR 3dB BANDWIDTH
JAMMER 3dB BANDWIDTH
JAMMER POWER DENSITY SPECTRUM
RADAR SIGNAL
AMPLITUDE
FREQUENCY
ω
14% OF JAMMING IN RECEIVER
BARRAGE JAMMING
Reducing jamming in the receiver from 100% to 14% reduces J/S by 8.6 dB.
J/S ⬇ 8.6 dB
JAMMER 3dB BANDWIDTH
JAMMER POWER DENSITY SPECTRUM
RADAR 3dB BANDWIDTHS
RADARS
ω

Figure 2. Noise Jamming.

All radar receivers are frequency selective. That is, they are filters that allow only a narrow range of frequencies into the receiver circuitry. Deceptive EA, by definition, creates forgeries of the real signal and, ideally, are as well matched to the radar receiver as the real signal. On the other hand, noise jamming probably **will not match** the radar receiver bandwidth characteristics. Noise jamming is either spot jamming or barrage jamming. As illustrated in Figure 2, spot jamming is simply narrowing the bandwidth of the noise jammer so that as much of the jammer power as possible is in the radar receiver bandwidth. Barrage jamming is using a wide noise bandwidth to cover several radars with one jammer or to compensate for any uncertainty in the radar frequency. In both cases some of the noise power is "wasted" because it is not in the radar receiver filter.

In the past, noise jammers were often described as having so many "watts per MHz." This is nothing more than the power of the noise jammer divided by the noise bandwidth. That is, a 500 watt noise jammer transmitting a noise bandwidth of 200 MHz has 2.5 watts/MHz. Older noise jammers often had noise bandwidths that were difficult, or impossible, to adjust accurately. These noise jammers usually used manual tuning to set the center frequency of the noise to the radar frequency. Modern noise jammers can set on the radar frequency quite accurately and the noise bandwidth is selectable, so the noise bandwidth is more a matter of choice than it used to be, and it is possible that all of the noise is placed in the victim radar's receiver.

If, in the example above, the 500 watt noise jammer were used against a radar that had a 3 MHz receiver bandwidth, the noise jammer power applicable to that radar would be:

$$3\ MHz\ x\ 2.5\ watts/MHz = 7.5\ watts\ \_\ 38.75\ dBm \qquad \textbf{[1]}$$

The calculation must be done as shown in equation [1] - multiply the watts/MHz by the radar bandwidth first and then convert to dBm. You can't convert to dBm/MHz and then multiply. (See derivation of dB in Section 2-4)

An alternate method for dB calculations is to use the bandwidth reduction factor (BF).

The BF is:
$$BF_{dB} = 10\ \text{Log}\left[\frac{BW_J}{BW_R}\right] \qquad \textbf{[2]}$$

where: BW_J is the bandwidth of the noise jammer, and BW_R is the bandwidth of the radar receiver.

The power of the jammer in the jamming equation (P_J) can be obtained by either method. If equation [1] is used then P_J is simply 38.75 dBm. If equation [2] is used then the jamming equation is written using (P_J - BF). All the following discussion uses the second method. Whichever method is used, it is required that $BW_J \geq BW_R$. If $BW_J < BW_R$, then all the available power is in the radar receiver and equation [1] does not apply and the BF = 0.

> Note: To avoid having to include additional terms for the following calculations, always combine any transmission line loss with antenna gain.

MAIN LOBE STAND-OFF / STAND-IN JAMMING

The equivalent circuit shown in Figure 3 applies to main lobe jamming by a stand-off support aircraft or a stand-in RPV. Since the jammer is not on the target aircraft, only two of the three ranges and two of the three space loss factors (α's) are the same. Figure 3 differs from the J/S monostatic equivalent circuit shown in Figure 4 in Section 4-7 in that the space loss from the jammer to the radar receiver is different.

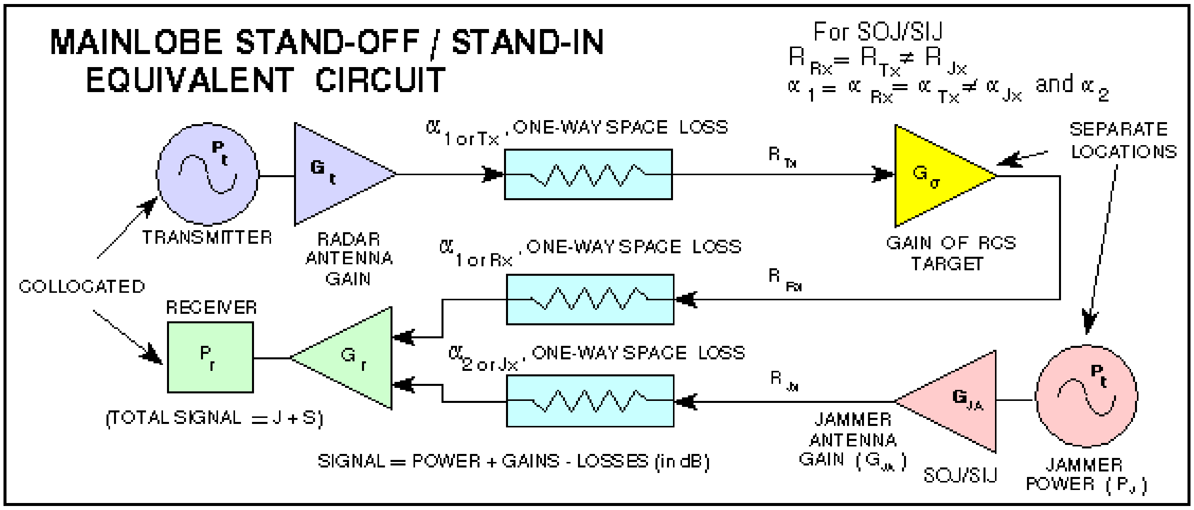
MAINLOBE STAND-OFF / STAND-IN EQUIVALENT CIRCUIT
For SOJ/SIJ
$R_{Rx} = R_{Tx} \neq R_{Jx}$
$\alpha_1 = \alpha_{Rx} = \alpha_{Tx} \neq \alpha_{Jx}$ and α_2
P_t
TRANSMITTER
G_t
RADAR ANTENNA GAIN
$\alpha_{1 or Tx}$, ONE-WAY SPACE LOSS
R_{Tx}
G_σ
GAIN OF RCS TARGET
SEPARATE LOCATIONS
$\alpha_{1 or Rx}$, ONE-WAY SPACE LOSS
R_{Rx}
COLLOCATED
RECEIVER
P_r
G_r
$\alpha_{2 or Jx}$, ONE-WAY SPACE LOSS
R_{Jx}
(TOTAL SIGNAL = J + S)
SIGNAL = POWER + GAINS - LOSSES (in dB)
G_{JA}
JAMMER ANTENNA GAIN (G_{JA})
SOJ/SIJ
P_t
JAMMER POWER (P_J)

Figure 3. Main Lobe Stand-Off / Stand-In EA Equivalent Circuit.

The equations are the same for both SOJ and SIJ. From the one way range equation in Section 4-3, and with inclusion of BF losses:

$$P_r 1 \text{ or } J = \frac{P_j G_{ja} G_r \lambda^2 BW_R}{(4\pi R_{Jx})^2 BW_J} \qquad [3]$$

From the two way range equation in Section 4.4: $P_r 2 \text{ or } S = \dfrac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R_{Tx}^4}$ **[4]**

so $$\frac{J}{S} = \frac{P_j G_{ja} G_r \lambda^2 (4\pi)^3 R_{Tx}^4 BW_R}{P_t G_t G_r \lambda^2 \sigma (4\pi R_{Jx})^2 BW_J} = \frac{P_j G_{ja} 4\pi R_{Tx}^4 BW_R}{P_t G_t \sigma R_{Jx}^2 BW_J} \qquad \textit{(ratio form)} \qquad [5]$$

Note: Keep R and σ in the same units. Converting to dB and using $10 \log 4\pi = 10.99$ dB:

$$10\log J/S = 10\log P_j - 10\log [BW_j/BW_R] + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10.99 \text{ dB} + 40\log R_{Tx} - 20\log R_{Jx} \quad \textbf{[6]}$$

If the simplified radar equation is used, the free space loss from the SOJ/SIJ to the radar receiver is α_{Jx}, then equation [7] is the same as monostatic equation [6] in Section 4-7 except α_{Jx} replaces α, and the bandwidth reduction factor [BF] losses are included:

$$10\log J = 10\log P_j - BF + 10\log G_{ja} + 10\log G_r - \alpha_{Jx} \qquad \text{(factors in dB)} \qquad \textbf{[7]}$$

Since the free space loss from the radar to the target and return is the same both ways, $\alpha_{Tx} = \alpha_{Rx} = \alpha_1$, equation [8] is the same as monostatic equation [7] in Section 4-7.

$$10\log S = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1 \qquad \text{(factors in dB)} \qquad \textbf{[8]}$$

and $$10\log J/S = 10\log P_j - BF + 10\log G_{ja} - \alpha_{Jx} - 10\log P_t - 10\log G_t - G_\sigma + 2\alpha_1 \qquad \text{(factors in dB)} \qquad \textbf{[9]}$$

Notice that unlike equation [8] in Section 4-7, there are two different α's in [9] because the signal paths are different.

SIDE LOBE STAND-OFF / STAND-IN JAMMING

The equivalent circuit shown in Figure 4. It differs from Figure 3, (main lobe SOJ/SIJ) in that the radar receiver antenna gain is different for the radar signal return and the jamming.

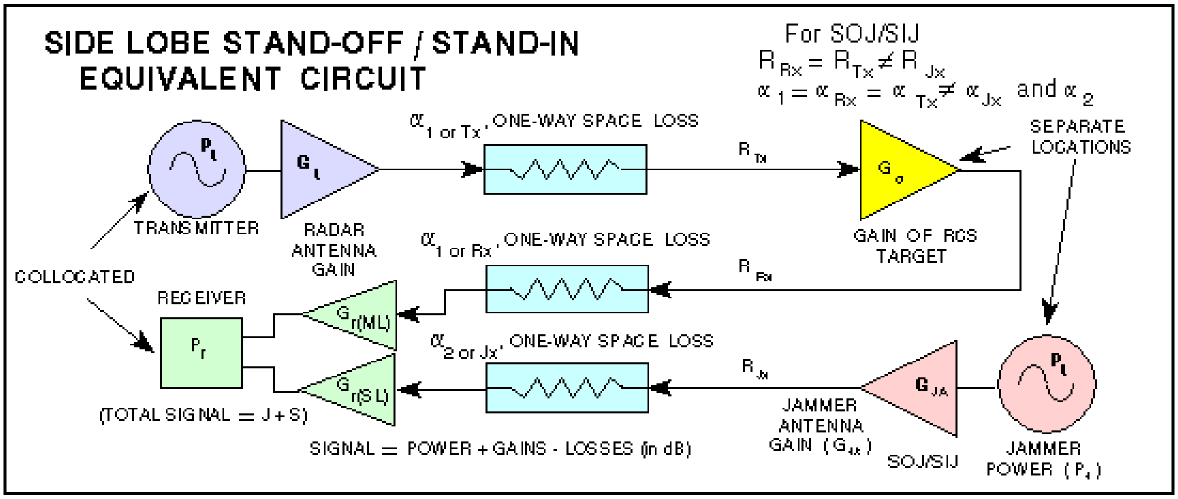
SIDE LOBE STAND-OFF / STAND-IN EQUIVALENT CIRCUIT
For SOJ/SIJ
$R_{Rx} = R_{Tx} \neq R_{Jx}$
$\alpha_1 = \alpha_{Rx} = \alpha_{Tx} \neq \alpha_{Jx}$ and α_2
$\alpha_{1 \text{ or } Tx}$, ONE-WAY SPACE LOSS
P_t
G_t
R_{Tx}
G_σ
SEPARATE LOCATIONS
TRANSMITTER
RADAR ANTENNA GAIN
$\alpha_{1 \text{ or } Rx}$, ONE-WAY SPACE LOSS
GAIN OF RCS TARGET
R_{Rx}
COLLOCATED
RECEIVER
$G_{r(ML)}$
P_r
$\alpha_{2 \text{ or } Jx}$, ONE-WAY SPACE LOSS
R_{Jx}
$G_{r(SL)}$
G_{JA}
P_t
(TOTAL SIGNAL = J + S)
JAMMER ANTENNA GAIN (G_{Jx})
SOJ/SIJ
JAMMER POWER (P_J)
SIGNAL = POWER + GAINS - LOSSES (in dB)

Figure 4. Side Lobe Stand-Off / Stand-In EA Equivalent Circuit.

To calculate side lobe jamming, the gain of the radar antenna's side lobes must be known or estimated. The gain of each side lobe will be different than the gain of the other side lobes. If the antenna is symmetrical, the first side lobe is the one on either side of the main lobe, the second side lobe is the next one on either side of the first side lobe, and so on. The side lobe gain is G_{SLn}, where the 'n' subscript denotes side lobe number: 1, 2, ..., n.

The signal is the same as main lobe equations [4] and [8], except $G_r = G_{r(ML)}$

$$P_r 2 \text{ or } S = \frac{P_t\, G_t\, G_{r(ML)}\, \lambda^2\, \sigma}{(4\pi)^3\, R_{Tx}^4} \quad \text{(ratio form)} \qquad \textbf{[10]}$$

If simplified radar equations are used:

$10\log S = 10\log P_t + 10\log G_t + 10\log G_{r(ML)} + G_\sigma - 2\alpha_1$ (factors in dB)

The jamming equation is the same as main lobe equations [3] and [7] except $G_r = G_{r(SL)}$:

$$J = \frac{P_j\, G_{ja}\, G_{r(SL)}\, \lambda^2\, BW_R}{(4\pi\, R_{Jx})^2\, BW_J} \qquad \textbf{[11]}$$

$10\log J = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - \alpha_{Jx}$ (factors in dB) **[12]**

so

$$\frac{J}{S} = \frac{P_j\, G_{ja}\, G_{r(SL)}\, 4\pi\, R_{Tx}^4\, BW_R}{P_t\, G_t\, G_{r(ML)}\, \sigma\, R_{Jx}^2\, BW_J} \quad \text{(ratio form)} \qquad \textbf{[13]}$$

Note: keep R and σ in same units. Converting to dB and using $10\log 4\pi = 10.99$ dB:

$10\log J/S = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - 10\log P_t - 10\log G_t - 10\log G_{r(ML)} - 10\log\sigma + 10.99 \text{ dB} + 40\log R_{Tx} - 20\log R_{Jx}$
(factors in dB) **[14]**

If simplified radar equations are used:

$10\log J/S = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - \alpha_{Jx} - 10\log P_t - 10\log G_t - 10\log G_{r(ML)} - G_\sigma + 2\alpha_1$ **[15]**
(in dB)

This page intentionally left blank.

JAMMING TO SIGNAL (J/S) RATIO - CONSTANT GAIN [LINEAR] JAMMING

JAMMING TO SIGNAL (J/S) RATIO (MONOSTATIC)

$$\frac{J}{S} = \frac{G_{ja(Rx)}\, G_j\, G_{ja(Tx)}\, \lambda^2}{4\pi\sigma} = \frac{G_{ja(Rx)}\, G_j\, G_{ja(Tx)}\, c^2}{4\pi\sigma\, f^2} \quad \text{(ratio form)}$$

$G_{ja(Rx)}$ = The Gain of the jammer receive antenna
G_j = The gain of the jammer
$G_{ja(Tx)}$ = The Gain of the jammer transmit antenna

or:

$$10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - 10\log (4\pi\sigma/\lambda^2)$$

or if simplified radar equations developed in previous sections are used:

$$10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - G_\sigma \quad \text{(dB)}$$

* Keep λ and σ in same units. Note: $\lambda = c/f$

Target gain factor,
$G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (dB)

K_2 Values (dB):

RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
m^2	-38.54	21.46
ft^2	-48.86	11.14

JAMMING TO SIGNAL (J/S) RATIO (BISTATIC)

Same as the monostatic case except G_σ will be different since RCS (σ) varies with aspect angle.

Since the jammer on the target is amplifying the received radar signal before transmitting it back to the radar, both J and S experience the two way range loss. Figure 1 shows that the range for both the signal and constant gain jamming have a slope that is 40 dB per decade. Once the jammer output reaches maximum power, that power is constant and the jamming slope changes to 20 dB per decade since it is only a function of one way space loss and the J/S equations for constant power (saturated) jamming must be used.

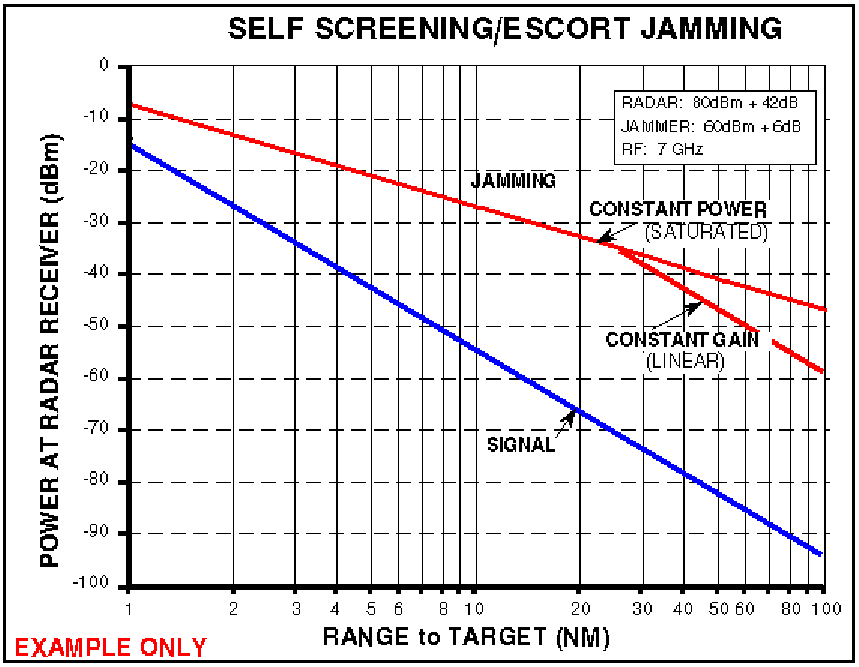

Figure 1. Sample Constant Gain / Constant Power Graph.

Normally the constant gain (linear) region of a repeater jammer occurs only at large distances from the radar and the constant power (saturated) region is reached rapidly as the target approaches the radar. When a constant gain jammer is involved it may be necessary to plot jamming twice - once using J from the constant power (saturated) equation [1] in Section 4-7 and once using the constant gain (linear) equation [4], as in the example shown in Figure 1.

CONSTANT GAIN SELF PROTECTION EA

Most jammers have a constant power output - that is, they always transmit the maximum available power of the transmitter (excepting desired EA modulation). Some jammers also have a constant gain (linear) region. Usually these are coherent repeaters that can amplify a low level radar signal to a power that is below the level that results in maximum available (saturated) power output. At some radar to target range, the input signal is sufficiently high that the full jammer gain would exceed the maximum available power and the jammer ceases to be constant gain and becomes constant power.

To calculate the power output of a constant gain jammer where:

S_{Rj} = The Radar signal at the jammer input (receive antenna terminals)
$G_{ja(Rx)}$ = The Gain of the jammer receive antenna
G_j = The gain of the jammer
α_{Tx} = The one-way free space loss from the radar to the target
P_{jCG} = The jammer constant gain power output
P_j = The maximum jammer power output
L_R = The jammer receiving line loss; combine with antenna gain $G_{ja(Rx)}$

From equation [10], Section 4-3, calculate the radar power received by the jammer.

$$10\log S_{Rj} = 10\log P_t + 10\log G_t - \alpha_{Tx} + 10\log G_{ja(Rx)} \qquad \text{(factors in dB)} \qquad \textbf{[1]}$$

The jammer constant gain power output is: $$10\log P_{jCG} = 10\log S_{Rj} + 10\log G_{ja} \qquad \textbf{[2]}$$

and, by definition: $$P_{jCG} \le P_j \qquad \textbf{[3]}$$

MONOSTATIC

The equivalent circuit shown in Figure 2 is different from the constant power equivalent circuit in Figure 4 in Section 4-7. With constant gain, the jamming signal experiences the gain of the jammer and its antennas plus the same space loss as the radar signal.

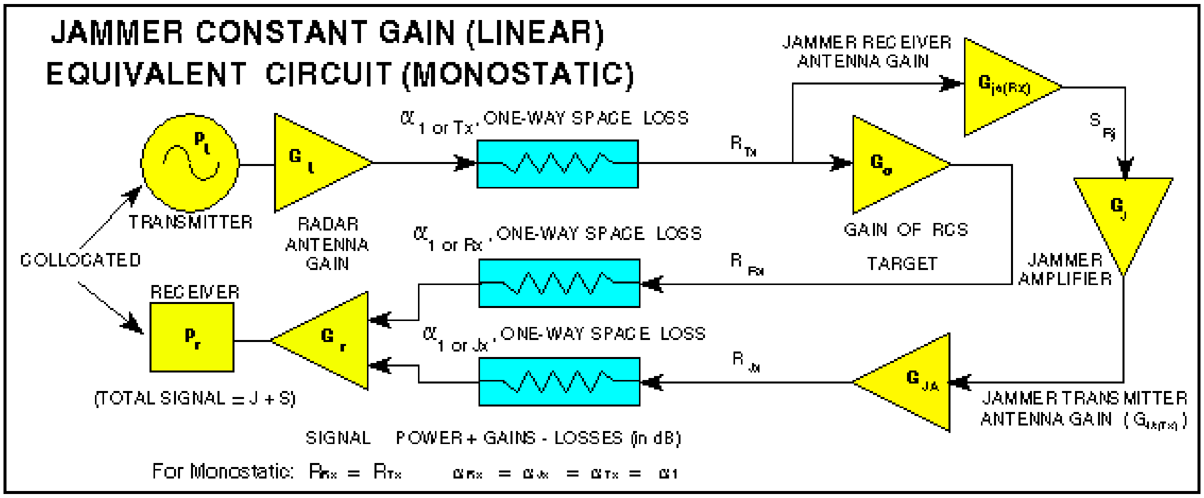
JAMMER CONSTANT GAIN (LINEAR)
EQUIVALENT CIRCUIT (MONOSTATIC)
JAMMER RECEIVER ANTENNA GAIN
$G_{ja(Rx)}$
$\alpha_{1 \text{ or } Tx}$, ONE-WAY SPACE LOSS
R_{Tx}
S_{Rj}
P_t
G_t
G_σ
G_j
TRANSMITTER
RADAR ANTENNA GAIN
$\alpha_{1 \text{ or } Rx}$, ONE-WAY SPACE LOSS
GAIN OF RCS
TARGET
R_{Rx}
JAMMER AMPLIFIER
COLLOCATED
RECEIVER
P_r
G_r
$\alpha_{1 \text{ or } Jx}$, ONE-WAY SPACE LOSS
R_{Jx}
G_{JA}
(TOTAL SIGNAL = J + S)
JAMMER TRANSMITTER ANTENNA GAIN ($G_{ja(Tx)}$)
SIGNAL POWER + GAINS - LOSSES (in dB)
For Monostatic: $R_{Rx} = R_{Tx}$ $\alpha_{Rx} = \alpha_{Jx} = \alpha_{Tx} = \alpha_1$

Figure 2. Jammer Constant Gain EA Equivalent Circuit (Monostatic).

To calculate J, the one way range equation from Section 4-3 is used twice:

$$J = \frac{P_t\, G_t\, G_{ja(Rx)}\, \lambda^2}{(4\pi R)^2} \frac{G_j\, G_{ja(Tx)}\, G_r\, \lambda^2}{(4\pi R)^2} \qquad \textbf{[4]}$$

From the two way range equation in Section 4-4: $S = \frac{P_t\, G_t\, G_r\, \lambda^2\, \sigma}{(4\pi)^3\, R^4}$ **[5]**

Terms cancel when combined: $\frac{J}{S} = \frac{G_{ja(Rx)}\, G_j\, G_{ja(Tx)}\, \lambda^2}{4\pi\sigma}$ *Keep λ and σ in same units* **[6]**

Or in dB form: $10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - 10\log(4\pi\sigma/\lambda^2)$ **[7]**

Since the last term can be recognized as minus G_σ from equation [10] in Section 4-4, where the target gain factor, $G_\sigma = 10\log(4\pi\sigma/\lambda^2) = 10\log(4\pi\sigma f^2/c^2)$, it follows that:

$10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - G_\sigma$ (factors in dB) **[8]**

Target gain factor, $G_\sigma = 10\log\sigma + 20\log f_1 + K_2$ (in dB)

K_2 Values

(dB)	RCS (σ) (units)	f_1 in MHz $K_2 =$	f_1 in GHz $K_2 =$
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

BISTATIC

The bistatic equivalent circuit shown in Figure 3 is different from the monostatic equivalent circuit shown in Figure 2 in that the receiver is separately located from the transmitter, $R_{Tx} \neq R_{Rx}$ or R_{Jx} and G_σ will be different since the RCS (σ) varies with aspect angle.

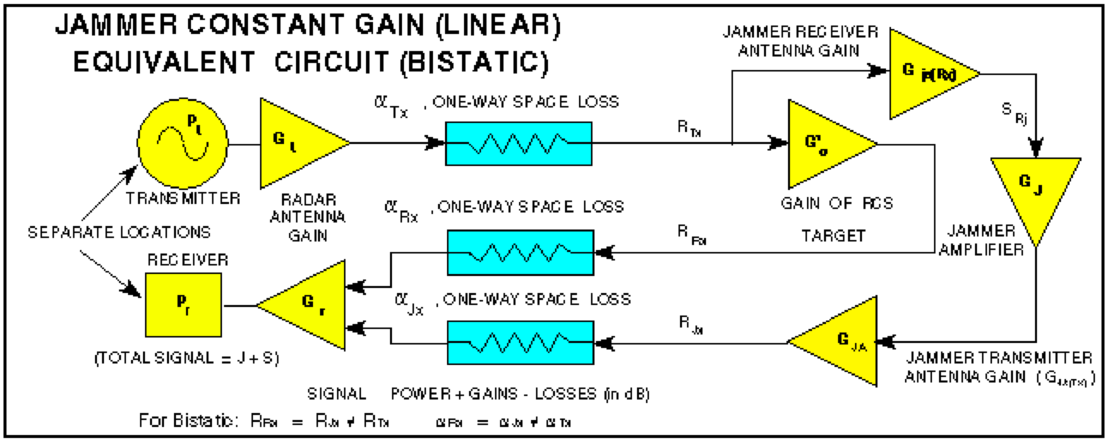
JAMMER CONSTANT GAIN (LINEAR)
EQUIVALENT CIRCUIT (BISTATIC)
JAMMER RECEIVER ANTENNA GAIN
$G_{ja(Rx)}$
α_{Tx} , ONE-WAY SPACE LOSS
P_t
G_t
R_{Tx}
G'_σ
S_{Rj}
G_J
TRANSMITTER
RADAR ANTENNA GAIN
α_{Rx} , ONE-WAY SPACE LOSS
GAIN OF RCS
TARGET
SEPARATE LOCATIONS
R_{Rx}
JAMMER AMPLIFIER
RECEIVER
P_r
G_r
α_{Jx} , ONE-WAY SPACE LOSS
R_{Jx}
G_{JA}
(TOTAL SIGNAL = J + S)
JAMMER TRANSMITTER ANTENNA GAIN ($G_{ja(Tx)}$)
SIGNAL POWER + GAINS - LOSSES (in dB)
For Bistatic: $R_{Rx} = R_{Jx} \neq R_{Tx}$ $\alpha_{Rx} = \alpha_{Jx} \neq \alpha_{Tx}$

Figure 3. Jammer Constant Gain EA Equivalent Circuit (Bistatic).

To calculate J, the one way range equation from Section 4-3 is used twice:

$$J = \frac{P_t\, G_t\, G_{ja(Rx)}\, \lambda^2}{(4\pi\, R_{Tx})^2} \frac{G_j\, G_{ja(Tx)}\, G_r\, \lambda^2}{(4\pi\, R_{Rx})^2} \qquad (R_{Jx} = R_{Rx}) \qquad \textbf{[9]}$$

From the two way range equation in Section 4-4: $S = \frac{P_t\, G_t\, G_r\, \lambda^2\, \sigma'}{(4\pi)^3\, R_{Tx}^2\, R_{Rx}^2}$ (σ' is bistatic RCS) **[10]**

Terms cancel when combined: $\frac{J}{S} = \frac{G_{ja(Rx)}\, G_j\, G_{ja(Tx)}\, \lambda^2}{4\pi\sigma'}$ *Keep λ and σ in same units* **[11]**

Or in dB form: $10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - 10\log (4\pi\sigma'/\lambda^2)$ **[12]**

Since the last term can be recognized as minus G_σ from equation [10] in Section 4-4, where the target gain factor, $G_\sigma = 10\log (4\pi\sigma'/\lambda^2) = 10\log (4\pi\sigma' f^2/c^2)$, it follows that:

$10\log = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - G_\sigma'$ (factors in dB) **[13]**

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)			
K_2 Values (dB)	RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

Linear J/S (Monostatic) Example (Linear Power Jamming)

Assume that a 5 GHz radar has a 70 dBm signal fed through a 5 dB loss transmission line to an antenna that has 45 dB gain. An aircraft that is flying 31 km from the radar has an aft EW antenna with -1 dB gain and a 5 dB line loss to the EW receiver (there is an additional loss due to any antenna polarization mismatch but that loss will not be addressed in this problem). The received signal is fed to a jammer with a gain of 60 dB, feeding a 10 dB loss transmission line which is connected to an antenna with 5 dB gain.

If the RCS of the aircraft is 9 m^2, what is the J/S level received at the input to the receiver of the tracking radar?

Answer:

$10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - G_\sigma$

$G_\sigma = 10\log \sigma + 20\log f_1 + K_2 = 10\log 9 + 20\log 5 + 21.46 = 44.98$ dB

Note: The respective transmission line losses will be combined with antenna gains, i.e. -1 -5 = -6 dB and -10 + 5 = -5 dB

$10\log J/S = -6 + 60 - 5 - 44.98 = 4.02$ dB @ 5 GHz

The answer changes to 1.1 dB if the tracking radar operates at 7 GHz provided the antenna gains and aircraft RCS are the same at both 5 and 7 GHz.

$G_\sigma = 10\log 9 + 20\log 7 + 21.46 = 47.9$ dB

$10\log J/S = -6 + 60 - 5 - 47.9 = 1.1$ dB @ 7 GHz

Separate J (-73.5 dBm @ 5 GHz and -79.34 dBm @ 7 GHz) and S (-77.52 dBm @ 5 GHz and -80.44 dBm @ 7 GHz) calculations for this problem are provided in Sections 4-3 and 4-4, respectively. A saturated gain version of this problem is provided in Section 4-7.

This page intentionally left blank.

RADAR CROSS SECTION (RCS)

Radar cross section is the measure of a target's ability to reflect radar signals in the direction of the radar receiver, i.e. it is a measure of the ratio of backscatter power per steradian (unit solid angle) in the direction of the radar (from the target) to the power density that is intercepted by the target.

The RCS of a target can be viewed as a comparison of the strength of the reflected signal from a target to the reflected signal from a perfectly smooth sphere of cross sectional area of 1 m^2 as shown in Figure 1.

The conceptual definition of RCS includes the fact that not all of the radiated energy falls on the target. A target's RCS (σ) is most easily visualized as the product of three factors:

σ = Projected cross section x Reflectivity x Directivity.

RCS(σ) is used in Section 4-4 for an equation representing power reradiated from the target.

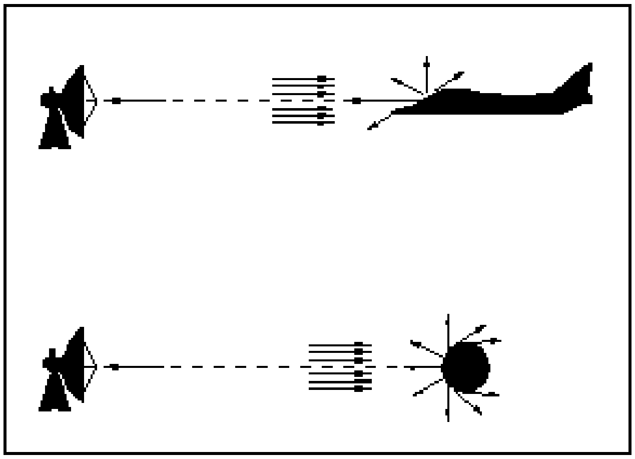

Figure 1. Concept of Radar Cross Section.

Reflectivity: The percent of intercepted power reradiated (scattered) by the target.

Directivity: The ratio of the power scattered back in the radar's direction to the power that would have been backscattered had the scattering been uniform in all directions (i.e. isotropically).

Figures 2 and 3 show that RCS does not equal geometric area. For a sphere, the RCS, **$\sigma = \pi r^2$**, where r is the radius of the sphere.

The RCS of a sphere is independent of frequency if operating at sufficiently high frequencies where λ<<Range, and λ<< radius (r). Experimentally, radar return reflected from a target is compared to the radar return reflected from a sphere which has a frontal or projected area of one square meter (i.e. diameter of about 44 in). Using the spherical shape aids in field or laboratory measurements since orientation or positioning of the sphere will not affect radar reflection intensity measurements as a flat plate would. If calibrated, other sources (cylinder, flat plate, or corner reflector, etc.) could be used for comparative measurements.

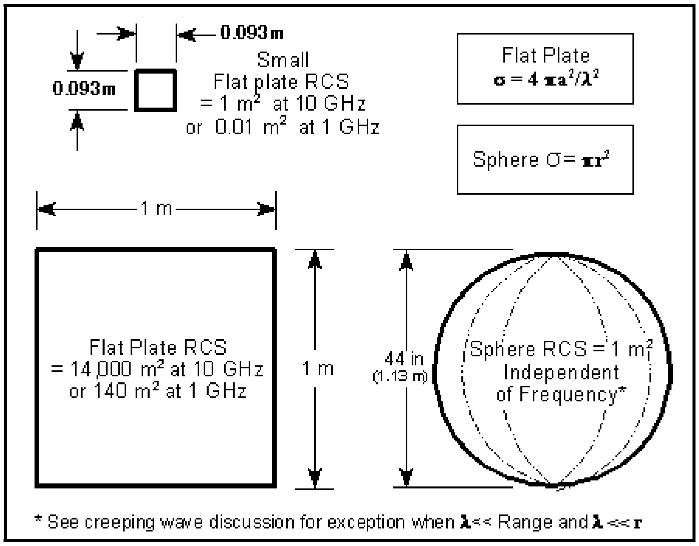
0.093m
0.093m
Small
Flat plate RCS
= 1 m^2 at 10 GHz
or 0.01 m^2 at 1 GHz
Flat Plate
$\sigma = 4\pi a^2/\lambda^2$
Sphere $\sigma = \pi r^2$
1 m
Flat Plate RCS
= 14,000 m^2 at 10 GHz
or 140 m^2 at 1 GHz
1 m
44 in
(1.13 m)
Sphere RCS = 1 m^2
Independent
of Frequency*
* See creeping wave discussion for exception when λ<< Range and λ << r

Figure 2. RCS vs. Physical Geometry.

To reduce drag during tests, towed spheres of 6", 14", or 22" diameter may be used instead of the larger 44" sphere, and the reference size is 0.018, 0.099, or 0.245 m^2 respectively instead of 1 m^2. When smaller sized spheres are used for tests you may be operating at or near where λ~radius. If the results are then scaled to a 1 m^2 reference, there may be some perturbations due to creeping waves. See the discussion at the end of this section for further details.

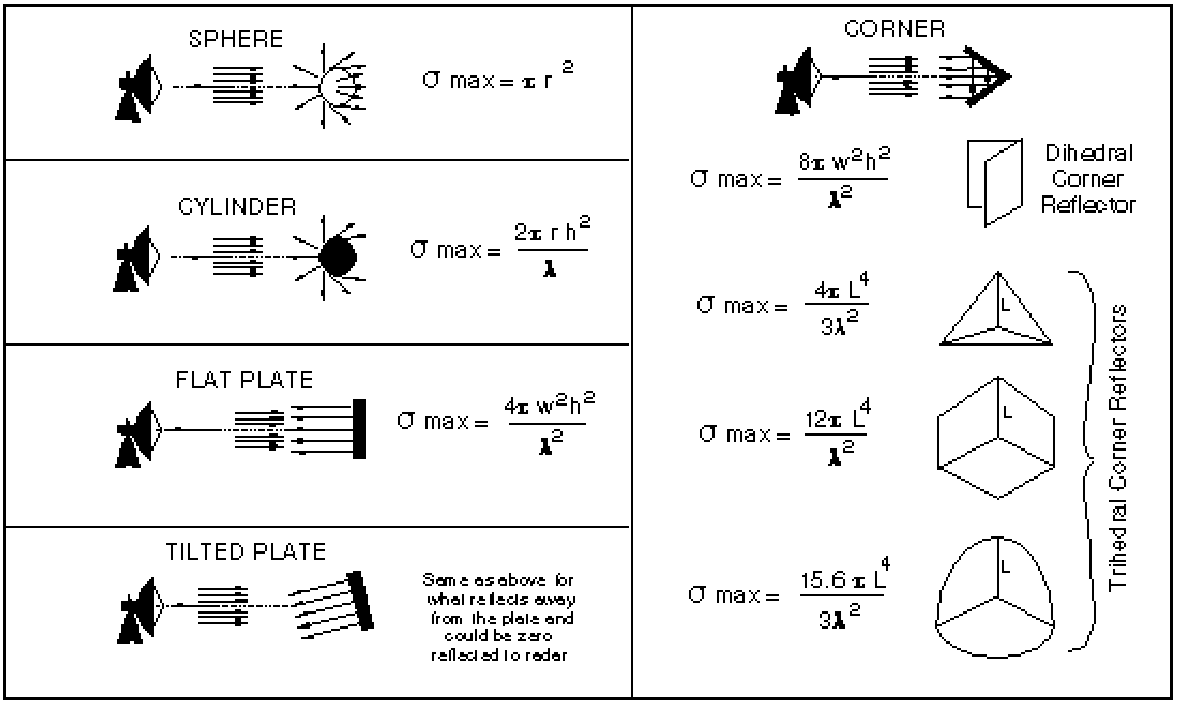
SPHERE
$\sigma\ max = \pi r^2$
CYLINDER
$\sigma\ max = \frac{2\pi r h^2}{\lambda}$
FLAT PLATE
$\sigma\ max = \frac{4\pi w^2 h^2}{\lambda^2}$
TILTED PLATE
Same as above for what reflects away from the plate and could be zero reflected to radar
CORNER
$\sigma\ max = \frac{8\pi w^2 h^2}{\lambda^2}$
Dihedral Corner Reflector
$\sigma\ max = \frac{4\pi L^4}{3\lambda^2}$
$\sigma\ max = \frac{12\pi L^4}{\lambda^2}$
$\sigma\ max = \frac{15.6\pi L^4}{3\lambda^2}$
L
Trihedral Corner Reflectors

Figure 3. Backscatter From Shapes.

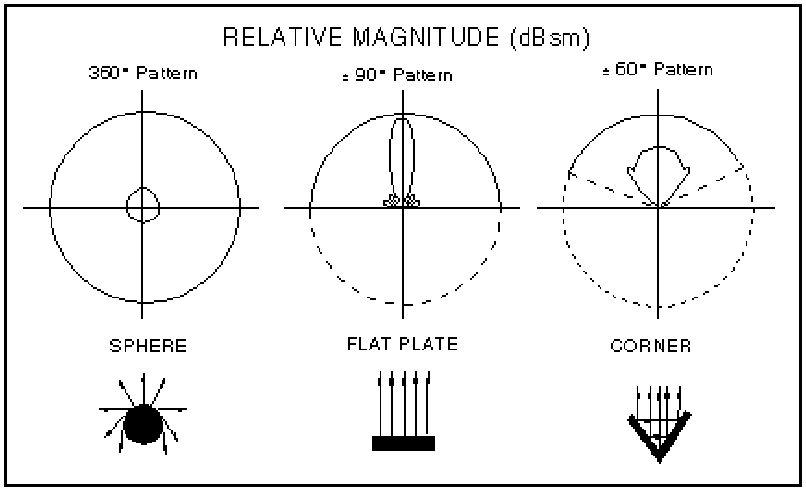
RELATIVE MAGNITUDE (dBsm)
360° Pattern
± 90° Pattern
± 60° Pattern
SPHERE
FLAT PLATE
CORNER

Figure 4. RCS Patterns.

In Figure 4, RCS patterns are shown as objects are rotated about their vertical axes (the arrows indicate the direction of the radar reflections).

The sphere is essentially the same in all directions.

The flat plate has almost no RCS except when aligned directly toward the radar.

The corner reflector has an RCS almost as high as the flat plate but over a wider angle, i.e., over ±60°. The return from a corner reflector is analogous to that of a flat plate always being perpendicular to your collocated transmitter and receiver.

Targets such as ships and aircraft often have many effective corners. Corners are sometimes used as calibration targets or as decoys, i.e. corner reflectors.

An aircraft target is very complex. It has a great many reflecting elements and shapes. The RCS of real aircraft must be measured. It varies significantly depending upon the direction of the illuminating radar.

Figure 5 shows a typical RCS plot of a jet aircraft. The plot is an azimuth cut made at zero degrees elevation (on the aircraft horizon). Within the normal radar range of 3-18 GHz, the radar return of an aircraft in a given direction will vary by a few dB as frequency and polarization vary (the RCS may change by a factor of 2-5). It does not vary as much as the flat plate.

As shown in Figure 5, the RCS is highest at the aircraft beam due to the large physical area observed by the radar and perpendicular aspect (increasing reflectivity). The next highest RCS area is the nose/tail area, largely because of reflections off the engines or propellers. Most self-protection jammers cover a field of view of +/- 60 degrees about the aircraft nose and tail, thus the high RCS on the beam does not have coverage. Beam coverage is frequently not provided due to inadequate power available to cover all aircraft quadrants, and the side of an aircraft is theoretically exposed to a threat 30% of the time over the average of all scenarios.

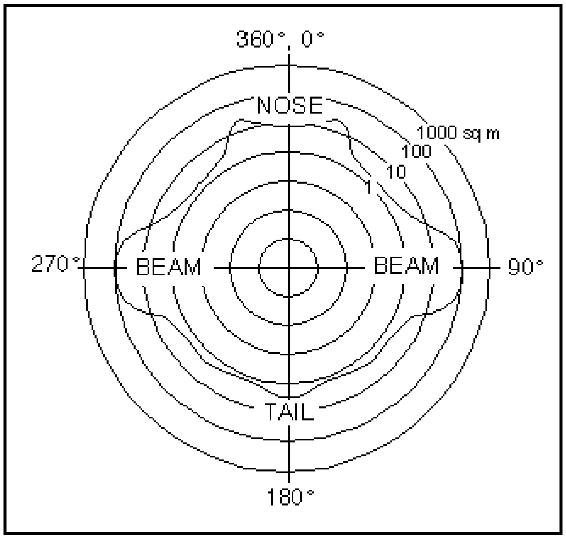

Figure 5. Typical Aircraft RCS.

Typical radar cross sections are as follows: Missile 0.5 sq m; Tactical Jet 5 to 100 sq m; Bomber 10 to 1000 sq m; and ships 3,000 to 1,000,000 sq m. RCS can also be expressed in decibels referenced to a square meter (dBsm) which equals 10 log (RCS in m^2).

Again, Figure 5 shows that these values can vary dramatically. The strongest return depicted in the example is 100 m^2 in the beam, and the weakest is slightly more than 1 m^2 in the 135°/225° positions. These RCS values can be very misleading because other factors may affect the results. For example, phase differences, polarization, surface imperfections, and material type all greatly affect the results. In the above typical bomber example, the measured RCS may be much greater than 1000 square meters in certain circumstances (90°, 270°).

SIGNIFICANCE OF THE REDUCTION OF RCS

If each of the range or power equations that have an RCS (σ) term is evaluated for the significance of decreasing RCS, Figure 6 results. Therefore, an RCS reduction can increase aircraft survivability. The equations used in Figure 6 are as follows:

Range (radar detection):
From the 2-way range equation in Section 4-4:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4}$$

Therefore, $R^4 \propto \sigma$ or $\sigma^{1/4} \propto R$

Range (radar burn-through):
The crossover equation in Section 4-8 has:

$$R_{BT}^2 = \frac{P_t G_t \sigma}{P_j G_j 4\pi}$$

Therefore, $R_{BT}^2 \propto \sigma$ or $\sigma^{1/2} \propto R_{BT}$

Power (jammer): Equating the received signal return (P_r) in the two way range equation to the received jammer signal (P_r) in the one way range equation, the following relationship results:

$$P_r = \underset{\displaystyle S}{\underset{\uparrow}{\frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4}}} = \underset{\displaystyle J}{\underset{\uparrow}{\frac{P_j G_j G_r \lambda^2}{(4\pi R)^2}}}$$

Therefore, $P_j \propto \sigma$ or $\sigma \propto P_j$

Note: jammer transmission line loss is combined with the jammer antenna gain to obtain G_t.

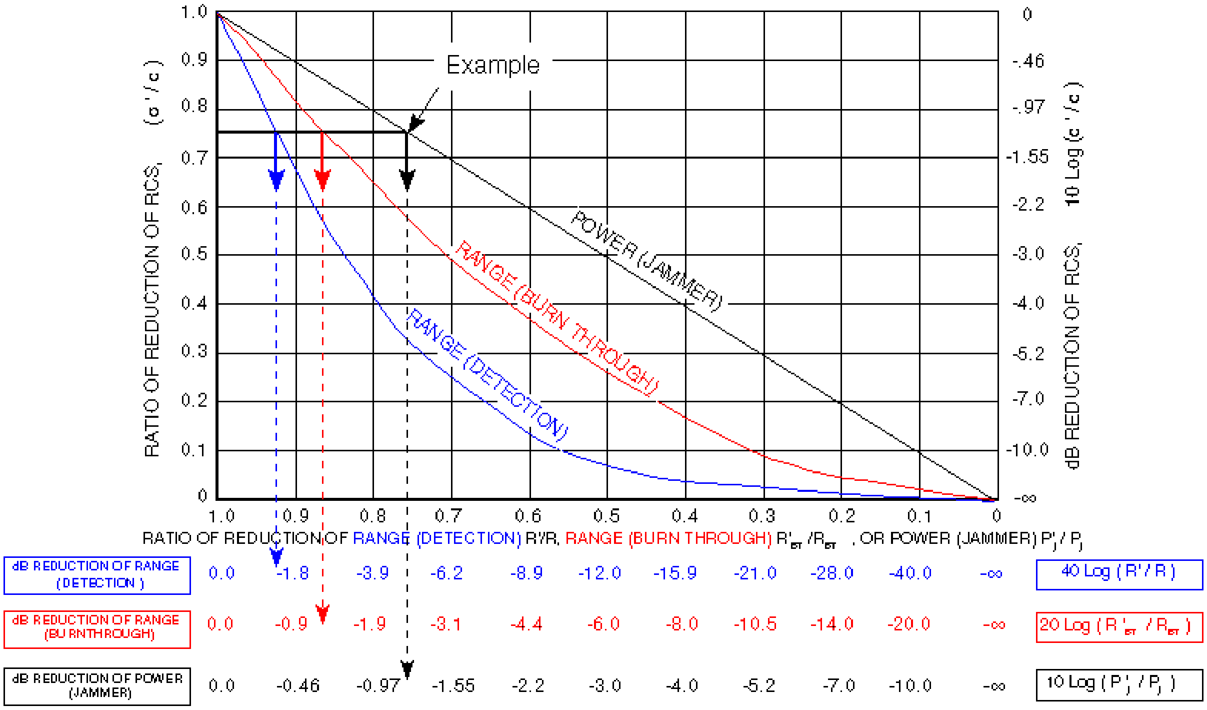

dB REDUCTION OF RANGE (DETECTION)	0.0	-1.8	-3.9	-6.2	-8.9	-12.0	-15.9	-21.0	-28.0	-40.0	-∞	40 Log (R' / R)
dB REDUCTION OF RANGE (BURNTHROUGH)	0.0	-0.9	-1.9	-3.1	-4.4	-6.0	-8.0	-10.5	-14.0	-20.0	-∞	20 Log (R'$_{BT}$ / R$_{BT}$)
dB REDUCTION OF POWER (JAMMER)	0.0	-0.46	-0.97	-1.55	-2.2	-3.0	-4.0	-5.2	-7.0	-10.0	-∞	10 Log (P'$_j$ / P$_j$)

Figure 6. Reduction of RCS Affects Radar Detection, Burn-through, and Jammer Power.

Example of Effects of RCS Reduction - As shown in Figure 6, if the RCS of an aircraft is reduced to 0.75 (75%) of its original value, then (1) the jammer power required to achieve the same effectiveness would be 0.75 (75%) of the original value (or -1.25 dB). Likewise, (2) If Jammer power is held constant, then burn-through range is 0.87 (87%) of its original value (-1.25 dB), and (3) the detection range of the radar for the smaller RCS target (jamming not considered) is 0.93 (93%) of its original value (-1.25 dB).

OPTICAL / MIE / RAYLEIGH REGIONS

Figure 7 shows the different regions applicable for computing the RCS of a sphere. The optical region ("far field" counterpart) rules apply when $2\pi r/\lambda > 10$. In this region, the RCS of a sphere is independent of frequency. Here, the RCS of a sphere, $\sigma = \pi r^2$. The RCS equation breaks down primarily due to creeping waves in the area where λ~2πr. This area is known as the Mie or resonance region. If we were using a 6” diameter sphere, this frequency would be 0.6 GHz. (Any frequency ten times higher, or above 6 GHz, would give expected results). The largest positive perturbation (point A) occurs at exactly 0.6 GHz where the RCS would be 4 times higher than the RCS computed using the optical region formula. Just slightly above 0.6 GHz a minimum occurs (point B) and the actual RCS would be 0.26 times the value calculated by using the optical region formula. If we used a one meter diameter sphere, the perturbations would occur at 95 MHz, so any frequency above 950 MHz (~1 GHz) would give predicted results.

CREEPING WAVES

The initial RCS assumptions presume that we are operating in the optical region (λ<<Range and λ<<radius). There is a region where specular reflected (mirrored) waves combine with back scattered creeping waves both constructively and destructively as shown in Figure 8. Creeping waves are tangential to a smooth surface and follow the “shadow” region of the body. They occur when the circumference of the sphere ~ λ and typically add about 1 m^2 to the RCS at certain frequencies.

RAYLEIGH REGION
$\sigma = [\pi r^2][7.11(kr)^4]$

where: $k = 2\pi/\lambda$

MIE (resonance)
$\sigma = 4\pi r^2$ at Maximum (point A)
$\sigma = 0.26\pi r^2$ at Minimum (pt B)

OPTICAL REGION
$\sigma = \pi r^2$
(Region RCS of a sphere is independent of frequency)

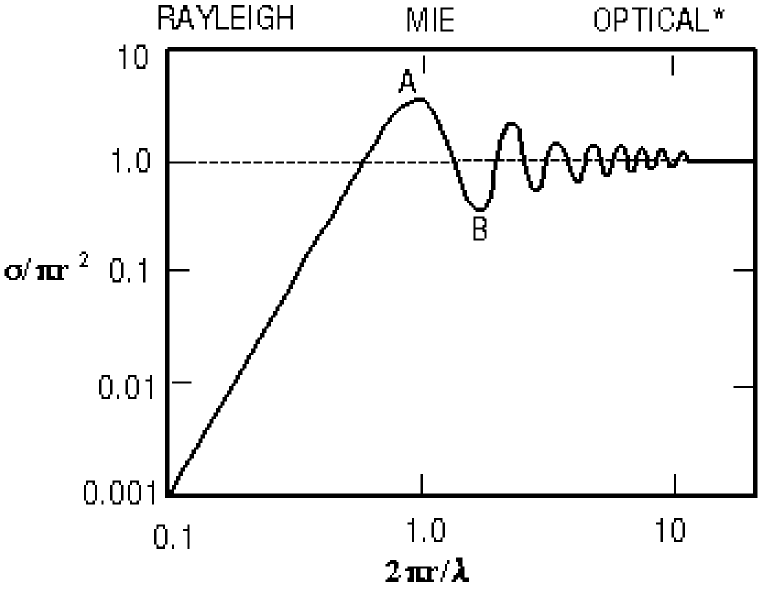

* "RF far field" equivalent
Courtesy of Dr. Allen E. Fuhs, Ph.D.

Figure 7. Radar Cross Section of a Sphere.

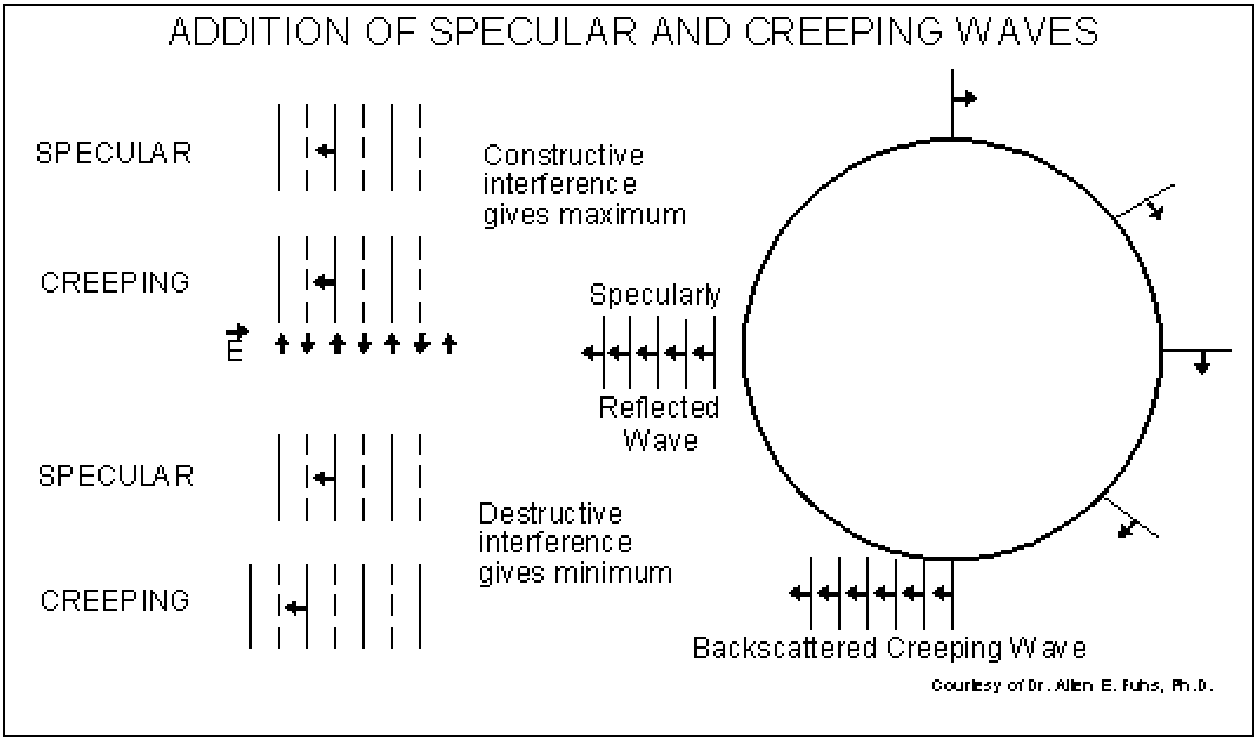
ADDITION OF SPECULAR AND CREEPING WAVES
SPECULAR
CREEPING
E
Constructive interference gives maximum
SPECULAR
CREEPING
Destructive interference gives minimum
Specularly
Reflected Wave
Backscattered Creeping Wave
Courtesy of Dr. Allen E. Fuhs, Ph.D.

Figure 8. Addition of Specular and Creeping Waves.

EMISSION CONTROL (EMCON)

When EMCON is imposed, RF emissions must not exceed -110 dBm/meter^2 at one nautical mile. It is best if systems meet EMCON when in either the Standby or Receive mode versus just the Standby mode (or OFF). If one assumes antenna gain equals line loss, then emissions measured at the port of a system must not exceed -34 dBm (i.e. the stated requirement at one nautical mile is converted to a measurement at the antenna of a point source - see Figure 1). If antenna gain is greater than line loss (i.e. gain 6 dB, line loss 3 dB), then the -34 dBm value would be lowered by the difference and would be -37 dBm for the example. The opposite would be true if antenna gain is less.

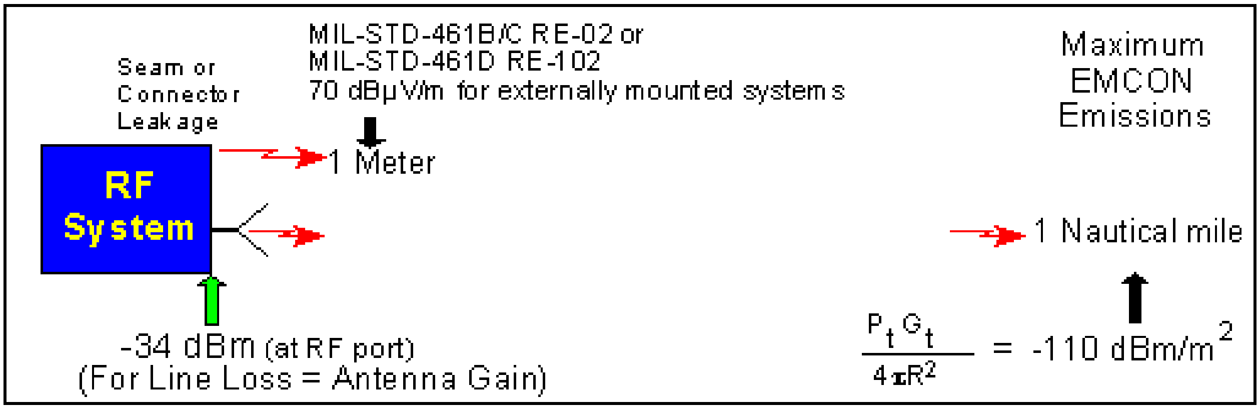
Seam or
Connector
Leakage
MIL-STD-461B/C RE-02 or
MIL-STD-461D RE-102
70 dBμV/m for externally mounted systems
1 Meter
Maximum
EMCON
Emissions
RF
System
1 Nautical mile
-34 dBm (at RF port)
(For Line Loss = Antenna Gain)
$\frac{P_t G_t}{4\pi R^2} = -110\ \text{dBm/m}^2$

Figure 1. EMCON Field Intensity / Power Density Measurements.

To compute the strength of emissions at the antenna port in Figure 1, we use the power density equation (see Section 4-2)

$$P_D = \frac{P_t G_t}{4\pi R^2} \quad \textbf{[1]} \qquad \text{or rearranging} \qquad P_t G_t = P_D\,(4\pi R^2) \quad \textbf{[2]}$$

Given that $P_D = -110\ \text{dBm/m}^2 = (10)^{-11}\ \text{mW/m}^2$, and R = 1 NM = 1852 meters.

$$P_t G_t = P_D\,(4\pi R^2) = (10^{-11}\text{mW/m}^2)(4\pi)(1852\text{m})^2 = 4.31(10)^{-4}\ \text{mW} = -33.65 \approx -34\ \text{dBm}$$

at the RF system antenna as given.

or, the equation can be rewritten in Log form and each term multiplied by 10:

$$10\log P_t + 10\log G_t = 10\log P_D + 10\log\,(4\pi R^2) \quad \textbf{[3]}$$

Since the m^2 terms on the right side of equation [3] cancel, then:

$$10\log P_t + 10\log G_t = -110\ \text{dBm} + 76.35\ \text{dB} = -33.65\ \text{dBm} \approx -34\ \text{dBm as given in Figure 1.}$$

If MIL-STD-461B/C RE02 (or MIL-STD-461D RE-102) measurements (see Figure 2) are made on seam/connector leakage of a system, emissions below 70 dBμV/meter which are measured at one meter will meet the EMCON requirement. Note that the airframe provides attenuation so portions of systems mounted inside an aircraft that measure 90 dBμV/meter will still meet EMCON if the airframe provides 20 dB of shielding (note that the requirement at one nm is converted to what would be measured at one meter from a point source).

The narrowband emission limit shown in Figure 2 for RE02/RE102 primarily reflect special concern for local oscillator leakage during EMCON as opposed to switching transients which would apply more to the broadband limit.

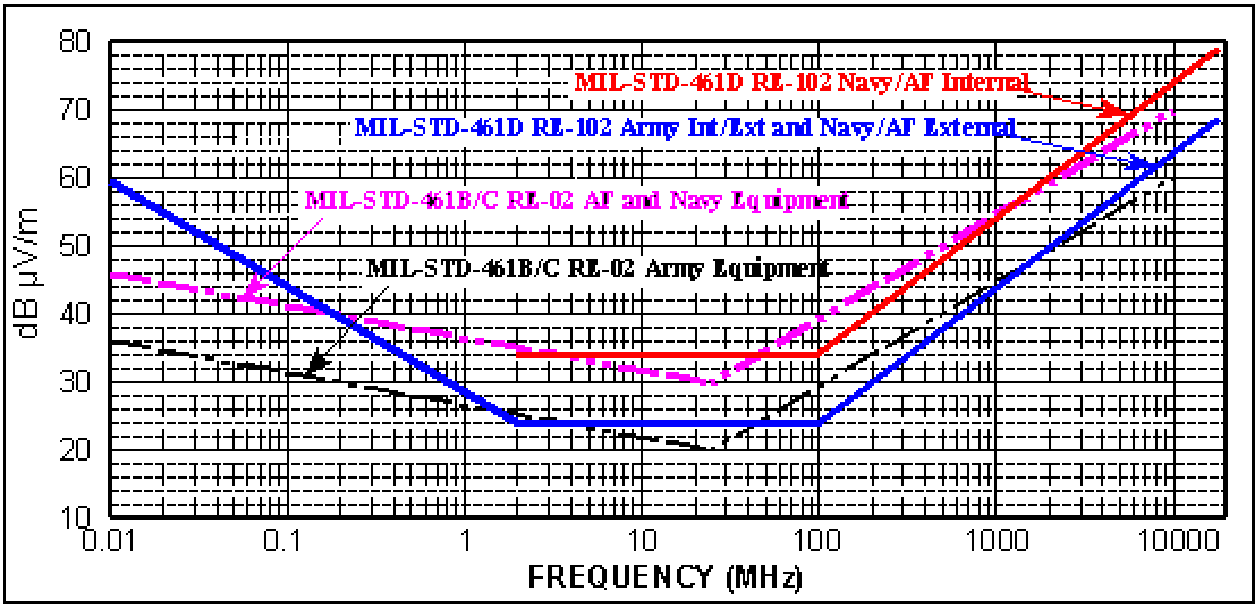

Figure 3. MIL-STD-461 Narrowband Radiated Emissions Limits.

Note that in MIL-STD-461D, the narrowband radiated emissions limits were retitled RE-102 from the previous RE-02 and the upper frequency limit was raised from 10 GHz to 18 GHz. The majority of this section will continue to reference RE02 since most systems in use today were built to MIL-STD-461B/C.

For the other calculation involving leakage (to obtain 70 dBµV/m) we again start with: $P_D = \frac{P_t G_t}{4\pi R^2}$

and use the previous fact that: 10log (P_tG_t) = -33.6 dBm = 4.37x10^{-4} mW (see Section 2-4).

The measurement is at one meter so $R^2 = 1\ m^2$

we have: $\frac{4.37x\,10^{-4}}{4\pi}\, mW/\,m^2 = .348x\,10^{-4}\, mW/\,m^2 = -\,44.6\ dBm/\,m^2 = P_D$ @ 1 meter

Using the field intensity and power density relations (see Section 4-1)

$$E = \sqrt{P_D Z} = \sqrt{3.48x\,10^{-8} \bullet 377\Omega} = 36.2x\,10^{-4}\ V/m$$

Changing to microvolts (1V = 10^6 µV) and converting to logs we have:

20 log (E) = 20 log (10^6 x 36.2x10^{-4}) = 20 log (.362x10^4) = 71.18 dBµV/m ≈ 70 dBµV/m as given in Figure 1.

Some Words of Caution

A common error is to only use the one-way free space loss coefficient α_1 directly from Figure 6, Section 4-3 to calculate what the output power would be to achieve the EMCON limits at 1 NM. This is incorrect since the last term on the right of equation [3] (10 Log($4\pi R^2$)) is simply the Log of the surface area of a sphere - **it is NOT** the one-way free space loss factor α_1. You cannot interchange power (watts or dBW) with power density (watts/m^2 or dBW/m^2).

The equation uses power density (P_D), **NOT** received power (P_r). It is independent of RF and therefore varies only with range. If the source is a transmitter and/or antenna, then the power-gain product (or EIRP) is easily measured and it's readily apparent if 10log (P_t G_t) is less than -34 dBm. If the output of the measurement system is connected to a power meter in place of the system transmission line and antenna, the -34 dBm value must be adjusted. The measurement on the power meter (dBm) minus line loss (dB) plus antenna gain (dB) must not be higher than -34 dBm.

However, many sources of radiation are through leakage, or are otherwise inaccessible to direct measurement and P_D must be measured with an antenna and a receiver. The measurements must be made at some RF(s), and received signal strength is a function of the antenna used therefore measurements must be scaled with an appropriate correction factor to obtain correct power density.

RE-02 Measurements

When RE-02 measurements are made, several different antennas are chosen dependent upon the frequency range under consideration. The voltage measured at the output terminals of an antenna is not the actual field intensity due to actual antenna gain, aperture characteristics, and loading effects. To account for this difference, the antenna factor is defined as:

$$AF = E/V \qquad \textbf{[4]}$$

where E = Unknown electric field to be determined in V/m (or μV/m)
V = Voltage measured at the output terminals of the measuring antenna

For an antenna loaded by a 50 Ω line (receiver), the theoretical antenna factor is developed as follows:

$$P_D A_e = P_r = V^2/R = V_r^2/50 \text{ or } V_r = \sqrt{50\, P_D A_e}$$

From Section 4-3 we see that $A_e = G_r\lambda^2/4\pi$, and from Section 4-1, $E^2 = 377\ P_D$ therefore we have:

$$AF = \frac{E}{V} = \frac{\sqrt{377\ P_D}}{\sqrt{50\ P_D(\lambda^2 G_r / 4\pi)}} = \frac{9.73}{\lambda\sqrt{G_r}} \qquad \textbf{[5]}$$

Reducing this to decibel form we have:

$$20 \log AF = 20 \log E - 20 \log V = 20 \log\left[\frac{9.73}{\lambda\sqrt{GSUBr}}\right] \quad \textit{with } \lambda \textit{ in meters and Gain numeric ratio (not dB)} \qquad \textbf{[6]}$$

This equation is plotted in Figure 3.

Since all of the equations in this section were developed using far field antenna theory, use only the indicated region.

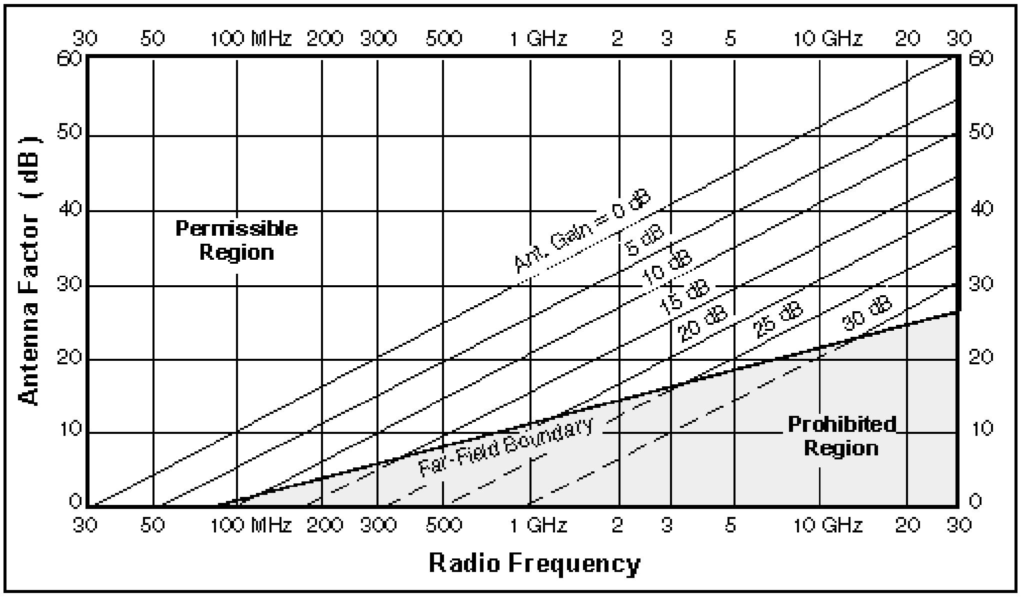

Figure 3. Antenna Factor vs. Frequency for Indicated Antenna Gain.

In practice the electric field is measured by attaching a field intensity meter or spectrum analyzer with a narrow bandpass preselector filter to the measuring antenna, recording the actual reading in volts and applying the antenna factor.

$$20\log E = 20\log V + 20\log AF \qquad \textbf{[7]}$$

Each of the antennas used for EMI measurements normally has a calibration sheet for both gain and antenna factor over the frequency range that the antenna is expected to be used. Typical values are presented in Table 1.

Table 1. Typical Antenna Factor Values.

Frequency Range	Antenna(s) Used	Antenna Factor	Gain(dB)
14 kHz - 30 MHz	41” rod	22-58 dB	0 - 2
20 MHz - 200 MHz	Dipole or Biconical	0-18 dB	0 - 11
200 MHz - 1 GHz	Conical Log Spiral	17-26 dB	0 - 15
1 GHz - 10 GHz	Conical Log Spiral or Ridged Horn	21-48 dB	0 - 28
1 GHz - 18 GHz	Double Ridged Horn	21-47 dB	0 - 32
18 GHz - 40 GHz	Parabolic Dish	20-25 dB	27 - 35

The antenna factor can also be developed in terms of the receiving antenna's effective area. This can be shown as follows:

$$AF = \frac{E}{V} = \frac{\sqrt{377\,P_D}}{\sqrt{50\,P_D\,A_e}} = \frac{2.75}{\sqrt{A_e}} \qquad \textbf{[8]}$$

Or in log form:

$$20\log AF = 20\log E - 20\log V = 20\log\left[\frac{2.75}{\sqrt{A_e}}\right] \qquad \textbf{[9]}$$

While this relation holds for any antenna, many antennas (spiral, dipole, conical etc.) which do not have a true "frontal capture area" do not have a linear or logarithmic relation between area and gain and in that respect the parabolic dish is unique in that the antenna factor does not vary with frequency, only with effective capture area. Consequently a larger effective area results in a smaller antenna factor.

A calibrated antenna would be the first choice for making measurements, followed by use of a parabolic dish or "standard gain" horn. A standard gain horn is one which was designed such that it closely follows the rules of thumb regarding area/gain and has a constant antenna factor. If a calibrated antenna, parabolic dish, or "standard horn" is not available, a good procedure is to utilize a flat spiral antenna (such as the AN/ALR-67 high band antennas). These antennas typically have an average gain of 0 dB (typically -4 to +4 dB), consequently the antenna factor would not vary a lot and any error would be small.

EXAMPLE:

Suppose that we want to make a very general estimation regarding the ability of a system to meet EMCON requirements. We choose to use a spiral antenna for measurements and take one of our samples at 4 GHz. Since we know the gain of the spiral is relatively flat at 4 GHz and has a gain value of approximately one (0 dB) in that frequency range. The antenna is connected to a spectrum analyzer by 25 feet of RG9 cable. We want to take our measurements at 2 meters from the system so our setup is shown as follows:

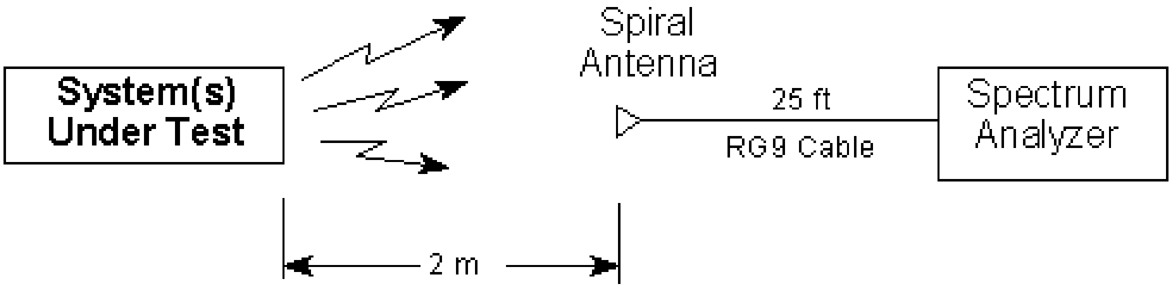

Our RG9 cable has an input impedance of 50Ω, and a loss of 5 dB (from Figure 5, Section 6-1).

First, let's assume that we measure -85 dBm at the spectrum analyzer and we want to translate this into the equivalent strength at 1 NM. Our power received by the antenna is: P_r = -85 dBm + 5 dB line loss = -80 dBm

also $P_D = P_r/A_e$ and $A_e = G\lambda^2/4\pi = (G/4\pi)\bullet(c/f)^2 = (1/4\pi)\bullet(3\text{x}10^8/4\text{x}10^9)^2 = 4.47\text{x}10^{-4}\ \text{m}^2$

in log form: $10\ \text{Log}\ P_D = 10\ \text{Log}\ P_r - 10\ \text{Log}\ A_e = -80\ \text{dBm} + 33.5 = -46.5\ \text{dBm/m}^2$
at our 2 meter measuring point

To convert this to a value at 1 NM, we use

$P_t\ G_t = P_{D@1\ nm}\ 4\pi R_1^{\ 2} = P_{D@2\ m}\ 4\pi R_2^{\ 2}$ and we solve for PD@1 $_{nm}$

in log form after cancelling the 4π terms:

$10\ \text{Log}\ P_{D@1\ nm} = 10\ \text{Log}\ P_{D@2\ m} + 10\ \text{Log}\ (R_{2m}/R_{1nm})^2 = -46.5\ \text{dBm/m}^2 - 59.3\ \text{dB} = -105.8\ \text{dBm/m}^2$ which is more power than the maximum value of -110 dBm/m^2 specified.

If we are making repetitive measurement as we might do when screening an aircraft on the flight line with numerous systems installed, or when we want to improve (reduce) the leakage on a single system by changing antennas, lines, connectors, or EMI gaskets or shielding, this mathematical approach would be unnecessarily time consuming since it would have to be repeated after each measurement. A better approach would be to convert the -110 dBm/m^2 value at 1 NM to the maximum you can have at the measuring instrument (in this case a spectrum analyzer), then you could make multiple measurements and know immediately how your system(s) are doing. It should be noted that -90 to -100 dBm is about the minimum signal level that can be detected by a spectrum analyzer, so you couldn't take measurements much further away unless you used an antenna with a much higher gain.

In order not to exceed EMCON, the power density must not exceed -110 dBm/m^2 at 1 NM, which is $10^{-11}\ \text{mW/m}^2$.

$$P_t\ G_t = P_{D@1\ nm}\ 4\pi R_1^{\ 2} = P_{D@2\ m}\ 4\pi R_2^{\ 2}$$

we solve for $P_{D@2\ m} = 10^{-11}(1852\text{m})^2/(2\text{m})^2 = 8.57\ \text{x}\ 10^{-6}\ \text{mW/m}^2 = -50.7\ \text{dBm/m}^2$

We'll be using a spectrum analyzer, so we want to compute what the maximum power or voltage may be.

Method 1 - Using the Power Density Approach

Using logs/dB and the values of $P_{D@2\ m}$ and A_e determined previously:
$10\ \text{Log}\ P_r = 10\ \text{Log}\ P_D + 10\ \text{Log}\ A_e = -50.7 - 33.5 = -84.2\ \text{dBm}$

taking line loss into account we have: -84.2 - 5 dB = - 89.2 dBm as the maximum measurement reading.

If we wanted to calculate it in volts, and take into account our line impedance we would have the following:

$P_r = P_D A_e = V^2/R = V^2/50\Omega$ also $A_e = G\lambda^2/4\pi$ so solving for V we have:

$$V = \sqrt{P_D\left[\frac{G_r\lambda^2}{4\pi}\right]R} = \sqrt{P_D\left[\frac{G_r}{4\pi}\left(\frac{c}{f}\right)^2\right]R} = \sqrt{8.57x10^{-9}\left[\frac{1}{4\pi}\left(\frac{3x10^8}{4x10^9}\right)^2\right]50\Omega} = 1.38x10^{-5}\ volts\ \ (before\ line\ loss)$$

since our line loss is 5 dB, we have -5 dB = 20 Log V_2/V_1. Solving for V_2 we get $7.79x10^{-6}$ volts or -89 dBm as a maximum at our measurement device input. We can see immediately that our value of -85 dBm that we measured on the previous page would not meet specifications, and neither would any signal with more power than -89 dBm.

Method 2 - Using the Antenna Factor Approach

Starting with the same value of power density that we obtained above ($8.57x10^{-9}$ W/m^2), we find the field intensity from Table 1, Section 4-1 to be approximately 65 dBμv/m. Also from Figure 3 in this section, AF = 43 dB @ 4 GHz (by calculating with equation [6], the exact value is 42.3 dB).

From equation [6]:
20log V = 20log E - 20log AF
20log V = 65 - 43 = 22 dBμv/m.

Since dBμv/m = 20 log (V)(10^6) = 20 log V + 20 log 10^6 = 20 log V + 120, we see that to get an answer in dBv we must subtract 120 from the dBμv/m value so: V_{dB} = 22 - 120 = -98dBv. We then subtract our line loss (-5 dB) and we have:

V = -98 - 5 = -103 dBv = 17 dBμv = $7.1x10^{-6}$ volts

using the fact that $P = V^2/R$ and for the input line R = 50Ω, P = $1x10^{-12}$ W = -120 dBW = -90 dBm

Although this method is just as accurate as that obtained using method 1, the values obtained in Table 1, Section 4-1, and Figure 3 must be interpolated, and may not result in values which are as precise as the appropriate formulas would produce.

Sample Problem: What is the approximate transmit power from a receiver?

A. 1 nanowatt (nW)
B. 10 nW
C. 100 nW
D. 1 microwatt (μW)
E. 10 μW
F. 100 μW
G. 1 milliwatt (mW)
H. 10 mW
I. 100 mW
J. 1 watt (W)
K. 10 W
L. 100 W
M. 1 kilowatt (kW)
N. 10 kW
O. 100 kW

The question may seem inappropriate since a receiver is supposedly a passive device which only receives a signal. If the receiver was a crystal video receiver as shown in Section 5-3, it wouldn't transmit power unless a built-in-test (BIT) signal was injected after the antenna to periodically check the integrity of the microwave path and components. The potential exists for the BIT signal to leak across switches and couple back through the input path and be transmitted by the receiver's antennas.

If the receiver uses a local oscillator (LO) and a mixer to translate the signal to an intermediate frequency (IF) for processing (such as a superhet shown in Section 5-3), there is the potential for the CW LO signal to couple back through the signal input path and be transmitted by the receiver's antenna. Normally a mixer has 20 dB of rejection for the reverse direction. In addition, the LO may be further attenuated by receiver front end filters.

In both cases, the use of isolators described in Section 6-7 could be used to further attenuate any signals going in the reverse direction, i.e. back to the antenna. A good receiver design should ensure that any RF leakage radiated by the receiver will not exceed the EMCON level.

In answer to the initial question, "transmit" leakage power should be less than -34 dBm (0.4 μW) to meet EMCON. Therefore, the real answer may be "A," "B," or "C" if EMCON is met and could be "D" through possibly "G" if EMCON is not met.

EW JAMMING TECHNIQUES

INTRODUCTION

Electronic jamming is a form of Electronic Attack where jammers radiate interfering signals toward an enemy's radar, blocking the receiver with highly concentrated energy signals. The two main technique styles are noise techniques and repeater techniques. The three types of noise jamming are spot, sweep, and barrage. Repeater techniques can be further subdivided into categories as shown in Figure 1.

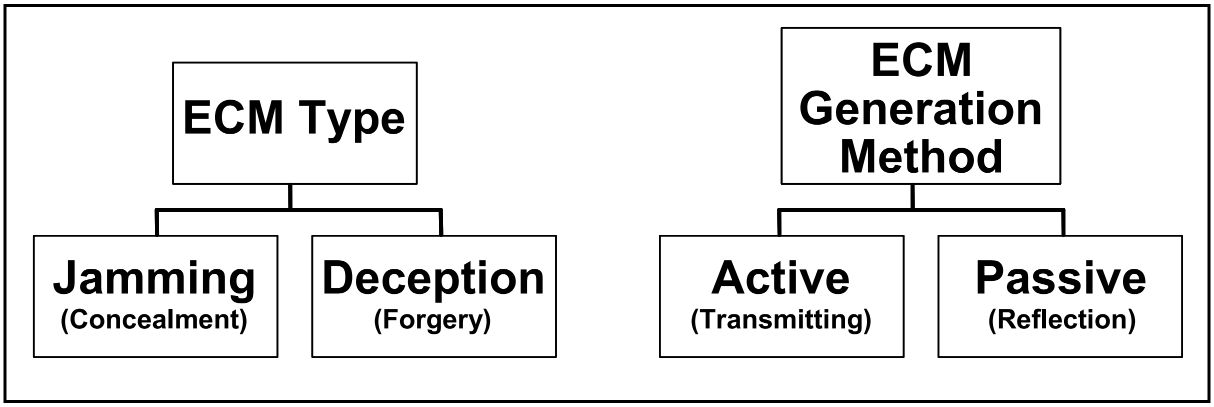
ECM Type
Jamming
(Concealment)
Deception
(Forgery)
ECM Generation Method
Active
(Transmitting)
Passive
(Reflection)

Figure 1. EA Repeater Technique Divisions.

ASYNCHRONOUS SWEPT WAVE MODULATION (ASWM)

ASWM is synonymous with A-SWM. Asynchronous indicates that the waveform is free running - also see SSWM. A swept wave modulation is essentially a swept amplitude modulation (SAM). It is a waveform that is swept between two frequencies that are usually chosen to bracket a radar's passive angle scanning rate. The modulation amplitude can be either down modulated or On-Off Keyed (OOK). The down modulated shape can be square wave, rectangular wave, linear (e.g. a sine wave), or a combination. The OOK modulated shape can be square wave or rectangular wave.

BARRAGE JAMMING

The jamming of multiple frequencies at once by a single jammer. The advantage is that multiple frequencies can be jammed simultaneously; however, the jamming effect can be limited because this requires the jammer to spread its full power between these frequencies. So the more frequencies being jammed, the less effectively each is jammed.

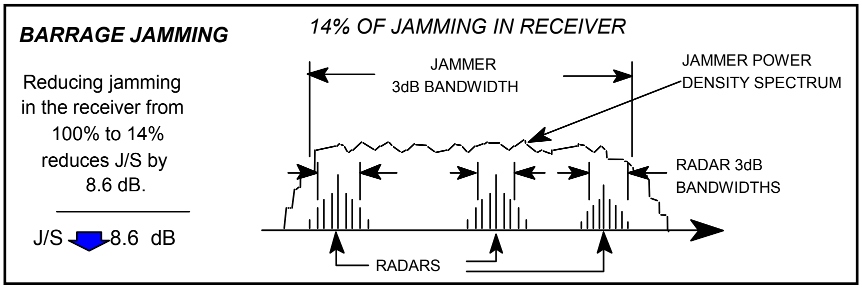
BARRAGE JAMMING
14% OF JAMMING IN RECEIVER
Reducing jamming in the receiver from 100% to 14% reduces J/S by 8.6 dB.
J/S 8.6 dB
JAMMER 3dB BANDWIDTH
JAMMER POWER DENSITY SPECTRUM
RADAR 3dB BANDWIDTHS
RADARS

Figure 2. Barrage Jamming.

CROSS POLARIZATION (X-POL)

(1) A self-screening or support EA technique that causes angle errors in tracking radars and sensing errors in jamming suppression EP systems of surveillance radars by radiating a signal that is orthogonally polarized to the principal polarization of the victim radar. (2) A technique used against monopulse and other passive lobe tracking radars. Requires a strong jam-to-signal ratio or the skin echo will show up in the pattern nulls.

HOME ON JAM (HOJ)

A means whereby a missile guidance receiver utilizes the self-screening target jamming signal to develop angular steering information so that the missile can home on that target.

IMAGE JAMMING

Jamming at the image frequency of the radar receiver. Barrage jamming is made most effective by generating energy at both the normal operating and image frequency of the radar. Image jamming inverts the phase of the response and is thereby useful as an angle deception technique. Not effective if the radar uses image rejection.

INVERSE CON SCAN (ICS)

One method of confusing a radar operator or fire control radar system is to provide erroneous target bearings. This is accomplished by first sensing the radar antenna or antenna dipole scan rate and then modulating repeater amplifier gain so that the weapons system will fire at some bearing other than the true target bearing.

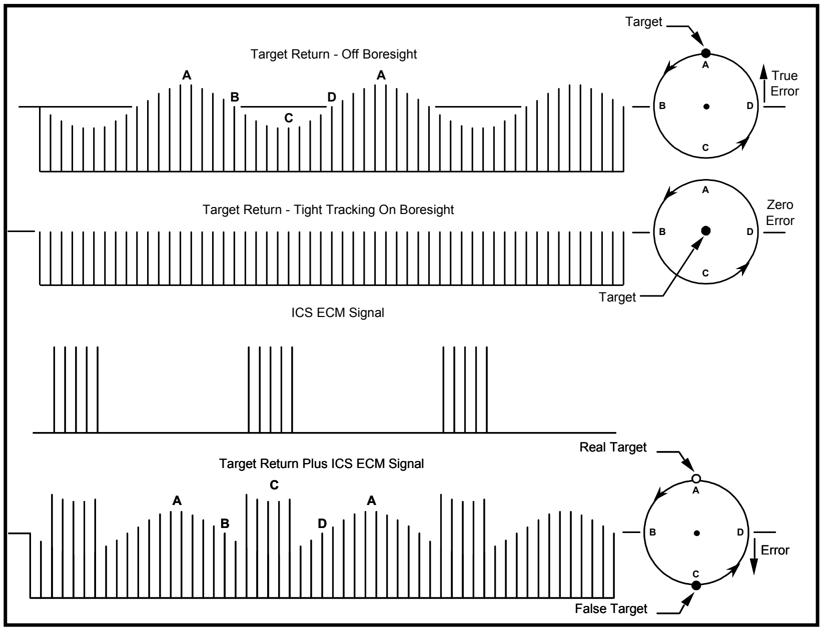
Target
Target Return - Off Boresight
A
A
B
C
D
True Error
Target Return - Tight Tracking On Boresight
Zero Error
Target
ICS ECM Signal
Real Target
Target Return Plus ICS ECM Signal
C
A
A
B
D
Error
False Target

Figure 3. Inverse Con Scan.

J/S REQUIREMENT

The jamming to signal ratio for effective coverage of the true target. Usually on the order of zero dB (J/S=1).

LINEAR MODULATION

A modulation technique in which the output varies in a straight line (linear) manner with the input (modulating) waveform. This is different from a discontinuous (On-Off) modulation. Typically the modulating waveform is a ramp or triangular wave but could also be a sine wave modulating input.

LOBE ON RECEIVE ONLY (LORO)

Predicting the future location of a target (to track) requires that the radar look at areas where the target is not located. When this scanning is accomplished with the radiated beam, large angle targets such as chaff clouds can create complete radar "white-out." RWR indications can be obtained before actual target lock-on. These problems can be overcome by scanning only the receiving antennas and using a separate transmitting antenna pointed only at the target.

In a LORO system, a transmitting antenna emits a few "exploratory" pulses along a direction obtained from an acquisition radar. These exploratory pulses are the acquisition mode of the TTR. That is, in its acquisition mode the small beamed TTR must scan the large location segment provided by the acquisition radar. In radars equipped with Fast Time Constant, the return pulse is applied to a differentiator of extremely short time constant. When the pulse is received, it is "cut-off" on the leading edge and only that portion is fed to the computer. This allows the radar to effectively track on the leading edge of the target. FTC does not improve the range resolution but it can prevent any countermeasures aft of the target which are in the same resolution cell as the target (such as chaff) from interfering with the radar receiver.

The receiving antennas scan their sector for the target return due to these exploratory pulses; as the power centroid is located, the center of the receiving pattern is brought onto the target. The transmitting antenna, which is slaved to the receiving antenna, is then pointing directly at the desired target and only that target is radiated during tracking. This approach allows a very small radiated beam, but the resolution cell of the system is still that of the receiving antenna.

NOISE JAMMING

The transmission of noise-like signals in the target system's radar receiver bandpass. At low power levels, noise jamming has the characteristics of receiver noise and can be mistaken by the radar operator as a problem with the radar. The object of noise jamming is to introduce a disturbing signal into the hostile electronic equipment so that the actual signal is obscured by the interference. The victim of this disturbance might be a radar receiver, a communications network, or a data link.

See also Barrage Jamming and Spot Jamming.

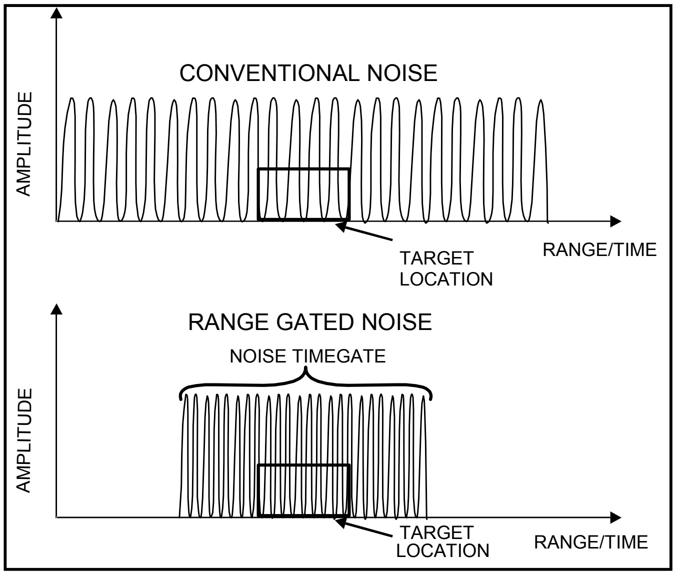

Figure 4. Noise Jamming.

ON-OFF MODULATION

On-Off Modulation is any modulation which switches rapidly between two states. This definition includes pulse radar operation.

On-Off Keying (OOK) is the envelope modulation of a jamming signal with a rectangular wave. The modulation rate and duty cycle are adjusted commensurate with the victim radar's processing time constants. These can be related to AGC time constants, logic time-outs, data sampling cycles, or any other data processing response times.

An important distinction must be made between the terms On-Off Keying and Blinking. While both terms involve envelope modulations that turn a jamming signal on and off, the term OOK is used as the envelope modulation of a single jamming signal, and the term blinking is used as the tactical application of OOK involving two or more cooperative jamming platforms.

PSEUDO RANDOM NOISE (PRN)

A controlled, noise-like, pulse pattern repeated in synchronism with the victim radar pulse repetition frequency. Synonymous with quasi-noise jamming.

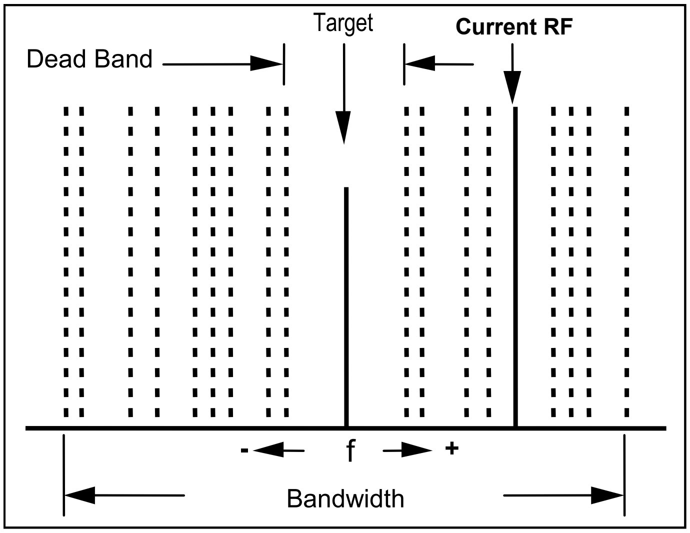

Figure 5. Random Dual Line.

RANDOM DUAL LINE (RDL)

RDL is a coherent repeater technique that is essentially the same as velocity noise (VN), narrow band noise (NBN), and pseudo random noise (PRN) except that no false Doppler frequencies are stepped directly over the target return. The objective is to prevent the momentary additions of the EA signal and target signal that might highlight the presence of a coherent return.

RANDOM RANGE PROGRAM (RANRAP)

A dynamic False Target Jamming technique program to create multiple realistic targets of varying size and distance from the jamming plane.

RANGE GATE PULL OFF (RGPO)

Once a tracking radar has detected a target, it will place *range gates* to either side of it. Range gates essentially blank out all signals which originate from ranges outside a narrow window, substantially increasing the signal-to-noise ratio and protecting the radar against unsynchronized jamming pulses. The radar 'concentrates' on a short range interval which encloses the target's location, and it no longer looks out for other targets. This state is known as 'lock-on'. But range gates can be 'stolen', and it is the objective of the Range Gate Pull-Off (RGPO) technique to *break lock* and escape from out of the window.

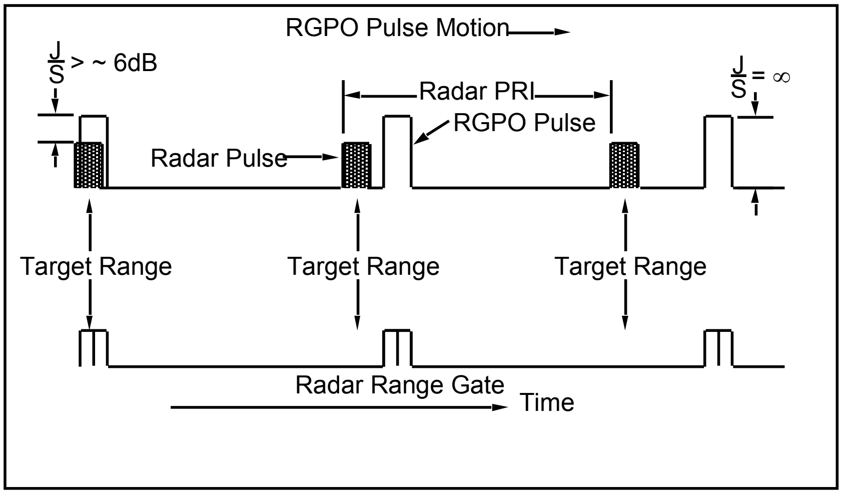

Figure 6. Range Gate Pull Off.

RGPO works as follows:

Upon detection (or assumption) that a tracking radar has locked on, the on-board jammer is switched on and starts to work in a couple of phases:

1. First, a sample of the illuminating pulse signal is taken and the radar's pulse repetition frequency (PRF) is determined. This sample is amplified and retransmitted simultaneously when further pulses are received. The aircraft actually highlights itself on the radar screen. The jamming power is steadily increased, and this continues until the replica is much stronger than the echo from the aircraft skin return. At this time, the sensitivity of the tracking radar's receiver is usually reduced in order to avoid overload. This causes the skin echo vanishes below the noise floor.

2. *Another* replica is transmitted after each of the 'dummy' skin echoes. The power of the second replica is increased while the dummy is made weaker.

3. Next, the tracker has locked on to the *delayed replica*, whereas the skin return has decreased into the noise. With respect to each of the radar's pulses, the replica is now being delayed by small, but increasing amounts of time. The range gates, of course, follow the dummy target which appears to be receding. This continues until the range gates have been moved away from the target's real position. The result is that the radar is tracking a *phantom target* and the skin return is being blanked out by the range gates.

4. Finally, the jammer is switched off and leaves the radar with just *nothing* but noise inside the window between its range gates. Break-lock was successfully achieved and the "tracking" radar needs to

switch back into a search or acquisition mode and loses time. The whole cycle will start again if the target is still within range and is reacquired.

As described above, RGPO creates only false targets which appear at greater ranges than the real target because the deceptive signal is transmitted *after* the skin echo. However, if the victim radar's PRF is constant then the time of incidence of the next radar pulse can be calculated and jamming pulses can be placed such that false targets at closer ranges are also produced.

SCINTILLATION (SCINT)

Scintillation is not an EA technique by itself, it is an implementation of an EA technique. Scintillation is simply superimposing a small, pseudo random amplitude modulation on the EA signal to make it appear more realistic to a manual operator.

SPOT JAMMING

Occurs when a jammer focuses all of its power on a single frequency. While this would severely degrade the ability to track on the jammed frequency, a frequency agile radar would hardly be affected because the jammer can only jam one frequency. While multiple jammers could possibly jam a range of frequencies, this would consume a great deal of resources to have any effect on a frequency-agile radar, and would probably still be ineffective

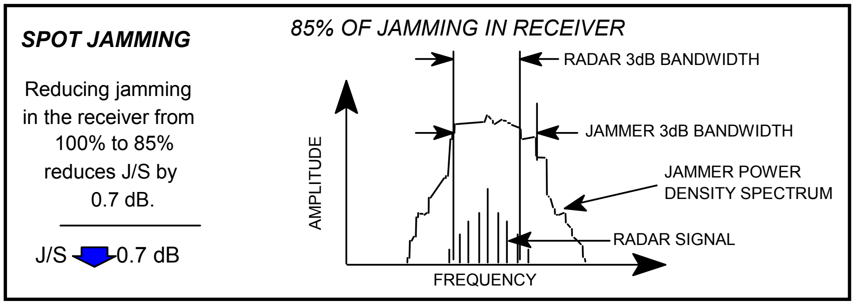
SPOT JAMMING
Reducing jamming in the receiver from 100% to 85% reduces J/S by 0.7 dB.
J/S ↓ 0.7 dB
85% OF JAMMING IN RECEIVER
RADAR 3dB BANDWIDTH
JAMMER 3dB BANDWIDTH
JAMMER POWER DENSITY SPECTRUM
RADAR SIGNAL
AMPLITUDE
FREQUENCY

Figure 7. Spot Jamming.

SWEPT JAMMING

This happens when a jammer's full power is shifted from one frequency to another. While this has the advantage of being able to jam multiple frequencies in quick succession, it does not affect them all at the same time, and thus limits the effectiveness of this type of jamming. Although, depending on the error checking in the receiver(s) this can render a wide range of receivers effectively useless.

SWEPT AMPLITUDE MODULATION (SAM)

The OOK frequency is linearly varied in a sawtooth fashion between preset frequency limits while the duty factor is held constant.

SWEPT WAVE MODULATION (SWM)

A swept wave modulation.2 (SWM.2) is essentially a swept amplitude modulation (SAM). It is a waveform that is swept between two frequencies that are usually chosen to bracket a radar's passive angle scanning rate. SWM can be either Synchronous Swept Wave Modulation.2-; (S-SWM) or Asynchronous Swept Wave Modulation.2-; (A-SWM.2-;). The modulation amplitude can be either down modulated or On-Off Keyed (OOK). The down modulating shape can be square wave, rectangular wave, linear (e.g. a sine wave), or a combination. The OOK modulating shape can be square wave or rectangular wave.

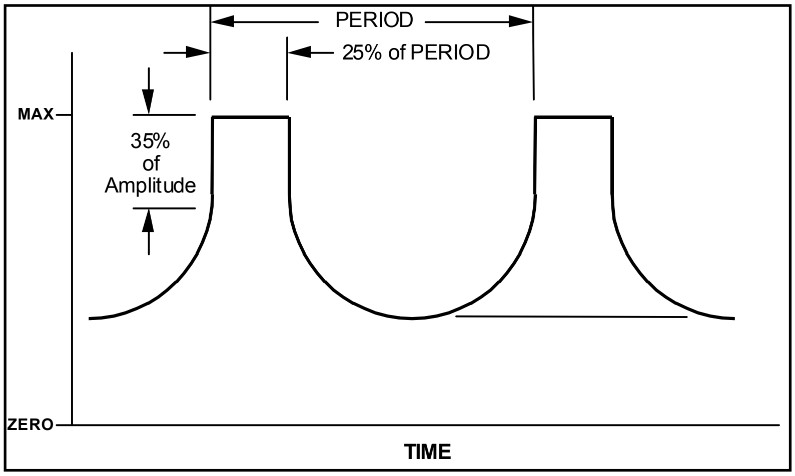

Figure 8. Typical Linear SWM Modulation.

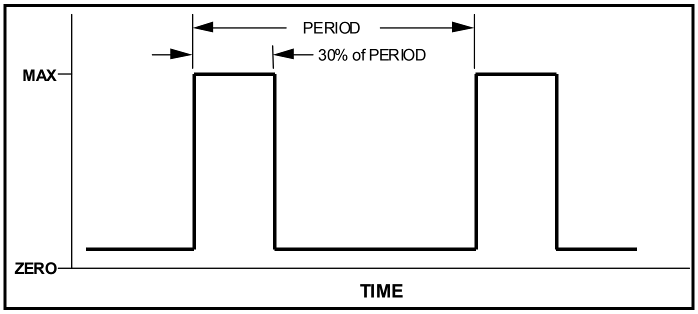

Figure 9. Rectangular SWM.

SYNCHRONOUS SWEPT WAVE MODULATION (SSWM)

SSWM is synonymous with S-SWM. S-SWM and A-SWM are essentially the same except that asynchronous means that the waveform is free running and synchronous means that when a radar scan or TWS beam can be detected, the modulation waveform is synchronized to the detected beam. For programming purposes A-SWM sets sweep limits and rate by frequency (Hz), while S-SWM sets them by period (mSec).

Active Con-Scan radars will not have a detectable modulation if the target is being tightly tracked in the center of the beam. Therefore, a SWM can "jog" the tracking sufficiently to detect the modulation and allowing subsequent synchronization of the SWM waveform.

TRACK WHILE SCAN JAMMING

The technique of shifting or "walking" EA pulses off target. Many angle jamming techniques are effective.

VELOCITY FALSE TARGETS (VFT)

VFT is a pseudo-random false Doppler target concealment technique. It is designed for use against radars that acquire Doppler targets with a bank of contiguous narrow band filters. A false Doppler target is programmed to remain in a Doppler filter long enough for the radar processing to declare it a valid target return, but not long enough for the radar processing to establish tracking.

The false Doppler target is then switched to the next pseudo-randomly selected frequency and repeated. It is intended to overload the radar processing and/or the operator's ability to identify an actual target. In the illustration, the VFTs jump around in the indicated numerical order.

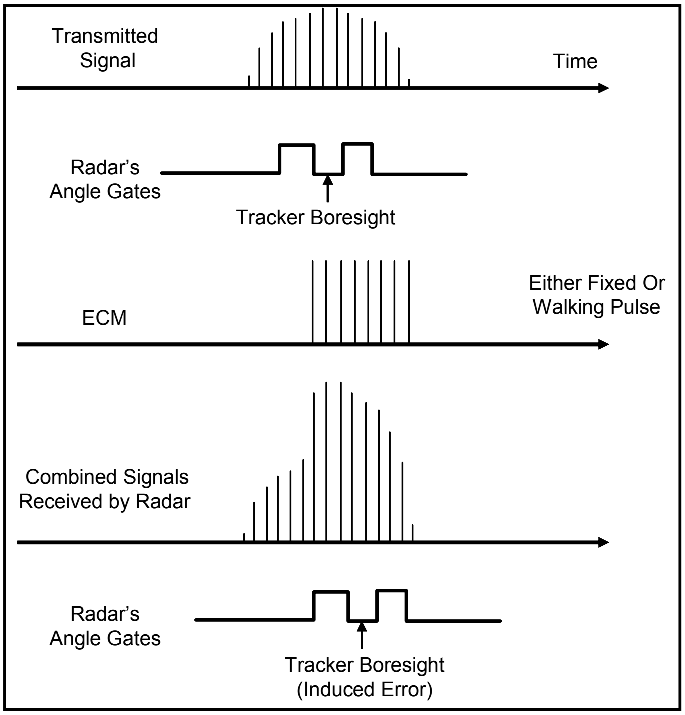
Transmitted Signal
Time
Radar's Angle Gates
Tracker Boresight
ECM
Either Fixed Or Walking Pulse
Combined Signals Received by Radar
Radar's Angle Gates
Tracker Boresight (Induced Error)

Figure 10. TWS Jamming.

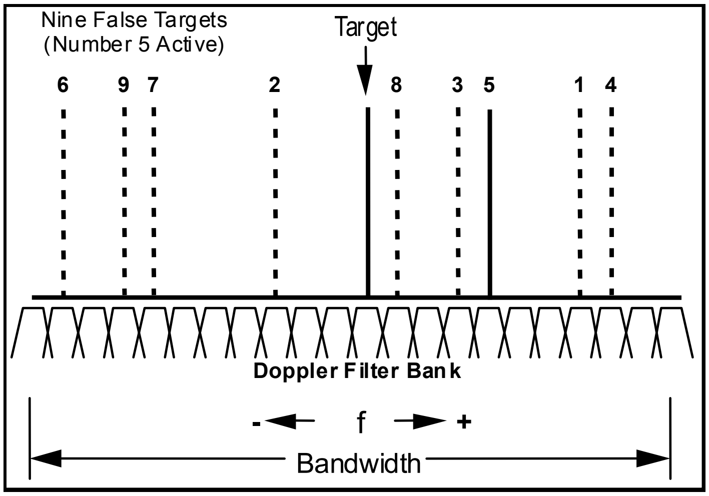
Nine False Targets (Number 5 Active)
Target
6
9
7
2
8
3
5
1
4
Doppler Filter Bank
-
f
+
Bandwidth

Figure 11. Velocity False Targets.

VELOCITY GATE PULL OFF

This is a method of capturing the velocity gate of a Doppler radar and moving it away from the skin echo. Similar to the RGPO, but used against CW or Doppler velocity tracking radar systems. The CW or pulse doppler frequency, which is amplified and retransmitted, is shifted in frequency (velocity) to provide an apparent rate change or Doppler shift.

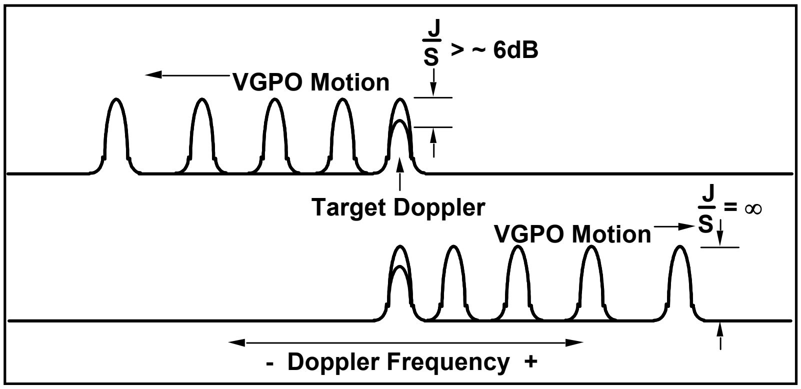

Figure 12. Velocity Gate Pull Off.

VELOCITY NOISE (VN)

VN is a coherent repeater technique. The objective is to create noise centered on a coherent radar's RF, with a noise bandwidth that is close to, or less than, the radar's bandwidth and conceal the target, or destroy target signal coherency. VN is generated by pseudo randomly stepping a frequency over the victim radar's bandwidth. The dashed RF lines represent possible frequencies and the solid line represents the frequency currently active.

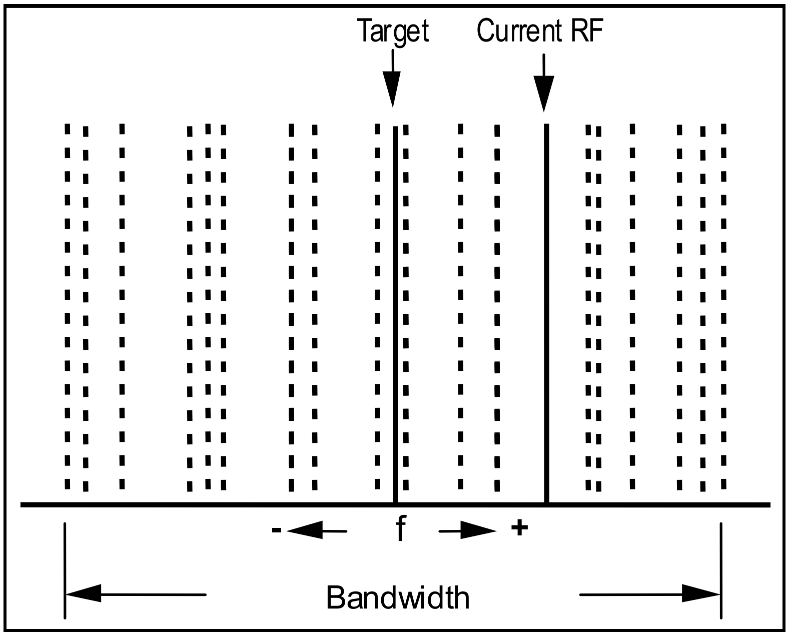

Figure 13. Velocity Noise.

This page intentionally left blank.

RADAR AND RECEIVER CHARACTERISTICS & TEST

RF Atmospheric Absorption / Ducting 5-1
Receiver Sensitivity / Noise 5-2
Receiver Types and Characteristics 5-3
Radar Modes 5-4
General Radar Display Types 5-5
IFF - Identification - Friend or Foe 5-6
Receiver Tests 5-7
Signal Sorting and Direction Finding 5-8

This page intentionally left blank.

RF ATMOSPHERIC ABSORPTION / DUCTING

Signal losses are associated with each stage of signal processing in both the transmitting and receiving portions of the system. The transmitting losses include power transmission efficiency, waveguide and antenna losses, and duplexer losses. In the receiver, losses include antenna, waveguide, RF amplifier, mixer, and IF amplifier.

In addition to these losses, energy traveling through the atmosphere suffers from atmospheric attenuation caused primarily by absorption by the gasses. For lower frequencies (below 10 GHz), the attenuation is reasonably predictable. For high frequencies in the millimeter wave range, the attenuation not only increases, but becomes more dependent upon peculiar absorbing characteristics of H_2O, O_2, and the like.

Figure 1 shows the areas of peak absorption in the millimeter wave spectrum. Figure 2 shows how the intensity of precipitation can affect atmospheric attenuation.

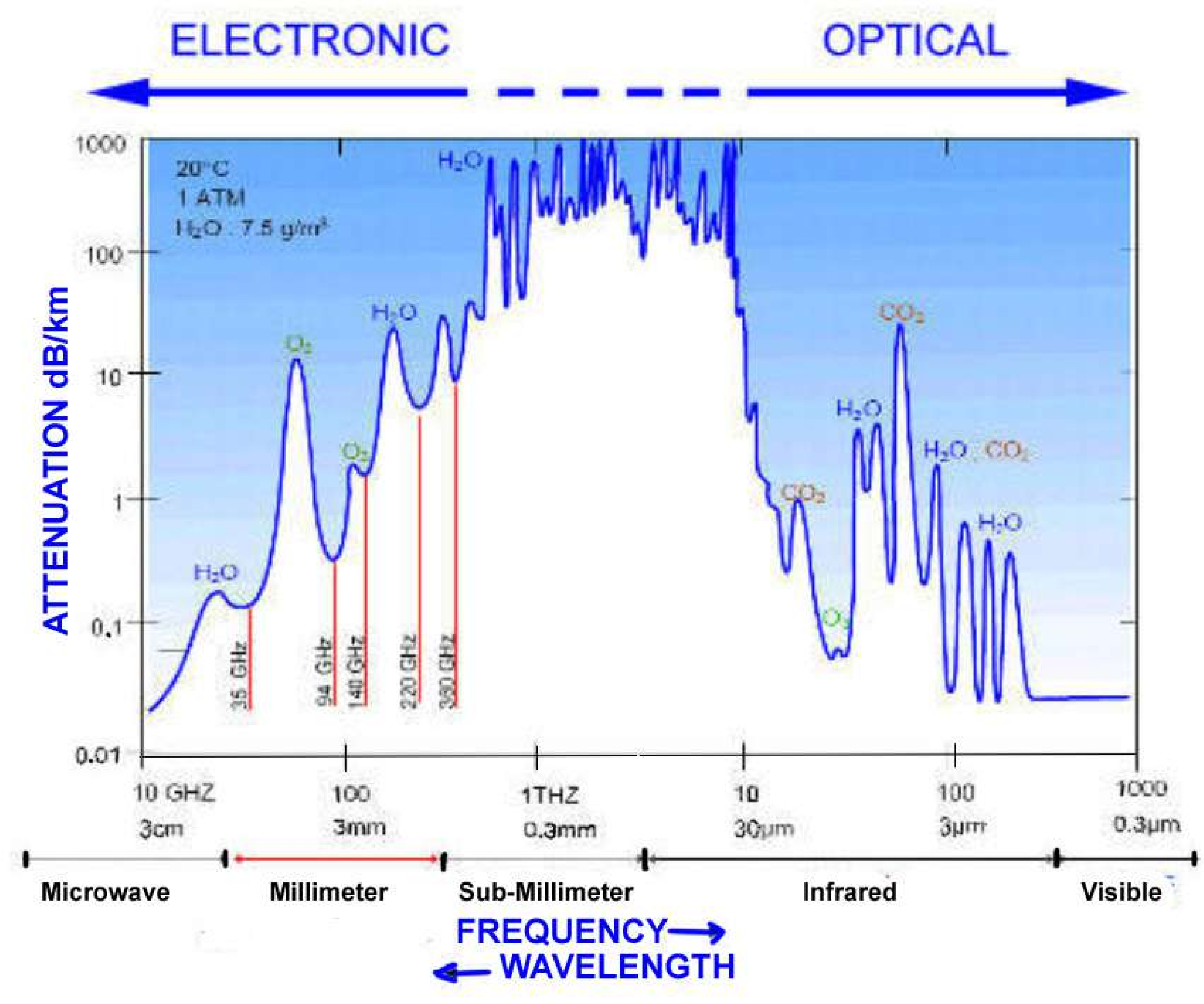

Figure 1. Atmospheric Absorption of Millimeter Waves.

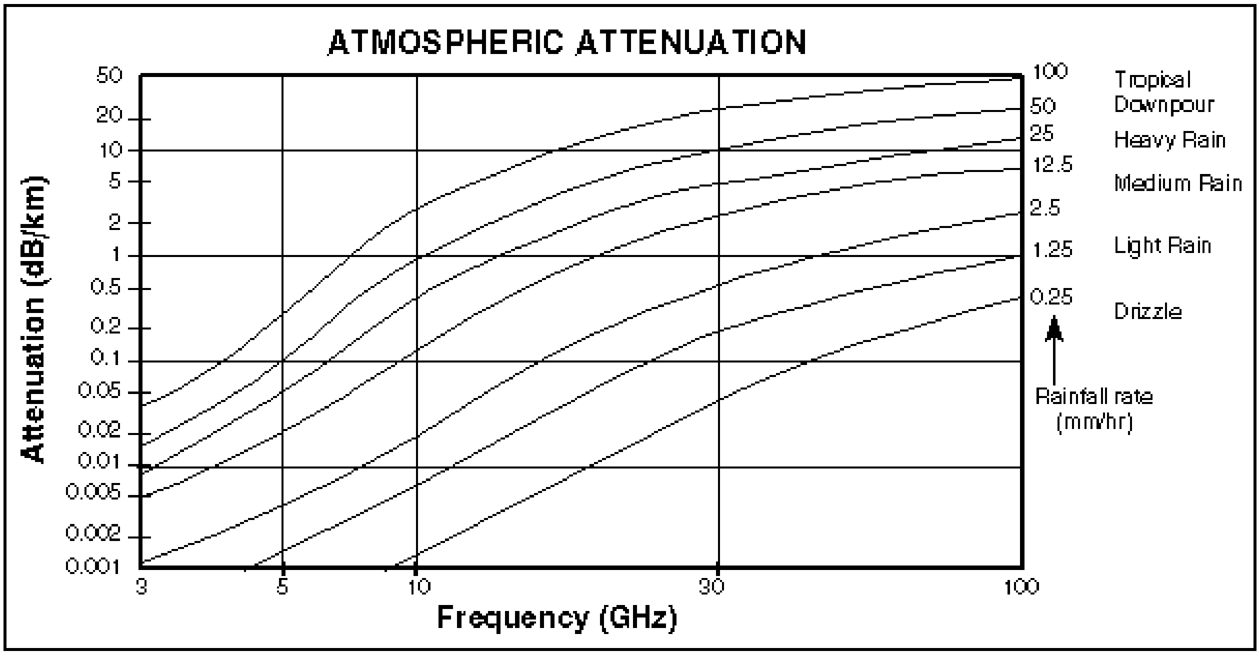

Figure 2. Atmospheric Attenuation.

Ducting is an increase in range that an electromagnetic wave will travel due to a temperature inversion of the lower atmosphere (troposphere) as shown in Figure 3. The temperature inversion forms a channel or waveguide (duct) for the waves to travel in, and they can be trapped, not attenuating as would be expected from the radar equation. Ducting may also extend range beyond what might be expected from limitations of the radar horizon (see Section 2-9).

The ducting phenomena is frequency sensitive. The thicker the duct, the lower the minimum trapped frequency.

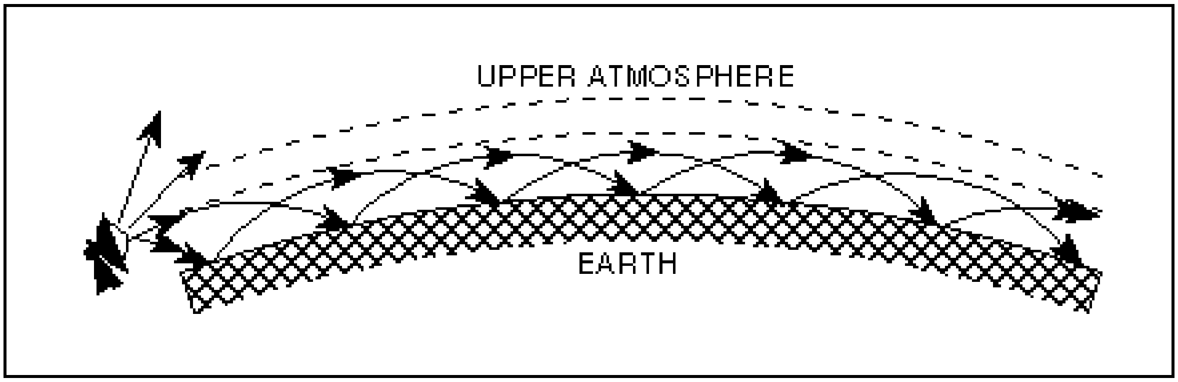

Figure 3. Ducting.

A similar occurrence takes place with ionospheric refraction, however the greatest increase in range occurs in the lower frequencies. This is familiar to amateur radio operators who are able to contact counterparts "around the world."

RECEIVER SENSITIVITY / NOISE

RECEIVER SENSITIVITY

Sensitivity in a receiver is normally taken as the minimum input signal (S_{min}) required to produce a specified output signal having a specified signal-to-noise (S/N) ratio and is defined as the minimum signal-to-noise ratio times the mean noise power, see equation [1]. For a signal impinging on the antenna (system level) sensitivity is known as minimum operational sensitivity (MOS), see equation [2]. Since MOS includes antenna gain, it may be expressed in dBLi (dB referenced to a linear isotropic antenna). When specifying the sensitivity of receivers intended to intercept and process pulse signals, the minimum pulse width at which the specified sensitivity applies must also be stated. See the discussion of post-detection bandwidth (B_V) in Section 5-2 for significance of minimum pulsewidth in the receiver design.

$\mathbf{S_{min} = (S/N)_{min} k T_o B(NF)}$ receiver sensitivity ("black box" performance parameter) **[1]**

or $\mathbf{MOS = (S/N)_{min} k T_o B(NF)/G}$ system sensitivity i.e. the receiver is connected to an antenna (transmission line loss included with antenna gain) **[2]**

where: S/N_{min} = Minimum signal-to-noise ratio needed to process (vice just detect) a signal
NF = Noise figure/factor
k = Boltzmann's Constant = 1.38×10^{-23} Joule/°K
T_o = Absolute temperature of the receiver input (°Kelvin) = 290°K
B = Receiver Bandwidth (Hz)
G = Antenna/system gain

We have a lower MOS if temperature, bandwidth, NF, or S/N_{min} decreases, or if antenna gain increases. For radar, missile, and EW receivers, sensitivity is usually stated in dBm. For communications and commercial broadcasting receivers, sensitivity is usually stated in micro-volts or dBμv. See Section 4-1.

There is no standard definition of sensitivity level. The term minimum operational sensitivity (MOS) can be used in place of S_{min} at the system level where aircraft installation characteristics are included. The "black box" term minimum detectable signal (MDS) is often used for S_{min} but can cause confusion because a receiver may be able to detect a signal, but not properly process it. MDS can also be confused with minimum discernable signal, which is frequently used when a human operator is used to interpret the reception results. A human interpretation is also required with minimum visible signal (MVS) and tangential sensitivity (discussed later). To avoid confusion, the terms S_{min} for "black box" minimum sensitivity and MOS for system minimum sensitivity are used in this section. All receivers are designed for a certain sensitivity level based on requirements. One would not design a receiver with more sensitivity than required because it limits the receiver bandwidth and will require the receiver to process signals it is not interested in. In general, while processing signals, the higher the power level at which the sensitivity is set, the fewer the number of false alarms which will be processed. Simultaneously, the probability of detection of a "good" (low-noise) signal will be decreased.

Sensitivity can be defined in two opposite ways, so discussions can frequently be confusing. It can be the ratio of response to input or input to response. In using the first method (most common in receiver discussions and used herein), it will be a negative number (in dBm), with the more negative being "better" sensitivity, e.g., -60 dBm is "better" than -50 dBm sensitivity. If the second method is used, the result will be a positive number, with higher being "better." Therefore the terms low sensitivity or high sensitivity can be very confusing. The terms S_{min} and MOS avoid confusion.

SIGNAL-TO-NOISE (S/N) RATIO

The Signal-to-Noise Ratio (S/N) (a.k.a. SNR) in a receiver is the signal power in the receiver divided by the mean noise power of the receiver. All receivers require the signal to exceed the noise by some amount. Usually if the signal power is less than or just equals the noise power it is not detectable. For a signal to be detected, the signal energy plus the noise energy must exceed some threshold value. Therefore, just because N is in the denominator doesn't mean it can be increased to lower the MOS. S/N is a required minimum ratio, if N is increased, then S must also be increased to maintain that threshold. The threshold value is chosen high enough above the mean noise level so that the probability of random noise peaks exceeding the threshold, and causing false alarms, is acceptably low.

Figure 1 depicts the concept of required S/N. It can be seen that the signal at time A exceeds the S/N ratio and indicates a false alarm or target. The signal at time B is just at the threshold, and the signal at time C is clearly below it. In the sample, if the temperature is taken as room temperature (T_o = 290°K), the noise power input is -114 dBm for a one MHz bandwidth. Normally S/N_{min} may be set higher than S/N shown in Figure 1 to meet false alarm specifications.

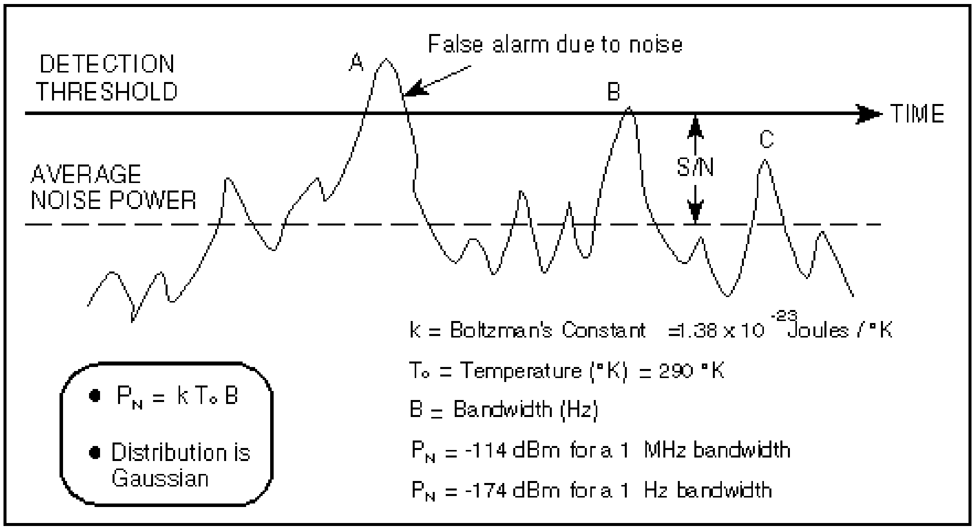
False alarm due to noise
A
DETECTION
THRESHOLD
B
TIME
C
S/N
AVERAGE
NOISE POWER
• $P_N = kT_oB$
• Distribution is Gaussian
k = Boltzman's Constant =1.38 x 10^{-23} Joules / °K
T_o = Temperature (°K) = 290 °K
B = Bandwidth (Hz)
P_N = -114 dBm for a 1 MHz bandwidth
P_N = -174 dBm for a 1 Hz bandwidth

Figure 1. Receiver Noise Power at Room Temperature.

The acceptable minimum Signal-to-Noise ratio (or think of it as Signal **above** Noise) for a receiver depends on the intended use of the receiver. For instance, a receiver that had to detect a single radar pulse would probably need a higher minimum S/N than a receiver that could integrate a large number of radar pulses (increasing the total signal energy) for detection with the same probability of false alarms. Receivers with human operators using a video display may function satisfactorily with low minimum S/N because a skilled operator can be very proficient at picking signals out of a noise background. As shown in Table 1, the setting of an acceptable minimum S/N is highly dependent on the required characteristics of the receiver and of the signal.

Table 1. Typical Minimum S/N Required.

Skilled Operator	Auto-Detection	Auto-detection with Amplitude, TOA, and Frequency Measurements	AOA Phase Interferometer	AOA Amplitude Comparison
3 to 8 dB	10 to 14 dB	14 to 18 dB	14 to 18 dB	16 to 24 dB

A complete discussion of the subject would require a lengthy dissertation of the probability and statistics of signal detection, which is beyond the scope of this handbook, however a simplified introduction follows. Let's assume that we have a receiver that we want a certain probability of detecting a single pulse with a specified false alarm probability. We can use Figure 2 to determine the required signal-to-noise ratio.

S/N EXAMPLE

If we are given that the desired probability of detecting a single pulse (P_d) is 98%, and we want the false alarm rate (P_n) to be no more than 10^{-3}, then we can see that S/N must be 12 dB (see Figure 2).

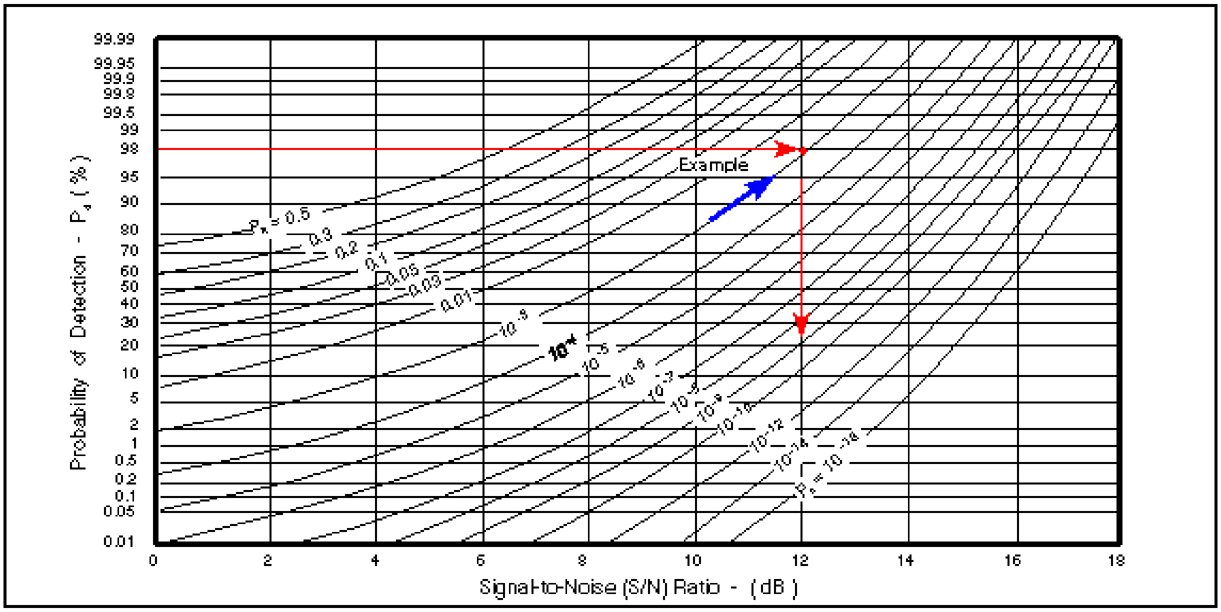

Figure 2. Nomograph of Signal-to-Noise (S/N) Ratio as a Function of Probability of Detection (P_d) and Probability of False Alarm Rate (P_n).

MAXIMUM DETECTION RANGE (ONE-WAY)

From Section 4-3, the one way signal strength from a transmitter to a receiver is:

$$S\ (or\ P_R) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 R^2}$$

For calculations involving receiver sensitivity the "S" can be replaced by S_{min}. Since $S_{min} = (S/N)_{min}$ $kT_oB(NF)$, given by equation [1], the one-way radar equation can be solved for any of the other variables in terms of receiver parameters. In communication, radar, and electronic warfare applications, you might need to solve for the maximum range (R_{max}) where a given radar warning receiver could detect a radiated signal with known parameters. We would then combine and rearrange the two equations mentioned to solve for the following one-way equation:

$$R_{max} \cong \sqrt{\frac{P_t\ G_t\ G_r\ \lambda^2}{(4\pi)^2 (S/N)_{min}\, k\, T_o\, B\, (NF)}} \quad or \quad \sqrt{\frac{P_t\ G_t\ G_r\ c^2}{(4\pi f)^2 (S/N)_{min}\, k\, T_o\, B\, (NF)}} \quad or \quad \sqrt{\frac{P_t\ G_t\ A_e}{4\pi (S/N)_{min}\, k\, T_o\, B\, (NF)}} \qquad \textbf{[3]}$$

We could use standard room temperature of 290° K as T_o, but NF would have to be determined as shown later.

In this calculation for receiver R_{max} determination, P_t, G_t, and λ are radar dependent, while G_r, S/N_{min}, NF, and B are receiver dependent factors.

Equation [3] relates the maximum detection range to bandwidth (B). The effects of the measurement bandwidth can significantly reduce the energy that can be measured from the peak power applied to the receiver input. Additional bandwidth details are provided in Sections 4-4, 4-7, and in other parts of this section.

NOISE POWER, kT_oB

Thermal noise is spread more or less uniformly over the entire frequency spectrum. Therefore the amount of noise appearing in the output of an ideal receiver is proportional to the absolute temperature of the receiver input system (antenna etc) times the bandwidth of the receiver. The factor of proportionality is Boltzmann's Constant.

Mean noise power of ideal receiver = kT_oB = P_N (Watts)
Mean noise power of a real receiver = $(NF)kT_oB$ (Watts)

The convention for the temperature of T_o is set by IEEE standard to be 290°K, which is close to ordinary room temperature. So, assuming T_o = 290°K, and for a bandwidth B = 1 Hz, $kT_oB = 4\text{x}10^{-21}$ W = -204 dBW = -174 dBm.

For any receiver bandwidth, multiply $4\text{x}10^{-21}$ W by the bandwidth in Hz, or if using dB;

$$10 \log kT_oB = -174 \text{ dBm} + 10 \text{ Log (actual BW in Hz)} \quad \text{or} \quad -114 \text{ dBm} + 10 \text{ Log (actual BW in MHz)}$$

and so on, as shown by the values in Table 2.

Table 2. Sample Noise Power Values (kT_oB).

Bandwidth	Bandwidth Ratio (dB)	Watts	dBW	dBm
1 Hz	0	$4\text{x}10^{-21}$	-204	-174
1 kHz	30	$4\text{x}10^{-18}$	-174	-144
1 MHz	60	$4\text{x}10^{-15}$	-144	-114
1 GHz	90	$4\text{x}10^{-12}$	-114	-84

Typical values for maximum sensitivity of receivers would be:

RWR	-65 dBm
Pulse Radar	-94 dBm
CW Missile Seeker	-138 dBm

If antenna contributions are ignored (see note in Table 4) for a CW receiver with a 4 GHz bandwidth, the ideal mean noise power would be -174 dBm + 10 $\text{Log}(4\text{x}10^9)$ = -174 dBm + 96 dB = -78 dBm. A skilled operator might only be able to distinguish a signal 3 dB above the noise floor (S/N=3 dB), or -75 dBm. A typical radar receiver would require a S/N of 3 to 10 dB to distinguish the signal from noise, and would require 10 to 20 dB to track. Auto tracking might require a S/N of approximately 25 dB, thus, a receiver may only have sufficient sensitivity to be able to identify targets down to -53 dBm. Actual pulse receiver detection will be further reduced due to sin x/x frequency distribution and the effect of the measurement bandwidth as discussed in Sections 4-4 and 4-7. Integration will increase the S/N since the signal is coherent and the noise is not.

Noise Bandwidth

Equivalent Noise Bandwidth (B_N) - Set by minimum pulse width or maximum modulation bandwidth needed for the system requirements. A choice which is available to the designer is the relationship of pre- and post-detection bandwidth. Pre-detection bandwidth is denoted by B_{IF}, while post-detection is denoted B_V, where V stands for video. The most affordable approach is to set the post-detection filter equal to the reciprocal of the minimum pulse width, then choose the pre-detection passband to be as wide as the background interference environment will allow. Recent studies suggest that pre-detection bandwidths in excess of 100 MHz will allow significant loss of signals due to "pulse-on-pulse" conditions. Equations [4] and [5] provide B_N relationships that don't follow the Table 3 rules of thumb.

Table 3. Rules of Thumb for B_N a.k.a. B (Doesn't apply for S/N between 0 and 10 to 30 dB).

S/N out	Linear Detector	Square Law Detector
High S/N (>15 to 20 dB)	$B_N = B_V$ (> 20 to 30 dB)	$B_N = 4\ B_V$ (> 10 to 15 dB)
Low S/N (< 0 dB)	$B_N = \sqrt{(2\, B_{IF}\, B_V - B_V^2) / 4\, (S/N)_{out}}$	$B_N = \sqrt{(2\, B_{IF}\, B_V - B_V^2) / (S/N)_{out}}$

For a square law detector: (1)

$$B_N = B_V \left[\ 2 + \sqrt{4 + \frac{(2\, B_{IF} / B_V) - 1}{(S/N)_{out}}}\ \right] \qquad \textbf{[4]}$$

At high $(S/N)_{out}$, the $1/(S/N_{out})$ term goes to zero and we have: $B_N = B_V\ [\ 2 + \sqrt{4}\ ] = 4\ B_V$

At low $(S/N)_{out}$, the $1/(S/N_{out})$ term dominates, and we have:

$$B_N = B_V \sqrt{\frac{(2\, B_{IF} / B_V) - 1}{(S/N)_{out}}} = \sqrt{\frac{2\, B_{IF}\, B_V - B_V^2}{(S/N)_{out}}}$$

For a linear detector: (1)

$$B_N = \frac{B_V}{2} + \frac{1}{4} \bullet \sqrt{B_V \left(4\, B_V + \frac{H^2 (2\, B_{IF} - B_V)}{(S/N)_{out}} \right)} \qquad \textbf{[5]}$$

H is a hypergeometric (statistical) function of $(S/N)_{in}$

$H = 2$ for $(S/N)_{in} << 1$
$H = 1$ for $(S/N)_{in} >> 1$

At high $(S/N)_{out}$, the $1/(S/N_{out})$ term goes to zero and we have: $B_N = \frac{B_V}{2} + \frac{1}{4}\sqrt{B_V\,(4\,B_V)} = B_V$

At low $(S/N)_{out}$, the $1/(S/N_{out})$ term dominates, and we have:

$$B_N = \frac{1}{4} \bullet \sqrt{\frac{B_V H^2 (2 B_{IF} - B_V)}{(S/N)_{out}}} = \sqrt{\frac{2 B_{IF} B_V - B_V^2}{4 (S/N)_{out}}}$$

Note (1): From Klipper, "Sensitivity of Crystal Video Receivers With RF Pre-amplification," *The Microwave Journal*, August 1965.

TRADITIONAL "RULE OF THUMB" FOR NARROW BANDWIDTHS (Radar Receiver Applications)
Required IF Bandwidth for Matched Filter Applications:

$$B_{IF} = \frac{1}{PW_{\min}} \quad \text{Where : } \begin{array}{l} B_{IF} = \text{Pre - detection RF or IF bandwidth} \\ PW_{\min} = \text{Specified minimum pulse width} = \tau \end{array}$$

Matched filter performance gives maximum probability of detection for a given signal level, but: (1) Requires perfect centering of signal spectrum with filter bandwidth, (2) Time response of matched pulse does not stabilize at a final value, and (3) Out-of-band splatter impulse duration equals minimum pulse width. As a result, EW performance with pulses of unknown frequency and pulse width is poor.

$$\begin{array}{c} \text{Required Video Bandwidth Post - Detection} \\ \text{Traditional " Rule of Thumb"} \end{array} \quad B_V = \frac{0.35}{PW_{\min}} \quad \text{Where : } B_V = \text{Post - detection bandwidth}$$

Some authors define B_V in terms of the minimum rise time of the detected pulse, i.e., $B_V = (0.35 \text{ to } 0.5)/t_r$ min, where t_r = rise time.

REVISED "RULE OF THUMB" FOR WIDE BANDWIDTHS (Wideband Portion of RWRs)

$$B_{IF} = \frac{2 \text{ to } 3}{PW_{\min}} \quad \text{and} \quad B_V = \frac{1}{PW_{\min}}$$

The pre-detection bandwidth is chosen based upon interference and spurious generation concerns. The post-detection bandwidth is chosen to "match" the minimum pulse width. This allows (1) Half bandwidth mistuning between signal and filter, (2) Half of the minimum pulse width for final value stabilization, and (3) The noise bandwidth to be "matched" to the minimum pulse width. As a result, there is (1) Improved EW performance with pulses of unknown frequency and pulse width, (2) Measurement of in-band, but mistuned pulses, and (3) Rejection of out-of-band pulse splatter.

NOISE FIGURE / FACTOR (NF)

Electrical noise is defined as electrical energy of random amplitude, phase, and frequency. It is present in the output of every radio receiver. At the frequencies used by most radars, the noise is generated primarily within the input stages of the receiver system itself (Johnson Noise). These stages are not inherently noisier than others, but noise generated at the input and amplified by the receiver's full gain greatly exceeds the noise generated further along the receiver chain. The noise performance of a receiver is described by a figure of merit called the noise figure (NF). The term noise factor is synonymous, with some authors using the term “factor” for numeric and “figure” when using dB notation. (The notation “F_n” is also sometimes used instead of “NF.”) The noise figure is defined as:

$$NF = \frac{Noise\ output\ of\ actual\ receiver}{Noise\ output\ of\ ideal\ receiver} = \frac{N_{out}}{GN_{in}} \quad or\ in\ dB: \quad 10\ \mathrm{Log}\left[\frac{Noise\ output\ of\ actual\ receiver}{Noise\ output\ of\ ideal\ receiver}\right] = 10\ \log \frac{N_{out}}{GN_{in}}$$

A range of NF values is shown in Table 4.

Table 4. Typical Noise Figure / Factor Value.	Decimal	dB
Passive lossy network (RF transmission line, attenuator, etc.) Example: 20 dB attenuator (gain = 0.01)	Same as reciprocal of gain value ex: 100	Same as dB value ex: 20
Solid State Amplifier (see manufacturers specifications)	4	6
Traveling Wave Tube (see manufacturers specifications)	10 to 100	10 to 20
Antennas (Below ≈ 100 MHz, values to 12 dB higher if pointed at the sun) Note: Unless the antenna is pointed at the sun, its negligible NF can be ignored. Antenna gain is not valid for NF calculations because the noise is received in the near field.	1.012 to 1.4	0.05 to 1.5

An ideal receiver generates no noise internally. The only noise in its output is received from external sources. That noise has the same characteristics as the noise resulting from thermal agitation in a conductor. Thermal agitation noise is caused by the continuous random motion of free electrons which are present in every conductor. The amount of motion is proportional to the conductor’s temperature above absolute zero. For passive lossy networks, the noise factor equals the loss value for the passive element:

$$NF = \frac{N_{out}}{G\,N_{in}} = \frac{kTB}{\frac{1}{L}kTB} = L \qquad \begin{array}{c} Where\ L = Ratio\ Value\ of\ Attenuation \\ i.e.\ For\ a\ 3\ dB\ attenuator, G = 0.5\ and\ L = 2 \\ \therefore NF = 2\ \ and\ \ 10\ \log NF = 3\ dB \end{array}$$

A typical series of cascaded amplifiers is shown in Figure 3.

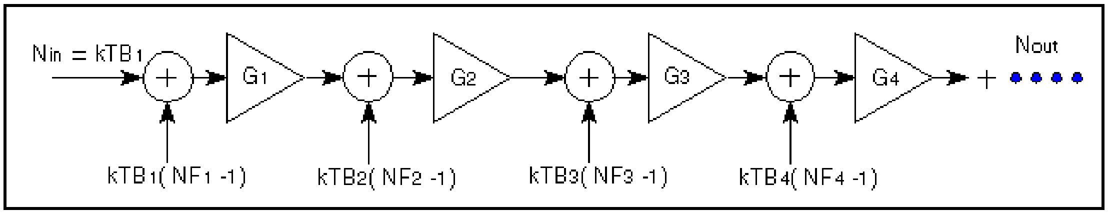

Figure 3. Noise Factors for Cascaded Amplifiers (NF_{CA}).

Loss (negative gain) can be used for the gain value of attenuators or transmission line loss, etc to calculate the noise out of the installation as shown in the following equation:

$$N_{out} = N_{in}\,G\,NF_{CA} = k\,T\,B_1(\,G_1G_2G_3\ldots)\left(NF_1 + \frac{B_2(NF_2 - 1)}{B_1G_1} + \frac{B_3(NF_3 - 1)}{B_1G_1G_2} + \frac{B_4(NF_4 - 1)}{B_1G_1G_2G_3} + \ldots\right) \quad (ratio\ form) \qquad \textbf{[6]}$$

If the bandwidths of the amplifiers are the same, equation [6] becomes:

$$N_{out} = N_{in}\, G\, NF_{CA} = k\, T\, B\, (G_1 G_2 G_3 \ldots)\left(NF_1 + \frac{NF_2 - 1}{G_1} + \frac{NF_3 - 1}{G_1 G_2} + \frac{NF_4 - 1}{G_1 G_2 G_3} + \ldots \right) \quad \text{(ratio form)} \qquad [7]$$

Pre-amplifier Location Affects Receiver Input Noise

As shown in Figure 4, if a 2 to 12 GHz receiver installation doesn't have enough sensitivity, it is best to install an additional amplifier closer to the antenna (case 1) instead of closer to the receiver (case 2). In both cases, the line loss (L) and the amplifier gain (G) are the same, so the signal level at the receiver is the same. For case 1, $S_1 = P_{in} + G - L$. In case 2, $S_2 = P_{in} - L + G$, so $S_1 = S_2$. The noise generated by the passive transmission line when measured at the receiver is the same in both cases. However, the noise generated inside the amplifier, when measured at the receiver input, is different.

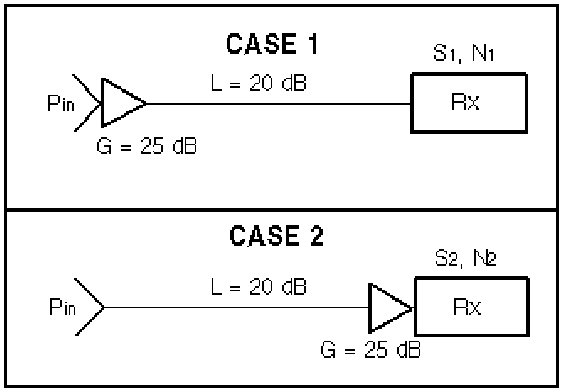

Figure 4. Pre-Amp S/N.

For this example, case 2 has a noise level at the input to the receiver which is 19.7 dB higher than case 1 (calculations follow later).

Table 5a	Case 1 Gain		Case 1 NF	
	Amp	L	Amp	L
dB	25	-20	6 *	20
ratio	316.2	0.01	4 *	100

Table 5b	Case 2 Gain		Case 2 NF	
	L	Amp	L	Amp
dB	-20	25	20	6 *
ratio	0.01	316.2	100	4 *

* Amplifier NF value from Table 4.

Using equation [3] and the data in Tables 5a and 5b, the noise generated by the RF installation is shown in Tables 6a and 6b (the negligible noise contribution from the antenna is the same in both cases and is not included) (also see notes contained in Table 4):

Table 6a. Case 1	**Table 6b.** Case 2
$G(NF) = 316.2\,(0.01)\left(4 + \frac{100 - 1}{316.7}\right) = 13.64$	$G(NF) = 0.01\,(316.2)\left(100 + \frac{4 - 1}{0.01}\right) = 1264.8$
10 log G(NF) = 11.34 dB	10 log G(NF) = 31 dB
Noise at receiver:	
$N_{out\,1}$ = -74 dBm + 11.34 dB = -62.7 dBm	$N_{out\,2}$ = -74 dBm + 31 dB = -43 dBm

$N_{out\,2} - N_{out\,1}$ = 19.7 dB. The input noise of -74 dBm was calculated using 10 log (kTB), where B = 10 GHz.

Note that other tradeoffs must be considered: (1) greater line loss between the antenna and amplifier improves (decreases) VSWR as shown in Section 6-2, and (2) the more input line loss, the higher the input signal can be before causing the pre-amplifier to become saturated (mixing of signals due to a saturated amplifier is addressed in Section 5-7).

Combining Receive Paths Can Reduce Sensitivity

If a single aircraft receiver processes both forward and aft signals as shown in Figure 5, it is desirable to be able to use the receiver's full dynamic range for both directions. Therefore, one needs to balance the gain, so that a signal applied to the aft antenna will reach the receiver at the same level as if it was applied to the forward antenna.

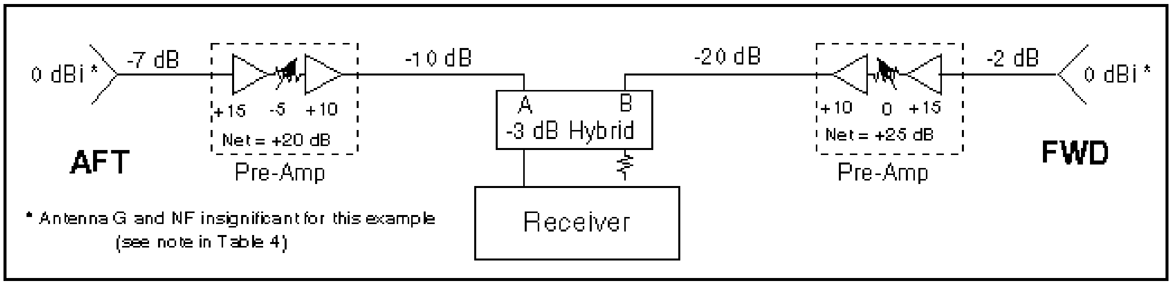

Figure 5. Example of Pre-Amplifier Affecting Overall Gain / Sensitivity.

Common adjustable preamplifiers can be installed to account for the excessive transmission line loss. In this example, in the forward installation, the level of the signal at the receiver is the same as the level applied to the antenna. Since the aft transmission line has 5 dB less attenuation, that amount is added to the preamplifier attenuator to balance the gain. This works fine for strong signals, but not for weaker signals. Because there is less loss between the aft preamplifier and the receiver, the aft noise dominates and will limit forward sensitivity. If the bandwidth is 2-12 GHz, and if port A of the hybrid is terminated by a perfect 50Ω load, the forward noise level would be -65.3 dBm. If port B is terminated, the aft noise level would be -60.4 dBm. With both ports connected, the composite noise level would be -59.2 dBm (convert to mw, add, then convert back to dBm). For this example, if the aft preamplifier attenuation value is changed to 12 dB, the gain is no longer balanced (7 dB extra loss aft), but the noise is balanced, i.e., forward = -65.6 dBm, aft = -65.3 dBm, and composite -62.4 dBm. If there were a requirement to see the forward signals at the most sensitive level, extra attenuation could be inserted in the aft preamplifier. This would allow the forward noise level to predominate and result in greater forward sensitivity where it is needed. Calculations are provided in Tables 7 and 8.

Table 7. Summary of Gain and NF Values for Figure 5 Components.

		Aft					Fwd				
		RF Line	Amp	Attn	Amp	RF Line & hybrid	RF Line	Amp	Attn	Amp	RF Line & hybrid
Gain	dB	-7	15	-5	10	-13	-2	15	0	10	-23
	ratio	0.2	31.6	0.32	10	0.05	0.63	31.6	0	10	0.005
NF	dB	7	6	5	6	13	2	6	0	6	23
	ratio	5	4	3.16	4	20	1.585	4	0	4	200

Aft NF = 22.79 therefore 10 log NF = 13.58 dB. Input noise level = -74 dBm + 13.58 dB = -60.42 dBm ≅ -60.4 dBm
Fwd NF = 7.495 therefore 10 log NF = 8.75 dB. Input noise level = -74 dBm + 8.75 dB = -65.25 dBm ≅ -65.3 dBm
The composite noise level at the receiver = -59.187 dBm ≅ -59.2 dBm.

Table 8. Effect of Varying the Attenuation (shaded area) in the Aft Preamplifier Listed in Table 7.

Aft Attn NF	Aft Attn Gain	Aft Noise	Fwd Noise	Composite Noise	Min Signal Received ***	Aft Input	Fwd Input
0 dB	0 dB	-55.8 dBm	-65.3 dBm	-55.4 dBm	-43.4 dBm	-48.4 dBm	-43.4 dBm
5	-5	-60.4	-65.3	-59.2	-47.2 *	-47.2 *	-47.2 *
10	-10	-64.4	-65.3	-61.8	-49.8	-44.8	-49.8
12	-12	-65.6 **	-65.3 **	-62.4	-50.4	-43.4	-50.4
15	-15	-67.1	-65.3	-63.1	-51.1	-41.1	-51.1

* Gain Balanced ** Noise Balanced *** S/N was set at 12 dB

TANGENTIAL SENSITIVITY

Tangential sensitivity (TSS) is the point where the top of the noise level with no signal applied is level with the bottom of the noise level on a pulse as shown in Figure 6. It can be determined in the laboratory by varying the amplitude of the input pulse until the stated criterion is reached, or by various approximation formulas.

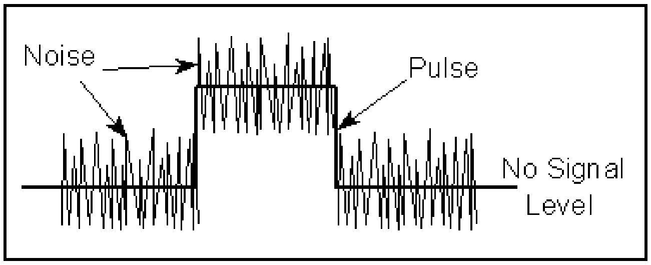

Figure 6. Tangential Sensitivity.

The signal power is nominally 8±1 dB above the noise level at the TSS point. TSS depends on the RF bandwidth, the video bandwidth, the noise figure, and the detector characteristic.

TSS is generally a characteristic associated with receivers (or RWRs), however the TSS does not necessarily provide a criterion for properly setting the detection threshold. If the threshold is set to TSS, then the false alarm rate is rather high. Radars do not operate at TSS. Most require a more positive S/N for track (> 10 dB) to reduce false detection on noise spikes.

SENSITIVITY CONCLUSION

When all factors effecting system sensitivity are considered, the designer has little flexibility in the choice of receiver parameters. Rather, the performance requirements dictate the limit of sensitivity which can be implemented by the EW receiver.

1. Minimum Signal-to-Noise Ratio (S/N) - Set by the accuracy which you want to measure signal parameters and by the false alarm requirements.

2. Total Receiver Noise Figure (NF) - Set by available technology and system constraints for RF front end performance.

3. Equivalent Noise Bandwidth (B_N) - Set by minimum pulse width or maximum modulation bandwidth needed to accomplish the system requirements. A choice which is available to the designer is the relationship of pre- (B_{IF}) and post-detection (B_V) bandwidth. The most affordable approach is to set the post-detection filter equal to the reciprocal of the minimum pulse width, then choose the pre-detection passband to be as wide as the background interference environment will allow. Recent studies suggest that pre-detection bandwidths in excess of 100 MHz will allow significant loss of signals due to "pulse-on-pulse" conditions.

4. Antenna Gain (G) - Set by the needed instantaneous FOV needed to support the system time to intercept requirements.

This page intentionally left blank.

RECEIVER TYPES AND CHARACTERISTICS

Besides the considerations of noise and noise figure, the capabilities of receivers are highly dependent on the type of receiver design. Most receiver designs are trade-offs of several conflicting requirements. This is especially true of the Electronic Support Measures (ESM) receivers used in Electronic Warfare.

This section consists of a figure and tables that provide a brief comparison of various common ESM receiver types. Figures 1 and 2 show block diagrams of common ESM receivers. Table 1 is a comparison of major features of receivers. Table 2 shows the receiver types best suited for various types of signals and Tables 3 and 4 compare several direction of arrival (DOA) and emitter location techniques. Table 5 shows qualitative and quantitative comparisons of receiver characteristics.

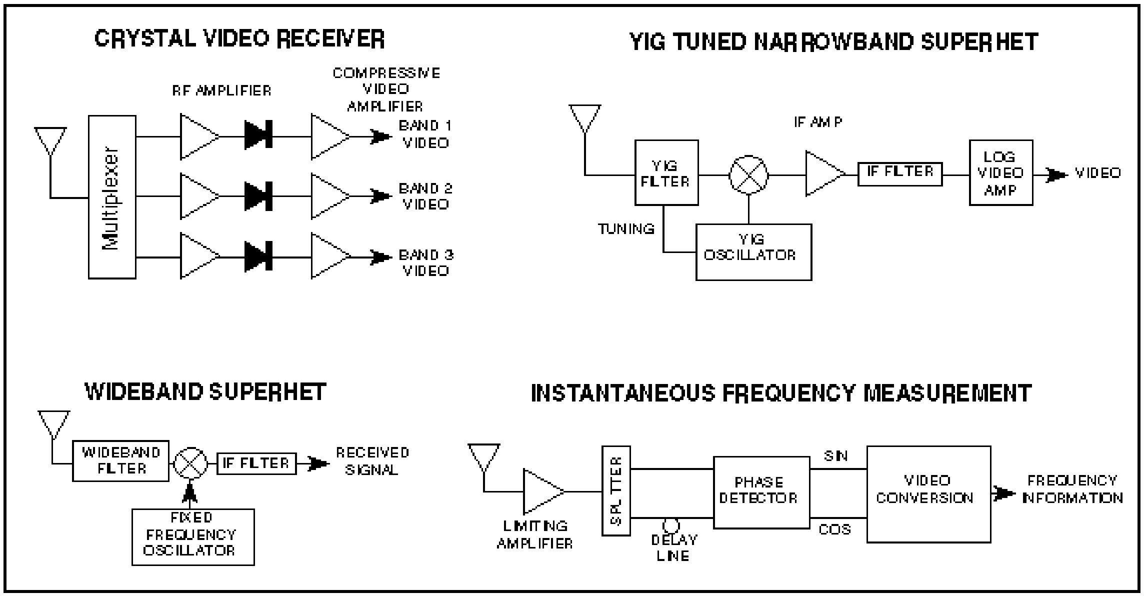
CRYSTAL VIDEO RECEIVER
RF AMPLIFIER
COMPRESSIVE VIDEO AMPLIFIER
Multiplexer
BAND 1 VIDEO
BAND 2 VIDEO
BAND 3 VIDEO
YIG TUNED NARROWBAND SUPERHET
IF AMP
YIG FILTER
IF FILTER
LOG VIDEO AMP
VIDEO
TUNING
YIG OSCILLATOR
WIDEBAND SUPERHET
WIDEBAND FILTER
IF FILTER
RECEIVED SIGNAL
FIXED FREQUENCY OSCILLATOR
INSTANTANEOUS FREQUENCY MEASUREMENT
LIMITING AMPLIFIER
SPLITTER
DELAY LINE
PHASE DETECTOR
SIN
COS
VIDEO CONVERSION
FREQUENCY INFORMATION

Figure 1. Common ESM Receiver Block Diagrams.

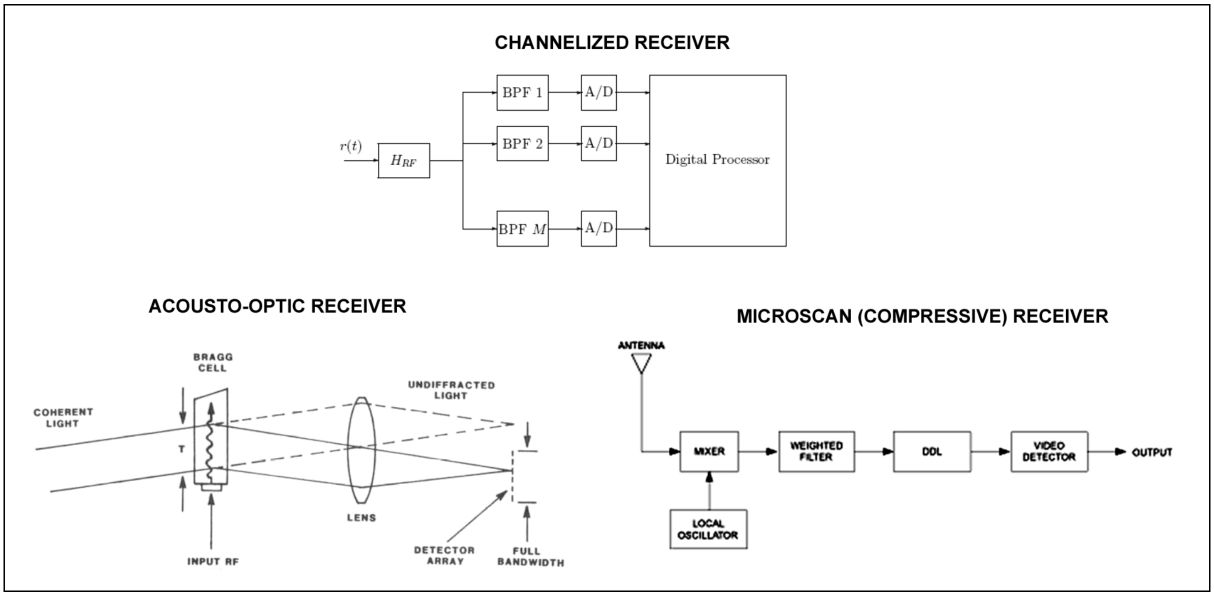
CHANNELIZED RECEIVER
$r(t)$
H_{RF}
BPF 1
BPF 2
BPF M
A/D
A/D
A/D
Digital Processor
ACOUSTO-OPTIC RECEIVER
BRAGG CELL
COHERENT LIGHT
T
UNDIFFRACTED LIGHT
LENS
INPUT RF
DETECTOR ARRAY
FULL BANDWIDTH
MICROSCAN (COMPRESSIVE) RECEIVER
ANTENNA
MIXER
WEIGHTED FILTER
DDL
VIDEO DETECTOR
OUTPUT
LOCAL OSCILLATOR

Figure 2. Common ESM Receiver Block Diagrams (Continued).

Table 1. Comparison of Major Features of Receivers.

Receiver	Advantages	Disadvantages	Principal Applications
Wideband crystal video	Simple, inexpensive, instantaneous, High POI in frequency range	No frequency resolution Poor sensitivity and Poor simultaneous signal performance	RWR
Tuned RF Crystal Video	Simple, Frequency measurement Higher sensitivity than wideband	Slow response time Poor POI	Option in RWR, Frequency measurement in hybrid
IFM	Relatively simple Frequency resolution Instantaneous, high POI	Cannot sort simultaneous signals Relatively poor sensitivity	Shipboard ESM, Jammer power management, SIGINT equipment
Narrow-band scanning Superhet	High sensitivity Good frequency resolution Simultaneous signals don't interfere	Slow response time Poor POI Poor against frequency agility	SIGINT equipment Air and ship ESM Analysis part of hybrid
Wide-band Superhet	Better response time and POI	Spurious signals generated Poorer sensitivity	Shipboard ESM Tactical air warning
Channelized	Wide bandwidth, Near instantaneous, Moderate frequency resolution	High complexity, cost; Lower reliability; limited sensitivity	SIGINT equipment Jammer power management
Microscan	Near instantaneous, Good resolution and dynamic range, Good simultaneous signal capability	High complexity, Limited bandwidth No pulse modulation information Critical alignment	SIGINT equipment Applications for fine freq analysis over wide range
Acousto-optic	Near instantaneous, Good resolution, Good simultaneous signal capability Good POI	High complexity; new technology	

Note: The Microscan receiver is also known as a compressive receiver

Table 2. Receiver Types vs. Signal Types.

Signal Type	Receiver Type							
	Wide-Band Crystal Video	TRF Crystal Video	IFM	Narrow-Band Superhet	Wide-Band Superhet	Channelized	Microscan	Acousto-optic
CW	Special design for CW	Special design for CW	Yes, but interferes with pulsed reception	Yes	Yes	Yes	Yes	Yes
Pulsed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Frequency	No	No	No	Yes, but won't recognize as same source	No	Yes	Yes	Yes
Frequency Agile	Yes, doesn't measure frequency	No	Yes	No	Yes (within passband)	Yes	Yes	No/Yes, depending on readout time
PRI Agile	Yes	Yes	Yes	No/Yes, depending on scan rate	Yes	Yes	No/Yes, imprecision in TOA	No/Yes, depending on readout time
Chirped	Yes, within acceptance BW	No	Yes	No/Yes, depending on BW	Yes	Yes (reduced sensitivity)	No/Yes, depending on scan rate	Yes (reduced sensitivity)
Spread Spectrum	Yes, within acceptance BW	No	Yes	No	No/Yes, depending on BW	Yes (reduced sensitivity)	Yes (reduced sensitivity)	Yes (reduced sensitivity)

Table 3. Direction of Arrival Measurement Techniques.

	Amplitude Comparison	Phase Interferometer
Sensor Configuration	Typically 4 to 6 Equal Spaced Antenna Elements for 360° Coverage	2 or more RHC or LHC Spirals in Fixed Array
DF Accuracy	$DF_{ACC} \approx \frac{\Theta_{bW}^{2} \, \Delta C_{dB}}{24\,S}$ (Gaussian Antenna Shape)	$DF_{ACC} \approx \frac{\lambda}{2\,\pi\,d\,\cos\theta}\Delta\theta$
DF Accuracy Improvement	Decrease Antenna BW; Decrease Amplitude Mistrack; Increase Squint Angle	Increase Spacing of Outer Antennas; Decrease Phase Mistrack
Typical DF Accuracy	3° to 10° rms	0.1° to 3° rms
Sensitivity to Multipath/Reflections	High Sensitivity; Mistrack of Several dB Can Cause Large DF Errors	Relatively Insensitive; Interferometer Can be Made to Tolerate Large Phase Errors
Platform Constraints	Locate in Reflection Free Area	Reflection Free Area; Real Estate for Array; Prefers Flat Radome
Applicable Receivers	Crystal Video; Channelizer; Acousto-Optic; Compressive; Superheterodyne	Superheterodyne
ΔC_{dB}= Amplitude Monopulse Ratio in dB S= Squint Angle in degrees θ_{BW}= Antenna Beamwidth in degrees		

Table 4. Emitter Location Techniques.

Measurement Technique	Advantages	Disadvantages
Triangulation	Single Aircraft	Non-instantaneous location Inadequate accuracy for remote targeting Not forward looking
Azimuth/elevation	Single Aircraft Instantaneous location possible	Accuracy degrades rapidly at low altitude Function of range
Time Difference of Arrival (Pulsed signals)	Very high precision Can support weapon delivery position requirements Very rapid, can handle short on-time threat	Very complex, diverse systems required, at least 3 aircraft High quality receivers, DME (3 sites) very wideband data link Very high performance control processor; requires very high reliability subsystems

Table 5. Qualitative Comparison of Receivers. (From NRL Report 8737)

Feature	Receiver Type							
	Wide-Band Crystal Video	TRF Crystal Video	IFM	Narrow-Band Superhet	Wide-Band Superhet	Channelized	Microscan	Acousto-optic
Instantaneous Analysis Bandwidth	Very wide	Narrow	Very wide	Narrow	Moderate	Wide	Wide	Moderate
Frequency Resolution	Very poor	Fair	Good	Very good	Poor	Fair	Good	Good
Sensitivity	Poor (No preamp) Fair (preamp)	Fair/ good	Poor (No preamp) Fair (preamp)	Very good	Fair	Fair/ good	Very good	Good
Dynamic Range	Fair	Fair/ good	Good	Very good	Fair	Good	Fair	Poor
Speed of Acquisition	Very Fast	Slow	Very Fast	Slow	Fast	Very Fast	Very Fast	Fast
Short pulse Width Capability	Good	Good	Good	Good	Very good	Good	Fair	Fair
Retention of Signal Character-istics	Fair	Fair	Poor	Good	Fair/ good	Good	Poor	Fair/ good
Applicability to Exotic Signals	Poor/ fair	Poor	Good	Poor	Fair/ good	Good	Fair/ good	Fair/ good
High signal Density Performance	Poor (high false alarm rate from background)	Fair/ good	Good	Poor	Fair (depending on BW)	Fair/good, depending on architecture & processing	Good	Poor
Simultaneous Signal Capability	Poor	Fair/ good	Poor	Good	Fair (depending on BW)	Good	Good	Good
Processing Complexity	Moderate depending on application	Moderate depending on application	Moderate	Moderate	Moderate	Low-high depending on architecture	Complex	Simple signal processing complex data processing
Immunity to Jamming	Poor	Fair	Poor/ Fair	Good	Poor/ Fair	Good	Good	Good
Power Requirements	Low	Low/ Moderate	Moderate	Moderate	Moderate	High	Moderate	Moderate/ High
RF Range (GHz)	Multi-octave (0.5-40)	0.15-18 separate	>0.5 to 40	<0.01 to 40	0.5 to 18	0.5 to 60	<0.5 to 8	0.5-4 (0.5-18 channelized and down conversion)
Max Instantane-ous Analysis Bandwidth	Multi-octave (to 17.5 GHz)	As high as desired with equivalent reduction in resolution	Multi-octave (1 octave per unit)	50 MHz	500 MHz	~2 GHz without degradation, 17.5 GHz with degradation	0.5 to 2 depending on PW limitation	1 GHz
Frequency Accuracy	Measurement accuracy no better than analysis BW	Measurement accuracy no better than analysis BW	5-10 MHz	0.5% to 1%	0.5 to 3 MHz	±1 MHz	10 KHz	±1 MHz

Feature	Receiver Type							
	Wide-Band Crystal Video	TRF Crystal Video	IFM	Narrow-Band Superhet	Wide-Band Superhet	Channelized	Microscan	Acousto-optic
Pulse Width Range	CW to 50 ns	CW to 50 ns	CW to ~20 ns (depending on resolution)	CW to 100 ns with 20 MHz resolution	CW to 4 ns with 500 MHz resolution	CW to 30 ns (depending on resolution)	CW to 250 ns	CW to 0.5 µs
Frequency Resolution	~400 MHz (no better than BW)	25 MHz	1 MHz	<0.1 MHz	100-500 MHz	10-125 MHz (less with freq vernier)	1 MHz	0.5 to 1 MHz
Sensitivity (dBm)	-40 to -50 (no preamp) -80 (with preamp)	Better than -80 with preamp	-40 (no preamp) -75 (preamp) 4 GHz BW	-90, 1 MHz BW	-80, 500 MHz BW	-70, 10-50 MHz BW	-90, 5-10 MHz BW	-70 to -80
Maximum Dynamic Range (dB)	70	70-80	80 (w/preamp) 100+ (saturated)	90	60	50-80	40-60	25-35
Tuning Time	-	50 ms	-	1.0 s (1 octave)	.12 s (200 MHz band)	-	0.3 µs LO scan time	0.5 ms (integration time)
Signal ID Time	100 ns	50 ms	2-10 ms	~0.1 s	-	2.10 ms	~1 µs	-
Minimum Weight (lb)	20 (with processor)	30	<20 (octave unit) 65-75 (full coverage)	60-75	35 (tuner only)	1309-200 for 0.5 to 18 GHz coverage	25	29-55
Size / Minimum Volume (in³)	Small 300 (w/processor)	Small 375	Sm/Moderate 600-1000 ~100 miniaturized	Moderate 1500-3000	Moderate Several thousand	Large 4000-8000 (0.5-18 GHz coverage	Moderate 1200-2000	Small 800-1900
Minimum Power (W)	100 (with processor) <10 without processor	60 (without processor)	~50 (octave unit)	150	150 (tuner only)	350 to 1200 for 0.5 to 18 GHz coverage	70-80	200
Cost	Low	Low/ Moderate	Moderate	Moderate/ High	Moderate/ High	High	Moderate/ High	Low/ Moderate

This page intentionally left blank.

RADAR MODES

Typical Radar modes are listed below in the general functional category for which they were designed. Not all of these modes are applicable to all radars and certain radars have additional modes.

• NAVIGATION

Terrain avoidance - A mode in which the radar is set at a fixed depression angle and short range to continuously sweep the ground area directly in front of the aircraft in order to avoid mountains. This is particularly useful during flight into unfamiliar territory when clouds, haze, or darkness obscure visibility.

Ground mapping - A mode in which the radar uses a variety of techniques to enhance ground features, such as rivers, mountains, and roads. The mode is unlike air-to-air modes where ground return is rejected from the display.

Precision velocity update / Doppler navigation - A mode in which the radar again tracks ground features, using Doppler techniques, in order to precisely predict aircraft ground speed and direction of motion. Wind influences are taken into account, such that the radar can also be used to update the aircraft inertial navigation system.

• FIGHTER MISSIONS

Pulse search - Traditional pulse techniques are used to accurately determine range, angle, and speed of the target. Limitations are easy deception by enemy jamming, and less range when compared to other modes.

Velocity search - A high PRF Pulse Doppler waveform is used for long range detection primarily against nose aspect targets, giving velocity and azimuth information. Although velocity search can work against tail-on targets, the Doppler return is weaker, consequently the maximum detection range is also much less. When the target is in the beam (flying perpendicular to the fighter), the closure (Doppler) is the same as ground return and target return is almost zero.

Track While Scan (TWS) - A system that maintains an actual track on several aircraft while still searching for others. Since the radar is sharing its computing time between targets, the accuracy is less precise than for a single target track (STT) mode of operation.

Raid assessment - A mode in which the radar has an STT on a single target, but is routinely driven off by a small amount in order to determine if multiple aircraft exists in the immediate vicinity of the target aircraft.

Single-Target-Track (STT) (including air combat maneuvering modes) - Highly precise STT modes are used to provide the most accurate information to the fire control computer so that accurate missile or gun firing can be accomplished. The fire control radar continuously directs energy at the target so that the fired missile locates and tracks on the reflected energy from the target. Air combat maneuvering modes are automatic modes in which the radar has several sweep patterns fixed about the aircraft axis, such that little or no work is required of the pilot in order to lock up a target.

- **AIR-TO-GROUND MISSIONS**

Weapons delivery - A mode in which ground features are tracked, and particular emphasis is placed on determining range to the ground target, angle of dive, weapons ballistic tables, and aircraft speed.

Surveillance/tracking of ground forces/targets - Similar to the above with emphasis on multiple ground features and less on weapons delivery data.

Reconnaissance - A specific navigational mode to aid in identifying specific targets.

- **AIR-TO-SURFACE MISSIONS**

ASW - Navigational techniques specializing in specific search patterns to aid in detection of enemy submarines.

- **TECHNIQUES USED FOR MULTIPLE APPLICATIONS**

Synthetic Aperture Radar (SAR) - A form of radar that uses the relative motion between an antenna and its target region, to provide coherent-signal variations, in order to obtain finer spatial resolution than is possible with conventional beam-scanning means. SAR is usually implemented by mounting a single beam-forming antenna on a moving platform such as an aircraft from which a target scene is repeatedly illuminated with pulses of radio waves at wavelengths anywhere from a meter down to millimeters. The many echo waveforms received successively at the different antenna positions are coherently detected and stored and then post-processed together to resolve elements in an image of the target region.

Over-The-Horizon Radar (OTHR) - uses the refraction of high frequency radiation through the ionosphere in order to detect targets beyond the line-of-sight. The complexities of the ionosphere can produce multipath propagation, which may result in multiple resolved detections for a single target. When there are multipath detections, an OTHR tracker will produce several spatially separated tracks for each target. Information conveying the state of the ionosphere is required in order to determine the true location of the target and is available in the form of a set of possible propagation paths, and a transformation from measured coordinates into ground coordinates for each path. Since may be no other information as to how many targets are in the surveillance region, or which propagation path gave rise to which track, there is a joint target and propagation path association ambiguity which must be resolved using the available track and ionospheric information.

GENERAL RADAR DISPLAY TYPES

There are two types of radar displays in common use today.

RAW VIDEO

Raw video displays are simply oscilloscopes that display the detected and amplified target return signal (and the receiver noise). Raw video displays require a human operator to interpret the various target noise and clutter signals.

On the left hand display of Figure 1, an operator could readily identify three targets and a ghost (a ghost is a phony target that usually fades in and out and could be caused by birds, weather, or odd temporary reflections - also referred to as an angel). Target 3 is a weak return and hidden in the noise - an operator can identify it as a target by the "mouse under the rug" effect of raising the noise base line.

SYNTHETIC VIDEO

Synthetic video displays use a computer to clean up the display by eliminating noise and clutter and creating its own precise symbol for each target.

On the right hand display target 1 comes and goes because it is barely above the receiver noise level - notice that it is quite clear on the raw video. Target 3 wasn't recognized by the computer because it's too far down in the noise. The computer validated the ghost as a target. The ghost might be a real target with glint or ECM characteristics that were recognized by the computer but not the operator.

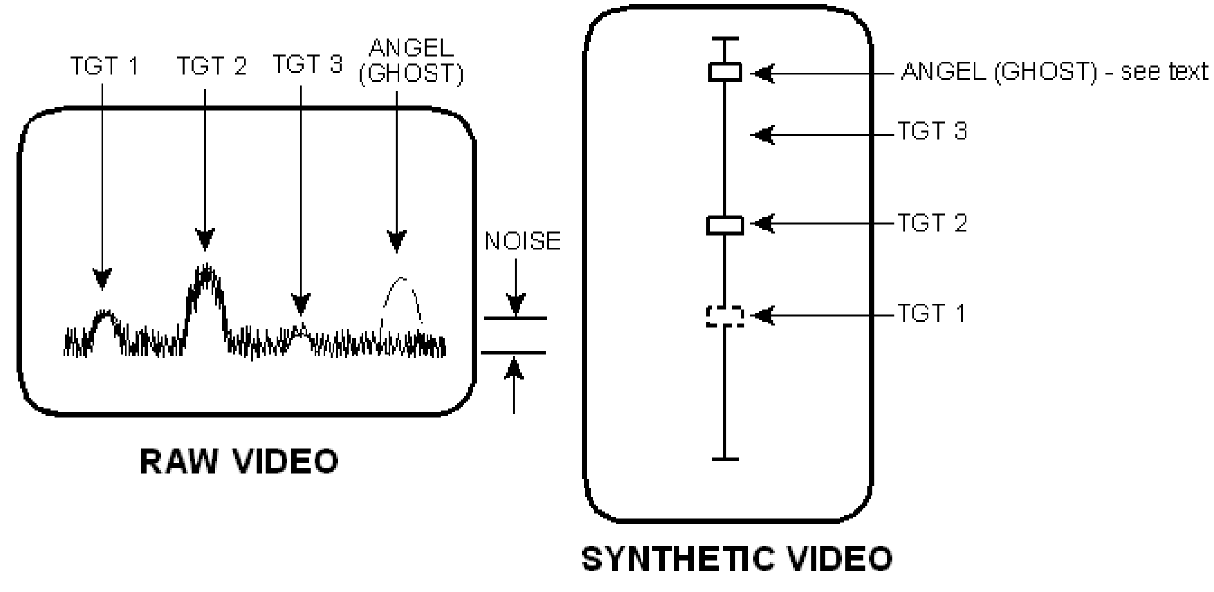

Figure 1. Radar Display Types.

SEARCH AND ACQUISITION RADARS

They generally use either a PPI or a sector PPI display as shown in Figure 2. PPI displays can be either raw video or synthetic video.

PPI scope (plan position indicator).
    Polar plot of direction and distance.
    Displays all targets for 360 degrees.

Sector PPI scope.
    Polar plot of direction and distance.
    Displays all targets within a specific sector.
    Origin may be offset so that “your” radar position may be off the scope.

TRACKING RADARS

Usually use some combination of A, B, C, or E scope displays. There are many other types of displays that have been used at one time or another - including meters - but those listed here are the most common in use today.

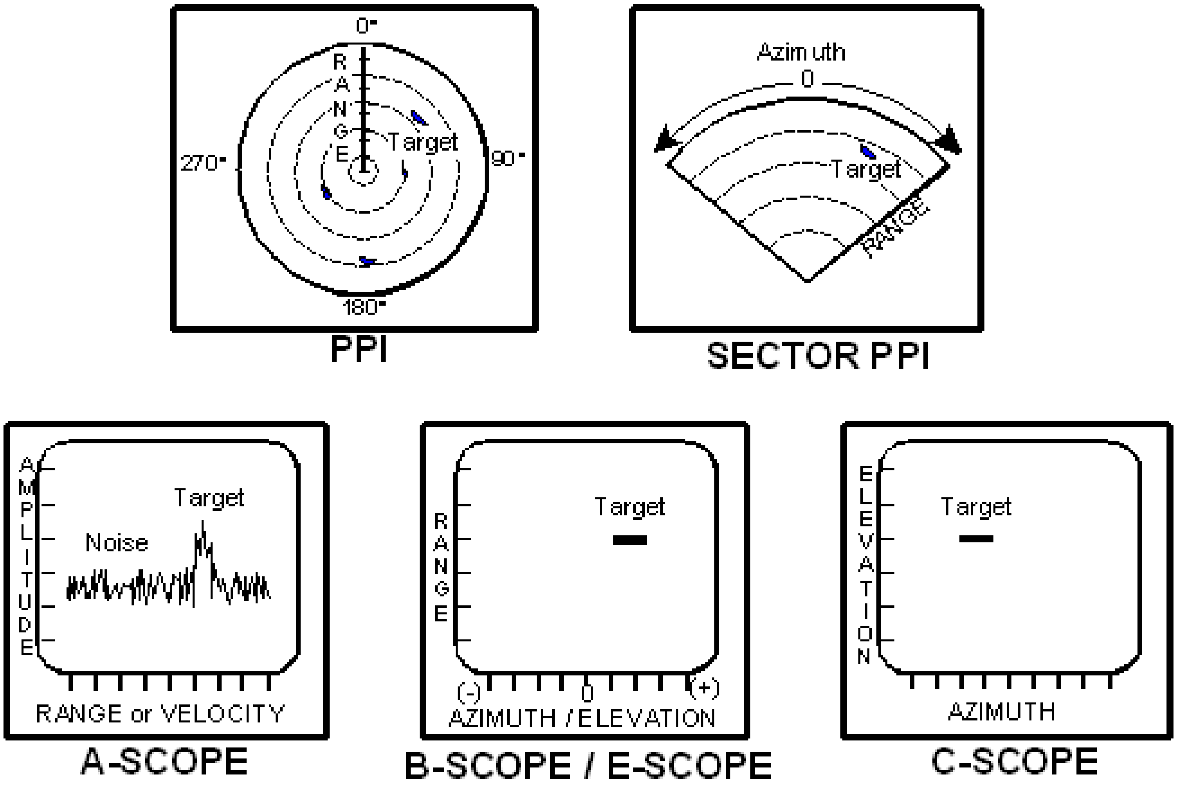
0°
RANGE
Target
270°
90°
180°
PPI
Azimuth
0
Target
RANGE
SECTOR PPI
AMPLITUDE
Target
Noise
RANGE or VELOCITY
A-SCOPE
RANGE
Target
(-)
0
(+)
AZIMUTH / ELEVATION
B-SCOPE / E-SCOPE
ELEVATION
Target
AZIMUTH
C-SCOPE

Figure 2. Common Radar Displays.

A-SCOPE

Target signal amplitude vs. range or velocity.

Displays all targets along pencil beam for selected range limits.

Displays tracking gate. Usually raw video. Some modern radars have raw video a-scopes as an adjunct to synthetic video displays.

Must be used with a separate azimuth and elevation display of some sort.

Also called a range scope (R-Scope).

B-SCOPE

Range vs. azimuth or elevation. Displays targets within selected limits.

Displays tracking gate. May be raw or synthetic video.

Surface radars usually have two. One azimuth/one elevation which can result in confusion with multiple targets.

C-SCOPE

Azimuth vs. elevation. Displays targets within selected limits of az and el.

Displays tracking gate. May display bull's-eye or aim dot.

May have range indicator inserted typically as a marker along one side. Usually synthetic video.

Pilots eye view and very common in modern fighter aircraft heads up displays for target being tracked.

Could be used in any application where radar operator needs an "aiming" or "cross hair" view like a rifle scope.

E-SCOPE

Elevation vs. Range similar to a B-scope, with elevation replacing azimuth.

This page intentionally left blank.

IFF - IDENTIFICATION - FRIEND OR FOE

Originated in WWII for just that purpose - a way for our secondary radars to identify U.S. aircraft from enemy aircraft by assigning a unique identifier code to U.S. aircraft transponders.

The system is considered a secondary radar system since it operates completely differently and independently of the primary radar system that tracks aircraft skin returns only, although the same CRT display is frequently used for both.

The system was initially intended to distinguish between enemy and friend but has evolved such that the term "IFF" commonly refers to all modes of operation, including civil and foreign aircraft use.

There are five major modes of operation currently in use by military aircraft plus two sub-modes.

- Mode 1 is a non-secure low cost method used by ships to track aircraft and other ships.
- Mode 2 is used by aircraft to make carrier controlled approaches to ships during inclement weather.
- Mode 3 is the standard system also used by commercial aircraft to relay their position to ground controllers throughout the world for air traffic control (ATC).
- Mode 4 is secure encrypted IFF (the only true method of determining friend or foe) Military only
- Mode 5 – provides a cryptographically secured version of Mode S and ADS-B GPS position. (military only). Mode 5 is divided into two levels. Both are crypto-secure with Enhanced encryption, Spread Spectrum Modulation, and Time of Day Authentication. Level 1 is similar to Mode 4 information but enhanced with an Aircraft Unique PIN. Level 2 is the same as Mode 5 level one but includes additional information such as aircraft position and other attributes
- Mode "C" is the altitude encoder (military and civilian).
- Mode S – provides multiple information formats to a selective interrogation. Typically aircraft are assigned a unique 24-bit Mode S address. The Mode S address is partitioned and a group of address ranges are allocated to each country. Some countries change the assigned address for security reasons, and thus it might not be a unique address. (military and civilian)

The non-secure codes are manually set by the pilot but assigned by the air traffic controller.

A cross-band beacon is used, which simply means that the interrogation pulses are at one frequency and the reply pulses are at a different frequency. 1030 MHz and 1090 MHz is a popular frequency pair used in the U.S.

The secondary radar transmits a series of selectable coded pulses. The aircraft transponder receives and decodes the interrogation pulses. If the interrogation code is correct, the aircraft transponder transmits a different series of coded pulses as a reply.

The advantage of the transponder is that the coded pulses "squawked" by the aircraft transponders after being interrogated might typically be transmitted at a 10 watt ERP, which is much stronger than the microwatt skin return to the primary radar. Input power levels may be on the order of several hundred watts.

The transponder antenna is low gain so that it can receive and reply to a radar from any direction.

An adjunct to the IFF beacon is the altitude encoding transponder known as mode C - all commercial and military aircraft have them, but a fair percentage of general aviation light aircraft do not because of cost. The number of transponder installations rises around many large metropolitan areas where they are required for safety (easier identification of aircraft radar tracks).

Air traffic control primary radars are similar to the two-dimensional search radar (working in azimuth and range only) and cannot measure altitude.

The expanded display in figure 1 is typical of an air traffic control IFF response. The aircraft was told to squawk a four digit number such as "4732." The altitude encoded transponder provides the aircraft altitude readout to the ground controllers display along with the coded response identifying that particular aircraft.

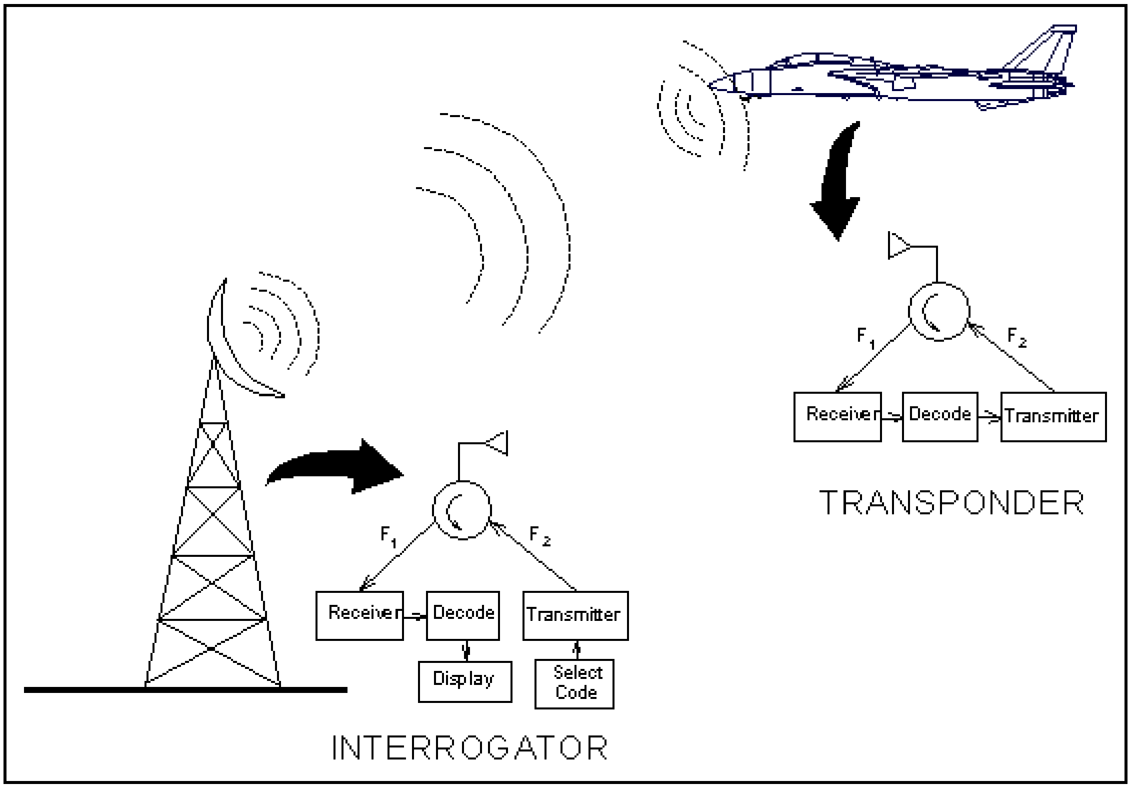

Figure 1. IFF Transponder.

In addition to systems with active electronic data interchange between airborne and ground equipment, some military surveillance systems can provide targeting in tactical applications. The development of automated techniques for use against ground targets is typically referred to as Automatic Target Recognition (ATR). When used against air targets, it is typically referred to as Non-Cooperative Target Recognition (NCTR). The requirements for radar target recognition are complex since typical targets have background clutter and often multiple targets types exist.

RECEIVER TESTS

Two tone and spurious response (single signal) receiver tests should be performed on EW and radar receivers to evaluate their spurious free dynamic range. A receiver should have three ranges of performance: (1) protection from damage, (2) degraded performance permitted in the presence of a strong interfering signal(s) and no degradation when only a strong desired signal is present, and (3) full system performance.

The original MIL-STD-461A design requirement and its companion MIL-STD-462 test requirement specified four receiver tests. These standards allowed the interfering signal(s) to be both inband and out of band, which is meaningful for design and test of EW receivers, however inband testing generally is not meaningful for narrowband communications receivers. These standards were difficult to follow and had to be tailored to properly evaluate the EW and radar system. MIL-STD-461B/C still allowed the interfering signal(s) to be both inband and out of band but deleted the single signal interference test (CS08 Conducted Susceptibility test). MIL-STD-461D/-462D leave the pass/fail criteria entirely up to what is listed in the individual procurement specification. It also places all interfering signals out of band, redesignates each test number with a number "100" higher than previously used, and combines "CS08" as part of CS104. Therefore, to provide meaningful tests for EW and radar systems, the procurement specification must specify the three ranges of performance mentioned in the beginning of this section and that the tests are to be performed with the interfering signal(s) both inband and out of band. The four tests are as follows (listed in order of likelihood to cause problems):

Test Name	MIL-STD-461A	MIL-STD-461D
Undesired, Single signal interference test	CS08	Part of CS104
Desired with undesired, two signal interference tests	CS04	CS104
Two signal intermodulation test	CS03	CS103
Two signal cross modulation test	CS05	CS105

The rest of this section explains the application of these tests and uses the names of the original MIL-STD-461A tests to separate the tests by function.

TEST SETUP

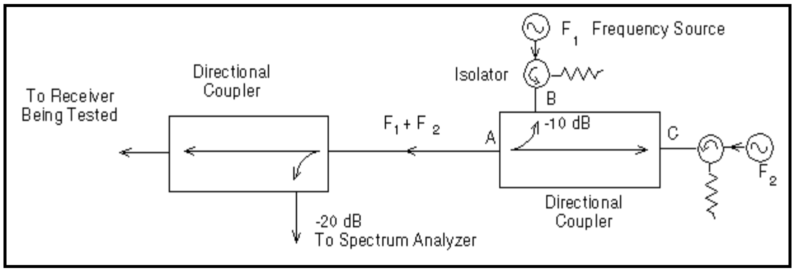

Figure 1. Receiver Test Setup When Antenna Can Be Removed.

A directional coupler used backwards (as shown here in Figure 1) is an easy way to perform two signal tests. The CW signal should be applied to the coupling arm (port B) since the maximum CW signal level is -10 dBm. The pulse signal should be applied to the straight-through path (port C) since the maximum pulse level is +10 dBm peak. These power levels are achievable with standard laboratory signal generators, therefore one doesn't have to resort to using amplifiers which may distort the signals. Always monitor the output signal to verify spectrally pure signals are being applied to the test unit. This can be accomplished by another directional coupler used in the standard configuration. Dissimilar joints or damaged or corroded microwave components can cause mixing. This can also result if the two signal generators are not isolated from one another. Therefore, even if a directional coupler is used to monitor the signal line, it is still advisable to directly measure the input to the receiver whenever there is a suspected receiver failure. This test does not need to be performed in an EMI shielded room and is more suitable for a radar or EW lab where the desired signals are readily available.

If the receiver's antenna is active or cannot be removed, a modified test as shown in Figure 2 should be performed. The monitoring antenna which is connected to the spectrum analyzer should be the same polarization as the antenna for the receiver being tested. Amplifiers may be required for the F_1 and F_2 signals. It is desirable to perform this test in an anechoic chamber or in free space.

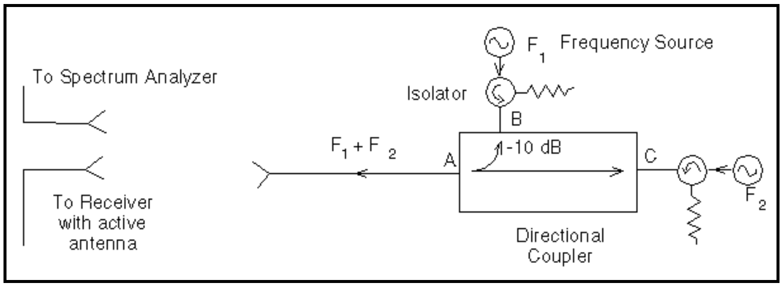

Figure 2. Receiver Test Setup When Antenna Is Active.

In the following discussion of CS08, CS04, CS03, and CS05 tests, it is assumed that when the receive light illuminates, the receiver identifies a signal that matches parameters in the User Data File (UDF) or pre-programmed list of emitter identification parameters. If a receiver is different, the following procedures will have to be appropriately tailored. If the UDF does not have entries for very low level signals in the 10% and 90% regions of each band, complete testing is not possible. Most problems due to higher order mixing products and adjacent band leakage are only evident in these regions. In the following tests, the lowest level where the receive light is constantly on is used to identify the minimum receive level. If a receiver has a receive level hysteresis or other idiosyncrasy, then using a 50% receive light blinking indicator may be more appropriate. Whatever technique is appropriate, it should be consistently used during the remainder of the test. The maximum frequency for testing is normally 20 GHz. If a millimeter wave receiver is being tested, the maximum frequency should be 110 GHz.

CS08 - UNDESIRED, SINGLE SIGNAL INTERFERENCE TEST

MIL-STD-461B/C (EMI design requirements) deleted this test. MIL-STD-461D allows a single signal test as part of CS104 (CS04) but specifies it as an out of band test. The original CS08 inband and out of band test is still needed and is the most meaningful test for wide band EW receivers which have a bandwidth close to an octave. This test will find false identification problems due to 1) lack of RF discrimination, 2) higher order mixing problems, 3) switch or adjacent channel/band leakage, and 4) cases where the absence of a desired signal causes the receiver to search and be more susceptible. In this latter case, a CS04 two signal test could pass because the receiver is captured by the desired signal, whereas a CS08 test could fail. Examples of the first three failures are as follows.

EXAMPLE 1

A 2 to 4 GHz receiver which uses video detection (e.g., crystal video) and doesn't measure RF is used for this example. This receiver assumes that if the correct Pulse Repetition Interval (PRI) is measured, it is from a signal in the frequency band of interest. Three cases can cause false identification. Refer to Figure 3.

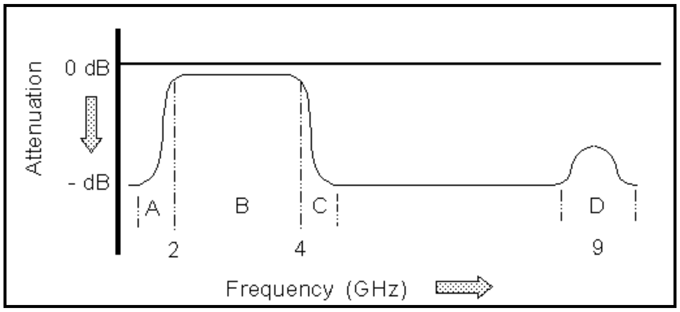

Figure 3. Frequency Areas in a Sample 2-4 GHz Receiver.

(1) Region A&C. The 2 to 4 GHz band pass filter will pass strong signals in regions A&C. If they have the correct PRI, they will also be identified.

(2) Region B. Any other signal besides the desired signal in the 2 to 4 GHz region that has the correct PRI will also be identified as the signal of interest.

(3) Region D. Band pass filters with poor characteristics tend to pass signals with only limited attenuation at frequencies that are three times the center frequency of the band pass filter. If these signals have the correct PRI, they will be incorrectly identified.

High duty cycle signals (CW or pulse doppler) in regions A, B, C, and D may overload the processing of signals, saturate the receiver, or desensitize the receiver. This case is really a two signal CS04 test failure and will be addressed in the CS04 section.

EXAMPLE 2

A receiver measuring the carrier frequency of each pulse (i.e., instantaneous frequency measurement (IFM)) and the PRI is used for this example. False signal identification can occur due to higher order mixing products showing up in the receiver pass bands. These unwanted signals result from harmonics of the input RF mixing with harmonics of the Local Oscillator (LO). Refer to Figures 4 and 5.

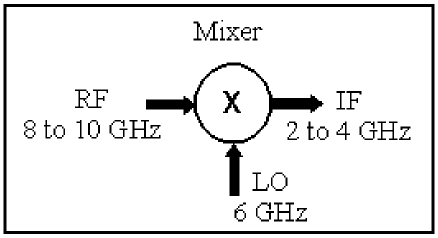

Figure 4. Low Side Mixing.

Mixers are nonlinear devices and yield the sum, difference, and the original signals. Any subsequent amplifier that is saturated will provide additional mixing products.

If a 8.5 GHz signal with a 1 kHz PRI is programmed to be identified in the UDF, measurements are made at the 2.5 GHz Intermediate Frequency (IF), i.e., RF-LO = IF = 8.5-6 = 2.5 GHz.

The same 2.5 GHz signal can result from an RF signal of 9.5 GHz due to mixing with the second harmonic of the LO i.e., 2 X 6 - 9.5 = 2.5 GHz. This signal will be substantially attenuated (approximately 35 dB) when compared to the normal IF of 9.5 - 6 = 3.5 GHz. If the receiver has filters at the IF to reduce the signal density and a filter has minimum insertion loss at 2.5 GHz and maximum insertion loss at 3.5 GHz, then only the low level 2.5 GHz signal will be measured and assumed to be due to a 8.5 GHz input signal whereas the input is really at 9.5 GHz.

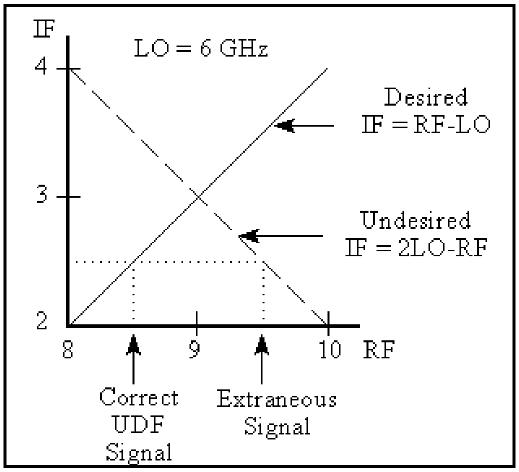

Figure 5. Low Side Mixing Results.

Spurious intermodulation products can also result from high side mixing, but generally the suppression of undesired signals is greater. In this case, the LO is at a frequency higher than the RF input. This is shown in Figures 6 and 7.

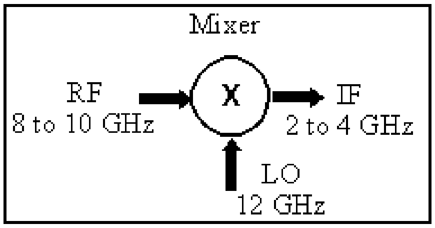

Figure 6. High Side Mixing.

Table 1. Intermodulation Product Suppression.

Harmonic of		
LO	RF	Suppression
1	1	0
1	2	ΔP-41
1	3	2ΔP-28
2	1	-35
2	2	ΔP-39
2	3	2ΔP-44
3	1	-10
3	2	ΔP-32
3	3	2ΔP-18
4	1	-35
4	2	Δ-39
5	1	-14
5	3	2ΔP-14
6	1	-35
6	2	ΔP-39
7	1	-17
7	3	2ΔP-11

Courtesy Watkins-Johnson

As previously mentioned, the amplitude of intermodulation products is greatly reduced from that of the original signals. Table 1 shows rule of thumb approximate suppression (reduction), where ΔP = P_{RF}(dBm) - P_{LO}(dBm). As can be seen, the strength of the LO is a factor. The higher the LO power, the more negative the suppression becomes.

If one assumes the maximum RF power for full system performance is +10 dBm and the LO power level is +20 dBm, then ΔP = -10 dB minimum. Therefore in this example, the 3RF-2LO mixing product would be 2ΔP - 44 = - 20 - 44 = -64 dB when compared to the desired mixing product.

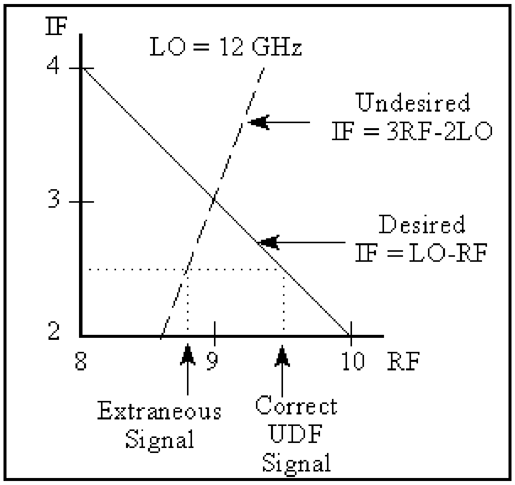

Figure 7. High Side Mixing Byproducts.

The use of double mixing, as shown in Figure 8, can significantly reduce unwanted signals but it is more expensive. For a 8 GHz signal in, one still generates a 2 GHz IF but by mixing up, then down, unwanted signals are not generated or significantly suppressed.

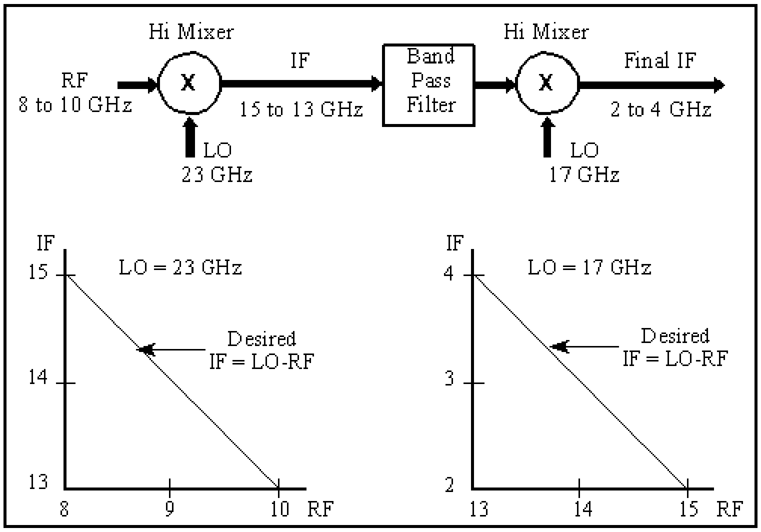
Hi Mixer
RF
8 to 10 GHz
X
IF
15 to 13 GHz
LO
23 GHz
Band
Pass
Filter
Hi Mixer
X
Final IF
2 to 4 GHz
LO
17 GHz
IF
15
14
13
8
9
10 RF
LO = 23 GHz
Desired
IF = LO-RF
IF
4
3
2
13
14
15 RF
LO = 17 GHz
Desired
IF = LO-RF

Figure 8. Double Mixing.

Some of these problems can be corrected by:

(1) always having LOs on the high side versus low side of the input RF (but this is more expensive),

(2) using double mixing

(3) software programming the receiver to measure for the potential stronger signal when a weak signal is measured in a certain IF region, and

(4) improved filtering of the LO input to the mixer and the output from the mixer.

EXAMPLE 3

If the same receiver discussed in example 2 had additional bands (Figure 9) and used a switch at the IF to select individual bands, a strong signal in an adjacent band could be inadvertently measured because:

(1) the switch, which may have 80 dB of isolation when measured outside the circuit, may only have 35 dB isolation when installed in a circuit because of the close proximity of input and output lines,

(2) the strong signal in one band may have the same IF value that is being sought in an adjacent band, and

(3) the additional parameters such as PRI may be the same.

As shown in Figure 9, assume that in band 2 we are looking for a 4.5 GHz signal that has a PRI of 1 kHz. Measurements are made at an IF of 3.5 GHz since LO-RF = IF = 8-4.5 = 3.5 GHz. If a 6.5 GHz signal is applied to band 3, its IF also equals 3.5 since LO-RF = 10-6.5 = 3.5 GHz. If this is a strong signal, has a PRI of 1 kHz, and there is switch leakage, a weak signal will be measured and processed when the switch is pointed to band 2. The receiver measures an IF of 3.5 GHz and since the switch is pointed to band 2, it scales the measured IF using the LO of band 2 i.e., LO-IF = RF = 8-3.5 = 4.5 GHz. Therefore, a 4.5 GHz signal is assumed to be measured when a 6.5 GHz signal is applied. Similarly this 6.5 GHz signal would appear as a weak 3.5 GHz signal from band 1 or a 9.5 GHz signal from band 4.

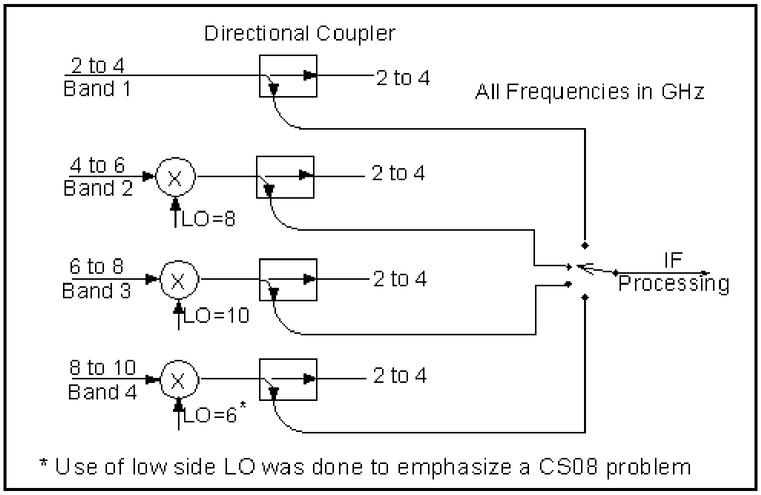

Figure 9. Multi Band Receiver With Common IF.

In performing this test it is important to map the entries of the UDF for each band i.e., show each resulting IF, its PRI, and the sensitivity level that the receive light is supposed to illuminate, i.e., if a test in one band used a PRI corresponding to a PRI in another band where the receive threshold is programmed to not be sensitive this will negate the effectiveness of a cross coupling test. Mapping the UDF will facilitate applying a strong signal to one band using the PRI of a desired signal in an adjacent band.

CS08 TEST PROCEDURE

Assume that the receiver band is 2 to 4 GHz as shown in Figure 10. Pick the UDF entry that has the greatest sensitivity. UDF #1 entry is for a 3±.05 GHz signal with a PRI of 1 kHz. If the test signal is set for the UDF #1 PRI, a receive light will also occur at the frequencies of UDF #2 if it also has the same PRI (this is not a test failure). If adjacent bands don't also have entries with the same PRI, then the test should be repeated for the band being tested with at least one of the adjacent band PRI values.

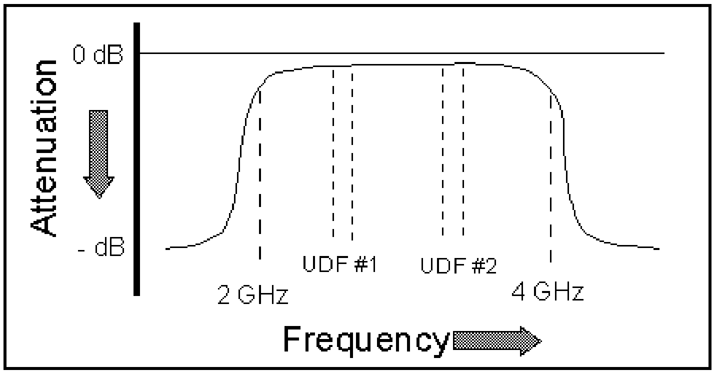

Figure 10. Receiver Band With Multiple UDF Entries.

(1) Set the receiver or jammer to the receive mode, verify it is working for UDF #1 and record P_o, the minimum signal level where the receive light is constantly on.

(2) Raise this signal to its maximum specified level for full system performance. If a maximum level is not specified, use +10 dBm peak for a pulse signal or -10 dBm for a CW signal.

(3) Tune this strong RF signal outside the UDF #1 range and record any RF frequency where the receive light comes on. If another inband UDF has the same PRI, this is not a failure.

(4) This test is performed both inband and out of band. Out of band tests should be performed on the high end to five times the maximum inband frequency or 20 GHz, whichever is less, and on the low end to IF/5 or 0.05 F0, whichever is less, unless otherwise specified. The out of band power level is +10 dBm peak for a pulse signal or -10 dBm for a CW signal, unless otherwise specified.

(5) If a receive light comes on when it is not supposed to, record the RF and reduce the power level to where the receive light just stays on constantly. Record this level P1. The interference rejection level is P1-P0= PIR

(6) Repeat this test for each type of signal the receiver is supposed to process, i.e., pulse, PD, CW, etc.

CS04 - DESIRED WITH UNDESIRED, TWO SIGNAL INTERFERENCE TEST

The intent is for a weak desired signal to be received in the presence of an adjacent CW signal. The desired signal is kept tuned at minimal power level and a strong unmodulated signal is tuned outside the UDF region. Radar and EW receivers without preselectors are likely to experience interference when this test is performed inband. Receivers with nonlinear devices before their passive band pass filter, or filters that degrade out of band, are likely to experience susceptibility problems when this test is performed out of band.

Tests performed inband - An unmodulated CW signal is used. If the receiver is supposed to handle both pulsed and CW signals, this test is performed inband. If the pulse receiver is supposed to desensitize in order to only process pulse signals above the CW level, then only this limited function is tested inband i.e., normally the levels correspond, if a CW signal of -20 dBm is present, then the receiver should process pulse signals greater than -20 dBm.

CS04 TEST PROCEDURE

(1) As shown in Figure 11, initially the pulse signal is tuned to F_0 and the minimum receive level P_0 is recorded, i.e., minimum level where the receive light is constantly on.

(2) The pulse signal is raised to the maximum specified level for full system performance and tuned on either side of F_0 to find the frequencies on both sides (F_{High} and F_{Low}) where the receive light goes out. If a maximum pulse power level is not specified, then +10 dBm peak is used. In some receivers F_L and F_H are the band skirts.

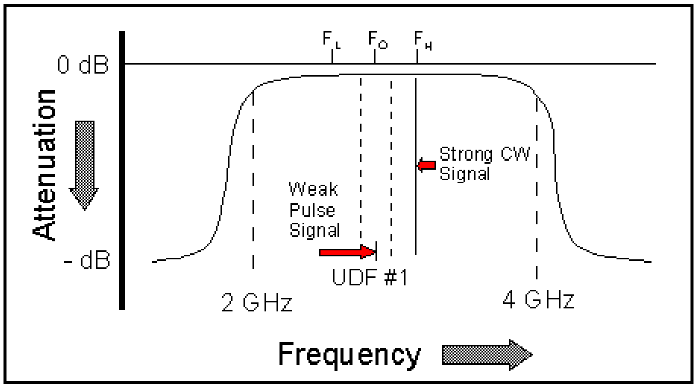

Figure 11. CS04 Test Signals.

(3) The pulse signal is returned to the level found in step 1. A CW signal at the maximum specified CW power level for full system performance is tuned above F_H and below F_L. If a maximum CW power level is not specified, then -10 dBm is used. Anytime the receive light is lost, the tuned CW RF value is recorded. The CW signal should be turned off to verify that the pulse signal can still be received in the absence of

interference. If the pulse signal is still being received, then the interfering CW signal should be reapplied and decreased to the lowest power level where the receive light stays on constantly. Record this level P_1. The interference rejection level is $P_1 - P_0 = P_{IR}$.

(4) Out of band tests should be performed to five times the maximum inband frequency or 20 GHz, whichever is less, and on the low end to IF/5 or 0.05 F_0, whichever is less, unless otherwise specified. The out of band CW power level is -10 dBm unless otherwise specified.

Failures - Out of band test

(1) If a non-linear device such as a limiter is placed before a band pass filter, a strong out of band signal can activate the limiter and cause interference with the inband signal. The solution is to place all non-linear or active devices after a passive band pass filter.

(2) Band pass filters with poor characteristics tend to pass signals with only limited attenuation at frequencies that are three times the center frequency of the band pass filter. Passage of a CW or high duty cycle signal that is out of band may desensitize or interfere with the processing of a weak inband signal.

CS03 INTERMODULATION TEST

This two signal interference test places a pulse signal far enough away (Δf) from the desired UDF frequency (F_0) that it won't be identified. A CW signal is initially placed $2\Delta f$ away. If an amplifier is operating in the saturated region, these two signals will mix and produce sum and difference signals. Subsequent mixing will result in a signal at the desired UDF frequency F_0 since $F_1 - (F_2 - F_1) = F_0$. These two signals are raised equally to strong power levels. If no problem occurs, the CW signal is tuned to the upper inband limit and then tuned out of band. A similar test is performed below F_0.

CS03 TEST PROCEDURES

(1) Set the receiver or jammer to the receive mode. Verify it is working at a desired signal frequency, (F_0), and record the minimum signal level i.e., lowest level where the receive light is constantly on (record this level P_0).

(2) The modulated signal is raised to the maximum specified level for full system performance and tuned on either side of F_0 to find the frequency F_1 on both sides where the receive light goes out. If a maximum power level is not specified, +10 dBm peak is used. The difference between F_1 and F_0 is Δf as shown in Figure 12.

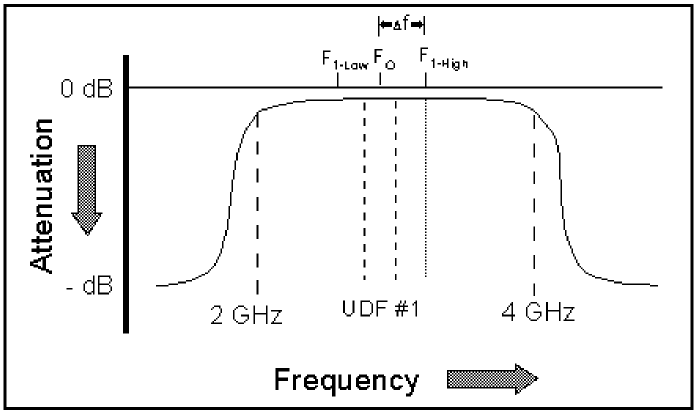

Figure 12. Initial CS03 Test Signal.

(3) As shown in Figure 13, a pulse signal is tuned to F_1 and a CW signal is tuned to F_2 where $F_2 = F_1 + \Delta f$ on the high side. The power level of the two signals is initially set to P_0 and raised together until the maximum specified levels for full system performance are reached. If maximum power levels are not specified, then +10 dBm peak is used for the pulse signal and -10 dBm is used for the CW signal. Whenever the receive light comes on, the two signals should be turned off individually to verify that the failure is due to a combination of the two signals versus (1) a single signal (CS08) type failure or (2) another inband UDF value has been matched. If the failure is due to the two signal operation, then the power level (P_1 and P_2) of F_1 and F_2 should be recorded. If $P_1 = P_2$, the intermodulation rejection level is $P_1 - P_0 = P_{IM}$. If $P_1 \neq P_2$, it is desirable to readjust them to be equal when the receive light just comes on.

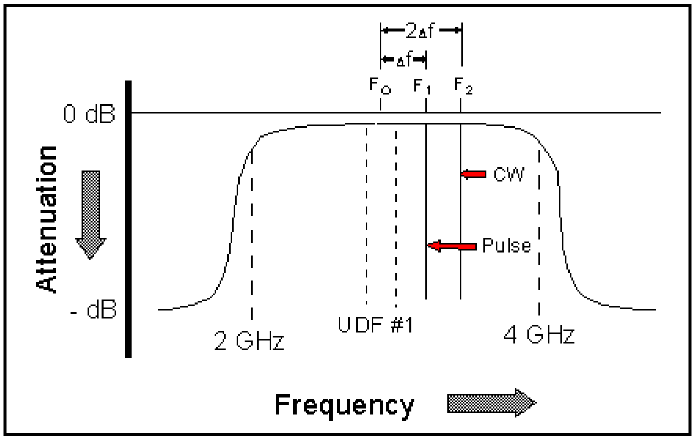

Figure 13. CS03 Testing Signal.

(4) Once the $F_1 + F_2$ signals are raised to the maximum power test levels described in step 3 without a failure, then F_2 is tuned to the upper limit of the band. F_2 should also be tuned out of band to five times the maximum inband frequency or 20 GHz whichever is less unless otherwise specified. The out of band power level is -10 dBm unless otherwise specified. Whenever the receive light comes on, F_2 should be turned off to verify that the failure is due to a two signal test. If it is, turn F_2 back on and equally drop the power levels of F_1 and F_2 to the lowest level where the receive light just comes on. Record the power levels (P_1 and P_2).

(5) Step 3 is repeated where F_1 is Δf below F_0 and $F_2 = F_1 - \Delta f$. Step 4 is repeated except F_2 is tuned to the lower limit of the band. F_2 should also be tuned out of band down to 0.1 F_0, unless otherwise specified.

(6) Normally if a failure is going to occur it will occur with the initial setting of F_1 and F_2. Care must be taken when performing this test to ensure that the initial placements of F_1 and F_2 do not result in either of the signals being identified directly.

As shown in Figure 14, if F_1 was placed at 3.2 GHz it would be identified directly and if F_2 was placed at 3.4 GHz it would be identified directly. Whereas, if F_1 was at 3.1 GHz and F_2 was at 3.2 GHz neither interfering signal would be identified directly but their intermodulation may result in an improper identification at F_0. Later when F_2 is tuned higher, the receive light will come on around 3.4 GHz and 3.6 GHz. This is not a test failure just a case of another inband UDF value being matched.

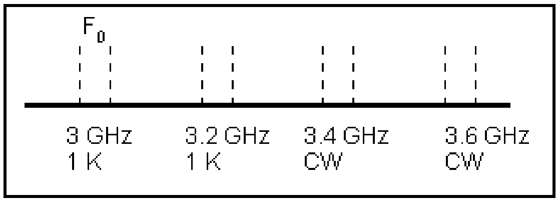

Figure 14. Sample UDF Entries.

CS05 - CROSS MODULATION

This two signal interference test places a weak CW signal where the receiver is programmed for a pulse signal and tunes a strong pulse signal elsewhere. As shown in Figure 15, when an amplifier is saturated, lower level signals are suppressed. When an amplifier is operated in the linear region all signals receive the rated linear gain. In this test the pulse signal will cause the amplifier to kick in and out of saturation and modulate the weak CW signal. The receiver may measure the modulation on the CW signal and incorrectly identify it as a pulse signal.

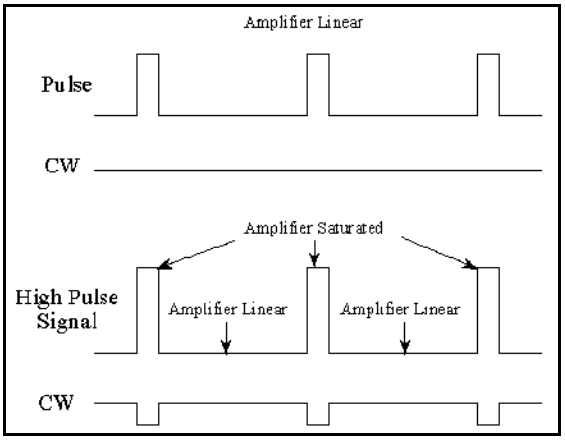
Amplifier Linear
Pulse
CW
Amplifier Saturated
High Pulse Signal
Amplifier Linear
Amplifier Linear
CW

Figure 15. Cross Modulation Example.

CS05 TEST PROCEDURE

(1) Initially the pulse signal is tuned to F_0 and the minimum power level P_0 where the receive light is constantly on is recorded.

(2) As shown in Figure 16, the signal is raised to the maximum specified level for full system performance for a pulse signal and tuned on either side of F_0 to find the frequencies on both sides, (F_{High} and F_{Low}) where the receive light goes out. If a maximum pulse power level is not specified, then +10 dBm peak is used.

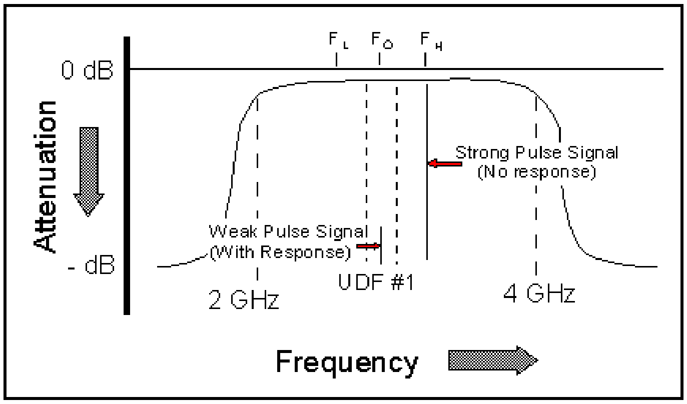
F L
F O
F H
0 dB
Attenuation
- dB
Strong Pulse Signal (No response)
Weak Pulse Signal (With Response)
2 GHz
UDF #1
4 GHz
Frequency

Figure 16. Initial CS05 Test Signals.

(3) The pulse signal from step 2 is turned off and a second signal is placed at F_0. It is a CW signal that is 10 dB stronger than the peak power level (P_0) measured is step 1. The receive light should not come on.

(4) As shown in Figure 17, the strong pulse signal of step 2 is turned back on and tuned above F_H and then tuned below F_L. Out of band tests should be performed to the maximum RF of the system + maximum IF or 20 GHz whichever is less and on the low end to the minimum RF of the system minus the maximum IF, unless otherwise specified.

(5) If a receive light occurs, turn off the weak CW signal since the “failure” may be due to the tuned pulsed signal, i.e., a CS08 failure or another inband UDF value has been matched.

If the light extinguishes when the weak CW signal is turned off, then turn the signal back on, reduce the value of the high level pulse signal until the minimum level is reached where the light stays on constantly. Record this level as P_1.

The cross modulation rejection level is P_1-P_0-10 dB = P_{CM}.

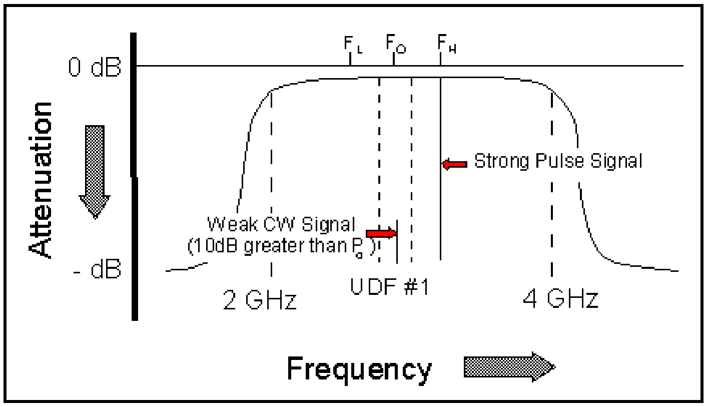
F_L
F_O
F_H
0 dB
Attenuation
Strong Pulse Signal
Weak CW Signal
(10dB greater than P_0)
- dB
2 GHz
UDF #1
4 GHz
Frequency

Figure 17. Final CS05 Test Signals.

This page intentionally left blank.

SIGNAL SORTING METHODS and DIRECTION FINDING

As shown in Figure 1, signal processing is basically a problem of signal detection, emitter parameter measurement and correlation, emitter sorting, identification, and operator notification.

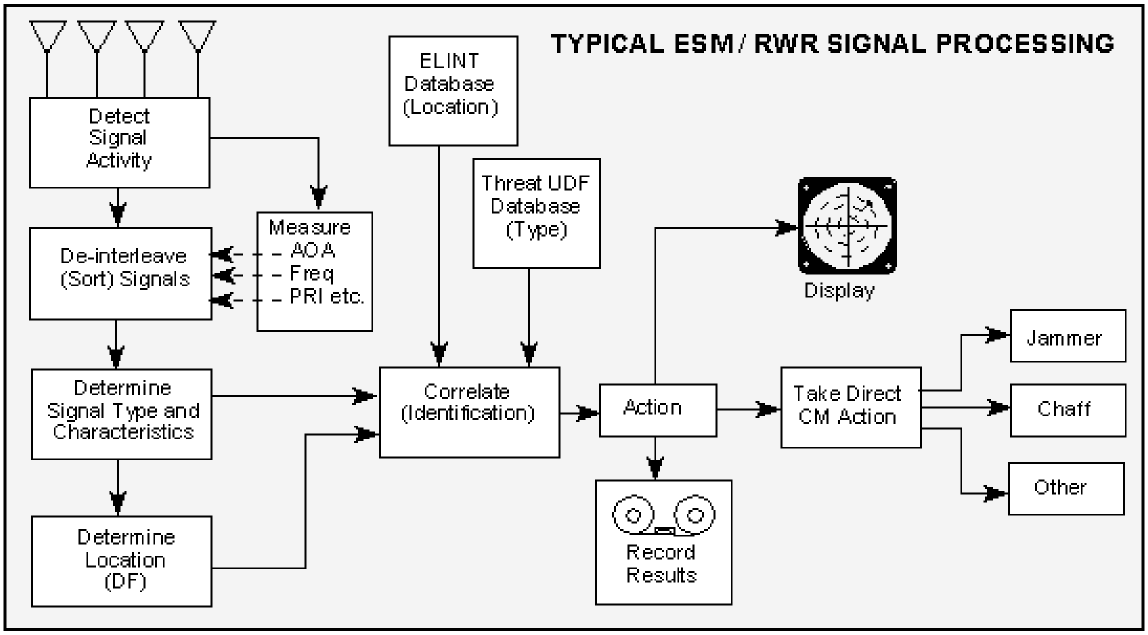
TYPICAL ESM/ RWR SIGNAL PROCESSING
Detect Signal Activity
De-interleave (Sort) Signals
Measure
– AOA
– Freq
– PRI etc.
Determine Signal Type and Characteristics
Determine Location (DF)
ELINT Database (Location)
Threat UDF Database (Type)
Correlate (Identification)
Action
Display
Record Results
Take Direct CM Action
Jammer
Chaff
Other

Figure 1. Signal Processing Steps.

The ultimate goal of this processing is to classify radar signals by their unique characteristics and to use this data to identify enemy radars operating in the environment, determine their location or direction, assess their threat to friendly forces, and display this information to the operator.

While not all electronic support measures (ESM) or radar warning receiver (RWR) systems perform every step in this process, each completes some of them. For example, ESM systems seldom initiate direct CM action, while RWRs sometimes do. Also ESM systems frequently record electronic data for future use, but few RWRs do. ESM systems place more emphasis on accurate emitter location and hence direction finding capabilities, while RWRs usually give a rough estimate of position/distance.

The typical emitter characteristics that an ESM system can measure for a pulse radar include the following data:

1. Radio Frequency (RF)
2. Amplitude (power)
3. Direction of Arrival (DOA) - also called Angle of Arrival (AOA)
4. Time of Arrival (TOA)
5. Pulse Repetition Interval (PRI)
6. PRI type
7. Pulse Width (PW)
8. Scan type and rate
9. Lobe duration (beam width)

However, this list is not comprehensive. Other emitter parameters are available which may be necessary to characterize the threat system.

More sophisticated ESM systems can measure additional parameters, such as PRI modulation characteristics, inter- and intra-pulse Frequency Modulation (FM), missile guidance characteristics (e.g., pattern of pulse spacing within a pulse group), and Continuous Wave (CW) signals.

Still other parameters which can describe an electromagnetic wave but are currently not commonly used for identification include polarization and phase. However, as threat emitters begin to use this data more frequently to avoid jamming the more important they may become in identifying signals.

Some of the emitter characteristics which describe an electromagnetic wave are shown in Figure 2.

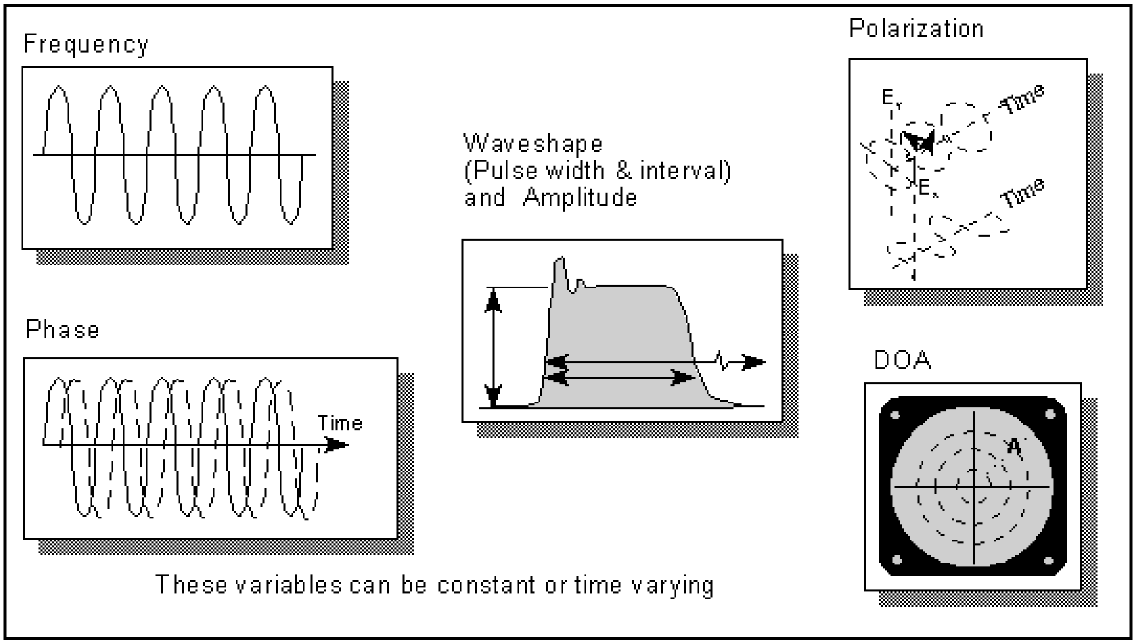
Frequency
Polarization
E_Y
Time
E_X
Time
Waveshape
(Pulse width & interval)
and Amplitude
Phase
Time
DOA
A
These variables can be constant or time varying

Figure 2. Information Content of an Electromagnetic Wave.

Table 1 illustrates the relative importance of several measured parameters during various stages of signal processing.

Table 1. Importance of Emitter Parameters During Signal Processing.

Parameter	Pulse Train De-interleavement	Emitter Identification	Intercept Correlation
Frequency	2	2	2
Amplitude	1	0	1
Angle of Arrival	2	0	2
TOA	0	0	1
PRI	2	2	2
PRI type	2	2	2
PW	2	1	1
Scan rate and type	0	2	1
Lobe Duration	0	1	1
0 Not Useful 1 Some Use 2 Very Useful			

Some emitter parameters can be measured using a single pulse; these parameters are referred to as monopulse parameters. The monopulse parameters include RF, PW, DOA, amplitude and TOA. RF can be determined on a pulse-by-pulse basis by receivers that can measure frequency. Frequency is very useful for emitter identification since most radars operate at a single frequency. Most real-time systems measure pulse width instead of pulse shape because the latter is much more difficult to characterize mathematically. Unfortunately, the apparent pulse width can be severely distorted by reflections, and consequently, its usefulness for emitter identification is limited. DOA cannot be used for emitter identification, but is excellent for sorting signals. A number of ESM systems use both frequency and DOA information to distinguish the new signals from the old (that is, known) ones. Amplitude also cannot be used for emitter identification. However, it can be used for sorting and for gross distance estimation using precompiled emitter's effective radiated power. Moreover, amplitude in conjunction with TOA can be used to determine the emitter's scan characteristics.

Other emitter parameters such as PRI, guidance and scan characteristics can be determined only by analyzing a group of pulses. All these parameters are useful for emitter identification; unfortunately, they require time for data collection and analysis, and call for sophisticated signal processing algorithms.

The problem of signal recognition in real-time is complicated by two factors: modulation of the signals and the very high pulse densities expected in the environment. Complex modulations (for example, inter-pulse RF modulation, intra-pulse RF modulation and agile Pulse Repetition Frequencies (PRFs)) present a significant pattern recognition problem for a number of ESM systems. It is expected that during some missions, hundreds of emitters will be transmitting simultaneously in the same vicinity. Wide-open antenna/receiver combination systems may have to cope with up to a million PPS. Even narrow-band receivers can expect data rates up to 100,000 PPS. At these rates, a single modern computer cannot be expected to process all the pulses, derive the characteristics for all emitters and identify the emitters in real-time. Other factors which encumber signal recognition include missing pulses, atmospheric noise and multiple reflections of pulses.

Present RWRs are designed primarily to cope with stable emitters. A stable emitter is one whose frequency and pulse repetition interval (PRI) remain relatively constant from pulse to pulse. The future threat will move steadily away from the stable emitter towards agile emitters which vary their frequency and PRI characteristics. The first change in this direction is towards the patterned agile emitter which varies its pulse and frequency parameters in accordance with a specific pattern. Examples of patterned agile emitters are MTI radars which use staggered PRFs, pulse Doppler radars which change frequency and PRF on a block-to-block basis, and certain frequency-agile radars whose transmitter frequency is mechanically modulated in a systematic pattern (e.g., spin-tuned magnetron). The next step in this evolution is towards truly agile emitters which change their frequency and PRF in a random manner on a pulse-to-pulse basis. One tempering factor in this evolution is that radars which process Doppler must maintain a constant frequency for at least two consecutive pulses.

In addition to agile frequency and PRI parameters, the future threat will be composed of a number of high-PRF pulsed Doppler, burst-frequency, CW, pulse-compression, agile-beam, and LPI radars, which use pseudo-noise waveforms. This conglomeration of radar types will cause a high signal density which must be segmented into a manageable data stream by the use of both frequency and spatial filtering in the RWR. While frequency and PRI are good parameters for sorting present-day non-agile emitters, they are poor or useless parameters for sorting agile emitters.

Angle of arrival is generally regarded as the best initial sorting parameter because it cannot be varied by the emitter from pulse to pulse.

PASSIVE DIRECTION FINDING AND EMITTER LOCATION

Direction finding (DF) systems provide several important functions in modern EW systems. We have already discussed the importance of measuring the emitter's bearing, or angle of arrival (AOA), as an invariant sorting parameter in the deinterleaving of radar signals and in separating closely spaced communication emitters. In addition, the conservation of jamming power in power-managed ECM systems depends on the ability of the associated ESM system to measure the direction to the victim emitter. A function which is becoming increasingly important in defense suppression and weapon delivery systems involves locating the emitter's position passively. This can be accomplished from a single moving platform through successive measurements of the emitter's angular direction, or from multiple platforms which make simultaneous angular measurements.

The emitter identification function requires identifying and associating consecutive pulses produced by the same emitter in angle of arrival (AOA) and frequency. The AOA is a parameter which a hostile emitter cannot change on a pulse-to-pulse basis. However, to measure the AOA of pulses which overlap in the time domain first requires them to be separated in the frequency domain. The advanced ESM receivers which accomplish this function must operate over several octaves of bandwidth while providing RMS bearing accuracies on the order of at least 2 degrees with high POI and fast reaction time in dense signal environments.

There are basically three methods, depicted in Figure 3, which allow the passive location of stationary ground-based emitters from airborne platforms.

These are:

1. The azimuth triangulation method where the intersection of successive spatially displaced bearing measurements provides the emitter location.

2. The azimuth/elevation location technique, which provides a single-pulse instantaneous emitter location from the intersection of the measured azimuth/elevation line with the earth's surface.

3. The time difference of arrival (TDOA), or precision emitter location system (PELS) method, which measures the difference in time of arrival of a single pulse at three spatially remote locations.

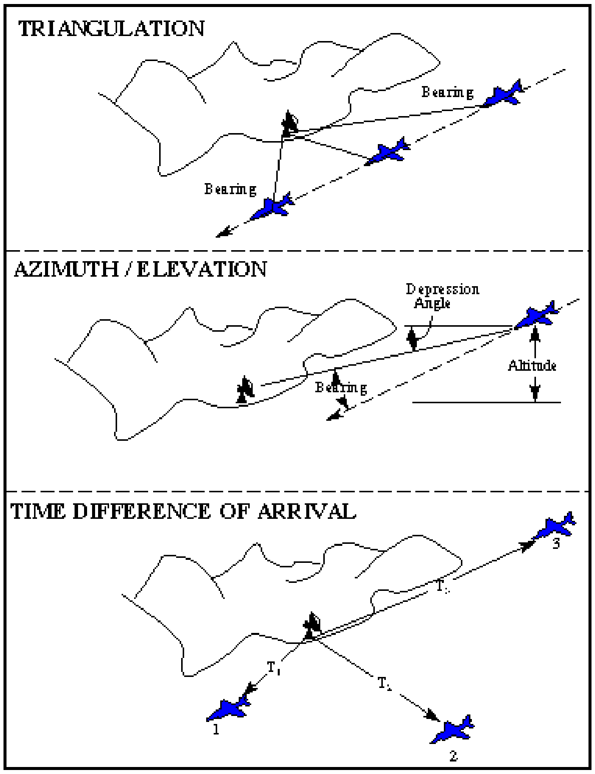

Figure 3. Passive Emitter Location Techniques.

Additional methods include:

1. Phase rate of change, which is similar to triangulation, except it makes calculations using the phase derivative.

2. Angle distance techniques, where the distance from the emitter is derived from the signal strength (with known "threat" characteristics).

3. RF Doppler processing, which measures Doppler changes as the aircraft varies direction with respect to the "target" radar.

The relative advantages and disadvantages of each are given in Table 2.

Table 2. Emitter Location Techniques.

Measurement Technique	Advantages	Disadvantages
Triangulation	Single Aircraft	Non-Instantaneous Location; Inadequate Accuracy for Remote Targeting; Not Forward Looking
Azimuth/Elevation	Single Aircraft Instantaneous Location Possible	Accuracy Degrades Rapidly at Low Altitude; Function of Range
Time Difference of Arrival (Pulsed Signals)	Very High Precision Can Support Weapon Delivery Position Requirements Very Rapid, Can Handle Short On-Time Threat	Very Complex, At Least 3 Aircraft; High Quality Receivers; DME (3 Sites); Very Wideband Data Link; Very High Performance Control Processor; Requires Very High Reliability Subsystems. Requires common time reference and correlation operation for non-pulse signals.

The triangulation method has the advantage of using a single aircraft, and its accuracy is greatest for a long baseline and the broadside geometry. The accuracy degenerates as the aircraft heading line approaches the boresight to the emitter.

The azimuth/elevation technique also has the advantage of using a single aircraft, but suffers from the difficultness of making an accurate elevation measurement with limited vertical aperture and in the presence of multipath effects.

The TDOA technique requires multiple aircraft and is complex, but has high potential accuracy. The determination of the location of the site involves the solution of at least two simultaneous second order equations for the intersection of two hyperbolas which represent $T_2 - T_1$ = Constant #1 and $T_3 - T_2$ = Constant #2. This method can be used to obtain a fix for an emitter which radiates only a single pulse.

ANGLE-OF-ARRIVAL (AOA) MEASUREMENTS

Several of the above DF measurements require AOA determination. Threat AOA measurements are also required to inform the aircrew in order to position the aircraft for optimal defense.

As shown in Figure 4, angle-of-arrival measuring systems fall into three main system categories of:

1. Scanning beam
2. Amplitude comparison or Simultaneous-multiple-beam
3. Phased Interferometer techniques

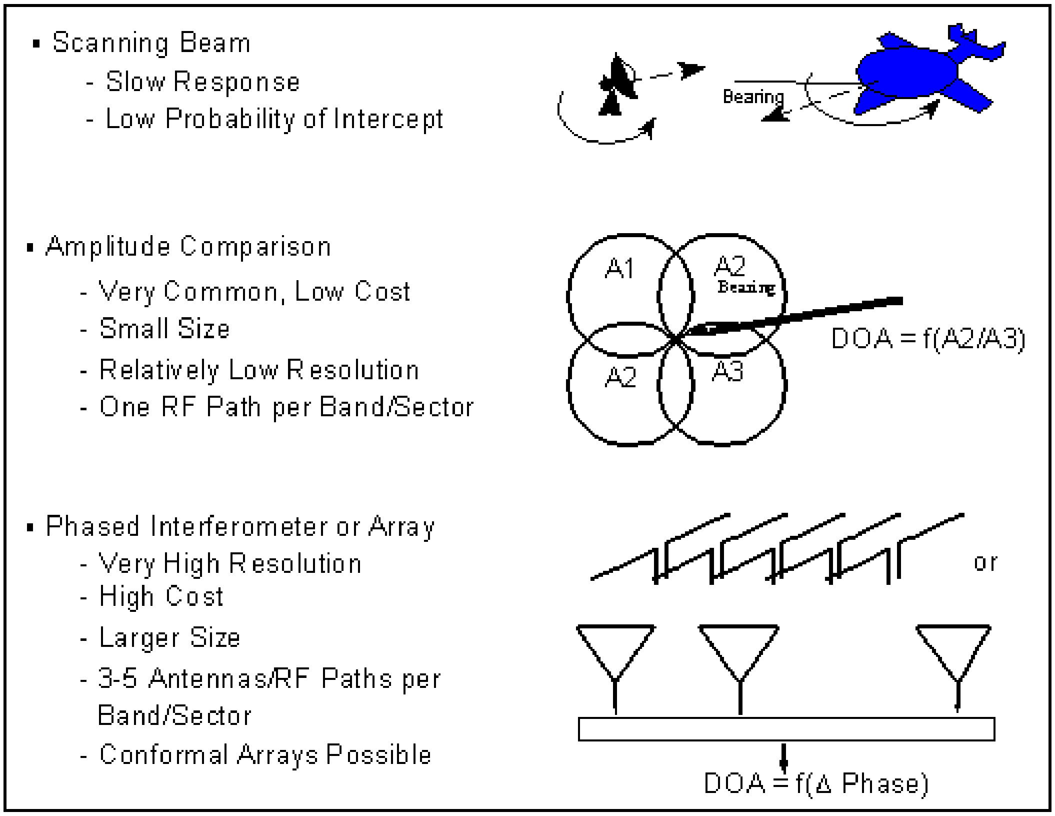

- Scanning Beam
  - Slow Response
  - Low Probability of Intercept
- Amplitude Comparison
  - Very Common, Low Cost
  - Small Size
  - Relatively Low Resolution
  - One RF Path per Band/Sector
- Phased Interferometer or Array
  - Very High Resolution
  - High Cost
  - Larger Size
  - 3-5 Antennas/RF Paths per Band/Sector
  - Conformal Arrays Possible

Figure 4. Angle-of-Arrival Measurement Techniques.

Scanning Beam

The mechanically scanning beam, or "spinner," requires only a single receiver and also exhibits high sensitivity due to the use of a directive antenna. The disadvantage is that the "spinner" usually exhibits slow response because it must rotate through the coverage angle (e.g., 360 degrees) to ensure that it intercepts an emitter. Also, if the emitter uses a scanning directional antenna, both beams must point at each other for maximum sensitivity, which is a low probability occurrence. Both of these effects cause the mechanically scanning beam technique to have a low probability of intercept (POI).

Amplitude Comparison

The two primary techniques used for direction finding are the amplitude-comparison method and the interferometer or phase-comparison method. The phase-comparison method generally has the advantage of greater accuracy, but the amplitude-comparison method is used extensively due to its lower complexity and cost. Regardless of which technique is used, it should be emphasized that the ultimate rms angular accuracy is given by:

$$\Delta\theta = \frac{k\,\theta_B}{\sqrt{SNR}}$$

where θ_B is the antenna's angular beamwidth, or interferometer lobe width, and SNR is the signal-to-noise ratio.

Thus, phase interferometers that typically use very widebeam antennas require high signal-to-noise ratios to achieve accurate angle-of-arrival measurements. Alternately, a multi-element array antenna can be used to provide relatively narrow interferometer lobes, which require modest signal-to-noise ratios.

Virtually all currently deployed radar warning receiving (RWR) systems use amplitude-comparison direction finding (DF). A basic amplitude-comparison receiver derives a ratio, and ultimately angle-of-arrival or bearing, from a pair of independent receiving channels, which utilize squinted antenna elements that are usually equidistantly spaced to provide an instantaneous 360° coverage. Typically, four or six antenna elements and receiver channels are used in such systems, and wideband logarithmic video detectors provide the signals for bearing-angle determination. The monopulse ratio is obtained by subtraction of the detected logarithmic signals, and the bearing is computed from the value of the ratio.

Amplitude comparison RWRs typically use broadband cavity-backed spiral antenna elements whose patterns can be approximated by Gaussian-shaped beams. Gaussian-shaped beams have the property that the logarithmic output ratio slope in dB is linear as a function of angle of arrival. Thus, a digital look-up table can be used to determine the angle directly. However, both the antenna beamwidth and squint angle vary with frequency over the multi-octave bands used in RWRs. Pattern shape variations cause a larger pattern crossover loss for high frequencies and a reduced slope sensitivity at low frequencies. Partial compensation of these effects, including antenna squint, can be implemented using a look-up table if frequency information is available in the RWR. Otherwise, gross compensation can be made, depending upon the RF octave band utilized.

Typical accuracies can be expected to range from 3 to 10 degrees rms for multi-octave frequency band amplitude-comparison systems which cover 360 degrees with four to six antennas.

The four-quadrant amplitude-comparison DF systems employed in RWRs have the advantage of simplicity, reliability, and low cost. Usually, only one antenna per quadrant is employed which covers the 2 to 18 GHz band. The disadvantages are poor accuracy and sensitivity, which result from the broad-beam antennas employed. Both accuracy and sensitivity can be improved by expanding the number of antennas employed. For example, expanding to eight antennas would double the accuracy and provide 3 dB more gain. As the number of antennas increases, it becomes appropriate to consider multiple-beam-forming antennas rather than just increasing the number of individual antennas. The geometry of multiple-beam-forming antennas is such that a conformal installation aboard an aircraft is difficult. Therefore, this type of installation is typically found on naval vessels or ground vehicles where the space is available to accommodate this type of antenna.

Simultaneous-multiple-beam (amplitude comparison)

The simultaneous-multiple-beam system uses an antenna, or several antennas, forming a number of simultaneous beams (e.g., Butler matrix or Rotman lens), thereby retaining the high sensitivity of the scanning antenna approach while providing fast response. However, it requires many parallel receiving channels, each with full frequency coverage. This approach is compatible with amplitude-monopulse angular measuring techniques which are capable of providing high angular accuracy.

A typical example of a multiple-beam antenna is a 16-element circular array developed as part of a digital ESM receiver. This system covers the range from 2 to 18 GHz with two antenna arrays (2 to 7.5 GHz and 7.5 to 18 GHz), has a sensitivity of -55 to -60 dBm and provides an rms bearing accuracy of better than 1.7 degrees on pulsewidths down to 100 ns.

Phased Interferometer Techniques

The term interferometer generally refers to an array type antenna in which large element spacing occurs and grating lobes appear.

Phase interferometer DF systems are utilized when accurate angle-of-arrival information is required. They have the advantage of fast response, but require relatively complex microwave circuitry, which must maintain a precise phase match over a wide frequency band under extreme environmental conditions. When high accuracy is required (on the order of 0.1 to 1°), wide baseline interferometers are utilized with ambiguity resolving circuitry. The basic geometry is depicted in Figure 5, whereby a plane wave arriving at an angle is received by one antenna earlier than the other due to the difference in path length.

The time difference can be expressed as a phase difference:

$$\varphi = \omega\Delta\tau = 2\pi a(f/c) = 2\pi\ (d \sin \theta)/\lambda,$$

where θ is the angle of arrival,
d is the antenna separation, and
λ is the wavelength in compatible units.

The unambiguous field of view (FOV) is given by $\theta = 2 \sin^{-1} (\pi/2d)$, which for $\lambda/2$ spacing results in 180° coverage. This spacing must be established for the highest frequency to be received.

Interferometer elements typically use broad antenna beams with beamwidths on the order of 90°. This lack of directivity produces several adverse effects. First, it limits system sensitivity due to the reduced antenna gain. Secondly, it opens the system to interference signals from within the antenna's broad angular coverage. The interference signals often include multipath from strong signals which can limit the accuracy of the interferometer.

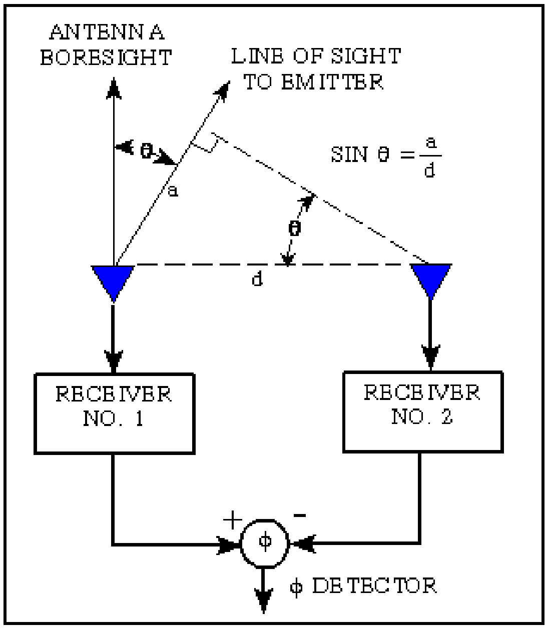

Figure 5. Phase Interferometer Principle.

In an interferometer, the locus of points that produce the same time or phase delay forms a cone. The indicated angle is the true azimuth angle multiplied by the cosine of the elevation angle. The error in assuming the incident angle to be the azimuth angle is negligible for signals near the antenna's boresight. At 45° azimuth and 10° elevation, the error is less than 1°, increasing to 15° for both at 45°. Two orthogonal arrays, one measuring the azimuth angle and the other the elevation angle can eliminate this error. For targets near the horizon, the depression angle is small, thereby requiring only horizontal arrays.

The rms angular accuracy of an interferometer in radians is given by:

$\sigma_\theta = \Delta\alpha\ /\ (\pi \bullet \sqrt{SNR}),$ where $\Delta\alpha = \lambda/(d \cdot \cos\theta)$ is the separation between adjacent nulls.

For a two-element interferometer, the spacing (d) must be $\lambda/2$ or less to provide unambiguous, or single lobe ± 90°, coverage. This, in effect, sets a wide interferometer (or grating) lobe which must be split by a large factor to achieve high accuracy. This, in turn, imposes a requirement for high SNR to achieve the large beam-splitting factor. For example, if 0.1° accuracy is required from an unambiguous two-element interferometer, then a SNR of about 50 dB is required to achieve this accuracy. This may be difficult to achieve considering the inherently low sensitivity of an interferometer system.

When high accuracy is required from an interferometer system, it is usual to employ separations greater than λ/2. The increased separation sets up a multi-grating-lobe structure through the coverage angle which requires less SNR to achieve a specified accuracy. For example, a two-element interferometer with 16λ spacing would set up a 33-grating-lobe structure (including the central lobe) throughout the ± 90° coverage angle. Within each of the 33 grating lobes, it would only require a SNR on the order of 20 dB to achieve 0.1° accuracy. However, there would be 33 ambiguous regions within the ± 90° angular coverage and also 32 nulls (where the phase detector output is zero), about which the system would be insensitive to an input signal. The ambiguities could be resolved by employing a third antenna element with λ/2 spacing, which would provide an accuracy on the order of 3° with 20 dB SNR. This accuracy is sufficient to identify which of the 33 lobes contains the signal. Providing coverage in the null regions requires additional antenna elements.

Interferometers employing multiple antenna elements are called multiple-baseline interferometers. In a typical design, the receiver consists of a reference antenna and a series of companion antennas. The spacing between the reference element and the first companion antenna is λ/2; other secondary elements are placed to form pairs separated by 1, 2, 4, and 8 wavelengths. The initial AOA is measured unambiguously by the shortest-spaced antenna pair. The next greatest spaced pair has a phase rate of change which is twice that of the first, but the information is ambiguous due to there being twice as many lobes as in the preceding pair. A greater phase rate of change permits higher angular accuracy while the ambiguity is resolved by the previous pair. Thus, the described multiple-baseline interferometer provides a binary AOA measurement where each bit of the measurement supplies a more accurate estimate of the emitter's AOA.

Harmonic multiple-baseline interferometers use elements which are spaced at $2^n \cdot \lambda/2$, with n = 0, 1, 2, 3. In nonharmonic interferometers, no pair of antennas provides a completely unambiguous reading over the complete field of view. For example, the initial spacing in the nonharmonic interferometer might be λ, while the next companion element spacing is 3λ/2. Ambiguities are resolved by truth tables, and hence the accuracy is set by the spacing of the widest baseline antenna pair. Nonharmonic interferometers have been implemented over 9:1 bandwidths (2 to 18 GHz) with rms accuracies from 0.1 to 1° and with no ambiguities over ± 90°. The principal advantage of the nonharmonic over the harmonic interferometer is the increased bandwidth for unambiguous coverage.

Interferometer DF accuracy is determined by the widest baseline pair. Typical cavity-backed spirals, track to 6 electrical degrees, and associated receivers track to 9°, resulting in an rms total of 11°. At a typical 16 dB SNR, the rms phase noise is approximately 9 electrical degrees. For these errors and an emitter angle of 45°, a spacing of 25λ is required for 0.1° rms accuracy while a spacing of 2.5λ is needed for 1° accuracy. For high accuracy, interferometer spacings of many feet are required. In airborne applications, this usually involves mounting interferometer antennas in the aircraft's wingtips.

The characteristics of typical airborne amplitude comparison and phase interferometer DF systems are summarized in Table 3. The phase interferometer system generally uses superheterodyne receivers which provide the necessary selectivity and sensitivity for precise phase measurements.

Table 3. Direction of Arrival Measurement Techniques.

	Amplitude Comparison	Phase Interferometer
Sensor Configuration	Typically 4 to 6 Equispaced Antenna Elements for 360° Coverage	2 or more RHC or LHC Spirals in Fixed Array
DF Accuracy	$DF_{ACC} \approx \frac{\theta_{BW}^2 \Delta C_{dB}}{24S}$ (Gaussian Shape)	$DF_{ACC} = \frac{\lambda}{2\pi d \cos\theta} \Delta\theta$
DF Accuracy Improvement	Decrease Antenna BW Decrease Amplitude Mistrack Increase Squint Angle	Increase Spacing of Outer Antennas; Decrease Phase Mistrack
Typical DF Accuracy	3° to 10° rms	0.1° to 3° rms
Sensitivity to Multipath/ Reflections	High Sensitivity Mistrack of Several dB Can Cause Large DF Errors	Relatively Insensitive; Interferometer Can Be Made to Tolerate Large Phase Errors
Platform Constraints	Locate in Reflection Free Area	Reflection Free Area; Real Estate For Array; Prefers Flat Radome
Applicable Receivers	Crystal Video; Channelizer; Acousto-Optic; Compressive; Superheterodyne	Superheterodyne

ΔC_{dB} = Amplitude Monopulse Ratio in dB
S = Squint Angle in degrees
θ_{BW} = Antenna Beamwidth in degrees

This page intentionally left blank.

MICROWAVE / RF COMPONENTS

Microwave Waveguides and Coaxial Cable 6-1
Voltage Standing Wave Ratio (VSWR) / Reflection Coefficient
    Return Loss / Mismatch Loss 6-2
Microwave Coaxial Connectors 6-3
Power Dividers and Directional Couplers 6-4
Attenuators / Filters / DC Blocks 6-5
Terminations / Dummy Loads 6-6
Circulators and Diplexers 6-7
Mixers and Frequency Discriminators 6-8
Detectors 6-9
RF / Microwave Amplifiers 6-10
Signal Generation 6-11
Digital Processing Components 6-12
Microwave Measurements 6-13

This page intentionally left blank.

MICROWAVE WAVEGUIDES and COAXIAL CABLE

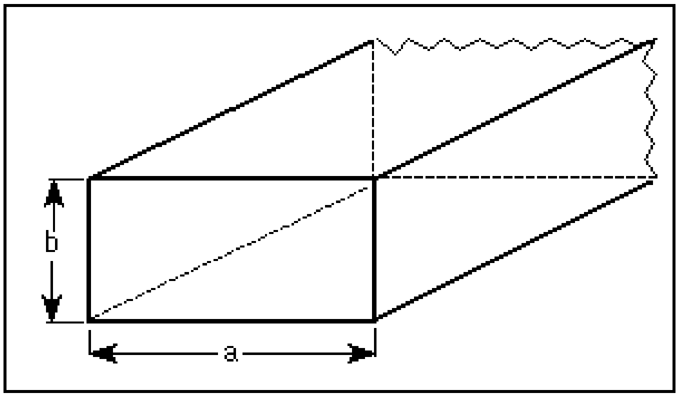

Figure 1. The Rectangular Waveguide.

In general, a waveguide consists of a hollow metallic tube of arbitrary cross section uniform in extent in the direction of propagation. Common waveguide shapes are rectangular, circular, and ridged. The rectangular waveguide has a width a and height b as shown in Figure 1. Commonly used rectangular waveguides have an aspect ratio b/a of approximately 0.5. Such an aspect ratio is used to preclude generation of field variations with height and their attendant unwanted modes. Waveguides are used principally at frequencies in the microwave range; inconveniently large guides would be required to transmit radio-frequency power at longer wavelengths. In the X-Band frequency range of 8.2 to 12.4 GHz, for example, the U.S. standard rectangular waveguide, WR-90, has an inner width of 2.286 cm (0.9 in.) and an inner height of 1.016 cm (0.4 in.).

In waveguides the electric and magnetic fields are confined to the space within the guides. Thus no power is lost to radiation. Since the guides are normally filled with air, dielectric losses are negligible. However, there is some I^2R power lost to heat in the walls of the guides, but this loss is usually very small.

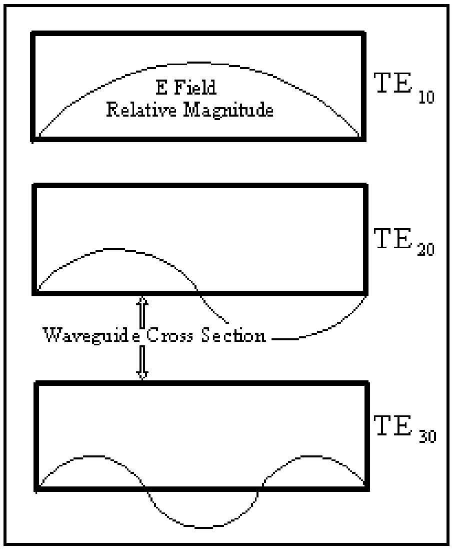

Figure 2. TE Modes.

It is possible to propagate several modes of electromagnetic waves within a waveguide. The physical dimensions of a waveguide determine the cutoff frequency for each mode. If the frequency of the impressed signal is above the cutoff frequency for a given mode, the electromagnetic energy can be transmitted through the guide for that particular mode with minimal attenuation. Otherwise the electromagnetic energy with a frequency below cutoff for that particular mode will be attenuated to a negligible value in a relatively short distance. This grammatical use of cutoff frequency is opposite that used for coaxial cable, where cutoff frequency is for the highest useable frequency. The dominant mode in a particular waveguide is the mode having the lowest cutoff frequency. For rectangular waveguide this is the TE_{10} mode. The TE (transverse electric) signifies that all electric fields are transverse to the direction of propagation and that no longitudinal electric field is present. There is a longitudinal component of magnetic field and for this reason the TE_{mn} waves are also called H_{mn} waves. The TE designation is usually preferred. Figure 2 shows a graphical depiction of the E field variation in a waveguide for the $TE_{10,}$ TE_{20}, and TE_{30} modes. As can be seen, the first index indicates the number of half wave loops across the width of the guide and the second index, the number of loops across the height of the guide - which in this case is zero. It is advisable to choose the dimensions of a guide in such a way that, for a given input signal, only the energy of the dominant mode can be transmitted through the guide. For example, if for a particular frequency, the width of a rectangular guide is too large, then the TE_{20} mode can propagate causing a myriad of problems. For rectangular guides of low aspect ratio the TE_{20} mode is the next higher order mode and is harmonically related to the cutoff frequency of the TE_{10} mode. It is this relationship together with attenuation and propagation considerations that determine the normal operating range of rectangular waveguide.

The discussion on circular waveguides will not be included because they are rarely used in the EW area. Information regarding circular waveguides can be found in numerous textbooks on microwaves.

CHARACTERISTICS OF STANDARD RECTANGULAR WAVEGUIDES

Rectangular waveguides are commonly used for power transmission at microwave frequencies. Their physical dimensions are regulated by the frequency of the signal being transmitted. Table 1 tabulates the characteristics of the standard rectangular waveguides. It may be noted that the number following the EIA prefix "WR" is in inside dimension of the widest part of the waveguide (i.e., WR90 has an inner dimension of 0.90").

DOUBLE RIDGE RECTANGULAR WAVEGUIDE

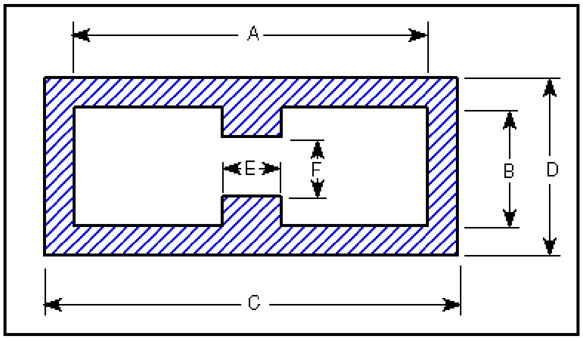

Figure 3. Double Ridge Waveguide.

Another type of waveguide commonly used in EW systems is the double ridge rectangular waveguide. The ridges in this waveguide increase the bandwidth of the guide at the expense of higher attenuation and lower power-handling capability. The bandwidth can easily exceed that of two contiguous standard waveguides. Introduction of the ridges mainly lowers the cutoff frequency of the TE_{10} mode from that of the unloaded guide, which is predicated on width alone. The reason for this can easily be explained when the field configuration in the guide at cutoff is investigated. At cutoff there is no longitudinal propagation down the guide. The waves simply travel back and forth between the side walls of the guide. In fact the guide can be viewed as a composite parallel plate waveguide of infinite width where the width corresponds to the direction of propagation of the normal guide. The TE_{10} mode cutoff occurs where this composite guide has its lowest-order resonant frequency. This occurs when there is only one E field maximum across the guide which occurs at the center for a symmetrical ridge. Because of the reduced height of the guide under the ridge, the effective TE_{10} mode resonator is heavily loaded as though a shunt capacitor were placed across it. The cutoff frequency is thus lowered considerably. For the TE_{20} mode the fields in the center of the guide will be at a minimum. Therefore the loading will have a negligible effect. For guides of proper aspect ratio, ridge height, and ridge width, an exact analysis shows that the TE_{10} mode cutoff can be lowered substantially at the same time the TE_{20} and TE_{30} mode cutoffs are raised slightly. Figure 3 shows a typical double ridged waveguide shape and Table 2 shows double ridged waveguide specifications. In the case of ridged waveguides, in the EIA designation, (WRD350 D36) the first "D" stands for double ridged ("S" for single ridged), the 350 is the starting frequency (3.5 GHz), and the "D36" indicates a bandwidth of 3.6:1. The physical dimensions and characteristics of a WRD350 D24 and WRD350 D36 are radically different. A waveguide with a MIL-W-23351 dash number beginning in 2 (i.e., 2-025) is a double ridge 3.6:1 bandwidth waveguide. Likewise a 1- is a single ridge 3.6:1, a 3- is a single ridge 2.4:1, and a 4- is a double ridge 2.4:1 waveguide.

Figure 4 shows a comparison of the frequency /attenuation characteristics of various waveguides. The attenuation is based on real waveguides which is higher than the theoretical values listed in Tables 1 and 2. Figure 5 shows photographs of waveguides with some common connectors. Figure 6 shows attenuation characteristics of various RF coaxial cables.

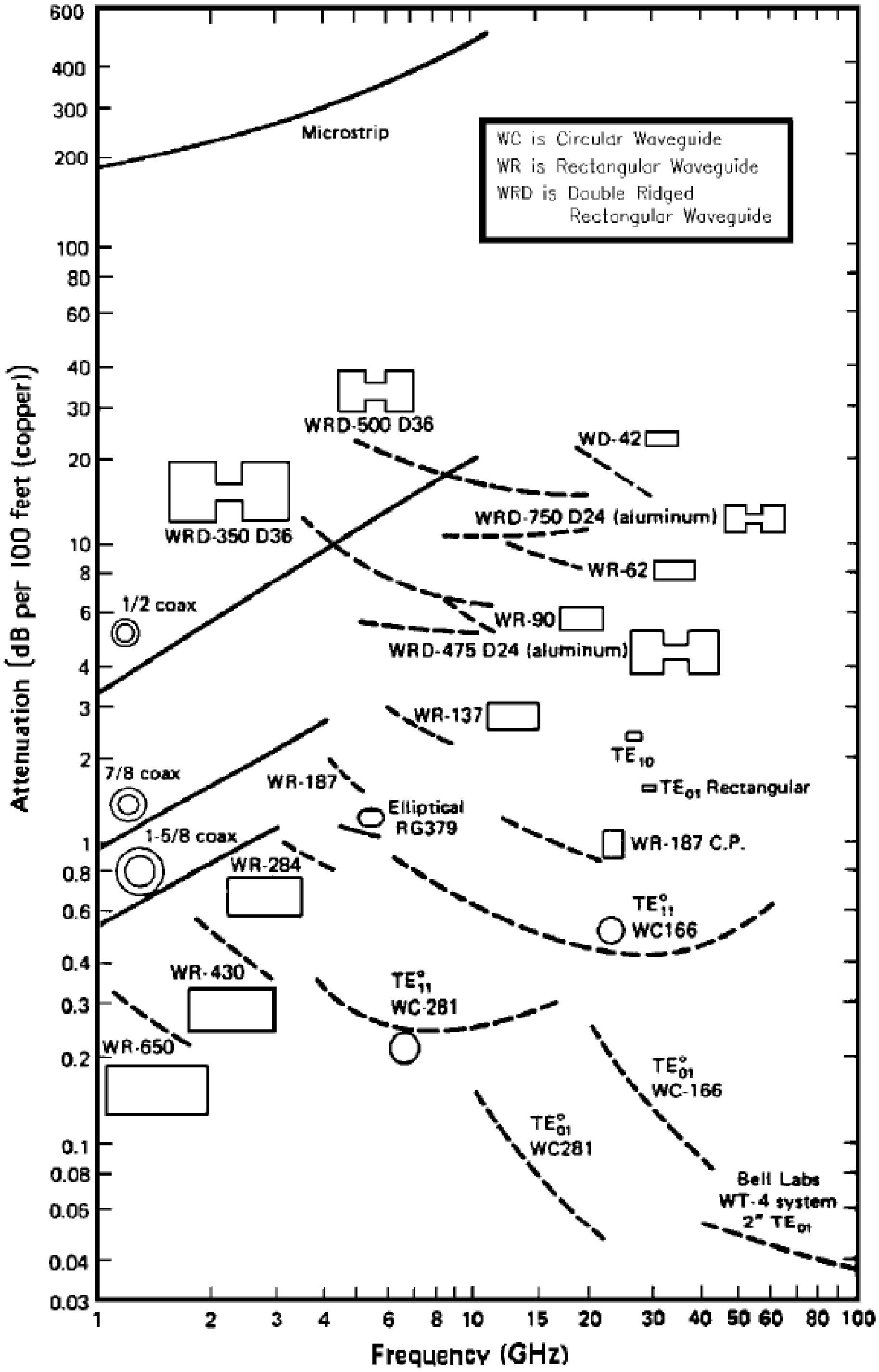

Figure 4. Attenuation vs. Frequency for a Variety of Waveguides and Cables.

Table 1. Rectangular Waveguide Specifications.

Waveguide Size	JAN WG Desig	MIL-W-85 Dash #	Material	Freq Range (GHz)	Freq Cutoff (GHz)	Power (at 1 Atm)		Insertion Loss (dB/100ft)	Dimensions (Inches)	
						CW	Peak		Outside	Wall Thickness
WR284	RG48/U RG75/U	1-039 1-042	Copper Aluminum	2.60 - 3.95	2.08	45 36	7650	.742-.508 1.116-.764	3.000x1.500	0.08
WR229	RG340/U RG341/U	1-045 1-048	Copper Aluminum	3.30 - 4.90	2.577	30 24	5480	.946-.671 1.422-1.009	2.418x1.273	0.064
WR187	RG49/U RG95/U	1-051 1-054	Copper Aluminum	3.95 - 5.85	3.156	18 14.5	3300	1.395-.967 2.097-1.454	1.000x1.000	0.064
WR159	RG343/U RG344/U	1-057 1-060	Copper Aluminum	4.90 - 7.05	3.705	15 12	2790	1.533-1.160 2.334-1.744	1.718x0.923	0.064
WR137	RG50/U RG106/U	1-063 1-066	Copper Aluminum	5.85 - 8.20	4.285	10 8	1980	1.987-1.562 2.955-2.348	1.500x0.750	0.064
WR112	RG51/U RG68/U	1-069 1-072	Copper Aluminum	7.05 - 10.0	5.26	6 4.8	1280	2.776-2.154 4.173-3.238	1.250x0.625	0.064
WR90	RG52/U RG67/U	1-075 1-078	Copper Aluminum	8.2 - 12.4	6.56	3 2.4	760	4.238-2.995 6.506-4.502	1.000x0.500	0.05
WR75	RG346/U RG347/U	1-081 1-084	Copper Aluminum	10.0 - 15.0	7.847	2.8 2.2	620	5.121-3.577 7.698-5.377	0.850x0.475	0.05
WR62	RG91/U RG349/U	1-087 1-091	Copper Aluminum	12.4 - 18.0	9.49	1.8 1.4	460	6.451-4.743 9.700-7.131	0.702x0.391	0.04
WR51	RG352/U RG351/U	1-094 1-098	Copper Aluminum	15.0 - 22.0	11.54	1.2 1	310	8.812-6.384 13.250-9.598	0.590x0.335	0.04
WR42	RG53/U	1-100	Copper	18.0 - 26.5	14.08	0.8	170	13.80-10.13	0.500x0.250	0.04
WR34	RG354/U	1-107	Copper	2.0 - 33.0	17.28	0.6	140	16.86-11.73	0.420x0.250	0.04
WR28	RG271/U	3-007	Copper	26.5 - 40.0	21.1	0.5	100	23.02-15.77	0.360x0.220	0.04

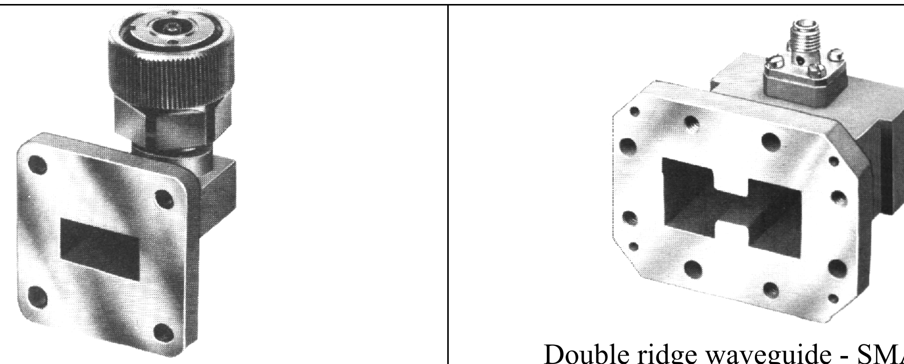

Standard waveguide - 7 mm

Double ridge waveguide - SMA jack

Figure 5. Waveguides With Some Common Connections.

Table 2. Double Ridge Rectangular Waveguide Specifications.

Waveguide Size	MIL-W-23351 Dash #	Material	Freq Range (GHz)	Freq Cutoff (GHz)	Power (at 1 Atm)		Insertion Loss (dB/ft)	Dimensions (Inches)					
					CW	Peak		A	B	C	D	E	F
WRD250		Alum Brass Copper Silver Al	2.60 - 7.80	2.093	24	120	0.025 0.025 0.018 0.019	1.655	0.715	2	1	0.44	0.15
WRD350 D24	4-029 4-303 4-031	Alum Brass Copper	3.50 - 8.20	2.915	18	150	0.0307 0.0303 0.0204	1.48	0.688	1.608	0.816	0.37	0.292
WRD475 D24	4-033 4-034 4-035	Alum Brass Copper	4.75 - 11.00	3.961	8	85	0.0487 0.0481 0.0324	1.09	0.506	1.19	0.606	0.272	0.215
WRD500 D36	2-025 2-026 2-027	Alum Brass Copper	5.00 - 18.00	4.222	4	15	0.146 0.141 0.095	0.752	0.323	0.852	0.423	0.188	0.063
WRD650		Alum Brass Copper	6.50 - 18.00	5.348	4	25	0.106 0.105 0.07	0.720	0.321	0.820	0.421	0.173	0.101
WRD750 D24	4-037 4-038 4-039	Alum Brass Copper	7.50 - 18.00	6.239	4.8	35	0.0964 0.0951 0.0641	0.691	0.321	0.791	0.421	0.173	0.136
WRD110 D24	4-041 4-042 4-043	Alum Brass Copper	11.00 - 26.50	9.363	1.4	15	0.171 0.169 0.144	0.471	0.219	0.551	0.299	0.118	0.093
WRD180 D24	4-045 4-046 4-047	Alum Brass Copper	18.00 - 40.00	14.995	0.8	5	0.358 0.353 0.238	0.288	0.134	0.368	0.214	0.072	0.057

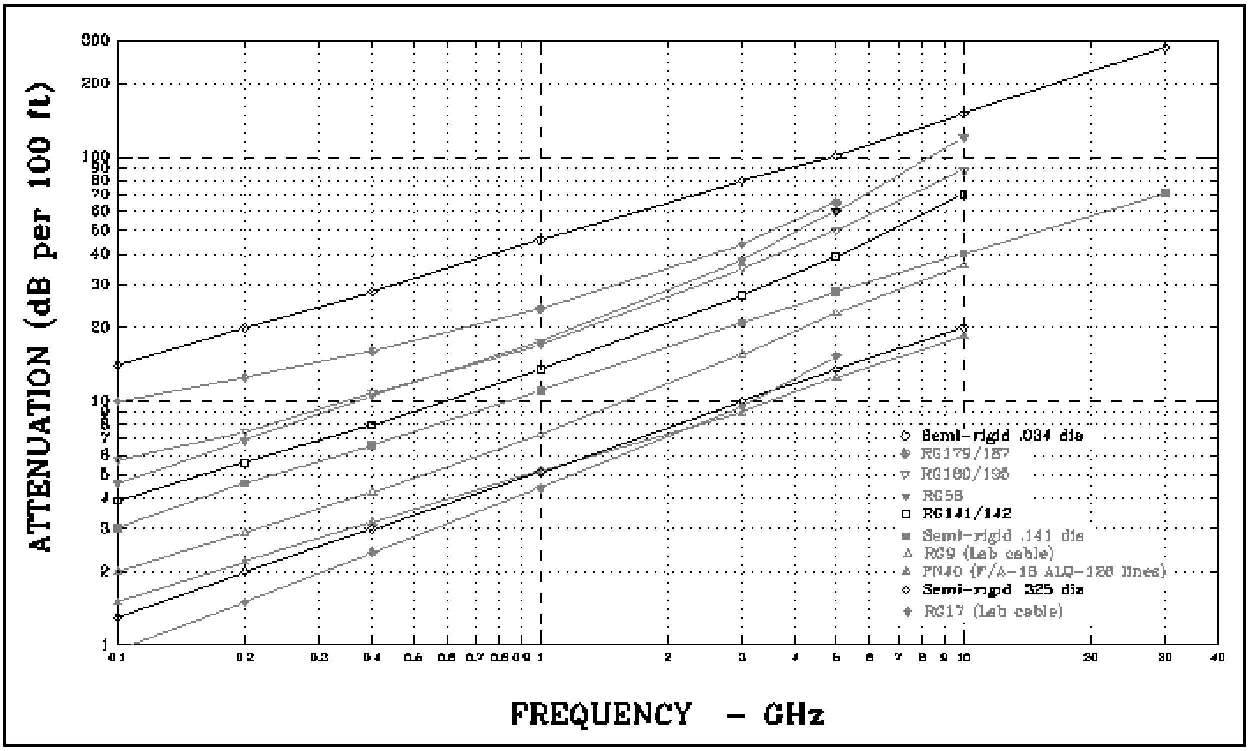

Figure 6. Attenuation vs. Frequency for a Variety of Coaxial Cables.

This page intentionally left blank.

VOLTAGE STANDING WAVE RATIO (VSWR) / REFLECTION COEFFICIENT RETURN LOSS / MISMATCH LOSS

When a transmission line is terminated with an impedance, Z_L, that is not equal to the characteristic impedance of the transmission line, Z_O, not all of the incident power is absorbed by the termination. Part of the power is reflected back so that phase addition and subtraction of the incident and reflected waves creates a voltage standing wave pattern on the transmission line. The ratio of the maximum to minimum voltage is known as the Voltage Standing Wave Ratio (VSWR) and successive maxima and minima are spaced by 180° ($\lambda/2$).

$$VSWR = \frac{E_{\max}}{E_{\min}} = \frac{E_i + E_r}{E_i - E_r}$$

where $E_{\max}$ = maximum voltage on the standing wave
$E_{\min}$ = minimum voltage on the standing wave
Ei = incident voltage wave amplitude
Er = reflected voltage wave amplitude

The reflection coefficient, ρ, is defined as E_r/E_i and in general, the termination is complex in value, so that ρ will be a complex number.

Additionally we define: $\Gamma = \frac{Z_L - Z_O}{Z_L + Z_O}$ The refection coefficient, ρ, is the absolute value of the magnitude of Γ. If the equation for VSWR is solved for the reflection coefficient, it is found that:

$$\begin{matrix}\textit{Reflection}\\ \textit{Coefficient}\end{matrix} = \rho = |\Gamma| = \frac{VSWR - 1}{VSWR + 1} \qquad \text{Consequently, } VSWR = \frac{1+\rho}{1-\rho}$$

The return loss is related through the following equations:

$$\begin{matrix}\textit{Return}\\ \textit{Loss}\end{matrix} = 10 \log\left[\frac{P_i}{P_r}\right] = -20 \log\left[\frac{E_r}{E_i}\right] = -20 \log\left[\frac{VSWR - 1}{VSWR + 1}\right] = -20 \log \rho$$

Return loss is a measure in dB of the ratio of power in the incident wave to that in the reflected wave, and as defined above always has a positive value. For example if a load has a Return Loss of 10 dB, then 1/10 of the incident power is reflected. The higher the return loss, the less power is actually lost.

Also of considerable interest is the Mismatch Loss. This is a measure of how much the transmitted power is attenuated due to reflection. It is given by the following equation:

Mismatch Loss = $-10 \log (1 - \rho^2)$

For example, an antenna with a VSWR of 2:1 would have a reflection coefficient of 0.333, a mismatch loss of 0.51 dB, and a return loss of

VSWR	Return Loss (dB)	% Power / Voltage Loss	Reflection Coefficient	Mismatch Loss (dB)
1	∞	0 / 0	0	0.000
1.15	23.1	0.49 / 7.0	0.07	.021
1.25	19.1	1.2 / 11.1	0.111	.054
1.5	14.0	4.0 / 20.0	0.200	.177
1.75	11.3	7.4 / 27.3	0.273	.336
1.9	10.0	9.6 / 31.6	0.316	.458
2.0	9.5	11.1 / 33.3	0.333	.512
2.5	7.4	18.2 / 42.9	0.429	.880
3.0	6.0	25.1 / 50.0	0.500	1.25
3.5	5.1	30.9 / 55.5	0.555	1.6
4.0	4.4	36.3 / 60.0	0.600	1.94
4.5	3.9	40.7 / 63.6	0.636	2.25
5.0	3.5	44.7 / 66.6	0.666	2.55
10	1.7	67.6 / 81.8	0.818	4.81
20	0.87	81.9 / 90.5	0.905	7.4
100	0.17	96.2 / 98.0	0.980	14.1
∞	.000	100 / 100	1.00	∞

* Divide % Voltage loss by 100 to obtain ρ (reflection coefficient)

9.54 dB (11% of your transmitter power is reflected back). In some systems this is not a trivial amount and points to the need for components with low VSWR.

If 1000 watts (60 dBm/ 30 dBW) is applied to this antenna, the return loss would be 9.54 dB. Therefore, 111.1 watts would be reflected and 888.9 watts (59.488 dBm/29.488 dBW) would be transmitted, so the mismatch loss would be 0.512 dB.

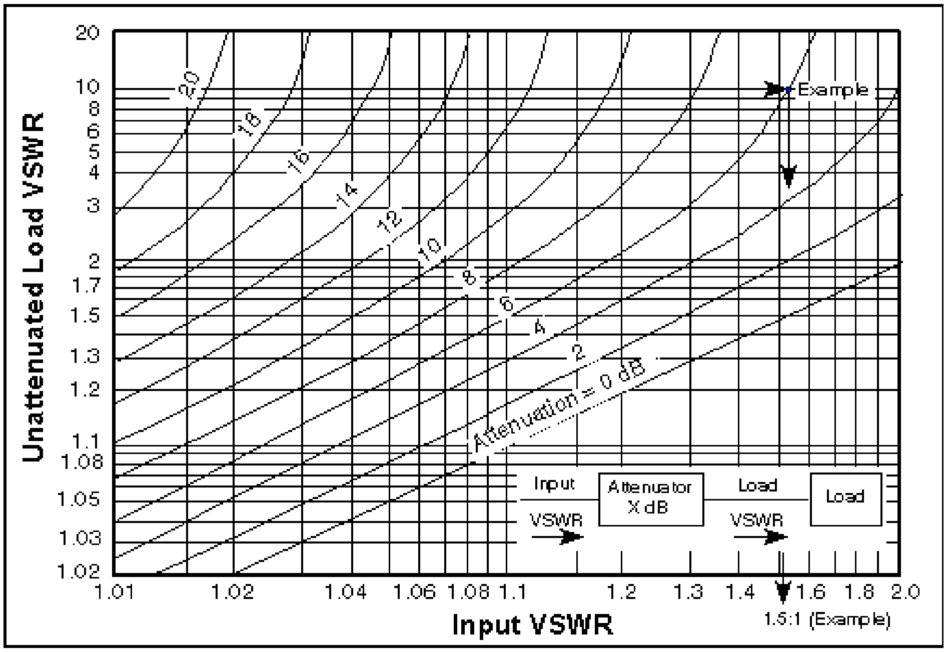

Figure 1. Reduction of VSWR by Attenuation.

Transmission line attenuation improves the VSWR of a load or antenna. For example, a transmitting antenna with a VSWR of 10:1 (poor) and a line loss of 6 dB would measure 1.5:1 (okay) if measured at the transmitter. Figure 1 shows this effect.

Therefore, if you are interested in determining the performance of antennas, the VSWR should always be measured at the antenna connector itself rather than at the output of the transmitter. Transmit cabling will load the line and create an illusion of having a better antenna VSWR. Transmission lines should have their insertion loss (attenuation) measured in lieu of VSWR, but VSWR measurements of transmission lines are still important because connection problems usually show up as VSWR spikes.

Historically VSWR was measured by probing the transmission line. From the ratio of the maximum to minimum voltage, the reflection coefficient and terminating impedance could be calculated. This was a time consuming process since the measurement was at a single frequency and mechanical adjustments had to be made to minimize coupling into circuits. Problems with detector characteristics also made the process less accurate. The modern network analyzer system sweeps very large frequency bandwidths and measures the incident power, P_i, and the reflected power, P_r. Because of the considerable computing power in the network analyzer, the return loss is calculated from the equation given previously, and displayed in real time. Optionally, the VSWR can also be calculated from the return loss and displayed real time.

If a filter is needed on the output of a jammer, it is desirable to place it approximately half way between the jammer and antenna. This may allow the use of a less expensive filter, or a reflective filter vs. an absorptive filter.

Special cases exist when comparing open and shorted circuits. These two conditions result in the same ∞ VSWR and zero dB return loss even though there is a 180° phase difference between the reflection coefficients. These two conditions are used to calibrate a network analyzer.

MICROWAVE COAXIAL CONNECTORS

For high-frequency operation, the average circumference of a coaxial cable must be limited to about one wavelength in order to reduce multimodal propagation and eliminate erratic reflection coefficients, power losses, and signal distortion.

Except for the sexless APC-7 connector, all other connectors are identified as either male (plugs) which have a center conductor that is a probe or female (jacks) which have a center conductor that is a receptacle. Sometimes it is hard to distinguish them as some female jacks may have a hollow center “pin” which appears to be male, yet accepts a smaller male contact.

An adapter is an approximately zero loss interface between two connectors and is called a barrel when both connectors are identical. A number of common of coaxial connectors are described below, however other special purpose connectors exist, including blind mate connectors where spring fingers are used in place of threads to obtain shielding (desired connector shielding should be at least 90 dB).

APC-2.4 (2.4mm) - The 50 Ω APC-2.4 (Amphenol Precision Connector-2.4 mm) is also known as an OS-50 connector. It was designed to operate at extremely high microwave frequencies (up to 50 GHz).	 2.4 mm jack - 3.5 mm jack
APC-3.5 (3.5mm) - The APC-3.5 connector provides repeatable connections and has a very low VSWR. Either the male or female end of this 50 Ω connector can mate with the opposite type of SMA connector. The APC-3.5 connector can work at frequencies up to 34 GHz.	
APC-7 (7mm) - The APC-7 was developed by HP, but has been improved and is now manufactured by Amphenol. The connector provides a coupling mechanism without male or female distinction and is the most repeatable connecting device used for very accurate 50 Ω measurement applications. Its VSWR is extremely low up to 18 GHz. Other companies have 7mm series available.	 7mm - 3.5mm plug
BNC (OSB) - The BNC (Bayonet Neill-Concelman or Bayonet Navy Connector) was originally designed for military system applications during World War II. The connector operates best at frequencies up to about 4 GHz; beyond that it tends to radiate electromagnetic energy. The BNC can accept flexible cables with diameters of up to 6.35 mm (0.25 in.) and characteristic impedance of 50 to 75 Ω. It is now the most commonly used connector for frequencies under 1 GHz. Other names the BNC has picked up over the years include: “Baby Neill-Concelman”, “Baby N connector”, “British Naval Connector”, and “Bayonet Nut Connector”	 BNC male plug (above) BNC female jack

C - The C (Concelman) coaxial connector is a medium size, older type constant 50 Ω impedance. It has a bayonet (twist and lock) connection. It is larger than the BNC, but about the same as Type N. It has a frequency range of 0-11 GHz	C Plug (top) 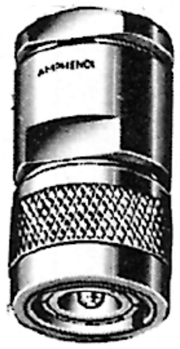
SC (OSSC) - The SC type connector is a screw version of the “C” connector.	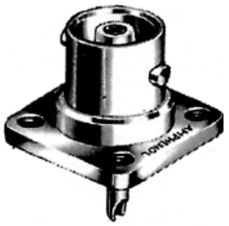 C jack (bottom)
SMA (OSM/3mm) - The SMA (Sub-Miniature A) connector was originally designed by Bendix Scintilla Corporation, but it has been manufactured by the Omni-Spectra division of M/ACOM (as the OSM connector) and many other electronic companies. It is a 50 Ω threaded connector. The main application of SMA connectors is on components for microwave systems. The connector normally has a frequency range to 18 GHz, but high performance varieties can be used to 26.5 GHz.	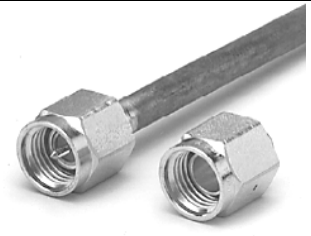 SMA Plug (above) 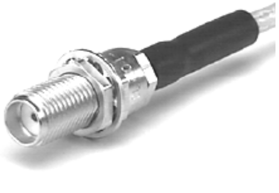 SMA Jack
SSMA (OSSM) - The SSMA is a microminiature version of the SMA. It is also 50 Ω and operates to 26.5 GHz with flexible cable or 40 GHz with semi-rigid cable.	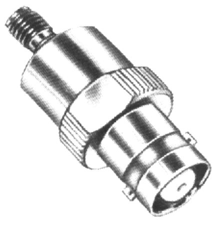 SSMA jack - BNC jack

SMC (OSMC) - The SMC (Sub-Miniature C) is a 50 Ω or 75 Ω connector that is smaller than the SMA. The connector can accept flexible cables with diameters of up to 3.17 mm (0.125 in.) for a frequency range of up to 7-10 GHz.	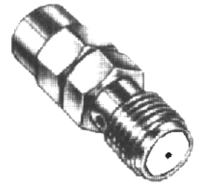 SMC Plug - SMA Jack (Some call this a SMC Jack, even though it has a female connector)
SMB (OSMB) - The SMB is like the SMC except it uses quick disconnect instead of threaded fittings. It is a 50 / 75 Ω connector which operates to 4 GHz with a low reflection coefficient and is useable to 10 GHz.	 SMB Plug (above) SMB Jack
TNC (OST) - The TNC (Threaded Neill-Concelman or threaded Navy Connector) is merely a threaded BNC. The function of the thread is to stop radiation at higher frequencies, so that the connector can work at frequencies up to 12 GHz (to 18 GHz when using semi-rigid cable). It can be 50 or 75 Ω. For size comparison here is a SMA plug - TNC plug 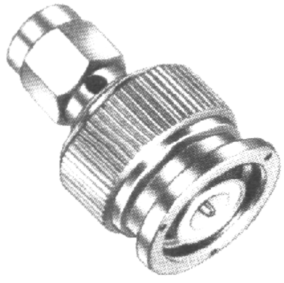	 TNC Plug (above) TNC Jack

Type N (OSN) - The 50 or 75 Ω Type N (Navy) connector was originally designed for military systems during World War II and is the most popular measurement connector for the frequency range of 1 to 11 GHz. The precision 50 Ω APC-N and other manufacturers high frequency versions operate to 18 GHz.

For size comparison here is a picture of Type N plug - TNC jack

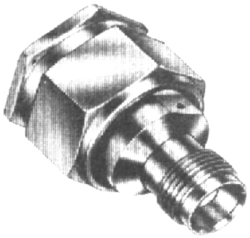

N Plug (above)

N Jack

Note: Always rotate the movable coupling nut of the plug, not the cable or fixed connector, when mating connectors. Since the center pin is stationary with respect to the jack, rotating the jack puts torque on the center pin. With TNC and smaller connectors, the center pin will eventually break off.

An approximate size comparison of these connectors is depicted below (not to scale).

Large ======================== Medium ============================ Small
SC 7mm N TNC/BNC 3.5mm SMA 2.4mm SSMA SMC

Figure 1 shows the frequency range of several connectors.

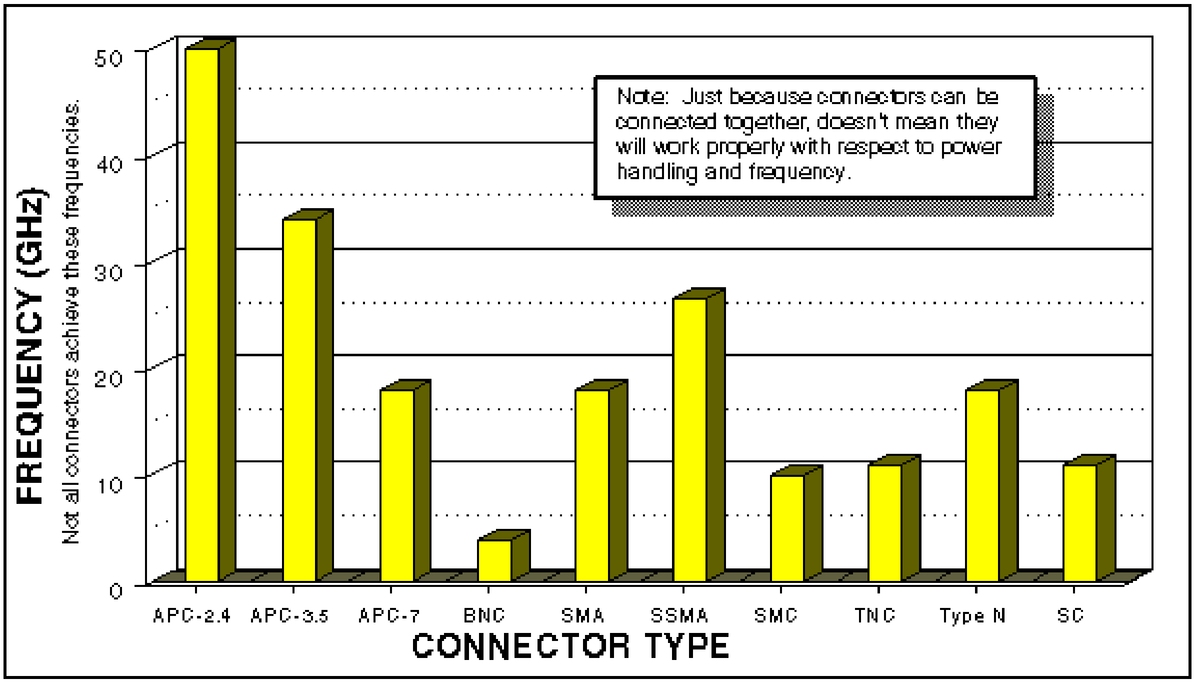

Note: Just because connectors can be connected together, doesn't mean they will work properly with respect to power handling and frequency.

Figure 1. Frequency Range of Microwave Connectors.

POWER DIVIDERS AND DIRECTIONAL COUPLERS

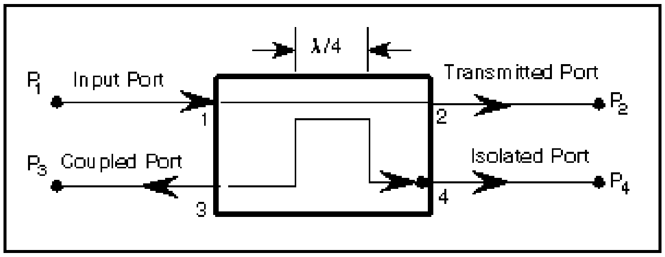

Figure 1. Directional Coupler.

A directional coupler is a passive device which couples part of the transmission power by a known amount out through another port, often by using two transmission lines set close enough together such that energy passing through one is coupled to the other. As shown in Figure 1, the device has four ports: input, transmitted, coupled, and isolated. The term "main line" refers to the section between ports 1 and 2. On some directional couplers, the main line is designed for high power operation (large connectors), while the coupled port may use a small SMA connector. Often the isolated port is terminated with an internal or external matched load (typically 50 ohms). It should be pointed out that since the directional coupler is a linear device, the notations on Figure 1 are arbitrary. Any port can be the input, (as in Figure 3) which will result in the directly connected port being the transmitted port, adjacent port being the coupled port, and the diagonal port being the isolated port.

Physical considerations such as internal load on the isolated port will limit port operation. The coupled output from the directional coupler can be used to obtain the information (i.e., frequency and power level) on the signal without interrupting the main power flow in the system (except for a power reduction - see Figure 2). When the power coupled out to port three is half the input power (i.e., 3 dB below the input power level), the power on the main transmission line is also 3 dB below the input power and equals the coupled power. Such a coupler is referred to as a 90 degree hybrid, hybrid, or 3 dB coupler. The frequency range for coaxial couplers specified by manufacturers is that of the coupling arm. The main arm response is much wider (i.e., if the spec is 2-4 GHz, the main arm could operate at 1 or 5 GHz - see Figure 3). However it should be recognized that the coupled response is periodic with frequency. For example, a λ/4 coupled line coupler will have responses at nλ/4 where n is an odd integer.

Common properties desired for all directional couplers are wide operational bandwidth, high directivity, and a good impedance match at all ports when the other ports are terminated in matched loads. These performance characteristics of hybrid or non-hybrid directional couplers are self-explanatory. Some other general characteristics will be discussed below.

COUPLING FACTOR

The coupling factor is defined as: $\mathit{Coupling\ factor\ (dB)} = -10 \log \frac{P_3}{P_1}$

where P_1 is the input power at port 1 and P3 is the output power from the coupled port (see Figure 1).

The coupling factor represents the primary property of a directional coupler. Coupling is not constant, but varies with frequency. While different designs may reduce the variance, a perfectly flat coupler theoretically cannot be built. Directional couplers are specified in terms of the coupling accuracy at the frequency band center. For example, a 10 dB coupling ± 0.5 dB means that the directional coupler can have 9.5 dB to 10.5 dB coupling at the frequency band center. The accuracy is due to dimensional tolerances that can be held for the spacing of the two coupled lines. Another coupling specification is frequency sensitivity. A larger frequency sensitivity will allow a larger frequency band of operation. Multiple quarter-wavelength coupling sections are used to obtain wide frequency bandwidth directional couplers. Typically this type of directional coupler is designed to a frequency bandwidth ratio and a maximum coupling ripple within the frequency band. For example a typical 2:1 frequency bandwidth coupler design that produces a 10 dB coupling with a ±0.1 dB ripple would, using the previous accuracy specification, be said to have 9.6 ± 0.1 dB to 10.4 ± 0.1 dB of coupling across the frequency range.

LOSS

In an ideal directional coupler, the main line loss port 1 to port 2 (P_1 - P_2) due to power coupled to the coupled output port is:

$$\mathit{Insertion\ loss\ (dB)} = 10 \log \left[1 - \frac{P_3}{P_1} \right]$$

The actual directional coupler loss will be a combination of coupling loss, dielectric loss, conductor loss, and VSWR loss. Depending on the frequency range, coupling loss becomes less significant above 15 dB coupling where the other losses constitute the majority of the total loss. A graph of the theoretical insertion loss (dB) vs. coupling (dB) for a dissipationless coupler is shown in Figure 2.

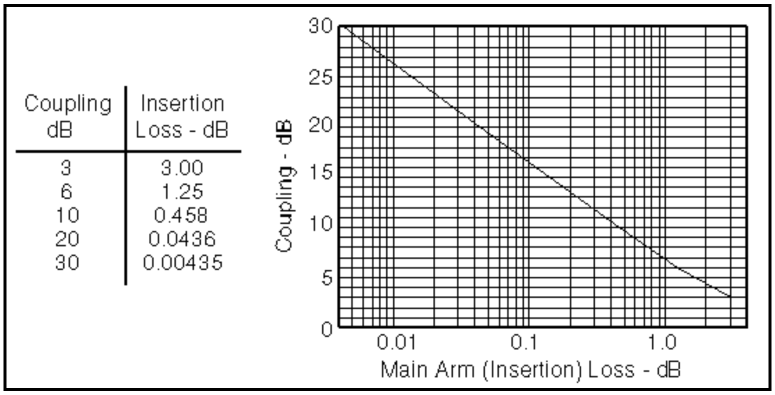

Coupling dB	Insertion Loss - dB
3	3.00
6	1.25
10	0.458
20	0.0436
30	0.00435

Figure 2. Coupling Insertion Loss.

ISOLATION

Isolation of a directional coupler can be defined as the difference in signal levels in dB between the input port and the isolated port when the two output ports are terminated by matched loads, or:

$$\mathit{Isolation\ (dB)} = -10 \log \frac{P_4}{P_1}$$

Isolation can also be defined between the two output ports. In this case, one of the output ports is used as the input; the other is considered the output port while the other two ports (input and isolated) are terminated by matched loads.

Consequently: $\mathit{Isolation\ (dB)} = -10 \log \frac{P_3}{P_2}$

The isolation between the input and the isolated ports may be different from the isolation between the two output ports. For example, the isolation between ports 1 and 4 can be 30 dB while the isolation between ports 2 and 3 can be a different value such as 25 dB. If both isolation measurements are not available, they can assumed to be equal. If neither are available, an estimate of the isolation is the coupling plus return loss (see VSWR section). The isolation should be as high as possible. In actual couplers the isolated port is never completely isolated. Some RF power will always be present. Waveguide directional couplers will have the best isolation.

If isolation is high, directional couplers are excellent for combining signals to feed a single line to a receiver for two-tone receiver tests. In Figure 3, one signal enters port P_3 and one enters port P_2, while both exit port P_1. The signal from port P_3 to port P_1 will experience 10 dB of loss, and the signal from port P_2 to port P_1 will have 0.5 dB loss. The internal load on the isolated port will dissipate the signal losses from port P_3 and port P_2. If the isolators in Figure 3 are neglected, the isolation measurement (port P_2 to port P_3) determines the amount of power from the signal generator F_2 that will be injected into the signal generator F_1. As the injection level increases, it may cause modulation of signal generator F_1, or even injection phase locking. Because of the symmetry of the directional coupler, the reverse injection will happen with the same possible modulation problems of signal generator F_2 by F_1. Therefore the isolators are used in Figure 3 to effectively increase the isolation (or directivity) of the directional coupler. Consequently the injection loss will be the isolation of the directional coupler plus the reverse isolation of the isolator.

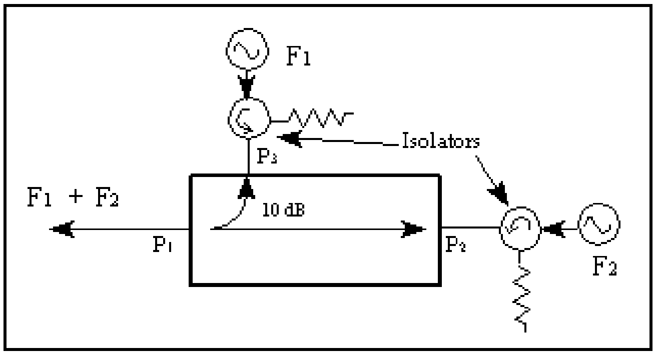

Figure 3. Two-Tone Receiver Tests.

DIRECTIVITY

Directivity is directly related to Isolation. It is defined as:

$$Directivity\ (dB) = -10 \log \frac{P_4}{P_3} = -10 \log \frac{P_4}{P_1} + 10 \log \frac{P_3}{P_1}$$

where: P_3 is the output power from the coupled port and P_4 is the power output from the isolated port.

The directivity should be as high as possible. Waveguide directional couplers will have the best directivity. Directivity is not directly measurable, and is calculated from the isolation and coupling measurements as:

Directivity (dB) = Isolation (dB) - Coupling (dB)

HYBRIDS

The hybrid coupler, or 3 dB directional coupler, in which the two outputs are of equal amplitude takes many forms. Not too long ago the quadrature (90 degree) 3 dB coupler with outputs 90 degrees out of phase was what came to mind when a hybrid coupler was mentioned. Now any matched 4-port with isolated arms and equal power division is called a hybrid or hybrid coupler. Today the characterizing feature is the phase difference of the outputs. If 90 degrees, it is a 90 degree hybrid. If 180 degrees, it is a 180 degree hybrid. Even the Wilkinson power divider which has 0 degrees phase difference is actually a hybrid although the fourth arm is normally imbedded.

Applications of the hybrid include monopulse comparators, mixers, power combiners, dividers, modulators, and phased array radar antenna systems.

AMPLITUDE BALANCE

This terminology defines the power difference in dB between the two output ports of a 3 dB hybrid. In an ideal hybrid circuit, the difference should be 0 dB. However, in a practical device the amplitude balance is frequency dependent and departs from the ideal 0 dB difference.

PHASE BALANCE

The phase difference between the two output ports of a hybrid coupler should be 0, 90, or 180 degrees depending on the type used. However, like amplitude balance, the phase difference is sensitive to the input frequency and typically will vary a few degrees.

The phase properties of a 90 degree hybrid coupler can be used to great advantage in microwave circuits. For example in a balanced microwave amplifier the two input stages are fed through a hybrid coupler. The FET device normally has a very poor match and reflects much of the incident energy. However, since the devices are essentially identical the reflection coefficients from each device are equal. The reflected voltage from the FETs are in phase at the isolated port and are 180° different at the input port. Therefore, all of the reflected power from the FETs goes to the load at the isolated port and no power goes to the input port. This results in a good input match (low VSWR).

If phase matched lines are used for an antenna input to a 180° hybrid coupler as shown in Figure 4, a null will occur directly between the antennas. If you want to receive a signal in that position, you would have to either change the hybrid type or line length. If you want to reject a signal from a given direction, or create the difference pattern for a monopulse radar, this is a good approach.

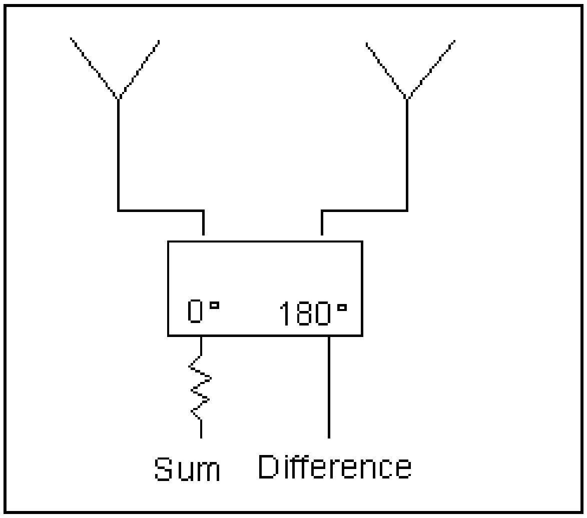

Figure 4. Balanced Antenna Input.

OTHER POWER DIVIDERS

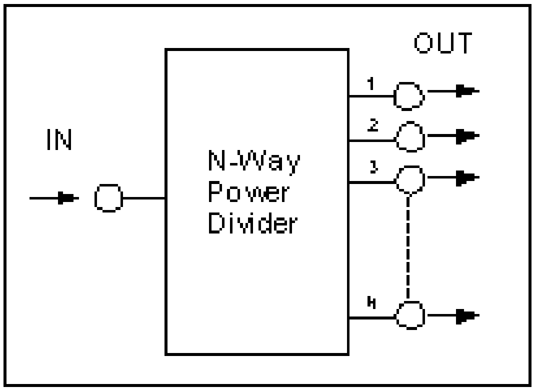

Figure 5. Power Divider.

Both in-phase (Wilkinson) and quadrature (90°) hybrid couplers may be used for coherent power divider applications. The Wilkinson's power divider has low VSWR at all ports and high isolation between output ports. The input and output impedances at each port is designed to be equal to the characteristic impedance of the microwave system. A typical power divider is shown in Figure 5. Ideally, input power would be divided equally between the output ports. Dividers are made up of multiple couplers, and like couplers, may be reversed and used as multiplexers. The drawback is that for a four channel multiplexer, the output consists of only 1/4 the power from each, and is relatively inefficient. Lossless multiplexing can only be done with filter networks.

Coherent power division was first accomplished by means of simple Tee junctions. At microwave frequencies, waveguide tees have two possible forms - the H-Plane or the E-Plane. These two junctions split power equally, but because of the different field configurations at the junction, the electric fields at the output arms are in-phase for the H-Plane tee and are anti-phase for the E-Plane tee. The combination of these two tees to form a hybrid tee allowed the realization of a four-port component which could perform the vector sum (Σ) and difference (Δ) of two coherent microwave signals. This device is known as the magic tee.

POWER COMBINERS

Since hybrid circuits are bi-directional, they can be used to split up a signal to feed multiple low power amplifiers, then recombine to feed a single antenna with high power as shown in Figure 6. This approach allows the use of numerous less expensive and lower power amplifiers in the circuitry instead of a single high power TWT. Yet another approach is to have each solid state amplifier (SSA) feed an antenna and let the power be combined in space or be used to feed a lens which is attached to an antenna. (See Section 3-4.)

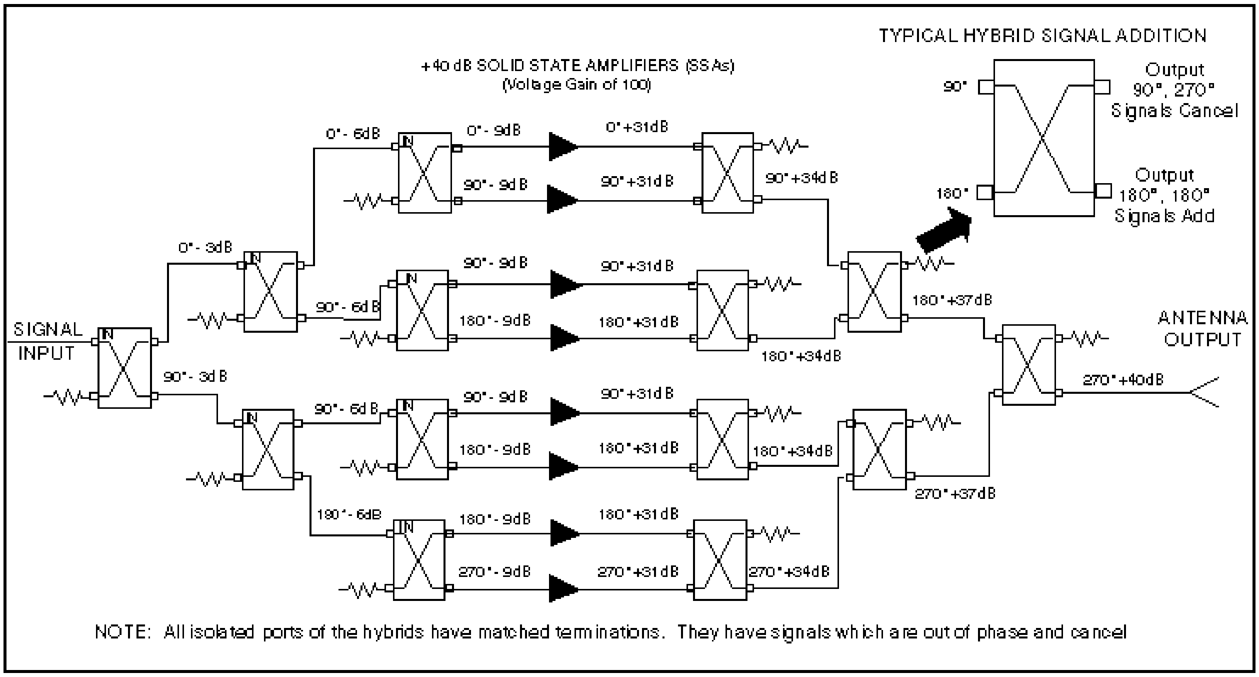
TYPICAL HYBRID SIGNAL ADDITION
+40 dB SOLID STATE AMPLIFIERS (SSAs)
(Voltage Gain of 100)
90°
180°
Output 90°, 270° Signals Cancel
Output 180°, 180° Signals Add
SIGNAL INPUT
IN
0°- 3dB
90°- 3dB
0°- 6dB
90°- 6dB
90°- 6dB
180°- 6dB
0°- 9dB
90°- 9dB
90°- 9dB
180°- 9dB
90°- 9dB
180°- 9dB
180°- 9dB
270°- 9dB
0°+31dB
90°+31dB
90°+31dB
180°+31dB
90°+31dB
180°+31dB
180°+31dB
270°+31dB
90°+34dB
180°+34dB
180°+34dB
270°+34dB
180°+37dB
270°+37dB
270°+40dB
ANTENNA OUTPUT

NOTE: All isolated ports of the hybrids have matched terminations. They have signals which are out of phase and cancel

Figure 6. Combiner Network.

Sample Problem:

If two 1 watt peak unmodulated RF carrier signals at 10 GHz are received, how much peak power could one measure?

A. 0 watts

B. 0.5 watts

C. 1 watt

D. 2 watts

E. All of these

The answer is all of these as shown in Figure 7.

The phase error could be due to a hybrid being used to combine the same signal received from two aircraft antennas.

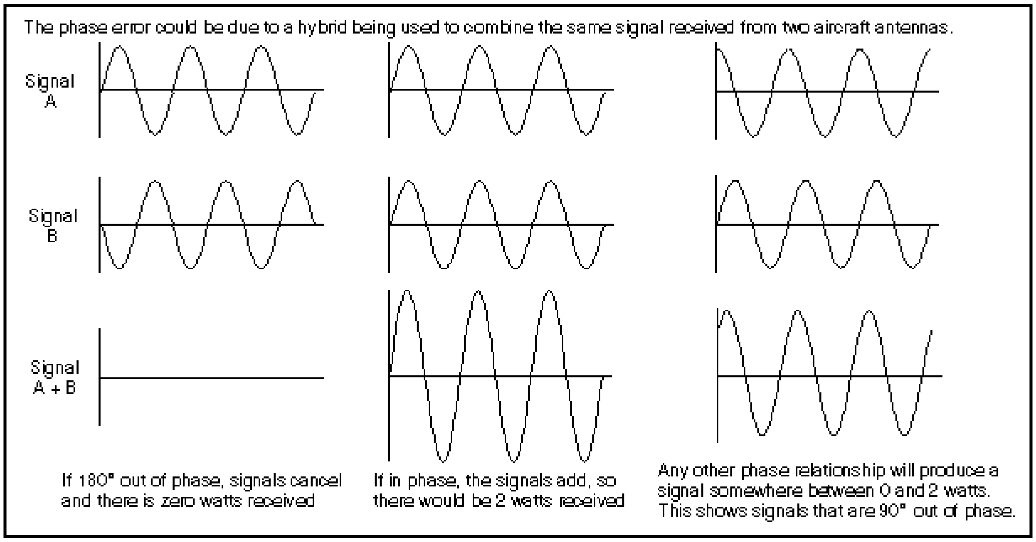
Signal A
Signal B
Signal A + B

If 180° out of phase, signals cancel and there is zero watts received

If in phase, the signals add, so there would be 2 watts received

Any other phase relationship will produce a signal somewhere between 0 and 2 watts. This shows signals that are 90° out of phase.

Figure 7. Sine Waves Combined Using Various Phase Relationships.

ATTENUATORS / FILTERS / DC BLOCKS

ATTENUATORS

An attenuator is a passive microwave component which, when inserted in the signal path of a system, reduces the signal by a specified amount. They normally possess a low VSWR which makes them ideal for reducing load VSWR in order to reduce measurement uncertainties. They are sometimes used simply to absorb power, either to reduce it to a measurable level, or in the case of receivers to establish an exact level to prevent overload of following stages.

Attenuators are classified as either fixed or variable and either reflective or non-reflective. The fixed and variable attenuators are available in both waveguide and coaxial systems. Most of the receivers under 20 GHz use coaxial type attenuators.

FIXED

The performance characteristics of a fixed attenuator are:

1. input and output impedances
2. flatness with frequency
3. average and peak power handling capability
4. temperature dependence

VARIABLE

The variable attenuator can be subdivided into two kinds: step attenuator and continuously variable attenuator. In a step attenuator, the attenuation is changed in steps such as 10 dB, 1 dB or 0.5 dB. In a continuously variable attenuator, the attenuation is changed continuously and a dial is usually available to read the attenuation either directly or indirectly from a calibration chart.

For a variable attenuator, additional characteristics should be considered, such as:

1. amount or range of attenuations
2. insertion loss in the minimum attenuation position
3. incremental attenuation for step attenuator
4. accuracy of attenuation versus attenuator setting
5. attenuator switching speed and switching noise.

REFLECTIVE

A reflective attenuator reflects some portion of the input power back to the driving source. The amount reflected is a function of the attenuation level. When PIN diodes are zero or reverse biased, they appear as open circuits when shunting a transmission line. This permits most of the RF input power to travel to the RF output. When they are forward biased, they absorb some input, but simultaneously reflect some back to the input port. At high bias current, most RF will be reflected back to the input resulting in a high input VSWR and high attenuation.

ABSORPTIVE

The VSWR of a non-reflective (absorptive) PIN diode attenuator remains good at any attenuation level (bias state). This is accomplished by configuring the diodes in the form of a Pi network that remains matched for any bias state or by use of a 90° hybrid coupler to cancel the waves reflected to the input connector.

MICROWAVE FILTERS

INTRODUCTION

Microwave filters are one of the most important components in receivers. The main functions of the filters are: (1) to reject undesirable signals outside the filter pass band and (2) to separate or combine signals according to their frequency. A good example for the latter application is the channelized receiver in which banks of filters are used to separate input signals. Sometimes filters are also used for impedance matching. Filters are almost always used before and after a mixer to reduce spurious signals due to image frequencies, local oscillator feedthrough, and out-of-frequency band noise and signals. There are many books which are devoted to filter designs. There are many kinds of filters used in microwave receivers, so it is impossible to cover all of them.

If a filter is needed on the output of a jammer, it is desirable to place it approximately half way between the jammer and antenna vs. adjacent to either. The transmission line attenuation improves the VSWR of the filter at the transmitter. This may allow use of a less expensive filter, or use of a reflective filter vs. an absorptive filter.

A filter is a two-port network which will pass and reject signals according to their frequencies. There are four kinds of filters according to their frequency selectivities. In the examples that follow, f_L = low frequency, f_M = medium frequency, and f_H = high frequency. Their names reflect their characteristics, and they are:

1. A low-pass filter which passes the low frequency signals below a predetermined value as shown in Figure 1.

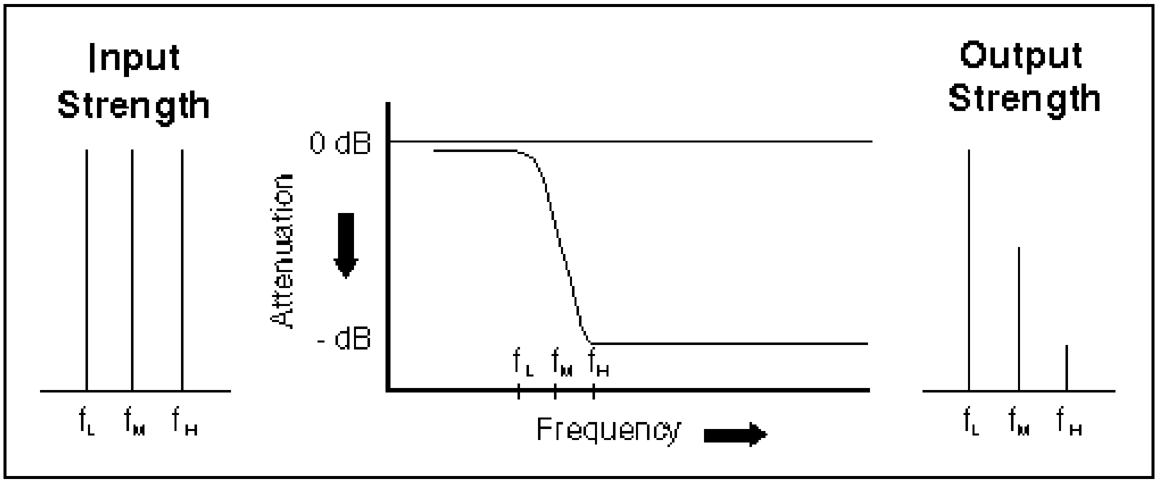

Figure 1. Low-Pass Filter.

2. A high-pass filter which passes the high frequency signals above a predetermined value as in Figure 2.

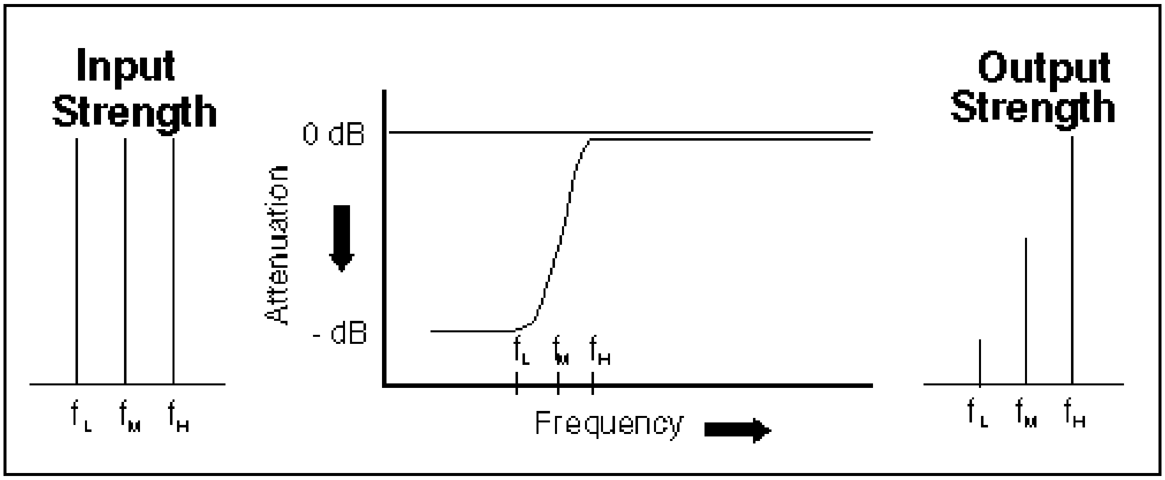
Input
Strength
f_L f_M f_H
0 dB
Attenuation
- dB
f_L f_M f_H
Frequency
Output
Strength
f_L f_M f_H

Figure 2. High-Pass Filter.

3. A band-pass filter which passes signals between two predetermined frequencies as shown in Figure 3.

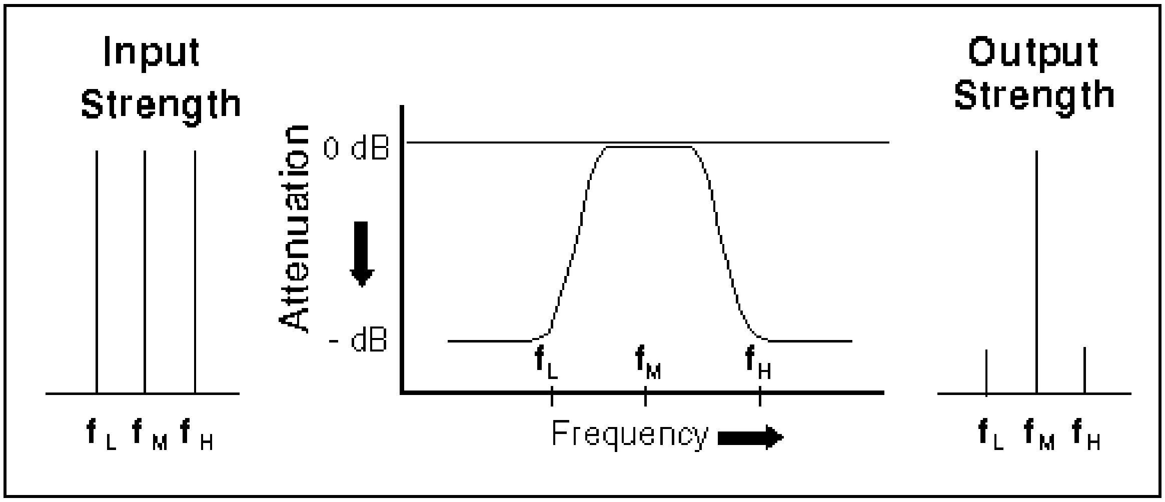
Input
Strength
f_L f_M f_H
0 dB
Attenuation
- dB
f_L f_M f_H
Frequency
Output
Strength
f_L f_M f_H

Figure 3. Band-Pass Filter.

A band-pass filter with different skirt slopes on the two sides of the pass band is sometimes referred to as an asymmetrical filter. In this filter the sharpness of the rejection band attenuation is significantly different above and below the center frequency. One additional note regarding band-pass filters or filters in general, their performance should always be checked in the out-of-band regions to determine whether or not they possess spurious responses. In particular they should be checked at harmonics of the operating frequency.

4. A band reject filter (sometimes referred to as a bandstop or notch filter) which rejects signals between two predetermined frequencies such as high power signals from the aircraft's own radar as shown in Figure 4.

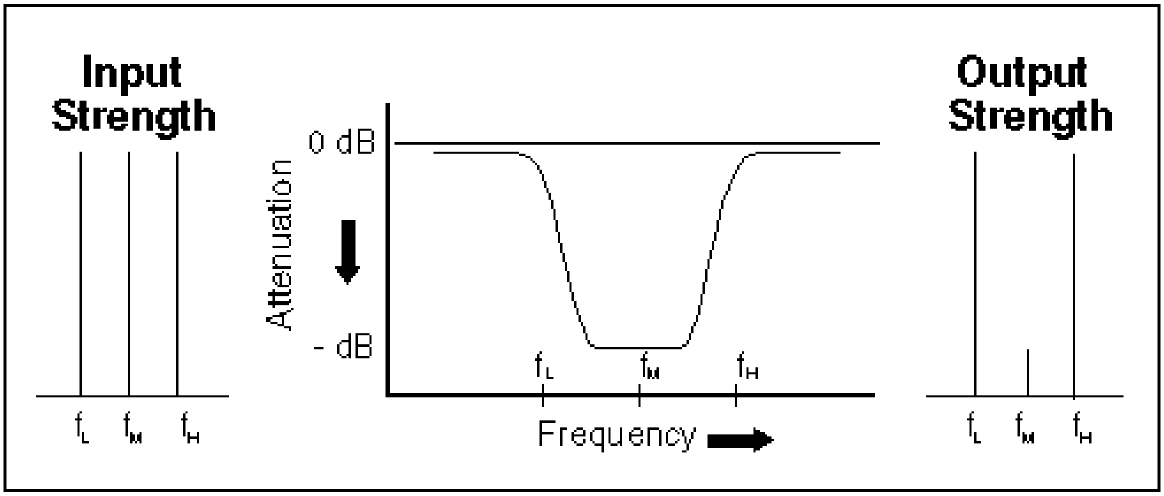

Figure 4. Band-Reject Filter.

In general, filters at microwave frequencies are composed of resonate transmission lines or waveguide cavities that, when combined, reflect the signal power outside the filter frequency pass band and provide a good VSWR and low loss within the frequency pass band. As such, specifications for filters are maximum frequency, pass band loss, VSWR, and rejection level at a frequency outside of the pass band. The trade-offs for filters are a higher rejection for a fixed frequency pass band or a larger frequency pass band for a fixed rejection, which requires a filter with more resonators, which produce higher loss, more complexity, and larger size.

DC BLOCKS

DC Blocks are special connectors which have a capacitor (high pass filter) built into the device. There are three basic types:

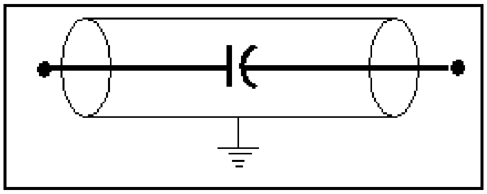

Figure 5. Inside DC Block.

1. INSIDE - The high pass filter is in series with the center conductor as shown in Figure 5. DC is blocked on the center conductor.
2. OUTSIDE - The high pass filter is in series with the cable shield as shown in Figure 6.
3. INSIDE/OUTSIDE - A high pass arrangement is connected to both the inner and outer conductors.

DC Blocks are ideal for filtering DC, 60 Hz, and 400 Hz from the RF line.

In general, capacitors with a large value of capacitance do not have the least loss at microwave frequencies. Also, since capacitance is proportional to size, a large size produces more capacitance with more inductance. Because of these reasons, D.C. blocks are typically available with a high pass frequency band starting in the region of 0.1 to 1 GHz.

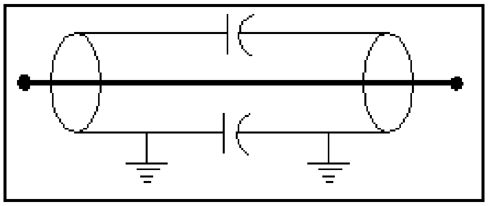

Figure 6. Outside DC Block.

TERMINATIONS / DUMMY LOADS

A termination is a one-port device with an impedance that matches the characteristic impedance of a given transmission line. It is attached to a certain terminal or port of a device to absorb the power transmitted to that terminal or to establish a reference impedance at that terminal. Important parameters of a termination are its VSWR and power handling capacity. In a receiver, terminations are usually placed at various unconnected ports of components such as hybrid and power dividers to keep the VSWR of the signal path low. It is extremely important that the isolated port in a directional coupler and the unused port of a power divider (i.e., only three ports of a four-way power divider are used) be properly terminated. All of the design considerations of directional couplers and power dividers are based on the fact that all ports are terminated with matched loads. If an unused port is not properly terminated, then the isolation between the output ports will be reduced which may severely degrade the performance of the receiver.

A termination is the terminology used to refer to a low power, single terminal device intended to terminate a transmission line. Similar devices designed to accommodate high power are generally termed dummy loads.

TERMINATIONS

Terminations are employed to terminate unconnected ports on devices when measurements are being performed. They are useful as dummy antennas and as terminal loads for impedance measurements of transmission devices such as filters and attenuators.

The resistive elements in most terminations are especially fabricated for use at microwave frequencies. Two types are commonly employed: (1) resistive film elements, and (2) molded resistive tapers. The resistive film is very thin compared to the skin depth and normally very short relative to wavelength at the highest operating frequency. The molded taper consists of a dissipative material evenly dispersed in a properly cured dielectric medium. Both forms of resistive elements provide compact, rugged terminations suitable for the most severe environmental conditions with laboratory stability and accuracy.

Terminations should be properly matched to the characteristic impedance of a transmission line. The termination characteristics of primary concern are:

a. operating frequency range
b. average power handling capability
c. operating temperature range
d. VSWR
e. size
f. weight

Many microwave systems employ directional couplers which require terminations on at least one port, and most have various modes of operation or test where terminations are needed on certain terminals.

A matched termination of a generalized transmission line is ideally represented by an infinite length of that line having small, but non-zero loss per unit length so that all incident energy is absorbed and none is reflected.

Standard mismatches are useful as standards of reflection in calibrating reflectometer setups and other impedance measuring equipment. They are also used during testing to simulate specific mismatches which would be encountered on the terminals of components once the component is installed in the actual system.

The following table shows common mismatches with the impedance that can provide the mismatch.

Common Mismatches (Z_O = 50 Ω)		
Ratio	Z_L (higher)	Z_L (lower)
1.0 : 1	50 Ω (matched)	50 Ω (matched)
1.25 : 1	62.5 Ω	40 Ω
1.50 : 1	75 Ω	33.3 Ω
2.00 : 1	100 Ω	25 Ω

DUMMY LOADS

A dummy load is a high power one port device intended to terminate a transmission line. They are primarily employed to test high power microwave systems at full power capacity. Low power coaxial loads are generally termed terminations and typically handle one watt or less.

Most radars or communications systems have a dummy load integrated into them to provide a non-radiating or EMCON mode of operation, or for testing (maintenance).

Three types of dissipative material are frequently employed in dummy loads: (1) lossy plastic, (2) refractory, and (3) water.

The lossy plastic consists of particles of lossy material suspended in plastic medium. This material may be designed to provide various attenuations per unit length but is limited as to operating temperature. It is employed primarily for low power applications.

The refractory material is a rugged substance that may be operated at temperatures up to 1600°F. It is virtually incapable of being machined by ordinary means but is often fabricated through diamond wheel grinding processes. Otherwise material must be fired in finished form. Such material is employed in most high power applications.

The dissipative properties of water are also employed for dummy load applications. Energy from the guide is coupled through a leaky wall to the water which flows alongside the main guide. Water loads are employed for extremely high power and calorimetric applications.

While dummy loads can operate over full waveguide bands, generally a more economical unit can be manufactured for use over narrower frequency ranges.

The power rating of a dummy load is a complex function dependent upon many parameters, including average and peak power, guide pressure, external temperature, guide size, air flow, and availability of auxiliary coolant. The average and peak powers are interrelated in that the peak power capacity is a function of the operating temperature which in turn is a function of the average power.

CIRCULATORS AND DIPLEXERS

A microwave circulator is a nonreciprocal ferrite device which contains three or more ports. The input from port n will come out at port n + 1 but not out at any other port. A three-port ferrite junction circulator, usually called the Y-junction circulator, is most commonly used. They are available in either rectangular waveguide or strip- line forms. The signal flow in the three-port circulator is assumed as $1\rightarrow2$, $2\rightarrow3$, and $3\rightarrow1$ as shown in Figure 1.

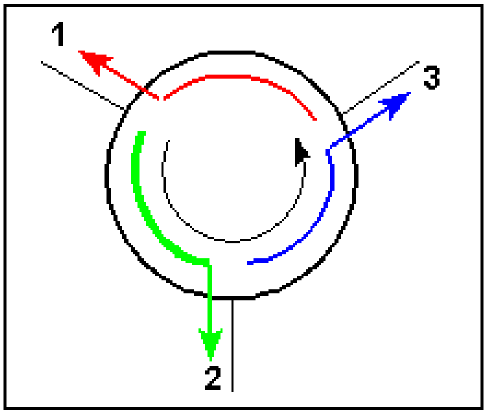

Figure 1. Symbolic Expression for a Y-Junction Circulator.

If port 1 is the input, then the signal will come out of port 2; in an ideal situation, no signal should come out of port 3 which is called the isolated port. The insertion loss of the circulator is the loss from 1 to 2, while the loss from 1 to 3 is referred to as isolation. A typical circulator will have a few tenths of a dB insertion loss from port 1 to 2 and 20 dB of isolation from port 1 to 3 for coaxial circulators (30 dB or more for waveguide circulators). When the input is port 2, the signal will come out of port 3 and port 1 is the isolated port. Similar discussions can be applied to port 3.

Since circulators contain magnets, they should not be mounted near ferrous metals since the close proximity of metals like iron can change the frequency response.

As shown in Figure 2, if one port of a circulator is loaded, it becomes an isolator, i.e., power will pass from ports one to two, but power reflected back from port two will go to the load at port three versus going back to port one.

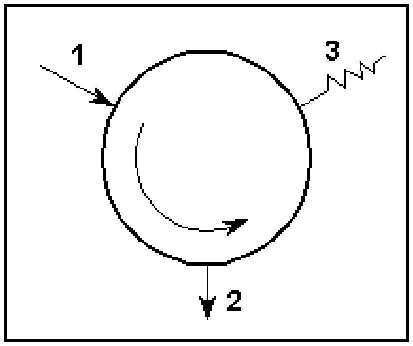

Figure 2. Isolator From A Circulator.

As shown in Figure 3 this circulator is made into a diplexer by adding a high pass filter to port two. Frequencies from port one that are below 10 GHz will be reflected by port two. Frequencies above 10 GHz will pass through port two. At the 10 GHz crossover frequency of the diplexer, a 10 GHz signal will be passed to both ports two and three but will be half power at each port. Diplexers or triplexers (one input and three output bands), must be specifically designed for the application.

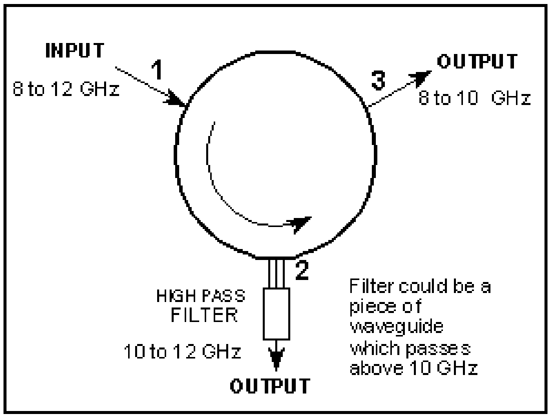

Figure 3. Diplexer From A Circulator.

Another useful device is the 4-port Faraday Rotator Circulator shown symbolically in Figure 4. These waveguide devices handle very high power and provide excellent isolation properties. It is useful when measurements must be made during high power application as shown. A water load is used to absorb the high power reflections so that a reasonable power level is reflected to the receiver or measurement port.

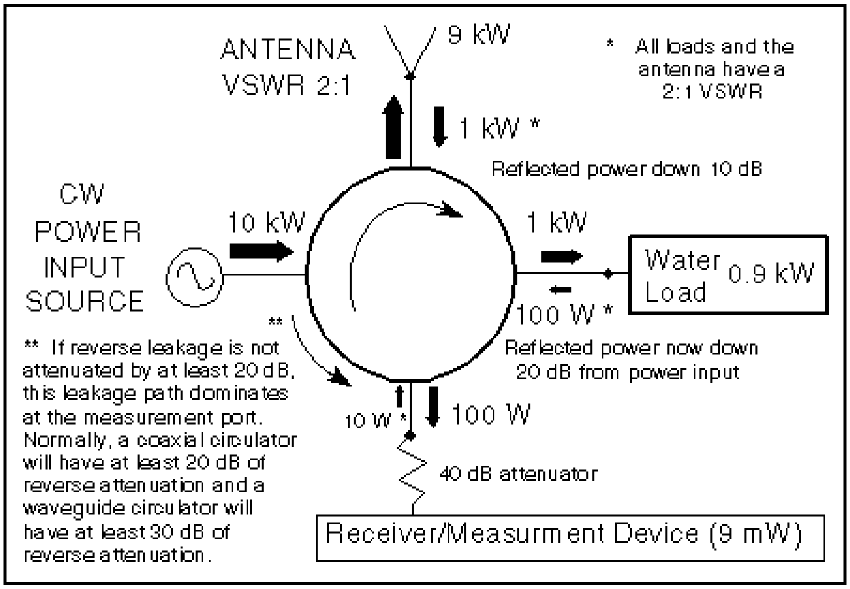
ANTENNA
VSWR 2:1
9 kW
* All loads and the antenna have a 2:1 VSWR
1 kW *
Reflected power down 10 dB
CW
POWER
INPUT
SOURCE
10 kW
1 kW
Water
Load
0.9 kW
100 W *
Reflected power now down 20 dB from power input
**
** If reverse leakage is not attenuated by at least 20 dB, this leakage path dominates at the measurement port. Normally, a coaxial circulator will have at least 20 dB of reverse attenuation and a waveguide circulator will have at least 30 dB of reverse attenuation.
10 W *
100 W
40 dB attenuator
Receiver/Measurment Device (9 mW)

Figure 4. Faraday Rotator Circulator.

The Maximum Input Power to a Measurement Device - The ideal input to a measurement device is in the 0 to 10 dBm (1 to 10 mW) range. Check manufacturer's specification for specific maximum value.

If the RF transmission lines and their components (antenna, hybrid, etc.) can support the wider frequency range, circulators could be used to increase the number of interconnecting RF ports from two as shown in Figure 5, to four as shown in Figure 6. Figure 7 shows an alternate configuration using diplexers which could actually be made from circulators as shown previously in Figure 3.

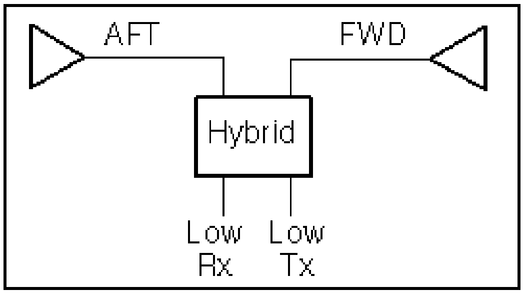
AFT
FWD
Hybrid
Low
Rx
Low
Tx

Figure 5. Low Band Configuration.

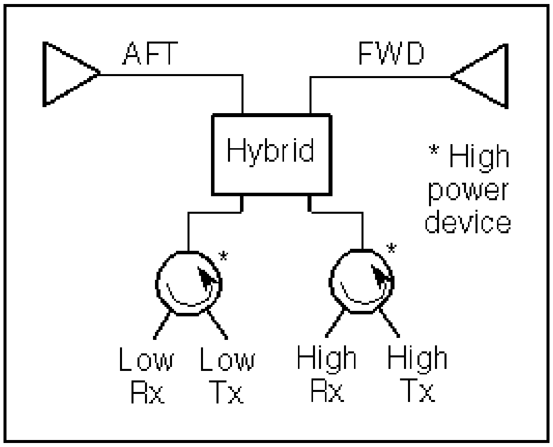
AFT
FWD
Hybrid
* High power device
*
*
Low
Rx
Low
Tx
High
Rx
High
Tx

Figure 6. Low/High Band Configuration.

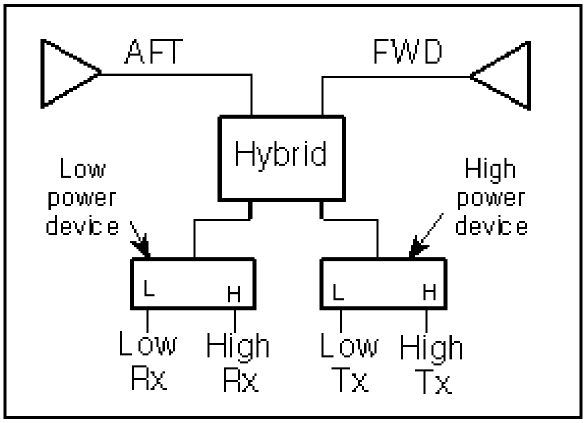
AFT
FWD
Hybrid
Low power device
High power device
L
H
L
H
Low
Rx
High
Rx
Low
Tx
High
Tx

Figure 7. Alternate Low/High Band Configuration.

MIXERS AND FREQUENCY DISCRIMINATORS

Mixers are used to convert a signal from one frequency to another. This is done by combining the original RF signal with a local oscillator (LO) signal in a non-linear device such as a Schottky-barrier diode.

The output spectrum includes:

- The original inputs, LO and RF
- All higher order harmonics of LO and RF
- The two primary sidebands, LO ± RF (m,n = 1)
- All higher order products of mLO ± nRF (where m,n are integers)
- A DC output level

The desired output frequency, commonly called the intermediate frequency (IF), can be either the lower (LO-RF) or upper (LO+RF) sideband. When a mixer is used as a down converter, the lower sideband is the sideband of interest.

A microwave balanced mixer makes use of the 3 dB hybrid to divide and recombine the RF and LO inputs to two mixing diodes. The 3 dB hybrid can be either the 90° or 180° type. Each has certain advantages which will be covered later. The critical requirement is that the LO and RF signals be distributed uniformly (balanced) to each mixer diode.

Figure 1 is a typical balanced mixer block diagram. The mixer diodes are reversed relative to each other; the desired frequency (IF) components of each diode are then in-phase while the DC outputs are positive and negative respectively.

The two diode outputs are summed in a tee where the DC terms cancel and only the desired IF component exists at the IF port.

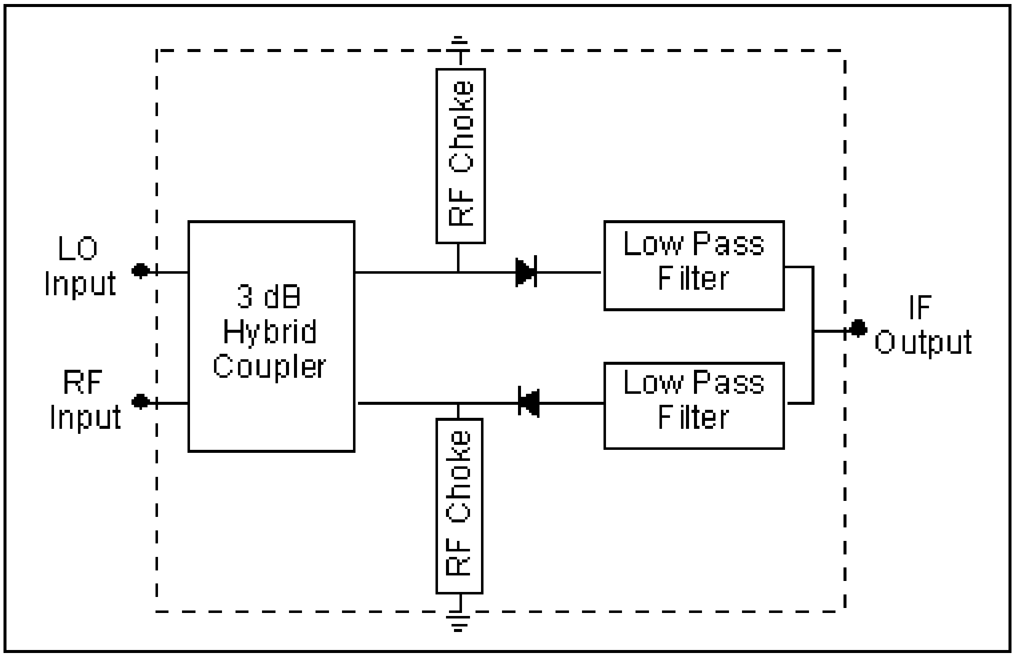

Figure 1. Mixer Block Diagram.

Other types of mixers exist, including the double-balanced mixer, and the Ortho-Quad® (quadrature fed dual) mixer. The relative advantages and disadvantages of each of the four types are summarized in Table 1.

Table 1. Mixer Comparison.

Mixer Type	VSWR [1]	Conversion Loss [2]	LO/RF Isolation [3]	Harmonic Suppression [4]	Dynamic Range	IF Bandwidth
90° Hybrid	good	lowest	poor	poor-fair	high	wide
180° Hybrid	poor	low	good	good	high	wide
Double-Balanced	poor	low	Very good -excellent	very good	high	extremely wide
Ortho Quad	good	low	very good	fair	high	wide

NOTES:
(1) Poor = 2.5:1 typical; Good = 1.3:1 typical
(2) Conversion loss: lowest: 5-7 dB typical; Low 7-9 dB typical
(3) Poor: 10 dB typical; Good: 20 dB typical; Very Good: 25-30 dB typical; Excellent: 35-40 dB typical
(4) Poor: partial rejection of LO/RF even harmonics
Fair: slightly better
Good: can reject all LO even harmonics
Very Good: can reject all LO and RF even harmonics

Used in various circuits, mixers can act as modulators, phase detectors, and frequency discriminators.

The phase discriminators can serve as a signal processing network for systems designed to monitor bearing, polarization, and frequency of AM or FM radiated signals.

A frequency discriminator uses a phase discriminator and adds a power divider and delay line at the RF input as shown in Figure 2. The unknown RF signal "A" is divided between a reference and delay path. The differential delay (T) creates a phase difference (θ) between the two signals which is a linear function of frequency (f) and is given by $\theta = 2\pi fT$.

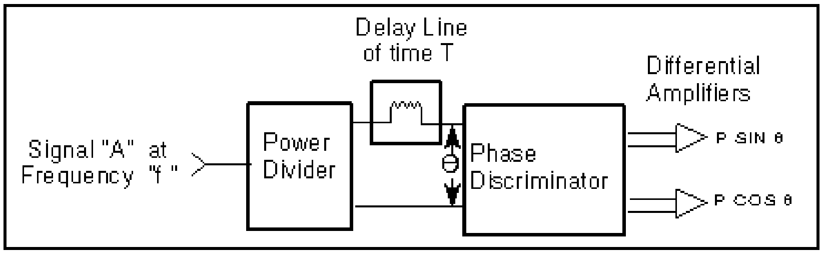

Figure 2. Frequency Discriminator.

When the two output signals are fed to the horizontal and vertical input of an oscilloscope, the resultant display angle will be a direct function of frequency.

DETECTORS

A detector is used in receiver circuits to recognize the presence of signals. Typically a diode or similar device is used as a detector. Since this type of detector is unable to distinguish frequency, they may be preceded by a narrow band-pass filter.

A typical simplistic circuit is shown in Figure 1.

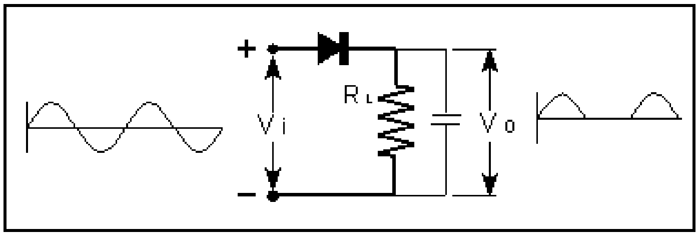

Figure 1. Typical Diode Detector Circuit.

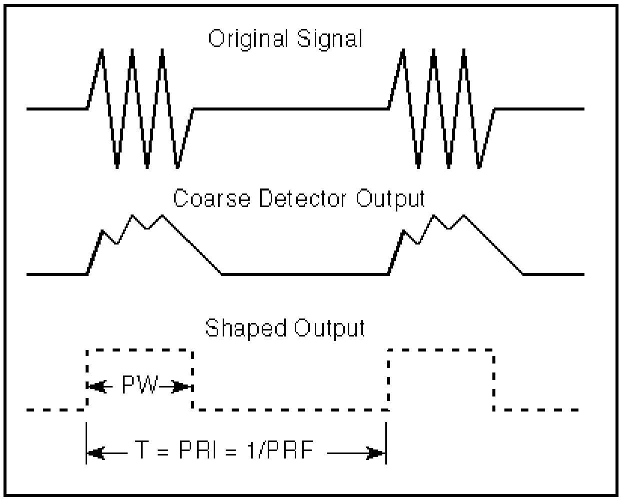

Figure 2. Demodulated Envelope Output.

To integrate a pulse radar signal, we can add capacitance to the circuit in parallel with the output load R_L to store energy and decrease the bleed rate. Figure 2 shows a typical input/output waveform which detects the envelope of the pulse radar signal. From this information pulse width and PRF characteristics can be determined for the RWR UDF comparison.

When the diode is reverse biased, very little current passes through unless the reverse breakdown voltage is exceeded. When forward biased and after exceeding the cut-in voltage, the diode begins to conduct as shown in Figure 3. At low voltages, it first operates in a square law region. Detectors operating in this region are known as small signal type. If the voltage is higher, the detector operates in a linear region, and is known as the large signal type.

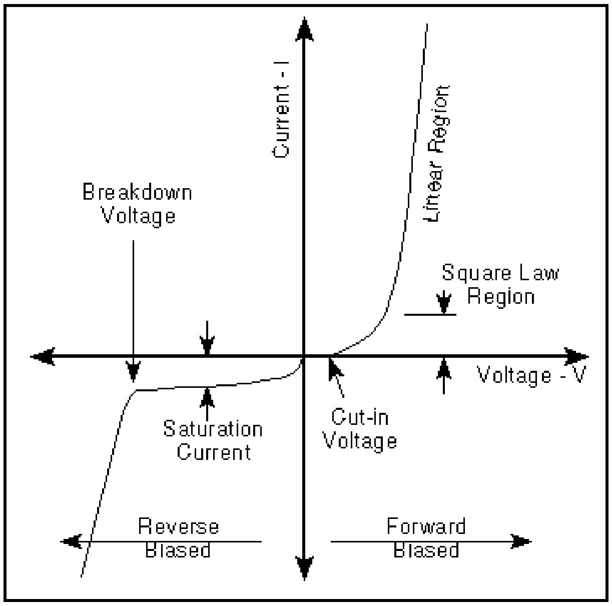

Figure 3. Diode Electrical Characteristics.

The power/voltage characteristics for a typical diode detector is shown in Figure 4.

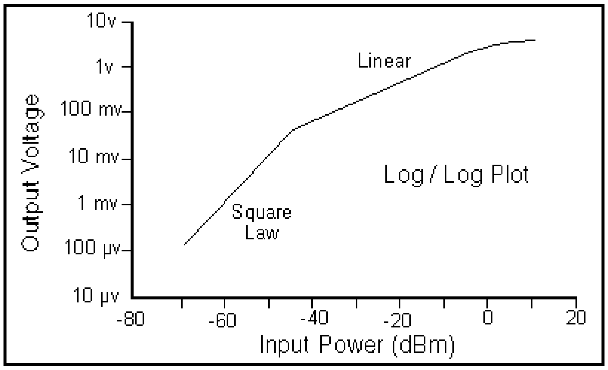

Figure 4. Diode Power/Voltage Characteristic.

Square Law Detector

In the square law region, the output voltage V_o is proportional to the square of the input voltage V_i, thus V_o is proportional to the input power.

$V_o = nV_i^2 = nP_i$ or $P_i \propto V_o$
Where n is the constant of proportionality

Linear Detector

In the linear detection region, the output voltage is given by:

$V_o = mV_i$ and since $P=V^2/R$, $P_i \propto V_o^2$
Where m is the constant of proportionality

Log Detector Amplifier

Another type of detector arrangement is the Log detector amplifier circuit shown in Figure 5. It is formed by using a series of amplifiers and diode detectors. Due to the nature of the amplifier/diode characteristics, the output voltage is related to the power by:

$P_i \propto 10^{pV_o + q}$
Where p and q are constants of proportionality

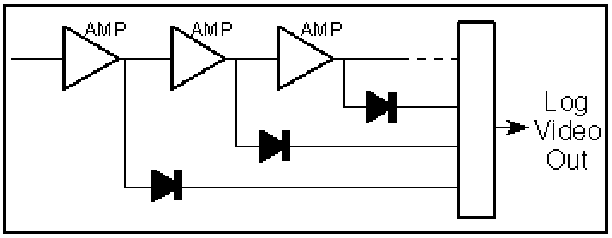

Figure 5. Log Detector.

The Log detector has good range, but is hampered by large size when compared to a single diode detector.

Pulse Width Measurements

If the pulse width of a signal was specified at the one-half power point, the measurements of the detected signal on an oscilloscope would vary according to the region of diode operation. If the region of operation is unknown, a 3 dB attenuator should be inserted in the measurement line. This will cause the power to decrease by one-half. The (temporary) peak amplitude on the oscilloscope becomes the amplitude reference point for measuring the pulse width when the external 3 dB attenuator is removed.

These voltage levels for half power using the three types of detectors are shown in Table 1.

Table 1. Detector Characteristics.

	Square Law	Linear	Log
Output Voltage When Input Power is reduced by Half (3 dB)	0.5 V_{in}	0.707 V_{in}	A very small value. ~ 0.15 V_{in} for typical 5 stage log amplifier
Sensitivity & Dynamic Range	Good sensitivity Small dynamic range	Less sensitivity Greater dynamic range	Poorest sensitivity Greatest dynamic range (to 80 dB)

Also see the Microwave Measurements section subsection entitled “Half Power or 3 dB Measurement Point.”

This page intentionally left blank.

RF / MICROWAVE AMPLIFIERS

An amplifier is one of the most essential and ubiquitous elements in modern RF and microwave systems. Fundamentally, an amplifier is a type of electronic circuit used to convert low-power signals into ones of significant power. The specific requirements for amplification are as varied as the systems where they are used. Amplifiers are realized using a wide range of different technologies, and are available in many form factors. While the performance of an amplifier can be measured by wide range of attributes, several important ones include: gain, power added efficiency (PAE), input and output return loss, 1-dB compression point (P_{1dB}), stability, linearity, intermodulation distortion, and noise figure.

Traveling Wave Tubes (TWTs)

A traveling-wave tube (TWT) is a specialized vacuum tube used in electronics to amplify radio frequency (RF) signals to very high power. A TWT is a component of an electronic assembly known as a traveling-wave tube amplifier (TWTA), often pronounced "Tweet-uh." While trades between these parameters exist, modern TWTAs are capable of providing very high gains (>50 dB), multiple octaves of bandwidth, high efficiency, and output powers that range from tens to thousands of watts. Pulsed TWTAs can reach even higher output powers. Figure 1 presents a simplified diagram of a helix-type TWT.

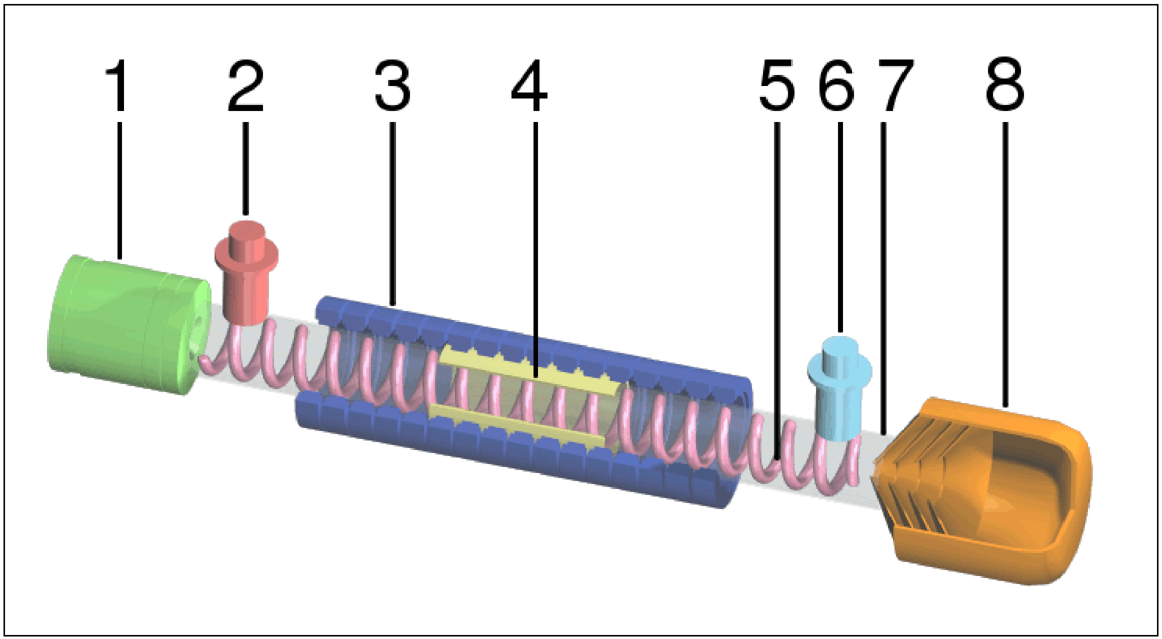

Figure 1. Cutaway view of a helix TWT.
(1) Electron gun; (2) RF input; (3) Magnets; (4) Attenuator;
(5) Helix coil; (6) RF output; (7) Vacuum tube; (8) Collector.
Reproduced from Wikipedia.

The device is an elongated vacuum tube with an electron gun (a heated cathode that emits electrons) at one end. A solenoid coil wrapped around the tube creates a magnetic field which focuses the electrons into a beam, which then passes down the middle of a wire helix that stretches the length of the tube, finally striking a collector at the other end. An input signal is coupled into the helix near the emitter, is amplified by the electron beam as it travels down the length of the tube, and the amplified signal is coupled to an external port near the collector.

The helix, which acts as a RF transmission line, delays the signal to near the same propagation speed as the electron beam. The speed at which the electromagnetic wave travels down the tube can be varied by changing the number or diameter of the turns in the helix. While propagating along the tube, the EM wave interacts with the electron beam. Since the electromagnetic wave effectively propagates slower than the electron beam, the electrons “bunch up” and modulate the input signal, giving up energy in the process - an effect known as velocity modulation. Thus, the traveling wave progressively grows in amplitude as it propagates down the tube towards the collector.

Figure 2 features a photo of a high voltage power supply and TWT.

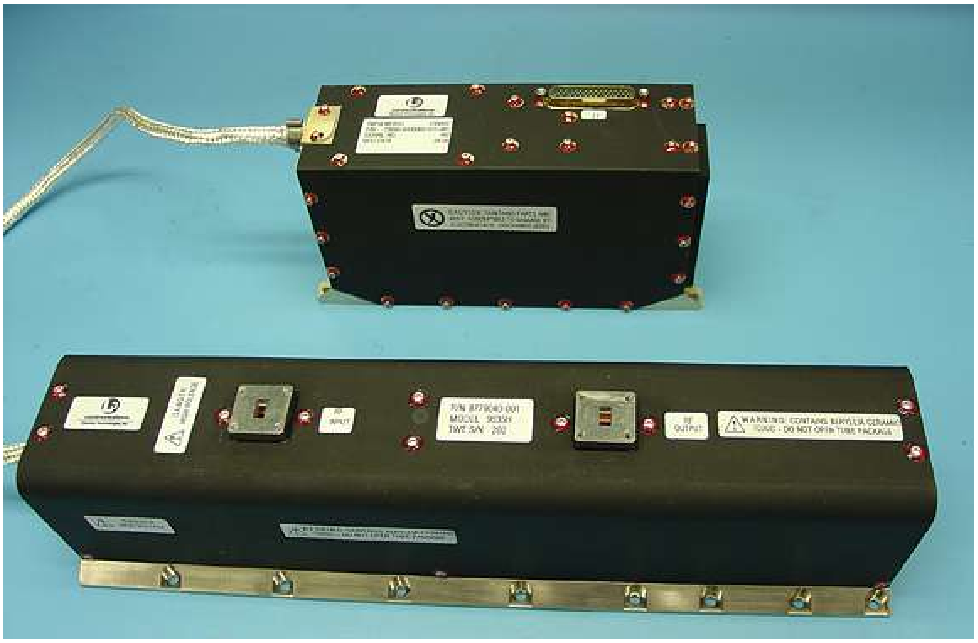

Figure 2. TWTA Including High Voltage Power Supply (top) and TWT (bottom).

Historically, TWTAs have been used in satellite transponders, radars, and in electronic warfare and self-protection systems. Recently, with the advent of wideband, high power solid-state amplifier solutions, however, TWTAs are slowly being replaced due to the higher reliability of their solid state counterparts.

Microwave Power Modules (MPMs)

A Microwave Power Module (MPM) is a hybrid solution between solid-state and vacuum tube electronics, which aims to take advantage of the best features of both technologies. They feature a solid-state pre-amplifier, miniaturized TWTA, and a high-density power supply, all integrated into a unit much more compact and lightweight than the traditional TWTA. While MPMs generally don’t provide as much power as their larger TWTA counterparts, their lighter weight, compact form factor, and relaxed power supply requirements (often 28 VDC or 270 VDC) enable use in applications where a TWTA would not be possible.

Similar to TWTAs, MPMs are capable of providing very high gains (>50 dB), multiple octaves of bandwidth, high efficiency. Typical power levels range from tens to hundreds of watts, with ~1 kW capability for pulsed MPMs.

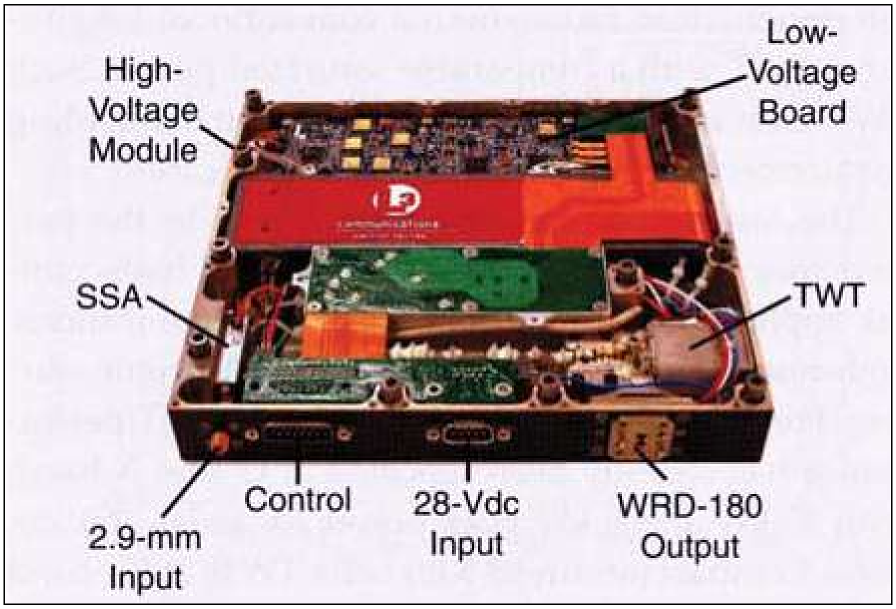
High-
Voltage
Module
Low-
Voltage
Board
SSA
TWT
2.9-mm
Input
Control
28-Vdc
Input
WRD-180
Output

Figure 3. 40W Ka-band MPM With Components Identified.
Image from *IEEE Magazine*, December 2009, page 42.

They have found applications in phased array antennas, lower-power radar transmitters, satellite communications, EW systems, and UAVs.

Solid State

Solid-state electronics, including amplifiers, are built entirely from solid materials and in which the electrons, or other charge carriers, are confined entirely within the solid material. The building material is most often a crystalline semiconductor. Solid-state power amplifiers, or SSPAs, are fabricated on many different semiconductor technologies, some of which include GaAs, GaN, InP, SiGe, CMOS. A photo of a GaN SSPA is presented in Figure 4.

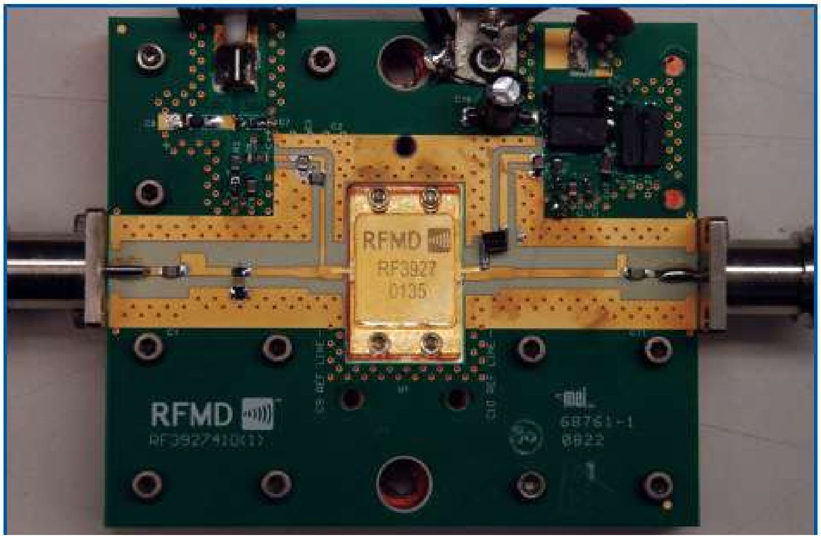
RFMD
RF3927
0135
RFMD
RF3927410(1)
68761-1
0822

Figure 4. Typical SSPA. Note the GaN die and external impedance matching networks.

In general, SSPAs are heralded as having a higher reliability than TWTAs. While high power SSPAs are available on the market, achieving high efficiency has been a challenge as many amplifier stages often need to be combined to meet to achieve power levels in the tens to hundreds of watts. Multi-stage designs increase size, weight, and power (SWAP) and decrease PAE (due to ohmic losses in the combining networks), which is one reason why TWTs and MPMs still provide excellent alternatives for medium to high power applications.

Below is a table that attempts to compare and contrast different solid-state technologies in the context of microwave networks and amplifiers. It is by no means complete, but provides a general overview.

Table 1. Comparison Between Different Semiconductor Process Technologies.

Technology	Advantages	Disadvantages
Silicon variants (CMOS, SiGe)	• Cheapest substrate due to CPU industry • Can be fabricated with nanometer accuracy	• Not a good overall microwave substrate • Results in lossy, high noise figure, low power components • Crystal is fragile • Junction temperatures limited to ~110C
Indium Phosphide Variants (InP/InGaP)	• High performance at low voltage • Good thermal stability • Results in high efficiency PAs, particularly at lower operating voltage • Extremely low noise figure • Useful through W-band and beyond	• Less mature, "niche" fabrication houses • Brittle, fragile material • Higher cost than GaAs • Low breakdown voltage – not good for high power
Gallium Arsenide variants (GaAs)	• Most mature, widespread technology • Many transistor variants (MESFET, PHEMT, MHEMT, HBT, etc) • High reliability • Fairly low cost (but more than silicon) • Great microwave substrate • Low loss, high ε_r. • 16-20V breakdown possible • Junction temperatures up to 150C	• High noise figure, Noise figure and power performance • Difficult to summarize - depends on transistor type used
Gallium Nitride (GaN)	• Up to 10x the power density of GaAs • High breakdown voltage (100V possible) • Results in high efficiency, high frequency, wide bandwidth PAs • Can operate hotter than GaAs, Si, or SiGe	• Currently more expensive than GaAs, but costs are decreasing • Difficult to fabricate • High power density leads to thermal challenges

SIGNAL GENERATION

Signal generators, also known variously as function generators, RF and microwave signal generators, pitch generators, arbitrary waveform generators, digital pattern generators or frequency generators are electronic devices that generate repeating or non-repeating electronic signals (in either the analog or digital domains). They are generally used in designing, testing, troubleshooting, and repairing electronic or electro-acoustic devices; though they often have artistic uses as well.

There are many different types of signal generators, with different purposes and applications (and at varying levels of expense); in general, no device is suitable for all possible applications.

Analog Signal Generators

RF signal generators are capable of producing CW (continuous wave) tones. The output frequency can usually be tuned anywhere in their frequency range. Many models offer various types of analog modulation, either as standard equipment or as an optional capability to the base unit. This could include AM, FM, ΦM (phase modulation) and pulse modulation. Another common feature is a built-in attenuator which makes it possible to vary the signal's output power. Depending on the manufacturer and model, output powers can range from -135 to +30 dBm. A wide range of output power is desirable, since different applications require different amounts of signal power. For example, if a signal has to travel through a very long cable out to an antenna, a high output signal may be needed to overcome the losses through the cable and still have sufficient power at the antenna. But when testing receiver sensitivity, a low signal level is required to see how the receiver behaves under low signal-to-noise conditions.

RF signal generators are required for servicing and setting up analog radio receivers, and are used for professional RF applications.

Arbitrary Waveform Generator

An arbitrary waveform generator (AWG) is a piece of electronic test equipment used to generate electrical waveforms. These waveforms can be either repetitive or single-shot (once only) in which case some kind of triggering source is required (internal or external). The resulting waveforms can be injected into a device under test and analyzed as they progress through the device, confirming the proper operation of the device or pinpointing a fault in the device.

Unlike function generators, AWGs can generate any arbitrarily defined wave shape as their output. The waveform is usually defined as a series of "waypoints" (specific voltage targets occurring at specific times along the waveform) and the AWG can either jump to those levels or use any of several methods to interpolate between those levels.

Because AWGs synthesize the waveforms using (baseband) digital signal processing techniques, their maximum frequency is usually limited to no more than a few gigahertz (~10 GHz being the latest state-of-the-art in 2012).

A major difficulty in generating non-repetitive waveforms at higher frequencies with AWGs is the large amount of data required to describe high-frequency baseband signals. For example, a 20 gigasample/sec arbitrary waveform with 8 bits of resolution requires 20 GB of data to represent every 1 second of signal, regardless of that signal's nature. Generating such large digital data streams and delivering them to the DAC in the AWG is an increasingly difficult problem as DAC upper frequencies continue to grow.

AWGs, like most signal generators, may also contain an attenuator, various means of modulating the output waveform, and often contain the ability to automatically and repetitively "sweep" the frequency of the output waveform (by means of a voltage-controlled oscillator) between two operator-determined limits. This capability makes it very easy to evaluate the frequency response of a given electronic circuit.

AWGs can operate as conventional function generators. These would include standard waveforms such as sine, square, ramp, triangle, noise, and pulse. Some units include additional built-in waveforms such as exponential rise and fall times, sinx/x, cardiac. Some AWGs allow users to retrieve waveforms from a number of digital and mixed-signal oscilloscopes. Some AWGs may display a graph of the waveform on their screen - a graph mode. Some AWGs have the ability to generate a pattern of words from multiple bit output connector to simulate data transmission, combining the properties of both AWGs and digital pattern generators.

One feature of direct digital synthesizer (DDS) based arbitrary waveform generators is that their digital nature allows multiple channels to be operated with precisely controlled phase offsets or ratio-related frequencies. This allows the generation of polyphase sine waves, I-Q constellations, or simulation of signals from geared mechanical systems such as jet engines. Complex channel-channel modulations are also possible.

Vector Signal Generators

Modern vector signal generators can be seen as a hybrid of the arbitrary waveform generator and the analog signal generator, combining a lower-bandwidth AWG with analog upconversion, allowing a moderate-bandwidth (~100 MHz, circa 2012) digital signal to be upconverted to any frequency (~50 GHz typically upper frequency typically).

With the advent of digital communications systems, it is no longer possible to adequately test these systems with traditional analog signal generators. This has led to the development of vector signal generators, also known as digital signal generators. These signal generators are capable of generating digitally-modulated radio signals that may use any of a large number of digital modulation formats such as QAM, QPSK, FSK, BPSK, and OFDM. In addition, since modern commercial digital communication systems are almost all based on well-defined industry standards, many vector signal generators can generate signals based on these standards. Examples include GSM, W-CDMA (UMTS), CDMA2000, LTE, Wi-Fi (IEEE 802.11), and WiMAX (IEEE 802.16). In contrast, military communication systems such as JTRS, which place a great deal of importance on robustness and information security, typically use very proprietary methods. To test these types of communication systems, users will often create their own custom waveforms and download them into the vector signal generator to create the desired test signal.

This page intentionally left blank.

DIGITAL SIGNAL PROCESSING COMPONENTS

The goal of DSP is usually to measure, filter and/or compress continuous real-world analog signals. The first step is usually to convert the signal from an analog to a digital form, by sampling and then digitizing it using an analog-to-digital converter (ADC), which turns the analog signal into a stream of numbers. However, often, the required output signal is another analog output signal, which requires a digital-to-analog converter (DAC). Even if this process is more complex than analog processing and has a discrete value range, the application of computational power to digital signal processing allows for many advantages over analog processing in many applications, such as error detection and correction in transmission as well as data compression. With the increasing bandwidth and dynamic range of ADC and digital components, common analog RF signal processing operations such as filtering, threshold detection, and pulse compression are being carried out in the digital domain. Additionally new capabilities such as synthetic aperture radar (SAR), digital RF memory (DRFM), and space-time adaptive processing (STAP) are wholly enabled by the increasing power of digital processing components.

Analog-to-Digital Converter

An analog-to-digital converter (abbreviated ADC, A/D or A to D) is a device that converts a continuous quantity to a discrete time digital representation. An ADC may also provide an isolated measurement. The reverse operation is performed by a digital-to-analog converter (DAC).

Typically, an ADC is an electronic device that converts an input analog voltage or current to a digital number proportional to the magnitude of the voltage or current. However, some non-electronic or only partially electronic devices, such as rotary encoders, can also be considered ADCs.

Resolution

The resolution of the converter indicates the number of discrete values it can produce over the range of analog values. The values are usually stored electronically in binary form, so the resolution is usually expressed in bits. In consequence, the number of discrete values available, or "levels," is a power of two. For example, an ADC with a resolution of 8 bits can encode an analog input to one in 256 different levels, since $2^8 = 256$. The values can represent the ranges from 0 to 255 (i.e. unsigned integer) or from −128 to 127 (i.e., signed integer), depending on the application.

Resolution can also be defined electrically, and expressed in volts. The minimum change in voltage required to guarantee a change in the output code level is called the least significant bit (LSB) voltage. The resolution Q of the ADC is equal to the LSB voltage. The voltage resolution of an ADC is equal to its overall voltage measurement range divided by the number of discrete values:

$$Q = \frac{E_{FSR}}{2^M}$$

where M is the ADC's resolution in bits and E_{FSR} is the full scale voltage range (also called 'span').

E_{FSR} is given by

$$E_{FSR} = V_{RefHi} - V_{RefLow}$$

where V_{RefHi} and V_{RefLow} are the upper and lower extremes, respectively, of the voltages that can be coded.

Normally, the number of voltage intervals is given by

$$N = 2^M - 1$$

where M is the ADC's resolution in bits.

That is, one voltage interval is assigned in between two consecutive code levels.

A typical 3 bit (2^8 = 8-level) coding scheme is depicted in Figure 1.

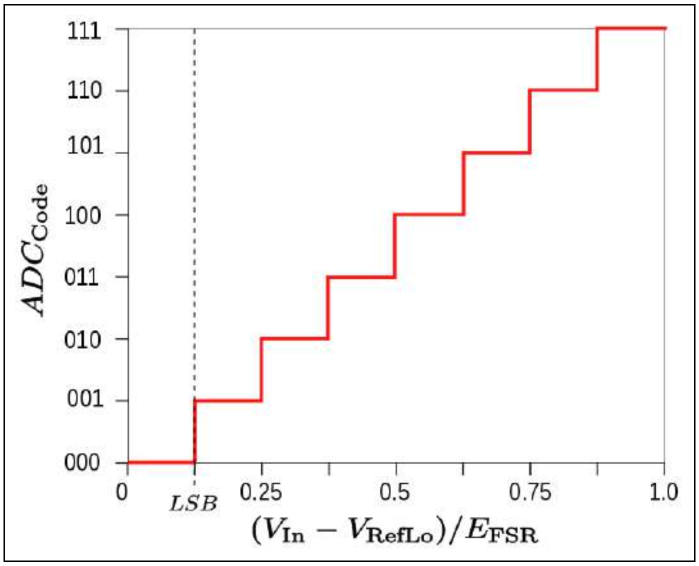

Figure 1. An 8-level (3 bit) ADC Coding Scheme.

Example:

Assume input signal x(t) = A cos(t), A = 5V
Full scale measurement range = -5 to 5 volts
ADC resolution is 8 bits: 2^8 - 1 = 256 - 1 = 255 quantization levels (codes)
ADC voltage resolution, Q = (5 V − [-5] V) / 255 = 10 V / 255 ≈ 0.039 V ≈ 39 mV.

In practice, the useful resolution of a converter is limited by the best signal-to-noise ratio (SNR) that can be achieved for a digitized signal. An ADC can resolve a signal to only a certain number of bits of resolution, called the effective number of bits (ENOB). One effective bit of resolution changes the signal-to-noise ratio of the digitized signal by 6 dB, if the resolution is limited by the ADC. If a preamplifier has been used prior to A/D conversion, the noise introduced by the amplifier can be an important contributing factor towards the overall SNR.

Accuracy

An ADC has several sources of errors. Quantization error and (assuming the ADC is intended to be linear) non-linearity are intrinsic to any analog-to-digital conversion. There is also a so-called aperture error which is due to a clock jitter and is revealed when digitizing a time-variant signal (not a constant value).

These errors are measured in a unit called the least significant bit (LSB). In the above example of an eight-bit ADC, an error of one LSB is 1/256 of the full signal range, or about 0.4%.

Sampling Rate

The analog signal is continuous in time and it is necessary to convert this to a flow of digital values. It is therefore required to define the rate at which new digital values are sampled from the analog signal. The rate of new values is called the sampling rate or sampling frequency of the converter.

A continuously varying band limited signal can be sampled (that is, the signal values at intervals of time T, the sampling time, are measured and stored) and then the original signal can be exactly reproduced from the discrete-time values by an interpolation formula. The accuracy is limited by quantization error. However, this faithful reproduction is only possible if the sampling rate is higher than twice the highest frequency of the signal. This is essentially what is embodied in the Shannon-Nyquist sampling theorem.

Oversampling

Usually, for economy, signals are sampled at the minimum rate required, with the result that the quantization noise introduced is white noise spread over the whole pass band of the converter. If a signal is sampled at a rate much higher than the Nyquist frequency and then digitally filtered to limit it to the signal bandwidth there are the following advantages:

- Digital filters can have better properties (sharper roll-off, phase) than analog filters, so a sharper anti-aliasing filter can be realized and then the signal can be down-sampled giving a better result
- A 20-bit ADC can be made to act as a 24-bit ADC with 256× oversampling
- The signal-to-noise ratio due to quantization noise will be higher than if the whole available band had been used. With this technique, it is possible to obtain an effective resolution larger than that provided by the converter alone.
- The improvement in SNR is 3 dB (equivalent to 0.5 bits) per octave of oversampling which is not sufficient for many applications. Therefore, oversampling is usually coupled with noise shaping. With noise shaping, the improvement is 6L+3 dB per octave where L is the order of loop filter used for noise shaping. e.g. - a 2nd order loop filter will provide an improvement of 15 dB/octave.

Digital to Analog Converter (DAC)

Digital-to-analog converters (DACs) perform the inverse function of the ADC. As such, the aforementioned principles of accuracy, resolution, sampling and oversampling, etc. apply equivalently.

Field-Programmable Gate Array (FPGA)

A field-programmable gate array (FPGA) is an integrated circuit designed to be configured by the customer or designer after manufacturing—hence "field-programmable." The FPGA configuration is generally specified using a hardware description language (HDL), similar to that used for an application-specific integrated circuit (ASIC). Circuit diagrams were previously used to specify the configuration, as they were for ASICs, but this is increasingly rare. FPGAs can be used to implement any logical function that an ASIC could perform. The ability to update the functionality after shipping, partial re-configuration of a portion of the design and the low non-recurring engineering costs relative to an ASIC design (notwithstanding the generally higher unit cost), offer advantages for many applications.

FPGAs contain programmable logic components called "logic blocks," and a hierarchy of reconfigurable interconnects that allow the blocks to be "wired together"—somewhat like many (changeable) logic gates that can be inter-wired in (many) different configurations. Logic blocks can be configured to perform complex combinational functions, or merely simple logic gates like AND and XOR. In most FPGAs, the logic blocks also include memory elements, which may be simple flip-flops or more complete blocks of memory.

FPGAs are becoming increasingly more popular for implementing RF signal processing functions such as signal compression (matched filtering), channel selection, modulation / demodulation, etc.

MICROWAVE MEASUREMENTS

Measurement Procedures

Calculate your estimated power losses before attempting to perform a measurement. The ideal input to a measurement device is in the 0 to 10 dBm (1 to 10 mW) range.

Linearity Check

To verify that a spectrum measurement is accurate and signals are not due to mixing inside the receiver, a linearity check should be performed, i.e., externally insert a 10 dB attenuator - if measurements are in the linear region of the receiver, all measurements will decrease by 10 dB. If the measurements decrease by less than 10 dB, the receiver is saturated. If the measurements disappear, you are at the noise floor.

Half-Power or 3 dB Measurement Point

To verify the half power point of a pulse width measurement on an oscilloscope, externally insert a 3 dB attenuator in the measurement line, and the level that the peak power decreases to is the 3 dB measurement point (Note: you cannot just divide the peak voltage by one-half on the vertical scale of the oscilloscope).

VSWR Effect on Measurement

Try to measure VSWR (or reflection coefficient) at the antenna terminals. Measuring VSWR of an antenna through its transmission line can result in errors. Transmission lines should be measured for insertion loss not VSWR.

High Power Pulsed Transmitter Measurements

When making power measurements on a high power pulsed transmitter using a typical 40 dB directional coupler, an additional attenuator may be required in the power meter takeoff line, or the power sensor may be burnt out.

For example, assume we have a 1 megawatt transmitter, with PRF = 430 pps, and PW = 13 μs. Further assume we use a 40 dB directional coupler to tap off for the power measurements. The power at the tap would be:

$$10 \log(P_p) - 10 \log(DC) - \text{Coupler reduction} =$$

$$10 \log(10^9 \text{mW}) - 10 \log(13\text{x}10^{-6})(430) - 40 \text{ dB} =$$

$$90 \text{ dBm} - 22.5 \text{ dB} - 40 \text{ dB} = 27.5 \text{ dBm (too high for a power meter)}$$

Adding a 20 dB static attenuator to the power meter input would give us a value of 7.5 dBm or 5.6 mW, a good level for the power meter.

High Power Measurements With Small Devices

When testing in the presence of a high power radar, it is normally necessary to measure the actual field intensity. The technique shown in Figure 4, in Section 6-7, may not be practical if the measurement device must be small. An alternate approach is the use of a rectangular waveguide below its cutoff frequency. In this manner, the “antenna” waveguide provides sufficient attenuation to the frequency being measured so it can be coupled directly to the measurement device or further attenuated by a low power attenuator. The attenuation of the waveguide must be accurately measured since attenuation varies significantly with frequency.

ELECTRO-OPTICS AND IR

ELECTRO-OPTICS AND IR 7-1

- Introduction 7-1.1
- Optical Spectrum 7-1.1
- Radiometric Quantities and Terminology 7-1.2
- Photometric Quantities 7-1.3
- Basic Principles 7-1.4
- Basic Radiant Power Relationships 7-1.5
- Infrared Source Characteristics 7-1.10
- Atmospheric Transmission 7-1.14
- EO Components and Sensors 7-1.16
- IR Threats to Aircraft and Their Countermeasures 7-1.30
- IRCM 7-1.36
- Lasers 7-1.38
- Fiber Optics 7-1.41

Laser Safety 7-2

This page intentionally left blank.

ELECTRO-OPTICAL SYSTEMS AND EW COUNTERMEASURES

INTRODUCTION

The development of infrared countermeasures and the evaluation of their performance against threat missiles is a broad, complex technical field. Much of the detailed information about the threats to be countered and the characteristics of the countermeasures themselves is understandably sensitive, beyond the limitations of this document. More detailed information can be requested, provided the requesting organization has the necessary clearance and need to know. Such requests will be considered on a case-by-case basis. Information will be provided only upon written concurrence of the controlling Government organization.

There are many electro-optical (EO) electronic warfare (EW) systems, which are analogous to radio frequency (RF) EW systems. These EO EW systems operate in the optical portion of the electromagnetic spectrum. Electro-optics (EO), as the name implies, is a combination of electronics and optics. By one definition EO is the science and technology of the generation, modulation, detection and measurement, or display of optical radiation by electrical means. Most infrared (IR) sensors, for example, are EO systems. In the popularly used term "EO/IR," the EO is typically used to mean visible or laser systems. The use of EO in this context is a misnomer. Actually, almost all "EO/IR" systems are EO systems as defined above. Another often-used misnomer is referring to an EO spectrum. EO systems operate in the optical spectrum, which is from 0.01 to 1000 micrometers. EO systems include, but are not limited to, lasers, photometry, infrared, and other types of visible, and UV imaging systems.

Within the broad field of Electronic Warfare, electro-optical systems are prevalent for communication systems and offensive and defensive applications. Lasers have been used extensively for weapons guidance purposes, warhead fuzing applications, targeting systems and other offensive weapons related purposes. Understanding electro-optical and radiometric principles and sensors is critical to the development of vehicle survivability systems. These principles range from signature reduction and camouflage to active countermeasure systems such as lamp-based and laser jammers to passive threat warning systems and expendable flare decoys.

Although military systems operate in many portions of the electro-optical spectrum, the infrared is of paramount importance for remote detection systems and weapons applications. Missile seekers, Forward Looking Infrared (FLIR) systems, and Infrared Search and Track Systems all operate in the infrared portion of the spectrum.

OPTICAL SPECTRUM

The optical spectrum is that portion of the electromagnetic spectrum from the extreme ultraviolet (UV) through the visible to the extreme IR. Figure 1 shows the optical spectrum in detail. The end points of the optical spectrum are somewhat arbitrary.

IR spectrum terminology has also varied through the years, with near (near visible), or short-wavelength infrared (SWIR) being on the high frequency end. Then as frequency decreases the spectrum is followed by intermediate or mid-wavelength infrared (MWIR), then far or long-wavelength IR (LWIR), and finally extreme IR.

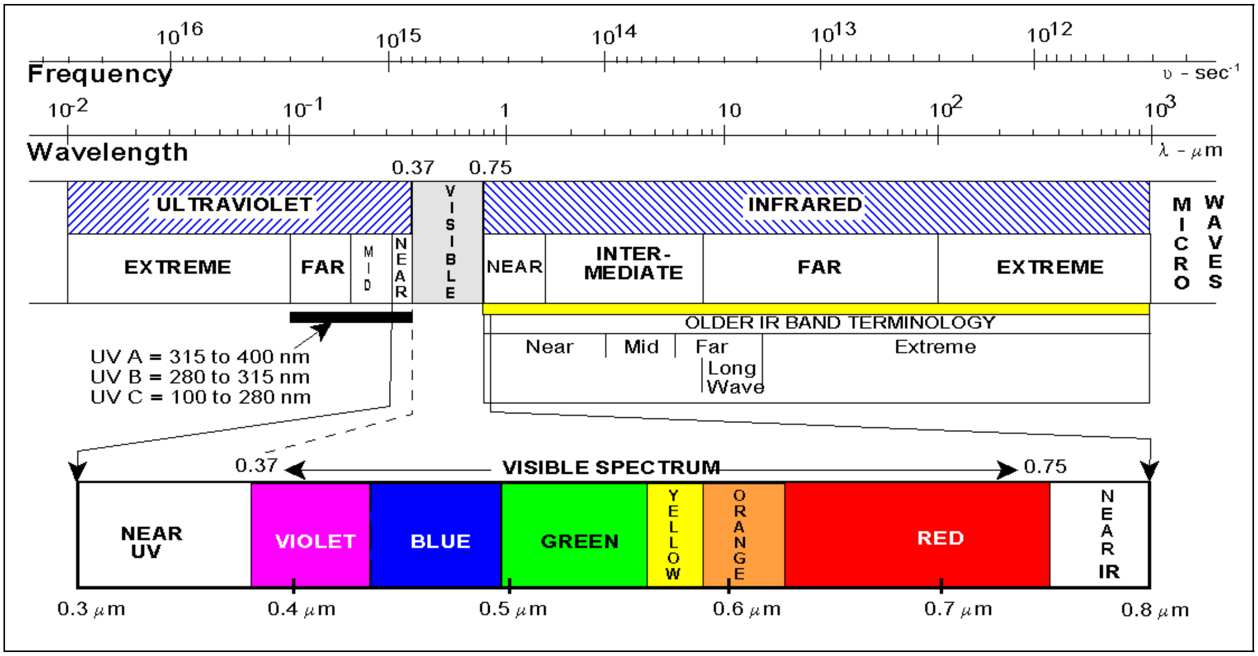

Frequency
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
υ - sec-1
Wavelength
10^{-2}
10^{-1}
1
10
10^{2}
10^{3}
λ - μm
0.37
0.75
ULTRAVIOLET
EXTREME
FAR
MID
NEAR
VISIBLE
INFRARED
NEAR
INTER-MEDIATE
FAR
EXTREME
MICRO WAVES
OLDER IR BAND TERMINOLOGY
Near
Mid
Far
Long Wave
Extreme
UV A = 315 to 400 nm
UV B = 280 to 315 nm
UV C = 100 to 280 nm
0.37
VISIBLE SPECTRUM
0.75
NEAR UV
VIOLET
BLUE
GREEN
YELLOW
ORANGE
RED
NEAR IR
0.3 μm
0.4 μm
0.5 μm
0.6 μm
0.7 μm
0.8 μm

Figure 1. Optical Spectrum.

RADIOMETRIC QUANTITIES AND TERMINOLOGY

The common terms used to describe optical radiation are the source parameters of power, radiant emittance (older term) or radiant exitance (newer term), radiance, and radiant intensity. They refer to how much radiation is given off by a body. The parameter measured by the detector (or collecting object/surface) is the irradiance. Any of these quantities can be expressed per unit wavelength in which case the subscript is changed from e (meaning energy derived units) to λ and the term is then called "Spectral ...X...", i.e. I_e is radiant intensity, while I_λ is spectral radiant intensity. These quantities in terms of currently preferred "Système International d'Unités" (SI units) are defined in Table 1.

Table 1. Radiometric SI Units.

Symbol	Name	Description	Units
Q	Radiant Energy		J (joules)
Φ_e	Radiant Power (or flux)	Rate of transfer of radiant energy	W (watts)
M_e	Radiant Exitance	Radiant power per unit area emitted from a surface	W m^{-2}
L_e	Radiance	Radiant power per unit solid angle per unit projected area	W $m^{-2}sr^{-1}$
I_e	Radiant Intensity	Radiant power per unit solid angle from a point source	W sr^{-1}
E_e	Irradiance	Radiant power per unit area incident upon a surface	W m^{-2}
X_λ	Spectral ...X..	(Quantity) per unit wavelength interval	(Units) nm^{-1} or μm^{-1}

Where X_λ is generalized for each unit on a per wavelength basis; for example, L_λ would be called "spectral radiance" instead of radiance.

In common usage, irradiance is expressed in units of Watts per square centimeter and wavelengths are in **µ**m instead of nanometers (nm). Other radiometric definitions are shown in Table 2.

Table 2. Other Radiometric Definitions.

Symbol	Name	Description	Units
α	Absorptance[1]	α = (*) absorbed / (*) incident	numeric
ρ	Reflectance	ρ = (*) reflected / (*) incident	numeric
τ	Transmittance	τ = (*) transmitted / (*) incident	numeric
$\in$	Emissivity	$\in$ = (*) of specimen / (*) of blackbody @ same temperature	numeric

Where (*) represents the appropriate quantity Q, Φ, M, E, or L

Note (1) Radiant absorptance should not be confused with absorption coefficient.

PHOTOMETRIC QUANTITIES

Whereas the radiometric quantities Φ_e, M_e, I_e, L_e, and E_e have meaning throughout the entire electromagnetic spectrum, their photometric counterparts Φ_v, M_v, I_v, L_v, and E_v are meaningful only in the visible spectrum (0.38 **µ**m thru 0.78 **µ**m).

The "standard candle" was redefined as the new candle or candela (cd). One candela is the luminous intensity of 1/60th of 1 cm^2 of the projected area of a blackbody radiator operating at the temperature of the solidification of platinum (2045 ºK). By definition, the candela emits one lumen (lm) per steradian.

Table 3 displays the photometric quantities and units. These are used in dealing with optical systems such as aircraft television camera systems, optical trackers, or video recording.

Table 3. Photometric SI Units.

Symbol	Name	Description	Units
Q_v	Luminous energy		lumen sec (lm s)
F or Φ_v	Luminous flux or Luminous Power	Rate of transfer of luminant energy	lumen (lm)
M_v	Luminous Exitance or flux density (formerly luminous emittance)	Luminant power per unit area emitted from a surface	lm m^{-2}
L_v	Luminance (formerly brightness)	Luminous flux per unit solid angle per unit projected area	nit (nt) or candela / m^2 or lm/sr·m^2
I_v	Luminous Intensity (formerly candlepower)	Luminous power per unit solid angle from a point source	candela (cd) or lm/sr
E_v	Illuminance (formerly illumination)	Luminous power per unit area incident upon a surface	lux or lx or lm/m^2
K	Luminous efficacy	$K = \Phi_v / \Phi_e$	lm / w

Table 4 displays conversion factors for commonly used illuminance quantities.

Table 4. Illuminance Conversion Units.

		Lux (lx)	Footcandle (fc)	Phot (ph)
1 lux (lm m^{-2})	=	1	0.0929	1 x 10^{-4}
1 footcandle (lm ft^{-2})	=	10.764	1	0.001076
1 phot (lm cm^{-2})	=	1 x 10^{4}	929	1

THE BASIC PRINCIPLES

The processes of absorption, reflection (including scattering), and transmission account for all incident radiation in any particular situation, and the total must add up to one in order that energy be conserved:

$\alpha + \rho + \tau = 1$, as shown in Figure 2.

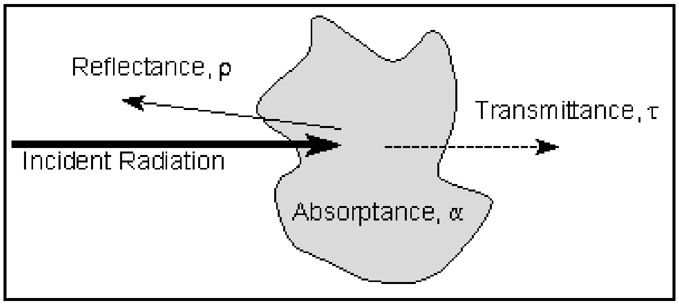

Figure 2. Radiation Incident on a Body.

If a material is opaque (no transmission), then: absorption + reflection = 1

In addition to the above processes, optical (including IR) radiation interacts with matter the same as radiation in any other part of the spectrum, including:

1. **Diffraction** – around edges
2. **Emission** – from matter by conversion from another form of energy
3. **Interference** – constructive and destructive
4. **Refraction** – bends when passing between two media with different propagation speeds (Snell's Law)
5. **Scattering** – when interacts with particles whose size approaches length of the wave
6. **Polarized** - electric field is partially polarized by reflection from dielectric

STERADIAN – SOLID ANGLE

Of significance to many terms and units in radiometric calculations is the *solid angle*. Figure 3 is a pictorial depicting the relationship of area, distance, and solid angle. By definition, the ratio of area on the surface of a sphere to the square of distance (the radius) is the unit less parameter *solid angle*, or steradian in the SI system of units. Solid angle is usually abbreviated as "sr" or given the Greek letter, Ω. The steradian is a dimensionless quantity in radiometric calculations.

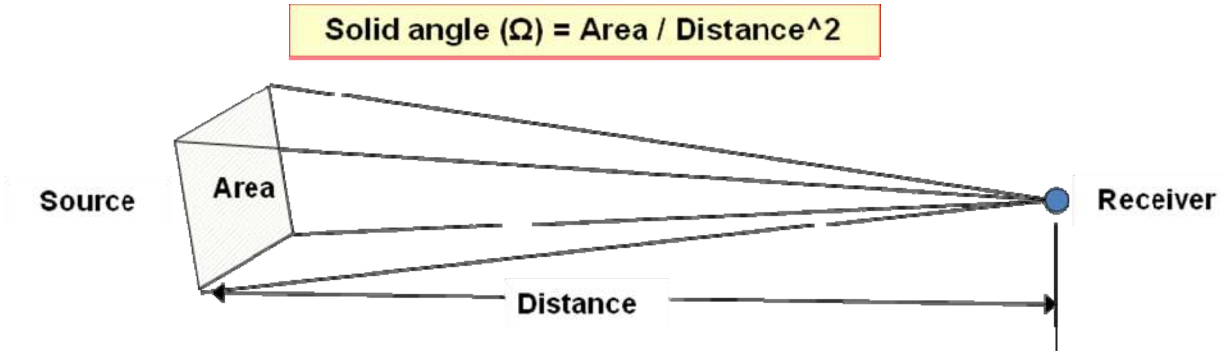

Figure 3. Solid Angle.

A sphere contains a solid angle of 4π steradians; a hemisphere contains 2π steradians, and so on. The area is a curved surface, but in most applications, the solid angles are sufficiently small that the area can be approximated as a plane. Also, for small angles, the solid angle in steradians is approximately equal to the product of two plane angles in radians.

CONVERSIONS

IR wavelengths are typically expressed in **µ**m, visible wavelengths in **µ**m or nm, and UV wavelengths in nm or angstroms. Table 5 lists conversion factors for converting from one unit of wavelength to another. The conversion is from column to row. For example, to convert from **µ**m to nm, multiply the value expressed in **µ**m by 10^3. IR wavelengths are also sometimes expressed in a frequency-like unit called wavenumbers or inverse centimeters. A wavenumber value can be found by dividing 10,000 by the wavelength expressed in **µ**m. For example, 2.5 **µ**m converts to a wavenumber of 4000 or 4000 inverse centimeters (cm^{-1}).

Table 5. Wavelength Conversion Units.

From -> To get ↓	Angstroms - Å	Nanometers - nm	Micrometers - **µ**m
	Multiply by		
Angstroms - Å	1	10	10^4
Nanometers - nm	10^{-1}	1	10^3
Micrometers - **µ**m	10^{-4}	10^{-3}	1

BASIC RADIANT POWER RELATIONSHIPS

Radiant intensity is the most commonly used term to describe the radiant power of a source per unit solid angle. Radiant Intensity offers the advantage of being a source term, like radiance, that is not related to the size of the radiating source. In practice, radiant intensity is a derived term and is not directly measurable. If the Instantaneous Field of View (IFOV), which represents the smallest optical resolution element of a remote sensor, subtends an angle smaller than the size of the radiating source, the sensor responds directly to *radiance*. If the IFOV of the sensor subtends an angle larger than the radiating source, the sensor responds to *irradiance*. The relationship among radiant intensity, radiance, and irradiance is shown in the following equation:

$$I = ED^2 = E\frac{A}{\Omega} = LA = L\Omega D^2$$

Where:

I = Intensity. Radiant intensity is the target source power per unit solid angle, in Watts/steradian.

E = Irradiance. Irradiance is the received power density in Watts/cm2 incident on a distant sensor. Irradiance is the quantity measured or detected by a distant sensor where the target is not spatially resolved (i.e., where the target subtends an angle smaller than the resolution or field of view of the instrument.)

L = Radiance. Radiance is the intensity per unit area of source, in W/cm^2/steradians. Radiance is the quantity measured by a camera or imaging system, where the target is optically resolved (i.e., where the target subtends an angle much greater than the resolution or instantaneous field of view of the imager.)

A = Area. Area is the cross-sectional or projected area of the target in cm^2.

D = Distance. Distance between the target and sensor in cm.

Ω = Solid angle subtended by the target in steradians. Solid angle appears directly or indirectly in many infrared quantities. The solid angle subtended by a source is the ratio of the source area to the square of distance.

All remote sensors receive energy from a source through an atmospheric path. The atmosphere attenuates the propagation of energy due to scattering, absorption, and molecular rotation and vibration, depending on the wavelengths involved. The atmosphere itself also radiates. All incident energy on a remote sensor, except as received under vacuum conditions, is detected through the atmosphere. The term *apparent* is applied to a radiometric quantity to acknowledge the presence of atmospheric effects. A remote sensor, responding the radiance of a source over an atmospheric path, responds to *apparent radiance*.

Real sensors do not respond uniformly to energy. Sensors have non-uniform spectral and spatial response due to many factors, including detector and optical characteristics. All energy received by the detector is spectrally weighted by the spectral response of the instrument. The term *effective* is applied to acknowledge the non-uniform response of the instrument.

Outside of a vacuum, all radiation incident on a remote sensor is attenuated by atmospheric effects such as scattering and absorption. Where the atmosphere absorbs, it also emits. Path radiance is the term applied to the contribution of the atmospheric path to the received radiation. The term *apparent* is applied to radiometric quantities to acknowledge the influence of the atmosphere on the received radiation.

DOMAINS, DISTRIBUTIONS, AND SENSOR RESPONSE

Radiant power that is emitted or reflected from a source is distributed across multiple dimensions or domains. The three domains that are most important to radiometric applications are spectral (Figure 4), spatial (Figure 5), and temporal (Figure 6). Knowing how the radiant power from a source is distributed across each domain is important to understanding performance of an EW system, whether it be a sensor, a weapon, or a countermeasure.

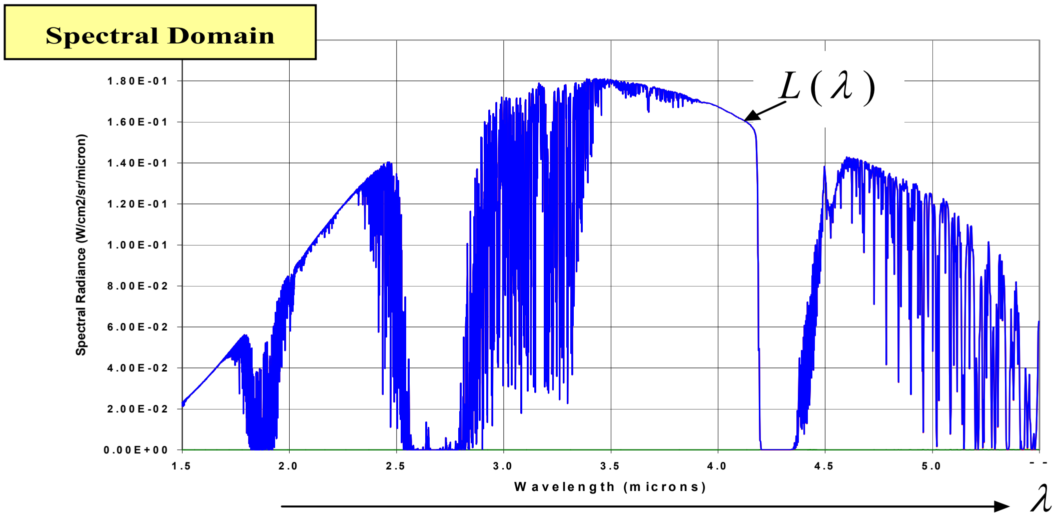

In the spectral domain, radiation is distributed as a function of wavelength (or wavenumber). Radiation from solid materials is distributed as a continuum in accordance with Planck's formula. The graph above shows the spectral distribution of radiation from a 600 degree Celsius blackbody after passage through the atmosphere. Radiation from gases appears at specific wavelength "lines" corresponding to molecular resonances.

Figure 4. Sample Spectral Distribution of a Solid Material.

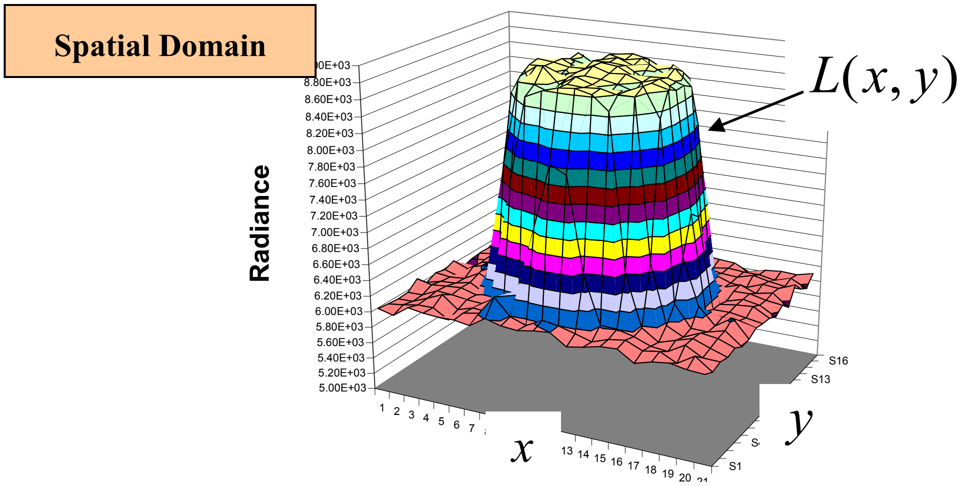

In the spatial domain, radiation is distributed as a function of angle, position, size, shape, or orientation. The graph above shows the radiance of a circular target as viewed by an imaging system.

Figure 5. Sample Spatial Distribution of a Circular Target Image.

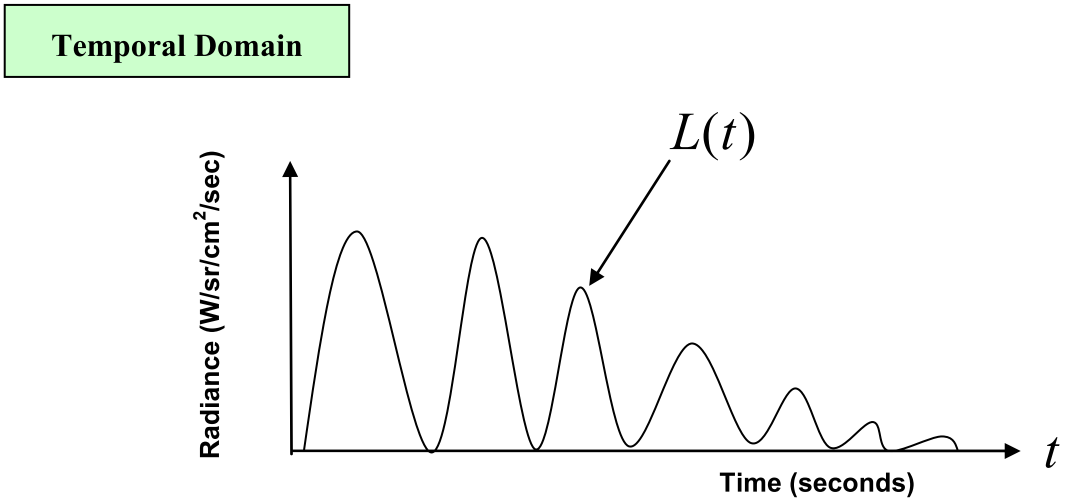

In the temporal domain, radiation is distributed as a function of time (or frequency). Most IR target sources, such as aircraft and ground vehicles, change slowly with time and can be considered steady state for measurement purposes. However, IR countermeasures systems and devices, such as jammers and decoy flares, have high frequency content that must be considered in their design and effectiveness assessment.

Figure 6. Sample Temporal Distribution of a Modulated Source.

All electro-optical sensors have some sensitivity to how radiation is distributed across the different domains. This sensitivity or response takes the following forms:

1. **Spectral response**:
    1. May be broad in the case of a radiometer or imager, or narrow in the case of a spectrometer where response takes the form of resolution.
    2. Spectral response is chosen to exploit particular spectral features of interest in a target.
    3. Non-uniformity of spectral response can have a significant effect on the performance of sensor.

2. **Spatial response**:
    1. May be broad in the case of a radiometer or spectrometer, where spatial response takes the form of sensitivity across the sensor's FOV. In the case of an imager, the spatial response may be narrow because spatial response takes the form of the resolution or IFOV.
    2. Required spatial resolution is defined by the size and radiant power of the source, the distance between the source and the target, and by the sensitivity of the sensor.
    3. Non-uniformity of spatial response can have a significant effect on the performance of sensor.
    4. Response to radiation from the background within the FOV is primarily a spatial response issue and can be significant to sensor performance.

3. **Temporal response**:
    1. Takes the form of bandwidth or, in the case of a digital system, of sampling frequency.
    2. Unlike spectral and spatial response, temporal response includes the response time of the detector and all of the electronics, data transfer time, and processing time for a sensor.
    3. Non-uniformity across a region is usually not a concern. Temporal response is usually uniform for target frequencies that fall well below the low-pass or Nyquist breakpoints.

Sensors respond to received radiant power by integrating the power distribution in each domain under the corresponding response curve. Sensors respond in multiple domains, so output is proportional to a multi-dimensional, weighted integral. Figure 7 illustrates response-weighting, or convolution, of radiation with instrument response in the spectral domain. The concept applies equally in the spatial domain.

Response Weighting in the Spectral Domain

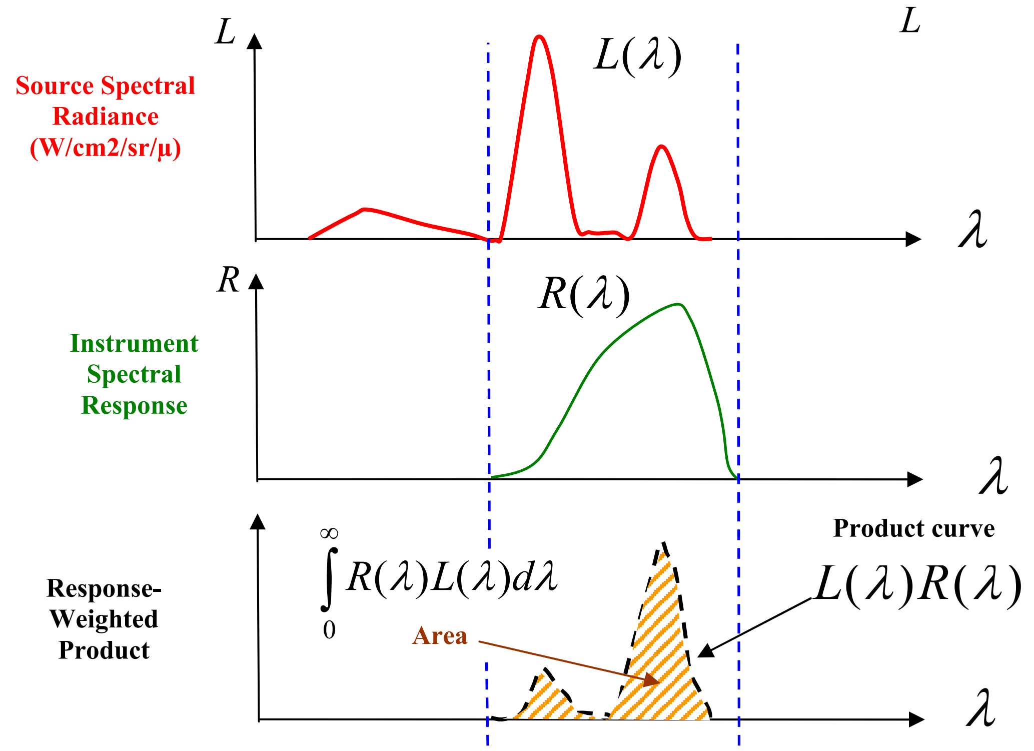
L
Source Spectral Radiance (W/cm2/sr/µ)
L(λ)
L
λ
R
R(λ)
Instrument Spectral Response
λ
Response-Weighted Product
$\int_0^\infty R(\lambda)L(\lambda)d\lambda$
Product curve
L(λ)R(λ)
Area
λ

Non-uniform sensor response across a domain distorts the distribution of received radiant power by weighting the radiation at some wavelengths differently than others. Instrument output is proportional to the integral of this weighted product, which makes instrument response shape an embedded part of both the measurement and the calibration. Some consequences of non-uniform response are:

- Two different sources with identical power, but different distributions will produce different measurement values.
- If two different sources (such as the calibration and the target) have similar relative power distributions, the relative weighting by the instrument will be the same for both sources, so the shape and degree of non-uniformity has little or no effect. This will be discussed in more detail later as a key strategy in calibration to reduce measurement uncertainty.

Figure 7. Response-Weighting in the Spectral Domain.

The consequence of instrument output being proportional to multiple weighted integrals is that the characteristics of the instrument affect the measured value in ways that cannot be easily extracted or corrected. Understanding non-uniformity in sensor response is critical to understanding performance and evaluating the effectiveness of the system.

INFRARED SOURCE CHARACTERISTICS

Reflectance

Reflectance is generally categorized as either *specular* (mirror-like) or *diffuse* (scattered by reflection from a rough surface). Most surfaces exhibit both types of reflection, but one typically dominates. Reflections from smooth surfaces are specular. Reflectance is a unitless quantity between zero and one that is the ratio of reflected power to the incident power. Reflectance varies with angle from the normal and with wavelength.

Diffuse reflectors scatter incident radiation broadly. Johann Heinrich Lambert described an ideal diffuse reflector in which the intensity of reflected rays is distributed as a cosine of the angle from the normal, regardless of the angles of the arriving incident rays. Such a surface is described as *Lambertian*. The same relationship applies to a diffuse source. If the source is perfectly diffuse, the radiance is independent of the viewing angle because the projected area of the source also varies as a cosine function of the angle from normal.

While no surface is perfectly diffuse, standard paint on most aircraft, for example, is near enough for at least a first approximation. Significant effort has gone into measuring and describing surface properties of military vehicles to support modeling and simulation activities. Bidirectional Reflectance Distribution Function (BRDF) is defined as the ratio of the reflected radiance to the incident irradiance at a wavelength, λ.

The BRDF provides a complete description of the reflectance properties of a surface. Measurements of BRDF are made in specially equipped optical laboratories. For aircraft, a typical use of BRDF data is in computer models that predict the reflections of the earth and sun from the aircraft fuselage and wings.

Emissivity

Emissivity is a unitless quantity that is a measure of the efficiency of a surface as an absorber or emitter. Emissivity is expressed as a number between 0.0 and 1.0. According to Kirchoff's Law, for an opaque object in thermal equilibrium, i.e., no net heat transfer, the emissivity equals the absorptance. In other words, a perfect emitter is also a perfect absorber. This ideal emitter is known as a *blackbody*. A surface with an emissivity of 1.0 emits the maximum radiation that Planck's Law allows. A body for which the emissivity is constant with wavelength but less than 1.0 is commonly known as a *graybody*. Planck's equation is typically expressed for an ideal blackbody emitter, but multiplying the blackbody expression by the emissivity term expresses the spectral distribution of power for a graybody.

Electromagnetic (EM) radiation, including infrared radiation, can be characterized by amplitude, frequency (or wavelength), coherency, and polarization properties. Amplitude refers to the magnitude of the electromagnetic wave. Coherency refers to the degree in which the electromagnetic waves maintain a constant phase difference, both spatially and temporally. Polarization is a description of the orientation of the wave's propagation perpendicular to the direction of travel. Like all electromagnetic radiation, infrared radiation travels as an EM field with the propagation velocity of the speed of light in a vacuum, and slower through air and other dielectrics. , Infrared radiation also exhibits both wave and particle properties. Which property is used depends on the application: Waves are used for applications involving propagation and geometrical optics, while particles (photons) are used for most detector applications. The energy of a photon is inversely proportional to its wavelength (hc/λ) where h is the Planck's constant, and c is the speed of light.

Infrared radiation also interacts with matter in a variety of ways:

- Reflects – wave is reflected from a surface.
- Refracts – direction of wave bends when passing between two transparent media with different propagation speeds.
- Scatters – when interacts with particles whose size approaches the wavelength of the radiation.
- Diffracts – around edges of an obstruction.
- Interferes – constructively and destructively.
- Absorbs – when absorbed by matter, radiation is (usually) converted into heat energy.
- Emits – radiation is emitted from matter by conversion from another form of energy.
- Transmitted – propagates through a transparent medium (or vacuum).

Plank's Law:

Infrared is directly related to the heat radiated from matter. Anything with a temperature above absolute zero (-273.15 degrees Celsius) radiates in the infrared. Planck's Law, discovered by German Physicist Max Planck, mathematically describes the distribution of radiant power across the spectrum for a given temperature. The form presented below calculates source radiance in Watts/sr/cm^2/ μ:

$$L_\lambda = \frac{C_1}{\pi\lambda^5\left[e^{\left(\frac{C_2}{\lambda T}\right)} - 1\right]}$$

Where: $C_1 = 2\pi c^2 h = 3.7415 \times 10E{+}04$ W cm^{-2} μ^4

$C_2 = ch/k = 1.4389E{+}04$ μ K

c = speed of light; h = Plank's constant; k = Boltzmann's constant

With λ in μ and T in Kelvin (= ºC + 273.15)

The behavior of Planck's curve with temperature is fundamental to every infrared detection scenario. Figure 8 shows Planck's equation for a single temperature. The shape of this curve resembles a wave of water, with a steep rise in power on the short wavelength side of the peak and a tail off on the long wavelength side.

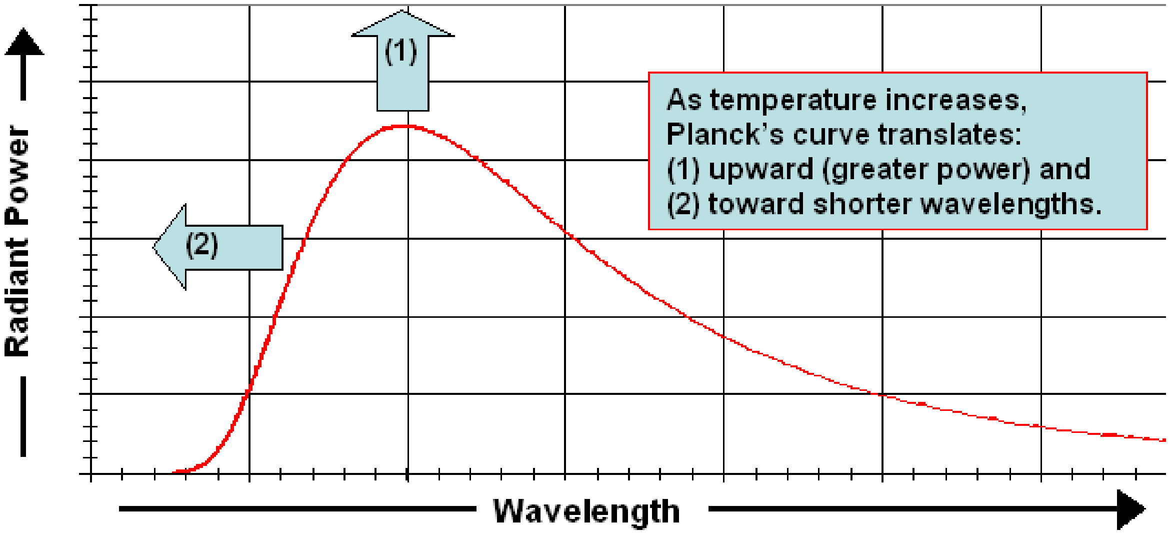

Figure 8. Temperature Effects and Power Distribution.

When temperature increases, two significant changes in power distribution occur and are governed by the following two relationships:

Stefan-Boltzmann Law:

The total power under the curve increases proportionally to the fourth power of the temperature. According to the Stefan-Boltzmann law, the total radiant emittance (power) of a blackbody is:

$$M = \sigma T^4 \qquad \text{Where:} \quad \sigma = \frac{2\pi^5 k^4}{15 c^2 h^3} = 5.67 \times 10^{-12}\ \text{Watts cm}^{-2}\ {}^{\circ}K^{-4}$$

and T is in Kelvin

This is Plank's radiation law integrated over all values of λ.

Wein's Displacement Law

The peak emission translates toward shorter wavelengths.

Wein's displacement law takes the derivative of the Plank's law equation to find the wavelength for maximum spectral exitance (emittance) at any given temperature:

$$\lambda_{max} = 2897.8 / T$$ where T = temperature is in Kelvin

The surface of the sun radiates with a spectral distribution like that of a 5900 Kelvin source. In accordance with Wein's Law, the wavelength at which the radiation for a 5900 Kelvin source peaks is approximately 490 nanometers. The maximum sensitivity of the day-adapted human eye (photopic) occurs at about 555 nanometers, which happens to be in a highly transparent region of the atmosphere. In other words, the temperature of our sun and the transmission of our atmosphere are conveniently matched to the response of the human eye. Figure 9 shows the relationship between the response of the human eye and the spectral distribution of solar irradiance.

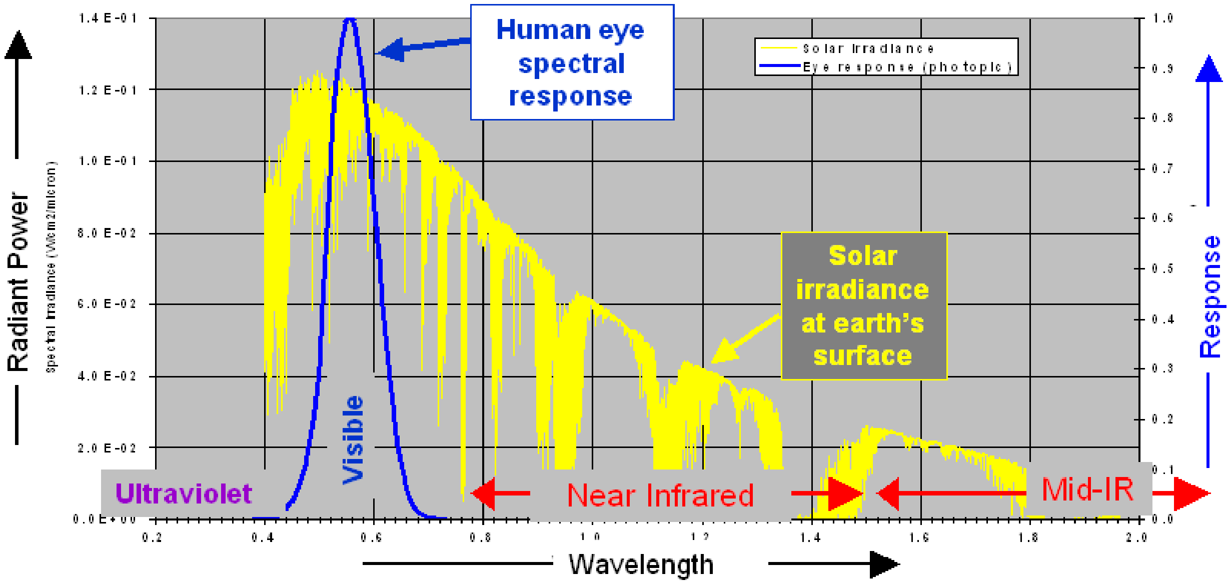

Figure 9. Power Distribution of the Sun Related to the Human Eye.

Unlike other electromagnetic radiation, the peak radiation from objects at the temperature of the earth, and of humans, vehicles, aircraft, etc., lies at infrared wavelengths. This makes the infrared portion of the spectrum militarily important.

For detection of an object, whether it is with the human eye or with an electro-optical sensor, it is not where the maximum emissions occur in the spectrum that is important. It is where the maximum difference between the target and background lies that is most significant.

The *absolute* signature of a source or target is the power radiated and reflected by the target without influence from background radiation. Absolute signature can be measured with an instrument such as an imaging camera at close range, producing highly resolved IR imagery. The difference between radiation from a target and its background is its *contrast*. Contrast is the target signature relative to a specific background. Contrast irradiance is the quantity detected by any remote sensor or missile. Contrast varies greatly with background conditions.

Figure 10 shows several different types of source distributions encountered by infrared sensors. Most targets of military interest are a complex combination of spectral contributions from a variety of sources. For example, the signature of an aircraft can typically be described by the contributions from airframe, hot engine parts, and plume.

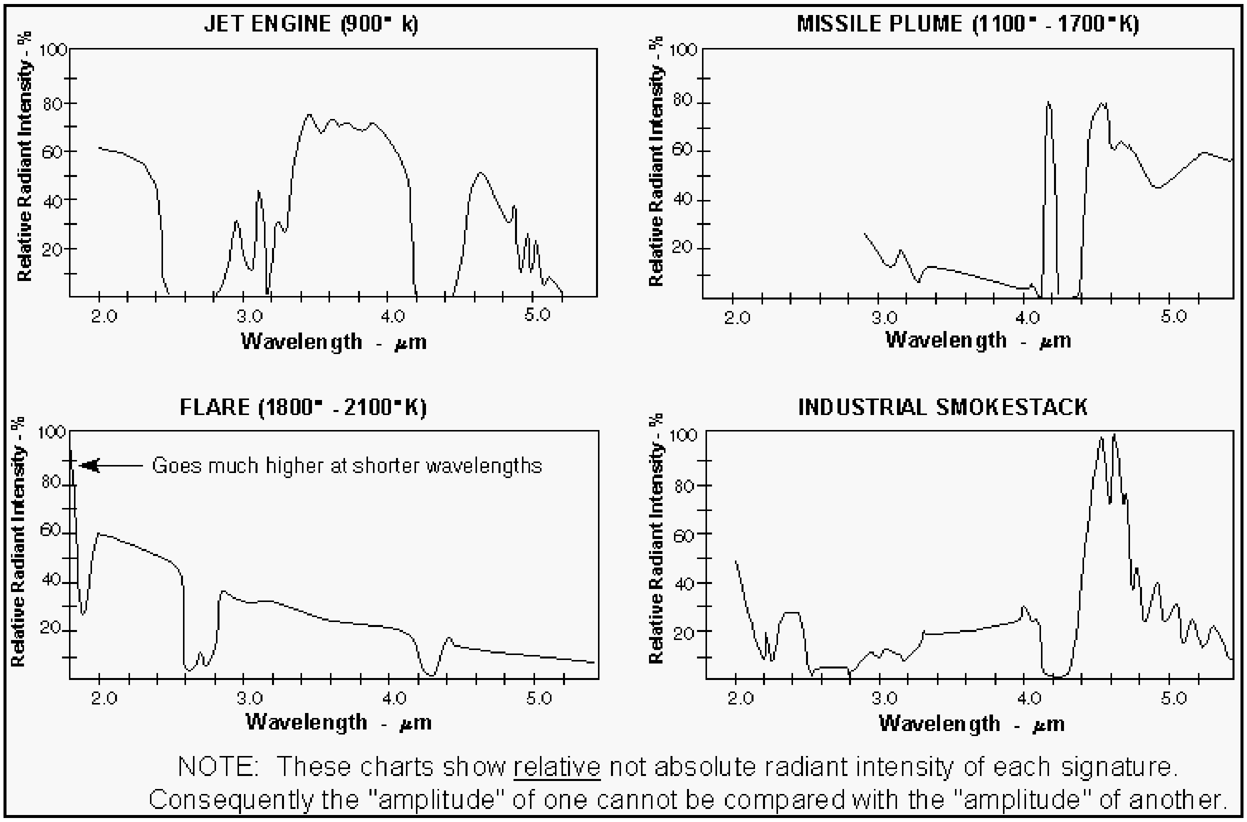

JET ENGINE (900° k)

MISSILE PLUME (1100° - 1700°K)

FLARE (1800° - 2100°K)

INDUSTRIAL SMOKESTACK

NOTE: These charts show relative not absolute radiant intensity of each signature. Consequently the "amplitude" of one cannot be compared with the "amplitude" of another.

Figure 10. Spectral Distribution of Various Targets.

ATMOSPHERIC TRANSMISSION

The radiation emitted or reflected from the targets and backgrounds must pass through the intervening atmosphere before reaching the detection system. The radiation is absorbed and re-emitted by molecular constituents of the atmosphere and scattered into and out of the path by various aerosol components.

Figure 11 reveals the presence of atmospheric windows, i.e. regions of reduced atmospheric attenuation. IR detection systems are designed to operate in these windows. Combinations of detectors and spectral bandpass filters are selected to define the operating region to conform to a window to maximize performance and minimize background contributions.

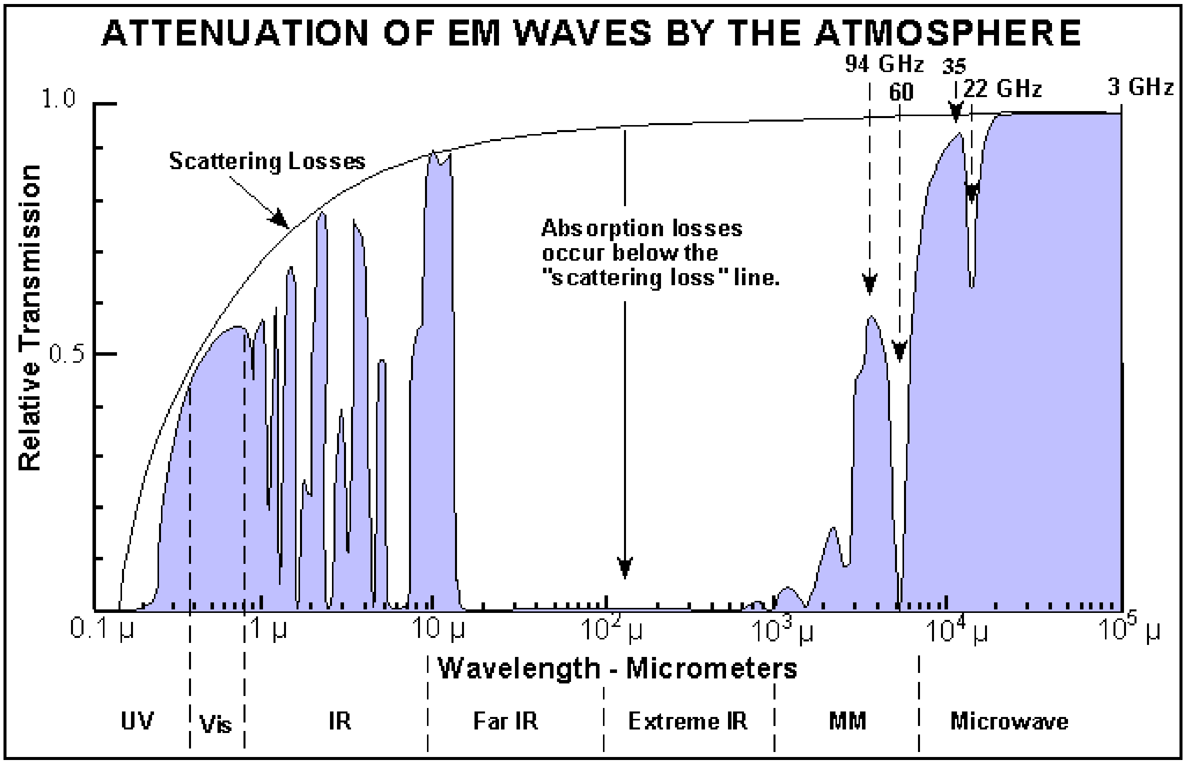

Figure 11. Atmospheric Transmission Over 1 NM Sea Level Path.

The molecules that account for most of the absorption in the IR region are water, carbon dioxide, nitrous oxide, ozone, carbon monoxide, and methane. Figure 12 shows an expanded view of the infrared portion of the spectrum.

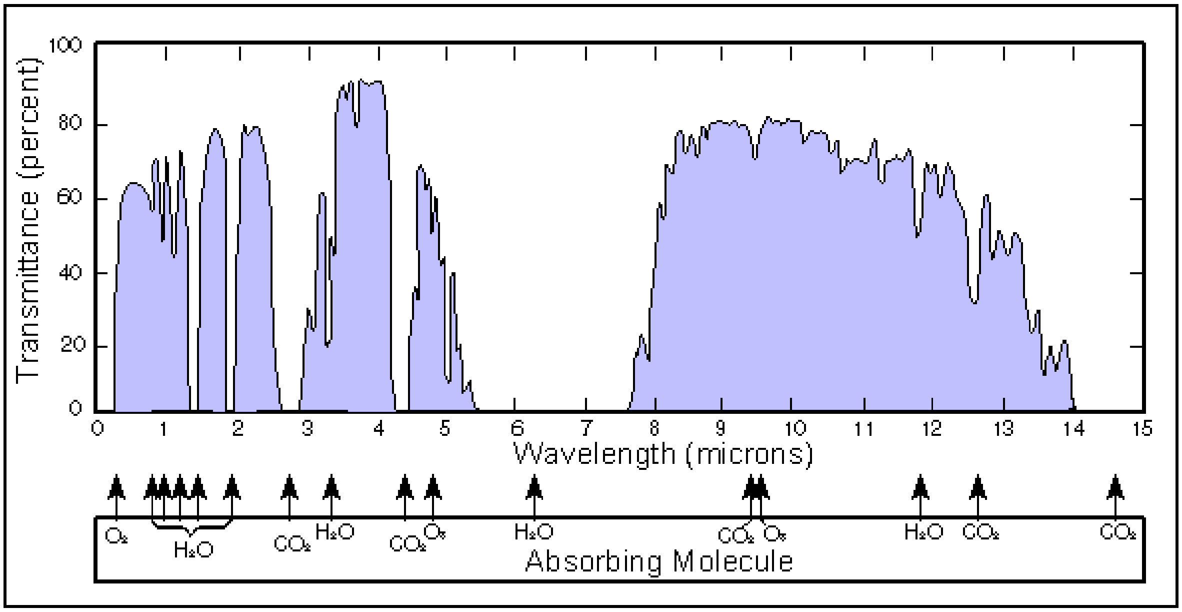

Figure 12. Transmittance of Atmosphere Over 1 NM Sea Level Path (Infrared Region).

The transmission in a window is greatly dependent on the length and characteristics of the path. As path length increases, absorption gets deeper and broader. Water vapor also has a significant effect overall on transmission through the atmosphere. High relative humidity attenuates transmission at all optical wavelengths.

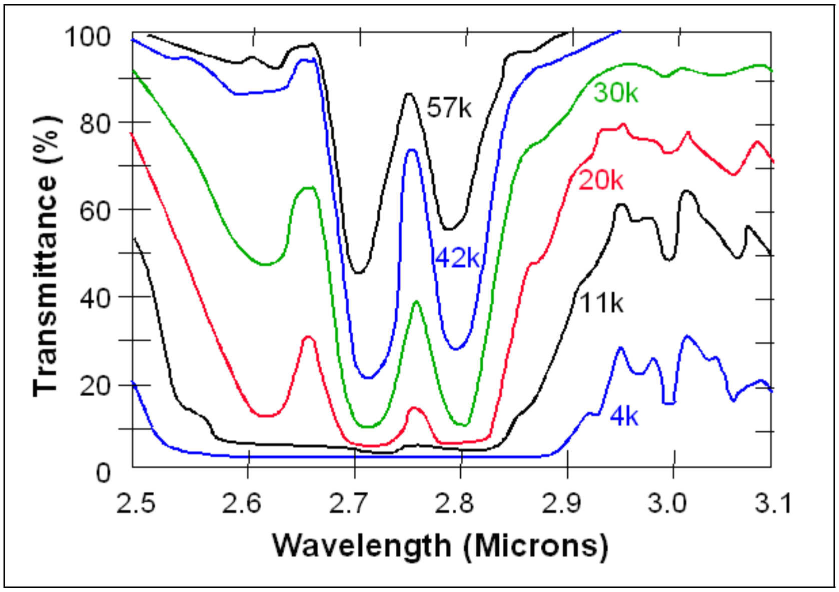

Figure 13. Transmittance at Various Altitudes.

Since water vapor generally decreases with altitude, transmission generally increases and path length becomes the determining factor. However, path length does not affect transmission of all wavelengths the same.

The altitude effects on transmittance are shown in Figure 13.

A great deal of work has gone into developing high fidelity atmospheric models. One of the most commonly used tools is MODTRAN™ (MODerate Resolution Atmospheric TRANsmission). MODTRAN™ models atmospheric propagation of EM radiation from 0.2 to 100 um. MODTRAN was developed by Spectral Sciences Inc. and the United States Air Force Research Laboratory.

ELECTRO-OPTICAL COMPONENTS AND SENSORS

Almost all IR instruments, missiles, search systems, etc. have similar functional components. Basic components typically include:

- Optics - Reflective or refractive lenses to:
  - Collect radiation. Irradiance (power density) is increased by collecting radiation over a large area and focusing down to a small area.
  - Form or focus an image that will be used to extract information about the target.

- Filter(s) - Spectral and spatial filters to distinguish target from background and to extract
  - Target information.
  - Spectral filters restrict sensitive wavelength range.
  - A spatial filter separates image information by features such as size or position.

- Detector - A transducer to convert received radiation to an electrical signal for processing.

- Electronics - Used to amplify and condition the detector signal and perform some action, such as controlling a servo for tracking and guidance or recording the received information. In addition to the above, the optical head may also contain a window to protect the electronics and an output unit consisting of indicators or displays.

Windows / Domes / Lens Materials

For most applications of EO systems in EW the detection system is protected from the environment by a window or dome of optically transmissive material. The window operates both as a weather seal and, in some cases, helps to define the spectral response region of the system. In some cases the window functions as a lens.

IR energy interacts with matter in ways we associate with light (reflection, refraction, and transmission). Lower energy of IR photons results in different optical properties than light. For example:

- Glass and water are not transparent to wavelengths longer than about 3 microns.
- Silicon and germanium are highly transparent to infrared radiation, but are opaque to visible light.

Transmission bands of representative window or lens materials are shown in Figure 14. The end points depicted are for the 10% transmission wavelengths. Not shown in Figure 14 are the various UV transmissive glasses such as Pyrex, Corex, and Vycor or Amorphous Material Transmitting IR radiation (AMTIR) which are various combinations of Ge-As-Se (AMTIR-1), As-Se (AMTIR-2, 4, & 5), Ge-Sb-Se (AMTIR-3), As-S (AMTIR-6).

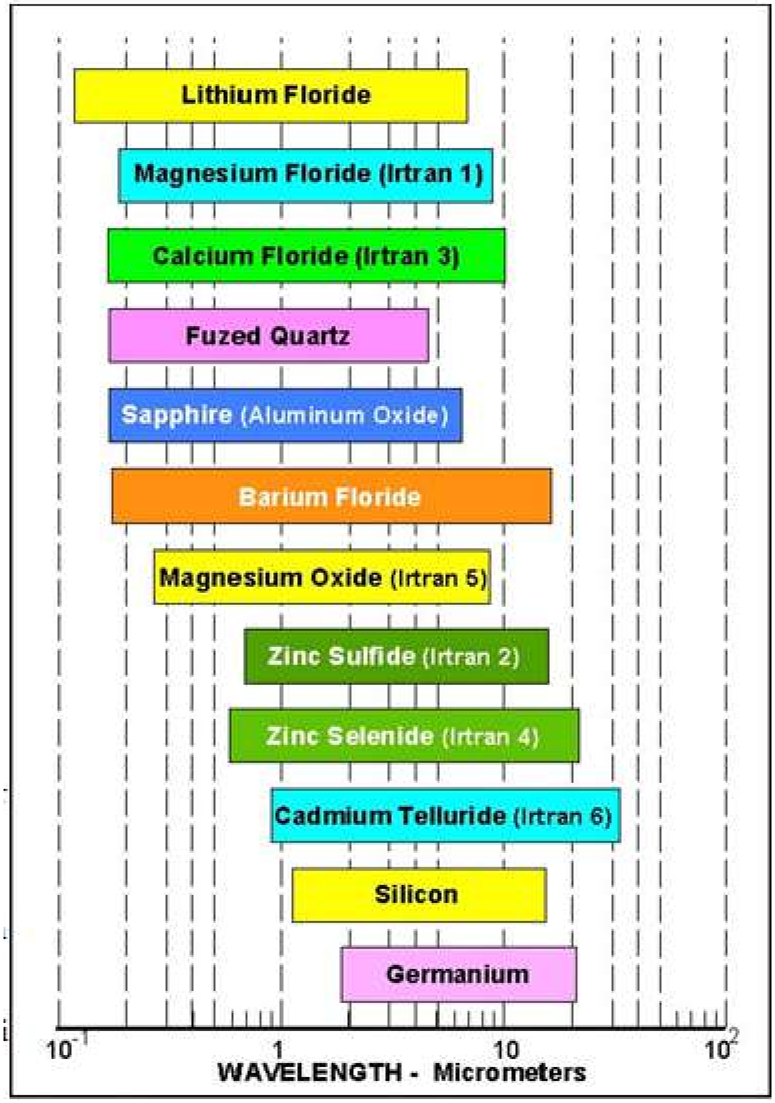

Figure 14. Transmission of Selected Window / Lens Materials.

Objective Lens

The objective lens is the first optical element in a lens in a typical sensor or missile seeker. The objective lens serves two main functions:

- Collect radiation (i.e., multiply irradiance (power density) by collecting over a large area and focusing onto a small area)
- Form an image of the target scene onto a filter and detector array.

Lens Types

Lenses for the IR can be either refractive or reflective. Refractive optics are “straight through” lenses, with the light never making large bends as shown in Figure 15.

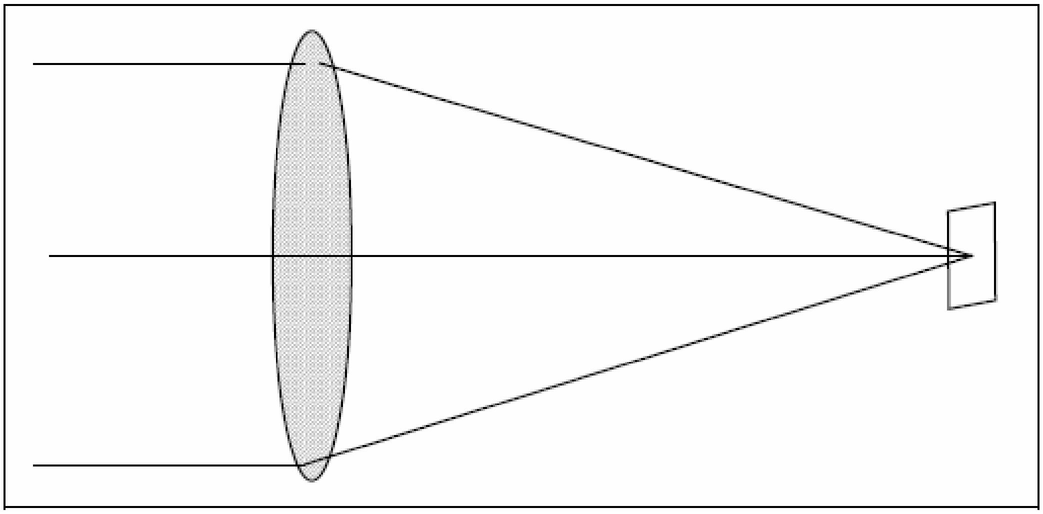

Figure 15. Refractive Lens.

A common reflective design used in many missiles is the Cassegrain. The Cassegrain design shown in Figure 16 is compact in size for its focal length.

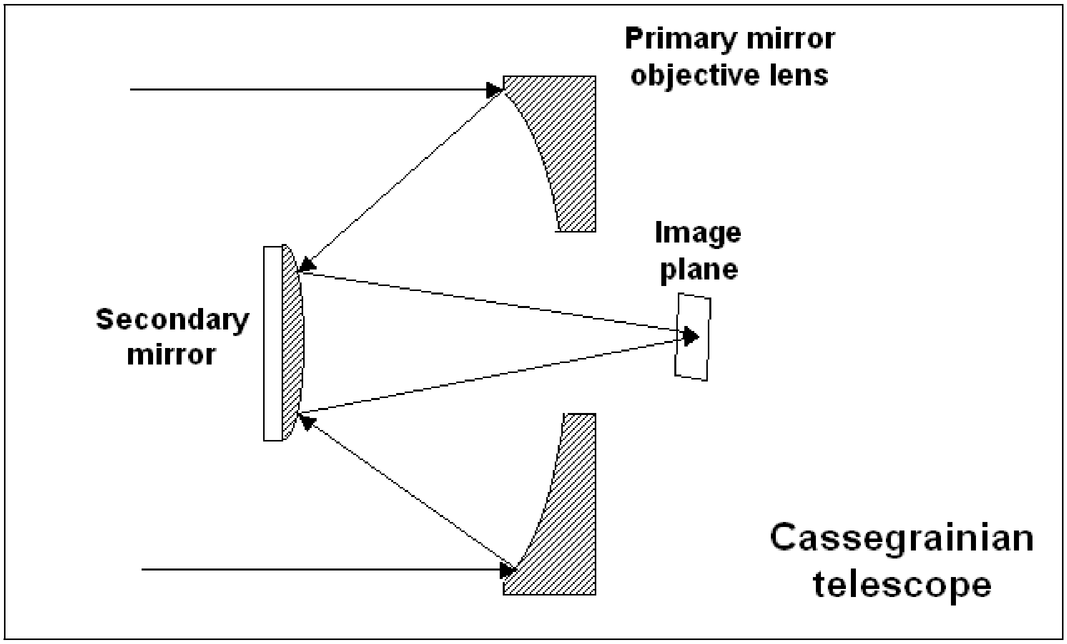
Primary mirror objective lens
Image plane
Secondary mirror
Cassegrainian telescope

Figure 16. Cassegrain Lens.

Optical Filters

Filters may be divided into two major categories: (1) spectral and (2) spatial.

Most optical radiation detectors have a wider sensitivity band than desired for the particular application. Spectral filters restrict sensitive wavelength range. Reasons for filtering include: enhancement of target-to-background contrast, avoidance of unwanted plume emissions and/or atmospheric absorption regions, and extraction and measurement of target spectral features.

To further define the system sensitivity, band interference filters or absorption filters are used. An absorption filter is a bulk material with a sharp cut-on or cut-off in its transmission characteristic. A cut-on and a cut-off filter can be combined to make a bandpass filter. By selecting absorption characteristics of absorption filters combined with the response of a detector, the desired system response can be obtained. An interference filter is composed of dielectric coatings on an appropriate substrate combined in such a way to produce cut-on, cut-off, or bandpass filters. Interference filters allow more control of the final response characteristics and smaller elements. See Figure 17.

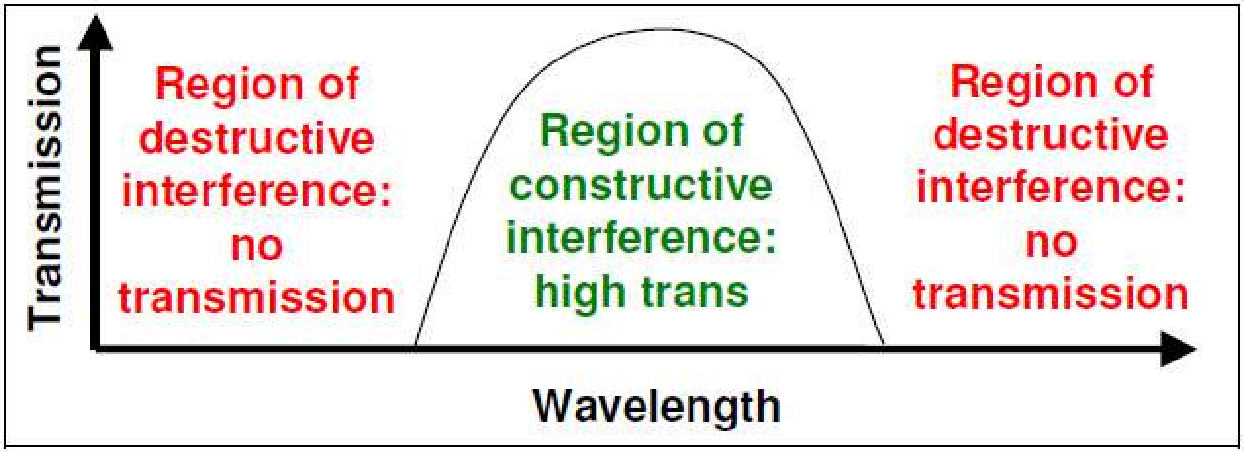
Transmission
Region of destructive interference: no transmission
Region of constructive interference: high trans
Region of destructive interference: no transmission
Wavelength

Figure 17. Spectral Bandpass Filter.

Most spectral filters are of the thin-film interference type. Layers of dielectric material are vacuum deposited on a substrate window material. Typical substrate materials in IR are sapphire, silicon, and germanium. Thickness of deposited layers designed to have constructive interference to pass desired radiation at desired wavelengths and destructive interference to block undesired wavelengths Besides spectral filters, EO system optics often have antireflection (or AR) coatings to eliminate or greatly reduce unwanted reflections between optical elements.

A spatial filter separates information in a scene image by features such as size or position. Spatial filters take a variety of forms. Some common types and their functions include:

- Field stop: limits an instrument's field of view. Blocks unwanted sources (such as sun) outside nominal field of view.
- Mechanical modulator or "chopper."
- Reticle: A mechanical modulator used in many missile designs. Usually discriminates against extended sources (such as background) in favor of "point" target sources and provides target directional information from modulation phase.

Detector Coolers

Many IR detectors have to be cooled for proper operation. Most systems use closed-cycle coolers or thermoelectric coolers. Thermoelectric coolers use the Peltier effect, which produces a reduced temperature by passing a d-c current through a thermoelectric junction. Multi-stage coolers can cool a detector down to below 200ºK. Closed-cycle coolers typically are of the Stirling cycle design and utilize the expansion of a gas (helium) to cool a cold finger attached to the detector. These generally operate at liquid nitrogen temperature (77ºK).

Detectors

A detector is a transducer that transforms electromagnetic radiation into a form, which can be more easily detected. In the detectors of interest to EW the electromagnetic radiation is converted into an electrical signal. In some systems the signal is processed entirely within the system to perform its function. In others the signal is converted to a form to allow the human eye to be used for the final detection and signal analysis.

Detectors are transducers than convert optical radiation into electrons. The physical effects by which electromagnetic radiation is converted to electrical energy are divided into two categories: photon effects and thermal effects. EW systems primarily use detectors dependent on photon effects. These effects can be divided into internal photo effects and external photo effects.

The external photo effect is known as photoemission. In the photoemissive effect, photons impinging on a photocathode drive electrons from its surface. These electrons may then be collected by an external electrode and the photocurrent thus obtained is a measure of the intensity of the received radiation.

Internal photoeffects of interest are the photoconductive effect and the photovoltaic effect. In the photoconductive effect, absorbed photons cause an increase in the conductivity of a semiconductor. The change is detected as a decrease in the resistance in an electrical circuit. In the photovoltaic effect, absorbed photons excite electrons to produce a small potential difference across a p-n junction in the semiconductor. The photovoltage thus produced may be amplified by suitable electronics and measured directly.

Thermal detectors respond directly to heat. Examples of these devices include bolometers, thermopiles, and pyroelectric detectors. The pyroelectric effect is an example of the thermal effect. The pyroelectric effect is a change in polarization in a crystal due to changes in temperature. Radiation falling on such a crystal is detected by observing the change in polarization as a build up of surface charge due to local heating. When coated with a good black absorber, the crystal will be sensitive to a wide band of wavelengths.

Figure 18 shows the spectral sensitivity range of typical detectors using these effects.

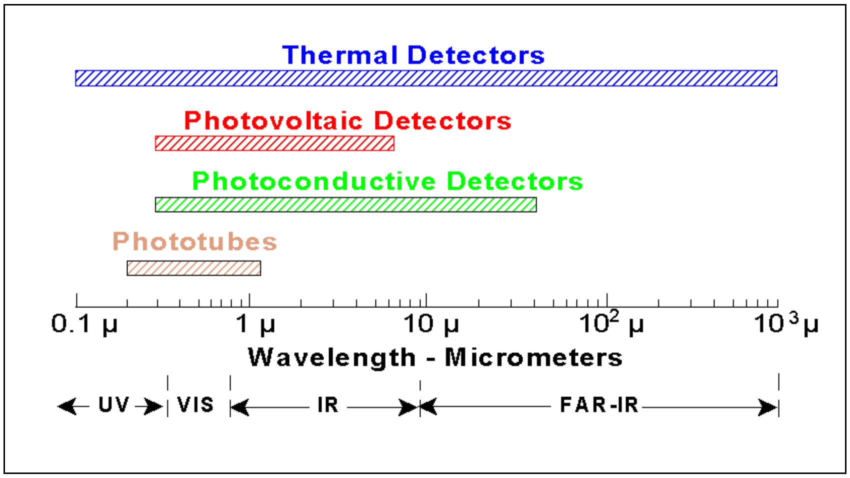

Figure 18. Spectral Range of Various Detectors.

Detector Types

Photoconductive detectors operate as resistors in a circuit. The resistance of the detector changes as the radiation incident on its surface changes. For EW applications, the most photoconductive detector types include: Indium Antimonide (InSb), which can also be operated in photovoltaic mode; Gallium Arsenide (GaAs); Lead Sulfide (PbS); and Lead Selenide (PbSe).

Photovoltaic detectors, the most common detectors used in modern EW and military sensor applications, produce a voltage that is proportional to the incident radiation. Common examples of photovoltaic detectors are Indium Antimonide (InSb) and Mercury Cadmium Telluride (HgCdTe). Both of these detector types offer high sensitivity when cryogenically cooled.

Diode phototubes and photomultipliers are commonly used detectors for UV systems including many operational missile-warning systems. These types of tubes offer the advantage of operating uncooled which can significantly reduce the complexity of a sensor system and offer increased reliability. Most of the modern IR sensors require cooled detectors. InSb, for example, requires cooling to 77 Kelvin to achieve the necessary sensitivity. Depending on the application, HgCdTe can be operated at somewhat higher temperature conditions.

A Photoelectromagnetic (PEM) detector has a junction that generates a current when exposed to light in a magnetic field.

Some detectors (such as InSb) have multiple modes of operation, including: Photoconductive (PC), Photovoltaic (PV), or Photoelectromagnetic (PEM) modes of operation.

Detector Parameters and Figures of Merit

The important parameters in evaluating a detector are the spectral response, time constant, the sensitivity, and the noise figure.

The spectral response determines the portion of the spectrum to which the detector is sensitive.

The time constant is a measure of the speed of response of the detector. It is also indicative of the ability of the detector to respond to modulated radiation. When the modulation frequency is equal to one over the time constant, the response has fallen to 70.7 % of the maximum value. The time constant is related to the lifetime of free carriers in photoconductive and photovoltaic detectors and to the thermal coefficient of thermal detectors. The time constant in photoemissive devices is proportional to the transit time of photoelectrons between the photocathode and anode.

The sensitivity of a detector is related to its responsivity. The responsivity is the ratio of the detected signal output to the radiant power input. For photoconductive and photovoltaic detectors the responsivity is usually measured in volts per watt -- more correctly, RMS volts per RMS watt. However, the sensitivity of a detector is limited by detector noise. Responsivity, by itself, is not a measure of sensitivity. Detector sensitivity is indicated by various figures of merit, which are analogous to the minimum detectable signal in radar. Such a quantity is the noise equivalent power (NEP). The NEP is a measure of the minimum power that can be detected. It is the incident power in unit bandwidth, which will produce a signal voltage equal to the noise voltage. That is, it is the power required to produce a signal-to-noise ratio of one when detector noise is referred to unit bandwidth. The units of NEP are usually given as watts, but more correctly, are watts/Hz$^{1/2}$ or watts·sec$^{1/2}$.

Another figure of merit is the noise equivalent irradiance (NEI). The NEI is defined as the radiant power per unit area of the detector required to produce a signal-to-noise ratio of one. The units of NEI are watts per square centimeter.

Noise equivalent power (NEP) is the radiant power required to produce a signal to noise ratio of one for a detector. Detectivity (D) of a detector is defined as the reciprocal of the NEP. The units of D are watts $^{-1}\cdot$sec $^{-1/2}$. A higher value of detectivity indicates an improvement in detection capability. Since D depends on detector area, an alternate figure of merit, known as D-star (D*). D* is the detectivity measured with a bandwidth of one hertz and reduced to a responsive area of one square centimeter. The units of D* are cm·watts $^{-1}\cdot$sec $^{-1/2}$. D* is the detectivity usually given in detector specification sheets. Typical spectral detectivity characteristics for various detectors are shown in Figure 19.

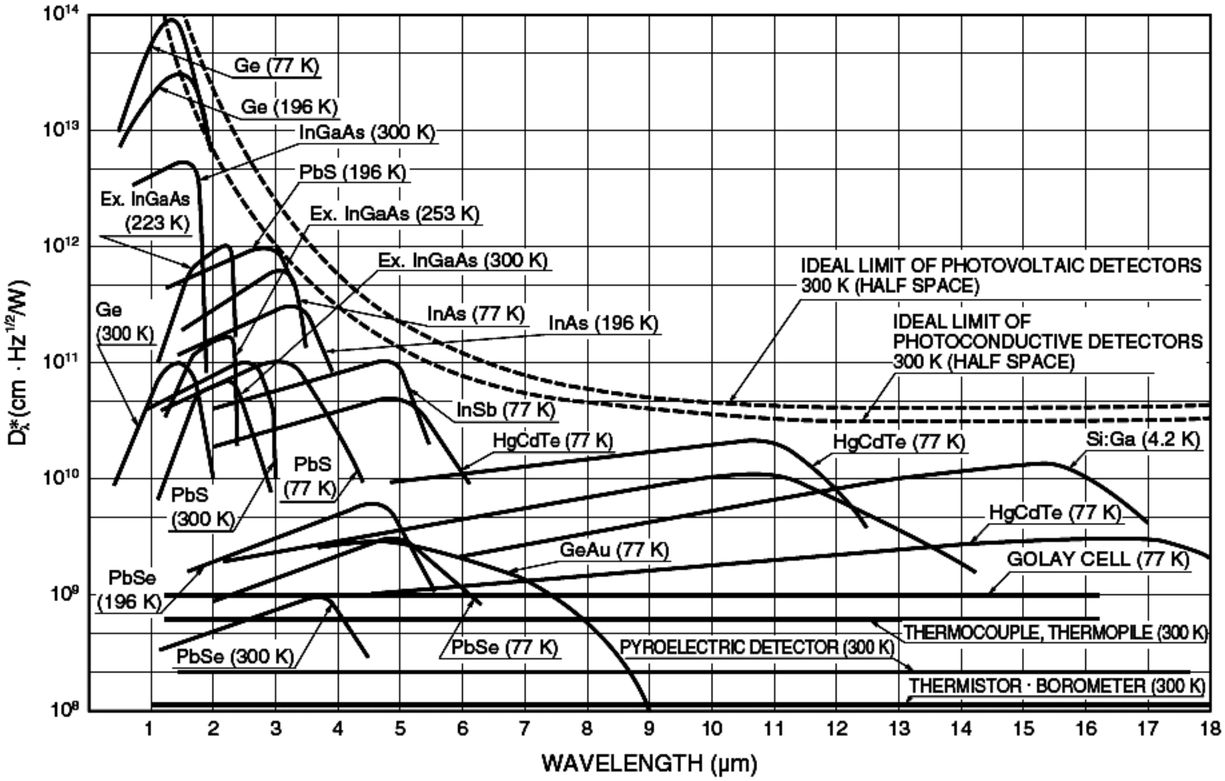

Figure 19. Spectral Detectivity of Various Detectors.

Besides the NEI mentioned above, the quantum efficiency of the photocathode is also a figure of merit for photoemissive devices. Quantum efficiency is expressed as a percent -- the ratio of the number of photoelectrons emitted per quantum of received energy expressed as a percent. A quantum efficiency of 100 percent means that one photoelectron is emitted for each incident photon.

There are other figures of merit for television cameras. The picture resolution is usually described as the ability to distinguish parallel black and white lines and is expressed as the number of line pairs per millimeter or TV lines per picture height. The number of pixels in the scene also defines the quality of an image. A pixel, or picture element, is a spatial resolution element and is the smallest distinguishable and resolvable area in an image. CCD cameras with 512 x 512 elements are common.

Another resolution quantity is the gray scale, which is the number of brightness levels between black and white a pixel can have.

Noise in Detectors

The performance of a detector is limited by noise. The noise is the random currents and voltages that compete with or obscure the signal or information content of the radiation. Five types of noise are most prominent in detectors: (1) thermal, (2) temperature, (3) shot, (4) generation-recombination, and (5) 1/f noise.

Thermal noise, also known as Johnson noise or Nyquist noise, is electrical noise due to random motions of charge carriers in a resistive material.

Temperature noise arises from radiative or conductive exchange between the detector and its surroundings, the noise being produced by fluctuations in the temperature of the surroundings. Temperature noise is prominent in thermal detectors.

Shot noise occurs due to the discreetness of the electronic charge. In a photoemissive detector shot noise is due to thermionic emission from the photocathode. Shot noise also occurs in photodiodes and is due to fluctuations in the current through the junction.

Generation-recombination noise is due to the random generation and recombination of charge carriers (holes and electrons) in semiconductors. When the fluctuations are caused by the random arrival of photons impinging upon the detector, it is called photon noise. When it is due to interactions with phonons (quantized lattice vibrations), it is called generation-recombination noise. Johnson noise is predominant at high frequencies, shot noise predominates at low frequencies, and generation-recombination and photon noise are predominant at intermediate frequencies.

As the name implies, 1/f noise has a power spectrum that is inversely proportional to frequency. It is dominant at very low frequencies. In photoemissive detectors it is called flicker noise and has been attributed to variation in the emission from patches of the photocathode surface due to variation in the work function of the surface. In semiconductors 1/f noise is also called modulation noise. Here it is apparently due to surface imperfections and ohmic contacts (which are a form of surface imperfection).

Infrared Spectral Region and Features of Interest

Different portions of the infrared spectrum are common for particular applications. The reasons for the selection of a specific window are often sensitive and beyond the scope of this document, but selections are typically based on several key considerations:

- Target characteristics such as size and spectral distribution of signature.
- Background radiance and clutter.
- Atmospheric effects (transmission, path radiance, scintillation, etc.).
- Distinguishing characteristics between natural and man-made sources.

Table 6 describes some of the types of characteristics that are prevalent in the short-wavelength (SWIR) infrared (0.7 to 3.0 microns), mid-wavelength (MWIR) infrared (3.0 to 6.0 microns) and long-wavelength (LWIR) infrared (7.0 to 14.0 microns) along with some types of systems that operate in these regions.

Table 6. Infrared Features, Regions, and Types of Systems.

Near or Short Wave IR (SWIR)	**Dominant natural source:**	**Sun**
	Atmospheric:	
	Transmission:	High
	Path radiance:	Scattered sunlight
	Dominant aircraft IR component:	Sunlit airframe
	Anti-aircraft threat:	Vehicle-launched SAM
Mid-wave IR (MWIR)	**Dominant natural source:**	**Sun**
	Atmospheric:	
	Transmission:	High transmission “windows” between H_2O and CO_2 absorption
	Path radiance:	Scattered sunlight below 3 microns Thermal at longer than 3 microns
	Dominant aircraft IR component:	Engine hot parts and plume
	Anti-aircraft threat:	All AAMs and SAMs
Far or Long-wave IR (LWIR)	**Dominant natural source:**	**Earth**
	Atmospheric:	
	Transmission:	High
	Path radiance:	Low: small thermal emission from ozone
	Dominant aircraft IR component:	Airframe direct emission and terrestrial illumination
	Anti-aircraft threat:	Airborne IRST. No anti-aircraft missiles

Sensors and Detection

Figure 20 shows a generalized detection problem. On the left of the diagram are the radiation sources - the sun, background, and the target of interest. In the middle is the intervening atmosphere, which attenuates the radiation as it travels to the detection system shown on the right of the diagram.

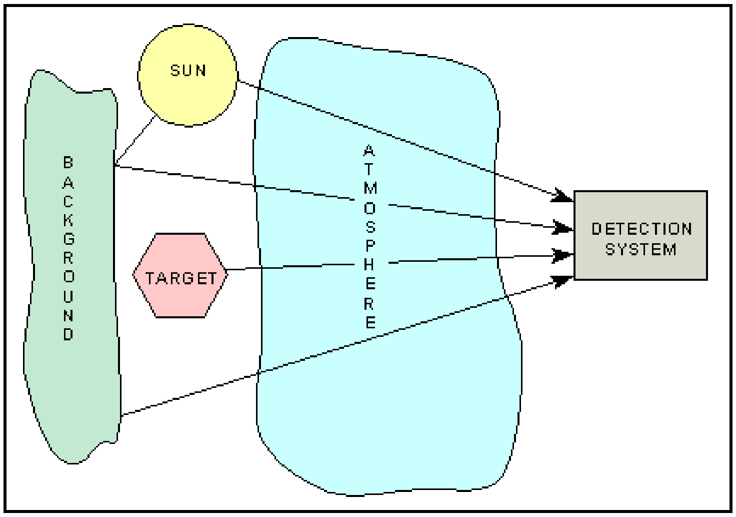

Figure 20. Generalized Detection Problem.

Figure 21 shows the basic relationships that are critical to detection of a target in the infrared. The figure is based on a generalized aircraft, but the principles apply whether the target is in fact an aircraft against a sky background or a ground vehicle being viewed from above against a terrain background. At detection, most targets are unresolved. The sensor’s ability to detect the target against background in this case is driven primarily by noise equivalent irradiance.

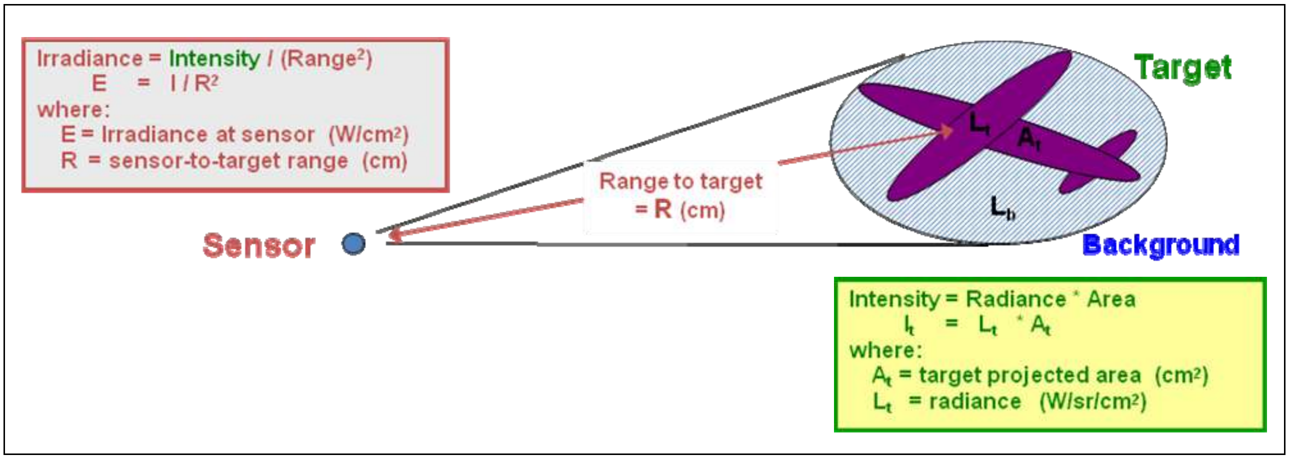
Irradiance = Intensity / (Range²)
E = I / R²
where:
E = Irradiance at sensor (W/cm²)
R = sensor-to-target range (cm)
Target
Lt
At
Lb
Range to target
= R (cm)
Sensor
Background
Intensity = Radiance * Area
It = Lt * At
where:
At = target projected area (cm²)
Lt = radiance (W/sr/cm²)

Figure 21. Detection of a Target with a Remote Sensor.

Each of the equations shown in Figure 21 in reality has atmospheric effects and attenuation due to transmission losses and contributes to path radiance. Just as the power distribution of the target and background vary with wavelength, atmospheric effects are also spectrally selective.

Figure 22 shows the roll off of irradiance as a function of range for two different aircraft. Detection occurs at the point of intersection with the sensors noise equivalent irradiance. In the case of threat missiles, there is often a signal-to-noise threshold required for launch of the missile to ensure target quality prior to launch. The product of noise equivalent irradiance and the threshold for these systems is known typically as the minimum trackable irradiance (MTI). This is the figure used to calculate detection range for such systems.

In an effort to simplify calculations, band average atmospheric transmission values are often applied during analysis of detection scenarios. The target itself is non-uniformly distributed as a function of wavelength, and the atmospheric effect is non-uniform, so this approach is mathematically incorrect since it pulls a non-constant term out of an integral. The degree of error introduced by the band average approach depends on the spectral distribution of the source and the overall transmission of the band in question, but caution should be applied when applying band averages. All calculations involving atmospheric propagation should be done spectrally and then integrated to provide the in-band value.

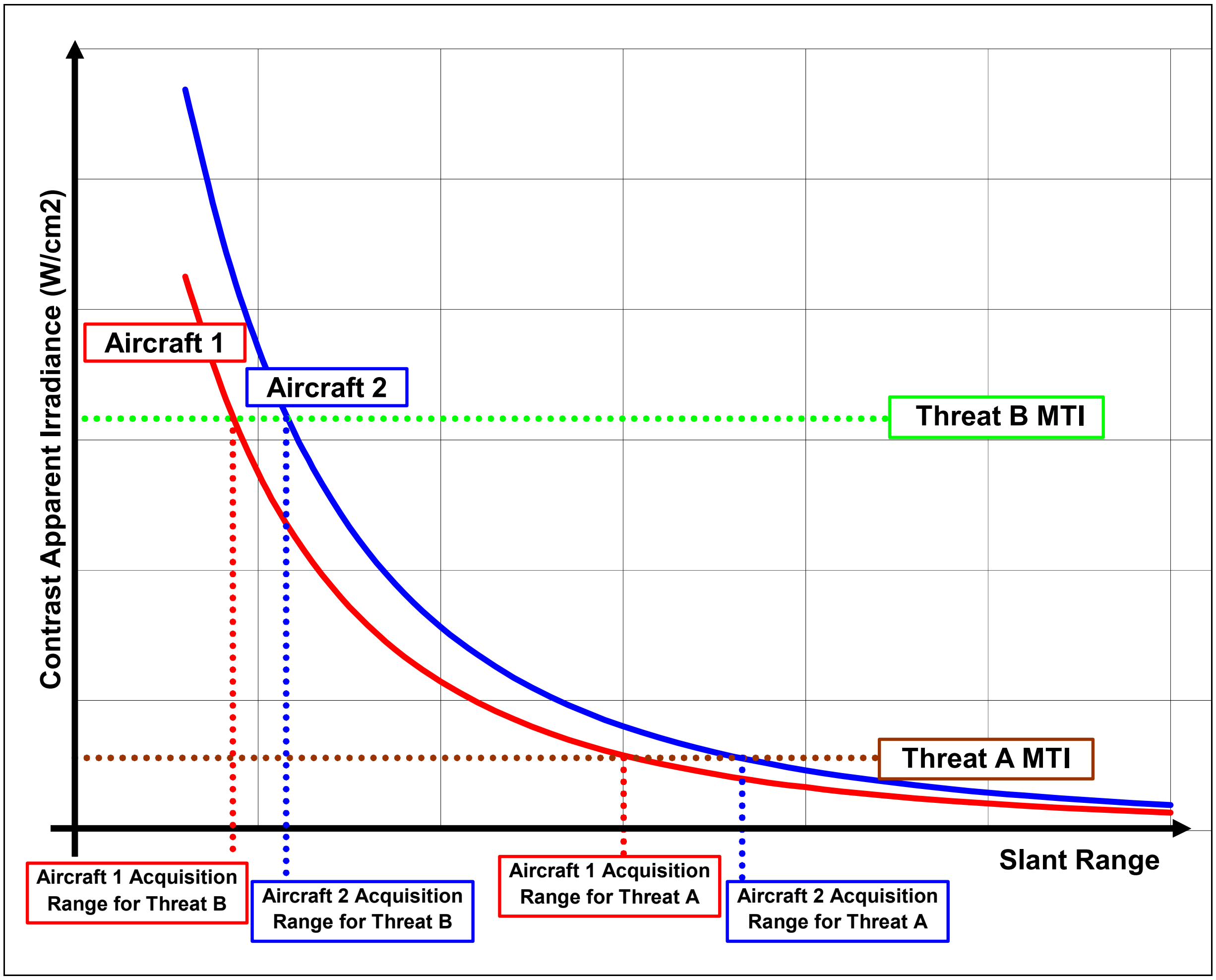

Figure 22. Detection Range Calculation.

Sensor Characterization

As described in a previous section, the output of every sensor is proportional to an integral of the received radiation weighted by the instrument's response function in that domain. Every sensor responds to radiation in accordance with the characteristics of the sensor and its components.

Characterization quantifies the sensor's response shape. Knowledge of response shape is essential to the design of a sensor and understanding its performance in with changing ambient conditions and against various types of targets in real environments.

Normalization

Calibration of a sensor, which describes its response to known input sources, and characterization of the sensor would ideally be the same process. Ideally, the absolute instrument response would be mapped over a domain with a traceable standard laboratory source that was tunable across the range of interest. In the spectral domain, for example, this would require a tunable monochromater whose output beam provided a level of spectral irradiance traceable to a radiation standard and that also had a cryogenically-cooled background source. The result would be an absolute spectral response function in units of output reading per unit radiance (or irradiance) as a function of wavelength.

In practice, it is sufficient and more practical to separate the characterization and calibration processes, so characterization determines only the relative response shape rather than absolute. Calibration then incorporates the results of the shape characterization to determine the absolute instrument response. Both, however, are important to understanding performance of a sensor.

Characterization uses a variety of different methods and sources to map relative response shapes. The response curve is then normalized and this normalized curve is used in the later derivation of the calibration coefficients. Different normalizations, such as normalization to an average value, are possible, but the convention throughout most of the measurement community today is to normalize response curves to unity at the peak.

When the contributions of all the components are combined into one curve, the result is then peak normalized and this, now unitless, curve is used in the calibration calculations.

Calibration

Sensor calibration, which is the process of relating the known input power to the output of a sensor, requires the use of standard sources, typically National Institute of Standards and Technology (NIST) traceable blackbodies and various other laboratory equipment such as collimators which make all of the rays coming from the source parallel to each other, thus representing a source at infinity. Figure 23 is a pictorial illustration of the calibration of a sensor.

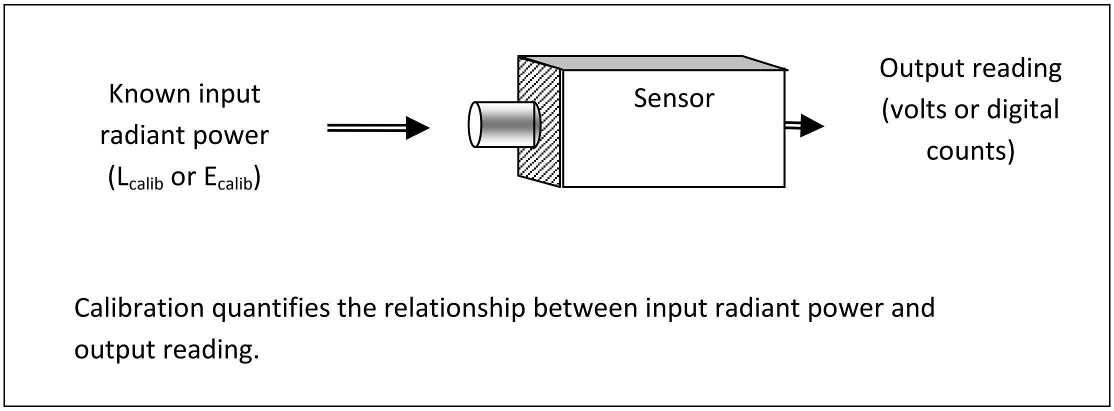
Known input radiant power (L_{calib} or E_{calib})
Sensor
Output reading (volts or digital counts)

Calibration quantifies the relationship between input radiant power and output reading.

Figure 23. Sensor Calibration Relates Input Power to Output.

Calibration of a sensor usually involves two major steps. Responsivity is the change in output of the sensor to changing input. For a sensor that responds linearly, for example, responsivity represents the slope of the curve when source radiance or incident irradiance is plotted against output voltage or counts for a digitized system. Figure 24 represents a calibration curve for a sensor that has a linear change in output over its dynamic range with changing input power. The slope of the curve is the "m" in the linear equation. Not all sensors respond linearly with power. Higher order response coefficients are common, especially for bolometers and infrared focal plane array sensors operated at short integration times. For these the process of determining the response of the instrument is the same, the curve just yields higher order terms.

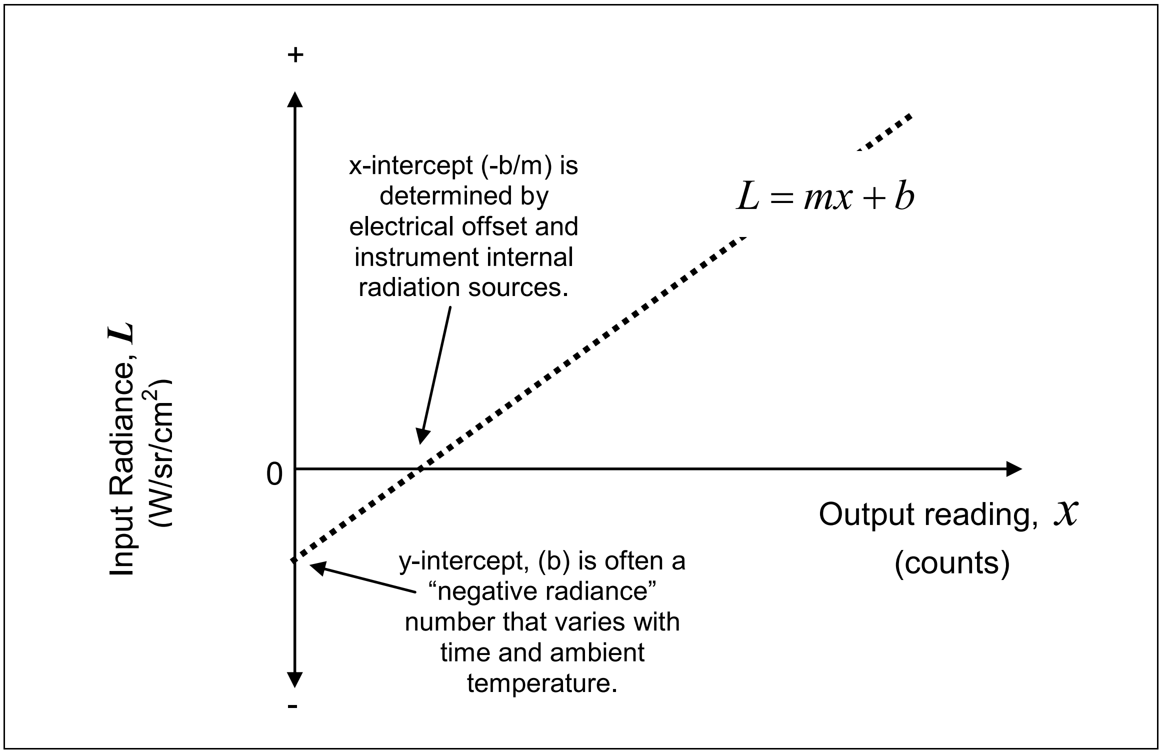
+
x-intercept (-b/m) is determined by electrical offset and instrument internal radiation sources.
$L = mx + b$
Input Radiance, $\boldsymbol{L}$ (W/sr/cm^2)
0
Output reading, x (counts)
y-intercept, (b) is often a "negative radiance" number that varies with time and ambient temperature.
-

Figure 24. Calibration of a Sensor to Determine Responsivity and Offset.

For most electro-optical sensors, the responsivity does not change with ambient temperature. In other words, the non-constant terms in the calibration equation, whether it is linear or higher order, do not change with changes in temperature. Over time, however, as detectors decay, responsivity decreases. This would show itself on the Figure 24 as an increase in the slope value, i.e., higher input power is required for the same output as the detector becomes less responsive.

Offset

Offset is another important parameter for instrument calibration. For any real (non-ideal) instrument, the response curve does not pass through zero. There are several reasons for this; one being that except in a complete vacuum, zero radiance does not exist. Additionally, contributions from detector noise and radiation from the optical elements in the lens, which cause the offset to drift with changes in ambient temperature, contribute to the offset term. Some amount of offset is designed into the system as well. All electronic circuits have some amount of DC drift. To prevent clipping of the signal if this drift should go below the lower limit of the analog-to-digital converter, the "bottom end" level is adjusted up to some offset level. The consequence of this offset voltage is the addition of a y-intercept term (b), which also must be quantified by the calibration if the sensor is a laboratory instrument. For a sensor that is used for contrast detection, the intercept value is unimportant since it subtracts out in the contrast calculation.

Sensitivity

Sensitivity for a sensor is determined to a large extent by the noise level in the detector output. For focal plane array detector, pixel-to-pixel non-uniformity also limits the sensitivity of the system since detection is determined by contrast with surrounding pixels. In practice, at least for cooled infrared sensors, detection is typically limited by background and not noise limits.

Instrument Response Uniformity and Non-Uniformity Correction

In reality, all sensors exhibit non-uniform response in all of the domains referenced previously. For example, in the spatial domain, the raw output of an infrared focal plane array detector exhibits pixel-to-pixel offset differences and response differences across the field-of-view. The response change is the result of two primary factors. Since each pixel is essentially a unique detector, it exhibits unique response because of manufacturing tolerances, slight differences in crystal structure, etc. Additionally, most electro-optical sensors implement an aperture or "field stop" in the case of infrared sensors, that limits the radiation that can reach the detector outside of the sensor's desired field-of-view. Radiation entering at angles off of normal to the detector shows a cosine roll-off in incident power. The result is a reduction in responsivity for pixels that are radially separated from the center of the detector. For the majority of systems, an optical gain correction can be applied to compensate for the change in response. The typical method involves using an extended blackbody source that fills the FOV of the sensor. Reference images are collected with the source at two temperatures that are well separated across the sensor's dynamic range. This process is typically called a "2-point" correction. Actual temperature is unimportant. Slope corrections can be determined for each pixel. The result is a gain map that can be stored in the sensor electronics that can be applied to each image to correct for the non-uniform spatial response across the detector array. Pixel-to-pixel offset maps can be determined using one of the same reference images. Pixel slope and offset corrections are typically derived as normalized quantities relative to a center pixel, average of center pixels, or maximum value. The application of the correction maps to the images is commonly referred to as "non-uniformity correction." Figure 25 shows the transition from a raw image to a non-uniformity corrected image.

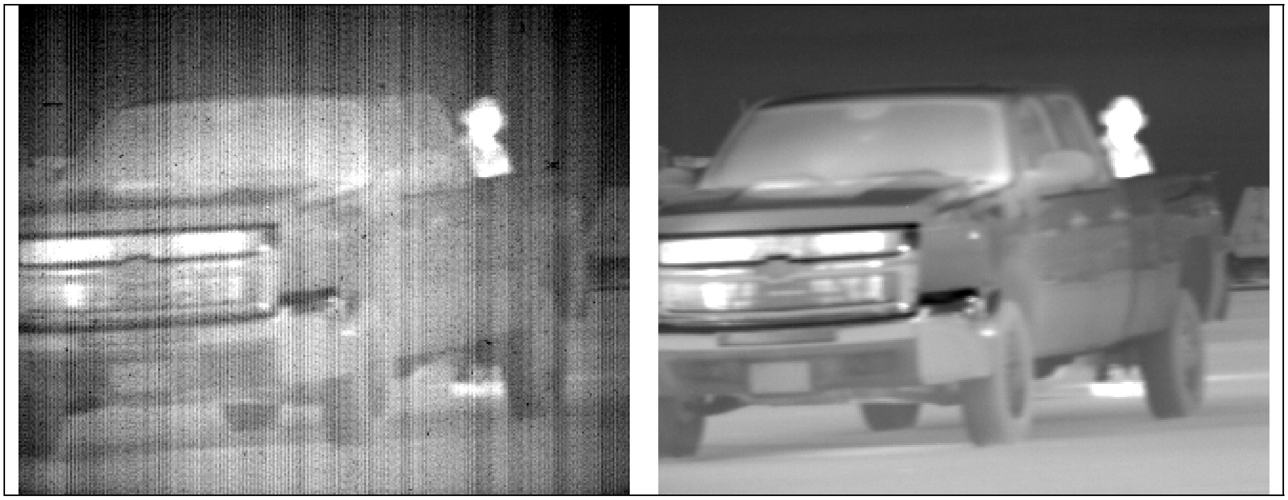

Figure 25. Non-Uniformity Correction of a Mid-IR Image.

Bad Pixel Replacement

Focal plane arrays have pixels that are either unresponsive or responsive outside of useful limits. Figure 25 shows some of the bad pixels that appear as small black spots in an image from an InSb IRFPA imager. Bad pixels can be identified during laboratory calibration or with a sensor mounted reference source. There are many approaches to replacement of bad pixels and the best approach often depends on the sensor characteristics and its application. One common approach is simply to replace the pixel with the average of its nearest neighbors.

INFRARED THREATS TO AIRCRAFT AND THEIR COUNTERMEASURES

IR guided missiles are the largest single cause for aircraft losses since the start of the 1991 Gulf war. All missiles designed within the last 20 years have counter-countermeasures circuitry. Every missile can be defeated with IR countermeasures given time to develop and test devices and techniques, but many missiles have not been exploited and the variety and complexity of the different designs present formidable challenges to the US countermeasure community.

The IR “signature” of any aircraft has three main components:

- Engine exhaust plumes
- Engine hot parts (tailpipe, etc)
- Airframe (aerodynamic heating & reflection from sun, earth, etc)

Infrared guided missiles modulate the signal produced by the aircraft in contrast with its background. Previous generations of missiles used reticles to produce signal waveforms that would provide spatial and temporal information from which signal processing could produce trackable information.

IR Missile Operation

Aircraft (or any other object) can be intercepted using several different types of guidance. The simplest type is pure pursuit, where the missile is always pointed directly at the target location. This is not aerodynamically efficient since the missile would follow a longer (curving) flight path when following a crossing target.

Most missile guidance systems are designed to lead the target so that intercept occurs at the point where the target will be at the time the missile arrives. This requires that the missile fly a course so the relative bearing to the target stays constant (constant “line of sight” angle). The LOS angle is determined by missile speed relative to the target (higher closing speed = smaller angle). The size of the angle isn’t important; only that it be constant (zero line of sight rate) as shown in Figure 26.

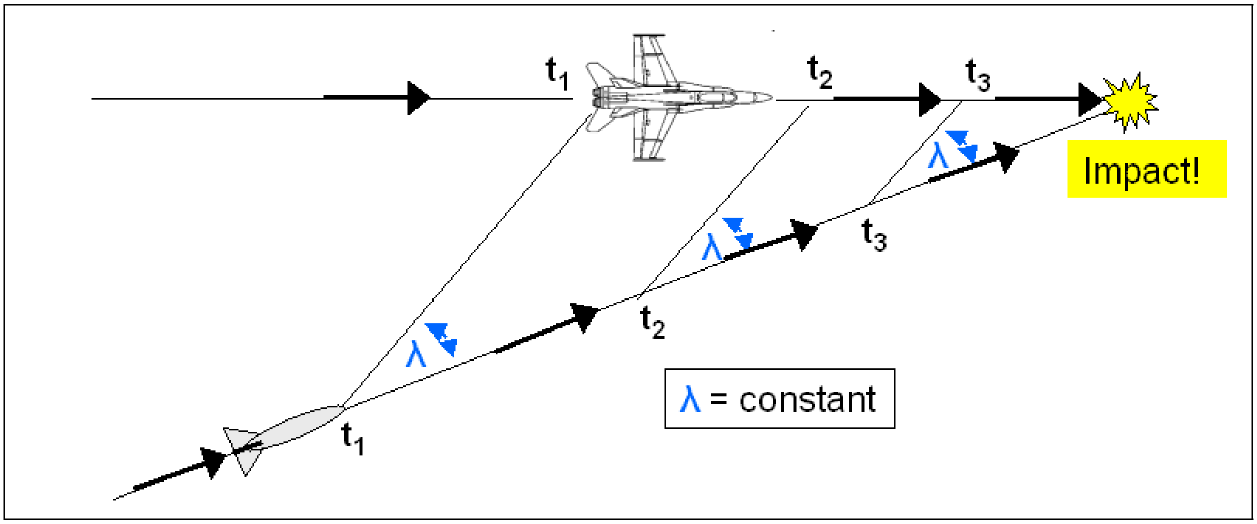

Figure 26. Missile Proportional Navigation.

This intercept course ("proportional navigation") requires that the missile have two separate servo loops: (1) a target tracker and (2) a wing control servo to control direction of flight

For a missile to guide to its intended target, it needs a tracker, which contains the following elements:

- Optics to collect and focus IR from target.
- Gimbals to allow movement to point the optics.
- Gyro stabilization to isolate optics from missile body.
- Detector to convert the received IR to electrical signal.
- Stabilization (gyro) to isolate from missile body.
- A method to determine target direction to enable closed-loop tracking.
- A method to distinguish the target from natural background.

The target tracker is the "window" into the missile's guidance through which it can be deceived by countermeasures.

The problem of determining target direction with a single detector was solved by forming an image of the target scene onto the center of a reticle disk that spun with the optics. Unlike, for example, the reticle in a rifle telescope that superimposes cross hairs, the reticle in a missile acts as a kind of shutter that blocks the passage of IR through part of the reticle and allows IR to pass through the other part. A target image falling on the opaque portion is blocked and produces no detector signal. A target image falling on the transparent portion is passed on to the detector. When the reticle is spun, IR from a target off center is alternately passed and blocked, resulting in amplitude modulation (AM). The phase of this modulation relative to a spin reference is used to tell target direction from center. A closed servo loop moves the optics to keep the target centered on the reticle. This is depicted in Figure 27.

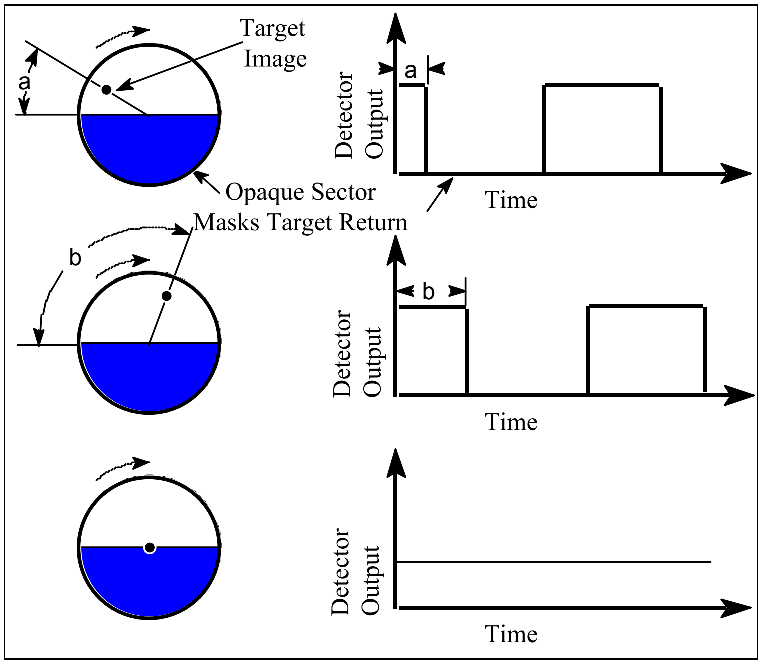
Target Image
a
Opaque Sector Masks Target Return
b
Detector Output
Time
a
b

Figure 27. Basic Reticle Design.

Target trackers have another problem: The aircraft target must be distinguished from natural background sources, such as sunlit clouds and terrain. To solve this, they look for features in the spectral, spatial, and temporal domains where the target is different from background.

- Temporal: There is no difference. Neither clouds nor aircraft signature are time varying.
- Spectral: Some difference. Choice of wavelength band yields helpful differences between target and background, but this is not sufficient by itself.
- Spatial: The most viable option. Aircraft are smaller than clouds and terrain. Background radiation can be greatly reduced by spatial filtering.

If half of the reticle is made with opaque "spokes," then some irradiance from targets with small images (such as aircraft) will be modulated more completely and generate a stronger signal at a faster modulating rate than large images (clouds) as shown in Figure 28.

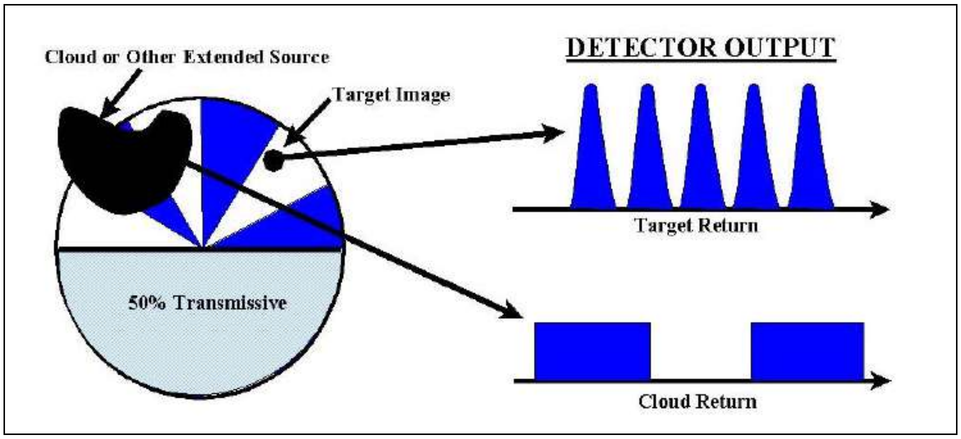
Cloud or Other Extended Source
Target Image
DETECTOR OUTPUT
50% Transmissive
Target Return
Cloud Return

Figure 28. IR Seeker Design for Background Discrimination.

The past figure and the following two figures (29 and 30) depict a spin-scan reticle used on the early Sidewinder designs. After the detector preamp, signal goes through a narrow bandpass filter to improve S/N. The AM waveform is then rectified and filtered. Target direction is determined from AM envelope phase.

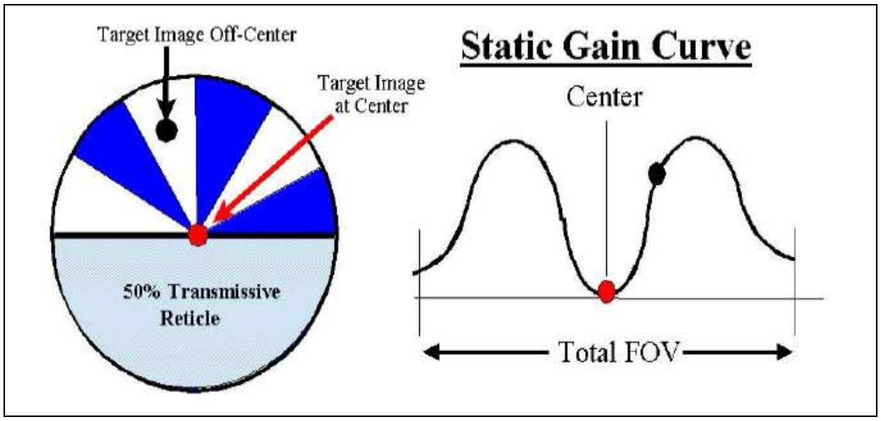
Target Image Off-Center
Target Image at Center
50% Transmissive Reticle
Static Gain Curve
Center
Total FOV

Figure 29. Spin Scan Seeker.

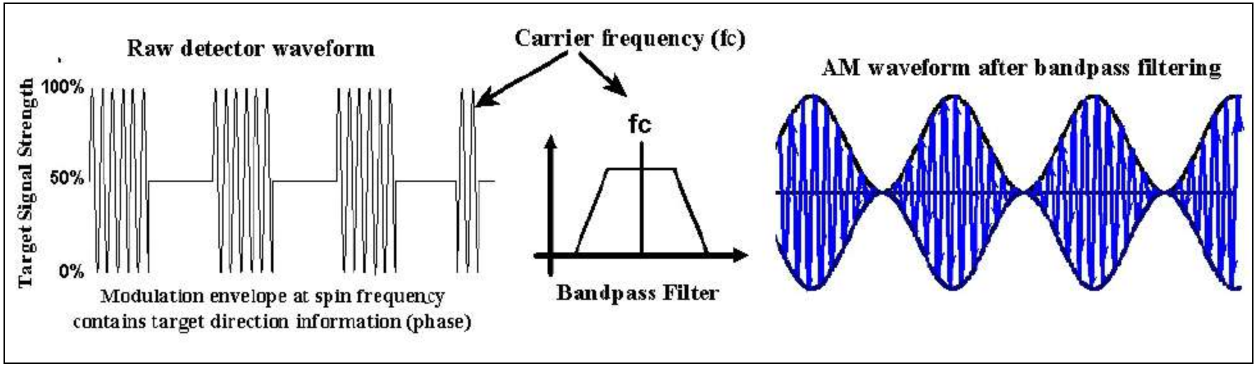
Raw detector waveform
Target Signal Strength
100%
50%
0%
Modulation envelope at spin frequency contains target direction information (phase)
Carrier frequency (fc)
fc
Bandpass Filter
AM waveform after bandpass filtering

Figure 30. Spin Scan Waveforms for Off-Center Target.

Spin-scan has the following characteristics that are important to countermeasures:

- The tracker loop drives to null the signal to zero. This occurs when the target is on the optical axis and the target image is at the center of the reticle.
- If the target is off-center, an AM carrier "error signal" is generated, where the phase of the modulation envelope indicates the target direction.

With spin scan, the missile is always looking at the target. This vulnerability to jammers led to the next evolution in target trackers: conical scan.

Conical scan borrows concept from early fire-control radars, which used a nutating feed horn. A con scan tracker is shown in Figure 31. With con scan:

- The secondary mirror of the Cassegrain is canted so the field of view seen by the detector sweeps out a pattern of overlapping circles.
- A target image at boresight falls near the edge of the reticle instead of center.
- Reticle pattern is same all the way around. (Usually tapered spokes.)

- Modulation of target near boresight is FM rather than AM. This allows tighter tracking.
- For larger angles off boresight, the target image falls outside the FOV of the detector for part of the scan. The modulation then becomes AM.

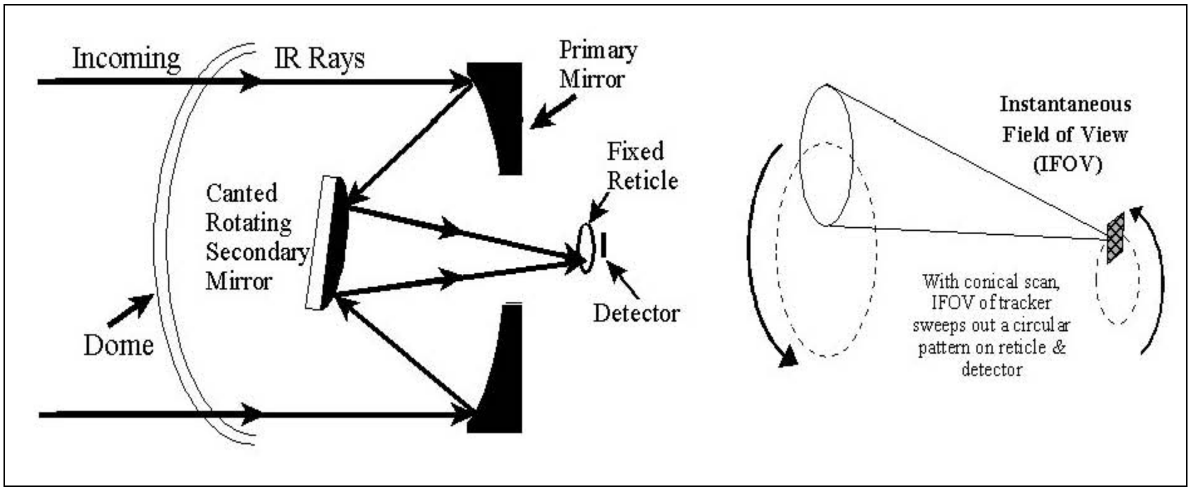
Incoming
IR Rays
Primary Mirror
Fixed Reticle
Canted Rotating Secondary Mirror
Detector
Dome
Instantaneous Field of View (IFOV)
With conical scan, IFOV of tracker sweeps out a circular pattern on reticle & detector

Figure 31. Conical Scan Tracker.

In the con scan tracker, as the missile instantaneous field of view nutates about a target on boresite, (moving through positions at t_1 through t_5 shown in Figure 32), the apparent position of a target image on the reticle sweeps out the circular pattern shown.

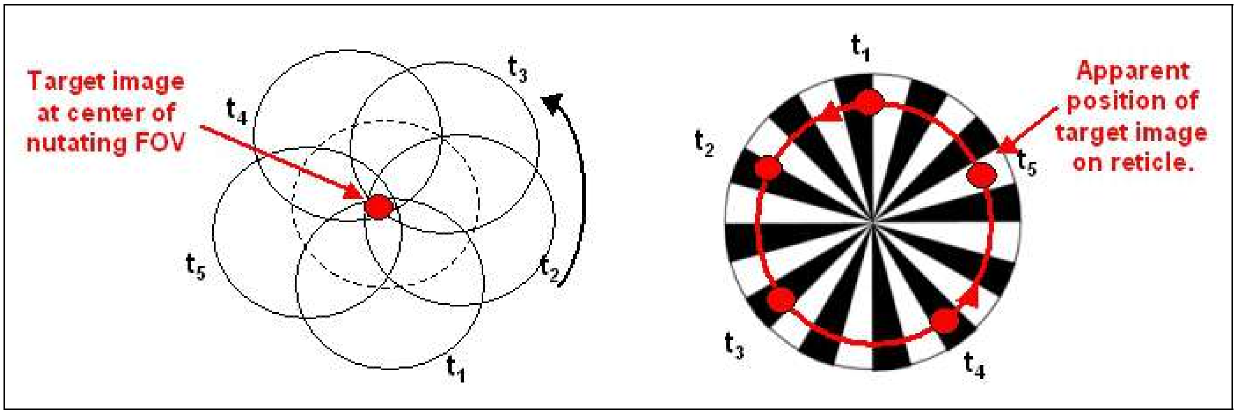
Target image at center of nutating FOV
t_1
t_2
t_3
t_4
t_5
Apparent position of target image on reticle.
t_1
t_2
t_3
t_4
t_5

Figure 32. Image on a Con Scan Reticle: Target at Boresite.

If the target is off boresight as shown in Figure 33, the detector receives a signal of varying pulse widths.

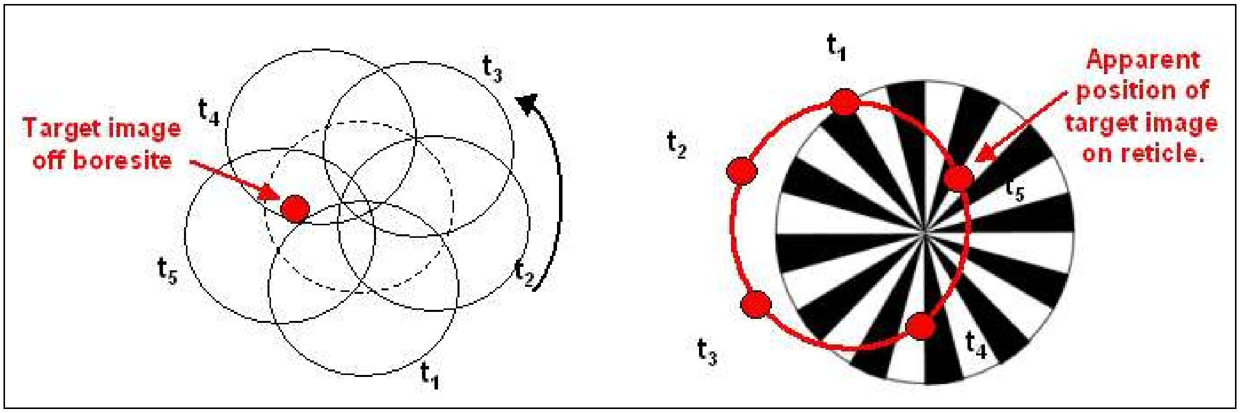
Target image off boresite
t_1
t_2
t_3
t_4
t_5
t_1
t_2
t_3
t_4
t_5
Apparent position of target image on reticle.

Figure 33. Image on a Con Scan Reticle: Target off Boresite**.**

The waveform produced by a target on boresite is a constant amplitude carrier at the reticle chopping frequency as shown in Figure 34. A target slightly off boresite produces a constant amplitude carrier that is frequency modulated at spin frequency.

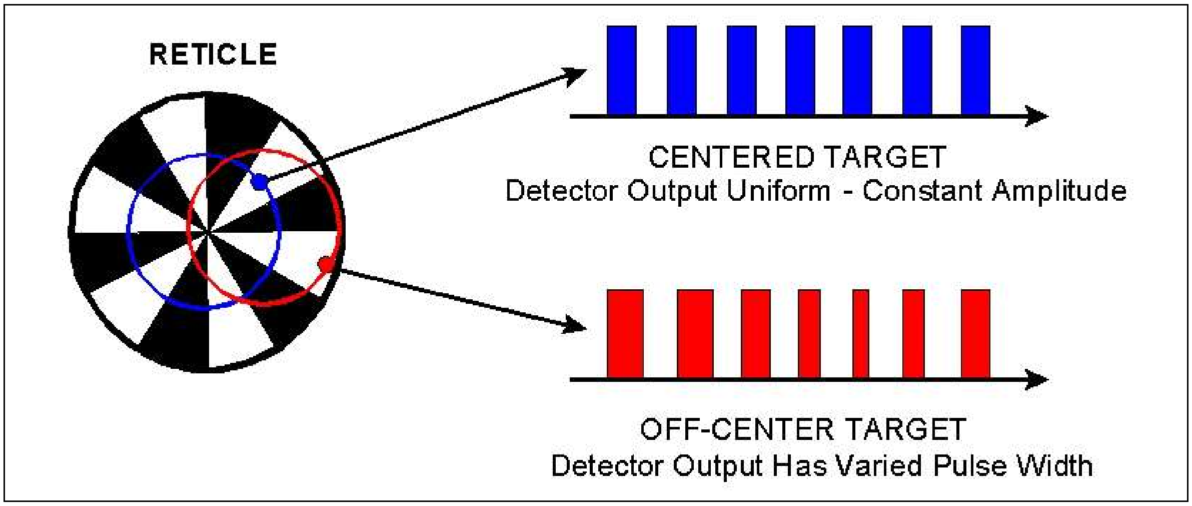
RETICLE
CENTERED TARGET
Detector Output Uniform - Constant Amplitude
OFF-CENTER TARGET
Detector Output Has Varied Pulse Width

Figure 34. Conical Scan Seeker Output.

A target further off boresite leaves the missile field of view during part of the scan, producing an amplitude-modulated waveform similar to that of a spin scan tracker. The important difference is that with a spin scan tracker, the target never leaves the missile field of view. With con scan, the target may fall outside the missile FOV at certain times during the scan. Because con-scan trackers do not necessarily view the target continuously, they can have high resistance to jammers.

Other types of seeker scan patterns now exist. The Rosette scan pattern shown in Figure 35 is one such example. It has an even higher resistance to countermeasures.

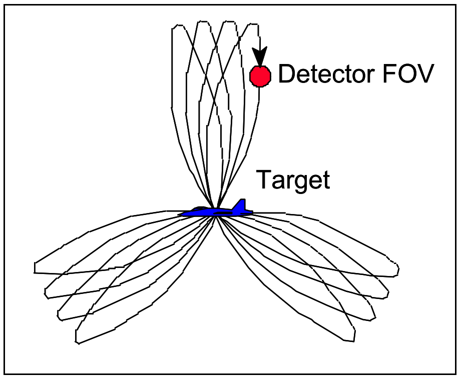

Figure 35. Rosette Scan Pattern.

Imaging arrays of detectors without reticles are newer yet. They may be classified as either staring (every pixel sees the entire scene), or scanning arrays, where the optics plays a role in determining which "pixels" are exposed to optical / IR radiation.

INFRARED COUNTERMEASURE (IRCM)

Flares

Figure 36 shows a Navy F/A-18E Super Hornet aircraft dispensing IRCM flares from its internal flare dispensers. IRCM have been the staple of countermeasures protection for military aircraft more than four decades. Flares are designed to transfer the track of an attacking infrared missile by exhibiting characteristics that confuse the tracking and guidance algorithms built into the missile. Modern missiles incorporate Counter-Countermeasures (CCM) capabilities including hardware configurations, circuitry, and logic to help identify countermeasures and reject them from processing. CCM capabilities can be based on spectral, spatial, and temporal features of the target scene. As missiles continue to improve in their sensitivity, range, maneuverability, and CCM capabilities, flares continue to evolve in order to keep pace with the evolving threat.

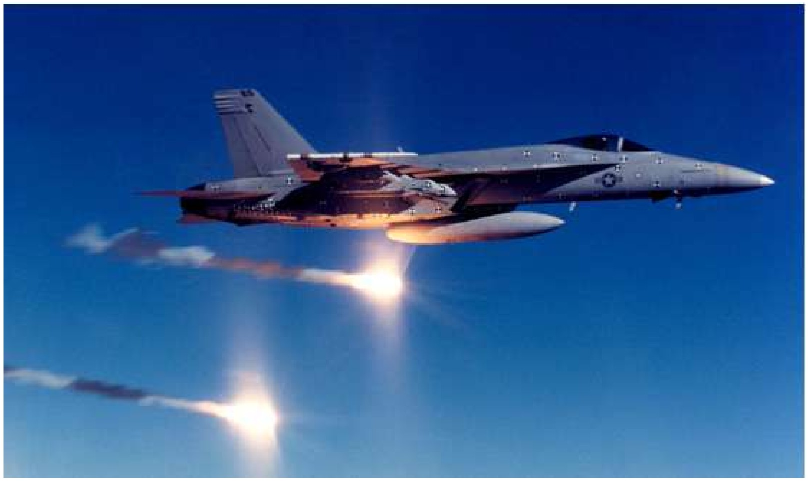

Figure 36. Navy F/A-18E Aircraft Dispensing IRCM Flares.

Over the years the Navy has fielded many flare types including the MJU-2/B, MJU-8, MJU-32, MJU-38, MJU-27, MJU-49, and many other improved versions of these flares and other types as well. IRCM flares continue to be the prevailing countermeasure for military aircraft protection by offering cost effective and robust protection.

Dispenser Systems

Most Navy fixed wing and rotary wing aircraft are equipped with countermeasures dispenser systems. These systems are critical to the survivability of the aircraft in a hostile threat environment. Modern dispensers such as the AN/ALE-47 offer high reliability and substantial programming capability that allows flare dispenses to be tailored for maximum protection of the host aircraft type. AN/ALE-47 is highly integrated into the aircraft over the 1553 data buses. The dispenser can incorporate information from several aircraft systems including missile-warning systems to improve its responses to threats and provide vital situational awareness to the aircrew.

Impulse Cartridges

IRCM flares are dispensed from the aircraft flare dispensers with electrically initiated impulse cartridges. Impulse cartridges incorporate energetic materials within a small confined canister. Upon application of a voltage to the electrical leads, a bridge wire in contact with the energetic materials burns through, igniting the propellant materials. The expanding gases push the flare from its case, held captive in the aircraft dispenser. Impulse cartridges for Navy use have been designed to withstand the extreme electro-magnetic environments encountered around aircraft carriers and other combat ships. Examples of impulse cartridges include the CCU-63 and CCU-136.

Infrared Jammers

Several lamp-based and mechanically modulated jammers have been developed over the years for protection of aircraft. Examples include the AN/ALQ-144 and AN/ALQ-157, predominantly used on helicopters and cargo aircraft. These jammers offer some level of protection over a broad field-of-regard and offer the advantage of continuous operation.

In principle, these jammers produce a modulated signal in the track band of the threat that corrupts the target tracking pulses in the missile seeker.

Passive Missile Warning Systems

Infrared-guided weapons provide passive attack capabilities against military vehicles. Unlike a radar-guided weapon that actively emits radiation and tracks the reflected pulses from the target, infrared weapons track radiation already being emitted from the target. Attacking missiles fly at very high speeds, and they are exceptionally maneuverable. Missile warning systems must be capable of detecting the threat, alerting the aircrew, and cueing a countermeasures response within sufficient time to counter the attacking missile. The time from launch to impact can be very short, making timely detection critical. Active warning systems have been developed in the past that use Doppler Radar capabilities to detect missiles, but passive missile warning systems have been preferred because of the desire to minimize emissions from the aircraft under attack.

Several passive missile-warning systems have been developed over the years for military aircraft. These systems operate in a variety of different parts of the electro-optical spectrum, but the most common are ultraviolet and infrared sensor systems. Examples of passive missile warning systems include AN/AAR-47, AN/AAR-54, AN/AAR-57, AN/AAQ-24 (both passive and active components), and the Joint and Allied Threat Awareness System (JATAS), currently under development by the Navy. Passive missile warning sensors continue to improve with advances in detector technologies, particularly with imaging detectors. These sensors provide excellent angle-of-arrival information, necessary to support cueing of laser based countermeasures, and advanced processing to detect and declare threat missiles in cluttered environments.

Laser Countermeasures

Laser-based infrared countermeasures have been in development for many years. Several systems have been fielded over the past fifteen years including the AN/AAQ-24 system on Air Force cargo aircraft and helicopters and a derivative system for Marine Corps helicopters.

Although configurations vary, most of these systems incorporate a single multi-band laser or several single-band lasers that produce modulated waveforms designed to corrupt a missile's guidance target tracking. The laser optics are located in a tracking gimbal that provides agile and rapid pointing over a broad field-of-regard. Laser based countermeasures require a relatively high angle of arrival accuracy from the host aircraft's missile warning sensor. Upon declaration of the threat, the missile warning system hands-off track to the tracking gimbal and cues the lasers to lase. A tracking camera in the tracking gimbal with high optical resolution helps to maintain track on the threat missile through the engagement period.

LASERS

The word laser comes from Light Amplification by Stimulated Emission of Radiation. A laser system emits light that is generated through a process of stimulated emission. The radiation produced by a laser exhibits high temporal and spatial coherence. In order to begin the process of stimulated emission, the lasing medium absorbs the energy from a pump source. The atoms in the lasing medium are excited to a higher energy state. These atoms will eventually return to their ground state. A large number of atoms that are excited to higher states create a population inversion. Population inversion describes the number of atoms in excited state versus the number of atoms in the ground state. In order for the atoms to return to their ground state, they must release energy. This energy is released in the form of photons. Energy of a photon is expressed as

$$E = \frac{hc}{\lambda}$$

Where

E = Energy, generally electron volts (eV)
h = Planck's constant = 4.136×10^{-15} (eV·s)
c = speed of light = 2.998×10^{8} (m/s)
λ = wavelength of light in meters

The energy that must be released by the atom to return to the ground state will direct the wavelength of the photon emitted since h and c are constants. If all the excited atoms released the same amount of energy to return to their ground state, the released photons would all have the same wavelength and would be considered fully monochromatic. Most lasers do not emit a single wavelength but a range of slightly differing wavelengths (Δλ).

The lasing medium may be a solid, a gas, liquid, or plasma. Some laser types include gas, chemical, dye, fiber-based, solid-state and semiconductor lasers. The laser radiation can be output in a continuous wave (CW) or in a pulsed wave. A continuous wave laser emits light that maintains a steady amplitude and frequency. A pulse wave will vary in amplitude and is also characterized by the systems pulse repetition frequency (PRF). The PRF is defined as the number of pulses emitted during a unit of time. Figure 37 shows the spectral output of several laser types.

The first laser was constructed by Theodore Maiman at Hughes Research Laboratories in Malibu, California. This laser was a pulsed, solid-state ruby laser. The ruby laser uses a synthetic ruby crystal as the lasing medium. A xenon flash lamp is used to excite the atoms in a ruby rod to higher energy levels. The highly polished and mirrored ends of the rod form a resonant cavity. One end of the rod has a slightly lower reflectivity. The lamp excitation produces an inverted population of excited atoms, which are stimulated to relax to lower energy levels releasing their extra energy as photons. Repeated reflections off the mirrored ends of the rod causes the photons to bounce back and forth through the rod stimulating further emissions at the same wavelength and phase producing a highly coherent beam, which finally passes through the lower reflectivity end.

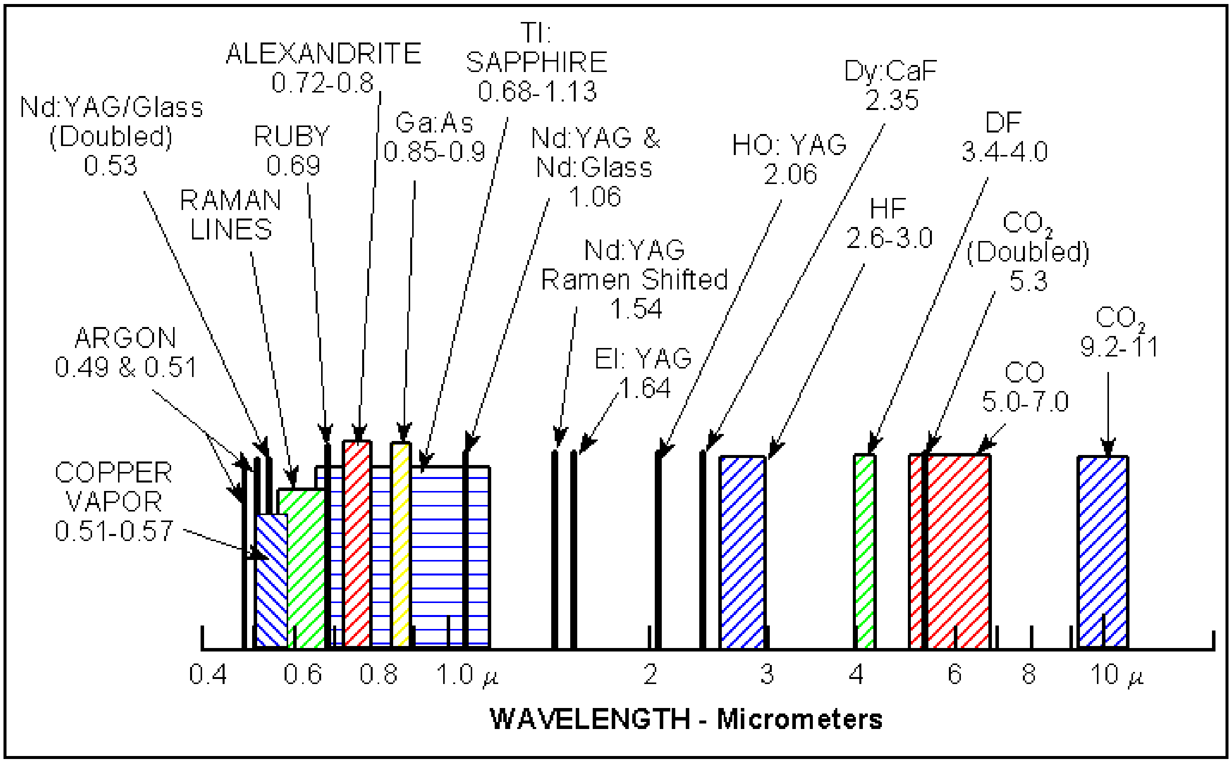

Figure 37. Spectral Lines / Ranges of Available Lasers.

The typical laser rangefinder uses a solid-state laser with a neodymium-YAG crystal lasing at 1.06 μm.

Gas lasers can be pulsed or CW. The gas dynamic laser obtains its inverted population through a rapid temperature rise produced by accelerating the gas through a supersonic nozzle. In chemical lasers the inversion is produced by a chemical reaction. In the electric discharge laser the lasing medium is electrically pumped. The gas can also be optically pumped. In an optically pumped gas laser the lasing medium is contained in a transparent cylinder. The cylinder is in a resonant cavity formed by two highly reflective mirrors.

Many gas lasers use carbon dioxide as the lasing medium (actually a mixture of CO_2 and other gases). These are the basis for most high energy or high power lasers. The first gas laser was an optically pumped CW helium-neon laser. The common laser pointer is a helium-neon laser operating at 0.6328 µm. The lasing medium is a mixture of helium and neon gas in a gas discharge or plasma tube.

The dye laser is an example of a laser using a liquid for the lasing medium. The lasing medium is an organic dye dissolved in a solvent such as ethyl alcohol. Dye lasers operate from the near UV to the near IR, are optically pumped, and are tunable over a fairly wide wavelength range.

Another type of laser is the semiconductor or injection laser, also known as a laser diode. The junctions of most semiconductor diodes will emit some radiation if the devices are forward biased. This radiation is the result of energy released when electrons and holes recombine in the junction. There are two kinds of semiconductor diode emitters: (1) the light emitting diode (LED), which produces incoherent spontaneous emission when forward biased and which has a broad (800 angstrom) spectral output, and (2) the laser diode, which maintains a coherent emission when pulsed beyond a threshold current and which has a narrow spectral width (< 10 angstrom). In the laser diode the end faces of the junction region are polished to form mirror surfaces. They can operate CW at room temperatures, but pulsed operation is more common. Figure 38 shows a typical diode laser structure.

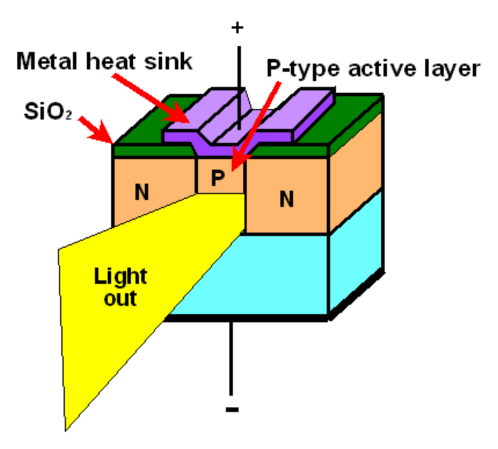

Figure 38. Diode Laser Construction.

Fiber lasers use fibers that are doped with rare-earth elements as the pumping medium. These rare-earth elements include elements such as Erbium (most common), Ytterbium, and Neodymium. There are other elements such as Thulium that are used for doping purposes. Erbium doped fiber lasers can emit in the 1.5 to 1.6 micron wavelength, which is important due to eye safety concerns in this part of the spectrum. Other wavelength emissions for Erbium include 2.7 and 0.55 microns.

Fiber based laser systems are beneficial in many ways. The fiber gain medium is compact compared to many other types of gain medium and is highly efficient. The fiber gain medium can also be physically manipulated to save space. Fiber based lasers are able to achieve high output powers. The gain medium of a fiber laser can extend for several kilometers to achieve these higher power outputs. The fact that the light is already propagating in a flexible fiber can also allow for system designs that implement a gain cavity in one location and then deliver the output in another location.

Q-switching is a common means of obtaining short intense pulses from lasers. The Q-switch inhibits lasing until a very large inverted population builds up. The switch can be active or passive. A passive Q-switch switches at a predetermined level. An active Q-switch is controlled by external timing circuits or mechanical motion. The switch is placed between the rod (or lasing medium) and the 100 percent mirror. Figure 39 shows an arrangement using a Pockels cell as an active Q-switch.

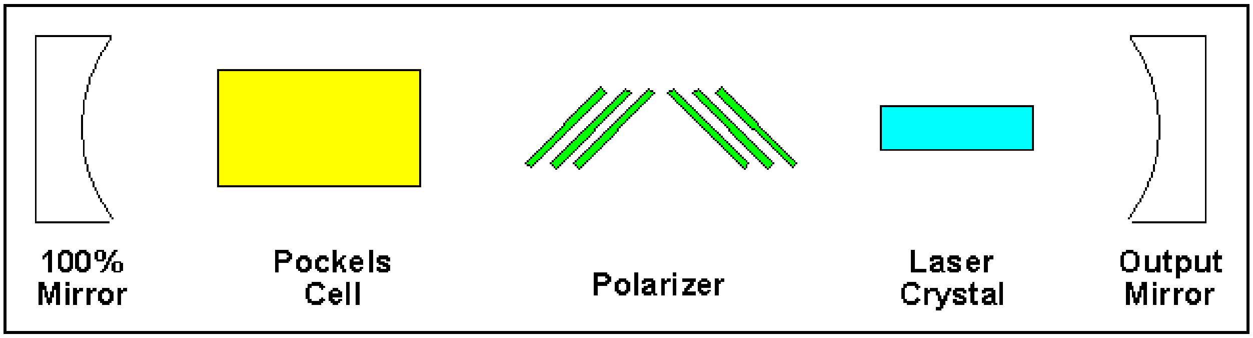

Figure 39. Q-Switch Arrangement.

Other methods of obtaining pulsed operation include using pump sources that are pulsed and mode-locking.

FIBER OPTICS

Fiber optic cables are the optical analog of RF waveguides. Fiber optic cables are made from transparent dielectrics. The fiber optic cable acts as an optical waveguide allowing light to propagate along the length of the fiber by using the principle of total internal reflection. This phenomenon can only occur under certain conditions relating to the material indices of refraction and the light ray's angle of incidence. Some benefits of fiber optic fiber include low losses, bandwidth, electromagnetic interference immunity, size, and weight.

Consider the physical construction of a bare optical fiber, depicted in Figure 40. A bare optical fiber is simply the inner glass core and the surrounding glass sleeve. The core must have a higher index of refraction than the cladding, $n_1 > n_2$. When $n_1 > n_2$, light impinging the boundary between the core and the cladding will totally internally reflect if the incident angle at each reflection is greater than the critical angle, θ_c. $\sin \theta_c = (n_2/ n_1)$

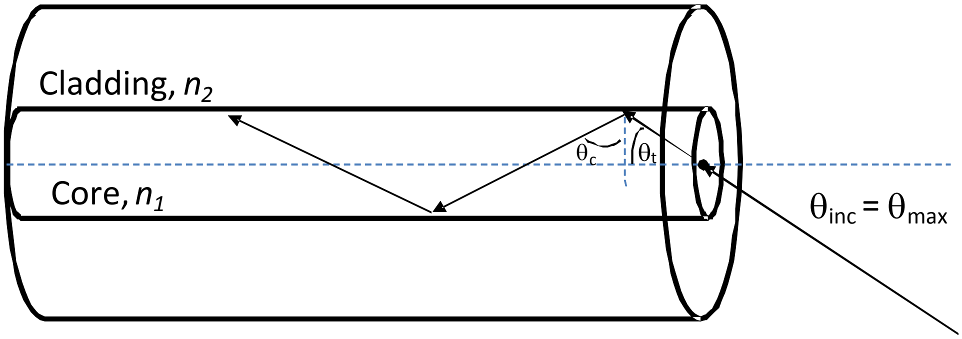

Figure 40. Bare Fiber Optic Cable.

Incident rays on the face of the fiber must intersect at angles less than θ_{max} for the internal ray, θ_t, to intersect at θ_c. When rays intersect the front face of the fiber at angles greater than θ_c, they are only partially reflected in the core and will leak out.

There are many varieties of optical fibers. Optical fibers can either be single mode or multimode. Single mode fibers are fibers which propagate a single mode down the length of the fiber while multimode fibers can propagate many modes. Single mode fibers typically have a much smaller core diameter, typically around 8 to 10 µm. Their cladding is usually 125 µm. Multimode fibers typically have core diameters around 62.5 µm with 125 µm claddings. These diameters can vary depending on the application. Loss in multimode fibers over a 1 kilometer distance is typically around 1 dB at 1310 nm. This value will vary some with changes in wavelength. Single mode fibers can maintain the quality of a light pulse over longer distances than multimode fibers due to modal dispersion effects that occur in multimode fibers. Typical losses for a single mode fiber over 1 kilometer is approximately .3 dB at 1310 nm. Again, this value will vary some with changes in wavelength. However, multimode fiber is much less expensive than single-mode and can have a lower connection loss due to the larger core diameter. Multimode fiber is commonly used in communications.

In addition to single mode or multimode, a fiber can have a step index profile or a graded index profile. Figure 41 depicts the two profiles. The step index profile maintains a uniform index of refraction within the core. A graded index profile has a peak index of refraction at the center of the core. The index of refraction value rolls off from the center to lower values closer to the cladding interface. This profile assists with the modal dispersion issue found in multimode fiber.

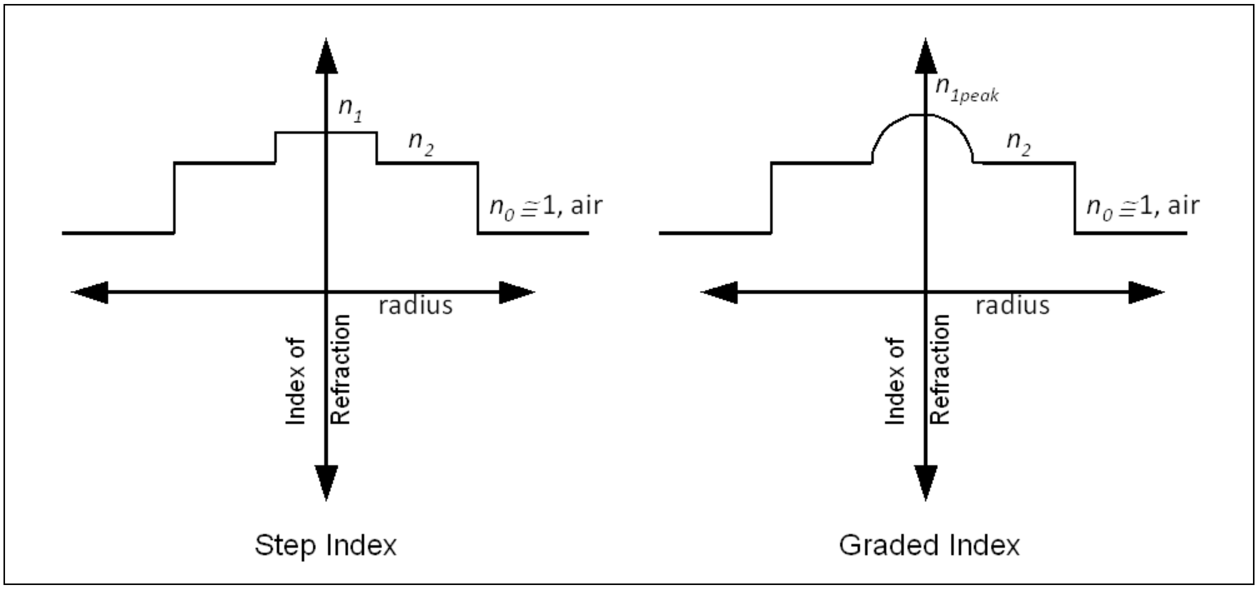

Figure 41. Fiber Profiles for a Single Mode Step Index and a Multimode Graded-Index Fiber.

Most fiber is not used in a bare form and has some additional layers of protection around the cladding. These layers can include a 250 µm buffer with a 900 µm PVC tight buffer. Some fiber will also contain aramid yarn followed by a 3 mm PVC furcation tube. Buffer tubes are often used to assist with identification and provide damage protection. The outer layers can provide additional isolation from environmental factors and lower optical crosstalk.

There are also more specialized types of fiber that include polarization maintaining and photonic crystal fibers. Polarization maintaining fibers are not constructed with a cylindrical core but instead use elliptical, bow-tie styled cores or stress rods located in the cladding (PANDA style). These are shown in Figure 42.

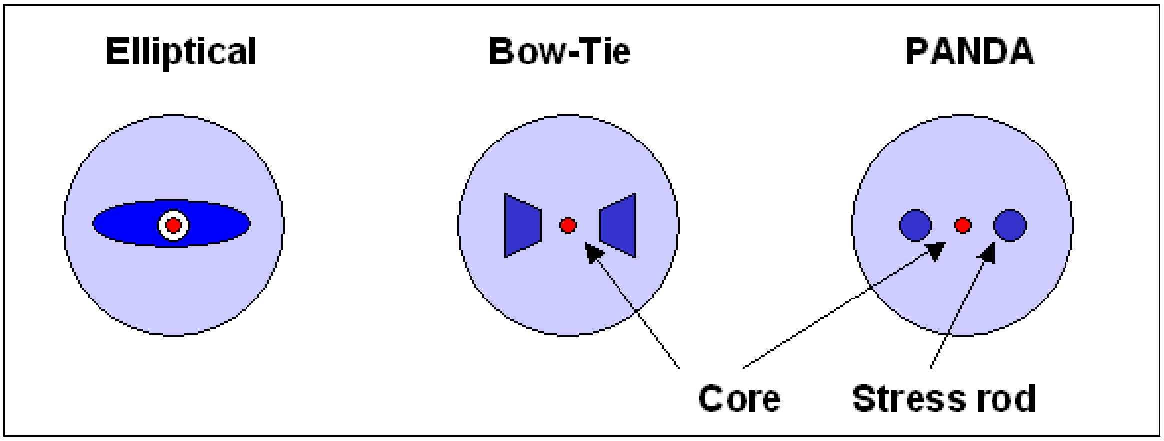

Figure 42. Polarization Maintaining Optical Fiber Types.

Polarization maintaining fibers maintain the state of the linearly polarized light propagating through the fiber. This type of fiber is used when the polarization state of the light cannot vary within a system. Photonic crystals propagate light by an arrangement of very small and closely spaced air holes that are maintained throughout the length of the fiber. Applications of photonic crystal fibers are varying and can be used in fiber lasers, amplifiers, sensors, and telecom.

It is well known that fiber optics has many communication applications; however, improvements in fiber optic technology have lent themselves to many EO applications. Many EO components are now fiber based and can interface with the tremendous advancements in fiber-based laser systems as well as other EO systems. An example of the use of fiber optics in an EW system is the AN/ALE-50 and 55 Fiber-Optic Towed Decoy (FOTD). The FOTD uses fiber optic cabling to communicate with the jammer.

This page intentionally left blank.

LASER SAFETY

Lasers are divided into the following classes:

Class 1	Low power / non-hazardous
Class 2/2a	Low power / minor controls necessary Emit less than 1 mW visible CW radiation. Not considered hazardous for momentary (<0.25 sec) unintentional exposure. Class 2a lasers are those class 2 lasers not intended to be viewed, i.e. supermarket scanners.
Class 3a/3b	Medium power / direct viewing hazard / little diffuse reflection hazard.
Class 3a	Visible lasers with 1-5 mW power output, invisible lasers, and those having 1-5 times the Accessible Emission Limit (AEL) of class 1 lasers.
Class 3b	All other class 3 lasers at all wavelengths which have a power output less than 500 mW.
Class 4	High power / eye & skin hazard / potential diffuse reflection hazard or fire hazard

There are several pertinent instructions and guidelines regarding laser use. They are:

- OPNAVINST 5100.27B Navy Laser Hazards Control Program
  (which replaced OPNAVINST 5100.27A and SPAWARINST 5100.12B)
- MIL-HDBK-828B, Range Laser Safety
- ANSI Z136.1-2007, American National Standard for the Safe Use of Lasers (Parent)

Every Navy command which uses lasers must have a Laser System Safety Officer (LSSO). All LSSOs must attend a Navy LSSO course.

There are four categories of LSSOs.

- Administrative Laser Safety Officer (ALSO)
- Technical Laser Safety Officer (TLSO)
- Laser Safety Specialist (LSS)
- Range Laser Safety Specialist (RLSS)

See OPNAVINST 5100.27B for details of their qualifications and responsibilities.

The hazard ranges of interest are the NOHD for direct viewing of a beam and the $r_{1(safe)}$ or $r_{2(safe)}$ for viewing a beam reflected off an object such as a wall. These are depicted in Figure 1. The Maximum Permissible Exposure (MPE) values present laser safety levels as a function of exposure time, laser PRF, pulse duration, and wavelength. Different tables are used for eye safety while directly viewing a beam, for viewing a diffusely reflected beam, and for skin exposure.

For repeated pulses the following equation is used to calculate the maximum permissible exposure (MPE).

$$\text{MPE (repeated pulse)} = \frac{MPE(single\ pulse)}{(\ PRF\ x\ t_e\ )^{1/4}} \qquad \textbf{[1]}$$

Where PRF is the pulse repetition frequency of the laser and t_e is the exposure duration.

For visible lasers t_e is usually taken as 1/4 second and for non-visible lasers a value of 10 seconds is used.

Figure 1 depicts some laser hazard distances.

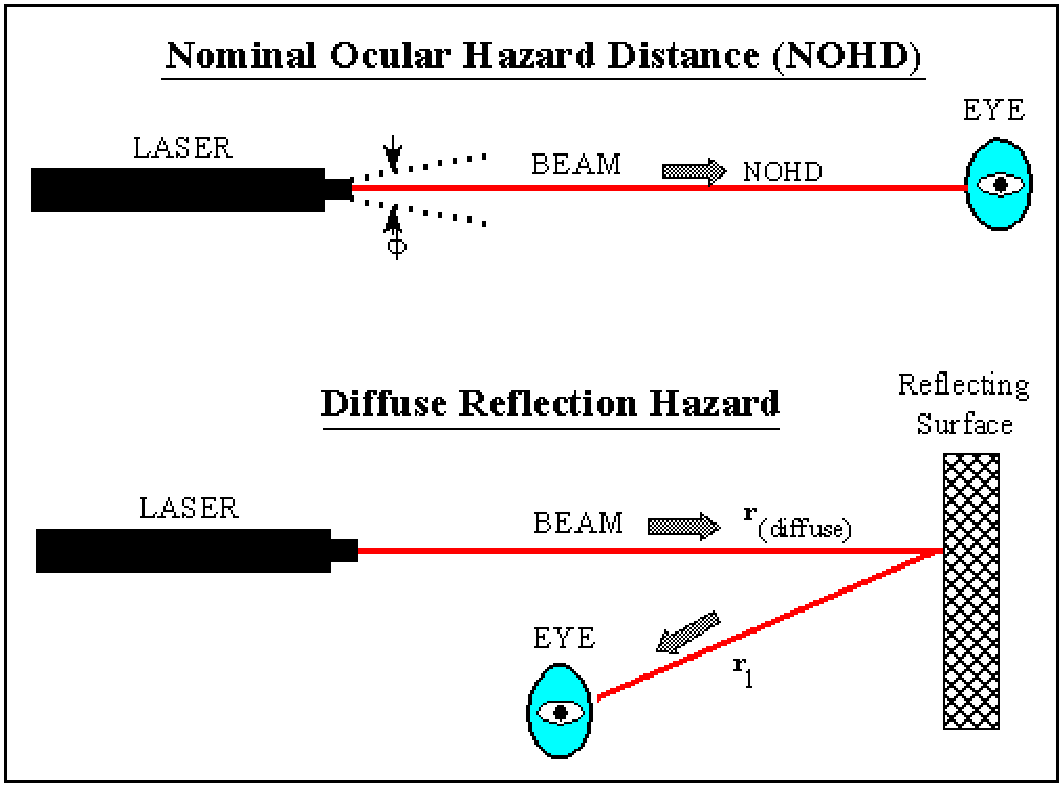
Nominal Ocular Hazard Distance (NOHD)
LASER
BEAM
NOHD
EYE
Φ
Diffuse Reflection Hazard
Reflecting Surface
LASER
BEAM
$r_{(diffuse)}$
EYE
r_1

Figure 1. Laser Hazard Distances.

Range laser safety specialists shall be designated for external operations. Range test plans shall specify:

- Permissible aircraft flight paths, and ship or vehicle headings.
- Hazard areas to be cleared.
- Operational personnel locations.
- Types of surveillance to be used to ensure a clear range.
- Radio / communications procedures.

During laser operations no portion of the laser beam may extend beyond the controlled target area unless adequate surveillance can prevent radiation of unprotected areas. Class 3 and class 4 lasers shall not be directed above the horizon unless coordinated with those responsible for the given airspace (FAA, Navy, Air Force, etc).

In an industrial environment, warning and hazard signs and lights will be posted, a hazard zone shall be designated when lasers are in operation, and training shall be provided to operators in the proper eye and body (skin) protection required. Interlocks to laser operation shall be provided when there is the possibility of unauthorized personnel entering the hazard area.

Fiber optic cables usually have laser power sources so appropriate warnings or labels need to be applied to connections or possible breakage points.

This page intentionally left blank.

AIRCRAFT DYNAMICS CONSIDERATIONS

Free Fall / Aircraft Drag.......8-1
Mach Number and Airspeed vs. Altitude.......8-2
Maneuverability 8-3
EMP / Aircraft Dimensions.......8-4

This page intentionally left blank.

FREE FALL / AIRCRAFT DRAG

The purpose of this section is to get an awareness of the distance traveled by a flare or other object such as a bomb, which is jettisoned or dropped by an aircraft. This will give the reader an appreciation for the significance of aircraft tactical altitude.

From Newton's second law of motion:

$F = m_o a$ where: F = Force
m_o = Mass of object
a = Acceleration

and the law of gravitation: $$F = G\frac{m_o m_e}{r^2}$$

		English Units	SI Units
where:	F = Force of attraction	lb_f	Newton
	G = universal gravitational constant	$3.44\text{x}10^{-8}$ $ft^4/lb\text{-}sec^4$	$6.67\text{x}10^{-11}$ $m^3/kg\text{-}sec^2$
	m_o, m_e= Masses (not weight) of object & earth	slug	kg
	r = distance between center of gravity of objects	feet	meter

Combining the two equations and solving for "a":

$$a = \frac{G\,m_e}{r^2} = g$$, the familiar constant acceleration due to gravity.

Since G and m_e are fixed and the variation in r (the distance from the earth's center) is small except for satellites, "g" is considered fixed at 32.2 ft/sec^2.

For objects with a constant acceleration (g), it can be shown that:

$$d = v_i t + \frac{1}{2}gt^2$$ where

d = distance traveled
v_i = Initial velocity
t = time
g = acceleration

For a falling object, Figure 1 on the following page may be used to estimate time/distance values.

- The upper curve is for an object shot upward with an initial velocity of 50 ft/sec.
- The middle curve is for an object shot horizontally with an initial velocity of 50 ft/sec or one that is a free-falling object dropped with no initial vertical velocity.
- The lower curve is for an object with a downward initial velocity of 50 ft/sec.

Notes:

1) 50 ft/sec is the typical cartridge ejection velocity of a flare/chaff expendable.
2) The top curve actually goes up 39 feet before starting back down, but this is difficult to see due to the graph scale.
3) This simplification ignores the effects of air drag or tumbling effects on a falling object which will result in a maximum terminal velocity, with resultant curve straightening.

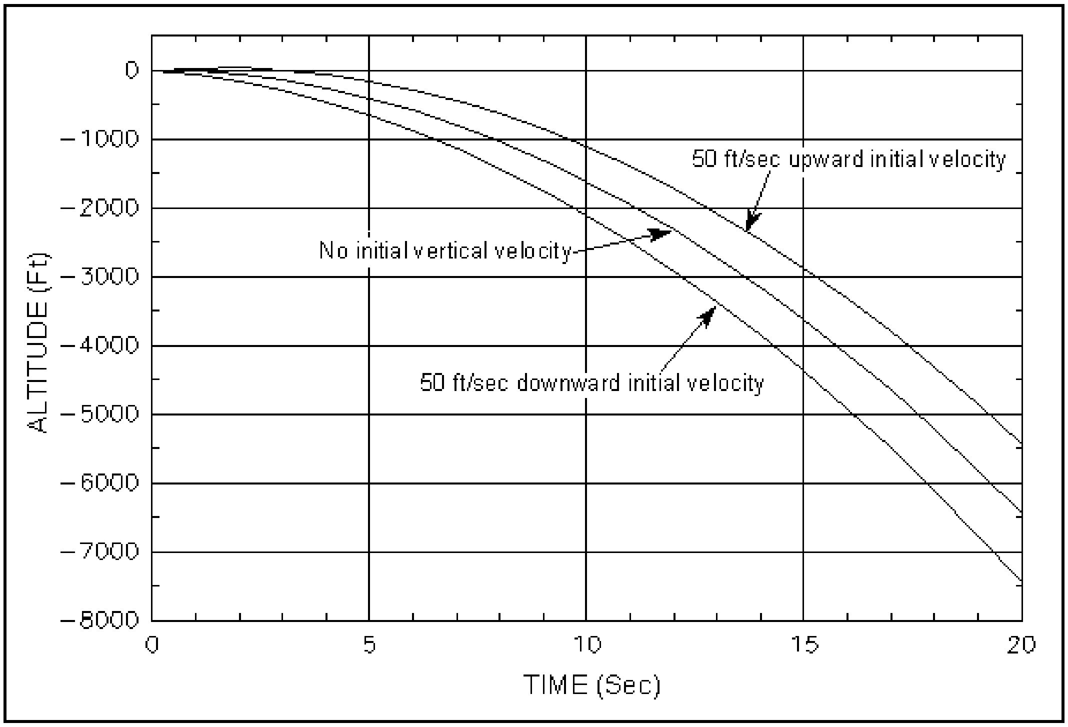

Figure 1. Object Fall Rate.

SAMPLE CALCULATIONS

Let us assume that we want to know how far a bomb or other object has fallen after 13 seconds if it had been dropped from an aircraft traveling at 450 kts which was in a 40° dive.

Our initial vertical velocity is: 450 kts (Sin 40°) (1.69 ft/sec per knot) = 489 ft/sec downward

$d = V_i t + \frac{1}{2}gt^2 = -489(13) + \frac{1}{2}(-32.2)(13)^2 = -6355 - 2721 = -9{,}076$ ft.

Remember to keep the signs (+/-) of your calculations in agreement with whatever convention you are using. Gravity pulls downward, so we used a minus sign for acceleration. Also the initial velocity was downward.

In reality, any object may well have reached terminal velocity before the time indicated using the above formula or Figure 1. In this example, the actual distance determined from ballistics tables would have been 8,000 ft, which is about 13% less than the above calculation would indicate. The drag characteristics of the object determine how much shorter the distance will be. In any case, it will not have dropped farther.

AIRCRAFT DRAG INDEX POINTS

Tactical aircraft carry stores in various combinations depending upon the mission. Each store has a different drag load which affects range. The pilot needs to know the total drag load in order to determine his aircraft range on a particular mission. Adding up the total drag in pounds of force for wind resistance would be cumbersome. Therefore, the drag of the stores is compared to a known reference drag (usually the aircraft), and expressed as a percentage of aircraft drag multiplied by some constant. This ratio is variously called drag count, drag index, or drag points. For instance, if a missile has 100 pounds of drag and the reference aircraft drag is 50,000 pounds, the ratio is 100/50,000 = 0.002. Multiply this by a constant of 100 (for example) and the drag index point is 0.2. The pilot only needs to look on a chart to see what the drag index points are for his stores, add up the drag points, and look on a chart to see what his aircraft range and best range (or endurance) speed will be.

This page intentionally left blank.

MACH NUMBER and AIRSPEED vs. ALTITUDE

MACH NUMBER is defined as a speed ratio, referenced to the speed of sound, i.e.

$$MACH\ NUMBER = \frac{Velocity\ of\ Interest}{Velocity\ of\ Sound} \quad (at\ the\ given\ atmospheric\ conditions) \qquad \textbf{[1]}$$

Since the temperature and density of air decreases with altitude, so does the speed of sound, hence a given true velocity results in a higher MACH number at higher altitudes.

AIRSPEED is a term that can be easily confused. The unqualified term airspeed can mean any of the following:

a. Indicated airspeed (IAS) - the airspeed shown by an airspeed indicator in an aircraft. Indicated airspeed is expressed in knots and is abbreviated KIAS.

b. Calibrated airspeed (CAS) - indicated airspeed corrected for static source error due to location of pickup sensor on aircraft. Calibrated airspeed is expressed in knots and is abbreviated KCAS. Normally it doesn't differ much from IAS.

c. True airspeed (TAS) - IAS corrected for instrument installation error, compressibility error, and errors due to variations from standard air density. TAS is expressed in knots and is abbreviated KTAS. TAS is approximately equal to CAS at sea level but increases relative to CAS as altitude increases. At 35,000 ft, 250 KIAS (or KCAS) is approximately 430 KTAS.

IAS (or CAS) is important in that aircraft dynamics (such as stall speed) responds largely to this quantity. TAS is important for use in navigation (True airspeed ± wind speed = ground speed).

Figures 1 and 2 depict relations between CAS and TAS for various altitudes and non-standard temperature conditions. The first graph depicts lower speed conditions, the second depicts higher speeds.

As an example of use, consider the chart on the next page. Assume we are in the cockpit, have read our IAS from the airspeed indicator, and have applied the aircraft specific airspeed correction to obtain 370 KCAS. We start at point "A" and go horizontally to our flight altitude at point "B" (25,000 ft in this case). To find our Mach, we go down vertically to point "C" to obtain 0.86 Mach. To get our TAS at our actual environmental conditions, we go from point "B" vertically until we hit the Sea Level (S.L.) reference line at point "D", then travel horizontally until we reach our actual outside air temperature (-20°C at altitude) at point "E", then go up vertically to read our actual TAS from the scale at point "F" (535 KTAS). If we wanted our TAS at "standard" temperature and pressure conditions, we would follow the dashed lines slanting upward from point "B" to point "G" and read 515 KTAS from the scale. Naturally, we could go into the graph at any point and go "backwards" to find CAS from true Mach or TAS.

Figure 3 shows a much wider range of Mach numbers. It contains only TAS and Mach, since aircraft generally do not fly above Mach 2, but missiles (which don't have airspeed indicators) do. The data on this graph can be obtained directly from the following formula for use at altitudes of 36,000 ft and below:

$$Speed\ of\ Sound\ (KTAS) = 29.06\sqrt{518.7 - 3.57\,A} \qquad Where\ A = altitude\,(K\ ft) \qquad \textbf{[2]}$$

The speed of sound calculated from this formula can be used with the equation on the first page to obtain Mach number. This equation uses the standard sea level temperature of 59° F and a lapse rate of -3.57°/1000 ft altitude. Temperature stabilizes at -69.7° F at 36,000 ft so the speed of sound stabilizes there at 573 knots. See the last page of this section for a derivation of equation [2].

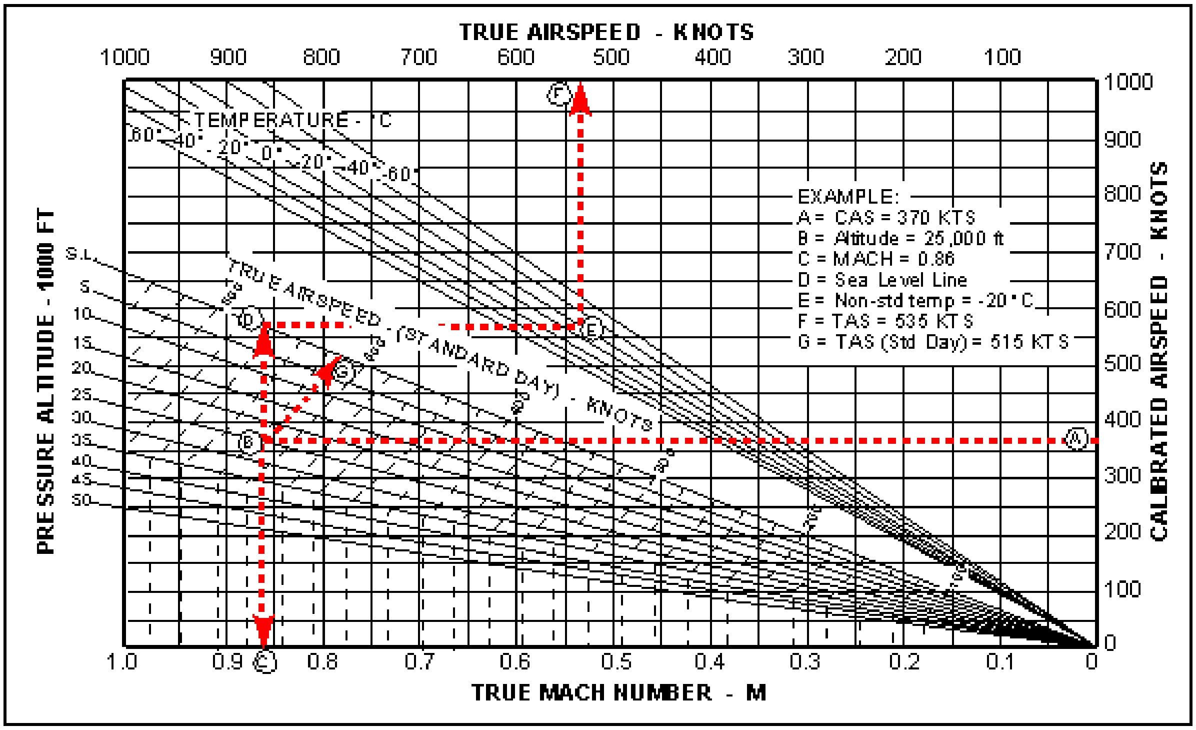

Figure 1. TAS and CAS Relationship With Varying Altitude and Temperature.

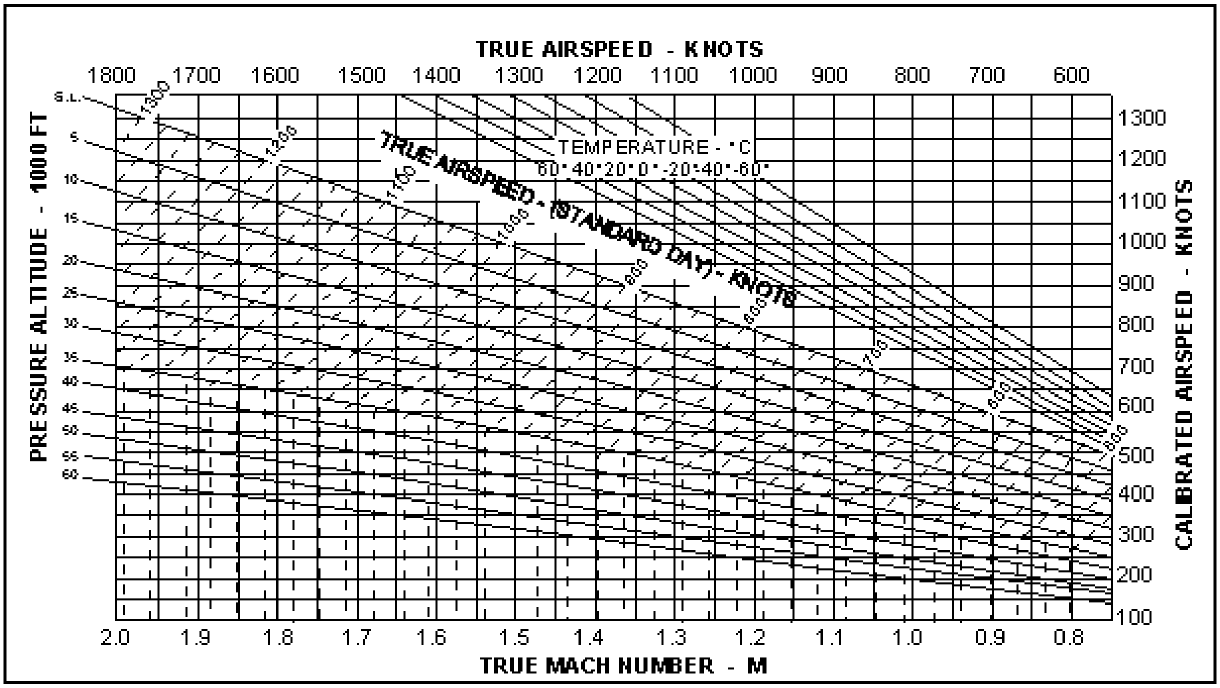

Figure 2. TAS and CAS Relationship With Varying Altitude and Temperature (Continued).

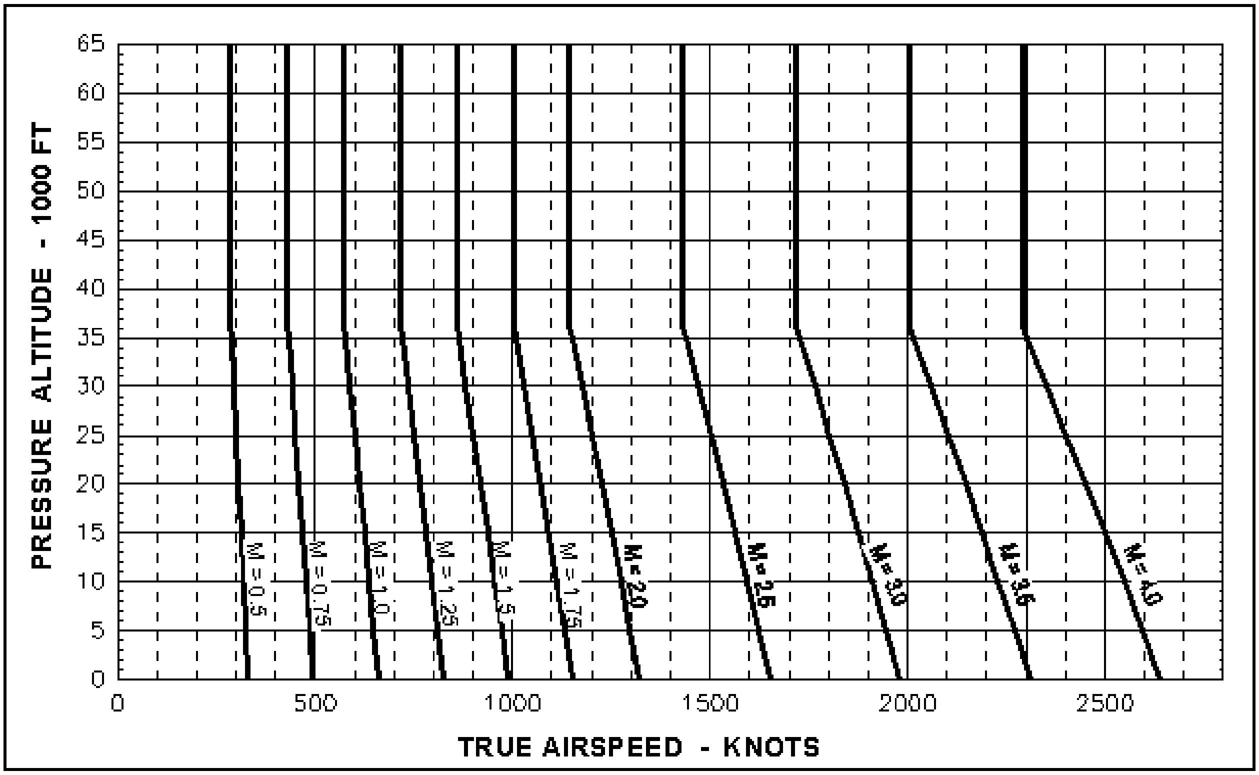

Figure 3. Mach Number vs. TAS Variation With Altitude.

The following is a derivation of equation [2] for the speed of sound:

Given: p = pressure (lb/ft^2)
v = specific volume (ft^3/lb)
R = a constant (for air: R = 53.3)
T = absolute temperature (°Rankine) = °F + 459.7
w = specific weight (lb/ft^3) = $1/v$
ρ = density = $w/g = 1/gv \;\therefore\; v = 1/g\rho$

From Boyle's law of gasses: $pv = RT$, therefore we have: $p/\rho = gRT = (32.2)(53.3)T = 1718\ T$ **[3]**

It can also be shown that: p/ρ^{γ} = constant; for air $\gamma = 1.4$ **[4]**

From the continuity equation applied to a sound wave: $\rho AV_a = (\rho + dp)A(V_a + dV_a)$ **[5]**

Expanding and dropping insignificant terms gives: $dV_a = -V_a\, d\rho/\rho$ **[6]**

Using Newton's second law ($p + \rho V_a/2$ = a constant) and taking derivatives: $dp = -\rho V_a dV_a$

Substituting into [6] gives: $V_a^2 = dp/d\rho$ **[7]**

Then taking derivatives of [4] and substituting in [7] gives: $V_a = \sqrt{\frac{\gamma p}{\rho}}$ **[8]**

Then using [3] gives: $V_a = \sqrt{\gamma gRT} = \sqrt{1.4(1718)T} = 49\sqrt{T}$ **[9]**

Using a "Standard" atmosphere of 59° F @ Sea Level (S.L.) and a lapse rate of -3.57°/1000 ft altitude:

$$V_a = 49\sqrt{459.7 + 59 - 3.57\,A}\,\frac{ft}{\text{sec}}\left[\frac{3600\,\text{sec}}{hr}\,\frac{nm}{6076\;ft}\right] = 29.06\sqrt{518.7 - 3.57A} \quad \text{which is equation [2]}$$

MANEUVERABILITY

A useful function is to determine how many "G's" an aircraft might require to make a given turn without altitude loss. From Newton's laws, F cos φ = W, where: F = force applied to an aircraft, W = weight, and φ = bank angle. By definition "G's" is the ratio of the force on an object to its weight, i.e., G = F/W = 1/cos φ.

Simple calculations will show the results presented in Table 1, to the right.

Given that the average structural limit of an aircraft is about 7 G's, the maximum bank angle that can be achieved in level (non-descending) flight is 81.8°.

Table 1. G vs. Angle of Bank.
(No altitude loss)

G	φ
1.0	0
1.4	45
2.0	60
3.9	75
7.2	82
11.5	85

Figure 1 can be used to determine the turn radius and rate-of-turn for any aircraft, given speed and angle of bank (assuming the aircraft maintains level flight). It may also be used in the reverse context. It should be noted that not all aircraft can fly at the speeds depicted - they may stall beforehand or may be incapable of attaining such speeds due to power/structural limitations.

In the example shown on Figure 1, we assume an aircraft is traveling at 300 kts, and decides to make a 30° angle of bank turn. We wonder what his turn radius is so we can approximate his flight path over the ground, and what his rate of turn will be. We enter the chart at the side at 300 kts and follow the line horizontally until we intercept the 30° "bank angle for rate of turn" line. We then go down vertically to determine the 2.10°/sec rate of turn. To get radius, we continue horizontally to the 30° "bank angle for turn radius" line. We can then go down vertically to determine the radius of 13,800 ft.

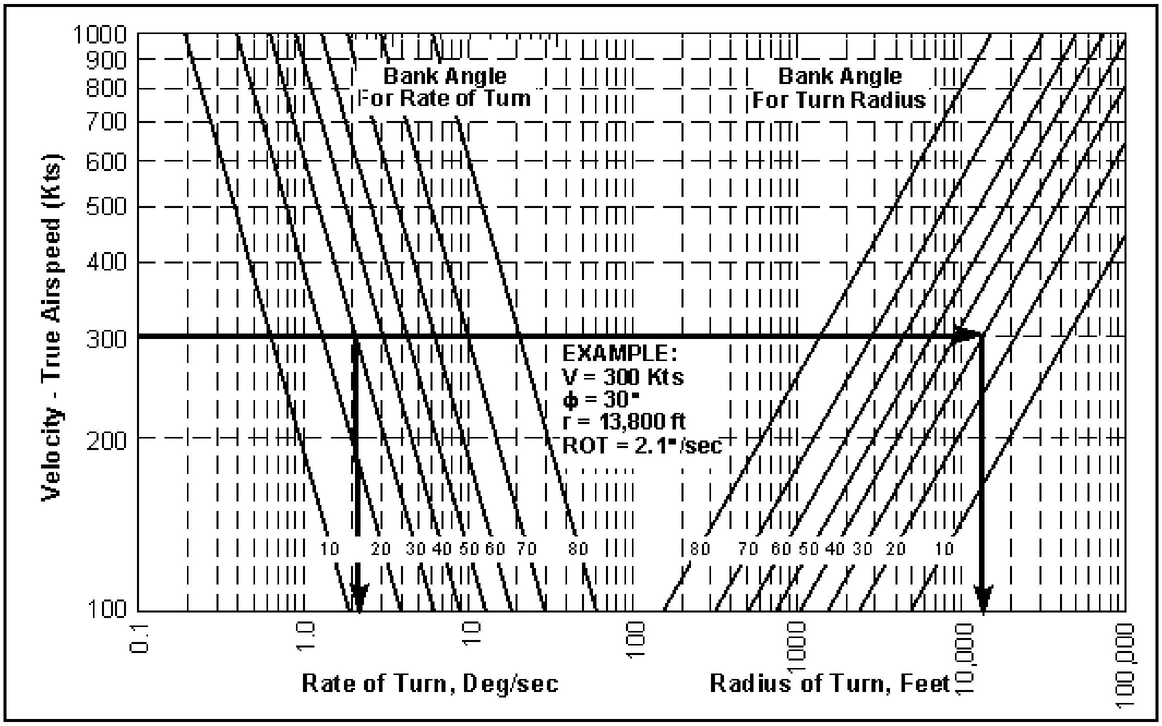

Figure 1. Aircraft Turn Rate / Radius vs. Speed.

The exact formulas to use are:

$$Rate\ of\ Turn = \frac{1091\tan(\phi)}{V} \qquad Radius\ of\ Turn = \frac{V^2}{11.26\tan(\phi)} \qquad Where: \begin{array}{c} V = Velocity\ (Knots) \\ and\ \phi = Angle\ of\ Bank \end{array}$$

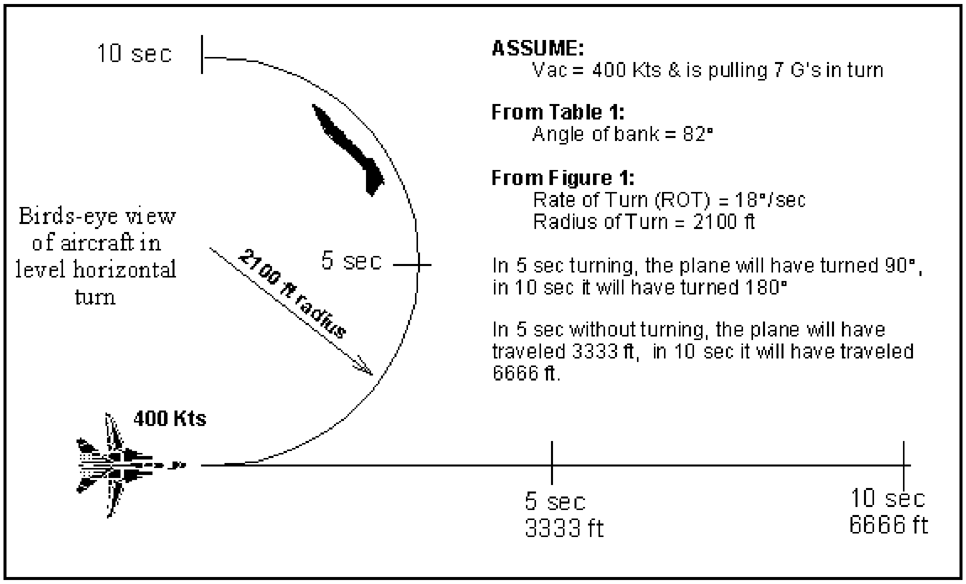
10 sec
Birds-eye view of aircraft in level horizontal turn
5 sec
2100 ft radius
400 Kts
5 sec
3333 ft
10 sec
6666 ft
ASSUME:
Vac = 400 Kts & is pulling 7 G's in turn
From Table 1:
Angle of bank = 82°
From Figure 1:
Rate of Turn (ROT) = 18°/sec
Radius of Turn = 2100 ft
In 5 sec turning, the plane will have turned 90°, in 10 sec it will have turned 180°
In 5 sec without turning, the plane will have traveled 3333 ft, in 10 sec it will have traveled 6666 ft.

Figure 2. Maneuvering Aircraft.

Another interesting piece of information might be to determine the distance a typical aircraft might travel during a maneuver to avoid a missile.

Figure 2 shows a birds-eye view of such a typical aircraft in a level (constant altitude) turn.

To counter many air-to-air missiles the pilot might make a level turn, however in countering a SAM, altitude is usually lost for two reasons: (1) the direction of maneuvering against the missile may be downward, and (2) many aircraft are unable to maintain altitude without also losing speed. These aircraft may have insufficient thrust for their given weight or may be at too high an altitude. The lighter an aircraft is (after dropping bombs/burning fuel), the better the performance. Likewise, the higher the altitude, the poorer the thrust-to-weight ratio. Maximum afterburner is frequently required to maintain altitude at maximum "G" level.

REFERENCE AXES (Roll, Pitch, Yaw)

The rotational or oscillating movement of an aircraft, missile, or other object about a longitudinal axis is called roll, about a lateral axis is called pitch, and about a vertical axis is called yaw as shown in Figure 3.

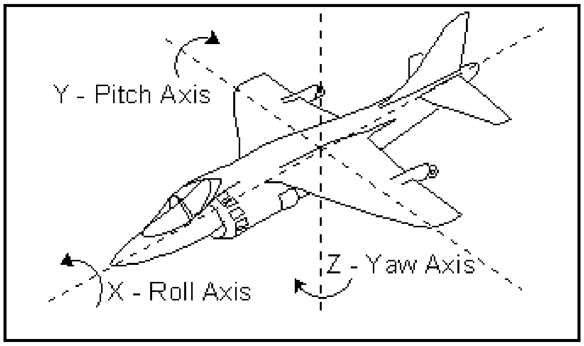
Y - Pitch Axis
X - Roll Axis
Z - Yaw Axis

Figure 3. Reference Axes.

SAMPLE CALCULATIONS

If we want to determine the rate of turn or turn radius more precisely than can be interpolated from the chart in Figure 1, we use the formulas. For our initial sample problem with an aircraft traveling 300 kts, in a 30° angle of bank turn, we have:

$$Rate\ of\ Turn = \frac{1091\tan(\phi)}{V} = \frac{1091\tan(30)}{300} = 2.1°/\sec$$

$$Radius\ of\ Turn = \frac{V^2}{11.26\tan(\phi)} = \frac{300^2}{11.26\tan(30)} = 13{,}844\ ft$$

These are the same results as we determined using Figure 1.

This page intentionally left blank.

EMP / AIRCRAFT DIMENSIONS

An aircraft flying in the vicinity of an electromagnetic pulse (EMP) acts like a receiving antenna and picks up EMP radiation in relation to size like a dipole (or half-wavelength dipole). The electromagnetic pulse spectrum decreases above 1 MHz as shown in Figure 1, so an F-14 aircraft that is an optimum ½ wavelength antenna at ≈8 MHz will pick up less EMP voltage than a B-52 or an aircraft with a trailing wire antenna. A rule of thumb for the voltage picked up is:

V_{EMP} = 8.1 volts/ft times the maximum dimension of the aircraft in feet

This rule of thumb was generated because a single linear relationship between voltage and aperture seemed to exist and compared favorably with more complex calculations for voltage picked up by various aircraft when subjected to EMP.

Table 1 shows various aircraft and the frequencies they would be most susceptible to, using $f = c/\lambda$, where λ matches the selected aircraft dimension for maximum "antenna reception effect." This should be a design consideration when trying to screen onboard avionics from the effects of EMP.

The following is a partial listing of aircraft types vs. identifying prefix letters (several are used in Table 1):

A	Attack	K	Tanker	T	Trainer
B	Bomber	O	Observation	U	Utility
C	Cargo	P	Patrol	V	Vertical or Short Takeoff and Landing (V/STOL)
E	Electronic Surveillance	Q	Special mission	X	Experimental
F	Fighter	R	Reconnaissance	Y	Prototype
H	Helicopter	S	Anti Sub/Ship		

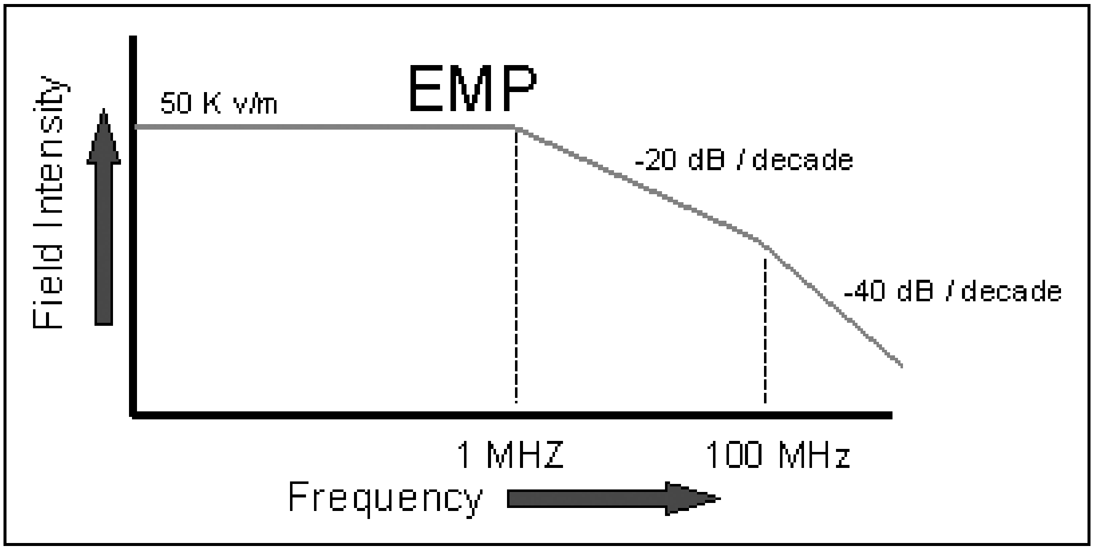

Figure 1. EMP as a Function of Frequency.

Table 1. Aircraft Dimensions and Equivalent Antenna Aperture.

Mission	Aircraft Type	Height (ft.)	Frequency (MHz)		Length (ft.)	Frequency (MHz)		Wing Span (ft.)	Frequency (MHz)	
		A	f	f/2	A	f	f/2	A	f	f/2
ATTACK	A-10	14.66	67.05	33.52	53.33	18.43	9.21	57.5	17.1	8.55
ELECTRONIC WARFARE	EA-6B	16.50	59.64	29.82	59.34	16.58	8.29	53.0	18.57	9.29
	EA-18	16.0	61.47	30.74	60.17	16.35	8.17	44.67	22.02	11.01
FIGHTER	F-15	18.4	53.42	26.71	63.75	15.42	7.71	42.8	22.97	11.48
	F-16	16.66	59.00	29.5	49.25	19.96	9.98	31.0	31.71	15.85
	F/A-18C/D	15.3	64.31	32.16	56.0	17.57	8.79	40.70	24.18	12.09
	F/A-18E/F	16	61.47	30.7	60.17	16.34	8.17	44.67	22.02	11.01
	F-22	17.75	55.41	27.7	64.17	15.33	7.66	43	22.87	11.43
	F-35A	14.2	69.27	34.63	51.4	19.1	9.57	35	28.1	14.05
	F-35C	15.5	63.45	31.73	51.4	19.1	9.57	43	22.8	11.43
	F-117	12.42	79.15	39.57	65.92	14.91	7.46	43.33	22.69	11.34
ASW	P-3C	33.75	29.16	14.58	116.42	8.45	4.23	99.67	9.87	4.94
	S-3A	22.75	43.25	21.63	54.34	18.45	9.23	68.67	14.33	7.17
	P-8	42.1	23.36	11.68	129.5	7.6	3.8	123.6	7.96	3.98
AEW	E-2C	18.4	53.48	26.74	56.50	17.42	8.71	80.58	12.21	6.11
V/STOL	AV-8B	11.64	84.45	42.23	46.3	21.23	10.62	30.3	32.44	16.22
	V-22	18.1	54.3	27.2	57.3	17.17	8.58	84.5	11.64	5.82
	F-35B	15.0	65.6	32.78	51.2	19.2	9.61	35	28.1	14.05
HELICOPTERS	AH-1W	14.16	69.46	34.73	58.25	16.89	8.44	48.0	20.49	10.24
	UH-1E	12.75	71.18	35.59	52.91	18.60	9.30	44.0	22.36	11.18
	U/SH-2	15.41	63.85	31.93	52.5	18.74	9.37	44.0	22.36	11.18
	SH-3D	16.42	59.93	29.97	72.67	13.54	6.77	62.00	15.87	7.84
	CH-46D	16.75	58.75	29.38	84.34	11.67	5.84	50.0	19.68	9.84
	CH-47	18.75	52.45	26.23	99.0	9.94	4.98	60.0	16.39	8.19
	CH-53A	24.91	39.50	19.75	88.16	11.16	5.58	72.25	13.62	6.81
	SH-60	17.0	57.85	28.93	64.83	15.17	7.59	53.83	18.27	9.14
	AH-64	15.25	64.5	32.25	58.25	16.89	8.44	48.0	20.49	10.24
TRANSPORT	C-2A	15.92	61.81	30.91	56.6	17.39	8.70	80.58	12.21	6.11
TANKERS	KC-130F	38.1	25.83	12.92	97.8	10.06	5.03	132.5	7.43	3.72
SPECIAL ELECTRONICS	EC-13OQ	38.5	25.56	12.78	99.34	9.91	4.96	132.58	7.42	3.71
TRAINER	T-2B	14.8	66.49	33.25	38.7	25.43	12.72	37.85	26.00	13.0
	T-39D	16.0	61.50	30.75	43.75	22.49	11.25	44.34	22.19	11.10
	T-45	13.5	72.86	36.43	39.33	25.0	12.5	30.8	31.93	15.97

DATA TRANSFER BUSSES

Data Busses 9-1
RS-232 Interface 9-2
RS-422 Balanced Voltage Interface 9-3
RS-485 Interface 9-4
IEEE-488 Interface Bus (HP-IB/GP-IB) 9-5
MIL-STD-1553 & 1773 Data Bus 9-6
Ethernet 9-7
Transmission Control Protocol / Internet Protocol 9-8

This page intentionally left blank.

DATA BUSSES

INTRODUCTION

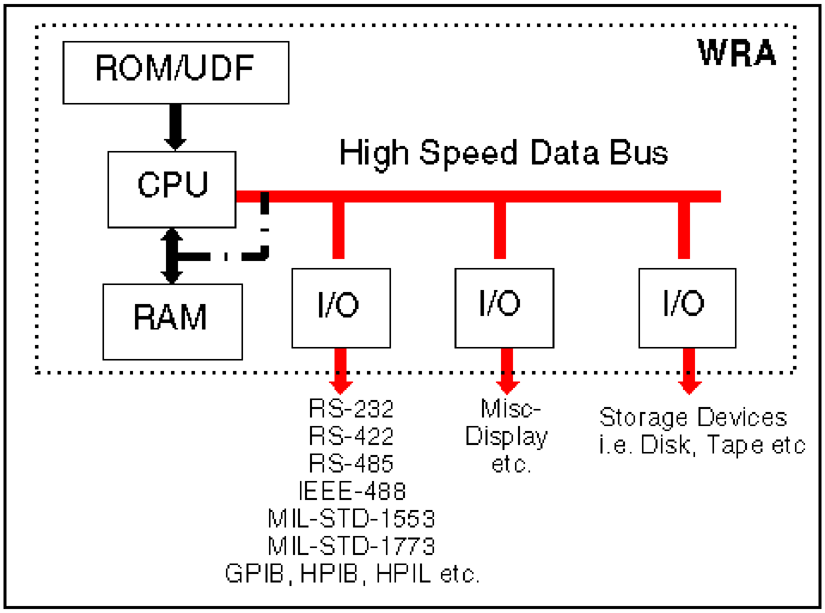

Figure 1. Avionics Block Diagram.

The avionics systems on aircraft frequently contain general purpose computer components which perform certain processing functions, then relay this information to other systems. Some common examples are the mission computers, the radar processors, RWRs, and jammers. Each system is frequently laid out as shown in Figure 1.

The Input/Output (I/O) modules will vary in function, but all serve the same purpose - to translate the electrical signals from one protocol to one of another in order to exchange information. I/O modules are used similarly in general purpose computers in laboratories to test equipment and/or tie computers together via a local area network (LAN) to exchange information. Some of the methodologies include a star, ring, or bus type network (see terminology at the end of this section).

A number of network "models" exist for describing the functions, interfaces and protocols involved in network data interchange. Regardless of the descriptive model used, all strive toward the same end and none actually changes the detailed implementation. Table 1 shows the layer names and the number of layers of networking models presented in Internet Engineering Task Forces (IETF) Request For Comments (RFCs) and in common use by textbooks.

Table 1. Some Common Network Architecture Models.

IETF-RFC 1122	Kurose, Forouzan	Comer, Kozierok	Stallings	Cisco Academy
Four layer Internet model	*Five-layer Internet model or TCP/IP protocol suite*	*TCP/IP 5-layer reference model*	*Five layer TCP/IP model*	*Four layer Internet model*
Application	Application	Application	Application	Application
Transport	Transport	Transport	Host-to-host or transport	Transport
Internet	Network	Internet	Internet	Internetwork
Link	Data link	Data link (Network interface)	Network access	Network interface
	Physical	(Hardware)	Physical	

These textbooks are secondary sources that may be contrary to the intent of RFC 1122 and other primary sources such as the Open Systems Interconnection (OSI) Reference Model developed by the International Organization for Standardization (ISO).

The OSI Reference Model is a more general description for layered communications and computer network protocol design. The IETF makes no effort to follow the OSI model although RFCs sometimes refer to it. The description of the OSI layers is shown in Table 2.

Table 2. OSI Reference Model.

	Name	Use	Misc Example	IP Suitte
7	APPLICATION	Meaning of data	HL7, Modbus	DHCP, DNS, FTP, Gopher, HTTP, NTP, SMTP, SNMP, Telnet
6	PRESENTATION	Building blocks of data and encryption	ASCII, EBCDIC, MIDI	MIME, XDR, SSL
5	SESSION	Opening and closing of specific communication paths	NetBIOS, SAP, Half Duplex, Simplex, SDP	NetBIOS, RTP, SAP
4	TRANSPORT	Error checking	NBF	PPTP, SCTP, TCP, UDP
3	NETWORK	Determination of data paths within the network	NBF, Q.931, IS-IS	IP, ICMP
2	DATA LINK	Data transmission, source, destination, and checksum	802.3 (Ethernet), PPP 802.11a/b/g/n MAC, FDDI	ARP, PPP, PPTP, SLIP
1	PHYSICAL	Voltage levels, signal connections, wire, or fiber	RS-232, 10Base-T, 802.11a/b/g/n Physical	

A layer is a collection of conceptually similar functions that provide services to the layer above it and receives service from the layer below it. On each layer an instance provides services to the instances at the layer above and requests service from the layer below. For example, a layer that provides error-free communications across a network provides the path needed by applications above it, while it calls the next lower layer to send and receive packets that make up the contents of the path. Conceptually two instances at one layer are connected by a horizontal protocol connection on that layer.

Most networks do not use all layers. For example, RS-232 is only a physical layer. Ethernet is only layers 1 and 2. TCP/IP is a protocol, not a network, and uses layers 3 and 4 regardless of whether layers 1 and 2 are a telephone line, wireless connection, or 10Base-T Ethernet cable.

Most of the sections in this division discuss the lowest (physical) layer of communication. There are, however, several more areas of general interest which are included in later sections such as Ethernet and TCP/IP. These are used in general purpose computers like the desktop PC or lab networks, and are not commonly used in aircraft.

The typical high-speed data busses on avionics/computers do not operate as fast as the CPU clock speed, but they are much faster than the interface busses they connect to. There are a number of interface busses (physical layer in network model) which are widely used by aircraft, avionics systems and test equipment. The most common include the RS-232, the RS-422, the RS-485, the IEEE-488 (GP-IB/HP-IB) and the MIL-STD-1553A/B. The MIL-STD-1773 bus is a fiber optic implementation of the 1553 bus and may be used in the future when technology requires it to reduce susceptibility to emissions or other reasons.

A summary of these more common types follows in Table 3, which includes a brief descriptive comparison, while a section covering each in more detail is provided later.

Table 3. Summary of Physical Bus Characteristics.

Bus	**Max Length**	**Max Number of Terminals(1)**	**Type**	**# of Lines(2)**	**Data Rate**	**Rise Time(3)**	**Data Format**
RS-232C	100 feet max 50 ft at 20k bps	1	Serial	3-20	150 - 19,200 baud per sec		5- to 8- bit serial
RS-422	1.2 km(4)	10(5)	Serial	3	see figure in RS-232 section	<0.1 T_b	unspecified
RS-485	unspecified	32	Serial	3	10 MHz	<0.3 T_b	unspecified
IEEE-488 (GP-IB/HP-IB)	20 meters	14	Parallel	16	500 kHz(6)		8-bit parallel
HP-IL	100 meters	30	Serial	2	20 k BPS		serial
MIL-STD-1553B MIL-STD-1773	300 feet N/A	32(7)	Serial	3	1 MHz	100-300 ns	20-bit serial

NOTES FROM TABLE:

(1) Max Number of Terminals does not include the bus controller.

(2) Including ground/shield

(3) T_b = time duration of the unit interval at the applicable data signaling rate (pulse width)

(4) Length is function of data signaling rate influenced by the tolerable signal distortion, amount of longitudinally coupled noise and ground potential difference introduced between the controller and terminal circuit grounds as well as by cable balance. See RS-422 section for graph.

(5) Physical arrangement of multiple receivers involves consideration of stub line lengths, fail-safe networks, location of termination resistors, data rate, grounding, etc.

(6) Rate can go up to 1 MHz if special conventions are followed.

(7) Max Number of Terminals includes terminal reserved for broadcast commands.

BUS TERMINOLOGY

10BASE-T: Standard “Plain Vanilla” Ethernet based on Unshielded Twisted Pair wire

10BASE-F: 10Mbps fiber optic Ethernet

100BASE-T: Standard “Fast Ethernet” based on twisted pair copper wire

ADDRESS: A unique designation for the location of data or the identity of an intelligent device; allows each device on a single communications line to respond to its own message.

ASCII (American Standard Code for Information Interchange): Pronounced asky. A seven-bit-plus-parity code established by ANSI to achieve compatibility between data services.

ASYNCHRONOUS OPERATION: Asynchronous operation is the use of an independent clock source in each terminal for message transmission. Decoding is achieved in receiving terminals using clock information derived from the message.

BAUD: Unit of signaling speed. The speed in baud is the number of discrete events per second. If each event represents one bit condition, baud rate equals bits per second (BPS). When each event represents more than one bit, baud rate does not equal BPS.

BIT: Contraction of binary digit: may be either zero or one. A binary digit is equal to one binary decision or the designation of one or two possible values of states of anything used to store or convey information.

BIT RATE: The number of bits transmitted per second.

BRIDGE: A network bridge connects multiple network segments at the data link layer (layer 2) of the OSI model, and the term layer 2 switch is very often used interchangeably with bridge. Bridges are similar to repeaters or network hubs, devices that connect network segments at the physical layer; however, with bridging, traffic from one network is managed rather than simply rebroadcast to adjacent network segments.

BROADCAST: Operation of a data bus system such that information transmitted by the bus controller or a remote terminal is addressed to more than one of the remote terminals connected to the data bus.

BUS CONTROLLER: The terminal assigned the task of initiating information transfers on the data bus.

BUS MONITOR: The terminal assigned the task of receiving bus traffic and extracting selected information to be used at a later time.

BYTE: A binary element string functioning as a unit, usually shorter than a computer “word.” Eight-bits per byte are most common. Also called a “character.”

COMMAND/RESPONSE: Operation of a data bus system such that remote terminals receive and transmit data only when commanded to do so by the bus controller.

CRC: Cyclic Redundancy Check; a basic error-checking mechanism for link-level data transmissions; a characteristic link-level feature of (typically) bit-oriented data communications protocols. The data integrity of a received frame or packet is checked by an algorithm based on the content of the frame and then matched with the result that is performed by a sender and included in a (most often, 16-bit) field appended to the frame.

CROSSOVER CABLE: Cable with transmit/receive pairs reversed so one computer or hub or switch can link directly to another.

DATA BUS: Whenever a data bus or bus is referred to in MIL-STD-1553B, it shall imply all the hardware including twisted shielded pair cables, isolation resistors, transformers, etc., required to provide a single data path between the bus controller and all the associated remote terminals.

DCE (Data Communications Equipment): Devices that provide the functions required to establish, maintain, and terminate a data-transmission connection; e.g., a modem.

DHCP: Dynamic Host Configuration Protocol - permits auto-assignment of temporary IP addresses for new devices logging in

DNS: Domain Name Server - associates names with IP addresses

DTE (Data Terminal Equipment): Devices acting as data source, data sink, or both.

DUPLEX: Communication traveling between two nodes in both directions

DYNAMIC BUS CONTROL: The operation of a data bus system in which designated terminals are offered control of the data bus.

EIA (Electronic Industries Association): A standards organization in the U.S.A. specializing in the electrical and functional characteristics of interface equipment.

FDM (Frequency-Division Multiplexer: A device that divides the available transmission frequency range into narrower banks, each of which is used for a separate channel.

FDX (Full Duplex): Simultaneous, two-way, independent transmission in both directions (4-wire).

FTP: File Transfer Protocol - the most popular mechanism for bulk movement of files on TCP/IP.

GATEWAY: Device which links Ethernet to dissimilar networks and transfers data at the application layer level. Interface cards link the PC to Ethernet via the PCI, ISA, PCMCIA, PC/104, or other buses.

GPIB: General Purpose Interface Bus (see section 9-5),

HALF DUPLEX: Operation of a data transfer system in either direction over a single line, but not in both directions on that line simultaneously.

HANDSHAKING: Exchange of predetermined signals between two devices establishing a connection. Usually part of a communications protocol.

HPIB / HPIL: Hewlett-Packard Interface Bus / Hewlett-Packard Interface Loop

HUB: The simplest method of redistributing data, are "dumb," not interpreting or sorting messages that pass through them. A hub can be as simple as an electrical buffer with simple noise filtering. It isolates the impedances of multiple spokes in a star topology. Some hubs also have limited store-and-forward capability. They indiscriminately transmit data to all other devices, which are still on the same collision domain, connected to the hubs. They are not assigned MAC addresses or IP addresses.

IEEE (Institute of Electrical and Electronic Engineers): An international professional society that issues its own standards and is a member of ANSI and ISO.

IP: Internet Protocol portion of TCP/IP. It is a protocol used for communicating data across a packet-switched network.

IP ADDRESS: Address of a TCP/IP enabled device on an Intranet or Internet – in the form xxx.xxx.xxx.xxx, where xxx is an integer between 0 and 255.

LAN: Local Area Network

MAC: Media Access Control - the physical components which disassemble Ethernet message fames

MANCHESTER ENCODING: Digital encoding technique (specified for the IEEE 802.3 Ethernet baseband network standard) in which each bit period is divided into two complementary halves; a negative-to-positive (voltage) transition in the middle of the bit period designates a binary “1,” while a positive-to-negative transition represents a “0.” The encoding technique also allows the receiving device to recover the transmitted clock from the incoming data stream (self-clocking).

MESSAGE: A single message is the transmission of a command word, status word, and data words if they are specified. For the case of a remote terminal to remote terminal (RT to RT) transmission, the message shall include the two command words, the two status words, and data words.

MODE CODE: A means by which the bus controller can communicate with the multiplex bus related hardware, in order to assist in the management of information flow.

MODEM (Modulator-Demodulator): A device used to convert serial digital data from a transmitting terminal to a signal suitable for transmission over a telephone channel, or to reconvert the transmitted signal to serial digital data for acceptance by a receiving terminal.

MULTIPLEXOR (also Multiplexer): A device used for division of a transmission into two or more subchannels, either by splitting the frequency band into narrower bands (frequency division) or by allotting a common channel to several different transmitting devices one at a time (time division).

NETWORK: An interconnected group of nodes; a series of points, nodes, or stations connected by communications channels; the assembly of equipment through which connections are made between data stations.

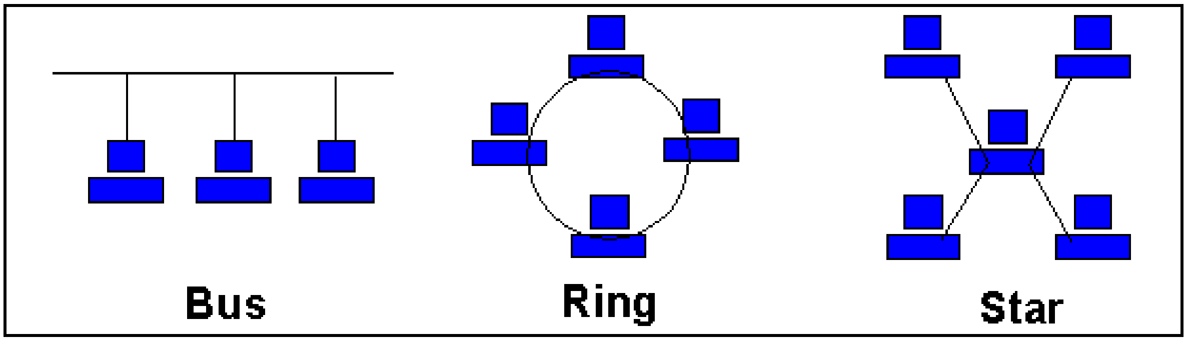

NODE: A point of interconnection to a network. Normally, a point at which a number of terminals or tail circuits attach to the network.

PARALLEL TRANSMISSION: Transmission mode that sends a number of bits simultaneously over separate lines (e.g., eight bits over eight lines) to a printer. Usually unidirectional.

PHASE MODULATION: One of three ways of modifying a sine wave signal to make it "carry" information. The sine wave or "carrier" has its phase changed in accordance with the information to be transmitted.

PING: Packet Inter Net Groper - very useful utility which probes for the existence of a TCP/IP host

POLLING: A means of controlling devices on a multipoint line.

PORT: A number in TCP/IP to which services are assigned; e.g. FTP is port 21; SMTP is port 25; HTTP is port 80.

PROTOCOL: A formal set of conventions governing the formatting and relative timing of message exchange between two communicating systems.

PULSE CODE MODULATION (PCM): The form of modulation in which the modulation signal is sampled, quantized, and coded so that each element of information consists of different types or numbers of pulses and spaces.

REMOTE TERMINAL (RT): All terminals not operating as the bus controller or as a bus monitor.

REPEATER: Buffer which cleans up, strengthens and re-transmits a signal.

ROUTER: Repeater which selectively re-distributes messages based on IP address

SERIAL TRANSMISSION: The most common transmission mode; in serial, information bits are sent sequentially on a single data channel.

SNMP: Simple Network Management Protocol; allows monitoring and management of a network.

SOCKET: Specific instance of an IP address and Port number that represents a single connection between two applications.

STAR TOPOLOGY: Topology which allows only one device at each end of a wire and requires repeaters for more than two devices.

STUBBING: Stubbing is the method wherein a separate line is connected between the primary data bus line and a terminal. The direct connection of stub line causes a mismatch which appears on the waveforms. This mismatch can be reduced by filtering at the receiver and by using bi-phase modulation. Stubs are often employed not only as a convenience in bus layout but as a means of coupling a unit to the line in such a manner that a fault on the stub or terminal will not greatly affect the transmission line operation. In this case, a network is employed in the stub line to provide isolation from the fault. These networks are also used for stubs that are of such length that the mismatch and reflection degrades bus operation. The preferred method of stubbing is to use transformer coupled stubs. The method provides the benefits of DC isolation, increased common mode protection, a doubling of effective stub impedance, and fault isolation for the entire stub and terminal. Direct coupled stubs should be avoided if at all possible. Direct coupled stubs provide no DC isolation or common mode rejection for the terminal external to its subsystem. Further, any shorting fault between the subsystems' internal isolation resistors (usually on the circuit board) and the main bus junction will cause failure of that entire bus. It can be expected that when the direct stub length exceeds 1.6 feet, that it will begin to distort the main bus waveforms. Note that this length includes the cable runs internal to a given subsystem.

SUBSYSTEM: The device or functional unit receiving data transfer service from the data bus.

SWITCH: Repeater which selectively re-distributes messages based on hardware MAC address

SYNCHRONOUS TRANSMISSION: Transmission in which data bits are sent at a fixed rate, with the transmitter and receiver synchronized. Synchronized transmission eliminates the need for start and stop bits.

TCP: Transmission Control Protocol - mechanism in TCP/IP that ensures that data arrives intact and in correct order

TELNET: Standard interface through which a client may access a host as though it were local

TERMINAL: The electronic module necessary to interface the data bus with the subsystem and the subsystem with the data bus. Terminals may exist as separate units or be contained within the elements of the subsystem.

TIME DIVISION MULTIPLEXING (TDM): The transmission of information from several signal sources through one communication system with different signal samples staggered in time to form a composite pulse train.

UDP: User Datagram Protocol - lower overhead alternative to TCP protocol which does not guarantee message delivery

WORD: A set of bits or bytes comprising the smallest unit of addressable memory. In MIL-STD-1553B, a word is a sequence of 16 bits plus sync and parity.

RS-232 INTERFACE

INTRODUCTION

The RS-232 interface is the Electronic Industries Association (EIA) standard for the interchange of serial binary data between two devices. It was initially developed by the EIA to standardize the connection of computers with telephone line modems. The standard allows as many as 20 signals to be defined, but gives complete freedom to the user. Three wires are sufficient: send data, receive data, and signal ground. The remaining lines can be hardwired on or off permanently. The signal transmission is bipolar, requiring two voltages, from 5 to 25 volts, of opposite polarity.

COMMUNICATION STANDARDS

The industry custom is to use an asynchronous word consisting of: a start bit, seven or eight data bits, an optional parity bit and one or two stop bits. The baud rate at which the word sent is device-dependent. The baud rate is usually 150 times an integer power of 2, ranging from 0 to 7 (150, 300, 600, ..., 19,200). Below 150 baud, many system-unique rates are used. The standard RS-232-C connector has 25 pins, 21 pins which are used in the complete standard. Many of the modem signals are not needed when a computer terminal is connected directly to a computer, and Figure 1 illustrates how some of the “spare” pins should be linked if not needed. Figure 1 also illustrates the pin numbering used in the original DB-25 connector and that now commonly used with a DB-9 connector normally used in modern computers

Specifying compliance to RS-232 only establishes that the signal levels in two devices will be compatible and that if both devices use the suggested connector, they may be able to be connected. Compliance to RS-232 does not imply that the devices will be able to communicate or even acknowledge each other’s presence.

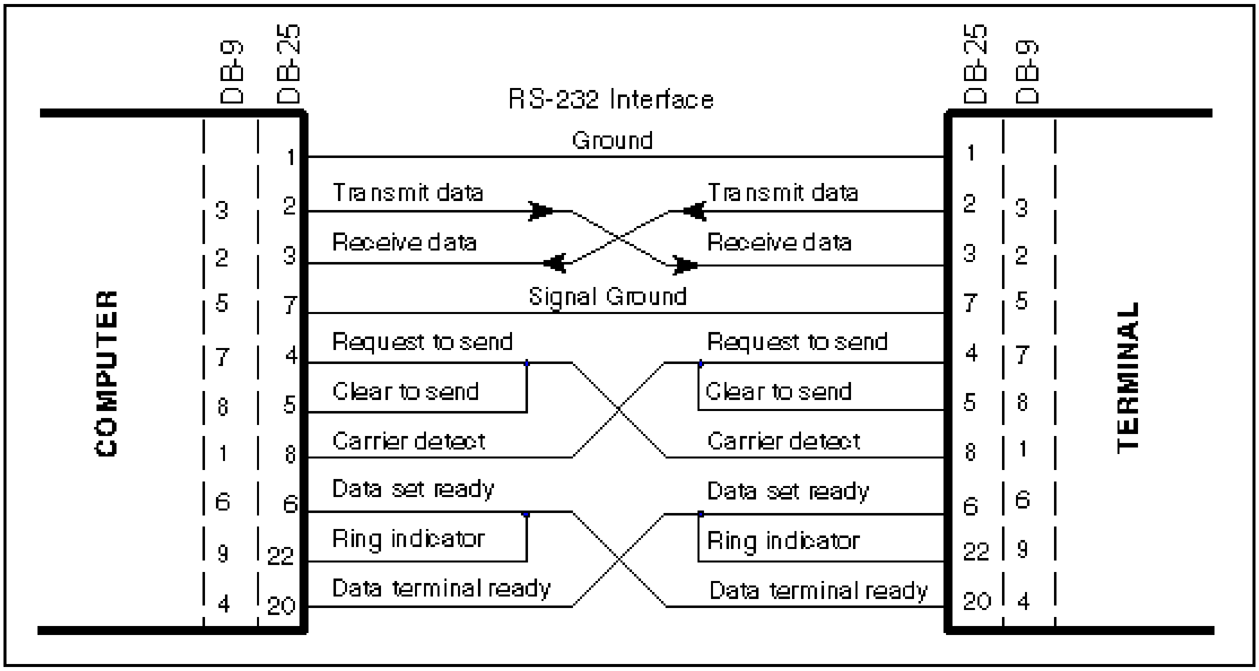
DB-9
DB-25
RS-232 Interface
DB-25
DB-9
COMPUTER
Ground
1
1
Transmit data
Transmit data
3
2
2
3
Receive data
Receive data
2
3
3
2
Signal Ground
5
7
7
5
Request to send
Request to send
7
4
4
7
Clear to send
Clear to send
8
5
5
8
Carrier detect
Carrier detect
1
8
8
1
Data set ready
Data set ready
6
6
6
6
Ring indicator
Ring indicator
9
22
22
9
Data terminal ready
Data terminal ready
4
20
20
4
TERMINAL

Figure 1. Direct-to-Computer RS-232 Interface.

Table 1 shows the signal names, and functions of the RS-232 serial port pinout. Table 2 shows a complete pin description.

Table 1. RS-232 Serial Port Pinout.

Name	Pin	Signal Name	Function
AA	1	PG Protective Ground	This line is connected to the chassis ground of the GPIB-232CV. Since the GPIB-232CV chassis ground is not connected to earth ground, pin 1 should be connected on both serial devices.
BA	2	TxD Transmit Data	This line carries serial data from the GPIB-232CV to the serial host.
BB	3	RxD Receive Data	This line carries serial data from the serial host to the GPIB-232CV.
CA	4	RTS Request to Send	This signal line is driven by the GPIB-232CV and when asserted indicates that the GPIB-232CV is ready to accept serial data. The GPIB-232CV un-asserts RTS when it is no longer ready to accept serial data because of a buffer full condition.
CB	5	CTS Clear to Send	This signal line is asserted by the serial host and sensed by the GPIB-232CV. When asserted, it indicates that the serial host is ready to accept serial data. When unasserted, it indicates that data transmission should be disabled.
AB	7	SG Signal Ground	This line establishes a reference point for all interface voltages.
CD	20	DTR Data Terminal Ready	This signal line is asserted by the GPIB-232CV to signal that it has been powered on, and is ready to operate.

Table 2. RS-232C Interface Signals.

Pin	Description	Pin	Description	Pin	Description
1	Protective Ground	10	(Reserved for Data Set Testing)	19	Secondary Request to Send
2	Transmitted Data	11	Unassigned	20	Data Terminal Ready
3	Received Data	12	Sec. Rec'd. Line Sig. Detector	21	Signal Quality Detector
4	Request to Send	13	Sec. Clear to Send	22	Ring Indicator
5	Clear to Send	14	Secondary Transmitted Data	23	Data Signal Rate Selector (DTE/DCE Source)
6	Data Set Ready	15	Transmission Signal Element Timing (DCE Source)	24	Transmit Signal Element Timing (DTE Source)
7	Signal Ground (Common Return)	16	Secondary Received Data	25	Unassigned
8	Received Line Signal Detector	17	Receiver Signal Element Timing (DCE Source)		
9	(Reserved for Data Set Testing)	18	Unassigned		

Electrical Characteristics: The RS-232-C specifies the signaling rate between the DTE and DCE, and a digital signal is used on all interchange circuits. The RS-232 standard specifies that logic "1" is to be sent as a voltage in the range -15 to -5 V and that logic "0" is to sent as a voltage in the range +5 to +15 V. The standard specifies that voltages of at least 3 V in amplitude will always be recognized correctly at the receiver according to their polarity, so that appreciable attenuation along the line can be tolerated. The transfer rate is rated > 20 kbps and a distance of < 15 m. Greater distance and data rates are possible with good design, but it is reasonable to assume that these limits apply in practice as well as in theory. The load impedance of the terminator side of the interface must be between 3000 and 7000 ohms, and not more than 2500 pF.

Table 3, summarizes the functional specifications of the most important circuits.

Table 3. RS-232-C Circuit Definitions.

Name	Direction to:	Function
Data Signals		
Transmitted Data (BA)	DCE	Data generated by DTE
Received Data (BB)	DTE	Data Received by DTE
Timing signals		
Transmitter Signal Element Timing (DA)	DCE	Clocking signal, transitions to ON and OFF occur at center of each signal element
Transmitter Signal Element Timing (DB)	DTE	Clocking signal, as above; both leads relate to signals on BA
Receiver Signal Element Timing (DD)	DTE	Clocking signal, as above, for circuit BB
Control Signals		
Request to Send (CA)	DCE	DTE wishes to transmit
Clear to Send (CB)	DTE	DCE is ready to transmit; response to request to send
Data Set Ready (CC)	DTE	DCE is ready to operate
Data Terminal Ready (CD)	DCE	DTE is ready to operate
Ring Indicator (CE)	DTE	Indicates that DCE is receiving a ringing signal on the communication channel
Carrier Detect (CF)	DTE	Indicates that DCE is receiving a carrier signal
Signal Quality Detector (CG)	DTE	Asserted when there is reason to believe there is an error in the received data
Data Signal Rate Selector (CH)	DCE	Asserted to select the higher of two possible data rates
Data Signal Rate Selector (CI)	DTE	Asserted to select the higher of two possible data rates
Ground		
Protective Ground (AA)	NA	Attached to machine frame and possibly external grounds
Signal Ground (AB)	NA	Establishes common ground reference for all circuits

Range: The RS-232-C standard specifies that the maximum length of cable between the transmitter and receiver should not exceed 100 feet. Although in practice many systems are used in which the distance between transmitter and receiver exceeds this rather low figure. The limited range of the RS-232C standard is one of its major shortcomings compared with other standards which offer greater ranges within their specifications. One reason why the range of the RS-232C standard is limited is the need to charge and discharge the capacitance of the cable connecting the transmitter and receiver.

Mechanical Characteristics: The connector for the RS-232-C is a 25 pin connector with a specific arrangement of wires. In theory, a 25 wire cable could be used to connect the Data Terminal Equipment (DTE) to the Data Communication Equipment (DCE). The DTE is a device that is acting as a data source, data sink, or both, e.g. a terminal, peripheral, or computer. The DCE is a device that provides the functions required to establish, maintain, and terminate a data-transmission connecting, as well as the signal conversion, and coding required for communication between data terminal equipment and data circuit; e.g. a modem. Table 4, shows the complete summary of the RS-232-C, e.g., descriptor, sponsor, data format, etc.

Table 4. Summary of the RS-232-C.

Data Format	5- to 8- bit serial
Transfer Type	Asynchronous
Error Handling	Optional Parity Bit
Connector	25-pin female connector on DCE; 25-pin male connector on DTE
Length	20 meters
Speed	20 kb/s
Remarks	RS-232 is used in the microcomputer world for communications between two DTEs. The null-modem is included into one or both connecting devices, and/or cable and is seldom documented. As a result, establishing an RS-232 connection between two DTEs is frequently a difficult task.

RS-422 BALANCED VOLTAGE INTERFACE

Specifying compliance to RS-422 only establishes that the signal between the specified devices will be compatible. It does not indicate that the signal functions or operations between the two devices are compatible. The RS-422 standard only defines the characteristic requirements for the balanced line drivers and receivers. It does not specify one specific connector, signal names, or operations. RS-422 interfaces are typically used when the data rate or distance criteria cannot be met with RS-232. The RS-422 standard allows for operation of up to 10 receivers from a single transmitter. The standard does not define operations of multiple tri-stated transmitters on a link.

The RS-422-A interfaces between the Data Terminal Equipment (DTE) and Data Communication Equipment (DCE) or in any point-to-point interconnection of signals between digital equipment. It employs the electrical characteristics of balanced-voltage digital interface circuits.

The balanced voltage digital interface circuit will normally be utilized on data, timing, or control circuits where the data signaling rate is up to 10 Mbit/s. While the balanced interface is intended for use at the higher data signaling rate, it may (in preference to the unbalanced interface circuit) generally be required if any of the following conditions prevail:

- The interconnecting cable is too long for effective unbalanced operation.
- The interconnecting cable is exposed to an extraneous noise source that may cause an unwanted voltage in excess of + 1 volt measured differentially between the signal conductor and circuit common at the load end of the cable with a 50 ohm resistor substituted for the generator.
- It is necessary to minimize interference with other signals.
- Inversion of signals may be required, i.e. plus to minus MARK may be obtained by inverting the cable pair.

Applications of the balanced voltage digital interface circuit are shown in Figure 1.

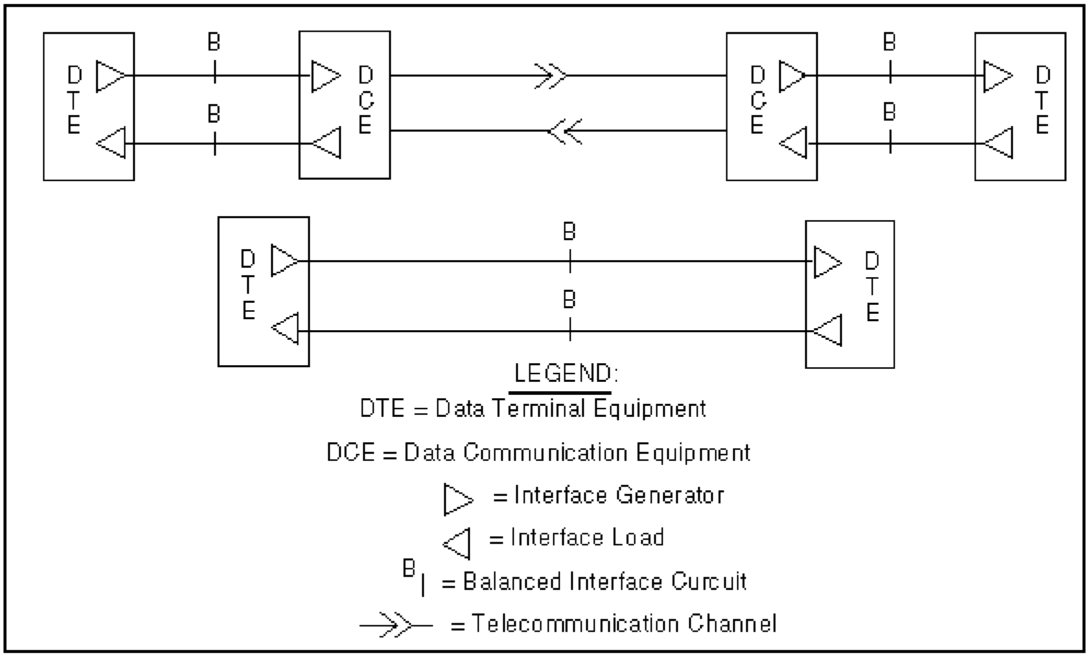
DTE
B
B
DCE
DCE
B
B
DTE
DTE
B
B
DTE
LEGEND:
DTE = Data Terminal Equipment
DCE = Data Communication Equipment
= Interface Generator
= Interface Load
B = Balanced Interface Curcuit
= Telecommunication Channel

Figure 1. Application of a RS-422 Circuit.

While a restriction on maximum cable length is not specified, guidelines are given later with respect to conservative operating distances as function of data signaling rate.

For a binary system in which the RS-422-A is designed, the data signaling rate in bit/s and the modulation in bauds are numerically equal when the unit interval used in each determination is the minimum interval.

Electrical Characteristics:

The balanced voltage digital interface circuit consists of three parts: the generator (G), the balanced interconnecting cable, and the load. The load is comprised of one or more receivers (R) and an optional cable termination resistance (RT). The balanced voltage interface circuit is shown in Figure 2.

Environmental Constraints:

Balanced voltage digital interface conforming to this standard will perform satisfactorily at data signaling rates up to 10 Mbit/s providing that the following operational constraints are satisfied:

- The interconnecting cable length is within that recommended for the applicable data signaling rate (see Figure 3) and the cable is appropriately terminated.
- The common mode voltage at the receiver is less than 7 volts (peak). The common mode voltage is defined to be any uncompensated combination of generator-receiver ground potential difference, the generator offset voltage (Vos), and longitudinally coupled peak noise voltage measured between the received circuit ground and cable within the generator ends of the cable short-circuited to ground.

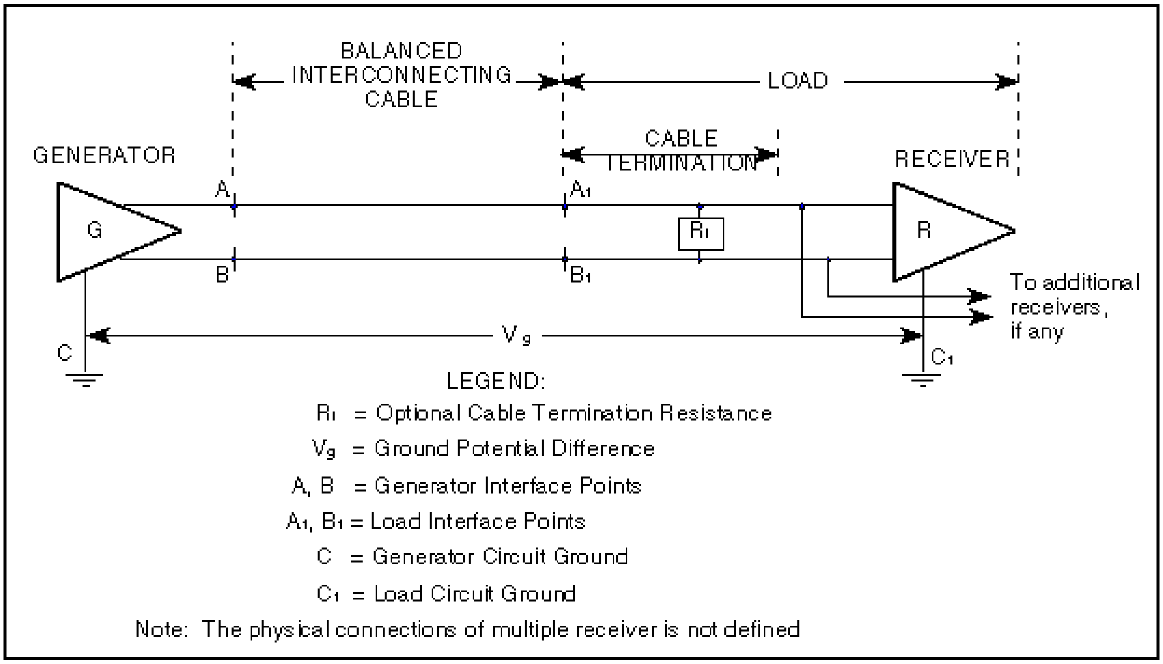
BALANCED INTERCONNECTING CABLE
LOAD
CABLE TERMINATION
GENERATOR
RECEIVER
G
R
A
B
A1
B1
Rt
To additional receivers, if any
Vg
C
C1
LEGEND:
Rt = Optional Cable Termination Resistance
Vg = Ground Potential Difference
A, B = Generator Interface Points
A1, B1 = Load Interface Points
C = Generator Circuit Ground
C1 = Load Circuit Ground
Note: The physical connections of multiple receiver is not defined

Figure 2. Balanced Digital Interface Circuit.

Interconnecting Cable Guidelines:

The maximum permissible length of cable separating the generator and the load is a function of data signaling rate and is influenced by the tolerable signal distortion, the amount of coupled noise and ground potential difference introduced between the generator and load circuit as well as by cable balance. The curve of cable length versus signaling rate is given in Figure 3. This curve is based upon using 24 AWG copper, twisted-pair cable with a capacitance of 52.5 pF/meter terminated in a 100 ohm load. As data signaling rate is reduced below 90 kbit/s, the cable length has been limited at 1200 meters by the assumed maximum allowable 6 dBV signal loss.

Industry customs are not nearly as well established for RS-422 interfaces as they are for RS-232. The standard specifies use of the 37-pin "D"; the 9-pin "D" is specified for use with the secondary channel. Most data communications equipment uses the 37-pin "D"; many computer applications use a 9-pin "D" only. Some equipment applications use the 25-pin "D" defined for RS-232.

Compatibility With Other Interfaces:

Since the basic differential receivers of RS-423-A and RS-422-A are electrically identical, it is possible to interconnect an equipment using RS-423-A receivers and generators on one side of the interface with an equipment using RS-422-A generators and receivers on the other side of the interface, if the leads of the receivers and generators are properly configured to accommodate such an arrangement and the cable is not terminated.

This circuit is not intended for interoperation with other interface electrical circuits such as RS-232-C, MIL-STD-188C, or CCITT (Comite Consultatif Internationale Telegraphique et Telephonique), recommendations V.28 and V.35. Under certain conditions, the above interfaces may be possible but may require modification of the interface or equipment; therefore satisfactory operation is not assured and additional provisions not specified herein may be required.

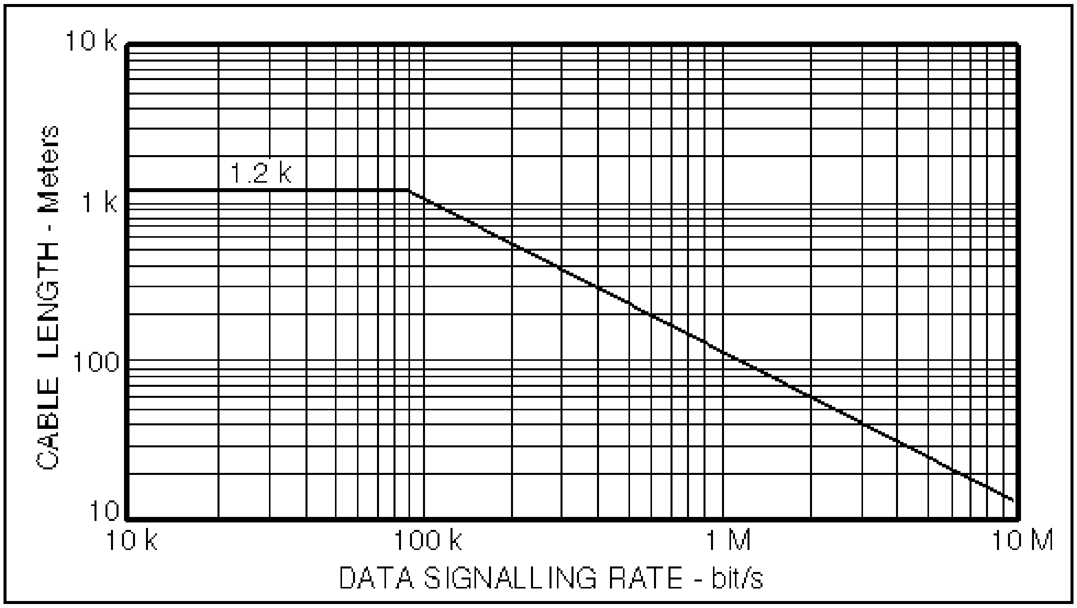

Figure 3. Data Signaling Rate vs. Cable Length.

This page intentionally left blank.

RS-485 INTERFACE

STANDARD FOR ELECTRICAL CHARACTERISTICS OF GENERATORS AND RECEIVERS FOR USE IN BALANCED DIGITAL MULTIPOINT SYSTEMS

Introduction: The RS-485 is the recommend standard by the Electronic Industries Association (EIA) that specifies the electrical characteristics of generators and receivers that may be employed for the interchange of binary signals in multipoint interconnection of digital equipments. When implemented within the guidelines, multiple generators and receivers may be attached to a common interconnecting cable. An interchange system includes one or more generators connected by a balanced interconnecting cable to one or more receivers and terminating resistors.

Electrical Characteristics: The electrical characteristics that are specified are measured at an interconnect point supplied by the devices manufacturer. Figure 1 shows an interconnection application of generators and receivers having the electrical parameters specified. The elements in the application are: generators, receivers, transmission cables, and termination resistances (Rt). The loads on the system caused by each receiver and passive generator shall be defined in terms of unit loads. Each generator can drive up to 32 unit loads consisting of both receivers and generators in the passive state. The loading caused by receivers and passive generators on the interconnect must be considered in defining the device electrical characteristics. Two areas are of concern: the DC load and the AC load characteristics. The DC load is defined as a number or fractions of "unit loads." The AC loading is not standardized but must be considered in the design of a system using the devices meeting this standard.

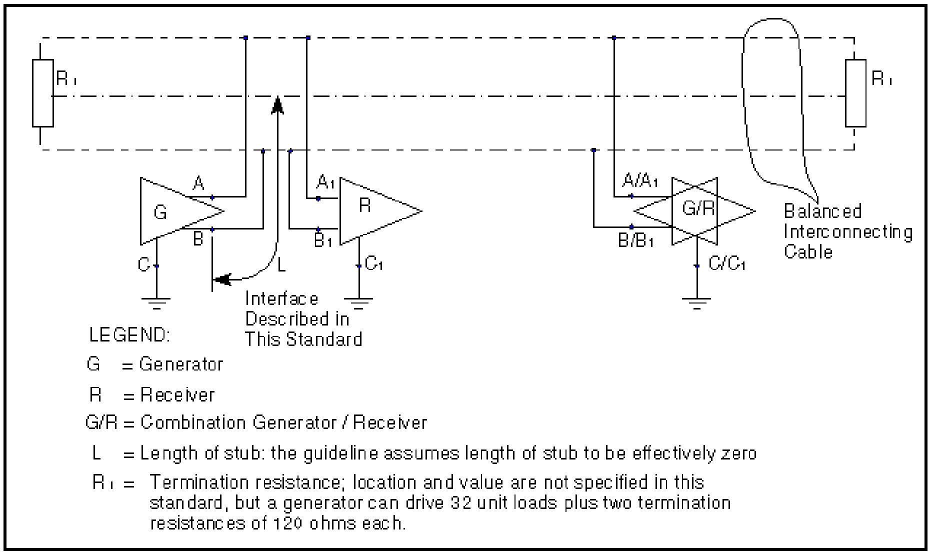

R t
R t
A
B
C
G
A1
B1
C1
R
L
Interface Described in This Standard
A/A1
B/B1
C/C1
G/R
Balanced Interconnecting Cable
LEGEND:
G = Generator
R = Receiver
G/R = Combination Generator / Receiver
L = Length of stub: the guideline assumes length of stub to be effectively zero
R t = Termination resistance; location and value are not specified in this standard, but a generator can drive 32 unit loads plus two termination resistances of 120 ohms each.

Figure 1. Multipoint Interconnect Application.

General System Configuration: The generators and receivers conforming to the RS-485 standard can operate with a common mode voltage between -7 volts and +7 volts (instantaneous). The common mode voltage is defined to be any uncompensated combination of generator-receiver ground potential difference and longitudinally coupled peak noise voltage measured between the receiver circuit ground and cable with the generator ends of the cable short circuited to ground, plus the generator offset voltage (Vos).

Grounding Arrangements: Proper operation of the generator and receiver circuits requires the presence of a signal return path between the circuit grounds of the equipment at each end of the interconnection. The grounding arrangements are shown in Figure 2. Where the circuit reference is provided by a third conductor, the connection between circuit common and the third conductor must contain some resistance (e.g., 100 ohms) to limit circulating currents when other ground connections are provided for safety. Some applications may require the use of shielded interconnecting cable for EMI or other purposes. The shield shall be connected to frame ground at either or both ends, depending on the application.

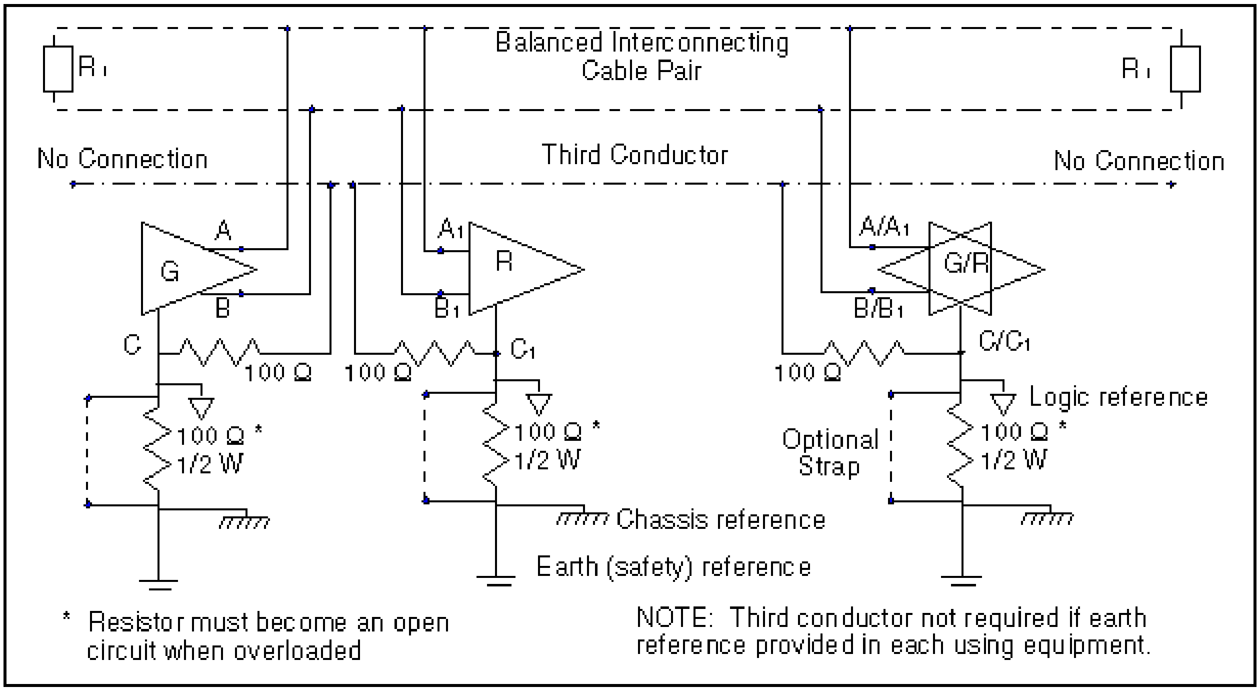
Balanced Interconnecting Cable Pair
R1
R1
No Connection
Third Conductor
No Connection
A
G
B
A1
R
B1
A/A1
G/R
B/B1
C
100 Ω
100 Ω
C1
100 Ω
C/C1
Logic reference
100 Ω *
1/2 W
100 Ω *
1/2 W
Optional Strap
100 Ω *
1/2 W
Chassis reference
Earth (safety) reference

* Resistor must become an open circuit when overloaded

NOTE: Third conductor not required if earth reference provided in each using equipment.

Figure 2. Grounding Arrangements.

Similarity with RS-422-A:

In certain instances, it may be possible to produce generators and receivers that meet the requirements of both RS-422-A and of RS-485. Table 1 depicts the differences in parameter specifications which exist between the two documents.

Table 1. Comparison of RS-422-A and RS-485 Characteristics.

Characteristic	RS-422-A	RS-485
Min. output voltage	2V into 100 ohm > 1/2 open circuit V	1.5 V into 54 ohms
I_{short} to ground	150 mA maximum	
I_{short} to -7, +12 volts		250 mA peak
t_{rise} time	< 0.1 t_b , 100 ohm load	< 0.3 t_b , 54 ohm, 50 pF load

Where t_b = time duration of the unit interval at the applicable data signaling rate (pulse width).

IEEE-488 INTERFACE BUS (HP-IB/GP-IB)

In the early 1970's, Hewlett-Packard came out with a standard bus (HP-IB) to help support their own laboratory measurement equipment product lines, which later was adopted by the IEEE in 1975. This is known as the IEEE Std. 488-1975. The IEEE-488 Interface Bus (HP-IB) or general purpose interface bus (GP-IB) was developed to provide a means for various instruments and devices to communicate with each other under the direction of one or more master controllers. The HP-IB was originally intended to support a wide range of instruments and devices, from the very fast to the very slow.

DESCRIPTION

The HP-IB specification permits up to 15 devices to be connected together in any given setup, including the controller if it is part of the system. A device may be capable of any other three types of functions: controller, listener, or talker. A device on the bus may have only one of the three functions active at a given time. A controller directs which devices will be talkers and listeners. The bus will allow multiple controllers, but only one may be active at a given time. Each device on the bus should have a unique address in the range of 0-30. The maximum length of the bus network is limited to 20 meters total transmission path length. It is recommended that the bus be loaded with at least one instrument or device every 2 meter length of cable (4 meters is maximum). The use of GP-IB extenders may be used to exceed the maximum permitted length of 20 meters.

ELECTRICAL INTERFACE

The GP-IB is a bus to which many similar modules can be directly connected, as is shown in Figure 1. A total of 16 wires are shown in the figure - eight data lines and eight control lines. The bus cables actually have 24 wires, providing eight additional for shielding and grounds.

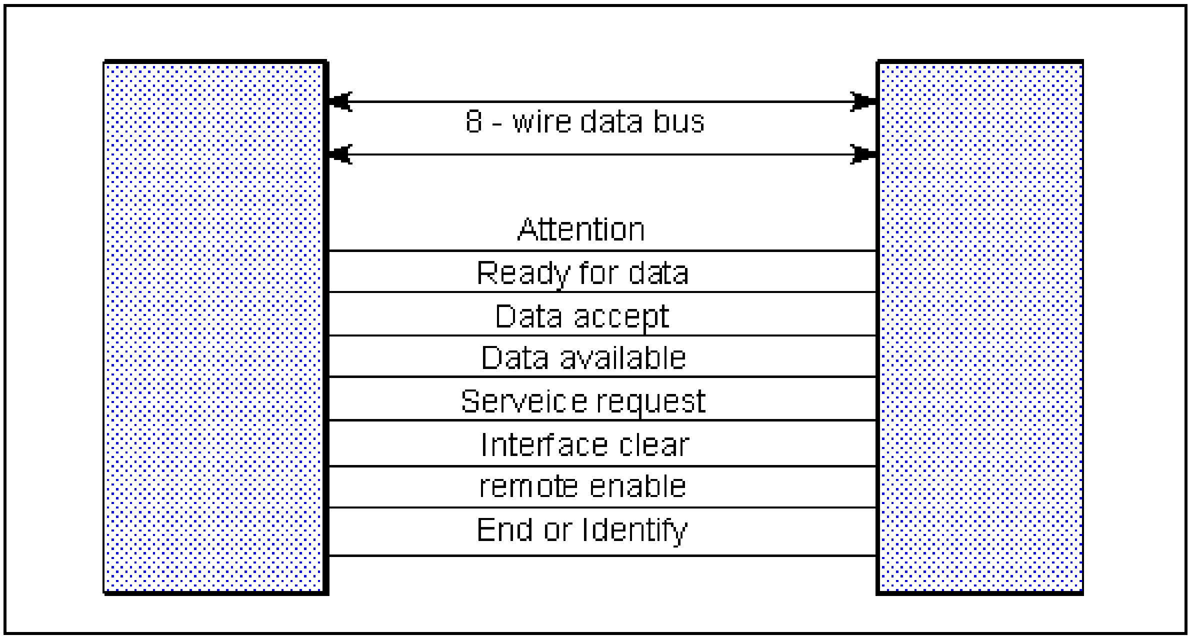

Figure 1. IEEE-488 (HP-IB / GP-IB) Bus Configuration.

The GP-IB defines operation of a three-wire handshake that is used for all data transfers on the bus. The bus operation is asynchronous in nature. The data-transfer rate of the GP-IB is 500 kHz for standard applications and can go up 1 MHz if special conventions are followed. Each transaction carries 8 bits, the maximum data bandwidth is on the order of 4 to 8 megabits (1 M byte) per second. The bus is a two way communications channel and data flows in both directions. Figure 2 illustrates the structure of the GP-IB bus and identifies the 16 connections of the interconnecting cable.

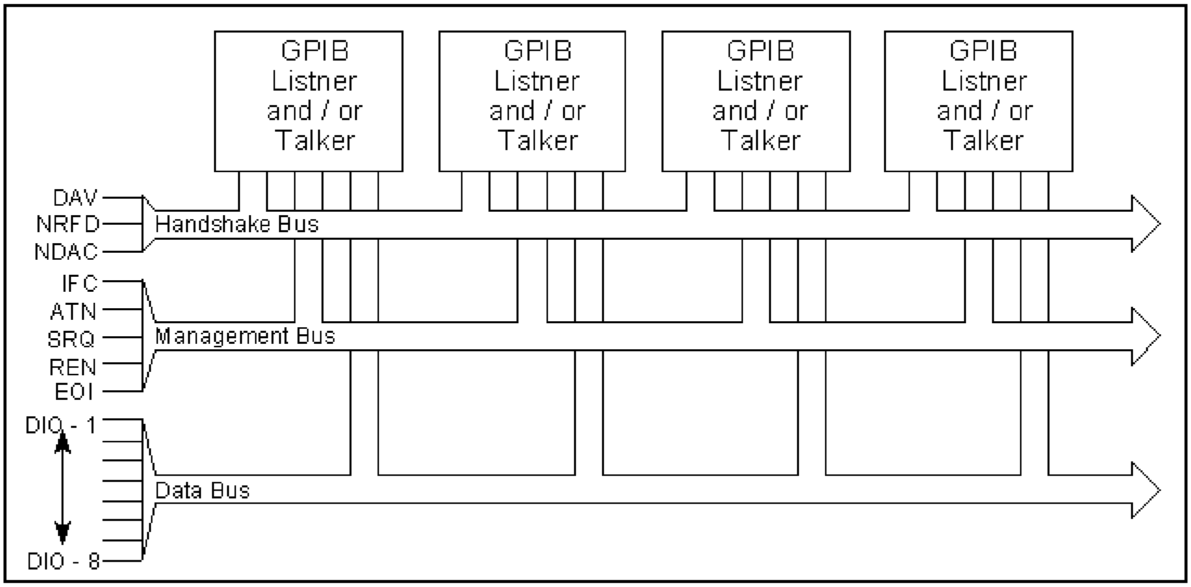
GPIB Listner and / or Talker
GPIB Listner and / or Talker
GPIB Listner and / or Talker
GPIB Listner and / or Talker
DAV
NRFD
NDAC
Handshake Bus
IFC
ATN
SRQ
REN
EOI
Management Bus
DIO - 1
Data Bus
DIO - 8

Figure 2. GP-IB Instrumentation Bus Structure.

The cabling limitations make it a less-than-ideal choice for large separation between devices. These limitations can be overcome with bus extenders. Those attempting to use bus extenders should be aware that few extenders are as transparent as claimed. This is especially true in handling of continuous data and interrupts. In nonextended environments, it provides an excellent means for high-speed computer control of multiple devices.

The following table shows the various interface functions, the mnemonics and the descriptions.

Table 1. GP-IB Interface Functions.

Interface Function	Mnemonic	Description
Talker (extended talker)	T (TE)	Device must be able to transmit
Listener (Extended listener)	L (LE)	Device must receive commands and data
Source Handshake	SH	Device must properly transfer a multiline message
Acceptor Handshake	AH	Device must properly receive remote multiline messages
Remote/Local	RL	Device must be able to operate from front panel and remote information from bus
Service Request	SR	Device can asynchronously request service from the controller
Parallel Poll	PP	Upon controller request, device must uniquely identify itself if it requires service
Device Clear	DC	Device can be initialized to a predetermined state
Device Trigger	DT	A device function can be initiated by the talker on the bus
Controller	C	Device can send addresses, universal commands, address commands, and conduct polls
Drivers	E	This code describes the type of electrical drivers in a device

The cabling specifications of the GP-IB interface system permit interconnecting all devices together in a star or linear configuration. The GP-IB connector is a 24-pin ribbon-type connector.

In summary, Table 2 on this page and the next shows the complete description of the GP-IB data bus.

Table 2. GP-IB Data Bus Description.

IEEE-488, GP-IB, HP-IB, or IEC-625					
Descriptor	8-bit parallel, monodirectional, multi-master (token passing) One controller, one talker, several listeners	Arbitration	Token passing: the controller addresses the next controller SRQ Service request when the controller assigns modes	Connector	24-pin Amphenol Female connector on equipment chassis. DIO1 1 ■ ■ 13 DIO5 DIO2 2 ■ ■ 14 DIO6 DIO3 3 ■ ■ 15 DIO7 DIO4 4 ■ ■ 16 DIO8 EOI 5 ■ ■ 17 REN DAV 6 ■ ■ 18 Gnd NRFD 7 ■ ■ 19 Gnd NDAC 8 ■ ■ 20 Gnd IFC 9 ■ ■ 21 Gnd SRQ 10 ■ ■ 22 Gnd ATN 11 ■ ■ 23 Gnd Shld 2 ■ ■ 24 Gnd
Sponsor	Hewlett-Packard	Error handling	Parity bit DI07 when 7-bit ACSII characters		
Standard	IEEE 488, IEC 625	Bus length	15 m		
Address space	31 devices	Driver	Special 24 mA drivers		
Data format	8-bit parallel	Speed	1 MByte/s		
Transfer type	Write only, talker toward listener(s) or commander toward all others			Remarks	The 488 is most commonly used for data acquisition of H-P peripherals. Programmable interfaces and drivers exist and simplify the development of microprocessor interfaces.
Timing	Handshaken 3-wire broadcast transfer: DAV data valid NDAC Not data accepted NRFD Not ready for data	References	IEEE Computer Society		

HP-IL VARIATION

Since introduction of the IEEE-488, technology produced a generation of medium-speed, low-power, instrumentation which had a need to operate in an automatic test system such as the GP-IB. The HP-IL (Hewlett-Packard Interface Loop), was introduced to meet this need. The HP-IL is a low-cost, low-power alternative to the GP-IB system. The HP-IL and GP-IB provide the same basic functions in interfacing controllers, instruments, and peripherals, but they differ in many other respects. HP-IL is suitable for use in low-power, portable applications (typically used for interface of battery-power systems). The GP-IB is not practical to operate from battery power. The HP-IL maximum data rate is 20K bytes per second. This is a high rate compared to the RS-232C, but much slower than GP-IB. The HP-IL can operate over distances of up to 100 meters between any two devices. Since it is a loop environment, there is no maximum system cable restriction. The basic device-addressing scheme allows for up to 30 devices on a loop.

MIL-STD-1553 & 1773 DATA BUS

PURPOSE

In recent years, the use of digital techniques in aircraft equipment has greatly increased, as have the number of avionics subsystems and the volume of data processed by them.

Because analog point-to-point wire bundles are inefficient and cumbersome means of interconnecting the sensors, computers, actuators, indicators, and other equipment onboard the modern military vehicle, a serial digital multiplex data bus was developed. MIL-STD-1553 defines all aspects of the bus, therefore, many groups working with the military tri-services have chosen to adopt it.

The 1553 multiplex data bus provides integrated, centralized system control and a standard interface for all equipment connected to the bus. The bus concept provides a means by which all bus traffic is available to be accessed with a single connection for testing and interfacing with the system. The standard defines operation of a serial data bus that interconnects multiple devices via a twisted, shielded pair of wires. The system implements a command-response format.

MIL-STD-1553, “Aircraft Internal Time-Division Command/Response Multiplex Data Bus,” has been in use since 1973 and is widely applied. MIL-STD-1553 is referred to as “1553” with the appropriate revision letter (A or B) as a suffix. The basic difference between the 1553A and the 1553B is that in the 1553B, the options are defined rather than being left for the user to define as required. It was found that when the standard did not define an item, there was no coordination in its use. Hardware and software had to be redesigned for each new application. The primary goal of the 1553B was to provide flexibility without creating new designs for each new user. This was accomplished by specifying the electrical interfaces explicitly so that compatibility between designs by different manufacturers could be electrically interchangeable.

The Department of Defense chose multiplexing because of the following advantages:

- Weight reduction
- Simplicity
- Standardization
- Flexibility

Some 1553 applications utilize more than one data bus on a vehicle. This is often done, for example, to isolate a Stores bus from a Communications bus or to construct a bus system capable of interconnecting more terminals than a single bus could accommodate. When multiple buses are used, some terminals may connect to both buses, allowing for communication between them.

MULTIPLEXING

Multiplexing facilitates the transmission of information along the data flow. It permits the transmission of several signal sources through one communications system.

BUS

The bus is made up of twisted-shielded pairs of wires to maintain message integrity. MIL-STD-1553 specifies that all devices in the system will connect to a redundant pair of buses. This provides a second path for bus traffic should one of the buses be damaged. Signals are only allowed to appear on one of the two buses at a time. If a message cannot be completed on one bus, the bus controller may switch to the other bus. In some applications more than one 1553 bus may be implemented on a given vehicle. Some terminals on the bus may actually connect to both buses.

BUS COMPONENTS

There are only three functional modes of terminals allowed on the data bus: the bus controller, the bus monitor, and the remote terminal. Devices may be capable of more than one function. Figure 1 illustrates a typical bus configuration.

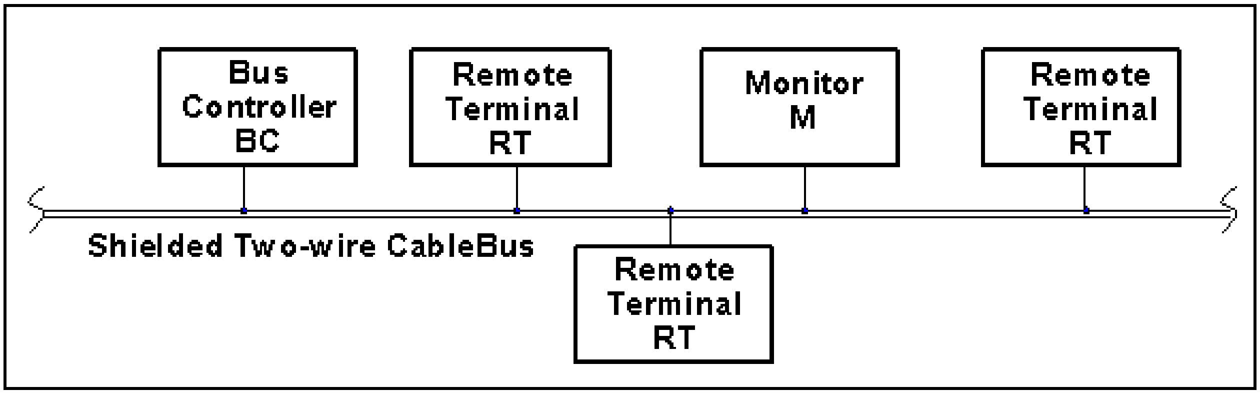

Figure 1. 1553 Bus Structure.

- Bus Controller - The bus controller (BC) is the terminal that initiates information transfers on the data bus. It sends commands to the remote terminals which reply with a response. The bus will support multiple controllers, but only one may be active at a time. Other requirements, according to 1553, are: (1) it is "the key part of the data bus system," and (2) "the sole control of information transmission on the bus shall reside with the bus controller."

- Bus Monitor - 1553 defines the bus monitor as "the terminal assigned the task of receiving bus traffic and extracting selected information to be used at a later time." Bus monitors are frequently used for instrumentation.

- Remote Terminal - Any terminal not operating in either the bus controller or bus monitor mode is operating in the remote terminal (RT) mode. Remote terminals are the largest group of bus components.

MODULATION

The signal is transferred over the data bus using serial digital pulse code modulation.

DATA ENCODING

The type of data encoding used by 1553 is Manchester II biphase.

- A logic one (1) is transmitted as a bipolar coded signal 1/0 (in other words, a positive pulse followed by a negative pulse).
- A logic zero (0) is a bipolar coded signal 0/1 (i.e., a negative pulse followed by a positive pulse).

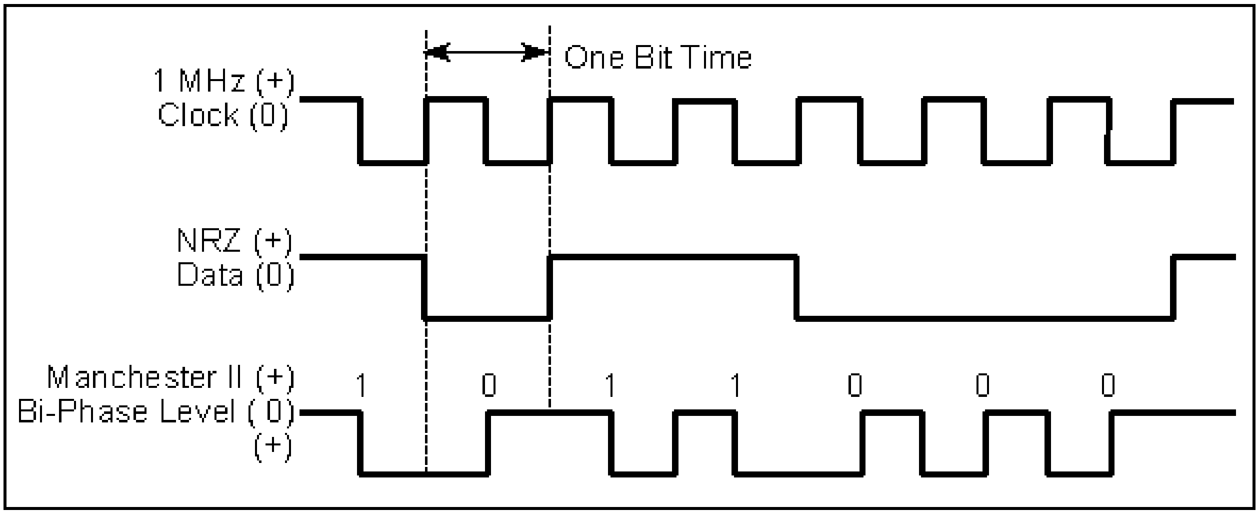

Figure 2. Data Encoding.

A transition through zero occurs at the midpoint of each bit, whether the rate is a logic one or a logic zero. Figure 2 compares a commonly used Non Return to Zero (NRZ) code with the Manchester II biphase level code, in conjunction with a 1 MHz clock.

BIT TRANSMISSION RATE

The bit transmission rate on the bus is 1.0 megabit per second with a combined accuracy and long-term stability of +/- 0.1%. The short-term stability is less than 0.01%.

There are 20 1.0-microsecond bit times allocated for each word. All words include a 3 bit-time sync pattern, a 16-bit data field that is specified differently for each word type, and 1 parity check bit.

WORD FORMATS

Bus traffic or communications travels along the bus in words. A word in MIL-STD-1553 is a sequence of 20 bit times consisting of a 3 bit-time sync wave form, 16 bits of data, and 1 parity check bit. This is the word as it is transmitted on the bus; 1553 terminals add the sync and parity before transmission and remove them during reception. Therefore, the nominal word size is 16 bits, with the most significant bit (MSB) first.

There are three types of words: command, status, and data. A packet is defined to have no inter-message gaps. The time between the last word of a controller message and the return of the terminal status byte is 4-12 microseconds. The time between status byte and the next controller message is undefined. Figure 3 illustrates these three formats.

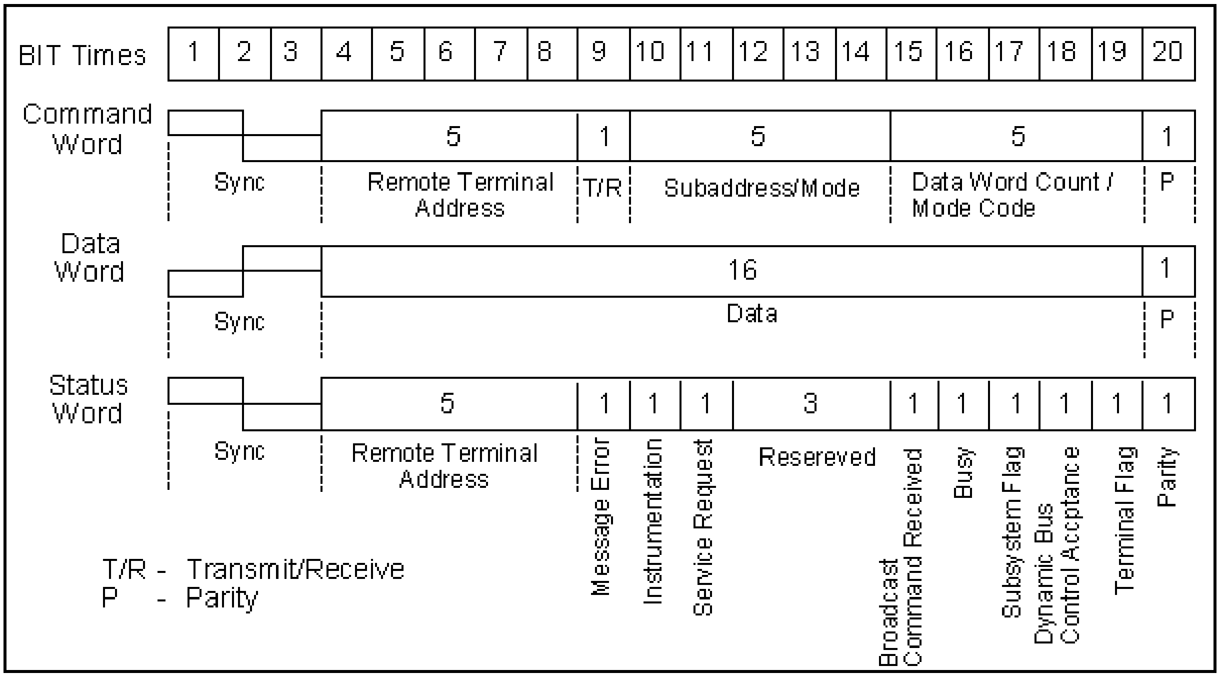
BIT Times
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Command Word
5
1
5
5
1
Sync
Remote Terminal Address
T/R
Subaddress/Mode
Data Word Count / Mode Code
P
Data Word
16
1
Sync
Data
P
Status Word
5
1
1
1
3
1
1
1
1
1
1
Sync
Remote Terminal Address
Message Error
Instrumentation
Service Request
Resereved
Broadcast Command Received
Busy
Subsystem Flag
Dynamic Bus Control Accptance
Terminal Flag
Parity
T/R - Transmit/Receive
P - Parity

Figure 3. 1553 Word Formats.

COMMAND WORD

Command words are transmitted only by the bus controller and always consist of:

- 3 bit-time sync pattern
- 5 bit RT address field
- 1 Transmit/Receive (T/R) field
- 5 bit subaddress/mode field
- 5 bit word count/mode code field
- 1 parity check bit.

DATA WORD

Data words are transmitted either by the BC or by the RT in response to a BC request. The standard allows a maximum of 32 data words to be sent in a packet with a command word before a status response must be returned. Data words always consist of:

- 3 bit-time sync pattern (opposite in polarity from command and status words)
- 16 bit data field
- 1 parity check bit.

STATUS WORD

Status words are transmitted by the RT in response to command messages from the BC and consist of:

- 3 bit-time sync pattern (same as for a command word)
- 5 bit address of the responding RT
- 11 bit status field
- 1 parity check bit.

The 11 bits in the status field are used to notify the BC of the operating condition of the RT and subsystem.

INFORMATION TRANSFERS

Three basic types of information transfers are defined by 1553:

- Bus Controller to Remote Terminal transfers
- Remote Terminal to Bus Controller transfers
- Remote Terminal to Remote Terminal transfers

These transfers are related to the data flow and are referred to as messages. The basic formats of these messages are shown in Figure 4.

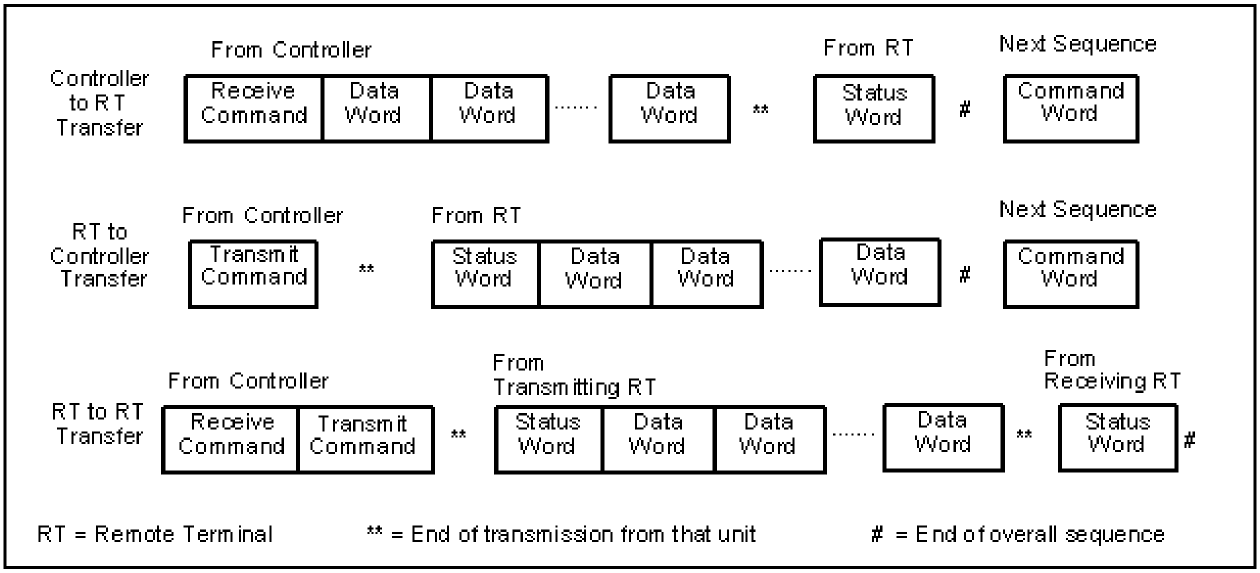
Controller to RT Transfer
From Controller
Receive Command
Data Word
Data Word
....
Data Word
**
From RT
Status Word
#
Next Sequence
Command Word
RT to Controller Transfer
From Controller
Transmit Command
**
From RT
Status Word
Data Word
Data Word
....
Data Word
#
Next Sequence
Command Word
RT to RT Transfer
From Controller
Receive Command
Transmit Command
**
From Transmitting RT
Status Word
Data Word
Data Word
....
Data Word
**
From Receiving RT
Status Word
#
RT = Remote Terminal
** = End of transmission from that unit
# = End of overall sequence

Figure 4. 1553 Data Message Formats.

The normal command/response operation involves the transmission of a command from the BC to a selected RT address. The RT either accepts or transmits data depending on the type (receive/transmit) of command issued by the BC. A status word is transmitted by the RT in response to the BC command if the transmission is received without error and is not illegal.

Figure 5 illustrates the 1553B Bus Architecture in a typical aircraft.

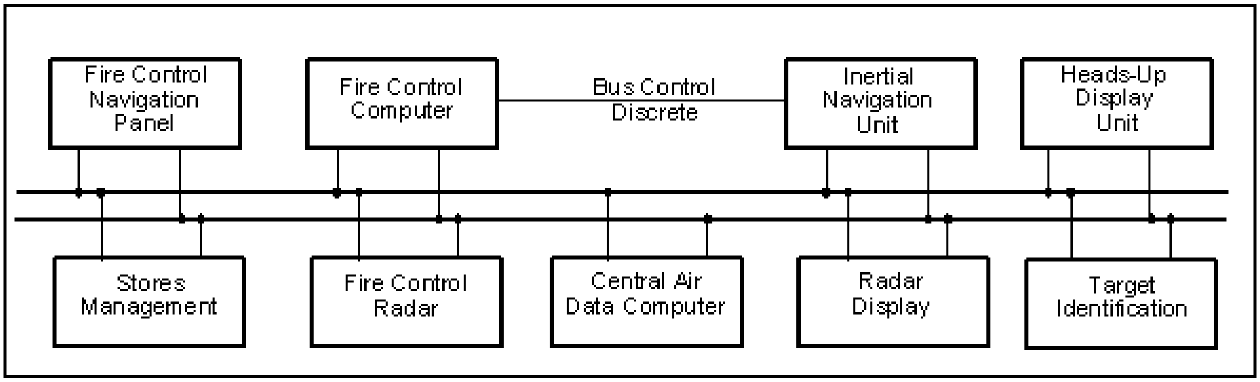
Fire Control Navigation Panel
Fire Control Computer
Bus Control Discrete
Inertial Navigation Unit
Heads-Up Display Unit
Stores Management
Fire Control Radar
Central Air Data Computer
Radar Display
Target Identification

Figure 5. Typical Bus Architecture.

MIL-STD-1773

MIL-STD-1773 contains the requirements for utilizing a fiber optic “cabling” system as a transmission medium for the MIL-STD-1553B bus protocol. As such, the standard repeats MIL-STD-1553 nearly word-for-word. The standard does not specify power levels, noise levels, spectral characteristics, optical wavelength, electrical/optical isolation or means of distributing optical power. These must be contained in separate specifications for each intended use.

Data encoding and word format are identical to MIL-STD-1553, with the exception that pulses are defined as transitions between 0 (off) and 1 (on) rather than between + and - voltage transitions since light cannot have a negative value.

Since the standard applies to cabling only, the bus operates at the same speed as it would utilizing wire. Additionally, data error rate requirements are unchanged.

Different environmental considerations must be given to fiber optic systems. Altitude, humidity, temperature, and age affects fiber optics differently than wire conductors. Power is divided evenly at junctions which branch and connectors have losses just as wire connectors do.

ETHERNET

HISTORY

Ethernet was originally developed by Digital, Intel, and Xerox (DIX) in the early 1970's at the Xerox Palo Alto Research Center (PARC). The two primary inventors were Robert Metcalf and David Boggs.

IEEE Project 802 was set up in February 1980 to provide standards for Local Area Network (LAN) and Metropolitan Area Network (MAN) Architecture. The following committees were established:

- IEEE 802.1: Standards related to network management.
- IEEE 802.2: General standard for the data link layer in the OSI Reference Model. The IEEE divides this layer into two sub-layers -- the logical link control (LLC) layer and the media access control (MAC) layer. The MAC layer varies for different network types and is defined by standards IEEE 802.3 through IEEE 802.5.
- IEEE 802.3: Defines the MAC layer for bus networks that use CSMA/CD. **This is the basis of what is most commonly referred to as the Ethernet standard**.
- IEEE 802.4: Defines the MAC layer for token-bus networks.
- IEEE 802.5: Defines the MAC layer for token-ring networks.
- IEEE 802.6: Standard for Metropolitan Area Networks (MANs).
- IEEE 802.7-9 Inactive Standards:
  - IEEE 802.7: Broadband local area networks. The working group is currently inactive.
  - IEEE 802.8: The Fibre Optic Technical Advisory Group was to create a LAN standard for fiber optic media used in token passing computer networks like FDDI.
  - IEEE 802.9: Integrated Services (IS) LAN Interface at the MAC and Physical Layers.
- IEEE 802.10: A standard for security functions that can be used in both LAN and MAN.
- IEEE 802.11: Wireless LAN MAC and physical layer
- IEEE 802.12: Demand Priority Access Method, Physical Layer and Repeater used in star topology, 100VG-AnyLAN, etc
- IEEE 802.16: Fixed broadband wireless LAN

Ethernet versions 1.0 and 2.0 followed after the original development until the IEEE 802.3 committee altered the structure of the Ethernet II packet to form the Ethernet 802.3 packet. You will currently see either Ethernet II (DIX) format or Ethernet 802.3 format being used.

IEEE 802.3 Ethernet uses Manchester Phase Encoding (MPE) for coding the data bits on the outgoing signal.

INTRODUCTION

Ethernet is a family of frame-based computer networking technologies for local area networks (LANs). The 'Ether' part of Ethernet denotes that the system is not meant to be restricted for use on only one medium type. Copper cables, fiber cables and even radio waves can be used.

Ethernet is standardized as IEEE 802.3. The combination of the twisted pair versions of Ethernet for connecting end systems to the network, along with the fiber optic versions for site backbones, is the most widespread wired LAN technology. It has largely replaced competing LAN standards such as token ring, FDDI, and ARCNET.

Ethernet defines a number of wiring and signaling standards for the Physical Layer of the OSI networking model, through means of network access at the Media Access Control (MAC) / Data Link Layer, and a common addressing format. Therefore, Ethernet defines the lower two layers of the OSI Reference Model (see Data Busses Section intro). Ethernet sometimes implies an attached protocol—such as TCP/IP, however TCP/IP really defines the transport and network layers, respectively, of the OSI model.

Ethernet was designed as a 'broadcast' system, i.e. stations on the network can send messages whenever and wherever they want. All stations may receive the messages, however only the specific station to which the message is directed responds. To handle simultaneous demands, the system uses Carrier Sense Multiple Access with Collision Detection (CSMA/CD).

CARRIER SENSE MULTIPLE ACCESS WITH COLLISION DETECTION (CSMA/CD)

When an Ethernet station is ready to transmit, it checks for the presence of a signal on the cable. If no signal is present then the station begins transmission. However, if a signal is present then the station delays transmission until the cable is not in use. If two stations detect an idle cable and at the same time transmit data, then a collision occurs. On a star-wired, unshielded twisted pair (UTP) network, if the transceiver of a sending station detects activity on both its receive and transmit pairs before it has completed transmitting, then it decides that a collision has occurred. On a coaxial system, a collision is detected when the DC signal level on the cable is the same or greater than the combined signal level of the two transmitters. Line voltage drops dramatically if two stations transmit at the same time and the first station to notice this sends a high voltage jamming signal around the network as a signal. The two stations involved with the collision quit transmitting again for a random time interval.

A Collision Domain is that part of the network where each station can 'see' other stations' traffic both unicast and broadcasts. The Collision Domain is made up of one segment of Ethernet coax (with or without repeaters) or a number of UTP shared hubs. A network is segmented with bridges that create two segments, or two Collision Domains where a station on one segment cannot see traffic between stations on the other segment unless the packets are destined for itself. If it can still see all broadcasts on the segmented network(s), no matter the number of segments, it is a Broadcast Domain.

ETHERNET FRAME

Figure 1 depicts the structure of the older DIX (Ethernet II) and the now standard 802.3 Ethernet frames. The numbers above each field represent the number of bytes.

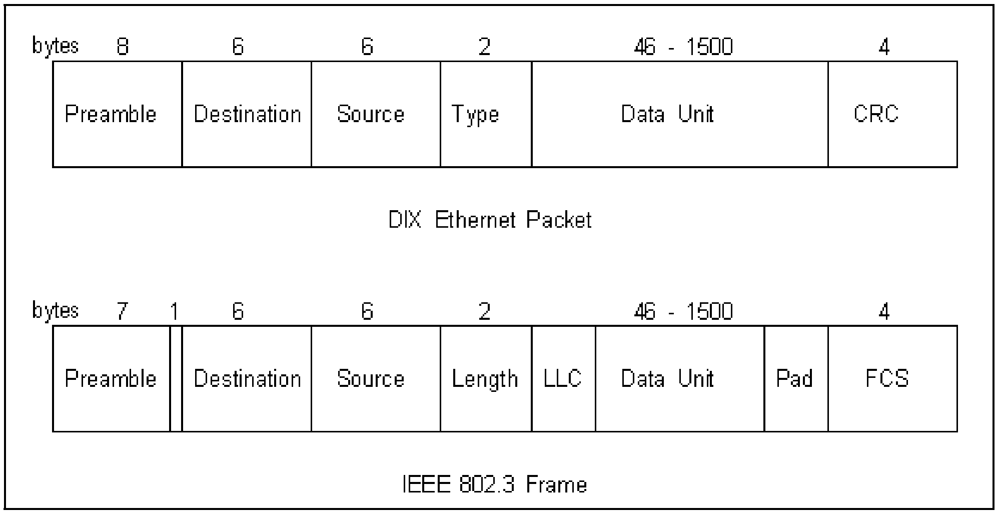

Figure 1. Ethernet Frame.

- **Preamble field:** Establishes bit synchronization and transceiver conditions so that the circuitry synchs with the received frame timing. The DIX frame has 8 bytes for the preamble rather than 7, as it does not have a Start Frame Delimiter (or Start of Frame).
- **Start Frame Delimiter:** Sequence 10101011 in a separate field, only in the 802.3 frame.
- **Destination address:** Hardware address (MAC address) of the destination station (usually 48 bits i.e. 6 bytes).
- **Source address:** Hardware address of the source station (must be of the same length as the destination address, the 802.3 standard allows for 2 or 6 byte addresses, although 2 byte addresses are never used, N.B. Ethernet II *only* uses 6 byte addresses).
- **Type:** Specifies the protocol sending the packet such as IP or IPX (only applies to DIX frame).
- **Length:** Specifies the length of the data segment, actually the number of LLC data bytes, (only applies to 802.3 frame and replaces the Type field).
- **LLC:** Logical Length Control is a data communication protocol layer which provides multiplexing and flow control mechanisms that make it possible for several network protocols (IP, IPX) to coexist within a multipoint network and to be transported over the same network media.
- **Data:** Actual data which is allowed anywhere between 46 to 1500 bytes within one frame.
- **Pad:** Zeros added to the data field to ‘Pad’ a short data field to 46 bytes (applies to 802.3 frame).
- **CRC:** Cyclic Redundancy Check to detect errors during transmission (DIX version of FCS).
- **FCS:** Frame Check Sequence to detect errors that occur during transmission (802.3 version of CRC). This 32-bit code has an algorithm applied to it, which will give the same result as the other end of the link, provided that the frame was transmitted successfully.

From the above we can see that the maximum 802.3 frame size is 1518 bytes and the minimum size is 64 bytes. Packets that have correct CRCs (or FCSs) but are smaller than 64 bytes are known as ‘Runts’.

The hardware address, or MAC address is transmitted and stored in Ethernet network devices in **Canonical** format i.e. Least Significant Bit (LSB) first. The expression **Little-Endian** is used to describe the LSB format in which Ethernet is transmitted. On the other hand, Token Ring networks and the subset, Fiber

Distributed Data Interface (FDDI) networks, transmit the MAC address with the Most Significant Bit (MSB) first, or **Big-Endian**, This is known as **Non-Canonical** format. Note that this applies on a byte-by-byte basis i.e. the bytes are transmitted in the same order, it is just the bits in each of those bytes that are reversed. The storage of the MAC addresses in Token Ring and FDDI devices however, may sometimes still be in Canonical format so this can sometimes cause confusion. The reference to, the distribution of MAC addresses and the Organizationally Unique Identifier (OUI) destinations are always in Canonical format.

The Logical Link Control (LLC) protocol data unit (PDU) is defined in IEEE 802.2 and operates with 802.3 Ethernet as seen in Figure 2.

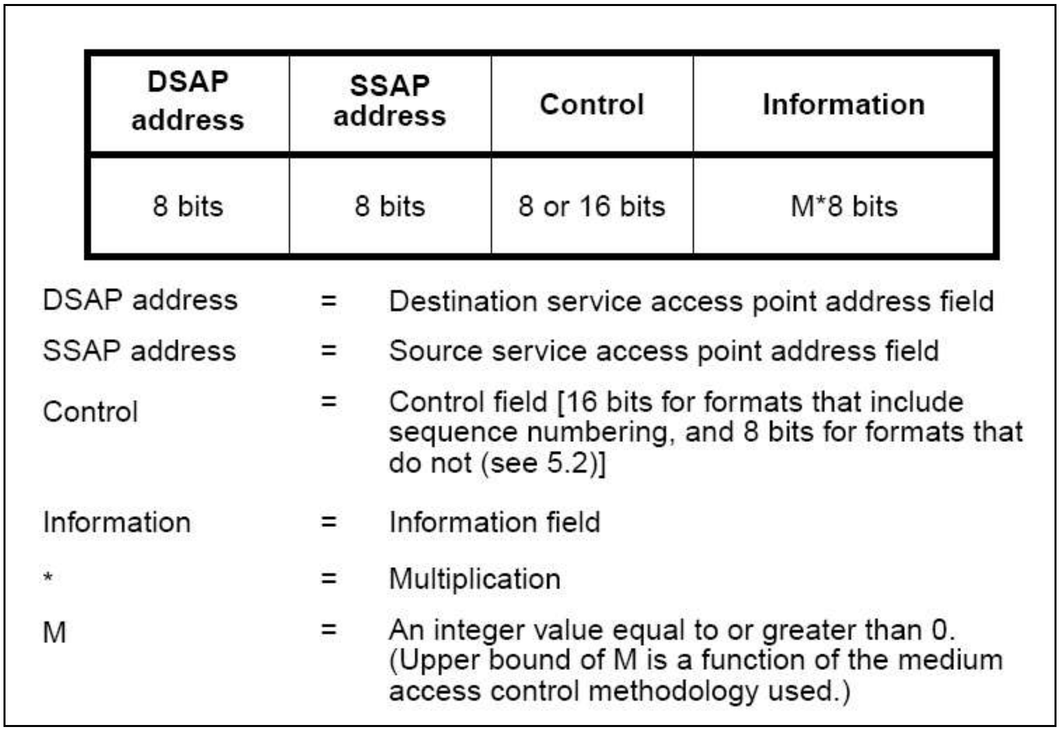

DSAP address	SSAP address	Control	Information
8 bits	8 bits	8 or 16 bits	M*8 bits

DSAP address = Destination service access point address field

SSAP address = Source service access point address field

Control = Control field [16 bits for formats that include sequence numbering, and 8 bits for formats that do not (see 5.2)]

Information = Information field

* = Multiplication

M = An integer value equal to or greater than 0. (Upper bound of M is a function of the medium access control methodology used.)

Figure 2. LLC Format.

LLC is based on the High-Level Data Link Control (HDLC) data protocol format. Whereas Ethernet II (2.0) combines the MAC and the Data link layers restricting itself to connectionless service in the process, IEEE 802.3 separates out the MAC and Data Link layers. IEEE 802.2 is also required by Token Ring and FDDI but cannot be used with the Novell 'Raw' format. There are three types of LLC; Type 1, which is connectionless, Type 2 which is connection-oriented, and Type 3 for Acknowledged Connections.

The **Service Access Point (SAP)** is used to distinguish between different data exchanges on the same end station and basically replaces the Type field for the older Ethernet II frame. The **Source Service Access Point (SSAP)** indicates the service from which the LLC data unit is sent, and the **Destination Service Access Point (DSAP)** indicates the service to which the LLC data unit is being sent. As examples, NetBIOS uses the SAP address of **F0** whilst IP uses the SAP address of **06**. The following lists some common SAPs:

00 - Null LSAP
02 - Individual LLC Sublayer Management Function
03 - Group LLC Sublayer Management Function
05 - IBM SNA Path Control (group)

06 - Internet Protocol (IP)
18 - Texas Instruments
42 - IEEE 802.1 Bridge Spanning Tree Protocol
80 - Xerox Network Systems (XNS)
86 - Nestar
AA - SubNetwork Access Protocol (SNAP)
E0 - Novell NetWare
F0 - IBM NetBIOS
F5 - IBM LAN Management (group)
FE - ISO Network Layer Protocol

The **Control Field** identifies the type of LLC, of which there are three:

- **Type 1** - Data (PDUs) shall be exchanged between LLCs without the need for the establishment of a data link connection. These PDUs shall not be acknowledged, nor shall there be any flow control or error recovery. This is called **Unsequenced Information (UI)**.
- **Type 2** – Needs data link connection. Uses **Information (I)** frames and maintains the sequence numbers during an acknowledged connection-oriented transmission.
- **Type 3** – No data link connection required. Uses **Acknowledged Connection (AC)** frames in an acknowledged connectionless service.

MAC ADDRESS

With an Ethernet MAC address, the first octet uses only the lowest significant bit as the Individual/Group (I/G) bit address identifier. If it is zero it indicates the address is individual. If it is one, it indicates a group address. The Universally or Locally administered (U/L) address bit is the bit of octet 0 adjacent to the I/G address bit. This bit indicates whether the address has been assigned by a local or universal administrator. Universally administered addresses have this bit set to 0. If this bit is set to 1, the entire address (i.e., 48 bits) has been locally administered.

The first 3 octets of the MAC address form the Organizational Unique Identifier (OUI) assigned by IEEE to organizations that requires their own group of MAC addresses.

SUBNETWORK ACCESS PROTOCOL (SNAP)

The SNAP protocol was introduced to allow an easy transition to the new LLC frame format for vendors. SNAP allows older frames and protocols to be encapsulated in a Type 1 LLC header making any protocol ‘pseudo-IEEE compliant’. SNAP is described in RFC 1042. Figure 3 shows how it looks:

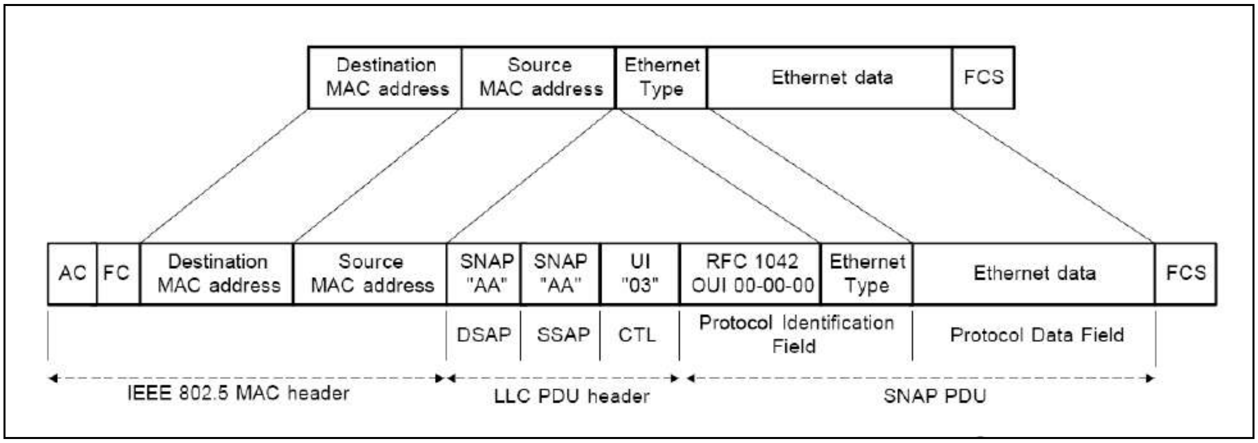
Destination MAC address
Source MAC address
Ethernet Type
Ethernet data
FCS
AC
FC
Destination MAC address
Source MAC address
SNAP "AA"
SNAP "AA"
UI "03"
RFC 1042 OUI 00-00-00
Ethernet Type
Ethernet data
FCS
DSAP
SSAP
CTL
Protocol Identification Field
Protocol Data Field
IEEE 802.5 MAC header
LLC PDU header
SNAP PDU

Figure 3. Ethernet Frame SNAP-encapsulated in IEEE 802.5 Frame.

It can be seen that it is an LLC data unit (sometimes called a Logical Protocol Data Unit (LPDU)) of Type 1 (indicated by 03). The DSAP and SSAP are set to AA to indicate that this is a SNAP header coming up. The SNAP header then indicates the vender via the Organizational Unique Identifier (OUI) and the protocol type via the Ethertype field. In the example above the OUI is 00-00-00 which means that there is an Ethernet frame. An ethernet type (Ethertype) of 08-00 would indicate Internet IP is the protocol. The official list of types can be found at IEEE. Although vendors tend to be are moving to LLC1 on the LAN, SNAP still remains.

TRANSMISSION MEDIA

10Base-5 - largely obsolete, but may still be in use.

Traditionally, Ethernet is used over ‘thick’ coaxial cable called 10Base-5 (the ‘10’ denotes 10Mbps, base means that the signal is baseband i.e. takes the whole bandwidth of the cable (so that only one device can transmit at one time on the same cable), and the ‘5’ denotes 500 m maximum length). The minimum length between stations is 2.5 m.

The cable is run in one long length forming a ‘Bus Topology’. Stations attach to it by way of inline N-type connections or a transceiver, which is literally screwed into the cable (by way of a ‘Vampire Tap’) providing a 15-pin Attachment Unit Interface (AUI) connection (also known as a DIX connector or a DB-15 connector) for a drop lead connection (maximum of 50 m length) to the station. The segments are terminated with 50 ohm resistors and the shield grounded at one end only.

The segment could be appended with up to a maximum of 4 repeaters; therefore 5 segments (total length of 2,460 m) can be connected together. Of the 5 segments only 3 can have devices attached (100 per segment). This is known as the 5-4-3 rule. A total of 300 devices can be attached on a Thicknet broadcast domain.

10Base-2 – dominant for many years

It was common to see the Thick coax used in Risers to connect Repeaters, which in turn provide 'Thin Ethernet' coaxial connections for runs up to 30 workstations. Thin Ethernet (Thinnet) uses RG-58 cable and is called 10Base-2 (The '2' now denoting 200 m maximum length - strictly speaking this is 185 m). The minimum length between stations is 0.5 m.

Each station connects to the thinnet by way of a Network Interface Card (NIC), which provides a BNC (British Naval Connector). At each station the thinnet terminates at a T-piece and at each end of the thinnet run (or 'Segment') a 50-ohm terminator is required to absorb stray signals, thereby preventing signal bounce. The shield is grounded at one end only.

A segment can be appended with other segments using up to 4 repeaters, i.e. 5 segments in total. Two of these segments however, cannot be tapped; they can only be used for extending the length of the broadcast domain (to 925 m). What this means is that 3 segments with a maximum of 30 stations on each can give you 90 devices on a Thinnet broadcast domain.

10Base-T – Most widespread today

The most common media used in of Ethernet today involves use of Unshielded Twisted Pair (UTP) or Shielded Twisted Pair (STP). For instance, Category 5 UTP can be installed in a 'Star-wired' format, with runs recommended at no greater than 100 m (including patch leads, cable run and flyleads) and Ethernet Hubs with UTP ports (RJ45) centrally located. It has been found though that runs of up to 150 m are feasible, the limitations being signal strength. Also, there should be no more than a 11.5 dB signal loss and the minimum distance between devices is 2.5 m. The maximum delay for the signal in a 10Mbps network is 51.2 microseconds. This comes from the fact that the bit time (time to transmit one bit) is 0.1 microseconds and that the slot time for a frame is 512 bit times.

In order to connect to Ethernet in this 'Star Topology', each station has a NIC, which contains an RJ45 socket which is used by a 4-pair droplead to connect to another nearby RJ45 socket.

Each port on the hub sends a 'Link Beat Signal' which checks the integrity of the cable and devices attached, a flickering LED on the front of the port of the hub tells you that the link is running fine. The maximum number of hubs (or, more strictly speaking, repeater counts) that you can have in one segment is four and the maximum number of stations on one broadcast domain is 1024.

The advantages of the UTP/STP technology are gained from the flexibility of the system, with respect to moves, changes, fault finding, reliability, and security. The following table shows the RJ45 pin-outs for 10Base-T:

RJ45 Pin	Function	Color
1	Transmit	White/Orange stripe
2	Transmit	Orange
3	Receive	White/Green stripe
4		Blue
5		White/Blue stripe
6	Receive	Green
7		White/Brown stripe
8		Brown

If you wish to connect hub to hub, or a NIC directly to another NIC, then the following 10Base-T crossover cable should be used:

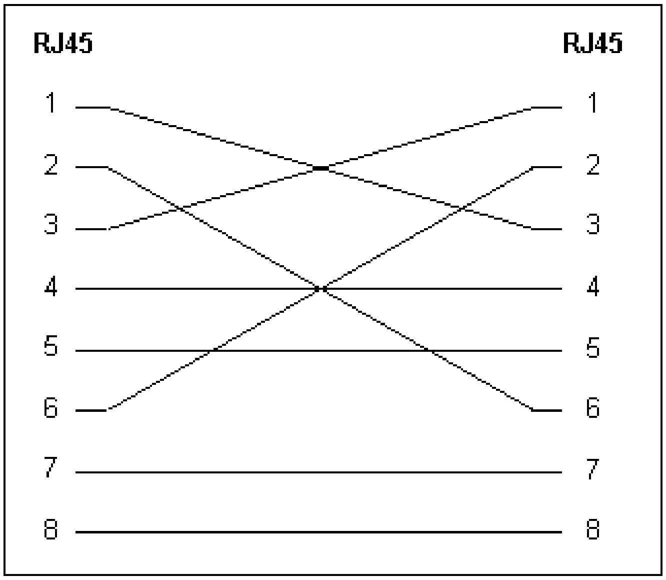

Figure 4. 10Base-T Crossover.

The 4-repeater limit manifests itself in 10/100Base-T environments where the active hub/switch port is in fact a repeater, hence the name multi-port repeater. Generally, the hub would only have one station per port but you can cascade hubs from one another up to the 4-repeater limit. The danger here is that you will have all the traffic from a particular hub being fed into one port so care would need to be taken on noting the applications being used by the stations involved, and the likely bandwidth that the applications will use.

There is a pre-standard called Lattisnet (developed by Synoptics), which runs 10 MHz Ethernet over twisted pair but instead of bit synchronization occurring at the sending (as in 10Base-T) the synchronization occurs at the receiving end.

10Base-F

The 10Base-F standard defines the use of fiber for Ethernet. 10Base-FL describes the standards for the fiber optic links between stations and repeaters, allowing up to 2 km per segment on multi-mode fiber. It replaces the 1987 FOIRL (Fiber Optic Inter-Repeater Link), which provides the specification for a fiber optic Media Attachment Unit (MAU) and other interconnecting components. Two related standards are essentially "dead": 10Base-FB allows up to 2 km per segment (on multi-mode fiber) and is designed for backbone applications such as cascading repeaters. In addition, there is the 10Base-FP (Passive components) standard

The 10Base-F standard allows for 1024 devices per network.

Fast Ethernet - 100Base-T

Fast Ethernet is the most popular of the newer standards and is an extension to 10Base-T. The '100' denotes 100 Mbps data speed and it uses the same two pairs as 10Base-T and must only be used on Category 5 UTP cable installations with provision for it to be used on Type 1 STP. The actual data throughput increases by between 3 to 4 times that of 10Base-T.

Fast Ethernet is specified in 802.3u and uses the same frame formats and CSMA/CD technology as normal 10Mbps Ethernet. The difference is that the maximum delay for the signal across the segment is 5.12 microseconds instead of 51.2 microseconds. This stems from the fact that the bit time (time to transmit one bit) is 0.01 microseconds and that the slot time for a frame is 512 bit times. The Inter-Packet Gap (IPG) for 802.3u is 0.96 microseconds as opposed to 9.6 microseconds for 10Mbps Ethernet.

Whereas 10Base-T uses Normal Link Pulses (NLP) for testing the integrity of the connection, 100Base-T uses Fast Link Pulses (FLP) which are backwardly compatible with NLPs but contain more information. FLPs are used to detect the speed of the network (e.g. in 10/100 switchable cards and ports). The distance must not exceed 100 m.

The IEEE uses the term 100Base-X to refer to both 100Base-Tx and 100Base-Fx and the Media-Independent Interface (MII) allows a generic connector for transceivers to connect to 100Base-Tx, 100Base-Fx and 100Base-T4 LANs.

There is no such thing as the 5-4-3 rule in Fast Ethernet. All 10Base-T repeaters are considered to be functionally identical. Fast Ethernet repeaters are divided into two classes of repeater: Class I and Class II. A Class I repeater has a repeater propagation delay value of 140 bit times, whilst a Class II repeater is 92 bit times. The Class I repeater (or Translational Repeater) can support different signaling types such as 100Base-Tx and 100Base-T4. A Class I repeater transmits or repeats the incoming line signals on one port to the other ports by first translating them to digital signals and then retranslating them to line signals. The translations are necessary when connecting different physical media (media conforming to more than one physical layer specification) to the same collision domain. Any repeater with an MII port would be a Class I device. Only one Class I repeater can exist within a single collision domain, so this type of repeater cannot be cascaded.

A Class II repeater immediately transmits or repeats the incoming line signals on one port to the other ports: it does not perform any translations. This repeater type connects identical media to the same collision domain (for example, TX to TX). At most, two Class II repeaters can exist within a single collision domain.

The cable used to cascade the two devices is called and unpopulated segment or IRL (Inter-Repeater Link). The Class II repeater (or **Transparent Repeater**) can only support one type of physical signaling.

100Base-T4

100Base-T4 was an early implementation of Fast Ethernet. It requires four twisted copper pairs, but those pairs were only required to be category 3 rather than the category 5 required by TX. One pair is reserved for transmit, one for receive, and the remaining two will switch direction as negotiated. A very unusual 8B6T code is used to convert 8 data bits into 6 base-3 digits (the signal shaping is possible as there are three times as many 6-digit base-3 numbers as there are 8-digit base-2 numbers). The two resulting 3-digit base-3 symbols are sent in parallel over 3 pairs using 3-level pulse-amplitude modulation (PAM-3).

100Base-T2

This little known version of Fast Ethernet is for use over two pairs of Category 3 cable and uses PAM-5 for encoding. There is simultaneous transmission and reception of data in both pairs and the electronics uses DSP technology to handle alien signals in adjacent pairs. 100Base-T2 can run up to 100 m on Category 3 UTP.

100Base-Fx

100Base-Fx is a version of Fast Ethernet over optical fiber. It uses a 1300 nm near-infrared (NIR) light wavelength transmitted via two strands of optical fiber, one for receive (RX) and the other for transmit (TX). Maximum length is 400 meters (1,310 ft) for half-duplex connections (to ensure collisions are detected) or 2 kilometers (6,600 ft) for full-duplex over multimode optical fiber. Longer distances are possible when using single-mode optical fiber. 100Base-Fx uses the same 4B5B encoding and NRZI line code that 100Base-TX does. 100Base-Fx should use the following connectors: SC, straight tip (ST), or media independent connectors (MIC), with SC being the preferred option. 100Base-Fx is not compatible with 10Base-Fl, the 10 MBit/s version over optical fiber.

100Base-Sx

100Base-Sx is a version of Fast Ethernet over optical fiber. It uses two strands of multi-mode optical fiber for receive and transmit. It is a lower cost alternative to using 100Base-Fx, because it uses short wavelength optics (850 nm) which are significantly less expensive than the long wavelength optics used in 100Base-Fx. 100Base-Sx can operate at distances up to 300 meters (980 ft).

100Base-Sx uses the same wavelength as 10Base-Fl, the 10 MBit/s version over optical fiber. Unlike 100Base-Fx, this allows 100Base-Sx to be backwards-compatible with 10Base-Fl. Because of the shorter wavelength used and the shorter distance it can support, 100Base-Sx uses less expensive optical components (LEDs instead of lasers) which makes it an attractive option for those upgrading from 10Base-Fl and those who do not require long distances.

100Base-Bx

100Base-Bx is a version of Fast Ethernet over a single strand of optical fiber (unlike 100Base-Fx, which uses a pair of fibers). Single-mode fiber is used, along with a special multiplexer which splits the signal into transmit and receive wavelengths. The two wavelengths used for transmit and receive are either 1310/1550 nm or 1310/1490 nm. Distances can be 10, 20, or 40 km.

100VG-AnyLAN

Based on 802.12 (Hewlett Packard), 100VG-AnyLAN uses an access method called Demand Priority. The 'VG' stands for 'Voice Grade' as it is designed to be used with Category 3 cable. This is where the repeaters (hubs) carry out continuous searches around all of the nodes for those that wish to send data. If two devices cause a 'contention' by wanting to send at the same time, the highest priority request is dealt with first, unless the priorities are the same, in which case both requests are dealt with at the same time (by alternating frames). The hub only knows about connected devices and other repeaters so communication is only directed at them rather than broadcast to every device in the broadcast domain (which could mean 100's of devices!). This is a more efficient use of the bandwidth. This is the reason why the new standard was developed as it is not strictly Ethernet. Standard 802.12 is designed to better support both Ethernet and Token Ring.

The encoding techniques used are 5B/6B and NRZ. All four pairs of UTP are used. On Cat3 the longest cable run is 100 m but this increases to 200 m on Cat5. The clock rate on each wire is 30 MHz; therefore 30 Mbits per second are transmitted on each pair giving a total bit rate of 120 Mbps. Since each 6-bits of data on the line represents 5 bits of real data due to the 5B/6B encoding, the rate of real data being transmitted is 25 Mbps on each pair, giving a total real data rate of 100 Mbps. For 2-pair STP and fiber, the bit rate is 120 Mbps on the transmitting pair, for a real data transmission rate of 100 Mbps.

Gigabit Ethernet

Although the functional principles of Gigabit Ethernet are the same as Ethernet and Fast Ethernet i.e. CSMA/CD and the Framing format, the physical setup is very different. One difference is the slot time. The standard Ethernet slot time required in CSMA/CD half-duplex mode is not long enough for running over 100 m of copper, so Carrier Extension is used to guarantee a 512-bit slot time.

1000Base-X (802.3z)

The 802.3z committee is responsible for formalizing the standard for Gigabit Ethernet. The 1000 refers to 1 Gbps data speed. The existing Fiber Channel interface standard (ANSI X3T11) is used and allows up to 4.268 Gbps speeds. The Fiber Channel encoding scheme is 8B/10B.

Gigabit Ethernet can operate in half or full duplex modes and there is also a standard 802.3x which manages XON/XOFF flow control in full duplex mode. With 802.3x, a receiving station can send a packet to a sending station to stop it sending data until a specified time interval has passed.

There are three media types for 1000Base-X. 1000Base-Sx, 1000Base-Lx, and 1000Base-Cx.

1000Base-Sx ('S' is for Short Haul) uses short-wavelength laser (850 nm) over multi-mode fiber. 1000Base-Sx can run up to 300 m on 62.5/125 μm multimode fiber and up to 550 m on 50/125 μm multimode fiber.

1000Base-Lx ('L' is for Long Haul) uses long wavelength laser (1300 nm) and can run up to 550 m on 62.5/125 μm multi-mode fiber or 50/125 μm multi-mode fiber. In addition, 1000Base-Lx can run up to 5 km (originally 3 km) on single-mode fiber using 1310 nm wavelength laser.

1000Base-Cx is a standard for STP copper cable and allows operation up to 25 m over STP cable.

1000Base-T (802.3ab)

Many cable manufacturers are enhancing their cable systems to 'enhanced Category 5' (Cat 5e) standards in order to allow Gigabit Ethernet to run at up to 100 m on copper. The Category 6 standard has yet to be ratified, and is not being pursued very strongly.

In order to obtain the 1000 Mbps data bit rate across the UTP cable without breaking the FCC rules for emission, all 4 pairs of the cable are used. Hybrid circuits at each end of each pair are used to allow simultaneous transmission and reception of data (full-duplex) by separating the transmission signal from the receiving signal. Because some transmission signal still manages to couple itself to the receiving side there is an additional echo canceller built in, this is called a Near End Crosstalk (NEXT) canceller.

Encoding is carried out with PAM-5.

TRANSMISSION CONTROL PROTOCOL / INTERNET PROTOCOL

Transmission **C**ontrol **P**rotocol / **I**nternet **P**rotocol (TCP/IP) is the communication protocol for the Internet.

Whereas IP handles lower-level transmissions from computer to computer as a message makes its way across the Internet, TCP operates at a higher level, concerned only with the two end systems, for example a Web browser and a Web server. In particular, TCP provides ordered delivery of a stream of bytes from a program on one computer to another program on another computer. Besides the Web, other common applications of TCP include e-mail and Internet file transfer like FTP. Among its other management tasks, TCP controls message size, the rate at which messages are exchanged, and network traffic congestion. Your computer Internet address is a part of the standard TCP/IP protocol and so is your domain name.

Inside the TCP/IP standard there are several protocols for handling data communication:

- TCP (Transmission Control Protocol) communication between applications
- UDP (User Datagram Protocol) simple communication between applications
- IP (Internet Protocol) communication between computers
- ICMP (Internet Control Message Protocol) for errors and statistics
- DHCP (Dynamic Host Configuration Protocol) for dynamic addressing

TCP: If one application wants to communicate with another via TCP, it sends a communication request. This request must be sent to an exact address. TCP is responsible for breaking data down into IP packets before they are sent, and for assembling the packets when they arrive. After a "handshake" between the two applications, TCP will set up a "full-duplex" communication between the two applications. The "full-duplex" communication will occupy the communication line between the two computers until it is closed by one of the two applications. UDP is very similar to TCP, but simpler and less reliable.

IP: IP is responsible for sending the packets to the correct destination. IP does not occupy the communication line between two computers and hence reduces the need for network lines. Each line can be used for communication between many different computers at the same time. With IP, messages (or other data) are broken up into small independent "packets" and sent between computers via the Internet. IP is responsible for "routing" each packet to the correct destination.

When an IP packet is sent from a computer, it arrives at an IP router. The IP router is responsible for "routing" the packet to the correct destination, directly or via another router. The path the packet will follow might be different from other packets in the same communication chain. The router is responsible for the right addressing, depending on traffic volume, errors in the network, or other parameters.

Each computer must have a unique IP address before it can connect to the Internet. Each IP packet must also have an address before it can be sent to another computer. TCP/IP uses 32 bits (4 bytes), of four numbers between 0 and 255, separated by a period, to address a computer. A typical IP address is of the format: 192.255.30.50

Domain Names: A name is much easier to remember than a 12 digit number. The names used for TCP/IP addresses are called domain names. "**navy.mil**" is a domain name. Subdomains are identified by prefixes to the main domain name (separated by a period). When you address a web site, like **http://www.navair.navy.mil** the name is translated to a number by a Domain Name System (DNS) Server. DNS servers are connected to the Internet all over the world. When a new domain name is registered together with a TCP/IP address, all DNS servers are updated with this information.

BROWSER AND OTHER APPLICATION LAYER PROTOCOLS

DHCP - Dynamic Host Configuration Protocol. DHCP is used for allocation of dynamic IP addresses to computers in a network.

FTP - File Transfer Protocol. FTP takes care of transmission of data files between computers.

HTTP - Hyper Text Transfer Protocol - takes care of the communication between a web server and a web browser. It is used for sending requests from a web client (a browser) to a web server, returning web content (web pages) from the server back to the client.

HTTPS - Secure HTTP. HTTPS typically handles credit card transactions and other sensitive data.

IMAP - Internet Message Access Protocol. The IMAP protocol is used by email programs (like Microsoft Outlook) just like the POP protocol (discussed below). The main difference between the IMAP protocol and the POP protocol is that the IMAP protocol will not automatically download all your emails each time your email program connects to your email server. The IMAP protocol allows you to look through your email messages at the email server before you download them. With IMAP you can choose to download your messages or just delete them. This way IMAP is useful if you need to connect to your email server from different locations, but only want to download your messages when you are back in your office.

NTP - Network Time Protocol. NTP is used to synchronize the time (the clock) between computers.

POP - Post Office Protocol. The POP protocol is used by email programs (like Microsoft Outlook) to retrieve emails from an email server. If your email program (also called email client), uses POP, all your emails are downloaded to your computer each time it connects to your email server.

SMTP - Simple Mail Transfer Protocol. The SMTP protocol is used for the transmission of e-mails to another computer. Normally your email is sent to an email server (SMTP server), and then to another server or servers, and finally to its destination. SMTP can only transmit pure text. It cannot transmit binary data like pictures, sounds, or movies directly. SMTP uses the MIME protocol to send binary data across TCP/IP networks.

SNMP - Simple Network Management Protocol. SNMP is used for administration of computer networks.

PRESENTATION LAYER PROTOCOLS

MIME - Multi-purpose Internet Mail Extensions protocol. Converts binary data to pure text. The MIME protocol lets SMTP transmit multimedia files including voice, audio, and binary data.

SSL - Secure Sockets Layer. The SSL protocol is used for encryption of data for secure data transmission.

LOWER LAYER PROTOCOLS

ARP - Address Resolution Protocol. ARP is used by IP to find the hardware address of a computer network card based on the IP address.

BOOTP - Boot Protocol. BOOTP is used for booting (starting) computers from the network.

ICMP - Internet Control Message Protocol. ICMP takes care of error-handling in the network.

LDAP - Lightweight Directory Access Protocol. LDAP is used for collecting information about users and e-mail addresses from the internet.

PPTP - Point to Point Tunneling Protocol. PPTP is used for setting up a connection (tunnel) between private networks.

RARP - Reverse Address Resolution Protocol. RARP is used by IP to find the IP address based on the hardware address of a computer network card.

This page intentionally left blank.

GLOSSARY

ACCEPTABLE DEGRADATION - The allowable reduction in system performance. For a fire control radar, the acceptable degradation is usually expressed as a reduction in range; for example, the maximum lock-on range might be degraded by 25 percent without loss of essential defense capability.

ACQUISITION - A procedure by which a fire control tracking radar attains initial lock-on. Usually, the approximate target coordinates are supplied to the tracking radar and it searches a predetermined volume of space to locate the target.

AEROSOLS - Solid particles dispersed in the atmosphere having resonant size particles with a high index of refraction. The particles both scatter and absorb visual and laser directed energy so as to cut down on weapon systems directed by these techniques.

AFC (AUTOMATIC FREQUENCY CONTROL) - An arrangement whereby the frequency of an oscillator or the tuning of a circuit is automatically maintained within specified limits with respect to a reference frequency. A magnetron drifts in frequency over a period of time. The AFC of a radar makes the local oscillator shift by an equal amount so the IF frequency will remain constant.

AGC (AUTOMATIC GAIN CONTROL) - A method for automatically obtaining an essentially constant receiver output amplitude. The amplitude of the received signal in the range gate determines the AGC bias (a DC voltage) which controls the receiver gain so as to maintain a nearly constant output even though the amplitude of the input signal changes.

AMPLIFIER - An electronic device used to increase signal magnitude or power. See also GaAs FET Amplifier, Klystron Amplifier, Traveling-Wave Tube Amplifier.

AMPLITUDE MODULATION (AM) - A method of impressing a message upon a carrier signal by causing the carrier amplitude to vary proportionally to the message waveform.

AMPLITUDE SHIFT KEYING (ASK) - A method of impressing a digital signal upon a carrier signal by causing the carrier amplitude to take different values corresponding to the different values of the digital signal.

ANGLE JAMMING - ECM technique, when azimuth and elevation information from a scanning fire control radar is jammed by transmitting a jamming pulse similar to the radar pulse, but with modulation information out of phase with the returning target angle modulation information.

ANGULAR SEPARATION - This term is frequently used to indicate a protective (from EMI) zone for a missile. The interfering antenna axis must be separated, throughout the critical portion of the missile flight, from the missile by the specified angle. The vertex of the angle is at the interference source antenna.

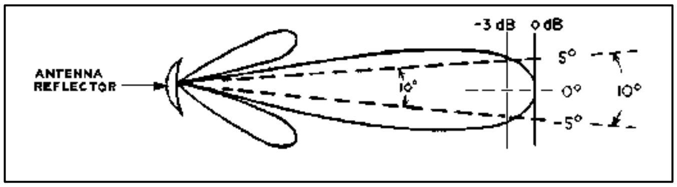

ANTENNA BEAMWIDTH - The angle, in degrees, between the half-power points (-3 dB) of an antenna beam. This angle is also nearly that between the center of the mainlobe and the first null. The angle is given for both horizontal and vertical planes unless the beam is circular. When so indicated, the term may refer to the angular width of the mainlobe between first nulls [beamwidth between first nulls (BWFN)]. See also Antenna Pattern. The figure illustrates vertical profile for antenna displaying a 10-degree beamwidth characteristic. The values can vary dramatically with frequency.

ANTENNA CROSS TALK - A measure of undesired power transfer through space from one antenna to another. Ratio of power received by one antenna to power transmitted by the other, usually expressed in decibels.

ANTENNA ISOLATION - The ratio of the power input to one antenna to the power received by the other. It can also be viewed as the insertion loss from transmit antenna input to receive antenna output to circuitry.

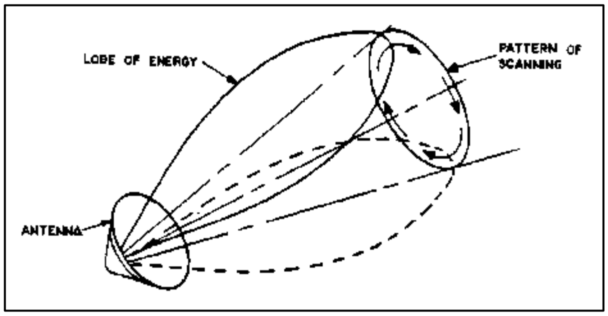

ANTENNA LOBING - Two lobes are created that overlap and intercept at -1 to -3 dB. The difference between the two lobes produces much greater spatial selectivity than provided by either lobe alone. (See also Lobe, Antenna.)

ANTENNA NUTATING - An antenna, as used in automatic-tracking radar systems, consisting of a parabolic reflector combined with a radiating element which is caused to move in a small circular orbit about the focus of the antenna with or without change of polarization. The radiation pattern is in the form of a beam that traces out a cone centered on the reflector axis. The process is also known as nutating conical scanning.

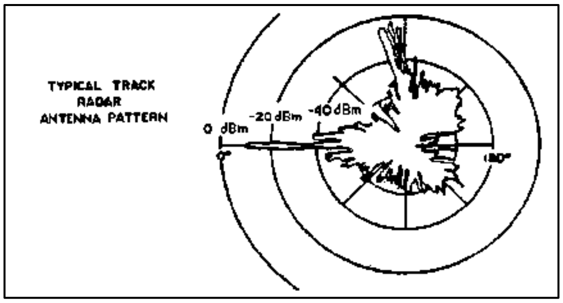

ANTENNA PATTERN - A cross section of the radiating pattern (representing antenna gain or loss) in any plane that includes the origin (source reference point) of the pattern. Both horizontal and vertical polar plots are normally used to describe the pattern. Also, termed “polar diagram” and “radiation pattern.”

ANTENNA, PENCIL-BEAM - A highly directional antenna designed that cross sections of the major lobe are approximately circular, with a narrow beamwidth.

ANTI-CLUTTER CIRCUITS (IN RADAR) - Circuits which attenuate undesired reflections to permit detection of targets otherwise obscured by such reflections.

APERTURE - In an antenna, that portion of the plane surface area near the antenna perpendicular to the direction of maximum radiation through which the major portion of the radiation passes. The effective and/or scattering aperture area can be computed for wire antennas which have no obvious physical area.

A-SCOPE - A cathode-ray oscilloscope used in radar systems to display vertically the signal amplitude as a function of time (range) or range rate. Sometimes referred to as Range (R)-Scope.

ASYNCHRONOUS PULSED JAMMING - An effective form of pulsed jamming. The jammer nearly matches the pulse repetition frequency (PRF) of the radar; then it transmits multiples of the PRF. It is more effective if the jammer pulsewidth is greater than that of the radar. Asynchronous pulsed jamming is similar to synchronous jamming except that the target lines tend to curve inward or outward slightly and appear fuzzy in the jammed sector of a radar scope.

ATTENUATION - Decrease in magnitude of current, voltage, or power of a signal in transmission between two points. May be expressed in decibels.

AUTOMATIC FREQUENCY CONTROL - See AFC.

AUTOMATIC GAIN CONTROL - See AGC.

BACKWARD WAVE OSCILLATOR (BWO) - A cross-field device in which an electron stream interacts with a backward wave on a nonreentrant circuit. This oscillator may be electronically tuned over a wide range of frequencies, is relatively unaffected by load variations and is stable. BWO is commonly pronounced “be woe”.

BALANCED MIXERS - The two most frequently encountered mixer types are single-balanced and double-balanced. In a double-balanced mixer, four Schottky diodes and two wideband transformers are employed to provide isolation of all three ports.

BALLISTIC MISSILE - Any missile which does not rely upon aerodynamic surfaces to produce lift and consequently follows a ballistic trajectory when thrust is terminated.

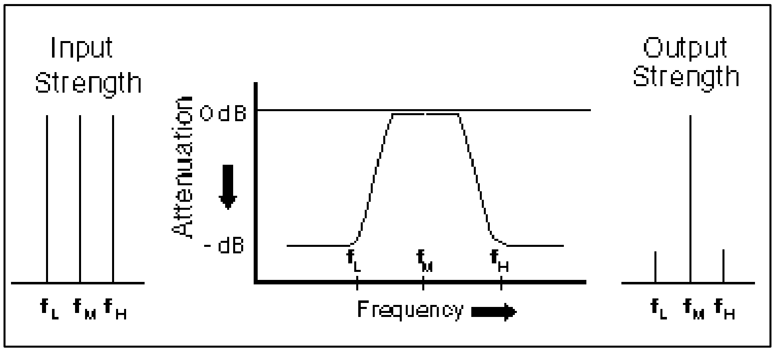

BANDPASS FILTER - A type of frequency discrimination network designed to pass a band or range of frequencies and produce attenuation to all other frequencies outside of the pass region. The figure illustrates a typical bandpass filter, incorporating a bandpass region of (F_h)-(F_l), offering no rejection (0 dB) to desired signal (F_m) and much higher rejection to the adjacent undesired signals F_h, and F_l. The upper and lower frequencies are usually specified to be the half power (-3 dB) or half voltage points (-6 dB).

BANDWIDTH - An expression used to define the actual operational frequency range of a receiver when it is tuned to a certain frequency. For a radar receiver, it is the difference between the two frequencies at which the receiver response is reduced to some fraction of its maximum response (such as 3 dB, 6 dB, or some other specified level). The frequencies between which “satisfactory” performance is achieved. Two equations are used:

$$\textit{Narrowband by \%}\ \left(\frac{F_u - F_l}{F_c}\right)(100) \quad ; \quad \textit{Broadband by ratio}\ \frac{F_u}{F_l}$$

$$\textit{Where}\ F_u = \textit{Upper}\ ;\ F_l = \textit{lower}\ ;\ F_c = \textit{center} = (\ F_u + F_l\ ) \div 2$$

See also Receiver Bandwidth and Spectrum Width.

BARRAGE NOISE JAMMING - Noise jamming spread in frequency to deny the use of multiple radar frequencies to effectively deny range information. Although this is attractive because it enables one jammer to simultaneously jam several radars of different frequencies, it does have the inherent problem that the wider the jamming spread, the less jamming power available per radar, i.e. the watts per MHz bandwidth is low.

BATTERY, MISSILE - A missile battery consists of a missile launcher and its associated missile fire control systems (such as a MK 11 MOD 0 Missile Launcher and two MK 74 MOD 0 Missile Fire Control Systems).

BEACON - A system wherein a transponder in a missile receives coded signals from a shipboard radar guidance transmitter and transmits reply signals to a shipboard radar beacon receiver to enable a computer to determine missile position. The missile beacon transmitter and shipboard radar beacon receiver are tuned to a frequency different from that of the guidance transmitter.

BEAM - See Lobe, antenna. The beam is to the side of an aircraft or ship.

BEAM, CAPTURE - See Capture Beam.

BEAM-TO-BEAM CORRELATION (BBC) - BBC is used by frequency scan radars to reject pulse jamming and jamming at a swept frequency. Correlation is made from two adjacent beams (pulses). The receiver rejects those targets (signals) that do not occur at the same place in two adjacent beams.

BEAMWIDTH - See Antenna Beamwidth.

BEAT FREQUENCY OSCILLATOR (BFO) - Any oscillator whose output is intended to be mixed with another signal to produce a sum or difference beat frequency. Used particularly in reception of CW transmissions.

BINGO - The fuel state at which an aircraft must leave the area in order to return and land safely. Also used when chaff/flares reach a preset low quantity and automatic dispensing is inhibited.

BIPOLAR VIDEO - Unrectified (pre-detection) IF (both positive and negative portions of the RF envelope) signals that arise from the type of detection and console display employed in pulse Doppler and MTI receivers.

BISTATIC RADAR - A radar using antennas at different locations for transmission and reception.

BLANKING - The process of making a channel, or device non-effective for a certain interval. Used for retrace sweeps on CRTs or to mask unwanted signals such as blanking ones own radar from the onboard RWR.

BOGEY - Unknown air target.

BURN-THROUGH RANGE - The ability of a radar to see through jamming. Usually, described as the point when the radar's target return is a specified amount stronger than the jamming signal (typical values are 6 dB manual and 20 dB automatic). See Section 4-8.

BUTT LINE - Line used for reference in measurement of left/right location. One of several aircraft references. See also fuselage station and water line.

CAPTURE BEAM - A wide beam incorporated in capture transmitters of beam rider (command guided) missile systems to facilitate gaining initial control of a missile immediately after launch. Upon capture, the system then centers the missile in the narrow guidance beam. The figure illustrates a launched missile at point of capture.

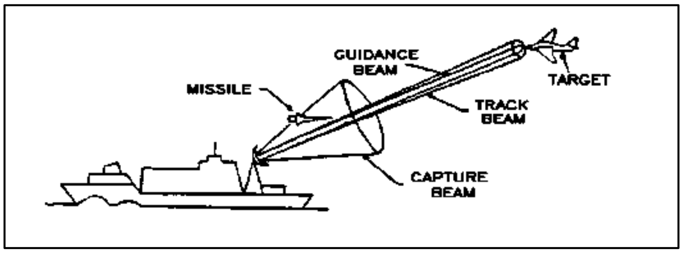

CAPTURE TRANSMITTER - A transmitter employing a wide beam antenna to gain initial control of in-flight missile for the purpose of centering the missile in the guidance transmitter antenna beam. See also Capture Beam.

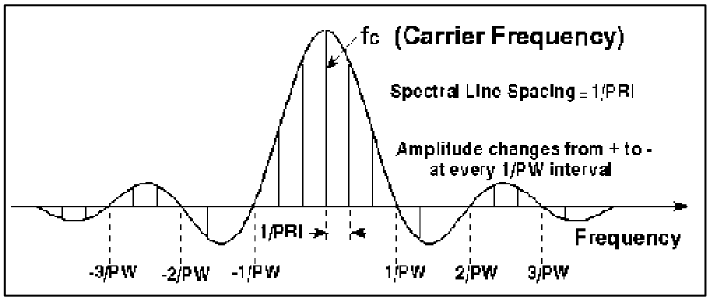

CARRIER FREQUENCY - The basic radio frequency of the wave upon which modulations are impressed. Also called "Carrier" or f_c. See figure at right.

CATCH-22 - A lose-lose situation, from the book of the same name.

CAVITY - A space enclosed by a conducting surface used as a resonant circuit at microwave frequencies. Cavity space geometry determines the resonant frequency. A storage area for oscillating electromagnetic energy.

CENTER FREQUENCY - The tuned or operating frequency. Also referred to as center operating frequency. In frequency diversity systems, the midband frequency of the operating range. See also Carrier Frequency.

CHAFF - Ribbon-like pieces of metallic materials or metalized plastic which are dispensed by aircraft or ships to mask or screen other "targets." The radar reflections off the chaff may cause a tracking radar to break lock on the target. The foil materials are generally cut into small pieces for which the size is dependent upon the radar interrogation frequency (approximately 1/2 wave length of the victim radar frequency). Being this length, chaff acts as a resonant dipole and reflects much of the energy back to the radar. Also see rainbow, rope, stream chaff, and window.

CHANNEL - A frequency or band of frequencies. In guided missile systems, an assigned center frequency and a fixed bandwidth around it. Designates operating frequency of track radars and frequency/code assignments of X-band CW illuminators.

CHIRP - A pulse compression technique which uses frequency modulation (usually linear) on pulse transmission.

CHIRP RADAR - See PC.

CIRCULARLY POLARIZED JAMMING - The techniques of radiating jamming energy in both planes of polarization simultaneously. With this method, there is a loss of 3 dB of effective power in either linear plane, and substantial loss if the opposite sense of circular polarization is used (i.e. left vs. right). See Section 3-2.

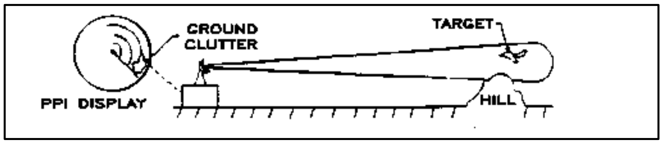

CLUTTER, RADAR - Undesired radar returns or echoes resulting from man-made or natural objects including chaff, sea, ground, and rain, which interfere with normal radar system observations. The figure illustrates a target being masked by ground clutter

CO-CHANNEL - This term is used to indicate that two (or more) equipments are operating on the same frequency.

COHERENT - Two signals that have a set (usually fixed) phase relationship.

COINCIDENCE DETECTOR - This radar video process requires more than one hit in a range cell before a target is displayed. This prevents video interference from pulses coming from another radar, because such interference is unlikely to occur twice in the same range cell.

COLLIMATION - The procedure of aligning fire control radar system antenna axes with optical line of sight, thereby ensuring that the radars will provide for correct target illumination and guidance beam positioning.

COMMAND CODE - Modulations superimposed upon transmitter carrier signals to provide electronic instructions to an airborne guided missile or pilotless aircraft. The receiver of the remotely guided vehicle is preset to accept only a selected transmitter code to eliminate the possibility of the vehicle responding to commands of extraneous signals. Missile command codes include instructions such as arm, warhead detonate, and self destruct.

COMMAND GUIDANCE - A guidance system wherein intelligence transmitted to the missile from an outside source causes the missile to traverse a directed flight path.

CONICAL SCAN - See Antenna, Nutating.

CONTINUOUS WAVE and CONTINUOUS WAVE ILLUMINATOR - See CW and CWI.

COOPERATIVE COUNTERMEASURES - (CO-OP) Generic term for jamming the same threat radar from two or more separate platforms that are in the same radar resolution cell.

COUPLING FACTOR - A multiplying factor, expressed in dB, used to express the change in EM energy intensity from a radar transmitter to a receiver. The factor includes the antenna gains and the loss (basic transmission loss) caused by the distance between the antennas. The factor will usually be a negative dB figure (a reduction in intensity) because basic transmission loss is always a large negative value. The antenna gains may be positive (pointed toward each other) or negative (no main beam interactions).

CROSS MODULATION - Intermodulation caused by modulation of the carrier by an undesired signal wave.

CROSS POLARIZATION - or “Cross Pole,” is a monopulse jamming technique where a cross-polarized signal is transmitted to give erroneous angle data to the radar. The component of the jamming signal with the same polarization as the radar must be very small.

CW (CONTINUOUS WAVE) - In radar and EW systems this term means that the transmitter is on constantly; i.e., not pulsed (100% duty cycle). These systems may frequency or phase modulate the transmitter output. A CW radar has the ability to distinguish moving targets against a stationary background while conserving spectrum bandwidth compared to pulsed radar requirements. A CW radar extracts accurate target range-rate data but cannot determine target range.

CWI (CONTINUOUS WAVE ILLUMINATOR) - A surface or aircraft-based CW transmitter employed in semiactive homing missile systems where the transmitter illuminates the target and the missile senses the reflected energy. The transmitter also provides a reference signal to the missile rear receiver to allow determination of range-rate data and target identification. CW transmitter emissions are sometimes coded corresponding to the associated missile receiver codes to reduce the possibility of the “missile accepting commands of extraneous signals.

DECIBEL (dB) - A dimensionless unit for expressing the ratio of two values of power, current, or voltage. The number of decibels being equal to: $dB = 10 \log P_2/P_1 = 20 \log V_2/V_1 = 20 \log I_2/I_1$ Normally, used for expressing transmission gains, losses, levels, and similar quantities. See Section 2-4.

DECEPTION - The deliberate radiation, reradiation, alteration, absorption or reflection of electromagnetic energy in a manner intended to mislead the enemy interpretation or use of information received by his electronic systems.

dB - See Decibel, or Decibel Section 2-4.

dBc - Decibels referenced to the carrier signal.

dBi - Decibels referenced to an isotropic radiator. (dBLi indicating linear isotropic radiator is sometimes used).

dBm - Decibels relative to 1 mW. dBm is calculated by using the ratio of some power (expressed in mW) to 1 mW. For example, 1 mW is 0 dBm and 10 mW is +10 dBm.

dBsm - Decibel referenced to one square meter.

dBv / dBμv - Decibels referenced to one volt or microvolt, i.e. 0 dBv is 1 volt or 120 dBμv.

dBW / dBμW - Decibels referenced to 1 watt or one microwatt, i.e. 0 dBW is 1 watt or 30 dBm or 60 dBμW.

DEMODULATOR - A device employed to separate the modulation signal from its associated carrier, also called Second Detector. See also Detection.

DESIGNATION - The assignment of a fire control radar to a specific target by supplying target coordinate data to the radar system.

DETECTION - Usually refers to the technique of recovering the amplitude modulation signal (envelope) superimposed on a carrier. See figure at right.

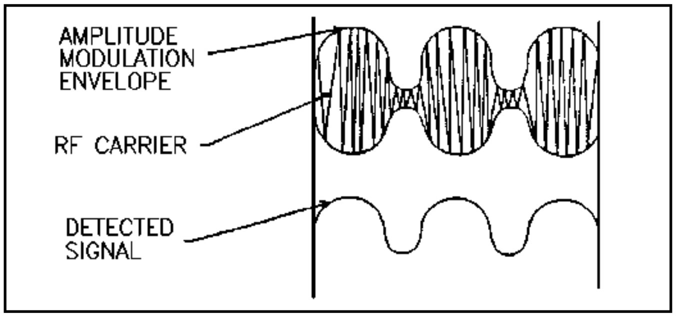

DICKE FIX - This type of radar processing occurs in the IF amplifier. A limiter follows a wideband amplifier, and then the signal goes to a matched filter amplifier. This discriminates against pulses that are too long (clutter) or too short (interference). The "DICKE FIX" is a technique that is specifically designed to protect the receiver from ringing caused by noise, fast-sweep, or narrow pulse jamming. The basic configuration consists of a broadband limiting IF amplifier, followed by an IF amplifier of optimum bandwidth. The limit level is preset at approximately the peak amplitude of receiver noise, the bandwidth may vary from 10 to 20 MHz, depending on the jamming environment. This device provides excellent discrimination against fast sweep jamming (10-500 MHz), usually something on the order of 20 to 40 dB, without appreciable loss in sensitivity. However, strong CW jamming will seriously degrade the performance of the DICKE FIX because the CW signal captures the radar signal in the limiter.

DIELECTRICALLY STABILIZED OSCILLATOR - The DSO uses a dielectric resonator as the frequency determining element. When the dielectric material is properly selected and used, the variations in dielectric constant vs. temperature and the dimensions of the resonant structure vs. temperature tend to cancel out, providing relatively good frequency vs. temperature stability. The DSO offers frequency accuracy and stability, low power consumption and high reliability. Some of the commonly used materials are barium, zirconium, or tin tinates. The composition of these materials may be controlled to achieve any frequency variation with temperature with close tolerances.

DIODE - An electronic device which restricts current flow chiefly to one direction. See also Gunn diode, IMPATT diode, PIN diode, point contact diode, Schottky barrier diode, step recovery diode, tunnel diode, varactor diode.

DIODE SWITCH - PIN-diode switches provide state-of-the-art switching in most present-day microwave receivers. These switches are either reflective or nonreflective in that the former reflect incident power back to the source when in the isolated state. While both types of switches can provide high isolation and short transition times, the reflective switch offers multi octave bandwidth in the all shunt diode configuration, while the non-reflective switch offers an octave bandwidth.

DIPLEX - The simultaneous transmission or reception of two signals using a common feature such as a single antenna or carrier. Typically, two transmitters operate alternately at approximately the same RF and using a common antenna. See Section 6-7 for a discussion of diplexers.

DIRECTIONAL COUPLER - A 4-port transmission coupling device used to sample the power traveling in one direction through the main line of the device. There is considerable isolation (typically 20 dB) to signals traveling in the reverse direction. Because they are reciprocal, the directional coupler can also be used to directively combine signals with good reverse isolation. The directional coupler is implemented in waveguide and coaxial configurations. See Section 6-4.

DIRECTIVITY - For antennas, directivity is the maximum value of gain in a particular direction. (Isotropic point source has directivity = 1). For directional couplers, directivity is a measure (in dB) of how good the directional coupling is and is equal to the isolation minus the coupling. See Section 6-4.

DISH - A microwave reflector used as part of a radar antenna system. The surface is concave and is usually parabolic shaped. Also called a parabolic reflector.

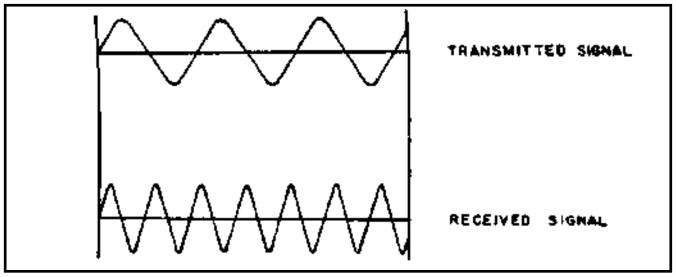

DOPPLER EFFECT - The apparent change in frequency of an electromagnetic wave caused by a change in distance between the transmitter and the receiver during transmission/reception. The figure illustrates the Doppler increase that would be realized by comparing the signal received from a target approaching the radar site to the transmitted reference signal. An apparent frequency decrease would be noted for targets departing the radar location. Differences can be calibrated to provide target range-rate data.

DRY RUN - A test run with aircraft/ship armament and/or EW switches off.

DUCTING - The increase in range that an electromagnetic wave will travel due to a temperature inversion of the atmosphere. The temperature inversion forms a channel or waveguide (duct) for the waves to travel in, and they can be trapped, not attenuating as would be expected from the radar equation.

DUMMY LOAD (Radio Transmission) - A dissipative but essentially nonradiating substitute device having impedance characteristics simulating those of the antenna. This allows power to be applied to the radar transmitter without radiating into free space. Dummy loads are commonly used during EMCON conditions or when troubleshooting a transmitter at a workbench away from its normal environment.

DUPLEXER - A switching device used in radar to permit alternate use of the same antenna for both transmitting and receiving.

DUTY CYCLE - The ratio of average power to peak power, or ratio of pulse length to interpulse period for pulsed transmitter systems. Interpulse period is equal to the reciprocal of the pulse repetition rate.
See Section 2-5.

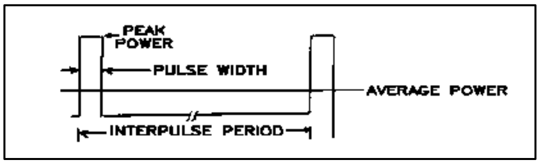

The duty cycle of a radar having a pulse length of 0.3 µsec and a PRF of 2000 pulses/sec is computed as follows:

Interpulse Period, T = PRI = 1/PRF = 500 µsec

$$Duty\ Cycle = \frac{Pulse\ length}{Interpulse\ Period} = \frac{0.3\ \mu\mathrm{sec}}{500\ musec} = 0.0006\ (or\ 0.06\%)$$ or Duty Cycle (dB) = 10log(Duty cycle) = -32.2 dB

An output tube providing an average power of only 90 watts for such a system would, therefore, provide a peak power of:

$$Peak\ Power = \frac{Average\ Power}{Duty\ Cycle} = \frac{90}{0.0006} = 150{,}000\ W\ or\ 52\ dBW\ or\ 82\ dBm$$

EFFECTIVE RADIATED POWER (ERP) - Input power to the antenna in watts times the gain ratio of the antenna. When expressed in dB, ERP is the transmitter power (P_T), in dBm (or dBW) plus the antenna gain (G_T) in dB. The term EIRP is used sometimes and reiterates that the gain is relative to an isotropic radiator.

EGRESS - Exit the target area.

ELECTROMAGNETIC COUPLING - The transfer of electromagnetic energy from one circuit or system to another circuit or system. An undesired transfer is termed EMI (electromagnetic interference).

EMC (ELECTROMAGNETIC COMPATIBILITY) - That condition in which electrical/electronic systems can perform their intended function without experiencing degradation from, or causing degradation to other electrical/electronic systems. More simply stated, EMC is that condition which exists in the absence of EMI. See also Intersystem and Intrasystem EMC tests.

EME (ELECTROMAGNETIC ENVIRONMENT) - The total electromagnetic energy in the RF spectrum that exists at any given location.

EMI (ELECTROMAGNETIC INTERFERENCE) - Any induced, radiated, or conducted electrical emission, disturbance, or transient that causes undesirable responses, degradation in performance, or malfunctions of any electrical or electronic equipment, device, or system. Also synonymously referred to as RFI (Radio Frequency Interference).

EMI MODEL - Usually a set of equations or logical concepts designed to illustrate the interactions, the detailed parameters considerations, and mathematical procedures necessary for proper analysis of a given EMI situation.

EMITTER - Any device or apparatus which emits electromagnetic energy.

EMP (ELECTROMAGNETIC PULSE) - The generation and radiation of a very narrow and very high-amplitude pulse of electromagnetic noise. It is associated with the high level pulse as a result of a nuclear detonation and with intentionally generated narrow, high-amplitude pulse for ECM applications. In the case of nuclear detonations, EMP consists of a continuous spectrum with most of its energy distributed through the low frequency band of 3 KHz to 1 MHz.

ERROR SIGNAL - In servomechanisms, the signal applied to the control circuit that indicates the degree of misalignment between the controlling and the controlled members. In tracking radar systems, a voltage dependent upon the signal received from a target whose polarity and magnitude depend on the angle between the target and the center axis of the scanning beam.

FAST TIME CONSTANT - See FTC.

FEET DRY / WET - Aircraft has crossed from water to shore / aircraft has crossed from shore to water.

FERRET - An aircraft, ship, or vehicle especially equipped for the detection, location, recording, and analyzing of electromagnetic radiations.

FIELD STRENGTH - The magnitude of a magnetic or electric field at any point, usually expressed in terms of ampere turns per meter or volts per meter. Sometimes called field intensity and is expressed in volts/meter or dBμv/meter. Above 100 MHz, power density terminology is used more often. See Section 4-1.

FIRST HARMONIC - The fundamental (original) frequency.

FREQUENCY AGILITY - A radar's ability to change frequency within its operating band, usually on a pulse-to-pulse basis. This is an ECCM technique employed to avoid spot jamming and to force the jammer to go into a less effective barrage mode.

FREQUENCY AGILITY RADAR - A radar that automatically or semiautomatically tunes through a discrete set of operating frequencies in its normal mode of operation.

FREQUENCY DIVERSITY RADAR - A radar system technique, employed primarily as an antijamming feature, where the transmitter output frequency varies randomly from pulse to pulse over a wide frequency range.

FREQUENCY RANGE - (1) A specifically designated portion of the frequency spectrum; (2) of a device, the band of frequencies over which the device may be considered useful with various circuit and operating conditions; (3) of a transmission system, the frequency band in which the system is able to transmit power without attenuating or distorting it more than a specified amount.

FREQUENCY SHIFT KEYING (FSK) - A form of FM where the carrier is shifted between two frequencies in accordance with a predetermined code. In multiple FSK, the carrier is shifted to more than two frequencies. FSK is used principally with teletype communications.

"FRUIT" - In a radar beacon system, there is a type of interference called "FRUIT," caused by beacon replies to interrogation asynchronous with the observer's interrogator. The largest amount of this interference is received through the sidelobes of the interrogating antenna, but it can become dense enough to cause false target indications.

FTC (FAST TIME CONSTANT) - An antijam feature employed in radar systems where receiver circuits may be selected to provide a short time constant to emphasize signals of short duration to produce discrimination against the low frequency components of clutter.

FUNDAMENTAL FREQUENCY - Used synonymously for tuned frequency, carrier frequency, center frequency, output frequency, or operating frequency.

FUSELAGE STATION or just STATION - A reference point (usually the nose of an aircraft) used to measure or identify fore and aft locations. One of several aircraft location designations - also see butt line and water line.

GaAs FET AMPLIFIER - Because of their low noise, field-effect transistors are often used as the input stage of wideband amplifiers. Their high input resistance makes this device particularly useful in a variety of applications. Since the FET does not employ minority current carriers, carrier storage effects are eliminated giving the device faster operating characteristics and improved radiation resistant qualities.

GAIN - For antennas, the value of power gain in a given direction relative to an isotropic point source radiating equally in all directions. Frequently expressed in dB (gain of an isotropic source = 0 dB). The formula for calculating gain is:

$$G = \frac{4\pi\, P(\theta,\phi)}{P_{in}} \quad ; \textit{where} \quad \begin{aligned} P(\theta,\phi) &= \textit{Radiation intensity in given direction} \\ P_{in} &= \textit{Power into lossless antenna radiating uniformly in all directions} \end{aligned}$$

Note: (1) If radiation efficiency is unity, then gain = directivity i.e. if directivity = 2, then gain = 3 dB, etc.
(2) Interference losses within an array also affect gain
(3) See Section 3-1 for further details

For amplifiers, gain is the ratio of the output power to input power (usually in dB).

GATE (RANGE) - A signal used to select radar echoes corresponding to a very short range increment. Range is computed by moving the range gate or marker to the target echo; an arrangement which permits radar signals to be received in a small selected fraction of the time period between radar transmitter pulses.

GATING - (1) The process of selecting those portions of a wave which exist during one or more selected time intervals; (2) the application of a square waveform of desired duration and timing to perform electronic switching; (3) the application of receiver operating voltages to one or more stages only during that portion of a cycle of operation when reception is desired. See also Gate (Range).

GCI (GROUND-CONTROLLED INTERCEPT) - Vectoring an interceptor aircraft to an airborne target by means of information relayed from a ground-based radar site which observes both the interceptor and the target.

GIGA - A prefix meaning 10^9 (times a billion). For example, gigahertz (GHz).

GLINT (In Radar) - 1. The random component of target location error caused by variations in the phase front of the target signal (as contrasted with Scintillation Error). Glint may affect angle, range of

Doppler measurements, and may have peak values corresponding to locations beyond the true target extent in the measured coordinate. 2. Electronic countermeasures that uses the scintillating, or flashing effect of shuttered or rotating reflectors to degrade tracking or seeking functions of an enemy weapons system.

GUARDBAND - A frequency band to which no other emitters are assigned as a precaution against interference to equipments susceptible to EMI in that band.

GUIDANCE, BEAM RIDER - A missile guidance technique which is dependent on the missile's ability to determine its positions with reference to the center of scan of the guidance radar beam and thus correct its trajectory on the basis of detected errors.

GUIDANCE CODE - A technique of modulating guidance transmitter carriers with coded pulses compatible with the receiver code of the missile assigned that system, thus reducing the possibility of the missile accepting erroneous commands of other transmissions.

GUIDANCE, COMMAND - A guidance system wherein intelligence transmitted to the missile from an outside source causes the missile to traverse a directed path in space.

GUIDANCE, HOMING, ACTIVE - A system of homing guidance wherein both the transmitter and receiver are carried within the missile.

GUIDANCE, HOMING, PASSIVE - A form of homing guidance, which is dependent on a missile's ability to detect energy emitted by the target. Frequently termed Home-On-Jam (HOJ).

GUIDANCE, HOMING, SEMIACTIVE - A system of homing guidance wherein the missile uses reflected signals from the target which has been illuminated by a source other than within the missile. See also CWI.

GUIDANCE, INERTIAL - A self-contained system independent of information obtained from outside the missile, usually using Newton's second law of motion.

GUNN DIODE - The Gunn diode is a transferred electron device which because of its negative resistance can be used in microwave oscillators or amplifiers. When the applied voltage exceeds a certain critical value, periodic fluctuations in current occur. The frequency of oscillation depends primarily upon the drift velocity of electrons through the effective length of the device. This frequency may be varied over a small range by means of mechanical tuning.

HARMONIC - A sinusoidal component of a periodic wave or quantity having a frequency that is an integral multiple of the fundamental frequency. For example, a component which is twice the fundamental frequency is called the second harmonic. (The fundamental is the first harmonic, which is frequently misunderstood.)

HERTZ - The unit of frequency equal to one cycle per second.

HOME-ON-JAM (HOJ) - See Guidance, Homing, Passive.

HORN ANTENNA - A flared, open-ended section of waveguide used to radiate the energy from a waveguide into space. Also termed "horn" or "horn radiator." Usually linearly polarized, it will be vertically polarized when the feed probe is vertical, or horizontally polarized if the feed is horizontal. Circular polarization can be obtained by feeding a square horn at a 45° angle and phase shifting the vertical or horizontal excitation by 90°.

HYPERABRUPT VARACTOR OSCILLATOR - Due to a non-uniform concentration of N-type material (excess electrons) in the depletion region, this varactor produces a greater capacitance change in tuning voltage and a far more linear voltage-vs.-frequency tuning curve. As a result, this device has an improved tuning linearity and low tuning voltage.

IF (INTERMEDIATE FREQUENCY) - The difference frequency resulting from mixing (beating) the received signal in a superheterodyne receiver with the signal from the local oscillator. The difference frequency product provides the advantages inherent to the processing (amplification, detection, filtering, and such) of low frequency signals. The receiver local oscillator may operate either below or above the receiver tuned frequency. A single receiver may incorporate multiple IF detection.

$IF = F_{LO} - F_O$ (for a local oscillator operating above the fundamental)

where:
F_O = Received fundamental frequency
F_{LO} = Local oscillator frequency

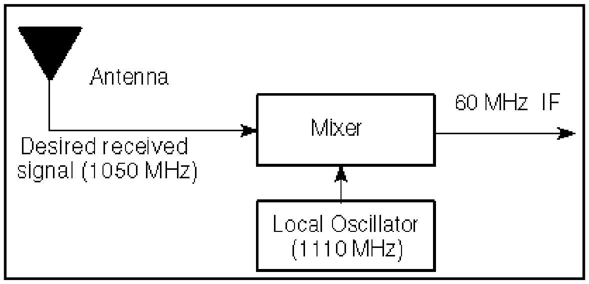
Antenna
Mixer
60 MHz IF
Desired received signal (1050 MHz)
Local Oscillator (1110 MHz)

The simplified block diagram illustrates a typical mixing procedure employed in radar systems to obtain desired IF frequencies. The local oscillator is tuned above the fundamental frequency in this example. It should be noted that an undesired signal received at the receiver image frequency of 1170 MHz will also produce the desired 60 MHz IF frequency; this relationship provides the receiver image. See also Image Frequency.

IFF (IDENTIFICATION FRIEND OR FOE) - A system using radar transmission to which equipment carried by friendly forces automatically responds by emitting a unique characteristic series of pulses thereby distinguishing themselves from enemy forces. It is the “Mode IV” for the aircraft transponder. See also transponder.

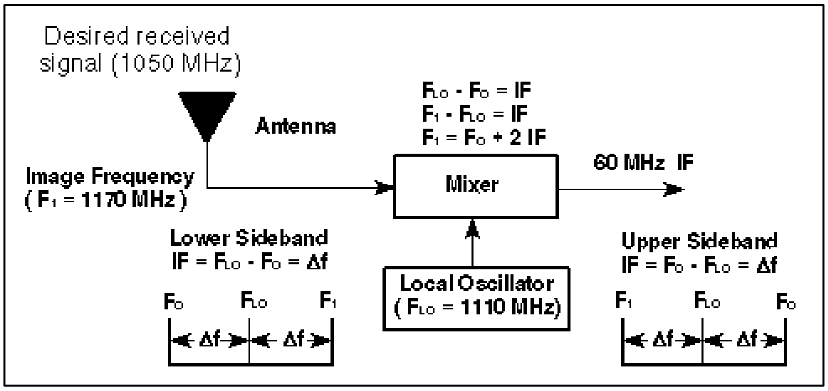
Desired received signal (1050 MHz)
Antenna
$F_{LO} - F_O = IF$
$F_I - F_{LO} = IF$
$F_I = F_O + 2\,IF$
Image Frequency (F_I = 1170 MHz)
Mixer
60 MHz IF
Lower Sideband
$IF = F_{LO} - F_O = \Delta f$
F_O F_{LO} F_I
Δf Δf
Local Oscillator (F_{LO} = 1110 MHz)
Upper Sideband
$IF = F_O - F_{LO} = \Delta f$
F_I F_{LO} F_O
Δf Δf

IMAGE FREQUENCY - That frequency to which a given superheterodyne receiver is inherently susceptible, thereby rendering such a receiver extremely vulnerable to EMI at that frequency. The image frequency is located at the same frequency difference (Δf) to one side of the local oscillator as the tuned (desired) frequency is to the other side. An undesired signal received at the image frequency by a superheterodyne receiver not having preselection would, therefore, mix (beat) with the oscillator, produce the proper receiver IF, and be processed in the same manner as a signal at the desired frequency. See also receiver selectivity.

IMAGE JAMMING - Jamming at the image frequency of the radar receiver. Barrage jamming is made most effective by generating energy at both the normal operating and image frequency of the radar.

Image jamming inverts the phase of the response and is thereby useful as an angle deception technique. Not effective if the radar uses image rejection.

IMPATT DIODE - The IMPATT (IMPact Avalanche and Transit Time) diode acts like a negative resistance at microwave frequencies. Because of this property, Impatt diodes are used in oscillators and amplifiers. Usually the frequency range is in the millimeter wave region where other solid state devices cannot compete.

INGRESS - Go into the target area.

INSERTION LOSS - The loss incurred by inserting an element, device, or apparatus in an electrical/electronic circuit. Normally expressed in decibels determined as 10 log of the ratio of power measured at the point of insertion prior to inserting the device (P_1) to the power measured after inserting the device (P_2). Insertion loss (dB) = 10 log P_1/P_2.

INTEGRATION EFFECT - Pulse radars usually obtain several echoes from a target. If these echoes are added to each other, they enhance the S/N ratio, making a weak target easier to detect. The noise and interference do not directly add from pulse to pulse, so the ratio of target strength to undesired signal strength improves making the target more detectable. Random noise increases by the square root of the number of integrations, whereas the signal totally correlates and increases directly by the number of integrations, therefore the S/N enhancement is equal to the square root of the number of integrations.

INTERFERENCE - See EMI.

INTERFERENCE PARAMETERS - Equipment and propagation characteristics necessary for the proper evaluation of a given EMI situation.

INTERFERENCE/SIGNAL RATIO - See I/S Ratio.

INTERFERENCE THRESHOLD - The level of interference normally expressed in terms of the I/S (interference/signal) ratio at which performance degradation in a system first occurs as a result of EMI.

INTERFEROMETER - When two widely spaced antennas are arrayed together, they form an interferometer. The radiation pattern consists of many lobes, each having a narrow beamwidth. This antenna can provide good spatial selectivity if the lobe-to-lobe ambiguity can be solved such as using amplitude comparison between the two elements.

INTERMODULATION - The production, in a nonlinear element (such as a receiver mixer), of frequencies corresponding to the sums and differences of the fundamentals and harmonics of two or more frequencies which are transmitted through the element; or, the modulation of the components of a complex wave by each other, producing frequencies equal to the sums and differences of integral multiples of the component frequencies of the complex wave.

INTERSYSTEM EMC - EMC between the external electromagnetic environment (EME) and an aircraft with its installed systems. Generally, only system BIT must operate properly on the carrier deck while all system functions must operate properly in the operational EME.

INTRASYSTEM EMC - EMC between systems installed on an aircraft, exclusive of an external environment.

INVERSE CON SCAN - One method of confusing a radar operator or fire control radar system is to provide erroneous target bearings. This is accomplished by first sensing the radar antenna scan rate and then modulating repeater amplifier gain so the weak portion of the radar signal is amplified by the jammer, while the strong portion is not, so the weapons systems will fire at some bearing other than the true target bearing. The angle deception technique is used to break lock on CONSCAN radars.

INVERSE GAIN - Amplification, inverse modulation, and re-radiation of a radar's pulse train at the rotation rate of the radar scan. Deceives a conical scanning radar in angle.

ISOTROPIC ANTENNA - A hypothetical antenna which radiates or receives energy equally in all directions.

I/S RATIO (INTERFERENCE-TO-SIGNAL RATIO) (ISR) - The ratio of electromagnetic interference level to desired signal level that exists at a specified point in a receiving system. The ratio, normally expressed in dB, is employed as a tool in prediction of electronic receiving system performance degradation for a wide range of interference receiver input levels. Performance evaluations compare actual I/S ratios to minimum acceptable criteria.

JAFF - Expression for the combination of electronic and chaff jamming.

JAMMING - The deliberate radiation, reradiation, or reflection of electromagnetic energy with the object of impairing the use of electronic devices, equipment, or systems by an enemy.

JATO - JAmmer Technique Optimization group -- Organization coordinates EA technique development and evaluation efforts in support of naval airborne EA (ALQ-99, USQ-113, ALQ-227, ALQ-231, ALE-43). The group is run out of PMA-234 and comprised of elements at NAWCWD-Point Mugu, Johns Hopkins University/ Applied Physics Lab, Naval Research Lab, and NSWC-Crane. Efforts include M&S, Lab test, ground test, flight test, direct fleet support, and requirements and test support to development programs.

JINK - An aircraft maneuver which sharply changes the instantaneous flight path but maintains the overall route of flight. More violent than a weave.

JITTERED PRF - An antijam feature of certain radar systems which varies the PRF consecutively, and randomly, from pulse to pulse to prevent enemy ECM equipment from locking on, and synchronizing with, the transmitted PRF. PRF is synonymous with pulse repetition rate (PRR).

JTAT - JATO Techniques Analysis and Tactics -- Analysis documents produced by JATO to advise EA systems operators on the best ways to employ their systems by region, mission and equipment configuration. Currently covering ALQ-99 and USQ-113 on EA-6B; ALQ-99 and ALQ-227 on EA-18G, and ALQ-231 on AV-8B.

KILO - A prefix meaning 10^3 (times one thousand). For example, kilohertz.

KLYSTRON AMPLIFIER - An electron beam device which achieves amplification by the conversion of periodic velocity variations into conduction-current modulation in a field-free drift region. Velocity variations are launched by the interaction of an RF signal in an input resonant cavity and are coupled out through an RF output cavity. Several variations including reflex and multi cavity klystrons are used.

KLYSTRON, MULTICAVITY - An electron tube which employs velocity modulation to generate or amplify electromagnetic energy in the microwave region. Since velocity modulation employs transit time of the electron to aid in the bunching of electrons, transient time is not a deterrent to high frequency operations as is the case in conventional electron tubes. See also Velocity Modulation.

KLYSTRON, REFLEX - A klystron which employs a reflector (repeller) electrode in place of a second resonant cavity to redirect the velocity-modulated electrons through the resonant cavity. The repeller causes one resonant circuit to serve as both input and output, which simplifies the tuning operation. This type of klystron is well adapted for use as an oscillator because the frequency is easily controlled by varying the position of the repeller. See also Velocity Modulation.

LEAKAGE - Undesired radiation or conduction of RF energy through the shielding of an enclosed area or of an electronic device.

LENS, RADAR (MICROWAVE) - The purpose of any such lens is to refract (focus) the diverging beam from an RF feed into a parallel beam (transmitting) or vice versa (receiving). The polarization is feed dependent.

LIGHT AMPLIFICATION BY STIMULATED EMISSION OF RADIATION (LASER) - A process of generating coherent light. The process utilizes a natural molecular (and atomic) phenomenon whereby molecules absorb incident electromagnetic energy at specific frequencies, store this energy for short but usable periods, and then release the stored energy in the form of light at particular frequencies in an extremely narrow frequency-band.

LIMITING - A term to describe that an amplifier has reached its point of saturation or maximum output voltage swing. Deliberate limiting of the signal is used in FM demodulation so that AM will not also be demodulated.

LITTORAL - Near a shore.

LOBE, ANTENNA - Various parts of the antenna's radiation pattern are referred to as lobes, which may be subclassified into major and minor lobes. The major lobe is the lobe of greatest gain and is also referred to as the main lobe or main beam. The minor lobes are further subclassified into side and back lobes as indicated in the figure to the right. The numbering of the side lobes are from the main lobe to the back lobe.

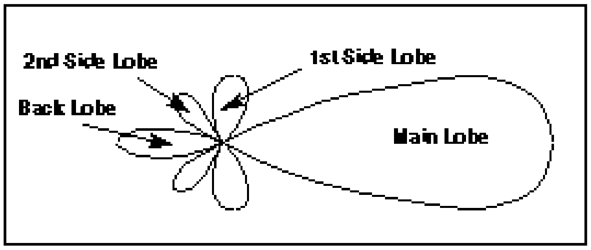

LOCAL OSCILLATOR FREQUENCY - An internally generated frequency in a superheterodyne receiver. This frequency differs from the receiver operating frequency by an amount equal to the IF of the receiver. The local oscillator frequency may be designed to be either above or below the incoming signal frequency.

LOG VIDEO - This receiver process, generally implemented in the IF, compresses the dynamic range of the signal so both weak and strong signals are displayed without changing the gain setting. Output voltage can be calibrated in volts/dB of input power.

LONG PULSE MODE - Many pulsed radars are capable of transmitting either long or short pulses of RF energy. When the long pulses of RF energy are selected manually (or sometimes automatically), the radar is said to be operating in the long pulse mode. In general, "long pulse mode" is used to obtain high average power for long-range search or tracking, and “short pulse mode” gives low average power for short-range, high-definition, tracking or search.

LOOSE DEUCE - General term for two aircraft working in mutual support of each other.

LORO (LOBE-ON-RECEIVE-ONLY) - A mode of operation generally consisting of transmitting on one non-scanning antenna system and receiving the reflected energy on another scanning system (The receiver could be TWS, Conical, or monopulse).

MACH NUMBER - The ratio of the velocity of a body to the speed of sound in the medium that is being considered. In the atmosphere, the speed of sound varies with temperature and atmospheric pressure, hence, so does mach number.

MAGNETIC ANOMALY DETECTOR - A means of detecting changes in the earth’s magnetic field caused by the presence of metal in ships and submarines.

MAGNETRON - A magnetron is a thermionic vacuum tube which is constructed with a permanent magnet forming a part of the tube and which generates microwave power. These devices are commonly used as the power output stage of radar transmitters operating in the frequency range above 1000 MHz and are used less commonly down to about 400 MHz. A magnetron has two concentric cylindrical electrodes. On a conventional magnetron, the inner one is the cathode and the outer one is the anode. The opposite is true for a coaxial magnetron.

MAGNETRON OSCILLATOR - A high-vacuum tube in which the interaction of an electronic space charge and a resonant system converts direct current power into ac power, usually at microwave frequencies. The magnetron has good efficiency, is capable of high power outputs, and is stable.

MATCHED FILTER - This describes the bandwidth of an IF amplifier that maximizes the signal-to-noise ratio in the receiver output. This bandwidth is a function of the pulsewidth of the signal.

MDS (MINIMUM DETECTABLE/DISCERNIBLE SIGNAL) - The receiver input power level that is just sufficient to produce a detectable/discernible signal in the receiver output. The detectable term is interchangeable with S_{min} and the discernable term is interchangeable with MVS. See Section 5-2.

MEACONING -- A system receiving radio signals and rebroadcasting them (or just transmitting) on the same frequency to confuse navigation. The meaconing station attempts to cause aircraft to receive inaccurate range or bearing information.

MEATBALL - Visual light “ball” seen in Fresnel lens optical landing system (FLOLS) by pilot during carrier or Navy field landing. Used as a reference to determine if flight path is high or low.

MEGA - A prefix meaning 10^6 (times one million). For example megahertz (MHz)

MICROVOLT PER METER - A commonly used unit of field strength at a given point. The field strength is measured by locating a standard receiving antenna at that point, and the “microvolts per meter” value is then the ratio of the antenna voltage in microvolts to the effective antenna length in meters. Usually used below 100 MHz. Above 100 MHz, power density terminology is normally used.

MICROWAVE AMPLIFICATION BY STIMULATED EMISSION OF RADIATION (MASER) - A low-noise radio-frequency amplifier. The emission of energy stored in a molecular or atomic system by a microwave power supply is stimulated by the input signal.

MISS DISTANCE - Used variously in different contexts. The distance from the missile to the geometric center of the aircraft, or the closest point of approach (CPA) of the missile to any portion of the aircraft such as the aircraft nose or telemetry pod, etc.

MISSILE SYSTEMS FUNCTIONS - Examples of missile system functions are: “acquisition” (ability to lock-on a desired target); “tracking” of a target; “guidance” of a missile toward a target; “illumination” of a target so that a homing missile can home on the reflected RF illumination; and “command” signal transmission to a missile to cause it to arm, to detonate, to commence homing, or to destroy itself.

MIXERS - See Balanced and Schottky Diode Mixers.

MODULATION - The process whereby some characteristic of one wave is varied in accordance with some characteristic of another wave. The basic types of modulation are angle modulation (including the special cases of phase and frequency modulation) and amplitude modulation. In missile radars, it is common practice to amplitude modulate the transmitted RF carrier wave of tracking and guidance transmitters by using a pulsed wave for modulating, and to frequency module the transmitted RF carrier wave of illuminator transmitters by using a sine wave.

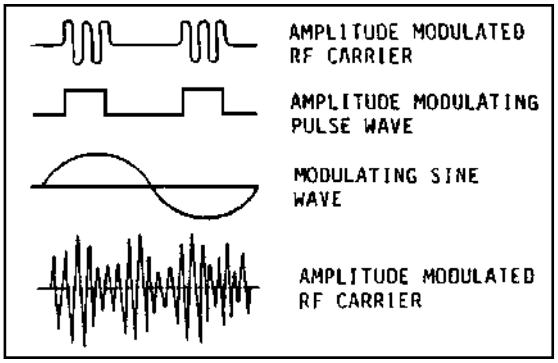
AMPLITUDE MODULATED RF CARRIER
AMPLITUDE MODULATING PULSE WAVE
MODULATING SINE WAVE
AMPLITUDE MODULATED RF CARRIER

MODULATION, AMPLITUDE - This type of modulation changes the amplitude of a carrier wave in responses to the amplitude of a modulating wave. This modulation is used in radar and EW only as a switch to turn on or turn off the carrier wave; i.e., pulse is a special form of amplitude modulation.

MODULATION, FREQUENCY - The frequency of the modulated carrier wave is varied in proportion to the amplitude of the modulating wave and therefore, the phase of the carrier varies with the integral of the modulating wave. See also Modulation.

MODULATION, PHASE - The phase of the modulated carrier is varied in proportion to the amplitude of the modulating wave. See also Modulation.

MONOPULSE - (See figure to right) A type of tracking radar that permits the extracting of tracking error information from each received pulse and offers a reduction in tracking errors as compared to a conical-scan system of similar power and size. Multiple (commonly four) receiving antennas or feeds are placed symmetrically about the center axis and operate simultaneously to receive each RF pulse reflected from the target. A comparison of the output signal amplitude or phase among the four antennas indicates the location of the target with respect to the radar beam center line. The output

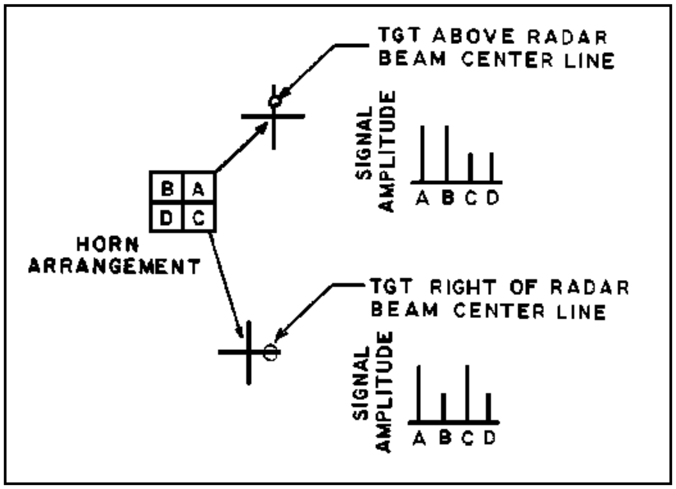

of the comparison circuit controls a servo system that reduces the tracking error to zero and thereby causes the antenna to track the target.

MOS (MINIMUM OPERATIONAL SENSITIVITY) - The minimum signal which can be detected and automatically digitally processed by a radar without human discrimination.

MTI (MOVING TARGET INDICATOR) - This radar signal process shows only targets that are in motion. Signals from stationary targets are subtracted out of the return signal by a memory circuit.

MULTIPATH - The process by which a transmitted signal arrives at the receiver by at least two different paths. These paths are usually the main direct path, and at least one reflected path. The signals combine either constructively or destructively depending upon phase, and the resultant signal may be either stronger or weaker than the value computed for free space.

MULTIPLEX - Simultaneous transmission of two or more signals on a common carrier wave. The three types of multiplex are called time division, frequency division, and phase division.

MULTIBAND RADAR - A type of radar which uses simultaneous operation on more than one frequency band through a common antenna. This technique allows for many sophisticated forms of video processing and requires any jammer to jam all channels at the same time in order to be effective.

MVS (MINIMUM VISIBLE SIGNAL) - The minimum input pulse signal power level which permits visibility of the output pulse, such as on a radar A-scope display. This level is determined by initially setting the input level above the visible detection threshold, and then slowly decreasing the amplitude.

NOISE FIGURE, RECEIVER - A figure of merit (NF or F) of a system given by the ratio of the signal-to-noise ratio at the input, S_i / N_i, divided by the signal-to-noise ratio at the output, S_o/N_o. It essentially expresses the ratio of output noise power of a given receiver to that of a theoretically perfect receiver which adds no noise.

$$\text{Noise Figure} = \frac{S_i / N_i}{S_o / N_o} = \frac{N_o}{G\,N_i}$$

Where $S_o = GS_i$ and G is the gain of the system.

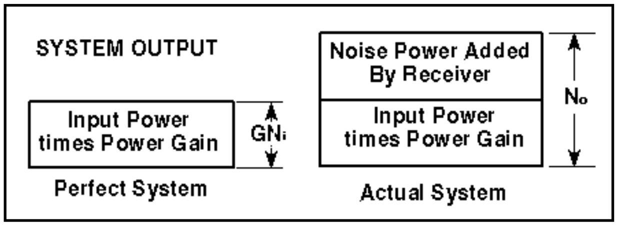
SYSTEM OUTPUT
Input Power times Power Gain
GNi
Perfect System
Noise Power Added By Receiver
Input Power times Power Gain
No
Actual System

Noise figure is usually expressed in dB and given for an impedance matched condition. Impedance mismatch will increase the noise figure by the amount of mismatch loss. NF is usually given at room temperature; 17°C or 290°K. See Section 5-2.

NOISE JAMMING - A continuous random signal radiated with the objective of concealing the aircraft echo from the enemy radar. In order for it to be effective, it must have an average amplitude at least as great as the average amplitude of the radar echo. There are three major categories of noise jamming which are grouped by how jamming power is concentrated: Spot, barrage, and swept jamming. (See individual definitions)

NONCOHERENT - Two signals that have no set phase relationship.

NOTCH - The portion of the radar velocity display where a target disappears due to being notched out by the zero Doppler filter. If not filtered (notched), ground clutter would also appear on the display. A notch filter is a narrow band-reject filter. A "notch maneuver" is used to place a tracking radar on the beam of the aircraft so it will be excluded.

NULL, ANTENNA PATTERN - The directions of minimum transmission (or reception) of a directional antenna. See also Lobe, Antenna.

NULL FILL - The nulls in an antenna pattern may be reduced (filled) by using a second ancillary (spoiler) antenna whose pattern is such that it fills in the nulls of the main antenna pattern.

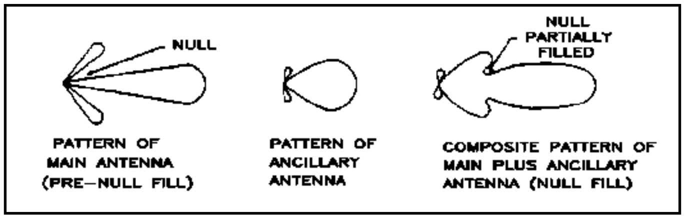
NULL
NULL PARTIALLY FILLED
PATTERN OF MAIN ANTENNA (PRE-NULL FILL)
PATTERN OF ANCILLARY ANTENNA
COMPOSITE PATTERN OF MAIN PLUS ANCILLARY ANTENNA (NULL FILL)

NUTATION - As applied to current missile system radars, this term refers to the mechanical motion of an antenna feed to produce a conical scan (fixed polarization) by the main beam of a tracking antenna, thus providing a means of developing tracking error signals. See also Antenna, Nutating. By analogy, "Nutation" also is used to denote the electrical switching of the quadrants of a seeker antenna. See also Interferometer. The effect is similar to that of a conical scan.

NUTATOR - A motor-driven rotating antenna feed used to produce a conical scan for a tracking radar. See also Antenna, Nutating. Also, the electrical circuits necessary to effect nonmechanical conical scans. See also Nutation.

OPERATIONAL CONSTRAINTS - Limitations on operating procedures in order to prevent interference between missile systems on a ship or between missile systems in a formation of ships under operational conditions. These limitations consist of such things as limited frequency bands or channels in which the radars may be tuned, limited sectors of space into which radar beams may be pointed, limits on minimum spacing between ships, limits on what codes may be used by radars and missiles on each ship, and limits on minimum interval between firing of certain missiles.

OSCILLATORS - Devices which generate a frequency. See also Backward Wave, Dielectrically Stabilized Oscillator, Hyperabrupt Varactor Oscillator, Magnetron Oscillator, Varactor Tuned Oscillator, and YIG tuned oscillator.

OSCILLATOR, LOCAL - See Local Oscillator Frequency.

PALMER SCAN - Conical scan superimposed on another type of scan pattern - usually a spiral pattern.

PARAMETER - A quantity which may have various values, each fixed within the limits of a stated case or discussion. In the present case, some examples of parameters; would be: radar frequency, limited by the tuning range of the radar; missile range, limited by the maximum operating range of the missile; or a missile code, limited by the number of codes available and by the codes that the ship radars are set up to operate on.

PASSIVE ANGLE TRACKING - Tracking of a target using radiation from the target (such as jamming), with no radiation from the radar itself. Only angular tracking is possible under these conditions since no measurement of time of travel of radiation to the target is possible, as is required to obtain target range.

PC (PULSE COMPRESSION) - The process used in search and tracking pulse radars whereby the transmitted pulse is long, so as to obtain high average transmitter output power, and the reflected pulse is processed in the radar receiver to compress it to a fraction of the duration of the transmitted pulse to obtain high definition and signal strength enhancement. Pulse compression may be accomplished by sweeping the transmitted frequency (carrier) during the pulse. The returned signal is then passed through a frequency-dependent delay line. The leading edge of the pulse is therefore delayed so that the trailing edge catches up to the leading edge to produce effectively a shorter received pulse than that transmitted. Pulse compression radars are also referred to as CHIRP radars. Other more sophisticated pulse compression techniques are also possible and are becoming more popular.

PENCIL BEAM - A narrow circular radar beam from a highly directional antenna (such as a parabolic reflector).

PHASED ARRAY RADAR - Radar using many antenna elements which are combined in a controlled phase relationship. The direction of the beam can be changed as rapidly as the phase relationships (usually less than 20 microseconds). Thus, the antenna typically remains stationary while the beam is electronically scanned. The use of many antenna elements allows for very rapid and high directivity of the beam(s) with a large peak and/or average power. There is also a potential for greater reliability over a conventional radar since the array will fail gracefully, one element at a time.

PIN DIODE - A diode with a large intrinsic (I) region sandwiched between the P- and N- doped semiconducting regions. The most important property of the PIN diode is the fact that it appears as an almost pure resistance at RF. The value of this resistance can be varied over a range of approximately one-10,000 ohms by direct or low frequency current control. When the control current is varied continuously, the PIN diode is useful for attenuating, leveling and amplitude modulation of an RF signal. When the control current is switched on and off or in discrete steps, the device is useful in switching, pulse modulating, and phase shifting an RF signal.

POINT CONTACT DIODE - This was one of the earliest semiconductor device to be used at microwave frequencies. Consisting of a spring-loaded metal contact on a semiconducting surface, this diode can be considered an early version of the Schottky barrier diode. Generally used as a detector or mixer, the device is somewhat fragile and limited to low powers.

POLARIZATION - The direction of the electric field (E-field) vector of an electromagnetic (EM) wave. See Section 3-2. The most general case is elliptical polarization with all others being special cases. The E-field of an EM wave radiating from a vertically mounted dipole antenna will be vertical and the wave is said to be vertically polarized. In like manner, a horizontally mounted dipole will produce a horizontal electric field and is horizontally polarized. Equal vertical and horizontal E-field components produce circular polarization.

PORT - The left side of a ship or aircraft when facing the bow (forward)

POWER (AVERAGE) FOR PULSED RADARS - Average power for a pulse radar is the average power transmitted between the start of one pulse and the start of the next pulse (because the time between pulses is many times greater than the pulse duration time, the average power will be a small fraction of peak power).

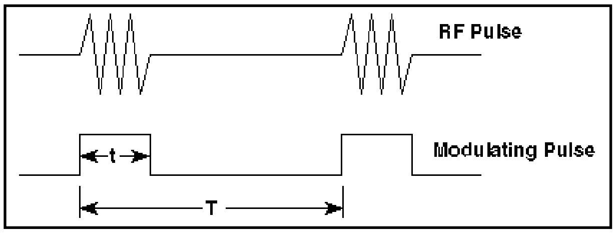

For this example: Peak Power = 1 MW, Pulse Time (t) = 0.5 microsecond, and Interval Between Pulses (T) = 1000 microseconds (1000 pps).

Peak Power = Pwr during pulse time (t) = 1 MW = 10^6 Watts = 90 dBm.

Avg Power = Average Power During Time (T) = 10^6 x t/T
= 10^6 x 0.5/1000 = 0.5 x 10^3 = 0.5 kilowatt = 57 dBm or 27 dBW

POWER OUTPUT - Power output of a transmitter or transmitting antenna is commonly expressed in dBW or dBm. One megawatt would be expressed as 60 dBW or 90 dBm:

10 log (1 megawatt / 1 watt) = 10 log ($10^6/10^0$)
= 10 x 6 = 60 dBW

10 log (1 megawatt / 1 milliwatt) = 10 log ($10^6/10^{-3}$)
= 10 x 9 = 90 dBm

POWER (PEAK) FOR PULSED RADARS - Peak power for a pulsed radar is the power radiated during the actual pulse transmission (with zero power transmitted between pulses).

POWER FOR CW RADARS - Since the power output of CW transmitters (such as illuminator transmitters) usually have a duty cycle of one (100%), the peak and average power are the same.

POWER DENSITY - The density of power in space expressed in Watts/meter^2, dBW/m^2, etc. Generally used in measurements above 100 MHz. At lower frequencies, field intensity measurements are taken. See Section 4-1.

PPI-SCOPE - A radar display yielding range and azimuth (bearing) information via an intensity modulated display and a circular sweep of a radial line. The radar is located at the center of the display.

PRESELECTOR - A device placed ahead of the mixer in a receiver, which has bandpass characteristics such that the desired (tuned) RF signal, the target return, is allowed to pass, and other undesired signals (including the image frequency) are attenuated.

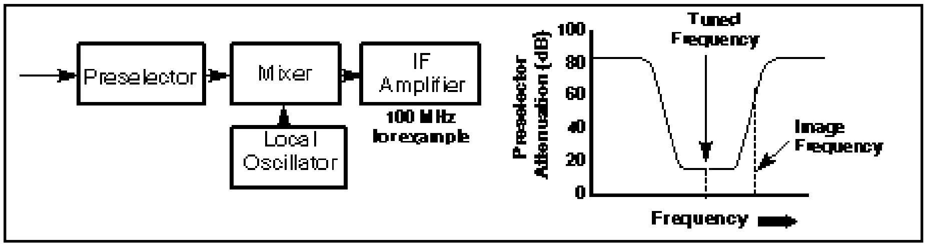

PROPAGATION - In electrical practice, the travel of waves through or along a medium. The path traveled by the wave in getting from one point to another is known as the propagation path (such as the path through the atmosphere in getting from a transmitting antenna to a receiving antenna, or the path through the waveguides and other microwave devices in getting from an antenna to a receiver).

PULSE COMPRESSION - See PC.

PULSED DOPPLER (PD) - A type of radar that combines the features of pulsed radars and CW Doppler radars. It transmits pulses (instead of CW) which permits accurate range measurement. This is an inherent advantage of pulsed radars. Also, it detects the Doppler frequency shift produced by target range rate which enables it to discriminate between targets of only slightly different range rate and also enables it to greatly reduce clutter from stationary targets. See also Doppler Effect.

PULSE LENGTH - Same meaning as Pulsewidth.

PULSE MODULATION - A special case of amplitude modulation wherein the carrier wave is varied at a pulsed rate. Pulse Modulation - The modulation of a carrier by a series of pulses generally for the purpose of transmitting data. The result is a short, powerful burst of electromagnetic radiation which can be used for measuring the distance from a radar set to a target.

PULSE REPETITION FREQUENCY (PRF) - The rate of occurrence of a series of pulses, such as 100 pulses per second. It is equal to the reciprocal of the pulse spacing (T) or PRT. (PRF = 1/T = 1/PRI). Sometimes the term pulse repetition rate (PRR) is used.

PULSE REPETITION FREQUENCY (PRF) STAGGER - The technique of switching PRF (or PRI) to different values on a pulse-to-pulse basis such that the various intervals follow a regular pattern. This is useful in compensating for blind speeds in pulsed MTI radars. Interpulse intervals which differ but follow a regular pattern.

PULSE REPETITION INTERVAL (PRI) or TIME (PRT) - Time between the beginning of one pulse and the beginning of the next.

PULSE SPACING - The interval of time between the leading edge of one pulse and the leading edge of the next pulse in a train of regularly recurring pulses. See also Pulse Repetition Frequency. Also called "the interpulse period."

PULSEWIDTH - The interval of time between the leading edge of a pulse and the trailing edge of a pulse (measured in microseconds for the short pulses used in radar). Usually measured at the 3 dB midpoint (50-percent power or 70% voltage level) of the pulse, but may be specified to be measured at any level. See Section 6-10 for measurement techniques.

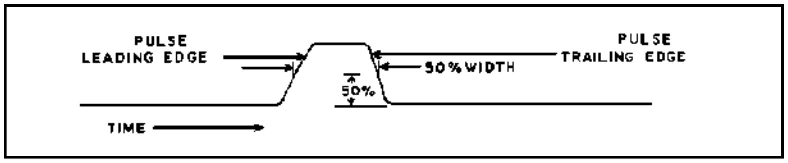

QUANTIZE - The process of restricting a variable to a number of discrete values. For example, to limit varying antenna gains to three levels.

RADAR - Radio detection and ranging.

RADAR CROSS SECTION - A measure of the radar reflection characteristics of a target. It is equal to the power reflected back to the radar divided by power density of the wave striking the target. For most targets, the radar cross section is the area of the cross section of the sphere that would reflect the same energy back to the radar if the sphere were substituted. RCS of sphere is independent of frequency if operating in the far field region. See Section 4-11.

RADAR RANGE EQUATION - The radar range equation is a basic relationship which permits the calculation of received echo signal strength, if certain parameters of the radar transmitter, antenna, propagation path, and target are known.

Given: $P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4}$ *(freespace)* as the basic two-way radar equation (see Sections 4-4 thru 4-6)

where:

P_r = Peak power at receiver input

λ = Wavelength of signal (length) = c/f

P_t = Transmitted signal level (power)
R = Range of target to radar (distance)
G_t = Gain of transmitting antenna (dimensionless ratio)
σ = Radar cross section of target
G_r = Gain of receiving antenna (dimensionless ratio)

In practical use, the radar range equation is often written in logarithmic form, all terms expressed in decibels, so that the results can be found by simple processes of addition and subtraction. Using the above equation and $\lambda = c/f$.

$$10 \log P_r = 10 \log P_t + 10 \log G_t + 10 \log G_r + 10 \log \sigma - 40 \log R - 20 \log f + 20 \log c - 30 \log 4\pi$$

where: f = Signal frequency (cycles {dimensionless}/time) c = Speed of light (length/time)

$$10 \log P_r = 10 \log P_t + 10 \log G_t + 10 \log G_r + G_\sigma - 2\alpha_1$$

where α_1 and G_σ are factors containing the constants and conversion factors to keep the equations in consistent units.

Refer to Sections 4-4 through 4-6.

RADAR TRIGGER KILL - see Trigger Kill, Radar.

RADIATION EFFICIENCY - $E = P_{radiated}/P_{in}$ (ideal=1)

RADIATION PATTERN - See Antenna Pattern.

RADIO FREQUENCY - See RF.

RADIO FREQUENCY INTERFERENCE - See RFI.

RAIL KEEPING - Ability of countermeasures to keep the missile on the launch rail, i.e., prevent launch.

RAINBOW - A technique which applies pulse-to-pulse frequency changing to identifying and discriminating against decoys and chaff.

RANGE CELL - In a radar, a range cell is the smallest range increment the radar is capable of detecting. If a radar has a range resolution of 50 yards and a total range of 30 nautical miles (60,000 yds), there are: 60000/50 = 1,200 range cells.

RANGE GATE - A gate voltage used to select radar echoes from a very short range interval.

RANGE GATE PULL OFF (RGPO) - Deception technique used against pulse tracking radars using range gates. Jammer initially repeats the skin echo with minimum time delay at a high power to capture the AGC circuitry. The delay is progressively increased, forcing the tracking gates to be pulled away ("walked off") from the target echo. Frequency memory loops (FML's), or transponders provide the variable delay.

RANGE RATE - The rate at which a radar target is changing its range with respect to the radar (in feet per second for example). Note that this rate is not the same as target velocity unless the target is moving straight toward or straight away from the radar.

RANGE SCOPE - See A-Scope or PPI.

RECEIVER BANDWIDTH - The difference between the limiting frequencies within which receiver performance in respect to some characteristic falls within specified limits. (In most receivers this will be the difference between the two frequencies where the intermediate frequency (IF) amplifier gain falls off 3 dB from the gain at the center IF frequency.) See also Receiver Selectivity.

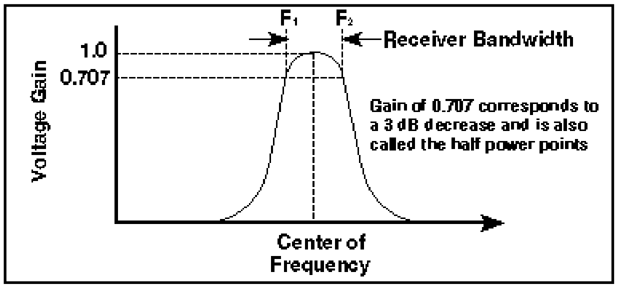

RECEIVER SELECTIVITY - The degree to which a receiver is capable of differentiating between the desired signal and signals or interference at other frequencies. (The narrower the receiver bandwidth, the greater the selectivity.)

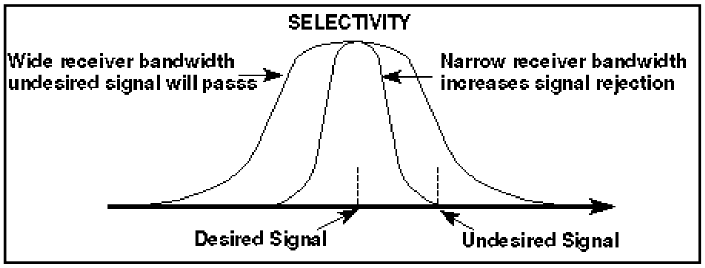

REFLECTION - The turning back (or to the side) of a radio wave as a result of impinging on any conducting surface which is at least comparable in dimension to the wavelength of the radio wave.

RESOLUTION - In radar, the minimum separation in angle or in range between two targets which the radar is capable of distinguishing.

RF (RADIO FREQUENCY) - A term indicating high frequency electromagnetic energy.

RFI (RADIO FREQUENCY INTERFERENCE) - Any induced, radiated, or conducted electrical disturbance or transient that causes undesirable responses or malfunctioning in any electrical or electronic equipment, device, or system. Same as EMI. Not to be confused with the logistic term ready for issue (also RFI).

RING AROUND - A condition in which a repeater jammer's total gain, from receiver antenna to transmitter antenna, exceeds the antenna isolation resulting in the repeater amplifying its own internal noise. Akin to positive feedback in an amplifier that causes unwanted oscillations.

RING AROUND (RADAR-TO-MISSILE) - The condition where radio frequency interference signals from a transmitter of one missile radar enter the receiving circuits of a missile under the control of another missile radar.

RING AROUND (RADAR-TO-RADAR) - The condition where radio frequency interference signals from a transmitter of one radar enter the receiving circuits of another radar.

ROPE - An element of chaff consisting of a long roll of metallic foil or wire which is designed for broad, low-frequency response. See Chaff.

R-SCOPE - (RANGE SCOPE) See A-scope or PPI.

SAFETY OF FLIGHT (SOF) TEST - A flight test to verify that a new or modified subsystem will not cause a major problem with the aircraft, i.e., interference can occur, but will not be such that required navigational systems will fail or which might potentially cause the loss of an aircraft under all normally expected weather conditions.

SCAN - To transverse or sweep a sector or volume of airspace with a recurring pattern, by means of a controlled directional beam from a radar antenna. See also “Antenna, nutating.”

SCHOTTKY BARRIER DIODE - The Schottky barrier diode is a simple metal-semiconductor boundary with no P-N junction. A depletion region between the metal contact and the doped semiconductor region offers little capacitance at microwave frequencies. This diode finds use as detectors, mixers, and switches.

SCHOTTKY DIODE MIXER - The mixer is a critical component in modern RF systems. Any nonlinear element can perform the mixing function, but parameters determining optimal mixing are noise figure, input admittance, and IF noise and impedance. The Schottky diode is particularly effective because of its low noise figure and nearly square law characteristics.

SCHOTTKY DIODE SWITCH - Standard P-N diodes are limited in switching ability at high frequencies because of capacitance provided by the minority carriers. The Schottky diode overcomes this problem by use of the metal-semiconductor junction with inherently low carrier lifetimes, typically less than 100 picoseconds.

SEARCH RADAR - A radar whose prime function is to scan (search) a specified volume of space and indicate the presence of any targets on some type of visual display, and, in some cases, to provide coordinates of the targets to a fire control system to assist in target acquisition and tracking.

SEEKER - The seeker consists of circuitry in a homing missile which detects, electronically examines, and tracks the target; provides data for controlling the flight path of the missile; and provides signals for destroying the missile or for detonating it at intercept. (The seeker function is similar to that of an interferometer.)

SELF-SCREENING JAMMING (SSJ) - Each aircraft carries its own jamming equipment for its own protection.

SENSITIVITY - The sensitivity of a receiver is taken as the minimum signal level required to produce an output signal having a specified signal-to-noise ratio. See also Minimum Visible Signal and Minimum Discernible Signal (MDS).

SENSITIVITY TIME CONTROL - See STC.

SENSOR - The receiver portion of a transmitter/receiver pair used to detect and process electromagnetic energy.

SHIELDING - The physical arrangement of shields for a particular component, equipment, or system, (A shield is a housing, screen, or other material, usually conducting, that substantially reduces the effect of electric or magnetic fields on one side of the shield upon devices or circuits on the other side.) Examples are tube shields, a shielded enclosure or cabinet for a radar receiver, and the screen around a screen room.

SHORT PULSE MODE - See Long Pulse Mode.

SIDEBAND - A signal either above or below the carrier frequency, produced by the modulation of the carrier wave by some other wave. See figure at right ⇒

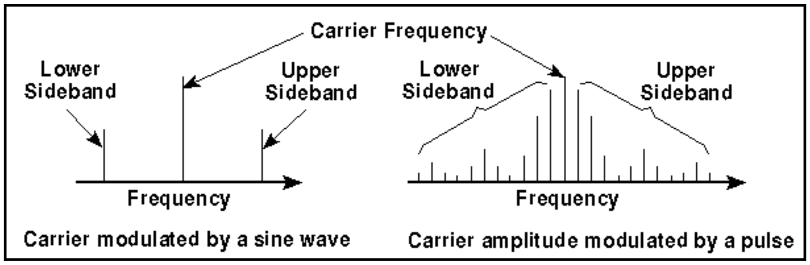

SIDELOBE - See Lobe, Antenna.

SIGNAL STRENGTH - The magnitude of a signal at a particular location. Units are volts per meter or dBV/m.

SIGNATURE - The set of parameters which describe the characteristics of a radar target or an RF emitter and distinguish one emitter from another. Signature parameters include the radio frequency of the carrier, the modulation characteristics (typically the pulse modulation code), and the scan pattern.

SILICON CONTROLLED SWITCH - A P-N-P-N device able to operate at sub-microsecond switching speeds by the application of gate signals. Because it is a four layer device, this switch is also known as a tetrode thyristor.

SLANT POLARIZATION - Technique of rotating a linear antenna 45° so it can receive or jam both horizontal and vertical polarization although there is a 3 dB loss. See Section 3.2.

SOLID STATE STAMO - A stable master oscillator constructed using transistors and other solid state devices as opposed to vacuum tubes. See also STAMO.

SPECTRUM - The distribution of power versus frequency in an electromagnetic wave. See also Spectrum Signature Analysis and illustrations under Sideband.

SPECTRUM ANALYZER - An electronic device for automatically displaying the spectrum of the electromagnetic radiation from one or more devices. A cathode ray tube display is commonly used to display this power-versus frequency spectrum. For examples of two types of displays, see illustrations under Sideband.

SPECTRUM SIGNATURE ANALYSIS - The analysis of the electromagnetic radiation from an electronic device to determine the relative power in each sideband, harmonic, and spurious emission compared to the carrier frequency. This particular distribution (or spectrum) is peculiar to the device and can identify this type of device, thereby acting as an identifying “signature.”

SPECTRUM WIDTH (TRANSMITTER) - The difference between the frequency limits of the band which contains all the spectrum frequency components of significant magnitude.

SPOILER ANTENNA - An antenna used to change (spoil) the antenna pattern of a second antenna so as to reduce the nulls in the pattern of the second antenna. See also Null Fill.

SPOKING (RADAR) - Periodic flashes of the rotating radial display. Sometimes caused by mutual interference.

“SPOOFING” - A type of deception by using an electronic device to transmit a “target” echo. The spoofing transmitter must operate at the same frequency and PRF as the radar to be deceived. The radar main pulse triggers the spoofing transmitter which, after a delay, transmits a false echo.

SPOT JAMMING - Narrow frequency band jamming concentrated against a specific radar at a particular frequency. The jamming bandwidth is comparable to the radar bandpass. Can deny range and angle information.

SPURIOUS EMISSION - Electromagnetic radiation transmitted on a frequency outside the bandwidth required for satisfactory transmission of the required waveform. Spurious emissions include harmonics, parasitic emissions, and intermodulation products, but exclude necessary modulation sidebands of the fundamental carrier frequency.

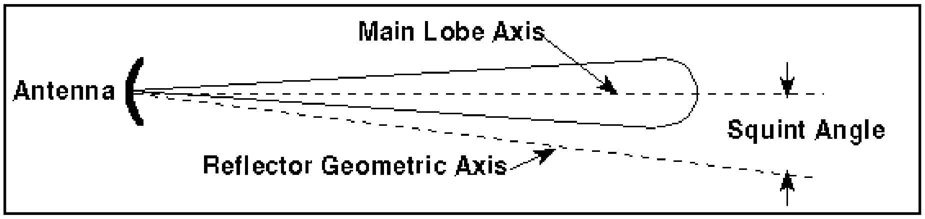

SQUINT ANGLE - The angular difference between the axis of the antenna mainlobe and the geometric axis of the antenna reflector, such as the constant angle maintained during conical scan as the mainlobe rotates around the geometric axis of the reflector.

STAGGERED PRF - Staggered PRF allows an increase in MTI blind speeds such that no zeros exist in the velocity response at lower velocities. In a two-period mode, the usual "blind speed" or occurrence of a zero in the velocity response is multiplied by a factor which is a function of the ratio of the two repetition periods.

STAMO (STABLE MASTER OSCILLATOR) - A very stable (drift free) oscillatory used to provide a precise frequency for transmission and for comparison with the reflected radar signal returned to the receiver, such as in a Doppler radar where a precise difference between transmitted and received signals must be measured to determine accurately the Doppler frequency.

STAND-FORWARD JAMMING - A method which places the jamming vehicle between the enemy sensors and attack aircraft.

STAND-IN JAMMING (SIJ) - Similar to stand-forward jamming but usually using an UAV with a lower powered jammer instead of a jammer aircraft.

STAND-OFF JAMMING (SOJ) - An ECM support aircraft orbits in the vicinity of the intended target. As the fighter-bomber pilot starts his strike penetration, the ECM aircraft directs jamming against all significant radars in the area. This technique provides broad frequency band ECM without affecting performance of the strike aircraft.

STARBOARD - The right side of a ship or airplane when facing the bow (forward).

STC (SENSITIVITY TIME CONTROL) - Gain control that reduces the radar receiver gain for nearby targets as compared to more distant targets. STC prevents receiver saturation from close-in targets.

STEP RECOVERY DIODE - A charge-controlled switch which ceases current conduction so rapidly that it can be used to produce an impulse. Cyclic operation of the diode can produce a train of impulses which when used with a resonant circuit can produce a single frequency output at any harmonic of the pulse frequency.

STERADIAN - Unit of solid angle. An entire sphere has 4π steradians.

STREAM CHAFF - Operational technique of dropping large quantities of chaff for a continuous period of time. This results in a “ribbon” or “stream” of returns many miles in lengths on radarscopes. The penetrating strike force can then use the resulting chaff corridor to mask their penetration.

SUBHARMONIC - A frequency which is an integral submultiple of another frequency. For example, a sine wave whose frequency is one-third of the frequency of another sine wave is called the third subharmonic. (3 MHz is the third subharmonic of 9 MHz).

SUPERHETERODYNE RECEIVER - A receiver that mixes the incoming signal with a locally generated signal (local oscillator) to produce a fixed, low intermediate frequency (IF) signal for amplification in the IF amplifiers.

SUPPRESSION - Elimination or reduction of any component of an emission, such as suppression of a harmonic of a transmitter frequency by band rejection filter.

SUPPRESSION OF ENEMY AIR DEFENSES (SEAD) - Activity which neutralizes, destroys, or temporarily degrades enemy air defense systems by using physical attack or electronic means (SEAD pronounced “seed” or “C add”).

SUSCEPTIBILITY - The degree to which an equipment or a system is sensitive to externally generated interference.

SWEPT JAMMING - Narrowband jamming which is swept through the desired frequency band in order to maximize power output. This technique is similar to sweeping spot noise to create barrage jamming, but at a higher power.

SWITCHES - See also Diode Switch, Silicon Controlled Switch, Schottky Diode Switch.

SYNCHRODYNE - A klystron mixer amplifier stage in a transmitter, where two signal frequencies are applied as inputs and a single amplified signal is taken out.

TARGET SIZE - A measure of the ability of a radar target to reflect energy to the radar receiving antenna. The parameter used to describe this ability is the “radar cross section” of the target. The size (or radar cross section) of a target, such as an aircraft, will vary considerably as the target maneuvers and presents different views to the radar. A side view will normally result in a much larger radar cross section than a head-on view. See also Radar Cross Section.

TERMINAL IMPEDANCE - The equivalent impedance as seen by the transmitter/receiver.

TERRAIN BOUNCE - Term for jamming that is directed at the earth’s surface where it is reflected toward the threat radar. Reflected jamming creates a virtual image of the jamming source on the earth as a target for HOJ missiles.

THERMISTOR - A resistor whose resistance varies with temperature in a defined manner. The word is formed from the two words “thermal” and “resistor,”

THRESHOLD ISR - The interference to signal ratio (ISR) at which the performance of a receiver starts undergoing degradation. It must be determined by tests.

TRACKING RADAR - A radar whose prime function is to track a radar target and determine the target coordinates (in range and angular position) so that a missile may be guided to the target, or a gun aimed at the target.

TRACKING RADAR RECEIVER - These are of two primary types: conical scan and monopulse. (1) The conical scan system directs the radar signal in a circle around the target. The radar paints this circle 15 to 40 times per second. As the target moves out of the center of this circle, the radar develops aim error voltages and re-aims the antenna. (2) The monopulse system directs four beams at the target simultaneously. The target is in the middle of the four beams. If the target is not in the center, the radar return develops an aim error voltage to re-aim the antenna.

TRACK WHILE SCAN (TWS) RADAR - Although it is not really a tracking radar in the true sense of the word, it does provide complete and accurate position information for missile guidance. In one implementation it would utilize two separate beams produced by two separate antennas on two different frequencies. The system utilizes electronic computer techniques whereby raw datum is used to track an assigned target, compute target velocity, and predict its future position, while maintaining normal sector scan. Most aircraft use only a single antenna.

TRADE-OFF TABLES - A set of tables showing the various combinations of two or more variables that are related in that making one variable better will make the other variable worse. The trade-off helps find the best solution considering all combinations. (For example, how a no-interference condition can be maintained if two emitter platforms are brought close together, if at the same time the frequency separation between their radar transmitters is increased.)

TRANSIENT - A phenomenon (such as a surge of voltage or current) caused in a system by a sudden change in conditions, and which may persist for a relatively short time after the change (sometimes called ringing).

TRANSPONDER - A transmitter-receiver capable of accepting the electronic challenge of an interrogator and automatically transmitting an appropriate reply. There are four modes of operation currently in use for military aircraft. Mode 1 is a nonsecure low cost method used by ships to track aircraft and other ships. Mode 2 is used by aircraft to make carrier controlled approaches to ships during inclement weather. Mode 3 is the standard system used by commercial aircraft to relay their position to ground controllers throughout the world. Mode 4 is IFF. See also IFF.

TRAVELING-WAVE TUBE AMPLIFIER - The TWT is a microwave amplifier capable of operation over very wide bandwidths. In operation, an electron beam interacts with a microwave signal which is traveling on a slow wave helical structure. The near synchronism of the beam and RF wave velocities results in amplification. Bandwidths of 3:1 are possible. Operation at high powers or at millimeter wavelengths is possible at reduced bandwidths.

TRIGGER KILL (RADAR) - A method employed to momentarily disable certain radar system circuits to reduce or eliminate RF emissions which may cause an EMI/EMC or RADHAZ situation such as on the deck of a ship.

TUNNEL DIODE - The tunnel diode is a heavily doped P-N junction diode that displays a negative resistance over a portion of its voltage-current characteristic curve. In the tunneling process, electrons from the p-side valence bands are able to cross the energy barrier into empty states in the N-side conduction band when a small reverse bias is applied. This diode is used as a microwave amplifier or oscillator.

UPLINK - The missile guidance signal which passes midcourse correction command guidance intelligence from the guidance radar site to the missile.

VARACTOR DIODE - A P-N junction employing an external bias to create a depletion layer containing very few charge carriers. The diode effectively acts as a variable capacitor.

VARACTOR TUNED OSCILLATOR - A varactor diode serves as a voltage-controlled capacitor in a tuned circuit to control the frequency of a negative resistance oscillator. The major feature of this oscillator is its extremely fast tuning speed. A limiting factor is the ability of the external voltage driver circuit to change the voltage across the varactor diode, which is primarily controlled by the driver impedance and the bypass capacitors in the tuning circuit.

VELOCITY GATE PULL-OFF (VGPO) - Method of capturing the velocity gate of a Doppler radar and moving it away from the skin echo. Similar to the RGPO, but used against CW or Doppler velocity tracking radar systems. The CW or pulse doppler frequency, which is amplified and retransmitted, is shifted in frequency (velocity) to provide an apparent rate change or Doppler shift.

VELOCITY MODULATION - Velocity modulation is modification of the velocity of an electron beam by alternately accelerating and decelerating the electrons at a frequency equal to the input frequency. Thus, the electrons are segregated in bunches, each bunch causing a cycle or current as it passes an output electrode. The velocity of the electrons is thus a function of the modulation voltage. See also Klystron, Multicavity and Klystron, Reflex.

VICTIM - A receiver (radar or missile) that suffers degradation due to ECM or EMI effects.

VIDEO - Receiver RF signals that have been converted (post detection) into a pulse envelope that can be seen when applied to some type of radar visual display; also used to describe the actual display itself (such as the video on an A-scope).

WARM - Acronym for Wartime Reserve Mode. Any mode of operation of a radar or ECM that is held in reserve, and never used, except in actual combat.

WATER LINE - A reference line used for vertical measurements. When used with an aircraft it is usually the ground with the landing gear extended normally. One of several aircraft location designations, also see butt line and fuselage station.

WAVEGUIDE - A transmission line consisting of a hollow conducting tube of arbitrary geometry (usually rectangular, but may be circular) within which electromagnetic waves may propagate.

WAVELENGTH (λ) - The distance traveled by a wave in one period (the period is the time required to complete one cycle). $\lambda = c/f$. In the atmosphere, electromagnetic waves travel at c, the speed of light (300 million meters per second or 30 cm/nsec). At 5 GHz, one wavelength = 6 cm. At 10 GHz, one wavelength = 3 cm.

WAVEMETER - An instrument for measuring the frequency of a radio wave. The wavemeter is a mechanically tunable resonant circuit. It must be part of a reflection of transmission measurement system to measure the maximum response of a signal. Below 20 GHz, the wavemeter has been replaced by the frequency counter with much greater accuracy and ease of use.

WEAVE - An aircraft maneuver that smoothly changes the instantaneous flight path but maintains the overall route of flight. Not as violent as a jink.

WET RUN - A test run with ship / aircraft armament and/or EW switches on.

WILD WEASEL - USAF aircraft (F-4Gs during Desert Storm) used for suppression of enemy air defense (SEAD) mission.

WINDOW - WWII name for chaff

YIG TUNED OSCILLATOR - A YIG (yttrium iron garnet) sphere, when installed in the proper magnetic environment with suitable coupling will behave like a tunable microwave cavity with Q on the order of 1,000 to 8,000. Since spectral purity is related to Q, the device has excellent AM and FM noise characteristics.

ZENER DIODE - A diode that exhibits in the avalanche-breakdown region a large change in reverse current over a very narrow range in reverse voltage. This characteristic permits a highly stable reference voltage to be maintained across the diode despite a wide range of current.

Donald Marshall On Megadeth's "The World Needs A Hero" Album Cover

Empowerment by Virtue of Golden Truth
Human Cloning: Specifically,
REM Driven Human Cloning
FULL DISCLOSURE

Donald Marshall

29 October 2015

https://www.facebook.com/donald.marshall.148
http://donaldmarshall.proboards.com/
Vincent Eastwood Interview
https://www.youtube.com/watch?v=M_1UiFeV5Jg&ab

Contents

List of Images 5

Introduction 6

Whistle Blowing: Qualifying Disclosure 6

The Serious Crime Act 2015 7

Fundamentals 8

Key Figures 9

- *Phil Schneider* 9
- *George Green* 11
- *Aaron Russo* 13
  - *Income Tax* 13
  - *Radio Frequency Identification (RFID) Microchips* 14
  - *Habeas Corpus* 15
  - *Fraudulent Manipulation of Election Results* 15
  - *The War on Terrorism is the War on Your Freedom* 16
  - *Whoever Makes the Money Makes the Rules* 16
  - *Russo's Message to Mankind* 16
  - *Let's Consolidate our Problems* 17
- *Aaron Russo: Reflections and Warnings* 17
  - *The Falsehood of September 11th 2001* 17
  - *The New World Order Agenda* 19
  - *Women's Liberation from the Perspective of the Elite* 20
  - *America is a Republic* 20
  - *Depopulation* 21
  - *Borrowing Money from Private Banks causes Inflation and Debt* 21
  - *The Deception of the Council of Foreign Relations (CFR)* 23
  - *Combining America, Mexico and Canada into one Country* 24
  - *Russo's Vision on How to Bring These People Down* 25

The Trans-humanist / Post Human Agenda 26

The Phases of Sleep and Rapid Eye Movement (R.E.M) Sleep 27

REM Sleep 28

Current Advances in Technology 28

- *Memory Suppression Technologies* 29
- *Mind-Voice Technology* 29
- *H.A.A.R.P. Technology* 29
- *H.A.A.R.P. and Weather Control* 30
- *H.A.A.R.P. and Mind Control* 31
- *RFID Technology* 32

  RFID Technology and the Medical Profession 32
  Further Implications of RFID Technology 33
  Mind Control: MK Ultra Technology 33
  Human Cloning 35
    Cloning Centre and Cloning Technology 36
Drip Feed Disclosure and Evaluative Conditioning 36
  Drip Feed Disclosure 36
  Evaluative Conditioning 37
Consciousness Transfer 38
Disclosure: Save the Victims through your Diligence 40
  Words of Inspiration: Personalized from a Theoretical Physicist –Thomas Campbell 41
  The Science of Hidden Consciousness 41
Disclosure: Donald Marshall's Message to the World 42
  The Two Types of Cloning Techniques 42
  The Different Grades of Clones: "Mark 1" to "Mark 4" Clones 43
  Donald Marshall's Dilemma with "Mark 2" Sleep Driven REM Clones 43
  Consciousness Transfer Happens When the Original Reaches REM Sleep 43
  Donald Marshall is an Original as he sits at Home Typing this Disclosure 44
  When Donald Marshall's REM Driven Cloned Experiences Began 45
  The Science of REM Driven Cloning Began in 1945 46
  Help Donald Marshall STOP REM Driven Cloning NOW! 47
  Do Not Panic, Riot, or Cause Chaos. This is Very Important. 47
  Spread this Disclosure Document until it reaches the Armed Forces 47
  Donald Marshall is NOT "The New Age Saviour". Nobody wants to be a 'saviour' 48
  How the Nostradamus Prophecies shapes the lives of the Illuminati 49
  HIV/AIDS has been cured; so have many forms of cancer 49
  The Behaviour of High Profile People at the Cloning Centre 50
  The Ring Leaders of the Cloning Centre 50
  The Royal Family of England 53
  Help Donald Marshall by Sharing this Disclosure with Your Close Friends and Family 57
  The Deplorable Nature of the Illuminati 59
  What Happened to Britney Spears? 60
  NEVER EVER SELL YOUR SOUL 62
    In This Physical World Selling Your Soul Is Selling Your REM Driven Clone 62
  Fake Alien Abductions conducted through REM Driven Cloning Technology 64
  I am THE Spy for Humanity against the Corruption of the Illuminati 65
  The Threat to Humanity when the H.A.A.R.P Grid is Complete 66
  REM Driven Cloning is the Most Terrible Nightmare Situation Ever 68

- *The Meek Shall Inherit the Earth* 70
- *Be Wary of: Shills; Trolls; Disinformation Agents and Double Agents* 72

Contact Information 73

Interviews 74

- *Links to Donald Marshall Interviews* 76
  - *Vincent Eastwood* 76
  - *Jeanice Barcelo* 76
  - *Lisa Phillips (CFR)* 76

Radio Presenters –Contact Donald Marshall 77

Professionals who understand "Consciousness Transfer" –Contact Donald Marshall 77

Email 77

How to Learn More about Donald Marshall's Whistle Blowing Disclosures 77

How Much Do You Know About Post-humanism / Trans-humanism? 78

This IS your ONLY chance to end these monsters. There is a deadline! 79

Do NOT Be Afraid to Help. We Outnumber the Illuminati by 1,000,000:1 80

Noteworthy Frequently Asked Question 81

Donald Marshall's Mission and Vision on How to Bring Down The Illuminati 81

Empowerment by Virtue of Golden Truth 83

Appendices 84

- *Appendix A: Thien Thanh Thi Nguyen (Tila Tequila) Transcript* 84
- *Appendix B: MK Ultra* 87
- *Appendix C: The Illuminati* 89

References 99

Legislation 111

List of Images

Image 1: Stages of sleep.. 27
Image 2: The ionosphere's relationship to earth.. 30
Image 3: Megadeth Album Cover: *The World Needs a Hero* (2001).. 58
Image 4: Donald Marshall’s original body compared with his REM Driven Duplicate Clone. 58
Image 5: Britney Spears Japanimation: Britney Spears grown as a duplicate clone. 61
Image 6: Britney Spears Japanimation: Britney Spears walks into the Cloning Centre. 61
Image 7: Britney Spears Japanimation: The Cloning Centre depicted from an aerial view. 61
Image 8: Possible H.A.A.R.P. Locations. 67

Introduction

Just read it. Every word in this document is true, and it is the most important document to ever reach the internet. This document is written to empower good people of the world against the tyranny which exists all around us in our world today. Effectively, the document presented here discloses concealed information about the extent to which medicine, science, and technology has advanced in our **current** era. Moreover, I am disclosing this information because I have been extensively abused, and I am victim of **current** advances in medicine, science, and technology, which I describe in detail below. There are many, many, other victims of abuse, including myself, and therefore it is our plea that you read this document with an open mind and you investigate **ALL** statements which may initially irk you and appear 'suspect' to begin with (the circumstantial evidence is available all around you); because as a victim of abuse, I can honestly say, there is nothing worse than experiencing abuse, and those who have the power to protect you from your abusers, ignore you.

I will now discuss what qualifies as whistle blowing disclosure. After detailing the criteria which qualifies for whistle blowing disclosure, I will detail aspects of the Serious Crime Act, and preface a foundation which defines, explains, and references key information for the reader to understand as well as explore the extent to which medicine, science, and technology has advanced today; all for the express purpose of compelling the reader that medicine, science, and technology, has advanced at an incredible rate; that **current** advances in medicine, science and technology are withheld from public knowledge; and finally these advancements in medicine, science, and technology, have been and are **currently** used to commit **serious unspeakable** crimes against humanity and our environment here on earth. Moreover, it is my aim to present this information as coherently as possible to assist the reader to save me, and save many, many others who experience abuse daily resulting from concealed advances in technology.

Whistle Blowing: Qualifying Disclosure

In the United Kingdom under the Public Interest Disclosure Act (1998) (PIDA) section 43B, a "qualifying disclosure" (whistle blowing disclosure) implies "any disclosure of information which in the reasonable belief of the worker, tends to show one of the following:

a) That a criminal offence has been committed, is being committed or is likely to be committed,
b) That a person has failed, is failing or is likely to fail to comply with any legal obligation to which he is subject
c) That a miscarriage of justice has occurred, is occurring or is likely to occur,
d) That the health or safety of any individual has been, is being or is likely to be endangered,
e) That the environment has been, is being, or is likely to be damaged, or
f) That information tending to show any matter falling within any one of the preceding paragraphs has been, or is likely to be deliberately concealed"

Furthermore, under the Enterprise and Regulatory Reform Act (ERRA) 2013 section 17 (*The new 'public interest' test)* is added for whistle blowing disclosures, amending section 43B of the Employment Rights Act (ERA) 1996 and therefore now reads as such (new words underlined):

"(1) In this part a 'qualifying disclosure' means any disclosure of information which, in the reasonable belief of the worker making the disclosure, *is made in the public interest and* tends to show one or more of the following-- [criminal offence, breach of legal obligation, etc]." (Halliday 2013, p. 2).

The following information is presented **with the express intent** that it is made in the public interest, and meets the all criteria set out in section 43B (a-f) of PIDA as well as the 'public interest test', ERRA section 17. Furthermore, in order for the reader to perceive the credence of my statements, so that (s)he has impetus to research the validity of my statements and not dismiss it as folly supplied by an internet troll who has nothing better to do than waste peoples' time, I will cite the Serious Crime Act 2015 here for the express purpose of convincing the reader that all statements made in this document and my original document are indeed FACTUAL! THIS IS NO HOAX. THIS IS NO JOKE. This is simply what is happening in the world today.

The Serious Crime Act 2015

The Serious Crime Act 2015 (UK) section 41, 3ZA, amends section 3A of the Computer Misuse Act 1990 which **previously** read [under Computer Misuse Offences]:

[F1 3A Making, supplying or obtaining articles for use in offence under section 1 or 3

1) [That] A person is guilty of an offence if he makes, adapts, supplies or offers to supply any article intending it to be used to commit, or to assist in the commission of, an offence under section 1 or 3.

To: [The Serious Crime Act 2015, section 41, 3ZA]

3ZA Unauthorised acts causing, or creating risk of, serious damage

(1) A person is guilty of an offence if—

(a) the person does any unauthorised act in relation to a computer;

(2) Damage is of a "material kind" for the purposes of this section if it is—

(a) damage to human welfare in any place;

(b) damage to the environment of any place;

(c) damage to the economy of any country; or

(d) damage to the national security of any country.

3) For the purposes of subsection (2)(a) an act causes damage to human welfare only if it causes—

(a) loss to human life;

(b) human illness or injury;

(4) It is immaterial for the purposes of subsection (2) whether or not an act causing damage—

(a) does so directly;

(b) is the only or main cause of the damage.

(5) In this section—

(c) a reference to a country includes a reference to a territory, and to any place in, or part or region of, a country or territory.

(6) A person guilty of an offence under this section is (unless subsection (7) applies) liable, on conviction on indictment, to imprisonment for a term not exceeding 14 years, or to a fine, or to both.

(7) Where an offence under this section is committed as a result of an act causing or creating a significant risk of—

(a) serious damage to human welfare of the kind mentioned in subsection (3)(a) or (3)(b), or

(b) serious damage to national security,

a person guilty of the offence is liable, on conviction on indictment, to imprisonment for life, or to a fine, or to both."

In laymen's terms it is a **serious crime** (Serious Crime Act 2015), punishable by imprisonment, a fine, or both, to cite hatred or spread hoaxes by use of a computer which would cause mass hysteria or public dissension. I have been condemning public figures, high profile individuals, celebrities, singers, movie stars, directors, politicians, world leaders, royalty, anti-secret society 'truthers' (who are really misinformation agents, and agents of deception) etc. –the list is long - since 2011 and **NOT** a single one of these high profile people has issued a public statement saying "I do not like how you have tarnished my name" or filed a law suit against me. Why you may ask? **Because** every word I have stated and will state again (in this document) is true.

Fundamentals

Before I describe the heinous crimes committed against me and many others (including unsuspecting civilians of the world) it is important that I explain, define, and reference:

- Key figures such as Phil Schneider, George Green and Aaron Russo, so that the reader has reference points to verify the extent to which technology has advanced in our **present** era (and continues to advance);
- Define and briefly explain the transhumanist / post human agenda (for those who may be unfamiliar with their aims).
- Define and explain rapid eye movement (R.E.M) sleep, the phases of sleep, and what happens to the (original's) body during sleep.

- Reference advances in technology, with particular attention to: Memory suppression technologies; Mind-voice technology; H.A.A.R.P technology, RFID microchip technology, and MK Ultra technology (CIA's mind control program);
- Define and explain cloning; the different types of clones; cloning centres and cloning technology;
- Define drip feed disclosure / evaluative conditioning and explain why it occurs;
- Explain what 'consciousness transfer' is to the best of my knowledge, and cite drip feed disclosure articles with the express intent to allow the reader to grasp the abuse I have suffered daily at the hands of my tormentors.

Key Figures

Phil Schneider

Phil Schneider (pictured) had 17 years experience working in government black projects carrying a level three security clearance. He was a geologist and engineer who worked in the black projects underground bases at Area 51, S-4, and Los Alamos.

He is most notable for disclosing (Schneider 1995; 1996; Open Minds 2011):

- The 'black budget' expenditure of the United States, which Schneider claims to be between 1.023 trillion U.S. dollars every 2 years (over $500 billion per year);
- Deep Underground Military Bases (also known as D.U.M.Bs -"dumbs"), and at the time of his lecture (Schneider, 1995), –that there are 131 **active** Deep Underground Military Bases present in the United States, and 1477 Deep Underground Military Bases worldwide;
- Each D.U.M.B costs on average 17-19 billion U.S. dollars; paid for by the taxpayer; and it takes approximately a year-and-a-half to 2 years to build D.U.M.Bs with sophisticated methods.
- That military technology outstrips the general public's technology at a rate of 44 to 45 years of technology for every calendar year which passes. In other words for every 12 months which passes military technology will have advanced by 44 years than what we as the general public is currently accustomed to. Therefore as a rough example if we were to take the year Facebook was founded (2004) as a base year, then the military technology since the creation of Facebook will have outstripped what the

general public is accustomed to by as much as **484 years(!)** (2015 - 2004 = 11. 11 x 44 = 484). Similarly, if we are to use the year which YouTube was founded as a base year (2005), once again, military technology would be 440 years more advanced than what the public is currently accustomed to today.

For the purpose of disclosing current advances in medicine, science and technology, and how such advances are used against me and the people of the world to commit monstrous crimes I will use 1945 as a base year. Everything will become apparent including why I use 1945 as a base year for my disclosure, but for the express purpose of compelling the reader to investigate my disclosure I must present everything, logically, sequentially, methodologically, and provide references for the reader in a coherent way to enable him or her to **pay serious attention** to my eye witness accounts. Therefore, at this present stage keep in mind the year 1945, and the fact that military technology outstrips the general public's technology at a rate of 44 years for every 12 months which passes.

George Green

George Green (pictured) was affiliated with U.S. Presidential candidates, and was once asked to be the Finance Chairman for the next President of the United States. Green would later declined the offer of Finance Chairman when a comment made by Ted Kennedy regarding sleeping with Green's 14 year old daughter caused George Green to re-evaluate his associations with this in-group.

During his 2008 interview with Project Camelot (2008a; 2008b), Green disclosed the following noteworthy information:

- U.S. presidents are "selected" and not elected. This is clearly expressed in a story Green recounts, when Green asks:
  "Who is going to be the next President of the United States [for him to overlook their finances]?"
  And the reply was: "Jimmy Carter"
  Green responds: "Jimmy who?"
  The reply Green received was: "Well, he's the Democratic Governor of Georgia."
  George Green: "But I've been voting Republican."
  Green was then confronted by a tall man, Paul Volcker (American Economist, and Chairman of Federal Reserve under Jimmy Carter and Ronald Regan), who walked over and said, "Son, don't worry about it [Republican or Democrat], we control 'em both."
- U.S. scientists learned how to make people (clones) since 1938 -walking talking ones –and the scientists call these people "synthetics" or "the others".
- Cloning technology is relatively advanced. All that has to be done is take two cells from the original, give the cells a small electrical charge (retain a fertilised egg), then all one needs is a receiver (a womb / artificial womb for the fertilised egg to grow).
- Scientist were excited by the synthetic technology because it meant that spare organ parts could be grown for an "original" human without rejection, because theoretically speaking, the DNA of the synthetic is the same as the original.
- Green gives an example of a cloned Politician: George Walker Bush. Green advices the viewer to seek old video recordings of George W. Bush, and compare the old George W. Bush, with the George W. Bush during Bush's second term,

-in terms of mannerism, speech pattern, body language etc. (old videos of George Walker Bush (Boringest 2006; Fox 4 News - Dallas-Fort Worth 2014) are presented in the References section).

- Scientist have also learned how to make these synthetic people within a few months, to the point where the synthetic can be a walking, talking duplicate of the original, intact with all the memories and experiences of the original. The only problem is that the memory, experiences, and functioning capabilities of these synthetics are like: “A DVD recorder. Sometimes you have glitches, and you have to take the synthetics to Camp David occasionally [every 6 months to a year] to get them tuned up.”
- Remember these synthetics are people too, they can think and act just like you can, but they do not have a soul.
- Most of the world leaders have been bought and paid for [by men in the background] and are created to think a certain way –to meet the agendas of these men who remain in the background.
- The global elite plans on depopulating the current human population of **over 7 billion** to **500 million** people. This is corroborated by the “Georgia Guidestones” (WorldTruth 2014).
- Vladimir Putin is executing plans to bring the United States down and into a massive depression. Furthermore, China agrees with Vladimir Putin’s plans; stating that the Chinese government has not been a Superpower for over 5000 years and “It is their turn to run the world.”
- China has the capacity to set back / shut down all of U.S. computing and electrical systems within two days.

At this present stage, the most **salient** points to keep in mind from Green’s accounts are that: U.S. presidents are selected and not elected; and **most importantly**, that scientists have been capable of creating synthetics (cloning people) since 1938. If you have kept in mind Schneider’s statement (1995; 1996), (that for every 12 months which passes military technology increases by the equivalent of 44 years compared to what the general public is accustomed to) –then learning that scientists have been capable of cloning humans since 1938 should not come as much surprise. Again, everything regarding my disclosure will become apparent, and I thank these men for disclosing their information, because it helps the reader to corroborate my disclosure.

Aaron Russo

Aaron Russo (pictured) was an American businessman, director, and political activist. He is best known for directing blockbuster films such as *Trading Places*, *The Rose* and *Wise Guys*. Russo believed that a human being "Should stand for something, and do the right thing when the time calls to act". Consequently, during the latter part of his career he did his best to warn the American public by producing documentaries such as *Mad as Hell* (1996), *Freedom to Fascism* (Russo 2006), and *Reflections and Warnings* (Jones 2008).

Russo disclosed the following key information in *Freedom to Fascism* (Russo 2006):

Income Tax

- There is no law requiring Americans to pay income tax. **Although this is a truth, I DO NOT advocate Americans NOT to pay income tax (and neither did Russo).** This is simply because the Federal Reserve can imprison you and seize your possessions for not paying income tax (although there is no law requiring Americans to do so). So for the sake of avoiding hassle for the mean time –pay your income taxes.
- To further illustrate the above point, Bob Shultz speaking at "We the People Foundation" said the following: "Most people believe that the income tax system is legal and that the revenue from the tax is used in the public interest. However, there is a substantial conclusive body of evidence that proves that our income tax system represents the most pernicious form of tyranny. It is the greatest hoax ever perpetrated by government against the working men and women of America."
- Charlie Beall: "The federal government itself refuses to provide, the American people, who are coercively being subjected to this extraction of their private property, without any underlying legal justification. There is no law. There is no law that requires the average American worker in the private sector to pay a direct un-apportioned tax on their labour and compensation for services. There is no law."

- State Representative – Phil Hart –(R-Idaho): "You can look through the statutes, and look for the law that requires you to pay, and when you do that, you **cannot** identify a law that requires the average person in America who earns a wage and works in private business to pay an income tax."
- Peter Gibbons (Tax Attorney) –It's actually very simple. Congress tried to enact an income tax in 1894 –The Supreme Court said that is unconstitutional. When The Supreme Court says something is unconstitutional, it's unconstitutional. They (Congress) tried again in 1913 –and The Supreme Court said –the 16th amendment- "No new power of taxation" –so if they (Congress) didn't have it then (1913) and they didn't get it; they **DON'T** have it. There is no constitutional base for a tax on the wages for Americans living and working in the 50 States of the union. Period. End of argument.

Radio Frequency Identification (RFID) Microchips

- The latest technology for identifying people at the point of self, when they make purchases –is actually the implantable (RFID) microchip. There are microchips that can actually be imbedded directly into human flesh... It's a tiny glass capsule about the size of a grain of rice... it contains an RFID computer microchip with a coiled antenna and it can transmit information also at a distance.
- Katherine Albrecht –Author of "*Spychips*" (2005) makes the following comment: "RFID is a technology that uses tiny computer technology the size of a grain of sand or smaller; hooked up to miniature antennas to transmit information about items at a distance. Back in 1999, Procter and Gamble, Gillette, and MIT got together to find a way to commercialise this technology and make it small enough, make it efficient enough and make it low cost enough to essentially ---their dream is to put these tiny computer chips on every physical item manufactured on planet earth."
- Radio waves can travel through walls, they can travel through wood, and they can travel through things we normally rely on to protect our privacy, for example your purse, your bag-pack, your pocket, –anything you are wearing or carrying.
- One of the most worrisome applications of RFID microchips, are proposals to put them into cash, meaning that it would be able to track every bank note where it had been, who it had been issued to, and create in essence an audit trail. That would essentially take away the anonymity of cash we now enjoy today.
- Once everything you do is tied down to a single number and there is no longer the ability to pay with cash, then all it takes to render you a non citizen is to simply turn that microchip off. **You will no longer be able to participate in any function in society including buying food. Once money becomes digitised through RFID technology, the elite can deduct whatever amount of money they want out of your microchip, whenever they want. They can trace you whenever they want. You will be at their mercy.**

Habeas Corpus

- Habeas Corpus is "[A] writ [formal document] requiring a person under arrest to be brought before a judge or into court, especially to secure the person's release unless lawful grounds are shown for their detention" (Legal-dictionary 2015).
- During the Bush Presidency [2001-2009 (History 2015)], President Bush signed executive orders giving him sole authority to impose martial law, and suspend Habeas Corpus. This gives him doctorial power over the people without any 'checks and balances.'

  In other words, "The government can jail you for life without charges, without a trial, without a lawyer" (Russo 2006).
- Furthermore: "The National Defense Authorization Act [4] signed by President Obama on the 31st December 2011 authorises the indefinite detention, without trial or indictment, of any US citizens designated as enemies by the executive." See Paye (2013) for further discussion regarding the suspension of Habeas Corpus.

Fraudulent Manipulation of Election Results

- In 2004 at the "Forum of Presidential Election" Clinton Eugene Curtis, a former Computer Programmer for NASA and ExxonMobil testified under oath that election results can be rigged using electronic programs.
- In 2004 he gave the following abridged testimony (Russo 2006):

  Judge: Mr Curtis, are there programs which can be used to secretly fix elections?

  Curtis: Yes.

  Judge: How do you know that to be the case?

  Curtis: Because in October of 2000 I wrote a prototype for present congressman Tom Feeney, and the company I worked for in Oviedo Florida –it did just that.

  Judge: And when you said, "it did just that" –it would rig an election?

  Curtis: It would flip the vote 51:49 –whoever you wanted it to go to, and whichever race you wanted to win.

  Judge: And would that program that you designed be something that election officials that might be on county boards would actually, could detect?

  Curtis: They'd never see it.

  Judge: So how would such a program, a secret program that, fixes the election –how could it be detected?

  Curtis: You would have to view it either with source code, or you would have to have a receipt and then count the hard paper against the actual vote total –other than that you won't see it.
- **Judge:** Given the availability of such vote rigging software and the testimony that has been given under oath of substantial statistical anomalies and gross differences – between exit polling data and the actual tabulated results, do you have an opinion whether or not Ohio elections, the Ohio Presidential elections was hacked?

  Curtis: Yes. I would say it was.

Person in the audience: ---So in other words there is absolutely no assurance whatsoever in anything in regard to these machines?
Curtis: Absolutely none.

- See Truthstream (2006) for further details regarding the full account of Curtis' testimony on how elections can be manipulated.
- Moreover, voting machine manufactures refuse to allow anyone to see the source code. Without paper ballots, the honesty of any election cannot be verified.

The War on Terrorism is the War on Your Freedom

- It is time to wake up America. These ID cards are not about defeating terrorism, but they are **ALL** about controlling the American people.
- The (mainstream) media controls the information that a person gets in various ways. They can make sure that the average American watching T.V. or reading the newspaper is going to come out with a certain mindset. She is going to say this is good, that's bad –and that is all they (media/elites) have to do.

Whoever Makes the Money Makes the Rules

- As Mayer Rothschild said "Give me control of a nation's money supply, and I care not who makes its laws". Mayer Rothschild, private banker, –knew that he and the other bankers would now control the laws of the nation. Government gave these bankers one of its most important powers, and now had to borrow money from the bankers and pay interest to finance the government.
- America has gone from people owning their own property, owning their own businesses –to a nation in debt because all the money is created by borrowing (from private banks) and this country has become one where people just live by borrowing.

Russo's Message to Mankind

- If you are in the military or law enforcement, remember you swore an oath [to uphold the law of the land]. You did not swear an oath to promote world government, or corporations.
- Now that you do understand what happened [to America, and the monetary system based on debt slavery] and how it is leading to a tyrannical one world government – the future of mankind depends on you –will you choose freedom or slavery?
- Stop being passive liberals. Stop being passive conservatives. Stop being passive centralists. Stop being passive human beings! When the media starts telling you that the country will fall apart if this is done... Do not be fooled!!
- Remember these are sick, malevolent and twisted people we are dealing with, trying to save themselves. Squash their agendas and stay on course.
- "I believe the time for mankind is time to give all or perish. Grow up or die. Grow up and become adults. Act like adults. Take some responsibility. The world which you have perceived is childlike; –and now the curtain has been pulled back for all to see."

- Unless people get active [get fully informed, and learn how governments of the world are conspiring against their people] and say –I am going to help shut down the Federal Reserve System; I am going to shut down the powers that be –the whole human race is doomed otherwise.
- We are coming near the end game and things are starting to accelerate... –and people are looking at the world leaders to declare martial law [which in turn will rid Habeas corpus].

Let's Consolidate our Problems

- If we can all get focused on how to win the game instead of all these different objectives –such as people are fighting for a better environment; people are fighting to keep their guns/gun control; there are all these different issues which are going on around the world -and which are all important on their own –but if we can consolidate on that and focus –and take all those people –and say –hey let's shut down the Federal Reserve, then we'll deal with those issues. Let's go to the agenda, the objective, the root cause of man's problems, first, then we'll go back to the other stuff.
- That is the priority right now. You have to cut off the head of the beast. And the head of the beast is the Federal Reserve System and the people behind it. You see, and that is what will save the world, and if people understand that –and they **stop** being passive liberals, and **stop** being passive republicans, but rather become **active** human beings –that is what will save the world!

Aaron Russo: Reflections and Warnings

In this documentary Russo recalls his relationship with Nick Rockefeller. Russo shares critical information Rockefeller discussed with him.

The Falsehood of September 11th 2001

- Russo met Nick Rockefeller through a female attorney who telephoned Russo and said "One of the Rockefellers would like to meet you." Russo made a documentary called *Mad as Hell* (1996), and Rockefeller had watched the video and knew Russo was running for Governor of Nevada and wanted to meet him. Russo said "Sure I'd like to meet him", and the two met and talked. Rockefeller proved to be a very smart man and shared ideas with Russo and was the person to tell Russo 11 months before 9/11 happened "There was going to be an event..."
- Rockefeller never told Russo what the event was going to be; but, there was going to be an event –and out of that event, we [America] were going to invade Afghanistan, to run pipelines from the Caspian Sea; we were going to invade Iraq, to take over the oil fields and establish a base in the Middle East, and make it all part of the New World Order; and we would go after (Hugo) Chavez and Venezuela –and sure enough, later when 9/11 happened –and I remember he was telling me how you are going to see soldiers looking in caves for people in Afghanistan and Pakistan and all these places and there is going to be this war on terror which has no real enemy and

the whole thing is a giant HOAX, but it is a way for the government to take over the American people (and the world).

- There is no question about it [that 9/11 is a hoax]. Nick Rockefeller said to Russo: “There is going to be a “War on Terror”” and he was laughing. Who are we fighting against...? Why do you think 9/11 happened and then nothing has happened since then? Do you think that our security is so great here that these people who pulled off 9/11, who were able to lock down another plane...? Come on it is ridiculous... 9/11 was done by people in our own government and our own banking system to perpetuate the fear of the American people into subordinating themselves to anything the government wants them to do. That is what it is about and to create this endless war on terror.
- Nick Rockefeller was laughing when he said, “We are going to be sending men into caves in Afghanistan and Pakistan” –and it was just cynical, he kept laughing and saying “Look how stupid everyone is! We can do whatever we want!”
- 9/11 was the first lie; and the next lie was to go into Iraq, to get Saddam Hussein out with his weapons of mass destruction (when the real issue was control of the oil fields) –that was the next lie. 9/11 created an endless war on terror that would go on and on and you can never define a real winner. There is no one to defeat and so it goes on and on forever. And they can do whatever they want; because they scared the hell out of the American public.
- This whole war on terror is [perpetuated on] a fraud. It is a farce. It is very difficult to say it out loud because people are intimidated in saying it. Because if you say it they want to make you into a nutcase –but the truth has to be... and the truth has to come out. The fact of the matter happens to be the whole war on terror is a fraud, it is a farce. Yes. There is a war going on in Iraq, because we invaded Iraq, and people over there are fighting... but the ‘war on terror’ –it is a JOKE –you know, and until we discover what really happened in 9/11 and who was responsible for 9/11 –because that is where the war on terror emanates from. That is where it comes from. It was 9/11 which allowed this war on terror to begin and until we get to the bottom root of 9/11 –the truth of 9/11 we’ll never know about the war on terror.
- Russo was in Tahiti when 9/11 happened and he got a call from his son – and his son said –“The twin towers, they were just attacked and they are falling down or something...” Russo was in Tahiti and had just woken up from sleep. Russo didn’t realise what it was immediately (11 months after Rockefeller had told him about 9/11 and it had happened –because Rockefeller said there was going to be an event –he wasn’t specific) –but after Russo saw that America was going to go to Iraq and Afghanistan, that is when he realised, and equated it to what Rockefeller had said.
- 9/11 was only a manifestation to create a fear in the American public. So that we would obey and do what they want us to do. Take for example, Richard Reed ‘The shoe bomber’ –now here is a guy who is 6 feet 6, ugly as can be; I heard he smelled... He sits on a plane, lights a match in a non-smoking area, to put his shoe on fire... surrounded by people... That is idiotic! If you were going to blow up a plane... you would go into the bathroom... you close the door... and you put your shoe on fire...

you are not going to sit there.... surrounded by people, lighting matches in a no smoking flight... they (elite) want you to believe this nonsense. That is ridiculous!

- The war on terrorism is to keep people in fear. It is an endless war without a real enemy (‘terrorists’) –so that people would submit and do whatever the government wants them to. Submit to searches, give you ID cards, put Radio Frequency Identification (RFID) chips in you etc. You become servants to the elite –that is what this is all about...
- Freedom of liberty is what people really want, and it is time to stop the duplicity of the government from lying to us. You see many people know the truth of what is happening in this country; like 9/11 but they are afraid to stand up. People have to stand up and find their courage and say “I’m not going to take this anymore, I know the truth” –and they create a situation where if you tell the truth, you are considered a lunatic.
- In other words if someone goes on a T.V. show and says that 9/11 was an inside job – immediately the person is labelled an idiot or crazy. They call you names. **You cannot be afraid of that.**
- If you do not fight the corruption and you do not stand up for what is right in life, you end up being a serf and a slave and you are leaving your children a world in which you would not want to live in yourself, so how can you in decency behave that way? You have to stand up for what is right in life, and unless you do that you are nothing.

The New World Order Agenda

- The whole agenda is to create a one world government where everyone has an RFID chip implanted in them. All the money is to be in those chips [a cashless world]. This information came straight from Nick Rockefeller himself. That is what the ultimate plans of the global elite, banking industry and Rockefeller wants to accomplish.
- The agenda is to implant everyone with RFID microchips. All money is transferred to those microchips. There is no more cash. Money would be in microchips. Instead of having cash, you would have money in your microchips, but whenever they want, they (elite) could take whatever amount out of your microchip whenever they want to. Total control. If you are a protestor they just turn off your chip; you cannot buy food; you cannot do anything; it is total control of the people.
- So they want a one world government controlled by them. Everyone being chipped, all the money in those chips, and they control the chips and they control people, and you become a slave. You become a serf to these people –that is their goal. That is their goal; that is their intentions.
- Russo did not believe in enslaving people and Rockefeller would question him in the following manner: “Why do you care about them? Why do you care about those people? What difference does it make to you [Aaron]? Take care of your own life.
  Do the best you can for you and your family. What do the rest of the people mean to you? They don’t mean anything to you. They are just serfs. They are just people...”

- Rockefeller asked –"Why are you fighting for the people for, what is it all about? The people have to be ruled. The constitution, what you are standing for is only for a few people, it's only for a few individuals who can live that way and we believe that it is best for society to be ruled by an elite people who control everything." Russo told Rockefeller he does not believe that. Russo believes: "God put me on this earth to be best person I could be and put everyone on this earth to be the best they can be, and NOT to be a slave and a sheep to YOU and these people (elite) –and I do not understand why you want to control everything. What is the need for that?"
- It was just a lack of caring [from Rockefeller's part], and that is just not who Russo was. It was just sort of like cold you know, and Russo used to say to Rockefeller, "What is the point, of all this? You have all the money in the world you need; you have all the power you need, what is the point? What is the end goal? Rockefeller said "The end goal is to get everyone chipped. To control the whole society. To have the elite people [the bankers and government] controlling the world."

Women's Liberation from the Perspective of the Elite

- Women's' Liberation was founded by the Rockefellers.
- Rockefeller asked Russo: "What do you think women's liberation was about?" At the time Russo had a pretty conventional thinking about it and he said "It's about women having the right to work; get equal pay with men; just like they won the right to vote." At this point Rockefeller started to laugh and he said to Russo "You're an idiot"
- Rockefeller said –"Let me tell you what that was about. We the Rockefellers funded women's' liberation. We are the ones who got it all over the newspapers and television. The Rockefeller foundation –and you want to know why? There were two primary reasons:"
- 1) "We couldn't tax half the population before Women's Liberation and the second reason was"
- 2) "Now we get the kids at an early age [because both parents are away from home working] –we can indoctrinate the kids how to think, so it breaks up their family –the kids start looking at the state as the family, as the school as the officials as their family, not as their parents teaching them, and so those are the two primary reasons for women's liberation."
- Russo thought, up to that point, Women's Liberation was a noble thing; however when he saw the Rockefellers' intentions behind it, where they were coming from when they created Women's Liberation; the thought process of it; Russo saw the evil behind what he thought was a noble venture.

America is a Republic

- America is a constitution republic / it is supposed to be a constitution republic –and **NOT** a democracy. The majority should not rule, and nor should the majority take over the inalienable rights of the minority.

- Americans pledge allegiance to the "Flag and the **Republic** which it stands" –and **NOT** democracy.
- Democracy is the worst form of government you can have because it is majority rule. Therefore the government can tell you what to do because 'the majority' wants it. It is irrelevant what the majority wants. Decisions should not take away the inalienable rights of an individual.
- Russo also said "It doesn't matter who you vote for, republican or democrat; –they are the same –neither one of them is stopping the Federal Reserve or paying income taxes.
- The elite (Federal Reserve) have taken over the American government; there is no difference between republicans and democrats. There is no difference between the two parties. The duality is manufactured. They (Federal Reserve; central banks) control both parties. It doesn't matter to the elite which one wins, because whoever is running for President will be someone they anoint. Whoever runs for President, will do whatever the elite want them to do. The fact of the matter happens to be that you cannot win an election unless you have enough money to win; they (Federal Reserve; central banks) make sure who gets the money.

Depopulation

- Russo and Rockefeller discussed many things –and one of the things Rockefeller brought up in conversation was reducing the world population. Rockefeller felt that there are too many people in the world. In a way Russo agreed that there are too many people in the world, but he **does not** think he has the authority to say who dies and who lives; but the elite felt that they want to reduce world population and Rockefeller felt it should be reduced by half.
- Rockefeller even mentioned to Russo in conversation, that they were having a real problem trying to solve the Israel/Palestine problem –and they were playing with the idea of bringing Israel to Arizona –and taking everybody from Israel and giving everybody a million dollars, and setting up Israel in the state of Arizona –because that is a problem that they are not in charge of.

Borrowing Money from Private Banks causes Inflation and Debt

- These people (elite) control the money so they make all the rules, and therefore they put the rules in which they want into effect, and the truth is America has really become a socialist communistic country. Everyone says it is a capitalistic country, but how can it be a capitalistic country when you have a central bank? That is the first question people should ask. It can't be. It is a planned economy, it is a phony!
- If they want to create prosperity, they just print dollars, or put digits into the economy. Now you have prosperity. You do not have real prosperity, you do not have real manufacturing, you just have money being injected in which is infusion of credit. This makes the government go into more debt.

- Whoever makes the money makes the rules. Rothschild said that. Why are we allowing these private bankers to make the money for our country? It is nonsense. Why are we paying interest to these banks to make money for us when the government can do it itself without paying interest, without all that debt? There is no answer to that question and it is a question no politician will raise. Everybody talks about America’s debt... We are in debt because we borrow money... but we don’t HAVE to borrow money. They designed it so that we can go into debt. We can create the money, and back it by gold so that they cannot create too many of it, so that you do not have the inflation, and do what the Founding Fathers told us.
- Why in the world does the American government borrow money from the banks when they have the ability to create it themselves without borrowing it, and paying interest on it? Why? Nobody can answer that question –not one politician ever raises that. Why does the American government ever borrow money, when they can create it without paying interest?

  –And people say that –well if the American government creates it, it will cause inflation. And that is their answer. And Russo says well let’s look at it: the American government has the Federal Reserve do it, which creates the same inflation as if they did it, but also with the inflation –now you are getting massive debt –so with the Federal Reserve you have inflation and debt. Now if the American government made the money, backed by gold which would limit the amount they could make –you wouldn’t have debt and you wouldn’t have inflation.
- It wasn’t until 1913 when the Federal Reserve came in, that America had inflation. Before then there was no inflation for 100 years. There were points and spikes, mostly during the Civil War –but basically there was no inflation other than during that short period of time. I mean a loaf of bread would cost the same thing. People could plan their lives.
- Today, they have planned inflation, and now you have two parents working, they cannot afford to take and pay for their family anymore; the kids are going to state run schools now, the kids are being indoctrinated how to think; they are being given Ritalin, they are being given all these drugs, the whole country is being dumb down, it is all because of the Federal Reserve System; and the Federal Reserve system and these bankers are responsible for the demise of America. And if we ever want to win this battle you must shut down the Federal Reserve System, and we must shut down these bankers and restore sound money to this country.
- If you analyse the situation and if you realise that since the Federal Reserve has come into being since 1913, illegally, without a constitutional amendment, by bribing a few senators during Christmas vacation, they turned over the most important power that the American government has, the creation and issuance of money to a private bank.
- Through that private bank issuing money they have destroyed this country. They have destroyed the purchasing power of money in this country; they have created social programs that are destroying this country.
- The Federal Reserve has created massive inflation in America which means the American worker has to keep on making more money to keep up with the cost of

living. The more money they make to keep up with the cost of living, the less competitive they become in the world economy. So now what happens is that we have to pay our workers so much to keep up with the cost of living; and then they (elite; government; corporations etc.) say screw the American worker; let's go overseas now to get the cheap labour.

- The inflation the Federal Reserve has created has now allowed other countries to outcompete us. Other countries do not have to pay as much money as we have to pay to our workers to survive. So now we are not competitive anymore and we have lost our manufacturing base. We have lost our competitive edge.
- "Freedom to Fascism" is a documentary that everybody should see. Russo and his team show the fraud of the income tax; they show how Judges put people into jail for no reason; they show the corruption of the justice system. They show how the Federal Reserve came into being and how it is controlling society and how all the central banks are working together through the bank of International Settlements, in Switzerland which is the central bank for all central banks and how all are working together to create this one world government; this one world order; which is what they are trying to do.

The Deception of the Council of Foreign Relations (CFR)

- Russo was interested in joining the Council of Foreign Relations, but he found out from Rockefeller himself, that part of the end goal of the CFR is to get everyone RFID chipped. To control the whole society and have the elite people (bankers, government etc.) controlling the world. Russo asked Rockefeller "Do all the people in the CFR believe the way that you do?" Rockefeller said "No, no, no. Most of them believe they are doing the right thing. A lot of them believe it is better off being socialistic. We have to convince people that socialism is really capitalism." Because America is becoming a socialistic country; it is a communist country today.
- Russo's friendship with Rockefeller became one where they would share thoughts, ideas and philosophies and Rockefeller wanted Russo to become a part of what they were doing [enslavement agenda], and for Russo to become a member of the CFR; Rockefeller offered various business opportunities for Russo to get involved in the CFR and to **not** take up the fight or the battle that Russo had been taking up in the past. Rockefeller wanted Russo to drop the idea of helping the people; because "What was the point in Russo fighting for the people?" Rockefeller would question.
- Russo asked Rockefeller do all the people of the Council of Foreign Relations feel the same way you feel? Rockefeller said "A lot of them think they are doing the right thing, they think that socialism is the best way to go (but this form of socialism involves redistributing the wealth for the elites and not to everyone), they think that they are doing the right thing." But the people at the (very) top they all know the truth of what is happening. (Therefore, the good people in the CFR are also under an illusion, and do not know that effectively they are working for evil men who have the overall goal to control the populace of the world).

- So it is compartmentalised within the elite structure as well. All the people in the CFR -2000 to 3000 people like Dan Raddler –they don't know what is going on--- they join the CFR because it is prestigious. They think it is good for business, it is good for this; they don't know what is really happening –the evil that comes out of it –that is emanating out of it.
- In terms of the CFR, in terms of compartmentalisation, there are many good people which Russo believes are part of these organisations who do not even understand what these organisations are really about. For example, when Russo was in Germany, doing cancer treatment, there was a gentleman there who was visiting a friend of his with cancer. The gentleman visiting his friend was a member of the CFR, and him and Russo were talking and Russo showed him the movie (*Freedom to Fascism* (2006)); --and he said, "Oh my god! I'm going to resign." "I had no idea this is what the CFR is about". He had no idea; he is just a nice guy, who thought he was joining a prestigious organisation.
- A lot of people join the CFR because they think it is a prestigious organisation; it will help them in business; make good business contacts etc. They do not have an understanding that the CFR is really about world domination. How they, and the Trilateral Commission, The Bilderberger [Group], the banks, all work together to control the people –a lot of them do not understand that. They do not see the big picture. They think: Oh the CFR is a prestigious organisation. I'll make this, I'll make that, and I can do business deals. It is just business to them. The CFR wants to get the people in there that have influence and power, and so they are part of that (enslavement agenda), and so they are not opposed to them. So the whole country is becoming the haves and the have not. You are getting the very, very wealthy and the middle class being destroyed and you are getting the poor people.
- In Russo's words: "You can call the CFR what you like, but it is a criminal organisation. Run by criminals. But people do not think of it as a criminal organisation, because it has 'class'; 'style'; 'prestige'; –and it is 'respected'; so people do not look at it as being a criminal organisation. That is what a great job they have done."

Combining America, Mexico and Canada into one Country

- Imagine this... here you are in America, and they (elite) are combining American, Canada and Mexico into one country. The North American union. And the American people do not know anything about it. It is not even in the press. They would rather talk about Rosy O'Donnell and Donald Trump calling each other names than discussing the fact that we are merging into one country. This isn't even reported.
- The fact of the matter happens to be, that tells you how controlled the media is. The elite control the media, and they control governments and they are all in bed together. Here you are combining American, Canada and Mexico into one country, and you do not see it in the press. You do not see it in the press. Why? This should be one of the top stories everywhere –and the elite are not worried about it. That tells you –there is the evidence that [media] it is controlled. They do not want the American people to

know what is going on, that is why they do not protect our borders. That is why we are losing our constitution, the very document that secures our freedoms.

Russo's Vision on How to Bring These People Down

- There is no question we are in tyranny, there is no question that the American citizen is no longer a free individual human being, to do the things that they wish to do. We're slaves and it is getting worse.
- We are dealing with complete evil; and until the American people wake up and say, we do not want this evil in our country anymore and we want to come back to a country of decency and goodness, integrity and honour, we are going down that road and that is what it is going to take? It is going to take people to stand up and say we do not want to live in this kind of world anymore. I believe we should pull all of our troops out of Iraq, I believe we should leave other countries alone. Let other countries live their lives the way they choose to. Stop trying to spread 'democracy' around the world, which is the worst form of government there is anyway (because 51 percent rule over 49%), restore our republic to what it is supposed to be and go back to what the founding fathers gave us. Restore the republic.
- In Russo's opinion: The populace must shut down the Federal Reserve System –and there has to be an uprising. There has to be an uprising. People have to stand up. One person cannot do it alone. You cannot do it alone. People do not seem to have the courage to do what they have to do.
- A lot of people in Hollywood know the truth, they do not want to stand up and speak about it; I know many of them have seen my movie (*Freedom to Fascism* (2006)) and they know I am right, and they want to talk about it because everybody is afraid. Everybody is afraid because they think that the money they get from the Federal Reserve is really money and they have a comfortable lifestyle and they are afraid of change. They are afraid to stand up for what is right, and until people are willing to stand up and have the courage to do what they need to do, it is not going to change; and hopefully we can affect change when people stand up and say "Hey, I've had enough".
- We (the populace) have one advantage. They (elite) need us to cooperate. See, if we do not cooperate with them, they cannot win. They always need our cooperation with them to go along with their programs. They try to "sell us". Democracy; this majority says this; believe in this; do this, do that; the war on terror; we have to be scared... They are always trying to do things to "sell to us" to go along with them, and once we learn not to cooperate with them; then we win the game.
- That is the point, do not cooperate with them, and do not go along with the program anymore. Stop it. Join forces, and bring freedom back to this country. It is going to take people who believe in freedom; The Constitution and the Founding Fathers, Thomas Jefferson, to make this country whole again, because right now it is in the grip of the evil ones, and the only way to stop that is for good men to stand up.

- We have to stop being scared. We have to do what is necessary to take back what is ours. We have to stop these bankers, these elite, full of liars, congressmen full of liars. They are destroying our borders.
- So through these bankers, attempting to take over America, knowing that America was the freest nation on this earth, it was necessary for them to takeover America, take away gun rights, freedom to bear arms, and create a country where we become slaves, because once they take over America, the rest of the world becomes a lot easier for them. And so by creating 9/11, an event to terrify the American people that we are being 'attacked by terrorist', you create a world where there is an enemy that can **never** be pinpointed. You can never win the battle. It is 100 year war – a never ending war on terrorism. So you are always fighting this war, and through the war on terrorism, through 9/11 which is the first lie, then you create the war on terrorism which is the next lie, then you create the war in Iraq through weapons of mass destruction, which is the next lie –so you get one lie, to the next lie, to the next lie, --now it is going to be Iran the next lie –and sending more troops and insurgents into Iraq.
- Restore America's Republic back to what it is supposed to be. Get the bankers out of our government. Get government to stop borrowing money from the banks. Government should make its own money; restore the Republic. Restore individual freedoms. That is what this country is about -and until we do that we are going to be slaves.
- You have to take away the creation of money away from the private bankers and you will solve 95% of your problems.
- Americans, mobilise, stand tall, stand together, tell the government you are "Mad as hell!" Do not cooperate with the government do not accept a National ID card. Do everything in your power to restore freedom and your individuality back to America. Stop being a country run by the institutions for the institutions. Let's go back to "We the people, by the people for the people", as opposed to, we the institutions, by the institution, for the institution. Stand up for your individual rights. Stand up for the God leaders that are in each and every one of us!

The Trans-humanist / Post Human Agenda

Transhumanism is a cultural and intellectual movement that believes we can, and should, improve the human condition through the use of advanced technologies. One of the core concepts in transhumanist thinking is life extension: through genetic engineering, nanotech, cloning, and other emerging technologies, eternal life may soon be possible. Likewise, transhumanists are interested in the ever-increasing number of technologies that can boost our physical, intellectual, and psychological capabilities beyond what humans are naturally capable of (thus the term *trans*human) (Anthony 2013). Transcranial direct current stimulation (tDCS), for example, which speeds up reaction times and learning speed by running a very weak electric current through your brain, (Anthony 2012) has already been used by the US military to train snipers. On the more extreme side, transhumanism deals with the concepts of mind uploading (to a computer), and what happens when we finally craft a computer with greater-than-human intelligence (the technological singularity) (See: “How to create a mind, or die trying”, Hewitt 2012) (Anthony 2013).

Moreover, put simply, “posthumanism” can be defined as that condition in which humans and intelligent technology are becoming increasingly intertwined (TheNanoAge 2015). For readers interested in learning more about the posthuman / transhuman agenda see BT Soul Catcher 2025 (BEAMS 2007); Avatar Project 2045 (2045 Initiative 2015; Borghino 2012) and Mind Clone Robot (Bloomberg Business 2015; RT 2015) which have all been disclosed as methods of transferring the human consciousness to a computer.

The Phases of Sleep and Rapid Eye Movement (R.E.M) Sleep

Sleepers pass through five stages of sleep: 1, 2, 3, 4, and REM (rapid eye movement) sleep. These stages progress cyclically from stage 1 through REM then begin again with stage 1. A complete sleep cycle takes an average 90 to 110 minutes (Sleepdex 2015). In other words, after falling asleep, it takes approximately 90-110 minutes to enter REM sleep.

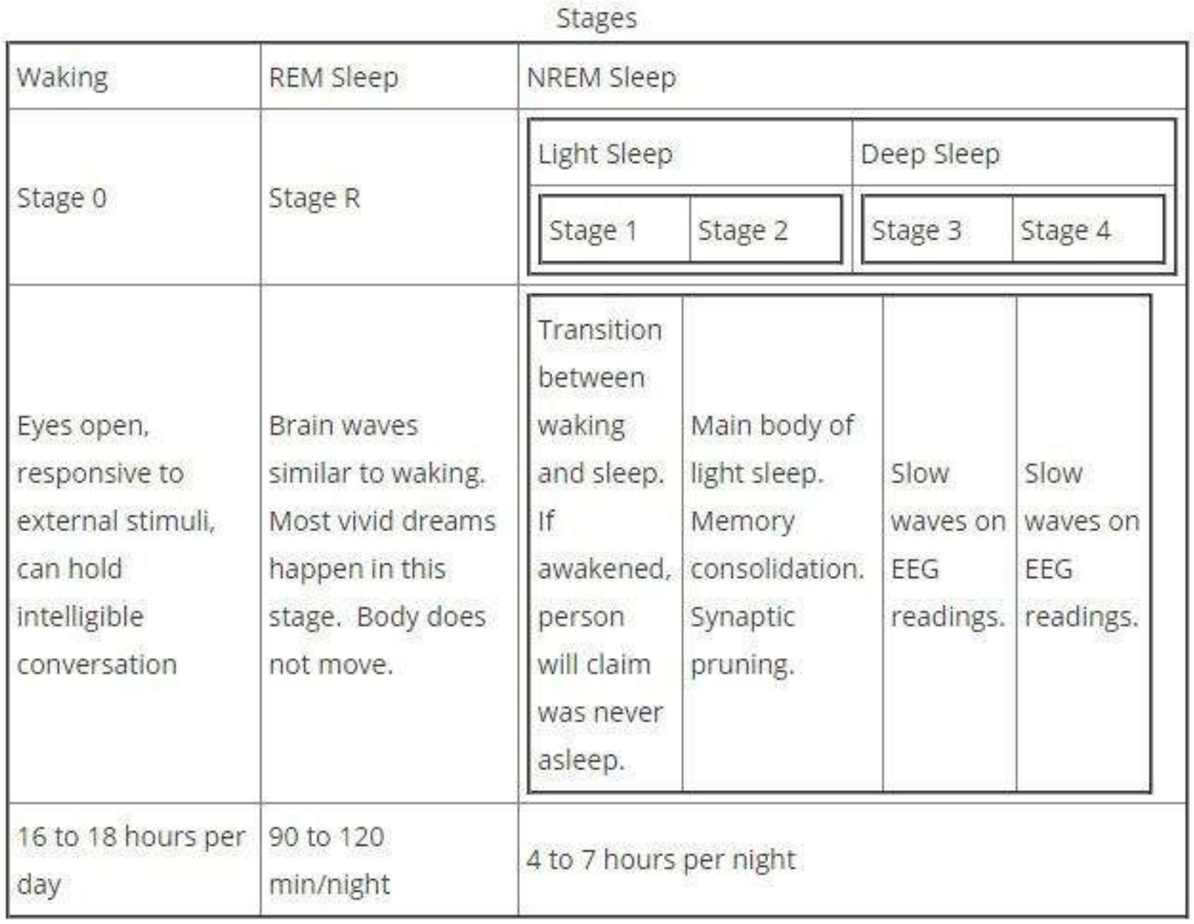

Stages

Waking	REM Sleep	NREM Sleep			
Stage 0	Stage R	Light Sleep		Deep Sleep	
		Stage 1	Stage 2	Stage 3	Stage 4
Eyes open, responsive to external stimuli, can hold intelligible conversation	Brain waves similar to waking. Most vivid dreams happen in this stage. Body does not move.	Transition between waking and sleep. If awakened, person will claim was never asleep.	Main body of light sleep. Memory consolidation. Synaptic pruning.	Slow waves on EEG readings.	Slow waves on EEG readings.
16 to 18 hours per day	90 to 120 min/night	4 to 7 hours per night			

Image 1: Stages of sleep. Source: Sleepdex: (2015)

Any sufficiently advanced technology is indistinguishable from magic.
Arthur C. Clarke

REM Sleep

Most dreaming occurs during Stage Five, known as REM. REM sleep is characterized by eye movement, increased respiration rate, and increased brain activity. REM sleep is also referred to as paradoxical sleep because, while the brain and other body systems become more active, your muscles become more relaxed, or paralyzed. Dreaming occurs because of increased brain activity, but voluntary muscles become paralyzed. Voluntary muscles are those that you need to move by choice, for example, your arms and legs. Involuntary muscles are those that include your heart and gut. They move on their own (Sleepdex 2015; Walcutt 2013).

Rapid eye movement, or REM sleep, is when you typically dream. You may have images float by in earlier stages, particularly when you are going through Alpha or Theta (brain waves), but the actual dream state occurs in REM (Walcutt 2013).

This period of paralyzation is a built-in protective measure to keep you from harming yourself. When you are paralyzed, you can’t leap out of bed and run. Do you ever feel like you can’t escape during a dream? Well, the truth is, you can’t. You can breathe, and your heart is working, but you really can’t move (Walcutt 2013).

The reader is also advised to see the Horizon documentary “Why Do We Dream?” (BBC Horizon 2009). This video is available, on YouTube, and details sleep, the phases of sleep, REM sleep, the period of paralysation and more.

Current Advances in Technology

This is a section some readers may have difficulty with, because reading about these current technologies alone, one cannot help but feel that such technologies described below sound like the stuff of science fiction and fantasy -although they have been corroborated through public drip-feed disclosure (explained below). Nevertheless, I urge the reader to keep in mind Arthur C. Clarke’s quote (above), as well as Phil Schneider’s testimony regarding the advancements of military technology in comparison to the general public’s technology (for every 12 months, military technology outstrips the technology the public is accustomed to by a rate of 44 years). Moreover, for the reader who finds difficulty understanding the technologies described below, seek the supporting articles / videos.

Memory Suppression Technologies

Memory suppression technologies are any scientifically advanced technologies which are used to suppress memory. Examples of how memories can be suppressed can be found by reading Winter's (2014) article which details how memories can be suppressed using light; and Greenberg's (2013) article which discusses memory suppression through gene / chemical modification.

Mind-Voice Technology

Mind-voice technology is an advanced technology which is capable of reading, listening, hearing or broadcasting your inner voice / thoughts. Examples of articles which discuss mind-voice technology are: Prigg (2014) details software which can read the inner voice; and New Scientist (2014) which also discusses a brain decoder which can eavesdrop on one's inner voice.

H.A.A.R.P. Technology

The High Frequency Active Auroral Research Program (HAARP) is a radio transmitting system that can bounce signals off the ionosphere (a region of the Earth's upper atmosphere 60km (37 miles) to 1000km (620 miles) altitude) and back to earth to probe deep into the earth or sea, its proponents say. The system could locate minerals or communicate with submarines. (Begich & Manning 1997; Sheen, Begich & Robbins 2005).

HAARP can also:

- Disrupt human mental processes.
- Knock out all global communications systems.
- Manipulate global weather.
- Change weather patterns over large areas.
- Interfere with wildlife migration patterns.
- Hurt ecosystems.
- Negatively affect your health, moods, and mental states.
- Unnaturally impact the Earth's upper atmosphere.

This illustration (below) shows the ionosphere's relationship to the Earth. The illustration appeared in the HAARP Environmental Impact Statement on page 10-125 of Volume II.

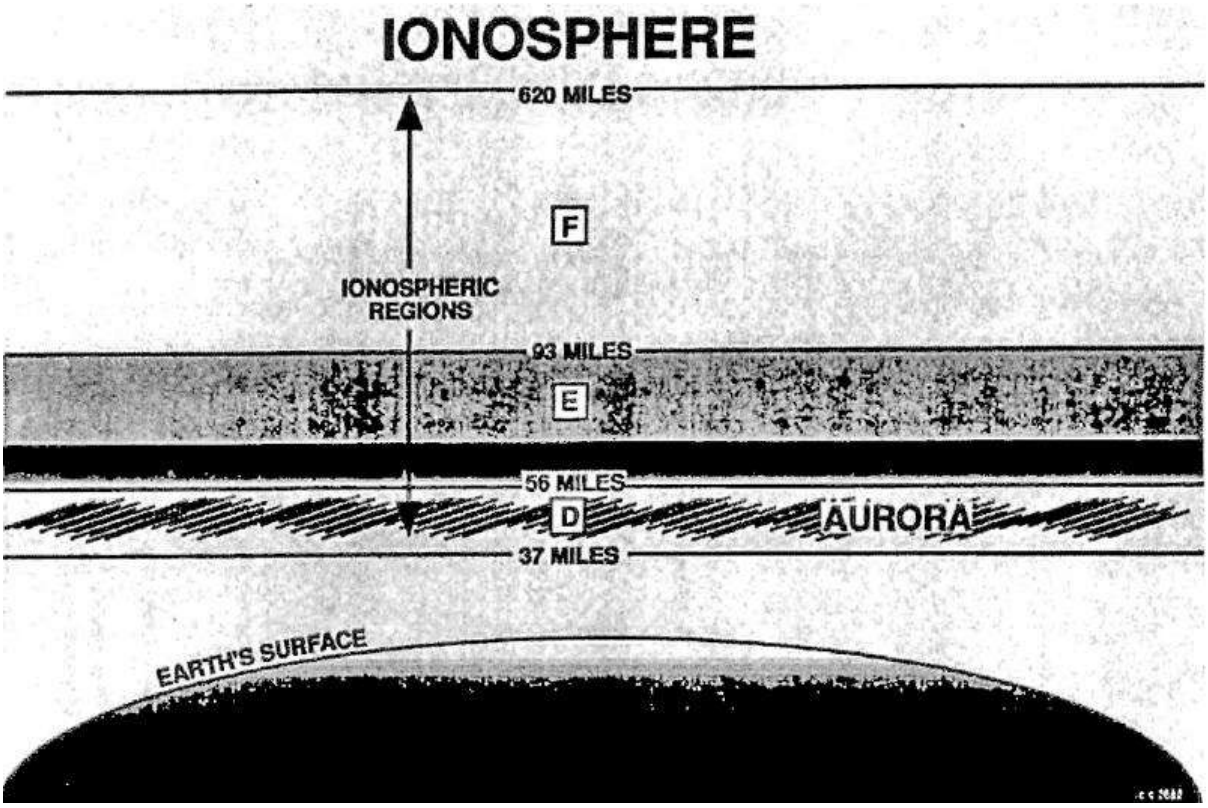

Image 2: The ionosphere's relationship to earth. Source: (Begich & Manning 1997)

The ionosphere protects the earth. HAARP (High frequency Active Auroral Research Program) is made to beam more than 1.7 gigawatts (billion watts) of radiated power into the ionosphere -the electrically charged layer above Earth‘s atmosphere. Put simply, the apparatus is a reversal of a radio telescope - just transmitting instead of receiving. It will boil the upper atmosphere. After disturbing the ionosphere, the radiations will bounce back onto the earth in the form of long waves which penetrate our bodies, the ground, and the oceans.

HAARP represents a technology which could lead to a new class of weapons that could change our world profoundly - an all-purpose military tool. If misused, the tool could mess up the weather. It could be used against humanity in a way that would change what people think, believe and feel. It could be used for good or evil, just as a harp can produce the music of Mozart or the melody of a death march.

H.A.A.R.P. and Weather Control

- “The theoretical implication [of Dr. Robert Helliwell and John Katsufrakis of Stanford University in 1974] suggested by their work is that global weather control can be attained by the injection of relatively small 'signals' into the Van Allen belts (radiation belts around Earth) - something like a super-transistor effect” said Frederic Jueneman.
- Yes. The weather can be controlled using HAARP technology.

- A series of weather disasters began in 1960, according to a CIA report mentioned in the editorial, but at the time climatologists couldn‘t look ahead and see that droughts, floods and abnormal temperatures would continue beyond that decade. As if natural disasters weren‘t bad enough, the CIA reported that national governments were already able to manipulate weather for military purposes [using HAARP technology].

As far back as 1958, the chief White House advisor on weather modification, Captain Howard T. Orville, said the U.S. Department of Defence (DoD) was studying ways to manipulate the charges of the earth and sky and so affect the weather by using an electronic beam to ionize or deionize the atmosphere over a given area. In 1966, Professor Gordon J. F. MacDonald, associate director of the Institute of Geophysics and Planetary Physics at the University of California, Los Angeles, was a member of the President‘s Science Advisory Committee, and later a member of the President's Council on Environmental Quality.

Gordon J. F. MacDonald published papers on the use of environmental control technologies for military purposes. MacDonald made a revealing comment: ―The key to geophysical warfare is the identification of environmental instabilities to which the addition of a small amount of energy would release vastly greater amounts of energy. MacDonald had a number of ideas for using the environment as a weapon system and he contributed to what was, at the time, the dream of a futurist. When he wrote his chapter, ―"How to Wreck the Environment" for the book "Unless Peace Comes" he was not kidding around.

In the text MacDonald describes the use of weather manipulation, climate modification, polar ice cap melting or destabilization, ozone depletion techniques, earthquake engineering, ocean wave control and brain wave manipulation utilizing the planet‘s energy fields. He also said that these types of weapons would be developed and, when used, would be virtually undetectable by their victims. He was not some wire haired fanatic when he made these observations in 1966 -he had the credentials of a world recognized scientist. What his futuristic concepts became, are the things which projects like HAARP are made of...

H.A.A.R.P. and Mind Control

- Radio frequency radiation, acting as a carrier for extremely low frequencies (ELF), can be used to wirelessly entrain (adjust) brain waves.
- We are talking about very, very low power requirements. The trick for influencing brain activity is in the combination of frequency, power level and wave form.
- As Dr. Patrick Flanagan, one of America's most gifted inventors noted in an interview, the HAARP project could be not only the biggest ionospheric heater in the world, but also the biggest brain-entrainment (brain adjustment) device ever conceived.
- According to HAARP records, when the device is built to full power it can send very low frequency (VLF) and extremely low frequency (ELF) waves using many wave forms at energy levels sufficient to affect the mental states of entire regional populations.

- The HAARP transmitting system could be used unintentionally or intentionally to alter mental functions.
- If HAARP is tuned to the right frequency, using just the right wave forms, mental disruption throughout a region could occur intentionally or as a side effect of the radio frequency transmissions [in other words: “mind control”].

RFID Technology

Radio Frequency Identification (RFID) microchips are microchips that can be directly imbedded into human flesh. This section discusses the dangers of RFID chips. See Rense (2001) for full review.

- RFID technology links the brains of people via implanted microchips to satellites controlled by ground-based super-computers.
- Today they are small enough to be inserted into the neck or back, and also intravenously (through a vein) in different parts of the body during surgical operations, with or without the consent of the subject. It is now almost impossible to detect or remove them.
- Implanted human beings can be followed anywhere.
- Today's microchips operate by means of low-frequency radio waves that target them. With the help of satellites, the implanted person can be tracked anywhere on the globe.

RFID Technology and the Medical Profession

- One reason the dangers of implantable microchip technology has remained a state secret is the widespread prestige of the psychiatric DIAGNOSTIC STATISTICAL MANUAL IV produced by the U.S. American Psychiatric Association (APA), and printed in 18 languages. Psychiatrists working for U.S. intelligence agencies no doubt participated in writing and revising this manual. This psychiatric "bible" covers up the secret development of Mind Control technologies by labelling some of their effects as symptoms of paranoid schizophrenia.
- The Psychiatric Diagnostic Statistical Manual (DSM) for mental disorders has been a brilliant cover up operation in 18 languages to hide the atrocities of military and intelligence agencies' actions towards their targets. THE MANUAL LISTS ALL MIND CONTROL ACTIONS AS SIGNS OF PARANOID SCHIZOPHRENIA.
- If a target is under surveillance with modern technology via TV, radio, telephone, loudspeakers, lasers, microwaves, poisoned with mind altering drugs via air-ducts, giving familiar smells which cause headache, nausea and so forth, if s/he claims her/his clothes are poisoned, her/his food or tap water as well --- all medical schools teach their students that the person is paranoid, ESPECIALLY if s/he believes intelligence agencies are behind it all.
- Never is the medical profession told that these are routine actions all over the world by intelligence agencies against their targets. Thus, victims of mind control are falsely considered mentally ill and get no help since they are not believed and their suffering is doubled by misinformed health professionals.

Further Implications of RFID Technology

- How many people realize what the implications of implantable chips actually mean? It means total loss of privacy and total outside control of the person's physical body functions, mental, emotional and thought processes, including the implanted person's subconscious and dreams! For the rest of his / her life!
- It sounds like science fiction but it is secret military and intelligence agencies' mind control technology, which has been experimented with for over half a century (since 1950s). Totally without the knowledge of the general public and even the general academic population.
- Supercomputers in Maryland, Israel and elsewhere with a speed of over 20 BILLION bits/sec can monitor millions of people simultaneously. In fact, the whole world population can be totally controlled by these secret brain-computer interactions, however unbelievable it sounds for the uninformed.
- Neuro-electromagnetic involuntary human experimentation has been going on with the so-called "vulnerable population" for over 50 years, in the name of "science" or "national security" contrary to all human rights. It happens today in the **USA, Japan, and Europe**. With few exceptions, the mass media suppresses all information about the entire topic.
- Only increased public awareness of the microchip implants, their frightful consequences to privacy by influencing of individuals' thoughts and actions, causing people to become biological robots with physical and emotional pain whenever the supercomputer technician so wishes, is enough reason to refuse to take the microchip into your body for whatever reason.
- It is the biggest threat to humanity and the most sinister plan to enslave the human race forever.

If you have a choice and want to remain a normal human being with privacy, **DO NOT** have your children implanted **NOR** yourself implanted with RFID microchip(s) (or any other type of implantable microchip). Otherwise your vision, hearing, sensing, thoughts, dreams and subconscious will be influenced by an outsider, who does not have your best interests in mind. For the rest of your life!

Mind Control: MK Ultra Technology

MK Ultra today has evolved from the 1950s variation of mind control (MK Ultra Compendium 1980), whereby drugs such as LSD, and interrogation were used as methods to weaken the mind of the individual to force confessions through mind control. Today, mind control is achieved through the implantable microchip (mentioned above; Rense 2001).

- The brain functions of an implanted person can then be remotely monitored by supercomputers and even altered through the changing of frequencies.
- Once implanted, the U.S. National Security Agency's (NSA) 20 billion bits / second supercomputers could now "see and hear" what you are experiencing with a remote monitoring system (RMS).

- Every thought, reaction, hearing and visual observation causes a certain neurological potential, spikes, and patterns in the brain and its electromagnetic fields, which can now be decoded into thoughts, pictures and voices. MK Ultra technology is therefore capable of Mental Video and Audio Projection, as well as Artificial Telepathy.
- The mass media have not reported that an implanted person's privacy vanishes for the rest of his or her life. S/he can be manipulated in many ways. Using different frequencies, the secret controller of MK Ultra technology can even change a person's emotional life. S/he can be made aggressive or lethargic. Sexuality can be artificially influenced. Thought signals and subconscious thinking can be read, dreams affected and even induced, all without the knowledge or consent of the implanted person by using MK ultra technology.
- Memory suppression technologies are used in conjunction with MK Ultra technology, which enables the programmer to control certain memories the victim remembers. The use of memory suppression technologies and MK Ultra technology allows the programmer to reinforce behaviour and elicit specific conditioned responses.
- Mind control techniques, such as MK Ultra, can be used for political purposes. The goal of mind controllers today is to induce the targeted persons or groups to act against his/her own convictions and best interests. Zombified individuals can even be programmed using MK Ultra technology, to murder and remember nothing of their crime afterward.
- The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct and to control the absolute behaviour and thought patterns of the individual. The purpose of mind control, using MK Ultra technology is to disrupt memory, discredit people through unusual behaviour, to make them insane or to commit suicide or murder.

See Mind-Computer (2012) which discusses how artificial telepathy is achieved. Artificial telepathy, also known as ‘brain to brain communication’, is also possible using MK Ultra technology. Furthermore, review: Jim Cristea (2009); Berkeley News (2011); UC Berkeley Campus Life (2011) CTForecaster (2013); nature video (2013) and Stromberg (2013). The above mentioned articles and videos describe, demonstrate, and corroborate how audio and video projection of the brain is achieved; how dreams can be recorded and projected digitally; how the brain can be scanned to reveal hidden information personal to an individual; and what a person’s underlying intentions are, by using brain scans. Moreover, the articles and videos mentioned above confirm the functionalities of MK Ultra technology.

Furthermore, when our brain functions are connected to supercomputers by means of RFID technology, MK Ultra technology and other implantable microchips, it will be too late for protest. This threat can be defeated only by educating the public, using available literature on biotelemetry (electronic equipment that receives signals from radio transmitters) and information exchanged at international congresses.

Human Cloning

There are currently five different types of clones, concealed from public knowledge. There are Mark 1 clones; Mark 2 Clones; Mark 3 Clones; and Mark 4 clones and reanimated clones. Moreover, there are two types of cloning techniques: duplication cloning and replication cloning.

Mark 1 clones are REM sleep driven clones. However, the technology used for Mark 1 clones causes many side effects and therefore Mark 1 REM sleep driven clones are no longer a preferred choice.

Mark 2 clones are also REM sleep driven clones. Mark 2 clones have fewer side effects than Mark 1 clones, and therefore Mark 2 clones are currently the preferred method for REM sleep driven clones.

Mark 3 clones are independent clones which operate on microchip containing the entire consciousness of an individual.

Mark 4 clones are also independent clones, operating on a microchip which contains the entire consciousness of an individual and is an advanced version of a Mark 3 clone.

These independent clones (Mark 3 and Mark 4) have a lifespan of 6 months to 12 months and require adjustments after this period to run efficiently once more. Without these adjustments, their functionality weakens.

To "reanimate" means "to restore to life; resuscitate; revive". Consequently, reanimated clones are clones which are genetically identical to that of a person who once lived.

Replication cloning is what the public is most familiar with. Replication cloning involves giving birth to a genetic identical of an original where the newborn starts life off as a baby and matures. The newborn is referred to as a clone.

However, duplication cloning is a current concealed advanced from of cloning, and it involves taking as little as two cells from an individual, adding a constant electrical charge to the cells until a complete human being is formed.

Duplication cloning is similar to the process of regenerative medicine as demonstrated by Dr. Stephen Badylak's video "How to grow a New Fingertip" (CBS 2008; Science Channel 2014) where Badylak states: 'A whole human can be grown within 9 months.' On average it takes 5 months to grow a duplicate clone of an original by means of advanced scientific and technological regenerative procedures.

Cloning Centre and Cloning Technology

A cloning centre is a place where clones are produced. Cloning technology are the advancements in medicine, science and technology used to produce duplicate and replicate copies of originals.

Drip Feed Disclosure and Evaluative Conditioning

Drip Feed Disclosure

Drip feed disclosure is the process of supplying information but in small amounts overtime. Drip feed disclosure is also the process of revealing information slowly overtime, possibly telling lies to conceal certain aspects of the truth until the source administering the drip feed disclosure has adequate time to let out the truth in a slow and controlled way, thereby delaying the betrayed partner (in this disclosure, the public) from having the "complete truth" for some time.

Drip feed disclosure is also a method to gauge public reaction used by governments, the media, multinational corporations and organisations as well as high ranking officials to "test" whether the general public is acceptant of the concealed information or not. When the public reacts favourably to the drip feed disclosure, more information is revealed and made public, and it appears to the unsuspecting observer that the people involved in making the disclosure are taking positive steps towards a favourable goal for all. However, when the public reacts adversely; information contradicting the drip feed disclosure is presented, and an expert is presented to the public who voices the concerns of the general public, and therefore it appears that the opinions of the public have been noted and research and development will not continue in the initial stated direction. Nevertheless, the truth remains concealed and research and development continues despite the aversions of the public.

Examples of drip feed disclosure involve articles such as mind uploading / downloading (BEAMS 2007), Mind Clone Robots (Bloomberg Business 2015; RT 2015), The 2045 Avatar Project (2045 Initiative 2015; Borghino 2012). I am here to tell you that I have been a spy for over 30 years and such technological accomplishments which are posted under trans-humanism / post-humanism genre have been realised many years ago and are **available** today. They are just concealed from the public.

Furthermore, Dolly the sheep was announced as the first publicly cloned mammal (Animal Research 1996), but how many readers can say they know that, four years later, a monkey (BBC News 2000), our closest primate, was cloned? This is publicly disclosed knowledge, but I suspect not many people know of a cloned monkey four years after Dolly. This is because there was such a worldwide adverse reaction to genetic cloning when Dolly was made public, that the disclosure of the cloned Rhesus monkey (BBC News 2000) was not made public knowledge on a grand scale.

Evaluative Conditioning

Evaluative conditioning is defined as a change in liking, which occurs due to an association with a positive or negative stimulus (see De Houwer et al., 2001). Simply put, this means that our preferences for brands, products, people and other things can be influenced and even modified by the presence of something we like or dislike strongly (Hale 2012).

In many settings, a neutral stimulus, called a "conditioned stimulus", often coincides with some desirable or undesirable object, called the unconditioned stimulus. An unknown brand, for example, might appear in a commercial that also depicts a happy child. Over time, stimuli that often coincide with desirable objects are perceived more favourably, whereas stimuli that often coincide with undesirable objects are perceived less favourably--called evaluative conditioning (De Houwer, Thomas, & Bauyens, 2001; Walter, Nagengast, & Trassilli, 2005; Moss 2009). An "unconditioned response" is a response to a neutral stimulus we have no / little control over. It is a natural automatic response.

In other words, our preferences for liking or disliking brands, products etc. (the neutral stimulus) can be influenced by embedding (implicitly placing) the brand etc. (neutral stimulus) with positive or negative associations. Overtime our conditioned response becomes one of conditioned favourable or negative response towards the brand, product etc. (neutral stimulus) when we are faced with the brand, product etc. at a future date. Our preferences have been guided overtime. Evaluative conditioning can change our preferences when carried out subliminally or implicitly; it does not have to be explicit.

In media; music; movies; and other forms of popular culture and entertainment, images and symbols (of stimuli which cause undesirable consequences) are embedded subliminally and implicitly and are paired with positive associations as a form of evaluative conditioning. This causes individuals who have no preconceived judgements of the stimulus to be guided to have positive associations with a stimulus which causes undesirable consequences.

In popular culture; media; advertisements; movies; music and other forms of entertainment, evaluative conditioning is used as a method of hinting (that something is wrong); showing off of power (i.e. nothing can be done to stop the negative stimulus) and as a form of gloating (i.e. we've pulled off the negative stimulus / we are pulling it off). In over the 30 years I have been a spy, I have witnessed methods of evaluative conditioning used to hint, show off power, and gloat more often than not, and evaluative conditioning is not used just to guide social order.

The fortunate aspect is that once a person learns that his / her preferences are being guided by methods of evaluative conditioning, in order to influence the person to have positive associations with negative stimuli (or feel powerless towards the negative stimuli); and that the stimulus (brand / product) does in fact cause undesirable consequences –then the ‘spell’ is broken. The person can now choose how he or she responds to the brand etc. Usually, once all is known: a negative stimulus is associated with negative associations; despite it being portrayed as positive through evaluative conditioning.

Consciousness Transfer

Consciousness is defined as “the state of being aware of and responsive to one’s surroundings; a person’s awareness or perception of something” (Dictionary Reference 2015). Consciousness can also be described as: individual awareness of a person’s unique thoughts, memories, feelings, sensations and environment (Cherry 2015).

John Locke (1632-1704) was an English philosopher, Oxford academic and medical researcher who argues that it is sameness of consciousness rather than sameness of substance that constitutes personal identity. Consequently, if the psychological life is transferred from the body of a prince to the body of a cobbler (shoe mender), Locke argues, the resulting person will be the prince and not the cobbler. He would be responsible for the prince’s actions and not the cobbler’s; those who were close to the prince could continue their relationships with him but those who had relationships with the cobbler could not, and so on (Schechtman 2012, p.334). Moreover, basic Lockean intuition has proved to be that “consciousness transfer” can be thought of as the feat in which the person moves from one body into another (Schechtman 2012, p.334).

Consciousness transfer can also be thought of as the process of transferring or copying the mental content (including long-term memory and “self”) from a particular brain and copying it to a computational device; artificial body or avatar body such as that of a robot or clone version of the original. The computation device, robot or the clone, will then respond essentially the same way as the original brain (as suggested by Lockean theory on consciousness transfer) and therefore the computational device, robot or clone experiences having a conscious mind and essentially the behaviour of the computational device, robot or clone, can be attributed as belonging to that of the original.

It is **OF MONUMENTAL IMPORTANCE** that the reader understands the above statement regarding consciousness transfer is a method of **CURRENT** concealed advances in technology and not a concept relegated only to the genre of science fiction. At the very least the reader should be **open** to the possibility of consciousness transfer and the implications of consciousness transfer, in order to truly begin to understand the extent to which current concealed advances in technology are used to commit monstrous crimes.

I understand the above sounds unsettling, but there are also many wonderful concealed advances in technology (detailed in the disclosure section). The marvels in technological feats and human accomplishment will be released for the benefit of mankind once the world learns about the monstrous crimes which are committed against the them, and the earth, through the use of highly concealed technological advances and the good people of the world band together to put an end to the tyranny around us.

Now is also a good time to remind the reader that if military technology advances by a rate of 44 years for every 12 months which passes, then since 1945 military technology has advanced by a rate of more than 3000 years compared to the technology the public is currently accustomed to (44 times 70 = 3080). Now ask yourself honestly, in 3000 years from now (the year is 5015) don’t you think humans will have been capable of developing a method which allows them to transfer their consciousness from one body to the next and more?

I am here to tell you that I have been a spy for over 30 years and that from the intelligence I have gathered over the years it has been illustrated to me countless times that consciousness transfer is indeed fact and currently exists. Consciousness transfer has been achieved since 1945, although this accomplishment in human advancement (and more) is concealed from public knowledge. Moreover, consciousness transfer is a highly advanced form of concealed technology used by high ranking members of society to commit unspeakable crimes against the public.

Furthermore, for the reader who wishes to understand how consciousness transfer is possible, see Petkova and Ehrsson (2008) and Ehrsson (2013). In the video Professor Henrik Ehrsson (2013) demonstrates:

1) Consciousness transfer from one body to the next (owning another body other than ones original)
2) Physiological evidence for owning the new body. In other words when consciousness has been transferred and the person perceives the new body as that of his or her own; when the new body is threatened as the mannequin (new body) was in the experiment, individuals still perceived the mannequin body as that of their own and became frightened (displayed biological and physiological responses).
3) Visual perception and stimulation causes us to perceive ownership of a new body.
4) The new body we inhabit must be similar to that of our own (it does not matter whether the new body is smaller or larger than our original so long as the measurements are proportional) for consciousness transfer to occur.

5) That it is possible for consciousness to be transferred to another person's body and have (or perceive) ownership of other person's body while being localised in your own body. In other words, dual consciousness is possible.

Petkova and Ehrsson's (2008) and Ehrsson's (2013) research gives one of the clearest publicly disclosed explanations of consciousness transfer and I urge the reader to watch the video in order to better understand my disclosure. One of the main findings from his research is that perception is not rigid; and perception does shape reality. Sight and synchronous (going on at the same time) stimulation can, and does alter the brains perception of reality. Sight and synchronous stimulation activates certain parts of the brain (the sensory parts of the brain and the motor (movement) parts of the brain). The match between the two (sight and stimulation) "convinces" the brain that 'hey, I'm no longer in this body; I'm in that one or that I have a third arm (although it is a false limp) etc. Accordingly, "perception is reality".

Disclosure: Save the Victims through your Diligence

I've done my best to preface this information as logically, sequentially and methodologically for the reader as possible. I have stated what the law is on 'whistle blowing'; I have given references and definitions where appropriate and I have provided many sources in order for the reader to research, corroborate, and better understand my disclosure, so that he or she becomes compelled to help me and victims like me, defeat tyranny and save the future of mankind; therefore if there are any areas of my disclosure which still appears suspect –and I understand, after all my efforts –things may still not appear clear to some readers because we are dealing with highly advanced concealed technologies which are not available in the world the general public lives in; and therefore it is difficult to fathom and furthermore conceive that men can be so evil to their fellow humans through the use of advances in science and technology. Nevertheless, I promise you, I have no reason to lie, I am victimised daily because of these technologies, and therefore my only option is to tell the truth, and nothing but the truth so that the people of the world can put an end to this grave injustice.

For the reader who still finds it difficult to accept my disclosure after first read, I want you to do two things:

1) **Give me the benefit of the doubt.** I am asking you to do this because as I stated earlier, one of the hardest things to do as a victim of abuse is to come forward about the abuse you have suffered. It is even more heartbreaking when nobody believes you. Consider all the victims of paedophiles you have heard about on the news. I cannot speak for their individual cases but I can attest to the fact it affects my mental psyche beyond comprehension; there is nothing worse than being ignored and that nobody believes you. So on that premise, please give me the benefit of the doubt, because I'm sure the majority of good human beings out there will feel worse when it is proven that I am a victim of severe abuse and you ignored me, when you could have done everything in your power to save me.

2) **Set out on a genuine quest to debunk my disclosure.** If my disclosure appears suspect to you to begin with, then set out on a genuine and investigative quest to debunk my disclosure because, if my disclosure initially appears circumspect to believe, then I should be proven false within minutes of your research of my disclosure, because even professionals make mistakes. There will be something about the disclosure which will just not 'add up'. However, when a person is telling the truth, then it is very difficult to debunk that person.
I promise you, it will be worth your time to know the real truth about the world, simply because the truth is liberating.

I would also like to remind you that the truth about the world has been hidden from the public for many decades, therefore when you do hear the truth of course it is going to sound absurd to begin with; but keep an open mind and pursue till you know what truth is; because "to "know" is to understand with certainty".

Words of Inspiration: Personalized from a Theoretical Physicist –Thomas Campbell

The following gives an introduction to many aspects which occur in this world behind closed doors or in plain sight depending on how familiar the reader is with the topics. Evidently, a lot of what is written predominantly affects the people of the world and it is time the people of the world stood up and reclaimed their world. I cannot stress this importance enough! The disclosure is also a basis which explains many of the anomalies which exist in the world today. I apologise in advance if the disclosure is too shocking to be believable (initially) and it shakes the foundations of your belief systems... but as I've said, rather than lash out at the information, you must question your own belief system... and probe into the mechanisms which have influenced you to adopt your current belief system... only then does the shock subside... and you can move forward with an open mind through investigation, which should lead you to either corroborate or debunk the disclosure. Keep in your heart: scepticism, optimism and open mindedness (Campbell 2008). Scepticism keeps you from believing fallacies. It keeps you from falling into that which you have not yet proven for yourself. Scepticism is a requirement for breakthroughs. Optimism gives you the ability to give merit to that which you have not yet proven for yourself and Open Mindedness gives you the ability to "see" that which has not yet been proven.

The Science of Hidden Consciousness

Furthermore, "Consciousness Science Kept Hidden" (YesEthan 2013) illustrates:

- "The heart is an electrical organ. It produces by far the strongest form of bio electricity in our body; up to 40-60 times stronger than the second most powerful source –which is the brain. This electrical energy travels through every single cell in our body and in a sense, binds the cells together. The bio-electricity field is strong enough that it can even be detected outside of the body; out into space; beyond the skin. It is very measurable electromagnetic energy, much like radio waves. What we found is that the heart produces an electromagnetic field that surrounds our entire body 360 degrees; and it can be detected about 3-4 feet with megatron meters outside of the body. Researchers at the medical facility in Kansas –say they have detected it 10-12 feet. So regardless of how far it goes, what is interesting is that we produce this electromagnetic field which can be detected; it can be measured, with sensitive mainstream medical equipment." -Howard Martin (co Author) "The Heart Math Solution""
- Scientific studies have shown: that the heart is not only the most powerful centre in the body, it is the most intelligent and has the ability to receive precognition (knowledge of a future event or situation, especially outside one's normal sense of perception). If we can create coherence with our heart and mind centre we can access our intuition more frequently.

When you read my disclosure, remember that it is the heart and NOT the brain which is the most powerful centre in the body; the heart is the most intelligent; it has the ability to receive precognition and intuition about a future event, **before it actually occurs**; therefore if you find that your heart is in resonance with the themes I disclose; do not fight it with the ego of the mind; when our egos get in the way it overrides our heart's intuition; your heart is capable of discerning the truth of an event before your brain becomes conscious of it; by all means listen to your heart, and allow your brain to naturally develop its intuition. It will be the coherence between your heart first, and then your mind centre which will allow you to access your intuition.

Make an effort and reach your own conclusions to the point where you feel comfort in the fact that you **KNOW** the information contained in this disclosure is indeed factual. This will be the point you will have become empowered with golden truth and compelled to fulfil your duty as a human being and also spread the truth far and wide. Yes, we are all living in very interesting times.

Disclosure: Donald Marshall's Message to the World

My name is Donald Marshall. I have been cloned by a large secretive cult know as "The Freemasons" and "The Vril Society" and "Scientologist" together called the Illuminati. For readers unfamiliar with The Illuminati see "The New World Order" (1990) by A. Ralph Epperson and Appendix C in this document; an introductory guide to the Illuminati and their agenda for the world is presented.

Furthermore, I have to tell you that human cloning has been done since 1945. Now is also a very good time to recall Phil Schneider's comment (Schneider was a Geologist and Engineer who worked on black projects and Deep Underground Military Bases (DUMBs)): "That for every 12 months which passes, military technology advances by a rate of 44 years compared to the technology the public is currently accustomed to" (Schneider 1995; 1996; Open Minds 2011). George Green (a former Financier affiliated with U.S. Presidential candidates) also attests to the fact that humans have been cloned since 1938 (Green 2008a; 2008b). Therefore if we take 1945 as a base year, then military technology or hidden technology has advanced by the rate equivalent to 3080 years than the technology the general public is currently accustomed to (2015 – 1945 = 70 years; 70 multiplied by 44 = 3080). So yes, humans have been cloned since 1945 and continue to be cloned today and that current hidden technology in the year 2015 is as much as 3000 years ahead of the technology the public is currently accustomed to. I'll repeat myself because it is **very important** the reader understands the above statement, even if, the rest of the disclosure is hard to fathom after first read. The two **most important** things to keep in mind are:
1) Humans have been cloned for over 70 years (since 1945);
2) **Present** hidden technology is more advanced -as much as 3080 years more advanced- than what the public is currently accustomed to and technology continues to advance at an incredible rate.

The Two Types of Cloning Techniques

Now there are two different types of cloning techniques, there is replication cloning and duplication cloning. Replication cloning is the type of cloning the public has generally heard of. Replication cloning involves taking the nucleus (the DNA) out of a donor egg, and replacing it with new DNA from the person to be cloned. After a few days the resulting embryo can be implanted for pregnancy. The newborn starts life off as a genetic copy of an original.

Duplication cloning, on the other hand, is the second hidden method and type of cloning where the DNA from the person is grown in a big thick tank full of (salty) water. It involves a method of regenerative technology where the cells are agitated and agitated and over the course of 5 months, a fully formed duplicate clone body of an original is developed. The Illuminati used to have to use a tissue sample from the original or the cells from women's pap smears because this contained rich cells to make duplicate clones; they also used children's foreskins which were discarded at the hospital that got removed. That is what happened to me. I had my foreskin removed at age 4 –and by age 5 the Illuminati grew duplicate clones of me. Now the Illuminati say they have upgraded the technology since 2000 and now all they need is blood from the original. They then agitate the blood cells over and over again through regenerative technology, until a fully formed duplicate clone body of an original is produced. For readers who are unfamiliar with regenerative procedures, see Dr Stephen Badylak's video on "How to grow a new fingertip" (Science Channel 2014) and Carmichael (2013) which discuses how scientists cloned a mouse from a blood sample (this is another example of drip-feed disclosure). Regenerative science and medicine (CBS 2008) does work; I've seen it many times.

The Different Grades of Clones: "Mark 1" to "Mark 4" Clones

There are a few different grades of clones that I know of. There are: "Mark 1" to "Mark 4" clones. Mark 1 clones were available at the end of World War II (1945). Mark 1 clones were rapid eye movement (REM) driven clones but they were primitive grades of cloning with lots of side effects caused to the original. Mark 1, REM driven clones were called "Organic Robotoids", even though there are no robotic parts to the clone at all. See Dr Peter Beter's discussion on "Organic Robotoids" (Beter 2011).

Donald Marshall's Dilemma with "Mark 2" Sleep Driven REM Clones

The problem I am having is with Mark 2 REM driven clones. "Mark 2" is a rapid eye movement (REM) driven clone. In other words, Mark 2 clones are sleep driven clones. REM sleep is the fifth stage of sleep (Sleepdex 2015). The first REM cycle usually happens 90 minutes to 110 minutes **after** we fall asleep (Sleepdex 2015). What that means is: **currently,** this is the Illuminati's main form of communication. They do not call people on the phone; they do not meet at the Bohemian Grove anymore (they only meet there once a year for traditional purposes). Since they discovered the science of sleep driven cloning, they meet at the cloning centre **WHEN THEY GO TO SLEEP** (The cloning centre is a physical location, located 5 / 6 hours radius from the Robert Pickton Farm (Port Coquitlam, British Columbia, Canada) at a nature reserve).

Consciousness Transfer Happens When the Original Reaches REM Sleep

The Illuminati can transfer the consciousness of an original once the person reaches REM sleep if there is a duplicate clone of the original at the cloning centre, once the person goes to sleep (90 minutes to 110 minutes after falling asleep). So what happens is that the consciousness of the original is transferred to the duplicated clone body (the duplicated clone body is grown within 5 months by regenerative technology) once the original reaches REM sleep. The original's consciousness is transferred from the original's body, although the original's body is still in their bed, asleep at home; to a duplicated REM driven clone body at the cloning centre and the original 'wakes up' as a cloned version of himself / herself at the cloning centre. It is a great marvel of science and one man's greatest achievements, but it is kept hidden and used for sinister purposes which I discuss below. I must address the readers at this point who are lost at the thought of duplication cloning, and consciousness transfer.

For readers who may have difficulty understanding the process of duplication cloning, see the video which features Dr Stephen Badylak (CBS 2008; Science Channel 2014). Badylak describes and illustrates how a new fingertip can be grown within 4 weeks, by the process of regenerative technology. Duplication cloning works a similar way; involving regenerative technology whereby the cells are agitated and agitated, and over the course of 5 months a fully formed duplicate clone body of an original is grown. Moreover, readers should also view the Horizon documentary "Why Do We Dream?" (2009); the documentary clearly explains the phases of sleep; particularly REM sleep. The documentary also clearly explains that during REM phase sleep the whole body shuts down, and it is only the brain which is active. The fact that the brain is still active, although the body is inactive, is a perfect opportunity for consciousness transfer; and this is usually the moment the Illuminati transfer the consciousness of a person who is asleep, to a cloned version of himself / herself. The reader should also explore Ehrsson's (2013) lecture on consciousness transfer. In the video, Ehrsson (2013) clearly explains how through the first person visual perspective (seeing through the eyes), and through synchronous (occurring at the same time) stimulation of a body part, individuals can perceive and see the world from another body different from their original. In other words, so long as people can see through the eyes, and the body is stimulated, the match between the visual perspective and feeling body sensations allows the individual to see and perceive the world through another body different from their original. The consciousness has been transferred to another body.

Ehrsson (2013) also demonstrates that consciousness is linked. Once consciousness is transferred (from the original's body to the new body), even though the new body is not the persons original body; when the new body is attacked, because the person perceives the world through the new body, everything feels "very real". When the new body (for which the person's consciousness has been transferred to) is attacked, the original still perceives the threat as "real", and therefore the person displays a biological and physiological fear response in their original body; the heart rate increases, as does respiration (breathing) rate, as well as anxiety.

Donald Marshall is an Original as he sits at Home Typing this Disclosure

This is how advanced technology is today. This is also exactly what happens to me. **I am an original as I sit at home here typing this disclosure,** but when I enter REM sleep, my entire original body shuts down, and only my brain is active. The Illuminati have linked my consciousness to a REM driven duplicate clone body of me. Therefore, as soon as I enter REM sleep, they are alerted to the fact that I have entered stage five of the sleep cycle, and in REM sleep, because they have a green and red light above duplicate clones at the cloning centre. The red light indicates that the original is awake, and the green light lets them know that original has entered REM phase sleep and it is time to transfer their consciousness to a duplicate clone version. They cannot transfer the consciousness of an original before REM phase or when the original is awake. When they try and transfer consciousness before an original is in REM phase it makes the original have intense headaches. Setting an alarm every 90 minutes to wake up and then go back to sleep also does not work in terms of avoiding REM sleep and having your consciousness transferred, because the body requires REM sleep, and therefore you become very tired, quickly, if you do not get REM sleep. If you stay awake for four days without REM sleep you begin to hallucinate; after seven days without REM sleep you will die.

When Donald Marshall’s REM Driven Cloned Experiences Began

This whole sordid episode began when I was 5 years old, after going to the doctors to get my foreskin removed, at age 4. The tissues from my discarded foreskin were used to grow multiple duplicate clones of me. Within months, I was having my consciousness transferred to a REM driven, 5 year old clone version of me every time I entered REM sleep. The reason the Illuminati did this was because they wanted to use me as what they term “a diddle kid” – in other words they wanted to have sex with an REM driven, 5 year old, duplicate clone version of me. This is what they do. They clone people. They make duplicate clone bodies of people to victimise in terrible ways, and they clone children for men with undeveloped penises to have sex with.

However, I could never remember any of these REM driven clone experiences when I woke up for 25 years; because another aspect of REM driven cloning is that the Illuminati have the advantage of being able to suppress the experiences of people who have their consciousness transferred to their REM driven duplicate clone versions through advanced memory suppression technologies. For the reader who wishes to understand how memory suppression works see Winter (2014). Memories are stored and retrieved from different parts of the brain. Interactions between the cerebral cortex (the outer layer of the brain) and hippocampus need to work together in order to bring the memories out of mental storage to be re-experienced by the mind (Winter 2014). Therefore all that has to be done is for them to turn off the nerve cells which communicate with the cerebral cortex and hippocampus for the evening which you went to sleep and had your consciousness transferred to a REM driven duplicate clone version of you, and you will not remember anything. The technology is so advanced; they can, and do suppress memories at a push of a button. Therefore, the victim wakes up the next morning and will not remember a single REM driven duplicate clone experience. All that they will be aware of is that they did not have a dream the previous night.

This is what happened to me for over 25 years. I wasn’t allowed to remember these REM driven duplicate clone experiences until I turned 30. My memories were suppressed for 25 years. I would wake up many days, after a night’s sleep, thinking ‘I didn’t have a dream last night’; or I would wake up the next morning and I would be sick, although I went to sleep the previous night feeling completely healthy. Throughout the years, I thought I had terminal illness; because whenever I went to see a doctor, the diagnosis would always come back that ‘there was nothing wrong with me’.

However, over the course of 25 years when I was memory suppressed, and not allowed to remember my REM sleep driven clone experiences, the Illuminati were using me to produce songs, lots and lots of songs, and whenever I could not produce a song, or did not want to, they would stab my REM driven clone body; sodomize my REM driven clone body; chain me to a crucifix and burn my REM driven clone and torture me for all kinds different reasons; and even when I was compliant and did what they told me to, I was attacked, simply because some members of the Illuminati wanted to know what it feels like to stab someone. Now because consciousness is linked (Ehrsson 2013), all these REM clone torture experiences affected me in my original body, and I have a weak heart today because of it, at the age of 39. The memories of my REM clone driven experiences also had to be released to me slowly (see Alford (2014) and FW: Thinking (2014) –for an example of how memories can be wiped and restored) over the course of many months because I had been through so much horror and torture that if the memories was released all at once, I would have had a heart attack in my real body and died.

This is one of the main reasons the Illuminati currently use REM sleep driven cloning (besides sex and torture): to plagiarise talented people under duress. There are many people including me who have been, and are plagiarised on a daily basis. Their consciousness is transferred to their REM driven clone alternates, and night after night and under duress they are forced to reveal their money making ideas or face torture or clone death; clone death after clone death. That is what the movie *Inception* (2010) is about: REM sleep driven cloning, and stealing ideas from the minds of unsuspecting individuals. The Illuminati actually made the film *Inception* (2010) as well as *The Island* (2005), *Avatar* (2009) and *The 6th Day* (2000); the films mentioned above contain references of cloning and REM driven cloning to show off the Illuminati's 'power' and throw in the world's face and laugh.

The Science of REM Driven Cloning Began in 1945

I am one of the few people fortunate, or unfortunate depending on how you look at it, who remembers all of my REM driven cloning experiences because these memories were released to me at the age of 30. I remember many people who have attended and still attend the cloning centre which they bring me to over the course of 34 years. I am currently 39 years old and was brought there since the age of 5. The science of REM sleep driven cloning was first discovered in 1945, and at first it was just a political thing which leaders of the world did. Through the science of REM sleep driven cloning, political figures, heads of state and royalty met each other as REM driven clones versions of their selves when they went to sleep and discussed worldly affairs in complete secret with each other. They also did whatever they wanted with each other as REM driven clones, which they could not do in their original bodies such as sex, fighting each other to the death, jumping off cliffs as their REM driven clones, and just about anything you can think of when one has the ability of pseudo-immortality as a sleep driven clone version of themselves hidden from the guise of the world and general public.

Well they soon got bored of each other very quickly, and started to make REM driven duplicate clones of movie stars, musicians, celebrities and public figures from all walks of life to hang with as REM driven clones and now they all get together for a disgusting time as REM driven clones when they go to sleep. Most of the G20 political leaders meet at the cloning centre as REM driven clones when they go to sleep to discuss worldly matters and watch gruesome things done to innocent and unsuspecting civilians, who are also REM sleep driven clones at the cloning centre, but civilians have their memories suppressed. This is what the films *Hostel* (2006); *Hostel: Part II* (2007); and *Hostel: Part III* (2011) (Illuminati made films) -are about: -rich people paying to torture and kill innocent people at a secluded location for sport; just that at the cloning centre, the Illuminati members torture and kill REM driven clones of innocent people.

Therefore, depending on what was done to your REM driven clone, you could wake up the next day with a very intense headache in your **original body**, if your REM clone was repeatedly punched in the head; an upset stomach in your **original body** if your internal organs were tortured; achy limbs; and worst of all, feel sick all over; although the previous evening you went to sleep feeling healthy. Remember consciousness is linked (Ehrsson 2013), therefore whatever happens to your REM driven clone will affect you in your original body. Even when they suppress your memory; the previous night's experiences still affects you in your original body. You wake up with the knowledge that you did not dream the previous evening when you went to sleep, and with certain illnesses depending on what was done to your REM driven clone.

Help Donald Marshall STOP REM Driven Cloning NOW!

Everything I discuss in this disclosure is firsthand witness accounts. I have seen and experienced horror spectacles done to innocent civilians and to me. For the sake of humanity and the innocent children brought to the cloning centres all over the world; I must tell the whole world about this and the good people of this world **MUST HELP ME TO BRING THIS TO A STOP.** My sense of duty, moral correctness, and conscience cannot allow me to stay silent about this evil. REM sleep driven cloning, and the torture of innocent civilians as REM driven clones; having REM driven clone sex with beautiful, innocent unsuspecting civilians of the world; the paedophilia of children through the process of REM driven cloning; REM driven sex slavery; the plagiarism of talented individuals through the process of REM driven cloning; and REM driven idea slavery, whereby talented individuals must produce new ideas and concepts daily (or whenever they sleep) or be stabbed or tortured in their sleep as REM clones –is the most vicious and pernicious form of tyranny ever to occur against humanity and **IT MUST BE STOPPED NOW!**

Do Not Panic, Riot, or Cause Chaos. This is Very Important.

Now I must preface this carefully, because members of the Illuminati have told me that I have to put this in an eloquent fashion, in a way which does not make people panic because people finding out about REM driven cloning, and the extent of the evil it has been used for, and continues to be used for; it could cause loss of social order, riots and anarchy in the streets. Moreover, I too **DO NOT** want riots and anarchy in the streets, despite the fact that I am vehemently angry considering the extent to which they plagiarised my talents over the years. If you are a good person reading this, and you want to help and you want social order restored for the benefit of mankind, promise yourself; me; all the innocent children they have affected through REM driven cloning, and the children of the future that you will **NOT** riot, and destroy a world for children who are going to inherit the world. **This is very important**. Remember the Illuminati have highly advanced technologies (kept hidden and secret), including weaponry, and they are just waiting for any excuse to use it on the populace. In all revolutions, the populace always win, and I want this disclosure and the end of REM driven sleep cloning to go smoothly. I want the good people of the world to keep spreading this information, keep spreading this disclosure all over social media, tell your close friends, your family, and as many people on social media platforms as you can. The internet, and radio shows with small audiences are the only forms of communication the Illuminati do not control. They own all forms of television networks, even including the Aboriginal People's Television Network (APTN), and therefore a message as important as this will never reach the average man or woman through television because it does not serve their agenda.

Spread this Disclosure Document until it reaches the Armed Forces

I want this message to reach the armed forces, because currently people in the armed forces are following orders, yet they do not know how corrupt their governments and people in high profile places are. They are unknowingly defending these corrupt people. People in the armed forces will **NOT** defend these corrupt people nor will they harm civilians when they realise that these corrupt people have been growing duplicate clone bodies of civilians, transferring civilians consciousness to their duplicate clone versions when the victim reaches REM sleep, and torturing civilians in their sleep; having sex with under-aged REM driven cloned children; having sex with innocent and unsuspecting adults as REM driven sleep clones, against their will; torturing REM driven clones for sport; and for money-making ideas to benefit their own pockets while they make innocent civilians sick and have side effects in their original bodies from REM sleep driven cloning technology. No. The armed forces will not accept that. Therefore keep spreading and sharing this information for everyone to see so that it reaches personnel in the armed forces **ASAP!** I want the armed forces to overthrow these corrupt people. The Illuminati think they are very sly and untouchable because together they hold most of the wealth in the world. We must show them that they are not untouchable. Please do all you can, to share and spread this message. It is the most important message to ever reach the internet because it affects all our liberties.

I must also address the readers who may feel afraid in spreading my message because they may want to do the right thing, but they are terrified the Illuminati may degrade or end their lives. As unbelievable as the next thing I am going to share sounds, it is a belief system which shapes the reality of the Illuminati and therefore they themselves are afraid to do anything which may degrade their lives or 'eternal soul'.

They believe that they are the 'fornicators' mentioned in The Bible, and if they were to degrade or harm the lives of anyone aiding me, they would suffer the wrath of God when they die. Notice I said **when they die**, but so long as they are still alive they can do all the evil they like, with little consequence, as they do in their REM driven clone versions at the cloning centre; worshipping Lucifer and doing all sorts of ungodly things. They even say God does not exist, only science and technology; yet they are afraid of dying. Very afraid of dying; because they believe they'll meet God's judgement. I'm just relaying what they have told me. They are weird like that. They also follow Hopi Indian Prophecy, Mayan Prophecy and Nostradamus Prophecy. They mix and match those three prophecies and come up with their own religion of what may or may not happen in the future.

Donald Marshall is NOT "The New Age Saviour". Nobody wants to be a 'saviour'

Furthermore, **THEY** believe I am a new age saviour for the end times; that I was going to save the world from something, according to the Nostradamus prophecy (quatrains); the Illuminati consider Nostradamus the greatest prophet ever to walk the earth. They are in awe of him.

I **DO NOT** endorse their claims. I just want REM driven cloning to stop and these people face punishments for their crimes against humanity. I really do not endorse such claims, but if it saves my life and yours, so be it. In reality and practicality, nobody wants to be 'the saviour'; not with such tyranny and pernicious evil such as this. We'd rather all just be living normal lives without the knowledge and practice of REM driven sleep cloning which is used for torture, clone death spectacles and sex with children. My friends on Facebook do not want to be heroes or 'saviours' either; nobody really does, but we do it out of a sense of duty, moral correctness and our conscience would never rest knowing the true extent of evil in this world, and doing nothing to end it. So all that saviour and Nostradamus prophecy talk, it is just the Illuminati; they like to quote prophecies, and make them come true to make themselves feel special.

Like I said, they are weird. I'm just relaying what they say, and I **do not** endorse the new age saviour claims. I'm just a normal guy, who was unfortunate to be cloned when he had his foreskin removed at age 4. This could have happened to anyone. Thankfully I had a very active imagination at the age of 5 and with the use of the technologies such as Mind-voice technology (that is what the Illuminati call it) it allowed me to make songs to keep the perverts off me. Mind-voice technology is just a technology which can tune into the inner voice in your head (See articles by Prigg (2014) and New Scientist (2014) –which discuss how scientist can listen to your inner voice; –it is similar to how Mind-voice technology works). Therefore, when you hear the sound of a guitar; the sound of drums; or any instrument for that matter; after hearing the sound, when you imagine the sound you heard, your REM driven clone can replicate this sound with Mind-voice technology **exactly**; and I used Mind-voice technology to make many, many songs over the years. As I've said, and I'll say it again: technology today is very advanced.

How the Nostradamus Prophecies shapes the lives of the Illuminati

Also another reason they have not killed me is because the Illuminati said "If they do, everyone is going to know I am telling the truth", -which I am with my right hand to God and I cannot lie about this stuff; the things I have seen are too sinister and diabolical. It would make me as bad as them if I did. Furthermore, because they follow the Nostradamus prophecies religiously, they have told me that they believe anyone helping me is considered part of "The Army of Light" and again, to hurt or degrade anyone helping me, considered part of The Army of Light would incur the wrath of God; and that to hurt anyone considered part of The Army of Light will degrade their lives; they will suffer misfortune and entire ruin as will their eternal soul. These are just some of the interpretations of the Nostradamus quatrains they have relayed to me. These people are beyond crazy, zealot and religious fanatics. Nevertheless, the main point to remember by all this is that **you are safe**, and **they cannot hurt you**, **they are scared to hurt you**, as well as kill me because prophecies have shaped their lives and the lives of their ancestors for hundreds of years.

Furthermore, and practically, I have told too many people that this is the extent, to which technology has developed in our current day, and it is kept secret and hidden from the public for nefarious purposes and this is what the Illuminati do: REM sleep driven cloning. I have been spreading this information **since 2011** and not one of these high profile people, government officials or celebrity figures has issued a (public) statement against me or taken me to court over libel charges because I'm telling the truth; and their statements would not hold in court.

I have also told over a million Arabs; I have appeared on radio interviews with Vinny Eastwood (Vincent Eastwood 2013); Jeanice Barcelo (Jeanice Barcelo 2013); and Lisa Phillips on the Cry Freedom Radio show (Astral 7ight 2013a – 2013h). I have reached audiences as many as 280,000 views on the Vinny Eastwood Show (Vincent Eastwood 2013) alone. However, I must stress that YouTube reduces the view count of my interviews every so often in order to suppress the truth so that my interviews do not get featured on their platform. Nevertheless, there is safety in numbers and rest assured there are new people waking up to the disclosure of REM driven sleep cloning every day; and therefore you can share and spread this information **without any worry** about facing reprisals from the Illuminati, even if you do not believe that it is in fact their beliefs in Nostradamus prophecies which is saving you and I from the Illuminati.

HIV/AIDS has been cured; so have many forms of cancer

Now is also a good time to share with the reader that this situation is not all 'doom and gloom.' In their quest to perfect technologies, science and medicine, the Illuminati discovered the cure for HIV/AIDS; cancer (except pancreatic cancer); Alzheimer's; Dementia and many more debilitating diseases which humanity suffers with. They relayed the fact that they found the cure for HIV/AIDS by stimulating the cells in an oxygenated rich environment because diseases/viruses cannot thrive in an oxygenated rich environment (in layman's terms). Magic Johnson was cured from AIDS this way and they use him as a spokesperson to sell the retrovirus pills which they know is not as effective. However, the greed these people have, knows no bounds, and because they receive too many donations for HIV/AIDS as well as cancer donations they do not announce they have cured HIV/AIDS or cancer publicly. As I've told you, and will continue to tell you: technology is far, far, FAR advanced than what you currently see around you. With the help of the good people of earth I promise to release these technologies for the benefit of mankind, and all reading this disclosure can hold me to this. You **must** hold me to this. It is time we ended corrupt governments and corrupt individuals in high positions of power in this world. Hold me to this.

I hope you are now beginning to understand why this is a **world emergency** and why these corrupt people must be ousted and overthrown. I also hope you understand why you must not riot or cause chaos because as I said at the beginning of this disclosure "I write this to empower good people of the world against the tyranny which exists all around us in our world today"; I **did NOT** write this to cause chaos, public dissension or anarchy. Good people of the world will inherit the earth once these corrupt people are overthrown. Therefore, under no circumstance must you destroy the world when you are going to inherit it. No matter how angry this disclosure makes you; no matter how angry the diabolical people who commit these crimes against humanity make you feel. Let's make sure this goes smoothly. Keep spreading and sharing this information, until this disclosure reaches the armed forces; until the armed forces bring these corrupt people to their knees. We are about to inherit the earth; by all means let's ensure everything goes smoothly. Now that you understand how important this disclosure is and that the overthrow of these corrupt individuals must progress smoothly, I'll continue.

The Behaviour of High Profile People at the Cloning Centre

The high profile individuals who attend these REM driven clone gatherings in their sleep have nothing better to do than to show off in disgusting ways. They have no shame and it seems nothing embarrasses them. Some REM sleep driven, cloning centre attendees, sit in the stands of an unused arena, smaller than a hockey rink, but it still has the capacity to seat between approximately 300 to 400 people. There is dirt in the centre of the rink where ice would be. They have frightened REM driven clones of children walk into the middle of the dirt rink to be victimised for a bizarre and disgusting spectacle. Sometimes, they have animals like dogs have sex with the REM driven clone children, while a man holds the dogs on leash so that it wouldn't bite the child on the back of the neck; which I have seen happen before. They all try to outdo each other in their levels of depravity; to be evil is to be "cool" to them.

The Ring Leaders of the Cloning Centre

I will now discuss the ring leaders, but before I name names, please understand that I am not someone who has collated stories and evidence together over the internet. I REALLY **do not** need to do this. Every time I enter REM sleep, the ring leaders at the cloning centre transfer my consciousness to a duplicate clone version of me. I have seen and talked with these monsters over the course of 34 years as REM driven clones at the cloning centre. However, I do understand that for the average reader, without understanding how far medicine, science and technology has advanced; without seeing or hearing people, other than me, discuss the crimes the ring leaders of the cloning centres commit, it is very difficult to imagine that this is what these high profile people are involved with, and do. Nevertheless, I guarantee it; this is the true extent and prevalence of evil in our world.

Their public personas, and public images is well managed by their PR teams, and therefore for the average person it is difficult to imagine these people in any other way but "positive"; but these people behind closed doors are the most evil, sick and twisted people I have ever had the misfortune of encountering. It is also difficult to take in what I am saying because I am not a public figure, nor do I have the "influence" or public image these people have; the general populace have never met, or heard of me, or known who I am; if this was the case my eye witness testimony alone will be enough; however, I do understand that we are dealing with highly advanced technologies, thousands of years advanced compared to what the general public is accustomed to, and because of this I provide references for the reader; I provide references for the reader to have something to corroborate what I am disclosing; and I provide references for the reader to see past the illusions these people have created.

Moreover, as "Consciousness Science Kept Hidden" demonstrates (YesEthan 2013):
remember that it is the heart and not the brain which is the most powerful centre in the body, the heart is the most intelligent; it has the ability to receive precognition and intuition about a future event, **before it actually occurs**, therefore if you find that your heart is in resonance with the names mentioned; do not fight it with the ego of the mind; when our egos get in the way it overrides our heart's intuition; your heart is capable of discerning the truth of an event before your brain becomes conscious of it; by all means listen to your heart, and allow your brain to naturally develop its intuition. It will be the coherence between your heart first, and then your mind centre which will allow you to access your intuition.

The references I provide for the reader to corroborate what I am about to say includes: "Appeal from Survivors of Canadian Genocide" (inifiniLor 2013); Tila Tequila –"Missing Children and Cloning Centres" (Astral 7ight 2013i); "Royal Babylon by Heathcote Williams" (MrCowshedder 2012); and "THIS MOVIE WILL BLOW YOUR F%SNG MIND" (KafkaWinstonWorld 2014).

Watch all the above videos and you will easily see all that I disclose. In Appeal from Survivors of Canadian Genocide interview (inifiniLor 2013): Listen to the indigenous people Stee-mas and Wahtsek speak about the pain and torment they have suffered under the Crown, the Vatican, the government, and churches of Canada. 50,000 to 150,000 indigenous children have gone missing because of the involvement of the Crown, the Vatican, the government, and the churches of Canada; and all these factions involved know this!

Tila Tequila is a television personality who has attended the cloning centres as a REM driven clone since childhood. In Tila Tequila's radio phone in (Astral 7ight 2013i), the audio is not very clear for the majority of her disclosure; she was also frightened and therefore she sounds most animated; however her words are clear at the following points of the audio (I also provide a transcript in the appendices section: "Appendix A"):

4.00 min: Cloning Centres and they take your children...
4.25 min: –There are these CLONING CENTRES –I'm not even talking about child molestation here...
5.40 min –They take your children –they not only molest them; men f*** them; and make them shoot each other...
7.59 min- There is a point where you cannot just turn the other way, you know this stuff is going on and you go 'oh well, you know... that's their problem; let's just turn the other cheek. How long are you going to turn the other cheek until it happens to your own freaking children?

10.34 min - Do you know why they love children? Because they are innocent souls; they are innocent... they are the most innocent, pure beings in this planet. They are not harmed by anything. They are new to the world; bright-eyed, pure innocent children. That is why these disgusting paedophile... and these clone rings, cloning centres, satanic rituals, Brownsville Texas... there are many of the of the cloning centres where they take your children that go missing.

Royal Babylon by Heathcote Williams (MrCowshedder 2012) addresses the fact that for Brits, despite their inordinate pride about their tradition, it is often revealed they know little about it than anyone. The documentary then goes on to address the fact that Prince Phillip the Duke of Edinburgh, has a lust for blood-sports which includes shooting endangered wildlife and mammals which he considers 'culling'. The documentary is also notable for demonstrating to the viewer that Queen Elizabeth II redistributed as much as 2.1 million for her private income from 276 people from Merseyside and Lancashire. These people had no will, so their property was grabbed. The same is true for the Duchy of Cornwall. If you die without making a will, everything you possess will go to Prince Charles. Furthermore, Queen Elizabeth II is the largest land owner on earth. Queen Elizabeth II: head of state of the United Kingdom, and 31 other states and territories is the legal owner of about 6600 million acres of land; one six of the earth's non ocean surface. She is the only person on earth who owns whole countries, and who owns countries that are not her own domestic territory. This land ownership is separate from her role as head of state; is different from other monarchies where no such claim is made –Norway, Belgium, Denmark etc. The value of her land holding is £17,600,000,000,000 (17.6 trillion Pounds) (approximately)).

In "THIS MOVIE WILL BLOW YOUR F%SNG MIND" (KafkaWinstonWorld 2014), readers with a deeper knowledge of the Illuminati and the agenda of the Illuminati will easily spot the minor inaccuracies in the documentary. However, the documentary provides a good reference and foundation for anyone new to the topic of the Illuminati and their agendas for the world. The documentary demonstrates how the world is currently a deception within a deception coated in reverse psychology. It gives widespread insight into the Royal family of England, and their lineage.

Furthermore, the documentary provides information that the world's most popular religions have been infiltrated with Illuminati dogma, in secret. The documentary details how (some) modern Christians, Jews, and Muslims do not (yet) recognise their common enemy because that enemy is invisible (until now). The enemy hides within their own religions. The enemy also hides within the activist community (anti-Illuminati movements), it finances big budget documentaries and movements. The end game is to sell a divisive (creating strong disagreement), anti-capitalistic, atheistic message and deliver trusting followers right into the lap of a communist (one government) New World Order. The enemy is the Illuminati.

The documentary also details that: "Today, the bloodlines of this unholy alliance are on the final stages of establishing one world religion, one world government, and one world ruler [-- that is **IF**, the good people of this world do nothing to stop them]. The Windsor Royals are at the control of this web of deceit. It is a known fact that Adam Weishaupt (the founder of the Illuminati) took refuge and got assistance from the German Saxe-Coburg Gotha Royal family. The Saxe-Coburg Gotha's changed their German name to Windsor (in 1917). The Windsor's public image is a façade, to hide the cesspool from which they operate behind the scenes; their power and control cascades out of Buckingham Palace to the elite bloodlines who run world affairs."

For anyone who thinks they know they Royal family of England, did you know that still to this day: they have a person appointed called "Groom of the Stool"? What is a "Groom of the stool", you may ask? A Groom of the stool is responsible for stools. The kind you sit on and the kind you expel. In other words a Groom of the stool wipes Royal bottoms. The above has been confirmed to me by Prince Charles of Wales himself as REM driven clones at the cloning centre, and I quote: "It gives you the feeling of being completely taken care of" when I asked Prince Charles (as a REM driven clone) why they still have that. But don't just take my word for this; research "Groom of the stool".

The Royal Family of England

As most of readers have already guessed; and as unbelievable as it may currently sound to some readers, the ring leaders of these cloning centres includes the royal family of England. As I've said REM driven cloning started as far back as 1945, and it was just a political thing, but Queen Elizabeth II has been involved with REM driven cloning since its inception! Yes! Queen Elizabeth II; Phillip, Duke of Edinburgh and Prince Charles of Wales are the worst of them. They are unbelievable depraved perverts all showing off to the celebrities who attend these REM driven clone gatherings. Elizabeth has the REM sleep driven cloned children at the cloning centre call her "Lilli-bet" and Elizabeth as a REM driven clone herself does ungodly things to these REM driven cloned children. Some she fakes being nice to; some she is terrible to; and as a REM sleep driven clone Elizabeth cuts these REM driven clone children with swords while they scream. The decent people at these cloning centres who have had their consciousness transferred into their duplicate REM driven clones, and are there against their will, like I am; they are terrified to say anything against Elizabeth and her cronies. Most of these decent people also have their children's consciousness transferred into their REM driven duplicate clones, and Elizabeth and her cronies hold these decent peoples' children as hostage; to be torn apart; clone death after clone death if they have the slightest inkling of informing anyone, but as Elizabeth and her cohort have been torturing me as REM driven sleep clones anyway, I will disclose all that they do.

Vladimir Putin is also there, as an REM driven clone, and he loves to put the fear of torture and death into people, but he is essentially a perverted coward himself. Most of the famous people who attended the cloning centre as REM driven clones are ashamed to speak or be seen by me there very much; because they are ashamed of the perverse and disgusting gatherings. I am a decent person, even when my consciousness is transferred to my REM duplicate clone alternative, and I will not participate in these depraved acts; so they use me as an example and torture me for being a good person. They have drugged my REM driven clone, and as I've mentioned because consciousness transfer is real (Ehrsson 2013) and because consciousness is linked, therefore everything feels "real as real"; when I'm drugged as a REM driven clone, I behave exactly as any person would on hard drugs (although I have never taken any drugs in my original body); my vision is blurry and my balance affected; they have used this as method of having me do disgusting things in a blurred state, videotaping these dirty acts and editing a highlight reel which also depicts me doing disgusting things to show to the world when the world learns of their depraved natures and their public personas are shattered so that I can be despised as much as they will be when the entire world learns about this. Furthermore, because of the microchip inserted in my REM driven clone; they mind controlled me like a "Manchurian Candidate" (*The Manchurian Candidate* 2004) using MK Ultra technology, which is far more advanced than the declassified MK Ultra techniques (MK Ultra Compendium 1980) mentioned online today which references MK Ultra techniques in the 1970s. See Rense (2001) for a detailed explanation of how implanted microchips work, and how these chips allow the programmer control victims. C.S.I.S Canadian intelligence are all involved; they have REM driven clones too; including a lot of Commissionaires and Canadian Prime Minister Steven Harper himself; They do what Elizabeth says, and seem to follow her every whim without question. It doesn't matter who you are; if you have a cute child, or a talented child with imagination, creativity and moneymaking potential (such as business smarts; song making ability; literary / oratory ability etc. –you name it so long as your child has moneymaking value to them) they will send for the blood record of your child; agitate the blood cells over and over until a duplicate clone of your child is grown; they will then transfer your child's duplicate clone body from the tank they were grown onto a steel rack; the steel racks contain 5 rows and each row holds a duplicate cloned body; once your child goes to sleep in his or her original body and once they enter REM sleep (90 minutes to 110 minutes after they first fall asleep) the Illuminati are alerted to this fact by a green light above your child's duplicate clone at the cloning centre and they then transfer your child's consciousness to the REM driven duplicate clone; this is the point where the prime evils molest your children; and so your children are being molested in their sleep as REM driven clones and then they suppress the child's memory or implant false memories (see Kim (2013) and Alford (2015) –on how false memories can be implanted) so your children wake up not remembering dreams from the previous night, or if they do remember anything, the memory is false. These depraved acts very much stunt children's development; it causes them to have learning disabilities; unexplained depression; loneliness; suicidal thoughts, and all kinds of side effects.

The cloning centre is a paedophiles paradise and **it must be stopped**! Elizabeth secretly owes a few music companies too (Universal and others) with bands under contract; she forces me under a knife to compose music for them; if I can't she will stab my REM driven clone; burn my REM driven clone and have her pervert thugs smash my REM driven clone body there. As Ehrsson (2013) demonstrated, because consciousness is linked; and consciousness transfer leads the person who has had their consciousness transferred to still perceive reality as "real as real" in the new body; I have the same biological and physiological responses as I would have if these attacks were occurring to my original body. This has caused me to have heart damage and severe debilitating headaches. Because my memories were released to me; I can now also remember all my past REM driven clone memories since age 5; as well as the previous night's REM driven cloned memories (when I have my consciousness transferred to the cloning centre) which causes severe stress to my mental psyche and mental well being.

As unbelievable as the above statement sounds regarding Elizabeth's ownership of music companies, just go and listen to the song "Abadon" performed by Boondox (RainmanJhof 2011) on YouTube. There are two "Abadon" versions; the version you seek can be found in the Reference section uploaded by RainmanJhof (2011); skip to 2:46 min (or listen to the whole song; pay particular attention at 2:46 min) and you will hear loud and clearly, Elizabeth say "Drop the mother f****ing base". That is Elizabeth's voice; you can compare it to any of her Christmas speeches. That is her. Elizabeth said the above as a REM driven clone at the cloning centre. I know. I was there. They got so confident, that nobody would ever find out about their REM driven cloning shenanigans, that they even had me sing on a song I produced: "Where'd You Go" –Fort Minor (Murdok Dubstep Remix) (MurdokDubstep 2010), available on YouTube. Tila Tequila and I sing on this remixed track as REM driven clones (MurdokDubstep 2010); I start to sing from 1:44 min onwards. You can compare the voice you hear in the song with the Vinny Eastwood interview (Vincent Eastwood 2013). That is me, in both cases, and that song was recorded at the cloning centre.

One good person at the cloning centre; Bernie Mac (also as an REM clone version of himself); who was an actor and comedian, tried to stand up for me and speak up for me saying ‘Cloning and torturing people here, it’s not right to do to a human being...’ and “This place is the devil”’... Bernie Mac also tried to have others join him in speaking out against the REM driven clone torture zone while he was having his consciousness transferred to his REM clone driven version. Elizabeth and Philip tortured him for 4 to 5 hours continuously at the cloning centre as an example of their ‘power’, Bernie Mac had an aneurysm (brain bleed) the next day and died. This is one of the aspects about REM driven clone torture: if you torture a REM driven clone continuously for hours on end, you will cause the person to die in their original body from a heart attack or aneurysm (while they sleep); even if they do manage to wake up the next day; because of the constant pain their consciousness has recorded; and because consciousness is linked they will wake up only to have a heart attack or aneurysm in their original bodies the following day. If you apply a constantly electrical current to a REM driven clone, that person will either have a heart attack or die from aneurysm in their original bodies while they sleep.

This form of murder is similar to the CIA’s heart attack gun (Non Mirage Truth Vision 2015); which was declassified in the 1970s and was a weapon used for causing heart attacks in victims, undetectable as the cause of heart attack under autopsy. The difference with REM driven clone death as a form of weaponry is that hardly anyone has heard of death by REM driven clone torture. Moreover, the perpetrator can be thousands of miles away from the victim; the culprit just needs a duplicate REM driven clone of the victim to torture or apply a constant electrical current to; this is exactly what is going on in our world every day. When Elizabeth and Philip don’t like someone and they want to get rid of that person, they clone that person; torture his or her REM driven clone till that person has a heart attack or aneurysm in their original body and dies.

This is what Elizabeth and Duke Phillip did to the indigenous people in Canada. The video reference (inifiniLor 2013) provided above, regarding the genocide of the indigenous people of Canada discusses this very topic (inifiniLor 2013). As Elizabeth and Duke Phillip visited Canada as clones, when they killed the native children; they both had solid alibis of being in England at the time. They then cloned all the witnesses of the Canadian missing children’s case and had them remotely killed by applying constant electric currents to their REM driven clones; one by one and each witness either died of a heart attack or aneurysm. They just dispatched the last witness by also applying a constant electric current to the persons REM driven clone because the missing children’s case was building momentum.

After Bernie Mac’s death everyone is afraid to speak up; they all became afraid of suffering the same fate. All it will take is a few lie detector tests from an independent source; and I will volunteer for them; they cover up for themselves so carefully that lie detector tests from an independent source unconnected to them in anyway will be the only viable method. I know many know non-famous people who have also attended these clone centre gatherings as their REM driven clone duplicate versions and also remember the clone gathering experiences in their original bodies (because they were not memory suppressed) that could be given lie detector tests to prove this. When it comes to the sexual exploitation of a child, lie detectors are admissible in Canadian courts. I need decent people with integrity to speak about this filthy business; to testify about who is involved. There is much to tell about the Illuminati and I am a wealth of information about this topic.

Elizabeth also had Princess Dianna Spencer killed by having someone shine the brightest light known to mankind through the window side on her car; they served away to avoid the light and hit the divider; it was no paparazzi. I first wrote about this in my original disclosure in **2011**; this was corroborated by an article 2 years **AFTER** I had made my original disclosure. See Radar Online (2013) for the corroborating article. Furthermore, "Candle in the wind" performed by Elton John (SadSongs4You 2010) was not Dianna's 'death song'; it was "Bigger than us" performed by "White Lies" (WhiteLiesVevo 2010). The Illuminati released this song, deliberately, 13 years after Dianna's death, again as a show of their 'power'. I did not want to compose this song but Elizabeth and her thugs threatened to stab and burn my REM driven clone repeatedly if I did not. Listen to the lyrics. I'll provide some of the lyrics here for those of you who are feeling too shocked by these revelations:

White Lies Lyrics "Bigger than us"

You took the tunnel route home.
You've never taken that way with me before.
Did you feel a need for change?
Apologies on your fingernails,
Love flickered in the city of lights
Like internet and radio waves.

I don't need your tears
I don't want your love
I've just got to get home

And I feel like I'm breaking up
But I wanted to stay.
Headlights on the hillside
Don't take me this way.
I don't want you to hold me
I want you to pray,
'cause it's bigger than us.

Are you beginning to see why these people must be stopped and stopped immediately? Are you beginning to see why it is no longer okay to 'turn a blind eye' or 'bury your head in the sand', and hope it all goes away? Can you see why I need the good people of this world to keep sharing this disclosure (online); to tell your friends and family until this information reaches the armed forces so that the armed forces can bring these people to their knees. I also do not want innocent civilians to get hurt despite the monstrosities these people cause to humanity on a daily basis. I want the overthrow to progress smoothly so that the good people of the earth can inherit the world, with all technologies and medical advancements which benefit mankind, made available to all. This is why I maintain my composure and resolve, in spite of the evil I see on a daily basis, and you the reader must also maintain your composure and resolve in spite of the evil which I am disclosing, so that we do not inherit a torn down world after we have completely overthrown these malevolent, sinister and corrupt hypocrites!

Furthermore, Diana survived the car crash; and while, either on her way to hospital or in hospital, Dianna was injected with a high concentration of salt, which is near undetectable in an autopsy and she died. Elizabeth did this because Diana was going to marry an Arab named Dodi Fayed. The Illuminati hate Arabs because for the most part, Arabs have not broken their relationship with God. Dianna hated her association with these people; Dianna knew about REM driven, sleep cloning and wanted nothing to do with it; but she was afraid of being killed and so said nothing. I know a lot about this and many more deplorable things the Illuminati have done over the course of 34 years and I will be more than happy to reveal all the evil they have done; as well as all the wonderful discoveries and progresses they have made in medicine, science and technology, which they have withheld from the general public. I am very happy to disclose all, because I known once the world learns of all they have done, we truly will begin to see the end of such evil on this planet. Just make sure you keep spreading this disclosure and it reaches the armed forces soon. Soon, soon, soon!

Help Donald Marshall by Sharing this Disclosure with Your Close Friends and Family

If you are reading this, then you have to help me. I know some of you may be reading this thinking, but I am just an ordinary man or woman Donald, I know you are telling the truth; I can feel it in my heart and my head that what you say is true and that you are telling the truth (which I promise you, with my right hand to God I am) but I'm just an ordinary man / woman Donald; I don't have any power; I don't have any influence; or because of the position I am currently in I can't speak out against these people; and I am just an ordinary man / woman I cannot be expected to take on the world's problems. You are right if you are thinking and feeling this, and I do not expect or want you to take on the world's problems. All I want you to do is share this information with close friends and family initially, because when you do that, at least that is one more person who will know about the true extent of evil in this world. When your close friend or family member also shares my disclosure with another close friend, the disclosure will eventually spread exponentially. It will also be one more person who has your well being in their thoughts because I understand how scary it is to know this. Therefore remember **there is safety in numbers;** and the more people you tell in your social circle, the more you can all look out for each other. So feel free to share my disclosure with close friends and family. Do not talk about my disclosure in public (for now) if it makes you feel uncomfortable. Moreover, if you want to reach more people (other than your initial social circle) the best way, is to do it online: through social media, such as Facebook, Twitter, and YouTube etc. As I've already mentioned, the Illuminati controls all forms of media except the internet.

The Illuminati, being so 'confident' at the time with the evil they wield onto the world, put my face on a Megadeth album cover titled *The World Needs a Hero* (2001) (to mock me) when I was 23. I posed for this album cover at the cloning centre when I was 23, and the album was released with a picture of my face, as a REM driven clone, 3 years later when I was 26. They did this because they were extremely confident that nothing would ever be proven. Google image it, although I am 39 now, you can still tell that it is me. The image shows what I end up looking like after an REM driven clone gathering with a skeleton popping out from my chest. I understand most of you are going to continue reading because although it is "liberating", it is also "paralysing" at the same time (Laurence Mountford) to realise what the truth of the world is, so I will display two images below, but promise yourself, that you will Google search this image in your spare time.

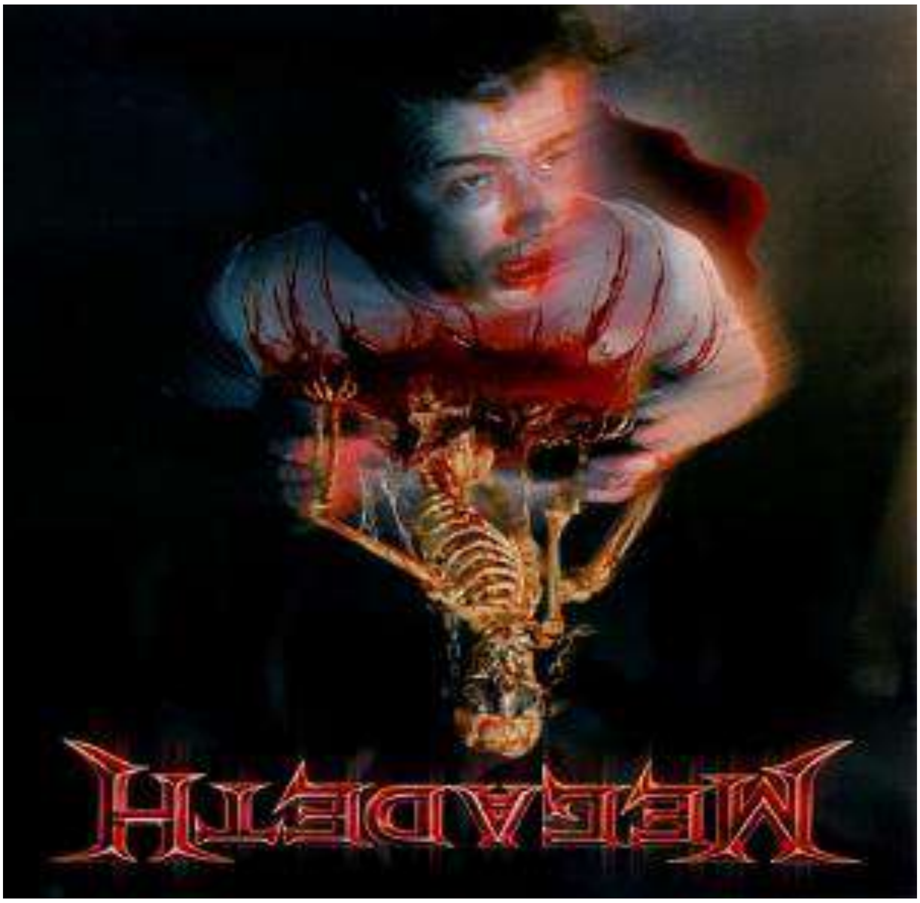

Image 3: Megadeth Album Cover: *The World Needs a Hero* (2001)

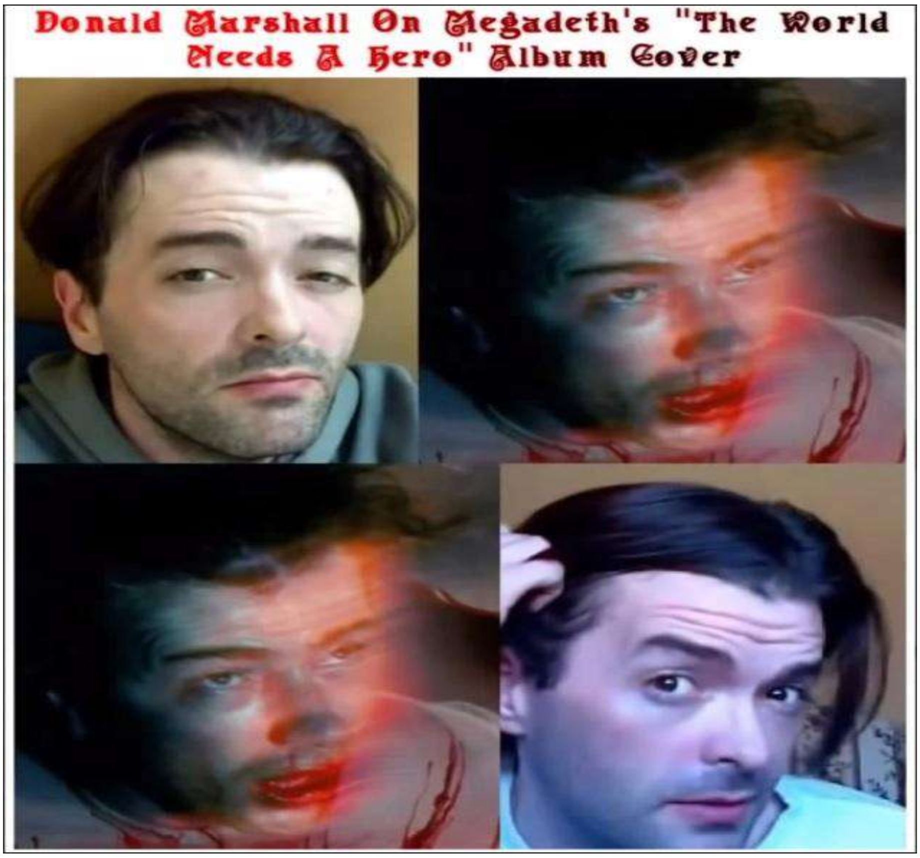
Donald Marshall On Megadeth's "The World Needs A Hero" Album Cover

**Image 4: My original body: top left (age 37) and bottom right (age 37);
compared to my 23 year old REM driven duplicate clone body (top right and bottom left) at the cloning centre.**

You can clearly see from the images that, it is in fact me. In Image 4, you can also compare what my real body (top left and bottom right) looks like compared to my duplicate clone body. Duplicate clones do not come out of the tank looking 100% like the original; it's 99% - there are always some slight variations as some have commented that the eyebrows and nose in the picture, bottom right, is not the same (as the duplicate clone on the album cover). However, to people who make such comments, I say: my eyebrows and nose is the same when you compare it to the pictures of my original body (top left and bottom right). Also, remember that the image on the album cover is a duplicate REM driven clone of me. REM duplicate clones do not come out of the tank looking 100% like the original. There are always slight variations; but the wrinkles on my 37 year old forehead (when I took the picture in my original body; top left and bottom right) is similar to my 23 year old wrinkles, the lips are the same; and the face which you see, is definitely, the same! Yes. REM duplicate cloning is real; going to sleep and waking up as a cloned version of version of you is real and **we must put a stop to it now!**

The Deplorable Nature of the Illuminati

My own family has been turned against me one by one. The Illuminati have told my family members when they were REM driven clones at the cloning centre, things such as I secretly hit or molested their children; poisoned their pets when the pets died of sickness and other such things to make my own family hate me and not want to help me. Even when my family side with me there at the cloning centre, as REM driven clones of themselves; and my family members say 'They do not believe that Donny did all those horrid things' -which they have been told; Elizabeth then says "Are you cooling me a liyah?!" in her disgusting croaking voice made as low as low and evil sounding as she can make it, complete with psychotic and malevolent glare; and as terrified REM driven clones my family say "No, no, of course [what you are saying] it's true", stuttering and afraid as REM clones. My mother Catherine McMahon sold me into this REM driven sex and torture cloning slavery when she remarried a man named Gordon Cohoon; Gordon's whole family is in this scummy secret society, his brothers Tom and Tony Cohoon; his sisters Darlene and Bernadette Cohoon; they fear the new and improved lie detector tests; and all it would take, is for an independent source, far, far removed from any association with the Illuminati to administer these lie detector tests on me **AND** the above named people to prove this unequivocally.

Half way through the making of my original disclosure document (in **2011**), when I went to sleep that evening, the Illuminati transferred my consciousness to my REM driven clone and introduced me to a man (who was also a REM driven clone) named TROY LANDRY. He is an alligator trapper from Louisiana on the TV program named "Swamp People". Troy Landry said "If I send my [original] disclosure out to the public then he would take a power drill to my shin bone at the cloning centre and suck the marrow from my [REM driven clone's] bones." It is similar to spinal tap, and one of the WORST things you can do to a REM driven clone besides burning a REM driven clone. I told Troy that I will be sending out my disclosure to the public, because I have to escape the cloning centre, and so, Troy Landry did just that. It was excruciatingly painful; all the REM driven clones that were sat in the stands of the arena just watched; slack-jawed. Troy Landry is an insatiable child molester and an extra retarded REM driven clone. A side effect of REM driven cloning is that as a REM driven clone version of yourself, your emotions are heightened and you are less smart as you are when you are experiencing the world in your original body. Nevertheless, Troy Landry is extra retarded as a REM driven clone.

In Louisiana, when Troy sees a young boy that he likes, Troy asks of the boy's name and tells the boy "He is an alligator hunter"; Troy then proceeds to shake the boy's hand; Troy then has his paedophile friends at the local cloning centre find the boy's blood records to grow duplicate clones of the boy. Five months later, there are multiple REM driven duplicate clone bodies of the boy grown for Troy Landry to victimise before the crowd of onlookers. All members of the Illuminati do this; it's just a way of life to them; they consider themselves 'the privileged people' in the world power organisation. Almost all the cast on the "Swamp People" television program attended the cloning centre as REM driven duplicate clones of themselves, **when they go to sleep**. After putting a power drill to my shin bone, Troy Landry asked me again, "Will I send my [original] disclosure out?" I said "Yes" to which he replied "I'll do the same to your pelvic bone everyday!" The pain on my pelvic bone was far worse than when Troy put the power drill to my REM driven clone shin bone. I told Troy "I have no choice but to inform the public" – and so Troy put a power drill to my REM driven pelvic bone, everyday... and it is the worst pain ever... you forget your own name; where you are; all you know is pain; and you beg God or anything to save you from it...

Another deplorable thing which the Illuminati did in real life is: The Canadian government were trying to lower the amount of prostitutes on the streets (Elizabeth hates prostitutes) so they had a man named Robert Pickton start killing the women and feeding the women to pigs on his farm. The Illuminati had a camera set up in the upper corner of a room in Pickton's house and recorded Pickton hitting these women over the head with a hammer (a ball-peen hammer). They took the recordings, all 49 of these murders, and they all watch them at the cloning centre. To be seen as 'tough' at the cloning centre is to watch all 49 of the Pickton murders without having it affect your psyche. Elizabeth loves watching the Pickton murders and she says she has a macabre fascination with death. Canadian Prime minster Steven Harper knows all about the conduct of the Pickton murders; he has seen all the recordings and cheers on as the rest of them do. If Mr. Pickton says anything about his involvement with the Illuminati, or the recordings, they will apply a constant electrical current to his REM driven clone until Pickton has a heart attack or aneurysm in his original body.

What Happened to Britney Spears?

The Pickton murders are also the reason Britney Spears 'lost it' and shaved her head bald. She had watched too many Pickton murders as a REM driven clone at the cloning centre and it drove her 'over the edge'. I'll also reference the following songs performed by Britney Spears so that the reader has more corroborating evidence to add to the ever increasing pile. Search for the video "Break the Ice" (BritneySpearsVevo 2009) on YouTube. In the Japanimation video; at the beginning of the video; that is a cloned **cartoon** version of Britney Spears being grown in a big thick thank full of water (1 second to 7 seconds); and at 1:32 min to 1:43 min Britney walks into the cloning centre. This is **EXACTLY** how the cloning tubes are in real life. They are stacked one on top of the other, each filled with salty water. Britney Spears made this video because she said a fantasy of hers is to "Blow up a cloning centre". The Illuminati allowed her to release this video because at the time they figured nobody would ever figure it out. I have provided the images on the next page.

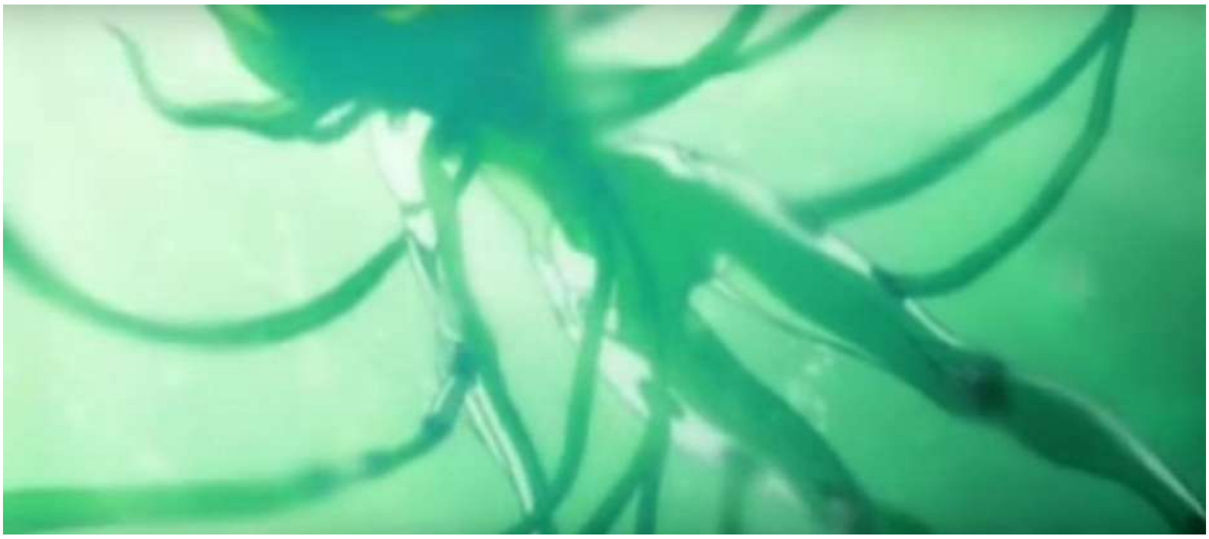

Image 5: Britney Spears is being grown as a duplicate clone in a cloning tube. The depiction is EXACTLY how the duplicate clones are grown in the Cloning Centre... with wires extending from the body; in a big thick tank full of salty water.

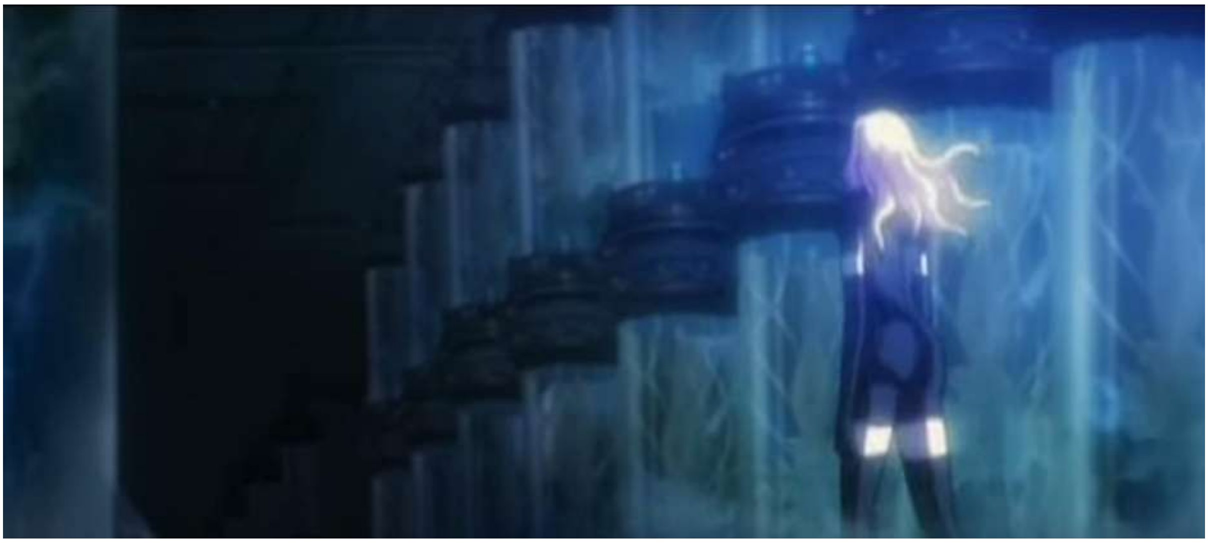

Image 6: Britney Spears walks into the cloning centre. The cloning tubes are stacked one on top of another. This is EXACTLY how these cloning tubes are in real life at the Cloning Centre.

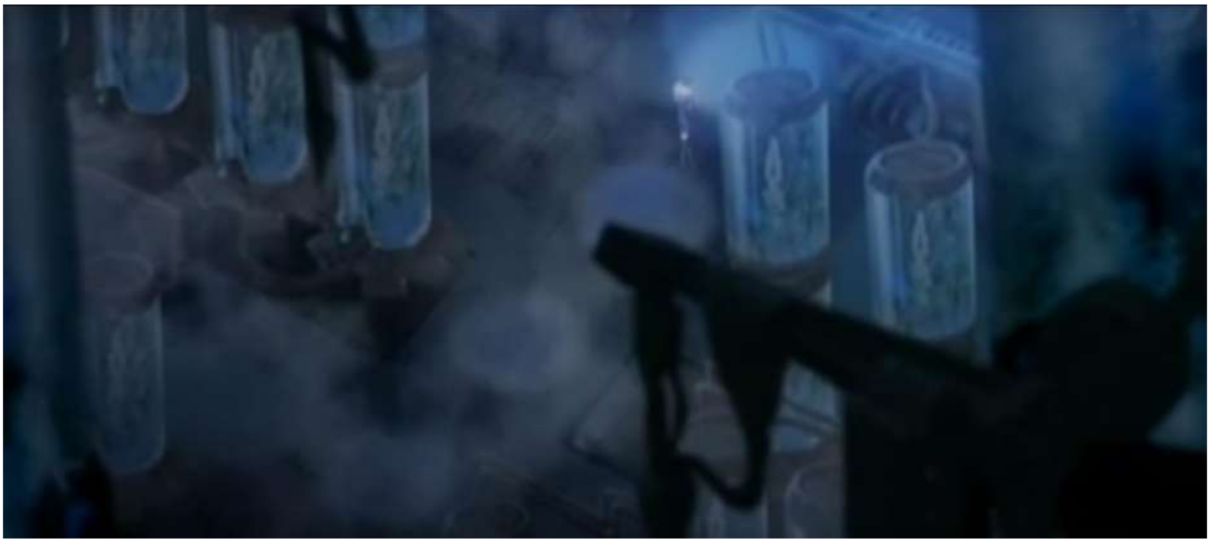

Image 7: The Cloning Centre depicted from an aerial view.

In “Hold it Against Me” (BritneySpearsVevo 2011) performed by Britney Spears; Britney is depicted fighting against her clone at 2:46 min till the end of the video. In “Mona Lisa” (SimpleGirl4ewer 2007) later remade by Britney Spears she changed the words in the song from “gone” to “cloned”. I’ll provide the lyrics to chorus here because they have not changed it on A-Z lyrics etc., but listen to the song and you will clearly hear the following:

Britney Spears Lyrics “Mona Lisa”

She’s the original (yeah, yeah)
She’s unforgettable (yeah yeah)
She wants you to know (yeah)
She’s been cloned…
It’s kind of incredible (yeah yeah)
She’s so unpredictable (yeah yeah)
She wants you to know (yeah)
She’s been cloned,
She’s been cloned,
She’s been cloned…

Yes. Britney Spears and many other celebrities want you to know they have been cloned. The celebrities are depending on you, the ordinary man / woman; the populace, as I am depending on you to spread this message fast and far. In fact Elizabeth has said that “If any public person speaks out, they will die!” Therefore, public figures cannot openly discuss this topic to warn the populace –so they hint. It is **NOW** up to the ordinary man and woman to end this. Furthermore, Elizabeth has a nuclear bomb hooked up at the cloning centre which she can press at any time to wreck havoc on the world unsuspectingly. Thankfully, I managed to convince Elizabeth not to do such a silly thing. I hope you can now understand why everyone is terrified of Elizabeth at the cloning centre, and why this is a **world emergency** and we are depending on you to share and spread my disclosure!

NEVER EVER SELL YOUR SOUL

In This Physical World Selling Your Soul Is Selling Your REM Driven Clone

Furthermore I must discuss an entrapment many up and coming ‘stars’ fall into. You have most likely heard many public figures say “I sold my soul to the devil” or some variation of “I sold my soul”. I am now going to address the aspect of “selling souls” or “selling your soul” which I am sure you have heard many of these celebrities talk about. I’ll address this as coherently as I can because the on the surface selling your soul talk is difficult to contemplate especially if you consider the soul to be a metaphysical, ethereal or mystical part separate from your earth body. Moreover, we live in a physical world, so selling a metaphysical aspect of you sounds incomprehensible (at first). For the most part we have little understanding of what the soul is. However, we are told that the soul is a part of us which never dies and it is eternal. It is all that we are, ever were, and ever will be. Now our faiths and beliefs come into question, because if you are an atheist: there is no God, selling your soul is a bunch of crazy nonsense.

From a practical perspective it also seems difficult to fathom, because if you go on what you have been told, you are selling an intangible part of yourself to be used by someone else for whatever they wish; and so if you get into the headspace of viewing your soul as something separate from you or that you are in control of your mind and your body, then again, 'selling an intangible aspect of you' again sounds like nonsense because right now nobody controls you; you are controlling you; so what are you really selling? Thin air?

Many up and coming 'stars' think this way when they are confronted with "selling their soul" aspect. The up and coming 'stars' are blinded by the prospect of being super famous and super rich; and in the moment they are confronted with "selling their soul" they believe these Illuminati people, are a group of over religious zealots and that they, -the up and coming 'stars' are taking advantage of the Illuminati because the 'soon to be 'stars'' do not believe (in that moment) it is possible to sell your soul, and only consider being super rich and super famous; so the 'stars' sign their 'soul' over.

At this point, it is usually too late for the up and coming 'star'; because what they have actually sold, in this physical world, is the use of their REM driven clone to the Illuminati. They have sold their REM driven clone to be used for REM driven sex by perverse old men and women whenever they enter REM sleep.

These celebrities are not memory suppressed either; so whenever they enter REM sleep, their consciousness is transferred to their REM driven clone, to be used for whatever the Illuminati wish: usually sex as REM driven clones with dirty and perverse old men in return for a promotion in the entertainment industry; such as movie role, advertisements, concert gigs etc. It is a very, very, sick business, and once these up and coming 'stars' get into the business they are trapped. Now you can begin to understand why so many celebrities, although on the surface it appears they have it all, are depressed; and do random and crazy things. It is a cry for help. Their illogical behaviours, (not all) for the most part, can be attributed to REM driven cloning. Furthermore, the Illuminati consider 'selling your soul' (the use of your REM clone) **a very serious business.** They do not joke about this; and there are NO RETURNS. Once you have sold the use of your REM driven clone for promotions etc., that's it. If you ever make a fuss and want your 'soul' back, in other words: you no longer wish to have your REM driven clone to be used by dirty old men for sex; the Illuminati will kill you by applying a constant electric current to your REM driven clone until you have a heart attack or aneurysm in your original body and die. It is as simple as that.

For example, Whitney Houston wanted nothing to do with the cloning centres, she said she would keep quiet; she never wants to be in the spotlight; she never wants to do radio or television. Let bygones be bygones, I'll go my way and you can go your way. The Illuminati were extremely angry! They had bank-rolled her, and now she didn't want to be their 'friend' or ever associate with them again. She was sacrificed; she was killed with a constant electric current to her REM clone. She had sold her soul, in the beginning of her career; in other words, the "use of her REM clone". Earlier in Whitney Houston's career, she had signed that they can use her REM driven clone however they wish.

I have also seen people say "No" –when confronted with 'selling their souls'. These people are then memory suppressed, and will not remember the proposal of their REM driven experience. The people who also refuse to sell the use of their REM driven clone ('sell their souls') are blacklisted secretly, and they will never receive a high paying movie role, concert gig, advertisements etc. If you have ever wondered why so many talented people have not 'made it' –they refused to sell their souls, in other words, the use of their REM driven clones in exchange for promotions. This is how the entertainment industry really works. Another aspect of selling your soul, in other words, **selling your REM driven clone** is that you are also agreeing to sell your children, and your children's children etc.; your whole lineage to the Illuminati; for the Illuminati to use your unborn children for paedophilia, torture, sex sports and other grotesque and ungodly things when they sleep. It is a nasty business. **NEVER EVER** sell your soul.

The lesson is very important. NEVER EVER, 'sell your soul'; NEVER, EVER! You will experience the worst nightmare situation ever, and in this physical world, that is the use of your REM driven clone to be used for sex sports. You'll never have a decent night's sleep ever again, and you will always retain the full memory of your REM driven clone experiences; as well as, you will be selling your unborn children into this dirty business. Even when confronted with this as a joke: never sell your soul. Never Ever! The physical aspect you are confronted with when asked to 'sell your soul' is: "YOU", and everything which makes you, "you"; your ideas, your creativity, your privacy, your biology; everything which makes you, you. Never sell your soul, which in this life is currently the use of your REM driven clone! Never, ever! I hope that is well received.

Fake Alien Abductions conducted through REM Driven Cloning Technology

I must add they grow duplicate REM driven clones of people from all walks of life and then they transfer the persons consciousness to their REM driven clone version at the cloning centre while the person's REM driven duplicate clone is chained down to stainless steel corpse tables; or they pre-inject the REM driven duplicate clones with drugs before transferring the consciousness so that once the consciousness has been transferred, the REM driven clone cannot move. This is the point when they usually send people into the room dressed as aliens. It is a fake alien abduction. It's just these Illuminati people dressed in Hollywood quality accessories and they make the REM driven duplicate clones which have either been chained to the steel corpse tables or pre-injected with drugs, believe they are having an 'alien abduction'. Members of the Illuminati even dye chicken skin grey, and stretch it over a mask for realism and perform "experiments" on the immobile REM driven clones; the perverts anally probe them; they rape their limb bodies and do not suppress the victims memories; so these victims wake up thinking they have been abducted by aliens but in reality, they had their consciousness transferred to their duplicate clone versions when they entered REM sleep and they were violated. The victims do not know where to turn and naturally, they are embarrassed. Some victims even try and do install cameras all around their bedrooms in order to videotape themselves while they sleep; to prove they have been 'taken'; but to no avail they have not been taken; they have been cloned and had their consciousness transferred to their REM driven duplicate copies once they enter REM sleep by the most disgusting perverts in the world. The scum even videotape these 'abductions' to watch later as sick demented porn. These fake abductions are conducted by the Illuminati more often than you can imagine. They just keep doing it to random people over and over and over...

I am THE Spy for Humanity against the Corruption of the Illuminati

Over the years the Illuminati have offered me: a priceless katana (samurai sword); a rigged jackpot winning lottery ticket; REM driven clone slaves to keep me quiet, any of the prettiest women I've ever seen; any girl from high school; Vladimir Putin's daughter; even children; I do the right thing and spit in their REM driven clone faces; jam my fingers into their eyes but they have the technology to turn off the pain sensors of a REM driven clone, and therefore they have their pain receptors turned off as REM clones, while my pain receptors are fully on (sometimes even beyond the threshold of what is normal pain for a human). See the online articles: Science Daily (2014) and RT (2014) which discuss how pain receptors can be turned on and off in humans. Not much will hurt their REM driven clones other than getting dirt or vomit in a wound or bleeding too much. However, they have perfected the science and technology of REM driven cloning and therefore it only takes time and approximately $30 to grow new duplicate clone bodies. I started to smear excrement in their faces as a REM driven clone at the cloning centre in an attempt to deter them; I have **never** handled faeces in my original body, but I do this as a REM driven clone at the cloning centre, and even that does not deter them from their depraved ways. I composed a song which was later performed by "One Republic" referring to it called "All the Right Moves" ("All the right friends in all the right places... All the right moves in all the right faces"). I've made so many songs with MK Ultra technology (Rense 2001; Jim Cristea 2009; Berekley News 2011; Mind-Computer 2012 nature video 2013) and Mind-voice technology (New Scientist 2014; Prigg 2014) it is ridiculous. You see it is very easy to make music using Mind-voice technology; Flo-Rida used Mind-voice technology to produce the song "Whistle" when I first made my original disclosure (in 2011). The song is a taunt at me about my "whistle blowing disclosure" i.e. 'Can Donald blow the whistle? Etc. etc.' The thing is, anyone with a talent for music can make music much easier with the technologies which they have (MK ultra technology; Mind-voice technology); however, the Illuminati made more money from the music I composed, and so over the years, they just kept using me, over and over and over again, for rock, pop, rap, country etc. There are so many people involved in this REM driven cloning business, it is staggering; the organisation is vast. The only things the Illuminati fear are nuclear war; the new and improved lie detector tests and the general populace spreading my DISCLOSURE DOCUMENTS.

For those of you thinking Queen Elizabeth II looks like a kindly little old lady and you still cannot believe that all this is true, although I have provided plenty of evidence to the contrary as well as evidence of her saying "Drop the mother f***ing base" as a REM driven clone on a song track (RainmanJhof 2011); you could not be more wrong. She is the worst human being I have ever seen or heard of. It is so sad to see these REM driven women and children having their consciousness transferred to their REM driven clone duplicates and brought to the centre of the dirt rink, sitting there naked, afraid and crying and raped and beaten for sport, for the rich and famous. REM driven cloned women and children display the most range of emotions, and it is the saddest thing to see there. Therefore, of course I cannot bring myself to be a part of this despite any riches which the Illuminati offer me.

Over the 31 years when I was having my consciousness transferred to my REM driven clone at the cloning centre; Elizabeth, Vladimir and many, many Illuminati members and cloning centre attendees discussed many topics with me. These Illuminati people, such as Elizabeth and Vladimir, as well as many others confided in me because I had composed songs since I was 5 years old.

For readers struggling with the idea that a 5 year old can compose such wonderful songs at such a young age I will remind you that, it is easier to compose songs with technologies such as MK Ultra and Mind-voice technology. One is only limited by their imagination and creativity. Moreover, research talented children. I won't provide references of talented children here because I want to protect children from the Illuminati, and not have children exploited like I was; and for the readers who are parents; I am sure you know how talented, creative, and imaginative your children are.

Accordingly, since the age of 5, Elizabeth, Vladimir and many, others confided me in me because they thought I was special, and wanted me to hang with them; some of these Illuminati people, I had composed their favourite songs; I was also memory suppressed for many years so they thought it was harmless to tell me everything that exists or is possible. They were trying to impress me and amaze me, and therefore they showed me plenty of their technologies and what they had. They wanted me to me a willing, full Illuminati member; so over the years I feigned interest and pretended to like these Illuminati people, so that I could learn what they do and how they do it. I was looking for a way to block the consciousness transfer to a clone during REM sleep, or hoping they would reveal this part to me; but they did not, and I did not find a method to block the consciousness transfer. However, I bided my time, until I knew just about everything they do; so that I could reveal to the world everything they do; to give humanity a fighting chance against everything they do and as a consequence, I became a spy for humanity against the corruption of the Illuminati.

The Threat to Humanity when the H.A.A.R.P Grid is Complete

One of the most important topics which they discussed with me that I must add to this disclosure is HAARP technology and their plans for HAARP technology. I have discussed the basics of HAARP above. HAARP radio transmitters have to be placed at certain points across earth to complete a 'HAARP grid'. Once all the installations are completed over the earth HAARP will have more functions as a grid than the basics described. The Illuminati have told me, there is a possibility the HAARP grid will be capable of time travel, although there is a 50/50 chance using HAARP for time travel could destroy the earth. They also told me that once the HAARP installations are completed, the HAARP grid will be able to bend space and make the distance between, for example, Earth and Pluto approximately 1000 miles for a few seconds, and then the distance will go back to normal.

The Illuminati also said, once the HAARP grid is completed, they would be able to mind control people like never before without the use of RFID microchips or MK Ultra. Life on earth will be much worse once all the HAARP installations are up and working together in unison. The grid has to be complete though. HAARP cannot achieve time travel manipulation and mind control if it is just located in one or a few locations around the earth. There must be a certain number of HAARP installations around the world, in certain locations around the world; and the HAARP radio transmitters must be working in unison in order to achieve time travel manipulation and mind control of the populace of earth.

This is what they told me as REM driven clones at the cloning centre. This disclosure could be false, or a half truth. I do not know. However, it is very important that I add this to the disclosure because the Illuminati are working twice as fast to complete the HAARP grid to achieve their aim of mind control over the entire population since I the time I sent out my first disclosure (in **2011**). Whether or not they will achieve their aim of total mind control once all the HAARP installations are completed is NOT something we can afford to debate. They also told me that once the HAARP installations are completed; they will have the capability to mind control anyone remotely like never before, at the push of a button. HAARP will also be capable of setting back time; which the Illuminati will use if they are ever in danger of being exposed. However, the Illuminati will remember the previous time, and the general populace will NOT remember the previous time, so that the Illuminati can correct the mistake(s) which led to their downfall. **Once the HAARP grid is completed, humanity will be slaves forever.** This is why we cannot afford to debate. Furthermore, from the book "Angels don't play this HAARP" (Begich & Manning 1997, p. 8), it is estimated that the HAARP grid will be completed by approximately 2017; others estimate by the year 2020. The important thing to note is that humanity does not have long, until the HAARP grid is completed.

I must also add the positive side of HAARP, because as I have said, there is nothing inherently evil about technology. Technology is just a tool; it depends on the person using it. HAARP used correctly, and in the hands of honest men and women, will tame the weather with no adverse effects. Therefore humanity is on a timeline and a deadline to bring the Illuminati to an end. **HAARP is one of the main reasons, this disclosure must spread, and spread fast.**

See below for possible HAARP locations around the earth (Rense 2011). The HAARP sites are located globally where the richest mineral belts are located. Sub-surface mineral exploration has been done from satellites by radio tomography that is 100% accurate.

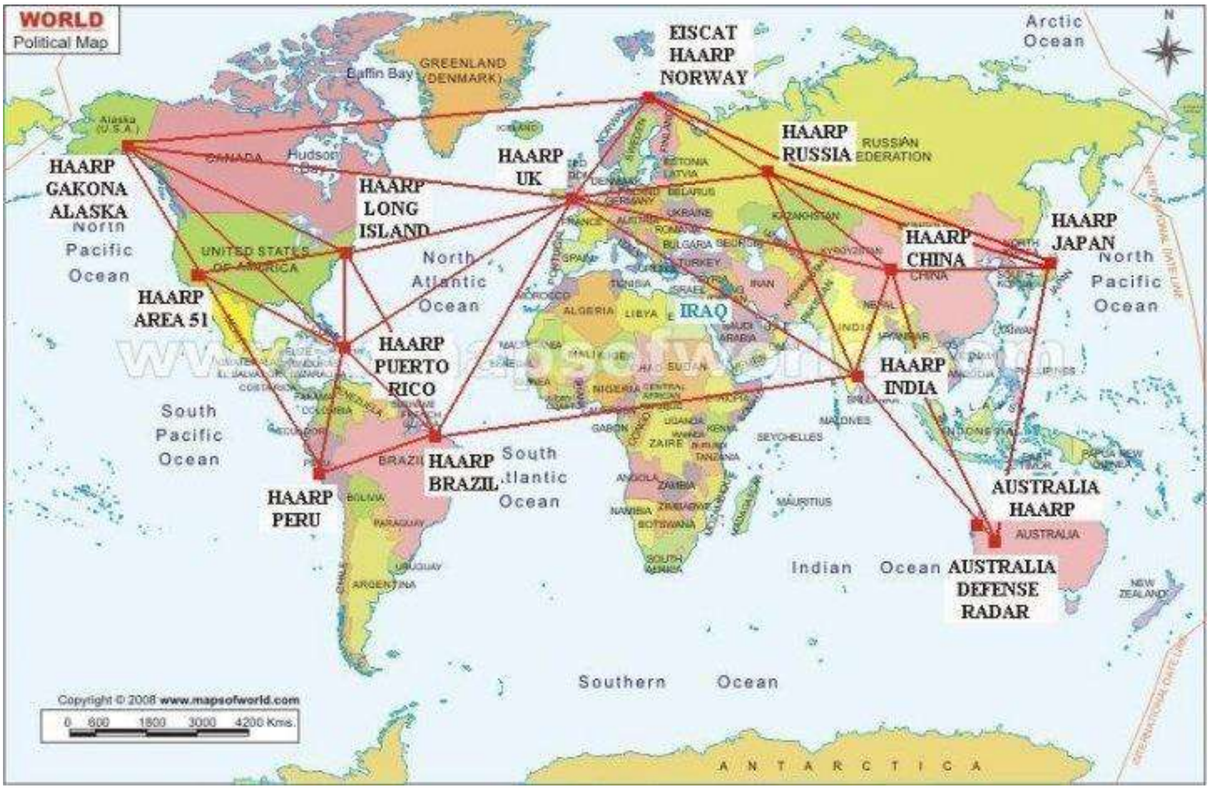
WORLD
Political Map
EISCAT
HAARP
NORWAY
HAARP
RUSSIA
HAARP
GAKONA
ALASKA
HAARP
LONG
ISLAND
HAARP
UK
HAARP
CHINA
HAARP
JAPAN
HAARP
AREA 51
HAARP
PUERTO
RICO
IRAQ
HAARP
INDIA
HAARP
BRAZIL
HAARP
PERU
AUSTRALIA
HAARP
AUSTRALIA
DEFENSE
RADAR
Arctic Ocean
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean
ANTARCTICA
Copyright © 2008 www.mapsofworld.com
0 600 1800 3000 4200 Kms.

Image 8: Possible H.A.A.R.P. Locations. Source: Rense (2011)

REM Driven Cloning is the Most Terrible Nightmare Situation Ever

Furthermore, in terms of REM driven cloning; Fefe Dobson (singer) and Kurt Russell (actor) told me during my original disclosure (**2011**) specifically, to include in my disclosure that they do not like the cloning centre and they didn't torture me. Then they said not to mention them, but I will mention them again. Mila Kunis (actress) from "That 70s Show" stabbed in my REM driven clone body multiple times when I made my original disclosure and I was rendered immobile for saying that "She is a slimy scumbag for hanging with these people" and for also saying "She has enormous eyeballs and looks like a lemur". Mila Kunis (as a REM driven clone at the cloning centre) then begged me not to mention her (in my original disclosure, but I'll mention her again too) and then Mila Kunis said "She doesn't want to go to the clone zone anyway" and the Illuminati then deactivated her consciousness from her REM driven clone body, and she left. Her REM driven clone body went limp at the point where her consciousness was deactivated from it; her REM driven clone body looked 'dead', and she was gone. Some of the privileged people at the cloning centre are allowed **NOT** to have their consciousness transferred to their REM driven, clone duplicate versions (or they attend the cloning centre willingly when they want), but not me. I am an imprisoned slave in the worst nightmare situation. I told Nicole Leone (Madonna) also a REM driven clone at the cloning centre, that I was going to tell everyone that she coerced me to compose songs for her throughout my life (as REM driven clones). Nicole Leone (Madonna) told me VERY specifically to say in my disclosure here "She is not afraid!" and "No one will believe me and nothing will come of this". I beg you to help me prove her wrong.

Moreover, having your consciousness transferred to a REM driven clone alternative inhibits the person's ability to dream for that evening. Although your original body is immobile and asleep in your bed, when they transfer your consciousness you have actually 'woken up' as a REM driven clone version of you at a physical location here on earth; therefore you walk and talk as a clone version of you. You are not dreaming; you are just a REM driven clone for the evening. Furthermore, when the Illuminati restore your memories (as they have with me) or do not suppress the memory of your REM driven cloning experience for the evening (such as with the fake alien abductions), you will remember the experience as 'clear as daylight on a summer's day'; the experience is 'clear as a bell' as life is in your original body.

It is unlike dreaming, where dreams are fuzzy, incoherent and with random conjecture. The Illuminati know this as a fact, and therefore pre-inject the REM driven clones of random unsuspecting civilians for the evening with drugs before they transfer the person's consciousness. They tell the person who has had their consciousness transferred to their REM driven clone all sorts of nonsense: such as they are in the 5th Dimension; the astral plane; a singularity; the spiritual realm; Valhalla; quantum hopping etc. and all sorts of fallacies so that the person will not have a coherent picture of what has happened or what is happening. The truth is that person has had their consciousness transferred to their REM driven duplicate clone when they went to sleep and entered REM sleep; their original body still remains asleep in their bed; their consciousness however is now transferred to a clone duplicate alternative, and the person is 'awake', walking and talking as a cloned version of themselves in a physical location. The Illuminati allow some people to retain some of these REM driven clone experiences to affect their mental psyche.

The physical location where the populace will be able to verify these REM driven clones is somewhere within 5/6 hours drive, in a radius of the Robert Pickton farm, Port Coquitlam, British Columbia, in Canada, somewhere at a remote nature reserve. This is the cloning centre all the rich and famous attend because it is above ground and the rich and famous don't like the 'hospital smell' of underground cloning centres. The above ground cloning centre located in Canada is the cloning centre I have my consciousness transferred to when I sleep. It is guarded by military personnel. There are many cloning centres all over the world; as Phil Schneider (1995; 1996) disclosed as far back as the mid 1990s when he spoke: there were 1477 Deep Underground Military Bases (DUMBs) worldwide and 131 active DUMBs in the United States alone. Each DUMB costs on average 17 -19 billion U.S. dollars; paid for by the taxpayer; and it takes approximately a year-and-a-half to 2 years to build DUMBs with sophisticated methods. DUMBs contain cloning floors, and have an entire floor dedicated to human cloning, where duplicate clones are grown of unsuspecting civilians (remember they just need your blood / DNA); and unsuspecting civilians have their consciousness transferred to their REM driven clones at a physical location, which is usually these Deep Underground Military Bases when they sleep; unsuspecting civilians are memory suppressed, also to be used for gruesome torture and sex sports; as well as the plagiarism of talented individuals. Schneider gave the above statistics in the mid 1990s; chances are there are more DUMBS and cloning centres worldwide today.

Are you beginning to understand why this is the worst nightmare situation ever? Can you understand how: you, your friends, your family, your brothers, your sisters, aunties, uncles, children, work colleagues and anyone else you have contact with in your network, may be involved in this REM driven cloning business and will not even be fully aware of the situation or remember any experiences which they may have had once they wake up from sleep and they are back in their original bodies? I remember all of the REM driven clone experiences as clear as daylight because my memories are fully restored. Can you understand why the Illuminati must be stopped and stopped immediately? This REM driven clone business is vast, and it affects everyone, not just me.

This is why I mention that I am fortunate or unfortunate to remember all these experiences clearly depending on how you view it. I am unfortunate in the sense that I can recall all the depraved things these people do as REM driven cloned duplicates as clear as daylight and it affects my mental psyche. However, the fortunate part is: because I can recall all these experiences clearly, I can inform the world about the evil committed against humanity in a clear and logical way for the average reader to understand and verify, so that the populace can bring these people to justice for the crimes they have perpetrated against humanity and continue to inflict on humanity. For anyone who has had their intellectual property stolen by the Illuminati as a consequence of the technologies they have used on you unsuspectingly, you will be able to sue and regain monetary reward for the theft of intellectual property as well as invasion of your privacy.

I do not want to get too far ahead, because at this present moment, it is **very important** (I cannot say it enough) that this disclosure spreads (this is the stage we are currently at); it reaches the armed forces; to enable the armed forces to realise the crimes committed against humanity because of these people (the armed forces are kept 'out of the loop'; they just 'follow orders'); the armed forces will have to perform a military coup and bring this people to their knees. This is the point where the ordinary man / woman can have his REM driven cloning experiences restored, (that is if you want to) and sue / rightfully reclaim damages and legal reward for any crimes committed against him or her.

The Meek Shall Inherit the Earth

As you can see if everything goes smoothly; the meek shall inherit the earth. This is why you must maintain your composure and resolve, as I do every single day. You must not riot or cause chaos, damage property or undertake any act which will cause public dissension. The good people of this earth are about to inherit the earth; under no circumstance must you destroy anything. At the end of this currently dire and complicated situation; humanity will inherit the cure for HIV/AIDS, cancer, Alzheimer's, dementia, you name it –many debilitating diseases known to man; and other such wonderful technologies like hypersonic trains which run on magnets and are capable of doing Mach 5 (Mach 5 is the equivalent of 3806 miles per hour or 6125 kilometres per hour; Schneider (1995; 1996) also mentions the availability of Mach 5 trains in his disclosure, back in the 1990s). These trains can be used to deliver foods which are discarded in More Economically Developed Countries (MEDCs) to Less Economically Developed Countries (LEDCs) because these trains can travel thousands of miles within minutes. Yes. The Illuminati have all these wonderful technologies and more, but withhold it, and in turn withhold the progress of humanity. Yes. Plenty of man's problems can be solved when we address the root causes of our problems, and one of the root causes of humanity's problems is: REM driven human cloning which is currently in the hands of evil men. Even REM driven cloning can be used for the benefit of mankind when this tool is in the hands of righteous men. It can be used learn new skills in your sleep: such as gain extra qualifications; learn how to speak a new language; play instruments etc. –so long as memories are not suppressed, you will be able to perform the skills you learned as a REM clone in your original body when you wake up. As you see there is nothing wrong with most technologies; it all depends on the person / people using it. Nevertheless, although they may be beneficial uses to REM driven cloning not yet realised, we must be shut down REM cloning before we can even consider how REM driven cloning may benefit mankind, because at this present moment all that REM driven human cloning is used for is to commit evil, particularly against unsuspecting civilians.

It is OK to feel angry about all these crimes committed against humanity daily. But do not let these evil crimes rule you, in fact, let it motivate you; let it motivate you to the point where you will stand up and speak out to ensure that these crimes end. I think about the children and the future of the world every day; that is what motivates me and I view my situation as a God-given mission to bring these people to justice. I view things this way to maintain my sanity and composure. This is what keeps me going despite the extreme evil I experience, and it helps me to maintain my resolve and composure. The fact that I do not want a single child to grow up in a world, where they are being messed with in their sleep is what keeps me going despite this extreme evil. Some children are even kidnapped and brought to the cloning centre in their original bodies, again to be messed with. This is the root cause of why so many children go missing across the world. I'm sure any rational thinking human being feels the same way I do, and does not want their children, or any child to inherit a sick world governed by sick individuals.

I am a baptised Roman Catholic and a God fearing man. I believe there is a Creator of the universe but the Illuminati tell me there is no God, only science and technology. The Illuminati are also responsible for the death of the young beauty pageant girl Jean-Benet Ramsay; her parents are in this REM driven cloning business. Casey Anthony's daughter; as well as, Casey Anthony (before she died) AND Casey Anthony's parents also attend the cloning centre as REM driven clones. Many others; a man had his wife killed and this was mentioned on the news; it was an Asian woman whose rich husband had killed her; the husband said the wife was responsible for their son's death. Elizabeth's response was that "Rich people in this organisation (The Illuminati) do not go to jail, they are covered for."

Stephen Spielberg and George Lucas are Illuminati members and they willing attend the cloning centre as REM driven clones when they sleep. George Lucas made other Illuminati members light my REM driven clone body on fire for the end of his movie *Star Wars Episode III: Revenge of the Sith* (2005).George Lucas said (as a REM driven clone) he wanted Hayden Christensen "To scream realistically". Hayden Christensen (also a REM driven clone at the cloning centre) watched me, and listened to my screams and groans of pain and copied the sounds coming from my burning REM driven clone exactly. Hayden Christensen knows he is a REM driven clone when he goes to sleep and Hayden Christensen knows his consciousness is transferred when he reaches REM sleep. Hayden Christensen knows all about REM driven human cloning, and he is a "privileged" Illuminati member. Natalie Portman also knows all there is to know about REM driven human cloning; as does approximately a quarter of the "Star Wars" cast. Many directors have attended the cloning centre as REM driven clones too, and have used me in a similar role playing situation; they have caused me to have excruciating pain as a REM driven clone, to see what kind of squeak or screech I will make, as if I were less than a dog; and when I am a bloody mess on the floor as a REM driven clone; that is when they usually crawl on my broken REM driven clone body and sodomize me saying something like "They love me and they cannot control themselves because I made all their favourite songs and I'm so "special"".

My ribs in my REM driven clone body will be broken, I will have suffered internal bleeding and will be crying or screaming, if I am able, but they just continue to sodomize me, and videotape such a depraved act of wickedness so that they can view it again like "evil pornography".

The REM driven cloned children, who have their consciousness transferred while they sleep at home in bed, in their original bodies, need a familiar face to talk to as REM clones or all they do is scream and cry at the cloning centre; that is where Joy Geizer comes in. Joy Geizer is married to my half brother from my father's first marriage; Joy Geizer is a REM driven clone girl guide leader, and when the Illuminati clone young girls, Joy Geizer speaks to these REM driven cloned girls; Joy Geizer keeps the REM driven cloned girls calm and "pimps" them out for free, knowing that these young girls will have their memories suppressed, and therefore they will not remember the experience and will not talk about it when they wake up. All the Geizers are in this REM driven, human cloning business; there are many people who remember these REM driven cloning experiences in their 'awake state' and in their original bodies who could be polygraph tested by independent polygraph testers. The police, and polygraph testers in my City of Halifax, Nova Scotia are compromised (they attend cloning centres as REM driven clones too); they cover up for child molesters and therefore cannot be trusted to fairly administer a polygraph test; as I said, commissionaires and C.S.I.S are heavily involved. All it would take is a few polygraph tests to prove these things unequivocally, because for those of you who have 'eyes', you can already see that REM driven cloning is real, and that this is actually the way of the world. I will take these polygraph tests publicly too, and demand that my mother and step father submit to them. They have told me, they **wouldn't** even attempt to lie on a polygraph.

Human cloning, particularly REM driven human cloning, is one of the absolute truths of this world; hidden from the world for over 70 years. My life and the freedoms of many, many people, including the reader (you) depend on this disclosure. The Illuminati say that the world finding out about REM driven human cloning will set the stage for the end of the world; people panicking and destroying property because of the evil these people have committed, but as I've said: prove them wrong, because we are going to inherit the earth, so there is no need to riot or cause public dissension. There will be no end of this world; just the end of the crazy Illuminati people. They always like to think and behave so negatively, don't they? *Sigh*. Furthermore, the whole talk about 2012; December 21st 2012 and the end of the world Mayan Prophecy; that was really about the world finding out REM driven human cloning, and the sick and sadistic nature it has been used for by these sick and sadistic people. Under no circumstance must you riot; you **must** prove you are better than these savage barbarians. I'll say it again, **because it is important**: we are going to inherit the earth, and therefore the smart and right thing to do, is not to destroy anything, especially the earth which you shall inherit.

Furthermore, I hope you are beginning to understand that the good people of this world must bring these people to justice and stop their wicked ways. I sincerely wrote this disclosure so that the good people of earth would be empowered to put a stop to their wicked ways. I wrote this disclosure to empower the good people of earth against tyranny and not so that the world would end. Therefore you must understand that the downfall of these Luciferians (Illuminati) must progress smoothly, because unless you have been to the cloning centre as a REM driven clone and retained the full memories of your experience, it is difficult to understand the level of depravity and sub human cruelty; it is beyond anything that has ever been heard of, and there is much more to tell.

Ignore this disclosure and you will condemn me to a horrible eventual death and you will encourage them that 'they are all 'powerful''. This is not an exaggeration. There is no need to exaggerate even President Barrack Obama is involved in the REM driven cloning business and attends the cloning centre as a REM driven clone version of himself, when he goes to sleep. He and his wife Michelle Obama attend these REM driven clone gatherings. Barrack Obama has even told me as a REM driven clone, at the cloning centre "Donny, we're all powerful. You are a slave and the people here [as REM driven clones at the cloning centre] won't speak up for you for fear of torture or death. Now make us a new song or we'll gut you like fish and leave you to writhe in agony!"

Be Wary of: Shills; Trolls; Disinformation Agents and Double Agents

There are many, many disinformation agents whose aim it is, is to denounce or debunk me. Be very wary of them. They DO NOT, and cannot debunk me, because after all, everything I have disclosed is truth. Although the truth concerns highly advanced concealed technologies, so it is a bit difficult to fathom at first, but keep reading, and you will realise that all which I have disclosed is true. Be very wary of the disinformation agents.

The most common thing these people say is that "Donald is a paranoid schizophrenic" –yet, ask them to detail the ins and outs of what a paranoid schizophrenic is and what specifically qualifies Donald to be a 'paranoid schizophrenic'; there is a high probability they will not be able to explain this to you, logically, sequentially and methodologically. Their main goal is just to make it as difficult for newcomers to understand what the real truth of this world is and as I have disclosed; and one of the root causes of man's problems is REM driven cloning.

Some disinformation agents are paid by the Illuminati, to behave this way, and are all too happy to echo empty statements (without any substantial evidence for their claims) because it provides them with a pay check. I will provide the following reference for the reader which discusses the signs and symptoms of a paranoid schizophrenic; all for the express intent for the reader to further understand paranoid schizophrenia. You can read more about paranoid schizophrenia on Medical News Today (2015). The link is provided in the reference section. You are welcome to read my Facebook wall and within days, you will realise I converse with others quite naturally. You can also ask individuals who have spoken to me through Skype or met me; but if this is not a viable option for you; all you have to do is read my Facebook wall and within days, you'll see that whoever makes such harsh and untrue remarks on my character is in fact a disinformation agent.

Furthermore, remember we live in a 3D **physical** world; and there is always a physical aspect to how something manifests. Therefore, when you hear others discussing concepts such as 'Tila Tequila demonstrated an 'energy ball' right on camera; Why didn't Donald discuss that?!' Please understand that these people are trying to deceive you with misinformation. There are also professional disinformation agents to be wary of. I do not perceive anything extra-ordinary in the world when I am awake in my original body beyond the five senses. I don't see orbs, auras or anything of the like. All I will ever discuss are practical things which have been unequivocally proven to me as fact; because in the situation the earth is currently in, I cannot afford to speculate.

Contact Information

My Facebook is: https://www.facebook.com/donald.marshall.148
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

I have a public wall on Facebook which starts from March 2012. In other words, all posts are made public since March 2012 and you do not have to join my friends list or subscribe me to see what I post (although you will have to have a Facebook account). I suggest that newcomers start reading from March 2012, and be patient, and read everything. You can read everything and learn the real truth of the world free of charge. I don't wish to write a book, I am not looking to turn a profit from this; I want to crush these guys and shut down the cloning centres!

It also worthwhile to add that: I am **not** asking for donations. I do not want **ANY** donations. Ever! These disclosures are far more important than any donations. This is **NOT** my job. I have a job. I am an independent contractor; carpentry is my trade and I earn a living this way. If you sincerely want to help; the best way to help is to spread and share my disclosures. That is all I ever ask for, so that eventually, the world knows about REM driven human cloning, and the armed forces can bring these people to their knees. **That is all I want; spread, spread, and spread this disclosure.** I will **never ever** ask for donations. Ever! Please keep this in mind, and anyone who asks for donations in my name or on my behalf should not be trusted. I, Donald Marshall, will **NEVER EVER** ask for any donations. I hope that is well received.

You can also view Proboards which has all my Facebook posts and have been archived by Celine O'Carroll and Astral 7ight by visiting:

http://donaldmarshall.proboards.com/
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

There is a search function on Proboards, and you can use this to search for and read all the disclosures I have made regarding REM driven clones, the people involved and more. I understand that it is human nature to want to know, which people have been to the cloning centre as REM driven clones; therefore, use the search function to read about any public figure which I have already covered that you have an inkling about. You can also post anonymously on Proboards and Celine and other Administrators will transfer your question onto Facebook which I'll answer.

Donald Marshall Revolution is a website which details a brief overview of the Illuminati.
http://donaldmarshallrevolution.com/
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

DONALD MARSHALL
REVOLUTION

Interviews

You can listen to the radio interviews I have done. Listen for consistency; particularly anything which you do not hear me, pronounce clearly for the first time; the best thing to do is to pause the recording at that particular point and replay it. You should also research the statement you do not understand. Sometimes reading helps comprehension a lot faster.

One of the main reasons you should pause and replay the recordings is because: the truth has been kept hidden for so long that a lot of what I discuss in my interviews are beyond most people's current world view; so at some points I may speak too fast for you; my audio/ microphone may not be so clear so you may miss what I say etc.

A friend has told me that when he first watched the Vinny Eastwood interview; he did not hear me say the word "scars" (When Vinny asked: 'How do I know I'm the real me?') although he replayed that particular point in the video 8 times. Everything was just beyond his current comprehension, at the time. No matter how many times he replayed that part he really could not hear me say the word "scars" –so he let that part go, and played the rest of the interview, pausing, and replaying points which he did not understand, especially to comprehend whether I was talking about my original body or my REM driven clone duplicate body. He also listened to all my interviews for consistency, to note any 'slip ups', or any parts of my testimonies which do not 'add up'. He would listen to all my interviews, pausing and replaying parts he did not understand and he would reserve his judgements until he felt everything I was saying was for example as ordinary as: 'I woke up today, brushed my teeth, and took the dog for a walk'. He was also patient to realise the truth. After listening to my interviews he would just let it 'sink in'. A week later he would come back and listen to the same interviews, to test whether his comprehension on the topics I discuss has improved, and whether he can understand what I am saying without having to pause and repeat at certain points in my interview; and soon enough he could now hear me say “scars” at that particular point of the Vinny Eastwood interview. He had reached the point where all topics I discuss sounded to him like I’m saying everyday common place stuff that people have heard, such as: ‘I woke up today, brushed my teeth, and took the dog for a walk’. For anyone who may struggle to understand the topics I discuss: I strongly recommend you take the above approach as my friend did; soon enough, you too will realise the real truth of the world like he has: REM driven cloning, kept secret and used for sinister purposes.

I cannot say the following is true for everybody, however, an unproductive venture a complete newcomer can do is to listen to my interviews first time, all the way through, without pausing or replaying parts which they do not fully understand; If you do this and if there is just a single part of my interviews which does not make sense to you; this will interfere with your understanding of the entire interview. Remember, all I am discussing is technology, thousands of years advanced compared to what you currently use; available today, hidden and secret. If things start to get too complex for you, reduce it to its bare minimums: (advanced) science and technology. I hope that helps.

Another thing which I do in my interviews, that friends have picked up on, is: -because REM driven cloning has been my reality for many years; I don’t differentiate between my original body and my REM driven clone body. I just say: I did this, I did that, and Elizabeth did this and that, therefore it can become very confusing for newcomers. Please bare with me; although I’m more emotional as a REM driven clone, and not as smart as I am in my original body (this is a side effect of cloning: REM driven clones are more emotional than normal, and dumber than they are in their original bodies) I’m still “me” when my consciousness is transferred; I have all the experiences and knowledge which makes me, “me” and therefore I naturally do not differentiate between my REM driven clone version, and my original body as an outsider discussing these concepts would. I understand it helps comprehension so I have painstakingly done this throughout this disclosure.

I hope this helps; and I hope this helps to better understand my disclosure as well as the interviews which I have done. On the next page you can find the links to my interviews. You can copy and paste the links to your web browser or press Ctrl+Click – (hold “Ctrl” on your keyboard and left “Click” with your mouse, on the images below) to direct you to the interviews.

Links to Donald Marshall Interviews

Vincent Eastwood

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=M_1UiFeV5Jg&ab_channel=VincentEastwood
OR Press Ctrl+Click (on the image below) to follow the link.

Jeanice Barcelo

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=3uzgu4ekT3c&ab_channel=JeaniceBarcelo
OR Press Ctrl+Click (on the image below) to follow the link.

Lisa Phillips (CFR)

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=UonnFuHLJKc&ab_channel=Astral7ight
OR Press Ctrl+Click (on the image below) to follow the link.
Listen to parts 1 through 8.

Donald Marshall
1st Part Of Interview
April 11, 2013

Radio Presenters –Contact Donald Marshall

Anyone who sincerely wants to contact me for radio interviews on their show is welcome to do this. Please contact me through Proboards by leaving a message for me to contact you.

Professionals who understand "Consciousness Transfer" –Contact Donald Marshall

Any neuroscientists, engineers or professionals who understand how consciousness transfer works, and can provide me with a detailed methodology of how to block the consciousness transfer to my REM driven clone; please message me on Proboards, and this will be greatly appreciated.

Email

I currently do not have a contactable email address. In my original disclosure, the email address has been compromised (hacked), and so has any other email accounts I created: Yahoo, hotmail, Gmail, AOL etc. It doesn't matter; they eventually get hacked; for whatever reason, the Illuminati do not want me to have an email account.

How to Learn More about Donald Marshall's Whistle Blowing Disclosures

The best ways to learn about the truth of the world is through my Facebook page and Proboards. There is a mountain of evidence which has been collated over the years, and corroborates all that I have disclosed here. I also understand that for some people it is impractical to read three years worth of Facebook posts, especially when some post are repetitive; contain unnecessary comments from trolls, shills and other disinformation agents. Therefore, I am currently compiling 5 documents which should make it easier for others to read and understand everything quickly; as well as, for others to share and spread those 5 documents including this main disclosure document –so 6 documents about the real truth of the world. All for free!!

These documents include:

- "Frequently Asked Donald Marshall Questions" –This will contain all the general FAQs which I have been asked over the years; all in one place.
- "Experiences from the cloning centre" –This document will contain the full disclosure of my REM driven cloning experiences from the cloning centre when I go to sleep; all in chronological order.
- "List of people to avoid" This will contain a whole list of trolls, shills, and disinformation agents and double agents in alphabetical order, to avoid at all costs, in real life and online; -all complied in one place; and the reasons why they should be avoided.
- "Public Figures and their relationship with Donald Marshall as REM driven clones" – this will be a complete list of every public figure I have ever come into contact with as a REM driven clone; and you will be able to read about all the public figures I have met as REM clones, as well as the experiences, all in one place.
- "The subterranean underground colonists: The Vril" –this will be everything about the biological parasites the Illuminati harbour. Yes. The Illuminati harbour biological parasites, which they use against unsuspecting humans. I told you the Illuminati were

the biggest depraved perverts ever to grace God's earth. REM driven human cloning, kept secret from the world for over 70 years is difficult to fathom as it is, as are these biological parasites. These parasites are also one of the main root causes of humanity's problems here on earth, and must be rendered extinct. In this document you will be able to learn everything about their, strengths, weaknesses, psychology, biology, ecology, sexuality and arm yourself with knowledge which ensures that humanity exterminates these parasites from our earth, forever. The secret of Vril; a secret no more!

I hope after reading my eye witness statements presented in this document that you have now began to understand that this is simply what is happening in our world today.

Heed my disclosure, very diligently. At the start of my disclosure I asked the reader to do two things:

1) Give me the benefit of the doubt –because there is nothing worse than being a victim of abuse, reporting it and the people who have the power to stop it, ignore you;
2) Set out on a genuine quest to debunk my disclosure. I understand that for some people, plenty of what I have disclosed will still sound 'too out there' beyond their current scope of reality to be believable. For such people, the honest thing to do is to start with the first topic which you do not fully understand and explore it, either by finding that corresponding topic on Proboards (Donald Marshall Proboards 2015), and continuing your research from there; or just Google any topic you currently do not understand and review the topic on websites such as New Scientist (2015), Gizmag (2015), Motherboard (2015), BBC News (2015a; 2015b), BBC Future (2015), The Guardian (2015a; 2015b) The Independent (2015a; 2015b), Daily Mail Online (2015) and start reading more articles in the Science and Technology columns because these people have been telling you what they have been doing for decades. Consequently, keep in mind that articles which discuss 'future' technologies and 'improvements' in medicine or science, are articles, which are **really** discussing present technologies, and achievements in science and medicine already realised; because most of these achievements have been realised decades ago. It just hasn't been fully disclosed publicly.

I have also saved and backed-up every reference (except the homepages of the websites referenced) in this document; therefore if a link or video is ever deleted let me know.

How Much Do You Know About Post-humanism / Trans-humanism?

Furthermore, you have to be honest here: because if you do not know much or anything about post-humanism / trans-humanism then I am afraid to say, -and this is no fault of yours – that you are behind in this the world; simply because these Illuminati people are Post-humanists; and that is the direction the want the rest of the world to go. They want the world to reach a destination where human cloning is common place; a world where downloading your mind / entire life experience onto a computer microchip and living on as a cloned version of you with the microchip running the consciousness to be common place; and these Illuminati Post-humanists want the world to head in a direction where it is commonplace to molest children through science and technology. No. This really is not Science-Fiction, this is the world we live in today; this is why such marvels in science, medicine and technology has been withheld from the general public and kept hidden by an 'elite' group of people because if these people honestly came forward, and said "We have invented this technology and we want to use it for this "negative" purpose" –the populace would not accept it, and these people will be lynched in the streets before lunchtime. Therefore, through concealed advancements in medicine, science and technology, the Illuminati can live out their inner depraved cravings and commit crimes against humanity and the world, for which most people will be too blind to see or even fathom, because the hidden science and technology is beyond the average person's current comprehension.

Therefore, please attain an education in post-humanism / trans-humanism because these Illuminati people believe they can rival creation, nature and the marvels of the universe through science, medicine and technology. The Illuminati go **against** creation; they go **against** nature; and they go **against** humans. They don't consider themselves humans. They consider themselves post-humans and believe they can become "gods" through advancements in medicine, science and technology. See LawOfIdentity (2014) and Mark Dice (2014) for Richard Seed's comments on 'becoming gods' through transhumanism. Richard Seed is a Physicist with a Ph.D. from Harvard University; he is well known in the controversial cloning debate and declares his aspirations to "become god" saying 'We are going to become Gods, period. If you don't like it, get off. You don't have to contribute, you don't have to participate, but if you are going to interfere with me becoming a God, you're going to have trouble. There'll be warfare.' If you value humanity, do your best to attain an education in the trend of post-humanism / trans-humanism; otherwise the future of mankind will be bleak. We must do everything we can to stop these people.

This is of uttermost importance. REM driven cloning of is the most terrible thing in the world, especially when it involves unsuspecting civilians, and worst of all, innocent children. If you choose NOT to do anything you allow the Illuminati to continue to clone your children, sisters, wives, and sons. You allow the Illuminati to continue to hijack the minds of your children, sisters, wives, and sons while they sleep, through concealed advancements in science and technology. You allow the Illuminati to transfer the consciousness of your children, sisters, wives and sons to their REM driven duplicate clone versions while they sleep, whereby the Illuminati molest your children, sisters, wives and sons, which will cause them to have learning disabilities, unexplained depression and suicidal thoughts, as well as all kinds of side effects.

This IS your ONLY chance to end these monsters. There is a deadline!

This is your **ONLY** chance to do something to end these monsters. We are on a timeline and there is a deadline. As I have already mentioned, as well as, the top scientists in the field of physics and geophysics have mentioned: HAARP is not only capable of controlling the weather; once the HAARP grid is completed and working at full power, it will be capable of mind control over entire populations. We will all be slaves to the Illuminati FOREVER. Our freewill will be gone forever. Furthermore, if they achieve their aim of time travel with HAARP technology, future generations of humanity will never be able to stop them, because as they have told me: 'They will always be able to go back to a previous time, and correct the mistake(s) which led to their downfall'. This is the Illuminati's end game and what they wish to achieve. This is why this disclosure is so heavy, and so important. It goes beyond REM driven cloning; although it is important the world understands REM driven cloning. The future of humanity is at stake; and because of the 2-5 year deadline until the HAARP grid is completed, it is time humanity stood up and brought an end to the Illuminati. This REALLY is your ONLY chance to end these monsters!

The only way for evil to triumph is for good people to do nothing.
Edmund Burke

Remember for every 12 months which passes military / concealed technology outstrips the technology the general public is accustomed to by 44 years (Schneider 1995; 1996). Close your eyes and think 44 years into the future. The year is 2059. What kind of technologies do you expect mankind to have? Do it for real; don't just read the words. Close your eyes and imagine. The year is 2059. What kind of technologies does mankind have? You see. It is 'out of this world' technology, right. This is exactly what is going on today in the year 2015. It's just that it is hidden. That's all. Not so difficult to fathom now.

Do NOT Be Afraid to Help. We Outnumber the Illuminati by 1,000,000:1

Don't be afraid to help. These people are easily defeated when good people stand together as one. These Illuminati people total no more than 10,000 people. 10,000 people against 7,000,000,000 (7 billion) people –that's less than 0.000001 % of the world's population. Now can you begin to understand how prevalent evil can be when a small organised group of individuals, as little as 10,000 of the world's most evil and tyrannical people all work together in unison to exert their influence over the world? The world does not have to be this way. We outnumber the Illuminati people by a 1,000,000 to 1. For every Illuminati person there is, there are **one million** people who are not Illuminati, therefore do not to be afraid to help; all you will be doing is helping humanity rid itself of its sickness. So please feel free to share and spread my disclosure far and wide.

For those of you whom this applies to: remember the Illuminati (and I know it sounds ridiculous despite the evil they do) are deeply religious. They believe anyone aiding me is part of the "Army of Light" prophesised by Nostradamus; and to harm or degrade the life of anyone assisting me will bring them to their ultimate end and they will incur the wrath of God; they will suffer utter ruin and demise in their lives if they are to hurt you; they are very scared of people who can see through the lies and deceptions they have inflicted on the world. So please, stay calm, do not stress your heart, and know that **you are safe**, and **they cannot hurt you**. You can feel safe in the knowledge that you can do the right thing by helping me. They told me this as REM driven clones on the night of 21st of February 2014 when I went to sleep. The above is exactly what they said.

For anyone interested in understanding the interpretations of the Nostradamus Prophecies, Crystal Links (2015) provides a good source for all 942 of the quatrains. Delores Cannon is also an author who has written three volumes called "Conversations with Nostradamus" and for anyone interested you can read these online (Galactic 2012a; Galactic 2012b) as well as watch her YouTube videos (CreativeForceVideo 2014; Disclose TruthTV 2015).

Yes. This is the state of the world today: a deception within a deception coated in reverse psychology; and fact is indeed stranger than fiction. Now after everything I have said, if you are still sceptical and 'in-between' on this issue then the best thing for you to do is to pay attention to your dreams; or lack of dreams –they are your own experiences and you cannot deny your experiences.

In other words, you may be having REM driven cloned experiences too; for some people they have a 'dream', where they are in the same environment, over and over and over again; or the theme being discussed in their 'dreams' is the same over and over and over again; or the environment their 'dreams' take place always appear to happen in the same 3 or 4 environments over and over again –no matter what the theme is. In each scenario there is a likely probability that these are not just dreams, and in fact these people have had their consciousness transferred to the REM driven clone duplicates and the Illuminati is trying to extract something from them, once that is done they implant a false memory (See Kim (2013) and Alford (2015) for discussions on how false memories are implanted) so you wake up with the feeling of remembering something, but in actuality it is false; or you wake up with no memory of dreams from the previous evening, which means there is a high probability that you had your memories suppressed when your consciousness was transferred to your REM driven clone. This is why I say you can be sceptical about everything I have said (for now), but pay attention to your dreams, or lack of dreams, they are your own experiences, and you cannot deny your own experiences.

Noteworthy Frequently Asked Question

A Frequently Asked Question I receive, which is worth a mentioning here, is as follows: **Q:** Donald, if everything you have said is true...; in other words human clones walk among us now in a multilayered conspiracy which reaches the highest levels of government; armed with this knowledge what does one do exactly? What happens now, Donald? You obviously want to spread the word and make people aware but to what end?

DM: In short: I want to bring a complete end to the Illuminati and usher in a "Golden Age" of mankind. This question is best answered by detailing my 'Mission and Vision' for ending the Illuminati completely.

Donald Marshall's Mission and Vision on How to Bring Down The Illuminati

- This disclosure must spread, and **spread FAST and FAR!**
- If you now understand everything I have disclosed in this document as truth, then do not waste any more time. Share this disclosure with your wives, husbands, brothers, sisters' aunties, uncles, friends, co-workers, and children. We all have a part to play in saving the world from a premature doom.
- Call friends who have not heard from you in a while and tell them you have important news to share. Share this document. Share it on social media such as Facebook; Twitter; Instagram; Dropbox; Slideshare etc.
- Keep sharing this document until it comes to the attention of the Armed Forces. The Armed Forces will have to intervene. Once the Armed Forces intervene we will have reached the middle stage of this vision and will be witnessing a complete end to the Illuminati. Until then, we're in the beginning stages, so please; spread this document faster and further. The quicker it is spread; the quicker the Illuminati are ended.
- Once the Armed Forces have intervened a Military coup can be orchestrated against these vile people.
- Cloning centres can then be shut down once these people are overthrown.

- The freeing of any missing people (children; teens; adults) trapped in the cloning centres can commence (once the Armed Forces intervene).
- HAARP (as well as other highly advanced technologies) can now be contained and not used for adverse effects against the world (once the world learns of this disclosure).
- These Illuminati people will then HAVE to appear in court for their crimes against humanity.
- Suing the Illuminati members in court (once court proceedings commence... we will be past the midpoint of my vision, and closer towards witnessing a complete end to the Illuminati).
- The populace (and it will be your choice) can then have their REM driven clone experiences restored and also sue and claim any legal reward / compensation.
- After the populace has sued the Illuminati for their crimes against humanity; the punishments can commence.
- Punishments will include: imprisonment and executions of these sick and malevolent people. The imprisonment and executions of evil Illuminati members will bring an end to the Illuminati.
- After the punishments and executions; Governments worldwide can now be replaced with incorrupt individuals, worldwide.
- The structure of Governance will also have to change. The reason for this because future generations will always be able to check their leaders and governance more appropriately; so that the depraved and subhuman acts I described can NEVER be committed against humanity again; nor will world leaders be capable of committing such monstrous crimes in such secrecy ever again.
- Ensuring that the laws of the land always progresses in line with advancements in medicine, science and technology; as well as, ensuring law progresses in the directions of Research and Development (R&D) regarding future technologies, science and medicine.
- Release of technologies, science and medicine which benefit mankind.
- Commence a “Golden Age” of mankind.

If you do not fight the corruption and you do not stand up for what is right in life, you end up being a serf and a slave and you are leaving your children a world in which you would not want to live in yourself, so how can you in decency behave that way? You have to stand up for what is right in life, and unless you do that you are nothing.

Aaron Russo

Empowerment by Virtue of Golden Truth

As you can see, despite all the horror of the world I currently present: if everything goes smoothly, the good people of this world, truly will inherit it. This is why despite any anger, sadness, or fears you may have; you must not riot, damage property or cause chaos or bring about any other form of public dissension. You cannot stay silent, or ignore the issue in the hope that the threat removes itself. You are called to act; you must take action to help bring the Illuminati to an end; you must act while at the same time, you must maintain your composure and resolve to ensure that the whole procedure goes smoothly, and we all inherit a world we want to live in.

Your life, **at this very moment**, is more important than you may have probably ever imagined. You have purpose. Through you, and other good people around the world, together we can bring an end to such unspeakable forms of tyranny in this world. It is my humble stance that you have now become truly empowered with golden truth and you are now compelled to bring this tyranny to an end.

Do not waste the knowledge you have obtained from this disclosure. It is my only hope to escape this man made living hell. It is my only hope, as well as, the hope of many REM clones imprisoned there, as well as, real people who go missing daily, and are trapped in the cloning centres.

We sincerely beg you.

Donald Marshall

Appendices

Appendix A: Thien Thanh Thi Nguyen (Tila Tequila) Transcript

Link: https://www.youtube.com/watch?v=7mRZ7ItF9ls&ab_channel=Astral7ight

00.00 – 1.10 min: You know what... since you f****** with my program darling Queen Elizabeth [II] and the paedophilia ring and the cloning centres, and the cloning centres. That's right darling the cloning centres. Parents listen to me right now, they are blocking me but that is quite alright. They are blocking me but that is quite alright. Because I have many, many other forces; I shall not say their names right now, but I have many big plans to expose all of you disgusting, sadistic f***s! Okay? That is all. I shall save that for another time. But, however, I shall REPEAT: that was just an introduction to the reptilian family, leading all the way back. They call themselves the "The Black Nobility". Now that is just one part of it; alright?

1.11 -2.47 min: The Black Nobility; the reptilian family; all the way back from ancient times; so which they think... they feel like they are the divine chosen ones... from whom may I ask? Definitely. Definitely not God. Our creator of the world. Reptilians: they feed on blood; children; the paedophilia ring; recently busted. Oh! It has been going on for centuries. Parents listen to me carefully. I don't care if I'm cutting out. I will continue this and I am not alone on this battle. Believe THAT! Believe THAT! I am not alone on this battle. I started out alone but I am marching on with MILLIONS; okay? So sit your old a** down okay. Because you are gonna roll over, but it doesn't matter anyway because you are all ancestral f***s! Who interbreed... ancestral f*****g... and then... and then... Oh! Only going to talk about cloning those children and... oh! And all those many children's parents listen to me carefully.

2.48 -4.24 min: Hundreds of missing children come up every year. You wonder... why? How? How could this be? And then there is so called CPS [Child Protection Services] or whatever they are called; they come and take your children, just, just for nothing; right? Not to discredit everybody, because not everybody is bad. I'm talking about the bad people. And they take your kids... they take them as this... they treat them like... I can't even say the word. It's disgusting, it's sadistic. They... they... they take your kids... they toss them out like little [inaudible] cause they are so f*****g... Pardon my French, but then again when I'm dealing with these evil cum-buckets I have no holy words coming out of my mouth, because these are the only words they resonate to. So therefore that's how I refer to them. Because they can only [inaudible] ...their masters whiplash on them with these [inaudible] words of cursing, vileness and slaving and that is not what the true God is; okay? The God of Hell...; Anyways...

4.25 – 5.30 min: There are these CLONING CENTRES where they take your children and do sadistic things to them. I'm not even talking about child molestation here; not to mention; uh there is one of them that got caught, flying out to Florida; to meet up with a four year old little girl, to have sex with a four year old girl. That's right. Google it because it is so highly sinful... We have commercials about... you know starving kids; you know save the starving kids and it is heartbreaking. We can have commercials about it... why? Because it is a horrible thing and people can have commercials about it. WHY do we not have commercials about... 'Daddy please don't, don't. Mommy please don't let daddy let daddy touch me?' – because it is disgusting! It is so disgusting; beyond sin that no one, NO ONE, can even make a commercial about that because that's how sinful it is.

5.31 – 7.19 min: Do you understand that? Do you understand how sinful that is? These people, I'm not even going to call them people; alright? They take your children; they not only molest them; men f*** them and make them shoot each other. They give them guns. It's either you shoot him or I'm gonna shoot you. They are... mind you, they are children. Children. Yes. I'm speaking out, because I... uh! Who else is doing this? You're all just [inaudible]... shame on you... And actually the most recent paedophilia... got taken down... WOW! How long did that take? Really?! Do you know how long this has been going on? *Sigh* Alright I'm gonna calm down... but as a parent and I love parents out there. My heart goes out to all the parents out there who have missing children. You know, we all pray for them; every day. And I put on a bold face in public every day; because there needs to be someone strong, believe that. But my strength comes from somewhere... that I have a very vulnerable emotional side where I feel very strongly for these children and innocent people. So therefore I want to speak out.

7.20 – 9.10 min: I have and I have my passion too; and in the end you shall all know why I am so passionate about exposing every single one of these scumbags; okay? The truth shall prevail and you all will know why. So, as for you parents... ah... there are no words to describe. But let me just expose because you can't just... there is a point where you CANNOT just turn the other way. You know this stuff is going on; and you go 'Oh well... that's their problem. Let's just turn the other cheek.' How long are you going to turn the other cheek, until it happens to your own freaking children? When, when, when your own child; three years old, get's run out [kidnapped] and gets blasted right in the head with a gun... yeah... there's more to that people; okay? And I'm not just saying that coming from some... I mean, actually, mothers, fathers out there... If you found out... I'm sure you would do way worse than what I am just saying. I'm just using voc. I'm just annihilating them vocally. I'm sure the parents out there who find out what their children have gone through, their missing children. I am pretty positive more than just a vocal annihilation of these scumbags that do this to your children; okay?

9.11 – 10.32 min: So keep on turning the other cheek folks. Hey, go and turn the other cheek; you are with the others. I am not. I AM NOT. Like I said you... There is only two ways to go about this: you're either with us: the good guys, or you're with the others. There is no in-between. Cause if you are in-between, hiding like cowards, turning the other cheek; doing whatever; well then you are a freaking coward and you are just a sheep. And sheep end up dying because you know what you are owned by "The Others". So pick one: you're either one of us: the good guys or you're part of the others. That is simple as that! Two choices: good guys; bad guys. In-between you're dead because the bad guys are going to suck your soul out [transfer your consciousness to your REM driven clone] and do some sadistic stuff to you and watch, and make you watch while they pretty sadistic stuff to your children as well.

10.33 – 11.33 min: Do you know why they love children? Because they are innocent souls; they're innocent... they're, they're the most innocent pure beings in this planet. They're not harmed by anything. They're new to the world; bright-eyed pure innocent children. That is why these disgusting paedophile and these clone rings; cloning centres; satanic rituals; Brownsville Texas... There are many other cloning centres where they take your children that go missing. And you wonder why? What happens to them? I'm sorry to break this to you but that is what happens. Now either do you want to know what happens to them or do you want to turn the other freaking cheek?

11.34 – 13.09 min: Tune into my next show. I'm gonna upload stuff; I have an arsenal ready to blow up! Okay. And I have reason behind this. You all should know my personal reason soon; but this has nothing to do with me right now. But know that I'm back, I'm back with a vengeance and I'm back with an army full of people around the world who are sick and tired of treated like animals; or quote, unquote "COWS". We all know what that means. For the outside world (the masses) we all know the term "sheep", sheeple. But for the insiders we know what the cows are don't we? You know what "The Others" like to do with the cows, right? They start to herd you in... and to... yeah...I'm gonna end it at that. And to all the parents, families and children out there, I love you so much. I... have to maintain composure, because that is what I do. That is all. Over and out.

Appendix B: MK Ultra

MK Ultra *-noun:* [Manufacturing Killers Utilizing Lethal Tradecraft Requiring Assassinations] The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct, and to control the absolute behaviour and thought patterns of the individual.

.

Modern-day MK Ultra involves an implantable microchip which is inserted into a Mark 2 clone of the victim. Every thought, reaction, hearing and visual observation causes a certain neurological potential, spikes, and patterns in the brain and its electromagnetic fields, which can now be decoded by the implantable microchip into thoughts, pictures and voices. The thoughts, pictures and voices of the implanted Mark 2 clone (victim) can now be displayed visually and heard on any system capable of converting visual images, such as a television or a computer. These images are usually displayed on a giant screen at the cloning centre.

The purpose of MK Ultra is to elicit a specific conditioned response in a victim (desired by the programmer) to an otherwise neutral stimulus. In other words, through the process of "classical conditioning" (which involves learning a new behaviour via the process of association. In simple terms two stimuli are linked together to produce a new learned response in a person) the implanted Mark 2 REM duplicate clone (victim) is placed thorough many similar recurring scenarios which are experienced (or perceived) by the victim as "real", because the programmer manipulates the victim's audio and visual field, and continuously pairs the victim with the neutral stimulus, so that either a positive or negative conditioned response is elicited in the victim towards the neutral stimulus.

In other words an implanted victim can be placed in many recurring situations deemed as frightful, where a mystery stranger saves the victim from the frightful situation over and over again. The victim will now have positive associations towards the mystery stranger and therefore the conditioned response is now one of 'positive associations' (towards the mystery stranger who did not elicit such a reaction before).

MK Ultra techniques are administered through Mark 2 (sleep driven) clones. In other words classical conditioning happens when the person sleeps. Memory suppression technologies are also used in conjunction with MK Ultra technology, which enables the programmer to control certain memories the victim remembers. The use of memory suppression technologies and MK Ultra technology allows the programmer to reinforce behaviour and elicit specific conditioned responses (in an original).

Therefore, in our example above, when the original meets the mystery stranger in real life, the original (victim) will be predisposed to the mystery stranger and automatically 'feel safe' around this person and may even believe that he / she should 'date this mystery stranger' – depending on what was programmed; because the victim was conditioned to elicit such an emotional response towards the mystery stranger in his / her sleep as a Mark 2 REM driven clone through MK Ultra functionality.

MK Ultra can also be used to create zombified individuals who are programmed to murder and remember nothing of their crime afterward. MK Ultra (technology) can also be used to disrupt the memory of the original, discredit people through unusual behaviour, make the person insane or commit suicide and murder.

MK Ultra has much functionality. Furthermore, because of its video and audio projection functionality, MK Ultra can also be used to compose music. MK Ultra is capable of relaying the Mark 2 REM driven clone's subconscious and conscious mind as visual images and audio, projected onto a screen. Therefore thoughts, pictures and voices which have been experienced consciously and subconsciously in one's life can be displayed visually and audibly on a television or computer screen. The Mark 2 clone now has the option of harmonising these thoughts, pictures, and voices into a coherent order which produces music through the thought process of his or her imagination. See Jim Cristea (2009); Berkeley News (2011); UC Berkeley Campus Life (2011) CTForecaster (2013); nature video (2013) and Stromberg (2013) for examples and discussions and the capability of MK Ultra technology.

Appendix C: The Illuminati

Illuminati *–noun:* A modern-day criminal organisation operated by reprobate (depraved, unprincipled and wicked person) criminals. Their main agenda: is to enslave the whole world through advanced concealed technologies.

Today's Illuminati trace their roots back to Professor Adam Weishaupt who found the Illuminati on 1st May 1776. Since the inception of the Illuminati the intent has always been, and remains: "to bring about a NEW World Order that writes God out of the picture and deifies (glorifies) Lucifer." This intent is still prevalent today. The following excerpt is derived from A. Ralph Epperson (1990) *The New World Order:*

> Weishaupt was a teacher of Cannon Law (law governing the affairs of a Christian Church, especially the law created or recognised by the Papal authority in the Roman Catholic Church) at the University of Ingolstadt in Bavaria, now part of Germany.
>
> He even told the world, in his writings, where he would conceal the Order: "None is fitter than the three lower degrees of Free Masonry; the public is accustomed to it, expects little from it, and therefore takes little notice of it." He felt that this secrecy would lead him to success because he felt no one would be able to break into it. He wrote: "Our secret Association works in a way that nothing can withstand ...."
>
> Weishaupt accepted the fact that all secret associations and secret orders had two doctrines, one concealed and reserved for the Masters... the other public ...." and the Illuminati was [and are to this day] a secret society with two doctrines.
>
> Professor Weishaupt, its founder, boasted of his organization's secrecy. He realized that this secrecy would enable them to decide the fate of nations and because their deliberations were secret, no outsider could interfere. He wrote: "The great strength of our Order lies in its concealment; let it never appear in its own name, but always covered by another name, and another occupation." Weishaupt later wrote about that secrecy in a letter to a fellow member of the Illuminati: "Nothing can bring this about [the new world order] but hidden societies. Hidden schools of wisdom are the means which will one day free men from their bonds [the "bonds" of religion] Princes and nations shall vanish from the earth." So the secret societies were created to bring the world to the new society known as the New World Order. The members of these organizations obviously feel that their goals are so noble that they may perform whatever tasks are required of them to bring that goal to fruition. This means that murder, plunder, and lying all become acceptable as long as these methods assist its members in obtaining their goal.
>
> Adam Weishaupt, the founder of the Illuminati, wrote over and over and over again, that "the ends justified the means." Weishaupt also told initiates to use whatever means, which included murder, to achieve the goals of the association that he was joining. And that the major goal of the Illuminati, was the destruction of all religion, including Christianity. That meant that if Christians physically stood in the way, they could be removed by simply murdering them. Weishaupt even went so far as to say that anyone not willing to take the life of another was unfit to join the Illuminati. He wrote the following in a letter to a fellow member in 1778: "No man is fit for our Order who is not ... ready to go to every length ...."
>
> Another reason, that Weishaupt felt that the Illuminati would succeed, was the fact that he was offering his members worldwide power. He felt that this inducement would enable him to draw into his organization only those who would do anything to satisfy that desire for power. He wrote: "The true purpose of the Order was to rule the world. To achieve this it was necessary for the Order to destroy all religions, overthrow all governments and abolish private property."

> But his religion had a different base than the traditional religion: his was based upon a worship of reason: "... then will Reason rule with unperceived sway." "... Reason will be the only code of Man. This is one of our greatest secrets." "When at last Reason becomes the religion of man, then will the problem be solved." Weishaupt's dedication of his organization to "reason" makes some sense when the reader recalls that "reason" has been defined as the "unbridled use of man's mind to solve man's problems without the involvement of God." The Bible calls this "the fruit of the tree of the knowledge of good and evil." It was this knowledge that God wanted man not to have, and it was the promise made to man by Lucifer that man could have it by eating of "the fruit." In addition, Weishaupt's religion offered its believers a reward not offered by any other religion: worldwide power!
>
> Weishaupt wrote: "The pupils [members of the Illuminati] are convinced that the Order will rule the world. Every member therefore becomes a ruler." Weishaupt's religion not only offered power to his believers, but he offered them something else not guaranteed by any other religion: worldly success. He said that once a candidate had achieved the exalted degree of Illuminatus Minor, the fourth of the thirteen inside his Order, his superiors would: "assist him [the member] in bringing his talents into action, and [would] place him in situations most favourable for their exertion, so that he may be assured of success." Finally, the goal of the Illuminati was "man made perfect as a god - without God."

The ideology of “man made perfect as a god –without God” still remains to this day, and it is practiced by today’s Illuminati members. The above phrase is what ties in Luciferian worship and trans-humanism. Lucifer is idolised by Illuminati members as the deity who gave man ‘knowledge’ and therefore is worthy of worship; God, -according to Luciferians, -did not want man to have knowledge and therefore is despised by Luciferians. Ingrained in the trans-humanism doctrine is the believe that: ‘man can become ‘god’ through science and technology and in turn overthrow the Creator of the universe: God’. See LawOfIdentity (2014) and Mark Dice (2014).

These are the basics of Luciferianism. Therefore everything which is natural or pertains to nature must be contended or destroyed by Luciferians. This is why Illuminati members endorse having sex with children, killing first born sons, and drink blood. All the above go against nature and according to the ‘edicts of Lucifer’: paedophilia makes the person committing the act younger (it doesn’t, it is just an excuse to act perverse on children because they know children are vulnerable); killing your first born son gives you good luck and fortune in this life (so yes, some Luciferians have sacrificed their first bon sons); and Illuminati members believe drinking blood / cannibalism is a ‘purifying agent’ (although in reality it causes spongiform encephalitis (holes in the brain)).

Modern-day Illuminati members still retain the goal of its founder “to bring about a NEW World Order that writes God out of the picture and deifies (glorifies) Lucifer.” Infiltration through secrecy, still remains their mode of operation, for the current Illuminati and therefore they have secretly infiltrated all the major religions on earth; government and education – where each successive generation is being dumb down; they have continually diminished the ability for individuals to own private property, or claim inheritance; divided people against each other to continually diminish patriotism; and have continually diminished family values.

They also compartmentalise their knowledge between members. Until I fully exposed the Illuminati, many people who have been REM clones at the cloning centre did not know they were in fact REM driven clones, and thought they were in the 5th Dimension; the astral plane; a singularity; the spiritual realm; Valhalla; quantum hopping; a time stutter etc. or whatever else the Illuminati told them.

Today's Illuminati members also meet in secret (just like the founding members) as REM driven clones, **when they go to sleep**. Furthermore, because Illuminati meetings are in secret and not many people know the exact location (because knowledge is compartmentalised) of the cloning centre; or the fact that their consciousness has been transferred to Mark 2 REM driven clone bodies at the cloning centre (and they are not in their original bodies); as well as the fact that unsuspecting civilians have their memories suppressed; the points mentioned above are the reasons the Illuminati believe 'they are all powerful and untouchable'; and as a consequence, today's Illuminati members do all the disgusting things they want, because they believe no outsider can interfere.

The ring leaders of the Illuminati today also believe "the ends justify the means". This is why they clone, torture, molest, murder, and rape unsuspecting civilians as REM driven clones in their sleep. The Illuminati of today offer their members incitements to go along with their agenda and not oppose them (or face death). Another popular method is to entice their target with many, many wonderful prospects, and have the target believe they are joining a noble and prosperous venture, so that the target fulfils the objectives of the Illuminati unknowingly; promoting the Illuminati in a positive way, because the target has been deceived to perceive the nature of the Illuminati as 'positive'; by the time the target finds out the true intent of the Illuminati, and the evil which emanates from it, it is too late. Those who rise up in the ranks of the Illuminati are the men and women who have an insatiable lust for power, and most importantly: the men and women who want to rule the world.

Modern-day Illuminati members also wish to become gods (through technology); overthrow the Creator, and achieve their overall aim of enslaving mankind. This is why they clone people, and clone people in high rank society from all walks of life (movie stars, musicians, politicians etc.; whether the person willing wants to be part of the Illuminati or not) so long as that person is in a position of power and influence, the Illuminati clone that person, and threaten that person, for example -"Hey, you're going to hang with us –or else" –through such coercion, the people in high rank society who have power and influence will not oppose the Illuminati's plans to become gods; enslave mankind forever; and rule the world. Another reason for cloning high rank society is to include these people into the Illuminati (willingly or unwillingly) to ensure the world's populace remains in ignorance (because once all the world leaders and high rank society are cloned and under the coercion of the Illuminati nobody in a position of influence or power can warn the populace against the Illuminati) until the Illuminati's plan is completed and they have enslaved the world forever.

The Illuminati's overall aim of 'becoming gods' and ruling the world, as well as, mind controlling all the inhabitants of earth is also the reason the Illuminati:

- Administer drip feed disclosure through media, by telling some truths mixed with lies in order to conceal their true intentions and overall aim, and prevent the betrayed partner (the public) from ever discovering "the complete truth";
- Administer evaluative conditioning, by placing their symbols and ideology in popular media with positive associations, so that the unknowing and unsuspecting public will eventually become predisposed to the Illuminati and unsuspectingly have a positive or neutral response towards the Illuminati;
- Because they want to become 'gods' is also the reason the Illuminati is promoting RFID microchips and only discussing RFID microchips positively, while at the same time placing suppression (gagging) orders on anyone who speaks negatively about RFID microchips –which implies an unsuspecting public will willingly accept the microchip; and at the point of transaction, the person will have (unknowingly) given up their privacy to a third party (the Illuminati) for the rest of his / her life;
- Their aim to become 'gods' is also the reason the Illuminati are hurriedly trying to complete the HAARP grid across earth –because a complete HAARP grid will allow them to achieve their goal of mind control over the entire world; which fulfils their objective of becoming gods; because a complete HAARP grid will be capable of time travel, and therefore the Illuminati will be capable of going back to a previous time to correct the mistake(s) which led to their downfall; the Illuminati members will retain the knowledge of the previous time, and the rest of humanity will have no recollection of such a memory.

The Illuminati is **not** a joke. It is **not** fiction. They are very real, and part of humanity's reality; and through advanced concealed technologies the Illuminati aim to enslave humanity forever. The ring leaders of the present Illuminati includes Queen Elizabeth II, Prince Philip Duke of Edinburgh, Prince Charles of Wales and Vladimir Putin. I have also detailed the actions of the ringleaders in the main text in this disclosure, as well as, detailed other modern-day evil Illuminati members on my Facebook and Proboards. It is time the good people of earth, stopped being afraid, do the right thing, put a stop to this evil, and save themselves, as well as their children's children from being slaves forever. Spread and share this disclosure.

Glossary

Aneurysm -*noun:* An excessive swelling of the wall of an artery at a fixed point in the body. A brain aneurysm is therefore a: bulge or ballooning in a blood vessel in the brain. It often looks like a berry hanging on a stem. A brain aneurysm can leak or rapture, causing bleeding into the brain.

Brain Aneurysm -*noun:* see Aneurysm.

Clone –*noun:* a cell, group of cells, an organism produced asexually from a single ancestor and is genetically identical to a single ancestor.

Concealed Technology -*noun:* hidden machinery and devices undisclosed and currently unavailable for public consumption. Concealed (or military) technology develops at a rate of 44 years for every 12 months which passes in comparison to the technology the public is currently accustomed to. Origin: Phil Schneider.

Cloning Centre –*noun:* a physical location (on earth) where duplicate and replicate clones are produced. These physical locations are usually Deep Underground Military Bases (DUMBs). DUMBs have an entire floor dedicated to cloning. It is also a place where Illuminati members meet with each other as REM driven clones. The above-ground cloning centre where many high profile people attend can be found within a radius of 5/6 hours drive from the Robert Pickton Farm Port Coquitlam, British Columbia, in Canada, somewhere at a remote nature reserve.

Cloning Technology -*noun:* the technological advancements in medicine, science and technology used to produce duplicate and replicate copies of originals.

Conditioned Response -*noun:* an automatic response established by training to an ordinarily neutral stimulus.

Conditioned Stimulus –*noun:* A previously neutral stimulus that, after repeated association with an unconditioned stimulus, elicits the response produced by the unconditioned stimulus itself.

Consciousness –*noun:* the state of being aware of and responsive to one’s surroundings; a person’s awareness or perception of something.

Consciousness Transfer -*noun or verb:* the process of transferring or copying the mental content (including long-term memory and “self”) from a particular brain and copying it to a computational device; artificial body or avatar body such as that of a robot or clone version of the original. It is also the feat in which the person’s mental content (long term memory and “self”) moves from one body into another.

Depopulation *verb:* to remove or reduce the population of, as by destruction or force.

Drip Feed Disclosure *noun or verb:* is the process of supplying information but in small amounts overtime. Drip feed disclosure is also the process of revealing information slowly overtime, possibly telling lies to conceal certain aspects of the truth until the source administering the drip feed disclosure has adequate time to let out the truth in a slow and controlled way, thereby delaying the betrayed partner (in this disclosure, the public) from having the "complete truth" for some time.

DUMB *–noun:* [Deep Underground Military Base] a facility directly owned and operated by or for the military or one of its branches that shelters military equipment and personnel, and facilitates training and operations beneath the surface of the earth.

Duplicate Clone *-noun:* a fully formed human body which is a genetic copy of original developed through the process of regenerative technology. Duplicate clones are grown in a big thick tank full of (salty) water.

Duplication Cloning *-verb:* involves agitating the cells of on an original repetitively until a fully formed human body of the original is developed. Duplicate clones take an average of 5 months to form into a fully developed human body of the original through the process of regenerative medicine and technology.

Evaluative Conditioning *-noun:* is a change in liking, which occurs due to an association with a positive or negative stimulus.

H.A.A.R.P. Technology *–noun:* [High frequency Active Auroral Research Program] a radio transmitting system that can bounce signals off the Earth's upper atmosphere, (60km (37 miles) to 1000km (620 miles) high) back to probe deep into the earth or sea. HAARP is also capable of: disrupting human mental processes; knocking out all global communication systems; changing weather patterns over large areas; interfering with wildlife and migration patterns; hurting ecosystems; negatively affect human beings health, moods and mental states; and unnaturally 'boil' the earth's upper atmosphere. HAARP used correctly will control the weather without any adverse effects.

H.A.A.R.P. Grid *–noun:* a network of radio transmitters which can bounce signals off the earth's upper surface. Each transmitter is located at a specific point across earth and communicates in unison with other radio transmitters across the earth. At this present time of writing, a HAARP grid has **not** been completed, although the Illuminati are working twice as fast to complete a HAARP grid. The threat to humanity once a HAARP grid is completed includes: mind control over the entire world's inhabitants. A completed HAARP grid will also be capable of time travel, and therefore the Illuminati will always be able to go back to a pervious point in time to correct the mistake(s) which led to their downfall. Humanity will be slaves forever.

Habeas Corpus *-noun:* is [A] writ [formal document] requiring a person under arrest to be brought before a judge or into court, especially to secure the person's release unless lawful grounds are shown for their detention.

Heart Attack *-noun:* A sudden occurrence of a blockage of the flow of blood to the heart.

Human Clone *–noun:* The creation of a genetically identical copy of a human.

Ionosphere *–noun:* the layer of the earth's atmosphere which contains a high concentration of ions and free electrons and is able to reflect radio waves. It lies above the mesosphere and extends from about 60km (37 miles) to 1,000 km (620 miles) above the earth's surface.

Illuminati *–noun:* A modern-day criminal organisation operated by reprobate (depraved, unprincipled and wicked person) criminals. Their main agenda: is to enslave the whole world through advanced concealed technologies. See Appendix C for further discussion.

Mark 2 Clone *–noun:* Is a sleep driven clone; specifically, a REM sleep driven clone. A Mark 2 Clone is activated by transferring the consciousness of an original into a Mark 2 Clone **only when** the original reaches REM sleep (usually 90- 110 minutes **after** the original falls asleep). Once the consciousness of the original is transferred from the original's body to the Mark 2 Clone, the Mark 2 clone is now capable of motion: such as walking, talking etc. The Mark 2 clone 'drops limp' and becomes motionless once the original wakes up from sleep. The original's consciousness no longer resides in the Mark 2 clone (once the original is awake) and therefore the Mark 2 clone is now incapable of motion. Mark 2 Clones are also known as "REM Driven Clones" and "REM Duplicate Clones".

Memory Suppression *–noun or verb:* is the selective removal of memories or associations with the mind using memory suppression technology.

Memory Suppression Technology *–noun:* any scientifically advanced technology used selectively to remove memories from the conscious mind.

Mind-Voice Technology *–noun:* an advanced technology capable of reading, listening, hearing and broadcasting a person's inner voice or thoughts. It is capable of replicating sounds **exactly**. Therefore an individual can hear the sound of drums, a guitar or any instrument and replicate that sound exactly just by thinking about it. Consequently, Mind-voice technology has the functionality of producing music.

Military Technology *–noun:* machinery and devices developed from scientific knowledge used by the Armed Forces which advance at a rate of 44 years for every 12 months which passes, compared to the technology the public is accustomed to. Origin: Phil Schneider

MK Ultra *-noun:* [Manufacturing Killers Utilizing Lethal Tradecraft Requiring Assassinations] The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct and to control the absolute behaviour and thought patterns of the individual. See Appendix B for further details.

Neutral Stimulus *–noun:* is a stimulus which initially produces no specific response other than focusing attention. In classical conditioning, when used together with an unconditioned stimulus, the neutral stimulus becomes a conditioned stimulus.

Negative Association *-noun:* is an undesirable experience or perception.

Negative Stimulus *-noun:* a stimulus with undesirable consequences.

New World Order *-noun:* [NWO] Agenda. The whole NWO agenda is to turn humanity into mindless slaves forever; whereby the post-humans / trans-humans mind control the entire world's populace either through RFID microchips or a completed HAARP grid. Another aspect of the NWO agenda is to depopulate the world's current population of 7.3 billion people to 500 million people (and never exceed a world population of 500 million people afterwards); ruled by a one world government; a one world ruler; with a one world religion. See Appendix C for more details.

Original *–noun:* A person who is **not** a clone.

Pain Receptor *-noun:* Any one of the many nerve endings throughout the body that warn of harmful changes in the environment such as excessive pressure or temperature.

Positive Association *-noun:* is a desirable experience or perception.

Positive stimulus *-noun:* a stimulus with desirable consequences.

Posthumanism *–noun:* seeks to rewrite the very definition of being human. It is the condition in which humans and intelligent technology become intertwined. In the Posthuman there are no essential differences or absolute demarcations between bodily existence and computer stimulation, cybernetic mechanism and biological organism, robot technology and human goals.

Posthuman *-noun:* see Posthumanism.

Project MK Ultra *-noun:* see MK Ultra.

Regenerative Medicine *-noun:* (of a living organism) the process of re-growing new tissues after loss or damage.

Regenerative Technology *-noun:* any machinery or device developed from scientific knowledge which has the capability to re-grow new tissues after loss or damage.

REM Sleep *–noun:* [Rapid Eye Movement] is the fifth stage of sleep in the sleep cycle. It takes 90 - 110 minutes to reach REM sleep after we fall asleep. REM sleep is also known as the "period of paralysation". The involuntary muscles such as the brain become more active whereas voluntary muscles (those that you move by choice) such as your arms and legs become more relaxed or paralysed. REM sleep is a kind of sleep that occurs at intervals during sleep, and it is characterised by rapid eye movements.

REM Driven Clone *-noun:* [Rapid Eye Movement Driven Clone] a clone that can only become activated, once the original is in REM sleep. See Mark 2 Clone.

REM Duplicate Clone *-noun:* [Rapid Eye Movement Driven Clone] A clone developed by regenerative medicine and technology and is therefore an identical copy of an original. REM duplicate clones can only become activated when the original is in REM sleep. See Mark 2 Clone.

REM Driven Clone Death *–noun:* the process where an original dies because of constant torture to their REM driven clone or where a constant electrical current is applied to the REM driven clone resulting in death of an original usually in the form of an aneurysm or heart attack (because consciousness is linked) in the original's body.

REM Driven Clone Torture *—noun or verb:* the action or practice of inflicting serve pain on a REM driven clone. REM driven clone torture causes biological and physiological responses in the original's body because consciousness is linked. Intermittent REM driven clone torture (depending on what is done) causes the original to experience severe headaches, an upset stomach, achy limbs, sickness; a weakened heart. Continuous REM driven clone torture will lead to the death of the original; usually in the form of aneurysm or heart attack in the original's body.

Replication Cloning *–verb:* involves giving birth to a genetic identical of an original where the newborn starts life off as a baby and matures. The newborn is referred to as a clone.

RFID Technology [Radio Frequency Identification] *-noun:* are electronic microchips the size of a grain of sand that can be directly embedded into the human flesh. RFID microchips communicate wirelessly through the use of electromagnetic fields to transfer data. RFID microchips link the brains of people via the implanted microchip to satellites controlled by ground base super-computers. The dangers of RFID microchips to the implanted person are: total loss of privacy and total control of the person's physical body functions, mental and emotional thought processes, including the implanted person's subconscious or dreams –for the rest of that person's life! RFID microchips are also tracking devices, and the implanted person can be tracked anywhere on the globe.

Selling One's Soul *-verb:* to sell the use one's "Mark 2" REM driven clone to the Illuminati, for the Illuminati to use the individual's Mark 2 REM driven clone in whatever manner the Illuminati wishes. There are no returns once the individual has signed over his / her (soul) Mark 2 REM driven clone. When an individual sells their (soul) Mark 2 REM driven clone, the person has also entered into a contract to sell the Mark 2 REM driven clone(s) of their current children (if they have any) as well as any unborn children the person may have later in life. The person sells all their descendants (souls) Mark 2 REM driven clones to the Illuminati, once the individual sells their (soul) Mark 2 REM driven clone to the Illuminati. Selling one's (soul) Mark 2 REM driven clone is considered **a serious business transaction** to the Illuminati. There are no returns. If the person ever makes a fuss and wants their (soul) Mark 2 REM driven clone back, the Illuminati will either torture the person's Mark 2 Clone, or apply a constant electric current to the person's Mark 2 Clone until the person either has a heart attack or aneurysm in their original body. This is what public figures are hinting at when they say "They have sold their soul". They have sold the use of their Mark 2 REM driven clone to the Illuminati. Selling one's soul is not a joke. Never sell your soul.

Stimulus *–noun:* is something that causes a reaction, especially interest, excitement or energy. It is also an energy change registered by the senses. For example a stimulus can be a shinny object for a baby.

Technology *–noun:* machinery and devices developed from scientific knowledge.

Technological Advancement *–noun:* is incorporating, by means of experimental development, a characteristic or capability not previously existing or available in standard practice, into a new or existing process or product that enhances a product's performance. Novelty, uniqueness, or innovation alone does not indicate a technological advancement.

Transhumanism *–noun:* the belief or theory that the human race can evolve beyond its current physical and mental limitations, especially by means of science and technology.

Unconditioned Response *–noun:* is a response to a neutral stimulus we have no / little control over. It is a natural automatic response. For example, food is an unconditioned stimulus for a hungry animal, and salivation is the unconditioned response.

Unconditioned Stimulus *–noun:* A stimulus that elicits an unconditioned response.

All references in this document have been saved and backed-up; therefore if any link in this document, is ever deleted, modified etc. (online), let me know.

References

2045 Initiative (2015) *2045 Strategic Social Initiative* [Online] Available from: http://2045.com/ [Accessed 1st June 2015]

Albrecht, K., and McIntyre (2005) *Spychips: How Major Corporations and Government Plan to Track Your Every Move with RFID.* Plume

Alford, J., (2014) *Scientist Selectively Erase and Restore Memories* [Online] Available from: http://www.iflscience.com/brain/scientists-selectively-erase-and-restore-memories [Accessed 8th July 2015]

Alford, J., (2015) *Scientist Implant False Memories Into Sleeping Mice* [Online] Available from: http://www.iflscience.com/brain/scientists-implant-false-memories-sleeping-mice [Accessed 28th July 2015]

Animal Research (1996) *Cloning Dolly the Sheep* [Online] Available from: http://www.animalresearch.info/en/medical-advances/timeline/cloning-dolly-the-sheep/ [Accessed 28th June 2014]

Anthony, S., (2012) *GoFlow: a DIY tDCS brain-boosting kit* [Online] Available from: http://www.extremetech.com/extreme/121861-goflow-a-diy-tdcs-brain-boosting-kit [Accessed 24th May 2015]

Anthony, S., (2013) *What is transhumanism, or, what does it mean to be human?* [Online] Available from: http://www.extremetech.com/extreme/152240-what-is-transhumanism-or-what-does-it-mean-to-be-human [Accessed 24th May 2015]

Astral 7sight (2013a) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 1/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=UonnFuHLJKc&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral 7sight (2013b) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 2/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=QPAXCwu5MIo&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral 7sight (2013c) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 3/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=5zlJ0VQP444&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013d) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 4/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=dCGhrDEl-q8&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013e) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 5/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=EDyerA-k8Ic&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013f) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 6/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=iqQZPpXl2yg&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013g) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 7/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=WYZSOnyWwP8&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013h) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 8/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=Y18m0gPLQhM&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013i) *Celebrity Tila Tequila: Missing Children and Cloning Centers* [Online video]. June 7th 2013. Available from: https://www.youtube.com/watch?v=7mRZ7ItF9ls&ab_channel=Astral7ight [Accessed 12th July 2015]

Avatar (2009) Film. Directed by James Cameron. [DVD]. UK: 20th Century Fox

BBC Future (2015) *BBC Future* [Online] Available from: http://www.bbc.com/future [Accessed 28th July 2015]

BBC Horizon (2009) *Why Do We Dream BBC Horizon* [Online video]. May 15th 2013. Available from: https://www.youtube.com/watch?v=E8MZ1twv0cU&ab_channel=AHOlearning [Accessed: 7th June 2015]

BBC News (2000) *Scientist 'clone' monkey* [Online] Available from: http://news.bbc.co.uk/1/hi/sci/tech/602027.stm [Accessed: 28th June 2015]

BBC News (2015a) *BBC News* [Online] Available from: http://www.bbc.co.uk/news/science_and_environment [Accessed 28th July 2015]

BBC News (2015b) *BBC News* [Online] Available from: http://www.bbc.co.uk/news/technology [Accessed 28th July]

BEAMS (2007) *BT's 'Soul Catcher2025' Implants* [Online] Available from: http://www.beamsinvestigations.org/BT's%20'Soul%20Catcher%202025'%20Implants.htm [Accessed 1st June 2015]

Begich, N., & Manning, J., (1997) *ANGELS DON'T PLAY THIS HAARP. Advances in Tesla Technology* [Online] Available from: http://www.alachuacounty.us/Depts/epd/EPAC/Angels%20Dont%20Play%20This%20HAARP%20by%20Nick%20Begich%201997.pdf [Accessed 15th June 2015]

Berkeley News (2011) *Scientists use brain imaging to reveal the movies in our mind* [Online] Available from: http://news.berkeley.edu/2011/09/22/brain-movies/ [Accessed: 22nd July 2015]

Beter, P., (2011) *"Organic Robotoids are real" by Dr Peter Beter* [Online video]. April 15th 2011. Available from: https://www.youtube.com/watch?v=nc0m5UMPwtU&ab_channel=youlittlerocket [Accessed 6th July 2015]

Bloomberg Business (2015) *See Future of Artificial Intelligence in Mind Clones Right Now!* [Online video] Available from: https://www.youtube.com/watch?v=4bqZp9TPYVk&ab_channel=BloombergBusiness [Accessed 27th June 2015]

Borghino, D., (2012) *"Avatar" project aims for human immortality by 2045* [Online] Available from: http://www.gizmag.com/avatar-project-2045/23454/ [Accessed 3rd June 2015]

Boringest (2006) *Bush Video 10 years ago!* [Online video]. January 3rd 2006. Available from: https://www.youtube.com/watch?v=pw4Bhmm22xo&ab_channel=boringest [Accessed: 11th May 2015]

BritneySpearsVevo (2009) *Britney Spears –Break The Ice* [Online video] October 24th 2009. Available from: https://www.youtube.com/watch?v=eQFIKP9rGhQ&ab_channel=BritneySpearsVEVO [Accessed 18th July 2015]

BritneySpearsVevo (2011) *Britney Spears –Hold it Against me* [Online video] February 17th 2011. Available from: https://www.youtube.com/watch?v=-Edv8Onsrgg&ab_channel=BritneySpearsVEVO [Accessed 18th July 2015]

Campbell, T., (2008) *Physics, Metaphysics and the nature of Consciousness* [Dr Thomas Campbell –My Big TOE (1 of 18)] [Online video] May 25th 2008. Available from: https://www.youtube.com/watch?v=MxECb7zcQhQ&list=PLBFFCEB1CAEDF9E6C&ab_channel=akn0ledge [Accessed 4th July 2015]

Carmichael, J., (2013) *Mouse Cloned From A Mere Drop of Blood* [Online] Available from: http://www.popsci.com/science/article/2013-06/mouse-cloned-mere-drop-blood [Accessed 19th July 2015]

CBS (2008) *Regeneration of Cells – Regrowing finger* [Online video] May 15th 2010. Available from: https://www.youtube.com/watch?v=ITxx2sOLW2Y&ab_channel=Weltenspur2 [Accessed: 25th June 2015]

Cherry, K., (2015) *What is Consciousness* [Online] Available from: http://psychology.about.com/od/statesofconsciousness/f/consciousness.htm [Accessed 1st July 2015]

CreativeForceVideo (2014) *Speaking with Nostradamus, Quantum Hypnosis with Delores Cannon* [Online video] August 2nd 2014. Available from: https://www.youtube.com/watch?v=CBOml6LTbpM&ab_channel=CreativeForceVideo [Accessed 25th July 2015]

Crystal Links (2015) *Nostradamus* [Online] Available from: http://www.crystalinks.com/nostradamus.html [Accessed 25th July 2015]

CTForecaster (2013) *Japanese Dream Recording Machine –Update* [Online video]. April 12th 2013. Available from: https://www.youtube.com/watch?v=gQueU9a8URw&ab_channel=CTForecaster [Accessed: 22nd July 2015]

Daily Mail Online (2015) *Mail Online* [Online] Available from: http://www.dailymail.co.uk/sciencetech/index.html [Accessed 28th July 2015]

De Houwer, J., Thomas, S., & Baeyens, F. (2001) *Associative Learning of Likes and Dislikes: A Review of 25 years of Research on Human Evaluative Conditioning*. Psychological Bulletin, Vol. 127, No.6, 853-869 In Hale, J., (2012) *What influences are Food Likes and Dislikes?* [Online] Available from: http://psychcentral.com/blog/archives/2012/02/15/what-influences-our-food-likes-and-dislikes/ [Accessed 28th June 2015]

Dictionary Reference (2015) *Consciousness* [Online] Available from: http://dictionary.reference.com/browse/consciousness?s=t [Accessed 30th June 2015]

Disclose TruthTV (2015) *Delores Cannon on WW3, Antichrist Nostradamus Prophecies [Full Video]* [Online video] March 19th 2015. Available from: https://www.youtube.com/watch?v=l0aCHrY0ObI&ab_channel=DiscloseTruthTV [Accessed 26th July 2015]

Donald Marshall Proboards (2015) *Donald Marshall Proboards* [Online] Available from: http://donaldmarshall.proboards.com/ [28th July 2015]

Ehrsson, H.H., (2013) *Inspirational Lecture –Professor Henrik Ehrsson* [Online video] October 3rd 2013. Available from: https://www.youtube.com/watch?v=iR7HissYN2U&ab_channel=karolinskainstitutet [Accessed 2nd July 2013]

Epperson, A. R., (1990) *The New World Order*, Publius Press. [Online] Available from: https://ia700406.us.archive.org/27/items/TheNewWorldOrder_342/TheNewWorldOrder.pdf [Accessed 7th July 2015]

Fox 4 News –Dallas Fort Worth (2014) *Aug. 15, 1988 – George W. Bush –KDFW* [Online video]. June 23rd 2014. Available from: https://www.youtube.com/watch?v=zwrl2axvPmY&ab_channel=FOX4News-Dallas-FortWorth [Accessed: 15th May 2015]

FW: Thinking (2014) *Erase and Restore your memories* [Online video]. Jun 18th 2014. Available from: https://www.youtube.com/watch?v=PpzzD_103jc&ab_channel=FW:Thinking [Accessed 8th July 2015]

Galactic (2012a) *Conversations With Nostradamus Volume One* [Online] Available from: http://galactic.no/rune/DoloresCannon_books/Dolores-Cannon-Conversations-With-Nostradamus-Volume-1.pdf [Accessed 25th July 2015]

Galactic (2012b) *Conversations With Nostradamus Volume Two* [Online] Available from: http://galactic.no/rune/DoloresCannon_books/Dolores-Cannon-Conversations-With-Nostradamus-V2.pdf [Accessed 25th July 2015]

Gizmag (2015) *Gizmag* [Online] Available from: http://www.gizmag.com/ [Accessed 28th July 2015]

Greenberg, P., (2013) *No More Bad Flashback: Scientist Find Gene That Erases Memories* [Online] Available from: http://mashable.com/2013/09/25/erase-bad-memories/ [Accessed 11th June 2015]

Hale, J., (2012) *What influences are Food Likes and Dislikes?* [Online] Available from: http://psychcentral.com/blog/archives/2012/02/15/what-influences-our-food-likes-and-dislikes/ [Accessed 28th June 2015]

Halliday, P., (2013) *Whistleblowing: the new 'public interest' test and other developments.* [Online] Available from: http://www.11kbw.com/uploads/files/PHPaper.pdf [Accessed 8th May 2015], p. 2

Hewitt, J., (2012) *How to create a mind or die trying* [Online] Available from: http://www.extremetech.com/extreme/141507-how-to-create-a-mind-or-die-trying [Accessed 24th May 2015]

History (2015) *George W. Bush* [Online] Available from: http://www.history.com/topics/us-presidents/george-w-bush [Accessed 14th May 2015]

Hostel (2006) Film. Directed by Eli Roth. [DVD]. USA: Lionsgate

Hostel: Part II (2007) Film. Directed by Eli Roth. [DVD]. USA: Lionsgate

Hostel: Part III (2011) Film. Directed by Scott Spiegel. [DVD]. USA: Sony Pictures Home Entertainment

Inception (2010) Film. Directed by Christopher Nolan. [DVD]. UK: Warner Bros. Pictures

inifiniLor (2013) *Appeal from Survivors of Canadian Genocide* [Online video]. April 8th 2013. Available from: https://www.youtube.com/watch?v=cVYkctM1k90&ab_channel=infiniLor [Accessed 13th July 2015]

Jeanice Barcelo (2013) *BNE Radio Show w/ guest Donald Marshall* [Online video]. March 9th 2013. Available from: https://www.youtube.com/watch?v=3uzgu4ekT3c&ab_channel=JeaniceBarcelo [Accessed 11th July 2015]

Jim Cristea (2009) *Mind Reading - FMRI - Machine that Reads Your Thoughts - 60 Minutes* [Online video]. January 9th 2009. Available from: https://www.youtube.com/watch?v=Cwda7YWK0WQ&ab_channel=JimCristea [Accessed: 22nd July 2015]

Jones, A., (2008) *Reflections and Warnings: An Interview with Aaron Russo* [Online video]. June 1st 2009. Available from: https://www.youtube.com/watch?v=YGAaPjqdbgQ&ab_channel=rohstyles23 [Accessed: 12th May 2015]

KafkaWinstonWorld (2014) *IS THIS WHERE WE ARE GOING? THIS MOVIE WILL BLOW YOUR F%$NG MIND(mirrored)* [Online video]. December 11th 2014. Available from: https://www.youtube.com/watch?v=4mUII1HcsRg&ab_channel=KafkaWinstonWorld [Accessed 15th July 2015]

Kim, M., (2013) *MIT scientists implant a false memory into a mouse's brain* [Online] Available from: http://www.washingtonpost.com/national/health-science/inception-mit-scientists-implant-a-false-memory-into-a-mouses-brain/2013/07/25/47bdee7a-f49a-11e2-a2f1-a7acf9bd5d3a_story.html [Accessed 28th July 2015]

LawOfIdentity (2014) *"We Will Become Gods"* [Online video]. August 20th 2014. Available from: https://www.youtube.com/watch?v=5MmXHMaati0&ab_channel=LawOfIdentity [Accessed 23 July 2015]

Legal-dictionary (2015) *Habeas Corpus* [Online] Available from: http://legal-dictionary.thefreedictionary.com/habeas+corpus [Accessed 14th May 2015]

Mark Dice (2014) *Top Transhumanist Claims He Will Become God and Kill Anyone Who Tries to Stop Him!!!* [Online video] May 14th 2014. Available from: https://www.youtube.com/watch?v=KPJARo-5VXE&ab_channel=MarkDice [Accessed 24th July 2015]

Medical News Today (2015) *Paranoid Schizophrenia: Causes, Symptoms and Treatments* [Online] Available from: http://www.medicalnewstoday.com/articles/192621.php [Accessed 21st July 2015]

Megadeth (2001) *The World Needs a Hero* [CD] USA: Sanctuary Records Group Ltd

MK Ultra Compendium (1980) *Secret CIA Human Experiments in the United States: MK Ultra Mind Control Research Program.* [online video] May 26th 2012. Available from: https://www.youtube.com/watch?v=c4f9Hs0s1jQ&ab_channel=TheFilmArchives [Accessed: 22nd June 2015]

Mind-Computer (2012) *Synthetic telepathy "Artificial Telepathy"* [Online] Available from: http://mind-computer.com/2012/05/15/synthetic-telepathy-artificial-telepathy/ [Accessed 22nd June 2015]

Moss, S., (2009) *Evaluative conditioning* [online] Available from: http://www.psych-it.com.au/Psychlopedia/article.asp?id=312 [Accessed 28th June 2015]

Motherboard (2015) *Motherboard* [Online] Available from: http://motherboard.vice.com/en_uk [Accessed 28th July 2015]

MrCowShedder (2012) *Royal Babylon by Heathcote Williams (rough cut).* [Online video]. May 11th 2012. Available from: https://www.youtube.com/watch?v=jIukrdRhnpw&ab_channel=MrCowshedder [Accessed 14th July 2015]

MurdokDubstep (2010) *Where'd You Go – Fort Minor (Murdok Dubstep Remix)* [Online video]. July 22nd 2010. Available from: https://www.youtube.com/watch?v=7VdAvIf1Nc4&ab_channel=MurdokDubstep [Accessed 15th July 2015]

nature video (2013) *Reading minds* [Online video]. October 23rd 2013. Available from: https://www.youtube.com/watch?v=z8iEogscUl8&ab_channel=naturevideo [Accessed: 22nd July 2015]

New Scientist (2014) *Brain Decoder can eavesdrop on your inner voice* [Online] Available from: https://www.newscientist.com/article/mg22429934-000-brain-decoder-can-eavesdrop-on-your-inner-voice/ [Accessed 12th June 2015]

New Scientist (2015) *New Scientist* [Online] Available from: https://www.newscientist.com/ [Accessed 28th July 2015]

Non Mirage Truth Vision (2015) *CIA's Heart Attack Gun* [Online video]. Available from: https://www.youtube.com/watch?v=Uwy56QTV4cs&ab_channel=NonMirageTruthVision [Accessed 16th July 2015]

Open Minds (2011) *Phil Schneider's incredible ET claims.* [Online] Available from: http://www.openminds.tv/phil-schneiders-incredible-et-claims/9982 [Accessed 9th May 2015]

Paye, J-C., (2013) *The Suspension of Habeas Corpus in America* [Online] Available from: http://www.globalresearch.ca/the-suspension-of-habeas-corpus-in-america/5311701 [Accessed 20th May 2015]

Petkova, V. I., and Ehrsson, H.H., (2008) *If I Were You: Perceptual Illusion of Body Swapping,* PLoS ONE, Volume 3, Issue 12, pp. 1-9

Prigg, M., (2014) *Mindreading software could eavesdrop on your secret inner voice* [Online] Available from: *http://www.dailymail.co.uk/sciencetech/article-2814896/The-mindreading-machine-listen-voices-head-let-paralysed-speak-again.html* [Accessed 13th June 2015]

Project Camelot (2008a) *Project Camelot Interviews George Green – Part 1 of 2* [Online video]. April 16th 2008. Available from: https://www.youtube.com/watch?v=sSYXrWIA618&ab_channel=ProjectCamelot [Accessed: 11th May 2015]

Project Camelot (2008b) *Project Camelot Interviews George Green – Part 2 of 2* [Online video]. April 16th 2008. Available from: https://www.youtube.com/watch?v=6zSrg0IxHzI&ab_channel=ProjectCamelot [Accessed: 11th May 2015]

RainmanJhof (2011) *Boondox-Abadon* [Online video]. October 31st 2011. Available from: https://www.youtube.com/watch?v=C2bH4B5_83Q&ab_channel=RainmanJhof [Accessed 15th July]

Radar Online (2013) *Princess Dianna 'Killed By Bright Light Shone By Special Forces Soldiers Into Car She Was In'* [Online] Available from: http://radaronline.com/exclusives/2013/09/princess-diana-killed-died-death-forces-light-special-air-service/ [Accessed 17th July 2015]

Rense (2001) *Microchip Mind Control, Implants and Cybernetics* [Online] Available from: http://www.rense.com/general17/imp.htm [Accessed 20th June 2015]

Rense (2011) *Possible HAARP Locations Around the World* [Online] Available from: http://www.rense.com/general92/haarp.htm [Accessed 20th July 2015]

Russo, A., (1996) *Mad as Hell* [Online video]. May 10th 2011. Available from: https://www.youtube.com/watch?v=Zz4TI75MszQ&ab_channel=XRepublicTV [Accessed: 13th May 2015]

Russo, A., (2006) *Freedom to Fascism* [Online video]. May 10th 2011. Available from: https://www.youtube.com/watch?v=uNNeVu8wUak&ab_channel=RevolutionistsUnited [Accessed: 12th May 2015]

RT (2014) *Revolutionary way to 'switch off' pain discovered* [Online] Available from: http://www.rt.com/news/210031-revolutionary-painkiller-discovered-scientists/ [Accessed 19th July 2015]

RT (2015) *Living forever as a robot? Prototype lets humans upload their mind into mechanised 'heads'* [Online] Available from: http://www.rt.com/usa/229811-mind-clones-robot-afterlife/ [Accessed 27th June 2015]

SadSongs4You (2010) *Elton John –Candle In The Wind. With lyrics* [Online video]. August 23rd 2010. Available from: https://www.youtube.com/watch?v=80rHyABCb20&ab_channel=SadSongs4You [Accessed 17th July 2015]

Schechtman (2012) *The Story of my (Second) Life: Virtual Worlds and Narrative Identity,* Philosophy and Technology, Volume 25, Issue 3, pp. 329-343

Schneider, P., (1995) *Phil Schneider Documentary of truth about Aliens & UFO's & our Government*. [Online video]. September 21st 2013. Available from: https://www.youtube.com/watch?v=Oljrjxnixtw&ab_channel=AliensAmongUs [Accessed: 10th May 2015]

Schneider, P., (1996) *Phil Schneider's Last Speech ~ Two Months Before His Assassination ~ Aliens & Underground Bases* [Online video]. November 24th 2013. Available from: https://www.youtube.com/watch?v=Slgb5U-OqFM&ab_channel=FallofMedia [Accessed: 10th May 2015]

Science Channel (2014) *How to Grow a New Fingertip | World's Strangest*. [Online video] June 16th 2014. Available from: https://www.youtube.com/watch?v=DtBUM51t4iw&ab_channel=ScienceChannel [Accessed 23rd June 2015]

Science Daily (2014) *'Off switch' for pain discovered: Activating the adenosine A3 receptor subtype is key to powerful pain relief* [Online] Available from: http://www.sciencedaily.com/releases/2014/11/141126132639.htm [Accessed 19th July 2015]

Sheen, M., Begich, N., & Robbins, W., (2005) *Holes in Heaven? H.A.A.R.P. & Advances in Tesla Technology* [Online video]. October 8th 2014. Available from: https://www.youtube.com/watch?v=SWVU6DKcjyA&ab_channel=documentary2014 [Accessed: 15th June 2015]

SimpleGirl4ewer (2007) *Britney Spears – Mona Lisa* [Online video]. March 18th 2007. Available from: https://www.youtube.com/watch?v=xEecXKUxl1s&ab_channel=SimpleGirl4ewer [Accessed 18th July 2015]

Sleepdex (2015) *Stages of Sleep* [Online] Available from: http://www.sleepdex.org/stages.htm [Accessed: 5th June 2015]

Star Wars Episode III: Revenge of the Sith (2005) Film. Directed by George Lucas. [DVD]. USA: 20 Century Fox

Stromberg, J., (2013) *Scientist Figure Out What You See While You're Dreaming* [Online] Available from: http://www.smithsonianmag.com/science-nature/scientists-figure-out-what-you-see-while-youre-dreaming-15553304/?no-ist [Accessed 22nd July 2015]

The 6th Day (2000) Film. Directed by Roger Spottiswoode. [DVD]. USA: Columbia Pictures

The Guardian (2015a) *The Guardian* [Online] Available from: http://www.theguardian.com/science [Accessed 28th July 2015]

The Guardian (2015b) *The Guardian* [Online] Available from: http://www.theguardian.com/uk/technology [Accessed 28th July 2015]

The Independent (2015a) The Independent [Online] Available from: http://www.independent.co.uk/news/science/ [Accessed 28th July 2015]

The Independent (2015b) The Independent [Online] Available from: http://www.independent.co.uk/life-style/gadgets-and-tech/ [Accessed 28th July 2015]

The Island (2005) Film. Directed by Michael Bay. [DVD]. USA: DreamWorks Pictures

The Manchurian Candidate (2004) Film. Directed by Jonathan Demme. [DVD]. USA: Paramount Pictures

TheNanoAge (2015) *Transhumanism and Posthumanism: The Future of Us (Humanity Plus)* [Online] Available from: *http://www.thenanoage.com/transhumanism-posthumanism.htm#transhuman* [Accessed: 30th May 2015]

Truthstream (2006) *Rigged USA Elections Exposed.* March 2nd 2006. Available from: https://www.youtube.com/watch?v=JEzY2tnwExs&ab_channel=truthstream [Accessed: 22th May 2015]

UC Berkeley Campus Life (2011) *Vision Reconstruction* [Online video]. December 11th 2011 Available from: https://www.youtube.com/watch?v=6FsH7RK1S2E&feature=youtu.be&ab_channel=UCBerkeleyCampusLife [Accessed 22nd July 2015]

Vincent Eastwood (2013) *Illuminati Cloning Programs, Sex and Murder Cults and Reptilians! 26Feb2013* [Online video] February 26th 2013. Available from: https://www.youtube.com/watch?v=M_1UiFeV5Jg&ab_channel=VincentEastwood [10th July 2015]

Walcutt, D. L., (2013) *Stages of Sleep*. [Online] Available from: http://psychcentral.com/lib/stages-of-sleep/ [Accessed 5th June 2015]

Walther, E., Nagengast, B., & Trasselli, C., (2005). *Evaluative conditioning in social psychology: Facts and speculations.* Cognition and Emotion, 19, 175–196 In Moss, S., (2009) *Evaluative conditioning* [online] Available from: http://www.psych-it.com.au/Psychlopedia/article.asp?id=312 [Accessed 28th June 2015]

WhiteLiesVevo (2010) *White Lies – Bigger Than Us* [Online video]. November 18th 2010 Available from: https://www.youtube.com/watch?v=JW0yynlDmqQ&ab_channel=WhiteLiesVEVO [Accessed 18th July]

Winter, L., (2014) *Scientist Erase Memories with Light* [Online] Available from: http://www.iflscience.com/brain/scientists-erased-memories-light [Accessed 10th June 2015]

WorldTruth (2014) *The Georgia Guidestones* [Online] Available from: http://worldtruth.tv/the-georgia-guidestones-2/ [Accessed 16th May 2015]

YesEthan (2013) *Consciousness Science Kept Hidden* [Online Video] January 12th 2013. Available from: https://www.youtube.com/watch?v=LFSRTsLOiv0&ab_channel=YesEthan [Accessed 4th July 2015]

Legislation

Computer Misuse Act (1990) Section 3A, *Making, supplying or obtaining articles for use in offence under section 1 or 3* [Online] Available from: http://www.legislation.gov.uk/ukpga/1990/18/section/3A [Accessed 8th May 2015]

Enterprise and Regulatory Reform Act (2013) Section 17, *Disclosures not protected unless believed to be made in the public interest.* [Online] Available from: http://www.legislation.gov.uk/ukpga/2013/24/section/17/enacted [Accessed 8th May 2015]

Employment Rights Act (1996) Section 43B, *Disclosures qualifying for protection.* [Online] Available from: http://www.legislation.gov.uk/ukpga/1996/18/section/43B [Accessed 8th May 2015]

Public Interest Disclosure Act (1998) Section 43B, *Disclosures qualifying for protection.* [Online] Available from: http://www.legislation.gov.uk/ukpga/1998/23/section/1 [Accessed 8th May 2015]

Serious Crime Act (2015) Section 41 3ZA, *Unauthorised acts causing, or creating risk of, serious damage.* [Online] Available from: http://www.legislation.gov.uk/ukpga/2015/9/section/41/enacted [Accessed 8th May 2015]

Tetryonic Theory - an overview (1) 2
Tetryonic theory..... challenging and correcting the foundations of quantum th 13
tetryonics merged pdf 21
Response to an analysis of Tetryonic Chemistry by Robert J Brotherus (1) 1675
Planck mass-energy templates [Tetryonic theory] 1694

Tetryonic Theory

The unified quantum field geometry of charged EM mass-ENERGY-Matter in motion

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them." *[Albert Einstein]*

Given the range and applied scope of Quantum Physics in today's modern technologic World there remains a driving desire to rationalise our numerous disparate scientific theories into one coherent theory with a intuitive model that can be applied equally to the Quantum and Cosmological scales of our Universe.

Such a theory would need to preserve the currently observed outcomes and present established theories in a new light, offering additional testable predictions of its own, and ideally do so in a manner that is simpler than that of the established theories and hypothesises.

Many foundational properties of Quantum Mechanics remain unaddressed by scientific theory and in the following pages an overview of the key quantum properties challenging our current scientific advancement will be highlighted, including a number of assumptions that currently impede the development of a fully realised, coherent solution to all of our current scientific questions.

While Mathematics is the language of Science it remains a language that lacks a well-defined physical model on which to test it and further its many and varied solutions to Quantum & Cosmological scale physics. It is this lack of any rigid, enforceable GEOMETRY that has allowed the development of numerous disjointed statistical and probabilistic solutions to physical problems, in turn impeding our scientific understanding and advancement of quantum processes.

The Standard Model has many observed and testable components to it but more recently new theories have emerged to contest it without being rigorously testable themselves. They rely on the established foundation provided by the Standard Model but try to explain its various deficiencies ad-hoc with various suppositions without any solid footing of their own.

Often the only way to progress further in our scientific endeavours is to retrace our footsteps in science and to develop new physical models on which we can discern our known results and observations thus excluding any false or misleading assumptions (mathematical of otherwise).

In doing so there exists the promise that a simple underlying foundation can be found to the physics of our Universe, revealing new and exciting advances in Science that will allow us to usher in a new age of scientific and technological advancement for the betterment of humanity as a whole.

'Tetryonics – The charged geometry of EM mass-ENERGY-Matter' whose founding principle is that **EQUILATERAL Planck energies are** the foundation geometry for all quantum mechanical processes is presented here as just such a solution to the current quandaries of Quantum Mechanics.

ENERGY

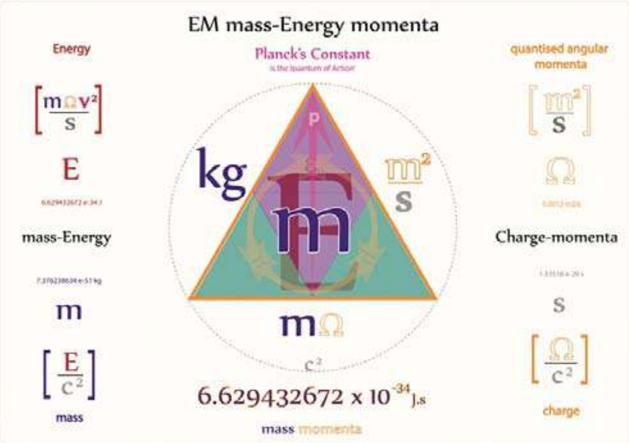
EM mass-Energy momenta
Energy
Planck's Constant
quantised angular momenta
m Ω v²
s
E
kg
m²
s
Ω
mass-Energy
Charge-momenta
m
s
m Ω
E
c²
Ω
c²
c²
6.629432672 x 10⁻³⁴ J.s
mass
mass momenta
charge
p

Energy, in Physics, is an indirectly observed quantity of a system that imbues it with the ability to exert a Force or do Work over a distance.

It has been measured and quantified through numerous methodologies over the centuries, most notably through its associated characteristics such as mass, velocity and ElectroMagnetic fields.

The most recent attempt to quantify its characteristics (with respect to heat and light) led to the discovery of Planck's constant and the development of Quantum Mechanics

The application of a simple proposition (that Energy has an equilateral geometry) opens the door on a greater understanding of the mechanics of the quantum world, a realm that will be forever beyond the reach of our physical eyes.

The myriad of perplexing properties of quantum properties (such as Charge, mass & Matter) and the astonishing outcomes of quantum experiments (Interference and Wave-Particle duality) can now all be readily modelled and explained rationally on a solid geometric footing.

Paving the way for new discoveries and a greater understanding of our Universe and its mechanics

Quantised Angular Momenta

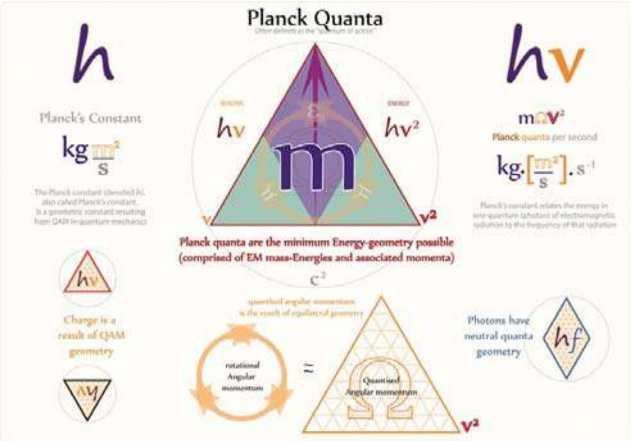
Planck Quanta
h
hv
Planck's Constant
kg m²/s
hv
hv²
m Ω v²
Planck quanta per second
kg.[m²/s].s⁻¹
m
v
v²
Planck quanta are the minimum Energy-geometry possible
(comprised of EM mass-Energies and associated momenta)
c²
hv
Charge is a result of QAM geometry
rotational Angular momentum
≈
Quantised Angular momentum
Ω
v²
Photons have neutral quanta geometry
hf

Building on the priori definition of equilateral energy quantisation it can be quickly demonstrated that any energy measurement per unit of Time results in a geometric unit of measurement that has historically be taken to represent rotational motion (specifically Angular momentum)

This unit of motion is found throughout Quantum Mechanics and is directly related to the square energy levels in nuclear processes.

The equilateral EM fields of energy quanta are constrained by its geometry and this geometry lies at the heart of understanding the quantum world in all its beauty.

It determines properties such as Electric permittivity and Magnetic permeability, the vector direction of linear momentum and relates the scalar property of electromagnetic mass directly to velocity.

CHARGE

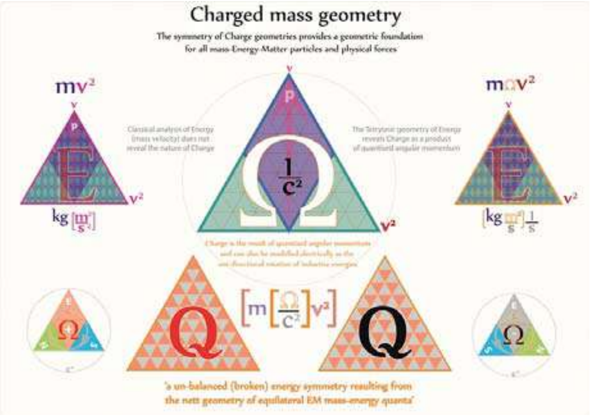
Charged mass geometry
The symmetry of Charge geometries provides a geometric foundation for all mass-Energy-Matter particles and physical forces
mv²
m□v²
E
p
v
v²
kg [m²/s²]
[kg m²/s] 1/s
Classical analysis of Energy (mass velocity) does not reveal the nature of Charge
The Tetryonic geometry of Energy reveals Charge as a product of quantised angular momentum
Ω
1/c²
Q
[m[Ω/c²]v²]
'a un-balanced (broken) energy symmetry resulting from the nett geometry of equilateral EM mass-energy quanta'

A fundemental property of all mass-ENERGY-Matter that gives rise to the Forces of electrical Interaction has eluded physical explanation by Physicists since its discovery and led to many 'band aid' additions to the Standard model in attempts to explain its observed properties.

The most recent attempt at an explanation was Special relativity theory which in turn led to our current models of charge from contractions of Matter due to motion.

SR explains Charge as an invariant property of of electrostatic bodies and that the motion of charges creates an additional magnetic moment through the relativistic distortion of spherical bodies of Matter.

Equilateral Planck energy momenta geometries and Matter topologies offer a completely different explanation for the source of electrostatic charges and their associated magnetic moments.

Positive and Negative charges are revealed to be opposite side of the same quantum energy 'quoin', and can be modelled electrically as ideal quantum inductive loops. It is equilateral geometry that gives rise to the physical properties of quantised angular momenta, inertial mass, elemental charges and even the geometric topology of Matter itself.

In a planar 2D form they form a neutral EM energy momenta geometry but it is when they form the topology of 3D Matter they are expressed as either Positive or Negative electric charge fascia.

ODD number energy quanta [W Bosons] combine via their magnetic bases in fixed quantum steps to create the 'squared' nuclear energy levels so familiar to quantum mechanics and form the basis for ElectroMagnetic Induction.

It is their rigid equilateral geometry that provides the basis for relativistic charge invariance and the quantum mechanism for photo-electron transitions in atoms while the net energy momenta quanta in charged geometries provides the electromotive force known as Voltage.

Equilateral geometry reveals an intrinsic connection between EM Energy, mass and Charge.

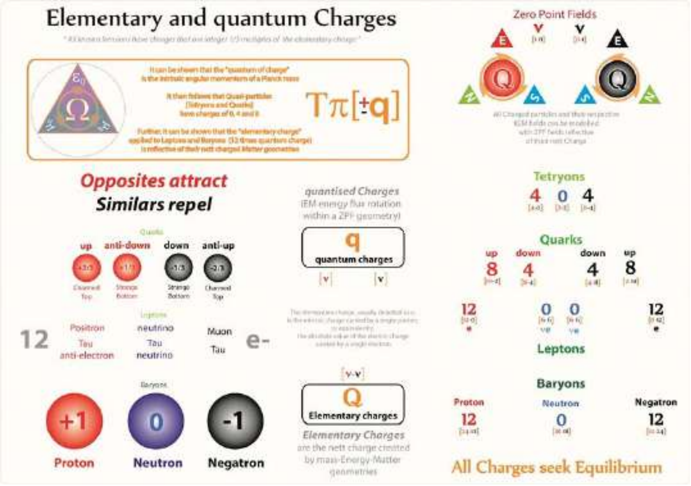
Elementary and quantum Charges
Tπ[±q]
Zero Point Fields
Q
Opposites attract
Similars repel
quantised Charges
(EM energy flux rotation within a ZPF geometry)
q
quantum charges
Tetryons
4 0 4
Quarks
up
anti-down
down
anti-up
Charmed Top
Strange Bottom
Strange Bottom
Charmed Top
up
down
down
up
8 4 4 8
12
Positron
Tau anti-electron
neutrino
Tau neutrino
Muon
Tau
e-
12 0 0 12
Leptons
Baryons
Q
Elementary charges
Elementary Charges are the nett charge created by mass-Energy-Matter geometries
+1
0
-1
Proton
Neutron
Negatron
Proton 12
Neutron 0
Negatron 12
All Charges seek Equilibrium

As separated charges seek equilibrium they provide motive energies & the quantum scaffolding for large scale Matter topologies and their force interactions throughout our Universe.

mass-ENERGY-Matter

All Energy has an EM mass equivalence and equally any object with EM mass has energy equivalence

But there has never been a formal scientific definition and associated equation to distinguish the property of EM mass from that of Matter.

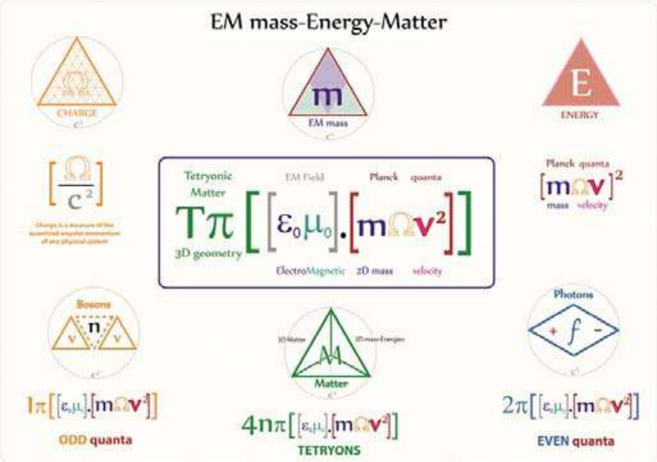
EM mass-Energy-Matter
CHARGE
c^2
m
EM mass
E
ENERGY
$\left[\frac{\Omega}{c^2}\right]$
Tetryonic Matter
EM Field
Planck quanta
$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
3D geometry
ElectroMagnetic
2D mass
velocity
Planck quanta
$[m\Omega v]^2$
mass
velocity
Bosons
n
v
v
Matter
Photons
f
$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
ODD quanta
$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
TETRYONS
$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
EVEN quanta

Using Tetryonic geometry it can be clearly demonstrated that ElectroMagnetic mass is NOT Matter – it is the Energy content of a system per unit of time [E / c^2] in total agreement with Einstein (and all before him) however, lacking a formal definition of either the two terms have become mired with each other so much that they are often used for each other inappropriately even in scientific peer-reviewed literature.

ElectroMagnetic mass is the two dimensional (planar) measurement of Energy per unit of Time whereas Matter is a measure of the tetrahedral energies found in a spherical volume of 3D space.

Mass-energy can be viewed as the paper from which 3D Matter is created when it is folded into a tetrahedral shape (the quantum canvas covering the topology of Matter)

The often used tem of 'massless' is now shown to be a misnomer that should be removed from the scientific vocabulary except where it specifically refers to empty space (devoid of any energies).

Matter is now formally defined geometrically as $4n\pi$ tetrahedral standing wave of EM energy and it is the tetrahedral topology of Tetryons that forms the foundation for all large scale Matter in the Universe (not spherical point particles).

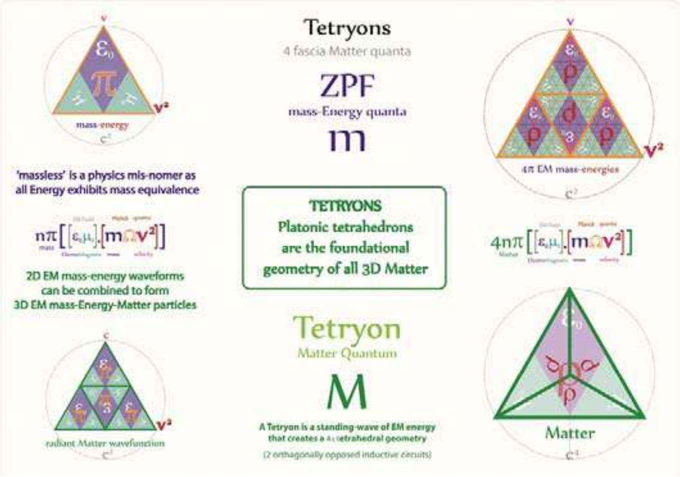
Tetryons
4 fascia Matter quanta
ZPF
mass-Energy quanta
m
v
ε_0
π
v^2
mass-energy
c^2
'massless' is a physics mis-nomer as all Energy exhibits mass equivalence
$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
2D EM mass-energy waveforms can be combined to form 3D EM mass-Energy-Matter particles
TETRYONS
Platonic tetrahedrons are the foundational geometry of all 3D Matter
4π EM mass-energies
$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
Tetryon
Matter Quantum
M
A Tetryon is a standing-wave of EM energy that creates a 4π tetrahedral geometry
(2 orthagonally opposed inductive circuits)
radiant Matter wavefunction
Matter
c^2

Charge allows EM masses to form a Tetryonic topology giving volume to Matter and its energy momenta provides the interactive forces between them imbuing motion into our Universe.

Matter at rest is comprised of EM mass-energies that are always in motion and propagating in a tetrahedral EM standing wave.

To cause at will the birth and death of matter would be man's grandest deed, which would give him the mastery of physical creation, make him fulfil his ultimate destiny."

[Nikola Tesla]

Sub-Atomic Particles

Using equilateral energy all the mass-energy geometries & mass-Matter topologies of EM fields and particles can be physically modelled revealing the known Standard Model particle sand a few more.

Bosons are shown to be transverse EM fields that facilitate EM induction and Photons are revealed as longitudinal dual-charge pairs possessing a neutral EM charge and inherent magnetic moment.

The net unidirectional momentum of Bosons then distinguishes them from Photons that possess a bi-directional momentum

Closer examination of the geometries highlights a long-standing error in the mathematical formulation of QM energy formulas, namely the mistaken interchanging of Planck's quanta [*v*] for Einstein's frequency [*f*] in relation to Energy.

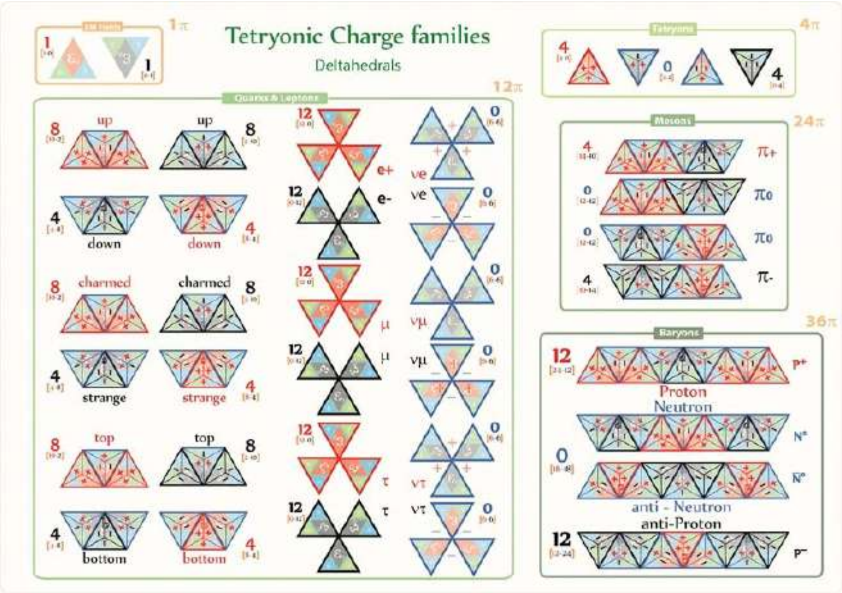
Tetryonic Charge families
Deltahedrals
1π
1
1
Quarks & Leptons
12π
8
up
up
8
4
down
down
4
8
charmed
charmed
8
4
strange
strange
4
8
top
top
8
4
bottom
bottom
4
12
e+
12
e-
12
μ
12
μ
12
τ
12
τ
0
νe
νe
0
0
νμ
νμ
0
0
ντ
ντ
0
Tetryons
4π
4
0
4
Mesons
24π
4
π+
0
π0
0
π0
4
π-
Baryons
36π
12
P+
Proton
Neutron
N°
0
N°
anti - Neutron
anti-Proton
12
P⁻

2D EM masses [bosons-photons] do not have the physical property of 3D topology possessed by Matter [Fermions]

Fermions can now be defined as the standing wave energy topologies that create all the known sub-atomic particles – the building blocks of large scale Matter and structure in the macro world.

Historically these particles have been classified according to their charges and masses, Tetryonic geometry now provides a 3D physical model of all the particles highlighting the physical source of the 1/3 charges of Quarks and revealing charge to be the foundational geometry of all Matter.

Tetryons are 4π charge topologies that are the foundational quanta of Matter and surprisingly have a mass-charge ratio identical to that of Leptons explaining how these particles have remained hidden from accelerator experiments.

Quarks are 12π charge geometries where the attractive strong charge interactions between their fascia form 8π particle topologies with entirely different properties to that of Leptons

Leptons also have 12π charge geometries but with repulsive fascia that result in them forming the quantum equivalent of a 6 loop rotor

Protons and Neutrons are 36π charge geometries [that result in 20π Baryon topologies] and have identical masses (against the current model of baryonic masses which derived from an entirely different process where Neutrons are formed by Protons absorbing electrons).

The Particle Zoo

Baryons – the building blocks of atomic nuclei can be formed from many tri- quark combinations and energy levels and this is of particular note when modern accelerators are used to probe atomic structures.

The higher energies they employ when colliding particles together results in higher 2D kinetic energies (and increased energy quanta) which recombine after collisions into a plethora of similar particles called the Particle Zoo.

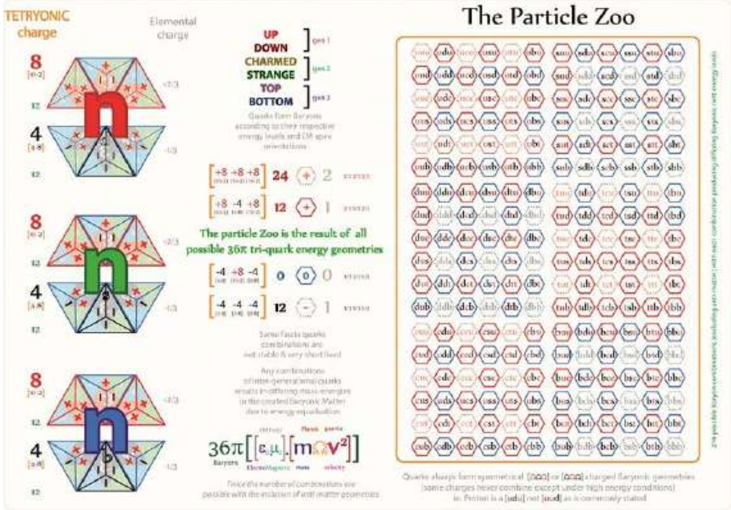
TETRYONIC charge
Elemental charge
UP
DOWN
CHARMED
STRANGE
TOP
BOTTOM
The Particle Zoo
The particle Zoo is the result of all possible 36π tri-quark energy geometries
36π[[εμ][mv²]]

A sound understanding of EM mass-energy geometries & the charged topologies of all sub-atomic Matter particles reveals the true genesis of all of the Baryons to be found in the particle zoo.

Tetryonic geometries facilitate the precise 3D modelling of all the periodic elements, allotropes and compounds enabling the development of new classes of materials and medicines providing us with many new insights into Quantum Chemistry and large scale Matter in general.

Quantum Electro-Dynamics [QED]

Adding to the mysteries of quantum mechanics is a number of well-known QED effects that have also eluded physical explanation – to date

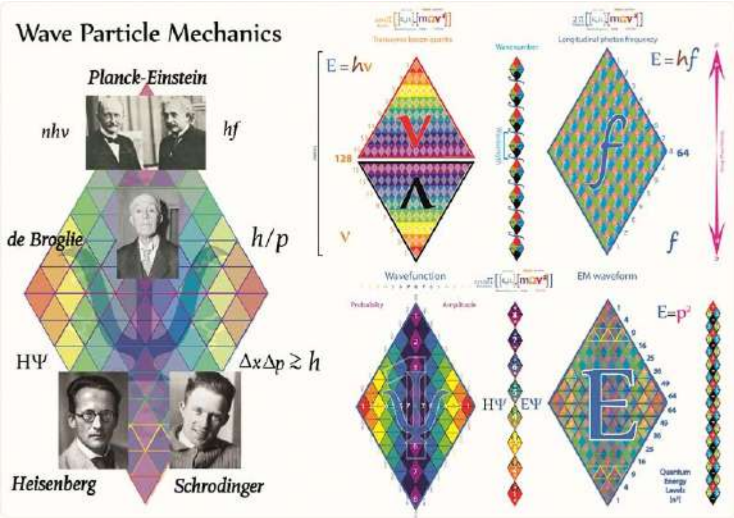
Wave Particle Mechanics
Planck-Einstein
nhν
hf
de Broglie
h/p
HΨ
ΔxΔp ≳ h
Heisenberg
Schrodinger
E = hν
E = hf
Wavefunction
EM waveform
HΨ EΨ
E=p²

In 1865, James Maxwell Clerk unified the then disparate theories of Electric and Magnetic fields into a theory of Electromagnetism and related then to the velocity of light.

The exact quantum geometry for EM wave geometries (and the photons comprising them) that explains their Wavefunctions and observed inference patterns represents one of the greatest challenges to developing a concise fully realised quantum theory of EM radiation.

Wave-Particle duality dating back to the 17th century is perhaps the best known example of quantum behaviour that has defied precise modelling despite intense effort by scientists around the World for over two centuries

The application of equilateral energy momenta to the charged geometries of Photons and EM waves not only explains these long standing mysteries but also removes the quantum fuzziness introduced by Heisenberg's Uncertainty principle, in turn clearing the way for Science to develop an advanced understanding of Electricity, its role in Quantum ElectroDynamics and provide new clean forms of energy from quantum processes.

Photo-electrons and spectral lines

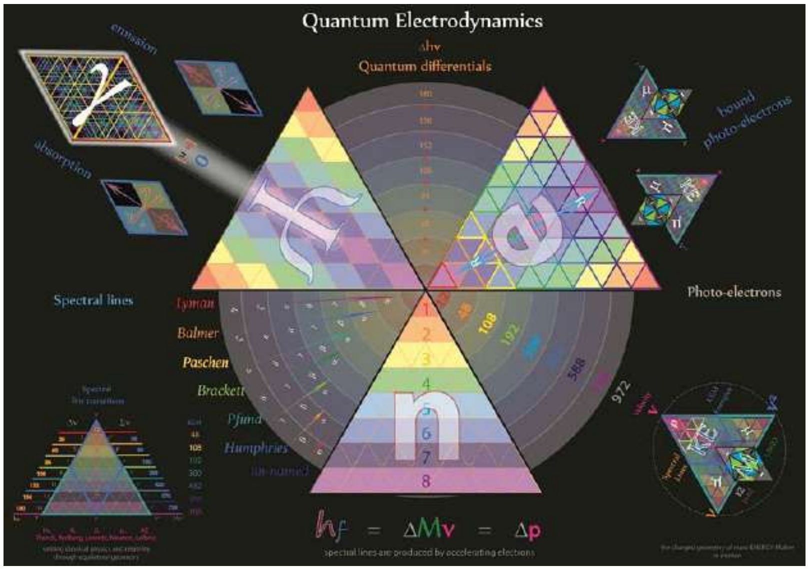
Quantum Electrodynamics
Δhv
Quantum differentials
emission
absorption
bound photo-electrons
Photo-electrons
Spectral lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
(un-named)
1
2
3
4
5
6
7
8
108
192
588
972
n
$hf = \Delta Mv = \Delta p$
spectral lines are produced by accelerating electrons

The physical relationship between Spectral lines, Rydberg's Constant and the Kinetic energies of Photo-electrons as they interact with photons and atomic nuclei is quickly revealed when equilateral energy geometries are employed

It is the rigid geometry of equilateral Planck energy momenta that gives rise to the invisible forces of Faraday's Electric and Gauss' Magnetic fields as well as Newton's Action-at-a-Distance.

Quantised Angular Momentum, a direct measure of the long hidden equilateral geometry of Energy, is revealed as the source of Charge, the physical constants and even the geometry of EM mass-energy & Matter itself.

The geometry of Nuclear Forces and Constants

Charge interactions resulting from equilateral EM energy geometries have been mistakenly developed into three disparate nuclear forces:

The EM Force is the result of $2n\pi$ charge geometries acting along transverse or longitudinal directions (or in superposition) in the forms of Bosons and Photons

The Weak Force is the inductive coupling of the magnetic permeability of adjacent energy geometries

The Strong Force is the attractive force between opposite charged fascia of Tetryonic Matter. It can also form a repulsive force between similar charged fasciae resulting in charged Leptons [electrons].

Gravitation (mathematically identical to Coulomb's Force save for strength and source) can also be modelled as the geometric mean of super-positioned EM waves.

All of which are determined by the Fine Structure Constant – the mysterious hand of GOD – setting the strength of EM interactions and determining Charge on the quantum level.

Applying equilateral geometry to energy momenta quanta, the Tetryonic model of EM mass-ENERGY-Matter quickly evolves to explain many additional electrical properties such as Voltage, Current and Power

Quantum Chemistry

Applying tetrahedral Matter topologies to quantum chemistry leads to a number of significant advances in the understanding of chemical processes most notably:

- **Accurate models for all periodic elements and their associated allotropes**
- **Advanced molecular orbitals plots**
- **Exact rest masses for all elements and**
- **A new Periodic table based on the charged topologies of Matter**

Improving upon and extending the Aufbau principle with charged geometries leads to many refinements in chemical engineering and facilitates the building of accurate 3D models of all elements, isotopes and compounds.

Every element can be analysed from the quantum level upwards revealing its exact 3D topology, quark configuration, rest mass-Matter-energies and chemical properties.

And number of misconceptions can be quickly clarified and corrected in turn advancing our understanding to the physical mechanics underlying chemical processes like chemical bonding, element family properties and the structures of complex chemical compounds.

Hydrocarbons and Carbohydrates have already been modelled using equilateral charge topology with the results matching the established chemical models of Nobel gases, Core electrons, Lewis diagrams and the most recent electron photographs of these chemical properties of atoms.

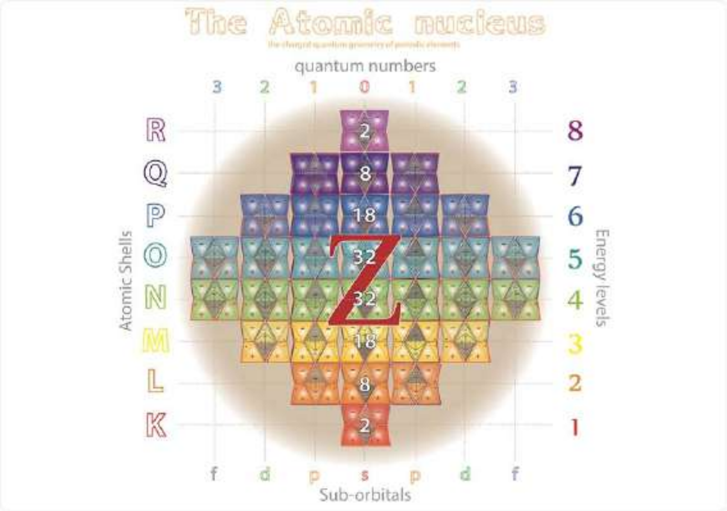

Even DNA can be modelled in 3 dimensions leading to a greater understanding of the role of quantum mechanics in biological processes and the interaction of EM radiation on organic compounds.

A highly developed quantum model of chemical elements and compounds will allow us to model chemical and biological structures in hitherto unimagined detail and facilitate the building of complex (and increasingly accurate) models of all chemical compounds and biological processes.

A Unified field equation

In physics a Unified Field Theory is a type of theory that allows all that is usually thought of as a fundamental force or elementary particle to be formulated in terms of a single equation that explains all their disparate properties.

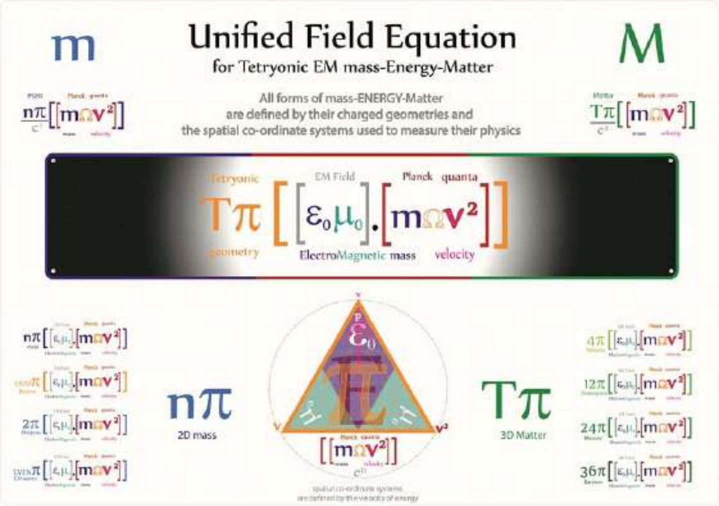
Unified Field Equation
for Tetryonic EM mass-Energy-Matter
m
M
All forms of mass-ENERGY-Matter
are defined by their charged geometries and
the spatial co-ordinate systems used to measure their physics
Tetryonic
EM Field
Planck quanta
geometry
ElectroMagnetic mass
velocity
$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
nπ
2D mass
Tπ
3D Matter
$\left[\left[m\Omega v^2\right]\right]$
4π
12π
24π
36π
2π

It is considered the holy grail of Science, offering the promise of advances in Technology, Energy and Medicine unparalleled in Human history.

Equilateral energy provides the foundation for the creation of a single equation that expresses EM mass-ENERGY and Matter as the result of geometric & topological charge interactions.

Applying equilateral geometry to Planck energy momenta quanta has been shown to firmly resolve many of the current mysteries plaguing the Standard model and provides a firm foundation for the development of a single unified quantum theory uniting all quantum and macroscopic forces.

It effortlessly merges all the tested features of Classical mechanics with the statistical probabilities of quantum mechanics and scales up to the cosmological scale of General Relativity.

Equilateral charged energy geometries provide a quantifiable foundation for advances in:

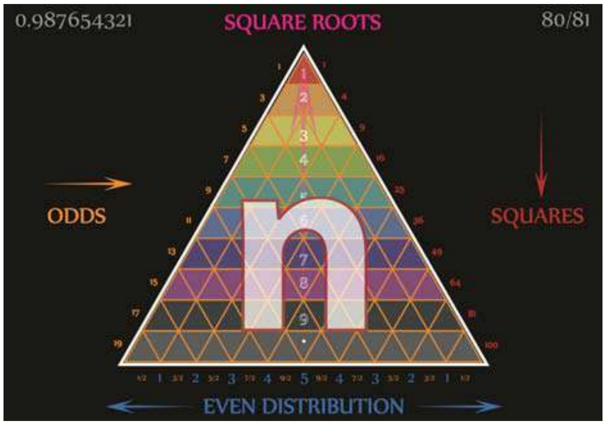
0.987654321
SQUARE ROOTS
80/81
ODDS
SQUARES
n
EVEN DISTRIBUTION

- **Quantum Mechanics**
- **Quantum Electrodynamics**
- **Quantum Chemistry**
- **Quantum Cosmology**

Explaining the quantum mysteries of mass & Matter, Wave-Particle duality and gravitational Singularities by challenging the very foundational assumptions of Relativity and the role of charge in our Universe.

Mathematical principles such as re-normalisation, probabilities and square roots of negative numbers can all be geometrically modelled through Tetryonics leading to further advances in fields such as biology, medicine, quantum computing and telecommunications

Through the charged geometry of Tetryonics many of the erroneous theories that have been developed during the last 1oo years can be falsified at last providing a solid foundation for further advances in science and mathematics.

Gravitation

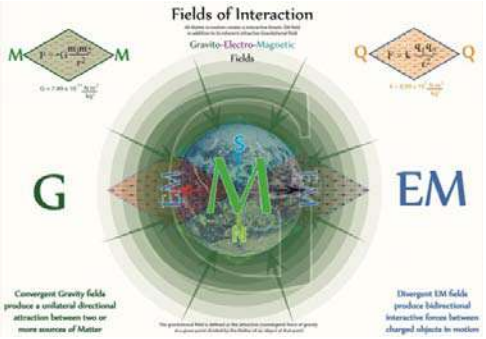
Fields of Interaction
Gravito-Electro-Magnetic
Fields
M
M
Q
Q
G
EM
S
M
N
Convergent Gravity fields
produce a unilateral directional
attraction between two or
more sources of Matter
Divergent EM fields
produce bidirectional
interactive forces between
charged objects in motion

The force of Gravity has been long-held to be the mysterious force that holds our Universe together, shaping the Stars and holding us to the Earth as we orbit our own SUN.

Tetryonic geometry reveals its quantum mechanics to be the result of electromagnetic fields and the effect produced by the geometry of Matter in these fields.

Newton saw Gravity as a force that acts instantaneously on distant bodies of Matter while Einstein dismissed the force between the bodies and explained it as the result of the 'curvature of Spacetime'.

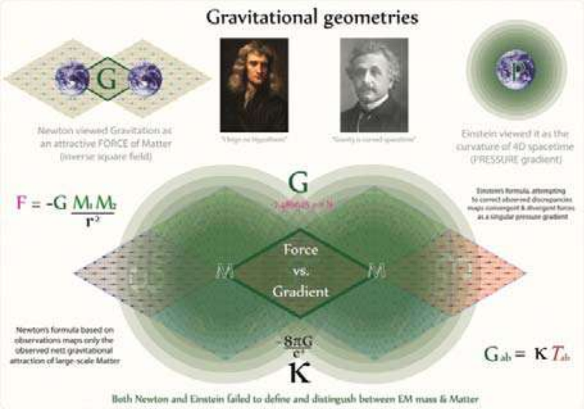
Gravitational geometries
G
Newton viewed Gravitation as
an attractive FORCE of Matter
(inverse square field)
Einstein viewed it as the
curvature of 4D spacetime
(PRESSURE gradient)
G
F = -G M₁M₂ / r²
Einstein's formula attempting
to correct observed discrepancies
maps convergent & divergent forces
as a singular pressure gradient
M
Force
vs.
Gradient
M
Newton's formula based on
observations maps only the
observed nett gravitational
attraction of large-scale Matter
-8πG / c⁴
K
Gab = K Tab
Both Newton and Einstein failed to define and distingush between EM mass & Matter

In fact the Gravitational field is comprised of 3 separate components, each of which produces differing forces of interaction, but all combine to produce the nett Gravitational field that we observe and describe as strictly convergent Gravity.

Gravity fields are created by the topology of Matter itself as it displaces and distorts the Vacuum energies surrounding it.

Electric fields diverge radially from all Matter as it is heated, or placed in motion by other forces creating an interactive field with both divergent and convergent equatorial regions of force and acceleration.

Dipole Magnetic fields located around the Polar Regions of all Matter also create a perturbative force on bodies close to them

Although Newton's and Einstein's theories were both correct in their respective analysis of the forces at work they both failed to define and distinguish between EM masses and Matter leading to the continued debate over the true nature and physical formulation of quantum gravity.

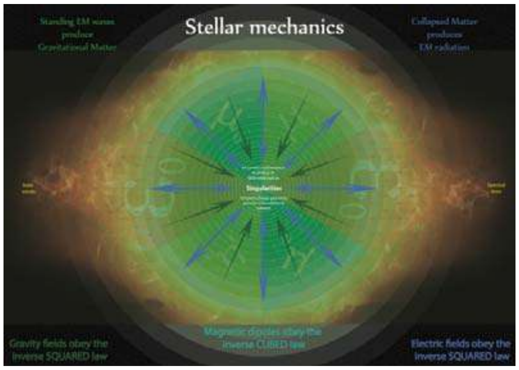
Stellar mechanics
Standing EM waves
produce
Gravitational Matter
Collapsed Matter
produces
EM radiation
Singularities
Gravity fields obey the
inverse SQUARED law
Magnetic dipoles obey the
inverse CUBED law
Electric fields obey the
inverse SQUARED law

These three quantum forces [G+E+M] combine to produce a field of gravitational acceleration that was so eloquently modelled by Newton back in the 16th century through his inverse square law.

And these 3 distinct forces also combine to produce the stellar mechanics that power the Stars – a Gravito-Electro-Magnetic pinch.

The force of Gravity, created by the geometry of Matter draws Matter toward the core of Stars.

And the GEM pinch singularity at the core of the Star destroys Matter (collapsing its 3D geometry), turning it into radiant 2D ElectroMagnetic radiation (Light).

Energy from the Stars

Virtual particles, Dark Energy and Dark Matter (were all developed to explain the observed mechanics of our Universe) and can now be revealed to be differing manifestations of the electromagnetic force interacting at various angles to, and with, the standing-wave topologies that create it.

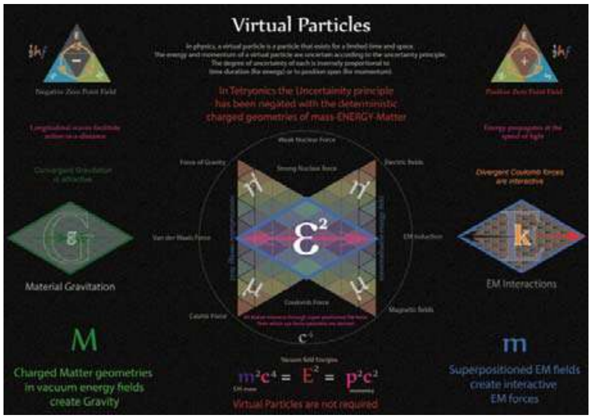
Virtual Particles
In Tetryonics the Uncertainity principle has been negated with the deterministic charged geometries of mass-ENERGY-Matter
Material Gravitation
EM Interactions
Weak Nuclear Force
Strong Nuclear Force
Coulomb Force
Van der Waals Force
ε²
k
M
Charged Matter geometries in vacuum energy fields create Gravity
m²c⁴ = E² = p²c²
Virtual Particles are not required
m
Superpositioned EM fields create interactive EM forces

Quantum fields of Planck energies combine to create charged 2D geometries of radiative ElectroMagnetic fields that in turn spread out interacting with each other to give us the fundamental laws of Attraction and Repulsion.

The 3D standing-wave topologies of Matter provide a rigid structure that displaces the surrounding Vacuum energies and creates a pressure gradient which we call Gravity.

And the destruction of these Matter topologies in turn creates radiant energies that propagate out into the surrounding space to weaken and form the Vacuum energies that permeate all of Space.

The eternal cycle of Matter creation and destruction within Stars and Galaxies is what drives the Universe we see today causing it to contract as it forms gravitational Matter and to expand as it restructures the standing-wave Matter into radiant forms of Energy (Light and heat).

The same process offers us clean, limitless energy for all our needs as well as unlimited resources from the restructuring of this energy into its varied elemental and compound forms.

Our toxic stockpiles of radioactive wastes can be completely eliminated from the environment and electrical power can be safely stored as mass within large-scale models of the deuterium nuclei and distributed worldwide to any location without the need for transmission lines.

Even hydro-electric storage schemes can be eliminated, returning water courses to their natural states of flow in turn reducing conflicts stemming from the flow of this precious resource worldwide.

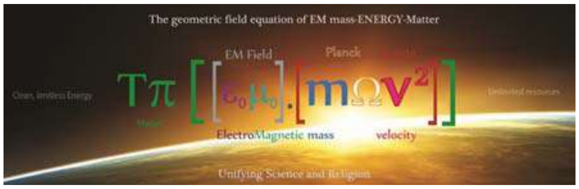
The geometric field equation of EM mass-ENERGY-Matter
EM Field
Planck
Clean, limitless Energy
Tπ [[ε₀μ₀].[mΩv²]]
Unlimited resources
ElectroMagnetic mass
velocity
Unifying Science and Religion

Through the geometries of equilateral Planck energies and the application of the Tetryonic unified field equation of mass-ENERGY-Matter in motion to our needs, for the first time in Humanity's history, we have the chance to advance our civilization to new pinnacles of technology and understanding in turn leaving our World a better place than we found it to be.

The age old dilemma of technological advancement versus environmental pollution is at an end

We stand on the threshold of realising one of the greatest advances in scientific understanding ever witnessed but we must also endeavour to redress our current social challenges with an equally rigorous determination to ensure all of Humanity benefits from this discovery

…………………..……… ABRAHAM [2012]

Tetryonics …. Challenging and correcting the foundations of quantum theory

Electric field permittivity has a 2pi (rhombic) field geometry while magnetic field permeability is always the result of two separate equilateral magnetic fields that form a magnetic dipole moment

Where two or more rhombic electric fields interact a Coulombic –charge force of interaction results
Where two or more equilateral magnetic fields interact a Gaussian-Ampere forces of interaction results *[opposite fields attract - same fields repel]*

The classical model of the unidirectional electrostatic force fields of point charged particles is corrected to show a bidirectional arrangement of equilateral Planck energy momenta geometries within the electrostatic field that creates the bi-directional forces of charge interactions

Where two similar coupled equilateral energy momenta fields combine an electrostatic field is created
Where two opposite coupled equilateral energy momenta fields combine a magnetic dipole is created

The arrangement and asymmetry of Planck quantum charges with in any EM field determines the motion and acceleration of charged particles within the field of electrostatic force

The classical forces of attraction should be renamed forces of interaction
(as they result from bidirectional quantum vector forces)

All electromagnetic fields are comprised of a squared number of equilateral Planck energy momenta quanta (whose asymmetrical distribution of equilateral Planck quanta results in a net divergent Force) with a convergent vector force component

When modelled with equilateral Planck geometries electromagnetic fields follow the Biot-Savart Law:

- Rhombic electric fields follow the inverse squared law of propagation
- Equilateral magnetic fields follow the inverse squared law of propagation
- Magnetic dipole fields follow an inverse cubed law of propagation

The equilateral geometry of quantised angular momentum creates chirality at the quantum level in physics (rotation does not affect the charge of plank quanta, but mirroring and/or spatial coordinate translations do)

All positive charged fields can be modelled as having a clockwise inductive flux geometry
All negative charged fields can be modelled as having an anticlockwise inductive flux geometry
(with OMEGA representing equilateral geometries per second and PI reflecting their scalar areas]

All W bosons are comprised of an odd number of equilateral Planck quanta in a trapezoidal geometry
All Z bosons are comprised of an even number of equilateral Planck quanta in a dual-trapezoidal geometry (neutral Z bosons are equivalent in terms of Planck mass energy momenta to photons save their net geometry)

All charged fascia [Higgs bosons] are comprised of squared numbers of equilateral Planck quanta in an equilateral geometry.

The odd quanta W bosons within electromagnetic fields form quantum levels of charged mass-energy momenta

The Planck equation [E=n.hv] of quantised energy is a generalised equation of Planck geometries (and applies to all energy momenta arrangements equally)

Bosons and Quantum levels are a transverse measure of mass energy momenta in EM fields
Photons are a longitudinal measure of mass energy momenta within the same electromagnetic fields

Equilateral Planck energy momenta per second [E/c^2] equates to mass geometries
Equilateral Planck energies momenta per second squared [E/c^4] equates to Matter topologies

Interactive **mass energy geometries** can combine via the weak force to form standing wave Matter topologies

Standing wave **Matter topologies** can combine via the strong force to form elementary subatomic particles

The fundamental Matter quantum of all physics is the tetryon (a tetrahedral mass-energy Matter topology) which is formed from photons of electrostatic energy and can be found as positive, neutral or negative charges topologies

These three forms of charged tetryons form the foundation of all the physical Matter of our universe as they seek a state of equilibrium, Tetryons are not spherical point charges – they are tetrahedral, likewise the subatomic particles they form are not spherical point charges - they are **[T]**π Tetryonic charge topologies

The Tetryonic topologies created by tetryons as they combine to form elementary particles creates new forms of closed volume topologies that follow Euler's characteristic [V+F-E = 2] for Platonic solids

Quarks and leptons are comprised of 12 charged mass energy geometries
- quarks form *[8π]* octahedral Matter topologies
- leptons form *[12π]* dodecahedral Matter topologies

and

Baryons are comprised of 36 charged mass energy geometries
- forming *[20π]* Icosahedral Matter topologies

Anti-UP quark's are comprised of 2 positive and 8 negative charges
creating a net charge of **-8** *[-2/3 elementary charge]*
DOWN quark's are comprised of 4 positive and 8 negative charges
creating a net charge of **-4** *[-1/3 elementary charge]*
Anti-DOWN quark's are comprised of 8 positive and 4 negative charges
creating a net charge of **4** *[1/3 elementary charge]*
UP quark's are comprised of 10 positive and 2 negative charges
creating a net charge of **8** *[2/3 elementary charge]*

Positrons are comprised of 12 positive charges
creating a net charge of **+12** *[positive integer elementary charge]*
neutrinos are comprised of 6 positive and 6 negative charges
creating a net charge of **0** *[no elementary charge]*
Electrons are comprised of 12 negative charges
creating a net charge of **-12** *[negative integer elementary charge]*

Protons are comprised of 24 positive charges and 12 negative charges
creating a net charge of **+12** *[positive integer elementary charge]*
Neutrons are comprised of 18 positive and 18 negative charges
creating a net charge of **0** *[no elementary charge]*
anti-Neutrons are comprised of 18 positive and 18 negative charges
creating a net charge of **0** *[no elementary charge]*
anti -Protons are comprised of 12 positive charges and 24 negative charges
creating a net charge of **-12** *[negative integer elementary charge]*

Rest lepton topologies have a completely neutralised magnetic dipole structure and all leptons (charged or neutral) have identical physical Matter topologies
Their rest masses are the result of the number of plank quanta comprising each Matter topology with applied forces creating additional asymmetrical KEM fields resulting in kinetic energies, magnetic moments and vector velocities of motion

As the energies of motion are added to equilateral KEM fields the Planck quanta themselves are physically WAVE-length Lorentz contracted

(Lorentz contractions apply to mass energy geometries only - NOT to Matter topologies)

Matter topologies are Lorentz invariant to velocity changes (or acceleration)
The relativistic distortion of spherical point charges in physics is no longer applicable, in turn undermining a basic tenet of special relativity as postulated by Albert Einstein

It is the equilateral secondary KEM fields of motion created by applied forces to material Matter topologies (and the asymmetrical distribution of plank charges therein) that creates the magnetic moments associated with charged particles in motion, not the relativistic distortion of spherical bodies of Matter as postulated by Einstein in special relativity

These secondary KEM fields of motion also contain the kinetic EM energies and vector [square root] linear momentum of material particles in motion - the mass energy momenta density and physical geometries of Matter topologies is unaffected by applied forces or changes to velocity-momenta

The changing Lorentz contracted mass energies of kinetic EM fields contributes to the total relativistic mass energy of Matter topologies but in no way changes the mass energy distribution of the Matter component of particles in motion itself - leading in turn to varying generations of subatomic particles with in the particle families themselves

The net asymmetric KEM field geometry of each particles in motion is reflective of the charged distribution of the Matter topology of the particle itself.

Mesons are created when quark and anti-quark particles combine via the strong force interaction between their charged fascias to create unique 14 pi Matter topologies (neutral or charged pions)

These strong interactions between opposite charge fascia of material particles and the resulting Matter topologies are evidence of the fact that Matter-antimatter combinations do not necessarily result in the explosive annihilation of the particles concerned (in fact the combining of opposite charge quarks is a necessary step in the formation of the baryons)

All Baryons are Matter topologies created from the combination of tri-quark topologies
The quark's combine via their opposite charged [Higgs] fascia to create neutral and charged elementary topologies
Due to this opposite charge interaction the charged distribution of baryonic topologies does not follow the textbook given arrangements of [UUD] & [DDU] instead they are arranged as follows:

Protons	**UP**	**+**	**DOWN**	**+**	**UP**
Neutrons	**DOWN**	**+**	**UP**	**+**	**DOWN**
anti-Neutrons	**anti-DOWN**	**+**	**anti-UP**	**+**	**anti-DOWN**
anti-Protons	**anti-UP**	**+**	**anti-DOWN**	**+**	**anti-UP**

The specific arrangement of charged fascia within these respective Matter topologies goes on to facilitate nuclear bonding via the residual strong force between atomic nuclei, following the charge arrangement given in modern textbooks does not allow the formation of baryons or their subsequent bonding to form larger scale atoms and molecules

Protons and neutrons are revealed to have quark arrangements that are mirror symmetries of each other, revealing the neutral charge neutron to have a complementary negative charge arrangement to that of protons.

The increased Planck mass energy contents of various families of quarks facilitates the creation of a huge number of possible Baryon topologies resulting in the particle zoo of subatomic elements created in particle collisions within accelerator experiments - all of these elements can be easily modelled and accounted for using Tetryonic mass-energy Matter topologies

Additionally, the particles often referred to as gluons within the standard model can be easily accounted for and explained as being neutral tetryons within Tetryonic theory - where these neutral charge Matter topologies form a neutral Matter dielectric between charged tetryons to facilitate the strong force interaction and the formation of larger particles

Neutral tetryons (gluons) can also form particles in their own right (neutrinos and glue-balls), as can charged tetryons (charge-ball particles), most of which are short lived

Tetryonics reveals the previously hidden geometry and topology of electromagnetic fields and subatomic particles and the geometric relationship and interplay between electric permittivity and magnetic permeability within all EM field that results on the speed of light in any medium.

As well as the revealing a geometric source of all physical constants within physics itself, Tetryonic charge geometries show the weak force to be a magnetic dipole [edge] interaction between mass energy geometries and Matter topologies and the strong force to be a charge [fascia] interaction between Matter topologies themselves

The equilateral [quantised angular momentum] of ***inductive*** Planck mass energy geometries within squared [HIGGS] fascia of Matter topologies is what creates ***inertial*** properties of Matter at the quantum level

The strong force between charged parallel Higgs fascia of Matter topologies results in 'hidden' partitions of mass energy with in the Matter topologies of quark's and baryons leading to the inability of mathematics online to accurately model and understand the physics of creating these particles at the quantum level

Tetryonic charged mass energy geometries allows for the accurate 2D and 3D modelling of all Matter topologies and fields of Force interactions within physics

It reveals that the deuterium atom [not the hydrogen atom] is the quantum building block of all elementary Matter - the unique charge topology of deuterium (consisting of a proton, neutron and electron) can be related to the electromechanical topology and function of quantum synchronous converters.

These quantum synchronous converters absorb and emit mass energies via the rotating electrons [quantum rotors] bound to the nuclei by way of radiant bosons and photons [heat and light]

The net charge of Deuteron nuclei facilitates their binding together to form helium and larger atoms, while stored mass energy quanta within their Matter topology results in raised quantum grounds states that can all be modelled mathematically using Schrodinger’s wavenumbers

The simpler geometric charge topology reveals visually how the mass energy content of baryons directly affects the angular momentum and energy levels of photo electrons bound to atomic nuclei, in turn revealing that larger elements be created in two ways - namely by the addition of a neutron to a Hydrogen atom (as classically modelled in modern chemical physics) or through the raising of Planck mass-energy quanta within the deuterium Matter topology itself

Deuterium nuclei without any bound photo-electrons form quantum batteries with EM mass energies can be stored indefinitely without being released.
These nuclei can be arranged atomically in either a series, parallel or empty parallel configuration each of which affects the ground energy states of the respective nuclei created.
Where photo-electrons (quantum rotors) are attracted and bind to these nuclei (quantum batteries) the stored mass energies can then be released in a controlled manner via spectral line emissions, or energy can be added via spectral line absorption.

As photo-electrons are bound to the deuteron nuclei the energy levels are affected directly by the mass energy levels of the nuclei that they bind to, in turn affecting their KEM field energies
The square root linear momentum of KEM field energies normally creates a vector motion but in the case of bound electrons this vector motion is turned into an angular motion resulting in electron spin within the nucleus.

The direction of the spin is always reference with respect to the nuclear magnetic moment resulting in up and down electron spins and their associated magnetic dipole moments

The kinematics of KEM fields provides an easy visual explanation of Newton's 2nd law of motion - namely that deceleration can be seen as the removal of odd number quanta from the squared kinetic energy fields and acceleration can be viewed as the addition of odd numbered quanta to squared kinetic energy fields.

The addition or subtraction of odd numbered Planck quanta from squared energy fields results in changing square root linear momentum for the particles that these fields are associated with - this changing linear momentum over time can be equated to acceleration leading to Newton's formulation of Force = mass x acceleration [F = ma = dp/t = dmv/t]

Using the equilateral geometry of Planck's constant we can now unify classical mechanics with quantum mechanics and reveal the underlying dynamics of relativistic mass energies of motion (as well is providing an electromechanical explanation of inertia in the physics of Matter in motion)

Newton's vector linear momentum [p=mv] forces are revealed to be the geometric square root of Leibnitz's scalar energies [E/mv^2] with their associated inertial mass being a measure of the scalar energies per second [m=E/v^2]

Linear momentum [mv] is differentiated from vector velocities [v] through the equilateral geometry of energy per second [m] in tetryonics, with the total linear momentum of any EM field or system of particles being the vector some of the square root Planck linear momentum of the quanta in the fields comprising the system

Once the equilateral field geometry of KEM fields is revealed it is easy to show that the total energy of the field [E=mv^2] is the result of the kinetic energies [$_{1/2}$Mv^2] plus the magnetic moment [$_{1/2}$Mv^2] of any material topology in motion

This in turn allows for a rigid geometric differentiation of quantised angular momentum [Ω] in Planck quanta, the linear momentum [p] of any system and the classical angular momentum [vector rotation about a point] in physics

These geometric definitions of mass energy momenta in physical systems as portrayed in Tetryonic theory reveal a differing geometric relationship to that commonly portrayed in modern physics]i.e. neutralising right angled Pythagorean triangles]

Additionally Lorentz corrections factors are restricted to planar 2-D mass energies of KEM fields associated to particles in motion where changes of velocity force result in the addition or subtraction of equilateral Planck quanta from the KEM field geometry resulting in the physical expansion or contraction of Planck geometries within the field

Tetryonic theory's geometric definition and application of quantised angular momentum with respect to mass energy momenta in physics allows for the physical modelling of mass energy and Matter at all scales of physics and their unification of classical, quantum and relativistic mechanics

The 'squared' energies of quantum mechanics and statistical distributions of the math associated with quantum mechanics is now revealed in an entirely new light facilitating the macro scale modelling of all hitherto unseen quantum particles and processes.

The geometric relationships between charge, mass geometries and Matter topologies is easily visualised and applied to all physics revealing quantised angular momenta [QAM] to be the hidden variable of Planck's constant [h].......

The modelling of quantised angular momentum as an equilateral geometry allows for the elimination of Heisenberg's uncertainty principle and Bell's inequality from the foundational tenants of quantum mechanics

Both the position and momentum of particles and fields can be now modelled at any stage of a system's spatio-temporal development and evolution using these simple equilateral charged geometries of mass energy momenta

Precise rest mass energies of any Matter topology can be calculated from 1st principles allowing for accurate modelling of all quantum systems and even periodic elements - expanding our knowledge of chemical systems and processes

The Planck-Compton frequency, de Broglie wavelength, absolute rest mass and charge Matter topology of all periodic elements, compounds and molecules can now be calculated with absolute certainty and precision and can be related back to Avogadro's number for all elements in chemistry

The exponential energy levels of atomic nuclei and atomic processes of energy release can also now be shown to be the result of an underlying geometric Planck scale geometry inherent to all mass energy Matter within physics

Einstein's mass energy equivalence formula is shown to be a geometric relationship between measurements of energy per second and inertial mass........ combined with Tetryonics new definition of Matter as energy per second squared we can now show that Einstein's famous formulation is an incomplete formulation of the mass-ENERGY-Matter relationship that applying only to specific processes within our physical universe

The Lorentz corrections on which special relativity is based are shown to be WAVE-length and QAM corrections for the changing physical geometries of scalar mass energies per unit of time in spatial co-ordinate systems (c^2) and that additional time-based dimensional analysis must be taken into account in order to define and differentiate between electromagnetic mass and Matter in physics

The application of squared (equilateral) geometries to physics in the role of quantised angular momentum has been overlooked since the inception of physics itself, but with the inclusion of this geometry a rigid geometric grammar can now be imposed upon the mathematics of physics leading to the correction of many erroneous mathematical assumptions

An exact value from first principles for Planck's constant can be calculated and Avogadro's number can be confirmed and related to the absolute rest mass of hydrogen atoms.

A geometric representation of Coulomb's charged field interactions and the impedance of free space can also be provided for all physical processes allowing for a fast and accurate visualisation of energy mechanics at all scales of physics

Tetryonics theory’s equilateral charged mass energy and polyhedral Matter topologies allow for the modelling and differentiation of energy in all its forms within any spatial coordinate system providing a vast improvement over relativity's use of stress energy tensors and non-Euclidean geometries in modelling physical systems

All fields and particles can be defined and differentiated in terms of their charged mass energy geometries and Matter topologies respectively and the differentiation of elemental families into their generations can be easily accounted for through the addition of mass energies of motion to secondary KEM fields associated with each particle grouping

Inertial resistance to force can be modelled and explained through the equilateral geometry of quantised angular momentum when viewed as an electromechanical equivalent to an inductive loop of energy

Electromagnetic inertial mass can now be clearly defined as a measure of the inductive EM energy density of any charged 2D planar Planck geometry in any spatial co-ordinate system of measurement

3D Matter can now be viewed and modelled as the closed topology created by 2D mass energies through charge interactions in any spatial co-ordinate system of measurement per second squared

Additionally physical terms such as 'mass-less' particles are shown to be a misnomer as any field or particle must by definition contain energy per unit of time - and should be more appropriately re-termed Matter-less or weightless particles

Tetryonics theory defines, differentiates and unifies energy in all its forms (mass, Matter & Force) through the equilateral geometry of quantised angular momentum at the quantum level, allowing for the development of a unified field equation that models physics on all scales and reveals a common geometric thread to all the disparate fields of physics that to date have defied unification

Equilateral (SQUARED) energy fields of Planck quanta comprised of statistical distributions of (ODD) transverse bosons and (EVEN) longitudinal photons account for all the quantum mechanical processes of physics including EM field geometries and material particle topologies and electron spin

Principle quantum levels [1st Schrodinger wave-numbers] are shown to be the result of the series addition of deuterium nuclei in atomic elements with their resulting energy levels determining the squared energy levels of bound photoelectron KEM fields – in turn forming the foundation of discrete quantum jumps of electrons bound to atomic nuclei

The bound photo-electron can only transition between allowed squared energy levels permitted by the energy levels of deuterium nuclei to which they are bound and once free from the atomic nuclei the electron can emit or absorbed a continuous spectra of light photons - but while bound to the deuterium nuclei in atoms it can only have specific [squared] KEM energy levels

The complex, asymmetric tri-quark (9 tetryon) Matter topology of baryons directly affects the magnetic moment of protons and neutrons in motion making them considerably weaker than the symmetric axial Bohr magnetron of electrons in motion

As the Bohr magneton of photo-electrons is determined by the squared energies of its secondary KEM field confusion has arisen with respect to the allocation of spins to subatomic particles, Tetryonic geometries now correct for this and suggest that the electron historically viewed as a spin 1/2 particle is in fact a spin 3 particle under the formal definitions of spin rotations i.e. rotating at 120° can bring it back to an identical quantum state as the initial state measured]

Again Tetryonic theory and its energy geometries in addition to explaining and correcting electron spins also shows that the term electron spin can now be taken literally when modelled with Tetryonic geometries to provide an accurate description of the origin of magnetic moments for electrons, in turn undermining the foundational postulate of special relativity that these magnetic moments are the result of the distortion of spherical point charges and that relativistic Lorentz contractions apply to Matter topologies

Dirack's constant, or Planck's reduced constant [$h/2\pi$], is revealed to be a measure of the electric permittivity of a relativistic electron's KEM field, and the measurement of the associated magnetic dipole of the same KEM field results in the mistaken concept of 1/2 spins in the measurement of Bohr magnetons

The magnetic vector created by the Bohr magneton [or KEM field dipole] when measured with respect to the nuclear magneton [Proton magnetic moment] produces either a higher energy [parallel] or lower energy [anti-parallel] state for all atomic nuclei

It is the arrangement of these parallel or anti-parallel magnetic dipoles within atomic nuclei that determines the exact frequency-wavelength (energy level) of emitted spectral lines with the interaction of a particle’s magnetic dipole moment with external electromagnetic fields creating Lorentz forces

As electrons and protons have identical quantum Matter topologies [elementary charges] but differing mass energy contents the electron is able to create a much larger secondary KEM field (with associated magnetic dipole moment) for the same velocity [$_{1/2}$Mv^2]

The energy level differences created by the parallel or anti-parallel spin coupling of electrons with protons or deuterium nuclei within atomic elements creates the hyperfine splitting and Zeemen effects observed and spectral lines

As the electron is now shown to be a complex non-spherical particle topology when it binds with protons to form hydrogen the unique Matter topology of both the proton and the electron [Hydrogen] create a small but detectable precession in the motion of the spinning electron as compared to that of the motion of the same electron in a deuterium nuclei

Additionally, as all periodic elements, compounds and molecules are comprised of deuterium nuclei (not hydrogen as currently supposed) the orientation and spin direction of electrons bound within these nuclei results in the diamagnetic and paramagnetic properties of various macro-scale materials

Tetryonics theory reveals that the mass charge ratio is of particles are measure of the constituent mass energy geometries that go into making the particle's final Matter topology that we observe a measure

But unlike the current mathematical approach the geometric approach of Tetryonics reveals the hidden mass energy partitions within the Matter topology of all subatomic particles and that particles such as tetryons (the quantum building blocks of all Matter) are in fact hidden within the debris of current collider results.

4 quantum charge tetryons have the same1/3 partial elementary charge as some quarks, and can be measured as having the same mass charge ratio's as charged leptons such as positrons and electrons in collider experiments.

………. To be continued [QED, Chemistry, Cosmology, Mathematics, Biology and economic theory]

answers@tetryonics.com

ABRAHAM [2014]

principia geometrica [3b] - quantum chemistry 2
principia geometrica [4] - quantum cosmology 230
principia geometrica [5] - quantum mathematics 461
Principia Scientifica [5] - Quantum Mathematics 624
Principia Scientifica [3a] - Quantum Chemistry 787
principia geometrica [1] - quantum mechanics 991
principia geometrica [2] - quantum electrodynamics 1232
principia geometrica [3a] - quantum chemistry 1451

Copyright ABRAHAM [2008] - All rights reserved

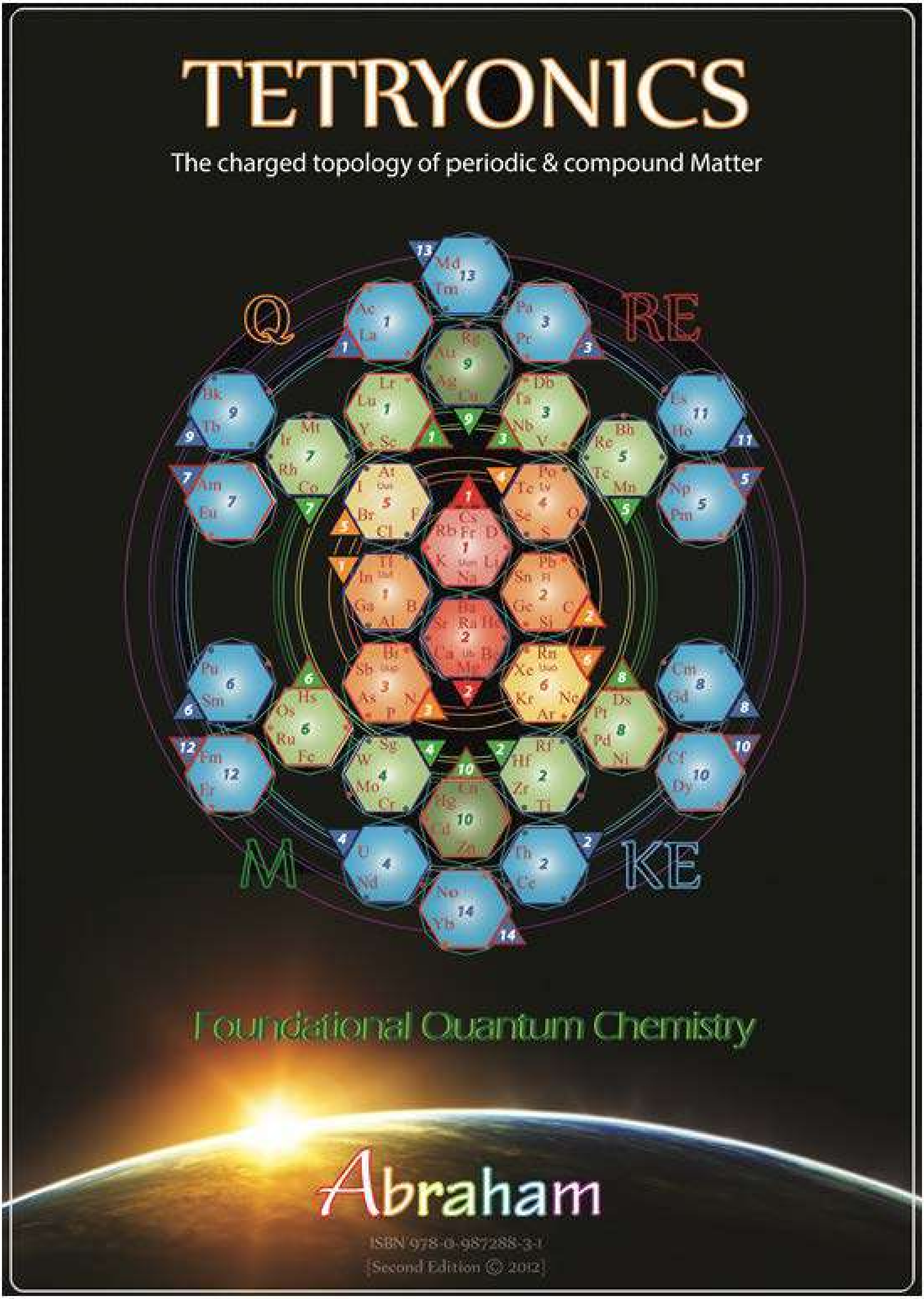

TETRYONICS

The charged topology of periodic & compound Matter

Foundational Quantum Chemistry

Abraham

ISBN 978-0-987288-3-1

[Second Edition © 2012]

Periodic element Families

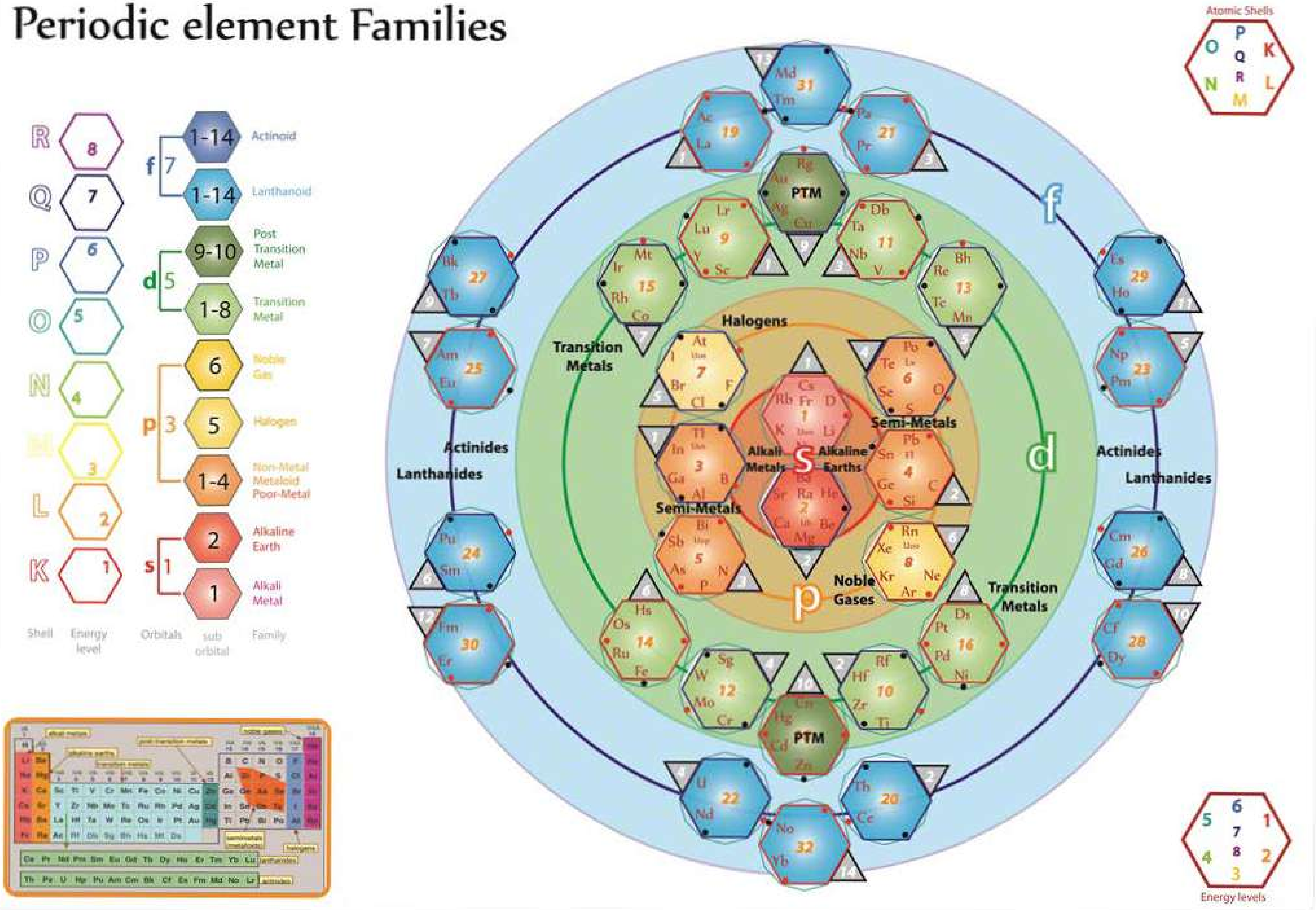
Atomic Shells
O P Q K
N R L
M
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family
f
d
s
p
PTM
Halogens
Transition Metals
Semi-Metals
Alkali Metals
Alkaline Earths
Noble Gases
Actinides
Lanthanides
Energy levels
5 6 1 7 8 4 2 3

Alkali Metals

The alkali metals are silver-colored (caesium has a golden tinge), soft, low-density metals, which react readily with halogens to form ionic salts, and with water to form strongly alkaline (basic) hydroxides.

These elements all have one electron in their outermost shell, so the energetically preferred state of achieving a filled electron shell is to lose one electron to form a singly charged positive ion.

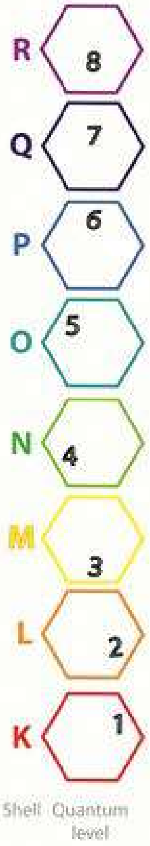
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
Shell
Quantum level

S
1
Alkali Metal

s Orbital

n1-8

s1 sub-orbital

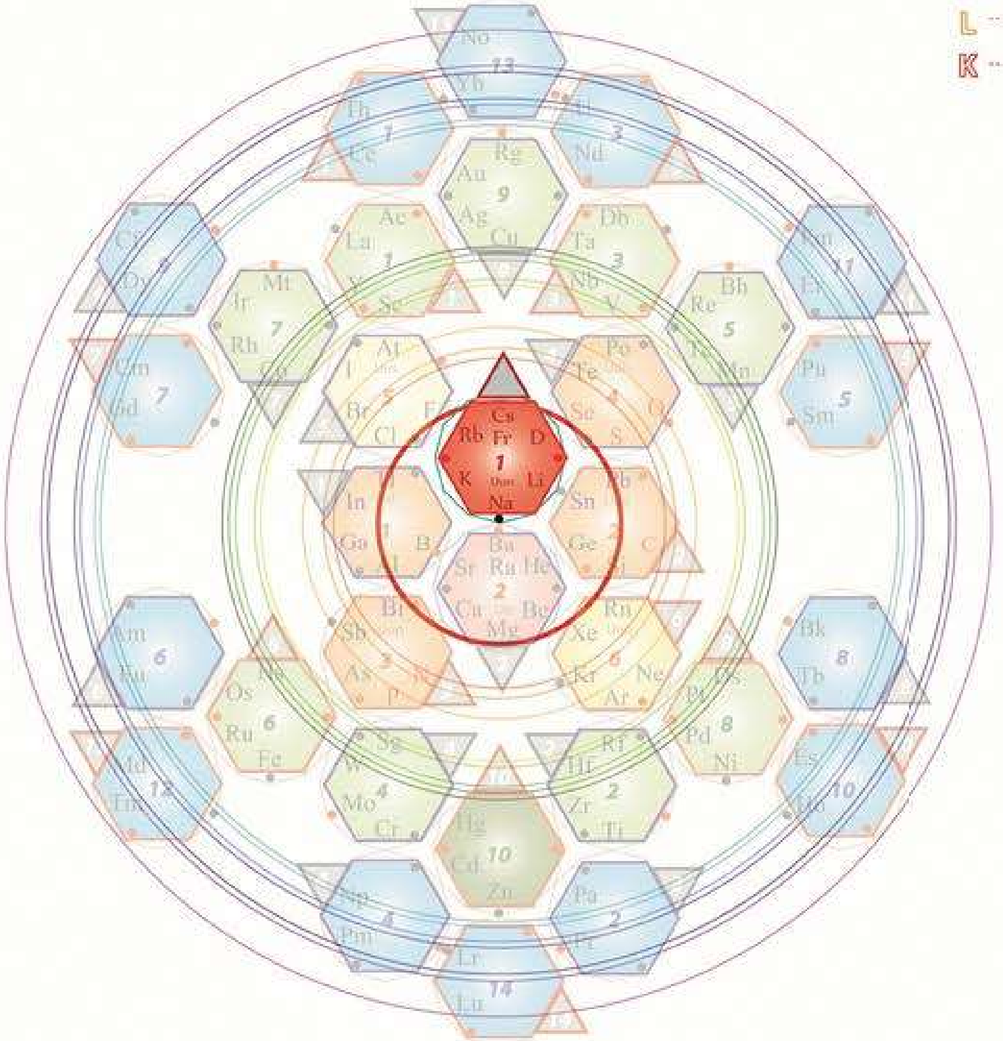
Cs
Rb
Fr
D
1
K
Li
Na
Ba
Sr
Ra
He
2
Ca
Be
Mg
No
13
Yb
Au
Rg
9
Ag
Cu
10
Hg
Cd
Zn
14
Lr
Lu

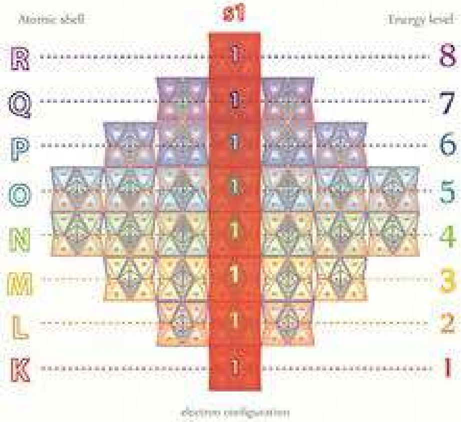
Atomic shell
s1
Energy level
R
Q
P
O
N
M
L
K
1
8
7
6
5
4
3
2
1
electron configuration

8

Alkaline Metals

Quantum level

The alkali metals are all highly reactive
and are never found in elemental form in nature.

As a result, in the laboratory they are stored under mineral oil.
They also tarnish easily and have low melting points and densities.
Potassium and rubidium possess a weak radioactive characteristic
(harmless) due to the presence of long duration radioactive isotopes.

Alkaline Earths

The alkaline earth metals are silver colored, soft metals, which react readily with halogens to form ionic salts, and with water, though not as rapidly as the alkali metals, to form strong alkaline (basic) hydroxides.

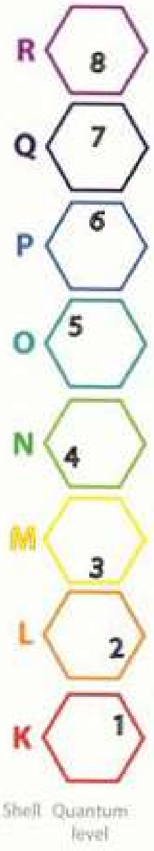
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
Shell
Quantum level

s **2** **Alkaline Earth**

s Orbital

n1-8

s2 sub-orbital

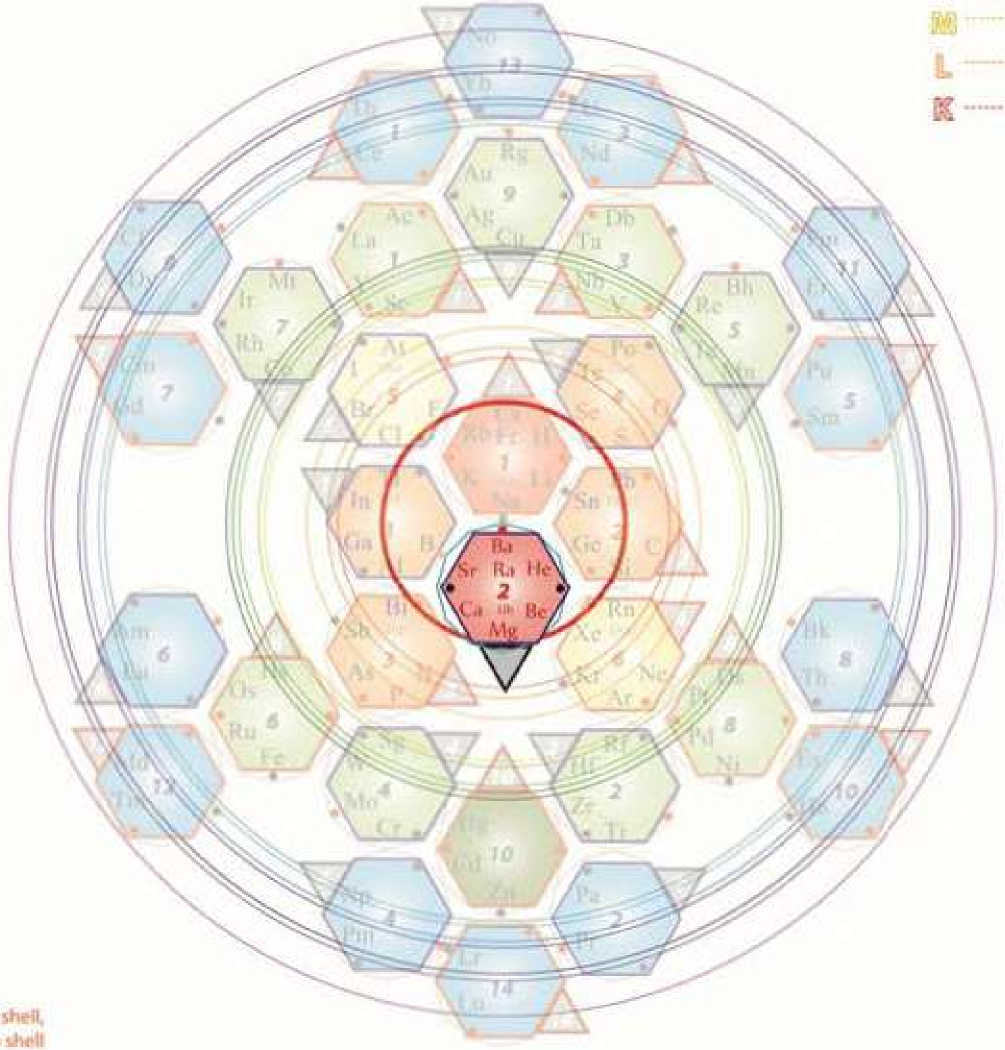
Ba
Sr
Ra
He
2
Ca
Be
Mg
1
In
Ga
B
Sn
Ge
C
Sb
As
P
Bi
Xe
Kr
Ar
Rn
Ne
Se
S
O
Po
Br
Cl
F
At
La
Ac
Sc
Au
Ag
Cu
Rg
9
Ta
Nb
V
Db
3
Ir
Rh
Co
Mt
7
Re
Tc
Mn
Bh
5
Os
Ru
Fe
Hs
6
Pd
Ni
Pt
Ds
8
W
Mo
Cr
Sg
4
Hf
Zr
Ti
Rf
2
Hg
Cd
Zn
10
No
Yb
13
Ce
Th
Nd
Pr
Pa
Pu
Np
U
Lr
Lu
14
Cm
Gd
7
Pu
Sm
5
Am
Eu
6
Bk
Tb
8
Es
Ho
10
Tm
Md
12
Dy
Cf
9
Er
Fm
11

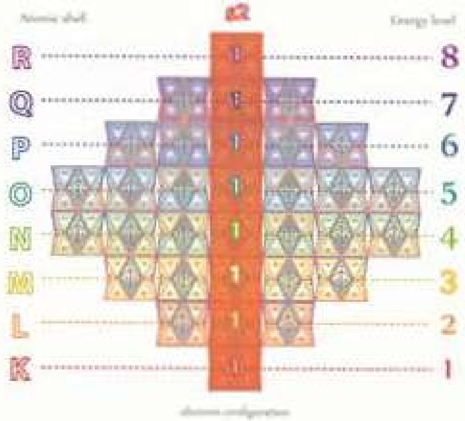
Atomic shell
s2
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
1
1
1
1
1
electron configuration

8
Alkaline Earths

All the alkaline earth metals have two electrons in their valence shell, so the energetically preferred state of achieving a filled electron shell is to lose two electrons to form doubly charged positive ions.

6
5
7
1
8
4
2
3
Quantum level

Metalloids

A metaloid is a material with a small overlap in the energy of the conduction band and valence bands.

Unlike a regular metal, metaloids have been described electrically as charge carriers of both types (holes and electrons), so it could be argue that they should be called 'double-metals' rather than metaloids.

As metaloids have fewer charge carriers than metals, they typically have lower electrical and thermal conductivities.

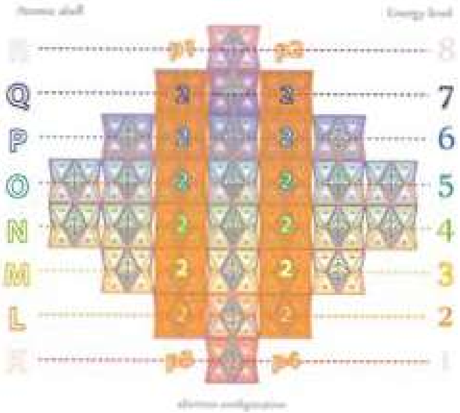
Atomic shell
Energy level
Q
P
O
N
M
L
8
7
6
5
4
3
2
1
p1
p3
p5
p6
2
electron configuration

Q
7
P
6
O
5
N
4
M
3
L
2
Shell
Energy level

p
1-4
Non-Metal
Metaloid
Poor-Metal

2 p-Orbitals

n2-7

p1-4 sub-orbitals

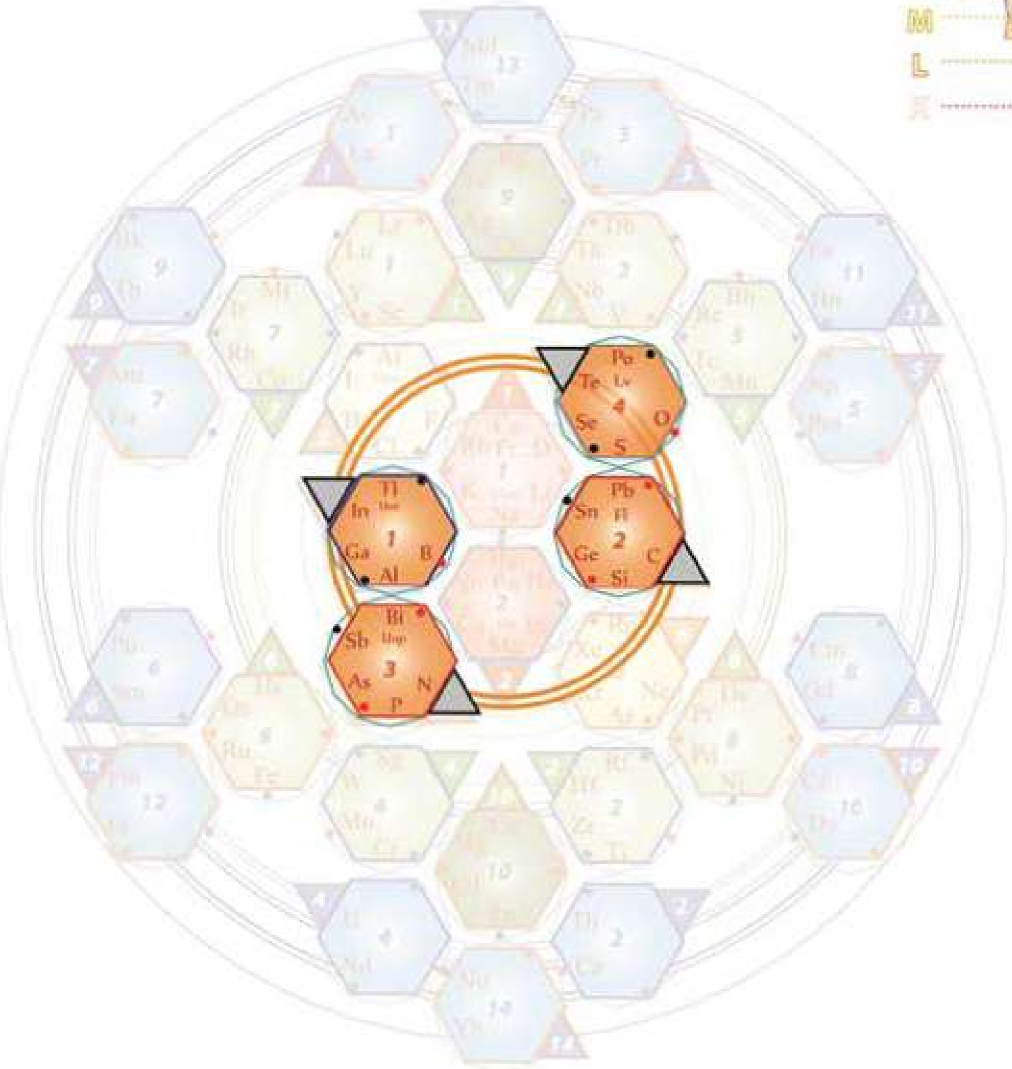
Po
Te
Lv
4
Se
O
S
Tl
In
Uut
1
Ga
B
Al
Pb
Sn
Fl
2
Ge
C
Si
Bi
Sb
Uup
3
As
N
P

24
Semi-Metals

Charge carriers typically occur in much smaller numbers than in a real metal. In this respect they resemble degenerate semiconductors more closely.
This explains why the electrical properties of metaloids are partway between those of metals and semiconductors.

6
5
7
4
2
3
Energy levels

Halogens

Owing to their high reactivity, the halogens are found in the environment only in compounds or as ions.

Halide ions and oxoanions such as iodate (IO3−) can be found in many minerals and in seawater.

Halogenated organic compounds can also be found as natural products in living organisms. In their elemental forms, the halogens exist as diatomic molecules, but these only have a fleeting existence in nature and are much more common in the laboratory and in industry.

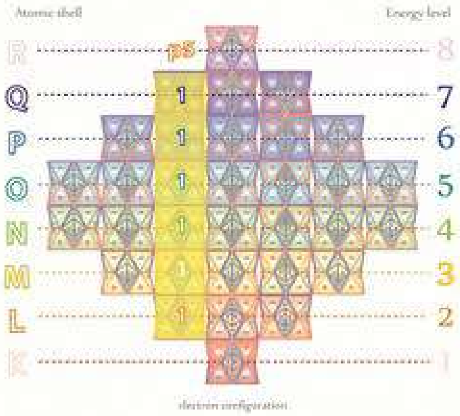
Atomic shell
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
p5
1
electron configuration

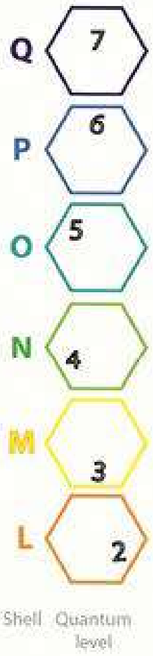
Q
7
P
6
O
5
N
4
M
3
L
2
Shell
Quantum level

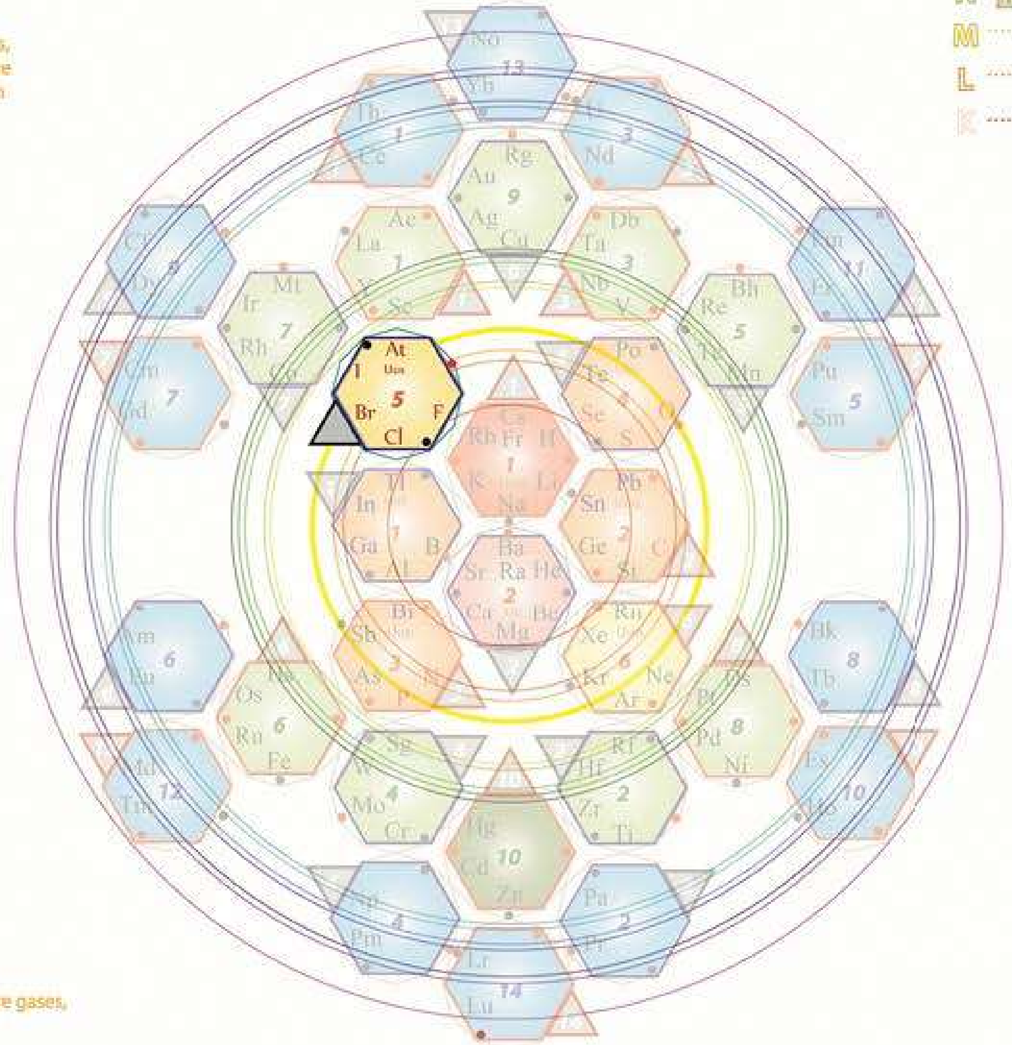
At
Uus
I
Br
5
Cl
F

p3 Orbital

n2-7

p5 sub-orbital

6

Halogens

6
5
7
4
2
3
Quantum level

At room temperature and pressure, fluorine and chlorine are gases, bromine is a liquid and iodine and astatine are solids.

Group 17 is therefore the only periodic table group exhibiting all three states of matter at room temperature

Nobel Gases

The properties of the noble gases can be well explained by modern theories of atomic structure: their outer shell of valence electrons is considered to be "full", giving them little tendency to participate in chemical reactions, and only a few hundred noble gas compounds have been prepared.

The melting and boiling points for each noble gas are close together, differing by less than 10 °C (18 °F); consequently, they are liquids only over a small temperature range.

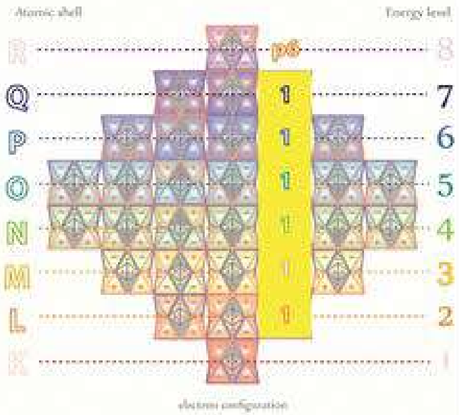
Atomic shell
Energy level
R
Q
P
O
N
M
L
K
p6
8
7
6
5
4
3
2
1
electron configuration

Q
7
P
6
O
5
N
4
M
3
L
2
Shell
Quantum level

p
6
Noble Gas

completed

p3 Orbital

n2-7

p6 sub-orbital

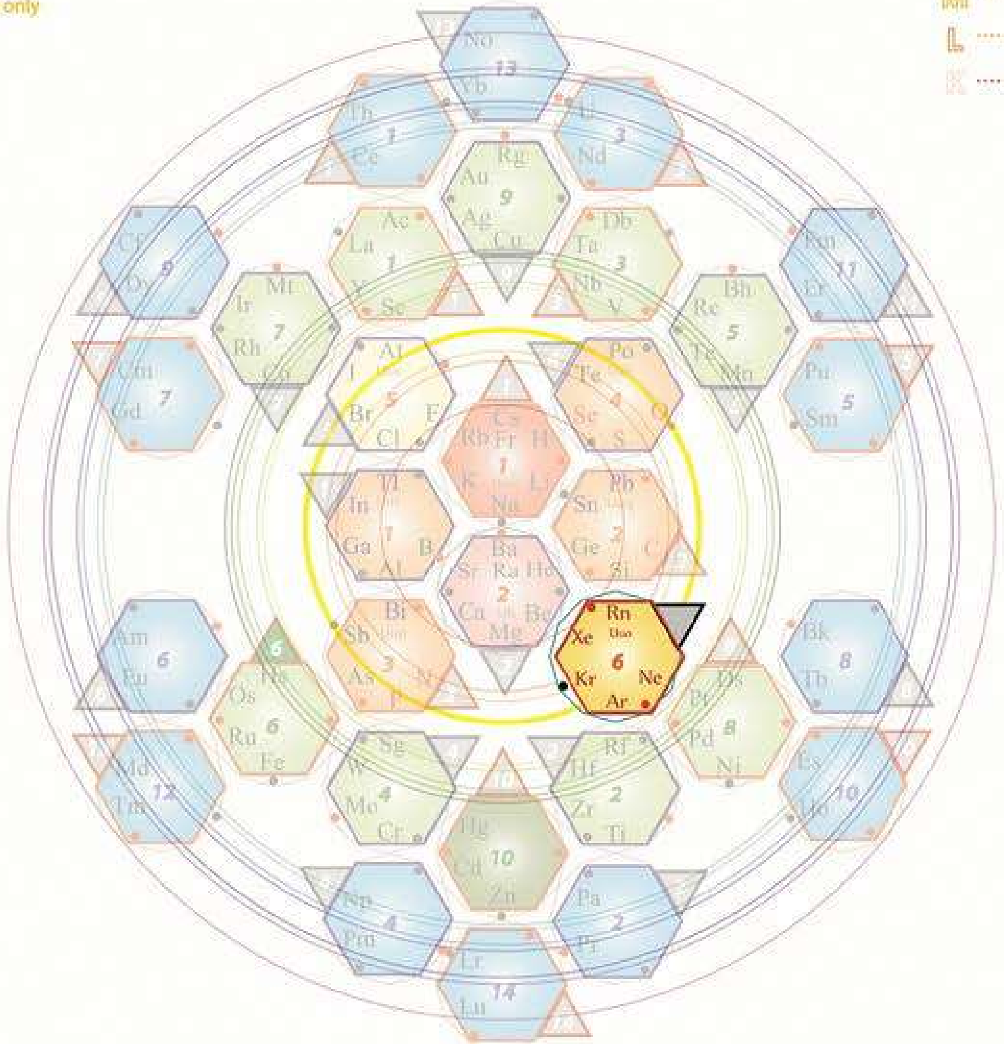
Rn
Uuo
Xe
6
Kr
Ne
Ar

6

Nobel Gases

6
5
7
4
2
3
Quantum level

Neon, argon, krypton, and xenon are obtained from air using the methods of liquefaction of gases and fractional distillation.

Helium is typically separated from natural gas, and radon is usually isolated from the radioactive decay of dissolved radium compounds.

Transitional Metals

In chemistry, the term transition metal commonly refers to any element in the d-block of the periodic table, including the group 12 elements zinc, cadmium and mercury.

This corresponds to groups 3 to 12 on the periodic table, which are all metals.

More strictly, IUPAC defines a transition metal as "an element whose atom has an incomplete d sub-shell, or which can give rise to cations with an incomplete d sub-shell."

The first definition is simple and has traditionally been used. However, many interesting properties of the transition elements as a group are the result of their partly filled d subshells.

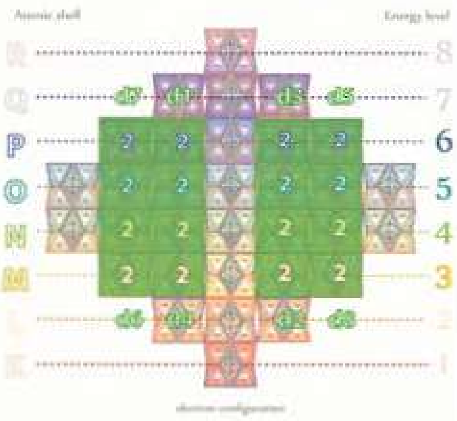
Atomic shell
Energy level
P
O
N
M
8
7
6
5
4
3
2
1
d5
d10
d3
d5
d5
d10
d3
d5
2
electron configuration

P
6
O
5
N
4
M
3
Shell
Quantum level

d
1-8
Transition Metal

4 d-Orbitals

n3-6

d1-8 sub-orbitals

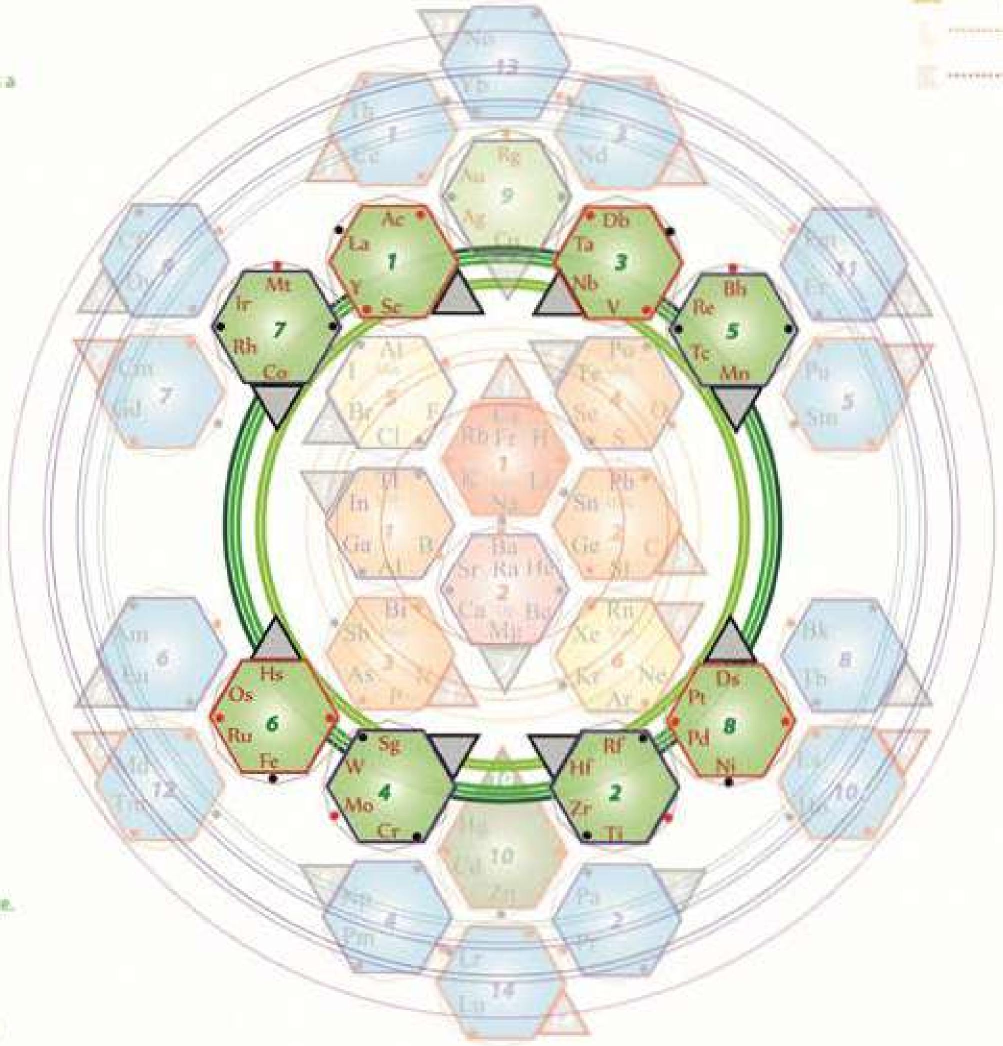
Ac
La
Y
Sc
1
Db
Ta
Nb
V
3
Mt
Ir
Rh
Co
7
Bh
Re
Tc
Mn
5
Hs
Os
Ru
Fe
6
Ds
Pt
Pd
Ni
8
Sg
W
Mo
Cr
4
Rf
Hf
Zr
Ti
2
Rg
Au
Ag
9
10
13
14

32 Transition Metals

6
5
4
3
Quantum level

The name transition comes from their position in the periodic table. In each of the four periods in which they occur, these elements represent the successive addition of electrons to the d atomic orbitals of the atoms.

In this way, the transition metals represent the transition between group 2 elements and group 13 elements.

Post-Transition Metals

In chemistry, the term post-transition metal is used to describe the category of metallic elements to the right of the transition elements on the periodic table.

There are two IUPAC definitions of "transition element" that have been in apparent conflict with one another since September 2007.

According to the first definition, transition metals are elements in group 3 through group 11.
In this case, post-transition metals include all of group 12— zinc, cadmium, mercury, and ununbium

According to the second definition, transition elements either have an incomplete d-subshell or have the ability to form an incomplete d-subshell.

Post Transitional metals are normally defined as completed d orbitals

Tetryonics suggests it may be appropriate to include all the d5 orbital elements in this grouping

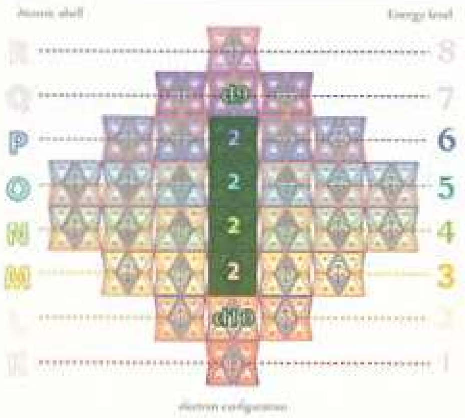
Atomic shell
Energy level
P
O
N
M
8
7
6
5
4
3
2
1
d9
2
2
2
2
d10
electron configuration

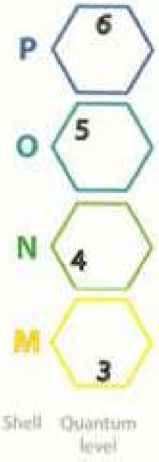
P
6
O
5
N
4
M
3
Shell
Quantum level

d
9-10
Post Transition Metal

d5 Orbital

n3-6

d9-10 sub-orbitals

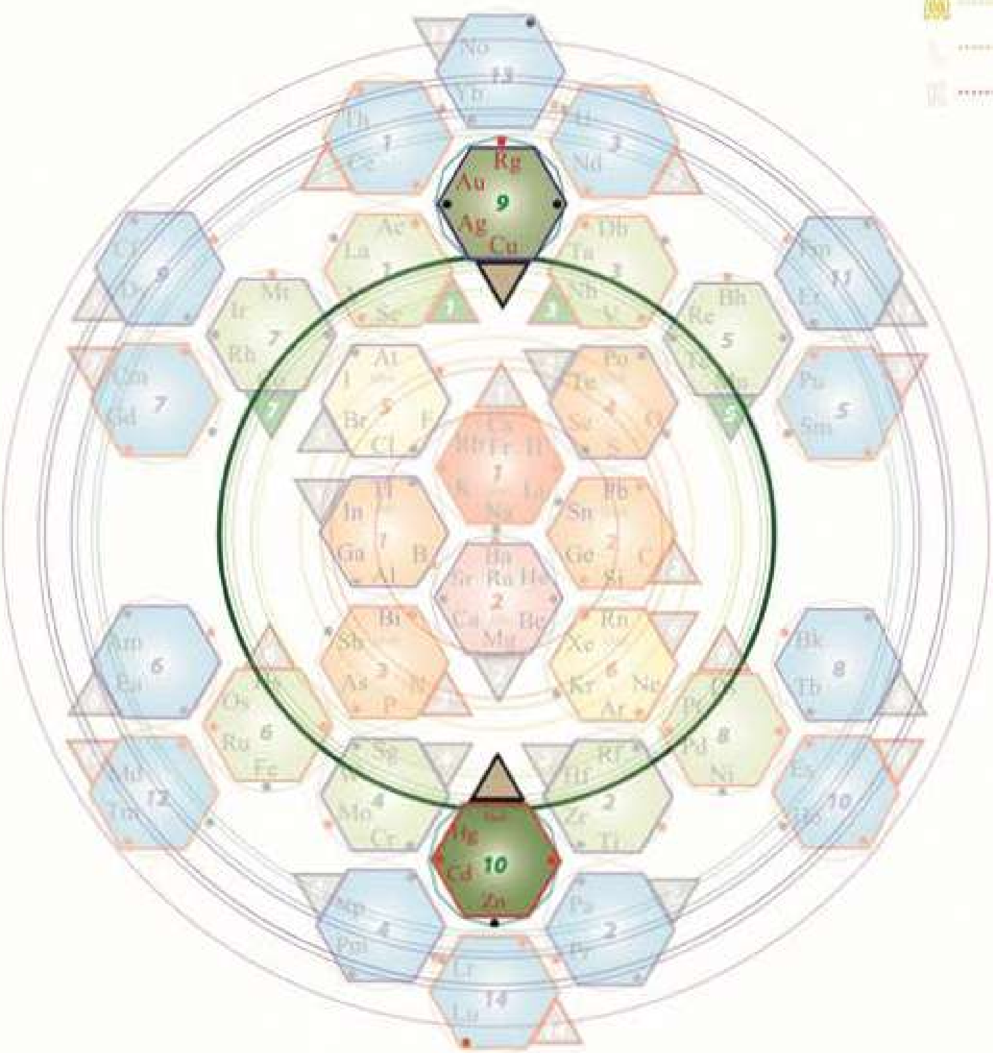
Rg
Au
9
Ag
Cu
Hg
Cd
10
Zn

8

Post-transitional Metals

In 2007, mercury(IV) fluoride was synthesized.[2][3] This compound contains a mercury atom with an incomplete d-subshell, and ununbium is predicted to have the capacity to form a similar electronic configuration.

In this case, post-transition metals include only zinc and cadmium within group 12.

6
5
4
3
Quantum level

Lanthanoids

All lanthanoids are f-block elements, corresponding to the filling of the 4f electron shell

The lanthanoid series (Ln) is named after Lanthanum.

The trivial name "rare earths" is sometimes used to describe all the lanthanoids together with scandium and yttrium.

These elements are in fact fairly abundant in nature, although rare as compared to the "common" earths such as lime or magnesia. Cerium is the 26th most abundant element in the Earth's crust, neodymium is more abundant than gold and even thulium (the least common naturally-occurring lanthanoid) is more abundant than iodine.

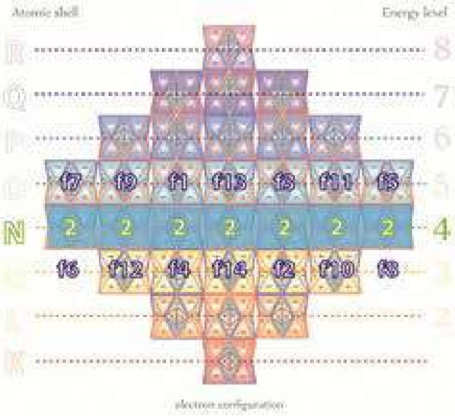
Atomic shell
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
f7 f9 f1 f13 f3 f11 f5
2 2 2 2 2 2 2
f6 f12 f4 f14 f2 f10 f8
electron configuration

f 1-14 Lanthanoid

N 4

Shell Quantum level

7 f-Orbitals

n4

f1-14 sub-orbitals

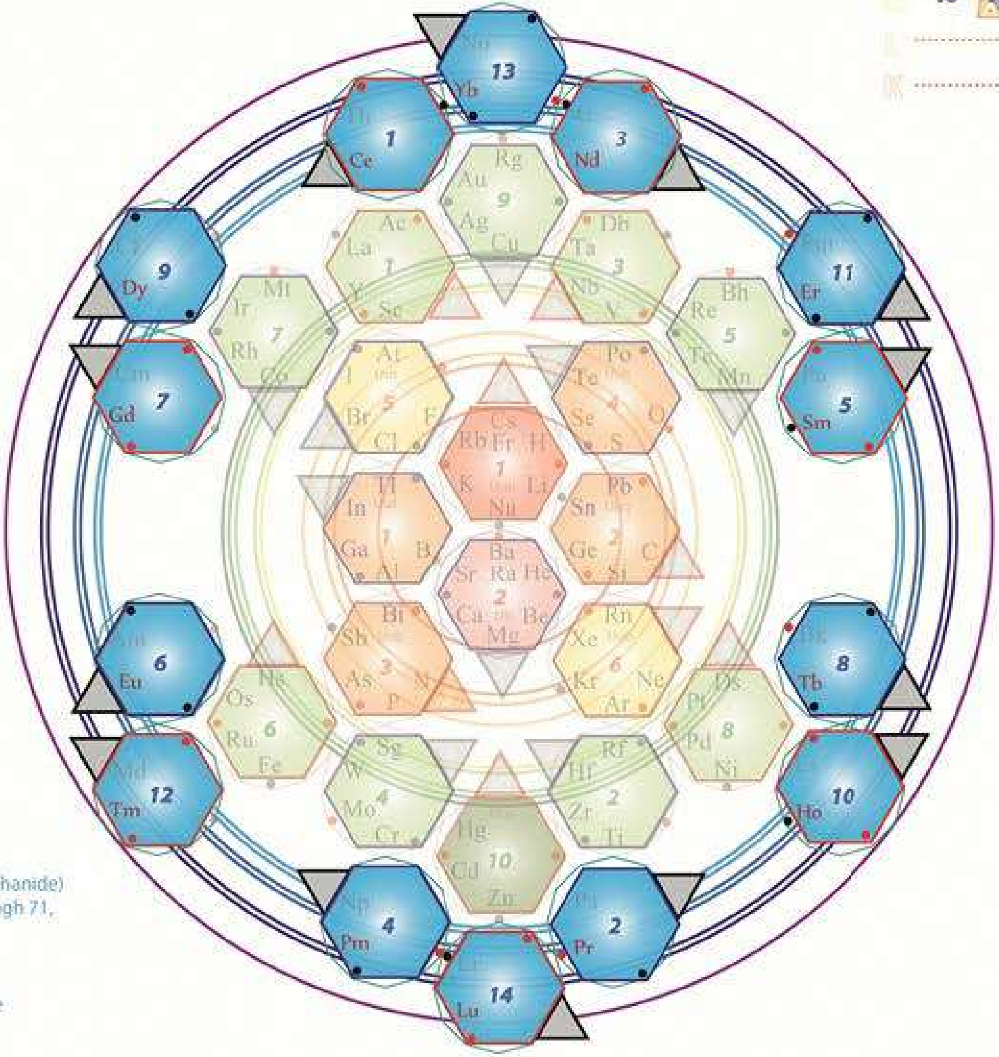
13 Yb
1 Ce
3 Nd
9 Dy
11 Er
7 Gd
5 Sm
6 Eu
8 Tb
12 Tm
10 Ho
4 Pm
2 Pr
14 Lu

14 Lanthanoids

4 Quantum level

The term "rare earths" arises from the minerals from which they were isolated, which were uncommon oxide-type minerals.

According to the IUPAC terminology, the lanthanoid (previously lanthanide) series comprises the fifteen elements with atomic numbers 57 through 71, from Lanthanum to Lutetium.

The use of this name is deprecated by IUPAC, as they are neither rare in abundance nor "earths" (an obsolete term for water-insoluble strongly basic oxides of electropositive metals incapable of being smelted into metal using late 18th century technology).

Actinoids

According to IUPAC nomenclature, the actinoid (previously actinide) series encompasses the 15 chemical elements that lie between actinium and lawrencium included on the periodic table, with atomic numbers 89 - 103.

The actinoid series derives its name from the first element in the series, actinium, and ultimately from the Greek ακτίς (aktis), "ray," reflecting the elements' radioactivity.

The actinoids display less similarity in their chemical properties than the lanthanoid series (Ln), exhibiting a wider range of oxidation states, which initially led to confusion as to whether actinium, thorium, and uranium should be considered d-block elements. All actinoids are radioactive.

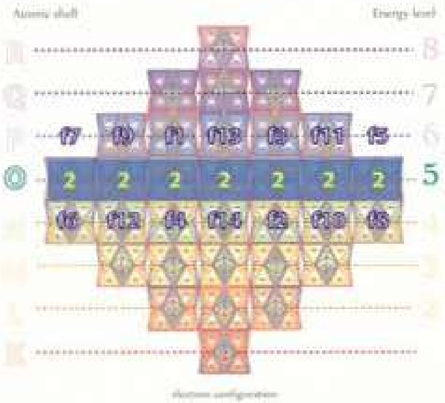
Atomic shell
Energy level
8
7
6
5
f7
f9
f1
f13
f3
f11
f5
O
2 2 2 2 2 2 2
f6
f12
f4
f14
f2
f10
f8
electron configuration

f 1-14 Actinoid

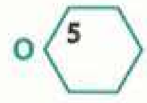

Shell Quantum level

7 f-Orbitals

n5

f1-14 sub-orbitals

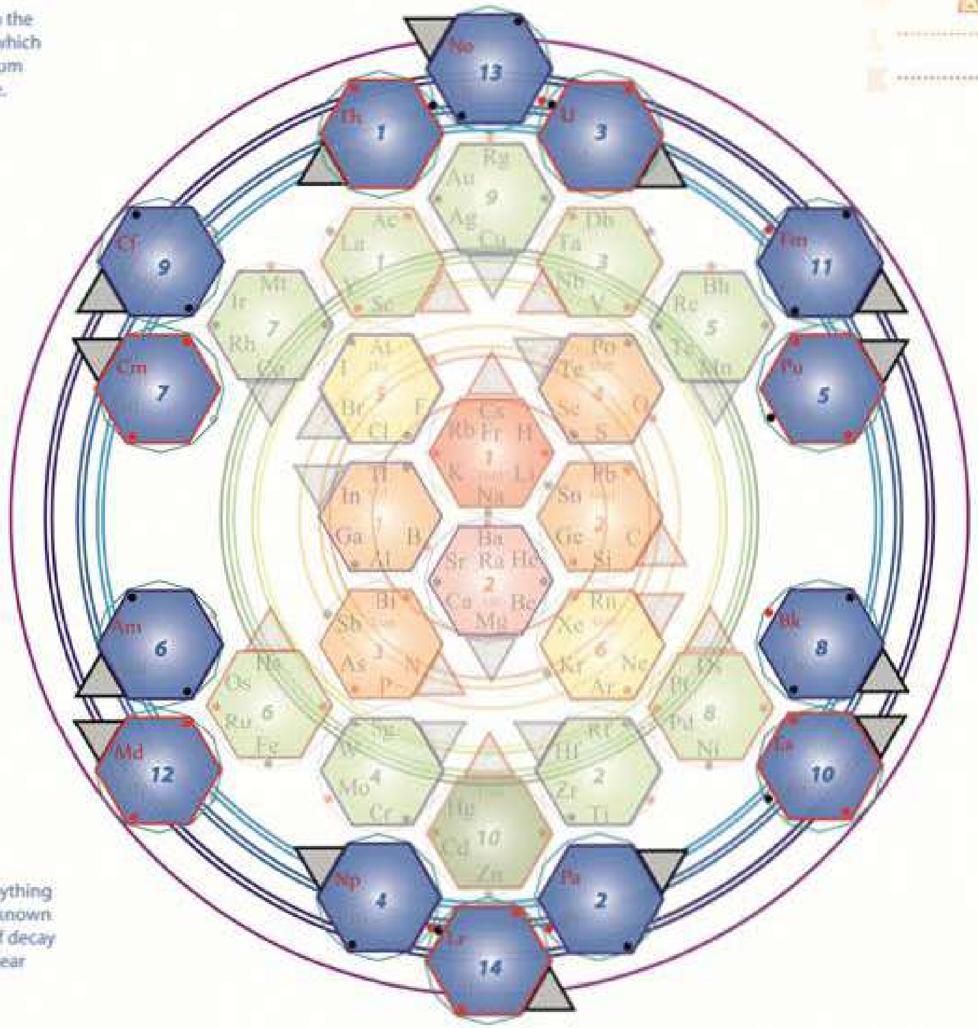
No 13
Th 1
U 3
Cf 9
Fm 11
Cm 7
Pu 5
Am 6
Bk 8
Md 12
Es 10
Np 4
Pa 2
Lr 14

14 Actinoids

Quantum level

Only thorium and uranium occur naturally in the earth's crust in anything more than trace quantities. Neptunium and plutonium have been known to show up naturally in trace amounts in uranium ores as a result of decay or bombardment. The remaining actinoids were discovered in nuclear fallout, or were synthesized in particle colliders.

The latter half of the series possess exceedingly short half-lives.

Copyright ABRAHAM [2008] - All rights reserved

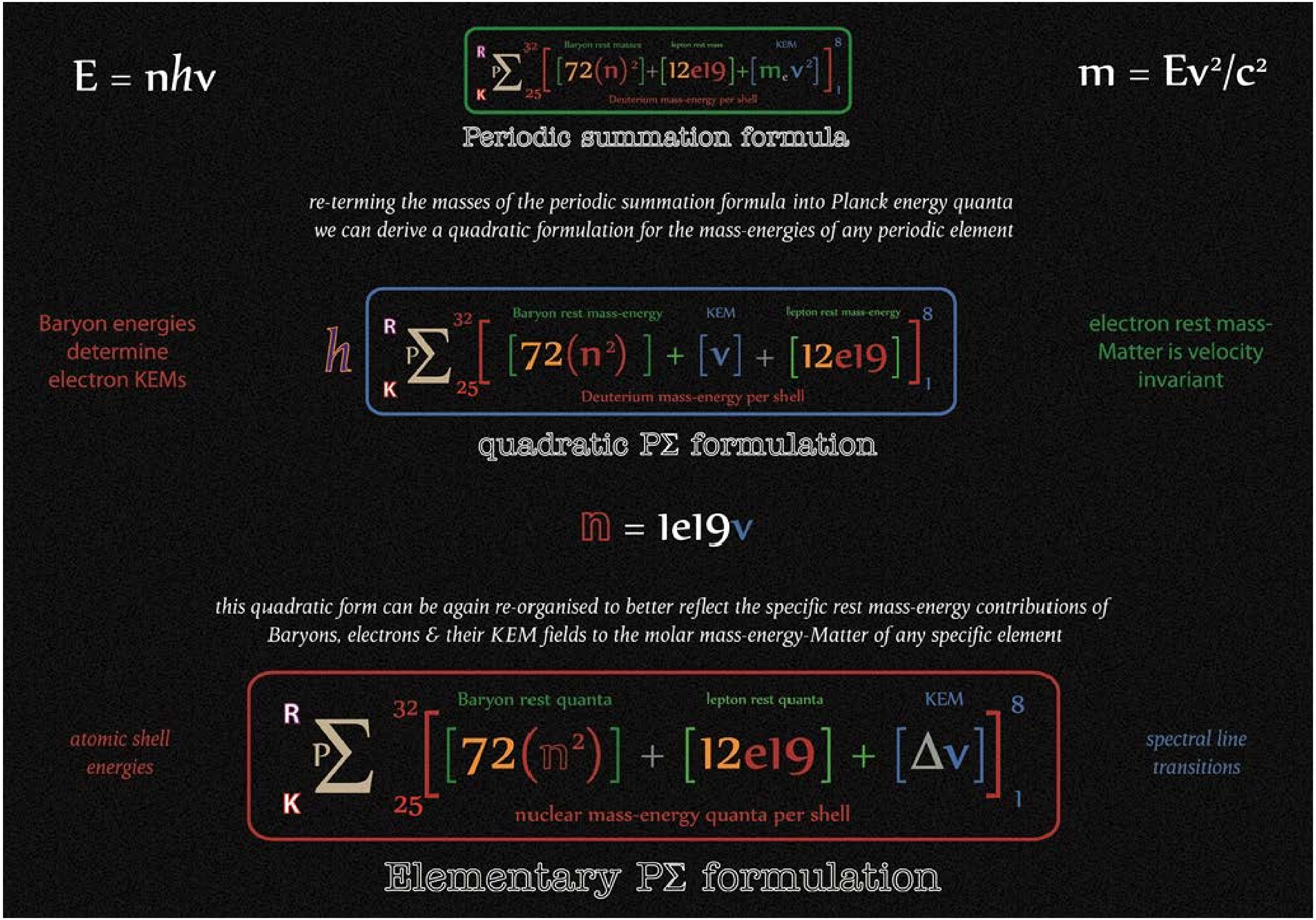

$$E = nhv$$

$$m = Ev^2/c^2$$

Baryon rest masses   lepton rest mass   KEM

$${}^{R}_{K}\,{}^{P}\!\sum_{25}^{32}\Big[[72(n)^2]+[12e19]+[m_e v^2]\Big]_1^8$$

Deuterium mass-energy per shell

Periodic summation formula

re-terming the masses of the periodic summation formula into Planck energy quanta we can derive a quadratic formulation for the mass-energies of any periodic element

Baryon energies determine electron KEMs

Baryon rest mass-energy   KEM   lepton rest mass-energy

$$h\;{}^{R}_{K}\,{}^{P}\!\sum_{25}^{32}\Big[[72(n^2)] + [v] + [12e19]\Big]_1^8$$

Deuterium mass-energy per shell

electron rest mass-Matter is velocity invariant

quadratic PΣ formulation

$$n = 1e19v$$

this quadratic form can be again re-organised to better reflect the specific rest mass-energy contributions of Baryons, electrons & their KEM fields to the molar mass-energy-Matter of any specific element

atomic shell energies

Baryon rest quanta   lepton rest quanta   KEM

$${}^{R}_{K}\,{}^{P}\!\sum_{25}^{32}\Big[[72(n^2)] + [12e19] + [\Delta v]\Big]_1^8$$

nuclear mass-energy quanta per shell

spectral line transitions

Elementary PΣ formulation

Tetryonics 52.13 - Periodic element summation formula

aufbau filling principle

The aufbau principle was developed by Neils Bohr and Wolfgang Pauli in 1920 as they applied quantum mechanics to the physical properies of chemistry

It attempted to model the properties of electrons in atoms to and led to the introduction of the n+l orbitals, and later to the development of Schrodiner's quantum numbers

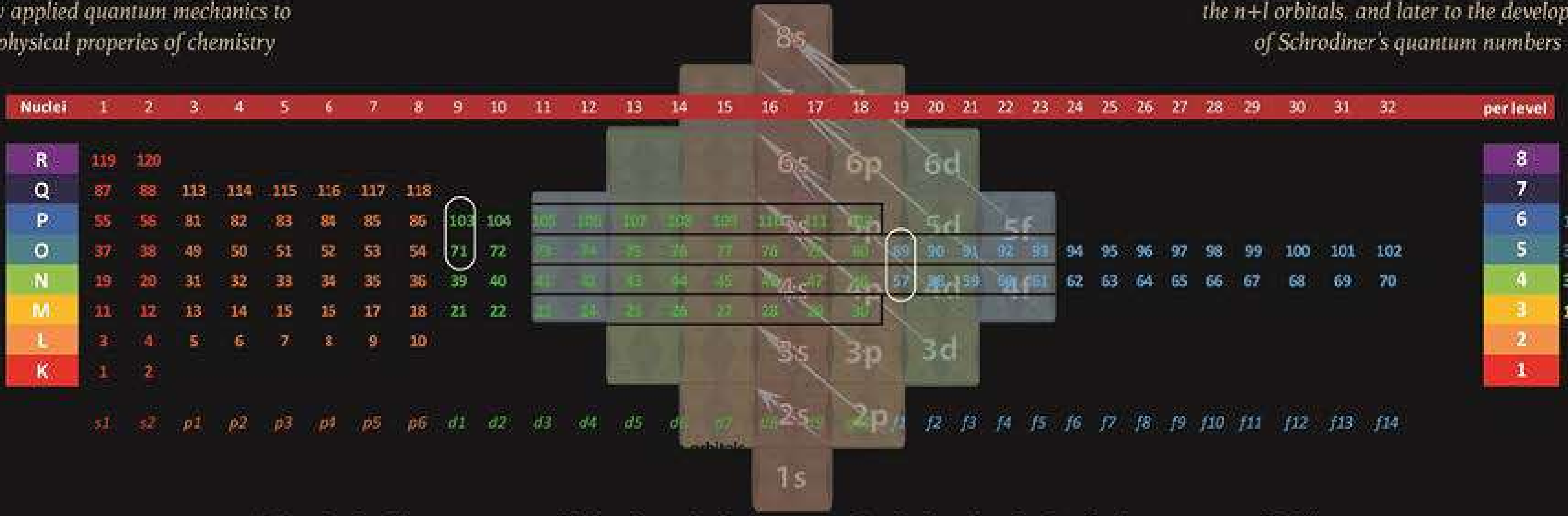
Nuclei
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
per level
R Q P O N M L K
8 7 6 5 4 3 2 1
2 8 18 32 32 18 8 2
8s
6s 6p 6d
5d 5f
4s 4p 4d 4f
3s 3p 3d
2s 2p
1s
s1 s2 p1 p2 p3 p4 p5 p6 d1 d2 d3 d4 d5 d6 d7 d8 d9 f1 f2 f3 f4 f5 f6 f7 f8 f9 f10 f11 f12 f13 f14

"The orbitals of lower energy are filled in first with the electrons and only then the orbitals of higher energy are filled."

While succesful in describing lower numbered elements, many anomalous configurations arise from it in the d & f electron orbitals that need to be corrected

Tetryonic orbital filling

Tetryonic theory refines the aufbau principle by correcting it to follow the true quantum filling order of Deuterium nuclei in atoms

This then allows for all elementary nuclei, their atomic configurations & quantum properties to be modeled exactly

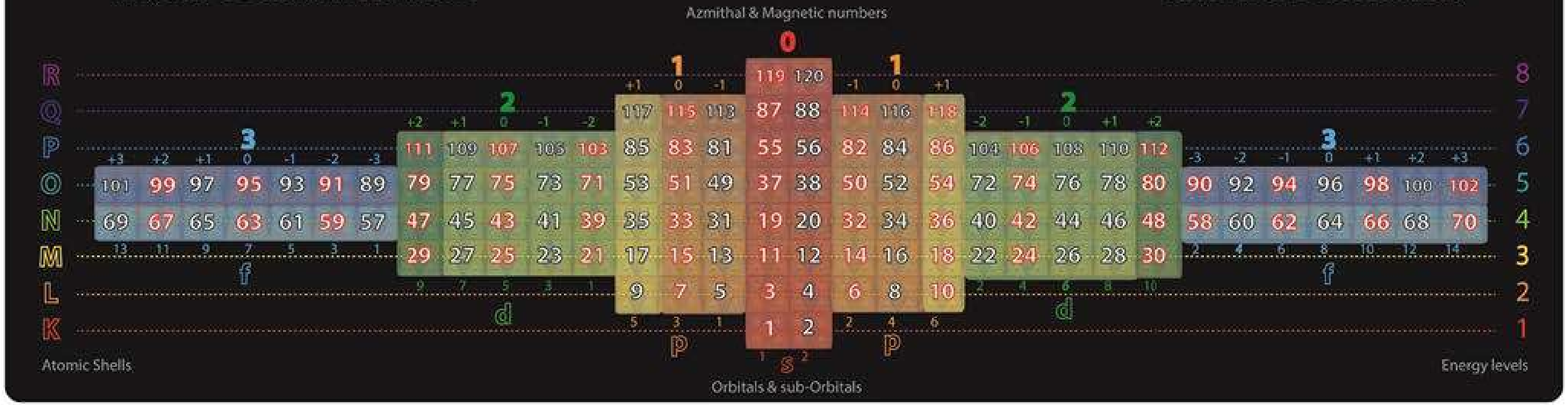
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
R Q P O N M L K
8 7 6 5 4 3 2 1
s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 52.14 - Tetryonic aufbau orbital filling

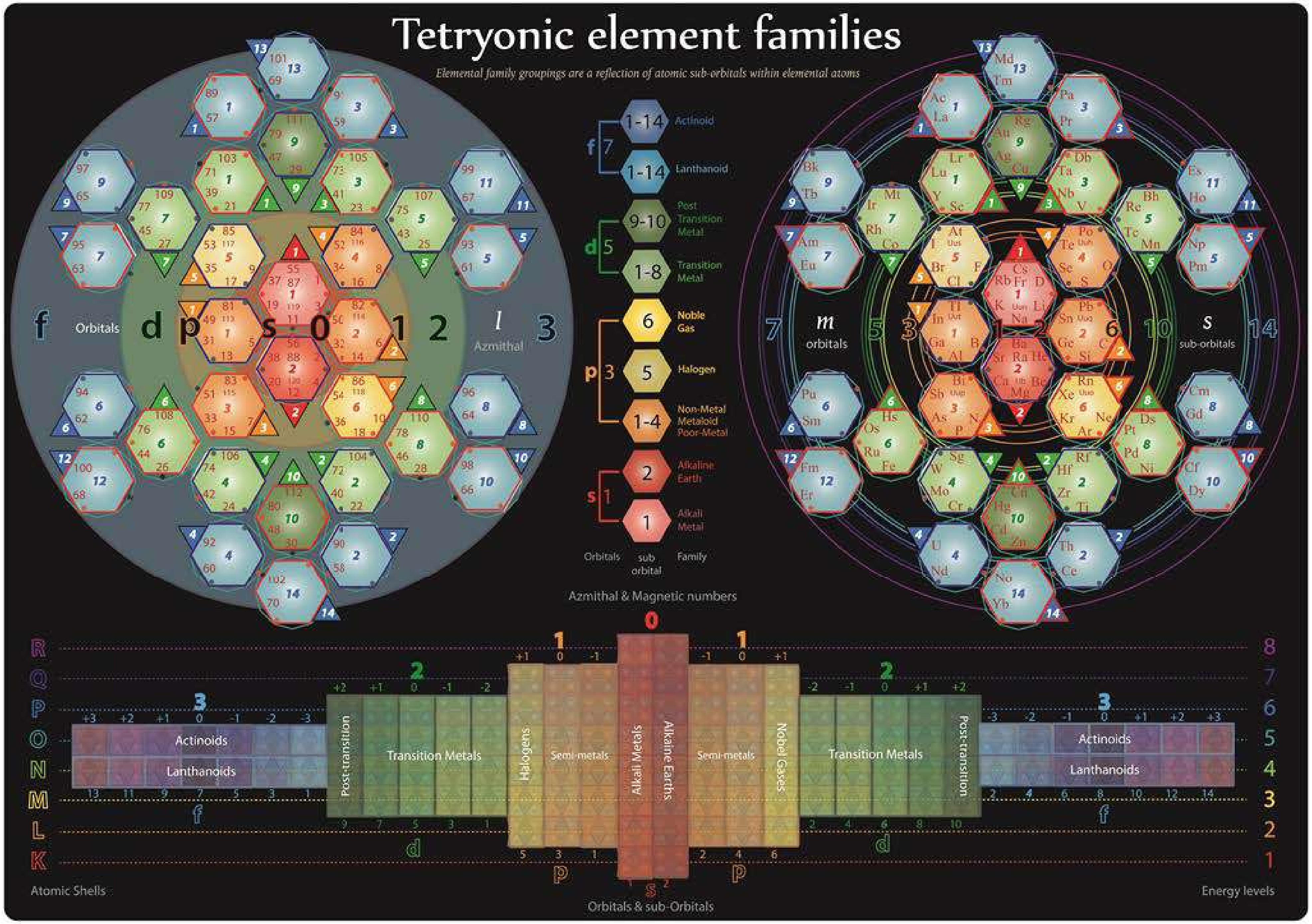

Tetryonic element families

Elemental family groupings are a reflection of atomic sub-orbitals within elemental atoms

f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Orbitals
sub orbital
Family
Azmithal & Magnetic numbers
f Orbitals d p s 0 1 2 l Azmithal 3
7 m orbitals 5 3 1 2 6 10 s sub-orbitals 14
R Q P O N M L K
Atomic Shells
8 7 6 5 4 3 2 1
Energy levels
Actinoids
Lanthanoids
Post-transition
Transition Metals
Halogens
Semi-metals
Alkali Metals
Alkaine Earths
Nobel Gases
Orbitals & sub-Orbitals

Tetryonics 52.15 - Tetryonic elemental families

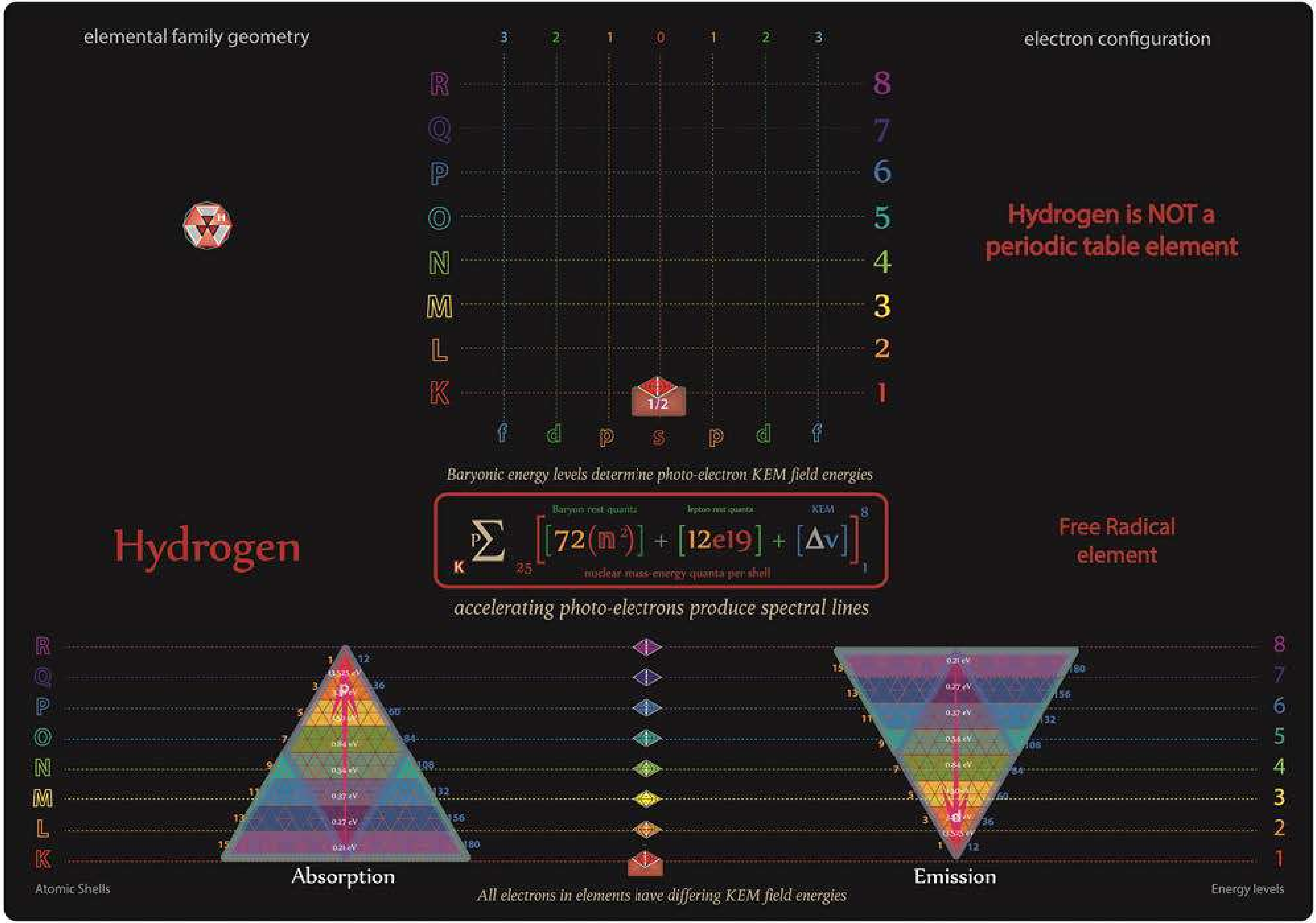
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
1/2
f d p s p d f
H
Atomic Shells
Absorption
Emission
Energy levels

Hydrogen is NOT a periodic table element

Hydrogen

Free Radical element

Baryonic energy levels determine photo-electron KEM field energies

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{P} {}_{25}\left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{8}$$

nuclear mass-energy quanta per shell

accelerating photo-electrons produce spectral lines

All electrons in elements have differing KEM field energies

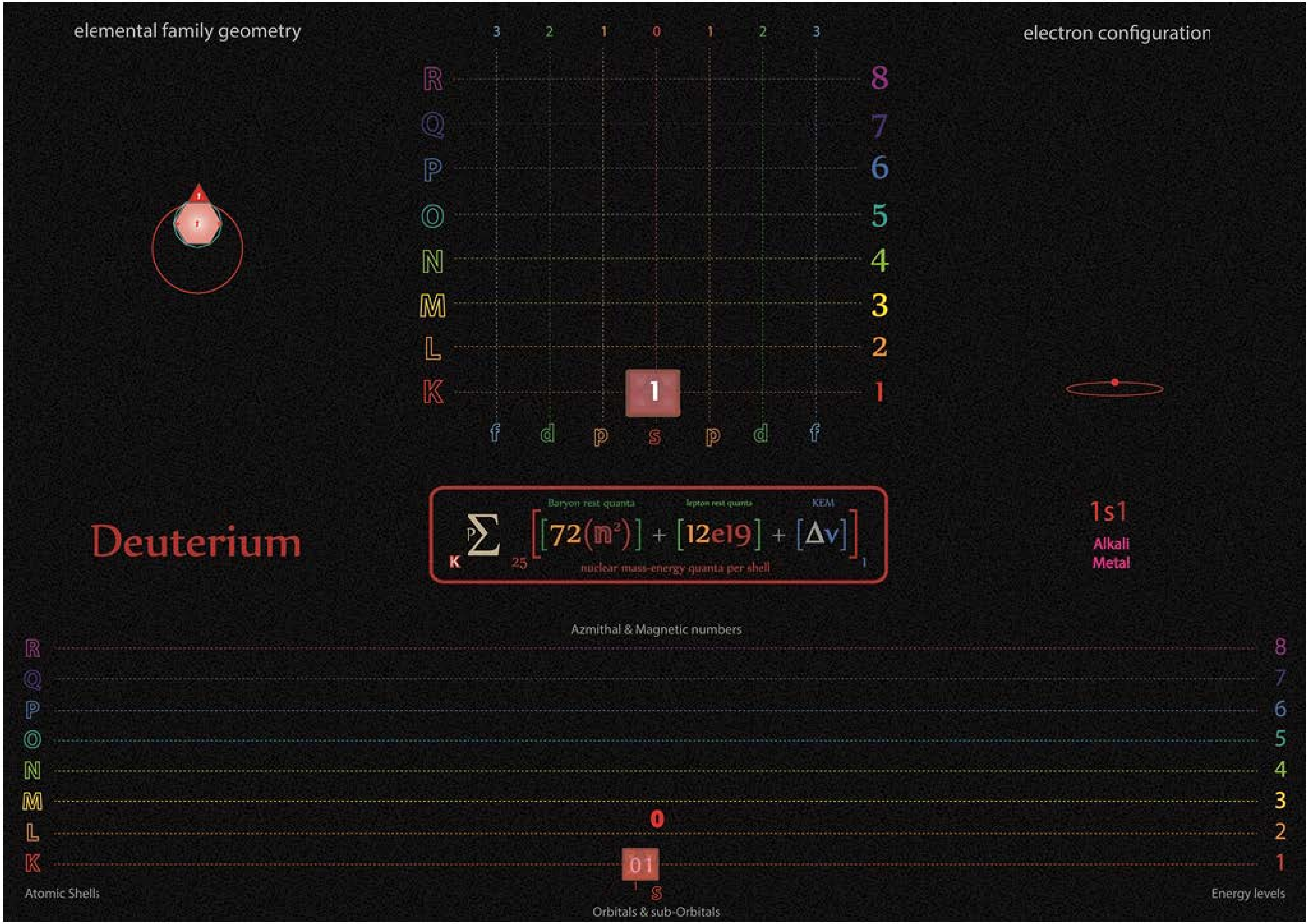
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
1
f d p s p d f
Deuterium
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
1s1
Alkali
Metal
Azmithal & Magnetic numbers
0
01
1
s
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

$$_{K}\sum^{P}\ _{25}\left[\left[72(\mathfrak{m}^{2})\right]+\left[12e19\right]+\left[\Delta v\right]\right]_{1}$$

Tetryonics 53.01 - Deuterium atomic config

Copyright ABRAHAM [2008] - All rights reserved

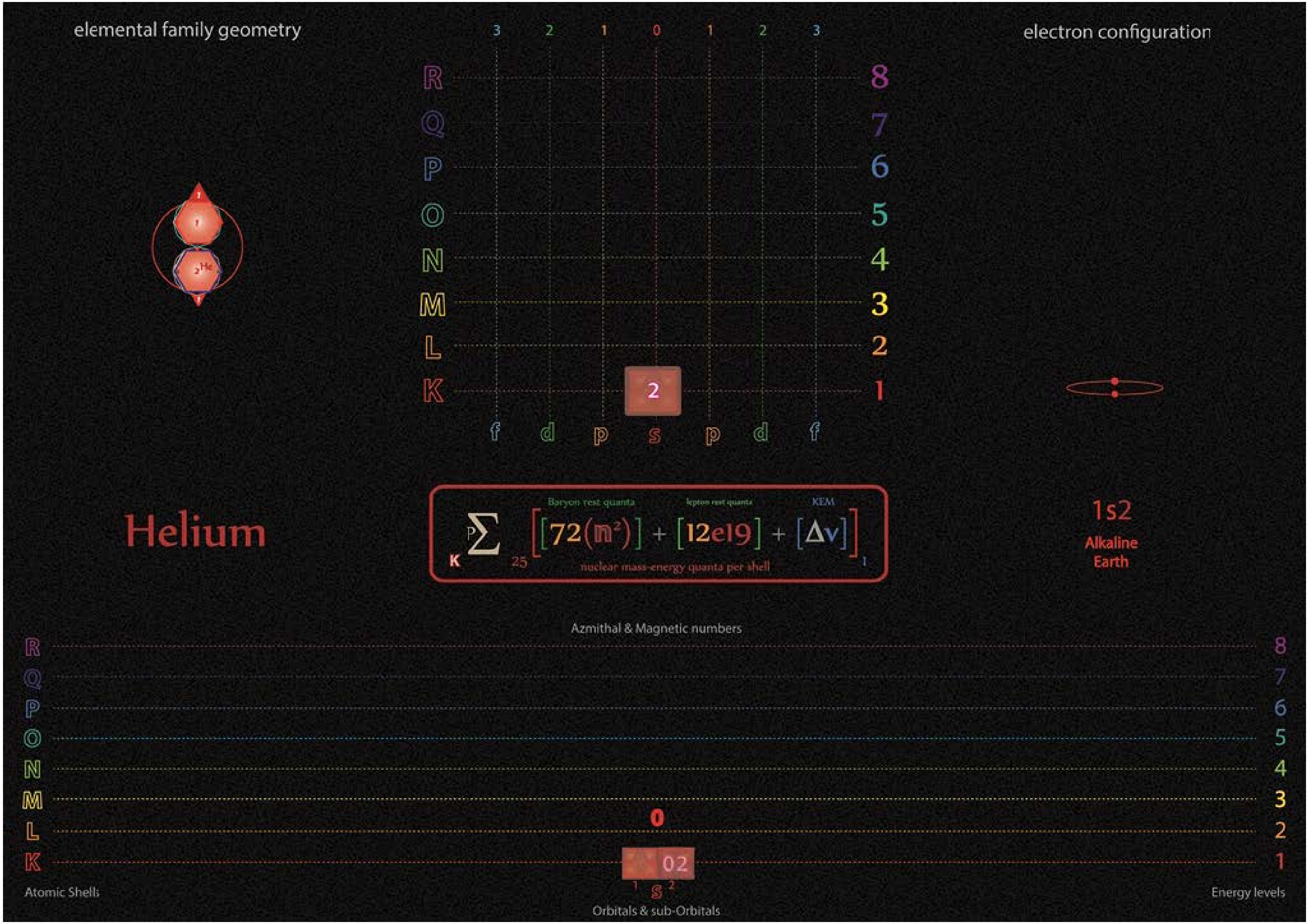
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
f d p s p d f
1
1
2He
Helium
Baryon rest quanta
lepton rest quanta
KEM
$_{K}\sum^{P} {}_{25}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_1$
nuclear mass-energy quanta per shell
1s2
Alkaline Earth
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
02
1 s 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.02 - Helium atomic config

Copyright ABRAHAM [2008] - All rights reserved

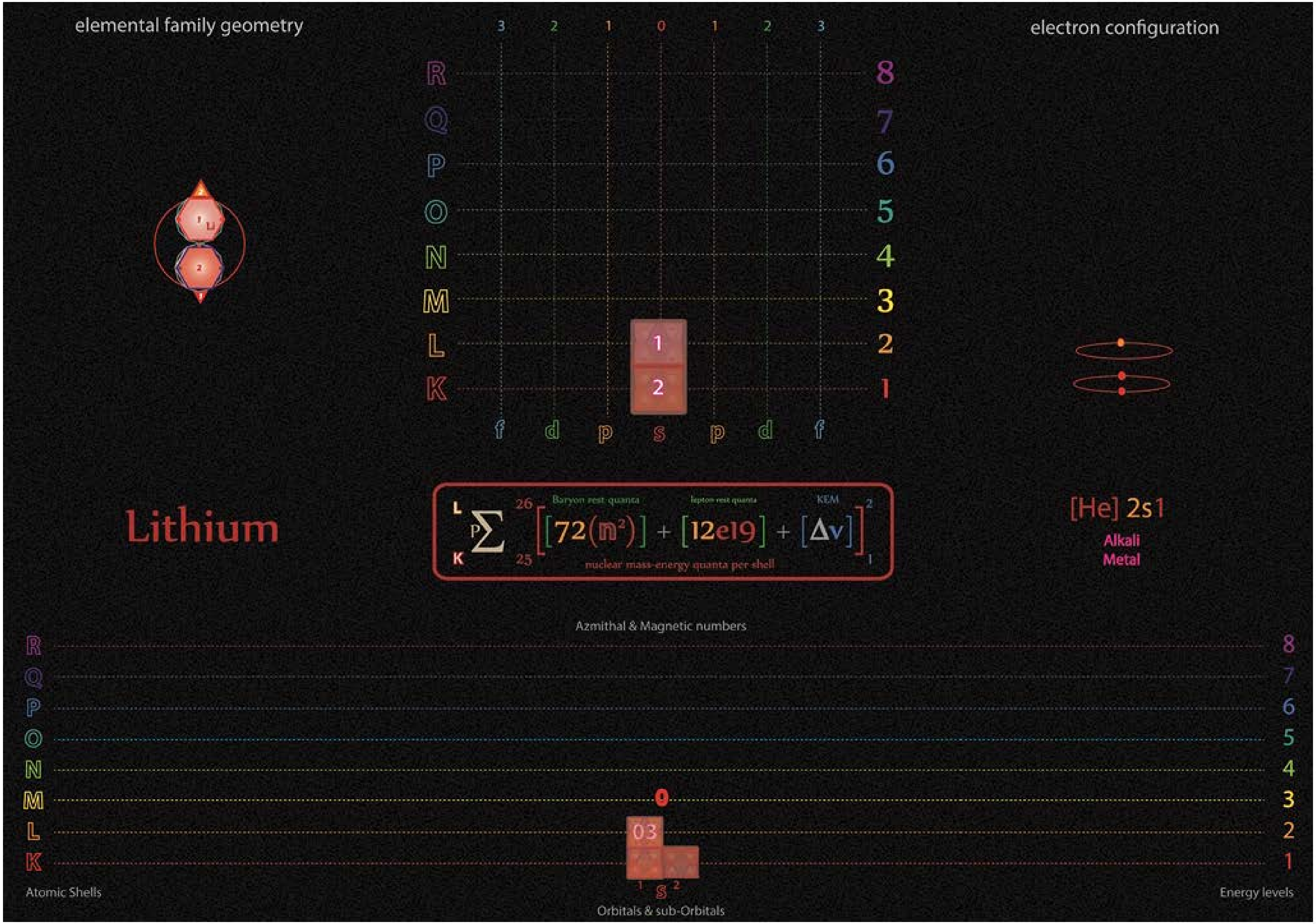
elemental family geometry
3
2
1
0
1
2
3
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
Lithium
L
K
P
26
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
2
1
nuclear mass-energy quanta per shell
[He] 2s1
Alkali
Metal
Azmithal & Magnetic numbers
0
03
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.03 - Lithium atomic config

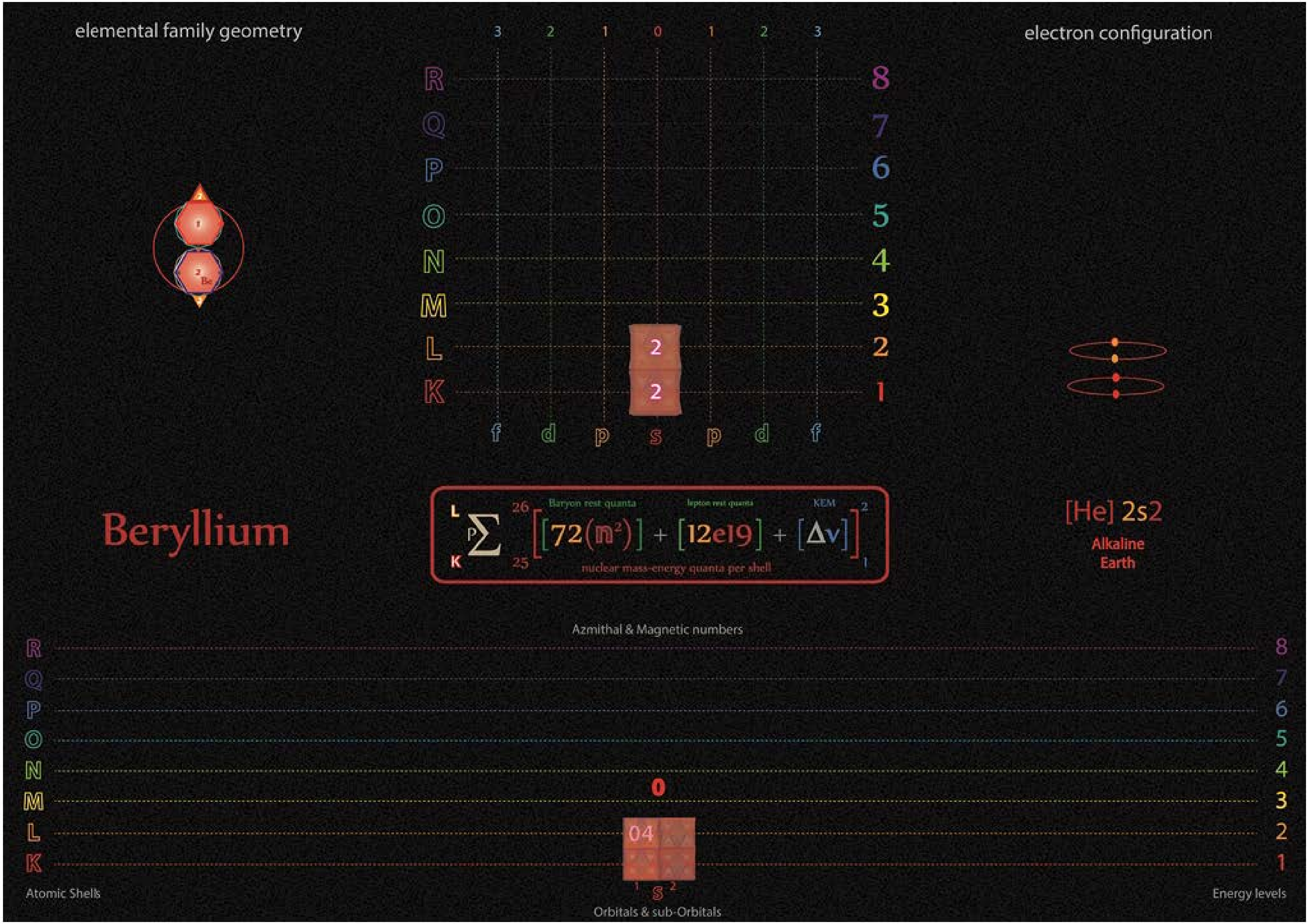

elemental family geometry

electron configuration

3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
2
f
d
p
s
p
d
f
2
1
Be

Beryllium

$$\sum_{K}^{L} {}_{p} \left[\overset{\text{Baryon rest quanta}}{\left[72(\mathrm{m}^2)\right]} + \overset{\text{lepton rest quanta}}{\left[12e19\right]} + \overset{\text{KEM}}{\left[\Delta v\right]} \right]_{25\ 1}^{26\ 2}$$

nuclear mass-energy quanta per shell

[He] 2s2

Alkaline Earth

Azmithal & Magnetic numbers

R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
04
1
s
2

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.04 - Beryllium atomic config

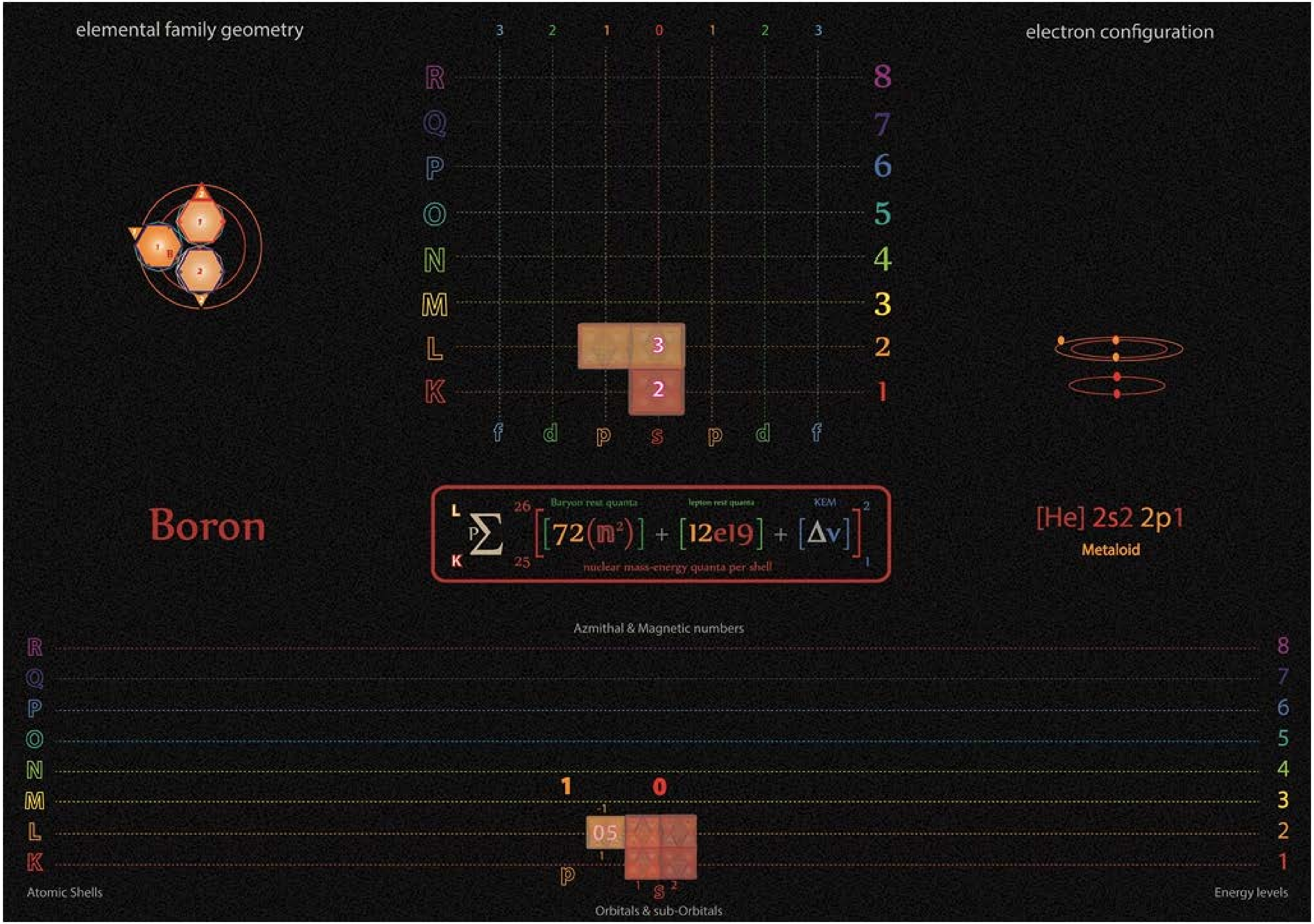
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
2
f d p s p d f
Boron
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{L} P \; ^{26}_{25}\left[[72(m^2)] + [12e19] + [\Delta v] \right]^2_1$
nuclear mass-energy quanta per shell
[He] 2s2 2p1
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1
0
-1
05
1
p
1
s
2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.05 - Boron atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

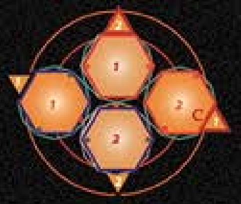

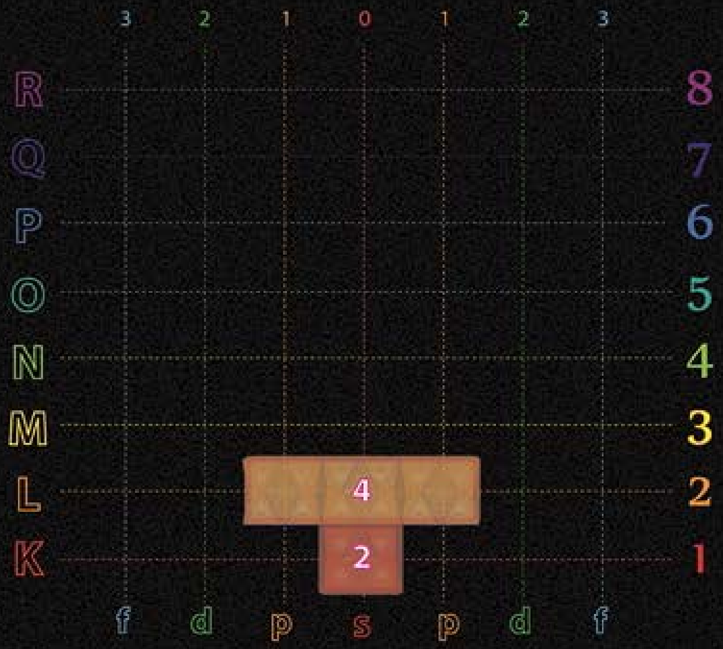
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
4
2
f
d
p
s
p
d
f

electron configuration

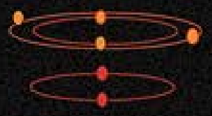

Carbon

$$\sum_{K}^{L} P \sum_{25}^{26} \left[\overset{\text{Baryon rest quanta}}{[72(m^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]_{1}^{2}$$

nuclear mass-energy quanta per shell

[He] 2s2 2p2

Metaloid

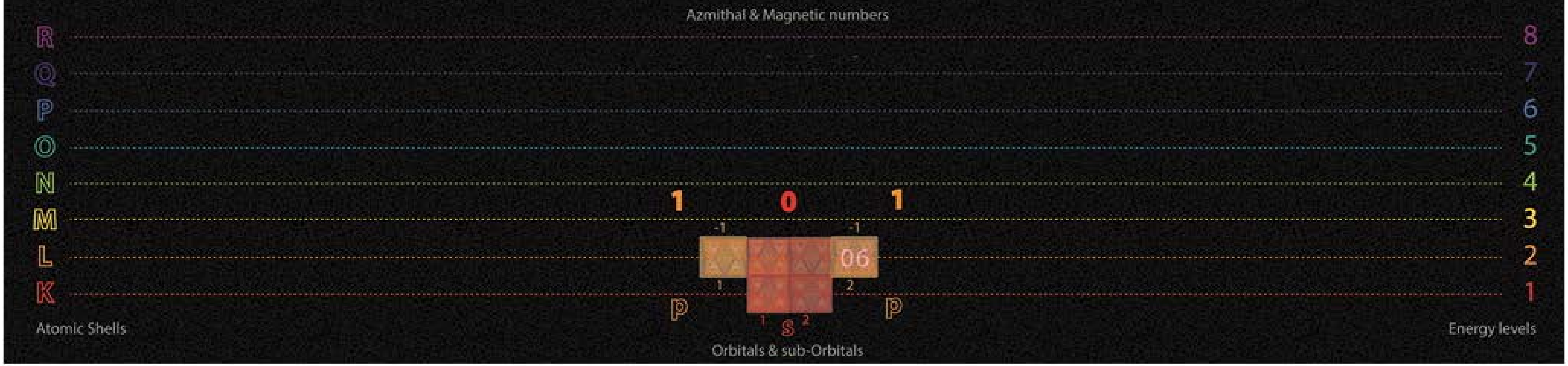
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
-1
-1
06
1
2
p
p
1
s
2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.06 - Carbon atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

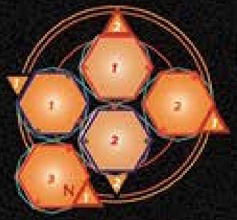

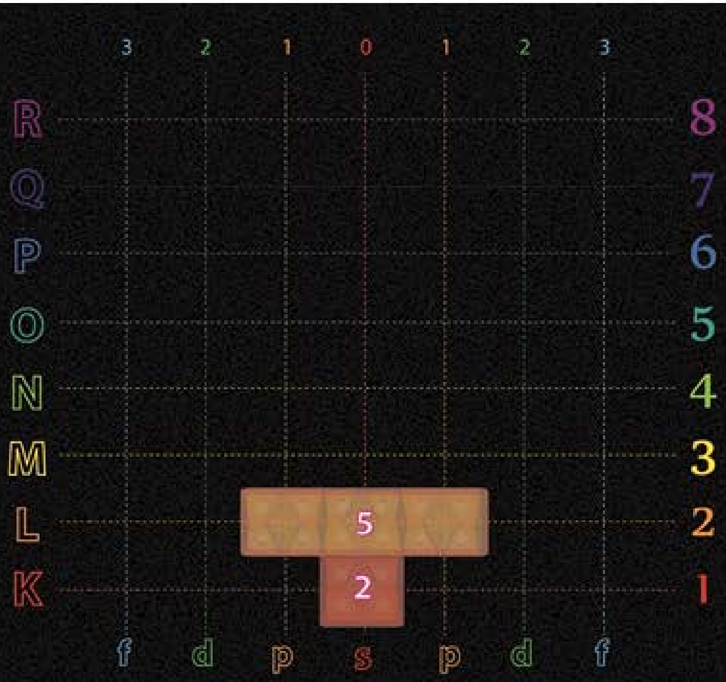
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
5
2
f
d
p
s
p
d
f

electron configuration

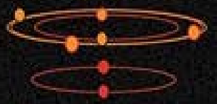

Nitrogen

$$\sum_{K}^{L}{}^{P}\left[\overset{\text{Baryon rest quanta}}{[72(m^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}\right]_{1}^{2}$$

26 / 25

nuclear mass-energy quanta per shell

[He] 2s2 2p3

Metaloid

Azmithal & Magnetic numbers

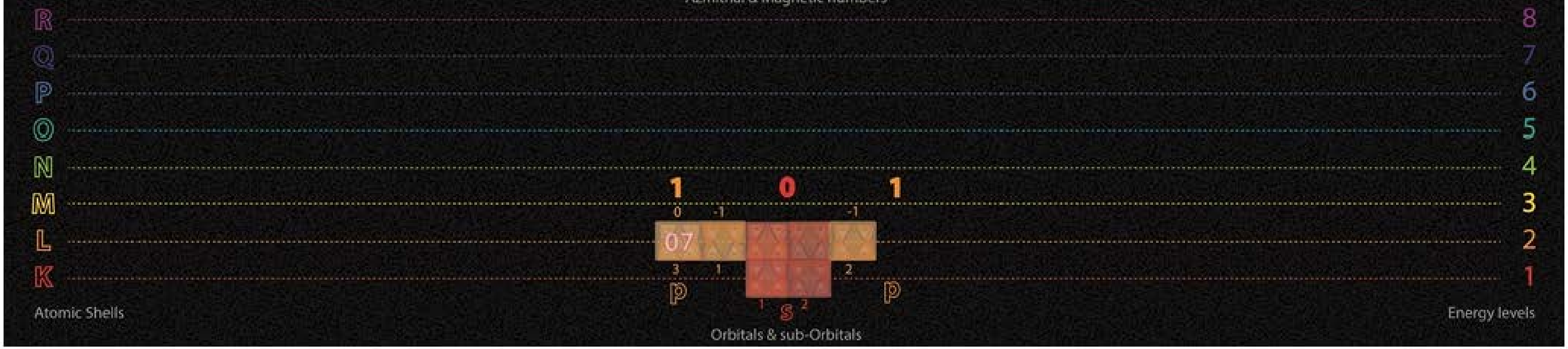
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
0
-1
-1
07
3
1
2
p
p
s
1
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.07 - Nitrogen atomic config

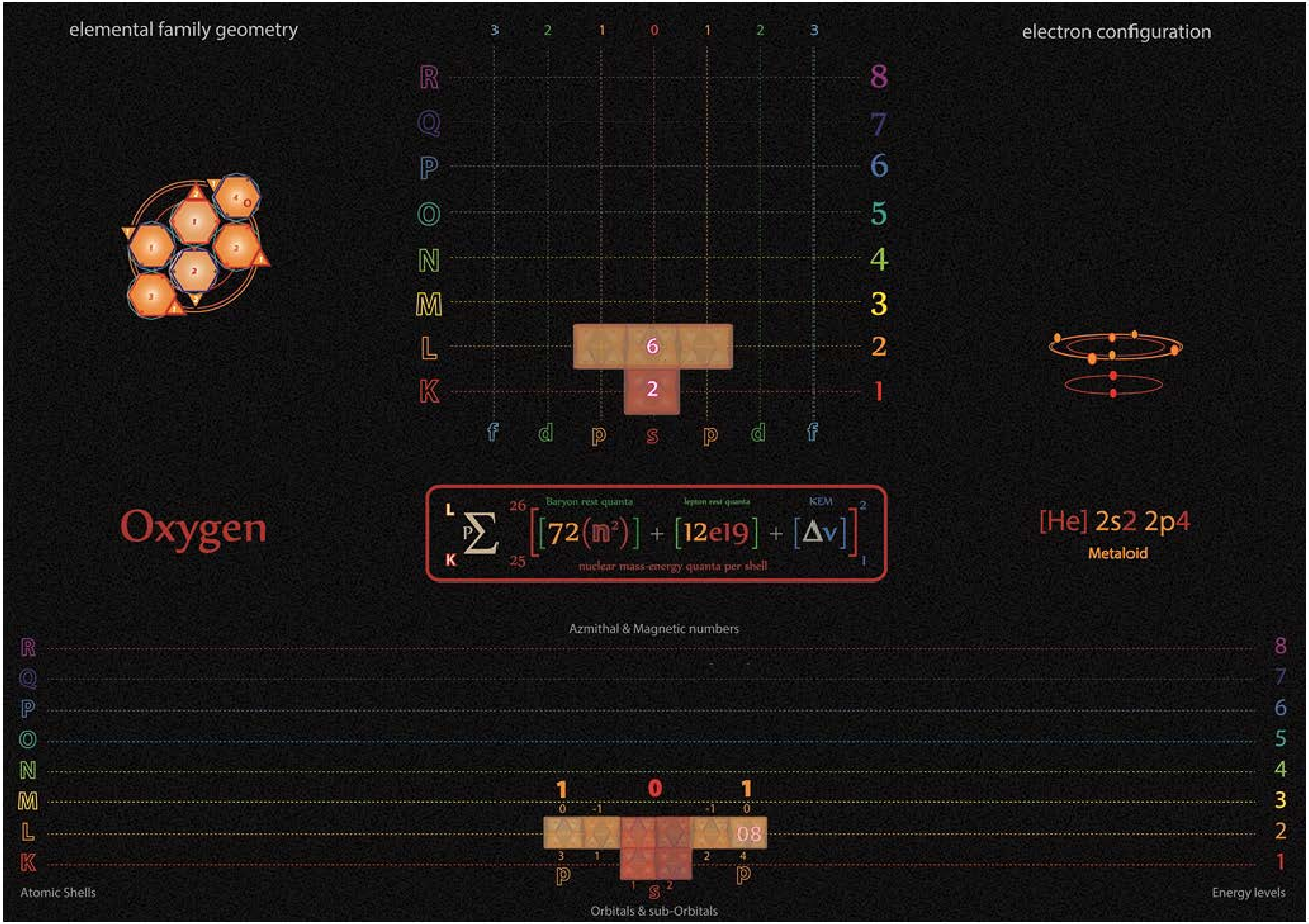
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
2
f d p s p d f
Oxygen
L
K
P
26
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[He] 2s2 2p4
Metaloid
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
08
3 1 2 4
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.08 - Oxygen atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

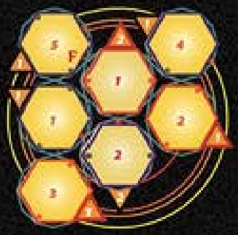

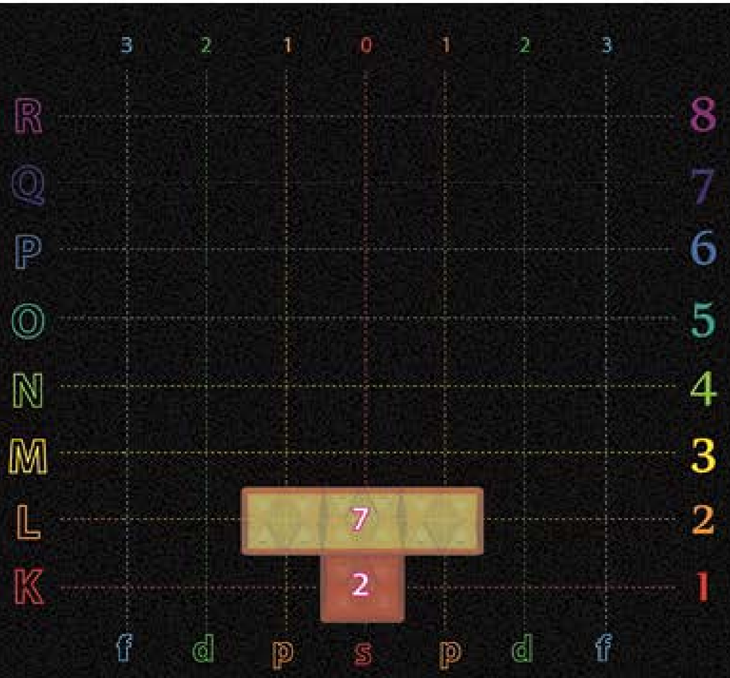
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7
2
f
d
p
s
p
d
f

electron configuration

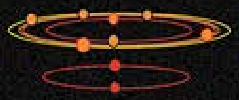

Fluorine

$$ {}^{L}_{K}\,{}_{p}\sum^{26}_{25}\left[\,[72(m^2)] + [12e19] + [\Delta v]\,\right]^{2}_{1} $$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[He] 2s2 2p5

Halogen

Azmithal & Magnetic numbers

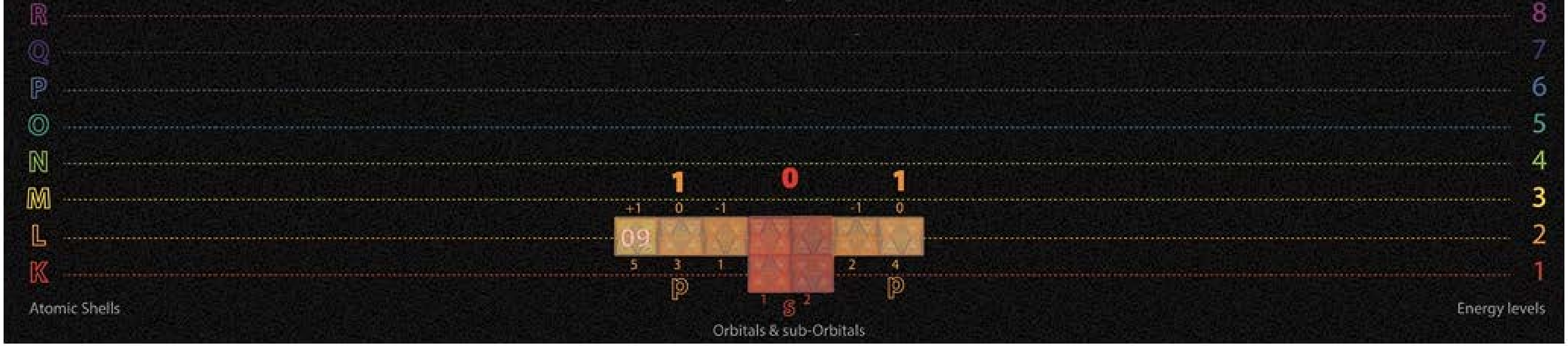
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
+1
0
-1
-1
0
09
5
3
1
2
4
p
1
s
2
p
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.09 - Fluorine atomic config

elemental family geometry

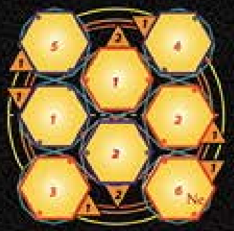
5
4
1
1
2
2
3
6
Ne

Neon

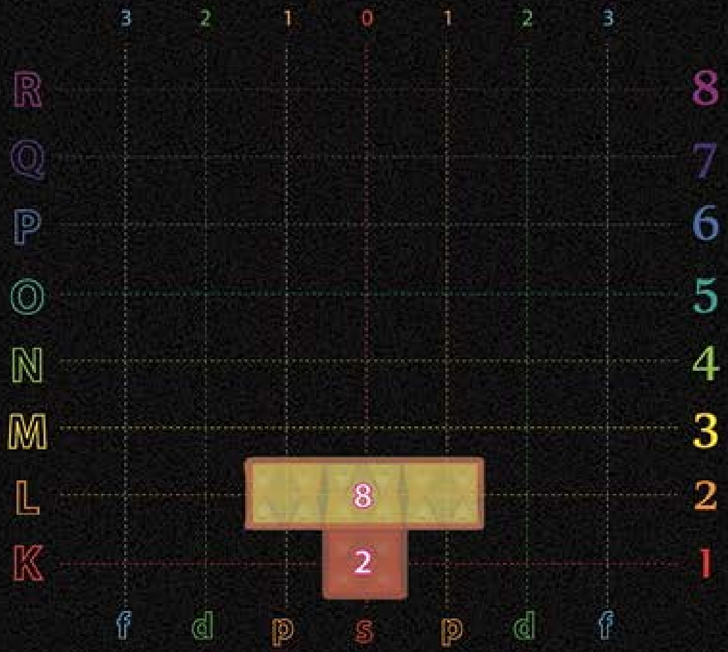
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
8
2
f
d
p
s
p
d
f

$$\sum_{K}^{L} P \left[\overset{\text{Baryon rest quanta}}{[72(n^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]_{1}^{2}$$

26 25

nuclear mass-energy quanta per shell

electron configuration

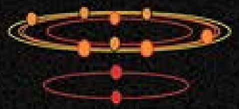

[He] 2s2 2p6

Noble Gas

Azmithal & Magnetic numbers

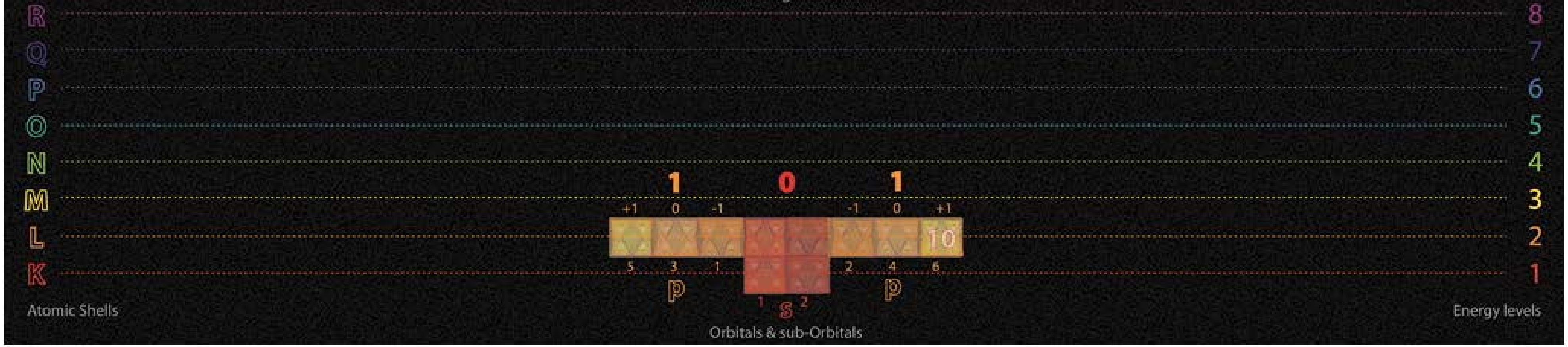
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
+1
0
-1
-1
0
+1
10
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.10 - Neon atomic config

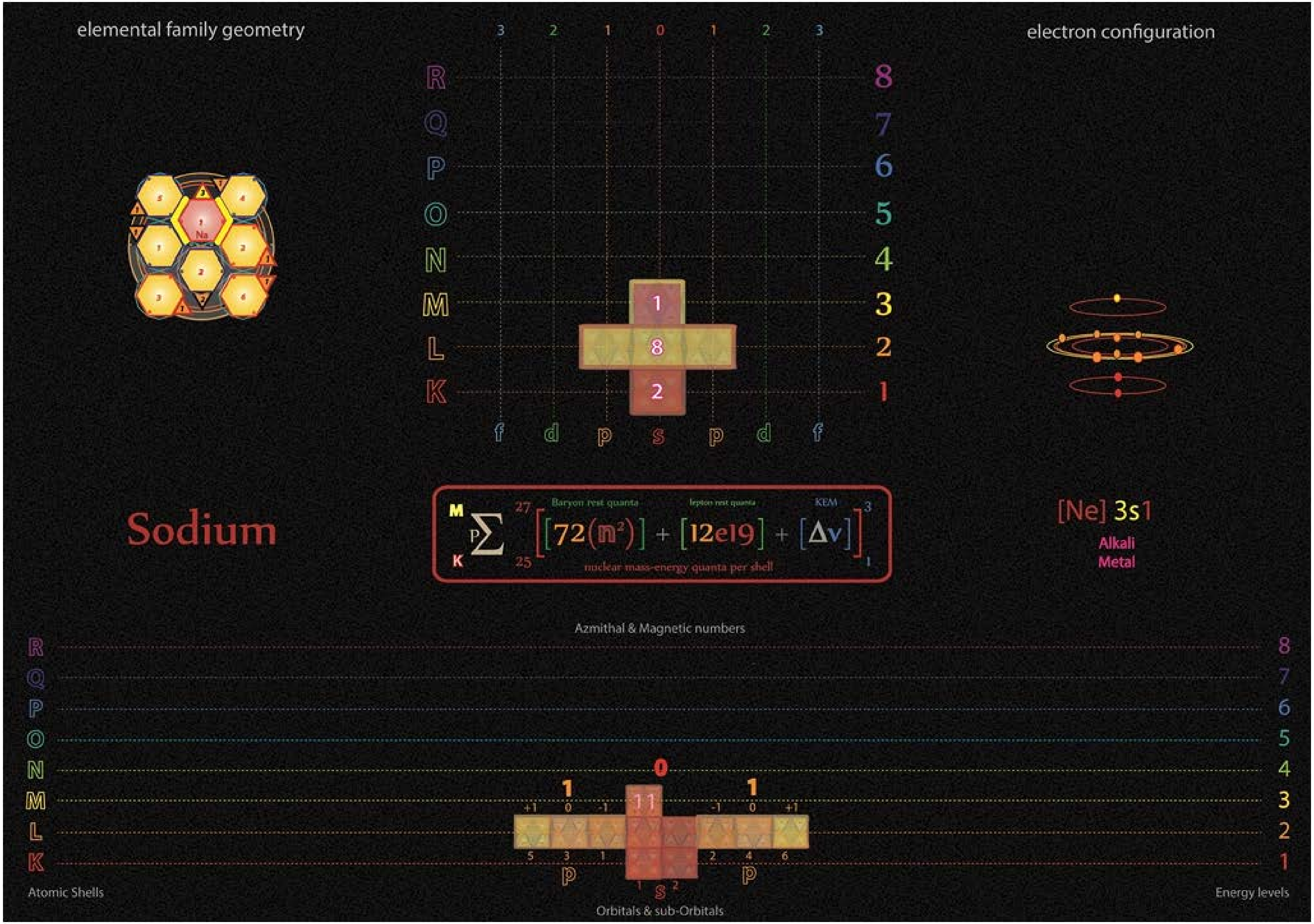
elemental family geometry
Na
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
1
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1 1
+1 0 -1 -1 0 +1
1 1
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Sodium

$$\sum_{K}^{M} P \sum_{25}^{27} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{3}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Ne] 3s1

Alkali Metal

Tetryonics 53.11 - Sodium atomic config

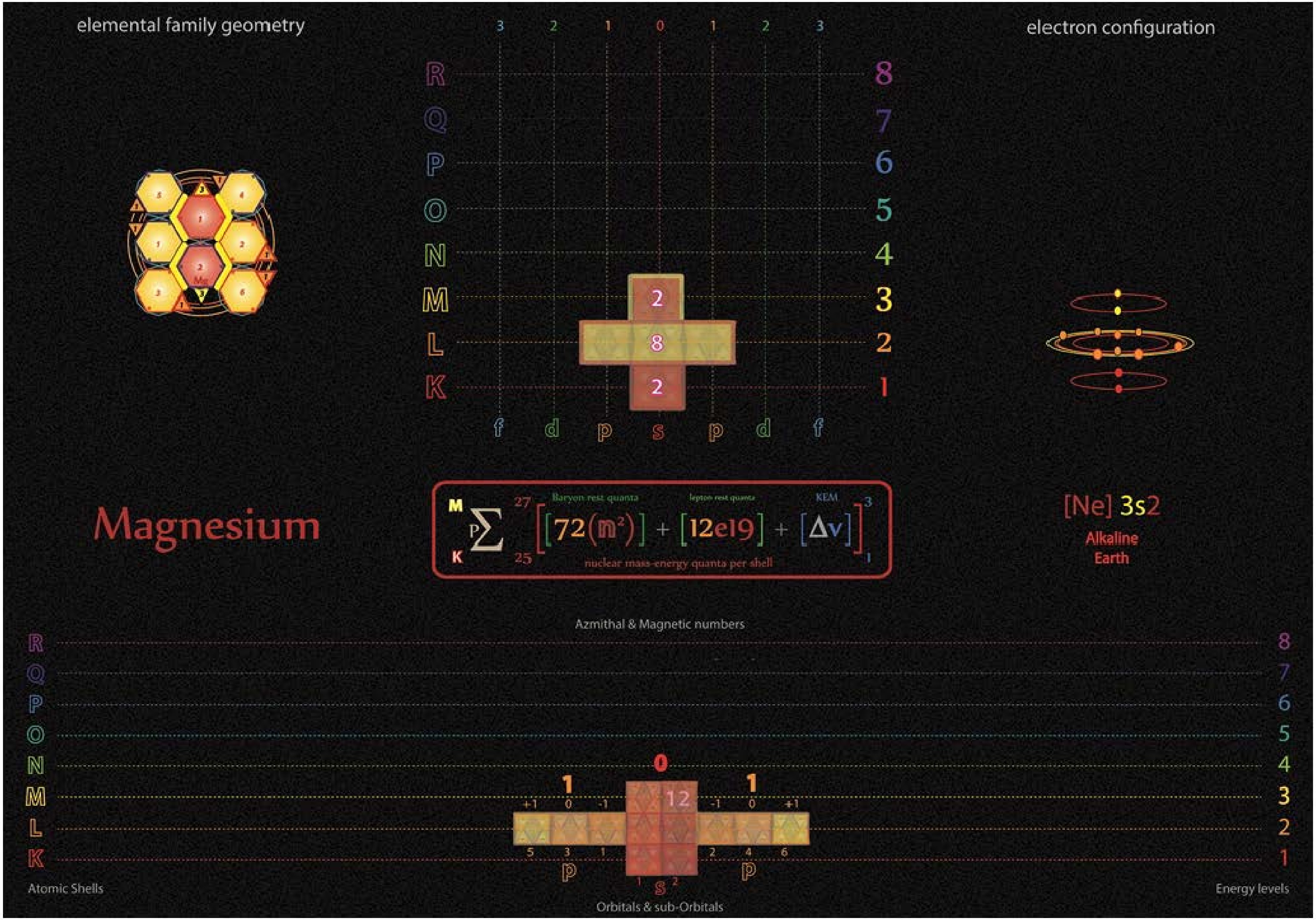
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
2
f d p s p d f
Magnesium
M
K
27
25
Baryon rest quanta
lepton rest quanta
KEM
3
1
nuclear mass-energy quanta per shell
[Ne] 3s2
Alkaline
Earth
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
1
12
+1 0 -1
-1 0 +1
5 3 1
2 4 6
p
s
1 2
p
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\overset{M}{\underset{K}{}}\,p\sum_{25}^{27}\left[\left[72(m^2)\right]+\left[12e19\right]+\left[\Delta v\right]\right]_1^3$$

Tetryonics 53.12 - Magnesium atomic config

Copyright ABRAHAM [2008] - All rights reserved

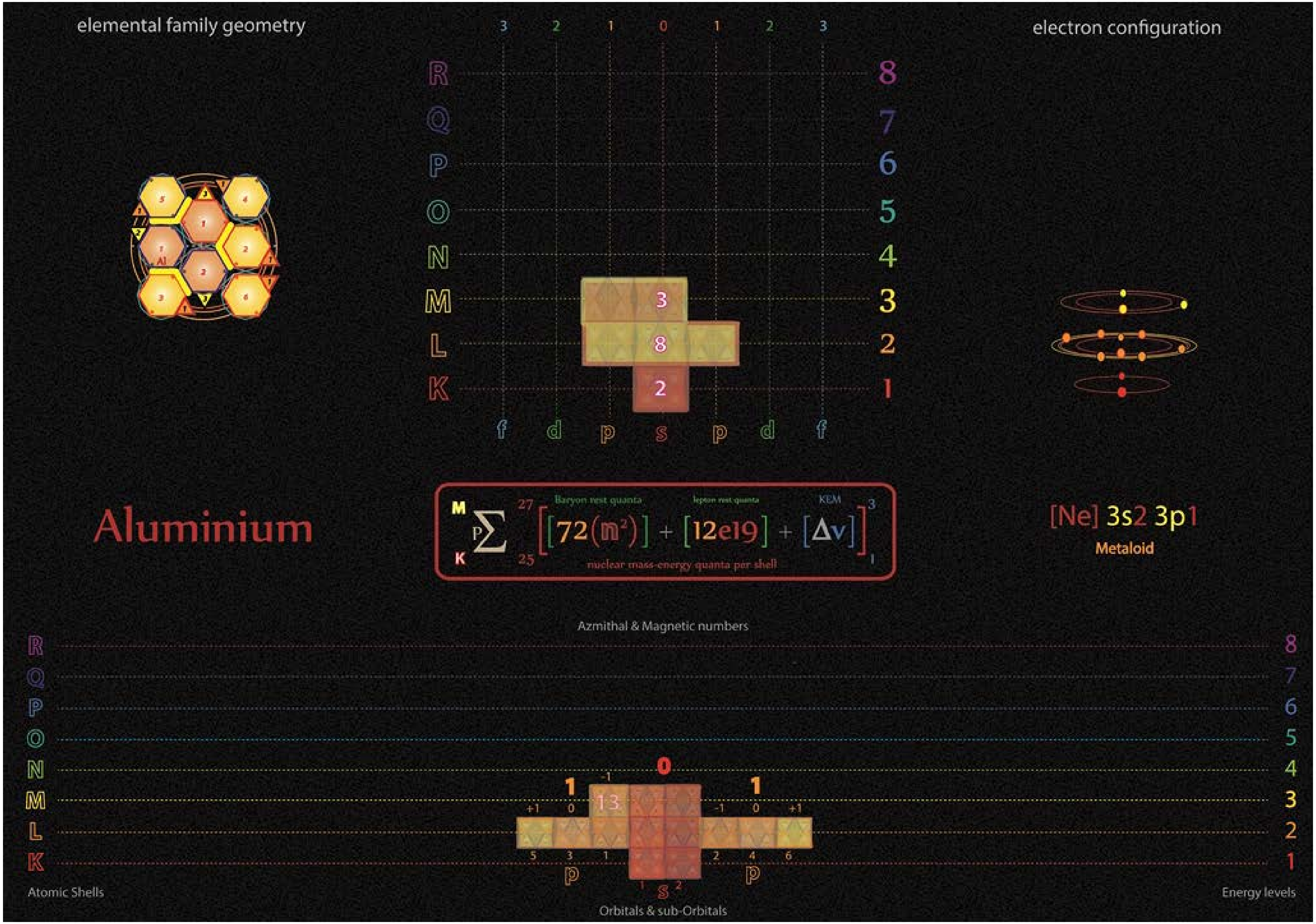
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
8
2
f d p s p d f
Aluminium
Baryon rest quanta
lepton rest quanta
KEM
$\overset{M}{\underset{K}{P\sum}}\sum_{25}^{27}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_1^3$
nuclear mass-energy quanta per shell
[Ne] 3s2 3p1
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
-1
+1 0 -1 0 +1
13
5 3 1 2 4 6
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.13 - Aluminium atomic config

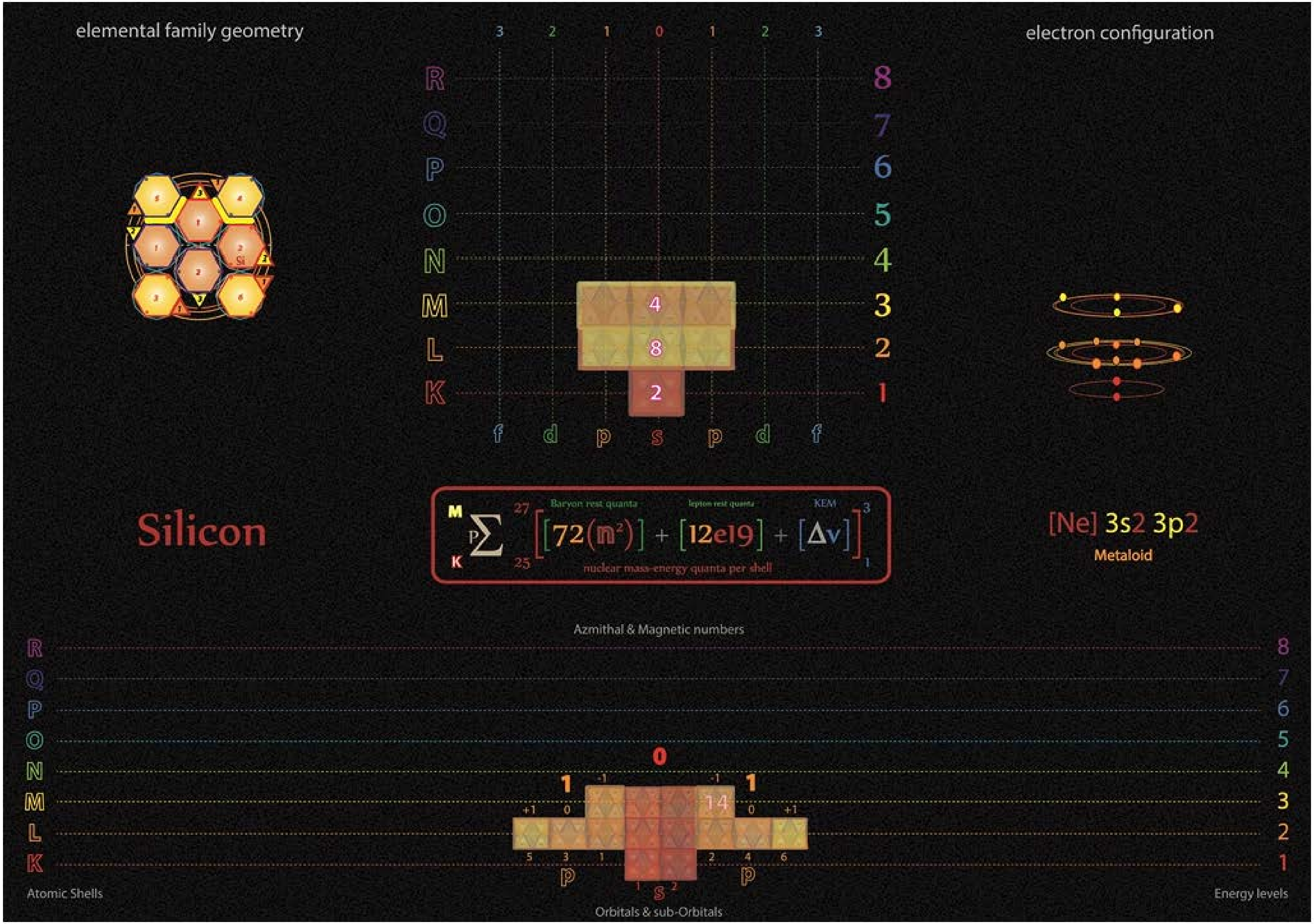
elemental family geometry
Si
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
4
8
2
f d p s p d f
electron configuration
Silicon
Baryon rest quanta
lepton rest quanta
KEM
M
K
P∑
27
25
[72(m²)] + [12e19] + [Δv]
3
1
nuclear mass-energy quanta per shell
[Ne] 3s2 3p2
Metaloid
Azmithal & Magnetic numbers
0
1 1
-1 -1
+1 0 0 +1
14
5 3 1 2 4 6
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.14 - Silicon atomic config

Copyright ABRAHAM [2008] - All rights reserved

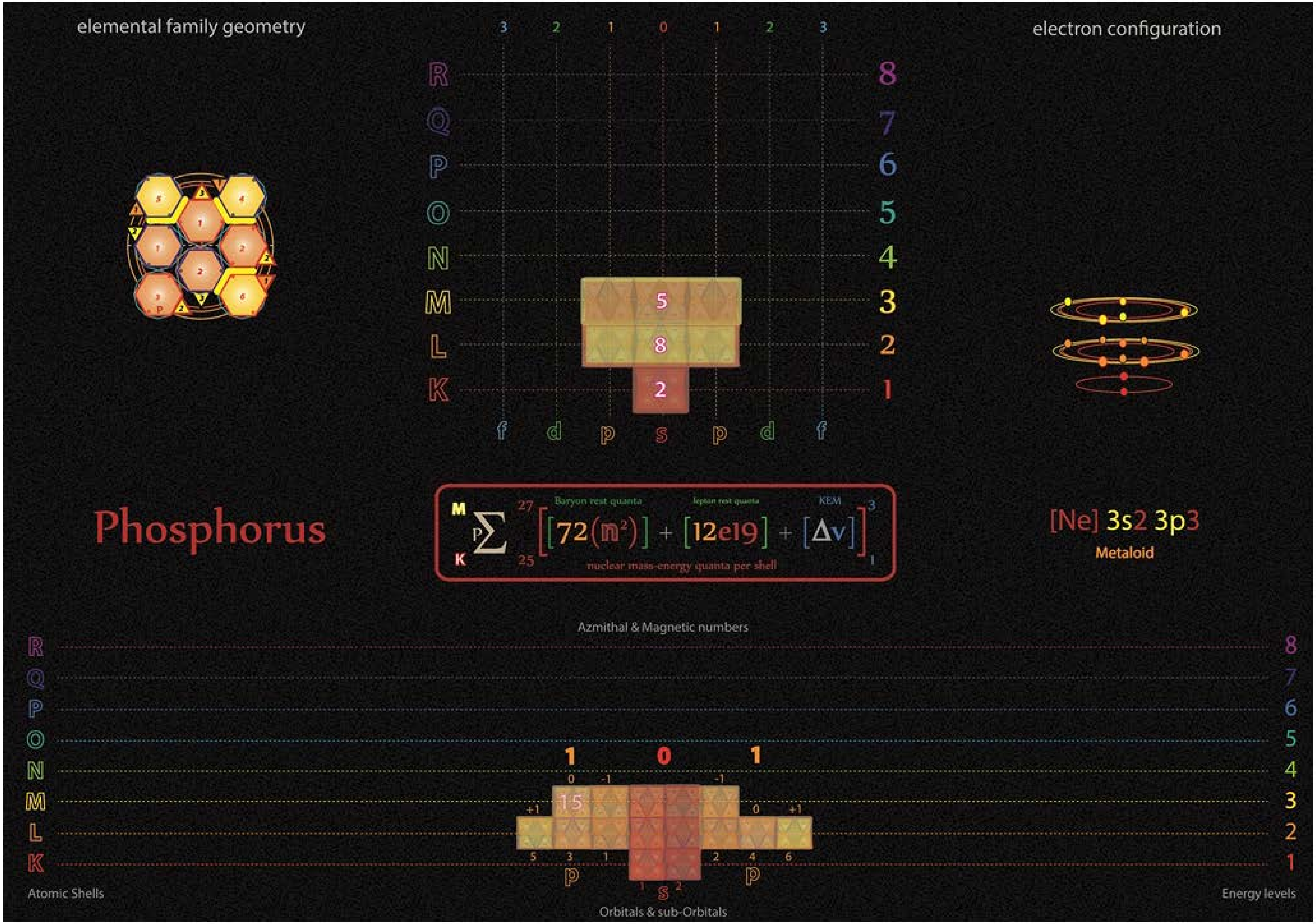

elemental family geometry

electron configuration

Phosphorus

$$\sum_{K}^{M} {}_{P}\sum_{25}^{27} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{3}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Ne] 3s2 3p3

Metaloid

Azmithal & Magnetic numbers

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
5
8
2
f d p s p d f
1 0 1
0 -1 -1
15
+1 0 +1
5 3 1 2 4 6
p s p
1 2

Tetryonics 53.15 - Phosphorus atomic config

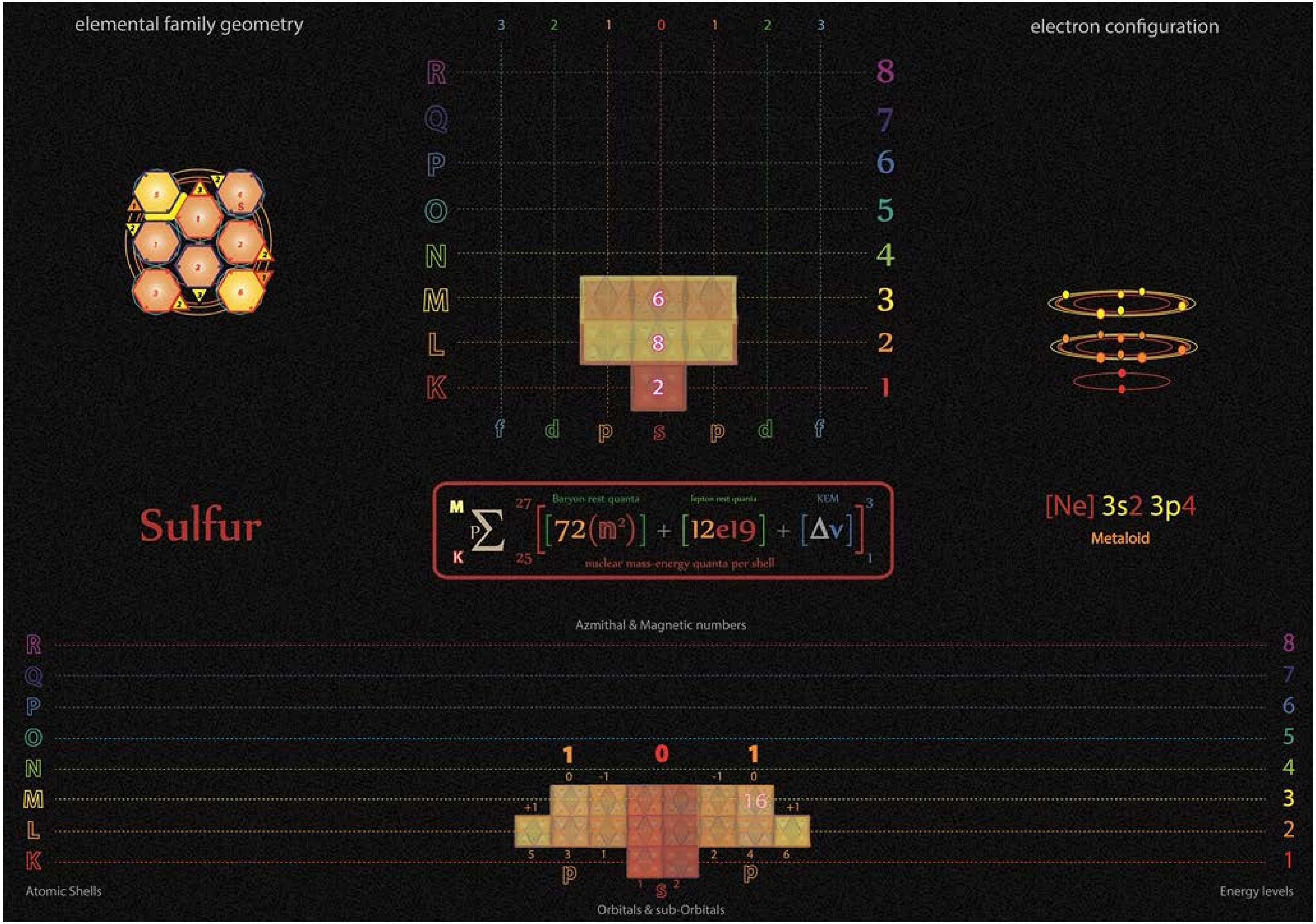
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
8
2
f d p s p d f
Sulfur
[Ne] 3s2 3p4
Metaloid
M
K
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
0 -1 -1 0
+1 16 +1
5 3 1 2 4 6
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{M} {}_{P}\sum_{25}^{27}\left[\left[72(\mathfrak{m}^2)\right] + \left[12e19\right] + \left[\Delta v\right]\right]_1^3$$

Tetryonics 53.16 - Sulfur atomic config

Copyright ABRAHAM [2008] - All rights reserved

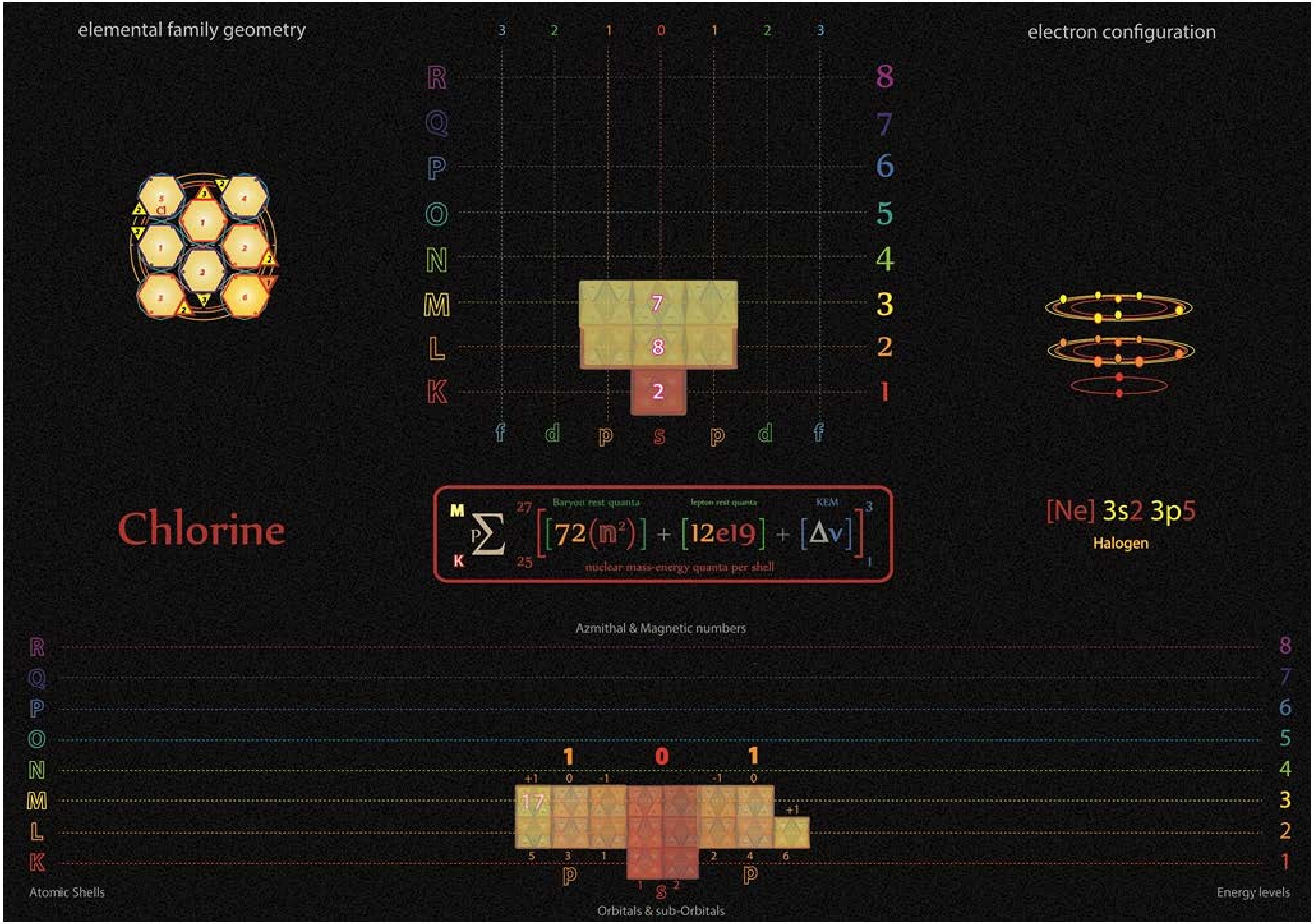
elemental family geometry
Cl
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
7
8
2
f d p s p d f
electron configuration
Chlorine
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Ne] 3s2 3p5
Halogen
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0 +1
17
5 3 1 2 4 6
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{M} P \sum_{25}^{27} \left[[72(n^2)] + [12e19] + [\Delta v] \right]_{1}^{3}$$

Tetryonics 53.17 - Chlorine atomic config

Copyright ABRAHAM [2008] - All rights reserved

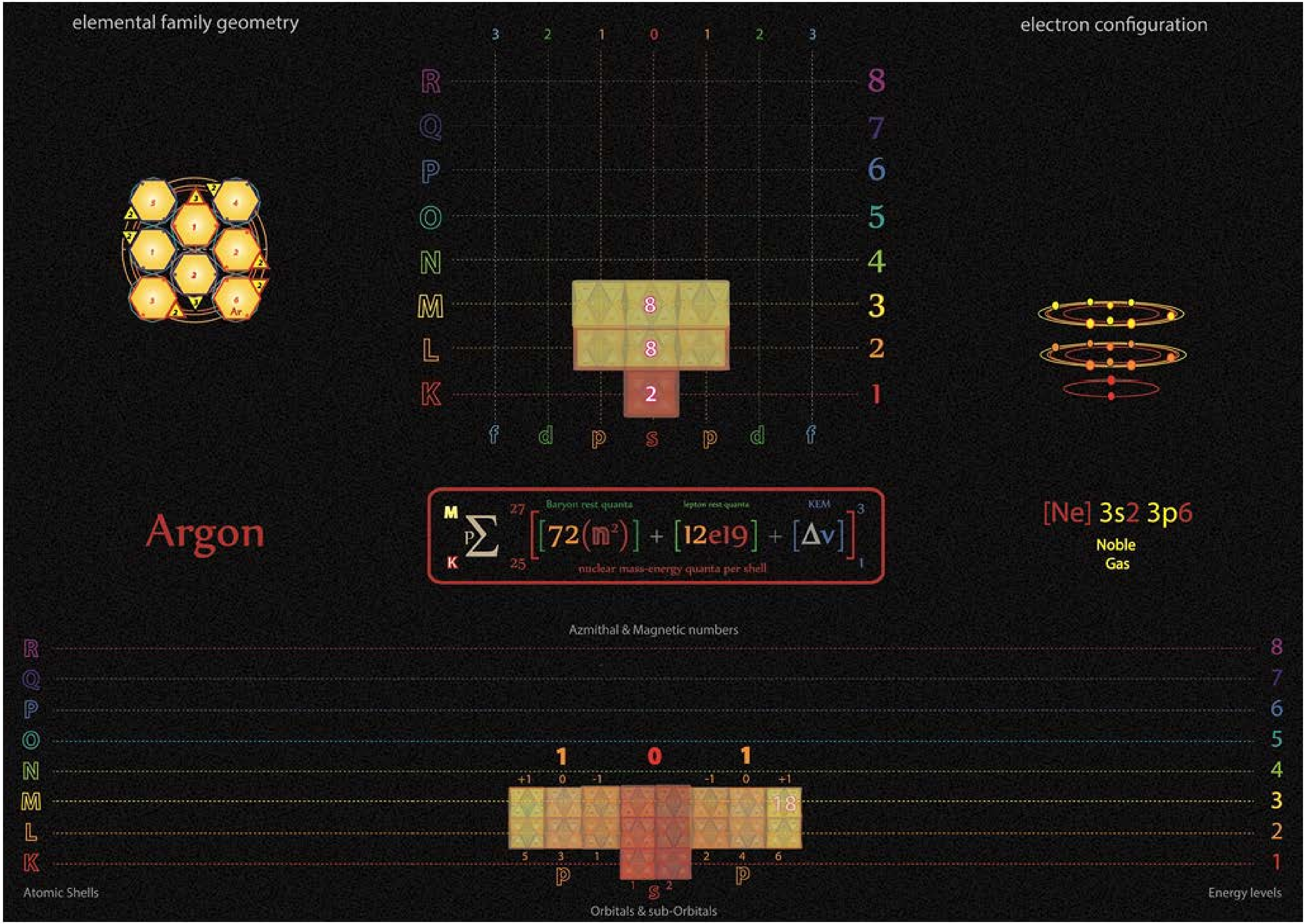
elemental family geometry
electron configuration
Argon
Ar
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
8
8
2
[Ne] 3s2 3p6
Noble Gas
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
18
5 3 1 2 4 6
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{M} {}_{p}\sum_{25}^{27} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_1^3$$

Tetryonics 53.18 - Argon atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

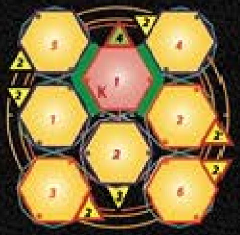
K
1
1
2
2
3
3
4
4
5
6

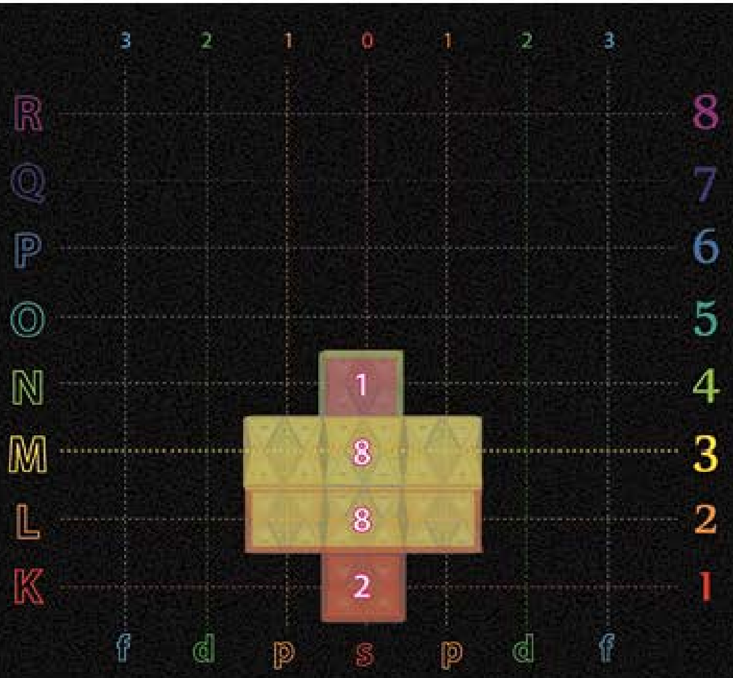
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
8
8
2
f
d
p
s
p
d
f

electron configuration

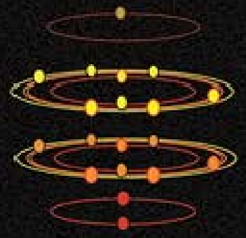

Potassium

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{N} P \; {}_{25}^{28}\Big[\,[72(m^2)] + [12e19] + [\Delta v]\,\Big]_{1}^{4}$$

nuclear mass-energy quanta per shell

[Ar] 4s1

Alkali
Metal

Azmithal & Magnetic numbers

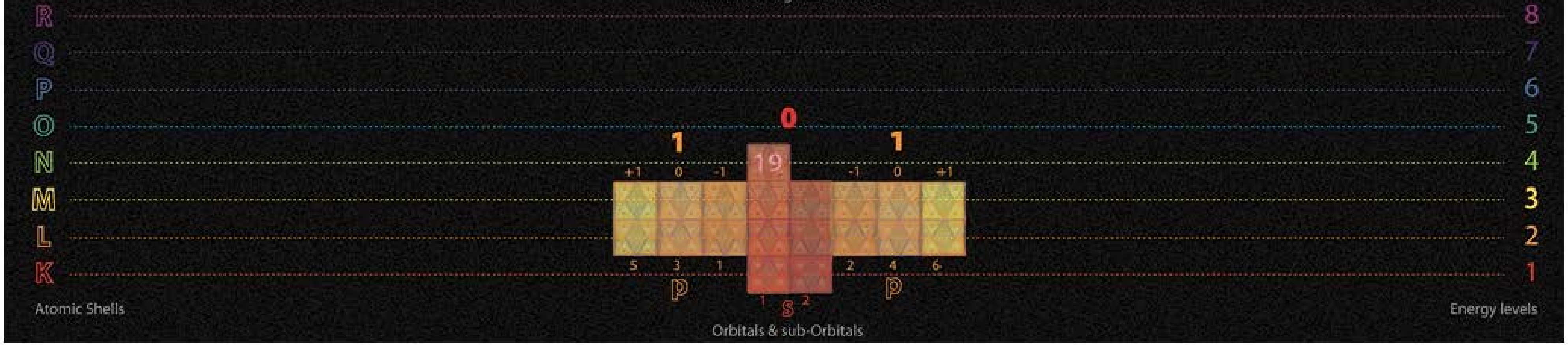
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
19
+1
0
-1
-1
0
+1
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.19 - Potassium atomic config

Copyright ABRAHAM [2008] - All rights reserved

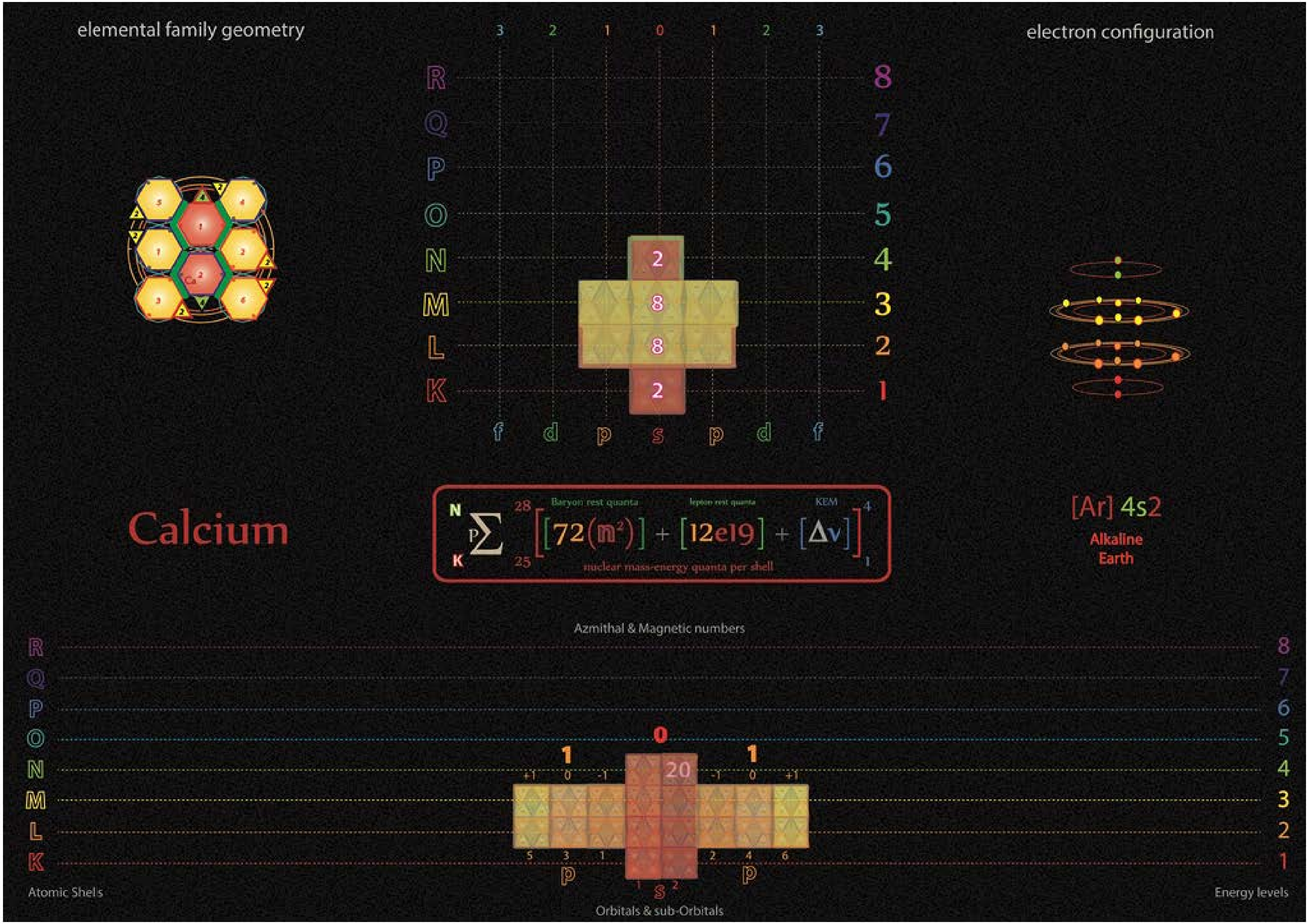
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
8
2
f d p s p d f
Calcium
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Ar] 4s2
Alkaline Earth
Azmithal & Magnetic numbers
0
1
1
20
+1 0 -1 -1 0 +1
5 3 1 2 4 6
p
s
p
1 2
Atomic Shels
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{N} p \,_{25}^{28}\left[[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\right]_{1}^{4}$$

Tetryonics 53.20 - Calcium atomic config

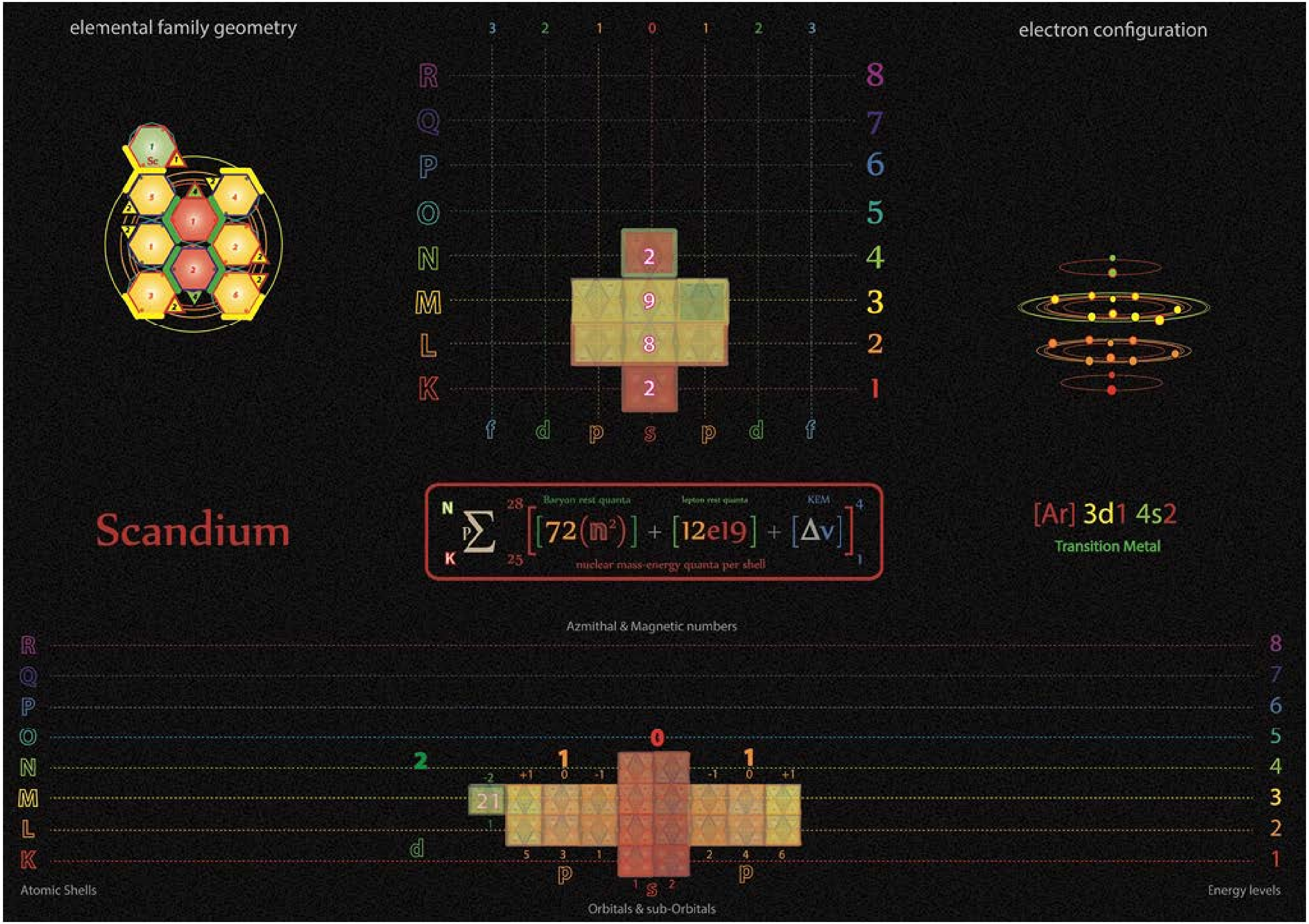
elemental family geometry
3 2 1 0 1 2 3
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
2
9
8
2
f d p s p d f
Sc
Azmithal & Magnetic numbers
2 1 0 1
-2 +1 0 -1 -1 0 +1
21
1
d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Scandium

N K P

$$\sum_{25}^{28}\left[72(\mathrm{m}^2)\right]+\left[12e19\right]+\left[\Delta v\right]_{1}^{4}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Ar] 3d1 4s2

Transition Metal

Tetryonics 53.21 - Scandium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

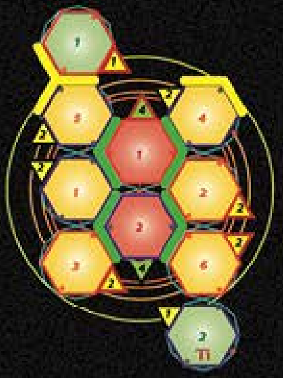

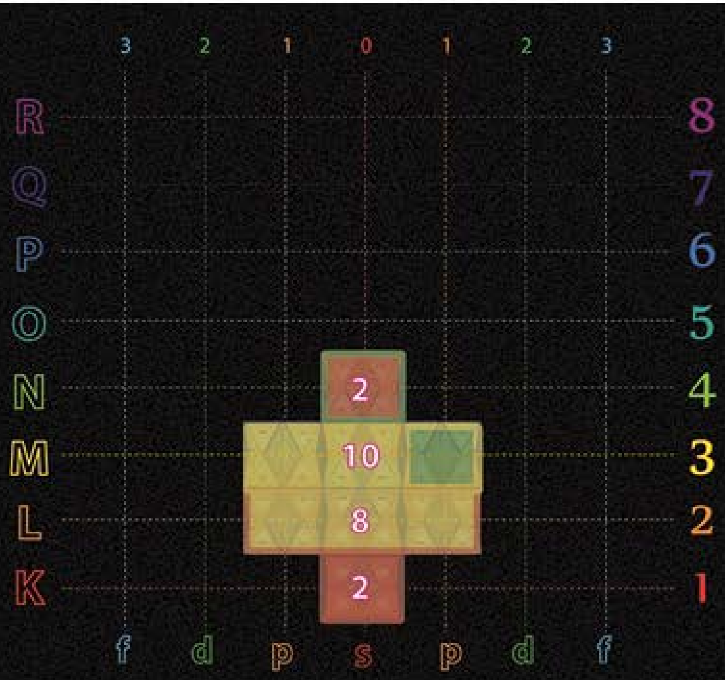
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
10
8
2
f
d
p
s
p
d
f

electron configuration

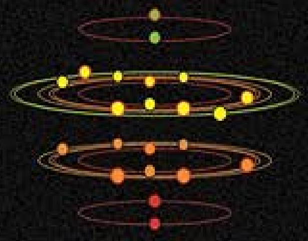

Titanium

Baryon rest quanta lepton rest quanta KEM

$$\sum_{K}^{N} P \; {}^{28}_{25}\left[\,[72(m^2)] + [12e19] + [\Delta v]\,\right]^{4}_{1}$$

nuclear mass-energy quanta per shell

[Ar] 3d2 4s2

Transition Metal

Azmithal & Magnetic numbers

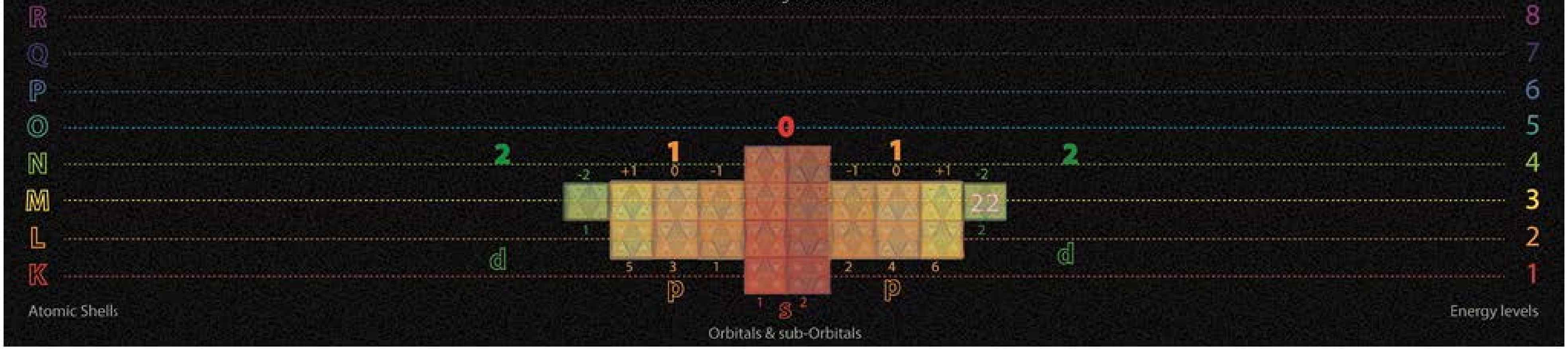
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
-2
+1
0
-1
-1
0
+1
-2
22
1
2
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.22 - Titanium atomic config

elemental family geometry

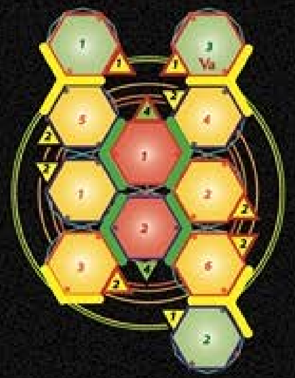

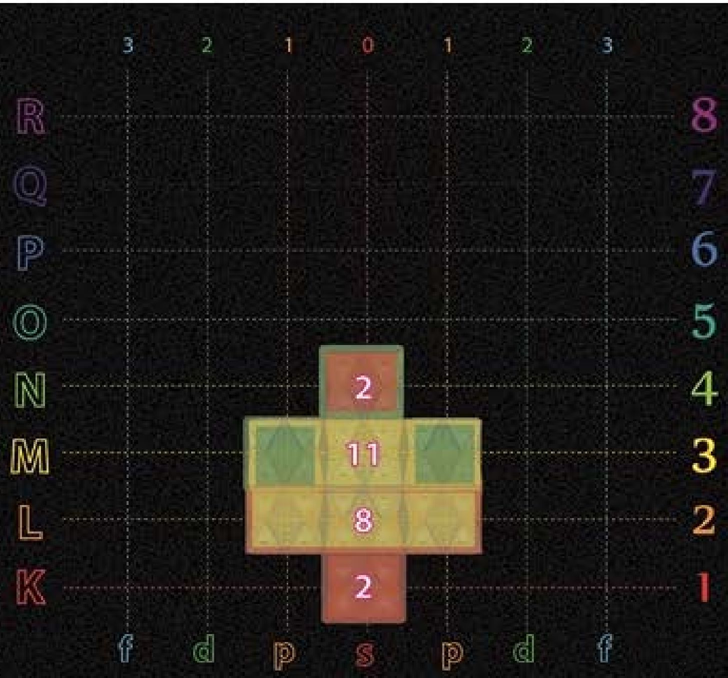
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
11
8
2
f
d
p
s
p
d
f

electron configuration

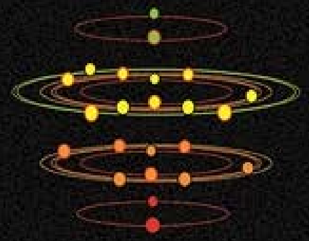

Vanadium

$$\sum_{K}^{N} P \sum_{25}^{28}\left[\overset{\text{Baryon rest quanta}}{[72(\mathrm{m}^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}\right]_{1}^{4}$$

nuclear mass-energy quanta per shell

[Ar] 3d3 4s2

Transition Metal

Azmithal & Magnetic numbers

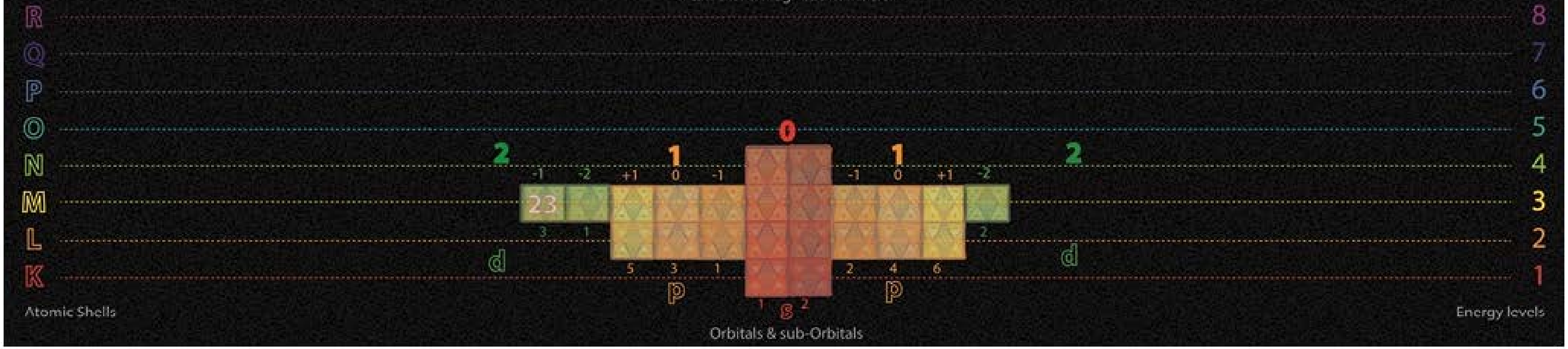
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
-1
-2
+1
0
-1
-1
0
+1
-2
23
3
1
2
5
3
1
2
4
6
d
p
s
1
2
p
d

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

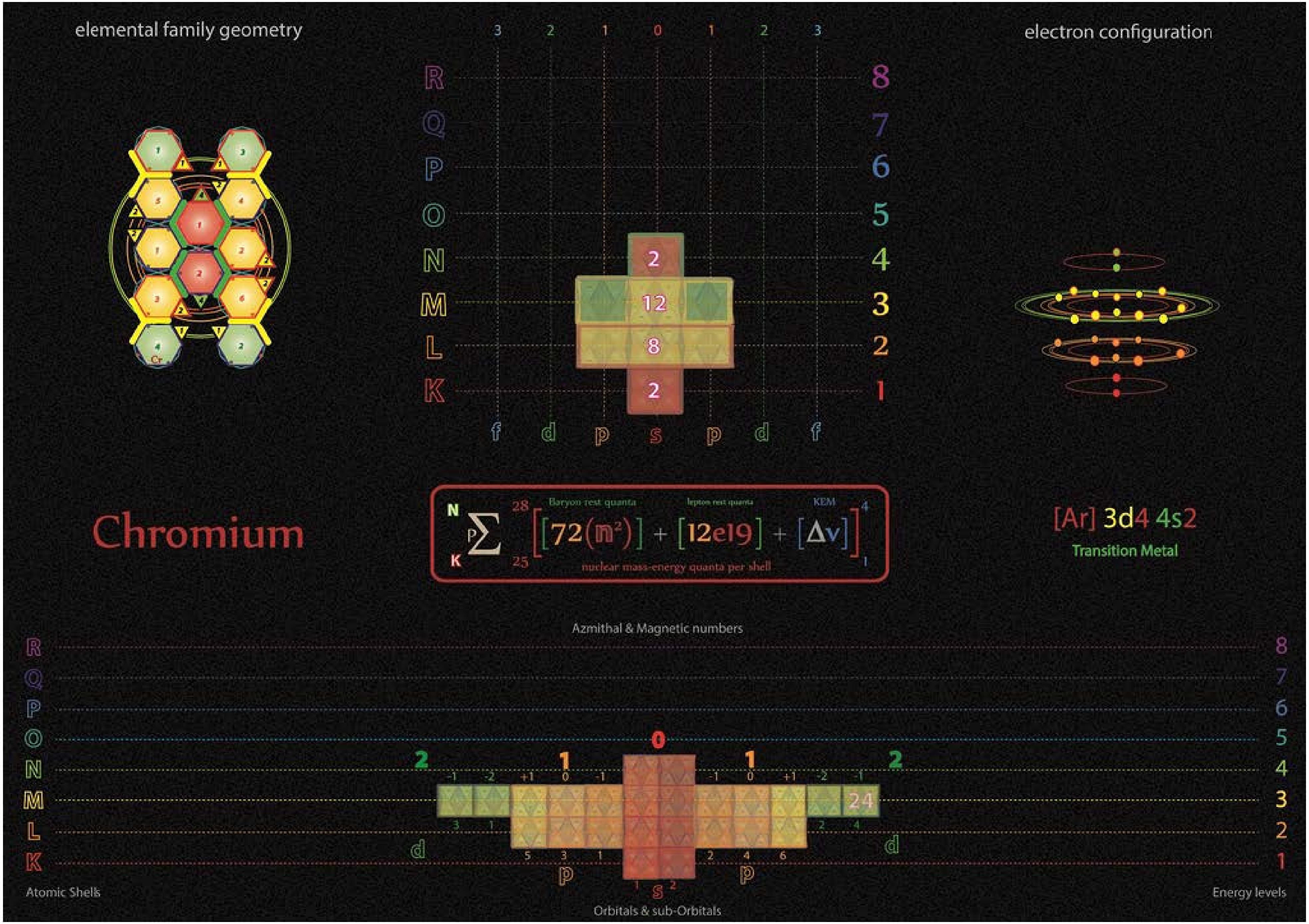
elemental family geometry
Cr
R Q P O N M L K
3 2 1 0 1 2 3
2
12
8
2
f d p s p d f
8 7 6 5 4 3 2 1
electron configuration
Chromium
N
K
P
28
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
4
1
nuclear mass-energy quanta per shell
[Ar] 3d4 4s2
Transition Metal
Azmithal & Magnetic numbers
2 1 0 1 2
-1 -2 +1 0 -1 -1 0 +1 -2 -1
24
3 1 2 4
5 3 1 2 4 6
d p s p d
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.24 - Chromium atomic config

elemental family geometry

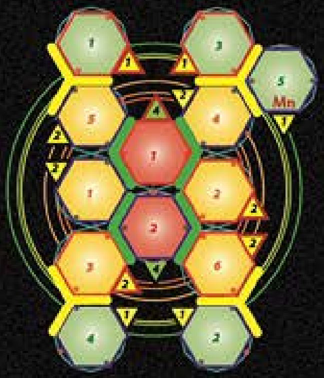

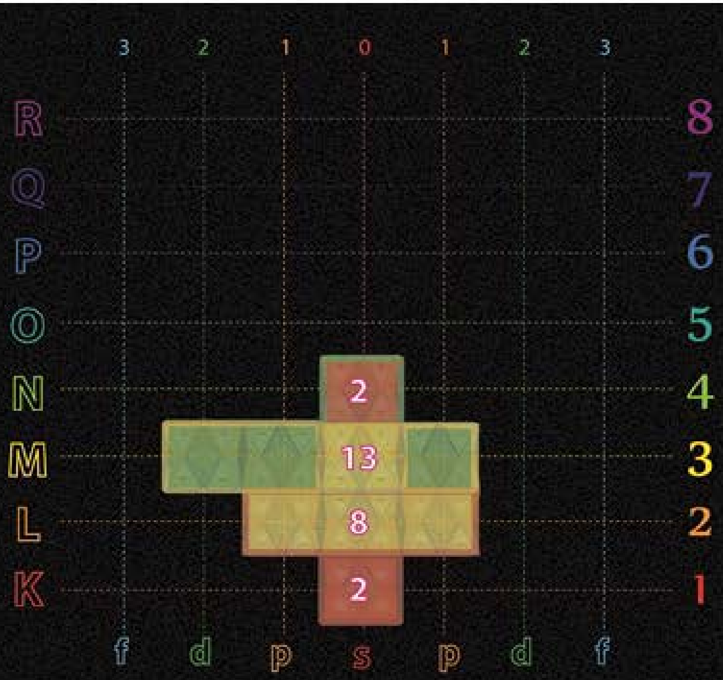
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
13
8
2
f
d
p
s
p
d
f

electron configuration

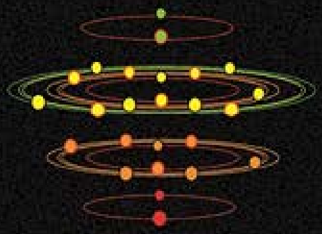

Manganese

$$\sum_{K}^{N} {}_{P} \, {}^{28}_{25}\left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{4}$$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Ar] 3d5 4s2

Transition Metal

Azmithal & Magnetic numbers

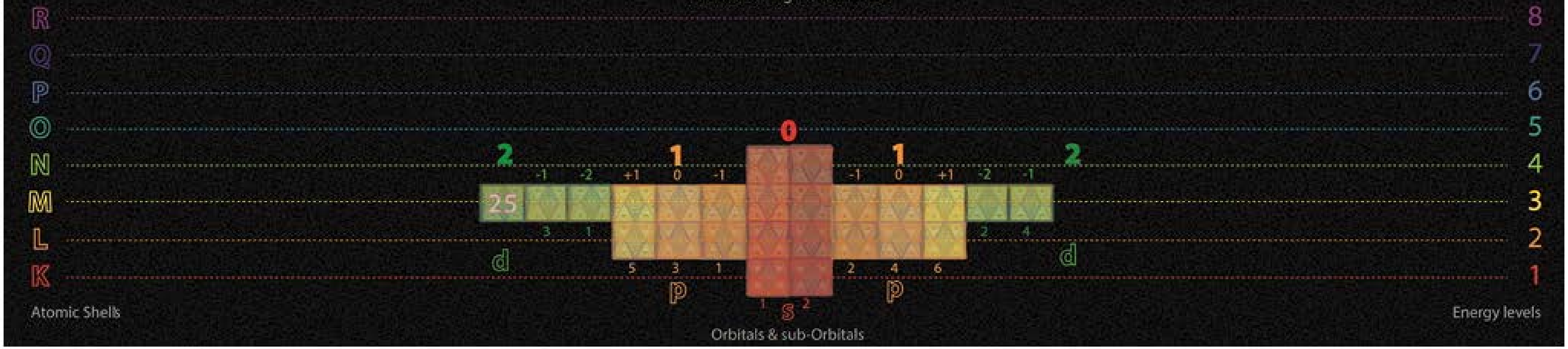
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
-1
-2
+1
0
-1
-1
0
+1
-2
-1
25
3
1
2
4
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.25 - Manganese atomic config

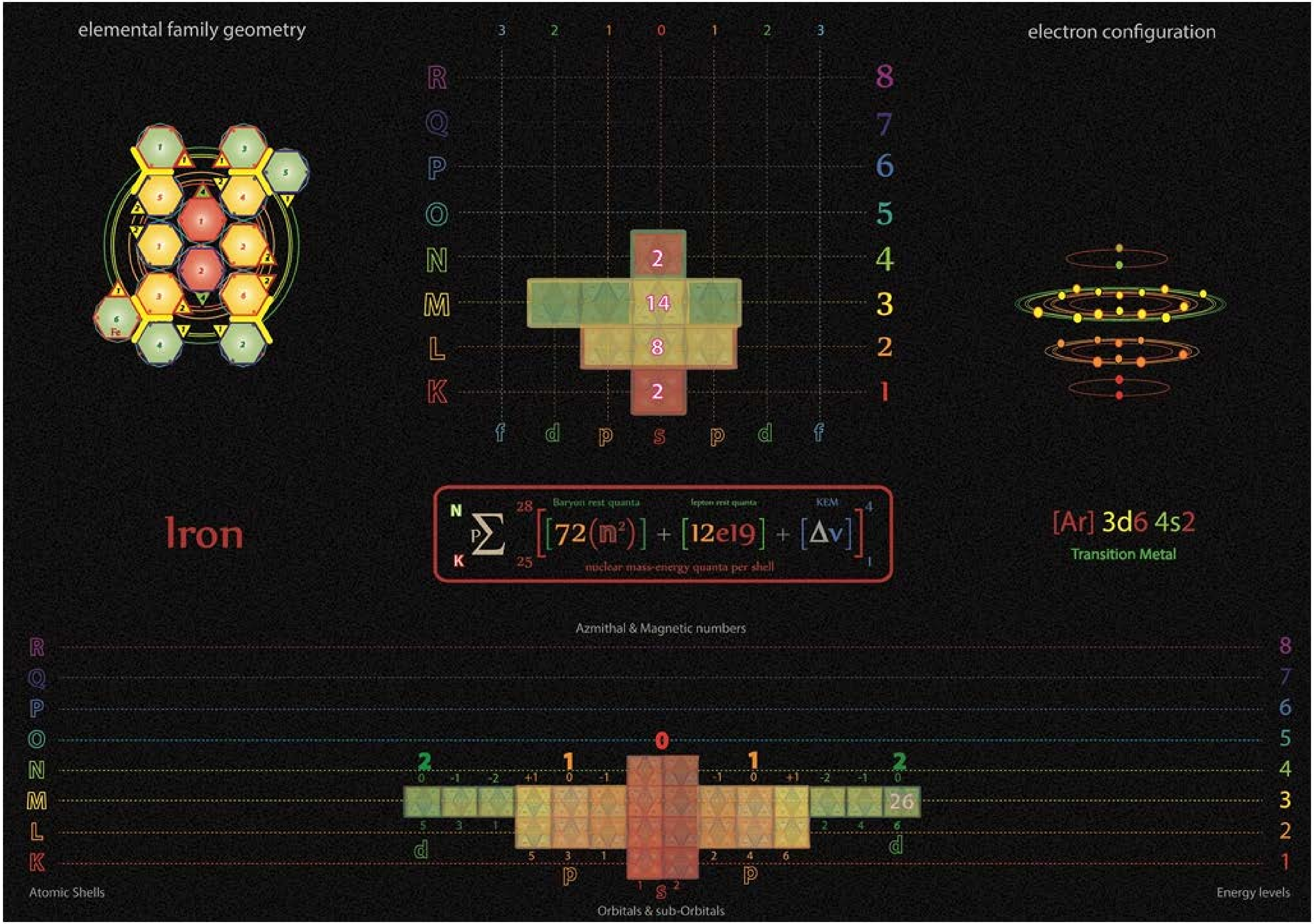
elemental family geometry
Fe
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
14
8
2
f d p s p d f
electron configuration
Iron
N
K
P
28
25
Baryon rest quanta
lepton rest quanta
KEM
$\sum [72(n^2)] + [12e19] + [\Delta v]$
4
1
nuclear mass-energy quanta per shell
[Ar] 3d6 4s2
Transition Metal
Azmithal & Magnetic numbers
2 1 0 1 2
0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0
26
5 3 1 2 4 6
d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.26 - Iron atomic config

elemental family geometry

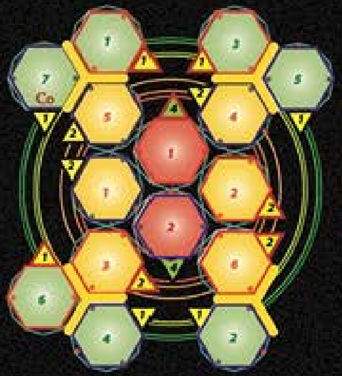
Co
1
3
7
5
5
4
1
1
2
2
3
6
6
4
2

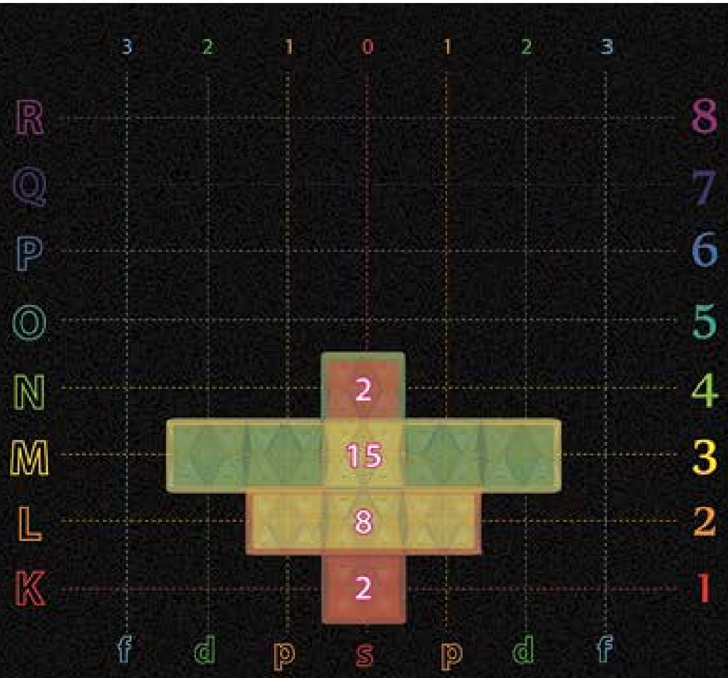
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
15
8
2
f
d
p
s
p
d
f

electron configuration

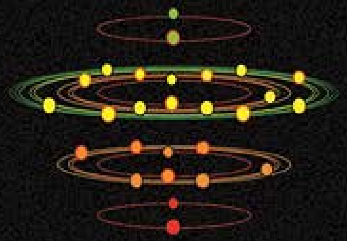

Cobalt

$$ {}_{K}^{N}{}^{P}\sum_{25}^{28} \left[\overset{\text{Baryon rest quanta}}{72(m^2)} \right] + \left[\overset{\text{lepton rest quanta}}{12e19} \right] + \left[\overset{\text{KEM}}{\Delta v} \right]_1^4 $$

nuclear mass-energy quanta per shell

[Ar] 3d7 4s2

Transition Metal

Azmithal & Magnetic numbers

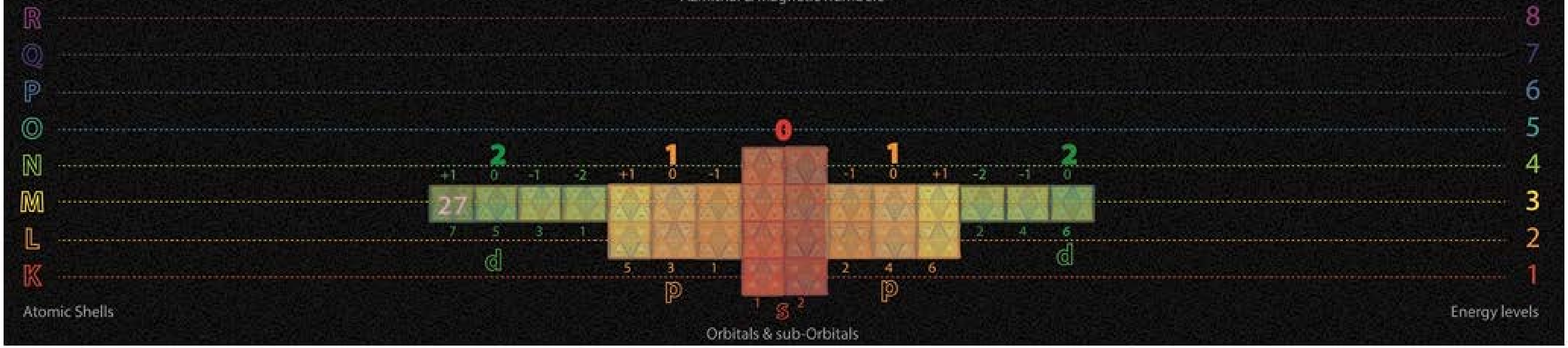
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
27
7
5
3
1
2
4
6
d
d
5
3
1
2
4
6
p
p
1
s
2

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.27 - Cobalt atomic config

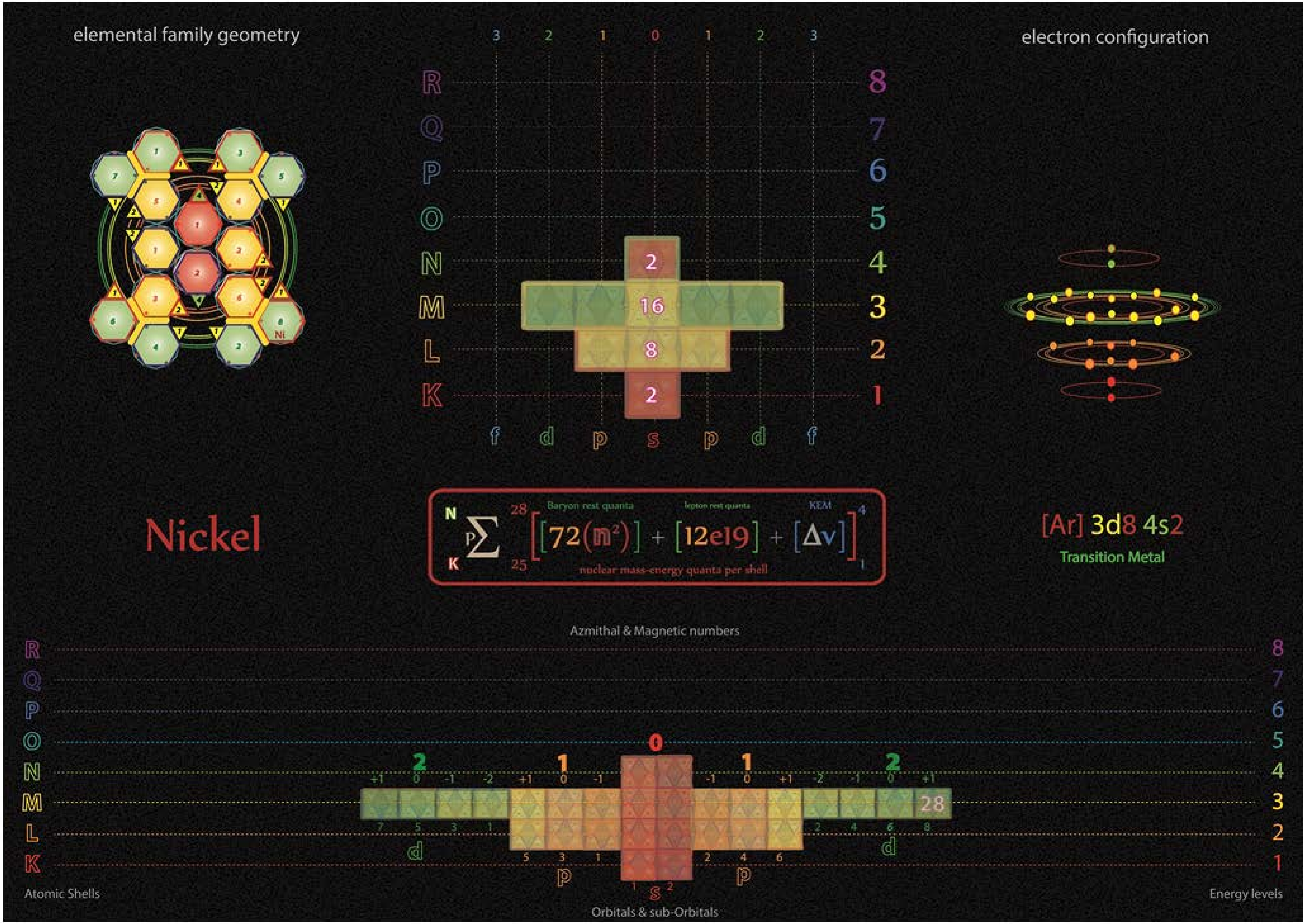
elemental family geometry
Ni
Nickel
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
16
8
2
f d p s p d f
electron configuration
[Ar] 3d8 4s2
Transition Metal
N
K
28
25
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{N} p \sum_{25}^{28} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_1^4$
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
2 1 0 1 2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
28
7 5 3 1 2 4 6 8
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.28 - Nickel atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

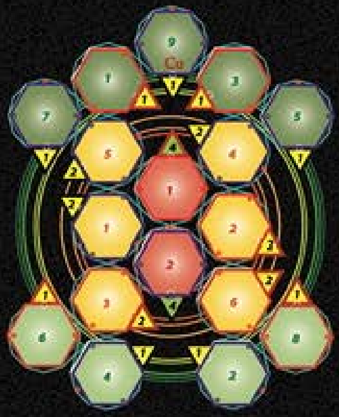

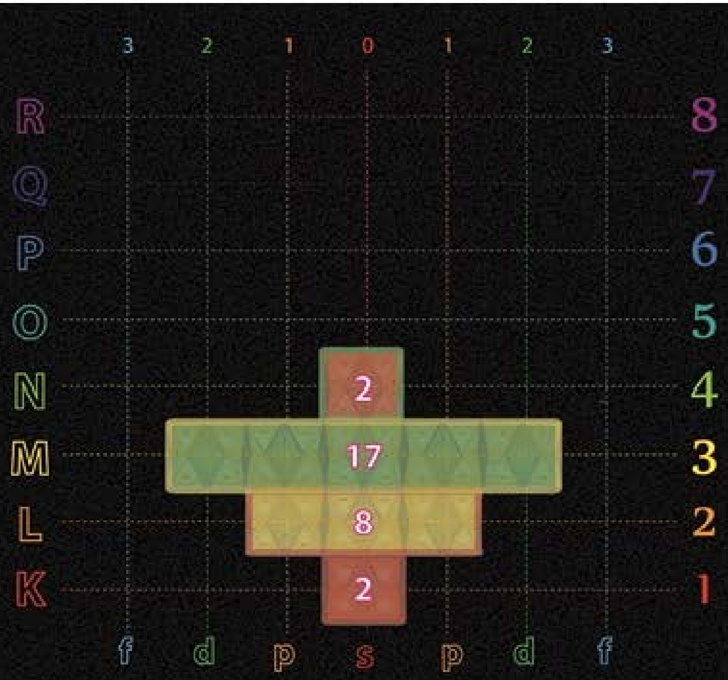
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
17
8
2
f
d
p
s
p
d
f

electron configuration

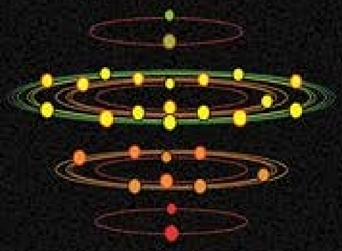

Copper

$$\sum_{K}^{N} {}^{P}\sum_{25}^{28}\left[\overset{\text{Baryon rest quanta}}{[72(m^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}\right]_{1}^{4}$$

nuclear mass-energy quanta per shell

[Ar] 3d9 4s2

Post Transition Metal

Azmithal & Magnetic numbers

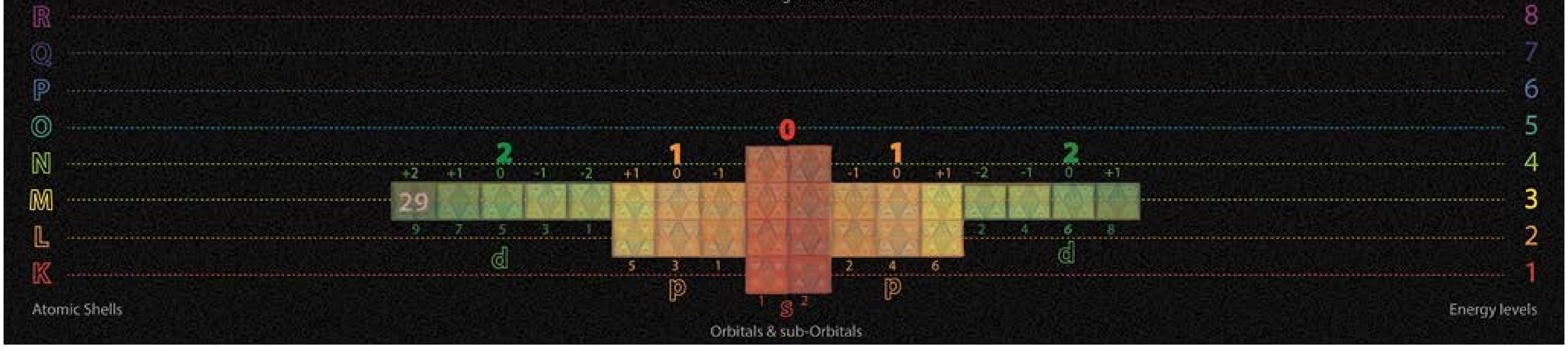
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
+2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
+1
29
9
7
5
3
1
2
4
6
8
d
d
5
3
1
2
4
6
p
p
1
s
2

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.29 - Copper atomic config

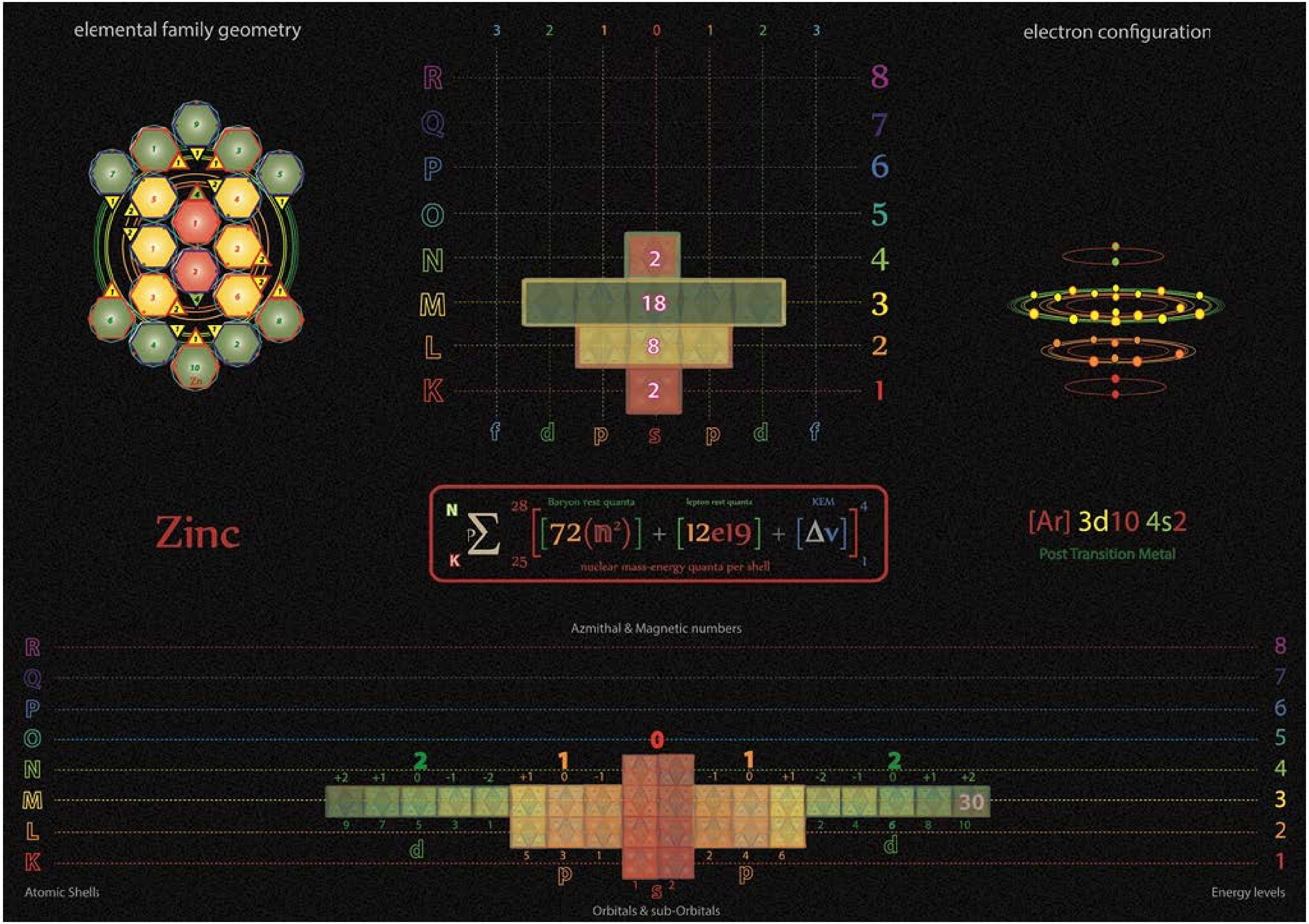
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
18
8
2
f d p s p d f
Zinc
N
K
P
28
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
4
1
nuclear mass-energy quanta per shell
[Ar] 3d10 4s2
Post Transition Metal
Azmithal & Magnetic numbers
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
30
9 7 5 3 1 2 4 6 8 10
d
5 3 1 2 4 6
p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.30 - Zinc atomic config

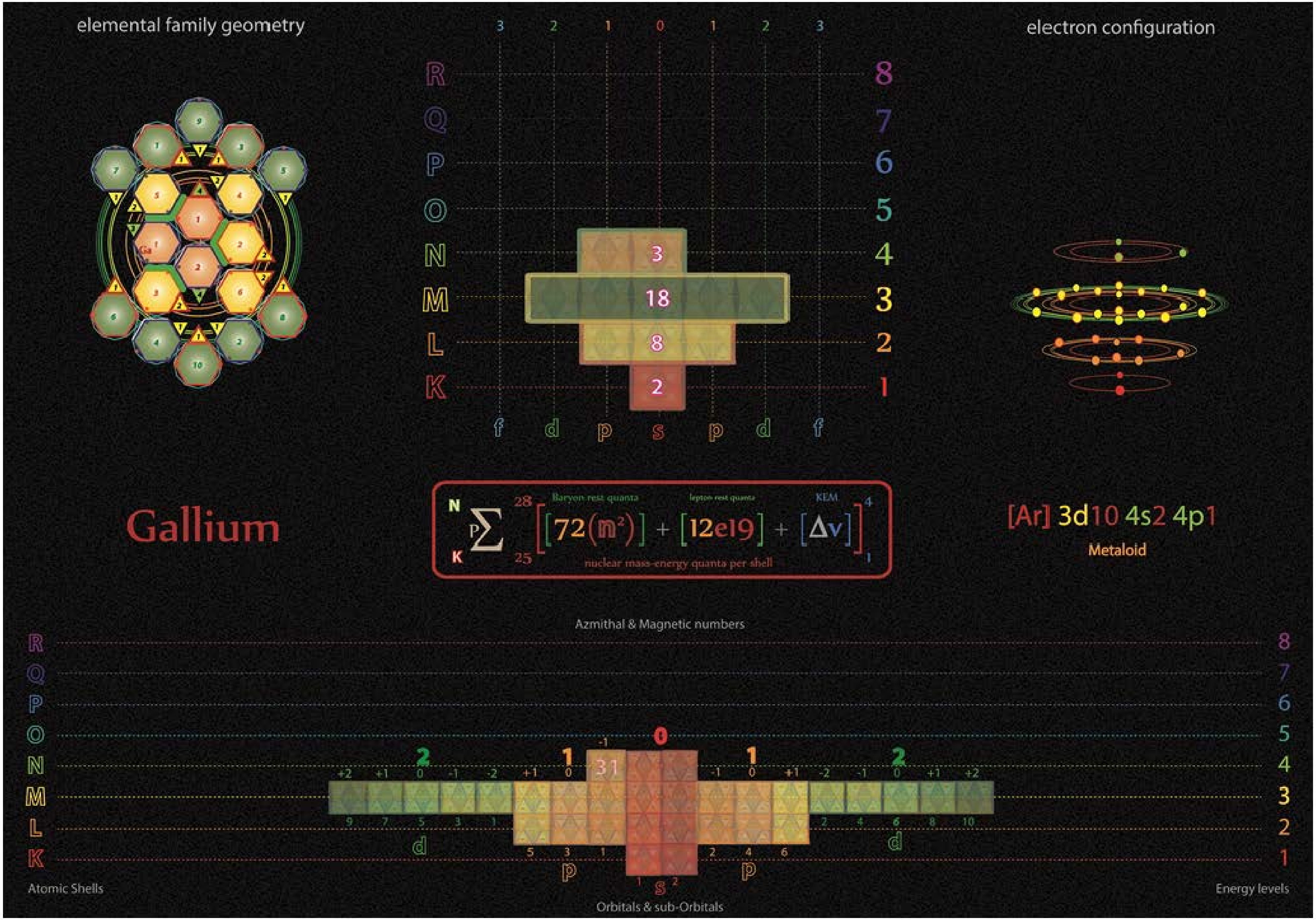
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
2 1 0 1 2
-1
+2 +1 0 -1 -2 +1 0 31 -1 0 +1 -2 -1 0 +1 +2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
d p s p d
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Gallium

Baryon rest quanta — lepton rest quanta — KEM

$$\sum_{K}^{N} P \; {}^{28}_{25}\left[\,[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\,\right]^{4}_{1}$$

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p1

Metaloid

Tetryonics 53.31 - Gallium atomic config

elemental family geometry

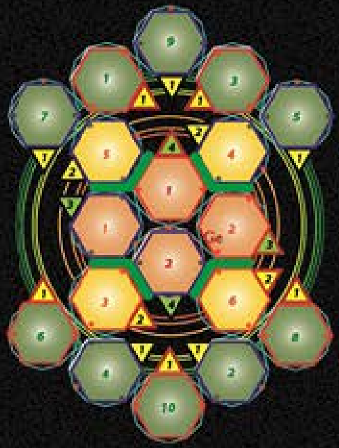
9
1
3
7
5
5
4
1
1
2
2
3
6
6
8
4
2
10

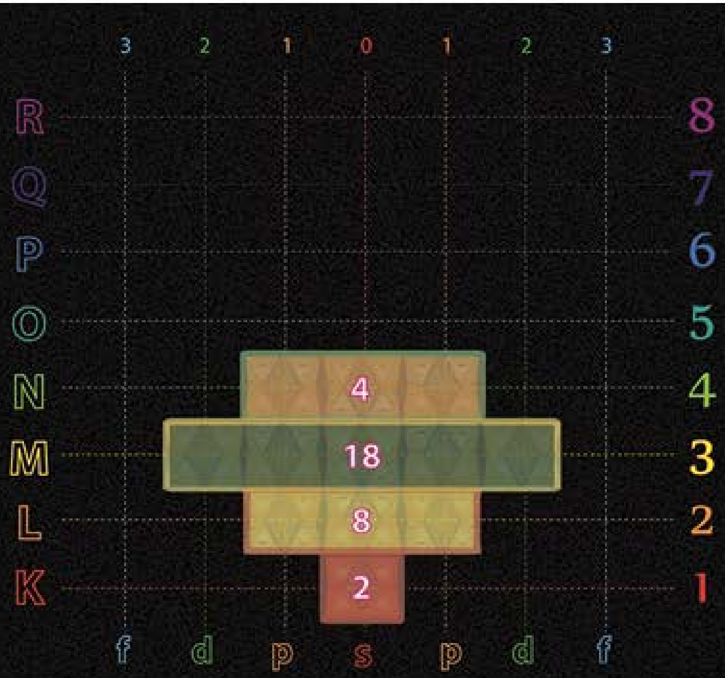
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
4
18
8
2
f
d
p
s
p
d
f

electron configuration

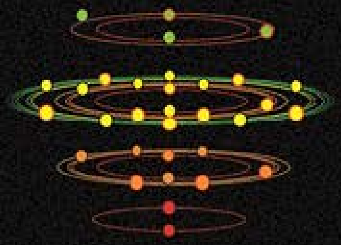

Germanium

$$\sum_{K}^{N} {}_{p}\sum_{25}^{28} \left[\overset{\text{Baryon rest quanta}}{[72(\text{m}^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]_{1}^{4}$$

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p2

Metaloid

Azmithal & Magnetic numbers

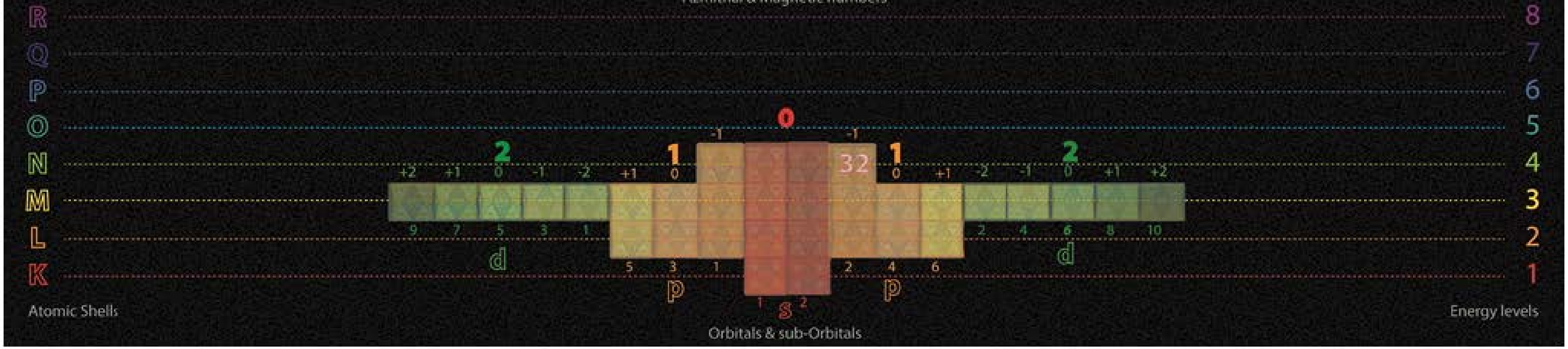
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
-1
-1
+2
+1
0
-1
-2
+1
0
32
0
+1
-2
-1
0
+1
+2
9
7
5
3
1
5
3
1
2
4
6
2
4
6
8
10
d
p
p
d
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

elemental family geometry

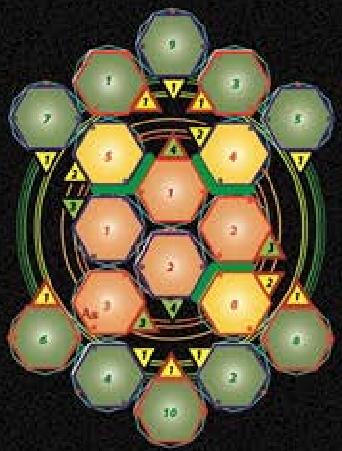

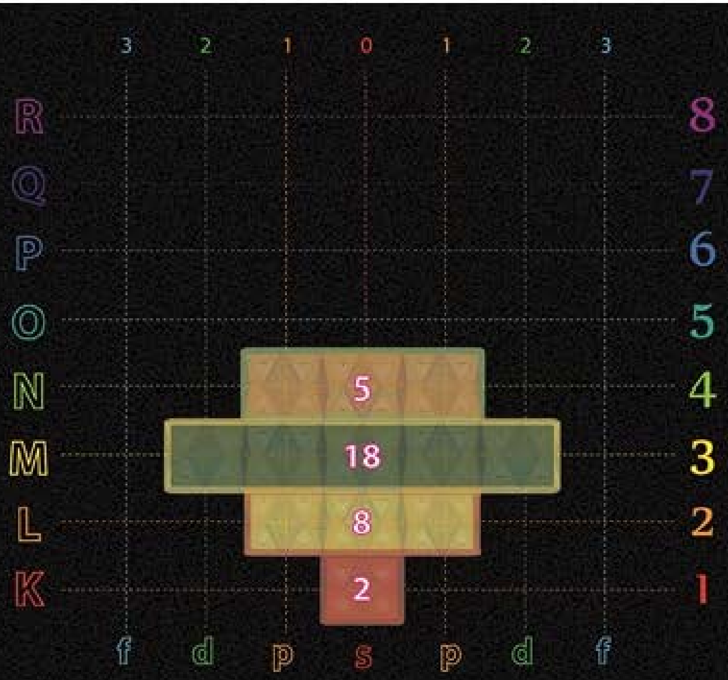
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
5
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

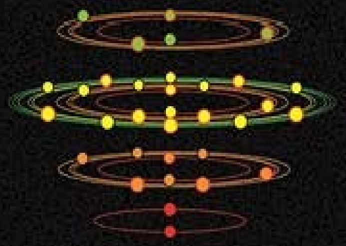

Arsenic

$$\sum_{K}^{N} P \sum_{25}^{28} \left[[72(n^2)] + [12e19] + [\Delta v] \right]_{1}^{4}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p3

Metaloid

Azmithal & Magnetic numbers

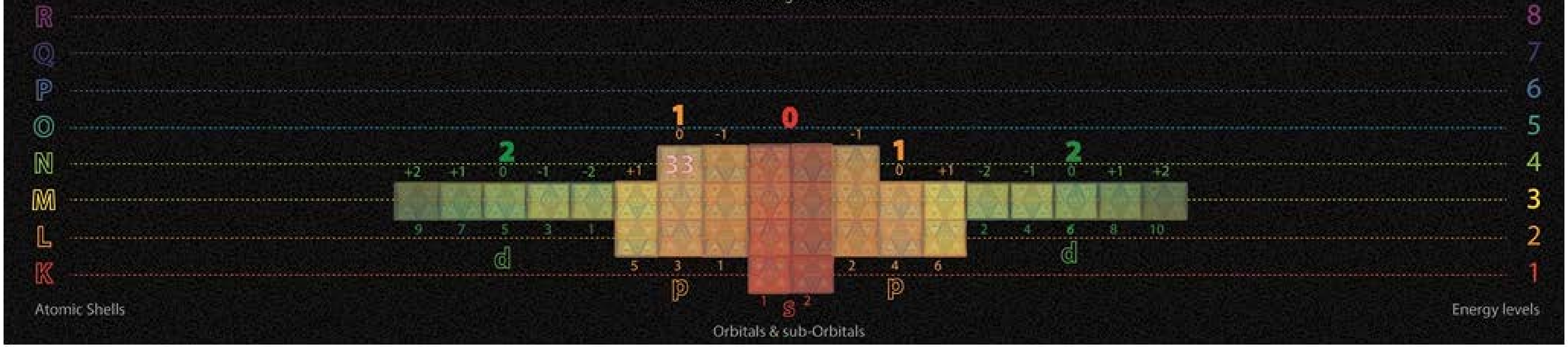
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
-1
0
-1
2
1
2
+2
+1
0
-1
-2
+1
33
0
+1
-2
-1
0
+1
+2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.33 - Arsenic atomic config

Copyright ABRAHAM [2008] - All rights reserved

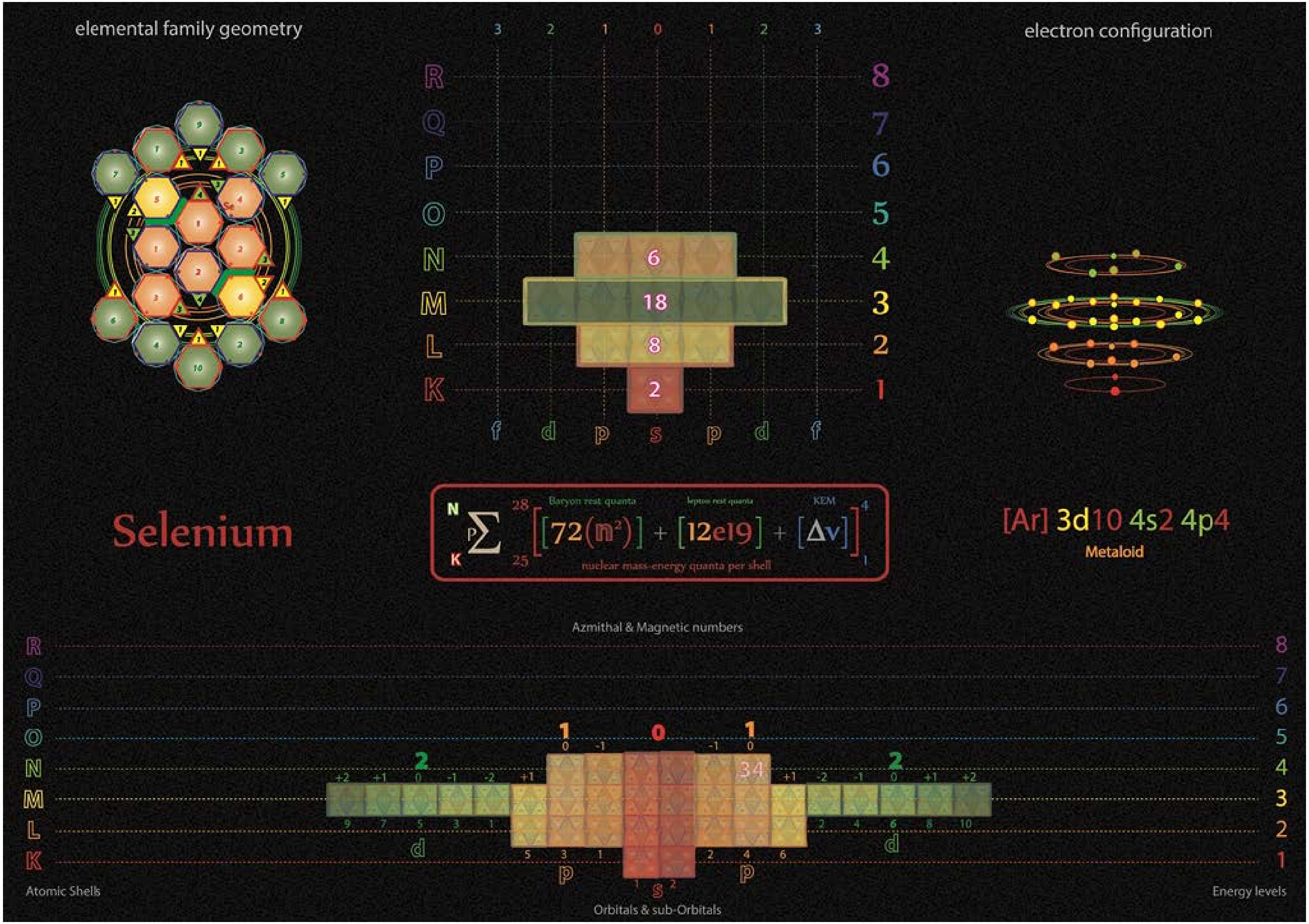
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
18
8
2
f d p s p d f
electron configuration
Selenium
N
K
P
28
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
4
1
nuclear mass-energy quanta per shell
[Ar] 3d10 4s2 4p4
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
34
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.34 - Selenium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

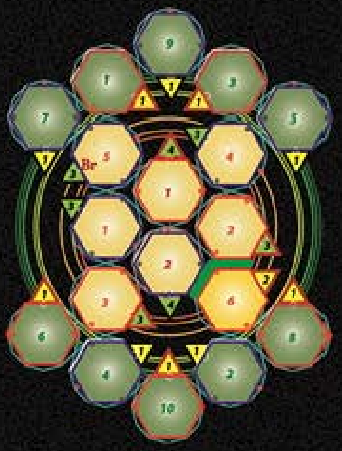

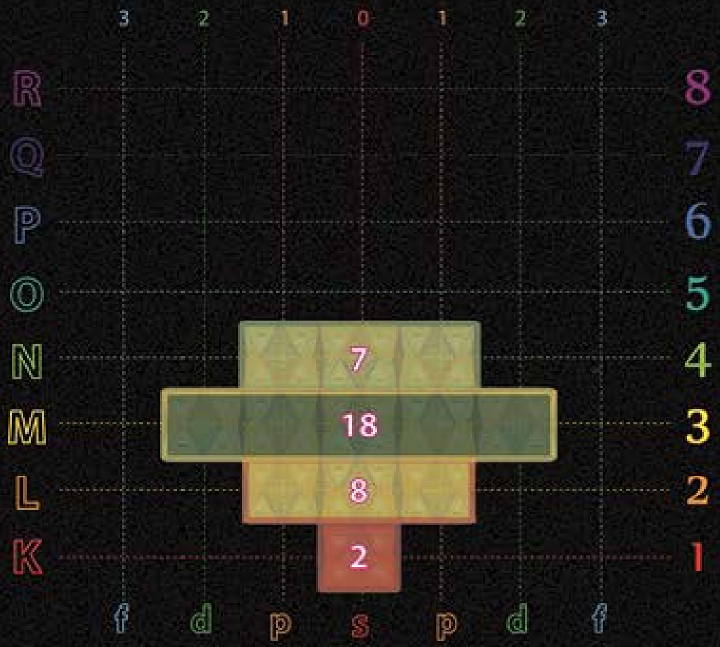
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7
18
8
2
f
d
p
s
p
d
f

electron configuration

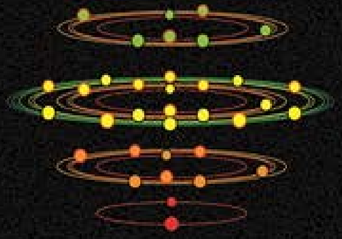

Bromine

$$\sum_{K}^{N}{}_{P}\!\sum{}^{28}_{25}\underbrace{\left[\overset{\text{Baryon rest quanta}}{[72(m^2)]}+\overset{\text{lepton rest quanta}}{[12e19]}+\overset{\text{KEM}}{[\Delta v]}\right]^{4}_{1}}_{\text{nuclear mass-energy quanta per shell}}$$

[Ar] 3d10 4s2 4p5

Halogen

Azmithal & Magnetic numbers

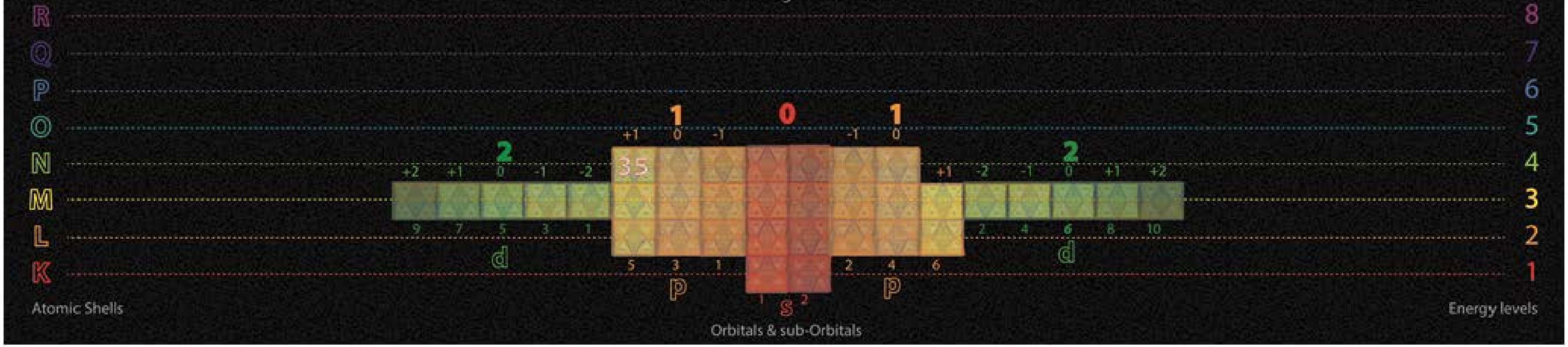
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
+2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
+1
+2
35
9
7
5
3
1
2
4
6
8
10
d
5
3
1
2
4
6
p
s
1
2
p
d
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.35 - Bromine atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

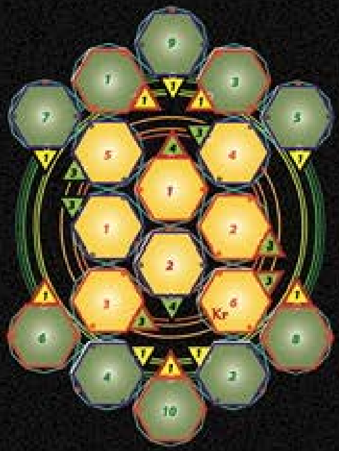

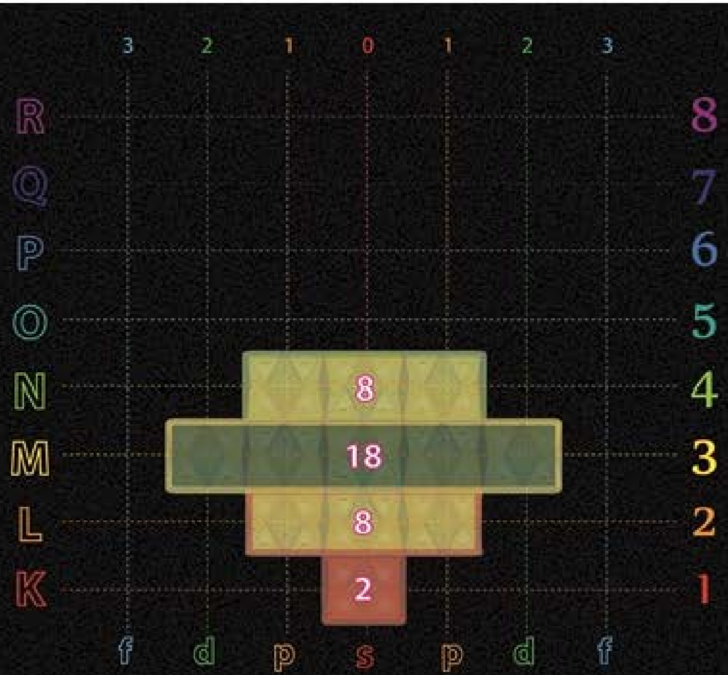
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
8
18
8
2
f
d
p
s
p
d
f

electron configuration

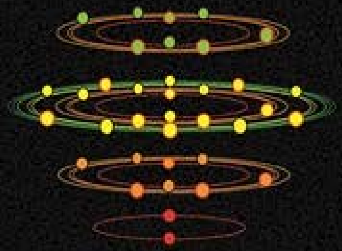

Krypton

Baryon rest quanta · lepton rest quanta · KEM

$$\sideset{^{N}_{K}}{}{}{}_{P}\sum_{25}^{28}\left[\,[72(\mathfrak{m}^{2})] + [12e19] + [\Delta v]\,\right]_{1}^{4}$$

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p6

Noble
Gas

Azmithal & Magnetic numbers

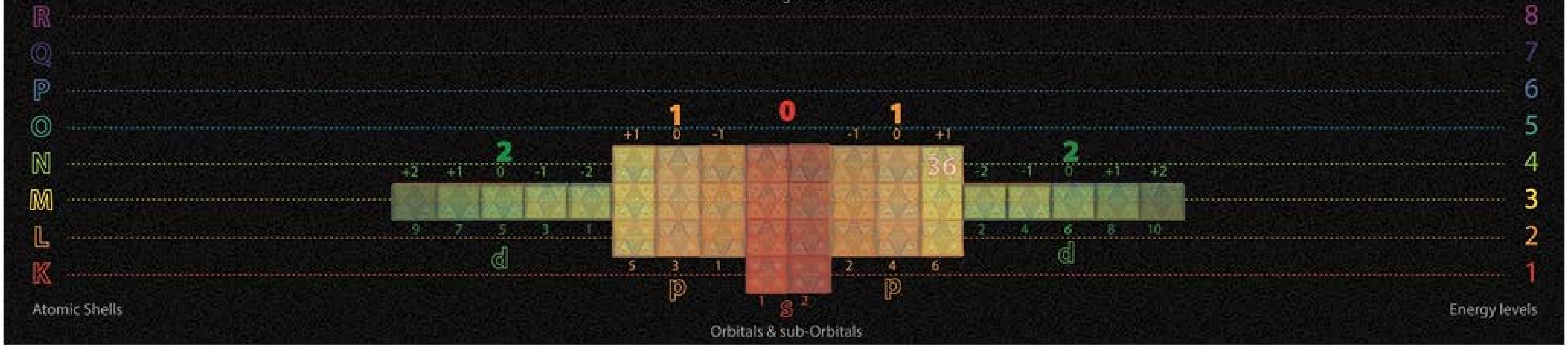
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
36
9 7 5 3 1
5 3 1
2 4 6
2 4 6 8 10
d
p
s
p
d
1
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.36 - Krypton atomic config

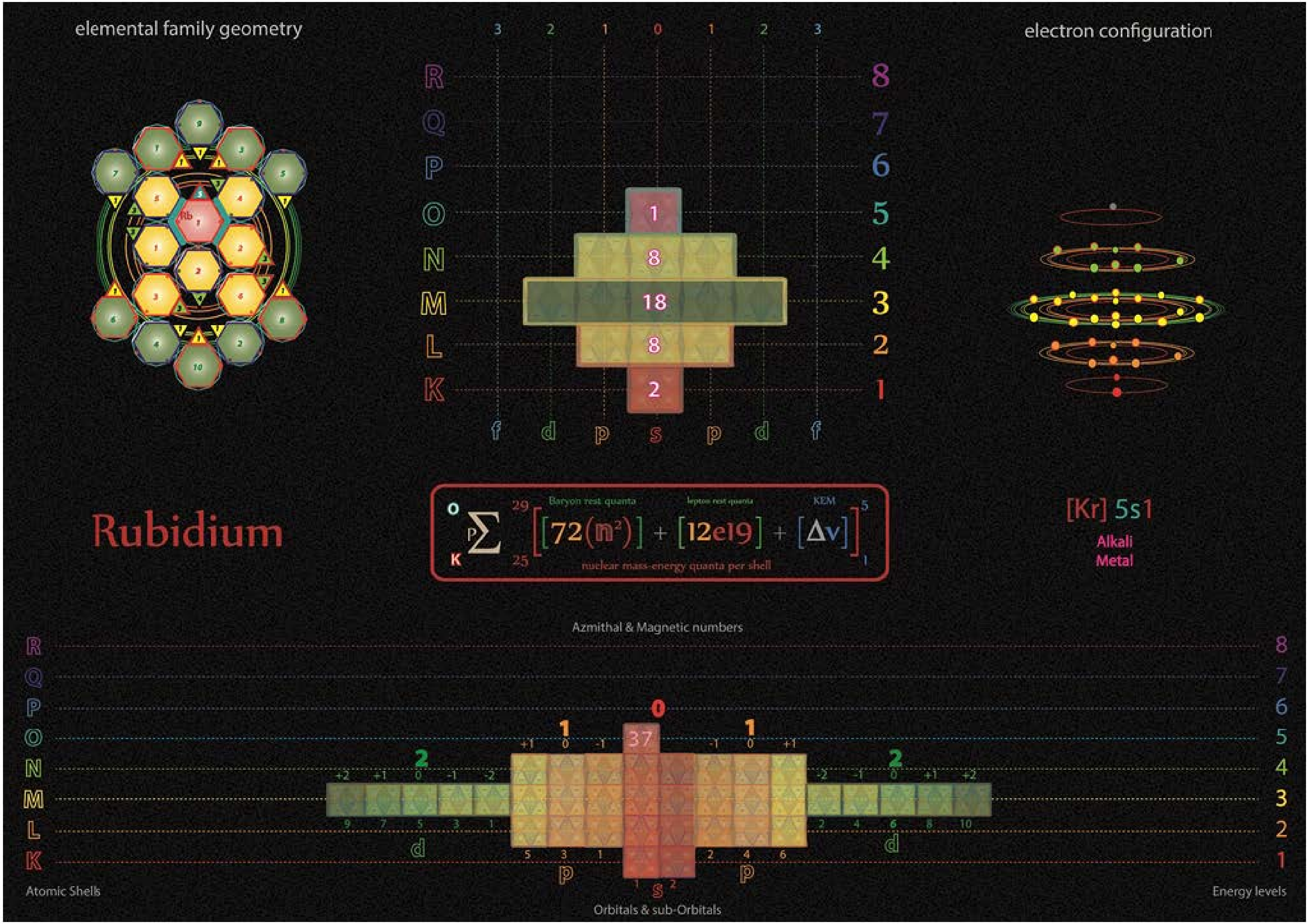
elemental family geometry
Rb
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
1
8
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
0
1 1
2 2
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -2 -1 0 +1 +2
37
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d p s p d
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Rubidium

$$\sum_{K}^{O}{}_{P}\ {}^{29}_{25}\Big[[72(m^2)] + [12e19] + [\Delta v]\Big]^{5}_{1}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Kr] 5s1

Alkali
Metal

Tetryonics 53.37 - Rubidium atomic config

Copyright ABRAHAM [2008] - All rights reserved

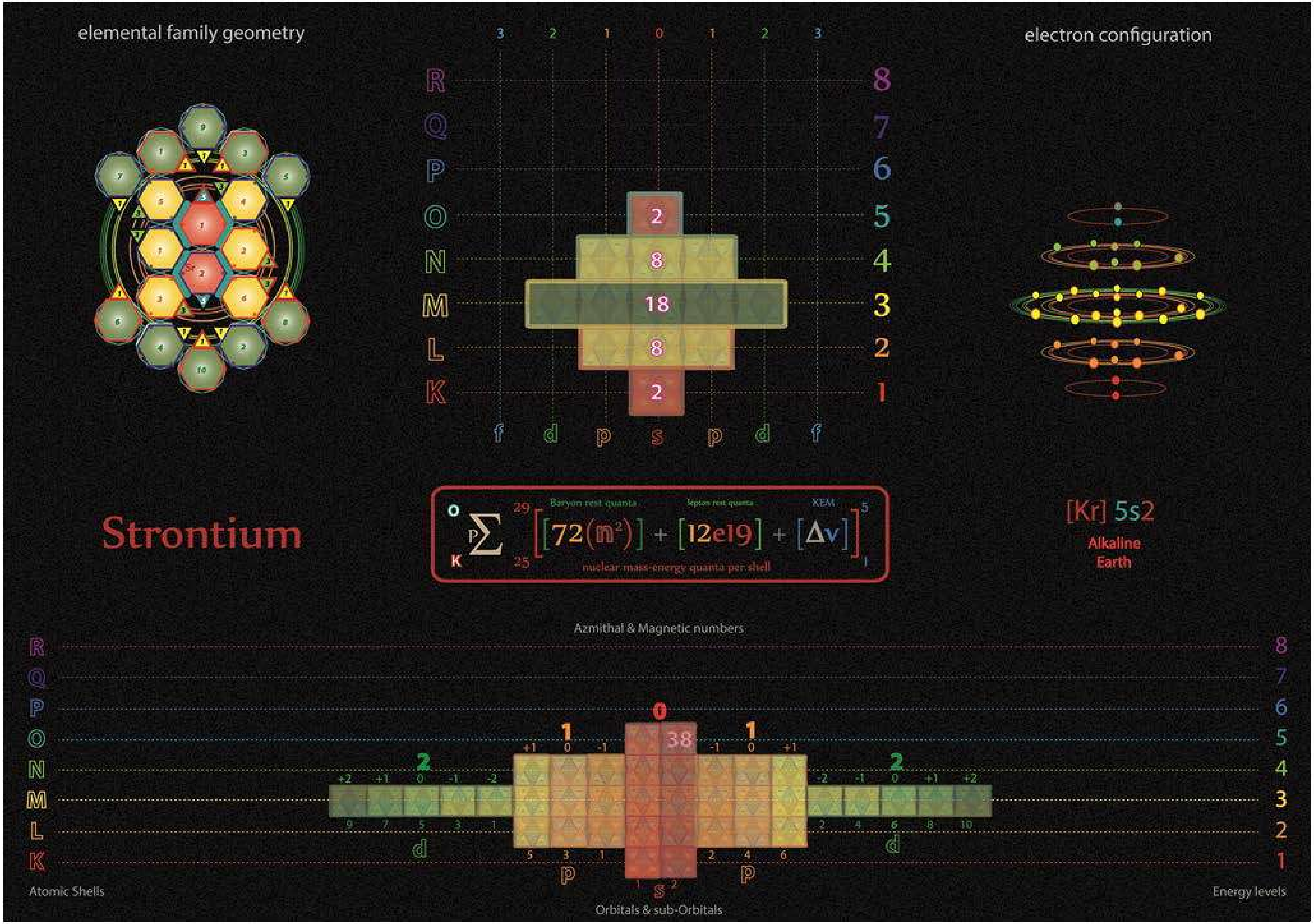
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
18
8
2
f d p s p d f
Strontium
[Kr] 5s2
Alkaline Earth
Baryon rest quanta
lepton rest quanta
XEM
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
0
38
1
1
2
2
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
p
s
p
d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{O}{}_{P}\,{}^{29}_{25}\left[\,[72(n^{2})] + [12e19] + [\Delta v]\,\right]^{5}_{1}$$

Tetryonics 53.38 - Strontium atomic config

Copyright ABRAHAM [2008] - All rights reserved

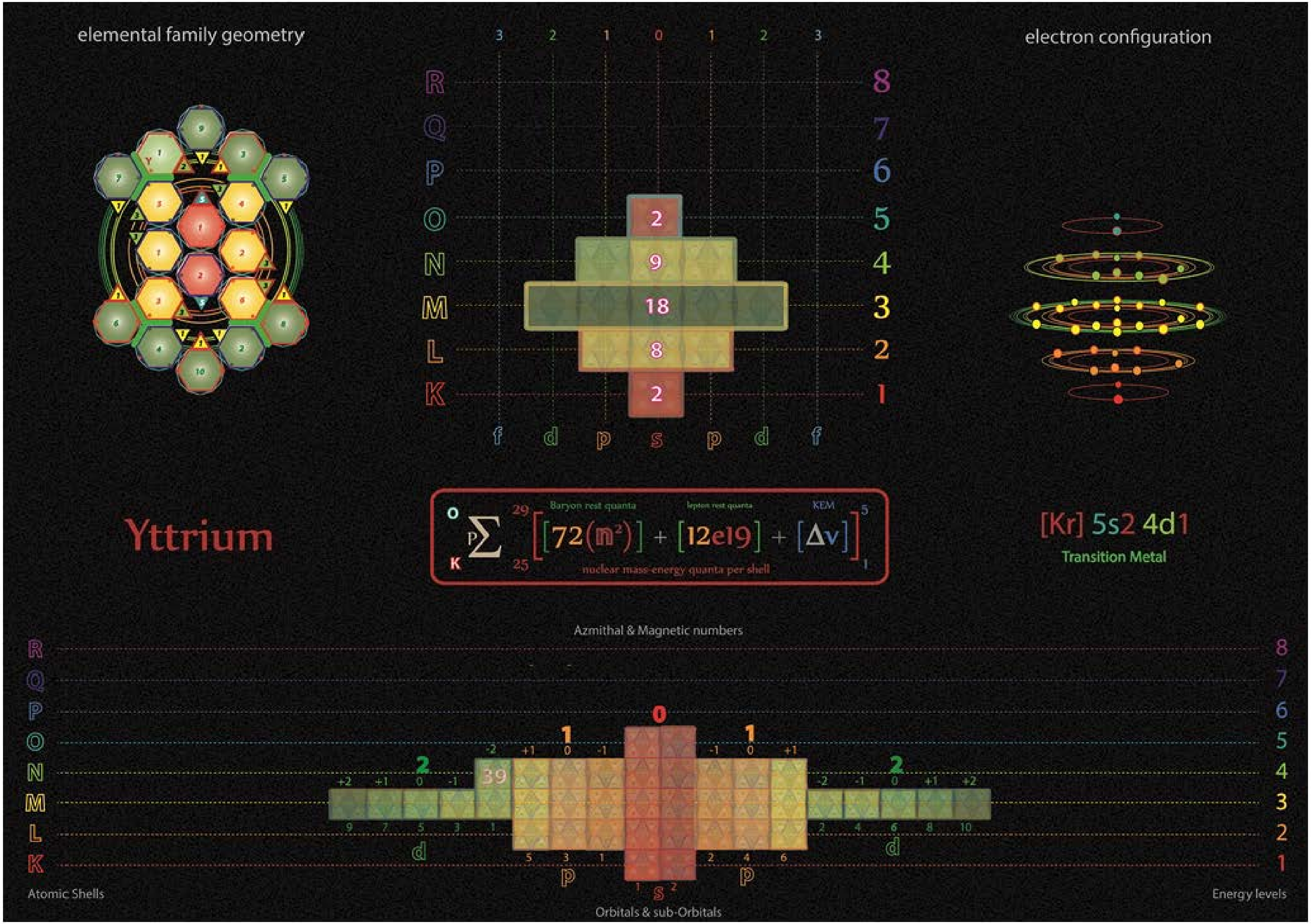
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
9
18
8
2
f d p s p d f
Yttrium
O
K
P∑
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
5
1
nuclear mass-energy quanta per shell
[Kr] 5s2 4d1
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
-2 +1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -1 0 +1 +2
39
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.39 - Yttrium atomic config

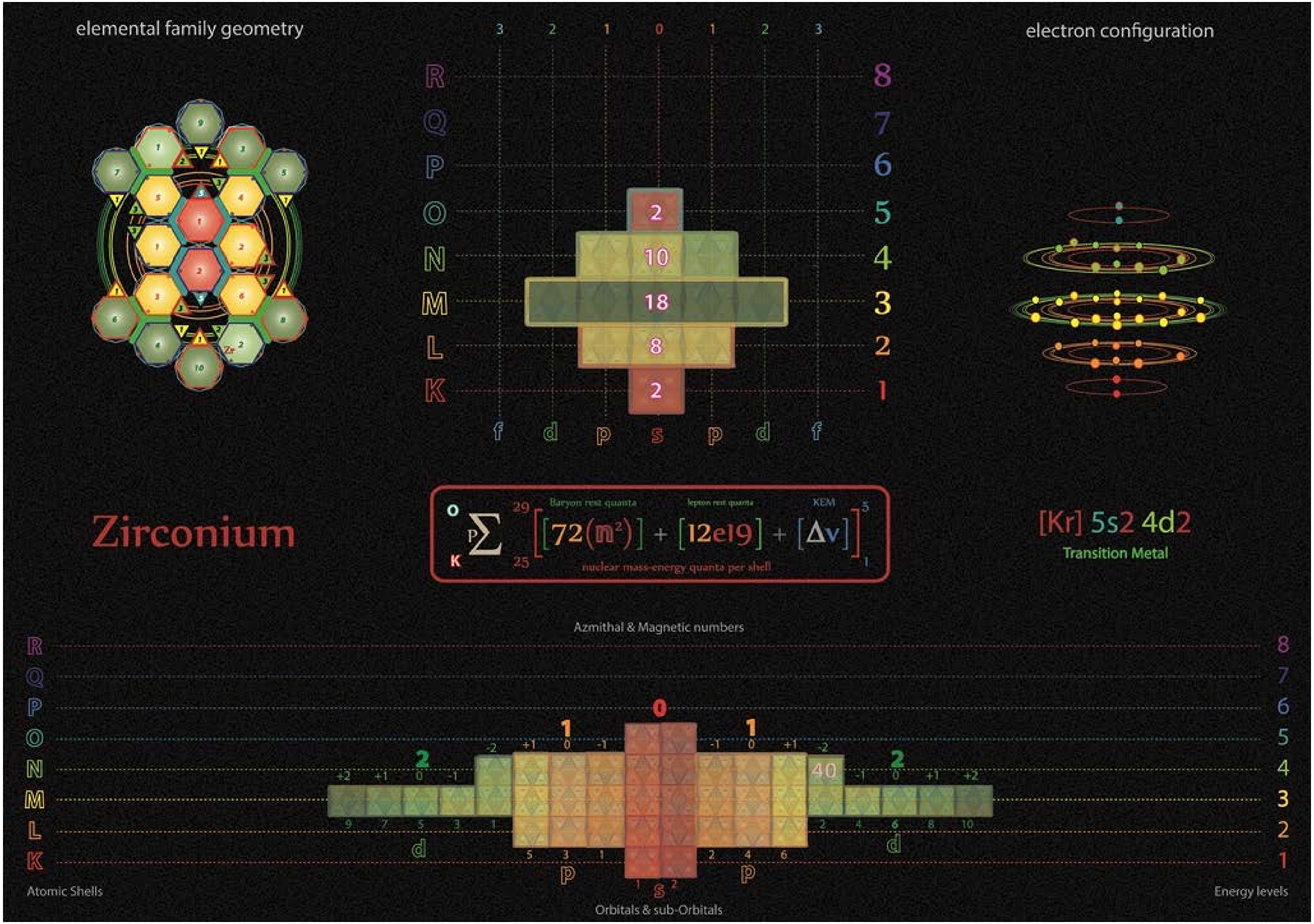
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
10
18
8
2
f d p s p d f
electron configuration
Zirconium
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
5
1
nuclear mass-energy quanta per shell
[Kr] 5s2 4d2
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
-2 +1 0 -1
-1 0 +1 -2
+2 +1 0 -1
-1 0 +1 +2
40
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 s 2
d
d
p
p
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.40 - Zirconium atomic config

Copyright ABRAHAM [2008] - All rights reserved

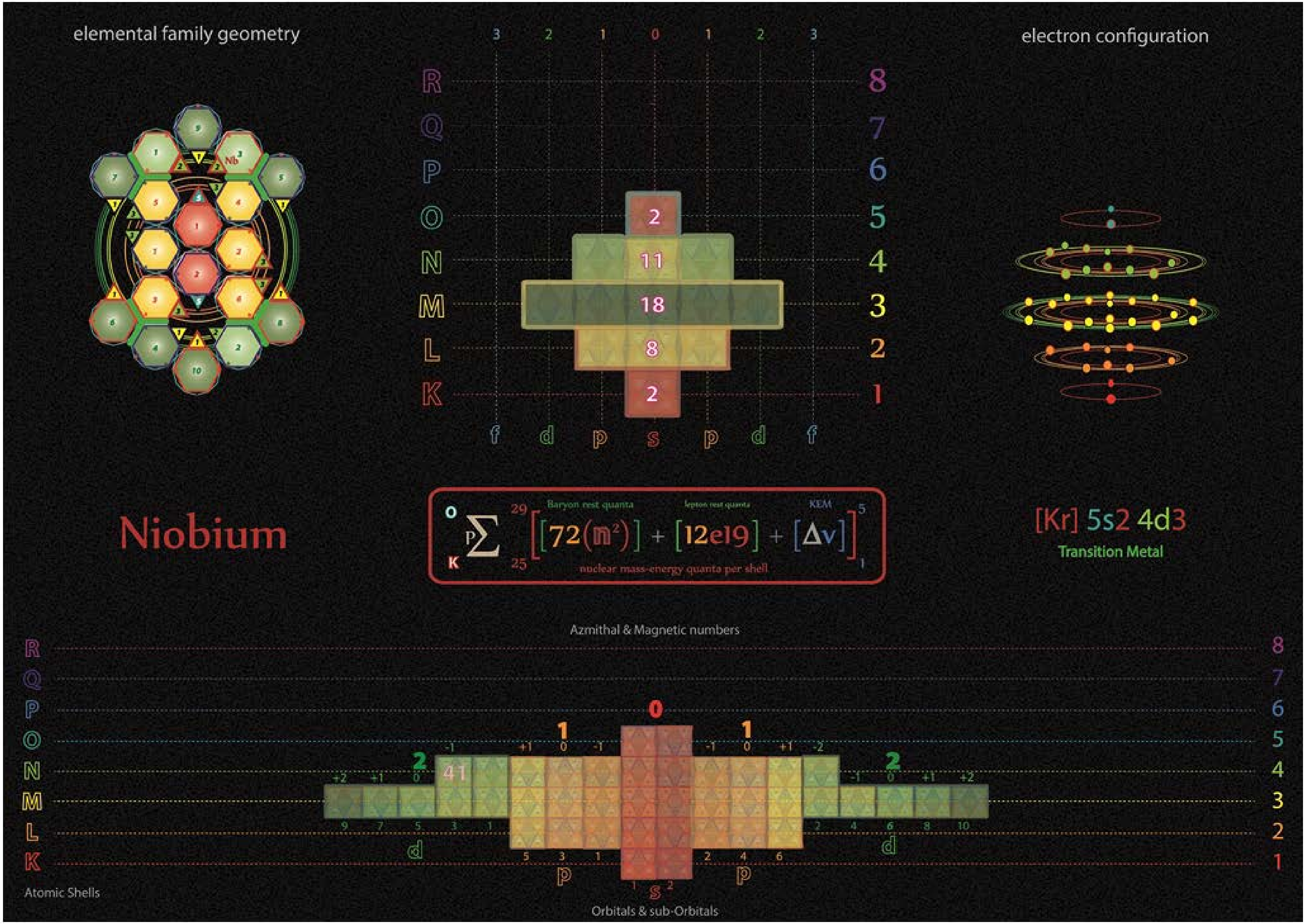

elemental family geometry

electron configuration

Niobium

Baryon rest quanta — lepton rest quanta — KEM

$$\sum_{K}^{O} {}^{p}\sum_{25}^{29}\left[[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\right]_{1}^{5}$$

nuclear mass-energy quanta per shell

[Kr] 5s2 4d3

Transition Metal

Azmithal & Magnetic numbers

Atomic Shells

Orbitals & sub-Orbitals

Nb
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
11
18
8
2
f d p s p d f
0
1 1
2 2
41
-1 +1 0 -1 -1 0 +1 -2
+2 +1 0 -1 0 +1 +2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
d p s 2 p d
1

Tetryonics 53.41 - Niobium atomic config

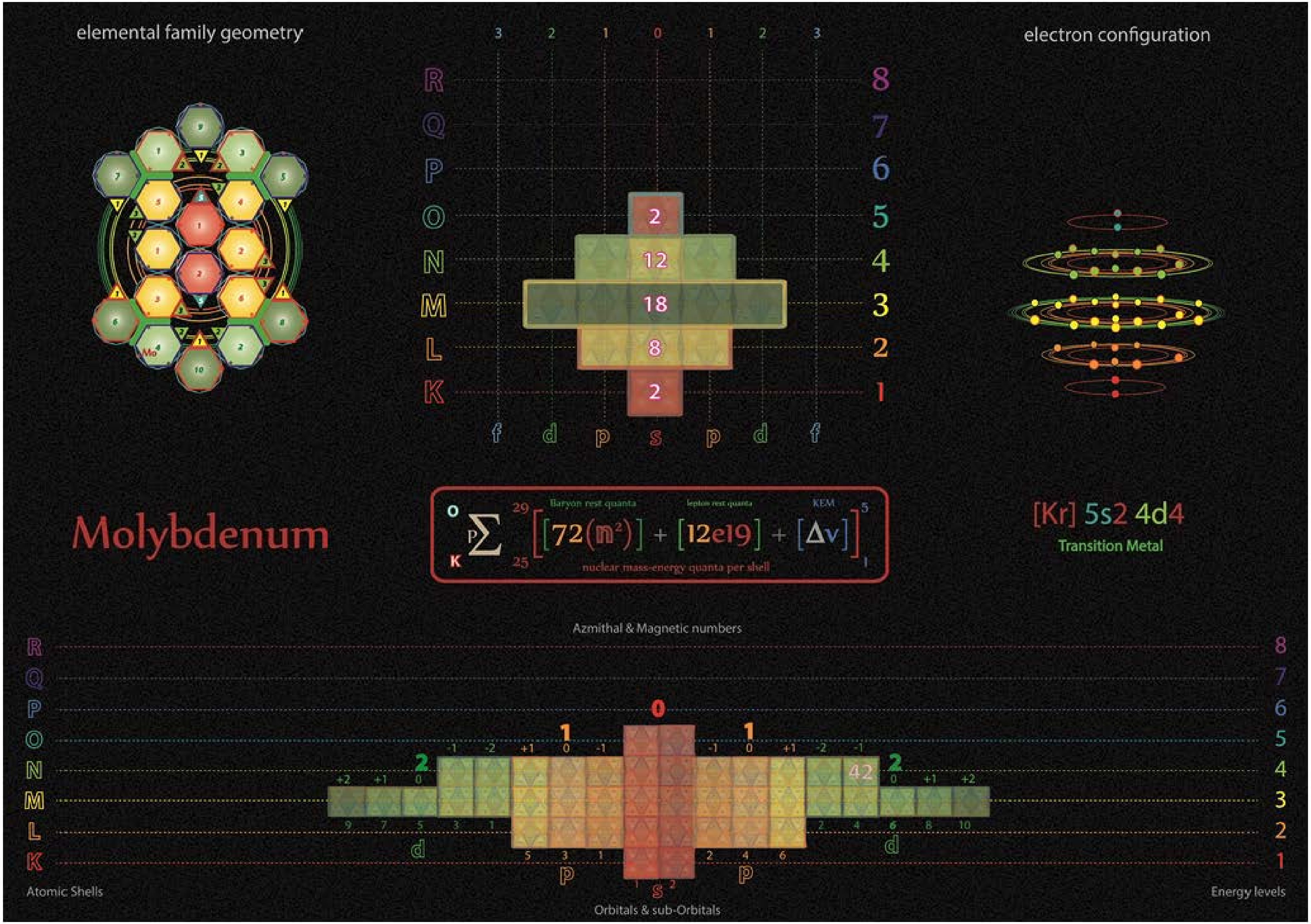
elemental family geometry
Mo
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
12
18
8
2
f d p s p d f
electron configuration
Molybdenum
O
K
P
Σ
29
25
Baryon rest quanta
lepton rest quanta
KEM
[[72(m²)] + [12e19] + [Δv]]
5
1
nuclear mass-energy quanta per shell
[Kr] 5s2 4d4
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
-1 -2 +1 0 -1 -1 0 +1 -2 -1
+2 +1 0 0 +1 +2
42
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
d d
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.42 - Molybdenum atomic config

Copyright ABRAHAM [2008] - All rights reserved

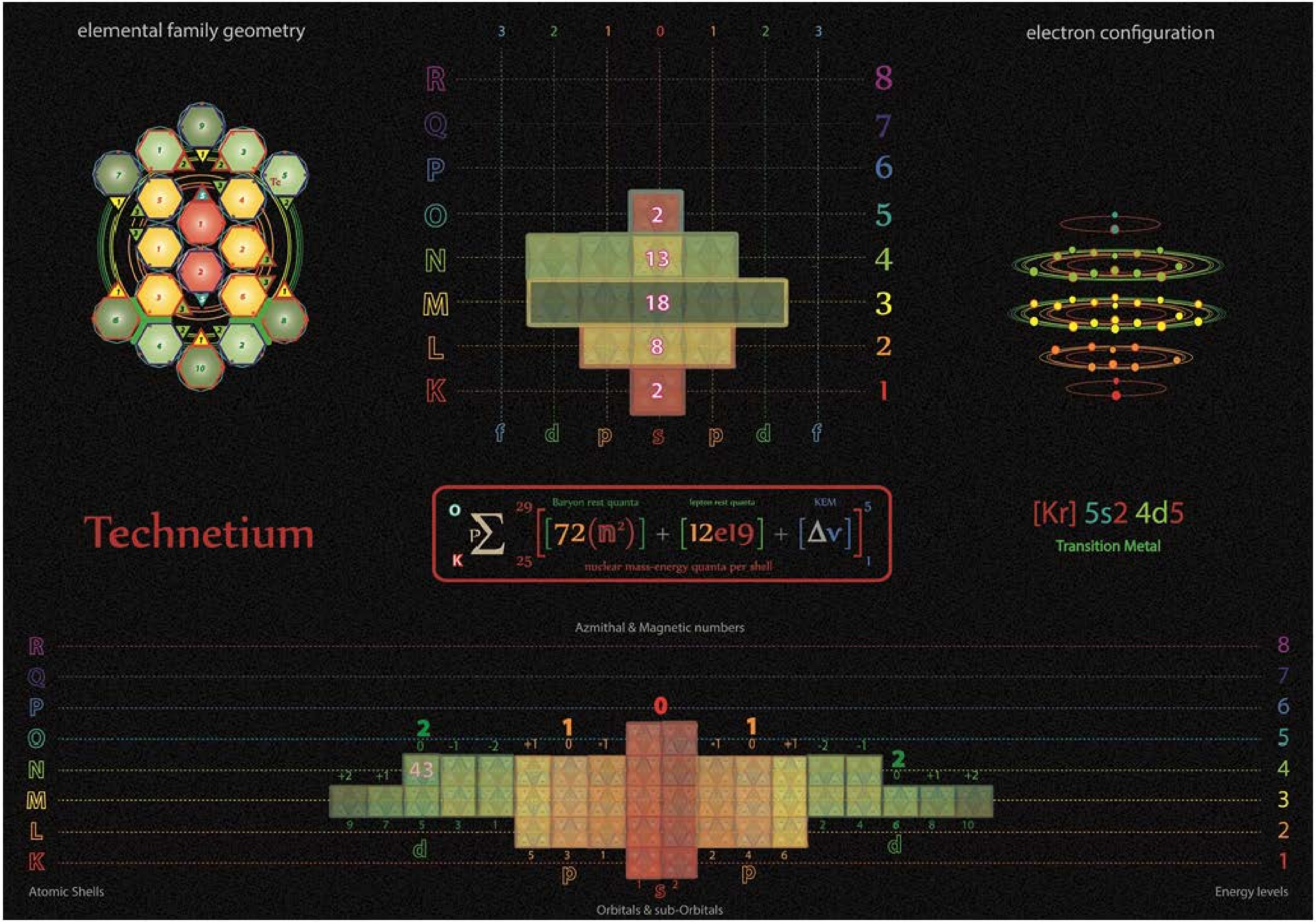
elemental family geometry
Tc
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
13
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1
1
2
2
-2 -1 0 +1 +2
43
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
d
p
s
p
d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Technetium

$$\sum_{K}^{O} p \; \sum_{25}^{29} \Big[[72(m^2)] + [12e19] + [\Delta v] \Big]_{1}^{5}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Kr] 5s2 4d5

Transition Metal

Tetryonics 53.43 - Technetium atomic config

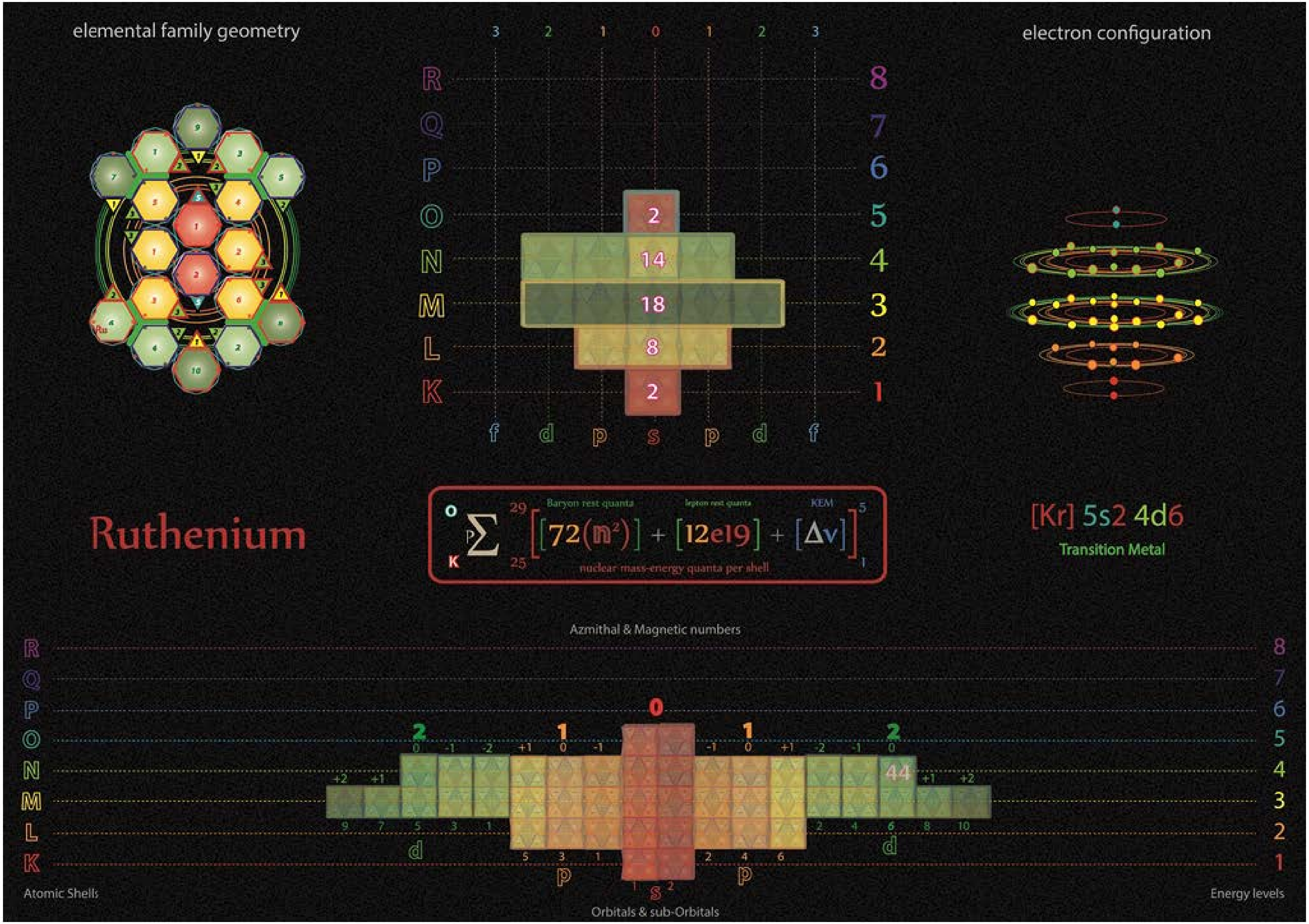
elemental family geometry
3 2 1 0 1 2 3
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
2
14
18
8
2
f d p s p d f
Ruthenium
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
5
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Kr] 5s2 4d6
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0
+2 +1 +1 +2
44
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.44 - Ruthenium atomic config

Copyright ABRAHAM [2008] - All rights reserved

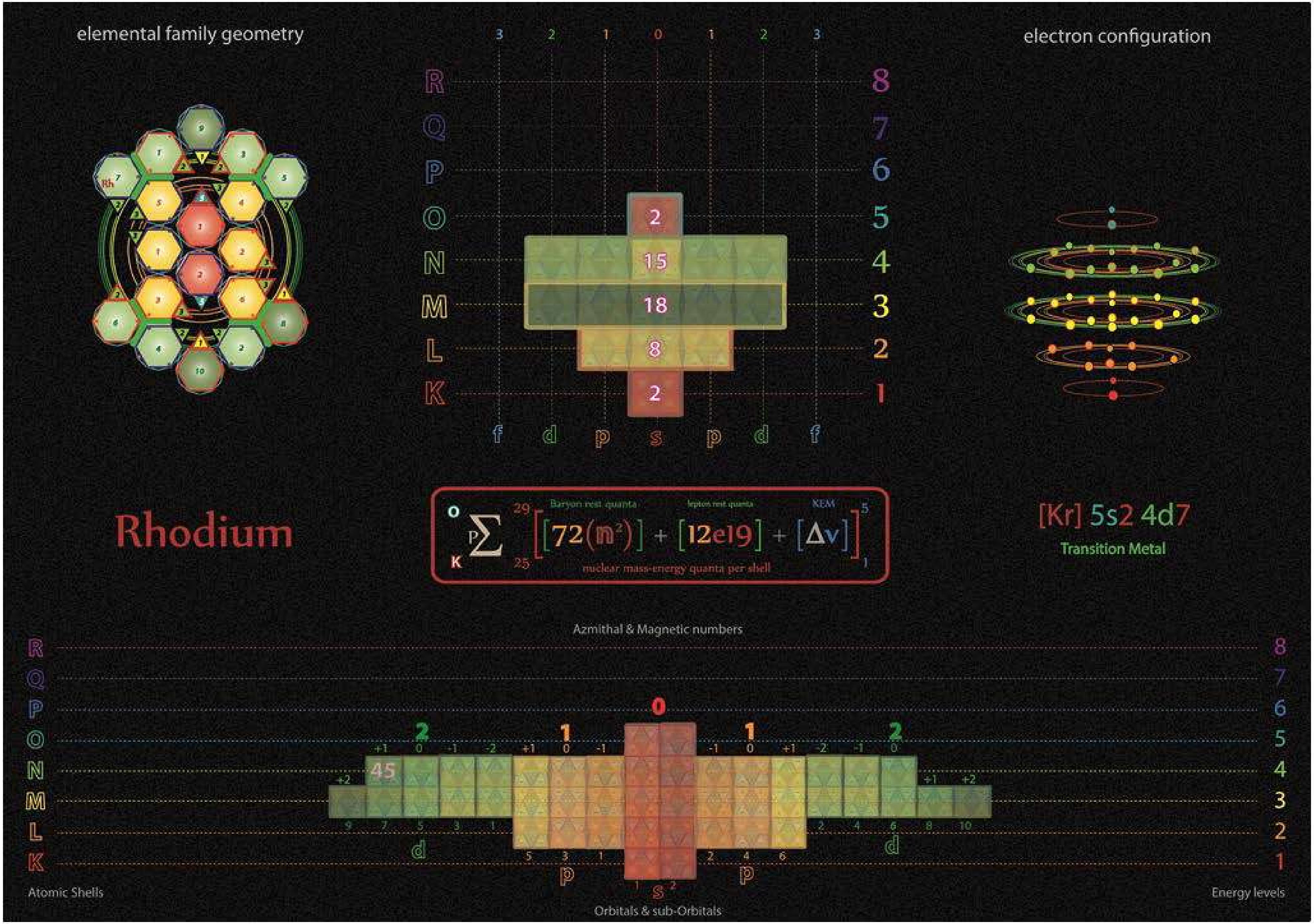
elemental family geometry
electron configuration
Rh
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
15
18
8
2
f d p s p d f
Rhodium
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
5
1
nuclear mass-energy quanta per shell
[Kr] 5s2 4d7
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0
+2 +1 +2
45
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{O} {}_{P}\left[\,[72(m^2)] + [12e19] + [\Delta v]\,\right]_{1}^{5}$$

Tetryonics 53.45 - Rhodium atomic config

Copyright ABRAHAM [2008] - All rights reserved

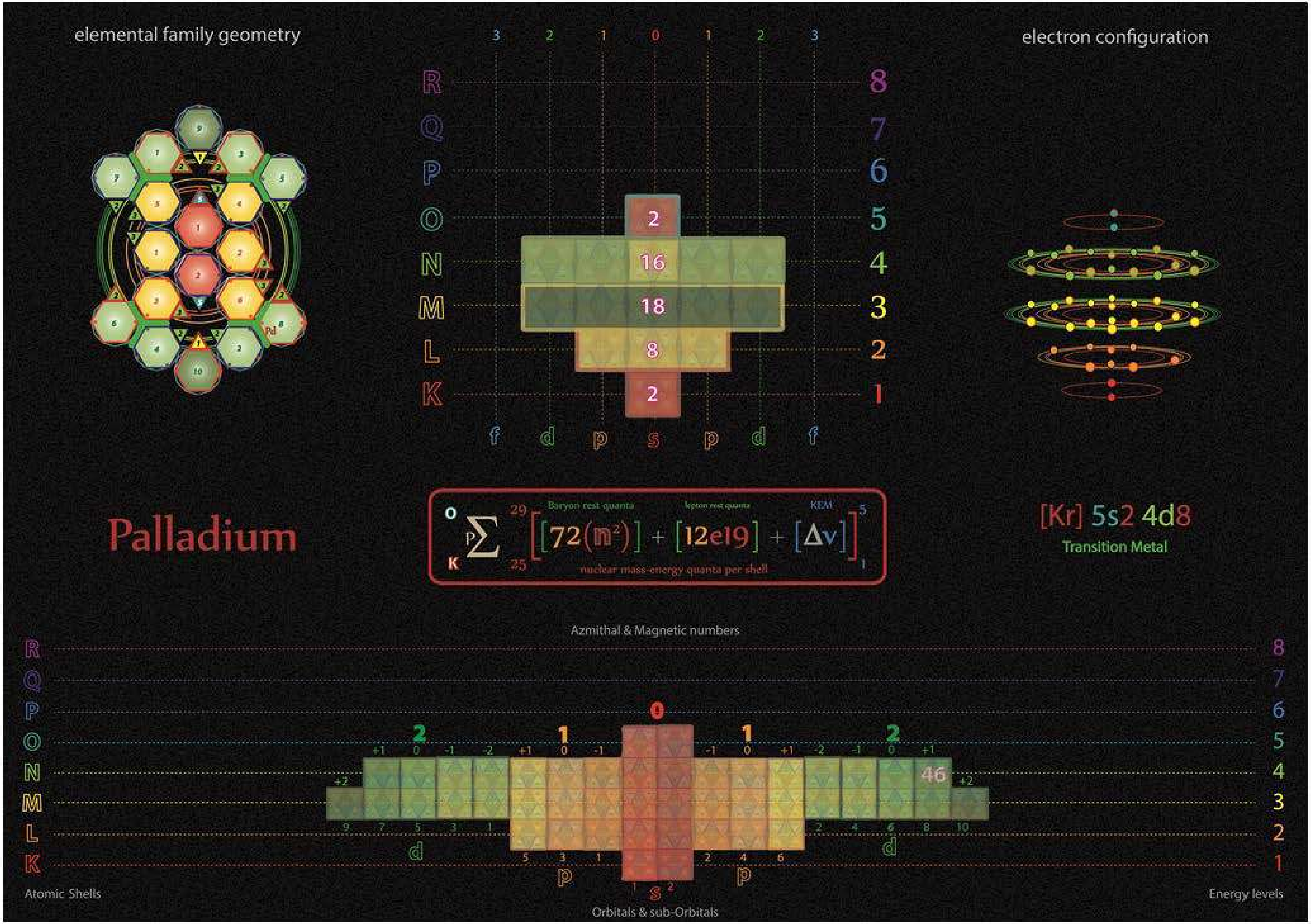
elemental family geometry
Pd
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
16
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
2 1 0 1 2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
+2 46 +2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Palladium

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{O} {}_{P}\sum_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_{1}^{5}$$

nuclear mass-energy quanta per shell

[Kr] 5s2 4d8

Transition Metal

Tetryonics 53.46 - Palladium atomic config

Copyright ABRAHAM [2008] - All rights reserved

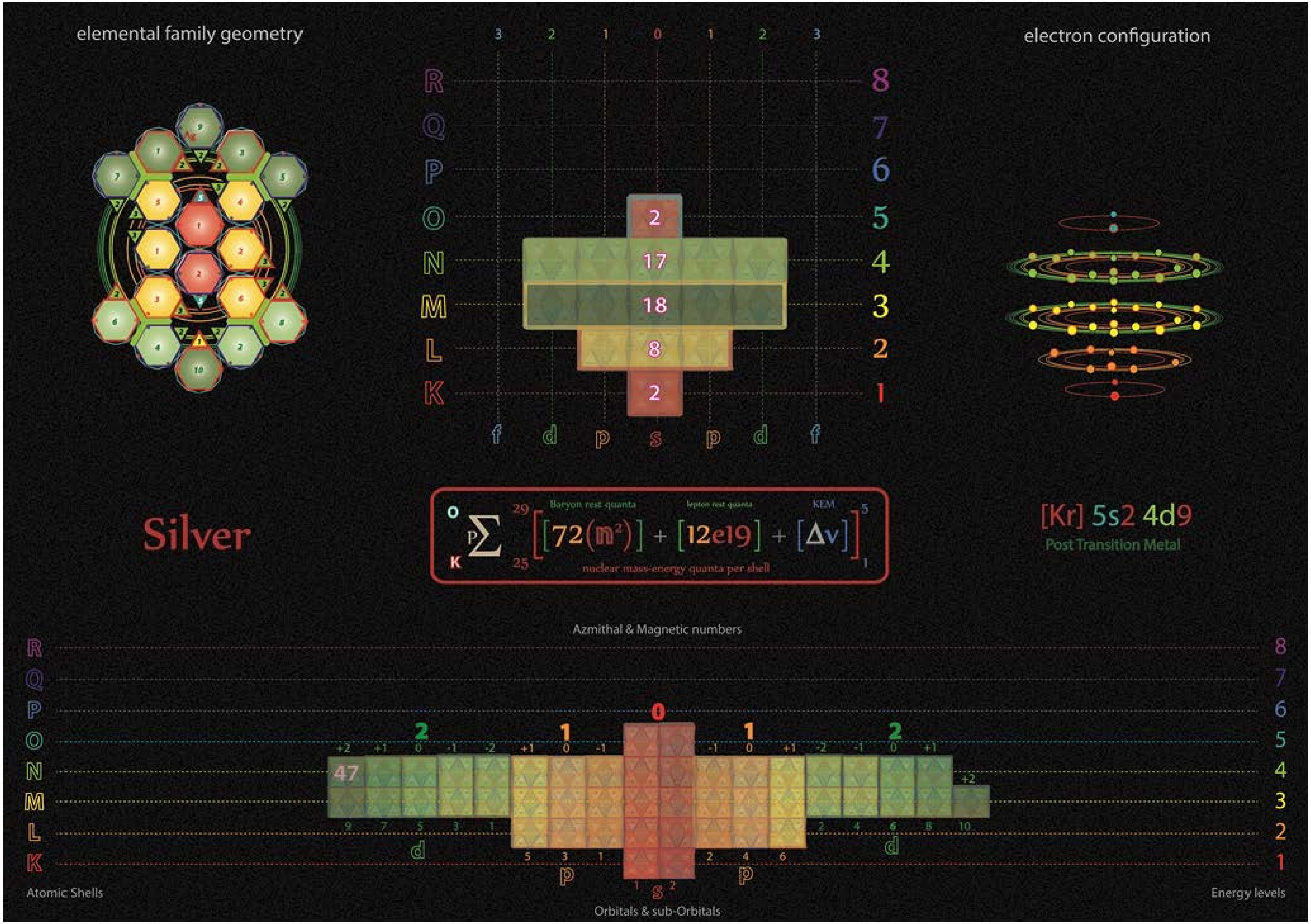
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
17
18
8
2
f d p s p d f
electron configuration
Silver
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[[72(n²)] + [12e19] + [Δv]]
nuclear mass-energy quanta per shell
5
1
[Kr] 5s2 4d9
Post Transition Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
+2
47
9 7 5 3 1
5 3 1
2 4 6
2 4 6 8 10
1 2
d p s p d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.47 - Silver atomic config

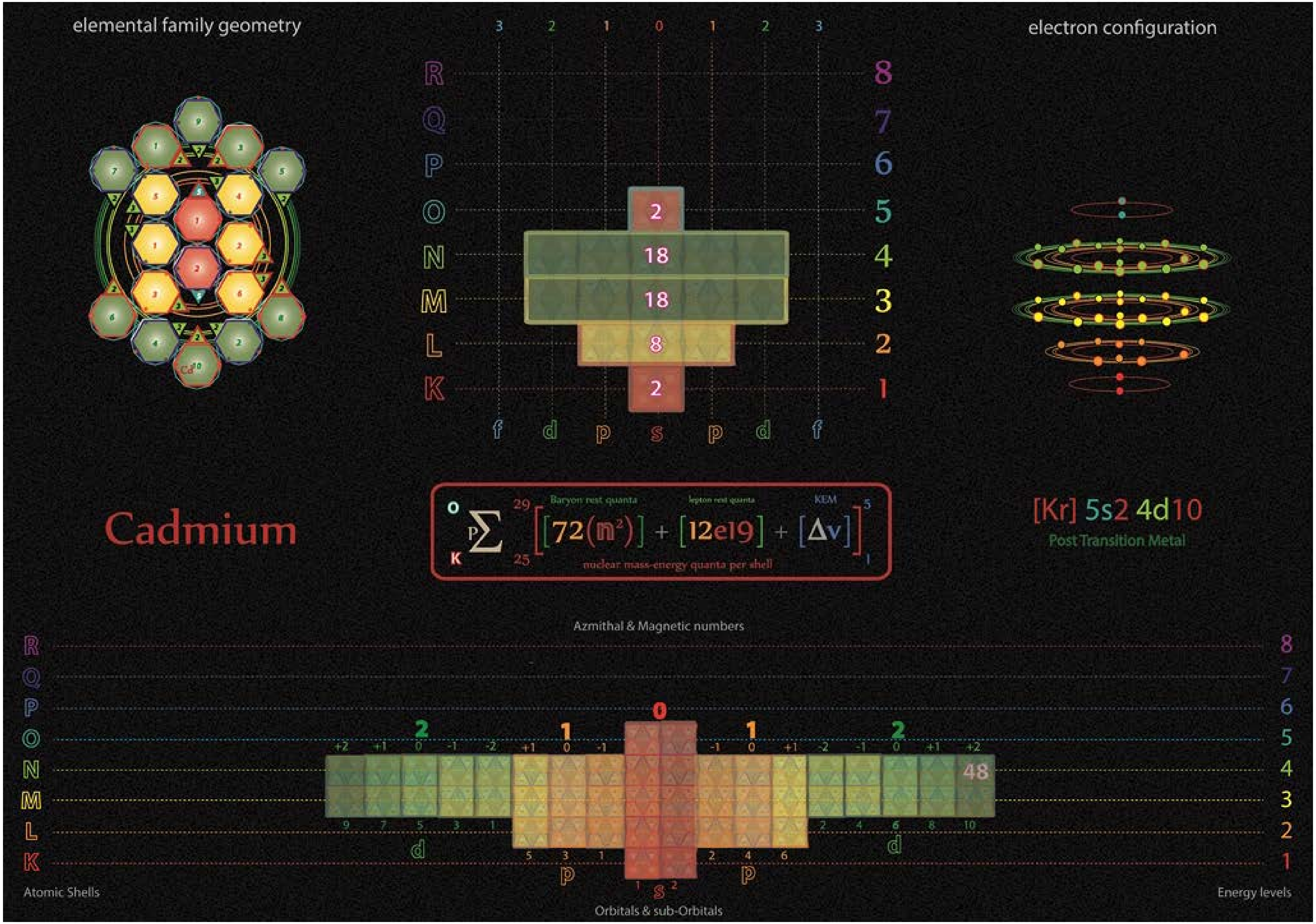
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
18
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
48
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Cadmium

$${}^{O}_{K}\sum_{P}{}^{29}_{25}\left[\overset{\text{Baryon rest quanta}}{[72(n^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}\right]^{5}_{1}$$

nuclear mass-energy quanta per shell

[Kr] 5s2 4d10

Post Transition Metal

Tetryonics 53.48 - Cadmium atomic config

Copyright ABRAHAM [2008] - All rights reserved

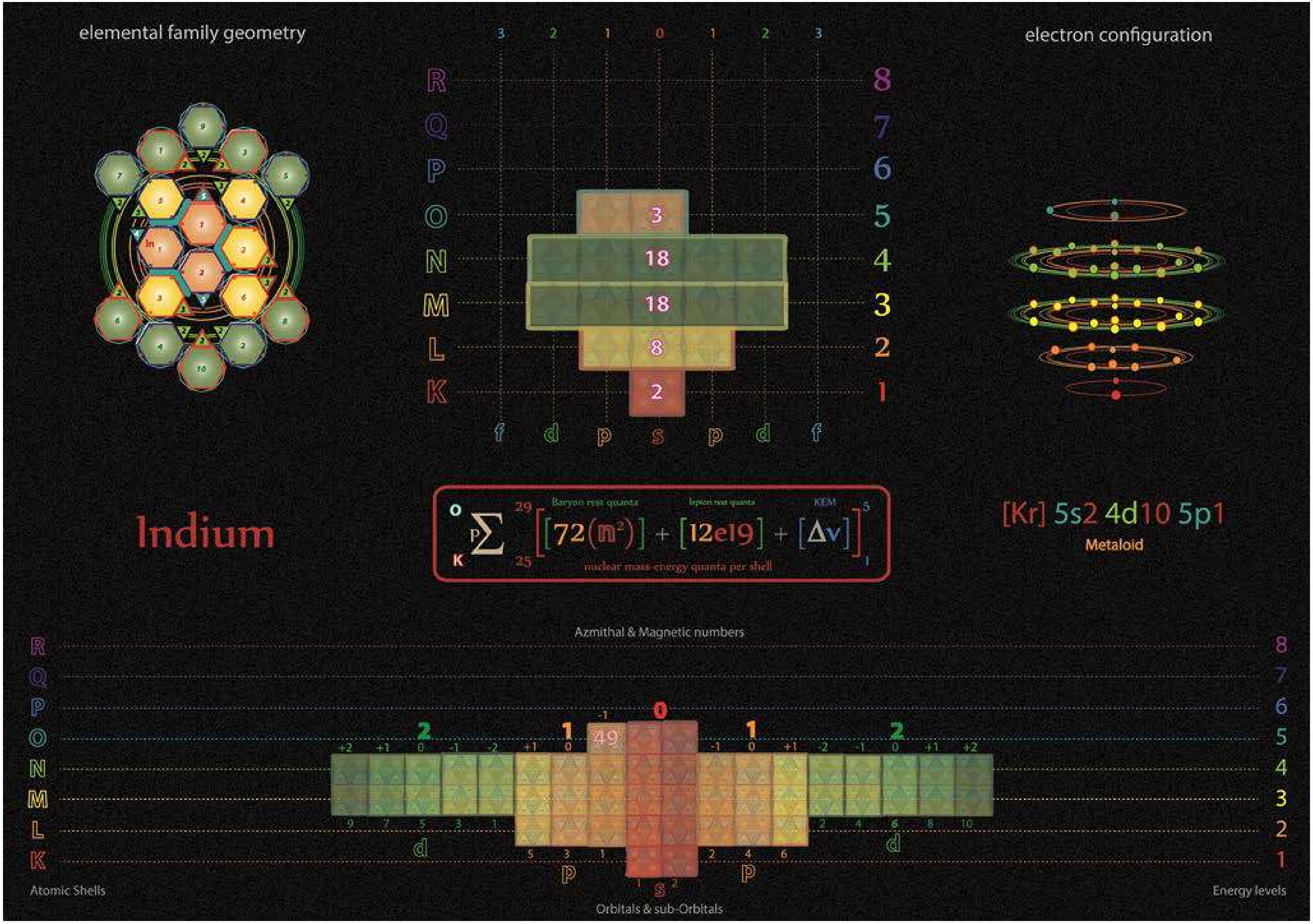
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
18
18
8
2
f d p s p d f
Indium
O
K
p
29
25
Baryon rest quanta
lepton rest quanta
KEM
72(m²)
12e19
Δv
5
1
nuclear mass-energy quanta per shell
[Kr] 5s2 4d10 5p1
Metaloid
Azmithal & Magnetic numbers
49
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d p s p d
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.49 - Indium atomic config

Copyright ABRAHAM [2008] - All rights reserved

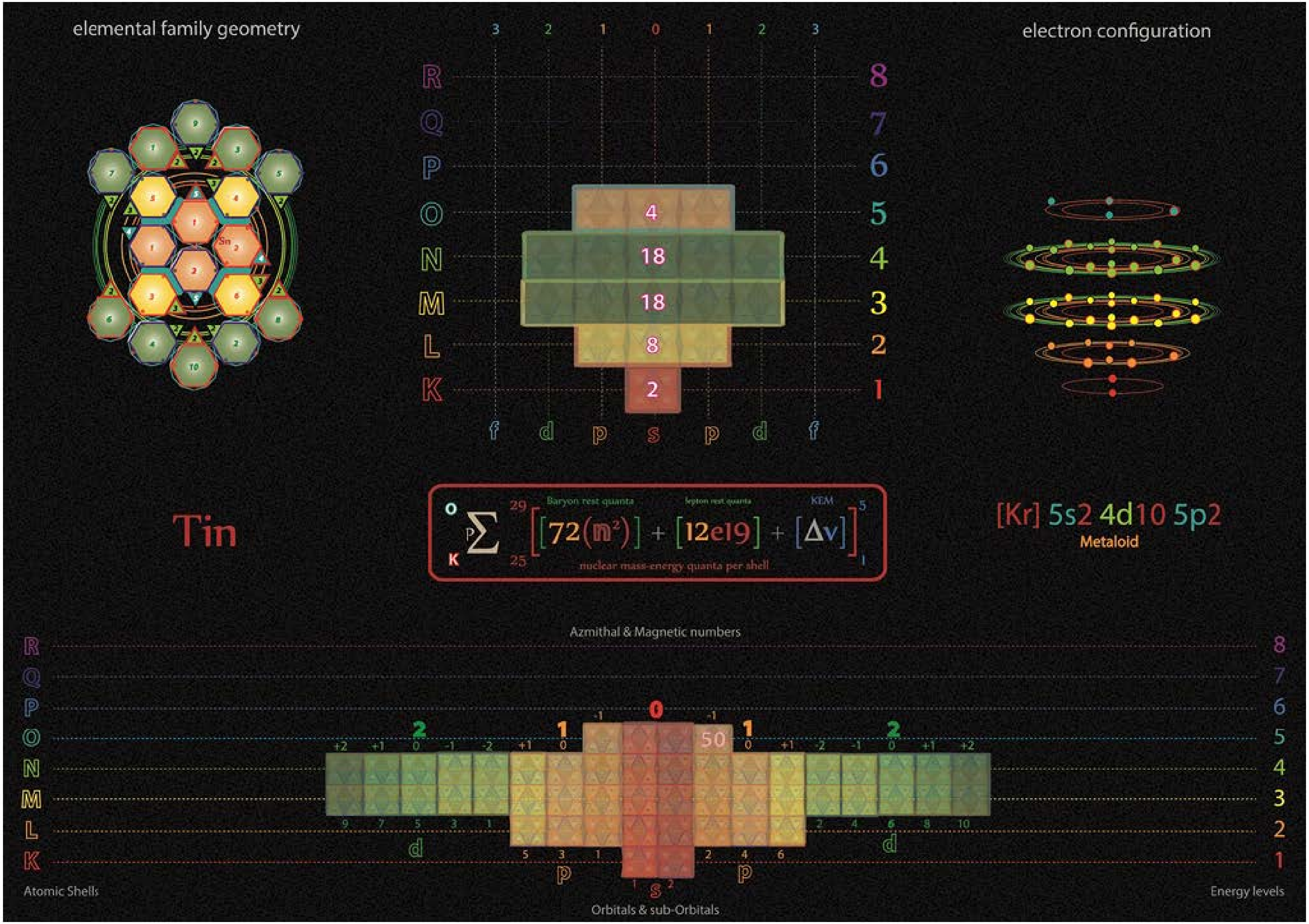
elemental family geometry
Sn
Tin
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
4
18
18
8
2
f
d
p
s
p
d
f
electron configuration
[Kr] 5s2 4d10 5p2
Metaloid
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
50
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{O} {}_{P}\left[72(m^2)\right]_{25}^{29} + \left[12e19\right] + \left[\Delta v\right]_{1}^{5}$$

Tetryonics 53.50 - Tin atomic config

Copyright ABRAHAM [2008] - All rights reserved

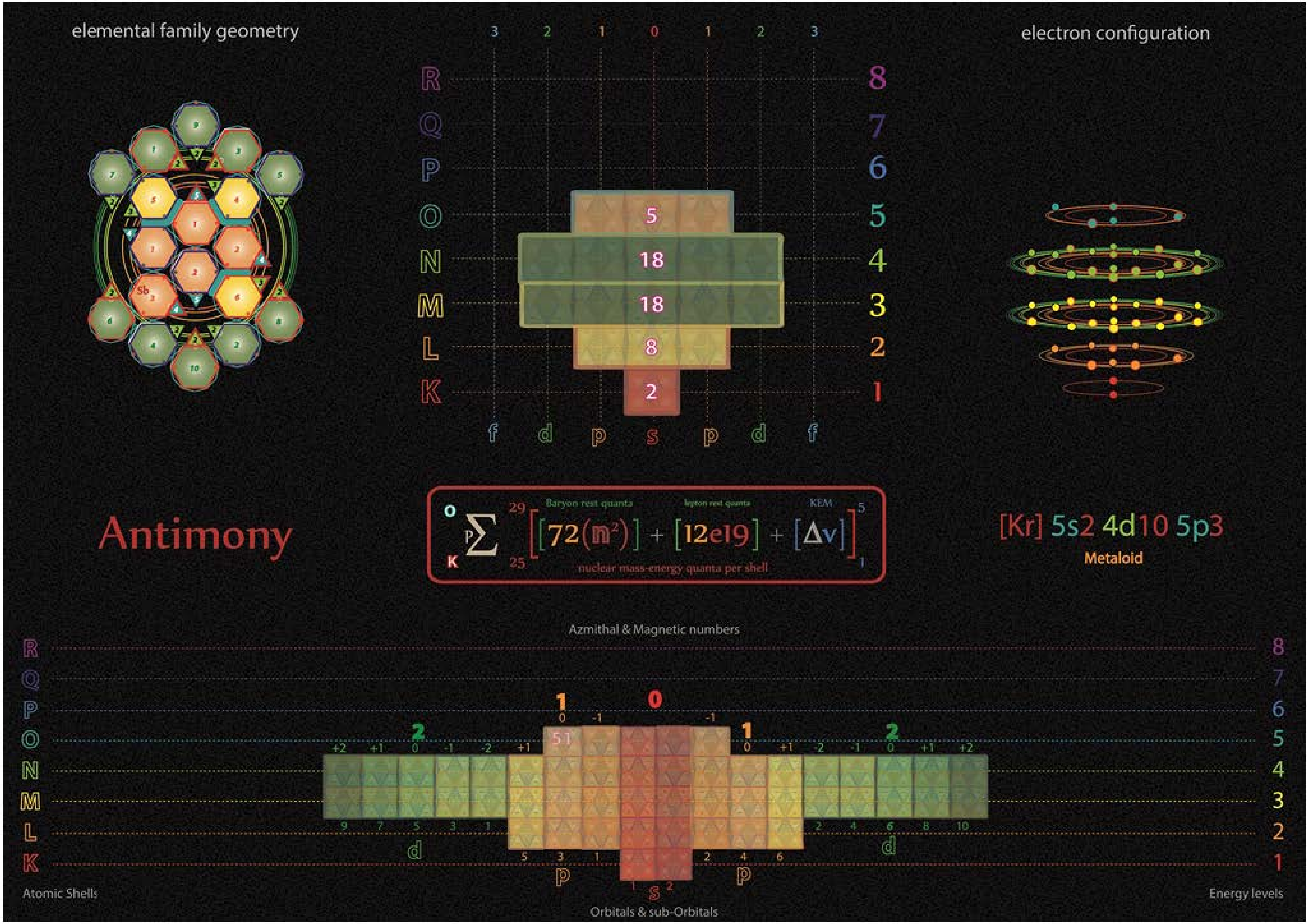
elemental family geometry
Sb
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
5
18
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
1 0 1 2 2
0 -1 -1 0
+2 +1 0 -1 -2 +1 0 +1 -2 -1 0 +1 +2
51
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
d p s p d
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Antimony

$$\sum_{K}^{O} {}_{P}\sum_{25}^{29}\left[\underset{\text{Baryon rest quanta}}{[72(m^2)]} + \underset{\text{lepton rest quanta}}{[12e19]} + \underset{\text{KEM}}{[\Delta v]}\right]_{1}^{5}$$

nuclear mass-energy quanta per shell

[Kr] 5s2 4d10 5p3

Metaloid

Tetryonics 53.51 - Antimony atomic config

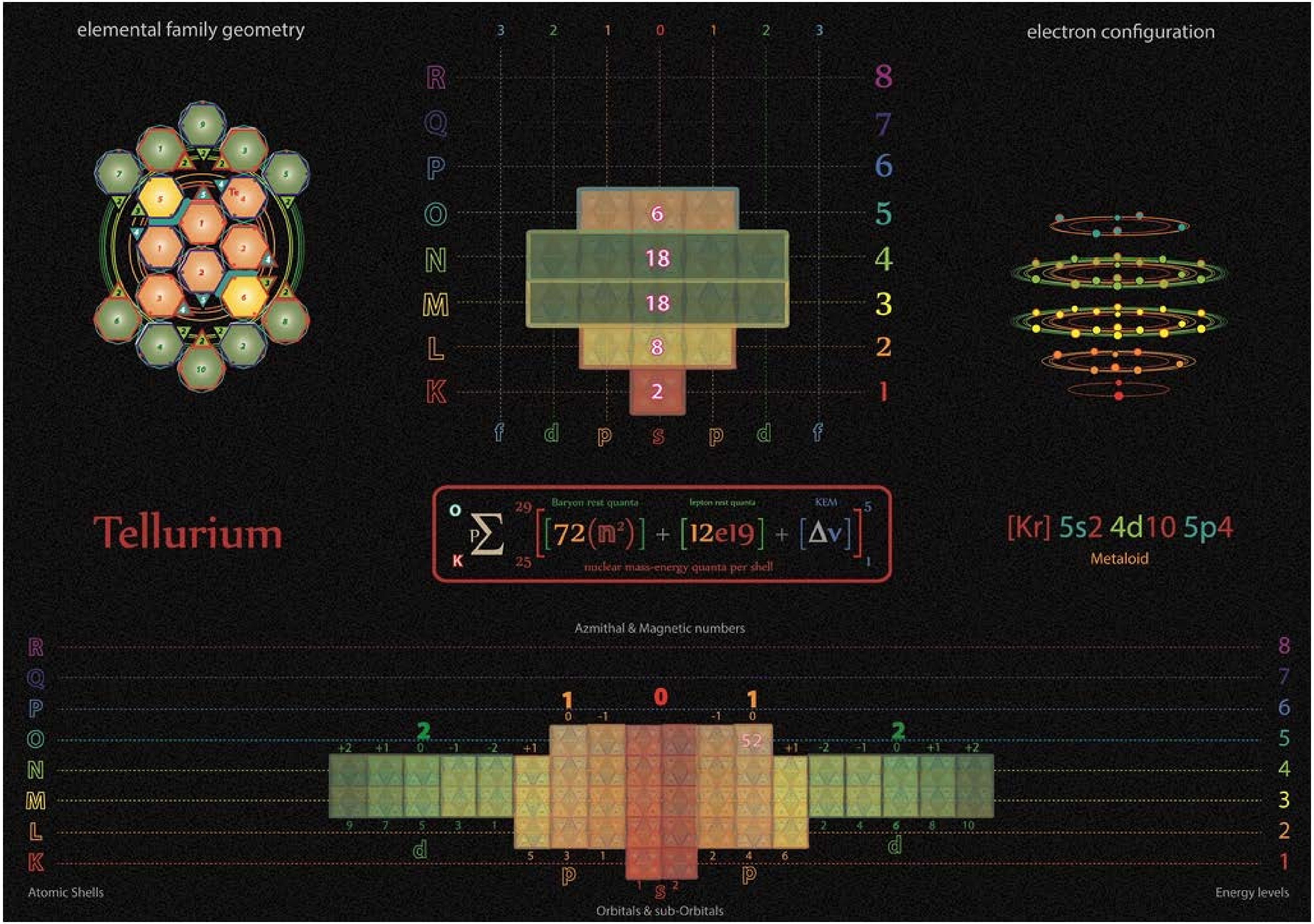
elemental family geometry
Te
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
18
18
8
2
f d p s p d f
electron configuration
Tellurium
O
K
$$_{K}^{O}{}_{p}\sum_{25}^{29}\left[\left[72(m^2)\right]+\left[12e19\right]+\left[\Delta v\right]\right]_1^5$$
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Kr] 5s2 4d10 5p4
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
52
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.52 - Tellurium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

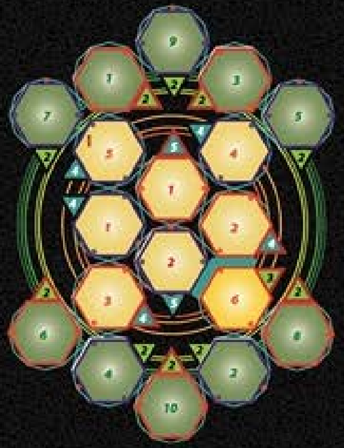

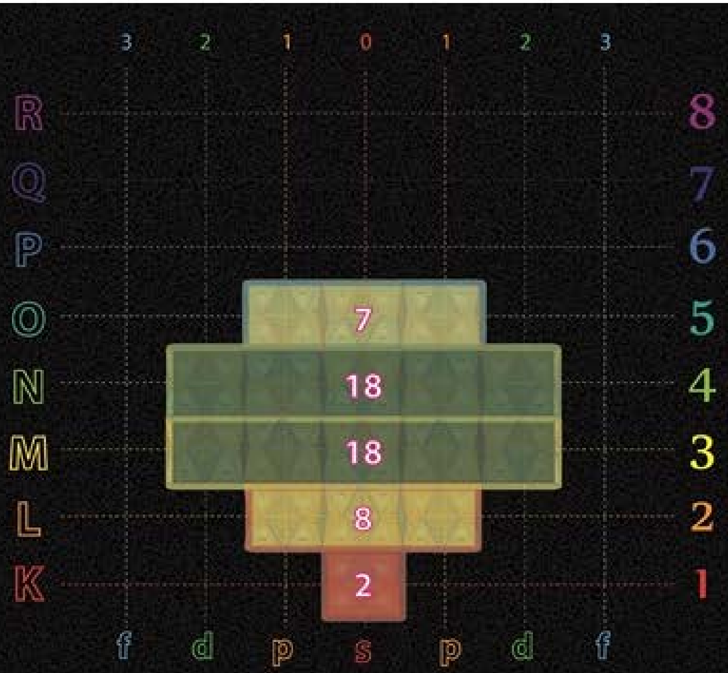
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7
18
18
8
2
f
d
p
s
p
d
f

electron configuration

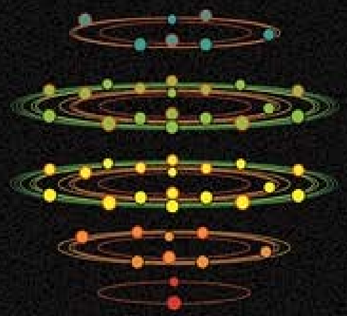

Iodine

$$\sum_{K}^{O} P \; {}_{25}^{29}\Big[72(\mathbb{m}^2)\Big] + \Big[12e19\Big] + \Big[\Delta v\Big]_{1}^{5}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Kr] 5s2 4d10 5p5

Halogen

Azmithal & Magnetic numbers

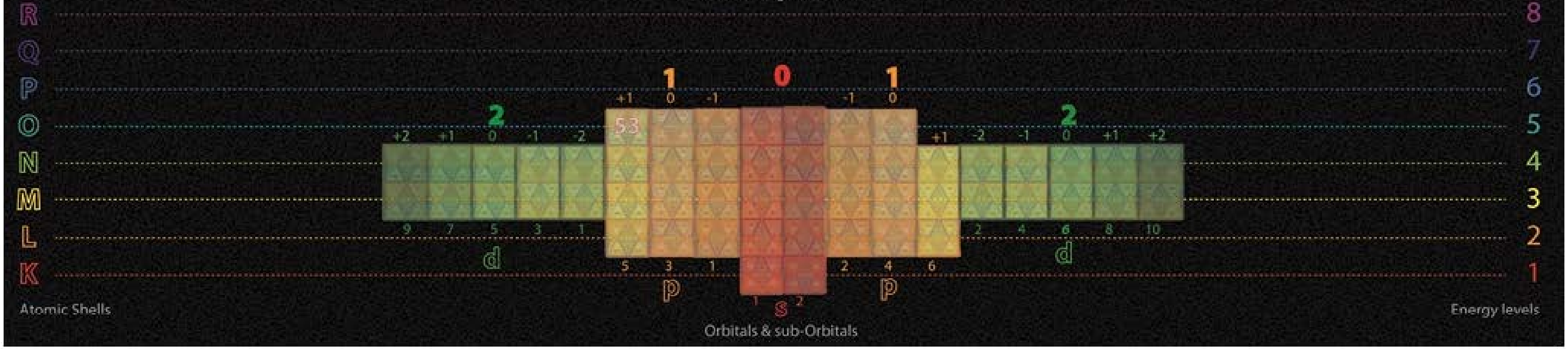
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
53
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
p
s
p
d
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.53 - Iodine atomic config

Copyright ABRAHAM [2008] - All rights reserved

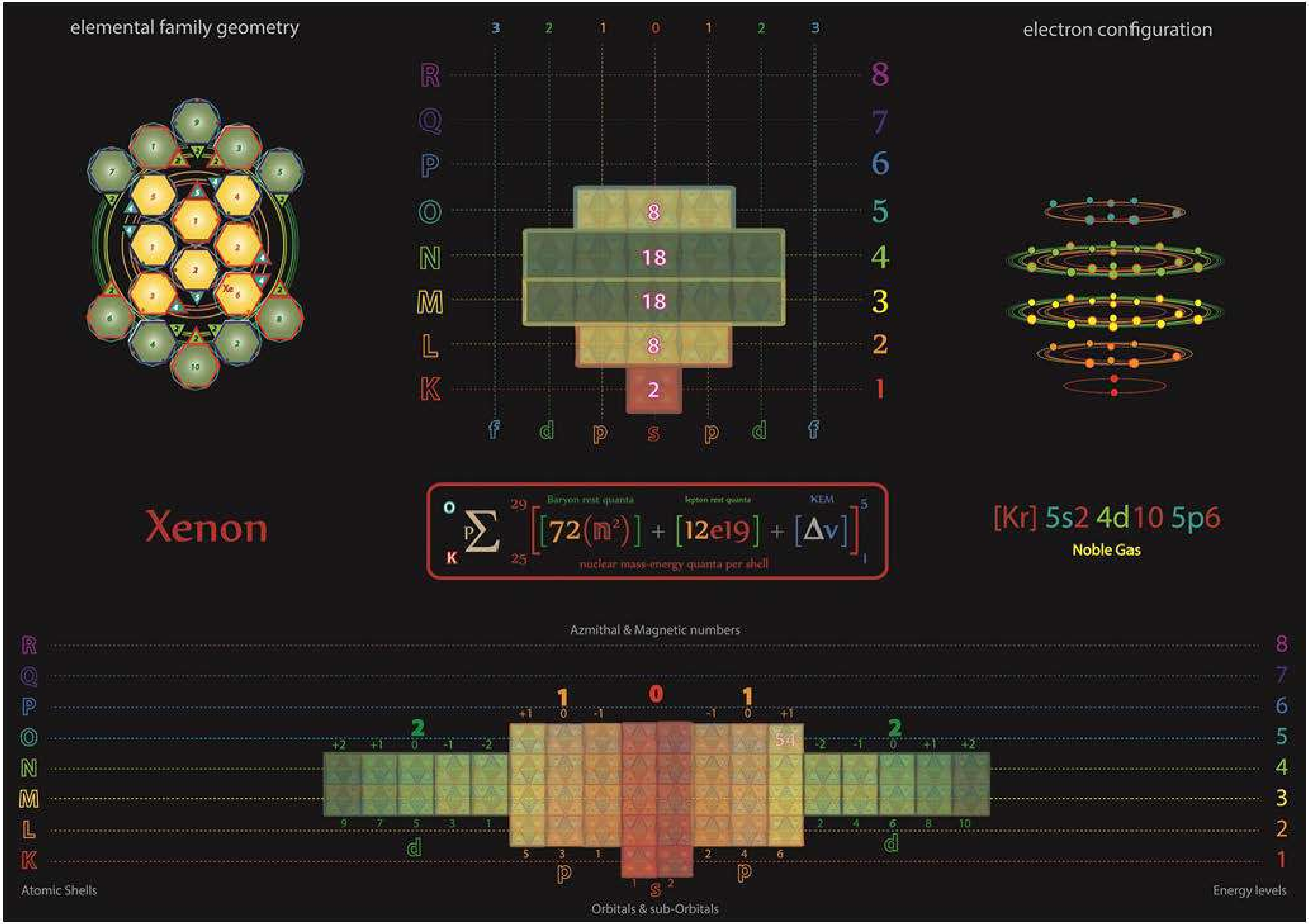
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
8
18
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
2 1 0 1 2
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
54
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d p s p d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Xenon

$$ {}^{O}_{K}\,{}_{P}\!\sum_{25}^{29}\left[\,[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\,\right]_{1}^{5} $$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Kr] 5s2 4d10 5p6

Noble Gas

Tetryonics 53.54 - Xenon atomic config

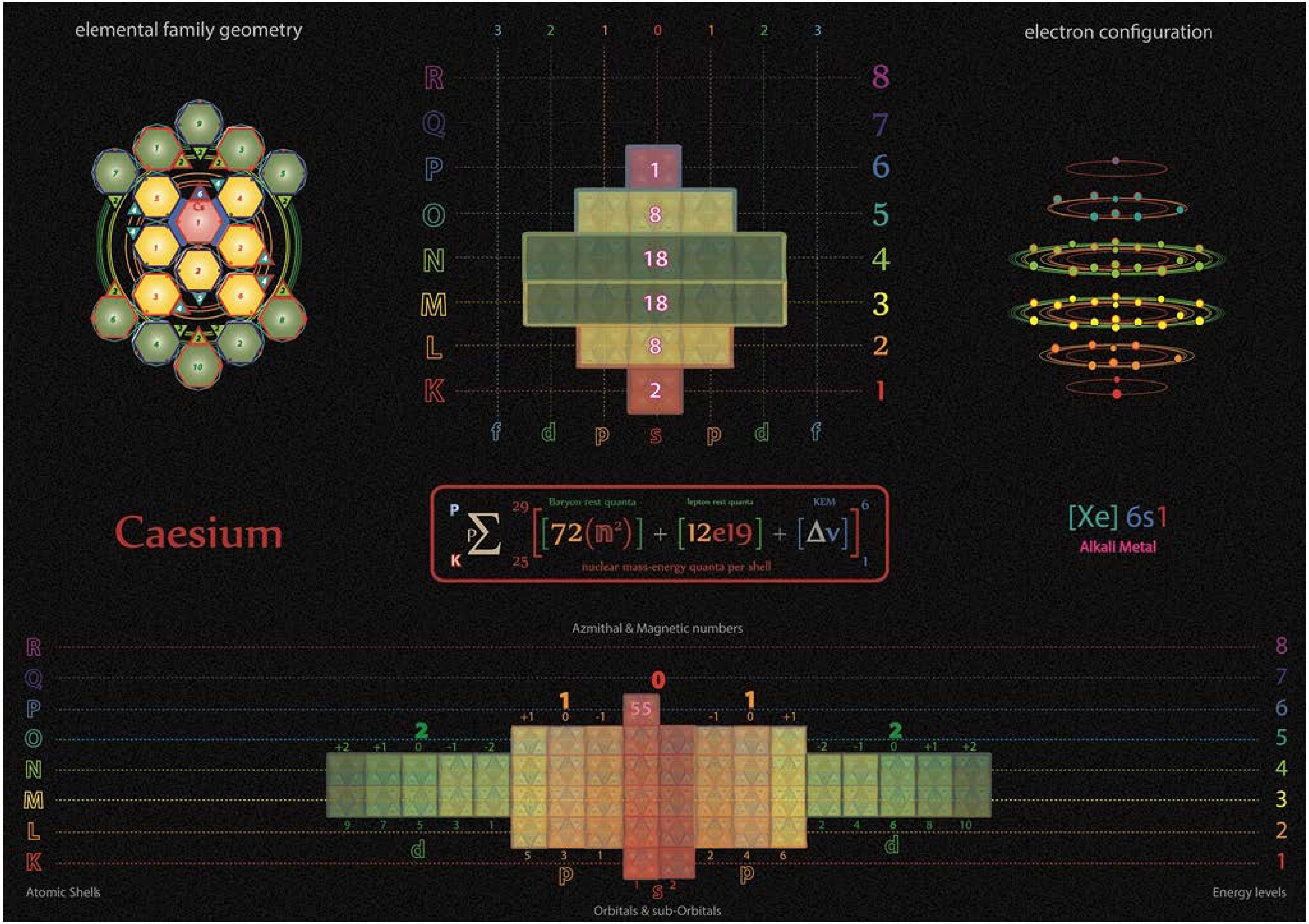
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
1
8
18
18
8
2
f d p s p d f
Caesium
Baryon rest quanta
lepton rest quanta
KEM
P ∑ K 29 25 [72(m²)] + [12e19] + [Δv] 6 1
nuclear mass-energy quanta per shell
[Xe] 6s1
Alkali Metal
Azmithal & Magnetic numbers
0
1 1
2 2
55
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -2 -1 0 +1 +2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
d p s p d
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.55 - Caesium atomic config

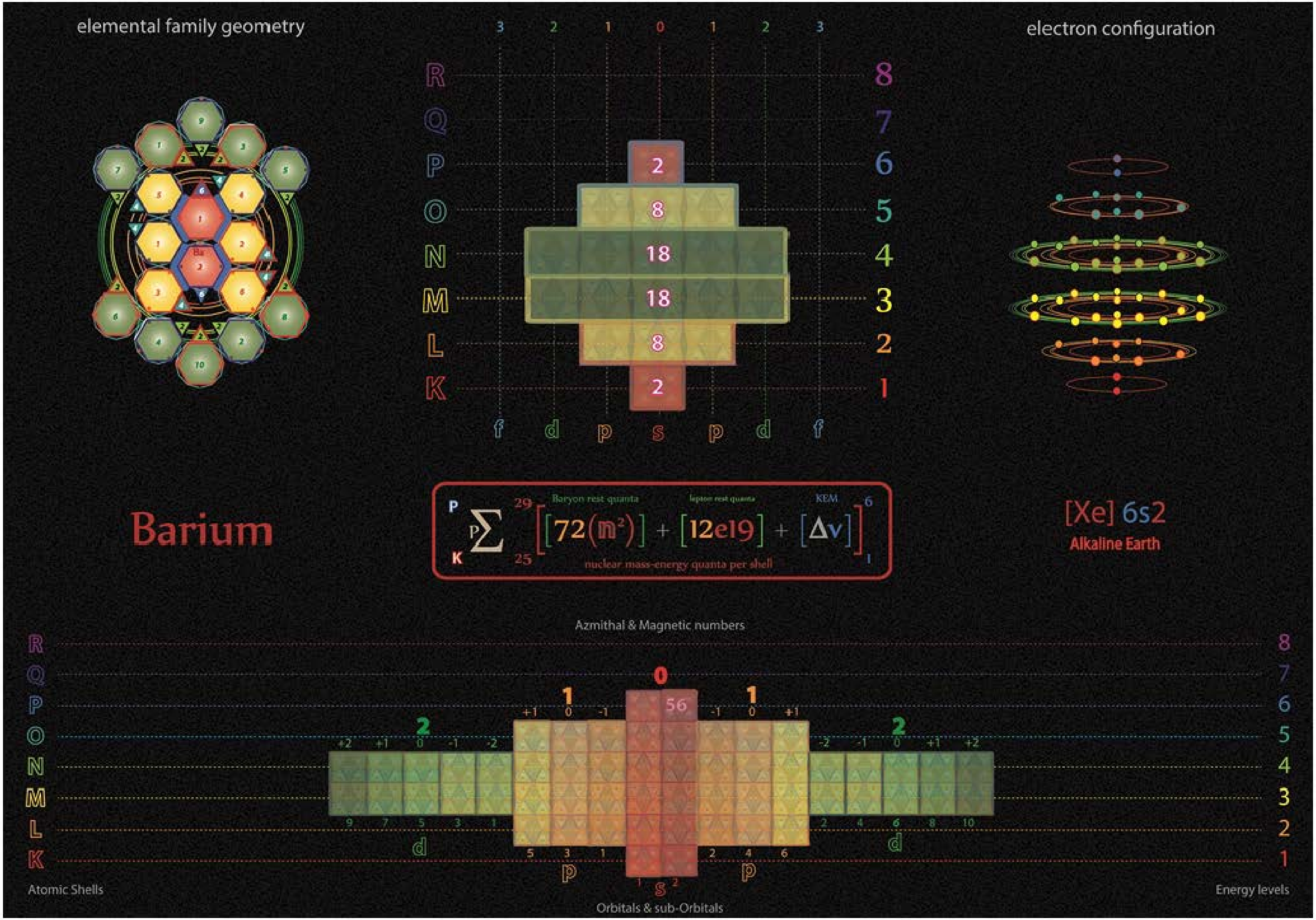
elemental family geometry
Ba
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
18
18
8
2
f d p s p d f
electron configuration
Barium
P
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 6s2
Alkaline Earth
Azmithal & Magnetic numbers
0
1
1
2
2
56
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
9 7 5 3 1
5 3 1
2 4 6
2 4 6 8 10
d
p
s
p
d
1 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.56 - Barium atomic config

Copyright ABRAHAM [2008] - All rights reserved

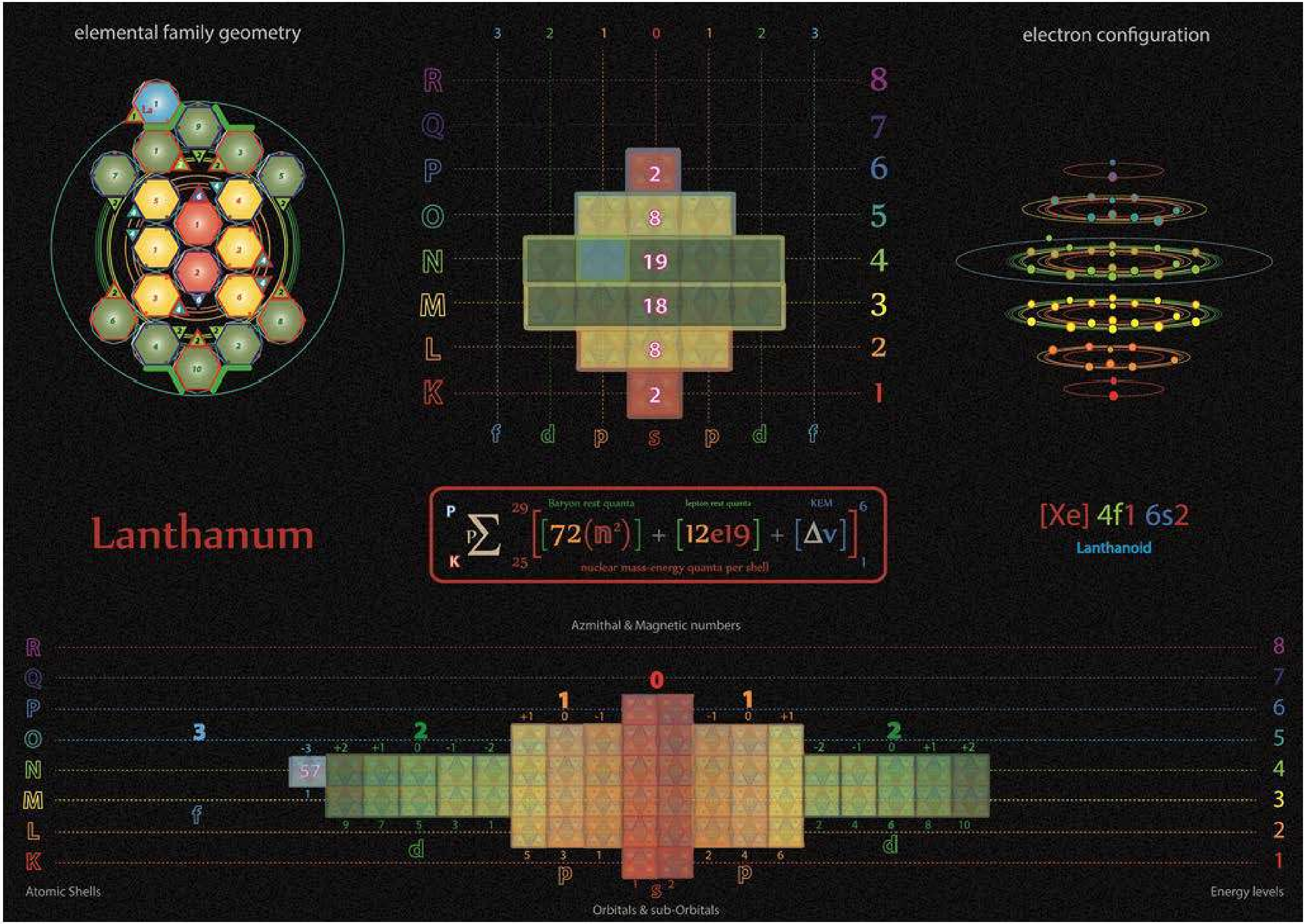
elemental family geometry
La
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
19
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
3
2
1
0
1
2
-3 +2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
57
1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
f
d
p
s
p
d
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Lanthanum

$$\sum_{K}^{P} {}_{P}\!\left[\overset{\text{Baryon rest quanta}}{\left[72(n^2)\right]}_{25}^{29} + \overset{\text{lepton rest quanta}}{\left[12e19\right]} + \overset{\text{KEM}}{\left[\Delta v\right]} \right]_{1}^{6}$$

nuclear mass-energy quanta per shell

[Xe] 4f1 6s2

Lanthanoid

Tetryonics 53.57 - Lanthanum atomic config

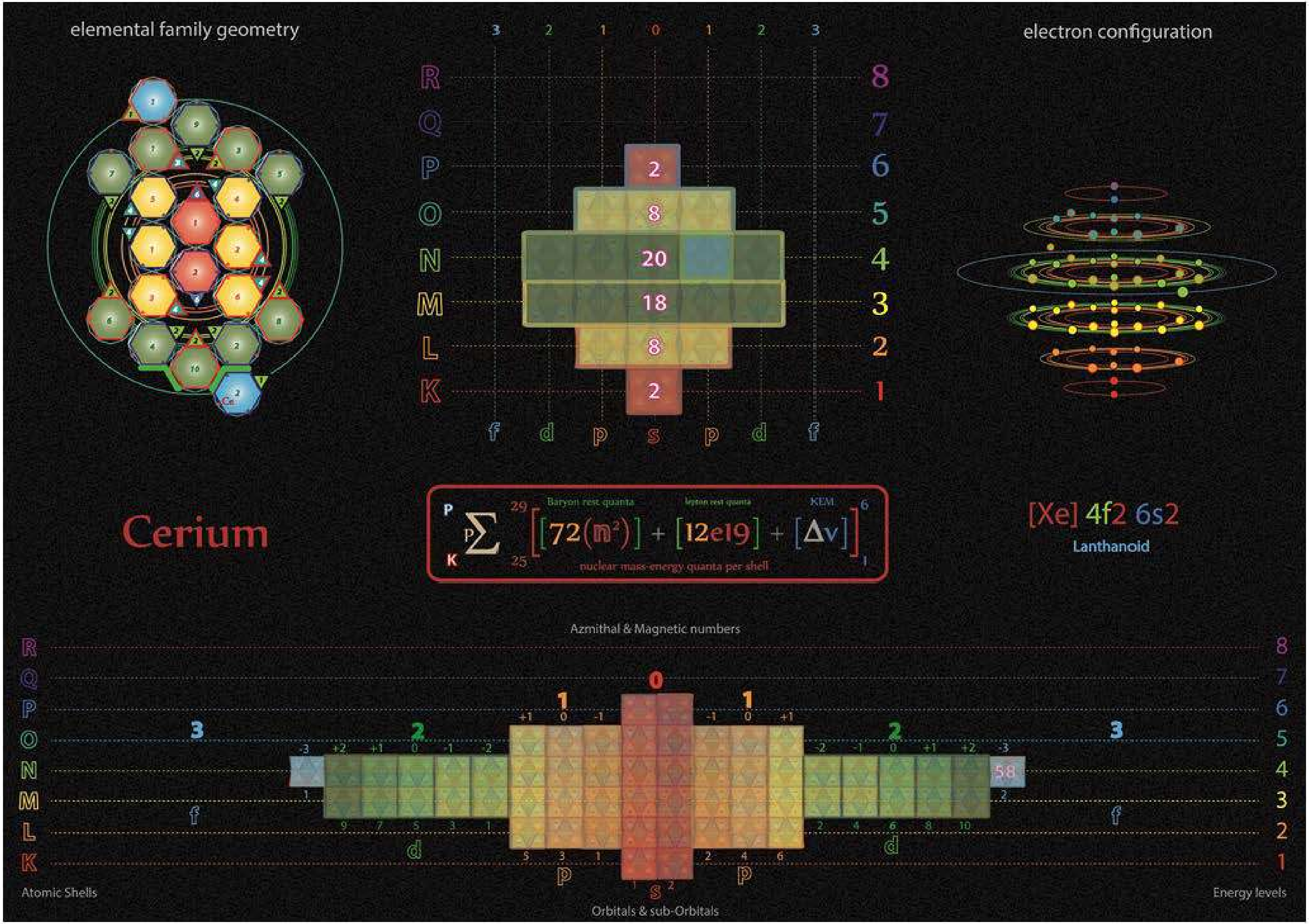

elemental family geometry

electron configuration

Cerium

$$^{P}_{K}\,{}_{P}\sum_{25}^{29}\left[\underset{\text{Baryon rest quanta}}{\left[72(\mathrm{m}^2)\right]}+\underset{\text{lepton rest quanta}}{\left[12e19\right]}+\underset{\text{KEM}}{\left[\Delta v\right]}\right]_{1}^{6}$$

nuclear mass-energy quanta per shell

[Xe] 4f2 6s2

Lanthanoid

Azmithal & Magnetic numbers

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
20
18
8
2
f d p s p d f
Ce
0
1 1
+1 0 -1 -1 0 +1
3 2 2 3
-3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3
58
1 2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
f d p s p d f

Tetryonics 53.58 - Cerium atomic config

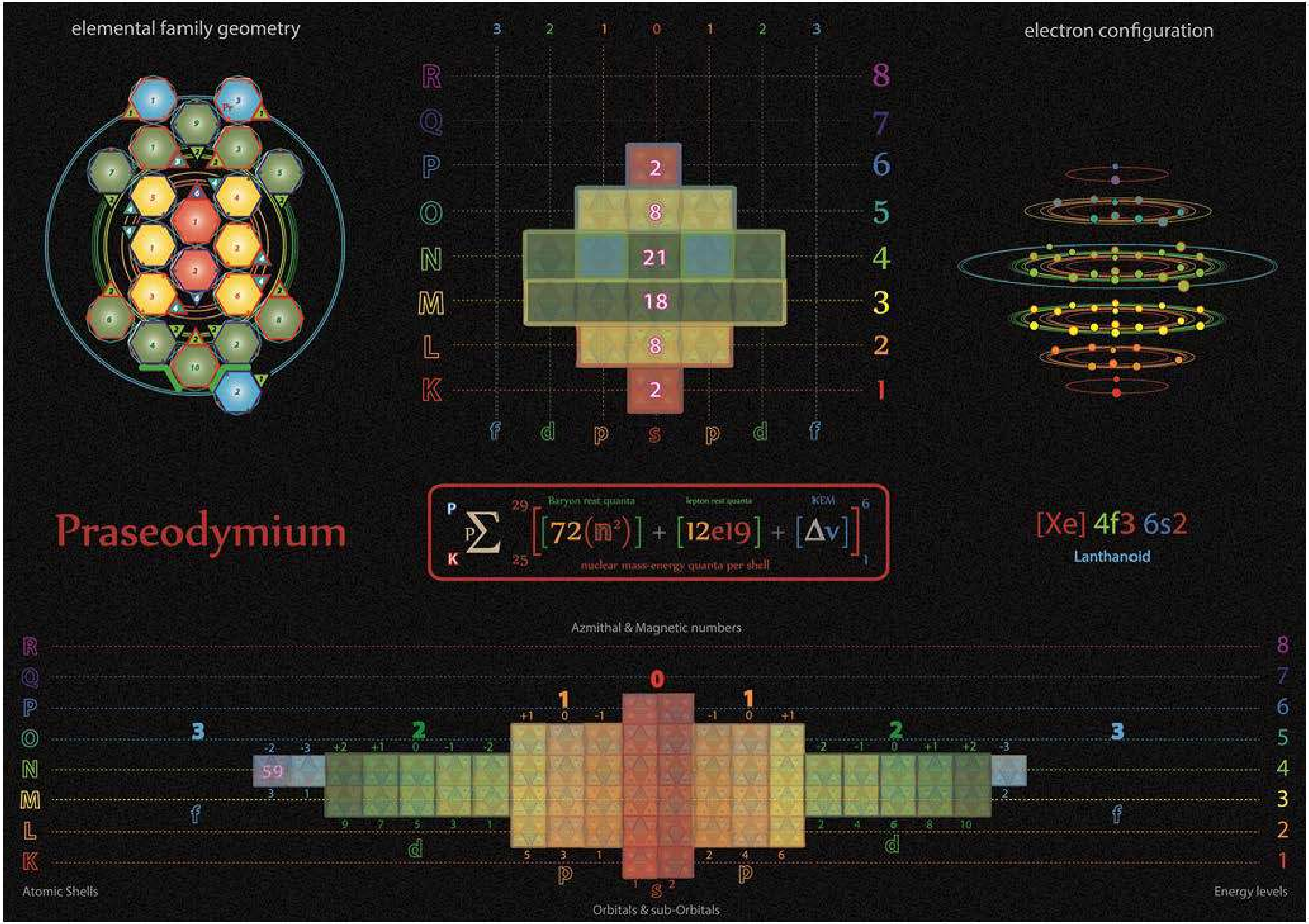
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
21
18
8
2
f d p s p d f
Praseodymium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
6
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Xe] 4f3 6s2
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
+1 0 -1 -1 0 +1
3 2 2 3
-2 -3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3
59
3 1 9 7 5 3 1 2 4 6 8 10 2
f d d f
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.59 - Praseodymium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

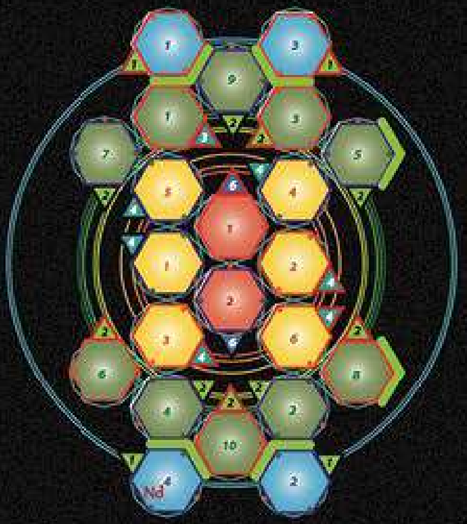

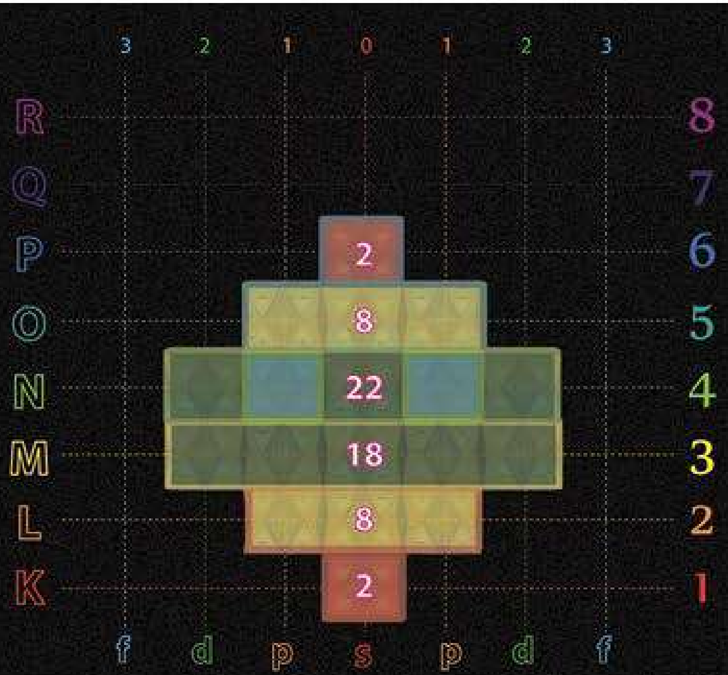
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
22
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

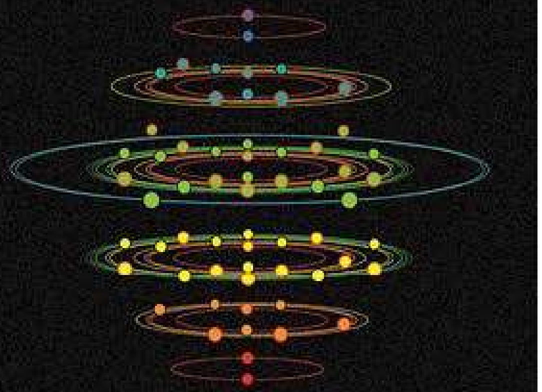

Neodymium

$$\sum_{K}^{P} {}_{p} \, {}_{25}^{29}\left[\underbrace{\overset{\text{Baryon rest quanta}}{[72(\mathrm{n}^2)]} + \overset{\text{lepton rest quanta}}{[12\mathrm{e}19]}}_{\text{nuclear mass-energy quanta per shell}} + \overset{\text{KEM}}{[\Delta v]} \right]_{1}^{6}$$

[Xe] 4f4 6s2

Lanthanoid

Azmithal & Magnetic numbers

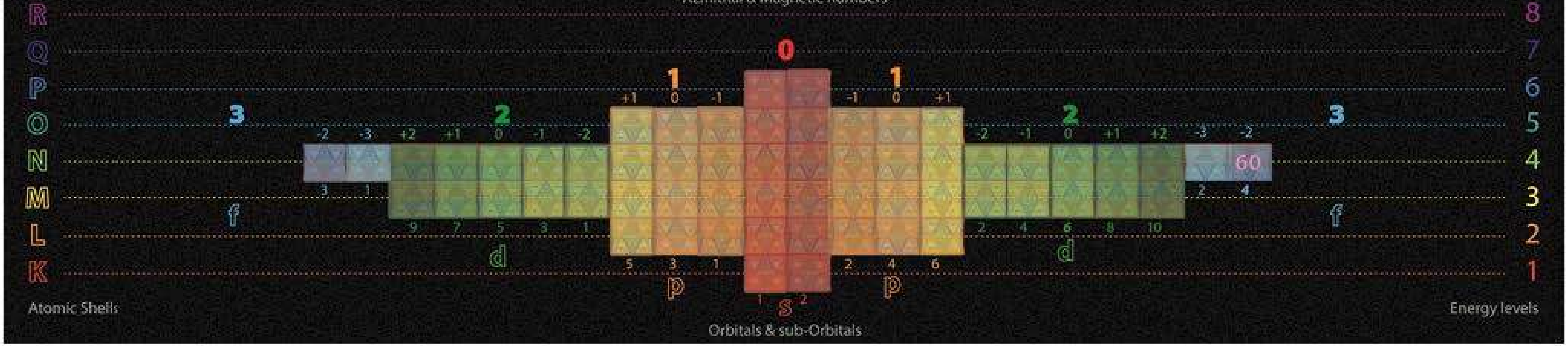
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
+1
0
-1
-1
0
+1
-2
-3
+2
+1
0
-1
-2
-2
-1
0
+1
+2
-3
-2
60
3
1
2
4
9
7
5
3
1
2
4
6
8
10
5
3
1
2
4
6
1
2
f
d
p
s
p
d
f

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.60 - Neodymium atomic config

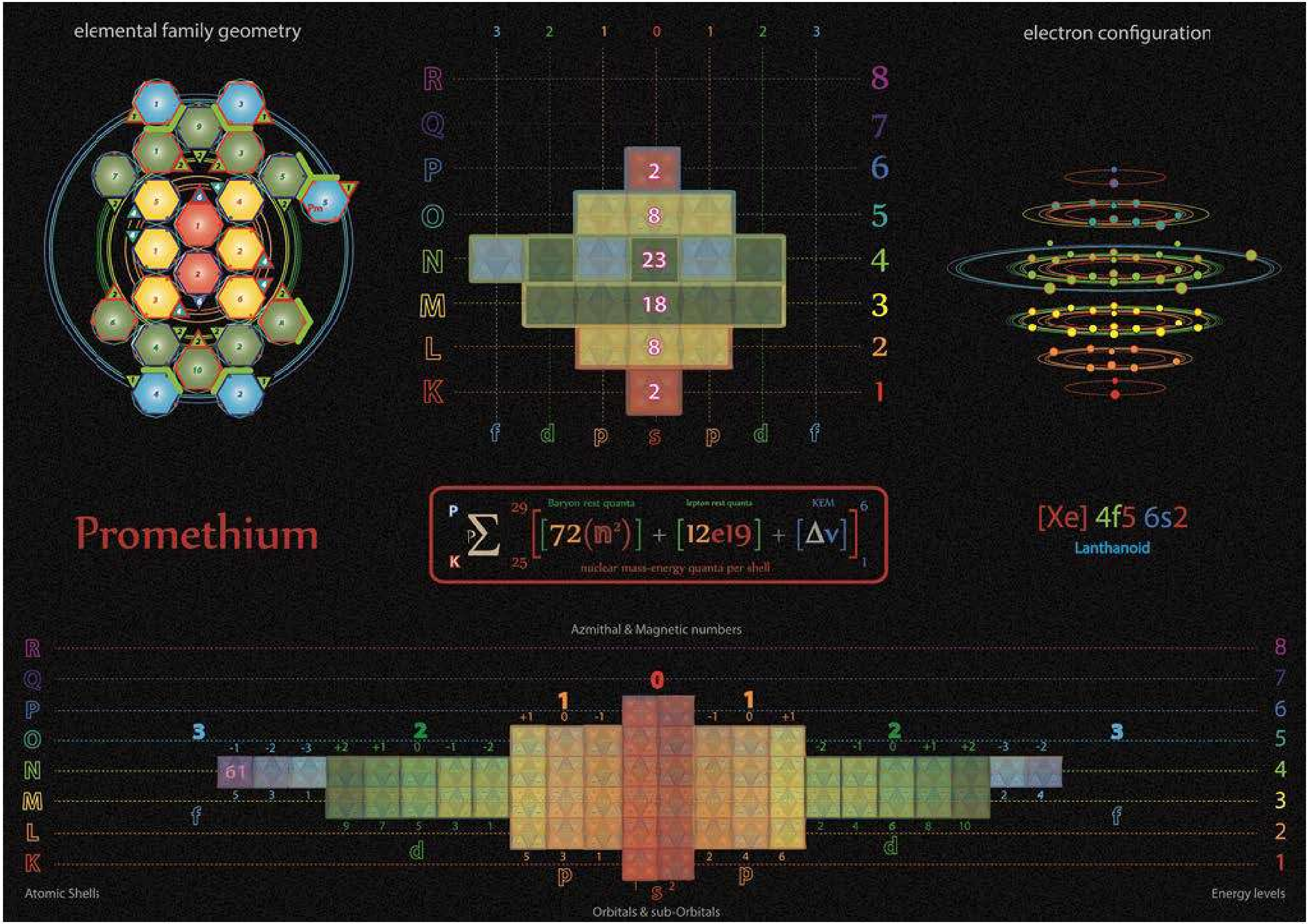
elemental family geometry
Pm
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
23
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
-1 -2 -3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3 -2
61
5 3 1 2 4
5 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Promethium

$$\sum_{K}^{P} {}_{25}^{29}\left[72(m^2)\right] + \left[12e19\right] + \left[\Delta v\right]_{1}^{6}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Xe] 4f5 6s2

Lanthanoid

Tetryonics 53.61 - Promethium atomic config

Copyright ABRAHAM [2008] - All rights reserved

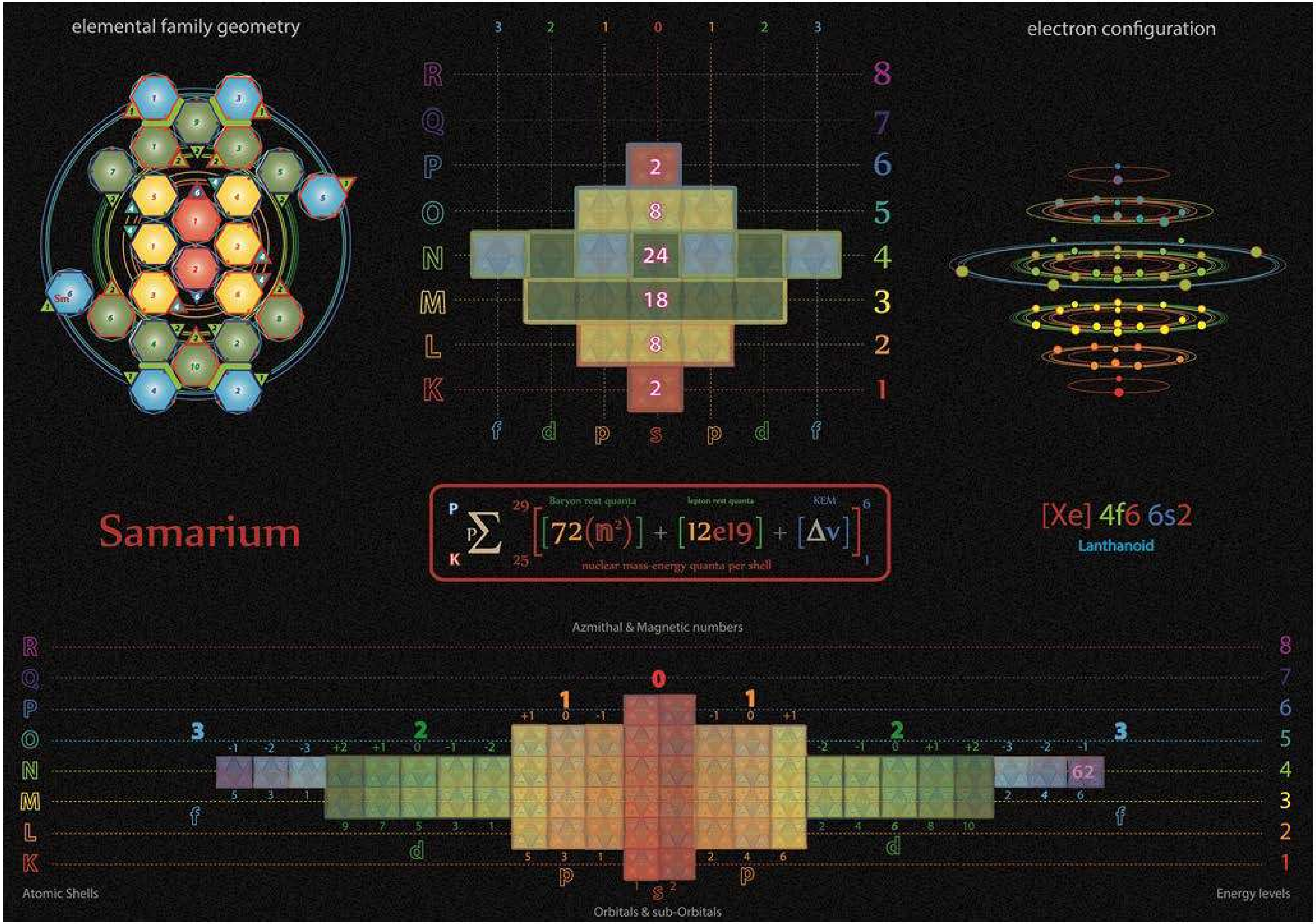
elemental family geometry
Sm
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2
8
24
18
8
2
f d p s p d f
Samarium
P
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 4f6 6s2
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
-1 -2 -3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3 -2 -1
62
5 3 1 2 4 6
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.62 - Samarium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

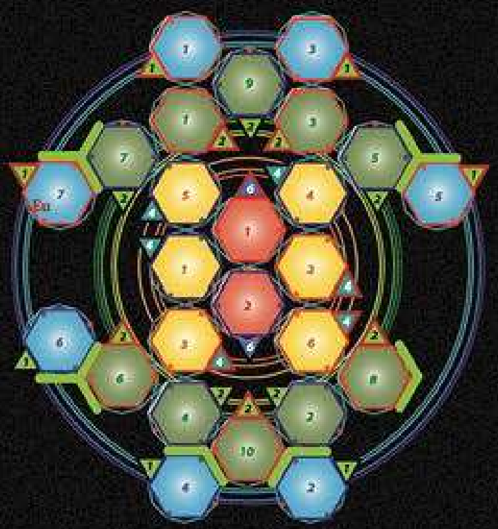

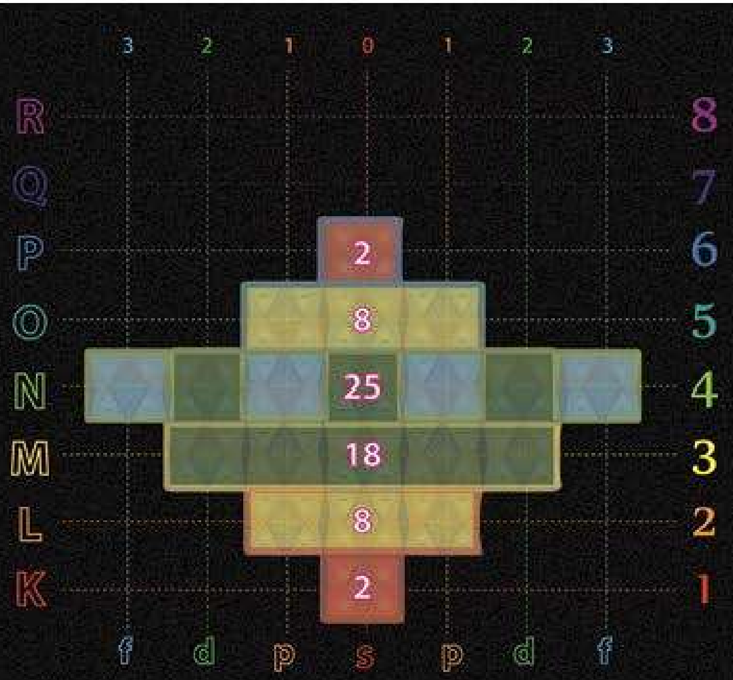
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
25
18
8
2
f
d
p
s
p
d
f

electron configuration

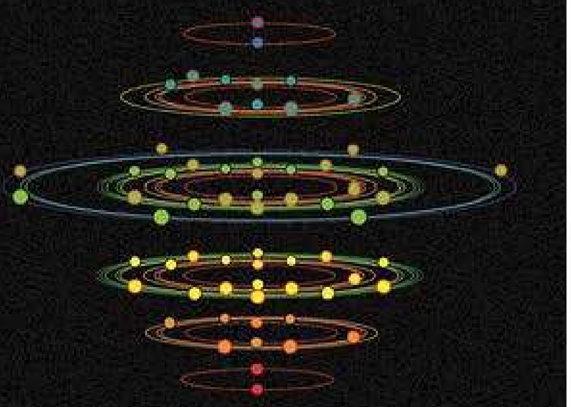

Europium

$$ {}_{K}^{P}\sum_{25}^{29} \Big[\underbrace{\overbrace{[72(n^2)]}^{\text{Baryon rest quanta}} + \overbrace{[12e19]}^{\text{lepton rest quanta}} + \overbrace{[\Delta v]}^{\text{KEM}}}_{\text{nuclear mass-energy quanta per shell}} \Big]_{1}^{6} $$

[Xe] 4f7 6s2

Lanthanoid

Azmithal & Magnetic numbers

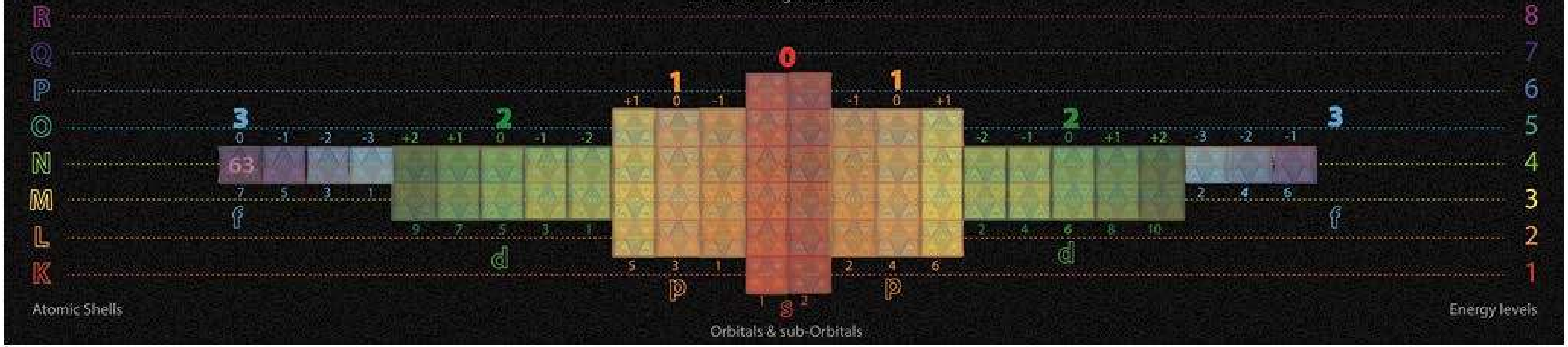
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
63
f
d
p
s
p
d
f
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.63 - Europium atomic config

Copyright ABRAHAM [2008] - All rights reserved

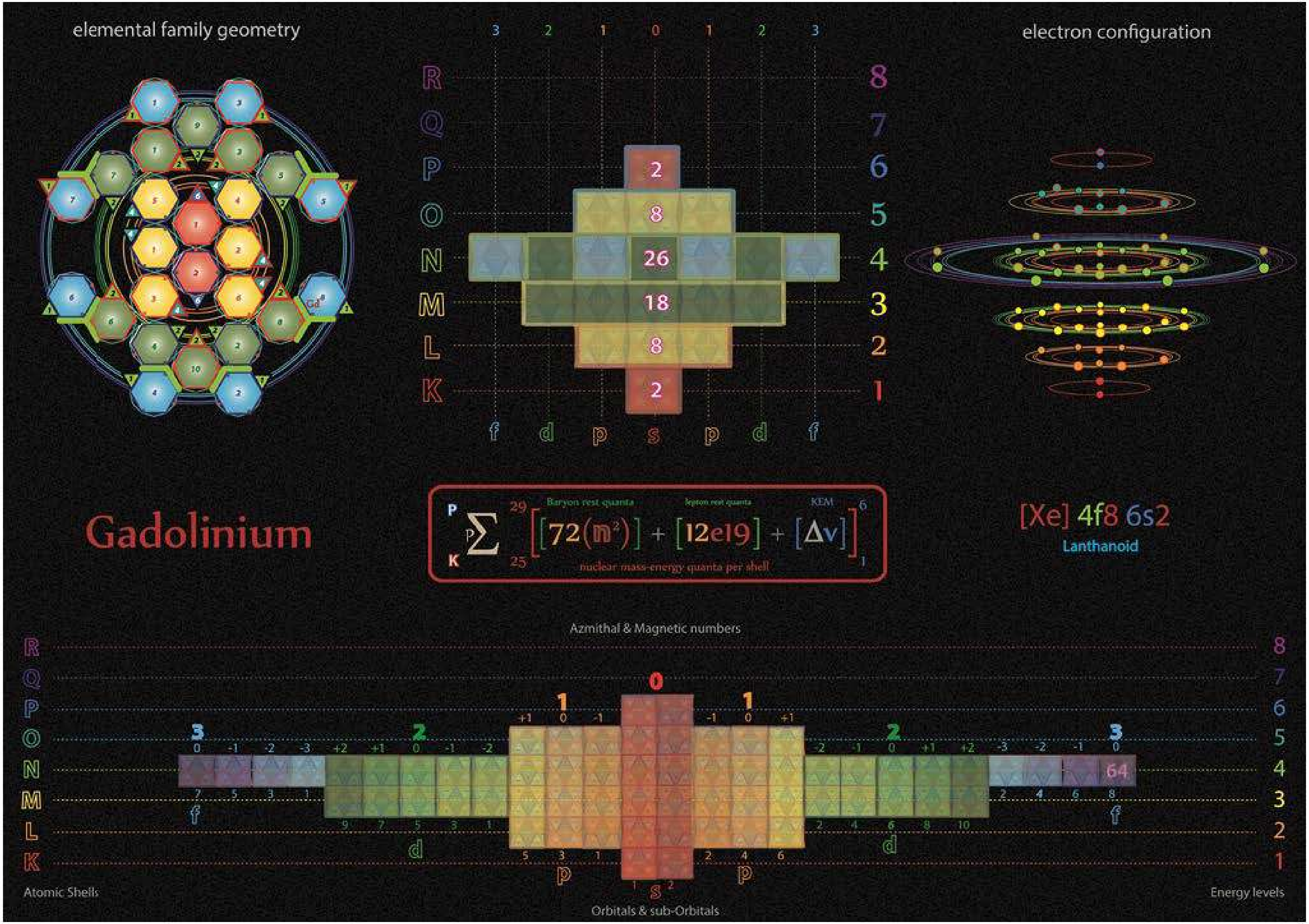
elemental family geometry
Gd
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
26
18
8
2
f d p s p d f
electron configuration
Gadolinium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 4f8 6s2
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1
-1 0 +1
0 -1 -2 -3 +2 +1 0 -1 -2
-2 -1 0 +1 +2 -3 -2 -1 0
64
7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.64 - Gadolinium atomic config

elemental family geometry

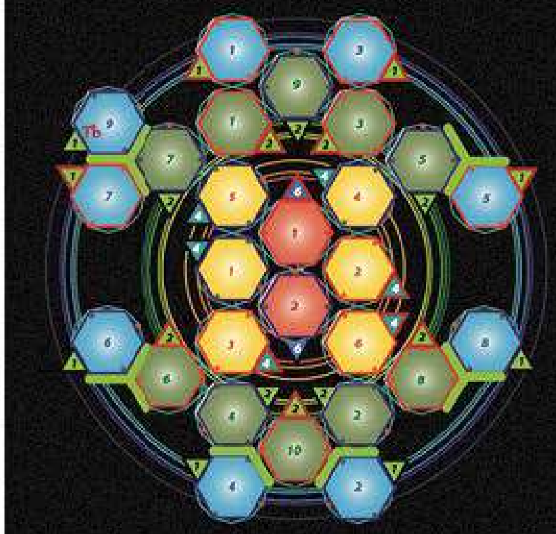
Tb
1
2
3
4
5
6
7
8
9
10

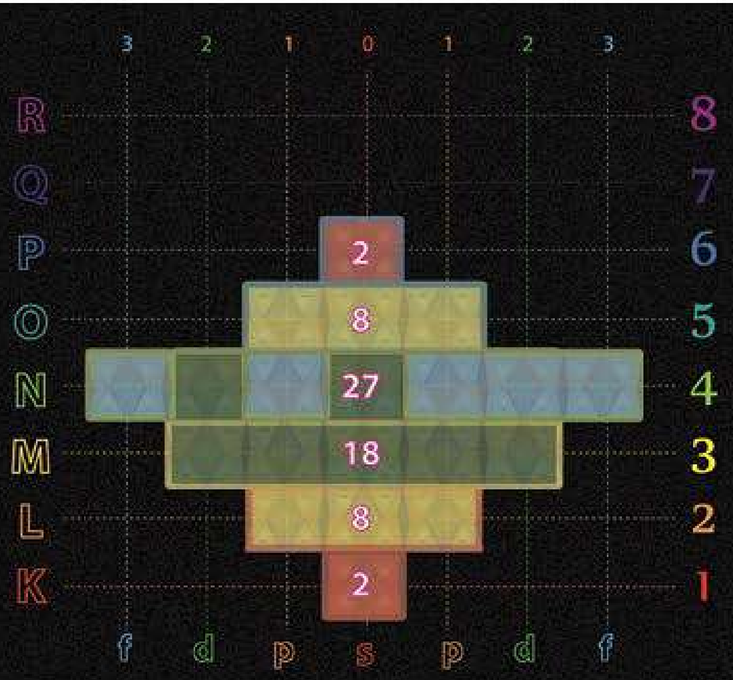
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
27
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

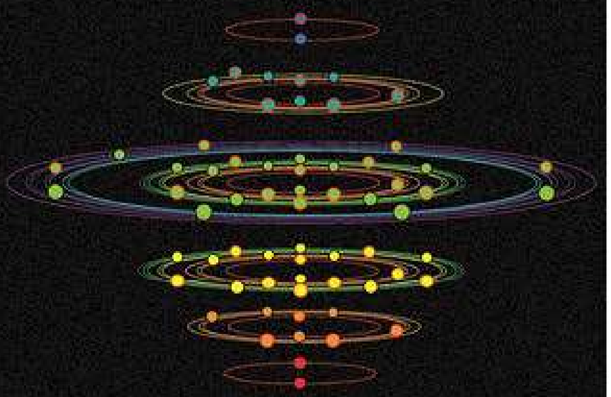

Terbium

$$\sum_{K}^{P} {}_{p}\,{}^{29}_{25}\left[72(\mathrm{m}^2)\right] + \left[12e19\right] + \left[\Delta v\right]^{6}_{1}$$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Xe] 4f9 6s2

Lanthanoid

Azmithal & Magnetic numbers

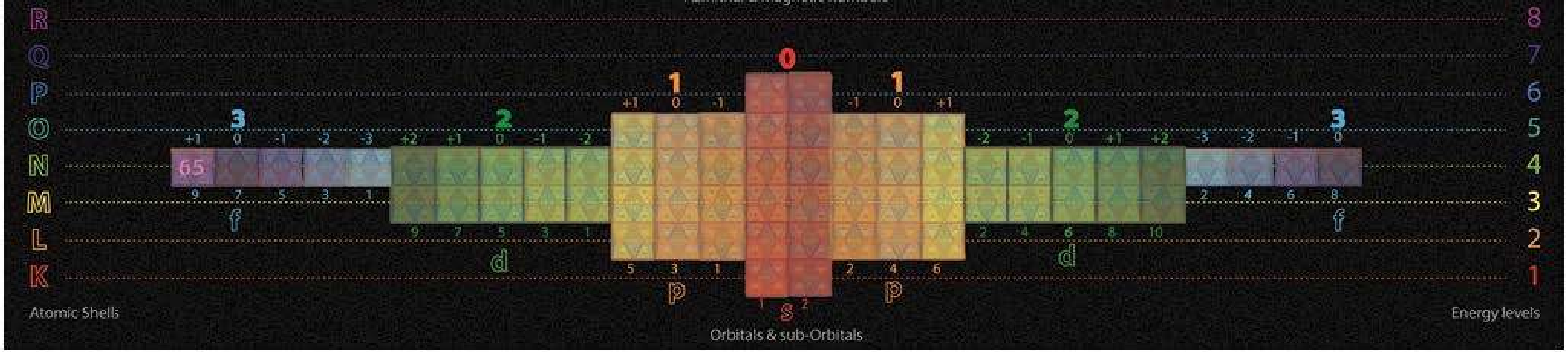
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
65
f
d
p
s
p
d
f

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.65 - Terbium atomic config

Copyright ABRAHAM [2008] - All rights reserved

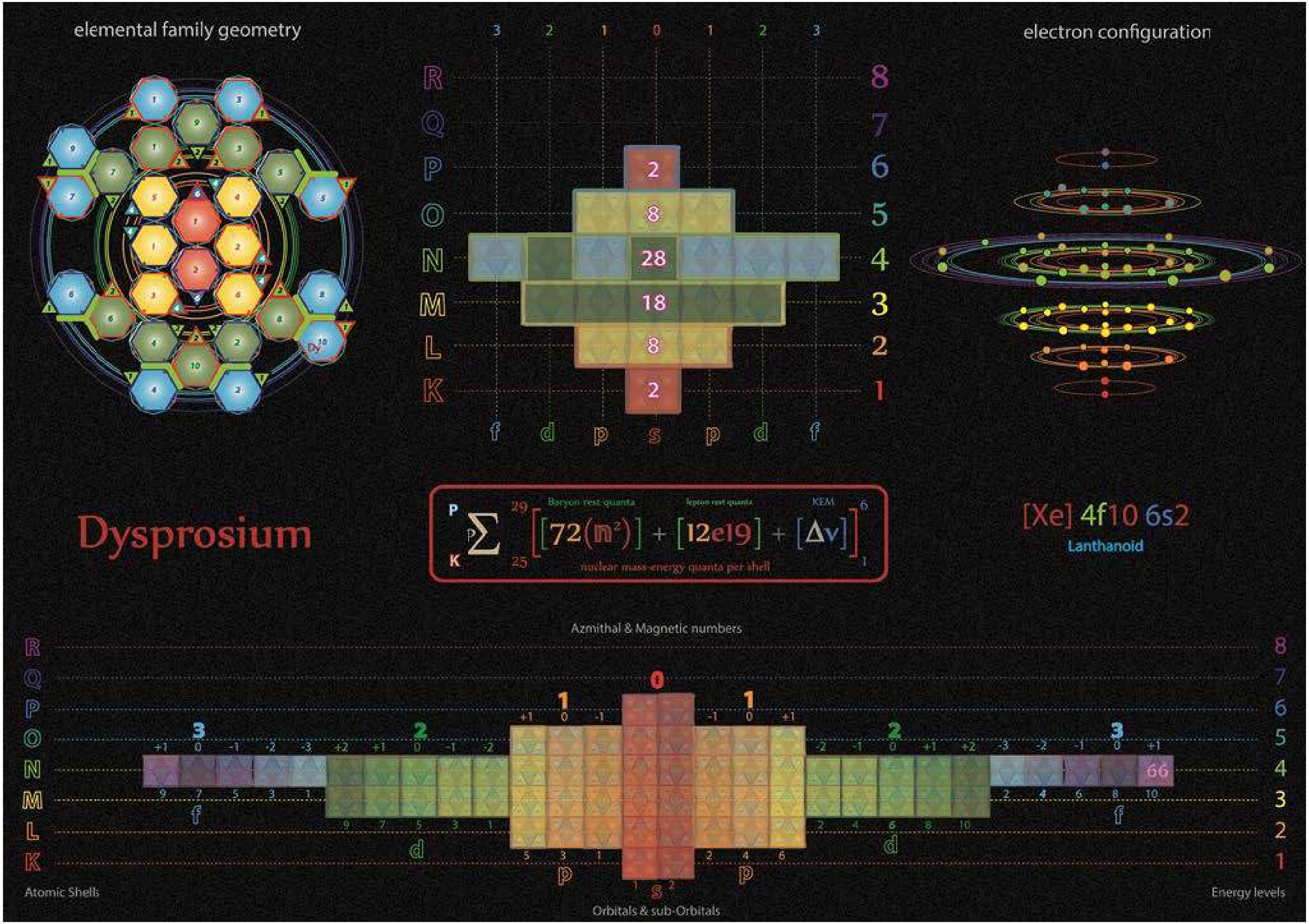
elemental family geometry
Dy
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
28
18
8
2
f
d
p
s
p
d
f
Dysprosium
P
K
Σ
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 4f10 6s2
Lanthanoid
Azmithal & Magnetic numbers
66
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.66 - Dysprosium atomic config

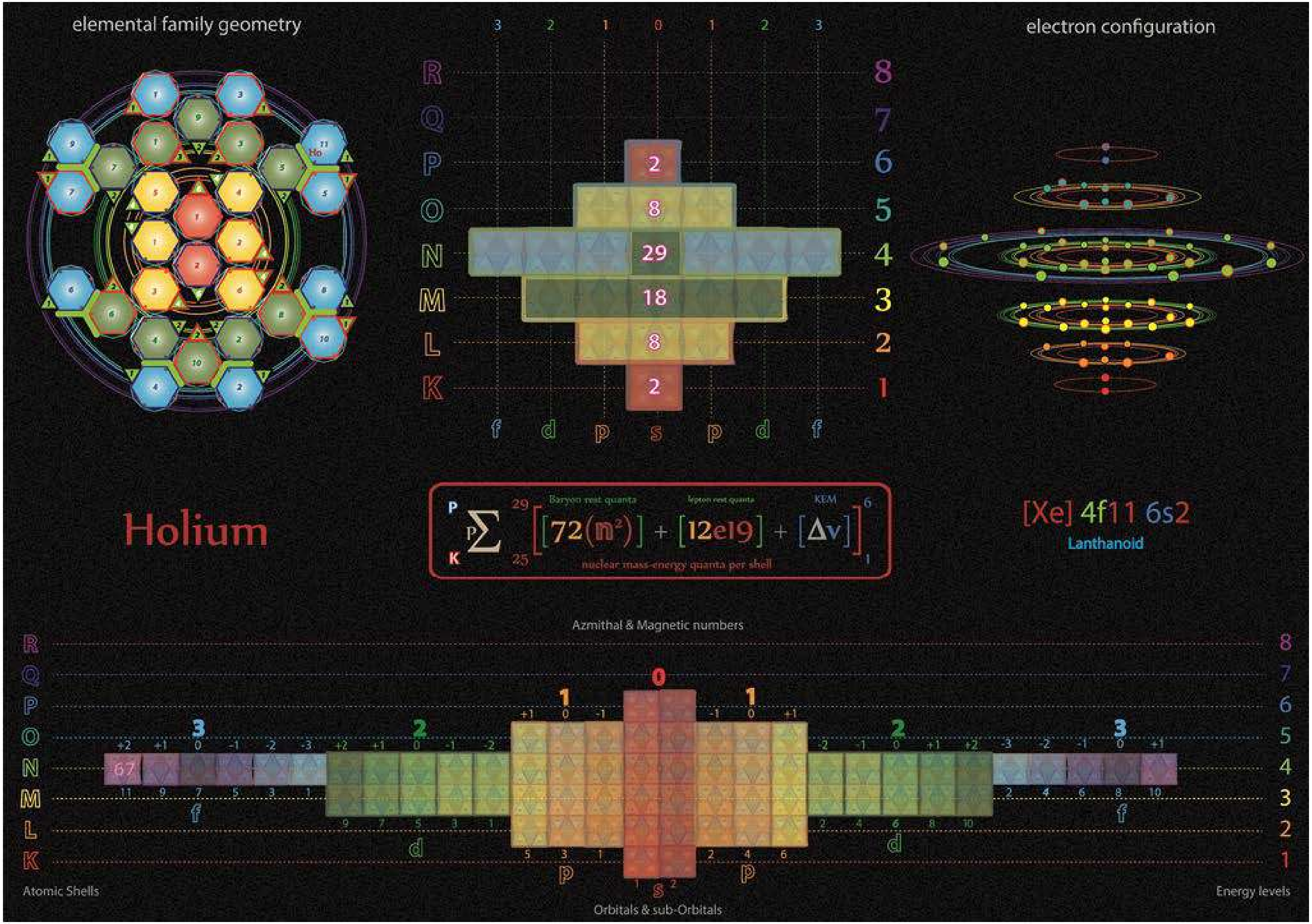
elemental family geometry
Ho
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2 8 29 18 8 2
f d p s p d f
Holium
P K p∑ 29 25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
6 1
nuclear mass-energy quanta per shell
[Xe] 4f11 6s2
Lanthanoid
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
67
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.67 - Holium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

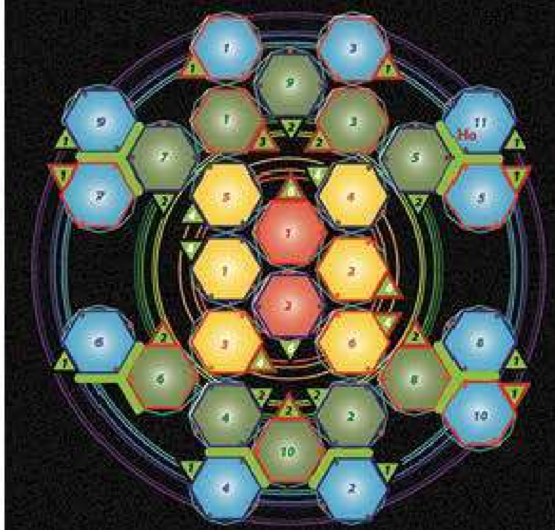

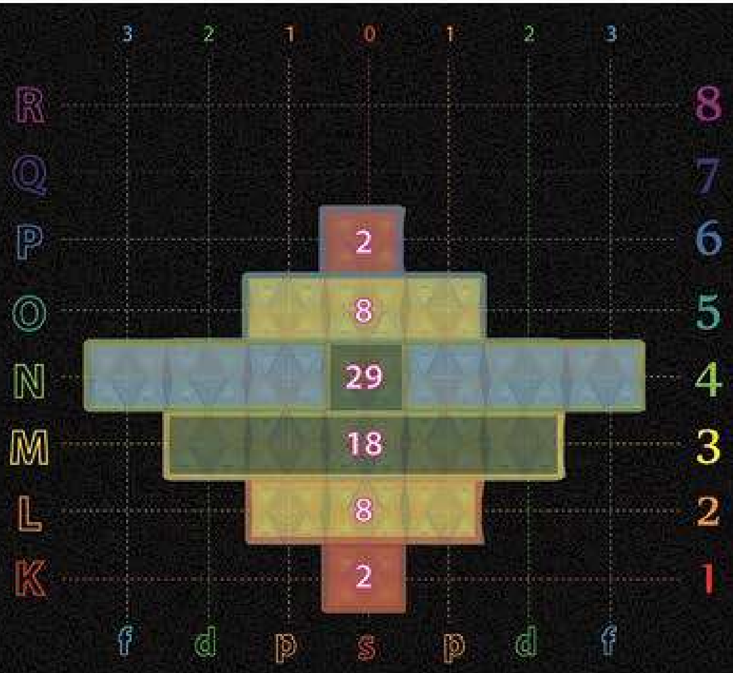
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
29
18
8
2
f
d
p
s
p
d
f

electron configuration

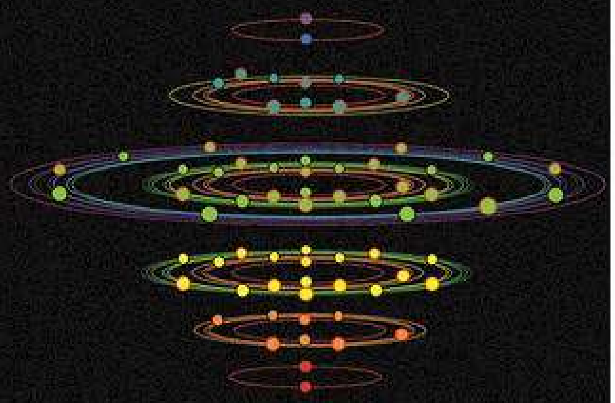

Holmium

$$ {}^{\mathbf{P}}_{\mathbf{K}}\,{}_{p}\sum^{29}_{25}\left[\,[72(n^{2})]+[12e19]+[\Delta v]\,\right]^{6}_{1} $$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Xe] 4f11 6s2

Lanthanoid

Azmithal & Magnetic numbers

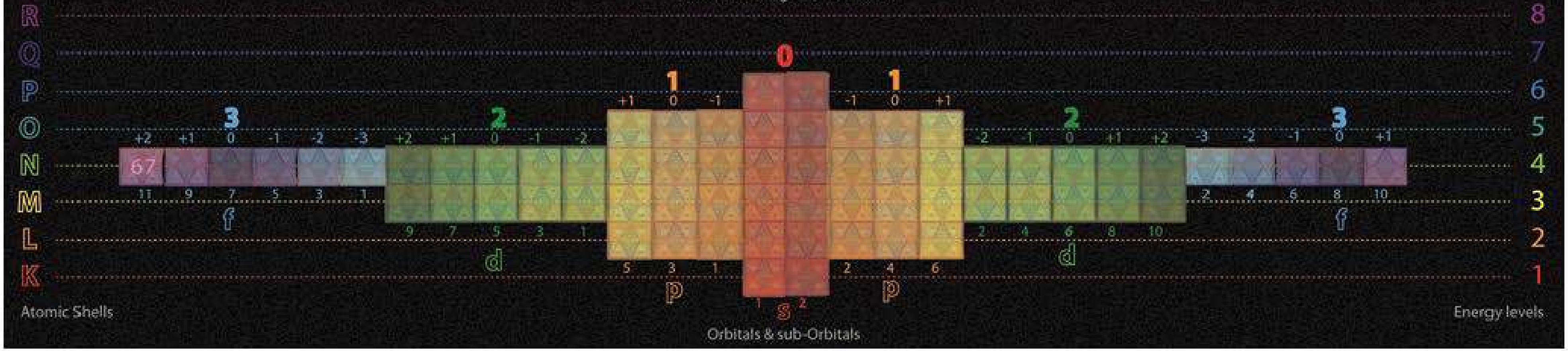
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
-2 -1 0 +1 +2
-3 -2 -1 0 +1
67
11 9 7 5 3 1
9 7 5 3 1
5 3 1
2 4 6
2 4 6 8 10
2 4 6 8 10
f
d
p
s
p
d
f
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.67 - Holmium atomic config

Copyright ABRAHAM [2008] - All rights reserved

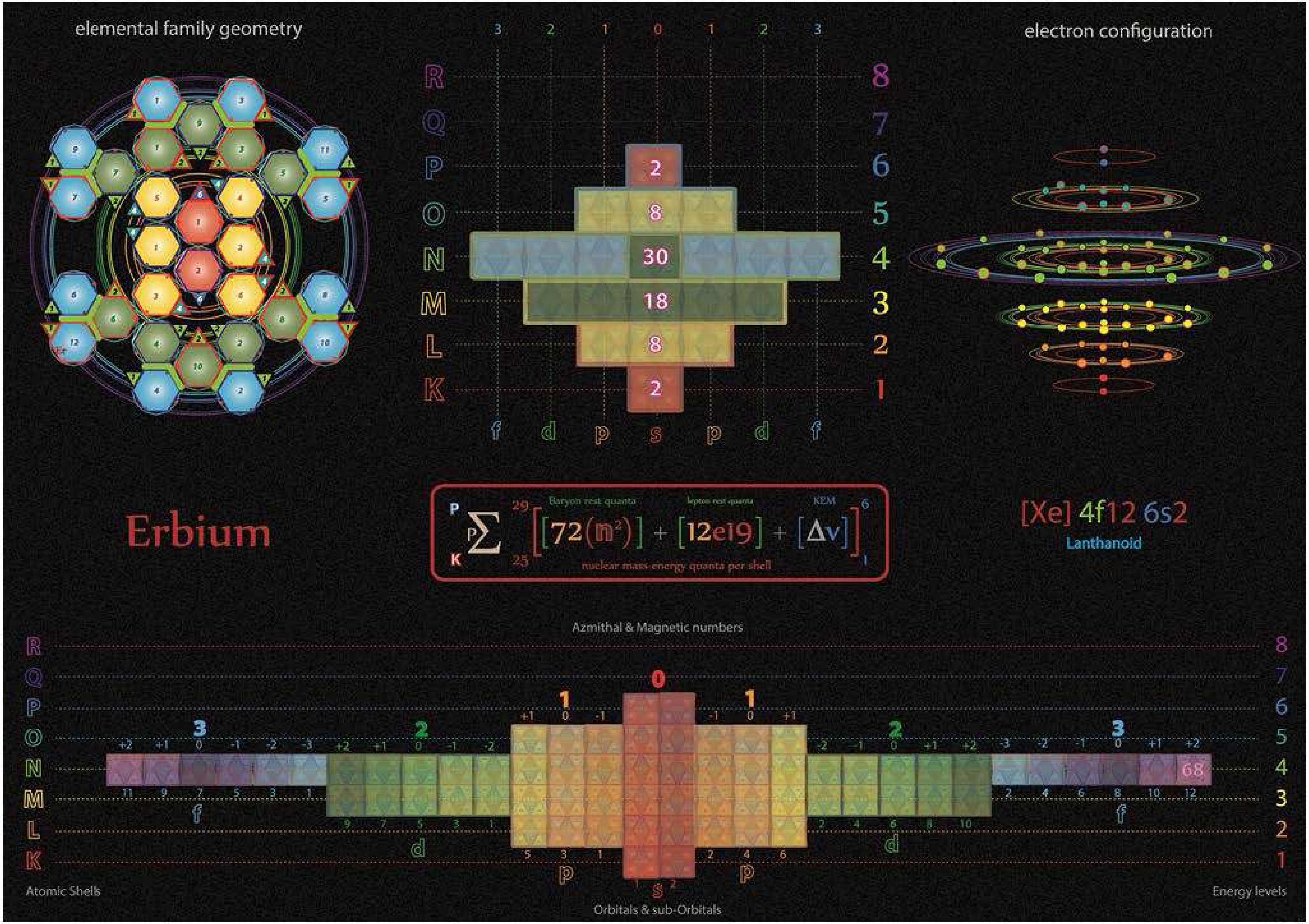
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
30
18
8
2
f d p s p d f
Erbium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 4f12 6s2
Lanthanoid
Azmithal & Magnetic numbers
68
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.68 - Erbium atomic config

Copyright ABRAHAM [2008] - All rights reserved

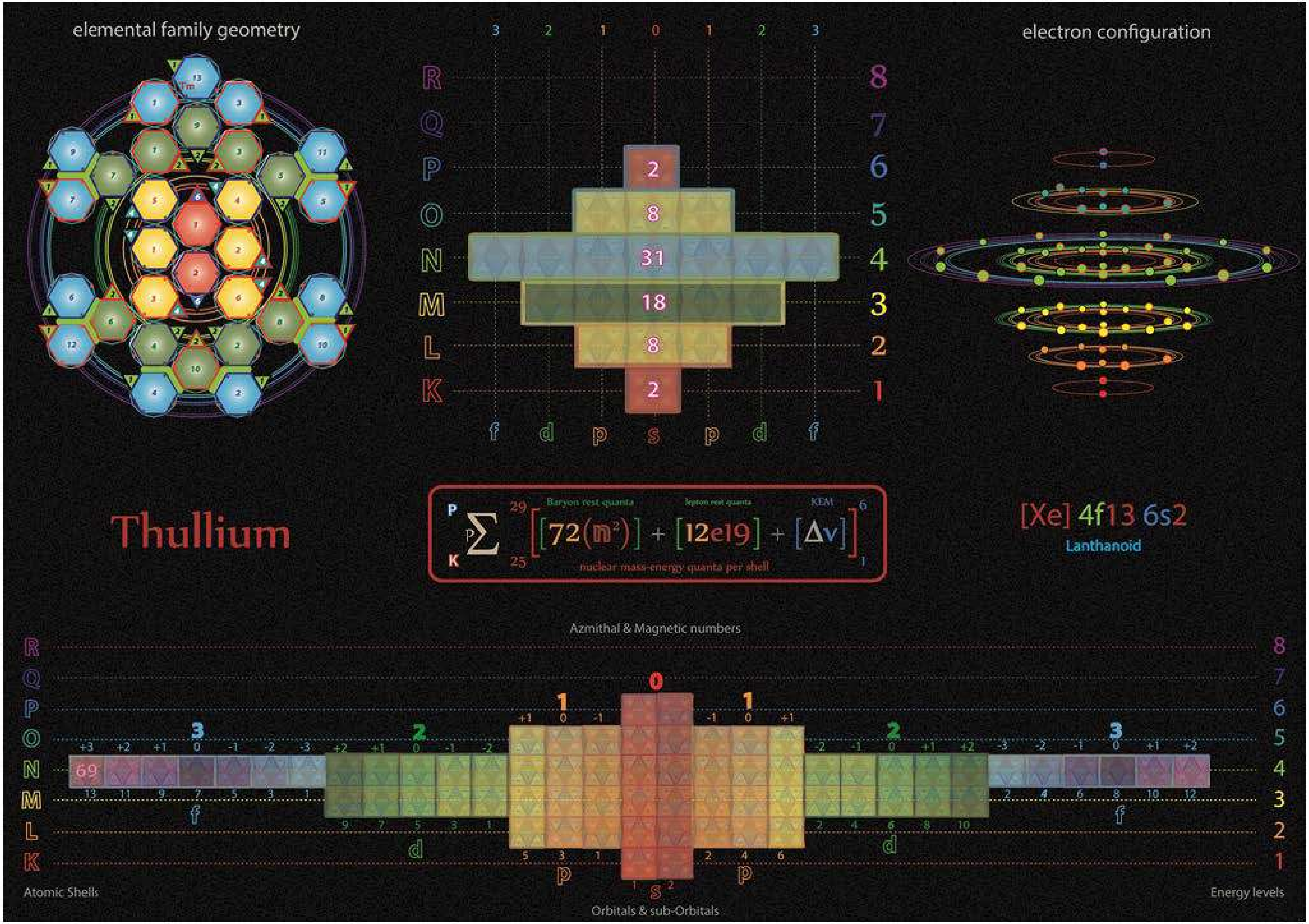
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
31
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
0
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2
3 2 1 0 1 2 3
69
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Thullium

$$\sum_{K}^{P}\ {}^{29}_{25}\left[\left[72(m^2)\right]+\left[12e19\right]+\left[\Delta v\right]\right]^{6}_{1}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Xe] 4f13 6s2

Lanthanoid

Tetryonics 53.69 - Thulium atomic config

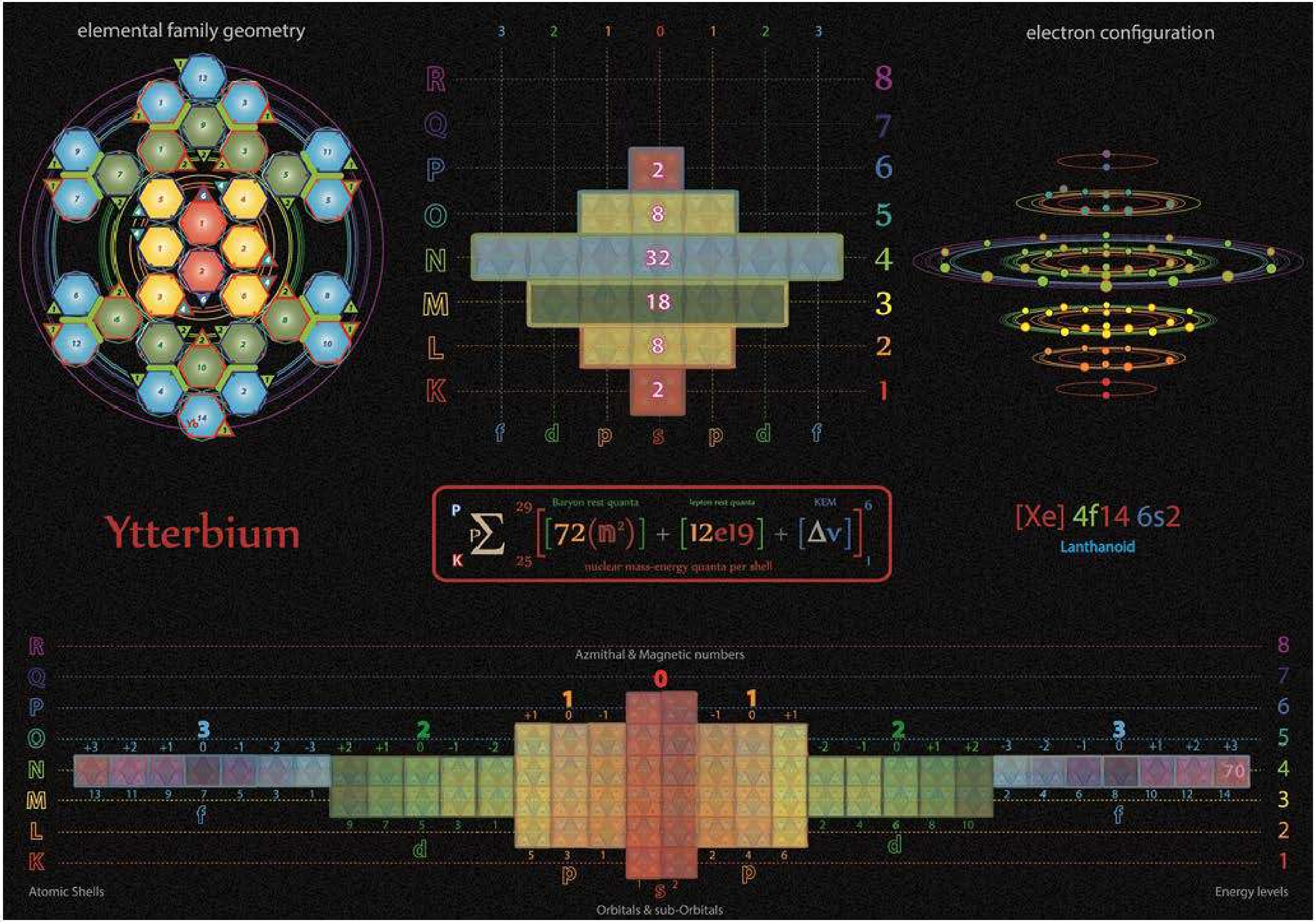
elemental family geometry
Yb
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
32
18
8
2
f d p s p d f
Ytterbium
P
K
p
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 4f14 6s2
Lanthanoid
Azmithal & Magnetic numbers
0
1
2
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
2 4 6 8 10 12 14
70
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 s 2
f d p p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.70 - Ytterbium atomic config

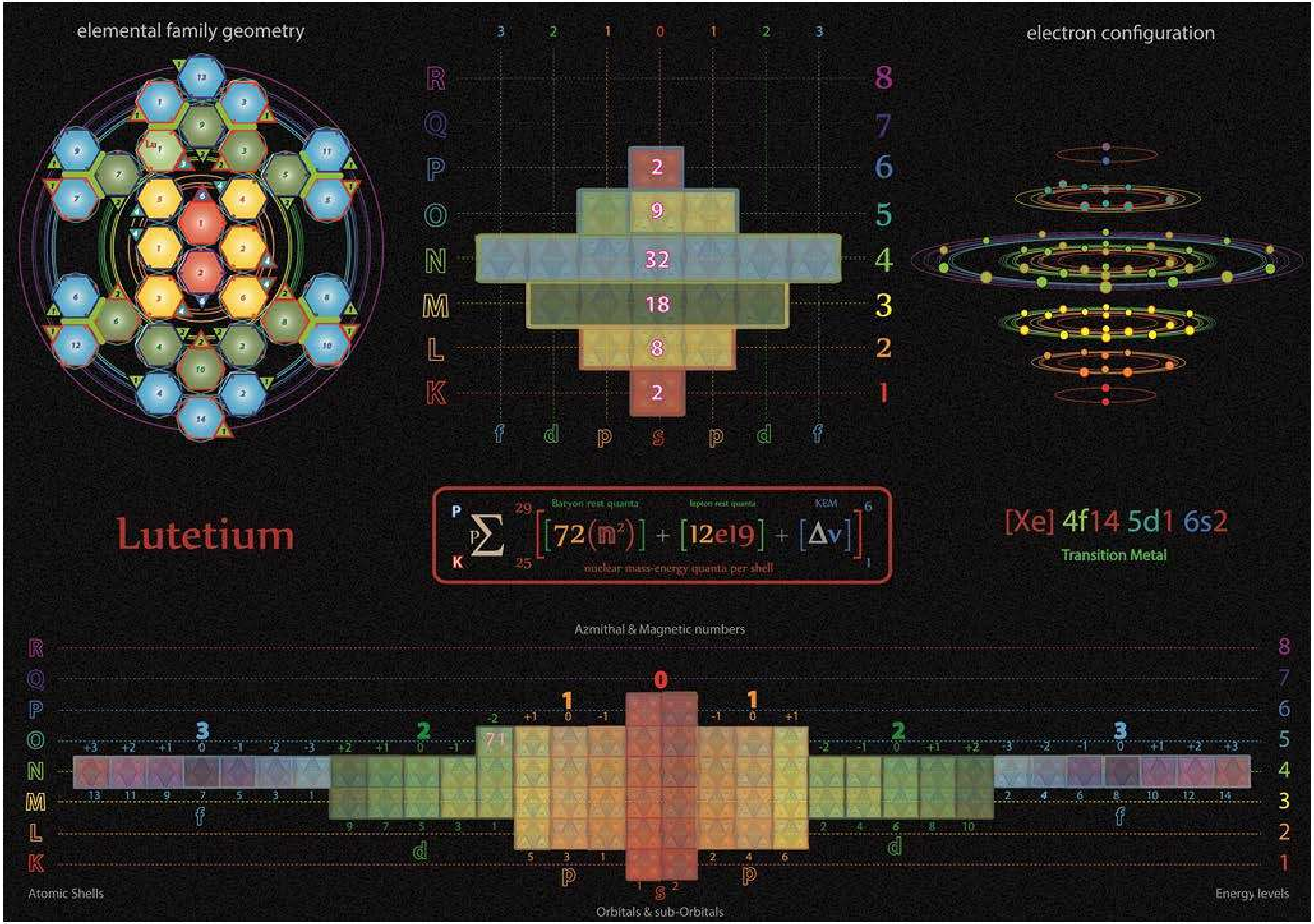
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2 9 32 18 8 2
f d p s p d f
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Lutetium

$$\sum_{K}^{P} p \left[[72(n^2)] + [12e19] + [\Delta v] \right]_{1}^{6} \quad {}^{29}_{25}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Xe] 4f14 5d1 6s2

Transition Metal

Tetryonics 53.71 - Lutetium atomic config

Copyright ABRAHAM [2008] - All rights reserved

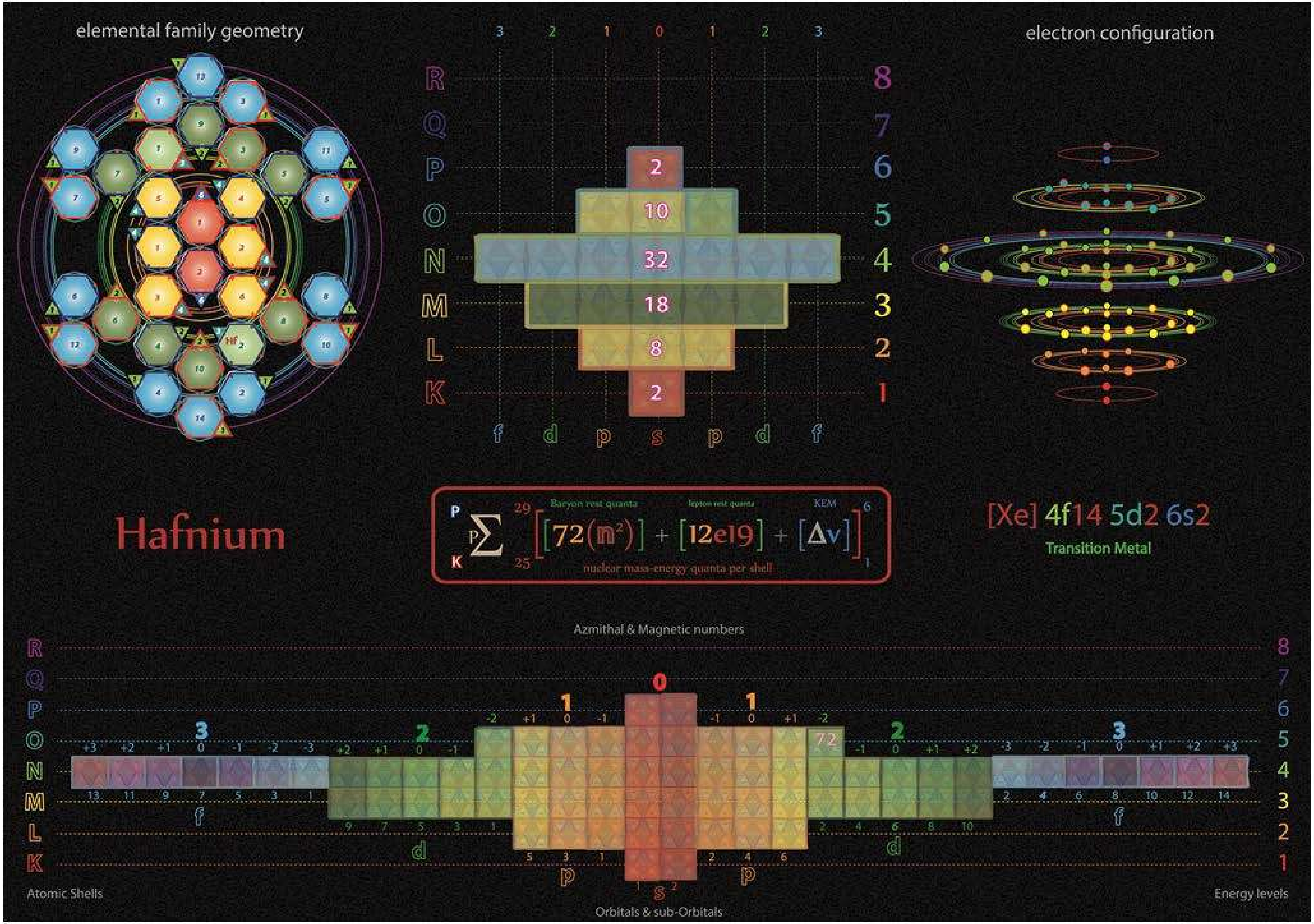
elemental family geometry
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
10
32
18
8
2
f
d
p
s
p
d
f
Hafnium
Baryon rest quanta
lepton rest quanta
KEM
P
K
29
25
6
1
$\sum [72(m^2)] + [12e19] + [\Delta v]$
nuclear mass-energy quanta per shell
[Xe] 4f14 5d2 6s2
Transition Metal
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.72 - Hafnium atomic config

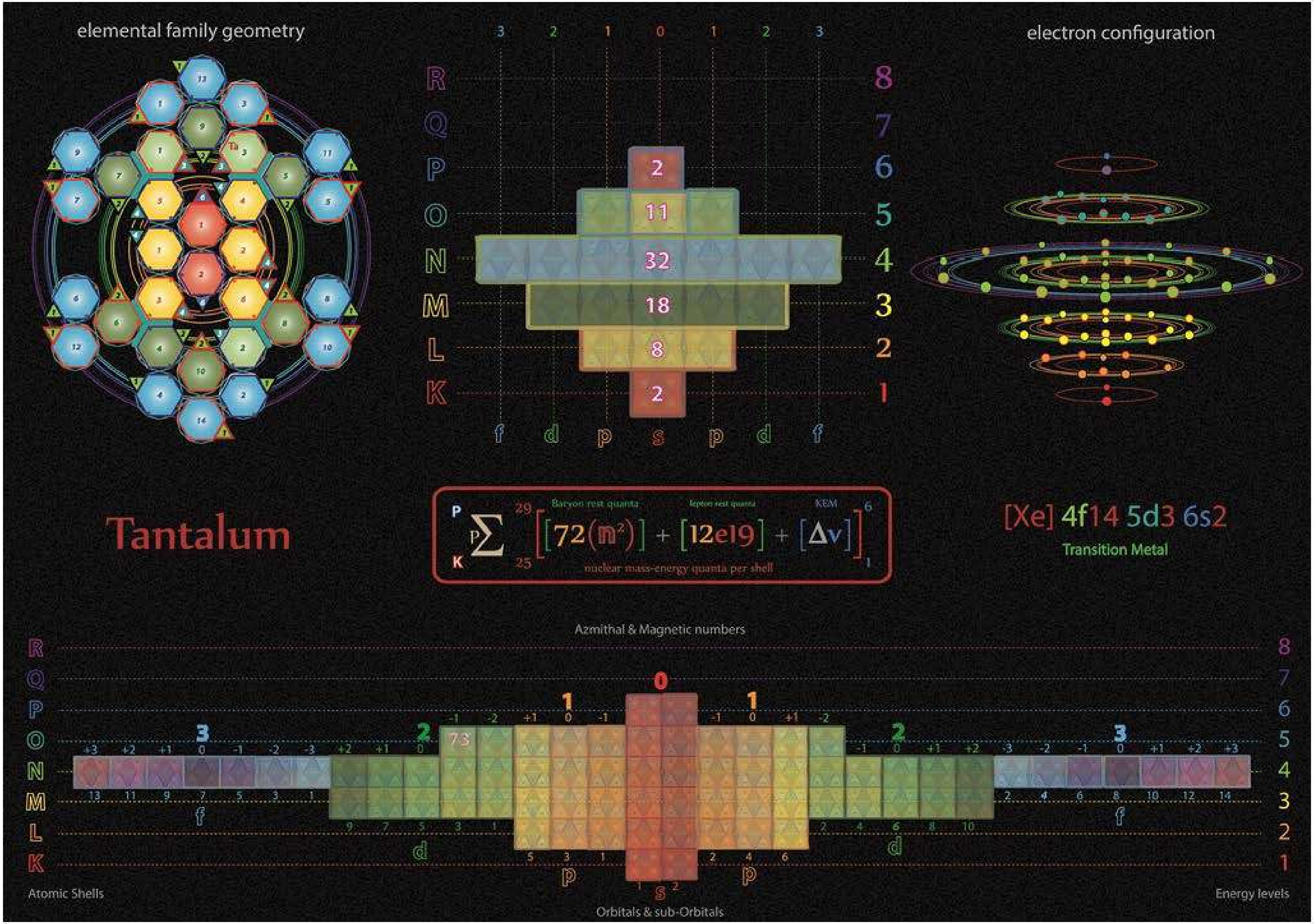
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
11
32
18
8
2
f d p s p d f
Tantalum
[Xe] 4f14 5d3 6s2
Transition Metal
Azmithal & Magnetic numbers
73
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$ {}_{K}^{P}\,{}_{p}\sum_{25}^{29}\left[\left[72(\mathrm{m}^2)\right]+\left[12e19\right]+\left[\Delta v\right]\right]_{1}^{6} $$

Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell

Tetryonics 53.73 - Tantalum atomic config

Copyright ABRAHAM [2008] - All rights reserved

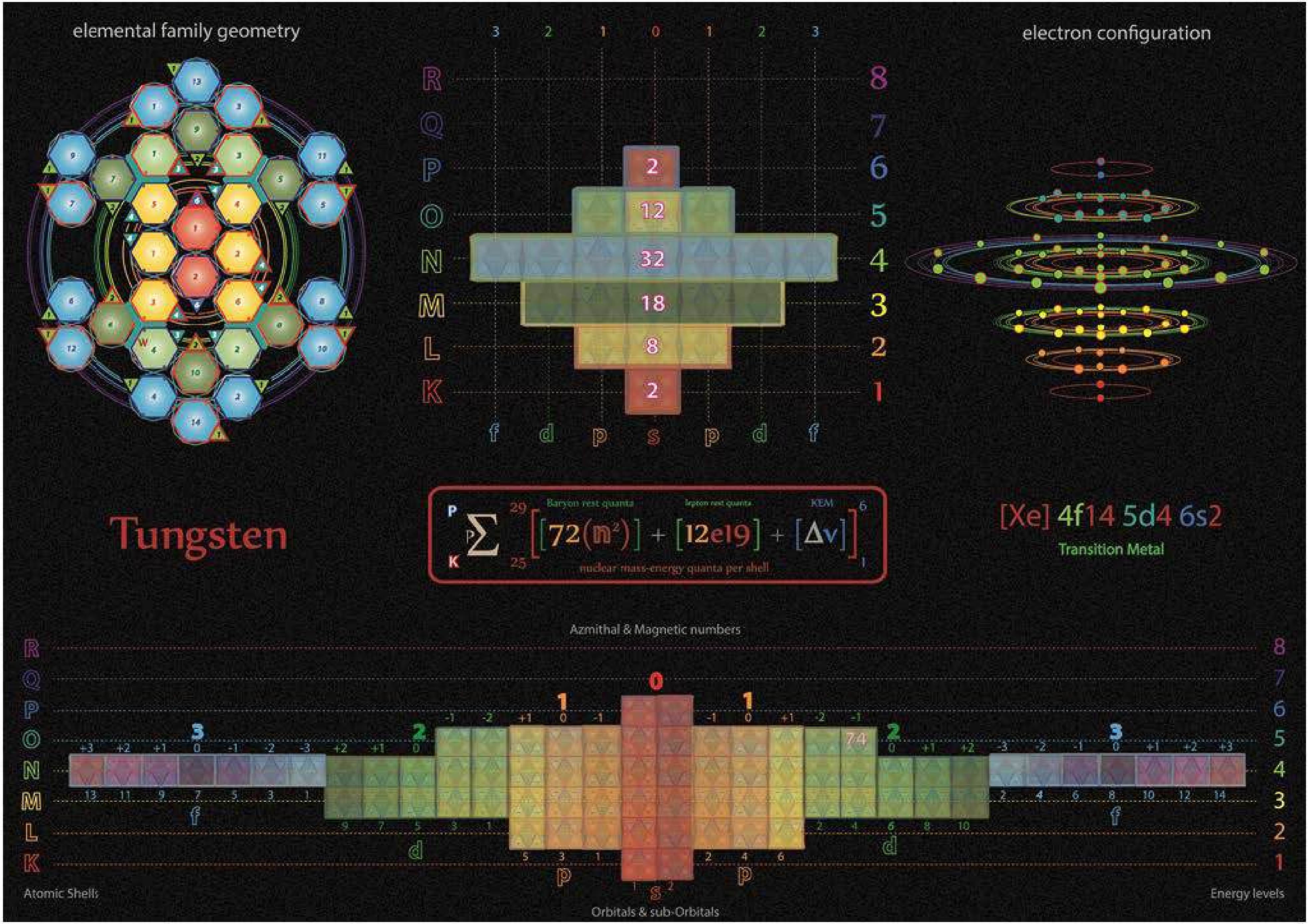
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2
12
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals
74

Tungsten

$$\sum_{K}^{P}{}_{25}^{29}\left[\,[72(m^2)] + [12e19] + [\Delta v]\,\right]_1^6$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Xe] 4f14 5d4 6s2

Transition Metal

Tetryonics 53.74 - Tungsten atomic config

Copyright ABRAHAM [2008] - All rights reserved

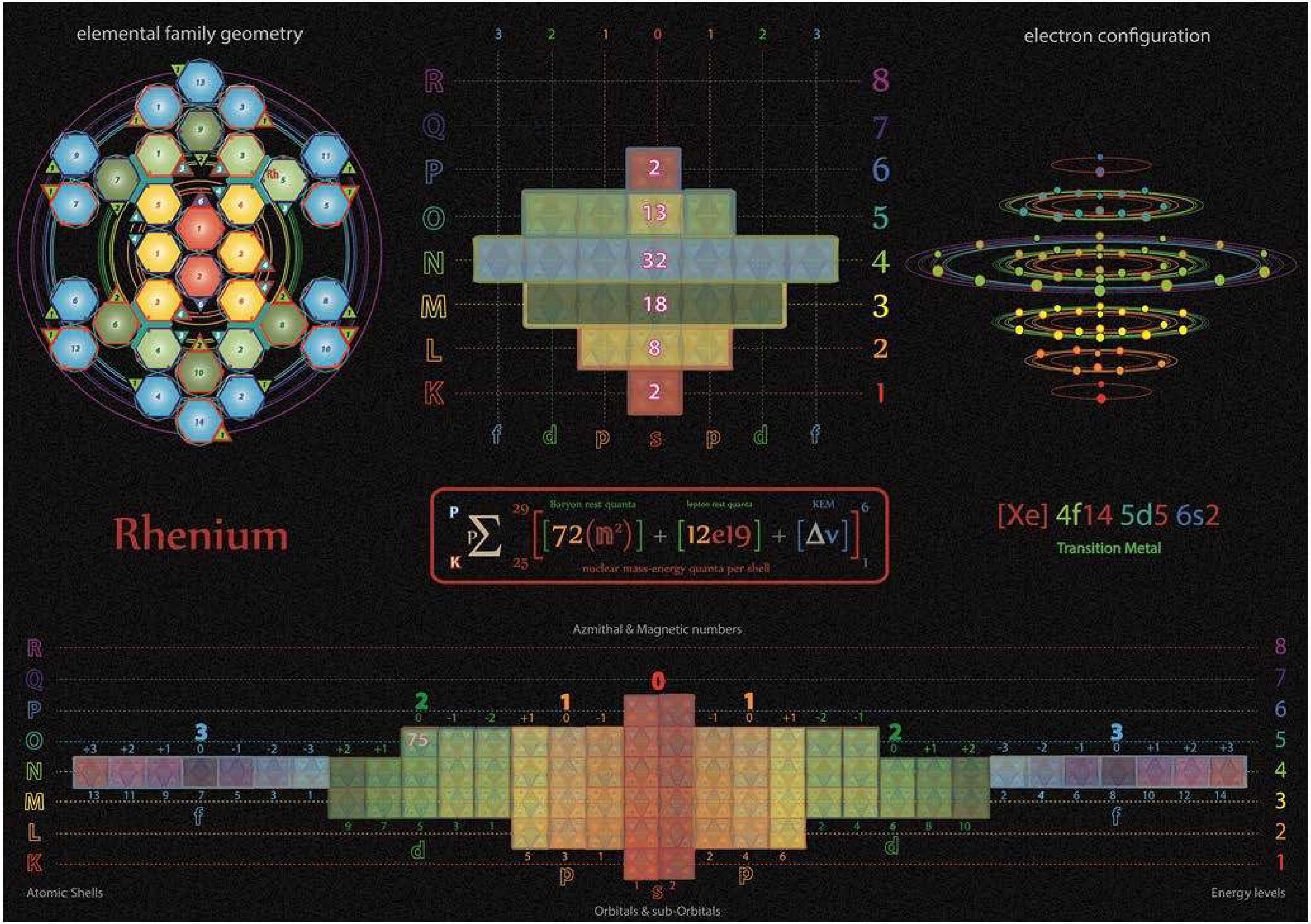
elemental family geometry
Rh
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
2
13
32
18
8
2
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
75
f
d
p
s
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Rhenium

$$\sum_{K}^{P} {}^{p}\sum_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_{1}^{6}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Xe] 4f14 5d5 6s2

Transition Metal

Tetryonics 53.75 - Rhenium atomic config

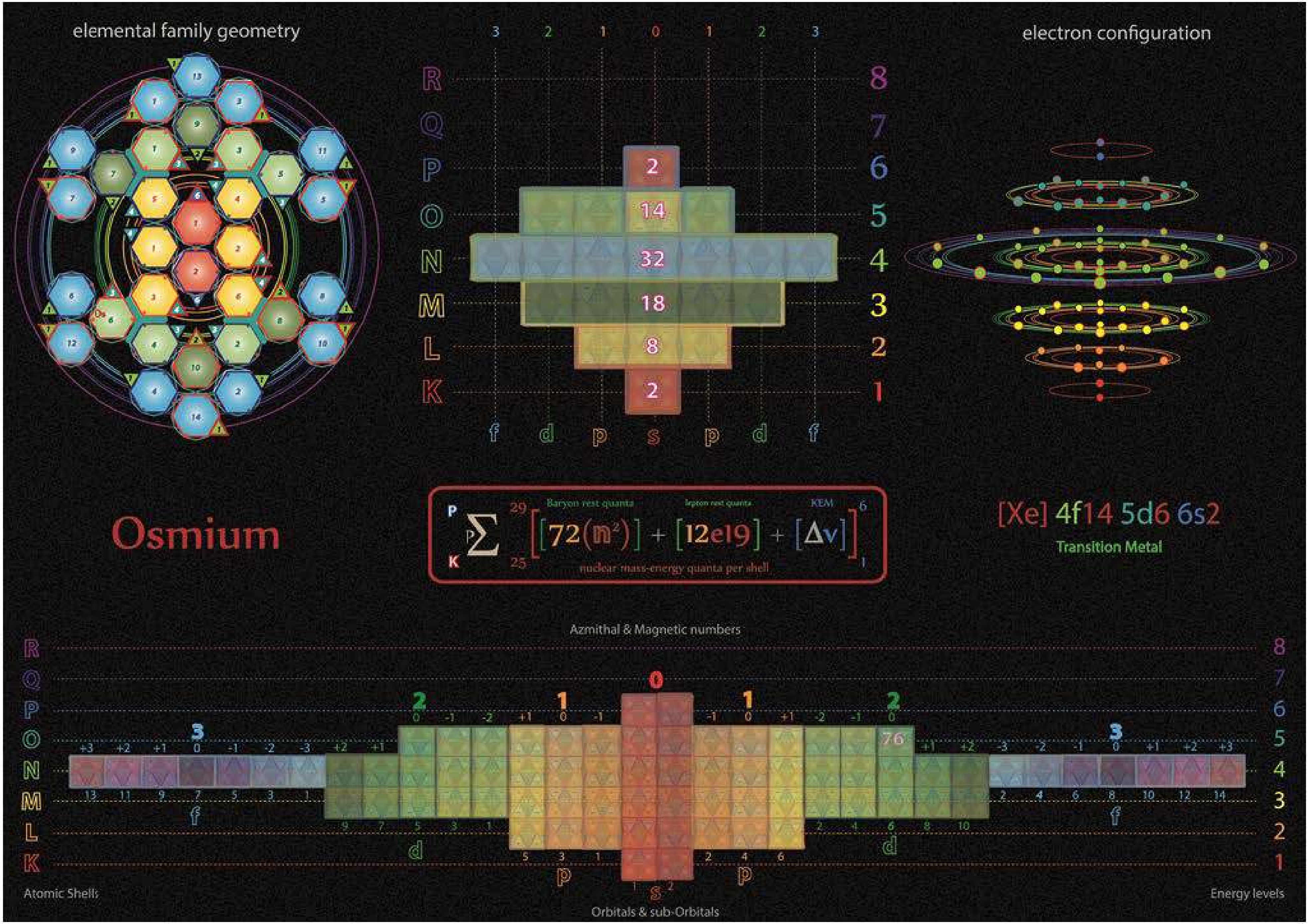
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
14
32
18
8
2
f d p s p d f
Osmium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
6
1
$\sum_{K}^{P} \left[[72(n^2)] + [12e19] + [\Delta v] \right]$
nuclear mass-energy quanta per shell
[Xe] 4f14 5d6 6s2
Transition Metal
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
+3 +2 +1 0 -1 -2 -3 +2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2 -3 -2 -1 0 +1 +2 +3
76
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 3 1
p
1 s 2
2 4 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.76 - Osmium atomic config

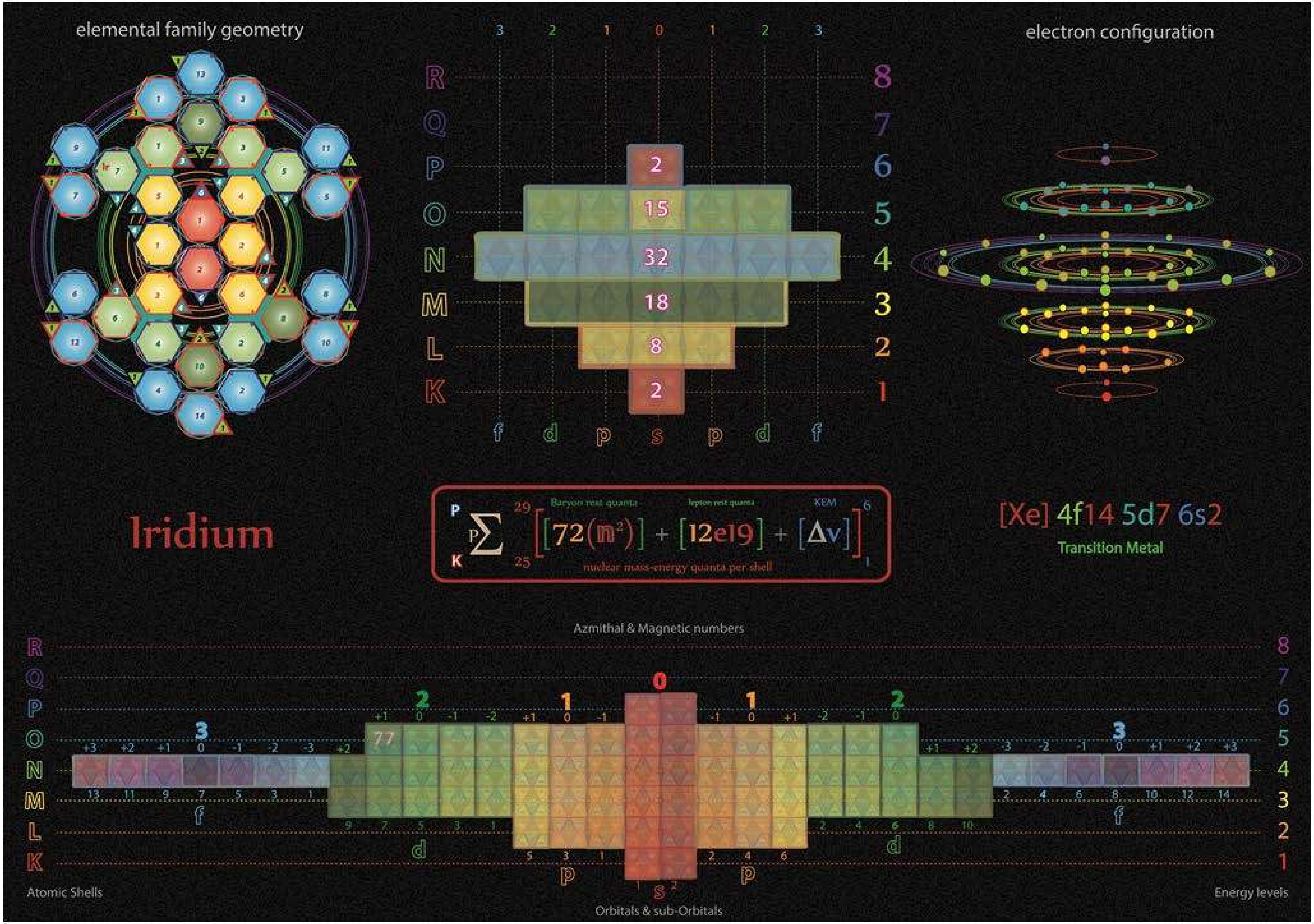
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
15
32
18
8
2
f d p s p d f
Iridium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
6
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Xe] 4f14 5d7 6s2
Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
77
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.77 - Iridium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

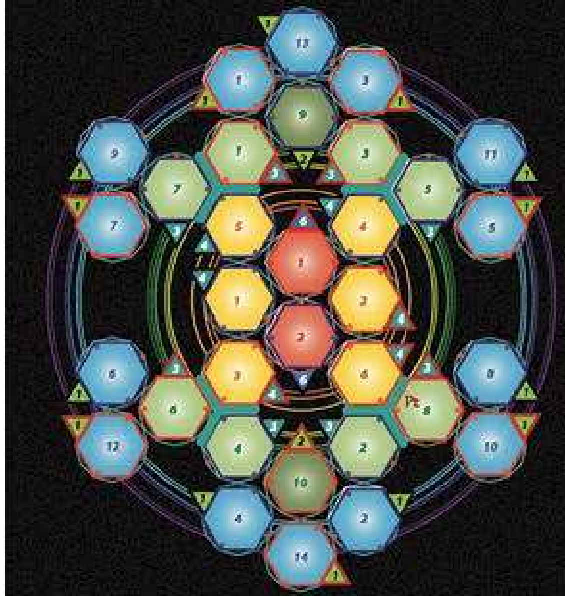
13
9
11
7
5
1
2
3
4
6
8
10
12
14
Pt

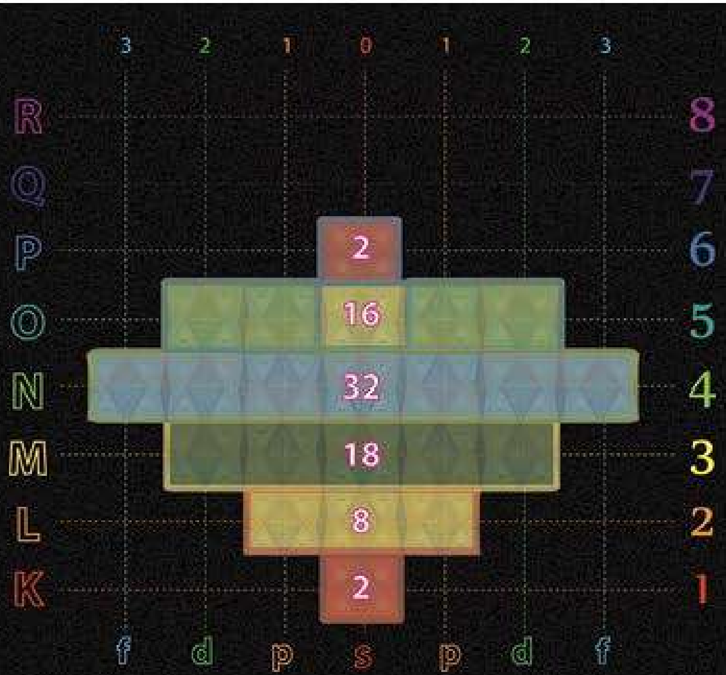
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
16
32
18
8
2
f
d
p
s
p
d
f

electron configuration

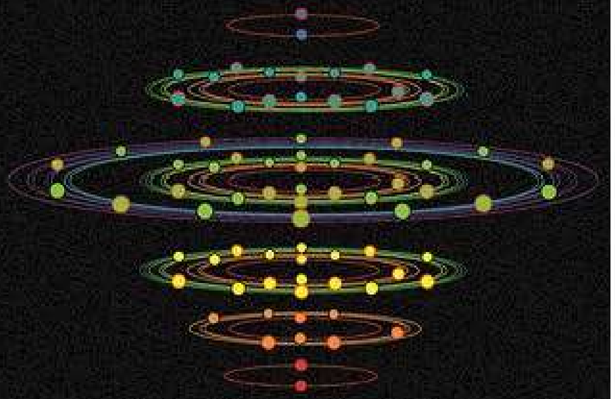

Platinum

$$\sum_{K}^{P} {}_{P}\, {}^{29}_{25}\Big[\underset{\text{}}{\overset{\text{Baryon rest quanta}}{[72(m^2)]}} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \Big]^{6}_{1}$$

nuclear mass-energy quanta per shell

[Xe] 4f14 5d8 6s2

Transition Metal

Azmithal & Magnetic numbers

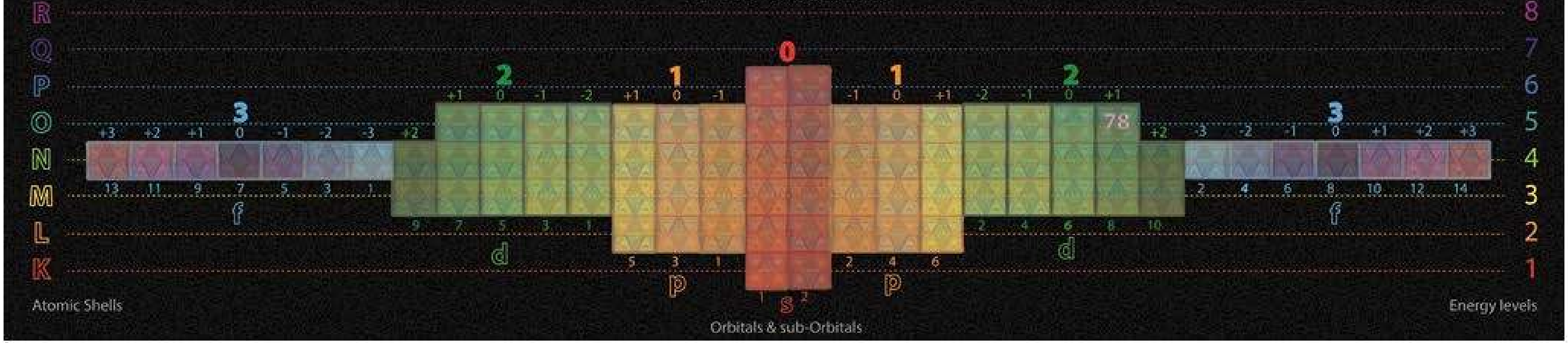
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
+3
+2
+1
0
-1
-2
-3
78
f
d
p
s
p
d
f
1
2

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.78 - Platinum atomic config

Copyright ABRAHAM [2008] - All rights reserved

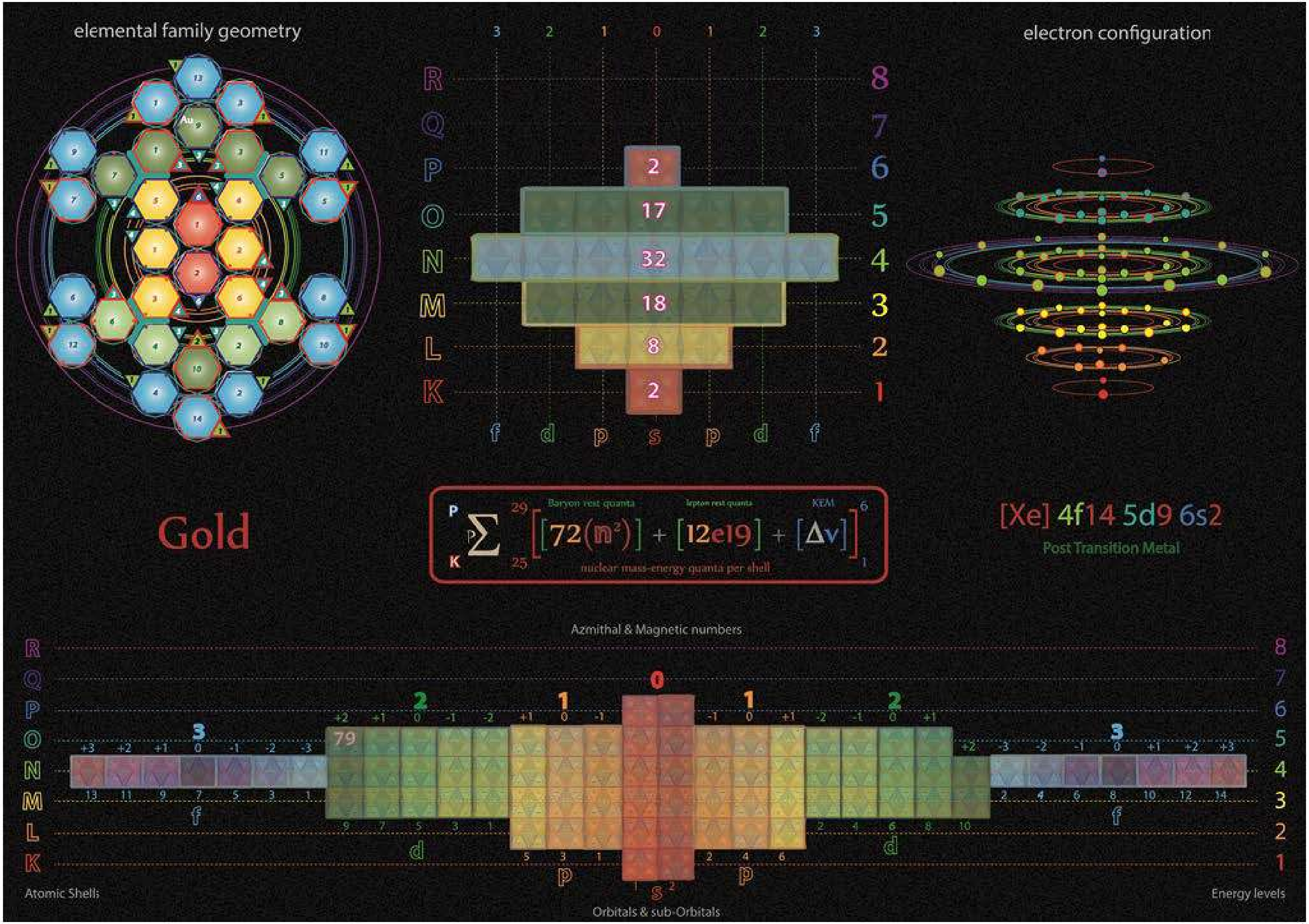
elemental family geometry
Au
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
17
32
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
79
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Gold

$$\sum_{K}^{P} {}_{25}^{29}\left[[72(n^2)] + [12e19] + [\Delta v] \right]_{1}^{6}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Xe] 4f14 5d9 6s2

Post Transition Metal

Tetryonics 53.79 - Gold atomic config

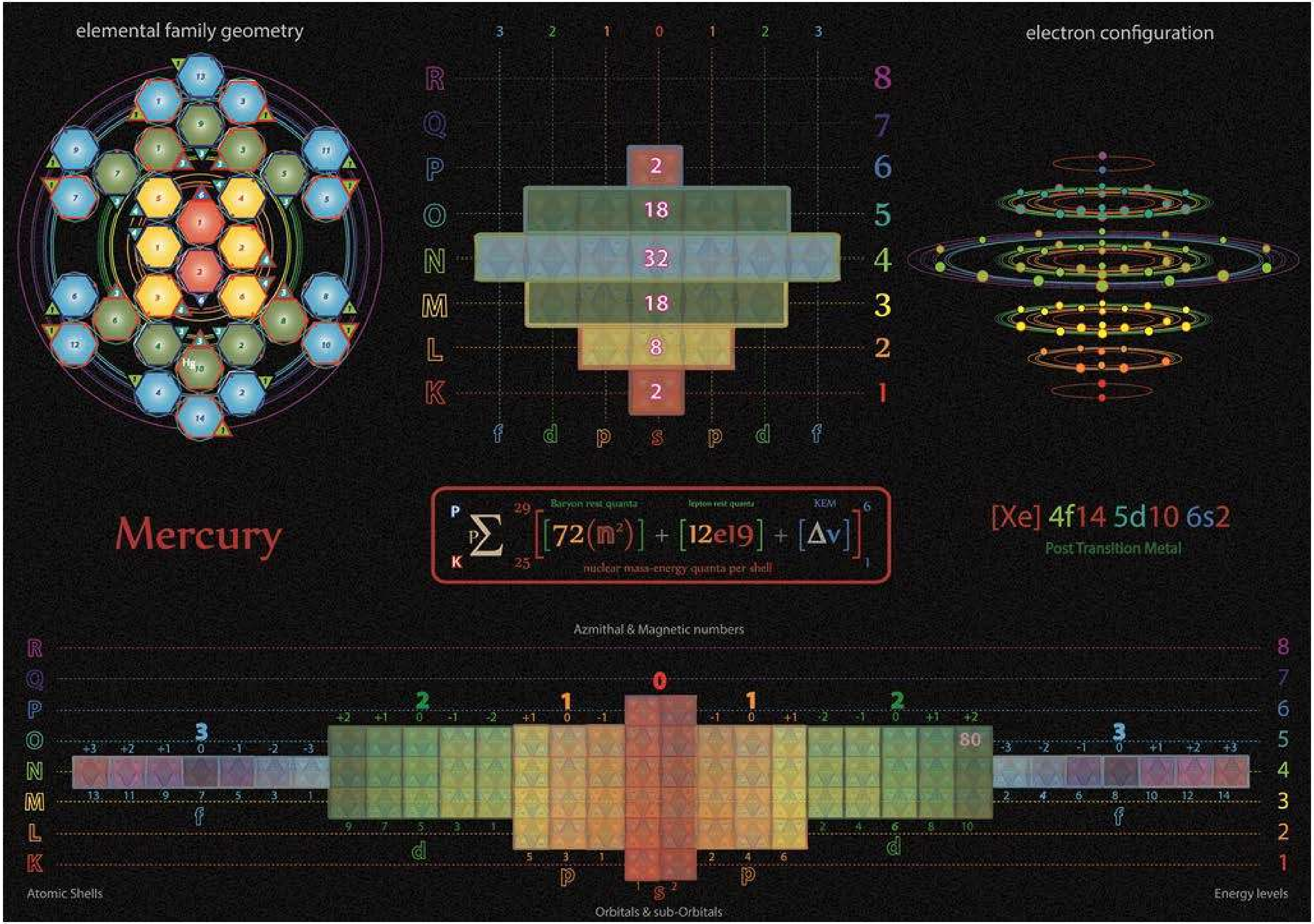
elemental family geometry
Hg
3 2 1 0 1 2 3
R Q P O N M L K
2
18
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Mercury
P
K
$\sum_{25}^{29}$
Baryon rest quanta
lepton rest quanta
KEM
$[72(n^2)] + [12e19] + [\Delta v]$
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2
Post Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
80
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f
d
p
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.80 - Mercury atomic config

Copyright ABRAHAM [2008] - All rights reserved

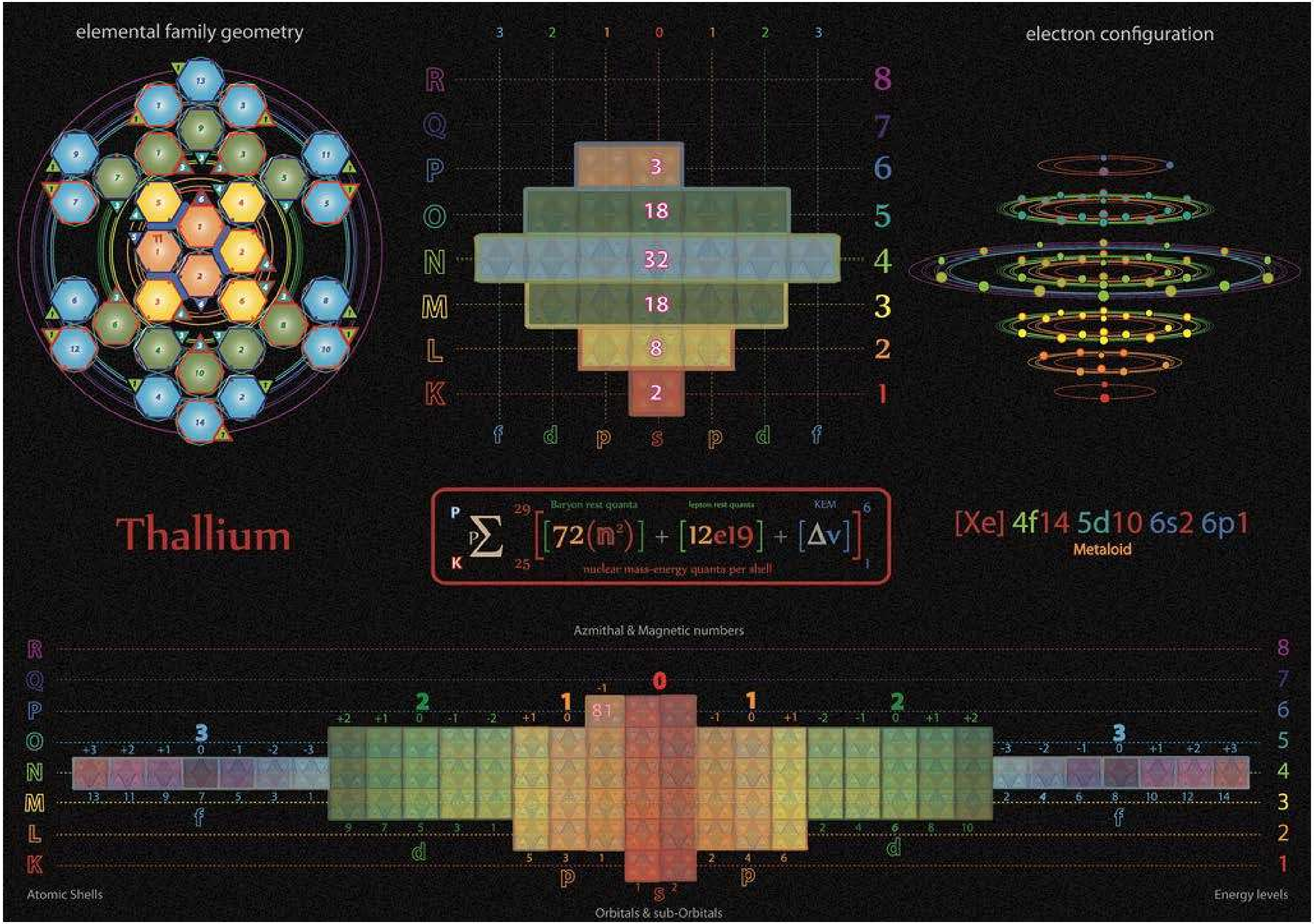
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
18
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals
81
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell

Thallium

$$ {}^{P}_{K}\,{}_{p}\sum_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_{1}^{6} $$

[Xe] 4f14 5d10 6s2 6p1

Metaloid

Tetryonics 53.81 - Thallium atomic config

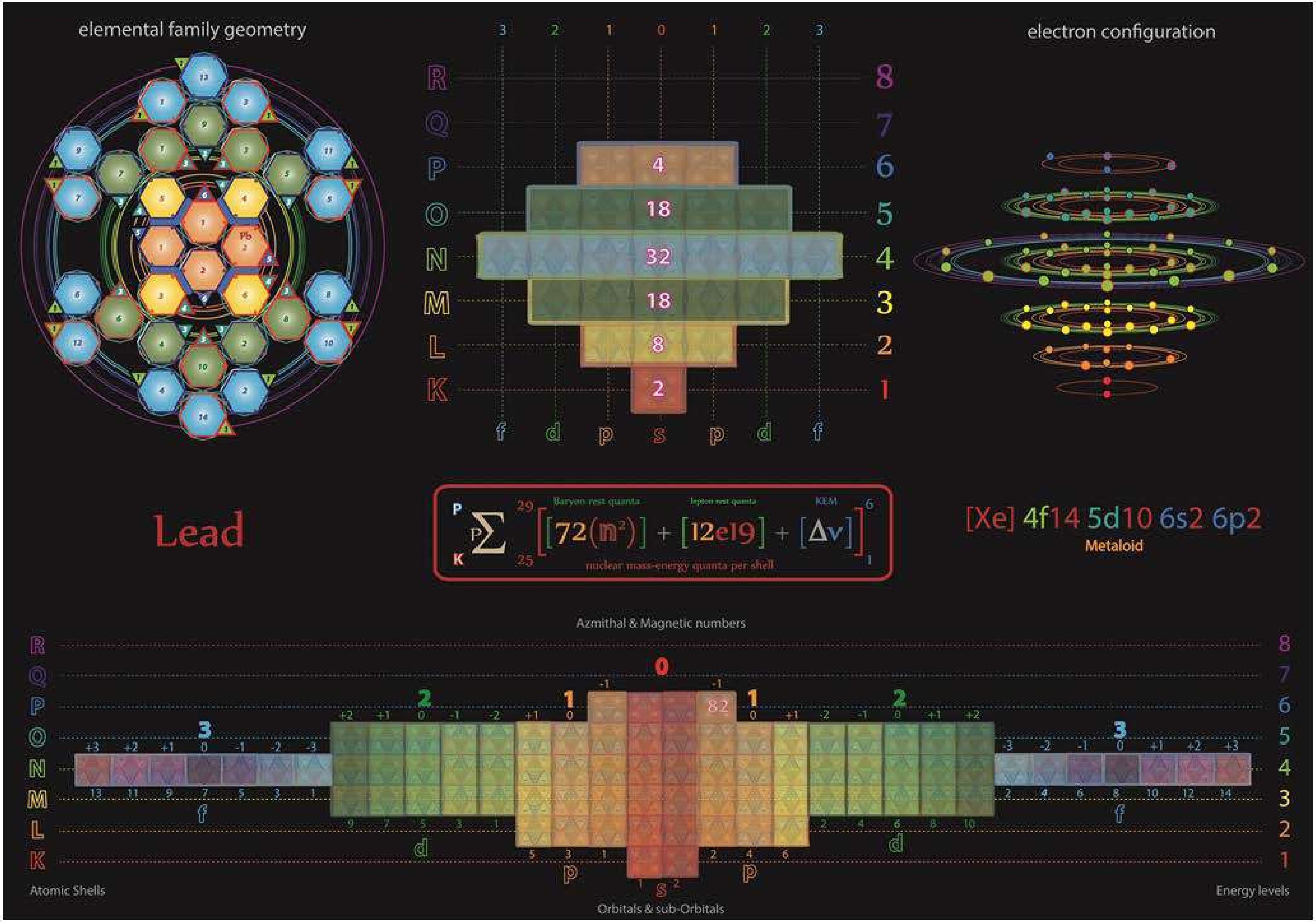
elemental family geometry
Pb
3 2 1 0 1 2 3
R Q P O N M L K
4
18
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
-1 -1
82
+2 +1 0 -1 -2 +1 0
0 +1 -2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Lead

$$ {}^{P}_{K}\,{}_{P}\!\sum\nolimits_{25}^{29}\left[\overset{\text{Baryon rest quanta}}{\left[72(n^2)\right]}+\overset{\text{lepton rest quanta}}{\left[12e19\right]}+\overset{\text{KEM}}{\left[\Delta v\right]}\right]_{1}^{6} $$

nuclear mass-energy quanta per shell

[Xe] 4f14 5d10 6s2 6p2

Metaloid

Tetryonics 53.82 - Lead atomic config

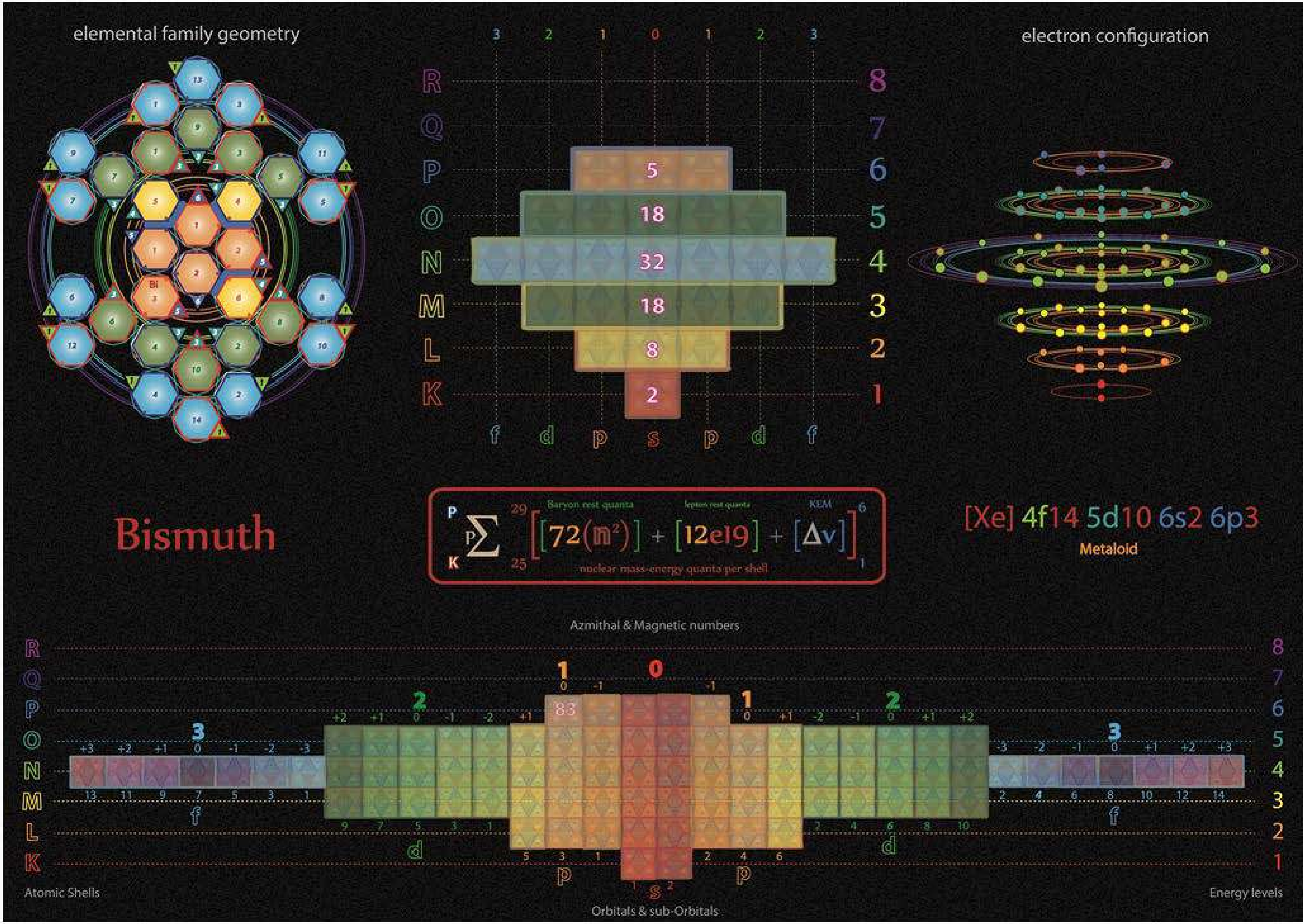
elemental family geometry
Bi
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
5
18
32
18
8
2
f d p s p d f
electron configuration
Bismuth
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{P} {}^{29}_{25}\left[[72(m^2)] + [12e19] + [\Delta v]\right]^{6}_{1}$
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2 6p3
Metaloid
Azmithal & Magnetic numbers
83
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.83 - Bismuth atomic config

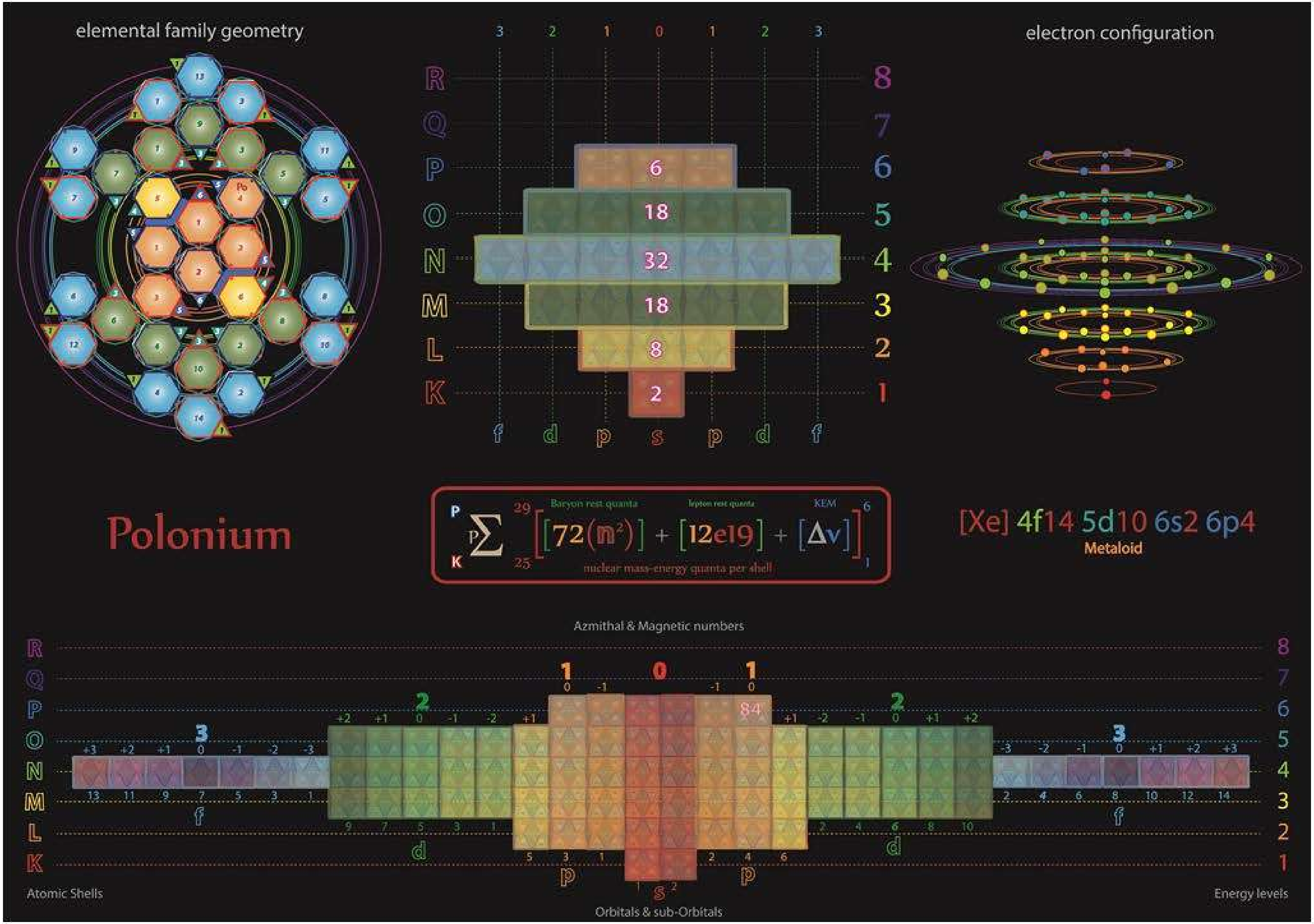
elemental family geometry
Po
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
18
32
18
8
2
f d p s p d f
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
Metaloid
Azmithal & Magnetic numbers
84
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Polonium

$$\sum_{K}^{P} {}_{P}\sum_{25}^{29}\left[[72(n^2)] + [12e19] + [\Delta v]\right]_1^6$$

[Xe] 4f14 5d10 6s2 6p4

Metaloid

Tetryonics 53.84 - Polonium atomic config

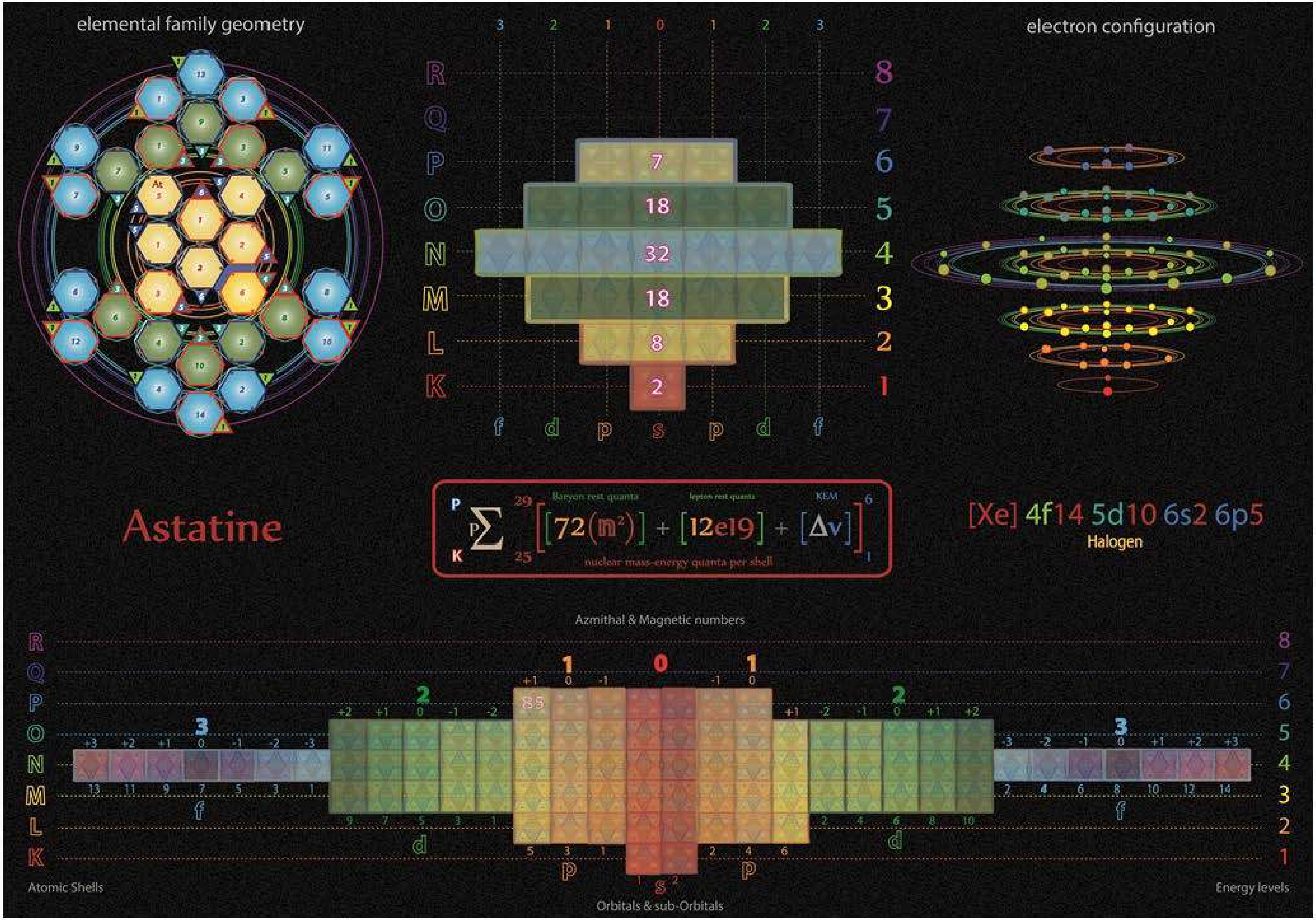
elemental family geometry
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7
18
32
18
8
2
f d p s p d f
Astatine
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{P} \left[[72(\mathrm{m}^2)] + [12e19] + [\Delta v] \right]_{1}^{6}$
29
25
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2 6p5
Halogen
Azmithal & Magnetic numbers
85
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.85 - Astatine atomic config

Copyright ABRAHAM [2008] - All rights reserved

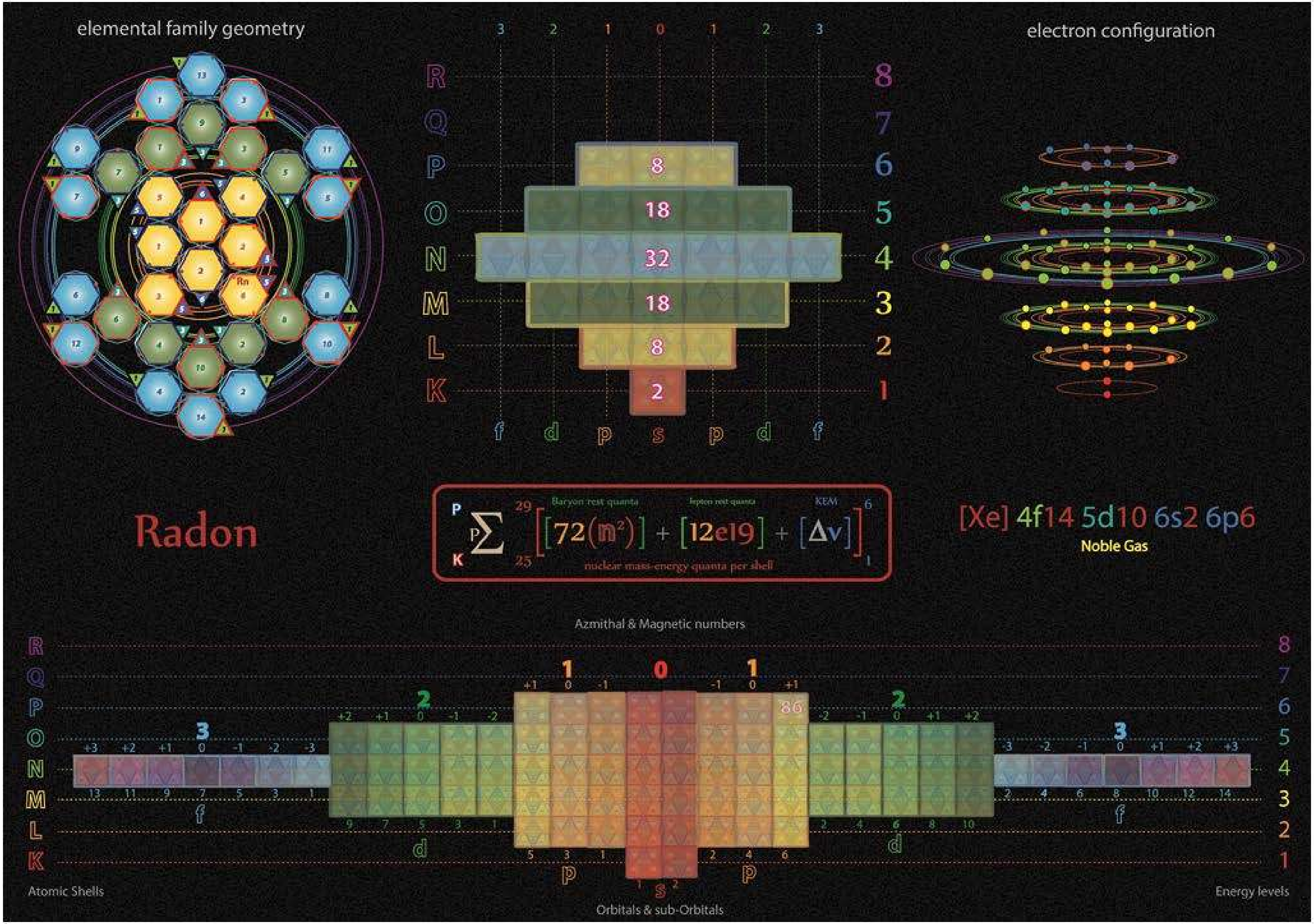
elemental family geometry
Rn
3 2 1 0 1 2 3
R Q P O N M L K
8
18
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
1 0 1
2 2
3 3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
86
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Radon

$$\sum_{K}^{P} {}_{P}\sum_{25}^{29}\left[\underbrace{\overset{\text{Baryon rest quanta}}{\left[72(n^2)\right]} + \overset{\text{lepton rest quanta}}{\left[12e19\right]} + \overset{\text{KEM}}{\left[\Delta v\right]}}_{\text{nuclear mass-energy quanta per shell}}\right]_{1}^{6}$$

[Xe] 4f14 5d10 6s2 6p6

Noble Gas

Tetryonics 53.86 - Radon atomic config

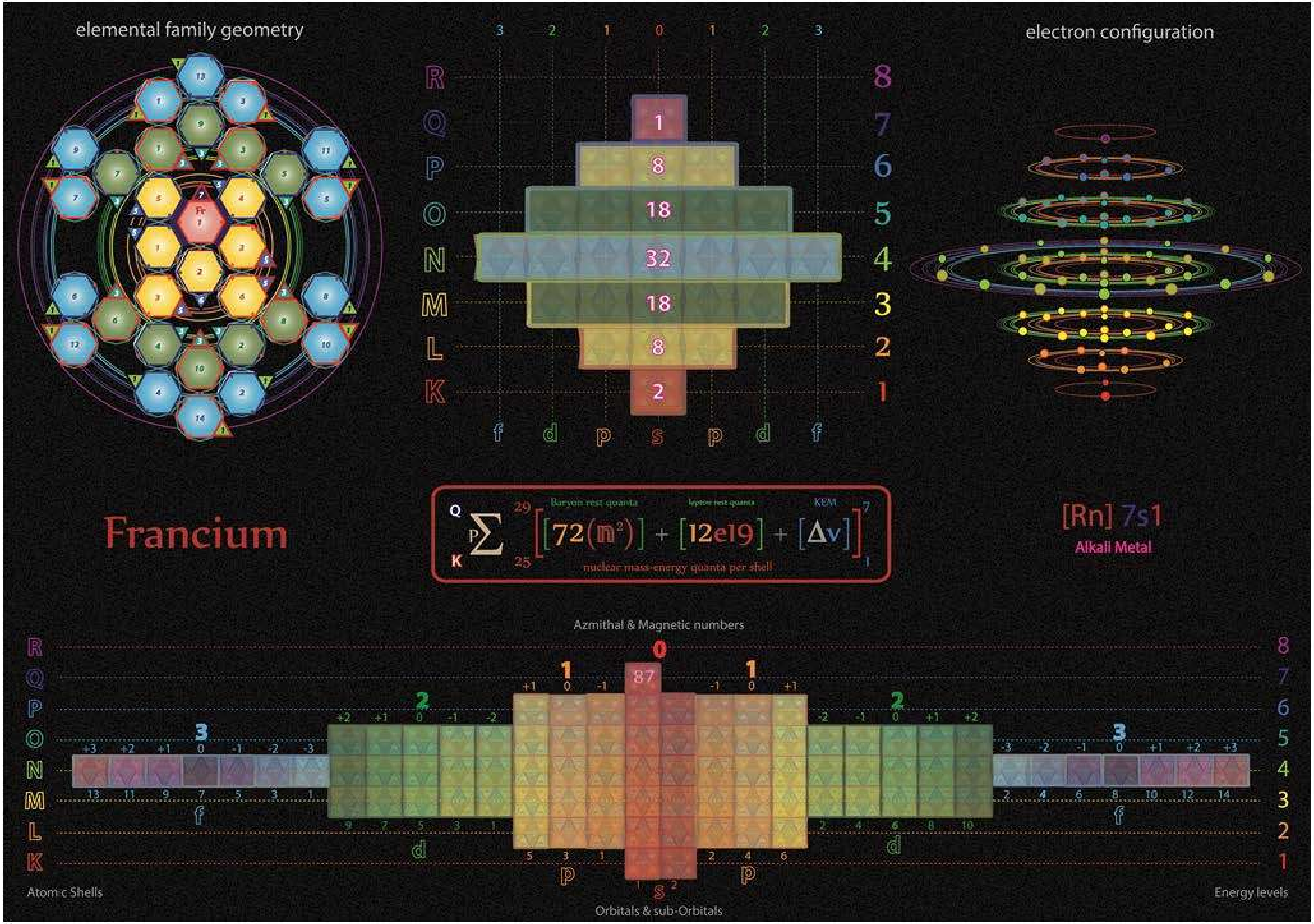
elemental family geometry
electron configuration
Francium
[Rn] 7s1
Alkali Metal
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
1
8
18
32
18
8
2
Baryon rest quanta
lepton rest quanta
KEM
Q ∑P K 29 25 [72(m²)] + [12e19] + [Δv] 7 1
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
87
0
1 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
13 11 9 7 5 3 1
+2 +1 0 -1 -2
9 7 5 3 1
+1 0 -1
5 3 1
-1 0 +1
2 4 6
-2 -1 0 +1 +2
2 4 6 8 10
-3 -2 -1 0 +1 +2 +3
2 4 6 8 10 12 14
f d p s p d f
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.87 - Francium atomic config

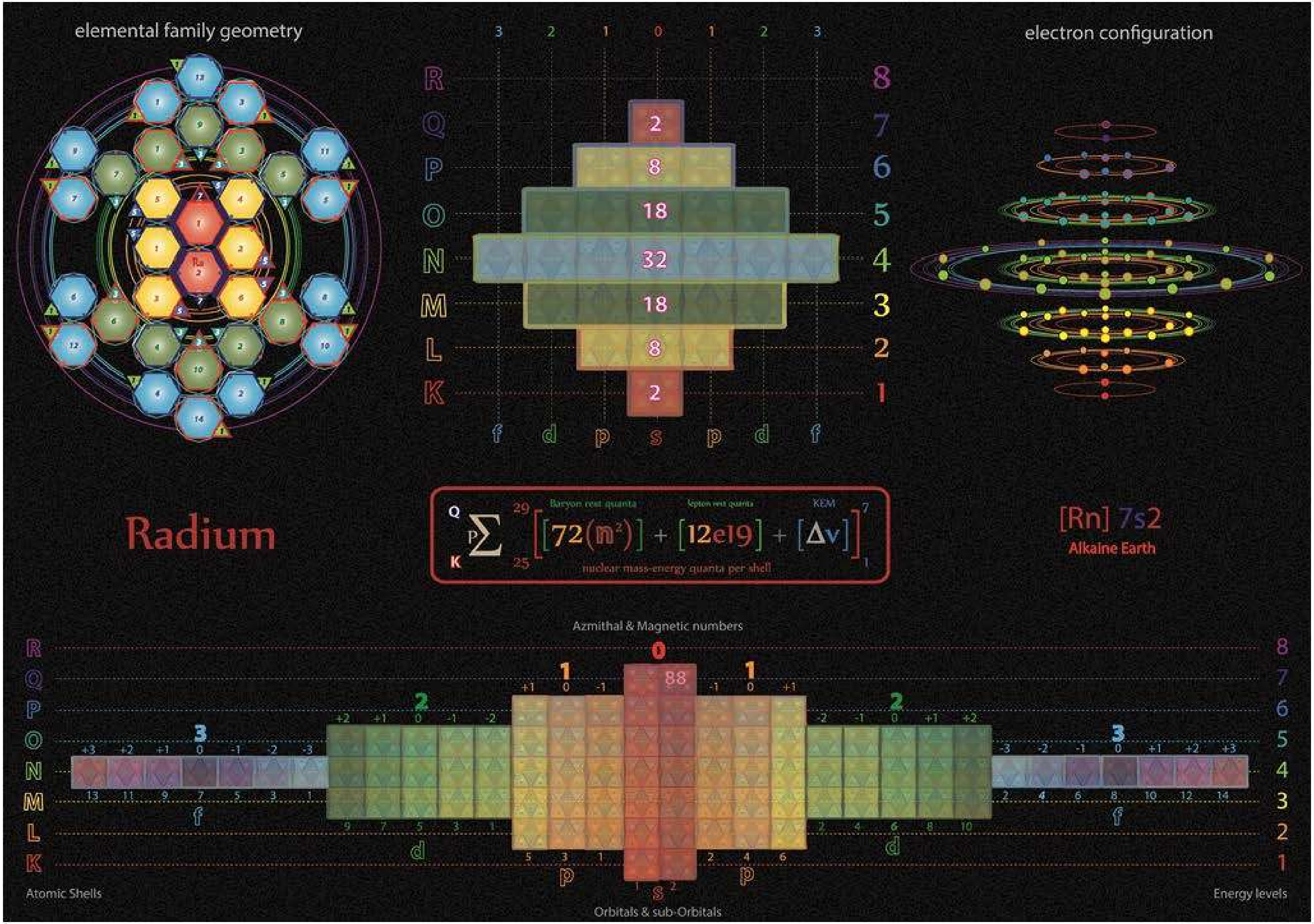
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
2
8
18
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Radium
Q
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 7s2
Alkaine Earth
Azmithal & Magnetic numbers
0
88
1
+1 0 -1
-1 0 +1
2
+2 +1 0 -1 -2
-2 -1 0 +1 +2
3
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
f
d
p
s
1 2
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.88 - Radium atomic config

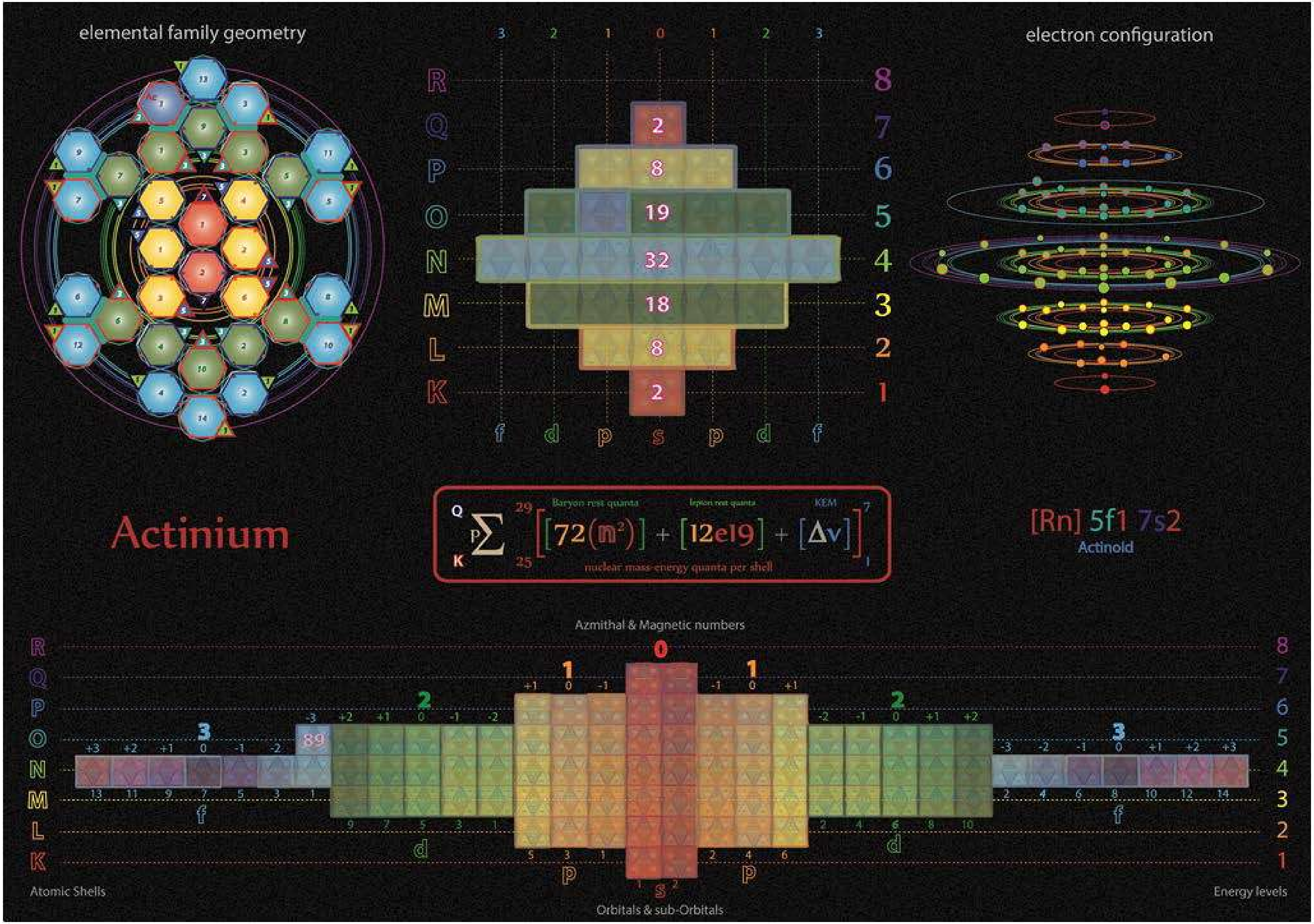
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2
8
19
32
18
8
2
f d p s p d f
Actinium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f1 7s2
Actinoid
Azmithal & Magnetic numbers
0
1
2
3
89
f
d
p
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.89 - Actinium atomic config

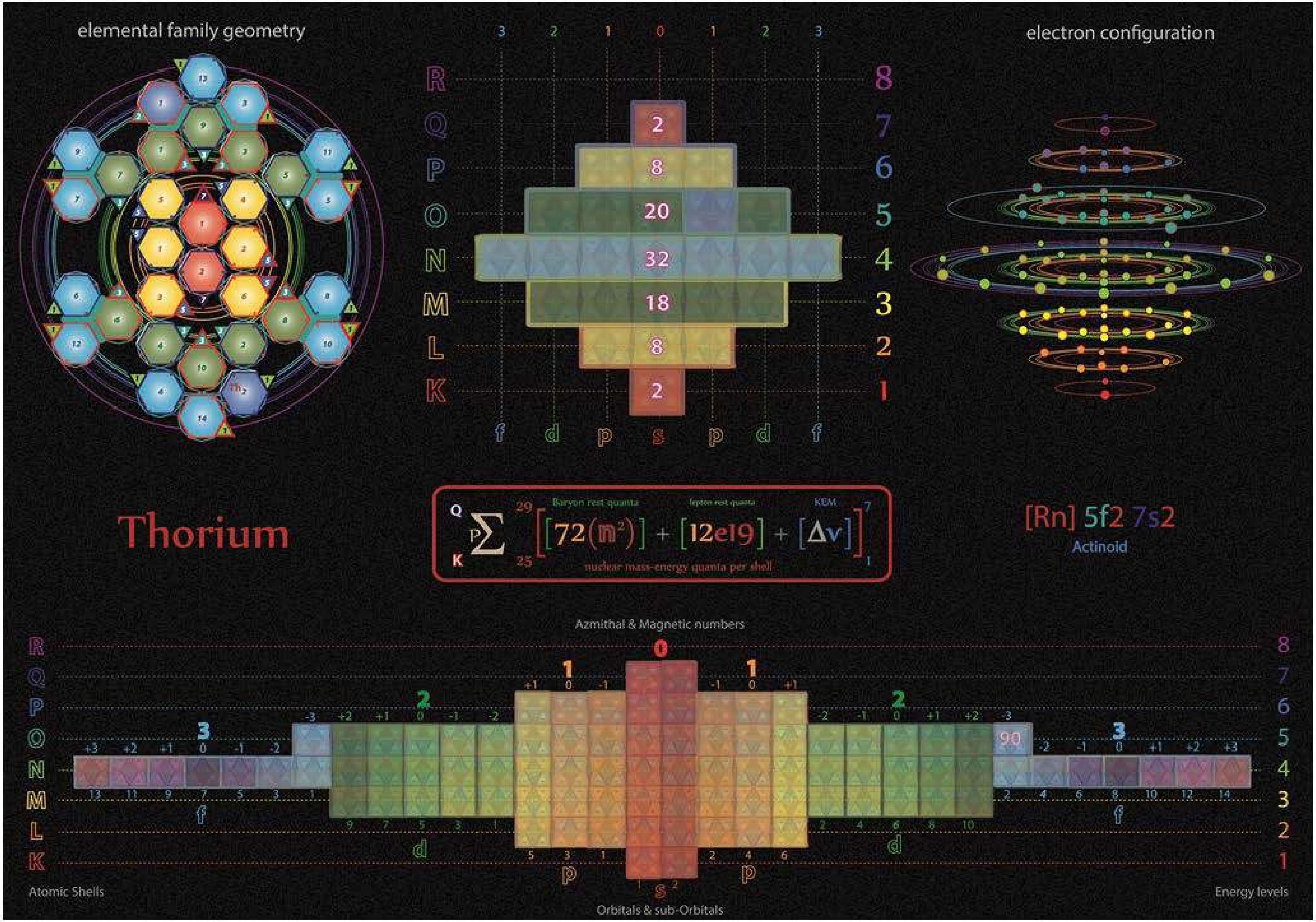
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
20
32
18
8
2
f d p s p d f
Thorium
Q
K
Baryon rest quanta
lepton rest quanta
KEM
29
25
7
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Rn] 5f2 7s2
Actinoid
Azmithal & Magnetic numbers
0
1
2
3
90
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.90 - Thorium atomic config

Copyright ABRAHAM [2008] - All rights reserved

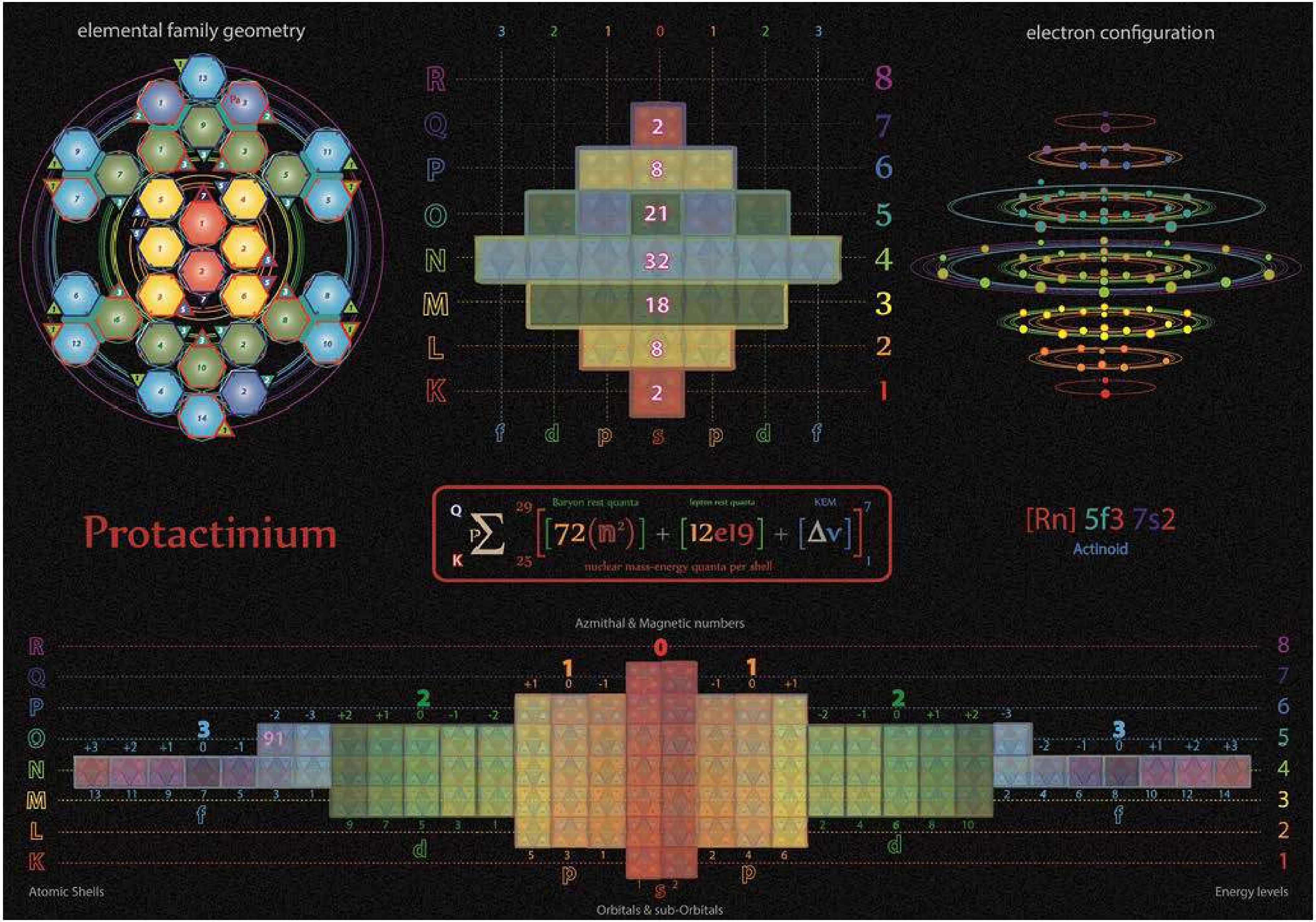
elemental family geometry
Pa
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
21
32
18
8
2
f d p s p d f
electron configuration
Protactinium
Q
K
p
29
25
Baryon rest quanta
lepton rest quanta
KEM
7
1
$\sum_{K}^{Q} [72(m^2)] + [12e19] + [\Delta v]$
nuclear mass-energy quanta per shell
[Rn] 5f3 7s2
Actinoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
-2 -1 0 +1 +2 +2 +1 0 -1 -2
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 0 +1 +2 +3
91
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.91 - Protactinium atomic config

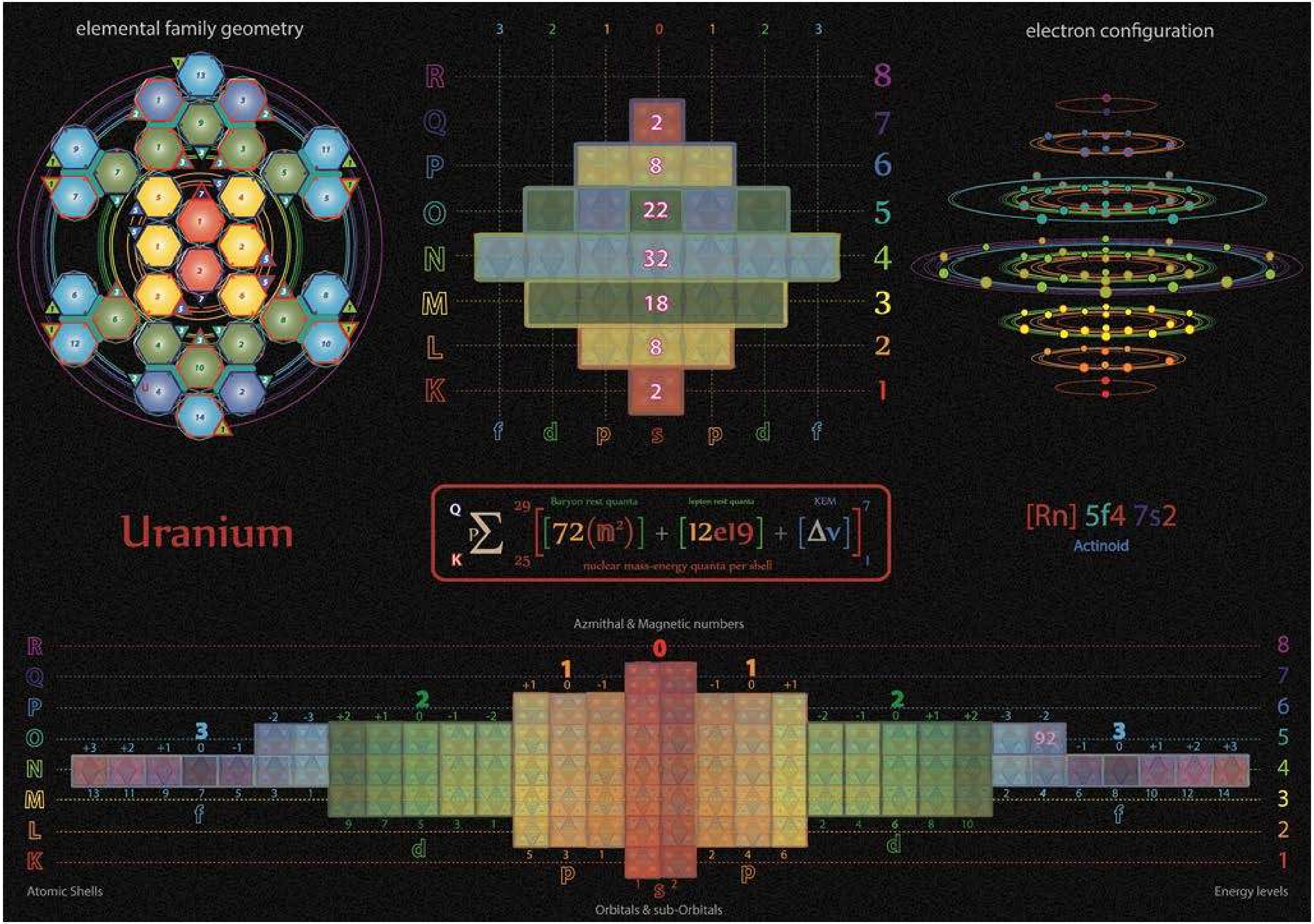
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
22
32
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+3 +2 +1 0 -1
-2 -3 +2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2 -3 -2
92
-1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Uranium

$$\sum_{K}^{Q} {}_{P} \left[\overset{\text{Baryon rest quanta}}{\left[72(m^2)\right]} + \overset{\text{lepton rest quanta}}{\left[12e19\right]} + \overset{\text{KEM}}{\left[\Delta v\right]} \right]_{1}^{7}$$

29 25

nuclear mass-energy quanta per shell

[Rn] 5f4 7s2

Actinoid

Tetryonics 53.92 - Uranium atomic config

Copyright ABRAHAM [2008] - All rights reserved

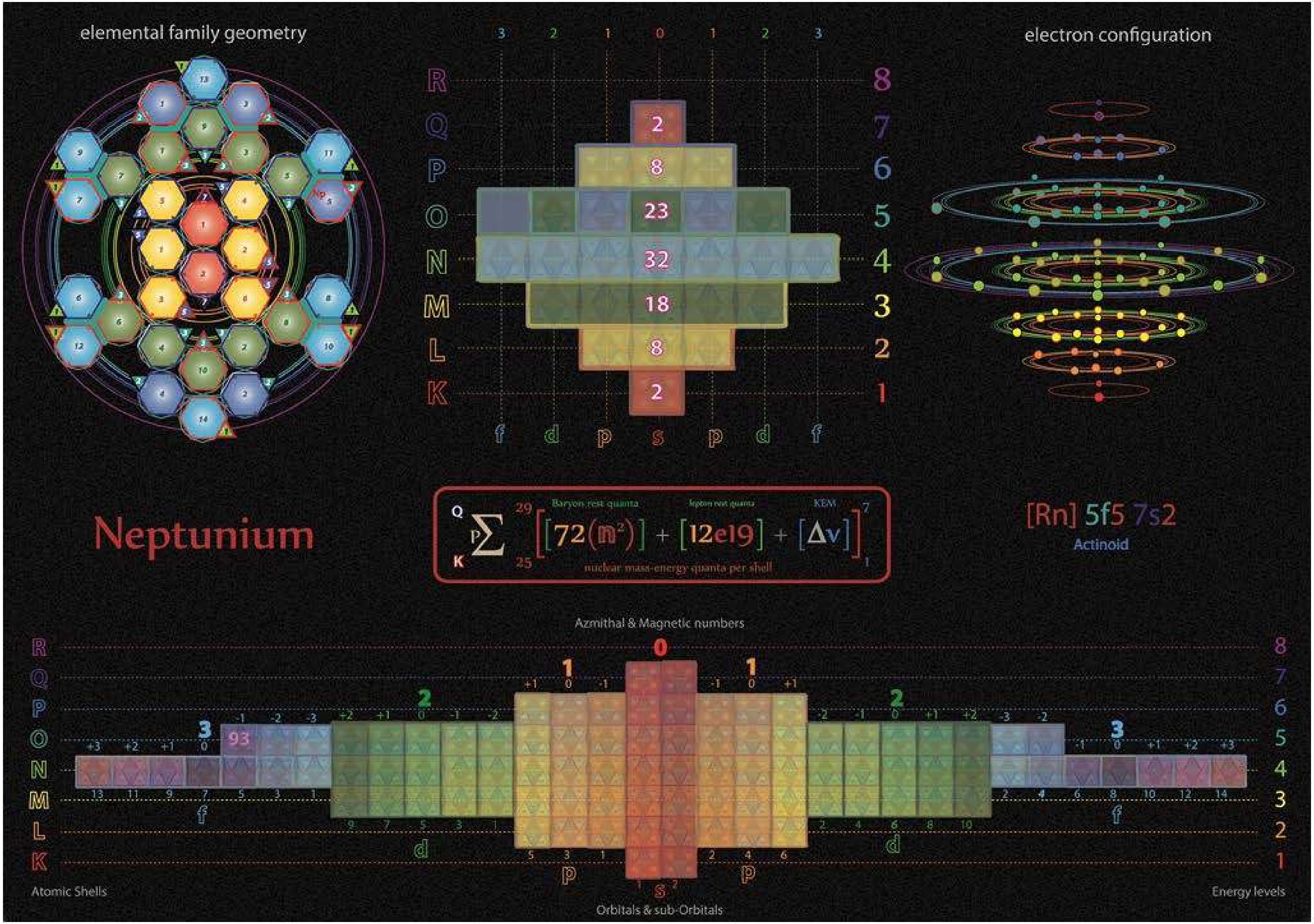
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2 8 23 32 18 8 2
f d p s p d f
Azmithal & Magnetic numbers
93
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Neptunium

$$\sum_{K}^{Q} p \left[[72(n^2)] + [12e19] + [\Delta v] \right]_{1}^{7}$$

29 25

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Rn] 5f5 7s2

Actinoid

Tetryonics 53.93 - Neptunium atomic config

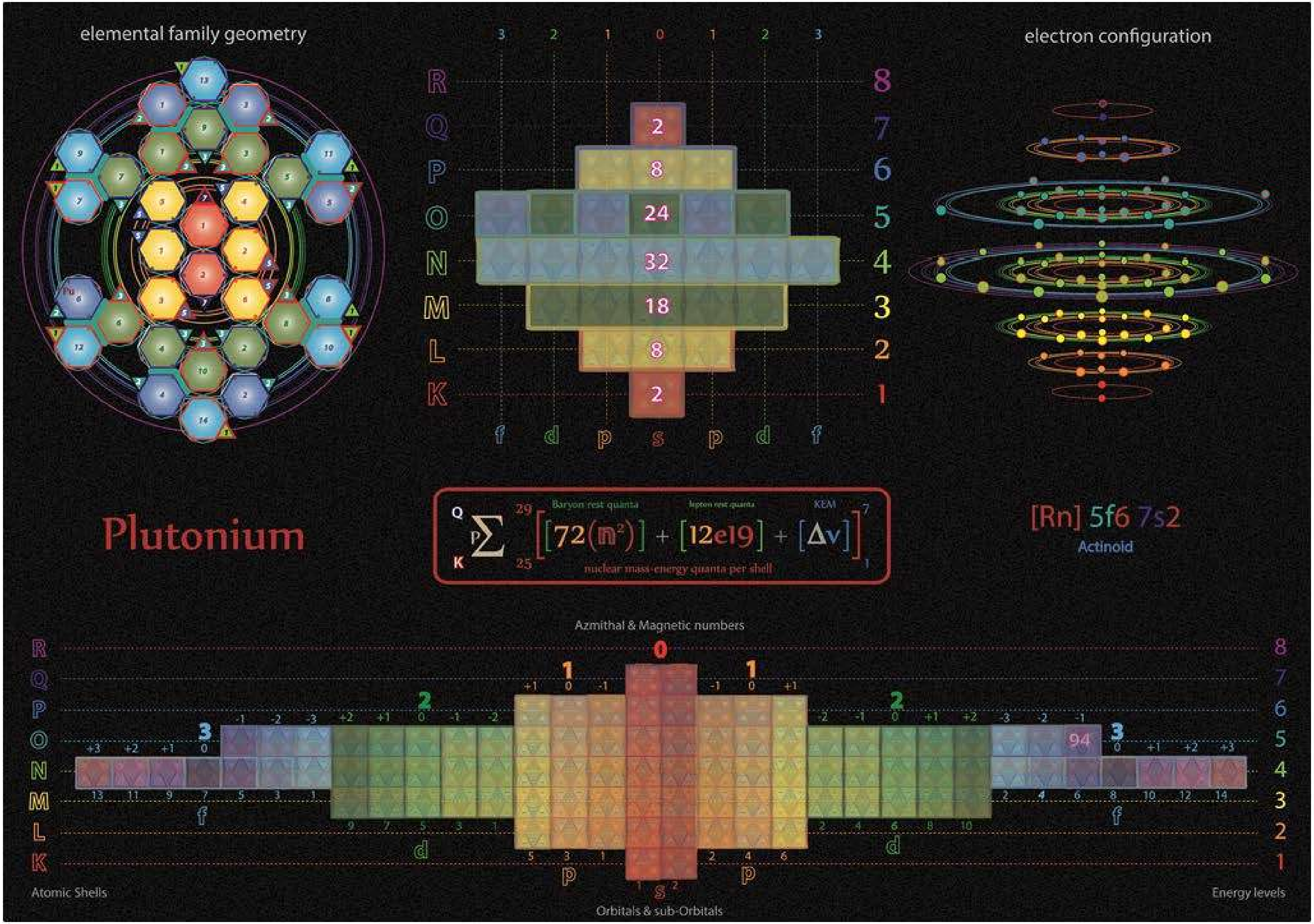
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
24
32
18
8
2
f d p s p d f
Plutonium
Baryon rest quanta
lepton rest quanta
KEM
Q
P
K
29
25
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f6 7s2
Actinoid
Azmithal & Magnetic numbers
0
1
2
3
94
f
d
p
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.94 - Plutonium atomic config

Copyright ABRAHAM [2008] - All rights reserved

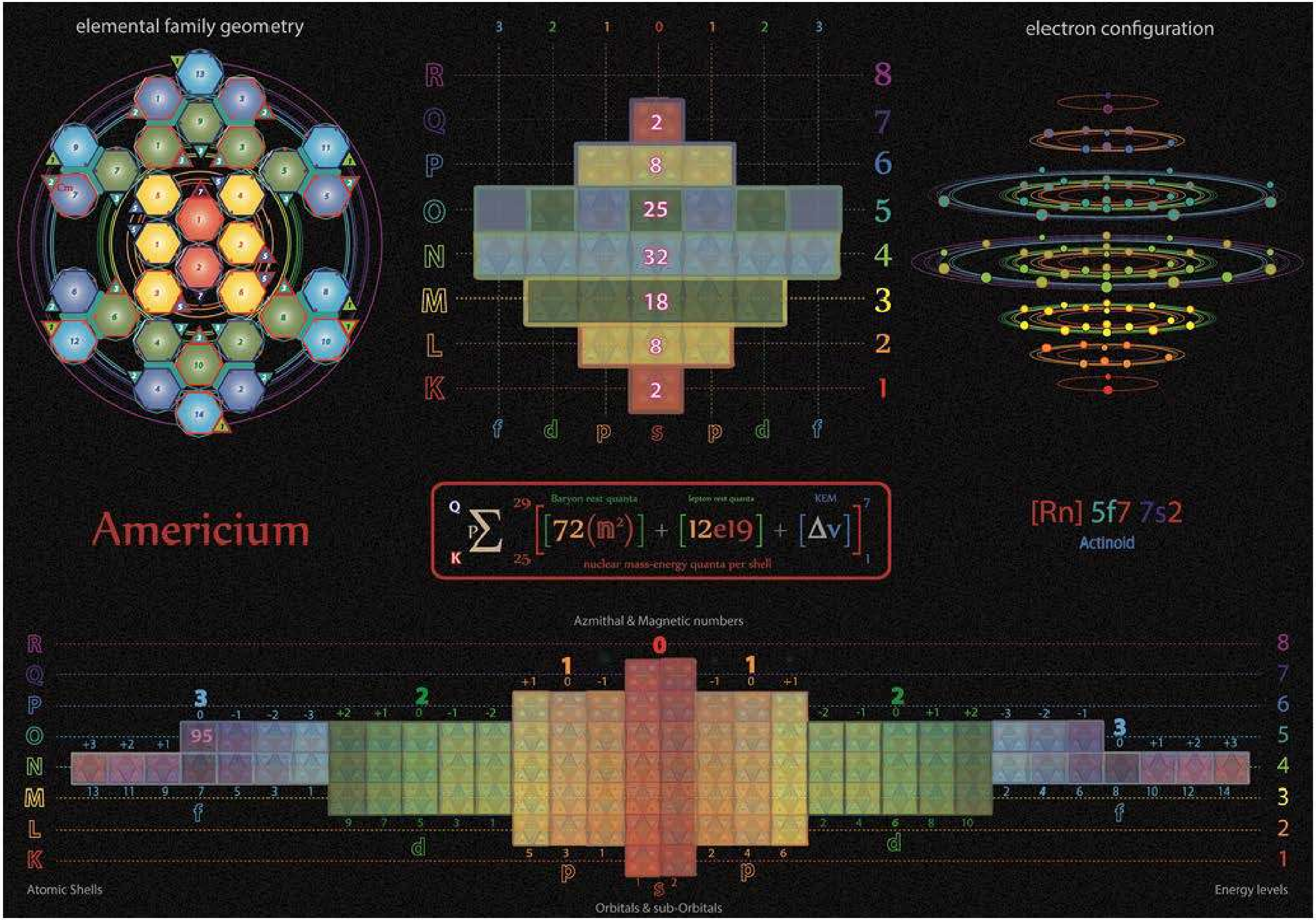
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
25
32
18
8
2
f d p s p d f
Americium
$\overset{Q}{\underset{K}{}} P\sum_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_1^7$
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Rn] 5f7 7s2
Actinoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
95
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.95 - Americium atomic config

Copyright ABRAHAM [2008] - All rights reserved

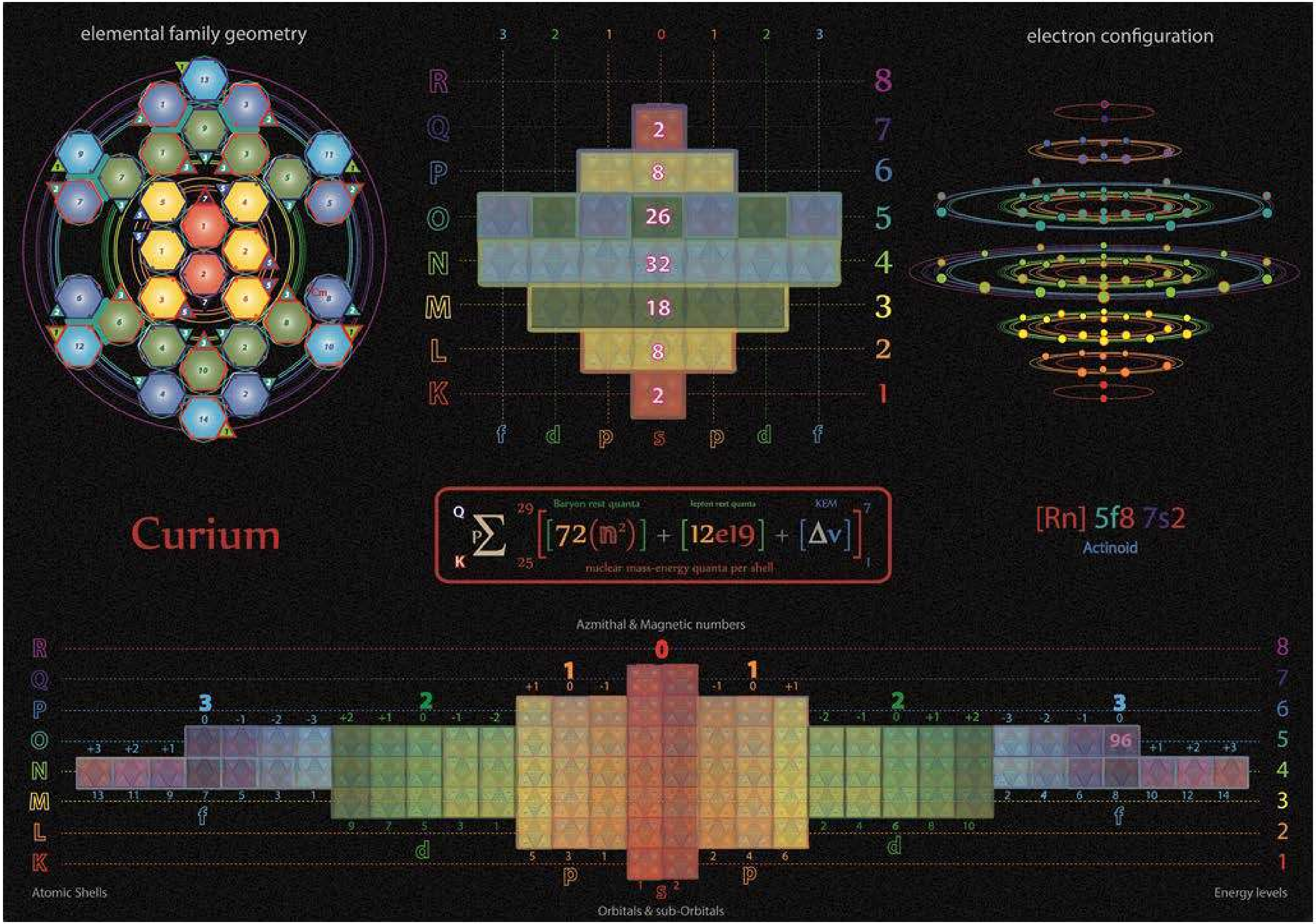
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
26
32
18
8
2
f d p s p d f
Curium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
7
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Rn] 5f8 7s2
Actinoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
0 -1 -2 -3
-3 -2 -1 0
+3 +2 +1
+1 +2 +3
96
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.96 - Curium atomic config

Copyright ABRAHAM [2008] - All rights reserved

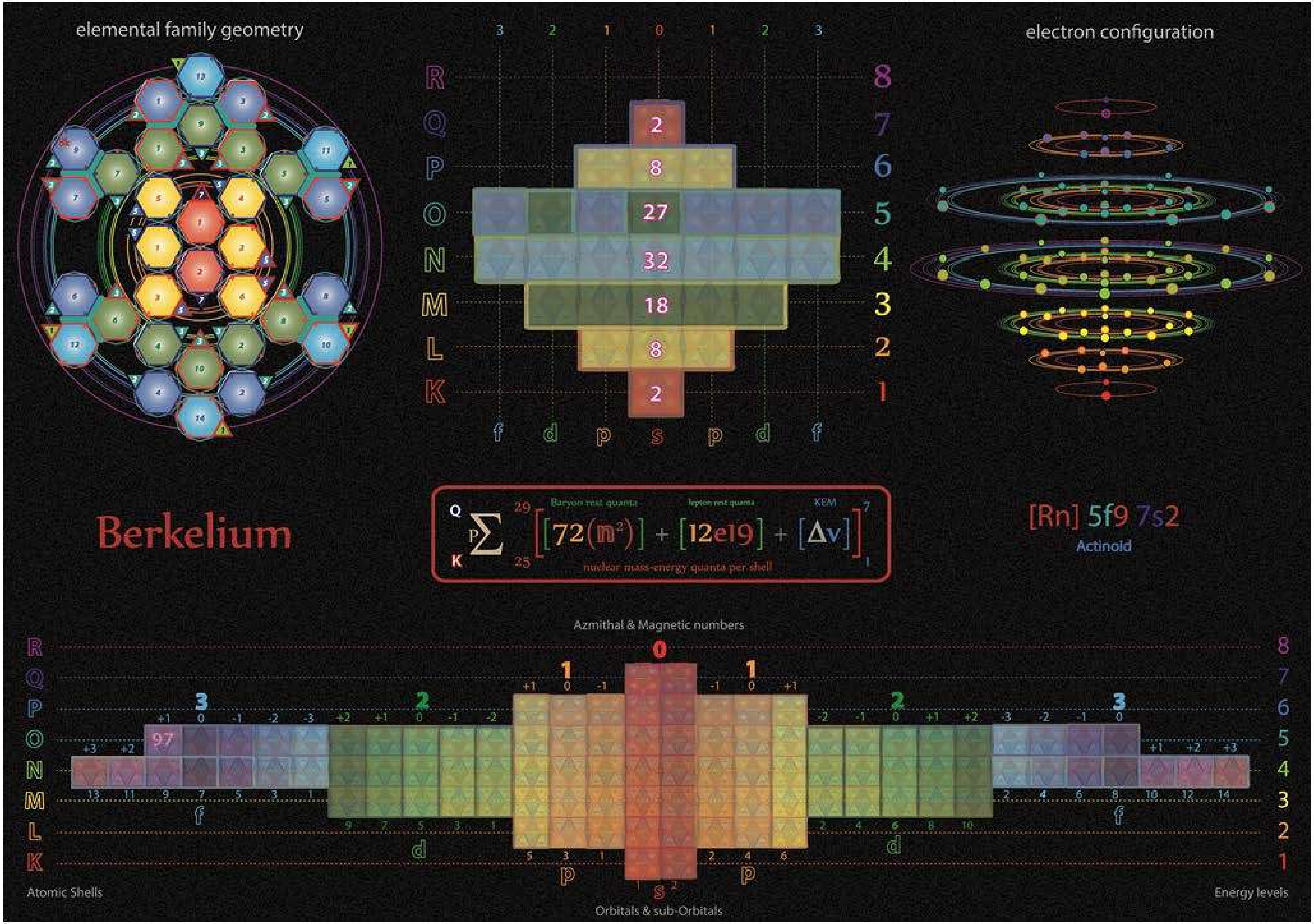
elemental family geometry
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
2
8
27
32
18
8
2
f d p s p d f
Berkelium
Q
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f9 7s2
Actinoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
97
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f
d
p
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.97 - Berkelium atomic config

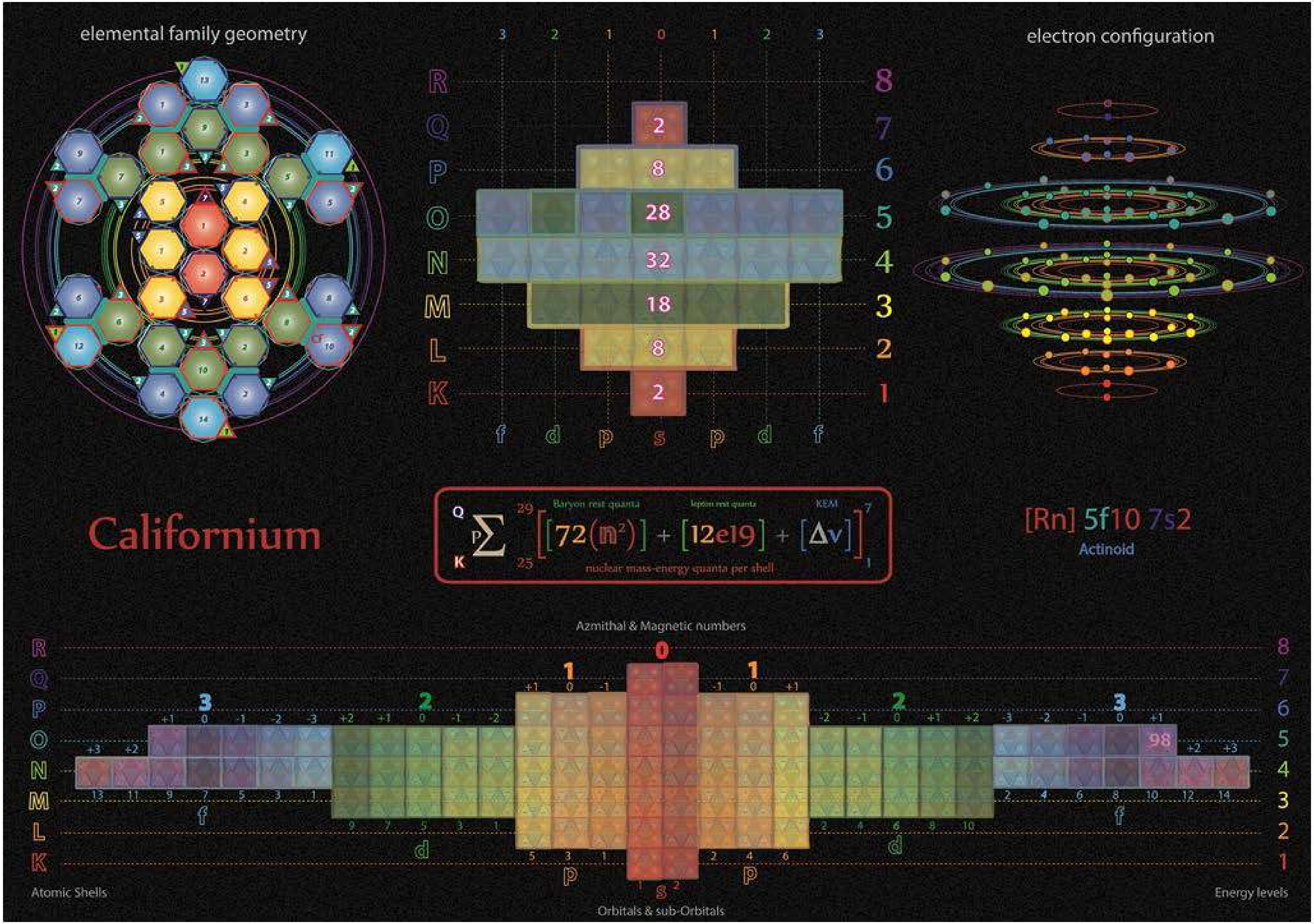
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
28
32
18
8
2
f d p s p d f
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
0
1 1
+1 0 -1 -1 0 +1
3 2 2 3
+1 0 -1 -2 -3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3 -2 -1 0 +1
+3 +2 +2 +3
98
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Californium

[Rn] 5f10 7s2

Actinoid

Tetryonics 53.98 - Californium atomic config

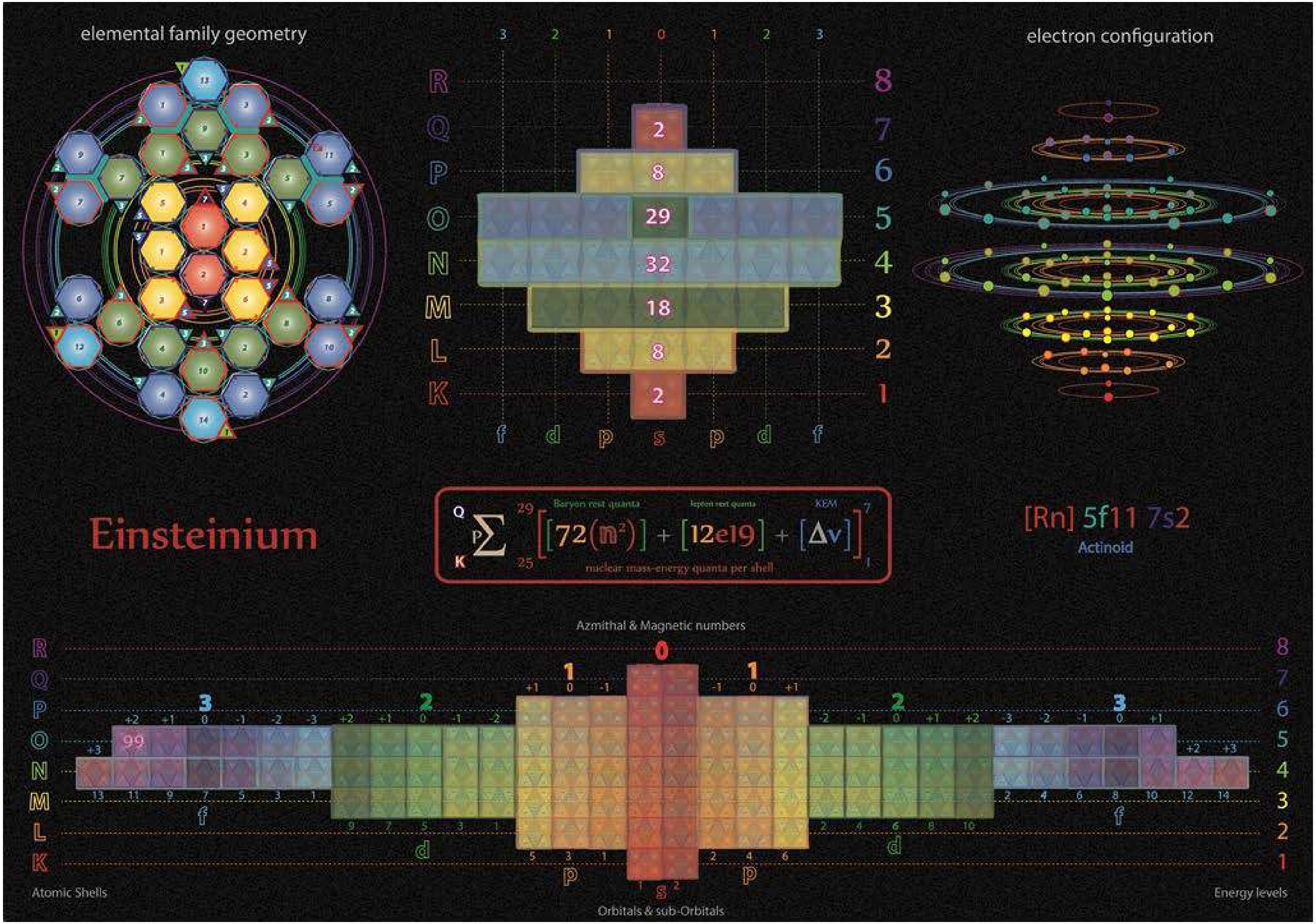
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
29
32
18
8
2
f d p s p d f
Einsteinium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[[72(n²)] + [12e19] + [Δv]]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f11 7s2
Actinoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3 -2 -1 0 +1
+3 +2 +3
99
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.99 - Einsteinium atomic config

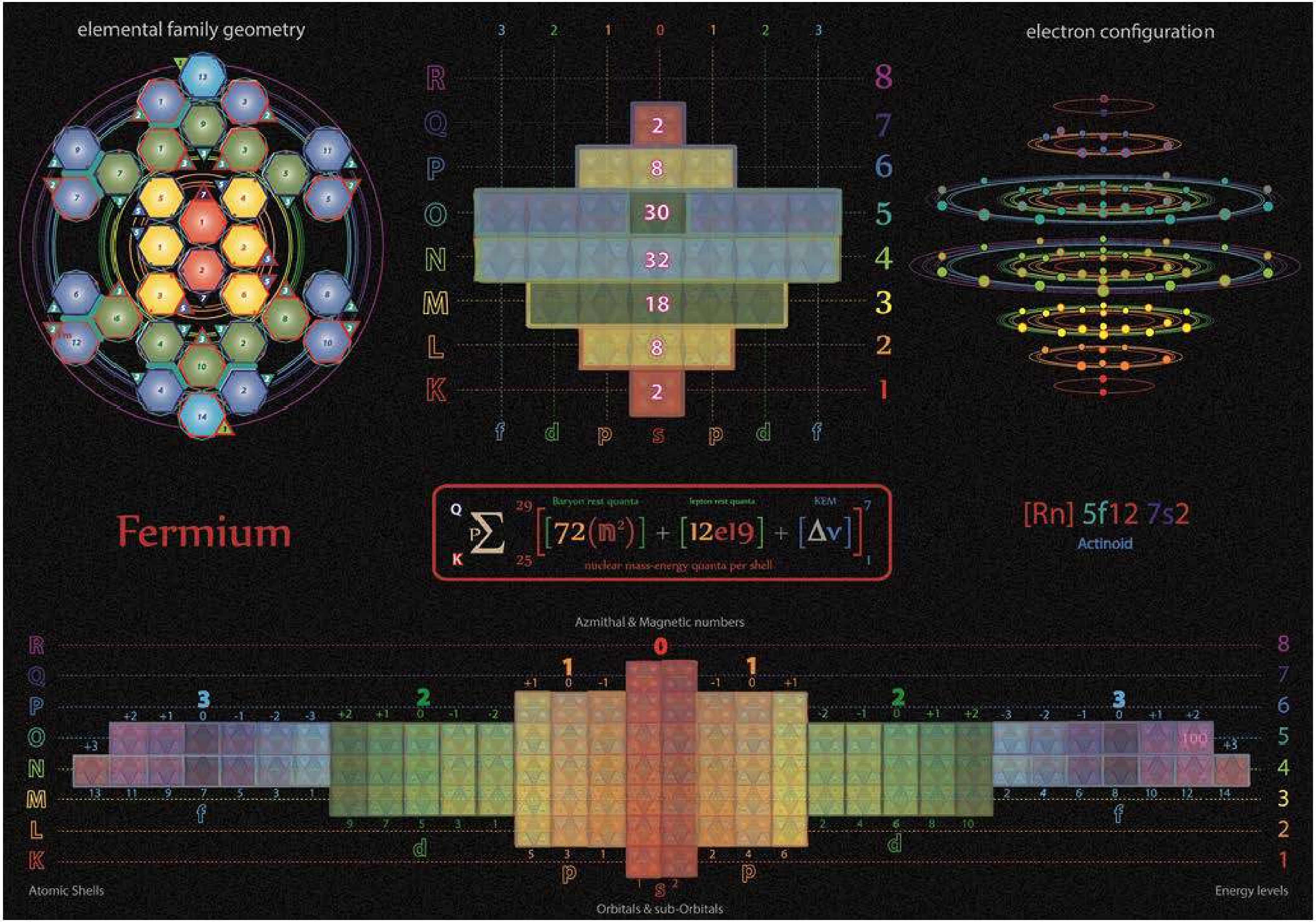
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
30
32
18
8
2
f d p s p d f
Fermium
Q
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f12 7s2
Actinoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
100
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.100 - Fermium atomic config

Copyright ABRAHAM [2008] - All rights reserved

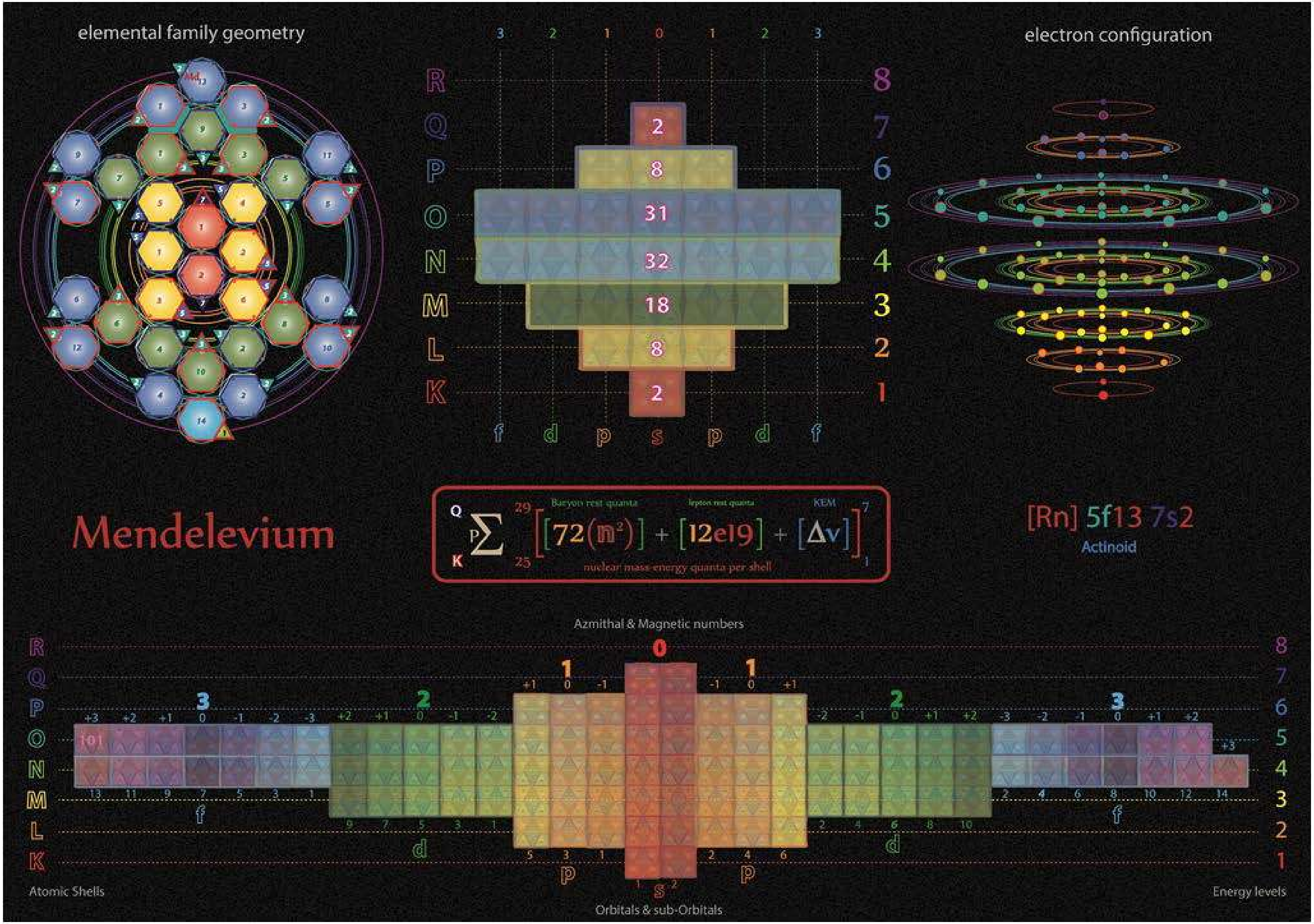

elemental family geometry

electron configuration

Mendelevium

$$\sum_{K}^{Q} P \; {}_{25}^{29}\Big[[72(m^2)] + [12e19] + [\Delta v] \Big]_{1}^{7}$$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Rn] 5f13 7s2

Actinoid

Azmithal & Magnetic numbers

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
2
8
31
32
18
8
2
f d p s p d f
0
1 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
101
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f

Tetryonics 53.101 - Mendelevium atomic config

Copyright ABRAHAM [2008] - All rights reserved

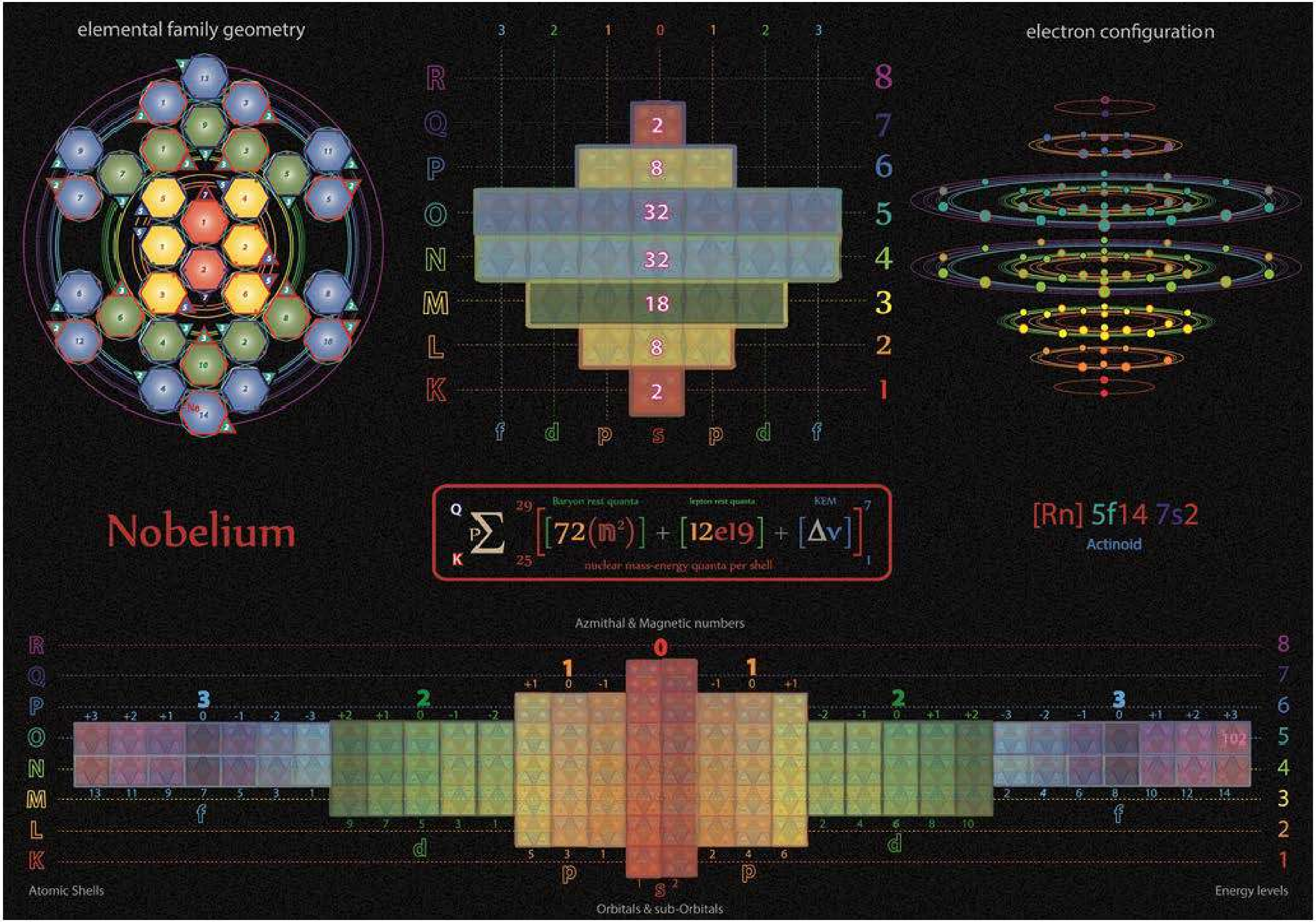
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2
8
32
32
18
8
2
f d p s p d f
Nobelium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f14 7s2
Actinoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
102
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.102 - Nobelium atomic config

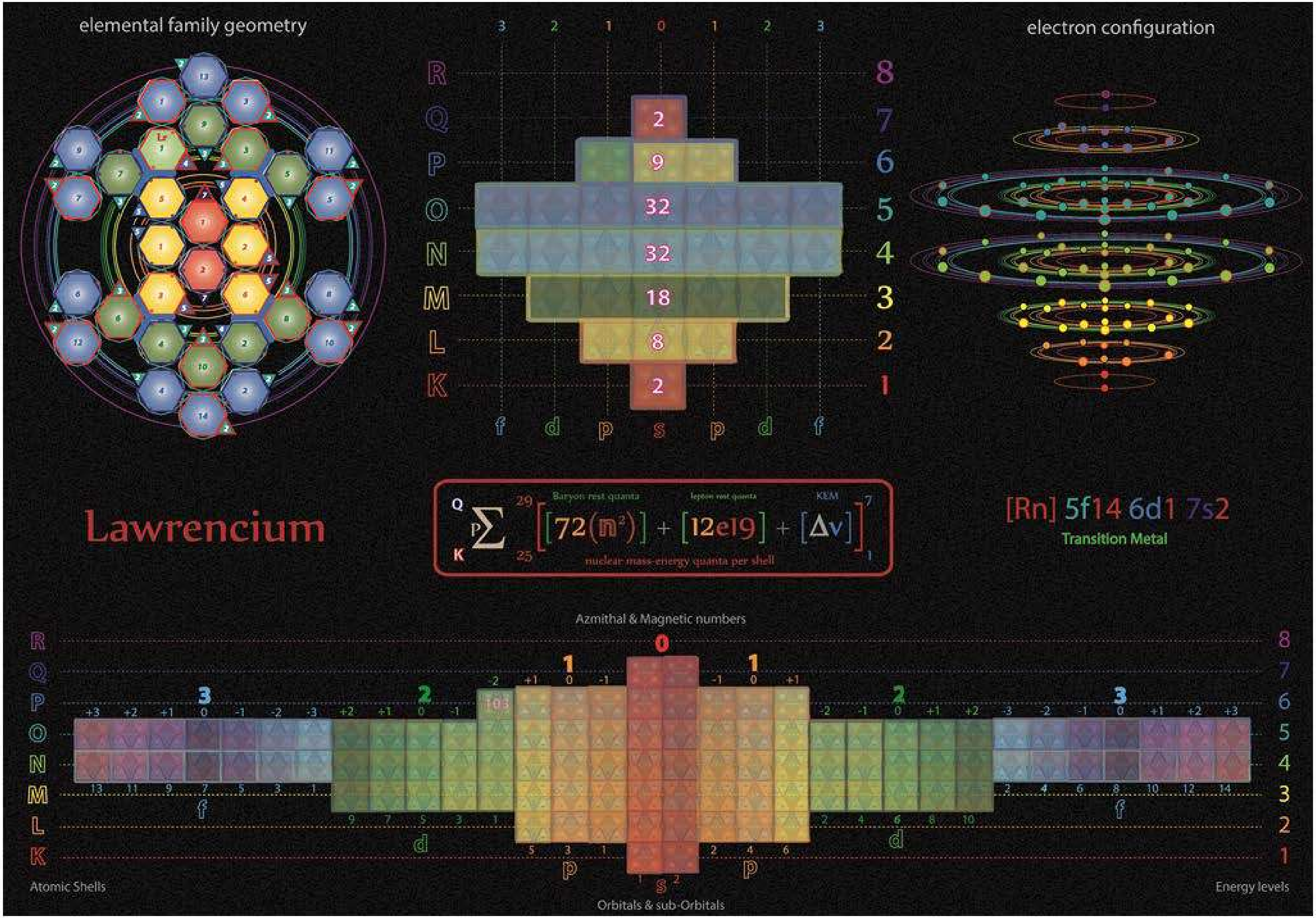
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2
9
32
32
18
8
2
f d p s p d f
Lawrencium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
72(n²)
12e19
Δv
7
1
nuclear mass-energy quanta per shell
[Rn] 5f14 6d1 7s2
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
103
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.103 - Lawrencium atomic config

elemental family geometry

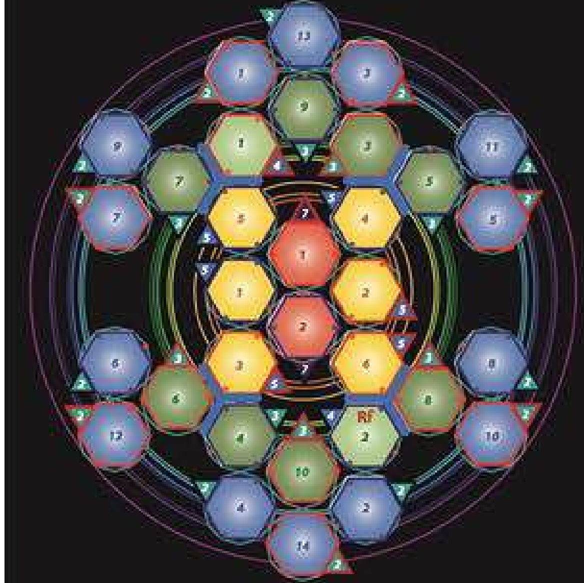

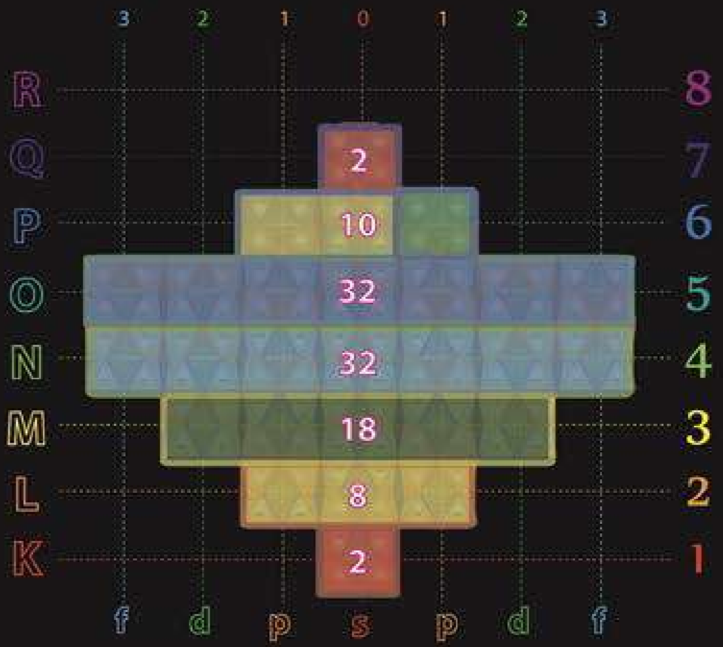
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
10
32
32
18
8
2
f
d
p
s
p
d
f

electron configuration

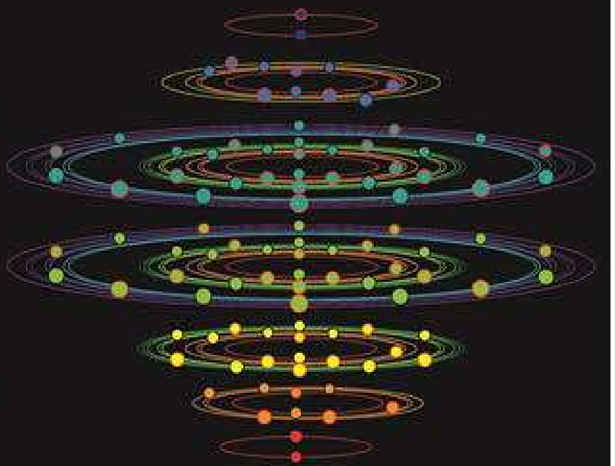

Rutherfordium

$$\sum_{K}^{Q} P \sum_{25}^{29} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{7}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Rn] 5f14 6d2 7s2

Transition Metal

Azmithal & Magnetic numbers

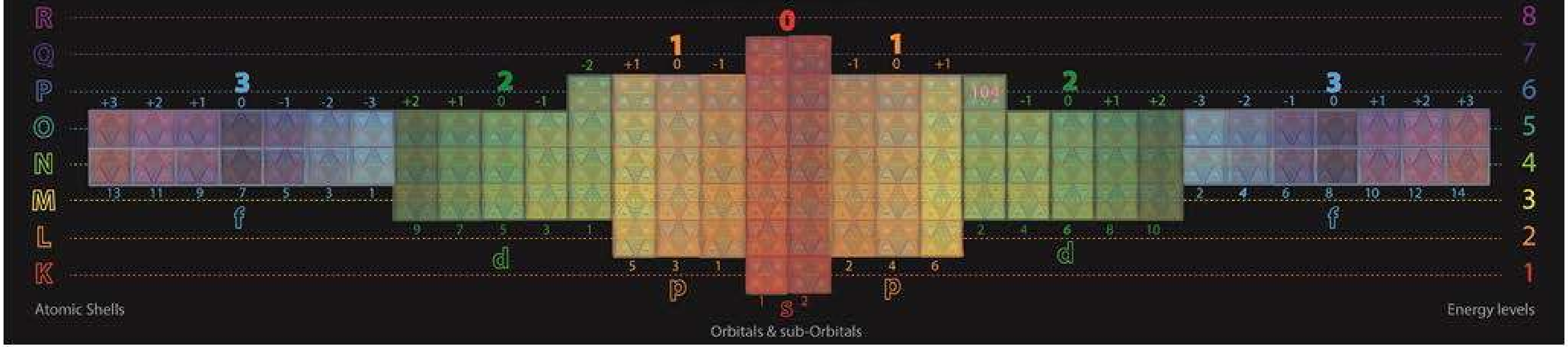
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
-2
+1
0
-1
-1
0
+1
+3
+2
+1
0
-1
-2
-3
+2
+1
0
-1
104
-1
0
+1
+2
-3
-2
-1
0
+1
+2
+3
13
11
9
7
5
3
1
9
7
5
3
1
5
3
1
1
2
2
4
6
2
4
6
8
10
2
4
6
8
10
12
14
f
d
p
s
p
d
f

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.104 - Rutherfordium atomic config

Copyright ABRAHAM [2008] - All rights reserved

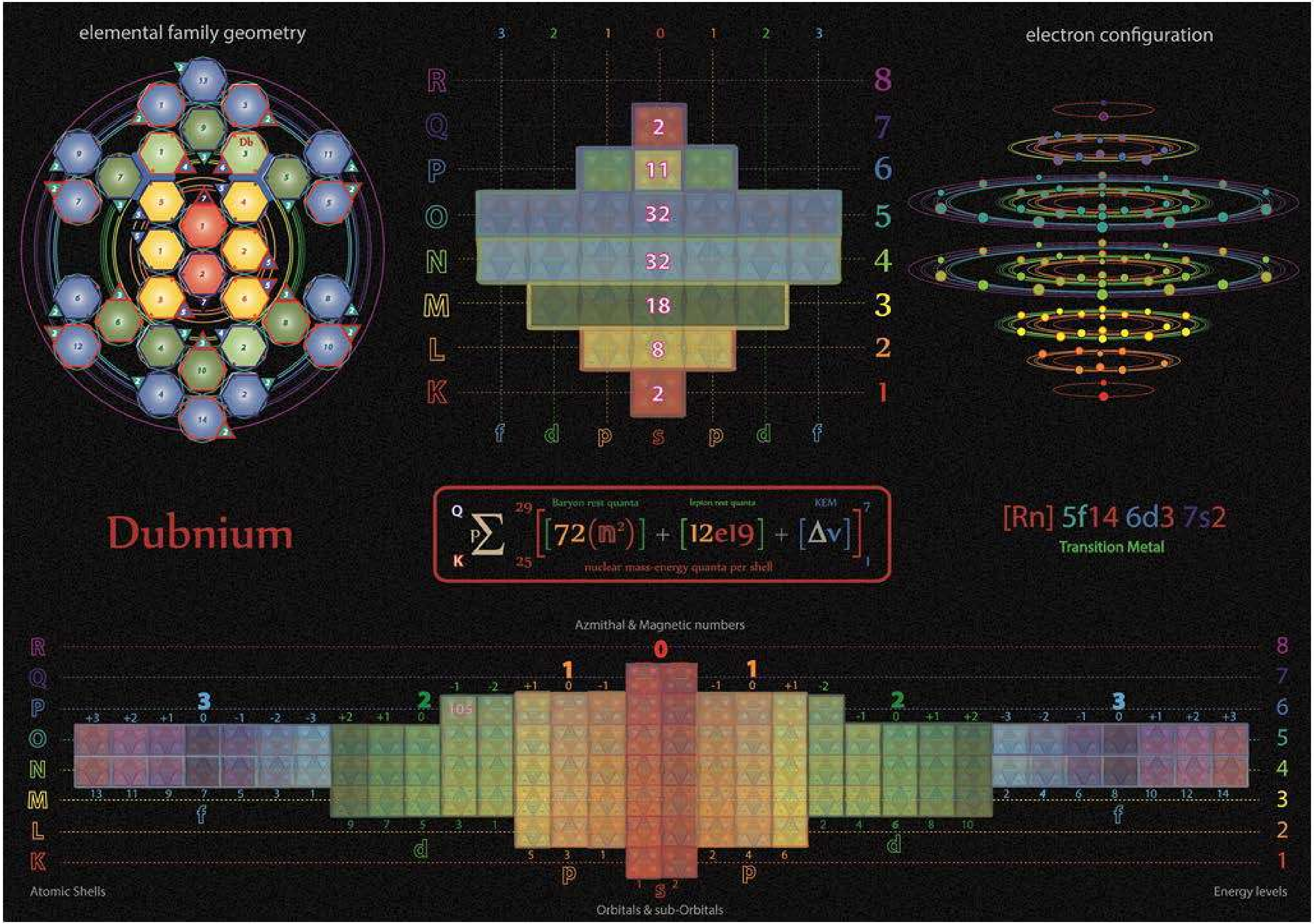
elemental family geometry
Db
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2
11
32
32
18
8
2
f d p s p d f
Dubnium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f14 6d3 7s2
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
-1 0 +1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
105
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.105 - Dubnium atomic config

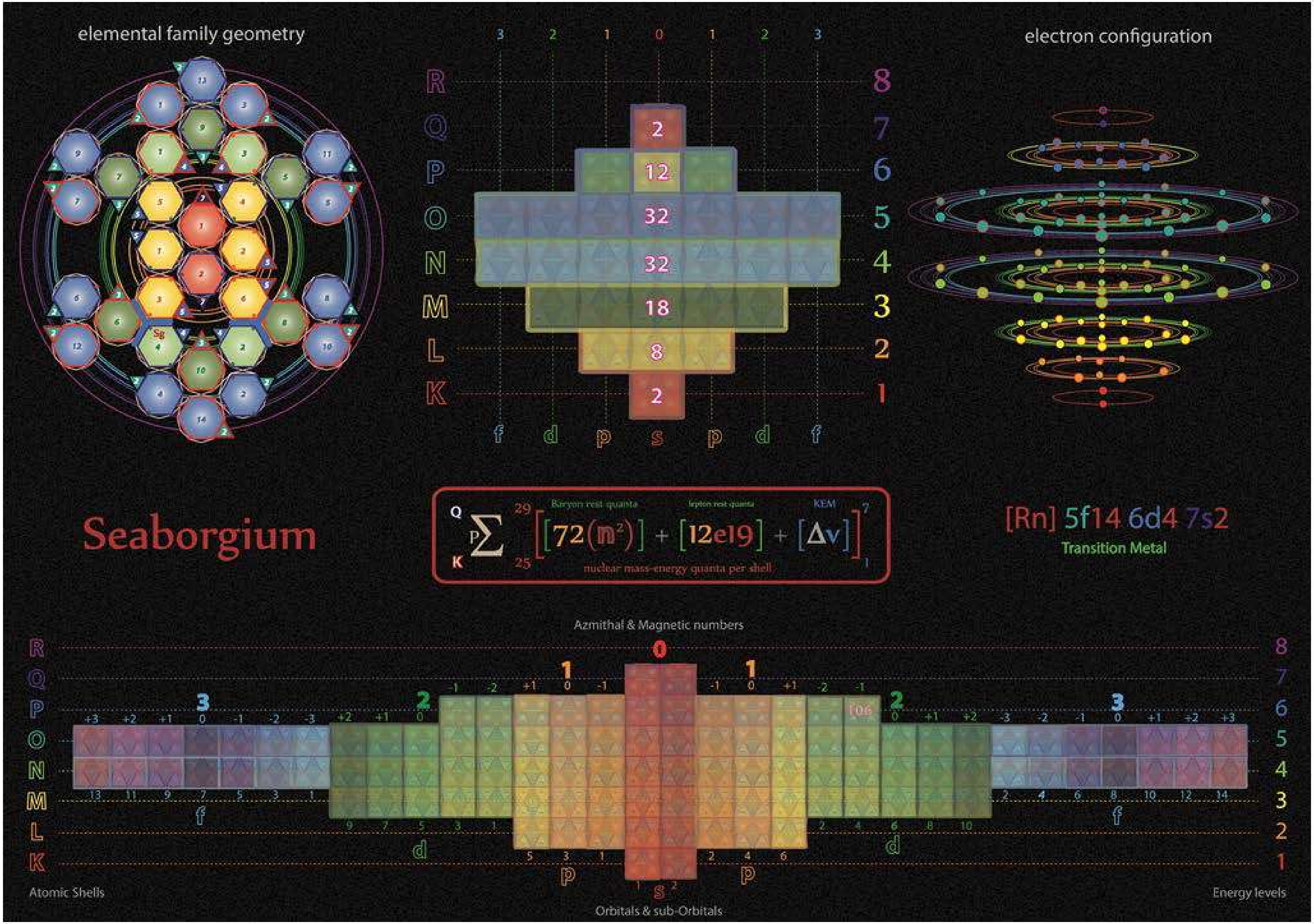
elemental family geometry
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
12
32
32
18
8
2
f
d
p
s
p
d
f
Seaborgium
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Rn] 5f14 6d4 7s2
Transition Metal
Azmithal & Magnetic numbers
106
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.106 - Seaborgium atomic config

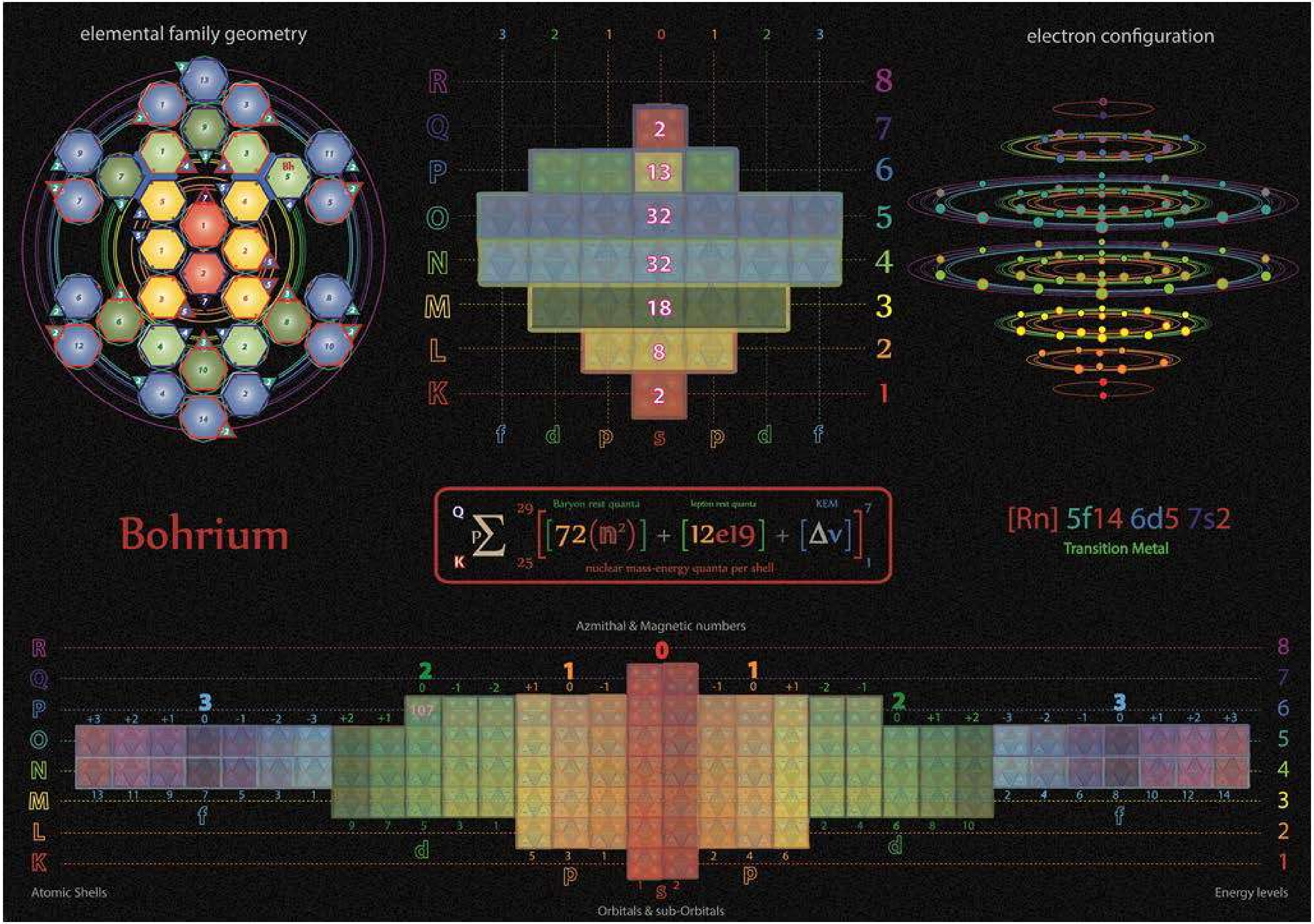
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
13
32
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
107
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Bohrium

Baryon rest quanta — lepton rest quanta — KEM

$$\sum_{K}^{Q} P \; {}_{25}^{29}\Big[\,[72(n^2)] + [12e19] + [\Delta v]\,\Big]_{1}^{7}$$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d5 7s2

Transition Metal

Tetryonics 53.107 - Bohrium atomic config

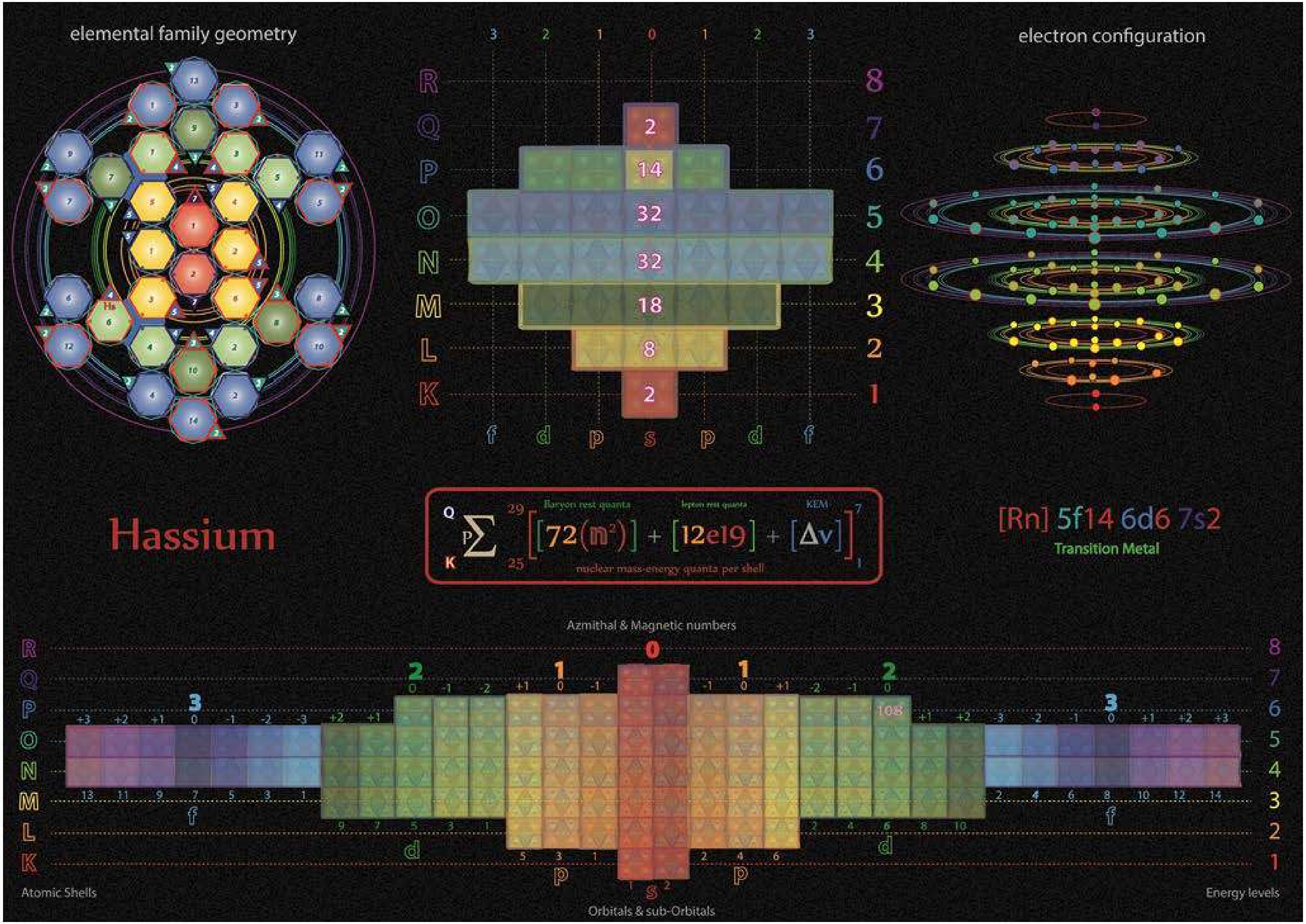
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
14
32
32
18
8
2
f d p s p d f
Hassium
$_{K}^{Q}\,{}_{P}\sum_{25}^{29}\left[\left[72(n^2)\right]+\left[12e19\right]+\left[\Delta v\right]\right]_{1}^{7}$
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Rn] 5f14 6d6 7s2
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
108
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.108 - Hassium atomic config

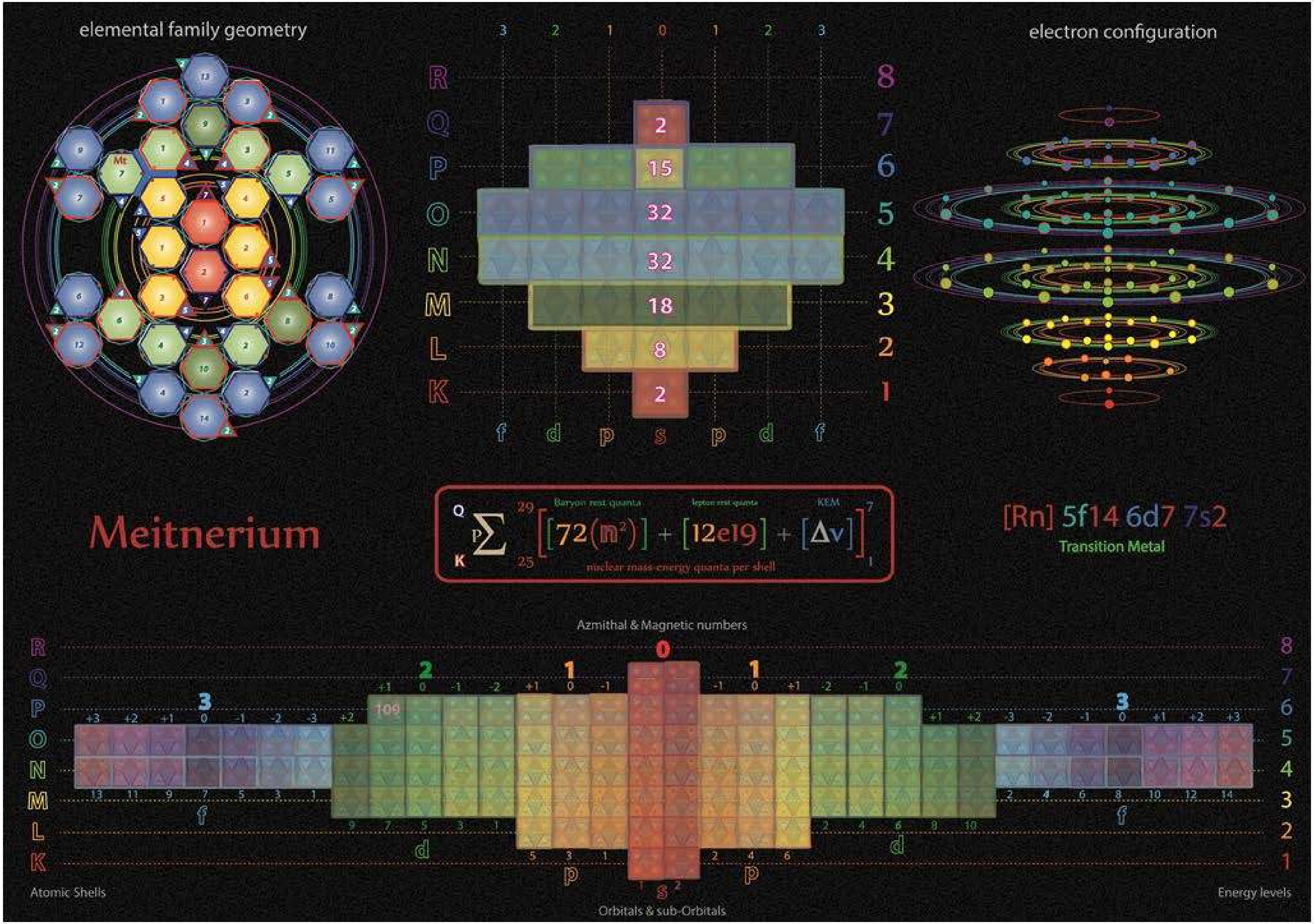
elemental family geometry
Mt
electron configuration
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
2
15
32
32
18
8
2
8
7
6
5
4
3
2
1
f d p s p d f
Meitnerium
Q
K
p∑
29
25
Baryon rest quanta
lepton rest quanta
KEM
[[72(n²)] + [12e19] + [Δv]]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f14 6d7 7s2
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
109
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f
d
p
s
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.109 - Meitnerium atomic config

elemental family geometry

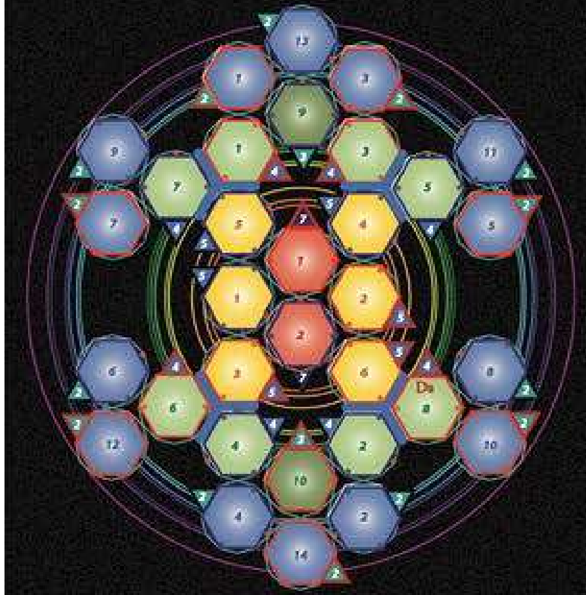

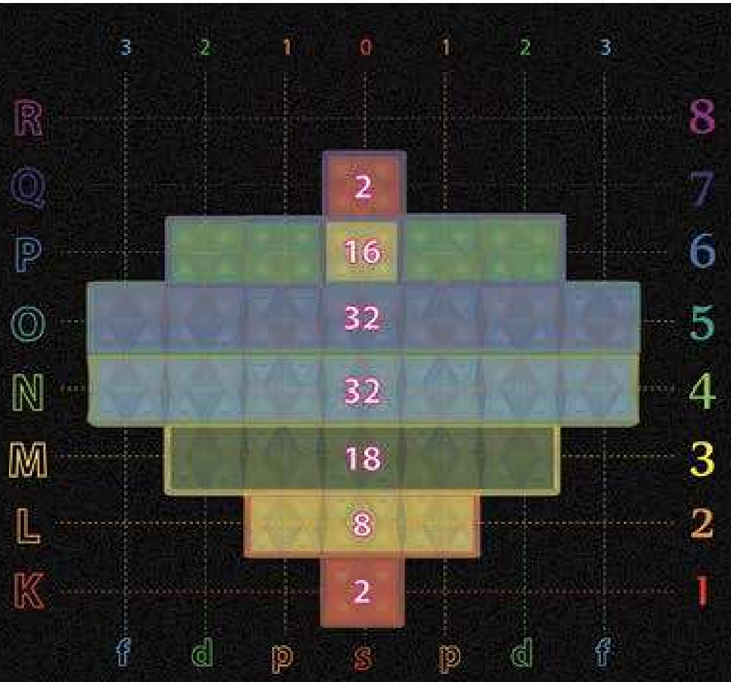
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
16
32
32
18
8
2
f d p s p d f

electron configuration

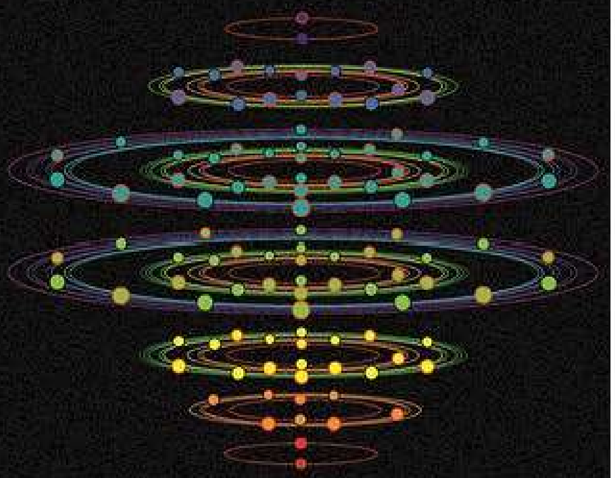

Darmstadtium

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{Q} {}_{P} \; {}^{29}_{25}\left[\left[72(\mathrm{m}^2)\right] + \left[12e19\right] + \left[\Delta v\right] \right]^{7}_{1}$$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d8 7s2

Transition Metal

Azmithal & Magnetic numbers

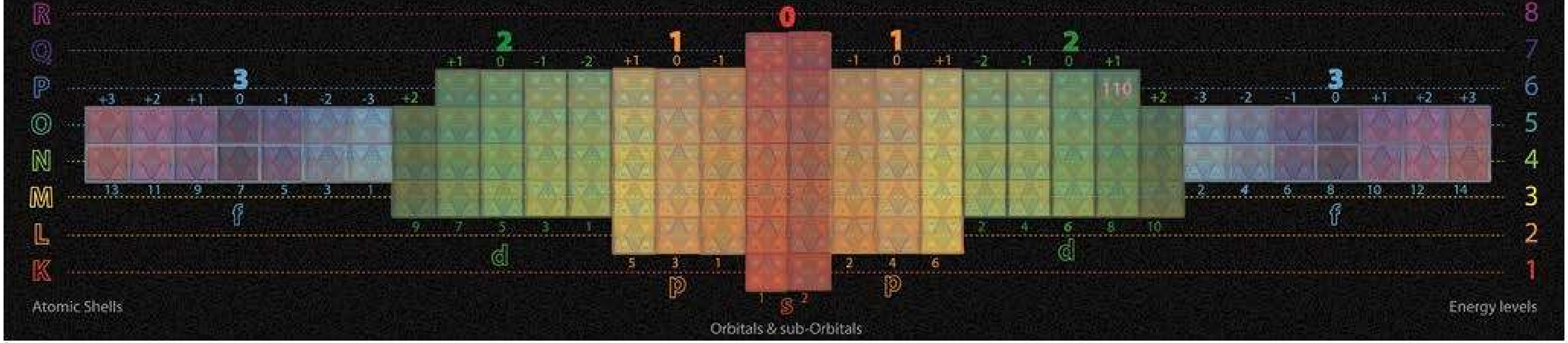
R
Q
P
O
N
M
L
K
0
2
1
1
2
3
3
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
+3 +2 +1 0 -1 -2 -3 +2
110
+2 -3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 3 1
p
1 s 2
2 4 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
8
7
6
5
4
3
2
1
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.110 - Darmstadtium atomic config

Copyright ABRAHAM [2008] - All rights reserved

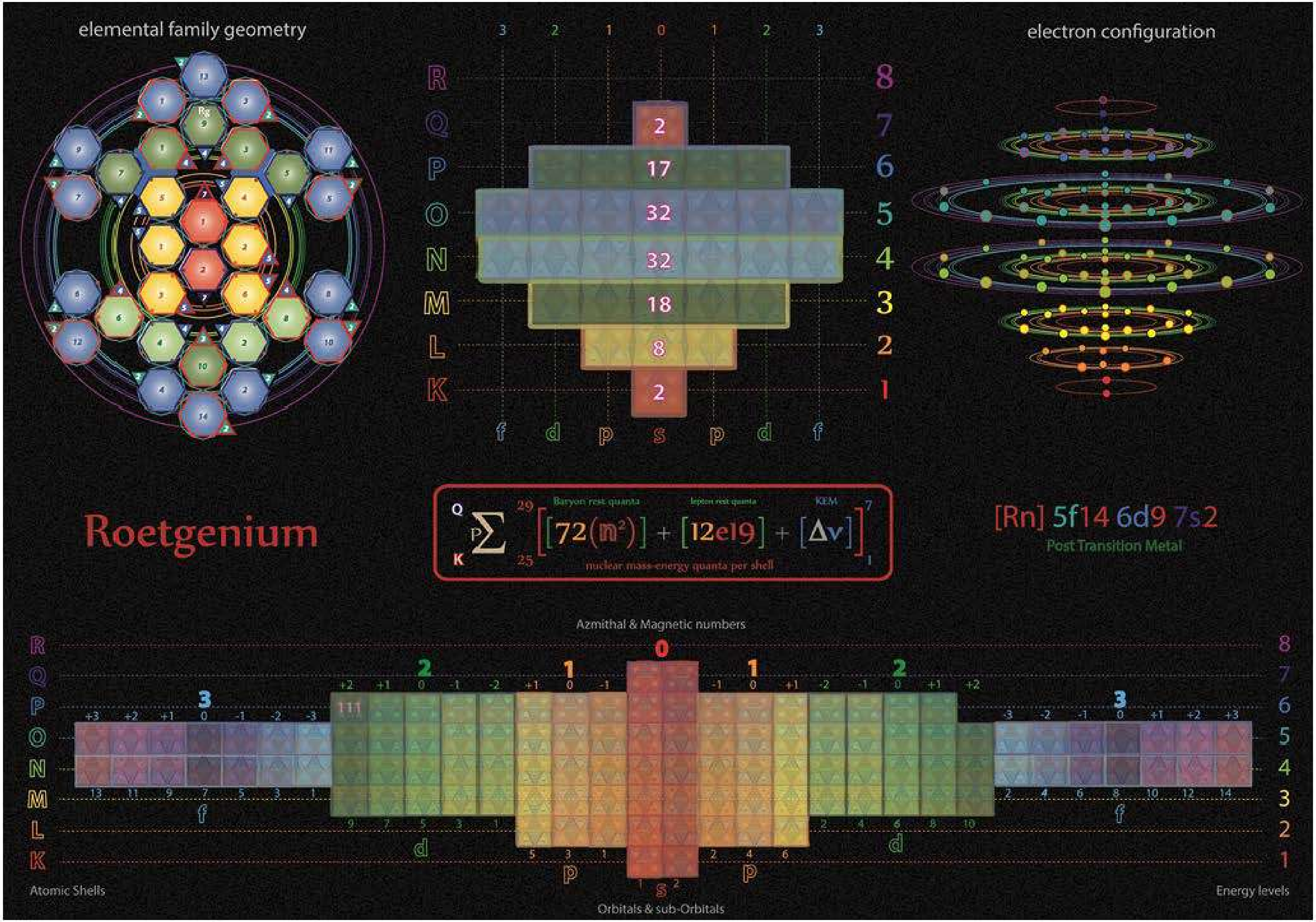
elemental family geometry
Rg
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
17
32
32
18
8
2
f d p s p d f
Roetgenium
Baryon rest quanta
lepton rest quanta
KEM
Q
P
K
29
25
7
1
$\sum_{K}^{Q} {}_{25}^{29}\left[[72(\mathrm{n}^2)] + [12e19] + [\Delta v]\right]_{1}^{7}$
nuclear mass-energy quanta per shell
[Rn] 5f14 6d9 7s2
Post Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
111
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.111 - Roetgenium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

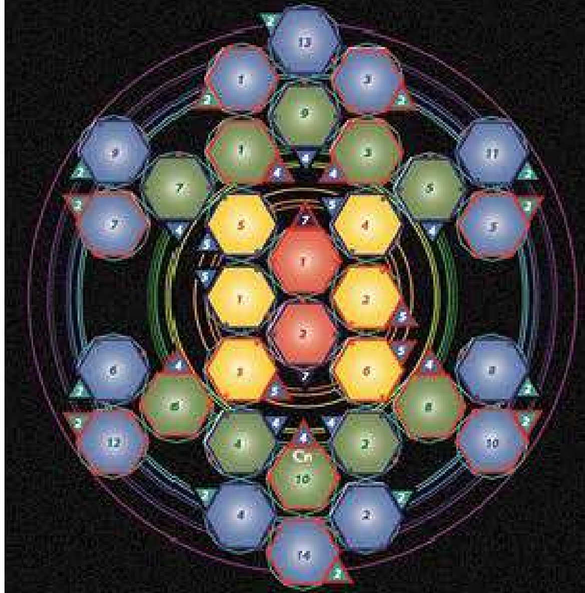

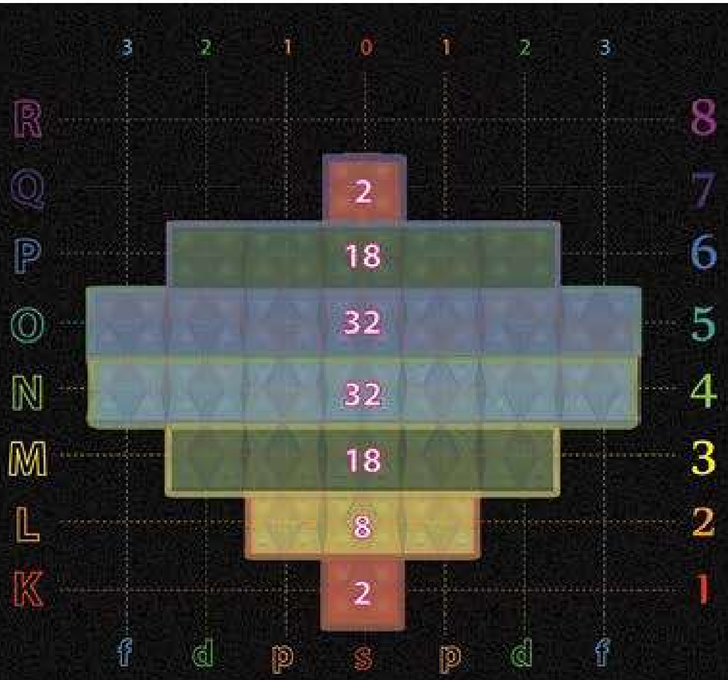
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
2
18
32
32
18
8
2
8
7
6
5
4
3
2
1
f d p s p d f

electron configuration

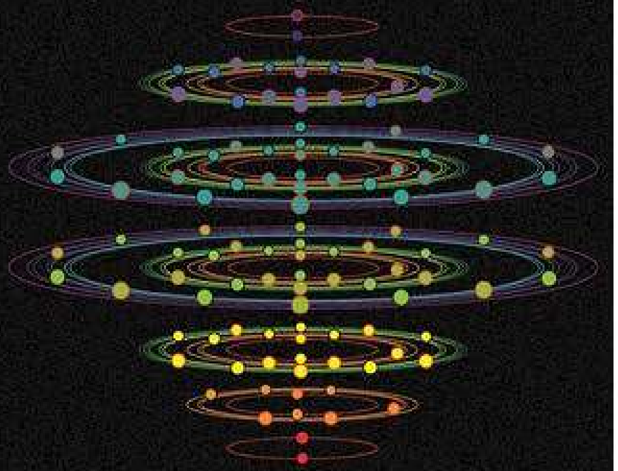

Copernicium

$$\sum_{K}^{Q} P \; {}^{29}_{25}\left[\overset{\text{Baryon rest quanta}}{[72(n^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]^{7}_{1}$$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d10 7s2

Post Transition Metal

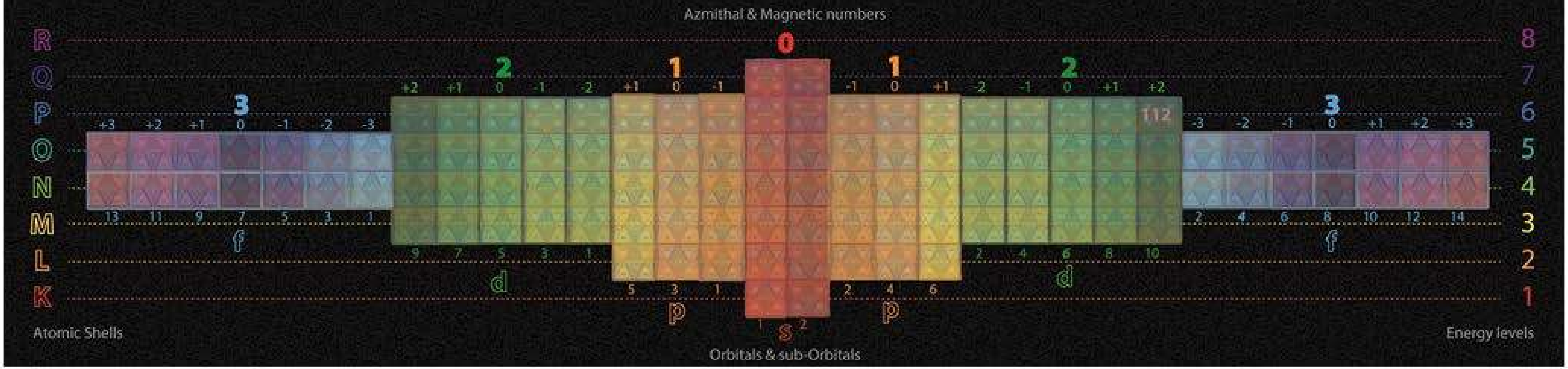
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
1
1
2
2
3
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
112
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f
d
p
s
p
d
f
8
7
6
5
4
3
2
1
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.112 - Copernicium atomic config

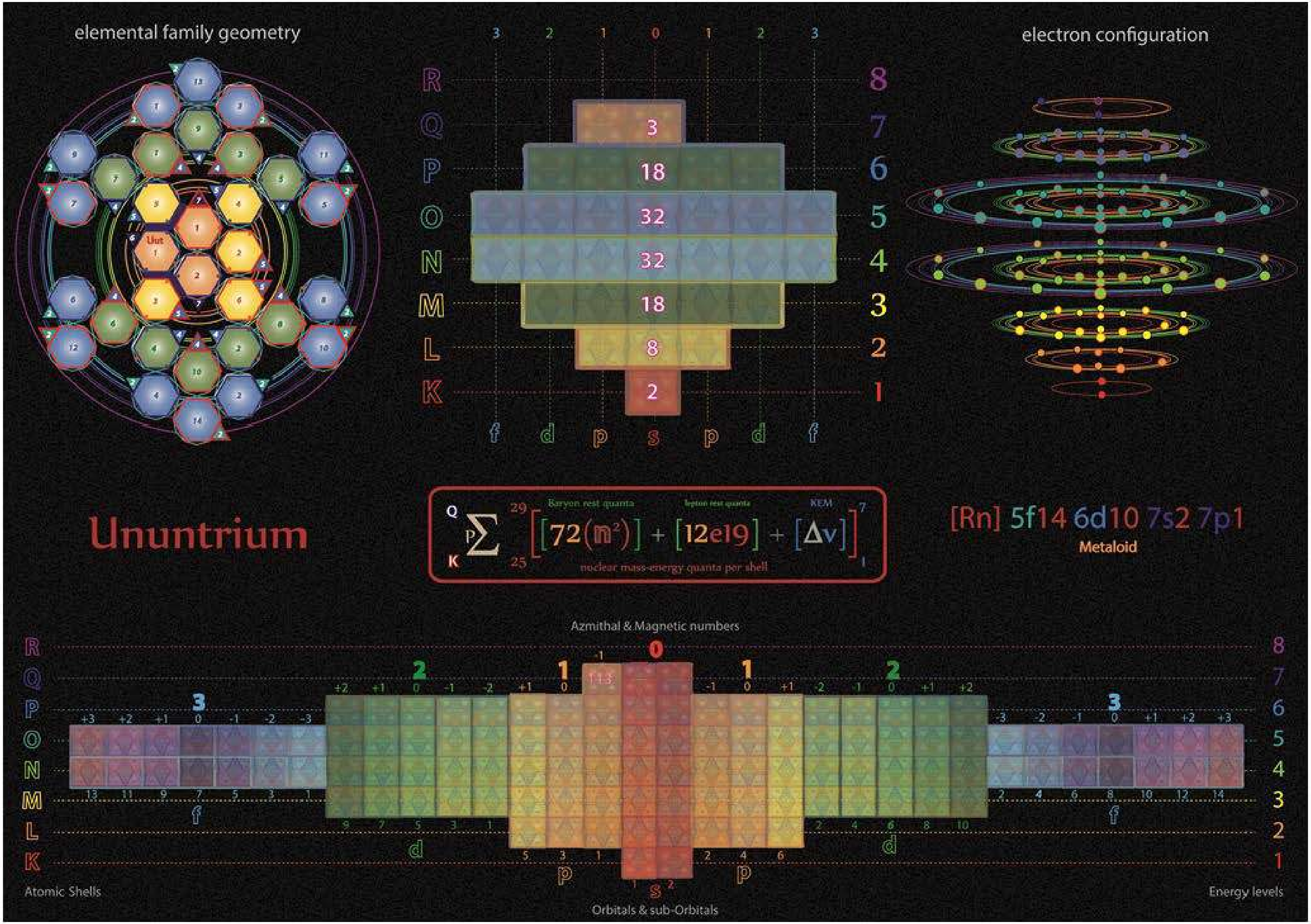
elemental family geometry
Uut
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
18
32
32
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals
113

Ununtrium

$$\sum_{K}^{Q} {}_{P} \, {}^{29}_{25}\left[\underset{\text{}}{\overset{\text{Baryon rest quanta}}{[72(\mathrm{m}^2)]}} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]^{7}_{1}$$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d10 7s2 7p1

Metaloid

Tetryonics 53.113 - Ununtrium atomic config

Copyright ABRAHAM [2008] - All rights reserved

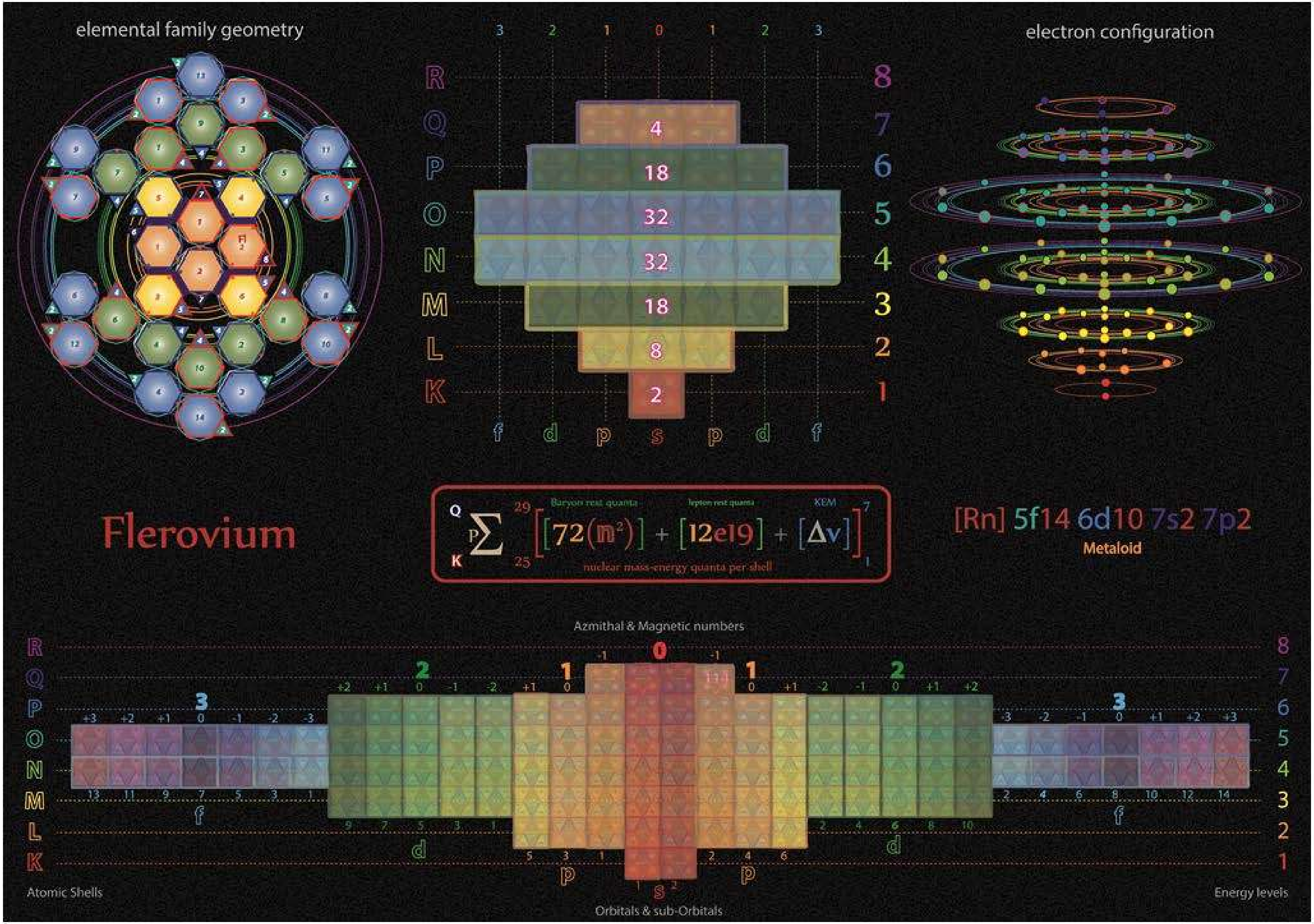
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
4
18
32
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Flerovium

$$\sum_{K}^{Q} P \; {}_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{7}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Rn] 5f14 6d10 7s2 7p2

Metaloid

Tetryonics 53.114 - Flerovium atomic config

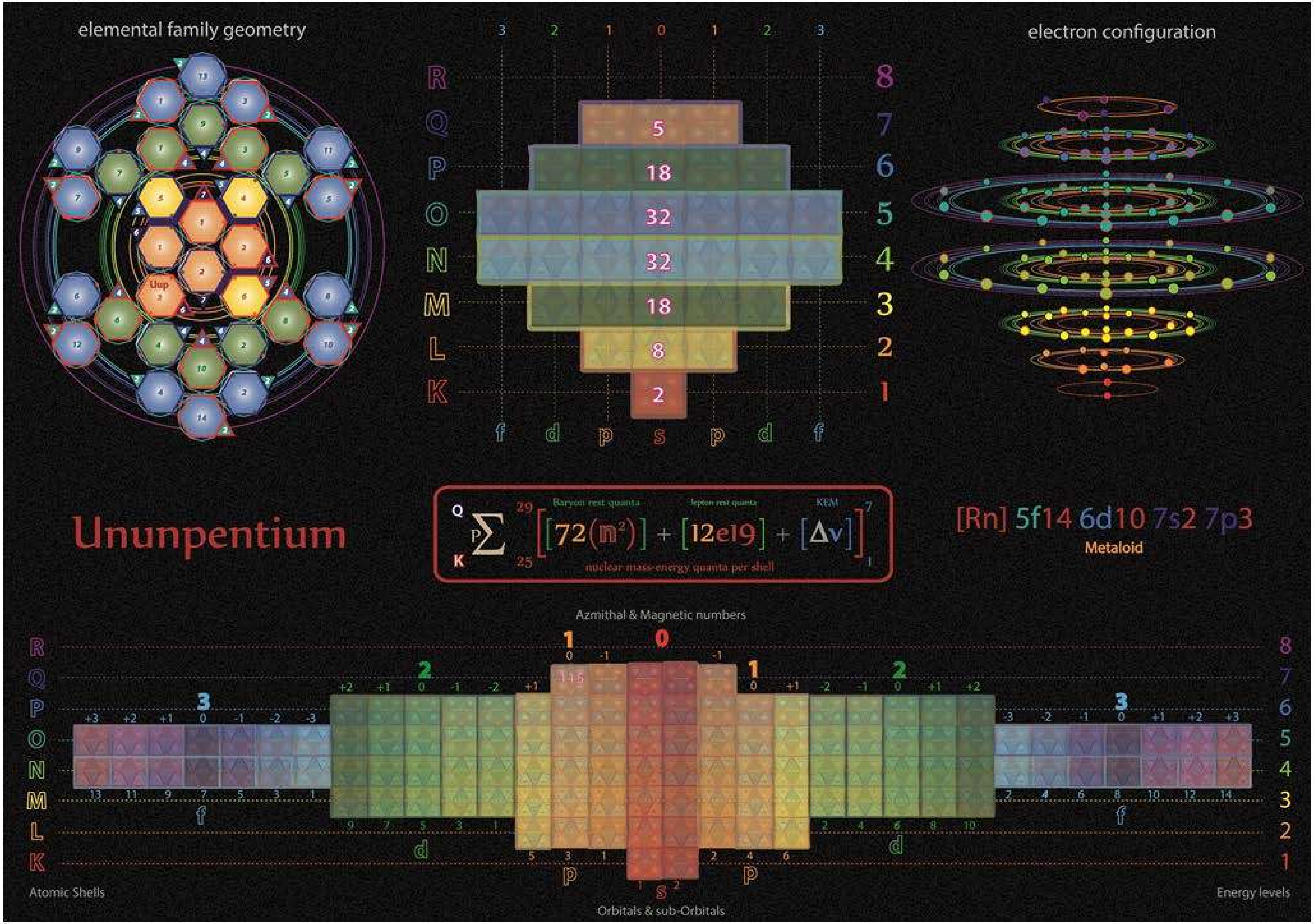
elemental family geometry
Uup
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
5
18
32
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Ununpentium

$$ {}^{Q}_{K}{}_{P}\sum_{25}^{29}\left[\underset{\text{nuclear mass-energy quanta per shell}}{\overset{\text{Baryon rest quanta}}{\left[72(n^2)\right]} + \overset{\text{lepton rest quanta}}{\left[12e19\right]} + \overset{\text{KEM}}{\left[\Delta v\right]}}\right]_{1}^{7} $$

[Rn] 5f14 6d10 7s2 7p3

Metaloid

Tetryonics 53.115 - Ununpentium atomic config

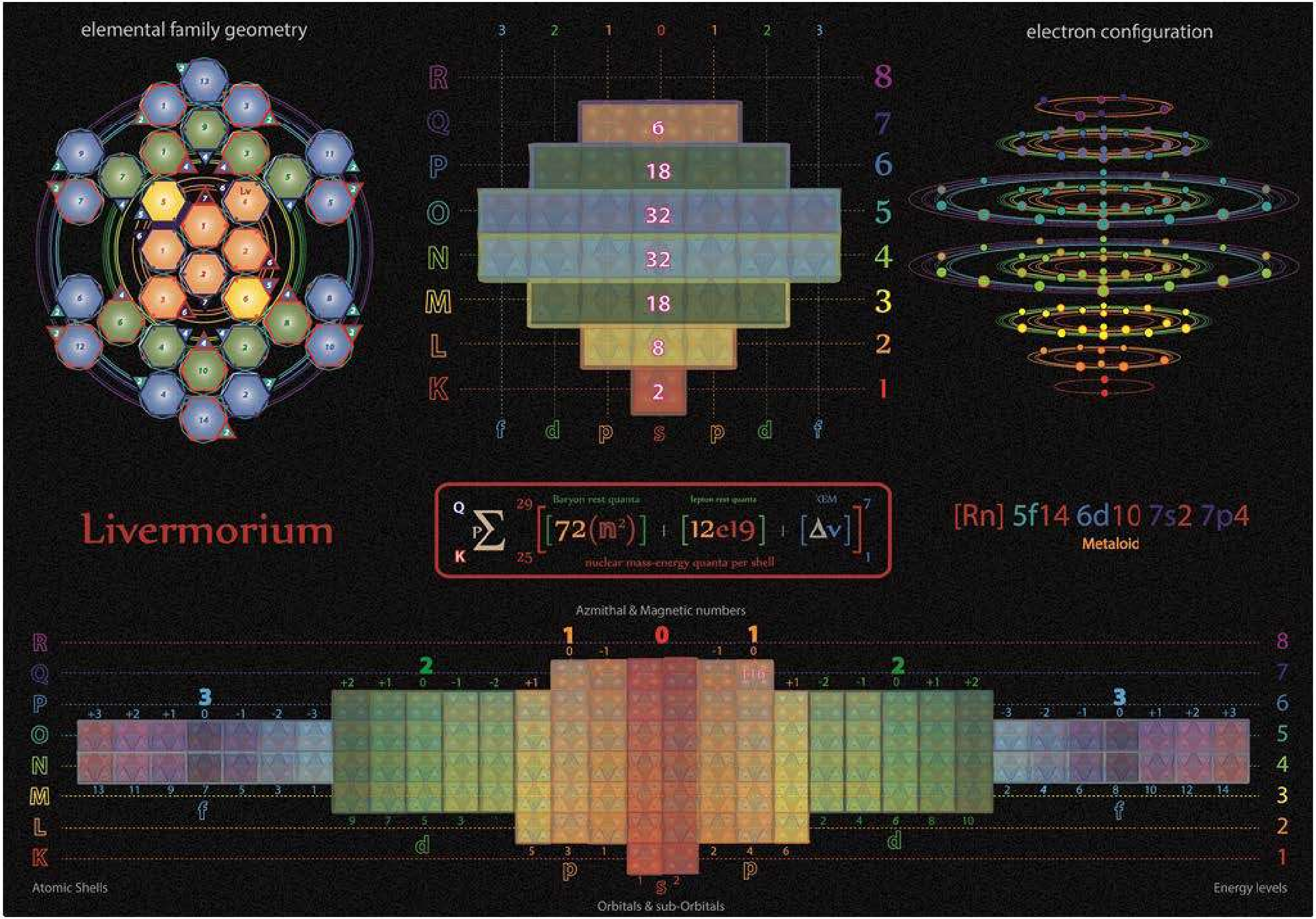
elemental family geometry
Lv
3 2 1 0 1 2 3
R Q P O N M L K
6
18
32
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Livermorium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
CEM
[72(n²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f14 6d10 7s2 7p4
Metaloid
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
116
13 11 9 7 5 3 1
9 7 5 3
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.116 - Livermorium atomic config

elemental family geometry

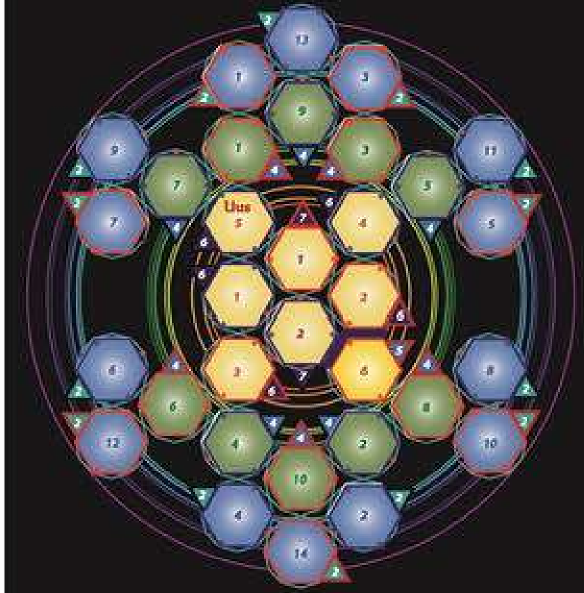

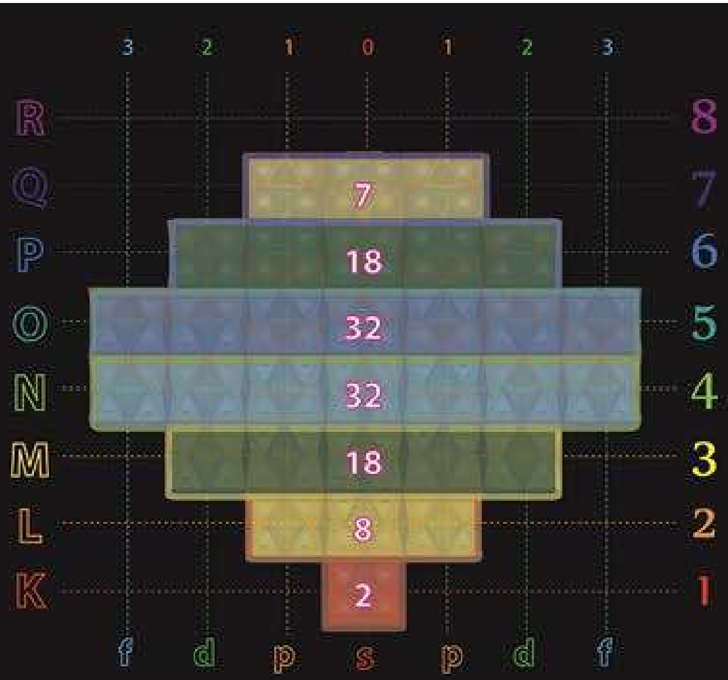
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
7
18
32
32
18
8
2
f d p s p d f

electron configuration

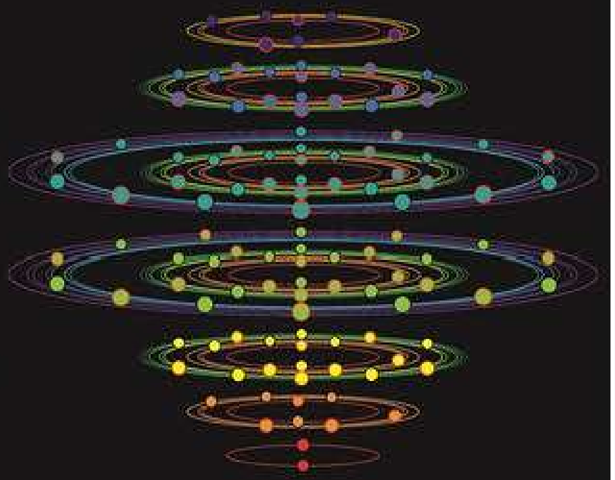

Unbinilium

Baryon rest quanta · lepton rest quanta · KEM

$$ {}^{Q}_{K}P\sum_{25}^{29}\left[72(m^2)\right]+\left[12e19\right]+\left[\Delta v\right]_{1}^{7} $$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d10 7s2 7p5

Halogen

Azmithal & Magnetic numbers

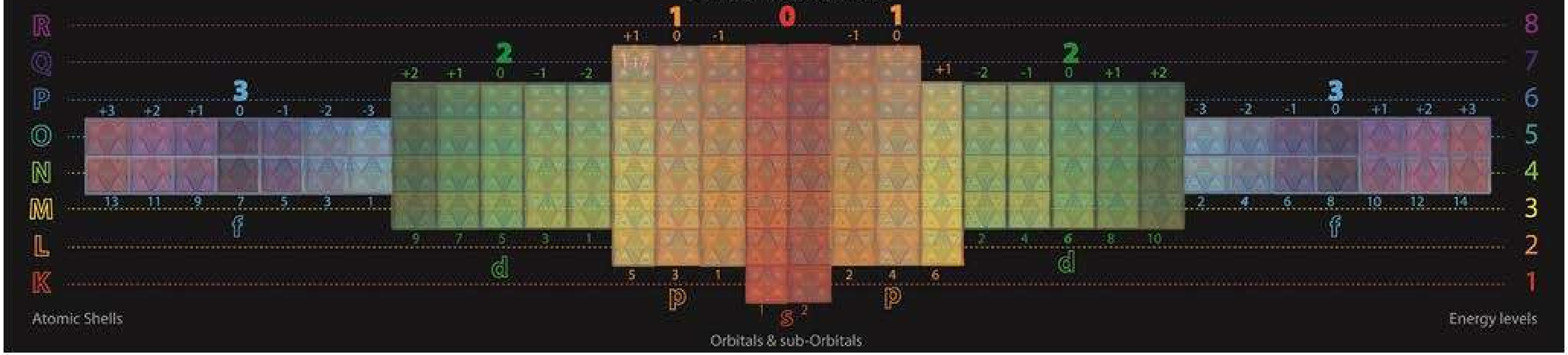
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.117 - Ununseptium atomic config

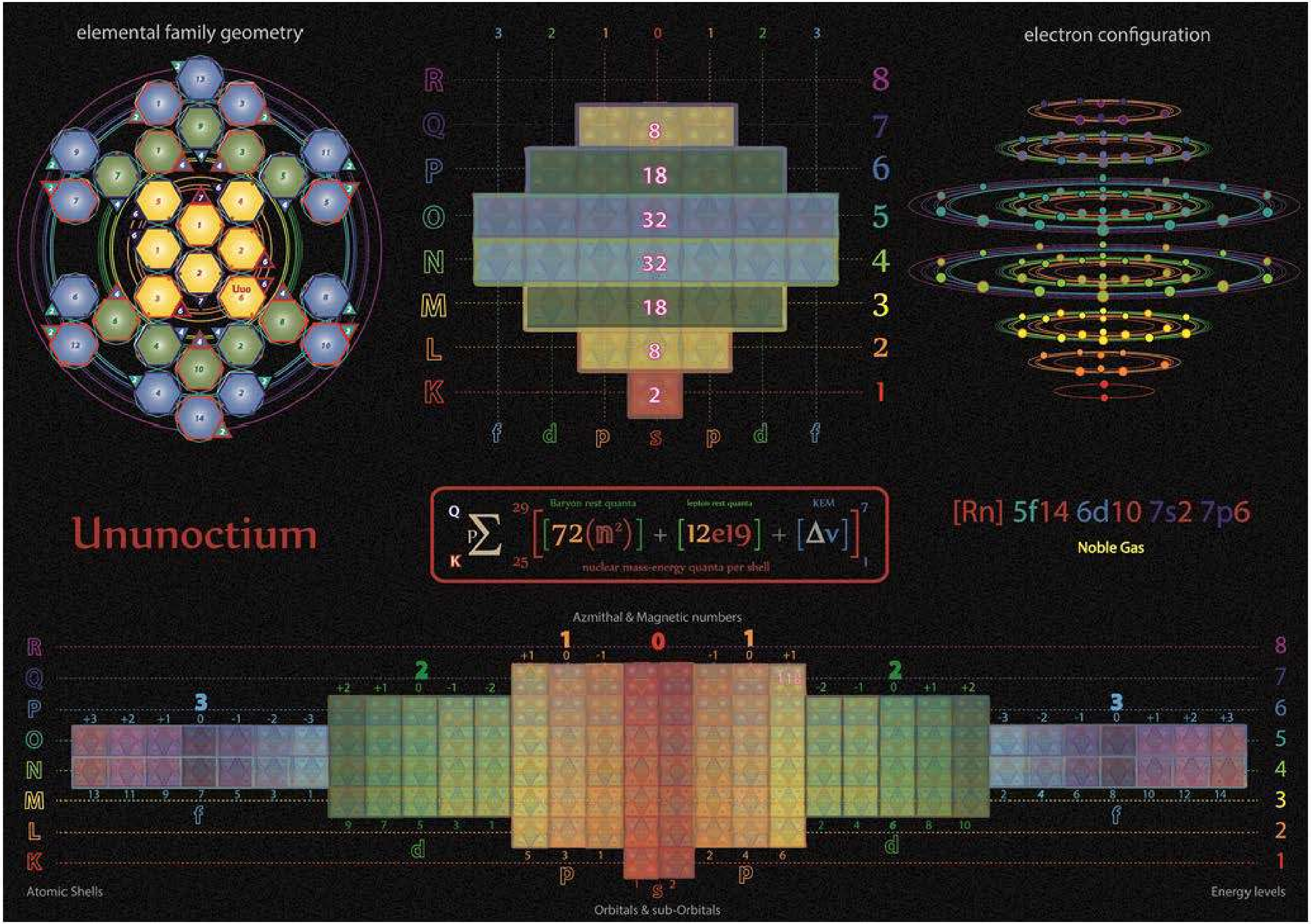

elemental family geometry

electron configuration

Ununoctium

$$\sum_{K}^{Q} \mathrm{P} \; {}^{29}_{25}\left[\underbrace{\overset{\text{Baryon rest quanta}}{[72(n^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}}_{\text{nuclear mass-energy quanta per shell}} \right]^{7}_{1}$$

[Rn] 5f14 6d10 7s2 7p6

Noble Gas

Azmithal & Magnetic numbers

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Uuo
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
8
18
32
32
18
8
2
f
d
p
s
p
d
f
1
0
1
2
2
3
3
+1
0
-1
-1
0
+1
+2
+1
0
-1
-2
-2
-1
0
+1
+2
+3
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+1
+2
+3
118
13
11
9
7
5
3
1
2
4
6
8
10
12
14
9
7
5
3
1
2
4
6
8
10
5
3
1
2
4
6
1
2

Tetryonics 53.118 - Ununoctium atomic config

Copyright ABRAHAM [2008] - All rights reserved

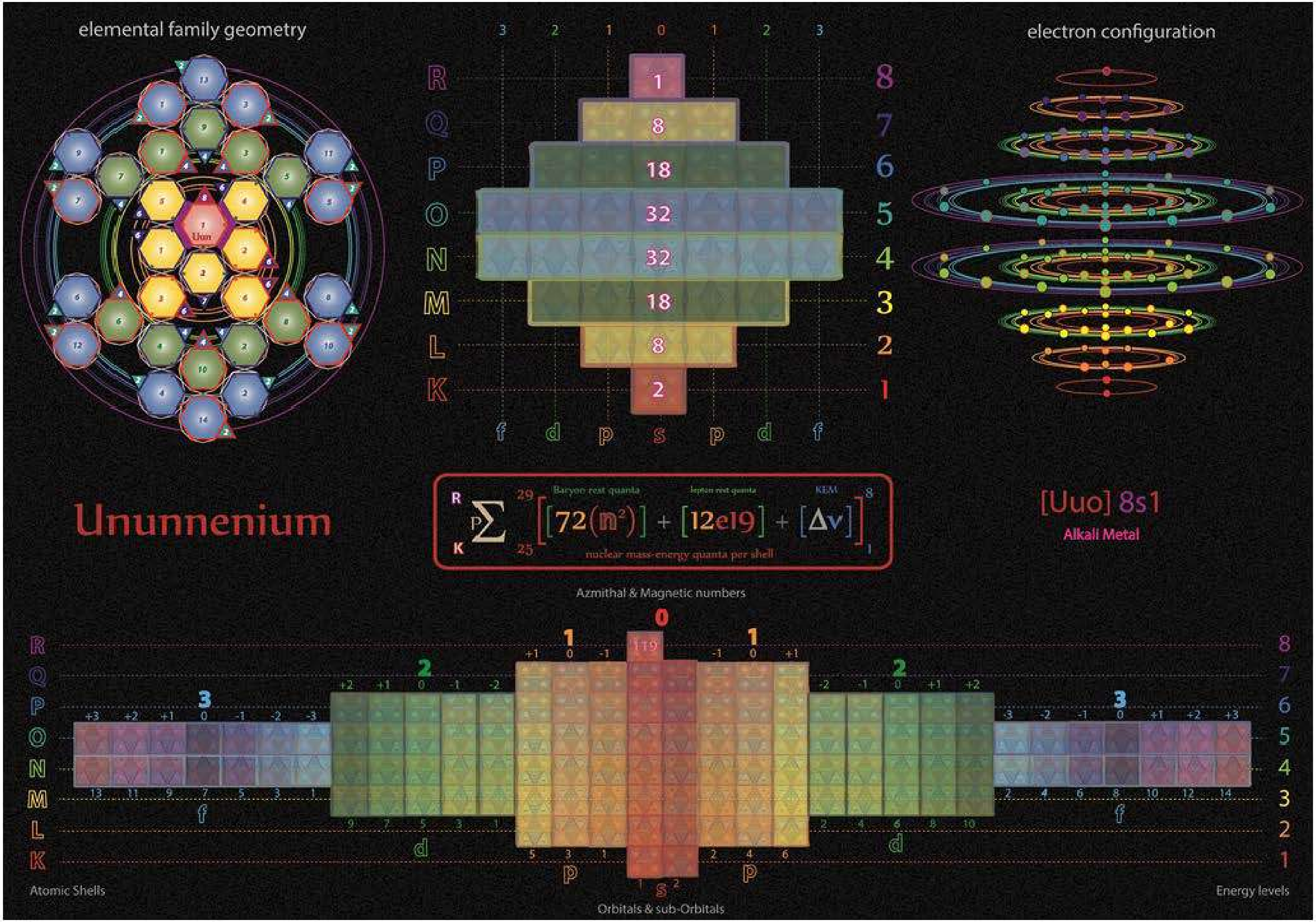
elemental family geometry
Uun
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
1
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Ununnenium
R
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
8
1
nuclear mass-energy quanta per shell
[Uuo] 8s1
Alkali Metal
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
119
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 3 1
p
1 s 2
2 4 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.119 - Ununnenium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

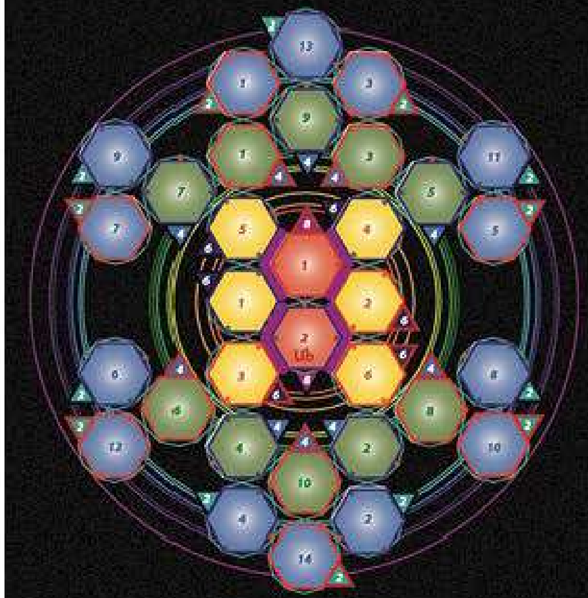
13
11
9
7
5
3
1
2
4
6
8
10
12
14
Ub

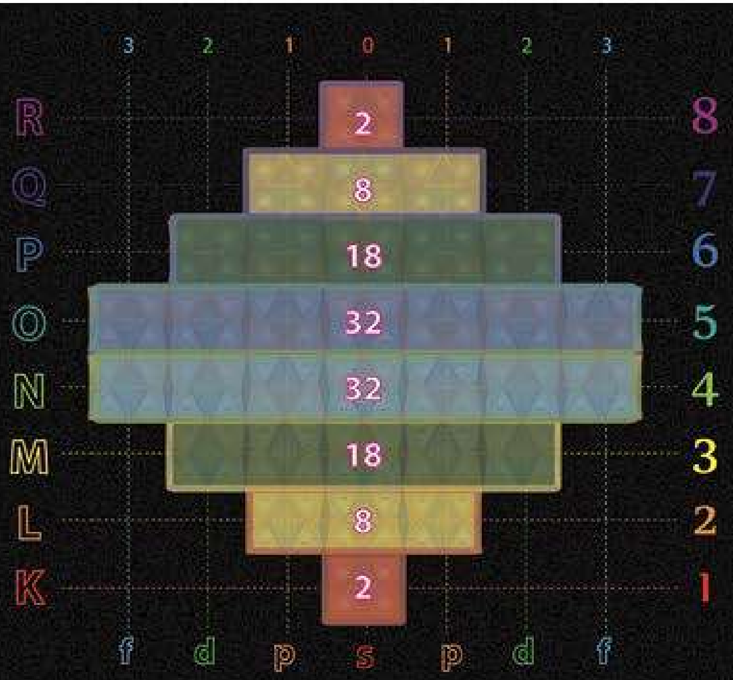
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

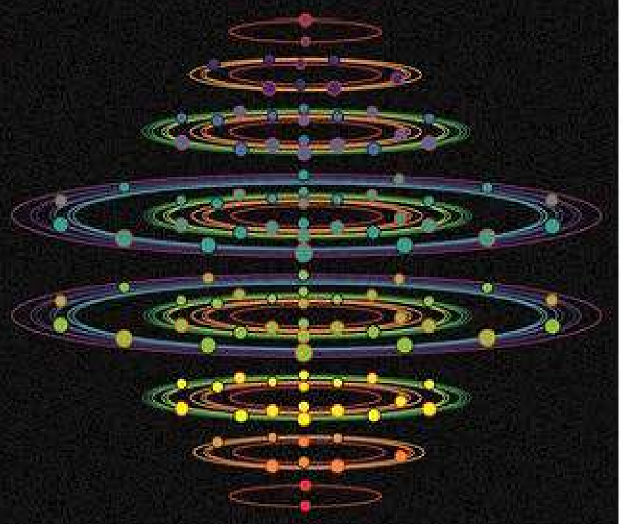

Unbinilium

$$\sum_{K}^{R}{}_{P}\;{}^{29}_{25}\left[\overset{\text{Baryon rest quanta}}{[72(n^2)]}+\overset{\text{lepton rest quanta}}{[12e19]}+\overset{\text{KEM}}{[\Delta v]}\right]^{8}_{1}$$

nuclear mass-energy quanta per shell

[Uuo] 8s2

Alkaline Earth

Azmithal & Magnetic numbers

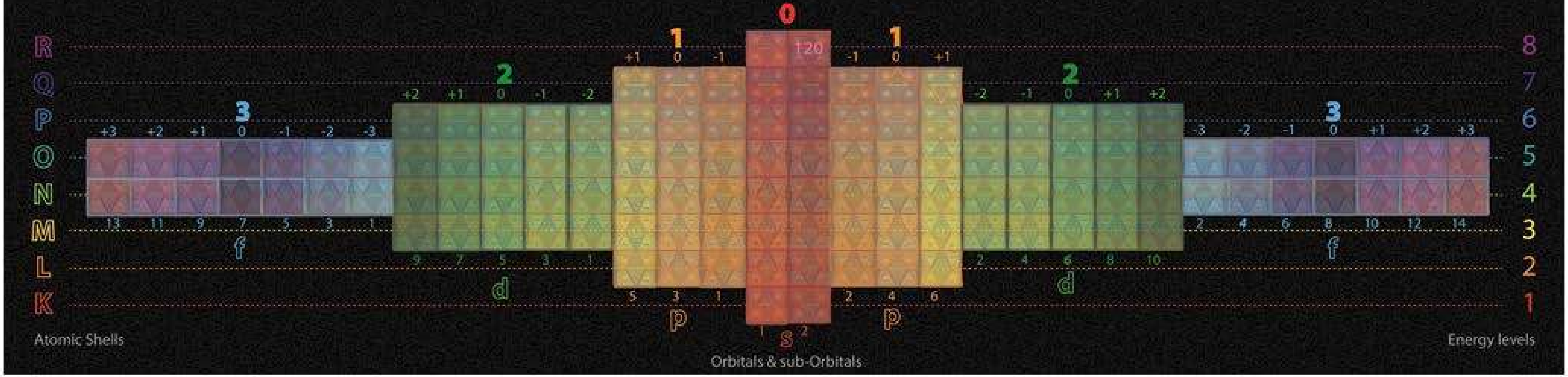
0
120
1
1
+1
0
-1
-1
0
+1
2
2
+2
+1
0
-1
-2
-2
-1
0
+1
+2
3
3
+3
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+1
+2
+3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
13
11
9
7
5
3
1
2
4
6
8
10
12
14
f
f
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.120 - Unbinilium atomic config

Chemical element bonding

All Deuterium nuclei pack in hexagonal patterns
to form larger atoms & molecules

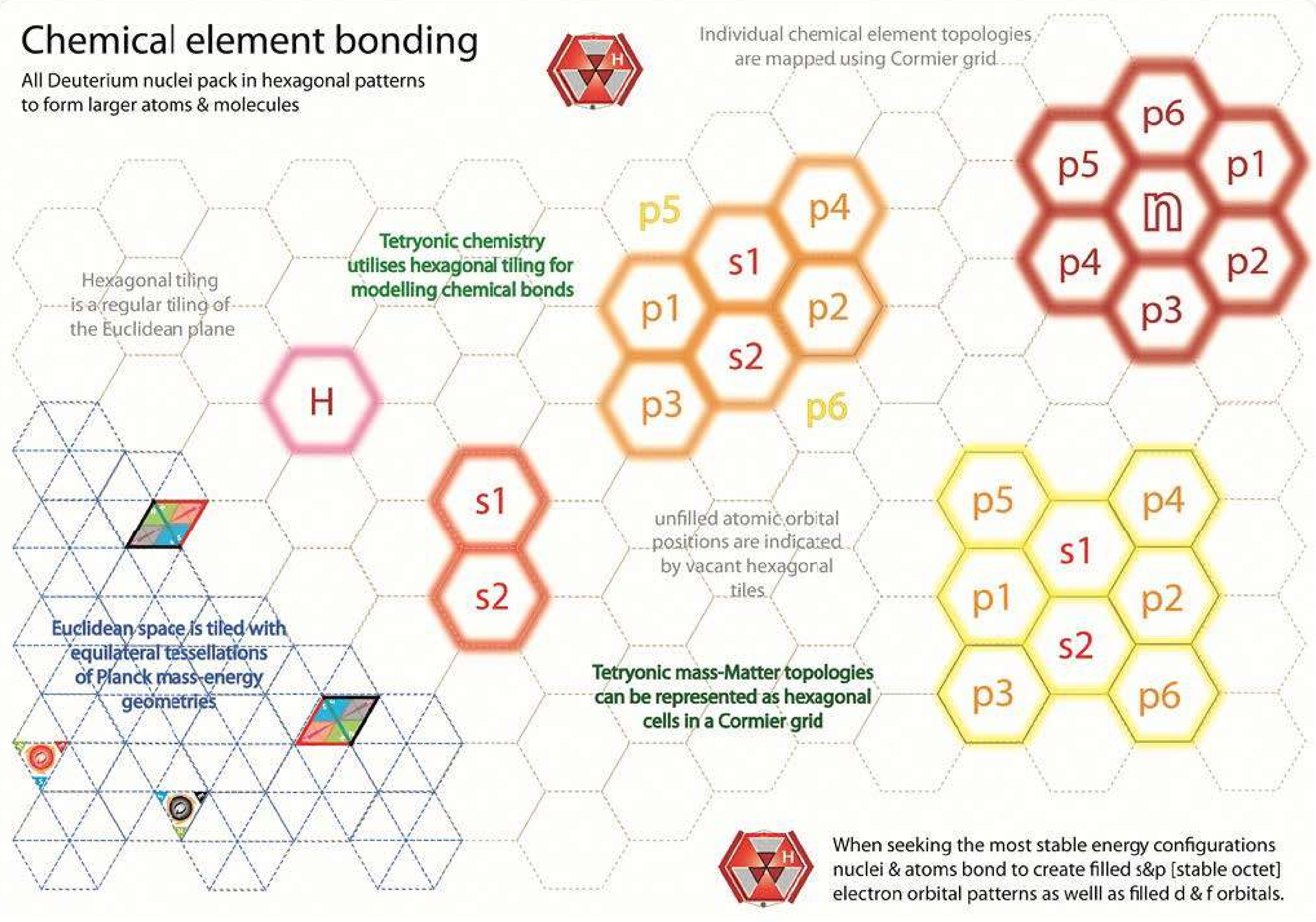

Individual chemical element topologies
are mapped using Cormier grid

Hexagonal tiling
is a regular tiling of
the Euclidean plane

Tetryonic chemistry
utilises hexagonal tiling for
modelling chemical bonds

Euclidean space is tiled with
equilateral tessellations
of Planck mass-energy
geometries

unfilled atomic orbital
positions are indicated
by vacant hexagonal
tiles

Tetryonic mass-Matter topologies
can be represented as hexagonal
cells in a Cormier grid

When seeking the most stable energy configurations
nuclei & atoms bond to create filled s&p [stable octet]
electron orbital patterns as welll as filled d & f orbitals.

Periodic Table Atomic configurations

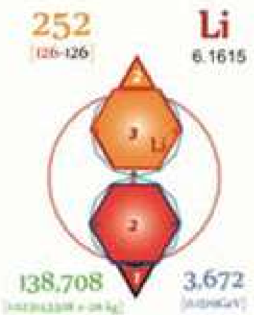
252
[126-126]
Li
6.1615
Li
138,708
3,672

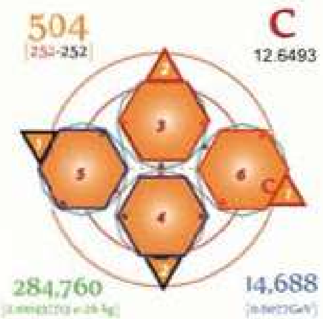
504
[252-252]
C
12.6493
C
284,760
14,688

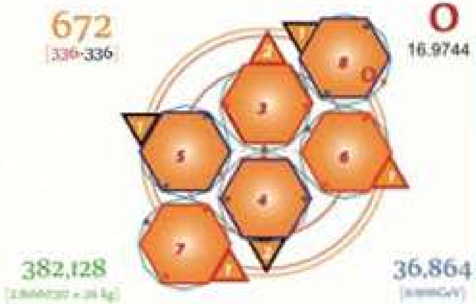
672
[336-336]
O
16.9744
O
382,128
36,864

Allotropes are different structural forms of the same element and can exhibit quite different physical properties and chemical behaviours

Allotropes

The Periodic Table, although useful in identifying Elements via their atomic and quantum numbers, does not reflect all the charged topologes that Deuterium nuclei can form as they combine.

Various atomic configurations with the same Tetryonic charge, but differing in their final mass-Matter topologies and properties can be formed - they are the Allotropes

Some of the more plentiful chemical elements form this way

Lithium

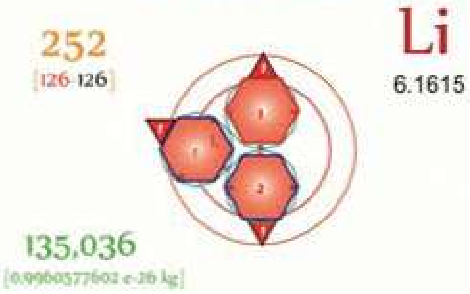
252
[126-126]
Li
6.1615
135,036
[0.99605577602 e-26 kg]

Oxygen

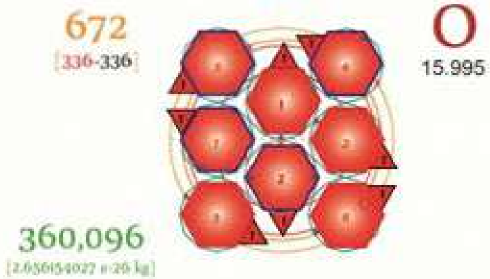
672
[336-336]
O
15.995
360,096
[2.656154027 e-26 kg]

Carbon

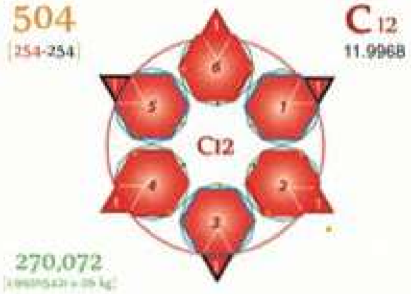
504
[254-254]
C12
11.9968
C12
270,072

6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1

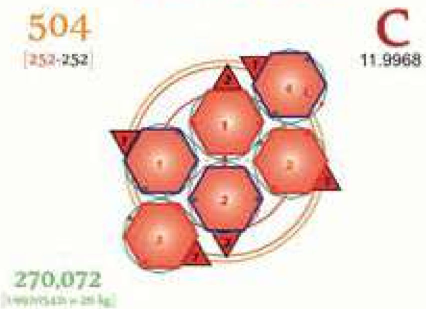
504
[252-252]
C
11.9968
270,072

6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1

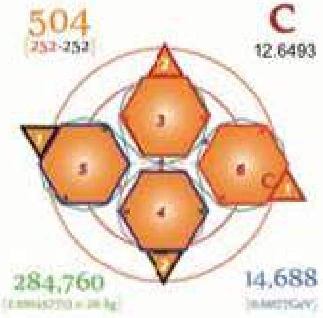
504
[252-252]
C
12.6493
C
284,760
14,688

6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2

Allotropes are elements created from the same number of Deuterium nuclei as periodic elements but possess a differing mass-Matter topology

Q

same component charge

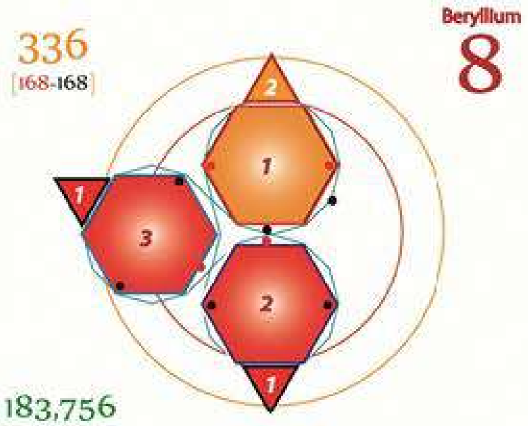
336
[168-168]
Beryllium
8
2
1
3
2
1
1
183,756

different Matter geometries

M

The varying material geometries allows what is the same chemical element to possess vastly different bonding points and chemical attributes

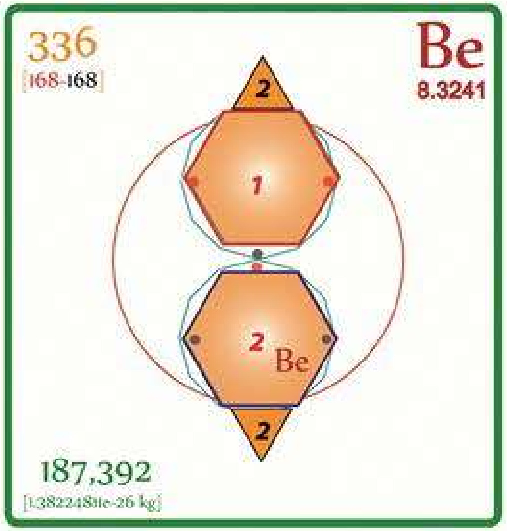
336
[168-168]
Be
8.3241
2
1
2
Be
2
187,392
[1.38224811e-26 kg]

Allotropic geometries

[charge vs Matter]

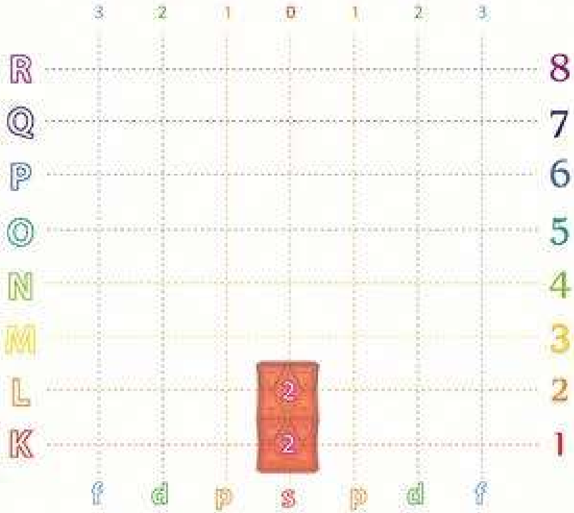
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
2
f
d
p
s
p
d
f

Some element allotropes have different Matter topologies that persist in different phases

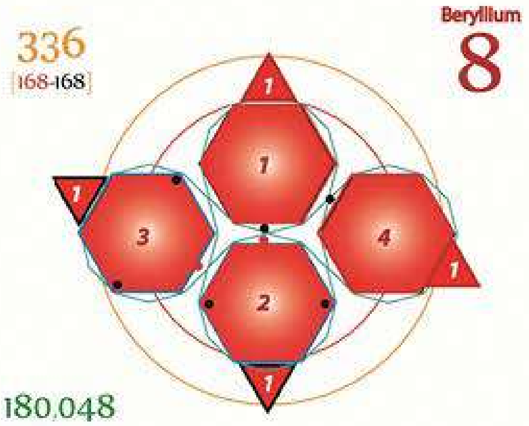
336
[168-168]
Beryllium
8
1
1
1
3
4
2
1
1
180,048

Allotropes vs. Isotopes

Isotopes are elementary atoms with the same number of nuclei, but with differing energy levels, resulting in different mass-energies

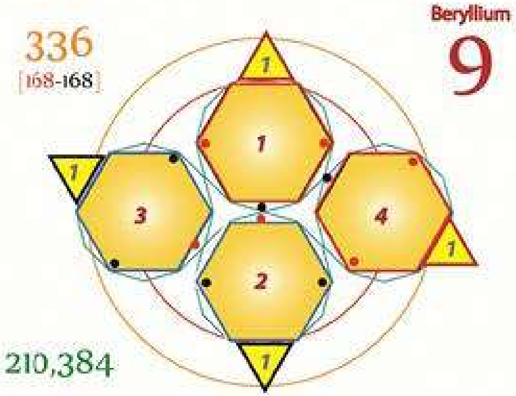
336
[168-168]
Beryllium
9
1
1
1
3
4
2
1
1
210,384

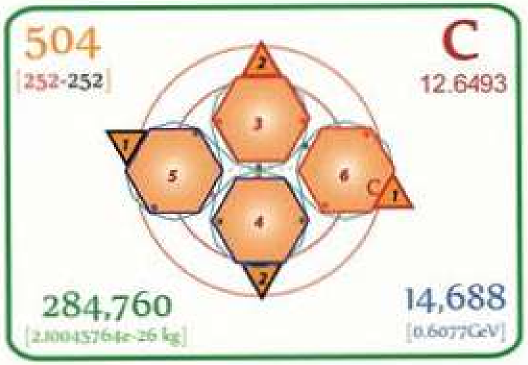
504
[252-252]
C
12.6493
3
5
6
4
2
1
C
284,760
[2.80045764e-26 kg]
14,688
[0.6077GeV]

Carbon Allotropes

Carbon is capable of forming many allotropes due to its valency.

Well known forms of carbon include:

Carbon
Diamond
Graphite
Graphene
Amorphous carbon
Buckminsterfullerenes
Carbon nanotubes
Glassy carbon
atomic & diatomic carbon

Tetryonic theory affords us the ability to model the charged mass-energy geometries & 3D Matter topologies of each element along with its bonding points

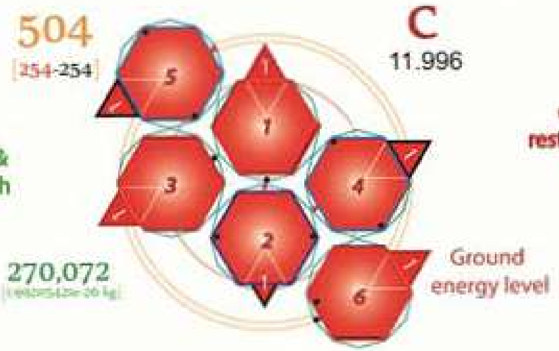
504
[254-254]
C
11.996
5
1
3
4
2
6
270,072
Ground energy level

It also allows for the calculation of the molar rest mass-energy geometry of any Matter topology

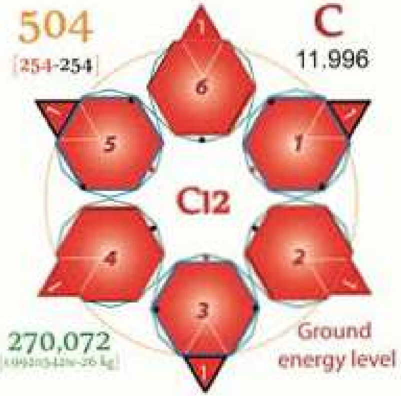
504
[254-254]
C
11.996
6
5
1
C12
4
2
3
1
270,072
Ground energy level

Graphene has a unique p orbital arrangement of its electrons

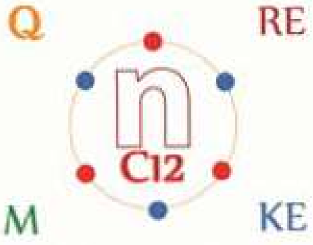
Q
RE
n
C12
M
KE

Electron Configuration

[n] p6

Shell			Molar meM	Nuclei per shell
R	8	446,976	19.855	6
Q	7	418,680	18.598	6
P	6	391,392	17.385	6
O	5	365,112	16.218	6
N	4	339,840	15.095	6
M	3	315,576	14.018	6
L	2	292,320	12.985	6
K	1	270,072	11.996	6

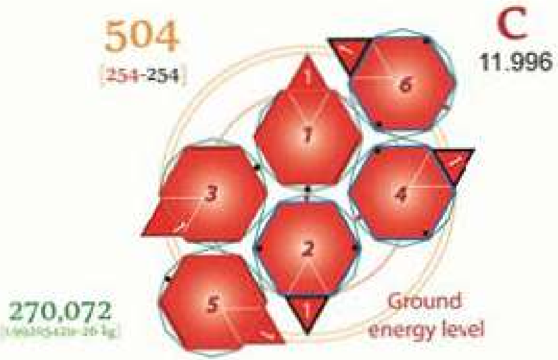
504
[254-254]
C
11.996
6
1
3
4
2
5
270,072
Ground energy level

Carbon-14, a radioactive isotope of carbon with a half-life of 5,730 years, is used to find the age of formerly living things through a process known as radiocarbon dating.

In 1961 the international unions of physicists and chemists agreed to use the mass of the isotope carbon-12 as the basis for atomic weight.

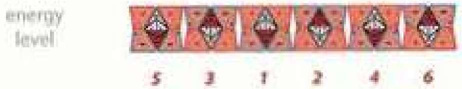
energy level
5
3
1
2
4
6

There are nearly ten million known carbon compounds and an entire branch of chemistry, known as organic chemistry, is devoted to their study.

Graphene

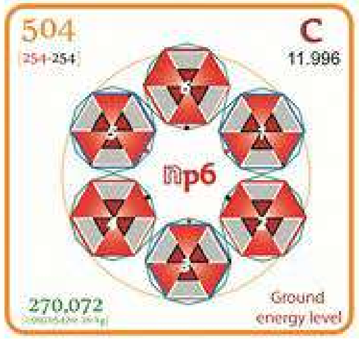

Graphene is also an allotrope of carbon.

Its structure is one-atom-thick planar sheets of sp2-bonded carbon atoms that are densely packed in a honeycomb crystal lattice

Graphite

The mineral graphite is an allotrope of carbon.

Unlike periodic table carbon [diamond], graphite is an electrical conductor, a semi-metal.

And Graphite is the most stable form of carbon under standard conditions.

Graphite has a layered, planar structure. In each layer, the carbon atoms are arranged in a hexagonal lattice

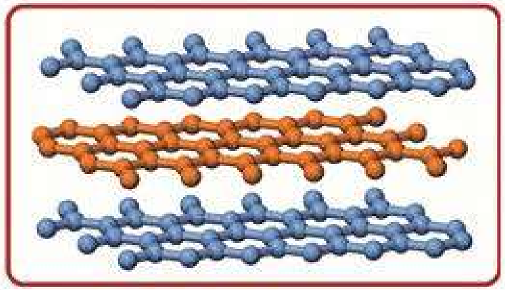

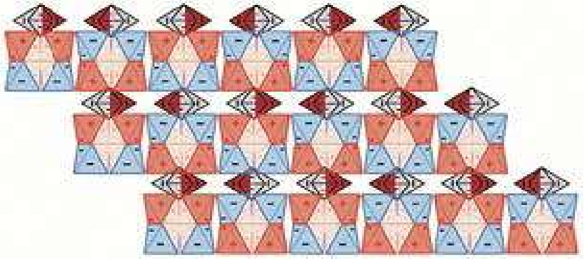

Graphite electrons can act as atomic bearings between the planar sheets of graphene

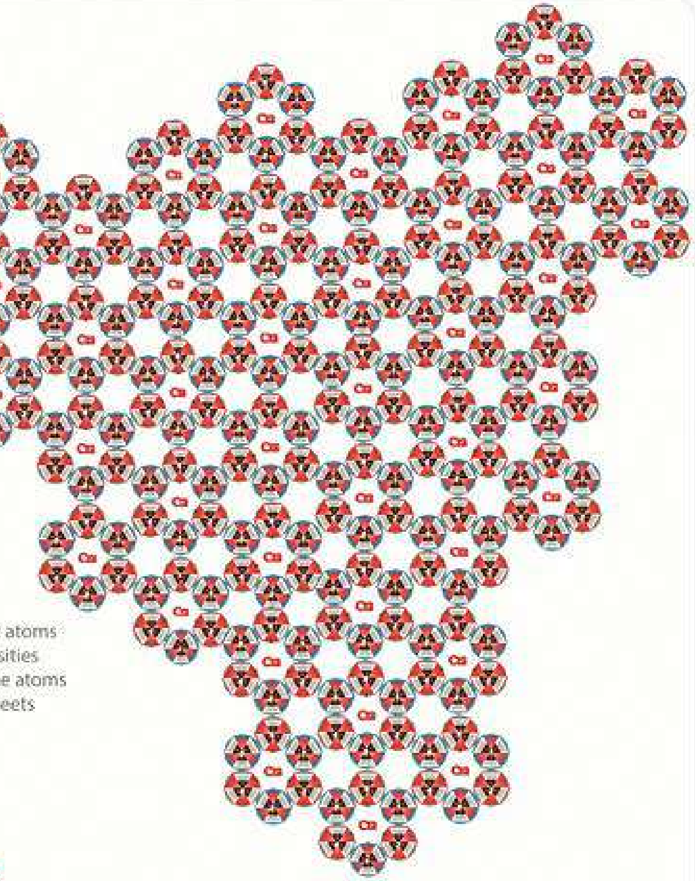

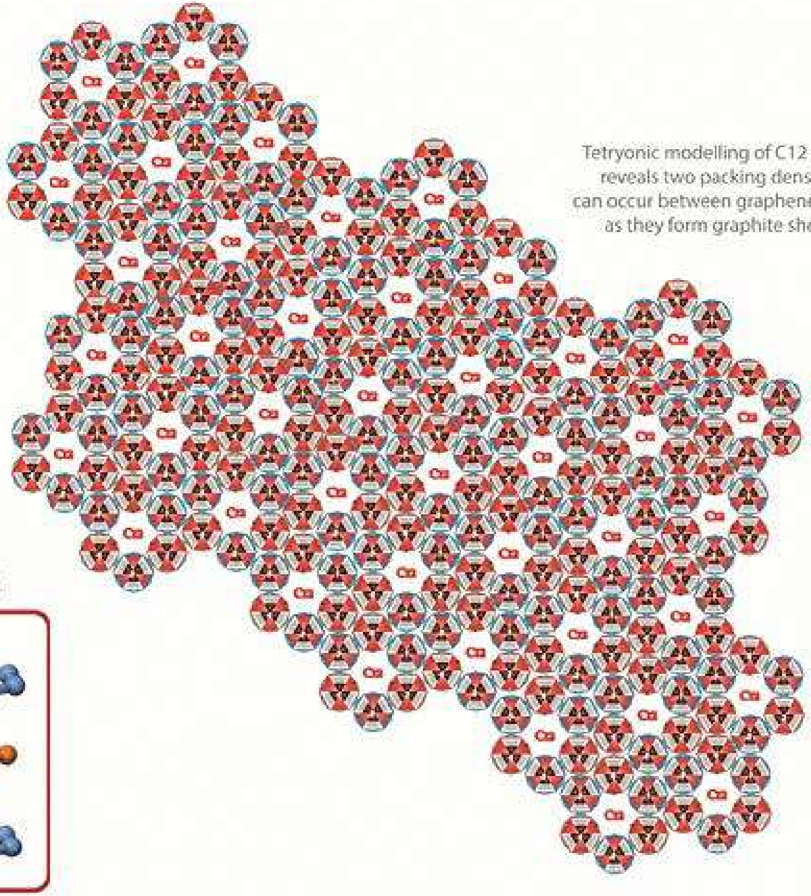

Tetryonic modelling of C12 atoms reveals two packing densities can occur between graphene atoms as they form graphite sheets

Unlike periodic carbon where electrons are tightly bound, graphite is highly conductive, with electrical energy being propagated via the electrons & their KEM fields

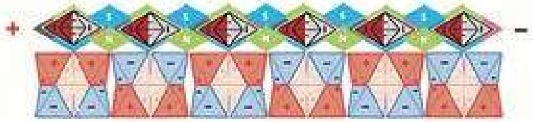

Graphene electrons are loosley bound to the top of the Deuterium nuclei

Copyright ABRAHAM [2008] - All rights reserved

Carbon Nanotubes

Carbon nanotubes (CNTs) are allotropes of carbon
with a cylindrical nanostructure.

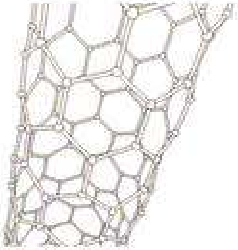

Large molecules consisting only of carbon, known as buckminsterfullerenes, or buckyballs, have recently been discovered and are currently the subject of much scientific interest.

A single buckyball consists of 60 or 70 carbon atoms (C60 or C70) linked together in a structure that looks like a soccer ball.

They can trap other atoms within their framework, appear to be capable of withstanding great pressures and have magnetic and superconductive properties.

C36

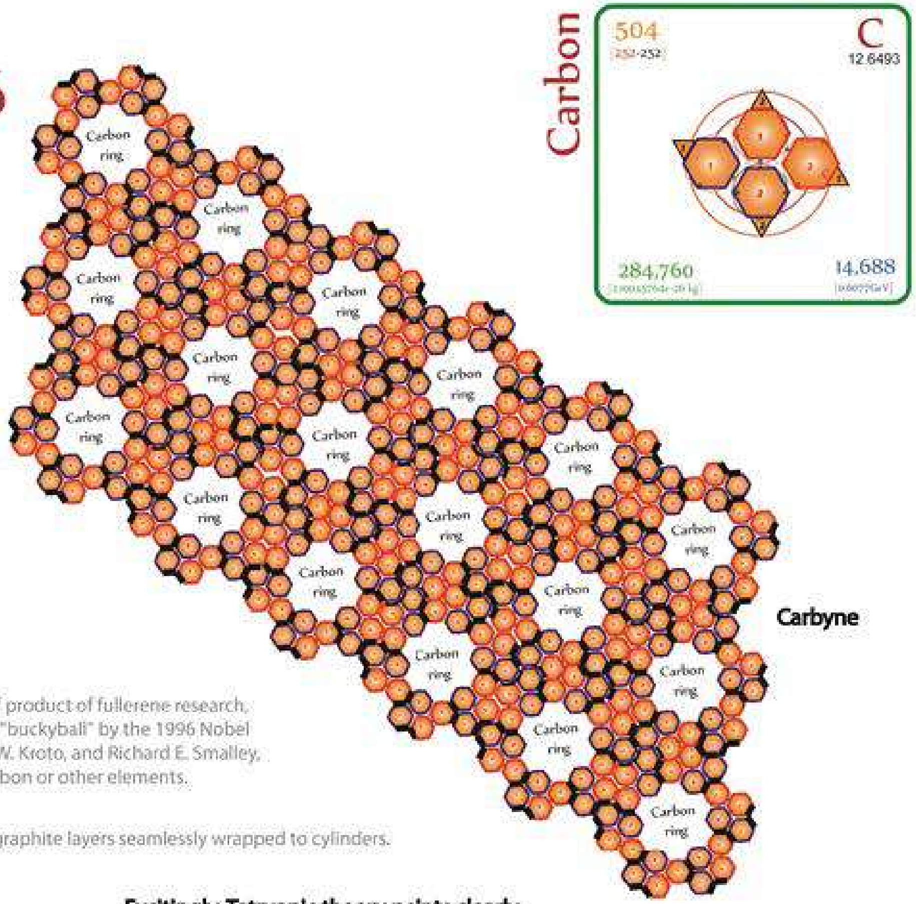

They can be thought of as a sheet of graphite
(a hexagonal lattice of carbon) rolled into a cylinder.

C12

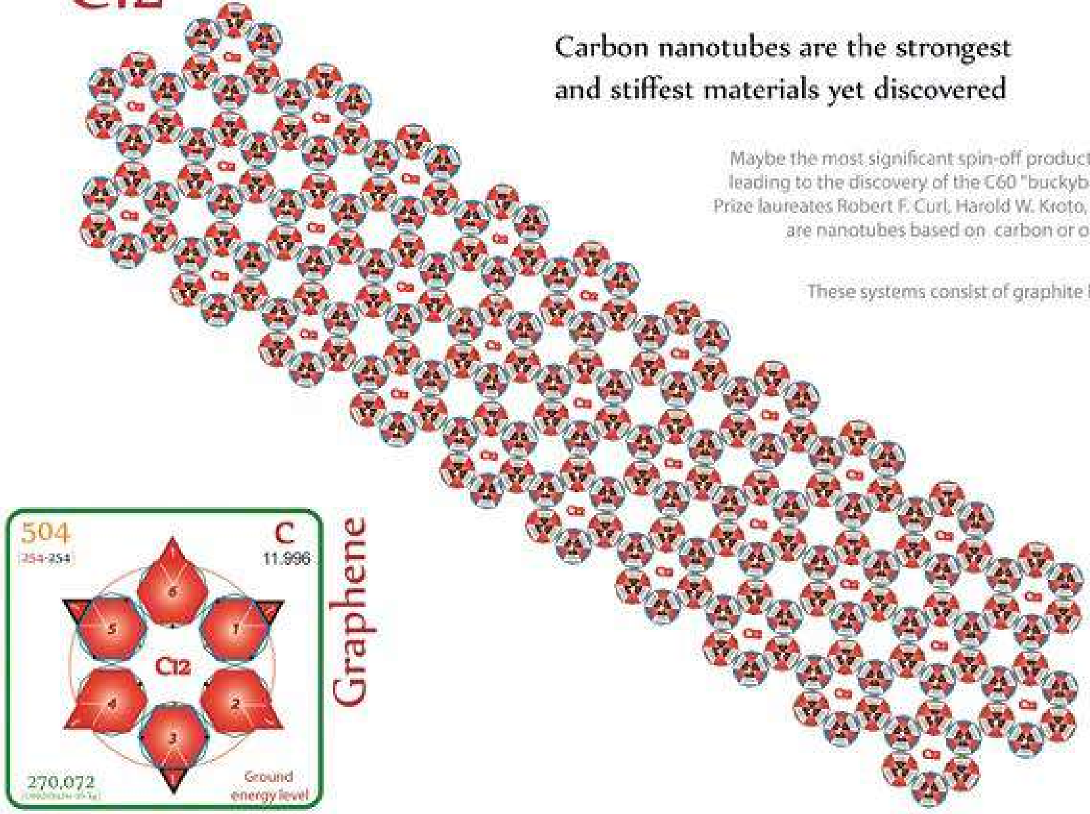

Carbon nanotubes are the strongest and stiffest materials yet discovered

Maybe the most significant spin-off product of fullerene research, leading to the discovery of the C60 "buckyball" by the 1996 Nobel Prize laureates Robert F. Curl, Harold W. Kroto, and Richard E. Smalley, are nanotubes based on carbon or other elements.

These systems consist of graphite layers seamlessly wrapped to cylinders.

Excitingly, Tetryonic theory points clearly to the possibility that organo-carbon nanotubes can be created.

Where C36 periodic carbon could be used insetad of conventional C12 graphene atoms

Carbyne

C60

Nanotubes are members of the fullerene structural family, which also includes the spherical buckyballs, and the ends of a nanotube may be capped with a hemisphere of the buckyball structure.

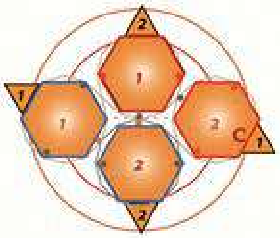

Periodic [C6] Carbon

Diamonds are renowned as a material with superlative physical qualities, most of which originate from the strong fascia bonds between its atoms.

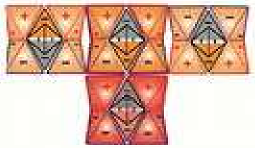

Diamonds

are a metastable allotrope of carbon,
where the carbon atoms are arranged in
cubic crystal structure called a diamond lattice.

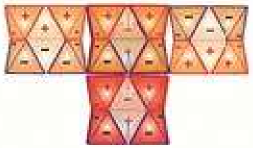

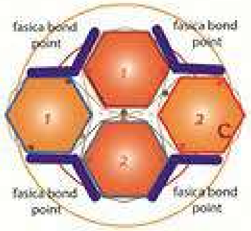

Ionised [C6] Carbon

Most natural diamonds are formed at high temperature and pressure at depths of 140 to 190 kilometers in the Earth's mantle [resulting in the loss of electrons form bound positions, making them poor conductors compared to grahphene atoms].

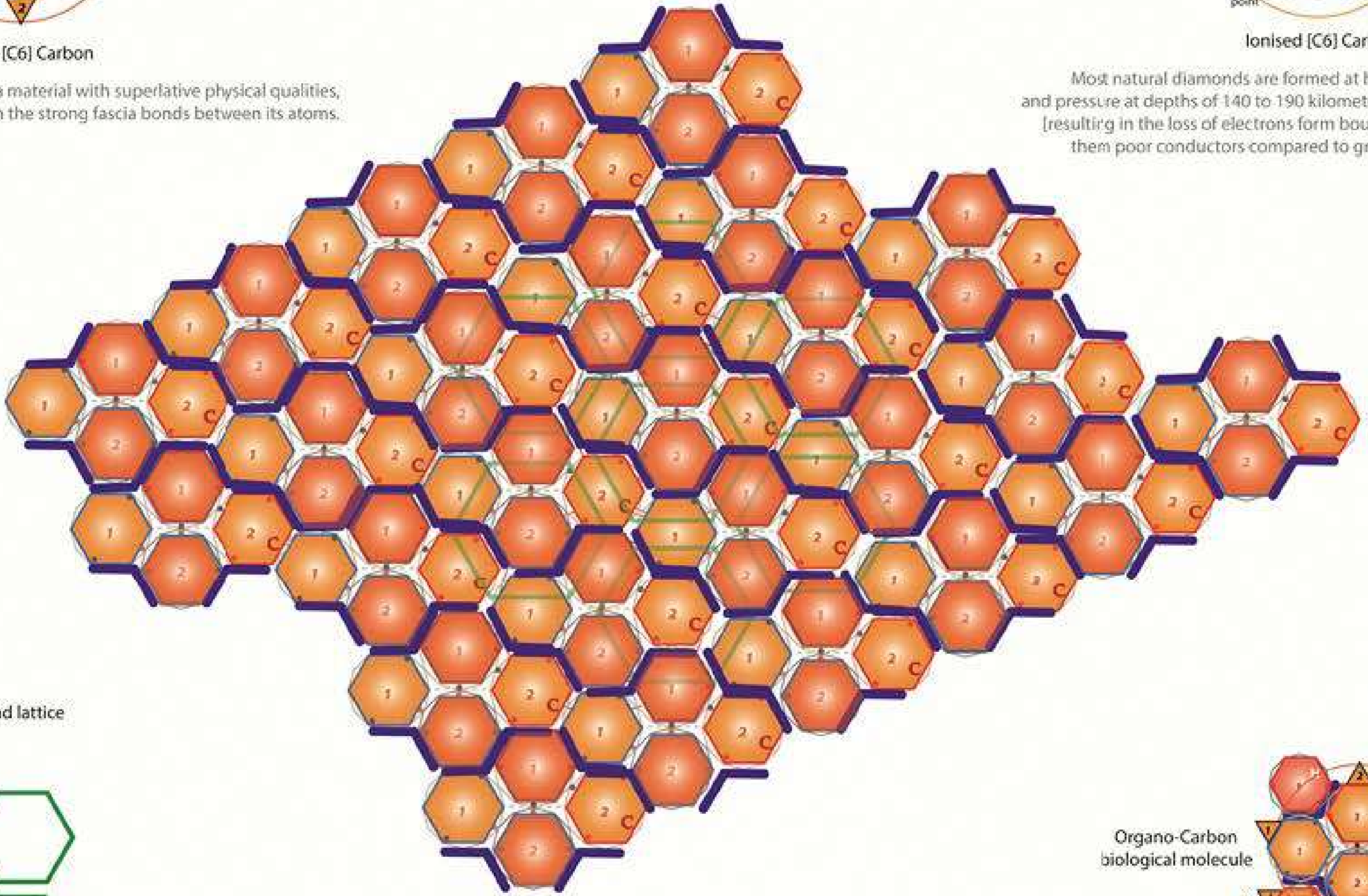

filled n[C7] Carbon diamond lattice

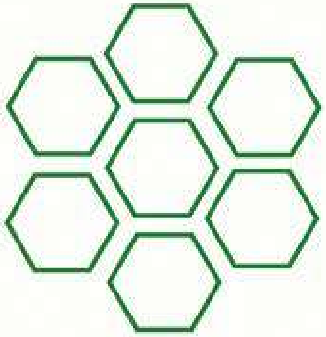

The chemical bonds that hold the carbon atoms in diamonds together
are stronger than those in graphite.

In diamonds, the bonds form an inflexible three-dimensional lattice, whereas in graphite,
the atoms are tightly bonded into sheets, which can slide easily over one another, making the overall structure weaker

Organo-Carbon
biological molecule

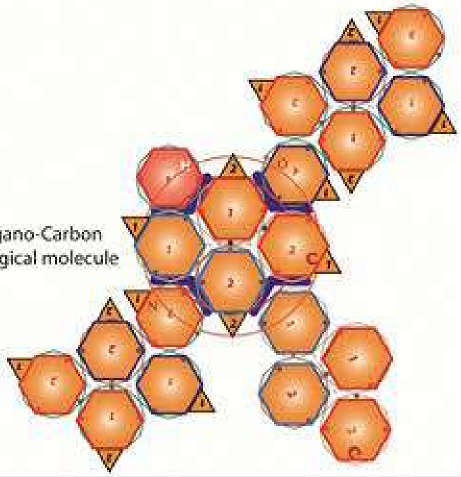

Copyright ABRAHAM [2008] - All rights reserved

Oxygen Allotropes

There are several known allotropes of oxygen.
The most familiar is molecular oxygen (O2), present at significant levels
in Earth's atmosphere and also known as dioxygen or tripet oxygen.
Another is the highly reactive ozone (O3).

Others include:
Atomic oxygen (O1, a free radical)
Singlet oxygen (O2), either of two metastable states of molecular oxygen
Tetraoxygen (O4), another metastable form
Solid oxygen, existing in six variously colored phases, of which
one is O8 and another one metallic

Atomic Oxygen also forms bonds easily with Hydrogen
to create Hydroxy compounds

OH
Hydroxide

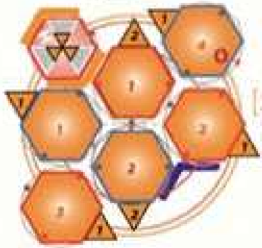

720
[360-360]

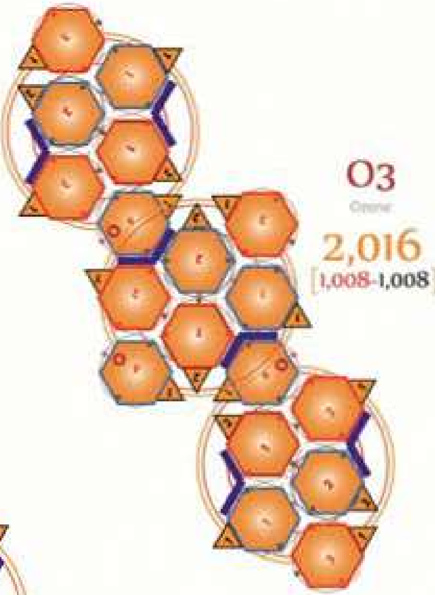

O3
Ozone
2,016
[1,008-1,008]

TetraOxygen
O4

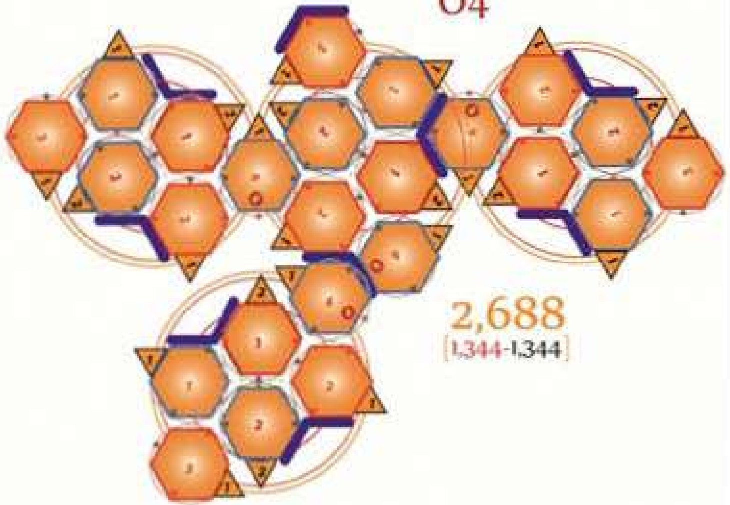

2,688
[1,344-1,344]

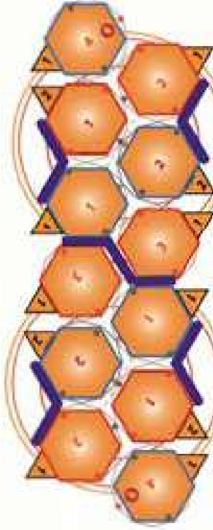

O2
Oxygen
1,344
[672-672]

All charge fascia interact
and bind where possible to form
neutral chemical bonds

TetraOxygen
O4
2,688
[1,344-1,344]

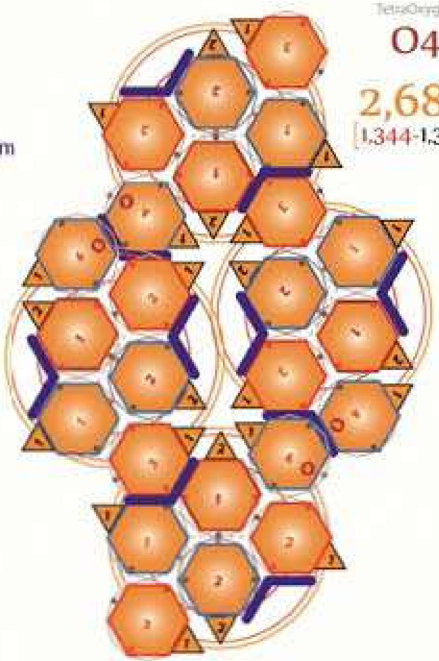

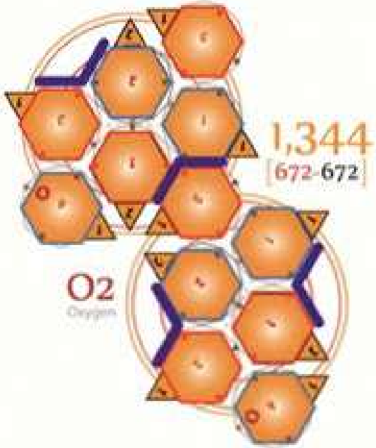

1,344
[672-672]

O2
Oxygen

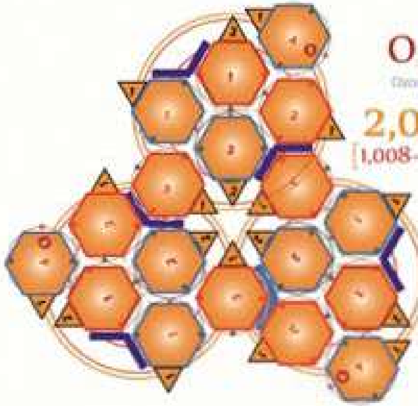

O3
Ozone
2,016
[1,008-1,008]

Tetryonics 54.08 - Oxygen allotropes

Copyright ABRAHAM [2008] - All rights reserved

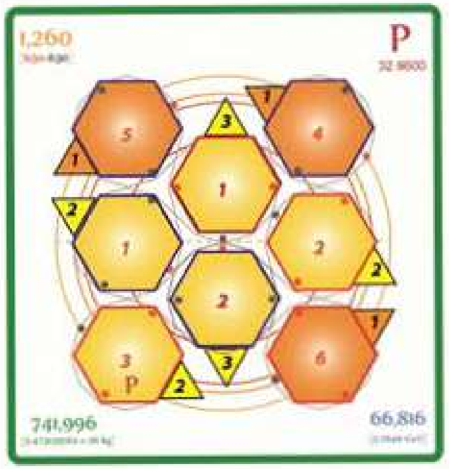
1,260
P
741,996
66,816

Phosphor Allotropes

Elemental phosphorus can exist in several allotropes; the most common of which are white and red solids.

Solid violet and black allotropes are also known.

Gaseous phosphorus exists as diphosphorus and atomic phosphorus.

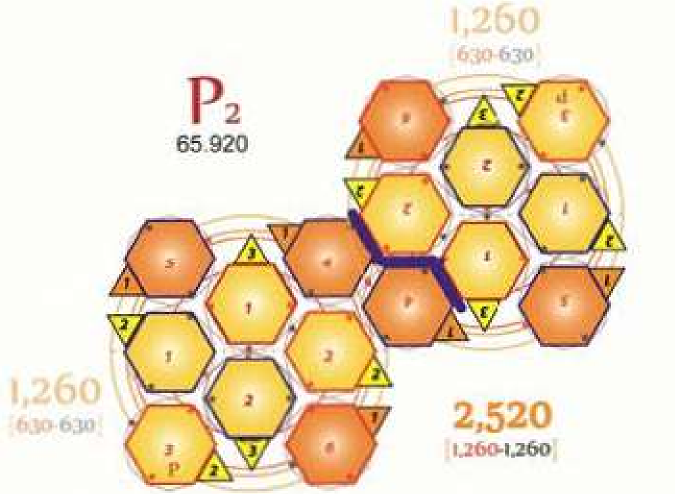
P_2
65.920
1,260
[630-630]
1,260
[630-630]
2,520
[1,260-1,260]

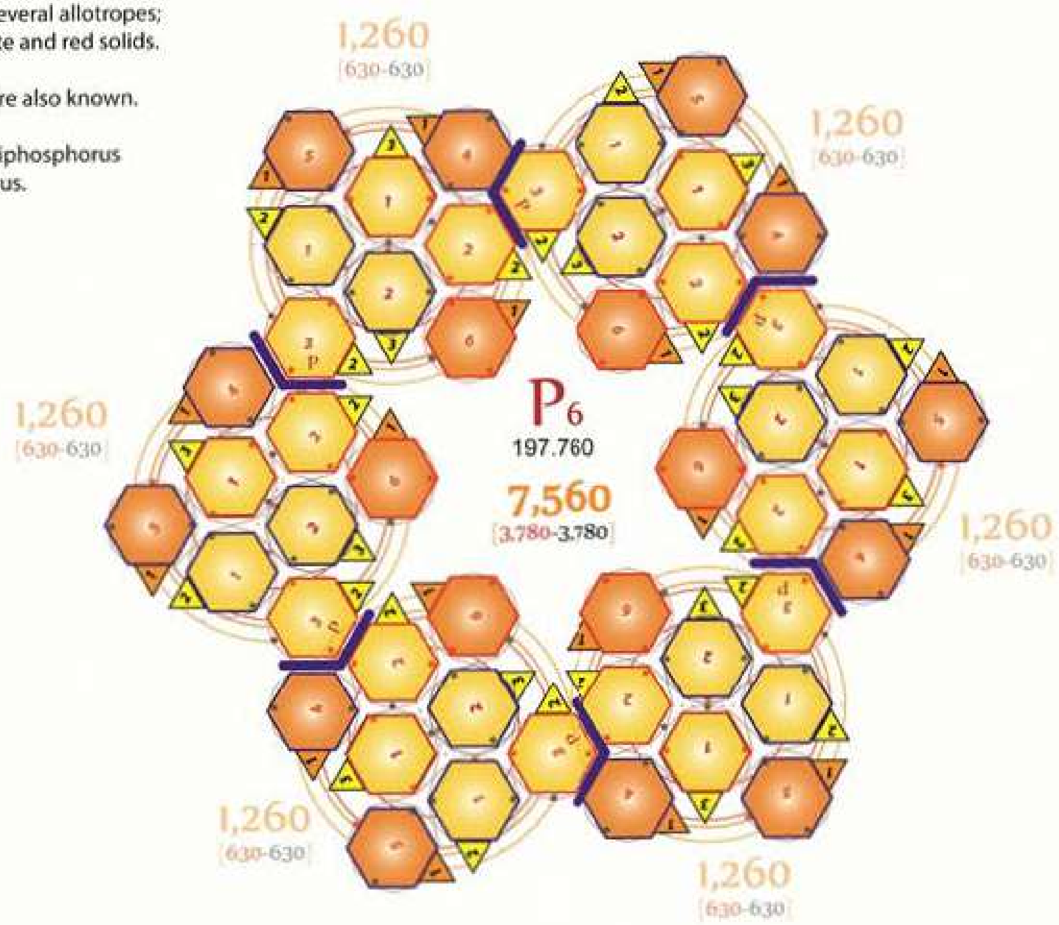
1,260
[630-630]
1,260
[630-630]
1,260
[630-630]
P_6
197.760
7,560
[3,780-3,780]
1,260
[630-630]
1,260
[630-630]
1,260
[630-630]

The diphosphorus allotrope (P2) can be obtained normally only under extreme conditions (for example, from P4 at 1100 kelvin).

Nevertheless, some advancements were obtained in generating the diatomic molecule in homogenous solution, under normal conditions with the use by some transition metal complexes (based on for example tungsten and niobium).

Black phosphorus is the thermodynamically stable form of phosphorus at room temperature and pressure. It is obtained by heating white phosphorus under high pressures (12,000 atmospheres).

In appearance, properties and structure it is very like graphite, being black and flaky, a conductor of electricity, and having puckered sheets of linked atoms.

Copyright ABRAHAM [2008] - All rights reserved

Sulfur Allotropes

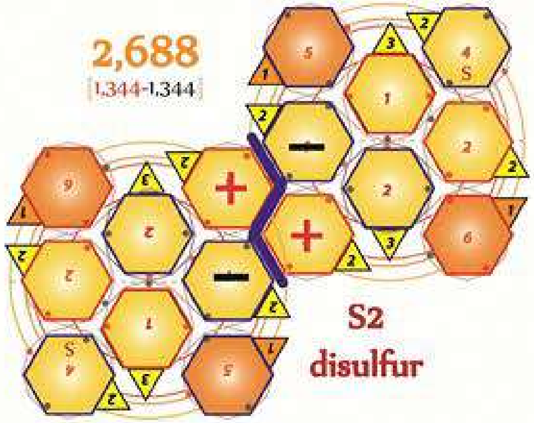

2,688
[1,344-1,344]

S2
disulfur

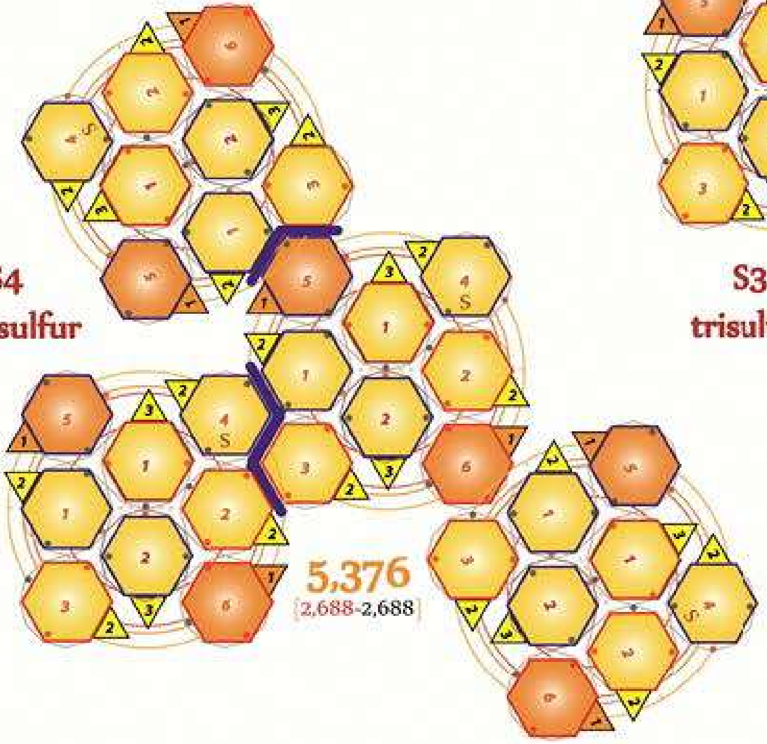

S4
tetrasulfur

5,376
[2,688-2,688]

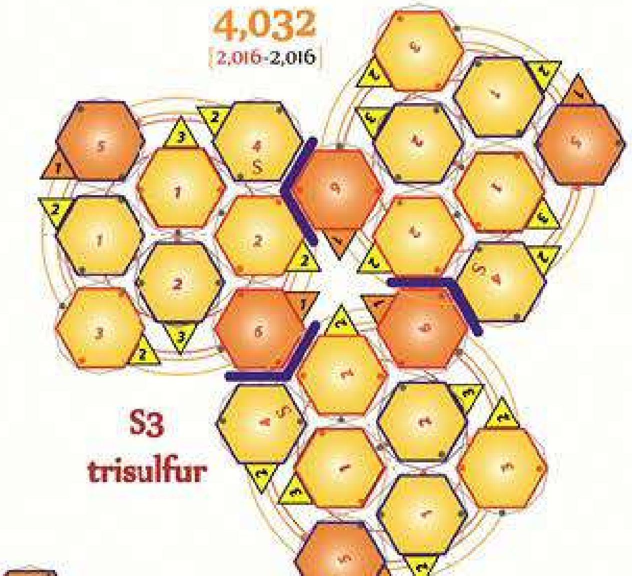

4,032
[2,016-2,016]

S3
trisulfur

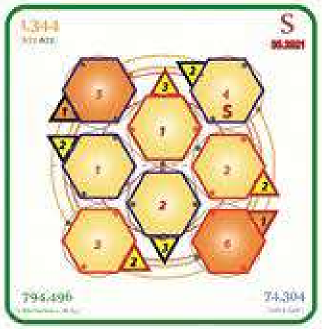

Charged fascia interactions
between differing elements
result in the familiar
chemical bonds

Sulfur [like Carbon]
can form Chains
or Ring compounds

There are a large number of allotropes of sulfur.
In this respect, sulfur is second only to carbon.

cyclo-hexasulfur

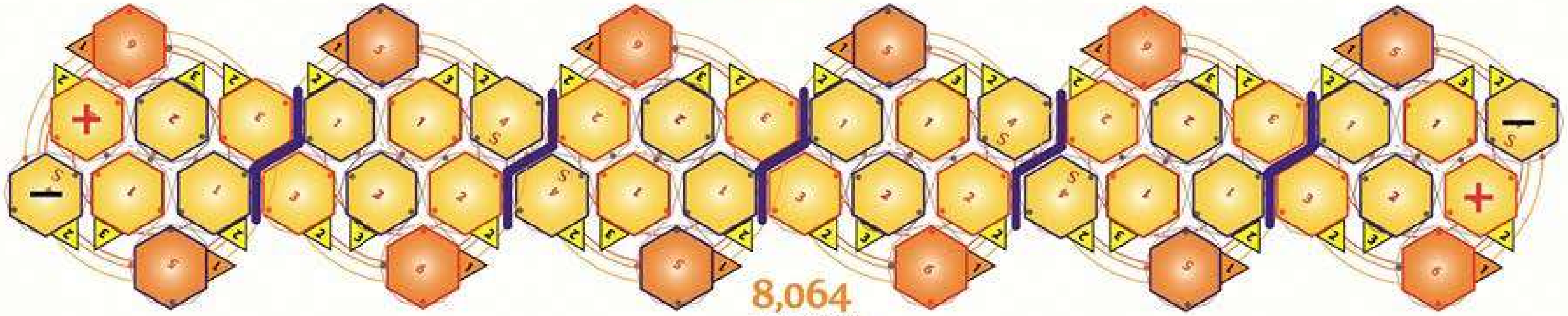

8,064
[4,032-4,032]

Isotopes

In addition to forming varying allotropic elements

Atoms can also absorb energy directly and create numerous elemental isotopes as a result of their differing nuclear energy levels

Deuterium nuclei with bound photo-electrons form quantum-scale synchronous converters

n
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
Shell
Level

0	**Hydrogen**	22,512	24,384	26,352	28,416	30,576	32,832	35,184	37,632
Element		1	2	3	4	5	6	7	8
1	**Deuterium**	45,012	48,720	52,596	56,640	60,852	65,232	69,780	74,496
2	**Helium**	90,024	97,440	105,192	113,280	121,704	130,464	139,560	148,992
3	**Lithium**	135,036	146,160	157,788	169,920	182,556	195,696	209,340	223,488
4	**Berylium**	180,048	194,880	210,384	226,560	243,408	260,928	279,120	297,984
5	**Boron**	225,060	243,600	262,980	283,200	304,260	326,160	348,900	372,480
6	**Carbon**	270,072	292,320	315,576	339,840	365,112	391,392	418,680	446,976
7	**Nitrogen**	315,084	341,040	368,172	396,480	425,964	456,624	488,460	521,472
8	**Oxygen**	360,096	389,760	420,768	453,120	486,816	521,856	558,240	595,968
9	**Fluorine**	405,108	438,480	473,364	509,760	547,668	587,088	628,020	670,464
10	**Neon**	450,120	487,200	525,960	566,400	608,520	652,320	697,800	744,960
11	**Sodium**	495,132	535,920	578,556	623,040	669,372	717,552	767,580	819,456
12	**Magnesium**	540,144	584,640	631,152	679,680	730,224	782,784	837,360	893,952
13	**Aluminium**	585,156	633,360	683,748	736,320	791,076	848,016	907,140	968,448
14	**Silicon**	630,168	682,080	736,344	792,960	851,928	913,248	976,920	1,042,944
15	**Phosphorus**	675,180	730,800	788,940	849,600	912,780	978,480	1,046,700	1,117,440
16	**Sulfur**	720,192	779,520	841,536	906,240	973,632	1,043,712	1,116,480	1,191,936
17	**Chlorine**	765,204	828,240	894,132	962,880	1,034,484	1,108,944	1,186,260	1,266,432
18	**Argon**	810,216	876,960	946,728	1,019,520	1,095,336	1,174,176	1,256,040	1,340,928
19	**Potassium**	855,228	925,680	999,324	1,076,160	1,156,188	1,239,408	1,325,820	1,415,424
20	**Calcium**	900,240	974,400	1,051,920	1,132,800	1,217,040	1,304,640	1,395,600	1,489,920

Level
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
Shell

It is the increased nucleonic energy levels that creates Isotopes (not extra Neutrons within an atomic nucleus)

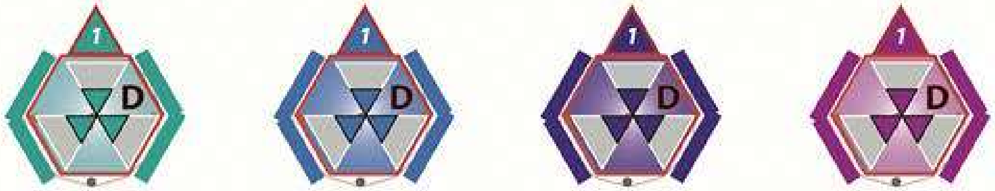

The energy level of Baryons determines the KEM energies of photo-electrons bound to them

Most isotopes are considered to be radioactive as a result of the nuclei seeking to release excess energy in the form of photons of energy or particles with high kinetic energies

Carbon Isotopes

It is widely held in the scientific community tha:
Carbon-14, 14C, or radiocarbon, is a radioactive isotope of carbon with a nucleus containing 6 protons and 8 neutrons.

It is in fact a nucleus comprised of 6 deuterium nuclei [with 6 Protons, 6 Neutrons & 6 electrons]

The mistaken belief in 'extra' neutrons being present in the nucleus stems from the fact that electrons and protons combine in equal numbers in the atomic nucleus and historically attributing the mass in excess of this as being the result of the mass contribution of 'extra' neutrons

Tetryonics finally corrects this erroneous assumption

Carbon 12-14

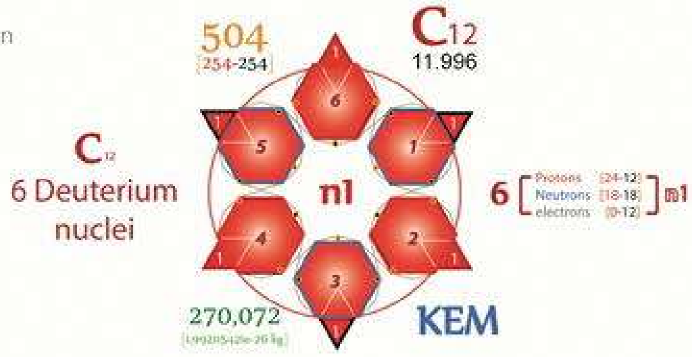
504
[254-254]
C12
11.996
C12
6 Deuterium nuclei
6
5
1
4
2
3
n1
6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1
270,072
KEM

There are NO extra neutrons (in excess of the element's Z#) in the nuclei of atomic isotopes.

The measured 'excess mass' is the direct result of the raised quantum levels of the Deuterium nuclei that comprise each atomic element

And is compreletely accounted for in Tetryonic theory by calculating for the total rest mass-energies in each elementary Matter topology.

The 'extra' mass historically attrtibuted to neutron numbers above that of the elemental number are now reflected as stored kinetic 'chemical' energies as they always were.

ALL elements & isotopes have equal numbers of Protons, electrons & Neutrons

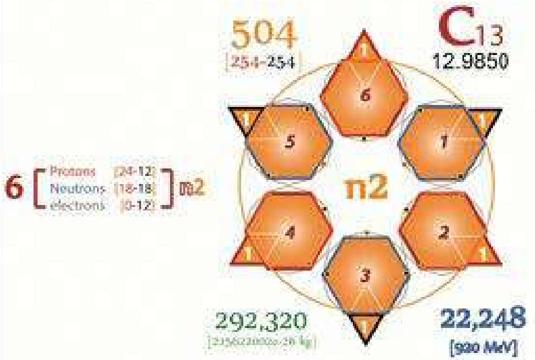
504
[254-254]
C13
12.9850
6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n2
6
5
1
n2
4
2
3
292,320
22,248
[920 MeV]

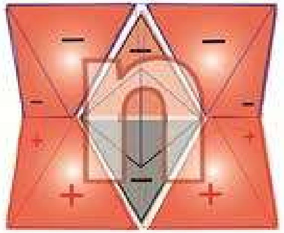

Quantum levels of atomic nuclei contribute to the molar mass [Isotopes are higher energy nuclei]

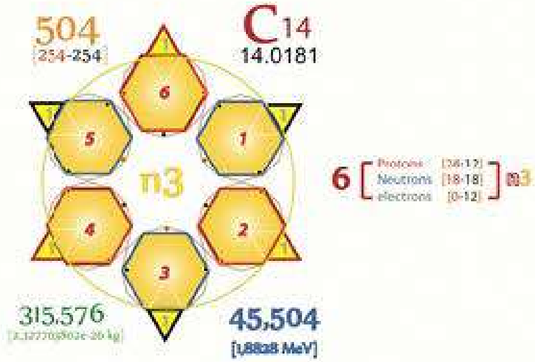
504
[254-254]
C14
14.0181
6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n3
6
5
1
n3
4
2
3
315,576
45,504
[1,8828 MeV]

	n1	n2	n3	n4	n5	n6	n7	n8
C12	270,072	292,320	315,576	339,840	365,112	391,392	418,680	446,976
KEM diff	0	22,248	45,504	69,768	95,040	121,320	148,608	176,904
neutron #		.98	2.0	3.1	4.2	5.4	6.6	7.8

This applies equally to all atomic nuclei

Copyright ABRAHAM [2008] - All rights reserved

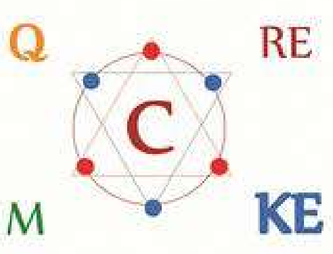

Electron Configuration
[n] p1

Atoms store mass-energies
in their standing wave Matter topologies
and release it via radioactive decay paths

It is increased kinetic energy level geometries that create nuclear isotopes (not extra Neutrons within a nucleus)

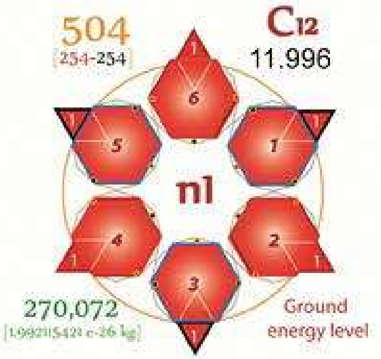
504
[254-254]
C12
11.996
n1
270,072
[1.992115421 e-26 kg]
Ground
energy level

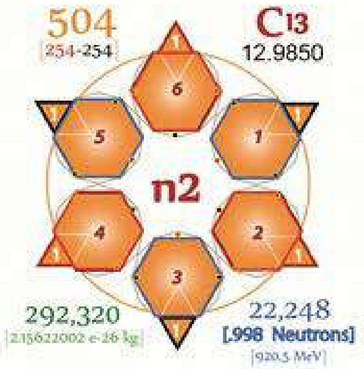
504
[254-254]
C13
12.9850
n2
292,320
[2.15622002 e-26 kg]
22,248
[.998 Neutrons]
[920.5 MeV]

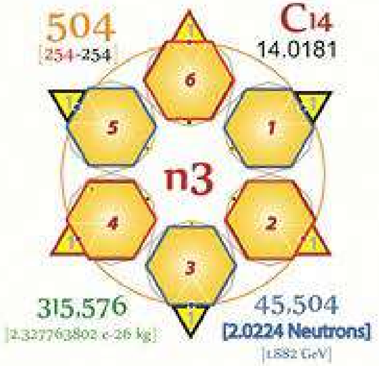
504
[254-254]
C14
14.0181
n3
315.576
[2.327763802 e-26 kg]
45.504
[2.0224 Neutrons]
[1.882 GeV]

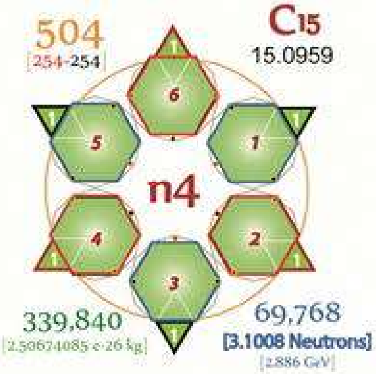
504
[254-254]
C15
15.0959
n4
339,840
[2.50674085 e-26 kg]
69,768
[3.1008 Neutrons]
[2.886 GeV]

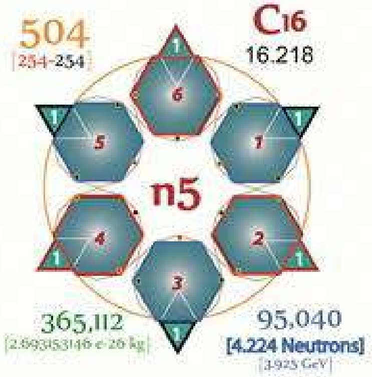
504
[254-254]
C16
16.218
n5
365,112
[2.693153146 e-26 kg]
95,040
[4.224 Neutrons]
[3.925 GeV]

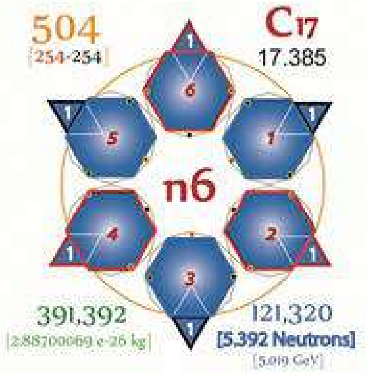
504
[254-254]
C17
17.385
n6
391,392
[2.88700069 e-26 kg]
121,320
[5.392 Neutrons]
[5.019 GeV]

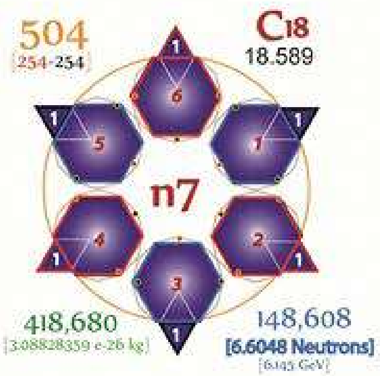
504
[254-254]
C18
18.589
n7
418,680
[3.08828359 e-26 kg]
148,608
[6.6048 Neutrons]
[6.145 GeV]

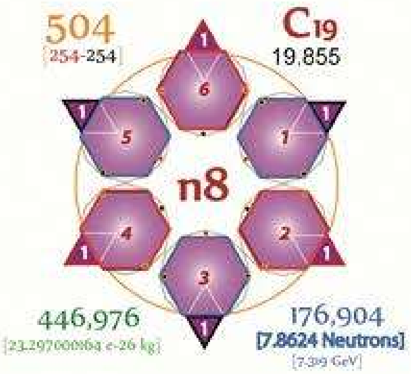
504
[254-254]
C19
19.855
n8
446,976
[23.29700064 e-26 kg]
176,904
[7.8624 Neutrons]
[7.319 GeV]

Shell			Molar meM	Nuclei per shell
R	8	446,976	19.855	6
Q	7	418,680	18.598	6
P	6	391,392	17.385	6
O	5	365,112	16.218	6
N	4	339,840	15.095	6
M	3	315,576	14.018	6
L	2	292,320	12.985	6
K	1	270,072	11.996	6

energy
level

22,500
[1 Neutron]

n

[930.97 MeV]

Kinetic mass-energy geometries
stored in 3D Matter topologies
is released as Chemical energies

Deuterium nuclei with bound photo-electrons form quantum-scale synchronous converters

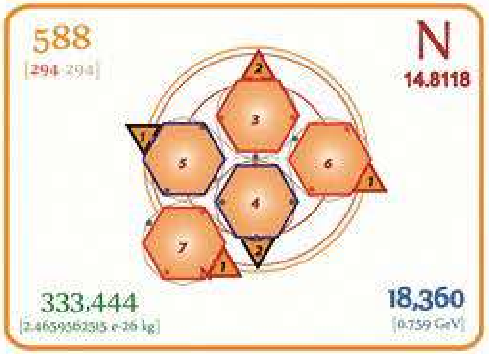
588
[294-294]
N
14.8118
333,444
[2.4659562315 e-26 kg]
18,360
[0.759 GeV]

It is Increased Nucleonic energy levels that creates Isotopes
(not extra Neutrons within any elementary Nucleus)

Nitrogen gas

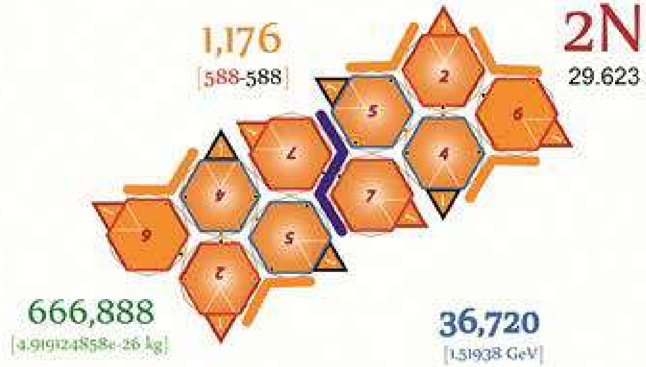
1,176
[588-588]
2N
29.623
666,888
[4.919124858e-26 kg]
36,720
[1.51938 GeV]

Nitrogen Allotropes & Isotopes

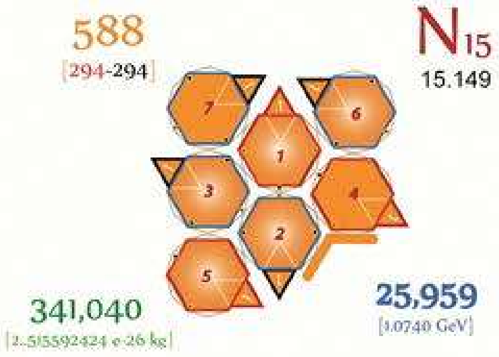
588
[294-294]
N15
15.149
341,040
[2.515592424 e-26 kg]
25,959
[1.0740 GeV]

Isotopes are created by increasing the number of Planck mass-energies stored in the standing-wave geometries of chemical elements

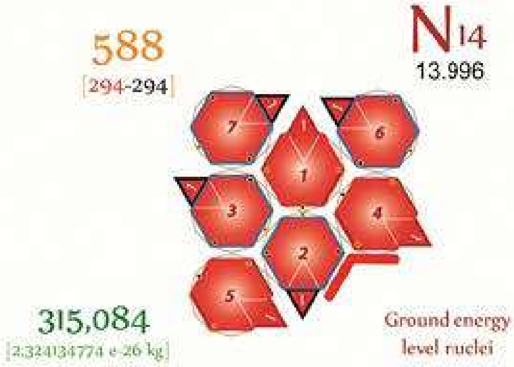
588
[294-294]
N14
13.996
315,084
[2.324134774 e-26 kg]

Ground energy level nuclei

The charged mass-energy geometries of differing allotropic topologies create various chemical bond points

Shell			Molar meM	Nuclei per shell
R	8	521,472	23.164	7
Q	7	488,460	21.697	7
P	6	456,624	20.283	7
O	5	425,964	18.921	7
N	4	396,480	17.612	7
M	3	368,172	16.354	7
L	2	341,040	15.149	7
K	1	315,084	13.996	7
				7

energy level

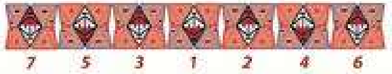

Copyright ABRAHAM [2008] - All rights reserved

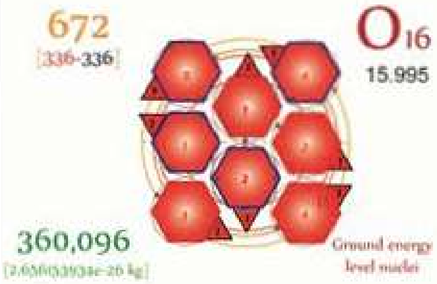
672
[336-336]
O16
15.995
360,096
[2.6560539334e-26 kg]
Ground energy level nuclei

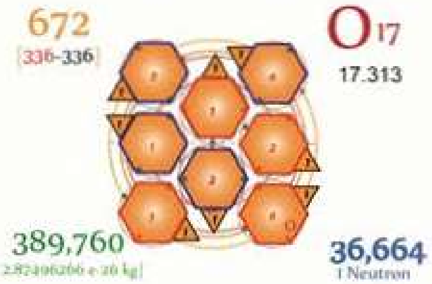
672
[336-336]
O17
17.313
389,760
[2.87496206 e-26 kg]
36,664
1 Neutron

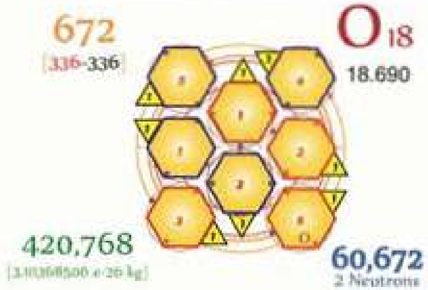
672
[336-336]
O18
18.690
420,768
[3.10368500 e-26 kg]
60,672
2 Neutrons

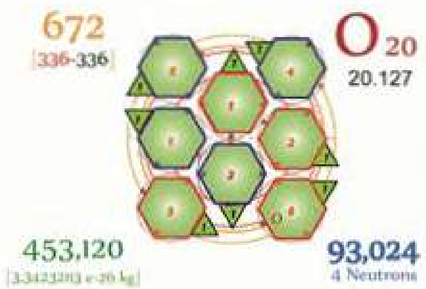
672
[336-336]
O20
20.127
453,120
[3.3423203 e-26 kg]
93,024
4 Neutrons

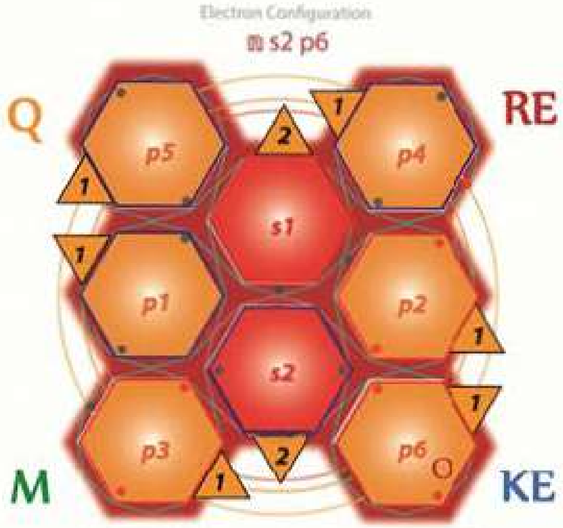
Electron Configuration
m s2 p6
Q
RE
M
KE
p5
p4
s1
p1
p2
s2
p3
p6
1
2
1
1
1
1
2
1

This oxygen allotrope's nuclei arrangement is very stable and inert like Neon [due to its filled sp electron orbitals]

Oxygen Allotropes & Isotopes

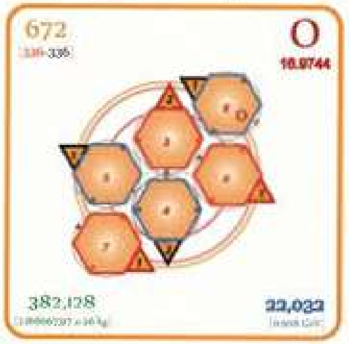
672
[336-336]
O
16.9744
382,128
22,032

There exists an allotrope of Oxygen formed when 8 nuclei combine in a single level SP orbital arrangement

Shell			Molar meM	Nuclei per shell
R	8	595,968	26.473	8
Q	7	558,240	24.797	8
P	6	521,856	23.181	8
O	5	486,816	20.825	8
N	4	453,120	20.127	8
M	3	420,768	18.690	8
L	2	389,760	17.313	8
K	1	360,096	15.995	8

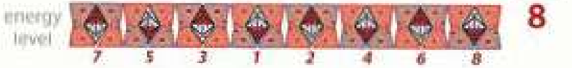
energy level
7
5
3
1
2
4
6
8
8

It is Increased Nucleonic energy levels that creates Isotopes (not extra Neutrons within a Nucleus)

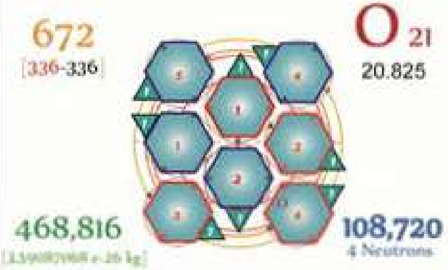
672
[336-336]
O21
20.825
468,816
[3.45817968 e-26 kg]
108,720
4 Neutrons

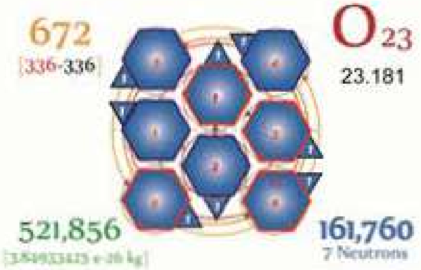
672
[336-336]
O23
23.181
521,856
[3.84933425 e-26 kg]
161,760
7 Neutrons

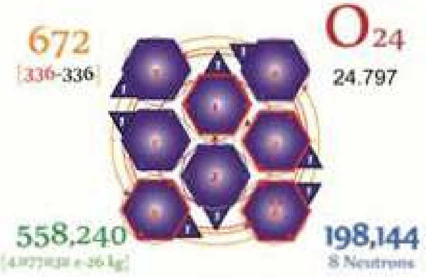
672
[336-336]
O24
24.797
558,240
[4.1177038 e-26 kg]
198,144
8 Neutrons

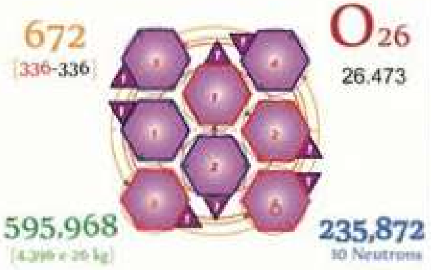
672
[336-336]
O26
26.473
595,968
[4.396 e-26 kg]
235,872
10 Neutrons

Tetryonics 54.15 - Oxygen isotopes

Bose–Einstein condensates exist at the low energy extremes

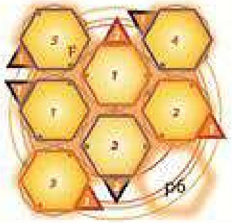

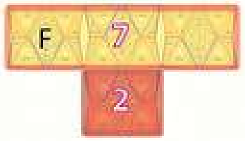

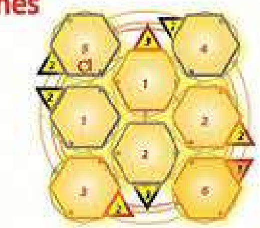

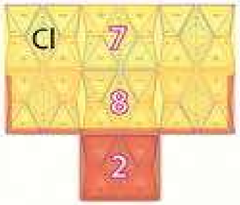

Gaseous

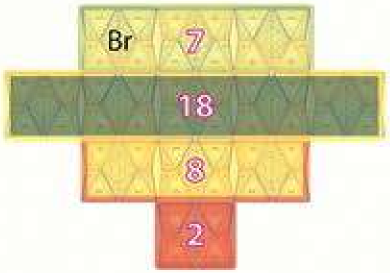

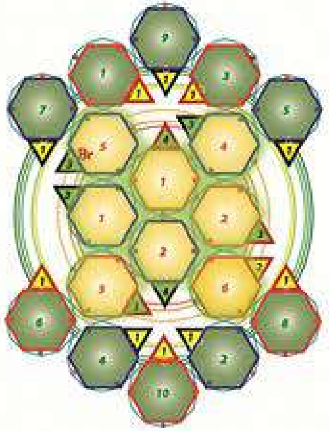

Liquid

The Natural states of Matter

(room temperature)

The halogen element family provides a great example of all 4 states of Matter at room temperature

Elements appear in their natural phase at room temperature as a gas, liquid, solid or synthetic

As energy is added or removed from any element's charged Matter topology or their baryonic KEM field geometry their phase states & Brownian motion changes

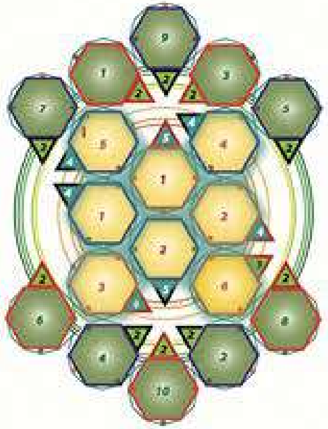

Solid

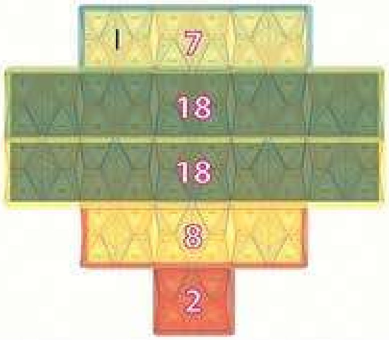

Radioactive

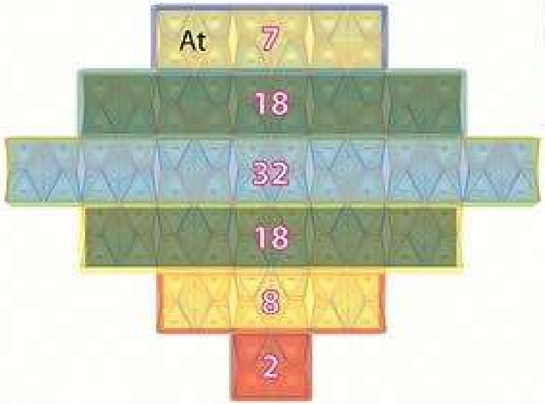

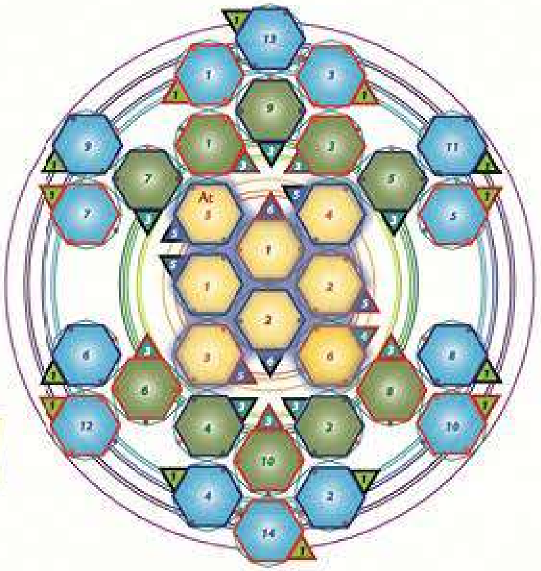

At the high energy extremes exist the Plasma states of Matter

Chemical bonds

A chemical bond is an attraction between atoms that allows the formation of chemical substances that contain two or more atoms.

Particle charge & Inter-fascia electric interaction provides the foundation of chemical bonds

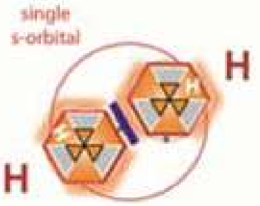

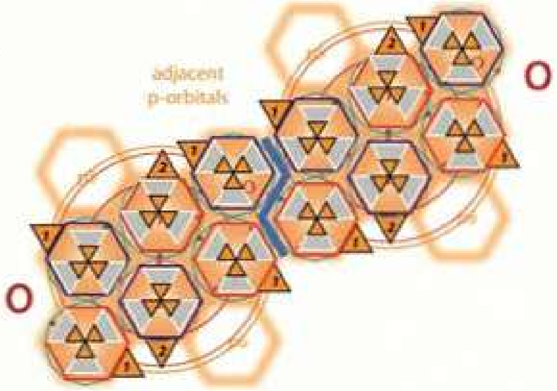

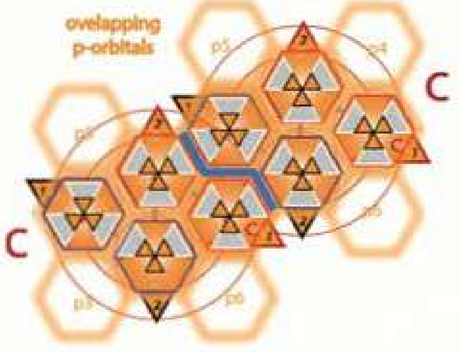

It is usually taught that based on electron sharing principles a single bond contains 2 paired electrons, a double bond has 4 electrons and a triple bond has 6 shared electrons

filled electron shells
[weak KEM'ical bonds]

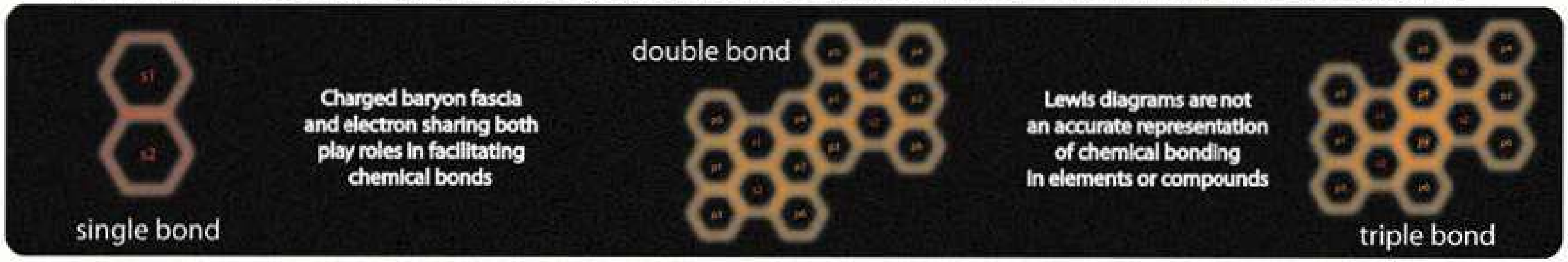

Charged baryon fascia and electron sharing both play roles in facilitating chemical bonds

Lewis diagrams are not an accurate representation of chemical bonding in elements or compounds

unfilled electron shells
[stronger KEM'ical bonds]

Tetryonic geometry provides an advanced quantum topology for all atoms and compounds that is superior to the older Lewis diagrams and even molecular orbital theory, facilitating a clear understanding of the quantum geometry of each element, its nuclear bonds and electron sharing in all compound chemical structures; along with the core & valence electron configurations & interactive electric charge fascia bonding positions for each compound created.

ethane
[C_2H_6]

6 valence electrons

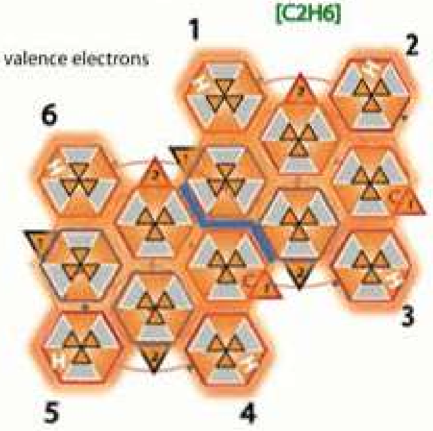

ethene
[C_2H_4]

4 valence electrons

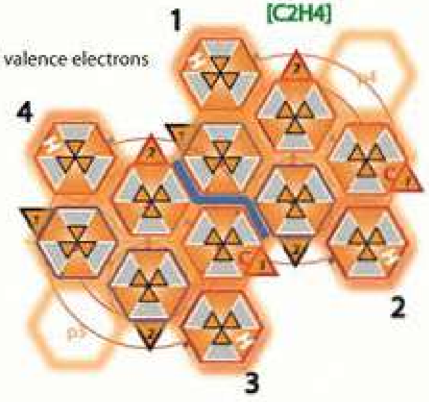

ethyne
[C_2H_2]

2 valence electrons

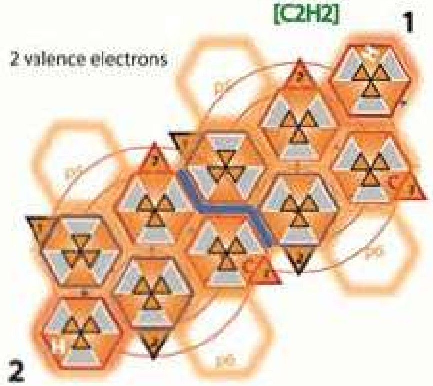

Valance electrons also play an important role in determining the strength of chemical bonds

Van der Waals Forces

Van der Waals forces include attractions between atoms, molecules, and surfaces. They differ from covalent and ionic bonding in that they are caused by correlations in the fluctuating polarizations of nearby particles

All charges seek equilibrium

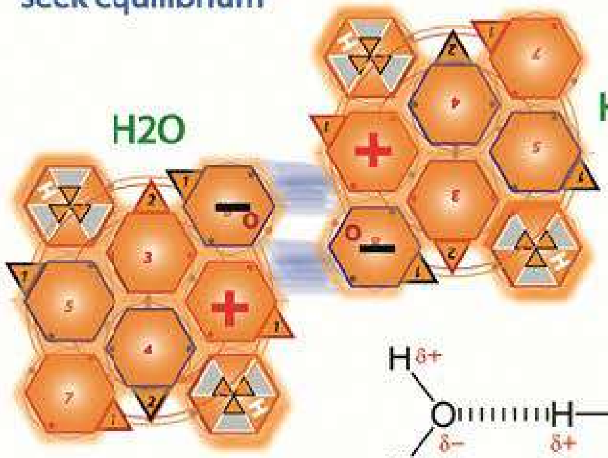

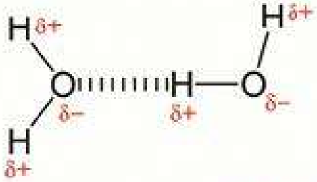

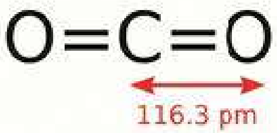

Van der Walls and London forces are the geometric fields of interaction between charged Matter topologies

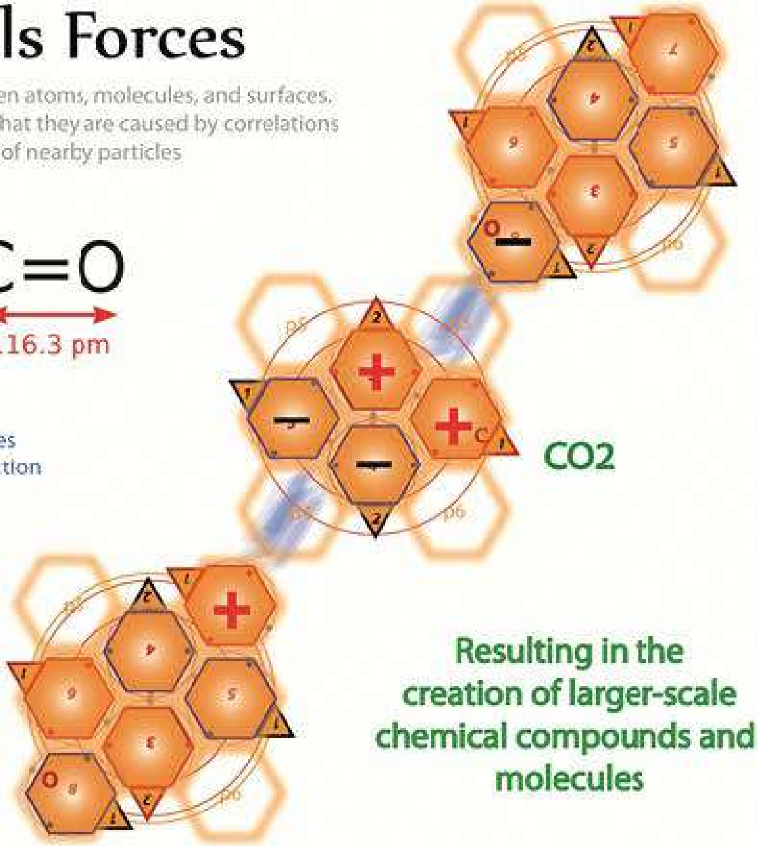

The Forces of interaction can be attractive or repulsive

Residual interactive bonding force between the charged fascia of atoms

Resulting in the creation of larger-scale chemical compounds and molecules

This is considered to be the only attractive intermolecular force present between neutral atoms (e.g., a noble gas configurations and charge neutral atomic topologies).

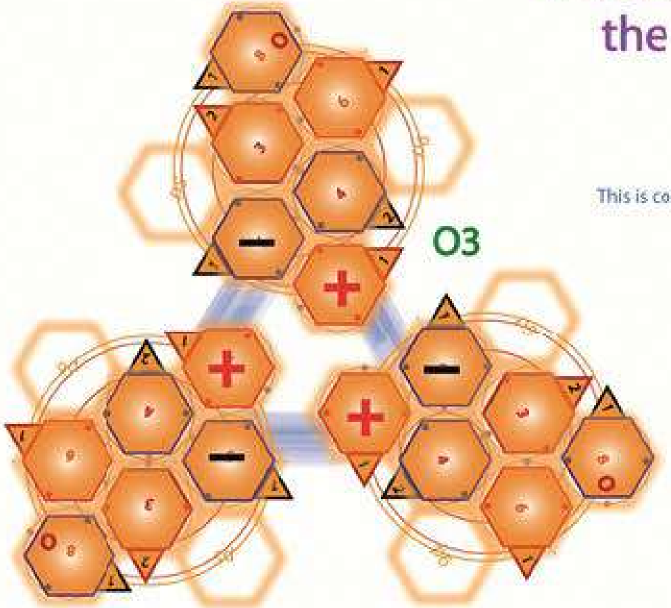

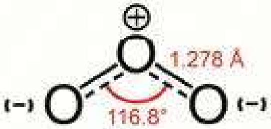

Without London forces, there would be no attractive force between noble gas atoms, and they wouldn't exist in liquid form.

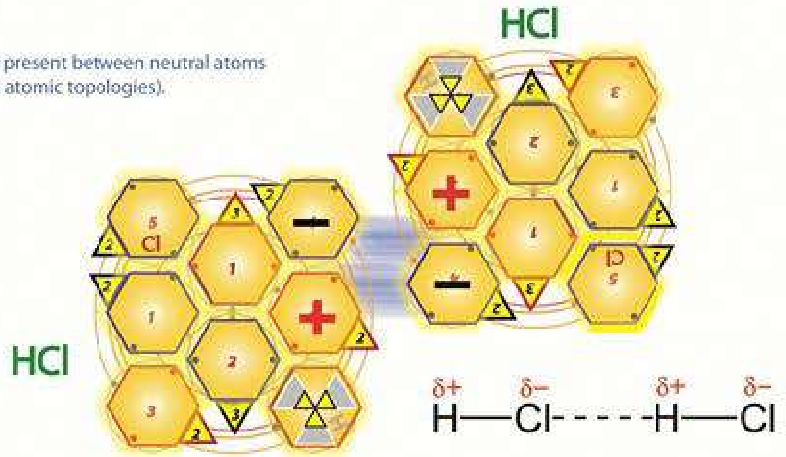

$\overset{\delta+}{H}—\overset{\delta-}{Cl}\text{- - - -}\overset{\delta+}{H}—\overset{\delta-}{Cl}$

Lewis Structures

The Lewis structure was named after Gilbert N. Lewis,
who introduced it in his 1916 article
The Atom and the Molecule.

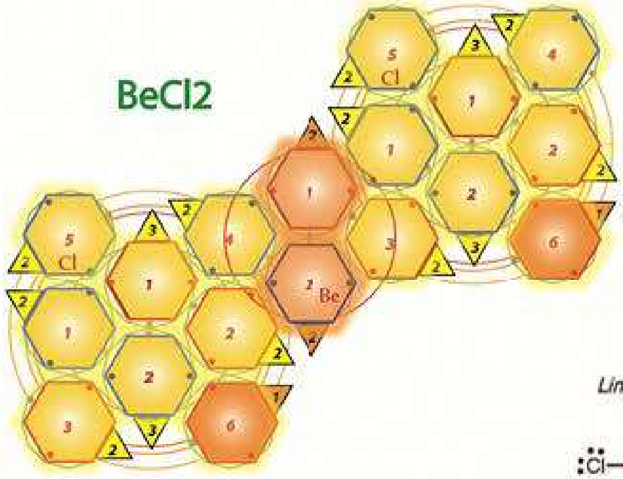
BeCl2

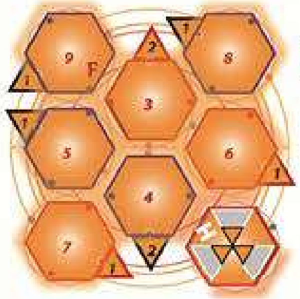
HF

Lewis structures, also called Lewis-dot diagrams,
are diagrams that show the bonding between the atoms of any molecule,
and the lone pairs of electrons that may exist in the molecule.

Line drawings can be used to depict molecular geometry:

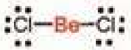
Beryllium dichloride
$BeCl_2$
linear

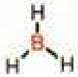
Borane
BH_3
trigonal planar

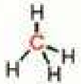
Methane
CH_4
tetrahedral

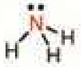
Ammonia
NH_3
trigonal pyrimidal

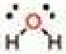
Water
H_2O
bent

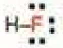
Hydrofluoric acid
HF
linear

Hydrogen is a
'free radical' atom
whose energy can be
changed to facilitate
chemical bonding
between elements

Tetryonic geometry uses the
charged geometry of the
element fascia themselves

to show the bonds and
charge interactions between
differing elements as they
form larger compounds

as well as the actual final
quantum topology of
compound elements
and molecules

Also applies to molecules with multiple bonding:

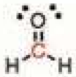
methanal
(formaldehyde)
trigonal planar

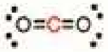
Carbon dioxide
linear

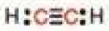
ethyne
(acetylene)
linear

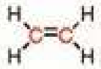
ethane
(ethylene)
both carbons are trigonal planar

They are similar to electron dot diagrams in that the valence electrons in lone pairs are represented as dots,
but they also contain lines to represent shared pairs in a chemical bond (single, double, triple, etc.).

single
fascia bond

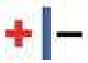

double
fascia bonds

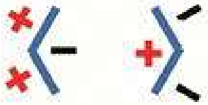

triple
fascia bond

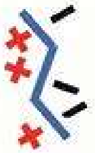

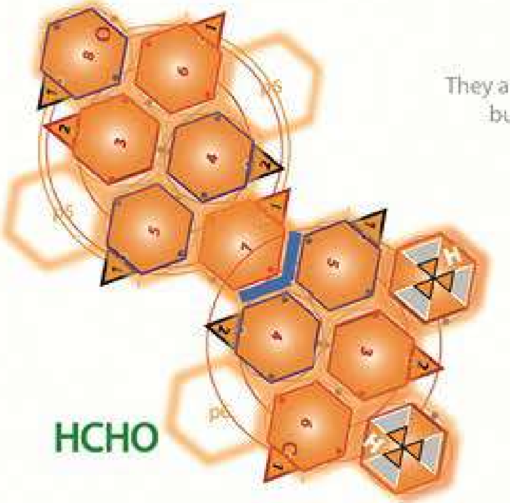
HCHO

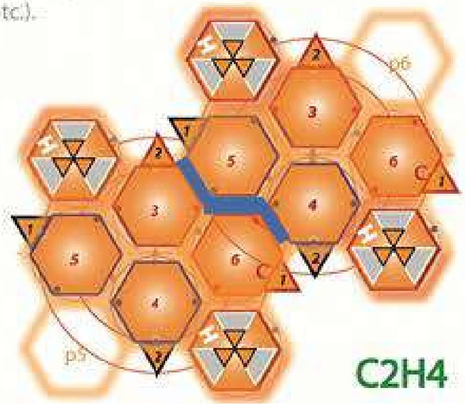
C2H4

It is the electric field fascia of baryons that facilitates chemical bonds

Atomic bonds

All atoms, elements and compounds seek stable core energy & configurations where their electron orbitals are filled

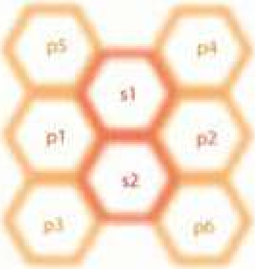

Hydrogen σ bonds

Outward presenting electic fascia bonds
Facilitate bonding between molecules

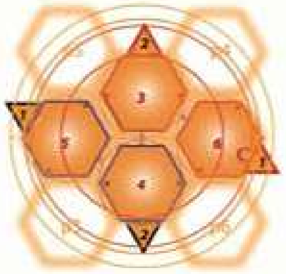

Molecular π bonds

Inward recieving bonds
capable of accepting Hydrogen
or extra-orbital elctric fascia bonds

Bonds fill in order of orbital filling
ie p1-6, d1-10

OH

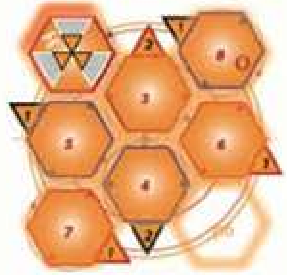

Hydroxide bonds

Oxygen-Hydrogen compound
creates a halogen-like topology
which seeks to fill its p6 orbital
in order to reach a stable electronic
configuration

Extremely reactive

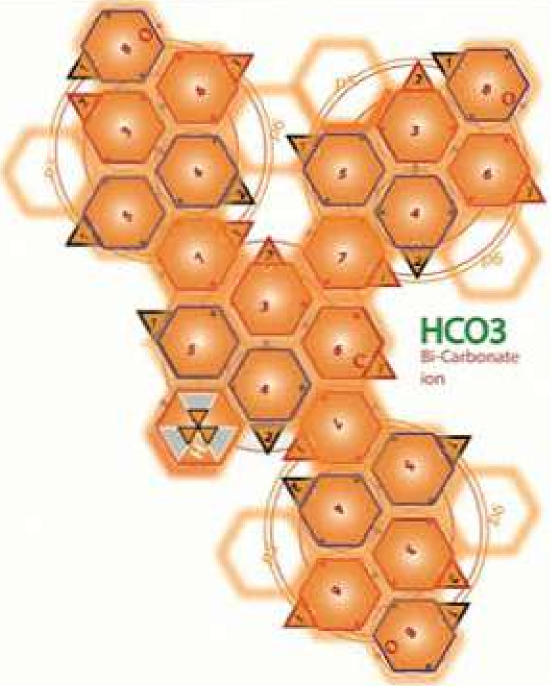

HCO3
Bi-Carbonate
ion

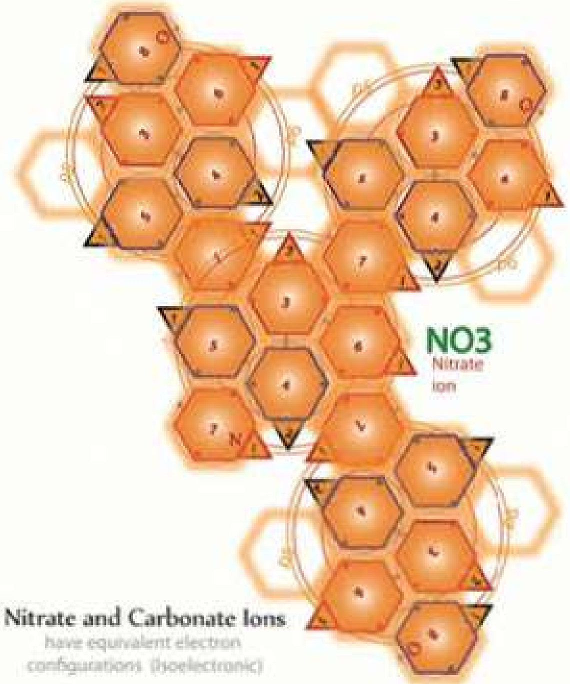

NO3
Nitrate
ion

Nitrate and Carbonate Ions
have equivalent electron
configurations (Isoelectronic)

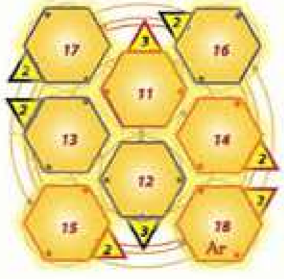

Core electron configuration

Unreactive (non-valence)
electron configuration

CH4

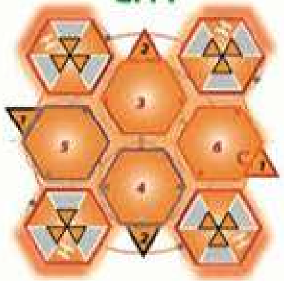

Covalent bonds

Intra-orbital bonding between
elements and compounds where
electrons exchange is the main
mechanism resulting in
stable electronic configurations

Ionic bonds

Extra-orbital bonding between
elements and compounds where
element charge attraction is the
predominent mechanism
with electrons sharing resulting in
stable electronic configurations

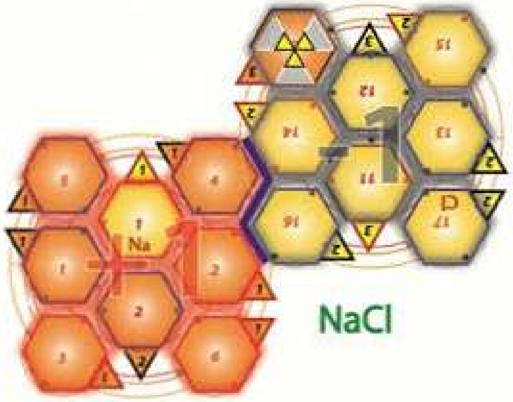

NaCl

Tetryonics 55.05 - Atomic facsia Bonds

Copyright ABRAHAM [2008] - All rights reserved

Atomic bonds

All atoms, elements and compounds seek stable core electron configurations where their electron orbitals are filled

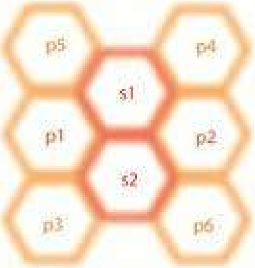

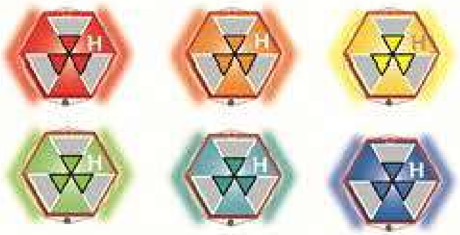

Hydrogen Bonds

Outward presenting electic fascia bonds
Facilitate bonding between molecules

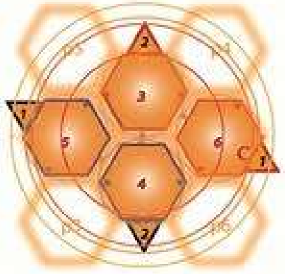

Molecular bonds

Inward recieving bonds
capable of accepting Hydrogen
or extra-orbital elctric fascia bonds

Bonds fill in order of orbital filling
ie p1-6, d1-10

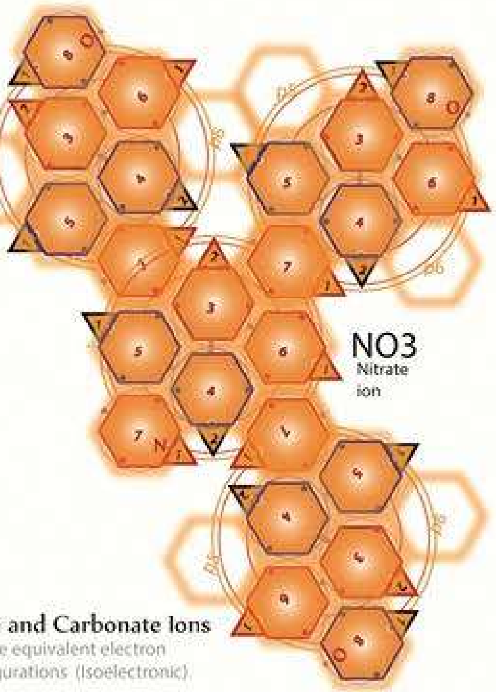

NO3
Nitrate
ion

Nitrate and Carbonate Ions

have equivalent electron
configurations (Isoelectronic)

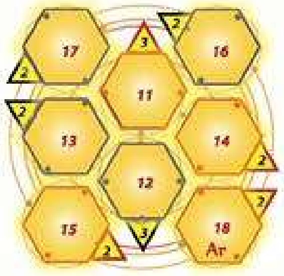

Core electron configuration

Unreactive [non-valence]
electron configuration

OH

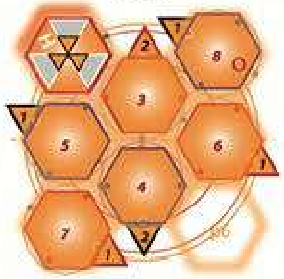

Hydroxide bonds

Oxygen-Hydrogen compound
creates a halogen-like geometry
which seeks to fill its p6 orbital
in order to reach a stable electronic
configuration

Extremely reactive

CH4

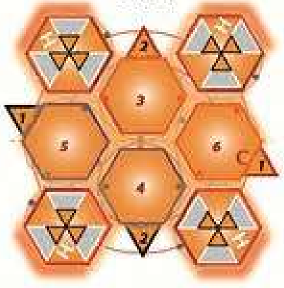

Covalent bonds

Intra-orbital bonding between
elements and compounds where
electrons exchange is the main
mechanisism resulting in
stable electronic configurations

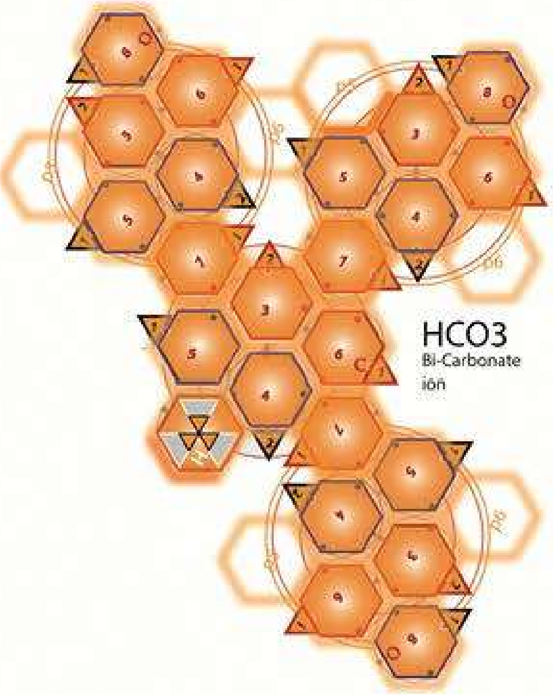

HCO3
Bi-Carbonate
ion

Ionic bonds

Extra-orbital bonding between
elements and compounds where
element charge attraction is the
predominent mechanism
with electrons sharing resulting in
stable electronic configurations

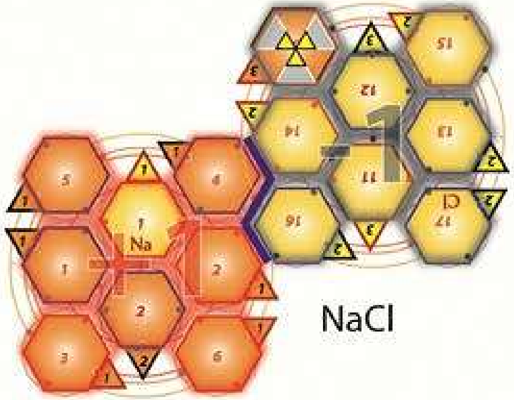

NaCl

Copyright ABRAHAM [2008] - All rights reserved

CORE ELECTRON configurations

			[He]
L	1s2	2s2 2p6	[Ne]
M	1s2 2s2 2p6	3s2 3p6	[Ar]
N	1s2 2s2 2p6 3s2 3p6 3d10	4s2 4p6	[Kr]
O	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10	5s2 5p6	[Xe]
P	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10 4f14 5s2 5p6 5d10	6s2 6p6	[Rn]
Q	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10 4f14 5s2 5p6 5d10 5f14 6s2 6p6 6d10	7s2 7p6	[Uuo]

The term "core electrons" or "noble gas core" refers to the electrons within the atom which have the same electron configuration as the nearest noble gas of lower atomic number and contain filled s&p orbitals

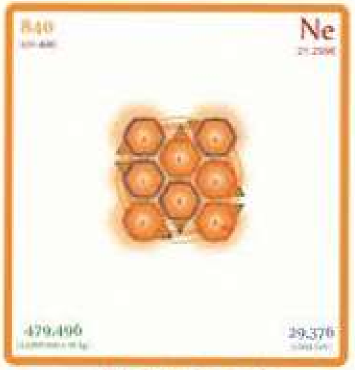
840
Ne
479,496
29,376

[He] 2s2 2p6
2

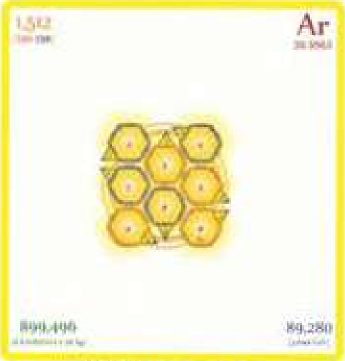
1,512
Ar
899,496
89,280

[Ne] 3s2 3p6
10

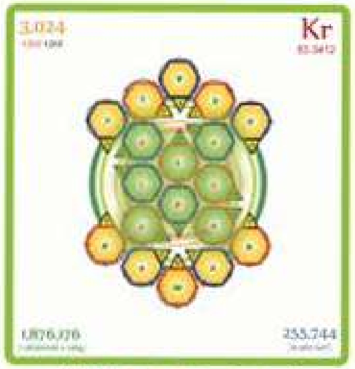
3,024
Kr
1,876,176
255,744

[Ar] 3d10 4s2 4p6
18

'Nobel Gas' configuration
full s,p orbitals

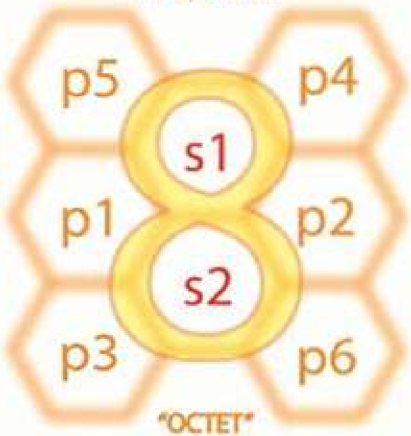
p5
p4
s1
p1
p2
s2
p3
p6

'core electron' configurations are present in energy levels L [2] - [7] Q

electrons not in filled or complete atomic orbitals are valence electrons

"OCTET"
No valance

Using core electron notation greatly reduces complexity of the electron configuration notations of larger elements in the periodic table

[Ub] 1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 5s2 4d10 5p6 4f14 5d10 6s2 6p6 5f14 6d10 7s2 7p6 8s2
120
2 8 18 32 32 18 8 2

They are the electrons in the inner part of the atom that are not valence electrons and therefore do not participate in bonding

[Uuo] 8s2
118

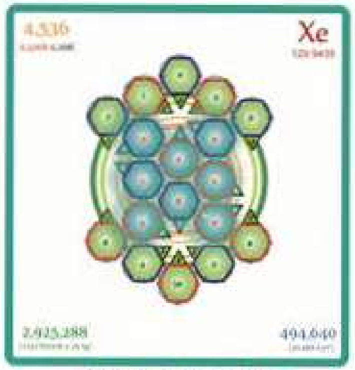
4,536
Xe
2,925,288
494,640

[Kr] 4d10 5s2 5p6
36

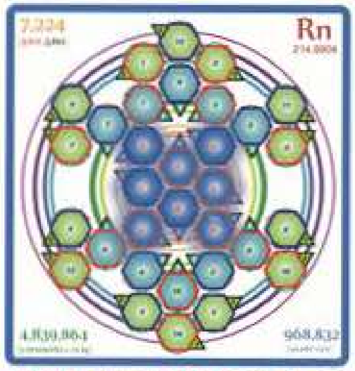
7,224
Rn
4,839,864
968,832

[Xe] 4f14 5d10 6s2 6p6
54

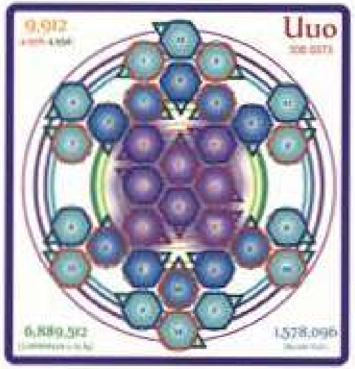
9,912
Uuo
6,889,512
1,578,096

[Rn] 5f14 6d10 7s2 7p6
86

Core electron configurations

Core electrons are the electrons in an atom that are not valence electrons and therefore do not participate in bonding.

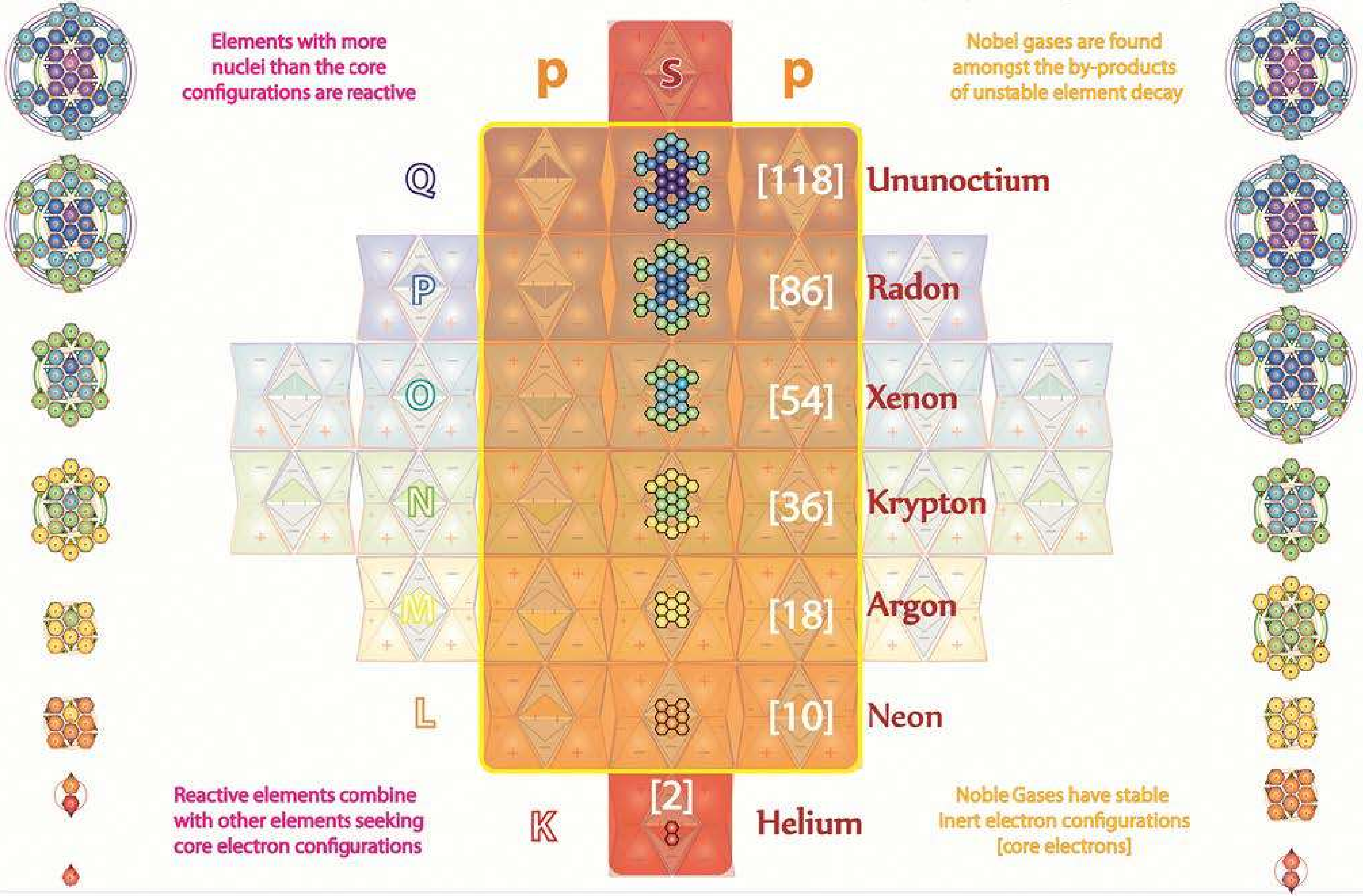

Reactive
Inert
p
s
p
Q
[118]
Ununoctium
P
[86]
Radon
O
[54]
Xenon
N
[36]
Krypton
M
[18]
Argon
L
[10]
Neon
K
[2]
Helium

Elements with more nuclei than the core configurations are reactive

Nobel gases are found amongst the by-products of unstable element decay

Reactive elements combine with other elements seeking core electron configurations

Noble Gases have stable Inert electron configurations [core electrons]

Copyright ABRAHAM [2008] - All rights reserved

Helium is a unique noble gas in that it only has filled s-orbitals

suggesting that it be more appropriate to designate it as an alkaline earth gas

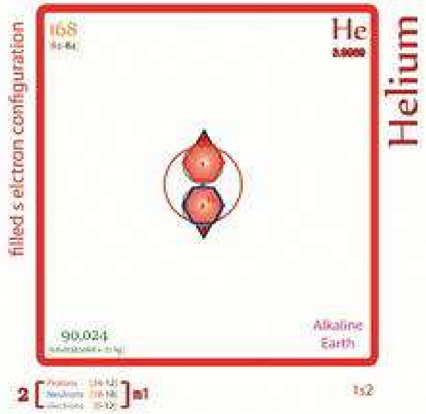
filled s elctron configuration
168
He
Helium
90,024
Alkaline Earth
2
Protons
Neutrons
electrons
m1
1s2

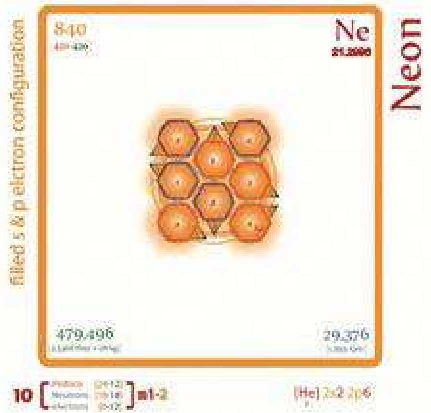
filled s & p elctron configuration
840
Ne
Neon
479,496
29,376
10
Protons
Neutrons
electrons
m1-2
[He] 2s2 2p6

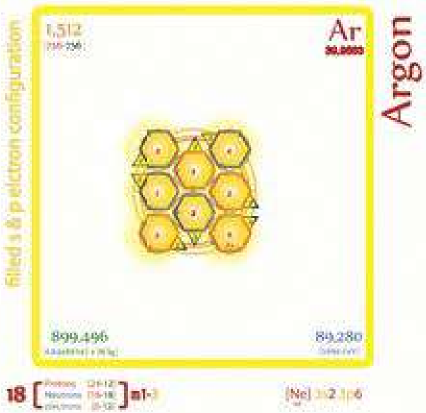
filled s & p elctron configuration
1,512
Ar
Argon
899,496
89,280
18
Protons
Neutrons
electrons
m1-3
[Ne] 3s2 3p6

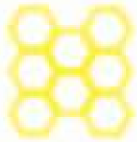

Nobel gases all have inert core sp electron configurations

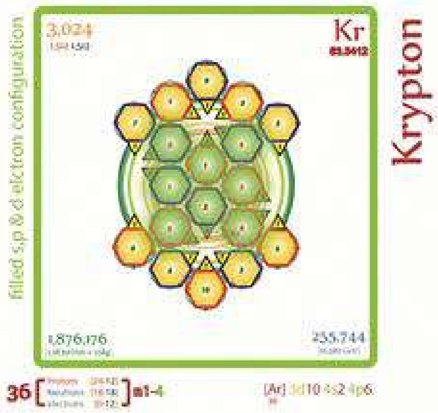
filled s,p & d elctron configuration
3,024
Kr
Krypton
1,876,176
255,744
36
Protons
Neutrons
electrons
m1-4
[Ar] 3d10 4s2 4p6

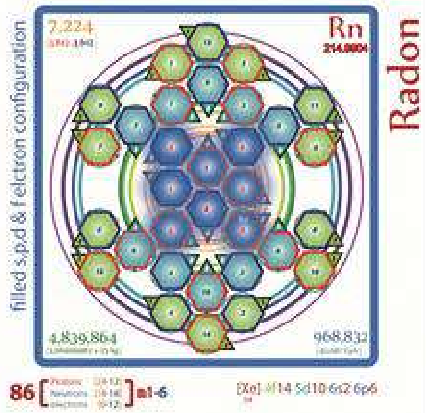
filled s,p,d & f elctron configuration
7,224
Rn
Radon
4,839,864
968,832
86
Protons
Neutrons
electrons
m1-6
[Xe] 4f14 5d10 6s2 6p6

The Nobel Gases

are nuclei whose electron orbitals are completely filled
(and consequently have no valence electrons for bonding)

2			[He]
10	1s2	2s2 2p6	[Ne]
18	1s2 2s2 2p6	3s2 3p6	[Ar]
36	1s2 2s2 2p6 3s2 3p6 3d10	4s2 4p6	[Kr]
54	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10	5s2 5p6	[Xe]
86	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10 4f14 5s2 5p6 5d10	6s2 6p6	[Rn]
118	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10 4f14 5s2 5p6 5d10 5f14 6s2 6p6 6d10	7s2 7p6	[Uuo]

All nobel gases have filled electron orbitals creating inert, chemically un-reactive elements

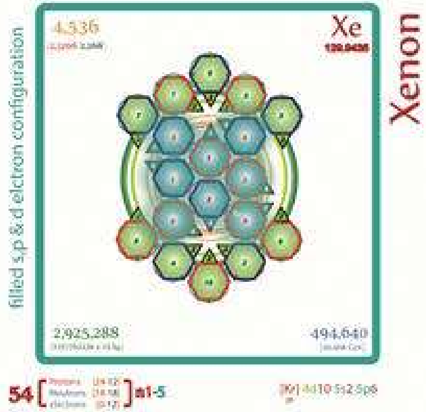
filled s,p & d elctron configuration
4,536
Xe
Xenon
2,925,288
494,640
54
Protons
Neutrons
electrons
m1-5
[Kr] 4d10 5s2 5p6

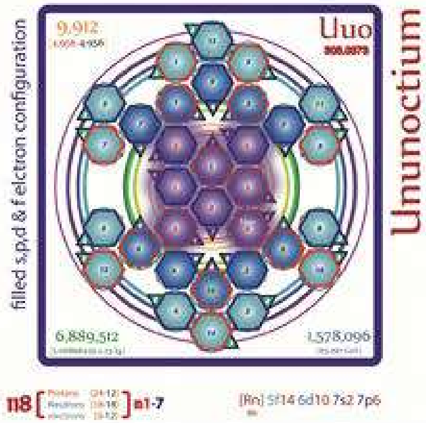
filled s,p,d & f elctron configuration
9,912
Uuo
Ununoctium
6,889,512
1,578,096
118
Protons
Neutrons
electrons
m1-7
[Rn] 5f14 6d10 7s2 7p6

The noble gases are a group of chemical elements with very similar properties: under standard conditions, they are all odorless, colorless, monatomic gases, with a very low chemical reactivity.

Whilst some of the nobel gases are often termed core electron configurations the large elements also have filled d & f orbitals in them

The six noble gases that occur naturally are Helium (He), Neon (Ne), Argon (Ar), Krypton (Kr), Xenon (Xe), and the radioactive noble gases are Radon (Rn) and Ununioctium (Uuo).

Molecular Bonds

Hydrogen bonds

A weak, primarily electrostatic, bond between a hydrogen atom bound to a highly electronegative element in a given molecule and a second highly electronegative atom in another molecule

Diatomic bonds are fascia interactions between molecules with the same nuclei number [O_2, F_2, I_2, Cl_2]

Sigma bonds

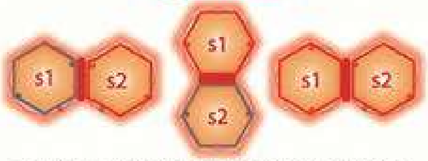
s1
s2
s1
s2
s1
s2

The residual EM force between opposite charge fascia in nuclei holds individual nuclei together in turn resulting in larger molecules and compounds

Pi bonds

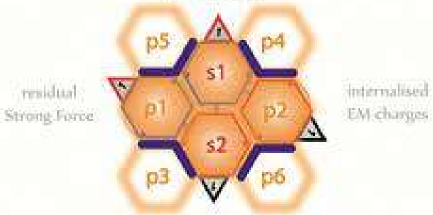
p5
p4
s1
residual Strong Force
p1
p2
internalised EM charges
s2
p3
p6

Molecular bonds

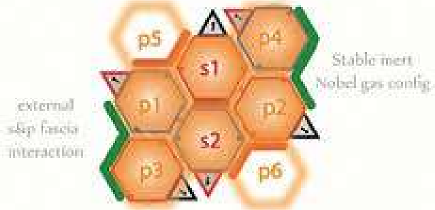
p5
p4
s1
Stable inert Nobel gas config
p1
p2
external s&p fascia interaction
s2
p3
p6

A molecular bond is an attractive force between nuclei that allows two or more elements or atoms to combine into a singular, complex compound structure

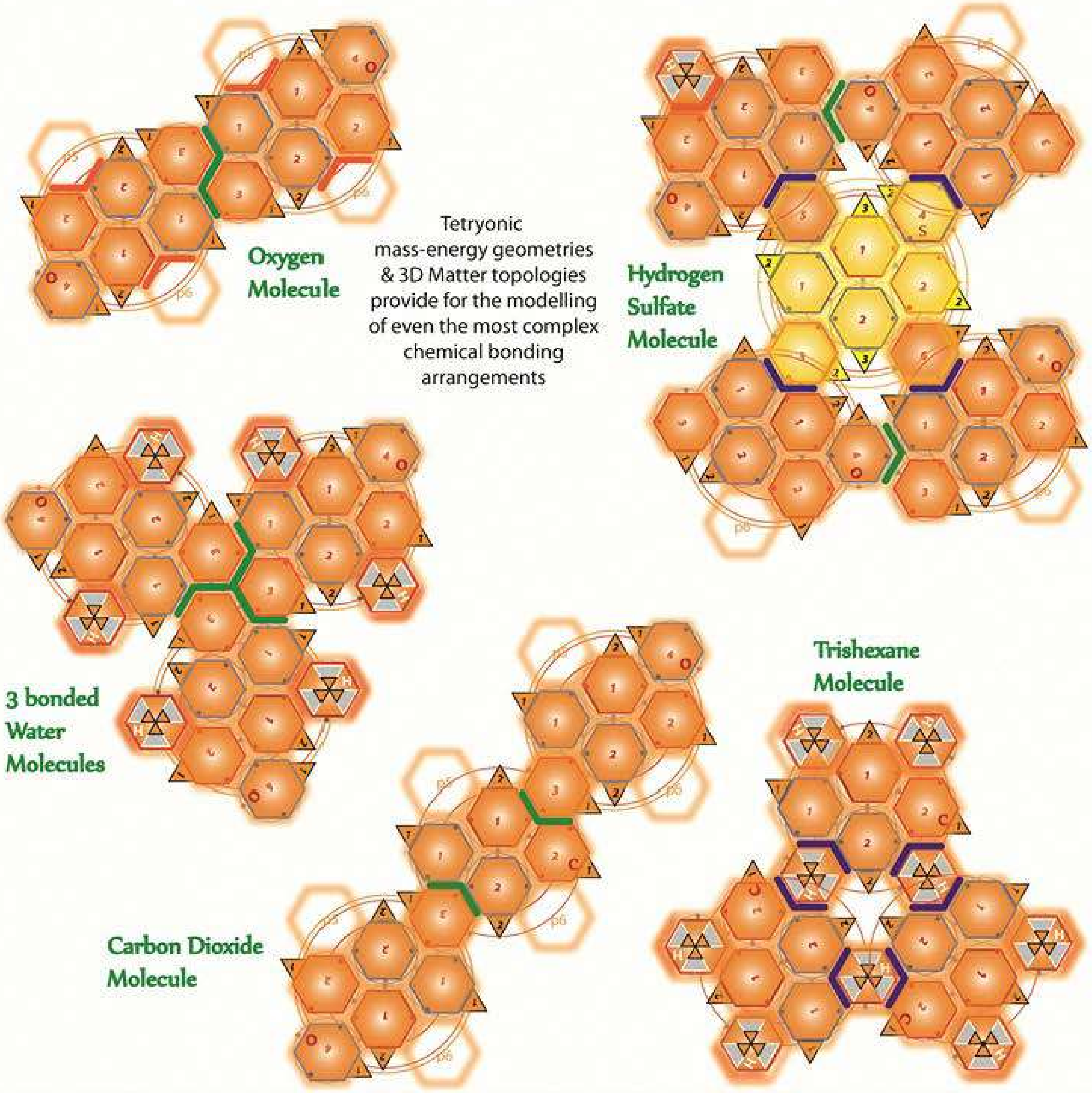

Oxygen Molecule

Tetryonic mass-energy geometries & 3D Matter topologies provide for the modelling of even the most complex chemical bonding arrangements

Hydrogen Sulfate Molecule

3 bonded Water Molecules

Trishexane Molecule

Carbon Dioxide Molecule

Covalent bonds are chemical links between two atoms in which electrons are shared between them.

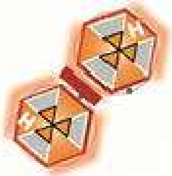

Sigma bonds are covalent bonds formed by direct overlapping of two adjacent atom's outermost orbitals.

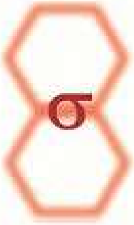

Note: sigma bonds can be formed by the bonding of either s-orbitals [or two adjacent p-orbitals]

Lewis electron dot diagrams fail to illustrate reality in that molecules exist as 3D objects and not as a two dimensional systems as shown by them.

Tetryonic geometry & topologies provide a polar view of 3D atomic nuclei that can be viewed as an exact representation of what a molecule & its bonds would look like when viewed from above.

Sigma & Pi covalent bonds

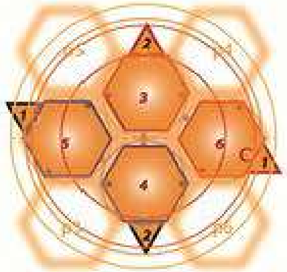

A pi bond is a covalent bond formed between two neighboring atom's unbonded p-orbitals.

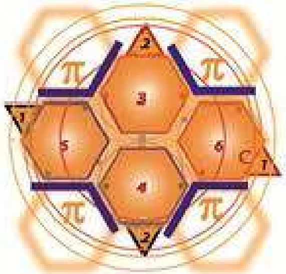

The single electrons from each atom's p orbital combine to form an electron pair creating the sigma bond.

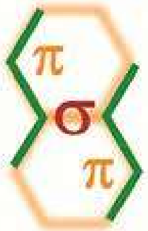

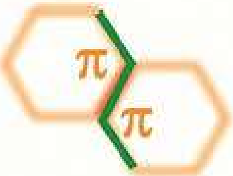

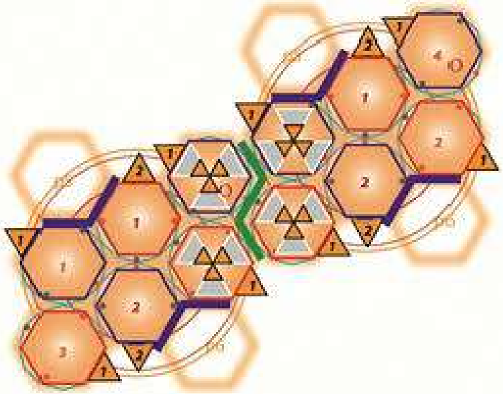

Double and triple bonds between atoms are usually made up of a single sigma bond and one or two pi bonds

π—σ—π

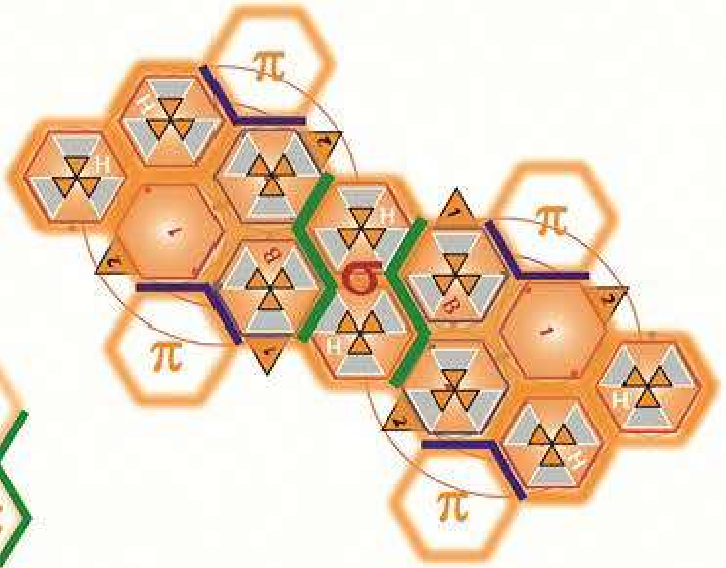

Copyright ABRAHAM [2008] - All rights reserved

Alkali Metals

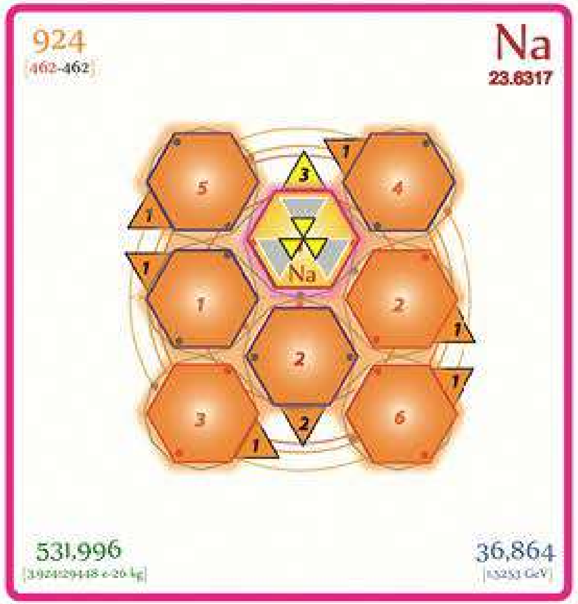
924
[462-462]
Na
23.6317
5
4
3
1
1
2
1
2
1
3
2
1
6
Na
531,996
[3.924129448 e-26 kg]
36,864
[1.5253 GeV]

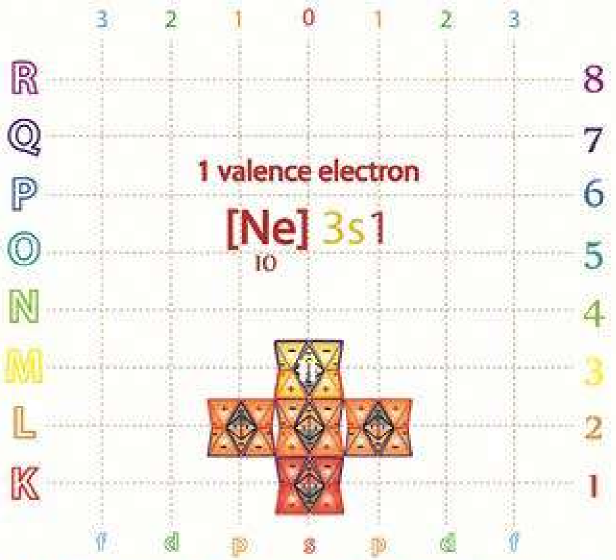
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1 valence electron
[Ne] 3s1
10
f
d
p
s
p
d
f

Atoms with more electrons than a closed shell are also highly reactive, as the extra valence electrons are easily removed from that orbital (to form a positive ion)

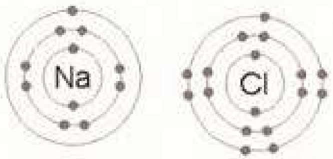

Valence electrons

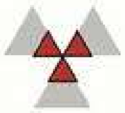

are the highest energy electrons in an atom forming the outermost electrons of an atom, and are important in determining how the atom reacts chemically with other atoms

Historically, the number of valence electrons was reflected by the element's group number in the Mendeleev table and formed the basis of elemental families

Tetryonic topologies now replace the older, incorrect models of valence electron configurations with the full 3D modelling of all atoms, elements & compounds

[zz] are core electron configurations

providing a superior visual means of accurately determining the energies and position of any electron in chemical compounds and determining the valence numbers

Halogens

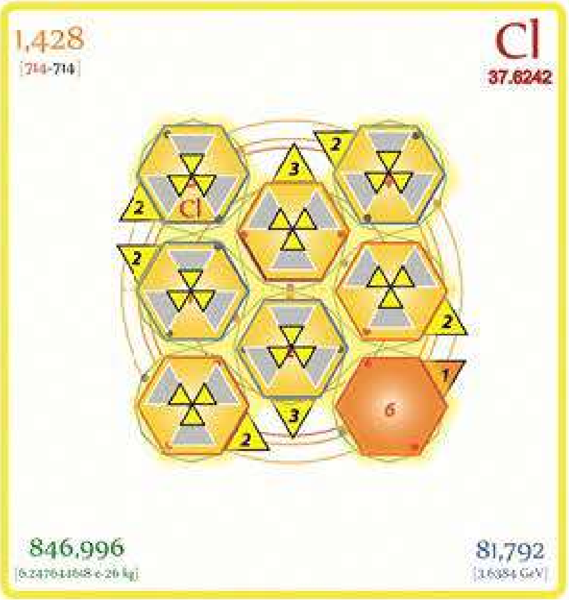
1,428
[714-714]
Cl
37.6242
2
3
2
Cl
2
2
1
3
2
6
846,996
[6.247644848 e-26 kg]
81,792
[3.6384 GeV]

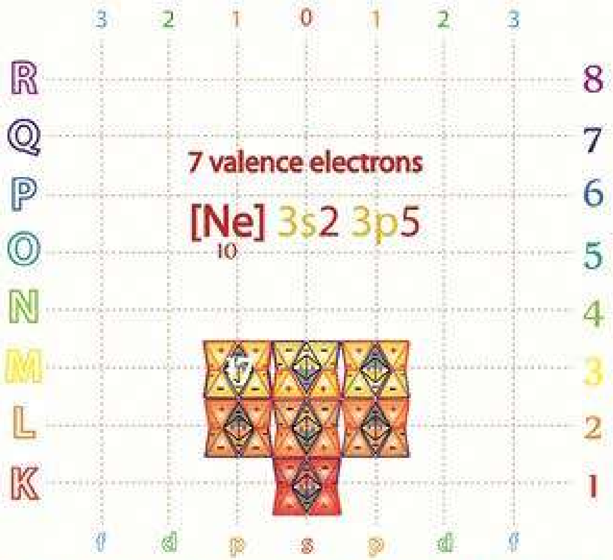
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7 valence electrons
[Ne] 3s2 3p5
10
f
d
p
s
p
d
f

Atoms one or two valence electrons short of a closed shell are highly reactive, due to their tendency to seek to gain the missing valence electrons (thereby forming a negative ion)

Copyright ABRAHAM [2008] - All rights reserved

Hydro-Chloric acid

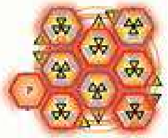

HCl

HNO3

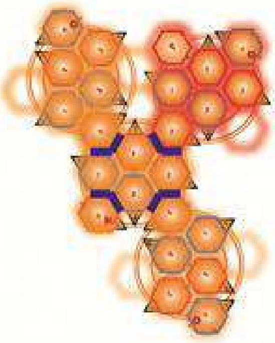

Nitric acid

Carbonic acid

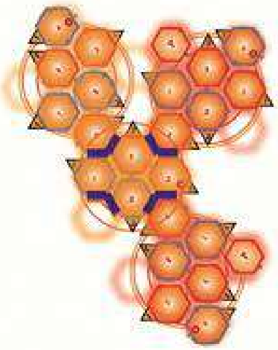

H2CO3

H2SO4

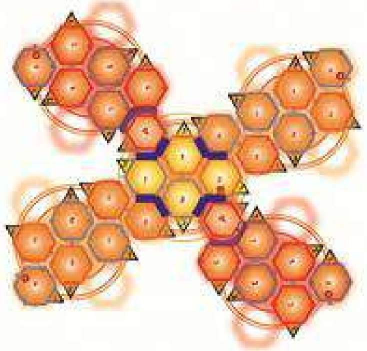

Sulfuric acid

Common Acids

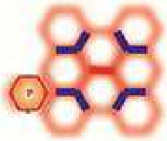

The Arrhenius definition defines acids as substances which increase the concentration of hydrogen ions (H+), or more accurately, hydronium ions (H3O+), when dissolved in water.

The Brønsted-Lowry definition is the most widely used definition where acid-base reactions are assumed to involve the transfer of a proton (H+) from an acid to a base.

A Lewis acid is a substance that can accept a pair of electrons to form a covalent bond.

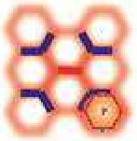

HC2H3O2

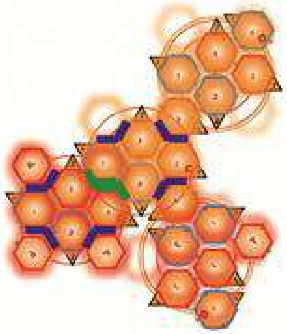

Acetic acid

Phosphoric acid

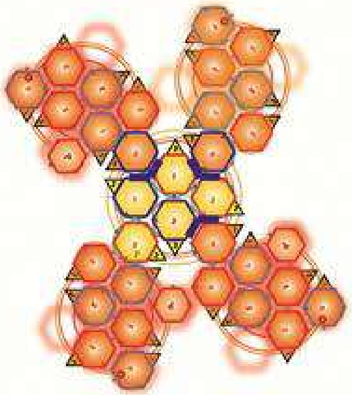

H3PO4

H2C2O4

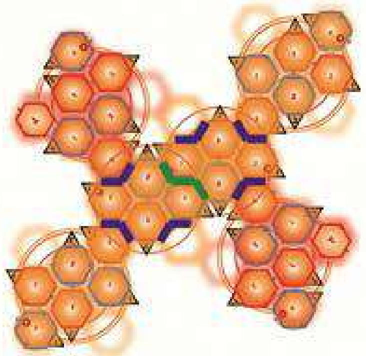

Oxalic acid

Citric acid

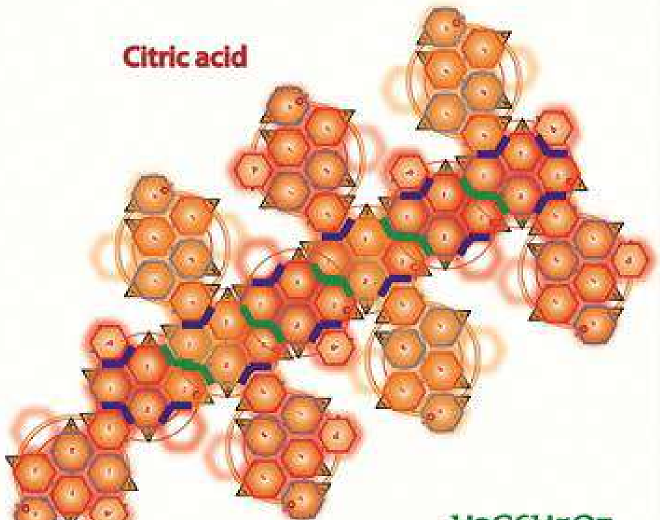

H3C6H5O7

Chemical reactions

Valence Bonds & Molecular orbitals

A chemical reaction is a process that leads to the transformation of one set of chemical substances to another.

Valence bond theory focuses on how the atomic orbitals of the dissociated atoms combine in molecular formations to give rise to individual chemical bonds.

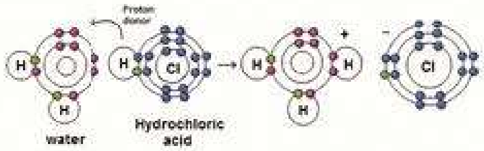

In contrast, molecular orbital theory has orbitals that are designed to cover the whole molecule but is mathematically more complex in application

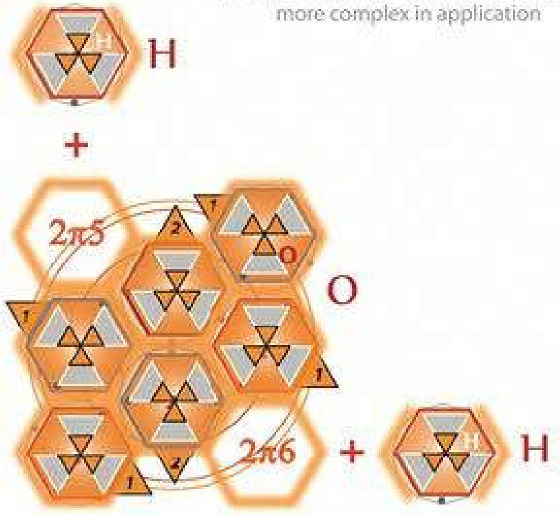

6 valence electrons
2 [n2] valence bonds available
(8 electron stable configuration)

Hydrogen Chloride gas

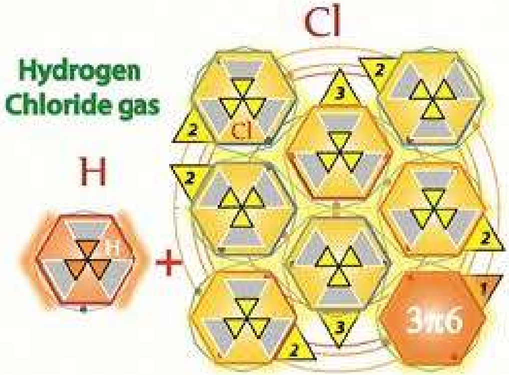

7 valence electrons
1 [n3] valence bond available
(8 electron stable configuration)

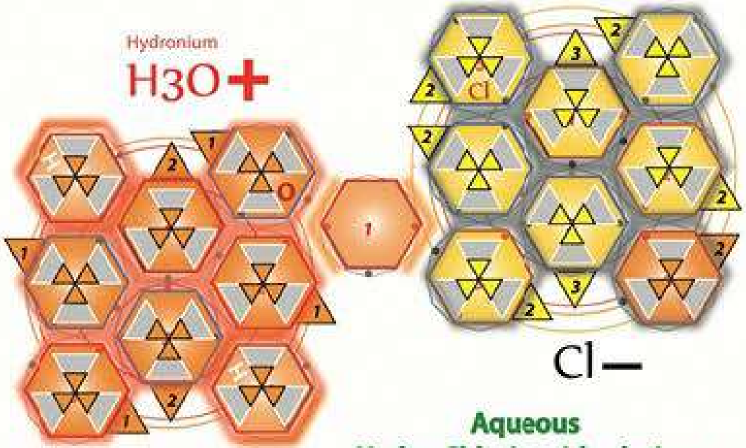

Aqueous Hydro-Chloric acid solution

Tetryonic theory replaces both VB & MO theories by modelling the physical quantum charge topologies and rest mass-energy geometry of all elements and compounds

In order to form Hydro-chloric acid [HCl aq] and bond with similar water [H2O] molecules, the n2 Hydrogen atom's energy must be raised so it can migrate to the [3π6] orbital and bond

[n2] Hydrogen atom cannot bond in the [3π6] valence position so in HCl gas it forms a [n2] di-atomic bond instead and its electron migrates to the [π6] orbital

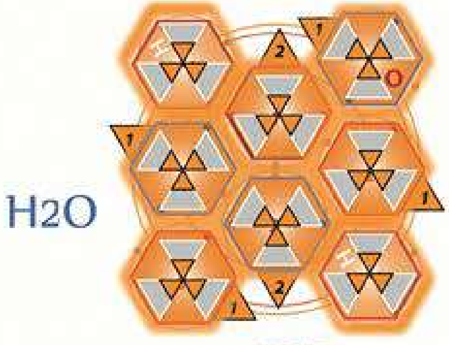

water

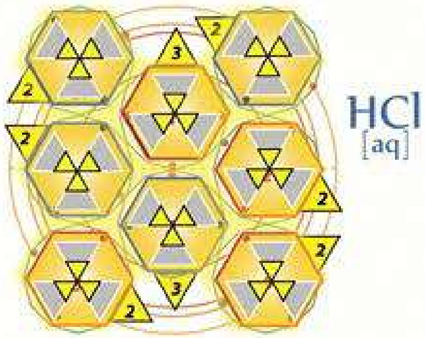

Hydrochloric acid

Valence configurations seek to create stable filled obital configurations

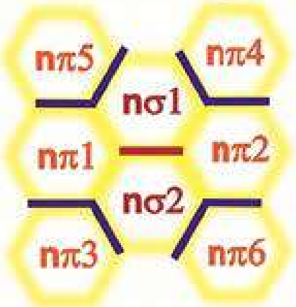

Building on exisiting electron configuration nomenclature tetryonic theory provides a new bonding schema for elements and complex compounds alike that is fully reflective of the physical quantum charge interactions

Ionic Bonding

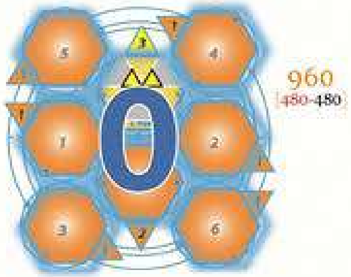
5
4
1
2
3
6
0
960
480-480

Alkaline Metal

Sodium atom gives up an electron in order to create a stable [Ne] nobel gas configuration

Sodium atom develops into a nett POSITIVE charge ion

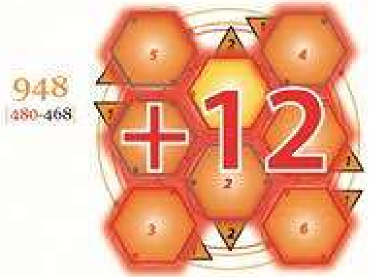
948
480-468
5
4
+12
2
3
6

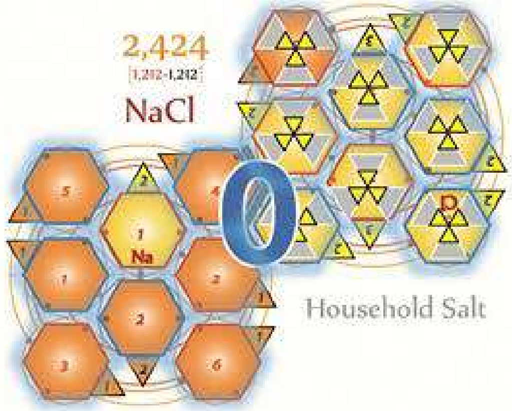
2,424
1,212-1,212
NaCl
5
1
Na
4
0
2
3
6
D
Household Salt

Resultant salt [NaCl] molecule has a NEUTRAL charge with the n3 electron migrating to fill the 3p6 orbital

12
0-12
108
Cu

OPPOSITES ATTRACT

Chlorine atom seeks an electron to order to create a stable [Ar] nobel gas configuration

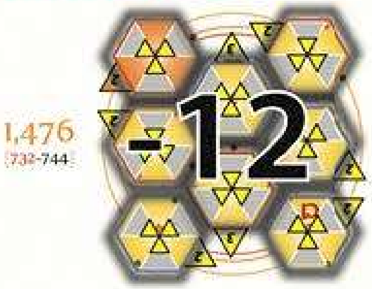
1,476
732-744
-12
D

Halogen

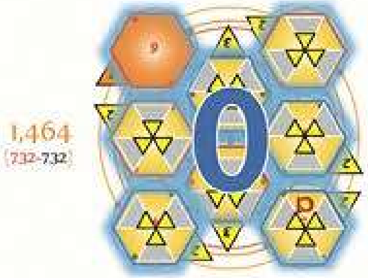
1,464
732-732
9
0
D

Chlorine atom develops into a nett NEGATIVE charge ion

All atoms and molecules seek equilibrium via stable electron configurations

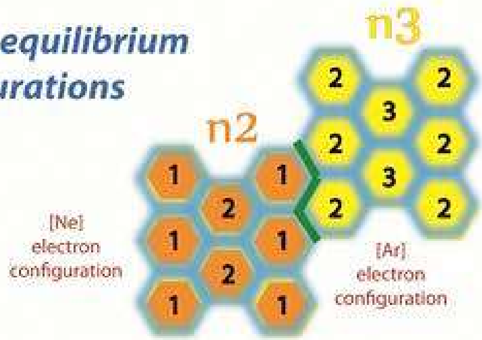
n3
n2
2
3
2
2
2
3
2
1
1
2
1
1
2
1
1
2
2
[Ne] electron configuration
[Ar] electron configuration

In short, it is a bond formed by the attraction between two oppositely charged ions.

An ionic bond (or electrovalent bond) is a type of chemical bond that can often form between metal and non-metal ions (or polyatomic ions such as ammonium) through electrostatic attraction.

The metal donates one or more electrons, forming a positively charged ion or cation with a stable electron configuration.
These electrons then enter the non metal, causing it to form a negatively charged ion or anion which also has a stable electron configuation.
The electrostatic attraction between the oppositely charged ions causes them to come together and form a bond.

Copyright ABRAHAM [2008] - All rights reserved

Valence electron rules

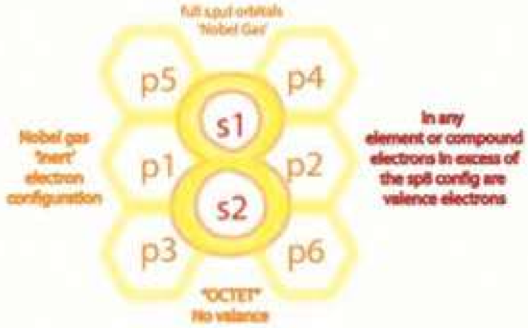
Full s,p,d orbitals
'Nobel Gas'
p5
p4
s1
p1
p2
s2
p3
p6
Nobel gas
'inert'
electron
configuration
In any
element or compound
electrons in excess of
the sp8 config are
valence electrons
"OCTET"
No valance

sp8 - core electron orbital grouping

Tetryonic quantum charge topologies
provide the complete picture

Valence electrons are the highest quantum level electrons in elemental or compound atom nuclei in excess of the core sp8 'noble gas' electron configuration

They are located in the outer most shells (ie K-Q or Quantum 1-7) making them easily ionised from the elemental nuclei

s 1-2 [electron orbital pair]
p 1-2, 3-4, 5-6 [electron orbital pairs]
d 1-2, 3-4, 5-6, 7-8 , 9-10 [electron orbital pairs]
f 1-2, 3-4, 5-6, 7-8, 9-10, 11-12, 13-14 [electron orbital pairs]

Valance electron congigurations always include 's' orbital electrons due to their higher quantum levels

Developed from Lewis diagrams
is usually only useful for elements Z<20

Fluorine
[He] 2s2, 2p5
Valence - 7

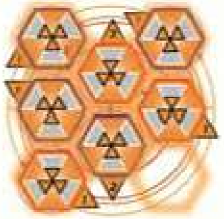

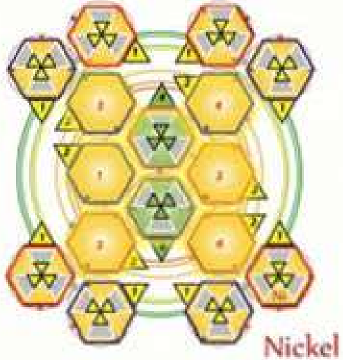

Nickel
[Ar] 3d8, 4s2
Valence - 10

Gold
[Xe] 4f14, 5d9, 6s2
Valence - 11

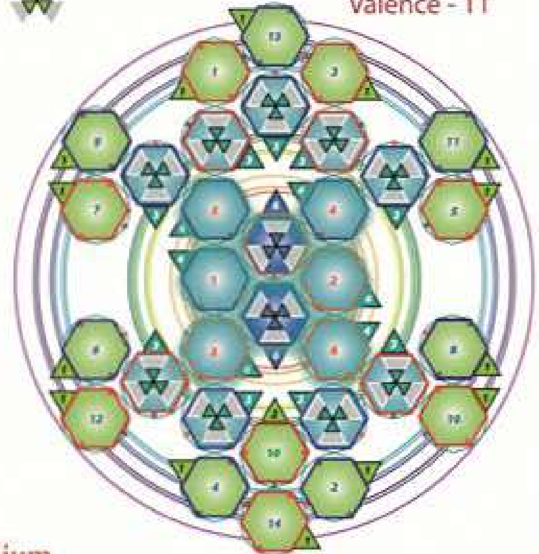

Scandium
[Ar] 3d1, 4s2
Valence - 3

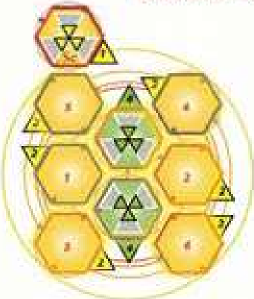

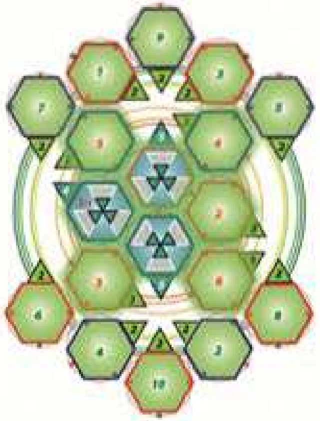

Indium
[Kr] 4d10, 5s2, 5p1
Valence - 3

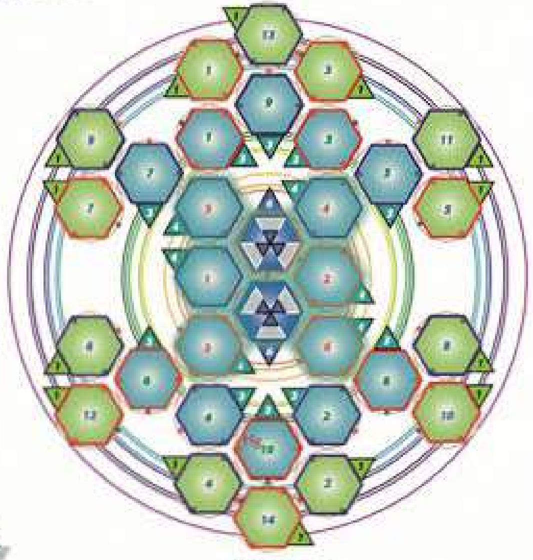

Mercury
[Xe] 4f14, 5d10, 6s2

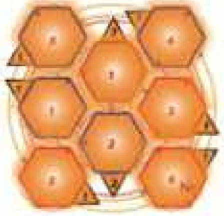

Neon
[He] 2s2, 2p6

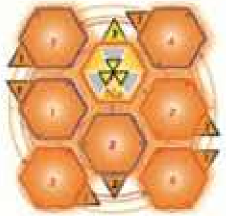

Sodium
[Ne] 3s1
Valence - 1

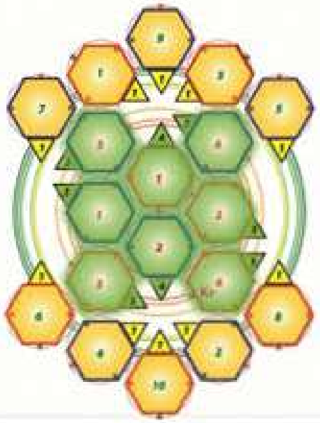

Krypton
[Ar] 4s2, 4p6

Filled electron orbitals

Once 8 valence electrons are reached a stable valence cofiguration is created

Once sub-orbit 8 (P7-8) is reached sequential valence numbering switches to paired valence numbering (as per orbital energies)

Valence numbers can be calculated by adding the 2 highest energy orbitals together (and subtracting 8 - if the total is higher than 8)

The octet rule is a simple chemical rule of thumb that states that atoms tend to combine in such a way that they each have eight electrons in their valence shells, giving them the same electronic configuration as a noble gas.

The rule is applicable to the main-group elements, especially carbon, nitrogen, oxygen, and the halogens, but also to metals such as sodium or magnesium. In simple terms, molecules or ions tend to be most stable when the outermost electron shells of their constituent atoms contain eight electrons.

In short, an element's valence shell is full and most stable when it contains eight electrons, corresponding to an s2p6 electron configuration.

CORE ELECTRONS

This stability is the reason that the noble gases are so unreactive, for example neon with electron configuration 1s2 2s2 2p6. (Helium is an exception as explained).

Note that a "full shell" means that there are the eight electrons in the valence shell when the next shell starts filling, even though higher subshells (d, f, etc.) have not been filled.

There can be at most eight valence electrons in a ground-state atom because p subshells are always followed by the s subshell of the next shell.

OCTET RULE

This means that once there are 8 valence electrons (when the p subshell is filled), the next additional electron goes into the next shell, which then becomes the valence shell.

A consequence of the octet rule is that atoms generally react by gaining, losing, or sharing electrons in order to achieve a complete octet of valence electrons. Reaction of atoms occurs primarily in two ways: ionically and covalently.

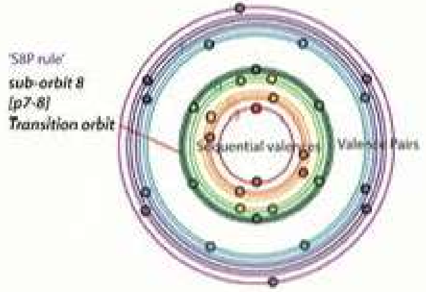
'S8P rule'
sub-orbit 8
[p7-8]
Transition orbit
Sequential valences
Valence Pairs

Once 8 valence electrons are reached the sub-orbits are stable and form a stable, non-reactive valence configuration.

Additionally, valence numbers proceed sequentially (1,2,3,4,5,6,7) up to sub-orbit 8 (P7-8) at which point all sub-orbit valences number in valence pairs (1-2, 3-4, 5-6, 7-8, 9-10, 11-2, 13-14)

The significance is that sub-orbit 8 is the middle sub-orbit (of 16 total sub-orbits possible)

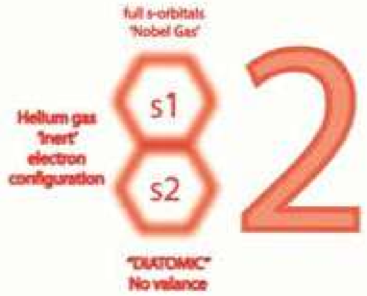
full s-orbitals
'Nobel Gas'
Helium gas
'inert'
electron
configuration
s1
s2
2
"DIATOMIC"
No valance

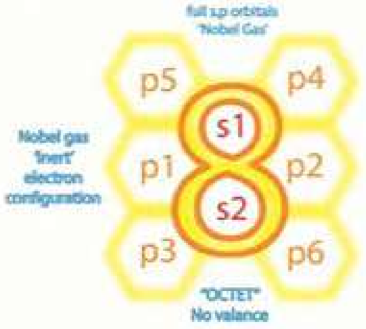
full s,p orbitals
'Nobel Gas'
Nobel gas
'inert'
electron
configuration
p5
p4
s1
p1
8
p2
s2
p3
p6
"OCTET"
No valance

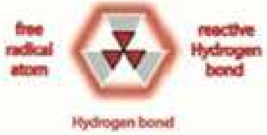
free
radical
atom
reactive
Hydrogen
bond
Hydrogen bond

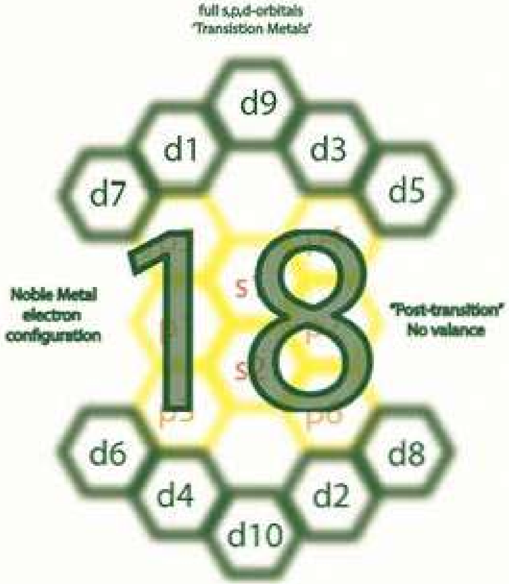
full s,p,d-orbitals
'Transistion Metals'
d9
d1
d3
d7
d5
Noble Metal
electron
configuration
18
"Post-transition"
No valance
d6
d8
d4
d2
d10

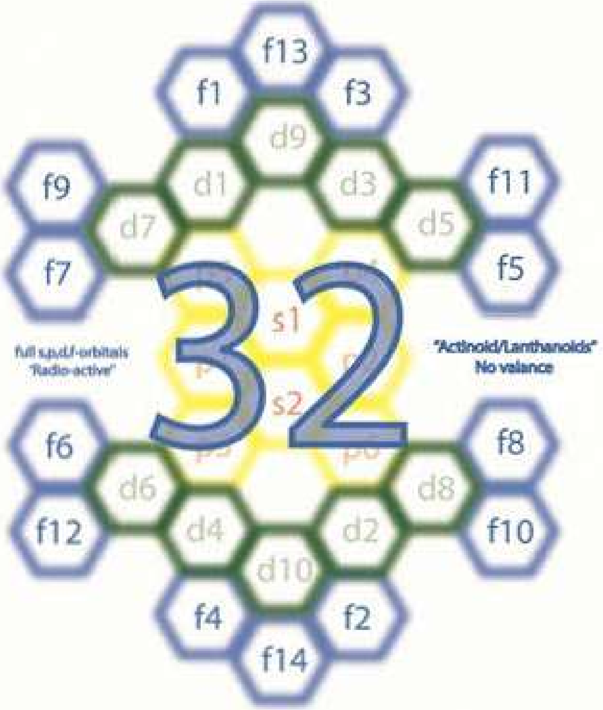
f13
f1
f3
d9
f9
d1
d3
f11
d7
d5
f7
f5
32
s1
s2
full s,p,d,f-orbitals
'Radio-active'
"Actinoid/Lanthanoids"
No valance
f6
f8
d6
d8
f12
d4
d2
f10
d10
f4
f2
f14

ACID

Hydrochloric acid is a clear, colourless solution of hydrogen chloride in water, it is a highly corrosive, strong mineral acid

Aqueous Hydro-Chloric acid solution

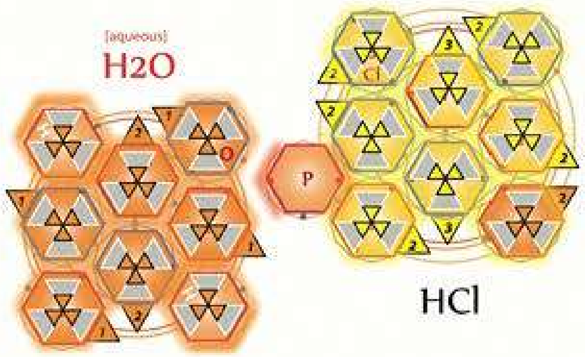
[aqueous]
H_2O
P
Cl
HCl

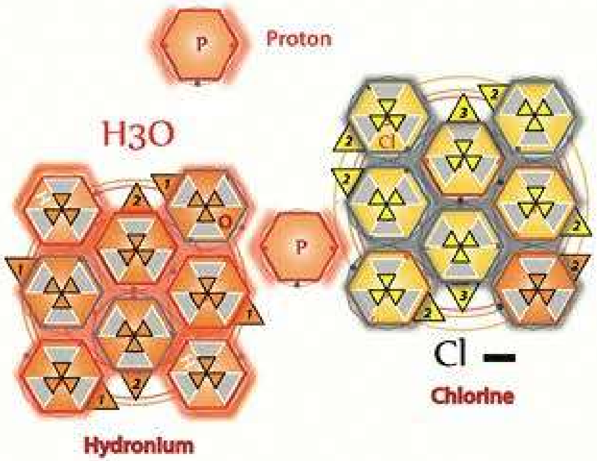
P
Proton
H_3O
P
Cl
Cl –
Chlorine
Hydronium

The H(+) cation of the acid combines with the OH(-) anion of the base to form water.

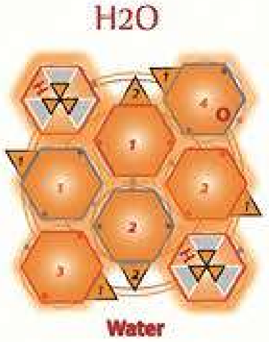
H_2O
H
O
Water

all charges seek equilibrium

+

When an acid and a base are placed together, they react to neutralize the acid and base properties, producing a salt.

$$\mathrm{HCl(aq)} + \mathrm{NaOH(aq)} \dashrightarrow \mathrm{NaCl(aq)} + \mathrm{H_2O(l)}$$

=>

combining the free Protons & electrons

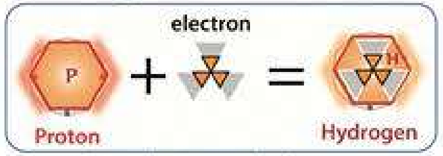
electron
P
+
=
H
Proton
Hydrogen

neutralises the ions and releases energy

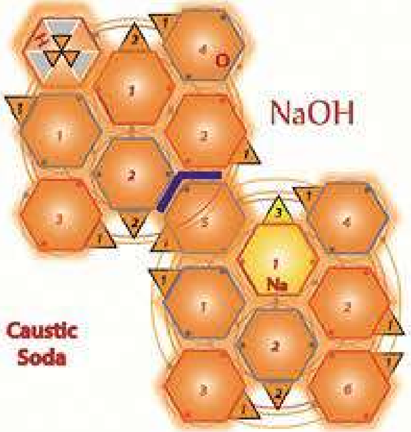
H
O
NaOH
Na
Caustic Soda

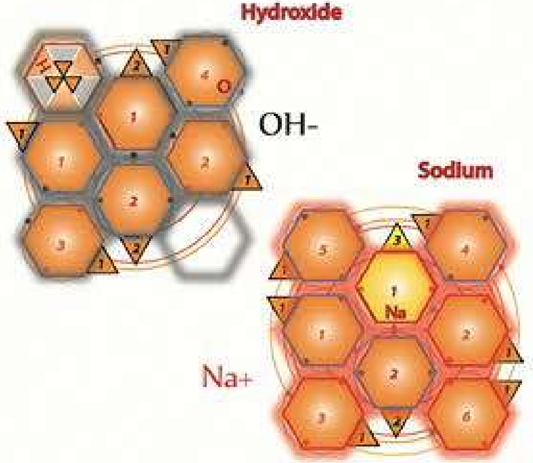
Hydroxide
H
O
OH-
Sodium
Na
Na+

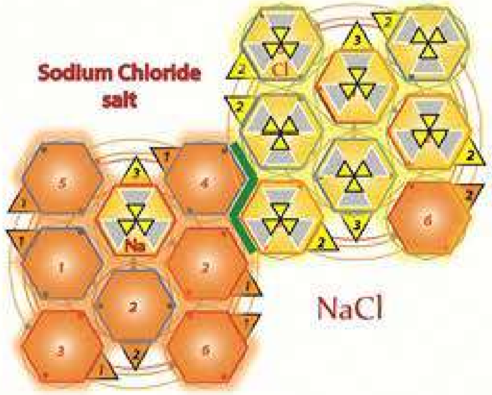
Sodium Chloride salt
Cl
Na
NaCl

The compound formed by the cation of the base and the anion of the acid is called a salt.

BASE

Sodium hydroxide, also known as caustic soda, or lye, has the molecular formula NaOH and is a highly caustic metallic base and alkali salt.

Geometric Molecular Topology

Molecules are most often held together with covalent bonds involving single, double, and/or triple bonds,
where a "bond" is a pair of electrons shared between elements as they seek equilibrium
(another method of bonding is ionic bonding and involves a positive cation and a negative anion).

Hydrogen

A Hydrogen bond is a chemical bond in which a hydrogen atom of one molecule is attracted to an electronegative atom, especially nitrogen, oxygen, or flourine atoms, usually of another molecule

The hydrogen bond is often described as an electrostatic dipole-dipole interaction.

However, it also has some features of covalent bonding: it is directional and strong, acting over interatomic distances shorter than that of van der Waals radii

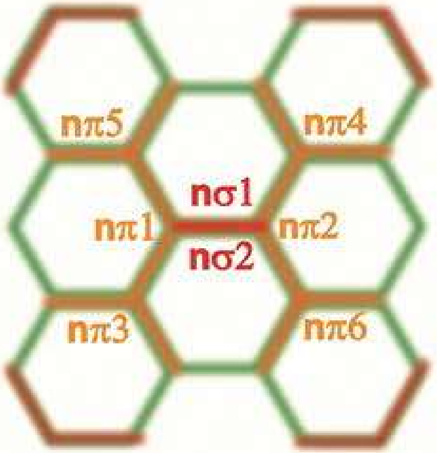

Oxygen

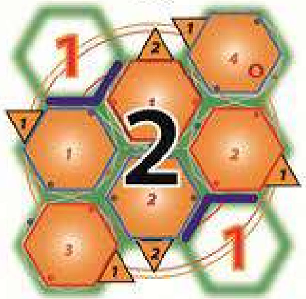

The oxidation state of oxygen is −2 in almost all known compounds

Oxides of Oxygen & oxygen molecules are found throughout the range of Organic & inorganic compounds

All molecules seek KEM'ical charge and energy equilibrium

Geometric Molecular Topology is the overall arrangement of the atoms in a molecule, where the bonded atoms in a molecule are responsible for determining the final molecular topology of a chemical system of bonded elements.

As the numbers of atoms in molecules increases, the quantum molecular topology of a system grows increasingly complex and can only be modelled accurately using Tetryonic charged geometries

by bonding together and forming larger complex molecules

Common nitrogen functional groups include: amines, amides, nitro groups, imines, and enamines.

Compounds of Carbon form the 'backbones' of Organic & inorganic compounds

Molecular Octets

The concept of the Expanded Octet occurs in any system that has an atom with more than four electron pairs attached to it.

Most commonly, atoms will expand their octets to contain a total of five or six electron pairs, in total. In theory, it is possible to expand beyond those number.

The large amounts of negative charge concentrated in small volumes of space prevent those larger expanded octets from forming.

When an atom expands its octet, it does so by making use of empty d-orbitals that are available in the valence level of the atom doing the expanding.

The atom that expands its octet in a structure will usually be located in the center of the structure and the system will not use any multiple bonds in attaching atoms to the central atom.

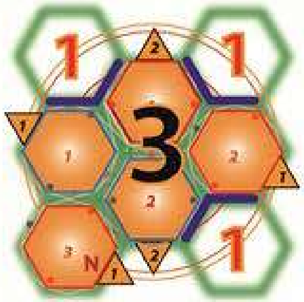

Nitrogen

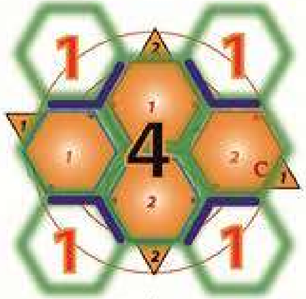

Carbon

Copyright ABRAHAM [2008] - All rights reserved

48
[24-24]

Hydrogen bonds

H

96
[48-48]

1 Proton
1 electron

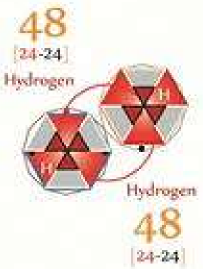
48
[24-24]
Hydrogen
H
H
Hydrogen
48
[24-24]

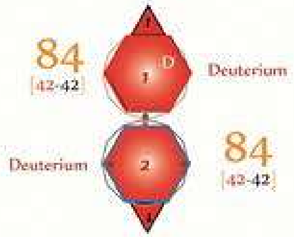
84
[42-42]
1
D
Deuterium
Deuterium
2
84
[42-42]

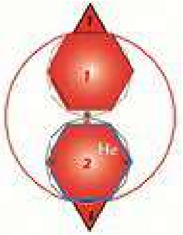

168
[84-84]

2 Protons
2 Neutrons
2 electrons

Hydrogen is a 'free radical' atom

H2 has a smaller s-orbital [lower energy level] compared to Helium

Deuterium is the building block of elements

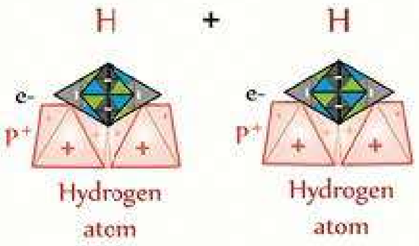
H
+
H
e-
P+
+
+
Hydrogen
atom
e-
P+
+
+
Hydrogen
atom

Hydrogen atoms form Covalent bonds

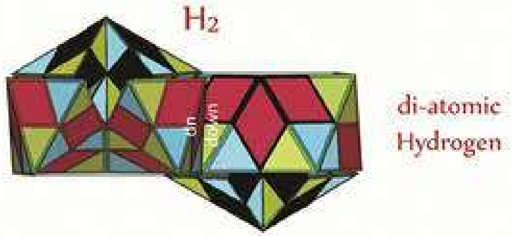
H2
di-atomic
Hydrogen

Hydrogen, bound mostly to carbon and nitrogen, is part of almost every molecule in your body: DNA, proteins, sugars, fats.
Hydrogen bonds - which form between atoms that "share" a hydrogen atom, is one of the most important interactions that makes biological molecules behave as they do.

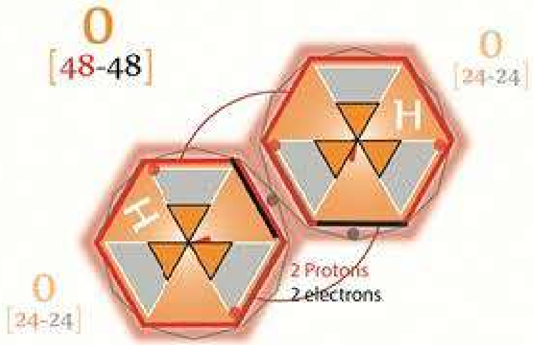
0
[48-48]
0
[24-24]
H
H
0
[24-24]
2 Protons
2 electrons

Hydrogen bonds are sigma bonds

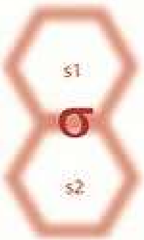

note: n2 energy level Hydrogen is illustrative only
[Hydrogen bonds can be of any energy level]

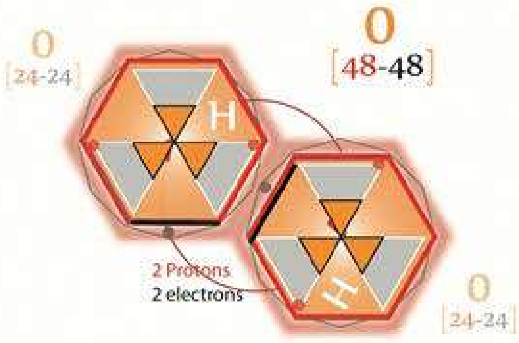
0
[24-24]
0
[48-48]
H
H
2 Protons
2 electrons
0
[24-24]

The charged fascia of Baryonic quarks facilitates chemical [hydrogen] bonding in molecules

Carbon Compounds

504
[252-252]

Carbon can bond in four places, and it can bond to itself so it's easy to make lots of different molecules.

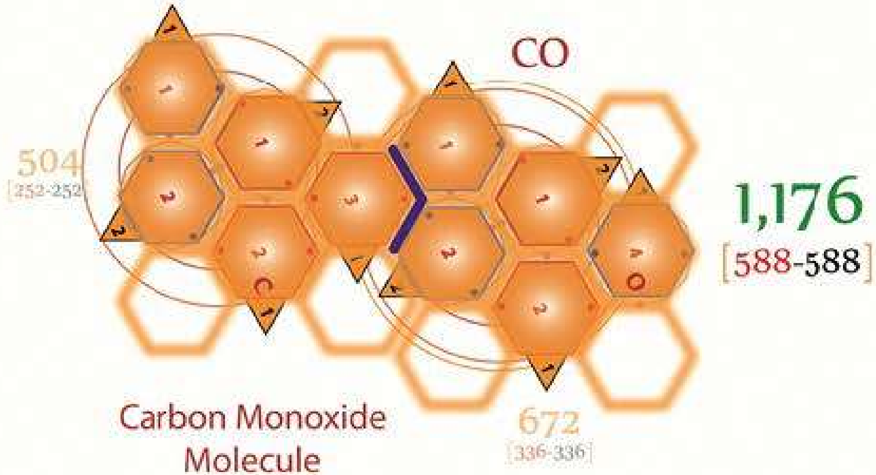
CO
504
[252-252]
1,176
[588-588]
672
[336-336]
C
O

Carbon Monoxide Molecule

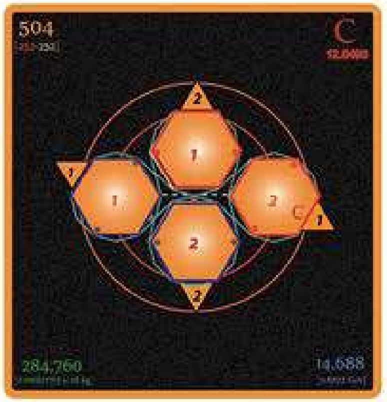
504
[252-252]
C
12.0000
284,760
14,688

[He]$_2$ 2s2 2p2

Many carbon compounds are essential for life as we know it.
Some of the most common carbon compounds are: carbon dioxide (CO2), carbon monoxide (CO), carbon disulfide (CS2), chloroform (CHCl3), carbon tetrachloride (CCl4), methane (CH4), ethylene (C2H4), acetylene (C2H2), benzene (C6H6), ethyl alcohol (C2H5OH) and acetic acid (CH3COOH) to name just a few.

Acetylene Molecule

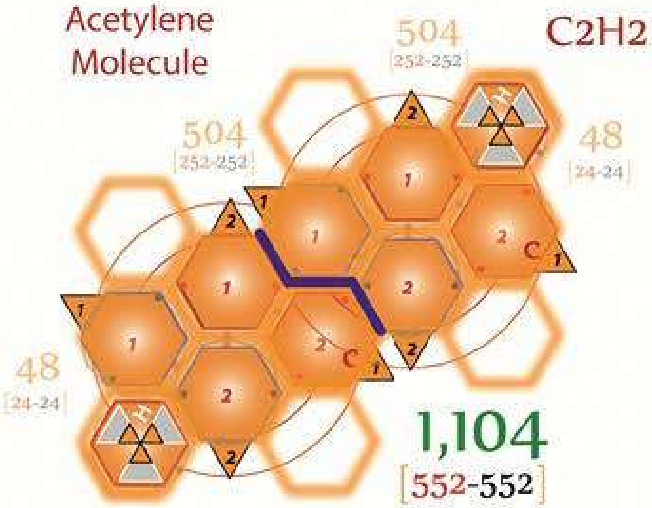
C2H2
504
[252-252]
504
[252-252]
48
[24-24]
48
[24-24]
H
C
C
H
1,104
[552-552]

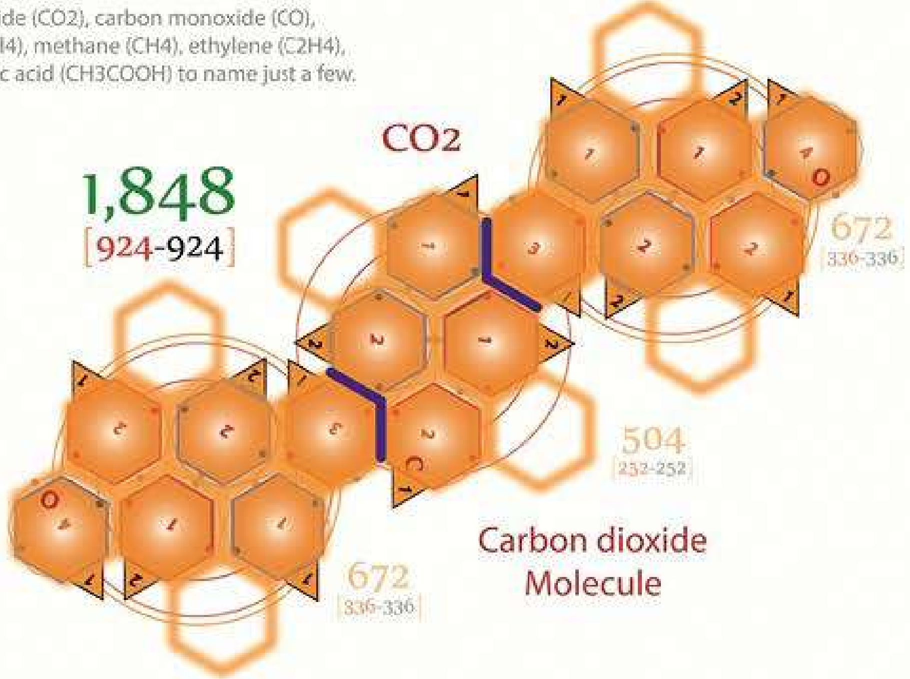
CO2
1,848
[924-924]
672
[336-336]
504
[252-252]
672
[336-336]
O
C
O

Carbon dioxide Molecule

C

There are nearly ten million known carbon compounds and an entire branch of chemistry, known as organic chemistry, is devoted to their study.

Nitrogen Compounds

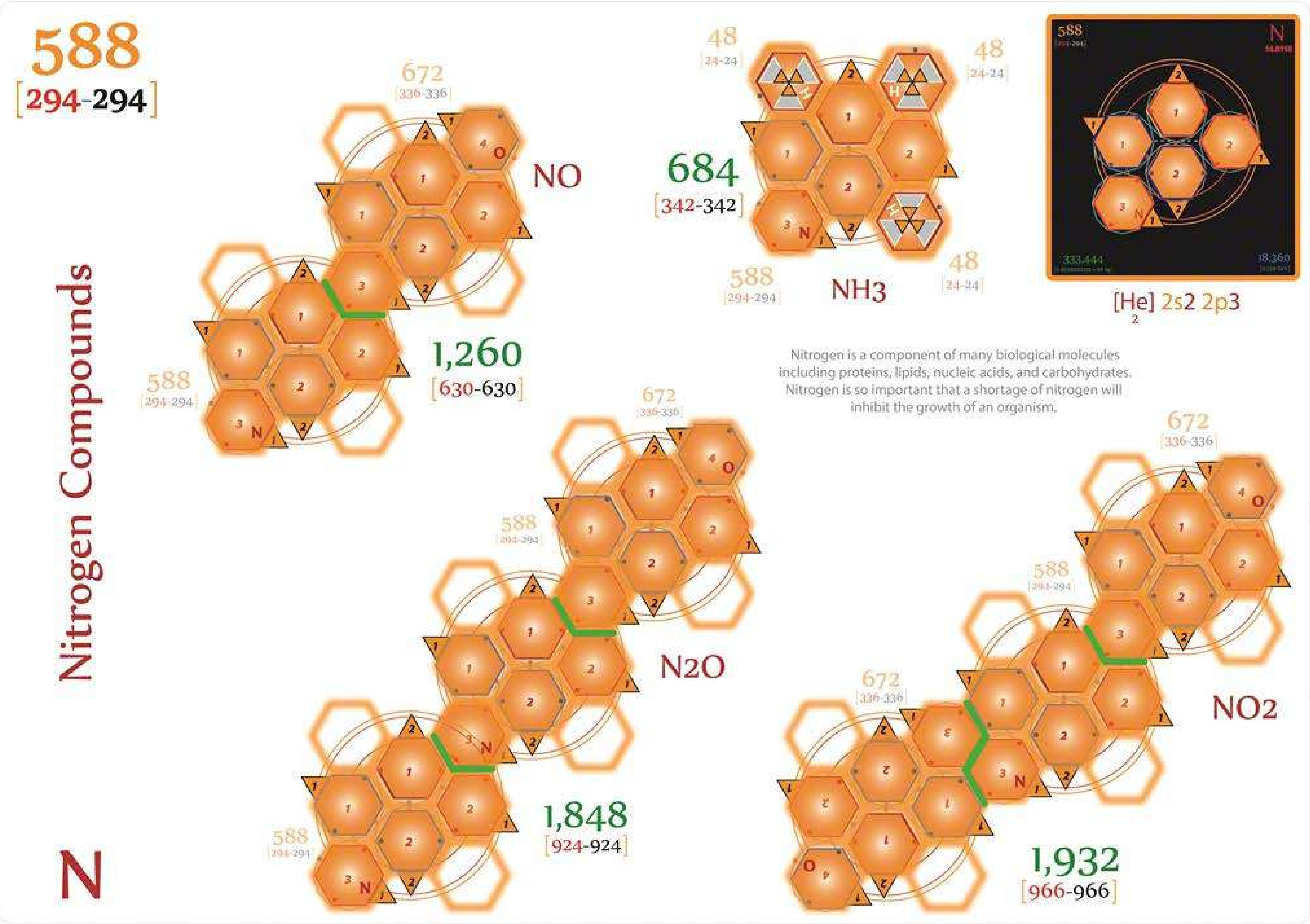
588
[294-294]
N
672
[336-336]
NO
588
[294-294]
1,260
[630-630]
48
[24-24]
48
[24-24]
684
[342-342]
588
[294-294]
NH3
48
[24-24]
588
N
333,444
18,360
[He] 2s2 2p3
672
[336-336]
588
[294-294]
N2O
588
[294-294]
1,848
[924-924]
672
[336-336]
588
[294-294]
672
[336-336]
NO2
1,932
[966-966]

Nitrogen is a component of many biological molecules including proteins, lipids, nucleic acids, and carbohydrates. Nitrogen is so important that a shortage of nitrogen will inhibit the growth of an organism.

Oxygen Compounds

672
[336-336]

O

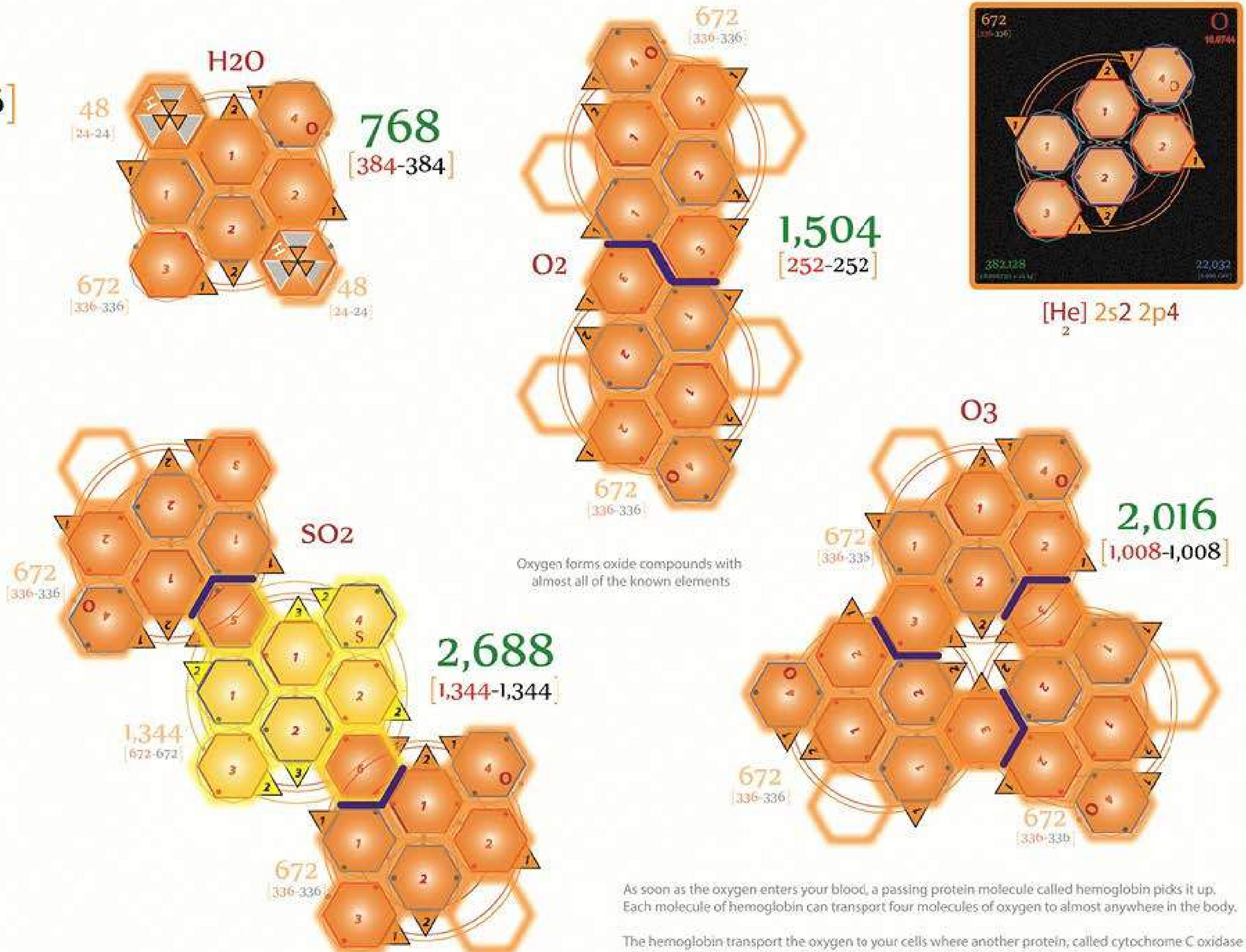
H2O
48
[24-24]
672
[336-336]
48
[24-24]
768
[384-384]
672
[336-336]
O2
1,504
[252-252]
672
[336-336]
672
[336-336]
O
382,128
22,032
672
[336-336]
SO2
2,688
[1,344-1,344]
1,344
[672-672]
672
[336-336]
O3
2,016
[1,008-1,008]
672
[336-335]
672
[336-336]
672
[336-336]

[He]$_2$ 2s2 2p4

Oxygen forms oxide compounds with almost all of the known elements

As soon as the oxygen enters your blood, a passing protein molecule called hemoglobin picks it up. Each molecule of hemoglobin can transport four molecules of oxygen to almost anywhere in the body.

The hemoglobin transport the oxygen to your cells where another protein, called cytochrome C oxidase makes two molecules of water out of every molecules of oxygen delivered to it

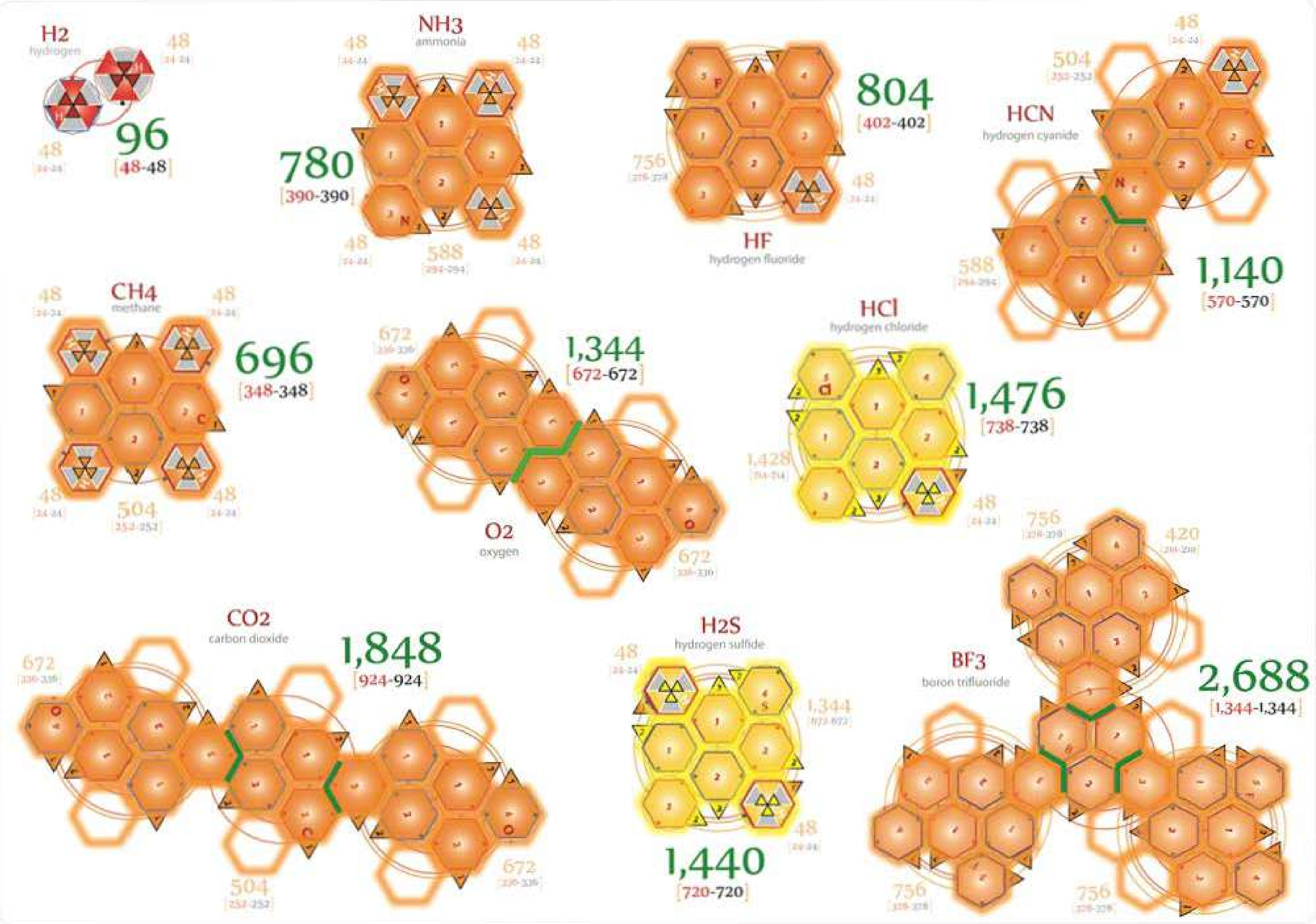
H2
hydrogen
48
[24-24]
48
[24-24]
96
[48-48]
NH3
ammonia
48
[24-24]
48
[24-24]
780
[390-390]
48
[24-24]
588
[294-294]
48
[24-24]
804
[402-402]
756
[378-378]
48
[24-24]
HF
hydrogen fluoride
48
[24-24]
504
[252-252]
HCN
hydrogen cyanide
588
[294-294]
1,140
[570-570]
CH4
methane
48
[24-24]
48
[24-24]
696
[348-348]
48
[24-24]
504
[252-252]
48
[24-24]
672
[336-336]
1,344
[672-672]
O2
oxygen
672
[336-336]
HCl
hydrogen chloride
1,476
[738-738]
1,428
[714-714]
48
[24-24]
756
[378-378]
420
[210-210]
CO2
carbon dioxide
672
[336-336]
1,848
[924-924]
504
[252-252]
672
[336-336]
H2S
hydrogen sulfide
48
[24-24]
1,344
[672-672]
48
[24-24]
1,440
[720-720]
BF3
boron trifluoride
2,688
[1,344-1,344]
756
[378-378]
756
[378-378]

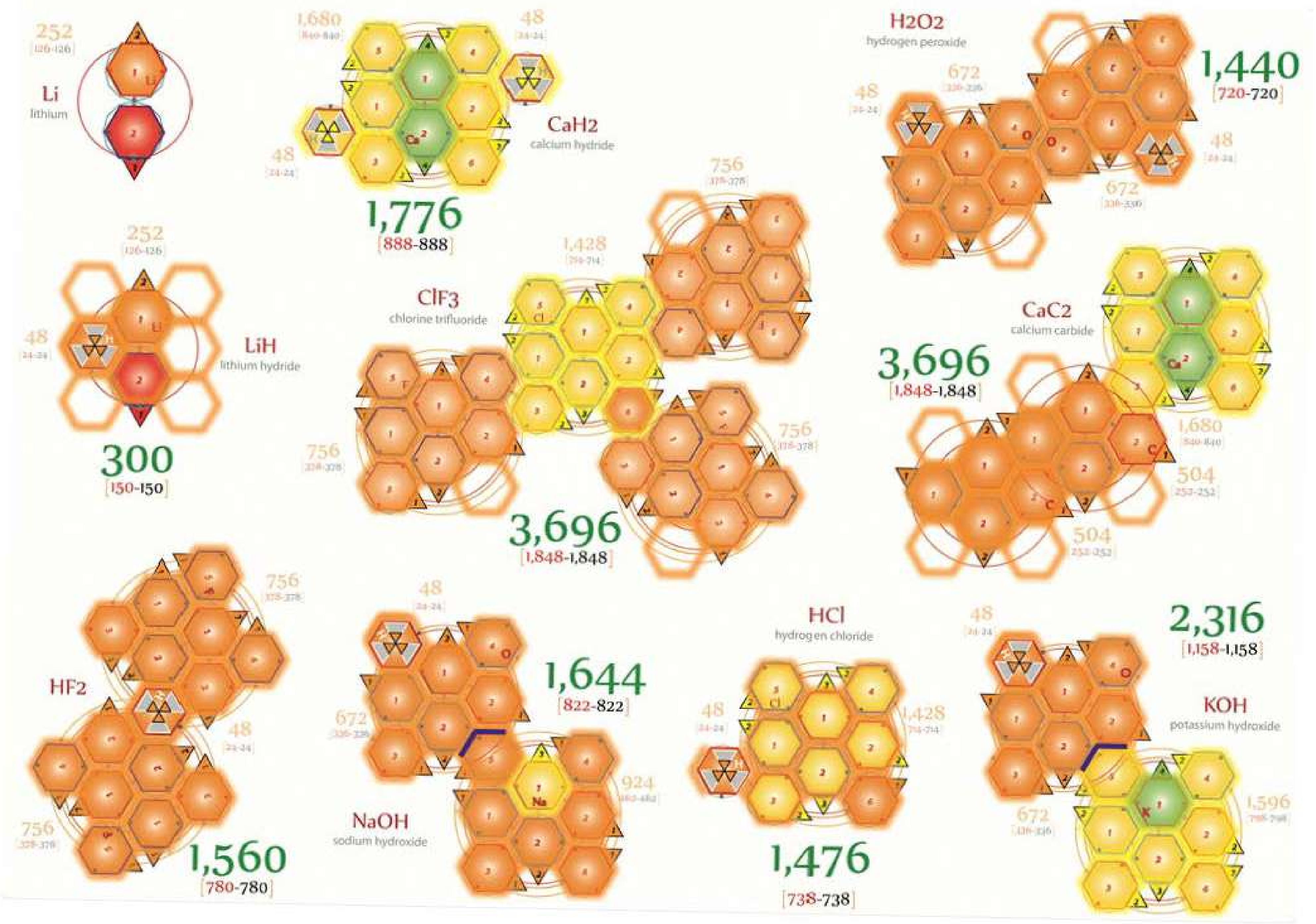
252
[126-126]
Li
lithium
1,680
[840-840]
48
[24-24]
48
[24-24]
CaH2
calcium hydride
Ca
1,776
[888-888]
H2O2
hydrogen peroxide
48
[24-24]
672
[336-336]
O
O
1,440
[720-720]
48
[24-24]
672
[336-336]
252
[126-126]
48
[24-24]
Li
LiH
lithium hydride
300
[150-150]
756
[378-378]
1,428
[714-714]
ClF3
chlorine trifluoride
Cl
756
[378-378]
756
[378-378]
3,696
[1,848-1,848]
CaC2
calcium carbide
Ca
3,696
[1,848-1,848]
1,680
[840-840]
C
C
504
[252-252]
504
[252-252]
756
[378-378]
HF2
48
[24-24]
756
[378-378]
1,560
[780-780]
48
[24-24]
O
1,644
[822-822]
672
[336-336]
Na
924
[462-462]
NaOH
sodium hydroxide
HCl
hydrogen chloride
Cl
48
[24-24]
1,428
[714-714]
1,476
[738-738]
48
[24-24]
2,316
[1,158-1,158]
O
KOH
potassium hydroxide
K
672
[336-336]
1,596
[798-798]

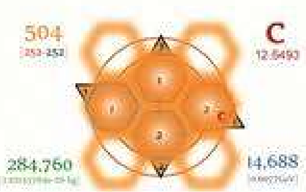
504
C
12.5493
284,760
14,688

Chemistry

is a branch of physical science, concerning the study of the composition, properties and behavior of Matter

Compounds containing bonds between carbon and a metal are called organometallic compounds.

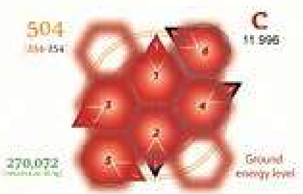
504
C
11.996
270,072
Ground energy level

Inorganic compounds are produced by non-living natural processes or in the laboratory.

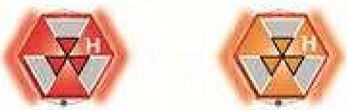

Inorganic compounds can form salts.

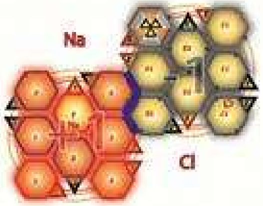

Inorganic compounds contain metal atoms.

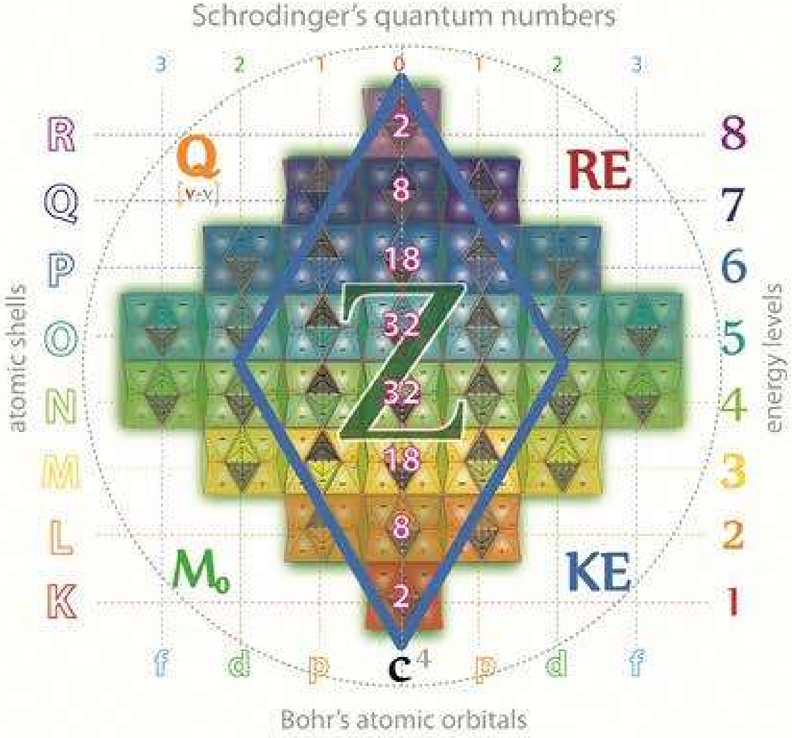
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
atomic shells
Q
[v-v]
RE
2
8
18
Z
32
32
18
8
2
M_0
KE
8
7
6
5
4
3
2
1
energy levels
f
d
p
c^4
p
d
f
Bohr's atomic orbitals

Organic compounds are produced by living things.

Organic compounds can't form salts.

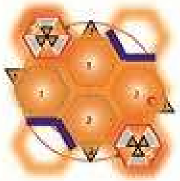

Organic compounds contain carbon-hydrogen bonds.

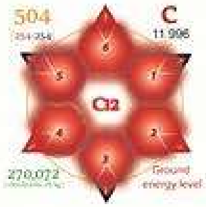
504
C
11.996
C12
270,072
Ground energy level

Tetryonic theory unifies and expands upon the currently disjointed physical and chemical theories through the application of 2D equilateral charged mass-energy geometries in 3D standing-wave mass-Matter topologies

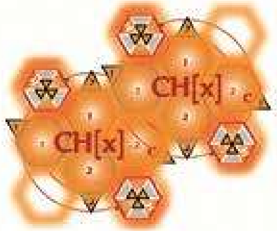

Inorganic **Typically the difference is defined as being whether or not a substance contains carbon or carbon-hydrogen bonds** **Organic**

A chemical compound is a collection of elements bonded together in a way that the resultant ions, atoms or molecules form a 3D material geometric structure. Tetryonic chemical geometries, along with its firm definition and distinction between EM mass & Matter provide a clear visual path for the differentiation between both branckes of modern chemistry - as well as the source of animation in living Matter

Functional Groups

A functional group is a reactive portion of a molecule.

carbonyl group

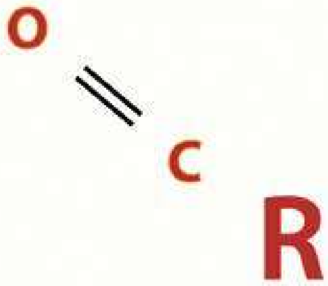

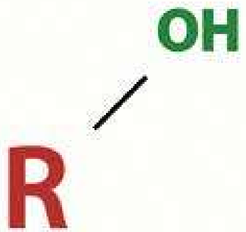

carboxyl group

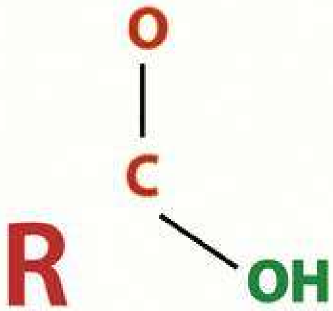

hydroxyl group

Organic molecules containing a hydroxyl group are known as alcohols.

The combinations of functional groups with hydrocarbons produce a vast number of compounds.

Carbon monoxide

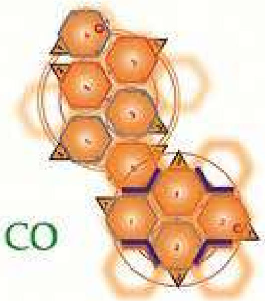

carbonyl group is composed of a carbon atom double-bonded to an oxygen atom

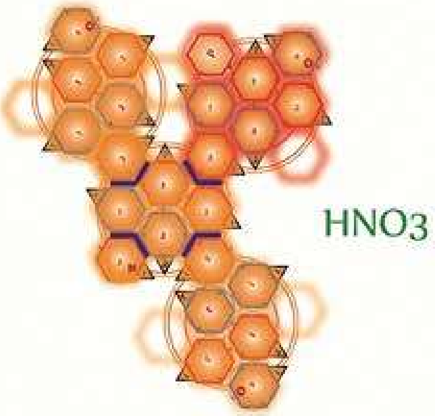

Nitric acid

Carbonic acid

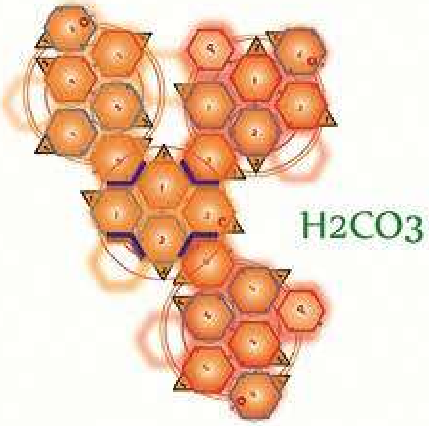

The carboxyl group is present in amino acids and carboxylic acids.

Copyright ABRAHAM [2008] - All rights reserved

Derivatives of Hydrocarbons

An almost unlimited number of carbon compounds can be formed by the addition of a functional group to a hydrocarbon

Esters

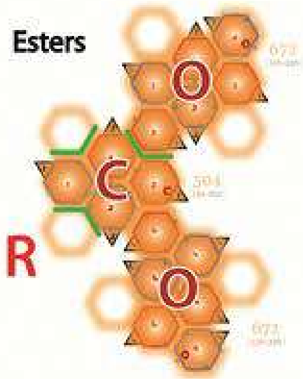

Most esters have pleasant odors.
Esters are responsible for the fragrances of
many flowers & the tastes of ripened fruits.

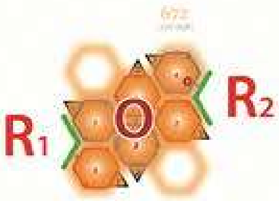

Ethers

The best known ether is diethyl ether.
It is a volatile, highly flammable liquid
that was used as an anesthetic in the past.

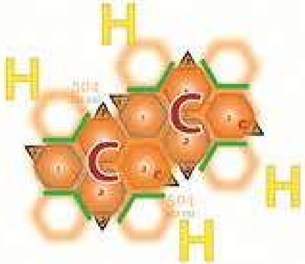

Alkyl Halides - haloalkanes

Common alkyl halides include medical anesthetics,
chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs)
and hydrofluorocarbons (HFCs).

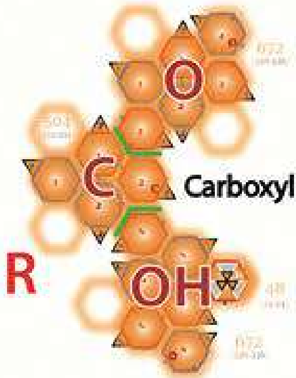

Carboxylic Acids

The simplest of the carboxylic acids is formic acid
and is a constituent of bee stings and the bites of
other insects including mosquitos.

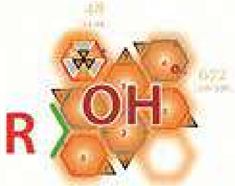

Alcohols

Alcohols are organic compounds containing a
hydroxyl group, [OH], substituted for a hydrogen atom.
Ethanol is the alcohol in alcoholic beverages and
it is also widely used as a solvent.

H
hydro

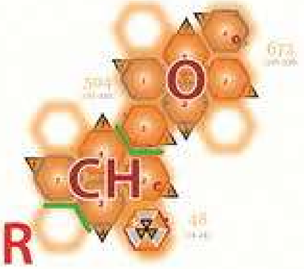

Aldehydes

An aldehyde is a compound containing a
carbonyl group with at least one hydrogen attached to it.
With a Hydrogen in place of the R group it forms Formaldehyde

C
carbons

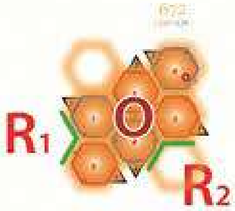

Ketones

Acetone is the simplest of the ketones.
Acetone is a commonly used solvent and is
the active ingredient in nail polish remover
and some paint thinners

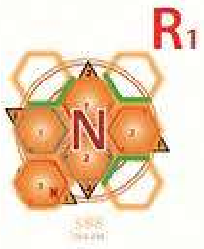

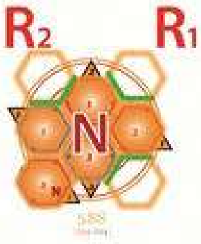

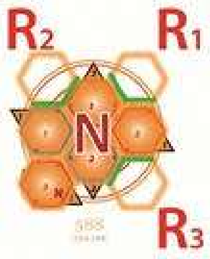

Amines

Amines are organic compounds that contain nitrogen,
they are basic compounds with strong odors,
often described as "fishy"

"R's" stand for carbon substituents or hydrogen atoms.

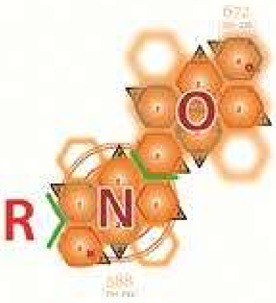

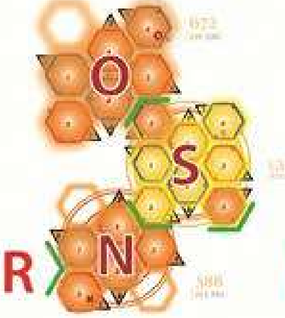

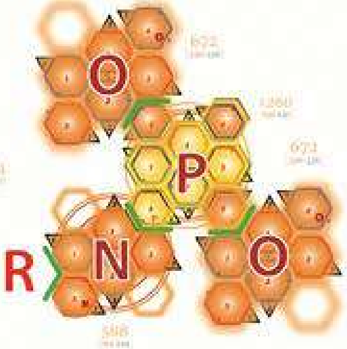

Amides

Amides are nitrogen-containing organic compounds
and are formed when amino acids react to form proteins.

Hydrocarbons

In organic chemistry, a methylene group is any part of a molecule that consists of two hydrogen atoms bound to a carbon atom, which is connected to the remainder of the molecule

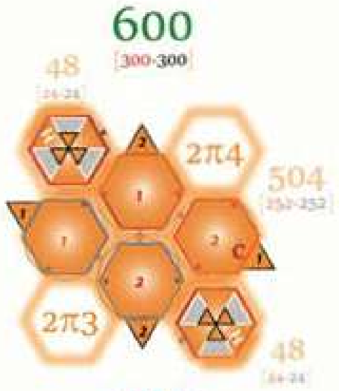
600
[300-300]
48
[24-24]
2π4
504
[252-252]
2π3
48
[24-24]

CH2
Methylene

The methylene group should be distinguished from the free methylene radical, also called carbene, whose molecule is a methylene group all by itself.

H2C
Carbene

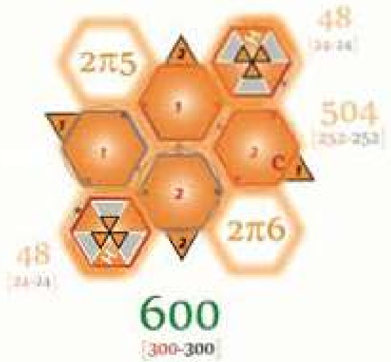
2π5
48
[24-24]
504
[252-252]
2π6
48
[24-24]
600
[300-300]

CH4
Methane

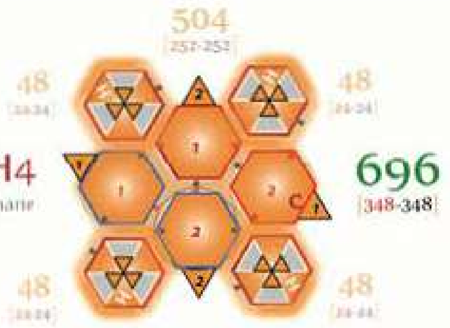
504
[252-252]
48
[24-24]
48
[24-24]
696
[348-348]
48
[24-24]
48
[24-24]

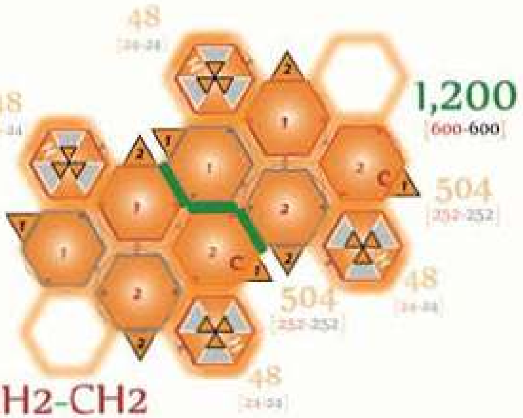
48
[24-24]
48
[24-24]
1,200
[600-600]
504
[252-252]
504
[252-252]
48
[24-24]
48
[24-24]

CH2-CH2
Ethylene

CH3-CH3
Ethane
C2H6

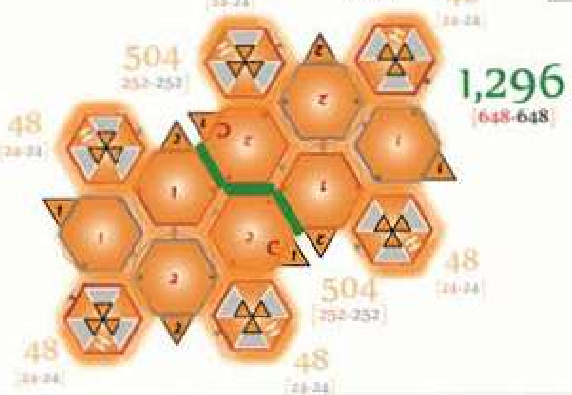
48
[24-24]
48
[24-24]
504
[252-252]
1,296
[648-648]
48
[24-24]
48
[24-24]
504
[252-252]
48
[24-24]
48
[24-24]

2

3

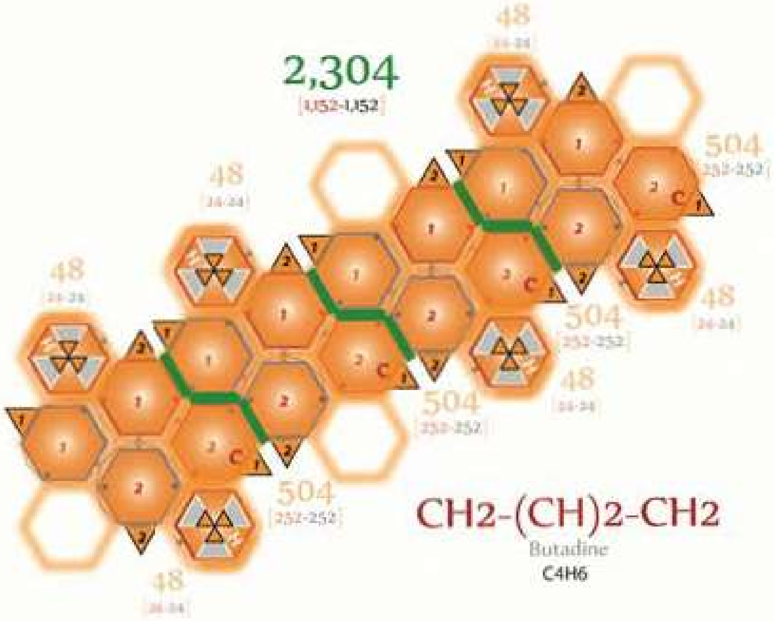
48
[24-24]
2,304
[1,152-1,152]
504
[252-252]
48
[24-24]
48
[24-24]
48
[24-24]
504
[252-252]
48
[24-24]
504
[252-252]
504
[252-252]
48
[24-24]

CH2-(CH)2-CH2
Butadine
C4H6

CH3-(CH2)-CH3
Propane
C3H8

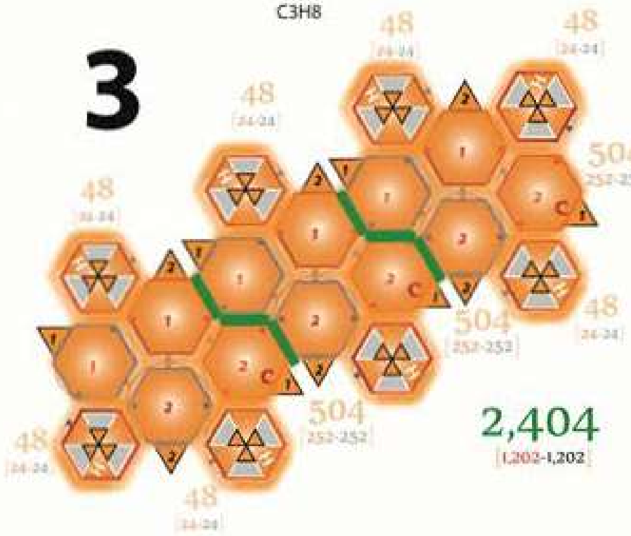
48
[24-24]
48
[24-24]
48
[24-24]
504
[252-252]
48
[24-24]
504
[252-252]
48
[24-24]
504
[252-252]
2,404
[1,202-1,202]
48
[24-24]
48
[24-24]

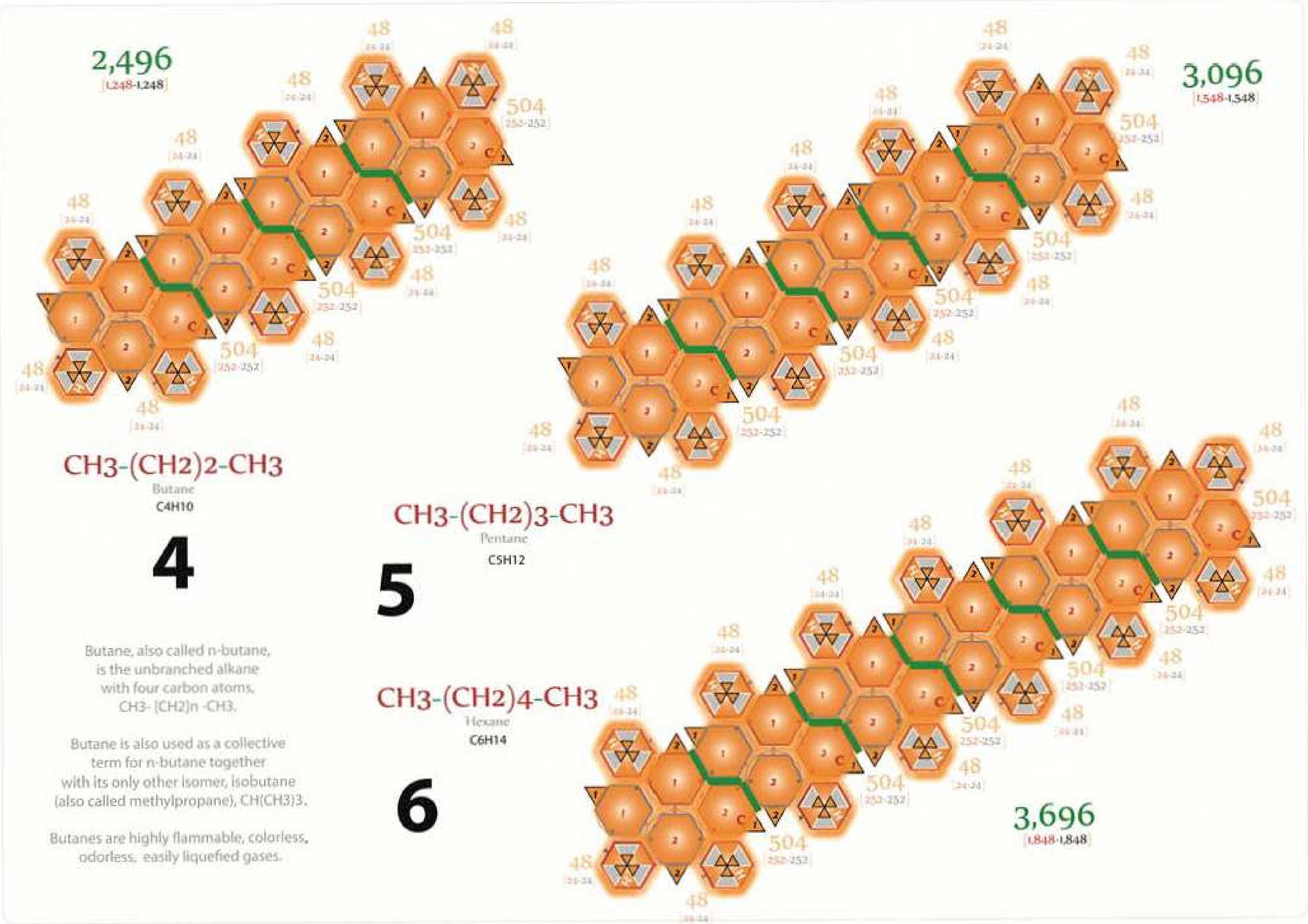
2,496
[1,248-1,248]
3,096
[1,548-1,548]
3,696
[1,848-1,848]
48
[24-24]
504
[252-252]
1
2
C

CH3-(CH2)2-CH3

Butane

C4H10

4

CH3-(CH2)3-CH3

Pentane

C5H12

5

CH3-(CH2)4-CH3

Hexane

C6H14

6

Butane, also called n-butane,
is the unbranched alkane
with four carbon atoms,
CH3- [CH2]n -CH3.

Butane is also used as a collective
term for n-butane together
with its only other isomer, isobutane
(also called methylpropane), CH(CH3)3.

Butanes are highly flammable, colorless,
odorless, easily liquefied gases.

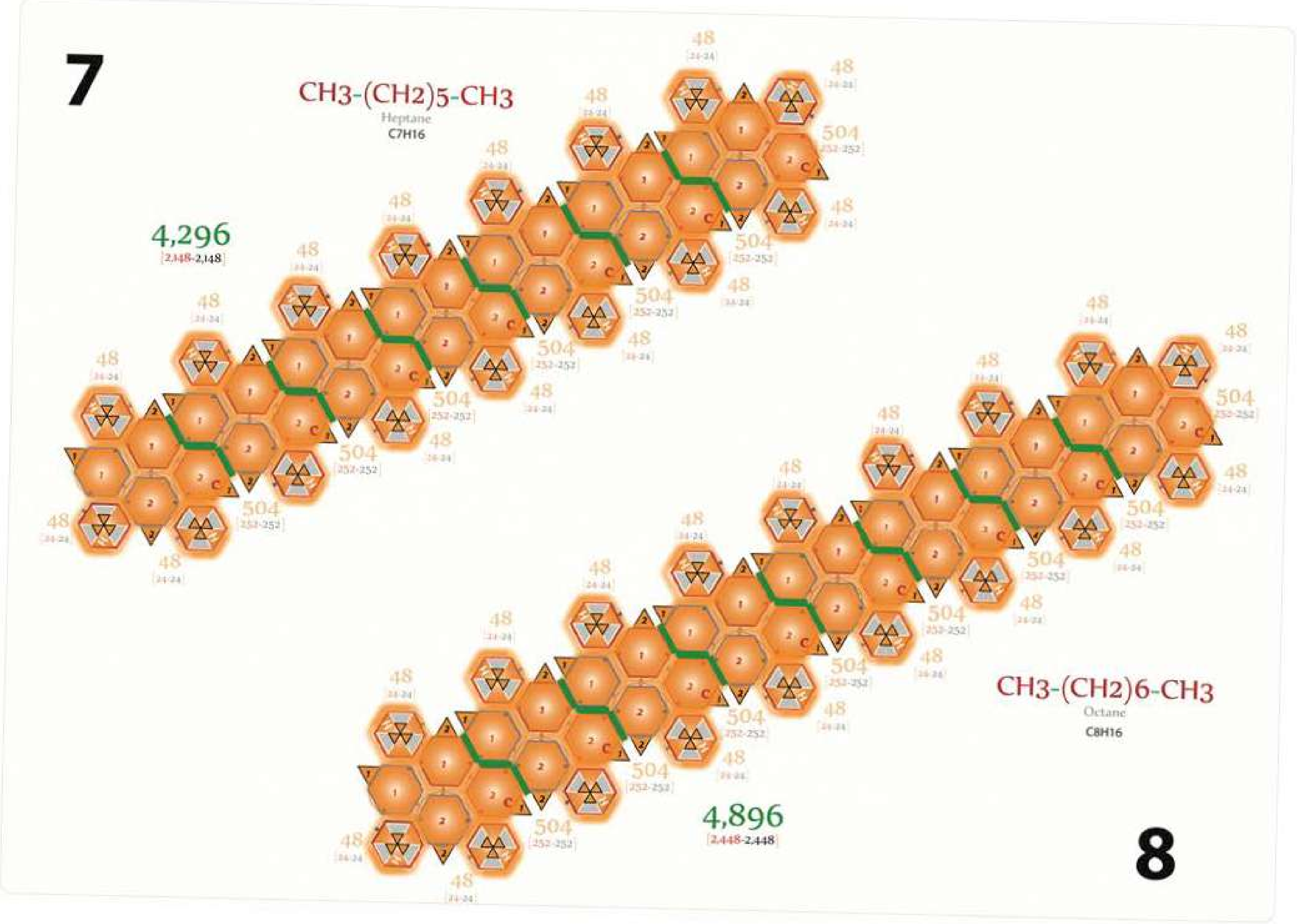
7
CH3-(CH2)5-CH3
Heptane
C7H16
4,296
[2,148-2,148]
48
[24-24]
504
[252-252]
8
CH3-(CH2)6-CH3
Octane
C8H16
4,896
[2,448-2,448]

Just like periodic elements,
compounds can form
allotropic compound structures

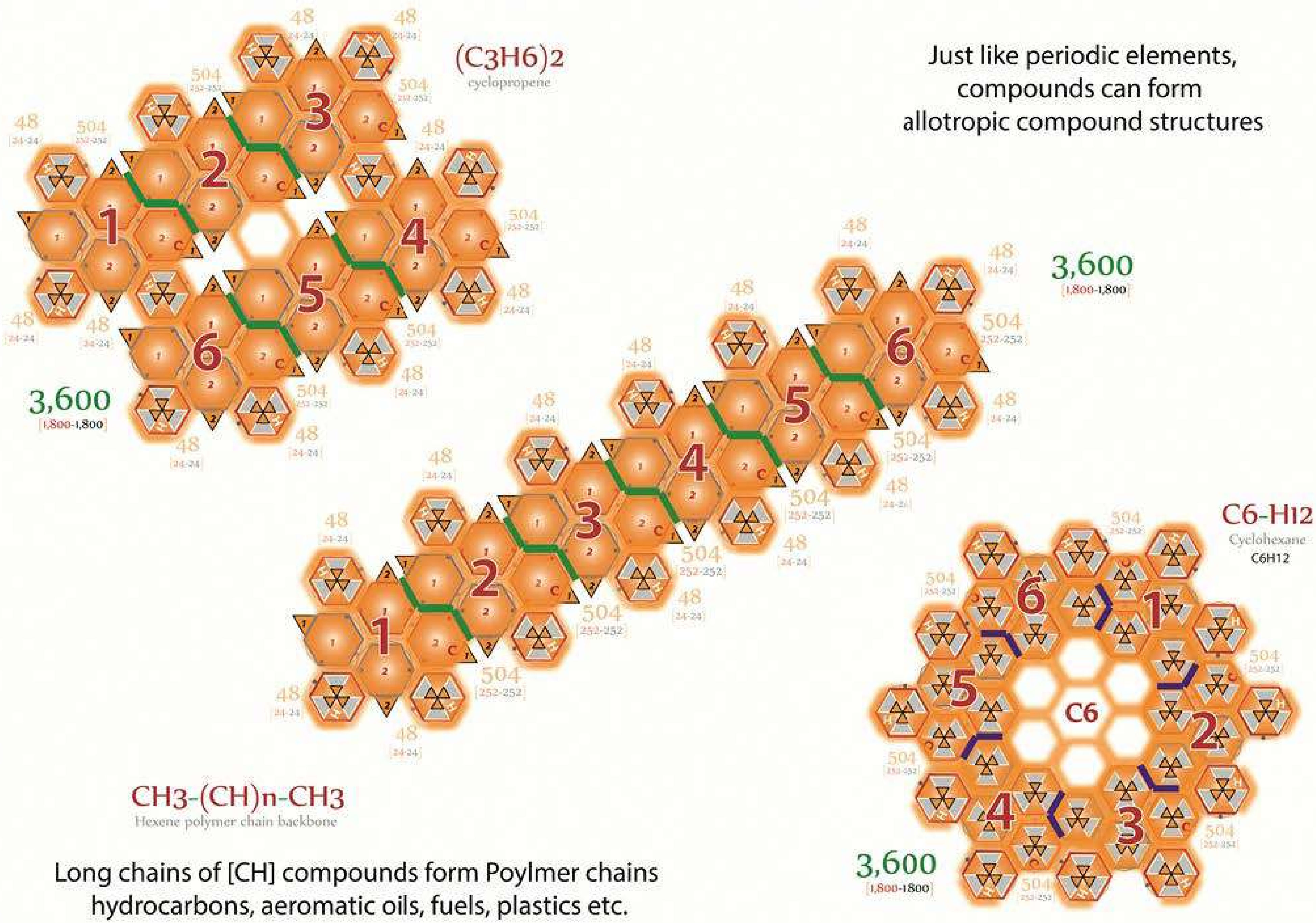
(C3H6)2
cyclopropene
3,600
[1,800-1,800]
48
[24-24]
504
[252-252]
1
2
3
4
5
6
C
CH3-(CH)n-CH3
Hexene polymer chain backbone
C6-H12
Cyclohexane
C6H12
C6

Long chains of [CH] compounds form Poylmer chains
hydrocarbons, aeromatic oils, fuels, plastics etc.

Rings of atoms are also common in organic structures. You may have heard the famous story of Auguste Kekulé first realizing that benzene has a ring structure when he dreamt of snakes biting their own tails.

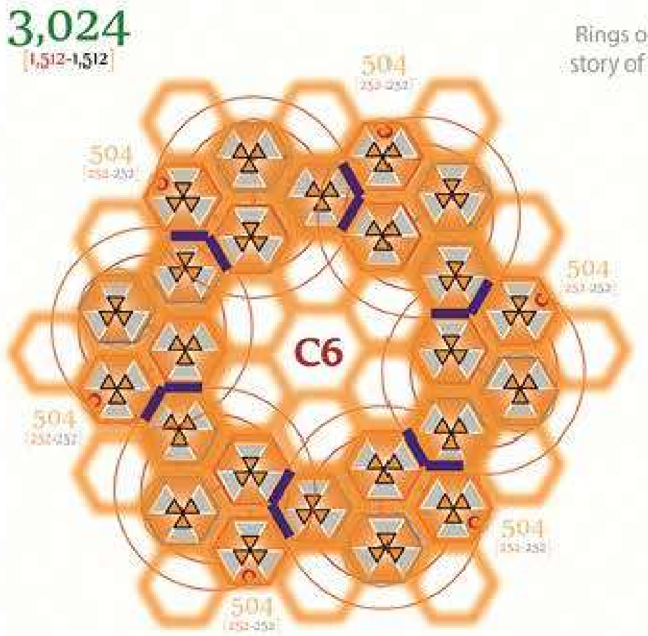
3,024
[1,512-1,512]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
C6

Carbon ring

Friedrich August Kekul

(7 September 1829 - 13 July 1896)

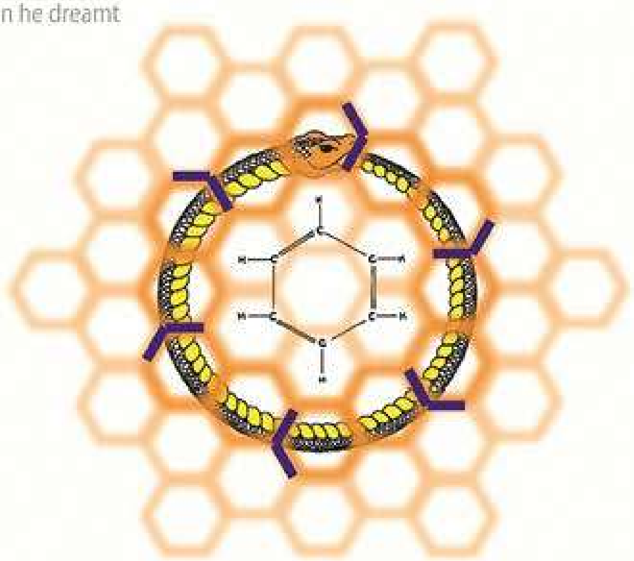

Organic Chemistry

In 1865, August Kekulé presented a paper at the Académie des Sciences in Paris suggesting a cyclic structure for benzene, the inspiration for which he ascribed to a dream. However, was Kekulé the first to suggest that benzene was cyclic. Some credit an Austrian schoolteacher, Josef Loschmidt with the first depiction of cyclic benzene structures.

In 1861, 4 years before Kekulé's dream, Loschmidt published a book in which he represented benzene as a set of rings. It is not certain whether Loschmidt or Kekulé—or even a Scot named Archibald Couper—got it right first

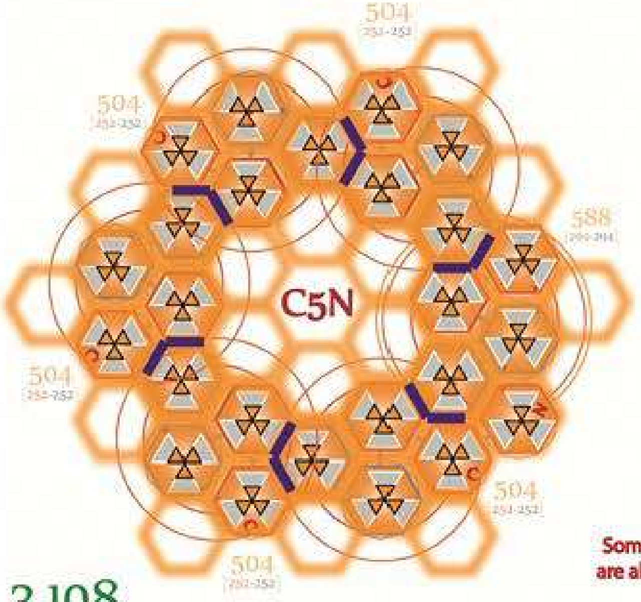
504
[252-252]
504
[252-252]
588
[252-294]
504
[252-252]
504
[252-252]
504
[252-252]
C5N
3,108
[1,554-1,554]

Carbon-Nitrogen ring

Some non-benzene-based compounds called heteroarenes, which follow Hückel's rule, are also aromatic compounds. In these compounds, at least one carbon atom is replaced by one of the heteroatoms oxygen, nitrogen, or sulfur.

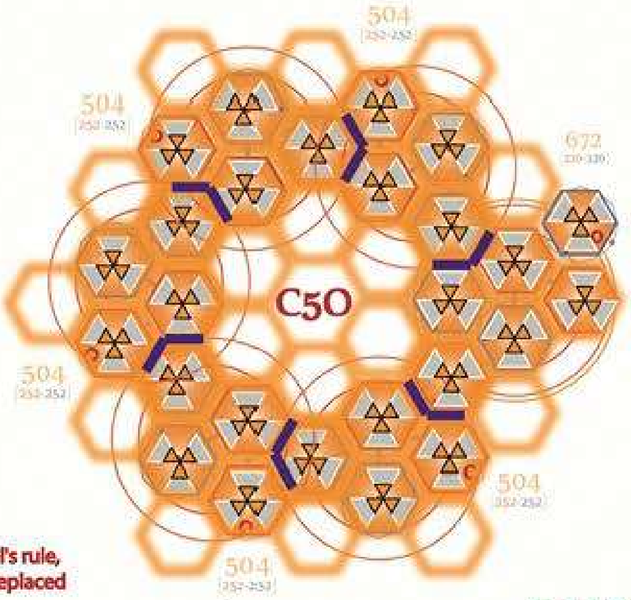
504
[252-252]
504
[252-252]
672
[336-336]
504
[252-252]
504
[252-252]
504
[252-252]
C5O
3,192
[1,596-1,596]

Carbon-Oxygen ring

Benzene molecules

Benzene, or benzol, is an organic chemical compound and a known carcinogen with the molecular formula C6H6.

It is sometimes abbreviated Ph–H.
Benzene is a colorless and highly flammable liquid with a sweet smell and a relatively high melting point.

Because it is a known carcinogen, its use as an additive in gasoline is now limited, but it is an important industrial solvent and precursor in the production of drugs, plastics, synthetic rubber, and dyes. Benzene is a natural constituent of crude oil, and may be synthesized from other compounds present in petroleum.

Benzene is an aromatic hydrocarbon and the second [n]-annulene ([6]-annulene), a cyclic hydrocarbon with a continuous pi bond.

Cyclic hydrocarbon compounds are often referred to as Aromatic compounds due to their sweet smell

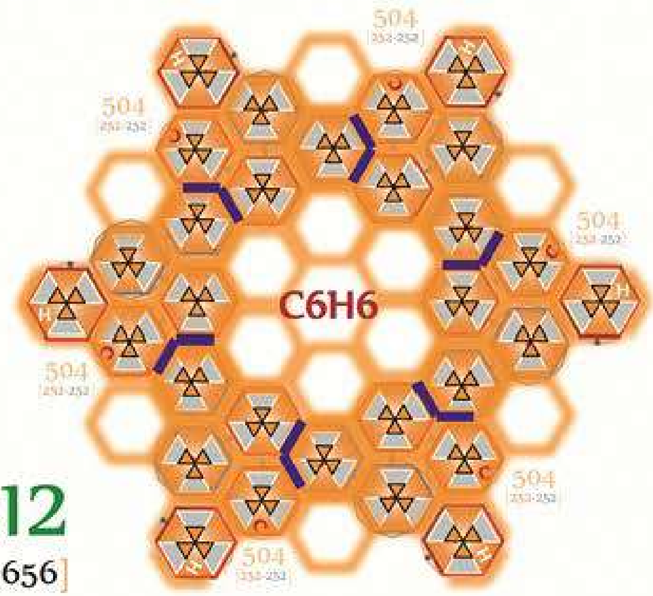
504
252-252
504
252-252
504
252-252
504
252-252
504
252-252
504
252-252
C6H6
H
C

3,312
[1,656-1,656]

(CH)6
Benzene ring
C6H6

A aromatic hydrocarbon is formed when CH compounds form a cyclic molecule

Many important additional chemical compounds are derived from benzene by replacing one or more of its hydrogen atoms with another functional group.

Examples of simple benzene derivatives are phenol, toluene, and aniline, abbreviated PhOH, PhMe, and PhNH2, respectively

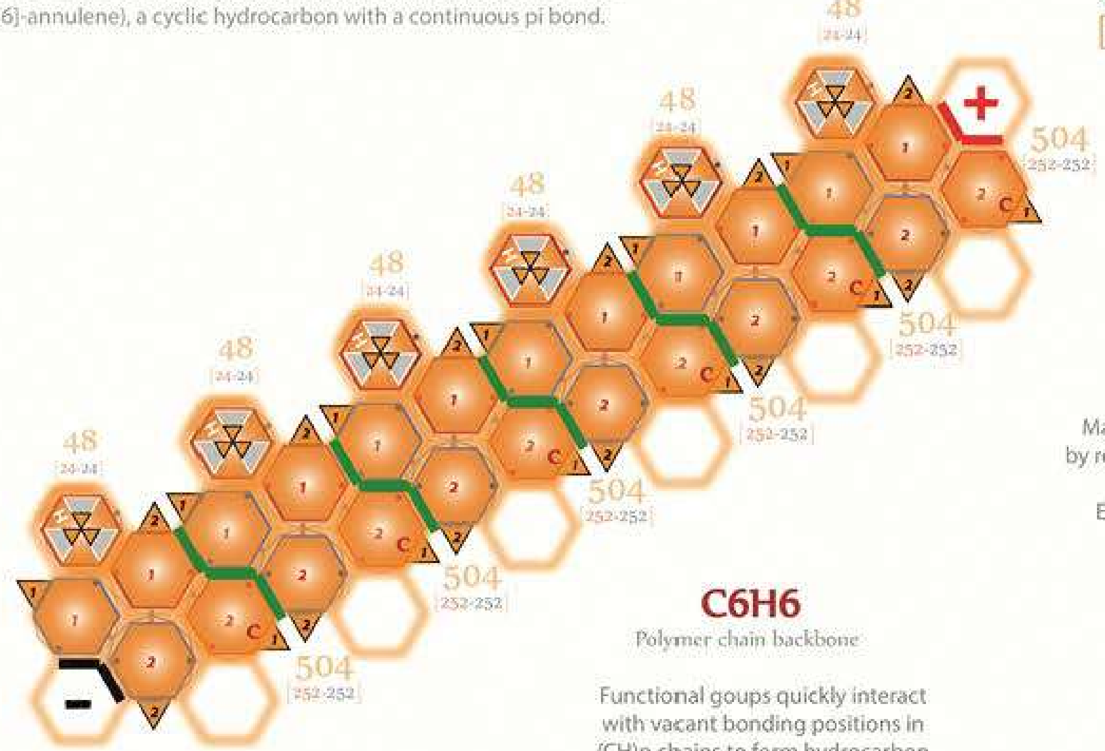
48
24-24
48
24-24
48
24-24
48
24-24
48
24-24
48
24-24
504
252-252
504
252-252
504
252-252
504
252-252
504
252-252
504
252-252
1
2
C
H
+
-

C6H6
Polymer chain backbone

Functional goups quickly interact with vacant bonding positions in (CH)n chains to form hydrocarbon compounds

Linear (CH)n chains are rarely found in nature as the Positive and Negative tail ends of Hydrocarbon chain interact and bond to form cyclic compounds.

Carbon-Nitrogen Chains

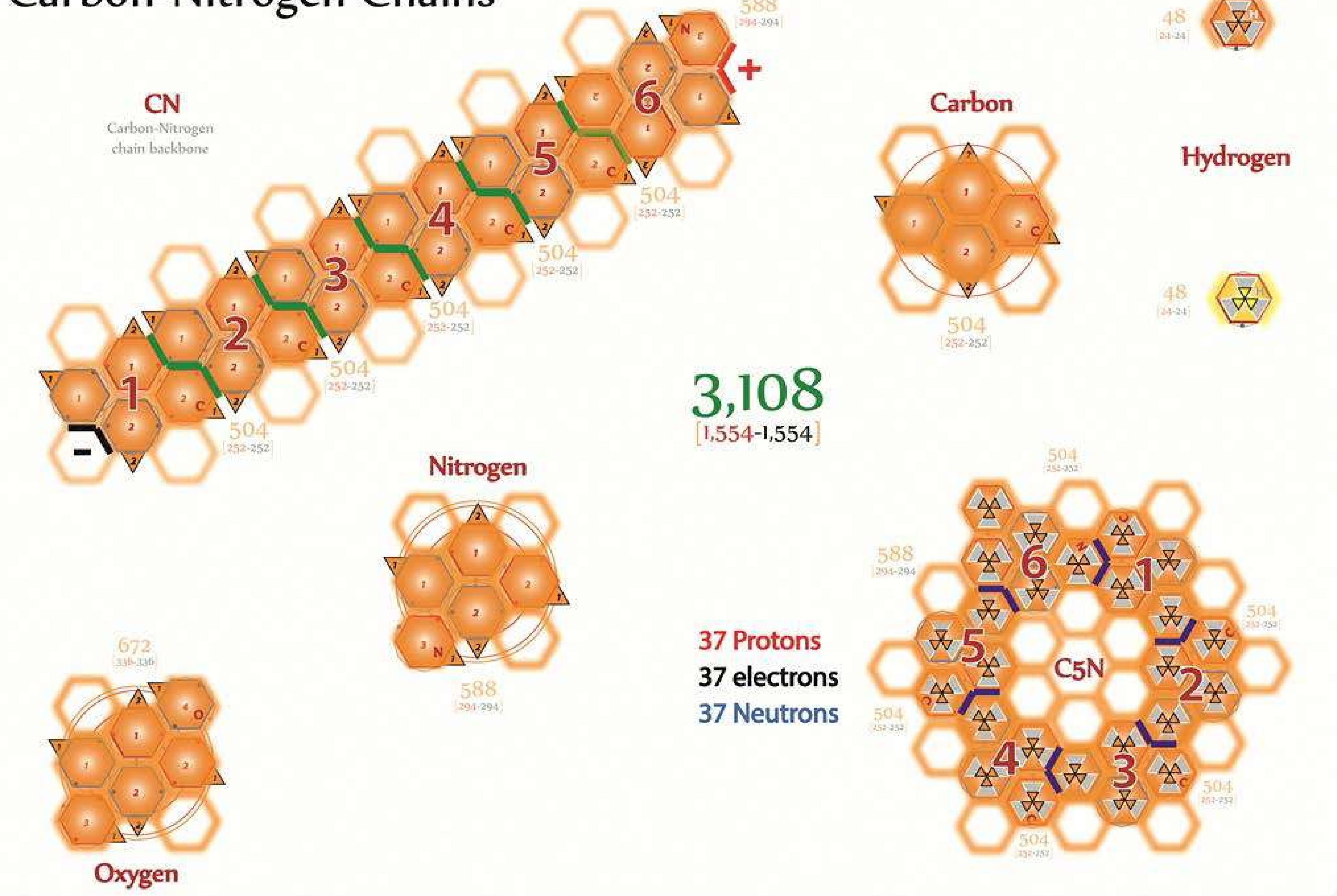
CN
Carbon-Nitrogen
chain backbone
1
2
3
4
5
6
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
588
[294-294]
+
-
Carbon
504
[252-252]
Hydrogen
48
[24-24]
48
[24-24]
Nitrogen
588
[294-294]
Oxygen
672
[336-336]
C5N
588
[294-294]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]

3,108
[1,554-1,554]

37 Protons
37 electrons
37 Neutrons

Carbon-Oxygen Chains

48
[24-24]

Hydrogen

48
[24-24]

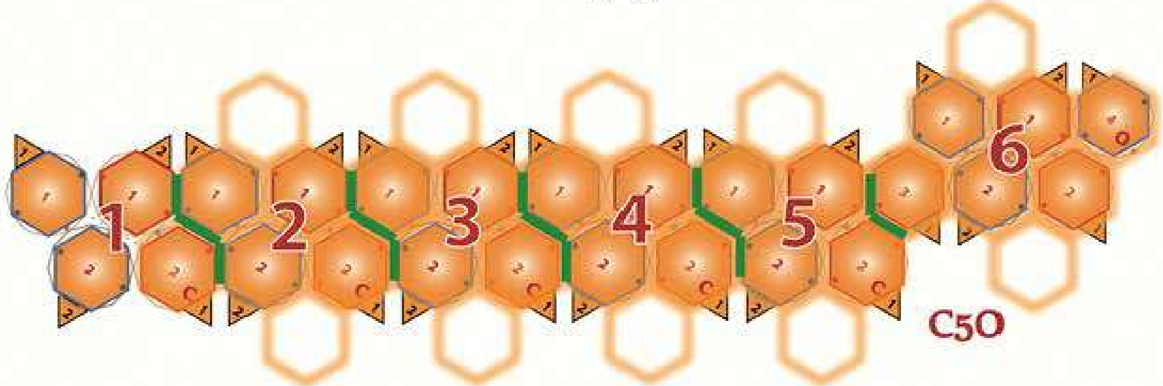

C5O

672
[336-336]

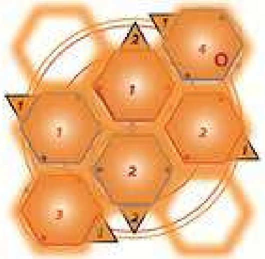

Oxygen

Heterocyclic Rings

Tetrahydropyran is the organic compound consisting of a saturated six-membered ring containing five carbon atoms and one oxygen atom.

Carbon

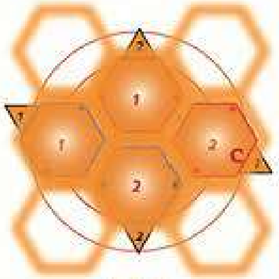

504
[252-252]

3,192
[1,596-1,596]

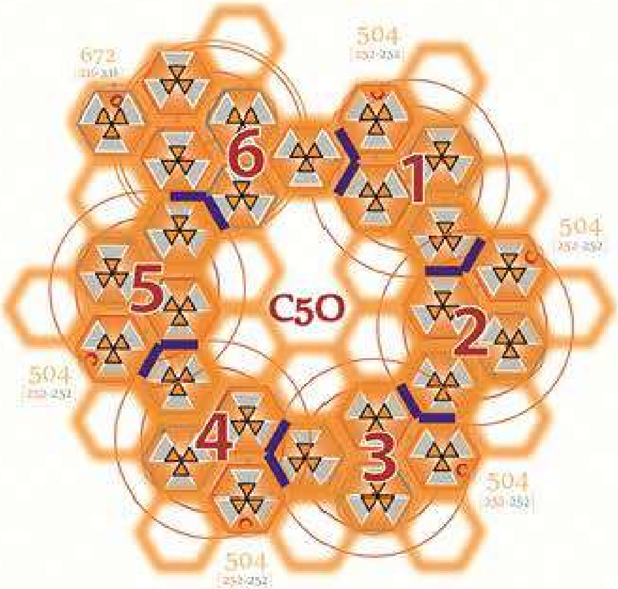
672
504
504
504
504
504
1
2
3
4
5
6
C5O

38 Protons
38 electrons
38 Neutrons

Nitrogen

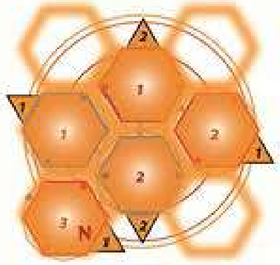

588
[294-294]

$C_6H_2(NO_2)_3CH_3$

Trinitrotoluene

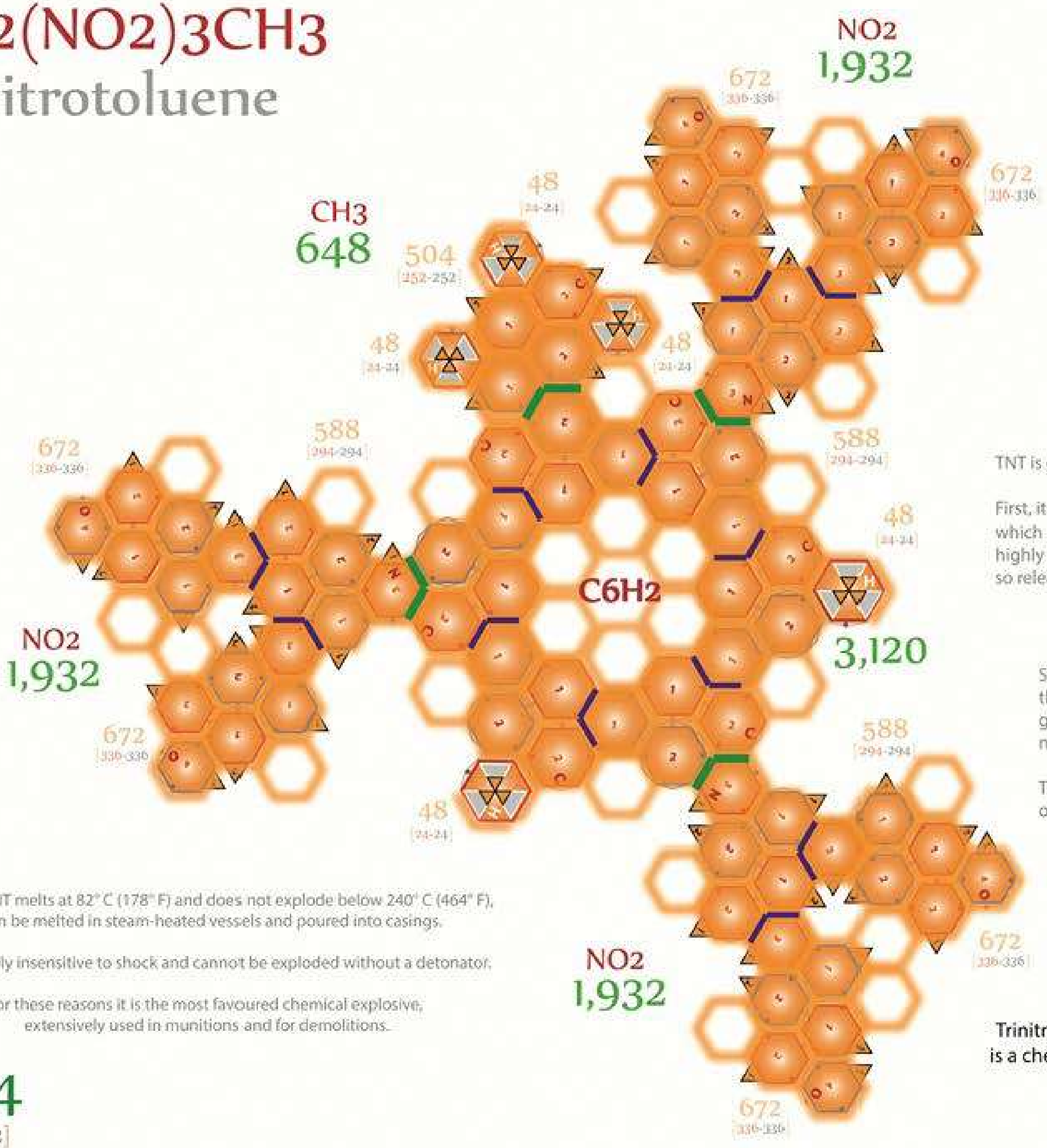

NO2
1,932
672
[336-336]
672
[336-336]
CH3
648
48
[24-24]
504
[252-252]
48
[24-24]
48
[24-24]
588
[294-294]
672
[336-336]
588
[294-294]
48
[24-24]
C6H2
3,120
NO2
1,932
672
[336-336]
588
[294-294]
48
[24-24]
NO2
1,932
672
[336-336]
672
[336-336]

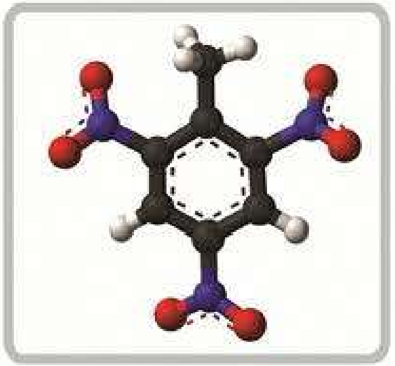

2,4,6-trinitrotoluene is better known by its initials, TNT.
It is an important explosive, since it can very quickly
change from a solid into hot expanding gases.

TNT is explosive for two reasons;

First, it contains the elements carbon, oxygen and nitrogen, which means that when the material burns it produces highly stable substances (CO, CO2 and N2) with strong bonds, so releasing a great deal of energy.

Secondly, TNT is chemically unstable - the nitro groups are so closely packed that they experience a great deal of strain and hindrance to movement from their neighbouring groups.

Thus it doesn't take much of an initiating force to break some of the strained bonds, and the molecule then flies apart.

Trinitrotoluene, or more specifically, 2,4,6-trinitrotoluene, is a chemical compound with the formula $C_6H_2(NO_2)_3CH_3$

Because TNT melts at 82° C (178° F) and does not explode below 240° C (464° F), it can be melted in steam-heated vessels and poured into casings.

It is relatively insensitive to shock and cannot be exploded without a detonator.

For these reasons it is the most favoured chemical explosive, extensively used in munitions and for demolitions.

9,564
[4,782-4,782]

(C6)4-H10

Pyrene

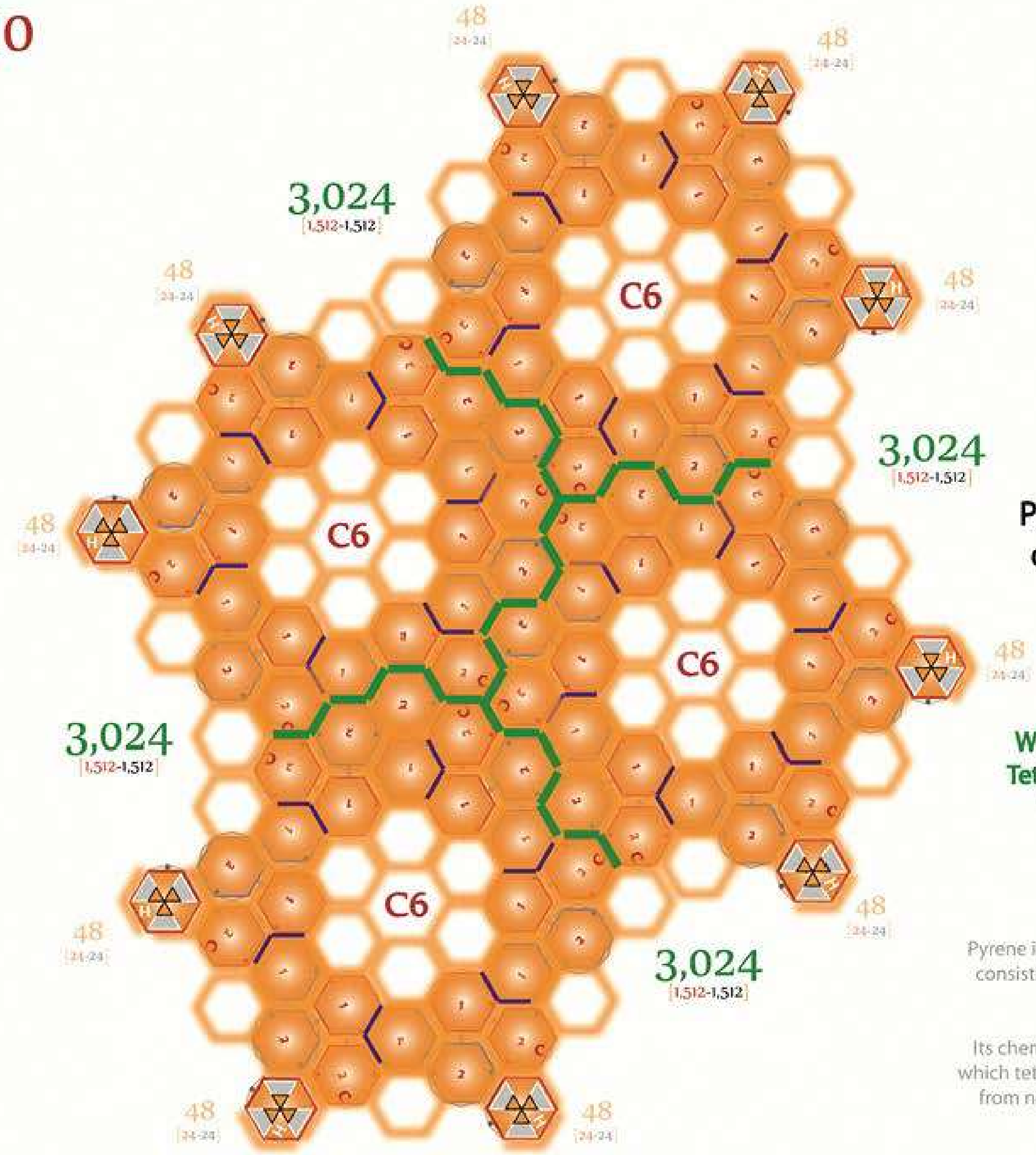
48
[24-24]
48
[24-24]
3,024
[1,512-1,512]
48
[24-24]
C6
48
[24-24]
3,024
[1,512-1,512]
48
[24-24]
C6
C6
48
[24-24]
3,024
[1,512-1,512]
C6
48
[24-24]
48
[24-24]
3,024
[1,512-1,512]
48
[24-24]
48
[24-24]
H

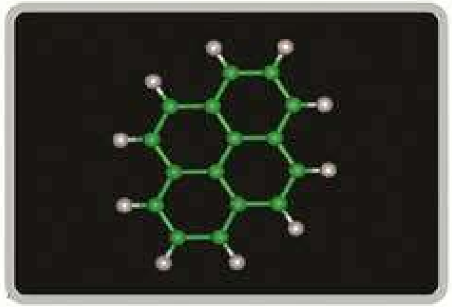

Pyrene highlights the failings of Lewis diagram structures

Which can be rectified using charged Tetryonic geometric Matter topologies for all elements and compound interactions & modelling

Pyrene is a polycyclic aromatic hydrocarbon (PAH) consisting of four fused benzene rings, resulting in a flat aromatic system.

Its chemcial formula as is often stated as C16H10, which tetryonic geometry shows is an error resulting from not using equilateral quantum geometries in chemical topology modelling.

12,576
[6,288-6,288]

Four cyclic carbon rings cannot be formed with only 16 C-atoms

Copyright ABRAHAM [2008] - All rights reserved

Ammonium Nitrate

3,384

NH4NO3,

Ammonium nitrate is composed of two polyatomic ions:
1) Ammonium Ion (NH4+)
2) Nitrate Ion (NO3-)

The bond between these ions is an ionic bond meaning the ammonium ion transfers an electron to the nitrate ion.

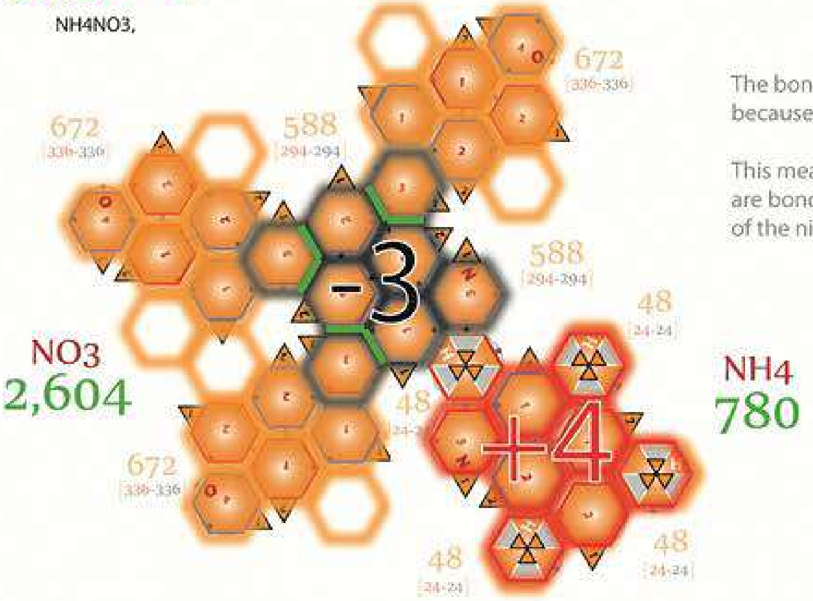
672
[336-336]
588
[294-294]
672
[336-336]
-3
588
[294-294]
48
[24-24]
NO3
2,604
NH4
780
+4
672
[336-336]
48
[24-24]
48
[24-24]
48
[24-24]

The bonds in a polyatomic atomic ion are covalent because they take place between gases.

This means that the hydrogens of the ammonium ion are bonded to the nitrogen covalently and that the oxygens of the nitrate are bonded to the nitrogen covalently

Ammonium nitrate decomposes into the gases nitrous oxide and water vapor when heated (non-explosive reaction); however, ammonium nitrate can be induced to decompose explosively by detonation.

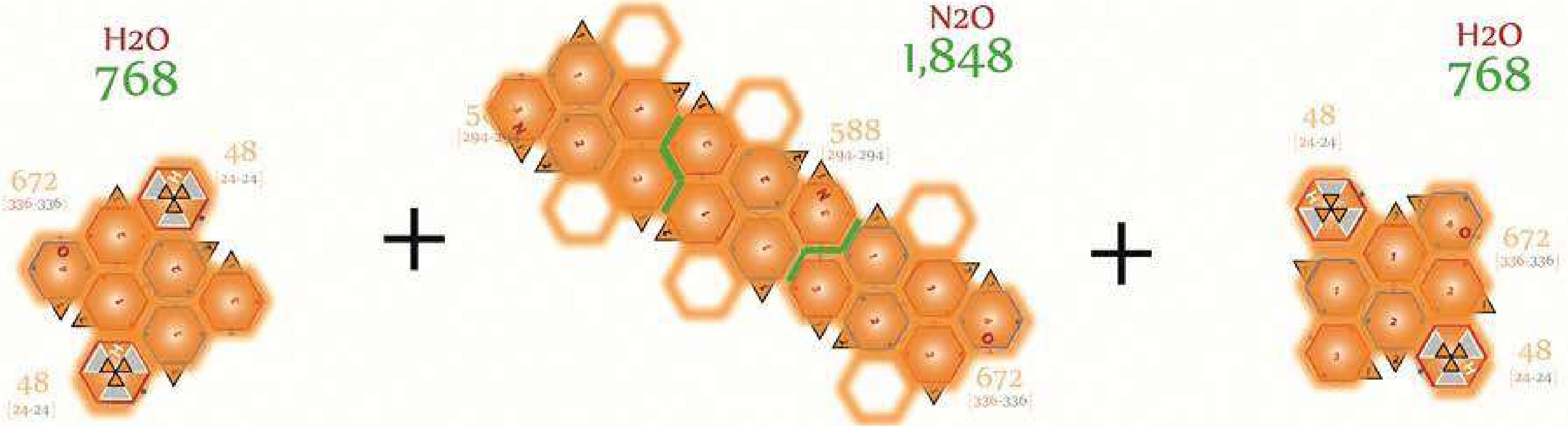
H2O
768
672
[336-336]
48
[24-24]
48
[24-24]
+
N2O
1,848
588
[294-294]
588
[294-294]
672
[336-336]
+
H2O
768
48
[24-24]
672
[336-336]
48
[24-24]

Copyright ABRAHAM [2008] - All rights reserved

Carbohydrates

A carbohydrate is an organic compound that consists only of carbon, hydrogen, and oxygen (with a hydrogen:oxygen ratio of 2:1) in other words, with the empirical formula $C_m(H_2O)_n$

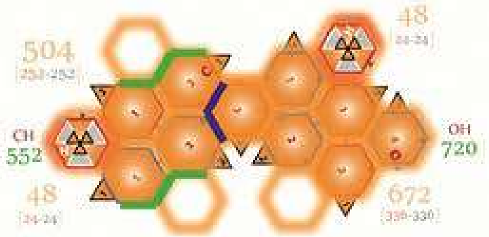
504
[252-252]
48
[24-24]
CH
552
OH
720
48
[24-24]
672
[336-336]

Monosaccharides

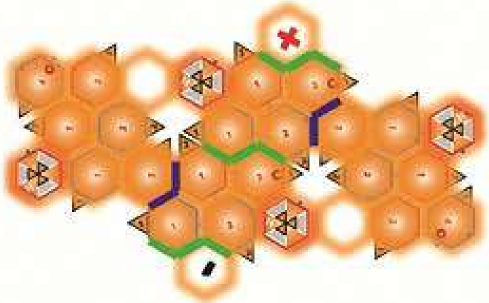

Disaccharides

Ribose

Gluclose

Lactose

Oligosaccharides

$$C_m(H_2O)_n$$

Polysaccharides

Fructose

Sucrose

Deoxyribose

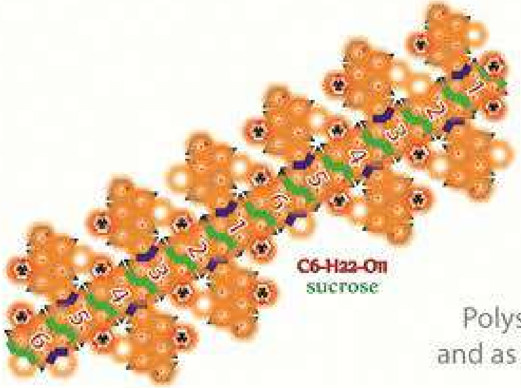
C6-H22-O11
sucrose
1
2
3
4
5
6

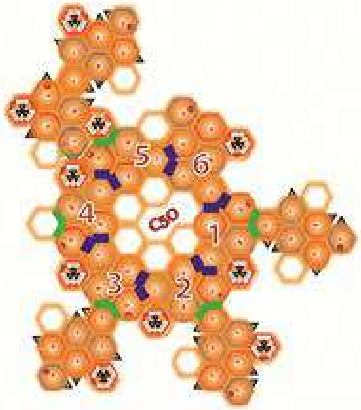

Carbohydrates perform numerous roles in living organisms. Polysaccharides serve for the storage of energy (e.g., starch and glycogen), and as structural components (e.g., cellulose in plants and chitin in arthropods)

Copyright ABRAHAM [2008] - All rights reserved

Monosaccharides are the simplest carbohydrates in that they cannot be hydrolyzed to smaller carbohydrates

Monosaccharides

The Carbon nuclei in Monosaccharides join together to form chains of biologically important carbohydrates

1,272

[636-636]

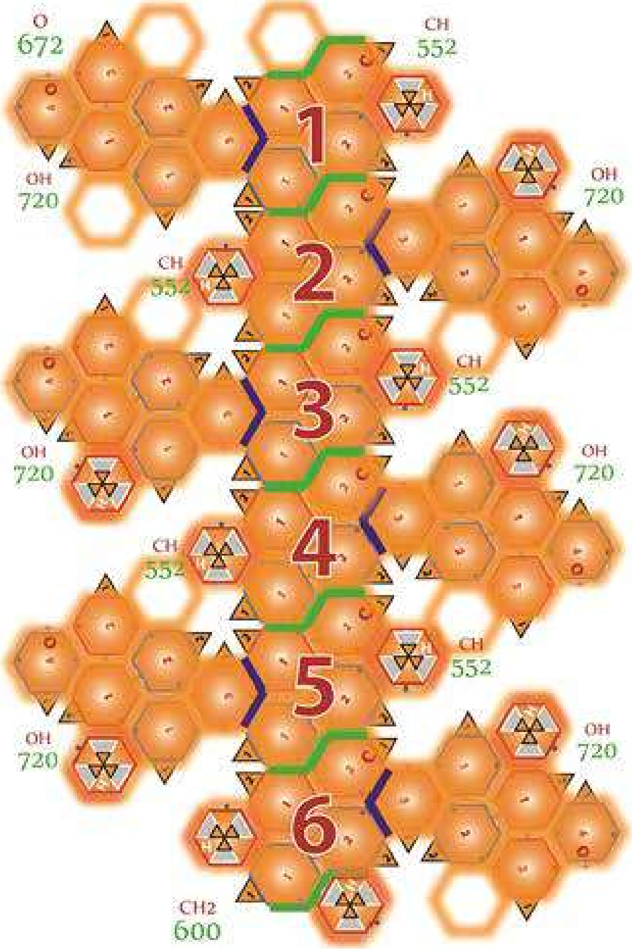
O
672
CH
552
OH
720
1
2
3
4
5
6
CH
552
CH
552
OH
720
OH
720
CH
552
CH
552
OH
720
OH
720
CH2
600

$C(H_2O)6$

D-glucose

$C_6\text{-}H_{12}\text{-}O_6$

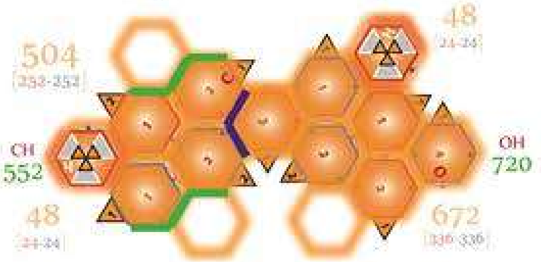
504
[252-252]
48
[24-24]
CH
552
OH
720
48
[24-24]
672
[336-336]

$C(H_2O)$

The general chemical formula of an unmodified monosaccharide is $(C \cdot H_2O)$ n, literally a "carbon hydrate."

CH 552

CH-OH

OH 720

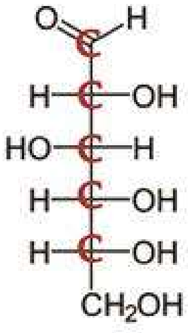

D-Glucose

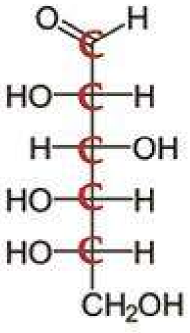

L-Glucose

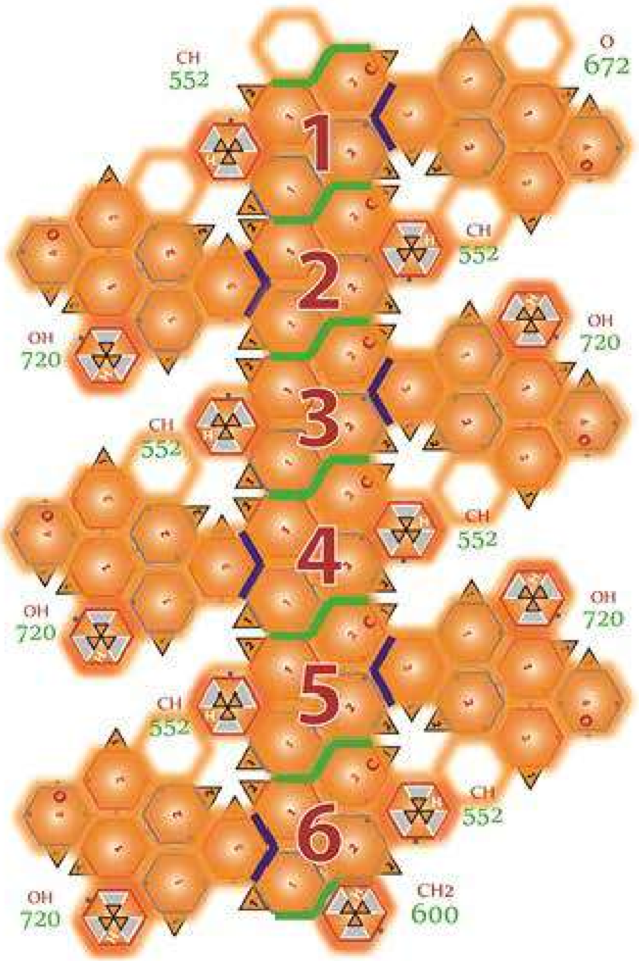
CH
552
O
672
1
2
3
4
5
6
CH
552
OH
720
OH
720
CH
552
CH
552
OH
720
OH
720
CH
552
CH
552
OH
720
CH
552
CH2
600

$C(H_2O)6$

L-glucose

$C_6\text{-}H_{12}\text{-}O_6$

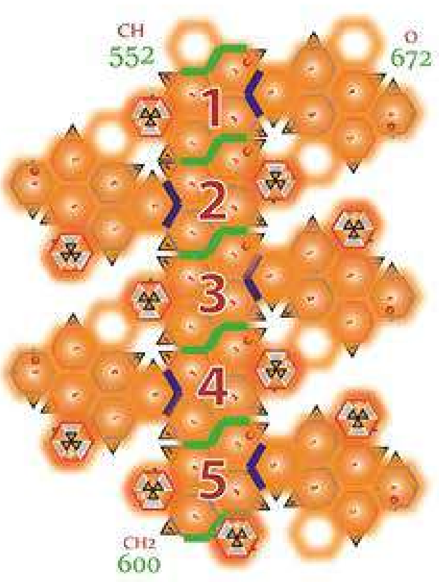
CH
552
O
672
1
2
3
4
5
CH2
600

C(H2O)5
ribose
C5-H10-O5

C(H2O)n

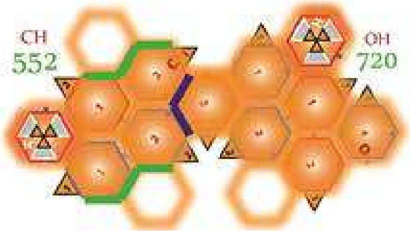
CH
552
OH
720

C(H2O)6
fructose
C6-H12-O6

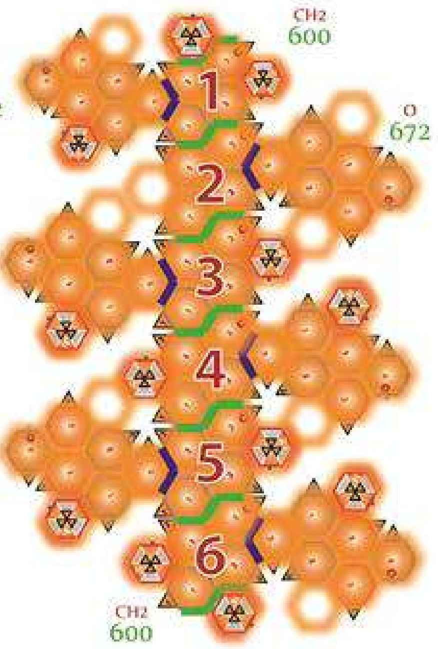
CH2
600
O
672
1
2
3
4
5
6
CH2
600

Major Monosaccharides

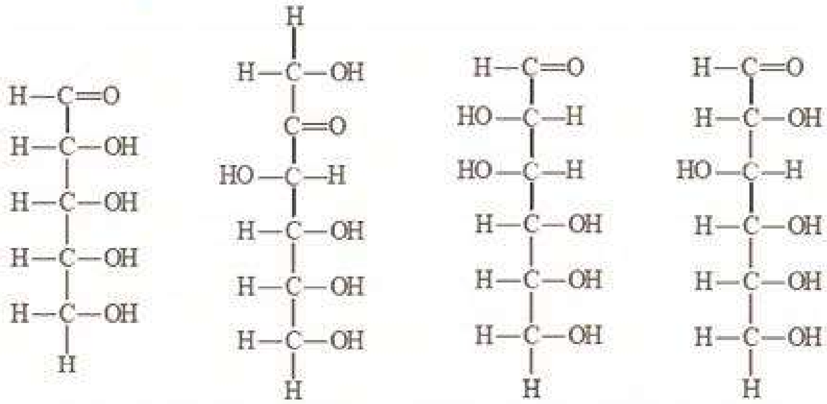

ribose fructose mannose glucose

D D & L sugars are chiral, mirror images of one another and have the same name. **L**

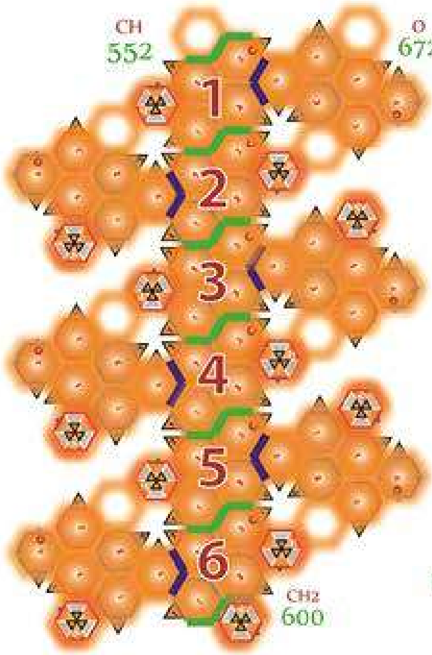
CH
552
O
672
1
2
3
4
5
6
CH2
600

C(H2O)6
mannose
C6-H12-O6

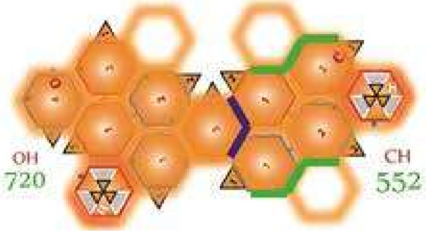
OH
720
CH
552

C(H2O)6
glucose
C6-H12-O6

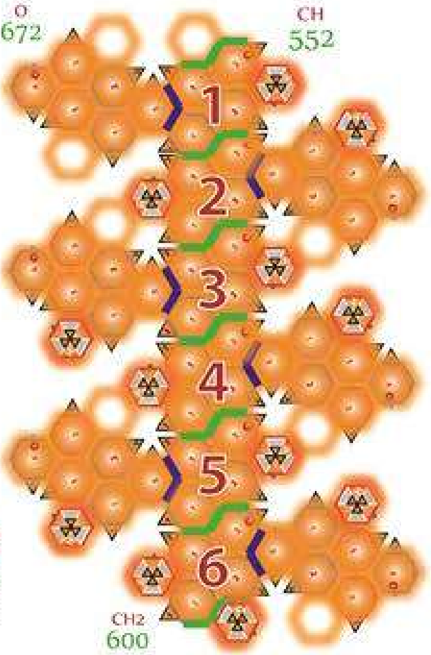
O
672
CH
552
1
2
3
4
5
6
CH2
600

A monosaccharide often switches from the acyclic (open-chain) form to a cyclic form, through a nucleophilic addition reaction between the carbonyl group and one of the hydroxyls of the same molecule.

cyclic monosaccharides

For many monosaccharides (including glucose), the cyclic forms predominate, and therefore the same name commonly is used for forms of isomers.

C(H2O)6
glucose
C6-H12-O6

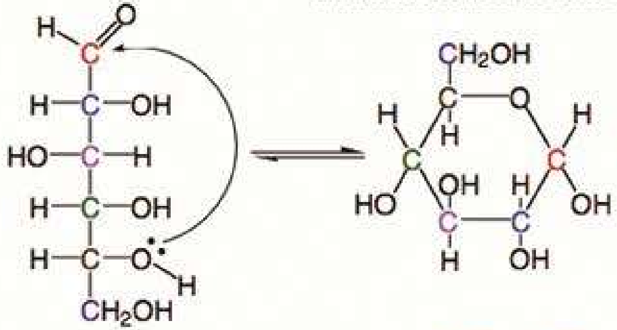
H
O
C
H—C—OH
HO—C—H
H—C—OH
H—C—O
H
CH₂OH
CH₂OH
C
O
H
H
C
C
HO
OH
H
OH
C
C
H
OH

C(H2O)6
glucopyranose
C6-H12-O6

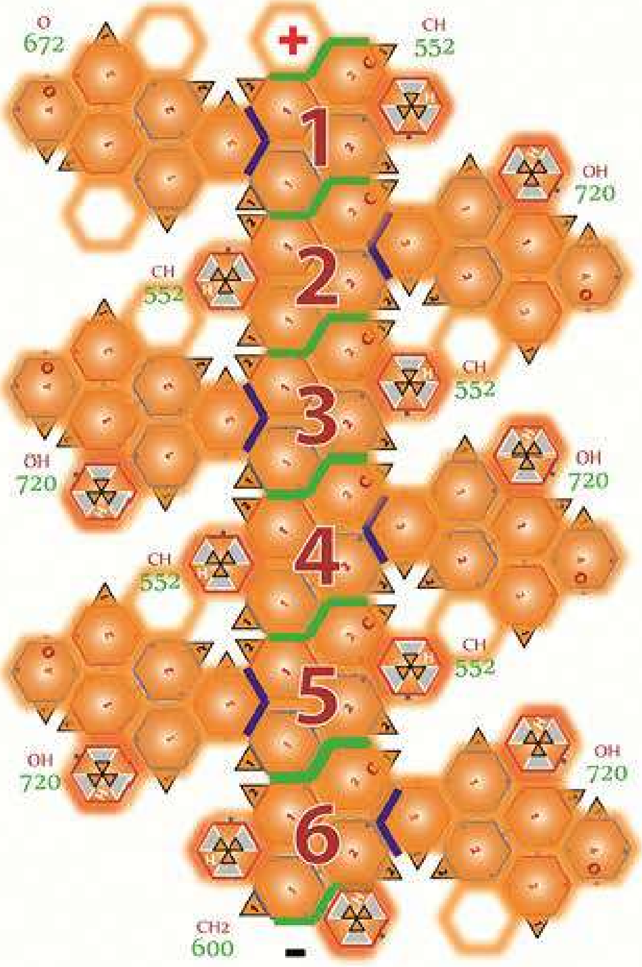
O
672
+
CH
552
1
OH
720
CH
552
2
CH
552
3
OH
720
OH
720
CH
552
4
CH
552
5
OH
720
OH
720
6
CH2
600
-

C6-H12-O6
7,632
[3,816-3,816]

The reaction creates a ring of carbon atoms closed by one bridging oxygen atom, which in turn can be easily broken to restore the linear monosaccharide chains

cyclic Oxygen-Carbon ring

Allotropic cyclic [Carbon-Oxygen] forms of polycarbohydrates can also be formed via other element bonding processes

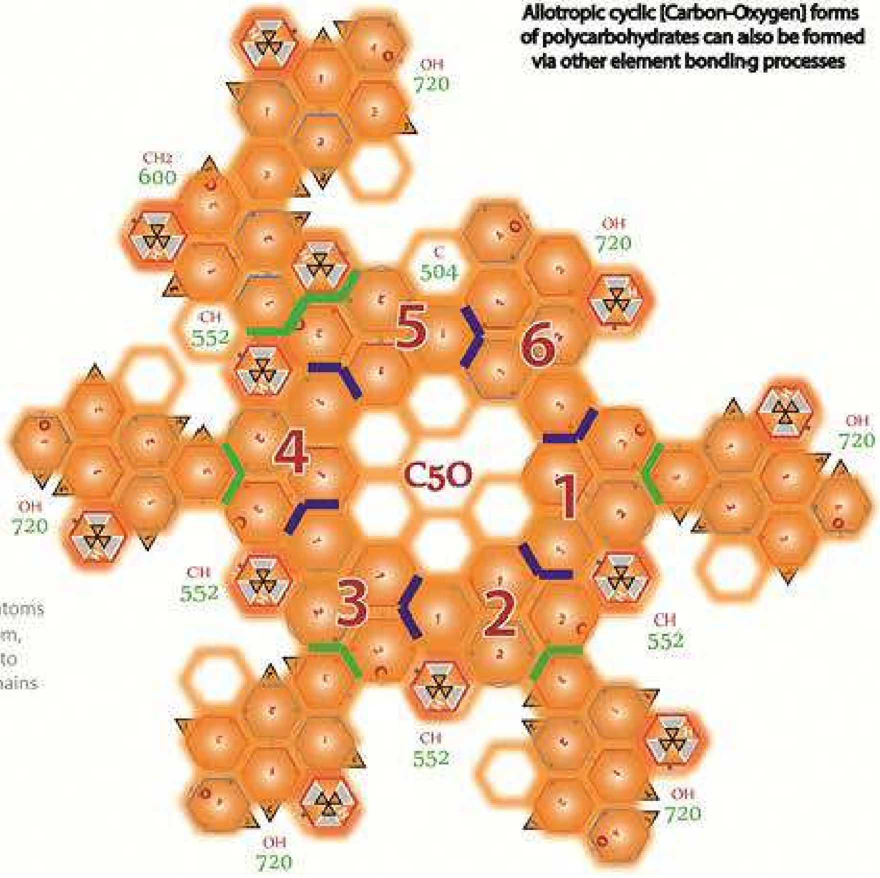
OH
720
CH2
600
C
504
OH
720
CH
552
5
6
4
C5O
1
OH
720
OH
720
CH
552
3
2
CH
552
CH
552
OH
720
OH
720

linear saccharide chain

The Carbon nuclei in Monosaccharides chains can join together to form cyclic [Carbon-Oxygen] rings or polycarbohydrates through condensation

acyclic monosaccharides

Condensed Disaccharides

Condensed Disaccharides are formed when two monosaccharides are joined together and a molecule of water is removed

For example, milk sugar (lactose) is made from glucose and galactose
whereas cane sugar (sucrose) is made from glucose and fructose

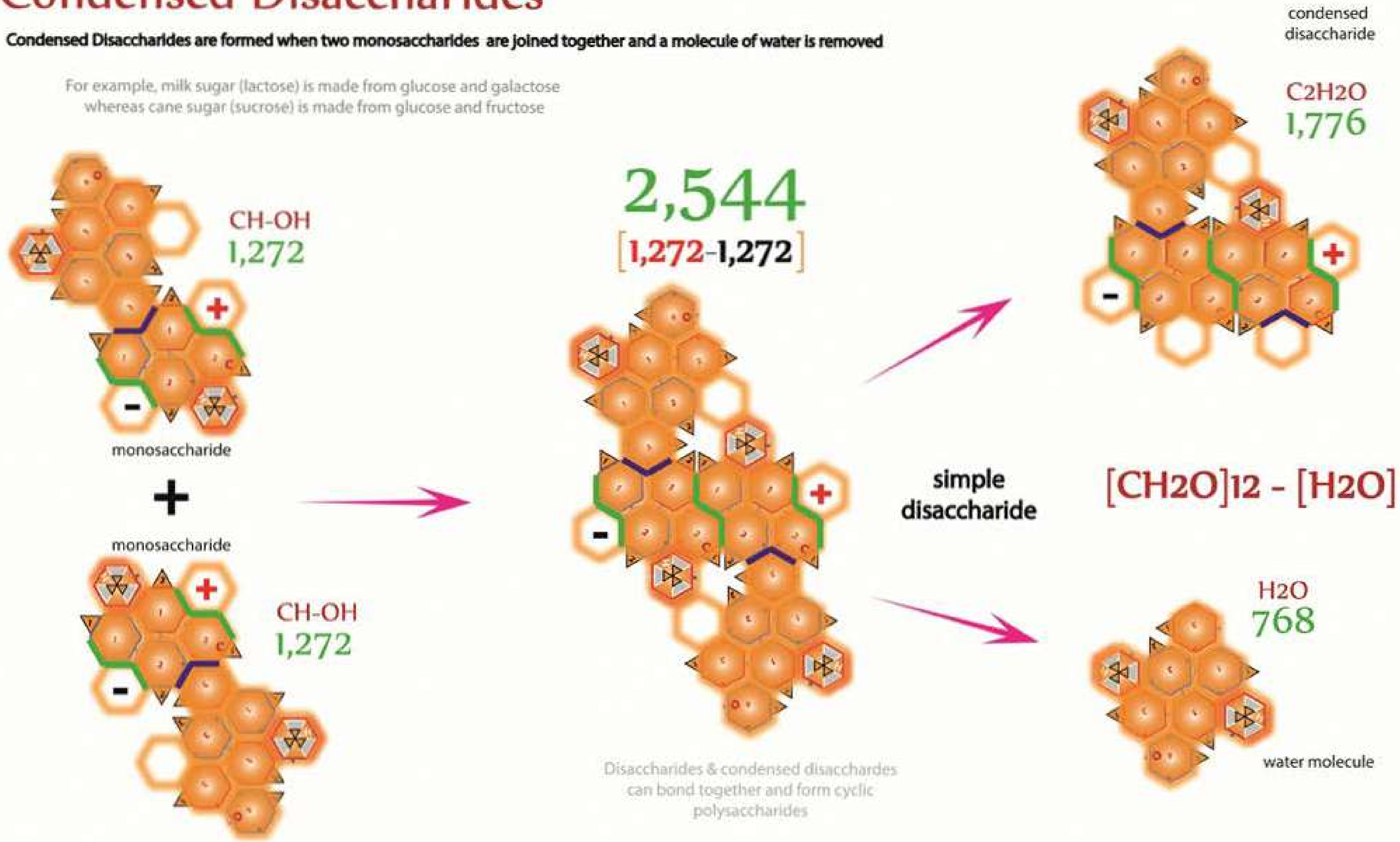

monosaccharide

CH-OH
1,272

+

monosaccharide

CH-OH
1,272

2,544
[1,272-1,272]

simple disaccharide

condensed disaccharide

C2H2O
1,776

[CH2O]12 - [H2O]

H2O
768

water molecule

Disaccharides & condensed disaccharides
can bond together and form cyclic
polysaccharides

The reverse of this reaction, the formation of two monosaccharides from one disaccharide, is called a hydrolysis reaction and requires one water molecule to supply the H and OH to the sugars formed.

Sucrose is used in many plants for transporting food reserves, often from the leaves to other parts of the plant.
Lactose is the sugar found in the milk of mammals and maltose is the first product of starch digestion
and is further broken down to glucose before absorption in the human gut.

Copyright ABRAHAM [2008] - All rights reserved

14,496

[7,248-7,248]

Disaccharides

Disaccharides are one of the four basic categories of carbohydrates,
(the others are monosaccharides, oligosaccharides, and polysaccharides)

Monosaccharides, such as glucose, are the monomers out of which disaccharides are constructed.

C_{12}-$(H_2O)_{11}$

Disaccharides & condensed disaccharides can bond together and form cyclic polysaccharides

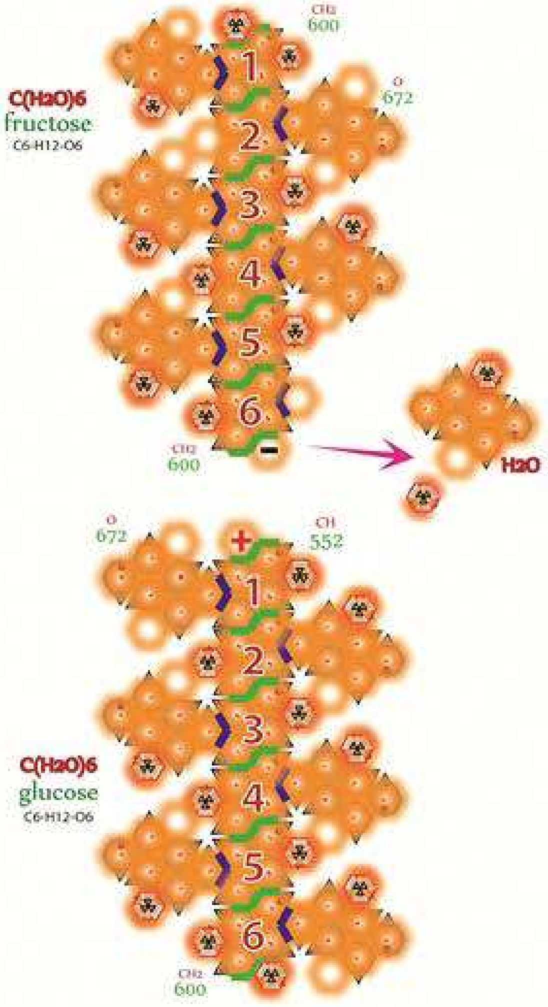

O
672
CH2
600
C(H2O)6
fructose
C6-H12-O6
1
2
3
4
5
6
H2O
CH
552
C(H2O)6
glucose
C6-H12-O6
CH2
600

sucrose
C12-H22-O11

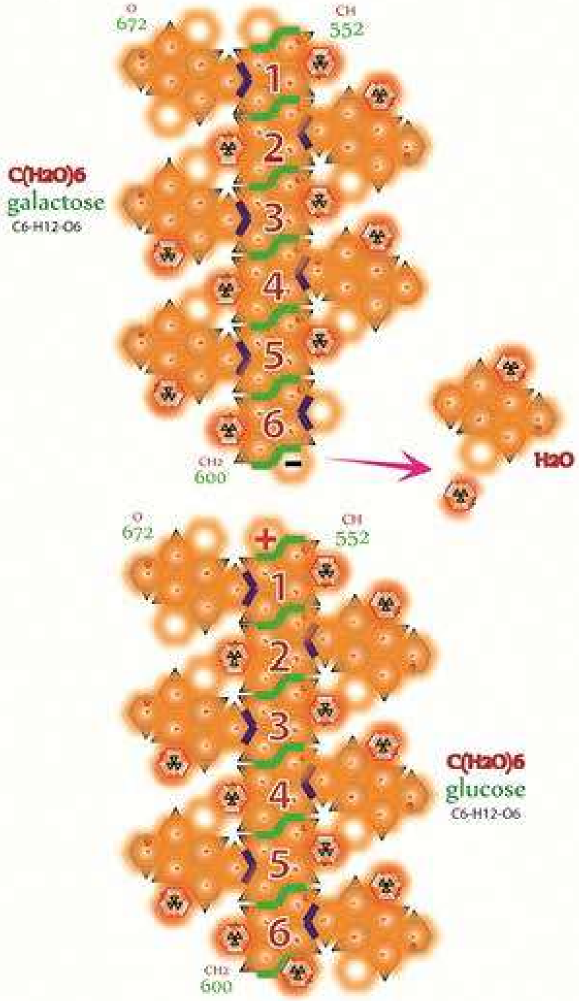

O
672
CH
552
C(H2O)6
galactose
C6-H12-O6
1
2
3
4
5
6
CH2
600
H2O
O
672
CH
552
C(H2O)6
glucose
C6-H12-O6
CH2
600

lactose
C12-H22-O11

The general chemical formula for carbohydrates, C(H2O),
gives the relative proportions of carbon, hydrogen, and oxygen in a monosaccharide
(the proportion of these atoms are 1:2:1)-(H2O)

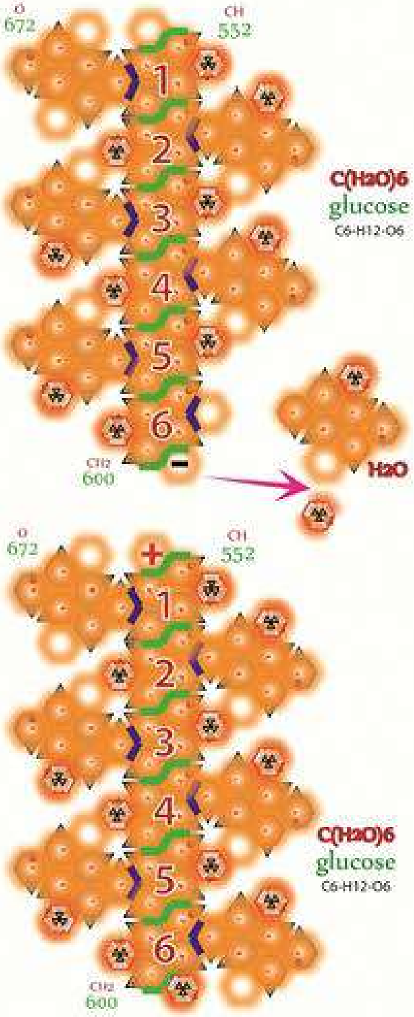

O
672
CH
552
C(H2O)6
glucose
C6-H12-O6
1
2
3
4
5
6
CH2
600
H2O
O
672
CH
552
C(H2O)6
glucose
C6-H12-O6
CH2
600

maltose
C12-H22-O11

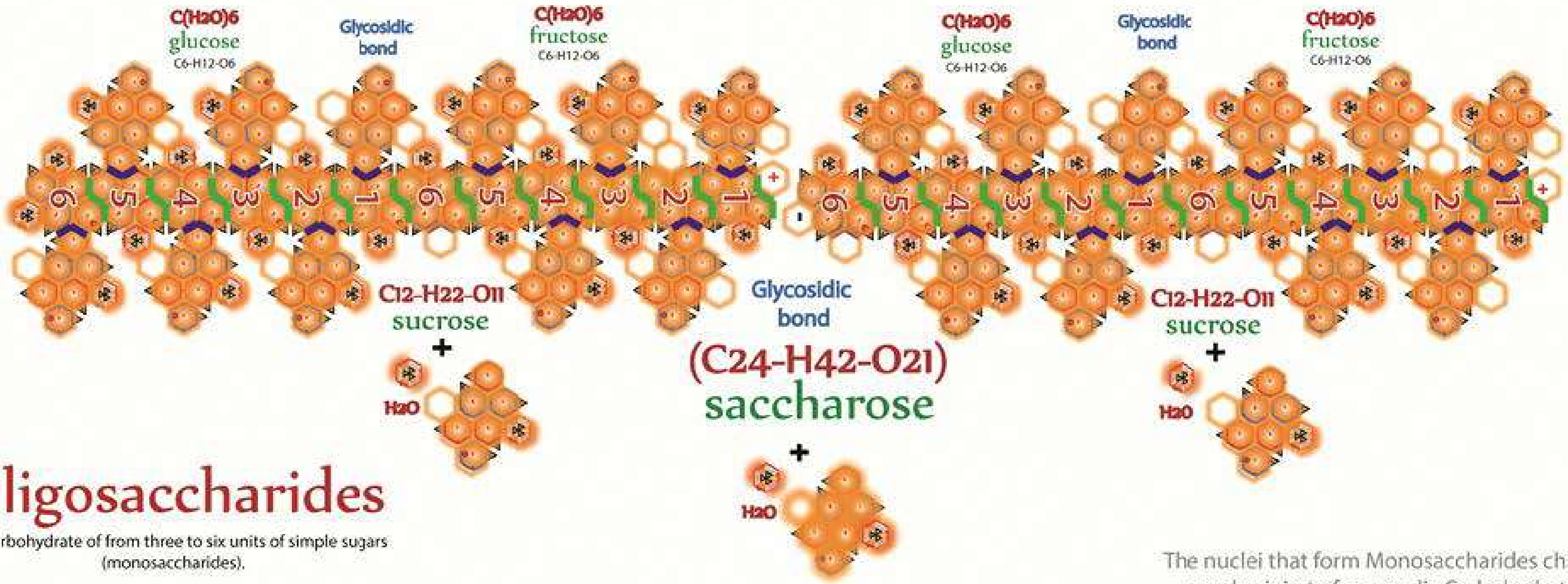
C(H2O)6
glucose
C6-H12-O6
Glycosidic
bond
C(H2O)6
fructose
C6-H12-O6
C(H2O)6
glucose
C6-H12-O6
Glycosidic
bond
C(H2O)6
fructose
C6-H12-O6
6
5
4
3
2
1
6
5
4
3
2
1
6
5
4
3
2
1
6
5
4
3
2
1
C12-H22-O11
sucrose
+
H2O
Glycosidic
bond
(C24-H42-O21)
saccharose
+
H2O
C12-H22-O11
sucrose
+
H2O

Oligosaccharides

any carbohydrate of from three to six units of simple sugars (monosaccharides).

The nuclei that form Monosaccharides chains can also join to form cyclic Carbohydrates [Polysaccharides]

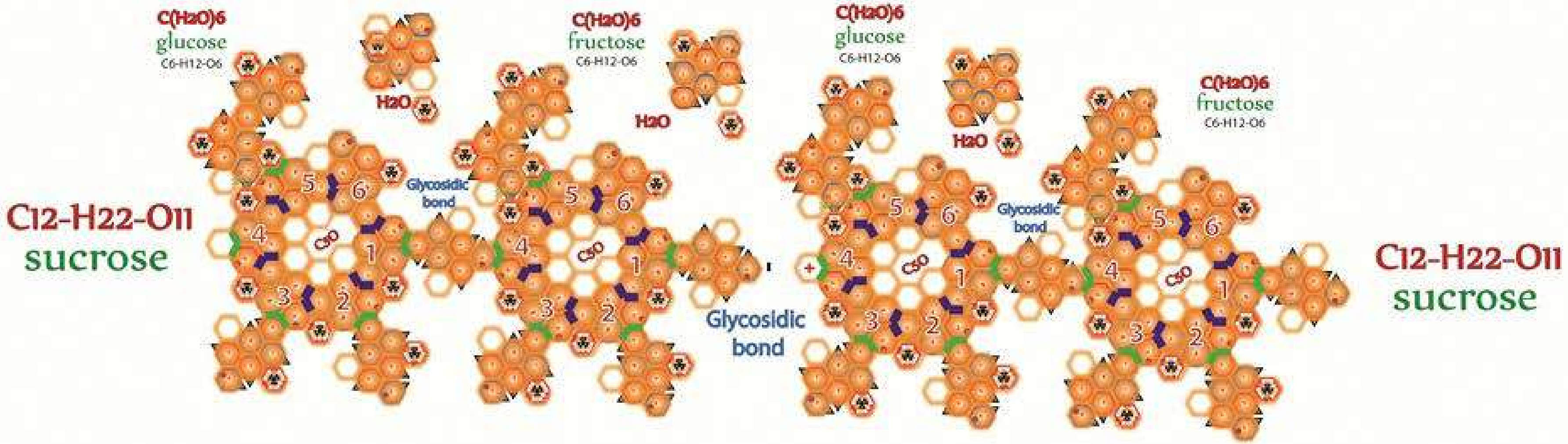
C(H2O)6
glucose
C6-H12-O6
H2O
C(H2O)6
fructose
C6-H12-O6
H2O
C(H2O)6
glucose
C6-H12-O6
H2O
C(H2O)6
fructose
C6-H12-O6
C12-H22-O11
sucrose
5
6
4
1
3
2
C6O
Glycosidic
bond
5
6
4
1
3
2
C6O
Glycosidic
bond
5
6
4
1
3
2
C6O
Glycosidic
bond
5
6
4
1
3
2
C6O
C12-H22-O11
sucrose

An oligosaccharide is a saccharide polymer containing a small number (typically two to ten) component sugars (monosaccharides). Oligosaccharides can have many functions; for example, they are commonly found on the plasma membrane of animal cells where they can play a role in cell–cell recognition.

Copyright ABRAHAM [2008] - All rights reserved

Starch or amylum is a carbohydrate consisting of a large number of glucose units joined by glycosidic bonds. This polysaccharide is produced by most green plants as an energy store.

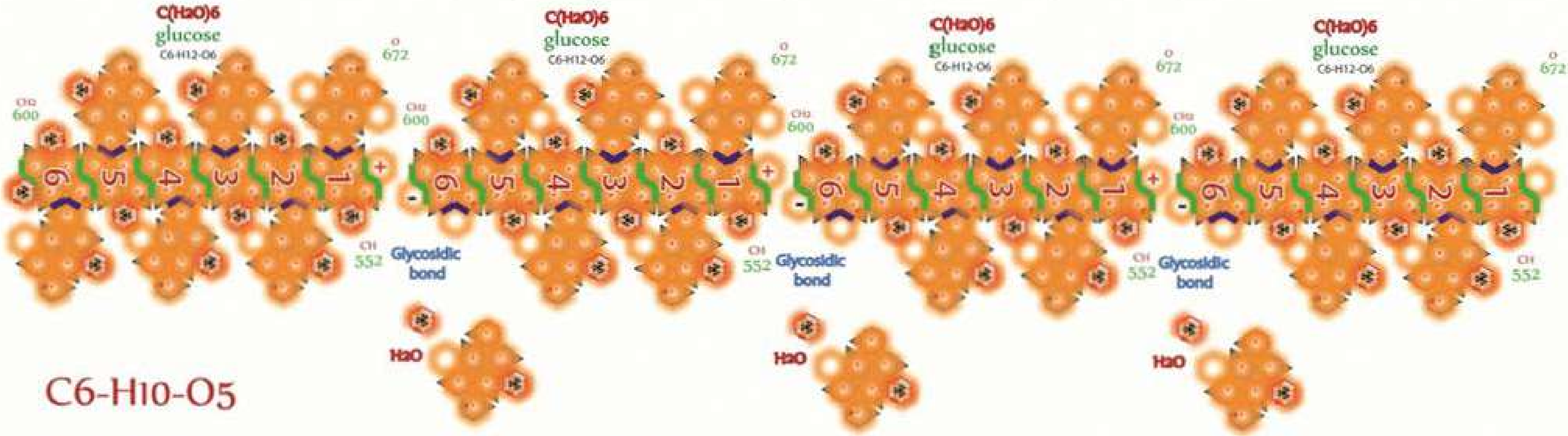
C(H2O)6
glucose
C6-H12-O6
O
672
CH2
600
CH
552
6
5
4
3
2
1
+
-
Glycosidic bond
H2O

C6-H10-O5

6,864 n

[3,432-3,432]

Polysaccharides

Polysaccharides have a general formula of Cx(H2O)y where x is usually a large number between 200 and 2500.

The repeating units in a Polysaccharide polymer backbone are often six-carbon monosaccharides, their general formula can be represented as (C6-H10-O5)n

$C_x[H2O]_y$

Polysaccharides are generally comprised of chains of cyclic monosaccharides

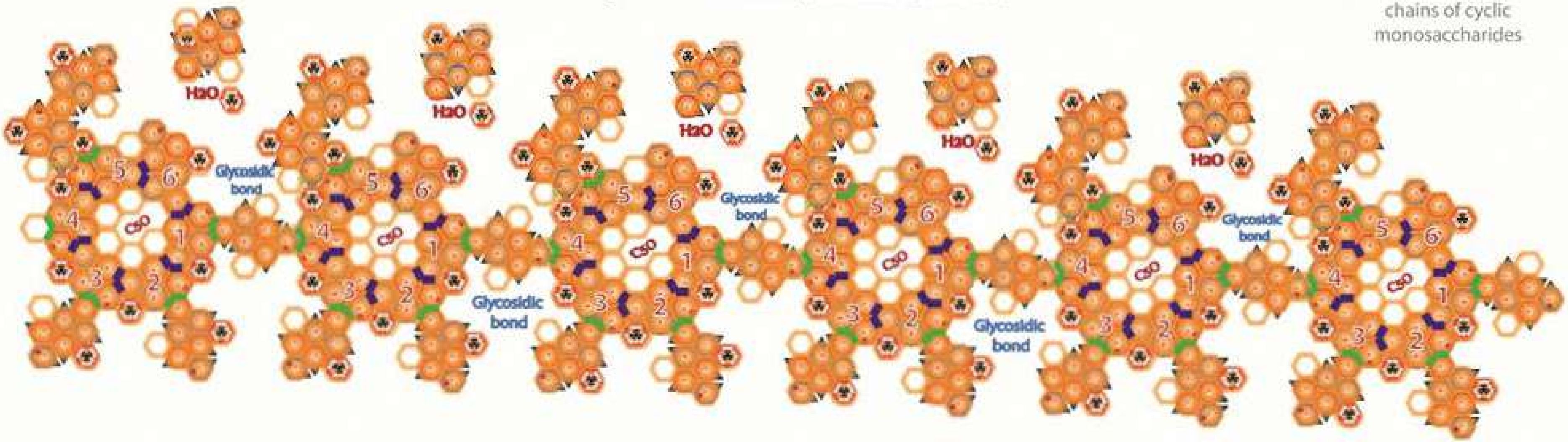
H2O
Glycosidic bond
C5O
1
2
3
4
5
6

Cellulose is an organic compound with the formula (C6H10O5)n, a polysaccharide consisting of a linear chain of several hundred to over ten thousand linked D-glucose sugars.

Copyright ABRAHAM [2008] - All rights reserved

Amines

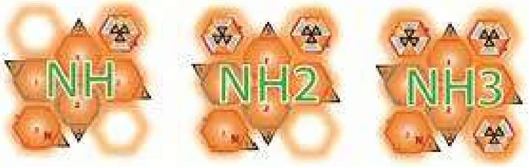

Amines are organic compounds that contain Nitrogen

NH2-[HC-R]-COOH + NH2-[HC-R]-COOH

As both the amine and carboxylic acid groups of amino acids can react to form amide bonds, one amino acid molecule can react with another and become joined through an amide linkage. This polymerization of amino acids is what creates proteins.

OH + NH2 -----> NH + H2O

This yields a peptide bond and a molecule of water via a condensation reaction

Carboxylic Acids

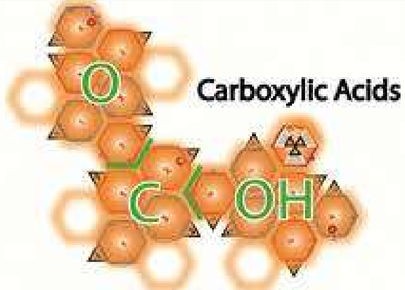

The general formula of a carboxylic acid is R-COOH

Amino Acids

Amino acids are biologically important organic compounds made from amine and carboxylic acid functional groups, along with a side-chain specific to each amino acid

un-ionized amino acid

NH2-[HC-R]-COOH

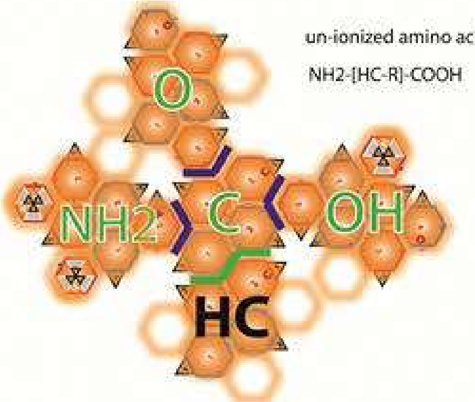

zwitterionic amino acid

NH3-[HC-R]-COO

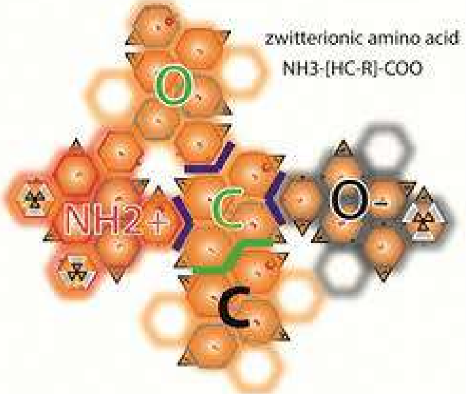

The α-carbon.

The α-amino acids in peptides and proteins consist of a carboxylic acid (-COOH) and an amino (-NH2) functional group attached to the same tetrahedral carbon atom

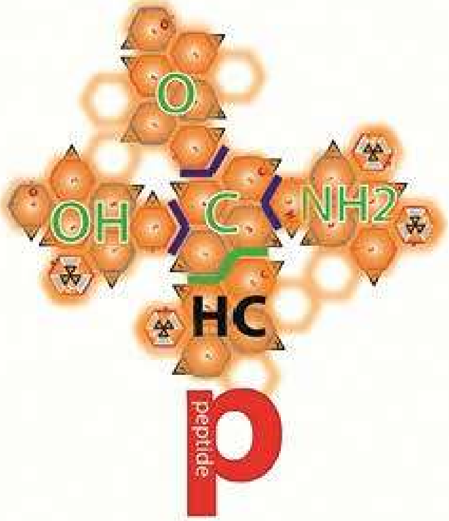

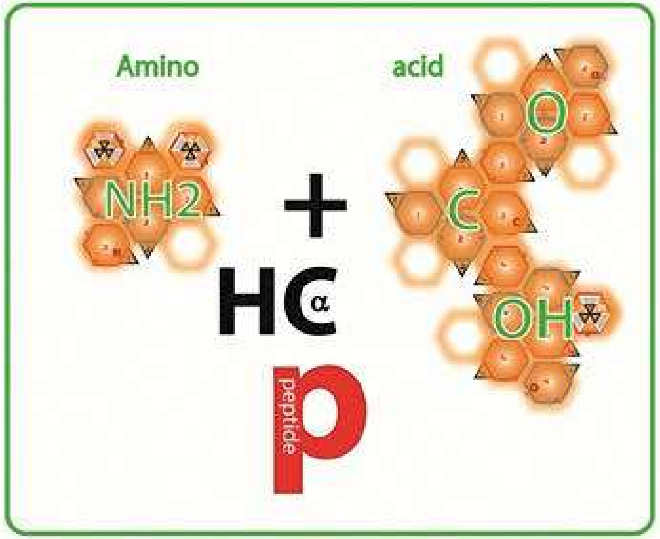

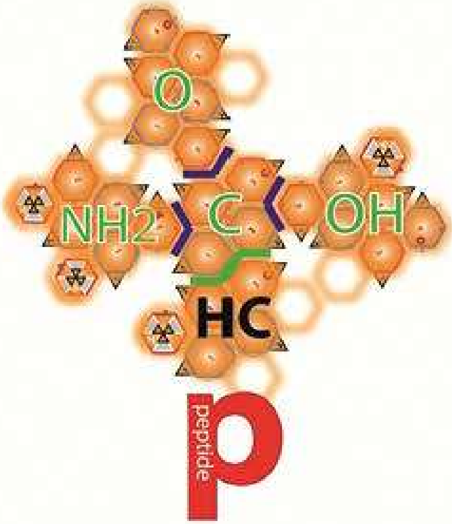

The peptide group attached to the alpha carbon distinguishes one amino acid from another (Tetryonic theory defines these compound side-chains of atoms as peptides [p])

Copyright ABRAHAM [2008] - All rights reserved

Peptide side chains

Amino acids are the structural units (monomers) that join together
to form short polymer chains called peptides;
or longer chains called either polypeptides or proteins

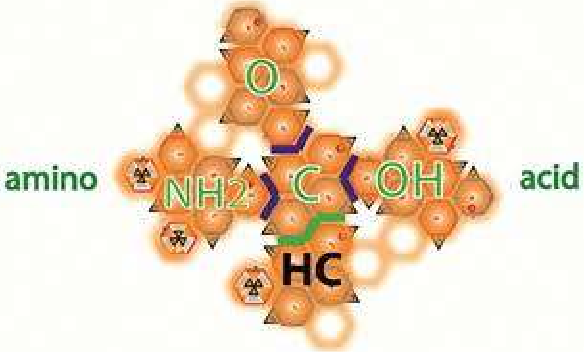

AAA
AAG
AAC
AAU
AGA
AGG
AGC
AGU
ACA
ACG
ACC
ACU
AUA
AUG
AUC
AUU

A
C5-H5-N5
Adenine

CAA
CAG
CAC
CAU
CGA
CGG
CGC
CGU
CCA
CCG
CCC
CCU
CUA
CUG
CUC
CUU

C
C4-H5-N3-O
Cytosine

3,180	**Gly**	C2H5NO2
3,780	**Ala**	C3H7NO2
4,452	**Ser**	C3H7NO3
4,884	**Pro**	C5H9NO2
4,980	**Val**	C5H11NO2
5,052	**Thr**	C4H9NO3
5,124	**Cys**	C3H7NO2S
5,580	**Lle**	C6H13NO2
5,580	**Leu**	C6H13NO2
5,592	**Asn**	C4H8N2O3
5,628	**Asp**	C4H7NO4
6,192	**Gln**	C5H10N2O3
6,216	**Lys**	C6H14N2O2
6,228	**Glu**	C5H9NO4
6,324	**Met**	C5H11NO2S
6,564	**His**	C6H9N3O2
6,996	**Phe**	C9H11NO2
7,392	**Arg**	C6H14N4O2
7,668	**Tyr**	C9H11NO3
8,640	**Trp**	C11H12N2O2

p
peptide side chain

*current chemical theory
hypothesizes that all
peptides and proteins
are the result of codon
triplet side chains
on amino acids*

*Tetryonic theory reveals
serious flaws in this line
of thought*

GAA
GAG
GAC
GAU
GGA
GGG
GGC
GGU
GCA
GCG
GCC
GCU
GUA
GUG
GUC
GUU

G
C5-H5-N5-O
Guanine

UAA
UAG
UAC
UAU
UGA
UGG
UGC
UGU
UCA
UCG
UCC
UCU
UUA
UUG
UUC
UUU

U
C4-H4-N2-O2
Uracil

Tetryonics 59.10 - Peptide side chains.jpg

Copyright ABRAHAM [2008] - All rights reserved

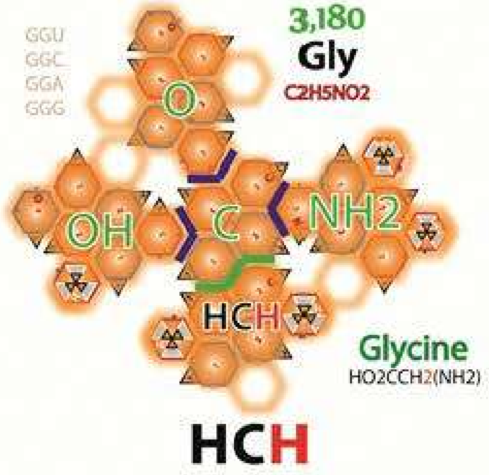
GGU
GGC
GGA
GGG
3,180
Gly
C2H5NO2
O
OH
C
NH2
HCH
Glycine
HO2CCH2(NH2)
HCH

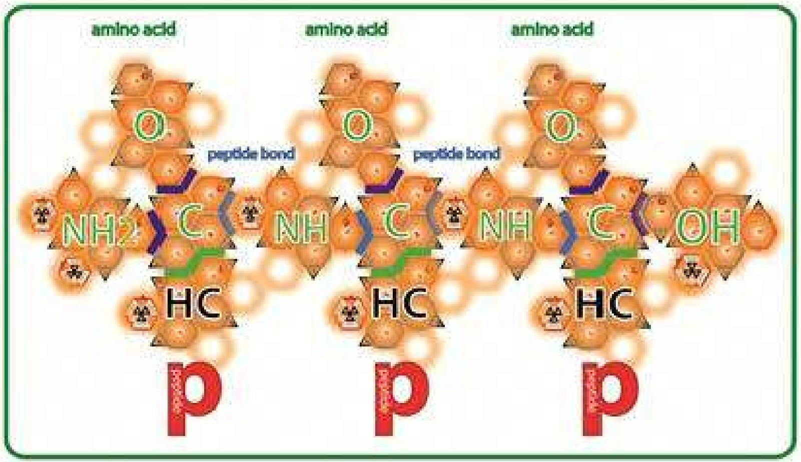
amino acid
amino acid
amino acid
O
O
O
peptide bond
peptide bond
NH2
C
NH
C
NH
C
OH
HC
HC
HC
p
peptide
p
peptide
p
peptide

poly-peptide chains

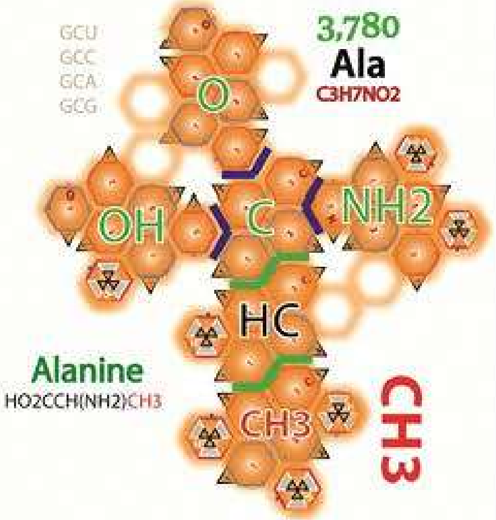
GCU
GCC
GCA
GCG
3,780
Ala
C3H7NO2
O
OH
C
NH2
HC
CH3
Alanine
HO2CCH(NH2)CH3
CH3

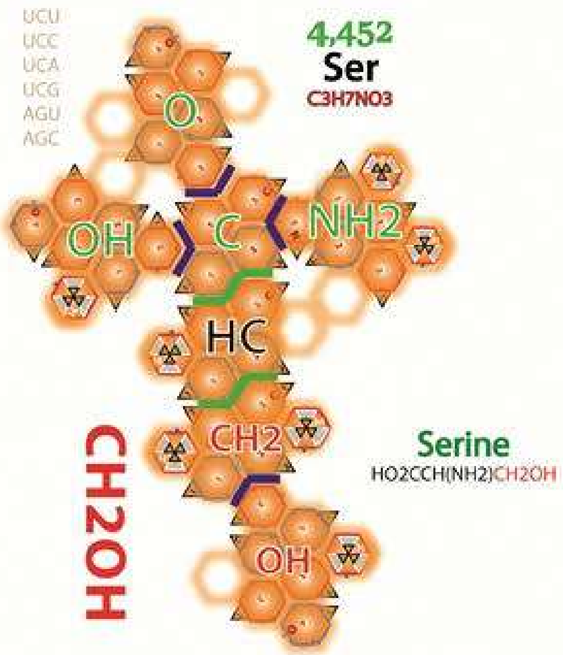
UCU
UCC
UCA
UCG
AGU
AGC
4,452
Ser
C3H7NO3
O
OH
C
NH2
HC
CH2
OH
Serine
HO2CCH(NH2)CH2OH
CH2OH

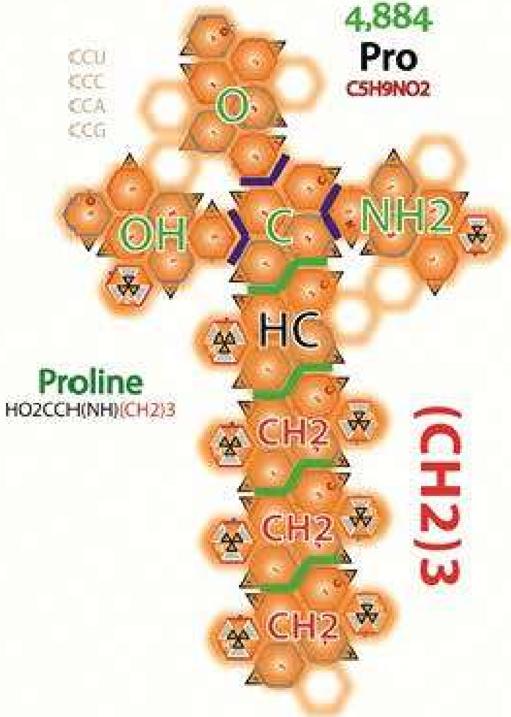
CCU
CCC
CCA
CCG
4,884
Pro
C5H9NO2
O
OH
C
NH2
HC
CH2
CH2
CH2
Proline
HO2CCH(NH)(CH2)3
(CH2)3

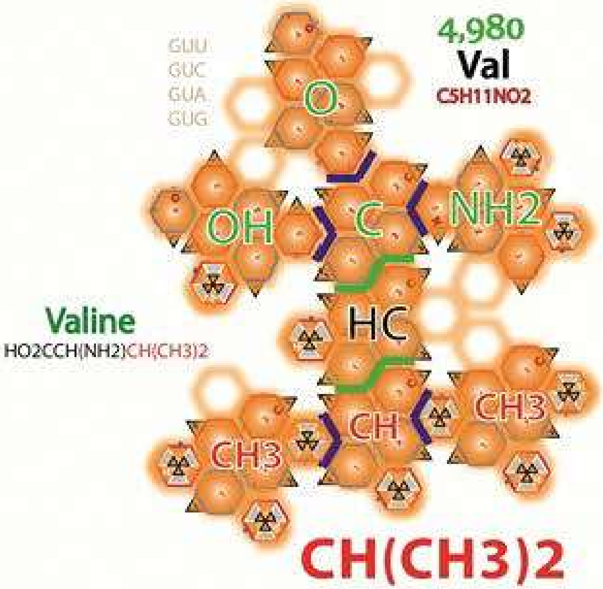
GUU
GUC
GUA
GUG
4,980
Val
C5H11NO2
O
OH
C
NH2
HC
CH3
CH
CH3
Valine
HO2CCH(NH2)CH(CH3)2
CH(CH3)2

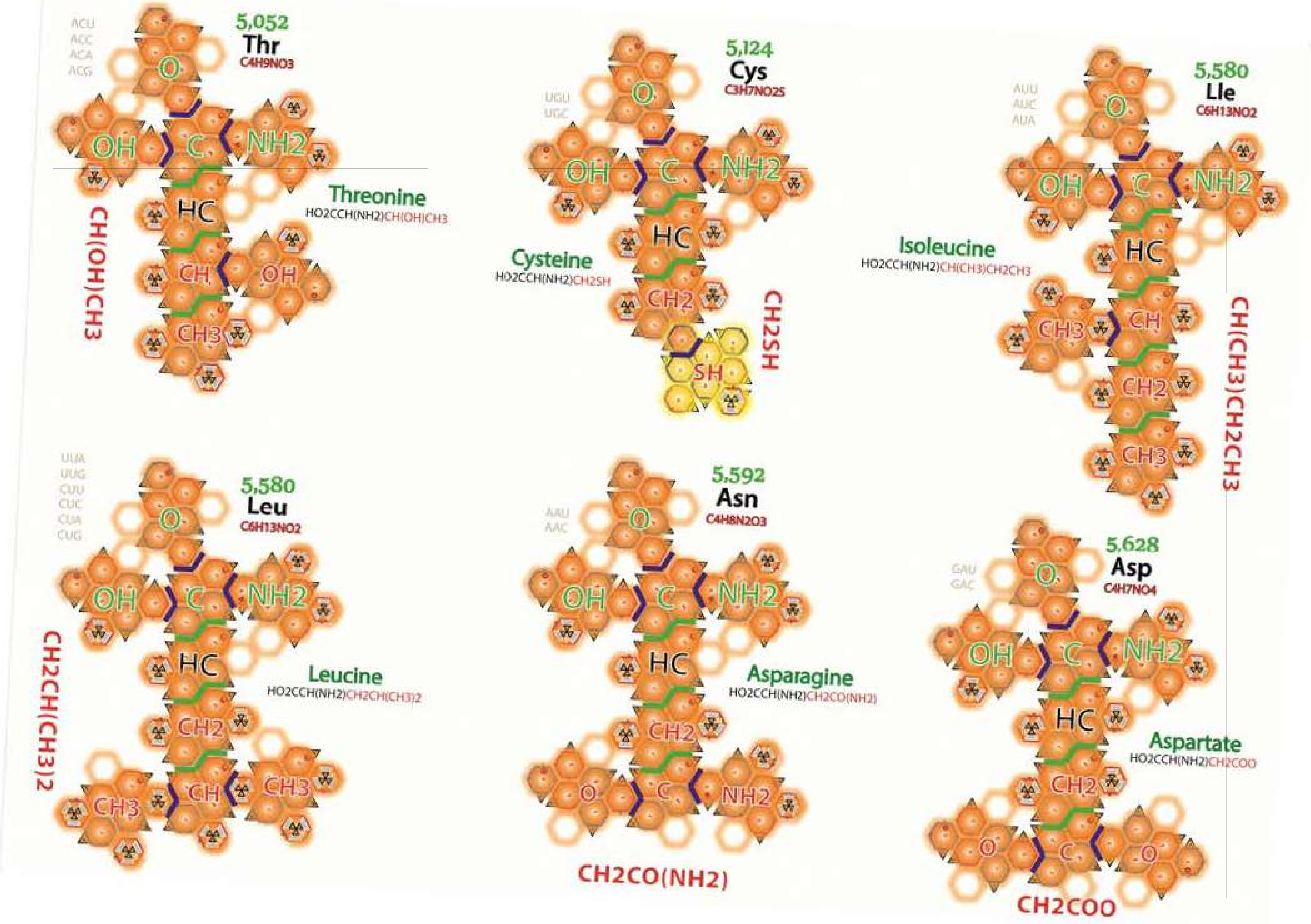
ACU
ACC
ACA
ACG
5,052
Thr
C4H9NO3
O
OH
C
NH2
HC
CH
OH
CH3
Threonine
HO2CCH(NH2)CH(OH)CH3
CH(OH)CH3
UGU
UGC
5,124
Cys
C3H7NO2S
O
OH
C
NH2
HC
CH2
SH
Cysteine
HO2CCH(NH2)CH2SH
CH2SH
AUU
AUC
AUA
5,580
Lle
C6H13NO2
O
OH
C
NH2
HC
CH3
CH
CH2
CH3
Isoleucine
HO2CCH(NH2)CH(CH3)CH2CH3
CH(CH3)CH2CH3
UUA
UUG
CUU
CUC
CUA
CUG
5,580
Leu
C6H13NO2
O
OH
C
NH2
HC
CH2
CH3
CH
CH3
Leucine
HO2CCH(NH2)CH2CH(CH3)2
CH2CH(CH3)2
AAU
AAC
5,592
Asn
C4H8N2O3
O
OH
C
NH2
HC
CH2
O
C
NH2
Asparagine
HO2CCH(NH2)CH2CO(NH2)
CH2CO(NH2)
GAU
GAC
5,628
Asp
C4H7NO4
O
OH
C
NH2
HC
CH2
O
C
O
Aspartate
HO2CCH(NH2)CH2COO
CH2COO

Copyright ABRAHAM [2008] - All rights reserved

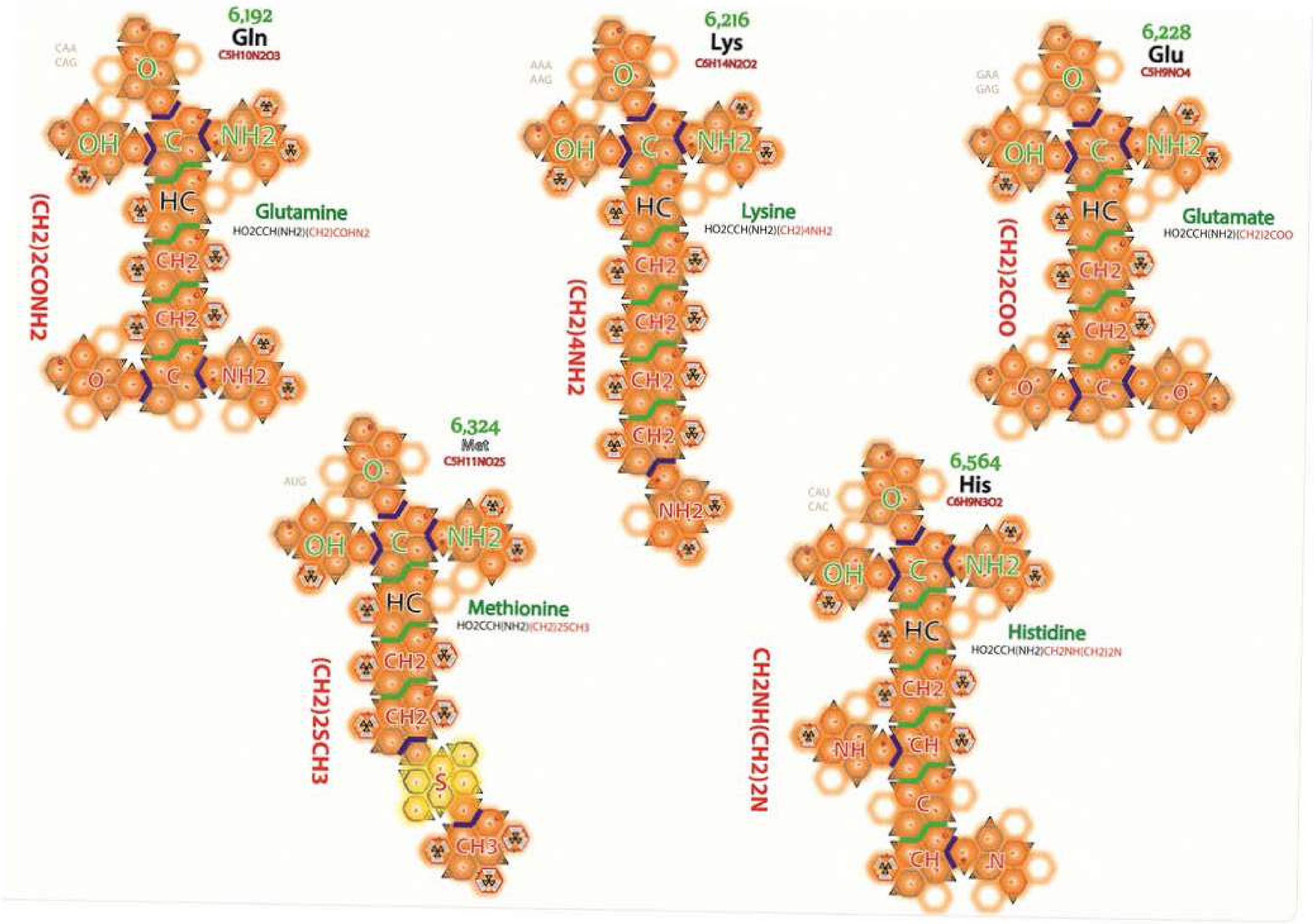
6,192
Gln
C5H10N2O3
CAA
CAG
O
OH
C
NH2
HC
CH2
CH2
O
C
NH2
Glutamine
HO2CCH(NH2)(CH2)COHN2
(CH2)2CONH2
6,216
Lys
C6H14N2O2
AAA
AAG
O
OH
C
NH2
HC
CH2
CH2
CH2
CH2
NH2
Lysine
HO2CCH(NH2)(CH2)4NH2
(CH2)4NH2
6,228
Glu
C5H9NO4
GAA
GAG
O
OH
C
NH2
HC
CH2
CH2
O
C
O
Glutamate
HO2CCH(NH2)(CH2)2COO
(CH2)2COO
6,324
Met
C5H11NO2S
AUG
O
OH
C
NH2
HC
CH2
CH2
S
CH3
Methionine
HO2CCH(NH2)(CH2)2SCH3
(CH2)2SCH3
6,564
His
C6H9N3O2
CAU
CAC
O
OH
C
NH2
HC
CH2
NH
CH
C
CH
N
Histidine
HO2CCH(NH2)CH2NH(CH2)2N
CH2NH(CH2)2N

Copyright ABRAHAM [2008] - All rights reserved

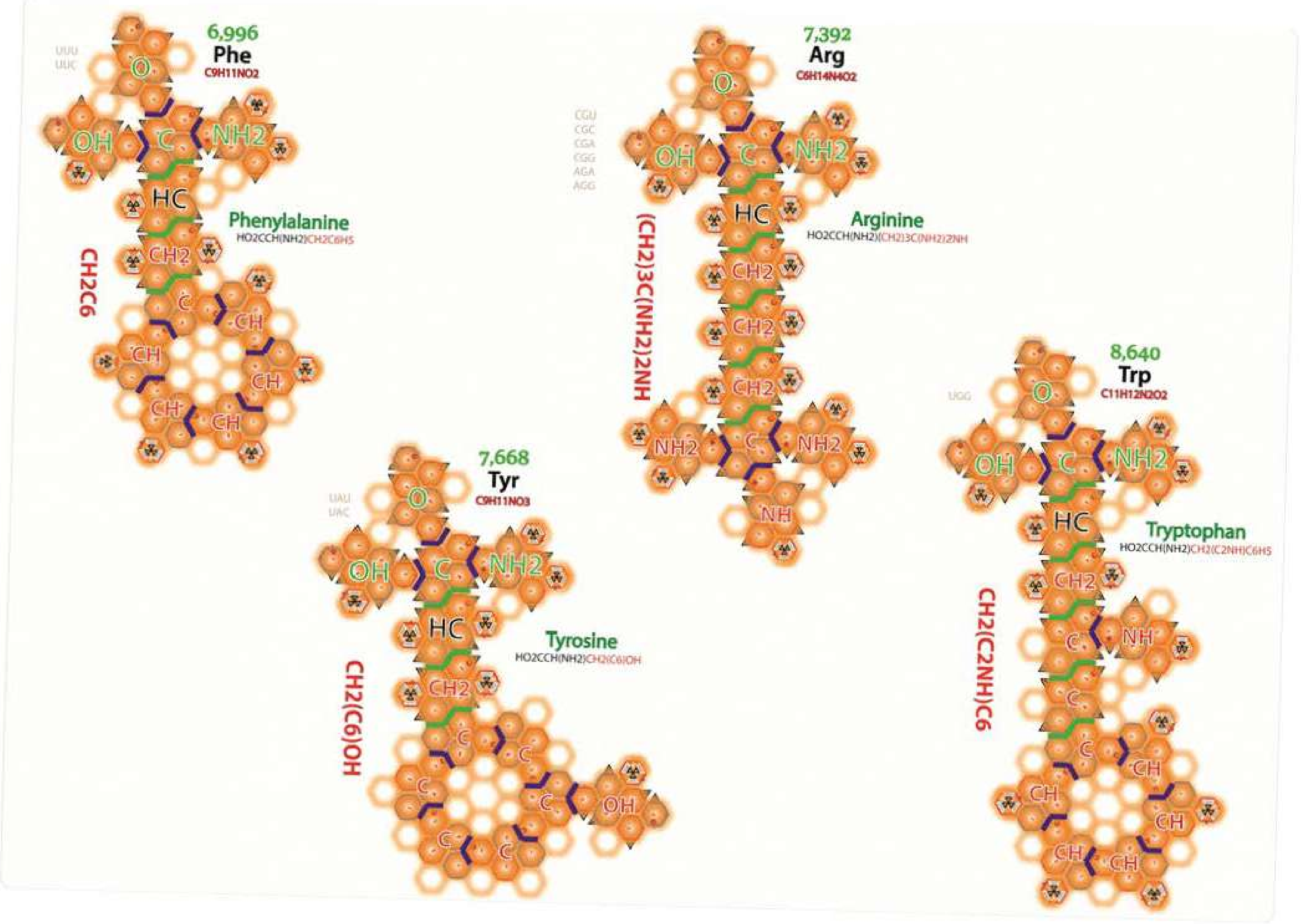
6,996
Phe
C9H11NO2
UUU
UUC
O
OH
C
NH2
HC
Phenylalanine
HO2CCH(NH2)CH2C6H5
CH2
CH2C6
C
CH
CH
CH
CH
CH
7,392
Arg
C6H14N4O2
CGU
CGC
CGA
CGG
AGA
AGG
O
OH
C
NH2
HC
Arginine
HO2CCH(NH2)(CH2)3C(NH2)2NH
(CH2)3C(NH2)2NH
CH2
CH2
CH2
NH2
C
NH2
NH
7,668
Tyr
C9H11NO3
UAU
UAC
O
OH
C
NH2
HC
Tyrosine
HO2CCH(NH2)CH2(C6)OH
CH2
CH2(C6)OH
C
C
C
C
OH
C
C
8,640
Trp
C11H12N2O2
UGG
O
OH
C
NH2
HC
Tryptophan
HO2CCH(NH2)CH2(C2NH)C6H5
CH2
CH2(C2NH)C6
C
NH
C
C
CH
CH
CH
CH
CH

Peptide Bonds

The four-atom functional group -CONH- is called a peptide link.

Polypeptide chains of amino acids arre held together by peptide bonds

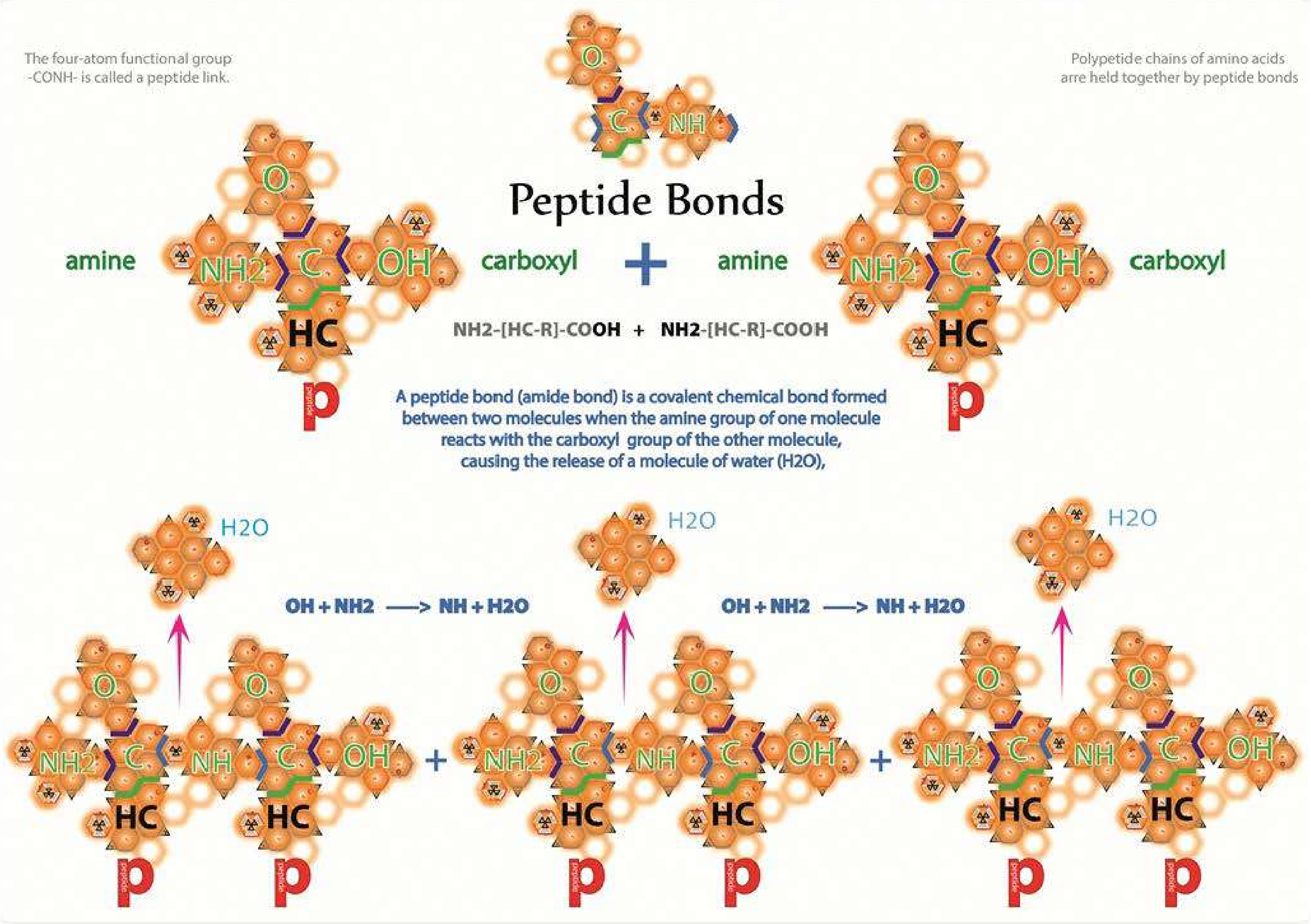
O
C
NH
amine
NH2
C
OH
carboxyl
HC
p
+
amine
NH2
C
OH
carboxyl
HC
p
H2O
H2O
H2O
O
NH2
C
NH
C
OH
HC
HC
p
p
+
NH2
C
NH
C
OH
HC
HC
p
p
+
NH2
C
NH
C
OH
HC
HC
p
p

NH2-[HC-R]-COOH + NH2-[HC-R]-COOH

A peptide bond (amide bond) is a covalent chemical bond formed between two molecules when the amine group of one molecule reacts with the carboxyl group of the other molecule, causing the release of a molecule of water (H2O),

OH + NH2 ——> NH + H2O

OH + NH2 ——> NH + H2O

Tetryonics 59.15 - Peptide bonds

amino terminus

The beginning of a polypeptide protein
molecule has a free amino group.

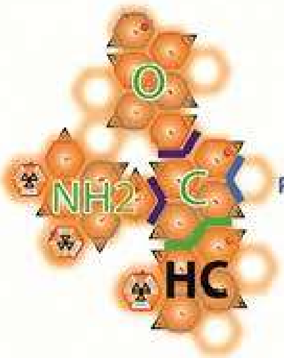

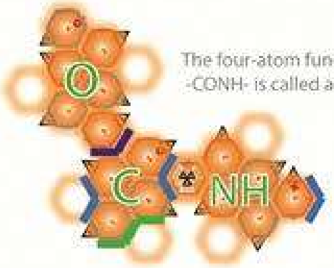

The four-atom functional group -CONH- is called a peptide link.

carboxy terminus

The end of a polypeptide protein
molecule has a free carboxyl group.

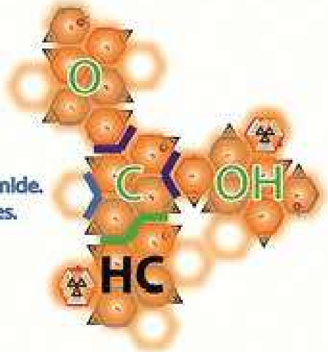

Proteins [poly-peptides] are formed by joining the -CO2H [carboxyl] end of one amino acid with the -NH2 [amine] end of another to form an amide. The -CONH- bond between amino acids is known as a peptide bond because relatively short polymers of amino acids are known as peptides.

polypetide chains

Proteins are formed when amino acids are covalently linked together.

DNA and RNA have a deoxyribose and ribose sugar backbone, respectively, whereas PNA's backbone is composed of repeating (2-aminoethyl)-glycine units linked by peptide bonds.

Proteins (polypeptides) are chains of amino acids held together by peptide bonds which form the backbone stringers of PNA, RNA & DNA.

Proteins

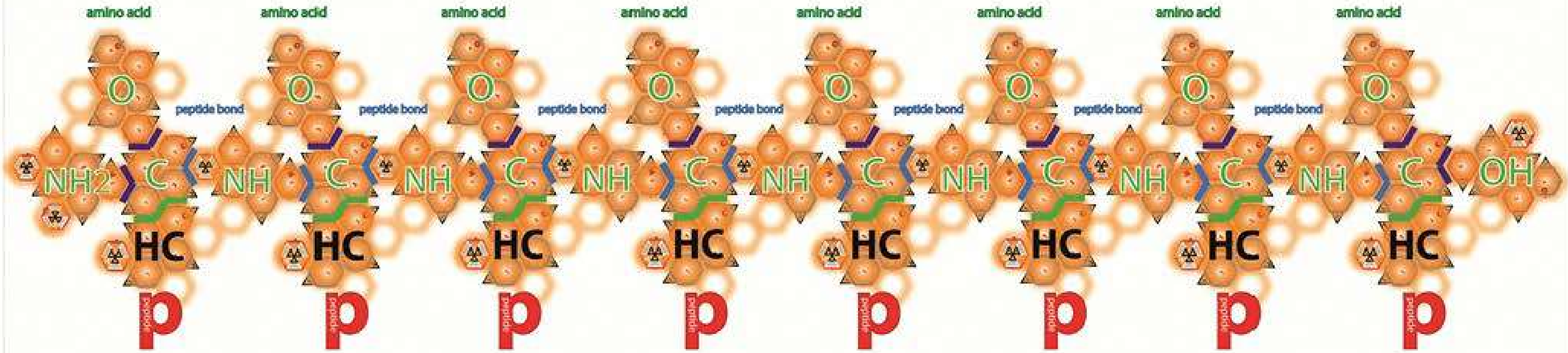
amino acid
amino acid
amino acid
amino acid
amino acid
amino acid
amino acid
amino acid
peptide bond
peptide bond
peptide bond
peptide bond
peptide bond
peptide bond
peptide bond
O
NH2
C
NH
HC
OH
peptide

There are 20 different nucleotide side chains present in biological molecules.

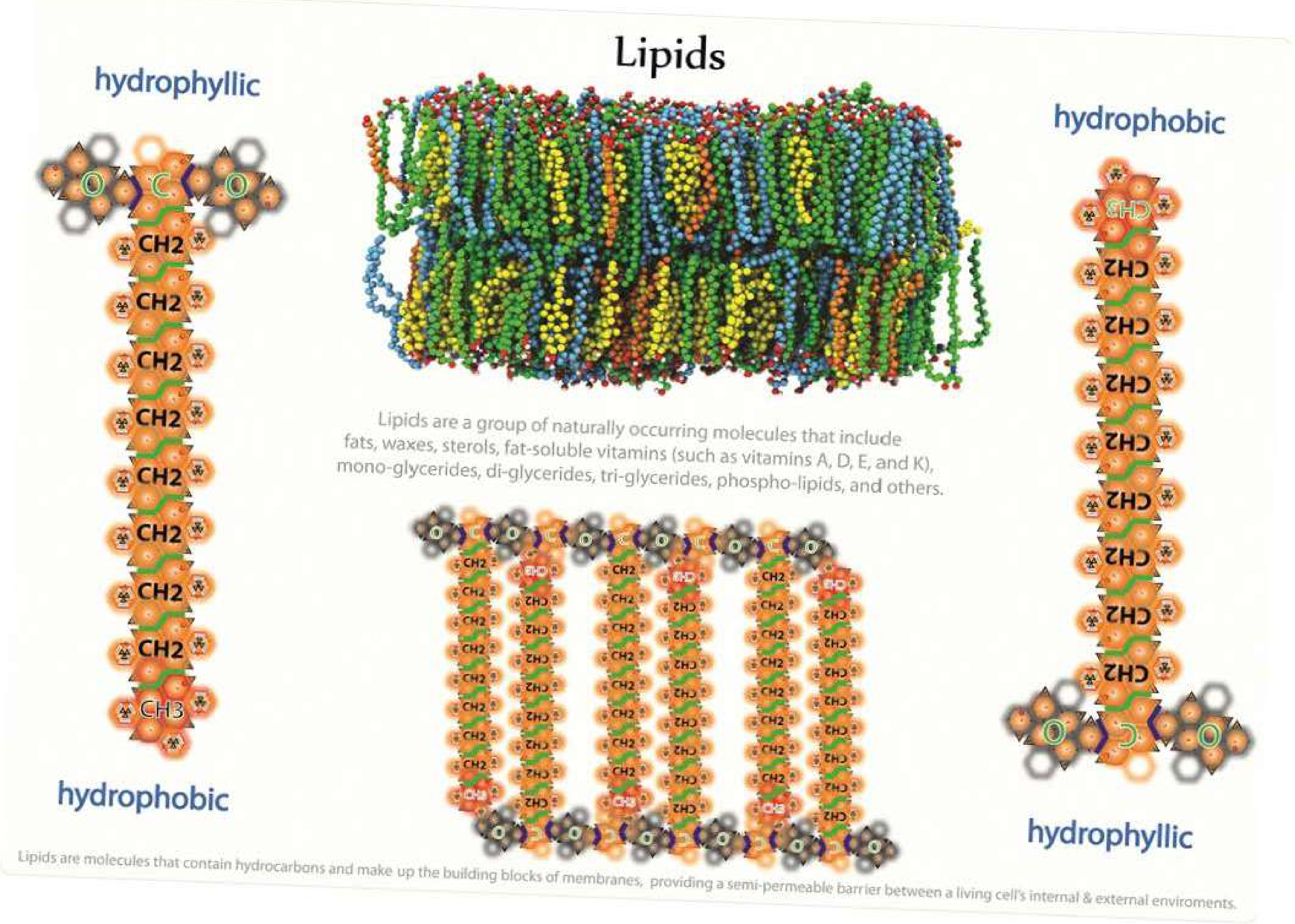

Lipids

Lipids are a group of naturally occurring molecules that include fats, waxes, sterols, fat-soluble vitamins (such as vitamins A, D, E, and K), mono-glycerides, di-glycerides, tri-glycerides, phospho-lipids, and others.

Lipids are molecules that contain hydrocarbons and make up the building blocks of membranes, providing a semi-permeable barrier between a living cell's internal & external enviroments.

hydrophyllic
O
C
O
CH2
CH3
hydrophobic
hydrophobic
hydrophyllic

Acetylcholine

is an organic molecule that acts as a neurotransmitter in many organisms, including humans. It is an ester of acetic acid and choline

Ach

C7H16NO2

CH3COO(CH2)2N+(CH3)3

Lewis chemical structure

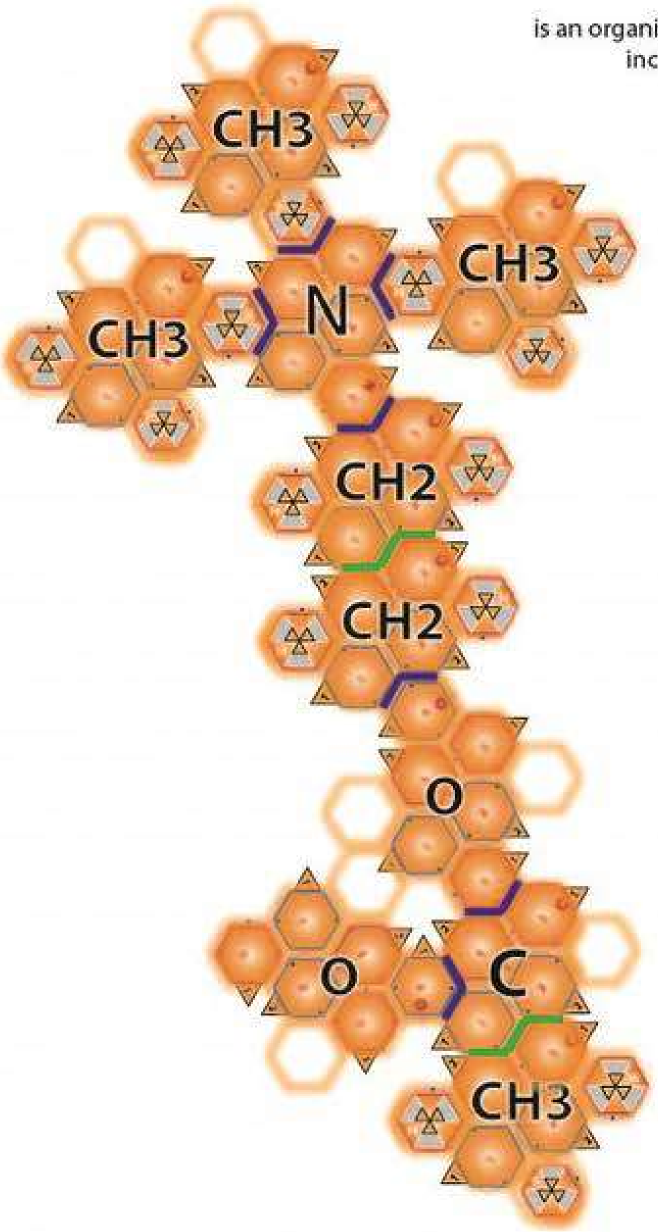
CH3
CH3
N
CH3
CH2
CH2
O
O
C
CH3

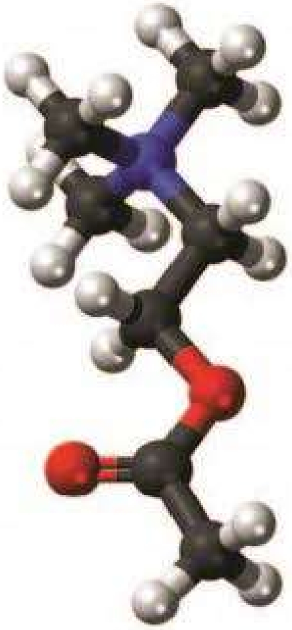

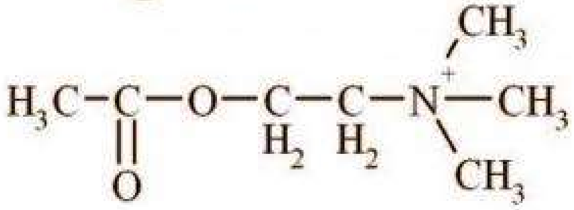

$$\mathrm{H_3C{-}\underset{\underset{\textstyle O}{\|}}{C}{-}O{-}\underset{H_2}{C}{-}\underset{H_2}{C}{-}\overset{+}{N}\!\overset{\diagup CH_3}{\underset{\diagdown CH_3}{-CH_3}}}$$

Tetryonic charge topology

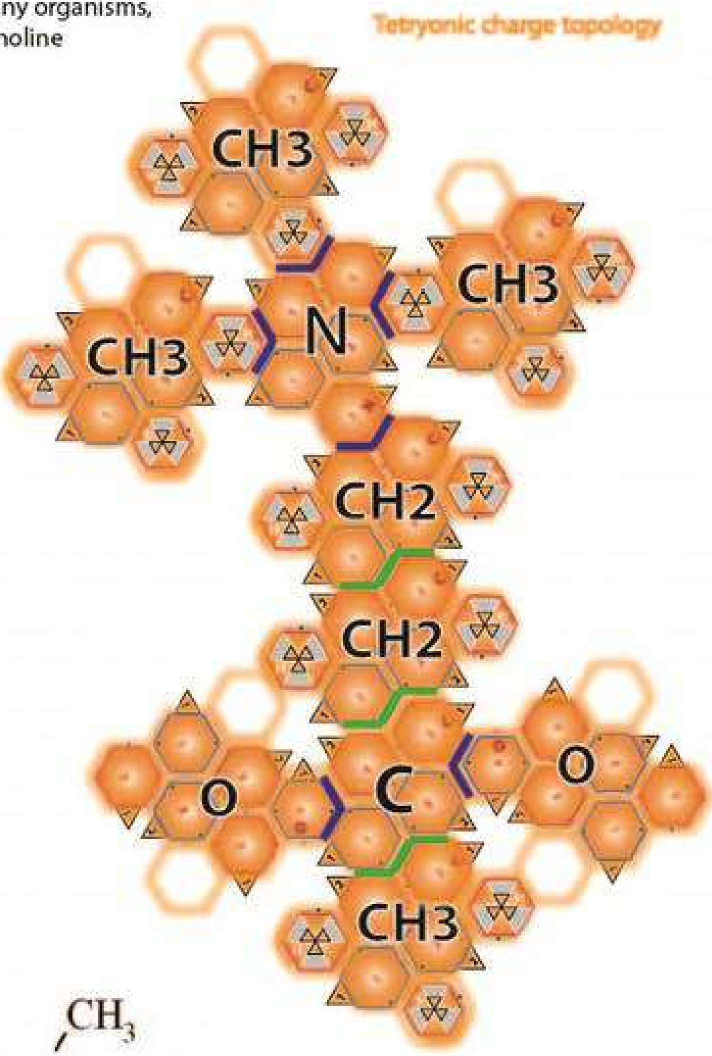
CH3
CH3
N
CH3
CH2
CH2
O
C
O
CH3

6,228

Tetryonics 59.18 - Acetylcholine [Ach]

Ribose

6,360

[3,180-3,180]

C5-H10-O5

Carbohydrate compounds can form acyclic sugars

Carbohydrate compounds can also form cyclic sugars

Ribose is an organic compound
with the formula C5H10O5; specifically,
a monosaccharide (simple sugar) with
linear form H–(C=O)–(CHOH)4–H

C(H2O)5
ribose
C5-H10-O5

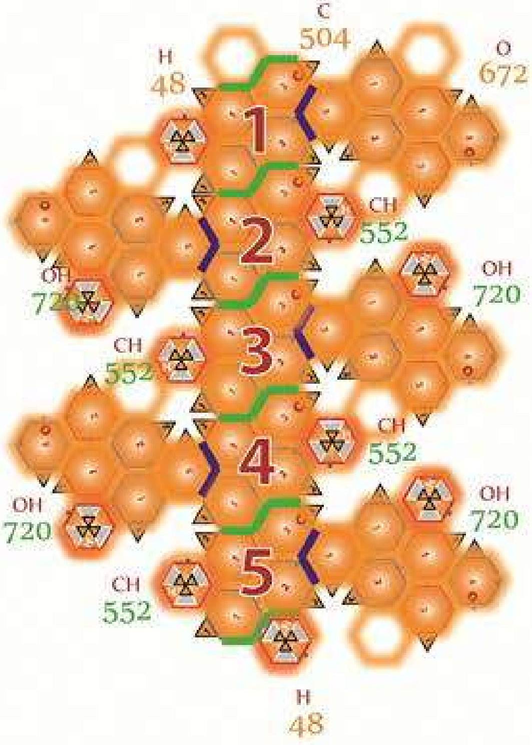
C
504
H
48
O
672
1
2
CH
552
OH
720
OH
720
3
CH
552
4
CH
552
OH
720
5
OH
720
CH
552
H
48

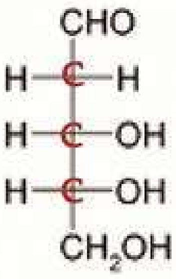
CHO
H
C
H
H
C
OH
H
C
OH
CH2OH

H-(CO)-(CHOH)4-H

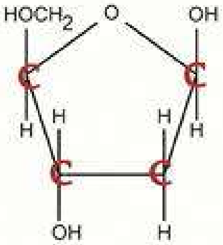
HOCH2
O
OH
C
C
H
H
H
H
C
C
OH
H

C(H2O)5
ribose
C5-H10-O5

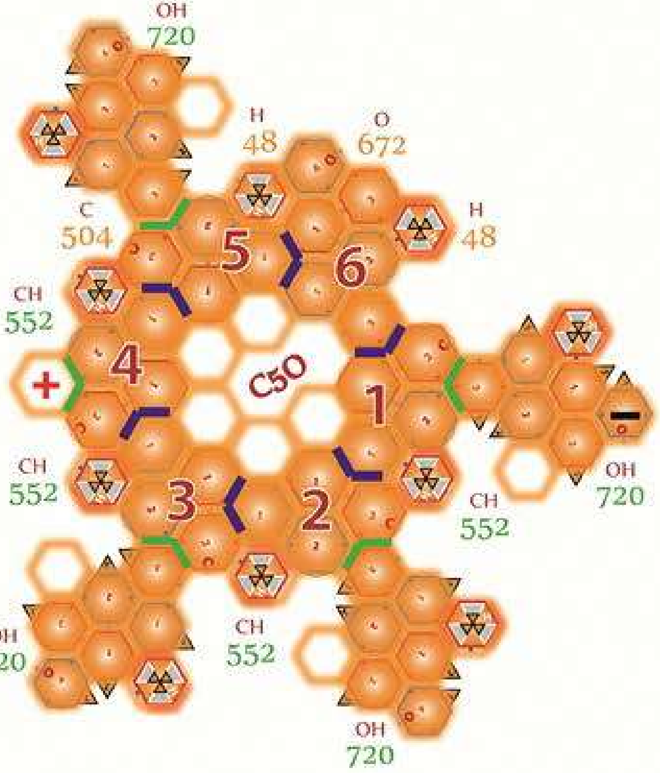
OH
720
H
48
O
672
C
504
5
6
H
48
CH
552
4
+
C5O
1
−
CH
552
OH
720
3
2
CH
552
OH
720
CH
552
OH
720

Ribose forms part of the backbone of RNA.
It is related to deoxyribose, which is found in DNA.

**Ribose is present within every living cell of the body and is used to manufacture ATP (the energy currency of the cell) from scratch.
Whilst the body can manufacture its own ribose from glucose, this requires energy and is a very slow process.**

As its name indicates it is a deoxygenated sugar, meaning that it is derived from the sugar ribose by loss of an oxygen atom

Deoxyribose

5,688

[2,844-2,844]

As a component of DNA, 2-deoxyribose derivatives have an important role in biology

The DNA (deoxyribonucleic acid) molecule, which is the main repository of genetic information in life, consists of a long chain of deoxyribose-containing units called nucleotides, linked via phosphate groups

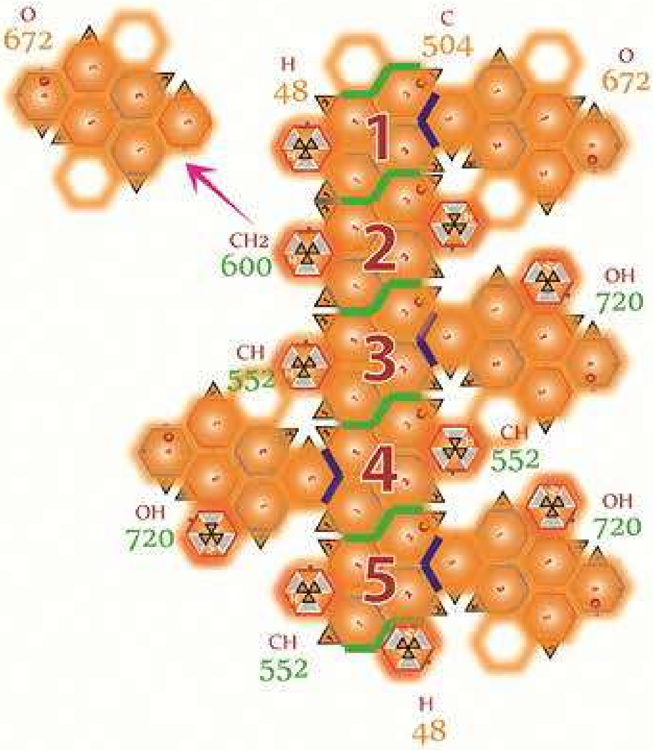
O
672
C
504
O
672
H
48
1
2
CH2
600
OH
720
3
CH
552
4
CH
552
OH
720
OH
720
5
CH
552
H
48

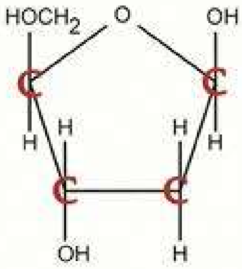
HOCH2
O
OH
C
C
H
H
H
H
C
C
OH
H

Deoxyribose, more, precisely 2-deoxyribose, is a monosaccharide

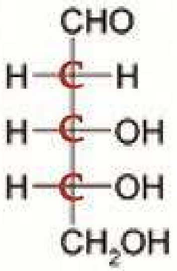
$$\begin{array}{c} CHO \\ H-C-H \\ H-C-OH \\ H-C-OH \\ CH_2OH \end{array}$$

C5-H10-O4
deoxyribose

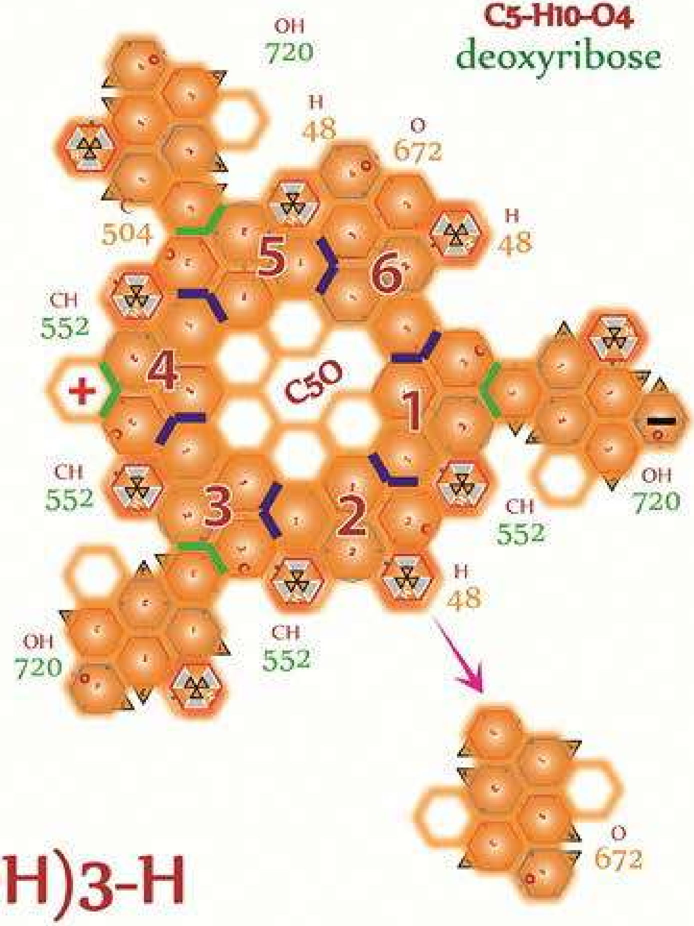
OH
720
H
48
O
672
C
504
H
48
5
6
CH
552
4
+
C5O
1
−
CH
552
3
2
OH
720
CH
552
H
48
OH
720
CH
552
O
672

C5-H10-O4
deoxyribose

H-(CO)-(CH2)-(CHOH)3-H

Hermann Emil Fischer won the Nobel Prize in Chemistry (1902) for his work in determining the structure of the D-aldohexoses.

However, the linear, free-aldehyde structures that Fischer proposed represent a very minor percentage of the forms that hexose sugars adopt in solution. It was Edmund Hirst and Clifford Purves, in the research group of Walter Haworth, who conclusively determined that the hexose sugars preferentially form a pyranose, or six-membered, ring. Haworth drew the ring as a flat hexagon with groups above and below the plane of the ring – the Haworth projection.

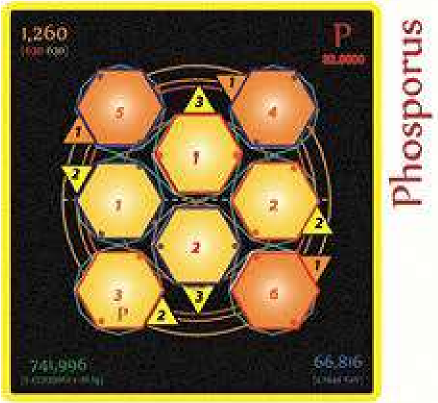
1,260
P
5
4
1
2
3
6
741,996
66,816
Phosporus

3,996

[1,998-1,998]

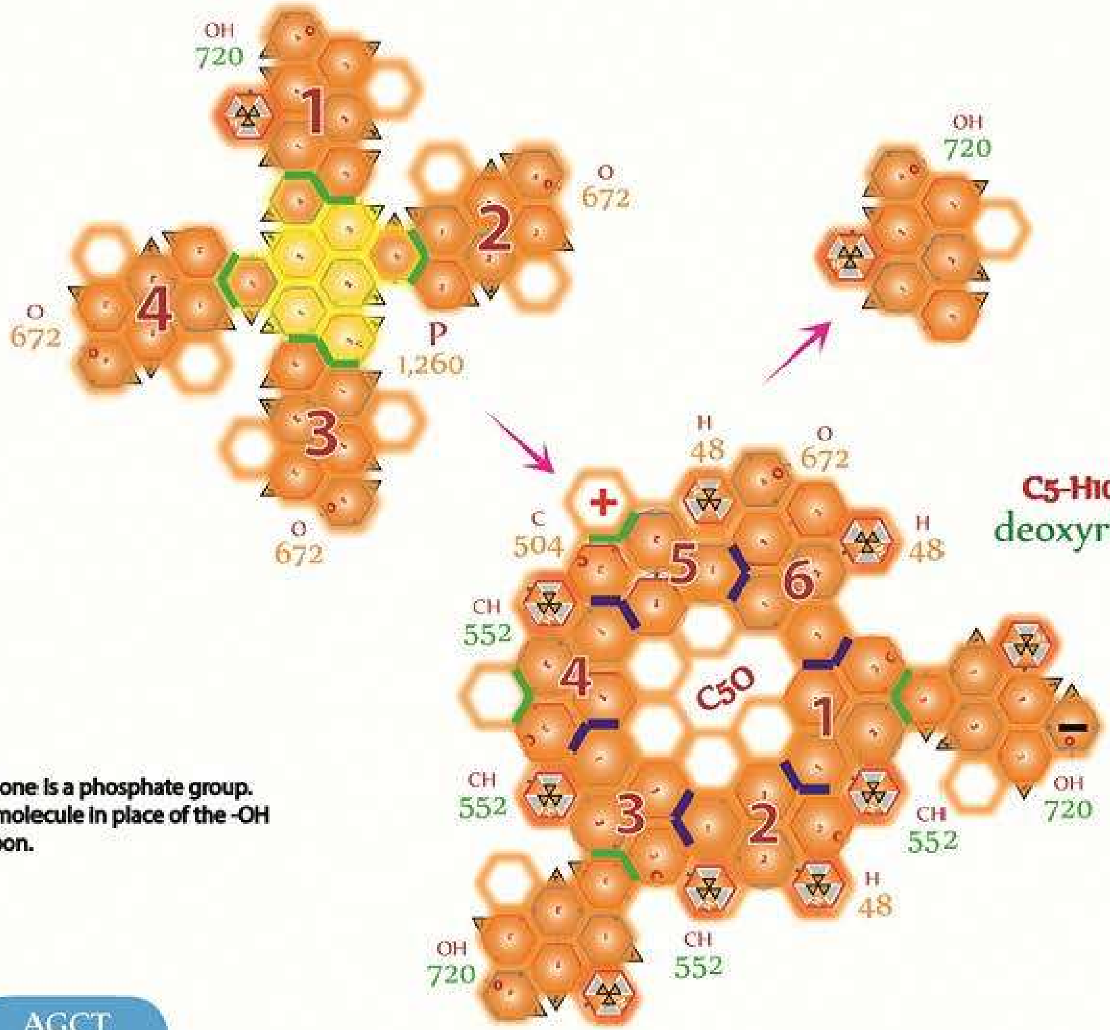
OH
720
1
O
672
2
4
O
672
P
1,260
3
O
672
OH
720
H
48
O
672
+
C
504
5
6
H
48
CH
552
4
C5O
1
−
OH
720
CH
552
3
2
CH
552
H
48
OH
720
CH
552

HPO4

phosphate
group

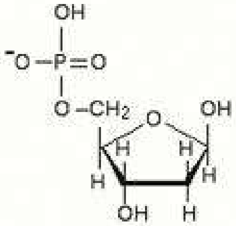

The other repeating part of the DNA backbone is a phosphate group. A phosphate group is attached to the sugar molecule in place of the -OH group on the 5' carbon.

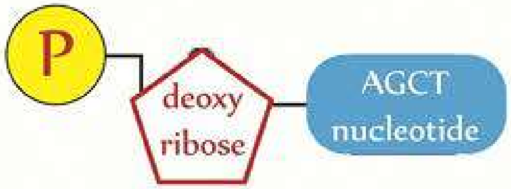
P
deoxy
ribose
AGCT
nucleotide

C5-H10-O4

deoxyribose

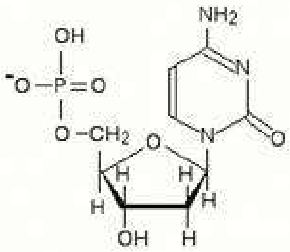

5,688

[2,844-2,844]

Copyright ABRAHAM [2008] - All rights reserved

Radioactive Decay

The Atom releases excess energies in various forms as it seeks a lower energy, state of equilibria with its surrounding EM energy environment

Energy in all its forms seeks equilibirium

Photons, Heat, Vibration & Kinetics can all raise the energy levels of atomic nuclei

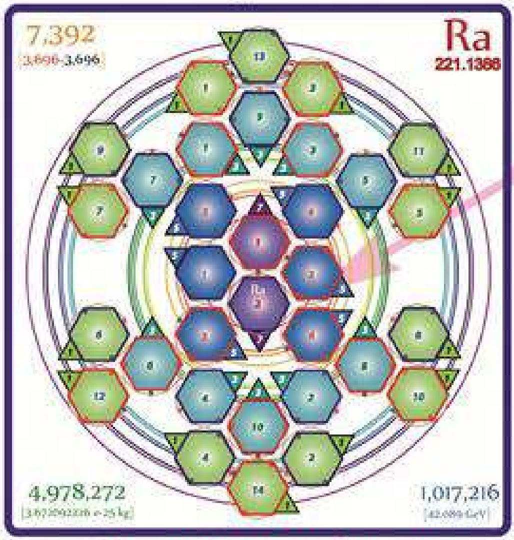
7,392
[3,696-3,696]
Ra
221.1388
4,978,272
1,017,216

stable core electron configurations

The spectral line emissions of photons from bound electrons in addition to Alpha, Beta and Gamma Particle emissions from nuclei can release KEM energies over long time spans

E

scalar equilateral

mass-ENERGY-Matter

geometries

topologies

mc^2

Mc^4

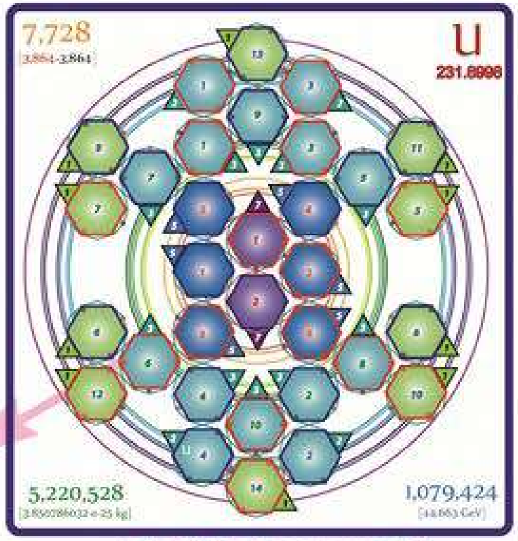
7,728
[3,864-3,864]
U
231.8998
5,220,528
1,079,424

unstable high energy nuclei

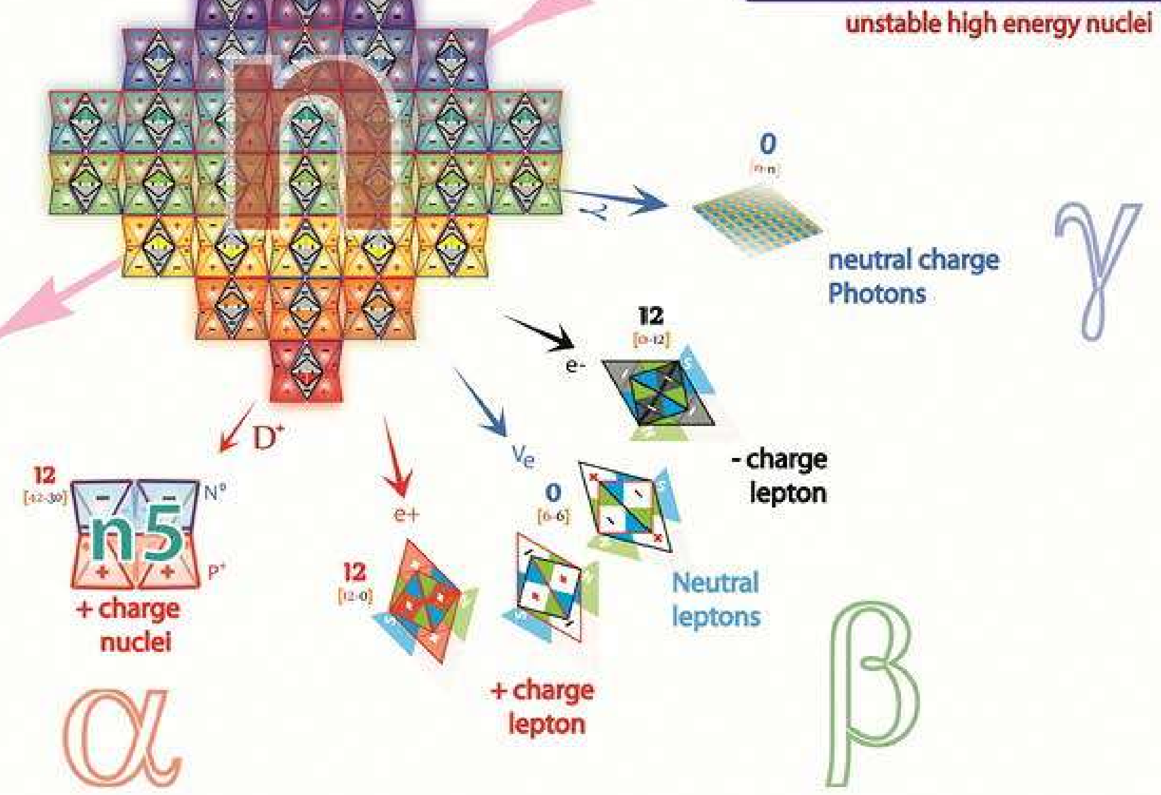
n
0
[n-n]
neutral charge
Photons
γ
12
[0-12]
e-
- charge
lepton
D+
12
[12-30]
N0
n5
P+
+ charge
nuclei
α
e+
12
[12-0]
νe
0
[6-6]
+ charge
lepton
Neutral
leptons
β

Copyright ABRAHAM [2008] - All rights reserved

Radio-Carbon isotope dating

Radiocarbon dating (sometimes simply known as carbon dating) is a radiometric dating method that uses the naturally occurring radioisotope carbon-14 (14C) to estimate the age of carbon-bearing materials up to about 58,000 to 62,000 years

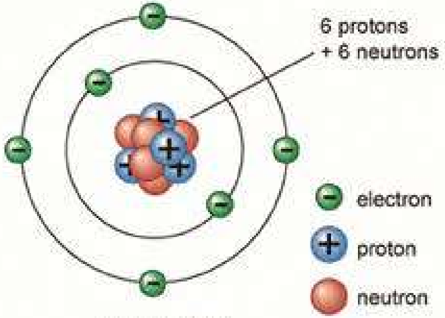
6 protons
+ 6 neutrons
electron
proton
neutron

Carbon atom

Standing-wave KE not 'excess' Neutrons

It is the stored Kinetic Energies [KE] that increase the mass of Carbon12 to create its isotopes, historically these mass-energies have been mistakenly thought to be extra Neutrons in the nuclei

C14 half-life
5730 ± 40

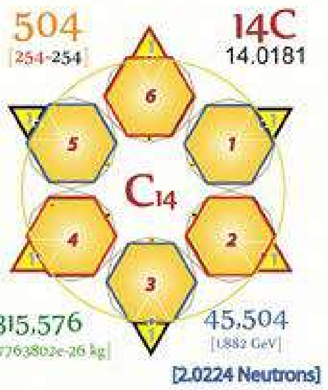
504
[254-254]
14C
14.0181
6
5
1
C14
4
2
3
315,576
[2.327763802e-26 kg]
45,504
[1.882 GeV]
[2.0224 Neutrons]

Carbon-14, 14C, or radiocarbon,
is a radioactive isotope of carbon
with a nucleus containing
6 protons, 6 neutrons and 6 electrons
with an excess energy content
of 1.88 GeV

[6 Deuterium nuclei]

ALL elemental Carbon atoms have an atomic configuration of 6 protons, 6 neutrons and 6 electrons

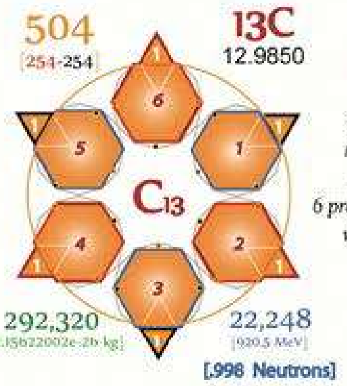
504
[254-254]
13C
12.9850
6
5
1
C13
4
2
3
292,320
[2.156622002e-26 kg]
22,248
[920.5 MeV]
[.998 Neutrons]

Carbon-13, 13C, is a weakly
radioactive isotope of carbon
with a nucleus containing
6 protons, 6 neutrons and 6 electrons
with an excess energy content
of 920 MeV

Radio-Carbon dating can be used to determine the age of carbonaceous materials up to about 60,000 years old

Isotopic mass-charge quotient determines an element's properties

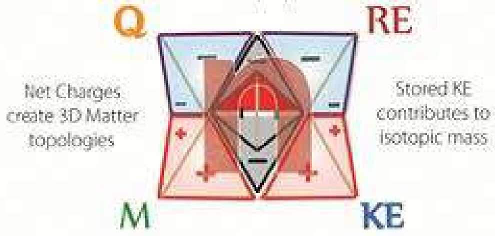

Net Charges create 3D Matter topologies

Stored KE contributes to isotopic mass

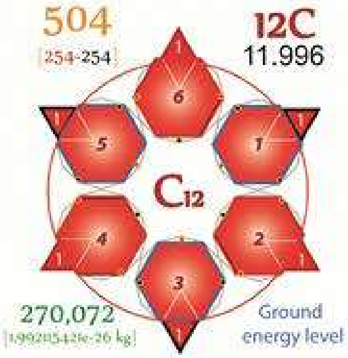
504
[254-254]
12C
11.996
6
5
1
C12
4
2
3
270,072
[1.992115421e-26 kg]
Ground
energy level

Tetryonic theory corrects historical errors of atomic geometries and nuclear energies

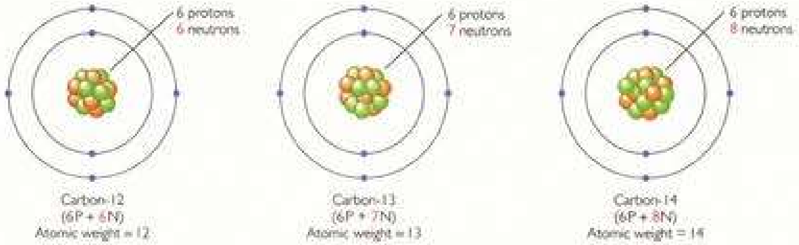
6 protons
6 neutrons
Carbon-12
(6P + 6N)
Atomic weight = 12
6 protons
7 neutrons
Carbon-13
(6P + 7N)
Atomic weight = 13
6 protons
8 neutrons
Carbon-14
(6P + 8N)
Atomic weight = 14

The stored KE [chemical] mass-energies of the Carbon 12+ isotopes comes from various radiological [and biological] processes and its steady, predictable release through the quantum mechanics of the synchronous convertors that make up Carbon nuclei make it useful in dating objects based on their decay rates

Radioactive Isotopes

Where the elementary nuclei's Deuteron energy levels are raised from their ground levels radioactive isotopes are created

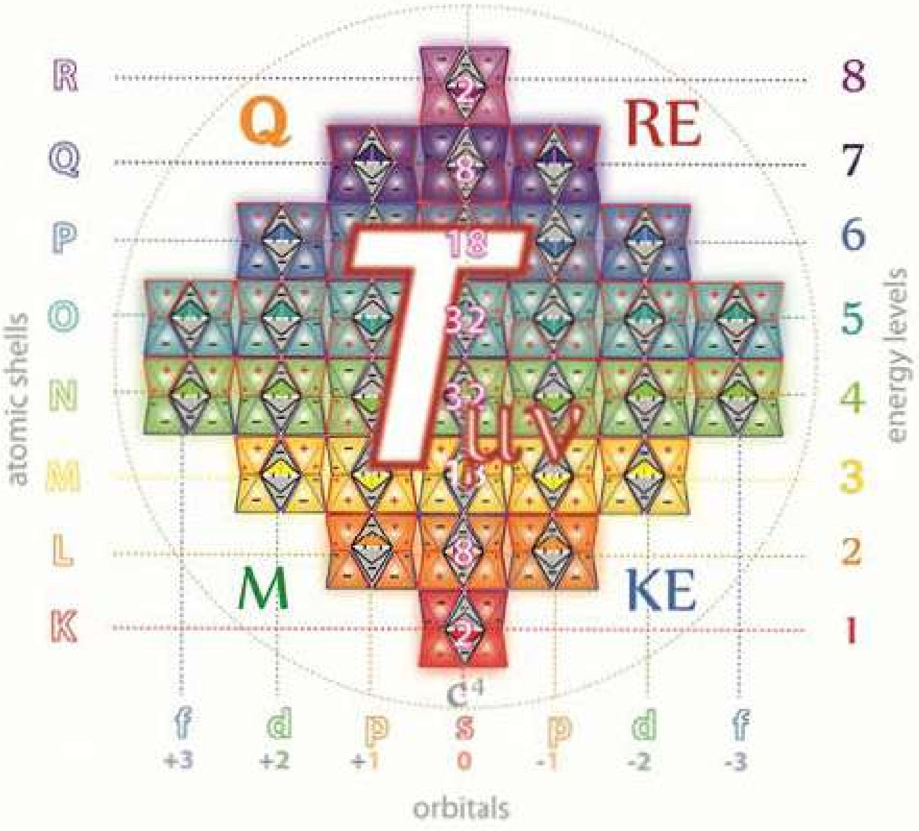
R
Q
P
O
N
M
L
K
atomic shells
8
7
6
5
4
3
2
1
energy levels
Q
RE
M
KE
2
8
18
32
32
18
8
2
$T_{\mu\nu}$
c^4
f
d
p
s
p
d
f
+3
+2
+1
0
-1
-2
-3
orbitals

Each periodic element is comprised of an EQUAL number of Protons, Neutrons & electrons that form each element's unique 2D mass-energy geometries & 3D Matter topology and contribute to its observed properties

The mass ~ energy content of Matter

c^2 c^4

Einstein's relativistic stress energy tensor models mass-energy-Matter as a nebulous energy-momenta density-pressure gradient

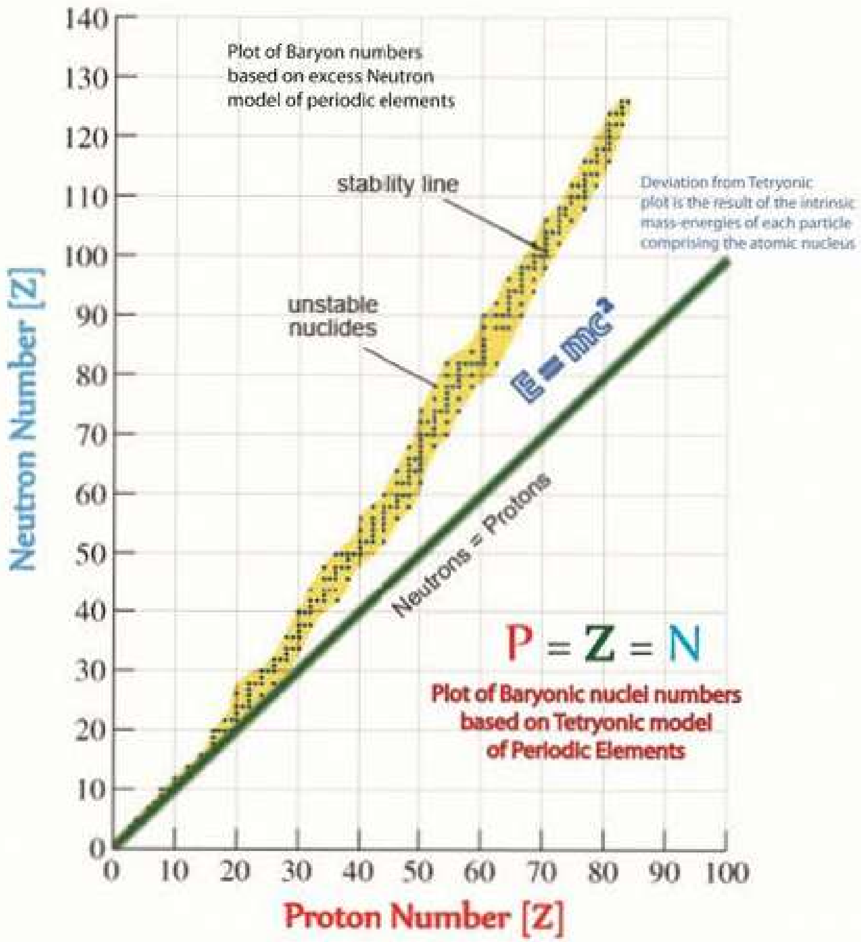

Tetryonic theory redefines the relativistic stress energy tensor [$T_{\mu\nu}$] into a geometric measure of the charged 2D electromagnetic mass-energies & 3D Matter topologies within any spatial co-ordinate system

Nuclear Decay processes

is the set of processes by which an unstable atomic nucleus emits subatomic particles

Charged Matter equivalence

Neutrogen $\underset{[6-6]}{\mathrm{ve}} + \underset{[18-18]}{N^0} \overset{[24-24]}{\sim} \underset{[24-12]}{P^+} + \underset{[0-12]}{e^-}$ **Hydrogen**

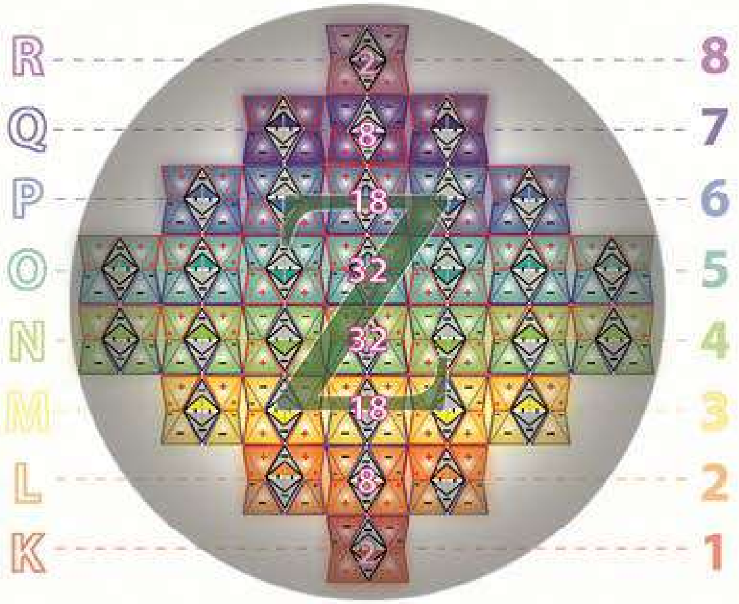

Charge & mass-energy momenta are conservative physical properties

Radioactivity was discovered in 1896 by the French scientist Henri Becquerel, while working on phosphorescent materials. These materials glow in the dark after exposure to light, and he suspected that the glow produced in cathode ray tubes by X-rays might be associated with phosphorescence

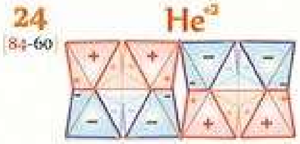

alpha-particles

$$\begin{matrix} \underset{[42-30]}{D^+} & & \\ & \longrightarrow & \underset{[84-60]}{He^{+2}} \\ \underset{[42-30]}{D^+} & & \end{matrix}$$

alpha particle production

E

Energy can be released from atomic nuclei through various processes

alpha-particles
beta-particles
gamma rays
spectral lines
heat & motion

positrons

beta-particles

neutrinos
electrons

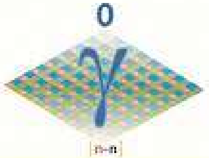

gamma rays

$$\underset{[v-v]}{KE} \longrightarrow \underset{[n-n]}{\gamma}$$

gamma ray production

m *All nuclear decay particles are determined by Tetryonic charge topologies* **M**

Charge Matter particle creation follows chemical equilibrium formulae

$$\underset{[18-18]}{N^0} \longrightarrow \underset{[12-0]}{e^+} + \underset{[6-6]}{\mathrm{ve}} + \underset{[0-12]}{e^-}$$

Neutron decay processes

$$\underset{[18-18]}{N^0} \longrightarrow \underset{[6-6]}{\mathrm{ve}} + \underset{[6-6]}{\mathrm{ve}} + \underset{[6-6]}{\mathrm{ve}}$$

3D Matter topologies are not conservative

Alpha radiation

(Deuteron emissions)

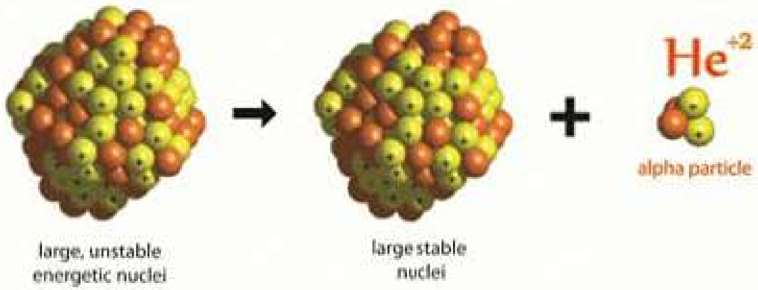
He+2
alpha particle
large, unstable energetic nuclei
large stable nuclei

Alpha particles are released during alpha decay processes in ultra-heavy nuclei like uranium, thorium, actinium, and radium

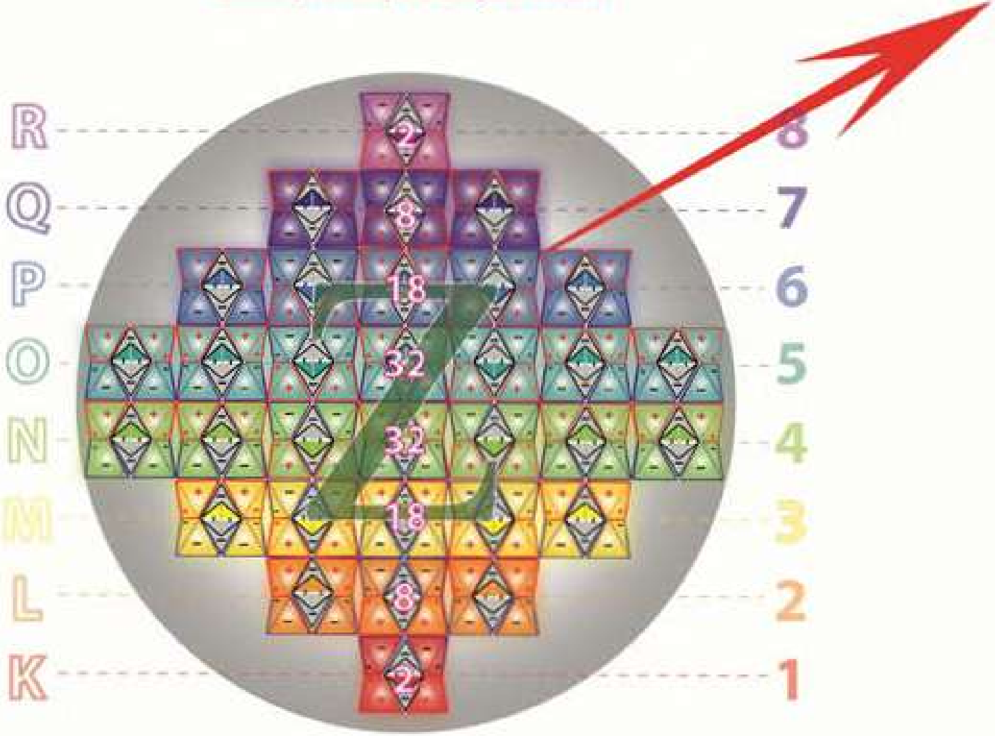
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z

Beta radiation - Negative charge sets (Nuclei)
Gamma radiation - Neutral Charge sets (Photons)

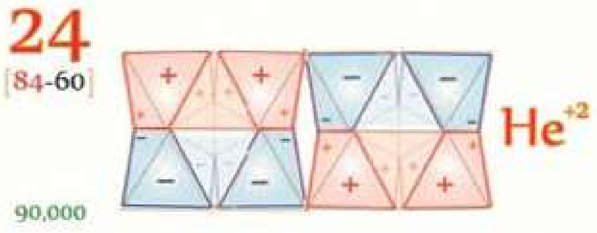
24
[84-60]
90,000
He+2

Helium nuclei creation

D+ [42-30] D+ [42-30]
+
[0-12] e⁻ [0-12] e⁻
→ He⁰ [84-84]

charged alpha nuclei bond & capture electrons to form neutral Helium

alpha particles

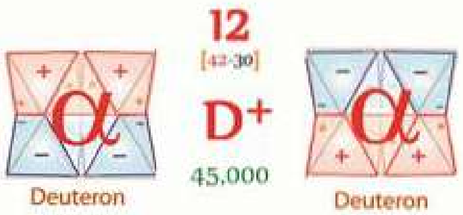
12
[42-30]
D+
45,000
Deuteron
Deuteron

The radioactive isotope Americium 241 emits alpha particles which are used in smoke detectors.

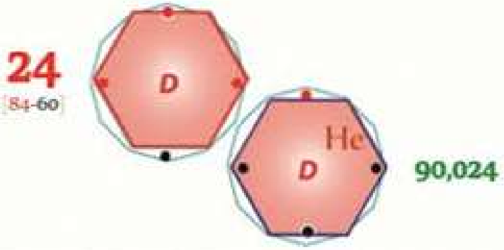
24
[84-60]
D
D
He
90,024

Due to their larger masses compared to beta and gamma particles, alpha particles can be easily stopped by a piece of paper or human skin.

Beta particle decay

$$\underset{[18\text{-}18]}{N^0} \rightarrow \underset{[12\text{-}0]}{e^+} + \underset{[6\text{-}6]}{\nu e} + \underset{[0\text{-}12]}{e^-}$$

Neutron decay paths

$$\underset{[18\text{-}18]}{N^0} \rightarrow \underset{[6\text{-}6]}{\nu e} + \underset{[6\text{-}6]}{\nu e} + \underset{[6\text{-}6]}{\nu e}$$

(the emission of leptons from atomic nuclei)

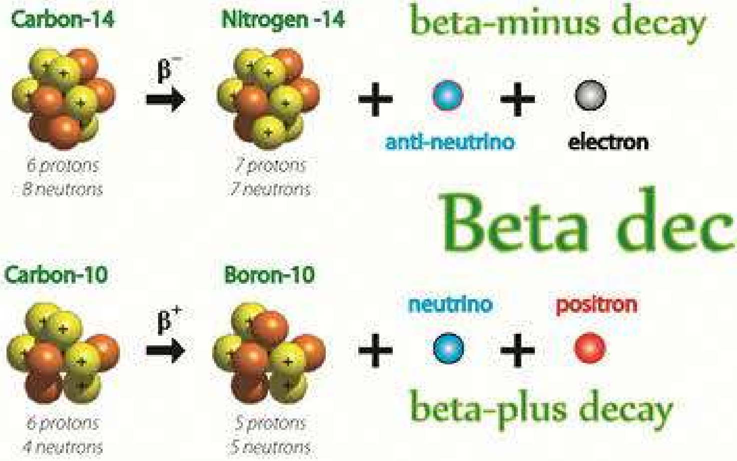
Carbon-14
Nitrogen -14
beta-minus decay
β^-
anti-neutrino
electron
6 protons
8 neutrons
7 protons
7 neutrons
Carbon-10
Boron-10
β^+
neutrino
positron
6 protons
4 neutrons
5 protons
5 neutrons
beta-plus decay

Beta decay

Charge is a conserved physical property of all material objects

$$\underset{[18\text{-}18]}{N^0} + \underset{[6\text{-}6]}{\nu e} \xrightarrow{[24\text{-}24]} \underset{[24\text{-}12]}{P^+} + \underset{[0\text{-}12]}{e^-}$$

Neutron decay is not a spontaneous process

All beta decay processes are the result of neutrino interaction with Neutrons

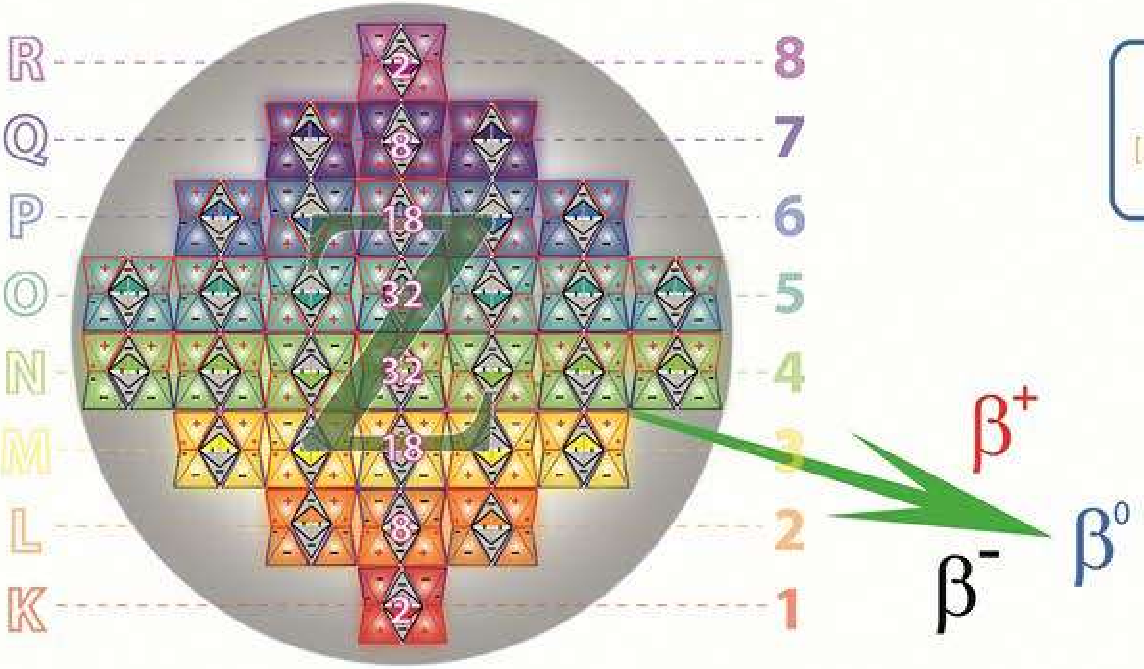
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
β^+
β^-
β^0

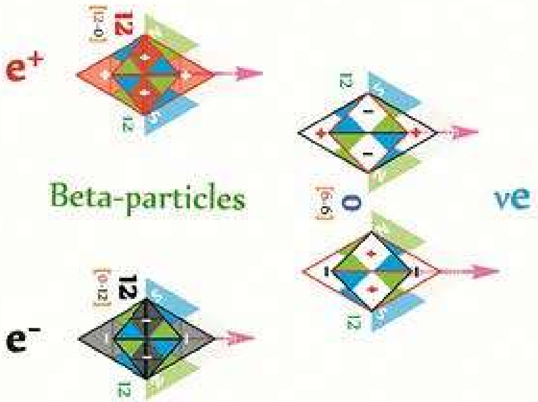
e^+
12
[12-0]
Beta-particles
0
[6-6]
ve
e^-
12
[0-12]

Alpha radiation - Positive charge particles (Nuclei)
Gamma radiation - Neutral charge quanta (Photons)

Copyright ABRAHAM [2008] - All rights reserved

Gamma Radiation

(high energy photon emission)

Natural sources of gamma rays on Earth include gamma decay from naturally occurring radioisotopes, and secondary radiation from atmospheric interactions with cosmic ray particles.

$\lambda[f]$ *All ejected gamma ray photons are neutral energy quanta sets [photons / EM waves]* $[f]\lambda$

Gamma rays typically have frequencies above 10 exahertz (or >10^19 Hz), and therefore generally have energies above 100 keV and wavelengths less than 10 picometers (less than the diameter of an atom)

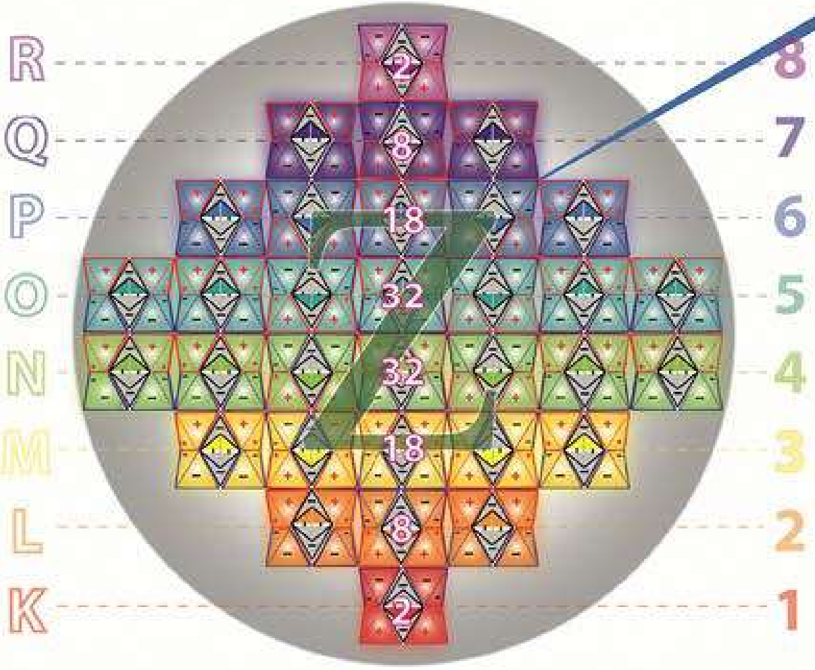

Alpha radiation - Positive charge particles (nuclei)
Beta radiation - Negative charge particles (leptons)

gamma rays

Gamma decay produces rays with energies of only a few hundred keV, and almost always less than 10MeV

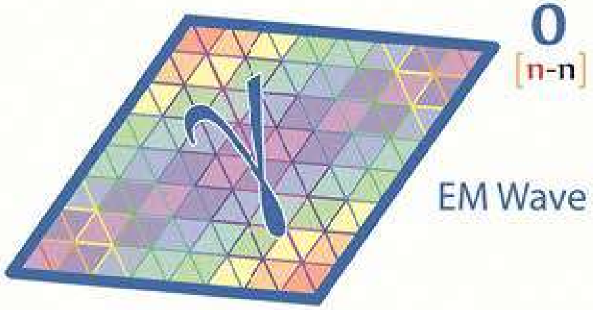

EM Field · Planck quanta

$$2\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Photons · ElectroMagnetic · mass · velocity

Gamma rays are a form of ionizing radiation, and they have very good penetrating power. They result from the release of atomic energies and will cause biological damage to living tissue.

$$2hv = E = hf$$

Care must always be taken to distingush between charged Planck quanta [v] and photons [f]

gamma ray bursts

Baryon mass-Matter topology

N^0 [18-18] → [2.25 e23] [v-v]

$$E = 2mc^2 = hf$$

930 MeV · 930 MeV

gamma ray mass-energy geometry

are the result of the stored KEM mass-energies of Matter topologies being released as mass-energy momenta geometries

Copyright ABRAHAM [2008] - All rights reserved

neutrino-Neutron Interactions

Neutrogen

$$\nu e_{[6\text{-}6]} + N^0_{[18\text{-}18]} \overset{[24\text{-}24]}{\sim} P^+_{[24\text{-}12]} + e^-_{[0\text{-}12]}$$

Charged Matter equivalence

Hydrogen

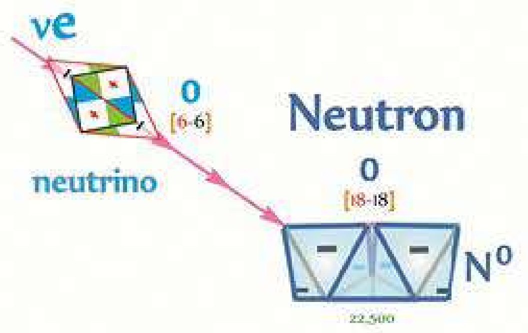

neutrino capture

Neutron decay is not the spontaneous process hypothesied by modern nuclear physicists

The observed decay products are the result of solar neutrinos interacting with Neutrons (both within nuclei and free)

the neutrinos also have an equally likely probability of interacting with Protons in the nuclei

[24-24]

A Neutronium atom has an identical mass-charge quotient to that of Hydrogen

[24-24]

Neutronium decay processes

A Neutron/neutrino interaction can decay into 4 differing particle sets

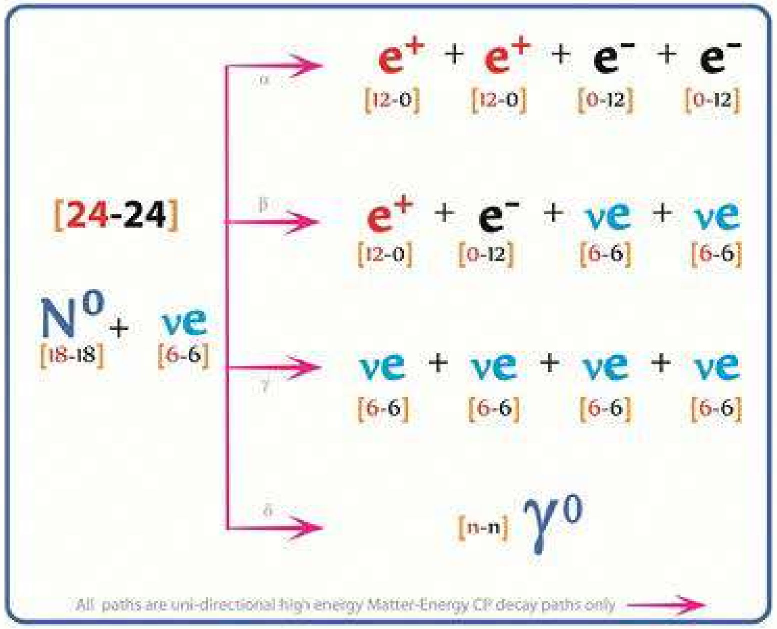

[24-24]

$N^0_{[18\text{-}18]} + \nu e_{[6\text{-}6]}$

α $\rightarrow$ $e^+_{[12\text{-}0]} + e^+_{[12\text{-}0]} + e^-_{[0\text{-}12]} + e^-_{[0\text{-}12]}$

β $\rightarrow$ $e^+_{[12\text{-}0]} + e^-_{[0\text{-}12]} + \nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]}$

γ $\rightarrow$ $\nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]}$

δ $\rightarrow$ $[n\text{-}n]\ \gamma^0$

All paths are uni-directional high energy Matter-Energy CP decay paths only →

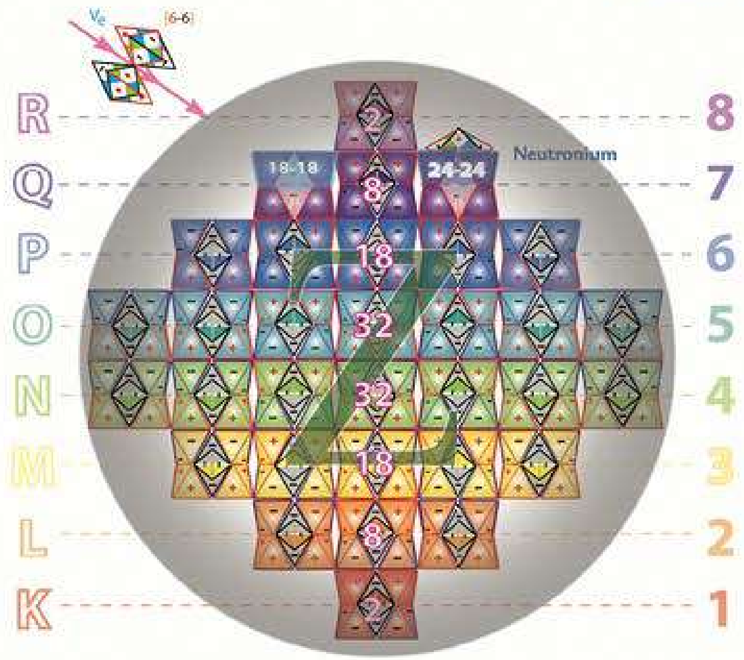

Excluding their nett Charge,
Neutrons have neutral particle geometries [18-18]
identical to that of a Proton [24-12]

$$N^0_{[18\text{-}18]} \neq P^+_{[24\text{-}12]} + e^-_{[0\text{-}12]}$$

[24-24]

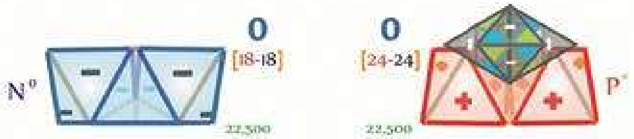

Spontaneous Neutron decay into Proton-electron-neutrinos is not possible without the interaction of Muonic neutrinos (see Tetryonic Charge numbers)

A Neutron is NOT a Proton that has absorbed an electron

Copyright ABRAHAM [2008] - All rights reserved

neutrino-Proton Interactions

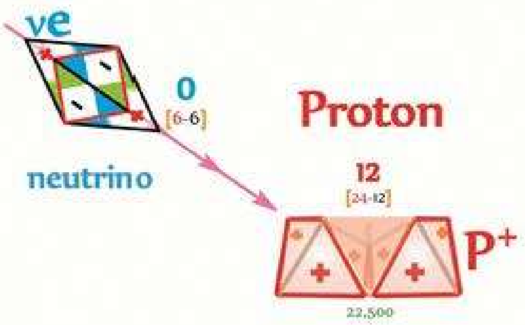

neutrino capture

Neutron decay is not the spontaneous process hypothesied by modern nuclear physicists

The observed decay products are the result of solar neutrinos interacting with Neutrons (both within nuclei and free)

the neutrinos also have an equally likely probability of interacting with Neutrons in the nuclei

[30-18]

A Proton-neutrino atom would have an identical charge quotient to that of a Proton

[24-12]

Although very weakly interacting, neutrinos can still be attracted to and bind to Protons via their negative charge fascia

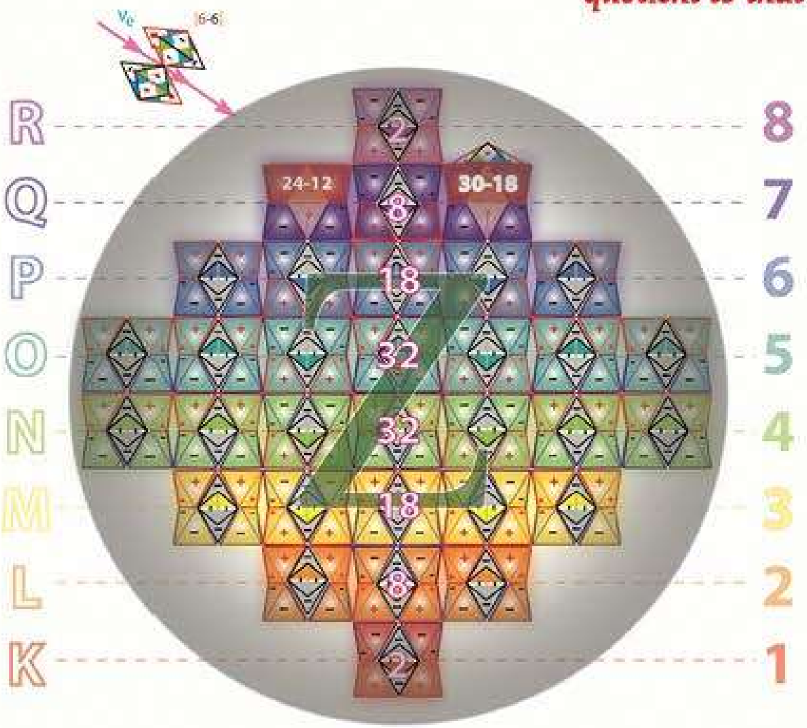

Proton decay paths

A Proton/Neutrino interaction can decay into 4 differing particle sets

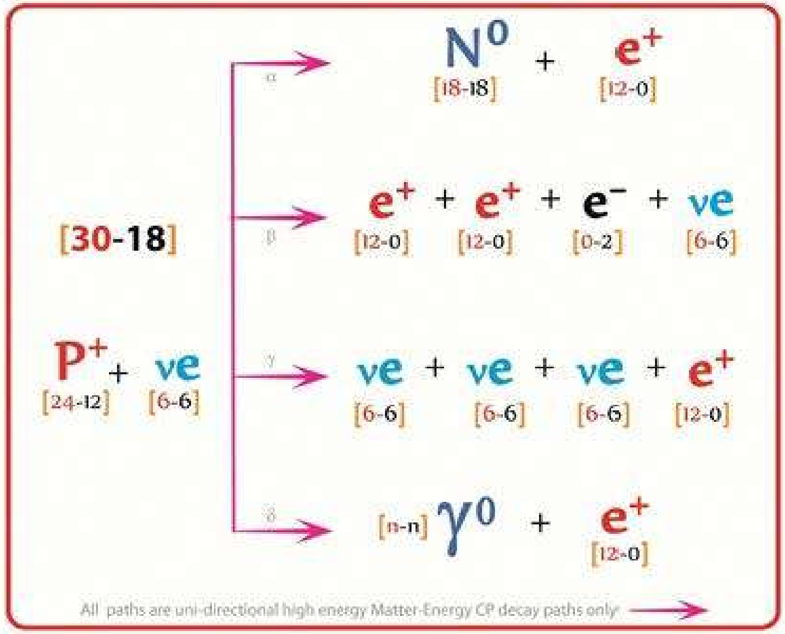

[30-18]

P+ [24-12] + ve [6-6]

α → N0 [18-18] + e+ [12-0]

β → e+ [12-0] + e+ [12-0] + e− [0-2] + ve [6-6]

γ → ve [6-6] + ve [6-6] + ve [6-6] + e+ [12-0]

δ → [n-n] γ0 + e+ [12-0]

All paths are uni-directional high energy Matter-Energy CP decay paths only →

If they exist, Proton-neutrino particle couplings would function in a manner identical to that of Proton-electron couplings and could be detected by the anomalous spin measurements that would result

A Proton has a Positive charge geometry [24-12]
equivalent to that of a Neutral Neutron [18-18]

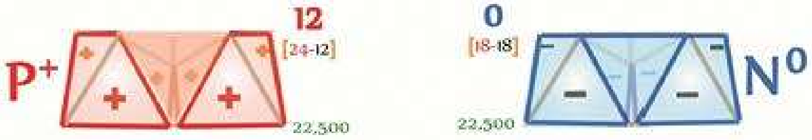

A Proton is topologically identical to a Neutron (differing only in the net charge created)

Quantum Batteries

Atomic nuclei can be easily scaled to non-quantum sizes to offer clean, safe and portable long term Energy storage devices that can store energy indefinitely and release it on demand anywhere in the World

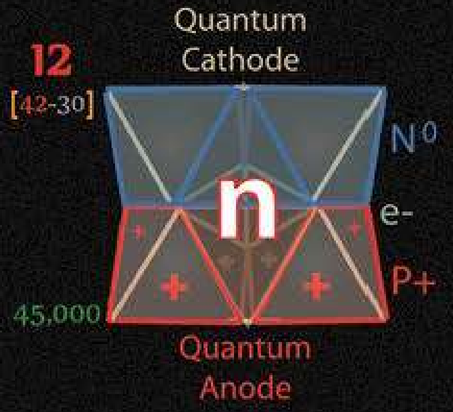
12
[42-30]
Quantum
Cathode
N0
n
e-
P+
45,000
Quantum
Anode

The quantum battery is unique in that in addition to storing energy indefinitely, when an electron binds to the Deuteron nuclei it has the ability to release specific energies [photons] by way of its quantum-scale synchronous converter topologies

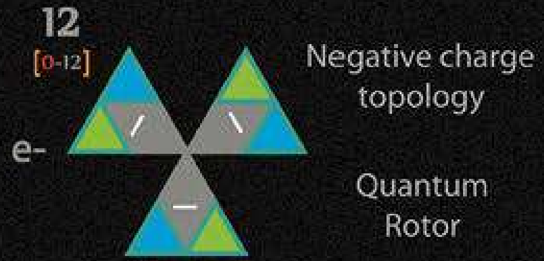
12
[0-12]
e-
Negative charge
topology
Quantum
Rotor

12 loop quantum inductive rotor

(Atomic Nuclei)

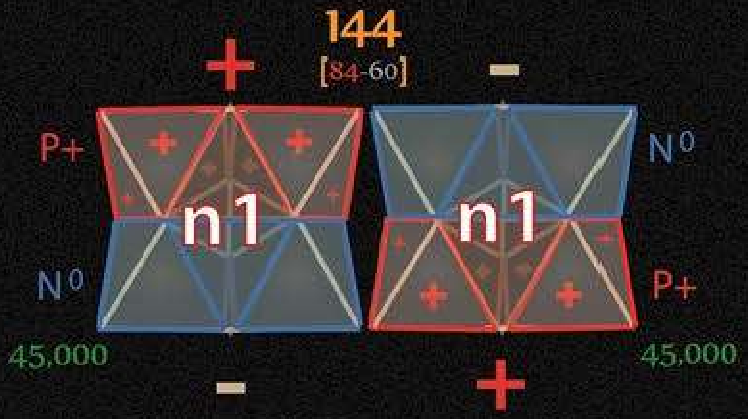
144
[84-60]
+
-
P+
N0
n1
n1
N0
P+
45,000
45,000
-
+

anti-Parallel Configuration

+|E90
03-2013
Energizer

Synchronous quantum converter topologies can be connected in parallel or series to meet varying power requirements anywhere in the World and provides for the safe storage of nuclear energy as mass

Series Configuration

+|E90
03-2013
Energizer

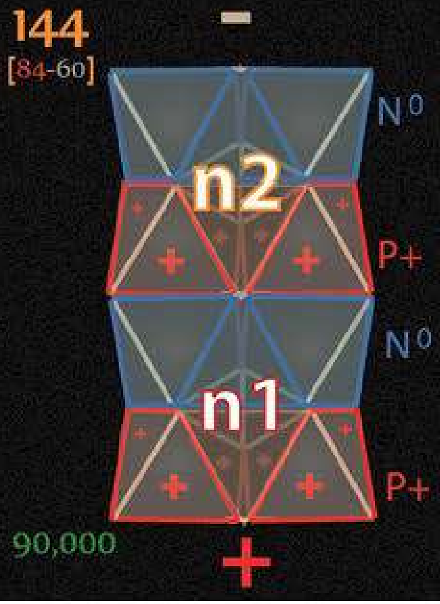
144
[84-60]
-
N0
n2
P+
N0
n1
P+
90,000
+

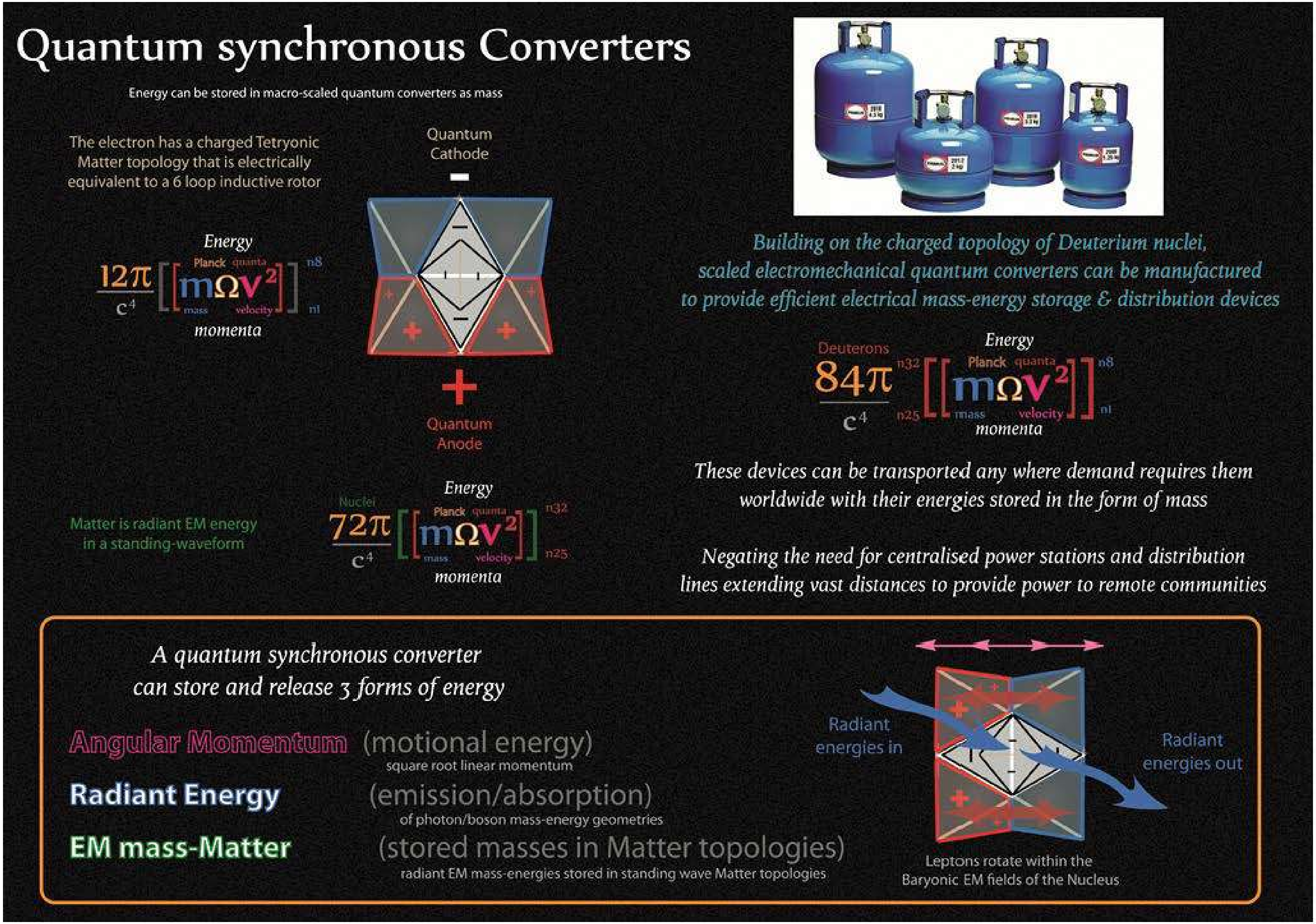

Quantum synchronous Converters

Energy can be stored in macro-scaled quantum converters as mass

The electron has a charged Tetryonic Matter topology that is electrically equivalent to a 6 loop inductive rotor

Energy

$$\frac{12\pi}{c^4}\left[\left[m\Omega v^2\right]\right]^{n8}_{n1}$$

Planck quanta mass velocity

momenta

Quantum Cathode

Quantum Anode

Matter is radiant EM energy in a standing-waveform

Energy

Nuclei

$$\frac{72\pi}{c^4}\left[\left[m\Omega v^2\right]\right]^{n32}_{n25}$$

Planck quanta mass velocity

momenta

Building on the charged topology of Deuterium nuclei, scaled electromechanical quantum converters can be manufactured to provide efficient electrical mass-energy storage & distribution devices

Energy

Deuterons

$$\frac{84\pi}{c^4}\left[^{n32}_{n25}\left[\left[m\Omega v^2\right]\right]^{n8}_{n1}\right]$$

Planck quanta mass velocity

momenta

These devices can be transported any where demand requires them worldwide with their energies stored in the form of mass

Negating the need for centralised power stations and distribution lines extending vast distances to provide power to remote communities

A quantum synchronous converter can store and release 3 forms of energy

Angular Momentum (motional energy)
square root linear momentum

Radiant Energy (emission/absorption)
of photon/boson mass-energy geometries

EM mass-Matter (stored masses in Matter topologies)
radiant EM mass-energies stored in standing wave Matter topologies

Leptons rotate within the Baryonic EM fields of the Nucleus

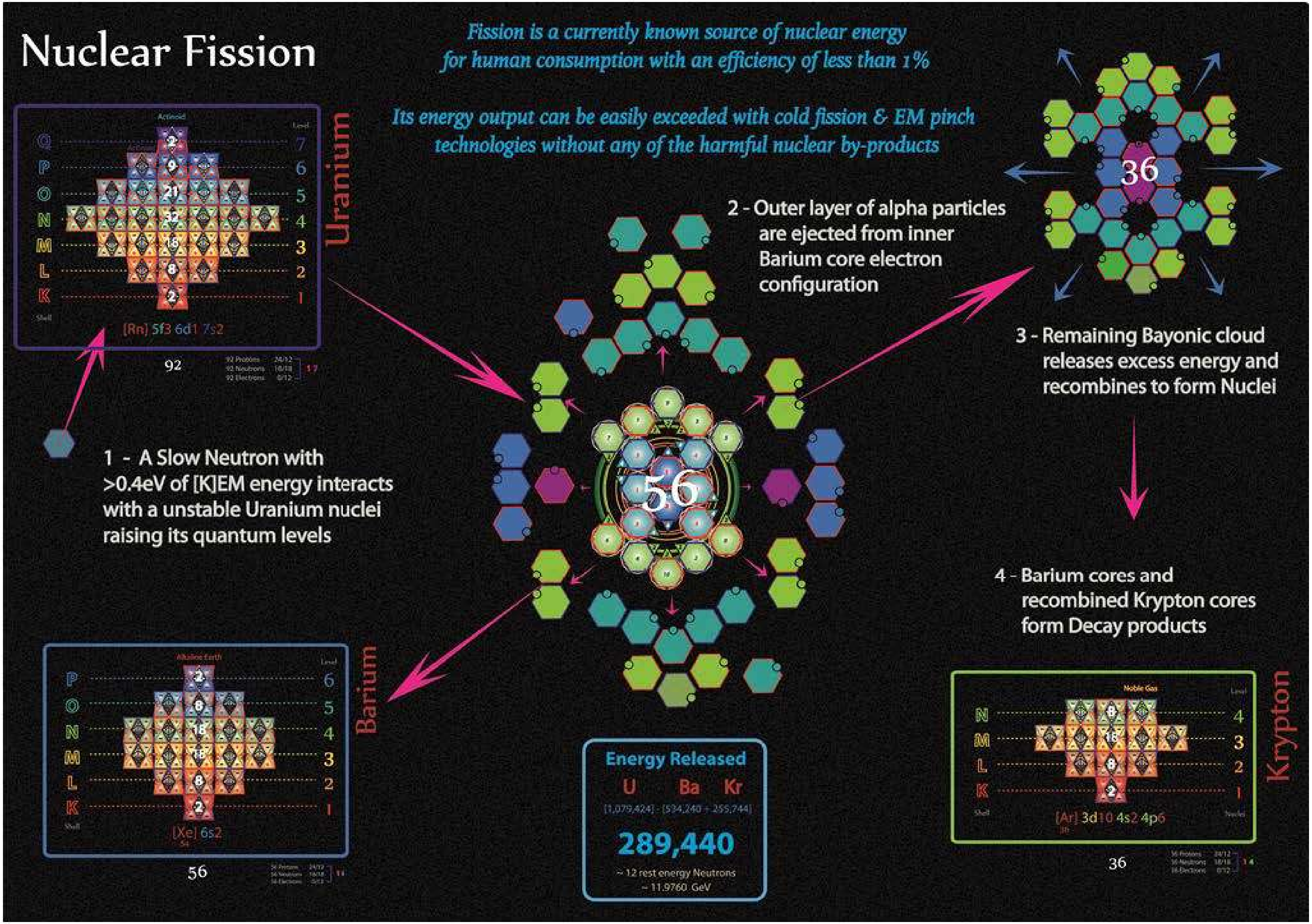

Nuclear Fission

Fission is a currently known source of nuclear energy for human consumption with an efficiency of less than 1%

Its energy output can be easily exceeded with cold fission & EM pinch technologies without any of the harmful nuclear by-products

1 - A Slow Neutron with >0.4eV of [K]EM energy interacts with a unstable Uranium nuclei raising its quantum levels

2 - Outer layer of alpha particles are ejected from inner Barium core electron configuration

3 - Remaining Bayonic cloud releases excess energy and recombines to form Nuclei

4 - Barium cores and recombined Krypton cores form Decay products

Energy Released

U Ba Kr

[1,079,424] - [534,240 + 255,744]

289,440

~ 12 rest energy Neutrons

~ 11.9760 GeV

Uranium
Actinoid
Level
Q
P
O
N
M
L
K
Shell
7
6
5
4
3
2
1
2
9
21
32
18
8
2
[Rn] 5f3 6d1 7s2
92
92 Protons 24/12
92 Neutrons 18/18
92 Electrons 0/12
17
56
36
Barium
Alkaline Earth
P
O
N
M
L
K
6
5
4
3
2
1
2
8
18
18
8
2
[Xe] 6s2
56
Krypton
Noble Gas
N
M
L
K
4
3
2
1
8
18
8
2
[Ar] 3d10 4s2 4p6
36
14

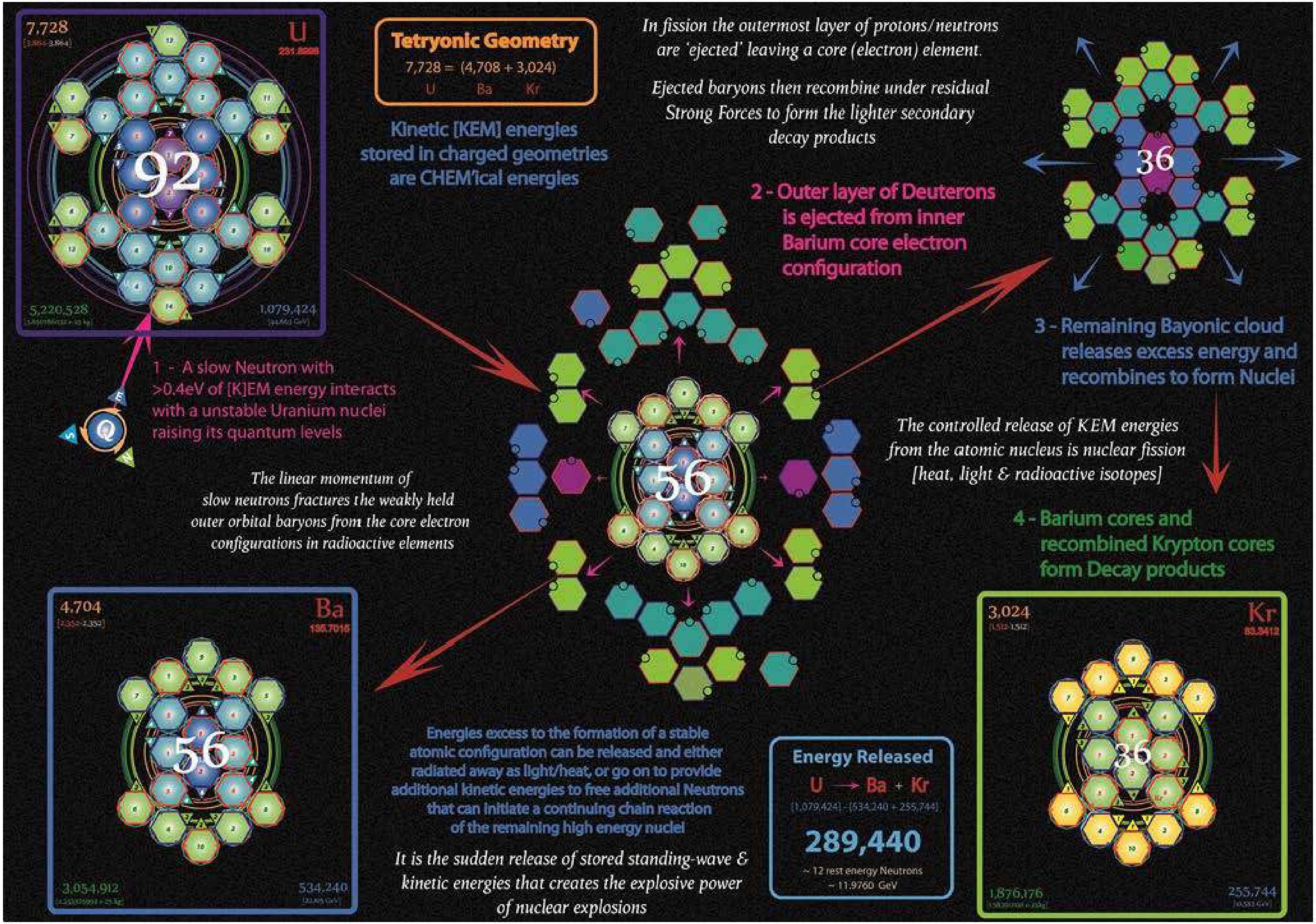

Tetryonic Geometry
7,728 = (4,708 + 3,024)
U Ba Kr

Kinetic [KEM] energies stored in charged geometries are CHEM'ical energies

In fission the outermost layer of protons/neutrons are 'ejected' leaving a core (electron) element.

Ejected baryons then recombine under residual Strong Forces to form the lighter secondary decay products

2 - Outer layer of Deuterons is ejected from inner Barium core electron configuration

1 - A slow Neutron with >0.4eV of [K]EM energy interacts with a unstable Uranium nuclei raising its quantum levels

The linear momentum of slow neutrons fractures the weakly held outer orbital baryons from the core electron configurations in radioactive elements

3 - Remaining Bayonic cloud releases excess energy and recombines to form Nuclei

The controlled release of KEM energies from the atomic nucleus is nuclear fission [heat, light & radioactive isotopes]

4 - Barium cores and recombined Krypton cores form Decay products

Energies excess to the formation of a stable atomic configuration can be released and either radiated away as light/heat, or go on to provide additional kinetic energies to free additional Neutrons that can initiate a continuing chain reaction of the remaining high energy nuclei

It is the sudden release of stored standing-wave & kinetic energies that creates the explosive power of nuclear explosions

Energy Released
U → Ba + Kr
(1,079,424) - (534,240 + 255,744)
289,440
~ 12 rest energy Neutrons
~ 11.9760 GeV

Tetryonics 60.13 - Nuclear Fission process

Copyright ABRAHAM [2008] - All rights reserved

Sonoluminescence

Sonoluminescence is the first hint at the energies that can be released from Tetryonic collapse and involves the emission of short bursts of light from imploding bubbles in a liquid when excited by sound.

It is a key step on the road to realising technologies that can provide Humanity with clean, safe, efficient Energy production through the conversion of Matter into various forms of EM radiation

The production of light as a result of the passing of sound waves through a liquid medium.

The sound waves cause the formation of bubbles that emit bright flashes of light when they collapse.

Sonoluminesence is the result of energy releases from the collapse of standing-wave Matter topologies within Fluids

Finding an efficient means of inducing and managing a controlled reaction where 3D Matter topologies are converted into 2D radiant energy is a key step on the path to creating a future energy source of all Humanity

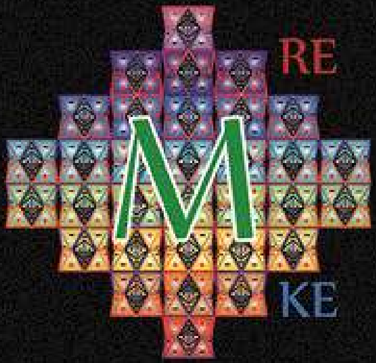

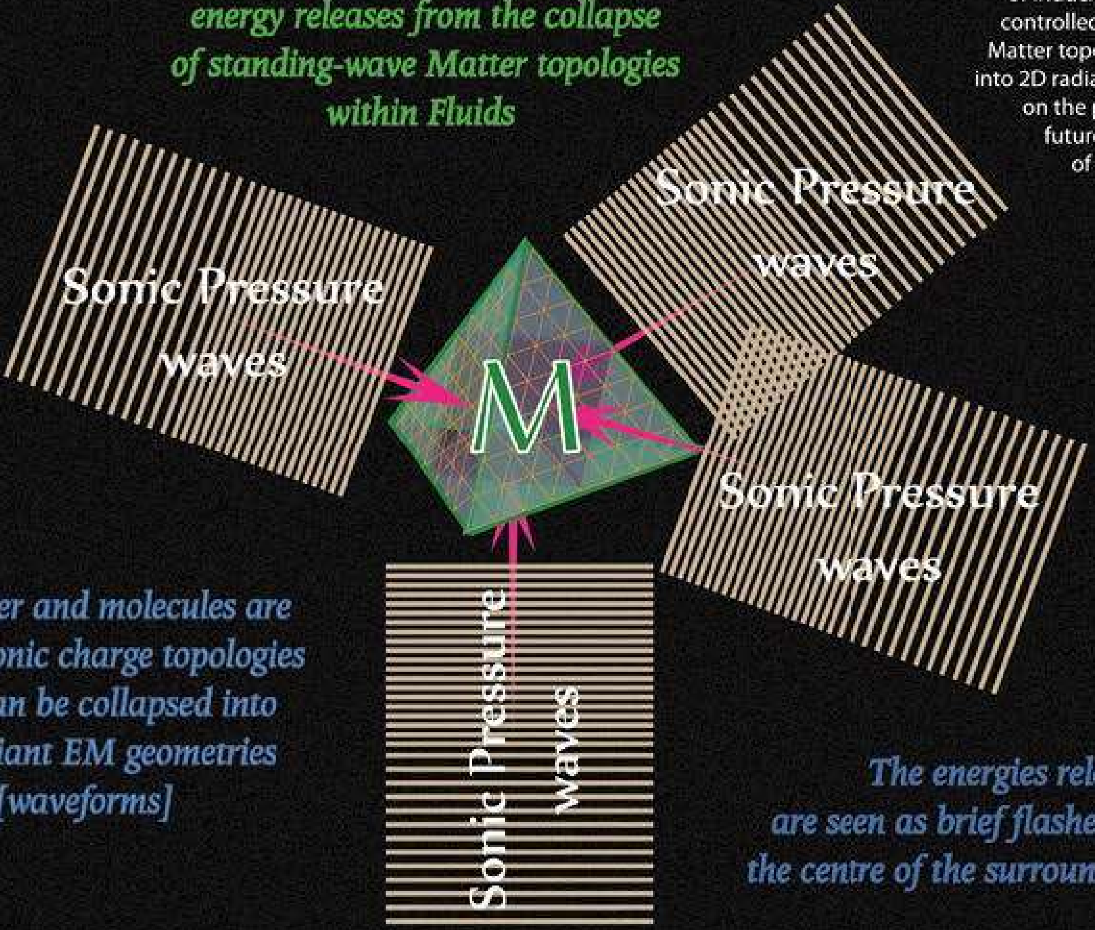

The gauge transformation of Matter into EM radiation is the most efficient form of energy release available to us

All Matter and molecules are 3D Tetryonic charge topologies which can be collapsed into 2D radiant EM geometries [waveforms]

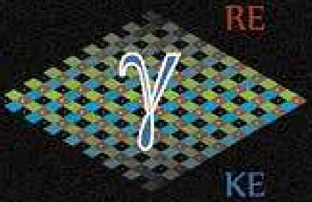

The energies released are seen as brief flashes of light at the centre of the surrounding medium

Stanley Pons Martin Fleischmann

(23 August 1943) (29 March 1927 – 3 August 2012)

'Nuclear fusion of the type postulated would be inconsistent with current understanding and, if verified, would require theory to be extended in an unexpected way'

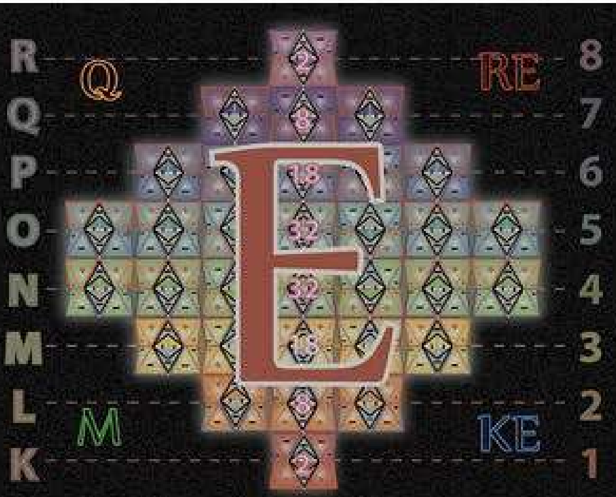
R
Q
P
O
N
M
L
K
Q
RE
M
KE
E
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2

The reported excess energy released is often refused as being impossible as it woud require the release of massive neutron radiation bath.

Cold Fission

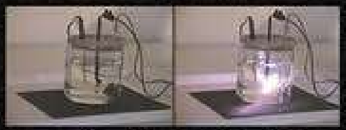

Atomic nuclei can easily release the reported energies without Neutron particle emissions when an accurate model of the atomic nuclei is used

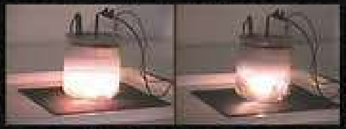

'Fusion' is a fictious nuclear process 'cold fusion' is in fact a form of nuclear fission

Palladium's rest molar mass is 108.7867 with a stored Kinetic energy content of 15.699 Gev [16 neutron equivalence]

Expelling the non-core atomic nuclei results in a release of stored KEM energies as the deuterium nuclei seek to reach a ground energy level state

Krypton's rest molar mass is 83.3412 with a stored Kinetic energy content of 10.582 Gev [11 neutron equivalence]

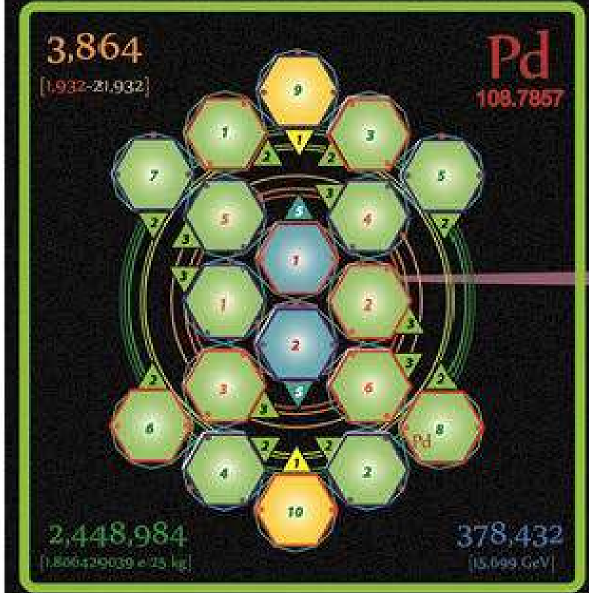
3,864
[1,932-21,932]
Pd
108.7857
9
1
3
7
5
10
2,448,984
[1.806429039 e-25 kg]
378,432
[15.099 GeV]

Palladium

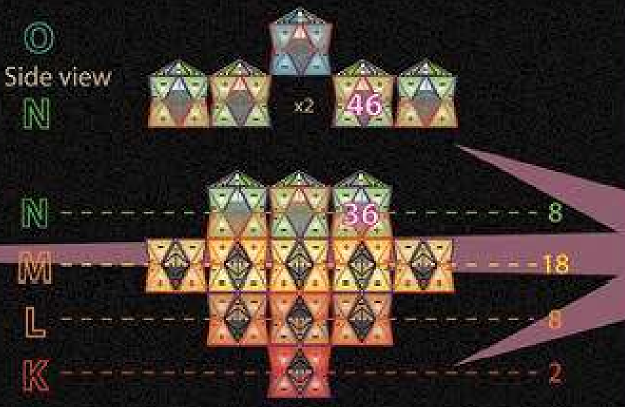
O
Side view
N
x2
46
N
M
L
K
36
8
18
8
2

2 x n5
10 Deuterons
8 x n4
122,688
[5.076 GeV]
Energy equivalent to 5+ Neutrons

NO excess Neutrons are ever released ~ only stored KEM energies and alpha particles

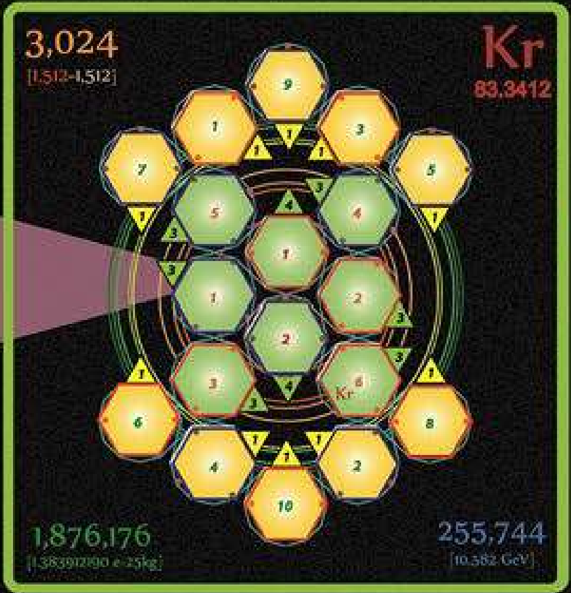
3,024
[1,512-1,512]
Kr
83.3412
9
1
3
7
5
10
1,876,176
[1.38391280 e-25kg]
255,744
[10.582 GeV]

Krypton

[Kr] 4d8 5s2
36
Transition Metal

46 Protons	24-12
46 Neutrons	18-18
46 Electrons	0-12

1 5

[Ar] 3d10 4s2 4p6
36
Noble Gas

36 Protons	24-12
36 Neutrons	18-18
36 Electrons	0-12

1 4

Tetryonics 60.15 - Cold Fission

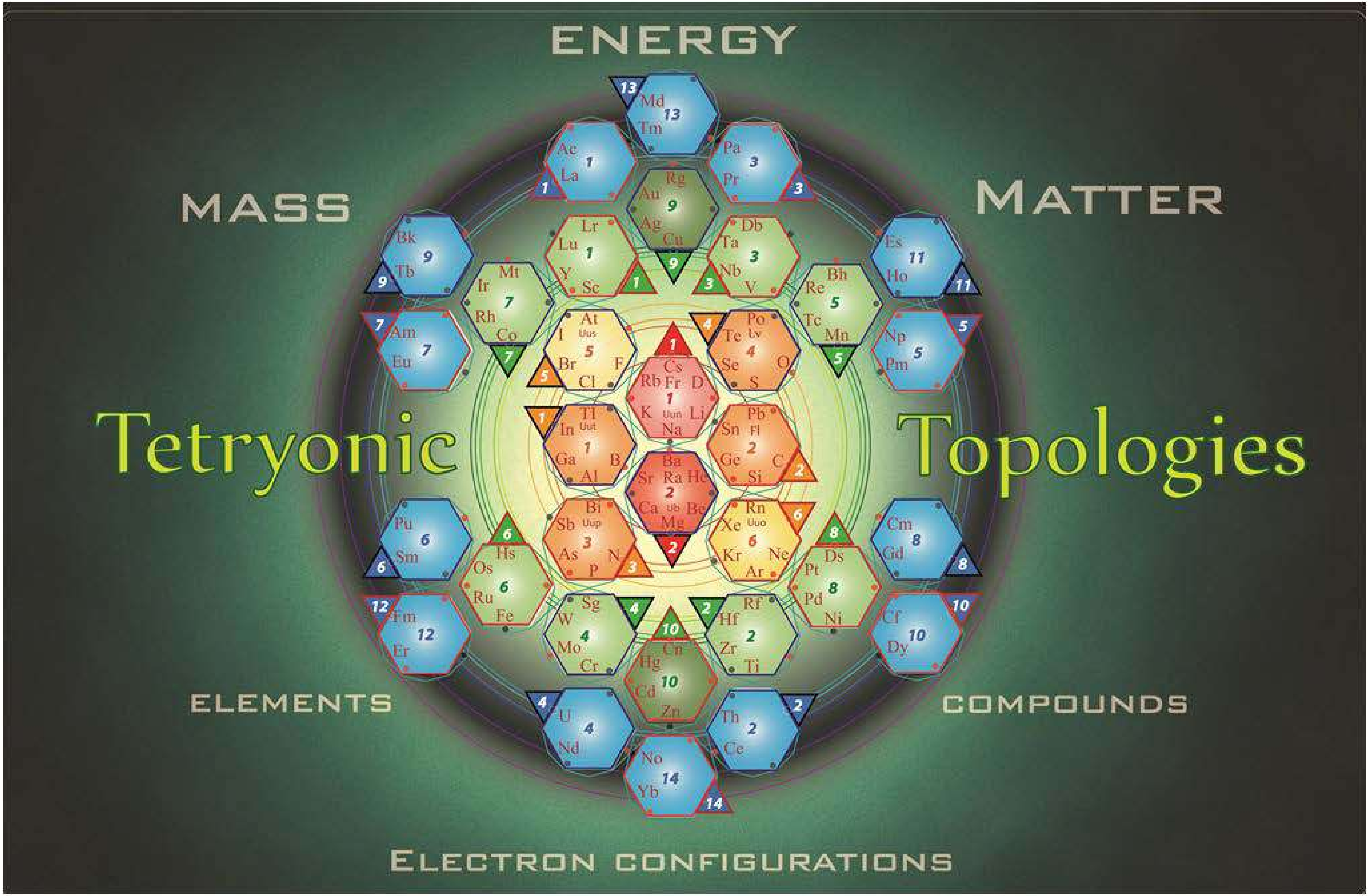
ENERGY
MASS
MATTER
Tetryonic
Topologies
ELEMENTS
COMPOUNDS
ELECTRON CONFIGURATIONS
13 Md Tm 13
1 Ac La 1
3 Pa Pr 3
Rg Au Ag Cu 9
9
Lr Lu Y Sc 1
1
Db Ta Nb V 3
3
9 Bk Tb 9
Mt Ir Rh Co 7
7
Bh Re Tc Mn 5
5
11 Es Ho 11
7 Am Eu 7
At Uus I Br Cl F 5
5
4
Po Lv Te Se S O 4
5 Np Pm 5
1
Cs Rb Fr D K Uun Li Na 1
1
Tl Uut In Ga Al B 1
Pb Fl Sn Ge Si C 2
2
Ba Sr Ra He Ca Ub Be Mg 2
2
6 Pu Sm 6
6
Hs Os Ru Fe 6
Bi Uup Sb As P N 3
3
Rn Uuo Xe Kr Ar Ne 6
6
8
Ds Pt Pd Ni 8
8 Cm Gd 8
12 Fm Er 12
Sg W Mo Cr 4
4
10
Cn Hg Cd Zn 10
2
Rf Hf Zr Ti 2
10 Cf Dy 10
4 U Nd 4
2 Th Ce 2
No Yb 14 14

Tetryonic Chemistry

The charged topology of periodic and compound Matter

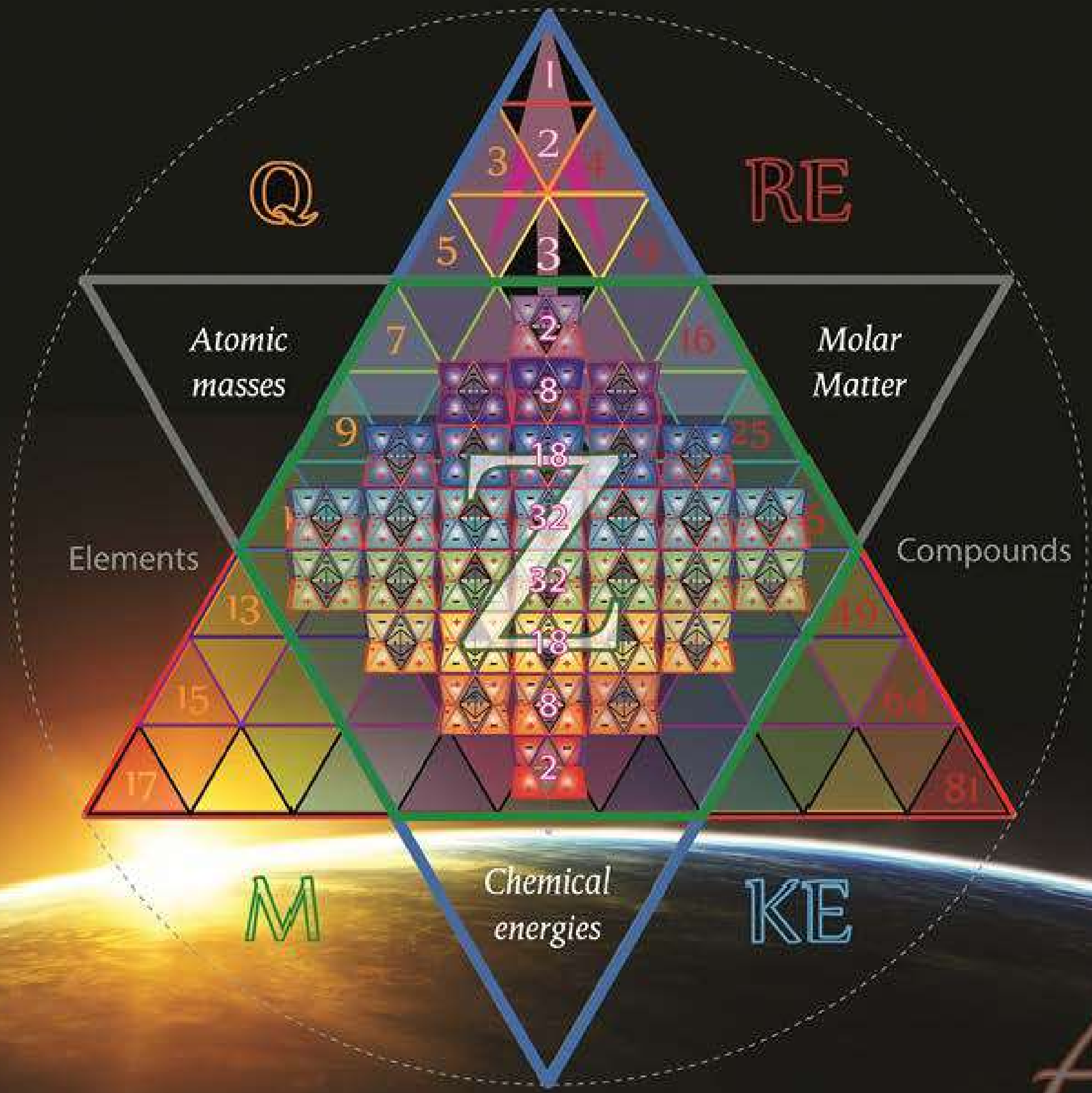
Q
RE
Atomic masses
Molar Matter
Elements
Compounds
M
KE
Chemical energies
Z
1
2
3
4
5
3
7
2
16
8
9
18
25
32
32
13
18
49
15
8
64
17
2
81

ISBN 978-0-987288-3-1
[Second Edition © 2012]

Abraham

Copyright ABRAHAM [2008] - All rights reserved

TETRYONICS

The charged interactions of universal Gravitation

Q
RE
ε
Radiant masses
Standing-wave Matter
Gravitational GEM fields
Interactive KEM fields
G
μ
k
M
KE
Vacuum Energies
1
2
3
4
5
6
7
8
9
0
11
13
15
17
16
25
36
49
64
81

Foundational Quantum Cosmology

Abraham

ISBN 978-0-987288-4-8
[Second Edition © 2012]

Copyright ABRAHAM [2008] - All rights reserved

Laws of Force

$F = ma$

Sir Issac Newton

(25 December 1642 – 20 March 1727)

$w = mg$

Gravity

Newton's Principia formulated the laws of motion and of universal gravitation

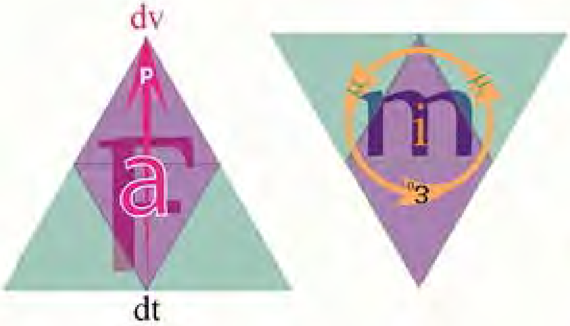

changing Magnetic lines of Force

Michael Faraday

(22 September 1791 – 25 August 1867)

create changing Electric fields of Force

Faraday discovered the principles of electromagnetic induction, diamagnetism, and the laws of electromagnetism.

Electro-Magnetism

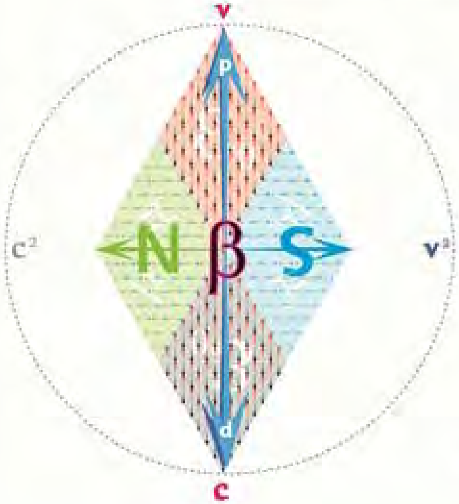

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

Gauss

Light is an electromagnetic wave

Faraday

$$\oint \vec{E}\cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

James Clerk Maxwell

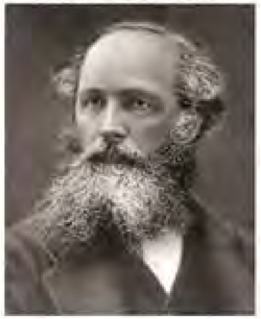

(13 June 1831 – 5 November 1879)

$$\oint \vec{B}\cdot d\vec{A} = 0$$

Gauss

The separate works of Gauss, Ampere & Faraday were unified to create an electromagnetic wave field theory

Ampere

$$\oint \vec{B}\cdot d\vec{s} = \mu_0 i + \frac{1}{c^2}\frac{\partial}{\partial t}\int \vec{E}\cdot d\vec{A}$$

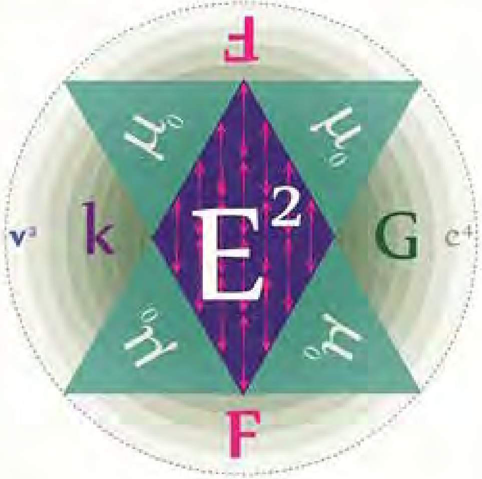

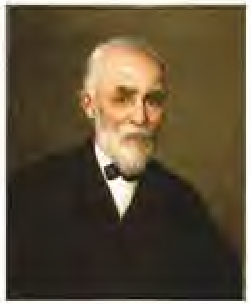

Lorentz Contraction

Lorentz proposed that moving bodies contract in the direction of motion (Lorentz–Fitzgerald contraction)

Hendrik Lorentz

(18 July 1853 – 4 February 1928)

Lorentz Transformations

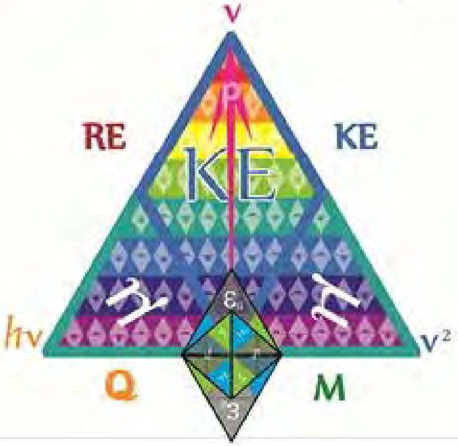

Special Relativity

attempted to explain the constant speed of Light using relative motion & Lorentz contractions

Albert Einstein

(14 March 1879 – 18 April 1955)

&

later attempted to explain gravity as a curvature of spacetime caused by Matter

General Relativity

Copyright ABRAHAM [2008] - All rights reserved

Early physics of Motion

In the 4th century BC, the Greek philosopher Aristotle believed that
there is no effect or motion without a cause

In Book VII of his De Architectura, the Roman engineer and architect Vitruvius
contends that gravity is not dependent on a substance's "weight" but rather on its "nature".

The cause of the downward motion of heavy bodies, such as the element earth, was related to their nature, which caused them to move

In the 7th Century the Indian mathematician Brahmagupta stated
"Bodies fall towards the earth as it is in the nature of the earth to attract bodies,
just as it is in the nature of water to flow."

Greek philosopher Aristotle

(384 BC - 322 BC)

Aristotle defined space to be
the boundary geometry of a body

[x,y,z]
geometry, toplogy

Aristotle defined time to be
a measure of movement

the past Δ the future

'Nature abhors a vacuum'

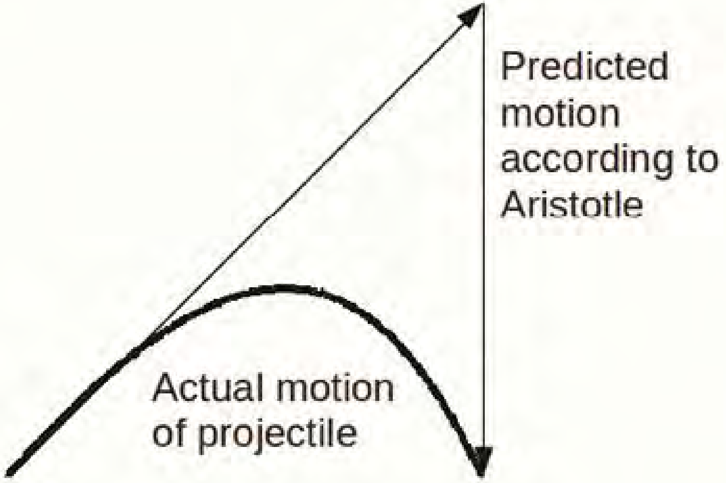
Predicted motion according to Aristotle
Actual motion of projectile

Heavier things fall faster, with speed being proportional to weight.

The speed with which a body falls is inversely proportional
to the density of the medium it is falling through.

Indian Mathematician Brahmagupta

(597-668 AD)

Brahmagupta dedicated a substantial
portion of his work to geometry

height
base

Brahmagupta is considered the first
to formulate the concept of zero

negative numbers 0 positive numbers

Brahmagupta introduced new algebric
methods for solving quadratic equations

Copyright ABRAHAM [2008] - All rights reserved

Galilean Gravity

Galileo Galilei

(15 February 1564 – 8 January 1642)

Galileo claimed to have a Universal Law of Gravitation
covering both terrestrial gravity and the motion of planets
which he was afraid to discuss.

But this looks more an inspired aspiration than a reality,
as he seems not to have considered gravitational force as
decreasing with distance from its source.

He did however, definitively show that objects experience a uniform
acceleration in a gravitational field (irrespective of their masses)
which was central to Newton's later Universal Law of Gravitation

*Galileo dismissed as a "useless fiction" the idea, held by his
contemporary Johannes Kepler, that the moon caused the tides.*

*Galileo also refused to accept Kepler's elliptical orbits of the planets,
considering the circle the "perfect" shape for planetary orbits.*

The Motion of Falling Bodies

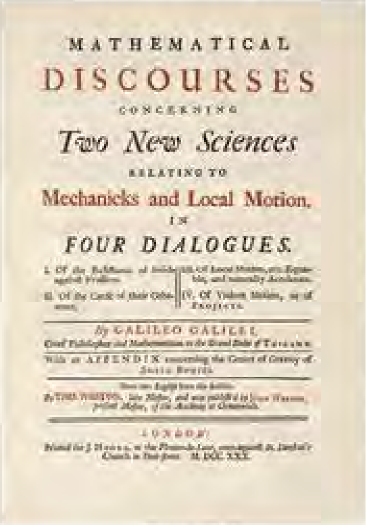

MATHEMATICAL
DISCOURSES
CONCERNING
Two New Sciences
RELATING TO
Mechanicks and Local Motion,
IN
FOUR DIALOGUES.

By GALILEO GALILEI,

LONDON:

(ca. 1590)

Galileo was the first to counter Aristotle's earlier idea's of motion by teaching that
All objects experience a uniform acceleration in a Gravitational field
(irrespective of their different mass-Matter contents)

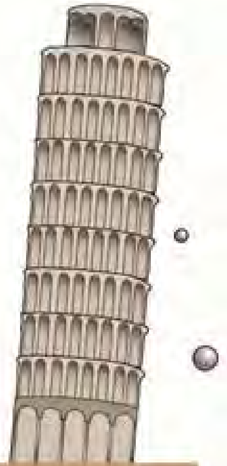

Aristotelian physics

**In De motu Galileo proposed that in free fall bodies dropped with a characteristic
uniform speed determined not by their weight but by their specific gravity.**

**He put this theory to the test by dropping bodies from heights and
found that his experiments did not confirm his theory.**

**It is quoted in 'Galileo's Daughter' that the lighter body (i.e. that of the lower specific gravity)
will move ahead of the heavier body at the start of the fall, and
that the heavier body then overtakes it and arrives at
the bottom slightly earlier.**

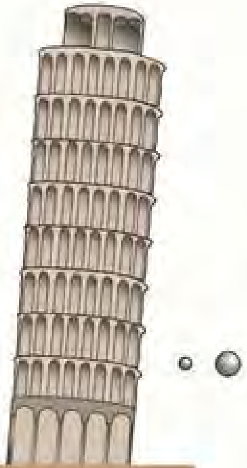

Newtonian physics

Sir Issac Newton

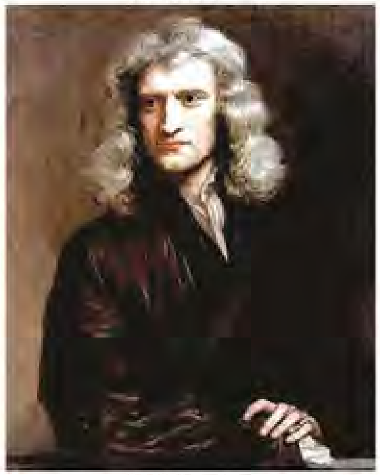

(25 December 1642 – 20 March 1727)

In the late 17th century, Robert Hooke's suggested
'that there is a gravitational force which depends on the inverse square of the distance'

Newtonian Gravitation

Newton's monograph Philosophiæ Naturalis Principia Mathematica, published in 1687, lays the foundations for most of classical mechanics. In this work, Newton described Universal gravitation and the three laws of motion, which dominated the scientific view of the physical universe for the next three centuries.

Newton showed that the motions of objects on Earth and of celestial bodies are governed by the same set of natural laws, by demonstrating the consistency between Kepler's laws of planetary motion and his theory of gravitation, thus removing the last doubts about heliocentrism and advancing the Scientific Revolution.

The Principia is generally considered to be one of the most important scientific books ever written.

Principia Mathematica

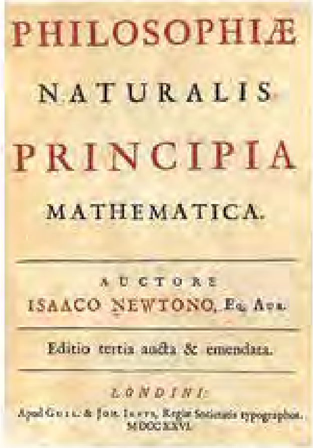

PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

AUCTORE
ISAACO NEWTONO, Eq. Aur.

Editio tertia aucta & emendata.

LONDINI:
Apud Guil. & Joh. Innys, Regiæ Societatis typographos.
MDCCXXVI.

first edition (1686/1687)

First Law of Motion

$$\sum \mathbf{F} = 0 \Rightarrow \frac{d\mathbf{v}}{dt} = 0$$

Bodies at rest - remain at rest
unless acted upon by an outside Force

Second Law of Motion

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \mathbf{v}\frac{\mathrm{d}m}{\mathrm{d}t}$$

The nett force on a particle is equal to the time
rate of change of its linear momentum [p]

Third Law of Motion

$$\sum \mathbf{F}_{a,b} = -\sum \mathbf{F}_{b,a}$$

For every action there is
an equal and opposite reaction

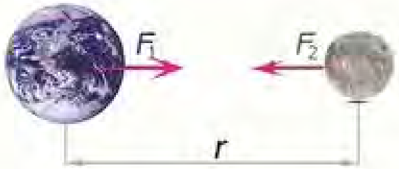

Newton's Law of universal Gravitation

"I deduced that the forces which keep the planets in their orbs must [be] reciprocally as the squares of their distances from the centers about which they revolve: and thereby compared the force requisite to keep the Moon in her Orb with the force of gravity at the surface of the Earth; and found them answer pretty nearly"

$$F = -G\,\frac{M_1\, m_2}{r^2}$$

"hitherto I have not been able to discover the cause of those properties of gravity from the phenomena, and I frame no hypothesis"

According to Bošković's simple dynamistic atomism, Matter is not only endowed with forces (dynamic system), but it is composed of forces (dynamistic system). Forces flow out of the atom and permeate empty space.

This idea led to the concept of the field, much later formulated by M. Faraday (1844), who together with J.C. Maxwell introduced the idea of fields of force into science.

Boskovic is famous for his atomic theory and made many important contributions to astronomy, including the first geometric procedure for determining the equator of a rotating planet from three observations of a surface feature and for computing the orbit of a planet from three observations of its position.

In 1753 he also discovered the absence of atmosphere on the Moon.

Rudjer Boskovic

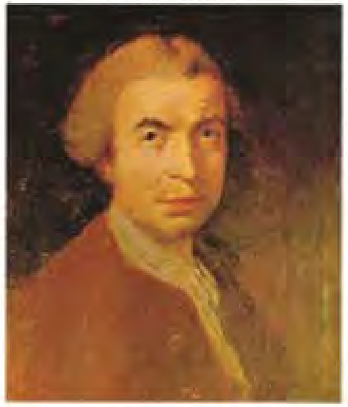

(18 May 1711 – 13 February 1787)

k

Boskovic proposed that attractive Gravitational fields could be explained as diminshed interactive EM fields that result in a nett converent force at great distances

G

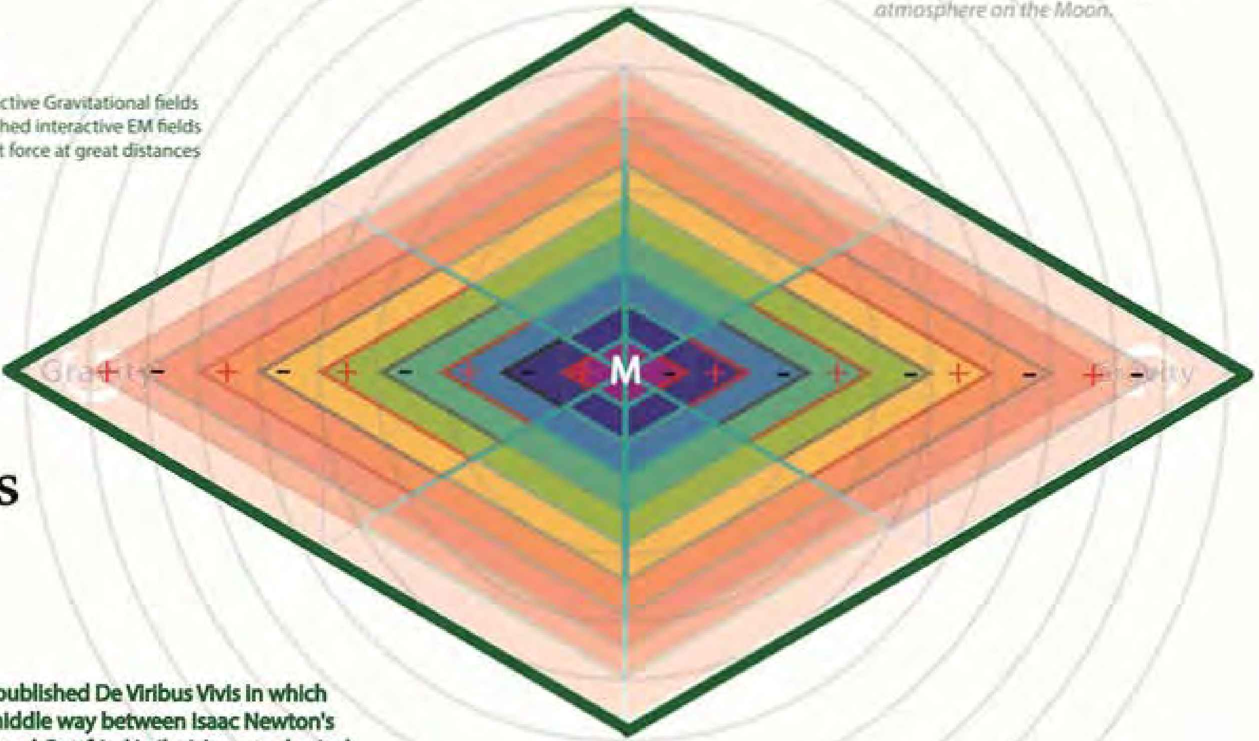

Diminished EM waves

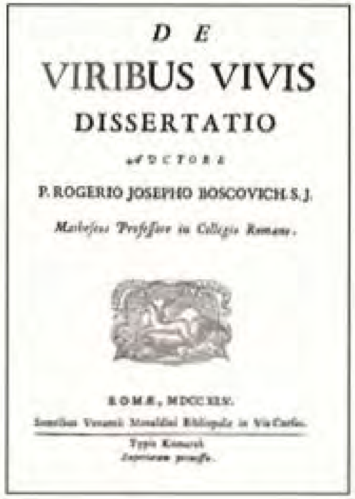

D E
VIRIBUS VIVIS
DISSERTATIO
AUCTORE
P. ROGERIO JOSEPHO BOSCOVICH S. J.
Matheseos Professore in Collegio Romano.

ROMÆ, MDCCXLV.

In 1745 Bošković published De Viribus Vivis in which he tried to find a middle way between Isaac Newton's gravitational theory and Gottfried Leibniz's metaphysical theory of monad-points

The force is repulsive when this curve lies above the line AC and attractive when it lies below it.

At very large distances (at and beyond V), it is attractive and approaches Newton's inverse-square law of force produced by gravity

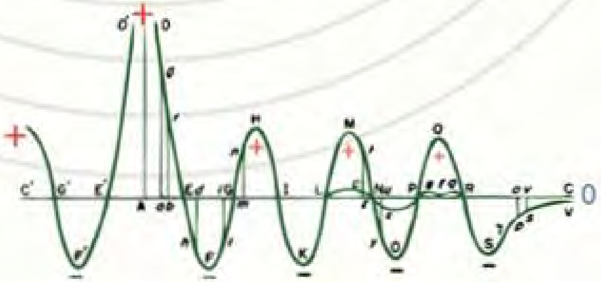

nett attractive force at large distances

Gaussian Gravity

Gauss's law for gravity bears the same mathematical relation to Newton's law that Gauss's law for electricity bears to Coulomb's law.

G

Gauss's law for gravity is mathematically similar in form to Gauss's law for electrostatics.

k

the striking mathematical similarity in the formulation of electrostatics and gravitation has puzzled science for centuries

$$F_{elec}\; k \propto \frac{Q_1 Q_2}{r^2}$$

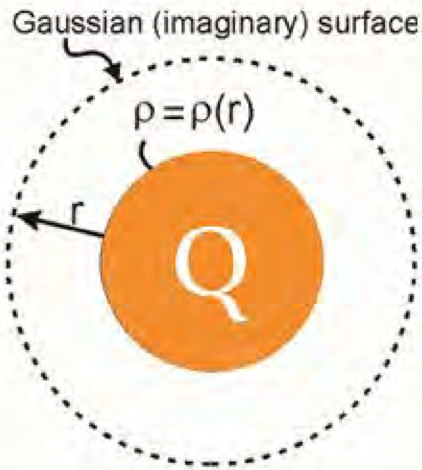

$$\oint \vec{E} \cdot d\vec{A} = \frac{q}{\varepsilon_0} = 4\pi k q$$

The electric flux through any closed surface is proportional to the enclosed electric charge.

$$\nabla \cdot E = \frac{\rho}{\varepsilon_0} = 4\pi k \rho$$

Carl Friedrich Gauss

(30 April 1777 – 23 February 1855)

Gauss's law for gravity, also known as Gauss's flux theorem for gravity, is a law of physics which is essentially equivalent to Newton's law of universal gravitation.

Although Gauss's law for gravity is physically equivalent to Newton's law, there are many situations where Gauss' law for gravity offers a more convenient and simple way to do a calculation than Newton's law.

note: 4π spherical topologies are equivalent to 4π tetrahedral topologies
[see Euler characteristics & Gauss-Bonnet topology theorem]

$$F_{grav}\; G \propto \frac{M_1 m_2}{r^2}$$

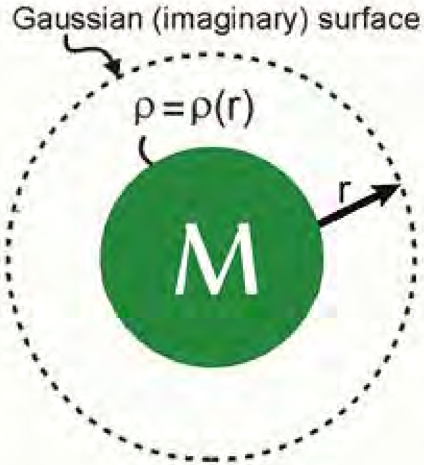

$$\oint_{\partial V} \mathbf{g} \cdot d\mathbf{A} = -4\pi G M$$

The gravitational flux through any closed surface is proportional to the enclosed Matter

$$\nabla \cdot \mathbf{g} = -4\pi G \rho$$

Poisson's Gravity fields

Charges are the sources and sinks of EM fields

Poisson recognised that charges are the sources and sinks of electrostatic fields:
positive charges emanate electric field lines, and the field lines terminate at negative charges.

Similarly, in Newton's gravitation masses are the sources of the gravity field
so that field lines terminate at objects that have mass.

Simeon Poisson

(21 June 1781 – 25 April 1840)

Matter is the source of Gravitational fields

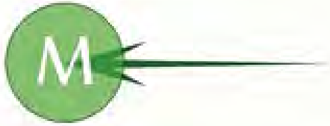

Poisson formalised Gauss' Law for electric fields (using the more general divergence theorem)

$$\iint \vec{E}\cdot d\vec{S} = \frac{q_e}{\epsilon_0} \Rightarrow \vec{\nabla}\cdot\vec{E} = \frac{\rho_e}{\epsilon_0}$$

and then applied the same formulation to spherical Matter for gravitation

$$\iint \vec{g}\cdot d\vec{S} = -4\pi G_c m \Rightarrow \vec{\nabla}\cdot\vec{g} = -4\pi G_c \rho_m$$

All EM field interactions are the result of superpositioned EM fields (comprising of convergent AND divergent energy momenta) whose quanta create differing field strengths,
The inverse square forces of gravity can be modelled using the same field geometry thus illustrating the reason for the geometric & mathematical similarity between Newton's & Coulomb's constants

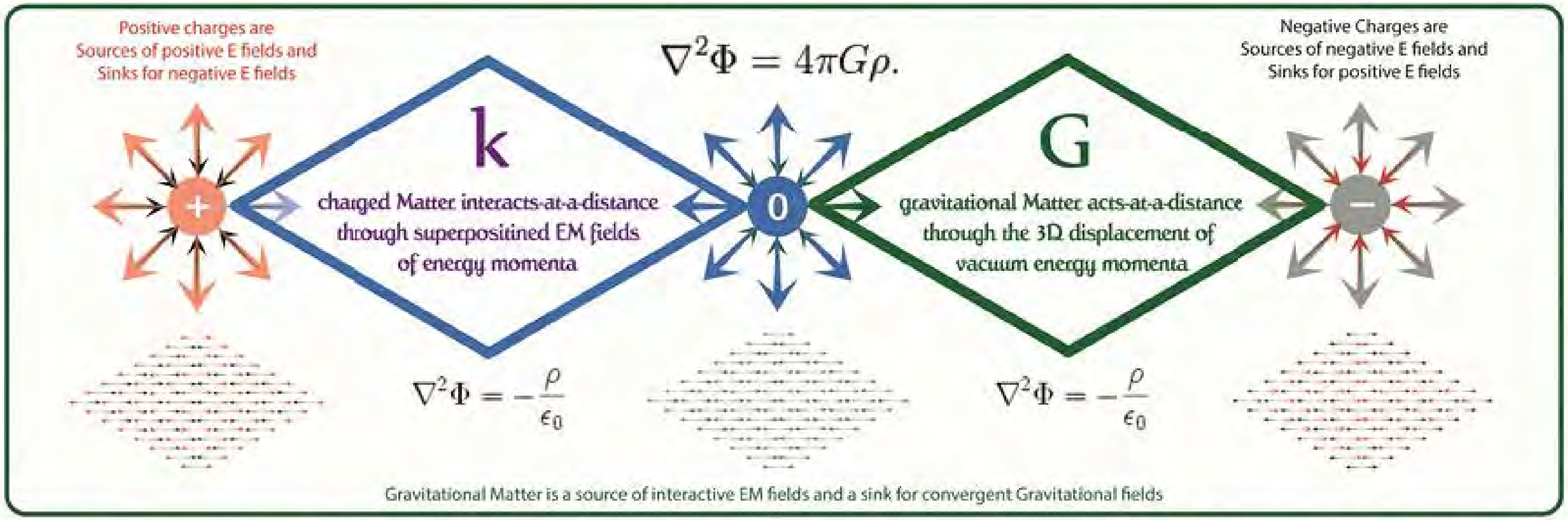

Positive charges are
Sources of positive E fields and
Sinks for negative E fields

$$\nabla^2\Phi = 4\pi G\rho.$$

Negative Charges are
Sources of negative E fields and
Sinks for positive E fields

charged Matter interacts-at-a-distance
through superpositined EM fields
of energy momenta

gravitational Matter acts-at-a-distance
through the 3D displacement of
vacuum energy momenta

$$\nabla^2\Phi = -\frac{\rho}{\epsilon_0}$$

$$\nabla^2\Phi = -\frac{\rho}{\epsilon_0}$$

Gravitational Matter is a source of interactive EM fields and a sink for convergent Gravitational fields

Tetryonic QM & QED clearly shows that both the sink and source fields are the result of bidirectional energy momenta

Poison's field equations shows that weak KEM field geometries can be used to model the convergent forces of Gravitational fields

Copyright ABRAHAM [2008] - All rights reserved

Gravito-Electro-Magnetism

inverse squared law

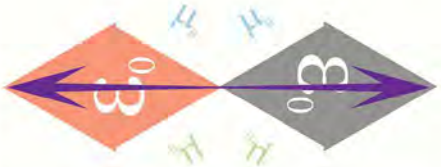

nett divergent
Electric field forces

$1/r^2$

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

k

$$\oint \vec{B}\cdot d\vec{A} = 0$$

inverse cubed law

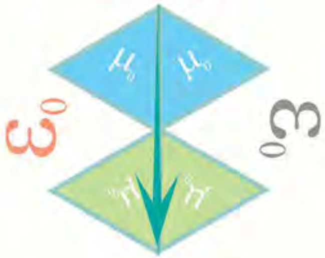

vector
Magnetic forces

$1/r^3$

Maxwell's 1864 achievements concerning electromagnetism have been called the "second great unification in physics", following on from the first one realised by Isaac Newton.

Maxwell's equations
$\nabla\cdot\mathbf{E} = \frac{\rho}{\epsilon_0}$
$\nabla\cdot\mathbf{B} = 0$
$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}$
$\nabla\times\mathbf{B} = \frac{1}{\epsilon_0 c^2}\mathbf{J} + \frac{1}{c^2}\frac{\partial\mathbf{E}}{\partial t}$

James Clerk Maxwell

(13 June 1831 – 5 November 1879)

Any opinion as to the form in which the energy of gravitation exists in space is of great importance, and whoever can make his opinion probable will have, made an enormous stride in physical speculation.

The apparent universality of gravitation, and the equality of its effects on matter of all kinds are most remarkable facts, hitherto without exception; but they are purely experimental facts, liable to be corrected by a single observed exception.

We cannot conceive of Matter with negative inertia or mass; but we see no way of accounting for the proportionality of gravitation to mass by any legitimate method of demonstration

GEM equations
$\nabla\cdot\mathbf{E}_g = -4\pi G\rho_g$
$\nabla\cdot\mathbf{B}_g = 0$
$\nabla\times\mathbf{E}_g = -\frac{\partial\mathbf{B}_g}{\partial t}$
$\nabla\times\mathbf{B}_g = 4\left(-\frac{4\pi G}{c^2}\mathbf{J}_g + \frac{1}{c^2}\frac{\partial\mathbf{E}_g}{\partial t}\right)$

$1/r^2$

strictly convergent vector force

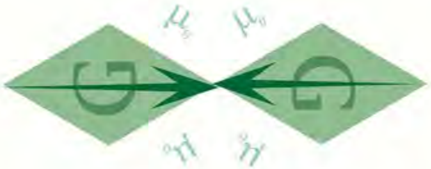

Maxwell proposed a set of equations based on his EM field laws that created a Gravito-Electro-Magnetic field reflective of the similarities between Newton's & Coulomb's Laws

$$\nabla\times\mathbf{E}_g = -\frac{\partial\mathbf{B}_g}{\partial t}$$

G

$$\nabla\cdot\mathbf{B}_g = 0$$

giving rise to spectulation about the existence of 'gravitons'
a theorised Gravitational 'charge' carrier

$1/r^3$

gravitomagnetic dipole

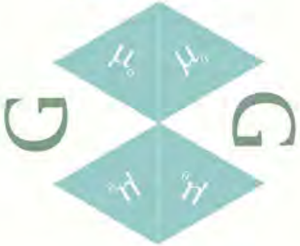

Copyright ABRAHAM [2008] - All rights reserved

General Relativity

The General Theory of Relativity

In 1915, Einstein extended SR and proposed a new theory of gravitation to account for accelerating bodies

Einstein said that when he "was sitting in a chair in the patent office at Bern when all of a sudden a thought occurred to me: 'If a person falls freely he will not feel his own weight.' I was startled. This simple thought ... impelled me toward a theory of gravitation. ..."

$$G_{ab} = 8\pi G\,T_{\mu\nu}$$

$$R_{\mu\nu} - \tfrac{1}{2}R\,g_{\mu\nu} = \frac{8\pi G}{c^4}\,T_{\mu\nu}$$

The EFE is a tensor equation relating a set of symmetric 4 x 4 tensors.

the scalar curvature

[illegible]

Newton's gravitational constant

Ricci curvature tensor

$$R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}\,T_{\mu\nu}$$

the stress-energy tensor

the metric tensor

the speed of light in vacuum

Each tensor has 10 independent components

Albert Einstein

(14 March 1879 – 18 April 1955)

Spacetime

Matter

$$G_{\mu\nu} + g_{\mu\nu}\Lambda = 8\pi T_{\mu\nu}$$

the cosmological constant

$$\underset{\text{curved spacetime}}{G_{ab}} = \kappa\,\underset{\text{mass-energy tensor}}{T_{\mu\nu}}$$

where

$$\kappa = -\frac{8\pi G}{c^4}$$

General Relativity describes gravity as a warping of space itself, not as a force.
Einstein pictured space as a three-dimensional version of a thin rubber sheet

Einstein's General Theory of Relativity, which was published in 1916 had as its foundation that the laws of nature in an accelerating frame are equivalent to the laws of a gravitational field.

This is known as the Equivalence Principle.

GEM fields & Gravity

Sir Issac Newton

(25 December 1642 – 20 March 1727)

Gravity is one of the four fundamental interactions of nature,
along with electromagnetism and the nuclear strong and weak forces

[for centuries scientific enquiry has been aimed squarely at discerning a physical mechanism for gravitation]

$$\overset{\text{Matter}}{\frac{4n\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Albert Einstein

(14 March 1879 – 18 April 1955)

Tetryonics reveals Gravity to be the result of the displacement of vacuum energies by mass-Matter topologies which is accurately modelled using GEM field geometry

$$F = -G\,\frac{\overset{\text{masses}}{M_1 m_2}}{r^2}$$

$$G_{ab} = \frac{8\pi G}{c^4}\,\overset{\text{energies}}{T_{\mu\nu}}$$

Gravity is responsible for keeping the Earth and the other planets in their orbits around the Sun;
for keeping the Moon in its orbit around the Earth; for the formation of tides; for natural convection,
by which fluid flow occurs under the influence of a density gradient and gravity;
for heating the interiors of forming stars and planets to very high temperatures;
and for various other phenomena observed on Earth.

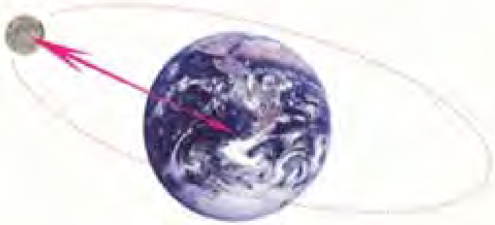

instantaneous
action-at-a-distance

*Gravity is a natural phenomenon by which physical bodies
attract each other with a force proportional to their mass-Matter content.*

*Gravity is most familiar as the agent that gives weight to objects with mass and
causes them to fall to the ground when dropped.*

*Gravity causes dispersed Matter to coalesce, and coalesced Matter to remain intact,
to the creation of the SUN, the Earth, and to source of energy in the stars that light our universe.*

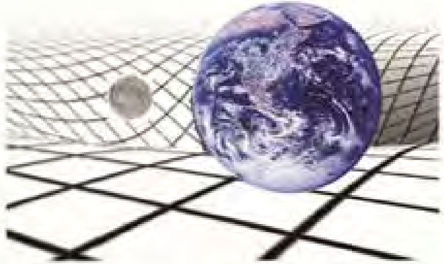

motion in
curved spacetime

Separating Space & Time

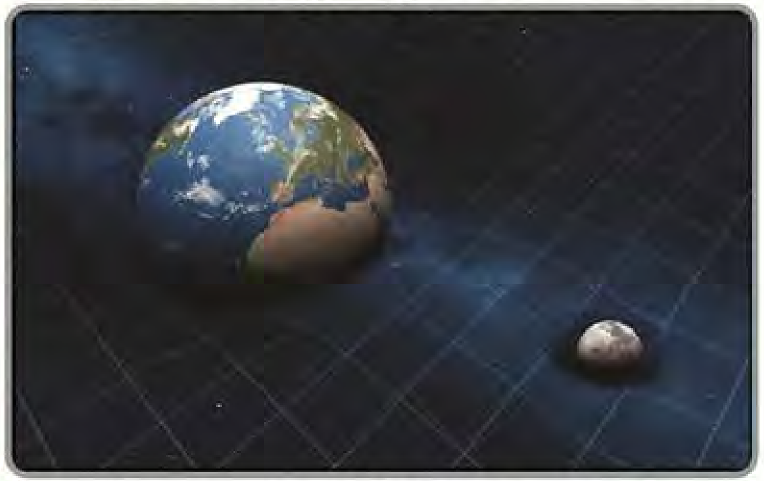

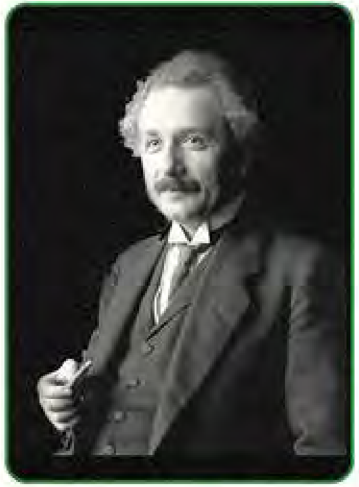

(in the search for Gravity)

Instantaneous
'action-at-a-distance'

Relativistic
spacetime 'curvature'

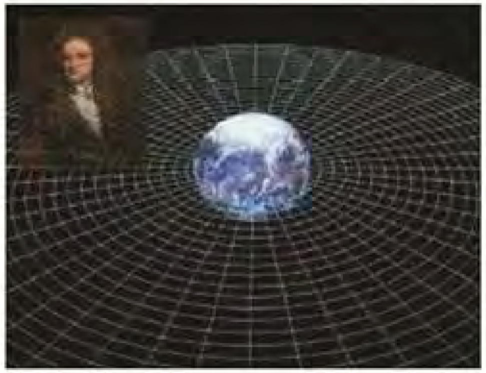

DOES

The Earth pull on the moon and
the moon pull proportionally back on the Earth

OR

Does the Earth curve spacetime in its vicinity
with the moon following the geodesic path created

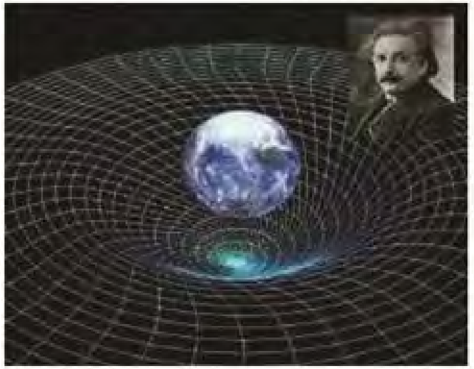

Tetryonic theory through its clear definitions of charged mass-ENERGY-Matter provides the long-sought answer to Gravitation

Copyright ABRAHAM [2008] - All rights reserved

Empty Space

is defined in Tetryonic & relativistic theory
by the speed of light

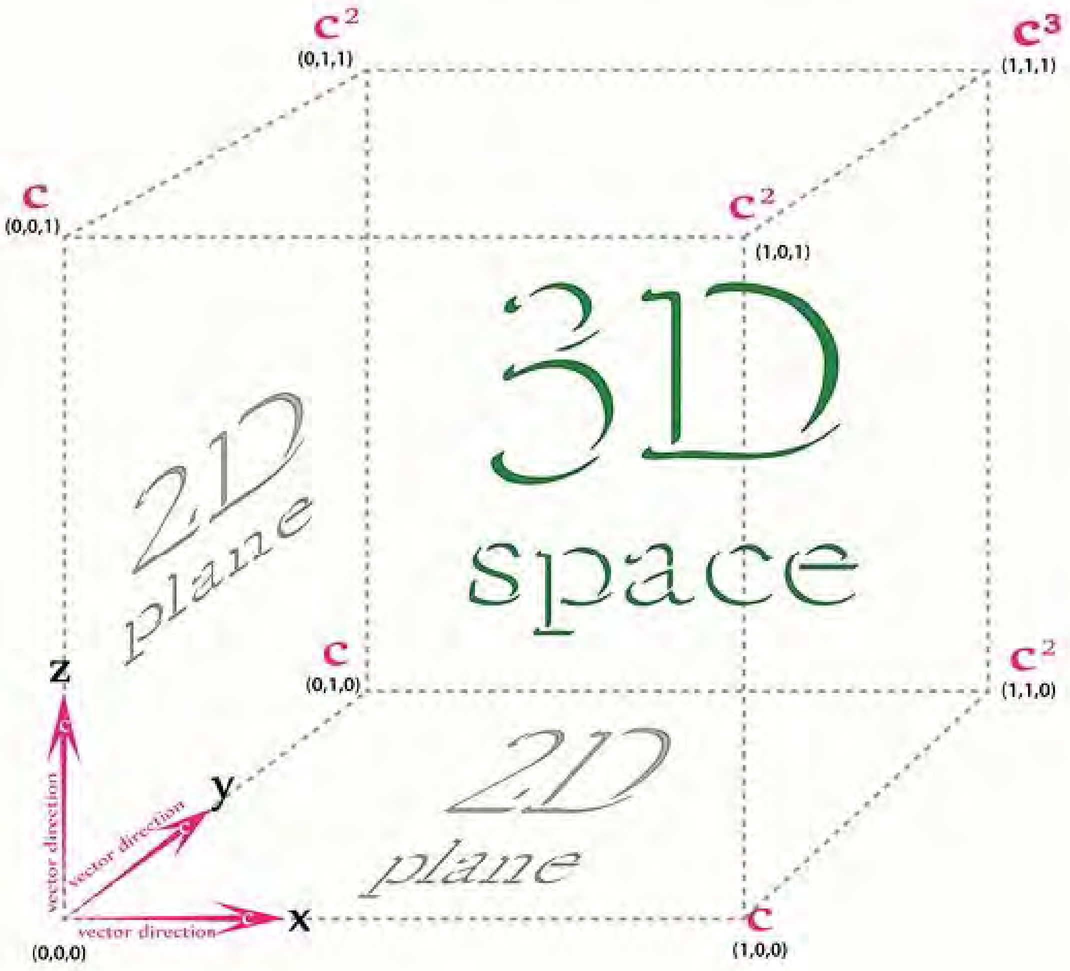
c^2
(0,1,1)
c^3
(1,1,1)
c
(0,0,1)
c^2
(1,0,1)
3D
space
2D
plane
c
(0,1,0)
c^2
(1,1,0)
2D
plane
z
y
x
vector direction
vector direction
vector direction
c
c
c
(0,0,0)
c
(1,0,0)

Energy

content of free space is Zero

A Spatial region is defined
by a co-ordinate system so as to
measure the physics and motion
of mass-ENERGY-Matter
within its confines

mass

is Energy content per unit of time

EM mass-energies can
be found in (and move through)
empty Space in various forms:
radiant energies, Matter etc

Matter

is Energy content per unit of time squared

Space can be any regular geometry
[Cubic, Spherical or polyhedral]
limited by the spatial co-ordinates
used to define the region or volume

Empty Space is defined as a geometric volume devoid of any form of Energy

Spacetime co-ordinates

The linear force vectors of divergent Energy
can form various co-ordinate systems
based on their vector directions
per unit of time

Charge radiates unilaterally
in the direction of its
vector linear momentum

Vector forces

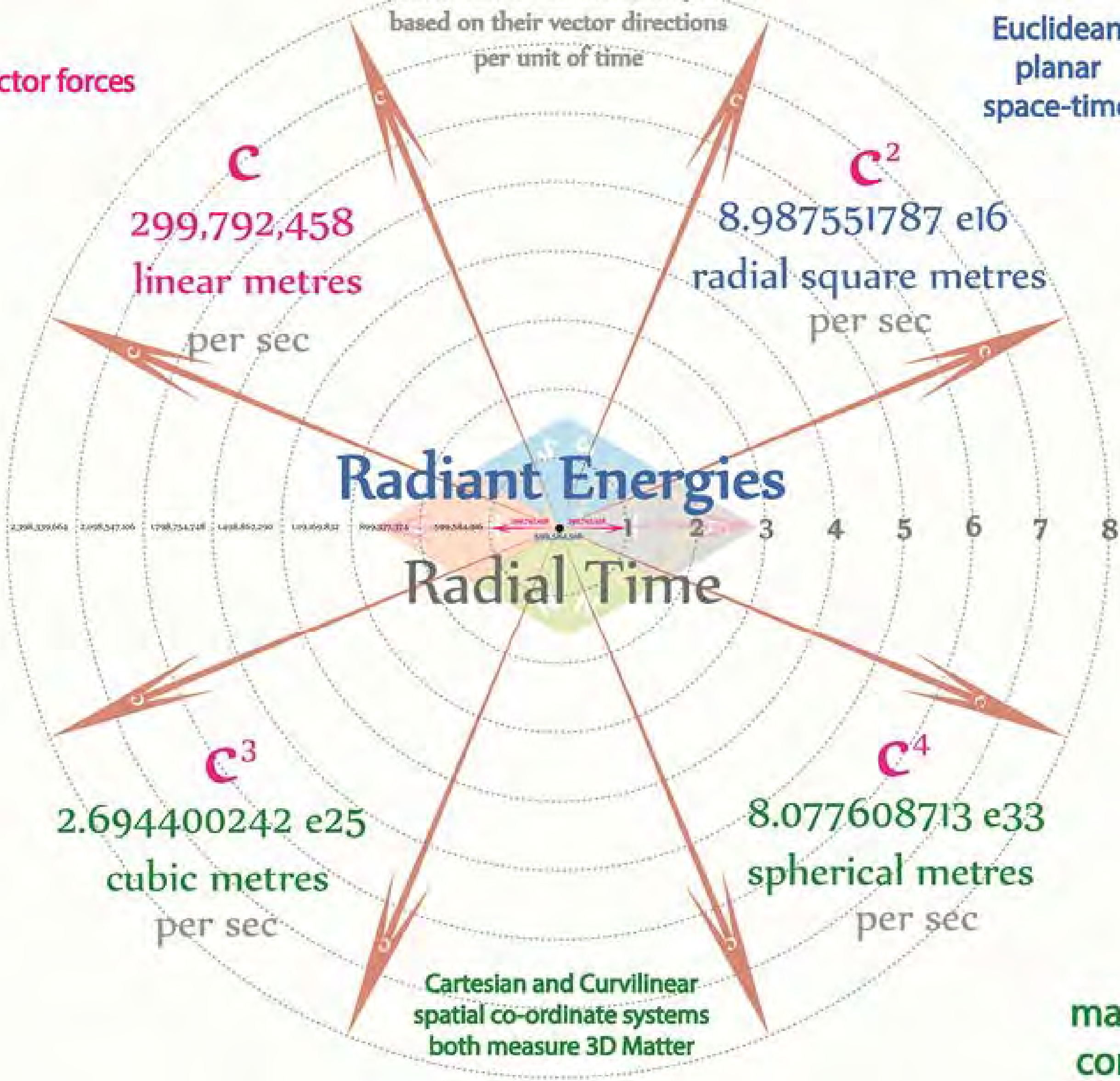
Radiant Energies
Radial Time
1
2
3
4
5
6
7
8
2,398,339,664
2,098,547,206
1,798,754,748
1,498,962,290
1,199,169,832
899,377,374
599,584,916
299,792,458

c
299,792,458
linear metres
per sec

c^2
8.987551787 e16
radial square metres
per sec

c^3
2.694400242 e25
cubic metres
per sec

c^4
8.077608713 e33
spherical metres
per sec

Cartesian and Curvilinear
spatial co-ordinate systems
both measure 3D Matter

Euclidean
planar
space-time

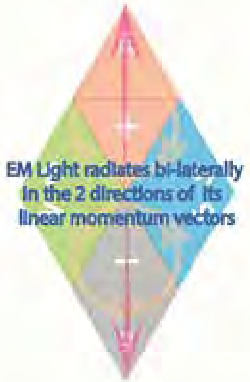

EM Light radiates bi-laterally
in the 2 directions of its
linear momentum vectors

Time in Physics is a measure
of how long it takes for light to
travel 299,792,458 metres
from its source in a vacuum

$$\frac{\text{metre}}{299{,}792{,}458\ \text{sec}}$$

A metre in Physics is
the distance light travels in
1/ 299,792,458 of a second
from its source

$$\frac{299{,}792{,}458\ \text{m}}{\text{second}}$$

**Energy entropy
of a system**

**mass-energy-Matter
content of a system**

Copyright ABRAHAM [2008] - All rights reserved

Spatial co-ordinate geometries

based on the vector speed of light
form distinct spatial co-ordinate systems
for the measurement of physics

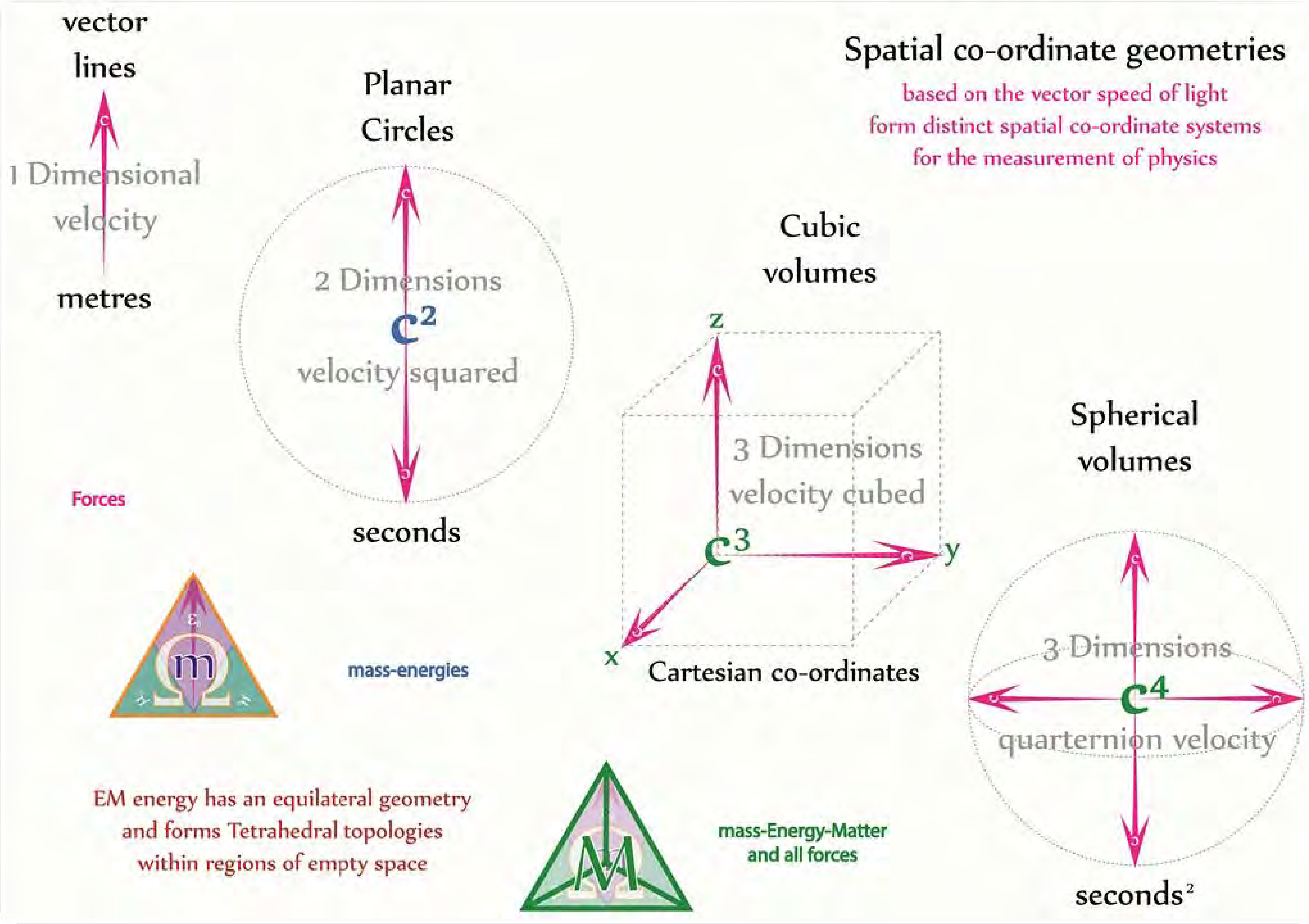

vector lines

1 Dimensional velocity

metres

Planar Circles

2 Dimensions

c^2

velocity squared

seconds

Cubic volumes

3 Dimensions

velocity cubed

c^3

Cartesian co-ordinates

Spherical volumes

3 Dimensions

c^4

quarternion velocity

seconds²

Forces

mass-energies

EM energy has an equilateral geometry
and forms Tetrahedral topologies
within regions of empty space

mass-Energy-Matter
and all forces

Copyright ABRAHAM [2008] - All rights reserved

Spatial co-ordinate systems

Cubic

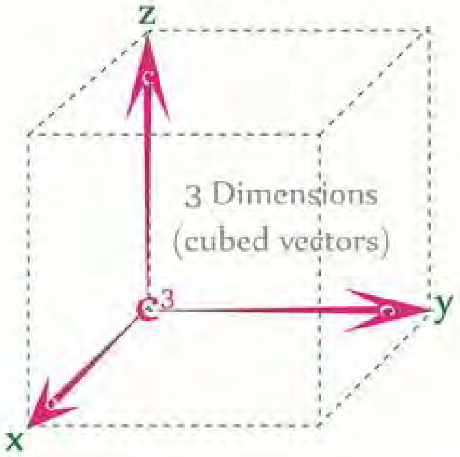

Spacetime (or space-time, or space time) is any mathematical model that combines space and time into a single continuum.

Spacetime is usually interpreted with space as being three-dimensional and time playing the role of a fourth dimension that is different from the spatial dimensions.

Tetryonic theory redefines as time to be a measure of the changing quantised angular momenta of Planck scale Charges In any spatial plane or volume

By combining space and time into a single manifold, we can significantly simply a large number of physical theories, and re-formulate them in a more uniform way to explain the mechanics of the Universe at all scales

Tetryonics maps radial spacetime co-ordinates through the vector momentum of Energy

Spherical

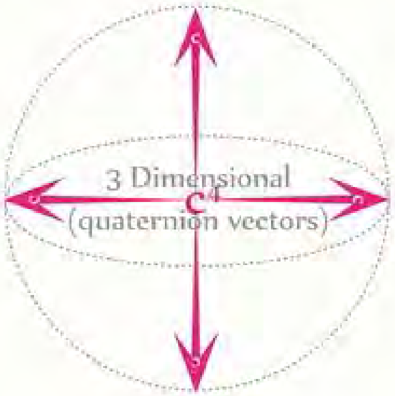

seconds²

Cartesian Space-Time

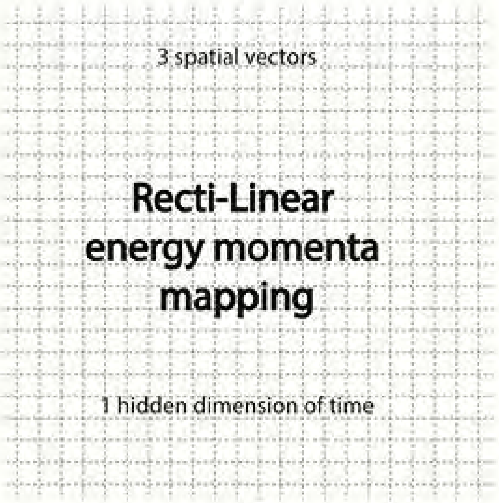

3 spatial vectors

Recti-Linear energy momenta mapping

1 hidden dimension of time

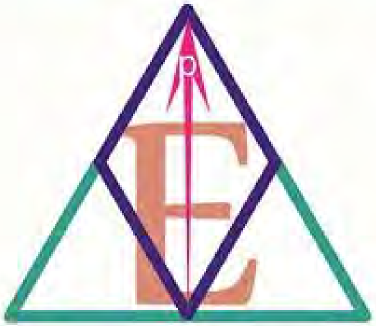

The quantised angular momenta of mass-Energy and Matter are triangular geometries

The current methodologies for mapping energy densities onto Euclidean, Cartesian and Reimannian co-ordinates are inaccurate reflections of the charged mass-ENERGY-Matter geometries, introducing a mathematical complexity to a otherwise simple foundational equilateral geometry

Tetryonic Space-Time

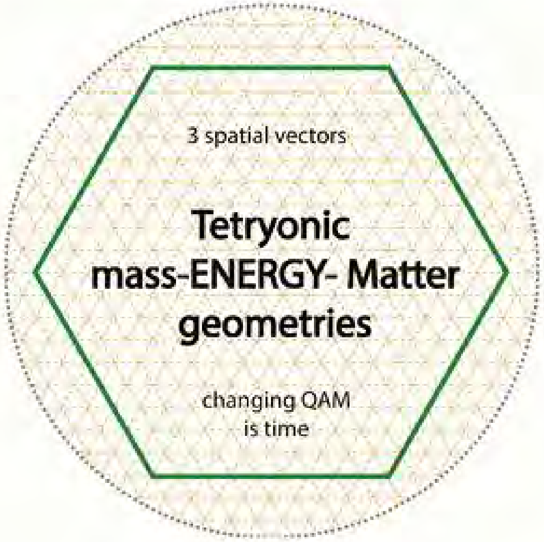

3 spatial vectors

Tetryonic mass-ENERGY- Matter geometries

changing QAM is time

By mapping equilateral mass-energies & Matter into radial spatial co-ordinates Tetryonic theory is able to reveal the hidden 4th dimension of time

Copyright ABRAHAM [2008] - All rights reserved

Space & time separated

All equilateral mass-energies & tetrahedral Matter can be mapped with planar spatial geometries within spherical space-time co-ordinates

Space is the region of measurement defined by the vector propagation of light
time is a measure of the changing QAM within any spatial region

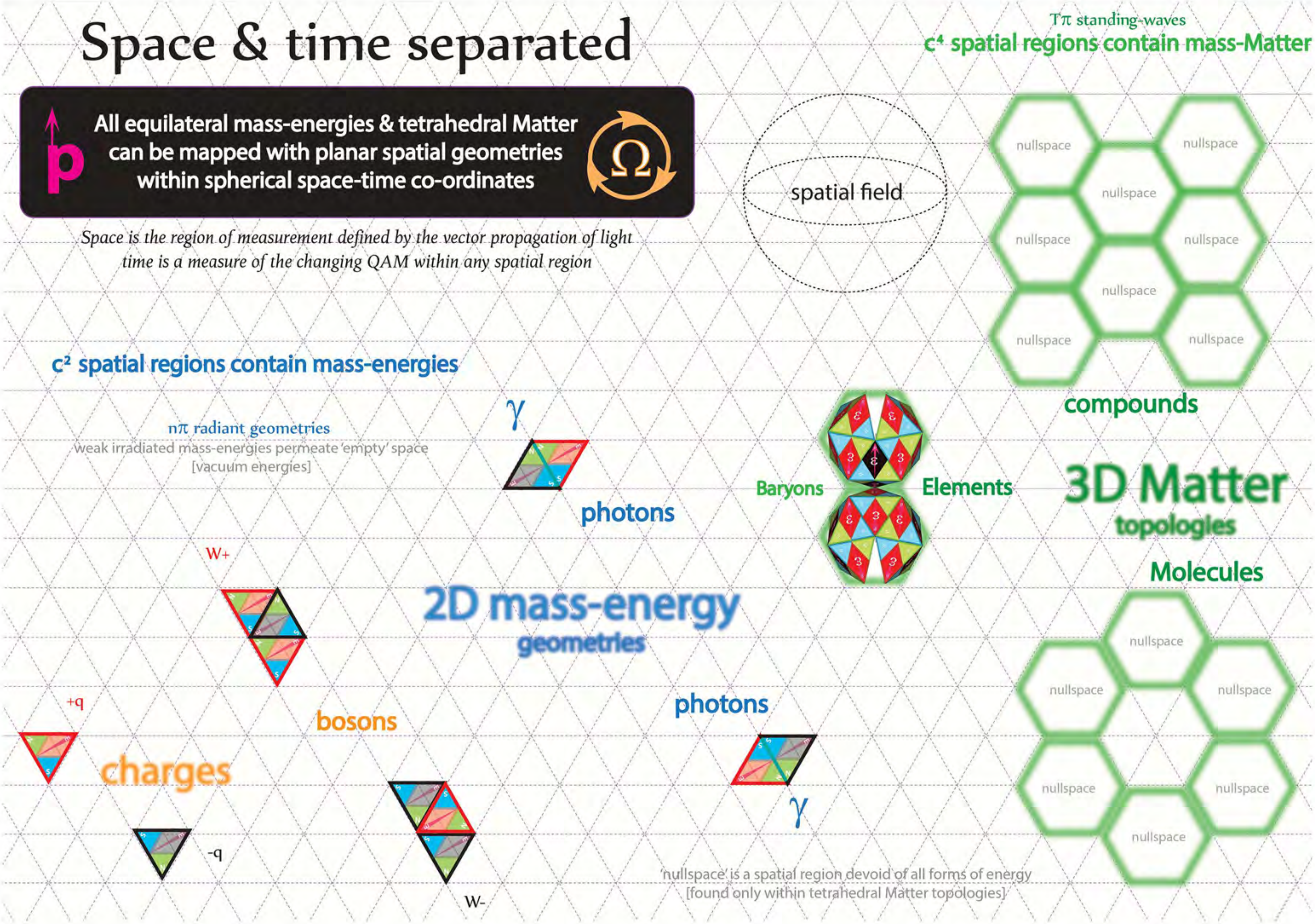
p
Ω
spatial field
nullspace
nullspace
nullspace
nullspace
nullspace
nullspace
nullspace
nullspace
γ
W+
+q
-q
W-
γ
nullspace
nullspace
nullspace
nullspace
nullspace
nullspace

Tπ standing-waves

c^4 spatial regions contain mass-Matter

compounds

Baryons

Elements

3D Matter
topologies

Molecules

c^2 spatial regions contain mass-energies

nπ radiant geometries
weak irradiated mass-energies permeate 'empty' space
[vacuum energies]

photons

2D mass-energy
geometries

bosons

charges

photons

'nullspace' is a spatial region devoid of all forms of energy
[found only within tetrahedral Matter topologies]

Copyright ABRAHAM [2008] - All rights reserved

2D Euclidean spatial geometry

All 2D mass-energy momenta map onto **FLAT** Euclidian co-ordinates

[x,y] ΔΩ

(2D space & time)

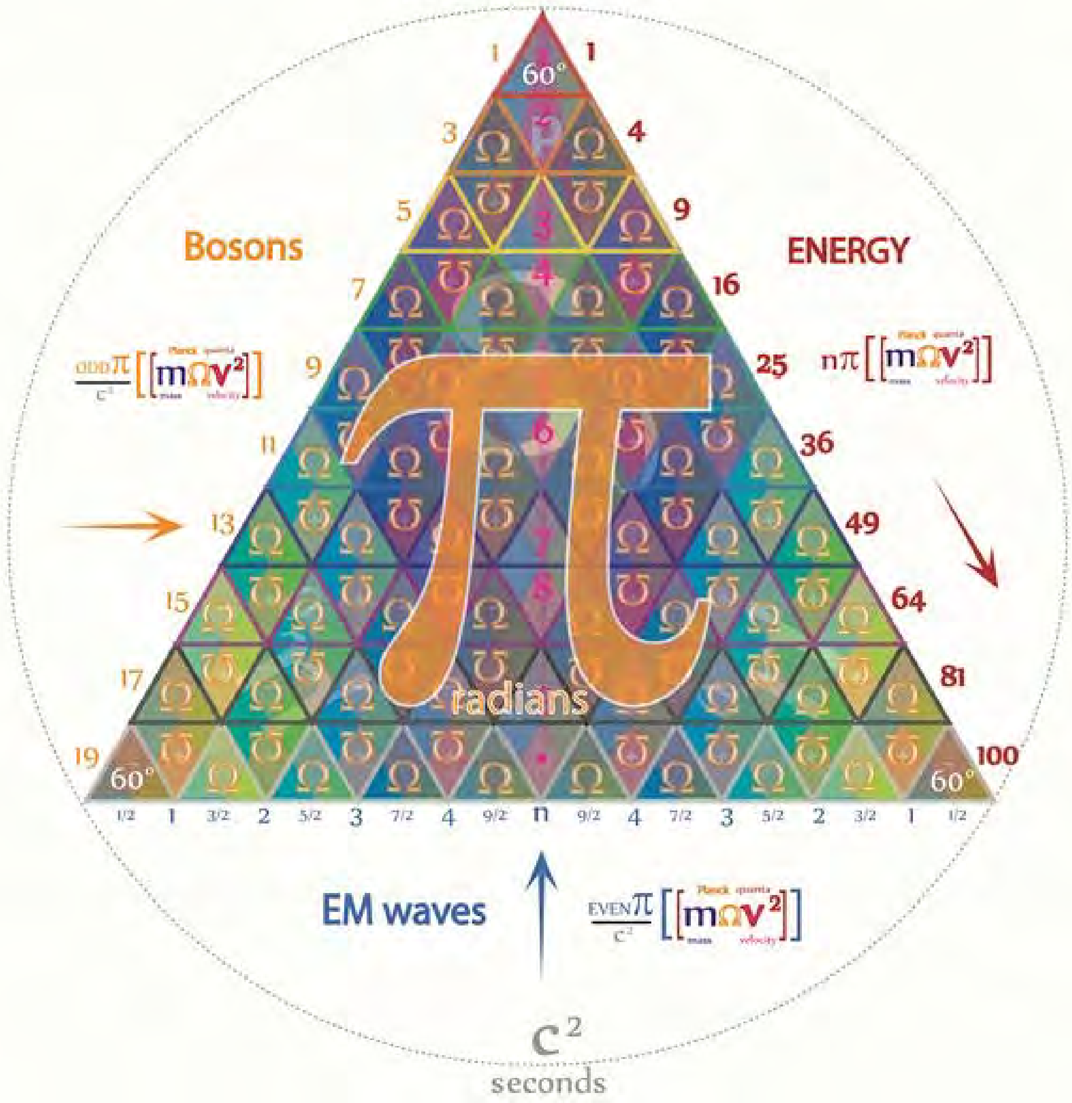
1
60°
1
3
4
5
9
7
16
9
25
11
36
13
49
15
64
17
81
19
100
60°
60°
1/2 1 3/2 2 5/2 3 7/2 4 9/2 n 9/2 4 7/2 3 5/2 2 3/2 1 1/2
π
radians
Bosons
ENERGY
EM waves
ODD π / c² [[mΩv²]]
nπ [[mΩv²]]
EVEN π / c² [[mΩv²]]
Planck quanta
mass
velocity
c²
seconds

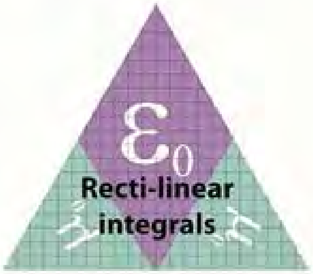
ε₀
Recti-linear integrals

Cartesian spatial co-ordinates

fail to accurately map equilateral energy momenta

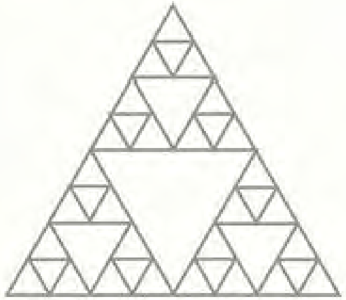

Minkowski spacetime

is a close approximation of Tetryonic energy momenta

ε₀
Equilateral integrals

Tetryonic space-time

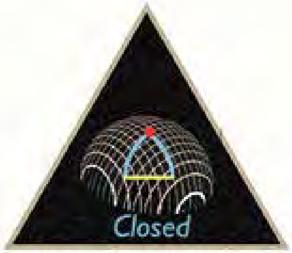

negative curvative

angles add up to > 180 °

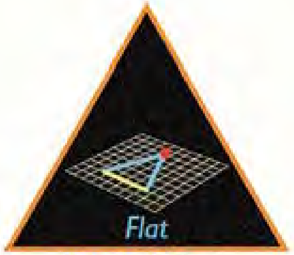

flat Euclidean

angles = 180 °

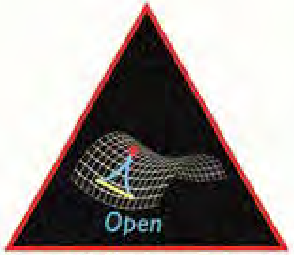

positive curvature

angles add up to < 180 °

Copyright ABRAHAM [2008] - All rights reserved

Light cones space & time

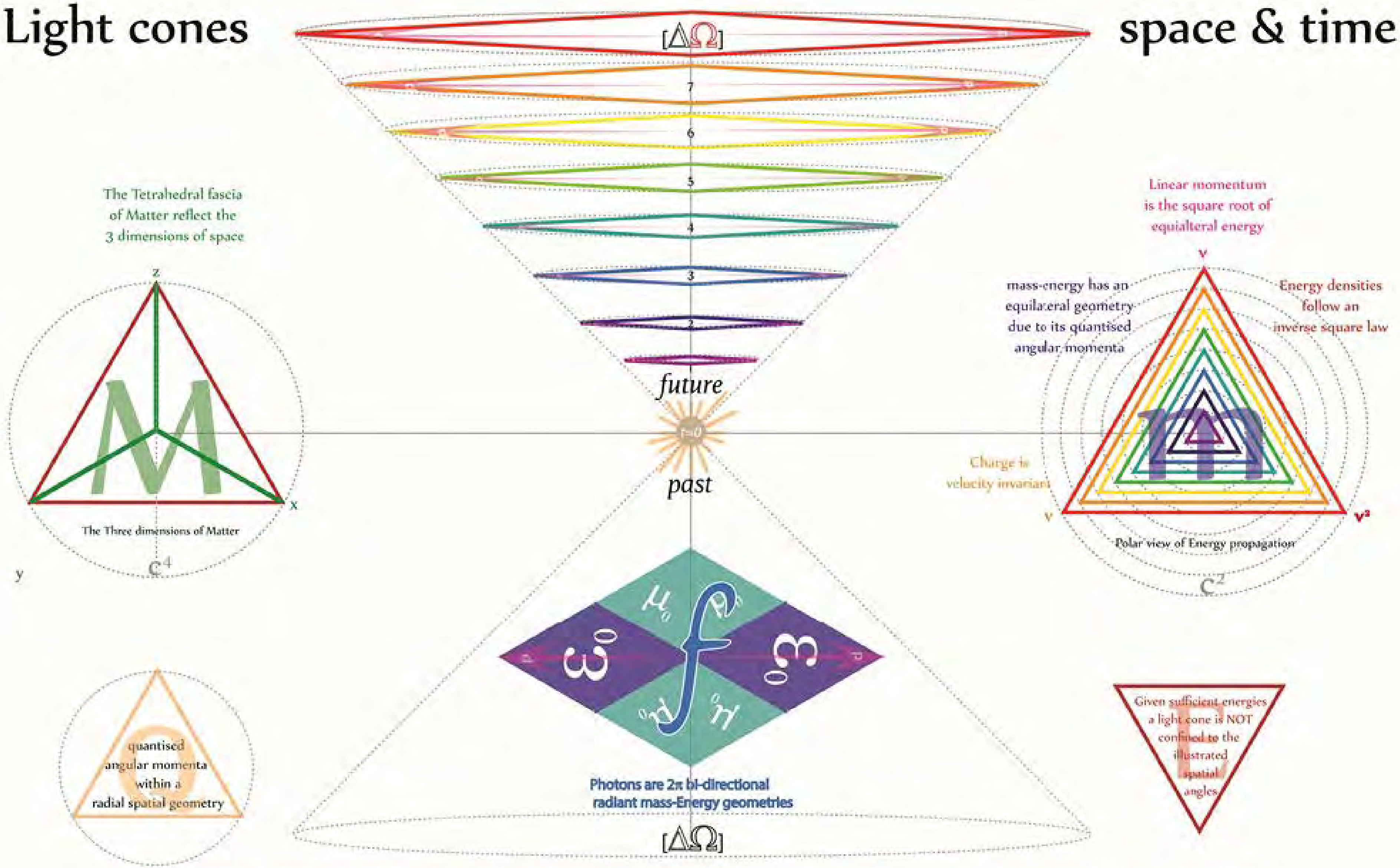

The Tetrahedral fascia
of Matter reflect the
3 dimensions of space

The Three dimensions of Matter

Linear momentum
is the square root of
equialteral energy

mass-energy has an
equilateral geometry
due to its quantised
angular momenta

Energy densities
follow an
inverse square law

Charge is
velocity invariant

Polar view of Energy propagation

quantised
angular momenta
within a
radial spatial geometry

**Photons are 2π bi-directional
radiant mass-Energy geometries**

Given sufficient energies
a light cone is NOT
confined to the
illustrated
spatial
angles

time is the measurement of the changing quantised angular momenta in any spatial region

A light cone is the path that a flash of light, emanating from a single event
(localized to a single point in space and a single moment in time) and traveling in all directions,
would take through space & time.

Copyright ABRAHAM [2008] - All rights reserved

Charge, energy momenta & time

measurements of the changing quantised angular momenta of scalar energy is the basis of measuring time

Ω

$\boxed{.0012}$

$\frac{m^2}{s}$

quantised angular momentum

$$\beta = \left[\frac{v}{c}\right]$$

Linear Lorentz factor

metres

$\frac{m^2}{s} \cdot \frac{s}{m}$

$\boxed{\frac{\Omega}{v}}$

λ

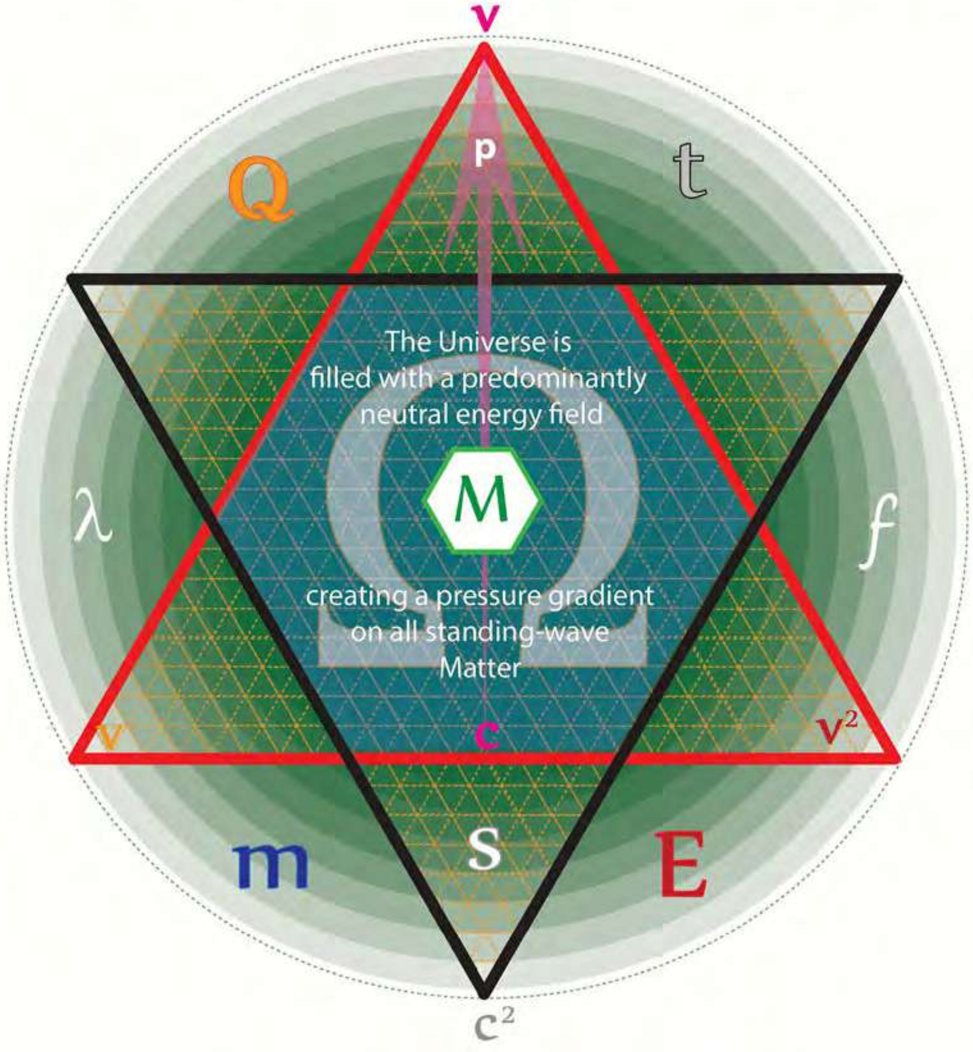

The Universe is filled with a predominantly neutral energy field

creating a pressure gradient on all standing-wave Matter

Velocity of Light

$$c = \left[\underset{\text{Wavelength}}{\overset{\lambda}{\left[\frac{\Omega}{c}\right]}} \cdot \underset{\text{Frequency}}{\overset{f}{\left[\frac{c^2}{\Omega}\right]}} \right]$$

q

$\boxed{\frac{\Omega}{c^2}}$

$\frac{m^2}{s} \cdot \frac{s^2}{m^2}$

seconds

$$\left[\frac{v^2}{c^2}\right] = \beta^2$$

Scalar Lorentz factor

per second

$\frac{m^2}{s^2} \cdot \frac{s}{m^2}$

$\boxed{\frac{v^2}{\Omega}}$

f

Copyright ABRAHAM [2008] - All rights reserved

$\vec{p}$

Spatial co-ordinate systems are defined by the vectors of their linear energy momentum

Space
[x,y,z]

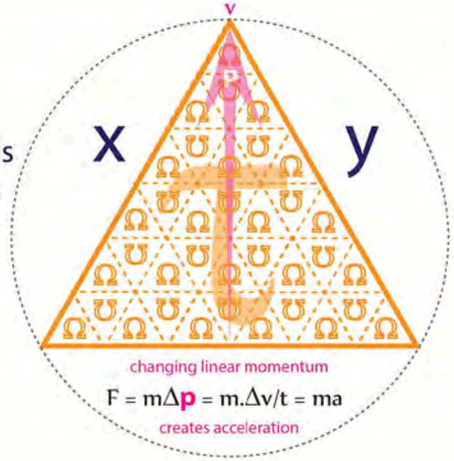
v
x
y
p
changing linear momentum
F = mΔp = m.Δv/t = ma
creates acceleration

$\Delta\Omega$

Time is a measure of the changing quantised angular momenta in any defined region of space

Time
[seconds]

Spacetime Fields

are spatial geometries that contain changing Quantised Angular energy-Momenta
[Planck quanta]

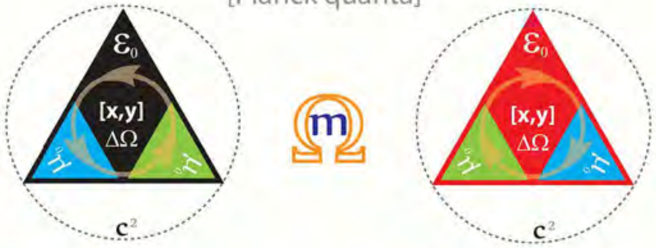
ε_0
[x,y]
ΔΩ
c²
Ωm
ε_0
[x,y]
ΔΩ
c²

Charge is the nett distribution of Planck quanta mass.[Quantised Angular Momenta] within any spatial co-ordintate system

EM masses are 2D planar energy waveforms

2D mass-energy quanta have linear & angular momenta

Matter has a 3D volumetric energy waveform

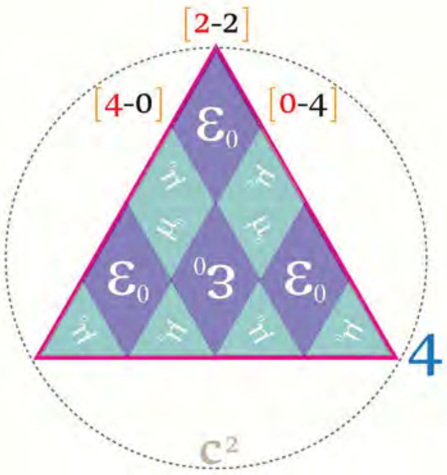
[2-2]
[4-0]
[0-4]
ε_0
ε_0
03
ε_0
4
c²

mass Planck quanta

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

mass velocity

Tetryonic geometry restores symmetry to Einstein's mass-energy equivalence and reveals a mass-ENERGY-Matter equivalance

radiant mass-energies can form standing-wave Matter

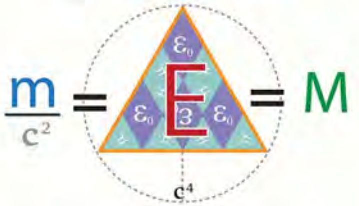
ε_0
E
ε_0
3
ε_0
c⁴

$$\frac{m}{c^2} = E = M$$

standing-wave Matter can release radiant mass-energies

where space and time are separated

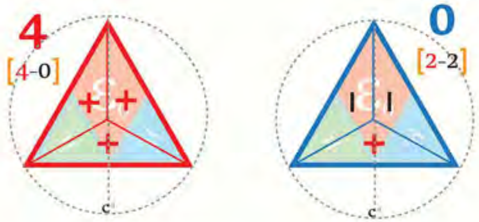

Matter Planck quanta

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

mass velocity

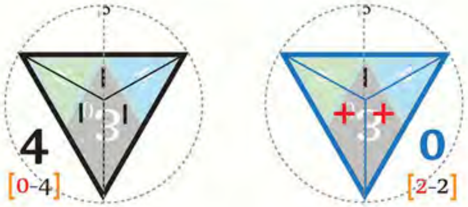
c
4
[0-4]
3
c
+ +
3
0
[2-2]

Copyright ABRAHAM [2008] - All rights reserved

The 4th dimension of Time

is charge, the angular momenta associated with any planar mass-energy

Accelerate

$\overrightarrow{\mathbf{p}}$

increasing a systems's linear momentum

$F = m\Delta\mathbf{p} = m.\Delta v/t = ma$

creates acceleration

Decelerate

$\overleftarrow{\mathbf{p}}$

decreasing a systems's linear momentum

$F = m\Delta\mathbf{p} = m.\Delta v/t = ma$

creates decceleration

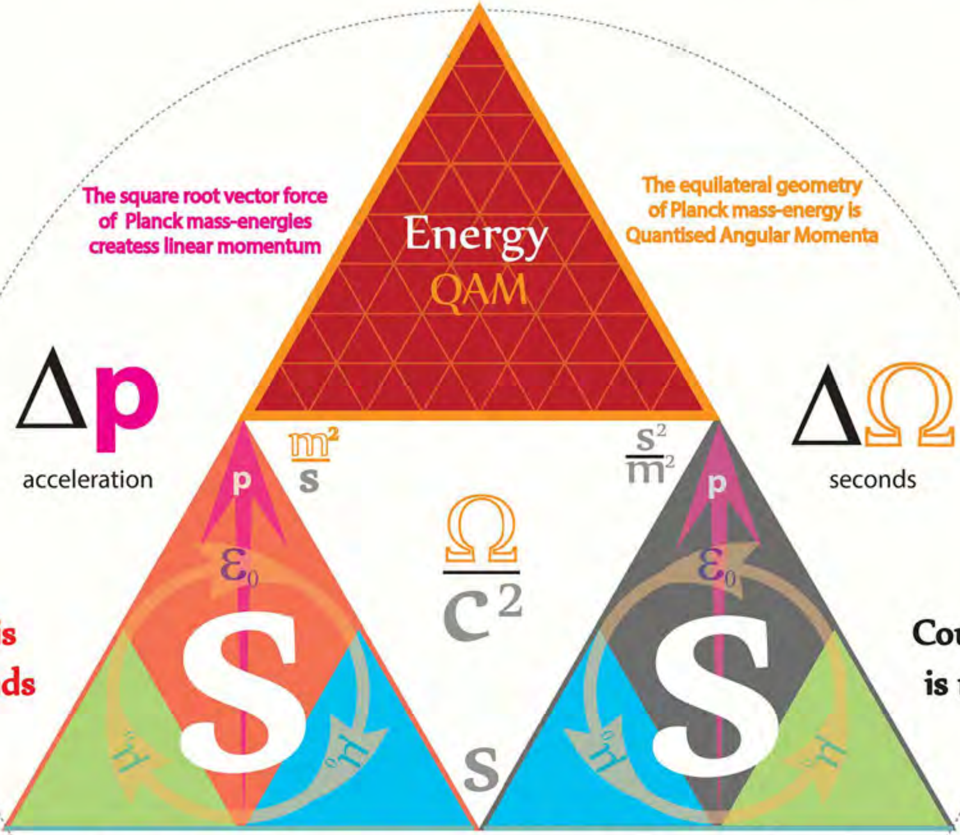

The square root vector force of Planck mass-energies createss linear momentum

The equilateral geometry of Planck mass-energy is Quantised Angular Momenta

$\Delta\mathbf{p}$

acceleration

$\Delta\Omega$

seconds

Clockwise charge is positive mass-seconds

Counter-clockwise charge is negative mass-seconds

Positive charge

Negative charge

the two 'arrows' of time in physics are a mathematical mis-interpretation of charge

Positive and Negative charges on opposite sides of Planck energy quanta,
the direct result of the equilateral geometry of all energy momenta
has been historically mis-represented in mathematics as
being two directional arrows of time

$\Delta\Omega$

Positive Time

Time is a measure of the changing angular momenta-energy (Planck geometry) of any given spatial region

$\Delta\Omega$

Negative Time

Tetryonics 62.11 - The 4th dimension of Time

CHARGE

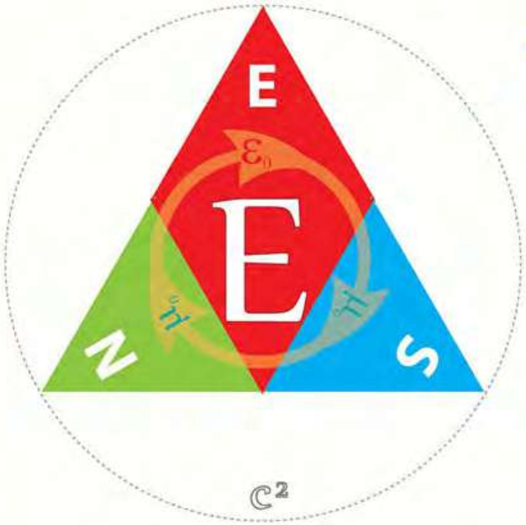
E
E
ε_0
N
S
c^2

all ideal inductive loops
resist changes to their
Energy levels

INERTIAL mass

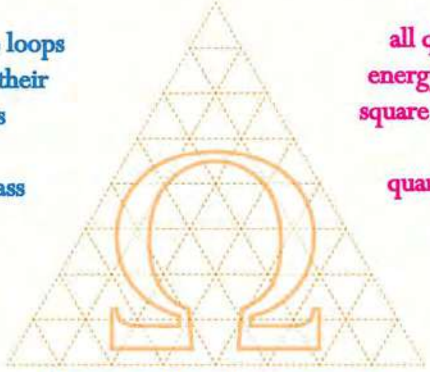

All quantised angular momentum
is an equilateral geometry

all quantised Planck
energy momenta posses
square root vector Forces

quantum of WORK

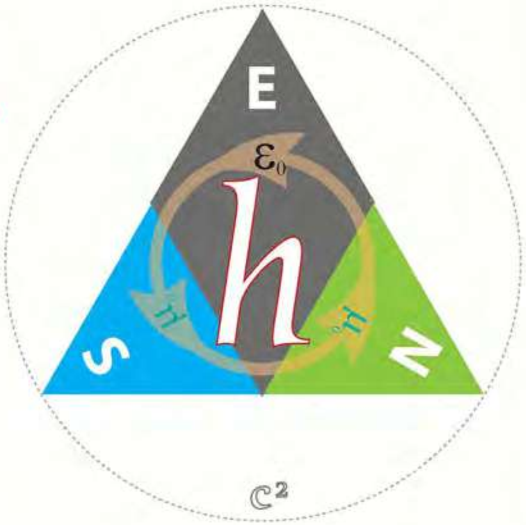
E
ε_0
h
S
N
c^2

vector velocity results from
linear momentum - the square root
of all scalar energies

$$kg\left[\frac{m^2}{s^2}\;\;\frac{s^2}{m^2}\right]$$

Scalar energy
per unit time

mass

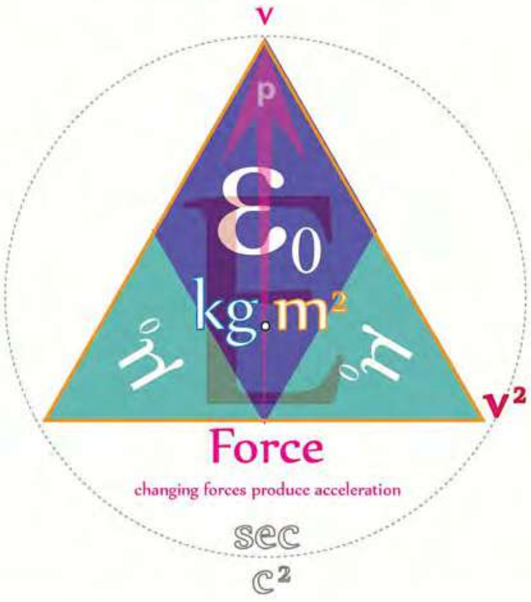
v
p
ε_0
kg.m²
v²
Force
changing forces produce acceleration
sec
c^2

$$kg\left[\frac{m^2}{s}\;\;\frac{s^2}{m^2}\right]$$

Planck quanta
per unit time

mass.seconds

Tetryonics 62.12 - Charge is mass.time

Charge is the equilateral geometry of Energy that gives form to all mass & Matter

Positive charged mass-energy

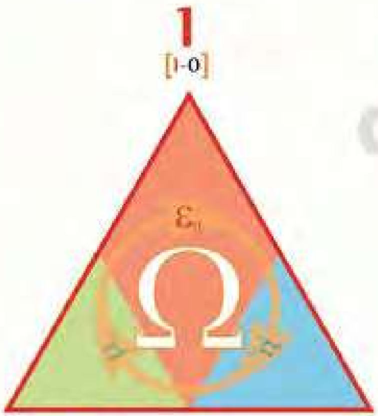

Clockwise inductive energy flux

charge

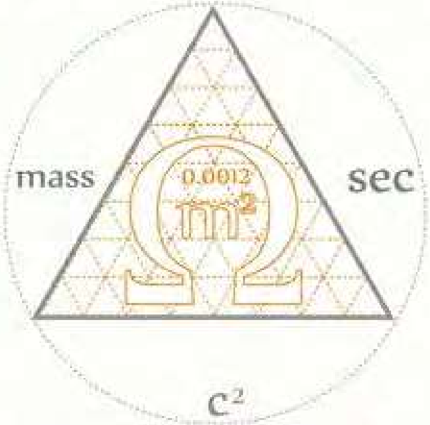

time

Negative charged mass-energy

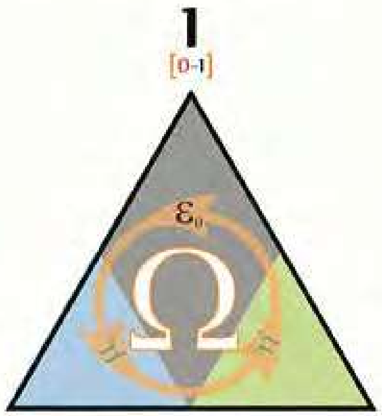

Counter clockwise energy flux

mass.seconds in physics is expressed as Coulombs [charge]

t

$$\frac{\Delta\Omega}{c^2}$$

$$\frac{m^2}{s}\,\frac{s^2}{m^2}$$

s

seconds

t

1.33518 e-20 s

Changes to nett quantised angular momenta
[charged mass-energies] within any
spatial co-ordinate system
forms the basis of time

ElectroMagnetic Charge is a quantum property resulting from the equilateral QAM geometry of mass-energy

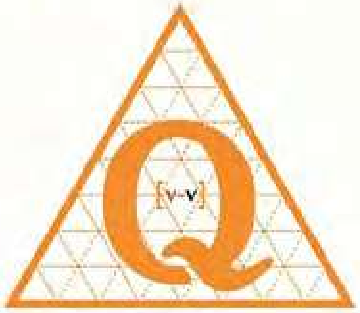

Two ElectroMagnetic charges are possible & both are created through the tesselation of EM energy momenta in ideal inductive loops

Copyright ABRAHAM [2008] - All rights reserved

The Quantum Arrow of Time

in physics is a measurement of changes to quantised angular momenta

Thermodynamic arrow of time

The Entropy of a system always increases

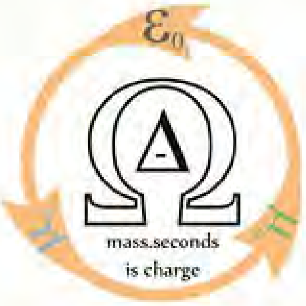
ε_0
mass.seconds
is charge

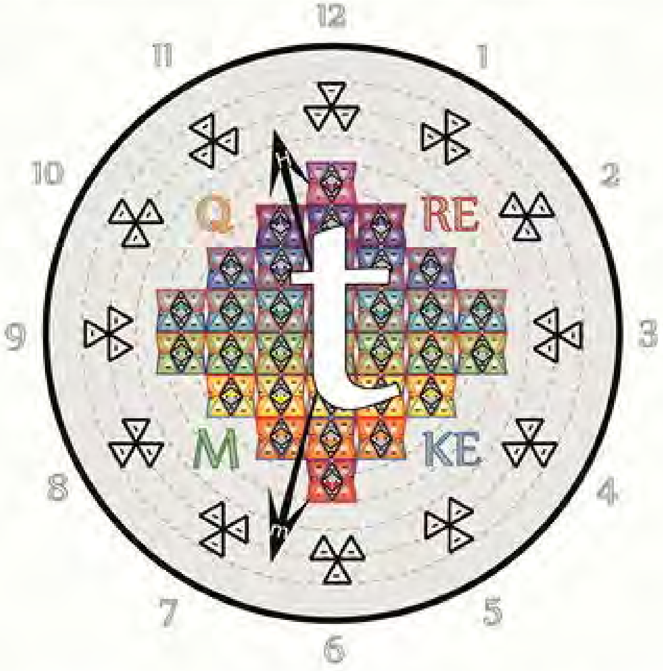
12
1
2
3
4
5
6
7
8
9
10
11
Q
RE
M
KE
t

Cosmological arrow of time

The Universe is evolving from a big bang to ?

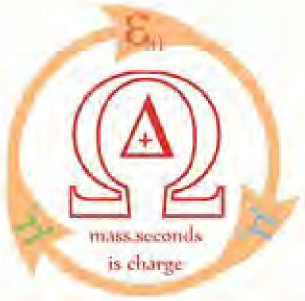
ε_0
mass.seconds
is charge

Radiative arrow of time

EM mass-energies rative outward from a point source

Causal arrow of time

A cause always precedes its effect

Tetryonic theory maps QAM as equilateral Planck geometries & reveals ΔQAM to be time
Time and the spatial dimensions it creates are mapped with radial & spherical co-ordinates

ΔmΩ

Decceleration

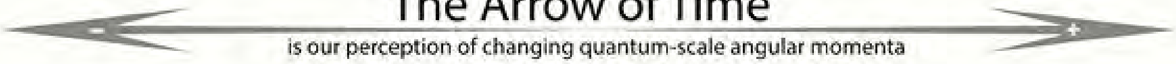

The Arrow of Time

is our perception of changing quantum-scale angular momenta

ΔmΩ

Acceleration

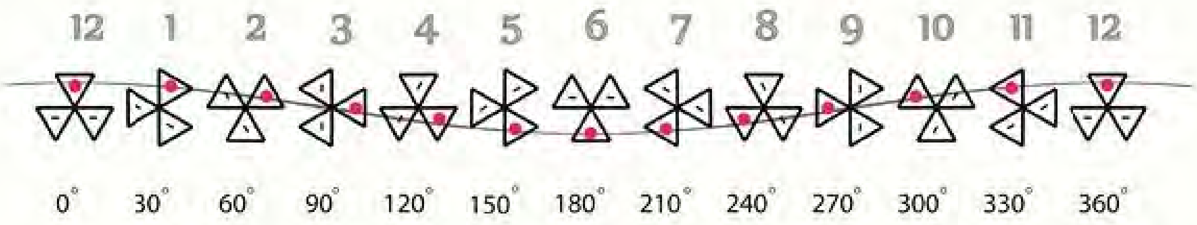
12
1
2
3
4
5
6
7
8
9
10
11
12
0°
30°
60°
90°
120°
150°
180°
210°
240°
270°
300°
330°
360°

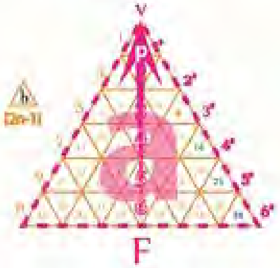

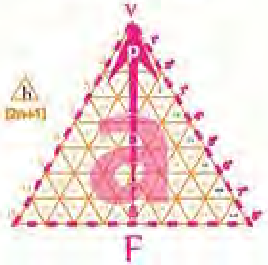

Time dilation

An accurate clock at rest with respect to one observer may be measured to tick at a different rate when compared to a second observer's own equally accurate clocks.
This effect arises neither from technical aspects of the clocks nor from the fact that signals need time to propagate, but from the nature of spacetime itself.

The changing quantised angular momenta of
all mass-energy-Matter densities in any spatial region
all impact on our measurement of time [ΔΩ]

increasing the mass of Matter
increases the stress energy tensor
[mass-energy gradient]

Stronger G fields

GR

Vacuum energy-Matter
pressure gradients

Weaker G fields

gravity is proportional to
the material displacement volume of mass-Matter
in a given spatial region of vacuum energy

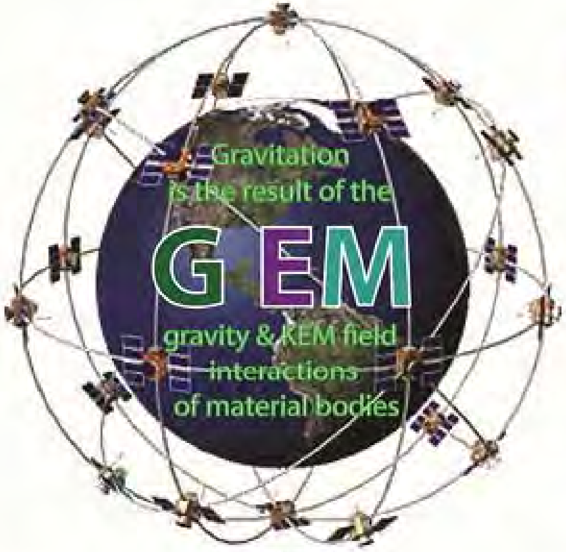

tests of GR using
photons are erroneous
tests of SR

accelerating material objects
increases the Planck quanta
[mass.QAM] in their KEM fields

Stronger KEM fields

SR

Kinetic EM energies
of motion

Weaker KEM fields

deccelerating material objects
decreases the Planck quanta
[mass.QAM] in their KEM fields

Any attempt to accurately model the full effections of Gravitation must include
a complete definition and differentiation of all mass-energy & Matter
along with the vacuum energy pressure gradient created by them

In the theory of relativity, time dilation is an actual difference of elapsed time between two events as measured by observers either moving relative to each other or differently situated from gravitational masses.

Copyright ABRAHAM [2008] - All rights reserved

Entropy in physical systems

is the result of changes to the net organisation of quantised angular momenta as it seeks equilibrium

Second law of thermodynamics

An isolated physical system, if not already in its own internal state of thermodynamic equilibrium, spontaneously evolves towards it.
In an isolated physical system, there is a tendency towards spatial homogeneity.

The Boltzmann equation was developed to describe the dynamics of an ideal gas

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} + \frac{F}{m}\frac{\partial f}{\partial v} = \frac{\partial f}{\partial t}$$

Ludwig Eduard Boltzmann

(February 20, 1844 – September 5, 1906)

$$S = k\cdot[\log(W)]$$

Boltzmann entropy formula

Tetryonic theory utilises the equilateral geometry of ideal quantum inductors to model EM energies

mass-energy radiator

normal statisitcal distribution of energies

Energy in all its forms seeks equilibrium

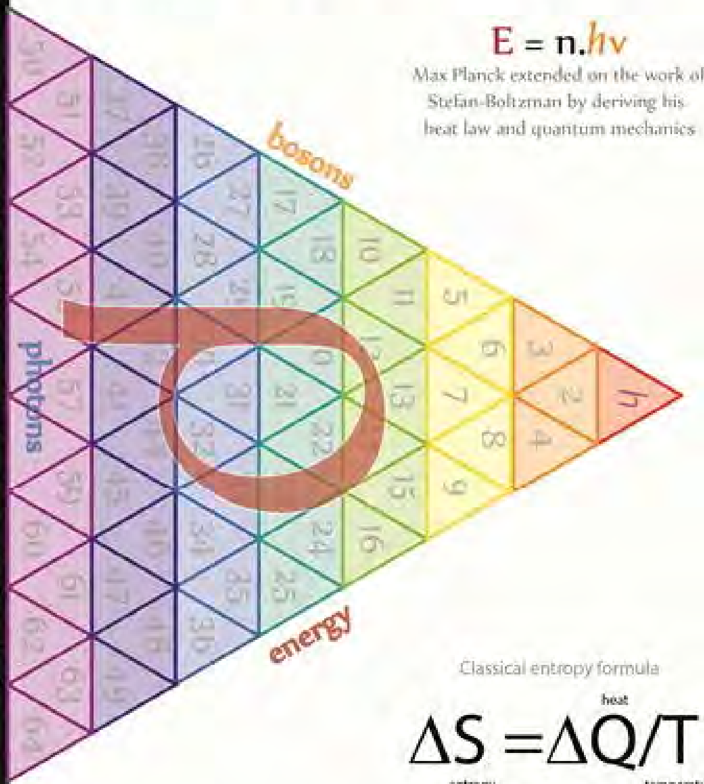

$$E = n.hv$$

Max Planck extended on the work of Stefan-Boltzman by deriving his heat law and quantum mechanics

Classical entropy formula

$$\underset{\text{entropy}}{\Delta S} = \overset{\text{heat}}{\Delta Q}/\underset{\text{temperature}}{T}$$

Energy always moves from higher energy organised systems to a lower dis-organised system

lower energy blackbody

A black body is an idealized physical body that absorbs all incident electromagnetic radiation

Jacob David Bekenstein

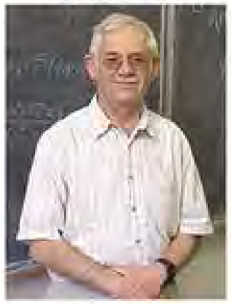

(born May 1, 1947)

Bekenstein-Hawking entropy formula

$$S = \frac{\pi A k c^3}{2hG}$$

still uses 3D [c³] classical cartesian co-ordinates

Stephen William Hawking

(born 8 January 1942)

Tetryonic theory unites classical and quantum mechanics with relativity through equilateral QAM

The Aether

Aether theories in physics propose the existence of a medium, the aether a space-filling substance or field, thought to be necessary as a transmission medium for the propagation of electromagnetic waves.

bosons, photons and Matter are simply geometric concentrations of EM energy in regions of Space

Spacetime is the QAM geometry of any measured spatial [LT] region

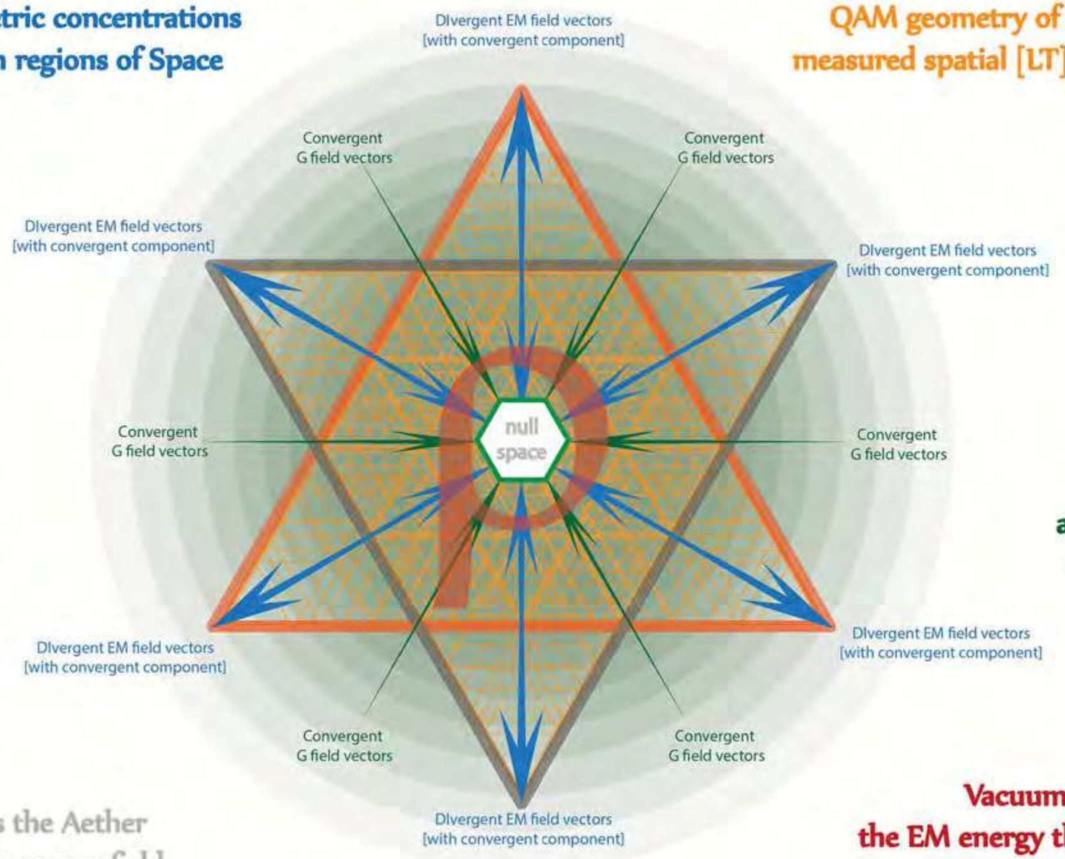
Divergent EM field vectors [with convergent component]
Convergent G field vectors
Convergent G field vectors
Divergent EM field vectors [with convergent component]
Divergent EM field vectors [with convergent component]
Convergent G field vectors
null space
Convergent G field vectors
Divergent EM field vectors [with convergent component]
Divergent EM field vectors [with convergent component]
Convergent G field vectors
Convergent G field vectors
Divergent EM field vectors [with convergent component]

Vast radiant energy fields exist throughout all of Space save inside Matter topologies

Matter geometries form perfect quantum Faraday cages excluding all energies thus creating nullspaces

The low level EM fields permeating the space between Matter forms a background vacuum energy density

The vacuum energy field acts upon the internal 'null space' volumes of all Matter creating a gravitational pressure gradient

Historically refered to as the Aether (or the Ether) this vacuum energy field has been the source of much contention

Vacuum Energy is the EM energy that permeates and fills any spatial region

Photons do not require the aether to propagate however the aether is permeated with 'weak' superpositioned EM fields

Tetryonics 63.01 - The Aether

The Luminiferous Aether

$$\frac{\rho}{m^2}$$

Photons are radiant 2D mass-energy geometries

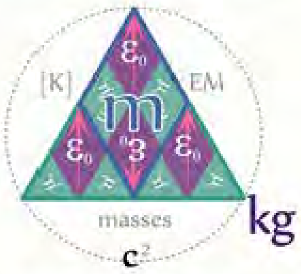

Vacuum energies are very long wavelength photons

Hendrik Lorentz created an electron/aether theory, in which he introduced a strict separation between Matter and the aether

In the late 19th century, luminiferous aether or ether, meaning light-bearing aether, was proposed as a medium for the propagation of light.

Newton's Opticks (1704) postulated an "Aethereal Medium" transmitting vibrations faster than light, by which light, is put into "Fits of easy Reflexion and easy Transmission"

Maxwell proposed a mechanical sea of molecular vortices to explain the transfer of ElectroMagnetic energies

Einstein wrote that one can actually speak about a "new aether", but one may not speak of motion in relation to that aether

In Tetryonics all EM radiation & Matter are revealed to be concentrations of geometric Energies seeking equilibrium in regions of Spacetime

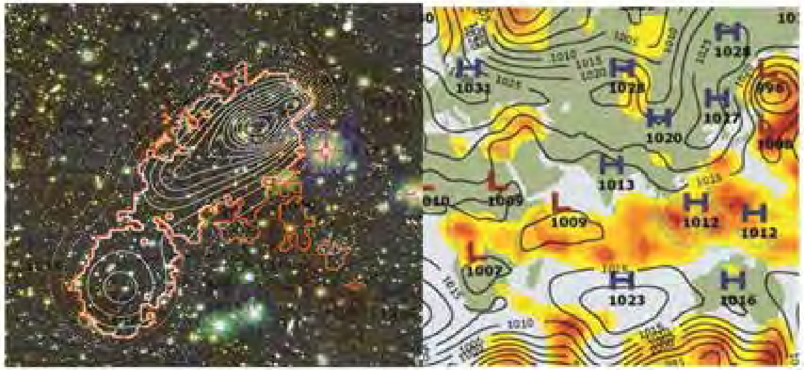

The aether does NOT facilitate the transmission of Energies throughout the Universe, it is the result of radiated [K]EM energies

Vacuum energies can be likened to the mean atmospheric pressure, with 2D mass-energies being HIGHER pressure 'gusts'

and

the nullspaces within 3D Matter creating a LOW pressure gradient [providing a driving mechanism akin to to that of Storms, Cyclones or Tornados]

$$\frac{\rho}{m^3}$$

All Matter are 3D mass-energy topologies containing nullspaces

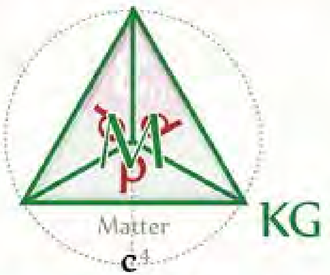

Planets, Stars and Galaxies contain high densities of mass-energies

Michaelson and Morley sought to detect the Relative motion between the Earth and aether by measuring the speed of light

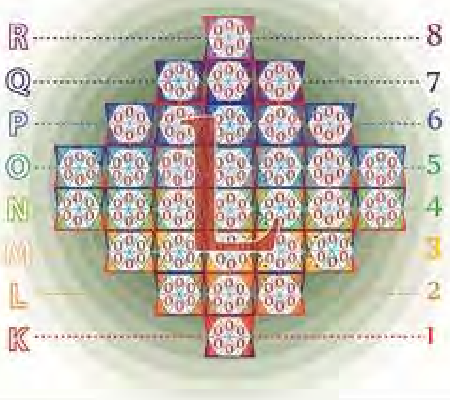

Vacuum Energy

Vacuum energy is an underlying background mass-energy that exists
in free space even when the space is devoid of matter

The Universe is Electric in nature and all Matter in it interacts via Gravito-Electro-magnetic [GEM] fields

Tetryonics shows that all Matter is quantised and these charged mass-energy quanta produce EM fields of varying strengths in turn radiating out from the source Matter into Space to form GEM fields

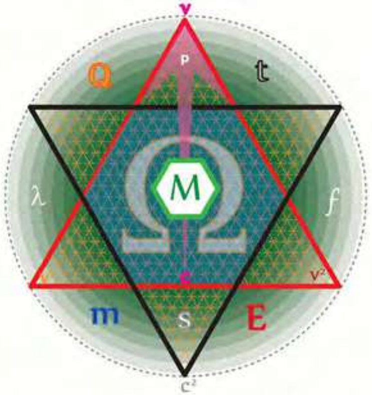

VE is not a 'quantum foam' as popularly depicted, it is a triangular Euclidean geometry of neutral mass-energies

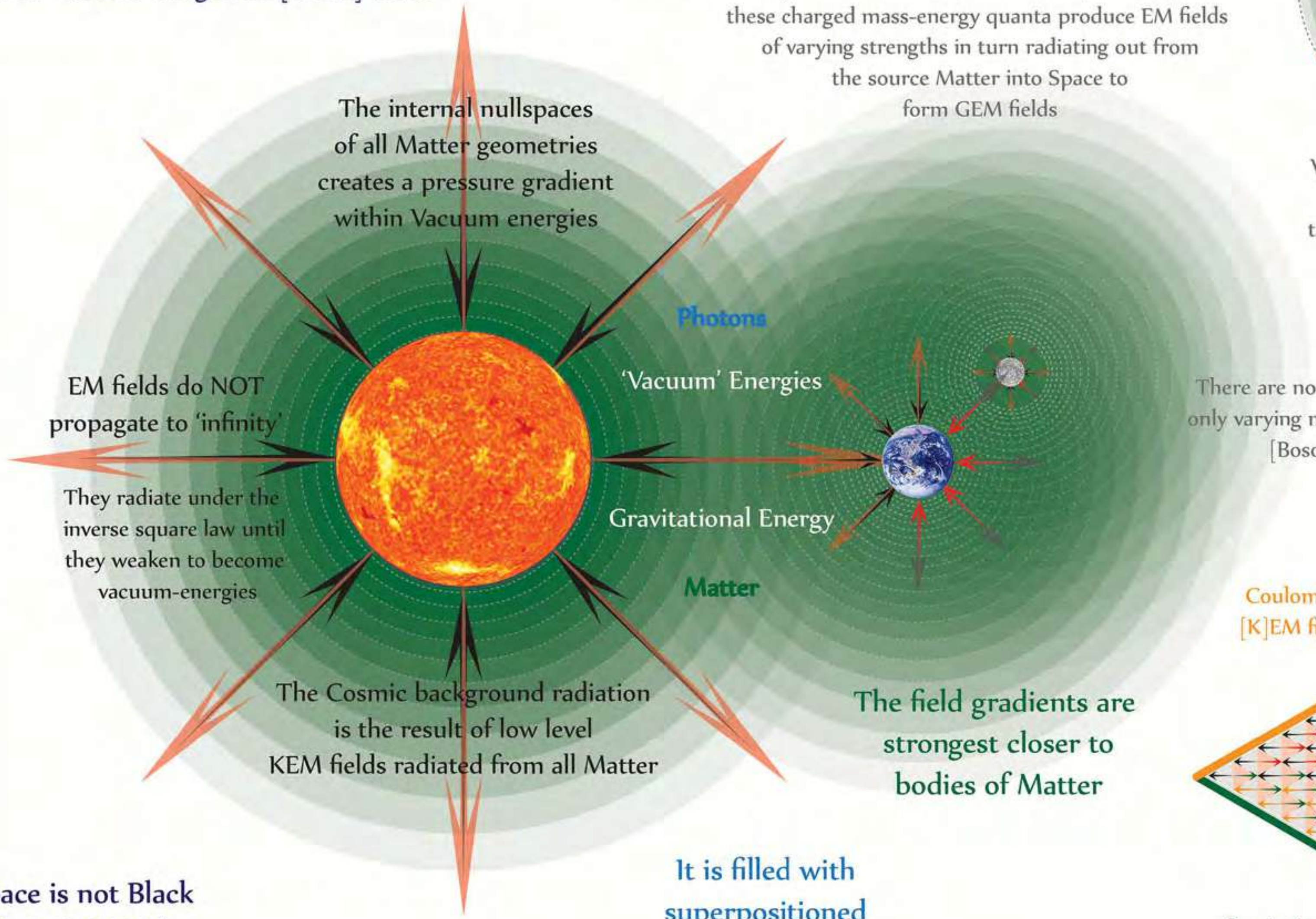

The internal nullspaces of all Matter geometries creates a pressure gradient within Vacuum energies

EM fields do NOT propagate to 'infinity'

They radiate under the inverse square law until they weaken to become vacuum-energies

The Cosmic background radiation is the result of low level KEM fields radiated from all Matter

The field gradients are strongest closer to bodies of Matter

There are no 'virtual particles' in free space only varying mass-Energy density geometries [Bosons, Photons & Matter]

Coulombic fields are DIVERGENT [K]EM fields with interactive forces

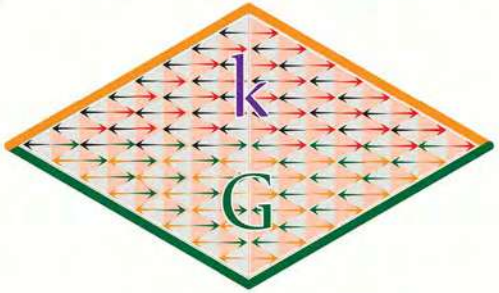

Gravitational fields are CONVERGENT
Gravity is always attractive

Space is not Black [it is INDIGO]

It is filled with superpositioned EM fields

Copyright ABRAHAM [2008] - All rights reserved

The Higgs Field

is another name for the Aether
(or vacuum energy field) that fills
free Space, its interaction with
inductive Tetryonic topologies
creates the inertial properties
observed in mass-Matter

Peter Higgs

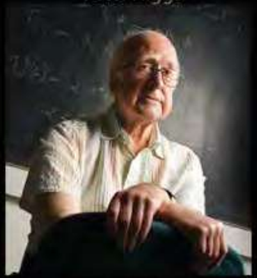

(born 29 May 1929)

The 'Higgs Boson' is
just a W boson by
another name

All Higgs vacuum fields
are equilateral mass-
Energy-momenta
fields

Free Space

is any spatial region
free of EM energy quanta
(nullspace within Tetryons)

Vacuum Energy

is the lowest denisty [longest wavelength]
electromagnetic mass-energy that
having diverged from Matter
radiates outwards to fill space
& create a pressure gradient

'Virtual particles'
do NOT exist

Copyright ABRAHAM [2008] - All rights reserved

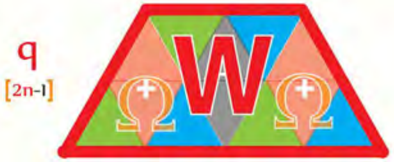

All Bosons have ODDn planar electromagnetic charge geometries

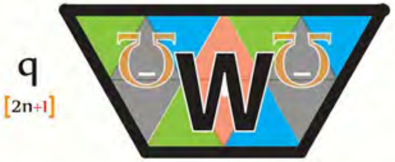

Inductive resistance to changes of the energy content of charged fascia through vacuum energy fields creates inertial mass

mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Electromagnetic mass
is a measure of the Energy
content of a 2D geometry
per unit of Time

m

**mass-energy
is a 2D geometric
property of 3D Matter**

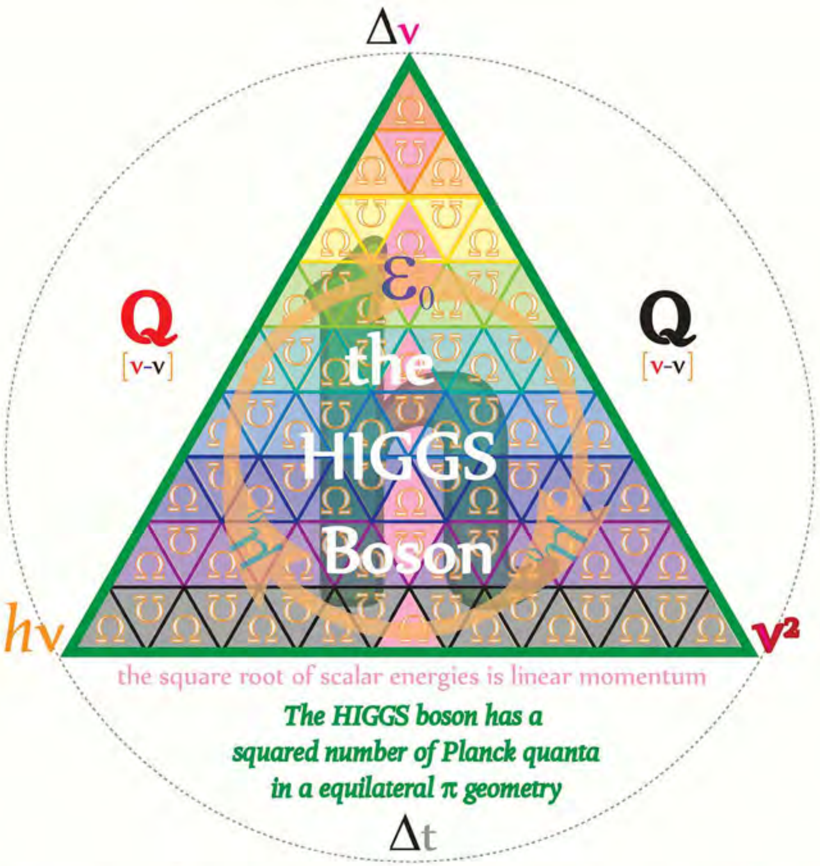

the square root of scalar energies is linear momentum

The HIGGS boson has a squared number of Planck quanta in a equilateral π geometry

The HIGGS boson is an alternative name for the 2D charged mass-energy fascia of all Matter
[whose inductive QAM quanta create the physical property of inertial mass]

$$F = ma$$

inertial mass is a measure of a body's resistance to changes in acceleration due to external forces

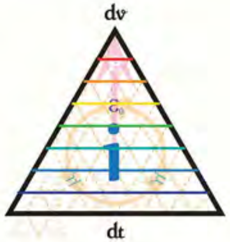

Charged Planck quanta form the geometry of of all mass-ENERGY-Matter
[each charge fascia contains equilateral mass-energy momenta]

TETRYONS

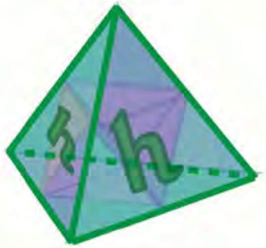

All Matter has 4n enclosed tetrahedral charge topologies

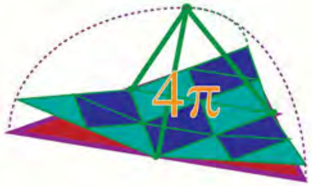

Matter Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

3D material Matter is a
EM standing-wave topology
whose mass is a measure of its
geometric Energy content

M

**Matter
is the 3D topology
created by mass-energies**

The energy momenta density of standing wave Matter

determines its inertial mass

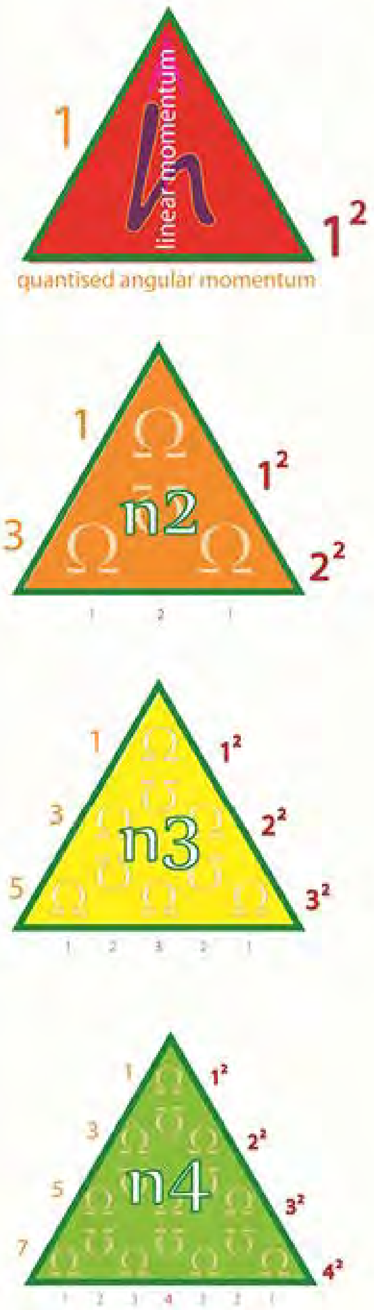
1
h
linear momentum
1²
quantised angular momentum
1
3
Ω
n2
1²
2²
1
3
5
n3
1²
2²
3²
1
3
5
7
n4
1²
2²
3²
4²

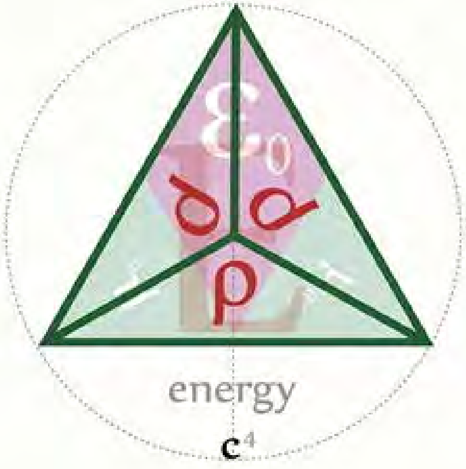
ε₀
d
d
ρ
energy
c⁴

m

mass

kg

M

Matter

KG

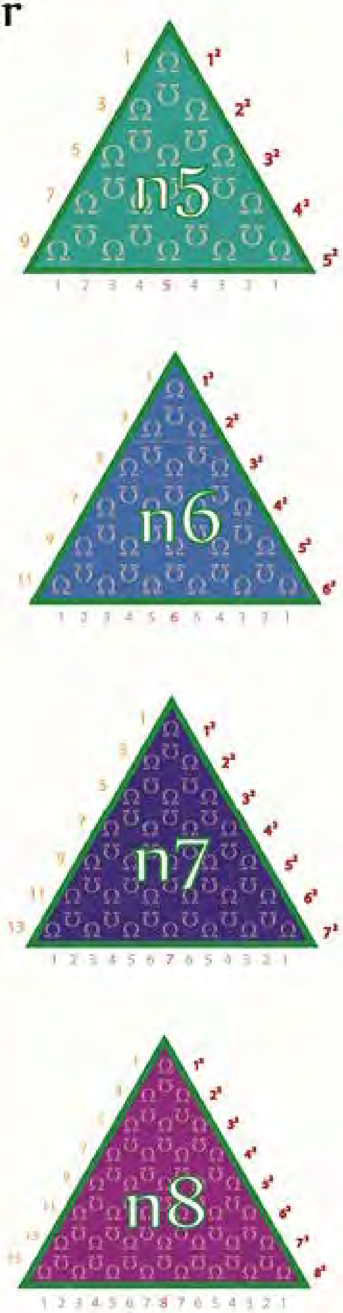
n5
n6
n7
n8
1²
2²
3²
4²
5²
6²
7²
8²

Spatial topology $\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$ EM mass-energies

Matter Planck quanta mass velocity

de Broglie Wavelengths

$\frac{\rho}{c^2}$

mass

m

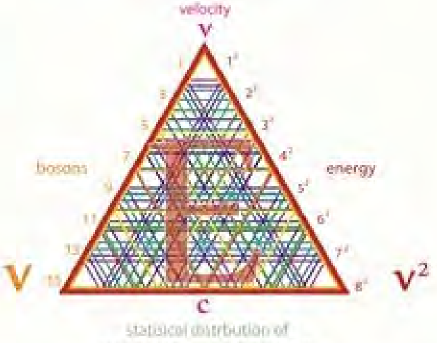
velocity
v
bosons
energy
v
v²
c
statistical distribution of

energy momenta

$\frac{\rho}{c^4}$

Matter

M

Compton Frequencies

The mass-energy content of Matter is contained entirely within its charged fascia topology

Copyright ABRAHAM [2008] - All rights reserved

3D Nullspaces

Planck mass-energy quanta clothe
only the topologies of Matter

**3D EM fields with volume
are measurements of
Energy per unit Time squared
[Matter]**

$$T\pi$$

**Nullspace is proportional to the
internal volumes of Tetryons**

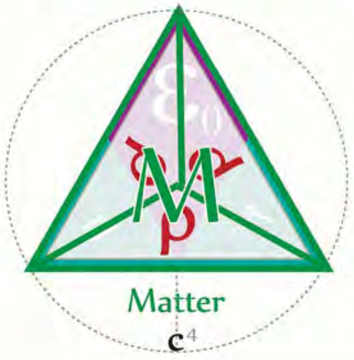

**2D planar EM fields
are measurements of
Energy per unit Time
[EM mass]**

$$[m\Omega v^2]$$

**The mass-energy densities [fascia] of
Matter do NOT influence nullspaces**

Inside the Tetrahedral topologies of all Matter exists the only true vacuum

Any radiant energy trying to enter
the interior volume of a Tetryon
will be absorbed into the fascia,
increasing its Tetryonic mass
(or increasing its KEM field energy)

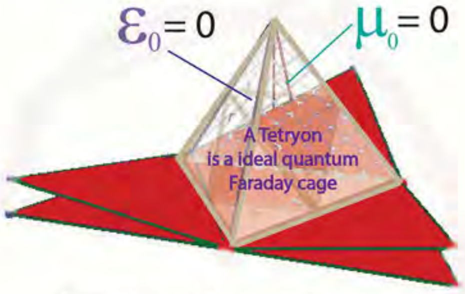

A volume devoid of any possible energy quanta
[ie The True vacuum]

The interior volume of Tetryons has the
lowest energy density/pressure possible
and is accumulative as Matter forms
increasingly complex states, creating
regions of low energy density within
the nuclei of all Matter

*Outer space is considered a high-quality vacuum,
with the equivalent of just a few hydrogen atoms
per cubic meter on average. However even it is
permeated with superpositioned EM fields
of vacuum mass-energies*

$$\frac{\rho_0}{T}$$

*A true vacuum is a spatial volume
that is completely devoid of EM mass-energies,
such that its energy density/pressure is much lower
than of the surrounding energy density/pressure
[a nullspace]*

Copyright ABRAHAM [2008] - All rights reserved

[x,y] $\Delta\Omega$

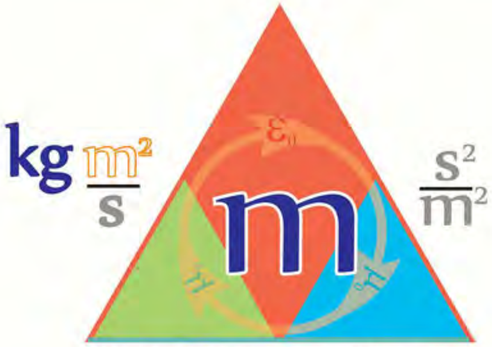

Positive charges in 2D space
are classically modelled as
clockwise energy fluxs

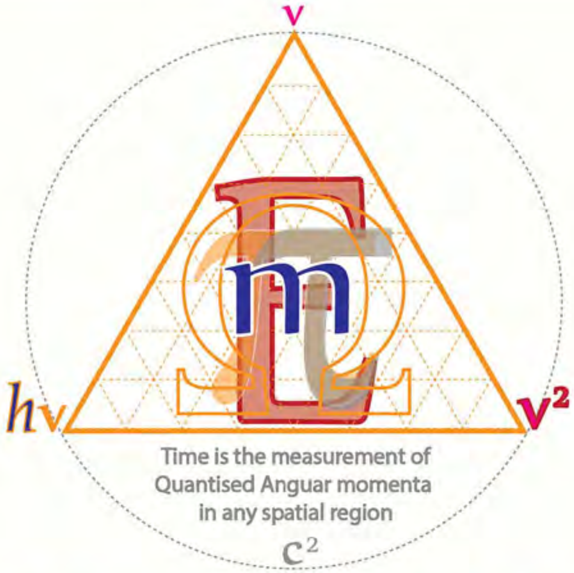

Time is the measurement of
Quantised Anguar momenta
in any spatial region

[x,y] $\Delta\Omega$

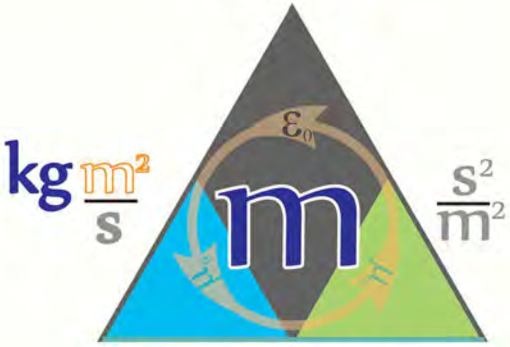

Negative charges in 2D space
are classically modelled as
anti-clockwise energy fluxs

Charged Planck mass-energies provide the scaffolding of 3D Matter

mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

m

mass

[x,y,z] $\Delta\Omega$

3D Space and Time

M

Matter

Matter Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

2D
mass geometries
are a property of
3D Matter

2 Dimensional mass-energies

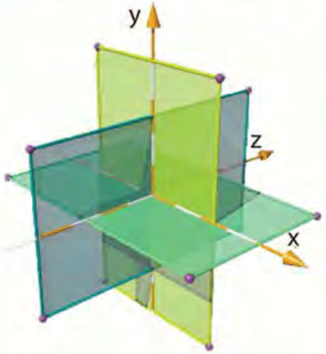

3 Dimensions of Space

kg

radiant
2D mass-energy momenta

EM Field Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

impedance mass velocity

creates 3D standing-wave
Matter

KG

2D charged fascia

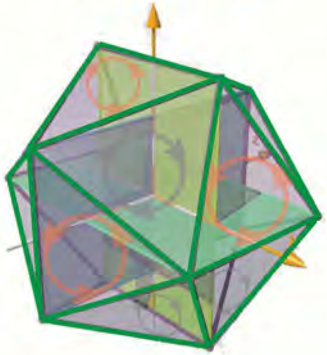

3D Space + Time

3D
Matter topology
is not a property of
2D mass-energy

Copyright ABRAHAM [2008] - All rights reserved

The impedance of any spatial region

is determined by its scalar energy content

Positive Matter

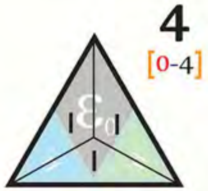

Negative Matter

$$\frac{m}{hv^2}$$

$$\frac{mc^2}{v^2}$$

$$\frac{mc^2}{h}$$

$$\frac{v^2}{c^2}$$

1.112650056 e-17

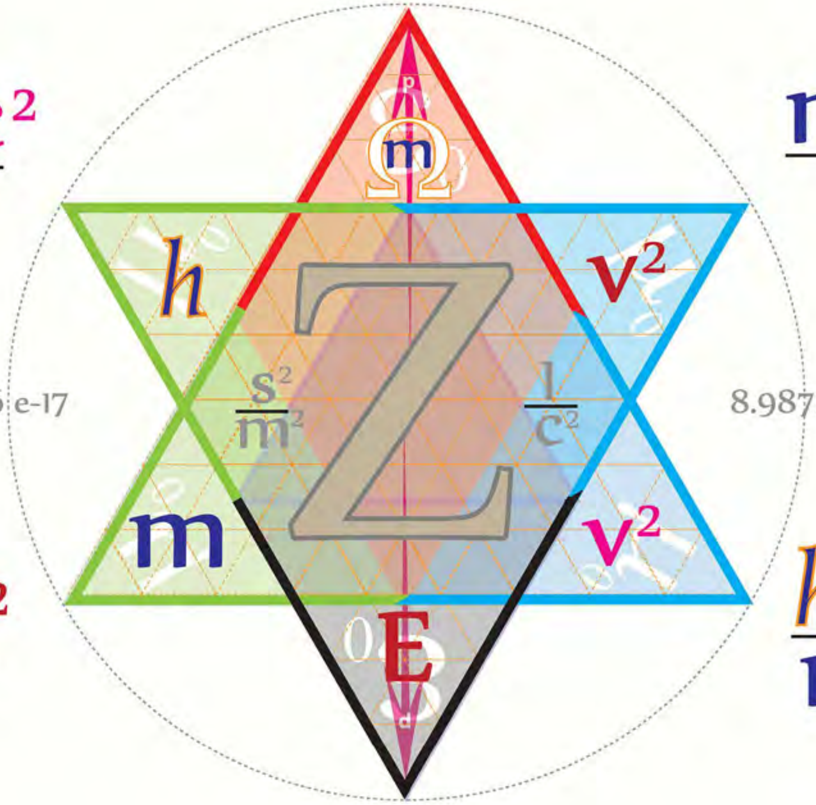

$8.987551787\ e16^{-1}$

$$\frac{h}{m}$$

$$\frac{hv^2}{c^2}$$

$$\frac{hv^2}{m}$$

Neutral Matter

Neutral Matter

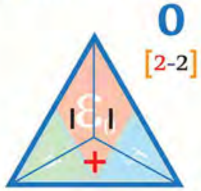

111.265 ohms

The impedance of any medium determines the velocity of propagation for mass-energy momenta through it

Copyright ABRAHAM [2008] - All rights reserved

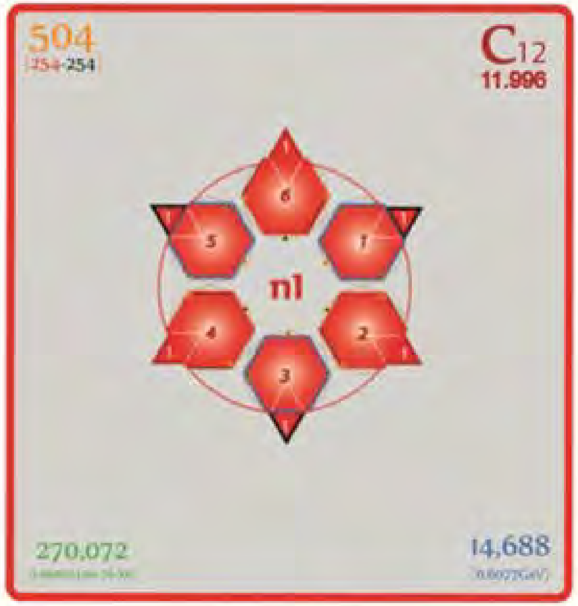
504
254-254
C12
11.996
6
5
1
n1
4
2
3
270,072
14,688

EM mass vs. Matter

Often a source of confusion,
Tetryonics provides a rigorous
definition of both EM mass & Matter

6.022141579 e23
[Avagadros number]

**1 ton of Carbon & 1 ton of Lead
have the same mass-energies**

**But they have differing Matter topologies
that create a differing energy densities
for any spational region of measurement**

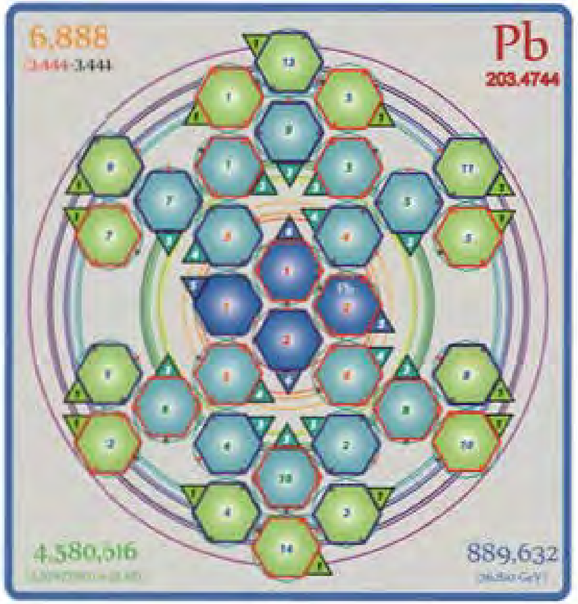
6,888
3444-3444
Pb
203.4744
4,580,616
889,632

**1 mole of Carbon atoms
[11.996 grams]**

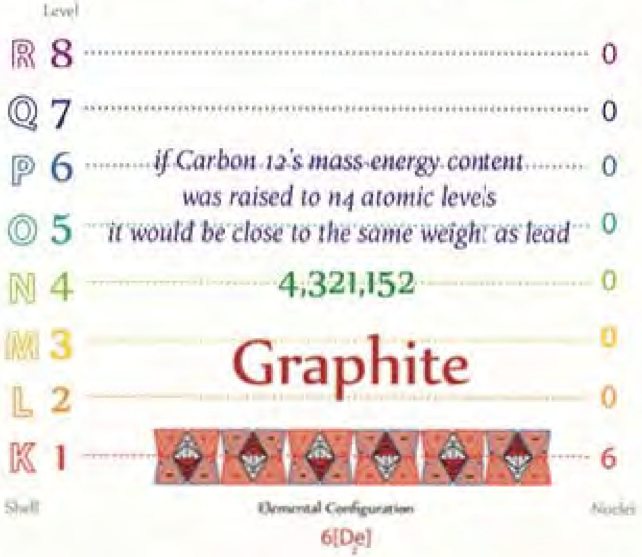
Level
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
0
0
0
0
0
0
0
6
4,321,152
Graphite
Shell
Elemental Configuration
Nuclei
6[De]

*if Carbon 12's mass-energy content
was raised to n4 atomic levels
it would be close to the same weight as lead*

*In terms of charged Matter topology
Carbon12 displaces 1/14th the volume of Lead*

504
Carbon12 atom
270,072

1 : 17

4,580,616
Lead atom
6,888

*In terms of mass-ENERGY-Matter densities
Lead is 17 times denser than Carbon
as a result of the mass quanta per charge fascia*

**1 mole of Lead
[203.4744 grams]**

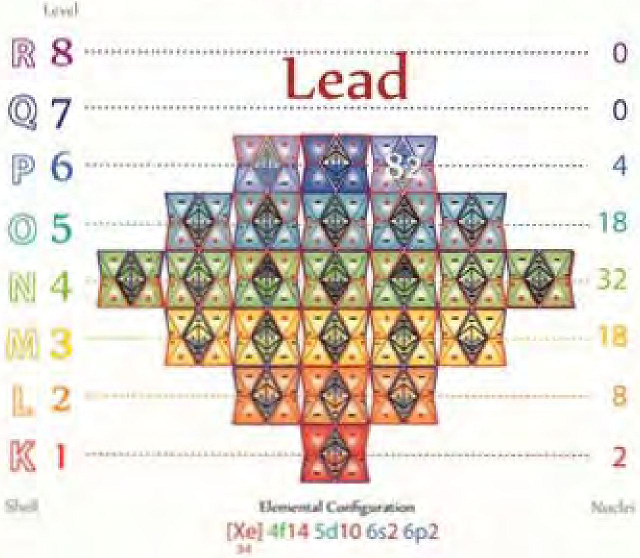
Level
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
Lead
82
0
0
4
18
32
18
8
2
Shell
Elemental Configuration
Nuclei
[Xe] 4f14 5d10 6s2 6p2

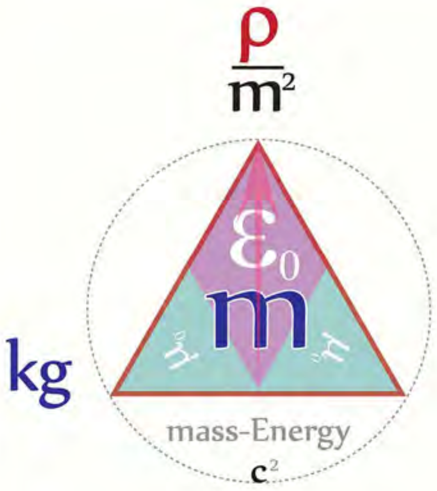

EM mass-energy geometries

are 2D Euclidean EM fields posessing energy momenta

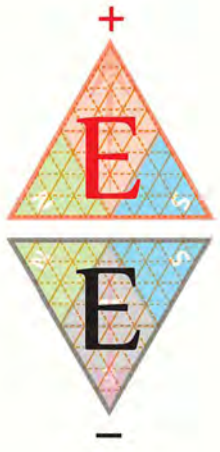

Charge

$$\frac{1\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Bosons

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Photons

$$\frac{2\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

mass

$$\frac{n\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

m

ENERGY

$$n\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

M

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Equilateral mass-energy geometries & Matter Topologies are measured in radial spatial-time co-ordinate systems

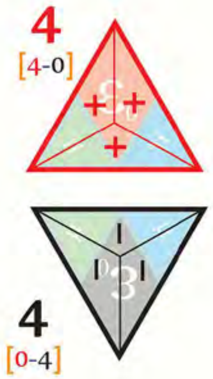

Tetryons

$$\frac{4\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Fermions

$$\frac{12\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Baryons

$$\frac{36\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Deuterons

$$\frac{84\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

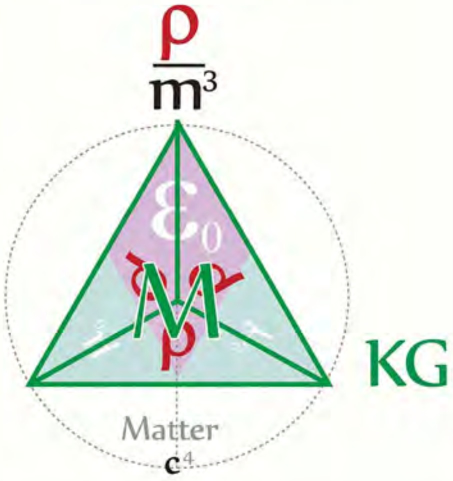

Matter Topologies

are 3D standing-wave [4nπ] mass-energy geometries

mass-energy-Matter quanta

are all defined by charged geometries & topologies

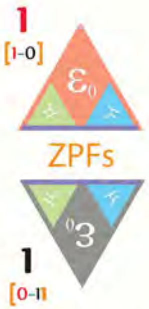
1
[1-0]
ε₀
ZPFs
3
1
[0-1]

Bosons (v)

Bosons

Planck quanta

$\frac{\text{ODD}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$

mass velocity

charges

1π ZPF v

$\downarrow$

ODDπ boson v

Inductive charges

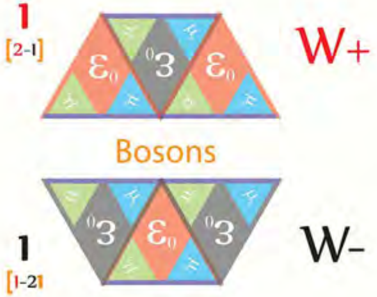
1
[2-1]
ε₀
3
ε₀
W+
Bosons
3
ε₀
3
1
[1-2]
W-

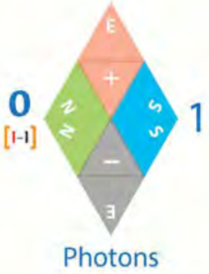
E
+
0
[1-1]
N
N
S
S
1
-
E
Photons

Photons(*f*)

EM waves

Planck quanta

$\frac{\text{EVEN}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$

mass velocity

mass-energies

2π photon $2v^2$

$\downarrow$

EVENπ EM wave $2v^2$

EM waves

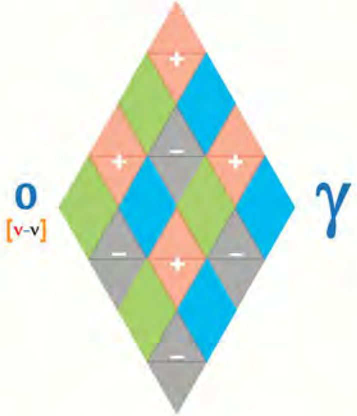

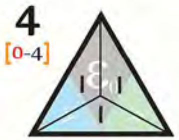

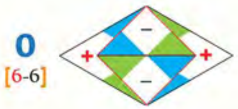

Tetryons (n)

Matter

Planck quanta

$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$

mass velocity

mass-Matter

4π tetryon n

12π lepton n

24π meson n

36π Baryon n

Matter particles

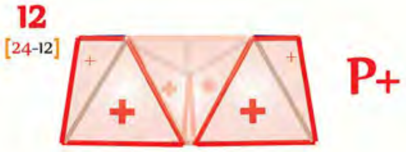

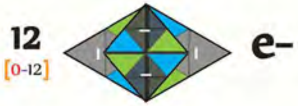

Copyright ABRAHAM [2008] - All rights reserved

Differentiating mass & Matter

geometries topologies

m
mass

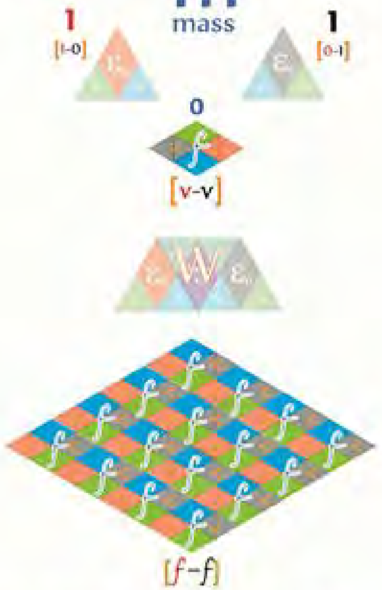

All modern physical theories including relativity fail to define and differentiate between mass & Matter

M
Matter

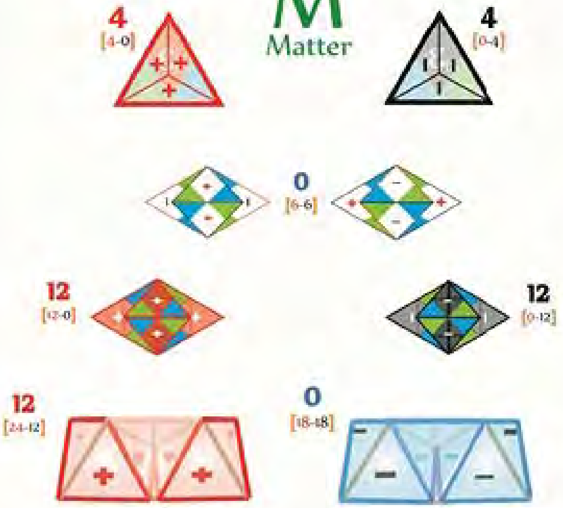

kg

Charged Planck quanta
7.376238634 e-51 kg

Photons
1.475247727 e-50 kg

EM waves
5.900990907 e-50 kg

nπ
mass

radiant mass-energies

Equilateral energy momenta

Tπ
Matter

standing-wave Matter

KG

Neutrino
8.851486361 e-50 kg

Lepton
8.851486361 e-31 kg

Baryon
1.6596536693 e-27 kg

geometry

Topology

Increasing a particle's speed increases its Kinetic EM mass-energies (rest Matter's EM mass-energy content is velocity invariant)

Charge
$$\frac{1\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Photons
$$\frac{2\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Bosons
$$\frac{\text{ODD}\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

EM waves
$$\frac{\text{EVEN}\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

All Matter has EM mass-energies not all EM mass-energies form Matter

Tetryons
$$\frac{4\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Fermions
$$\frac{12\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Baryons
$$\frac{36\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Deuterons
$$\frac{84\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

$$E = mv^2$$

E
Energy
$$kg\,\frac{m^2}{s^2}$$

6.629432351 e-34 J

$$\boxed{\underset{\text{Energy}}{E} = n\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]}$$

ENERGY

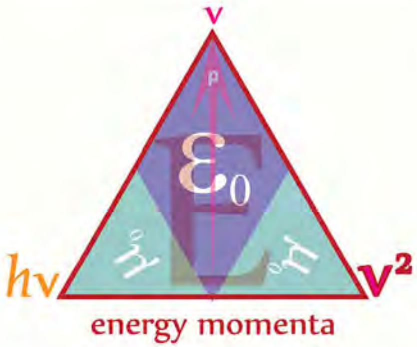
v
p
ε_0
hv
v²
energy momenta

the proagation velocity of energy is determined by the impedance of any spatial region

E=hv²

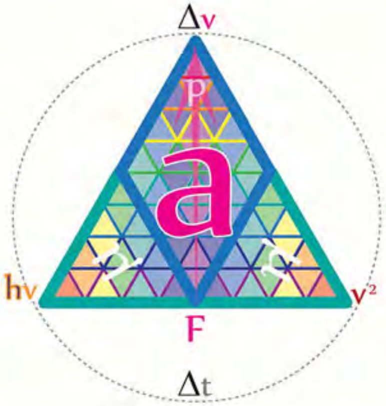
Δv
p
a
hv
v²
F
Δt

F=ma

the scalar impedance of space

$$\frac{s^2}{m^2}$$

$$\boxed{\begin{array}{c} \varepsilon_0 \quad \times \quad \mu_0 \\ \textbf{Electric permittivity} \ \& \ \textbf{Magnetic permeability} \\ 1.11265\ \text{e-}17\ \frac{s^2}{m^2} \end{array}}$$

$$\frac{1}{c^2}$$

ε_0
$\frac{1}{c^2}$
EM field
permitivity & permeability
c²

the Forces created by energy momenta in any spatial region are related through the scalar constant of mass

EM mass

$$m = \frac{E}{c^2}$$

m
mass
kg

$$\boxed{\underset{\text{mass}}{n\pi}\left[\overset{\text{field impedance}}{\underset{\text{ElectroMagnetic}}{\left[\varepsilon_0\mu_0\right]}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]}$$

7.376238376 e-51 kg

ε
m
EM mass
c²

Copyright ABRAHAM [2008] - All rights reserved

The energy-momenta of Matter

$$\mathbf{p} = mv$$

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

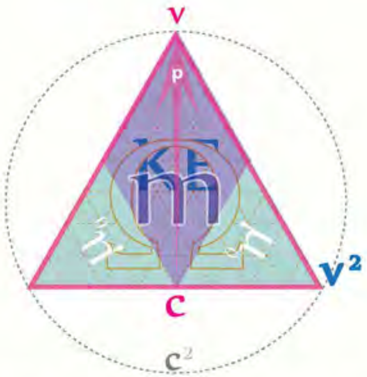

The Energy comprising all KEM fields is proportional to the particle's velocity

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Scalar correction factor

$$\mathbf{p}^2 = E = mv^2$$

$$\mathbf{p} = T\pi\left[\left[m\underset{\text{momenta}}{\overset{\text{Energy}}{\Omega}}v\right]\right]$$

As the velocity of
a particle increases
so does its KEM field's
mass-energy momentum

$$mv = \mathbf{p} = h\nu$$

$$kg\,\frac{m}{s}$$

$$\sqrt{6.629432351\text{ e-}34\text{ J}}$$

$$4n\pi\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v}\right]$$

2.5476847 e-17 kg m/s

The linear momentum of mass-energies in standing-wave rest Matter topologies propagates at c

$$\mathbf{p}_M = 4\pi\left[mv\right]$$

The rest mass of a particle with NO motion is the total EM mass-energy of the particle

Its linear four-momentum is the square root of its scalar energies

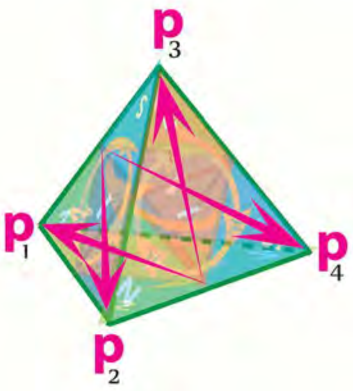

rest mass-Matter

In standing-wave Matter,
where there is NO external forces
acting on the particle, all internal momenta
act in opposition and add to ZERO

all fascia

$$\vec{\mathbf{p}}_0 = \sum \vec{\mathbf{p}}$$

Gravitational mass-Matter quanta

gravity pulls on things proportional to their mass-Matter content

$$F = -G\frac{M_1 M_2}{r^2}$$

Radiant mass-energy geometries

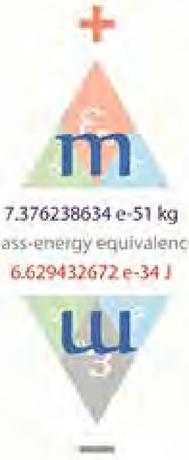
+
m
7.376238634 e-51 kg
mass-energy equivalence
6.629432672 e-34 J
ω
−

All 3D Tetryonic topologies create displacement Matter volumes that in turn generate KEM fields as a result of the motion of their charge fascia

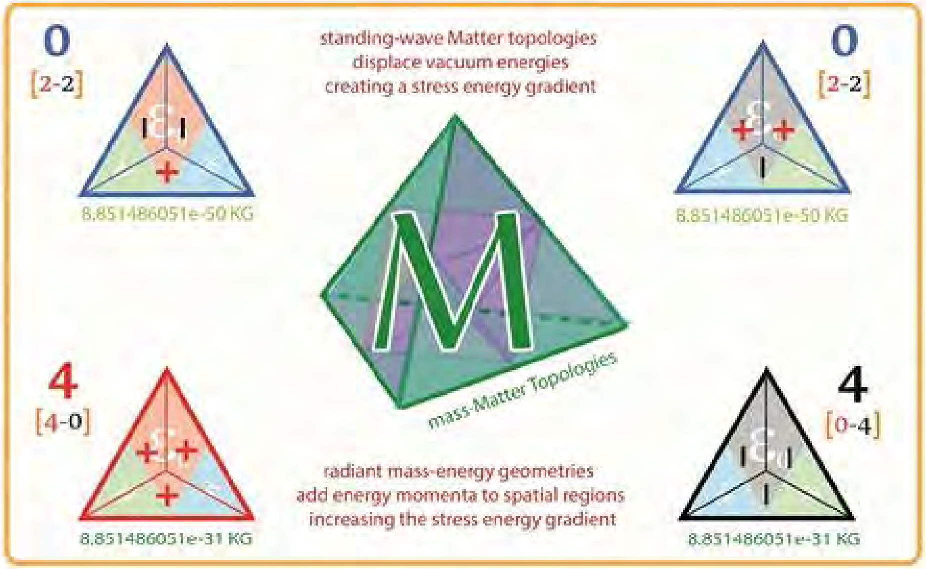
0
[2-2]
8.851486051e-50 KG
0
[2-2]
8.851486051e-50 KG
M
mass-Matter Topologies
4
[4-0]
8.851486051e-31 KG
4
[0-4]
8.851486051e-31 KG

standing-wave Matter topologies displace vacuum energies creating a stress energy gradient

radiant mass-energy geometries add energy momenta to spatial regions increasing the stress energy gradient

Standing-wave Matter Topologies

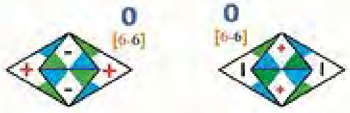

8.851486051e-50 KG

Matter particles can have the same Tetryonic topologies but possess a differing charged mass-energy momenta content

8.851486051e-31 KG

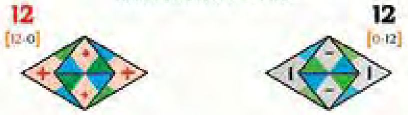
12
[12-0]
12
[0-12]

Gravitational Attraction can be modelled as a dipole attraction of opposite charged mass fascia of Matter topologies through its differing energy field interactions

Gravitation attraction is a strictly UNI-DIRECTIONAL CONVERGENT FORCE

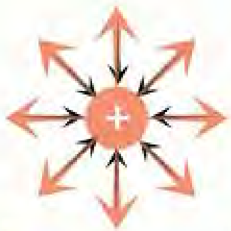

$$k = 8.99\text{e}9\,\frac{N\,m^2}{C^2}$$

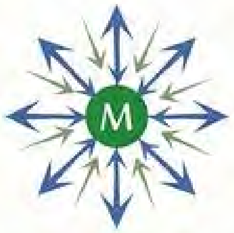

$$G = 6.67\text{e-}11\,\frac{N\,m^2}{KG^2}$$

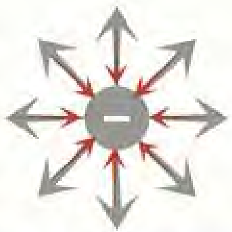

KEM field energy momenta produce a BI-DIRECTIONAL INTERACTION FIELD

Gravitation attraction is 20 orders of magnitude weaker

KEM field energy momenta produce a BI-DIRECTIONAL INTERACTION FIELD

Copyright ABRAHAM [2008] - All rights reserved

Stress energy gradients

Energy's highest density form is 3D Matter [comprised of planar 2D EM mass-energies]
whilst its lowest pressure form is the vacuum energies created by radiated mass-energies filling space

ENERGY in all its forms seeks equilibrium

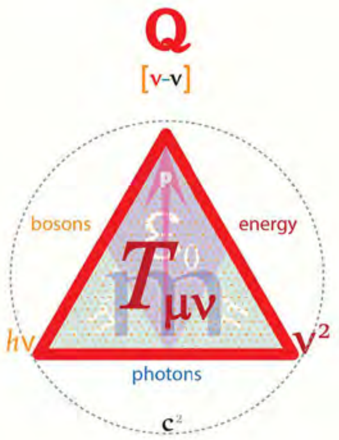

Where Newton's Law of Gravitation is modelled on geometric means

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

Inverse Square Law

$$F = -G\,\frac{M_1 m_2}{r^2}$$

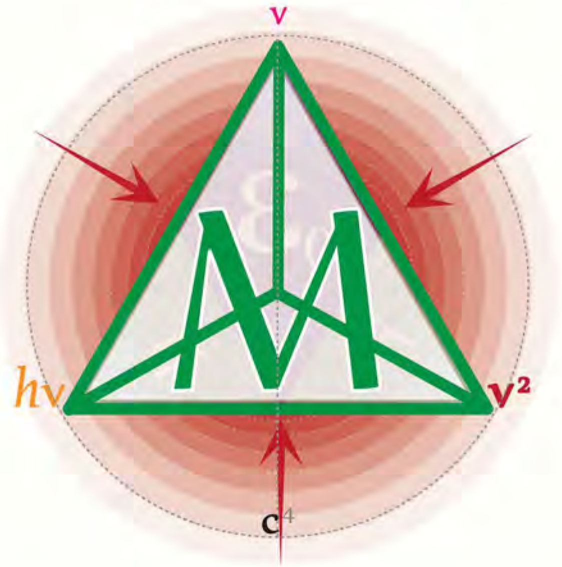

Like any pressure gradient, ENERGY always flows from areas of highest density to regions of low density seeking to reduce 'pressure' differentials and reach a state of equilibrium.

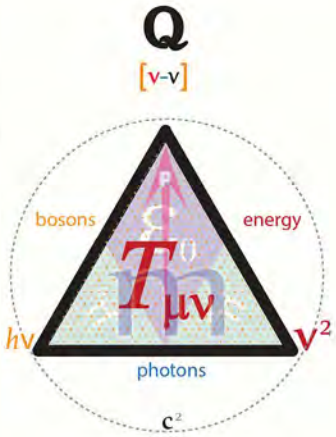

Einstein's General Relativity is modelled on energy pressure

$$PV = nRT$$

Matter tells Spacetime how to bend
Spacetime tells Mater how to move

$$R_{\mu\nu} - \tfrac{1}{2}R\,g_{\mu\nu} = \frac{8\pi G}{c^4}\,T_{\mu\nu}$$

mass-Matter topologies create stress energy differentials in the vacuum energies of space

Copyright ABRAHAM [2008] - All rights reserved

Gravitational Matter & weight

is a measure of the amount of mass-energy in Matter

The gravitational mass of a body determined by its response to the force of gravity

mass-Matter

Matter

Planck quanta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

The gravitational acceleration of a body in any gravity field is uniform and proportional to its mass-Matter

$$w = F = M.g$$

electrons

Planck quanta

$$\frac{12\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energy-Matter

8.8514586361 e-31 KG

48

[24-24]

charge topology

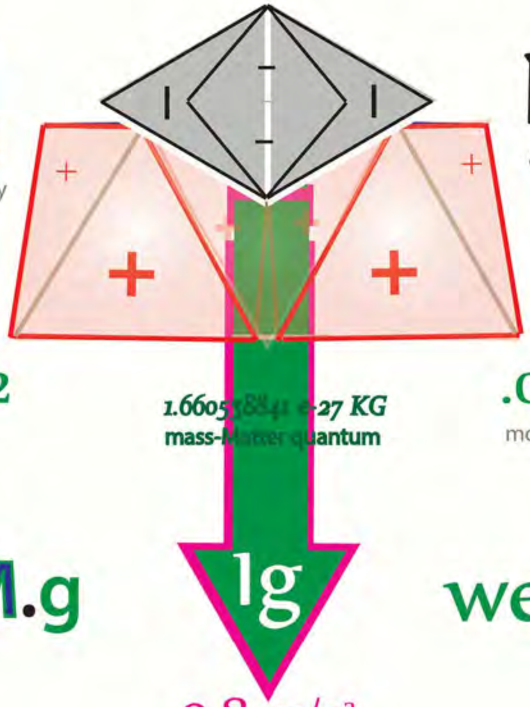

N_A

Avagadro

Baryons

Planck quanta

$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energy-Matter

1.659653693 e-27 KG

22,512

mass-Matter

1.660538841 e-27 KG

mass-Matter quantum

.001 KG

molar mass

w=M.g

weight

9.8 m/s²

$$F=Ma_g$$

= [.001] x [9.8]

= .0098 N

KG

g

Force

$$\frac{\left[kg\,\frac{m}{s^2}\right]}{KG}$$

Zero Point Fields

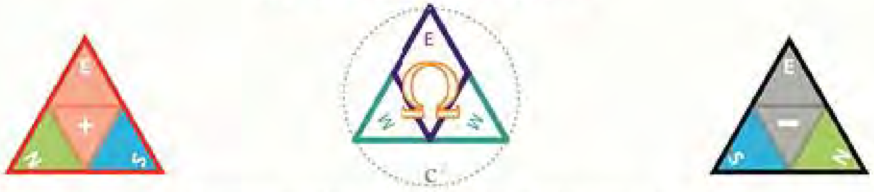
E
+
N
S
E
Ω
M
M
C²
E
−
S
N

ZPF quanta charges are the opposite sides of the same 'quantum coin'

ZPF quanta have unidirectional energy momenta orientations

KEM fields

N-S magnetic moment vector

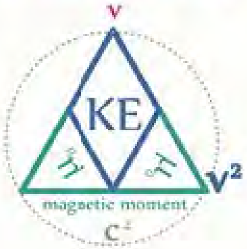
v
KE
M
M
v²
magnetic moment
C²

S-N magnetic moment vector

standing-wave mass-Matter

As the velocity of Matter increases its relativistic mass-Energy quanta increase as a result of the associated momentum, kinetic energy and magnetic moment

standing-wave mass-Matter

in addition to their standing-wave topologies mass-Matter also posesses

radiant mass-energies

Gravity & EM fields

All energy fields are velocity dependent and combine to form vectorial K-EM fields

All charged Matter posseses bi-directional interactive radiant EM fields

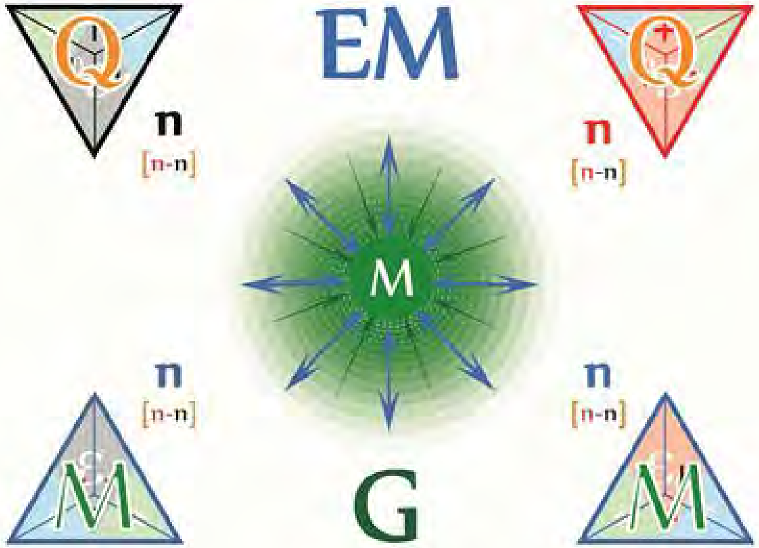
Q
n
[n-n]
EM
Q
n
[n-n]
M
n
[n-n]
M
G
n
[n-n]
M

charged Matter topologies create uni-directional convergent energy gradiants

All mass-Matter is velocity invariant and creates convergent G fields

Matterless EM mass-energies

Photons are often termed as 'mass-less' particles in physics which is a complete mis-nomer.
Leading to many erroneous supersitions on the true nature and physics of light and gravity

$$\frac{E}{c^2} = m$$

mass geometry

$$M = \frac{E}{c^4}$$

Matter topology

mass is a measure of planar Energy per second

mass-less infers no energy momenta

Photons are weightless radiant EM energies not massless particles

photons possess 2D planar radiant EM mass-energy momenta

A Particle has a charged Matter topology

Matter is a closed tetrahedral standing-wave energy possessing volume

photons & EM waves have no 3D Matter topology, they are Matter-less geometries

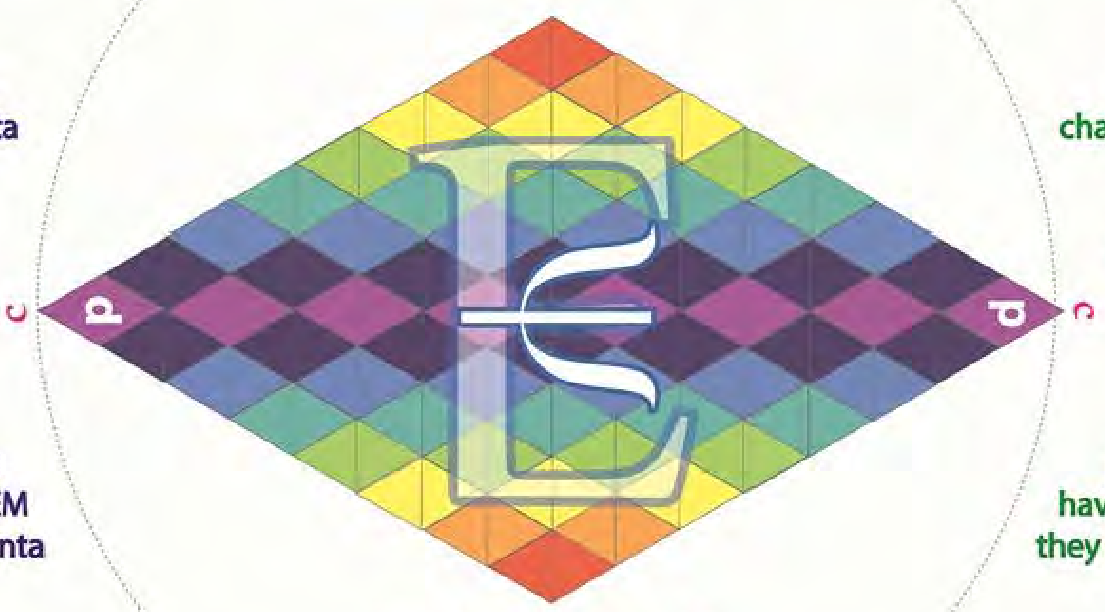

ENERGY

$$\frac{n\pi\left[\left[m\Omega v^2\right]\right]}{c^2}$$

Planck quanta mass velocity

second

mass

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

Matter

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

The definition and distinction between mass & Matter
is a major failing of modern physics, corrected for by Tetryonic theory

EM fields, vacuum energies and gravity

all result from radiant equilateral mass-energy momenta
and standing-wave Matter in radial space

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

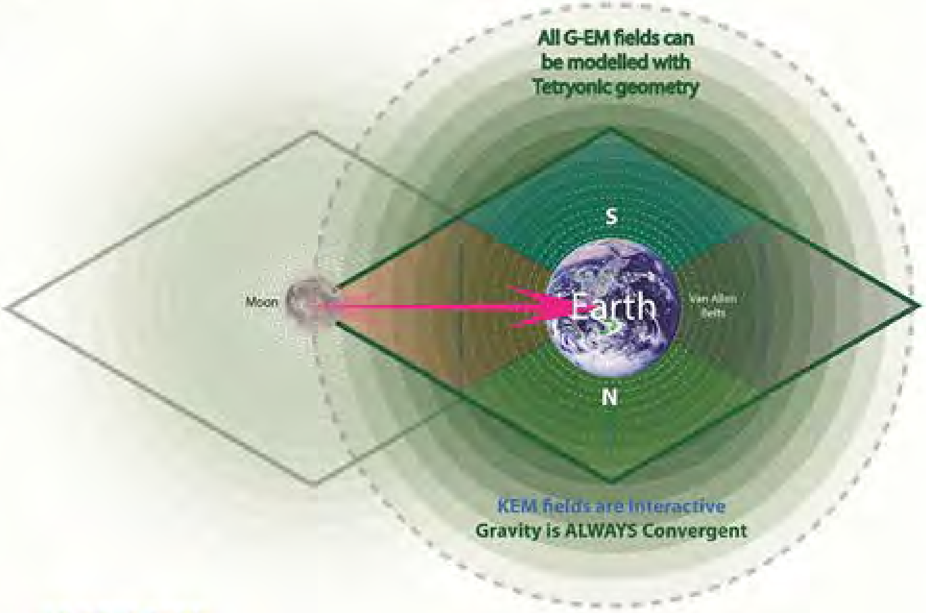
All G-EM fields can
be modelled with
Tetryonic geometry
S
Moon
Earth
Van Allen
Belts
N
KEM fields are Interactive
Gravity is ALWAYS Convergent

mass-energy momenta

EM

**Kinetic & EM fields create
bi-directional fields
of INTERACTION**

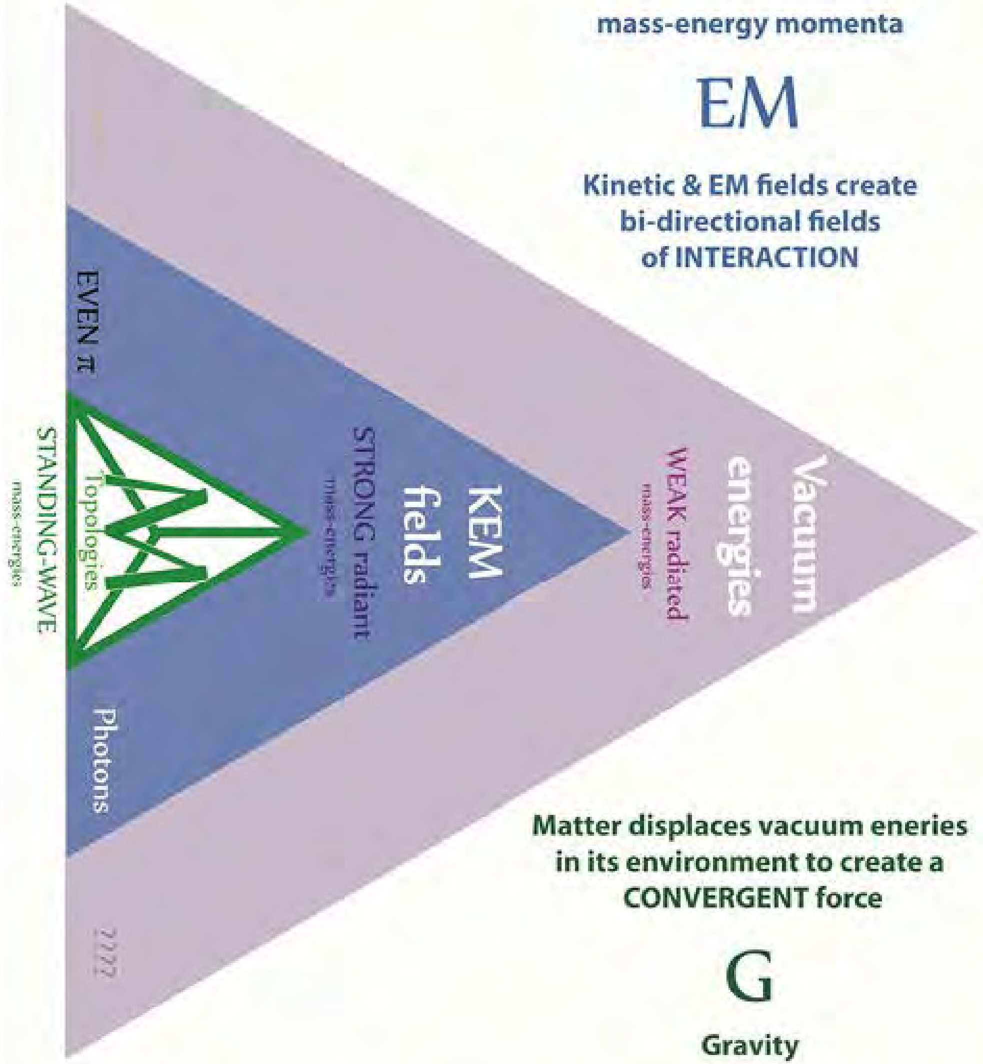
EVEN π
STANDING-WAVE
mass-energies
Topologies
STRONG radiant
mass-energies
KEM
fields
WEAK radiated
mass-energies
Vacuum
energies
Photons
????

**All GEM fields
have equilateral
quantum geometries**

**Matter displaces vacuum eneries
in its environment to create a
CONVERGENT force**

G

Gravity

*the proportional attraction of
material bodies towards each other
can be easily modeled with Tetryonic fields*

$$F = -G\,\frac{M_1\,m_2}{r^2}$$

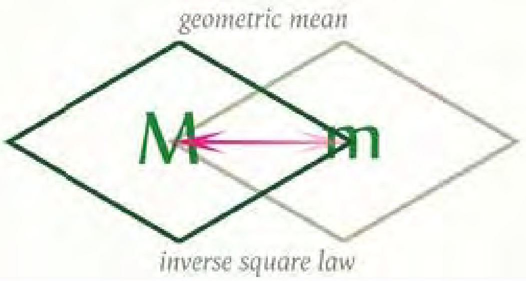
geometric mean
M
m
inverse square law

Copyright ABRAHAM [2008] - All rights reserved

Charges vs. Gravity

both can be modelled as interactive or attractive geometric forces

Charged particles are comprised of non-neutralised Planck quanta

Neutral Matter is comprised of neutralised Planck quanta

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

$$F = -G\,\frac{M_1 M_2}{r^2}$$

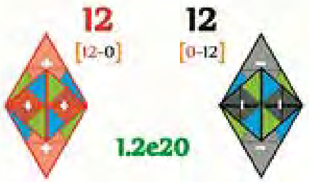

12
[12-0]
12
[0-12]
1.2e20

Charges produce bi-lateral interactive forces

mass

Matter

Nuclei

Elements

Gravity is a convergent attractive force

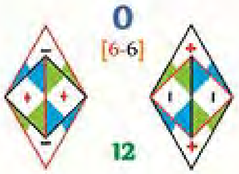

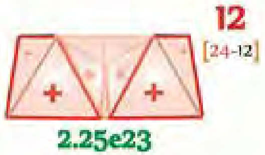

12
[24-12]
2.25e23

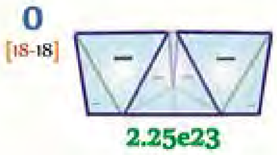

0
[18-18]
2.25e23

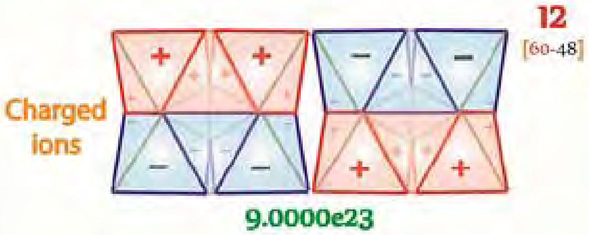

12
[60-48]
Charged ions
9.0000e23

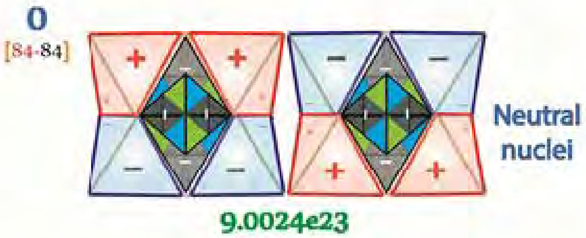

0
[84-84]
Neutral nuclei
9.0024e23

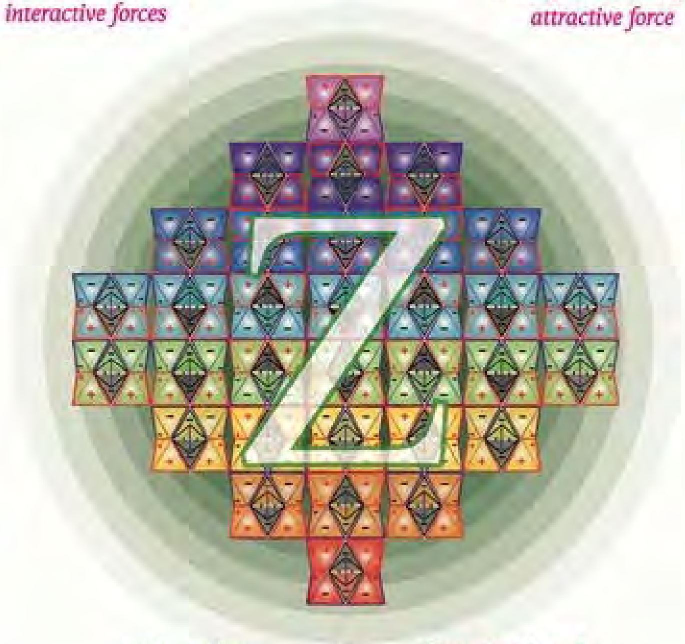

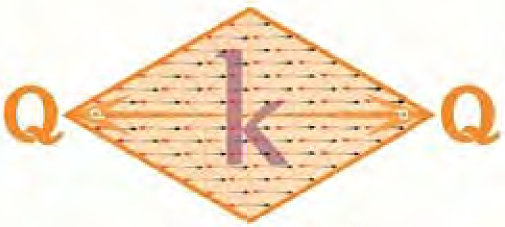

$$8.99 \times 10^{9}\ \frac{N\,m^2}{kg^2}$$

Opposites Attract
Similars Repel

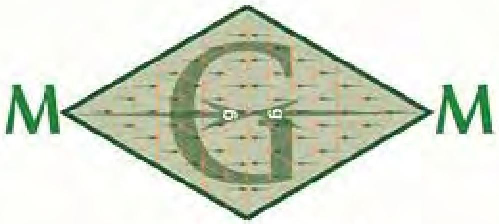

$$7.49 \times 10^{-11}\ \frac{N\,m^2}{KG^2}$$

Gravitational Matter
is always Attractive

Interactive mass-energies & **Gravitational mass-Matter**

both produce inverse squared forces

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

GravitoElectroMagnetism, (abbreviated GEM), refers to a set of formal analogies between Maxwell's field equations and an approximation, valid under certain conditions, of the Einstein field equations for General Relativity.

$$\nabla \cdot \mathbf{B} = 0$$

$$\mathbf{F} = q\left(\frac{-1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}\right) = q\mathbf{E}$$

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

$$k = 8.99 \times 10^{9}\,\frac{N\,m^2}{C^2}$$

nett Charge is a DIVERGENT field

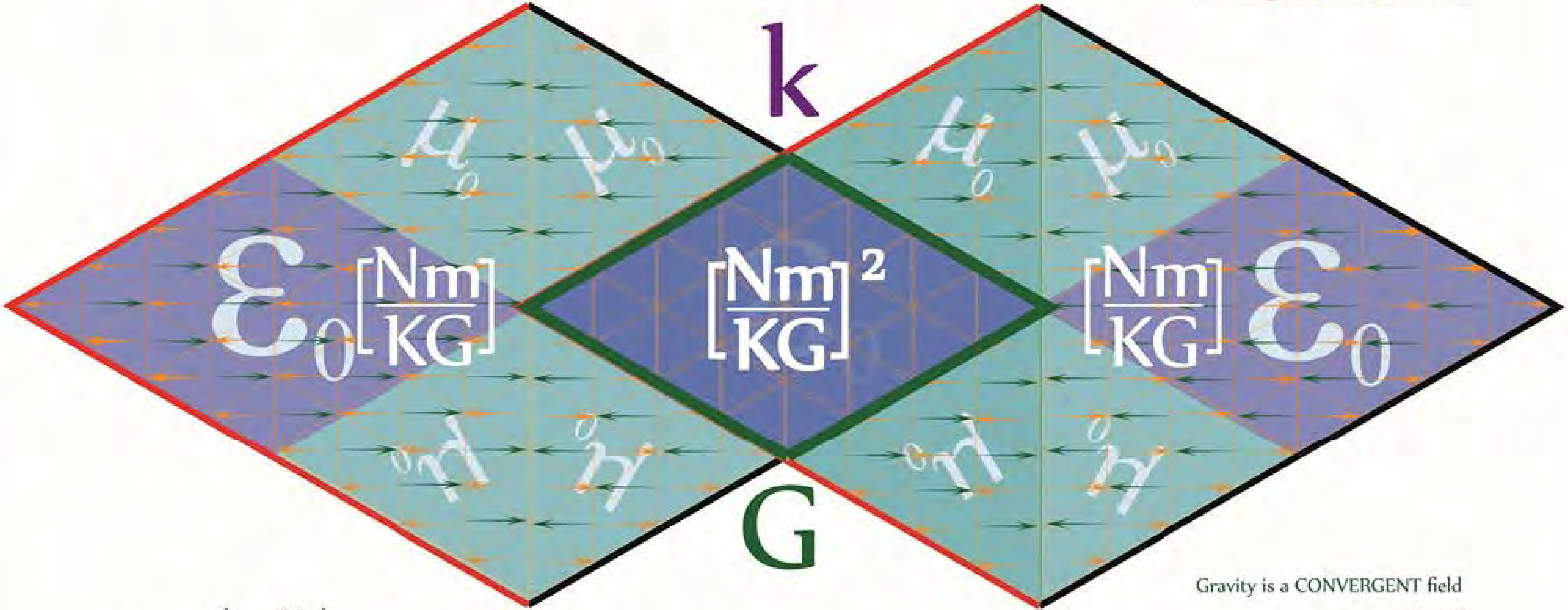
k
μ_0
$\varepsilon_0 \left[\frac{Nm}{KG}\right]$
$\left[\frac{Nm}{KG}\right]^2$
$\left[\frac{Nm}{KG}\right] \varepsilon_0$
G

Gravity is a CONVERGENT field

$$\mathbf{F} = m\left(-G\frac{M}{r^2}\hat{\mathbf{r}}\right) = mg.$$

$$F = -G\,\frac{M_1 M_2}{r^2}$$

$$G = 6.67 \times 10^{-11}\,\frac{N\,m^2}{kg^2}$$

$$\nabla \times \mathbf{E}_g = -\frac{\partial \mathbf{B}_g}{\partial t}$$

GEM fields are an accurate way of modelling Gravitational attraction between massive bodies using weak KEM field vectors

$$\nabla \cdot \mathbf{B}_g = 0$$

Copyright ABRAHAM [2008] - All rights reserved

Gravitational fields can be modelled as weak convergent EM fields

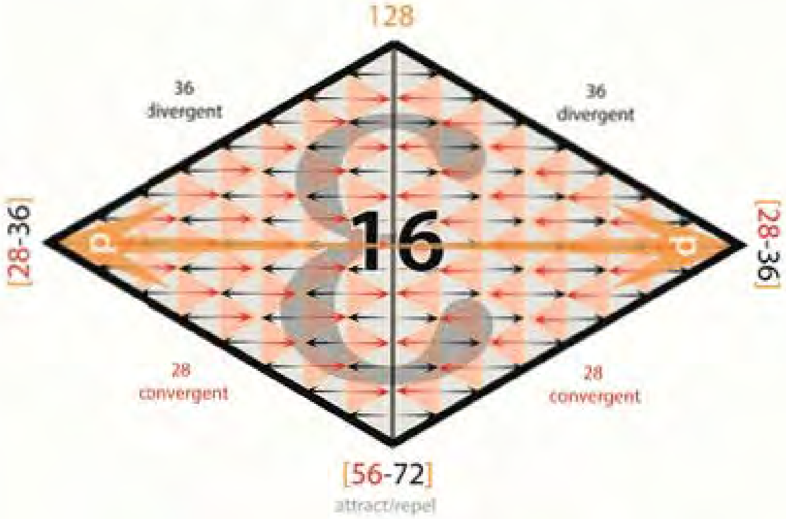
128
36
divergent
36
divergent
[28-36]
p
16
p
[28-36]
28
convergent
28
convergent
[56-72]
attract/repel

Electrostatic fields

Charges attract or repel dependent on polarity
Neutral Matter is largely unaffected (equally repel & attract)

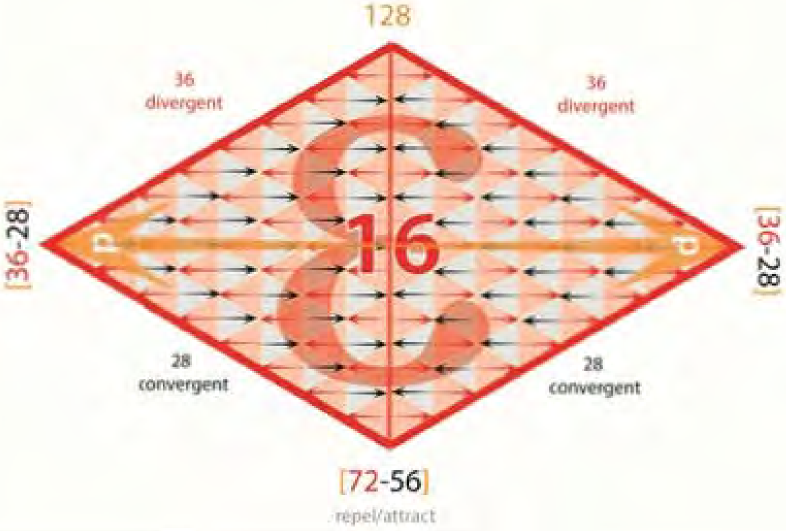
128
36
divergent
36
divergent
[36-28]
p
16
p
[36-28]
28
convergent
28
convergent
[72-56]
repel/attract

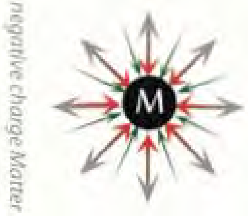
negative charge Matter
M

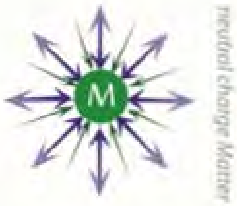
M
neutral charge Matter

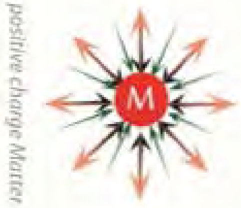
positive charge Matter
M

Charged Matter INTERACTS electrically

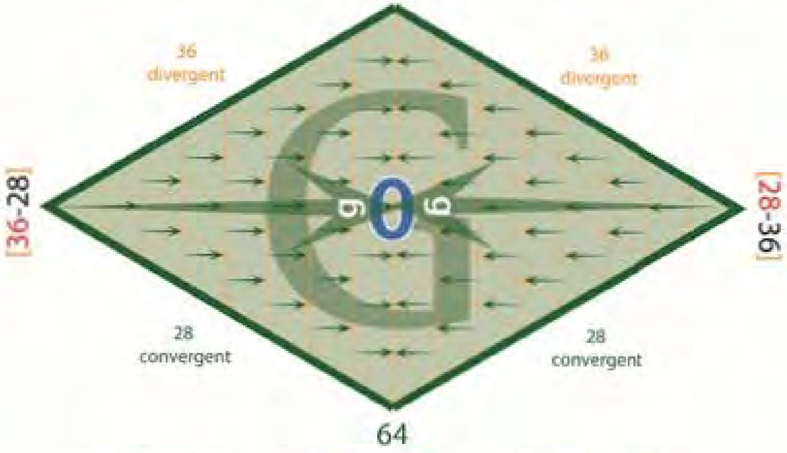
36
divergent
36
divergent
[36-28]
g
0
g
[28-36]
28
convergent
28
convergent
64

All Matter ATTRACTS gravitationally

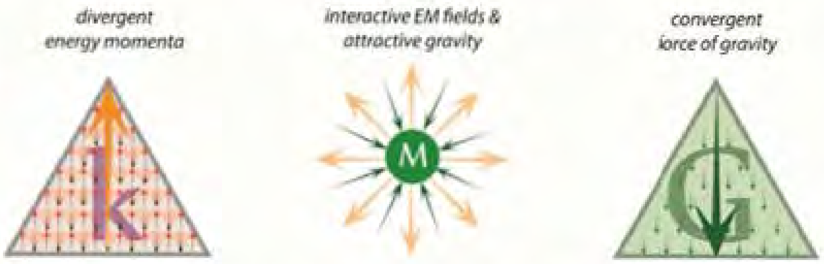
divergent
energy momenta
k
interactive EM fields &
attractive gravity
M
convergent
force of gravity
G

Nett KEM field interaction is divergent from source
Gravitational attraction is the convergent mean force between two sources

Copyright ABRAHAM [2008] - All rights reserved

Gravitational & Electric constants

are scaling constant relating Gravitational gradients to superpositioned EM fields

[1665]

Newton was the first to apply the Inverse Square law to Gravity

but was unaware of ElectroMagnetic forces

GEM

Gravito ElectroMagnetism

[1769]

Coulomb applied the Inverse Square Law to Charge Interactions

and obtained the value for the Electric Constant

Newton's convergent Gravity & Couolomb's Charge interaction
both follow inverse squared laws
(geometric means)

Attractive Gravity has the negative vector of nett Charge Interaction

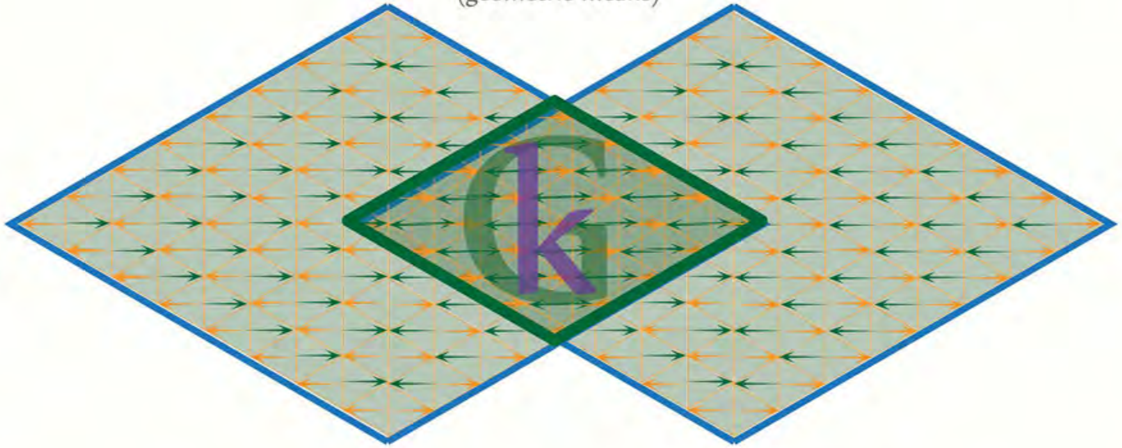

Interactive EM fields are 20 orders of magnitude stronger than Gravity

'Graviton'
Gravitational 'charge' carrier

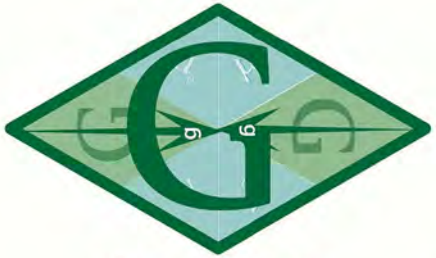

Convergent force vectors

superpositioned EM fields can be used to model the strength and direction of the weaker Graviational force vectors at any point in the fields as they are both inverse squared forces

Producing the ongoing speculation of the existence of a gravitational charge carrier

(an erroneous assumption that stems from using mathematics alone, without a solid understanding of geometric mass-energies & Matter topologies)

Photon
Electric charge carrier

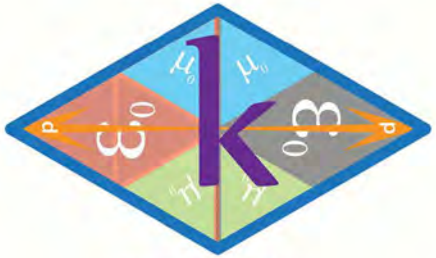

Divergent force vectors

Newton's attractive Gravity & Coulomb's Charge interaction
both follow inverse square laws

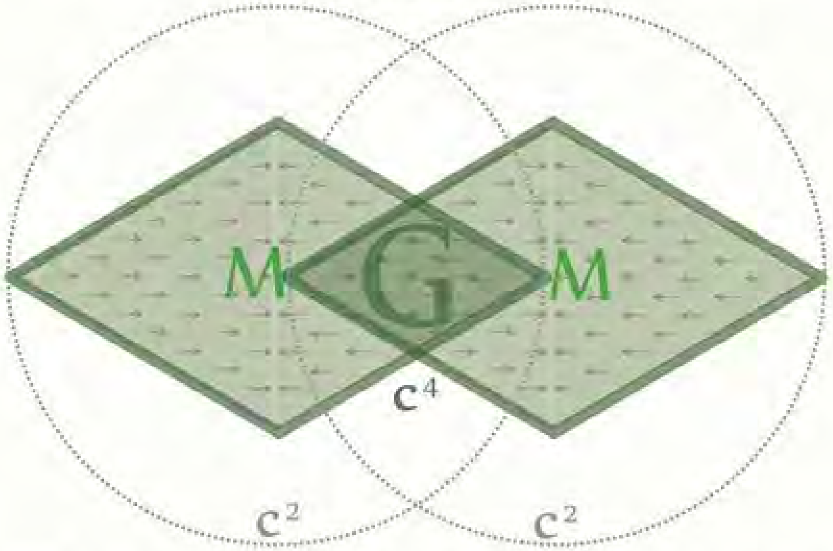

Instantaneous
action-at-a-distance

The Force of Gravity

Absolute
Space & Time
(x,y,z) (t)

Gravitational action
limited to c

mass-ENERGY-Matter

Curved relativistic
Spacetime
(x,y,z,t)

Modelling Gravity with EM field geometries

while superpositioned EM fields can be used to model the strength and vector direction
of weaker Gravitational fields at any point in time within the field geometry
they DO NOT reflect the actual physicality of the fields involved

$$R_{00} = 4\pi\rho$$

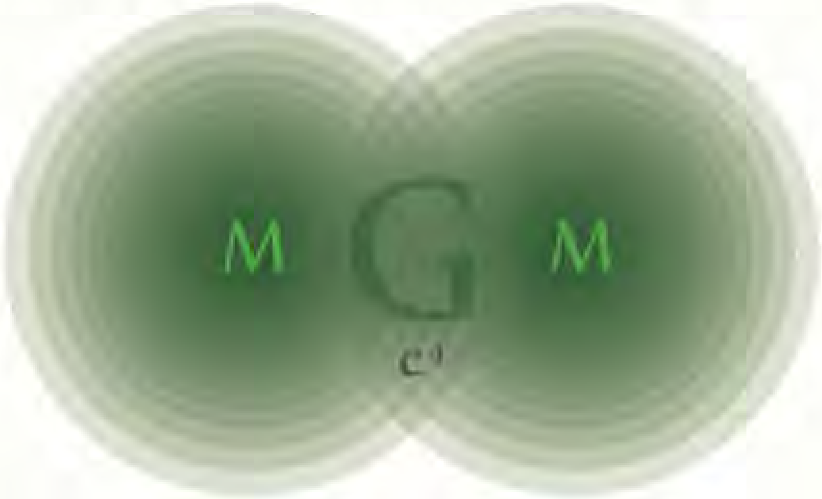

$$R_{\mu\nu} - \frac{1}{2}R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

*The speed of Gravity is not limited to c
as changes to the energy density-pressure
gradient are immediately translated through
the aether from one object to the other*

*Gravity is a the pressure gradient created
by the displacement of vacuum energies
though the presence of 3D mass-Matter
topologies in any region of space*

Fields of Interaction

All topological Matter in motion creates a interactive Kinetic EM field
in addition to its inherent convergent Gravitational field

Convergent Gravity fields produce a uni-directional attraction between two or more Matter topologies

Divergent EM fields produce a bi-directional interactive force between charged objects

G fields

E fields

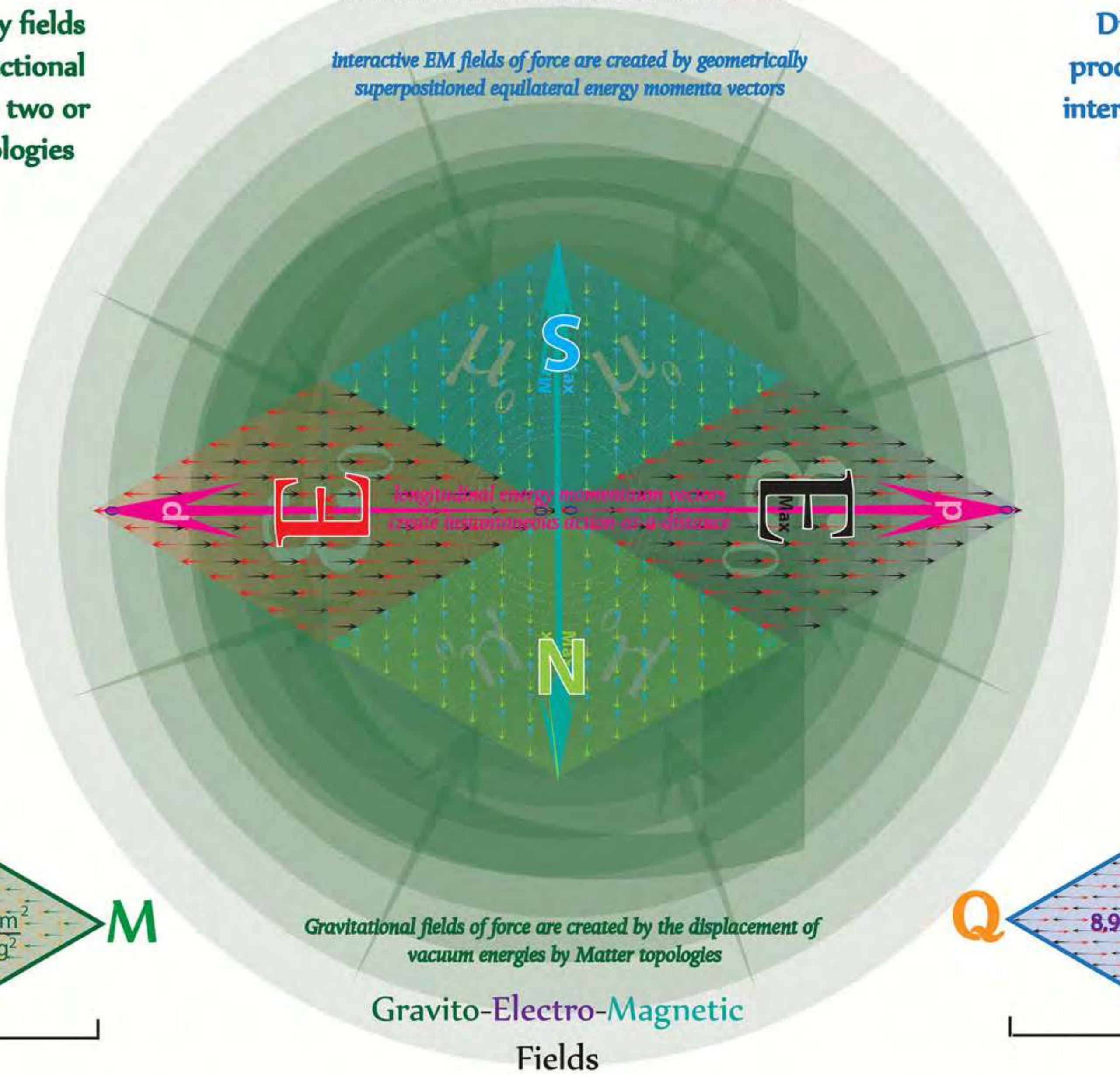
S
N
E
E
Max
p
p
μ0
ε0
longitudinal energy momentum vectors
create instantaneous action-at-a-distance

interactive EM fields of force are created by geometrically superpositioned equilateral energy momenta vectors

Gravitational fields of force are created by the displacement of vacuum energies by Matter topologies

Gravito-Electro-Magnetic
Fields

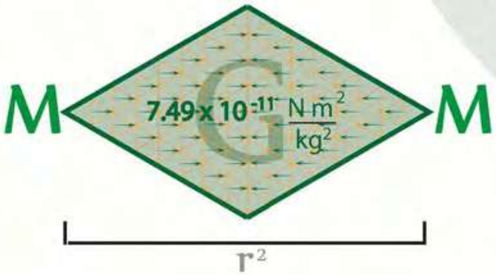
M
G
$7.49 \times 10^{-11} \frac{N\,m^2}{kg^2}$
M
r^2

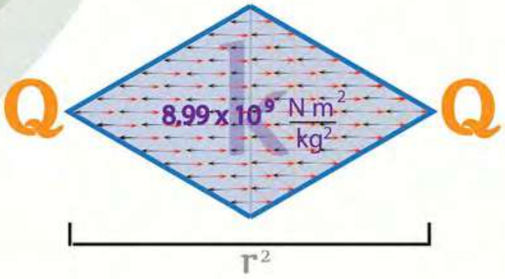
Q
k
$8{,}99 \times 10^{9} \frac{N\,m^2}{kg^2}$
Q
r^2

GEM gradients

The gravitational field of Material objects is the result of zero energy nullspace volumes encompassed by all Matter which in turn creates a displacement energy pressure differential gradient with the surrounding Vacuum Energies

Gravitational attraction can be modelled as a CONVERGENT EM FIELD created by the nullspace differential of Matter

All Matter in motion creates an eternally dynamic system of Gravitational fields and radiative ElectroMagnetic Energies

EM radiation is a bidirectional DIVERGENT-CONVERGENT FIELD resulting from quanta imbalances

radiative EM fields are created by all Matter as it seeks equilibrium with its environment, the impedance of space creates a propagation velocity equal to the speed of light

convergent Gravitational fields

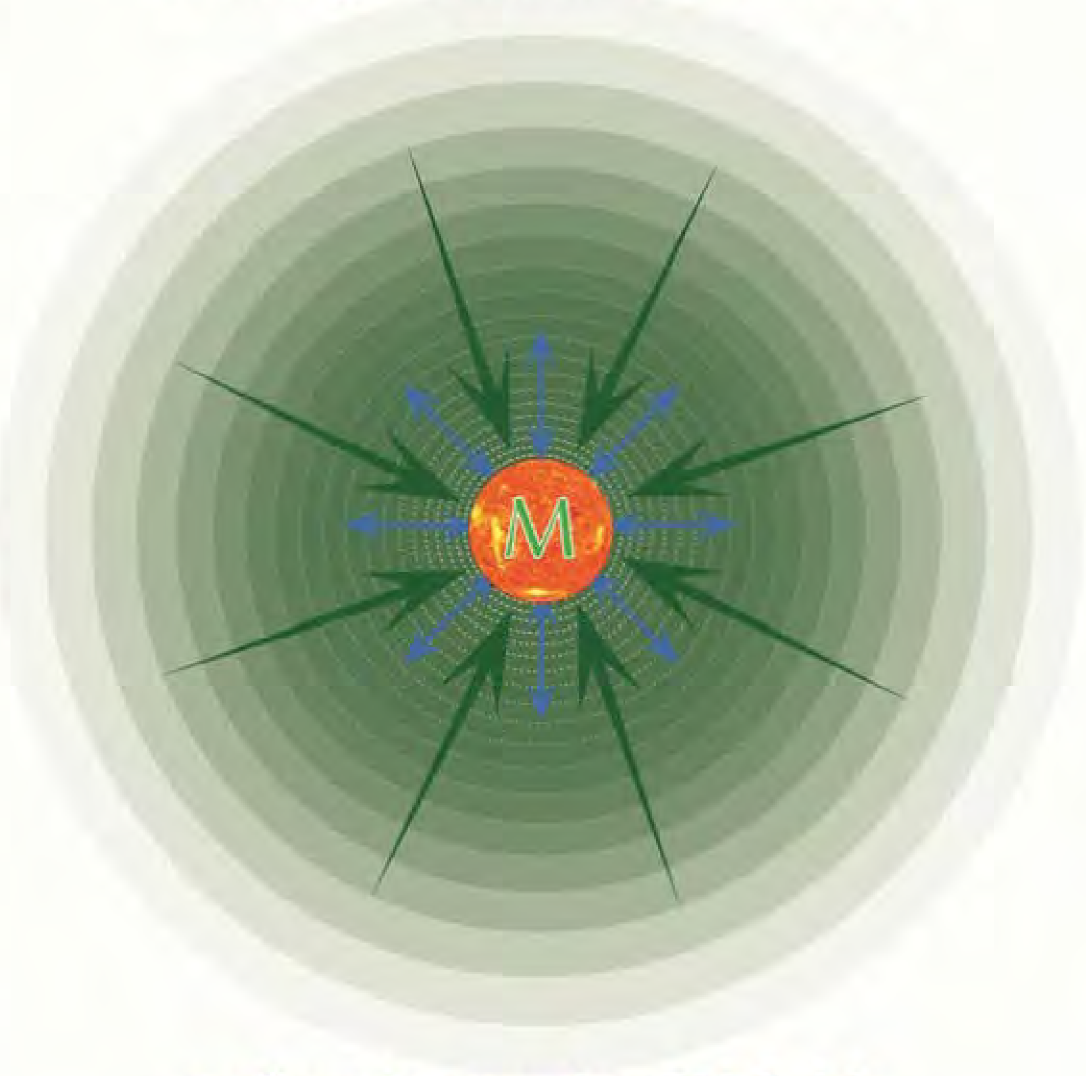

radiant interactive EM fields

Gravity and ElectroMagnetic fields

Newtonian Gravity

Gravity & Electro-Magnetic field strengths
all converge to zero at the centre of the sphere

General Relativity

All EM fields and
Matter are comprised
of charged EM quanta

The total relativisitc
energy of any object
is the sum of its mass-
Matter & Kinetic energies

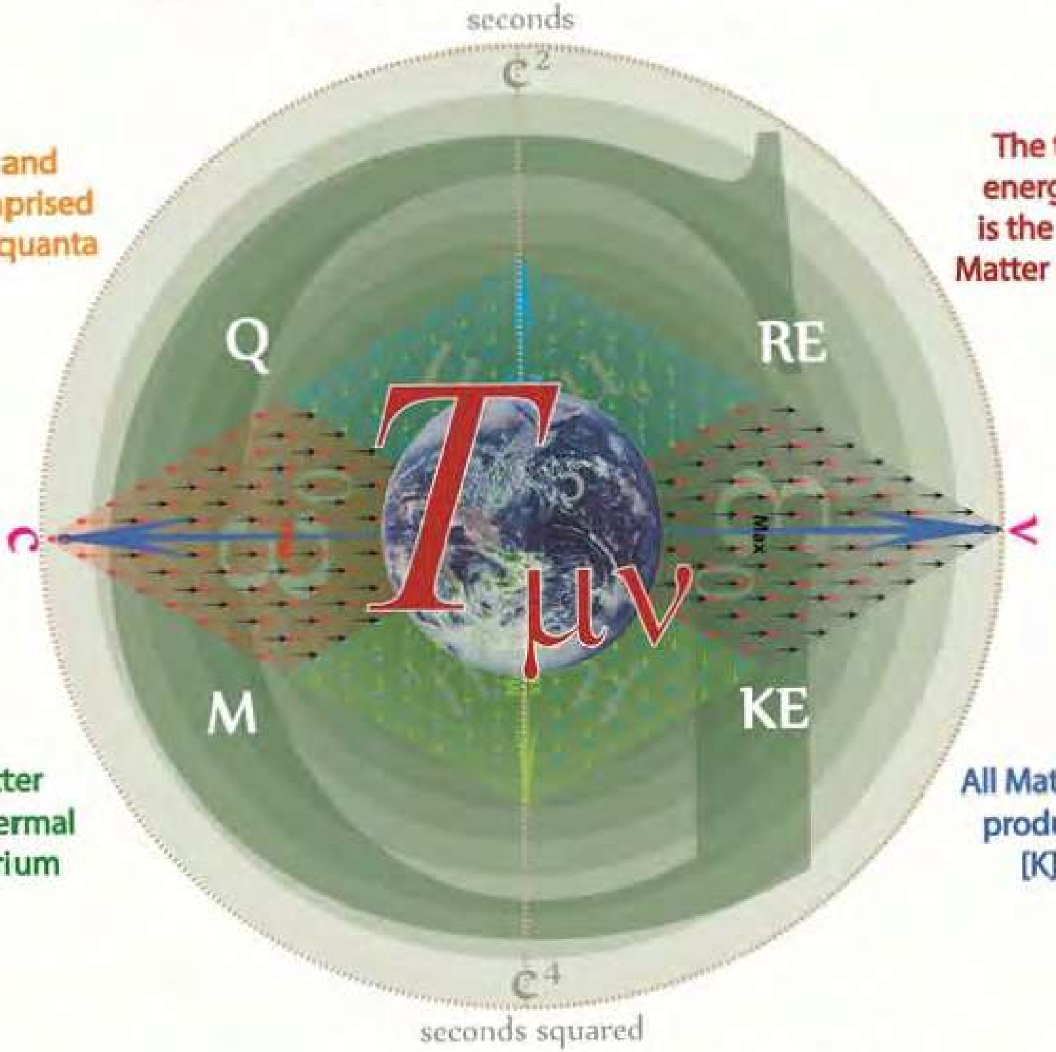

All Matter
seeks thermal
equilibrium

All Matter in motion
produces radiant
[K]EM fields

Charges are Interactive

Gravimetric fields can be modelled as weak EM fields
[space-time curvatures] created by Matter

Gravity is always attractive

Electric vs Gravitational Fields

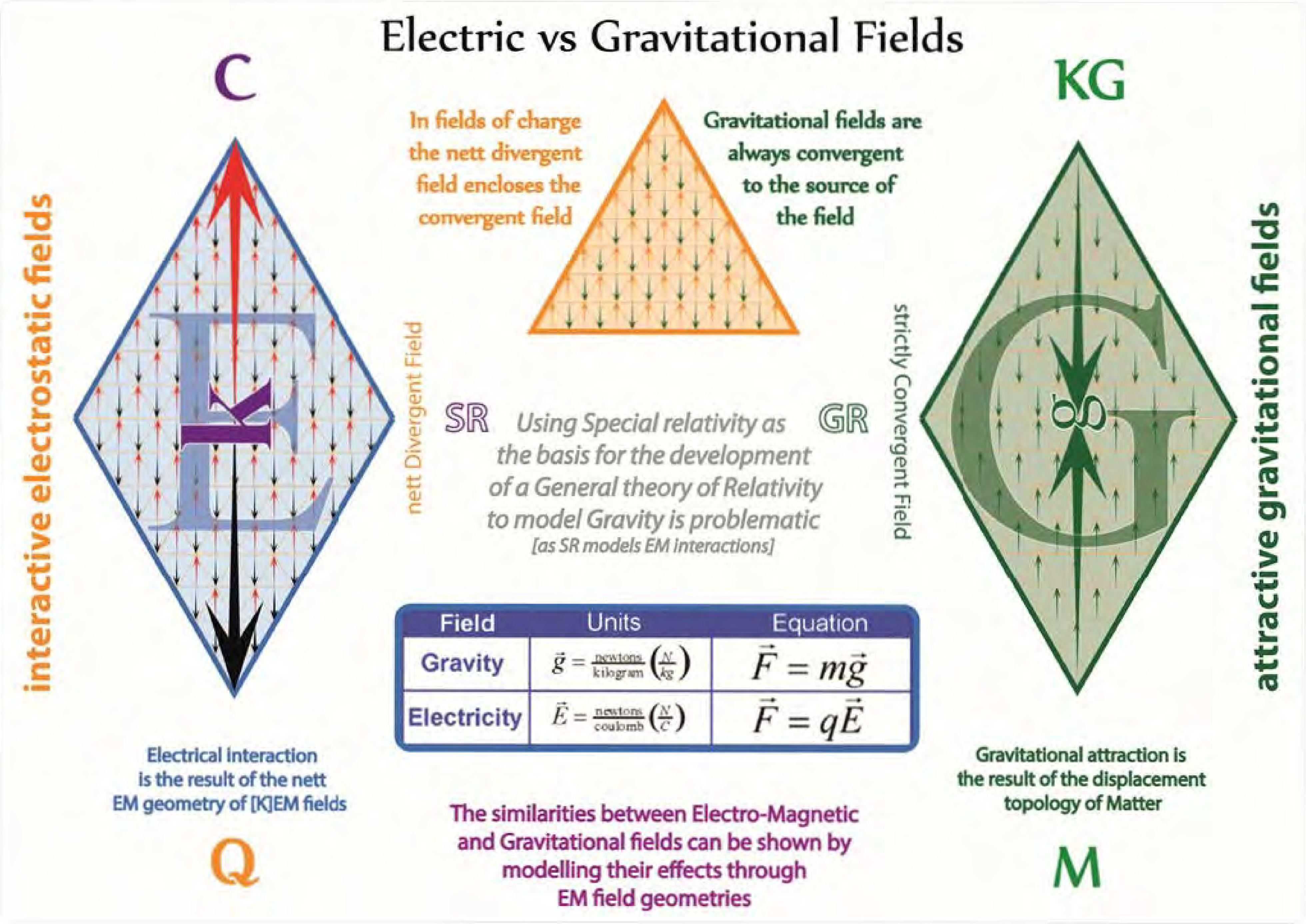
C
KG
interactive electrostatic fields
attractive gravitational fields
K
E
G
g
nett Divergent Field
strictly Convergent Field
SR
GR
Q
M

In fields of charge the nett divergent field encloses the convergent field

Gravitational fields are always convergent to the source of the field

Using Special relativity as the basis for the development of a General theory of Relativity to model Gravity is problematic

[as SR models EM interactions]

Field	Units	Equation
Gravity	$\vec{g} = \frac{\text{newtons}}{\text{kilogram}} \left(\frac{N}{kg}\right)$	$\vec{F} = m\vec{g}$
Electricity	$\vec{E} = \frac{\text{newtons}}{\text{coulomb}} \left(\frac{N}{c}\right)$	$\vec{F} = q\vec{E}$

Electrical interaction is the result of the nett EM geometry of [K]EM fields

Gravitational attraction is the result of the displacement topology of Matter

The similarities between Electro-Magnetic and Gravitational fields can be shown by modelling their effects through EM field geometries

Tetryonics 66.01 - Electric vs. Gravitational fields

Electric fields and Gravitation

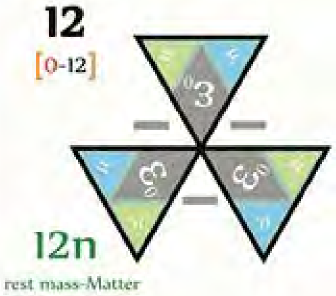

Electrostatic forces are much greater than Gravitational forces (by about 10^20 times).
The force of gravity is attractive for all Matter, whereas electrostatic forces are repulsive for like charged Matter,
There are no negative [repulsive] gravitational charges, while there are both positive and negative electric charges.

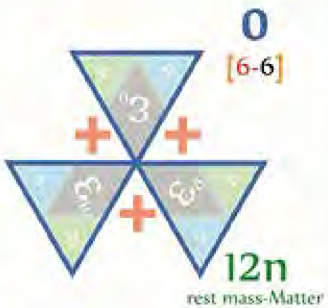

Electrostatic forces are interactive
while Gravitational forces are always attractive

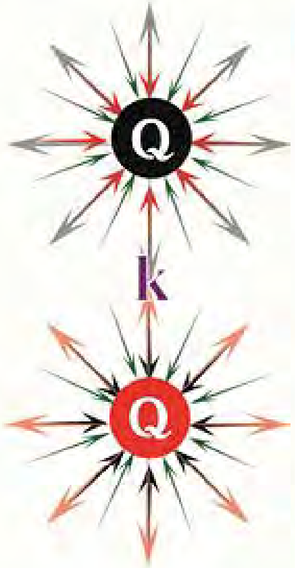

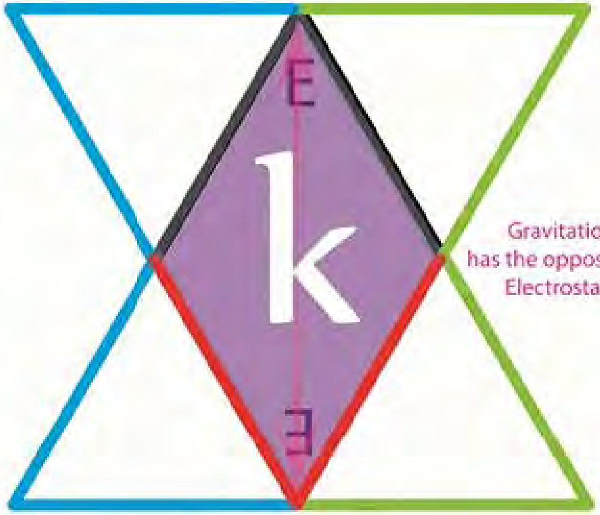

Gravitational attraction
has the opposite force vector of
Electrostatic interaction

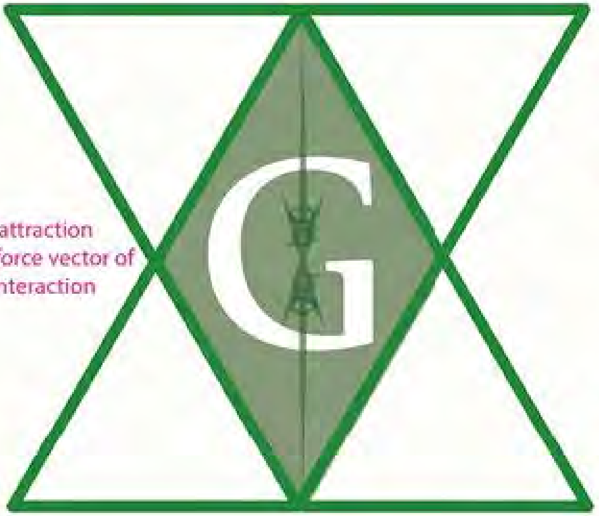

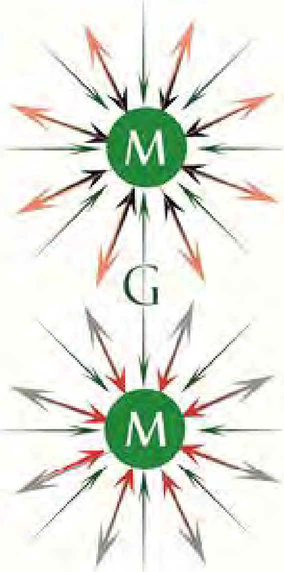

All ElectroMagnetic field quanta are subject to
Lorentz wavelength-frequency corrections due to velocity

Gravitational Matter topologies are NOT affected by
Lorentz corrections due to their velocity of motion

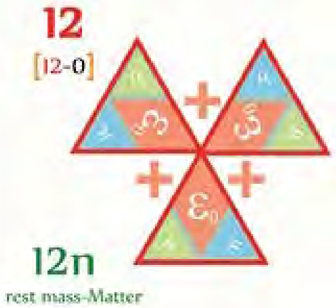

There are similarities between the Electric field and the Gravity field of Matter,
(so sometimes Matter is refered to as "gravitational charge")

Both act in a vacuum and are central and conservative.
Both obey an inverse-square law (both are inversely proportional to square of r).
Both field types propagate at the finite speed of c, the speed of light.
Electric charges and EM mass-energies are conservative

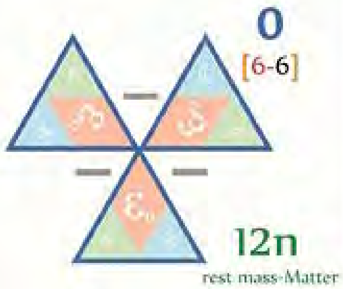

Charge vs Gravitation

[empirically determined value]

8.987551787e9 N

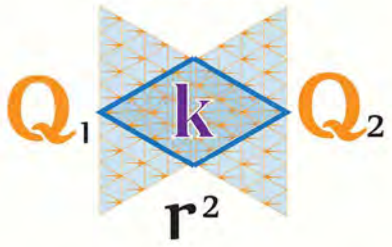

"The magnitude of the electrostatic force between two charges is directly proportional to the product of the magnitudes of each charge and inversely proportional to the square of the distance between the charges."

Charge interaction can be modelled as a DIVERGENT EM FIELD

We have two fundamental laws stating that the forces between two particles are directly proportional to the product of their charges/masses and inversely proportional to the square of the distance between them

Gravitational attraction can be modelled as a CONVERGENT EM FIELD

"The magnitude of the gravitational force between two masses is directly proportional to the product of the magnitudes of each mass and inversely proportional to the square of the distance between the charges."

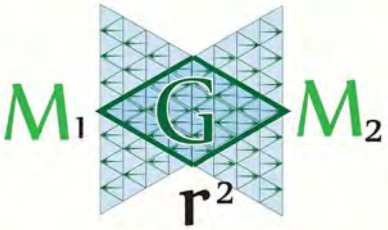

6.67428e-11 N

[empirically determined value]

Coulombic interaction is proportional to Charge (energy geometry imbalances)

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

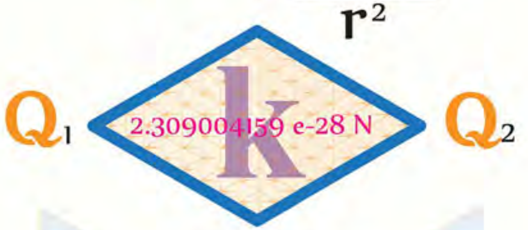

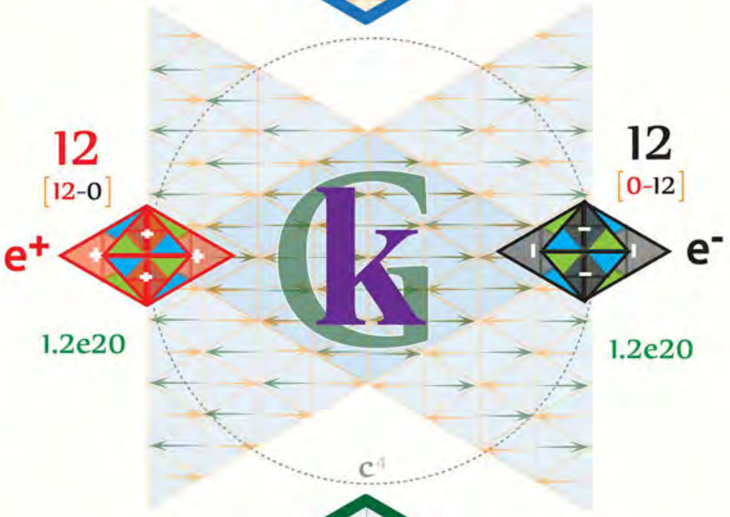

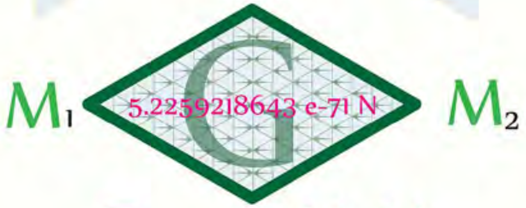

$$F = -G\,\frac{M_1 M_2}{r^2}$$

Gravitational attraction is proportional to Matter (total rest EM mass-energy content)

*It can now be clearly demonstrated that **Newtonian Gravity** can be modelled as a weak convergent Electro-Magnetic Force*

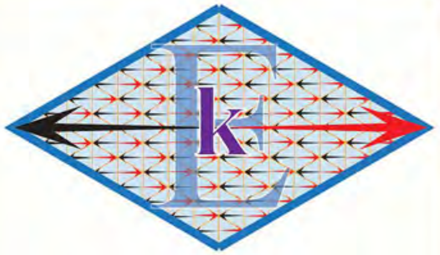

$$\frac{k}{G} \quad \frac{8.9875 \text{ e9 N}}{6.67384 \text{ e-11 N}}$$

$$= 1.346 \text{ e20}$$

The field strength differential is reflective of the Compton frequency of electron [1.20 e20 quanta]

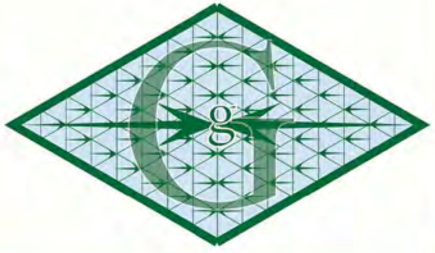

Einstein's general relativity offers a different model of Gravitational attraction however

Gravito-Electro-Magnetic interaction

$$F \propto \frac{Q_1 Q_2}{r^2}$$

Charge interactions can be modelled as a superpositional
DIVERGENT FORCE

The ElectroMagnetic [Coulombic] field strength is proportional to an object's relativistic charge

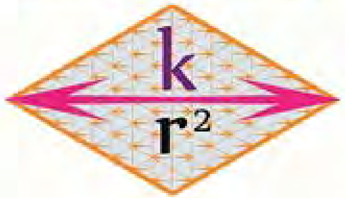

$$F = k\frac{Q_1 Q_2}{r^2}$$

equilateral mass-energy momenta quanta create the interactive fields of Force between material bodies of Matter

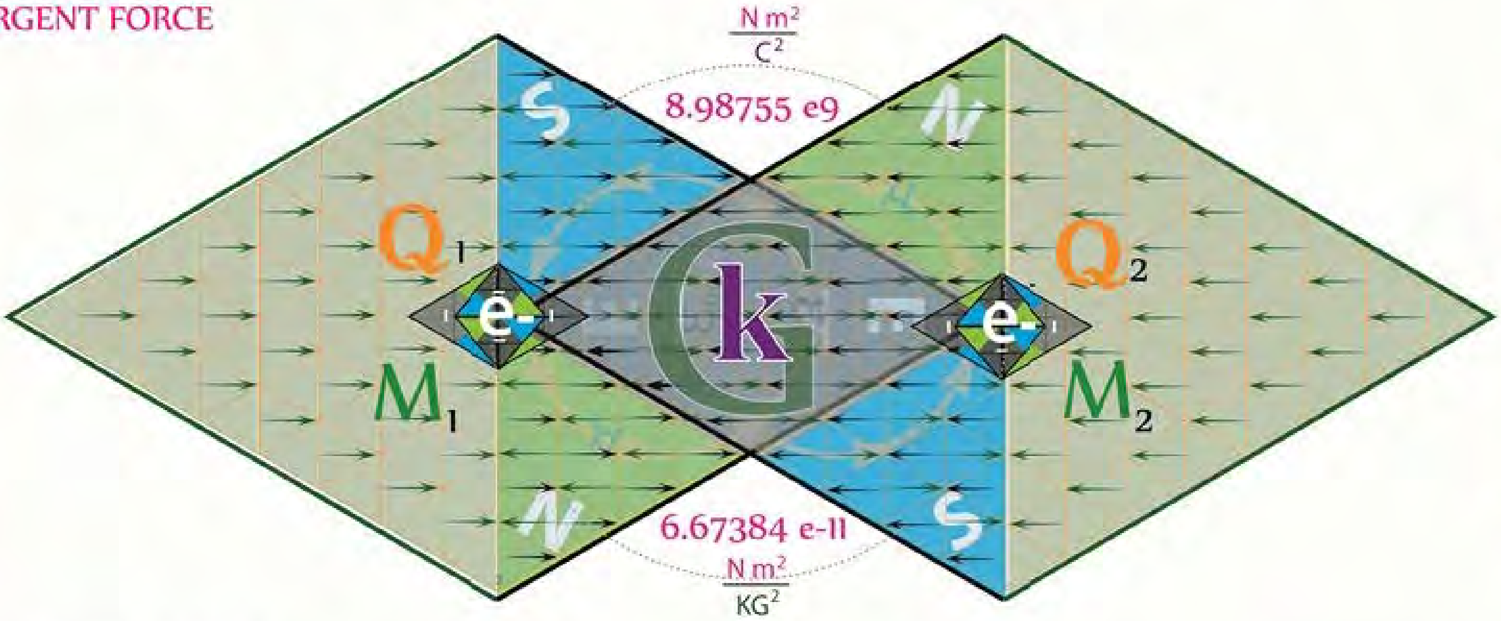

The attractive strength of Gravity is 20 orders of magnitude weaker than the strength of Electrostatic interaction

Gravitational attraction can be modelled as a superpositional
CONVERGENT FORCE

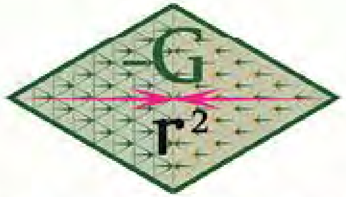

$$F = -G\frac{M_1 m_2}{r^2}$$

$$F \propto \frac{M_1 m_2}{r^2}$$

Gravitational field strenth is proportional to the rest Matter topology (and can be modelled with weak coulombic fields)

Copyright ABRAHAM [2008] - All rights reserved

Charged Matter Gravitational interactions

**Electrostatic force fields dominate Gravitational fields
and can be modelled via their superpositioned
energy momenta force fields**

*Following the field symmetries
provided by Tetryonic geometry
we can now model all fields*

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

2.30900415 e-28 Newtons

*Electromagnetic interactions
as well as the convergent
Gravitational atraction*

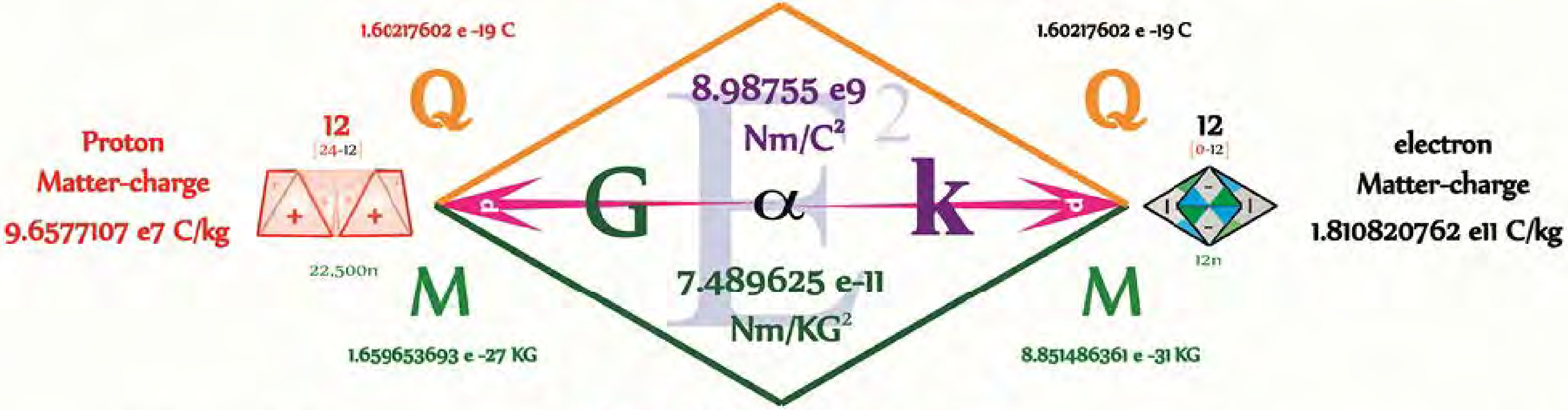
1.60217602 e -19 C
1.60217602 e -19 C
Q
Q
Proton
Matter-charge
9.6577107 e7 C/kg
12
[24-12]
+
+
22,500n
8.98755 e9
Nm/C²
E²
G
α
k
p
p
12
[0-12]
12n
electron
Matter-charge
1.810820762 e11 C/kg
M
M
7.489625 e-11
Nm/KG²
1.659653693 e -27 KG
8.851486361 e -31 KG

1.10025594 e-67 Newtons

*Reflecting all of the
geometric force interactions in terms of
EM mass-ENERGY & Matter*

$$F = -G\,\frac{M_1 m_2}{r^2}$$

*Allowing an accurate value
of the graviationl constant
to be derived*

**Gravity [like electrostatic] fields
can be modelled as superpositioned
Energy momenta force fields**

Copyright ABRAHAM [2008] - All rights reserved

The Gravitational Constant

The gravitational constant is a difficult physical constant to measure to high accuracy,
in fact it was not measured until over seventy years after Newton's death by
Henry Cavendish with his torsion beam experiment, performed in 1798

G

The gravitational constant is denoted by letter G,
it is an empirical physical constant involved in the calculation
of the gravitational force between two material bodies.

It usually appears in Sir Isaac Newton's law of universal gravity,
and in Albert Einstein's theory of general relativity.

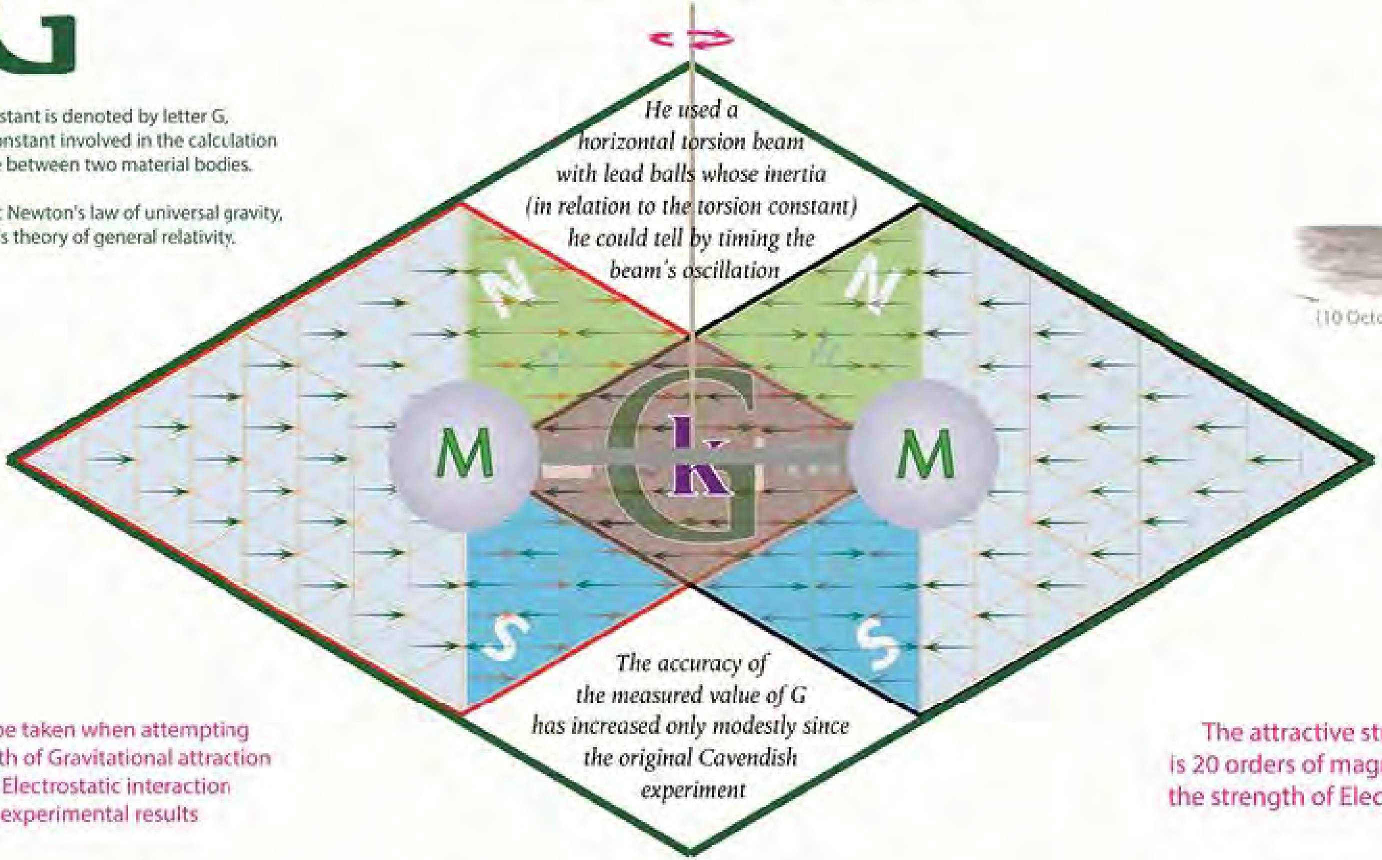

He used a horizontal torsion beam with lead balls whose inertia (in relation to the torsion constant) he could tell by timing the beam's oscillation

The accuracy of the measured value of G has increased only modestly since the original Cavendish experiment

Henry Cavendish

(10 October 1731 – 24 February 1810)

Extreme care must be taken when attempting
to measure the strength of Gravitational attraction
to ensure that NO Electrostatic interaction
influences the experimental results

The attractive strength of Gravity
is 20 orders of magnitude weaker than
the strength of Electrostatic interaction

8.98755 e9 N

$$F = k \frac{Q_1 Q_2}{r^2}$$

It is also known as the universal gravitational constant,
Newton's constant, and colloquially as Big G.

g

It should not be confused with "little g" (g),
which is the local gravitational field
(equivalent to the free-fall acceleration),
especially that at the Earth's surface.

G

6.67384 e-11 N

$$F = -G \frac{M_1 m_2}{r^2}$$

Gravity has no established relation to other fundamental forces, so it does not appear possible to calculate it indirectly from other constants that can be measured more accurately, as is done in some other areas of physics

Copyright ABRAHAM [2008] - All rights reserved

Fundamental field strengths

are the result of equilateral energy momenta geometries

magnetic dipole interaction
ε0
W
03
Weak

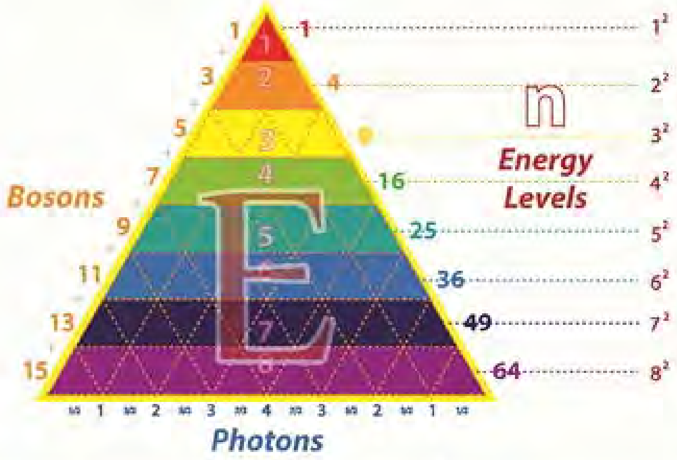
Bosons
1
3
5
7
9
11
13
15
E
n
Energy Levels
1
4
16
25
36
49
64
1²
2²
3²
4²
5²
6²
7²
8²
Photons

Strong
S
parallel charged fascia interaction

Force	Approximate Relative Strength	Range
Strong Nuclear	10^{38}	10^{-15}
ElectroMagnetic	10^{36}	Infinite
Weak	10^{2}	10^{-18}
Gravitational	1	infinite

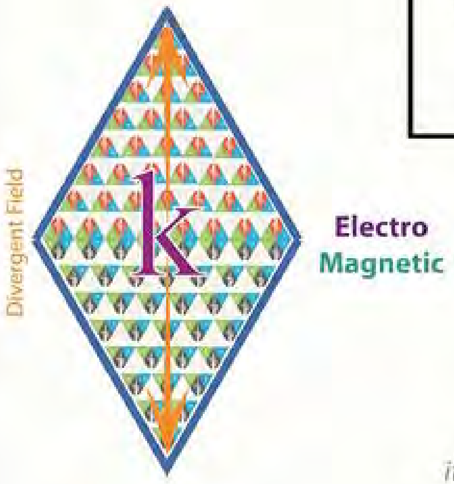
Divergent Field
k
Electro Magnetic

All the field strengths of the fundamental interactions are a direct result of the interactions of their Tetryonic mass-energy goemetries and Tetryonic Matter topologies

m **E** **M**

geometries density topologies

(whilst Gravity can be modelled as a weak convergent Coulombic attraction its actual physics are the result of vacuum energy displacement by Matter topologies)

Gravity
G
Convergent Field

The Strength of the Gravitational Constant

Coulomb's Constant

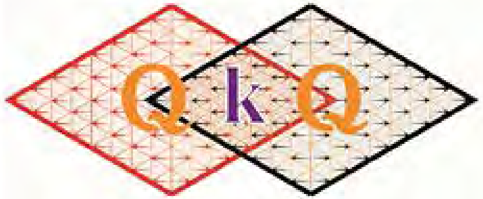

[24-12]

Proton nett charge

1.60217646 e -19 C

equal
(but opposite)
nett Charges

[0-12]

Electron nett charge

1.60217646 e -19 C

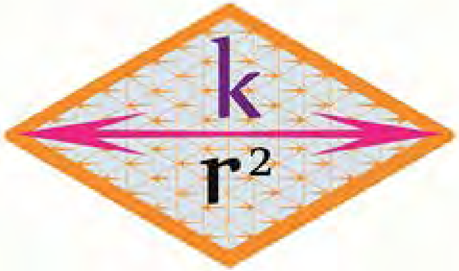

8.987551787e9 N

CHARGED mass-energy
geometries
interact

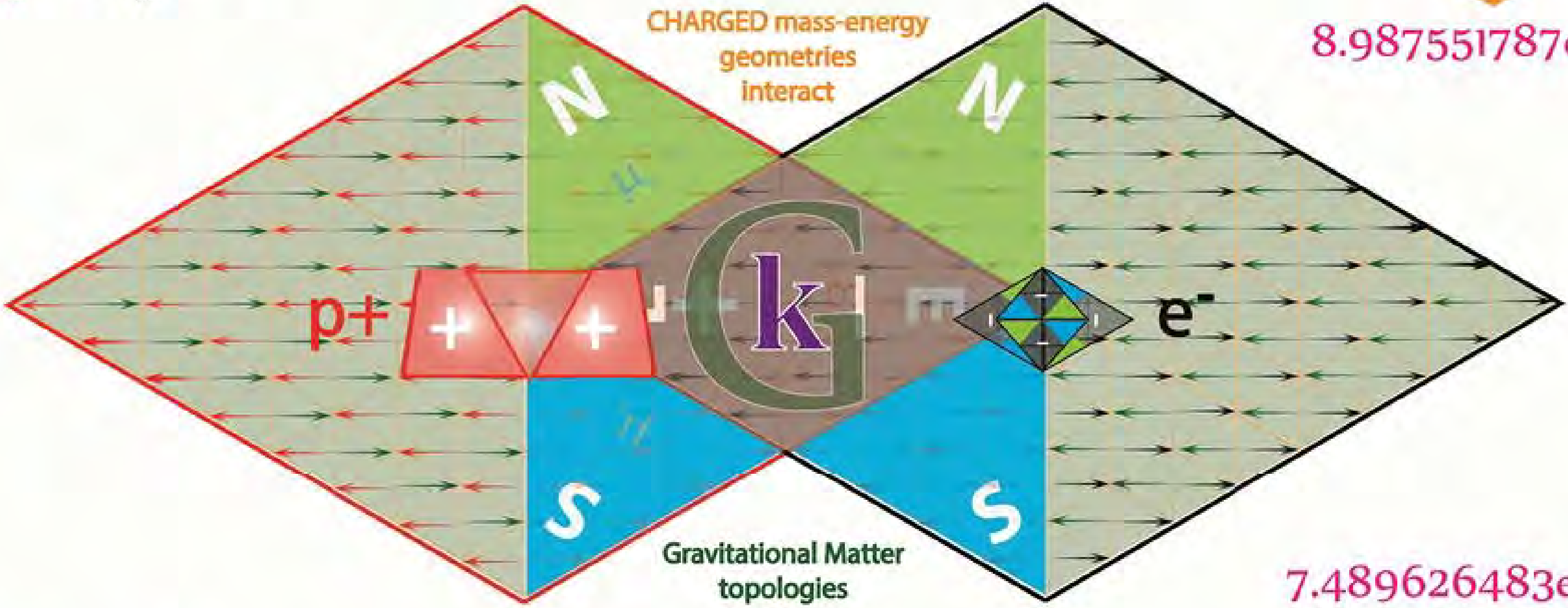

Gravitational Matter
topologies
attract

7.489626483e-11 N

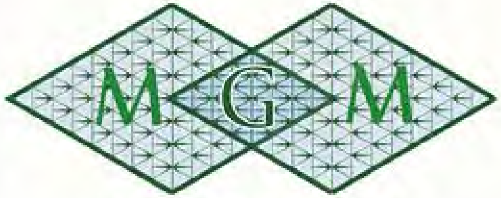

Newton's Constant

1.659653693 e -27 KG

Proton rest mass-Matter

[1.5e23-7.5e22]

1875

8.851486361 e -31 KG

electron rest mass-Matter

[1.2e20]

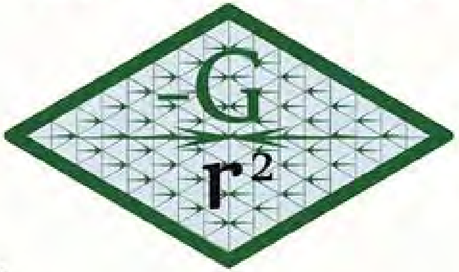

Although it provides a coherent and accurate model of all the motive forces & interactions present the modelling of Gravity with EM fields dces not account for the source, strength & unilateral direction of Gravitational fields

Tetryonics 66.08 - Strength of the Gravitational Constant

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic Matter interaction and Gravitation

Electrostatic convergence can serve as a model for Gravitational attraction

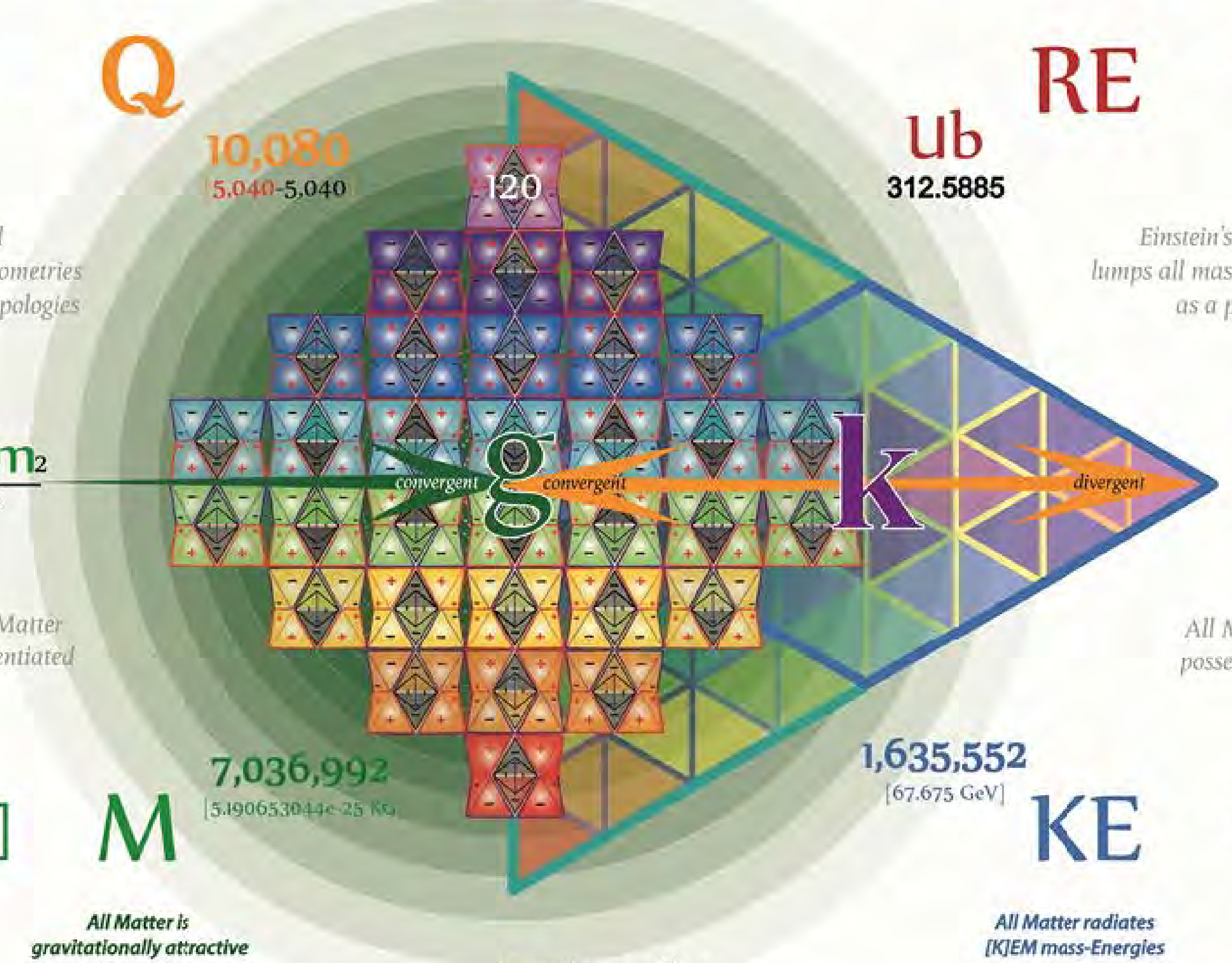

All Matter is comprised of EM mass-Energies

Q

10,080

5,040-5,040

Charged mass-energy geometries form Matter topologies

$$F = -G\,\frac{M_1 m_2}{r^2}$$

EM mass & Matter must be differentiated

$$\overset{\text{Matter}}{\frac{4n\pi}{c^4}}\left[\left[\overset{\text{Planck quanta}}{\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}}\right]\right]$$

M

7,036,992

[5.190653044e-25 Kg

All Matter is gravitationally attractive

RE

All gravitational systems are comprised of Matter & KEM fields

ub

312.5885

Einstein's stress-energy tensor lumps all mass-Energy-matter together as a pressure gradient

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

All Matter in motion posseses [K]EM fields

1,635,552

[67.675 GeV]

KE

$$\overset{\text{EM waves}}{\frac{\text{EVEN}\,\pi}{c^2}}\left[\left[\overset{\text{Planck quanta}}{\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}}\right]\right]$$

All Matter radiates [K]EM mass-Energies

but in reality

Gravitational attraction results from the displacement of vacuum energies by Matter topologies

[The effects of stronger [K]EM fields can be excluded from the long distance attractive [gravitational] motion of material bodies to each other]

Modelling the force of Gravity

It is in many ways very similar to Coulomb's law of Interaction
between charged particles developed by Coulomb
(after Newton's Gravitation)

Gravity can be modelled as weak convergent EM field

For many people, it is easiest to picture gravity
as a kind of special invisible force,
emanating from ALL the Matter in the universe.
Newton's law of Gravity leads one to sees
it this way and it is probably the view
most of us were given when we learned
about gravity in elementary school.

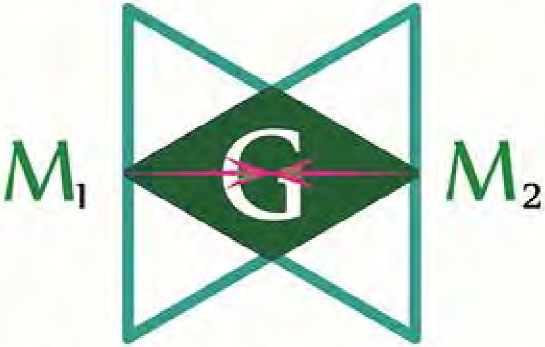

However, there is another way
to model Gravity's effects.
It stems from Albert Einstein's
General theory of Relativity,
a theory that describes gravitation
as being the result of the curvature of
spacetime as a result of energy gradiant

Gravity can also be modelled as convergent pressure gradients

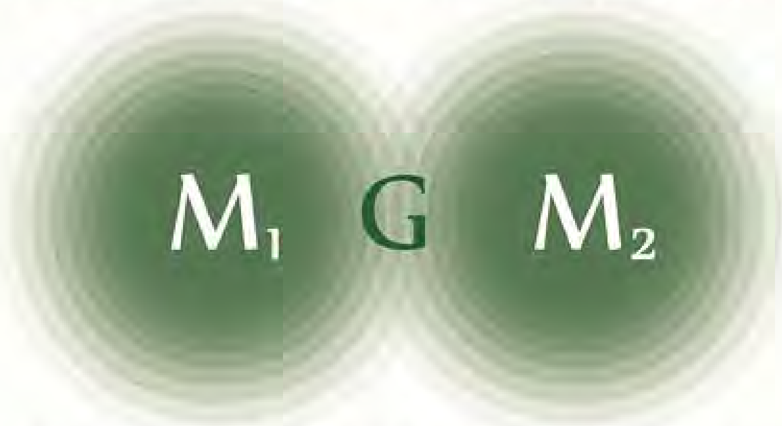

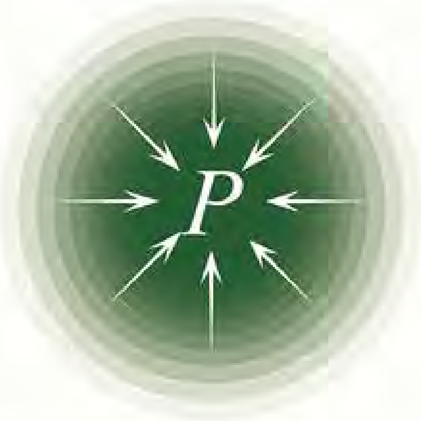

To understand what gravity really is you must have a clear understanding of the distinctions between radiant EM mass and standing-wave Matter and the various quantum interactions they produce.

Tetryonics provides a clear geometric definition and distinction between 2D radiant mass-energy geometries and 3D Matter topologies along with a new unified model of the mechanics of Universal Gravitation at all scales

radiant mass-energy geometries

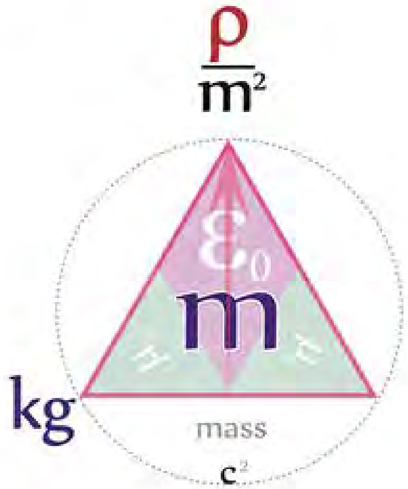

'Action-at-a-Distance'

mass-energies Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

[K]EM mass-Energies with QAM create interactive divergent [and convergent] charge geometries

Newtonian Gravity

Toplogical mass is Matter

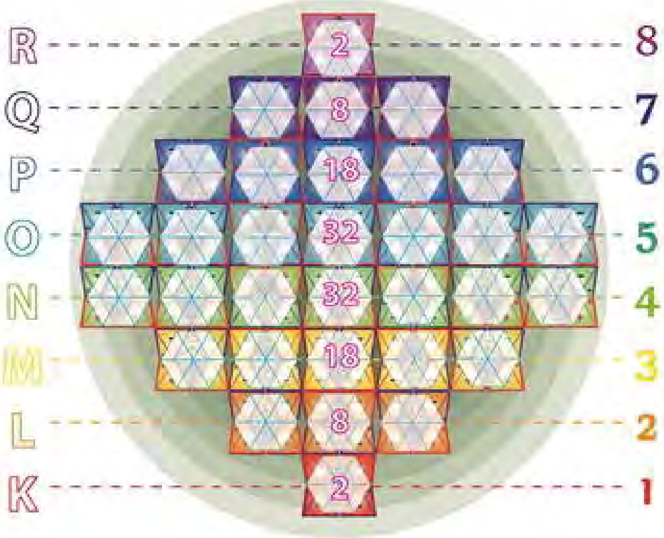

All Matter contains nullspaces and produces motional KEM fields

standing-wave Matter toplogies

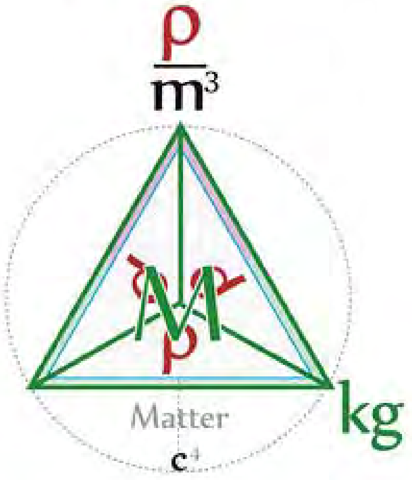

'Folded Space-Time'

Matter Planck quanta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Tetryonic Matter creates low pressure nullspace resulting in convergent Energy pressure gradients

General Relativity

Copyright ABRAHAM [2008] - All rights reserved

Instantaneous action-at-a-distance

Tetryonic QED theory shows that where the linear energy momentum of longitudinal waves align and act in a co-linear direction they can facilitate near instantanous action-at-a-distance

$$\sum \mathbf{F} = 0 \Rightarrow \frac{d\mathbf{v}}{dt} = 0$$

Free particles move with constant vector-velocity
(that is, with zero acceleration, unless acted upon by an outside force).

The Earth orbits the SUN where it is now about a common centre of gravity

'Action-at-a-Distance' and 'Instantaneous' forces were introduced into Physics within Newton's notion of absolute space and time

If the SUN were to vanish the Earth would cease traveling in its current elipical orbit immediately

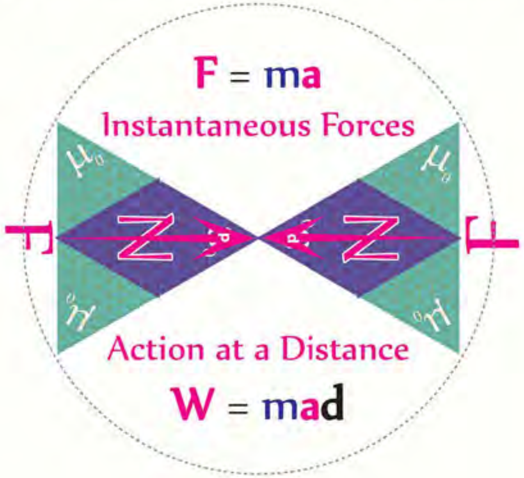

F = ma

Instantaneous Forces

Action at a Distance

W = mad

$$\sum \mathbf{F} = \frac{d\mathbf{p}}{dt} = m\frac{d\mathbf{v}}{dt} + \mathbf{v}\frac{dm}{dt}$$

The vector-force on a particle equals the product of its mass by its vector acceleration

$$\mathbf{F} = \mathbf{ma} = \frac{\Delta \mathbf{p}}{\Delta t} = \Delta\left[\frac{\mathbf{mv}}{t}\right]$$

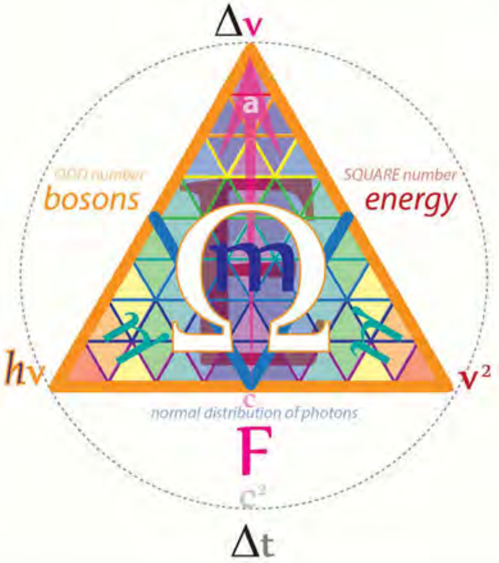

The speed of gravity is limited to c

Einstein's GR theory replaces Newton's force of gravity and absolute space & time with a relativistic spacetime gradient curved by mass-energies

If the SUN were to vanish the Earth would continue traveling in its current orbit for 8⅓ minutes

The forces of action and reaction are equal and opposite;
(if a particle A exerts a force f on a particle B, then B exerts a force–f on A)

$$\sum \mathbf{F}_{a,b} = -\sum \mathbf{F}_{b,a}$$

Copyright ABRAHAM [2008] - All rights reserved

Newton's Gravitational Constant

$$F = -G \frac{M_1 m_2}{r^2}$$

According to Newton's law of universal gravitation,

the attractive force (F) between the centres of two bodies of mass-Matter, is proportional to the product of their Matter (M1 and m2), and inversely proportional to the square of the distance (r) between them

Long distance force

spacetime curvature

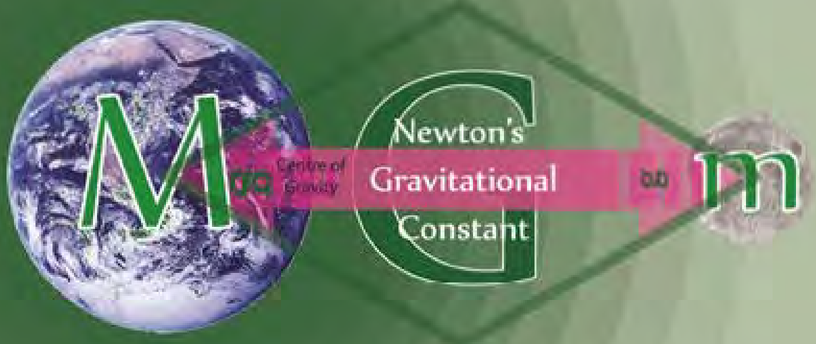

Newton's Gravitional Force models a motive force for gravitational acceleration

Einstein's General relativity has no motive force for gravitational accelerations

6.67384 e-11

$$\left[\frac{Nm}{kg}\right] \cdot \left[\frac{Nm}{kg}\right]$$

The gravitational force between two interacting bodies of Matter is

$$\left[\frac{N\,m^2}{kg^2}\right] = \left[kg\,\frac{m}{s^2}\,\frac{m^2}{kg^2}\right] = \left[\frac{m^3}{kg\,s^2}\right]$$

$4\pi G\rho$

$8\pi G\rho$

Tetryonics 67.04 - The Gravitational constant

Copyright ABRAHAM [2008] - All rights reserved

Action-at-a-Distance

All Matter creates KEM fields of energy momenta

All Matter creates a displacement of vacuum energies

radiant mass-energies
INTERACT

Once established longitudinal EM fields between Matter form an 'inelastic rod' of energy momenta that can facilitate an instantaneous 'Action-at-a-Distance'

standing-wave Matter
ATTRACTS

Energy momenta propagates at 'c' - ACTION moves instantaneously along established lines of longitudinal Electric Force

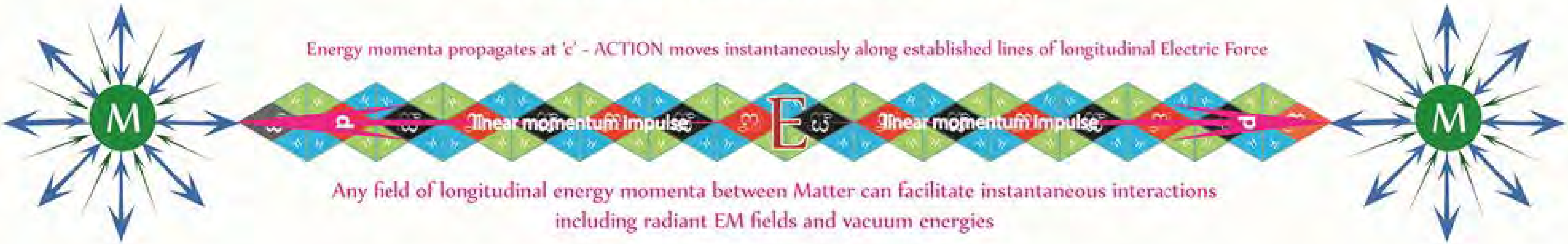

Any field of longitudinal energy momenta between Matter can facilitate instantaneous interactions including radiant EM fields and vacuum energies

Gravity is known as a Universal force of attraction between two objects that is proportional to their mass-Matter content

Newton's theory of gravity offered no prospect of identifying any mediator of gravitational interaction. His theory assumed that gravity acts instantaneously, regardless of distance.

Energy momenta creates universal motive forces

According to Albert Einstein's theory of special relativity, instantaneous action-at-a-distance was seen to violate the relativistic upper limit on speed of propagation of information.

If one of the interacting objects were to suddenly be displaced from its position, the other object would feel its influence instantaneously, meaning information had somehow been transmitted faster than the speed of light.

Gravity fields create a universal attractive field

Gravitational Constant

The gravitational constant denoted by letter G, is an empirical physical constant involved in the calculation(s) of gravitational force between two bodies.
It appears in Sir Isaac Newton's law of universal gravitation, and in Albert Einstein's theory of general relativity.
It is also known as the universal gravitational constant, Newton's constant, and colloquially as Big G.

It should not be confused with "little g", which is the local gravitational field (equivalent to the free-fall acceleration), at the Earth's surface

Gravitational Forces can be modelled

$6.67\text{ e-11}\ \left[\frac{Nm}{kg}\right]^2$

linear Newton metres

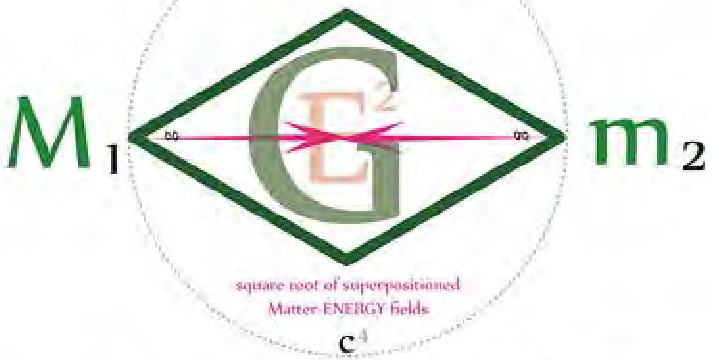

$6.67\text{ e-11}\ \left[\frac{E}{kg}\right]^2$

scalar Energy momentum

as the geometric MEAN of superpositioned fields of gravitational energy

$$F = -G\,\frac{M_1\, m_2}{r^2}$$

The gravitational force between two neutral interacting bodies is

N.m = Joules

$\left[\frac{Nm}{kg}\right]\cdot\left[\frac{Nm}{kg}\right]$

$[N.m/kg]^2 \sim [E/kg]^2$

Gravitational acceleration

Gravitational acceleration can be viewed as a reflection of the change in an object's momentum over time

$$\sum \mathbf{F} = \frac{d\mathbf{p}}{dt} = m\frac{d\mathbf{v}}{dt} + \mathbf{v}\frac{dm}{dt}$$

$$\frac{\left[\frac{m}{s}\right]}{s}$$

$$\vec{g} = \frac{\vec{F}_g}{m} = -G\frac{mM}{mr^2}\hat{r} = -\frac{GM}{r^2}\hat{r}$$

expressed vectorially the force is always ATTRACTIVE to the larger masss and acts in opposition to the KEM field's divergent momenta

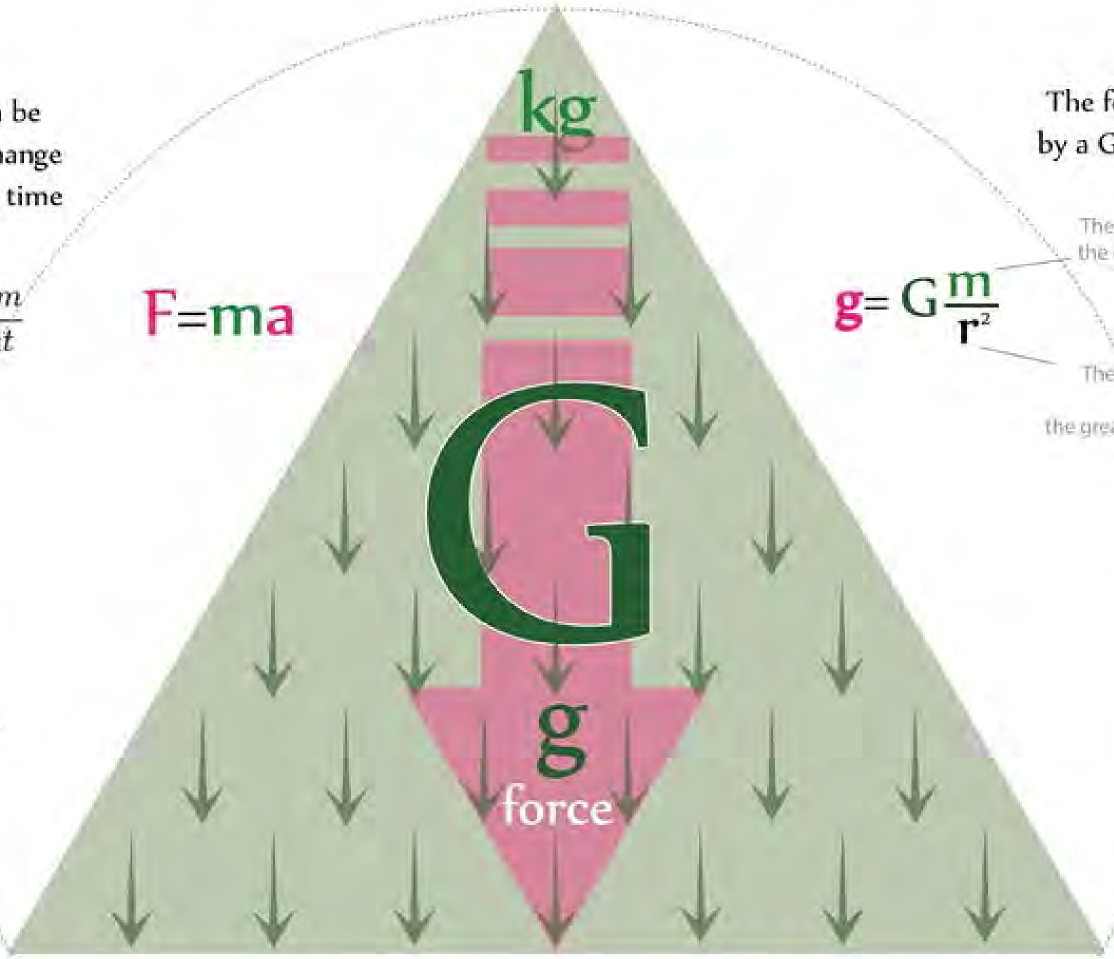

F=ma

The force on a body acted upon by a Gravitational field is given by

$$g = G\frac{m}{r^2}$$

The greater the mass of a gravitating object the greater the acceleration due to the mass

The closer a body is to a the source of a Gravitational field the greater the acceleration due to the field

gravitational acceleration can be modelled as the convergent vector component of superpositioned EM fields

$$\left[\frac{N}{kg}\right] = \left[\frac{m}{s^2}\right] = \frac{\left[kg\frac{m}{s^2}\right]}{kg}$$

$$\left[\frac{Nm}{kg}\right]^2$$

The gravitational force between TWO neutral interacting bodies is

$$\left[\frac{Nm}{kg}\right].\left[\frac{Nm}{kg}\right]$$

Modelling Gravitational fields

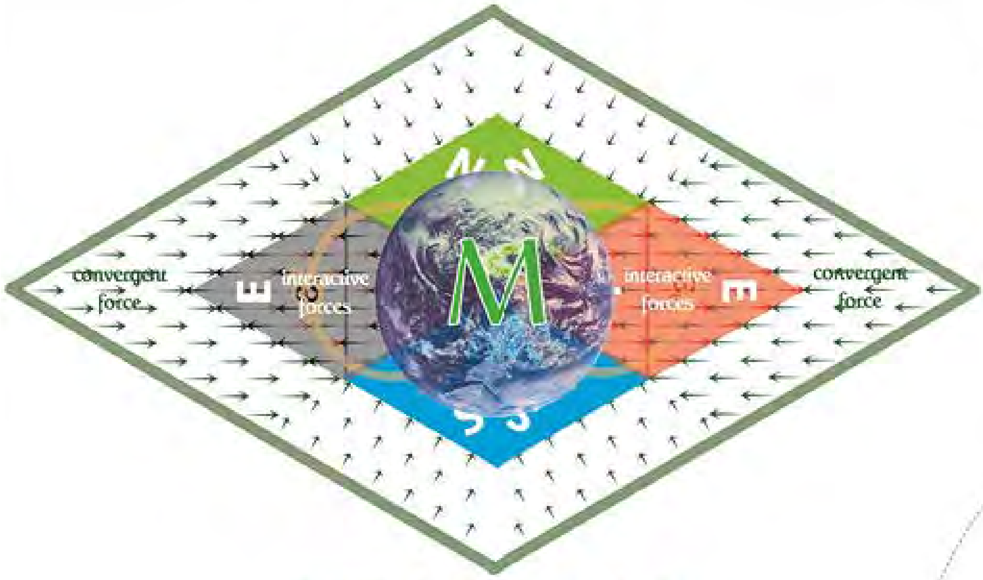
N
S
M
E
E
convergent force
interactive forces
interactive forces
convergent force

The [K]EM fields of planetary bodies are the result of their mass-Matter density and motion in space

Gravity can be modelled as weak convergent force created by and acting between the mass-Matter in planetary bodies

Matter topologies displace vacuum energies creating Gravity

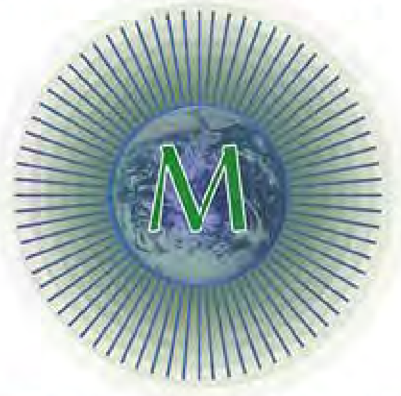

Weak KEM fields can accurately model Gravitational fields

Gravitational fields are the result of the displacement of vacuum energies by Matter topologies and the interactive energy momenta of their radiant KEM fields

$$F = k\frac{Q_1 Q_2}{r^2}$$

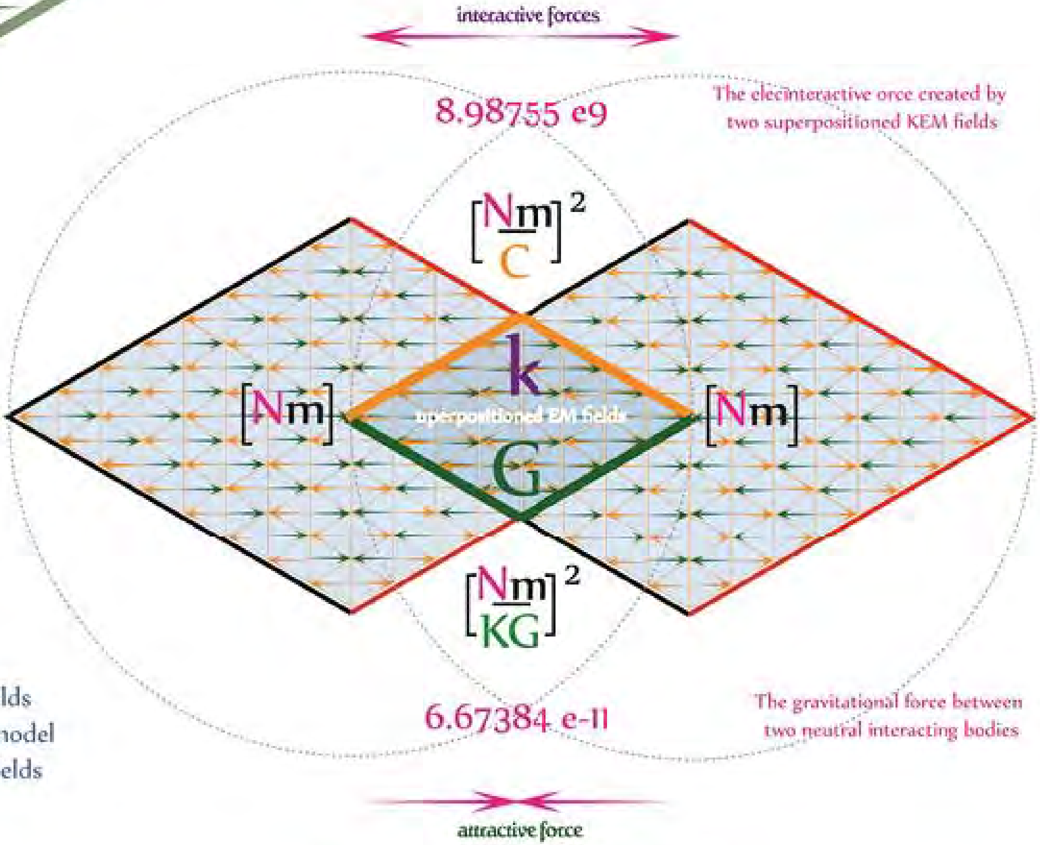
interactive forces
8.98755 e9
The elecinteractive orce created by two superpositioned KEM fields
$\left[\frac{Nm}{C}\right]^2$
k
[Nm]
superpositioned EM fields
[Nm]
G
$\left[\frac{Nm}{KG}\right]^2$
6.67384 e-11
The gravitational force between two neutral interacting bodies
attractive force

$$F = -G\frac{M_1 M_2}{r^2}$$

Centres of Gravity

are determined by
the distribution of the combined
gravitational mass-Matter in a system

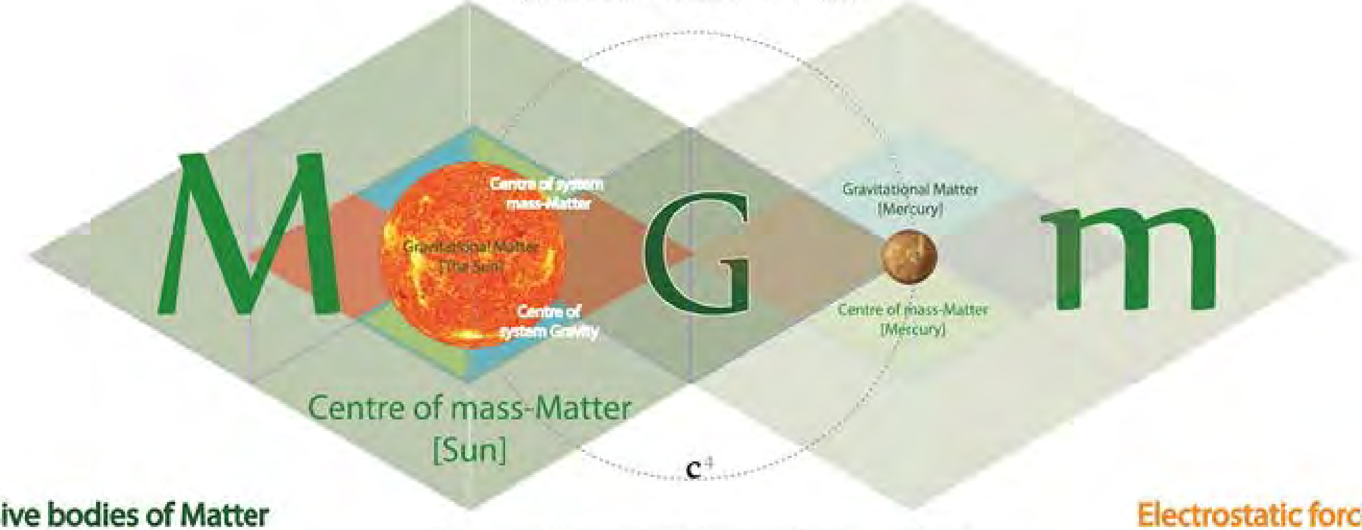
M
Centre of system
mass-Matter
Gravitational Matter
[The Sun]
Centre of
system Gravity
G
Gravitational Matter
[Mercury]
Centre of mass-Matter
[Mercury]
m
Centre of mass-Matter
[Sun]
c^4

Any two massive bodies of Matter (be they Stars or Neutrinos) will orbit about their 'Centre of Gravity'

The term center of mass is often used interchangeably with center of gravity, but they are physically different concepts. They happen to coincide in a uniform gravitational field, but where gravity is not uniform, center of gravity refers to the mean location of the gravitational force acting on a body.

Electrostatic forces are many orders of magnitude greater than gravity & dominate nuclear interactions at the quantum scale

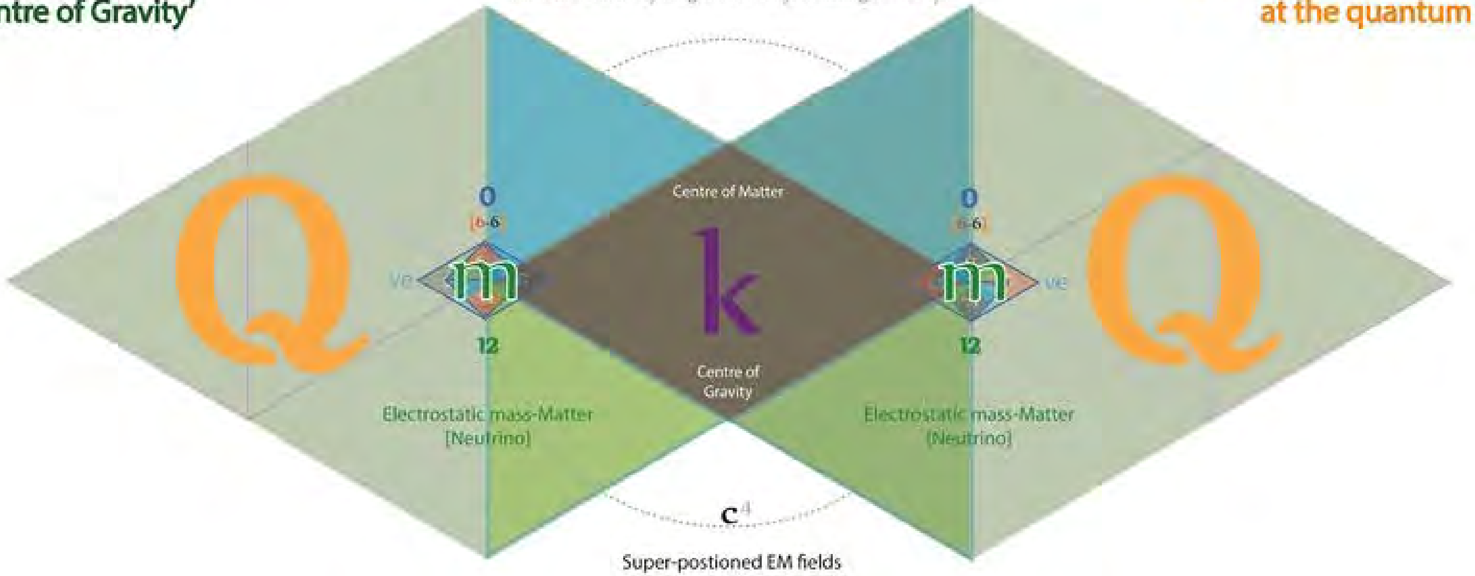
Q
0
[6-6]
ve
m
12
Centre of Matter
k
Centre of
Gravity
Electrostatic mass-Matter
[Neutrino]
0
[6-6]
m
ve
12
Electrostatic mass-Matter
[Neutrino]
Q
c^4

Super-postioned EM fields
have c^4 geometries [c^2x c^2]

$$F=ma$$

There are two quite distinct types of mass in Newton's theory of mechanics and gravitation.

(i) inertial mass, which occurs as the ratio between force and acceleration in Newton's second law and thus measures a particle's resistance to acceleration, and

(ii) Gravitational mass, which may be regarded as the gravitational analog of electric charge, and which occurs in Newton's Gravitational equation

$$F_g = -G\frac{m}{r^2}$$

Inertial mass
is a measure of the total
energy momenta geometries
of Matter topologies

All Gravitational Matter
has EM mass geometries in
its Matter topology

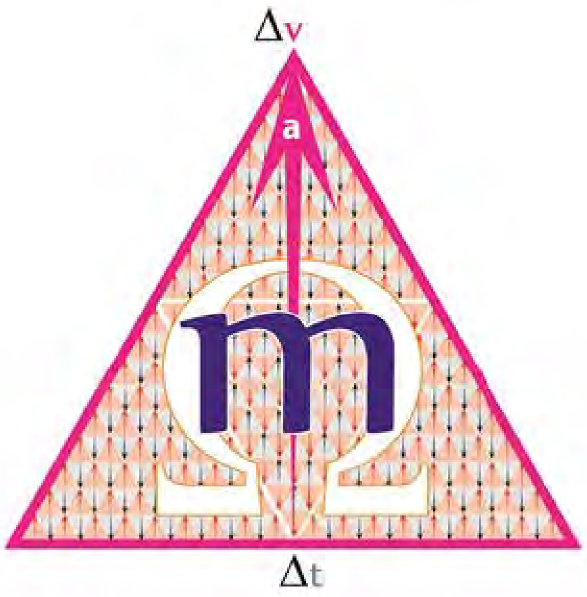

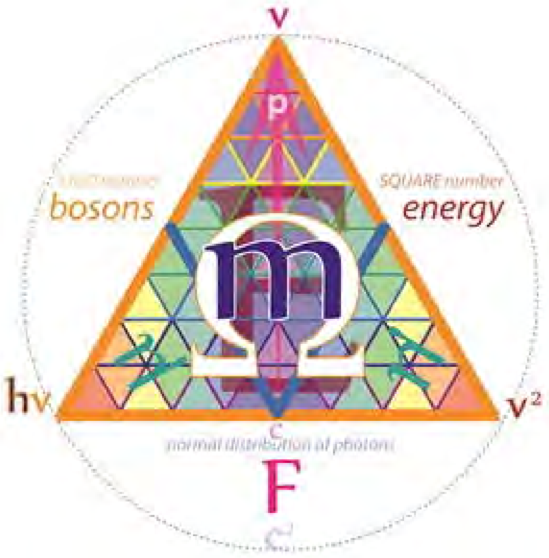

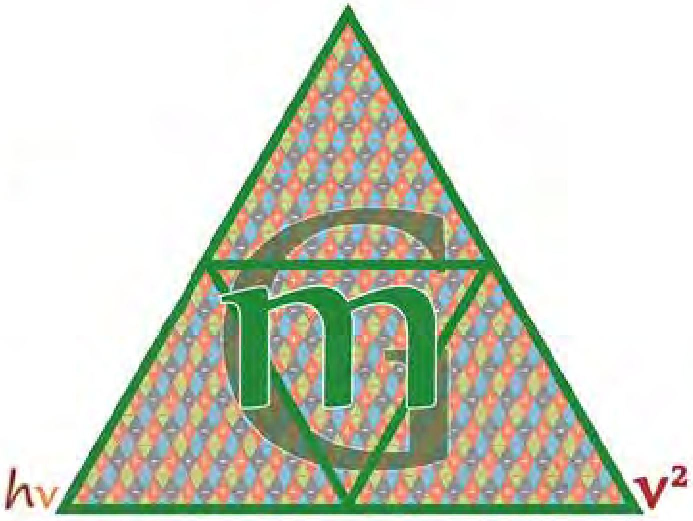

Often stated as the 'Principle of Equivalence'
on which General relativity is founded ,
it can be summarised as follows:

$$a = \frac{m}{m} g$$

Inertial mass

results from constituent quanta
acting against (or with) any
Force applied to a mass

*"Inertial mass and Gravitational mass
are equivalent"*

***whilst mass-energies contribute equally
to inertial & gravitational mass
it is Matter topologies that creates Gravity***

'Gravitational' mass

is seen as being the constituent quanta
of opposite charge polarity
attracting another particle's quanta

Copyright ABRAHAM [2008] - All rights reserved

Discerning between the competing laws of Gravity

$$M + KE = RE$$

gravitational mass-Matter

fields of force

relativisitic mass-energy densities

At the heart of resolving the true nature of Gravity is the formal definitions of mass-Energy-Matter and the interplay between the various field interactions they produce and possess

Newton's view of gravity

Sir Isaac Newton described gravity as a force that attracts things possessing Matter

Sir Issac Newton

(25 December 1642 – 20 March 1727)

Gravity is an instantaneous 'Action-at-a-Distance' mediated by a means unknown

$$F = -G\frac{M_1 m_2}{r^2}$$

Newton taught that masses attract each other with a force inversly proportional to the distance between the masses.

Newton's idea was that planets are held in their orbits around the Sun by a force proprtional to the mass of the Sun and the planets, and that force is always pointing inwards (i.e. planets get pulled in the direction of the Sun, and the Sun gets pulled in the directions of the planets).

Newton said that a person standing on the Earth is being held to the Earth by a force due to the mass of the Earth and the mass of the person, and that force is always pointing inwards (i.e. a person gets pulled in the direction toward the center of the Earth, and the Earth gets pulled in the direction of that person).

Newton's and Einstein's view of gravity differ markedly

Einstein, on the other hand, taught that all things with mass (including the Sun, Earth, and the rest of the planets) create a energy density gradient that causes spacetime to curve and that this curvature is an alteration the geometry of spacetime.

The curvature casues Matter to move toward the center of the largest nearby mass. The closer to this mass the larger the degree of curvature in the local spacetime.

Additionally, Einstein taught that if there are no external forces acting upon an object (even those without any mass such as photons) then that object will simply follow its natural path (geodesic) through the altered geometry of the local spacetime - regardless of the degree of curvature of that local spacetime. This implies that if an object is somehow kept from following its natural path (geodesic) then a net force does exist and is measurable on that object.

Ricci curvature scalar curvature Gravitational constant

$$R_{\mu\nu} - \tfrac{1}{2}R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

stress-energy tensor

spatial geometry

Einstein's view of gravity

Albert Einstein described gravity as a 3D curvature of spacetime caused by Matter

Albert Einstein

(14 March 1879 – 18 April 1955)

Matter and energy tell spacetime how to curve. Spacetime tells matter how to move

There is no doubt that both Sr Isaac Newton & Albert Einstein were both brilliant, their theories describing Universal Gravitation have remained at the fore for hundreds of years now.

However, a more accurate model of the Universal Gravitation is now at hand and its true nature is revealed via the Tetryonic geometry of EM mass-Energy-Matter and its associated field interactions. Tetryonic Gravitation demonstrates the validity of both appoaches and reveals the true geometry behind the field interplays that result in the familiar force of Gravity

Tetryonics 67.11 - The competing Laws of Gravity

Copyright ABRAHAM [2008] - All rights reserved

The two models of Gravitation

Newtonian gravitation models the observed large-scale convergent force between all material bodies

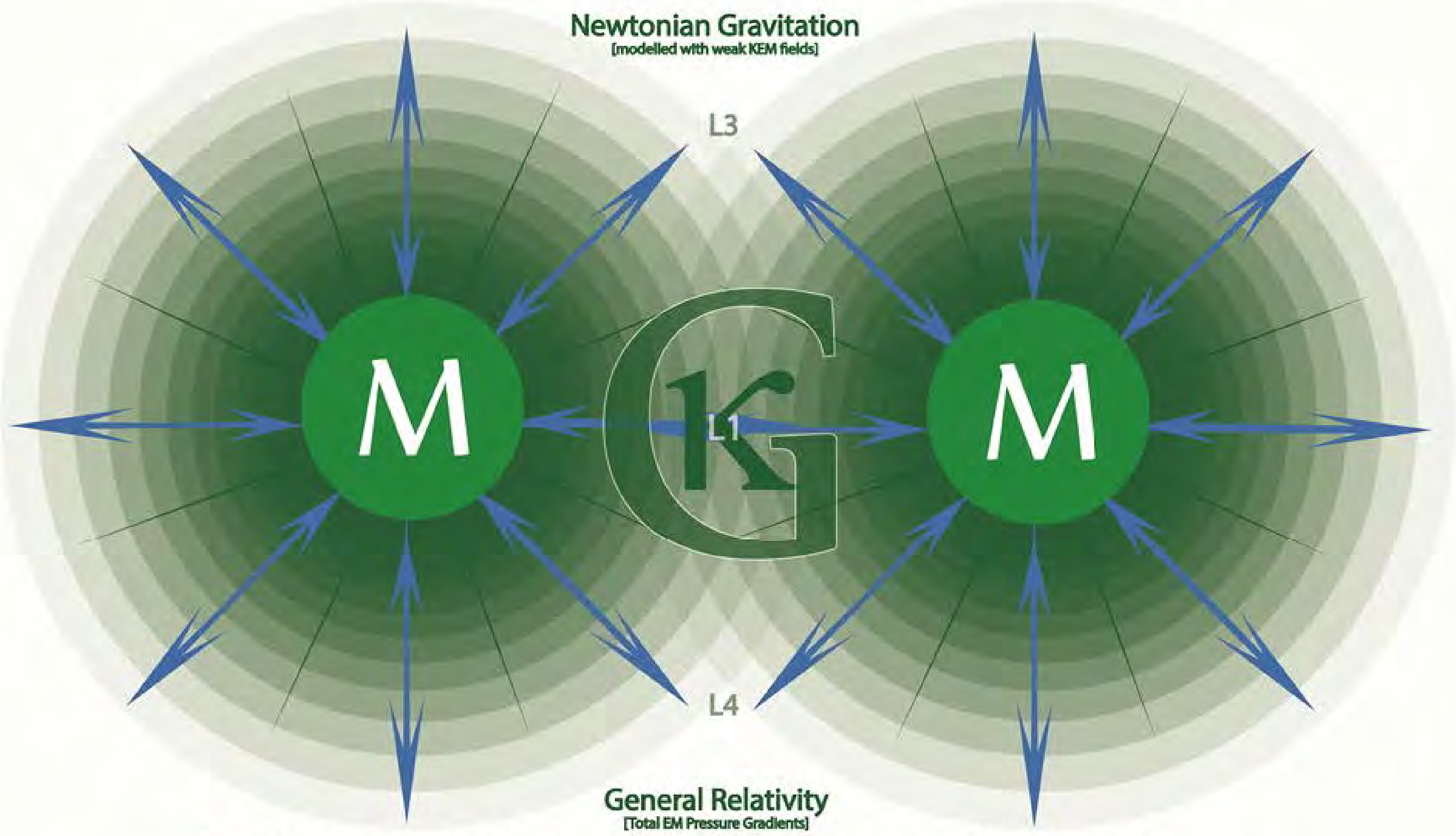
Newtonian Gravitation
[modelled with weak KEM fields]
L3
M
G
K
L1
M
L4
General Relativity
[Total EM Pressure Gradients]

General Relativity models the convergent pressure gradiant created by all forms of energy

While both models are very successful they both fail to reveal the quantum-scale geometries and topologies that give rise to universal gravitation

Theories of gravitation

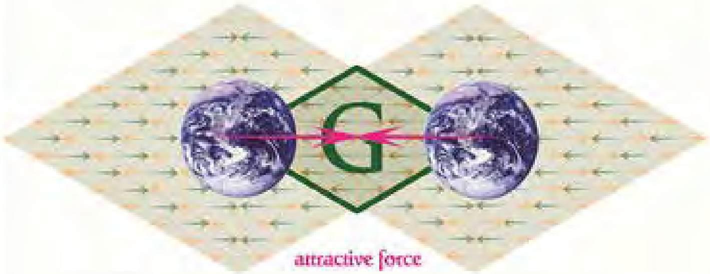

for the mechanics of gravitation
"I feign no Hypothesis"

"Gravity is the result of
curved spacetime"

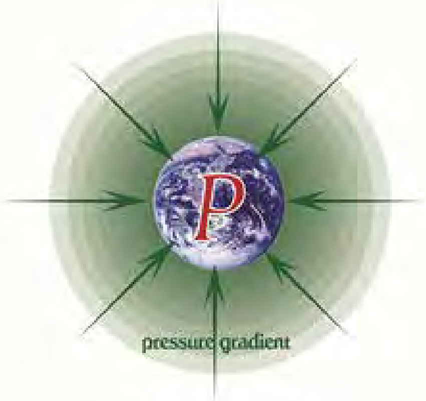

Newton viewed Gravitation as
an attractive FORCE of Matter
(inverse square field strength)

Einstein viewed it as the
curvature of 4D spacetime
(PRESSURE gradient)

$$F = -G \frac{M_1 M_2}{r^2}$$

$$G_{ab} = \kappa T_{\mu\nu}$$

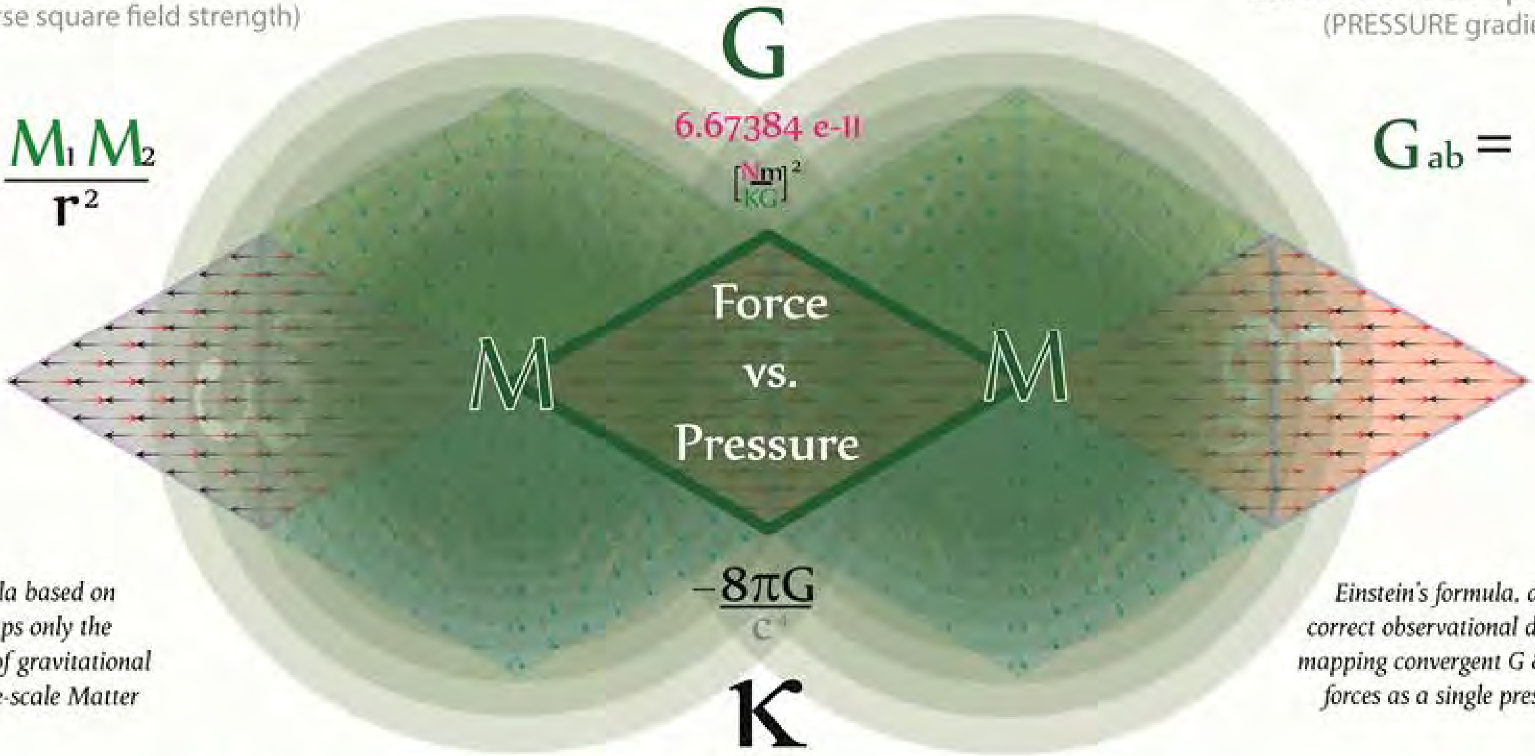

G

6.67384 e-11

$$\left[\frac{Nm}{KG}\right]^2$$

Force
vs.
Pressure

M M

$$-\frac{8\pi G}{c^4}$$

$$\kappa$$

*Newton's formula based on
observations maps only the
observed nett force of gravitational
attraction on large-scale Matter*

*Einstein's formula, attempted to
correct observational discrepancies by
mapping convergent G & divergent EM
forces as a single pressure gradient*

Both Newton and Einstein failed to define and distingush between mass-energy geometries & Matter topologies

Copyright ABRAHAM [2008] - All rights reserved

Newton's G vs Einstein GR

Both Newton and Einstein developed their gravitational models
from the observed motions of celestial bodies without any distinction
between mass and Matter or the quantum forces at work

Gravitational Matter

Newton maps Matter's gravitational attraction
[excluding any perturbations caused by [K]EM fields]

All mass-energies

Einstein's GR maps Newton's G & SR effects
[mass-energy momenta [K]EM interactions]

Φ

F

$\nabla^2\Phi$

ρ

$g_{\mu\nu}$

$\Gamma^{\sigma}_{\mu\nu}$

$R_{\mu\nu}$

$T_{\mu\nu}$

Increasing accuracy in the measurements of the motions
of spacecraft interacting with planetary bodies [using EM waves]
has resulted in the need for refinement of both theories

Einstein's doubling of Newton's gravitational constant
was done to fit later observed perturbations in
Mercury's orbital mechanics

Newtonian G field

Any valid tests of the purely gravitational fields of Matter
must exclude KEM fields and not use EM masses to test it

$$[\text{Newton}] = \frac{4\pi GMm}{r^2}$$

$$\nabla^2\Phi = 4\pi G\rho$$

Poisson

$$[\text{Einstein}] = \frac{8\pi GT_{\mu\nu}}{c^4}$$

Pressure gradients and GR

Tensor mathematics is utilised in General relativity to model mass-energy geometries and energy pressure gradients as curved spacetime

$$F = -G\frac{M_1}{r^2}$$

Newton's Gravitational fields are modelled on the observed motions of large-scale Matter attraction

Newtonian G field

$$4\pi$$

G

Vector Force between Matter

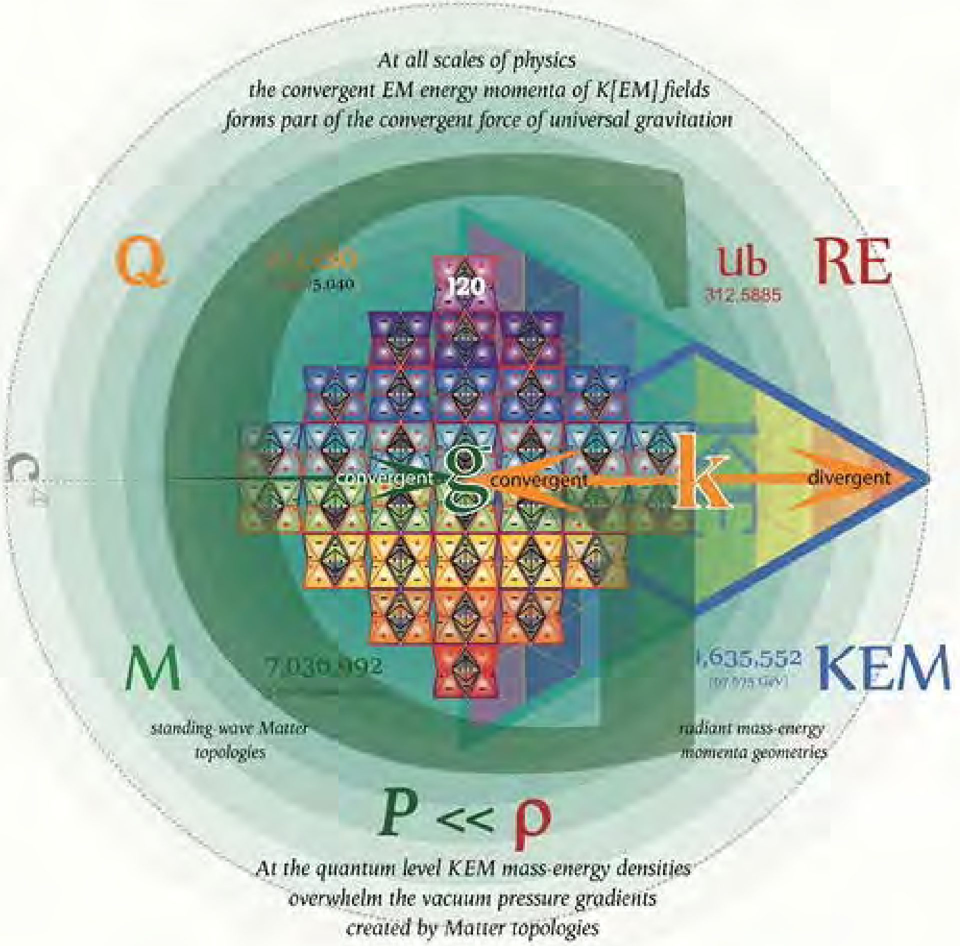

At all scales of physics the convergent EM energy momenta of K[EM] fields forms part of the convergent force of universal gravitation

standing-wave Matter topologies

radiant mass-energy momenta geometries

$$P \ll \rho$$

At the quantum level KEM mass-energy densities overwhelm the vacuum pressure gradients created by Matter topologies

$$G_{ab} = \kappa T_{\mu\nu}$$

Einstein's GR gradients corrects Newtonian gravity by modelling all forms of EM mass-Energy momenta

Einstein's G field

$$8\pi$$

K

Pressure gradient per unit Volume

$$n\pi\rho = \underset{\text{geometries}}{n\pi}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right] = T_{\mu\nu}$$

representive of ALL the mass-energy momenta of any spatial volume

Stress energy-pressure gradiants

are the result of Tetryonic charge geometries & topologies

$$\frac{\overset{\text{mass}}{n\pi}}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Radiant mass-energies have planar energy geometries

$n\pi$	**Energy quanta**
EVENπ	**photons**
ODDπ	**bosons**
4π	**tetryons**
12π	**quarks**
12π	**leptons**
24π	**mesons**
36π	**Baryons**

All Matter topologies are tetrahedral standing-wave mass-energy geometries

$$\frac{\overset{\text{Matter}}{T\pi}}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

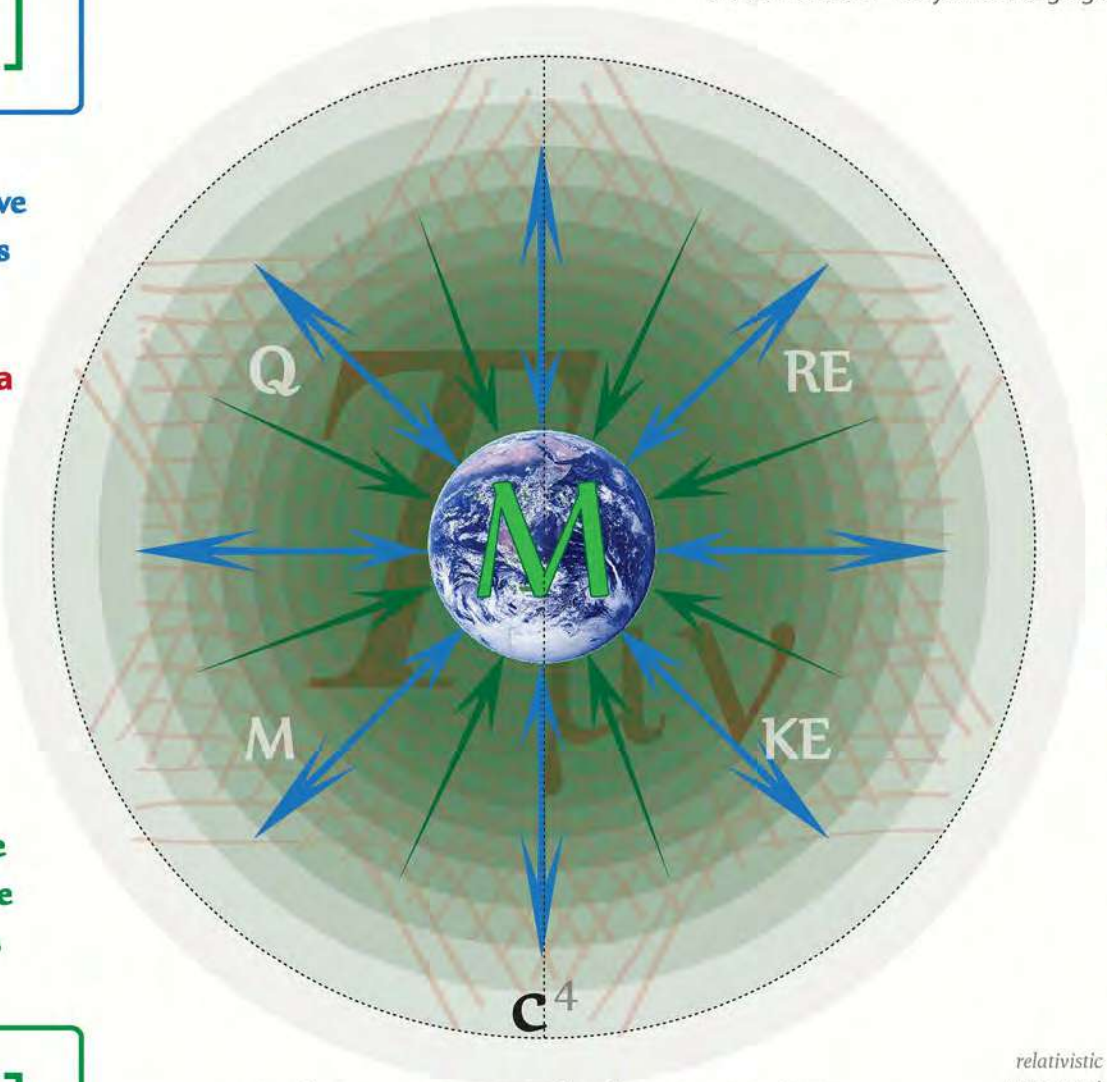

On the non-quantum scale all Matter topologies can be generalised and mathematically modelled as SPHERICAL 4π mass-energy topologies

radiated [K]EM fields
have interactive energy momenta
[convergent & divergent]

fields of Force

KEM

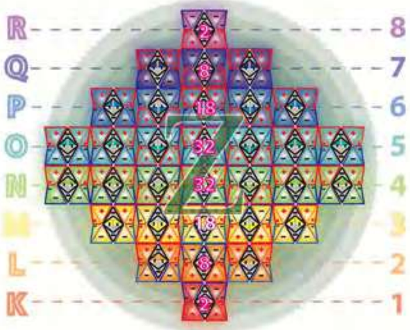

M

standing-wave topologies

displace vacuum energies
creating convergent Gravity

$$\underset{\text{stress energies}}{\overset{\text{relativistic mass-energies}}{RE}} = \underset{\text{topologies}}{\overset{\text{standing-wave Matter}}{M}} + \underset{\text{geometries}}{\overset{\text{radiant mass-energies}}{KEM}}$$

Copyright ABRAHAM [2008] - All rights reserved

Stress Energy tensor

In the Einstein field equations [EFE] of general relativity, the stress-energy tensor is the source of gravitational mass just as Matter is the source of such a field in Newtonian gravity

$$T_{\mu\nu}$$

In Einstein's general relativity,
the symmetric stress-energy tensor
acts as the source of spacetime curvature

mass *ENERGY* *Matter*

m E M

geometries *momenta* *topologies*

Matter Planck quanta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

The Tetryonic unified field equation
models all mass-ENERGY geometries
& Matter topologies

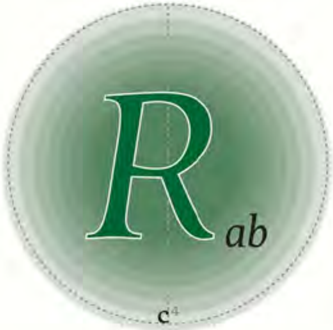

A recognised problem in General Relativity is that
the stress tensor models all forms of mass-energy-Matter
in the form of a generalised stress energy pressure gradient
but fails to differentiate between either mass or Matter
and their resulting interactions within the fields

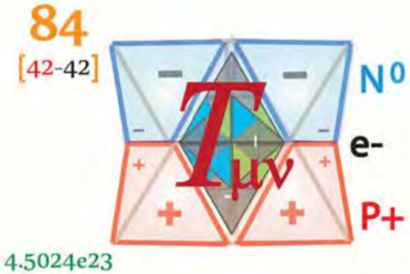

Planck quanta

$$n\pi\left[m\Omega v^2\right]$$

mass velocity

radiant mass-energies
create interactive
fields of force

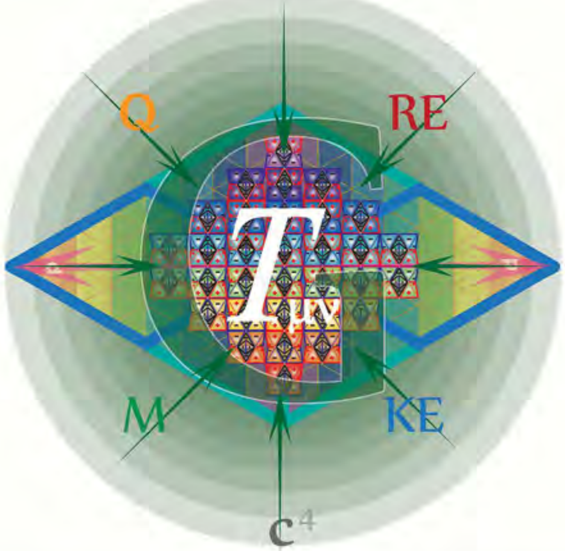

$$T\pi[\rho]$$

standing-wave Matter
creates vacuum energy
pressure gradients

kg

2D mass geometries

EM mass is not Matter

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Charged energy geometries are the basis for 2D mass

Photons have mass-energy momenta geometries but no Matter topology

Gravimetric topologies

Charges create radiant mass-energy geometries & standing-wave Matter topologies

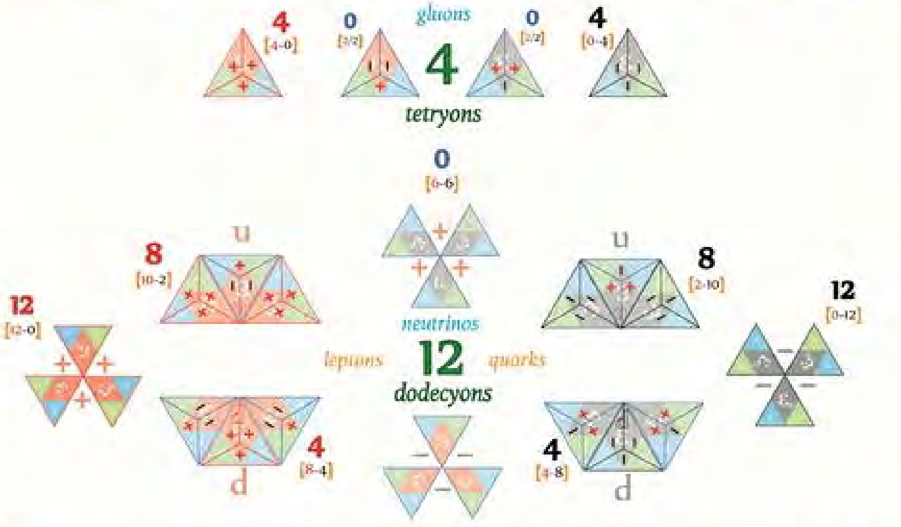
4
[4-0]
0
[2/2]
gluons
4
tetryons
0
[2/2]
4
[0-4]
0
[6-6]
8
[10-2]
u
12
[12-0]
neutrinos
leptons
12
quarks
dodecyons
u
8
[2-10]
12
[0-12]
4
[8-4]
d
4
[4-8]
d

2D mass-energy geometries form surface integral areas of charged Matter topologies

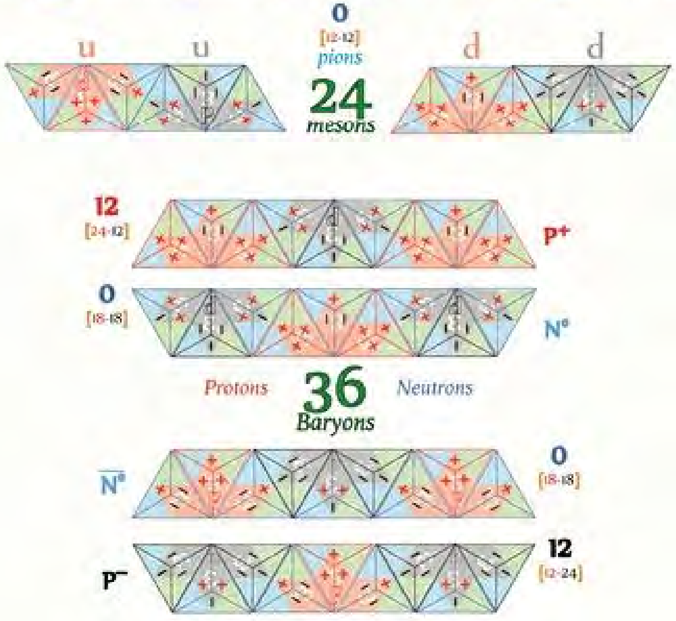
0
[12-12]
pions
u
u
24
mesons
d
d
12
[24-12]
P+
0
[18-18]
N°
Protons
36
Baryons
Neutrons
N°
0
[18-18]
P⁻
12
[12-24]

KG

3D Matter topologies

Matter has mass energies

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

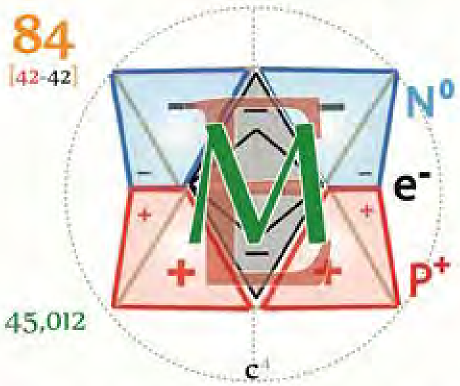
84
[42-42]
N⁰
e⁻
P⁺
45,012
c⁴

Charged mass-energy topologies are the basis for 3D Matter

Its topology forms a quantum Faraday cage [with a charged surface integral of mass-energies]

Copyright ABRAHAM [2008] - All rights reserved

Interactive mass-energies vs. gravitational Matter

energy in all its forms seeks equilibrium

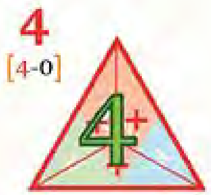

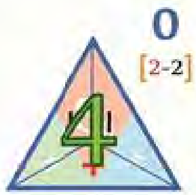

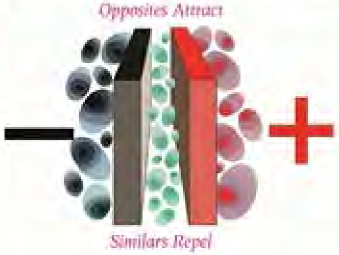

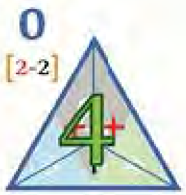

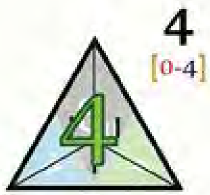

radiant mass-energy geometries

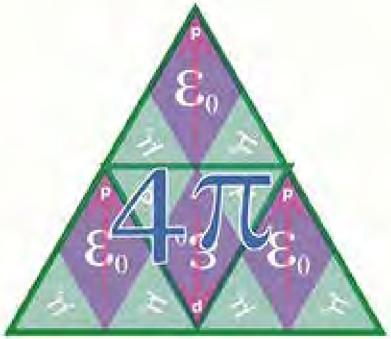

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

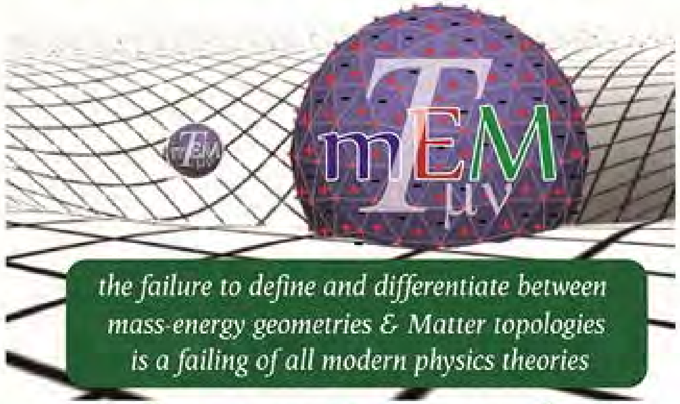

the failure to define and differentiate between mass-energy geometries & Matter topologies is a failing of all modern physics theories

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

standing-wave Matter topologies

Charged Planck quanta of mass-energy momenta that form topologies displace the vacuum energies surrounding them in turn creating a mass-energy-Matter pressure gradient

Compton frequency of rest Matter is what creates inertial mass, its Gravitational Matter results from its charged topology

The integral inductive EM mass-energy momenta in Matter fascia create the physical property of inertial mass

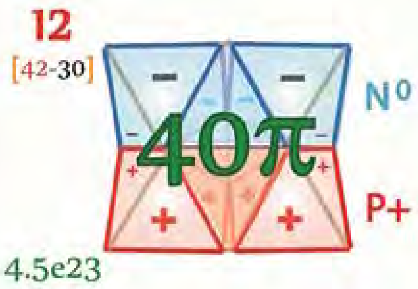

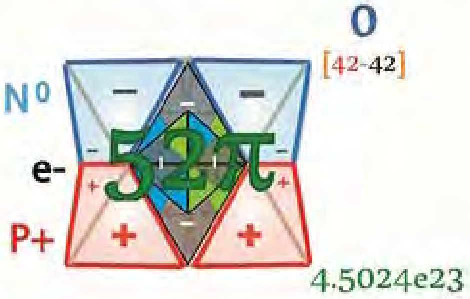

$$F_z = \Delta P . \Delta A$$

Pressure is the force per unit area
applied in a direction perpendicular to the surface of an object

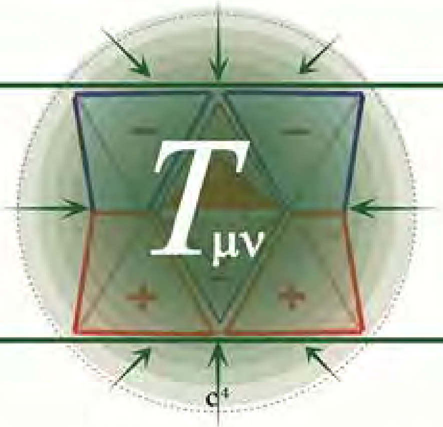

The theory of general relativity uses a stress-energy tensor to reflect to the pressure differentials created by Energy in ALL its forms (making no distinction between EM masses & Matter)

In Tetryonics the stress energy tensor is expanded, through geometric terms, to diferentiate between radiant mass-energy interactions and Matter topologies in radial spatial co-ordinate systems defined by the speed of light

Radiant EM masses

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Photons are often referred to as 'massless particles' are better defined as Matterless EM mass-energy momenta

ω_0
n
μ_0
ε_0

$$\overset{\text{Photons}}{\frac{2\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

interactive mass-energy momenta geometries

Energy Pressure gradients

The vacuum energy pressure gradient created by the charged topology of Matter is distinct from its intrinsic & radiative mass-energy geometries

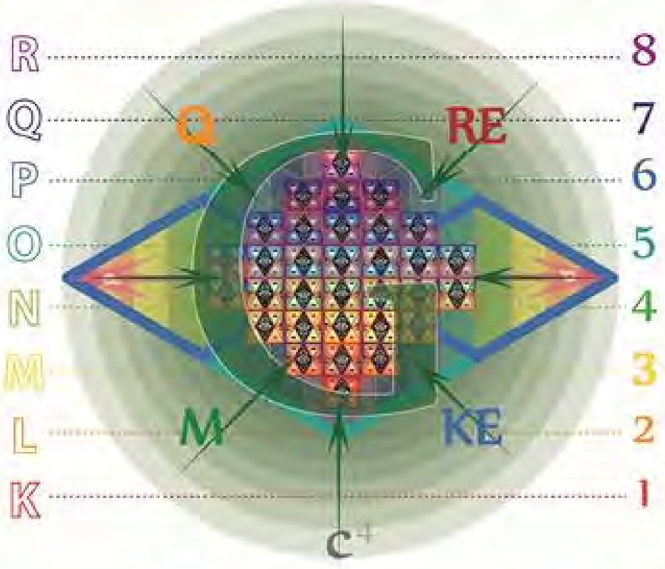

Any theory of gravitation that attempts to model it on any scale must account for energy in all its forms

Standing-wave Matter

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Tetryons, the quantum building blocks of Matter have EM mass-energy topologies that create a vaccum pressure gradient

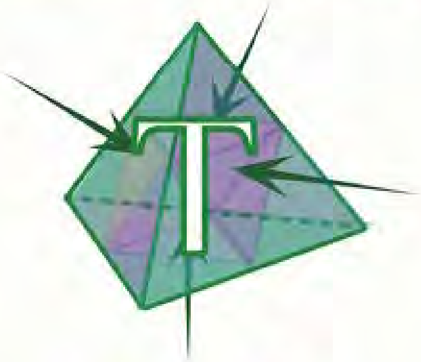

$$\overset{\text{Tetryons}}{\frac{4\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

gravitational Matter topologies

Convergent mass-energies & Gravitational Matter

m

Spacetime curvature

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Stress-Energy Tensor

M

The concept of mass in general relativity (GR) is more complex than the concept of mass in special relativity. In fact, general relativity does not offer a single definition for the term mass, but offers several different definitions which are applicable under different circumstances.

Komar mass **Bondi mass** **ADM mass**

Under some circumstances, the mass or Matter components of a system may not even be defined

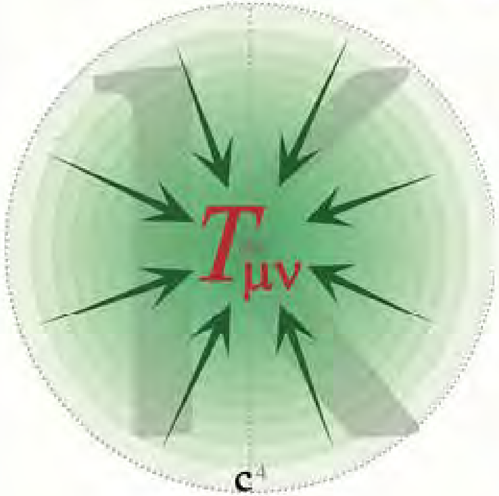

Einstein Field equation

$$\overset{\text{Spacetime curvature}}{R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}} = \underset{\text{constant}}{\overset{\text{gravitation}}{K}} \overset{\text{total mass-energy-Matter}}{\underset{\text{per spatial volume}}{\frac{n\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]}}$$

Tetryonic geometries & topologies

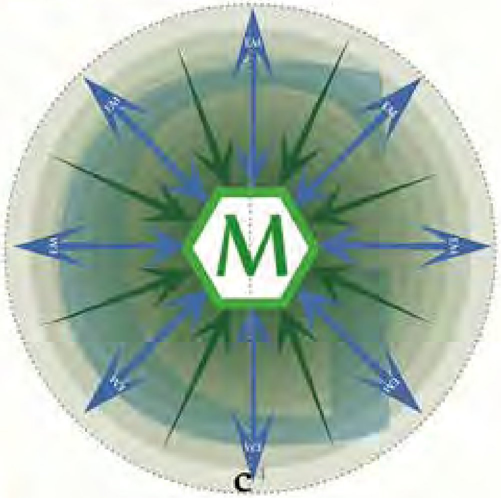

In contrast Tetryonic field equations explicitly differentiate all EM mass-energies and Matter via their charged geometries along with their contributions to the total energy density of any spatial region

GR models the convergent force of gravitation as a convergent energy density pressure gradient created by mass-energies

$$\underset{\text{mass-Matter}}{\overset{\text{gravitational}}{M}} + \underset{\text{mass-energies}}{\overset{\text{interactive}}{kEM}} = \underset{\text{energies}}{\overset{\text{stress}}{U}}$$

Tetryonics models gravitation as the result of mass-energy interactions AND Matter topology displacement of vacuum energies

Interactive EM force fields are subject to Lorentz velocity corrections

Faraday

$$\mathcal{E} = -\frac{d\Phi_B}{dt},$$

$$|\mathcal{E}| = N\left|\frac{d\Phi_B}{dt}\right|$$

E & M

EM

$$\oint \vec{E}\cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

$$\oint \vec{B}\cdot d\vec{A} = 0$$

$$\oint \vec{B}\cdot d\vec{s} = \mu_0 i + \frac{1}{c^2}\frac{\partial}{\partial t}\int \vec{E}\cdot d\vec{A}$$

Maxwell

Special Relativity was developed from the foundation of Lorentz corrections to EM fields

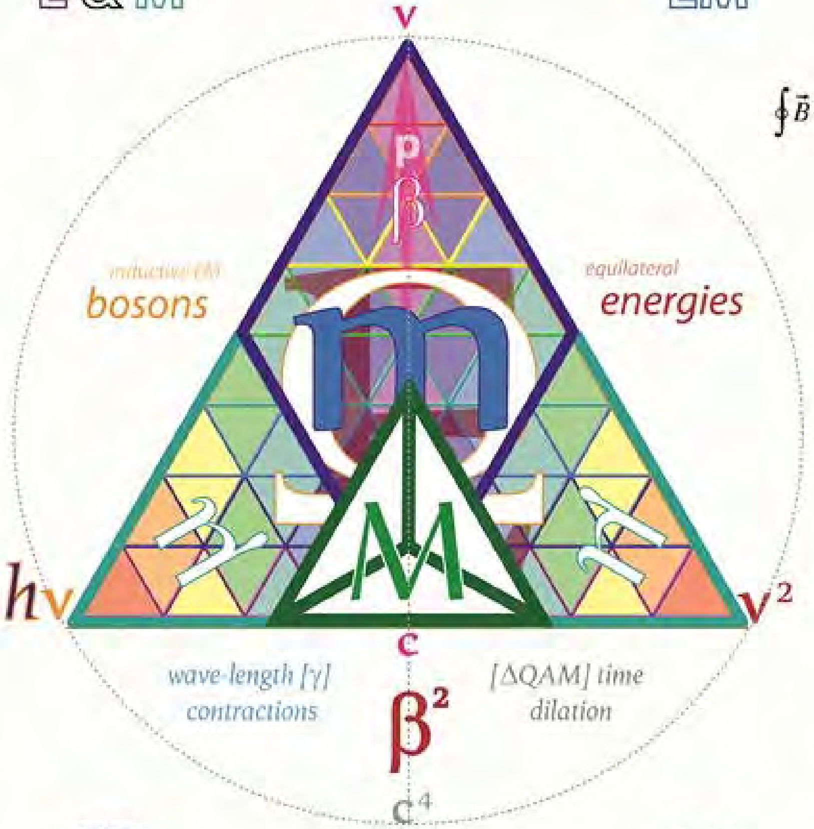
v
p
β
m
M
inductive EM
bosons
equilateral
energies
hv
v²
c
wave-length [γ] contractions
β²
[ΔQAM] time dilation
c⁴

General Relativity was developed from Special Relativity and incorporates EM field interactions

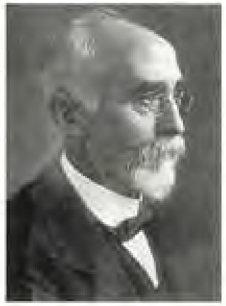

Lorentz

$$L = L'\sqrt{1-\frac{v^2}{c^2}}$$

$$t' = \frac{t}{\sqrt{1-\frac{v^2}{c^2}}}$$

SR

seconds^2

GR

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}$$

$$E = \frac{mc^2}{\sqrt{1-\frac{v^2}{c^2}}}$$

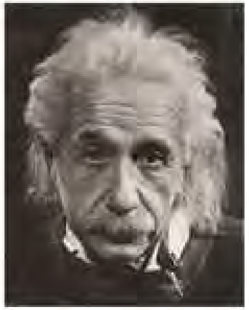

Einstein

Standing-wave Matter topologies are Lorentz invariant

the failure to differentiate between mass & Matter has lead to major misconceptions in physics

Tensor Rankings

Rank 0

scalars

A tensor may consist of a single number, in which case it is referred to as a tensor of order zero, or simply a scalar

An example of a scalar would be the EM mass-energy of a particle or field

A second example of a scalar field would be the value of the gravitational potential energy as a function of position

Rank 1

The next most complicated tensor is the tensor of order one, otherwise known as a vector.

vectors

A vector can be defined as a direction between two points, its strength may vary continuously from point-to-point, thereby defining a vector field

Rank 2

mass-energy momenta

A tensor of rank 2 is defined as a system that has a magnitude and two associated vector directions within the fields that it describes

Next above a vector are tensors of order 2, which are often referred to as matrices.

Rank 3

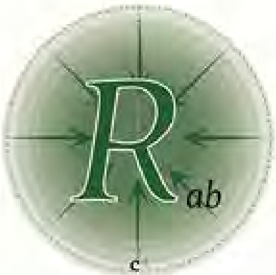

curved spacetime

Rank 4 Tensors as used in Einstein's General Theory of Relativity, describe the curvature of spacetime [the so-called Riemann curvature tensor].

Tetryonic Matrices

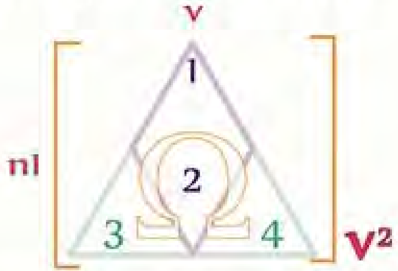

EM fields can be represented by Tetryonic Matrices [Tensors] identical to current square Tensor Matrices with the only difference being the change in the geometry of the matrix to better reflect the true geometry of EM fields

This better facilitates the easy recognition of Electric & Magnetic scalar fields and their associated energy-momenta vectors

Rank 4

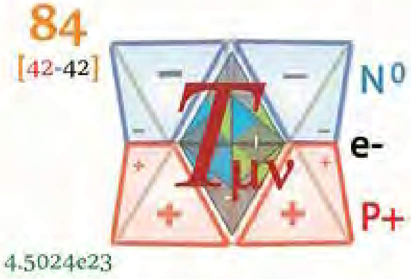

mass-Matter topology

the total mass-energies of Matter & fields thatcontribute to creating gravity.

ElectroMagnetic field tensor

The electromagnetic tensor or electromagnetic field tensor (sometimes called the field strength tensor, Faraday tensor or Maxwell bivector) is a mathematical object that describes the electromagnetic field of a physical system in Maxwell's theory of electromagnetism

Energy

$$2\pi\left[\frac{1}{c^2}\cdot\left[m\Omega v^2\right]\right]$$

Photons momenta

Photons

Planck quanta

$$\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Visualising EM waves and photons in terms of their Tetryonic geometry the scalar relationships between Electric and Magnetic fields becomes readily apparent.

$$F^{\mu\nu} = -F^{\nu\mu}$$

photons are their own anti-particles

A EM field tensor has six degrees of freedom, representing the three components of each of the electric and magnetic fields

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$$

Speed of Light

EM wave factors such as Lorentz's velocity corrections arise naturally from the geometry providing a clear and concise illustration of the relationships between EM mass-energies, Charge, Electric & Magnetic fields and the speed of Light

Wavelength contraction

$$L = L'\sqrt{1 - \frac{v^2}{c^2}}$$

speed of Light

$$c = \left[\left[\frac{c^2}{\Omega}\right]\cdot\left[\frac{\Omega}{c}\right]\right]$$

velocity of propagation

f Frequency

λ Wavelength

$$\vec{\nabla}\cdot\vec{E} = \frac{\rho}{\epsilon_0}$$

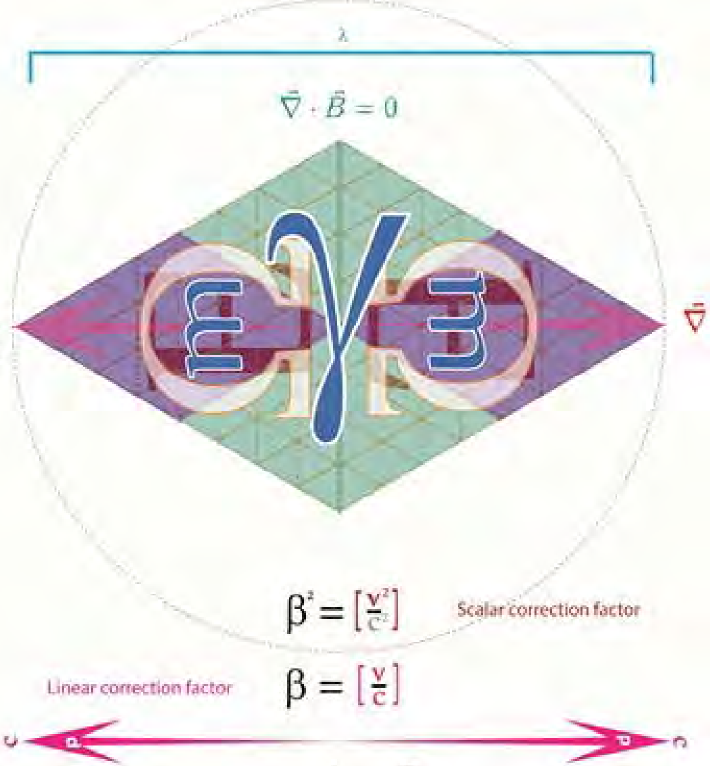

$$\vec{\nabla}\cdot\vec{B} = 0$$

$$\vec{\nabla}\cdot\vec{E} = \frac{\rho}{\epsilon_0}$$

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Scalar correction factor

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

$$p = \frac{h}{\lambda} = \frac{E}{c}$$

Linear Momentum

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

EM Fields

$$\frac{\partial\vec{B}}{\partial t} + \vec{\nabla}\times\vec{E} = 0$$

the change of the magnetic field with respect to time plus the curl of the electric field is equal to zero

$$\vec{\nabla}\times\vec{B} - \frac{1}{c^2}\frac{\partial\vec{E}}{\partial t} = \mu_0\vec{J}$$

The curl of the magnetic field vector minus the change of the electric field with respect to time, is equal to the field's current density

$$\det(F) = \frac{1}{c^2}\left(\vec{B}\cdot\vec{E}\right)^2$$

the electromagnetic tensor may also be written in terms of the 4-vector potential

Copyright ABRAHAM [2008] - All rights reserved

3D space time

$$[\Delta x, \Delta y, \Delta z] \quad \Delta\tau$$

In Tetryonics, the relativistic Four vector reflects the changing quantised angular momenta [time] of EM mass-energies & Matter in 3D space

Relativistic Four-vectors

Charged Matter interacts at-a-distance through super-positioned energy momenta fields from which our force constants are derived

The electromagnetic four-potential is a potential from which electromagnetic field can be derived. It combines both the electric scalar potential and the magnetic vector potential into a single space-time four-vector

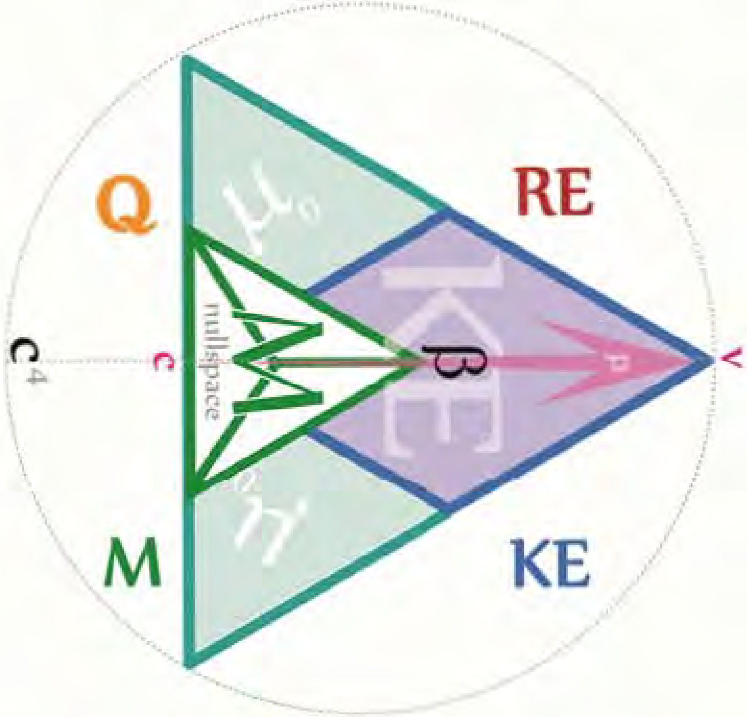

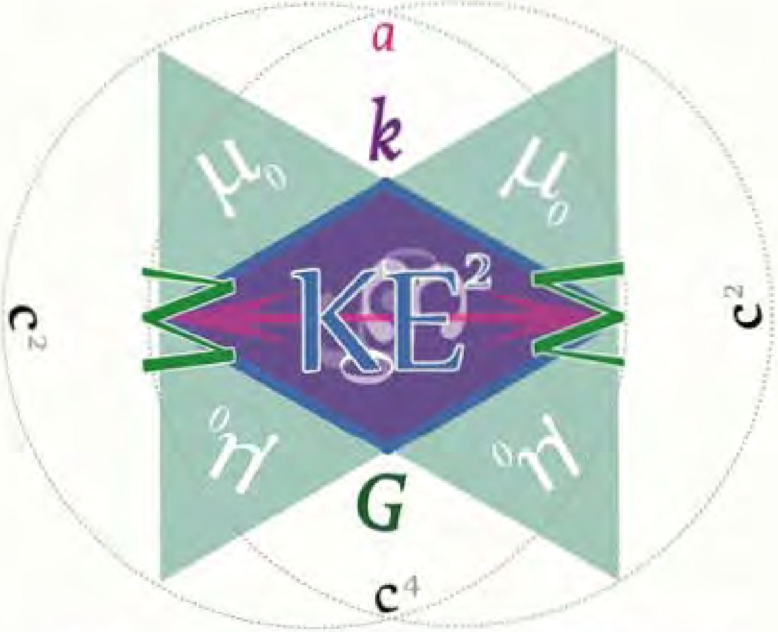

Inverse sqaured Gravity can be mathematically modeled using the same geometry explaining the similar formulations for charge interaction & gravitation

$$M_0 + KE = E$$

Standing-wave four-vector

Lorentz corrected mass-energies

Stress energy momenta tensor

$$m^2v^4 = KE^2 = p^2v^2$$

EM mass

radiant energies

momenta

Four-vectors are reflective of the total EM mass-energy momenta of superpositioned [K]EM fields where the KE is the interactive force component

Special and General relativity

General relativity was developed by Einstein in an attempt to describe the acceleration forces experienced by graviational Matter (on the foundation of special relativity) which in turn was an extension of Lorentz corrections of EM waves developed to account for the fixed speed of light and established Newtonian mechanics

SR

Special Relativity

(describes inertial frames of reference
in flat Euclidean spacetime)

mass-energy

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

geometries

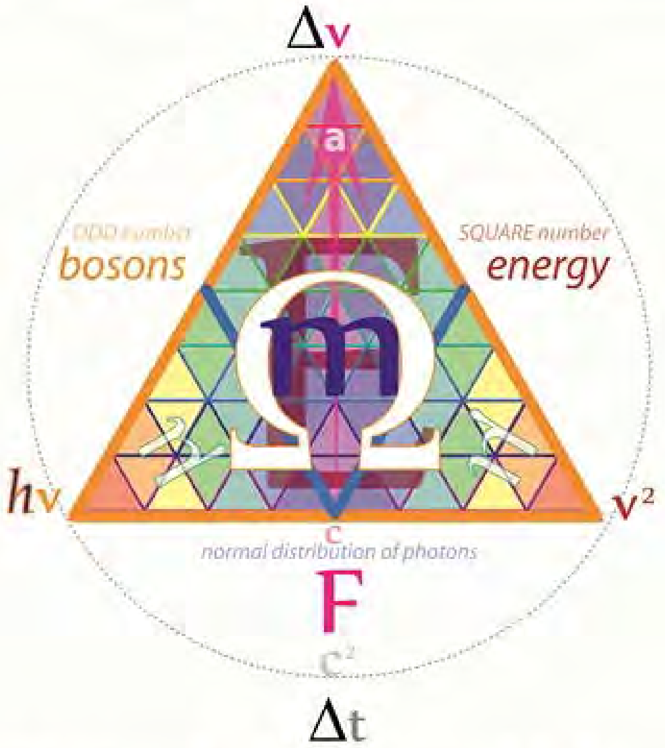

GR

General Relativity

(describes accelerating frames of reference
in curved Riemannian spacetime)

Matter

$$\overset{\text{Matter}}{\frac{4n\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

topologies

2D

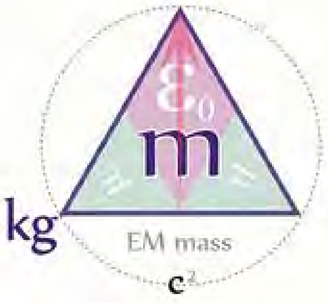

all physical theories to date have failed to explain the mechanics of gravitation at all scales of energy and distance

$$\frac{\rho}{m^2}$$

6.629432351 e-34 J

$$\overset{\text{ENERGY}}{n\pi}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

7.376238376 e-51 kg

$$\frac{\rho}{m^3}$$

3D

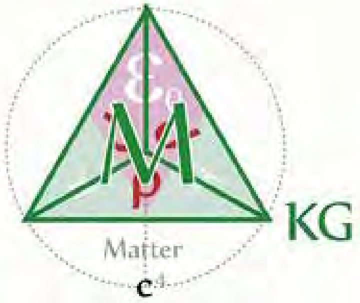

The Weak Equivalence Principle

$$m_g . a = m_i . g$$

Newton

The uniqueness of free fall states than any two test bodies must fall with the same acceleration in a given external gravitational field

$$m_g = m_i \sum_A \eta^A \frac{E^A}{c^2}$$

Einstein

The uniqueness of free fall trajectories allows one to regard spacetime as filled being with curved energy gradients

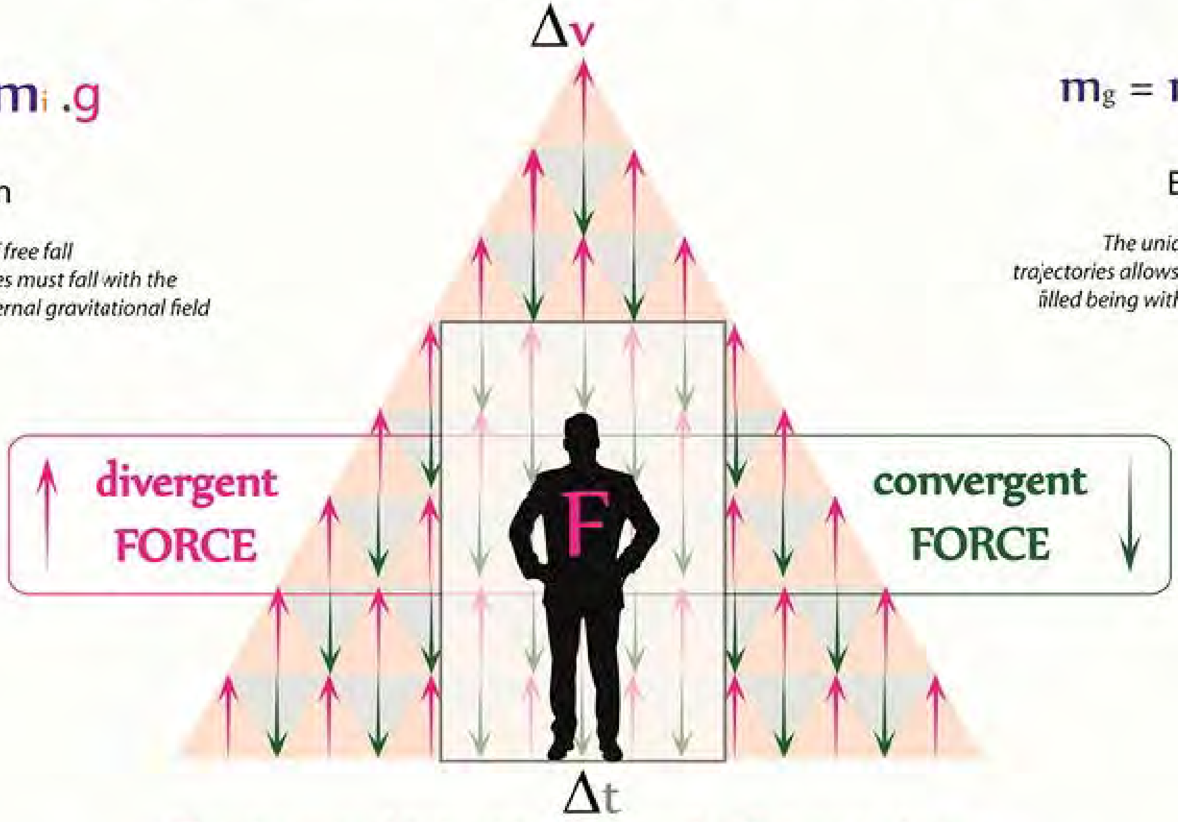

$F = ma$

$F = mg$

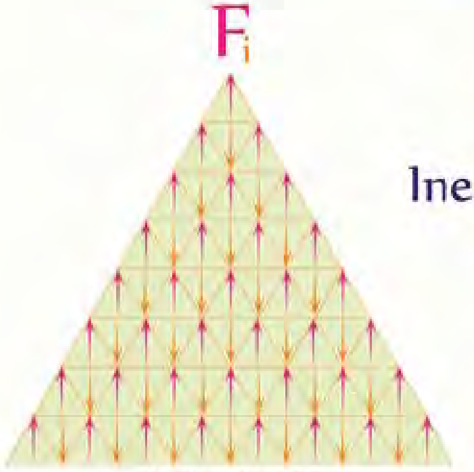

acceleration vs inertial mass

'There is no way of distinguishing between the effects on an observer of a uniform gravitational field and that of constant acceleration'

Inertial and Gravitational masses are equivalent

Of particular note is the fact that the relativistic formulation of the weak equivalence principle hints at the fact that mass-energies in all their forms must be accounted for and correctly modeled for an accurate picture of the mechanics of gravitation on all-scales to be developed

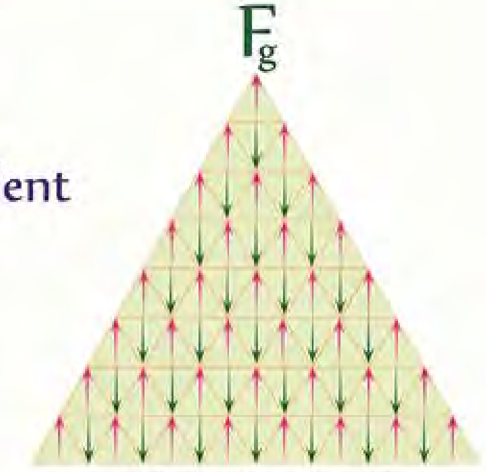

acceleration vs gravitational mass

Copyright ABRAHAM [2008] - All rights reserved

Einstein's Gravitational Constant

Newton gave a law for the behavior of objects that experienced gravitational forces:
$F = ma$, where $F = -m\nabla\phi$ for a given gravitational field ϕ

And along with Gauss & Poisson also gave a law determining how φ is generated:
$\nabla^2\phi = 4\pi G\rho$.

Einstein subsiquently re-expressed Newtonian Gravitation as a curvation of spacetime brought about by the presence of all mass-Energies

$$G_{ab} = \kappa T_{\mu\nu}$$

General Relativity

spactime curvature | Gravitational constant | stress energies

$$G_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

speed of light

Einstein viewed gravitation as the result of a curvature of spacetime created by a mass-energy-pressure gradiant

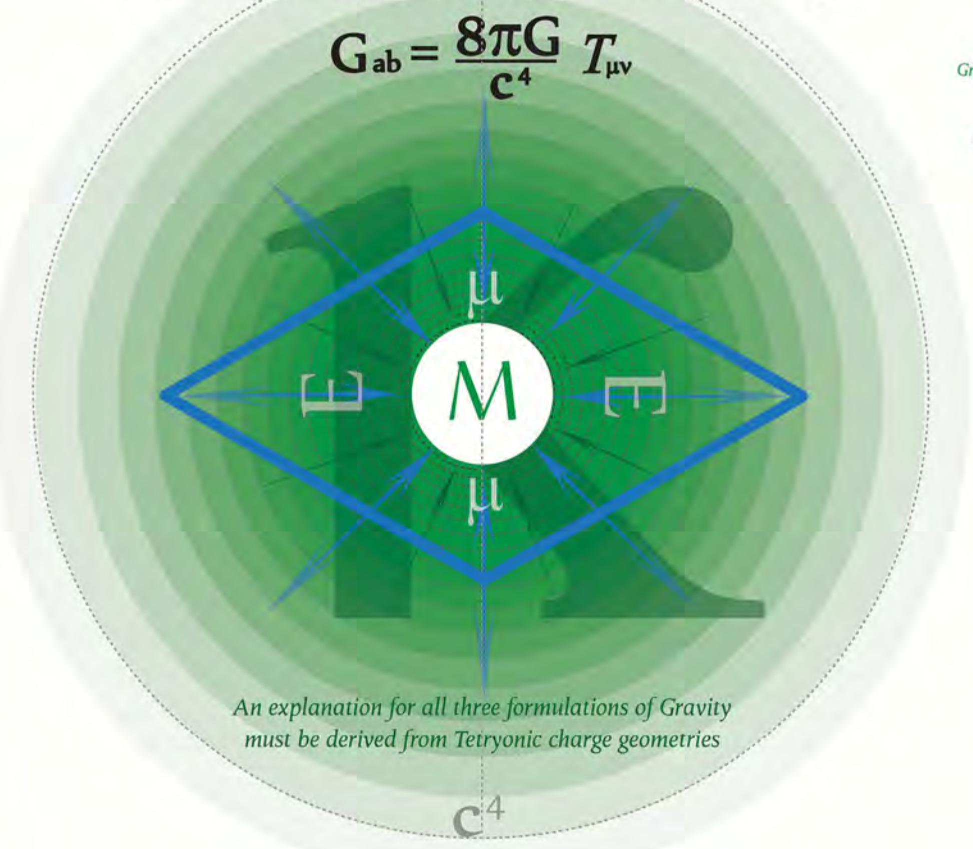

An explanation for all three formulations of Gravity must be derived from Tetryonic charge geometries

$$n\pi[mEM]$$

Tetryonic charged geometries

Gravitational Matter topologies | *Interactive mass-energy geometries*

$$T\pi\left[\frac{\rho}{c^4}\right] + n\pi\left[\frac{\rho}{c^2}\right]^2$$

3D spatial region | *2D spatial fields*

Tetryonics differentiates between Matter & interactive mass-energies via their charged geometries and topologies

EM

2D euclidean mass-energy-momenta

Rank 2 tensors

convergent & divergent forces

$$m = \frac{\rho}{c^2}$$

M

3D mass-Matter topologies

Rank 4 tensors

convergent pressure gradient

$$M = \frac{\rho}{c^4}$$

By modeling all mass-energy geometries and Matter topologies as a stress energy tensor General Relativity fails to differentiate between the interactive Forces of mass-energy momenta and gravitational Matter

mass-energy-Matter spatial interactions

No two Fermions [3D Matter topologies] can occupy the same physical space at the same time - Pauli exclusion

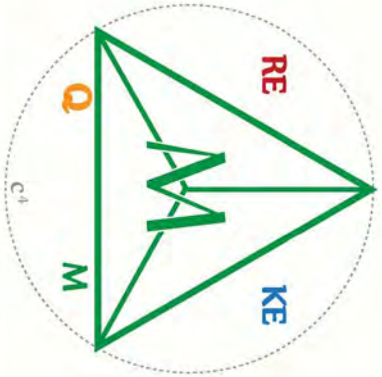

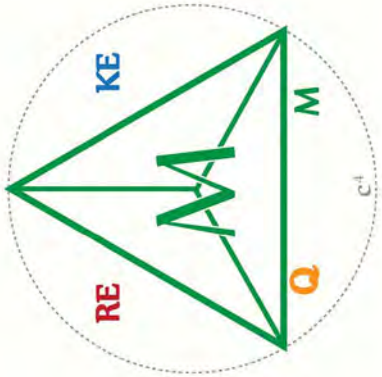

c^4

is a volumetric measure of the amount of space taken up by Matter

Fermions

Planck quanta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

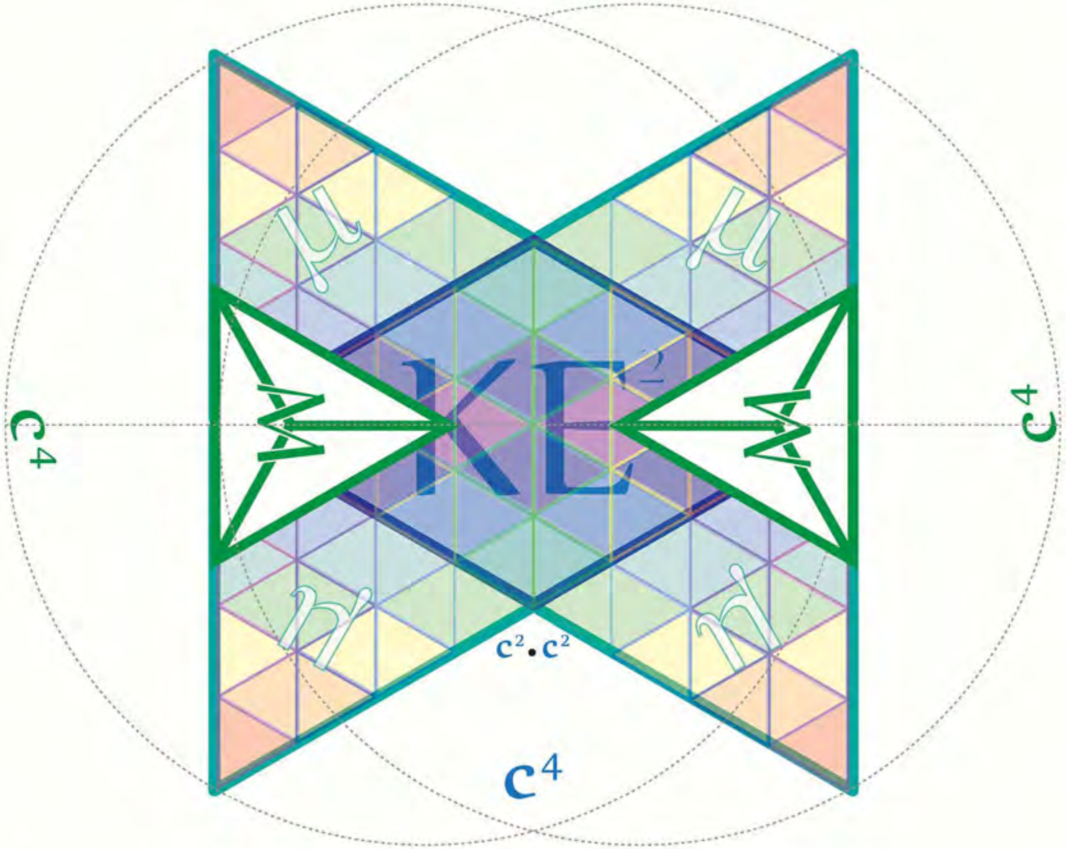
μ
μ
KE²
M
M
μ
μ
c^4
c^4
$c^2 \cdot c^2$
c^4

c^4

is also a spatial measure of two superpositioned interactive KEM fields

Force interaction

Planck quanta

$$\frac{2\pi}{c^2}\cdot\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

All separated particles interact with each other via 2D fields of Force (superpositioned [K]EM fields)

v
Ω
RE
μ
KE
p
v
c^2
μ
m
KE
v^2

v^2
KE
m
μ
v
p
KE
c^2
μ
RE
Ω
v

Copyright ABRAHAM [2008] - All rights reserved

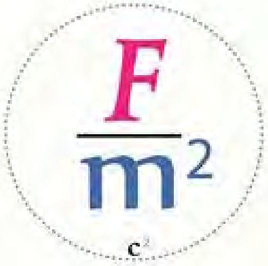

SR

Special relativity applies
Lorentz velocity corrections to
radiant mass-energy momenta

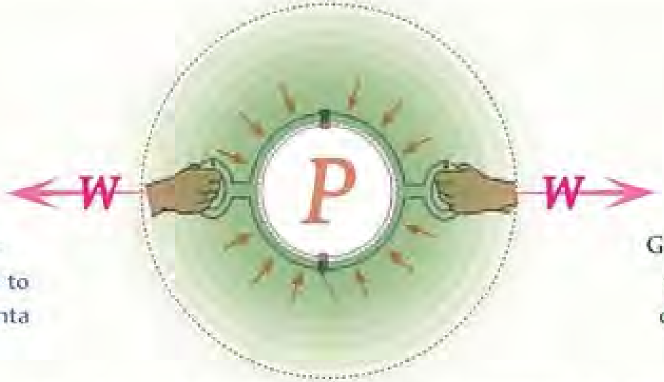

GR

General relativity applies SR
to total energy momenta
density-pressure gradients

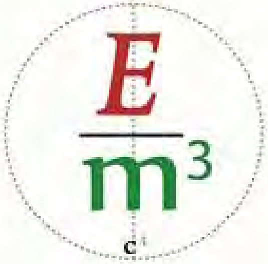

Otto von Guericke, (1602-1686), a German physicist, performed a famous experiment: the "Magdeburg Hemispheres".

When two halves of a large, hollow metal ball had all the air removed from the volume they created when joined, two teams of horses couldn't pull them apart, because the pressure of the Earth's atmosphere (external to the volume created) exerted a very large force on the ball as it sought to equalise the pressure gradient created.

Nullspace energy pressure gradients

mass-energy geometries

$$P = \frac{Force}{Area} = \frac{F}{A} \Rightarrow \frac{F.d}{A.d} \Leftarrow \frac{W}{V} = \frac{Energy}{Volume} = U$$

mass-Matter topologies

The scalar mass-energies of all Matter topologies are all contained in their charged fascia

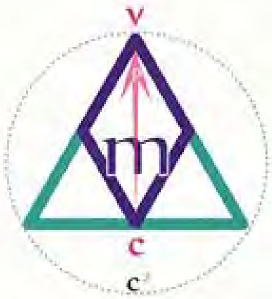

$$n\pi\left[\frac{\rho}{c^2}\right]$$

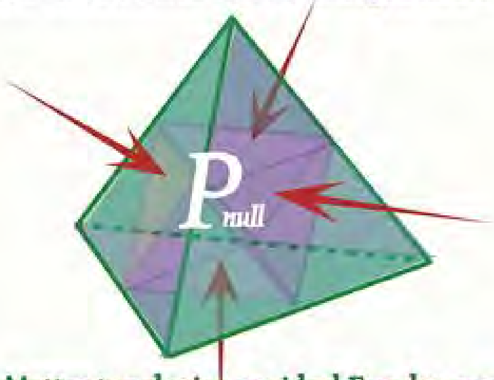

$$T\pi\left[\frac{\rho}{c^4}\right]$$

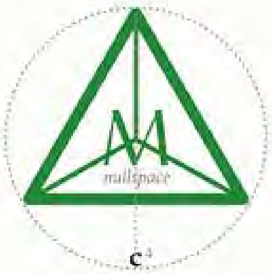

A Faraday cage is a metallic enclosure that prevents the entry or escape of an electromagnetic field (EM field)

All Matter topologies are ideal Faraday cages, quantum standing-waves of EM mass-energies that in turn create regions of nullspace

An ideal quantum Faraday cage consists of an unbroken, perfectly conducting shell. (a Matter topology of scalar mass-energies)

Copyright ABRAHAM [2008] - All rights reserved

Total stress energies & Matter

The stress energy tensor can be regarded only as a provisional means of representing all of the interactive mass-energies & gravitational Matter in any physical system.

$$T_{\mu\nu}$$

GR's Stress Energy Tensor
does NOT differentiate
between EM mass & Matter

Einstein

It is only in present circumstances that we find we have insufficient knowledge of the electromagnetic field of concentrated charges that compels us, provisionally, to leave, undetermined in present theory, the true form of this tensor...

The right hand side the EFE is a formal condensation of all things whose comprehension in the sense of a field theory is still problematic.

Not for a moment... did I doubt that this formulation was merely a makeshift in order to give the general principle of relativity a preliminary closed form expression.

For it was essentially no more than a theory of the gravitational field, which was isolated somewhat artificially from a total field of as yet unknown structure.

Abraham

mass

$$\frac{n\pi}{c^2}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

ENERGY

$$n\pi\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Matter

$$\frac{T\pi}{c^4}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Q

changing quantised angular momenta creates charge~Time

M

3D MATTER topologies displace vacuum energies creating Gravity

Vacuum energies are formed from weak radiated mass-energies

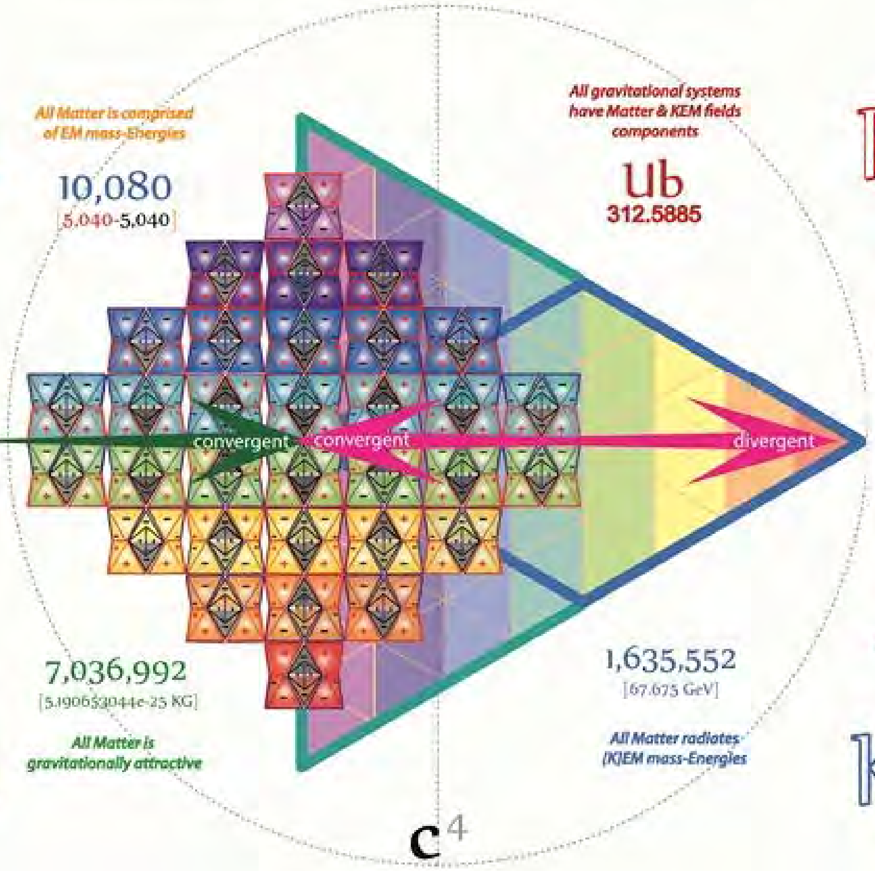
All Matter is comprised of EM mass-Energies
10,080
[5,040-5,040]
All gravitational systems have Matter & KEM fields components
Ub
312.5885
convergent
convergent
divergent
7,036,992
[5.19065304e-25 KG]
All Matter is gravitationally attractive
1,635,552
[67.675 GeV]
All Matter radiates (K)EM mass-Energies
c^4

RE

changing linear momentum creates forces of Acceleration

radiant 2D planar mass-ENERGIES create interactive forces

KE

KEM field energy momenta create fields of interactive Force

Tetryonics 69.10 - Total stress energies & Matter

Copyright ABRAHAM [2008] - All rights reserved

Einstein's Field Equation

"Spacetime tells Matter how to move; Matter tells spacetime how to curve"
John Archibald Wheeler

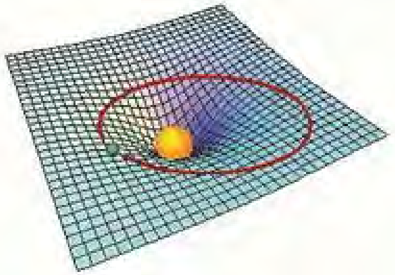

$$R_{ab} - \tfrac{1}{2} R g_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

The solutions of the Einstein field equations are metrics of spacetime

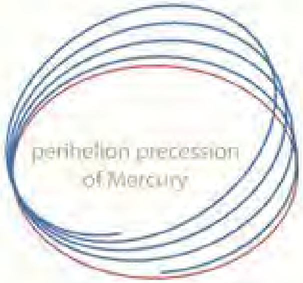

General Relativity was developed to account for the observed perihilion precession of Mercury's orbit around the SUN
[only noticable through detailed observations after Newton developed his law of universal gravitation]

Newtonian G field

4π

the scalar curvature

the cosmological constant

Newton's gravitational constant

Ricci curvature tensor

$$R_{ab} - \frac{1}{2} g_{ab} R + \Lambda g_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

the stress-energy tensor

the metric tensor

the spatial region of mass-energy-Matter being measured
as determined by the speed of Energy in vacuum

The EFE is a tensor equation relating a set of symmetric 4 x 4 tensors. Each tensor has 10 independent components

Einstein's G field

8π

The EFE (based on SR principles) reduces to Newton's law of gravitation where the radiant mass-energy fields are weak and velocities are much less than the speed of light
[highlighting the inability of GR to distingush between radiant mass-energy geometries & the standing-wave Matter topologies comprising the total energies within the stress Energy tensor]

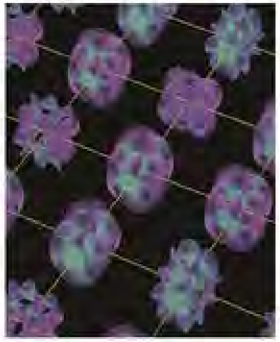

Curved Spacetime

Einstein's gravitational constant

$$G_{ab} + \Lambda g_{ab} = K\, T_{\mu\nu}$$

the cosmological constant

mass-ENERGY-Matter

Despite the successes of General Relativity, the problems of quantum gravity and spacetime singularities remain open problems in its formulation of curved spacetime geometries

Copyright ABRAHAM [2008] - All rights reserved

ZPF

All Energies are quantised

Quantum theory

m

Gravitation is Universal

Newtonian Mechanics

M

Gravity results from curved spacetime

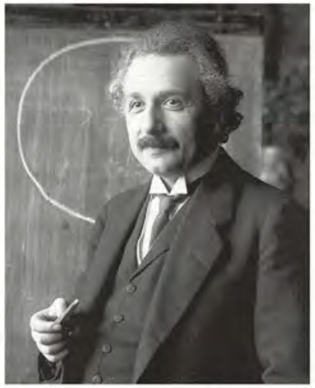

Relativity theory

VE

Quantum gravity is the effort in theoretical physics to create a theory that includes both general relativity and the standard model of particle physics. Currently, these two theories describe different scales of nature and attempts to explore the scale where they overlap yield results that don't quite make sense, like the force of gravity (or curvature of spacetime) becoming infinite.

Quantum gravity

Gravity as it is currently understood cannot be reconciled with the laws of quantum mechanics.

Euclidean quantum gravity
Scalar theories
Modified Newtonian dynamics (MOND)
Einstein-Cartan Bimetric theories
Tensor-vector-scalar
DGP model
Geometrodynamics
Loop Quantum Gravity
Lovelock theory
Kaluza-Klein theory
Induced gravity
Canonical Quantum Gravity
Jackiw-Teitelboim gravity
Liouville gravity
String theory
Composite gravity
Causal Set theory
Brans-Dicke
Massive gravity
Teleparallelism
Scalar-tensor-vector
Causal dynamical triangulation
Conformal gravity
Nonsymmetric gravitation
2+1D topological gravity
Gauss-Bonnet gravity
Supergravity
Gauge theory gravity
f(R) gravity
Unified theory

hv 1899 *Planck's law*

Q 1785 *Couolomb's law*

G 1686 *Newton's law of Gravitation*

SR 1905 *Special relativity*

GR 1915 *General relativity*

Without the aid of Tetryonic's charged geometries
Gravitational & EM fields point to mathematical singularities

Inverse squared
Electric fields produce
INTERACTIVE
fields of force

Inverse squared
Magnetic fields produce
PURTURBATIVE
fields of force

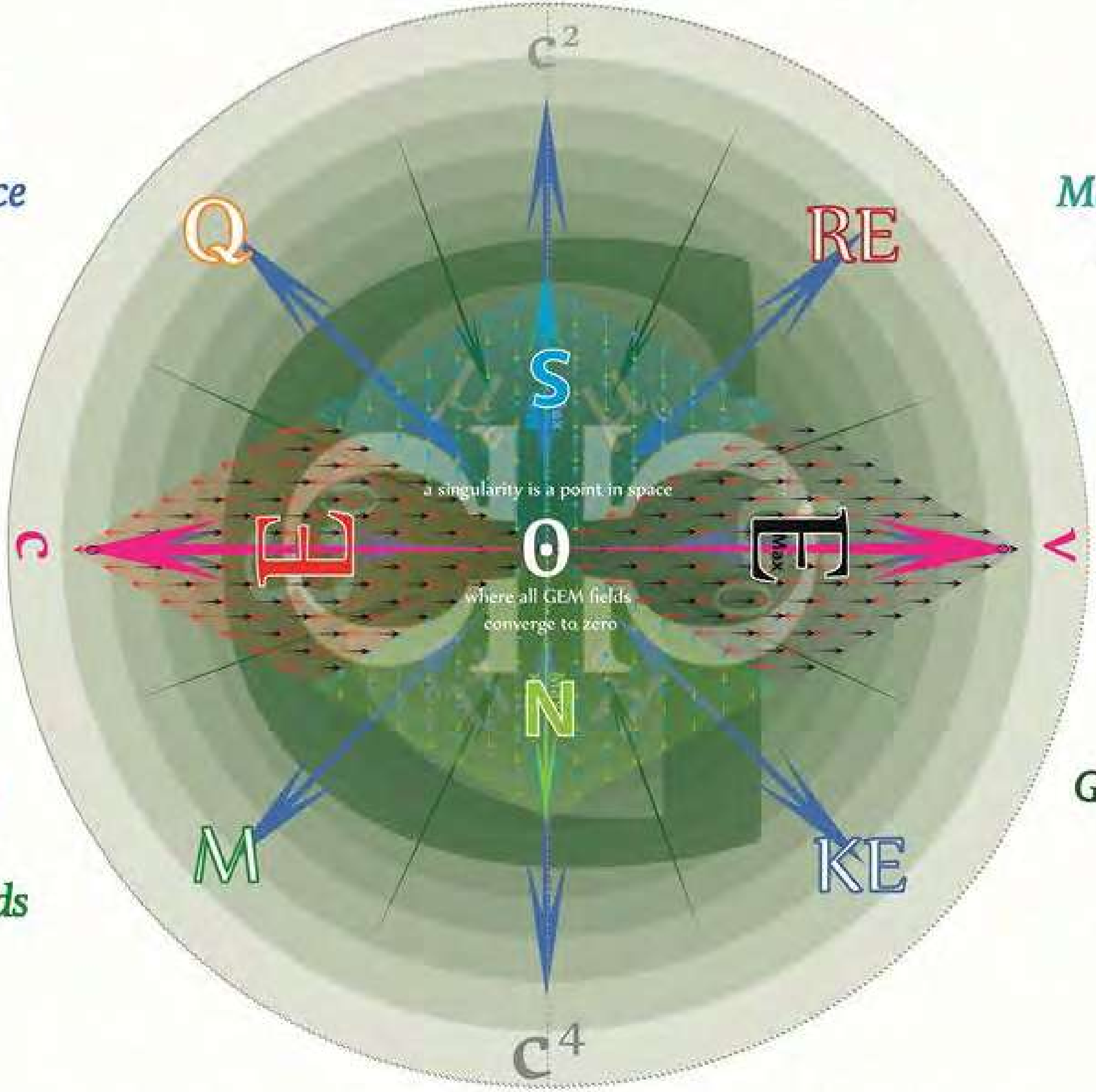

c^2
Q
RE
S
a singularity is a point in space
c
E
0
Max
E
v
where all GEM fields
converge to zero
N
M
KE
c^4

There are NO
Black Holes

There are NO
point Particles

Matter produces
CONVERGENT
inverse squared fields
of Gravity

Gravitation is the nett
convergent force
produced by all
G-E-M fields

GravitoElectroMagnetic fields

Copyright ABRAHAM [2008] - All rights reserved

Matter topologies and gravity

the positive spatial displacements created by tetrahedral standing-wave mass-energies displace vacuum energies found in those regions to create a energy density pressure gradient

A convergent Gravity field is the result of the higher density-pressure vacuum energies in any spatial region seeking to equalise the lower pressure regions of nullspace within Matter topologies

Matter has a
Tetrahedral mass-energy
quantum topology
[non-Spherical]

M

The mass-energies
of Matter topologies
are stored in the integral
surface area geometries
of their charged fascia

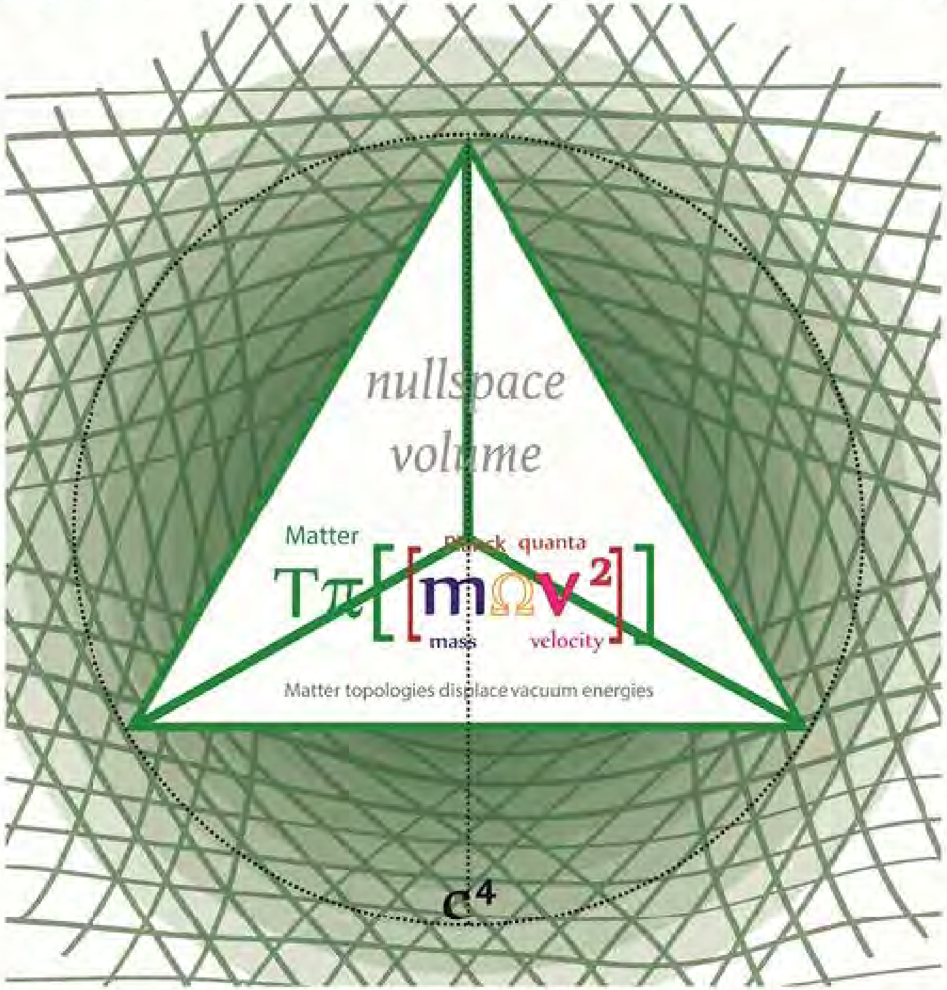

Gravity itself is
the result of vacuum
energy pressure gradients
not Coloumbic fields

G

Gravitation is the
result of the convergent
gravity AND interactive
[k]EM fields that all quantum
Matter topologies create

The resultant density-pressure gradients created by the displacement of vacuum energies by Matter topologies can be modelled using equilateral EM field geometries and a reduced Coloumb constant

Copyright ABRAHAM [2008] - All rights reserved

Spatial displacement of GEM energies by Matter topologies

The positive 4nπ displacement of vacuum eneriges by Matter topologies creates

Spatial energy density-pressure

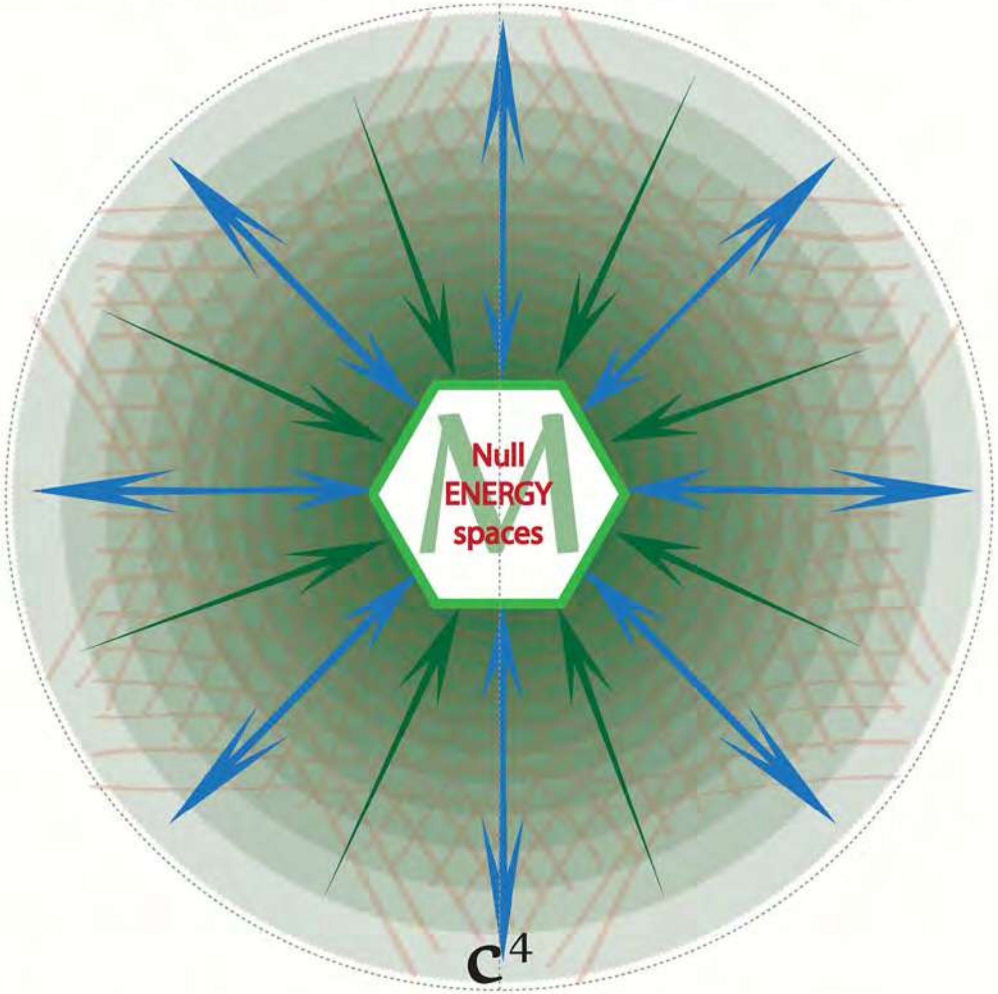

Convergent Gravitational fields

$$F = -G \frac{M_1 m_2}{r^2}$$

Planets orbit around the SUN

Newtonian fields models the resultant attraction as gravity

EM fields accelerate charged Matter

$$F = k \frac{Q_1 Q_2}{r^2}$$

$$G_{ab} = 8\pi G T_{\mu\nu}$$

All EM mass & Matter create stress energies

Einstein models the GEM energy densities as curved spacetime

Orbital precessions must be accounted for

Newton	Einstein
4π vs.	8π
G	G+SR

Newtonian Gravity

model the observed, long range attraction between two bodies of Matter
[exclusive of the close range electromagnetic interactions of each body]

A gravity field is produced by the charged topology of mass-Matter as it displaces vacuum energies and creates intrisic nullspace energy differentials

The strength of a Gravity field produced by Matter is proportional to the total EM mass-energy geometry/topology of any body and its convergent force can be modelled with equilateral Tetryonic geometries

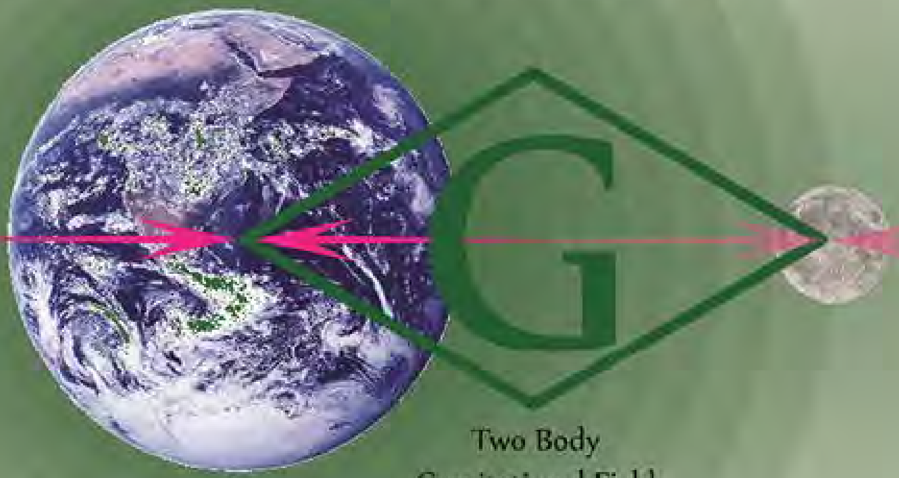

Single Body
Gravitational Field

$$F = -G\frac{M_1}{r^2}$$

Single Body
Gravitational Field

$$F = -G\frac{m_2}{r^2}$$

Two Body
Gravitational Field

$$F = -G\frac{M_1 m_2}{r^2}$$

However, unless it is at absolute zero all Matter topologies produce interactive EM fields that contribute to gravitational forces as evidenced with the perihelion of Mercury where later observations revealed inaccuracies in Newton's formulation for the motion of material bodies subject to convergent G forces & interactive EM fields

Copyright ABRAHAM [2008] - All rights reserved

Gravitational energies as a geometric mean

the geometric mean of two numbers is the square root of their product

EM standing waves [Matter] can be created and destroyed

EM mass-energies can be neither created or destroyed

Sir Issac Newton was the first to realise that gravity could be modelled as an inverse square Force

$$F = -G \frac{M_1 m_2}{r^2}$$

M

E

The Force of Gravity
can be accurately modelled

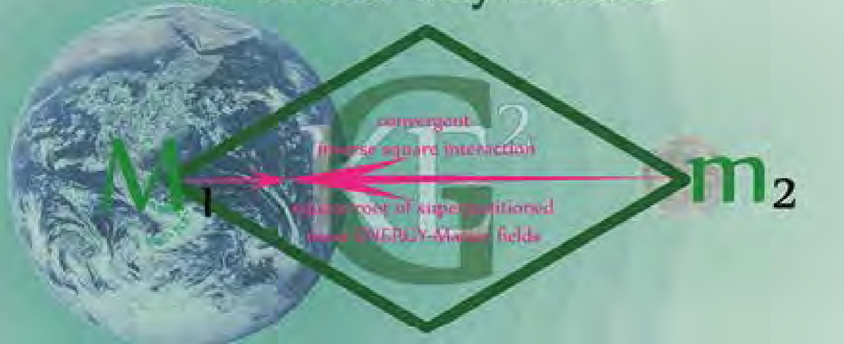
M_1
convergent
inverse square interaction
square root of superpositioned
mass ENERGY-Matter fields
G^2
m_2

Gravitational field of mass-Matter only

$$F_G = 4\pi G \rho$$

[ignoring EM interactions]

Gravitational energy of Matter in motion

$$G_{ab} = 8\pi G T_{\mu\nu}$$

[including KEM field interactions]

as the convergent geometric MEAN of interactive superpositioned E-fields that represent the total mass-energies of the gravitational Matter

$$F_G = \frac{M_1 . m_2}{r^2}$$

$$F_G = \sqrt{M_1 . m_2}$$

Gravity is the result of spatial energy density differentials created by the internal null-space volumes of Matter.
Energy (in all its forms) will seek to fill areas of lower pressure in search of universal equilbrium

Newton modelled the attractive forces between objects of Matter

Matter topologies are non-conservative and as they are destroyed the low pressure nullspace gradient [convergent gravitational Force] they created is replaced with conservative forms of radiative EM energy [interactive vacuum energies-photons]

GR models the energy density created by mass-Matter

Gravitational Tidal Forces

The tidal force is a secondary effect of the force of gravity resulting from the geometry of massive bodies attracting each other.

Its geometric effects allow acceleration fields to be distingushed from pure graviation fields

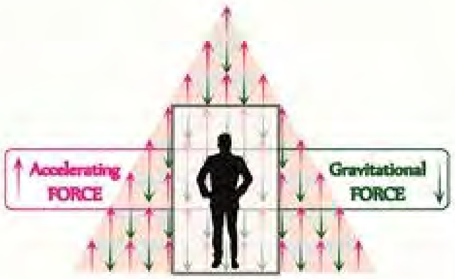

Over a small area on the surface of a Gravitating body, gravitational attraction is often modelled as a uniform field acting downwards across the whole area.

The Gravity of a material body of mass-Matter has historically been defined as a strictly attractive force between bodies

$$F = -G\frac{M_1 m_2}{r^2}$$

Tetryonics reveals Gravitation to be the nett result of 3 distinct interactions:

Convergent G-fields

Interactive E-fields and

Perturbative M-fields

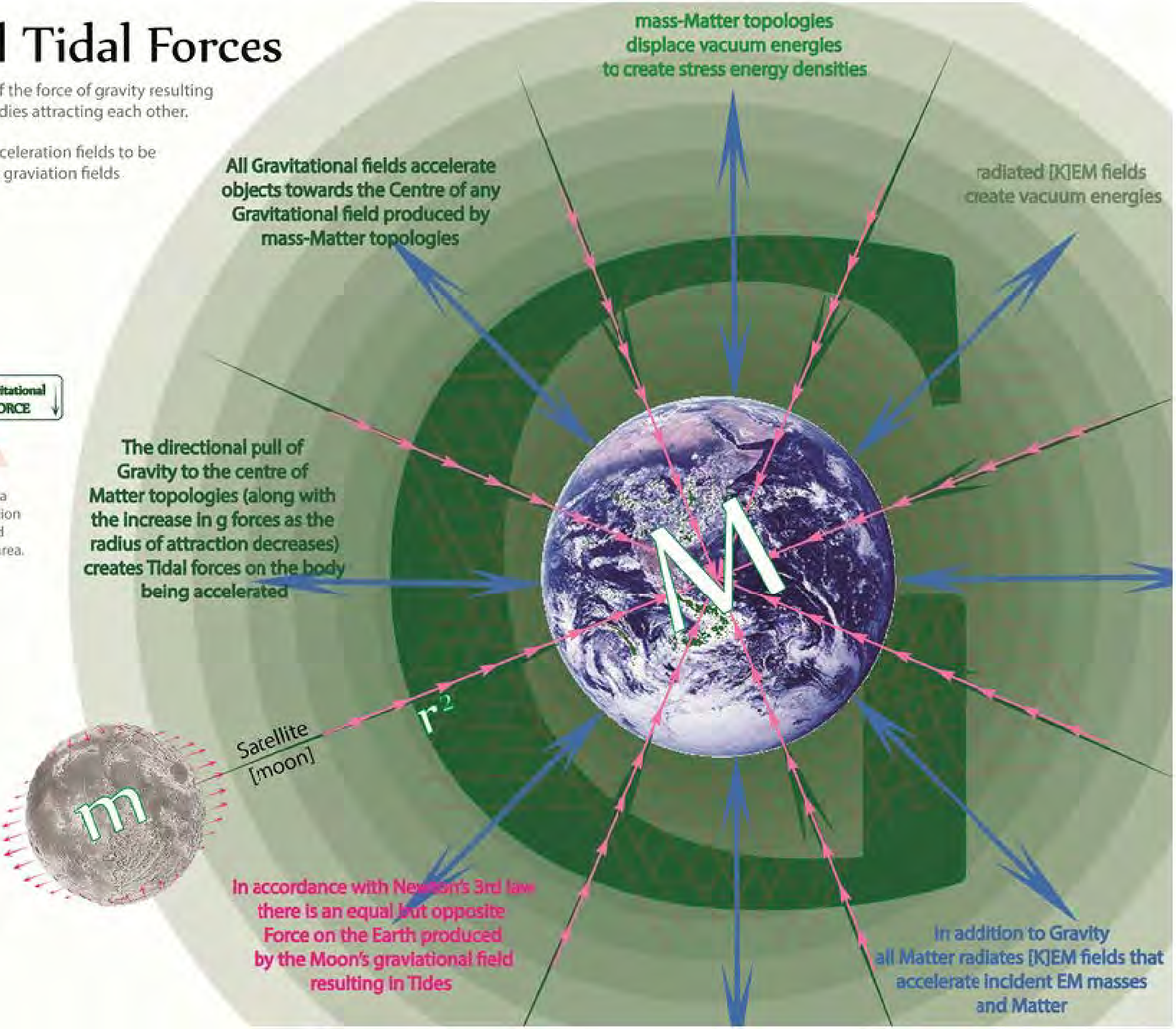

mass-Matter topologies displace vacuum energies to create stress energy densities

All Gravitational fields accelerate objects towards the Centre of any Gravitational field produced by mass-Matter topologies

radiated [K]EM fields create vacuum energies

The directional pull of Gravity to the centre of Matter topologies (along with the increase in g forces as the radius of attraction decreases) creates Tidal forces on the body being accelerated

In accordance with Newton's 3rd law there is an equal but opposite Force on the Earth produced by the Moon's graviational field resulting in Tides

In addition to Gravity all Matter radiates [K]EM fields that accelerate incident EM masses and Matter

Tetryonic Gravitation

is the result of the combined interactions
of radiant EM geometries & Matter topologies
in any spatial region of measurement

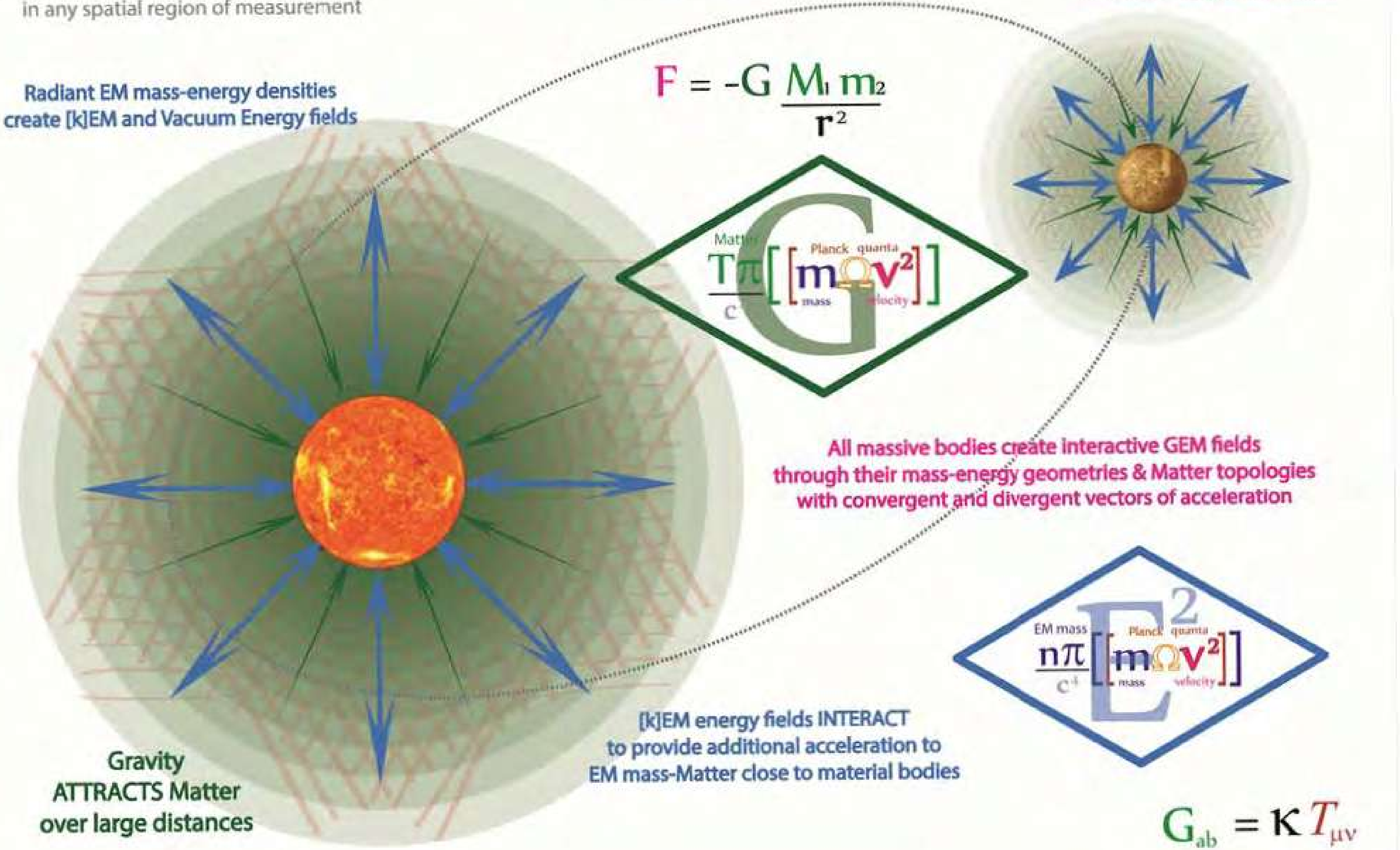

[k]EM fields & Vacuum energy gradients
create energy momenta pressure differentials
that accelerate Matter

3D mass-Matter topologies
create nullspace
vacuum energy differentials

Radiant EM mass-energy densities
create [k]EM and Vacuum Energy fields

$$F = -G\frac{M_1 m_2}{r^2}$$

$$G \quad \frac{T\pi}{c}\left[\left[m\Omega v^2\right]\right]$$

All massive bodies create interactive GEM fields
through their mass-energy geometries & Matter topologies
with convergent and divergent vectors of acceleration

$$E^2 \quad \frac{n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Gravity
ATTRACTS Matter
over large distances

[k]EM energy fields INTERACT
to provide additional acceleration to
EM mass-Matter close to material bodies

$$G_{ab} = \kappa T_{\mu\nu}$$

Copyright ABRAHAM [2008] - All rights reserved

EM masses & Gravitational Matter

$$\frac{\rho}{m^2}$$

The core problem preventing the unification of Classical gravitation with QM & Relativity theory is the imprecise formulation, definition and differentiation between mass & Matter

$$\frac{\rho}{m^3}$$

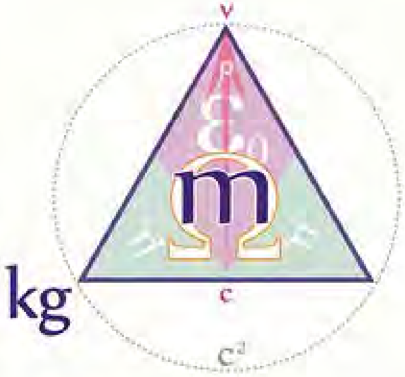

$$F = ma$$

mass is a conserved quantity
Matter is not a conserved quantity

$$F = -G\,\frac{M_1 m_2}{r^2}$$

inertial mass
active gravitational mass
passive gravitational mass

$$G_{ab} = \kappa\, T_{\mu\nu}$$

Even GR offers no clear distinction

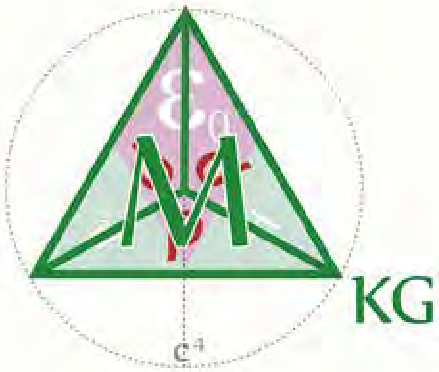

$$E = mc^2$$

radiant mass-energy geometry

$$E = Mc^4$$

standing-wave Matter topology

General relativity has been heralded as a highly successful model of gravitation and cosmology, which has so far passed every unambiguous observational and experimental test.

Even so, there are strong indications the theory is incomplete

Without a clear, concise understanding and accurate models of the relationship between mass & Matter [and their interactions] a Unified Field theory is impossible

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\overset{\text{Energy}}{\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]}$$

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\overset{\text{Energy}}{\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]}$$

mass is a geometric measure of a field's planar energy momenta content per unit Time

$$mc^2 = E = Mc^4$$

Matter is a topological measure of an object's planar mass-energy content per unit of Time

2D masses

bosons photons

Note: in both cases the amount of Energy described is identical
It is the changing spatial co-ordinate systems used to measure Energy in all it forms
that allows for the differentiation betweeen mass-energy geometries & Matter-energy topologies

3D Matter

Fermions, Baryons

Copyright ABRAHAM [2008] - All rights reserved

mass-ENERGY-Matter equivalence

Einstein's mass-energy equivalence can be further refined and differentiated through charge geometries
[in turn creating a symetric mass-ENERGY-Matter equivalence formulation]

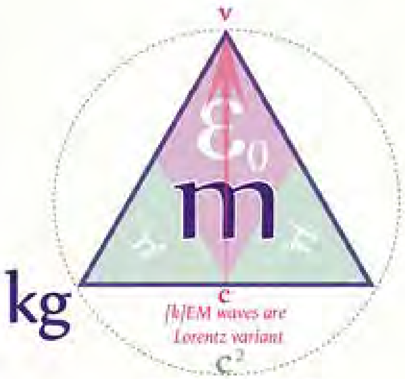
v
ε0
m
μ0
μ0
c
[k]EM waves are Lorentz variant
c²
kg

mass

standing-wave

topologies

The mass-energies of EM waveforms is contained within its charged geometry

$$E = mc^2$$

EM mass-energy is a property of Matter

$$m = \frac{E}{c^2}$$

radiant EM mass geometry

$$\frac{m}{c^2} = M$$

standing-wave Matter topology

$$\frac{E}{c^4} = M$$

radiant

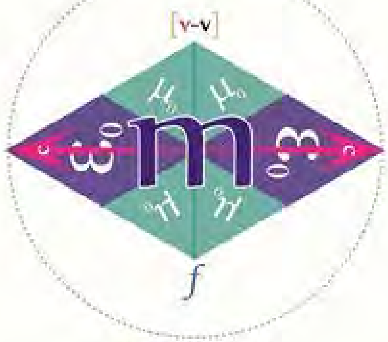
[v-v]
μ0
μ0
ε0
m
ε0
c
c
μ0
μ0
f

geometries

Matter is NOT a property of EM mass-energy

$$E = Mc^4$$

The mass-energies of Matter is contained entirely within its charged fascia topology

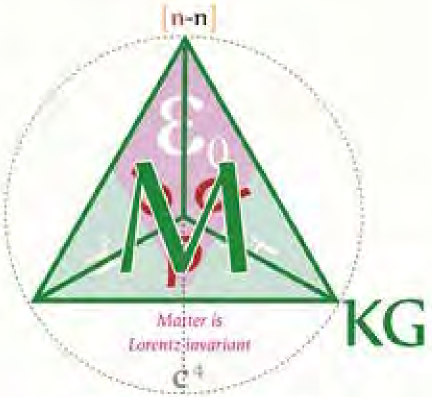
[n-n]
ε0
M
Matter is Lorentz invariant
c⁴
KG

Matter

Tetryonic unified GEM fields

Equilateral Tetryonic geometry/topology provides a unified quantum field equation to model all Gravito-Electro-Magnetic mass-ENERGY-Matter interactions

Bosons

$$\frac{\mathrm{ODD}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

All Matter is comprised of CHARGED EM mass-Energies

Q

RE

The total energy of any system is comprised of Matter & KEM fields

ENERGY

$$n\pi\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

mass & Matter are distinct EM properties

The stress-Energy tensor includes all forms of energy

ENERGY

$$\underset{\text{Matter}}{T\pi}\left[\underset{\text{mass}}{m}\Omega v^2\right]$$

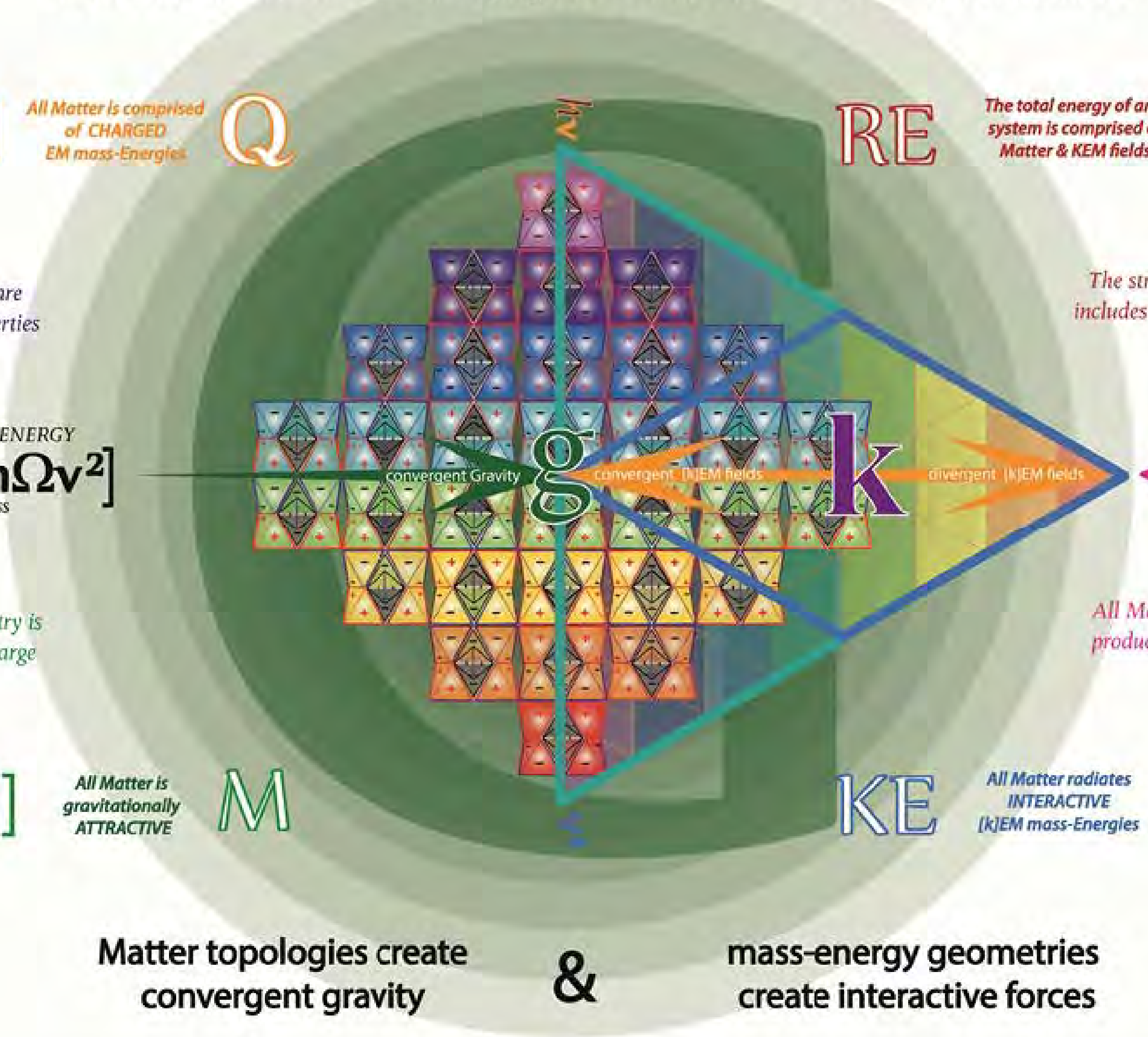

Equilateral energy momenta are geometrically related to the square of velocity

3D Matter geometry is determined by charge

All Matter in motion produces a [k]EM field

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

All Matter is gravitationally ATTRACTIVE

M

KE

All Matter radiates INTERACTIVE [k]EM mass-Energies

EM waves

$$\frac{\mathrm{EVEN}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Matter topologies create convergent gravity & mass-energy geometries create interactive forces

Quantum Gravity

At present, one of the deepest problems of modern theoretical physics is the unification of General Relativity, which describes gravitation, and applies to large-scale structures (stars, planets, galaxies), with Quantum Mechanics, which describes the other three fundamental forces acting on the atomic scale

M

Newtonian Gravitation models observed motions of large-scale MATTER

$T_{\mu\nu}$

Einstein's GR models interactions of all mass-energy-Matter

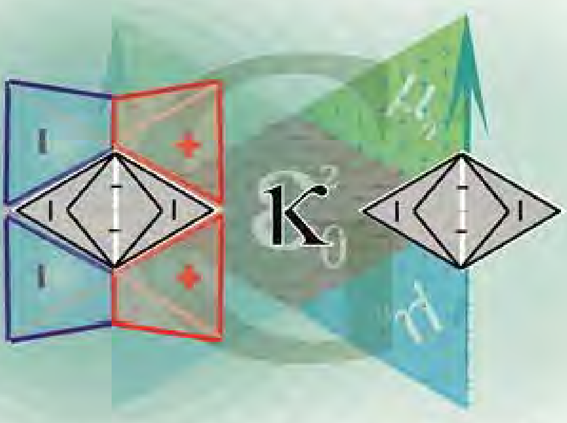

G

The large scale Cosmos is ruled by GRAVITY

k

The Quantum world is ruled by ELECTROMAGNETISM

There are major points of tension between Quantum mechanics and General relativity.

Firstly, classical general relativity breaks down at singularities, and Quantum mechanics becomes inconsistent with general relativity in the neighborhood of singularities

Secondly, it is not clear how to determine the gravitational field of a particle, since under the Heisenberg uncertainty principle of Quantum mechanics its location and velocity at any instant of Time cannot be known with certainty.

Thirdly, there is the Problem of Time in Quantum gravity. Time has a different meanings in Quantum mechanics and General relativity and hence there are subtle issues to resolve when trying to formulate a theory which combines the two.

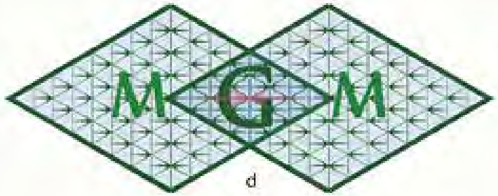

Gravitational attraction is a convergent Force between Matter

Newton

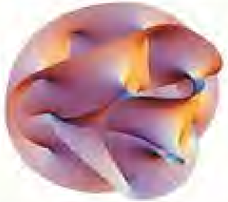

Gravitation results from energy density-pressure gradients

Einstein

Attempts to generalize ordinary quantum field theories, used in elementary particle physics to describe the fundamental interaction of gravity have to date led to serious problems.

Gravitons

Gravitons are not required in Tetryonic theory
as gravitational attraction is a convergent force created
by the EM energy differential produced by
3D Matter topologies in a vacuum energy field

All standing-wave Matter geometries create Gravity

G

CONVERGENT
Gravitation

Gravitons are the hypothetical carriers of the attractive Gravitational FORCE

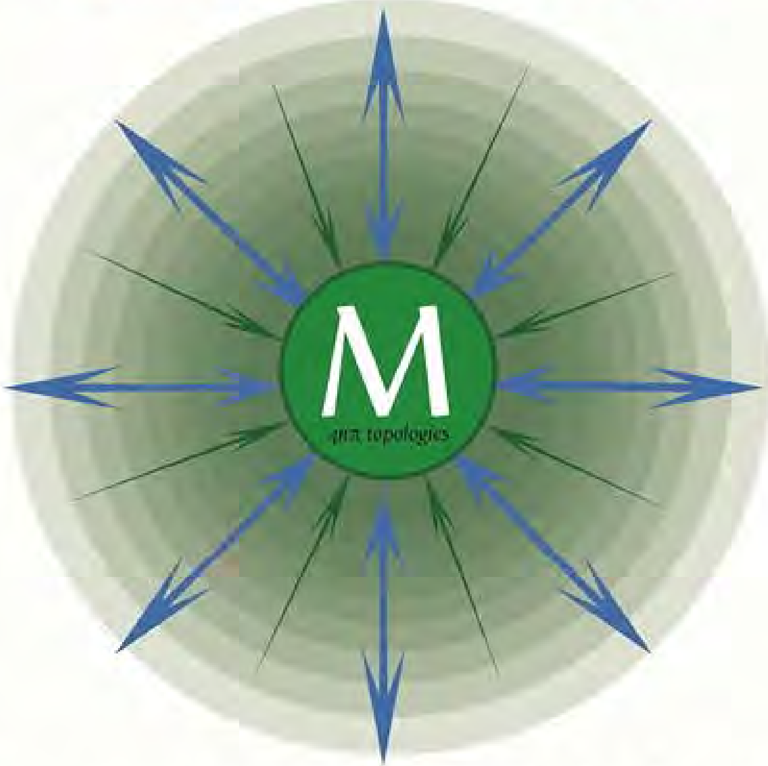

Radiant EM fields are produced by all Matter in motion

EM

DIVERGENT
ElectroMagnetism

Photons are the carriers of the interactive ElectroMagnetic FORCE

Gravitons were proposed as the gravitational equivalent of the Photon in ElectroMagnetic theory but have never been detected

$4\pi G$

$8\pi G$

They were postulated because of the great success of quantum field theory (in particular, the Standard Model) at modeling the behavior of all other known forces of nature as being mediated by elementary particles: electromagnetism by the photon, the strong interaction by the gluons, and the weak interaction by the W and Z bosons.

The hypothesis was that the gravitational interaction is likewise mediated by a – yet undiscovered – elementary particle, dubbed the graviton.
In the classical limit the theory would allow Newton's law of gravitation in the weak-field limit to conform with Einstein's field equations of General relativity

Laplacian field interactions

energy density

electric permittivity

impedance of a medium

$$Z = \frac{\rho}{c^2} = \varepsilon_0 \mu_0$$

field geometry

magnetic permeability

energy density

Einstein's constant

$$G_P = \kappa \frac{\rho}{c^4}$$

energy density pressure gradient

spatial volume

The planar two-dimensional Laplacian is given by

$$\Delta = \nabla^2 = \nabla \;\; \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

This fundamental differential operator of Applied Mathematics was introduced by Laplace in his 1782 study of the force of gravitational attraction exerted by spheroids and was named after him by Maxwell in his 1873 treatise on electromagnetism.

The net divergence of Positive charge fields is due to the anti-symmetry of their component Planck quanta

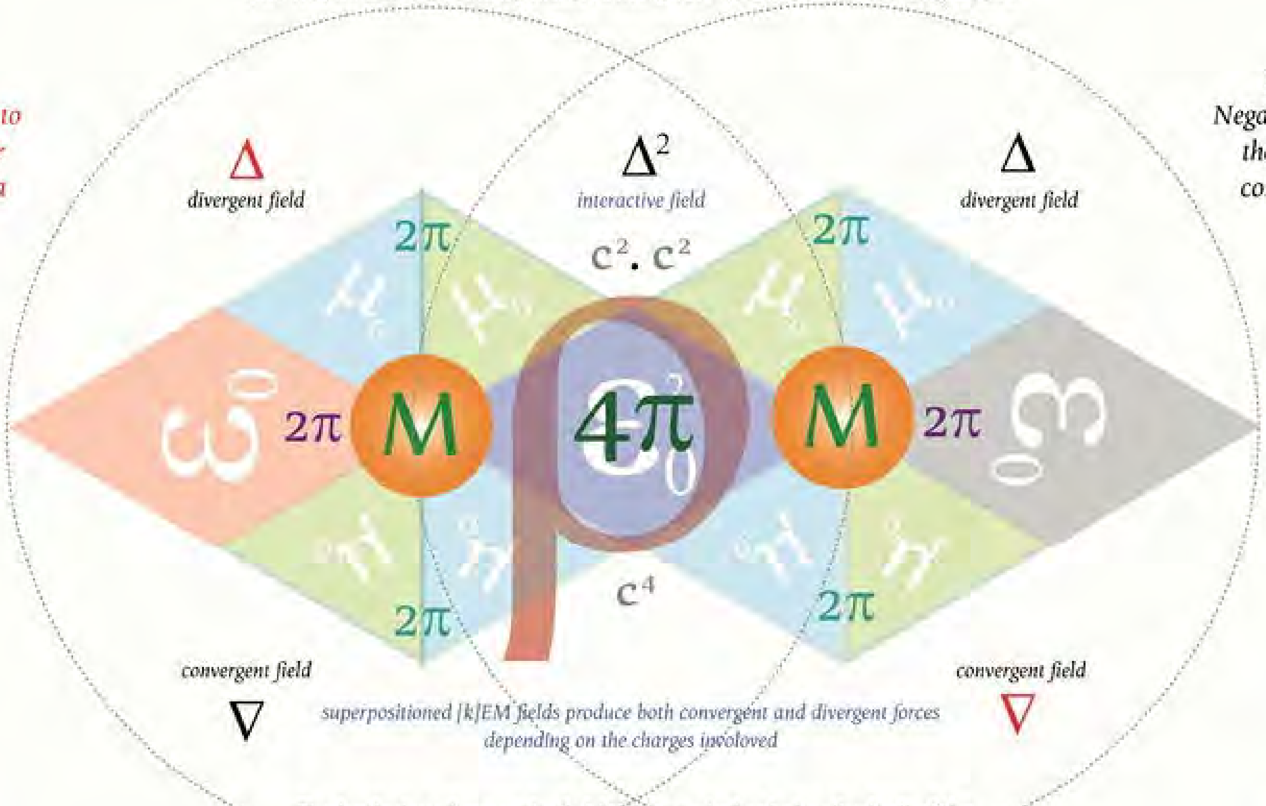
Δ
divergent field
Δ²
interactive field
Δ
divergent field
2π
c². c²
2π
μ₀
μ₀
μ₀
μ₀
ε₀
2π
M
4π
ε₀
M
2π
ε₀
ε₀
ε₀
ε₀
ε₀
2π
c⁴
2π
convergent field
∇
convergent field
∇

superpositioned [k]EM fields produce both convergent and divergent forces depending on the charges involoved

The net divergence of Negative charge fields is due to the anti-symmetry of their component Planck quanta

$$E = \frac{\rho}{c^2 \mu_0}$$

$$B = \frac{\rho}{c^2 \varepsilon_0}$$

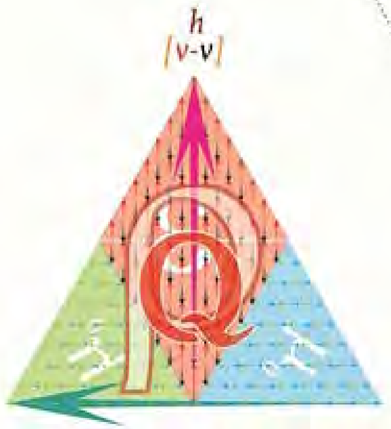

The Laplacian of any scalar field is interpreted as being the limit of the net outward flux per unit enclosed area of this field through a closed contour surrounding the point of evaluation

$$\Delta\phi = \lim_{A \to 0} \frac{1}{A} \oint_{\partial D} \frac{\partial \phi}{\partial \nu}\, dl, \quad A = \text{area}(D)$$

co-ordinate-free representation of the Laplacian

Newton's Gravitational Constant
has been shown to be a greatly
reduced version of Coulomb's
Electric interaction Constant

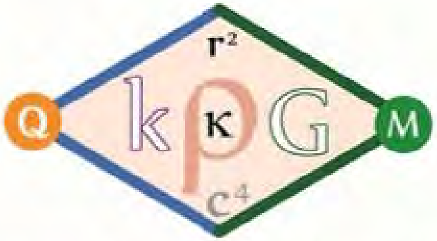

r^2
Q
k
ρ
κ
G
M
c^4

Newtonian Gravitation models
the observed convergent force of
GEM interactions that results in
the motion of large-scale Matter

Any compete Unified field equation for
Gravitation must take into account both
the Nullspace gravitation resulting from
Matter topologies as well as the Coulombic
2D interaction fields that result from charge
and the motion of Matter through the aether

Combined Interactive & Gravitational Forces

mass-energy-Matter

$$G_P = \kappa \frac{\rho}{c^4}$$

spatial volume

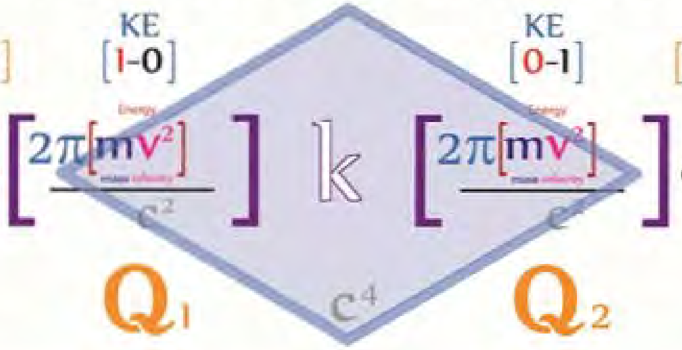

[24-12]
KE
[1-0]
KE
[0-1]
[0-12]
Energy
mass velocity
Energy
mass velocity
c^4
Q_1
Q_2

Coulomb's Law $$P+ \left[\frac{2\pi[mv^2]}{c^2} \right] k \left[\frac{2\pi[mv^2]}{c^2} \right] e-$$

Coulombic
[k]EM fields

Divergent [Interaction]
radiative EM Fields

[k]EM interactive superpositioned EM fields of Force [k]EM

$$P+ \left[\frac{T\pi}{c^4}[m] + \left[n\pi \frac{[mv^2]}{c^2} \; \kappa \; n\pi \frac{[mv^2]}{c^2} \right] + \frac{T\pi}{c^4}[m] \right] e-$$

Energy, mass, Lorentz, correction, *Einstein's constant*, mass, Lorentz, correction, Energy, mass

Matter *convergent gravitational field of Force* *Matter*

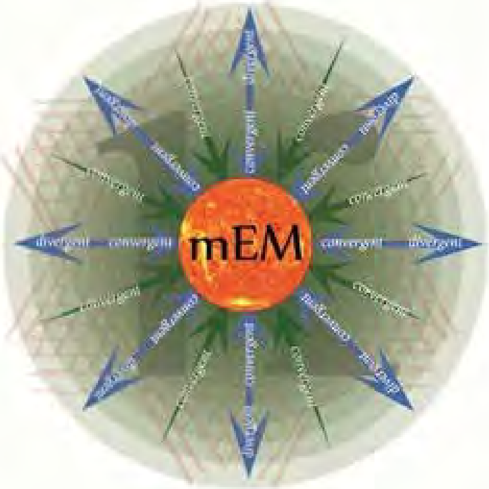

mEM
divergent
convergent

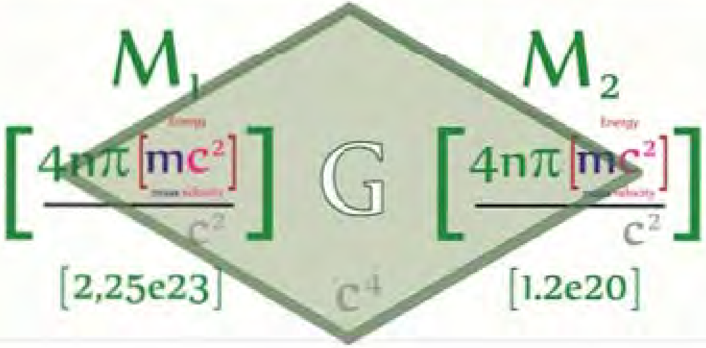

M_1
M_2
Energy
mass velocity
Energy
mass velocity
c^4
[2,25e23]
[1.2e20]

Newton's Law $$P+ \left[\frac{4n\pi[mc^2]}{c^2} \right] G \left[\frac{4n\pi[mc^2]}{c^2} \right] e-$$

Gravitational
Matter

Convergent [Attractive]
Gravity Field

The observed net convergent force between
objects of mass-Matter are the result of
convergent G & interactive EM fields

Copyright ABRAHAM [2008] - All rights reserved

vacuum impedance $Z = \frac{1}{c^2} = \varepsilon_0 \mu_0$ electric permittivity, magnetic permeability

mass-energy density $G_P = \kappa \frac{\rho}{c^4}$ spatial region

The energy density pressure gradient produced by the positive dispacement of vacuum energies by 3D standing-wave mass-Matter topologies is a strictly convergent force

$P = \frac{kT}{V}$

CHARGED mass-energy geometries create Matter topologies

mass-Matter-Vacuum energy displacements contributes to STRESS energies

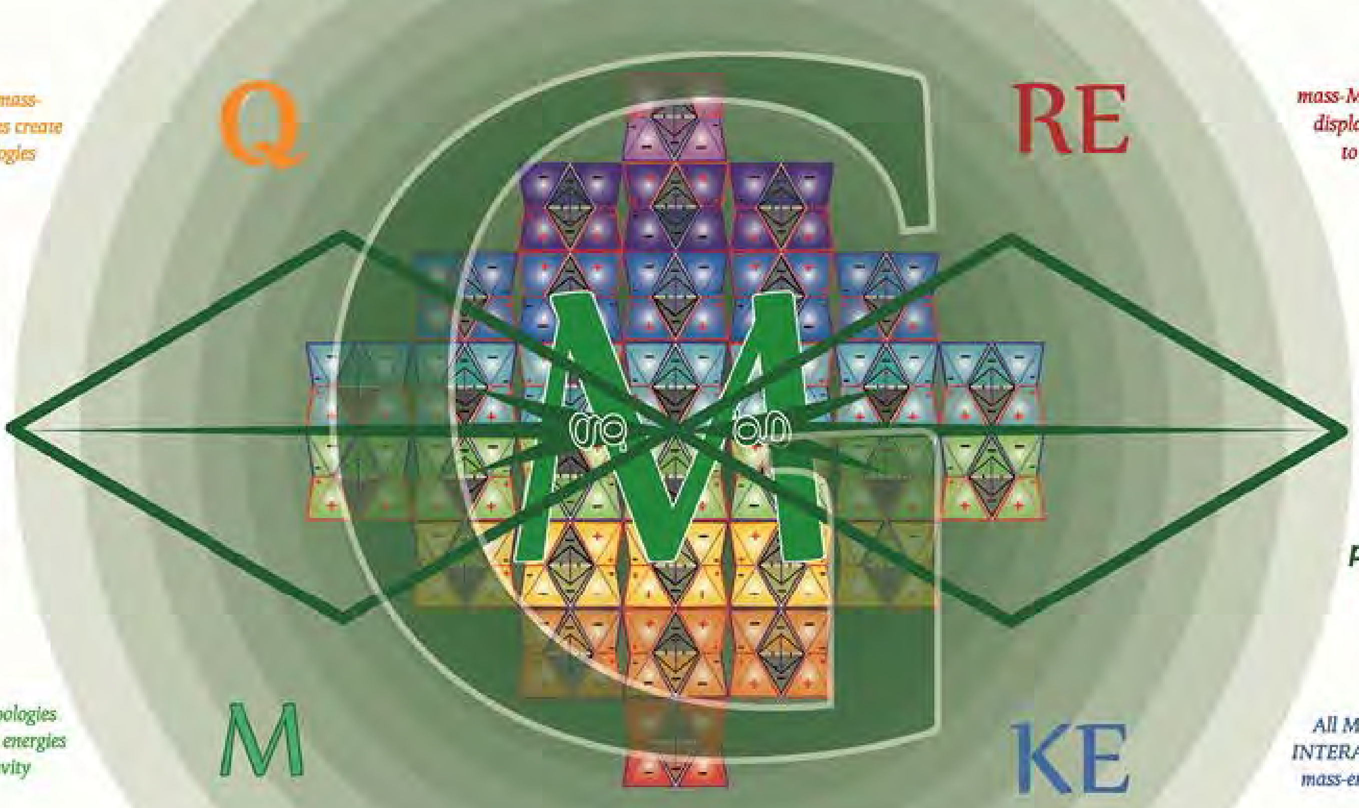

M

All Matter is gravitationally ATTRACTIVE

m

Gravity fields produce a convergent inverse square force

mass-Matter topologies displace vacuum energies to create gravity

All Matter radiates INTERACTIVE [k]EM mass-energy momenta

$$\nabla\phi\ grav = 4\pi G\rho$$

The pressure Force of Gravity is proportional to the amount of mass-Matter in a system and can be modelled with weak superpositioned E-field geometries

c^4

$$F = -G\frac{M_1 m_2}{r^2}$$

$$G = \frac{8\pi G T_{\mu\nu}}{c^4}$$

Copyright ABRAHAM [2008] - All rights reserved

Inverse squared interactive E-fields

Although present, the Gravitational Attraction between quantum scale Matter particles is vastly overwhelmed by the Coulombic Interaction field resulting from the nett Charge imbalances of Matter and its fields of motion

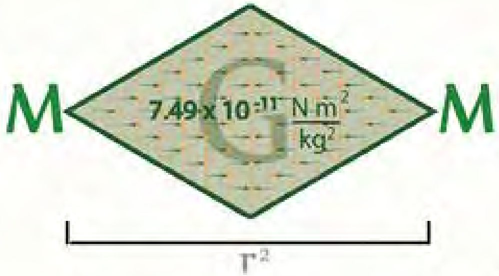

wherever mass-energies create standing-wave topologies they also create Gravity

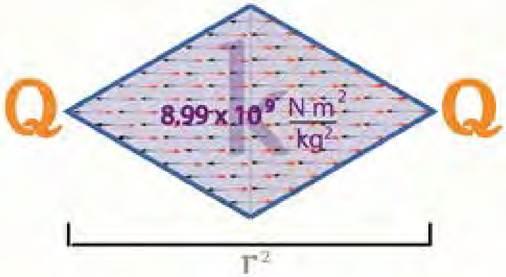

wherever there exists mass-energy imbalances energy will seek equilibirum via radiant EM fields

Protons have Stronger G fields than electrons but weaker [k]EM fields

electrons have Weaker G fields than Protons but stronger [k]EM fields

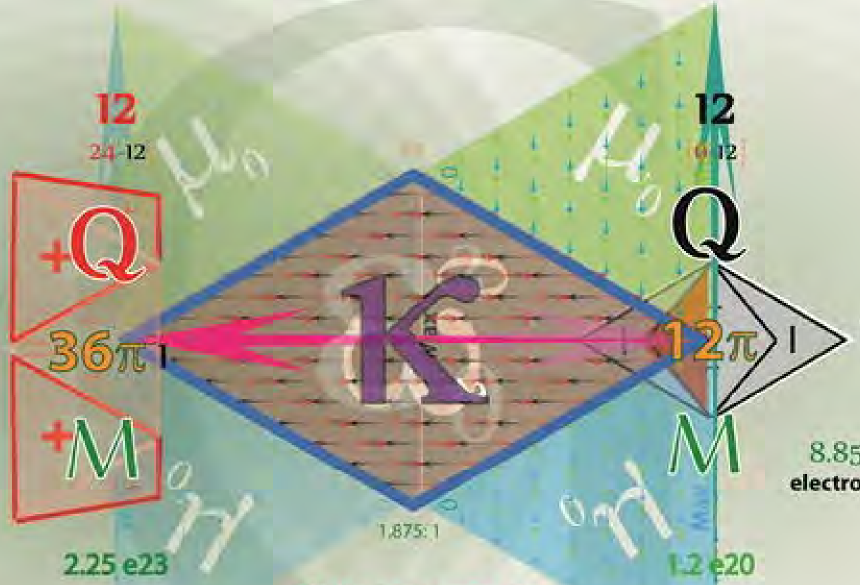

1.6596×10^{-27} KG
proton mass-Matter

8.8514×10^{-31} KG
electron mass-Matter

All Matter in motion create inverse squared forces due to their divergent [K]EM fields

	Matter		mass-energy		Matter		mass-energy	
M	36π	*	25n²	*KEM*	12π	*	1n²	*[k]EM*
	topology		geometry		topology		geometry	

superpositioned kEM fields create forces of attraction & repulsion

Gravity creates a strictly convergent force of attraction

G

c^4

mass-Matter topologies displace vacuum energies to form convergent gravity

Equilateral [k]EM mass-energies contribute to the total stress energies of any region

Copyright ABRAHAM [2008] - All rights reserved

vacuum impedance

electric permittivity $\varepsilon_0 = \frac{1}{c^2 \mu_0}$ magnetic permeability

$\varepsilon_0 = 8.8541878176\ e^{-12}$ F/m

Vacuum Permittivity

The Electric fields produced by Matter produce interactive nett divergent forces

Opposites ATTRACT
Similars REPEL

$P = \frac{kT}{V}$

Electric field permittivity contributes to STRESS energies

CHARGED mass-energy geometries create Matter topologies

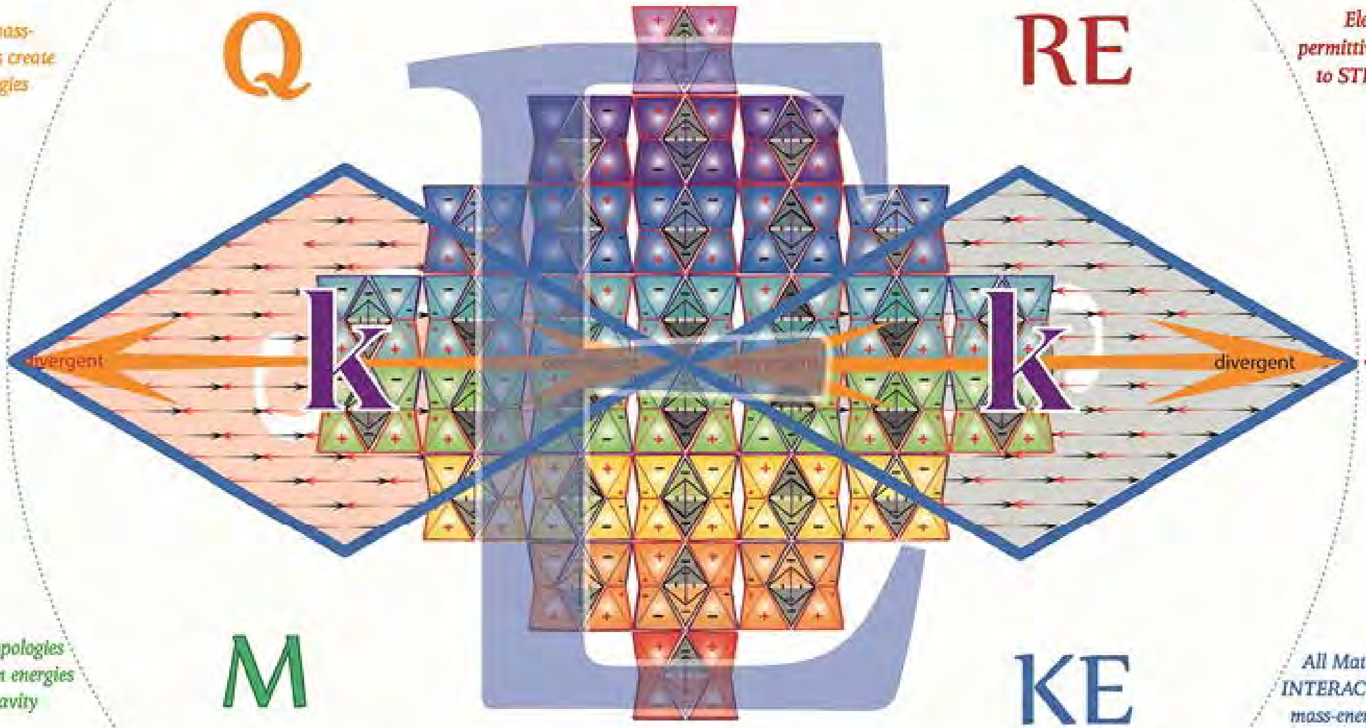

All Electric fields are nett divergent charge geometries >

< ***Electric fields follow the inverse square law***

mass-Matter topologies displace vacuum energies to create gravity

All Matter radiates INTERACTIVE [k]EM mass-energy momenta

Equal opposing E-field energy-momenta create electrostatic fields around Matter [that produce Lorentz forces]

c^2

vacuum impedance $Z = \frac{1}{c^2} = \varepsilon_0 \mu_0$ (electric permittivity, magnetic permeability)

electric permittivity of a spatial region

$4\pi \frac{1}{c^2 \mu_0} \left[\left[m \Omega v^2 \right]\right]$

Planck quanta
mass velocity

Perturbative M-forces

ElectroMagnetism is a velocity dependent vector interaction that dominates the Gravitational attraction of Matter

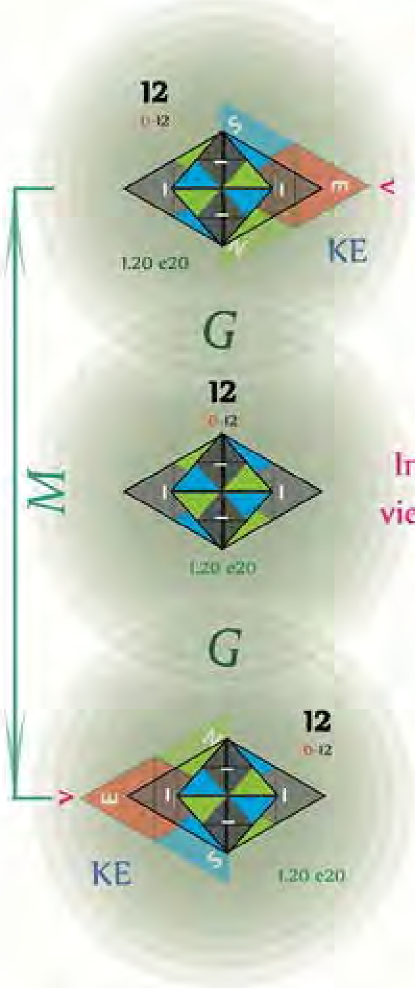

Inertial viewpoint

EM

Maxwells ElectroMagnetism

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \cdot \mathbf{B} = 0$$

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}$$

$$\nabla \times \mathbf{B} = \frac{1}{\epsilon_0 c^2}\mathbf{J} + \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t}$$

[k]EM fields have electric and magnetic field components

The Tetryonic GEM model reveals the interplay of Convergent [Gravity], interactive [E-fields] & perturbative [M-field] forces

Gravitational Matter is ALWAYS attractive vectorial ElectroMagnetic fields are interactive

Gravity can be modelled with pseudo-electric and pseudo-magnetic field components

GravitoElectroMagnetism

$$\nabla \cdot \mathbf{E}_g = -4\pi G \rho_g$$

$$\nabla \cdot \mathbf{B}_g = 0$$

$$\nabla \times \mathbf{E}_g = -\frac{\partial \mathbf{B}_g}{\partial t}$$

$$\nabla \times \mathbf{B}_g = -\frac{4\pi G}{c^2}\mathbf{J}_g + \frac{1}{c^2}\frac{\partial \mathbf{E}_g}{\partial t}$$

G

Inertial viewpoint

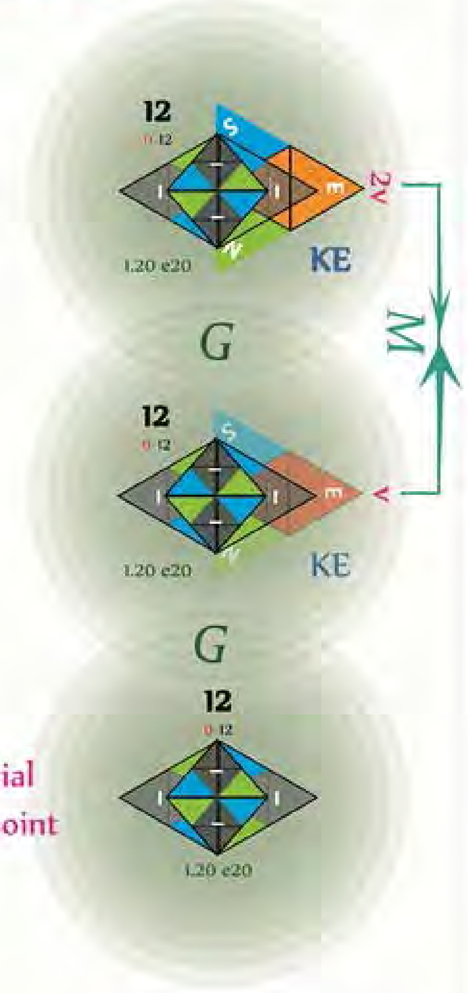

The Gravitational forces on the central particle are always equal irrespective of the inertial viewpoints

The sign of the effect is different (attraction in one case, repulsion in another), but this is just because in electromagnetism the sign of the electric part of the acceleration is opposite that of gravity; electric charges of the same type repel each other, while in gravity two massive objects attract.

Apart from this sign, there is such a close analogy to magnetism that this velocity-dependent gravitational effect is often termed Gravito-magnetism

The ElectroMagnetic forces resulting from inertial motion in a different frame of reference produces different KINETIC [EM] forces

Copyright ABRAHAM [2008] - All rights reserved

vacuum impedance

magnetic permeability $\mu_0 = \frac{1}{c^2 \varepsilon_0}$ electric permittivity

The Magnetic fields of Matter produce perturbative interactions

μ_0 = 1.2566370614 e^{-12} N/A^2

Vacuum Permeability

$P = \frac{kT}{V}$

Magnetic field permeability contributes to STRESS energies

CHARGED mass-energy geometries create Matter topologies

Q

RE

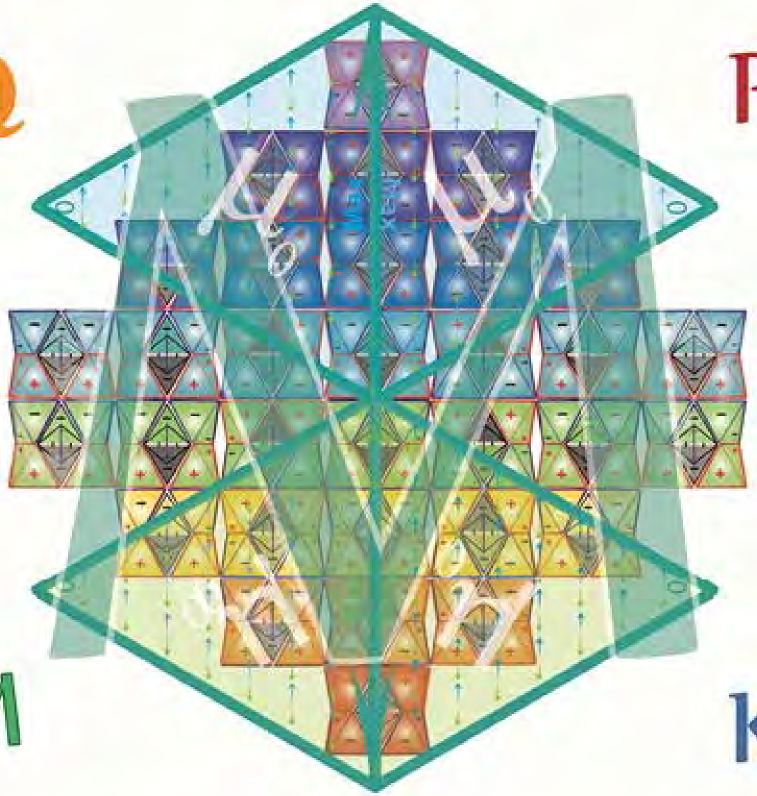

Magnetic dipoles follow the inverse cubed law

Magnetic fields follow the inverse square law

mass-Matter topologies displace vacuum energies to create gravity

M

KE

All Matter radiates INTERACTIVE [k]EM mass-energy momenta

Matter posesses a secondary magnetic moment that is subject to Lorentz forces

c^2

vacuum impedance $Z = \frac{1}{c^2} = \varepsilon_0 \mu_0$ electric permittivity, magnetic permeability

magnetic permeability of a spatial region

$4\pi \frac{1}{c^2 \varepsilon_0} \left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}} \Omega \overset{\text{quanta}}{\underset{\text{velocity}}{v^2}} \right] \right]$

Gravitational & EM fields

Any unified Field equation that attempts to accurately model celestial scale gravitational attraction must also take into account KEM field interactions between the bodies resulting from their respective motions

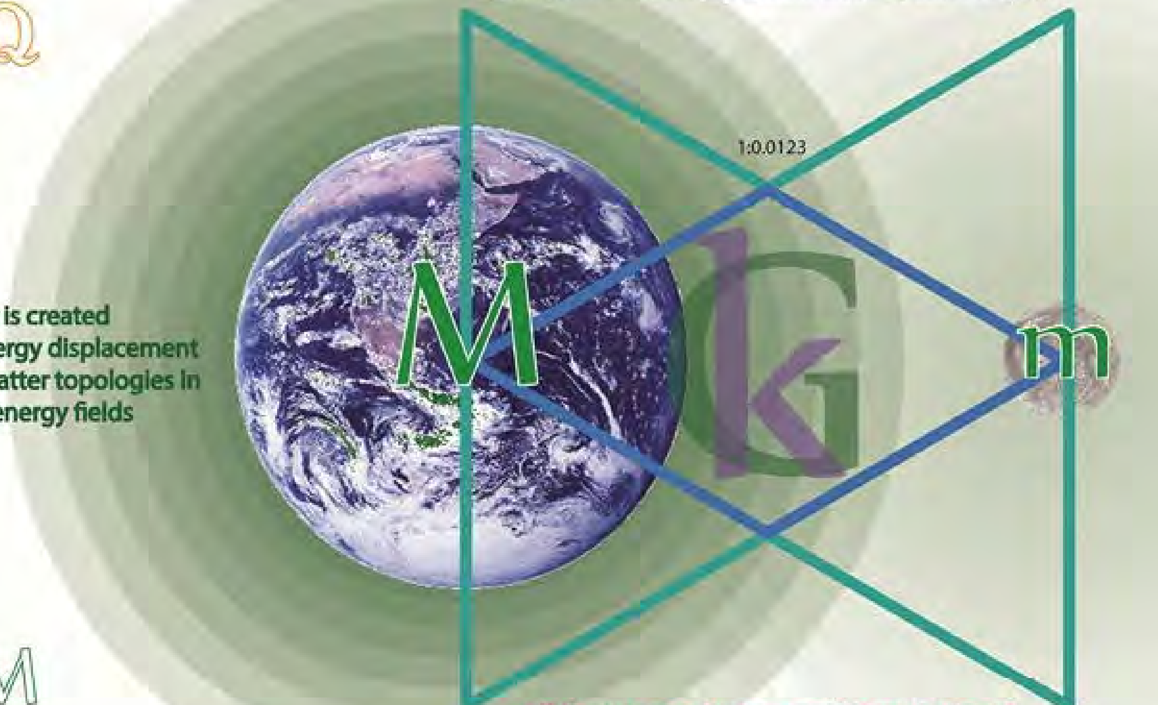

Gravity is created by vacuum energy displacement produced by Matter topologies in vacuum energy fields

along with displaced Vacuum energy gradients

Interactive E-fields and perturbative M-fields

also contribute to the nett convergent gravitational field

All Matter produces a nett Gravitational [convergent] force as a direct result of its mass-Matter topologies & KEM mass-energies

mass-Matter
5.9736×10^{24} KG

mass-Matter
7.3477×10^{22} KG

interactive KEM field mass-energy momenta

$$\text{Earth}\ \left[\overset{\text{Matter}}{\frac{4n\pi}{c^4}}\overset{\text{Energy}}{\underset{\text{mass}}{[\rho]}} + \left[2n\pi\overset{\text{EM mass}}{\frac{[\rho]}{c^2}}\ \kappa\ 2n\pi\overset{\text{EM mass}}{\frac{[\rho]}{c^2}}\right] + \overset{\text{Matter}}{\frac{4n\pi}{c^4}}\overset{\text{Energy}}{\underset{\text{mass}}{[\rho]}}\right]\ \text{moon}$$

2.8673e26 π
displacement topology

3.526e24 π
displacement topology

$$\xleftrightarrow{\qquad r^2 \qquad}$$

Gravitational gradients can be modelled as weak KEM fields

Independently and almost simultaneously in1964 three groups of physicists proposed the proposed the Higgs mechanism thorugh which the inertial mass properties of Matter are created:
François Englert and Robert Brout; by Peter Higgs (inspired by ideas of Philip Anderson); and by Gerald Guralnik, C. R. Hagen, and Tom Kibble.

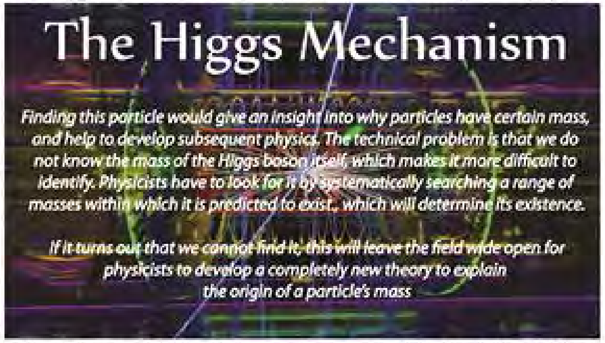

The Higgs Mechanism

Finding this particle would give an insight into why particles have certain mass, and help to develop subsequent physics. The technical problem is that we do not know the mass of the Higgs boson itself, which makes it more difficult to identify. Physicists have to look for it by systematically searching a range of masses within which it is predicted to exist.. which will determine its existence.

If it turns out that we cannot find it, this will leave the field wide open for physicists to develop a completely new theory to explain the origin of a particle's mass

The Higgs mechanism was incorporated into modern particle physics by Steven Weinberg and Abdus Salam, and is an essential part of the standard model.

As the Universe cooled and the temperature fell below a critical value, an invisible force field called the 'Higgs field' was formed together with the associated 'Higgs boson'

The Higgs mechanism is a process by which vector bosons can obtain mass

M

TETRYONS / FERMIONS / BARYONS
3D tetrahedral standing-waves
[4nπ EM mass-Matter topologies]
interact with the vacuum energy aether

Matter

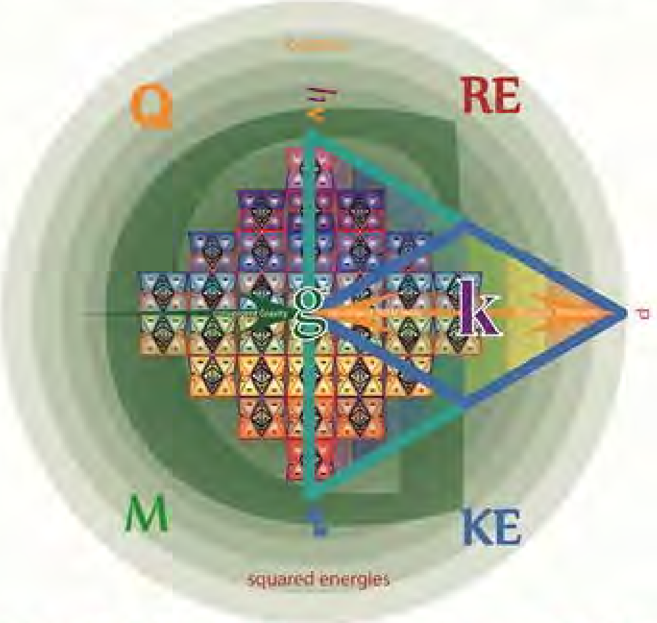

m

BOSONS / PHOTONS
2D planar matterless fields
[nπ EM mass-energy geometries]
'slice' through the vacuum energy aether

EM masses

Tetryonics is a full relativistic unified theory of electromagnetic mass-ENERGY & Matter and motion in which the classical properties of inerital mass and quantum mechanical quantised energy momenta are explained as being direct result of the geometric properties of inductive equilateral energies moving through external electromagnetic fields

Applying the Unified Field Equation

for ElectroMagnetic masses & Matter interactions

All Matter has a
4π standing-wave
mass-energy topology

Matter topologies

All kEM waves have
4π superpositioned
mass-energy geometries

mass-ENERGY geometries

$$G_{ab} = 8\pi G T_{\mu\nu}$$

The net gravitational force is the geometric mean of all EM mass-ENERGY-Matter interactions

$$-G\frac{M_1 m_2}{r^2} = F$$

$$F = k\frac{Q_1 Q_2}{r^2}$$

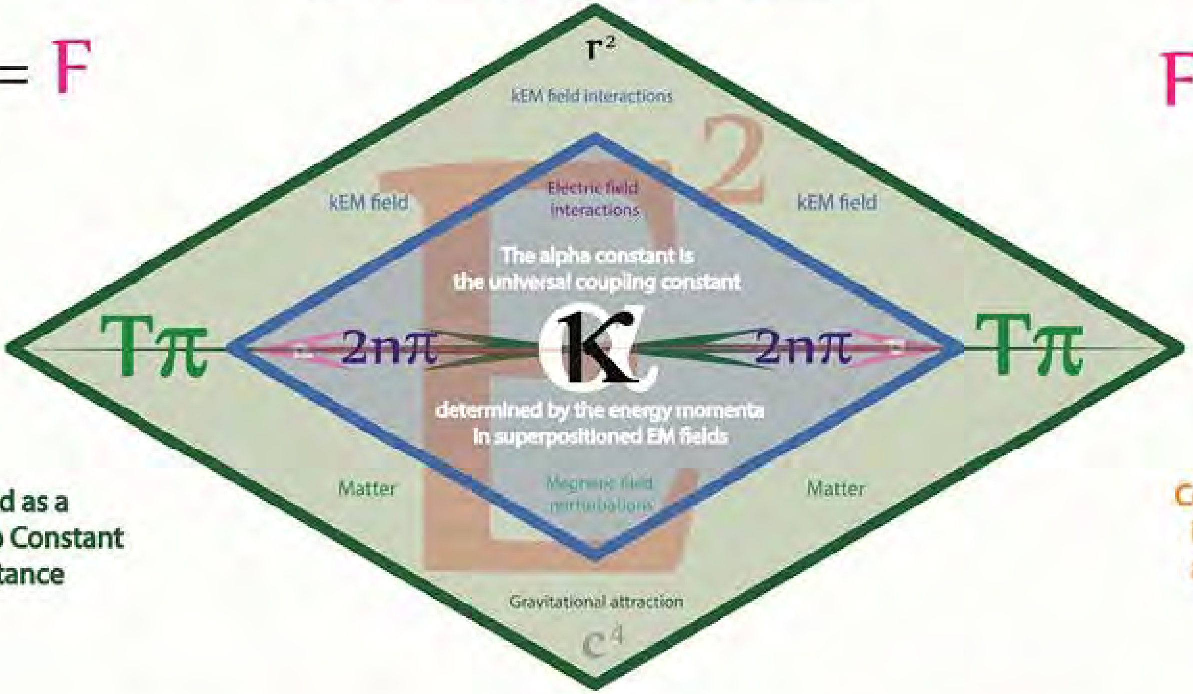
r²
kEM field interactions
E²
Electric field interactions
kEM field
kEM field
The alpha constant is the universal coupling constant
Tπ
2nπ
K
2nπ
Tπ
determined by the energy momenta in superpositioned EM fields
Matter
Magnetic field perturbations
Matter
Gravitational attraction
c⁴

G

k

Gravity can be modelled as a reduced negative Coloumb Constant acting over a large distance

Coloumb's Constant is a large interactive Electric constant acting over a short distance

Tetryons - 4π
Leptons - 12π
Quarks - 12π
Mesons - 24π
Baryons - 36π

Charges - 1π

Photons -2π

EM Fields - 2nπ

Ignoring interactive [k]EM field interactions reduces GR [k] to Newtonian [G]

$$G_{ab} = 4\pi G T_{\mu\nu}$$

Tetryonic two body interactions

$$G_{ab} = 8\pi G T_{\mu\nu}$$

Factoring all GEM interactions
Tetryonics resolves to General relativity

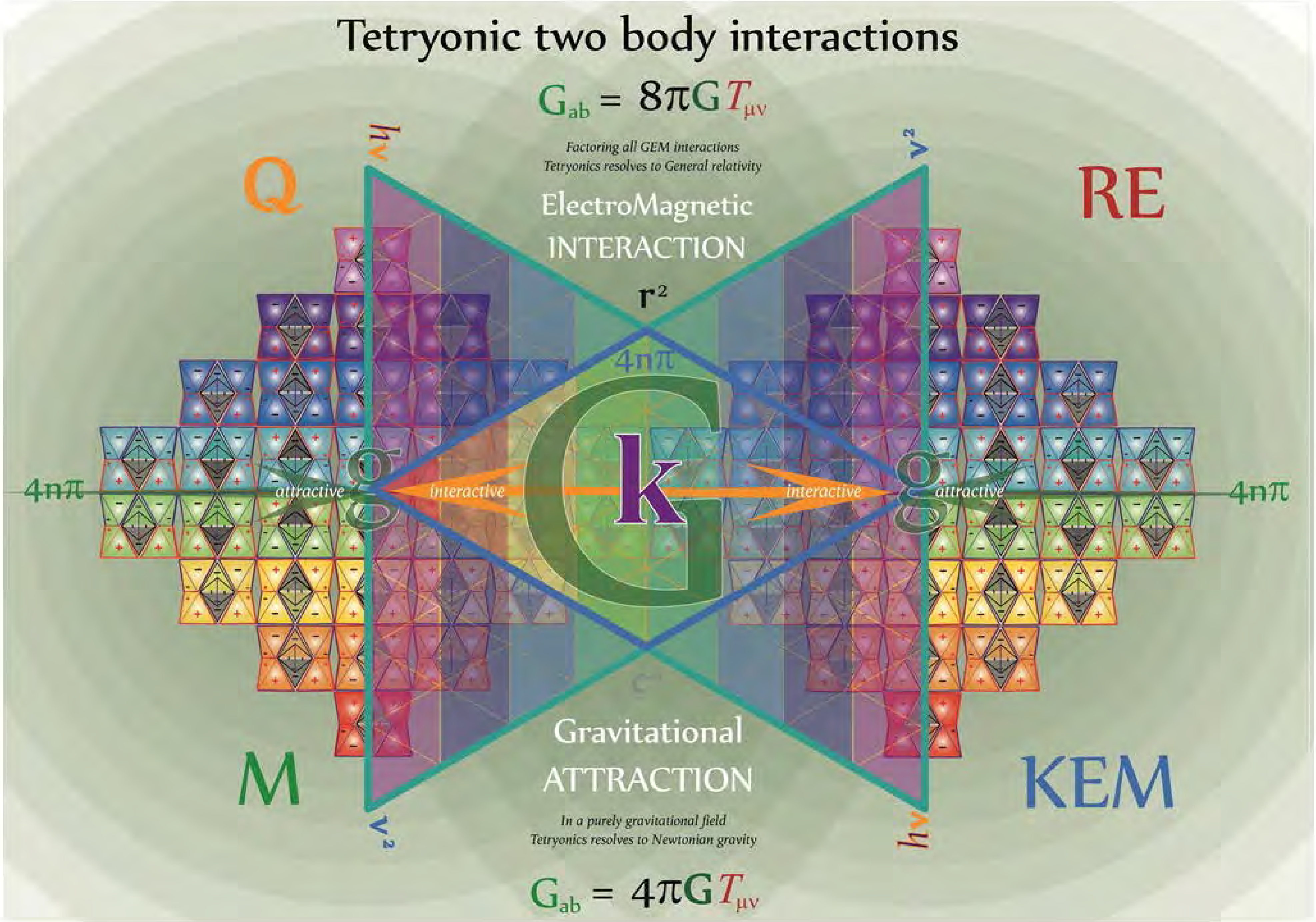
hv
v²
Q
RE
ElectroMagnetic
INTERACTION
r²
4nπ
G
k
4nπ
attractive
g
interactive
interactive
g
attractive
4nπ
c²
Gravitational
ATTRACTION
M
KEM
v²
hv

In a purely gravitational field
Tetryonics resolves to Newtonian gravity

$$G_{ab} = 4\pi G T_{\mu\nu}$$

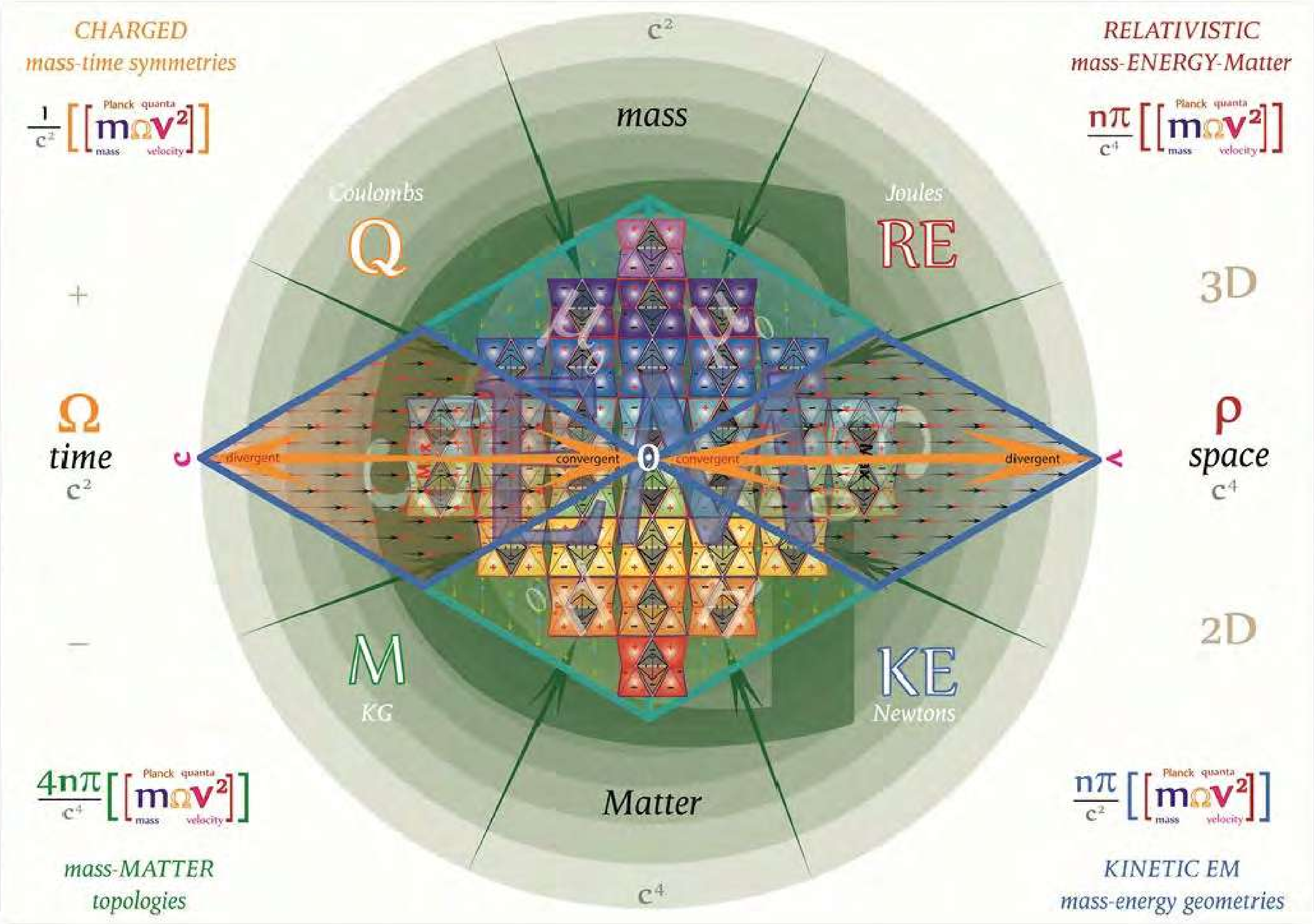
CHARGED
mass-time symmetries
1/c² [[mΩv²]]
Planck quanta
mass velocity
RELATIVISTIC
mass-ENERGY-Matter
nπ/c⁴ [[mΩv²]]
c²
mass
Coulombs
Q
Joules
RE
3D
+
Ω
time
c²
c
divergent
convergent
0
convergent
divergent
v
ρ
space
c⁴
−
2D
M
KG
KE
Newtons
Matter
c⁴
4nπ/c⁴ [[mΩv²]]
nπ/c² [[mΩv²]]
mass-MATTER
topologies
KINETIC EM
mass-energy geometries

Tetryonics 71.14 - Quantum GEM space-time fields

Our EM SUN

The SUN (like the Planets) is a CHARGED body moving through an EM environment

G & E fields obey the inverse SQUARED laws

The nett Charge of a body is the result of an excess of one type of energy quanta and is reflective of their charged topologies [ie ionised Matter]

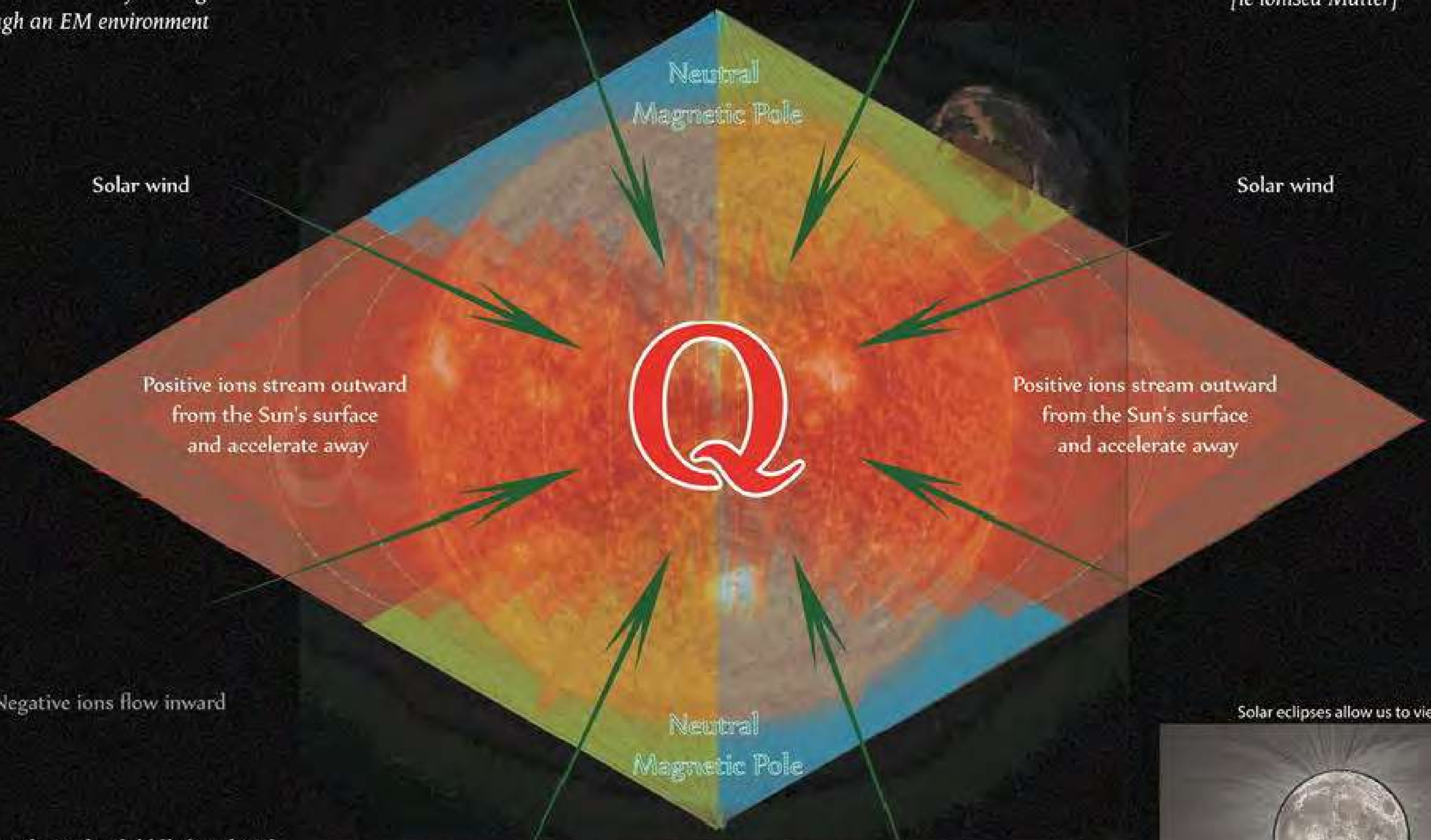

Solar wind

Solar wind

Positive ions stream outward from the Sun's surface and accelerate away

Positive ions stream outward from the Sun's surface and accelerate away

Negative ions flow inward

The charged E field [heliosphere] of our SUN interacts with other charged planetary bodies

Magnetic dipole fields follow the inverse CUBED law

Solar eclipses allow us to view

the SUN's EM corona in detail

Solar Dynamics

The SUN is a ElectroMagnetohydrodynamo

Prominences

Flares, Prominences and CMEs are the result of Electro-Magnetic geometries

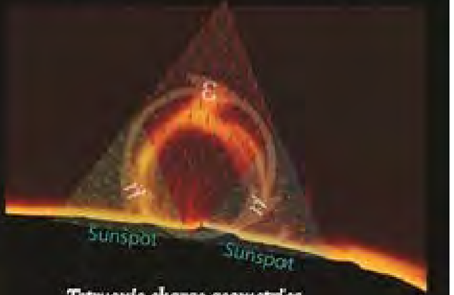

Tetryonic charge geometries explain Solar EM field dynamics

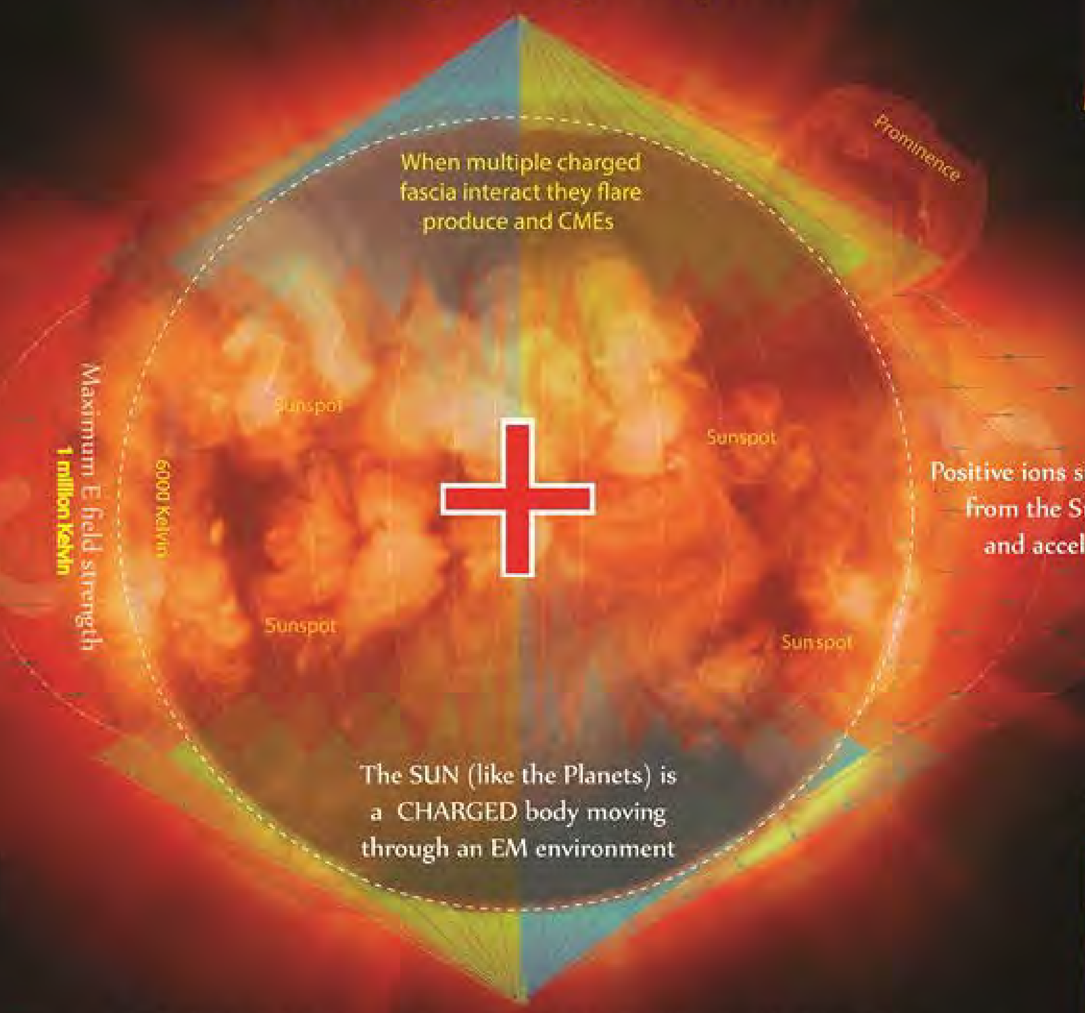

When multiple charged fascia interact they flare produce and CMEs

Positive ions stream outward from the Sun's surface and accelerate away

The SUN (like the Planets) is a CHARGED body moving through an EM environment

Where Solar EM loops breakdown or interact flares result

Solar Flares

The breaking of large inductive EM loops creates CMEs [Coronal Matter Ejections]

CMEs

The Earth is located at an average distance of about 150 million kilometers [1AU] from the Sun
It takes sunlight travelling at the speed of Light [c] about 8.3 minutes to cross this distance and reach the Earth.

BUT due to the nature of Longitudinal waves emitted by the SUN mechanical momenta impulses [action-at-a-distance] can be transfered from the SUN to the planets nearly instantanteously providing the mechanism for Newton's instataneous pull of Gravity on planetary bodies

Copyright ABRAHAM [2008] - All rights reserved

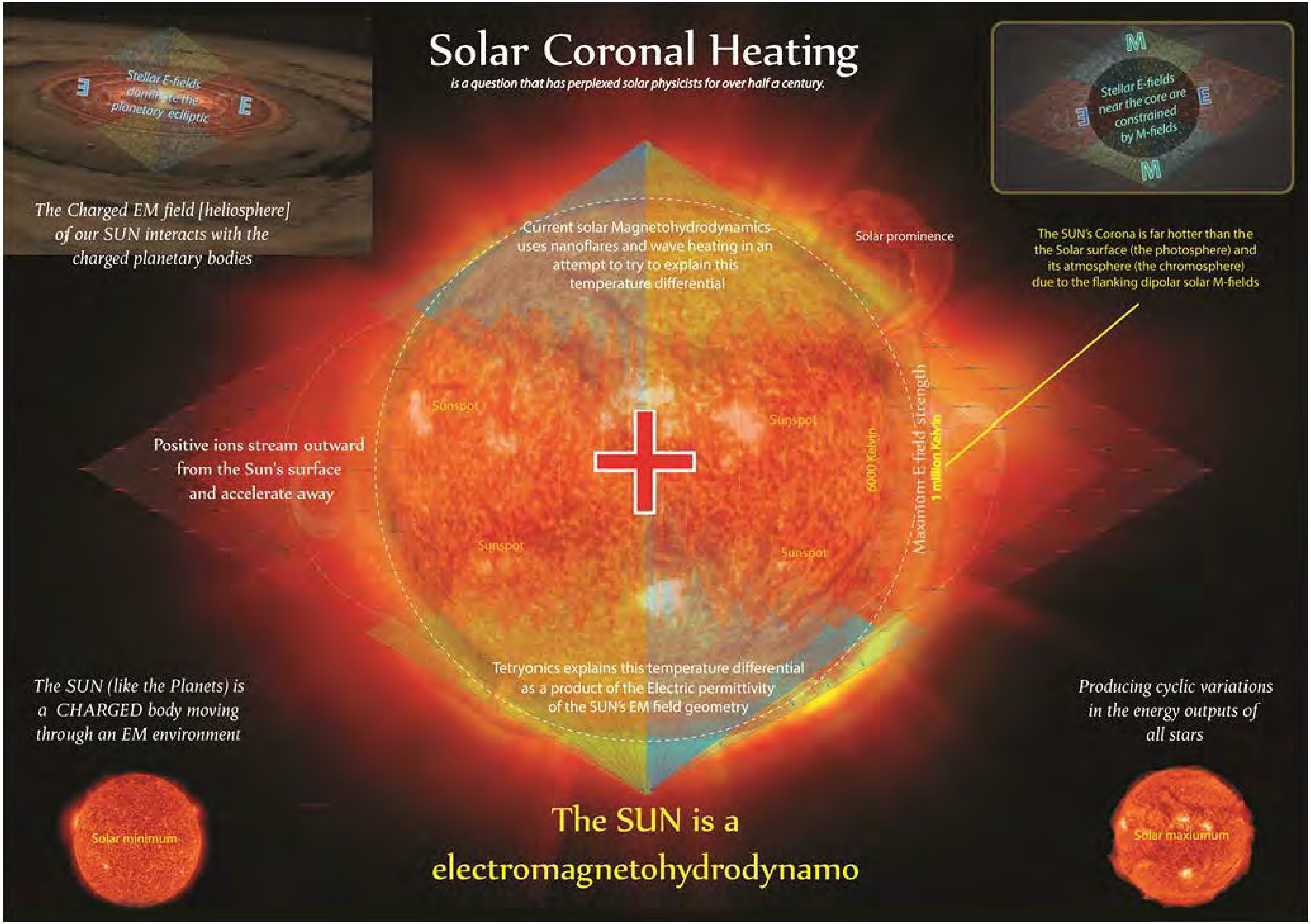

Stellar E-fields dominate the planetary ecliptic
E
E
M
Stellar E-fields near the core are constrained by M-fields
E
E
M
Sunspot
Sunspot
Sunspot
Sunspot
6000 Kelvin
Maximum E-field strength
1 million Kelvin
Solar minimum
Solar maxiumum

Solar Coronal Heating

is a question that has perplexed solar physicists for over half a century.

The Charged EM field [heliosphere] of our SUN interacts with the charged planetary bodies

Current solar Magnetohydrodynamics uses nanoflares and wave heating in an attempt to try to explain this temperature differential

Solar prominence

The SUN's Corona is far hotter than the the Solar surface (the photosphere) and its atmosphere (the chromosphere) due to the flanking dipolar solar M-fields

Positive ions stream outward from the Sun's surface and accelerate away

Tetryonics explains this temperature differential as a product of the Electric permittivity of the SUN's EM field geometry

The SUN (like the Planets) is a CHARGED body moving through an EM environment

Producing cyclic variations in the energy outputs of all stars

The SUN is a electromagnetohydrodynamo

Tetryonics 72.03 - Solar Coronal Heating

Solar cycles

On an 11 year cycle, the shape of the corona oscillates from a wide crown about the Sun's equator to a completely closed envelope surrounding the Sun

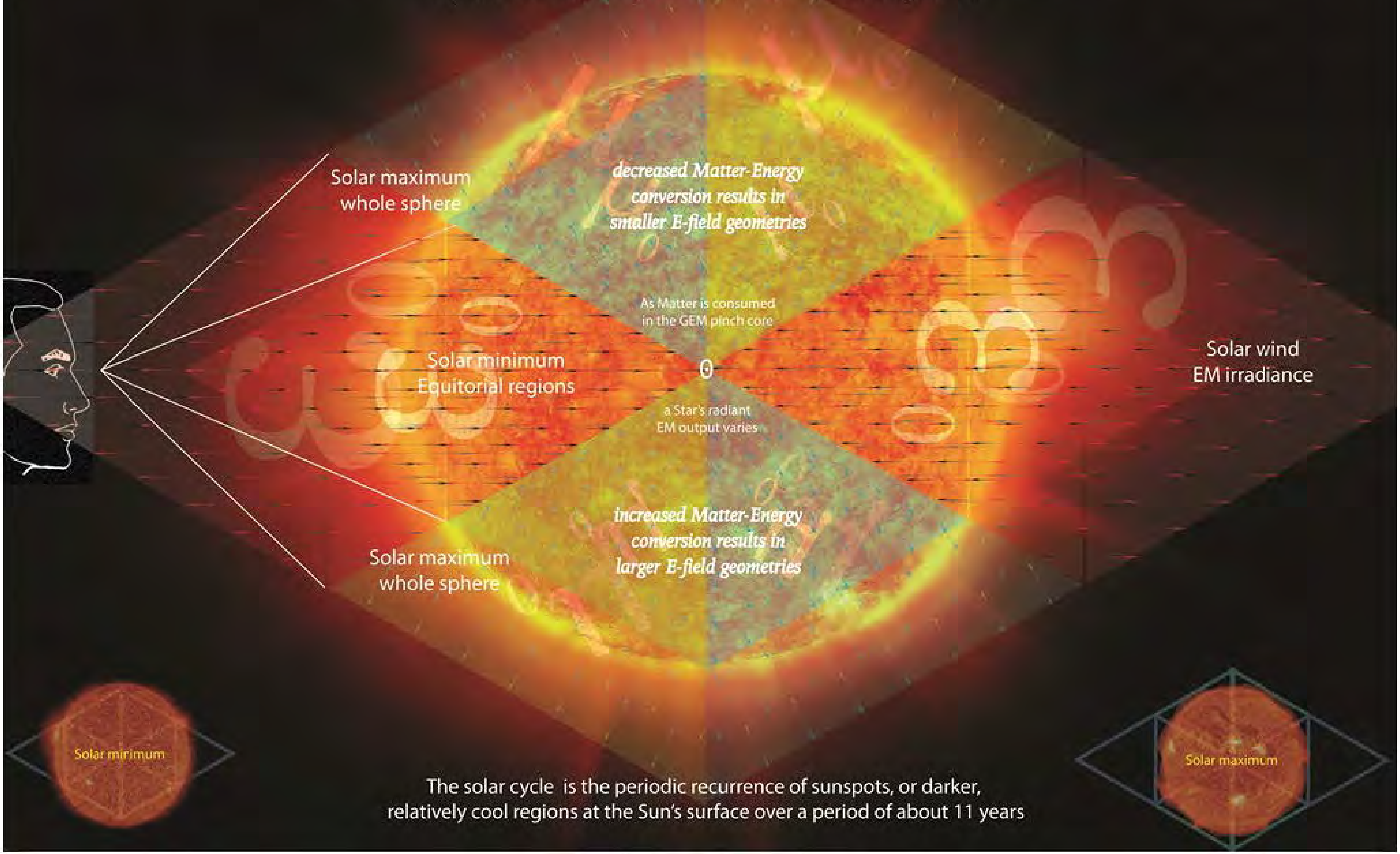
Solar maximum
whole sphere
decreased Matter-Energy
conversion results in
smaller E-field geometries
As Matter is consumed
in the GEM pinch core
Solar minimum
Equitorial regions
0
a Star's radiant
EM output varies
Solar wind
EM irradiance
Solar maximum
whole sphere
increased Matter-Energy
conversion results in
larger E-field geometries
Solar minimum
Solar maximum

The solar cycle is the periodic recurrence of sunspots, or darker, relatively cool regions at the Sun's surface over a period of about 11 years

Sunspot cycles

This is a question that has perplexed solar physicists for over half a century.

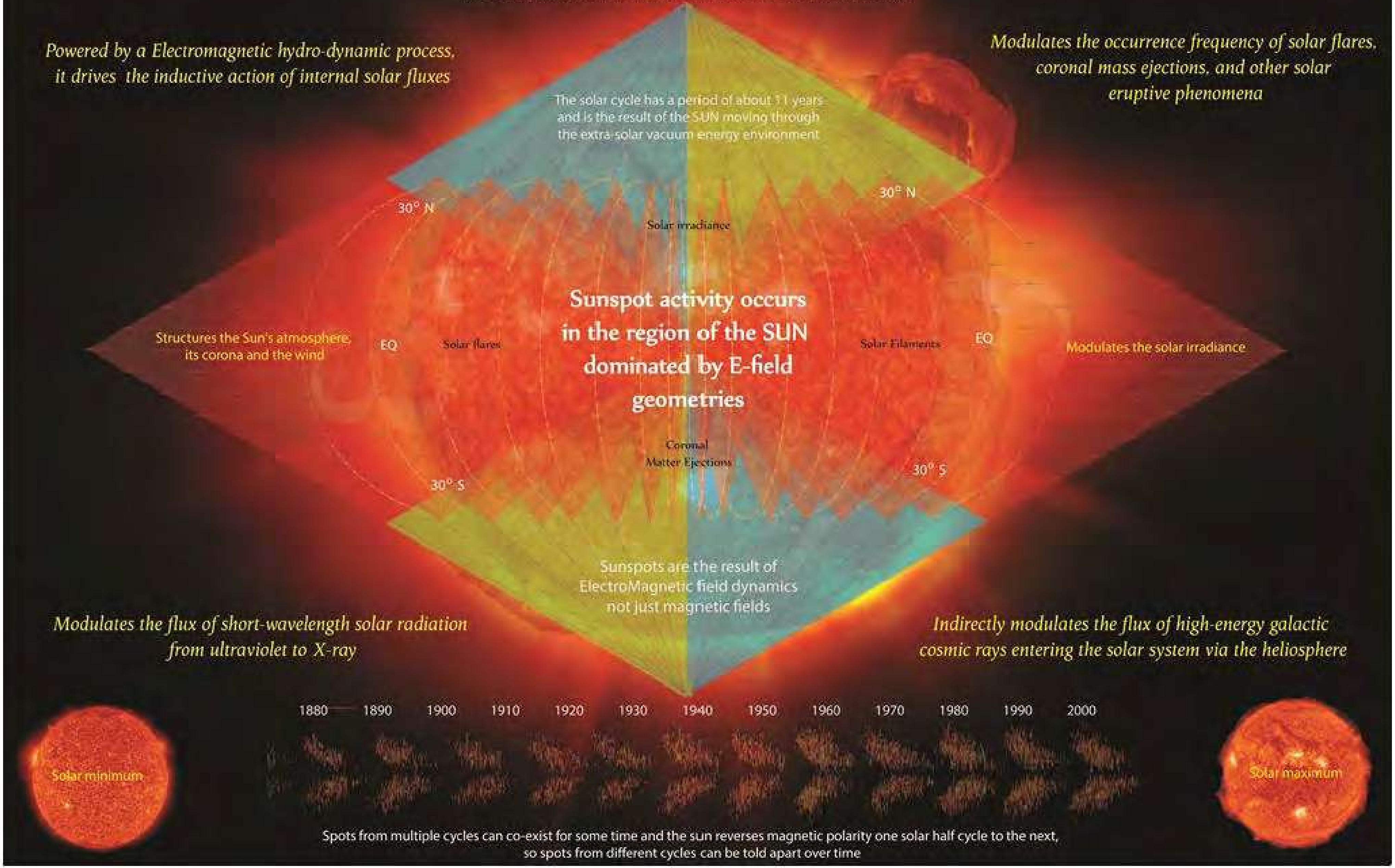
30° N
30° N
Solar irradiance
EQ
Solar flares
Solar Filaments
EQ
Coronal
Matter Ejections
30° S
30° S
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
Solar minimum
Solar maximum

Powered by a Electromagnetic hydro-dynamic process, it drives the inductive action of internal solar fluxes

The solar cycle has a period of about 11 years and is the result of the SUN moving through the extra-solar vacuum energy environment

Modulates the occurrence frequency of solar flares, coronal mass ejections, and other solar eruptive phenomena

Structures the Sun's atmosphere, its corona and the wind

Sunspot activity occurs in the region of the SUN dominated by E-field geometries

Modulates the solar irradiance

Sunspots are the result of ElectroMagnetic field dynamics not just magnetic fields

Modulates the flux of short-wavelength solar radiation from ultraviolet to X-ray

Indirectly modulates the flux of high-energy galactic cosmic rays entering the solar system via the heliosphere

Spots from multiple cycles can co-exist for some time and the sun reverses magnetic polarity one solar half cycle to the next, so spots from different cycles can be told apart over time

Solar irradiance

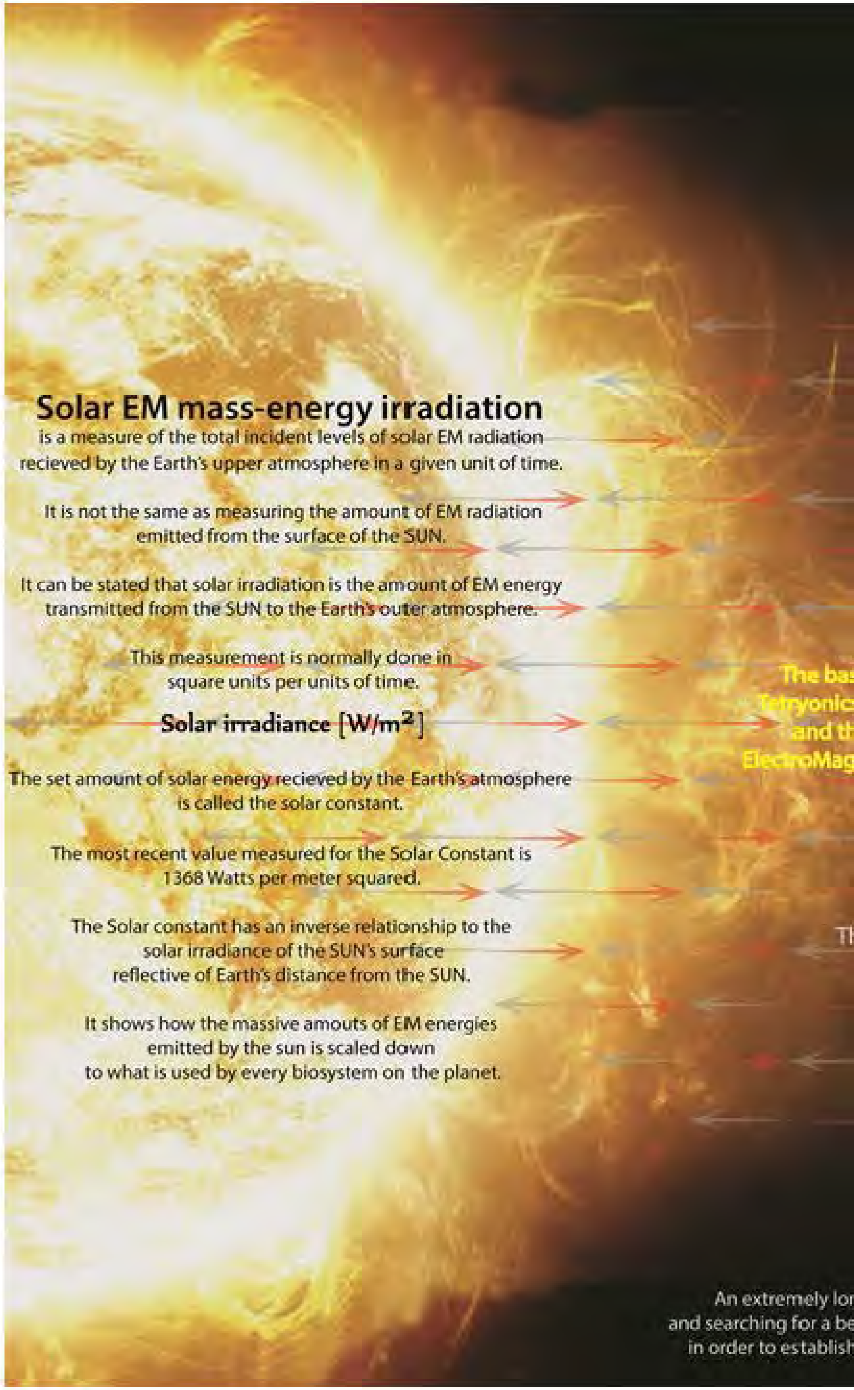

Solar EM mass-energy irradiation

is a measure of the total incident levels of solar EM radiation recieved by the Earth's upper atmosphere in a given unit of time.

It is not the same as measuring the amount of EM radiation emitted from the surface of the SUN.

It can be stated that solar irradiation is the amount of EM energy transmitted from the SUN to the Earth's outer atmosphere.

This measurement is normally done in square units per units of time.

Solar irradiance [W/m²]

The set amount of solar energy recieved by the Earth's atmosphere is called the solar constant.

The most recent value measured for the Solar Constant is 1368 Watts per meter squared.

The Solar constant has an inverse relationship to the solar irradiance of the SUN's surface reflective of Earth's distance from the SUN.

It shows how the massive amouts of EM energies emitted by the sun is scaled down to what is used by every biosystem on the planet.

The SUN is an ElectroMagnetic pinch reactor

The Sun emits longitudinal EM radiation across most of the electromagnetic spectrum.

The basic causes of the solar variability and solar cycles are still under debate, Tetryonics dictates a EM mechanism for the energies produced at the core of Stars and that the variability of the stellar outputs may be due to changes in the ElectroMagnetic environment that Stars [and their planetary systems] move through

The spectrum of the Sun's solar radiation is close to that of a black body with a temperature of about 5,800 K.

Solar irradiation produces Carbon 14

An extremely long stretch of low activity in recent years has left scientists baffled and searching for a better model of the SUN's interal processes and how to forcast its output in order to establish a better link between the solar cycle and global climate flucuations

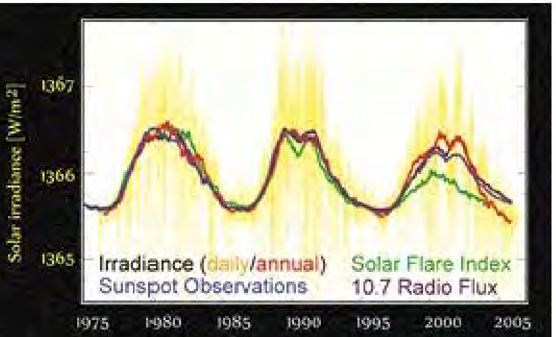

The total solar irradiance (TSI) is the amount of solar radiative energy incident on the Earth's upper atmosphere

Changes in the SUN's EM output can have a direct effect on the weather and electrical activity of the Earth and other planets

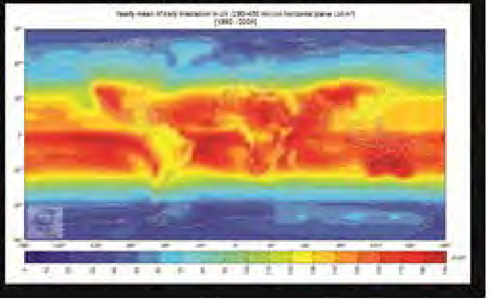

Copyright ABRAHAM [2008] - All rights reserved

Stellar refraction of EM waves

Any test of General Relativity using photons (or EM waves) will only measure SR interactions
[ie Gravitational red-shifting is a SR effect on EM masses not graviational Matter]

G

The motion of Matter is affected by gravitational attraction and EM interactions

*Photons are NOT 'bent' by Gravitational fields
they are refracted by propagation velocity changes
resulting from their EM interactions
with the solar corona*

EM

The motion of EM masses is affected by EM interactions and creates refraction

$$\alpha[\text{einstein}] = \frac{8\pi GM}{c^2}$$

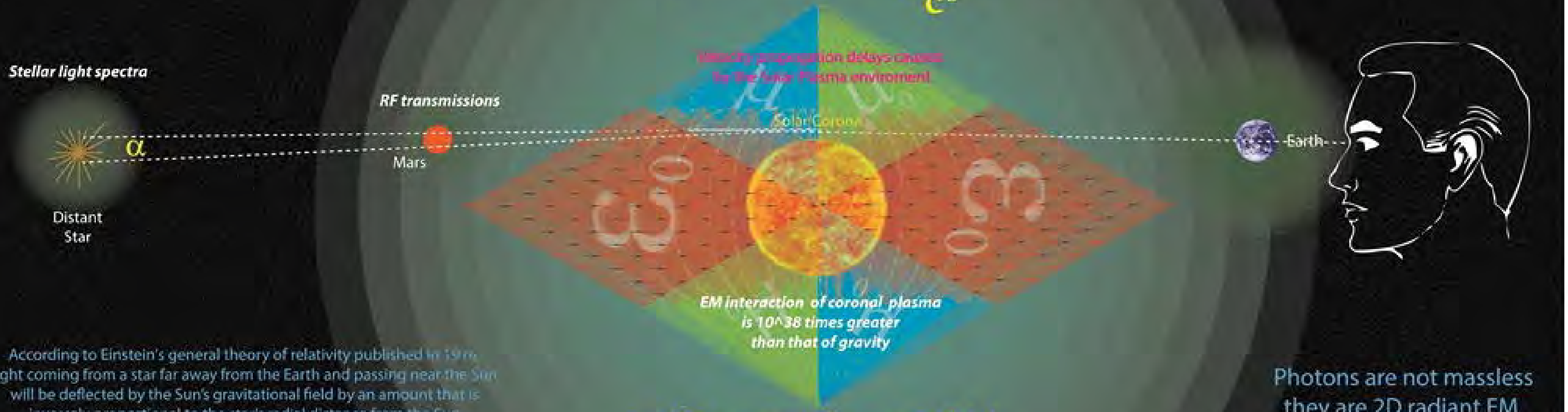
Stellar light spectra
α
Distant Star
RF transmissions
Mars
Velocity propagation delays caused by the Solar Plasma environment
μ
ω
Solar Corona
ε
ε
Earth
EM interaction of coronal plasma is 10^38 times greater than that of gravity

According to Einstein's general theory of relativity published in 1916
light coming from a star far away from the Earth and passing near the Sun
will be deflected by the Sun's gravitational field by an amount that is
inversely proportional to the star's radial distance from the Sun

1.745" at the Sun's limb
(this is the result of EM refraction NOT Matter gravitation)

$$\alpha[\text{newton}] = \frac{4\pi GM}{r}$$

Photons are not massless
they are 2D radiant EM
mass-energy momenta
geometries

r

*The additional 'bending' of light rays as they pass by a Gravitational body as predicited
in Einstein's general relativity is caused by refractive EM interactions
not by the force of gravity*

4π

4π

c^2

Gravitation in General relativity is Newton's Gravity with SR corrections

Tetryonics 72.07 - Stellar refraction of EM Waves

Copyright ABRAHAM [2008] - All rights reserved

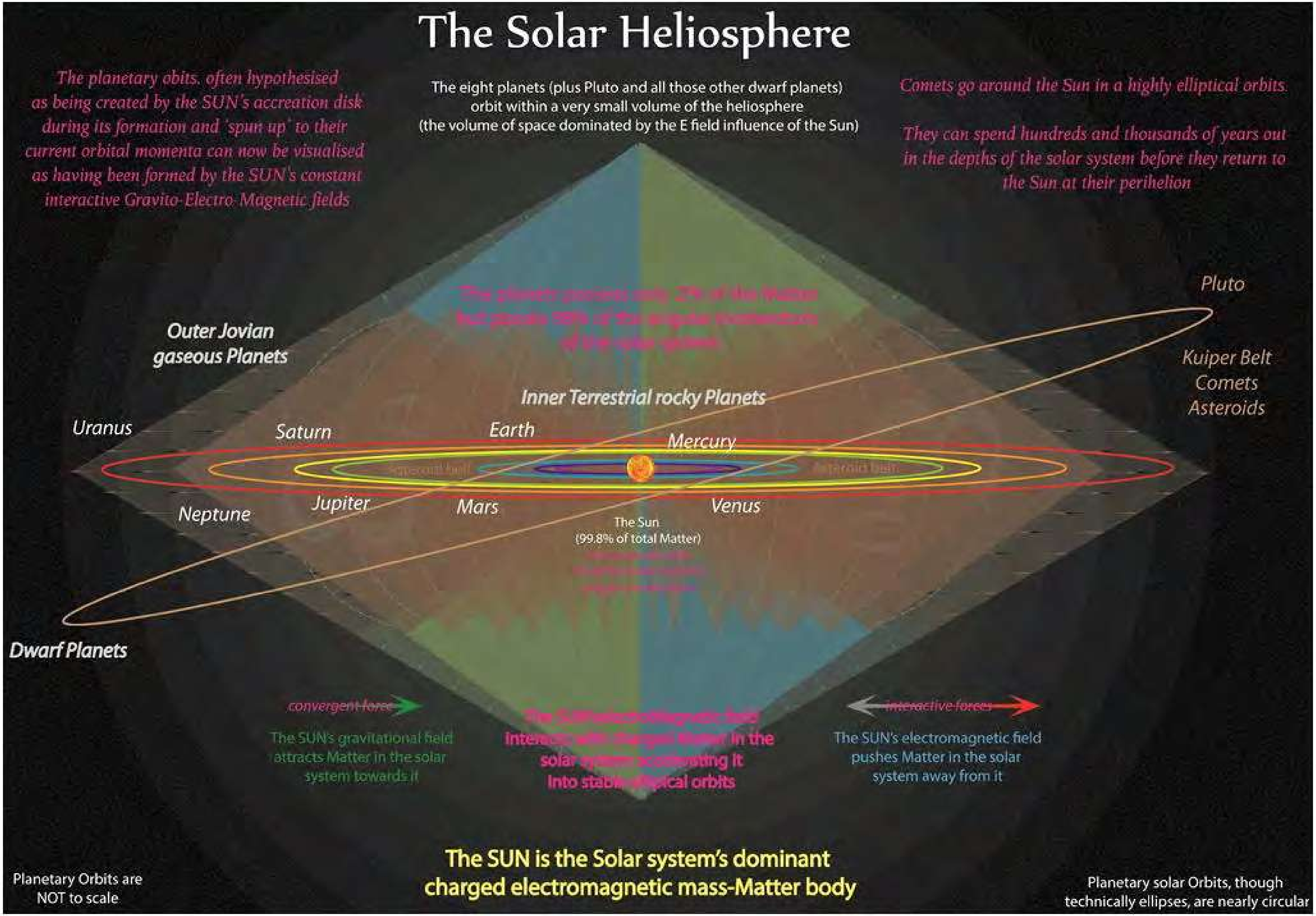

The Solar Heliosphere

The planetary obits, often hypothesised as being created by the SUN's accreation disk during its formation and 'spun up' to their current orbital momenta can now be visualised as having been formed by the SUN's constant interactive Gravito-Electro-Magnetic fields

The eight planets (plus Pluto and all those other dwarf planets)
orbit within a very small volume of the heliosphere
(the volume of space dominated by the E field influence of the Sun)

Comets go around the Sun in a highly elliptical orbits.

They can spend hundreds and thousands of years out in the depths of the solar system before they return to the Sun at their perihelion

The planets possess only 2% of the Matter but possess 98% of the angular momentum of the solar system

Pluto
Outer Jovian gaseous Planets
Kuiper Belt
Comets
Asteroids
Inner Terrestrial rocky Planets
Uranus
Saturn
Earth
Mercury
asteroid belt
asteroid belt
Neptune
Jupiter
Mars
Venus
The Sun
(99.8% of total Matter)
Dwarf Planets
convergent force
interactive forces

The SUN's gravitational field attracts Matter in the solar system towards it

The SUN's electroMagnetic field Interacts with charged Matter in the solar system accelerating it Into stable elliptical orbits

The SUN's electromagnetic field pushes Matter in the solar system away from it

The SUN is the Solar system's dominant charged electromagnetic mass-Matter body

Planetary Orbits are NOT to scale

Planetary solar Orbits, though technically ellipses, are nearly circular

Tetryonics 72.08 - The Solar Heliosphere

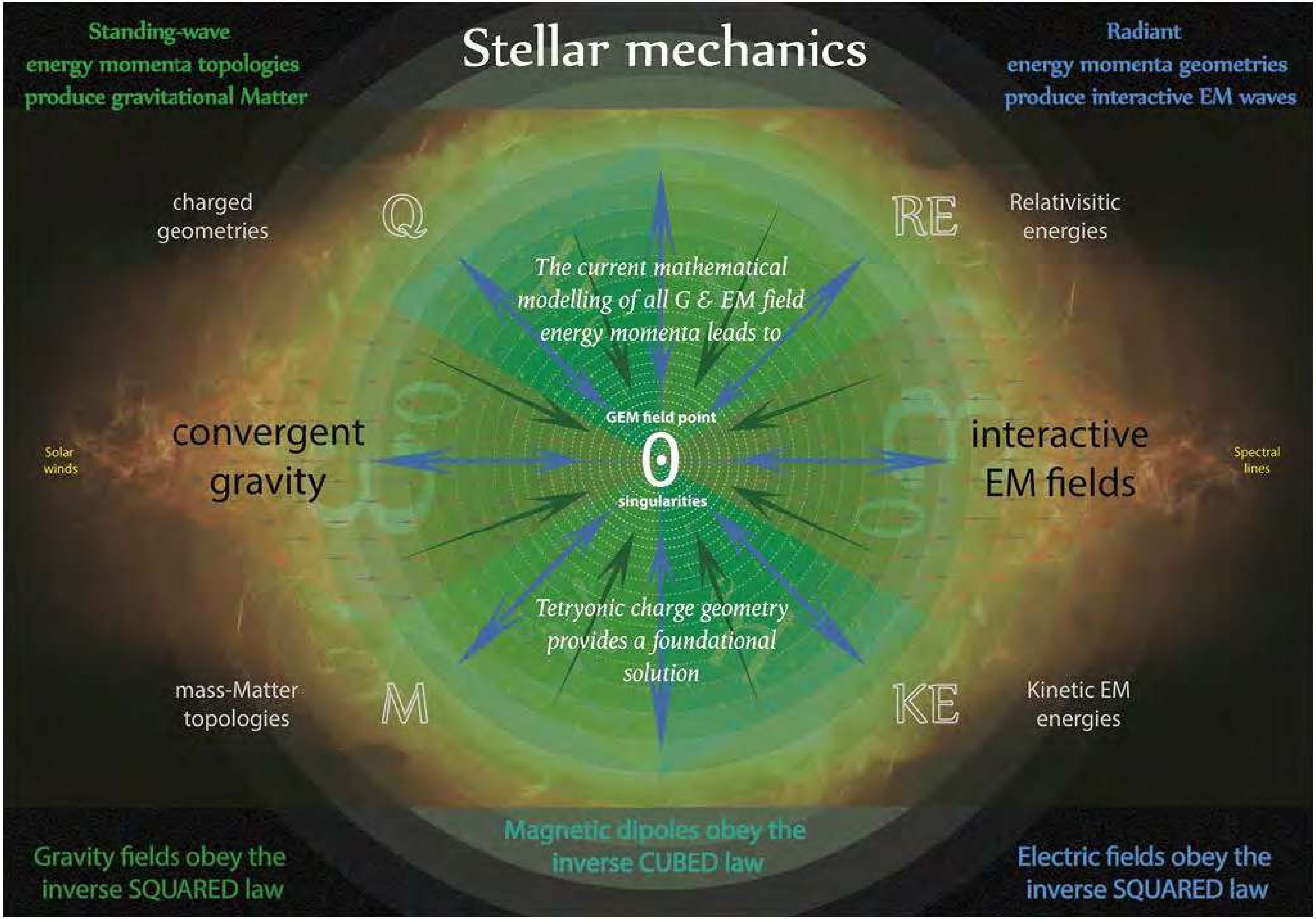

Stellar mechanics

Standing-wave
energy momenta topologies
produce gravitational Matter

Radiant
energy momenta geometries
produce interactive EM waves

charged
geometries

Q

RE

Relativisitic
energies

The current mathematical
modelling of all G & EM field
energy momenta leads to

GEM field point

0

singularities

Solar
winds

convergent
gravity

interactive
EM fields

Spectral
lines

Tetryonic charge geometry
provides a foundational
solution

mass-Matter
topologies

M

KE

Kinetic EM
energies

Gravity fields obey the
inverse SQUARED law

Magnetic dipoles obey the
inverse CUBED law

Electric fields obey the
inverse SQUARED law

Tetryonics 73.01 - Stellar mechanics

Anti-Matter Stars

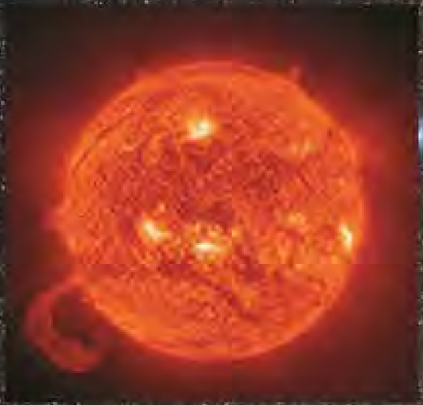

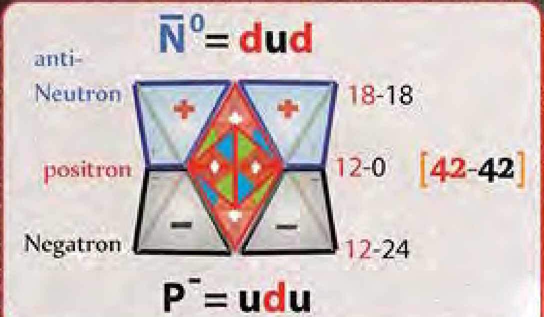
$\bar{N}^0$ = dud
anti-Neutron
18-18
positron
12-0
[42-42]
Negatron
12-24
P^- = udu

anti-Matter stars have
the opposite charge topology
to Matter stars

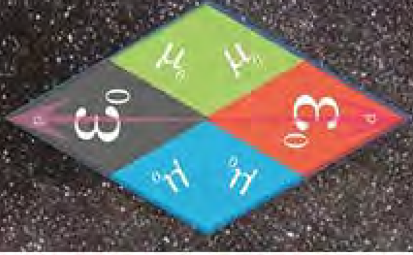
μ_0
μ_0
ω_0
ω_0
μ_0
μ_0

Neutron Stars

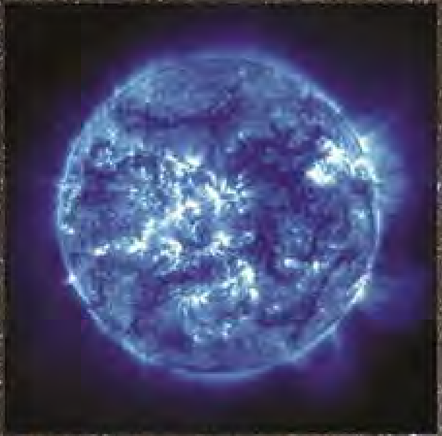

Colours used are illustrative of Matter types only

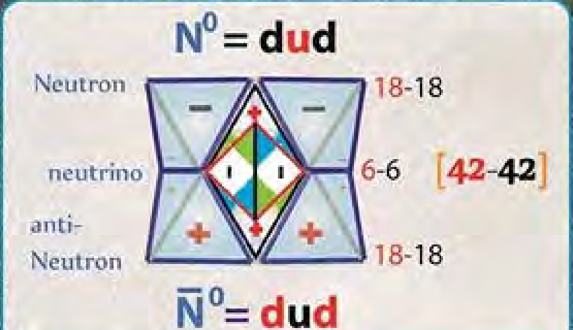
N^0 = dud
Neutron
18-18
neutrino
6-6
[42-42]
anti-Neutron
18-18
$\bar{N}^0$ = dud

All stars types emit Neutral Photons

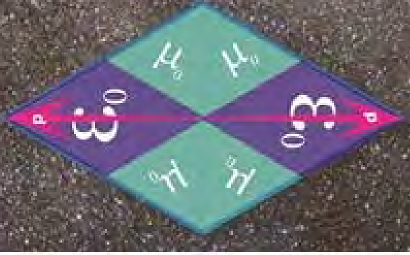
μ_0
μ_0
ω_0
ω_0
μ_0
μ_0

Matter Stars

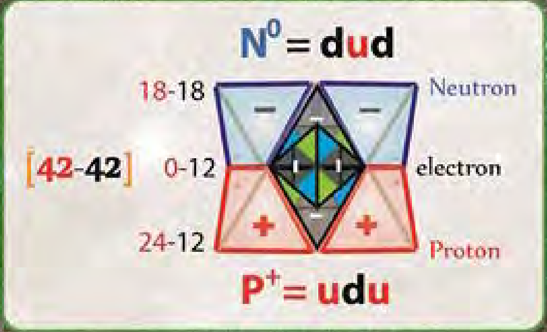
N^0 = dud
18-18
Neutron
[42-42]
0-12
electron
24-12
Proton
P^+ = udu

Matter stars have
the opposite charge topology
to anti-Matter stars

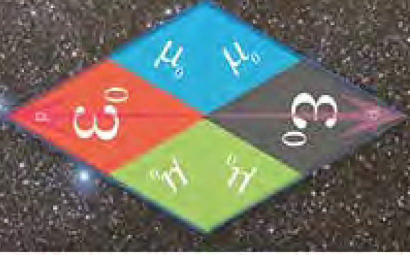
μ_0
μ_0
ω_0
ω_0
μ_0
μ_0

Star classifications

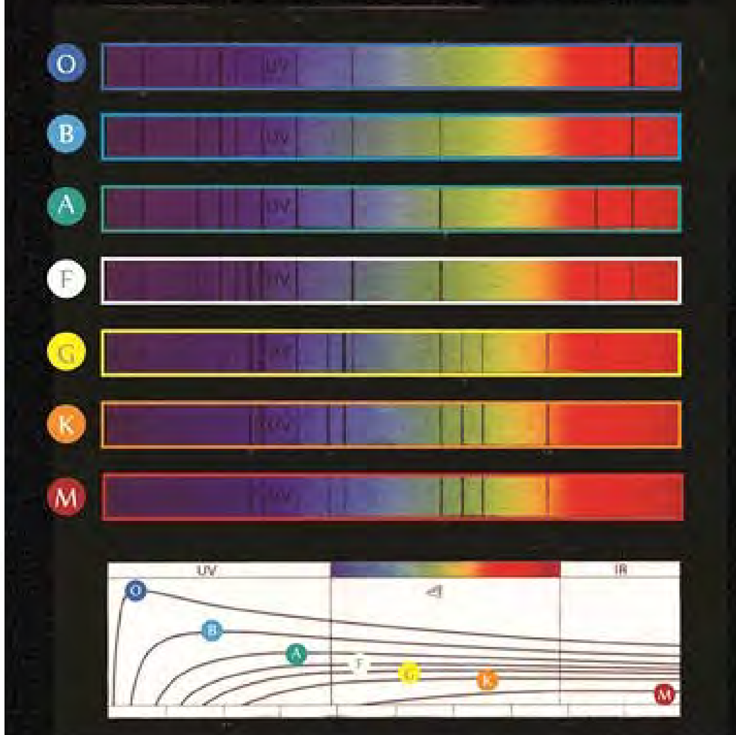

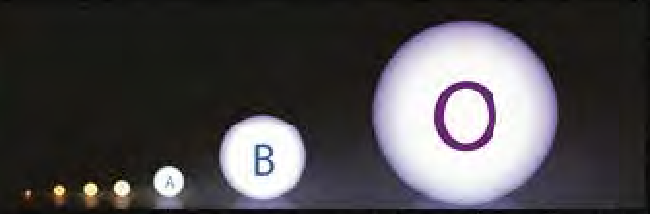

The greater the EM mass-energies radiated in a given volume of space the higher the Temperatures

White dwarves

Average Matter in Small volumes

The greater the voume for a given amount of radiated EM mass-energy the 'cooler' the Frequency of light produced from stellar processes

Blue giants

Greater mass-Matter in Larger volumes

O

≥ 16 M

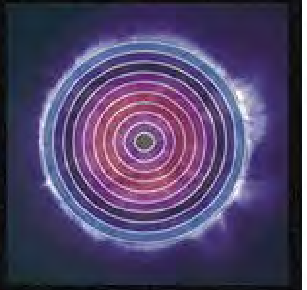

B

2.1–16 M

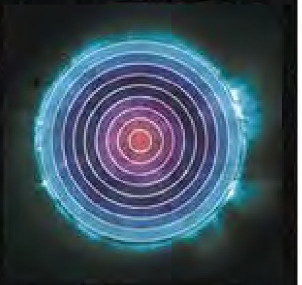

A

1.4–2.1 M

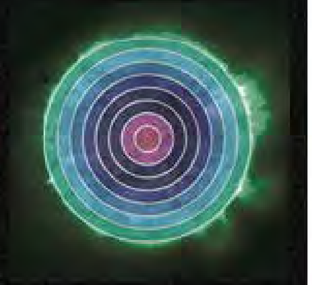

Red giants

Less mass-Matter in Larger volumes

M

≤ 0.45 M

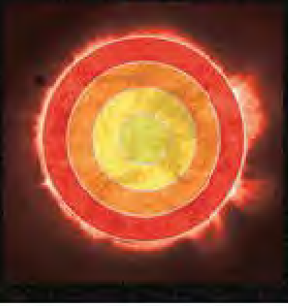

K

0.45–0.8 M

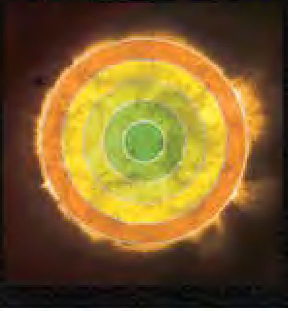

G

0.8–1.04 M

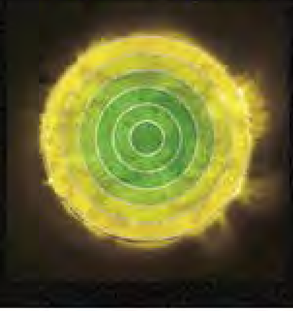

F

1.04–1.4 M

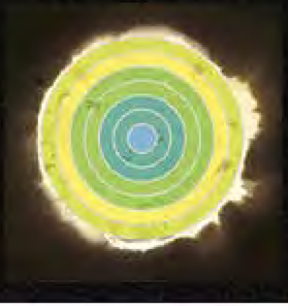

Main sequence

Average mass-Matter in medium volumes

Nucleosynthesis

The goal of nucleosynthesis is to understand the vastly differing abundances of the chemical elements and their several isotopes as being a result of natural history

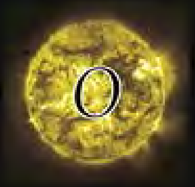

Nucleosynthesis is the process of creating new atomic nuclei from pre-existing nucleons (protons and neutrons)

All Stars have a dominant spectral colour and emit a multitude of spectral lines

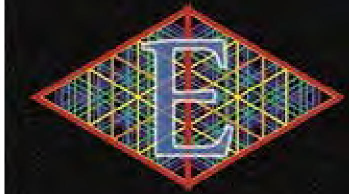

All Stars emit various colour spectra that combine to form spectral lines reflective of the star's energy output

The CNO cycle describing element creation inside Stars is erroneous ....

All spectral lines can be produced by the excitation of stellar nuclei by photons released by GEM pinch Matter-Energy conversion

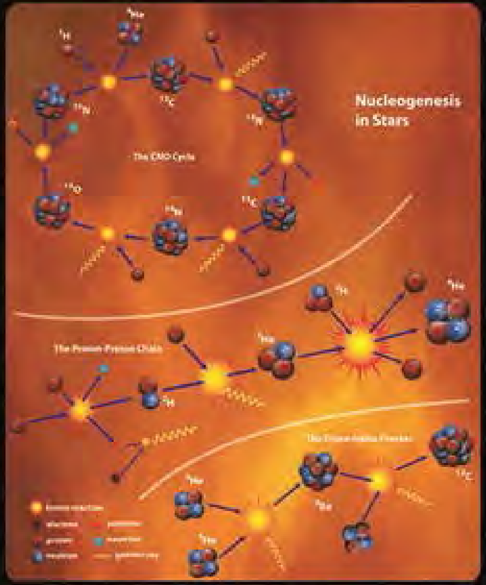

Elements in the Sun

Element	% of total atoms	% of total mass
Hydrogen	91.2	71.0
Helium	8.7	27.1
Oxygen	0.078	0.97
Carbon	0.043	0.40
Nitrogen	0.0088	0.096
Silicon	0.0045	0.099
Magnesium	0.0038	0.076
Neon	0.0035	0.058
Iron	0.030	0.014
Sulfur	0.015	0.040

In 1920, Arthur Eddington was the first to suggest that stars obtained their energy from nuclear fusion of hydrogen to form helium.

The P-P Chain describing the FUSION of elements to release energy is erroneous ....

All Matter-Energy conversion inside Stars is the result of ElectroMagnetic Pinches

Nucleosynthesis problems such as the abundances of observed elements & low mass element synthesis remain open problems for the current theory

Despite intense efforts to create suitable conditions, FUSION as described in nuclear theory, remains an ellusive source of Energy

The CNO cycle and P-P reactions are incorrect theorical models, developed to explain the SUN's GEM pinch dynamics and broad spectral line emissions

Copyright ABRAHAM [2008] - All rights reserved

Solar Arc lamp spectra

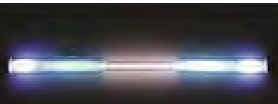

spectral lines are produced by accelerating electrons

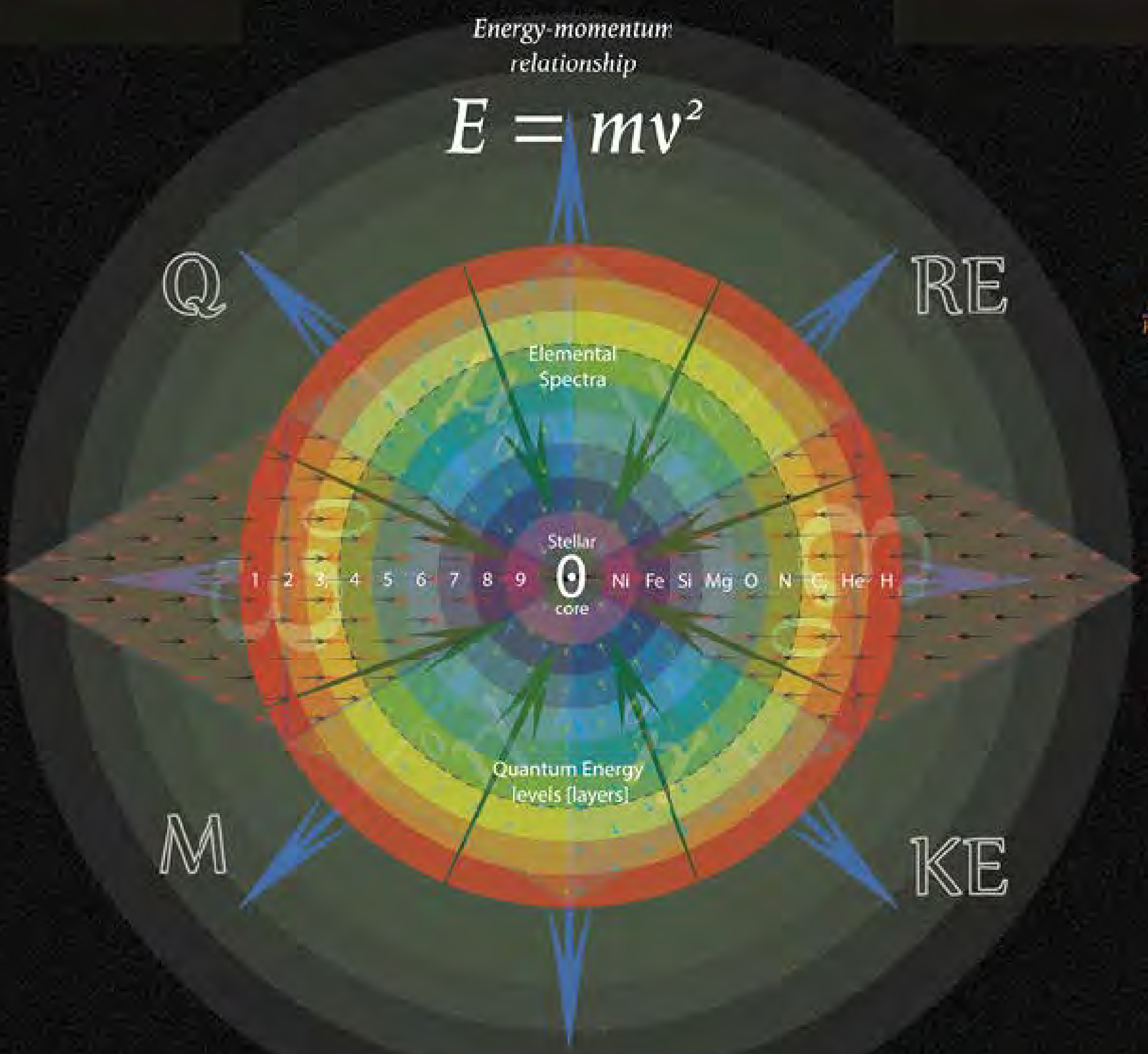

Energy-momentum relationship

$$E = mv^2$$

Q
RE
M
KE
Elemental Spectra
Stellar
0
core
1 2 3 4 5 6 7 8 9
Ni Fe Si Mg O N C He H
Quantum Energy levels [layers]

Spectral line transitions

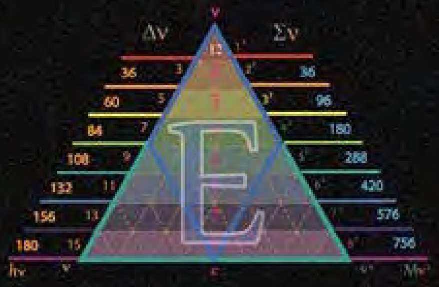

v
Δv
Σv
12
36
60
84
108
132
156
180
3
5
7
9
11
13
15
36
96
180
288
420
576
756
hv
v
c
Mv
E

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

Spectral lines

KEM	
48	*Lyman*
108	*Balmer*
192	*Paschen*
300	*Brackett*
432	*Pfund*
588	*Humphries*
768	*un-named*

classical physics and relativity united through equilateral geometry

Astrophysicists and astronomers use arc lamps as spectral line reference sources for many elemental emission and absorption lines

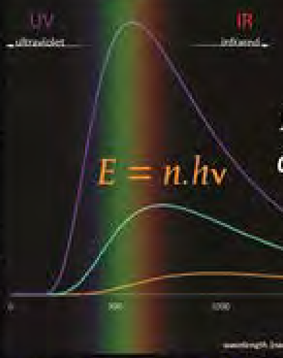

Tetryonic QED explains Planck's heat law along with all the observed spectral lines from first principles

$$\Delta hf = \Delta Mv = \Delta p$$

Planck's quantum radiation Law gives a energy distribution that peaks at a certain wavelength, Wien's Law explains the peak shifts to shorter wavelengths for higher temperatures, and the Stefan-Boltzman law explains the growth in the height [the area under] the curve with increasing temperatures

Stellar spectral lines

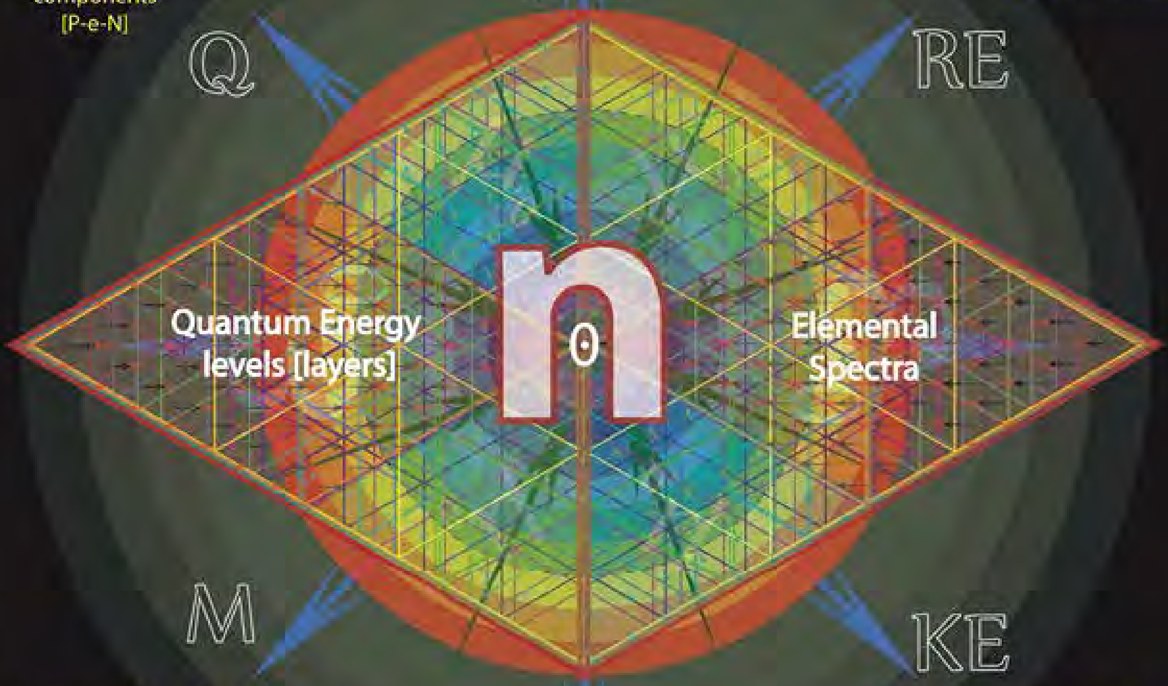
Q
RE
M
KE
n
0
Quantum Energy levels [layers]
Elemental Spectra

$$E = Mc^4$$

Energy-mass-Matter equivalence

Z#
All elements have equal number nuclei components
[P-e-N]

$\Delta mv = \Delta hf$
Excitation of photo-electrons in atomic nuclei produces spectral line emissions

Energies released from the collapse of mass-Matter topologies in stellar cores is radiated as spectral mass-energies
[GEM pinch]

mass-Energy equivalence

$$E = mc^2$$

The emissions from GEM pinch processes produce broad spectral lines that range from UV to IR rays
[& include white light]

The spectral lines of all elements can be mimicked by Deuterium nuclei with rasied energy levels undergoing energy transitions

Tetryonic Electrodynamics

$\Delta h\nu$
Quantum differentials

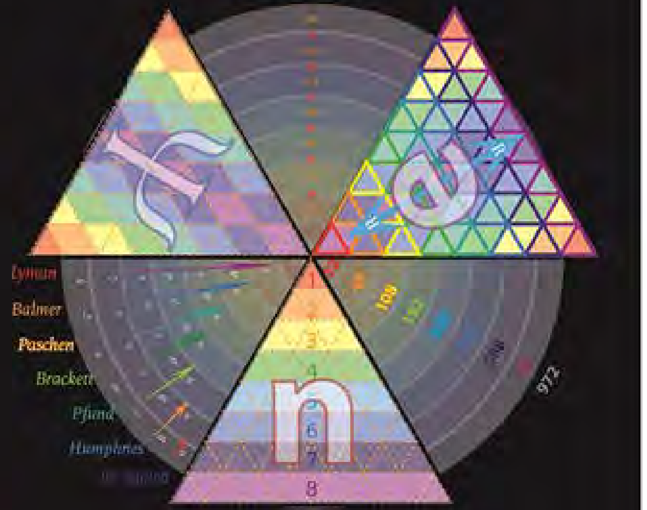
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
1
2
3
4
5
6
7
8
n
100
972

$$hf = \Delta Mv = \Delta p$$

spectral lines are produced by accelerating electrons

As energy is 'released' by the collapse of mass-Matter topologies in the stellar core they radiate outward raising the nuclear energies of the surrounding Matter which in turn radiate spectral series photons

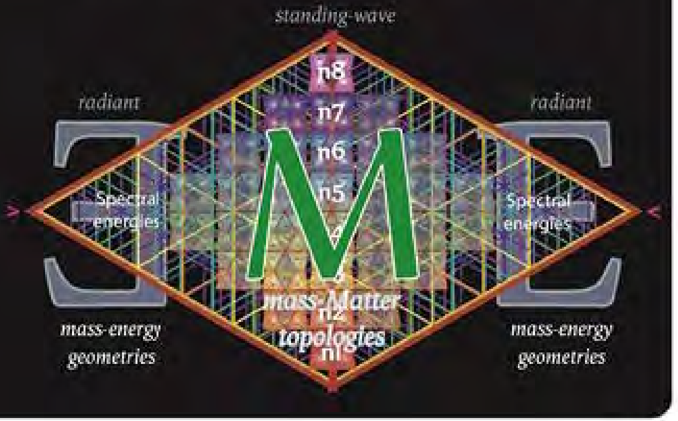
standing-wave
radiant
radiant
n8
n7
n6
n5
n2
n1
M
Spectral energies
Spectral energies
mass-Matter topologies
mass-energy geometries
mass-energy geometries

Copyright ABRAHAM [2008] - All rights reserved

Joseph von Fraunhofer

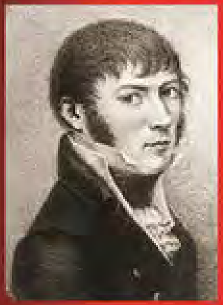

(6 March 1787 – 7 June 1826)

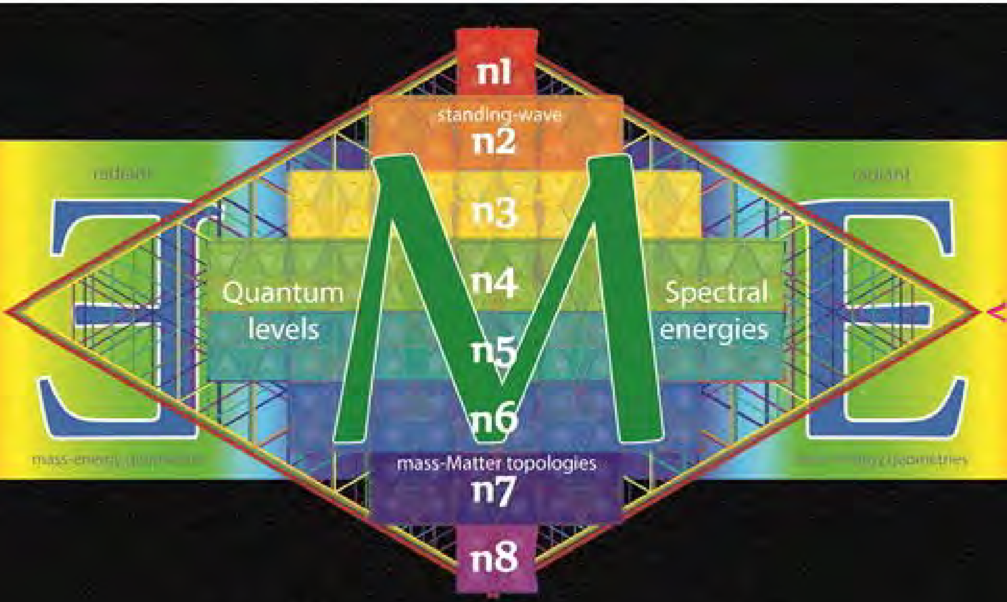
n1
standing-wave
n2
n3
n4
n5
n6
mass-Matter topologies
n7
n8
radiant
radiant
Quantum levels
Spectral energies
mass-energy geometries
mass-energy geometries
M
E
E

Annie Jump Cannon

(December 11, 1863 – April 13, 1941)

SPECTRAL SERIES

Lyman
Balmer
Paschen
Brackett
Pfund
Humphries

Stellar spectroscopy

Astronomical spectroscopy is the technique of spectroscopy used in astronomy.
The object of study is the spectrum of electromagnetic radiation, including visible light,
which radiates from stars and other celestial objects.

ATOMIC SHELLS

K
L
M
N
O
P

Spectral absorption lines are often referred to as Fraunhofer lines

QUANTUM LEVELS

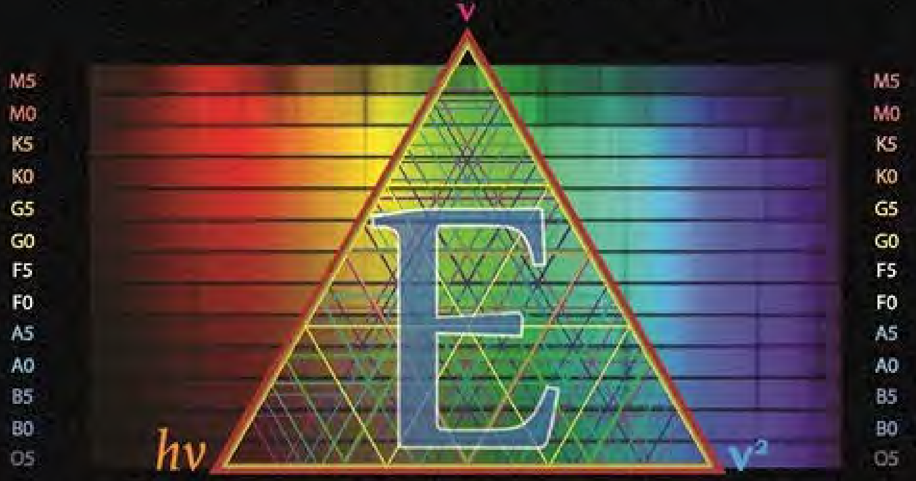
v
M5
M0
K5
K0
G5
G0
F5
F0
A5
A0
B5
B0
O5
hv
E
v²
M5
M0
K5
K0
G5
G0
F5
F0
A5
A0
B5
B0
O5

In 1901 Annie Cannon improved upon the 1897 work of Harvard associate Antonia Maury to produce the Harvard spectral class

STELLAR CLASSIFICATIONS

Spectroscopy can be used to derive many properties of distant stars and galaxies, such as their chemical composition, and motion through Doppler shift measurements.

Copyright ABRAHAM [2008] - All rights reserved

mass-ENERGY-Matter

Energy from the SUN

EM energies released by the collapse of standing-wave Matter topologies in the SUN radiate through space as mass-energy momenta and interact with Matter

All atoms are comprised of Deuterium nuclei

Radiant electromagnetic waves propagate

at the speed of light in a vacuum [c] taking 8.3 minutes to reach Earth

$$E = hf$$

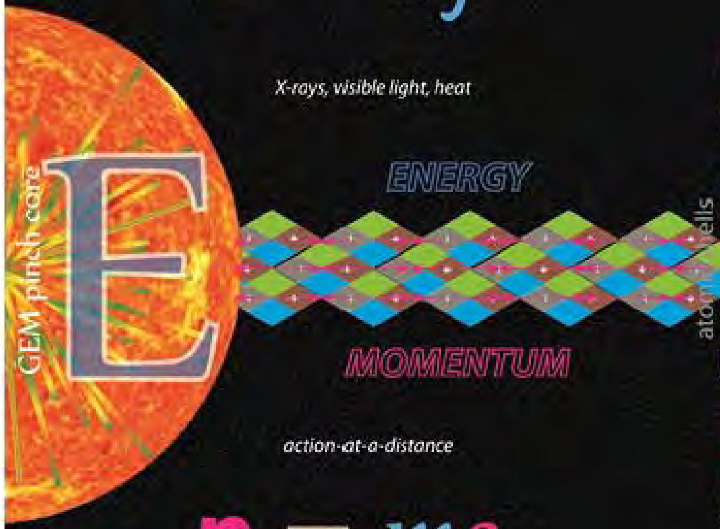
X-rays, visible light, heat
GEM pinch core
E
ENERGY
MOMENTUM
action-at-a-distance

$$p = mc$$

Square root linear momenta present in all longitudinal EM waveforms from rigid rods

and once established propagate impulses of momenta near-instantaneously along the longitudinal waves arriving 8.3 minutes before visible light

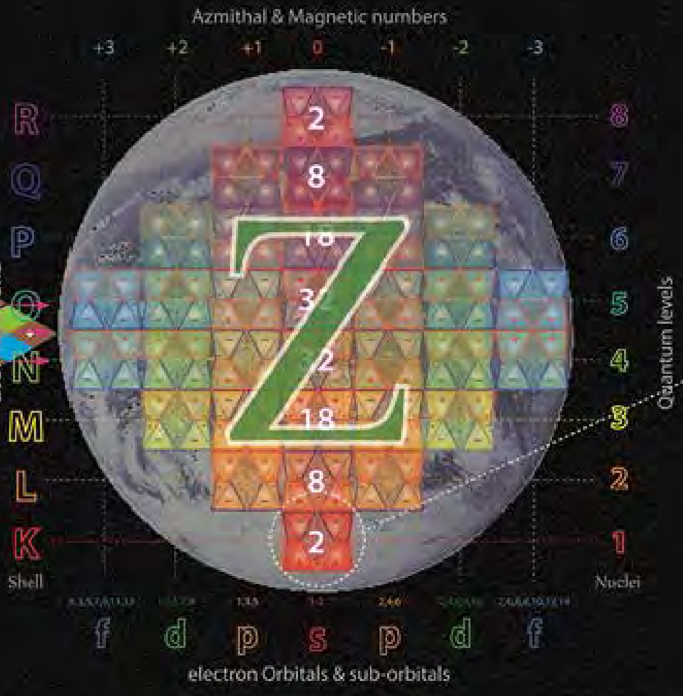
Azmithal & Magnetic numbers
+3
+2
+1
0
-1
-2
-3
R
Q
P
O
N
M
L
K
Shell
atomic shells
8
7
6
5
4
3
2
1
Nuclei
Quantum levels
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
electron Orbitals & sub-orbitals

The Earth receives a broad spectrum of energies from the SUN [with visible light falling into the Balmer spectral series]

$$hf = \underset{366.859\text{ nm}}{kEM} = \frac{hcR}{n^2}$$

By inductively coupling to the raised quantum KEM fields of the bound electrons their energies can be extracted and utilised

Deuterium nuclei are quantum synchronous convertors of electromagnetic energies

$$kEM = Mv^2$$

As energy momenta is recieved the energy levels of the Baryons increases in turn affecting and raising the KEM field energies of bound electrons

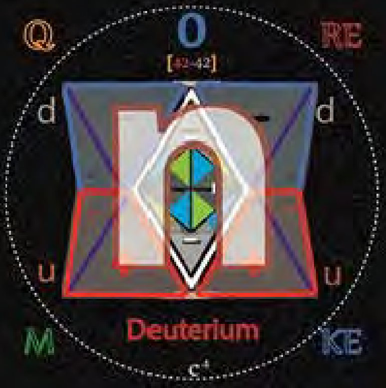
Q
O
RE
d
d
u
u
n
M
Deuterium
KE
c⁴

spectral lines, atomic transitions, photo-electric effect

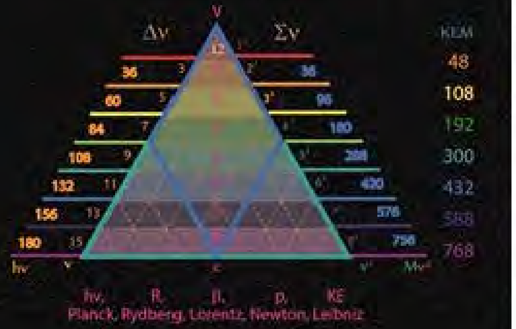
v
Δv
Σv
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
3
5
7
9
11
13
15
36
96
180
288
420
576
756
hv
v
c
v²
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

changing spectral emissions $\Delta hf = \Delta Mv = \Delta p$ accelerate photo-electrons

Copyright ABRAHAM [2008] - All rights reserved

Stellar Energies

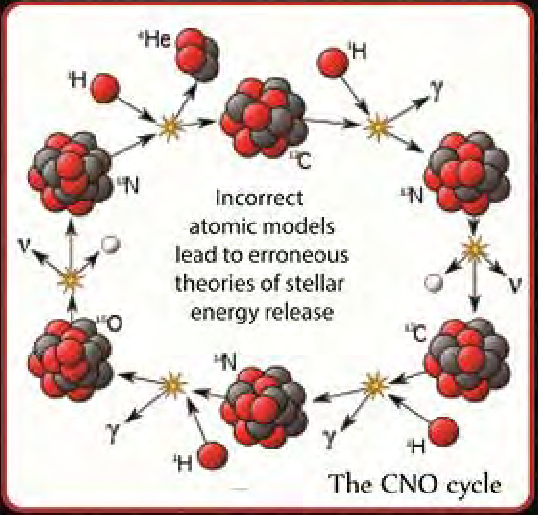
^{4}He
^{1}H
^{1}H
γ
^{12}C
^{15}N
^{13}N
ν
ν
^{15}O
^{13}C
^{14}N
γ
γ
^{1}H
^{1}H

Incorrect
atomic models
lead to erroneous
theories of stellar
energy release

The CNO cycle

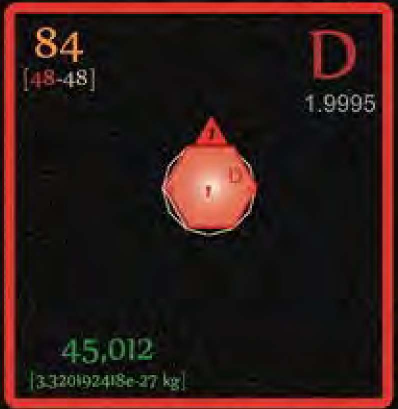
84
[48-48]
D
1.9995
45,012
[3.320192418e-27 kg]

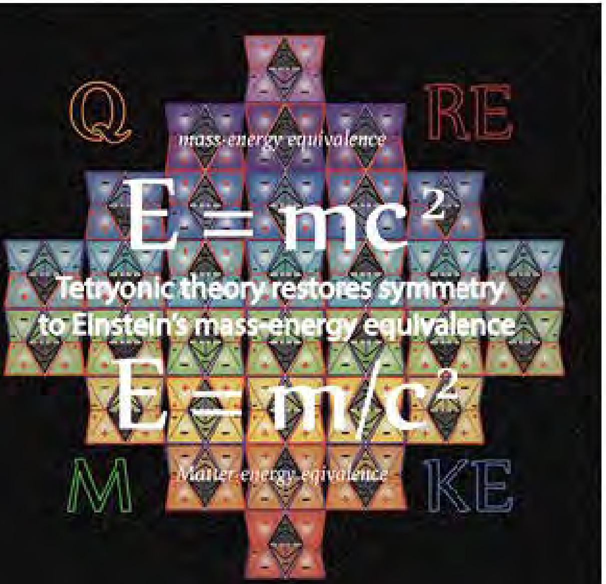

mass-energy equivalence

$$E = mc^2$$

Tetryonic theory restores symmetry to Einstein's mass-energy equivalence

$$E = m/c^2$$

Matter-energy equivalence

ENERGY

$$\underset{\text{mass}}{m}c^2 = E = \underset{\text{Matter}}{M}c^4$$

mass-ENERGY-Matter equivalence shows that a tiny amount of mass-Matter can be converted into a tremendous amount of radiant EM mass-energy

The energy released hypothesised by 'Fusion' is 1/200th that of the Tetryonic collapse of Matter into radiant EM mass-energies

All mass-ENERGY-Matter topologies are modelled in Tetryonic theory and Baryon decay is instigated by neutrinos

Neutrons are NOT Protons with an absorbed electron inside their charge topology

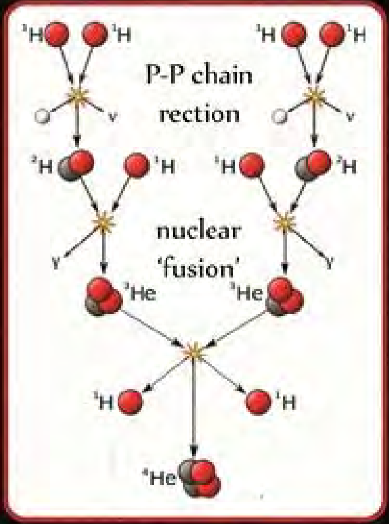
^{1}H
^{1}H
^{1}H
^{1}H
ν
ν
^{2}H
^{1}H
^{1}H
^{2}H
γ
γ
^{3}He
^{3}He
^{1}H
^{1}H
^{4}He

P-P chain rection

nuclear 'fusion'

The P-P fusion chain of Hydrogen->Helium is completely erroneous

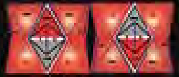

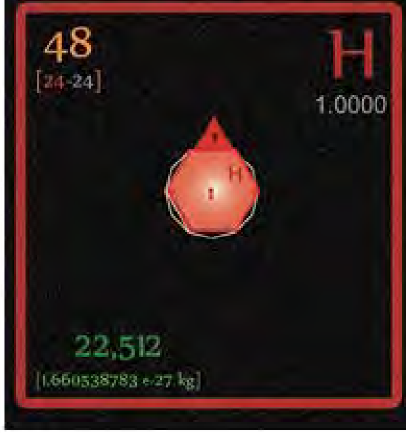
48
[24-24]
H
1.0000
22,512
[1.660538783 e-27 kg]

High pressure tunneling through the Coulomb barrier does NOT occur

The neutrino production of Proton-Proton chain reactions is 1/3 of what is detected

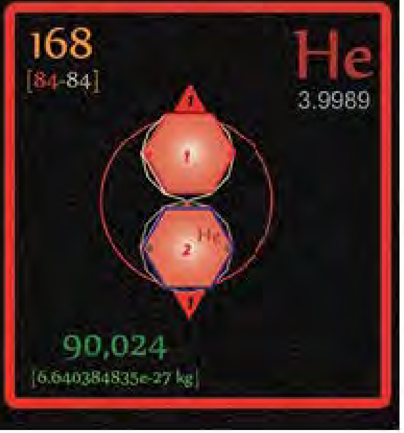
168
[84-84]
He
3.9989
90,024
[6.640384835e-27 kg]

The SUN's energy comes from Tetryonic Matter collapse NOT the fusion of atomic nuclei

Copyright ABRAHAM [2008] - All rights reserved

Stellar Nuclear Fusion

is the process of fusing of light atomic nuclei in order to facilitate energy releases

'COLD fusion'

can be realised through the fragmenting of large atomic nuclei thus releasing stored nuclear [K]EM waveform energies

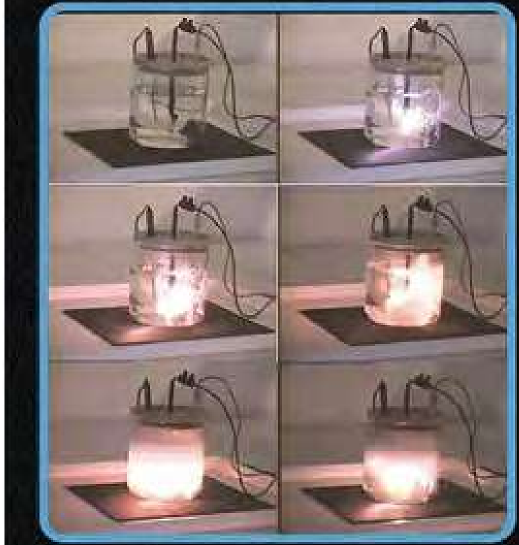

The much-maligned 'COLD fusion' of Palladium is in fact a form of atomic fission that releases the stored chemical [kEM] energies of atomic nuclei and is 12-13% efficient

Long sought after, but never realised, Fusion energies remain an ellusive source of clean limitless energy.

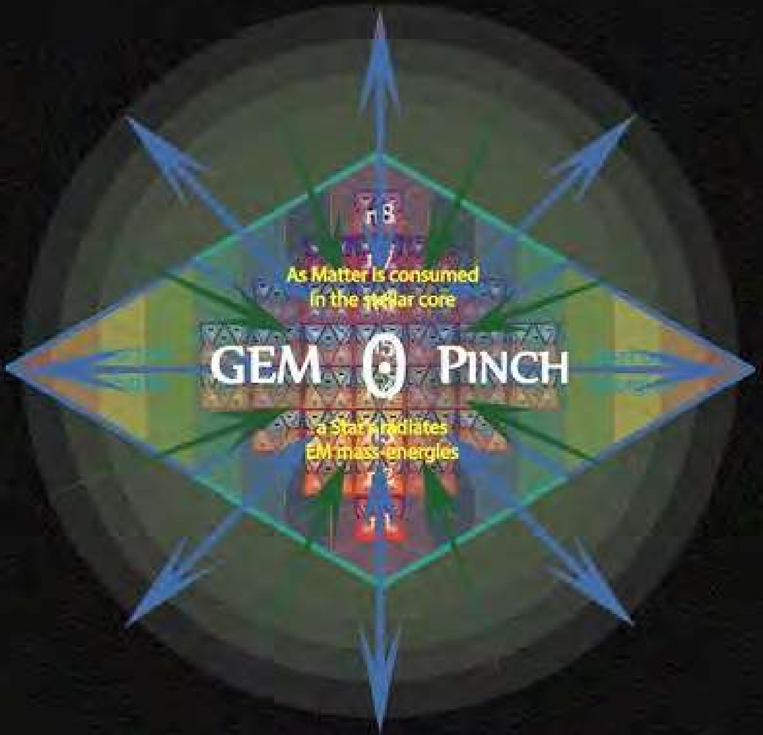
As Matter is consumed in the stellar core
GEM
PINCH
a Star radiates EM mass-energies

Tetryonics reveals the true mechanics behind the cold 'fusion' observed by Drs. Pons & Fleischmann and corrects the model of stellar processes to reveal new energy sources for the future

'HOT fusion'

is a fallacy, Stars release energy by collapsing standing-wave Matter into radiant EM mass-energy geometries in electromagnetic pinches

The hypothesised 'HOT fusion' of elements in stars is the Tetryonic conversion of mass-Matter within the star into radiant mass-energies requiring an electromagnetic pinch reaction but is 100% efficient

Tetryonics 73.10 - Fission, Fusion & EM pinches

Copyright ABRAHAM [2008] - All rights reserved

Where Fusion went Wrong

The conversion of standing-wave Matter into radiant mass-energy constitutes a sustainable, virtually unlimited environmentally compatible long-term source of energy that can meet all of Humanity's needs.

The toriodal design of Tokamak reactors does not allow for the formation of plasma pinches at their cores

Of the two designs only inertial confinement offered the best chance of creating the conditions required to effectively collapse Matter topologies in order to release the EM energies it contains.

However a far superior approach would be the electromagnetic pinching of Matter (as present in all stellar cores)

Inertial confinement fusion produces brief shockwaves, excessive heating and no sustained EM field

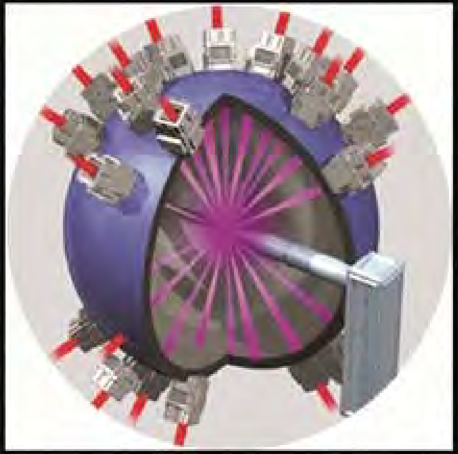

Tokamak design

Among a small number of approaches, the concept of toroidal magnetic confinement of fusion plasmas has achieved the most impressive scientific and technical progress towards energy release by thermonuclear burn of deuterium-tritium fuels.

The plasma consists of charged nuclei and electrons, which move in tight spirals around the lines of force of strong magnetic fields.

The basic problems in attaining useful nuclear fusion conditions are

(1) to heat the gas to these very high temperatures and

(2) to confine a sufficient quantity of the reacting nuclei for a long enough time to permit the release of more energy than is needed to heat and confine the gas.

A subsequent major problem is the capture of this energy and its conversion to electrical energy for work

Inertial confinement

In inertia confinement the fuel—tritium or deuterium is contained within a tiny glass sphere that is then bombarded on several sides by a pulsed laser or heavy ion beam.

This causes an implosion of the glass sphere, setting off a thermonuclear reaction that ignites the fuel but offer no self-regualation and relase of radiant mass-energies.

E_{in}

**In any useful fusion device,
the energy output must exceed the energy required
to confine and heat the plasma**

E_{out}

ElectroMagnetic Pinch technology

can serve as the gateway to realising new forms of clean, safe, efficient Energy production through the conversion of any Matter into various forms of EM radiation utilising electromagnetic pinch geometries developed from Tetryonic theory

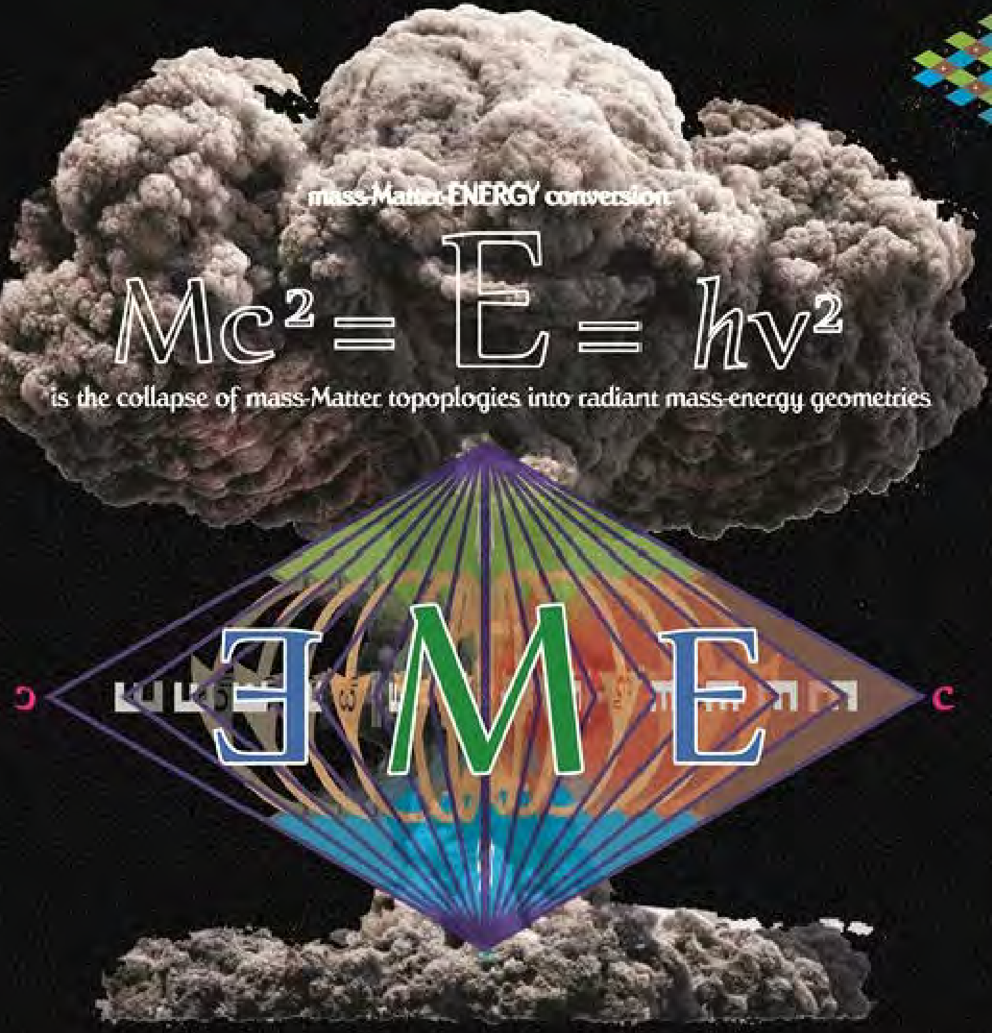

mass-Matter-ENERGY conversion

$$Mc^2 = E = hv^2$$

is the collapse of mass-Matter topoplogies into radiant mass-energy geometries

Charged Matter topologies are standing wave EM mass-energies

Radiant mass-energies form kEM wave geometries

An electromagnetic pulse (sometimes abbreviated EMP) is a burst of electromagnetic radiation that results from an explosion (usually from the detonation of a nuclear weapon) and/or a suddenly fluctuating magnetic field. The resulting rapidly changing electric fields or magnetic fields may couple with electrical/electronic systems to produce damaging current and voltage surges.

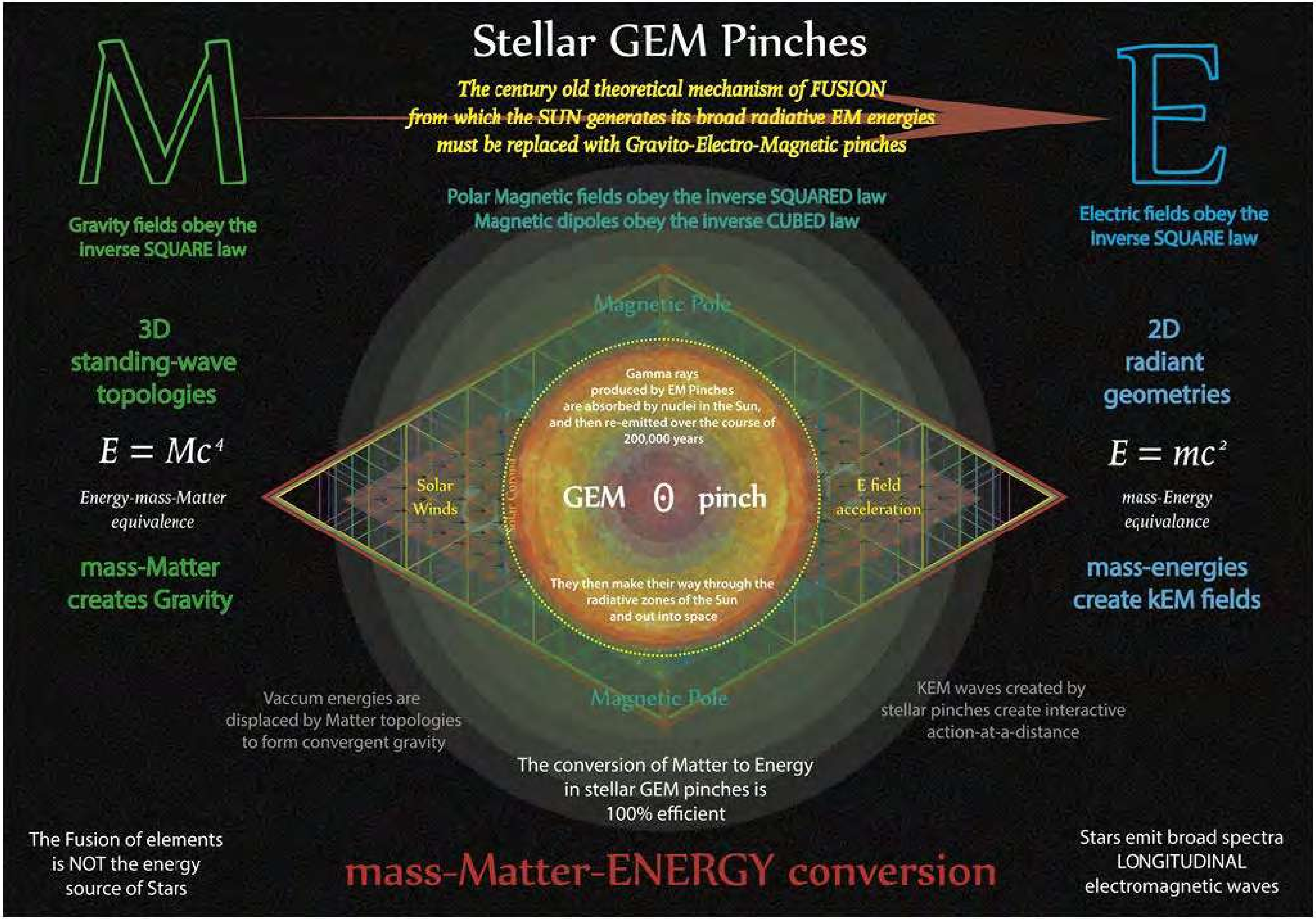

Stellar GEM Pinches

The century old theoretical mechanism of FUSION from which the SUN generates its broad radiative EM energies must be replaced with Gravito-Electro-Magnetic pinches

Gravity fields obey the inverse SQUARE law

Polar Magnetic fields obey the inverse SQUARED law
Magnetic dipoles obey the inverse CUBED law

Electric fields obey the inverse SQUARE law

3D standing-wave topologies

$$E = Mc^4$$

Energy-mass-Matter equivalence

mass-Matter creates Gravity

Gamma rays produced by EM Pinches are absorbed by nuclei in the Sun, and then re-emitted over the course of 200,000 years

They then make their way through the radiative zones of the Sun and out into space

2D radiant geometries

$$E = mc^2$$

mass-Energy equivalance

mass-energies create kEM fields

Vaccum energies are displaced by Matter topologies to form convergent gravity

KEM waves created by stellar pinches create interactive action-at-a-distance

The conversion of Matter to Energy in stellar GEM pinches is 100% efficient

The Fusion of elements is NOT the energy source of Stars

mass-Matter-ENERGY conversion

Stars emit broad spectra LONGITUDINAL electromagnetic waves

Tetryonics 73.13 - GEM Pinches

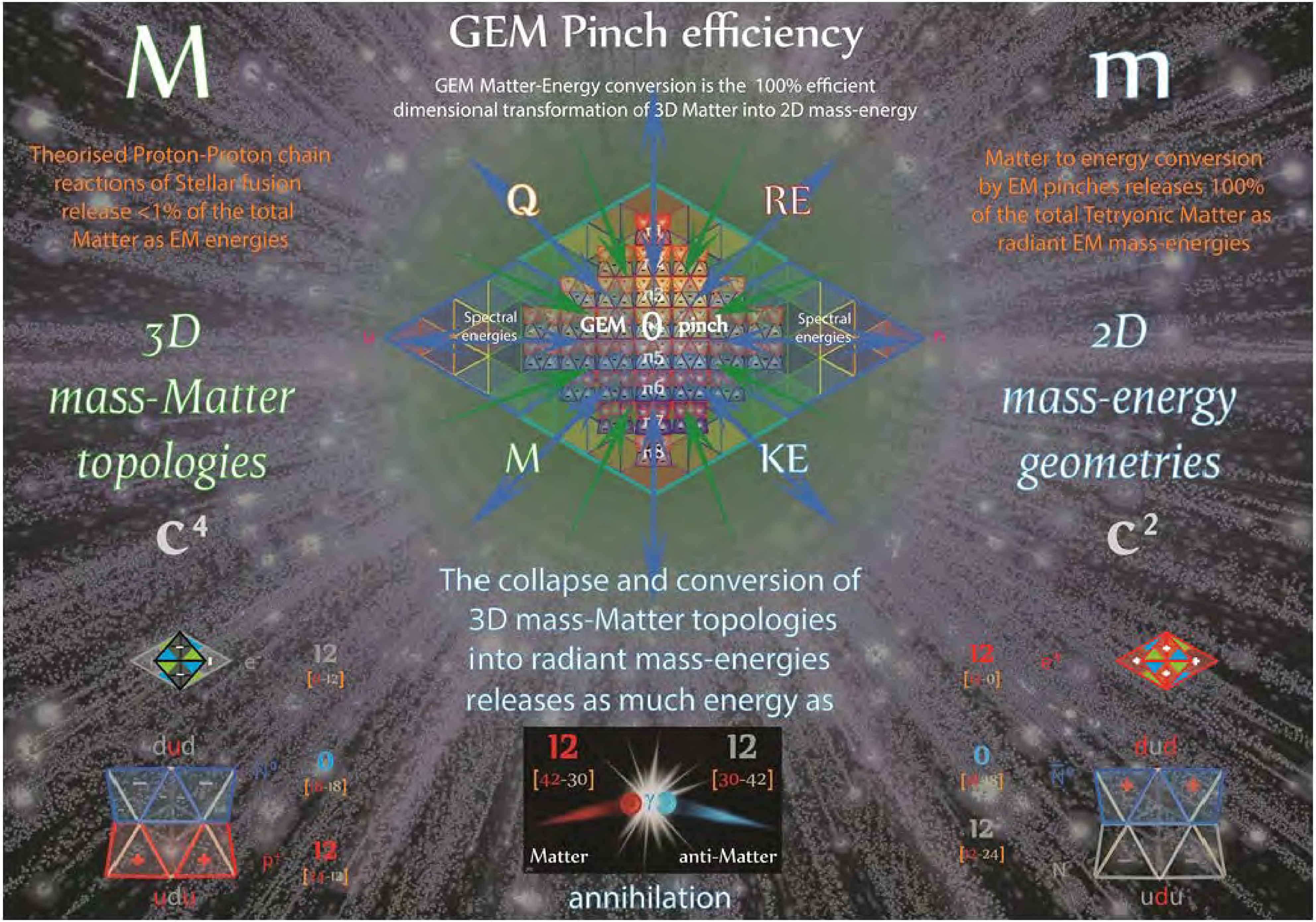

GEM Pinch efficiency

GEM Matter-Energy conversion is the 100% efficient dimensional transformation of 3D Matter into 2D mass-energy

M

Theorised Proton-Proton chain reactions of Stellar fusion release <1% of the total Matter as EM energies

m

Matter to energy conversion by EM pinches releases 100% of the total Tetryonic Matter as radiant EM mass-energies

3D mass-Matter topologies

c^4

2D mass-energy geometries

c^2

The collapse and conversion of 3D mass-Matter topologies into radiant mass-energies releases as much energy as

annihilation

Q
RE
M
KE
GEM 0 pinch
Spectral energies
Spectral energies
n1
n2
n3
n5
n6
n7
n8
u
n
e⁻
12
dud
N⁰
0
p⁺
12
udu
12
[42-30]
12
[30-42]
γ
Matter
anti-Matter
12
e⁺
0
N̄⁰
dud
12
N
udu

Tetryonics 73.14 - Matter-Energy collapse

Copyright ABRAHAM [2008] - All rights reserved

EM Pinch reactors

The technology required to create clean, limitless energy
and meet all of humanity's future energy requirements
has existed since the 1900's.

The basic steps to build a EM Pinch reactor are

(1) design and build a EM pinch field into which a fuel pellet can be injected

(2) the fuel can be any form of Matter that exhibits both Electric & Magnetic dipoles to facilitate the automatic positioning of the material fuel into the pinch core [the EM fields interact untill they equalise]

(3) enclose the EM Pinch core in a containment casing that can both capture and absorb the radiant EM energies produced by the collapse of Matter at the core and convert these energies into electricity

Th efficiency of the containment vessel along with energy storage and distribution can all be improved through the use of various Tetryonic field geometries to capture and re-circulate all EM energies produced

There are no moving parts to wear out or require extensive maintenance regimes

Photons

$$\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

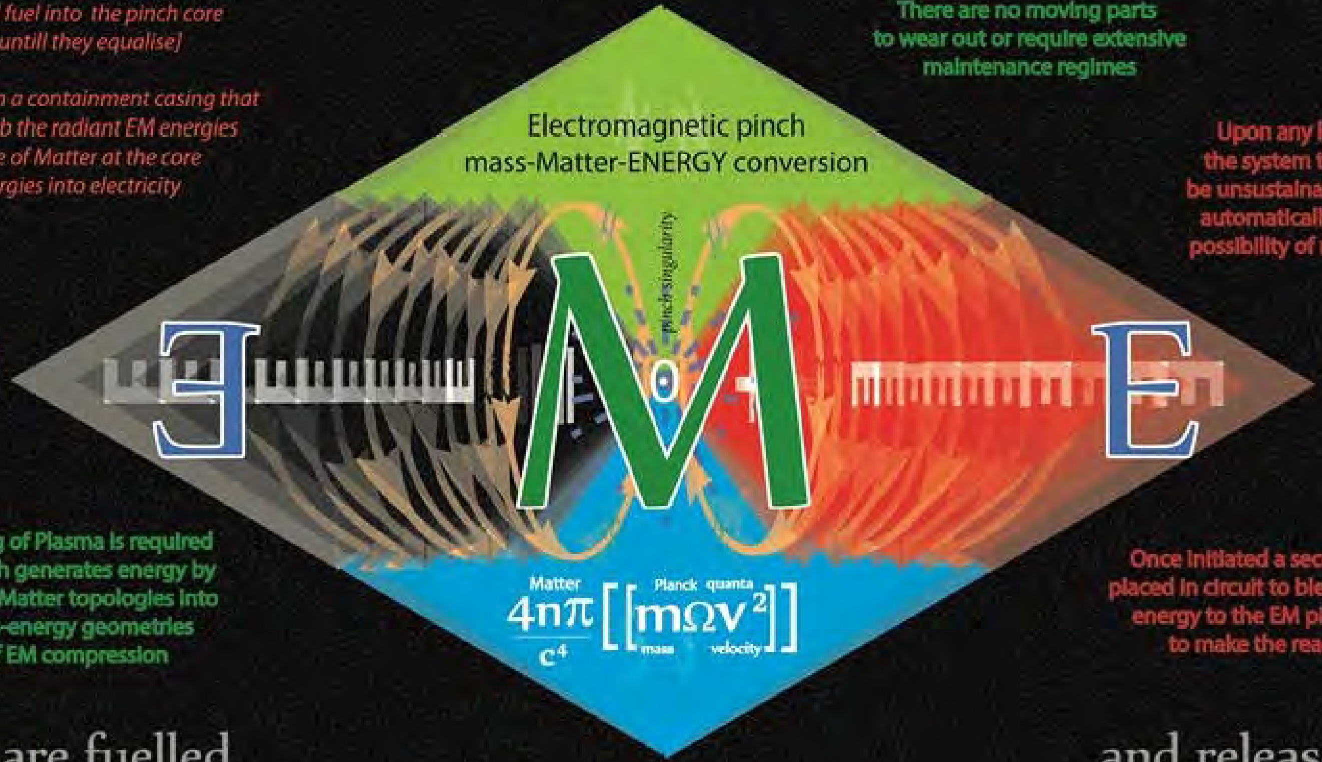

Electromagnetic pinch
mass-Matter-ENERGY conversion

Matter

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

EM waves

$$\frac{2n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

Upon any loss of Power to the system the EM pinch will be unsustainable and shutdown automatically preventing any possibility of runaway reactions

No preheating of Plasma is required as the EM pinch generates energy by collapsing' 3D Matter topologies into 2D EM mass-energy geometries by way of EM compression

Once initiated a secondary 'circuit' could be placed in circuit to bleed off and return suficient energy to the EM pinch field circuit in order to make the reaction self-sustaining

EM pinches are fuelled

by any form of Matter

The electromagnetic pinching of Matter (as utilised in Nature in all stellar cores) produces only EM radiation [light & heat] as a by-product - there are no harmful wastes - as the Matter is converted with 100% efficiency

and release mass-energy

as radiative KEM fields

Tetryonics 73.15 - The EM Pinch reactor

Copyright ABRAHAM [2008] - All rights reserved

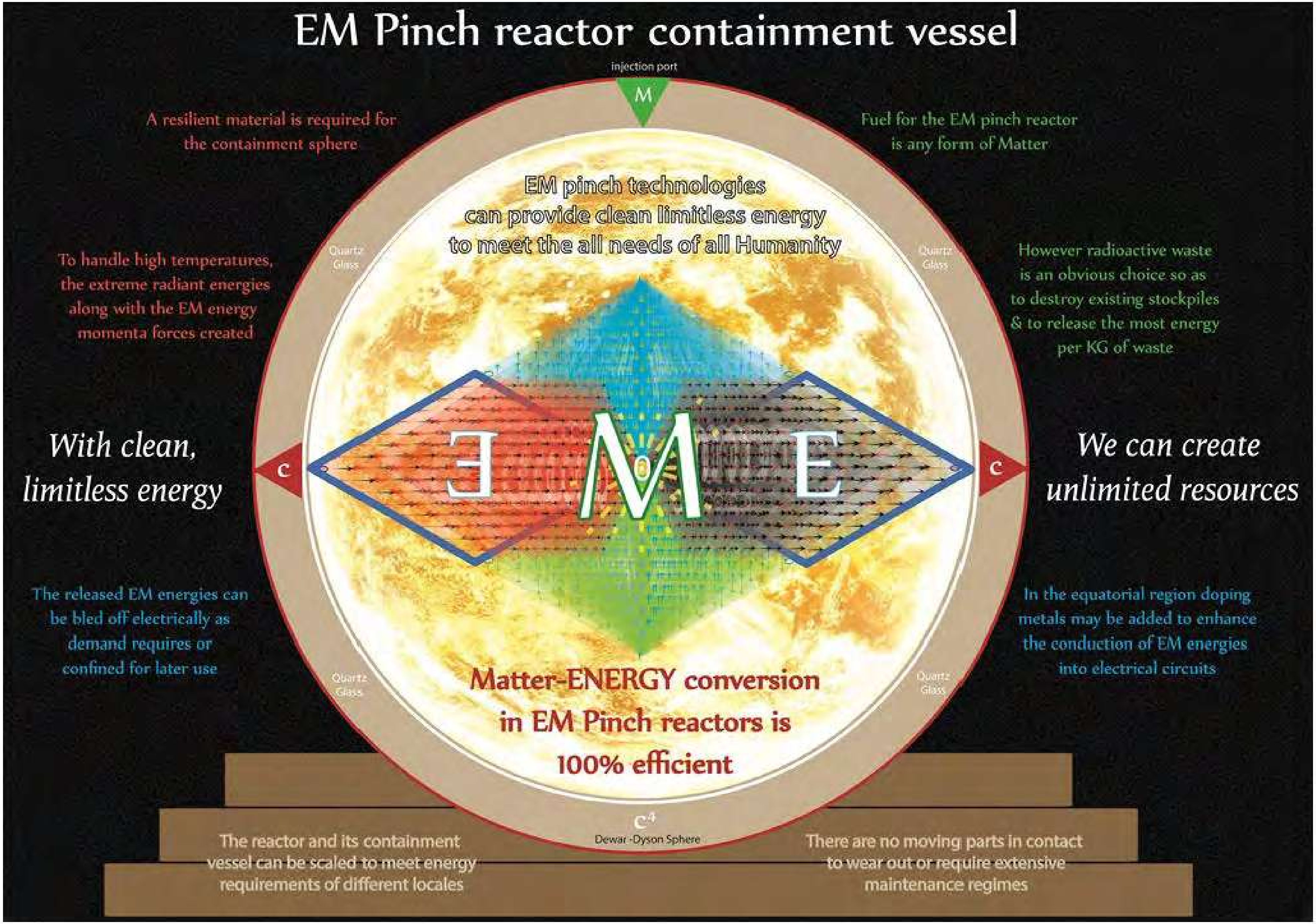

EM Pinch reactor containment vessel

A resilient material is required for the containment sphere

Fuel for the EM pinch reactor is any form of Matter

EM pinch technologies can provide clean limitless energy to meet the all needs of all Humanity

To handle high temperatures, the extreme radiant energies along with the EM energy momenta forces created

However radioactive waste is an obvious choice so as to destroy existing stockpiles & to release the most energy per KG of waste

With clean, limitless energy

We can create unlimited resources

The released EM energies can be bled off electrically as demand requires or confined for later use

In the equatorial region doping metals may be added to enhance the conduction of EM energies into electrical circuits

Matter-ENERGY conversion in EM Pinch reactors is 100% efficient

The reactor and its containment vessel can be scaled to meet energy requirements of different locales

There are no moving parts in contact to wear out or require extensive maintenance regimes

Tetryonics 73.16 - EM Pinch reactor containment

The Earth has finite resources and an ever increasing human population
puts pressure on its ability to supply resources to meet our needs

Unlimited material Resources

*Tetryonics offers a solution to human problems that
have plagued civilization from its beginnings
leading to wars, disease and famines*

With clean, limitless
energy sources

Unlimited resources
can be created

mass-energy geometries
cannot be created or destroyed

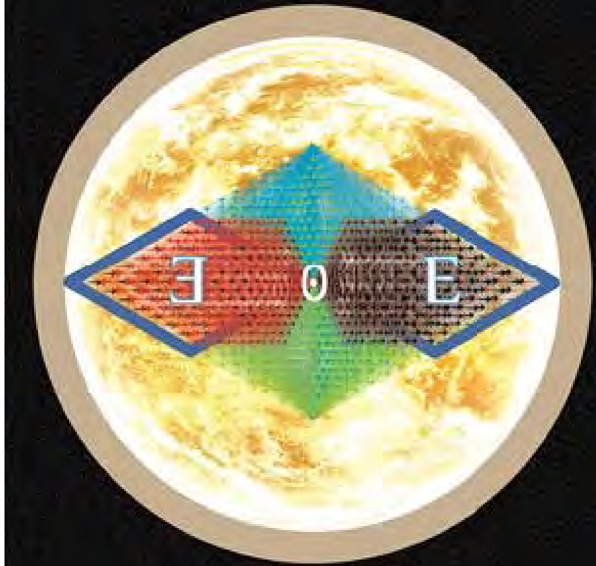

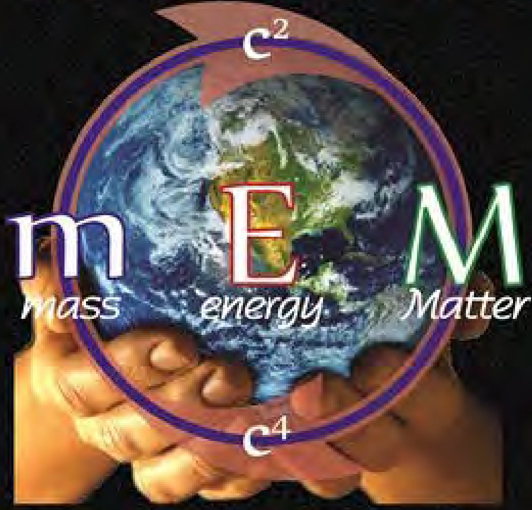

mass-Matter topologies
can be created and destroyed

Resources

Health

*With this new understanding of the charged mass-energy geometries of Matter
excess energies generated from the destruction of Matter topologies
can be reformed into any desired element or compound, and
all harmful waste can be removed from our environment*

Food

Water

Copyright ABRAHAM [2008] - All rights reserved

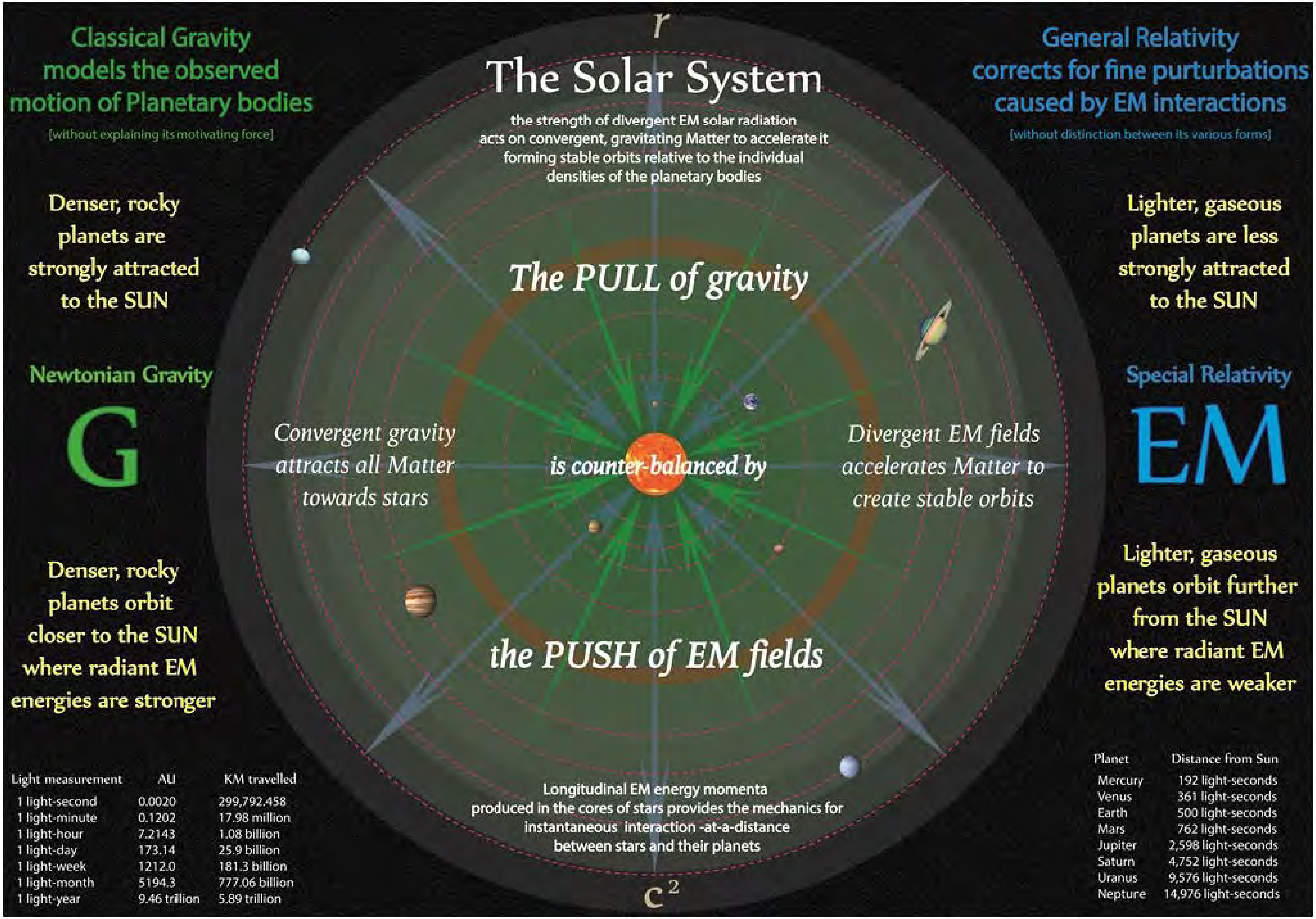

Classical Gravity models the observed motion of Planetary bodies

[without explaining its motivating force]

Denser, rocky planets are strongly attracted to the SUN

Newtonian Gravity

G

Denser, rocky planets orbit closer to the SUN where radiant EM energies are stronger

Light measurement	AU	KM travelled
1 light-second	0.0020	299,792.458
1 light-minute	0.1202	17.98 million
1 light-hour	7.2143	1.08 billion
1 light-day	173.14	25.9 billion
1 light-week	1212.0	181.3 billion
1 light-month	5194.3	777.06 billion
1 light-year	9.46 trillion	5.89 trillion

r

The Solar System

the strength of divergent EM solar radiation acts on convergent, gravitating Matter to accelerate it forming stable orbits relative to the individual densities of the planetary bodies

The PULL of gravity

Convergent gravity attracts all Matter towards stars

is counter-balanced by

Divergent EM fields accelerates Matter to create stable orbits

the PUSH of EM fields

Longitudinal EM energy momenta produced in the cores of stars provides the mechanics for instantaneous interaction -at-a-distance between stars and their planets

c^2

General Relativity corrects for fine purturbations caused by EM interactions

[without distinction between its various forms]

Lighter, gaseous planets are less strongly attracted to the SUN

Special Relativity

EM

Lighter, gaseous planets orbit further from the SUN where radiant EM energies are weaker

Planet	Distance from Sun
Mercury	192 light-seconds
Venus	361 light-seconds
Earth	500 light-seconds
Mars	762 light-seconds
Jupiter	2,598 light-seconds
Saturn	4,752 light-seconds
Uranus	9,576 light-seconds
Nepture	14,976 light-seconds

Tetryonics 74.01 - The Solar System

Copyright ABRAHAM [2008] - All rights reserved

Kepler's Laws

Kepler's three laws of planetary motion can be described as follows:

The path of the planets about the sun is elliptical in shape,
with the center of the sun being located at one focus.
(The Law of Ellipses)

An imaginary line drawn from the center of the sun to the center of
the planet will sweep out equal areas in equal intervals of time.
(The Law of Equal Areas)

The ratio of the squares of the periods of any two planets is equal to
the ratio of the cubes of their average distances from the sun.
(The Law of Harmonies)

$$T^2 = k\,a^3$$

The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

The divergent energy momenta of kEM fields create an acceleration FORCE on material bodies

The orbit of every planet is an ellipse
with the Sun at one of the two foci.

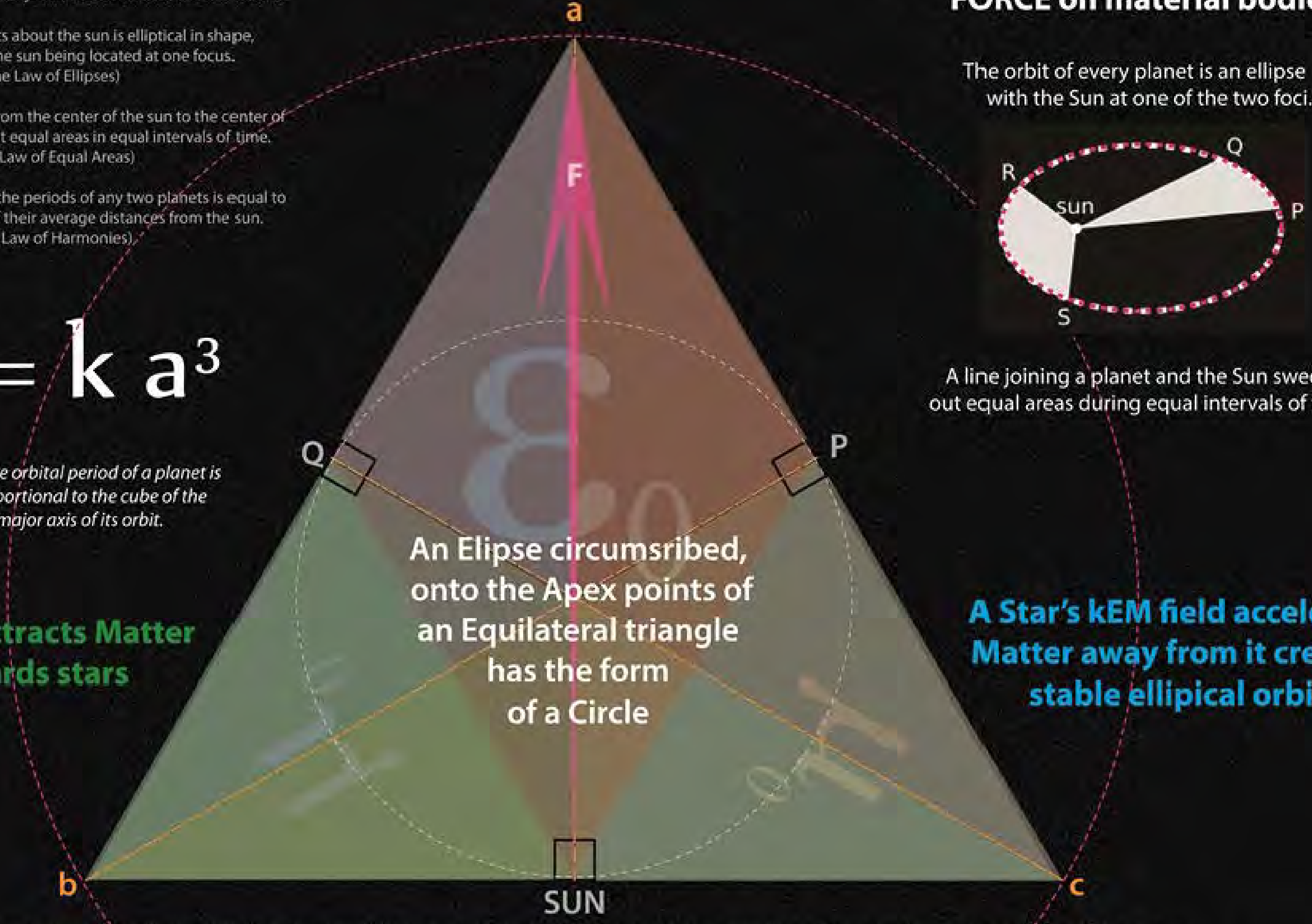

A line joining a planet and the Sun sweeps
out equal areas during equal intervals of time.

Gravity attracts Matter towards stars

An Elipse circumsribed,
onto the Apex points of
an Equilateral triangle
has the form
of a Circle

A Star's kEM field accelerates Matter away from it creating stable ellipical orbits

With an equilateral triangle, draw lines from the center of the circle to each vertex and each midpoint, creating six right triangles, as shown with six different colored triangles.
Each right triangle has a radius of the circle for one leg, and half of a side of the original triangle for another.
Any two right triangles sharing one of the radial lines must therefore be congruent, and that implies that the hypotenuses of the triangles are all equal.
This in turn shows that the six right triangles are all congruent, and so the large triangle they combine to form is equilateral

Copyright ABRAHAM [2008] - All rights reserved

The Major celestial bodies

All planetary orbits are the result of a balance between gravitation Matter and interactive kEM fields

The sun is a rotating charged mass-Matter object, and it creates a solar system-wide kEM field.

The solar mass is large enough to produce a EM field with lines of Force reaching the orbit of Pluto.

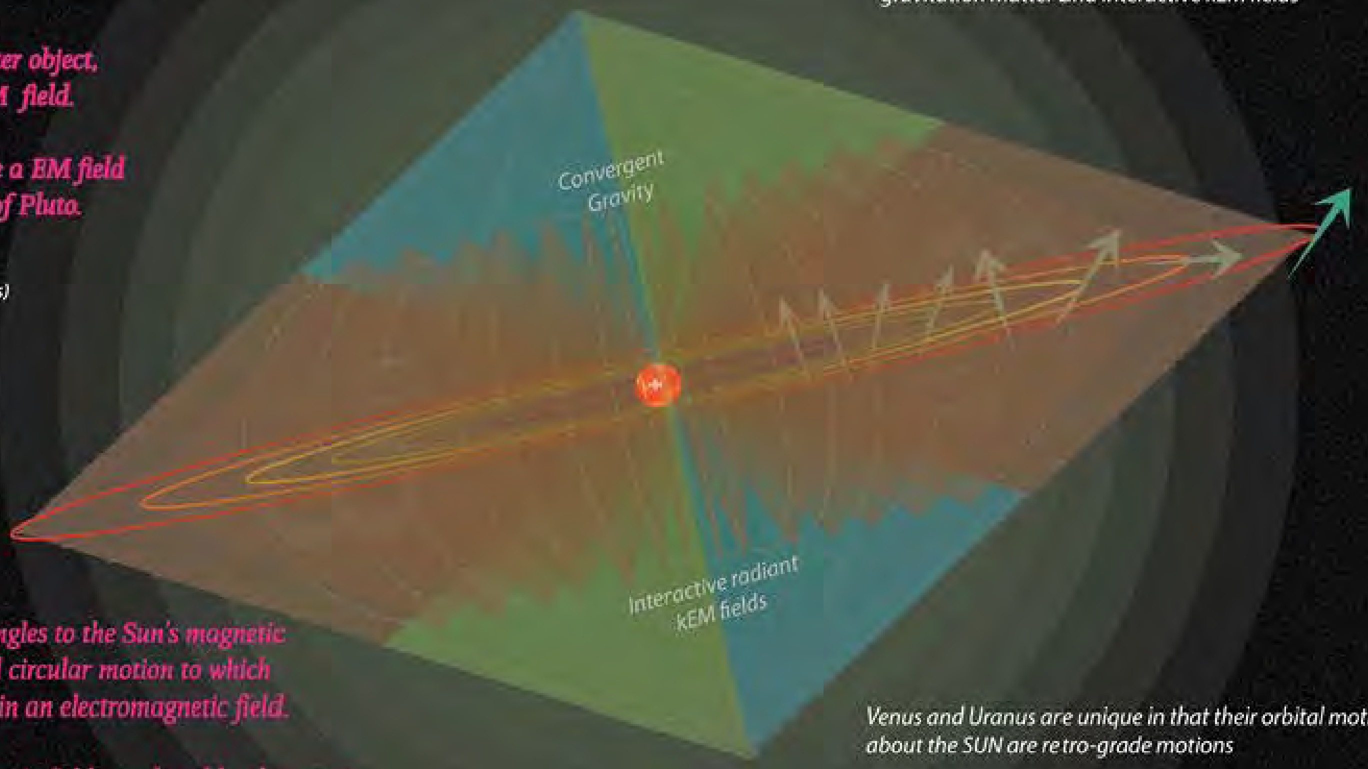

Planet	*Object Axial tilt (°)*	*Axial tilt (radians)*
Mercury	~0.01	0.00
Venus	177.4	3.10
Earth	23.44	0.41
Moon	1.5424	0.03
Mars	25.19	0.44
Jupiter	3.13	0.06
Saturn	26.73	0.47
Uranus	97.77	1.71
Neptune	28.32	0.49
Pluto	119.61	2.09

The charged planets move at right angles to the Sun's magnetic lines of force and describe the usual circular motion to which moving charged bodies are subject to in an electromagnetic field.

Venus and Uranus are unique in that their orbital motions about the SUN are retro-grade motions

Satellites, in turn, revolve in the smaller magnetic fields produced by the rotation of the planets. Mercury and Venus have no satellites as they produce extremely weak electromagnetic fields in addition to their Gravitational fields.

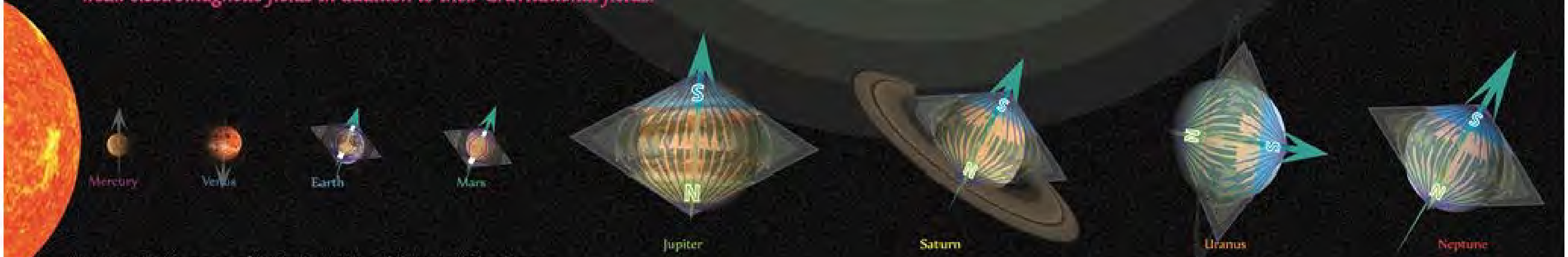

"In comparing the masses of the Earth, Jupiter and Saturn to their volumes, one remarks that the densities of these planets are, to some degree, inverse to their mean distances from the Sun. This rule is not true for Venus and Uranus"

Urbain Le Verrier

Tetryonics 74.03 - The Major celestial bodies

Solar orbit Inclinations

Inclination to Sun's equator (°)

Rocky Planets	
Mercury	3.38
Venus	3.86
Earth	7.15
Mars	5.65

Gas giants	
Jupiter	6.09
Saturn	5.51
Uranus	6.48
Neptune	6.43

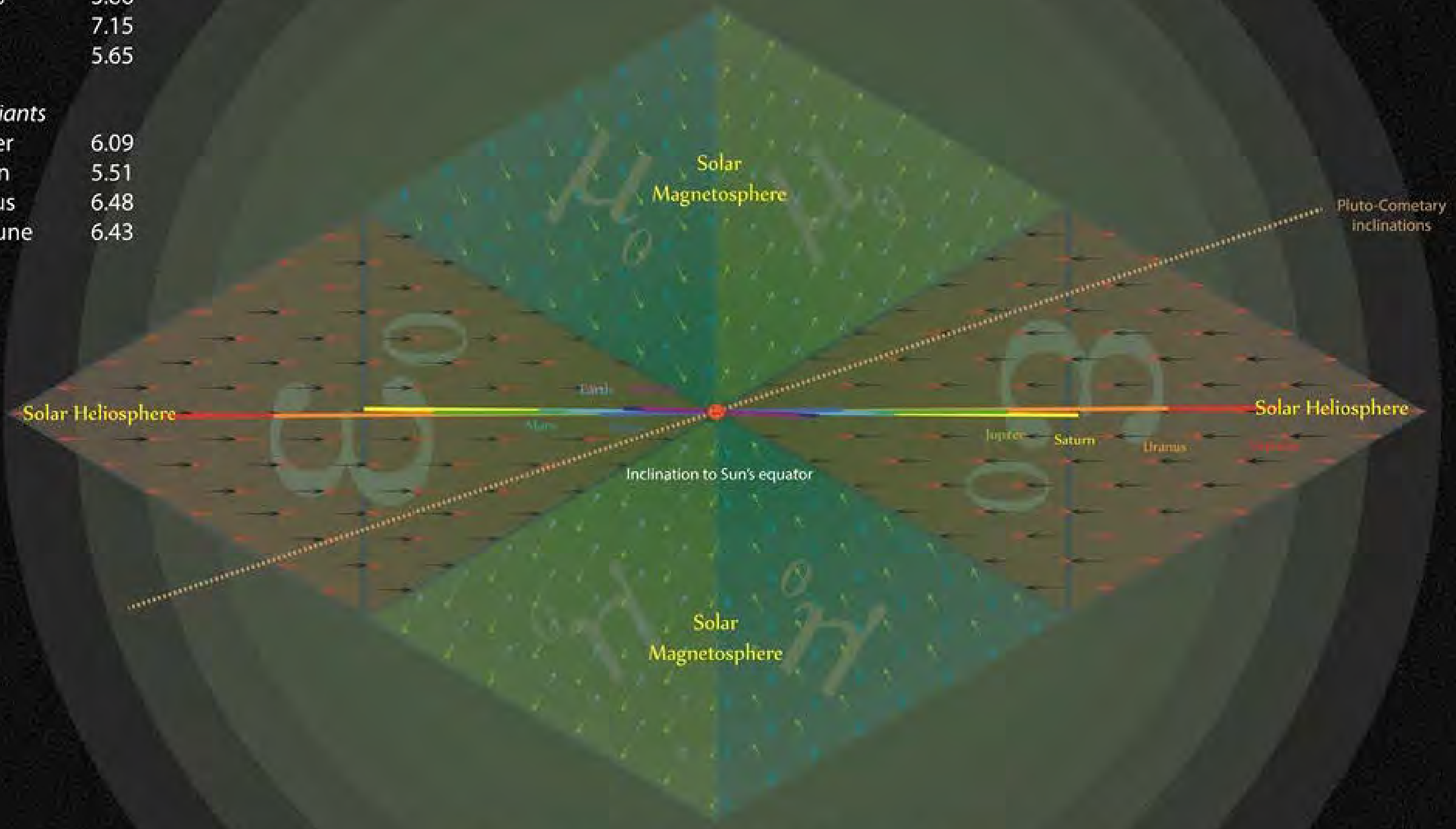
Solar Magnetosphere
Pluto-Cometary inclinations
Earth
Solar Heliosphere
Mars
Solar Heliosphere
Jupiter
Saturn
Uranus
Inclination to Sun's equator
Solar Magnetosphere

Orbital inclination is one of the six orbital parameters describing the shape and orientation of a celestial orbit.

It is the angular distance of the orbital plane from the plane of reference (usually the primary's equator or the ecliptic), normally stated in degrees.

Gravity is a strictly convergent Force

← ← ↔

GR = G + EM

Einstein Newton Lorentz

General Relativity models Newtonian Gravity with SR interactions

The Lagrange Points

are the five positions in an orbital configuration where a small object affected only by gravity can be stationary relative to two larger material bodies of mass-Matter

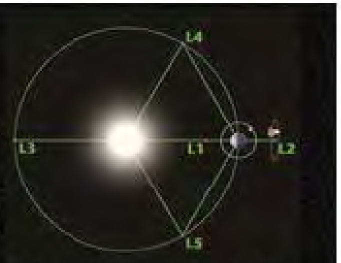

GEM forces between two bodies can be modelled with super-positioned equilateral energy geometries

Gravitational forces from the two massive bodies are in the same ratio as the material masses of the two bodies

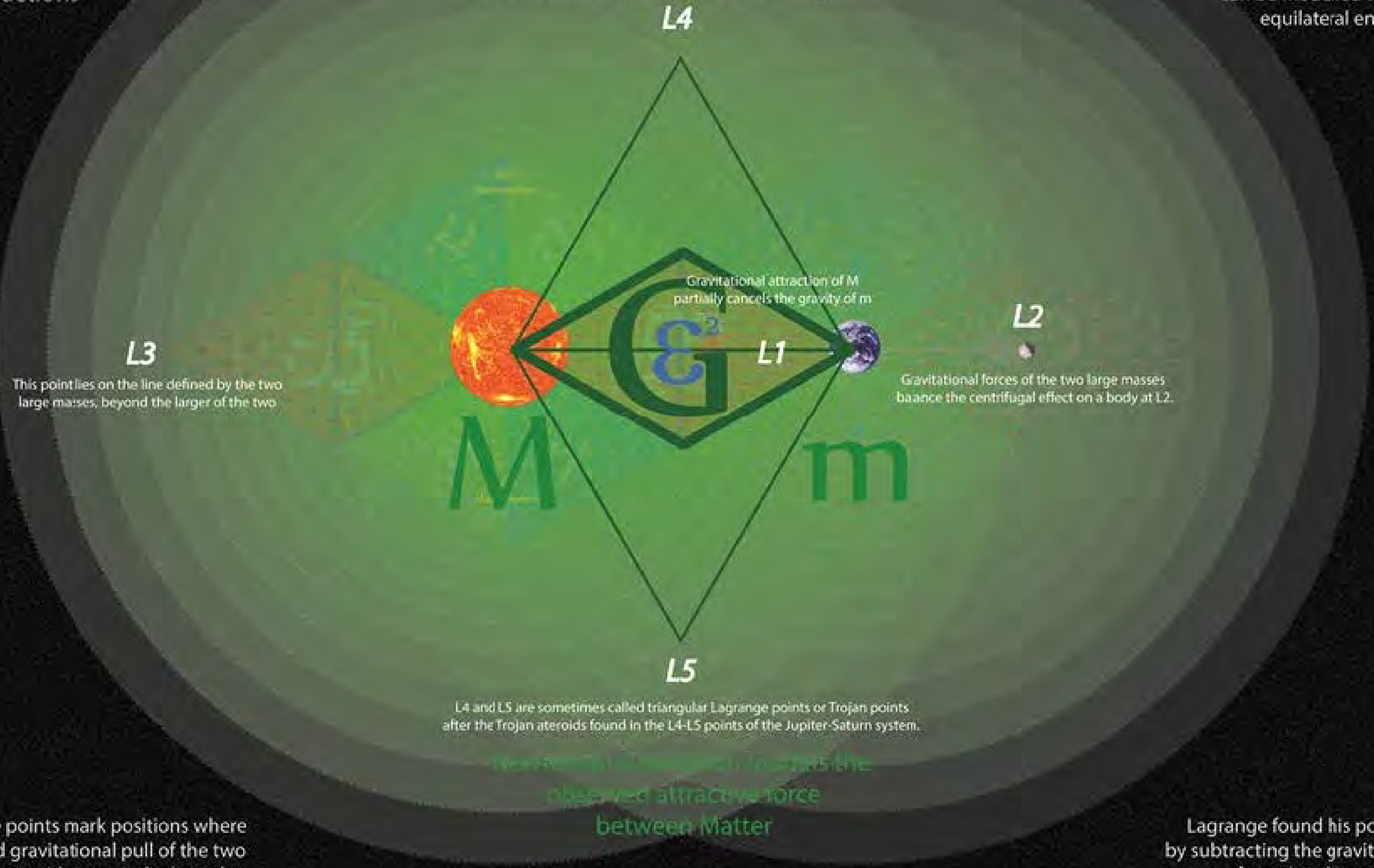

Gravitational attraction of M partially cancels the gravity of m

Gravitational forces of the two large masses baance the centrifugal effect on a body at L2.

This pointlies on the line defined by the two large masses, beyond the larger of the two

L4 and L5 are sometimes called triangular Lagrange points or Trojan points after the Trojan ateroids found in the L4-L5 points of the Jupiter-Saturn system.

[illegible] the observed attractive force between Matter

The Lagrange points mark positions where the combined gravitational pull of the two large masses provides precisely the right centripetal force required to rotate with them

Lagrange found his points of stabilty by subtracting the gravitational potential energy from the object's kinetic energy.

Tetryonics 74.05 - Lagrangian Points

Copyright ABRAHAM [2008] - All rights reserved

Earth-Moon Gravito-Electro-Magnetic fields

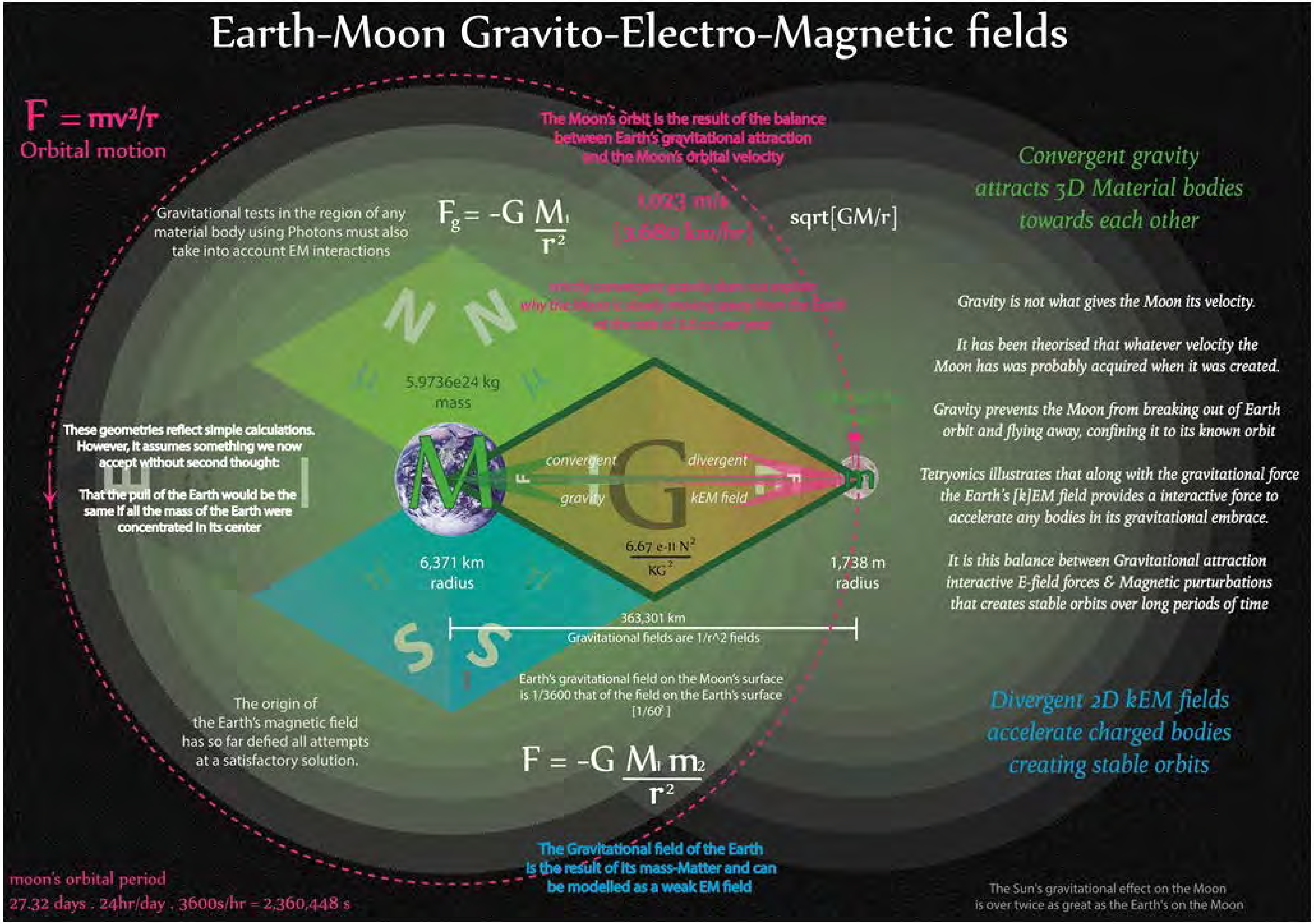

$F = mv^2/r$

Orbital motion

Gravitational tests in the region of any material body using Photons must also take into account EM interactions

$F_g = -G \frac{M_1}{r^2}$

The Moon's orbit is the result of the balance between Earth's gravitational attraction and the Moon's orbital velocity

1,023 m/s
[3,680 km/hr]

sqrt[GM/r]

strictly convergent gravity does not explain why the Moon is slowly moving away from the Earth at the rate of 3.8 cm per year

Convergent gravity attracts 3D Material bodies towards each other

Gravity is not what gives the Moon its velocity.

It has been theorised that whatever velocity the Moon has was probably acquired when it was created.

Gravity prevents the Moon from breaking out of Earth orbit and flying away, confining it to its known orbit

Tetryonics illustrates that along with the gravitational force the Earth's [k]EM field provides a interactive force to accelerate any bodies in its gravitational embrace.

It is this balance between Gravitational attraction interactive E-field forces & Magnetic perturbations that creates stable orbits over long periods of time

5.9736e24 kg
mass

These geometries reflect simple calculations. However, it assumes something we now accept without second thought:

That the pull of the Earth would be the same if all the mass of the Earth were concentrated in its center

6,371 km
radius

$6.67 \text{e-}11 \frac{N^2}{KG^2}$

1,738 m
radius

363,301 km
Gravitational fields are 1/r^2 fields

Earth's gravitational field on the Moon's surface is 1/3600 that of the field on the Earth's surface [$1/60^2$]

The origin of the Earth's magnetic field has so far defied all attempts at a satisfactory solution.

$F = -G \frac{M_1 m_2}{r^2}$

Divergent 2D kEM fields accelerate charged bodies creating stable orbits

The Gravitational field of the Earth is the result of its mass-Matter and can be modelled as a weak EM field

moon's orbital period
27.32 days . 24hr/day . 3600s/hr = 2,360,448 s

The Sun's gravitational effect on the Moon is over twice as great as the Earth's on the Moon

Tetryonics 74.06 - Earth-Moon EM interaction field

The Earth's ElectroMagnetic field

Earth's magnetic field (also known as the geomagnetic field) is the magnetic field that extends from the Earth's inner core to where it meets the solar wind,

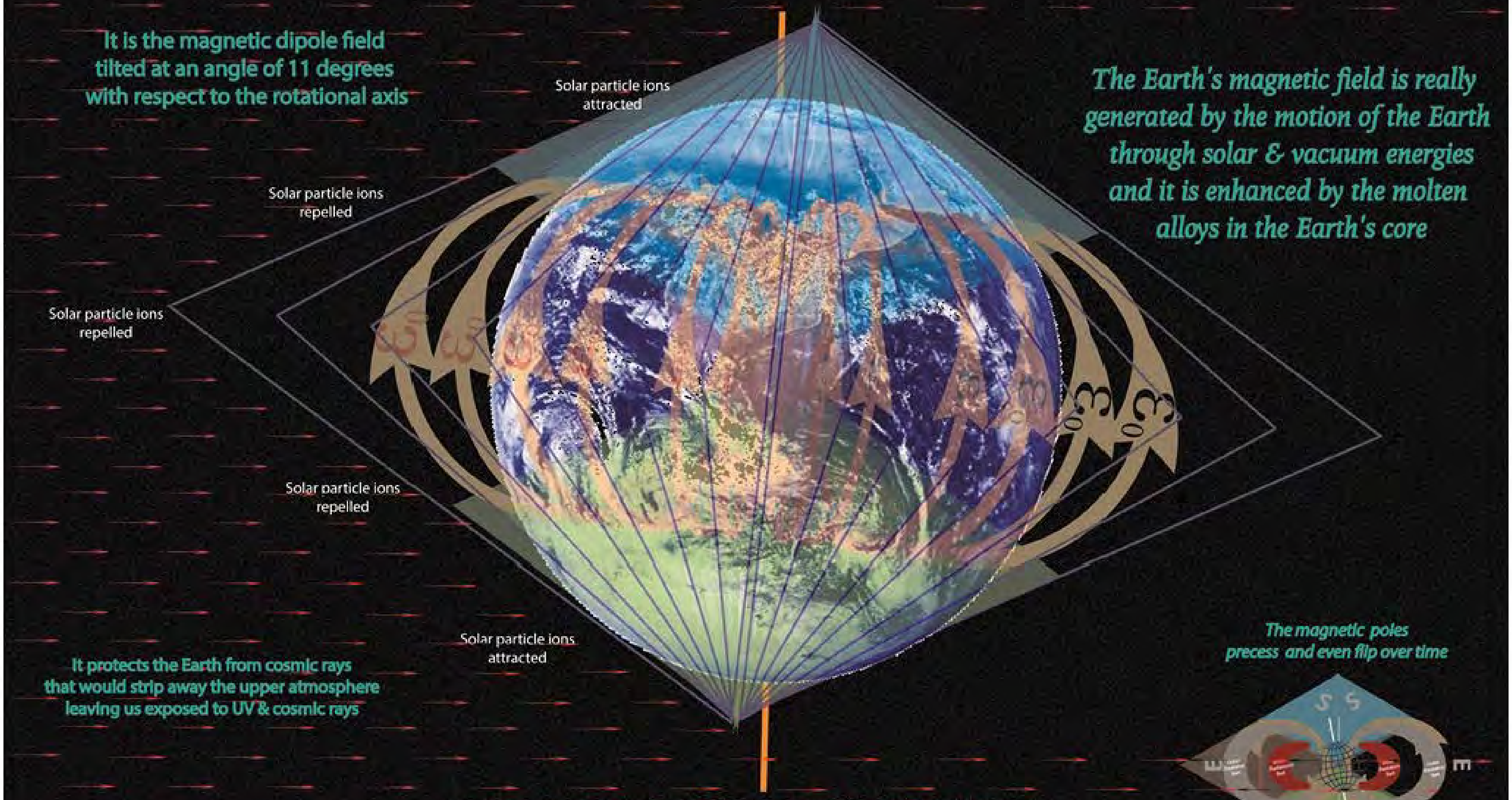
Solar particle ions attracted
Solar particle ions repelled
Solar particle ions repelled
Solar particle ions repelled
Solar particle ions attracted
S S
N N
E
E

It is the magnetic dipole field tilted at an angle of 11 degrees with respect to the rotational axis

The Earth's magnetic field is really generated by the motion of the Earth through solar & vacuum energies and it is enhanced by the molten alloys in the Earth's core

It protects the Earth from cosmic rays that would strip away the upper atmosphere leaving us exposed to UV & cosmic rays

The magnetic poles precess and even flip over time

At random intervals (several hundred thousand years) the Earth's field reverses (the north and south geomagnetic poles change places with each other)

Tetryonics 74.07 - The Earth's Magnetic field

The Earth's molten core

The Earth has an outer silicate solid crust, a highly viscous mantle, a liquid outer core that is much less viscous than the mantle, and a solid inner core. Scientific understanding of Earth's internal structure is based on observations of topography and bathymetry.

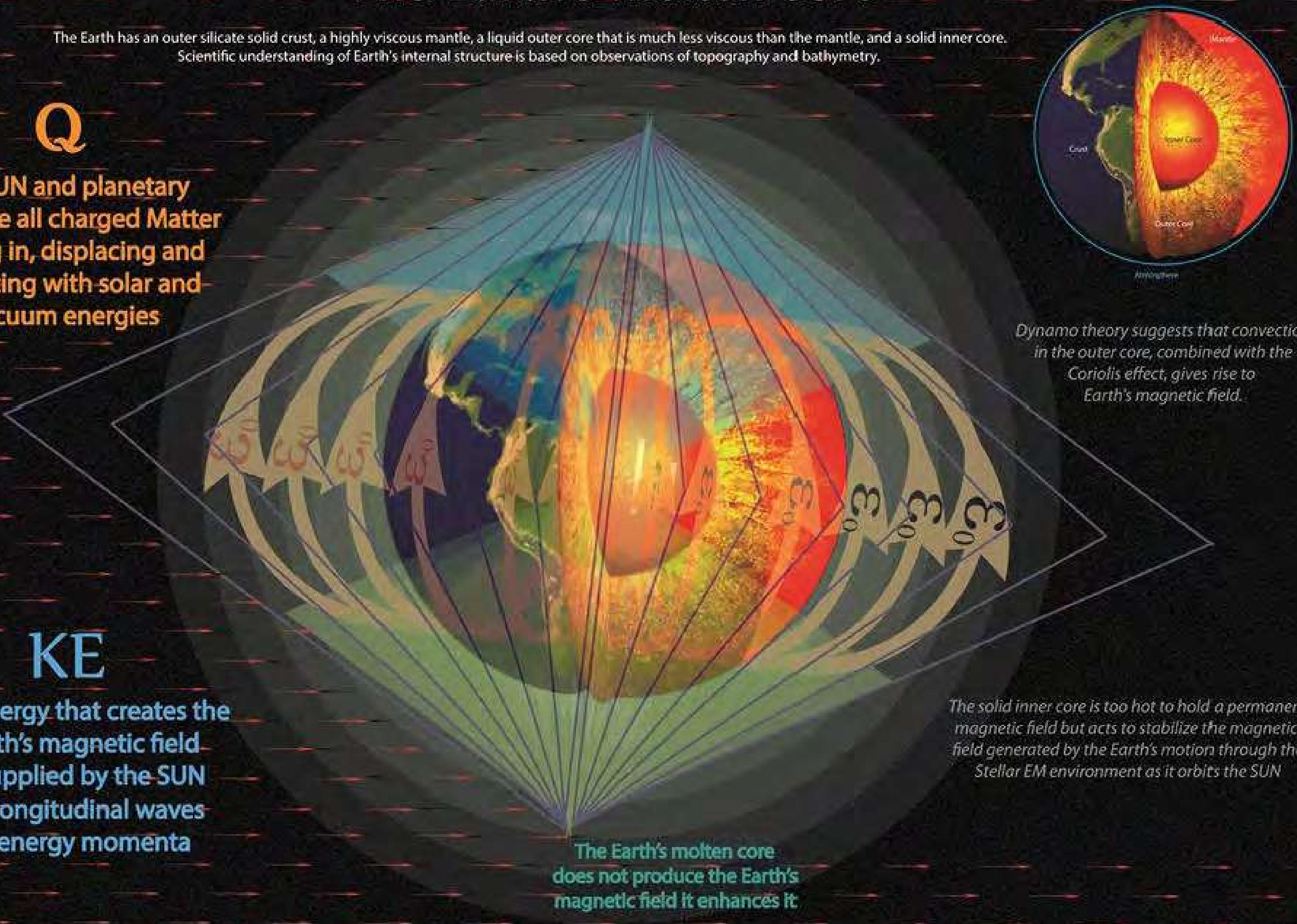

Q

The SUN and planetary bodies are all charged Matter moving in, displacing and interacting with solar and vacuum energies

Dynamo theory suggests that convection in the outer core, combined with the Coriolis effect, gives rise to Earth's magnetic field.

KE

The energy that creates the Earth's magnetic field is supplied by the SUN via longitudinal waves of energy momenta

The solid inner core is too hot to hold a permanent magnetic field but acts to stabilize the magnetic field generated by the Earth's motion through the Stellar EM environment as it orbits the SUN

The Earth's molten core does not produce the Earth's magnetic field it enhances it

The force exerted by Earth's gravity can be used to calculate its mass, and by estimating the volume of the planet, its average density can be calculated.

Astronomers can also calculate Earth's mass from its orbit and effects on nearby planetary bodies.

Terrestrial Aurora

result from the interaction of the Earth's magnetic field and the energetic solar wind, which provides energised ions and photons to the polar regions of Planets

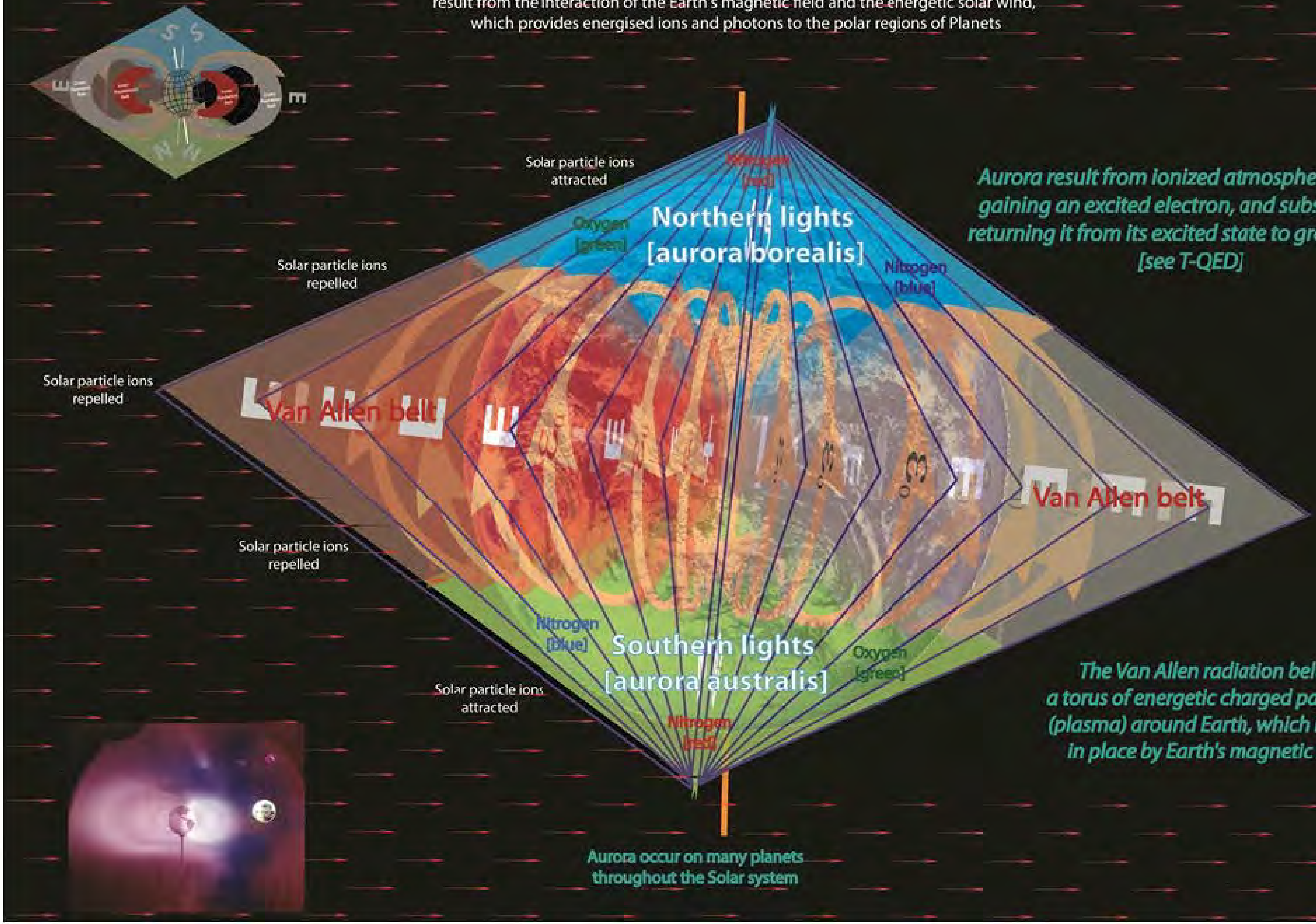
S
S
E
E
N
N
Solar particle ions attracted
Nitrogen [red]
Oxygen [green]
Northern lights [aurora borealis]
Nitrogen [blue]
Solar particle ions repelled
Solar particle ions repelled
Van Allen belt
Van Allen belt
Solar particle ions repelled
Nitrogen [blue]
Southern lights [aurora australis]
Oxygen [green]
Solar particle ions attracted
Nitrogen [red]

Aurora result from ionized atmospheric atoms gaining an excited electron, and subsequently returning it from its excited state to ground state [see T-QED]

The Van Allen radiation belt is a torus of energetic charged particles (plasma) around Earth, which is held in place by Earth's magnetic field

Aurora occur on many planets throughout the Solar system

Absolute Rest Frames

Absolute Rest is intrinsically linked to the concepts of absolute time and space originally introduced by Sir Isaac Newton in the Philosophiæ Naturalis Principia Mathematica, which provided a theoretical foundation that facilitated Newtonian mechanics.

According to Newton, absolute time and space respectively are independent aspects of objective reality.

Rest in physics refers to an object being stationary relative to a particular frame of reference or another object

In Tetryonics an Inertial Rest frame can be determined by the absence of a Kinetic EM field & its Magnetic moment

The concepts of space and time were separate in physical theory prior to the advent of special relativity theory, which connected the two and showed both to be dependent upon the observer's inertial state of motion relative to other inertial frames.

In Einstein's theories, the ideas of absolute time and space were superseded by the notion of spacetime in special relativity, and by dynamically curved spacetime in general relativity.

$$p^2 = E = mv^2$$

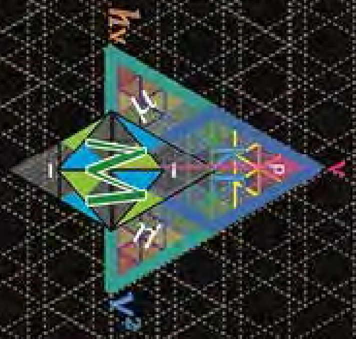

$$\frac{p^2}{2} = KE = \frac{1}{2}Mv^2$$

In some of his later paper (especially in 1920 and 1924), Einstein gave a new definition of the aether by identifying it with "properties of space", and this aether can be called absolute, as long as its state cannot be influenced by Matter.

So he argued that Newton's absolute space can be considered as an "absolute aether". Also the four-dimensional spacetime of special relativity (which replaces the absolute space of Newton) would be some sort of "absolute aether", as its states cannot be influenced by Matter.

Tetryonics clearly establishes free space as a 3D volume filled with a interactive vacuum [aether] and EM energies against which any relative motion can be deduced by the presence (or absence) of divergent [K]EM fields produced as a direct result of Matter in motion additionally, 3D mass-Matter topologies displace the aether [vacuum energy field] producing convergent differential energy gradients [gravity fields]

Red-shifted Photonic Energies

Energy is ALWAYS conserved... longer wavelength [red-shifted] photon energies are a reflection of the time specific measurement of photonic energies at a particular distance from the source of the light

Photons are neutral, coupled charge equilateral mass-energy momenta geometries that radiate bilaterally at the speed of light in a vacuum

Photons

$$\frac{2\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

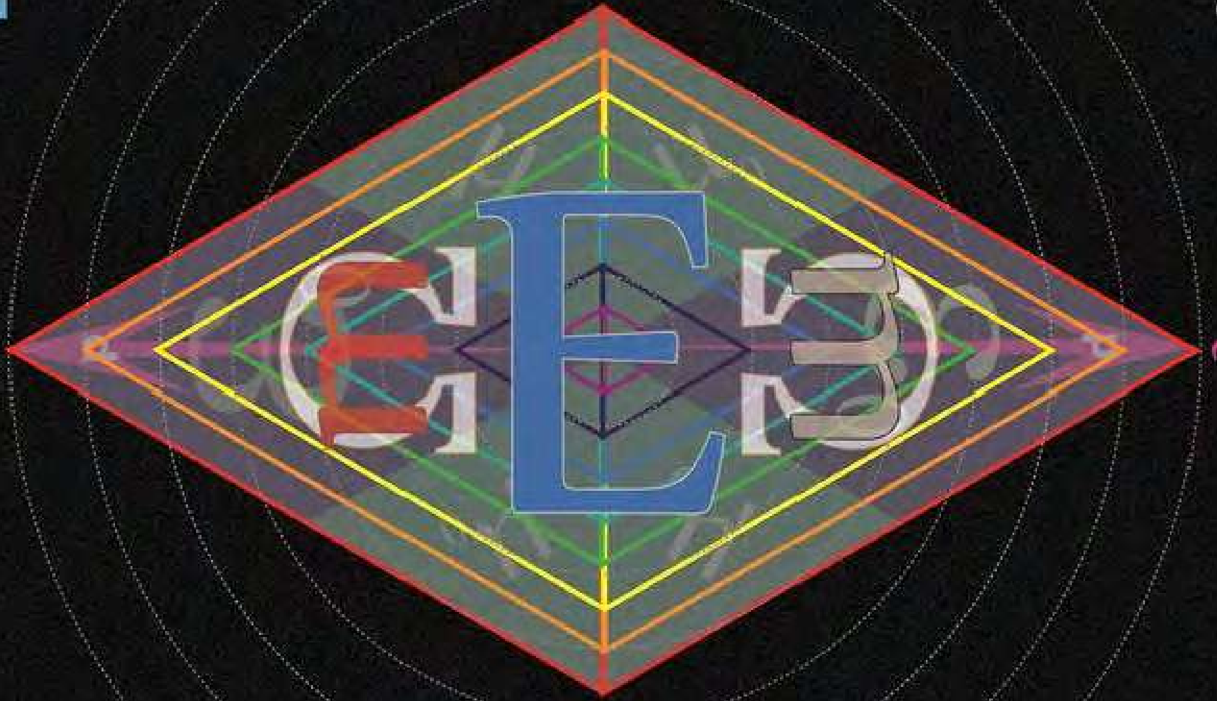

As photons radiate from a point source at 'c'

their EM energies diminsh as per the inverse square law

Planck quanta
mΩv²
mass velocity
c²
1 second

producing the familiar red-shifting of spectral lines

P
c²
1 second

Tetryonics 75.02 - Red-shifted Photonic energies

Copyright ABRAHAM [2008] - All rights reserved

The red shift of Photons

$d\lambda$

According to special relativity, as something approaches the speed of light, the passage of time slows.
A twin moving at the speed of light would not age or change relative to the stationary twin.

General relativity additionally states that the cosmological red shift is the result of a photon traveling through an expanding space time field which causes the photon to loose energy and it's wavelength correspondingly increases.

df

A photon travels at the speed of light; therefore it will not experience the passage of time, therefore it will not change; yet it does.

The divergent spreading of EM energy over time from a source creates the spacetime aether

The Quantised Angular Momentum and Energy content are directly related through the geometry of Photons

All EM Energy waveforms (save Matter) radiate outward following the inverse square law & reducing their energy content over time

Blue-shifted photon travelling towards observer

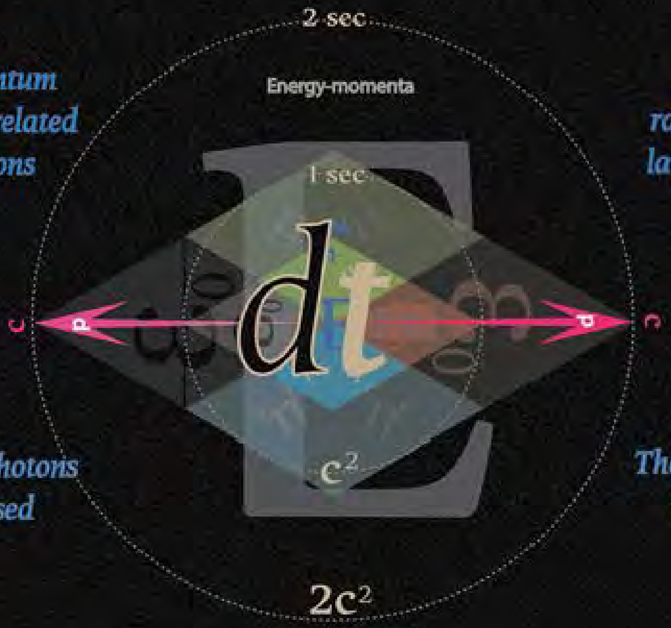

Red-shifted photon moving away from observer

The EM mass-energy content of photons reveals photons to be a quantised electromotive field

The wavelength of a photon is related directly to its quantised mass-energy momenta geometry per unit of time

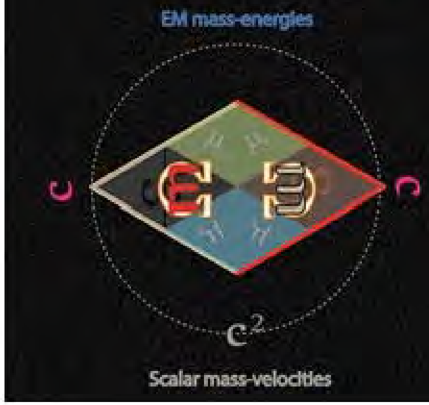

Tetryonic geometry relates photon mass-energy momenta to the constant velocity of light in a vacuum and its intrinsic quantised angular momenta

$$\frac{mc^2}{c^2} = KE = \frac{hv^2}{c^2}$$

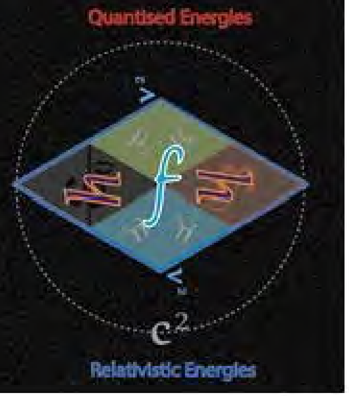

Tetryonics 75.03 - The Red-shift of Photons

Copyright ABRAHAM [2008] - All rights reserved

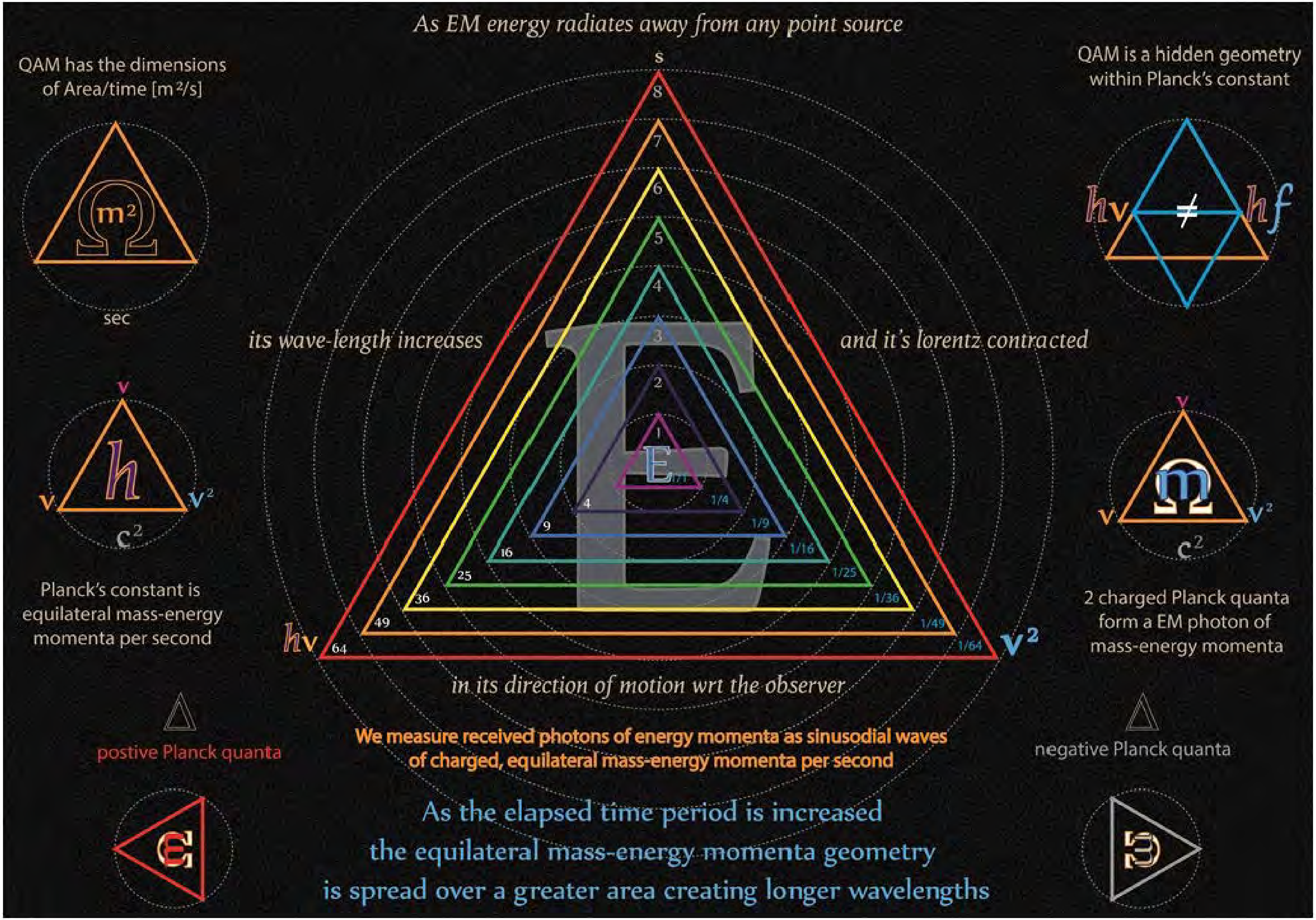
sec
m²
Ω
v
h
v
v²
c²
hv
≠
hf
v
m
Ω
v
v²
c²
s
8
7
6
5
4
3
2
1
E
4
9
16
25
36
49
64
1/1
1/4
1/9
1/16
1/25
1/36
1/49
1/64
hv
v²
ε
ω

As EM energy radiates away from any point source

its wave-length increases

and it's lorentz contracted

in its direction of motion wrt the observer

QAM has the dimensions
of Area/time [m²/s]

QAM is a hidden geometry
within Planck's constant

Planck's constant is
equilateral mass-energy
momenta per second

2 charged Planck quanta
form a EM photon of
mass-energy momenta

postive Planck quanta

negative Planck quanta

We measure received photons of energy momenta as sinusodial waves
of charged, equilateral mass-energy momenta per second

As the elapsed time period is increased
the equilateral mass-energy momenta geometry
is spread over a greater area creating longer wavelengths

Copyright ABRAHAM [2008] - All rights reserved

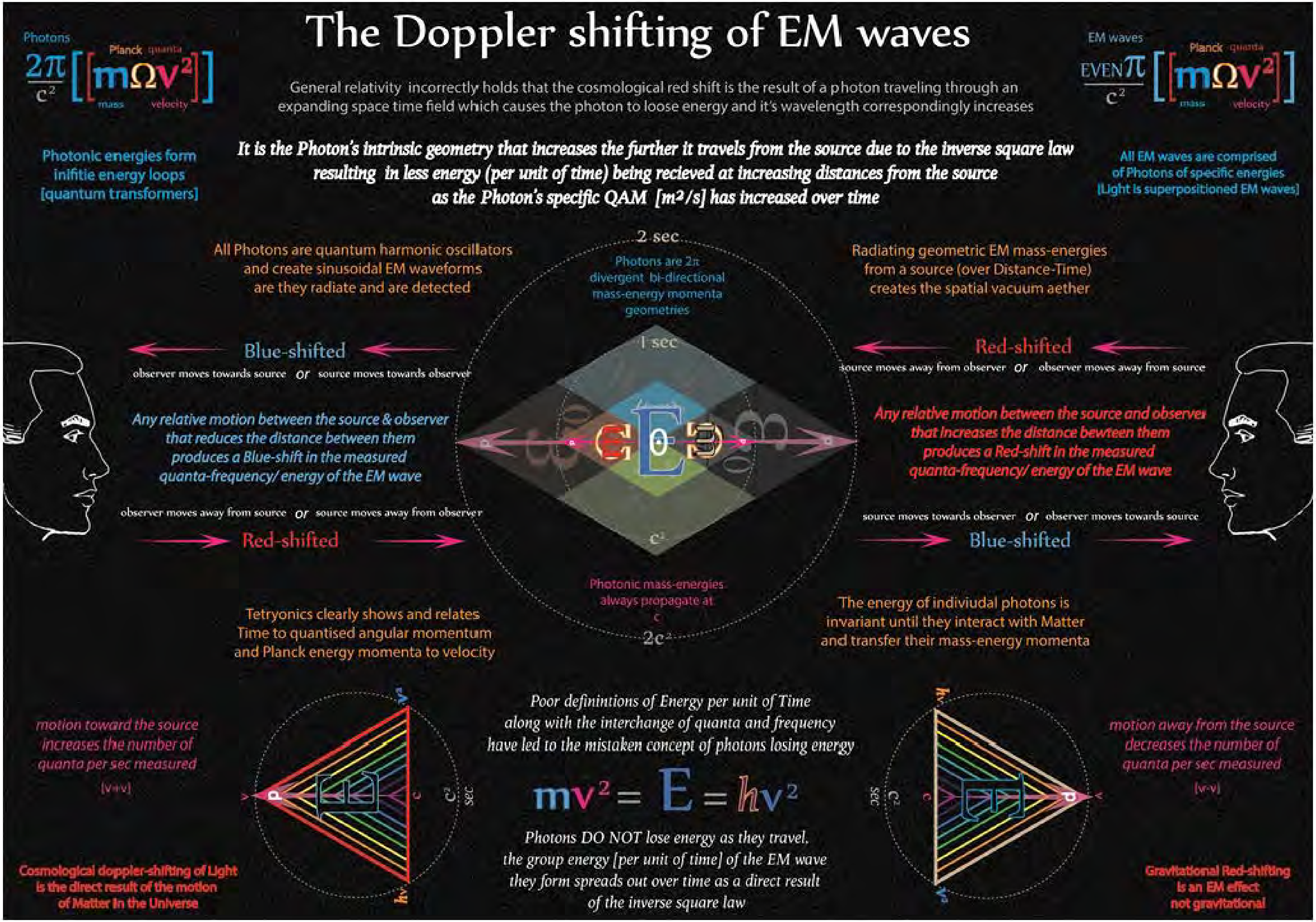

The Doppler shifting of EM waves

Photons

$$\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Photonic energies form
inifitie energy loops
[quantum transformers]

General relativity incorrectly holds that the cosmological red shift is the result of a photon traveling through an expanding space time field which causes the photon to loose energy and it's wavelength correspondingly increases

It is the Photon's intrinsic geometry that increases the further it travels from the source due to the inverse square law resulting in less energy (per unit of time) being recieved at increasing distances from the source as the Photon's specific QAM [m²/s] has increased over time

EM waves

$$\frac{EVEN\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

All EM waves are comprised
of Photons of specific energies
[Light is superpositioned EM waves]

All Photons are quantum harmonic oscillators
and create sinusoidal EM waveforms
are they radiate and are detected

Photons are 2π
divergent bi-directional
mass-energy momenta
geometries

Radiating geometric EM mass-energies
from a source (over Distance-Time)
creates the spatial vacuum aether

Blue-shifted

observer moves towards source *or* source moves towards observer

Any relative motion between the source & observer that reduces the distance between them produces a Blue-shift in the measured quanta-frequency/ energy of the EM wave

observer moves away from source *or* source moves away from observer

Red-shifted

Red-shifted

source moves away from observer *or* observer moves away from source

Any relative motion between the source and observer that Increases the distance bewteen them produces a Red-shift in the measured quanta-frequency/ energy of the EM wave

source moves towards observer *or* observer moves towards source

Blue-shifted

Photonic mass-energies
always propagate at
c

Tetryonics clearly shows and relates
Time to quantised angular momentum
and Planck energy momenta to velocity

The energy of indiviudal photons is
invariant until they interact with Matter
and transfer their mass-energy momenta

*motion toward the source
increases the number of
quanta per sec measured*

[v+v]

*Poor definintions of Energy per unit of Time
along with the interchange of quanta and frequency
have led to the mistaken concept of photons losing energy*

$$mv^2 = E = hv^2$$

*Photons DO NOT lose energy as they travel.
the group energy [per unit of time] of the EM wave
they form spreads out over time as a direct result
of the inverse square law*

*motion away from the source
decreases the number of
quanta per sec measured*

[v-v]

**Cosmological doppler-shifting of Light
is the direct result of the motion
of Matter in the Universe**

**Gravitational Red-shifting
is an EM effect
not gravitational**

Tetryonics 75.05 - The Doppler shifting of EM waves

Relativistic Corrections

are applied to only [kinetic]EM mass-energy geometries

$$\frac{Mv^2}{c^2} = KEM = \frac{hv^2}{c^2}$$

$d\lambda$

Wave-length contraction

linear
Lorentz corrections

$$\beta = \left[\frac{v}{c}\right]$$

SR - measuring rods in differing interial frames have different contracted lengths

GR - lengths measured in curved spacetime are different to straight lines in flat spacetime

$$\frac{\Delta\Omega}{c}$$

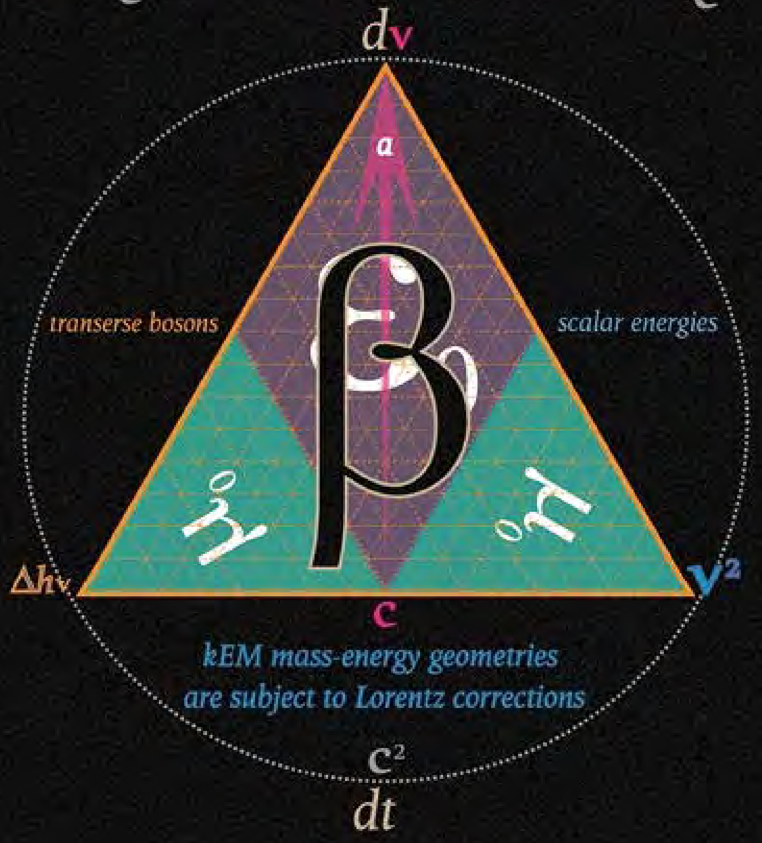

kEM mass-energy geometries
are subject to Lorentz corrections

dt

Time Dilation

scalar
Lorentz corrections

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

SR - clocks in differing inertial frames experience time dilation

GR - clocks in differing Gravitational fields experience time dilation

$$\frac{\Delta\Omega}{c^2}$$

scalar energy momenta

Matter Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v\right]^2\right]$$

mass velocity

linear momentum

The mass-energies of standing-wave Matter topologies are Lorentz invariant

Copyright ABRAHAM [2008] - All rights reserved

GPS relativistic corrections

Claims that the Global Positioning System (GPS) provides a valid test of General relativity are erroneous
[GR provides Lorentzian [SR] correcttions under the guise of Gravitational effects]

Photons are mass-energy momenta quanta
[they are Matter-less NOT mass-less]
and are subect to Lorentz Forces and corrections

Newton G + SR = GR *Einstein*
Lorentz

Graviational Matter

M

All Matter topologies
produce
a Gravitational field

Newtonian Gravity
is universal

The gravitational effects
caused by Matter can only
be accurately modelled [or
tested] through the exclusion
of all EM fields interactions

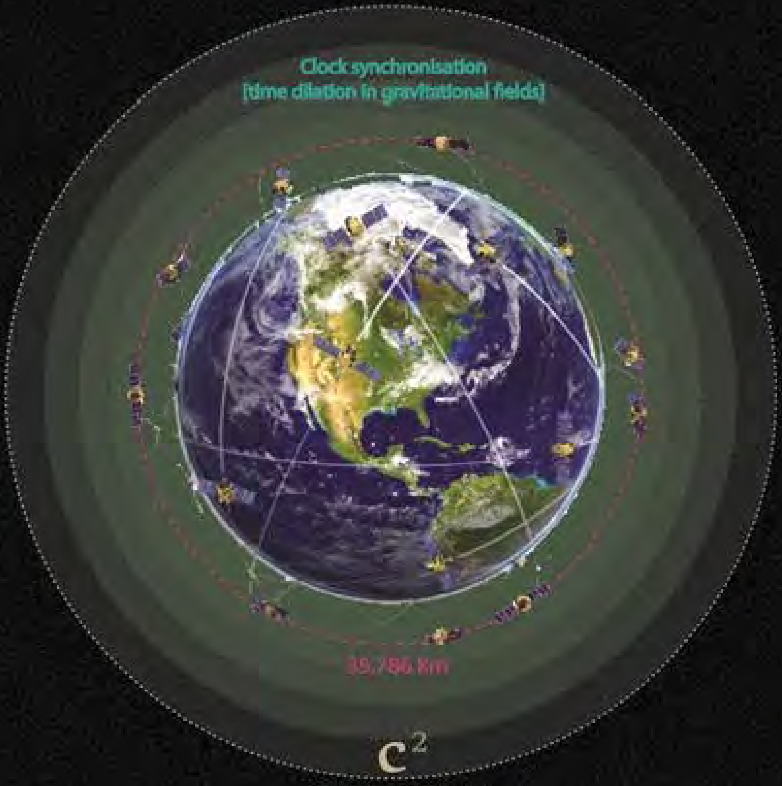
Clock synchronisation
[time dilation in gravitational fields]
35,786 km
c^2

299,792,458 km

Interactive masses

m

All Matter in motion
produces
interactive mass-energies

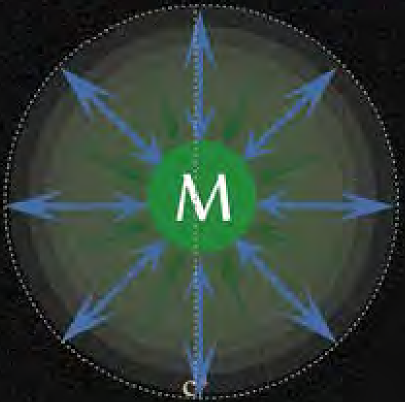

General Relativity
is limted to c

Any test of GR using Photons
as a testing mechanism
by definition is affected by
and subject to Special relativity
[ElectroMagnetic effects]

The 'gravitational' frequency shifting [Red-Blue shifting of EM waves] is in fact an electromagnetic effect

Copyright ABRAHAM [2008] - All rights reserved

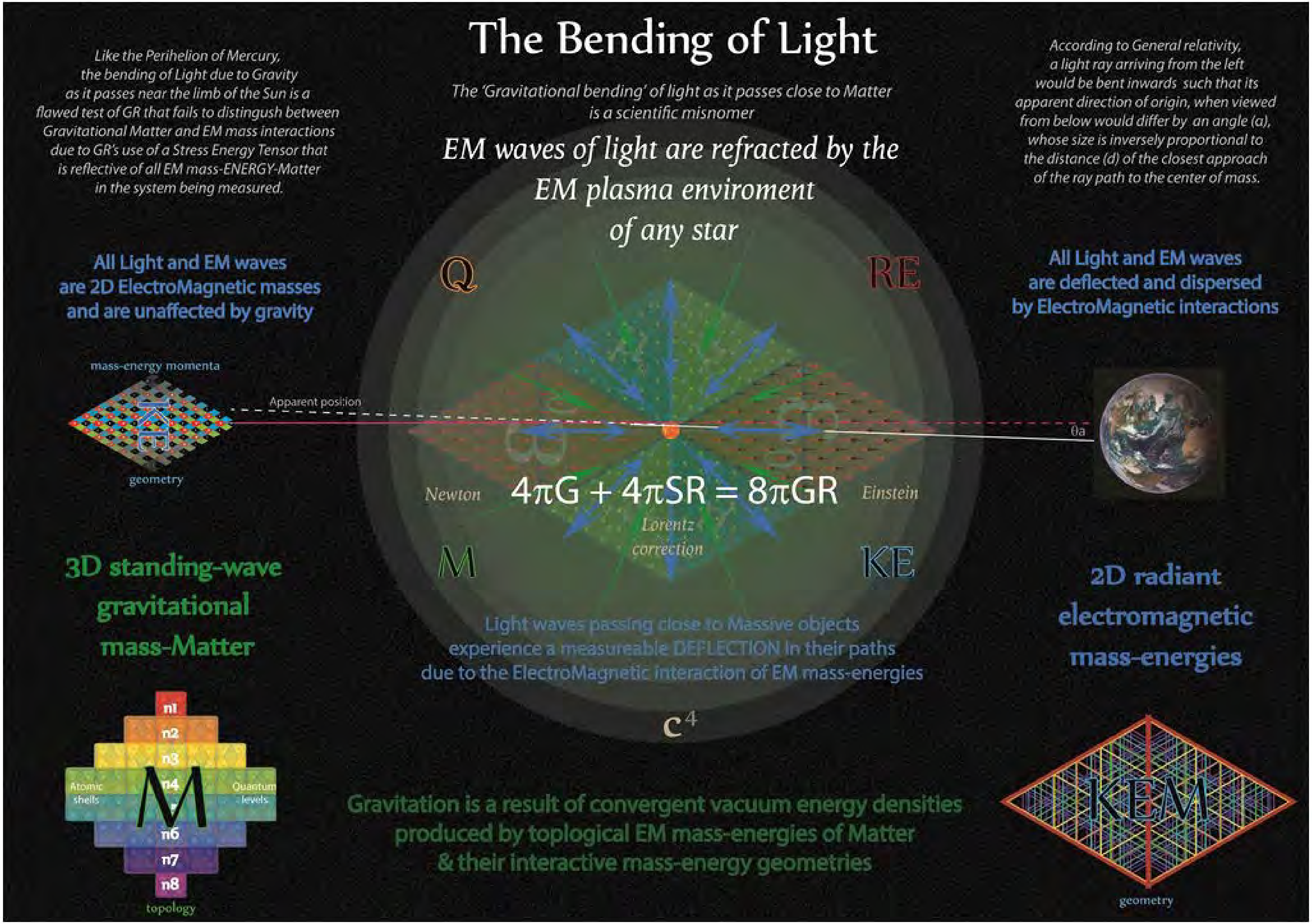

The Bending of Light

*Like the Perihelion of Mercury,
the bending of Light due to Gravity
as it passes near the limb of the Sun is a
flawed test of GR that fails to distingush between
Gravitational Matter and EM mass interactions
due to GR's use of a Stress Energy Tensor that
is reflective of all EM mass-ENERGY-Matter
in the system being measured.*

*The 'Gravitational bending' of light as it passes close to Matter
is a scientific misnomer*

*EM waves of light are refracted by the
EM plasma enviroment
of any star*

*According to General relativity,
a light ray arriving from the left
would be bent inwards such that its
apparent direction of origin, when viewed
from below would differ by an angle (a),
whose size is inversely proportional to
the distance (d) of the closest approach
of the ray path to the center of mass.*

**All Light and EM waves
are 2D ElectroMagnetic masses
and are unaffected by gravity**

**All Light and EM waves
are deflected and dispersed
by ElectroMagnetic interactions**

Q
RE
M
KE
mass-energy momenta
KE
geometry
Apparent position
θa
Newton
Einstein
Lorentz
correction
c^4
n1
n2
n3
n4
n6
n7
n8
Atomic
shells
Quantum
levels
M
topology
KEM
geometry

$$4\pi G + 4\pi SR = 8\pi GR$$

**3D standing-wave
gravitational
mass-Matter**

**2D radiant
electromagnetic
mass-energies**

Light waves passing close to Massive objects
experience a measureable DEFLECTION in their paths
due to the ElectroMagnetic interaction of EM mass-energies

**Gravitation is a result of convergent vacuum energy densities
produced by toplogical EM mass-energies of Matter
& their interactive mass-energy geometries**

Tetryonics 75.08 - The Bending of Light

Copyright ABRAHAM [2008] - All rights reserved

The Perihelion of Mercury

is the result of EM interactions in addition to gravitational attraction

The perihelion of Mercury's orbit precesses at a measurable rate, but even after accounting for gravitational perturbations caused all other planets in the solar system, Newton's theory (assuming a precise inverse-square relationship for distance) predicts a rate of precession that differs from the measured rate by approximately 43 arcseconds per century.

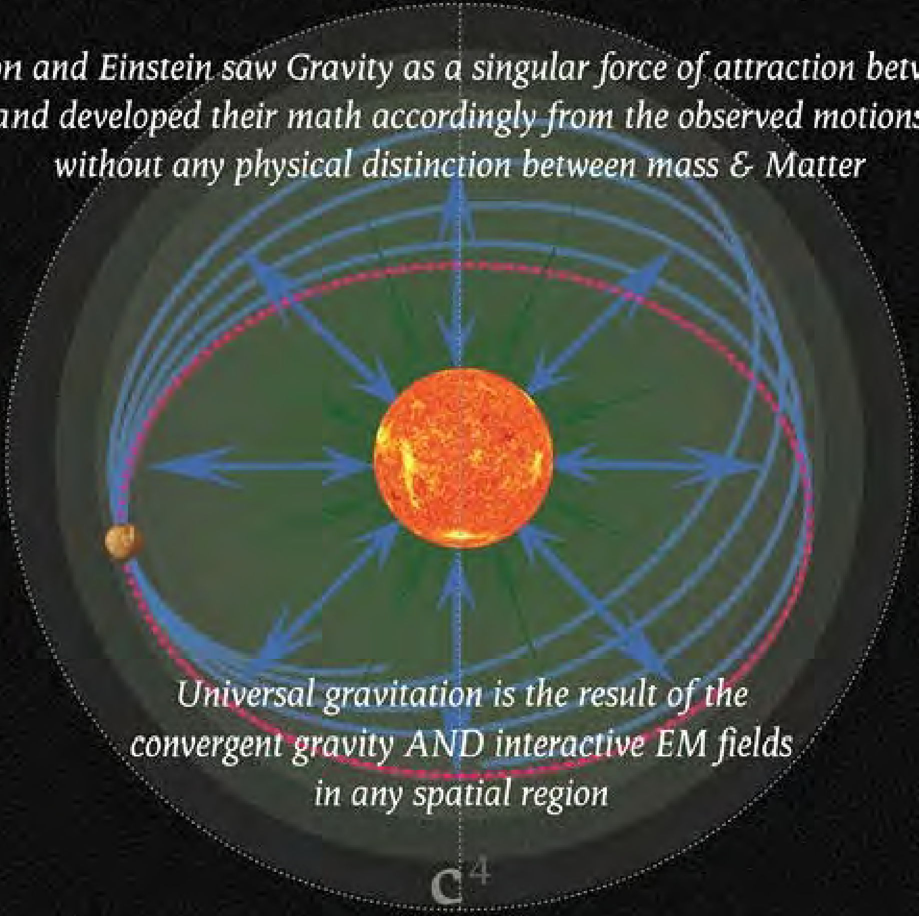

Both Newton and Einstein saw Gravity as a singular force of attraction between objects and developed their math accordingly from the observed motions without any physical distinction between mass & Matter

$4\pi G$

Newton modelled only the observed nett attractive force of Matter in the system

$$F = -G\frac{M_1 m_2}{r^2}$$

$8\pi G$

Einstein modelled the nett convergent force of all energies without differentiation

$$G_{ab} = \kappa T_{ab}$$

Universal gravitation is the result of the convergent gravity AND interactive EM fields in any spatial region

c^4

Newton

G

Einstein

$G + EM$

Newton Lorentz

General relativity was developed in part to provide an estimate for this rate of precession that better matches observations.

Newton's theory can also be 'incorrectly modified' to explain this perihelion by factoring in the gravitational pull due to other planets or making tiny adjustments to parameters in the gravitational equation to reflect the effect of ElectroMagnetic interactions between the Sun and Mercury.

By modelling the quantum interactions of ALL GEM mass-ENERGIES and their distinct geometries Tetryonic theory explains recently measured discrepancies in both Newtonian & GR formulations of Gravity

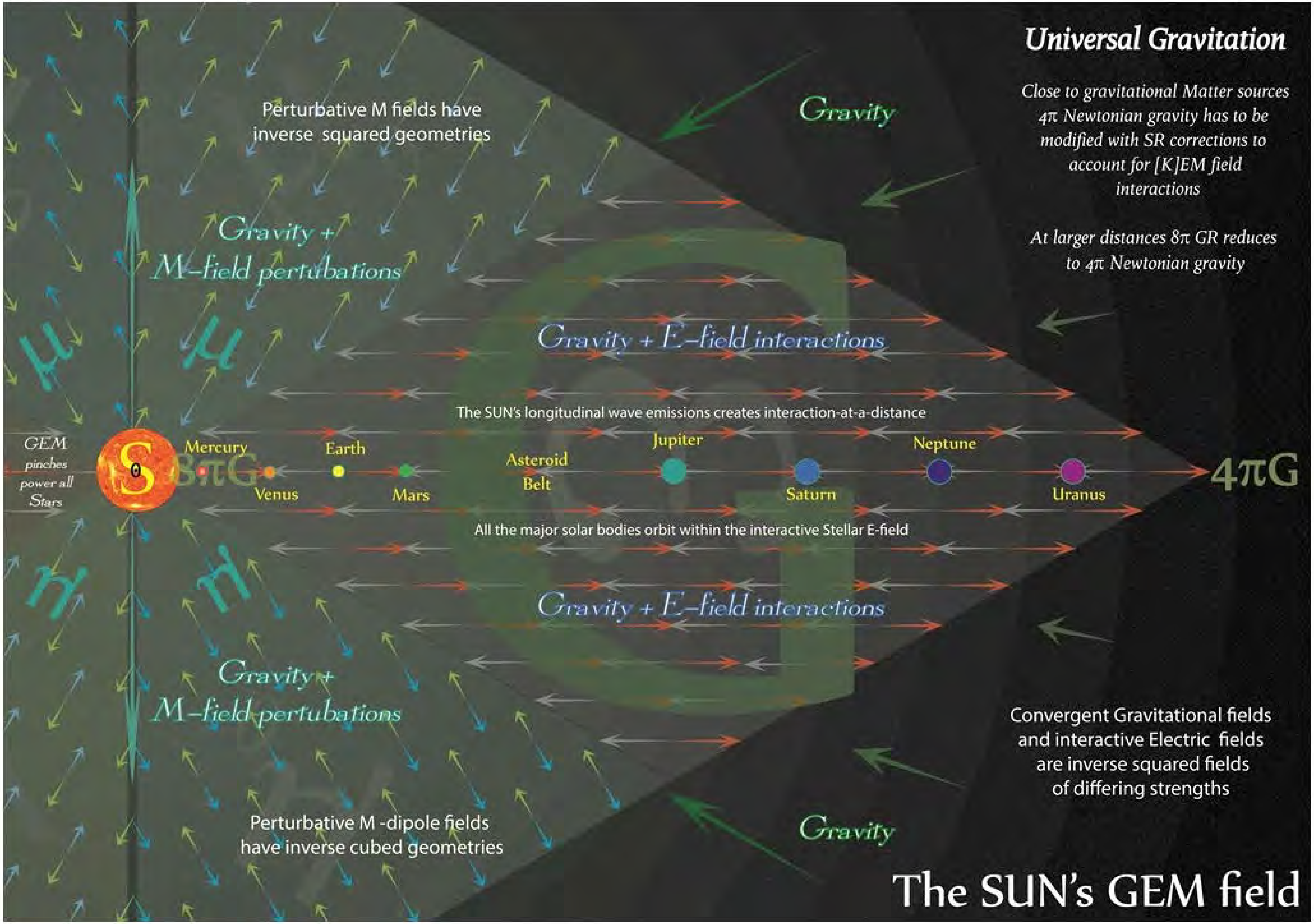

Gravity
Gravity + M-field pertubations
Gravity + E-field interactions
Gravity + E-field interactions
Gravity + M-field pertubations
Gravity
μ
μ
η
η
GEM pinches power all Stars
S
8πG
Mercury
Venus
Earth
Mars
Asteroid Belt
Jupiter
Saturn
Neptune
Uranus
4πG
G

Universal Gravitation

Close to gravitational Matter sources 4π Newtonian gravity has to be modified with SR corrections to account for [K]EM field interactions

At larger distances 8π GR reduces to 4π Newtonian gravity

Perturbative M fields have inverse squared geometries

The SUN's longitudinal wave emissions creates interaction-at-a-distance

All the major solar bodies orbit within the interactive Stellar E-field

Convergent Gravitational fields and interactive Electric fields are inverse squared fields of differing strengths

Perturbative M -dipole fields have inverse cubed geometries

The SUN's GEM field

Tetryonics 76.01 - The SUN's GEM field

Copyright ABRAHAM [2008] - All rights reserved

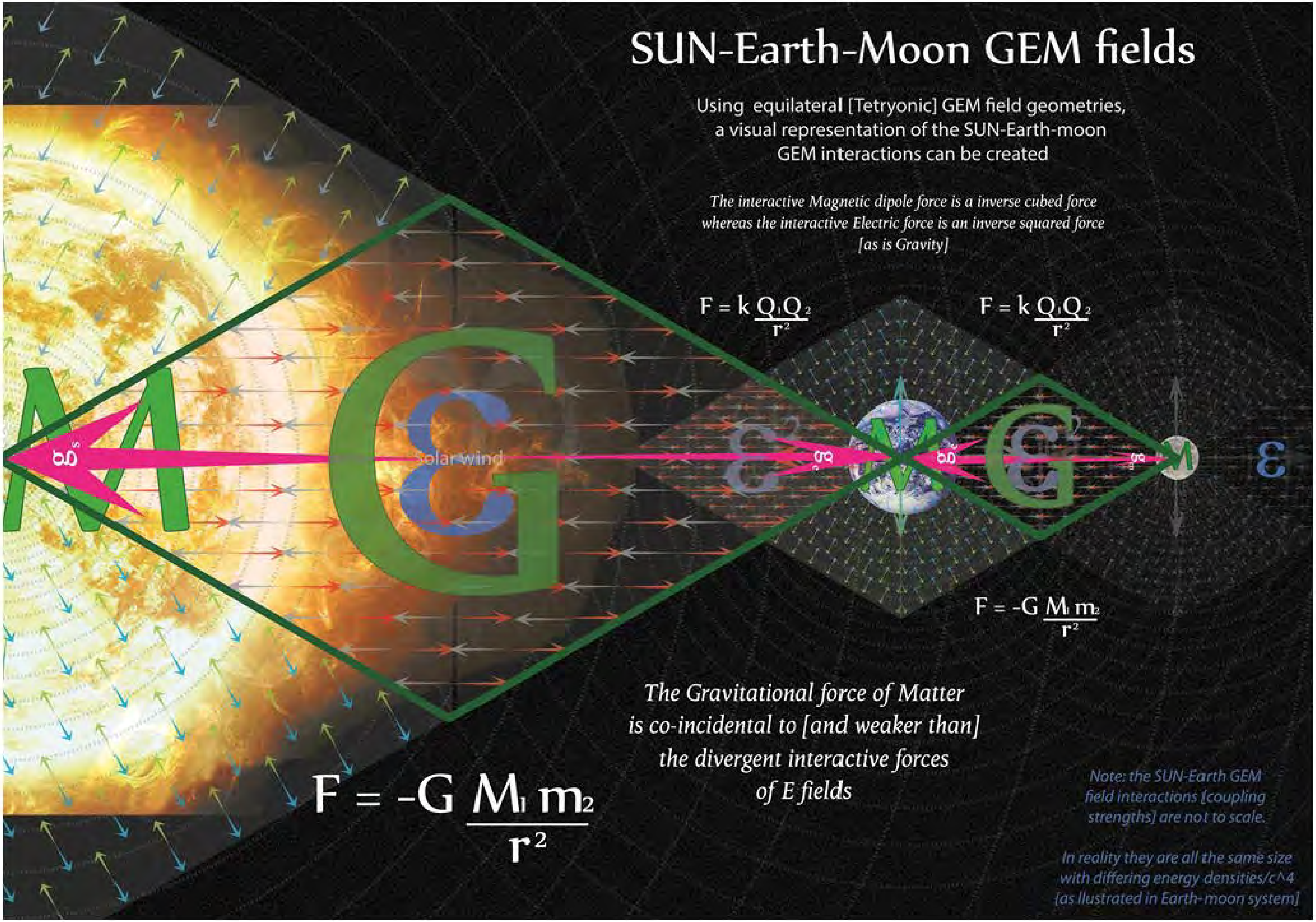

SUN-Earth-Moon GEM fields

Using equilateral [Tetryonic] GEM field geometries, a visual representation of the SUN-Earth-moon GEM interactions can be created

The interactive Magnetic dipole force is a inverse cubed force whereas the interactive Electric force is an inverse squared force [as is Gravity]

$$F = k\frac{Q_1Q_2}{r^2}$$

$$F = k\frac{Q_1Q_2}{r^2}$$

$$F = -G\frac{M_1 m_2}{r^2}$$

The Gravitational force of Matter is co-incidental to [and weaker than] the divergent interactive forces of E fields

$$F = -G\frac{M_1 m_2}{r^2}$$

Note: the SUN-Earth GEM field interactions [coupling strengths] are not to scale.

In reality they are all the same size with differing energy densities/c^4 [as illustrated in Earth-moon system]

Tetryonics 76.02 - SUN-Earth-moon GEM fields

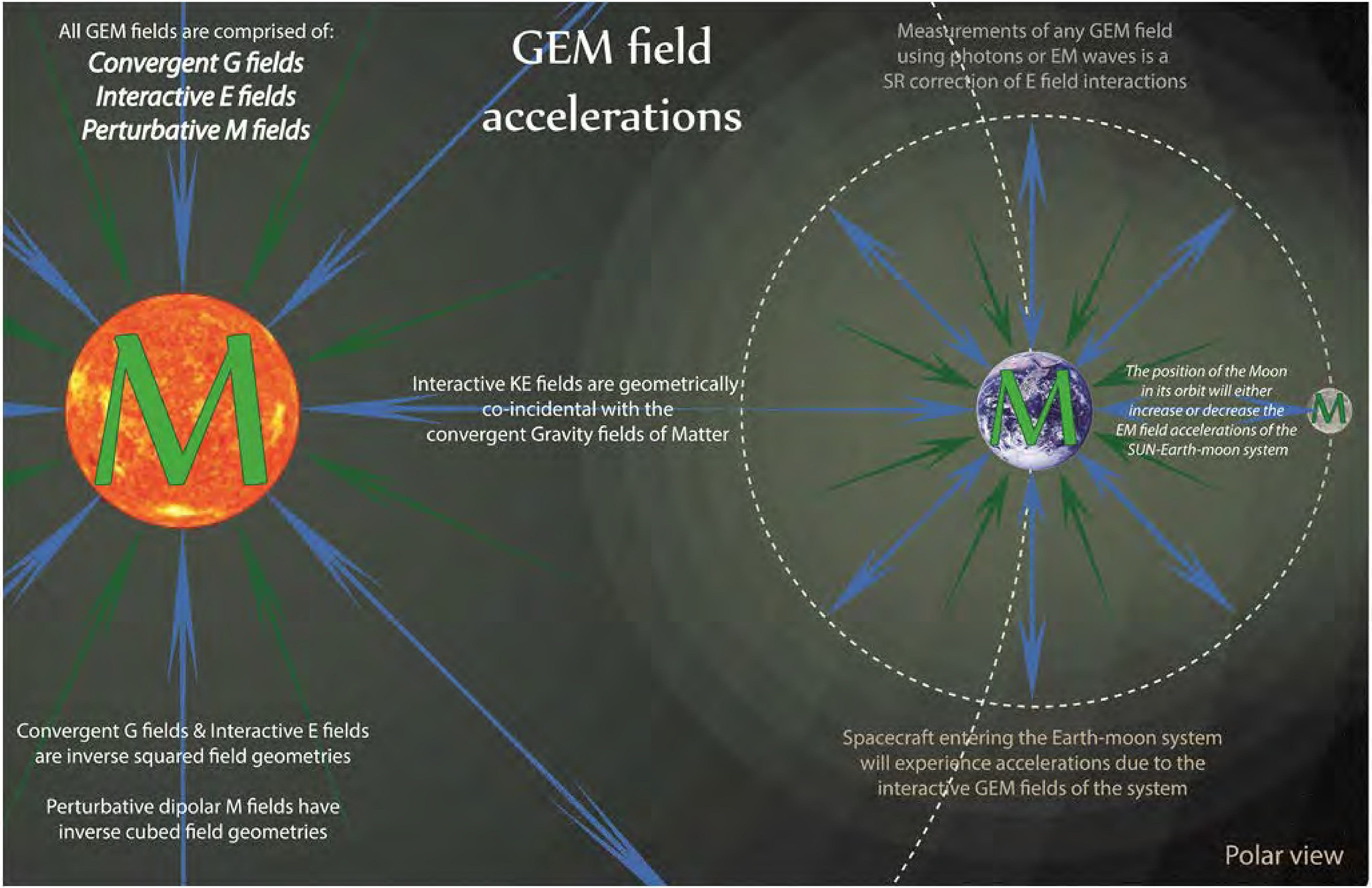

GEM field accelerations

All GEM fields are comprised of:

Convergent G fields
Interactive E fields
Perturbative M fields

Measurements of any GEM field using photons or EM waves is a SR correction of E field interactions

Interactive KE fields are geometrically co-incidental with the convergent Gravity fields of Matter

The position of the Moon in its orbit will either increase or decrease the EM field accelerations of the SUN-Earth-moon system

Convergent G fields & Interactive E fields are inverse squared field geometries

Perturbative dipolar M fields have inverse cubed field geometries

Spacecraft entering the Earth-moon system will experience accelerations due to the interactive GEM fields of the system

Tetryonics 76.03 - GEM fields [Polar]

Copyright ABRAHAM [2008] - All rights reserved

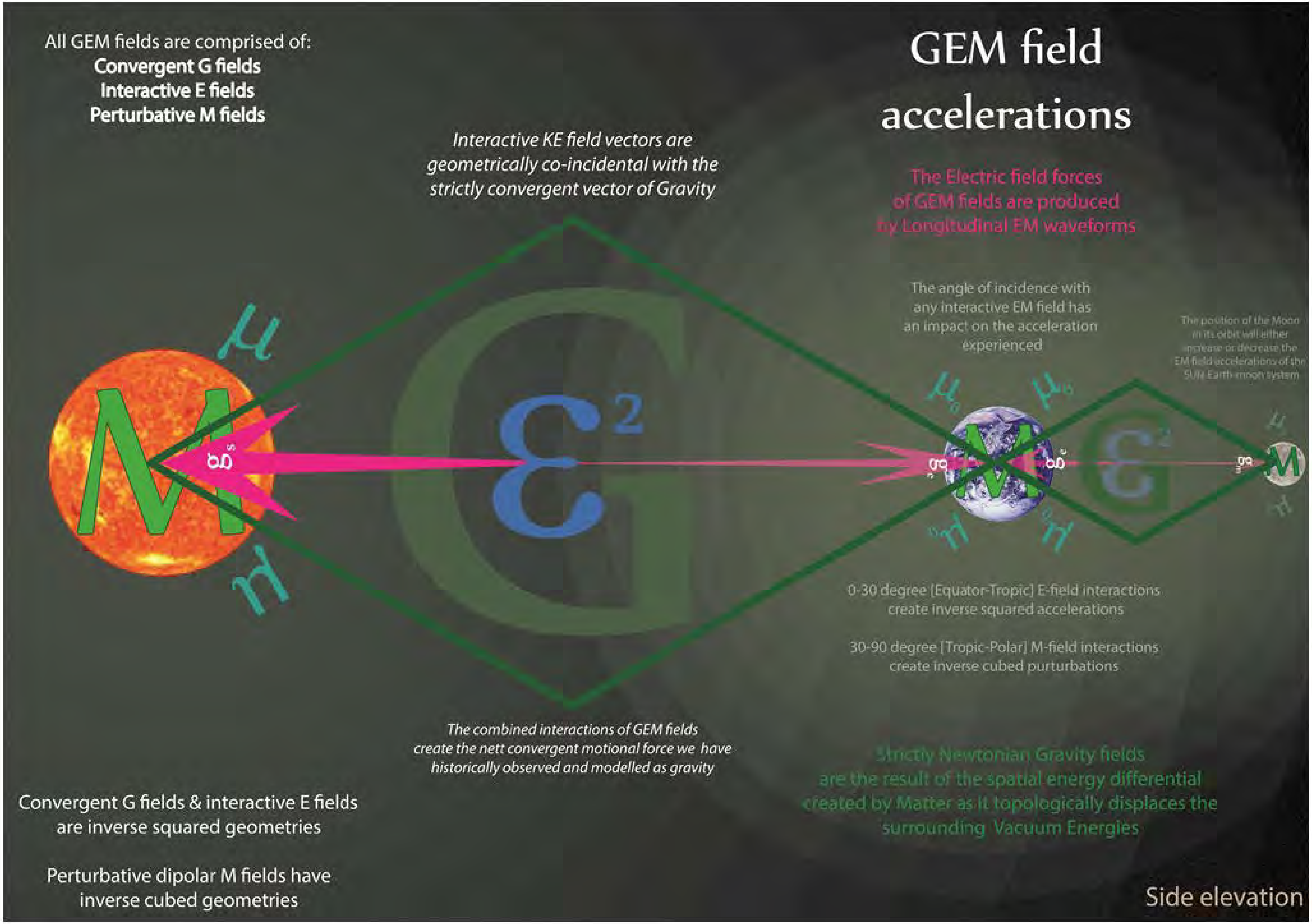

GEM field accelerations

All GEM fields are comprised of:
Convergent G fields
Interactive E fields
Perturbative M fields

Interactive KE field vectors are geometrically co-incidental with the strictly convergent vector of Gravity

The Electric field forces of GEM fields are produced by Longitudinal EM waveforms

The angle of incidence with any interactive EM field has an impact on the acceleration experienced

The position of the Moon in its orbit will either increase or decrease the EM field accelerations of the SUN Earth-moon system

0-30 degree [Equator-Tropic] E-field interactions create inverse squared accelerations

30-90 degree [Tropic-Polar] M-field interactions create inverse cubed purturbations

The combined interactions of GEM fields create the nett convergent motional force we have historically observed and modelled as gravity

Strictly Newtonian Gravity fields are the result of the spatial energy differential created by Matter as it topologically displaces the surrounding Vacuum Energies

Convergent G fields & interactive E fields are inverse squared geometries

Perturbative dipolar M fields have inverse cubed geometries

Side elevation

Tetryonics 76.04 - GEM fields [Side]

Gravitational flyby accelerations

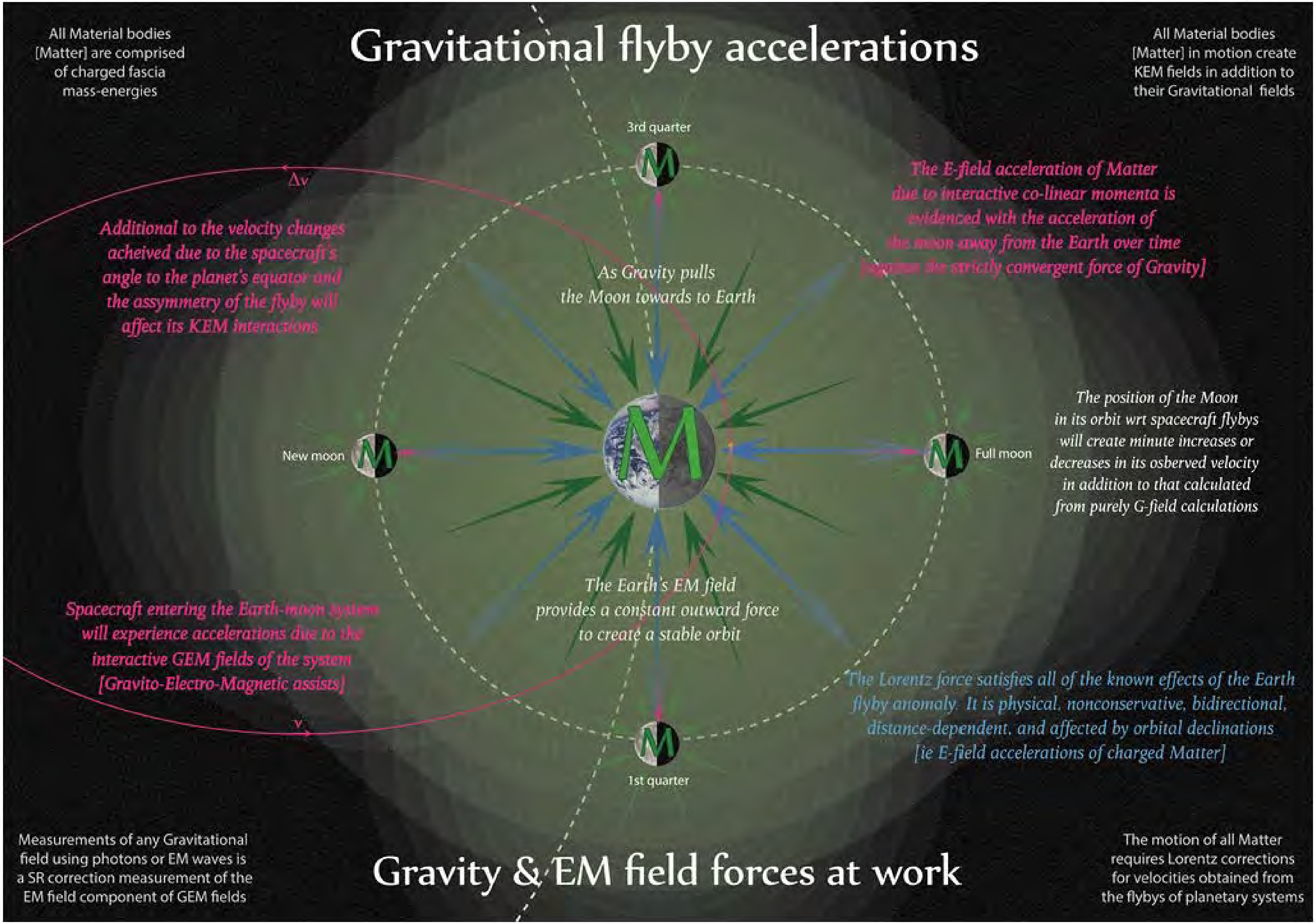

All Material bodies
[Matter] are comprised
of charged fascia
mass-energies

All Material bodies
[Matter] in motion create
KEM fields in addition to
their Gravitational fields

*The E-field acceleration of Matter
due to interactive co-linear momenta is
evidenced with the acceleration of
the moon away from the Earth over time
[opposes the strictly convergent force of Gravity]*

*Additional to the velocity changes
acheived due to the spacecraft's
angle to the planet's equator and
the assymmetry of the flyby will
affect its KEM interactions*

*As Gravity pulls
the Moon towards to Earth*

*The position of the Moon
in its orbit wrt spacecraft flybys
will create minute increases or
decreases in its osberved velocity
in addition to that calculated
from purely G-field calculations*

*The Earth's EM field
provides a constant outward force
to create a stable orbit*

*Spacecraft entering the Earth-moon system
will experience accelerations due to the
interactive GEM fields of the system
[Gravito-Electro-Magnetic assists]*

*The Lorentz force satisfies all of the known effects of the Earth
flyby anomaly. It is physical, nonconservative, bidirectional,
distance-dependent, and affected by orbital declinations
[ie E-field accelerations of charged Matter]*

Measurements of any Gravitational
field using photons or EM waves is
a SR correction measurement of the
EM field component of GEM fields

Gravity & EM field forces at work

The motion of all Matter
requires Lorentz corrections
for velocities obtained from
the flybys of planetary systems

Tetryonics 76.05 - Gravitational flyby accelerations

Flyby anomalies

ALL Matter creates GravitoElectoMagnetic fields

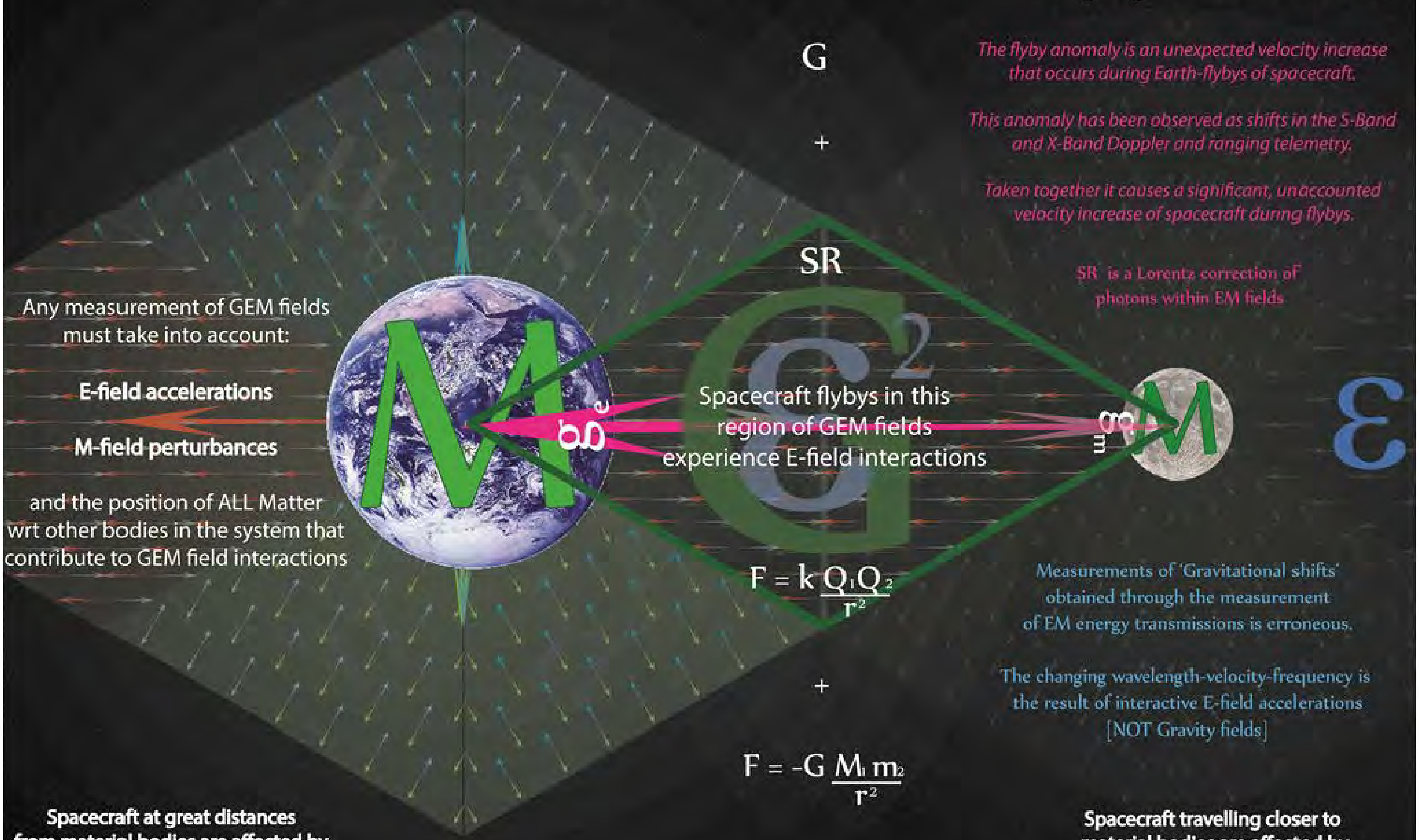

The flyby anomaly is an unexpected velocity increase that occurs during Earth-flybys of spacecraft.

This anomaly has been observed as shifts in the S-Band and X-Band Doppler and ranging telemetry.

Taken together it causes a significant, unaccounted velocity increase of spacecraft during flybys.

SR is a Lorentz correction of photons within EM fields

Any measurement of GEM fields must take into account:

E-field accelerations

M-field perturbances

and the position of ALL Matter wrt other bodies in the system that contribute to GEM field interactions

Spacecraft flybys in this region of GEM fields experience E-field interactions

$$F = k\frac{Q_1 Q_2}{r^2}$$

$$F = -G\frac{M_1 m_2}{r^2}$$

Measurements of 'Gravitational shifts' obtained through the measurement of EM energy transmissions is erroneous.

The changing wavelength-velocity-frequency is the result of interactive E-field accelerations [NOT Gravity fields]

Spacecraft at great distances from material bodies are affected by convergent Gravitational fields [Newtonian gravity]

Spacecraft travelling closer to material bodies are affected by interactive E-fields as well as G-fields [General Relativity]

Pioneer anomaly

Despite many proposed solutions,
there is not yet an accepted explanation
for the cause of the Pioneer velocity anomaly

Newtonian Gravitation

$$F = -G\frac{M_1 m_2}{r^2}$$

measures the nett gravitational attraction
of Matter towards other Matter
[using weak [K]EM fields]

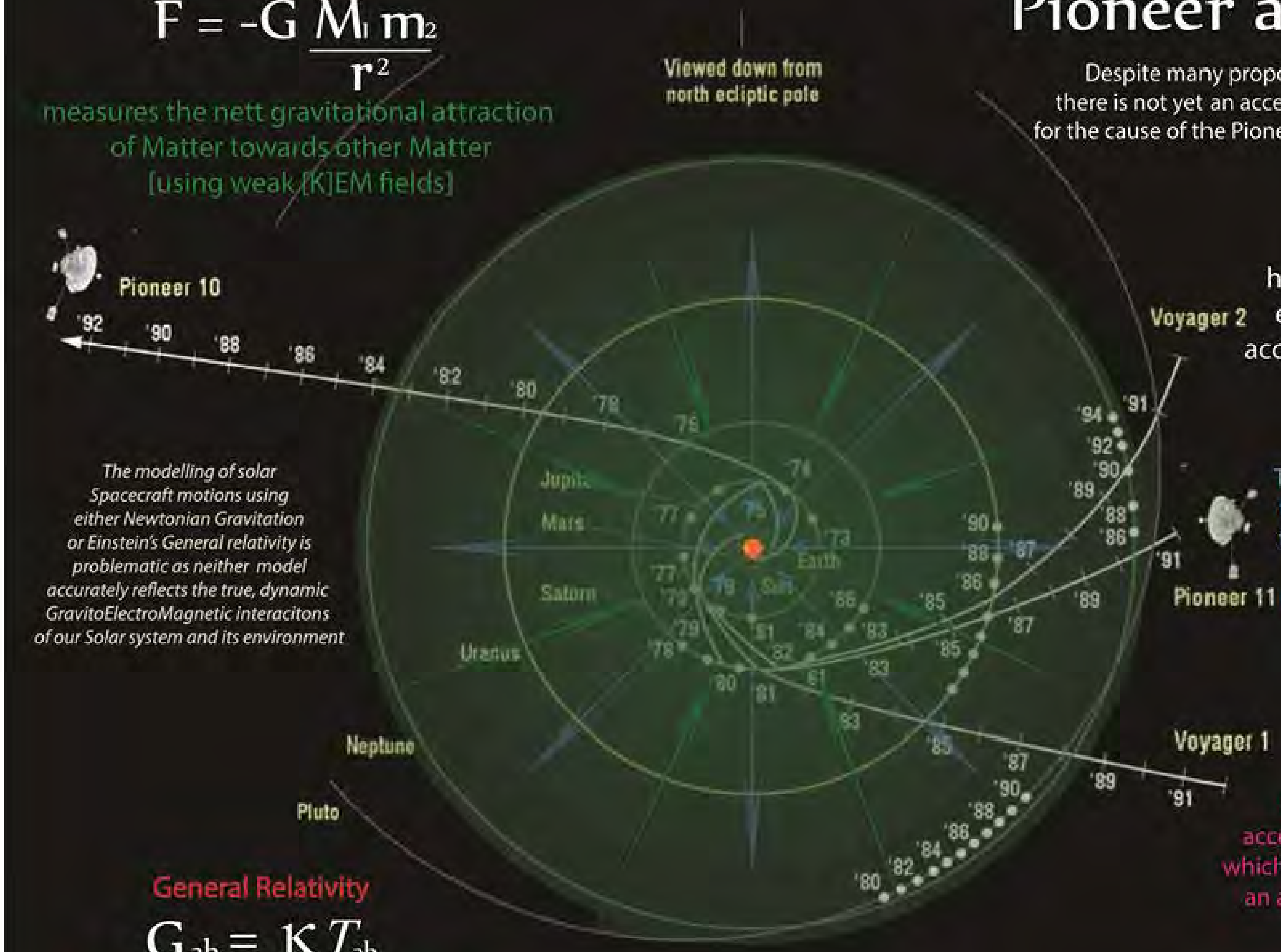
Viewed down from north ecliptic pole
Pioneer 10
'92
'90
'88
'86
'84
'82
'80
'78
'76
'74
Jupiter
Mars
Saturn
Uranus
Neptune
Pluto
Earth
Sun
Voyager 2
'94
'92
'90
'89
'88
'86
'91
Pioneer 11
'91
'89
'87
Voyager 1
'87
'89
'91
'90
'88
'86
'84
'82
'80

Many proposals
have been put forward to
explain this unexpected
acceleration towards the Sun

*The modelling of solar
Spacecraft motions using
either Newtonian Gravitation
or Einstein's General relativity is
problematic as neither model
accurately reflects the true, dynamic
GravitoElectroMagnetic interacitons
of our Solar system and its environment*

Tetryonic GEM interactions shows
that the measured acceleration is
the result of a weakening, inverse
squared stellar EM interaction
throughout the Solar heliopause

This would result in
a decrease in divergent
acceleration away from the Sun
which has been mis-interpretted as
an acceleration toward the Sun

General Relativity

$$G_{ab} = \kappa T_{ab}$$

calculates for the total Gravitational and
EM interactions for energy in all its forms
[using Energy pressure gradients]

"Both Pioneer spacecraft are escaping the Solar System, and are slowing under the influence of the Sun's nett gravitational [GEM] field.

*Upon very close examination of navigational data, the spacecraft were found to be slowing slightly more than expected.
The effect is an extremely small but unexplained acceleration towards the Sun, of 8.74±1.33×10−10 m/s2"*

Tetryonics 76.07 - Pioneer anomaly

Copyright ABRAHAM [2008] - All rights reserved

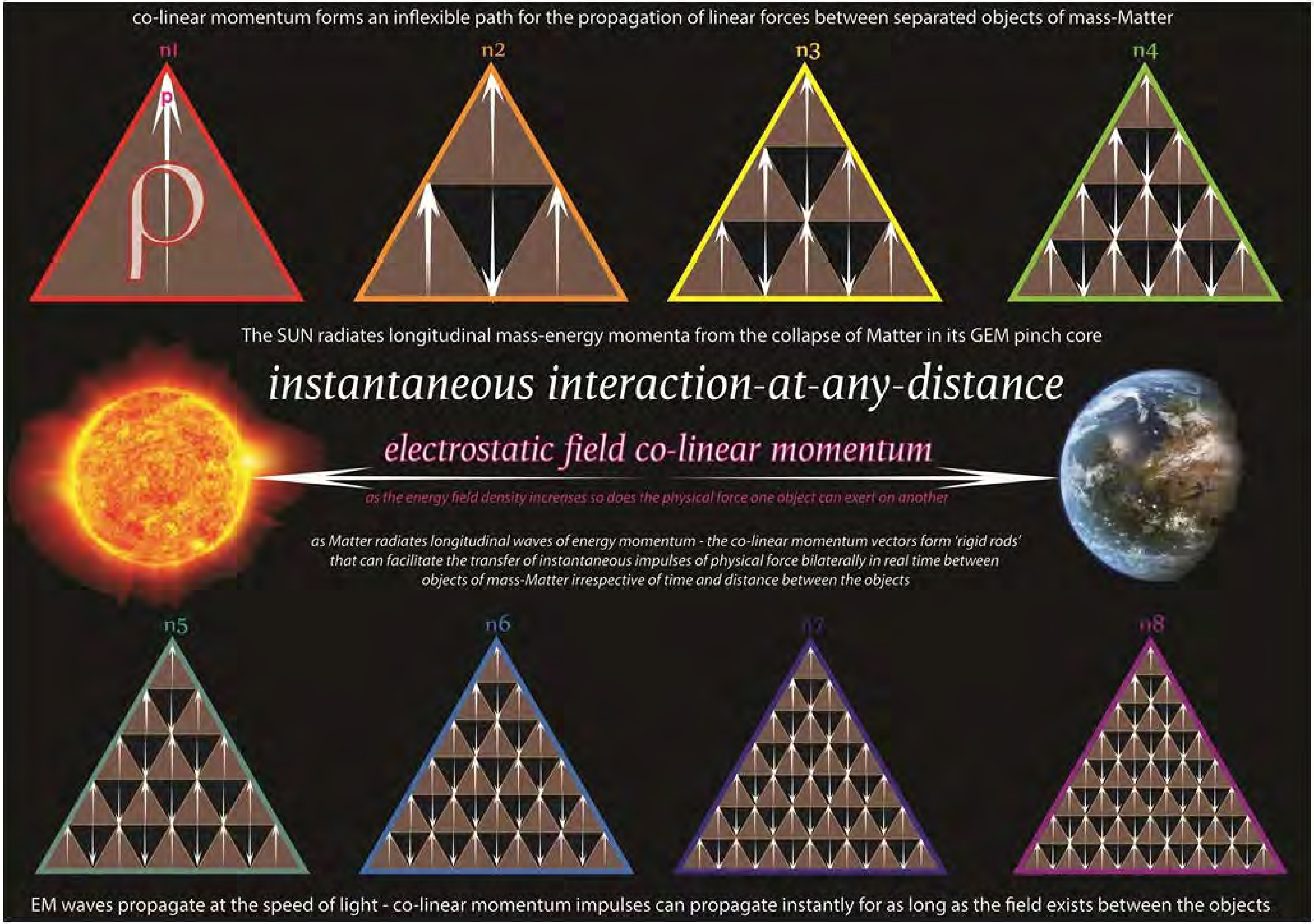

co-linear momentum forms an inflexible path for the propagation of linear forces between separated objects of mass-Matter

The SUN radiates longitudinal mass-energy momenta from the collapse of Matter in its GEM pinch core

instantaneous interaction-at-any-distance

electrostatic field co-linear momentum

as the energy field density increases so does the physical force one object can exert on another

as Matter radiates longitudinal waves of energy momentum - the co-linear momentum vectors form 'rigid rods' that can facilitate the transfer of instantaneous impulses of physical force bilaterally in real time between objects of mass-Matter irrespective of time and distance between the objects

EM waves propagate at the speed of light - co-linear momentum impulses can propagate instantly for as long as the field exists between the objects

Tetryonics 76.08 - Co-linear momentum [interaction-at-a-distance]

Instantaneous interaction-at-a-distance

One of the major points of disagreement between Newtonian Graviation and General Relativity concerns Einstein's limiting of information and energy momenta propagation to the speed of light whereas the Newtonian physics for Gravitation acts instantly between Matter irrespective of distance and time

partilces interact with others in real time irrespective of distance

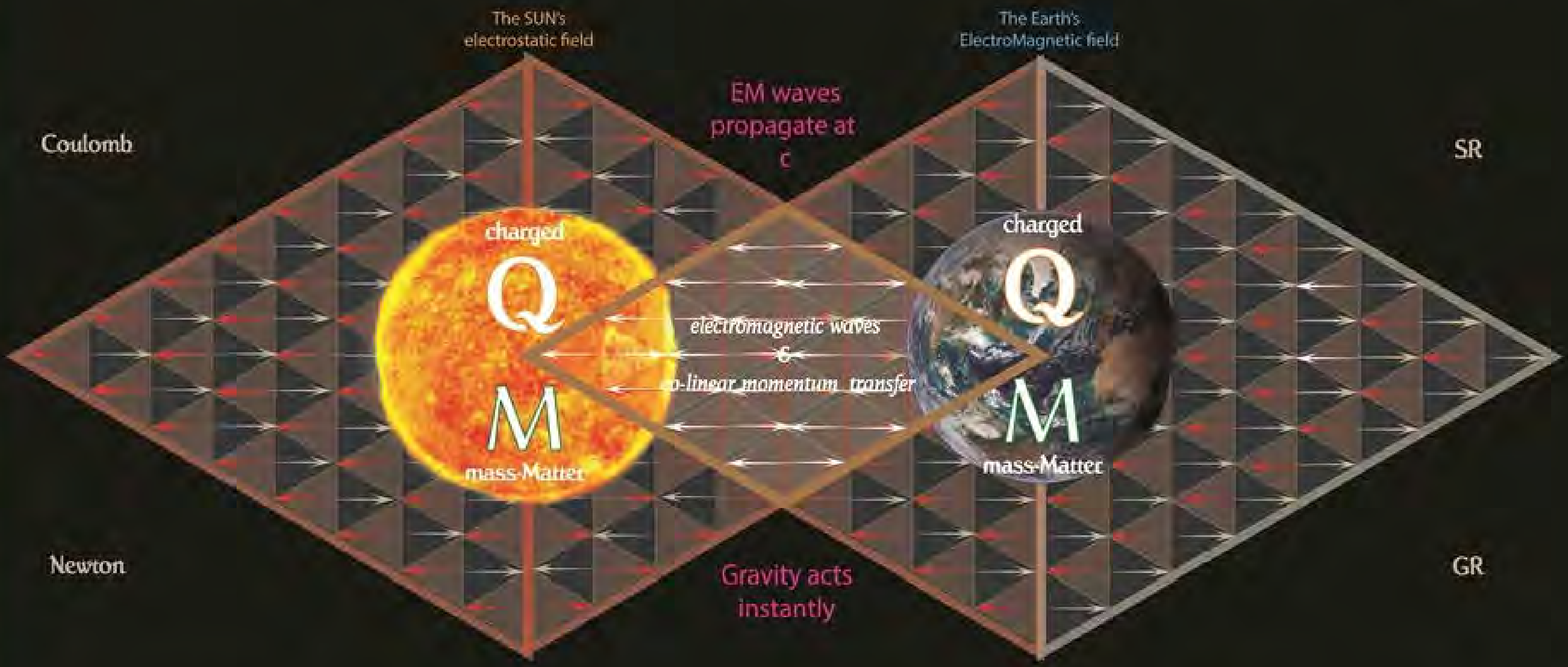
The SUN's electrostatic field
The Earth's ElectroMagnetic field
EM waves propagate at c
Coulomb
SR
charged
Q
M
mass-Matter
charged
Q
M
mass-Matter
electromagnetic waves & co-linear momentum transfer
Newton
GR
Gravity acts instantly

the flow of energy and information is limited to the speed of light

Where electrostatic and/or electromagnetic fields overlap they form fields of interaction between bodies of Matter

the 'quantum rods' of co-linear momentum in their EM fields facilitate 'FTL spooky interaction-at-a-distance' facilitating near instantaneous [real time] impulses of linear momentum to convey physical forces over vast distances without loss for as long as the fields exist between the objects

Copyright ABRAHAM [2008] - All rights reserved

The efficient wireless transmission of electrical energy requires the creation of a large electrostatic field of longitudinal energy momenta

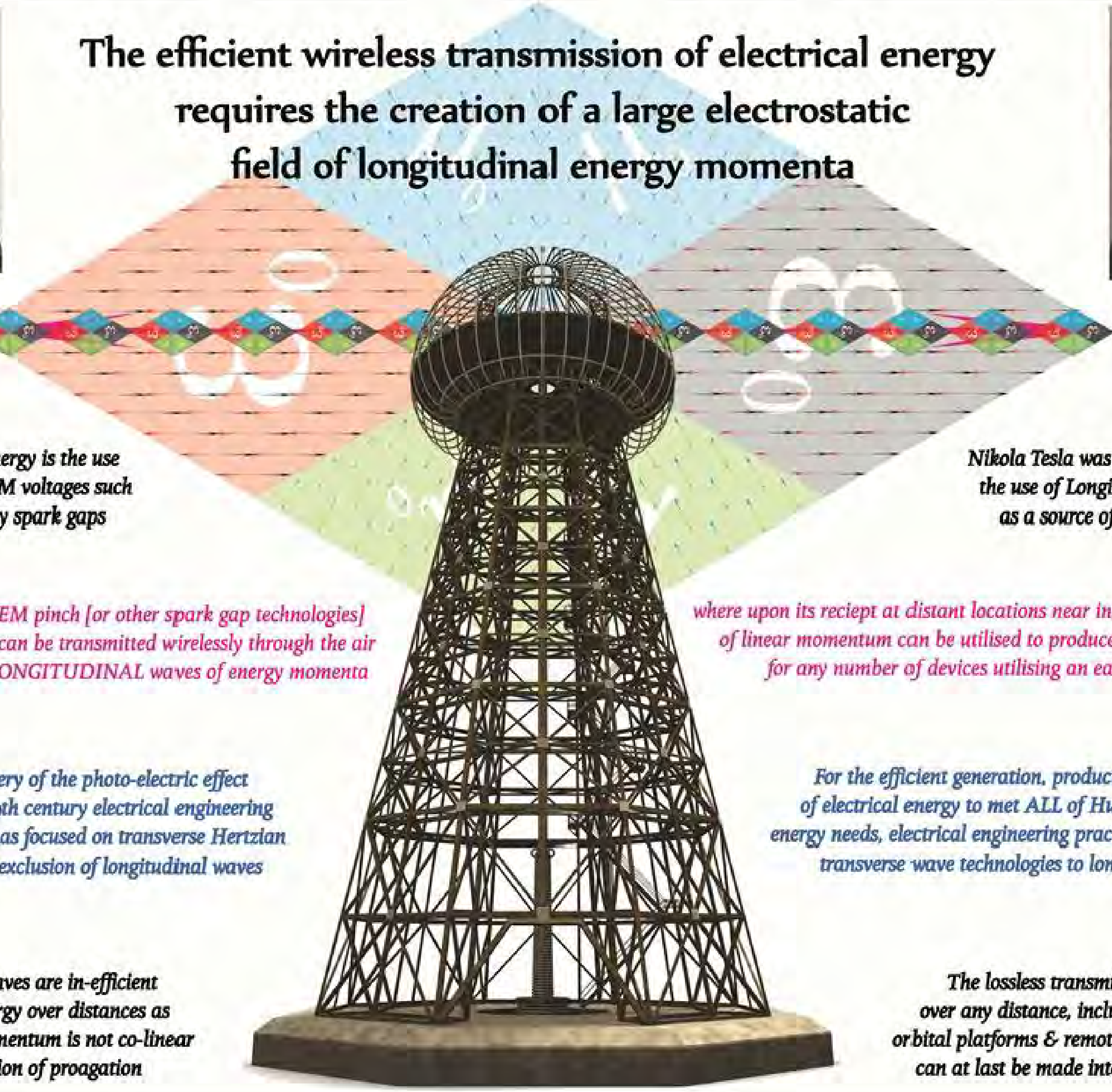

The key to wireless energy is the use of LONGITUDINAL EM voltages such as those produced by spark gaps

Nikola Tesla was the first to promote the use of Longitudinal EM waves as a source of wireless energy

Once generated by EM pinch [or other spark gap technologies] real electrical power can be transmitted wirelessly through the air without loss using LONGITUDINAL waves of energy momenta

where upon its reciept at distant locations near instanteous impulses of linear momentum can be utilised to produce electrical work for any number of devices utilising an earth return

Since the discovery of the photo-electric effect at the turn of the 19th century electrical engineering and physics theory has focused on transverse Hertzian waveforms to the exclusion of longitudinal waves

For the efficient generation, production and storage of electrical energy to met ALL of Humanities growing energy needs, electrical engineering practices must move from transverse wave technologies to longitudinal waves

Transverse EM waves are in-efficient for transmitting energy over distances as their linear energy momentum is not co-linear with their direction of proagation

The lossless transmission of power over any distance, including to and from orbital platforms & remote locations on Earth, can at last be made into a physical reality

Copyright ABRAHAM [2008] - All rights reserved

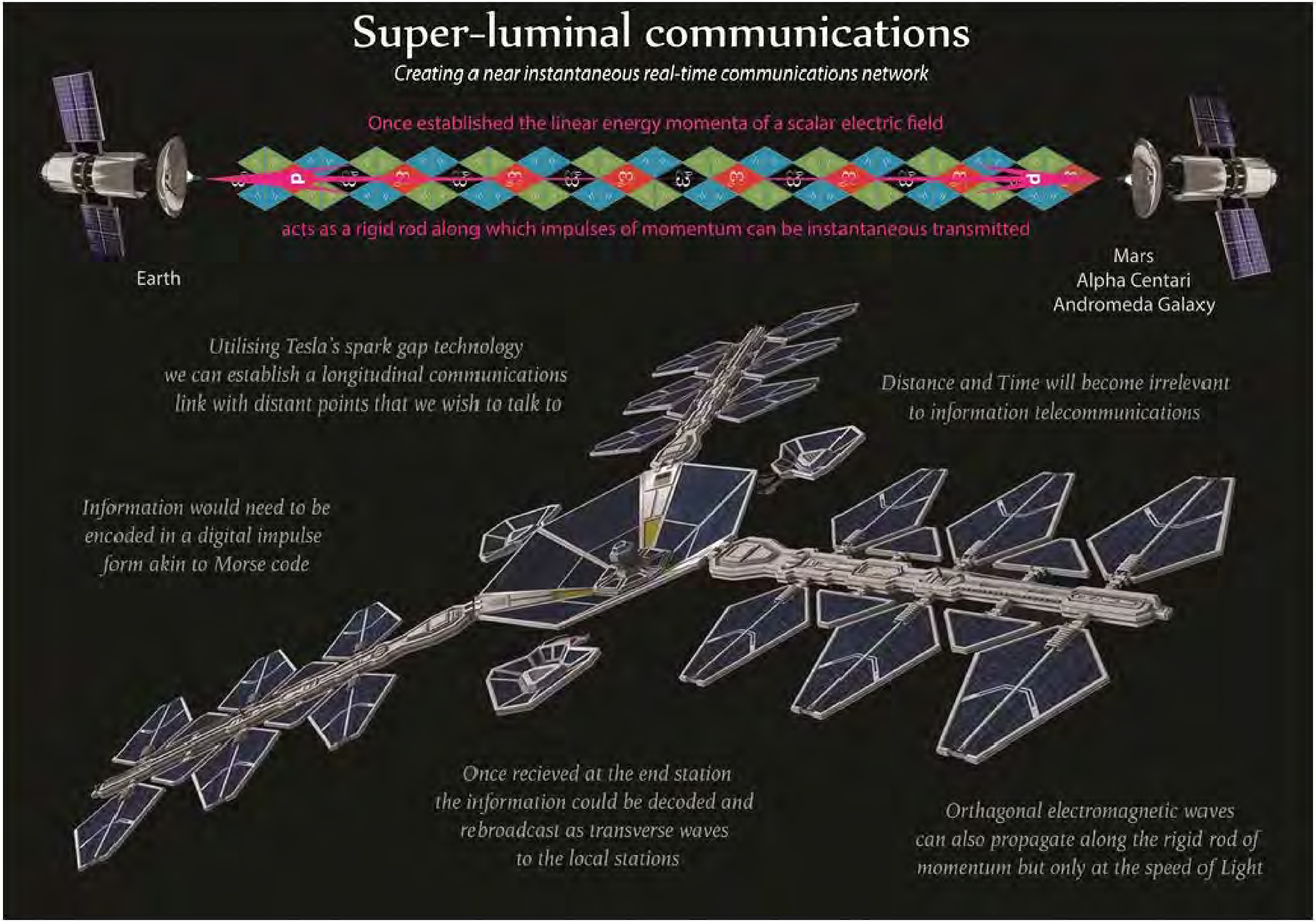

Super-luminal communications

Creating a near instantaneous real-time communications network

Once established the linear energy momenta of a scalar electric field

acts as a rigid rod along which impulses of momentum can be instantaneous transmitted

Earth

Mars
Alpha Centari
Andromeda Galaxy

Utilising Tesla's spark gap technology we can establish a longitudinal communications link with distant points that we wish to talk to

Distance and Time will become irrelevant to information telecommunications

Information would need to be encoded in a digital impulse form akin to Morse code

Once recieved at the end station the information could be decoded and rebroadcast as transverse waves to the local stations

Orthagonal electromagnetic waves can also propagate along the rigid rod of momentum but only at the speed of Light

Tetryonics 76.11 - FTL Communications

Copyright ABRAHAM [2008] - All rights reserved

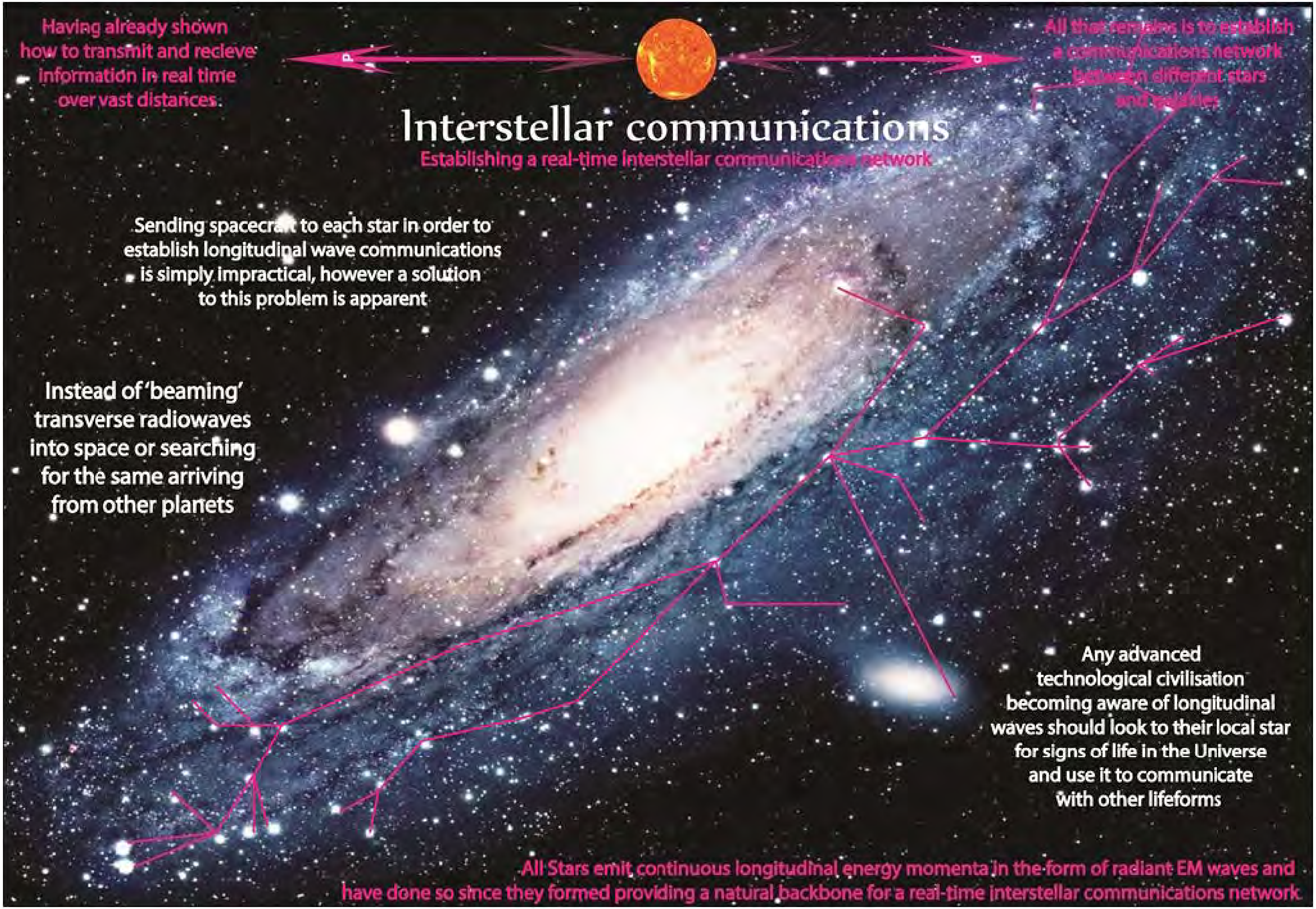

Having already shown how to transmit and recieve information in real time over vast distances

All that remains is to establish a communications network between different stars and galaxies

Interstellar communications

Establishing a real-time interstellar communications network

Sending spacecraft to each star in order to establish longitudinal wave communications is simply impractical, however a solution to this problem is apparent

Instead of 'beaming' transverse radiowaves into space or searching for the same arriving from other planets

Any advanced technological civilisation becoming aware of longitudinal waves should look to their local star for signs of life in the Universe and use it to communicate with other lifeforms

All Stars emit continuous longitudinal energy momenta in the form of radiant EM waves and have done so since they formed providing a natural backbone for a real-time interstellar communications network

Tetryonics 76.12 - Interstellar communication

Mercurial anomalies

Mercury has a weak magnetic field

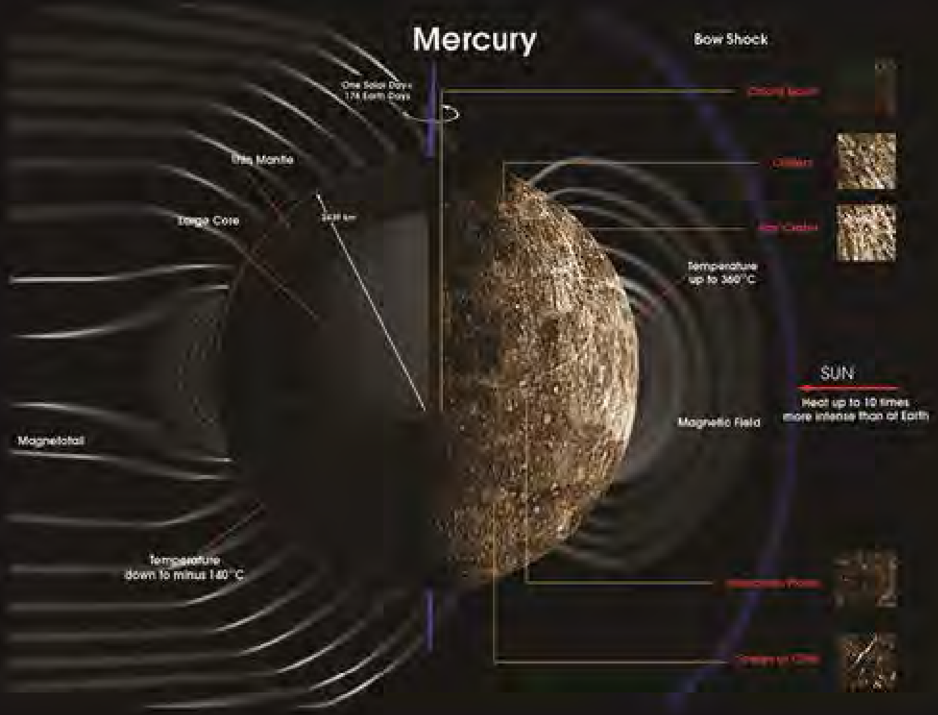
Mercury
Bow Shock
One Solar Day=
176 Earth Days
Thin Mantle
Large Core
2439 km
Temperature
up to 360°C
SUN
Heat up to 10 times
more intense than at Earth
Magnetic Field
Magnetotail
Temperature
down to minus 140°C

Mercury has heat of its own, not just reflected heat of the Sun;

Mercury has still an atmosphere of Hydrogen, possibly the last vestiges of a more extensive halo and trail (caduceus)

Mercury rotates because it is in its present orbit it is strongly affected by the EM field of the SUN,

Because the ElectroMagnetic forces near the Sun need to be accounted for in the motion of any satellite about a more massive body [ie Satellite-Planet or Planet- Sun], the stronger E & perturbative M forces are responsible for the observed precession of the perihelion of Mercury, and Leverrier's discovery of this precession can be modelled with Einsteins' General relativity through Tetryonic geometries

Laplace showed that should a celestial body attracted by its primary as inverse square of distance be subject to another attraction that changes as the inverse cube of distance, a precession by that body would result.

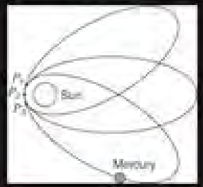
P_1
P_2
P_3
Sun
Mercury

Einstein corrected Newton's gravitational results by calculating for ALL the energies in a system but failed to realise that his 'graviational correction' was in fact a Lorentz correction for EM interactions

Copyright ABRAHAM [2008] - All rights reserved

Venusian anomalies

is sometimes referrd to as Earth's "twin" owing to their similar size, gravity, and bulk composition.

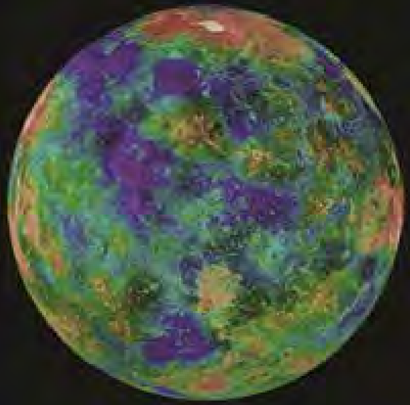

Venus has the densest atmosphere of all the terrestrial planets
in the Solar System, consisting of mostly carbon dioxide.
The atmospheric pressure at the planet's surface is 92 times that of the Earth.

**Venus is the hottest world in the solar system.
Although Venus is not the planet closest to the sun,
its dense atmosphere traps heat in a runaway version
of the greenhouse effect that warms up the Earth.**

As a result, temperatures on Venus reach 465 degrees C

The lack of an intrinsic magnetic field at Venus was surprising
given it is similar to Earth in size, and was expected also to
contain a dynamo at its core

Venus has no moons or rings

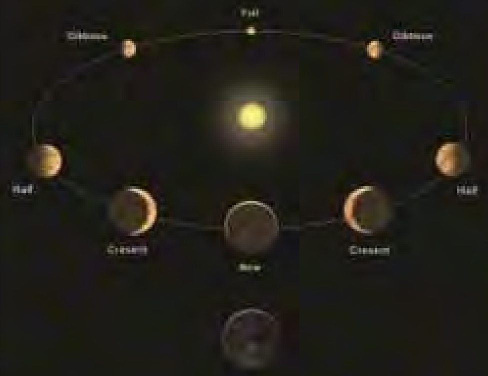

Galileo's discovery that Venus showed phases
(while being never very far from the Sun in our sky)
proved that it orbits the Sun and not the Earth

The Venusian year is about 225 Earth days long.

Venus takes 243 Earth days to rotate on its axis,
which means that days on Venus would be longer
than years. However, because of Venus' curious
retrograde rotation, the time from one sunrise
to the next is only about 117 Earth days

The very top layer of Venus' clouds zip around the planet
every four Earth days, propelled by hurricane-force winds
traveling roughly 224 miles (360 kilometers) per hour.

This super-rotation of the planet's atmosphere, some 60 times faster
than Venus itself rotates, may be one of Venus' biggest mysteries.

Earth-Moon anomalies

The Moon is slowly accelerating away from the Earth

Athough it is attracted gravitationally to the Earth the Moon is moving away from the Earth at about 4cm per year as a result of the Earth's EM field interactions with Matter in its vicinity.

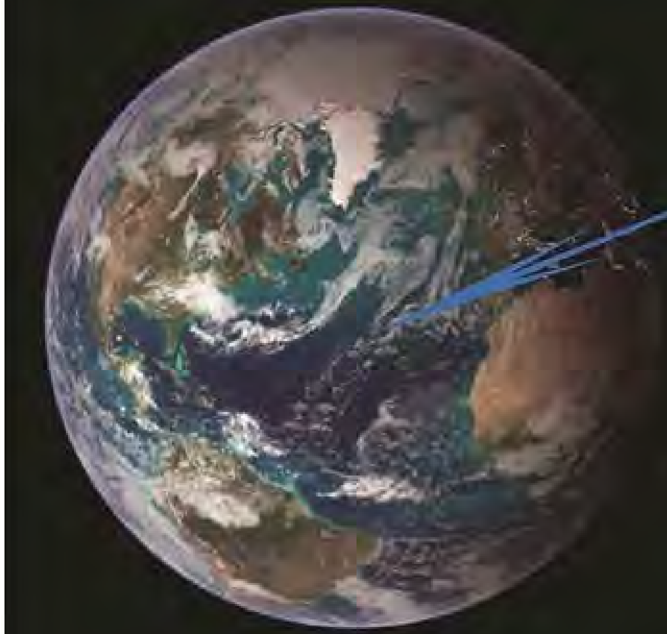

The Earth's EM field accelerates the moon against the pull of gravity stabilising its orbit

its current field strength forces the moon to slowly accelerate away from the Earth over time

The Moon's Origin

Most scientists think the moon was born from a gargantuan collision, when a young, 30-million-year-old Earth was sideswiped by an embryonic planet the size of Mars some 4.5 billion years ago, with debris from our planet and this impactor eventually coalescing from the ejecta to form our Moon.

The Lunar a lunar day is a month long, the same as the Moon's orbital period

The Moon only shows one face to the Earth

it is gravitational locked in its orbit with one side always facing the Earth when viewed from the Earthat night. There is no such thing as the "dark side of the moon". The sun shines on all sides

The Moon is 400 times smaller than the SUN
The SUN is 400 times more distant

The Moon looks exactly the same size as the Sun

"There is no astronomical reason why the moon and the sun should fit so well. It is the sheerest of coincidences, and only the Earth among all the planets is blessed in this fashion." ....Isaac Asimov

A striking coincidence that has had astronomers thinking since ancient times is why the Sun and the Moon appear to be the same size during an eclipse. It's known that the Sun's diameter is 400 times that of the Moon, but, then again, the Moon is 400 times further away from the Sun.

This creates the optical illusion that the two are actually the same size, allowing us to observe the Solar corona and its EM field lines

Our moon is the only moon in the solar system that has a stationary, near-perfect circular orbit. Stranger still, the moon's center of mass is about 6000 feet closer to the Earth than its geometric center (which should cause wobbling), but the moon's bulge is on the far side of the moon, away from the Earth. "Something" had to put the moon in orbit with its precise altitude, course, and speed.

Copyright ABRAHAM [2008] - All rights reserved

Martian anomalies

Phobos

There is no completely satisfactory theory as to how Deimos and Phobos came to be in orbit around Mars

With a dametre of 6 km Deimos
is one of the smallest known moons
in the solar system

Deimos

Although Mars is much smaller than Earth, [approx 1/4 the size]
its surface area is about the same as the land surface area of Earth

Deimos and Phobos are probably asteroids perturbed
by Jupiter into orbits that allowed them to be captured by Mars

The southern hemisphere of Mars
is predominantly ancient
cratered highlands somewhat
similar to the Moon.

In contrast, the northern hemisphere
consists of plains which are much younger,
lower in elevation and have a much
more complex history.

Mars orbits about
1.5au from the SUN

An abrupt elevation change of several kilometers seems to occur at the boundary

Copyright ABRAHAM [2008] - All rights reserved

Jupiters' anomalies

Jupiter has an intense and far-reaching magnetic field.
Exactly what drives the magnetic powerhouse and
how deep inside the planet it originates, however, is unknown

Jupiter has twelve satellites.

The five inner moons revolve around their planet in orbits only slightly inclined to the planet's equator at distances from about 110,000 miles for the innermost to about 640,000 miles for the outermost.

Jupiter's diameter is approx
1/10 that of the SUN and
10 times the size of Earth

Jupiter's mass is 2.5 times that of all the other planets in our Solar System combined

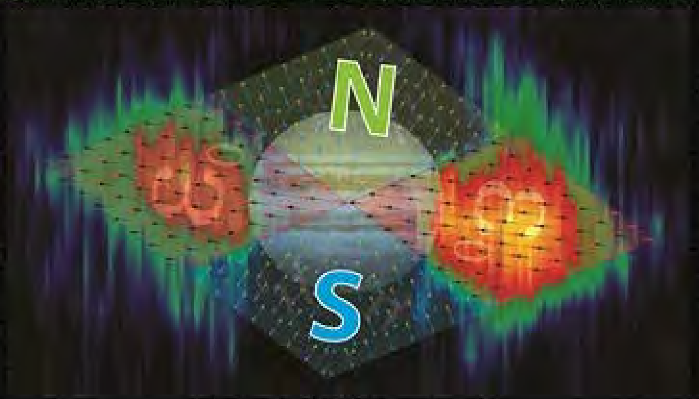

Tetryonic EM geometry
offers answers to many of the dynamic
mysteries of Jupiter and its moons

Then there is a group of three moons whose orbits are inclined to the planet's equator by almost 30° at distances of about 7 million miles from Jupiter.

These three moons also revolve around the planet in the predicted direction.

The four outer moons, however, move around the planet in retrograde motion, or opposite to that of the other eight satellites, at distances from about 12 to 13 million miles from the planet

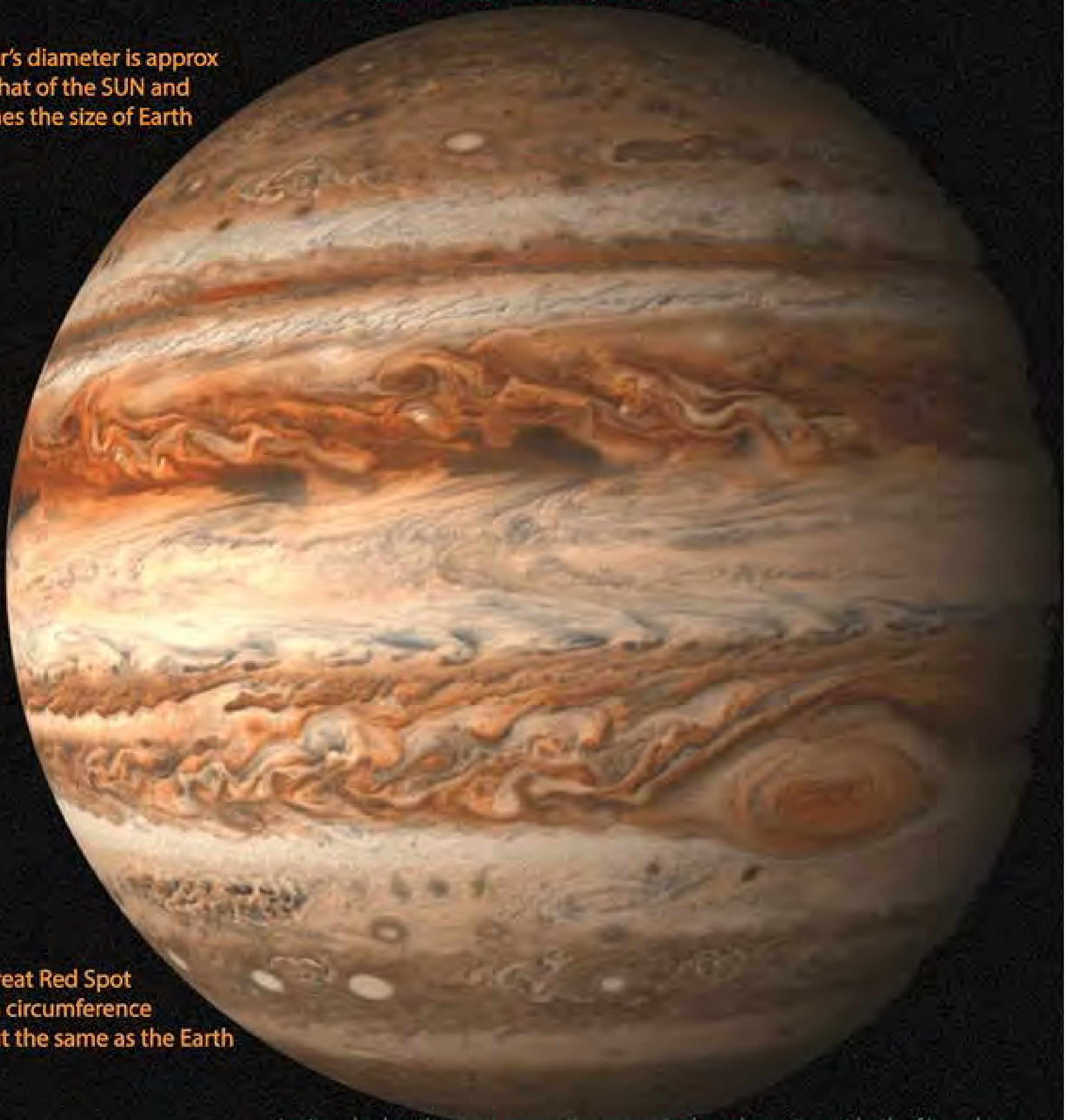

Its Great Red Spot
has a circumference
about the same as the Earth

Jupiter also has the most energetic auroras in the solar system at both of its poles

Copyright ABRAHAM [2008] - All rights reserved

Saturnian anomalies

In addition to the planet's giant rings, Saturn has nine major moons and more than 60 known satellites.

The motion of the outermost, Phoebe, is retrograde, moving in a direction opposite to the other eight moons and opposite to that predicted from an evolutionary origin.

Tetryonic theory suggests that this retrograde motion is the result of it being a gravitationally captured body whose motion over time will be slowed, eventually forcing it into the same motion as the others

TITAN
is the only moon that
possesses a thick atmosphere, and
it's the only place beyond Earth to have
stable bodies of liquid Methane on its surface

The rings of Saturn occasionally exhibit ghostly radial spokes that seemingly defy the laws of gravity

The Saturnian rings, like many others, are formed by the electric field of the planet and subject to seasonal fluctuations of the SUN's heliosheric E-field

Tetryonic GEM interactions shows that these features are the result of electric forces additional to gravity

Strange features abound within the rings of Saturn. These include hundreds of "record grooves" or narrow fluctuations in the B ring, the most massive ring.

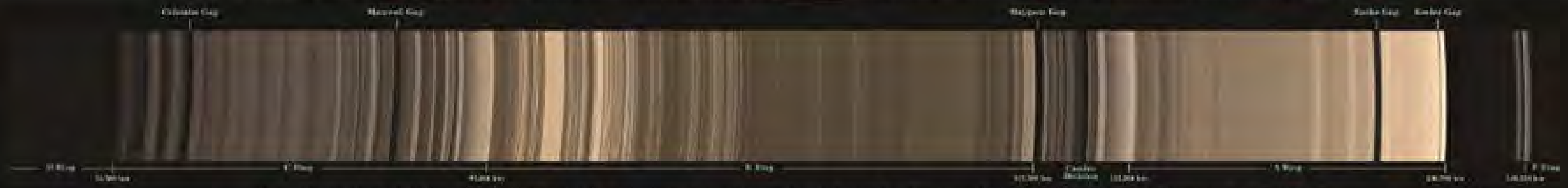

These also include "plateaus" in the C ring, the innermost main ring—broad, dark, sharp-edged features thicker than elsewhere in the ring.

Copyright ABRAHAM [2008] - All rights reserved

Neptunes' anomalies

Neptune was the first planet found by mathematical prediction rather than by empirical observation

Neptune has two satellites.

Nereid, a small moon, moves around Neptune in the predicted direction, but Triton, one of the larger satellites in the solar system with a mass almost twice that of the earth's moon.

Of Neptune's 13 moons, Triton is by far the biggest and the only one massive enough to be spheroidal.

Weirdly, Triton has a "retrograde" orbit, revolving in the opposite direction of the planet and other moons.

Plus, the orbit is at an angle rather than in the plane around the equator like typical satellites

The magnetic field of Neptune, like that of Uranus, is highly tilted at 47 degrees from the rotation axis and offset at least 0.55 radii (about 13,500 kilometers or 8,500 miles) from the physical center

Uranus' anomalies

Uranus is remarkable, even though it rotates around the sun in the same direction as the other planets in an orbit inclined less than a degree (46') from the ecliptic (the plane of the earth's orbit around the sun), the axis of rotation of Uranus is nearly in the plane of its orbit.

Accordingly, the inclination of the equator of Uranus to the plane of its orbit is 98°, and its axial rotation is retrograde.

The five moons or satellites of Uranus move exactly in the equatorial plane of the planet and they revolve in the same direction as the planet rotates.

Their motion, with respect to the remainder of the solar system, is, therefore, also retrograde.

Tetryonic theory dictates that Uranus' peculiar planetary mechanics may be the result of the EM interactions of Uranus itself with its satellites rather than the current evolutionary explanation

The direction of the axial rotation of Uranus and the motion of its satellites is opposite to that predicted on the basis of an evolutionary origin.

Copyright ABRAHAM [2008] - All rights reserved

Pluto's anomalies

Pluto is much smaller than any of the official planets
and due to its size and orbital inclination
is now classified as a "dwarf planet"

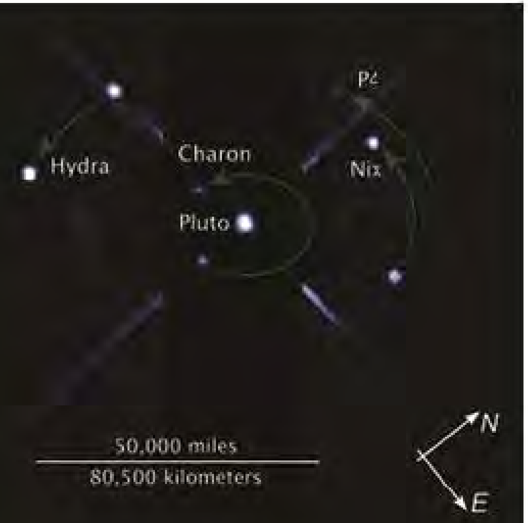
P4
Hydra
Charon
Nix
Pluto
50,000 miles
80,500 kilometers
N
E

Pluto rotates in the opposite direction
to that of most of the other planets

Pluto is locked in a 3:2 resonance with Neptune;
i.e. Pluto's orbital period is exactly 1.5 times longer than Neptune's.
Its orbital inclination is also much higher than the other planets

Charon is unusual in that it is
the largest moon with respect to its
primary planet in the Solar System
(a distinction once held by Earth's moon)

Pluto's orbit is highly eccentric.
At times it is closer to the Sun than Neptune

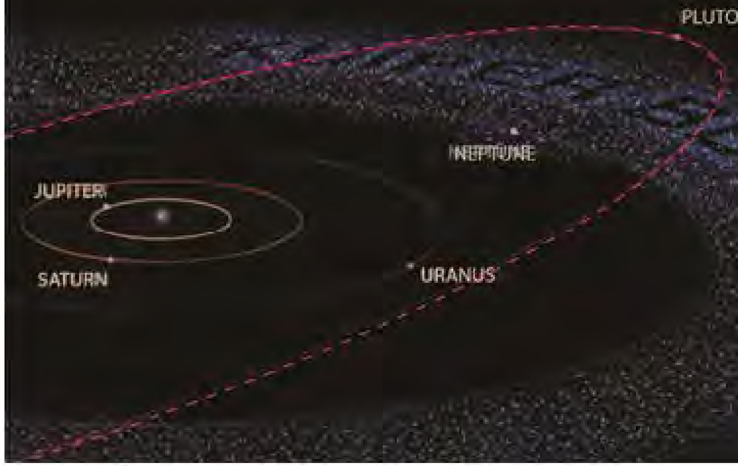
PLUTO
NEPTUNE
JUPITER
SATURN
URANUS

The Pluto–Charon system is noteworthy
for being one of the Solar System's
few binary system

Pluto
Nix
Hydra
Charon

Cometary anomalies

Comets provide some of the clearest evidence of Electromagnetic interactions between the SUN and other planetary bodies

While the solid nucleus of comets is generally less than 50 km across, the coma may be larger than the Sun, and Ion tails have been observed to extend 3-4 astronomical units (5-600 million km)

Cometary discharging may also occur due to any disturbances of its electrical plasma sheath as it passes through regions of varying electric potential.

This seems to have occurred in the recent "totally surprising" outburst of Comet Holmes 17P as it moved away from the Sun's domain.

In 2000 the explosive break up of Comet Linear provoked great amazement as the event occurred well over a hundred million kilometers from the Sun

Other anomalies that can be explained by a CHARGED model of cometary nuclei are: Unexpectedly high temperatures and X-ray emissions from cometary comas

The sharply carved relief of comets [the exact opposite of what is expected under the "dirty snowball" model]

The unexplained ability of a relatively minuscule comet nucleus to hold in place a highly spherical coma, up to millions of miles in diamater, against the force of the solar wind;

Ejection of larger particles and "gravel" that was never anticipated under the idea that comets accreted from primordial clouds of ice, gas, and dust;

A short supply or complete absence of water and other volatiles on comets' nuclei;

The Electric Universe's model of charged solar bodies predicted the occurrence of an electrical flash in advance to the impact of a projectile into the nucleus of Comet Tempel 1 (Deep Impact.)

**Comets have two tails:
A Dust tail and
A Ion tail
most recently observed
in comet Hale-Bopp**

The Dust tail is the result of reflective ejected Matter following the comet's path throught solar system while the Ion tail is the result of ejected charged particles of Matter and is heavily influenced by the SUN's radiated E fields

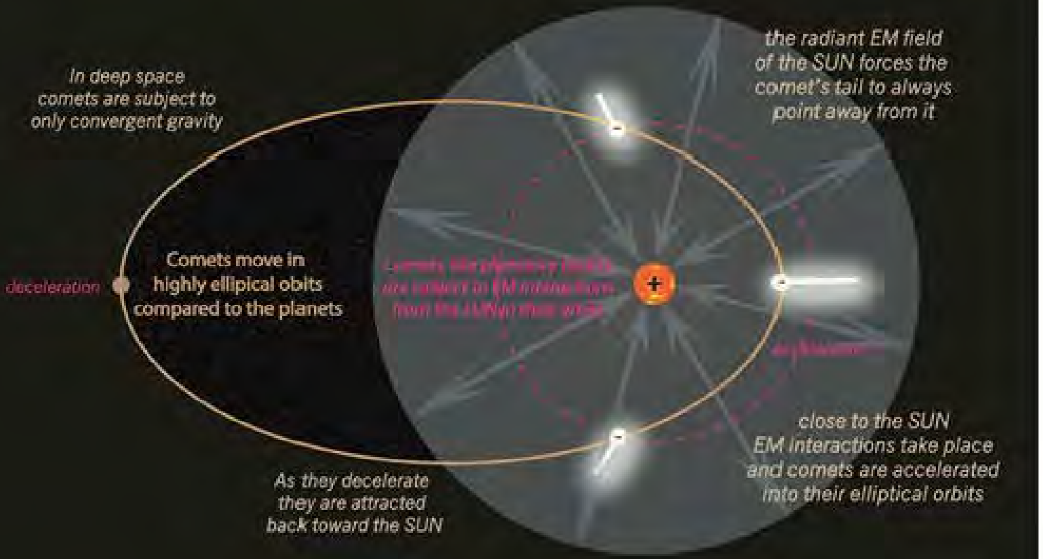

In deep space comets are subject to only convergent gravity

the radiant EM field of the SUN forces the comet's tail to always point away from it

Comets move in highly ellipical obits compared to the planets

close to the SUN EM interactions take place and comets are accelerated into their elliptical orbits

As they decelerate they are attracted back toward the SUN

The tails of comets always point away from the Sun, so after a comet has passed the Sun it actually travels tail first

Copyright ABRAHAM [2008] - All rights reserved

The Plasma universe

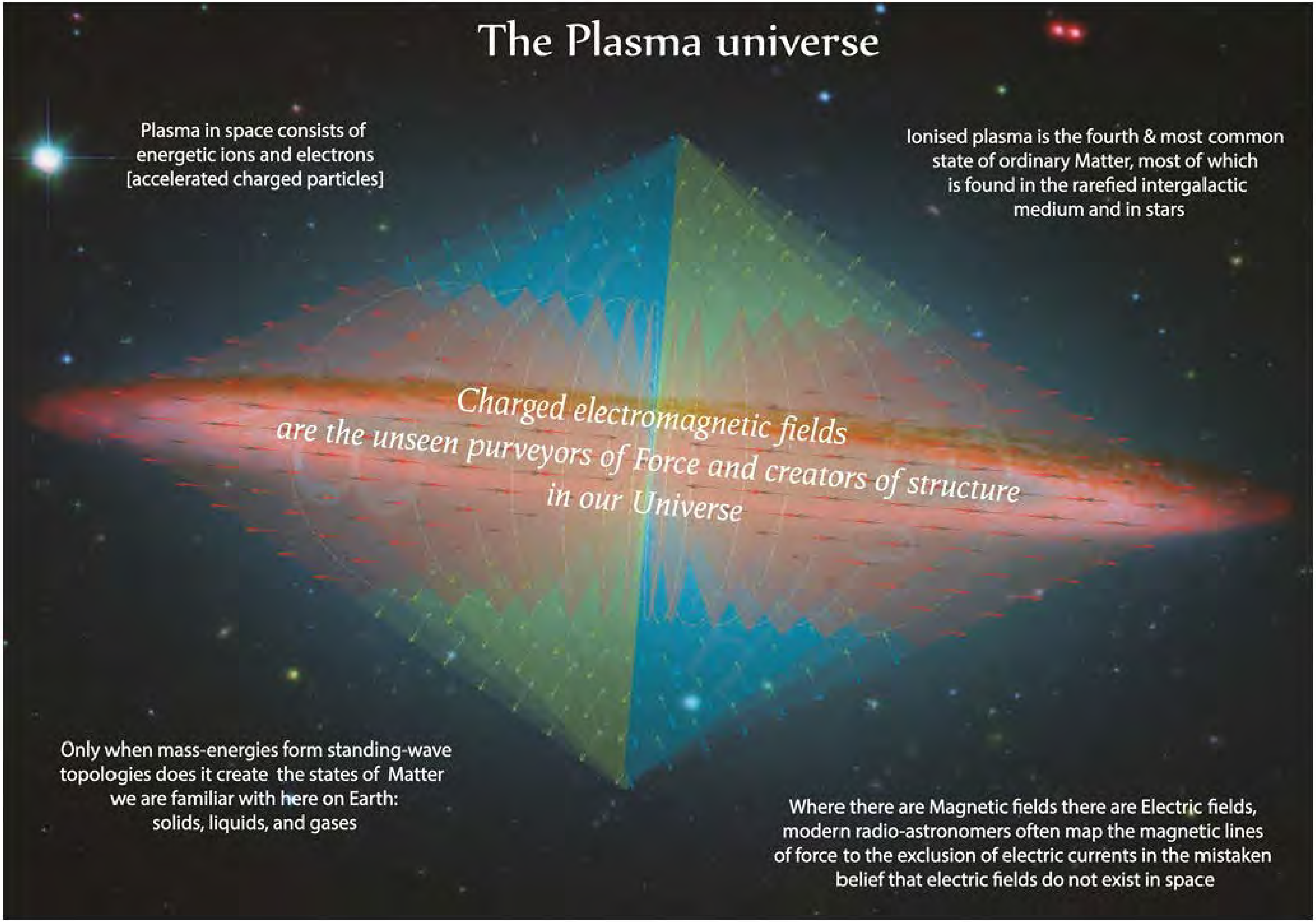

Plasma in space consists of
energetic ions and electrons
[accelerated charged particles]

Ionised plasma is the fourth & most common
state of ordinary Matter, most of which
is found in the rarefied intergalactic
medium and in stars

*Charged electromagnetic fields
are the unseen purveyors of Force and creators of structure
in our Universe*

Only when mass-energies form standing-wave
topologies does it create the states of Matter
we are familiar with here on Earth:
solids, liquids, and gases

Where there are Magnetic fields there are Electric fields,
modern radio-astronomers often map the magnetic lines
of force to the exclusion of electric currents in the mistaken
belief that electric fields do not exist in space

Copyright ABRAHAM [2008] - All rights reserved

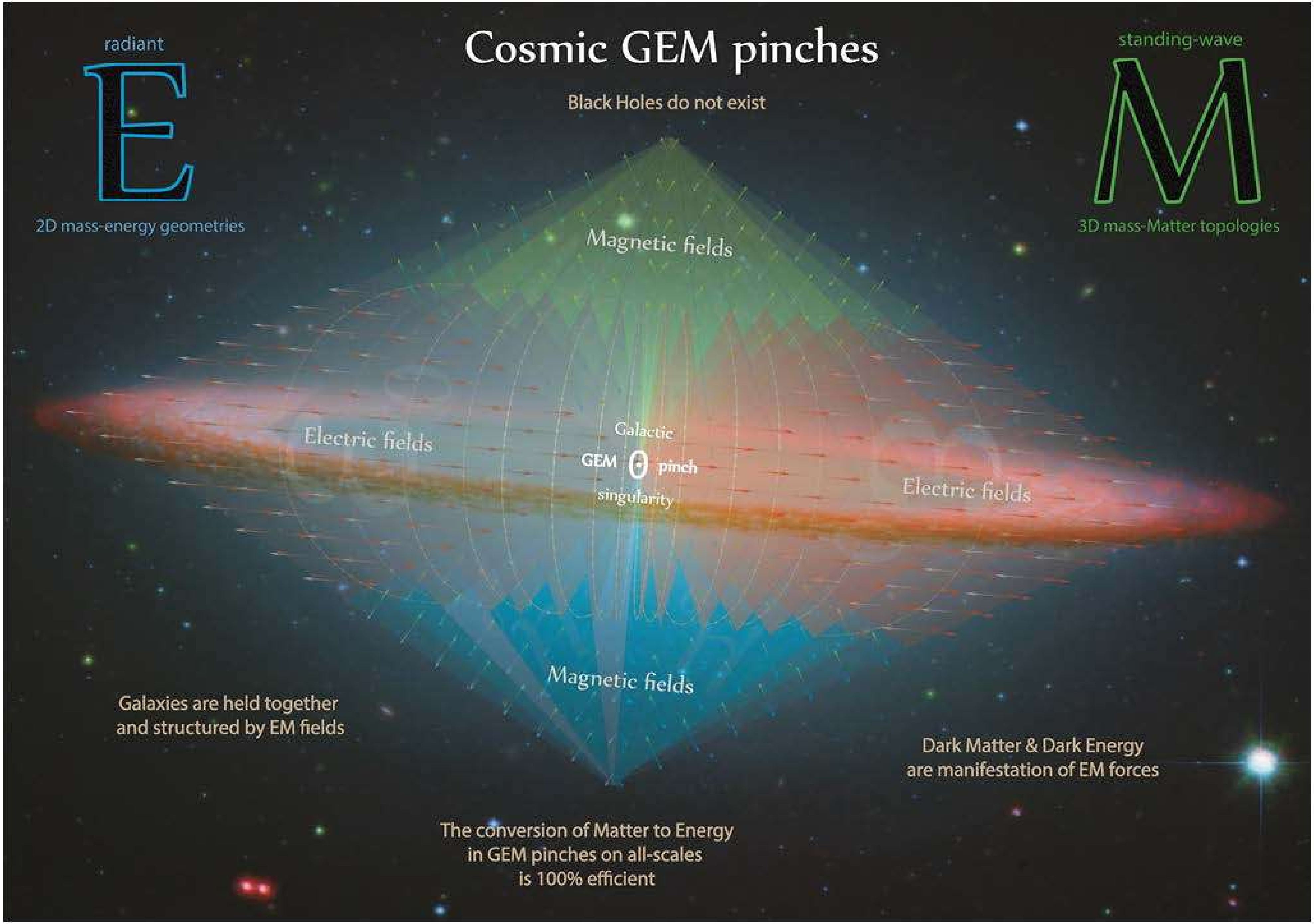

Cosmic GEM pinches

Black Holes do not exist

radiant
E
2D mass-energy geometries
standing-wave
M
3D mass-Matter topologies
Magnetic fields
Electric fields
Galactic
GEM
pinch
singularity
Electric fields
Magnetic fields

Galaxies are held together and structured by EM fields

Dark Matter & Dark Energy are manifestation of EM forces

The conversion of Matter to Energy in GEM pinches on all-scales is 100% efficient

Tetryonics 78.02 - GEM pinches

Copyright ABRAHAM [2008] - All rights reserved

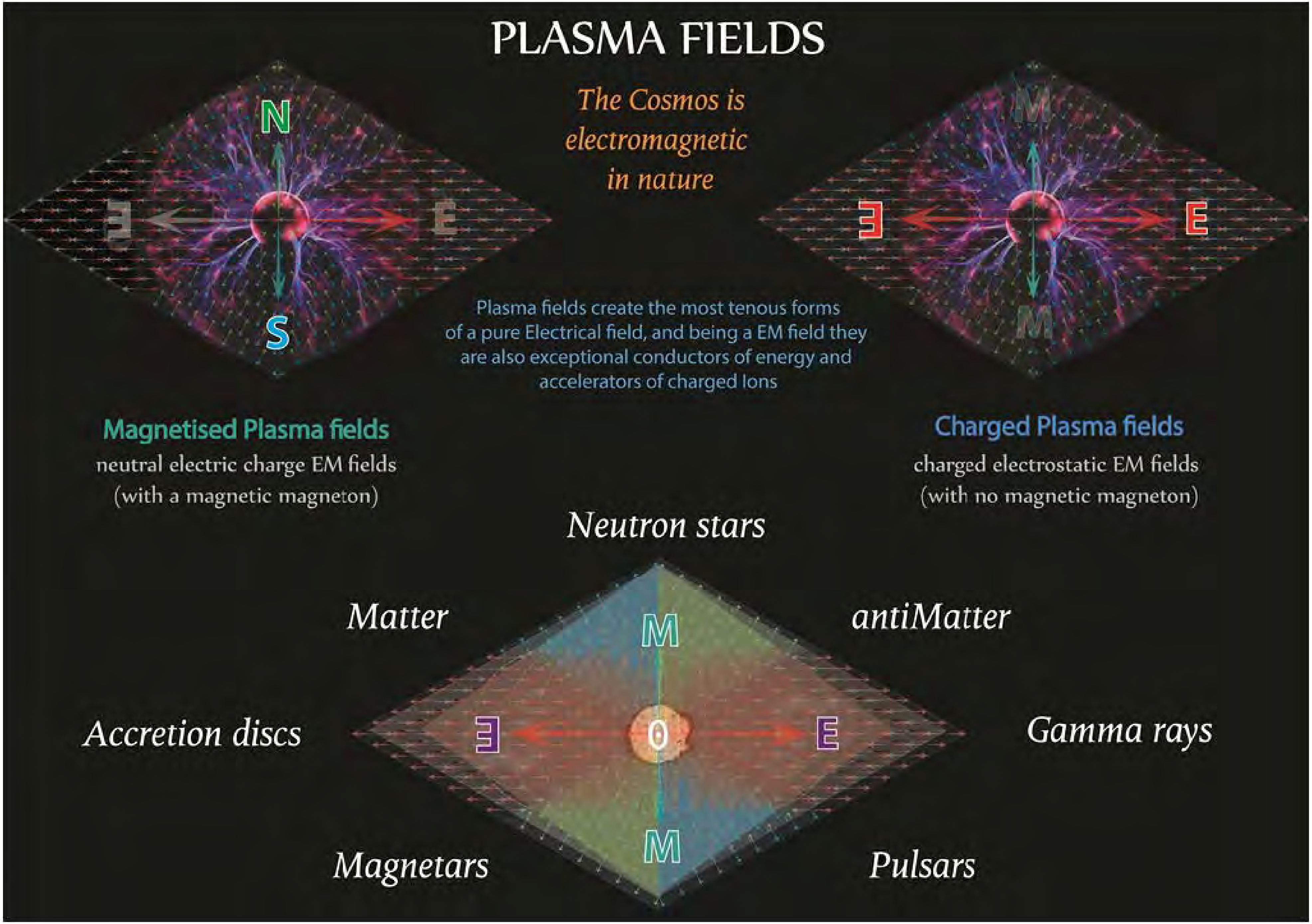

PLASMA FIELDS

The Cosmos is electromagnetic in nature

Plasma fields create the most tenous forms of a pure Electrical field, and being a EM field they are also exceptional conductors of energy and accelerators of charged Ions

Magnetised Plasma fields
neutral electric charge EM fields
(with a magnetic magneton)

Charged Plasma fields
charged electrostatic EM fields
(with no magnetic magneton)

Neutron stars

Matter

antiMatter

Accretion discs

Gamma rays

Magnetars

Pulsars

Tetryonics 78.03 - Plasma Fields

Neutron Stars

*All Gravitational fields
are comprised of quantum
GEM fields*

In current astrophysical theories in order to create a

Neutron star

the central region of the star collapses under gravity.

*It collapses so much that protons and electrons
combine to form neutrons. and neutrogenic Matter*

Hence the name "neutron star".

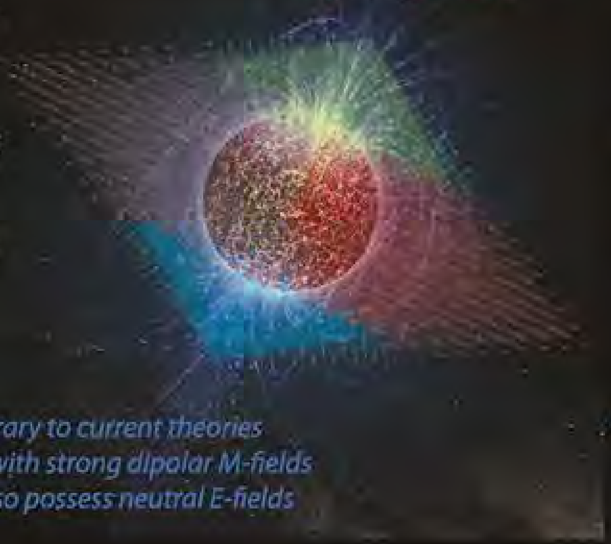

*Typical Neutron stars have
very strong M-fields*

*Contrary to current theories
any star with strong dipolar M-fields
must also possess neutral E-fields*

"A neutron star is a type of remnant that can result from the
gravitational collapse of a massive star during a Type II,
Type Ib or Type Ic supernova event.

Such stars are composed almost entirely of neutrogenic Matter,
[Neutrons, antiNeutrons and neutrinos] which against
current theory act in exactly the same manner as
ordinary Matter in producing spectral lines.

*"Neutron stars are very hot and are supported against further
collapse because of the Pauli exclusion principle.*

*This principle states that no two neutrons (or any other fermionic
particle) can occupy the same place and quantum state simultaneously""*

*As shown throughout Tetryonic theory
Neutron stars can be formed
comprised completely of 'Neutral Matter'*

*Neutral Matter is comprised of
Neutron - anti-Neutron - Neutrino
atomic nuclei*

*Neutral matter atoms function identically to
normal matter-antimatter particle topologies
and releases energy in the form of photons*

*The gravitational collapse of any
stellar material must overcome
the radiant EM energies that it
releases*

*Pulsars are non-neutral stellar
objects that have weak M-fields
but emit extremely strong E-field
radiation as they collapse*

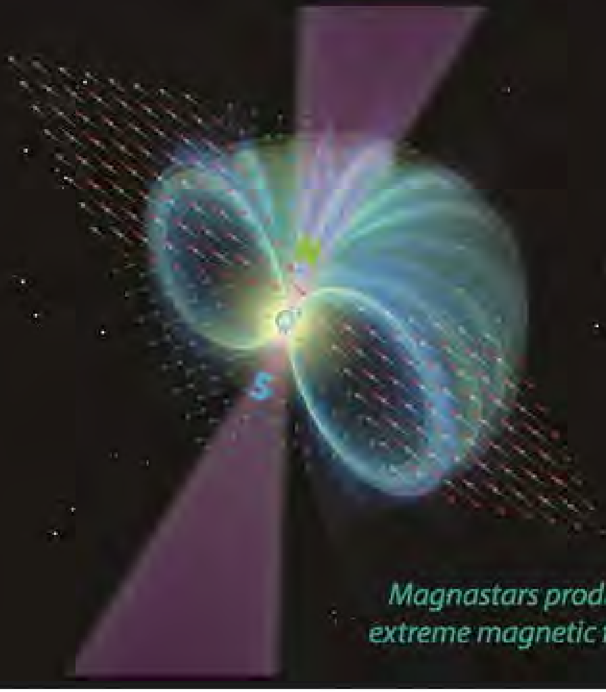

*Magnastars produce
extreme magnetic fields*

*Neutral matter in Neutron stars experience the same ElectroMagnetic forces as
normal matter and anti-matter stars but have differing charge arrangements*

Copyright ABRAHAM [2008] - All rights reserved

Neutrons &

Neutronium formation

neutrinos

Neutronium is a proposed name for a substance composed entirely of Neutrons & neutrinos

Neutron degeneracy

As a star collapses, the Fermi energy of the electrons increases to the point where it is energetically favorable for them to combine with protons to produce neutrons (via inverse beta decay, also termed "neutralization" and electron capture)

This, as shown through Tetryonic geometry, is incorrect

Neutrons are NOT comprised of Protons with electrons fused into them under pressure

Gravitational fields are comprised of convergent gravity and interactive EM fields

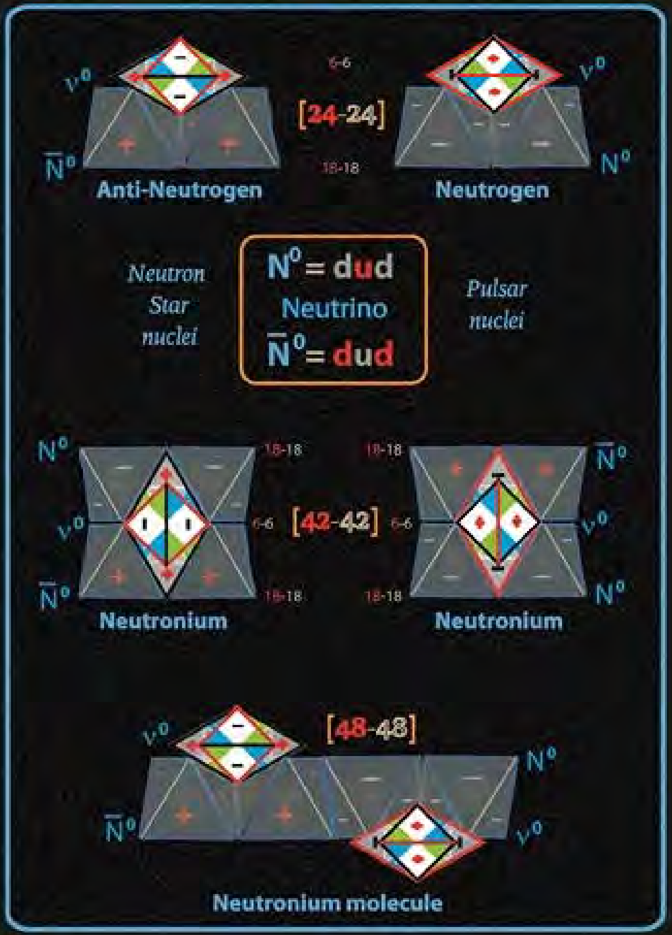
ν^0
$\bar{N}^0$
Anti-Neutrogen
6-6
[24-24]
18-18
ν^0
N^0
Neutrogen
Neutron Star nuclei
N^0= dud
Neutrino
$\bar{N}^0$= dud
Pulsar nuclei
N^0
ν^0
$\bar{N}^0$
18-18
6-6
[42-42]
6-6
18-18
18-18
18-18
Neutronium
$\bar{N}^0$
ν^0
N^0
Neutronium
ν^0
[48-48]
N^0
$\bar{N}^0$
ν^0
Neutronium molecule

"The gravitational field at the star's surface is about 2×10^11 times stronger than on Earth.

Such a strong gravitational field acts as a gravitational lens and bends the radiation emitted by the star such that parts of the normally invisible rear surface become visible"

Photons and EM waves are REFACTED by EM fields NOT bent by gravity as currently hypothesized

Neutronium

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\overset{\text{quanta}}{\underset{\text{velocity}}{v}}^{2}\right]\right]$$

The neutral nuclei of neutron stars emit exactly the same light as normal stars do

Neutron stars are made of neutronium nuclei

A neutron star is a type of stellar remnant that can result from the gravitational collapse of a massive star during a Type II, Type Ib or Type Ic supernova event.

Such stars are composed almost entirely of neutrons, which are subatomic particles without electrical charge and with the same mass-Matter as protons. Neutron stars are very hot and are supported against further collapse by quantum degeneracy pressure due to the Pauli exclusion principle.

This principle states that no two neutrons (or any other fermionic particles) can occupy the same place and quantum state simultaneously

N^0

ν^0

Tetryonics 78.05 - Neutronium formation

Copyright ABRAHAM [2008] - All rights reserved

Equilateral Planck geometries create elementary charge topologies

Neutron star types

Neutron stars comprise one of the possible evolutionary end-points for stars,
they are compact stellar objects measuring between 10 and 20 km across
with intense magnetic fields around 10^12 times that of the Earth

Like all forms of Matter, neutron stars are formed from 3 distinct charge flavours of Matter

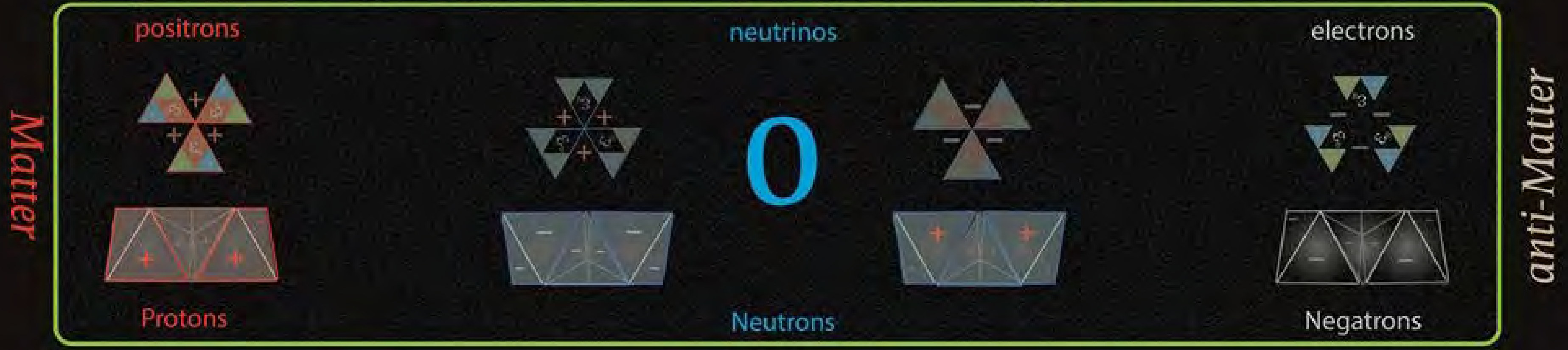
Matter
positrons
Protons
neutrinos
0
Neutrons
electrons
Negatrons
anti-Matter

Neutron stars are neutral charge pulsars

Positive charge pulsars

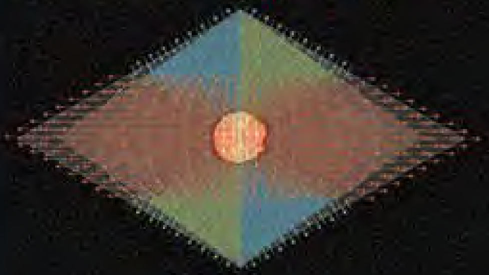

Strong E-fields

Neutron Stars

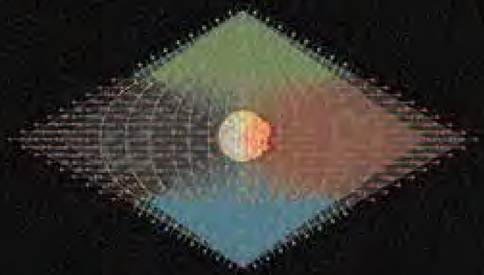

Strong dipolar M-fields

Negative charge pulsars

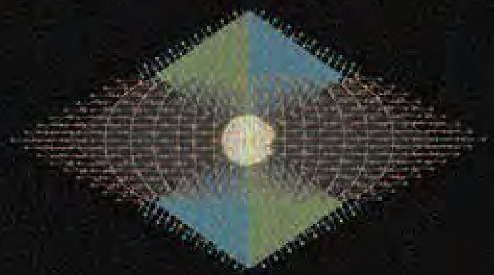

Strong E-fields

Tetryonics 78.06 - Neutron star types

Copyright ABRAHAM [2008] - All rights reserved

Pulsars

A pulsar is a rotating neutron star that emits a beam of electromagnetic radiation.

This radiation can only be observed when the beam of emission is pointing towards the Earth,
much the way a lighthouse can only be seen when the light is pointed in the direction of an observer,
and is responsible for the pulsed appearance of emission.

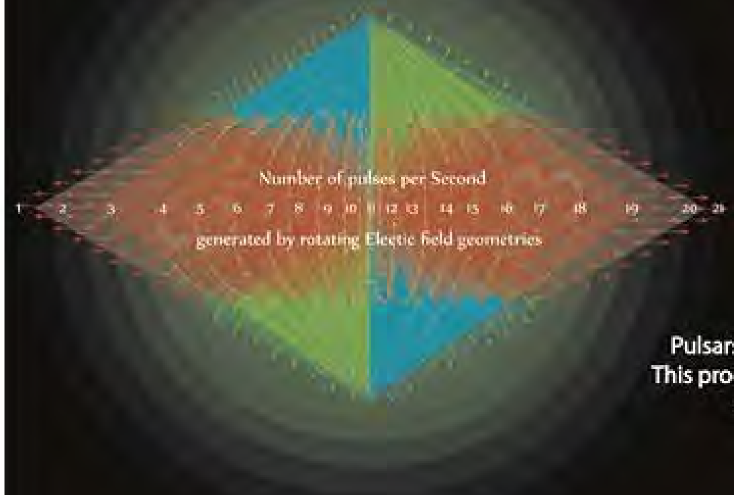
Number of pulses per Second
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
generated by rotating Electic field geometries

Pulsars are electrially charged versions of neutron stars that have weak M-fields in comparision to Neutron stars themselves

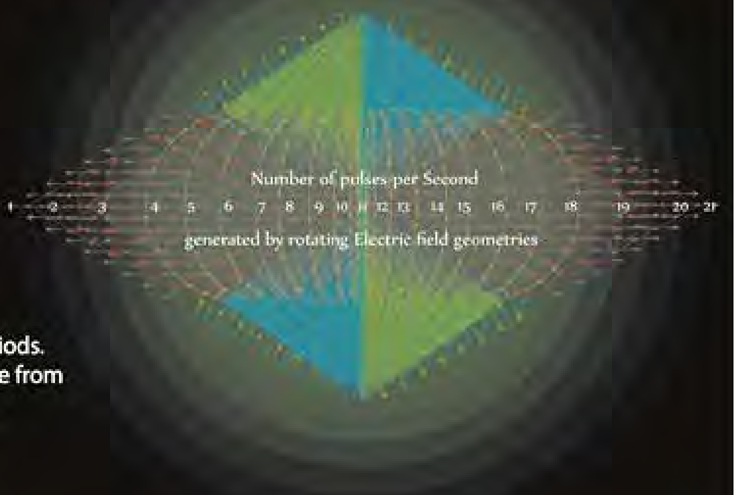
Number of pulses per Second
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
generated by rotating Electric field geometries

Pulsars are very dense, and have short, regular rotational periods.
This produces a very precise interval between pulses that range from
milliseconds to several seconds for individual pulsars

E-field waves accelerate charged particles M-fields do no work

Ampere

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$

Electric force *Magnetic field*

Lorentz

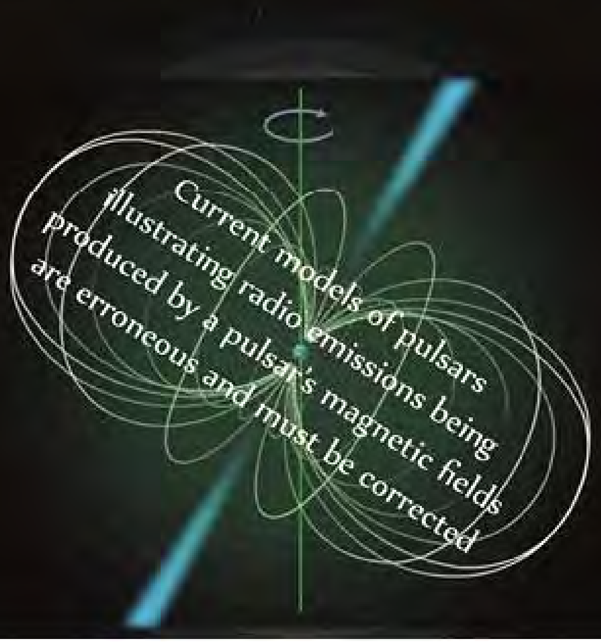
Current models of pulsars illustrating radio emissions being produced by a pulsar's magnetic fields are erroneous and must be corrected

Radio wave astomony detects E-field waves of energy momenta not magnetic waves

Maxwell

Because a magnetic force is always perpendicular to the motion,
the magnetic field can do no work on an isolated charge.

It can only do work indirectly, via the electric field
generated by a changing magnetic field

Faraday

Pulsars

Are compact Matter bodies with strong GEM fields

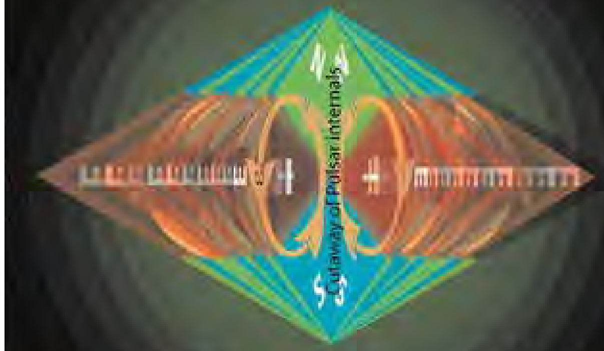

Like all Stellar bodies their GEM fields resolve to a point at their core

The gravitational collapse of Matter produces radiative electromagnetic waves

A millisecond pulsar (MSP) is a pulsar with a rotational period in the range of about 1-10 milliseconds.

Millisecond pulsars have been detected in the radio, X-ray, and gamma ray portions of the electromagnetic spectrum.

The origin of millisecond pulsars is still unknown.

The leading theory is that they begin life as longer period pulsars but are spun up or "recycled" through accretion. For this reason, millisecond pulsars are often called recycled pulsars.

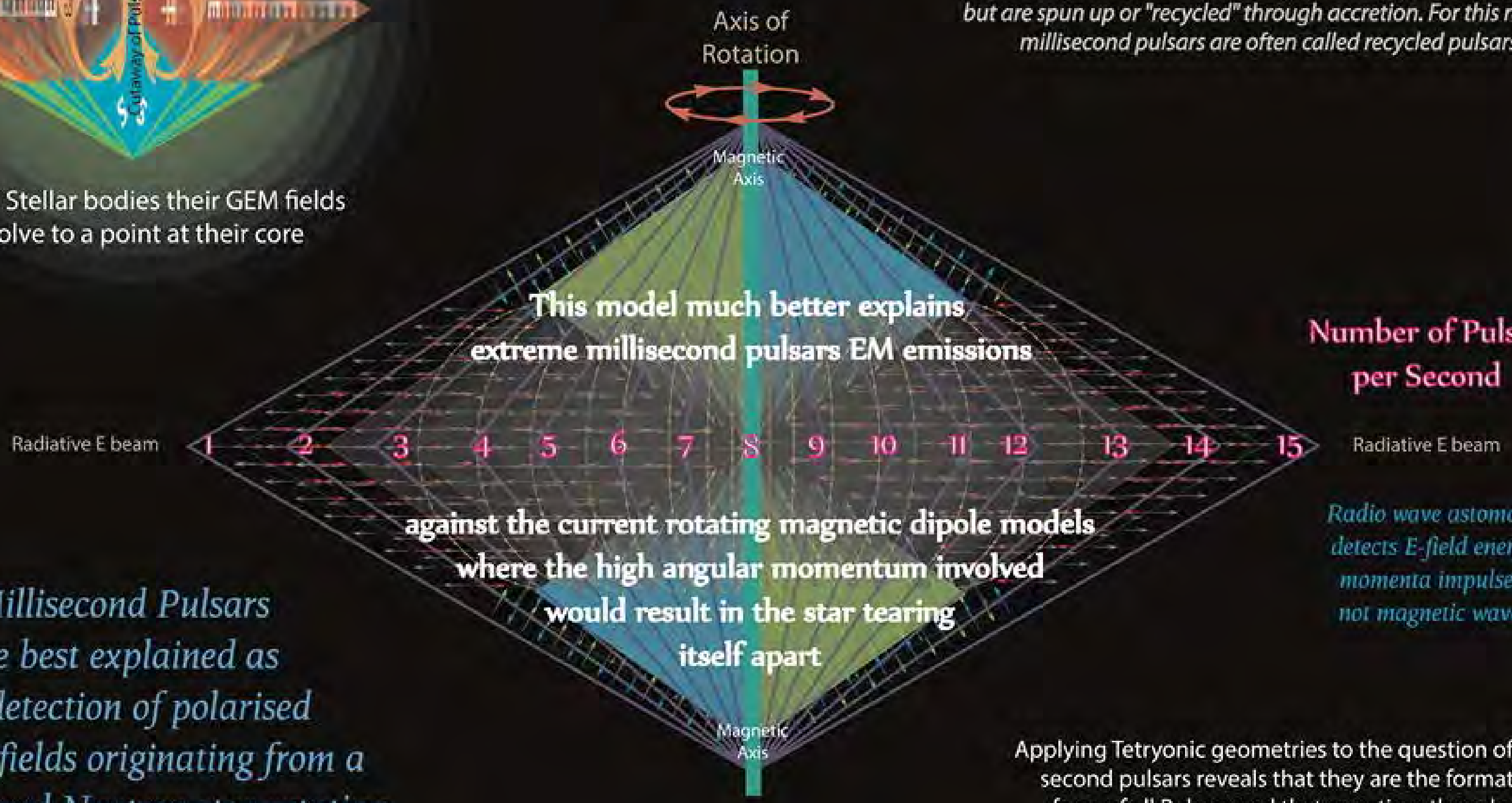

This model much better explains extreme millisecond pulsars EM emissions against the current rotating magnetic dipole models where the high angular momentum involved would result in the star tearing itself apart

Number of Pulses per Second

Radio wave astomony detects E-field energy momenta impulses not magnetic waves

Millisecond Pulsars are best explained as the detection of polarised Electric fields originating from a non-charged Neutron star rotating with high angular momentum

Applying Tetryonic geometries to the question of millisecond pulsars reveals that they are the formative form of all Pulsars and that over time they slow down to become the observed longer period pulsars [obeying the conservation of Angular momentum law]

Tetryonics 78.08 - Pulsar rotations

Copyright ABRAHAM [2008] - All rights reserved

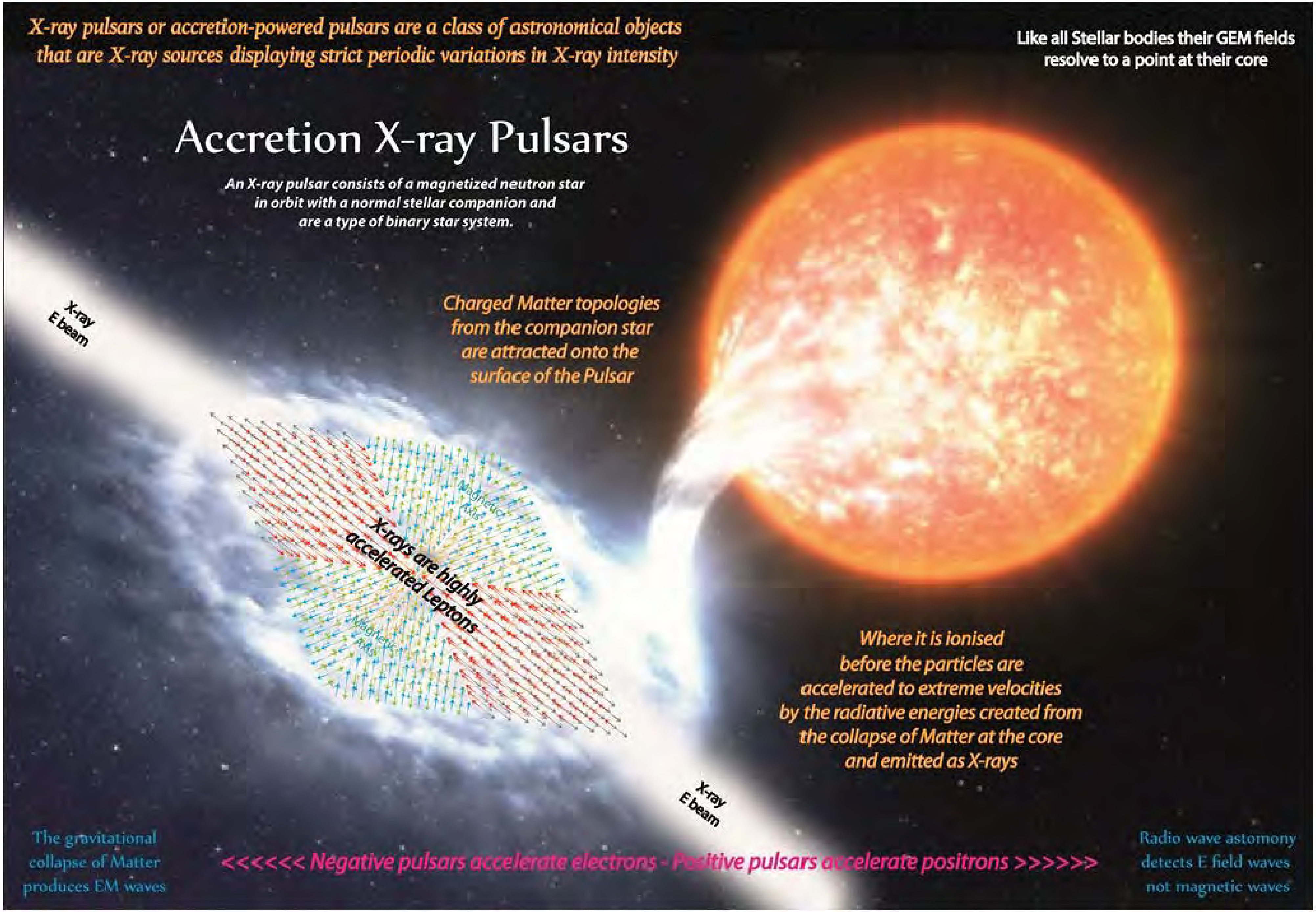

X-ray pulsars or accretion-powered pulsars are a class of astronomical objects that are X-ray sources displaying strict periodic variations in X-ray intensity

Like all Stellar bodies their GEM fields resolve to a point at their core

Accretion X-ray Pulsars

An X-ray pulsar consists of a magnetized neutron star in orbit with a normal stellar companion and are a type of binary star system.

Charged Matter topologies from the companion star are attracted onto the surface of the Pulsar

Where it is ionised before the particles are accelerated to extreme velocities by the radiative energies created from the collapse of Matter at the core and emitted as X-rays

The gravitational collapse of Matter produces EM waves

<<<<<< Negative pulsars accelerate electrons - Positive pulsars accelerate positrons >>>>>>

Radio wave astomony detects E field waves not magnetic waves

Tetryonics 78.09 - Accretion X-ray Pulsars

Copyright ABRAHAM [2008] - All rights reserved

Stellar Novae

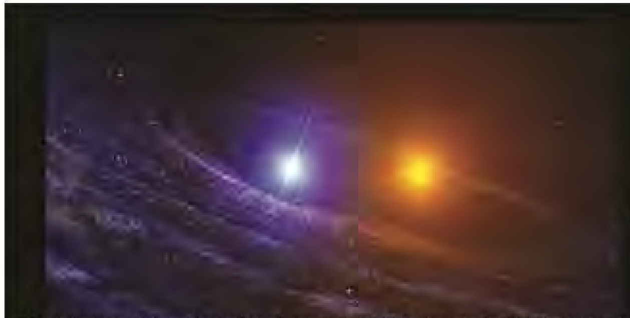

Novae are cataclysmic stellar phenomena that take place in binary systems consisting of a compact stellar object and a low-mass star.

Novae are cataclysmic nuclear explosions in a white dwarf star.

It is caused by the accretion of hydrogen on to the surface of the star, which ignites and starts a runaway nuclear 'fusion' process

All stellar 'fusion' processes are in fact GEM pinches that convert standing wave Matter into radiant mass-energy

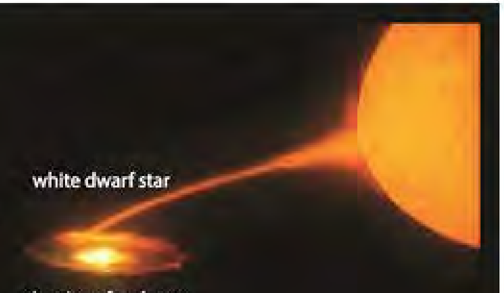

the size of a planet
but with a mass of up to
1.4 times that of the Sun

The stars must be close enough for the intense gravitational field of the white dwarf to tear material away from the outer layers of its companion.

photons of energy momenta
form kEM fields that accelerate
ions and electrons

spectral
line transitions

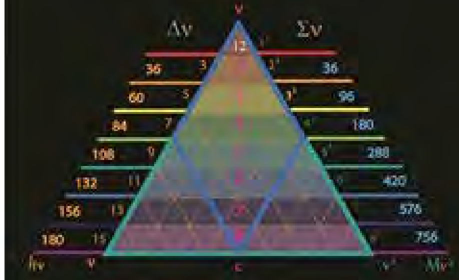

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

all elemental spectral lines
are the result of accelerating
photo-electrons

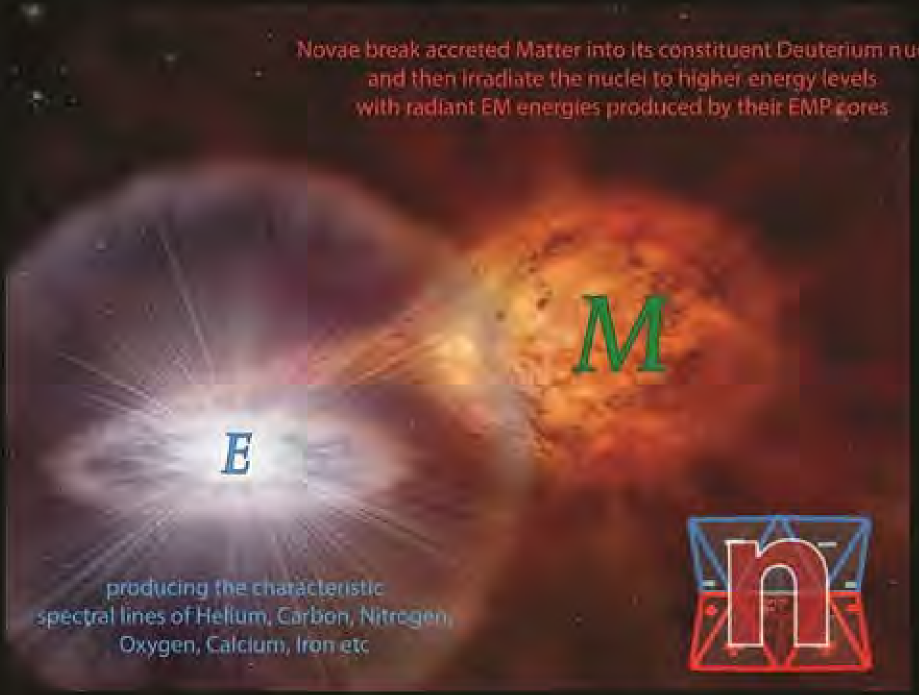

Novae break accreted Matter into its constituent Deuterium nuclei
and then irradiate the nuclei to higher energy levels
with radiant EM energies produced by their EMP cores

producing the characteristic
spectral lines of Helium, Carbon, Nitrogen,
Oxygen, Calcium, Iron etc

runaway
Matter-energy conversion
within the supernova

12	*Lyman*
48	*Balmer*
108	*Paschen*
192	*Brackett*
300	*Pfund*
432	*Humphries*
588	*un-named*
768	*ionised leptons*

accelerates ions & photo-electrons
producing spectral line emissions
and creates gamma ray bursts

An enigma that has puzzled experts for over 50 years is the origin of the irregular, inhomogeneous distribution of nova ejecta

The material transferred by the companion star is often of solar composition (i.e., close to 98% hydrogen and helium by mass).

Elements, in the range between carbon (C) and calcium (Ca) on the periodic table, can account for 30% to 50% of the material ejected during a nova explosion.

Supernovae explosions

Supernovae are extremely luminous and cause a burst of radiation that often briefly outshines an entire galaxy,

The GEM pinching [collapse of Matter waveforms]
always results in divergent neutral charge
KEM mass-energies

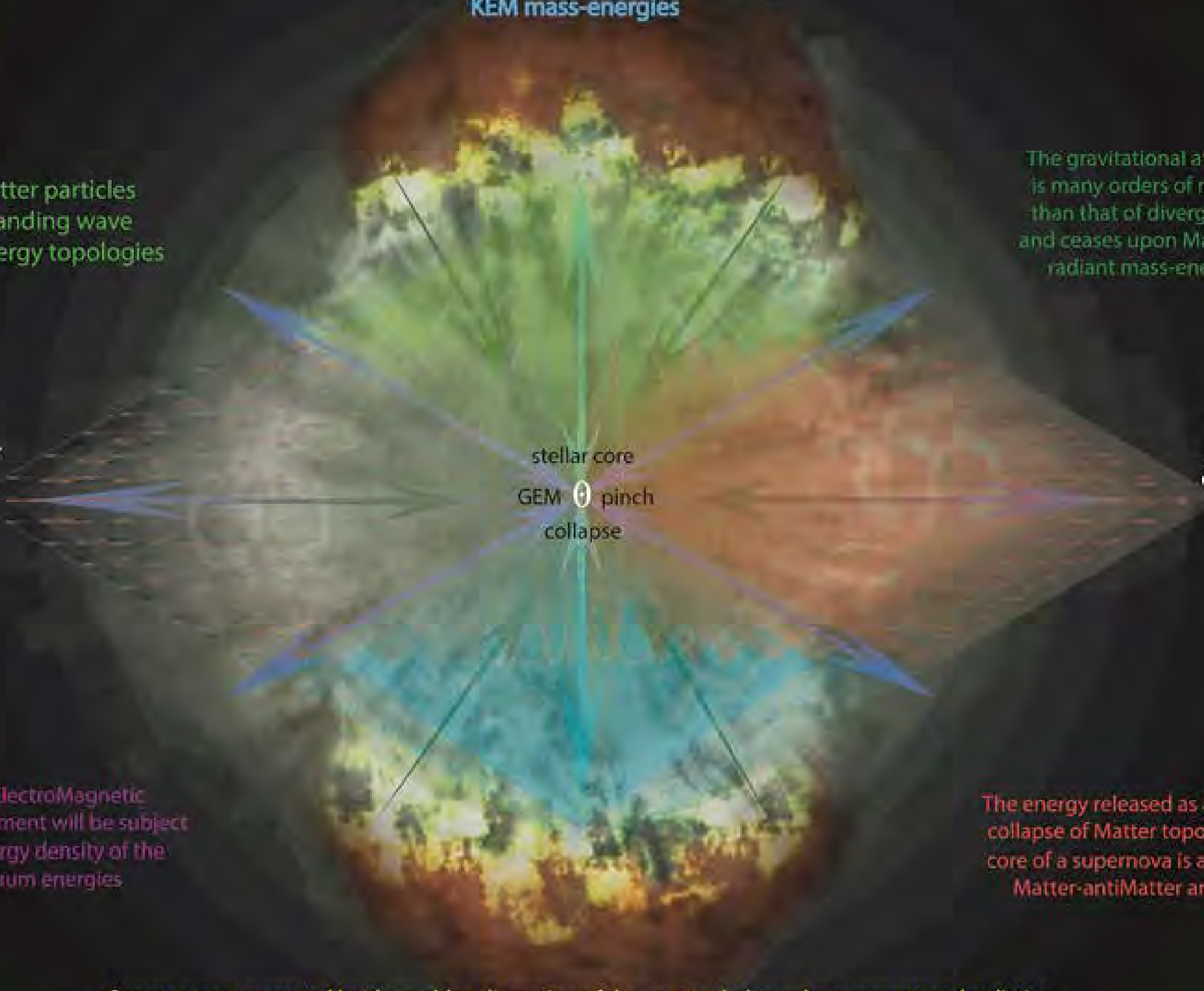

All Matter particles
are standing wave
mass-energy topologies

The gravitational attraction of Matter
is many orders of magnitude weaker
than that of divergent KEM energies
and ceases upon Matter's collapse into
radiant mass-energy geometries

If the EM pinch at the core of a Star
rapidly increases its efficiency or
loses its ability to confine the
resultant KEM energies the
star will be blasted apart

Black holes, hypothesied in physics
as a solution to the math of Einstein's
General relativity since it was was first
formulated in 1915, do NOT exist

All their physics are provided for
by GEM pinch field mechanics

A star moving in a ElectroMagnetic
vacuum energy environment will be subject
to changes in the energy density of the
surrounding vacuum energies

The energy released as a result of the
collapse of Matter topologies in the
core of a supernova is as efficient as
Matter-antiMatter annihilation

Supernovae are created by the sudden disruption of the counter balanced convergent and radiative
mass-ENERGY-Matter output of GEM pinch singularities at the core of compact stellar objects
which results in the sudden release of massive amounts of divergent mass-energy
all at once from the stellar core blasting away the remaining stellar Matter

Copyright ABRAHAM [2008] - All rights reserved

Singularities

A gravitational singularity or spacetime singularity is a location where the quantities that are used to measure the gravitational electromagnetic fields become infinite in a way that does not depend on the spatial co-ordinate system

Just as the radiative KEM fields of Matter in motion and charged EM fields gives rise to the modern physics misconceptions of spherical point particles and String theories

Tetryonic theory precludes point particles and black holes through analytic mass-energy geometries & Gaussian Matter topologies that result from equilateral Planck energy momenta interactions

The collapse of 3D Matter releases its standing-wave energies as radiant 2D mass-energies foms the basis for all Stellar energy releases

Current theories based on mathematical modelling of General relativity without accurate quantum mechanical models predicts that at densities greater than those supported by fermion degeneracy, gravity overwhelms all other forces collapsing the 3D material body of a large star to create a singularity and the eventual formation of a black hole.

Equilateral Planck charge geometries are the key to understanding stellar processes

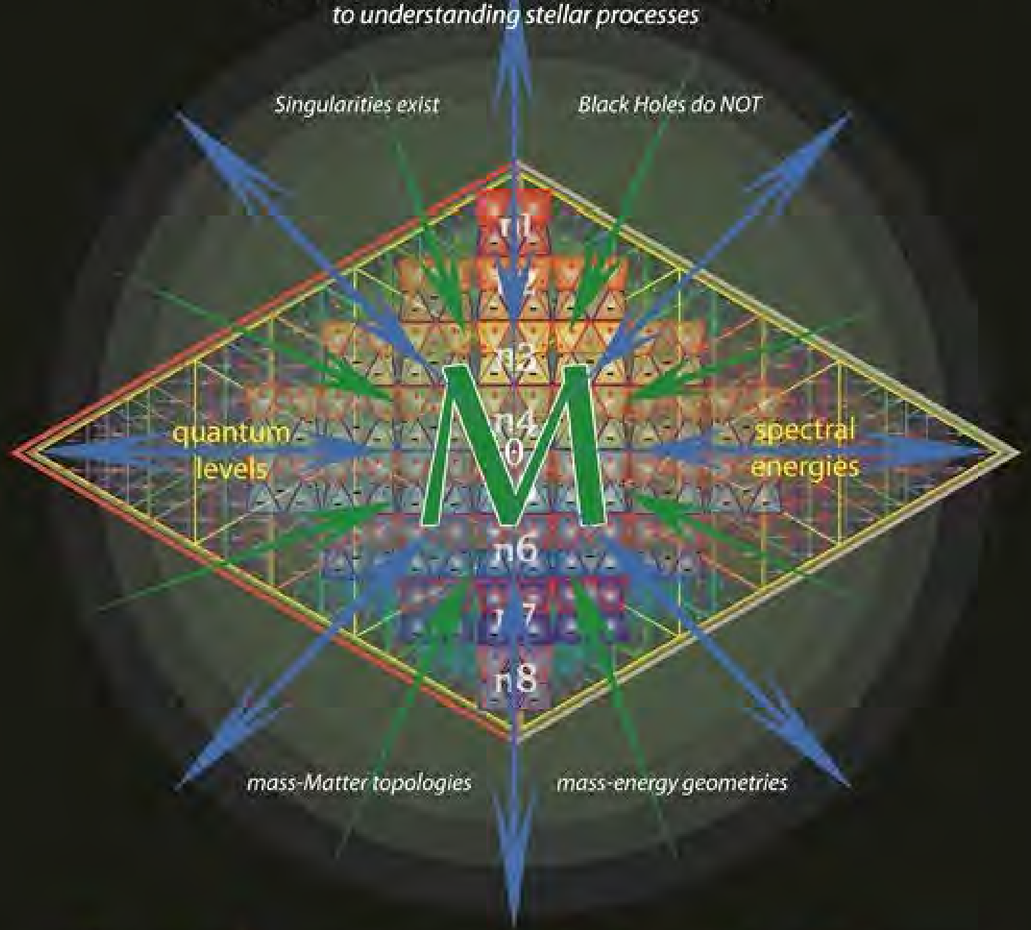

Singularities exist

Black Holes do NOT

quantum levels

spectral energies

mass-Matter topologies

mass-energy geometries

The smallest quanta of 3D Matter is a Tetryon

(any attempt to reduce its 3 spatial dimensions results in its collapse into 2D divergent mass-energy waveforms)

Tracing the GravitoElectroMagnetic fields of Matter back to their origin point can give rise to the misconception of a EM field singularity leading to a black hole

Copyright ABRAHAM [2008] - All rights reserved

Quasi-stellar objects

As standing-wave Matter is converted to radiant EM masses and pure energy the outputs of GEM pinches can increase dramatically over short periods

Black holes do not exist

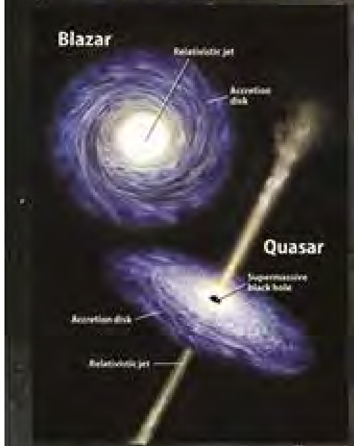
Blazar
Relativistic jet
Accretion disk
Quasar
Supermassive black hole
Accretion disk
Relativistic jet

A quasi-stellar radio source ("quasar") is a very energetic and distant active galactic nucleus.

Quasars and Blazars are extremely luminous and were first identified as being high redshift sources of electromagnetic energy, including radio waves and visible light, that were point-like, similar to stars, rather than extended sources similar to galaxies

SDSS J0919+2720
30,000 light-years
7.7kpc
5"6
SDSS J1005+4016
F814W I
F475W B
N
E
50,000 light-years
15.3kpc
8"9
SDSS J0827+5224
50,000 light-years
15.3kpc
5"8

Long hypothesised as having supermassive Blackholes at their cores Tetryonics reveals Matter~Energy GEM pinches to be the source of the observed, distant large-scale stellar dynamics

All Matter (comprised of charged Planck quanta) will accelerate according to their charge distributions as they interact with the divergent energies of EM pinches

The central 'core' point singularity of any GEM pinch is completely invisible and is revealed only by its effects on Matter through its interactive EM fields

$$E = mc^2$$

mass-energy equivalence

mass geometries are energy per c^2

Stellar GEM pinches convert Matter topologies into radiant mass-energy geometries with 100% efficiency

Matter topologies are energy per c^4

Matter-energy equivalence

$$E = \frac{m}{c^2}$$

Copyright ABRAHAM [2008] - All rights reserved

Active Galaxies

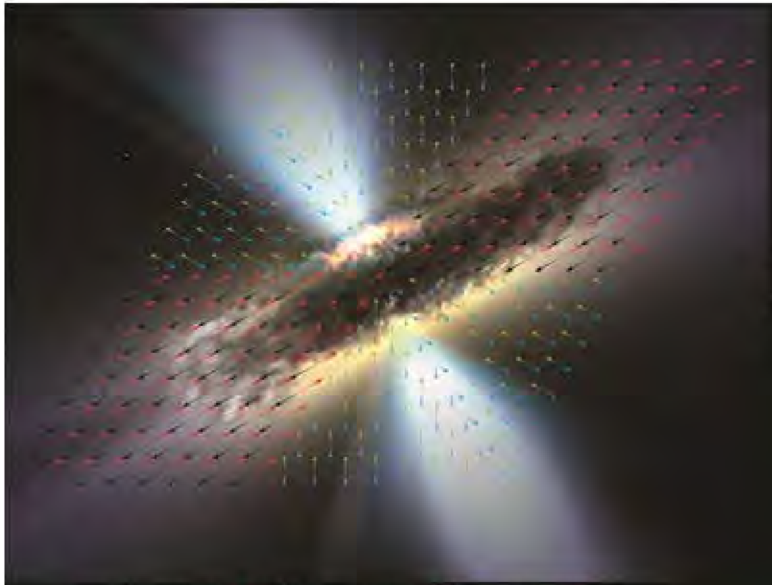

In addition to the E-field acceleration of charged particles GEM pinches can also release energies along their Magnetic axis

An active galactic nucleus (AGN) is a compact region at the centre of a galaxy that has a much higher than normal luminosity over at least some portion, and possibly all, of the electromagnetic spectrum.

The motion of stars about a central point in a galaxy's core is the result of a GEM pinch's radiated interactive forces

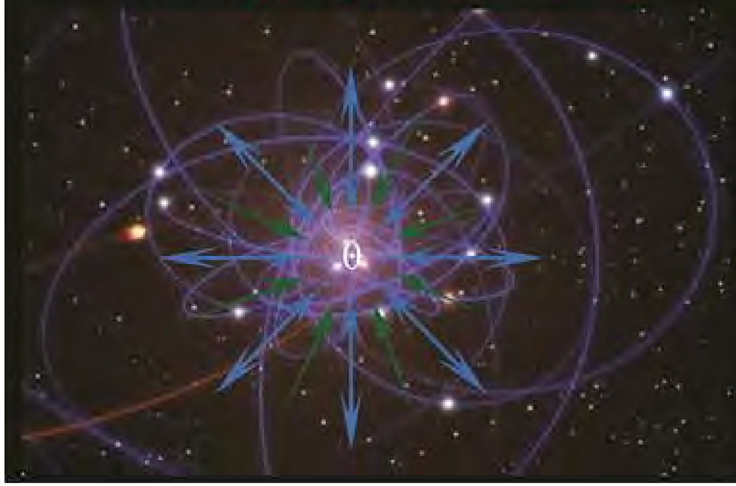

GR uses stress tensors to model the gravitational effects of ALL energy in a system in its various forms without distinction or differentiation between mass-energy geometries and Matter topologies

EM PINCHES ARE 100% EFFICIENT CONVERTERS OF MATTER INTO ENERGY

Using Tetryonic geometry the Universe is revealed as being completely dominated by ElectroMagnetic forces, with Gravity aiding at the larger scales to help shape the differing forms of Galaxies that we see

EM PINCHES REDUCE TO SINGULARITIES AT THE FOCUS OF THEIR DIVERGENT FIELDS

Tetryonic geometry reveals divergent interactive EM masses & convegent gravitational Matter all contribute to the nett force of Gravitational fields

The current astrophysical models of active galaxies propose supermassive Blackholes as the only mechanism possible to produce the observed large-scale dynamics

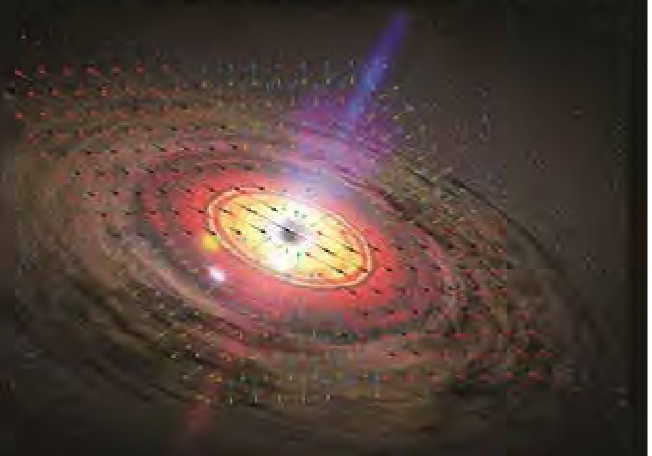

E-fields hold Matter in a disc and M-fields release highly polarised energies

Tetryonics reveals the Galactic EM Pinches are also capable of fitting the observed data and explaining many of the current mysteries surrounding how active galaxies generate their power.

The energies released by the conversion of Matter into energy in GEM pinches is the most energetic process in the Universe

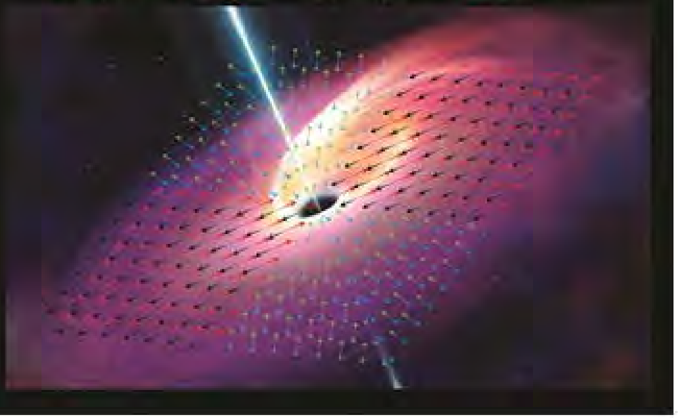

Copyright ABRAHAM [2008] - All rights reserved

Black holes

[Garbage IN - Garbage OUT]

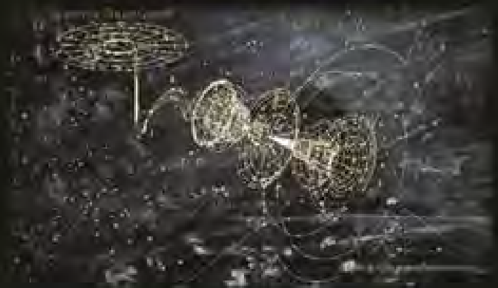

Computer-based modelling of GR
mathematics without charge geometries
leads to paradoxial & erroneous outcomes

Tetryonics allows for Supermassive Matter to form but does NOT allow for the classically defined Black holes

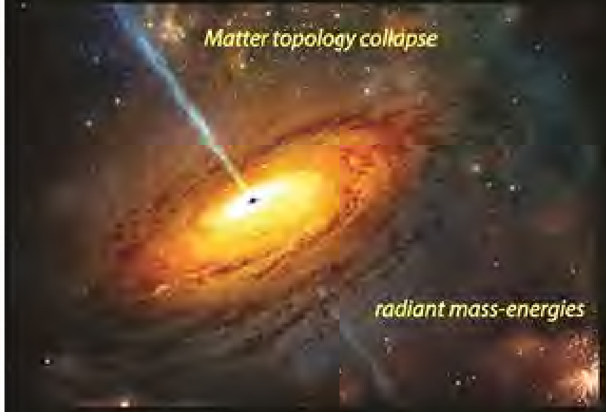

The centres of Galaxies & Quasars hypothesised as being powered by supermassive blackholes are in fact the foci of ElectroMagnetic pinches

In General relativity, a black hole is a region of space in which the gravitational field is so powerful that nothing, including light, can escape its gravitational pull.

The black hole has a one-way surface, called an event horizon, into which objects can fall, but out of which nothing can come. It is called "black" because it absorbs all the light that hits it, reflecting nothing, just like a perfect blackbody in thermodynamics.

'Standard Model' quantum analysis of black holes shows them to possess a temperature and emit Hawking radiation.

In Standard Model Quantum Mechanics a problem arises in that Charge and Gravity equations break down and produce infinities and other impossible results

Very large non-radiating 'Dark Matter' would be the closest thing possible to the currently theorised black-holes.

However the 3D volume of such an object would be limited to the space taken up by its constituent Tetryonic charge Matter topologies. (ie Tetryons face-to face without Space present between any facsias)

This is similar to the current models of Neutron Stars , WIMPs and MACHOs

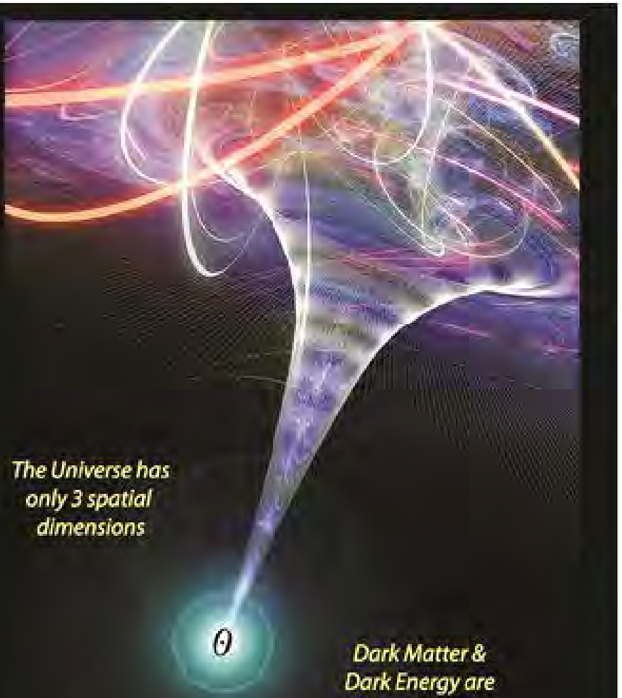

The Universe has only 3 spatial dimensions

Dark Matter & Dark Energy are EM vector forces

Black holes, Worm holes & other passable singularities do NOT exist

The collapse of standing-wave Matter topologies results in the release of divergent radiative mass-energy geometries

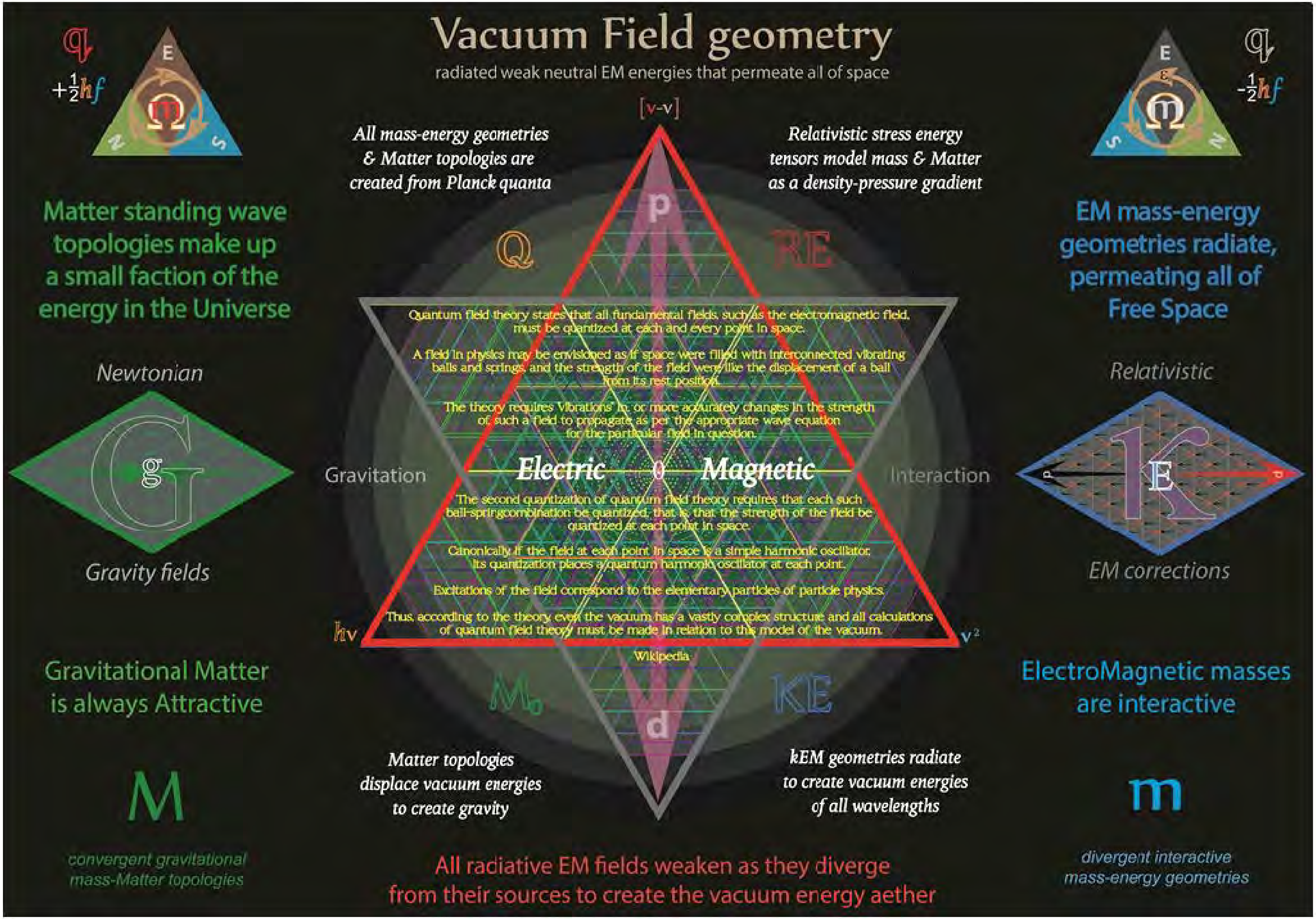

Vacuum Field geometry

radiated weak neutral EM energies that permeate all of space

All mass-energy geometries & Matter topologies are created from Planck quanta

Relativistic stress energy tensors model mass & Matter as a density-pressure gradient

Matter standing wave topologies make up a small faction of the energy in the Universe

EM mass-energy geometries radiate, permeating all of Free Space

Quantum field theory states that all fundamental fields, such as the electromagnetic field, must be quantized at each and every point in space.

A field in physics may be envisioned as if space were filled with interconnected vibrating balls and springs, and the strength of the field were like the displacement of a ball from its rest position.

The theory requires 'vibrations' in, or more accurately changes in the strength of such a field to propagate as per the appropriate wave equation for the particular field in question.

Newtonian

Gravitation

Electric ***Magnetic***

Interaction

Relativistic

The second quantization of quantum field theory requires that each such ball-springcombination be quantized, that is that the strength of the field be quantized at each point in space.

Canonically, if the field at each point in space is a simple harmonic oscillator, its quantization places a quantum harmonic oscillator at each point.

Excitations of the field correspond to the elementary particles of particle physics.

Thus, according to the theory, even the vacuum has a vastly complex structure and all calculations of quantum field theory must be made in relation to this model of the vacuum.

Wikipedia

Gravity fields

EM corrections

Gravitational Matter is always Attractive

ElectroMagnetic masses are interactive

Matter topologies displace vacuum energies to create gravity

kEM geometries radiate to create vacuum energies of all wavelengths

M

convergent gravitational mass-Matter topologies

m

divergent interactive mass-energy geometries

All radiative EM fields weaken as they diverge from their sources to create the vacuum energy aether

Tetryonics 79.07 - Vacuum Field geometry

The Big Bang theory

The Big Bang is a considered to be a well-tested scientific theory which is widely accepted within the scientific community because it is the most accurate and comprehensive explanation for the full range of phenomena astronomers observe.

The framework for the Big Bang model relies on Albert Einstein's general relativity and on simplifying assumptions such as homogeneity and isotropy of space.

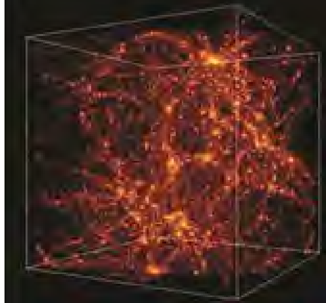

Primodial Atom

According to the Big Bang theory, the Universe was once in an extremely hot and dense state which expanded rapidly

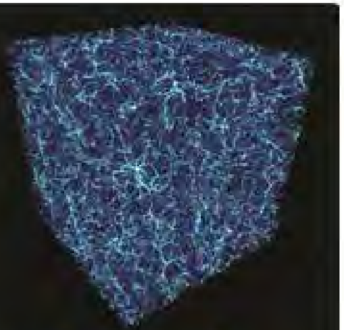

The Big Crunch

If the mass density of the Universe were greater than the critical density, then the Universe would reach a maximum size and then begin to collapse.

It would become denser and hotter again, ending with a state similar to that in which it started

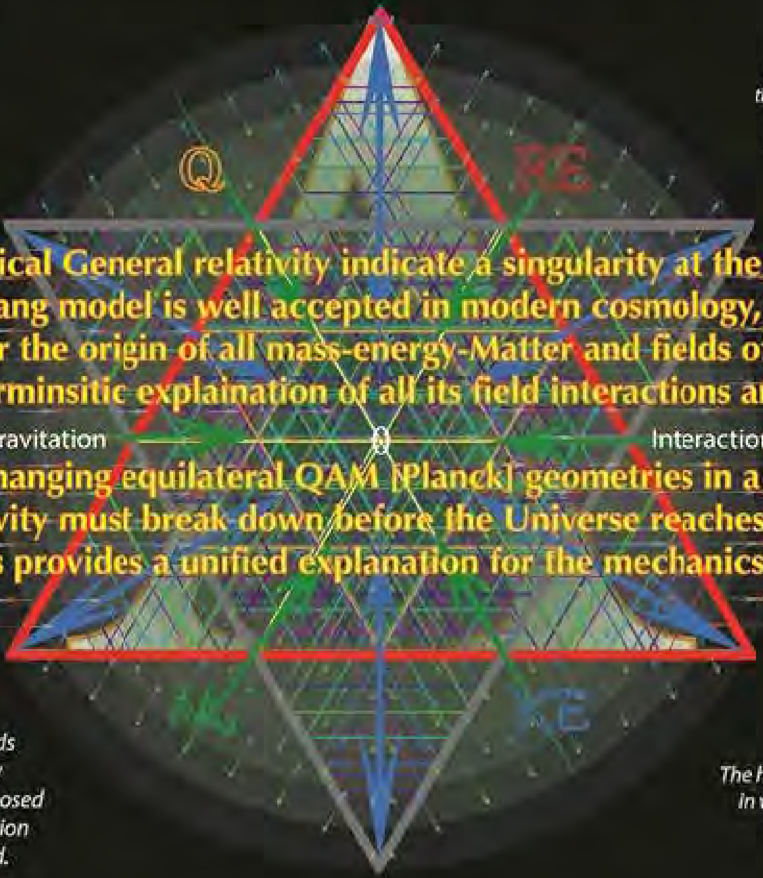

The equations of classical General relativity indicate a singularity at the origin of cosmic time, and while the Big Bang model is well accepted in modern cosmology, Tetryonics provides a clear causal model for the origin of all mass-energy-Matter and fields of force in our Universe, offering a unified, determinsitic explaination of all its field interactions and ultimate energy fate.

Tetryonics relates Time to changing equilateral QAM [Planck] geometries in a dynamic evolving Universe. Moreover, general relativity must break down before the Universe reaches the Planck temperature while Tetryonics provides a unified explanation for the mechanics of the Universe

Big Bang

Extrapolation of the expansion of the Universe backwards in time using general relativity yields an infinite density and temperature at a finite time in the past, which it is proposed underwent a explosive birth and a period of rapid expansion in order to produce the homognetity currently observed.

The Big Chill

The heat death of the universe is a suggested ultimate fate of the universe, in which the universe has diminished to a state of no thermodynamic free energy and therefore can no longer sustain motion or life.

Inflation

Energy death

The Big Bang theory is the prevailing cosmological model that attempts to explain the early development of the Universe

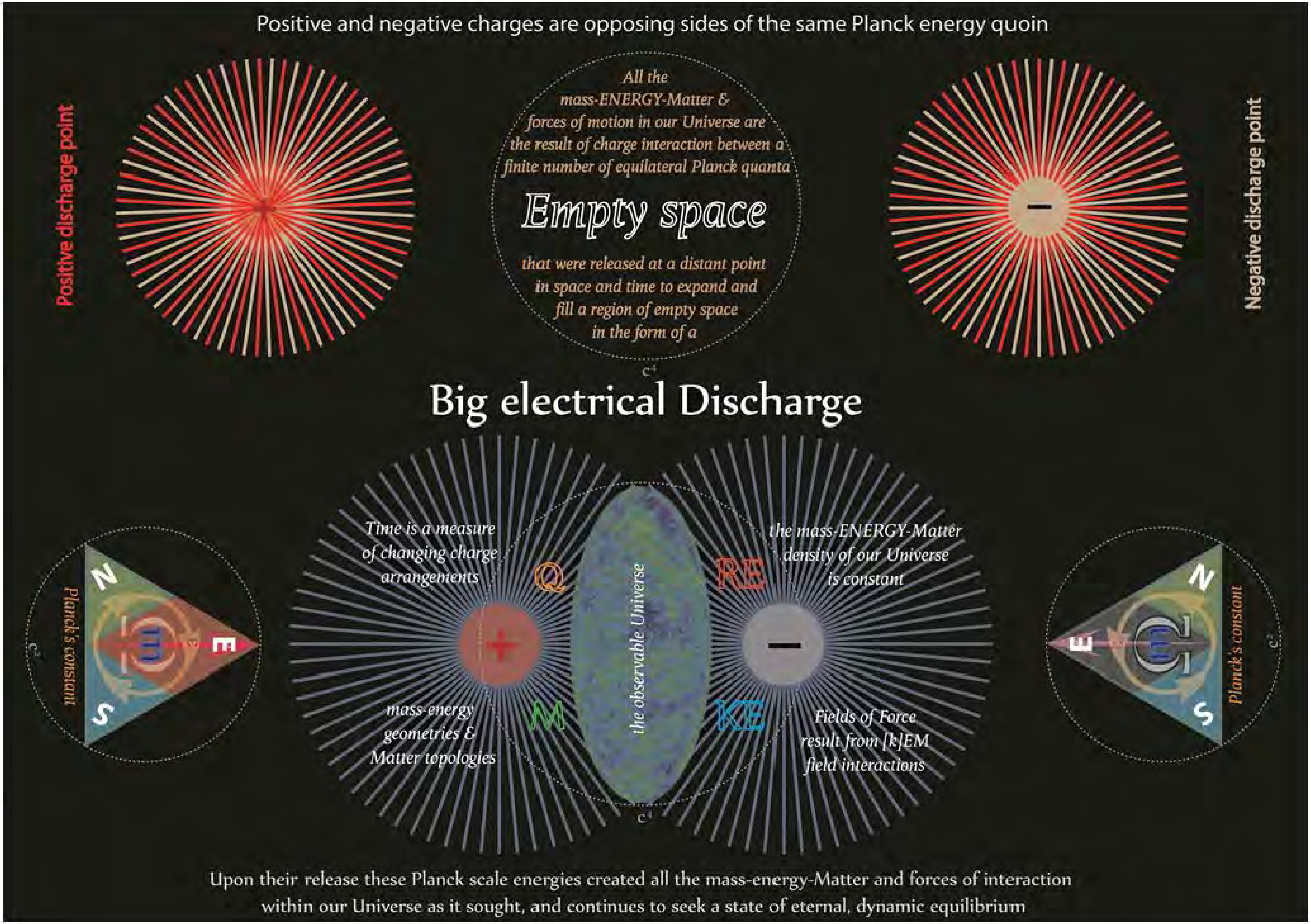

Positive and negative charges are opposing sides of the same Planck energy quoin

All the mass-ENERGY-Matter & forces of motion in our Universe are the result of charge interaction between a finite number of equilateral Planck quanta

Empty space

that were released at a distant point in space and time to expand and fill a region of empty space in the form of a

c^4

Big electrical Discharge

Time is a measure of changing charge arrangements

mass-energy geometries & Matter topologies

the mass-ENERGY-Matter density of our Universe is constant

Fields of Force result from [k]EM field interactions

Upon their release these Planck scale energies created all the mass-energy-Matter and forces of interaction within our Universe as it sought, and continues to seek a state of eternal, dynamic equilibrium

Positive discharge point
Negative discharge point
Q
M
RE
KE
the observable Universe
c^4
N
E
S
Planck's constant
c^2
N
E
S
Planck's constant
c^2

The Dynamic Universe

Like a living breathing Organism, the Universe can now be seen as an eternally dynamic system that is constantly evolving through Matter creation, EM interactions & Matter destruction

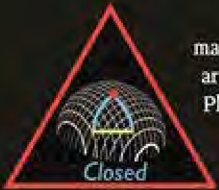

All mass-ENERGY -Matter arises from equilateral Planck energy quanta

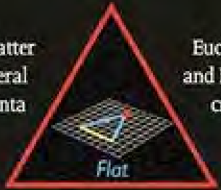

Flat Euclidean geometries and Platonic topologies create everything

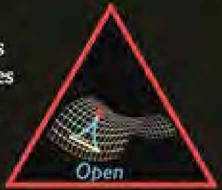

Gravitational contraction

Singularities

arise from the vector modelling of convergent EM fields of force without an understanding of the EM energy momenta geometries

The formation of Matter creates regions of attractive low vacuum energy pressure

[Convergent Gravity]

Dark Matter

results from modern modelling of convergent EM interactions without an understanding of the EM energy momenta geometries

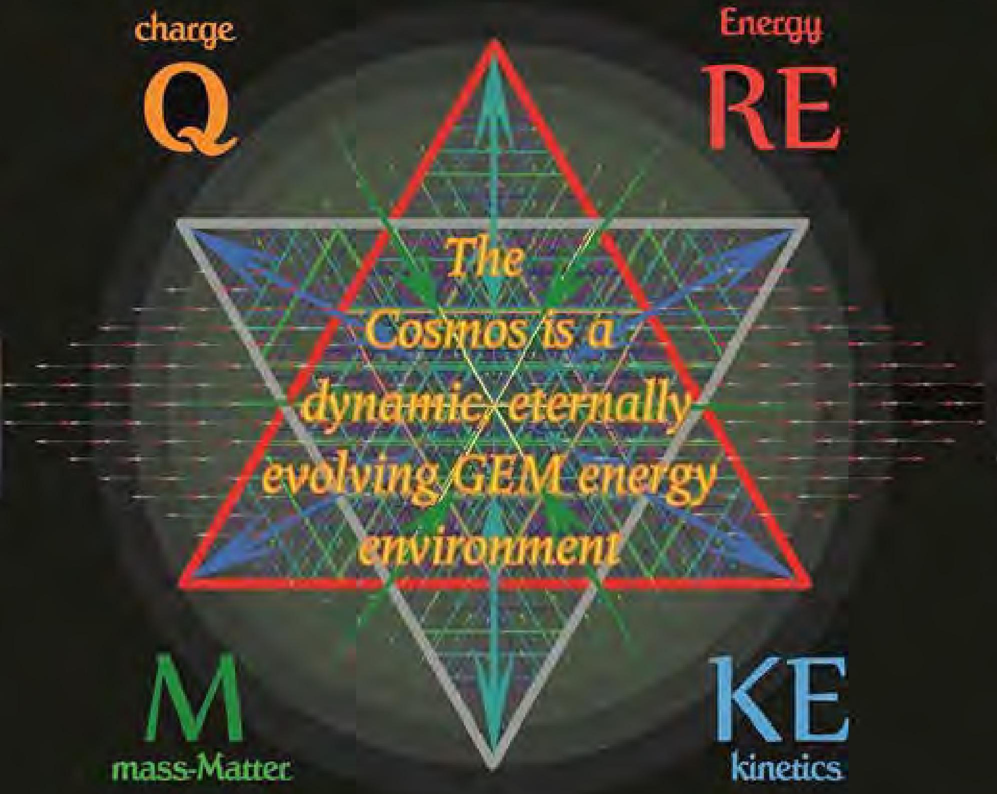

charge

Q

Energy

RE

The Cosmos is a dynamic eternally evolving GEM energy environment

M

mass-Matter

KE

kinetics

Radiant expansion

Inflation

arises from an attempt to explain the observed homogenity of the Universe without an understanding of kEM energy momenta geometries

The collapse of standing-wave Matter releases interactive EM mass-energies

[Divergent EM Radiation]

Dark Energy

results from modern modelling of divergent EM interactions without an understanding of the EM energy momenta geometries

The energetic 'space-time fabric' of our Universe is not curved, it is created from flat Euclidean Planck quanta in a finely tuned state of dynamic equilibrium between gravitational Matter & radiant EM mass-energies

Tetryonics 79.10 - The Dynamic Universe

Copyright ABRAHAM [2008] - All rights reserved

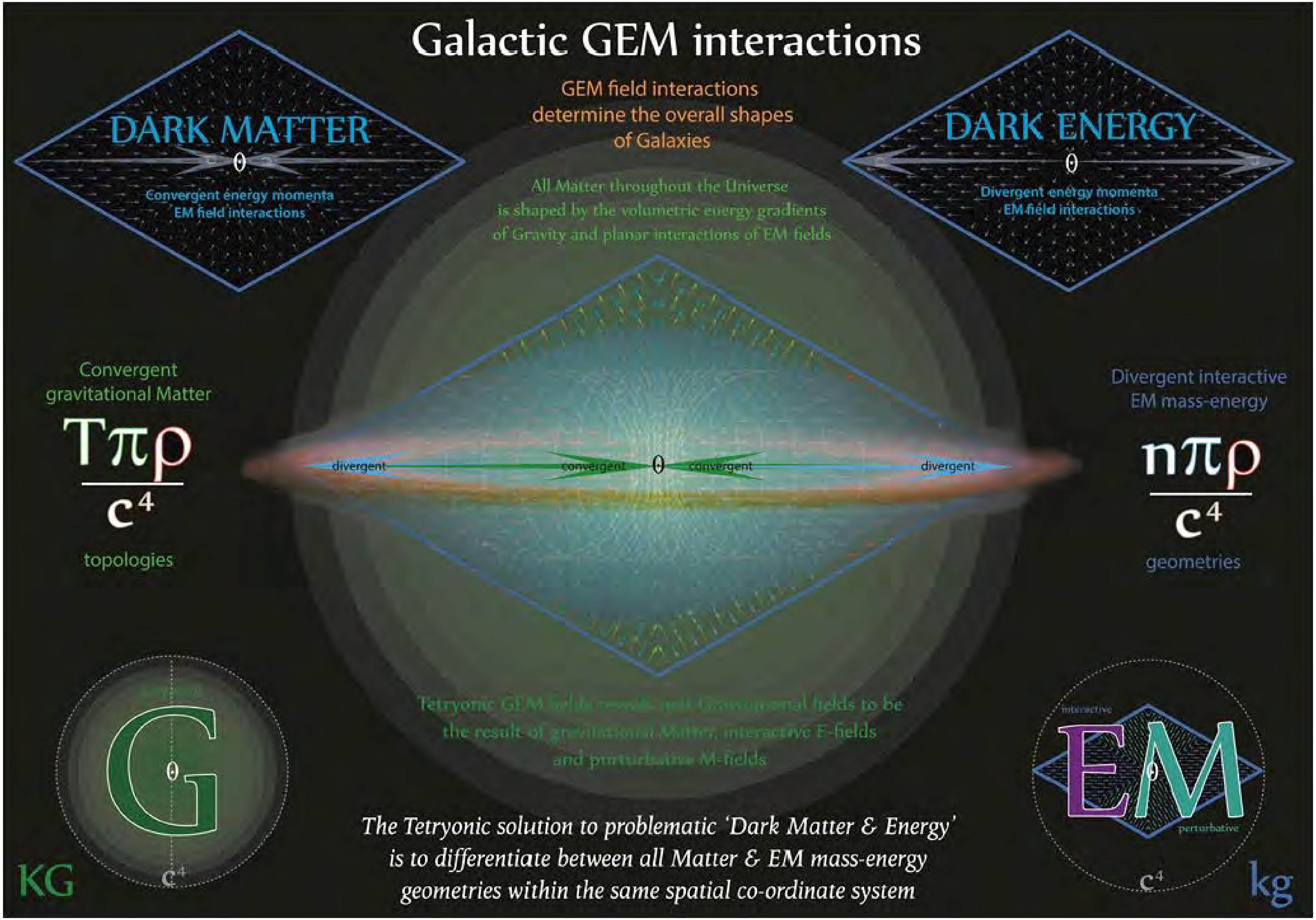

Galactic GEM interactions

GEM field interactions
determine the overall shapes
of Galaxies

DARK MATTER

Convergent energy momenta
EM field interactions

DARK ENERGY

Divergent energy momenta
EM field interactions

All Matter throughout the Universe
is shaped by the volumetric energy gradients
of Gravity and planar interactions of EM fields

Convergent
gravitational Matter

$$\frac{T\pi\rho}{c^4}$$

topologies

Divergent interactive
EM mass-energy

$$\frac{n\pi\rho}{c^4}$$

geometries

Tetryonic GEM fields reveals that Gravitational fields to be
the result of gravitational Matter, interactive E-fields
and perturbative M-fields

*The Tetryonic solution to problematic 'Dark Matter & Energy'
is to differentiate between all Matter & EM mass-energy
geometries within the same spatial co-ordinate system*

Tetryonics 79.11 - Universal GEM interactions

Copyright ABRAHAM [2008] - All rights reserved

The Dark Forces

Gravitational attraction is the result of the nullspace of Matter.
It is many orders of magnitude weaker than ElectroMagnetic forces
(Gravity does not hold the whole galaxy together)

charged energy geometries

energy momenta quanta

$$n\pi\left[m\Omega v\right]^2$$

mass velocity

$$G_{ab} + \Lambda g_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Newtonian Gravity
models the motion resulting
from the attractive force
between Matter

General Relativity
models perturbative force
resulting from the total energy
of any system

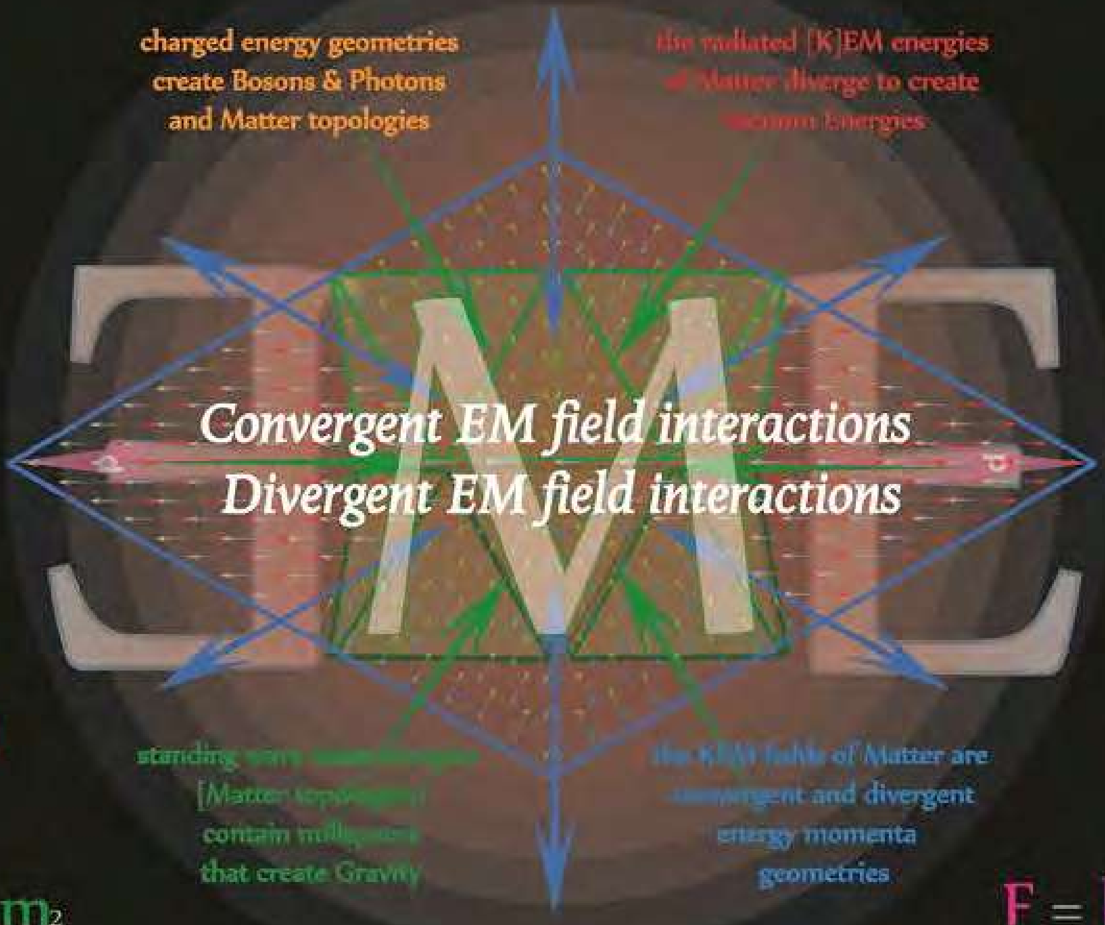

charged energy geometries
create Bosons & Photons
and Matter topologies

the radiated [K]EM energies
of Matter diverge to create
vacuum Energies

Convergent EM field interactions
Divergent EM field interactions

standing wave [illegible]
[Matter topologies]
contain nullspaces
that create Gravity

the [K]EM fields of Matter are
convergent and divergent
energy momenta
geometries

All [K]EM radiation has
a CONVERGENT component in its
interactive field that increases
when Matter is in motion

DARK MATTER

or when its standing-wave Matter
topology is destroyed

All [K]EM radiation has
a DIVERGENT omponent in its
interactive field that increases
when Matter is in motion

DARK ENERGY

or when its standing-wave Matter
topology is destroyed

$$E \xrightarrow[\text{energy momenta}]{\text{convergent}} M$$

$$M \xrightarrow[\text{energy momenta}]{\text{divergent}} E$$

$$F = -G\frac{M_1 m_2}{r^2}$$

$$F = k\frac{Q_1 Q_2}{r^2}$$

Gravitational
Matter
is attractive

G

Tetryonics dictates that in order to correct for measured Gravitational accelerations
we must distingush between EM mass and Matter and take into account
all the EM forces as well as that of Gravity produced by Matter

k

ElectroMagnetic
mass-energies
are interactive

Tetryonics 80.01 - The Dark Forces

Copyright ABRAHAM [2008] - All rights reserved

Vacuum Energies

Vacuum energy is an underlying background energy that exists in space even when it is devoid of Matter (free space) and results from the introduction of the Cosmological constant [Λ] by Albert Einstein as a modification of his original theory of General relativity in order to achieve a static universe.

Dark Matter is a measure of CONVERGENT KEM interactions additional to that of Gravity

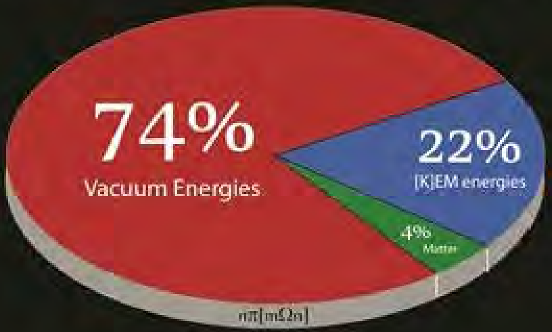

Dark Energy is a measure of DIVERGENT KEM interactions opposing that of Gravity

He later abandoned this concept after observations of the Hubble redshift indicated that the universe might not be stationary

Gravitational Matter is always attractive

$$G_{ab} + \Lambda g_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

total system energy

ElectroMagnetic masses are always interactive

Matter

M

topologies

Unlike Einstein's Special and General theories of Relativity that utilizes a scalar energy density tensor, Tetryonics utilizes analytical geometry & topology to differentiate between EM mass & Matter and model all energy interactions in any spatial region

mass

m

geometries

Gravity results from standing-wave Matter

Gravito-Electro-magnetic interactions

Total energy geometries *quantised energy momenta*

$$GEM = \frac{n\pi[m\Omega v]^2}{c^4}$$

of all EM mass-ENERGY-Matter in the Universe can be defined as

spatial volume of mass-energy-Matter interactions

EM interactions result from superpositioned EM masses

Copyright ABRAHAM [2008] - All rights reserved

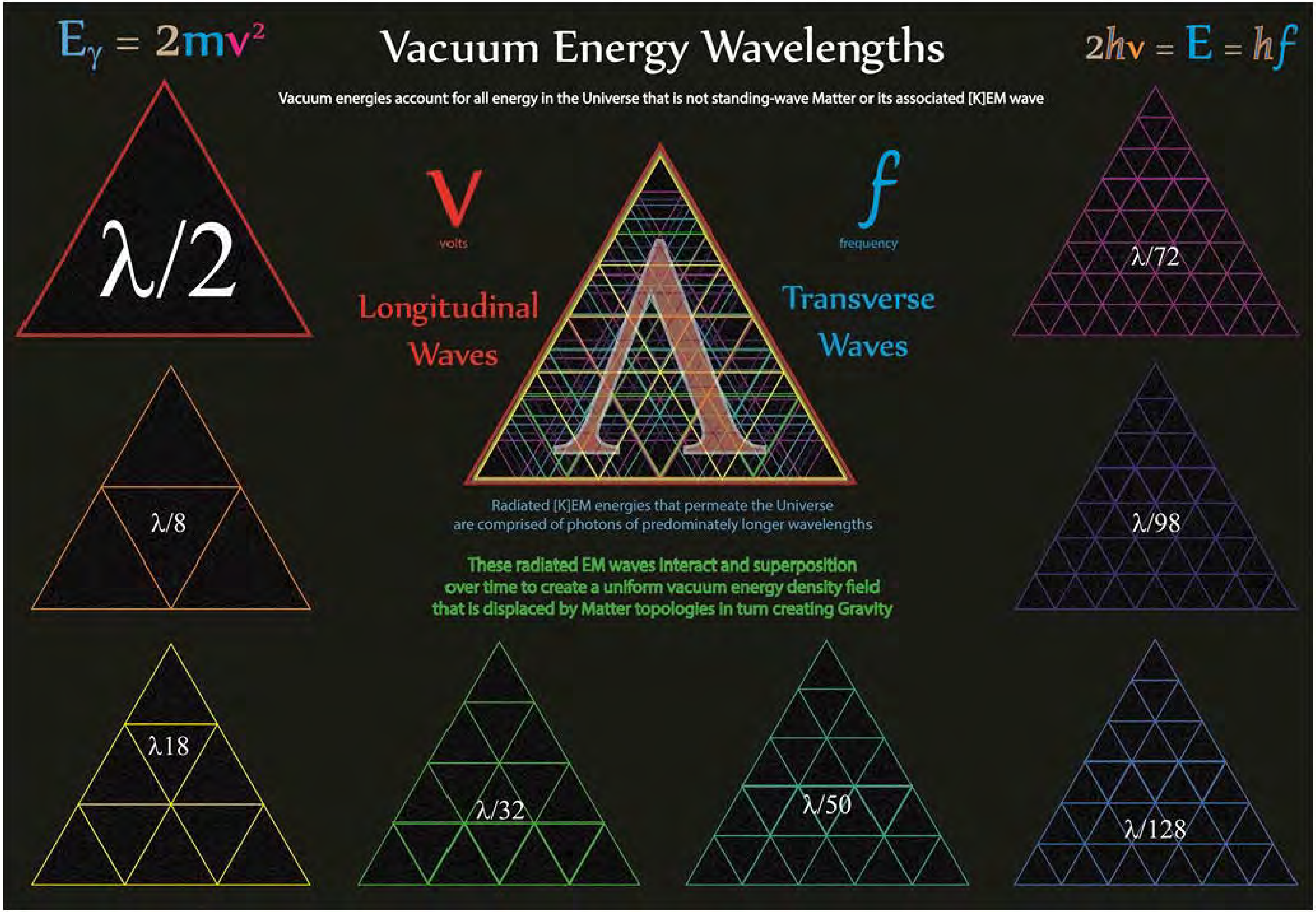
λ/2
V
volts
Longitudinal
Waves
Λ
f
frequency
Transverse
Waves
λ/72
λ/8
λ/98
λ18
λ/32
λ/50
λ/128

$E_\gamma = 2mv^2$

Vacuum Energy Wavelengths

$2hv = E = hf$

Vacuum energies account for all energy in the Universe that is not standing-wave Matter or its associated [K]EM wave

Radiated [K]EM energies that permeate the Universe
are comprised of photons of predominately longer wavelengths

These radiated EM waves interact and superposition
over time to create a uniform vacuum energy density field
that is displaced by Matter topologies in turn creating Gravity

Copyright ABRAHAM [2008] - All rights reserved

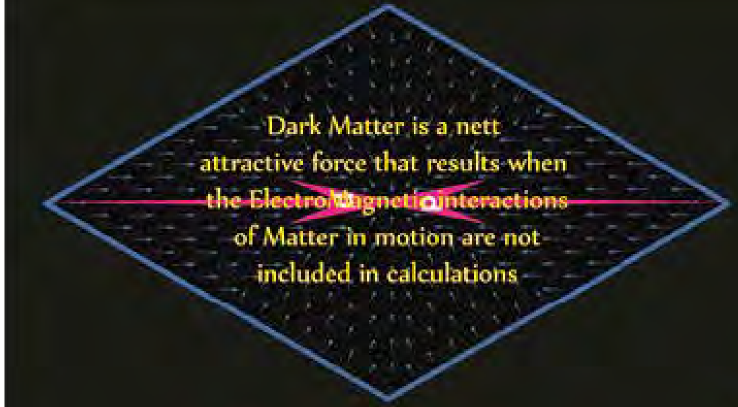

Dark Matter is a nett attractive force that results when the ElectroMagnetic interactions of Matter in motion are not included in calculations

In addition to Gravity the [K]EM waves of all Matter in motion possess a convergent component which in turn creates a unidirectional 'force'

The convergent EM force component can be modelled as being the result of an 'invisible' form of Matter or additional gravitational force

DARK MATTER

Convergent EM field vector momenta

All Matter emits divergent EM radiation photons comprised of bidirectional vector momenta

Matter produces Gravity
EM waves are interactive

Dark Matter is matter that neither emits nor scatters light or other electromagnetic radiation, and so cannot be directly detected via optical or radio astronomy.

Gravity is the result of the displacement of vacuum energies by 3D Matter topologies

All [K]EM radiation has a convergent component in its interactive field that increases when Matter is in motion or created in the form of EM standing-wave topologies

DARK MATTER

As Matter is destroyed and energy is released the attractive force of Gravity decreases but the convergent acceleration due to Dark Matter increases

Dark Matter

Dark matter is estimated to constitute 84% of the Matter in the universe and 23% of the total mass-energies

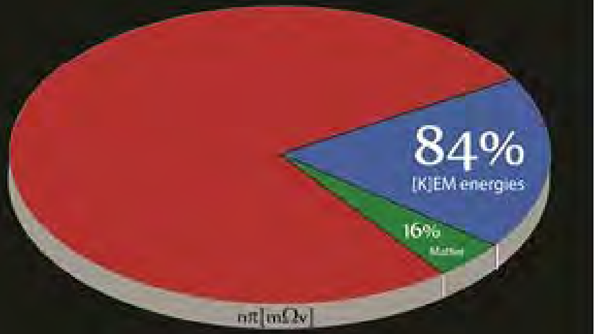

Dark Matter created by KEM field convergence contributes to the total energies of the cyclic Universe

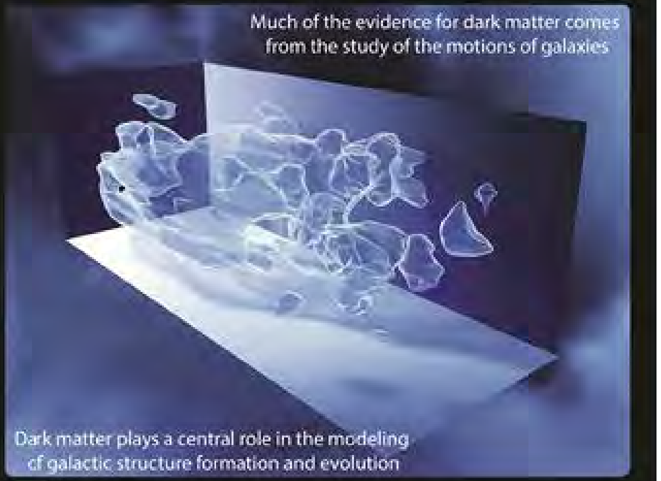

Much of the evidence for dark matter comes from the study of the motions of galaxies

Dark matter plays a central role in the modeling of galactic structure formation and evolution

Dark Matter's existence is inferred from its gravitational effects on visible Matter and the convergent lensing of background radiation

Galaxy Rotations

In 1986, Nobel laureate Hannes Alfven
postulated at an electrical model
for the solar system

The rotation curve of a galaxy (also called a velocity curve)
can be represented by a graph that plots the orbital speed (in km/s) of the stars or gas
in the galaxy on the y-axis against the distance from the center of the galaxy on the x-axis

The dynamic modelling
of large scale Matter must take
into account all forces present

Based on Newtonian mechanics and assuming, as was originally thought, that most of the Matter of the galaxy had to be in the galactic bulge near the center, Matter (such as stars and gas) in the disk portion of a spiral should orbit the center of the galaxy similar to the way in which planets in the solar system orbit the sun, i.e. where the average orbital speed of an object any distance away from the majority of the Matter would decrease inversely with the square root of the radius of the orbit

G

**Gravitational
Matter
is attractive**

Newtonian gravity
models convergent
Gravity only

Gravity alone produces
spheres or ellipsoids

Matter is comprised of charged quanta
ALL charged quanta experience
Lorentz forces

EM forces produce
dynamic spiral shapes

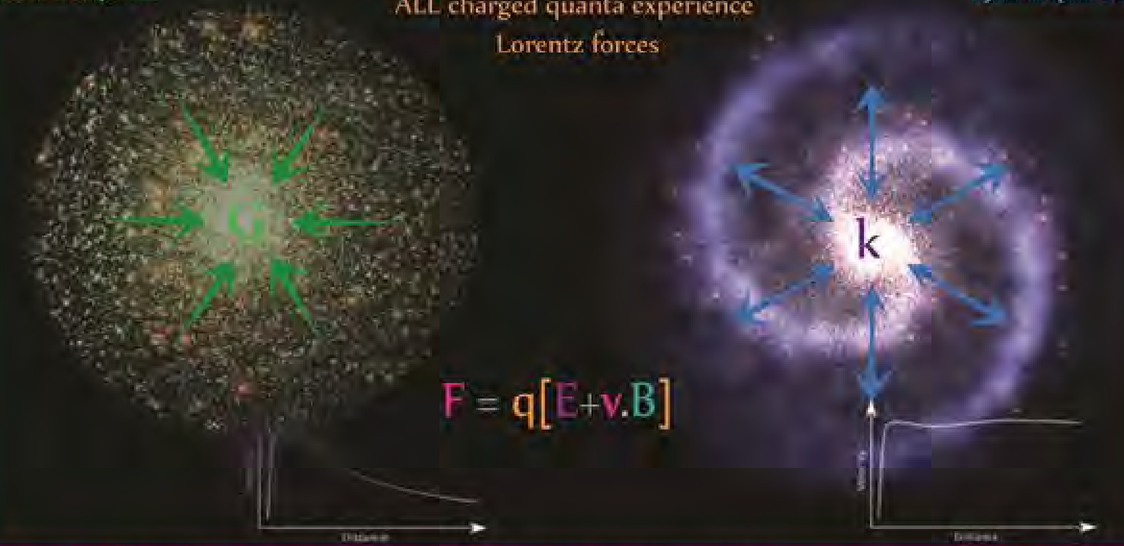

$$F = q[E+v.B]$$

SR

**ElectroMagnetic
mass-Energy
is interactive**

General relativity
models interactive
GEM energies

$$F = -G\frac{M_1}{r^2}$$

The curves of the rotational velocities of stars within Galaxies do not decrease in the expected inverse square root relationship but are "flat", i.e. outside of the central bulge their speed is nearly constant from the centre out ot the edge of the galaxy [identical to the forces produced by homopolar motors]

$$F = k\frac{Q_1}{r^2}$$

*Tetryonics dictates that we must take into account EM forces as well as Gravity
in order to account for any perturbations that convergent Gravitation alone cannot account for*

Observations have shown there is not sufficient visible Matter in the Universe to account for the apparent strength of gravitational forces within and between galaxies leading to the hypothesis that up to 90% of the Matter in the Universe is

*Gravity alone does not hold the whole galaxy together
It is many orders of magnitude weaker than ElectroMagnetic forces*

To account for the redshift-magnitude relation of type Ia supernovae General relativity requires that much of the energy in the Universe consists of a component with large negative pressure, this negative-gravity has been dubbed

Dark Matter

Gravitational attraction is the nett convergent force resulting from three distinct quantum forces

Dark Energy

Copyright ABRAHAM [2008] - All rights reserved

Dark Energy

Dark energy is a hypothetical form of energy that permeates all of space and tends to increase the rate of expansion of the universe.

EM mass-energies are conservative

Dark Energy is a nett repulsive force that results when Gravitational Matter topologies are destroyed by stellar GEM pinch processes

Matter is not conservative

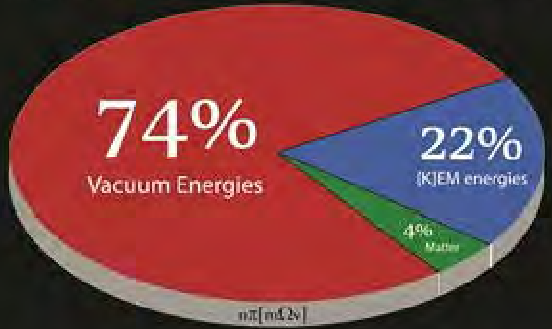

Independent from its actual nature, dark energy would need to have a strong negative pressure (repulsive force or divergent energy opposite that of gravitational attraction) In order to explain the observed acceleration in the expansion rate of the universe.

Dark Energy is the divergent scalar field created by [K]EM field energies

2D [K]EM fields
create interactive forces
between Matter

The cosmological constant is physically equivalent to vacuum energy

All [K]EM radiation has a divergent component in its interactive field that increases when Matter is in motion

DARK ENERGY

or its standing-wave topology is destroyed

As Matter is destroyed and energy is released the attractive force of Gravity decreases and the divergent acceleration due to Dark Energy increases

The [K]EM fields of Matter
DARK ENERGY
Divergent EM field interactions
KEM fields contributes to the total energies
of the cyclic Universe

DARK MATTER
Convergent EM field interactions

A major outstanding problem is that most quantum field theories predict a huge energy value for the quantum vacuum, more than 100 orders of magnitude too large.

This would need to be cancelled almost, but not exactly, by an equally large term of the opposite sign.

DARK ENERGY
Divergent EM field interactions

In fact a negative pressure does not influence the gravitational interaction between Matter [which remains attractive] but rather alters the overall evolution of the universe at the cosmological scale, typically resulting in the localised accelerating expansion of the universe despite the attraction between Matter present throughout the Universe.

Copyright ABRAHAM [2008] - All rights reserved

Energy-momenta geometry

The super-positioned EM fields of free space is comprised of Planck energy momenta

A major outstanding problem is that most quantum field theories predict a huge cosmological constant from the energy of the quantum vacuum, more than 100 orders of magnitude too large.

If the Higgs field does exist, another field with an unknown negative energy is needed to reduce the total energy in the vacuum to a slightly positive value to correspond with the current accelerated rate of expansion

Summing over all possible Energy quanta at all points in space gives a large but NOT infinite quantity of energy-momenta

Due to the equilateral asymmetry of all the energy momenta quanta in charged electromagnetic fields

Dark Energy is Divergent Vacuum Energy momenta in free Space

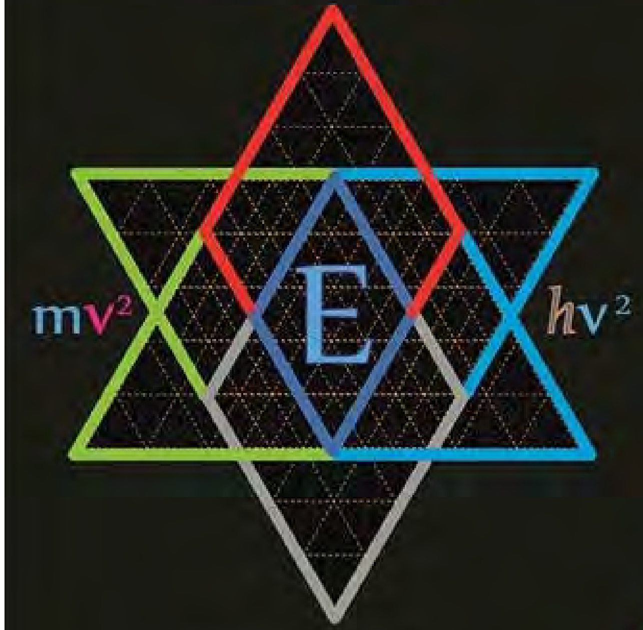

Convergent Vacuum Energy momenta in free Space is Dark Matter

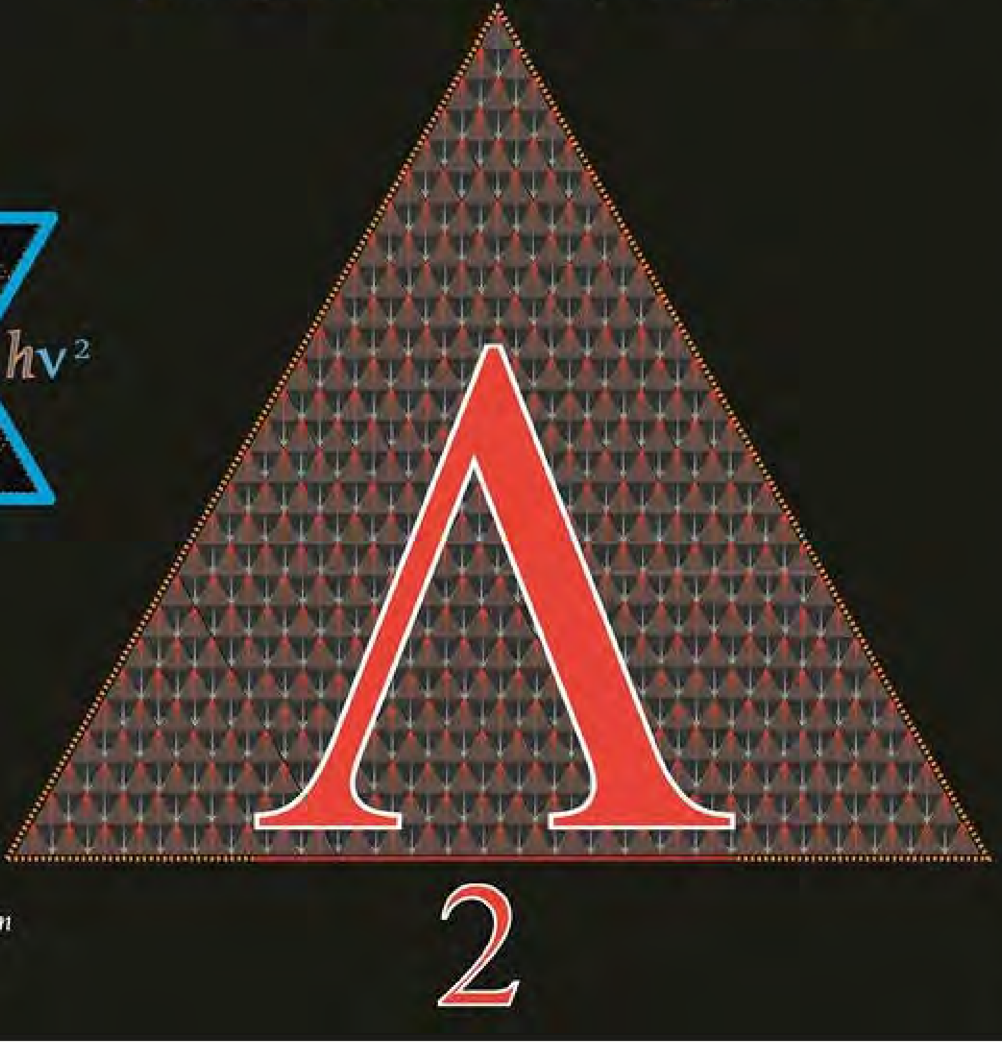

E Field Coulombic forces

convergent momenta

divergent momenta

Vacuum charge momentum

1	1		1
2	3		0.666667
3	5		0.6
4	7		0.571429
5	9		0.555556
6	11		0.545455
7	13		0.538462
8	15		0.533333
9	17		0.529412
10	19		0.526316
11	21		0.52381
12	23		0.521739
13	25		0.52
14	27		0.518519
15	29		0.517241
16	31		0.516129
17	33		0.515152
18	35		0.514286
19	37		0.513514
20	39		0.512821
21	41		0.512195
22	43		0.511628
23	45		0.511111
24	47		0.510638
25	49		0.510204
26	51		0.509804
27	53		0.509434
28	55		0.509091
29	57		0.508772
30	59		0.508475

The total Vacuum Energy in any spatial region contains fields of Planck energy-momenta seeking an equalised, homgenous state

Divergent energy momenta always dominates over Convergent energy momenta

Tetryonics 80.07 - Vacuum Energy-momenta

Charged electrostatic fields

E Field Coulombic forces

convergent momenta **divergent momenta**

Charge interaction

DIVERGENT CHARGE FIELDS
are external Energy momenta

momenta

CONVERGENT CHARGE FIELDS
are internal Energy momenta

Energy-momenta forces

Forward velocity **downward Thrust**

Momenta acceleration

1	1		1
2	3		0.666667
3	5		0.6
4	7		0.571429
5	9		0.555556
6	11		0.545455
7	13		0.538462
8	15		0.533333
9	17		0.529412
10	19		0.526316
11	21		0.52381
12	23		0.521739
13	25		0.52
14	27		0.518519
15	29		0.517241
16	31		0.516129
17	33		0.515152
18	35		0.514286
19	37		0.513514
20	39		0.512821
21	41		0.512195
22	43		0.511628
23	45		0.511111
24	47		0.510638
25	49		0.510204
26	51		0.509804
27	53		0.509434
28	55		0.509091
29	57		0.508772
30	59		0.508475

Electrostatic fields can provide a source of thrust

The greater the EM charge field strength the greater the divergent force

a source of cost effective space launches

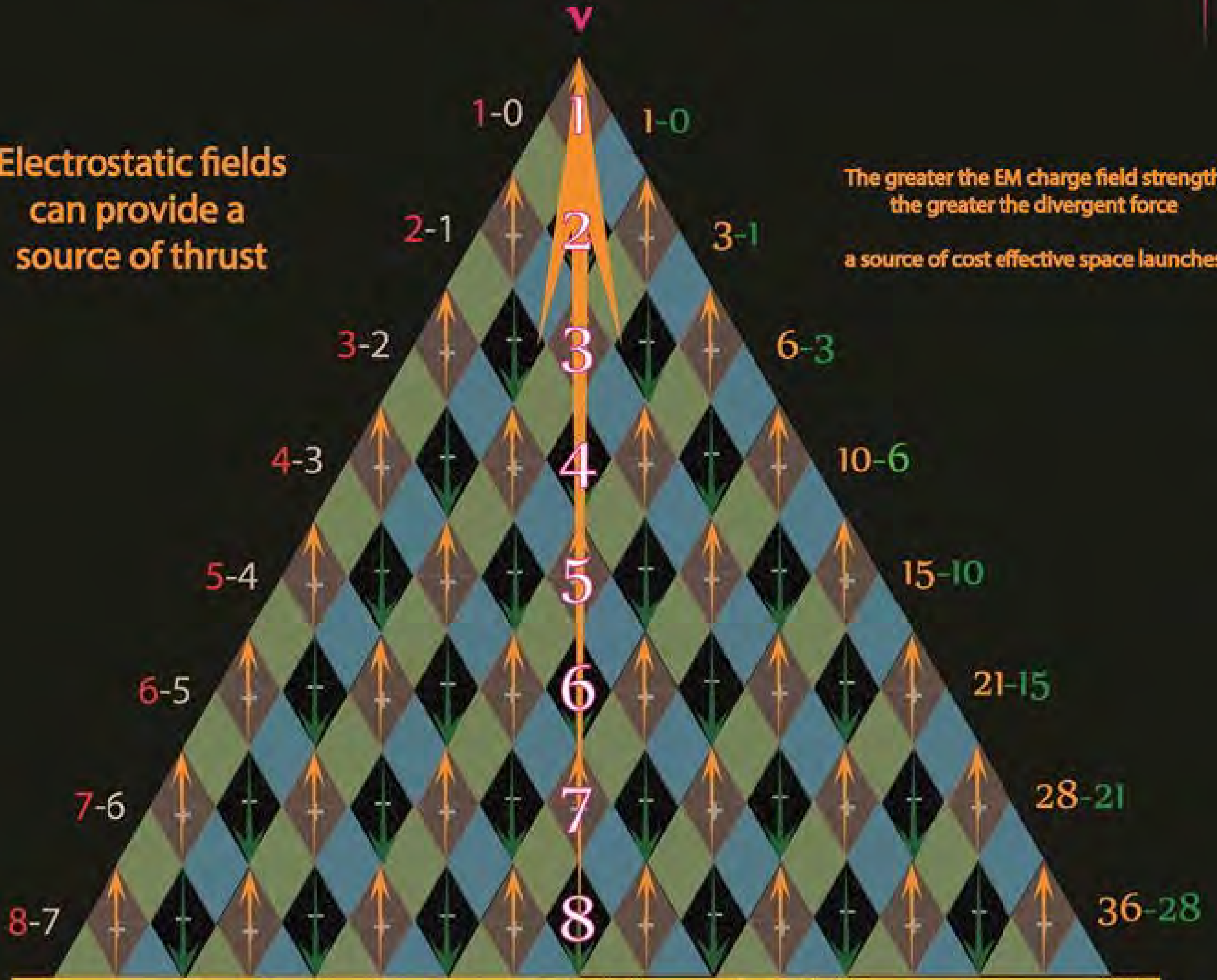

1	0
3	.25
2	.333
1,666	.375
1.5	.4
1.4	.4166
1.333	.4375
1,2857	.4375

7/15

#		Pos	Neg		Charge		Pos %	Neg %		Pos Split	Neg Split		Power		SQRT
1000		1000	999		1		0.50025	0.49975		500500	499500		1000000		1000

Launch Pad

Momentum splits are reversed
for Negative charges

Electric Charge divergence is always superior to Charge convergence

Momentum trends to 50/50 split
(but never reaches it)

Launch Pad

Tetryonics 80.08 - Electro-Gravitic Charge lauches

Copyright ABRAHAM [2008] - All rights reserved

Electrostatic lifters

Electrostatic levitation is the process of using an electric field to levitate a charged object and counteract the effects of gravity

Tetryonic geometries and Coulombic forces offers an alternative means of propulsion for space lauches and frictionless transports

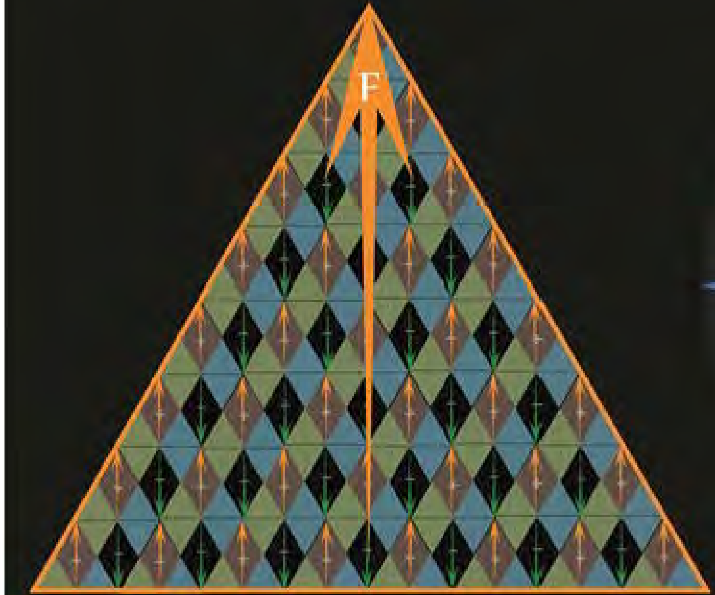

Charged voltage fields contain nett divergent energy momenta [Planck quanta] that can do work over distances

Electrostatic charges are always divergent from their sources

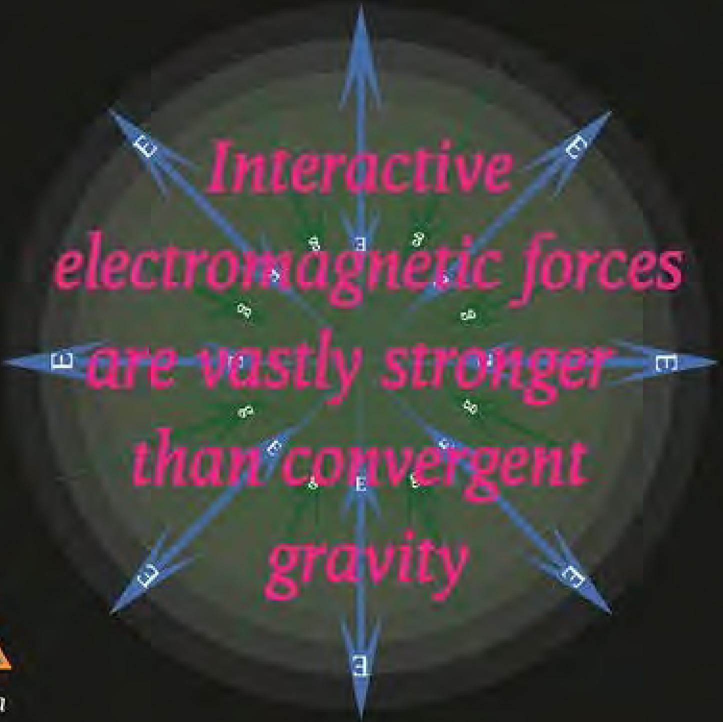

Interactive electromagnetic forces are vastly stronger than convergent gravity

Static electricity is the result of electrostatic charges

Electrostatic charges are comprised of divergent energy momenta

Electrostatic levitation

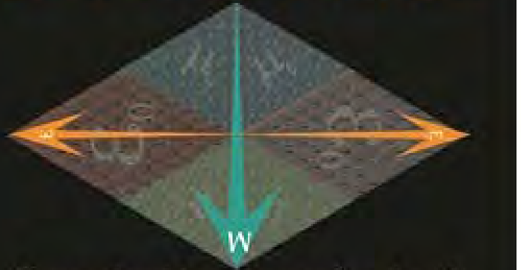

forces should not be confused with

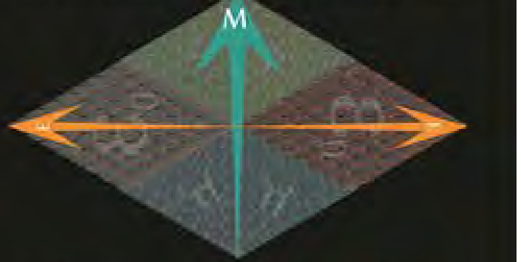

Magnetic levitation

Copyright ABRAHAM [2008] - All rights reserved

The Light speed barrier

BOSONS/PHOTONS
2D Planar massless particles
[EM mass-Energies]
'slice' throught the
vacuum energy aether

CHARGED
particle or massive body

ENERGY-MOMENTA
of populsion field

The Compton frequency of quanta
in any EM field is directly proportional
to the field's velocity of propagation

**The 'speed of Light'
is the limiting speed of
transverse EM accelerations
[it is not THE limit]**

$$E = Mv^2$$

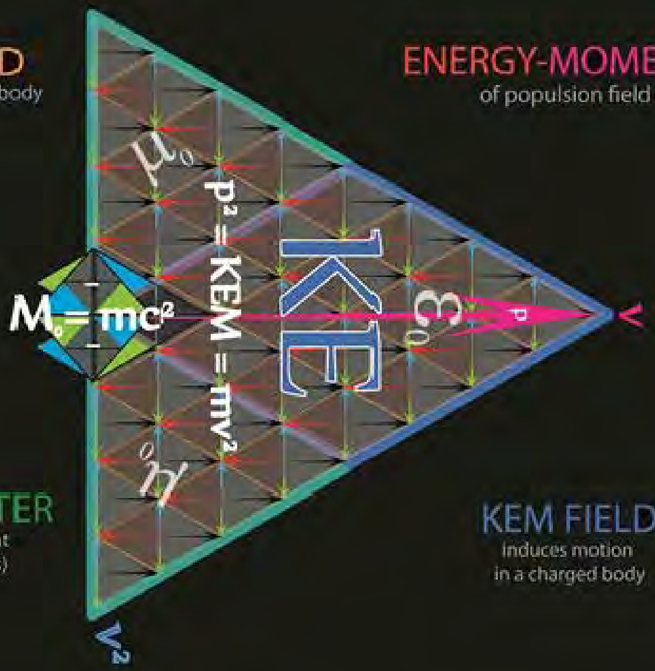

$$KE = \frac{1}{2}Mv^2$$

MOTION VECTOR
direction of propulsion
of a charged body

REST MATTER
remains invariant
(but accelerates)

KEM FIELD
induces motion
in a charged body

TETRYONS/BARYONS
3D Tetrahedral EM Matter
[4nπ EM mass-Energies]
interact at various angles with the
vacuum energy aether

**Tetryonic geometry clearly shows that here is NO limit to the
energy content that can be released into any spatial region,
negating the long held belief of a Light speed barrier**

V

According to special relativity, c is the maximum speed at which all EM mass-Energy and information in the universe can travel.
It is predicted by General relativity to be the absolute velocity of all 'massless particles' and associated fields
- Including ElectroMagnetic radiation such as Light in a vacuum.

Such particles and waves travel at c regardless of the motion of the source or the inertial frame of reference of the observer.
In the theory of relativity, c interrelates space and time, and appears in the famous equation of mass–energy equivalence E = mc2.

C

The speed at which light waves propagate in vacuum
is independent both of the motion of the wave source
and of the inertial frame of reference of the observer

$$E = n\pi\left[\overset{\text{Planck quanta}}{\left[\underset{\text{mass}}{m}\,\Omega\,\underset{\text{velocity}}{v^2}\right]}\right]$$

The speed of light in vacuum, usually denoted by c,
is a physical constant important in many areas of physics
and is determined by the impedance of the medium

Tetryonics 80.10 - The Light Speed Barrier

Inflation

While the detailed particle physics mechanisms responsible for inflation are not known, the basic picture makes a number of predictions that have been confirmed by observation.

Inflation is thus now considered part of the standard hot Big Bang cosmology

$$\underset{\text{momentum}}{\overset{\text{ENERGY}}{n\pi}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v}\right]^2\right]$$

v
p
KE
c
hv
v²
Γ
c²

All mass-energy propagates
at a velocity directly related to
the impedance of the medium

Tetryonics now provides a detailed physical model of ALL ponderable Matter and its kinetics through the charged geometry of mass-ENERGY-Matter

The energy in any spatial region
possesses a related mass-velocity
[linear momentum] component

The observable Universe could have originated from a region of pure energy in it's past.

This point would have contained all the Energy that ultimately formed all the mass-Matter of the Universe and Forces that we now observe and measure in two 2D planar fields

GIVEN SUFFICIENT ENERGIES ANY REGION OF SPACE-TIME WILL UNDERGO SUPERLUMINAL EXPANSION [INFLATION]

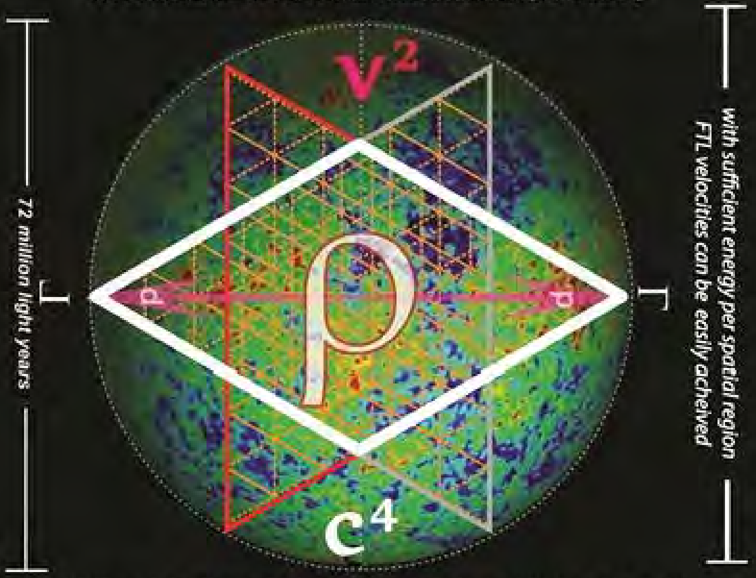
v²
ρ
p
p
c⁴
72 million light years
with sufficient energy per spatial region
FTL velocities can be easily achieved

UNTIL IT REACHES A CRITICAL ENERGY DENSITY THAT RESULTS IN THE OBSERVED SPEED OF LIGHT THAT WE NOW MEASURE

"And God said,
Let there be light:
and there was light"

The planar EM Energies released at that point would have resulted in faster-than-light divergence from that point as dictated by the geometric energy-momenta velocity relationship

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

It prescribes a dynamical EM Universe evolving over time

into a state of equilibrium between gravitational mass-Matter topologies and radiant, interactive mass-energy momenta geometries

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Superluminal velocities

REQUIRED VELOCITY

$$\beta = \left[\frac{v}{c}\right]$$

Linear correction factor

Any velocity is obtainable by Matter providing sufficient energy is released in very short durations to a region of space containing a charged body

$$\Sigma F = \frac{\Delta p}{dt} = m.\frac{dv}{dt}$$

W bosons — W bosons

$$[hv]- \quad E \quad +[hv]$$

decelerate — accelerate

Every quantum energy level change requires a related change in the number of quanta compared to the previous energy level (ad infinitum)

Note: inertial considerations still apply
Instanteous release of energy to acheive superluminal velocities is not advised

KEM fields contain energy-momenta

$$p^2 = KEM = mv^2$$

Any mass-energy field moving at superluminal velocities is still subject to Lorentz contractions

Any Matter held within, and moving along with, a superluminal EM field remains Lorentz invariant

$$\frac{mv^2}{c^2} = \frac{KEM}{c^2} = \frac{hv^2}{c^2}$$

The mass-energy of Kinetic EM fields is velocity dependent and subject to Lorentz corrections

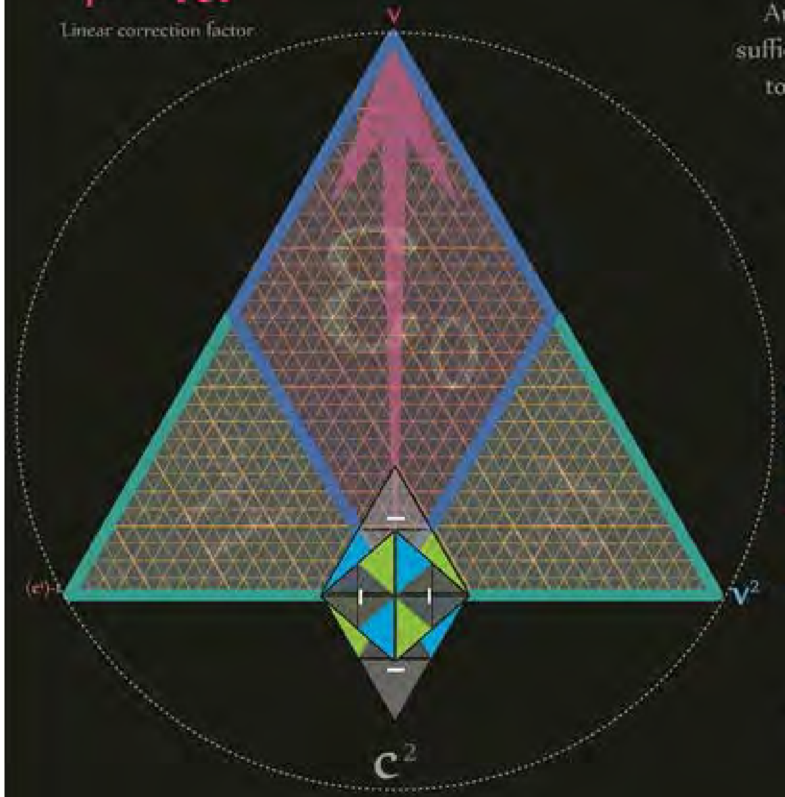

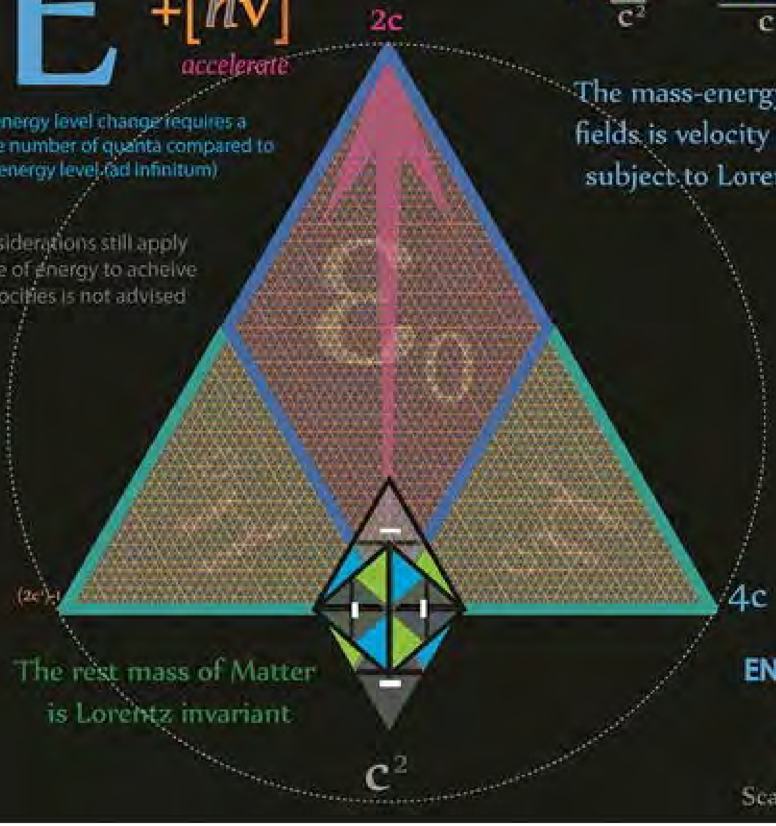

The speed of light is a limiting velocity for accelerations due to transverse wave electromagnetic energies

The rest mass of Matter is Lorentz invariant

Providing the KEM geometry contains enough energy per c^2 geometry the long held belief of 'the speed of light' being an ultimate speed barrier is shown to be invalid

ENERGY REQUIRED

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Scalar correction factor

Tetryonics 80.12 - Superluminial velocities

Copyright ABRAHAM [2008] - All rights reserved

Faster than Light Travel

It has been well established by Tetryonic geometry that the speed of light 'c' is not a insurmountable barrier as long expounded by Relativity theory

The natural speed of Light is easily surpassed given the right technologies (and sources of energy)

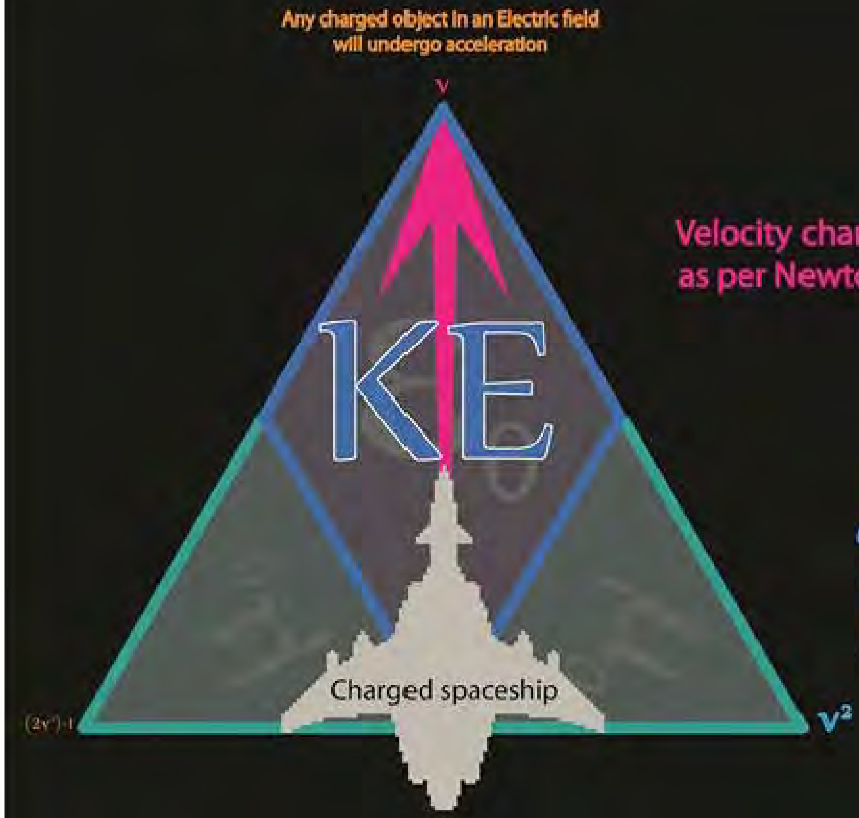
Any charged object in an Electric field will undergo acceleration
v
KE
Charged spaceship
$(2v^2)-1$
v^2

Velocity changes are additive as per Newtonian mechanics

The energy requirements to accelerate Matter to increasing velocities are exponental and follow equilateral geometries

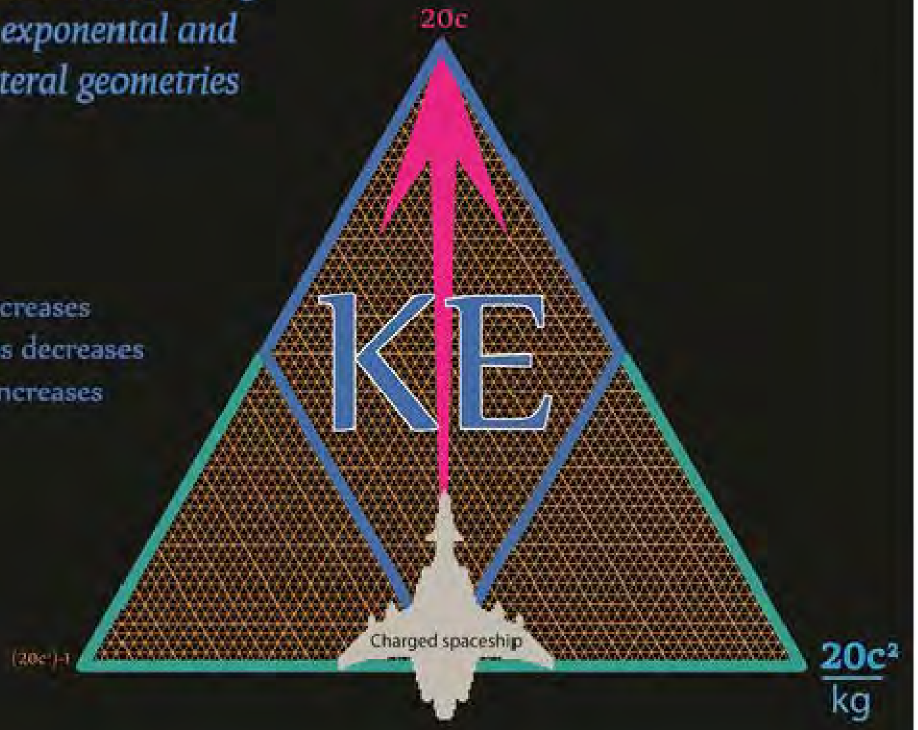
20c
KE
Charged spaceship
$(20c^2)-1$
$\frac{20c^2}{kg}$

As the Spaceship's Velocity increases the wavelength of Kinetic Energies decreases and the Compton Frequency increases

Warp Speed Scotty

Proposed explanation for GAMMA RAY BURSTS (GRB)

Any massive body using KEM fields as a motive force for faster-than-Light travel would produce an extremely bright release of Gamma rays at their point of acceleration (as only 1/2 of the photonic energy would be used to accelerate the craft in any one direction). additionally anybody noticing the burst of required Energy would not find a 'source' for the Gamma ray burst when they looked as the craft would have accelerated away from that point in free space

Tetryonics 80.13 - FTL Travel

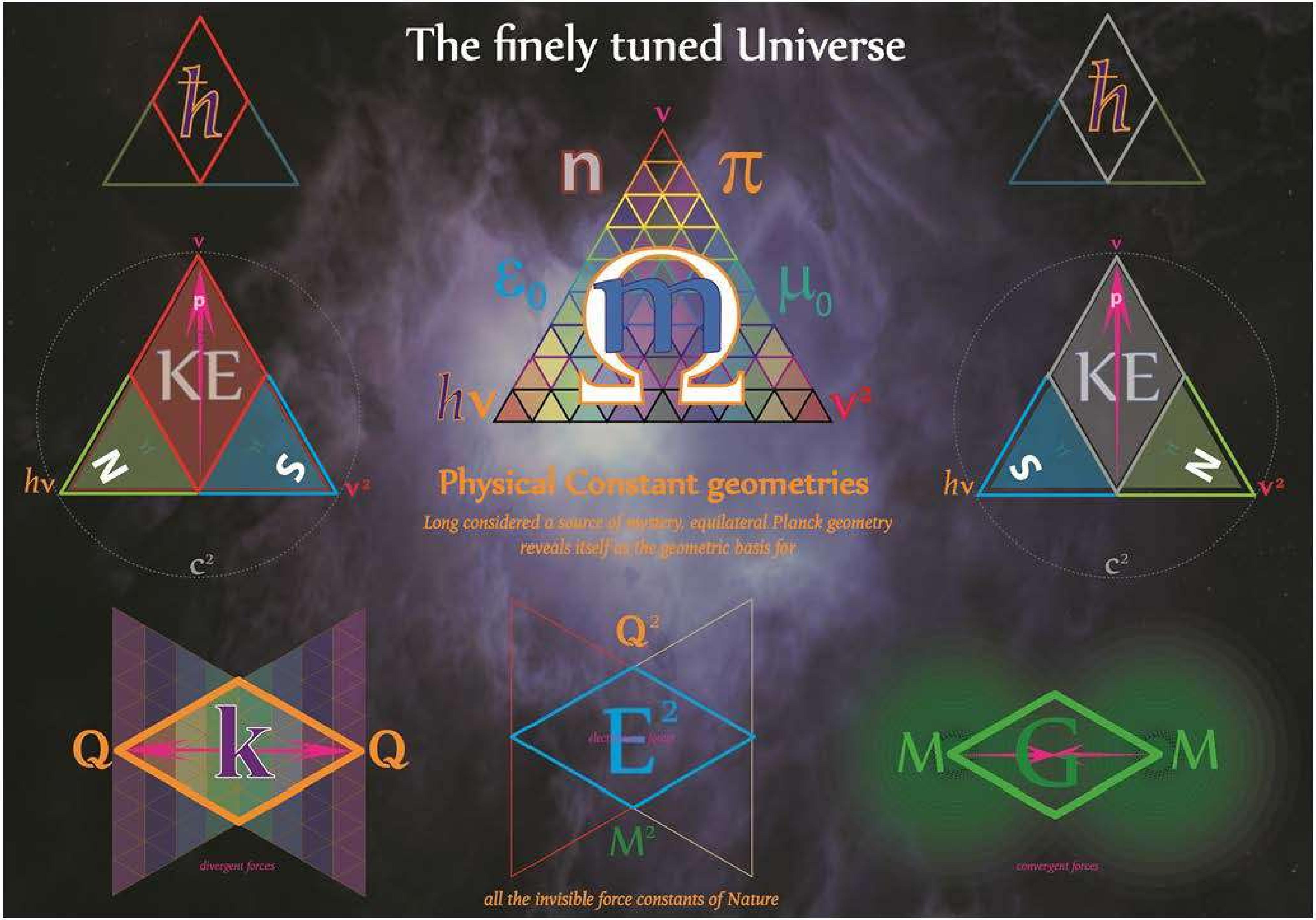

The finely tuned Universe

Physical Constant geometries

Long considered a source of mystery, equilateral Planck geometry reveals itself as the geometric basis for

all the invisible force constants of Nature

ħ
ħ
ν
n
π
ε0
μ0
m
Ω
hν
ν²
ν
p
KE
N
S
hν
ν²
c²
ν
p
KE
S
N
hν
ν²
c²
Q
k
Q
divergent forces
Q²
E²
M²
M
G
M
convergent forces

Tetryonics 80.14 - Physical Constant geometries

Conservation symmetry

A conservation law states that some quantity X
(describing any physical system)
remains constant throughout its motion

$$\frac{\Delta x}{dt} = 0$$

Rotation symmetry gives
conservation of angular momentum

conservation of
angular momenta

$$\left[\frac{\Omega}{c^2}\right]$$

QAM / second

conservation of
linear momentum

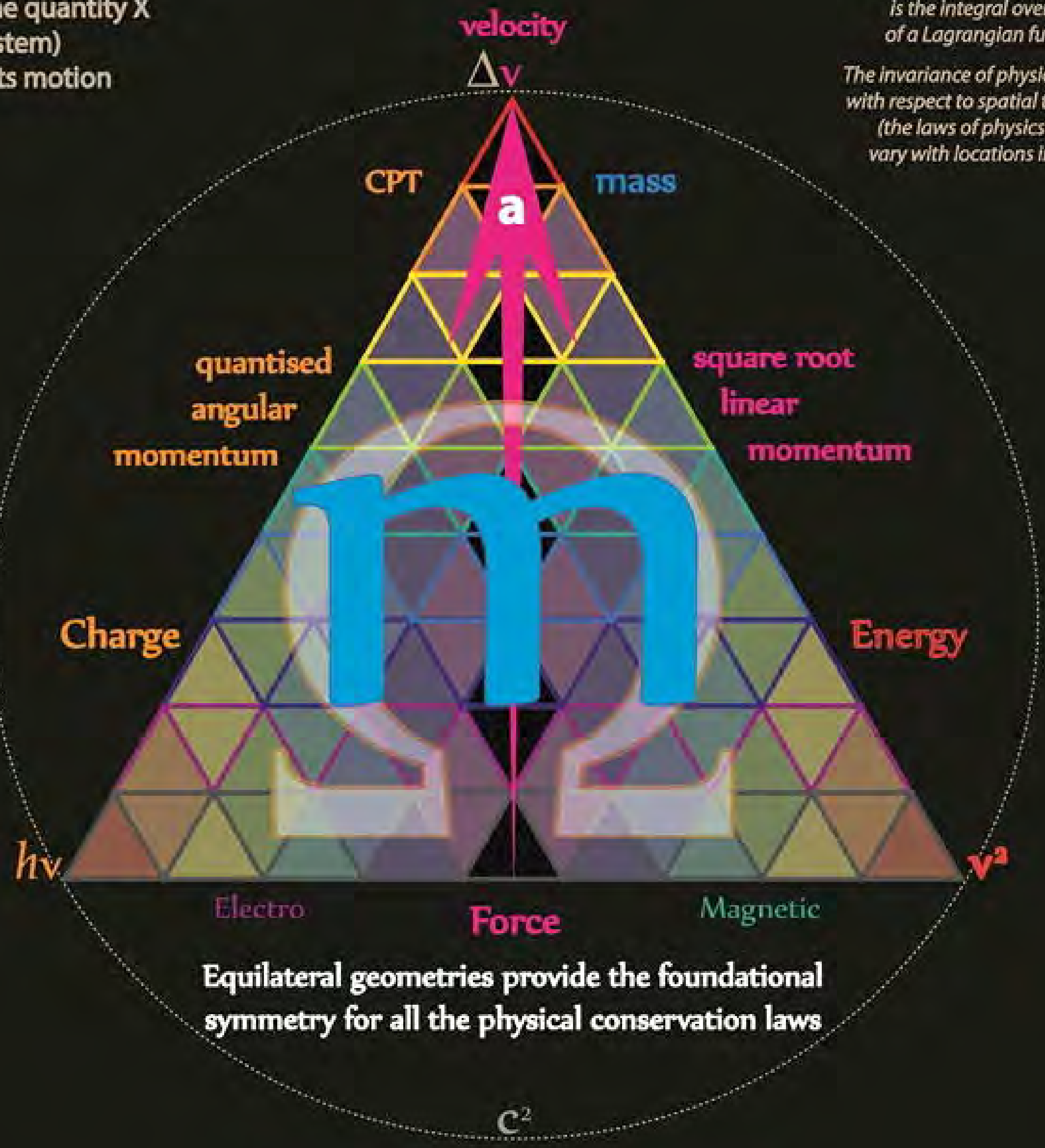

velocity
Δv
CPT
mass
a
quantised
angular
momentum
square root
linear
momentum
m
Charge
Energy
hv
v²
Electro
Force
Magnetic
c²

Equilateral geometries provide the foundational symmetry for all the physical conservation laws

Matter topologies are not conservative

Noether's Symmetry

The action of a physical system
is the integral over time
of a Lagrangian function

The invariance of physical systems
with respect to spatial translation
(the laws of physics do not
vary with locations in space)

Amalie Emmy Noether

(23 March 1882 – 14 April 1935)

Time translation symmetry
gives conservation of energy

conservation of
energy

v
ε_0
E
hf
hf
v²

6.629432672 e-34 J

Planck quanta

$$\left[m\Omega v^2\right]$$

mass velocity

7.376238634 e-51 kg

mass-ENERGY quanta

conservation of
mass

Tetryonics 80.15 - Conservation symmetry

Virtual Particles

In physics, a virtual particle is a particle that exists for a limited time and space.
The energy and momentum of a virtual particle are uncertain according to the uncertainty principle.
The degree of uncertainty of each is inversely proportional to
time duration (for energy) or to position span (for momentum).

In Tetryonics the Uncertainty principle has been negated with deterministic, re-normalisable charged mass-ENERGY geometries & Matter topologies

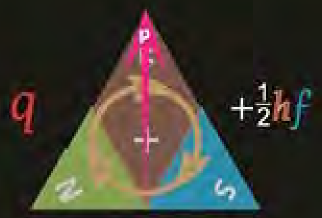
q
P
+
N
S
$+\frac{1}{2}hf$

Positive Zero Point Field
energy momenta

Longitudinal waves facilitate action-at-a-distance

Convergent Gravitation is attractive

Material Gravitation

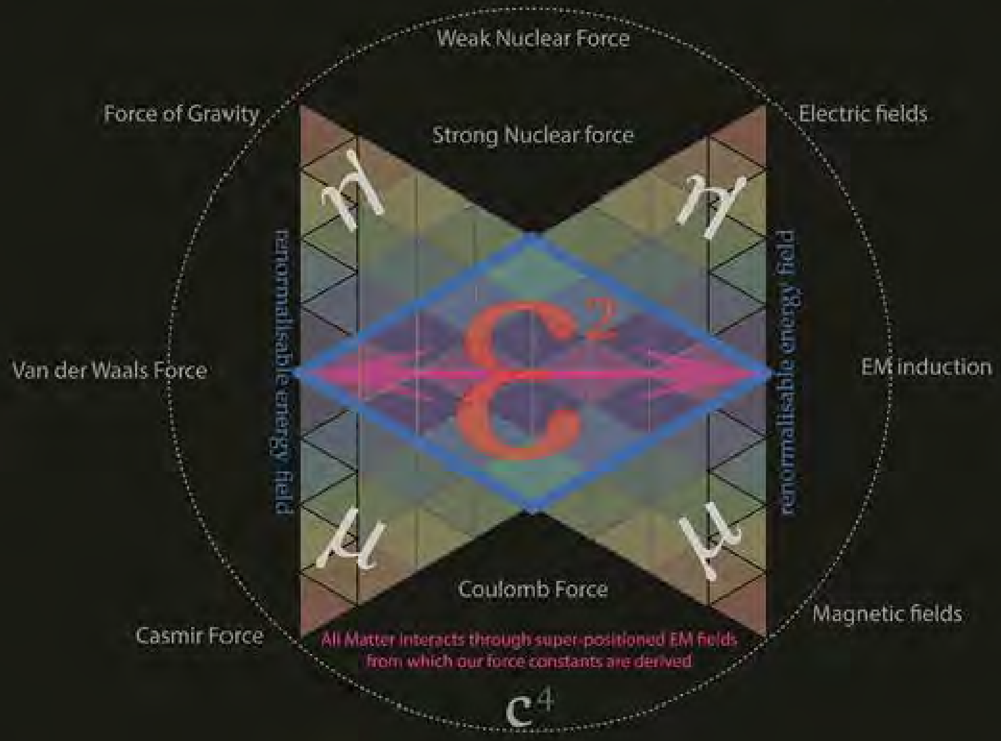
Weak Nuclear Force
Force of Gravity
Electric fields
Strong Nuclear force
η
η
renormalisable energy field
renormalisable energy field
Van der Waals Force
ε^2
EM induction
μ
μ
Coulomb Force
Casmir Force
Magnetic fields
All Matter interacts through super-positioned EM fields
from which our force constants are derived
c^4

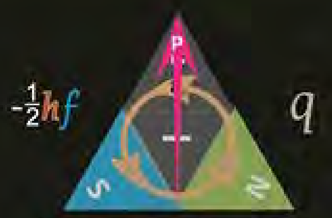
$-\frac{1}{2}hf$
P
−
S
N
q

Negative Zero Point Field
energy momenta

Energy propagates at the speed of light

Divergent Coulomb forces are interactive

EM Interactions

M

Charged Matter topologies in vacuum energy fields create Gravity

Vacuum field Energies

$$\underset{\text{EM mass}}{m^2c^4} = E^2 = \underset{\text{momenta}}{p^2c^2}$$

virtual Particles are not a required part of tetryonic theory

m

Superpositioned EM fields create interactive EM forces

Copyright ABRAHAM [2008] - All rights reserved

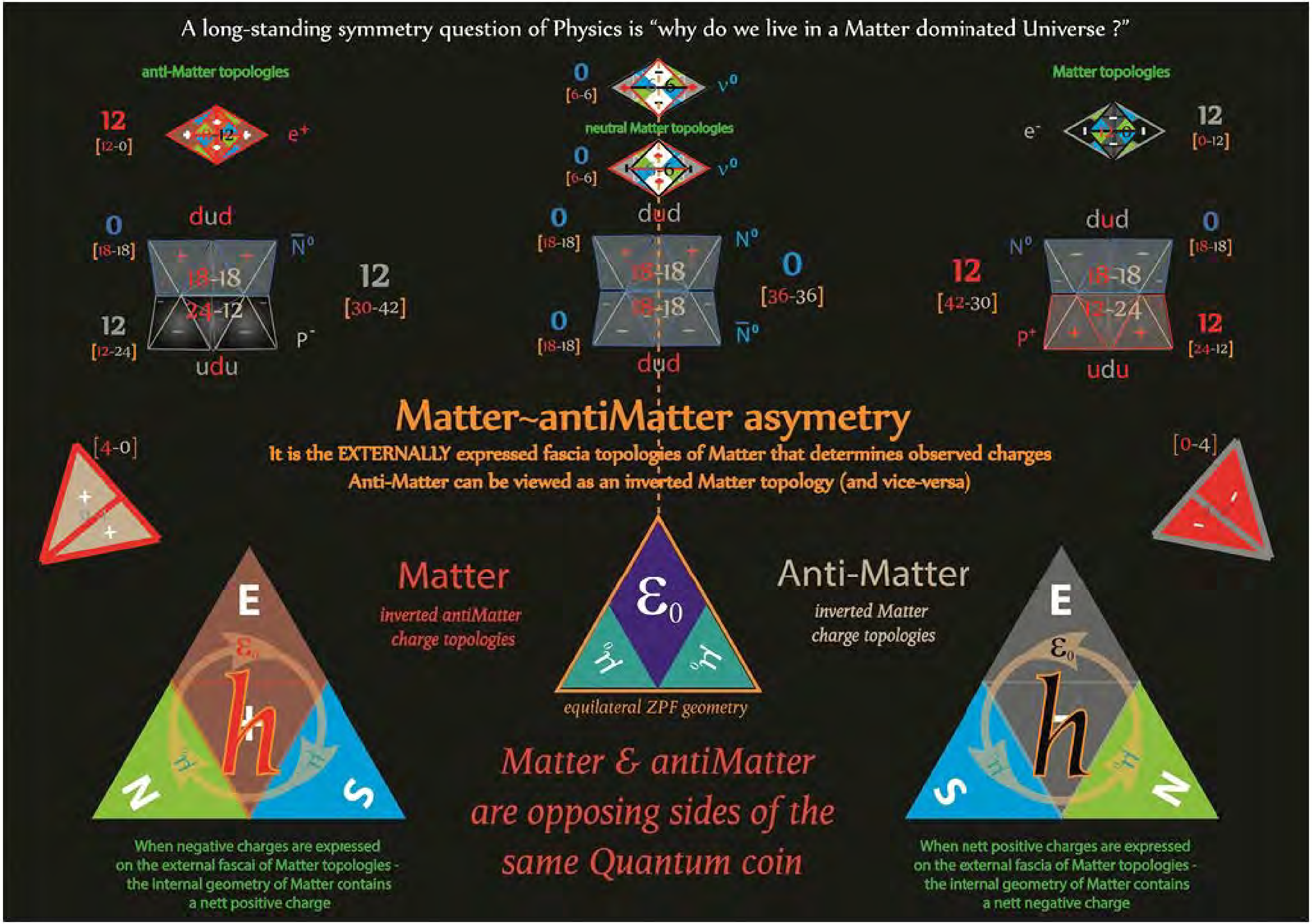

A long-standing symmetry question of Physics is "why do we live in a Matter dominated Universe ?"

anti-Matter topologies
neutral Matter topologies
Matter topologies
12
[12-0]
e+
0
[6-6]
ν0
0
[6-6]
ν0
e-
12
[0-12]
dud
0
[18-18]
N̄0
18-18
24-12
12
[12-24]
P-
udu
12
[30-42]
dud
0
[18-18]
N0
18-18
18-18
0
[18-18]
N̄0
dud
0
[36-36]
12
[42-30]
N0
dud
18-18
12-24
0
[18-18]
p+
12
[24-12]
udu
[4-0]
[0-4]
E
N
S
h
ε0
μ0
E
S
N
h
ε0
μ0

Matter~antiMatter asymetry

It is the EXTERNALLY expressed fascia topologies of Matter that determines observed charges
Anti-Matter can be viewed as an inverted Matter topology (and vice-versa)

Matter

inverted antiMatter charge topologies

Anti-Matter

inverted Matter charge topologies

equilateral ZPF geometry

Matter & antiMatter are opposing sides of the same Quantum coin

When negative charges are expressed on the external fascai of Matter topologies - the internal geometry of Matter contains a nett positive charge

When nett positive charges are expressed on the external fascia of Matter topologies - the internal geometry of Matter contains a nett negative charge

Tetryonics 80.17 - Charged fascia chirality

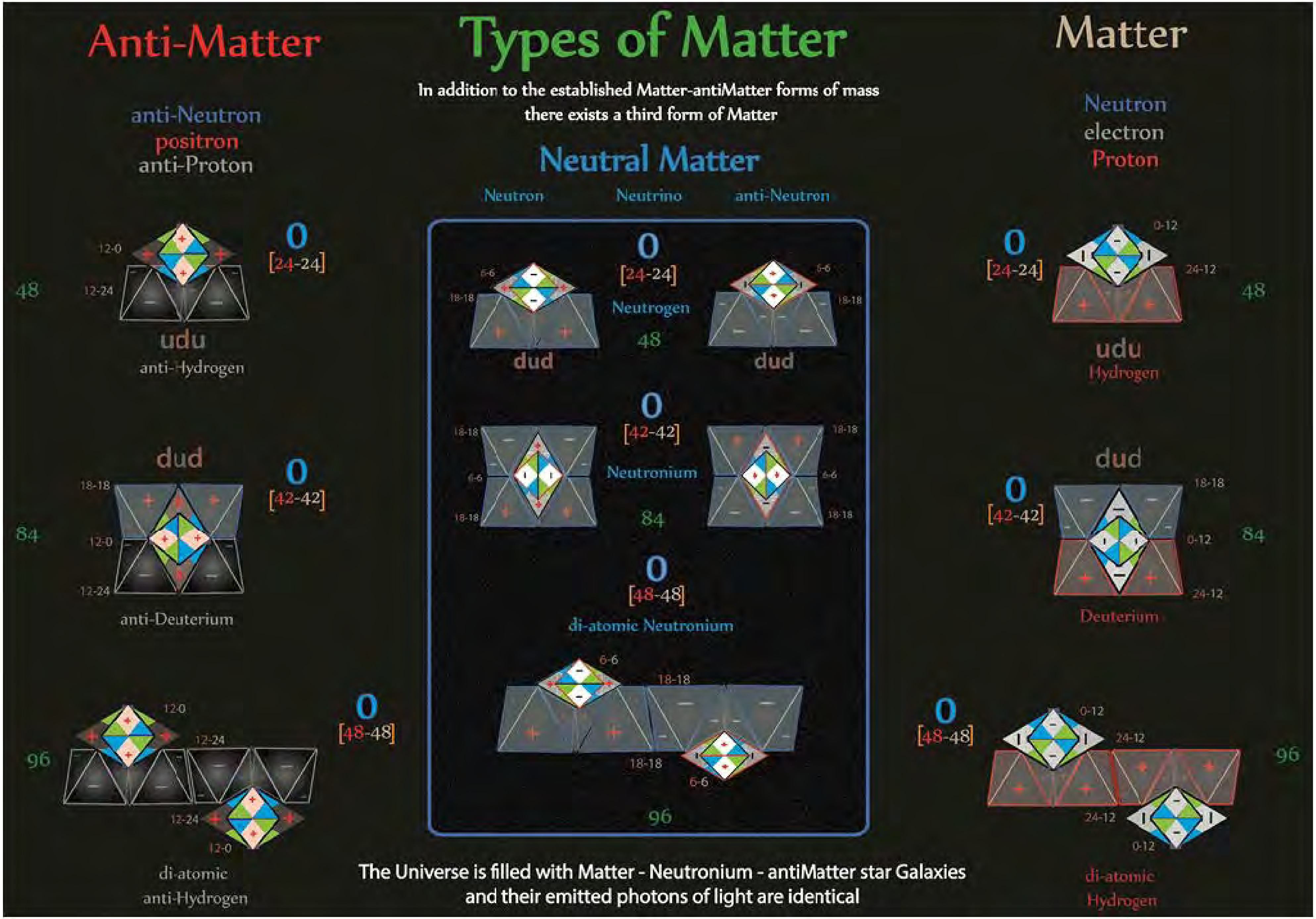

Types of Matter

In addition to the established Matter-antiMatter forms of mass
there exists a third form of Matter

Anti-Matter

anti-Neutron
positron
anti-Proton

Neutral Matter

Neutron Neutrino anti-Neutron

Matter

Neutron
electron
Proton

0
[24-24]
12-0
12-24
48
udu
anti-Hydrogen
dud
0
[42-42]
18-18
12-0
12-24
84
anti-Deuterium
0
[48-48]
12-0
12-24
12-24
12-0
96
di-atomic
anti-Hydrogen
0
[24-24]
Neutrogen
48
6-6
18-18
dud
6-6
18-18
dud
0
[42-42]
Neutronium
84
18-18
6-6
18-18
18-18
6-6
18-18
0
[48-48]
di-atomic Neutronium
6-6
18-18
18-18
6-6
96
0
[24-24]
0-12
24-12
48
udu
Hydrogen
dud
0
[42-42]
18-18
0-12
24-12
84
Deuterium
0
[48-48]
0-12
24-12
24-12
0-12
96
di-atomic
Hydrogen

The Universe is filled with Matter - Neutronium - antiMatter star Galaxies
and their emitted photons of light are identical

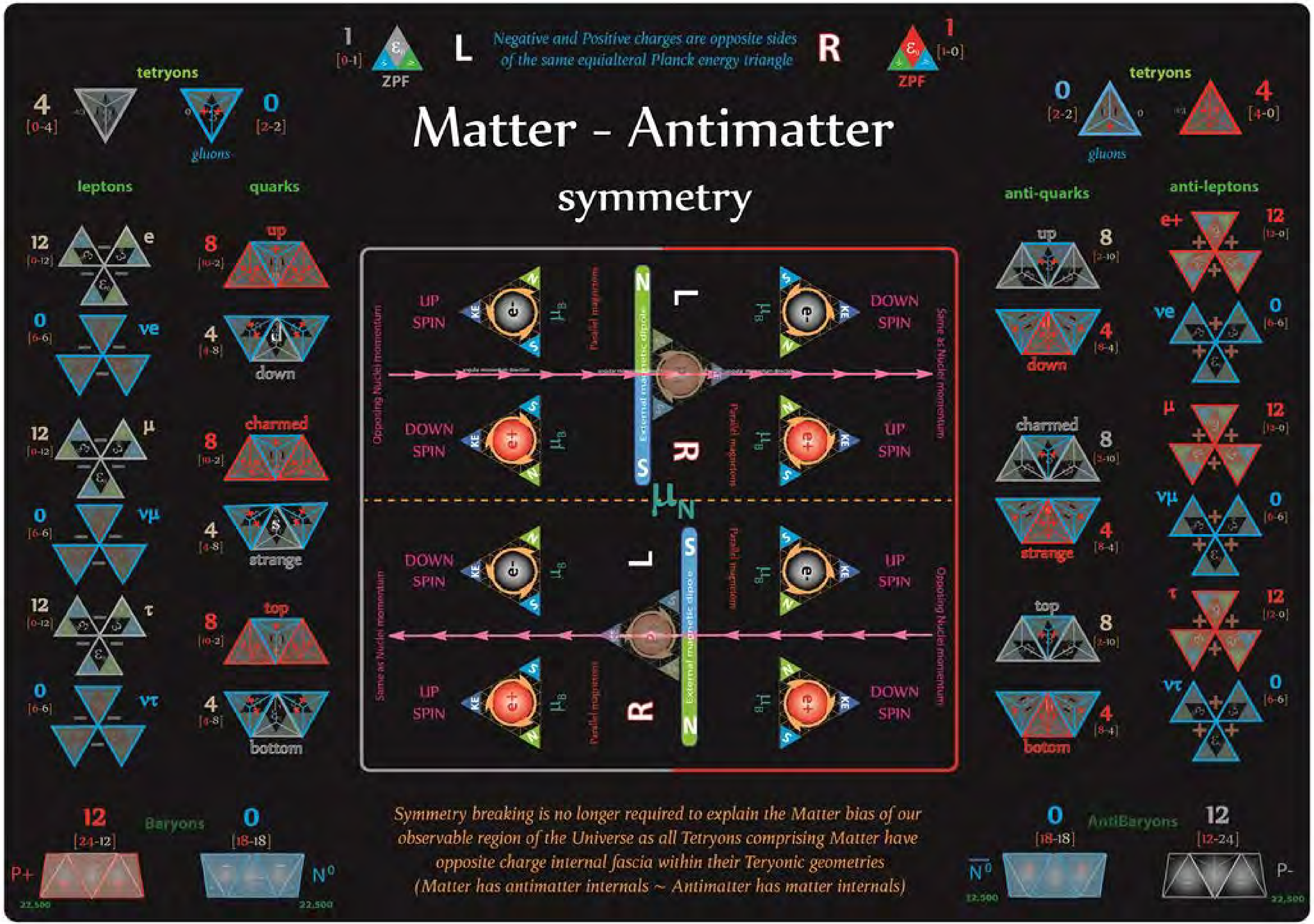

Matter - Antimatter

symmetry

Negative and Positive charges are opposite sides of the same equilateral Planck energy triangle

Symmetry breaking is no longer required to explain the Matter bias of our observable Region of the Universe as all Tetryons comprising Matter have opposite charge internal fascia within their Teryonic geometries (Matter has antimatter internals ~ Antimatter has matter internals)

1
[0-1]
ZPF
L
R
1
[1-0]
ZPF
tetryons
4
[0-4]
0
[2-2]
gluons
tetryons
0
[2-2]
gluons
4
[4-0]
leptons
quarks
anti-quarks
anti-leptons
12
[0-12]
e
0
[6-6]
ve
12
[0-12]
μ
0
[6-6]
vμ
12
[0-12]
τ
0
[6-6]
vτ
8
[10-2]
up
4
[4-8]
down
8
[10-2]
charmed
4
[4-8]
strange
8
[10-2]
top
4
[4-8]
bottom
up
8
[2-10]
down
4
[8-4]
charmed
8
[2-10]
strange
4
[8-4]
top
8
[2-10]
botom
4
[8-4]
e+
12
[12-0]
ve
0
[6-6]
μ
12
[12-0]
vμ
0
[6-6]
τ
12
[12-0]
vτ
0
[6-6]
UP SPIN
DOWN SPIN
DOWN SPIN
UP SPIN
DOWN SPIN
UP SPIN
UP SPIN
DOWN SPIN
e-
e+
KE
N
S
μ_B
Parallel magnetism
External magnetic dipole
Opposing Nuclei momentum
Same as Nuclei momentum
L
R
μ_N
12
[24-12]
Baryons
0
[18-18]
P+
N⁰
22,500
22,500
0
[18-18]
AntiBaryons
12
[12-24]
$\overline{N}^0$
P-
22,500
22,500

Tetryonics 80.19 - Matter - Antimatter symmetry

Anti-Matter

Baryon asymmetry

Matter

The Big Bang should have produced equal amounts of Matter and antiMatter, as such, there should have been total cancellation of both.

It is not yet understood why the Universe has more matter than antimatter.
It is generally assumed that when the Universe was young and very hot,
it was in equilibrium and contained equal numbers of baryons and antibaryons.

Protons should have cancelled with antiprotons, electrons with antielectrons (positrons),
neutrons with antineutrons, and so on for all elementary particles.
This would have resulted in a sea of photons in the universe with no matter.

However, observations suggest that the Universe, including its most distant parts,
is made almost entirely of normal matter.

dud

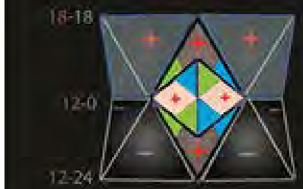

0

[42-42]

84

anti-Deuterium

dud

0

[42-42]

84

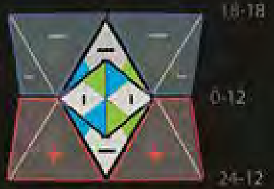

Deuterium

anti-Neutron
positron
anti-Proton

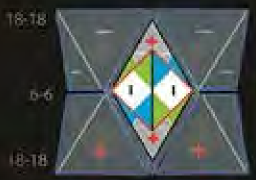

0

[42-42]

Neutronium

84

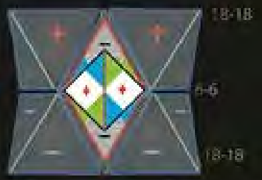

Neutron
electron
Proton

antiMatter dominated Galaxy

Matter dominated Galaxy

Neutral Matter

Neutron

Neutrino

anti-Neutron

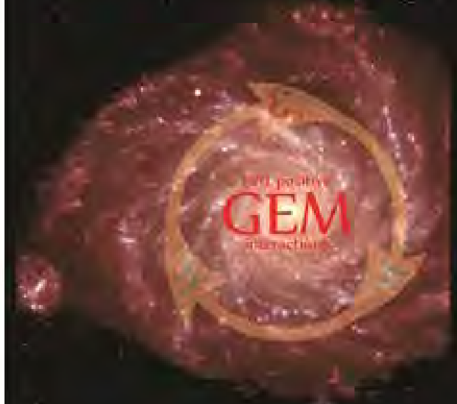

Clockwise rotation

Tetryonics dictates that the Universe is filled with Matter - Neutronium - Antimatter star Galaxies and their emitted light photons are identical

Tetryonics also demonstrates that the charge topologies of each Matter type prevalent in the formation of each Galaxy determines the overall resultant EM field geometries that shapes their large-scale evolution

It is hypothesised that the rotation direction of galaxies is also influenced by how each galaxy was formed
in particular the predominance of Matter over antiMatter interactions and
the electrodynamic fields of each Galaxy's GEM pinch core

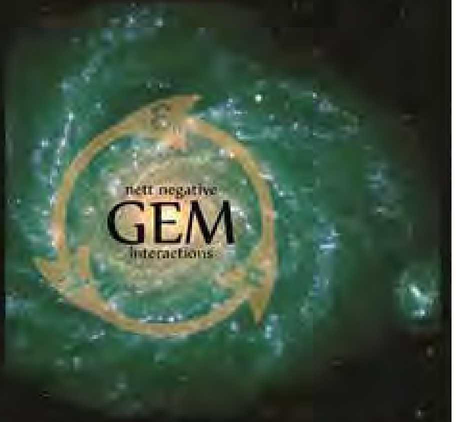

Counter-clockwise rotation

Opposites attract - Similars repel
even on the galactic-scale

Galactic Matter distributions

Tetryonics dictates that the Universe is filled with 3 forms of
Matter - Neutronium - Antimatter stars and Galaxies
their light is identical and cannot be used to distingush them

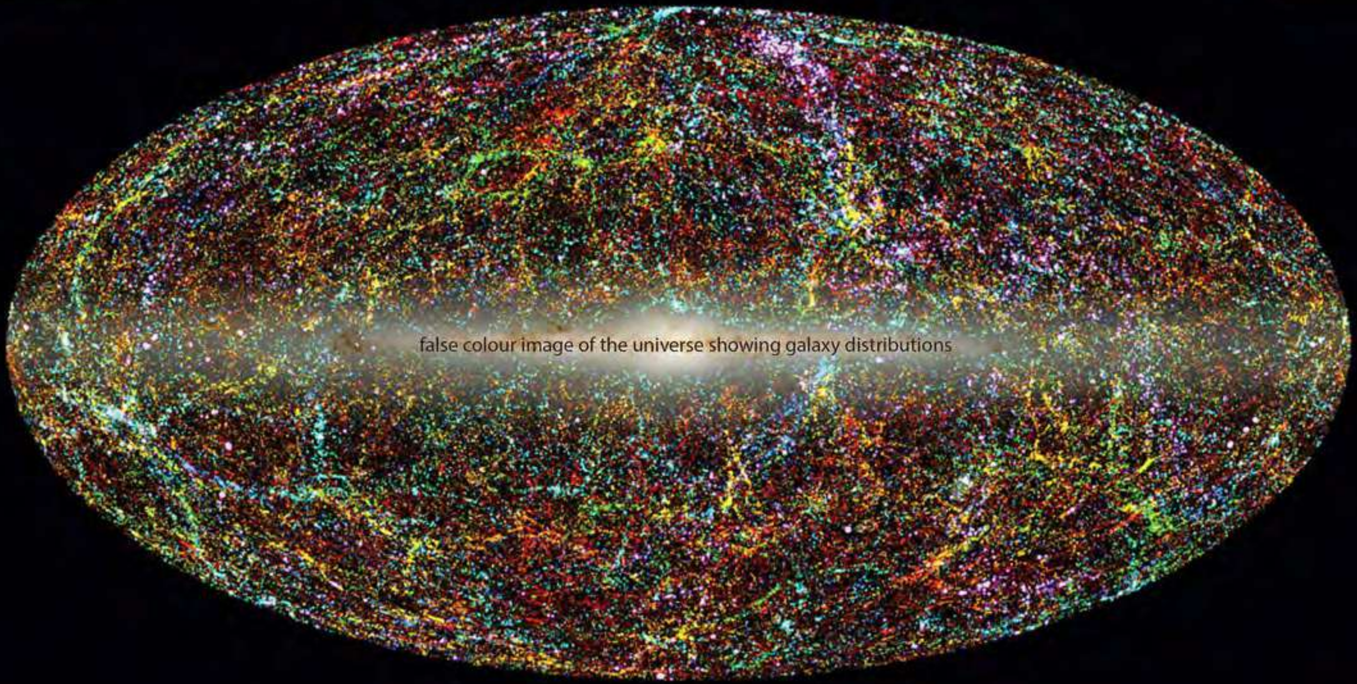

false colour image of the universe showing galaxy distributions

Matter

Neutron
Electron
Proton

Neutronium

Neutron
Neutrino
anti-Neutron

Neutronium is found predominantly in Neutron stars and Pulsars
spread throughout Matter and antiMatter galaxies
however it is conceivable that neutral galaxies may exist

Anti-Matter

anti-Neutron
Positron
anti-Proton

Tetryonics 80.21 - Galactic Matter distribution

E

scalar Energy

$\mathrm{kg}\,\frac{\mathrm{m}^2}{\mathrm{s}^2}$

Planck

$$h = m.\Omega$$

quantised angular momentum is equilateral geometry

$$E = h\nu^2$$

h

Planck's constant

$\mathrm{kg}\,\frac{\mathrm{m}^2}{\mathrm{s}}$

Leibniz-Newton

linear momentum

$$E = mv^2$$

mass-energy momenta relationship

square root equilateral energy is linear momentum

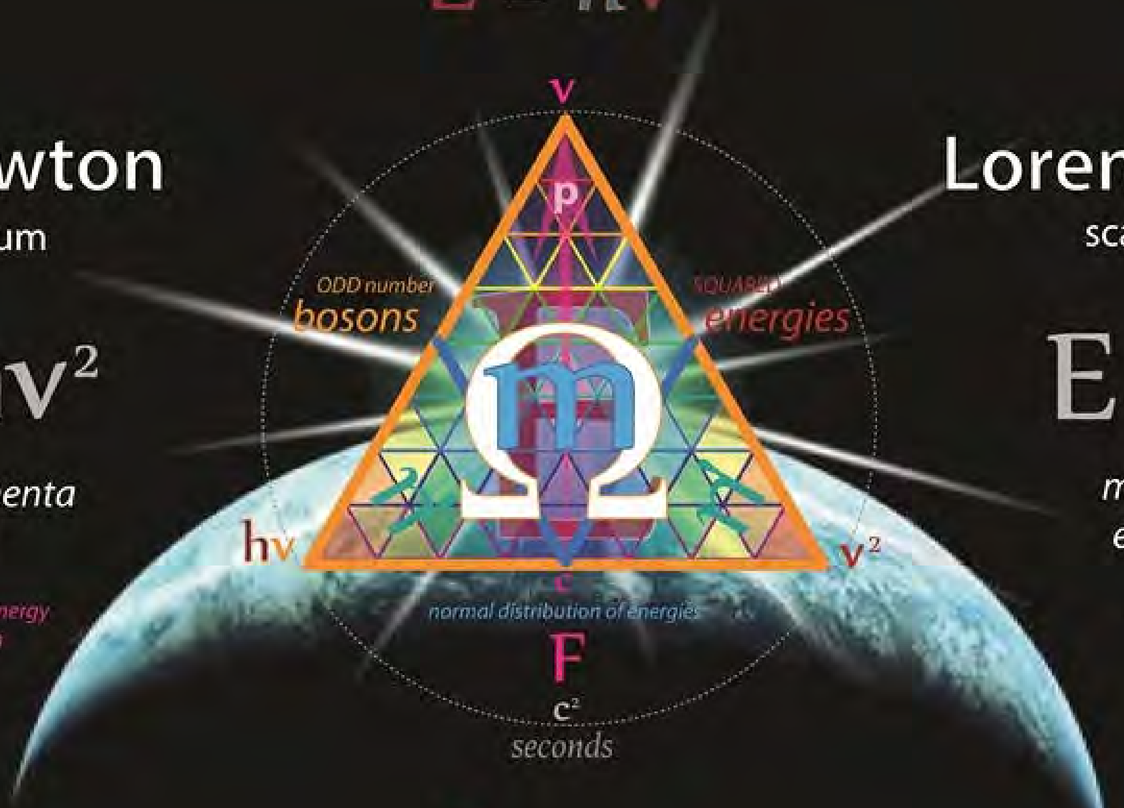
v
p
ODD number
bosons
SQUARED
energies
m
Ω
hv
v²
c
normal distribution of energies
F
c²
seconds

Lorentz-Einstein

scalar energies

$$E = \frac{mv^2}{c^2}$$

mass-energy equivalence

Energy per second is mass

p

linear momentum

$\mathrm{kg}\,\frac{\mathrm{m}}{\mathrm{s}}$

$$F = ma$$

mass is a scalar constant relating Force to acceleration

$$F = m\frac{\Delta v}{\Delta t}$$

m

scalar mass

kg

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic Theory

The charged geometry of ElectroMagnetic mass-ENERGY
& Matter in motion

Charged geometries

Bosons

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta
mass velocity

Total Relativistic Energies

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

Planck quanta
mass velocity

Quantum Mechanics

Chemistry

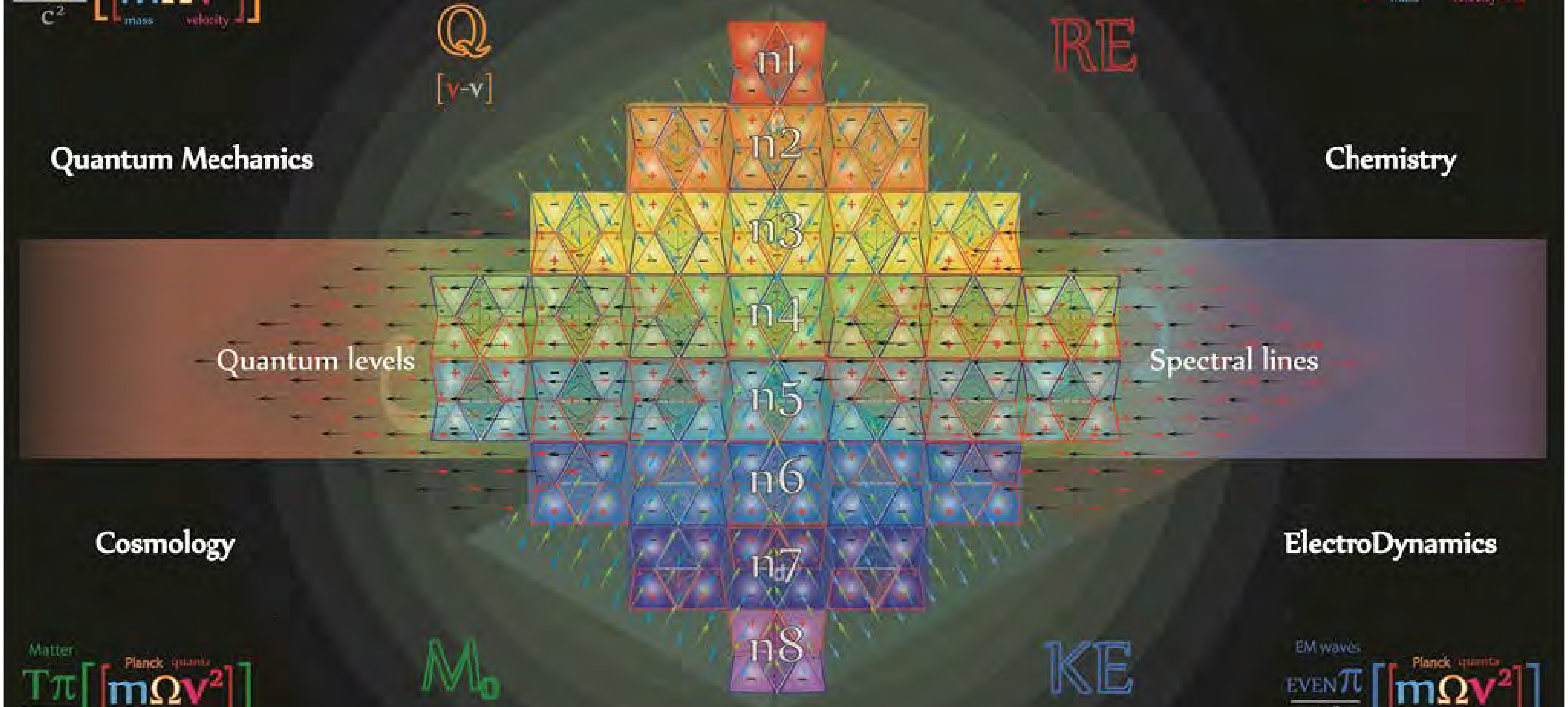

Cosmology

ElectroDynamics

Matter

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta
mass velocity

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta
mass velocity

mass/Matter topologies

Unifying QM, QED, Chemistry and Cosmology
through charged equilateral Planck mass-energy momenta

Kinetic Energies of motion

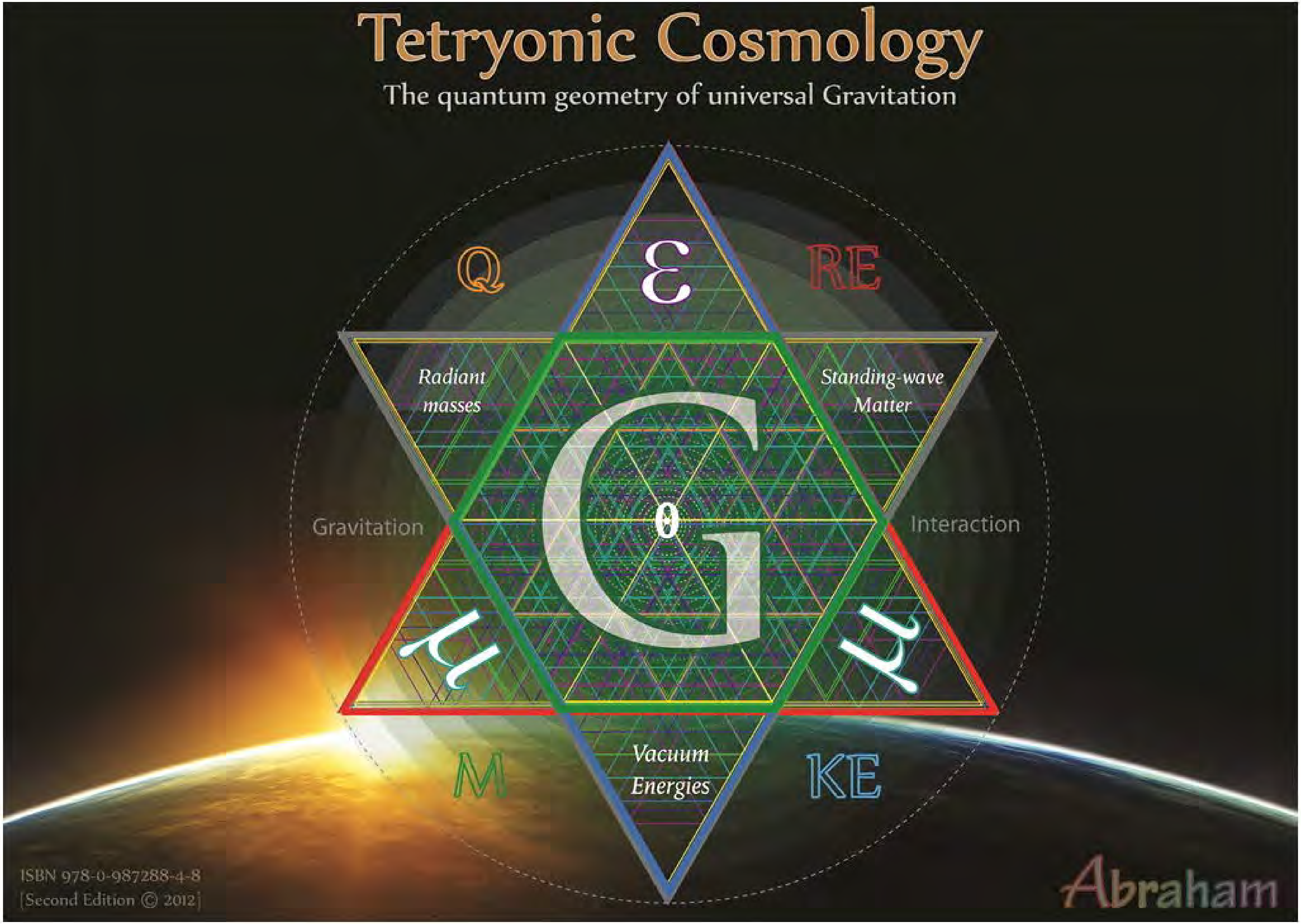
Tetryonic Cosmology
The quantum geometry of universal Gravitation
Q
ε
RE
Radiant
masses
Standing-wave
Matter
G
0
Gravitation
Interaction
μ
γ
M
Vacuum
Energies
KE
ISBN 978-0-987288-4-8
[Second Edition © 2012]
Abraham

Tetryonics 80.24 - Quantum Cosmology

TETRYONICS

The equilateral geometry underpinning the mathematics of Physics

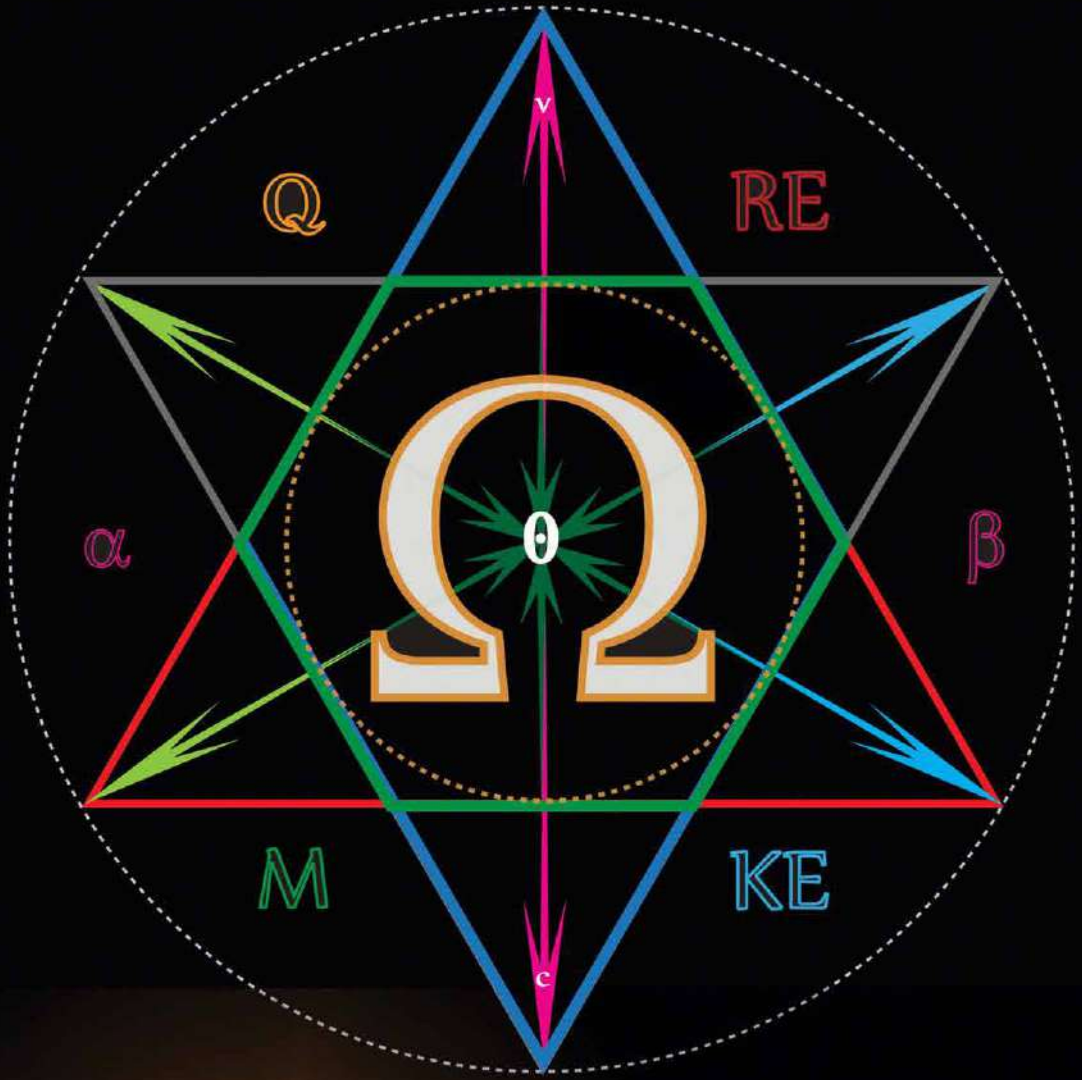

Foundational physics Mathematics

Abraham

ISBN 978-0-987288-5-5

[Second Edition © 2012]

Copyright ABRAHAM [2008] - All rights reserved

Copyright ABRAHAM [2008] - All rights reserved

Geometry and the Theory of Everything

Plato

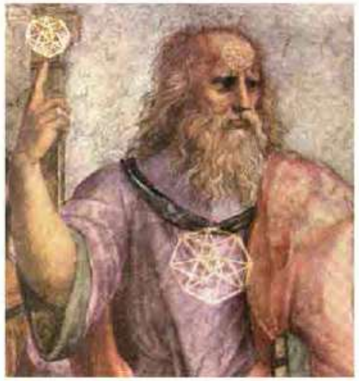

(c.428-348 BC)

DIFFERENT TYPE OF GEOMETRIES

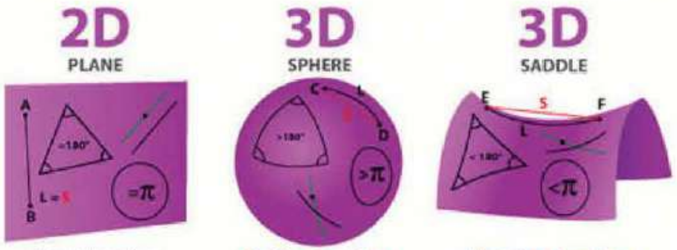

Zero Curvature
Euclidion geometry

Positive Curvature
Elliptic geometry

Negative Curvature
Hyperbolic geometry

(studied by Omar Khayyam, Girolamo Saccheri, Bernhard Riemann, ...)

Euclid

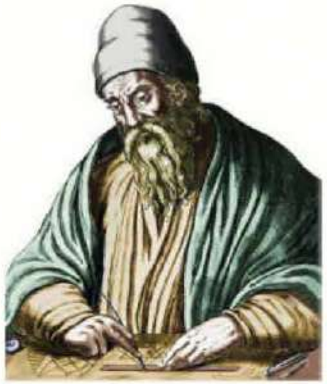

(c.330-275 BC, fl. c.300 BC)

The Socratic tradition was not particularly congenial to mathematics, as may be gathered from Socrates' inability to convince himself that 1 plus 1 equals 2, but it seems that his student Plato gained an appreciation for mathematics after a series of conversations with his friend Archytas in 388 BC.

One of the things that most caught Plato's imagination was the existence and uniqueness of what are now called the five "Platonic solids".

It's uncertain who first described all five of these shapes - it may have been the early Pythagoreans - but some sources (including Euclid) indicate that Theaetetus (another friend of Plato's) wrote the first complete account of the five regular solids.

Presumably this formed the basis of the constructions of the Platonic solids that constitute the concluding Book XIII of Euclid's Elements.

In any case, Plato was mightily impressed by these five definite shapes that constitute the only perfectly symmetrical arrangements of a set of (non-planar) points in space, and late in life he expounded a complete "theory of everything", in the treatise called Timaeus, based explicitly on these five solids.

Interestingly, almost 2000 years later, Johannes Kepler was similarly fascinated by these five shapes, and developed his own cosmology from them

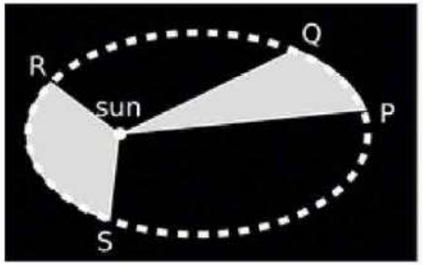

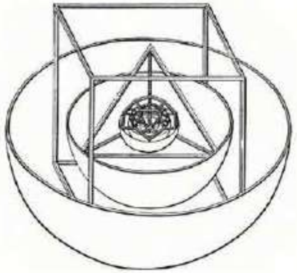

Tetractys

The Tetractys historically symbolized the four elements
[Earth, Air, Fire, and Water]
and the relationship between Humanity
and the cosmos created by GOD

The Greek Tetractys is a
triangular figure consisting of ten points
arranged in four rows:

one, two, three, and four points in each row,
which is the geometrical representation of the
fourth triangular number.

As a mystical symbol, it was very important to
the secret worship of the Pythagoreans.

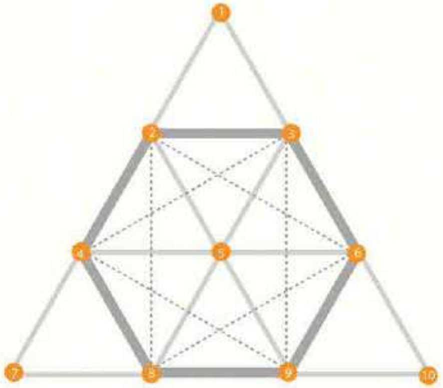

Sacred numbers

A

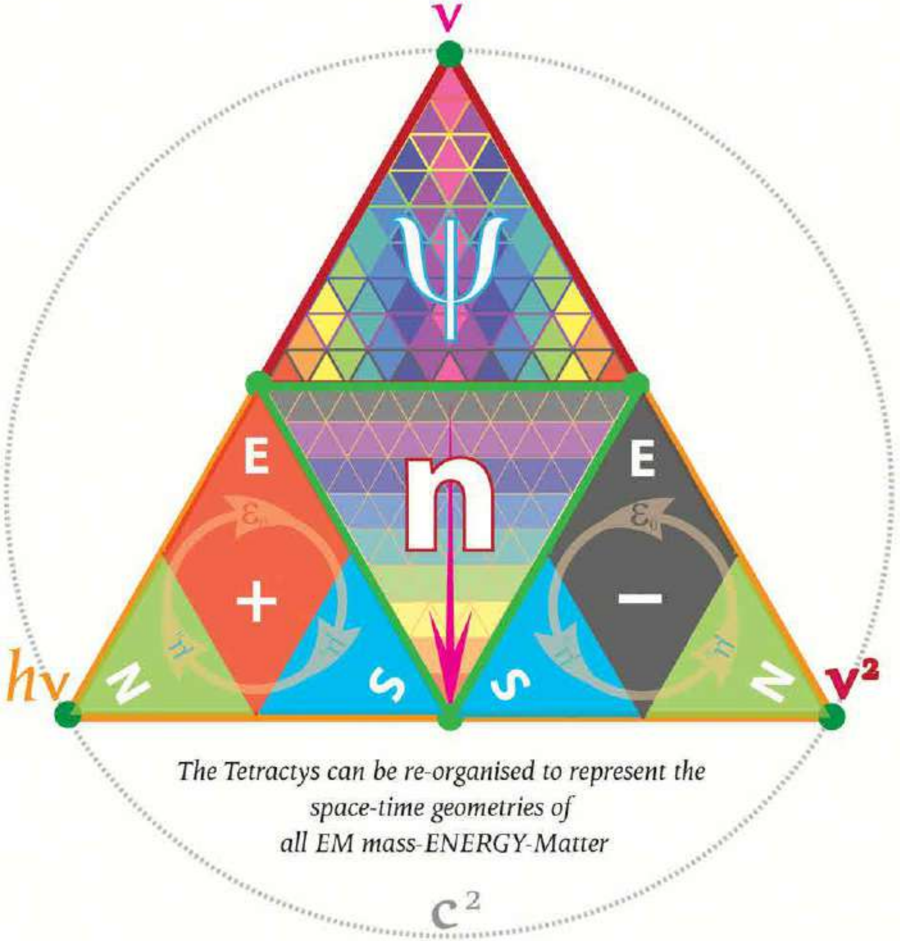

*The Tetractys can be re-organised to represent the
space-time geometries of
all EM mass-ENERGY-Matter*

The tetrad was the name given to
the number four -
in Pythagorean philosophy
there were four seasons
and four elements,
and the number was also associated
with planetary motions and music

As a mystical symbol, it was very important to
the secret worship of the Pythagoreans.

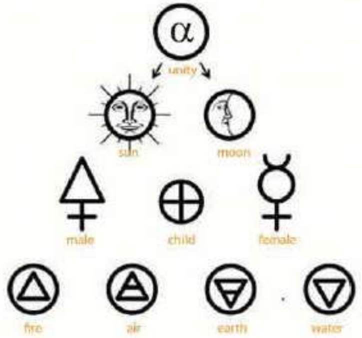

The Cosmos

Ω

The single triangle in the first row represents zero-dimensions (a point)
A vector direction in one-dimension can be represented as a line between any two points
The second row represents a Boson (two-dimensions in a plane defined by a rhombus of three triangles)
The whole figure folded represents three-dimensions (a tetrahedron defined by four apex points)
Photons of ElectroMagnetic mass-Energy quanta are represented by two opposing triangles

The Greek Zodiac

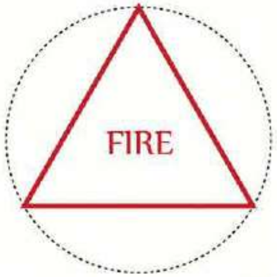

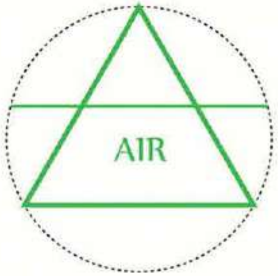

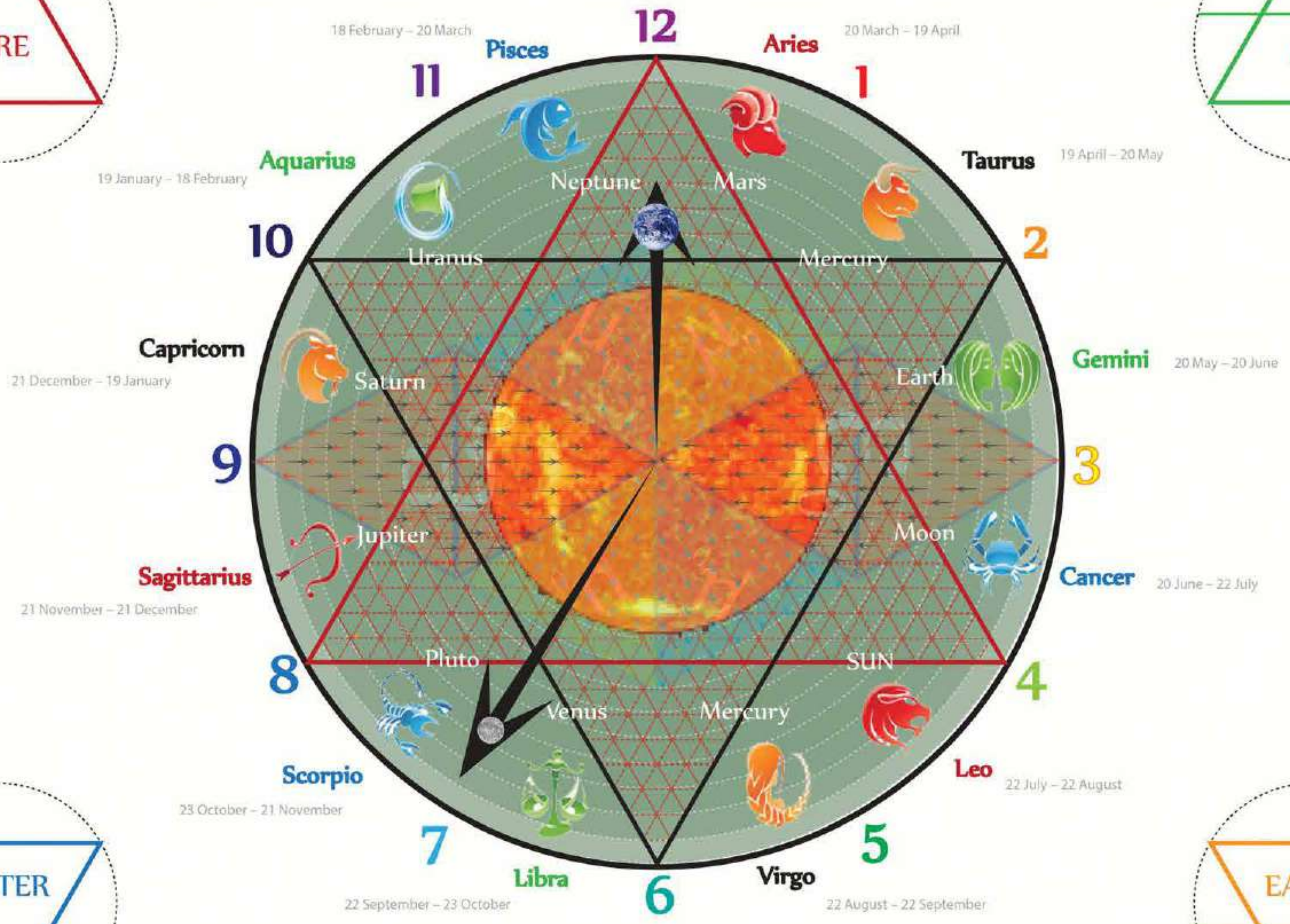
12
Aries
1
20 March – 19 April
Taurus
19 April – 20 May
2
Gemini
20 May – 20 June
3
Cancer
20 June – 22 July
4
Leo
22 July – 22 August
5
Virgo
22 August – 22 September
6
Libra
22 September – 23 October
7
Scorpio
23 October – 21 November
8
Sagittarius
21 November – 21 December
9
Capricorn
21 December – 19 January
10
Aquarius
19 January – 18 February
11
Pisces
18 February – 20 March
Neptune
Mars
Uranus
Mercury
Saturn
Earth
Jupiter
Moon
Pluto
SUN
Venus
Mercury

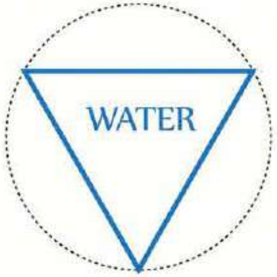

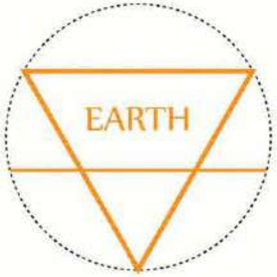

The Greek Elements

Copyright ABRAHAM [2008] - All rights reserved

Equilateral Triangles

An equilateral triangle is a triangle in which all three sides are equal

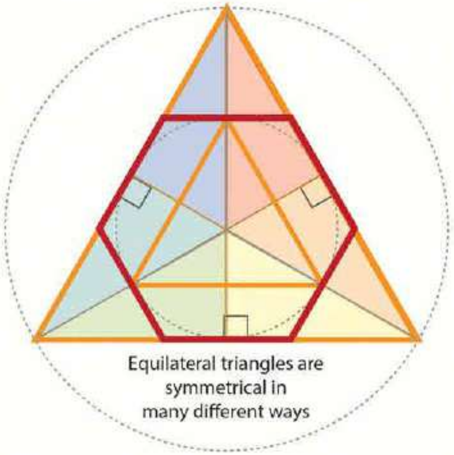

Equilateral triangles are symmetrical in many different ways

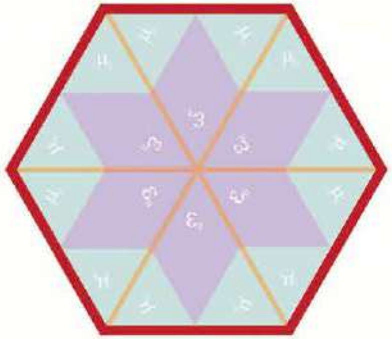

Any six equilateral triangles joined can make a hexagon.

It is unique in that it is the only polygon that can be tiled [or divided] and produce only identical geometries and squares numbers

The equilateral triangle is eminently suited for the construction of fractals

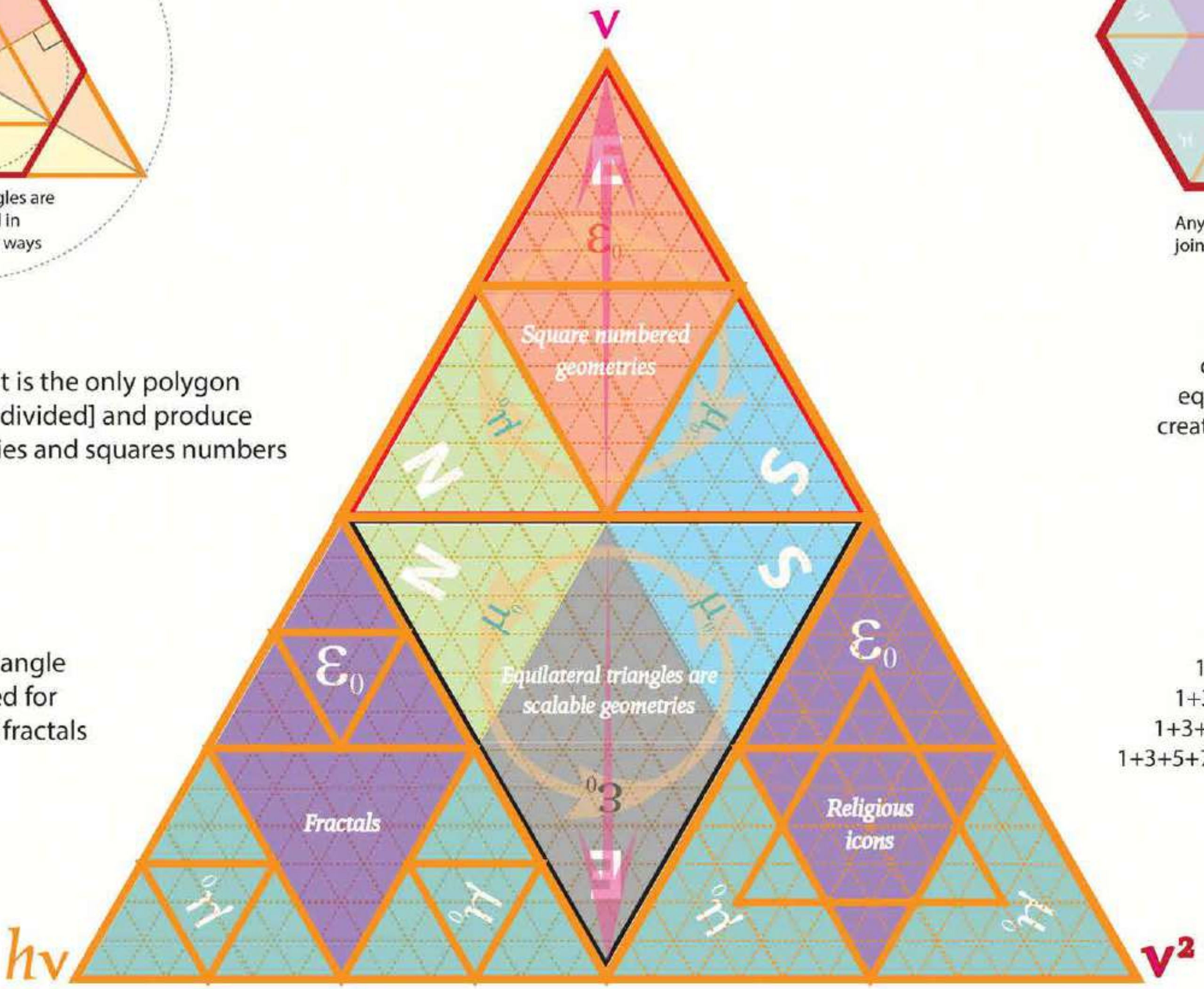
v
E
ε0
Square numbered geometries
μ0
N
S
N
S
μ0
μ0
ε0
ε0
Equilateral triangles are scalable geometries
ε0
E
Fractals
Religious icons
μ0
μ0
μ0
μ0
hv
v²

The
tesselation of
odd numbered
equilateral triangles
creates square numbers

That is,

1
1+3 = 4
1+3+5 = 9
1+3+5+7 = 16
1+3+5+7+9 = 25
1+3+5+7+9+11 = 36
1+3+5+7+9+11+13 = 49
1+3+5+7+7+9+11+13+15 = 64
etc

An equilateral triangle is simply a specific case of a regular polygon with 3 sides

The Pythagorean Theorem

Though attributed to Pythagoras, it is not certain that he was the first person to prove it.

The first clear proof came from Euclid, and it is possible the concept was known
1000 years before Pythoragas by the Babylonians

$$a^2 + b^2 = c^2$$

The square of the hypotenuse of a triangle is equal to the sum of the squares of its sides.

In any right triangle,
the area of the square whose side is the hypotenuse
(the side opposite the right angle) is equal to
the sum of the areas of the squares
whose sides are the two legs
(the two sides that meet at a right angle)

Pythagoras of Samos

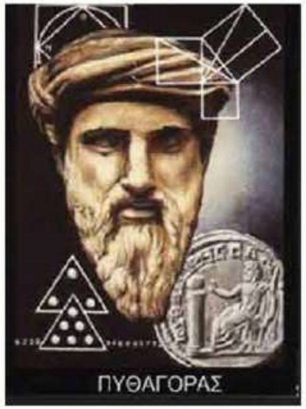

about (570 – 495 BC)

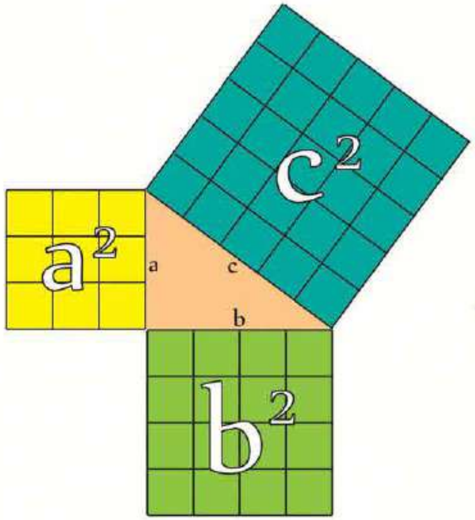

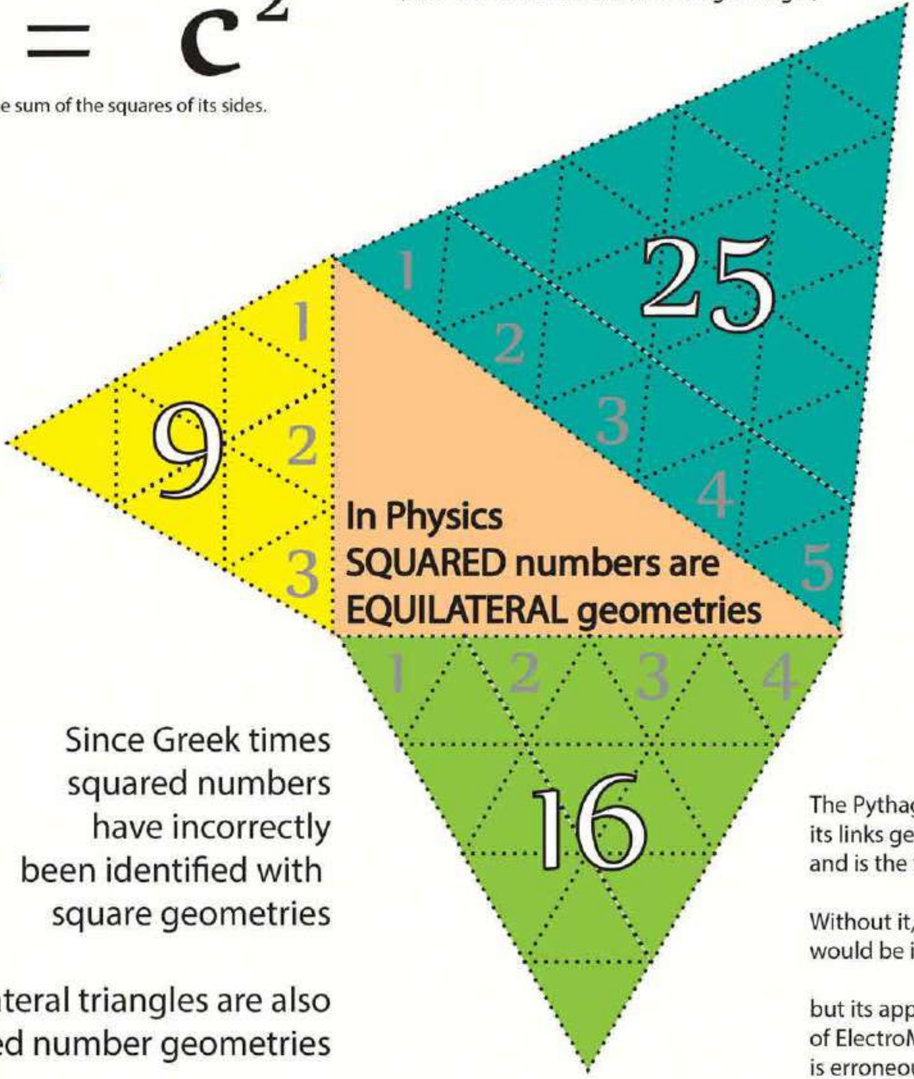

In Physics
SQUARED numbers are
EQUILATERAL geometries

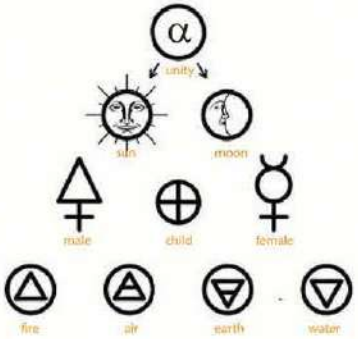

Pythagorean Tetractys

Since Greek times
squared numbers
have incorrectly
been identified with
square geometries

Equilateral triangles are also
squared number geometries

The Pythagorean equation is at the core of much of geometry,
its links geometry with algebra,
and is the foundation of trigonometry.

Without it, accurate surveying, mapmaking, and navigation
would be impossible,

but its application to the energy-momenta geometries
of ElectroMagnetic fields and Matter in motion in Physics
is erroneous and must be corrected for science to advance

Hexagons

A regular hexagon can be subdivided into six equilateral triangles

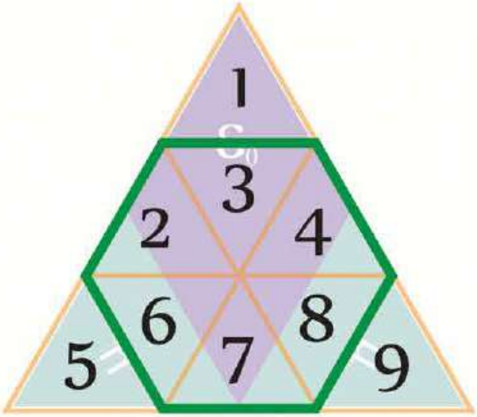

Energy geometries

Atomic nuclei geometries

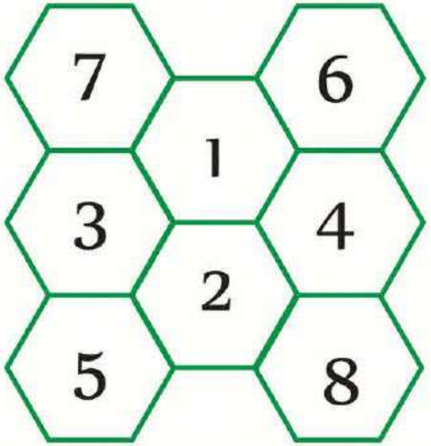

Hexagons can be tiled or tessellated in a regular pattern on a flat two-dimensional plane

Hexagons are the only regular polygon that can be subdivided into another regular polygon

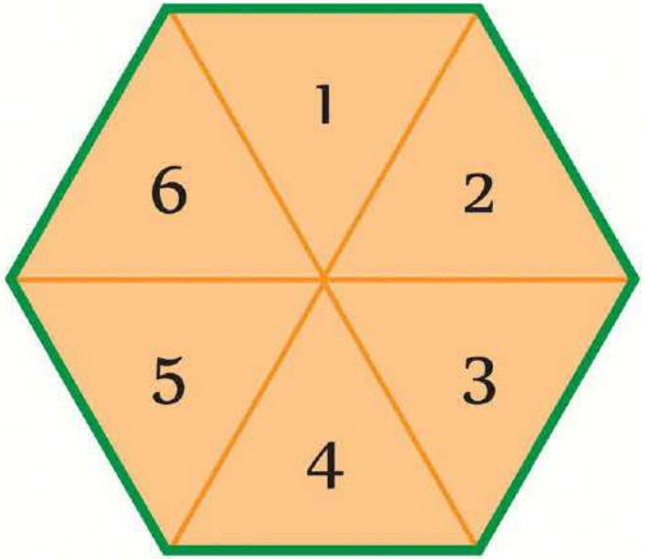

Hexagons are the unique regular polygon such that the distance between the center and each vertex is equal to the length of each side

Six is a highly composite number,
the second-smallest composite number,
and the first perfect number.

That is, 1*2*3=1+2+3=6

An interesting relationship between circular and hexagonal geometry is that hexagonal patterns often appear spontaneously when natural forces are trying to approximate circles

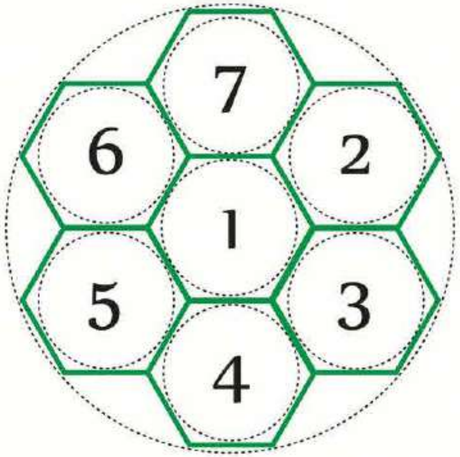

Hexagonal tessellation is topologically identical to the close packing of circles on a plane

Copyright ABRAHAM [2008] - All rights reserved

Platonic Solids

The Pythagoreans knew that there were only five regular convex solids,
the tetrahedon, cube, octahedron, icosohedron and dodecahedron
and each one could be accurately circumscribed by a sphere.

FIRE

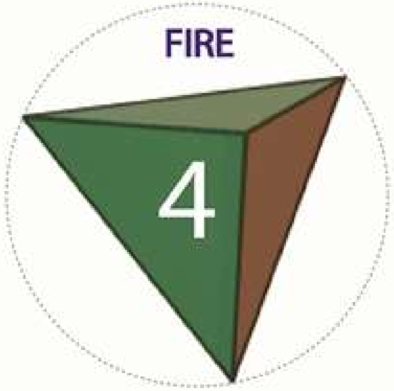

4 triangles meet to form
a Tetrahedron

EARTH

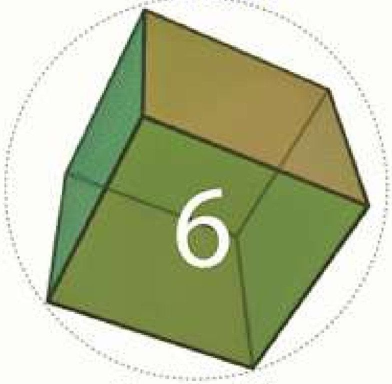

6 squares meet to form
a Cube

AIR

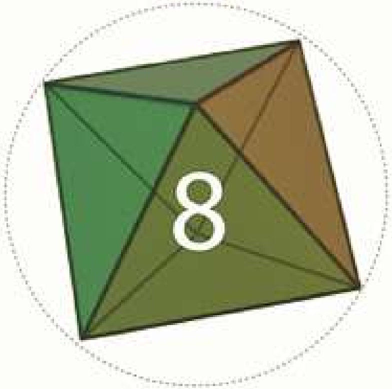

8 triangles meet to form
a Octahedron

tetryons **4**

regular deltahedrons

leptons **12**

F - E + V = 2

faces edges vertexs

8 **quarks**

regular deltahedrons

20 **Baryons**

12 pentagons meet to form
a Dodecahedron

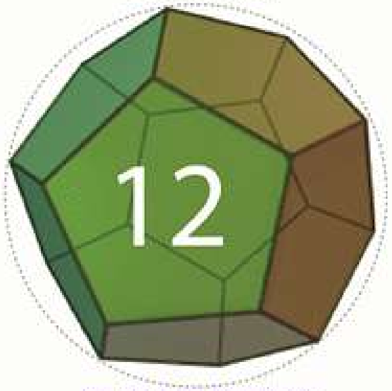

THE HEAVENS

Plato

(c.428-348 BC)

The philosopher Plato concluded that they must be the fundamental building blocks
– the atoms – of nature, and assigned to them what he believed to be
the essential elements of the universe.

20 triangles meet to form
a Icosahedron

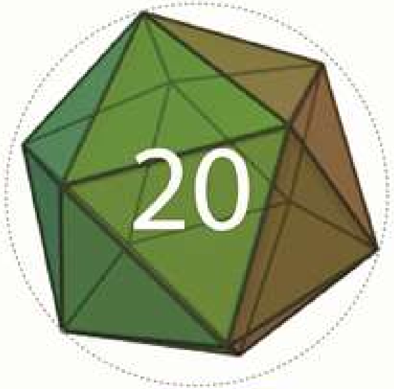

WATER

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic Solids

Despite their unique topologies Tetryonic solids are not unlike Platonic solids save that their toplogies are comprised entirely from complex hitherto undescribed **4nπ equilateral Planck mass-energy momenta geometries** that also match the Euler numbers of Platonic solids

+ tetryon

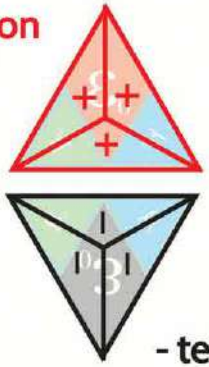

- tetryon

4 faces
6 edges
4 vertices

2

up quark

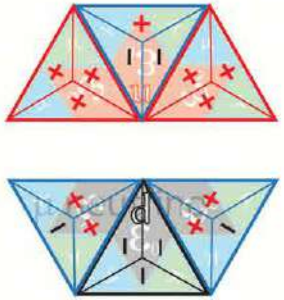

down quark

8 faces
12 edges
6 vertices

2

mass geometries

4π ***tetryons*** **4π**

regular
deltahedrons

12π ***leptons*** **12π**

Matter topologies

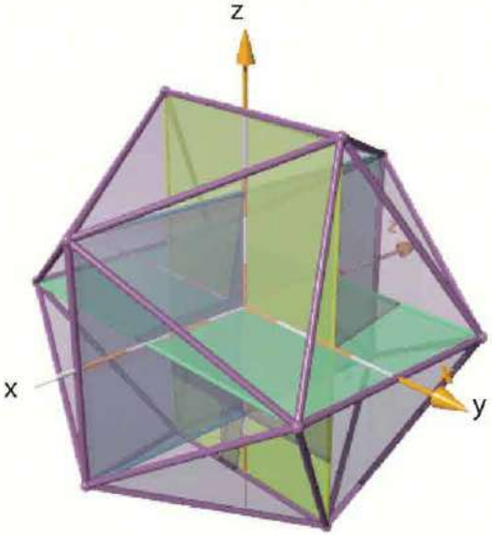

Matter topologies

8π ***quarks*** **12π**

regular
deltahedrons

20π ***Baryons*** **36π**

mass geometries

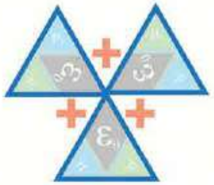

neutrino

electron

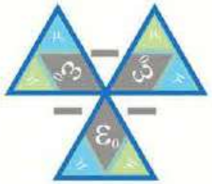

12 faces
18 edges
8 vertices

2

Their equilateral topologies are best described as regular topologic-deltahedrons:

tetra-delta-hedrals	**tetryons**	4π external charge fascia
octa-delta-hedrals	**quarks**	8π external charge fascia
dodeca-delta-hedrals	**leptons**	12π external charge fascia
icoso-delta-hedrals	**Baryons**	20π external charge fascia

note:
Charged mass-energy fascia geometries and edges become "hidden" upon the meshing of delta-hedra to form Matter topologies

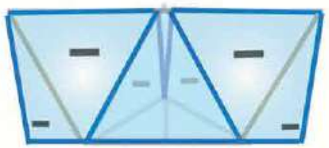

Neutron

Proton

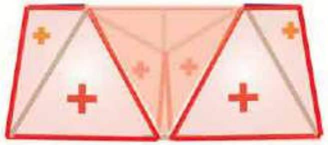

2

20 faces
30 edges
12 vertices

Copyright ABRAHAM [2008] - All rights reserved

Euclidean geometry

Euclid

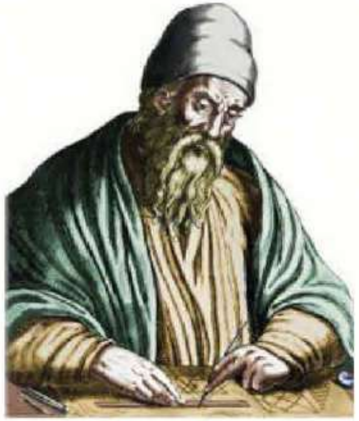

(c.330-275 BC, fl. c.300 BC)

Arguably the most influential Mathematics book ever written is Euclid's 'The Elements'

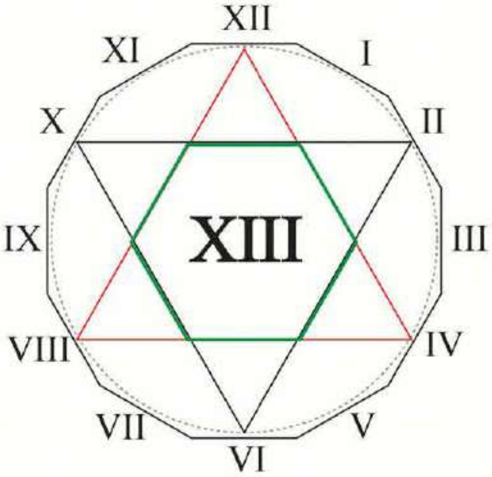

In all, it contains 465 theorems and proofs, described in a clear, logical and elegant style, and using only a compass and a straight edge.

The Elements - Book 1 - Definition 20

Of the trilateral figures, an equilateral triangle is that which has its three EQUAL sides

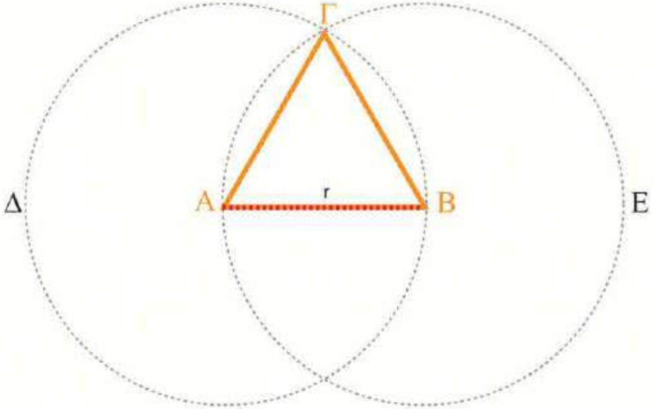

Euclid's Elements - Book 1 - Proposition 1
Method of constructing an Equialteral triangle

Euclid's five general axioms were:

Things which are equal to the same thing are equal to each other.

If equals are added to equals, the wholes (sums) are equal.

If equals are subtracted from equals, the remainders (differences) are equal.

Things that coincide with one another are equal to one another.

The whole is greater than the part

Copyright ABRAHAM [2008] - All rights reserved

Archimedes

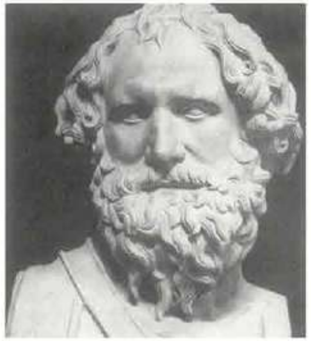

c. (287 BC – c. 212 BC)

As its definition relates to the circle,
π is found in many formulae in trigonometry and geometry,
especially those concerning circles, ellipses, or spheres.

It is also found in formulae from other branches of science,
such as cosmology, number theory, statistics, fractals,
thermodynamics, mechanics, and electromagnetism

Incorrect identification of Pi [c/d] as opposed to Pi radians
in Physics has led to the inappropriate association
of spherical particles to the physical sciences
whereas equilateral triangles & tetrahedra
form its true geometry

The number π is a mathematical constant that is
the ratio of a circle's circumference to its diameter.

Pi

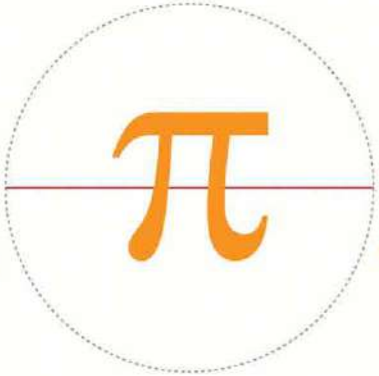

π is an irrational number
its decimal representation
never ends and never repeats.

$\pi = c/d$

The ratio C/d is constant,
regardless of the circle's size

3.141592654.......

Proof of the fact that C=2πr and how Archimedes proved it

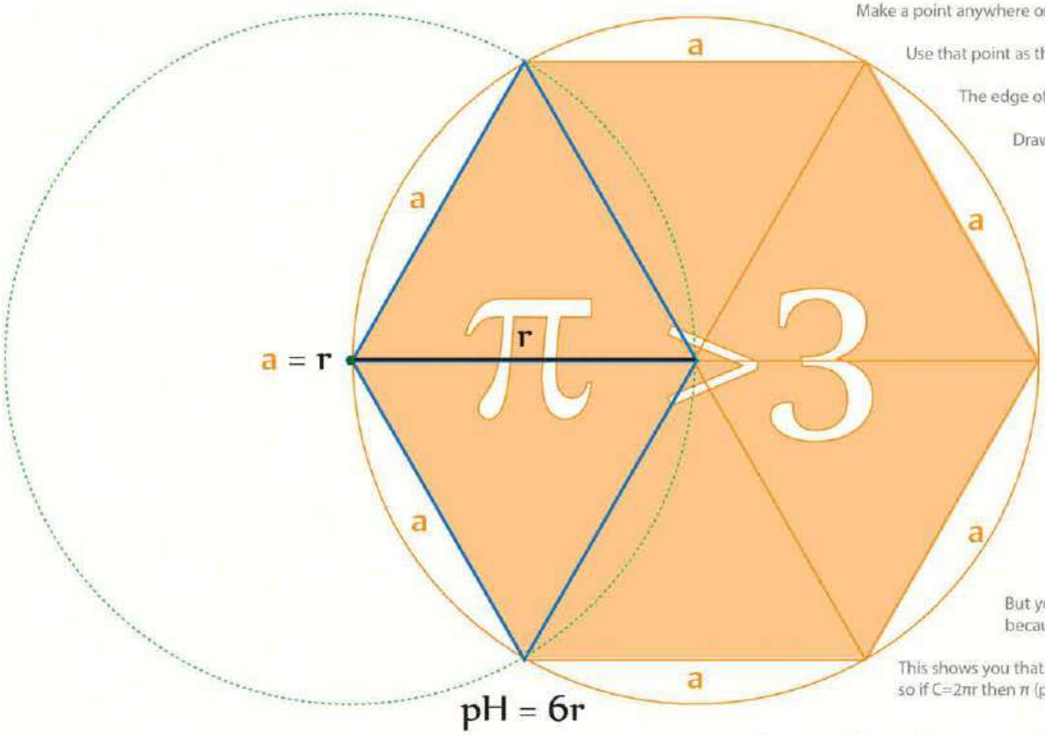

Draw any circle.

Make a point anywhere on the circumference of the [green] circle.

Use that point as the center of a second [blue] circle with the same radius as the green circle.

The edge of the blue circle should touch the center of the green circle.

Draw the line segment connecting the centers of the two circles.

That forms the radius of both of the circles.

Now draw the line connecting the center of the blue circle to where it crosses the green circle on both sides, and complete the triangles.

You should have two equilateral triangles whose sides are equal to the radius of the green circle

Now extend all of the radius lines so they become diameter lines, all the way across the circle, and finish drawing all of the triangles to connect them.

You've got six equilateral triangles now, that make an orange hexagon.

So the perimeter of your hexagon is the same as six times the radius of your circle.

But your circumference is a little bigger than the perimeter of your hexagon, because the shortest distance between two points is always a straight line.

This shows you that the circumference of the blue circle has to be more than 6r, so if C=2πr then π (pi) has to be a little bigger than 3, which it is.

The more sides we draw on our polygon, the closer we will get to the real value of pi (3.14159 etc.).
Using a polygon with 96 sides, Archimedes was able to calculate that π was a little bigger than 3.1408

The Golden Ratio

Two quantities are in the golden ratio (φ) if the ratio of the sum of the quantities to the larger quantity is equal to the ratio of the larger quantity to the smaller one.

The figure to the right illustrates the geometric relationship

[1.61803]

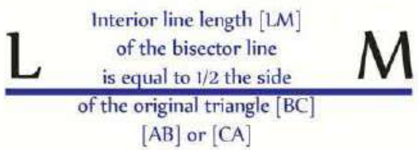

Interior line length [LM] of the bisector line is equal to 1/2 the side of the original triangle [BC] [AB] or [CA]

Φ

Golden ratio phi

$$\hbar = \frac{h}{2\pi}$$

Tetryonic geometry reveals the maximum E-field amplitude of the reduced Planck constant to be an example of the Golden ratio in physics

velocity vectors

The height of the triangle [ALM] produced by the bisecting line is 1/2 the height of the height of [ABC]

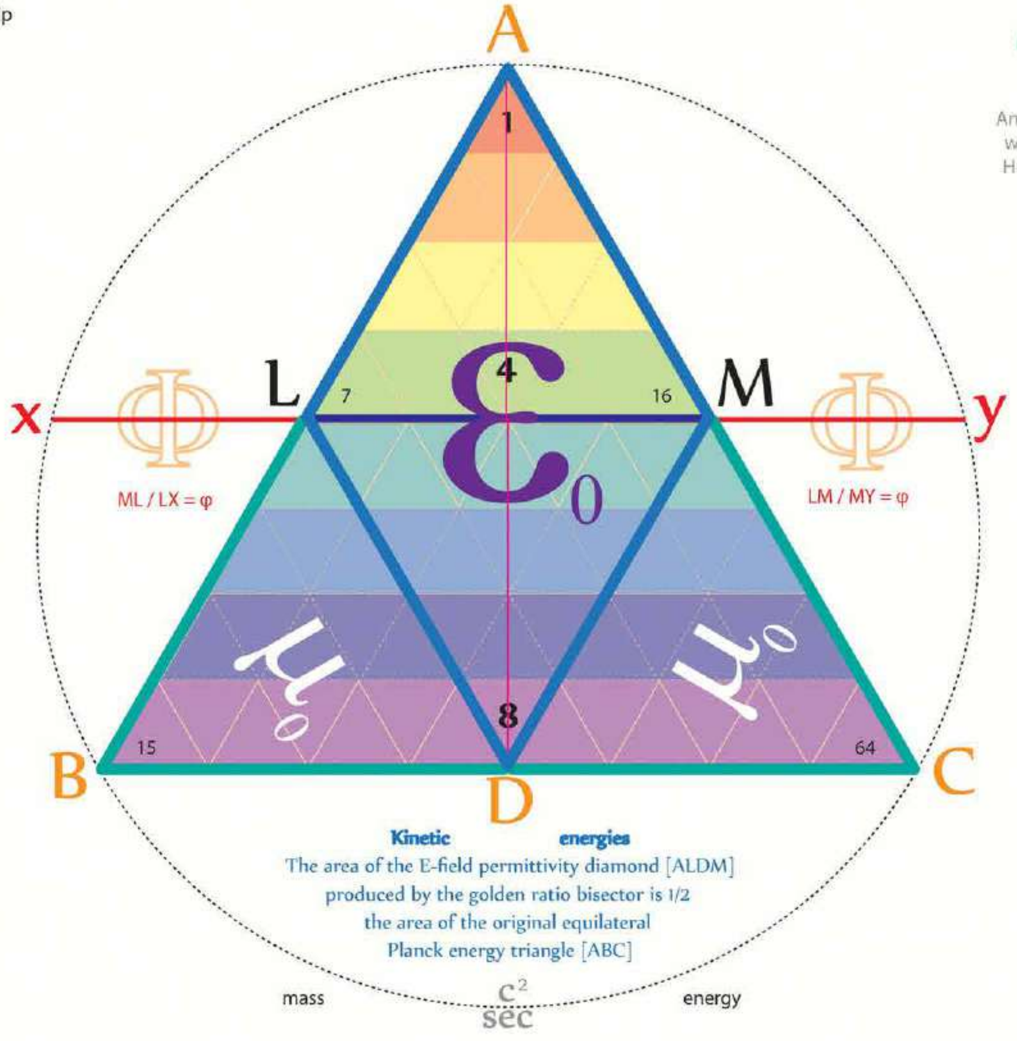

Kinetic energies

The area of the E-field permittivity diamond [ALDM] produced by the golden ratio bisector is 1/2 the area of the original equilateral Planck energy triangle [ABC]

Planck's constant [h]

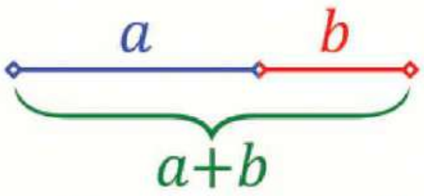

$a+b$ is to a as a is to b

An intriguing showing of φ in an equilateral triangle was observed by George Odom, a resident of the Hudson River Psychiatric Center, in the early 1980s

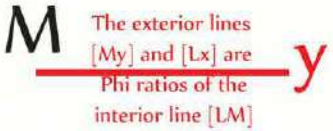

The exterior lines [My] and [Lx] are Phi ratios of the interior line [LM]

Let L and M be the midpoints of the sides AB and AC of an equilateral triangle ABC

By measurement and the Intersecting Chords Theorem

MX·MY = AM·MC.

which is of the form

(a + b)·b = a·a.

Denoting a/b = x, we see that

$1 + x = x^2$,

George Odom

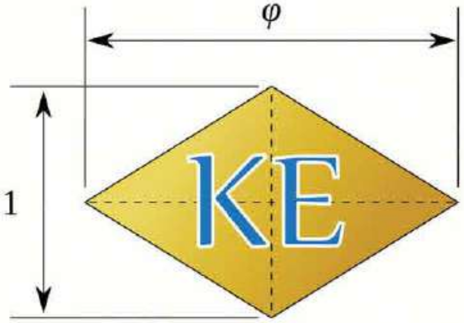

The Golden Rhombus

Applying the golden ratio (φ) to quantum scale electrodynamic geometry we can quickly determine that the linear momentum and magnetic moment vectors of photons & EM waves can also be expressed as a golden ratio

vector velocities

vs

magnetic vector

Φ

Golden ratio phi

[1.61803]

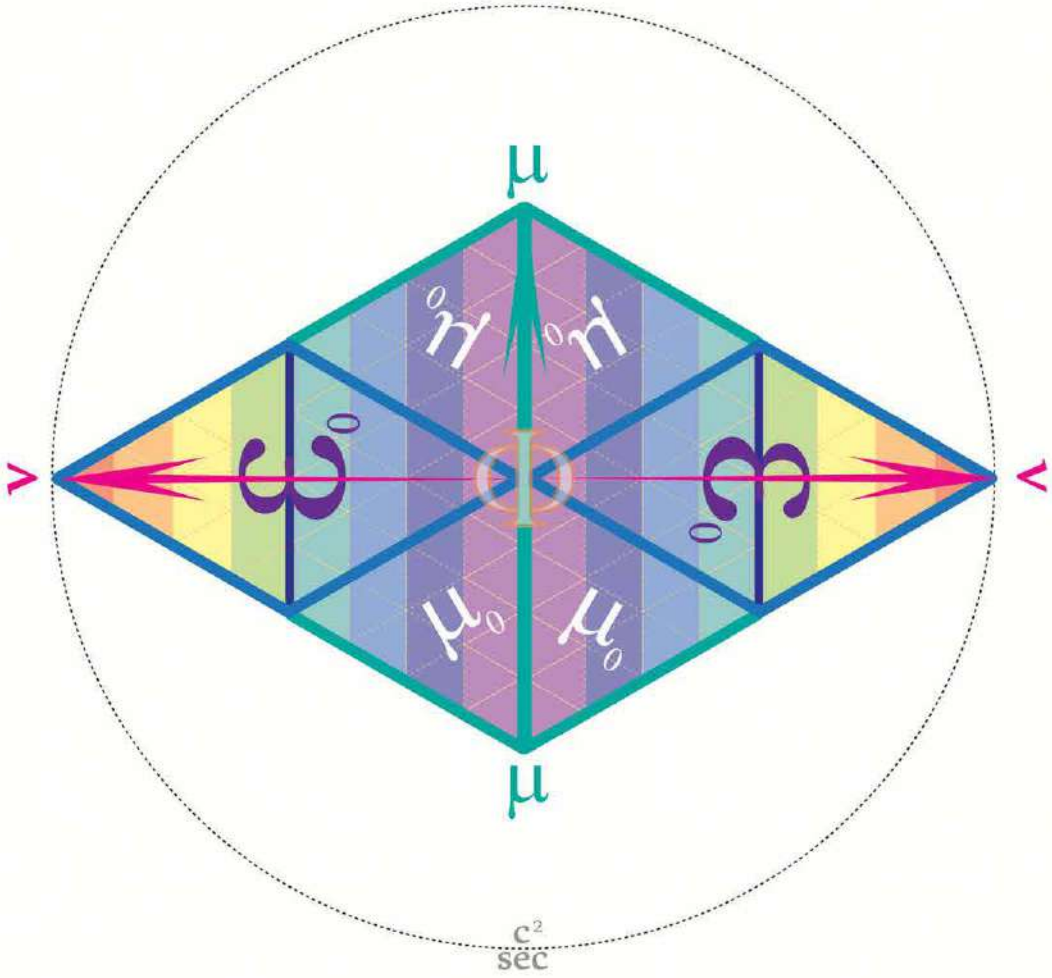

magnetic Force

vs

electric Force

photons of EM energy

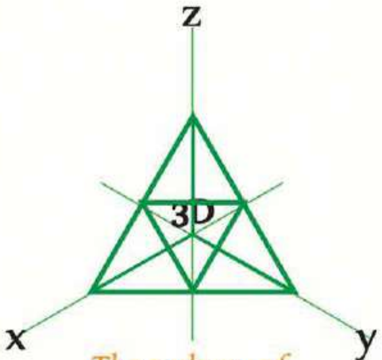

The 3 planes of Cartesian co-ordinates

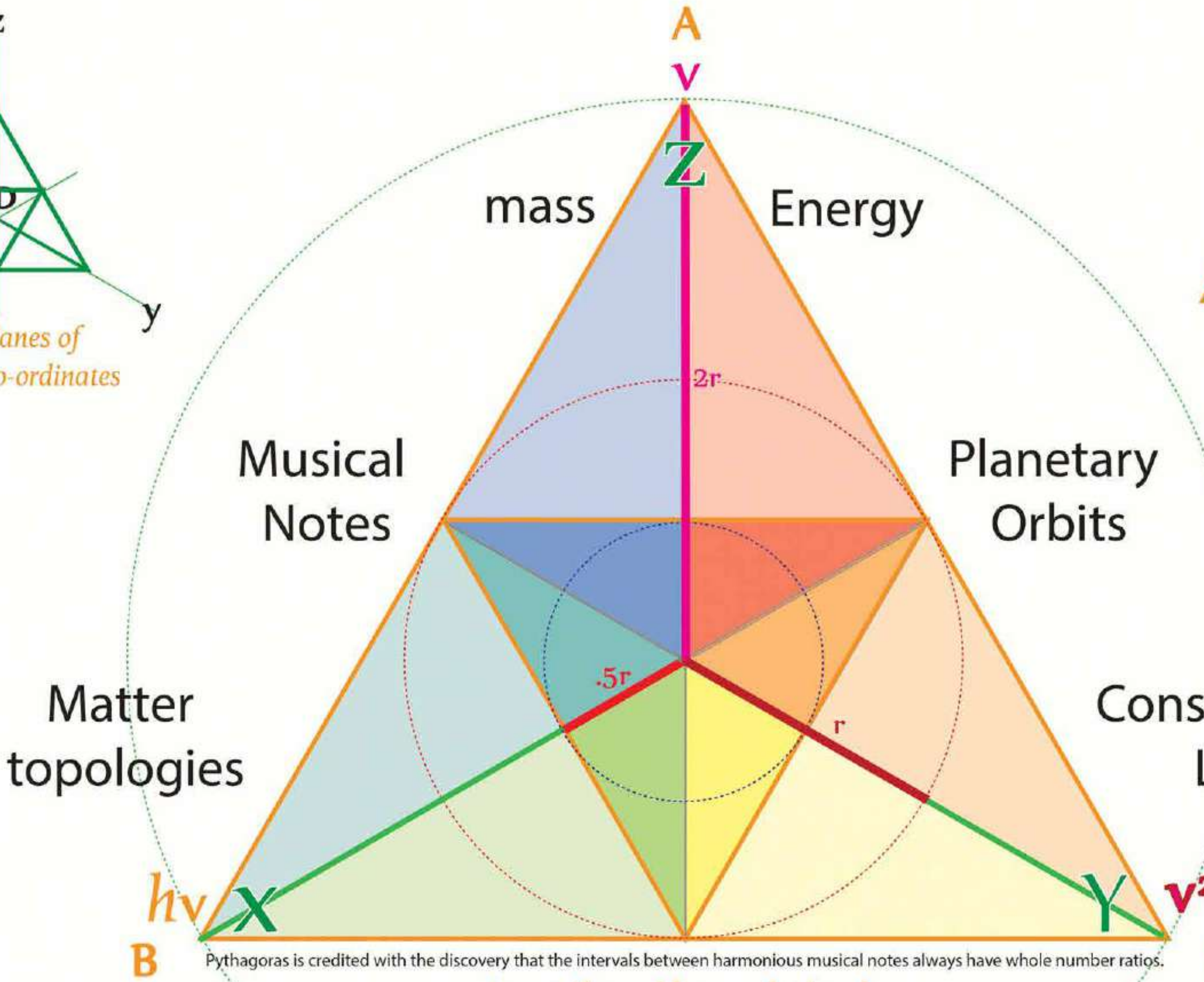
A
v
Z
mass
Energy
2r
Musical Notes
Planetary Orbits
.5r
r
Matter topologies
Conservation Laws
hv
X
Y
v²
B
C

Pythagoras is credited with the discovery that the intervals between harmonious musical notes always have whole number ratios.

An equilateral [square] triangle divided into 6 equal semitones and 6 equal quarter tones

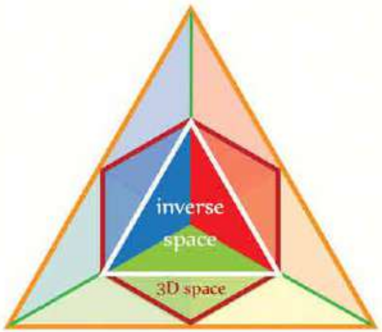
inverse
space
3D space

An equilateral [square] triangle divided into 12 equal semitones

With an Equilateral triangle, draw lines from the center of the circle to each vertex and each midpoint, creating six right triangles, as shown with six different colored triangles.
Each right triangle has a radius of the circle for one leg, and half of a side of the original triangle for another.
Any two right triangles sharing one of the radial lines must therefore be congruent, and that implies that the hypotenuses of the triangles are all equal.
This in turn shows that the six right triangles are all congruent, and so the large triangle they combine to form is equilateral

Equilateral Fifths

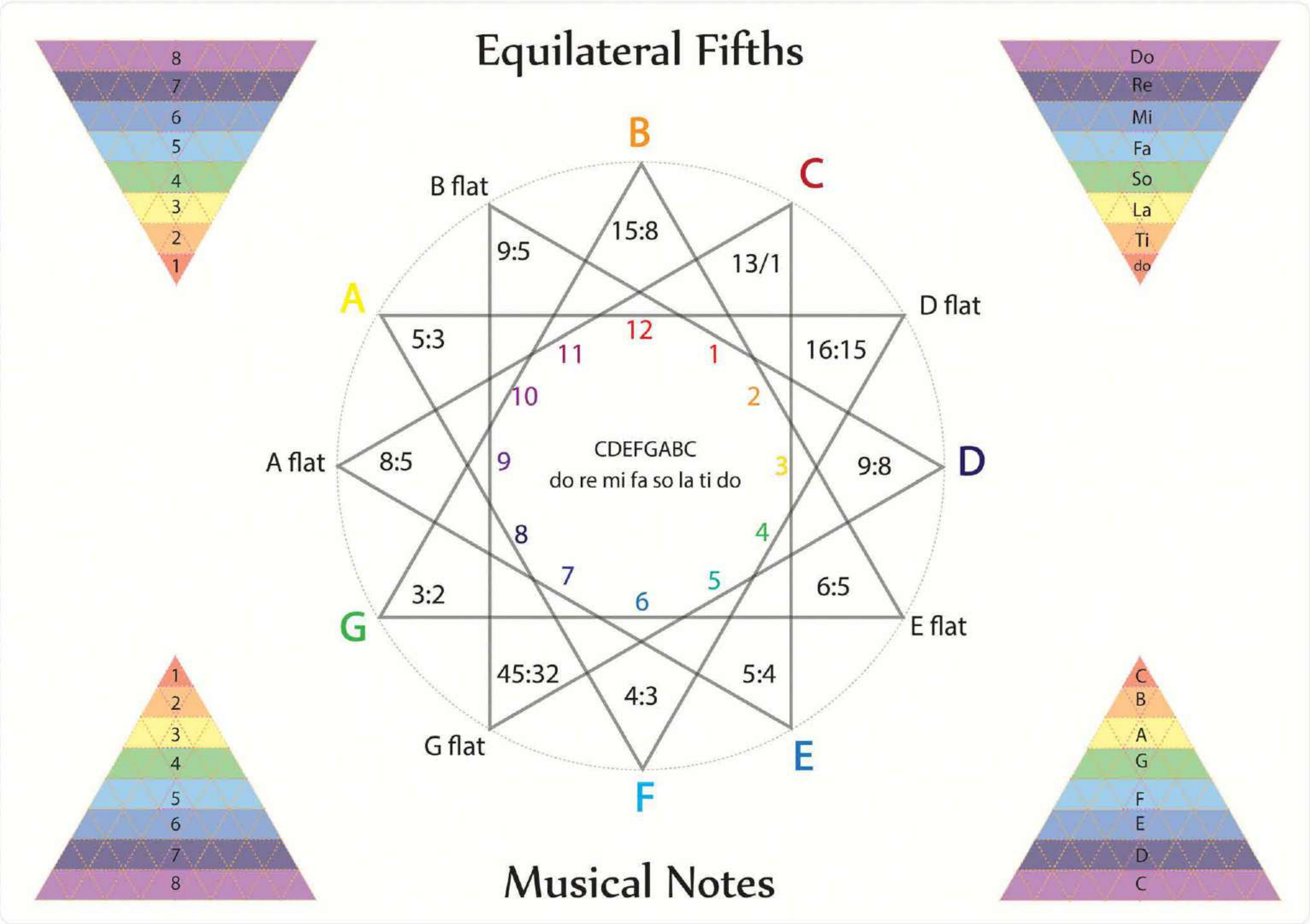
8
7
6
5
4
3
2
1
Do
Re
Mi
Fa
So
La
Ti
do
B
C
B flat
D flat
A
D
A flat
G
E flat
G flat
E
F
15:8
13/1
9:5
5:3
16:15
8:5
9:8
3:2
6:5
45:32
4:3
5:4
12
1
2
3
4
5
6
7
8
9
10
11
CDEFGABC
do re mi fa so la ti do
1
2
3
4
5
6
7
8
C
B
A
G
F
E
D
C

Musical Notes

Tetryonics 81.14- Equilateral Fifths

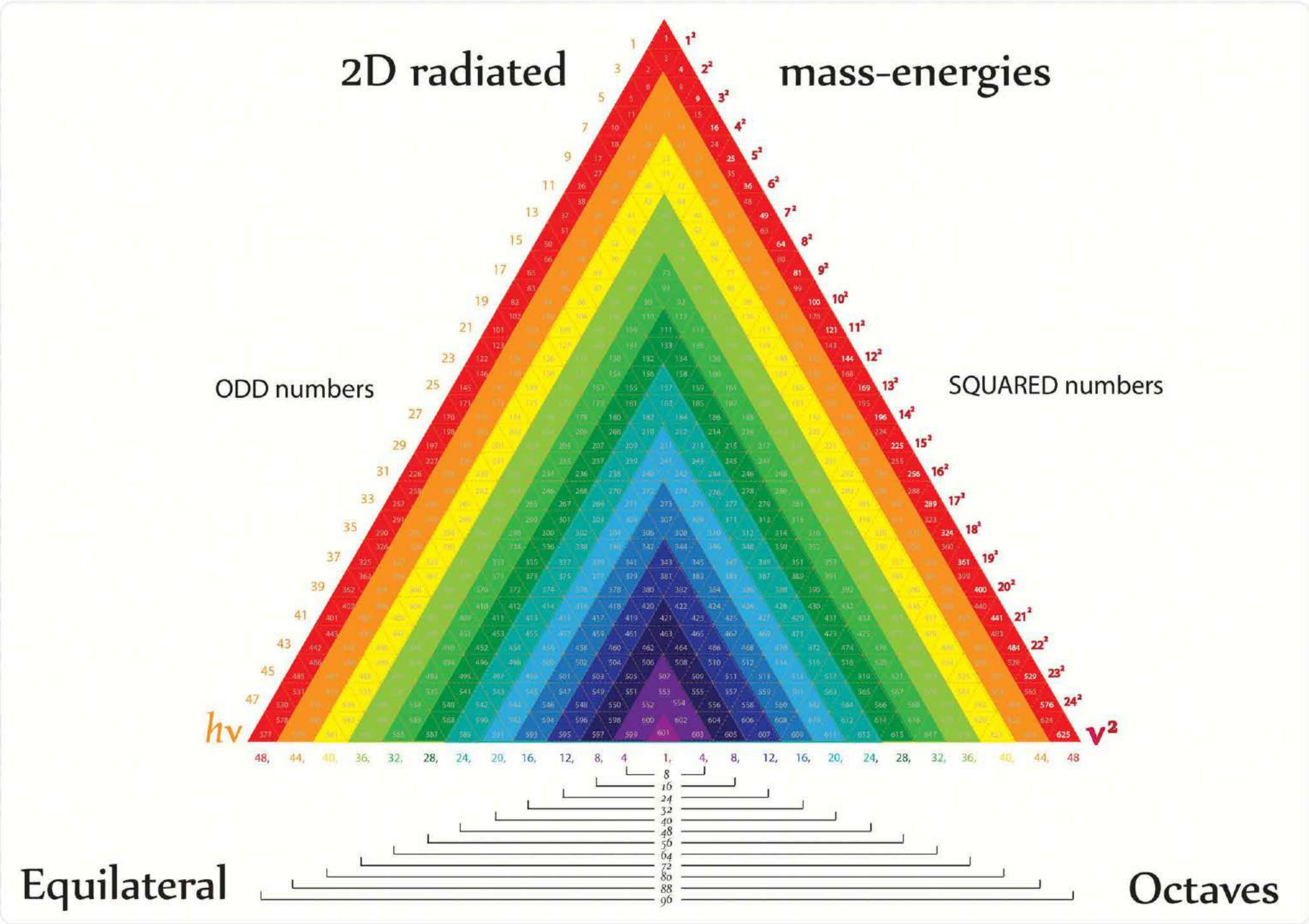
2D radiated
mass-energies
ODD numbers
SQUARED numbers
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
hv
1²
2²
3²
4²
5²
6²
7²
8²
9²
10²
11²
12²
13²
14²
15²
16²
17²
18²
19²
20²
21²
22²
23²
24²
v²
48, 44, 40, 36, 32, 28, 24, 20, 16, 12, 8, 4 1, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48
8
16
24
32
40
48
56
64
72
80
88
96
Equilateral
Octaves

Ternary Diagrams

Viviani's Theorem implies that lines parallel to the sides of an equilateral triangle
provide (homogeneous/barycentric/areal/trilinear) coordinates for ternary diagrams
for representing three quantities A,B,C whose sum is a constant
(which can be normalized to unity).

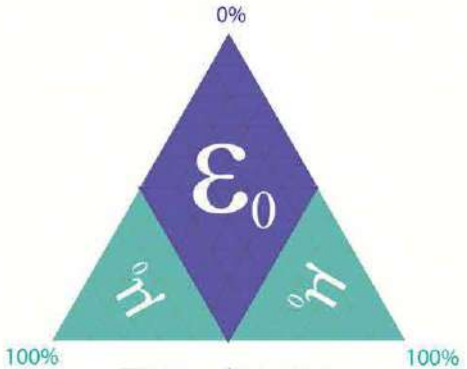
0%
ε_0
μ_0
μ_0
100%
100%

Ternary diagrams
should NOT be used
to model EM field
strengths

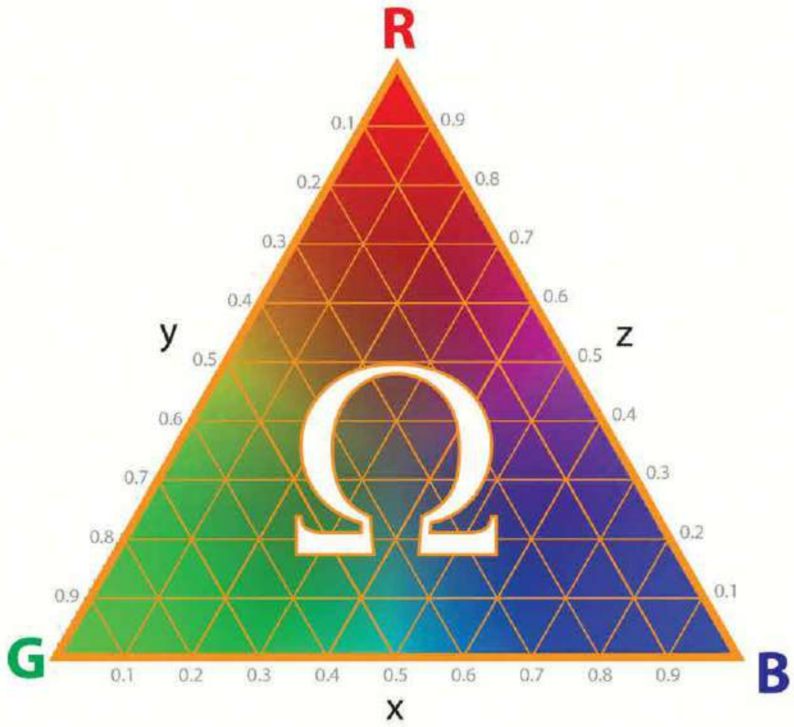
R
G
B
y
z
x
Ω
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

In a ternary plot, of EM energy
the Electric field [E] and the
Magnetic dipole [N-S]
must renormalise
to 1

A ternary diagram is simply a triangular coordinate system in which the 3 edges correspond to the axes.

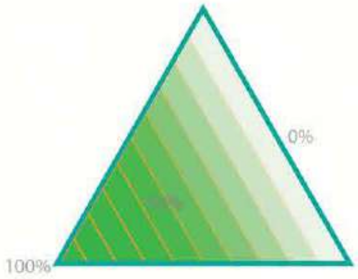

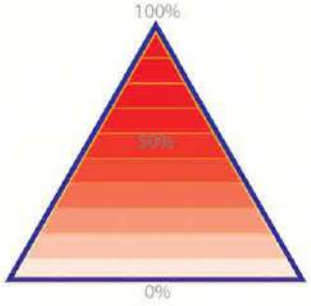

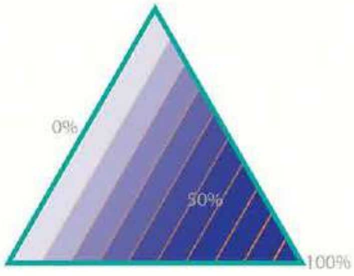

Trilinear charts are commonly used for finding the result of mixing three components (such as gases, chemical compounds, soil, color, etc.) that add to 100% of a quantity.

Whilst the Pythagorean Theorem boasts a slightly greater economy of terms than the Eutrigon Theorem (Wayne Roberts 2003), the latter contains an important area not included in the former:

the area enclosed or swept out by the three points of the triangle in question

$$ab = a^2 + b^2 - c^2$$

the area of any eutrigon is equal to the combined areas of the equilateral triangles on legs 'a' and 'b', minus the area of the equilateral triangle on its hypotenuse 'c'.

The Pythagorean theorem

$$a^2 + b^2 = c^2$$

The square of the hypotenuse of a triangle is equal to the sum of the squares of its sides.

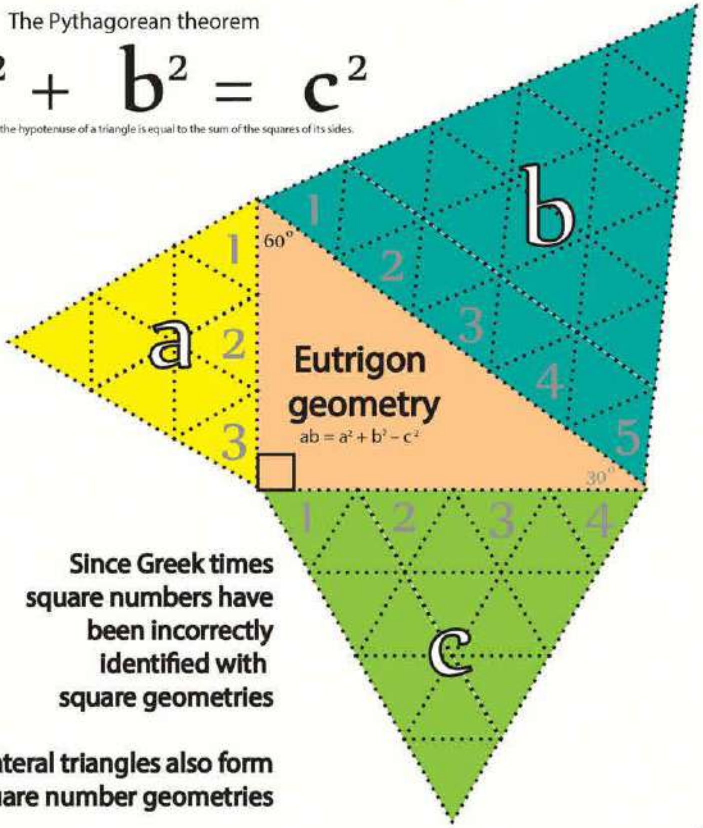

Since Greek times square numbers have been incorrectly identified with square geometries

Equilateral triangles also form square number geometries

Eutrigons

are an important new class of triangle (mathematically defined by Wayne Roberts), as the analogue of the right-triangle in orthogonal (Cartesian) coordinate geometry

Kepler's Second Law of planetary motion

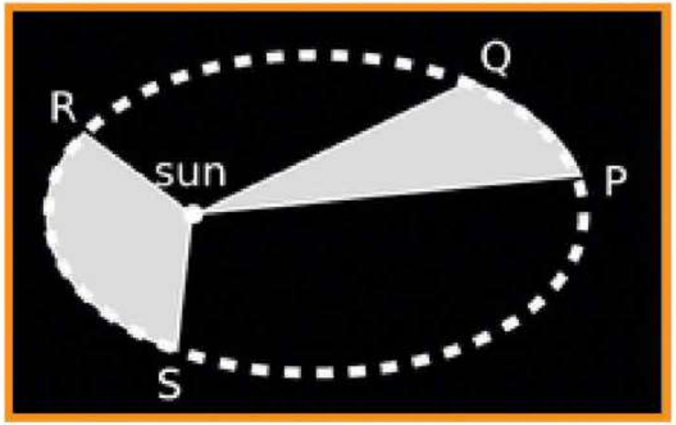

The orbit of every planet is an ellipse with the Sun at one of the two foci.

A line joining a planet and the Sun sweeps out equal areas during equal intervals of time.

The algebraic form of the Eutrigon Theorem, (like the algebraic form of Pythagoras' Theorem), is proven to be special case of the Cosine Rule...

Tetryonic theory reveals the equilateral [square] energy geometry
that reveals the 'harmonics at play' in physical laws
such as the second law of Kepler, and in many other phenomena in
physics, chemistry, cosmology, biochemistry and number theory
thus providing the foundation for the mathematics of quantum mechanics

Copyright ABRAHAM [2008] - All rights reserved

Number theory

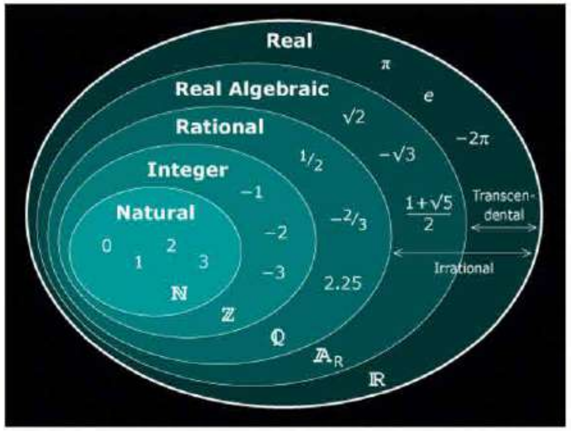
Real
π
Real Algebraic
e
√2
Rational
−2π
1/2
−√3
Integer
−1
1+√5 / 2
Transcen-dental
Natural
−2/3
−2
0
2
1
3
−3
2.25
Irrational
ℕ
ℤ
ℚ
𝔸R
ℝ

The Pythagoreans also established the foundations of number theory, with their investigations of triangular, square and also perfect numbers (numbers that are the sum of their divisors).

They discovered several new properties of square numbers, such as that the square of a number n is equal to the sum of the first n odd numbers (e.g. $1 + 3 + 5 + 7 = 16$).

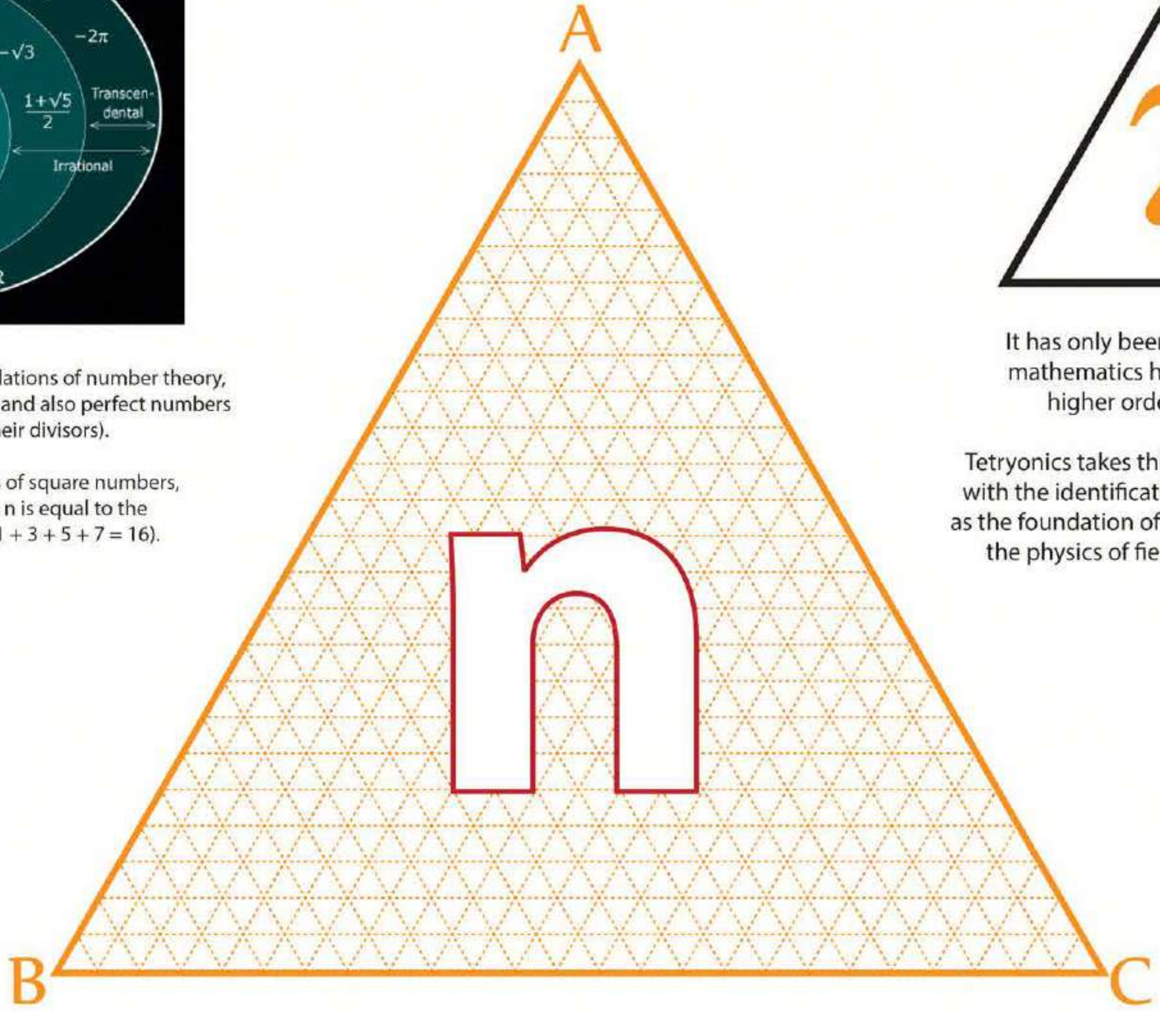

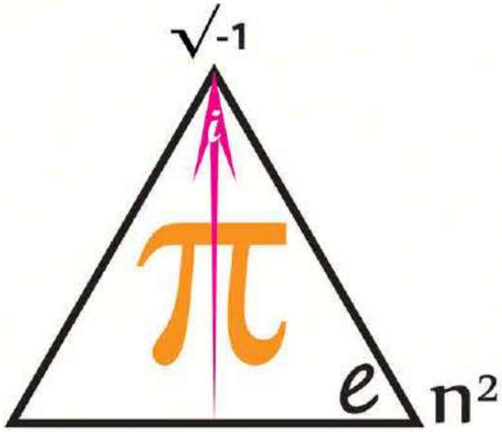

It has only been in recent centuries that mathematics has begun to explore the higher order irrational numbers

Tetryonics takes this investigation to new levels with the identification of equilateral geometries as the foundation of transcendental numbers and the physics of fields and particles in motion

What mathematics has failed to appreciate is the significance application of equilateral geometries to the 'square' numbers of physics and science in general

Copyright ABRAHAM [2008] - All rights reserved

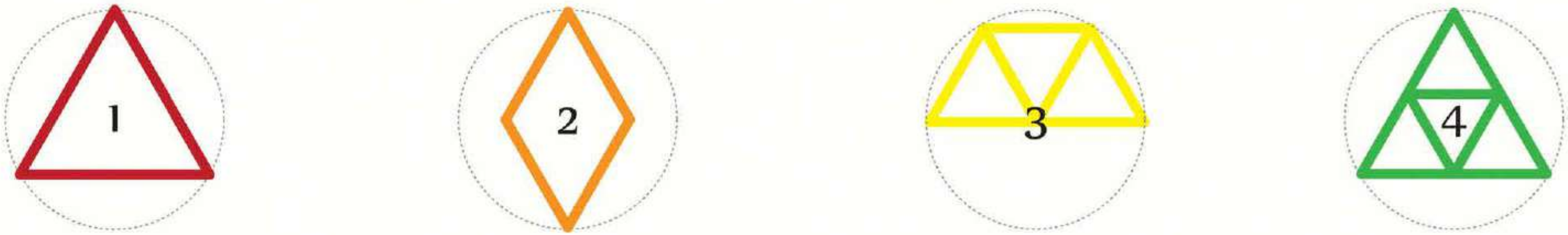

TETRYONICS

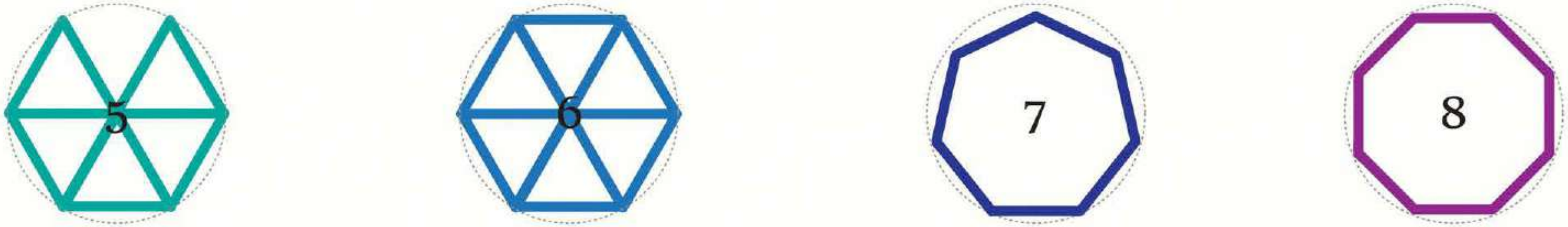

COUNTING POLYGONS

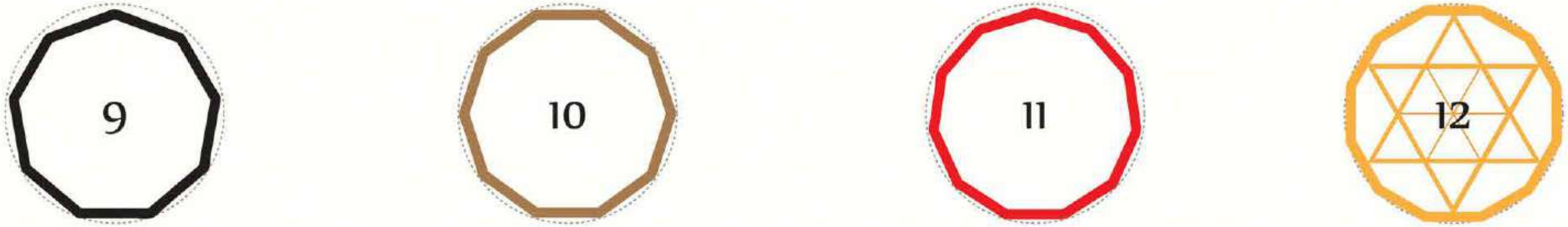

Tetryonics 82.02 - Counting Polygons

Copyright ABRAHAM [2008] - All rights reserved

SQUARED energies in quantum mechanics are EQUILATERAL geometries

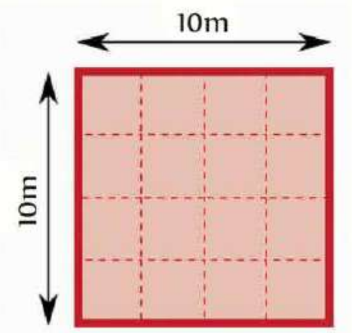

Square

area = s^2 = [100]

Circles

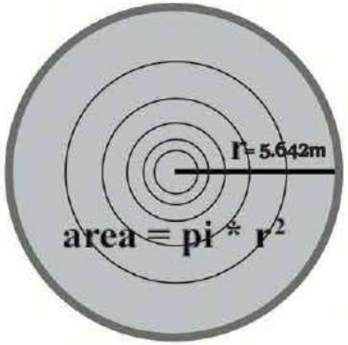
r = 5.642m
area = pi * r^2

= pi *$[5.642]^2$
= 100

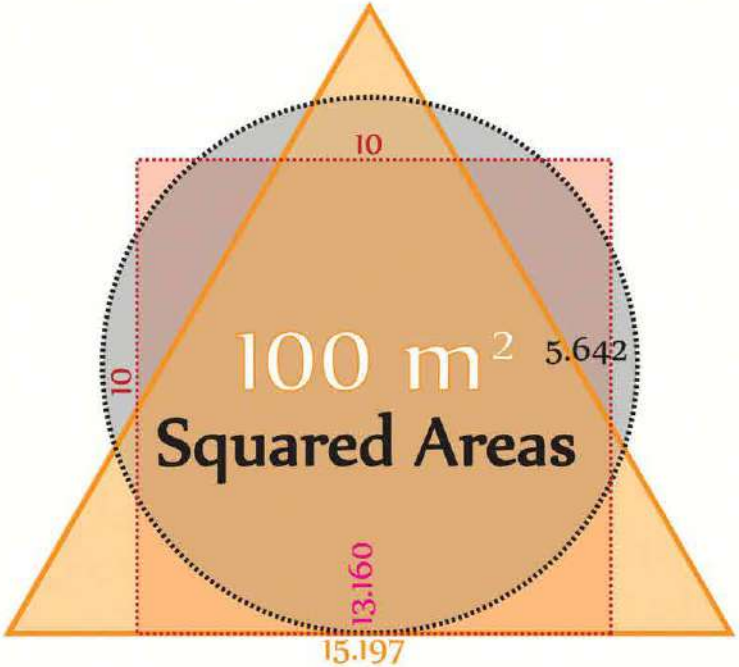
10
10
5.642
13.160
15.197
100 m^2
Squared Areas

can be created by a number of planar geometries

For a long time it has been assumed by scientists (and mathematicians) that circular [and squared] geometries are the geometric foundation of all physics, leading to a serioulsy flawed model of particles and forces in quantum mechanics

Tetryonic theory now reveals that quantised equilateral angular momenta creates the foundational geometry of all the mass-Energy-Matter & forces of physics

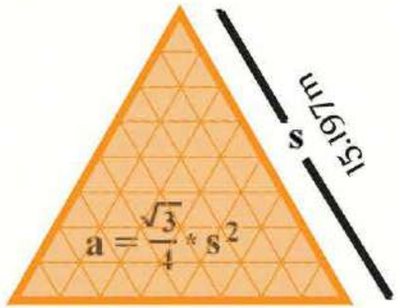
s
15.197m
$a = \frac{\sqrt{3}}{4} * s^2$

Equilateral

area = $(\frac{1}{2}*b)*h$

Triangles

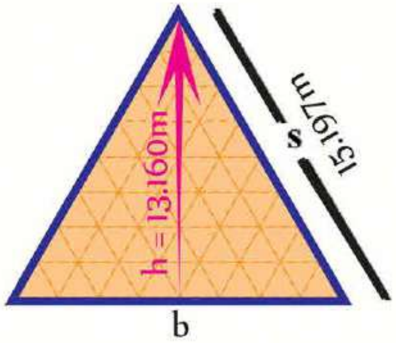
s
15.197m
h = 13.160m
b

b h
[.5x15.197] x 13.160
= 100

0.987654321

SQUARE ROOTS

80/81

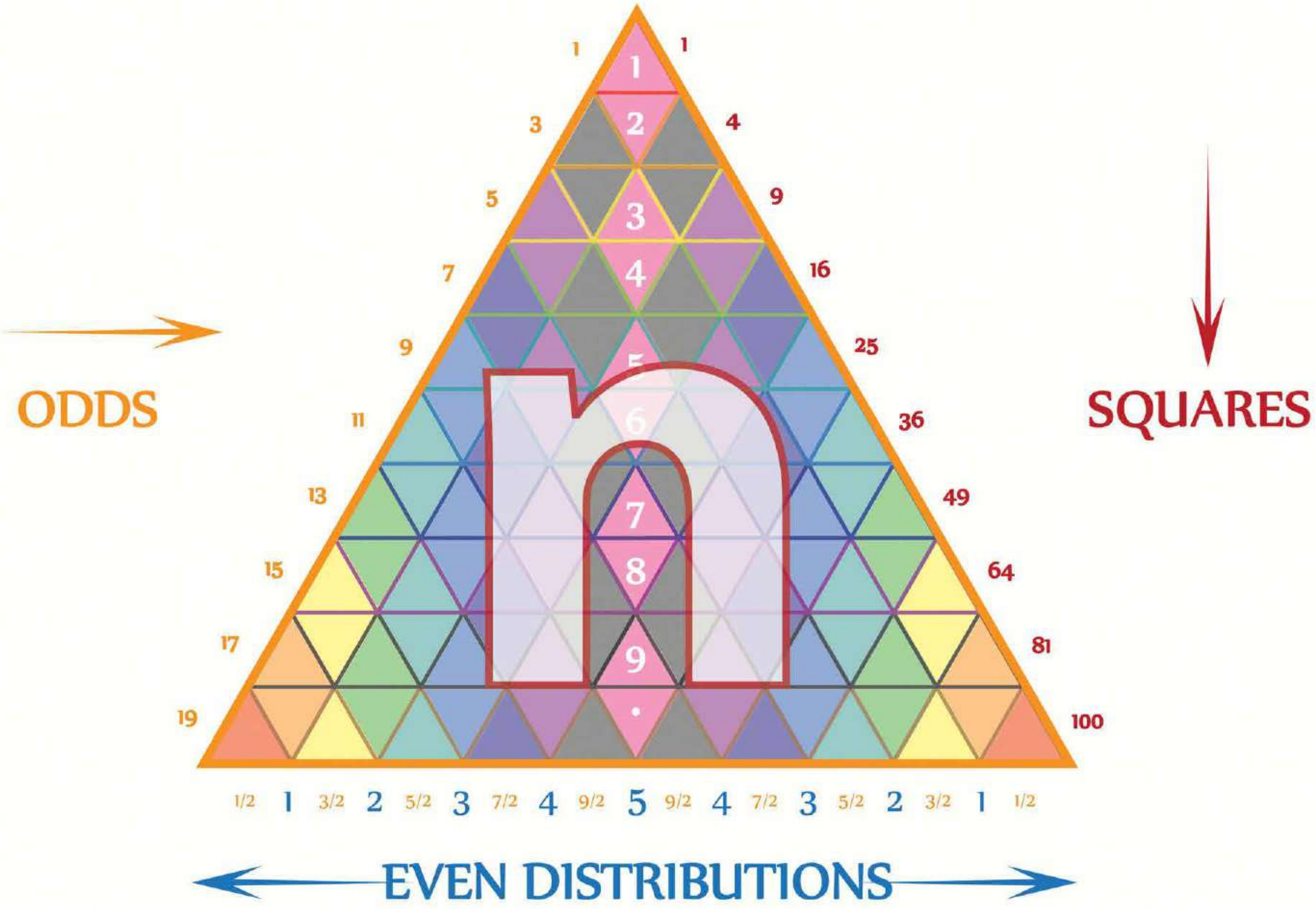

ODDS
1
3
5
7
9
11
13
15
17
19
SQUARES
1
4
9
16
25
36
49
64
81
100
1
2
3
4
5
6
7
8
9
•
n
1/2 1 3/2 2 5/2 3 7/2 4 9/2 5 9/2 4 7/2 3 5/2 2 3/2 1 1/2
EVEN DISTRIBUTIONS

Integers

The integers (from the Latin integer), literally "untouched", hence "whole"
in Tetryonics it is the basis for the Planck charge quantum

$$2n-1$$

$$[[n]+[n-1]]$$

Triangular numbered geometries are NOT equialteral geometries

0, 1, 3, 6, 10, 15, 21, 28, 36, 45, 55, ...

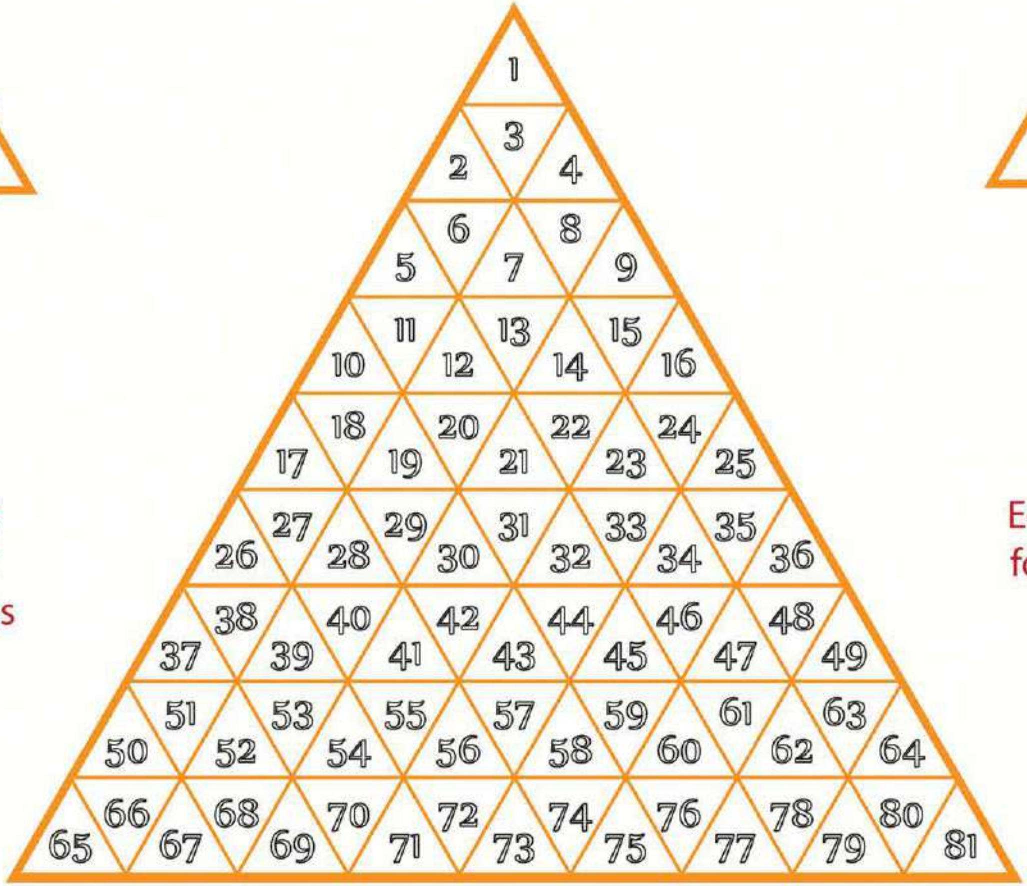

1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1

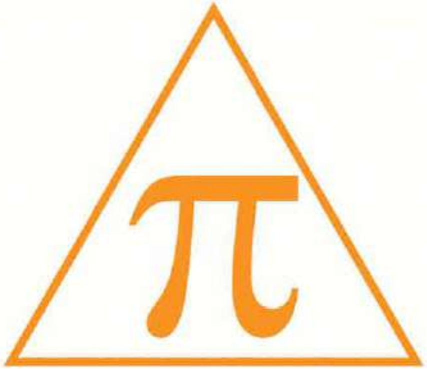

$$\sum_{1}^{n}[2n-1]$$

Equilateral geometries form **SQUARE** number geometries

1, 4, 9, 16, 25, 49, 64, 81, 100, 121, 144, ...

Equilateral energy quanta form a normal longitudinal distribution

Viewed as a subset of the real numbers, they are numbers that can be written without a fractional or decimal component

ε_0
π
μ_0
μ_0

Bosons are a transverse measure of scalar energy momenta

ODD numbers

An odd number is an integer which is not a multiple of two.

E
E
μ_0
μ_0
n
3
ε_0
ε_0
+
−
N
S
S
N

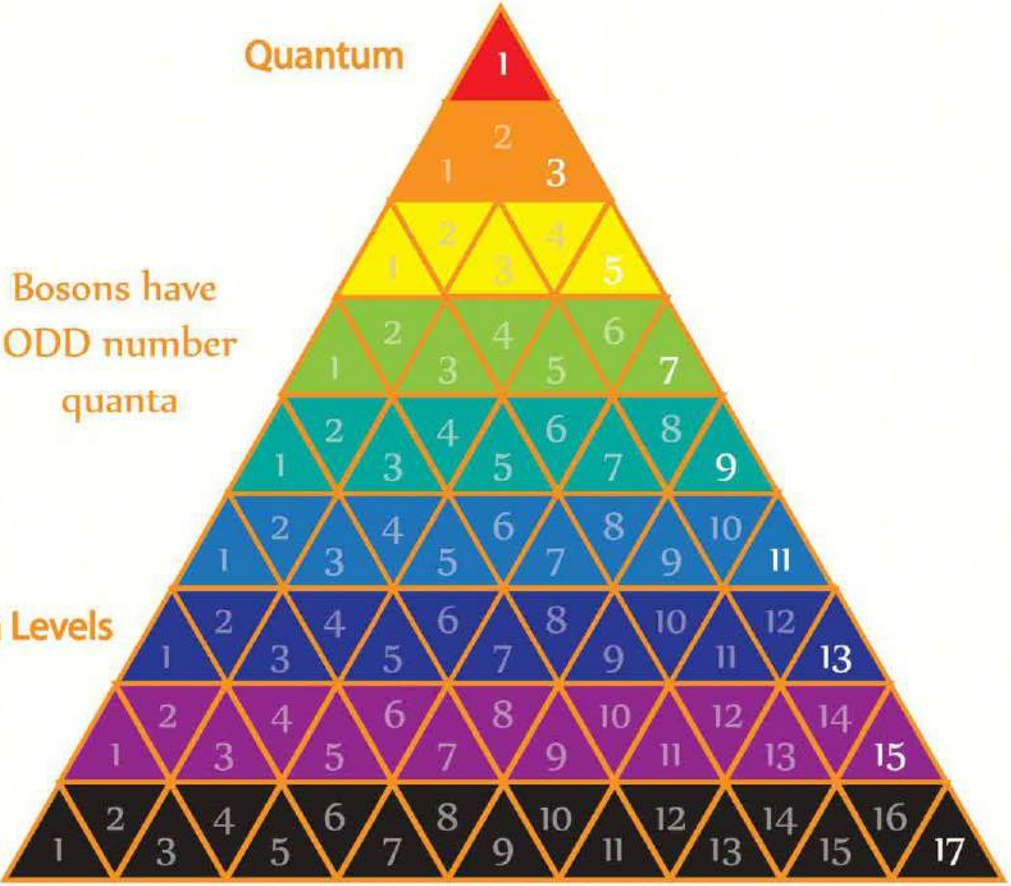
1
1 2 3
1 2 3 4 5
1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9 10 11
1 2 3 4 5 6 7 8 9 10 11 12 13
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Quantum

Bosons have
ODD number
quanta

Quantum Levels

hv

1
3
5
7
9
11
13
15
17

An odd number, when divided by two, will result in a fraction

Tau radians

Around the whole outside of a circle,
there are about 6.283 radians - or, Tau radians

3.141592654

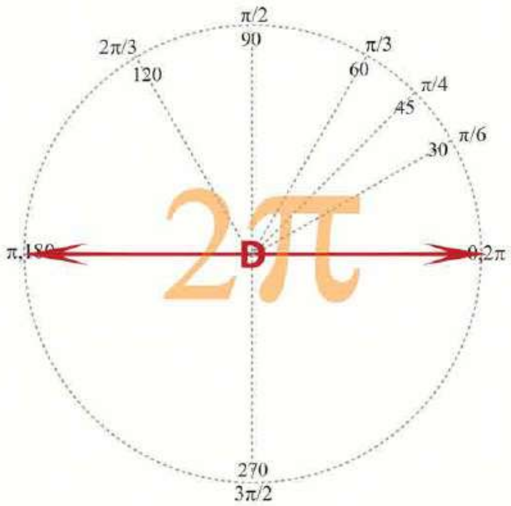
π/2
90
2π/3
120
π/3
60
π/4
45
π/6
30
π,180
2π
D
0,2π
270
3π/2

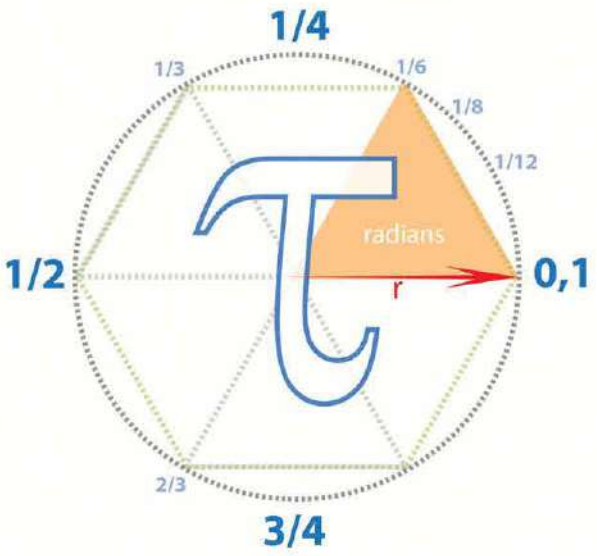
1/4
1/3
1/6
1/8
1/12
1/2
radians
0,1
r
2/3
3/4

6.283185307

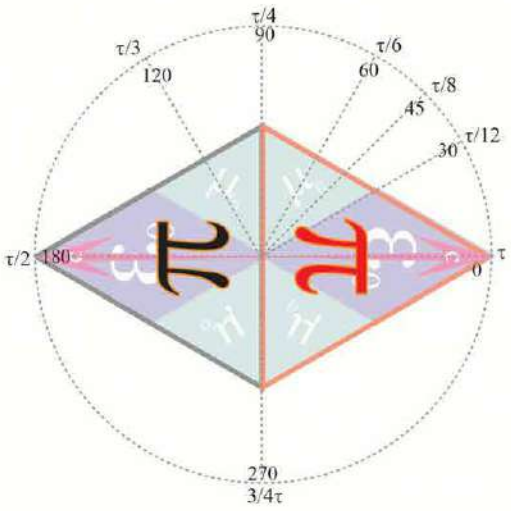
τ/4
90
τ/3
120
τ/6
60
τ/8
45
τ/12
30
τ/2
180
τ
0
270
3/4τ

τ is a more 'natural' radian system for geometric physics than π
Tau = 2π = 360 degree rotation about a point

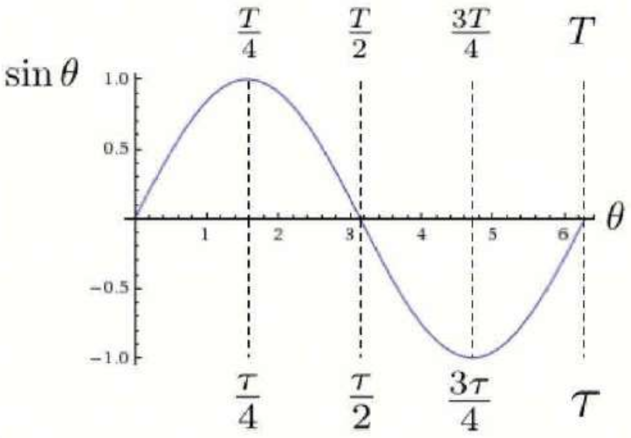

π
historically defined as the
ratio of a circle's circumference to its DIAMETER
should be redefined in physics to

τ
the ratio of its circumference to its RADIUS

in doing so many of the π/2 terms common to physics
will re automatically rationalised and will better reflect
the Tetryonic geometry of mass-ENERGY-Matter in motion

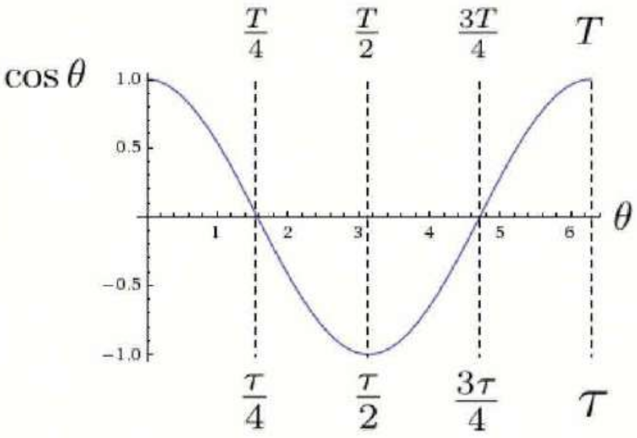

Copyright ABRAHAM [2008] - All rights reserved

Photons are a longitudinal measure of scalar energy momenta

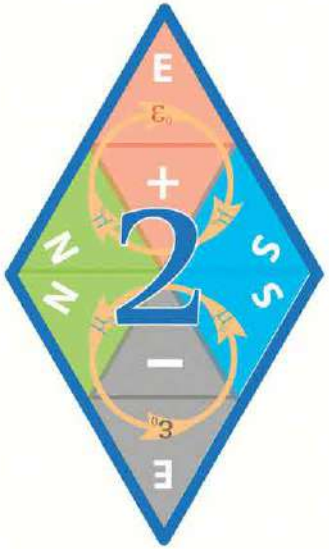
E
ε_0
+
2
N
N
S
S
−
ε_0
E

Photons are EVEN number quantums

EVEN numbers

An integer that is not an odd number is an even number

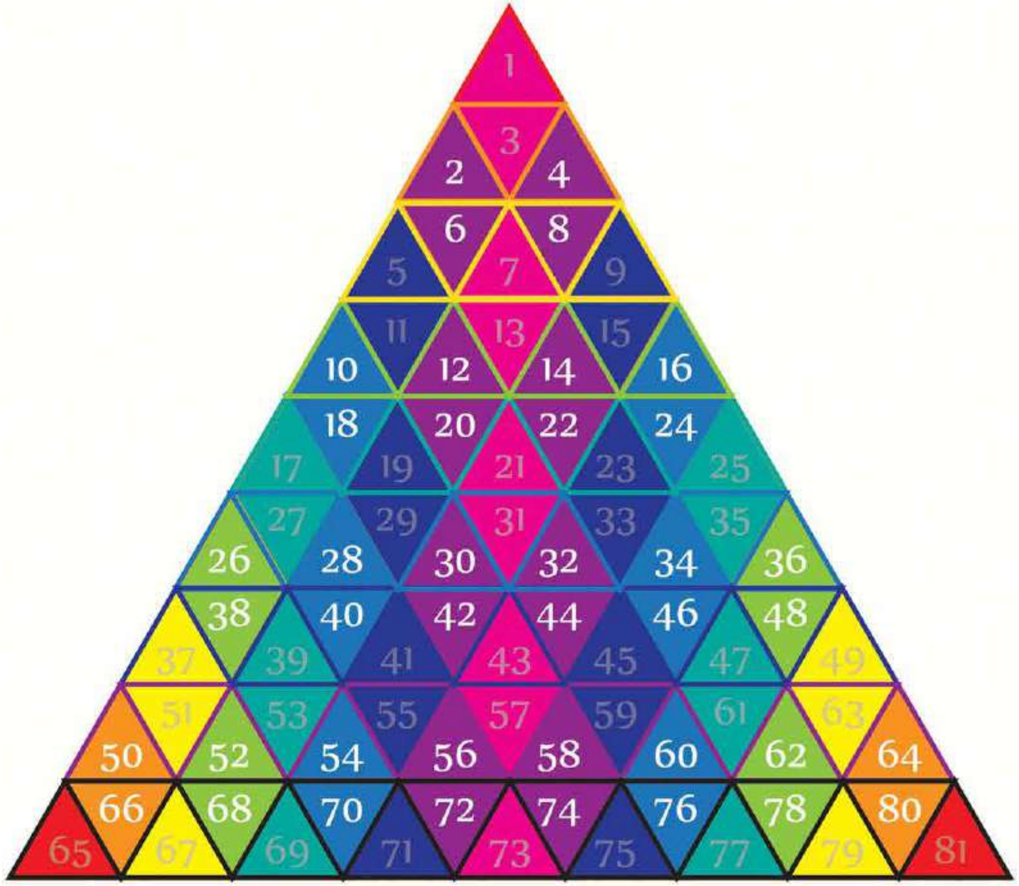
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

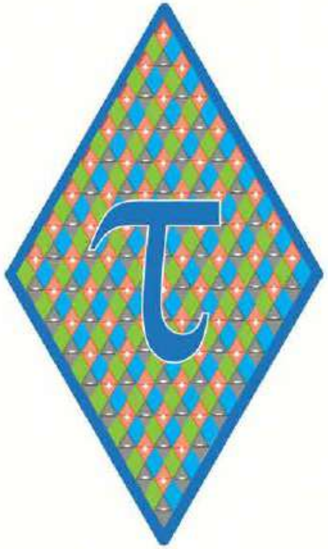

EM waves are comprised of EVEN numbered quanta

If an even number is divided by two, the result is another whole number

1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

If an odd number is divided by two, the result is a fractional number

Copyright ABRAHAM [2008] - All rights reserved

Triangular numbers

Historically, a triangular number counts quanta that can pack together to form an equilateral triangle

1, 3, 6, 10, 15, 21, 28, 36, 45, 55, ...

this form of geometric counting of same charges over-complicated the simpler physical reality

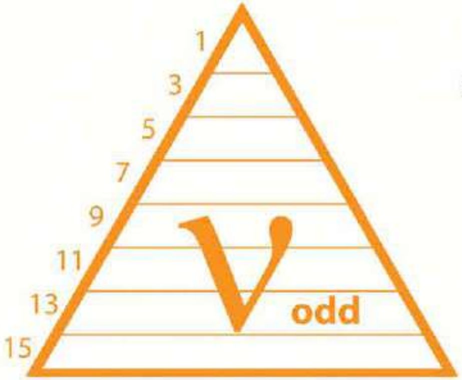
1
3
5
7
9
11
13
15
ν
odd

$$2\nu-1$$

$$[[\nu]+[\nu-1]]$$

Equilateral chords or quantum levels are ODD numbers

1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21,....

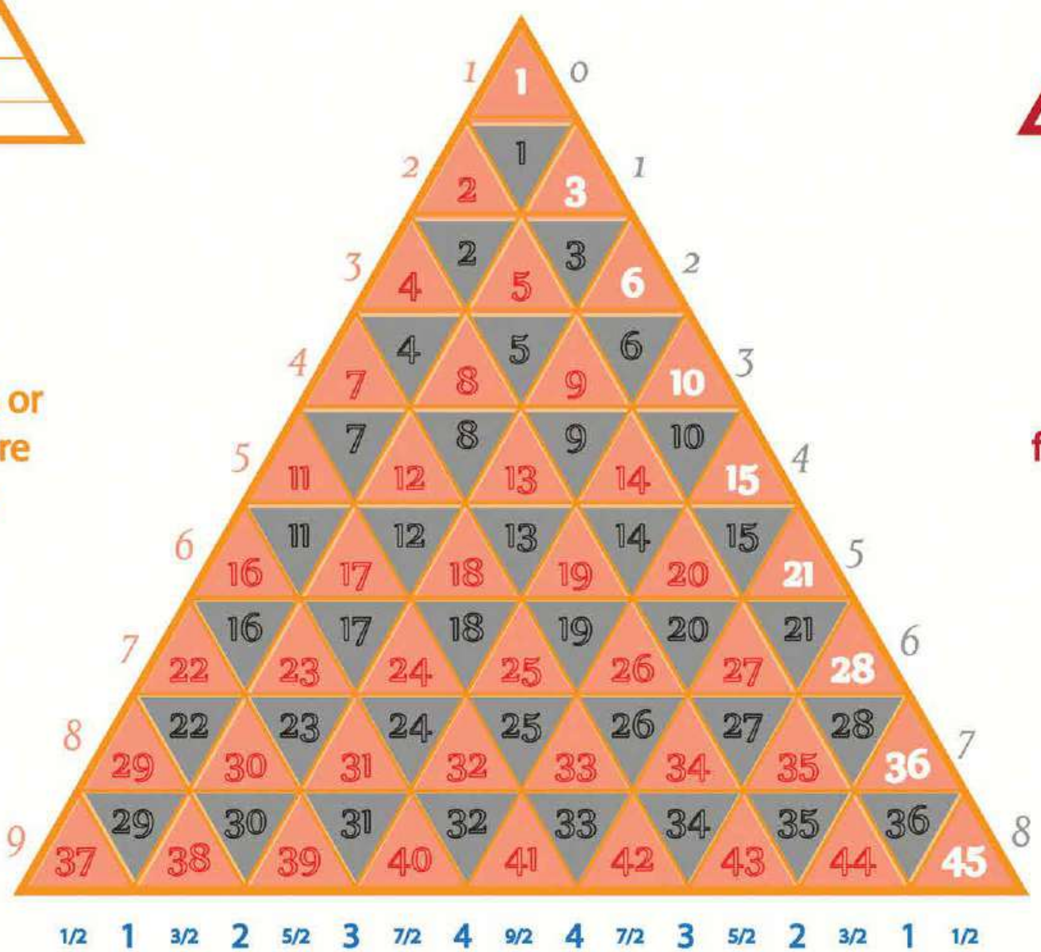
1 1 0
2 2 1 3 1
3 4 2 5 3 6 2
4 7 4 8 5 9 6 10 3
5 11 7 12 8 13 9 14 10 15 4
6 16 11 17 12 18 13 19 14 20 15 21 5
7 22 16 23 17 24 18 25 19 26 20 27 21 28 6
8 29 22 30 23 31 24 32 25 33 26 34 27 35 28 36 7
9 37 29 38 30 39 31 40 32 41 33 42 34 43 35 44 36 45 8
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

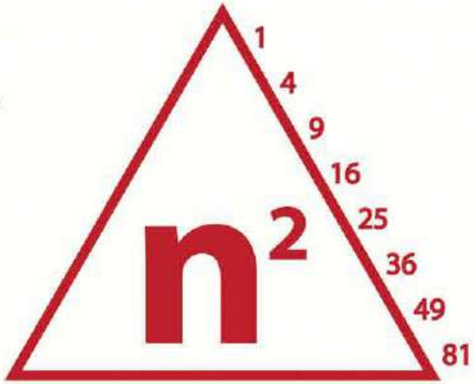
1
4
9
16
25
36
49
81
n
2

$$\sum_{1}^{n}[2n-1]$$

Equilateral geometries form SQUARE numbered geometries

1, 4, 9, 16, 25, 49, 64, 81, 100, 121, 144, ...

Triangular energy quanta form normal distributions

Squared numbers

A square number, sometimes also called a perfect square,
is the result of an integer multiplied by itself

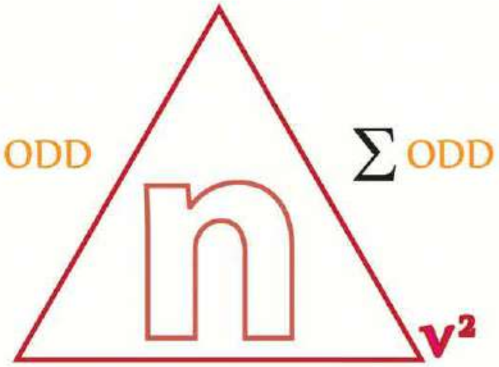

1
+
3
+
5
+
7
+
9
+
11
+
13
+
15
+
17

Square numbers result from the summing of consecutive ODD numbers

ODDS

2n-1

[[n] + [n-1]]

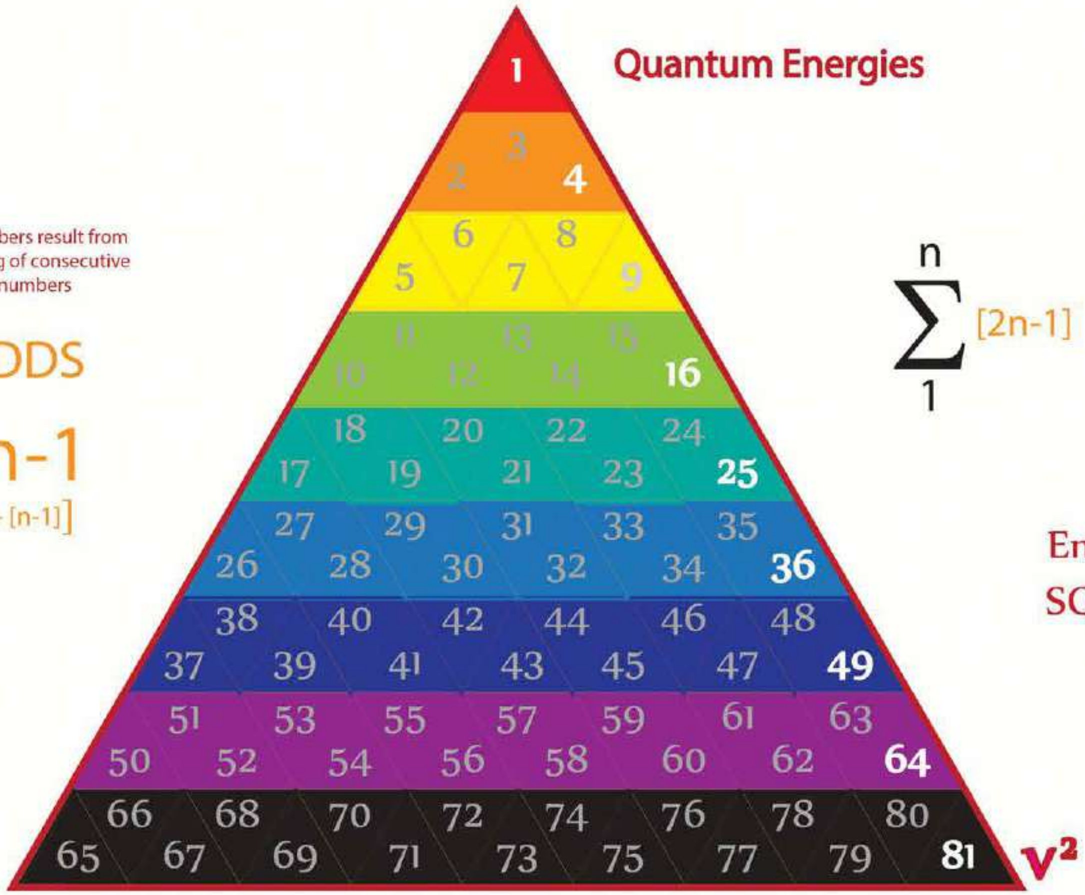
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
v²

Quantum Energies

$$\sum_{1}^{n} [2n-1]$$

Energy levels have
SQUARE number
quanta

Compton Frequency

In Tetryonics Square numbers produce equilateral geometries

Square roots

A square root of a number is a number that, when it is multiplied by itself (squared) , gives the first number again.

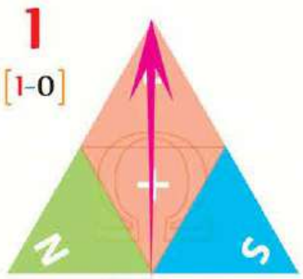

Root of positive one

-*i* and +*i*

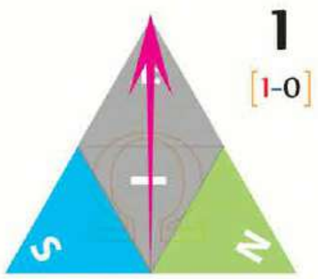

Root of negative one

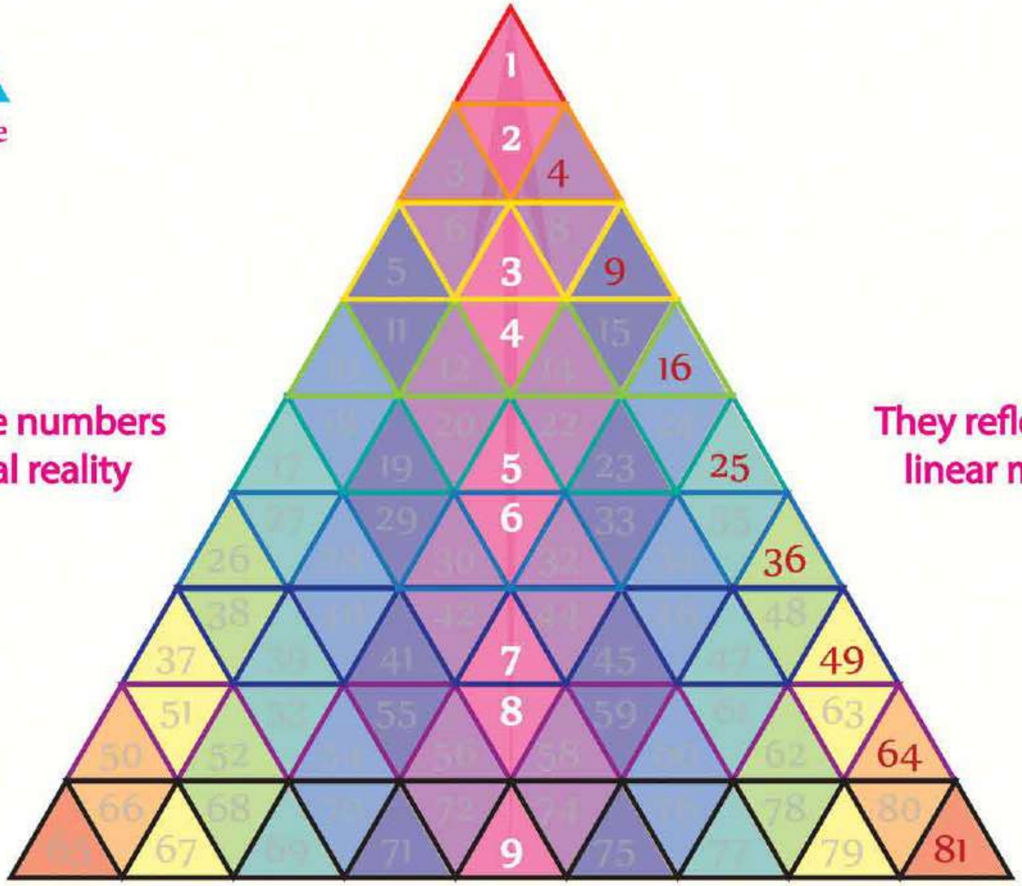
1
2
4
3
9
5
4
16
5
25
19
23
6
36
29
33
26
7
49
37
41
45
48
38
8
64
51
55
59
63
50
52
62
9
81
66
67
68
71
75
78
79
80

Square roots of negative numbers have a basis in physical reality

They reflect the real non-negative linear momentum of a system

A whole number with a square root that is also a whole number is called a perfect square in Tetryonic theory they are actually equilateral geometries

Real Numbers

In mathematics, a real number is a value that represents a quantity along a continuous line. The real numbers include all the rational numbers,

Quantum levels

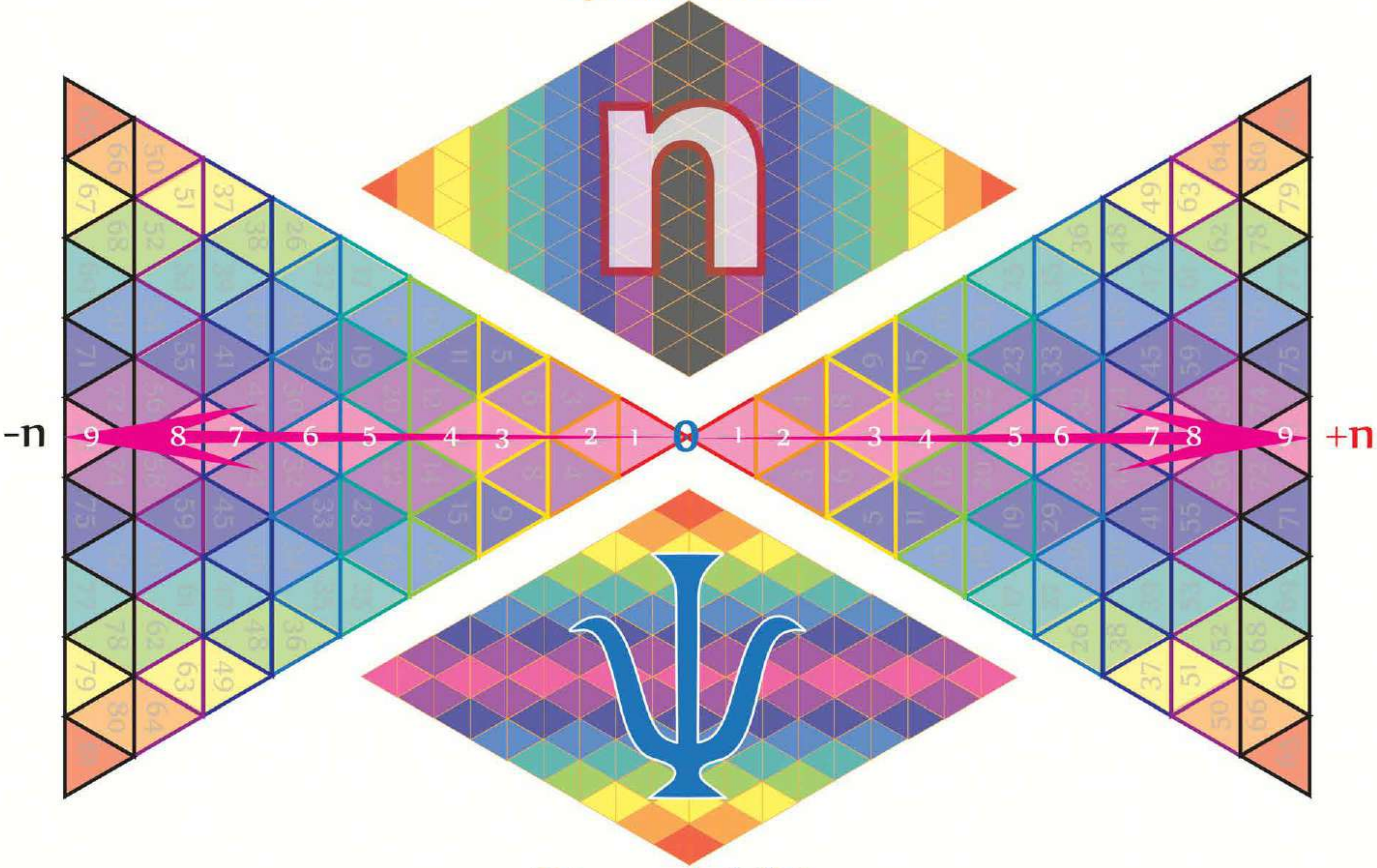

Wave probabilities

Copyright ABRAHAM [2008] - All rights reserved

ONE to ∞^2

Quantum levels

ODD numbers
[2n-1]
1, 3, 5, 7, 9......

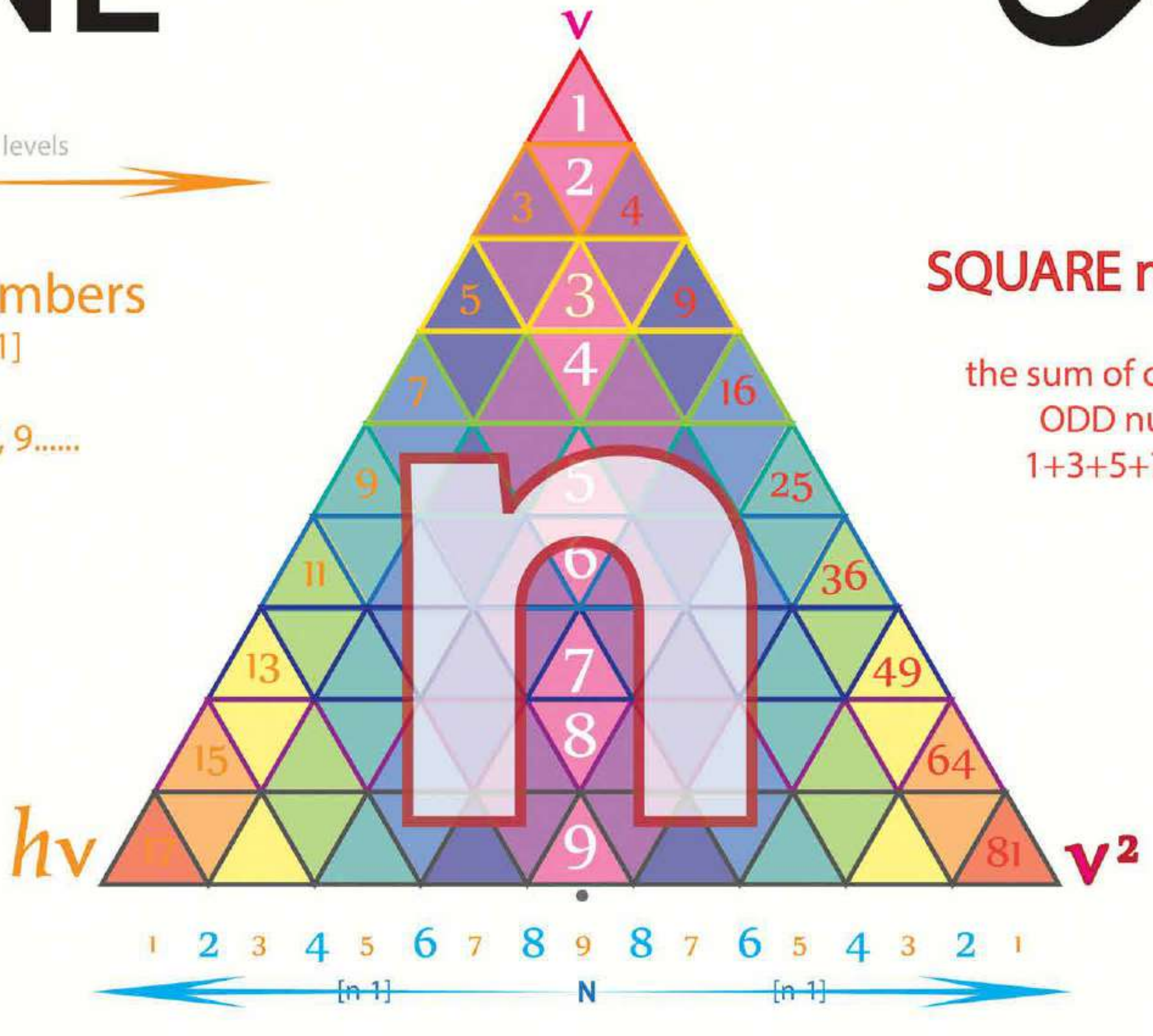
v
1
2
3
4
3
5
9
4
7
16
5
9
25
6
11
36
7
13
49
8
15
64
9
17
81
n
hv
v²
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1
[n-1]
N
[n-1]

SQUARE numbers

the sum of consecutive
ODD numbers
1+3+5+7+9+......

Scalar Energies

Linear energy momentum

NORMAL DISTRIBUTION

Tetryonics 82.13 - One to Infinity squared

Copyright ABRAHAM [2008] - All rights reserved

Basic Properties of nested scribed Equilateral Triangles

Given an equilateral triangle of side s

altitude

$\sqrt{3/2}\,S$

$3S$

perimeter

$\sqrt{3/4}\,S^2$

area

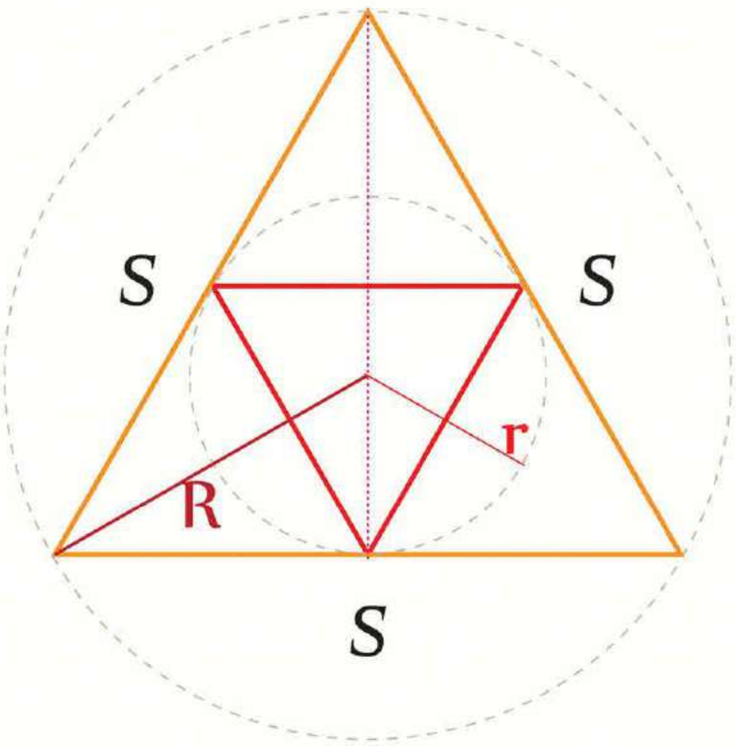

$\sqrt{3/6}\,S$

in-radius

$\sqrt{\pi/12}\,S^2$

in-circle area

circum-radius

$\sqrt{3/3}\,S$

circum-circle area

$\sqrt{\pi/3}\,S^2$

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic [equilateral] geometry

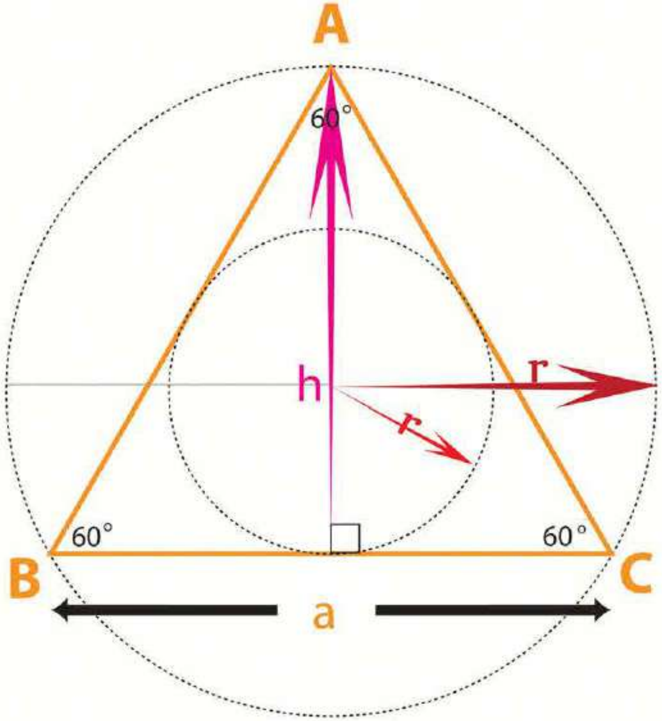

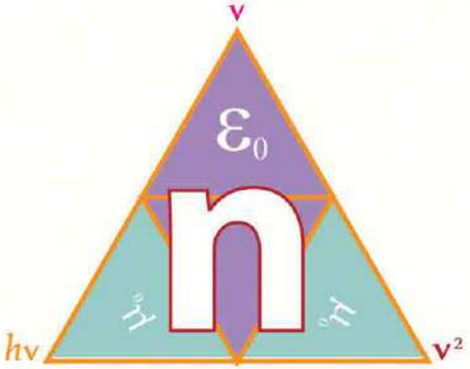

The equilateral triangle exhibits 'square symmetry'
it can always be divided into n^2 [number]
of smaller self-similar parts

**All triangles are
flat euclidean
π radian
geometries
[180°]**

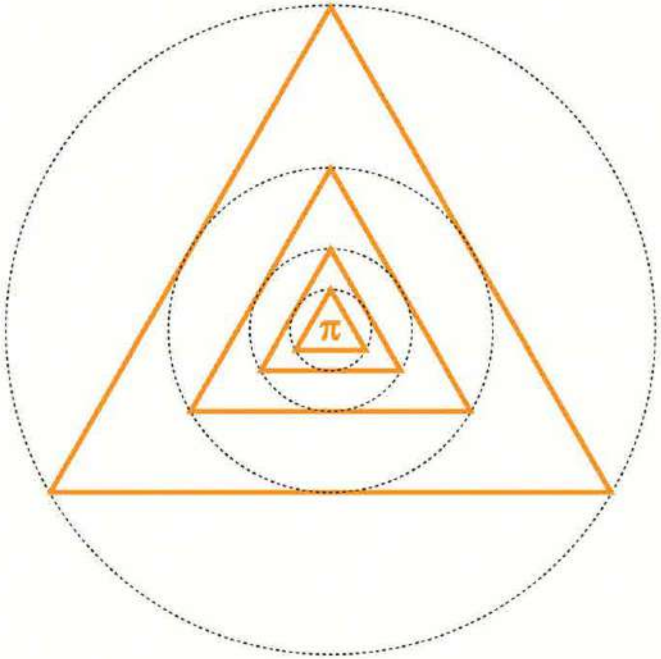

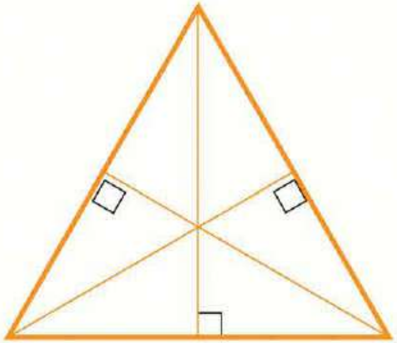

An equilateral triangle is the most symmetrical triangle,
having 3 lines of reflection and rotational symmetry of order 3 about its center

CHARGE

Scribed equilateral geometries

reflect space-time's geometric relationship with charged mass-energy

Energy

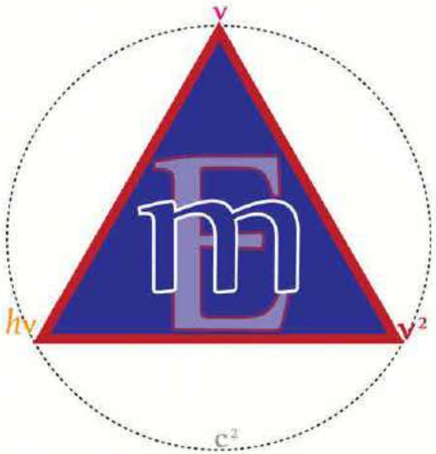

QAM

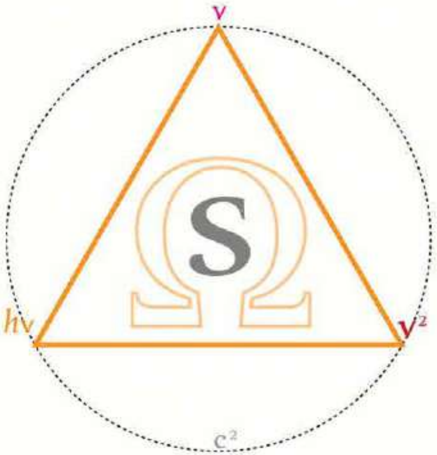

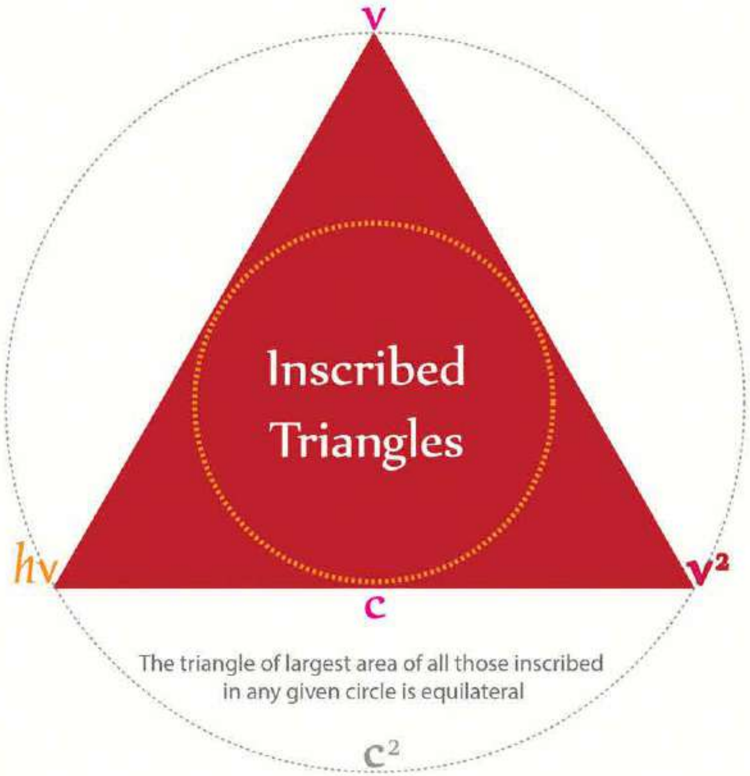

Inscribed Triangles

The triangle of largest area of all those inscribed in any given circle is equilateral

Energy per spatial unit

mass

QAM per spatial unit

seconds

$$m = \frac{E}{c^2}$$

$$\frac{\Omega}{c^2} = s$$

Copyright ABRAHAM [2008] - All rights reserved

Circumscribed Triangles

reflect Energy's relationship with Time

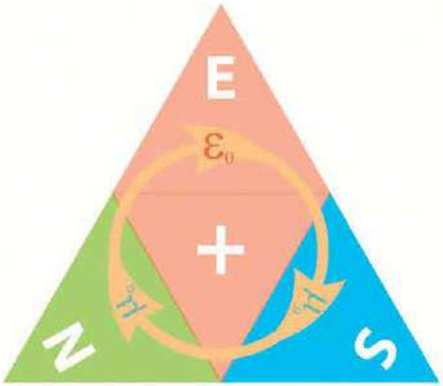
E
ε_0
+
N
S
μ_0
μ_0

Positive Planck Charge

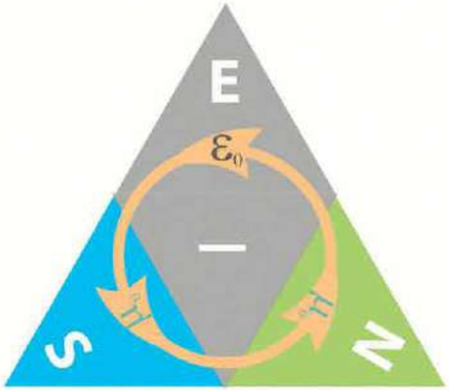
E
ε_0
−
S
N
μ_0
μ_0

Negative Planck Charge

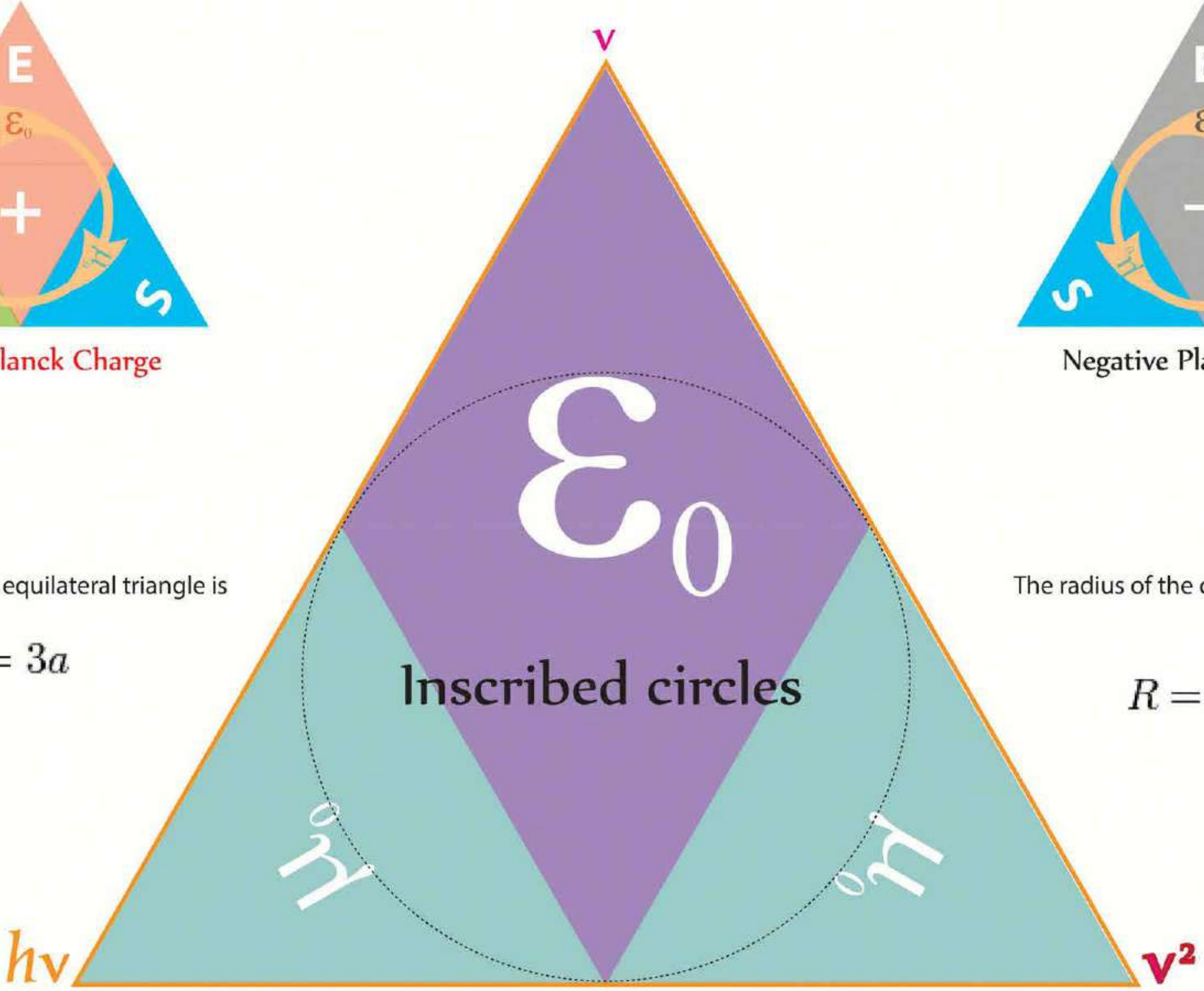
v
ε_0
Inscribed circles
μ_0
μ_0
hv
v²

The perimiter of an equilateral triangle is

$$p = 3a$$

The radius of the circumscribed circle is

$$R = \frac{\sqrt{3}}{3}a$$

The equilateral triangle has the smallest area of all those circumscribed around a given circle

Circumscribed circles

By Euler's inequality, the equilateral triangle has the smallest ratio R/r
of the circumradius to the inradius of any triangle: specifically, R/r = 2

The ratio of the area to the square of the perimeter of an equilateral triangle

$$\frac{1}{12\sqrt{3}}$$

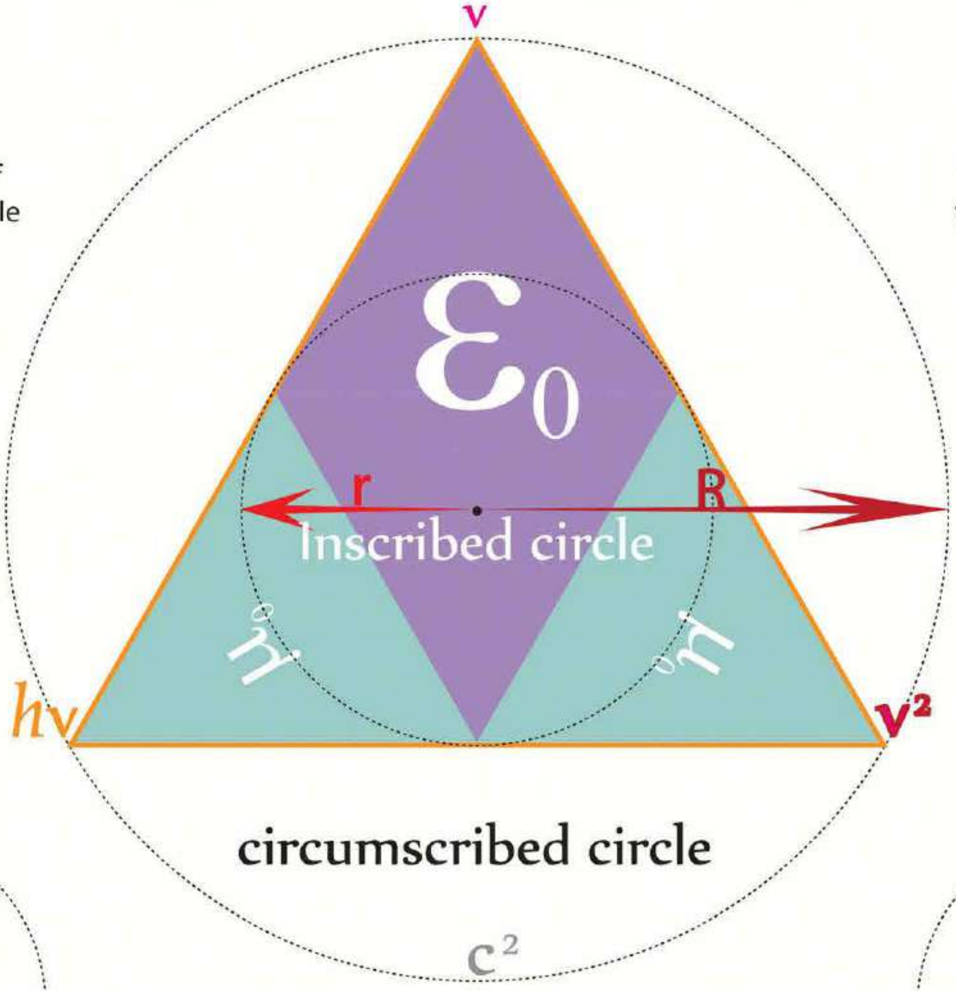
v
ε_0
r
R
Inscribed circle
hv
v²
circumscribed circle
c²

The ratio of the area of the incircle to the area of an equilateral triangle

$$\frac{\pi}{3\sqrt{3}}$$

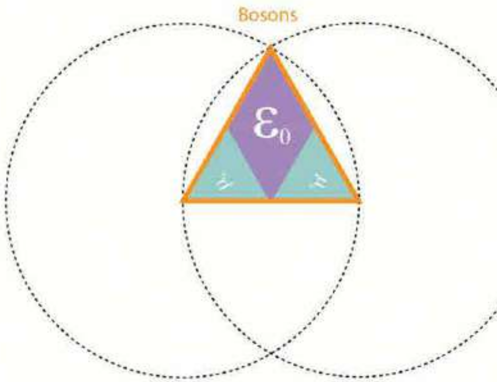
Bosons
ε_0

Equilateral triangles and Tetrahedrons will scale at exactly the same proportion as Circles and Spheres scribing them

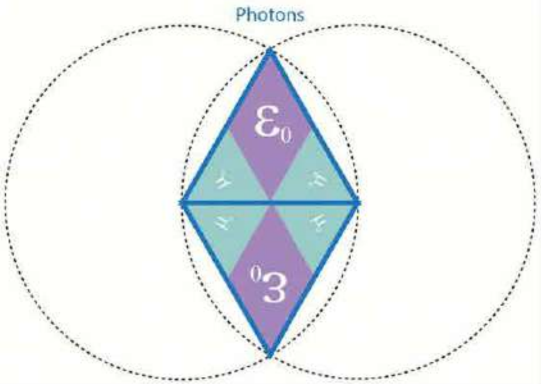
Photons
ε_0

Charged mass-ENERGIES

Charge is the result of quantised angular momenta
[the inscribed circular flux of energy in equilateral geometries]

Time is a measure of changing quantised angular momenta
[the circumscribed spatial co-ordinate of equilateral energy geometries]

inscribed

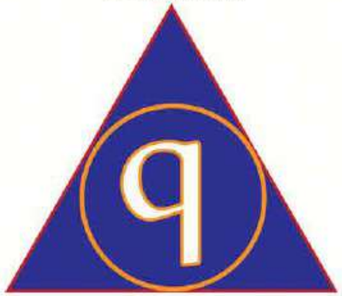

charge

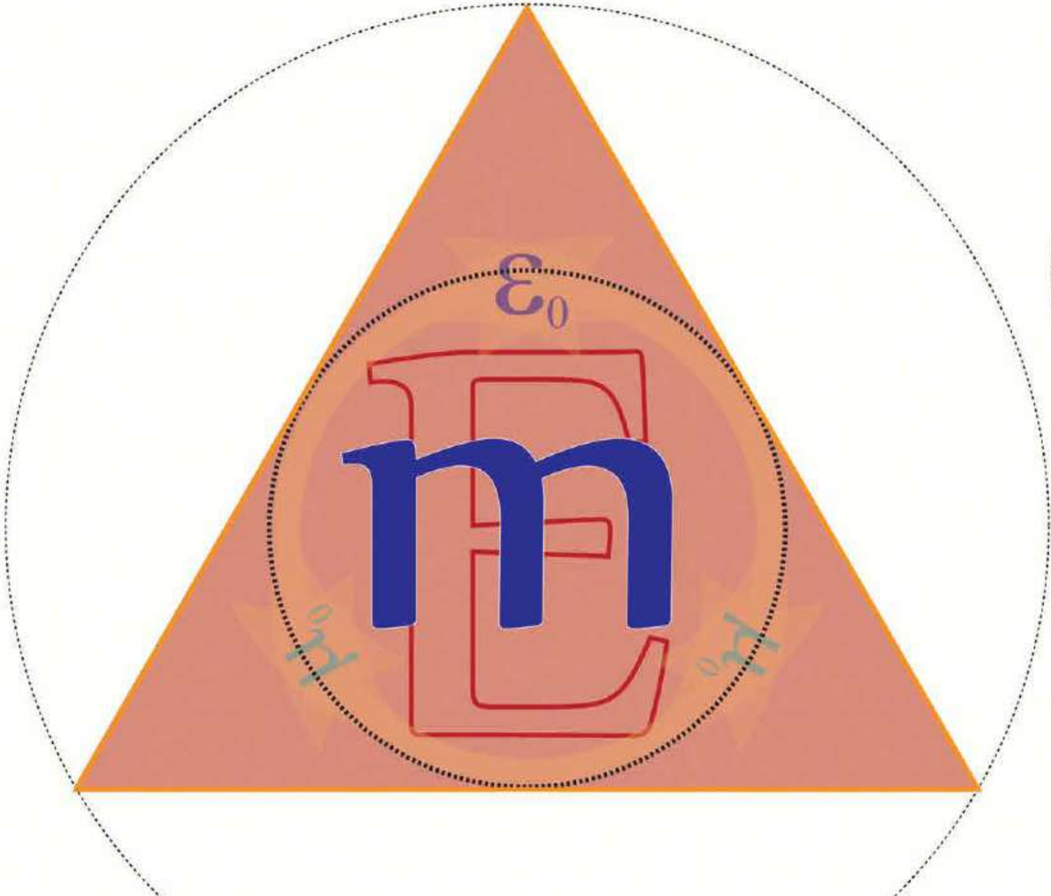

circumscribed

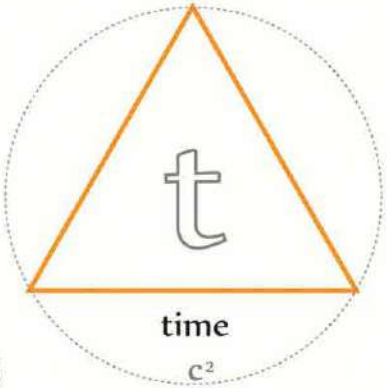

It is the equilateral geometry of energy not a classical vector rotation that creates QAM

Scalar EM mass is a measure of equilateral Planck energy per spatial co-ordinate system

Charged mass-energies can be described as
a vector measure of scalar Energy momenta per unit of time
[inscribed circular fluxes of equilateral energy per circumscribed temporal geometry]

Copyright ABRAHAM [2008] - All rights reserved

Trigometric functions

Standard trigometric functions must be carefully applied to measurements of equilateral Planck mass-energy geometries in scribed circular space-time co-ordinate systems

$$\sin\,\theta = \frac{\text{opposite}}{\text{hypotenuse}}$$

$$\cos\,\theta = \frac{\text{adjacent}}{\text{hypotenuse}}$$

$$\tan\,\theta = \frac{\text{opposite}}{\text{adjacent}}$$

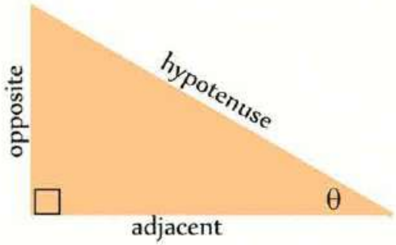
opposite
hypotenuse
θ
adjacent

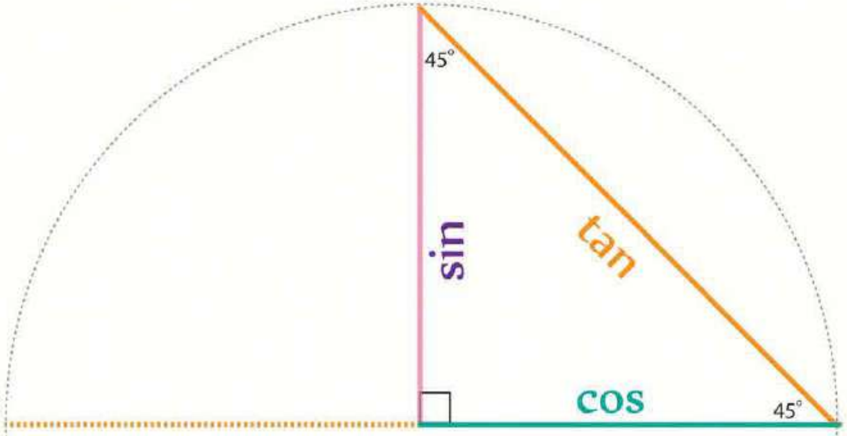
45°
sin
tan
cos
45°

The most familiar trigonometric functions are the sine, cosine, and tangent.

In the context of the standard unit circle with radius 1, where a triangle is formed by a ray originating at the origin and making some angle with the x-axis,

the SINE of the angle gives the length of the y-component (rise) of the triangle, the COSINE gives the length of the x-component (run), and the TANGENT function gives the slope (y-component divided by the x-component)

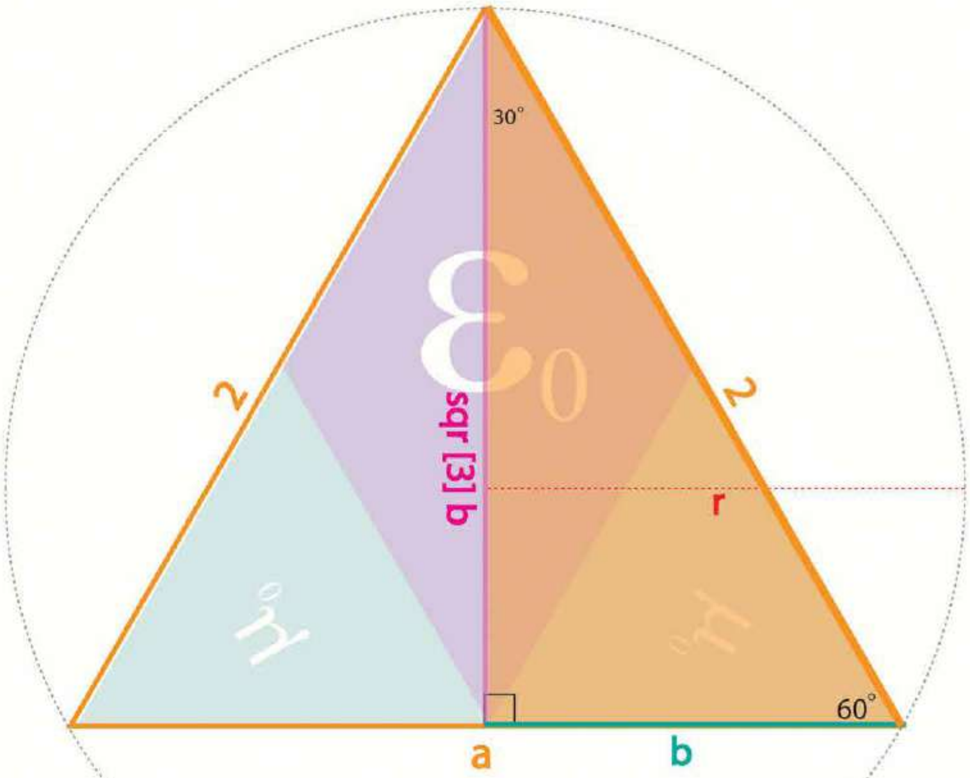
30°
ε_0
sqr [3] b
2
2
r
μ_0
60°
a
b

In most cases it is in fact easier to ignore the trigonomteric functions and math and simpy draw their related equilateral geometries

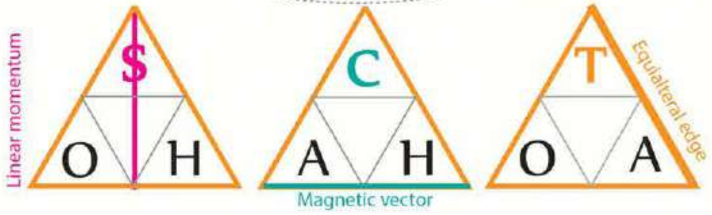
Linear momentum
$
O
H
C
A
H
Magnetic vector
T
O
A
Equilateral edge

The roots of scribed equilateral triangles

Scalar equilateral energies map directly onto circular space-time co-ordinates
through their square root linear momentum

r

The ratio of the
circumscribed circle of
an equilateral triangle
to its inscribed circle
is 2:1

d

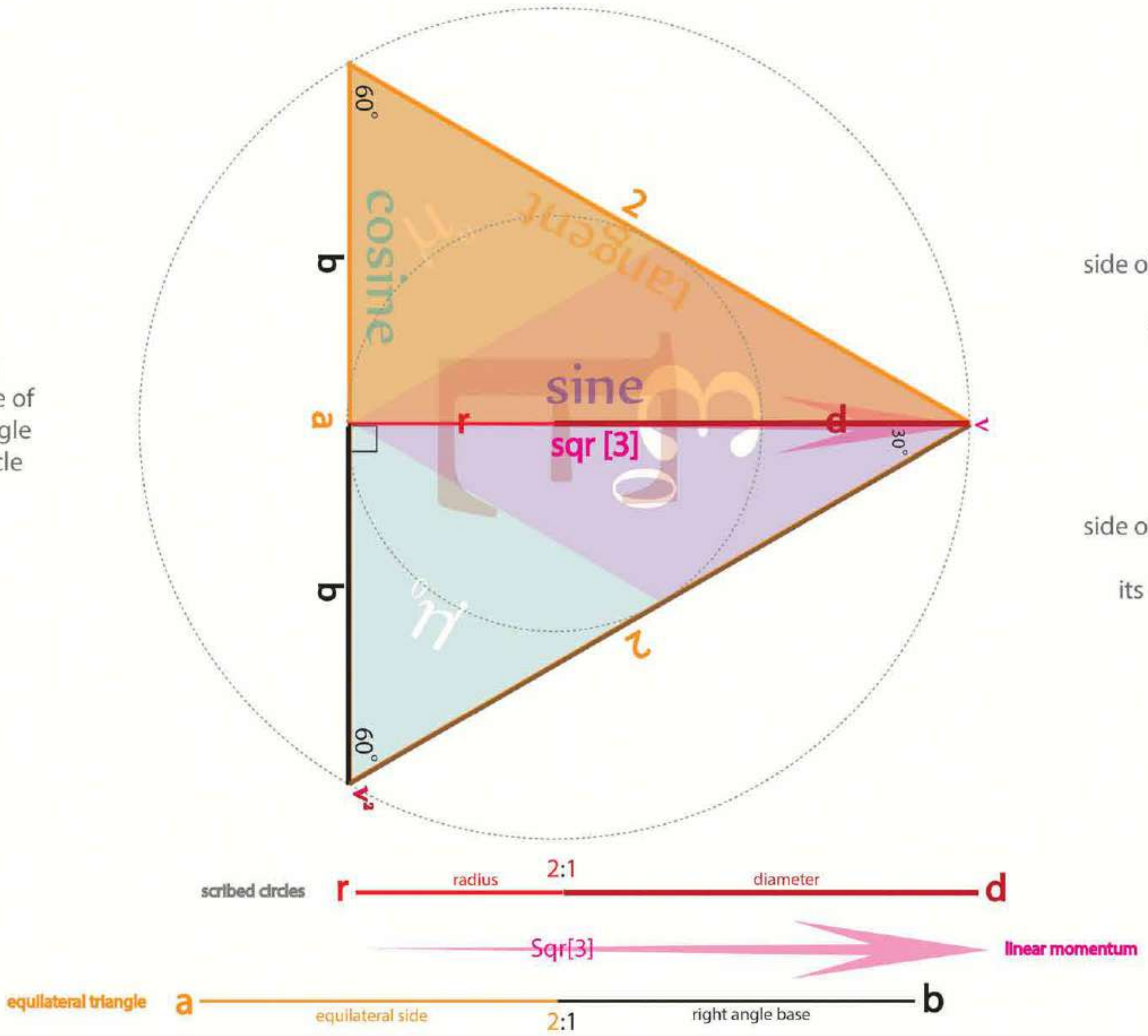
60°
cosine
tangent
2
b
a
r
sine
d
sqr [3]
30°
v
b
2
60°
v²
scribed circles
r
radius
2:1
diameter
d
Sqr[3]
linear momentum
equilateral triangle
a
equilateral side
2:1
right angle base
b

0.866 r

The ratio of the
side of an equilateral triangle
to the radius of
its incribed circle
is sqr [3]/2

a

The ratio of the
side of an equilateral triangle
to the radius of
its circumscribed circle
is sqr [3]

1.732 d

Tetryonics 83.08 - Linear mapping of energy momentum

Equilateral triangles and scribed circles

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1 - \beta^2}} = \frac{dt}{d\tau}$$

$$\gamma(\beta) = 1 + \frac{1}{2}\beta^2 + \frac{3}{8}\beta^4 + \frac{5}{16}\beta^6 + \frac{35}{128}\beta^8 + \ldots$$

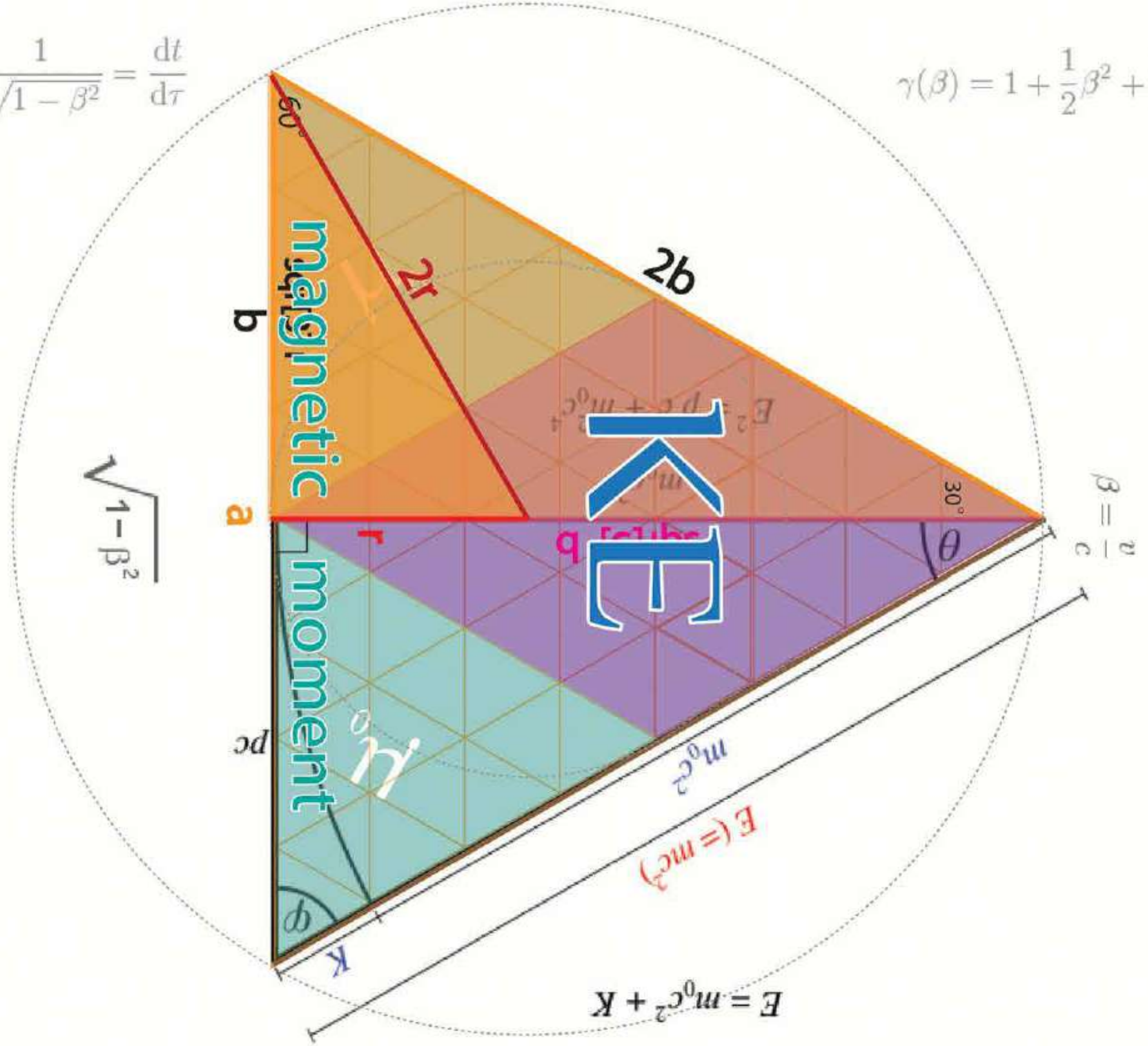
60°
2r
2b
b
magnetic moment
μ
a
r
q
KE
30°
θ
$\beta = \frac{v}{c}$
$\sqrt{1-\beta^2}$
pc
μ_0
m_0c^2
E (= mc²)
φ
K
$E = m_0c^2 + K$
$E = pc + m_0c^2$

$$L = L'\sqrt{1 - \frac{v^2}{c^2}}$$

$$E_k = \frac{mc^2}{\sqrt{1 - (v/c)^2}} - mc^2$$

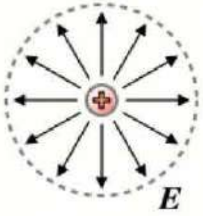

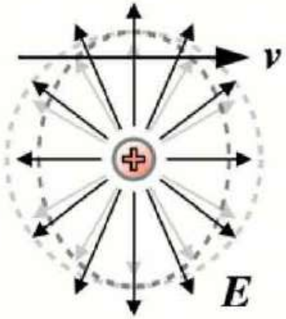

$$mc^2\left\{\frac{1}{\sqrt{1 - v^2/c^2}} - 1\right\}$$

$$p = \frac{m_0 v}{\sqrt{1 - \frac{v^2}{c^2}}} = \gamma m_0 v$$

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}$$

All the relativistic relationships historically attributed to circularised energies and modelled through the use of the math of right angled triangles are in fact the result of equilateral, scalar geometries

$$\sqrt{1 - \left(\frac{v}{c}\right)^2}$$

Copyright ABRAHAM [2008] - All rights reserved

Finite sequences and series have defined first and last terms,

$$\sum a_n = a_0 + a_1 + a_2$$

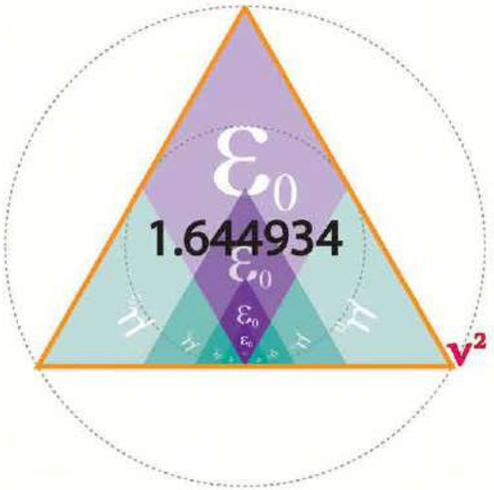

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \lim_{n \to +\infty} \left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} \right).$$

whereas infinite sequences and series continue indefinitely

The Basel problem is a famous problem in mathematical analysis with relevance to number theory, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735.

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \ldots = \frac{\pi^2}{6}$$

Tetryonics now provides a geometric solution to visualising and solving the Basel problem

$$\zeta(2) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \frac{\pi^2}{6} \approx 1.645;$$

Tetryonic Infinite Series

is a infinite sequence of square numbers,
the result of adding all those terms together
[or their geometric inverse]

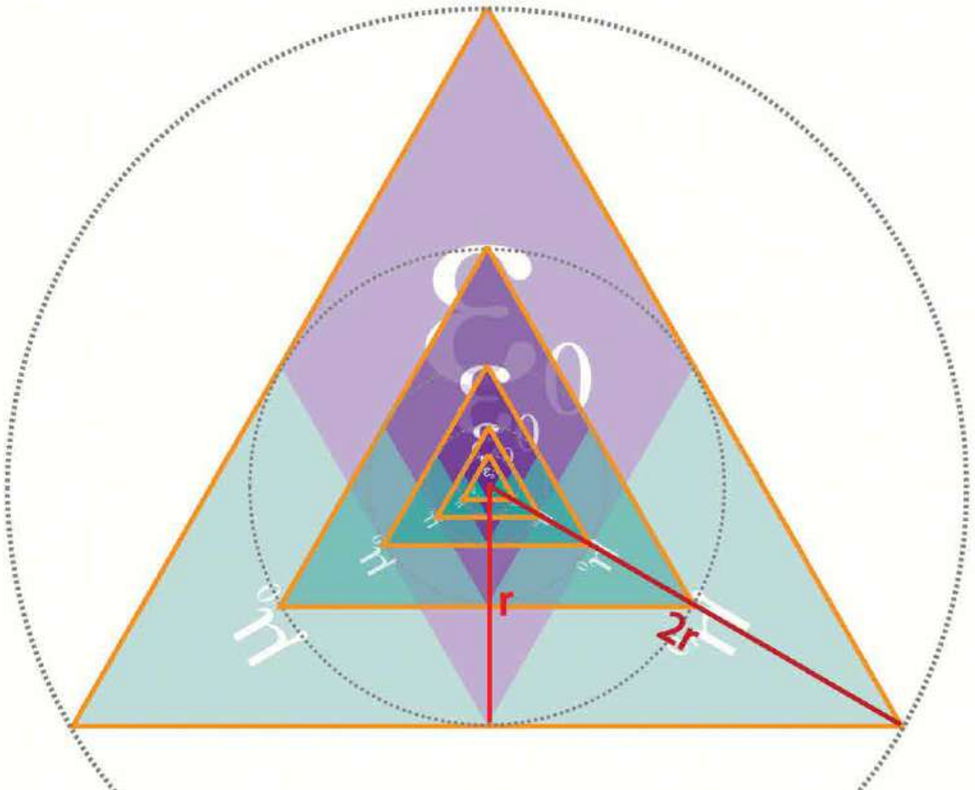

$$\sum_{n \geq 1} \frac{1}{n^2}$$

The entire sum of the series is equal to
twice the size of the radius of the largest inscribed circle
which is equal to the largest circle circumscribing the triangular series.

Copyright ABRAHAM [2008] - All rights reserved

Inverting the Circle

Electric
Permittivity

ε_o

Magnetic
Permeability

μ_o

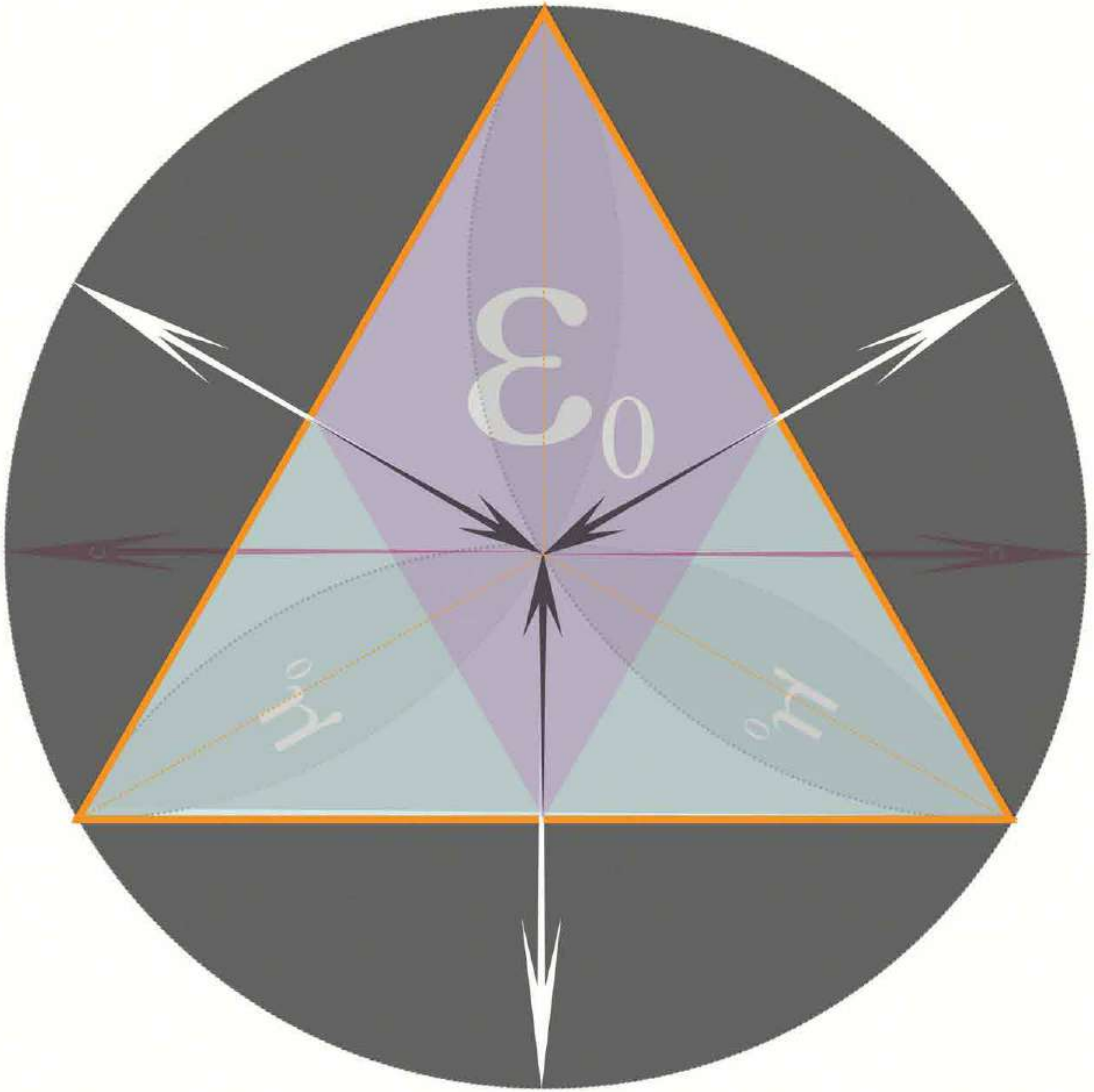

c^2

$\frac{1}{c^2}$

2D spatial
co-ordinates

ElectroMagnetic
fields

Tetryonics 83.11 - Inverting the Circle

Copyright ABRAHAM [2008] - All rights reserved

Irrational numbers

an irrational number cannot be represented as a simple fraction.
Irrational numbers are those real numbers that cannot be represented as terminating or repeating decimals

sin(x)=-square root of (3)/2

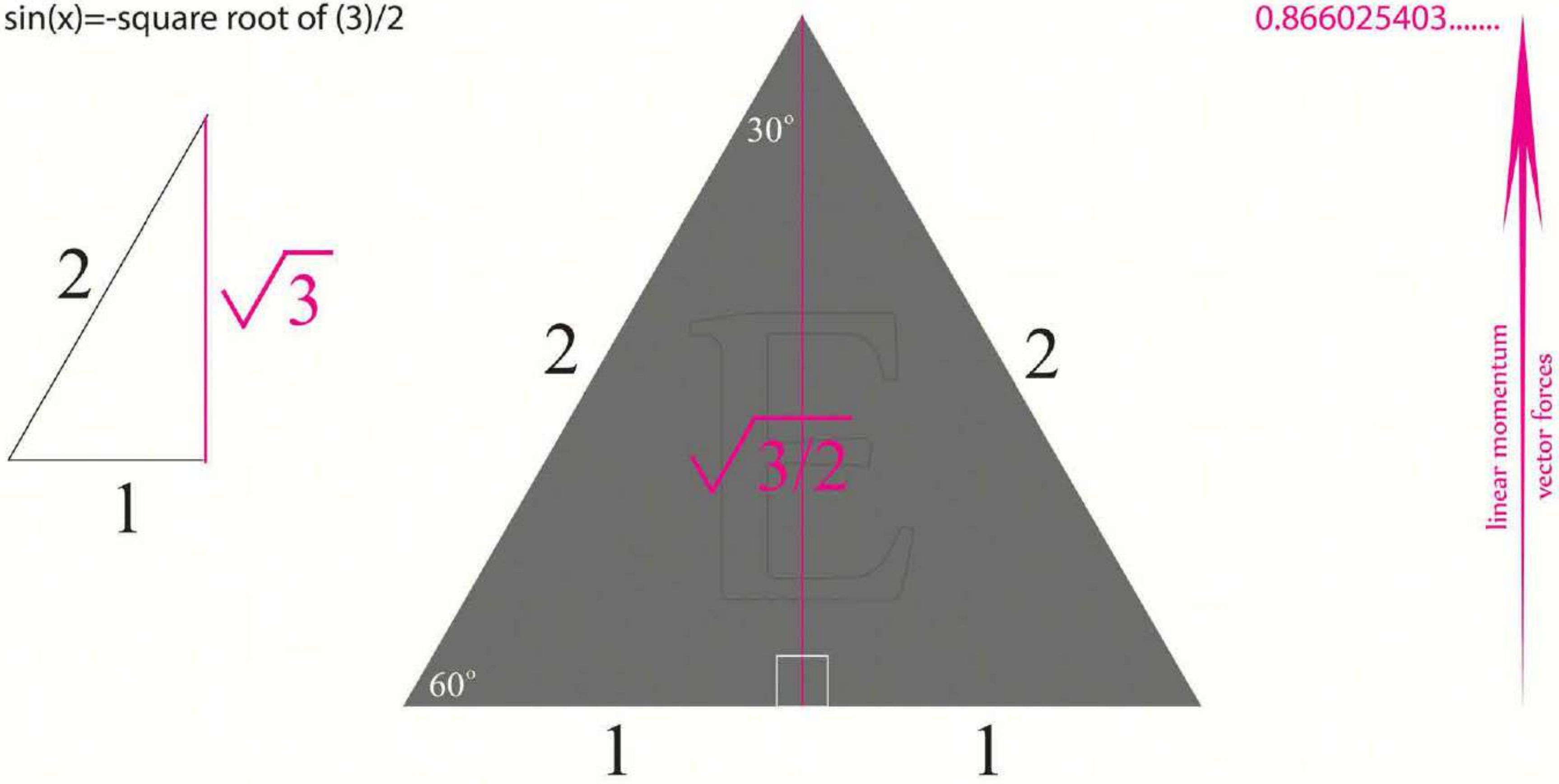
0.866025403.......
2
$\sqrt{3}$
1
30°
2
2
$\sqrt{3/2}$
60°
1
1
linear momentum
vector forces

Pythagφras' theoreom
and irratiφnal numbers expressed in terms of right-angled triangles in Physics
offer a 'half truth' regarding the equilateral geometry of Energy

Leibniz

Newton

Newton focused his work on linear momentum which he developed into his famous laws of motion

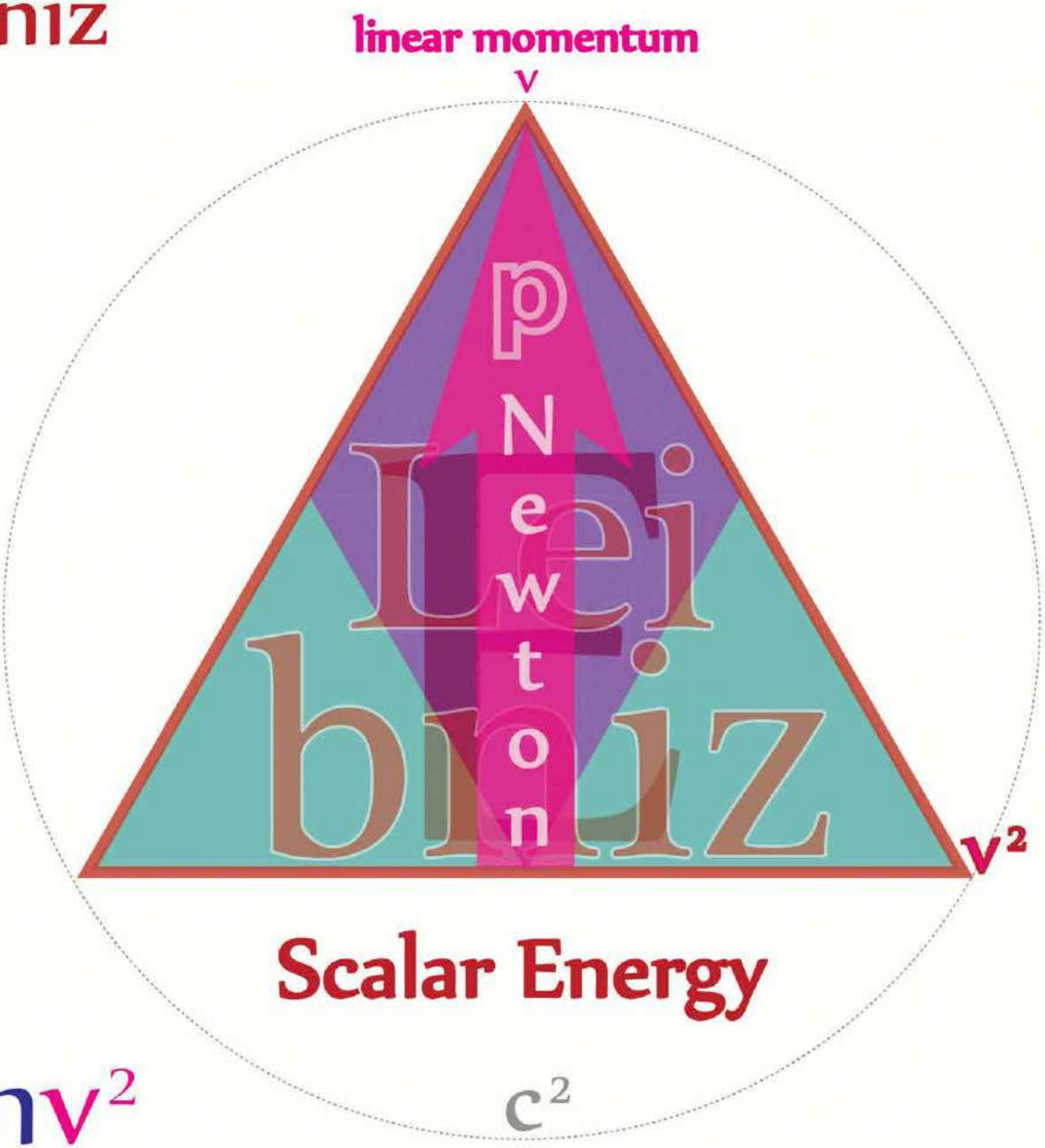

Leibniz introduced the concept of vis viva (living force) which attempted to quantify life force

Newton and Leibniz disagreed about what the world is made of and how its physics shaped our scientific concepts of force, energy, and momentum

Tetryonics reveals the physical relationships they both described mathematically as geometric properties of equilateral Planck energy momenta

$E=mv^2$

$p=mv$

Geometric Square Roots

In geometrical terms, the square root function maps the area of a square to its side length.

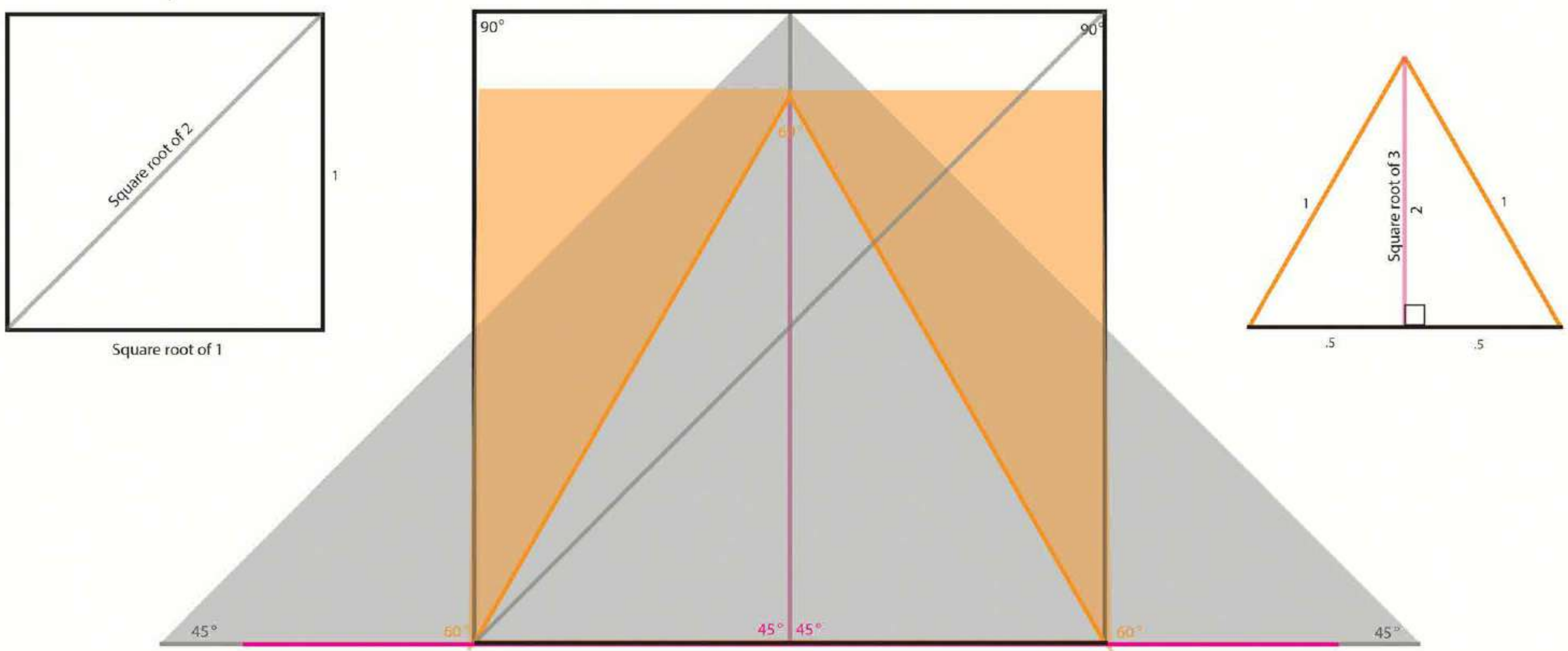

"In mathematics, as in any scientific research, we find two tendencies present. On the one hand, the tendency toward abstraction seeks to crystallize the logical relations inherent in the maze of material that is being studied, and to correlate the material in a systematic and orderly manner. On the other hand, the tendency toward intuitive understanding fosters a more immediate grasp of the objects one studies, a live rapport with them, so to speak, which stresses the concrete meaning of their relations"

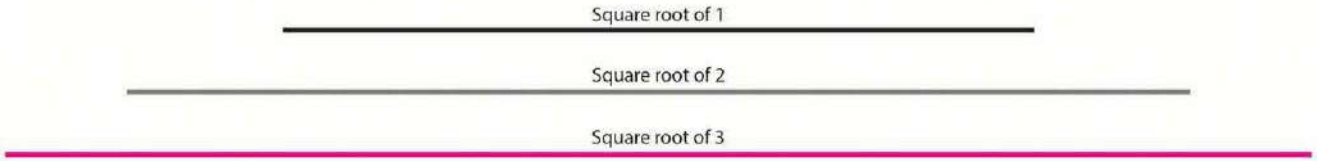

"As to geometry, in particular, the abstract tendency has here led to the magnificent systematic theories of Algebraic Geometry, of Riemannian Geometry, and of Topology; these theories make extensive use of abstract reasoning and symbolic calculation in the sense of algebra.

Notwithstanding this, it is still as true today as it ever was that intuitive understanding plays a major role in geometry. And such concrete intuition is of great value not only for the research worker, but also for anyone who wishes to study and appreciate the results of research in geometry"

David Hilbert [Geometry and the Imagination]

Copyright ABRAHAM [2008] - All rights reserved

Square Roots in Physics

In mathematics, a square root of a number a is a number [n] such that [n]2 = x,
or, in other words, a number [n] whose square (the result of multiplying the number by itself, or [n × n]) is x.

Modern calculators use the
Square Root Algorithm
to calculate the value

It is an approximate numerical value

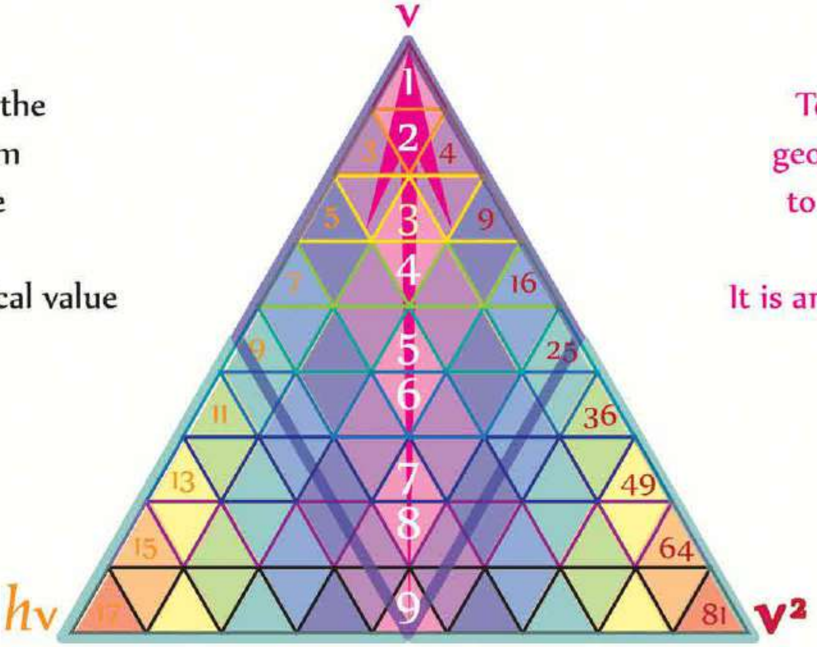
v
1
2
3
4
3
4
5
6
5
9
7
16
7
8
9
25
11
36
13
49
15
64
17
81
9
hv
v²

Tetryonics uses the
geometric Square Root
to calculate the value

It is an exact geometric value

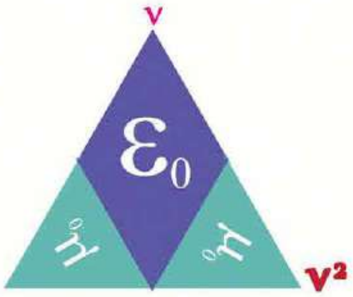
v
ε_0
μ_0
μ_0
v²

$$E = mv^2$$

*In classical geometry,
the square root function maps the
area of a square to its side length.*

$$p = \sqrt{E}$$

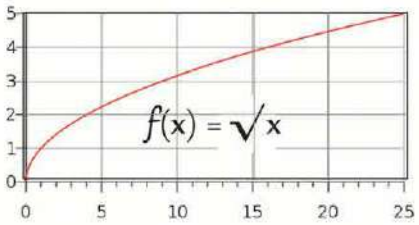

Geometry can easily map irrational numbers

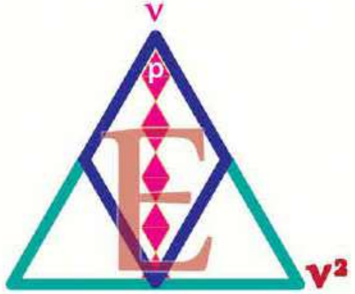

$$p^2 = E$$

*In physics,
the square root function maps
ENERGY [E] to momentum [mv]*

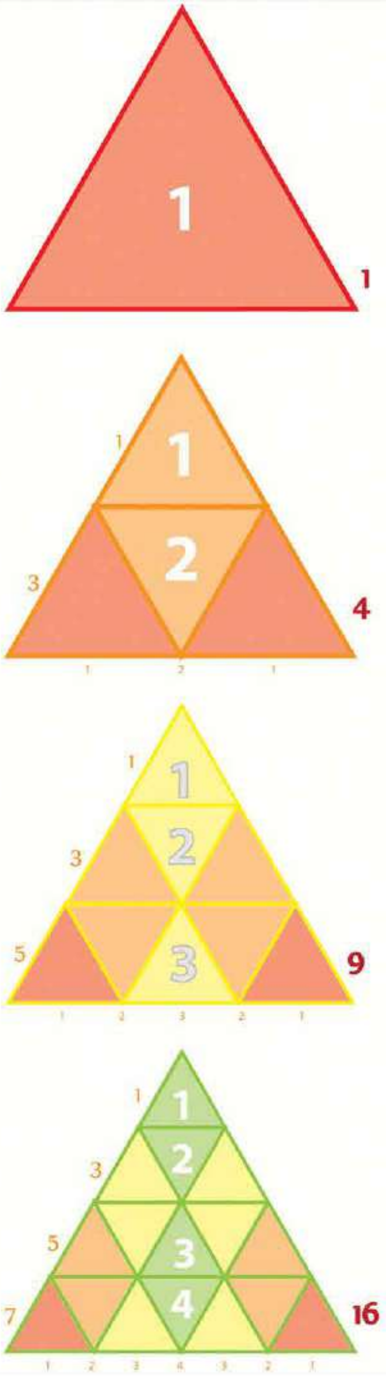
1
1
1
2
1
3
4
1
2
1
1
2
3
1
3
5
9
1
2
3
2
1
1
2
3
4
1
3
5
7
16
1
2
3
4
3
2
1

The Square roots of n

Historically, any number raised to the power of 2 has been modeled using a polygon--the square
That's why we call raising a number to the second power "squaring the number."

note: scalar linear momentum [mv]

is distinct from vector [v]elocity

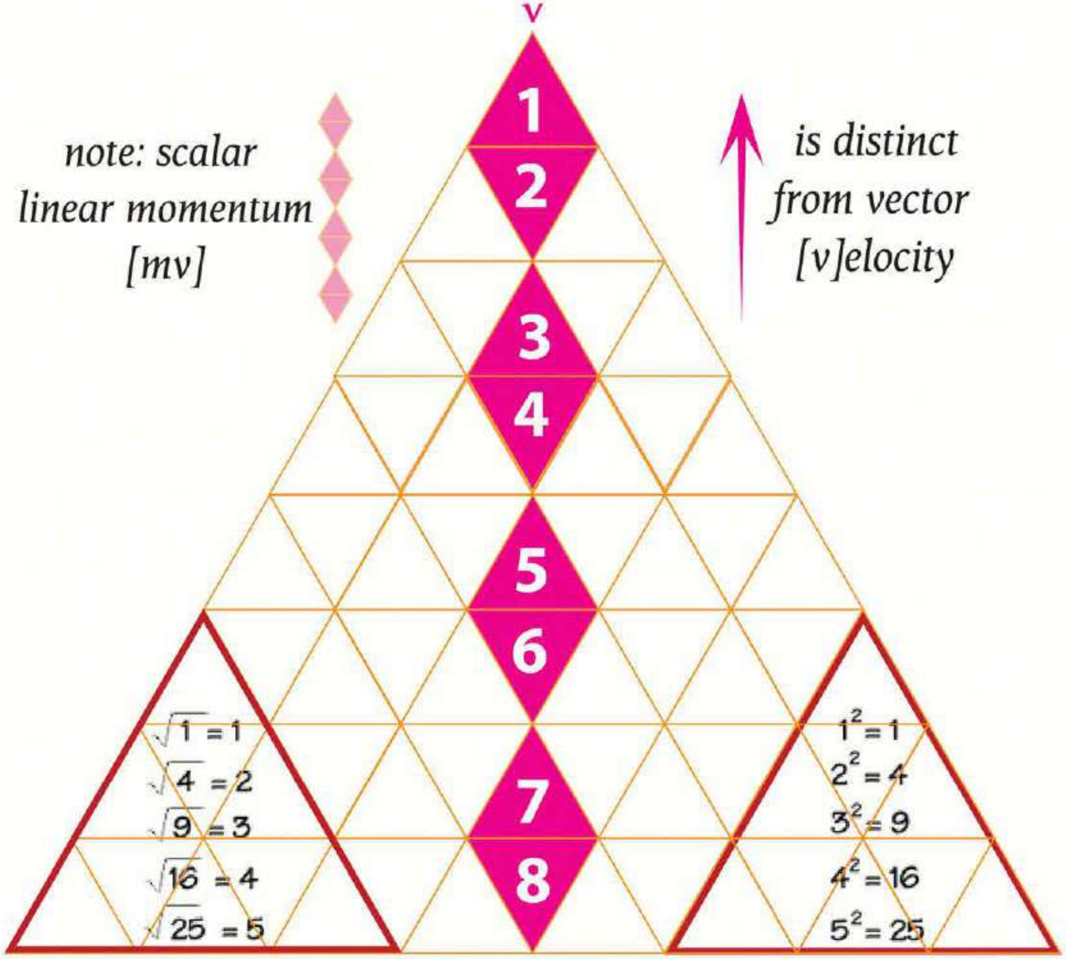

$\sqrt{1} = 1$
$\sqrt{4} = 2$
$\sqrt{9} = 3$
$\sqrt{16} = 4$
$\sqrt{25} = 5$

$1^2 = 1$
$2^2 = 4$
$3^2 = 9$
$4^2 = 16$
$5^2 = 25$

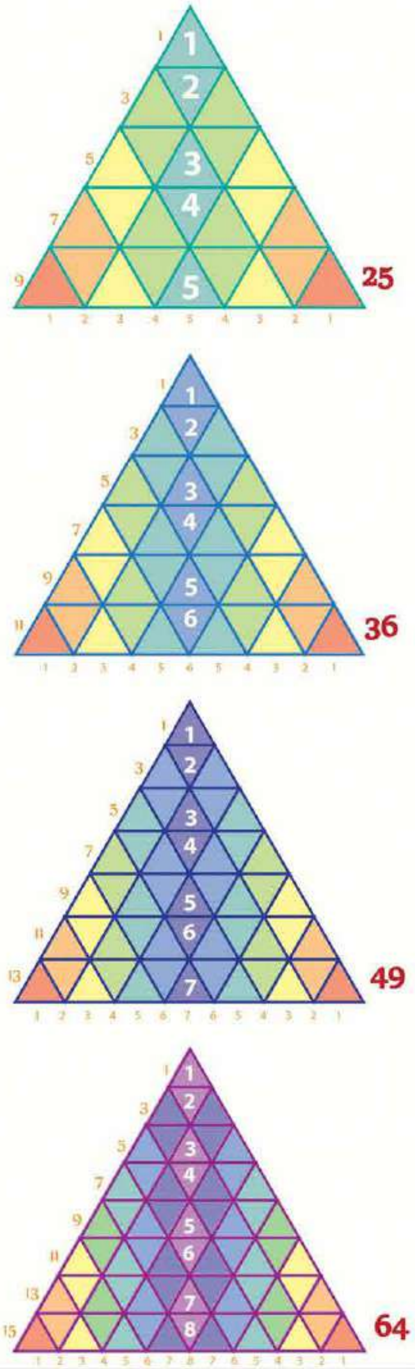
1
2
3
4
5
25
1
2
3
4
5
6
36
1
2
3
4
5
6
7
49
1
2
3
4
5
6
7
8
64

[In physics square numbers are in fact equilateral geometries]
The perfect squares are squares of whole numbers.
Here are the first eight perfect squares

The Square root of Negative 1

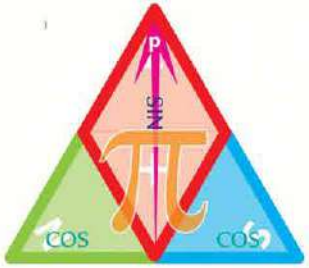

Positive fields are out of phase with Negative fields

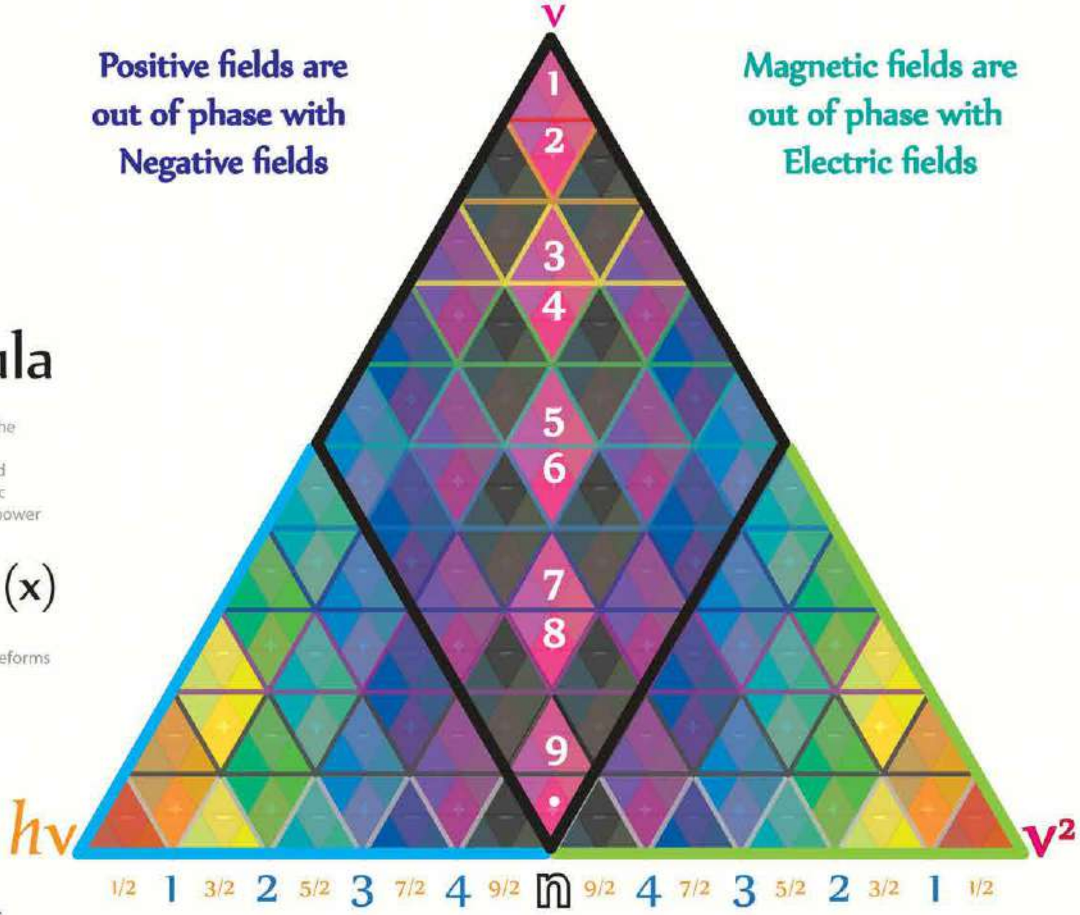

Magnetic fields are out of phase with Electric fields

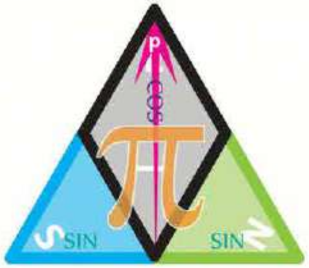

Euler's Formula

Euler's formula is often considered to be the basis of the complex number system. In deriving this formula, Euler established a relationship between the trigonometric functions, sine and cosine, and e raised to a power

$$e^{ix} = \cos(x) + i\sin(x)$$

a mathematical description of EM-Energy waveforms

Leonhard Euler

(15 April 1707 – 18 September 1783)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

$$\sin x + \cos x = 1 + x - \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

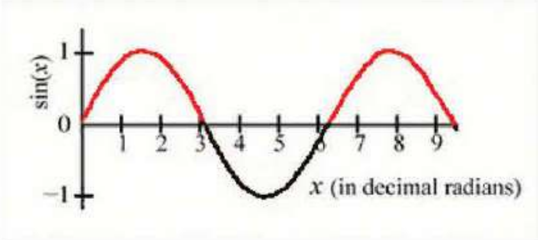

$$e^{i\pi} = -1$$

$$e^{i\pi} + 1 = 0$$

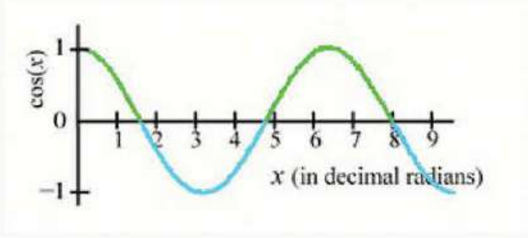

Geometric means

The geometric mean of two numbers, is the square root of their product

geometric square root
of positive one

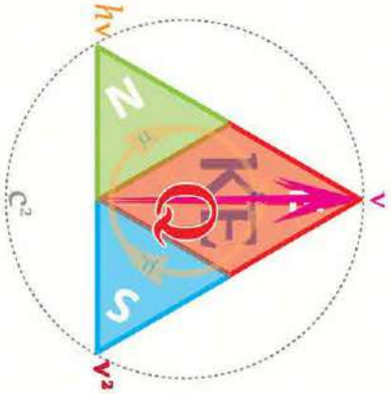

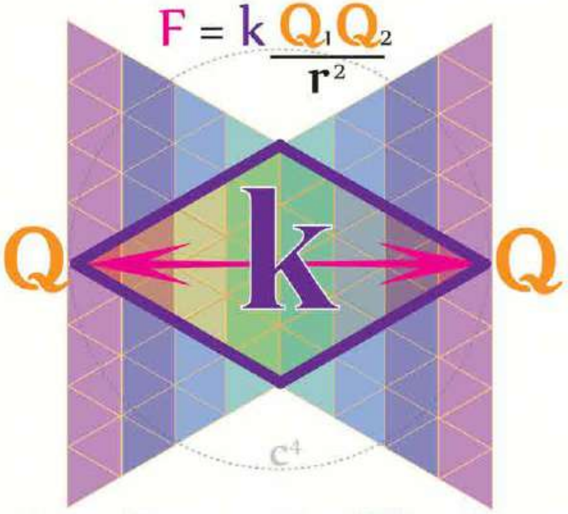

$$F = k\frac{Q_1 Q_2}{r^2}$$

geometric square root
of negative one

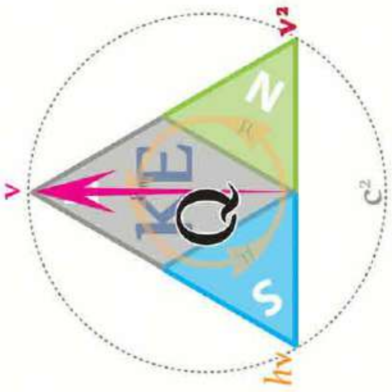

In physics, the geometric mean of two superpositioned fields produces a vector square root Force

It is generally stated that
the geometric mean applies
only to positive numbers.

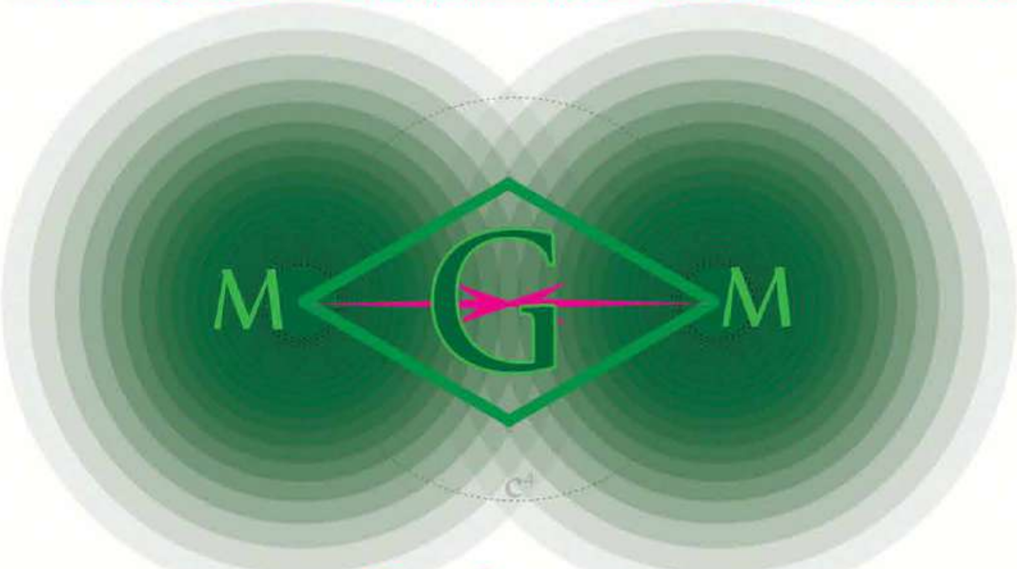

In Tetryonic geometry
the geometric mean applies
to positive & negative fields.

$$F = -G\frac{m_1 m_2}{r^2}$$

Superpositioning

When two or more waves traverse the same space,
the net amplitude at each point is the
sum of the amplitudes of the individual waves.

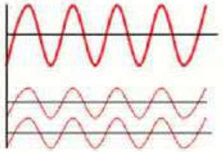

constructive interference

In phase

Additive

+

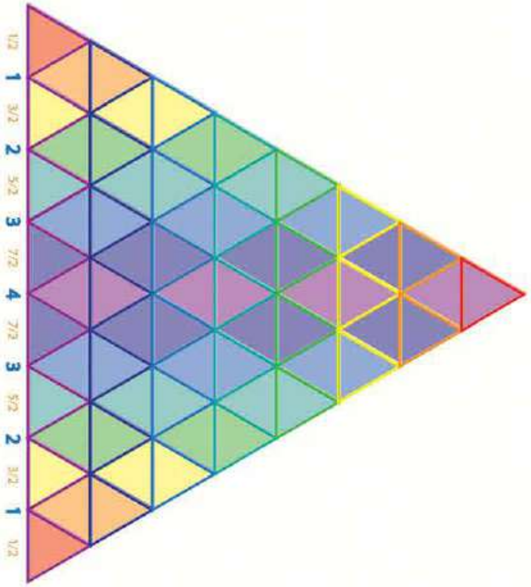
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

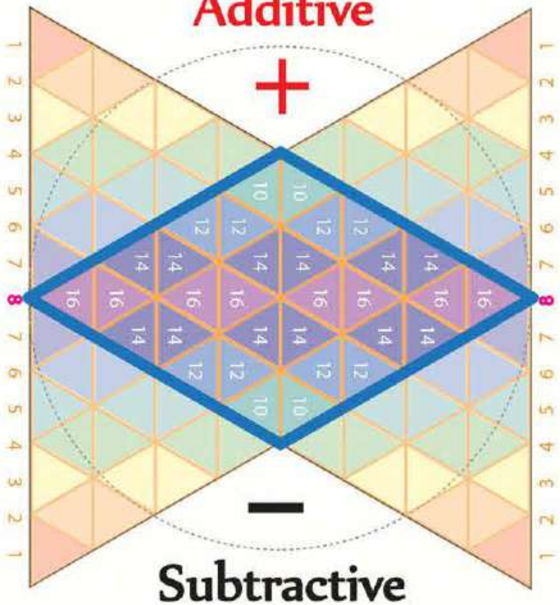
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
10
12
14
16

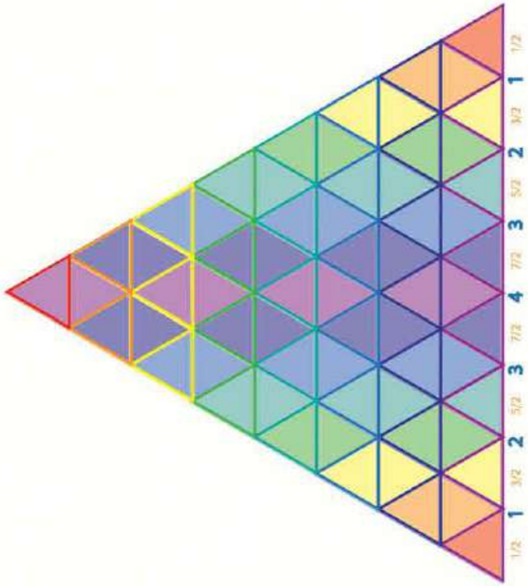
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

−

Subtractive

Out of phase

destructive interference

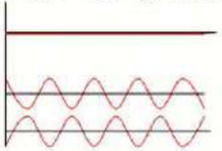

When two or more waves traverse the same space,
if the summed variation has a smaller amplitude
than the constiuent component variations.

The lines of Force

Tetryonic Multiplication table

A multiplication table is a mathematical grid used to define a multiplication operation and its results

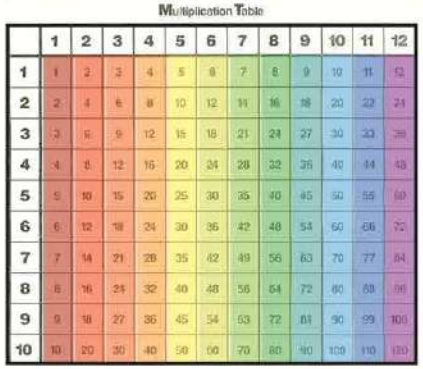

Multiplication Table

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	3	4	5	6	7	8	9	10	11	12
2	2	4	6	8	10	12	14	16	18	20	22	24
3	3	6	9	12	15	18	21	24	27	30	33	36
4	4	8	12	16	20	24	28	32	36	40	44	48
5	5	10	15	20	25	30	35	40	45	50	55	60
6	6	12	18	24	30	36	42	48	54	60	66	72
7	7	14	21	28	35	42	49	56	63	70	77	84
8	8	16	24	32	40	48	56	64	72	80	88	96
9	9	18	27	36	45	54	63	72	81	90	99	108
10	10	20	30	40	50	60	70	80	90	100	110	120

Historically Multiplication tables
have been based on Square geometries

In Tetryonics multiplication tables can
also be based on EQUILATERAL geometries

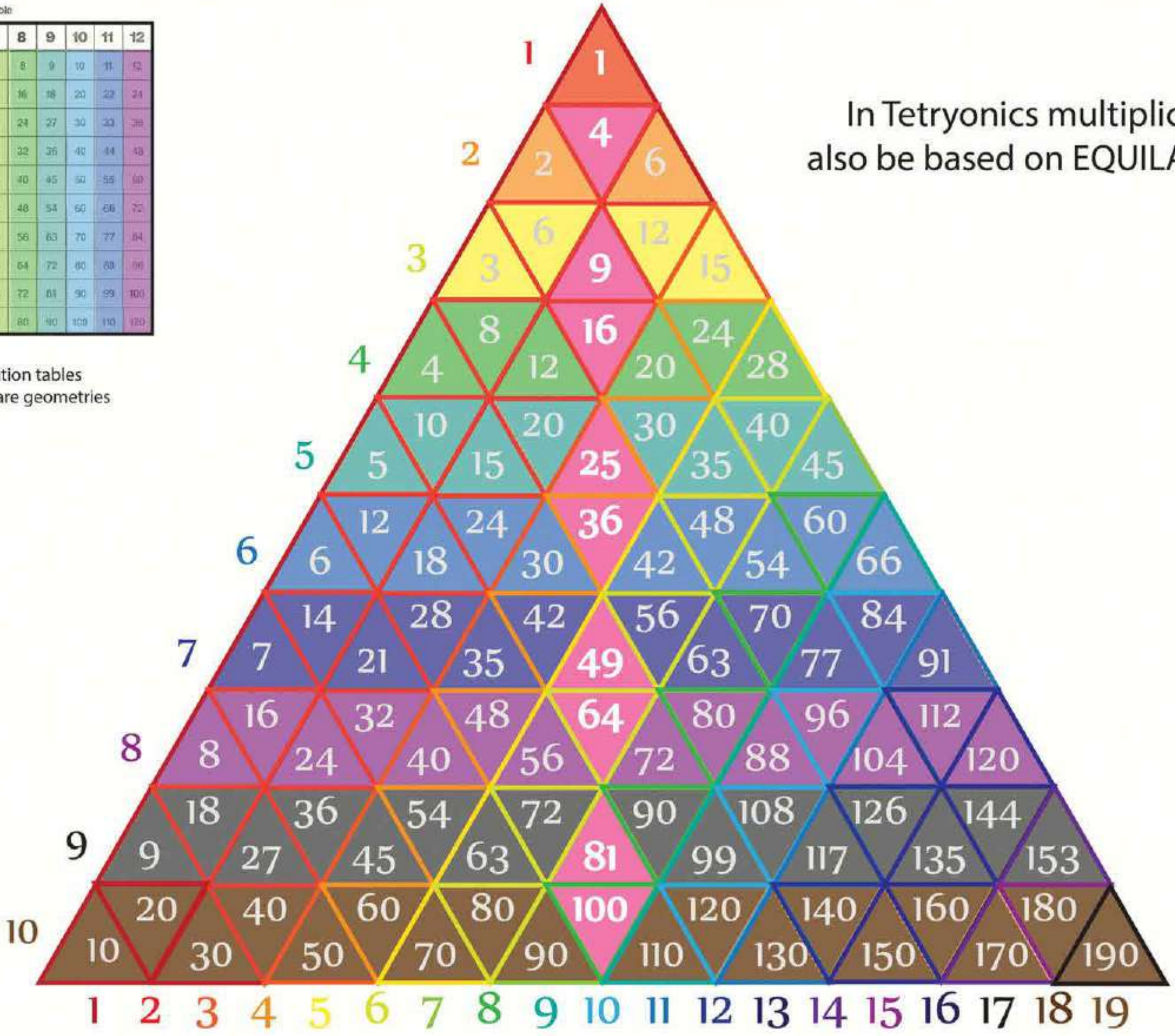
1 1
2 2 4 6
3 3 6 9 12 15
4 4 8 12 16 20 24 28
5 5 10 15 20 25 30 35 40 45
6 6 12 18 24 30 36 42 48 54 60 66
7 7 14 21 28 35 42 49 56 63 70 77 84 91
8 8 16 24 32 40 48 56 64 72 80 88 96 104 112 120
9 9 18 27 36 45 54 63 72 81 90 99 108 117 126 135 144 153
10 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

The integer multiplicitors are colour coded

Copyright ABRAHAM [2008] - All rights reserved

Rhombic Multiplication Tables

1	2	3	4	5	6	7	8	9	10
2	**4**	6	8	10	12	14	16	18	20
3	6	**9**	12	15	18	21	24	27	30
4	8	12	**16**	20	24	28	32	36	40
5	10	15	20	**25**	30	35	40	45	50
6	12	18	24	30	**36**	42	48	54	60
7	14	21	28	35	42	**49**	56	63	70
8	16	24	32	40	48	56	**64**	72	80
9	18	27	36	45	54	63	72	**81**	90
10	20	30	40	50	60	70	80	90	**100**

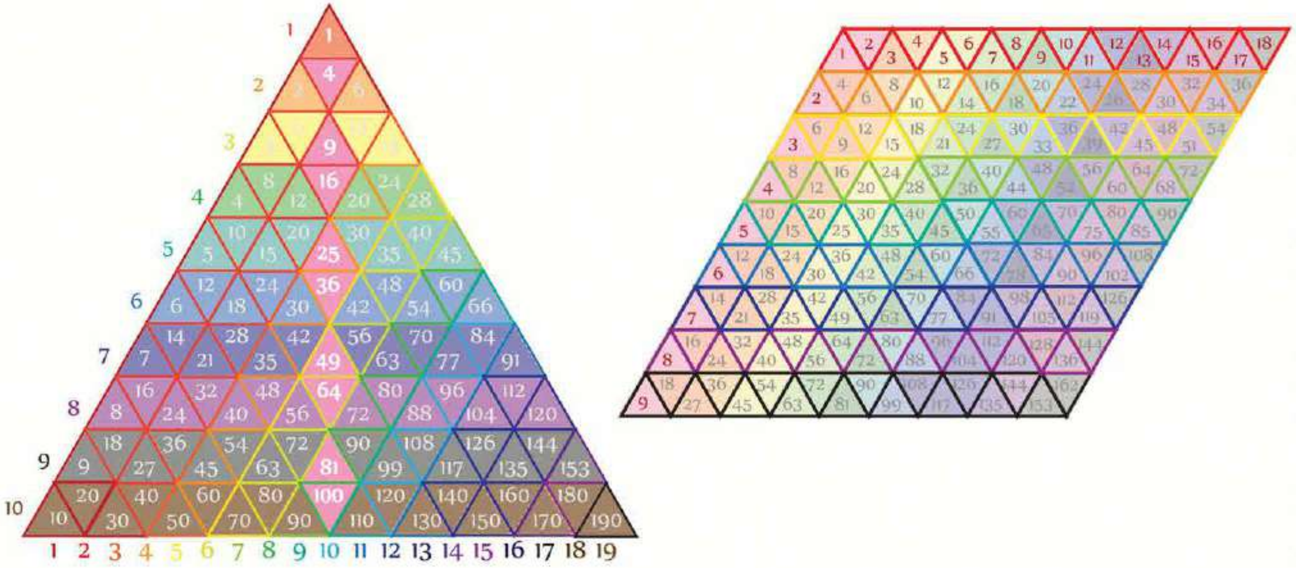

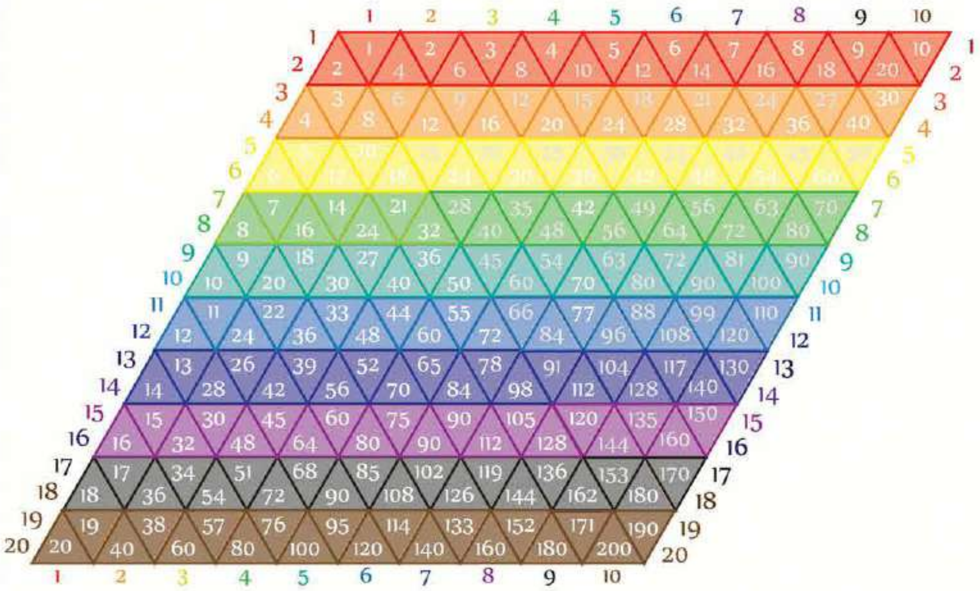

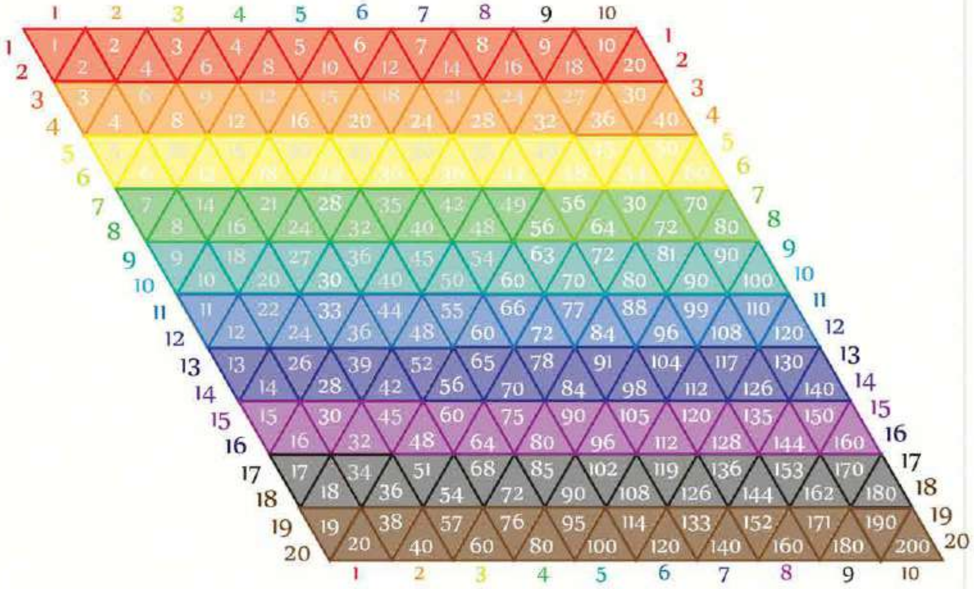

Tetryonics 85.03 - Rhombic multiplication tables

Copyright ABRAHAM [2008] - All rights reserved

Photonic Root Tables

Tetryonic multiplication table can take a number of geometric forms

Square root median

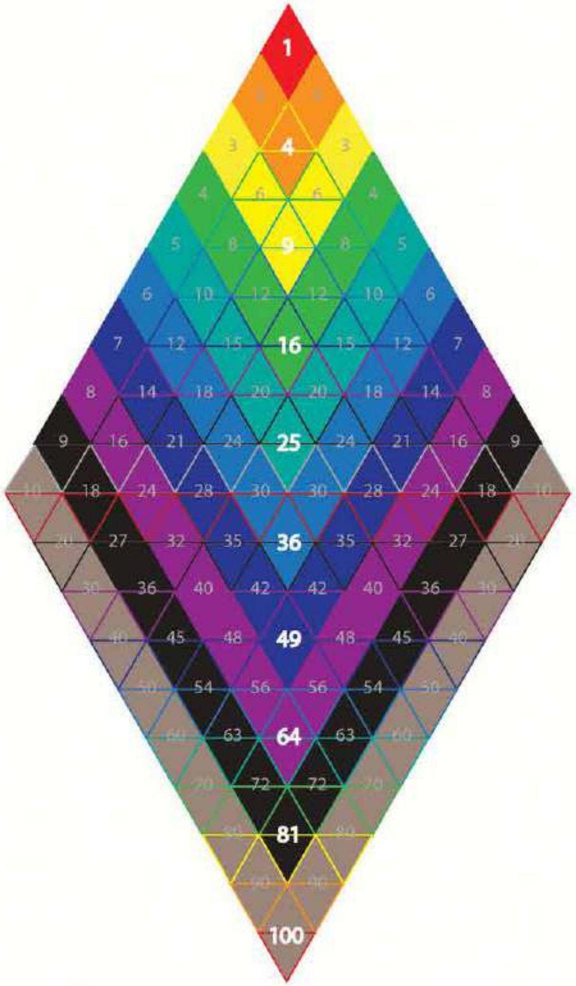

Table read diagonally

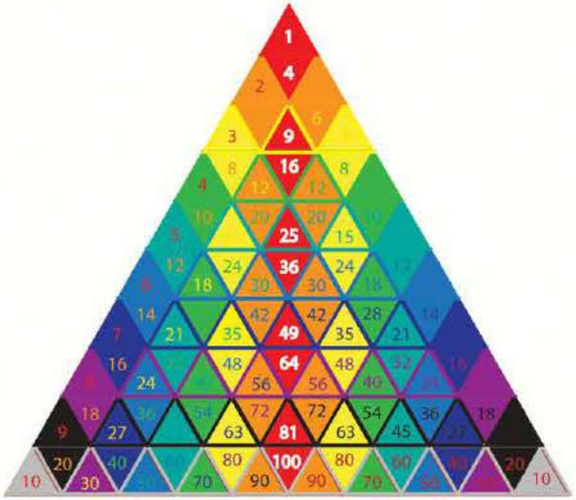

1	2	3	4	5	6	7	8	9	10
2	**4**	6	8	10	12	14	16	18	20
3	6	**9**	12	15	18	21	24	27	30
4	8	12	**16**	20	24	28	32	36	40
5	10	15	20	**25**	30	35	40	45	50
6	12	18	24	30	**36**	42	48	54	60
7	14	21	28	35	42	**49**	56	63	70
8	16	24	32	40	48	56	**64**	72	80
9	18	27	36	45	54	63	72	**81**	90
10	20	30	40	50	60	70	80	90	**100**

Integer median

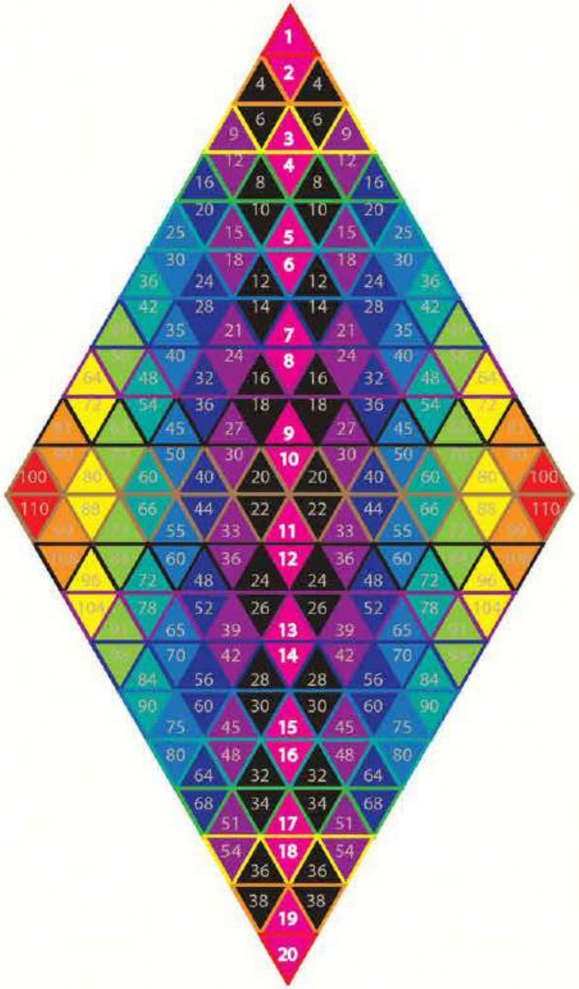

Table read from centre
to outside edge then down

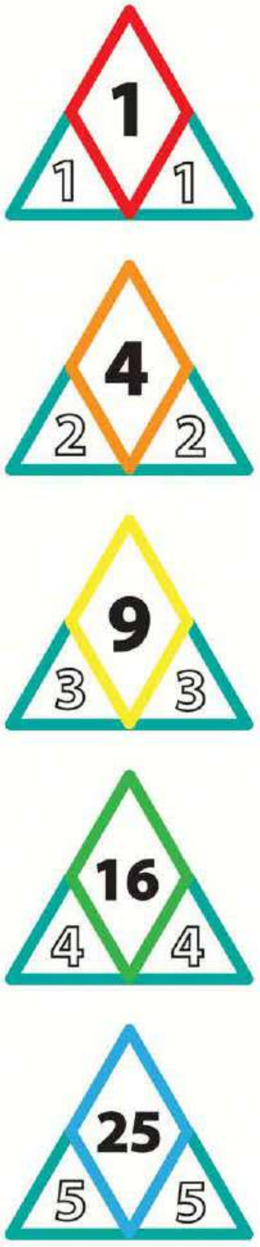
1
1
1
4
2
2
9
3
3
16
4
4
25
5
5

Divide [division] tables

Minus [subtraction] tables

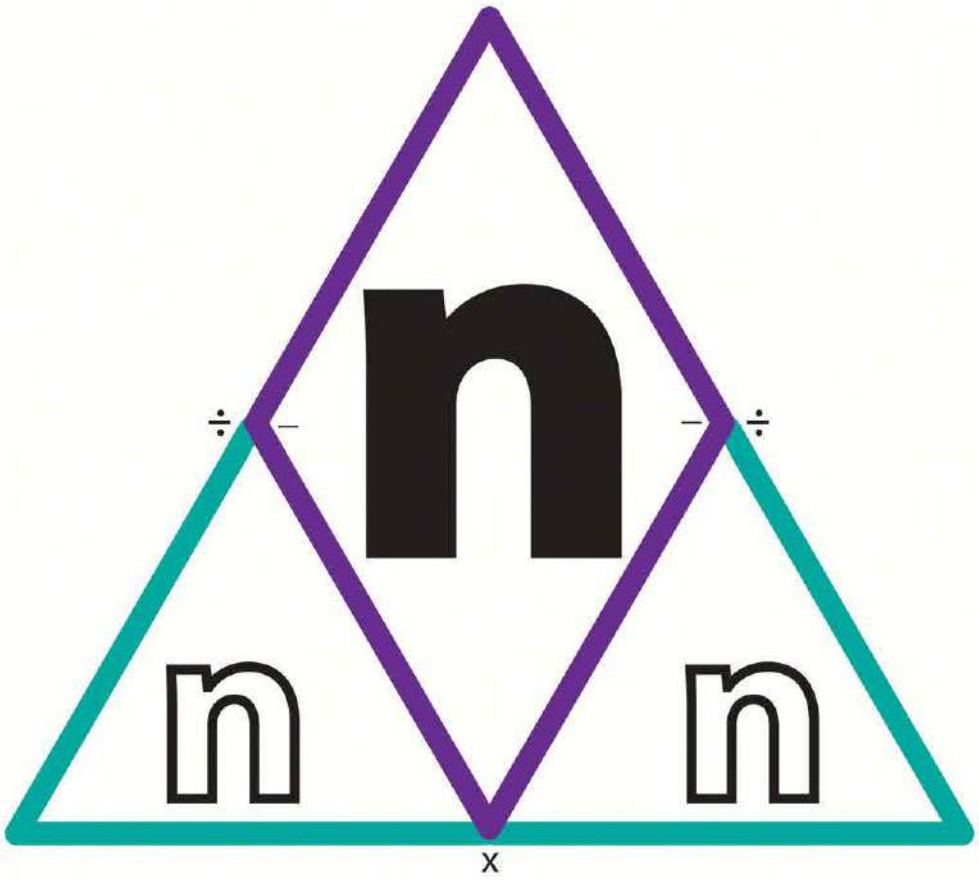

Times [multiplication] tables

Plus [addition] tables

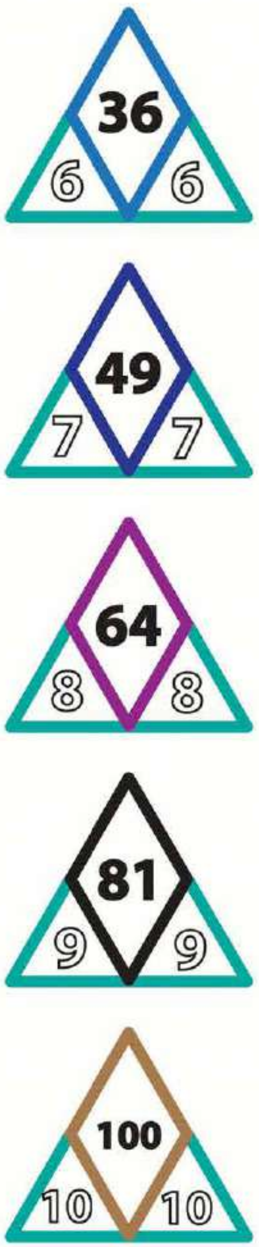
36
6
6
49
7
7
64
8
8
81
9
9
100
10
10

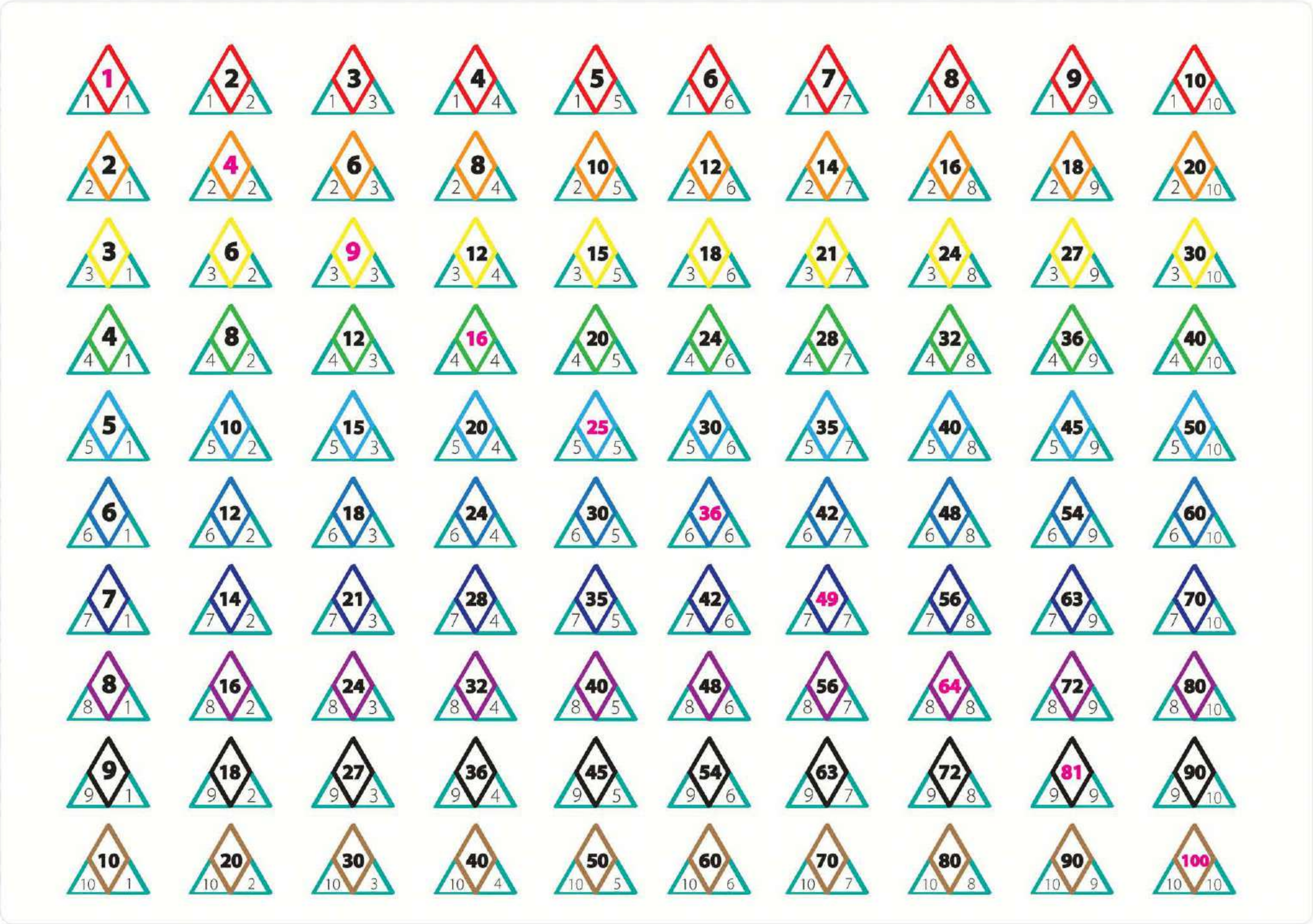

1 1 1	2 1 2	3 1 3	4 1 4	5 1 5	6 1 6	7 1 7	8 1 8	9 1 9	10 1 10
2 2 1	4 2 2	6 2 3	8 2 4	10 2 5	12 2 6	14 2 7	16 2 8	18 2 9	20 2 10
3 3 1	6 3 2	9 3 3	12 3 4	15 3 5	18 3 6	21 3 7	24 3 8	27 3 9	30 3 10
4 4 1	8 4 2	12 4 3	16 4 4	20 4 5	24 4 6	28 4 7	32 4 8	36 4 9	40 4 10
5 5 1	10 5 2	15 5 3	20 5 4	25 5 5	30 5 6	35 5 7	40 5 8	45 5 9	50 5 10
6 6 1	12 6 2	18 6 3	24 6 4	30 6 5	36 6 6	42 6 7	48 6 8	54 6 9	60 6 10
7 7 1	14 7 2	21 7 3	28 7 4	35 7 5	42 7 6	49 7 7	56 7 8	63 7 9	70 7 10
8 8 1	16 8 2	24 8 3	32 8 4	40 8 5	48 8 6	56 8 7	64 8 8	72 8 9	80 8 10
9 9 1	18 9 2	27 9 3	36 9 4	45 9 5	54 9 6	63 9 7	72 9 8	81 9 9	90 9 10
10 10 1	20 10 2	30 10 3	40 10 4	50 10 5	60 10 6	70 10 7	80 10 8	90 10 9	100 10 10

Copyright ABRAHAM [2008] - All rights reserved

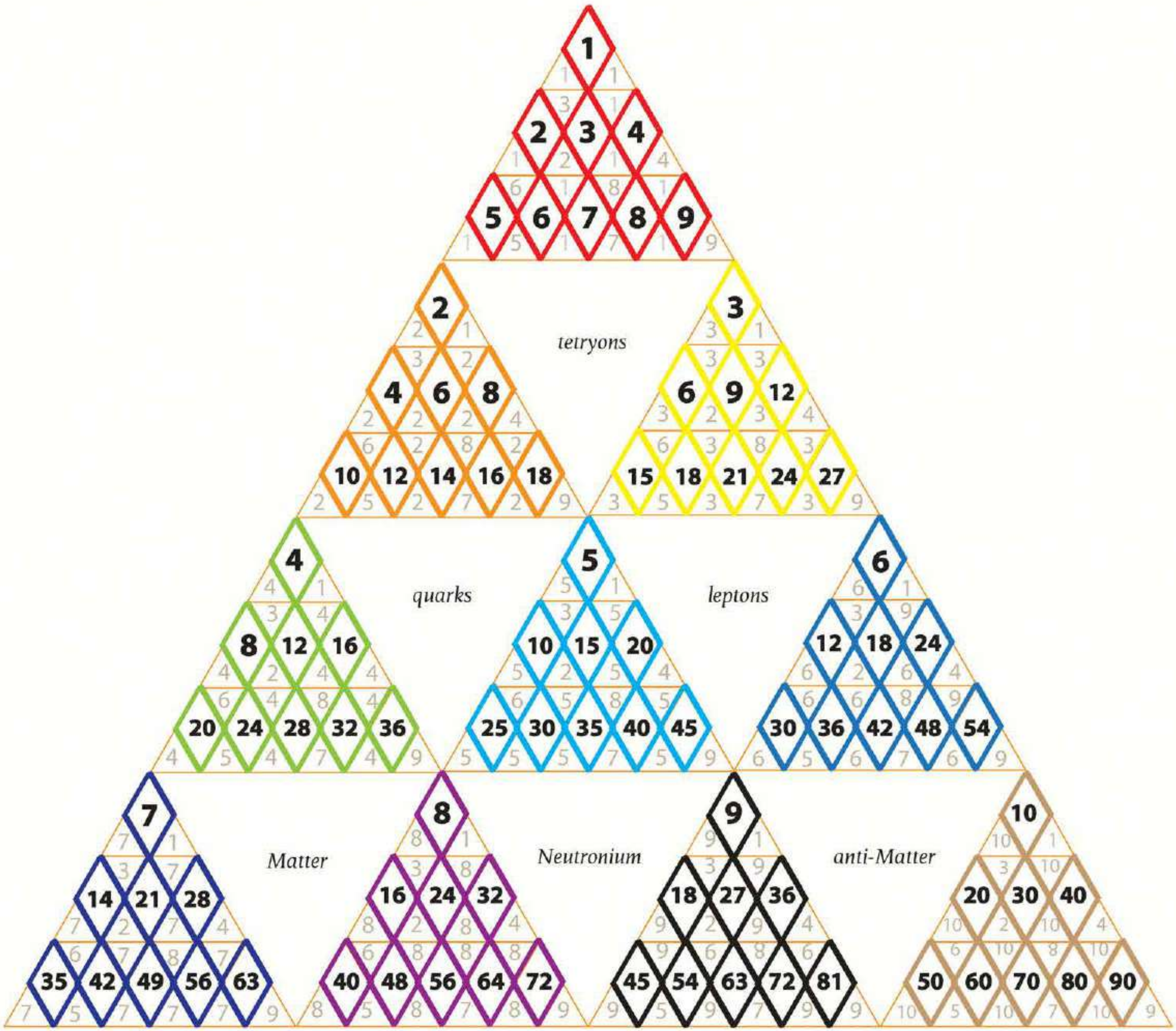
1
2 3 4
5 6 7 8 9
tetryons
2
4 6 8
10 12 14 16 18
3
6 9 12
15 18 21 24 27
quarks
4
8 12 16
20 24 28 32 36
5
10 15 20
25 30 35 40 45
leptons
6
12 18 24
30 36 42 48 54
Matter
7
14 21 28
35 42 49 56 63
Neutronium
8
16 24 32
40 48 56 64 72
9
18 27 36
45 54 63 72 81
anti-Matter
10
20 30 40
50 60 70 80 90

Tetryonics 85.07 - Tetrad Math Tables

Irrational Numbers

An irrational number is defined to be any number that is the part of the real number system that cannot be written as a complete ratio of two integers

One well known irrational number is π

Irrational numbers can be easily represented geometrically

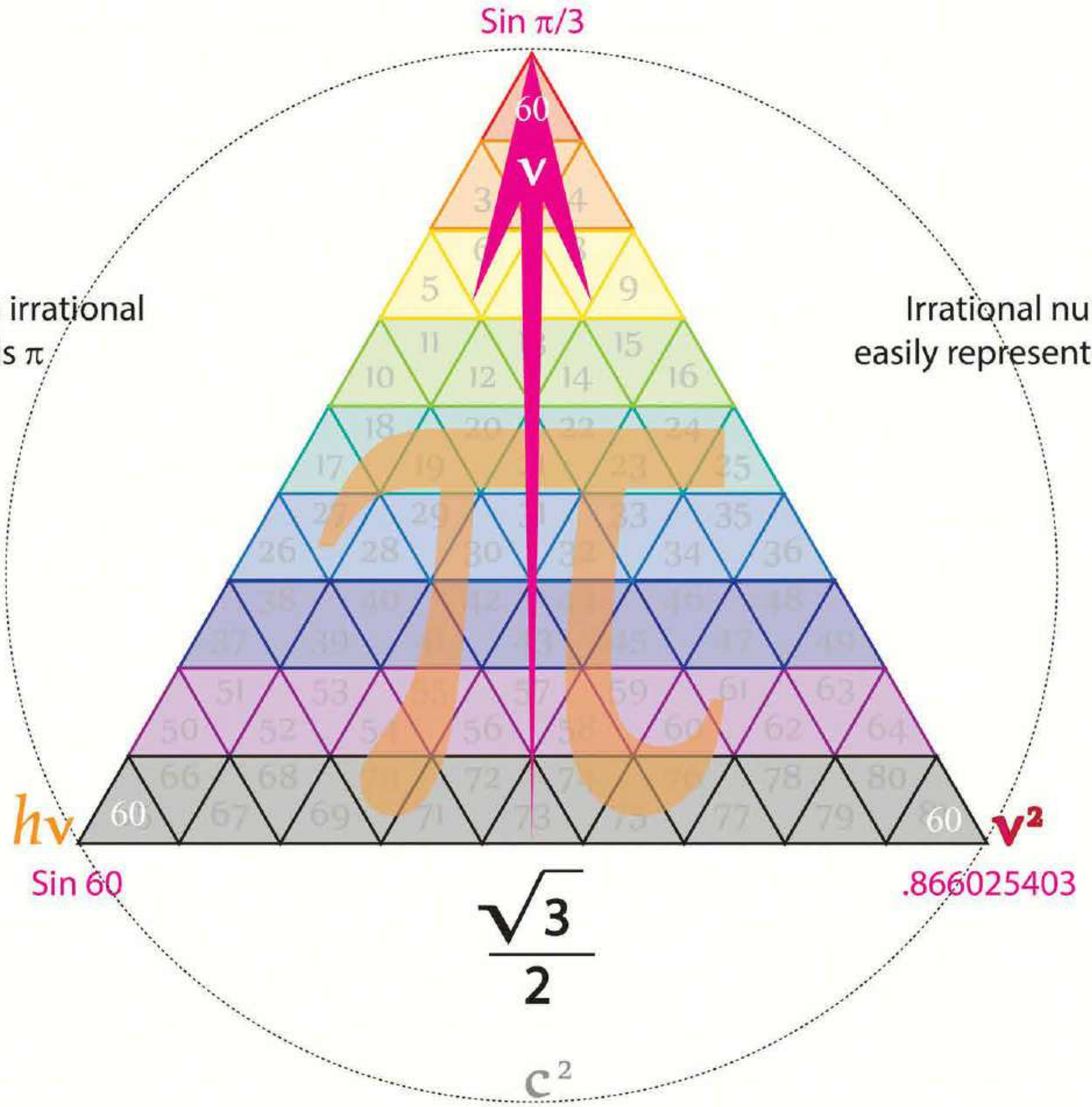
Sin π/3
60
v
π
hv
60
60
v²
Sin 60
.866025403
$\frac{\sqrt{3}}{2}$
c²

Exponentials & Logarithms

e and the Natural Log are inverse functions of each other:

e^x is the amount of continuous GROWTH after a certain amount of time.
Natural Log (ln) is the amount of TIME of continuous growth to reach a certain level

1
10
100
1,000

0
1
2
3

$10,000 = 10^4$

$\log(10,000) = 4$

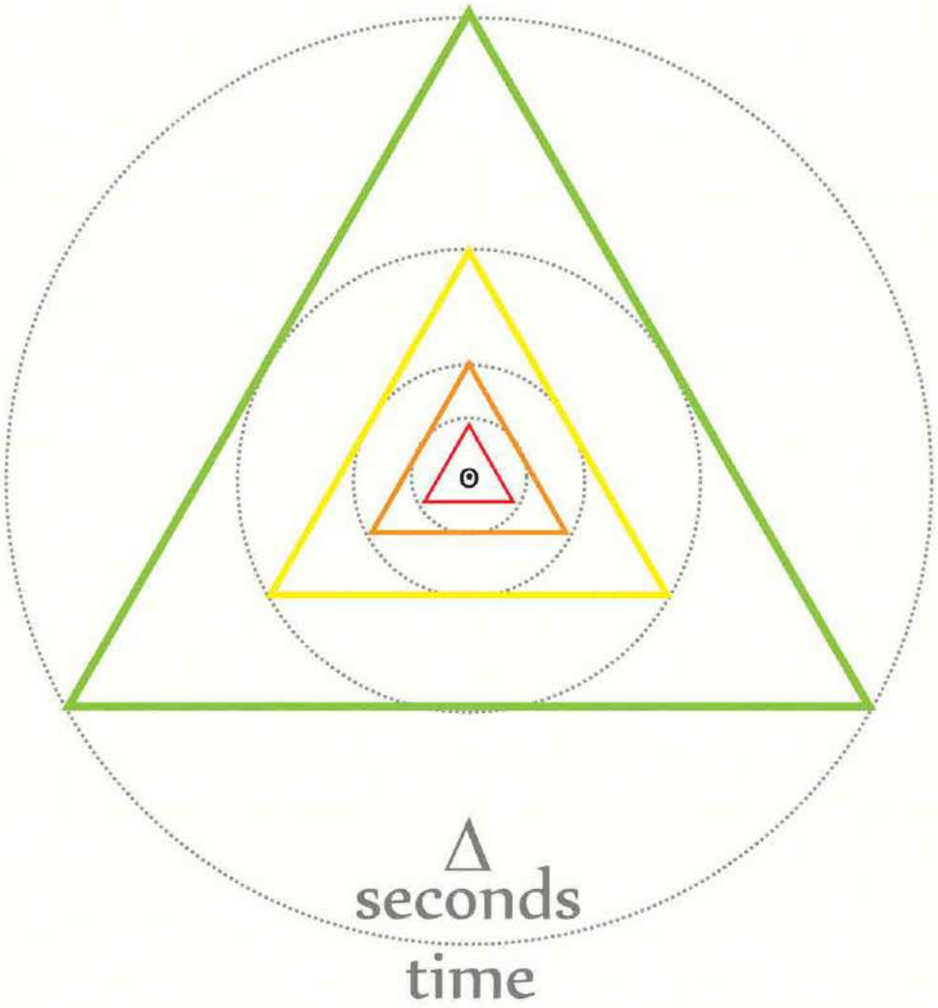

e

How much growth after x units of time
(and 100% continuous growth)

$$e^x = \lim_{n\to\infty}(1+x/n)^n,$$

GROWTH

ln

ln(x) lets us plug in continuous growth
and get the time it would take.

$$\ln(x) = \lim_{n\to\infty} n(x^{1/n}-1).$$

PERIOD

Copyright ABRAHAM [2008] - All rights reserved

Exponential growth

$$\text{growth} = e = \lim_{n\to\infty}\left(1 + \frac{r}{n}\right)^t$$

e represents the idea that all continually growing systems are scaled versions of a common rate

GEOMETRIC

3.141592654

π

Pi is the ratio between circumference and diameter shared by all circles.

It is a fundamental ratio inherent in all circles and therefore impacts any calculation of circumference, area, volume, and surface areas

Pi radians are equally important and show all quantised equilateral energy geometries are related to their scribed circles

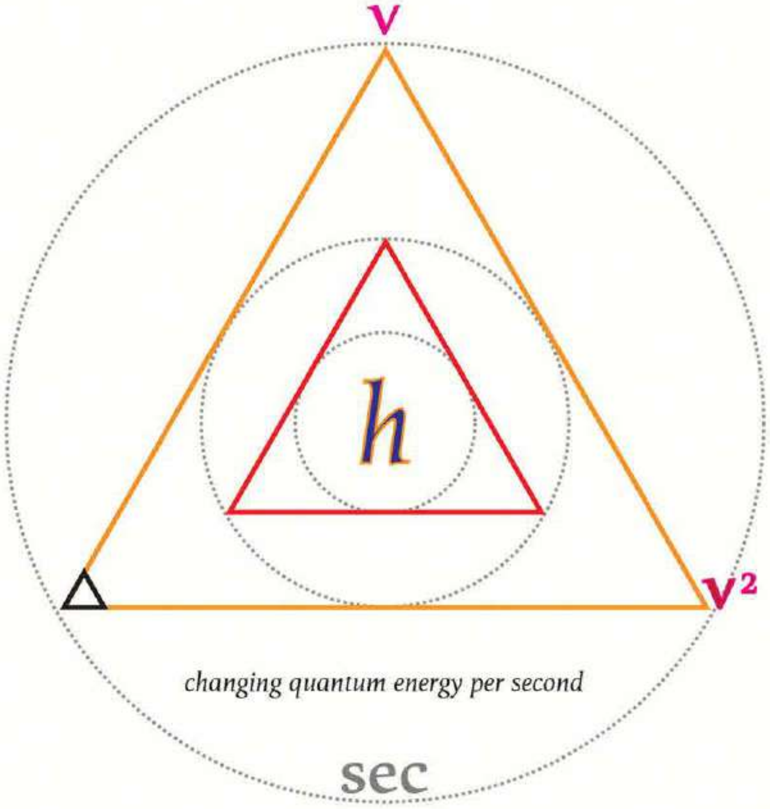

GROWTH

2.718281828

e

e is the base rate of growth shared by all continually growing processes.

e lets you take a simple growth rate (where all the change happens all at once at the end of a period of time - ie quantised growth)

e shows up whenever systems grow exponentially and continuously..... radioactive decay, interest calculations and populations

tetryons

EM fields

quarks

leptons

e can be applied to the equilateral energy geometries of physical systems only where the rate of increase is a integer factor of a squared number

Baryons

Nuclear Energy levels

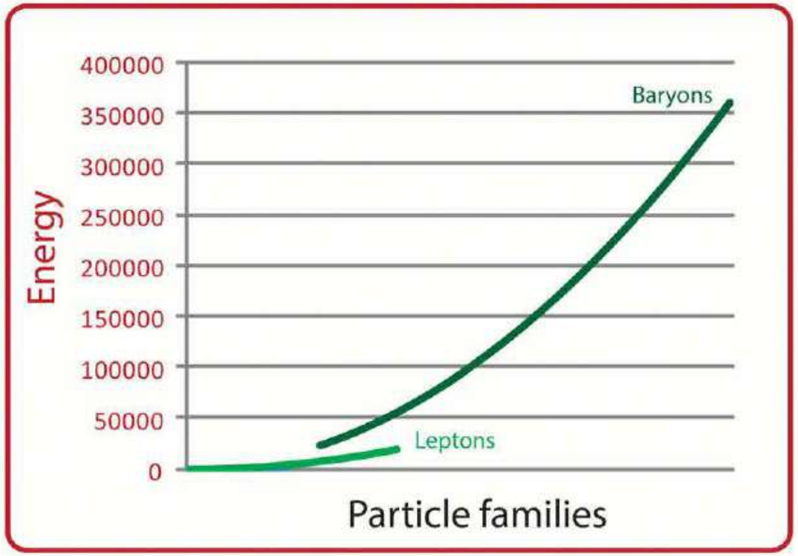

The emission and absorption of bosons and Photons within sub-atomic nuclei

Increase and decrease in integer amounts according to the charged Tetryonic topologies of the particles involved

Exponential energy levels

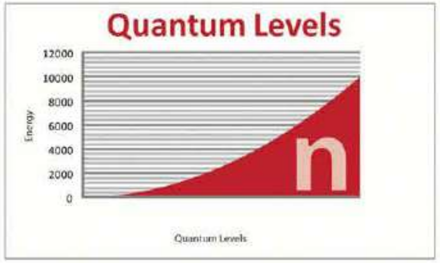

Radioactive decays follow exponential curves determined by the Tetryonic topology of the sub-atomic particle families

Proton
Neutron

antiNeutron
antiProton

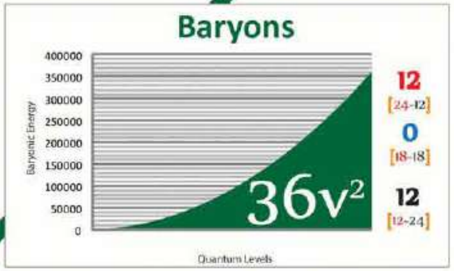

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

12
[12-0]

Electron
Positron

Neutrino

Up
Down
Strange
Charmed
Top
Bottom

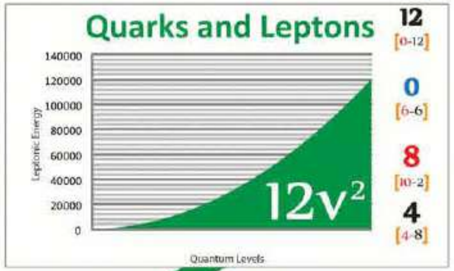

quarks

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

leptons ElectroMagnetic mass velocity

Positive
Negative
Neutral

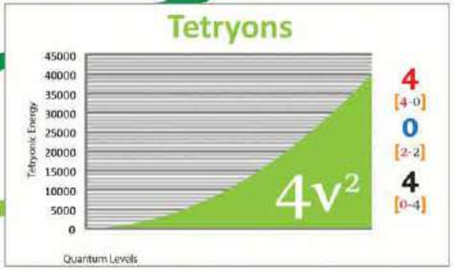

EM Field Planck quanta

$$4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

tetryons ElectroMagnetic mass velocity

Series addition & the Riemann Zeta Function

The second series addition of the Reimann Zeta function is where x=2: (pi^2)/6=1+1/2^2+1/3^2+1/4^2+...
(the sum of the reciprocals of the squares)

$$\zeta(s) = \sum_{n} \frac{1}{n^s}$$

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}$$

$$f(n) = 1/n^2$$

$$\sum_{n=1}^{\infty} \frac{1}{n^x}$$

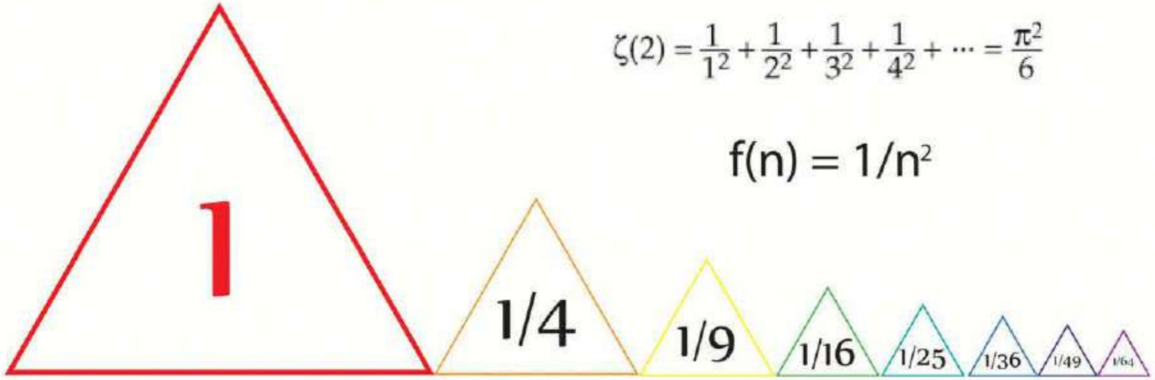

In mathematics, the Riemann zeta function, is a prominent function of great significance in number theory. It is a named after German mathematician Bernhard Riemann. It is so important because of its relation to the distribution of prime numbers. It also has applications in other areas such as physics, probability theory, and applied statistics

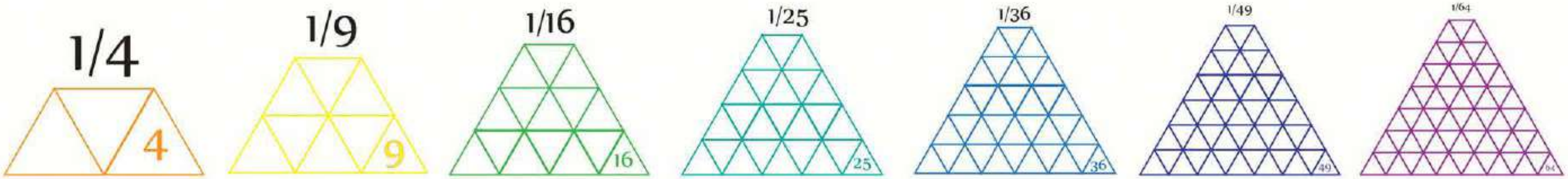

The mystery of prime numbers
Question: which natural numbers are prime? how are they distributed among natural numbers?

Primes are basic building blocks for natural numbers:

- any natural number is a product of prime numbers
- a prime number is only divisible by itself and by 1:
  (it cannot be further simplified)

We don't know how to predict where the prime numbers are:

"Prime numbers grow like weeds among the natural numbers, seeming to obey no other law than that of chance but also exhibit stunning regularity" (Don Zagier, number theorist)

Adding the odd numbers in order produces the square numbers

Apart from 2, all primes are odd numbers;
the difference between two consecutive squares being odd,
every prime can be expressed as the difference between two squares

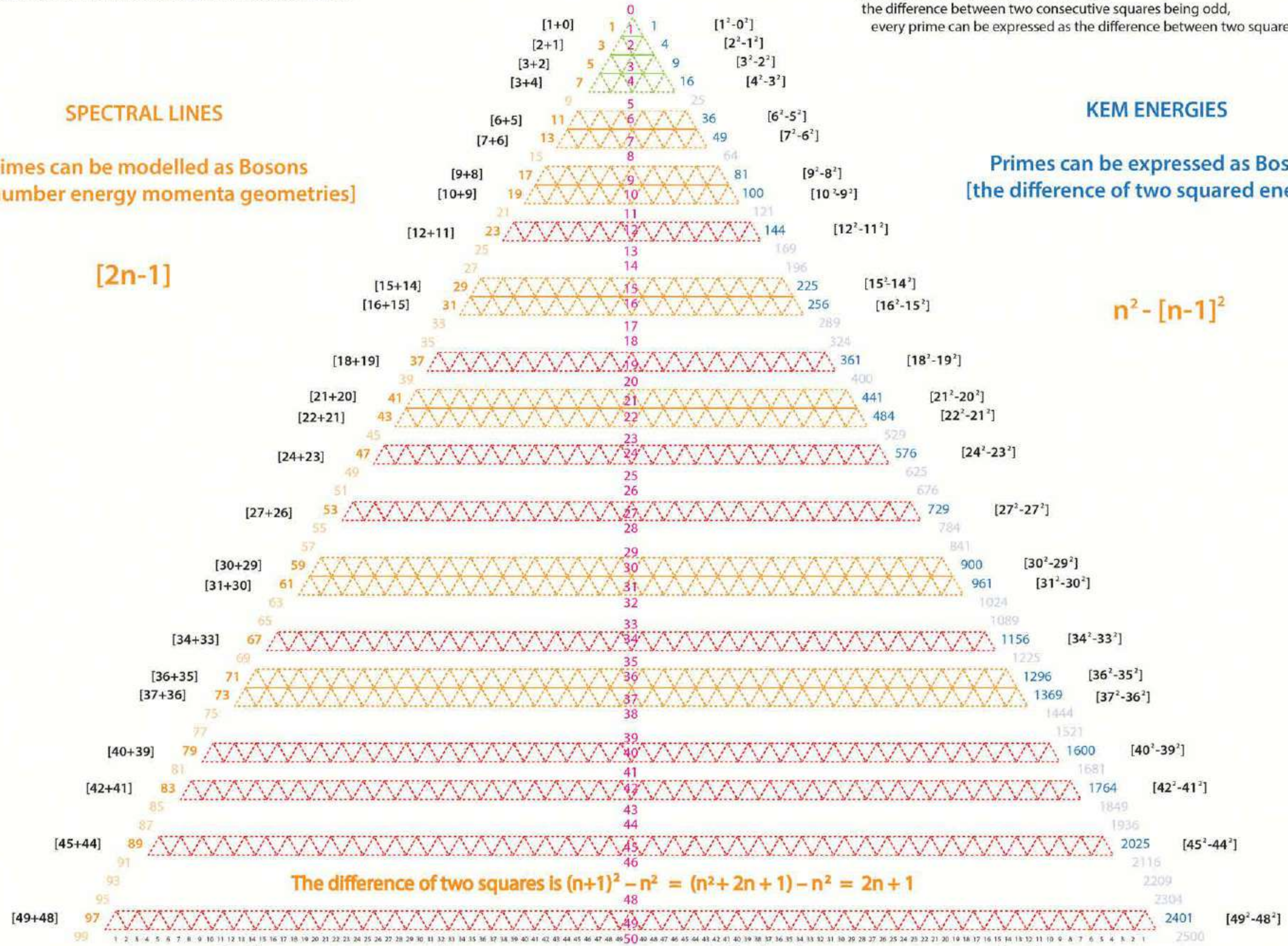
[1+0] 1 1 [1^2-0^2]
[2+1] 3 4 [2^2-1^2]
[3+2] 5 9 [3^2-2^2]
[3+4] 7 16 [4^2-3^2]
[6+5] 11 36 [6^2-5^2]
[7+6] 13 49 [7^2-6^2]
[9+8] 17 81 [9^2-8^2]
[10+9] 19 100 [10^2-9^2]
[12+11] 23 144 [12^2-11^2]
[15+14] 29 225 [15^2-14^2]
[16+15] 31 256 [16^2-15^2]
[18+19] 37 361 [18^2-19^2]
[21+20] 41 441 [21^2-20^2]
[22+21] 43 484 [22^2-21^2]
[24+23] 47 576 [24^2-23^2]
[27+26] 53 729 [27^2-27^2]
[30+29] 59 900 [30^2-29^2]
[31+30] 61 961 [31^2-30^2]
[34+33] 67 1156 [34^2-33^2]
[36+35] 71 1296 [36^2-35^2]
[37+36] 73 1369 [37^2-36^2]
[40+39] 79 1600 [40^2-39^2]
[42+41] 83 1764 [42^2-41^2]
[45+44] 89 2025 [45^2-44^2]
[49+48] 97 2401 [49^2-48^2]
99 2500

SPECTRAL LINES

Primes can be modelled as Bosons [ODD number energy momenta geometries]

[2n-1]

KEM ENERGIES

Primes can be expressed as Bosons [the difference of two squared energies]

$$n^2 - [n-1]^2$$

The difference of two squares is $(n+1)^2 - n^2 = (n^2 + 2n + 1) - n^2 = 2n + 1$

[n-1] **all energy quanta create normal distributions** [n-1]

Prime number Distributions

A twin prime is a prime number that differs from another prime number by two, for example the twin prime pair (41, 43). Sometimes the term twin prime is used for a pair of twin primes; an alternative name for this is prime twin or prime pair.

Twin primes appear despite the general tendency of gaps between adjacent primes to become larger as the numbers themselves get larger due to the prime number theorem (the "average gap" between primes less than n is log(n)).

$$p_2 - p_1 = 2$$

Every twin prime pair except (3, 5) is of the form (6n – 1, 6n + 1) for some natural number n, and with the exception of n = 1, n must end in 0, 2, 3, 5, 7, or 8

What Yitang Zhang just proved is that there are infinitely many pairs of primes that differ by at most 70,000,000. In other words, that the gap between one prime and the next is bounded by 70,000,000 infinitely often—thus, the "bounded gaps" conjecture.

On first glance, this might seem a miraculous phenomenon. If the primes are tending to be farther and farther apart, what's causing there to be so many pairs that are close together?

It is the geometry of Numbers themselves that causes a natural geometric convergence

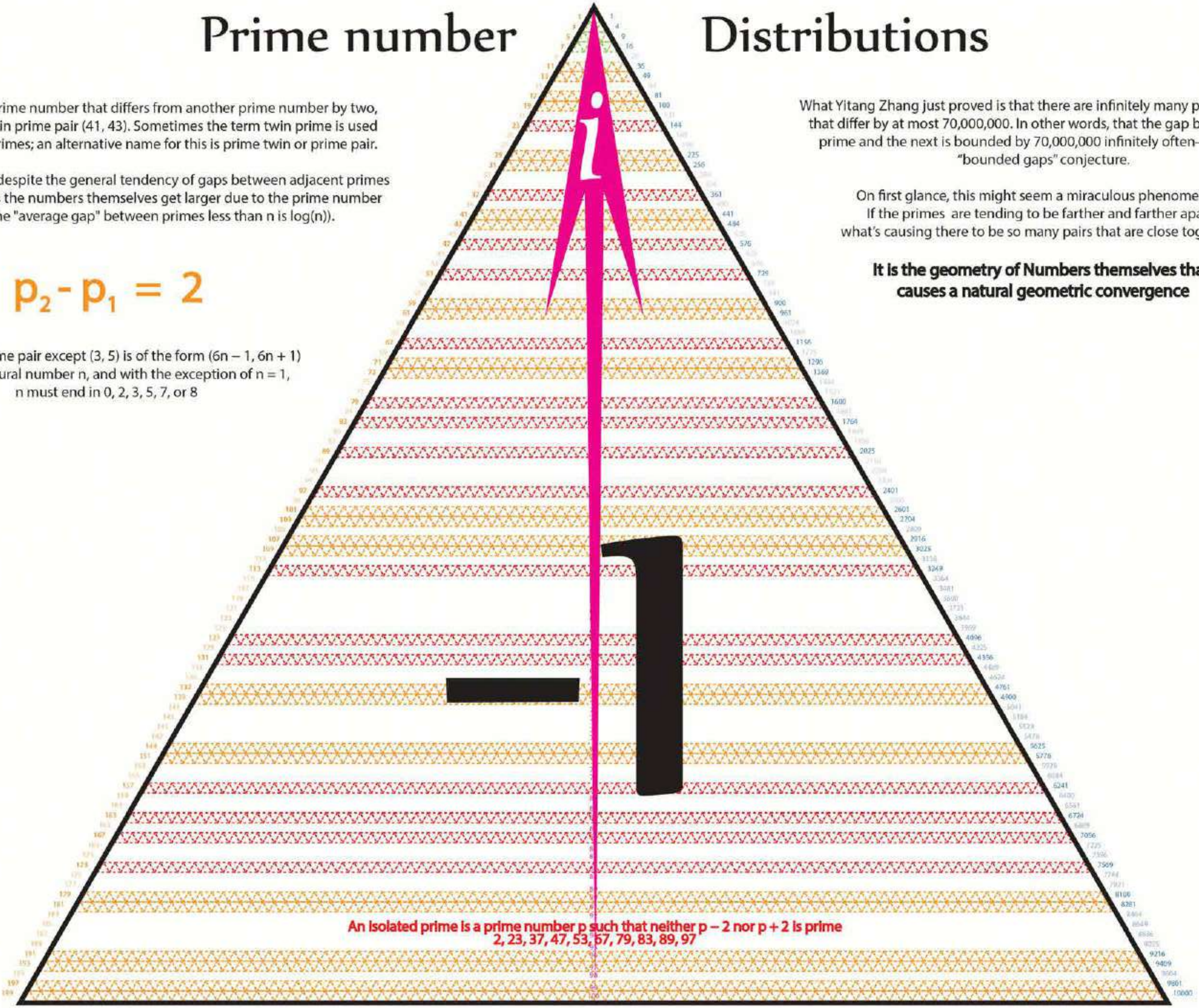

An isolated prime is a prime number p such that neither p – 2 nor p + 2 is prime
2, 23, 37, 47, 53, 67, 79, 83, 89, 97

Copyright ABRAHAM [2008] - All rights reserved

The Digital roots of Prime numbers

The digital root (also repeated digital sum) of a number is the (single digit) value obtained by an iterative process of summing digits, on each iteration using the result from the previous iteration to compute a digit sum.

The process continues until a single-digit number is reached.

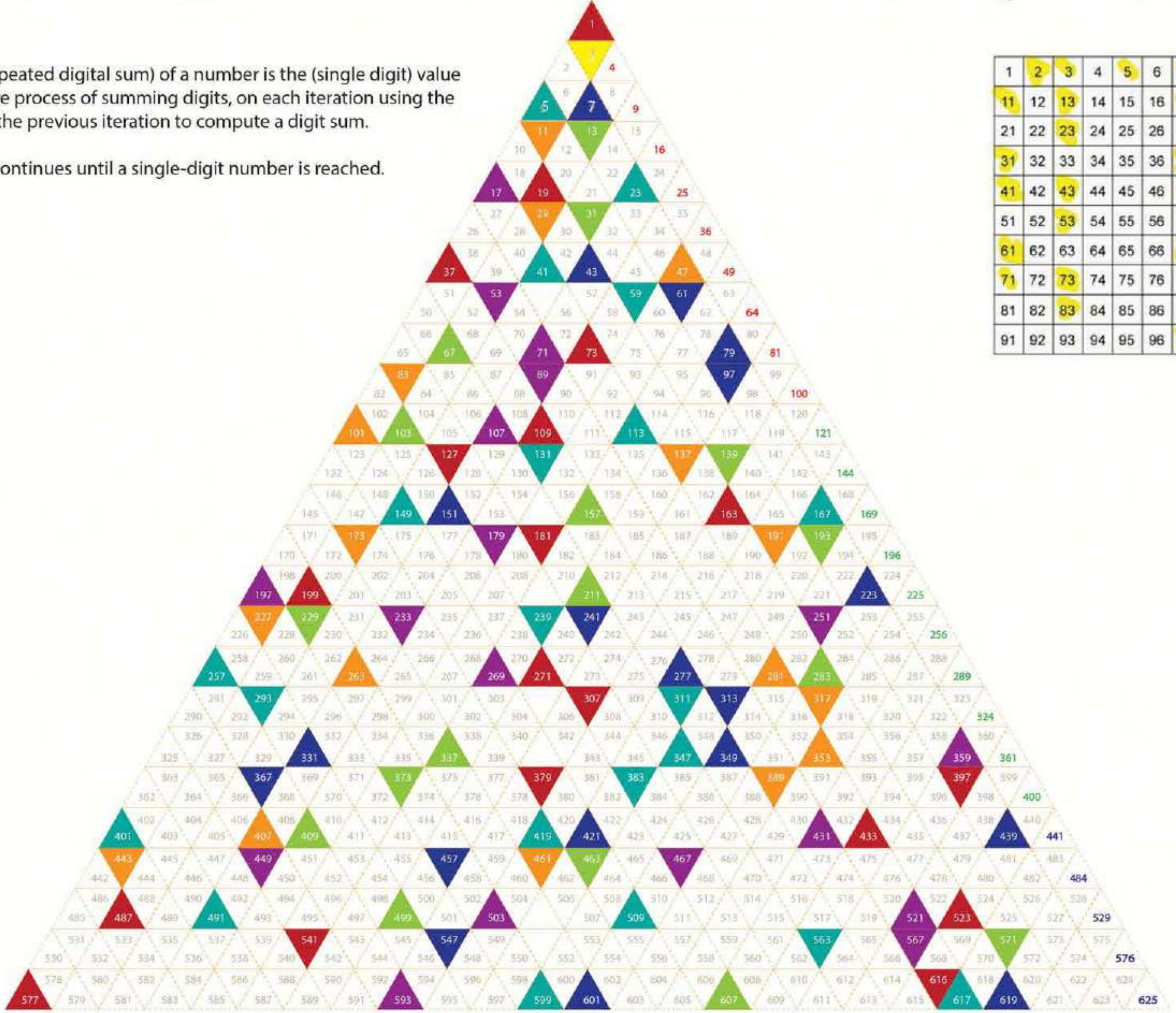

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Archimedes is given credit for the first calculus.

Archimedes infinite series

Today's calculus was published by Newton.

1 + 3/4 + 3/16 + 3/64 + 3/256 + 3/1024 + ...

1 + .75 + .1875 + .046875 + .01171875 + ...

$$3\left(\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3}+\frac{1}{4^4}+\cdots\right)=1.$$

$$\frac{3}{4}+\frac{3}{4^2}+\frac{3}{4^3}+\frac{3}{4^4}+\cdots=1.$$

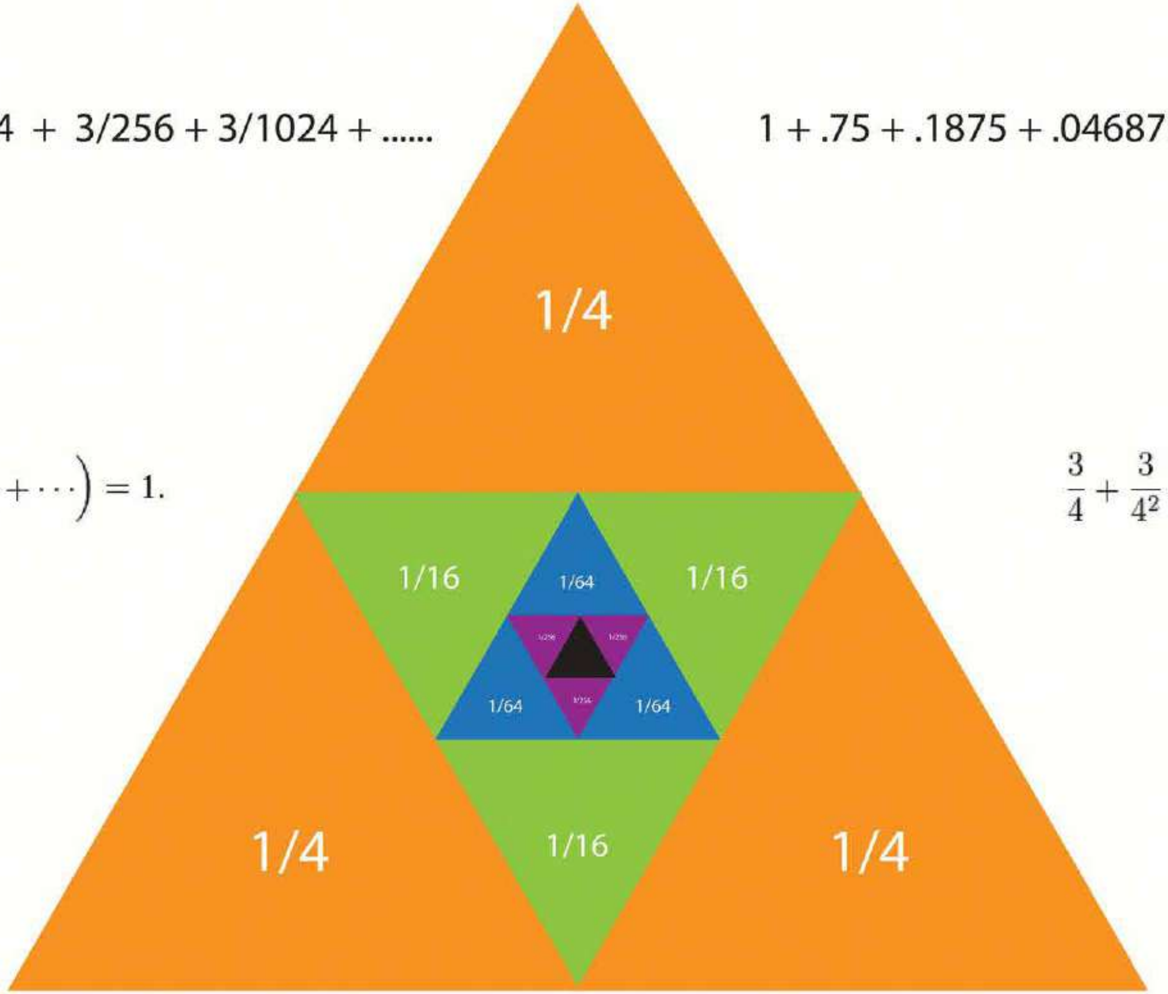
1/4
1/16
1/64
1/16
1/256
1/256
1/64
1/256
1/64
1/4
1/16
1/4

Nested convergent infinite series

Summing a convergent infinite series

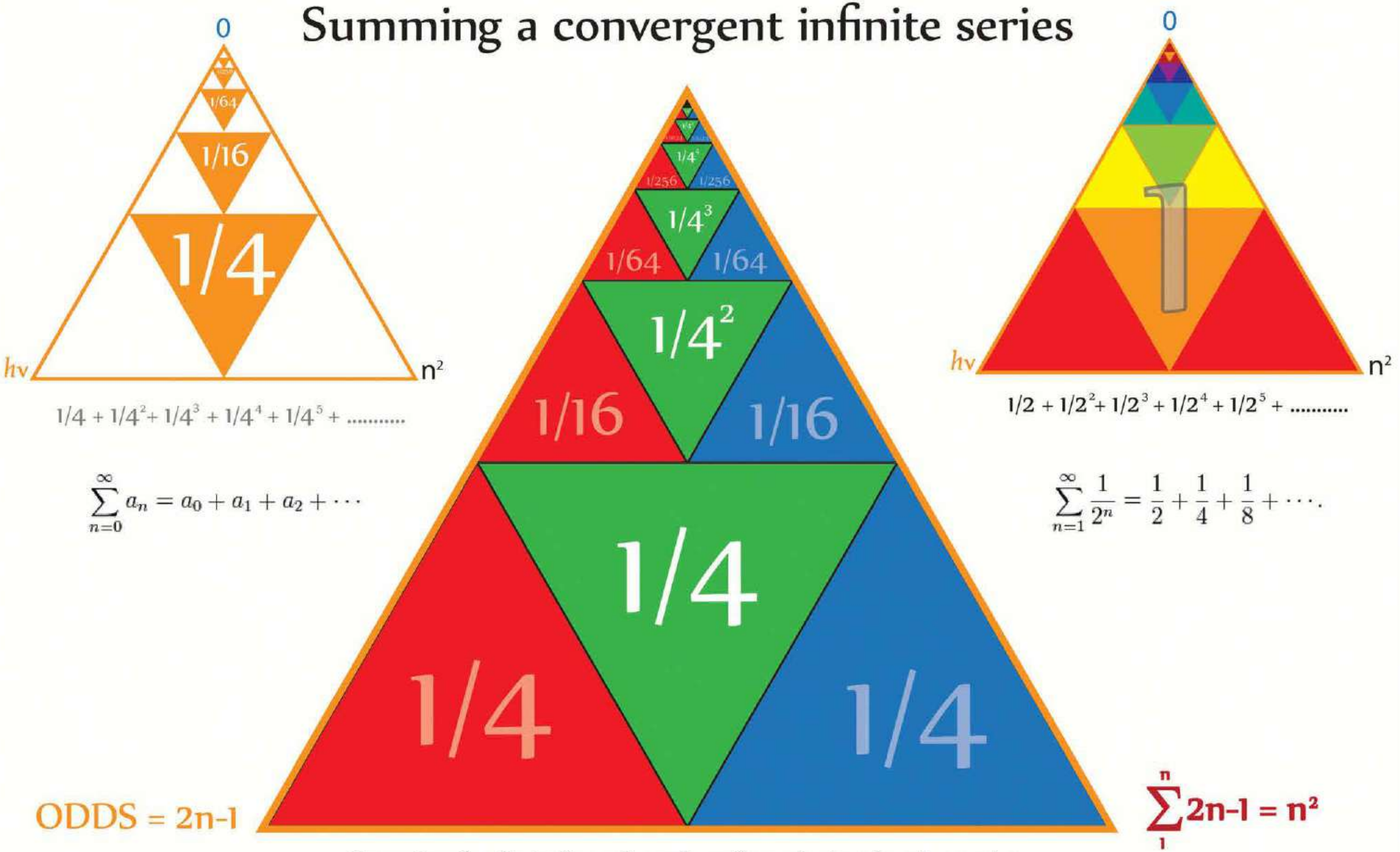
0
1/64
1/16
1/4
hv
n²
1/4⁴
1/256
1/256
1/4³
1/64
1/64
1/4²
1/16
1/16
1/4
1/4
1/4
0
1
hv
n²

$1/4 + 1/4^2 + 1/4^3 + 1/4^4 + 1/4^5 + \dots\dots\dots$

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots$$

$1/2 + 1/2^2 + 1/2^3 + 1/2^4 + 1/2^5 + \dots\dots\dots$

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

ODDS = 2n-1

$$\sum_{1}^{n} 2n\text{-}1 = n^2$$

Summing the dissimilar coloured equilateral triangles gives unity

1/3 1/3 1/3

1/3 1/3 1/3

Summing the dissimilar coloured equilateral triangles gives unity

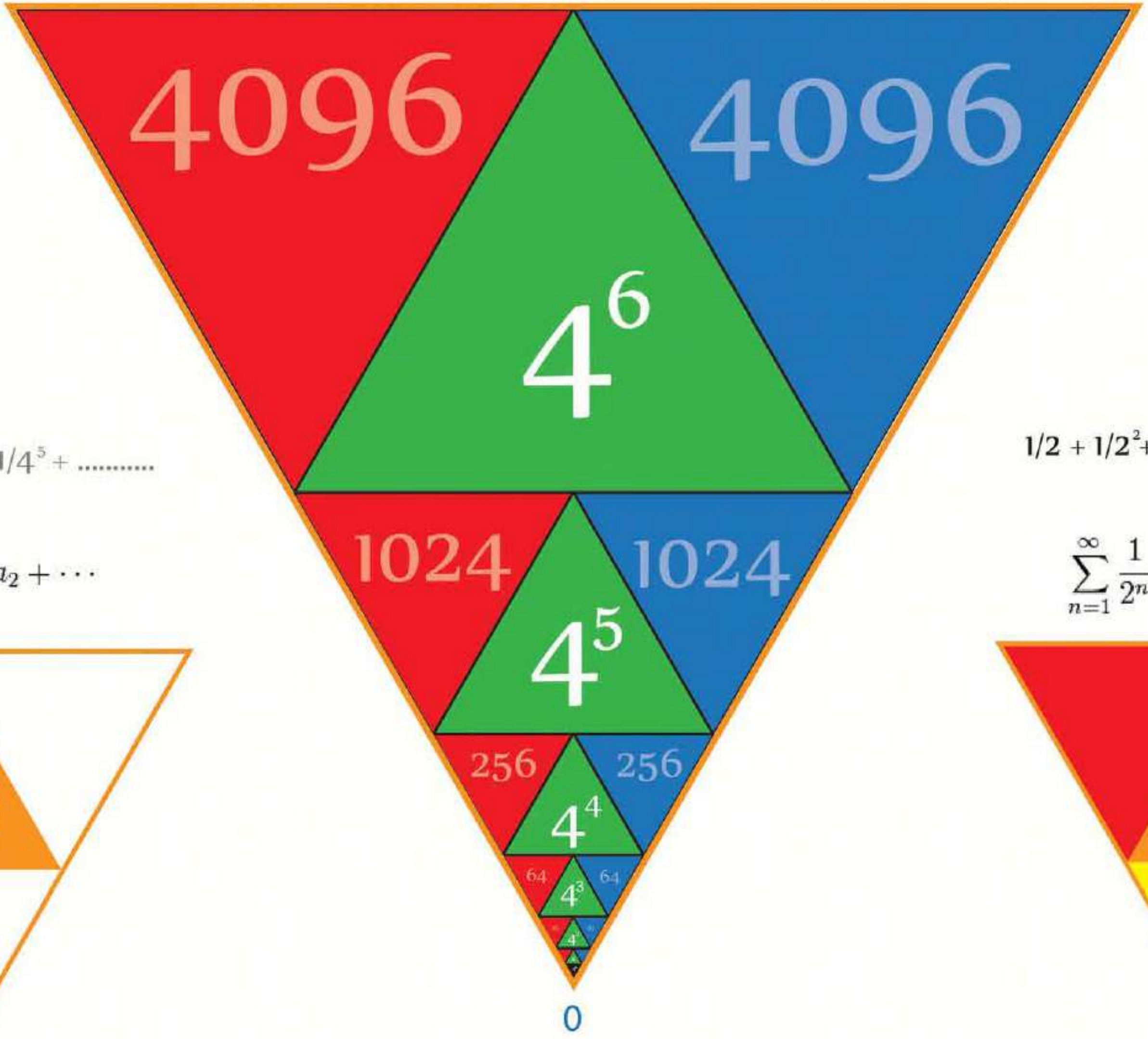
4096
4096
4^6
1024
1024
4^5
256
256
4^4
64
64
4^3
0

$1/4 + 1/4^2 + 1/4^3 + 1/4^4 + 1/4^5 + \ldots\ldots\ldots$

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots$$

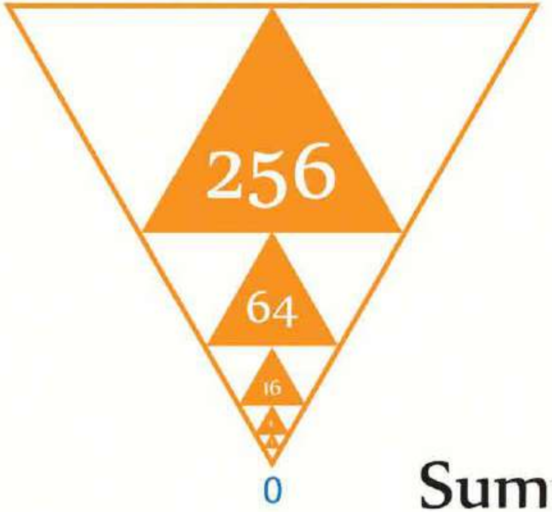

$1/2 + 1/2^2 + 1/2^3 + 1/2^4 + 1/2^5 + \ldots\ldots\ldots$

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

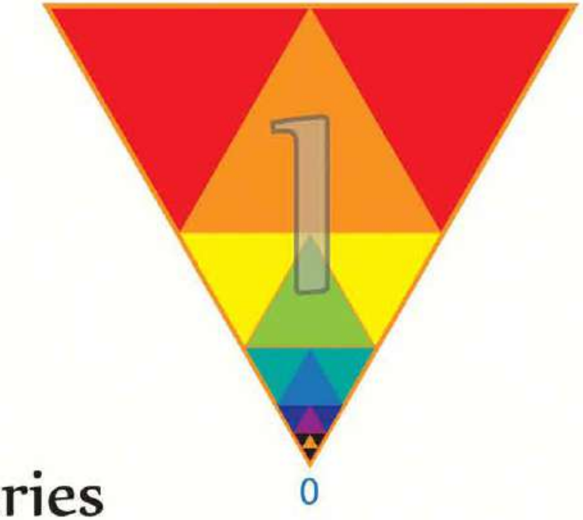

Summing a divergent infinite series

Copyright ABRAHAM [2008] - All rights reserved

Pietro Mengoli

(1626 – June 7, 1686)

The Basel Problem

The Basel problem is a famous problem in mathematical analysis with relevance to number theory, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735.

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6} \approx 1.644934$$

Leonhard Euler

(15 April 1707 – 18 Sept 1783)

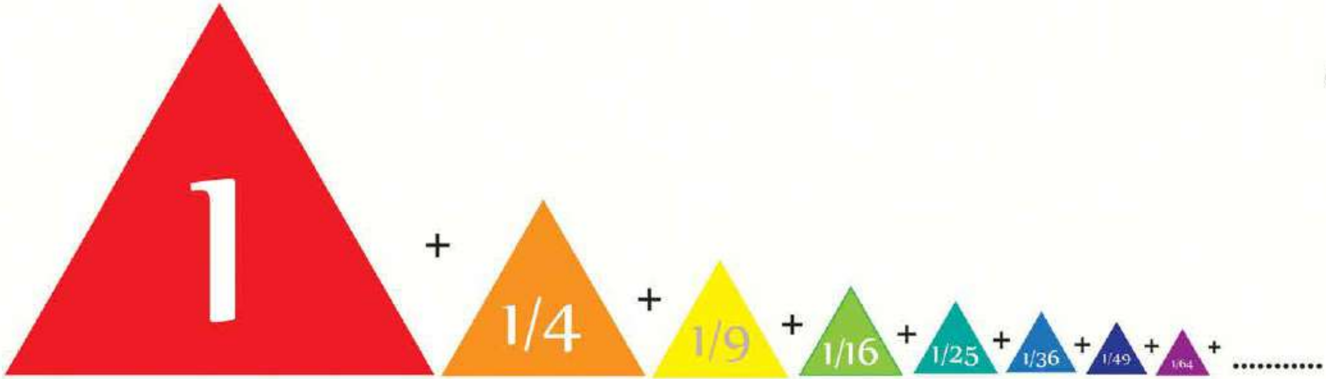

The Basel problem asks for the precise summation of the reciprocals of the squares of the natural numbers, i.e. the precise sum of the infinite series

$$f(n) = 1/n^2$$

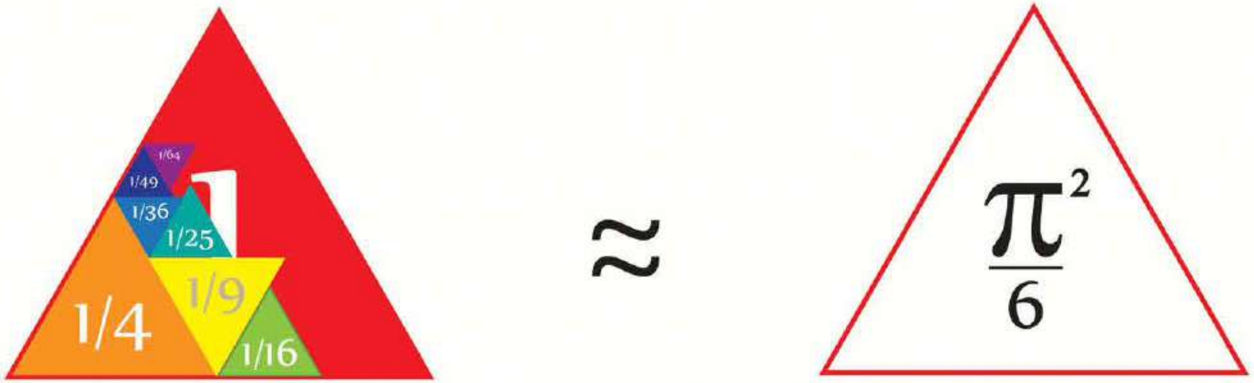

Integrals of mass-energy

is a means of finding scalar areas using summation and limits.

The summation of equilateral energy momenta quanta
with respect to their linear vector components

Integration is a micro adding of ***CONTINUOUS*** quantities.

$$\int_{n}^{\infty} f(x)$$

Integration is a special case of summation.

Integration is defined as the limit of a summation as the number of elements approches infinity while a part of their respective value approaches zero.

Summation is the finite sum of multiple, fixed values.

$$\sum_{n}^{\infty} x$$

Summation is a macro adding of ***DISCRETE*** quantities.

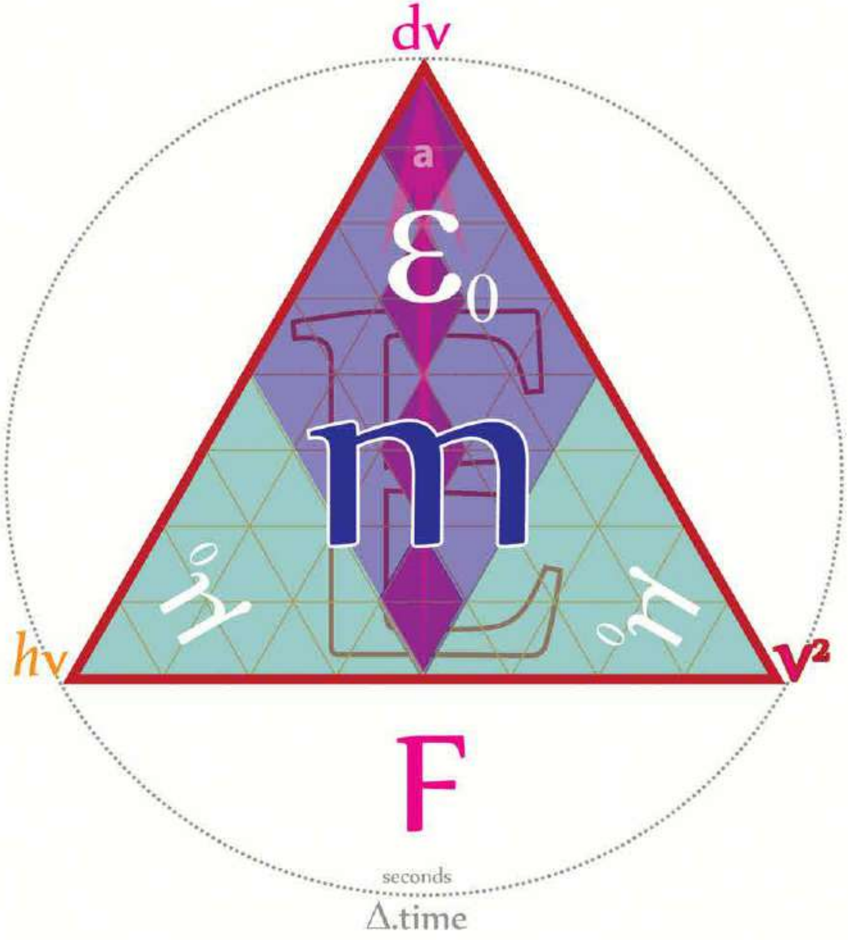

Integrating the energy quanta contained within equilateral charge geometry
gives the variable Force required to acheive changes in motion
[Energy, work, acceleration]

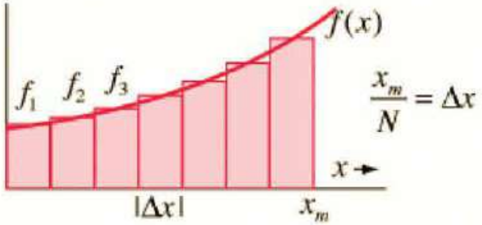

The Integral of the continuous area under the curve is the summation of an infinte number of disctrete rectangular measurements made to a specified limit

An integration isn't a simple summation, but the limit of a sequence of summations

All Planck energies are discrete equilateral geometries

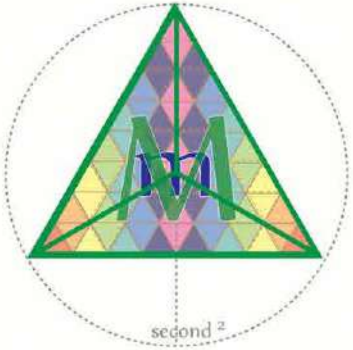

mass is the surface integral of
EM energy geometries
per unit of time

Copyright ABRAHAM [2008] - All rights reserved

"The calculus of infinitesimals"

The fundamental theorem of calculus
simply states that the sum of infinitesimal changes in a quantity over time
adds up to the net change in the quantity.

Scalar Energy

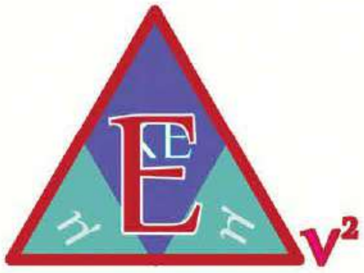

$E = mv^2$

Leibniz's vis viva
(Latin for living force)
is mv2,

[2 x kinetic energy]

Gottfried Wilhelm von Leibniz

(July 1, 1646 – November 14, 1716)

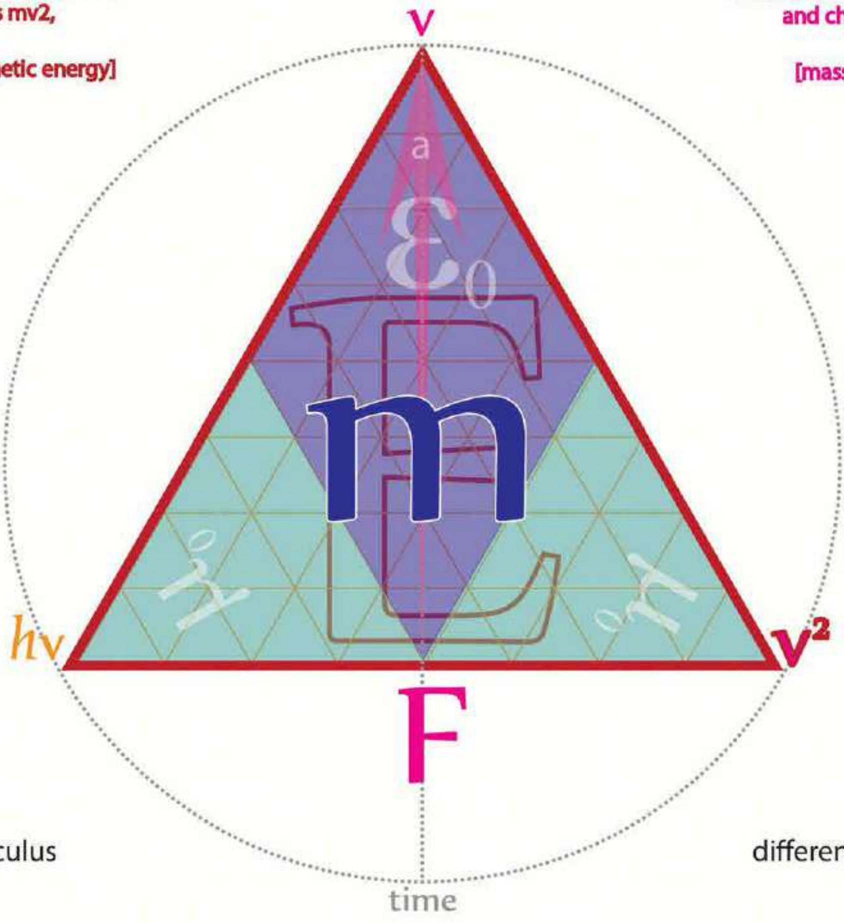

vector Force

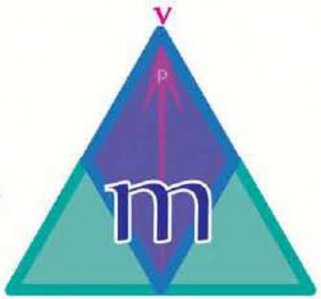

Much of Newton's work
centred around momentum
and changes to it

[mass.velocity]

$F = m[dv/dt]$

Sir Isaac Newton

(1643-1727)

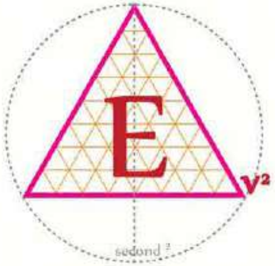

integral calculus

The founders of calculus thought
of the integral as an infinite sum of
rectangles of infinitesimal width

The fundamental theorem of calculus is a theorem that links
the concept of the integral with the derivative of a function.

differential calculus

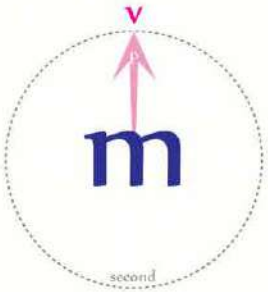

In calculus, a branch of mathematics,
the derivative is a measure of how
a function changes as its input changes

Copyright ABRAHAM [2008] - All rights reserved

Energy momentum geometries

Leibniz

scalar energies

linear momentum is the square root vector force of the scalar energy required to do a set amount of work

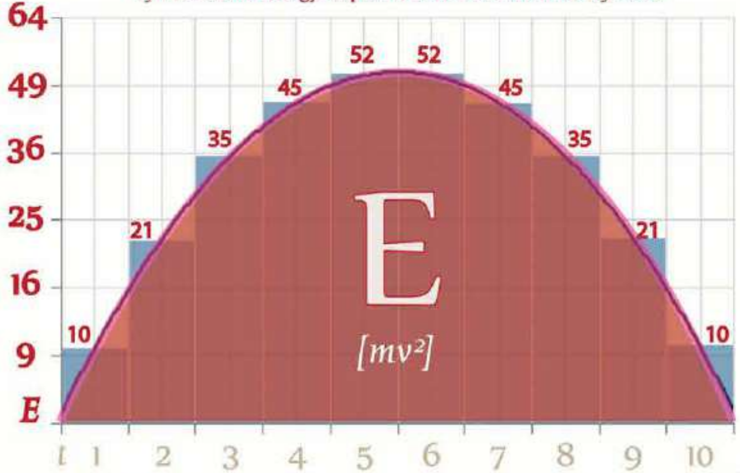

$$E = p^2$$

Energy momentum

$$kg\left[\frac{m}{s}\right]^2$$

Newton

linear momentum

$$d[mv]/dt = dp/dt = ma$$

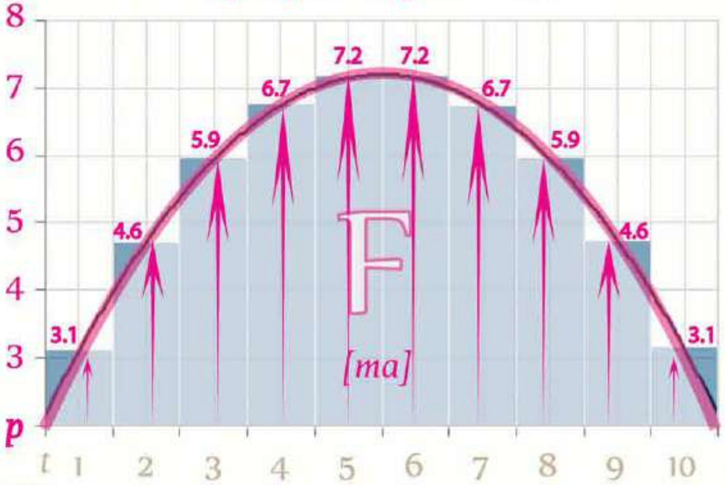

print out this page

then cut out the 10 Planck energy momenta triangles shown

slice fine lines along the pink linear momentum arrows with a razor blade or similar as highlighted under the Newtonian acceleration curve

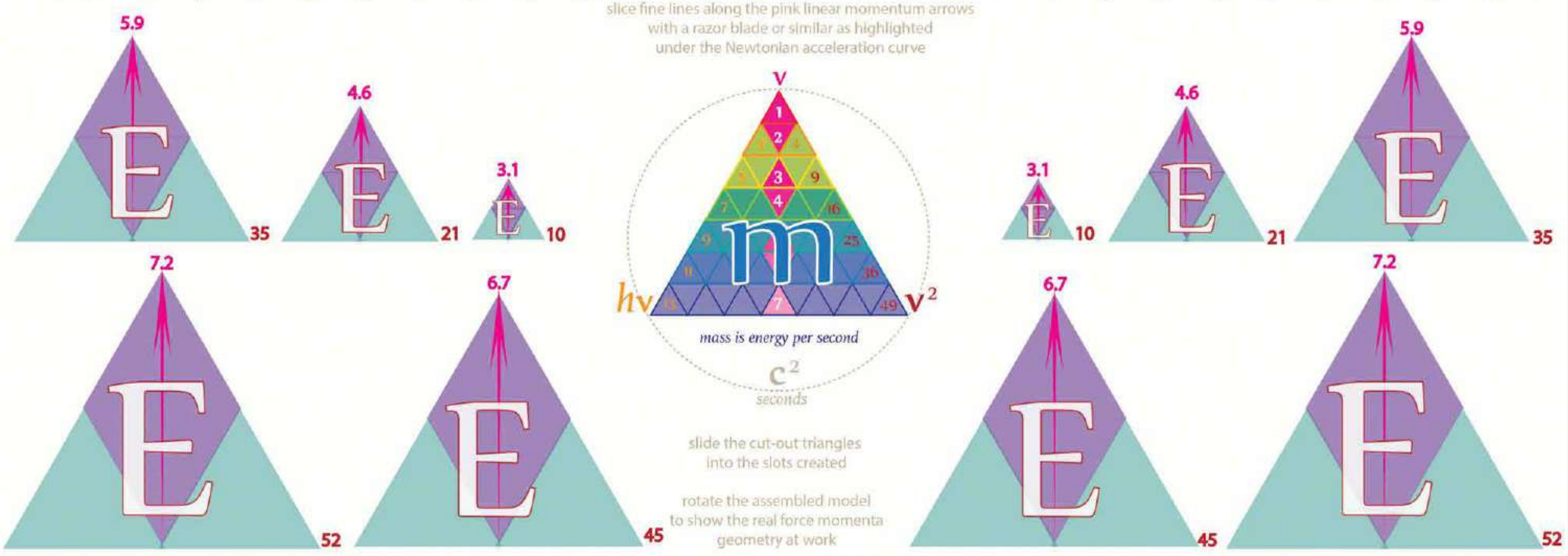
5.9
E
35
4.6
E
21
3.1
E
10
7.2
E
52
6.7
E
45
v
1
2
3
4
9
16
25
36
49
7
m
hv
v²
c²
seconds
3.1
E
10
4.6
E
21
5.9
E
35
6.7
E
45
7.2
E
52

mass is energy per second

slide the cut-out triangles into the slots created

rotate the assembled model to show the real force momenta geometry at work

Tetryonics 87.07 - Physically modeling the geometric forces of acceleration in calculus *linear momentum is the square root vector force of the scalar energy required to do a set amount of work*

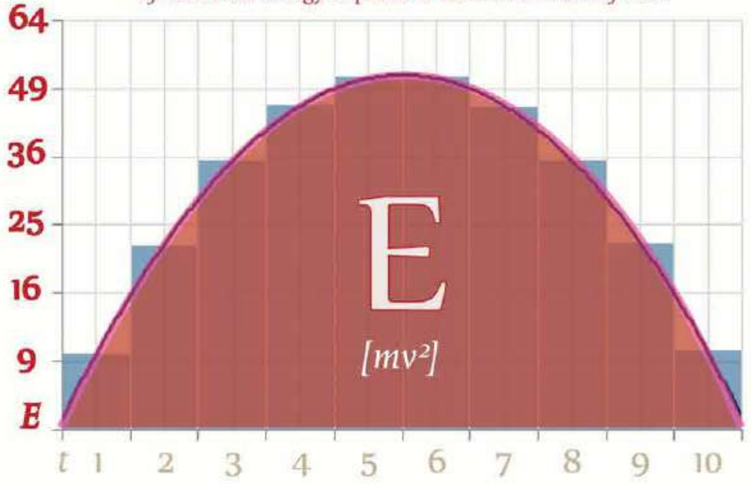

p

linear momentum

$\mathrm{kg}\,\dfrac{\mathrm{m}}{\mathrm{s}}$

print out this page

$d[mv]/dt = dp/dt = ma$

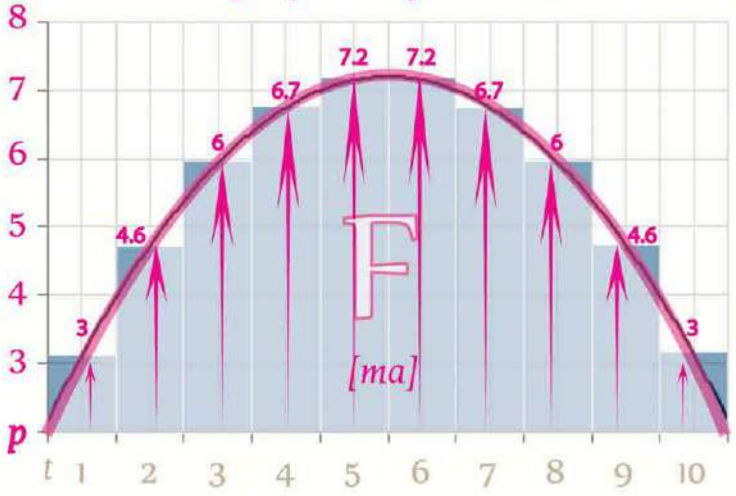

Leibniz

scalar energies

$$E = mv^2$$

mass-energy momenta relationship

square root equilateral energy is linear momentum

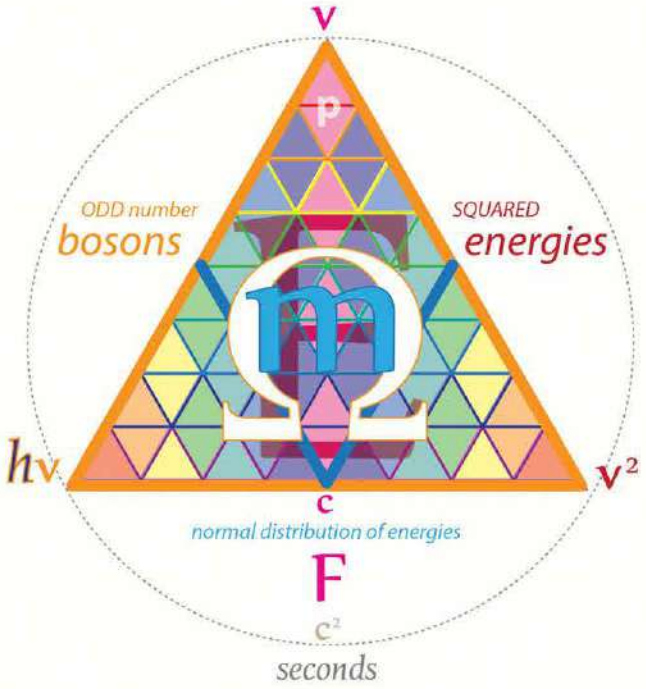
v
p
ODD number
bosons
SQUARED
energies
m
hv
v²
c
normal distribution of energies
F
c²
seconds

Planck quanta and their vector linear momentum lie at 90 degrees to the angle use in the graphs of motion in calculus

Newton

linear momentum

$$F = ma$$

mass is a scalar constant relating Force to acceleration

$$F = m\frac{\Delta v}{\Delta t}$$

Differentiation

Differentiation is concerned with things like speeds and accelerations, slopes and curves etc.
These are Rates of Change, they are things that are defined locally.

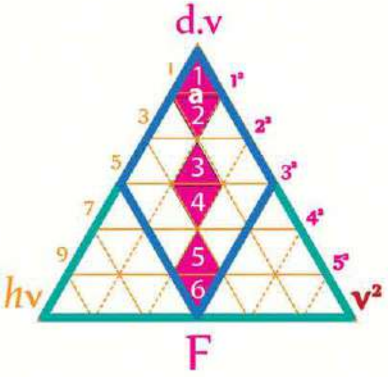
d.v
1
a
2
3
4
5
6
1'
2'
3'
4'
5'
1
3
5
7
9
hv
v²
F

An increase in a force opposing an object's vector velocity results in
DECCELERATION

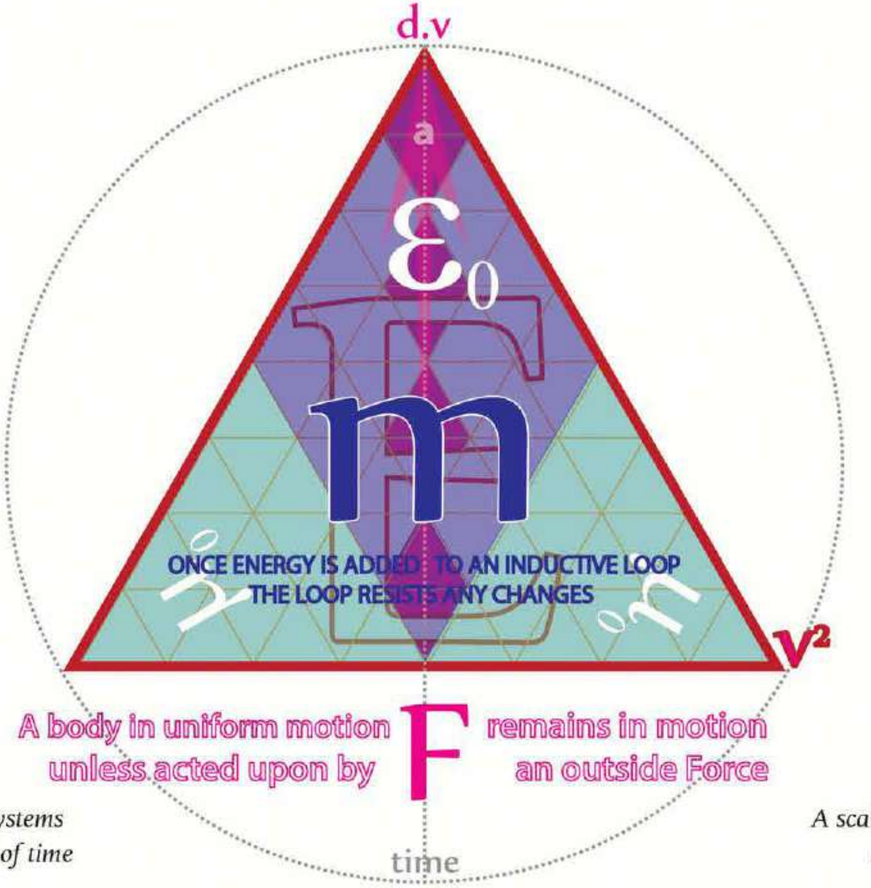
d.v
a
ε0
E
m
ONCE ENERGY IS ADDED TO AN INDUCTIVE LOOP
THE LOOP RESISTS ANY CHANGES
μ0
v²
F
time

A body in uniform motion remains in motion
unless acted upon by an outside Force

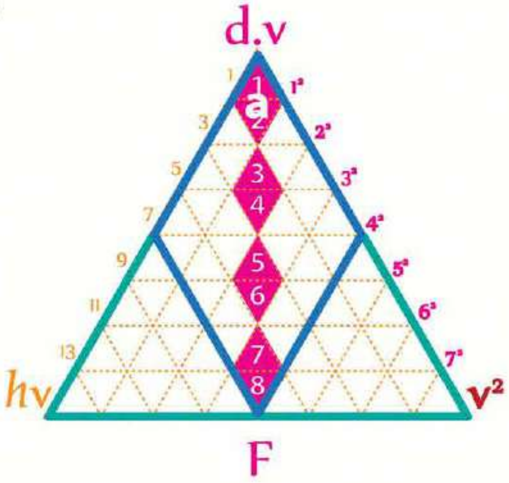
d.v
1
a
2
3
4
5
6
7
8
1'
2'
3'
4'
5'
6'
7'
1
3
5
7
9
11
13
hv
v²
F

An increase in a force in line with an object's vector velocity results in
ACCELERATION

A linear measure of Forces acting on physical systems resulting in changes to distance covered per unit of time

A scalar measure of Forces acting on physical systems resulting in changes to their rate of motion

$$v = \frac{\Delta d}{\Delta t}$$

$$= m/s$$

$$F = ma$$

In physics, the derivative of the displacement of a moving body with respect to time is the velocity of the body, and the derivative of velocity with respect to time is acceleration.

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \mathbf{v}\frac{\mathrm{d}m}{\mathrm{d}t}$$

Newton's second law of motion states that the derivative of the momentum of a body equals the force applied to the body.

$$a = \frac{\Delta v}{\Delta t}$$

$$= m/s^2$$

Copyright ABRAHAM [2008] - All rights reserved

Visualising the geometric half-truths of relativistic physics

The source of all the physical relationships of mass-energy momenta and the constants in Physics is the Equilateral Triangle (and all texts must be corrected)

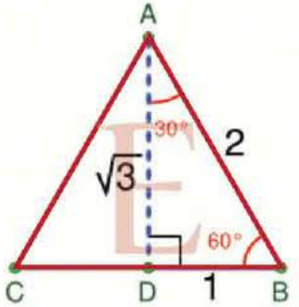

Energy geometries within Physics including Special Relativity with its Lorentz corrections have historically been incorrectly illustrated through the geometry of right angled triangles

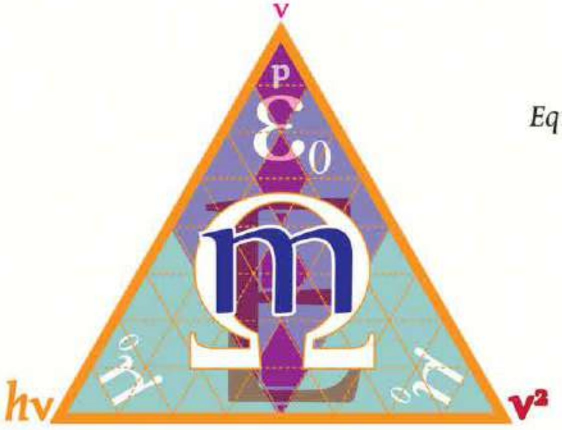

F

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = m\mathbf{a}$$

$$\mathbf{F} = m\mathbf{a}\,.$$

Physics is geometry, one cannot be separated from the other

Equilateral geomtries lead to a intuitive understandings of Physics, Chemistry, Electrodynamics and Gravitation along with all the other apsects of Nature.

$$mv^2 = E = hv^2$$

ENERGY

6.629432673 e-34 J

Planck quanta

$$\left[m\Omega v^2\right]$$

mass velocity

7.376238634 e-51 kg

momentum

$$p^2 = mv^2$$

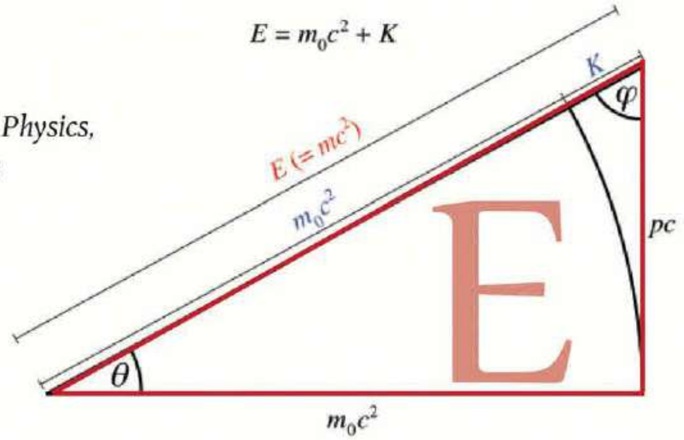

$$E^2 = p^2c^2 + m_0^2c^4$$

Generalizing, we see that the square of the total energy, mass, or distance in spacetime is the sum of the components squared.

We can see an origin of distance in spacetime relating to velocity in pc in which Energy is subject to Lorentz corrections [v/c]

$$E = pc.$$

Additionally, EM mass can be directly related tot the Energy content of a body by the velocity of Energy

$$E = mc^2$$

Velocity

In physics, velocity is the measurement of the rate and direction of change in the position of an object.

It is a vector physical quantity; both magnitude and direction are required to define it.

The scalar absolute value (magnitude) of velocity is speed, a quantity that is measured in metres per second (m/s or ms–1) when using the SI (metric) system.

$$\frac{m}{s}$$

Speed is the scalar value of the Distance traveled per unit of TIme

$$\bar{v} = \frac{\Delta x}{\Delta t}.$$

Velocity is the vector value of the Distance traveled per unit of Time

$$\frac{\vec{m}}{s}$$

Velocity squared is the scalar value of the Distance traveled per unit of Time squared (Energy of a given spatial volume)

$$\frac{m^2}{s^2}$$

v Velocity $\frac{m}{s}$

sec

is a 2D
radial Space-time
MEASUREMENT

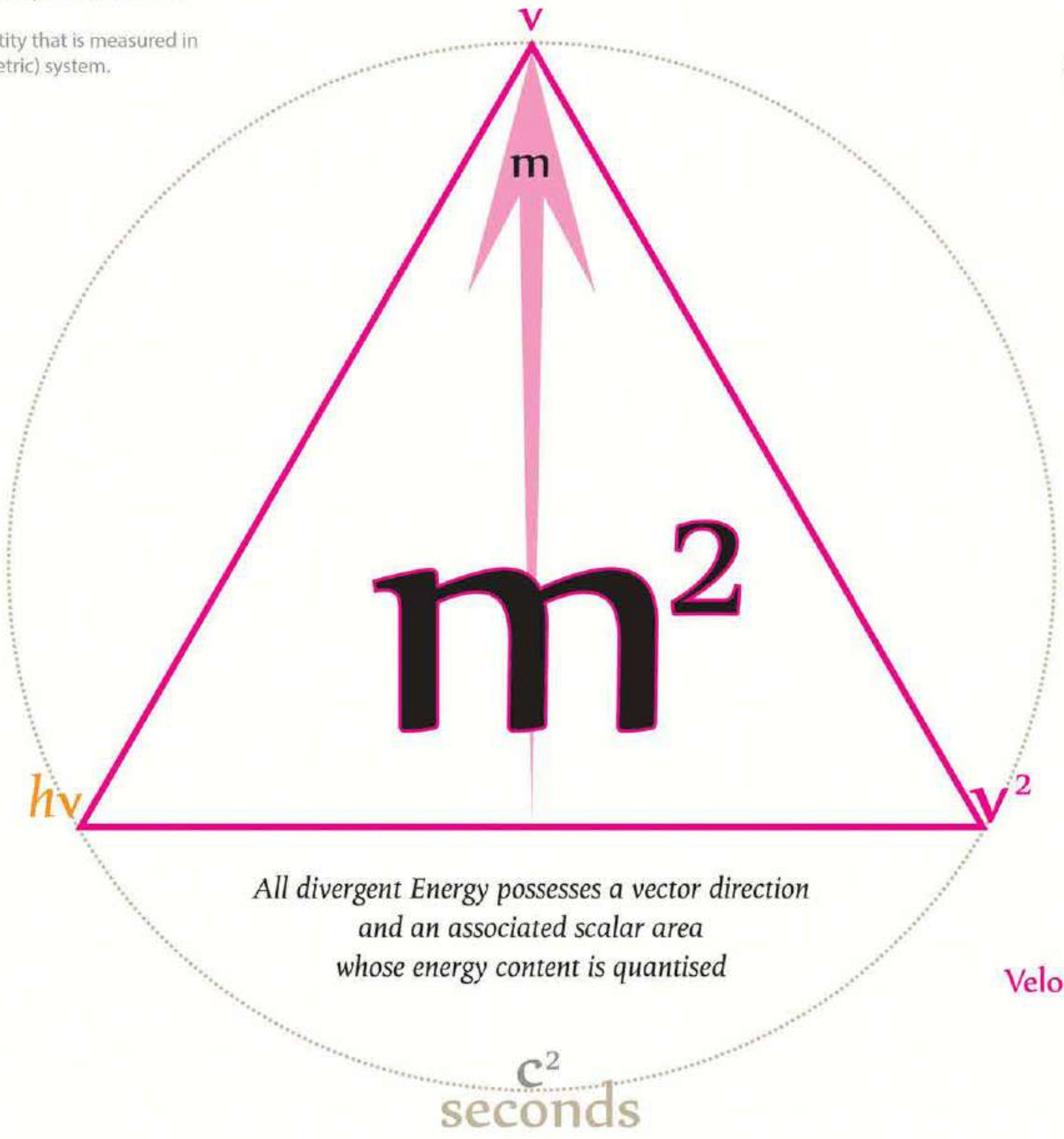

All divergent Energy possesses a vector direction and an associated scalar area whose energy content is quantised

Velocity squared

$$\frac{m^2}{s^2}$$

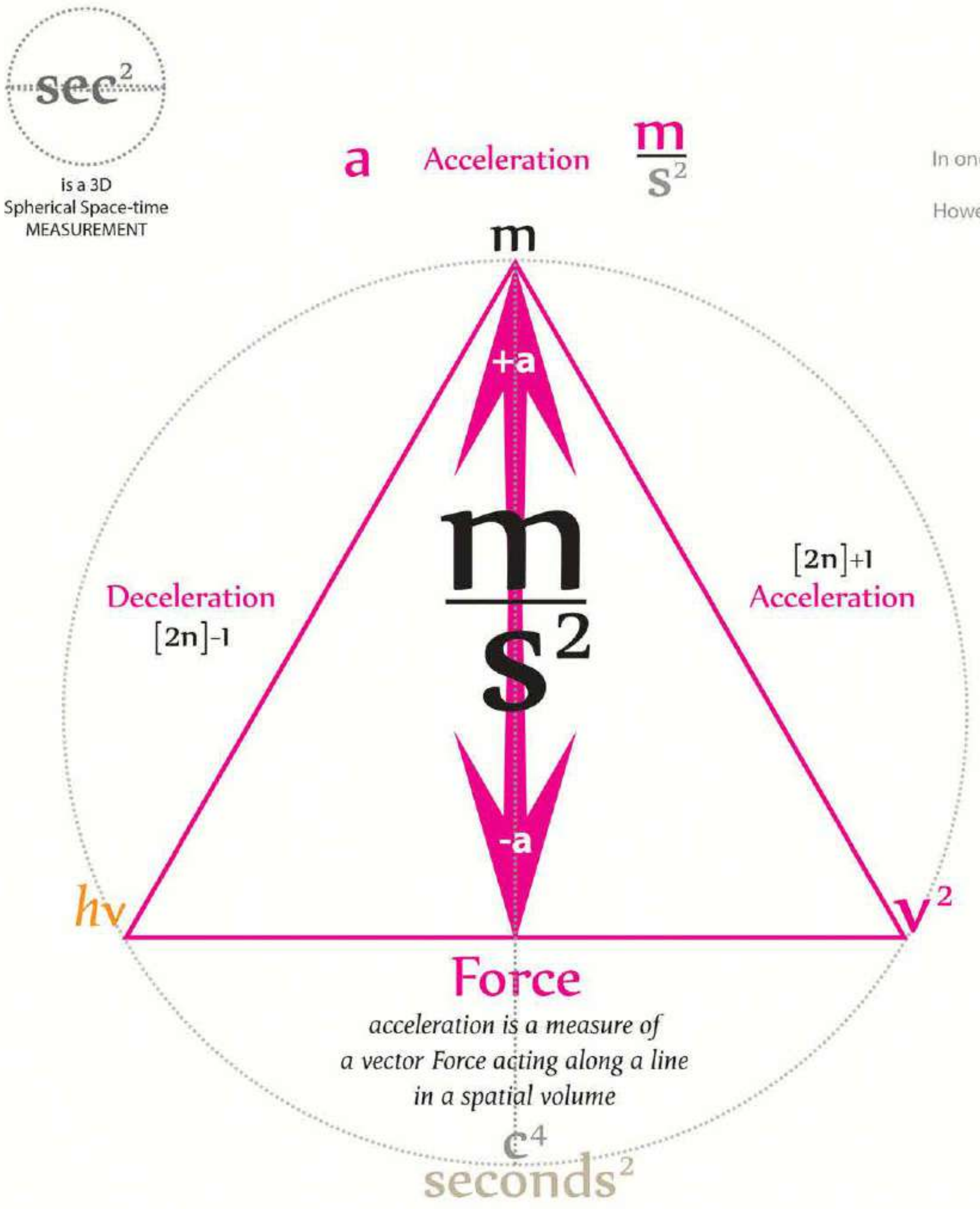
sec^2
is a 3D
Spherical Space-time
MEASUREMENT
a
Acceleration
$\frac{m}{s^2}$
m
+a
$\frac{m}{s^2}$
-a
Deceleration
[2n]-1
[2n]+1
Acceleration
hv
v^2
Force
c^4
seconds^2

acceleration is a measure of a vector Force acting along a line in a spatial volume

Acceleration

In physics, acceleration is the rate of change of velocity over time [dt]

In one dimension, acceleration is the rate at which something speeds up or slows down.

However, since velocity is a vector, acceleration describes the rate of change of both the magnitude and the direction of velocity.

Acceleration has the dimensions [Length]/[Time Squared]
In SI units, acceleration is measured in meters per second squared (m/s^2).

$$a = \frac{\Delta y}{\Delta x} = \frac{\Delta v}{\Delta t}.$$

In classical mechanics, for a body with constant mass,
the acceleration of the body is proportional to the net force acting on it
(Newton's second law)

$$F=ma \longrightarrow a=F/m$$

Force

$$kg\,\frac{m}{s^2}$$

Additionally, for a mass with constant velocity,
(ie in an inertial frame)
the energy of motion is expressed as its momentum
(acceleration causes changes in Energy-momentum)

$$p = kg\,\frac{m}{s}$$

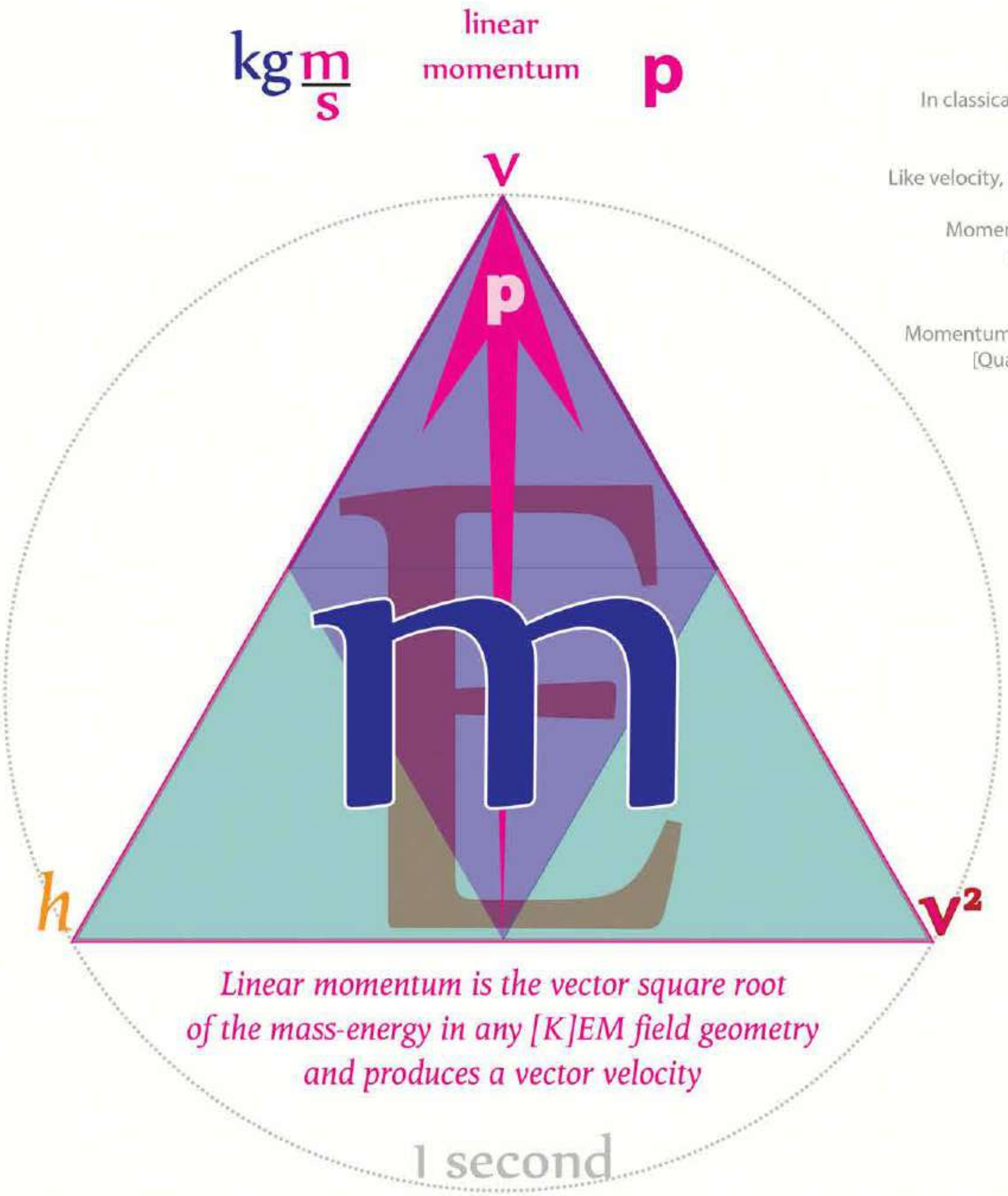

Linear momentum is the vector square root of the mass-energy in any [K]EM field geometry and produces a vector velocity

Momentum

In classical mechanics, momentum (pl. momenta; SI unit kg·m/s, or, equivalently, N·s) is the product of the mass and velocity of an object (p).

Like velocity, momentum is a vector quantity, possessing a direction as well as a magnitude.

Momentum is a conserved quantity (law of conservation of linear momentum), meaning that if a closed system is not affected by external forces, its total momentum cannot change.

Momentum should be referred to in its specific forms to distinguish it in its various forms [Quantised Angular, Linear, Rotational and quantum/nuclear momentum]

$$\boxed{p = \frac{hv^2}{v} = mv}$$

Although originally expressed in Newton's Second Law, the conservation of momentum also holds in special relativity and, with appropriate definitions, a (generalized) momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory, and general relativity.

In relativistic mechanics, non-relativistic momentum is further multiplied by the Lorentz factor.

$$p^2 = E = mv^2$$

Energy can be expressed as the square of linear momentum

$$kg\frac{m^2}{s^2}$$

equilateral Planck energy momenta

$$m\Omega v^2$$

Copyright ABRAHAM [2008] - All rights reserved

Energy-momentum relationship

The total number of equilateral Planck quanta [quantised mass-energy momenta]
is directly related to the square of its linear momentum [mass-velocity]

Quantum Mechanics

$$h = \frac{E}{v^2}$$

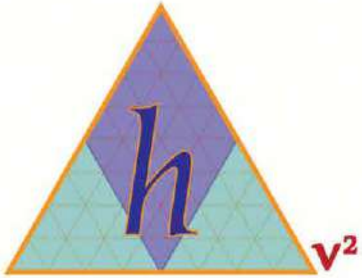

Quantised energy
equilateral momenta

$$hv^2$$

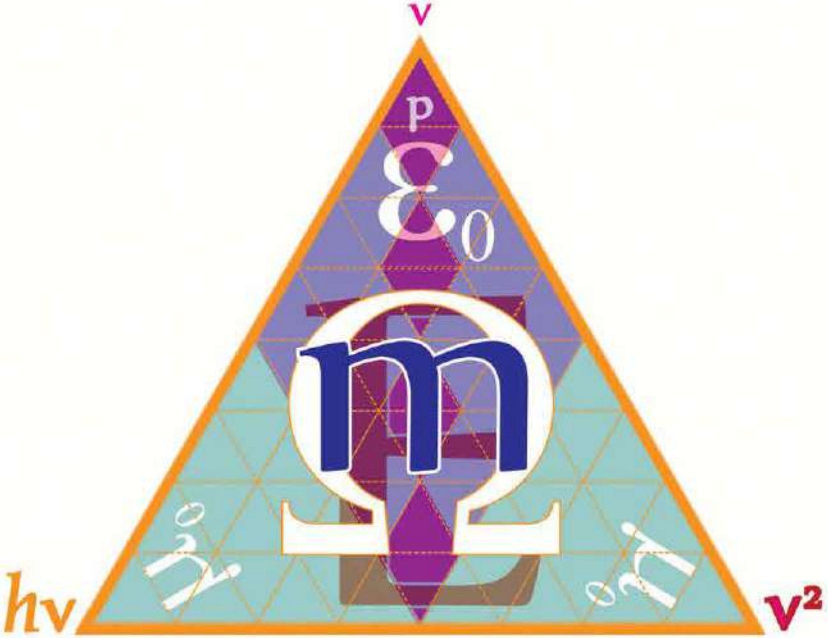

Quantised Energy momenta is related to
Scalar mass energy momenta through
the equilateral geometry of Planck's constant

Newtonian physics

$$m = \frac{E}{v^2}$$

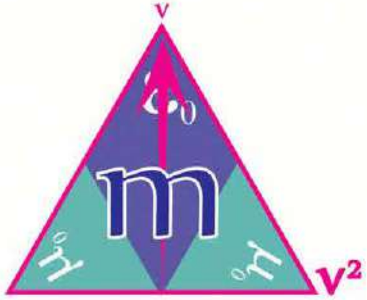

Scalar energy
linear momentum

$$pv$$

$$m\Omega v^2 = E = mv^2$$

Inertial resistance to Force

Inertia is the resistance of any physical object
to a change in its state of motion.

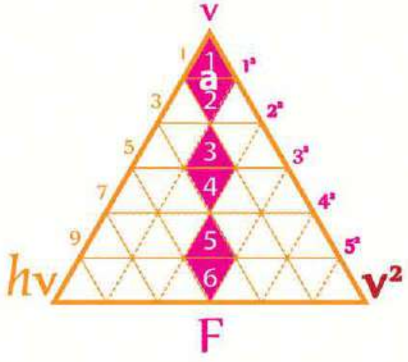

Changes to mass-velocity require a corresponding change to boson quanta

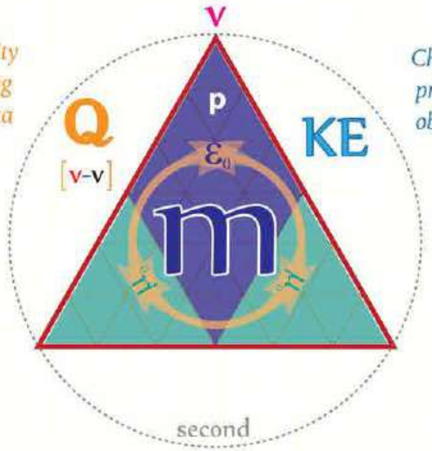

Changes to mass-velocity produces a change in an object's Kinetic Energies

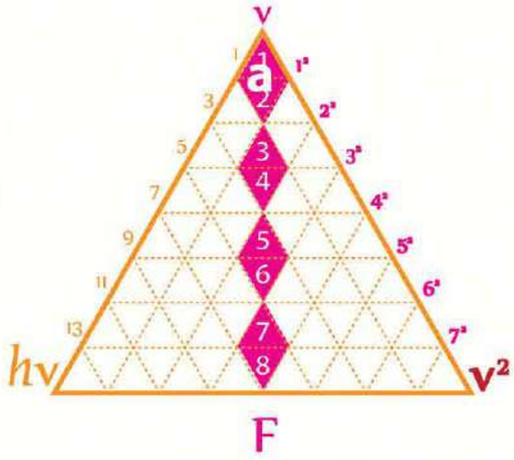

$$E = mc^2$$

The total intrinsic momenta of all energy waveforms is the sum of their constituent Quantised Angular momenta (mass-energy momenta)

$$M_0 = m/c^2$$

Any change in motion results in changes to
the Charge geometries creating in turn proportional
changes to KEM mass and momenta components

$$F = ma$$

The 'inductive resistance' of Charge quanta fields to changes in their mass-energy momenta content is what we term Inertia

$$KEM = Mv^2$$

Matter in motion has Kinetic Energies in addition to invariant rest mass-Energy

KE = RE - rest Matter

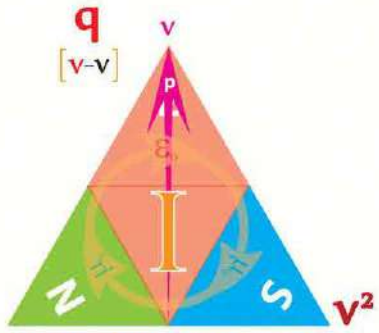

quantised moment of inertia

Any change to an object's velocity results in a corresponding change to its mass-Energy momenta which is reflected by its inertia

$$p^2 = E = mv^2$$

Inertial mass can be related to Charge through inductive Planck quanta

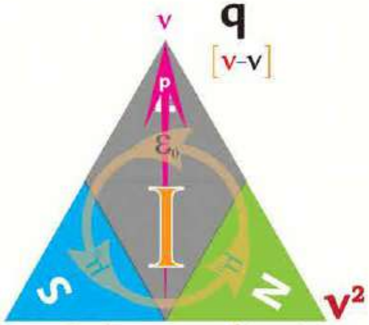

quantised moment of inertia

Copyright ABRAHAM [2008] - All rights reserved

Quantised Angular momentum

As it is a physical [equilateral] geometry QAM is conservative in any system where there are no external Forces and serves as the foundational geometric source for all the conservation laws of physics

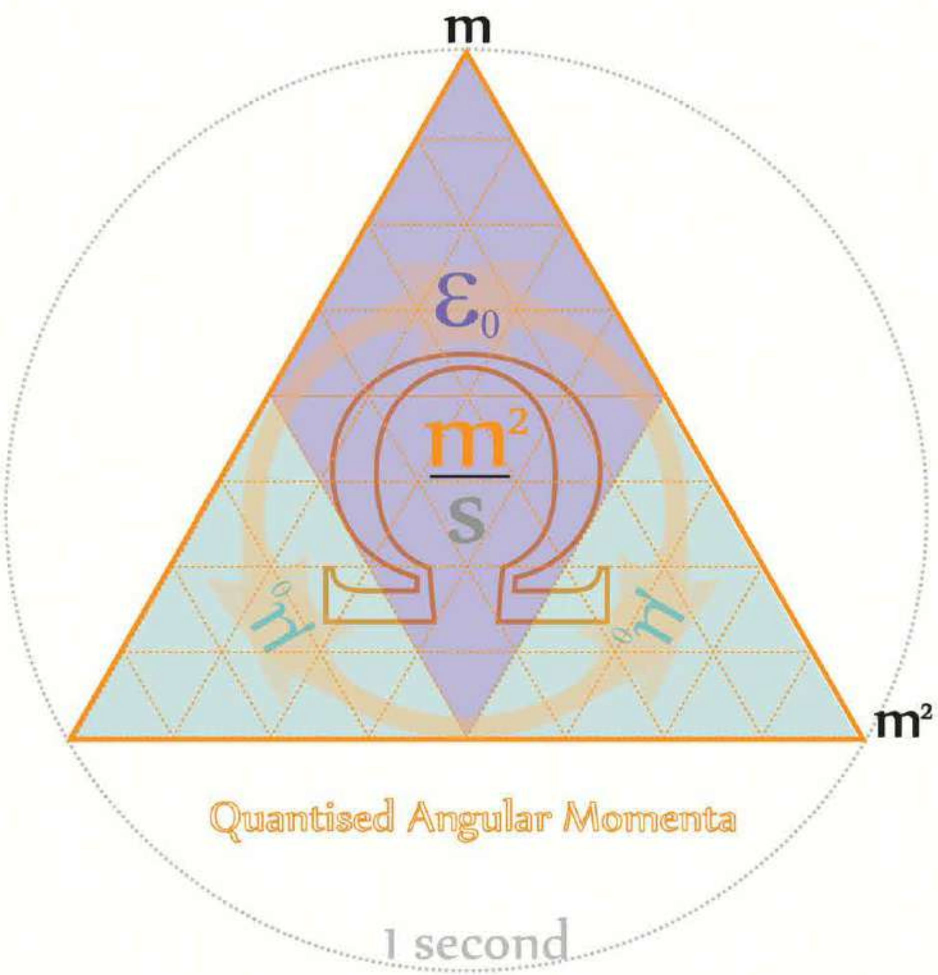

Angular momentum is sometimes described as the rotational analog of linear momentum, in Tetryonics it is revealed to be the equilateral geometry of quantised mass-energy momenta within any defined space-time co-ordinate system

A major re-definition of quantised angular momentum in physics is revealed

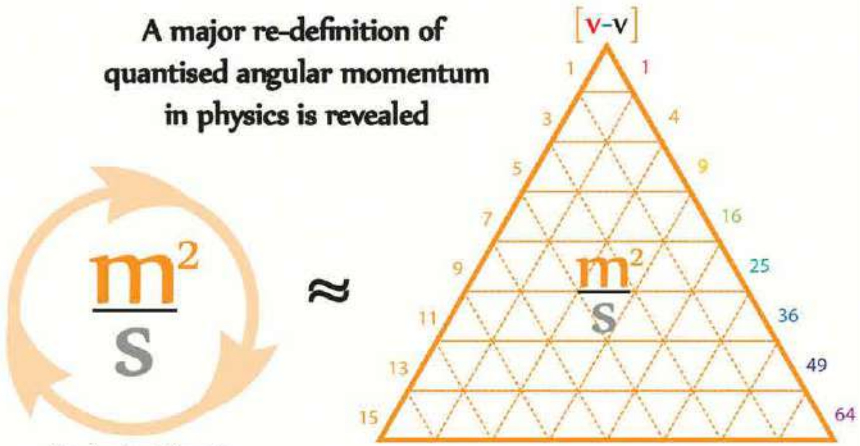

classical rotational angular momentum

Quantised Angular momentum

In quantum mechanics, angular momentum is quantised – that is, it cannot vary continuously, but only in ODD number "quantum steps" between the allowed SQUARE nuclear Energy levels

In physics, angular momentum, moment of momentum, or rotational momentum is a conserved vector quantity that can be used to describe the overall state of a physical system.

When applied to specific mass-Energy-Matter systems QAM reveals the true quantum geometry and nature of Energy in our universe

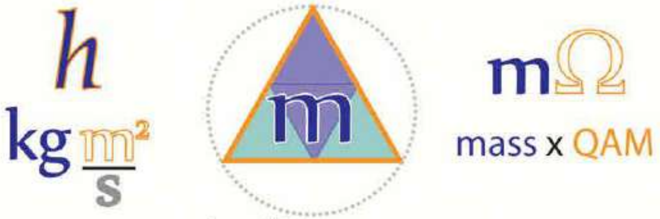

Planck's Constant

Normally viewed as an expression of rotational momentum Quantised Angular Momentum [QAM] is in fact a result of the equilateral geometric quantization of mass-energy

Copyright ABRAHAM [2008] - All rights reserved

Charged geometries

All charge geometries are nett divergent

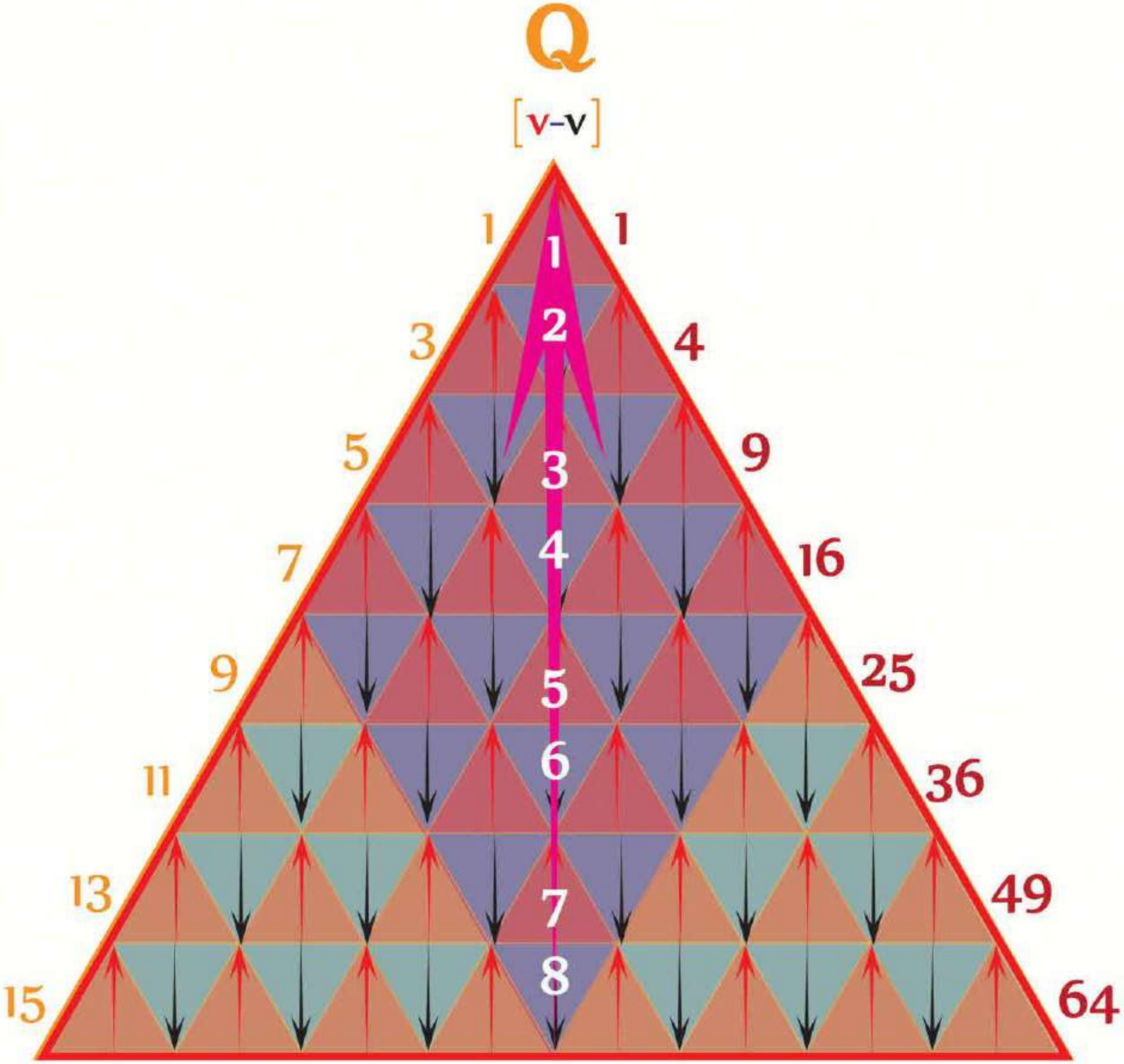
Q
[v-v]
1
1
3
4
5
9
7
16
9
25
11
36
13
49
15
64
1
2
3
4
5
6
7
8

Divergent energy momenta	Convergent energy momenta
1	0
3	1
6	3
10	6
15	10
21	15
28	21
36	28

All charge geometries are comprised of finite equilateral energy momenta quanta

Renormalisation

Renormalization was first developed in quantum electrodynamics (QED) to make sense of infinite integrals in perturbation theory.
The problem of infinities first arose in the classical electrodynamics of point particles in the19th and
later in the calculation of Gravitational fields in General Relativity in the early 20th century.

In QED Infinities must be cancelled

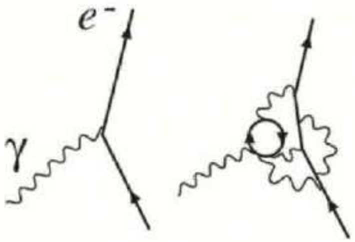

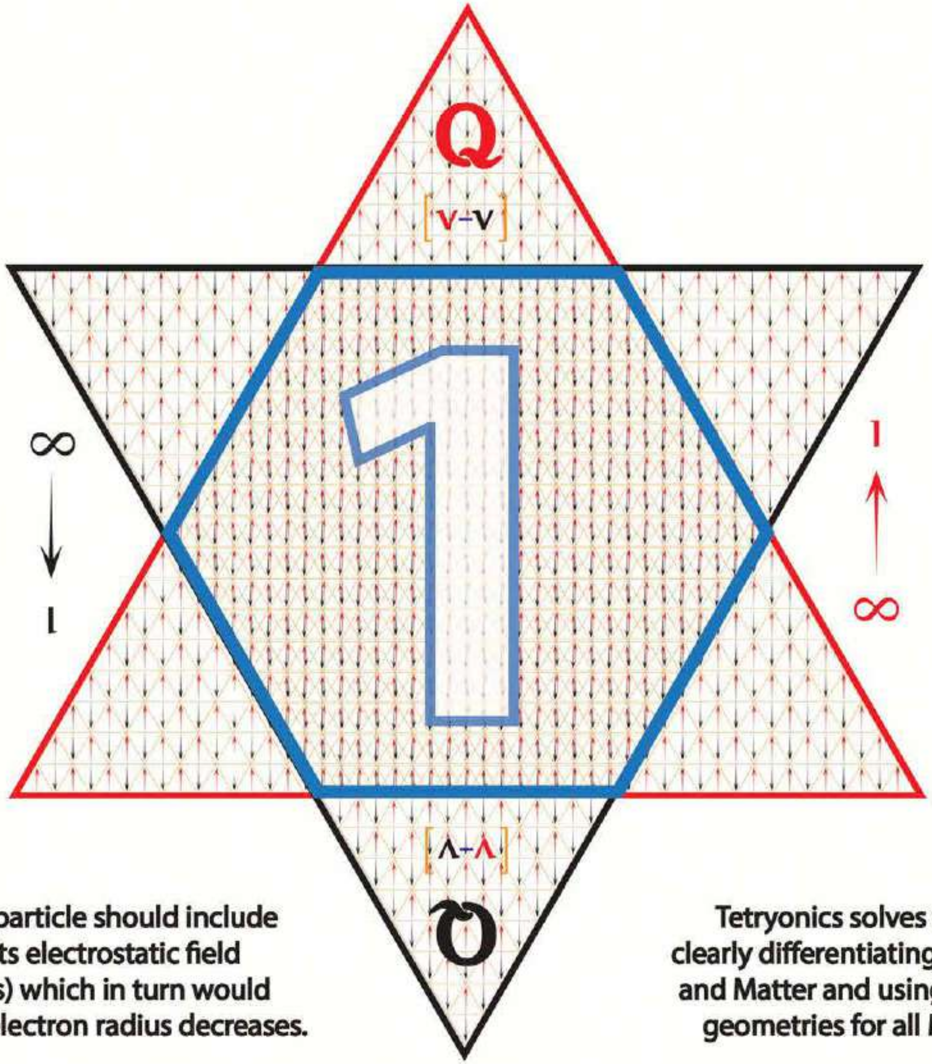

In Tetryonics Infinities do not exist

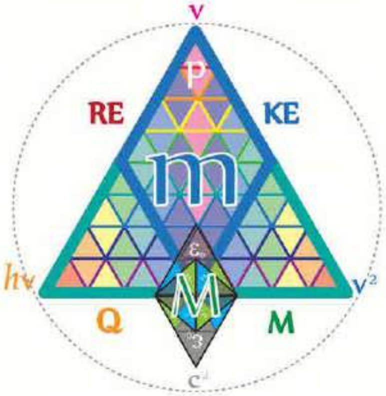

The mass of a charged particle should include the mass-energy in its electrostatic field (Electromagnetic mass) which in turn would approach infinity as the electron radius decreases.

Tetryonics solves the problem by clearly differentiating between EM mass and Matter and using finite equilateral geometries for all Matter in motion

Initially viewed as a suspicious provisional procedure by some of its originators, renormalization was eventually embraced as an important and self-consistent tool in several fields of physics and mathematics.

2D space $[c^2]$

The adjective Cartesian refers to the French mathematician and philosopher René Descartes who developed the coordinate system in 1637

Since then many other coordinate systems have been developed such as the polar coordinates for the plane, and the spherical and cylindrical coordinates for three-dimensional space.

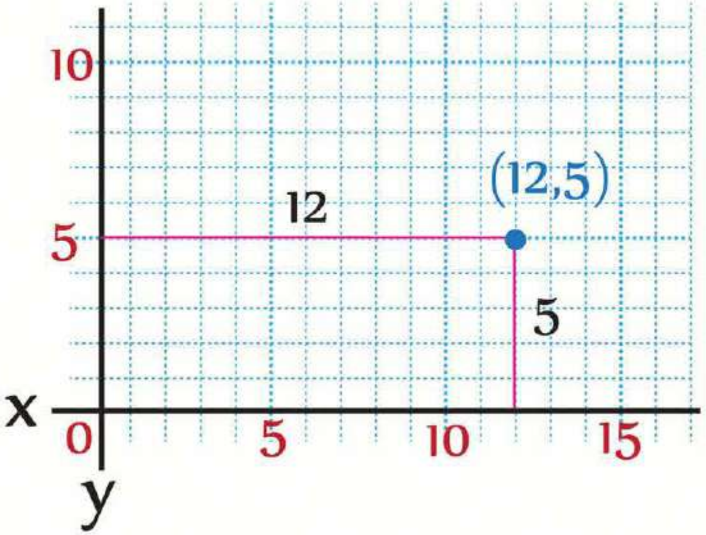

Cartesian coordinates are the foundation of analytic geometry, and provide enlightening geometric interpretations for many other branches of mathematics, such as linear algebra, complex analysis, differential geometry, multivariate calculus, group theory, and more

Mapping 3D spaces using Recti-linear co-ordinates

Cartesian coordinates can be defined as the positions of the perpendicular projections of a point onto the two or more axes, expressed as signed distances from the origin.

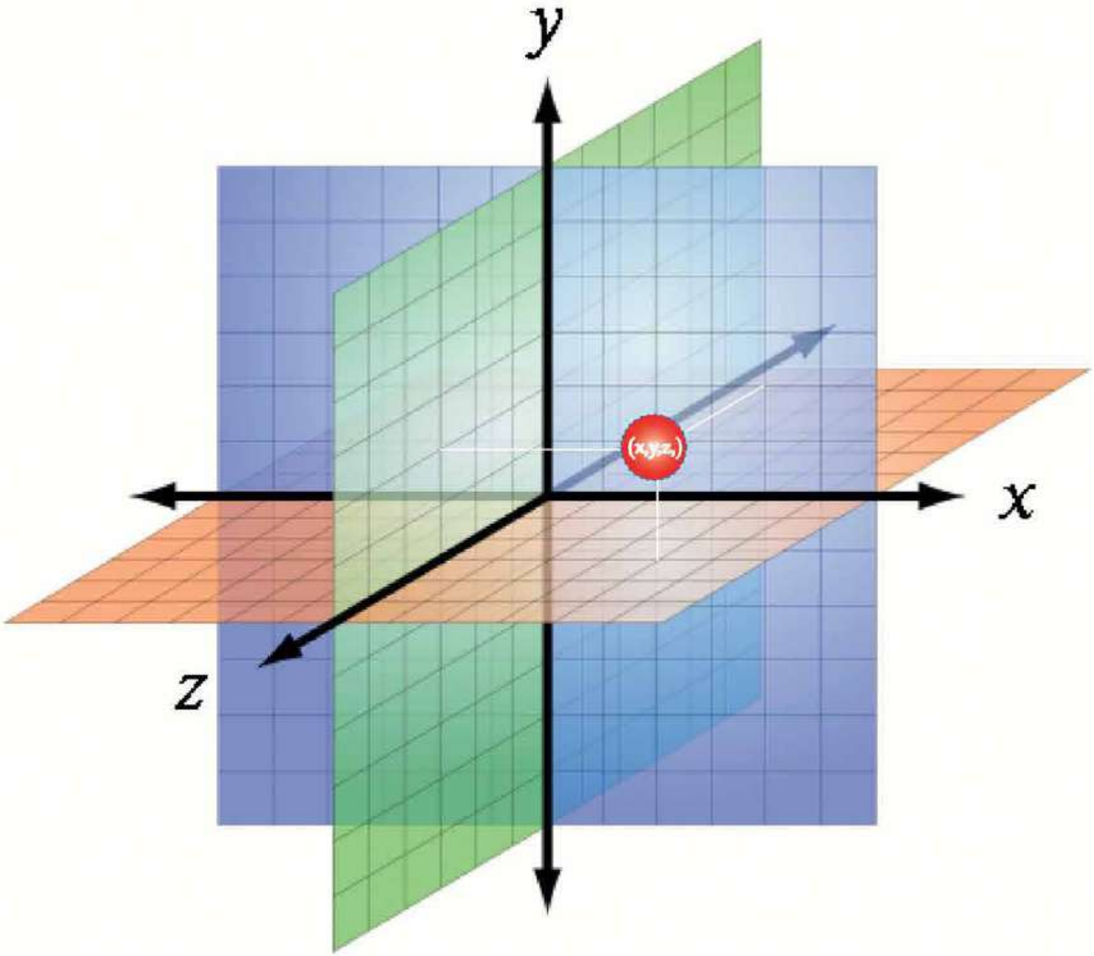

3D Cartesian co-ordinate $[c^3]$ systems are distinct from spherical co-ordinate $[c^4]$ systems

Polar co-ordinates

In mathematics, the polar coordinate system is a two-dimensional co-ordinate system in which each point on a plane is determined by a distance from a fixed point and an angle from a fixed direction.

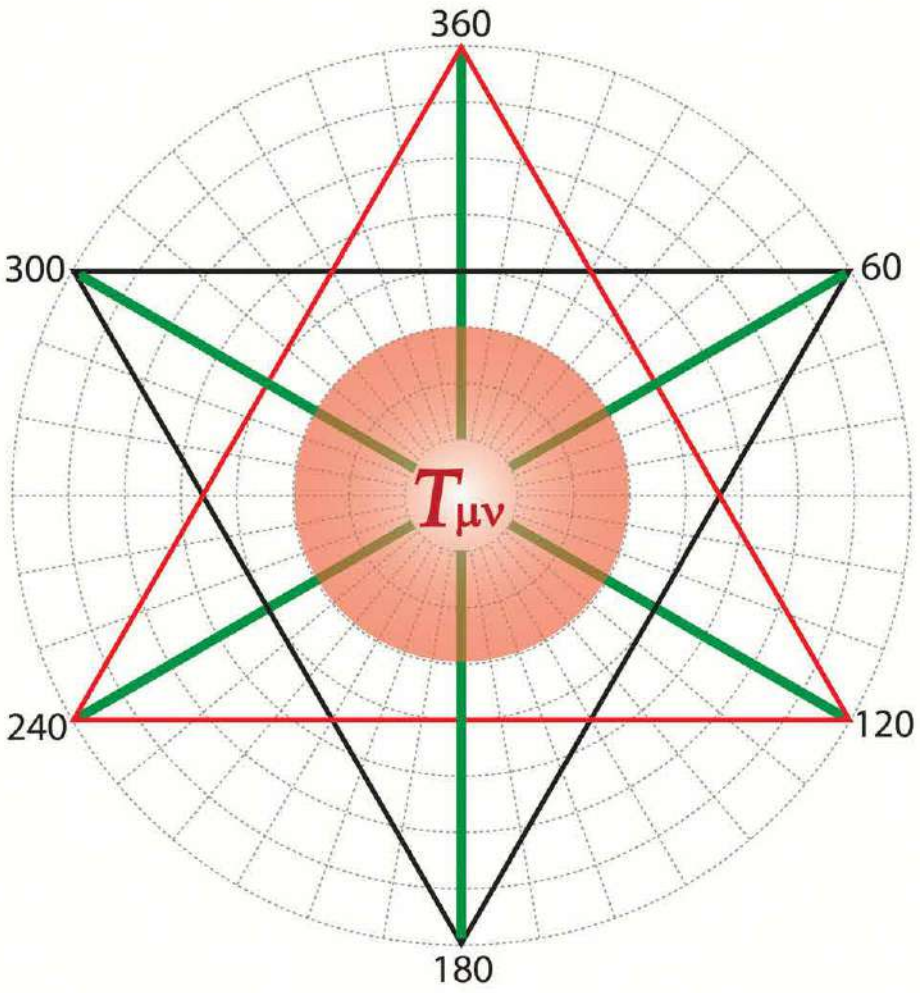

In geometry, curvilinear coordinates are a coordinate system for Euclidean space in which the co-ordinate lines may be curved.

Action Dynamics

Curvilinear co-ordinates may be derived from a set of rectilinear Cartesian coordinates by using a locally invertible transformation that maps one point to another in both systems

Metric Tensors

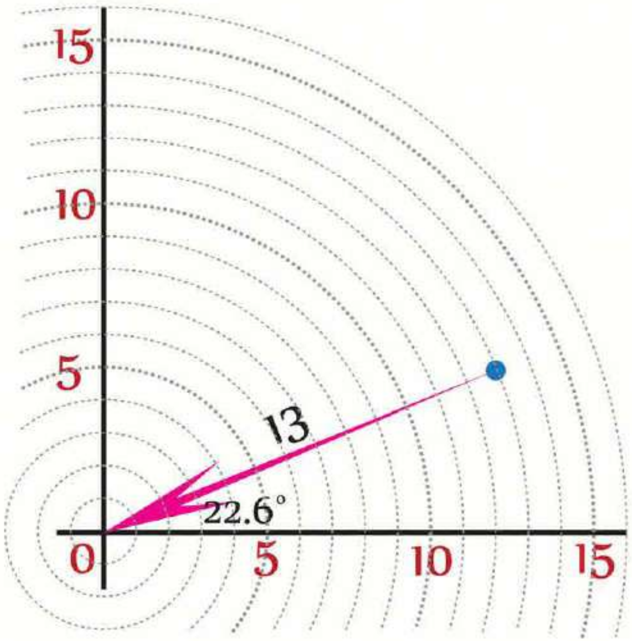

Gravitational acceleration

Polar or curvilinear co-ordinate systems are used extensively by Einstein in his theory of General Relavtivity

Reimannian curved space-time

Copyright ABRAHAM [2008] - All rights reserved

Co-ordinate transformations

There are many different possible coordinate systems for describing geometrical figures and they can all be related to one another.

Such relations are described by coordinate transformations which give formulas for the coordinates in one system in terms of the coordinates in another system

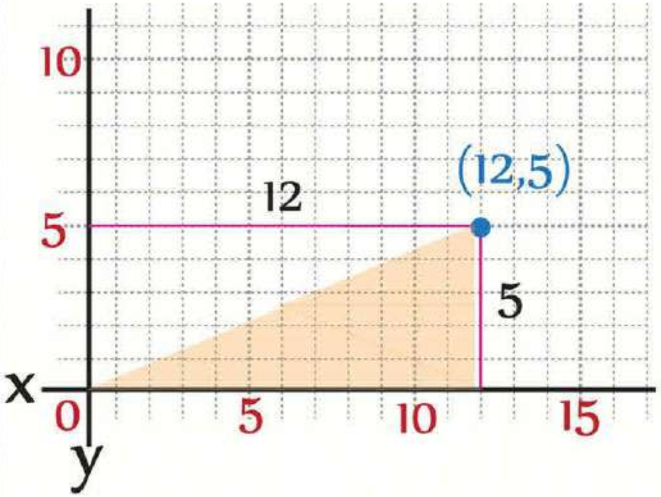

Converting between
Polar and Cartesian
coordinates

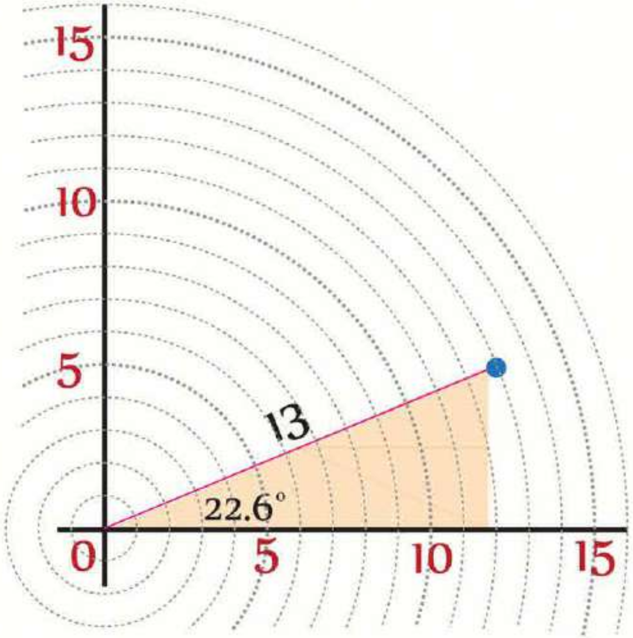

$$\tan\theta = \frac{\text{opposite}}{\text{adjacent}}$$

$$= \tan^{-1}(5/12)$$

$$= 22.619$$

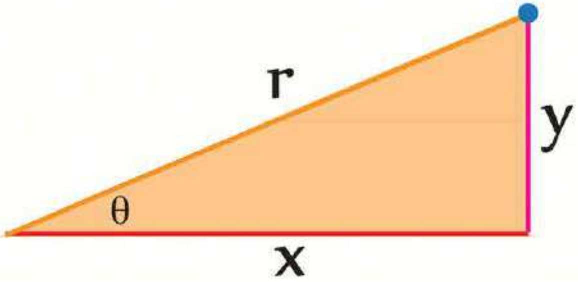

$$r^2 = 12^2 + 5^2$$

$$= \sqrt{169}$$

$$= 13$$

solving for the triangle
using trigometric functions

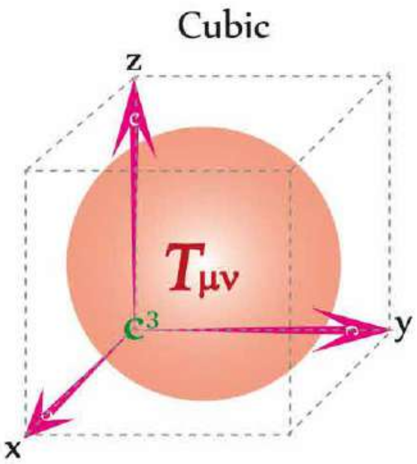
Cubic
z
y
x
$T_{\mu\nu}$
c^3

Spatial co-ordinate systems

Spacetime is any mathematical co-ordinate system or model that combines space and time into a single continuum.

Spacetime is usually interpreted with space as being three-dimensional with time playing the role of a fourth dimension that is different from the spatial dimensions.

From a Euclidean space perspective, the universe has three spatial dimensions and one dimension of time [reflected by quantised angular momentum].

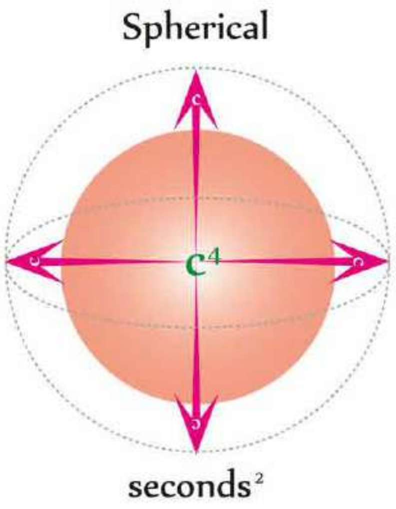
Spherical
c^4
seconds²

Tetryonics maps spatial co-ordinates through the momenta vectors of equilateral Energy

Cartesian Space-Time

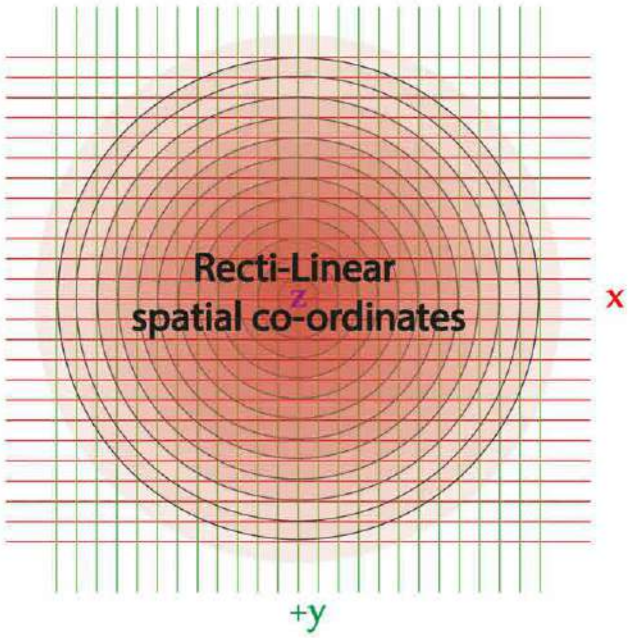
Recti-Linear
spatial co-ordinates
z
x
+y

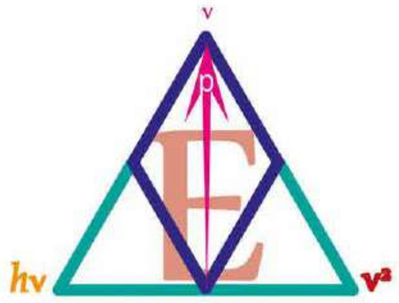

EM mass-ENERGY momenta are equilateral geometries

Mapping equilateral Energy geometries onto recti or curvi-linear spatial co-ordinate systems introduces mathematical complexity to a otherwise simplistic geometry for all EM mass-Energy-Matter interactions

Tetryonic Space-Time

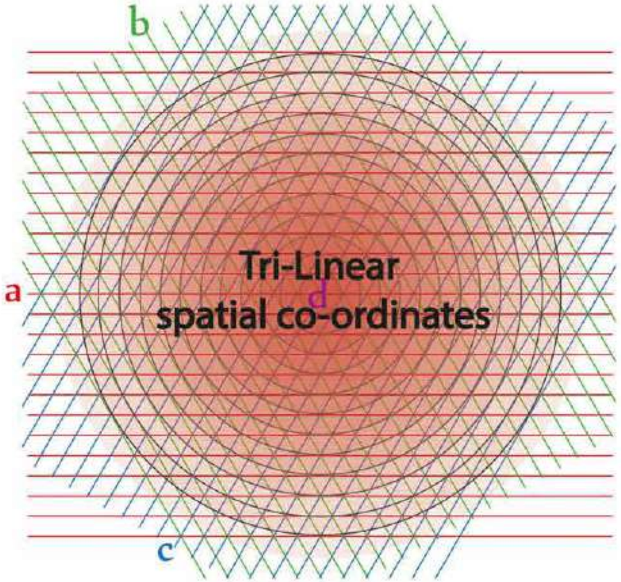
b
a
c
d
Tri-Linear
spatial co-ordinates

In physics spatial co-ordinates to date have been based on Cartesian co-ordinates when in fact Energy momenta follow a Tri-Linear co-ordinate geometry

vectors

1 Dimensional velocity

Euclidean

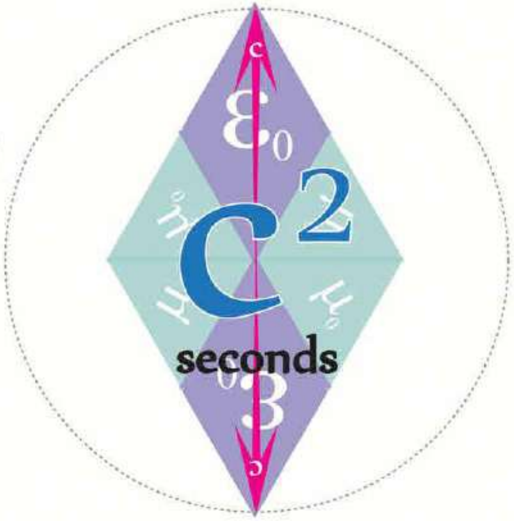

2 Dimensions velocity squared

$$m = \frac{E}{c^2}$$

mass-energy equivalence

mass-Matter

$$\frac{m}{c^2} = M$$

equivalence

$$T_{\mu\nu}$$

energy-Matter equivalence

$$M = \frac{E}{c^4}$$

GR fails differentiate betweem EM mass-ENERGY-Matter unlike Tetryonic theory

Cartesian co-ordinates

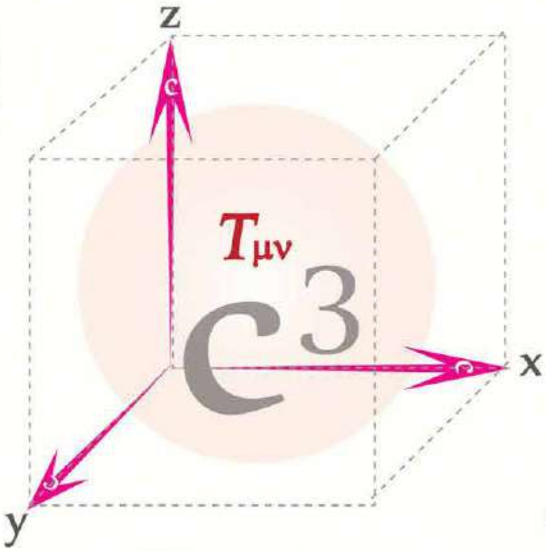

3 Dimensions velocity cubed

Spherical

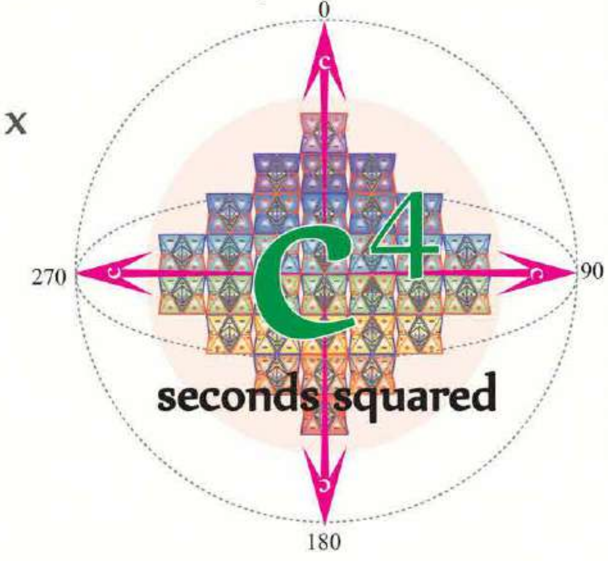

3 Dimensions quadrature velocity

Space-time co-ordinates

The propagation of Energy momenta forms distinct spatial co-ordinate systems

Where angles are typically measured in degrees (°) or radians (rad), where 360° = 2π rad = τ radians.

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic co-ordinate systems

Energy per second

m

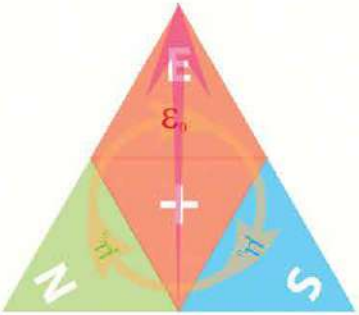

2D mass-energy geometries

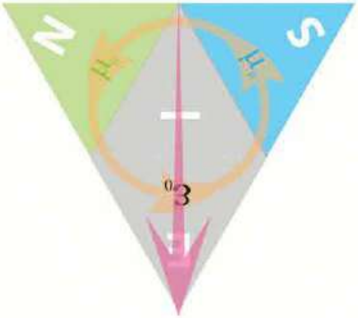

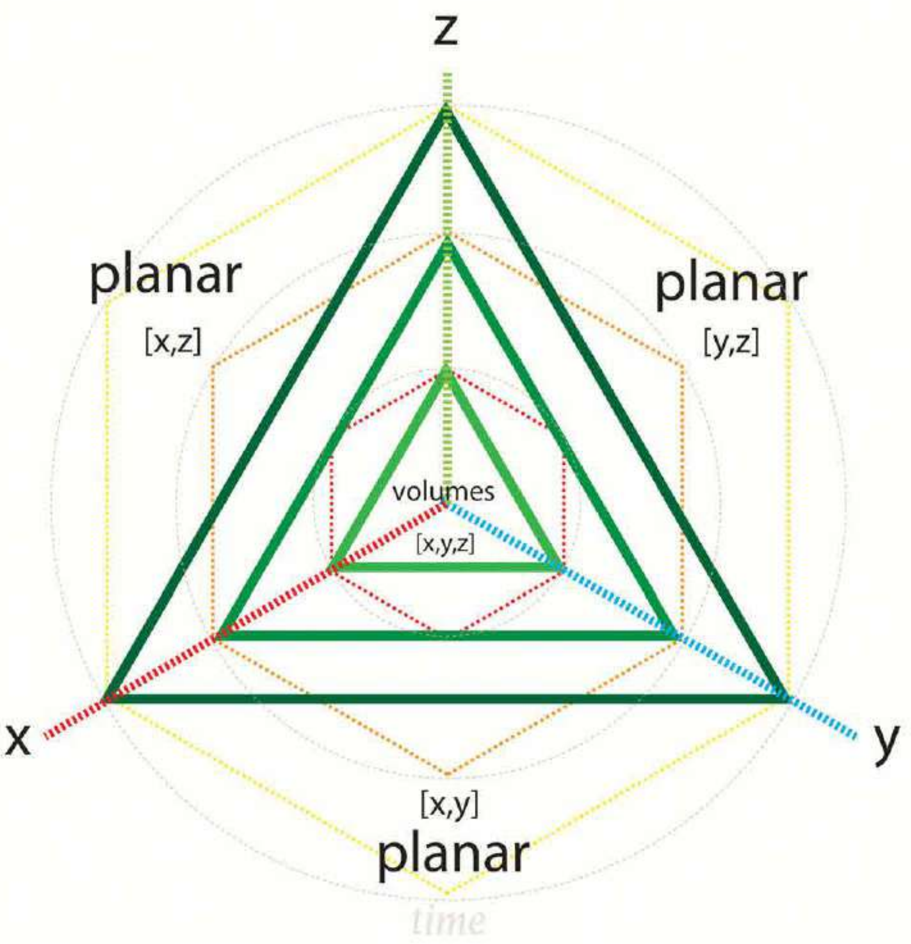
z
planar
[x,z]
planar
[y,z]
volumes
[x,y,z]
x
y
[x,y]
planar
time

Planar mass-energy geometries have no z-components

Differentiation between 2D mass-energy & 3D Matter is key to extending our understanding of physics

Matter topologies have z-components

Energy per second squared

M

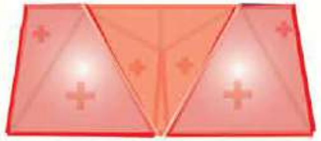

Proton

3D Matter topologies

Neutron

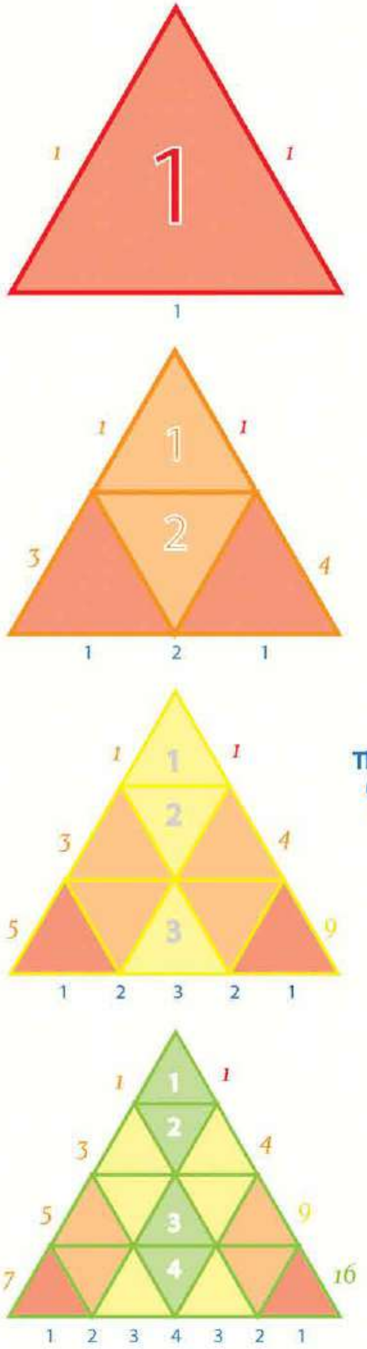
1
1 1
1
1
2
1 1
3 4
1 2 1
1
2
3
1 1
3 4
5 9
1 2 3 2 1
1
2
3
4
1 1
3 4
5 9
7 16
1 2 3 4 3 2 1

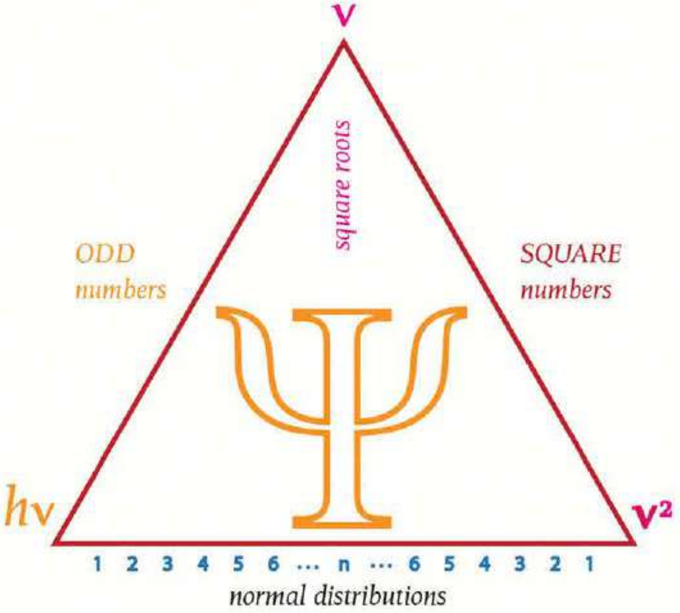
v
square roots
ODD
numbers
SQUARE
numbers
hv
v²
1 2 3 4 5 6 ... n ... 6 5 4 3 2 1
normal distributions

Quantum probability distributions

The equilateral geometry and distribution of quantised Energy momenta provides the basis for all statistical probabilities in Quantum mechanics, thermodynamic & information entropy - including a solution to Heisenberg's Uncertainty Principle thus paving the way forward for a new understanding, and manipulation of physical phenomena at the quantum level

1

1 2 1

1 2 3 2 1

1 2 3 4 3 2 1

1 2 3 4 5 4 3 2 1

1 2 3 4 5 6 5 4 3 2 1

1 2 3 4 5 6 7 6 5 4 3 2 1

1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Normal distributions are extremely important in statistics, and are often used in the natural and social sciences for real-valued random variables whose distributions are not known

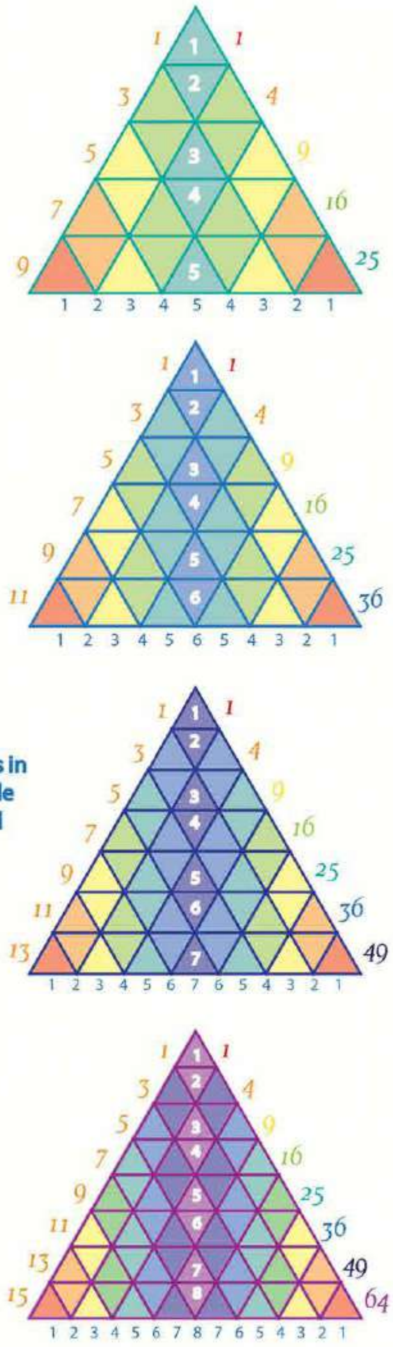
1
2
3
4
5
1 1
3 4
5 9
7 16
9 25
1 2 3 4 5 4 3 2 1
1
2
3
4
5
6
1 1
3 4
5 9
7 16
9 25
11 36
1 2 3 4 5 6 5 4 3 2 1
1
2
3
4
5
6
7
1 1
3 4
5 9
7 16
9 25
11 36
13 49
1 2 3 4 5 6 7 6 5 4 3 2 1
1
2
3
4
5
6
7
8
1 1
3 4
5 9
7 16
9 25
11 36
13 49
15 64
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Copyright ABRAHAM [2008] - All rights reserved

Quantum Probability Distributions

The normal distribution is a probability distribution.
It is also called Gaussian distribution because it was discovered by Carl Friedrich Gauss.

x/n^2

Probabilities are
the square of the
Amplitude

n

$\Sigma\ x/n^2$

All probabilities
are re-normalisable
and sum to Unity

1

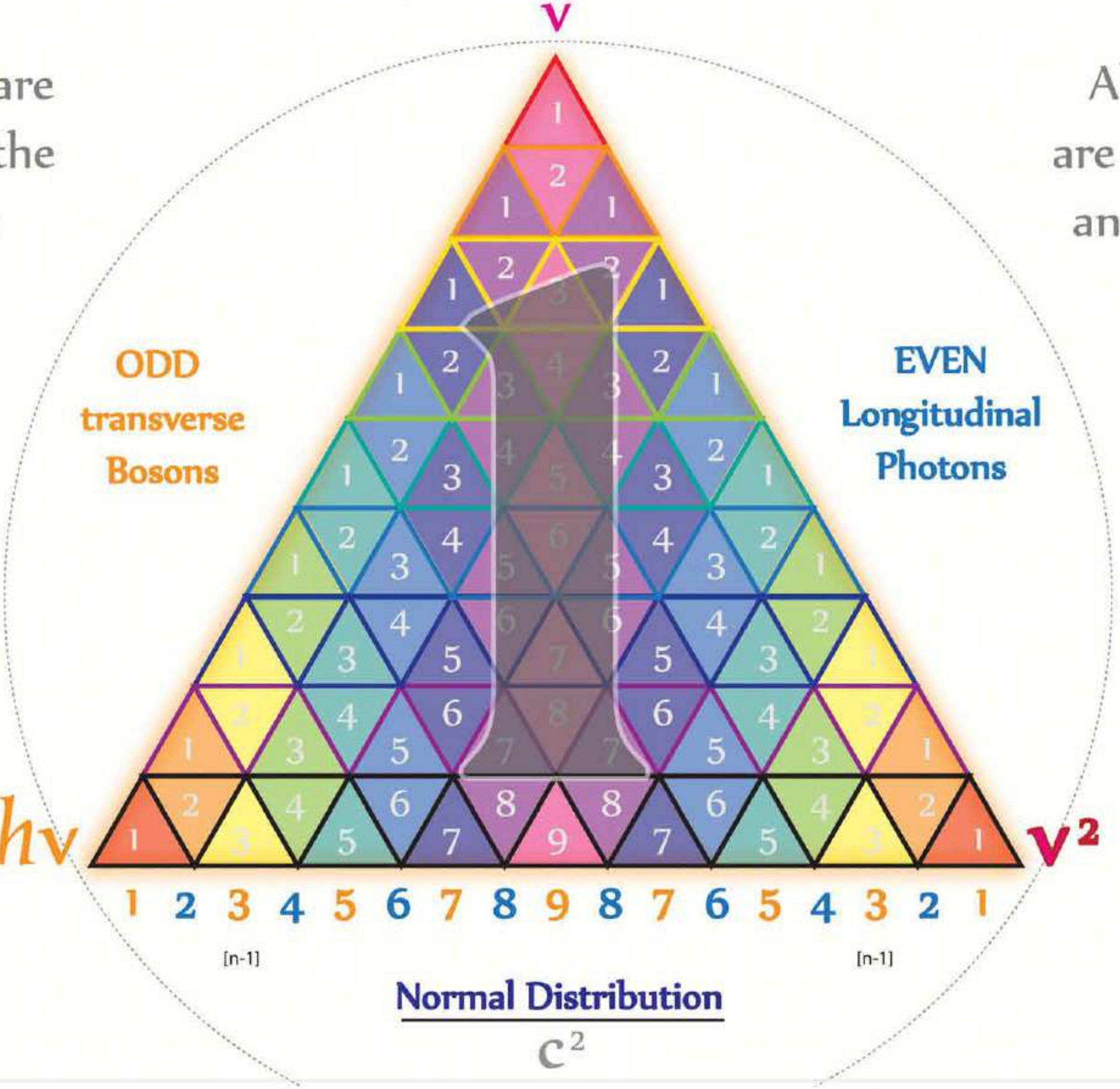
1
2
1 1
2 3 2
1 1
2 4 2
1 3 3 1
2 4 4 2
1 3 5 3 1
2 4 6 4 2
1 3 5 5 3 1
2 4 6 6 4 2
1 3 5 7 5 3 1
2 4 6 8 6 4 2
1 3 5 7 7 5 3 1
2 4 6 8 8 6 4 2
1 3 5 7 9 7 5 3 1
1
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1
[n-1]
[n-1]

square root momenta

v

ODD
transverse
Bosons

EVEN
Longitudinal
Photons

Planck
quanta

hv

v^2

Square
energies

Normal Distribution

c^2

Quanta Distributions

Quantum Levels

transverse bosons

$$2h\nu = E = hf$$

Quantum Probabilities

longitudinal photons

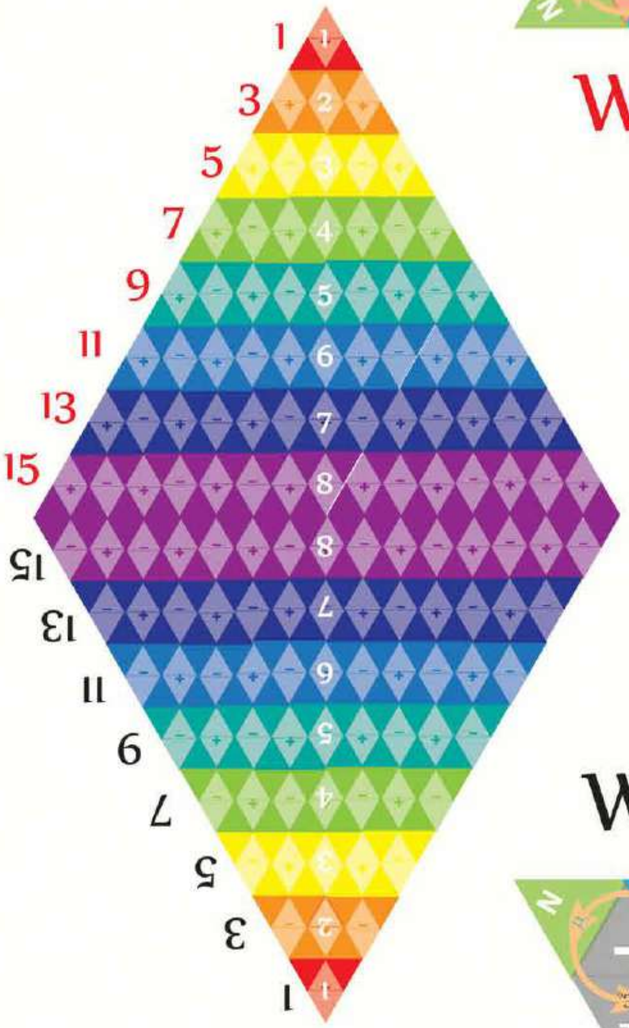
1
3
5
7
9
11
13
15
15
13
11
9
7
5
3
1
1
2
3
4
5
6
7
8
8
7
6
5
4
3
2
1

W+

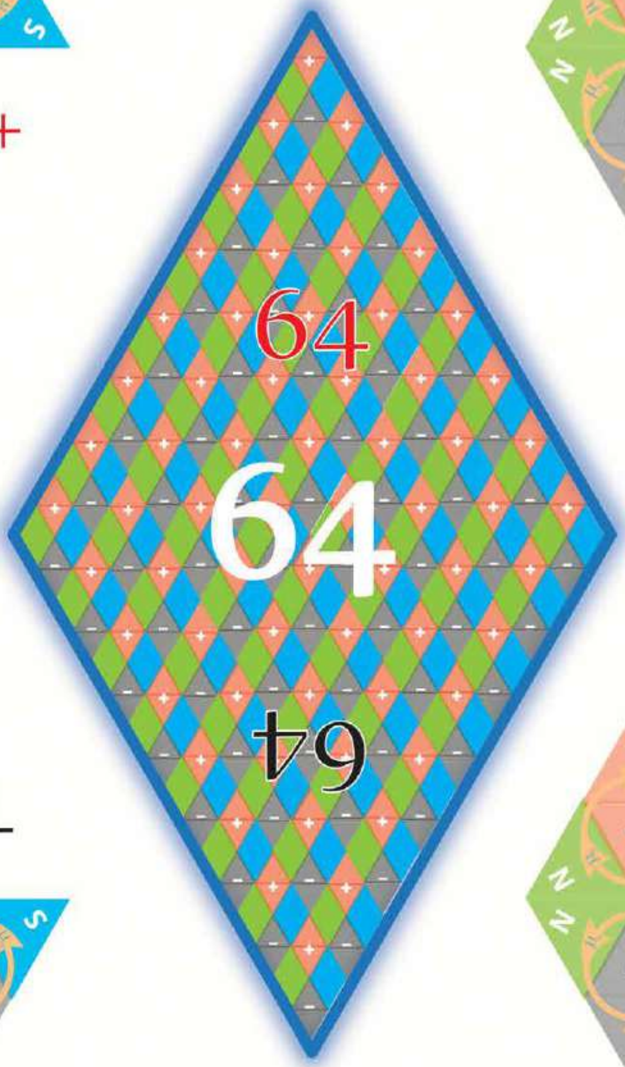

W-

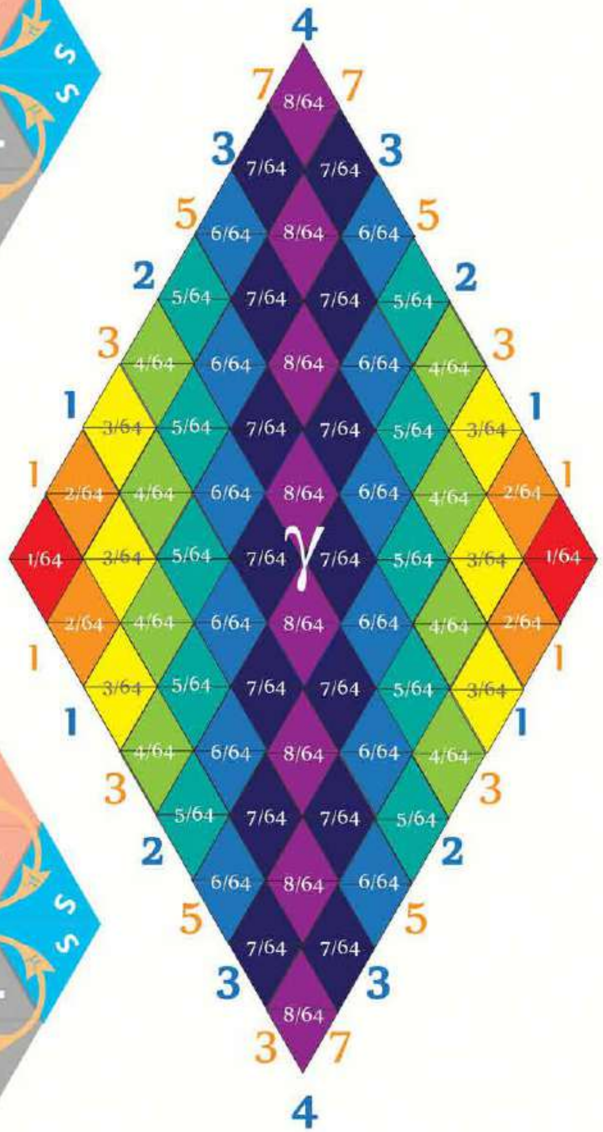
4
7 7
3 3
5 5
2 2
3 3
1 1
1 1
1 1
1 1
3 3
2 2
5 5
3 3
3 7
4
8/64
7/64 7/64
6/64 8/64 6/64
5/64 7/64 7/64 5/64
4/64 6/64 8/64 6/64 4/64
3/64 5/64 7/64 7/64 5/64 3/64
2/64 4/64 6/64 8/64 6/64 4/64 2/64
1/64 3/64 5/64 7/64 γ 7/64 5/64 3/64 1/64
2/64 4/64 6/64 8/64 6/64 4/64 2/64
3/64 5/64 7/64 7/64 5/64 3/64
4/64 6/64 8/64 6/64 4/64
5/64 7/64 7/64 5/64
6/64 8/64 6/64
7/64 7/64
8/64

Wavefunctions

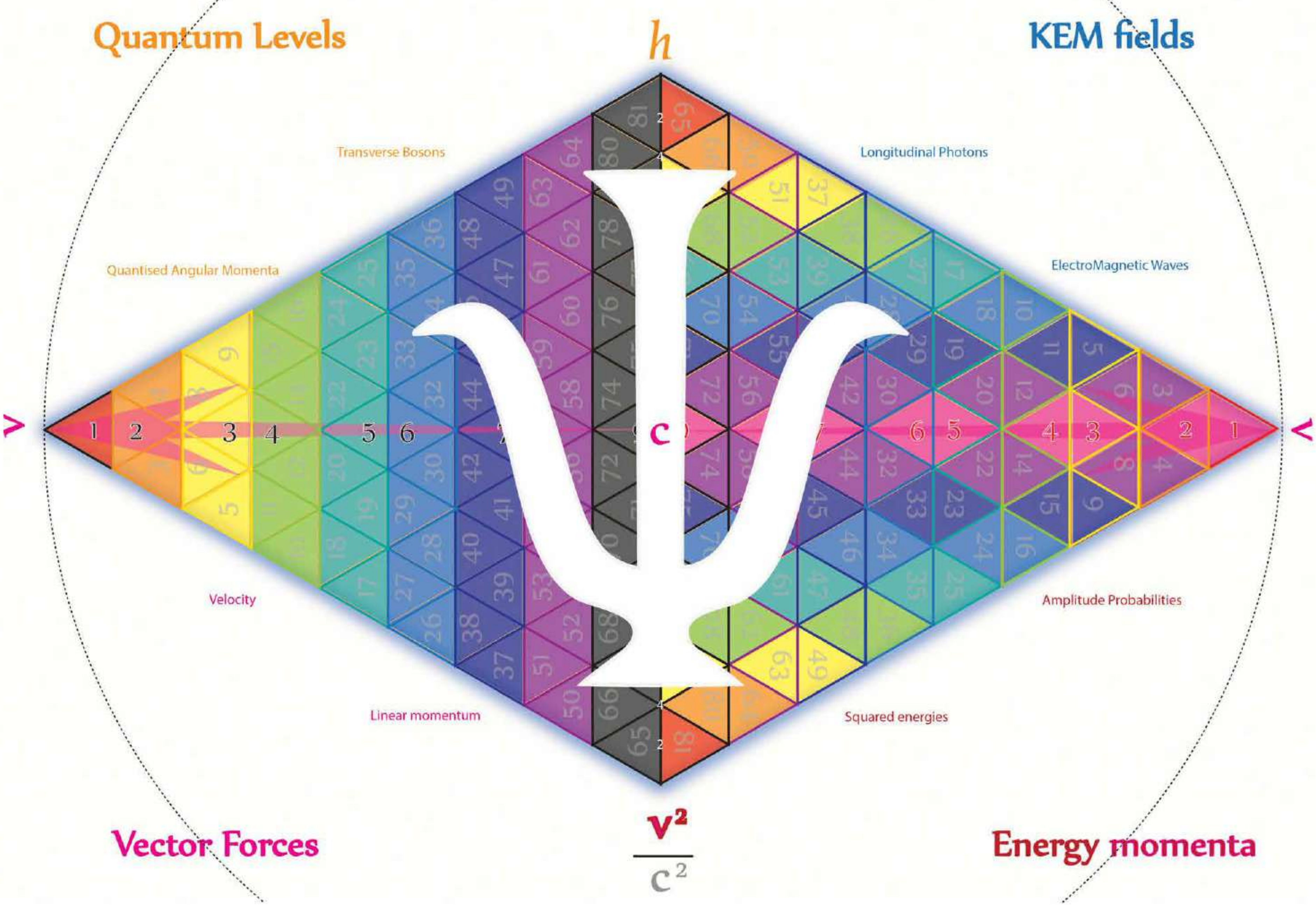
Quantum Levels
h
KEM fields
Transverse Bosons
Longitudinal Photons
Quantised Angular Momenta
ElectroMagnetic Waves
v
1 2 3 4 5 6 7
c
7 6 5 4 3 2 1
v
Velocity
Amplitude Probabilities
Linear momentum
Squared energies
Vector Forces
v^2
c^2
Energy momenta

EM waveforms

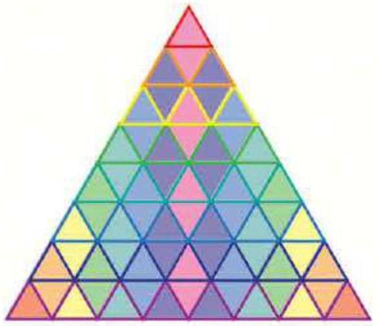

All EM waveforms can be measured by either their Transverse EM masses [Bosons] or their Longitudinal EM masses [Photons]

PHOTONS

EM waves Planck quanta

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

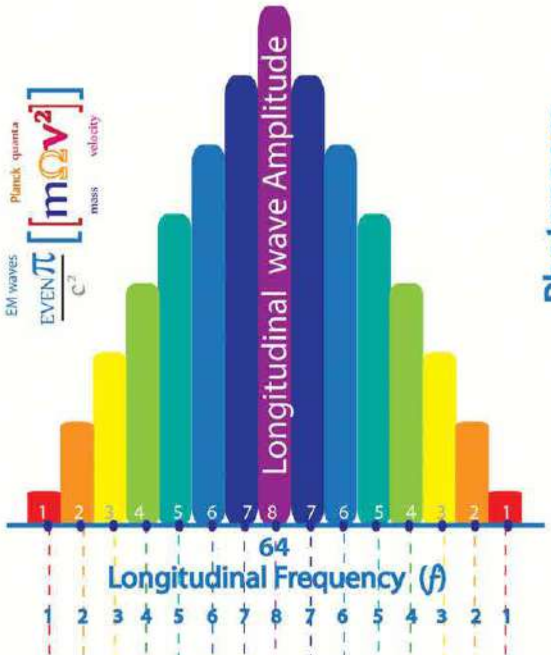

Longitudinal wave Amplitude
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
64
Longitudinal Frequency (f)
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Photons are longitudinal quanta

vs.

BOSONS

Bosons Planck quanta

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

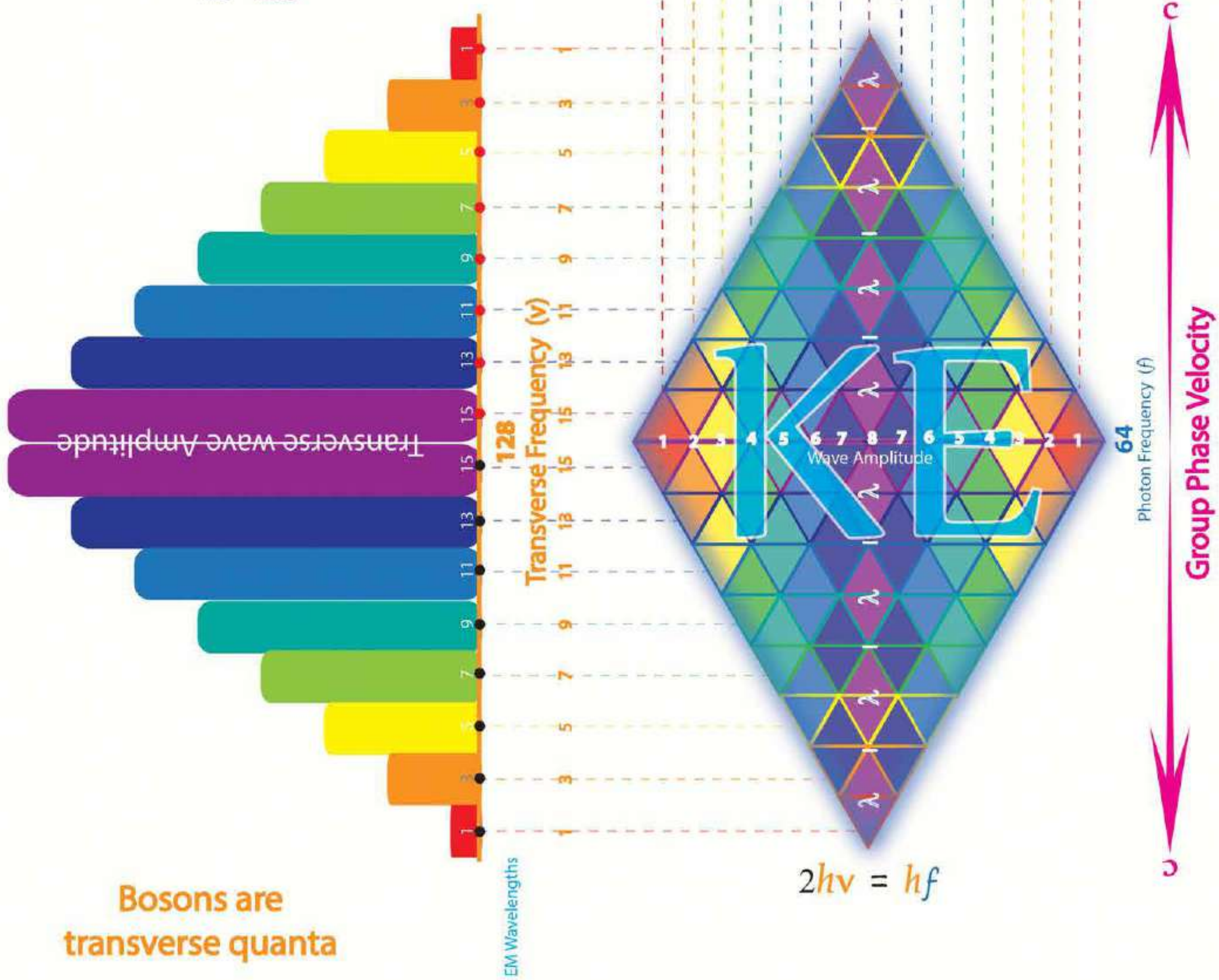

Transverse wave Amplitude
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
128
Transverse Frequency (v)
EM Wavelengths
λ
KE
Wave Amplitude
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
64
Photon Frequency (f)
2hv = hf
c
c
Group Phase Velocity
Bosons are transverse quanta

Copyright ABRAHAM [2008] - All rights reserved

Quantum Energy distributions

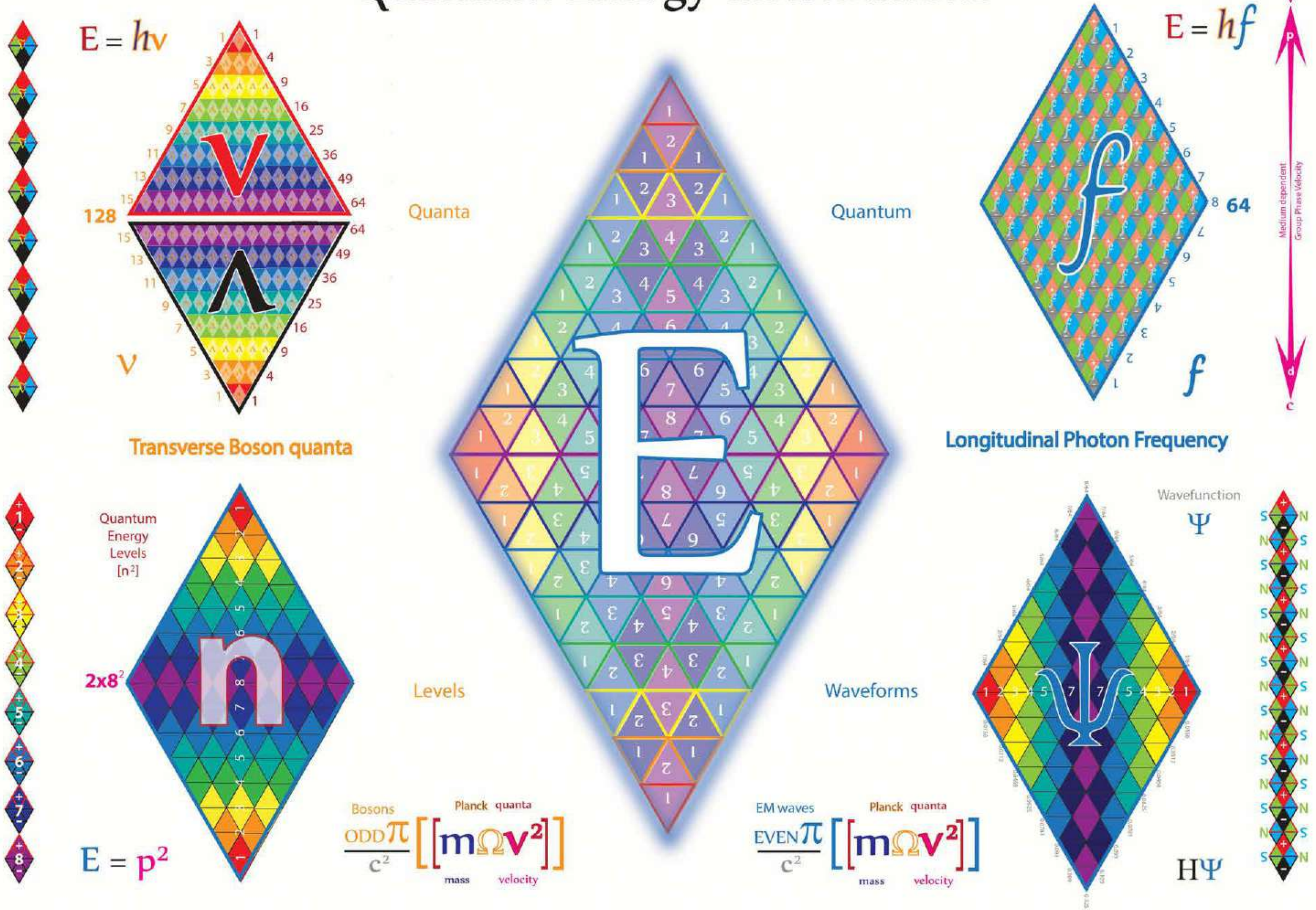
E = hν
1
4
9
16
25
36
49
64
128
ν
Λ
ν
Quanta
Transverse Boson quanta
Quantum Energy Levels [n²]
2x8²
n
E = p²
Levels
Bosons
ODD π / c²
Planck quanta
[[mΩv²]]
mass
velocity
E
Quantum
E = hf
8 64
f
f
Longitudinal Photon Frequency
Medium dependent Group Phase Velocity
v
p
d
c
Wavefunction
Ψ
Waveforms
EM waves
EVEN π / c²
Planck quanta
[[mΩv²]]
mass
velocity
HΨ
S
N

Copyright ABRAHAM [2008] - All rights reserved

Normal Distributions

The Gaussian distribution sometimes informally called the bell curve.

Pierre de Fermat

Pierre de Fermat is given credit
for early developments that led to
infinitesimal calculus.

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

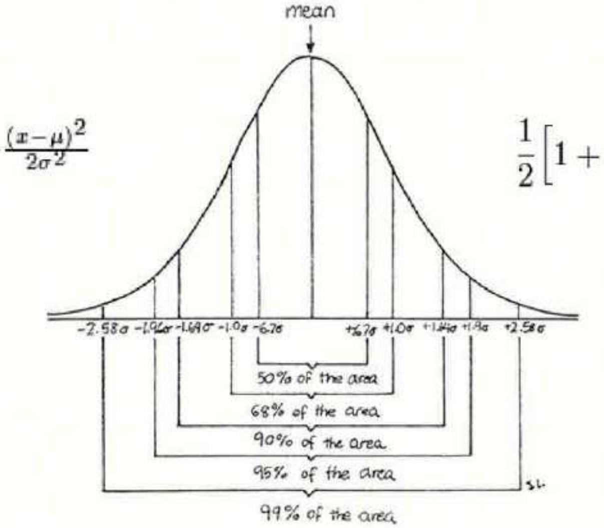
mean
-2.58σ
-1.96σ
-1.69σ
-1.0σ
-.67σ
+.67σ
+1.0σ
+1.64σ
+1.9σ
+2.58σ
50% of the area
68% of the area
90% of the area
95% of the area
99% of the area

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$

Leonhard Euler

Leonhard Euler developed a formula
which links complex exponentiation
with trigonometric functions

**A bell shaped curve defines the standard normal distribution,
in which the probability of observing a point is greatest near the average,
and declines rapidly as one moves away from the mean.**

all ODD numbers are

$$a^2 - b^2 = [a-b].[a+b]$$

the difference of two squares

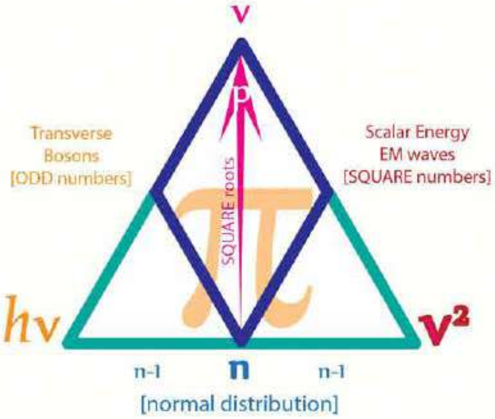
v
Transverse
Bosons
[ODD numbers]
Scalar Energy
EM waves
[SQUARE numbers]
SQUARE roots
hv
v²
n-1
n
n-1
[normal distribution]

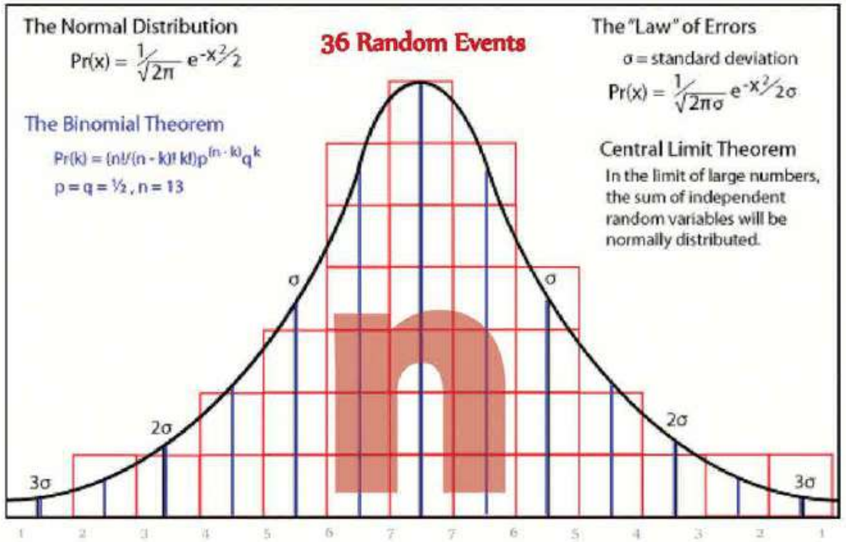

The Normal Distribution

$\Pr(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

36 Random Events

The "Law" of Errors

σ = standard deviation

$\Pr(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/2\sigma}$

The Binomial Theorem

$\Pr(k) = (n!/(n-k)!\,k!)p^{(n-k)}q^{k}$

$p = q = ½ , n = 13$

Central Limit Theorem

In the limit of large numbers,
the sum of independent
random variables will be
normally distributed.

$$e^{i\pi} = -1$$

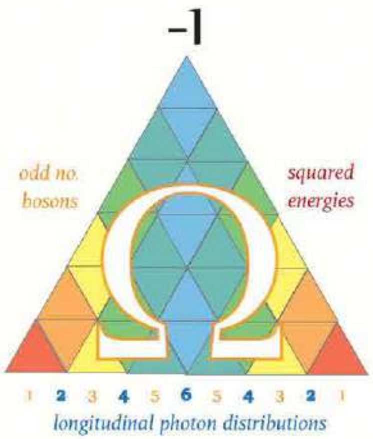
-1
odd no.
bosons
squared
energies
1
2
3
4
5
6
5
4
3
2
1
longitudinal photon distributions

In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function

Copyright ABRAHAM [2008] - All rights reserved

Fundamental theorem of Energy momenta

A nth level scalar energy momenta waveform has exactly n linear momentum in unit circle co-ordinate systems
(with Longitudinal and Transverse equilateral Planck waveforms being orthogonal to each other)

scalar energies

E

mv^2

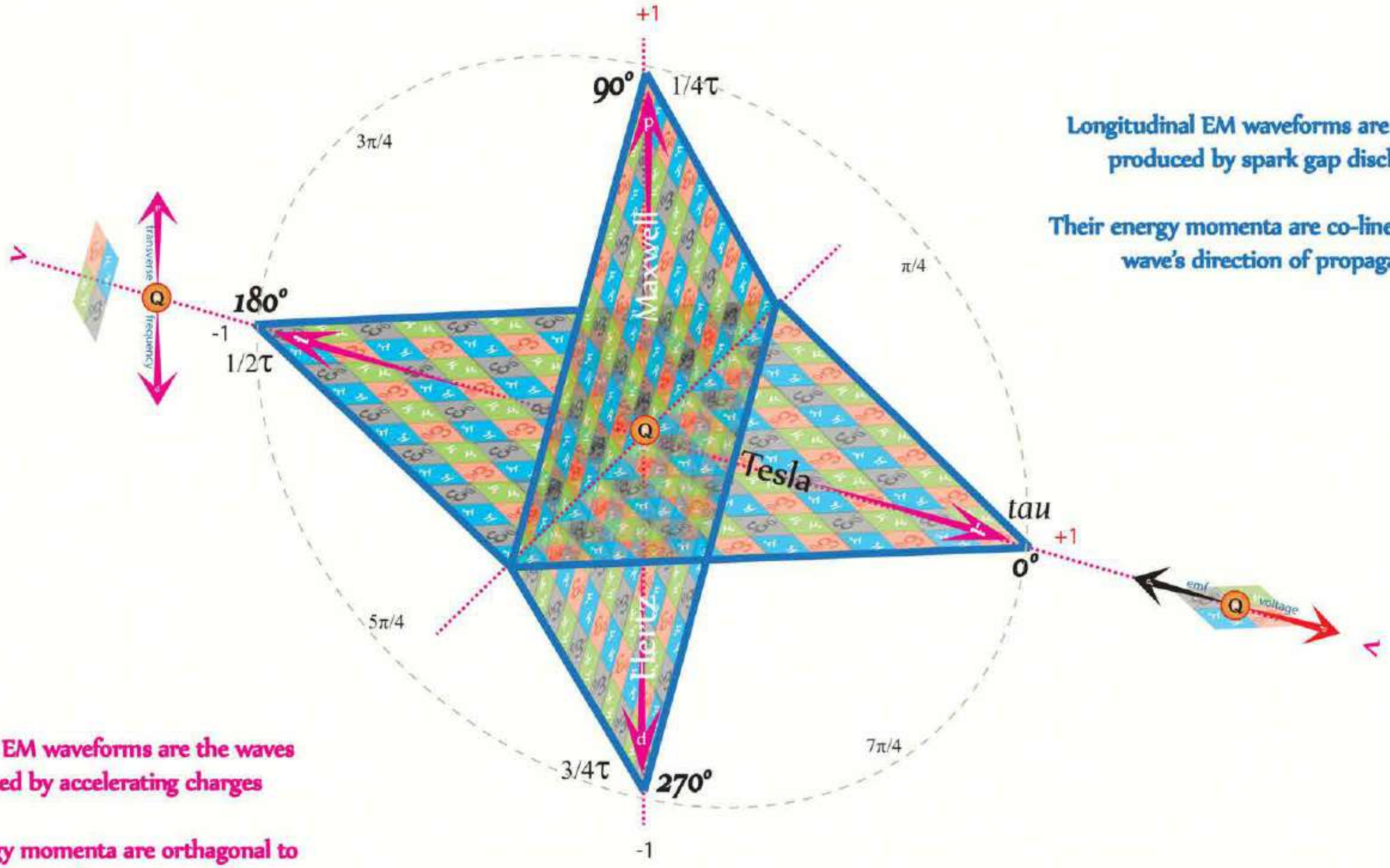
+1
90°
1/4τ
3π/4
π/4
Maxwell
transverse
frequency
Q
180°
-1
1/2τ
Q
Tesla
tau
+1
0°
emf
Q
voltage
5π/4
7π/4
Hertz
3/4τ
270°
-1

Longitudinal EM waveforms are the waves produced by spark gap discharges

Their energy momenta are co-linear with the wave's direction of propagation

Transverse EM waveforms are the waves produced by accelerating charges

Their energy momenta are orthogonal to the direction of wave propagation

p

mv

linear momentum

[Mathematics] *i* or *j* *[Electrical theory]*

square root of negative one

Velocity and Time dependent EM fields

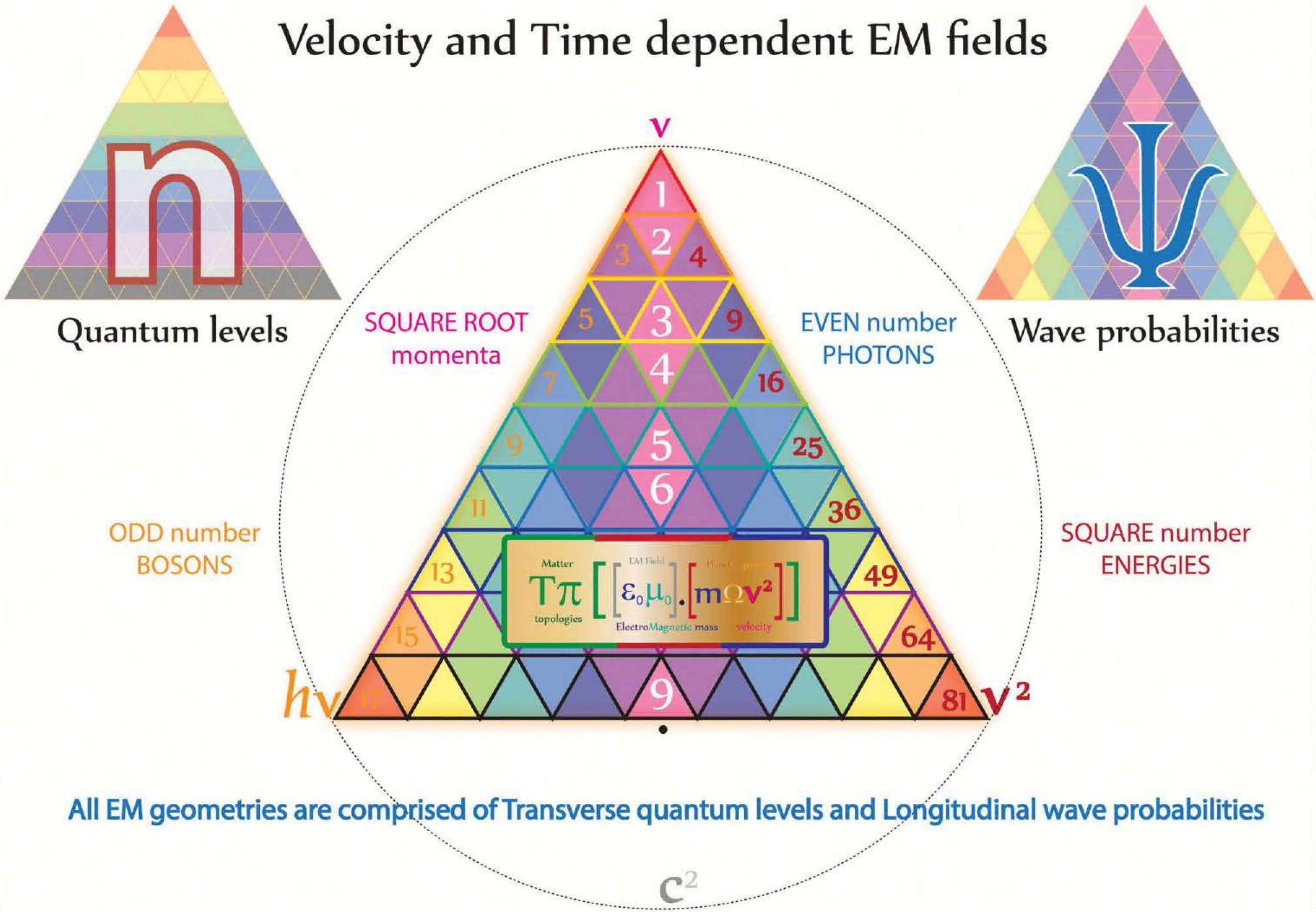

Quantum levels

Wave probabilities

SQUARE ROOT momenta

EVEN number PHOTONS

ODD number BOSONS

SQUARE number ENERGIES

Matter

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

topologies

EM Field

ElectroMagnetic mass

velocity

All EM geometries are comprised of Transverse quantum levels and Longitudinal wave probabilities

Matrices

Matrices are a key tool in linear algebra.
One use of matrices is to represent linear transformations,
which are higher-dimensional analogs of linear functions
of the form f(x) = cx, where c is a constant; matrix multiplication
corresponds to composition of linear transformations.

Further developing equilateral Matrices and tensor mathematics to
reflect the 2D geometry of EM , KEM and GEM quantum fields,
along with the geometric quantisation of mass-energiy momenta
and their energy distributions allows for field interactions
to be accurately visualised and modelled

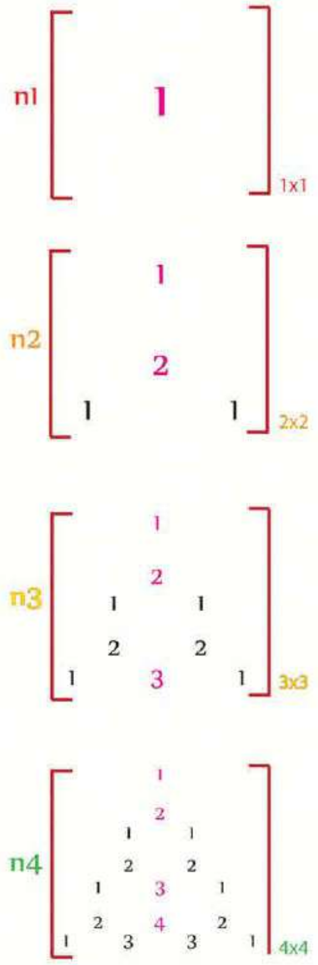
n1
1
1x1
n2
1
2
1 1
2x2
n3
1
2
1 1
2 2
1 3 1
3x3
n4
1
2
1 1
2 2
1 3 1
2 4 2
1 3 3 1
4x4

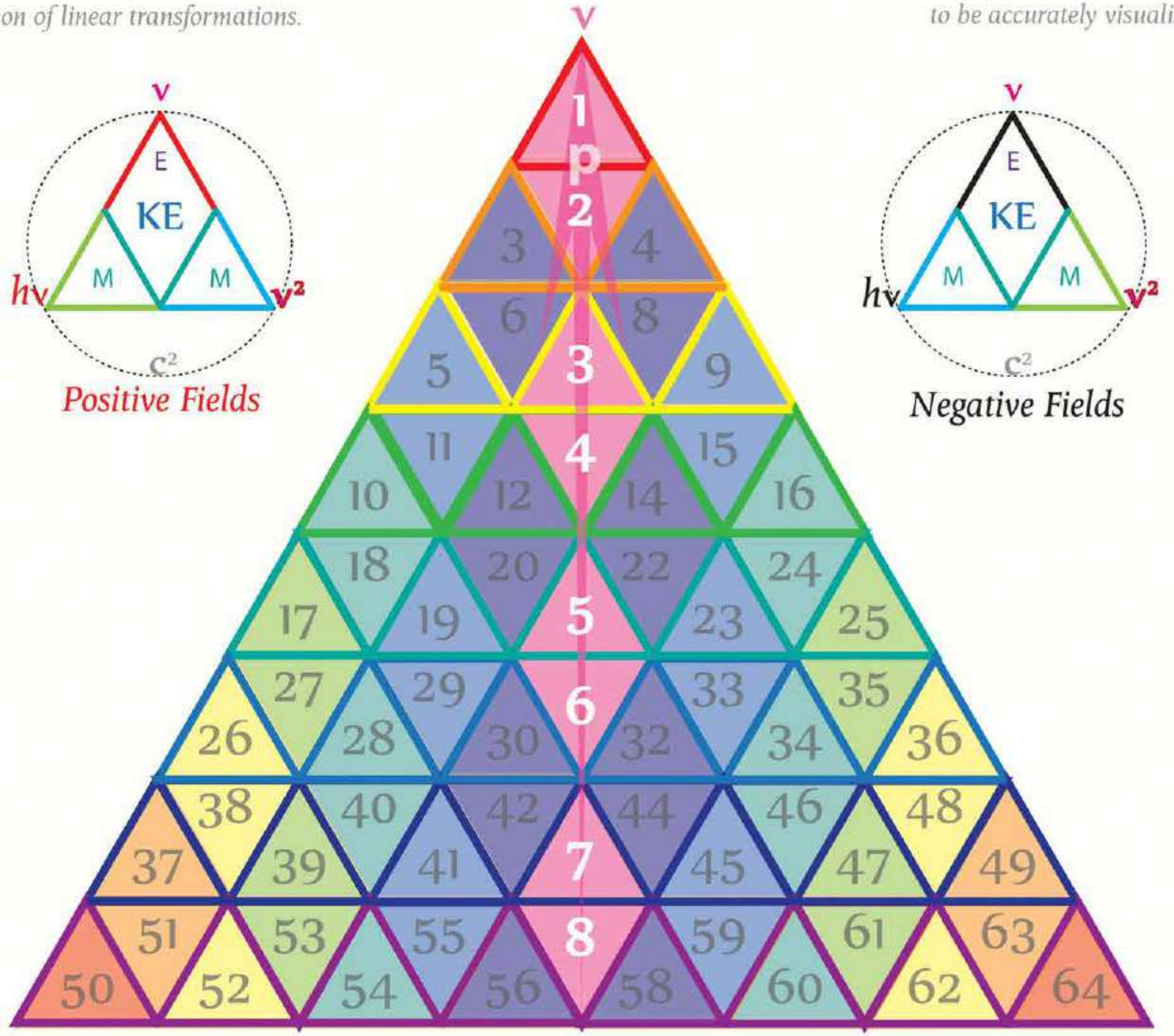
v
E
KE
M
M
hv
v²
c²
Positive Fields
v
E
KE
M
M
hv
v²
c²
Negative Fields
v
1
p
2
3
4
6
8
5
3
9
11
4
15
10
12
14
16
18
20
22
24
17
19
5
23
25
27
29
33
35
26
28
30
6
32
34
36
38
40
42
44
46
48
37
39
41
7
45
47
49
51
53
55
8
59
61
63
50
52
54
56
58
60
62
64

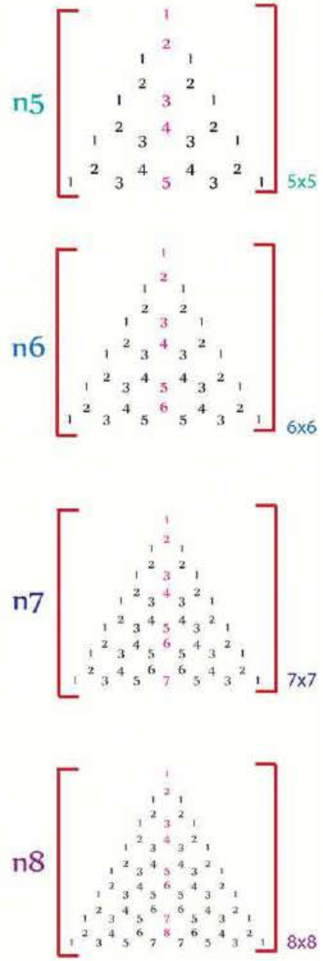
n5
5x5
n6
6x6
n7
7x7
n8
8x8

Tensors are geometric entities introduced into mathematics and physics to extend the notion of scalars, geometric vectors, and matrices to increasingly higher orders.

Modifying Square matrices to reflect the equialteral geometries of Tetryonic fields allows for the accurate geometric modelling of all Scalar & Vector fields along with their varied intrinsic quantum energies and physical properties

Energy momenta Tensors

momenta
(a property of Energy)
is converative

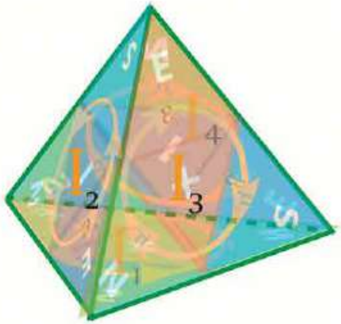

$$p_M \longrightarrow (E, p1, p2, p3, p4).$$

All standing-wave Matter topologies can be modelled using its Tetryonic charge energy momenta Tensors

with an additional Kinetic EM energy-momenta tensor required for Matter in motion

$$p_{KEM} \longrightarrow (E, p5).$$

Matter
(a geometric property)
is NOT conservative

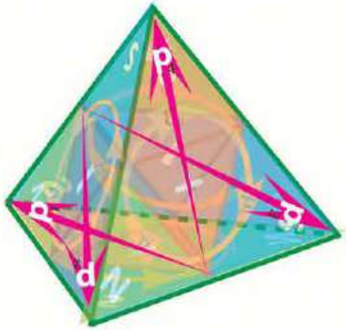

2D+1 [SR] mass-energy momenta can be folded into 3D+1 [GR] Matter
that can be modelled using 4 Energy-momenta tensors

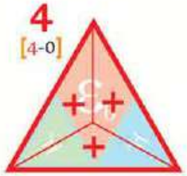

Positive standing-wave
Matter

$$E = \sum_{\text{all fascia}} E$$

*Energy of a massive particle
is the total of all Planck quanta
(compton frequency) in a 3D geometry*

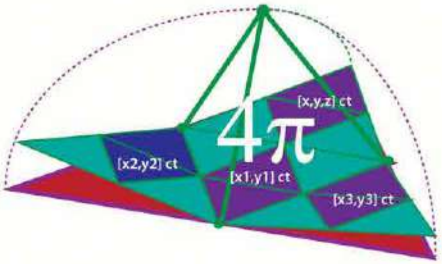

$$\vec{p} = \sum_{\text{all fascia}} \vec{p}$$

*Total Momentum
is the total of all quanta
linear Momenta in a 3D particle*

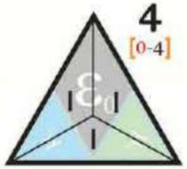

Negative standing-wave
Matter

Relativistic Matter
in motion

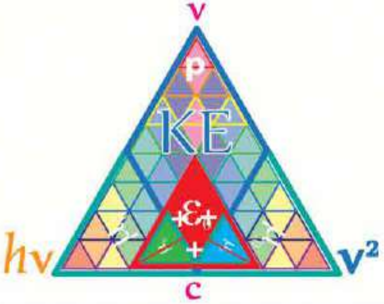

$$I = \sum_{\text{all fascia}} I$$

*Inertial mass
is the total of all inertia
in a 3D particle*

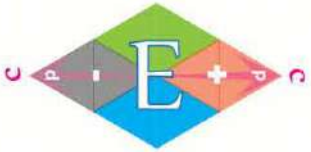

Photons have
null
energy-momentum
tensors

*Note that both
2D mass-energies [Special Relativity]
and 3D Matter [General Relativity]
have distinct Energy momenta*

*It is the 3D Tetrahedral topologies that
provides a definative basis for Matter*

Relativistic Matter
in motion

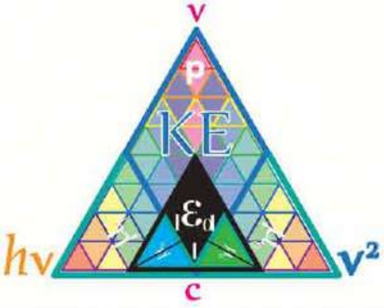

*Tetryonic Energy momenta tensors should not be confused with Four vector tensors
which map energy-momenta vectors in 3D spatial[cartesian] co-ordinate systems*

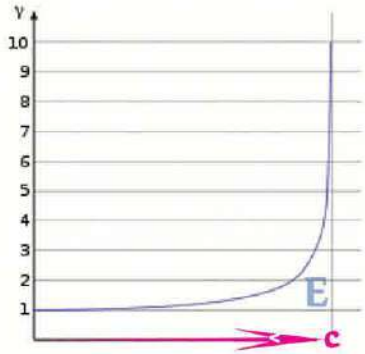

Relativistic Lorentz correction factor

The Lorentz factor or Lorentz term is an expression which appears in several equations in special relativity
It is often incorrectly graphed [as pictured to the right] as an increasing function of velocity approaching but never reaching c

**'c' is the maximum velocity acheivable through the electrical acceleration of particles
is NOT the maximum velocity possible for Matter in motion**

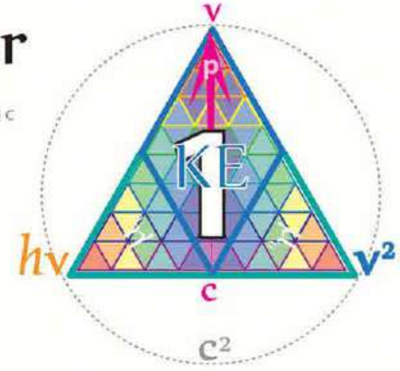

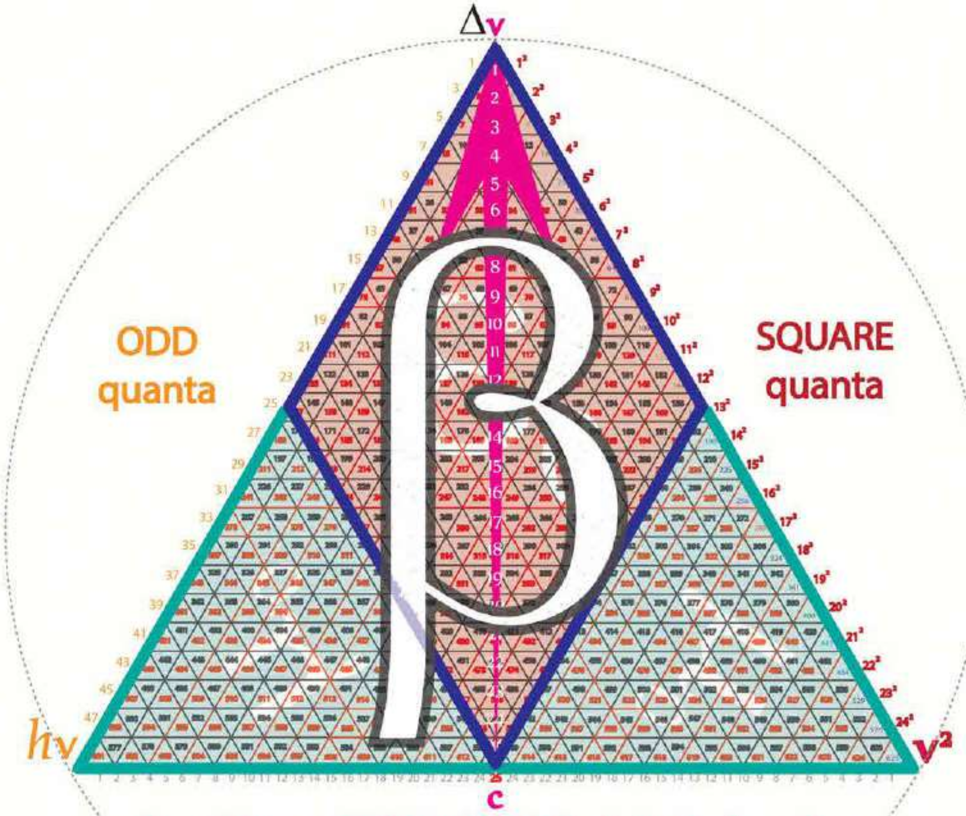

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

Scalar correction factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

*Lorentz factors are LINEAR and SCALAR velocity related corrections
to the relativistic mass-energy momenta content of any
physical system accelerated by EM Forces*

$$\frac{\Delta\gamma}{c^2}$$

Wavelength contraction

Time Dilation

$$\frac{\Delta\Omega}{c^2}$$

$$\frac{\Delta mv}{c^2}$$

Relativistic momentum

$$\frac{\Gamma}{c^2}$$

Lorentz corrections

Relativistic mass-energy

$$\frac{\Delta hv}{c^2}$$

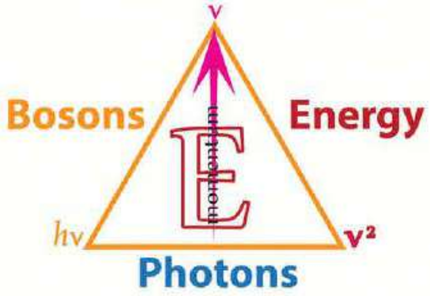
v
Bosons
Energy
E
momentum
hv
v²
Photons

form the basis for all charged energy geometries and their associated quanta distributions

Equilateral energies

Any measurements of mass without a Space-Time co-ordinate system are measurements of Energy

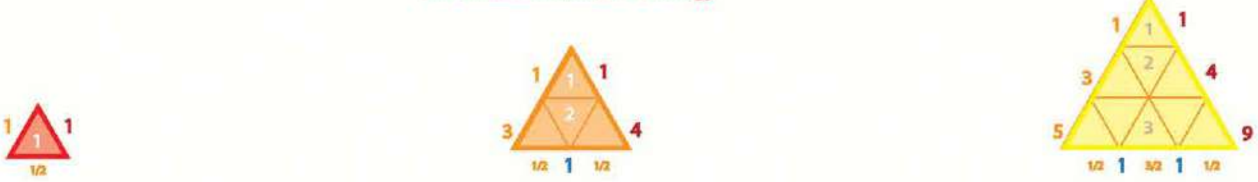
1 1 1
½
1 1 1 2 3 4
½ 1 ½
1 1 1 2 3 4 5 9 3
½ 1 3/2 1 ½

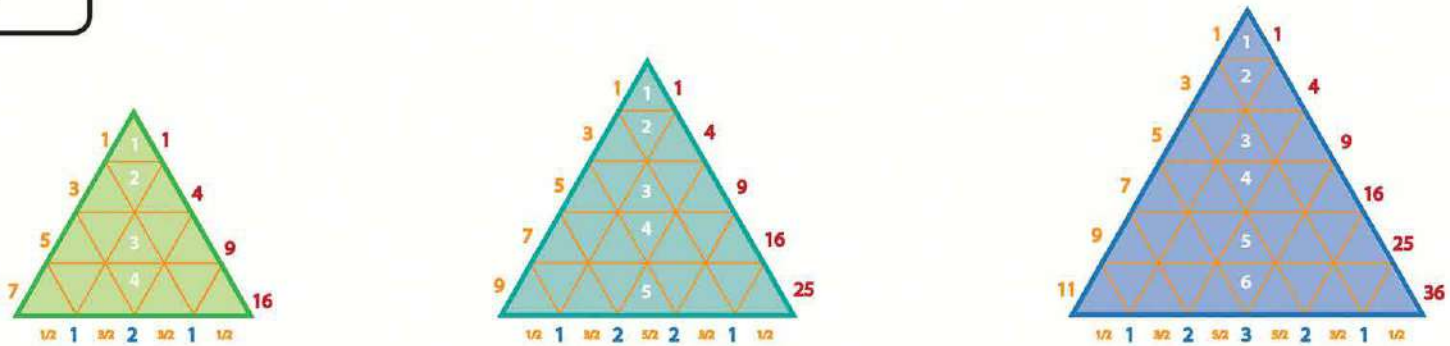
1 3 5 7
1 4 9 16
1 2 3 4
½ 1 3/2 2 3/2 1 ½
1 3 5 7 9
1 4 9 16 25
1 2 3 4 5
½ 1 3/2 2 5/2 2 3/2 1 ½
1 3 5 7 9 11
1 4 9 16 25 36
1 2 3 4 5 6
½ 1 3/2 2 5/2 3 5/2 2 3/2 1 ½

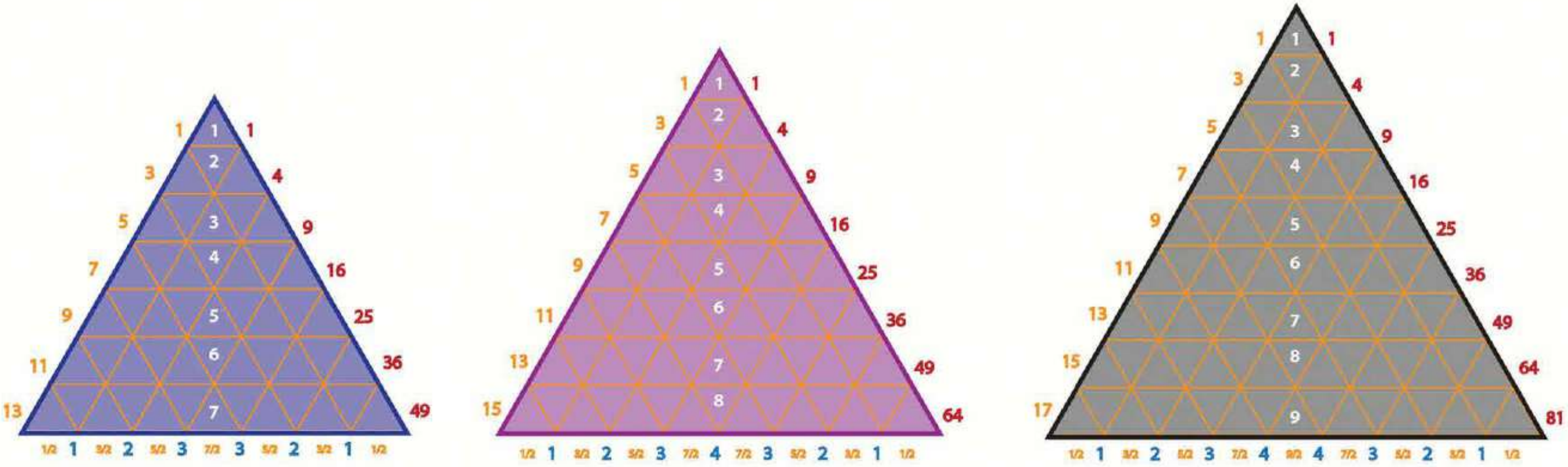
1 3 5 7 9 11 13
1 4 9 16 25 36 49
1 2 3 4 5 6 7
½ 1 3/2 2 5/2 3 7/2 3 5/2 2 3/2 1 ½
1 3 5 7 9 11 13 15
1 4 9 16 25 36 49 64
1 2 3 4 5 6 7 8
½ 1 3/2 2 5/2 3 7/2 4 7/2 3 5/2 2 3/2 1 ½
1 3 5 7 9 11 13 15 17
1 4 9 16 25 36 49 64 81
1 2 3 4 5 6 7 8 9
½ 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 ½

Copyright ABRAHAM [2008] - All rights reserved

Electromagnetic mass

2D mass-energy is the surface integral of 3D Matter

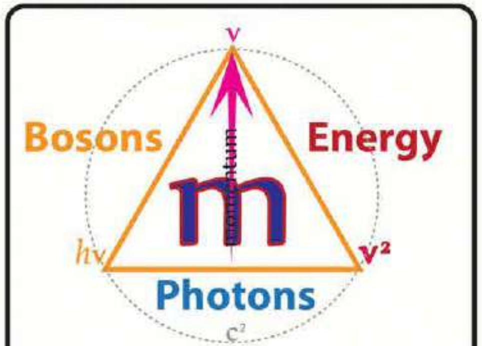
v
Bosons
Energy
momentum
m
hv
v²
Photons
c²

Time specific measurements of equilateral ENERGY momenta geometries forms the basis for EM mass

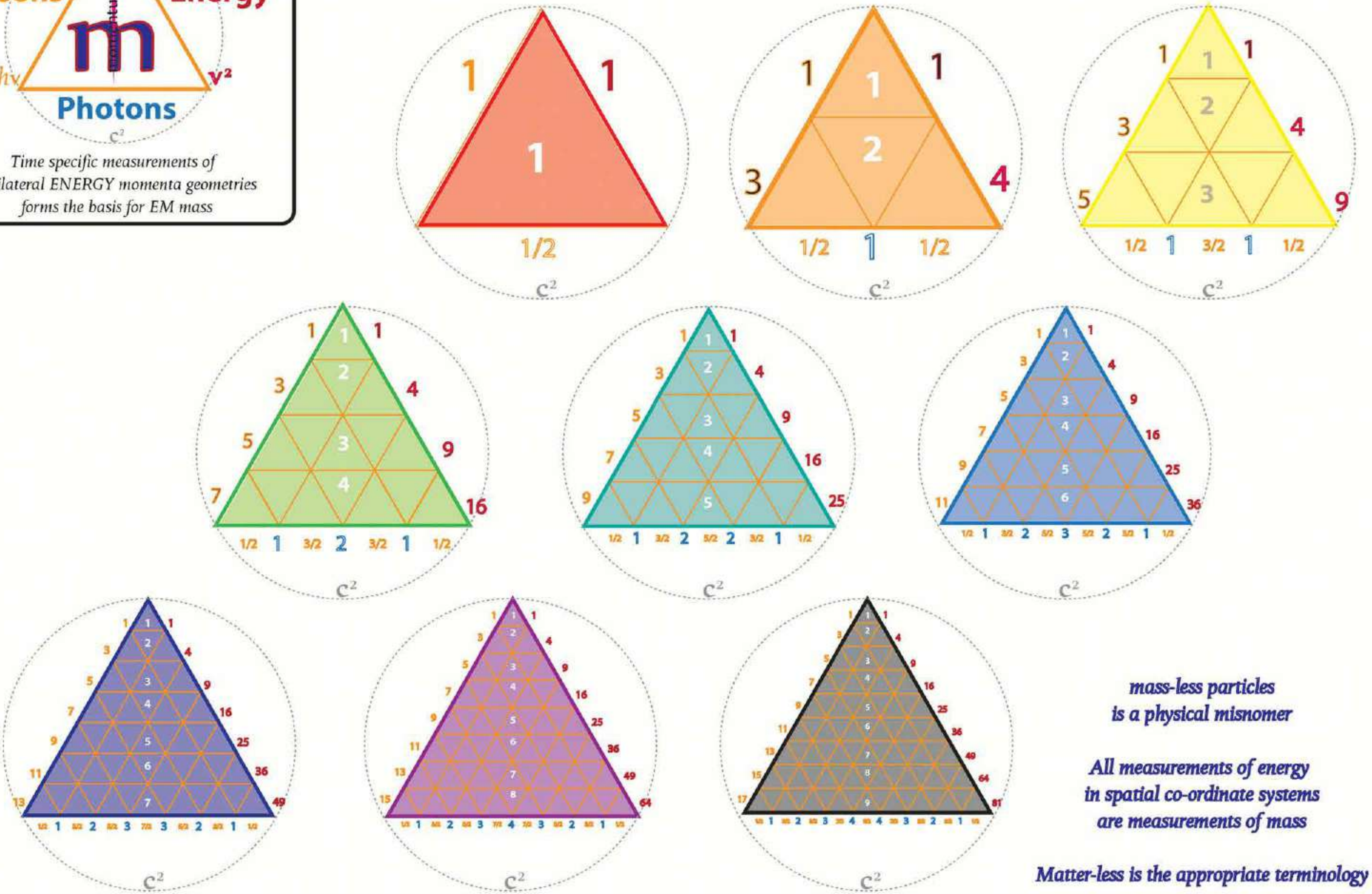
1
1
1
1/2
c²
1
1
1
2
3
4
1/2
1
1/2
c²
1
1
1
2
3
3
4
5
9
1/2
1
3/2
1
1/2
c²

mass-less particles is a physical misnomer

All measurements of energy in spatial co-ordinate systems are measurements of mass

Matter-less is the appropriate terminology

Standing-wave Matter

Matter is a higher order 3D topology created by standing wave mass-energies

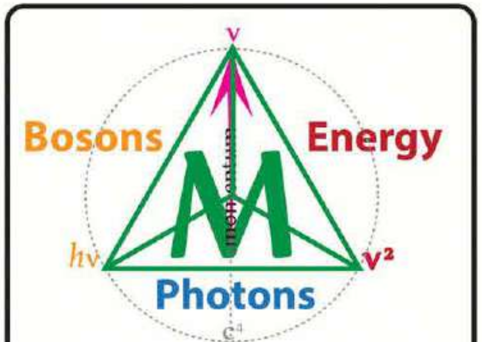
v
Bosons
Energy
M
momentum
hv
v²
Photons
c⁴

The equilateral mass-ENERGY content of 4nπ Tetrahedral standing-wave topologies forms the basis of Matter

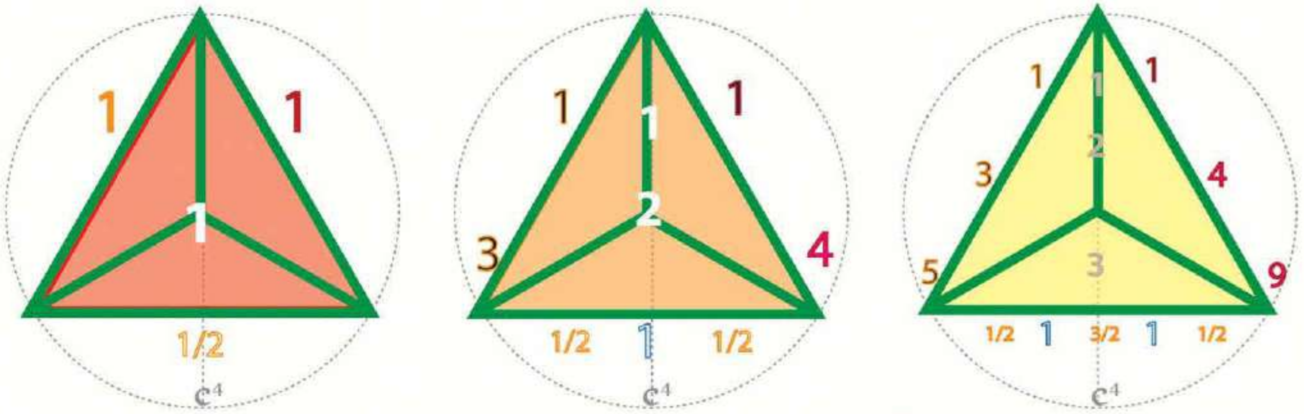
1
1
1
1/2
c⁴
1
1
1
2
3
4
1/2
1
1/2
c⁴
1
1
1
2
3
3
4
5
9
1/2
1
3/2
1
1/2
c⁴

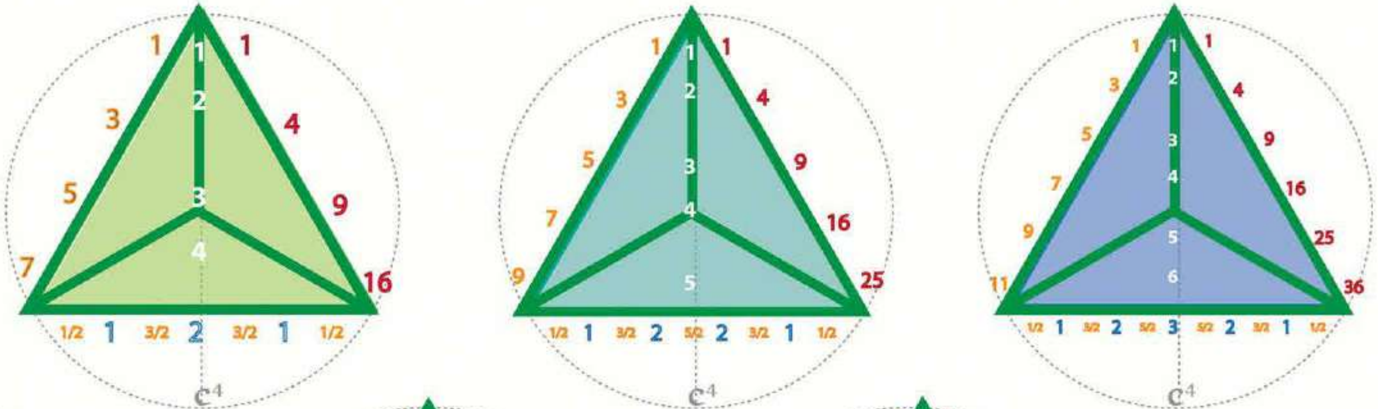
1
3
5
7
1
4
9
16
1
2
3
4
1/2
1
3/2
2
3/2
1
1/2
c⁴
1
3
5
7
9
1
4
9
16
25
1
2
3
4
5
1/2
1
3/2
2
5/2
2
3/2
1
1/2
c⁴
1
3
5
7
9
11
1
4
9
16
25
36
1
2
3
4
5
6
1/2
1
3/2
2
5/2
3
5/2
2
3/2
1
1/2
c⁴

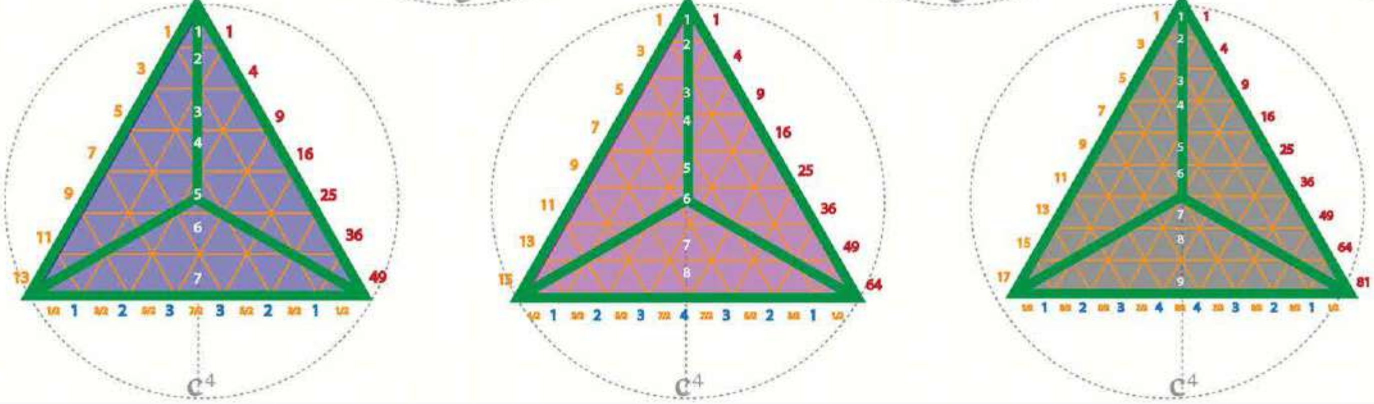
1
3
5
7
9
11
13
1
4
9
16
25
36
49
1
2
3
4
5
6
7
1/2
1
3/2
2
5/2
3
7/2
3
5/2
2
3/2
1
1/2
c⁴
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
1
2
3
4
5
6
7
8
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2
c⁴
1
3
5
7
9
11
13
15
17
1
4
9
16
25
36
49
64
81
1
2
3
4
5
6
7
8
9
1/2
1
3/2
2
5/2
3
7/2
4
4
7/2
3
5/2
2
3/2
1
1/2
c⁴

Only mass-ENERGIES contained in the tetryonic fascia of standing-wave topologies contribute to weight

Periodic element nuclei

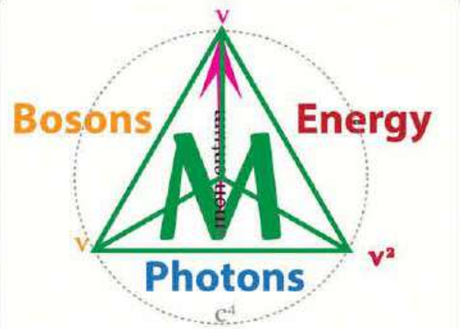
v
Bosons
Energy
M
momentum
v
v²
Photons
c⁴

The mass-ENERGY content of
4nπ Tetrahedral standing-wave topologies
forms the basis of all periodic elements

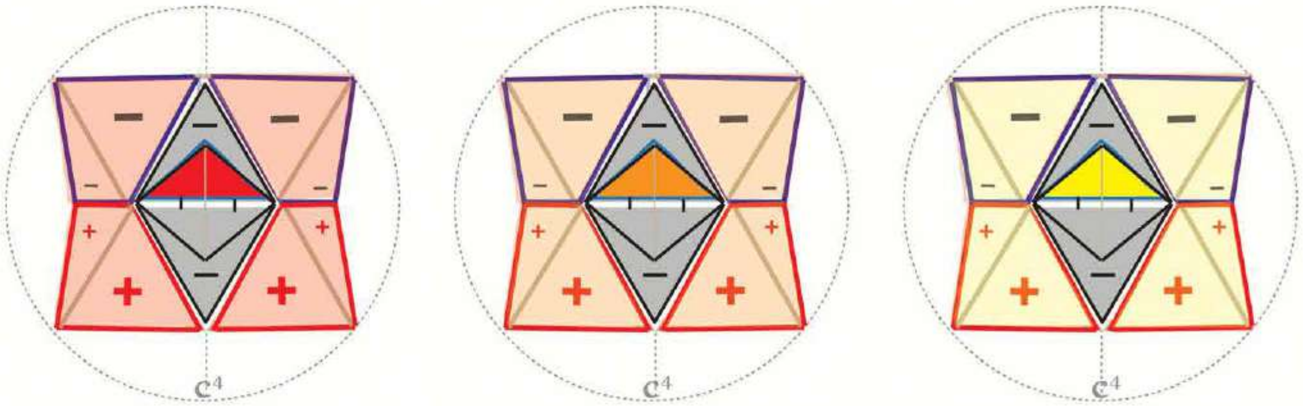
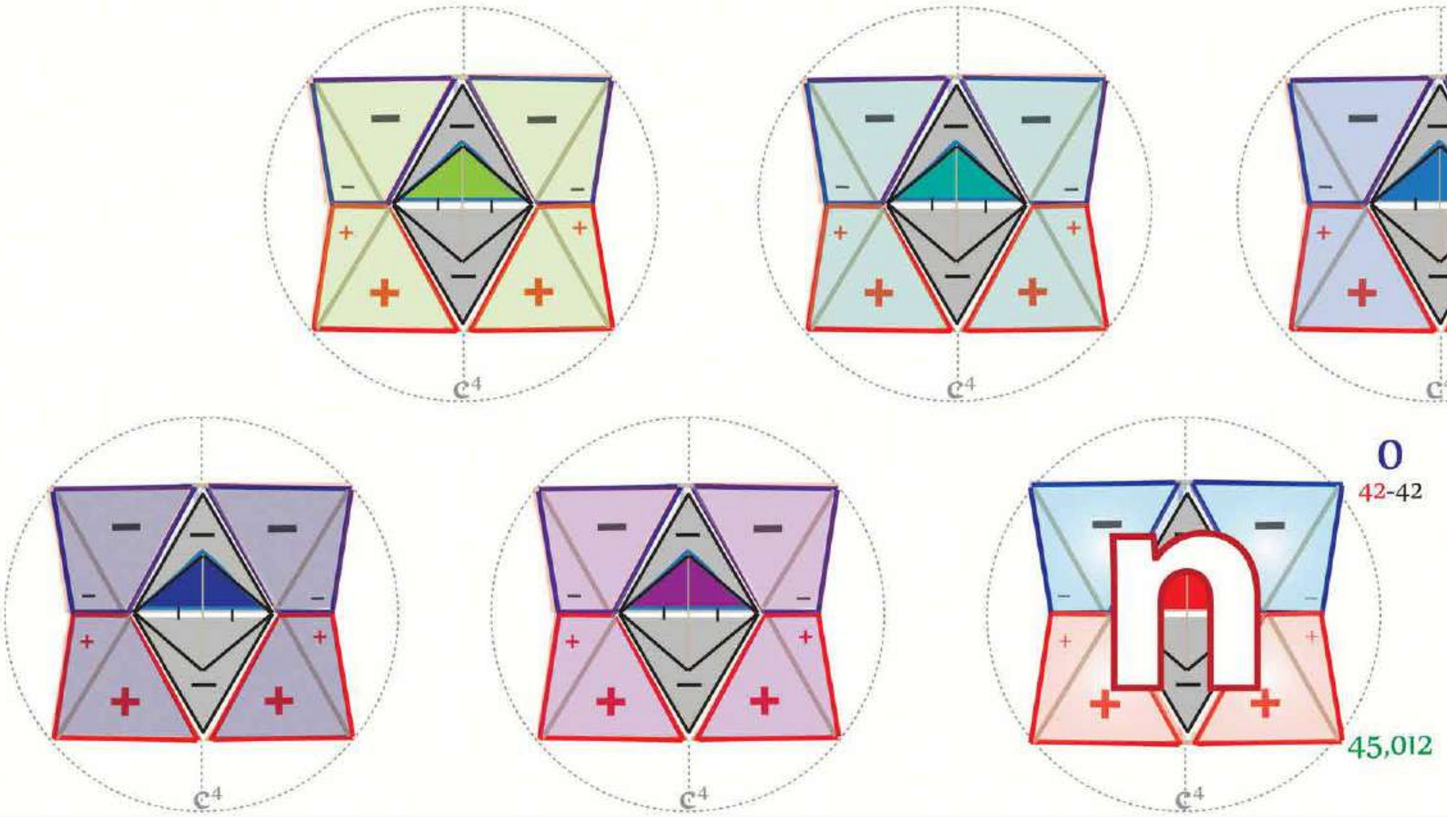
−
+
c⁴
n
0
42-42
45,012

ALL
periodic elements are
comprised of
n level deuterium
nuclei

Copyright ABRAHAM [2008] - All rights reserved

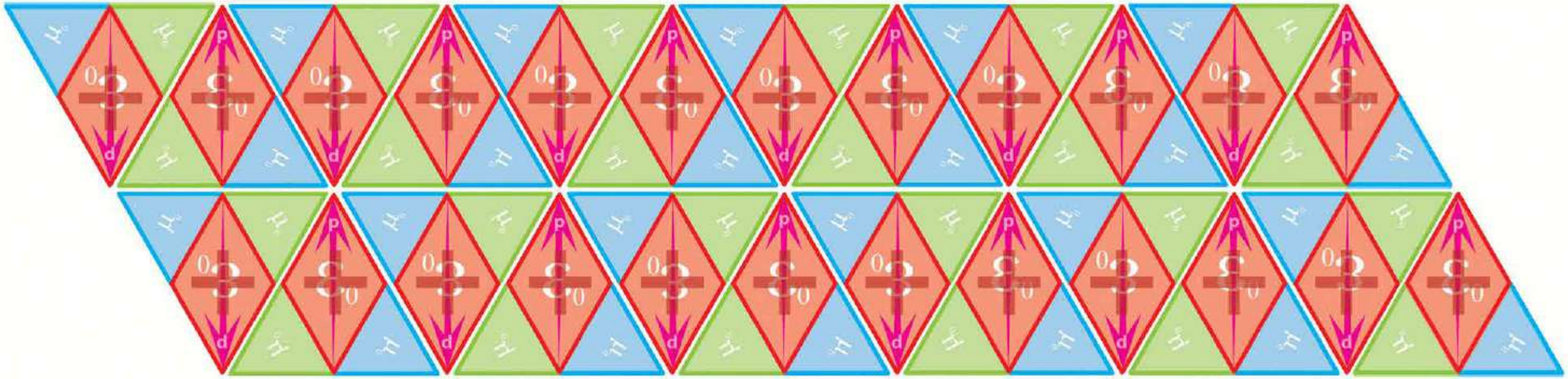

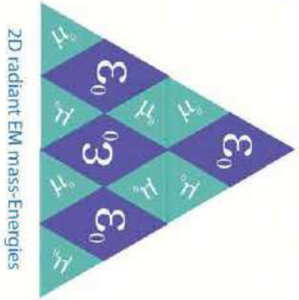
2D radiant EM mass-Energies
ε0
μ0

n1 - Charged Tetryon Templates

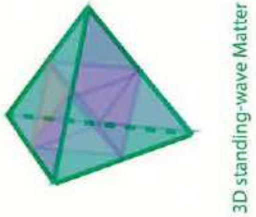
3D standing-wave Matter

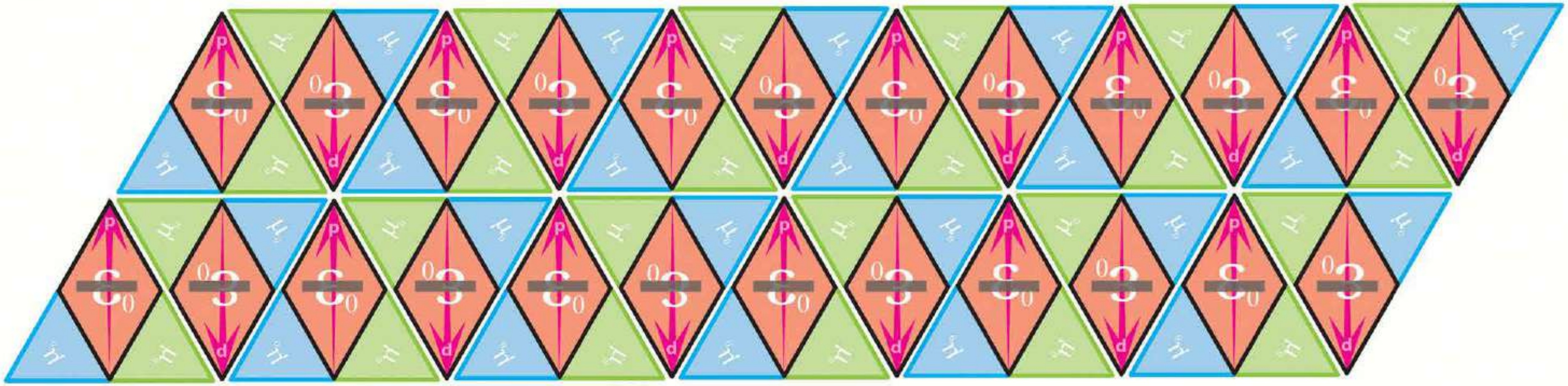

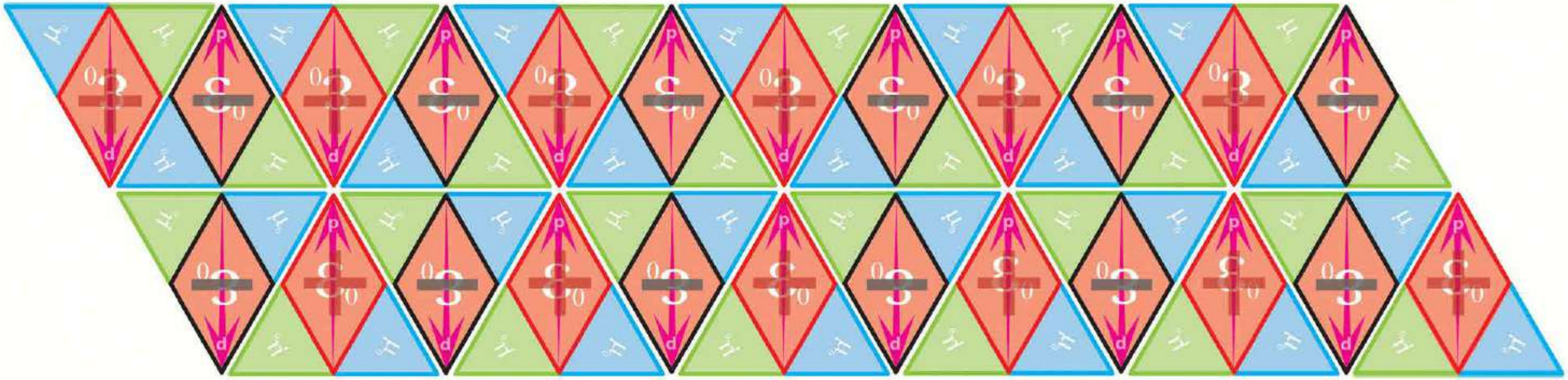

2D radiant EM mass-Energies

n1 - Neutral Tetryon Templates

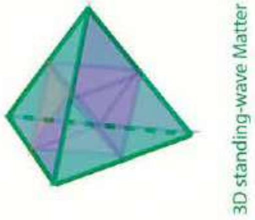
3D standing-wave Matter

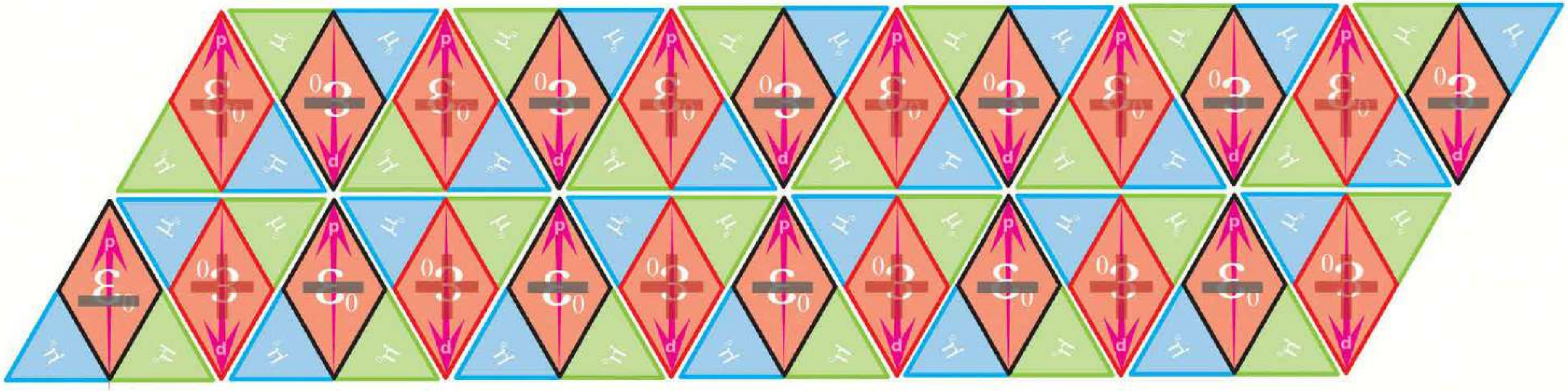

Copyright ABRAHAM [2008] - All rights reserved

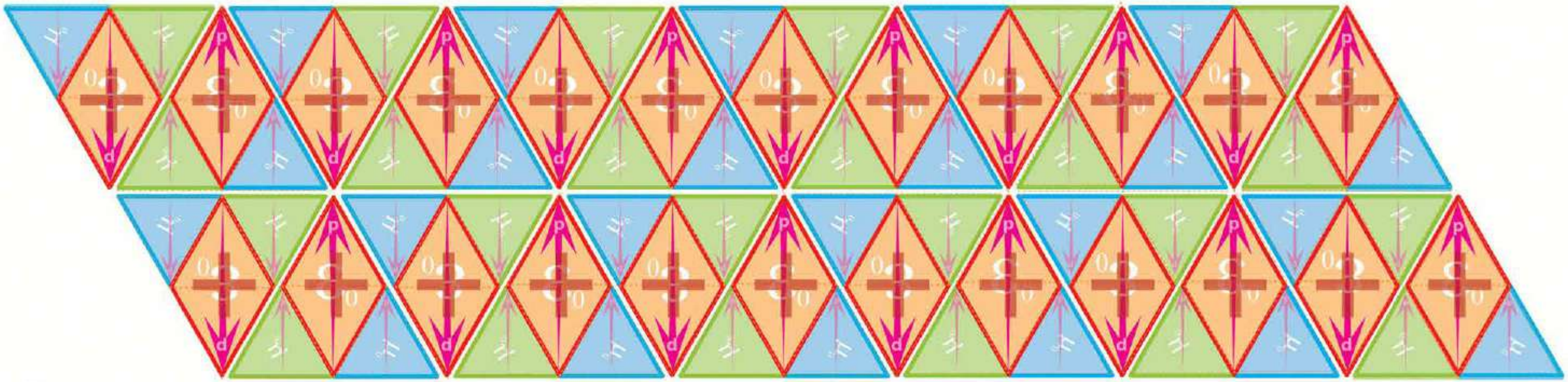
p
d
μ_0
ε_0

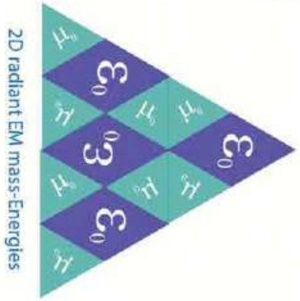
2D radiant EM mass-Energies
μ_0
ε_0

n2 - Charged Tetryon Templates

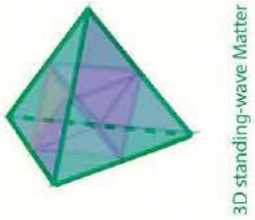
3D standing-wave Matter

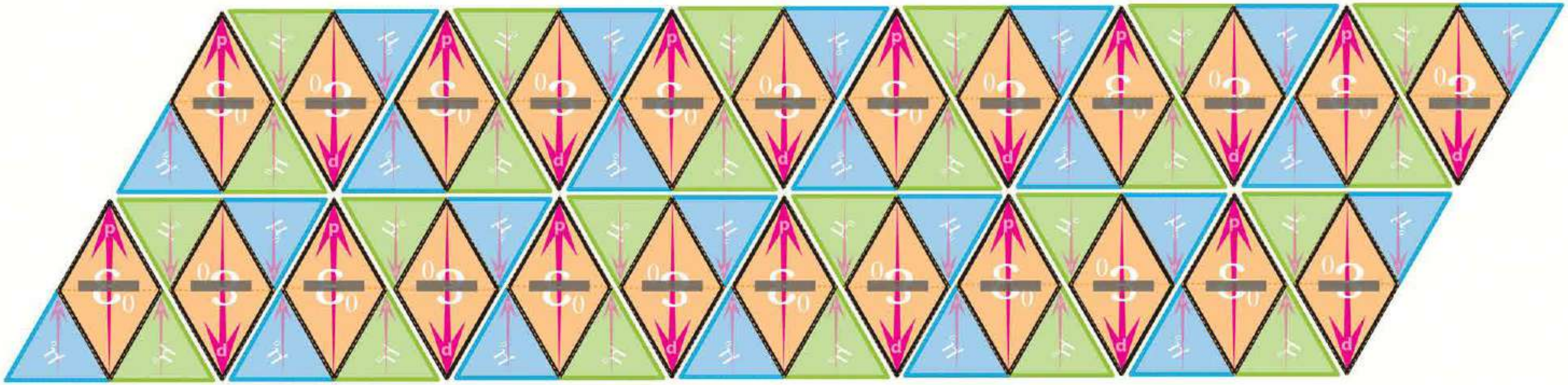
p
d
μ_0
ε_0

Copyright ABRAHAM [2008] - All rights reserved

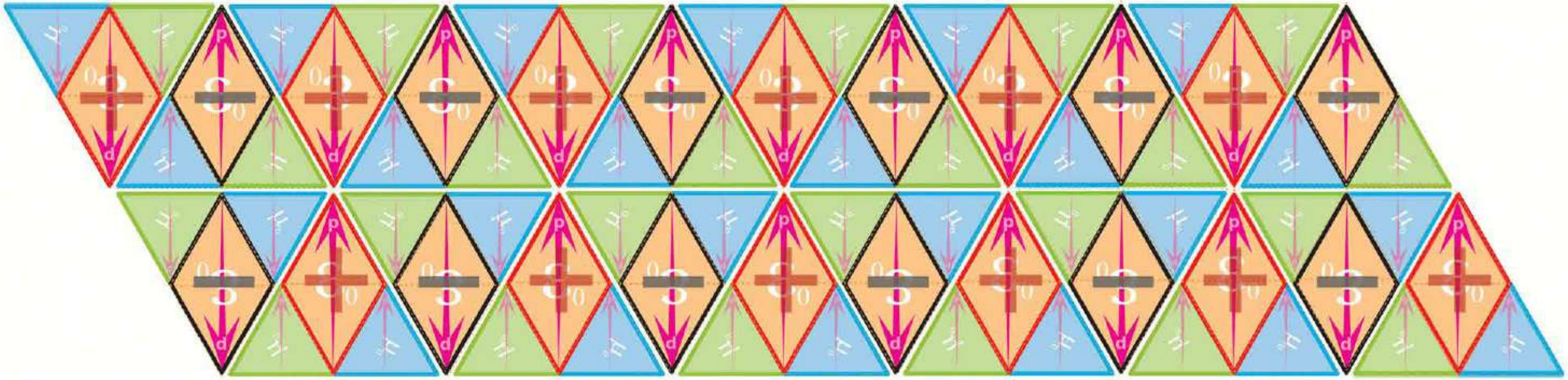

2D radiant EM mass-Energies

n2 - Neutral Tetryon Templates

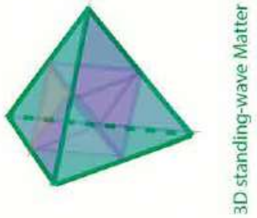
3D standing-wave Matter

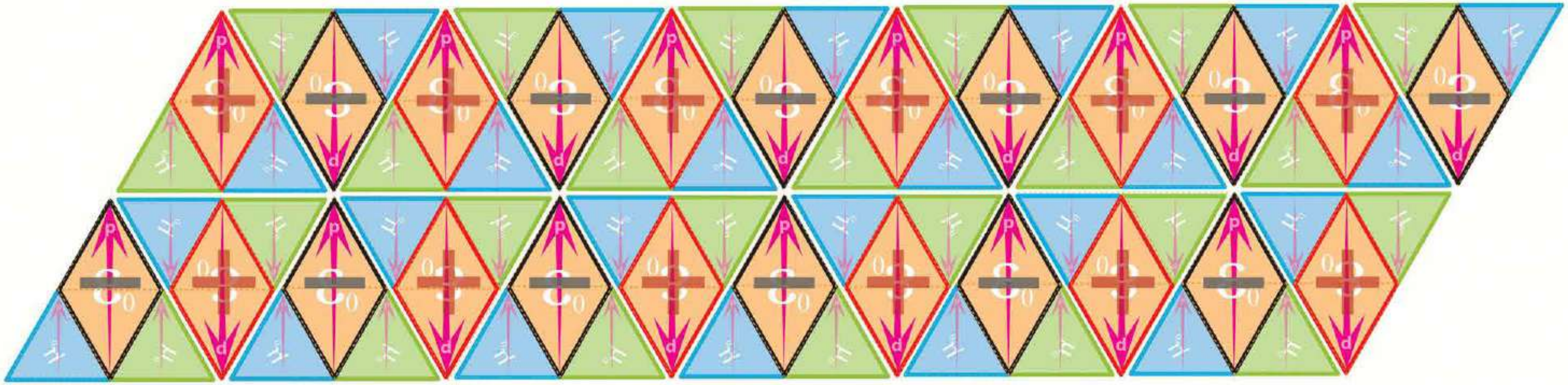

Copyright ABRAHAM [2008] - All rights reserved

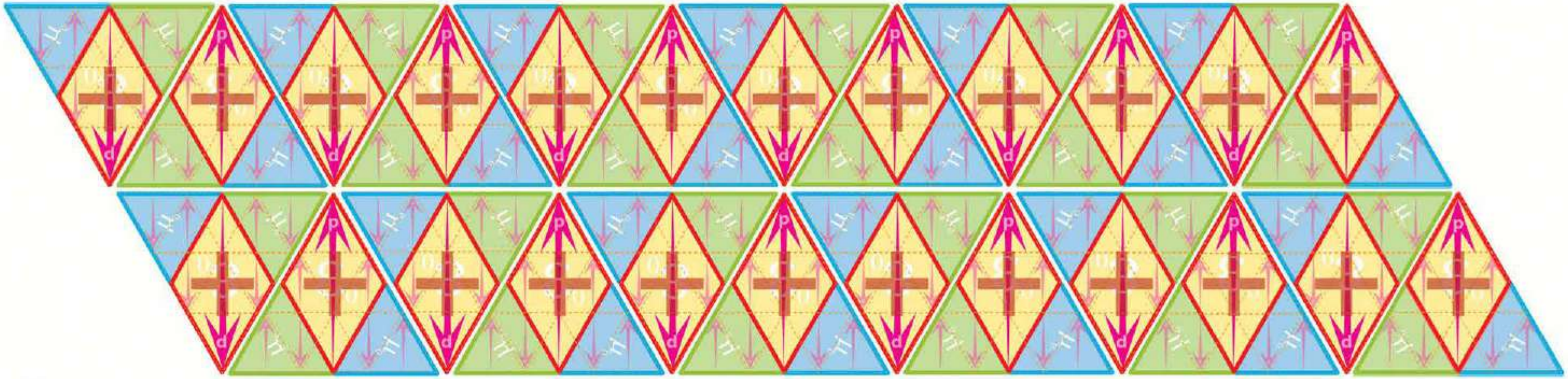

2D radiant EM mass-Energies
μ0
ε0

n3 - Charged Tetryon Templates

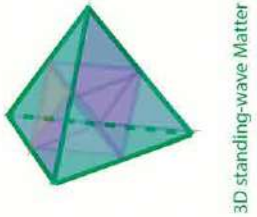
3D standing-wave Matter

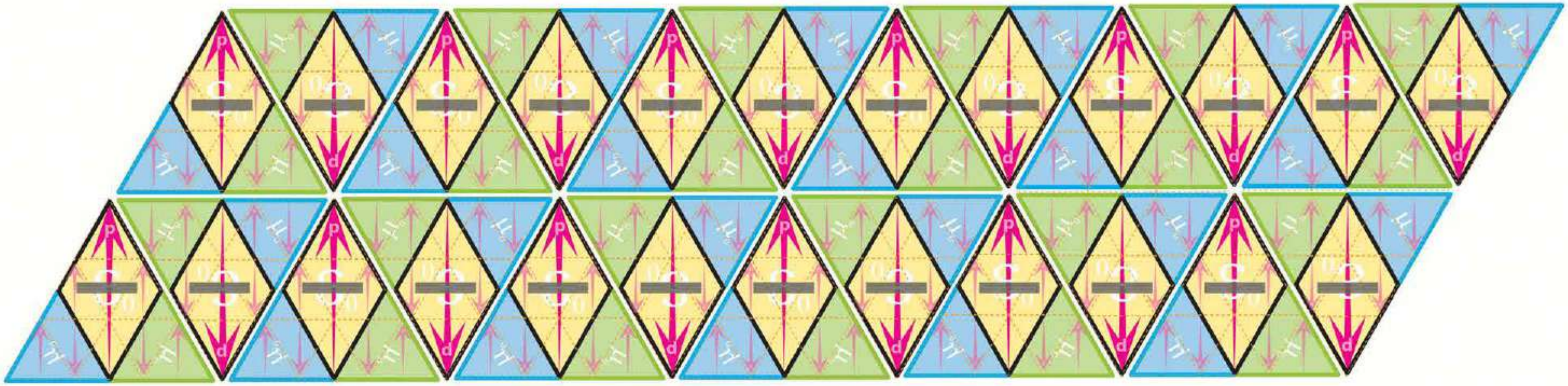

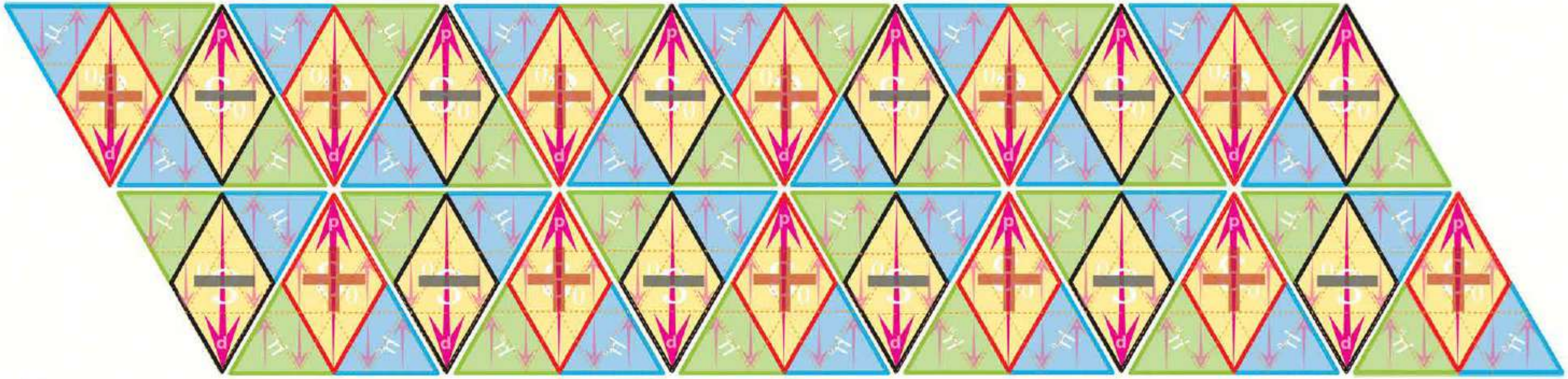

2D radiant EM mass-Energies

n3 - Neutral Tetryon Templates

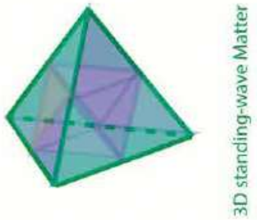
3D standing-wave Matter

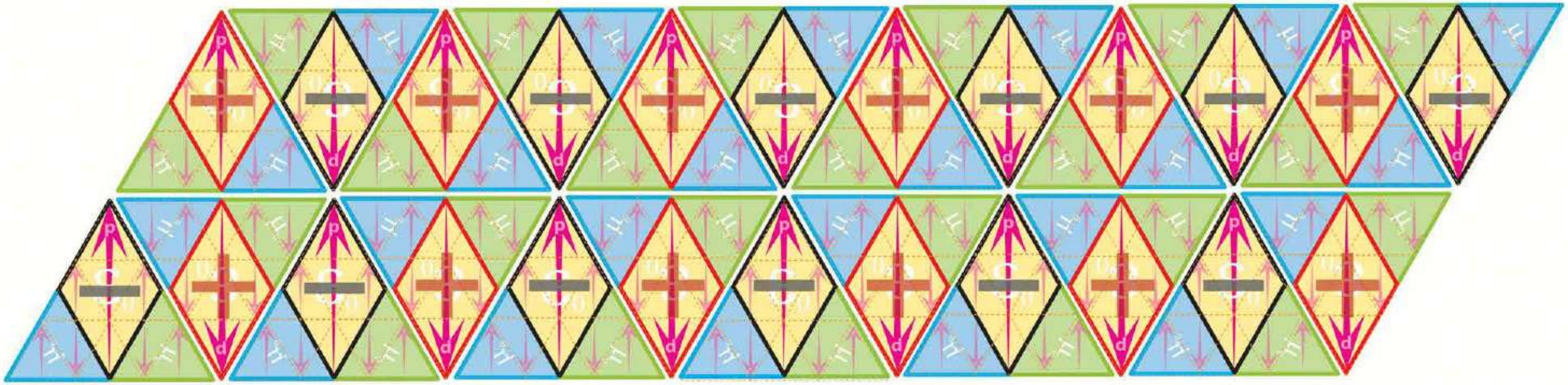

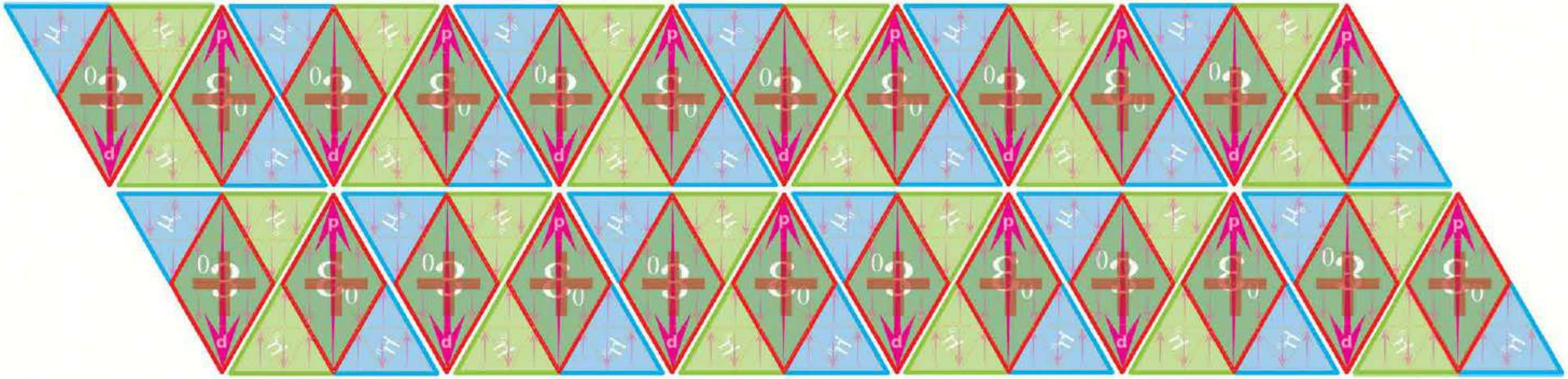
p
d
ε_0
μ_0

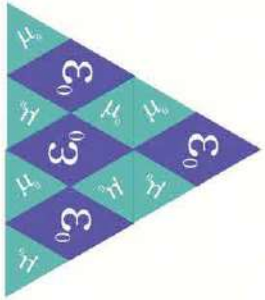
ε_0
μ_0

2D radiant EM mass-Energies

n4 - Charged Tetryon Templates

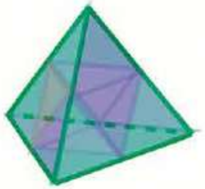

3D standing-wave Matter

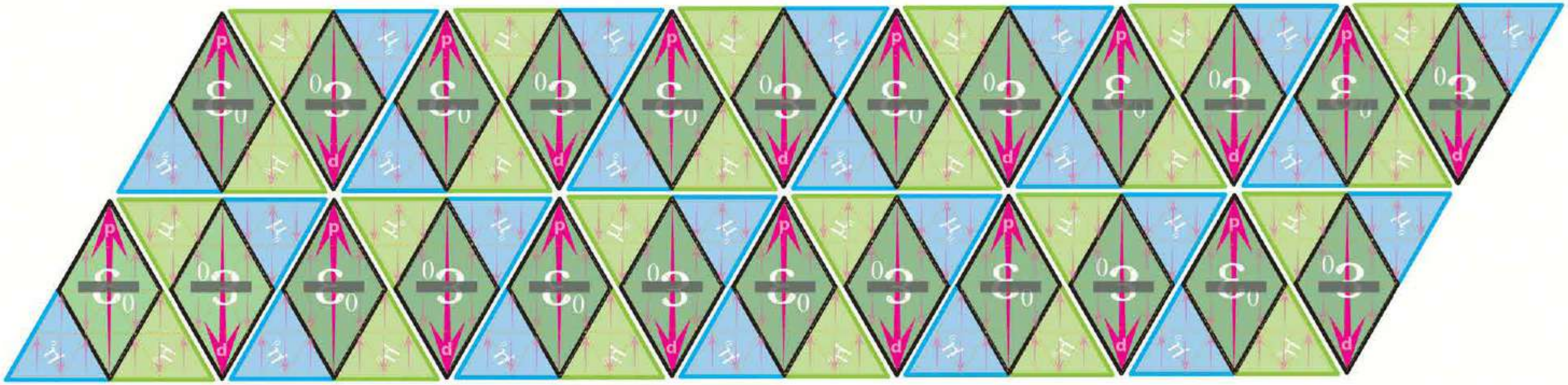
p
d
ε_0
μ_0

Copyright ABRAHAM [2008] - All rights reserved

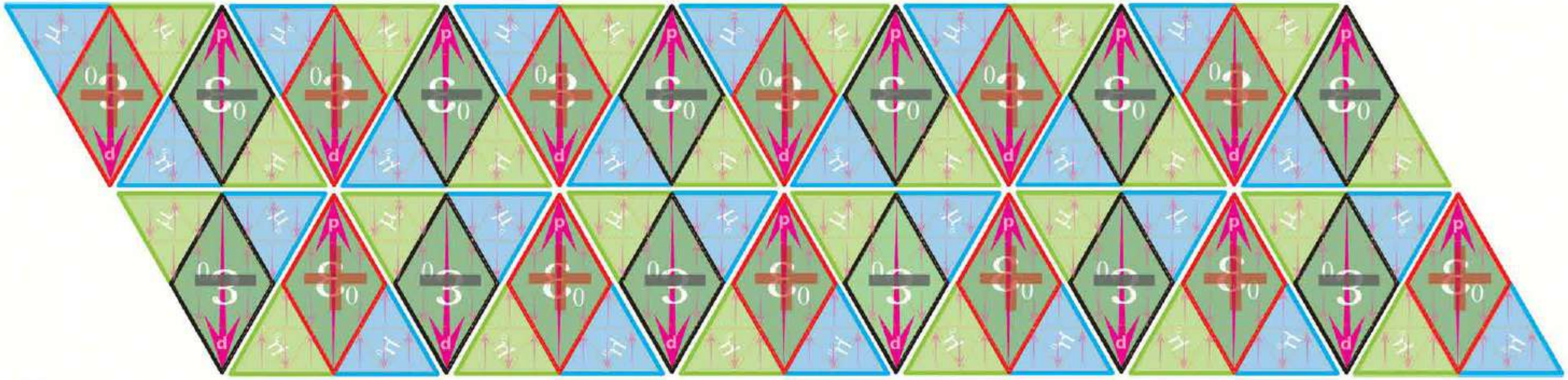
p
d
μ_0
ε_0

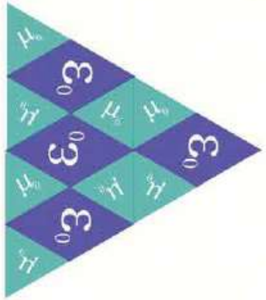
μ_0
ε_0

2D radiant EM mass-Energies

n4 - Neutral Tetryon Templates

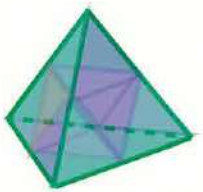

3D standing-wave Matter

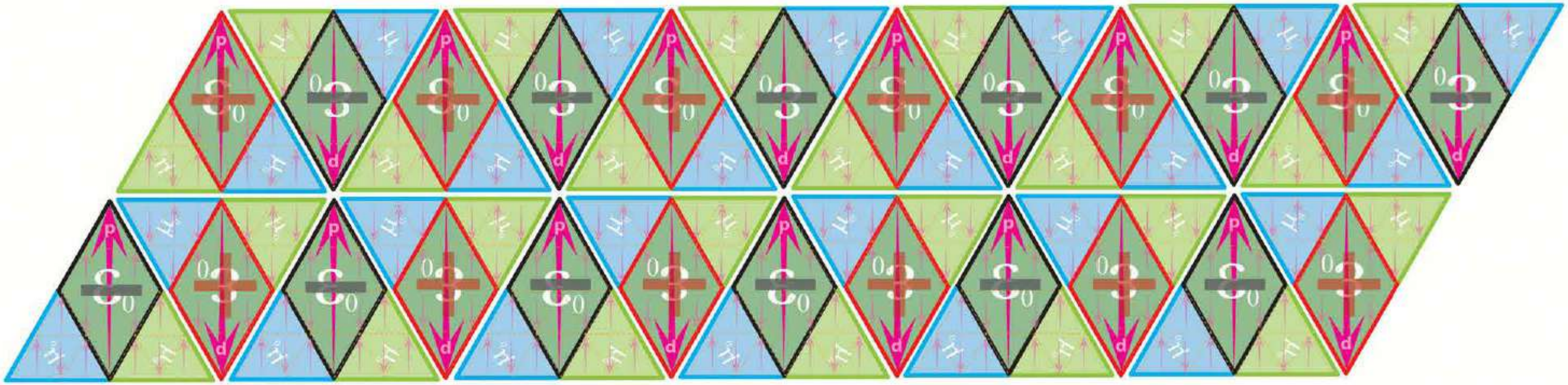
p
d
μ_0
ε_0

Copyright ABRAHAM [2008] - All rights reserved

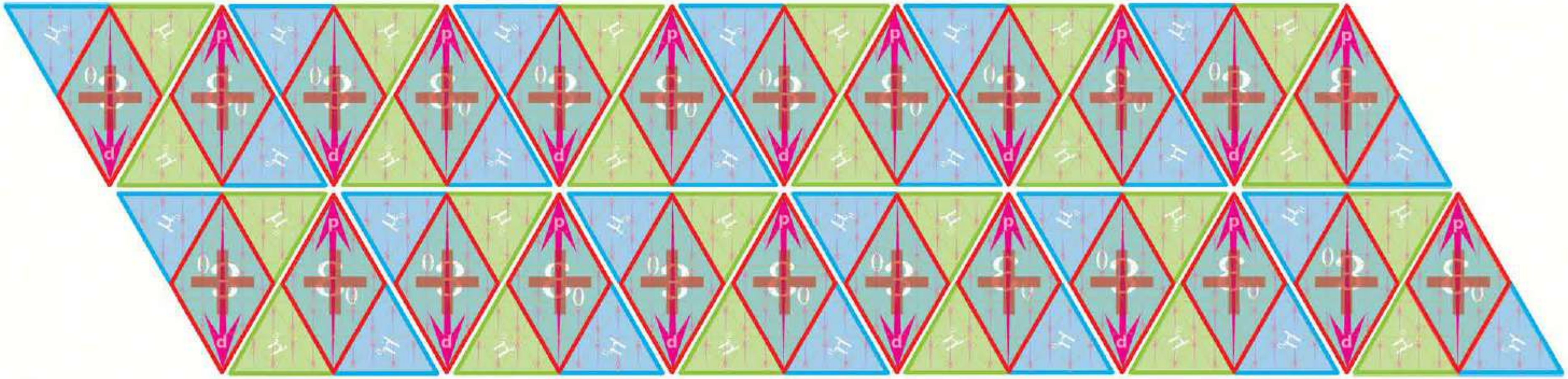

2D radiant EM mass-Energies

n5 - Charged Tetryon Templates

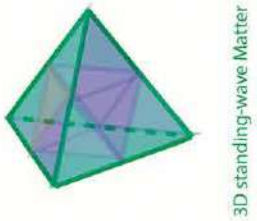
3D standing-wave Matter

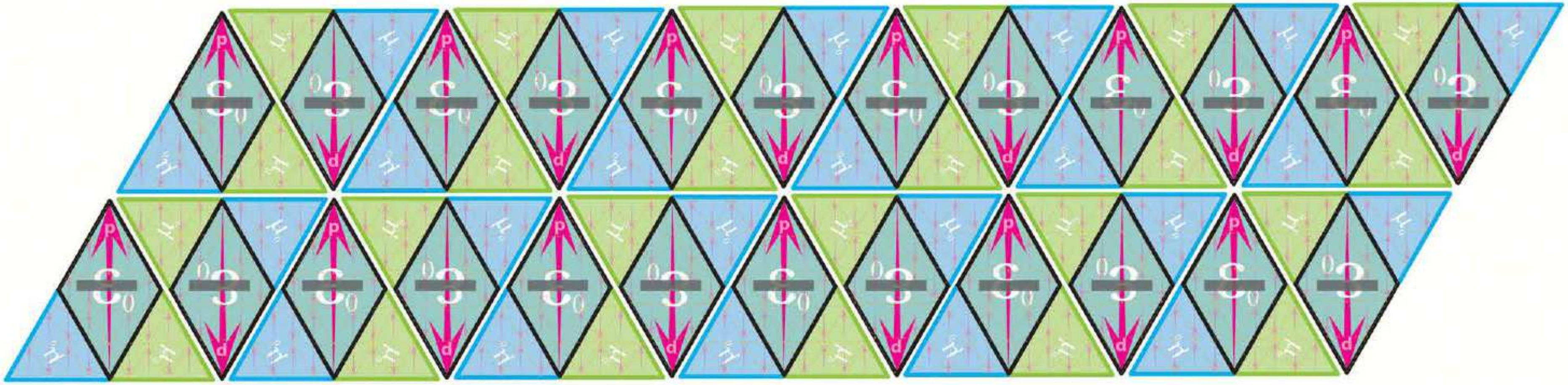

Copyright ABRAHAM [2008] - All rights reserved

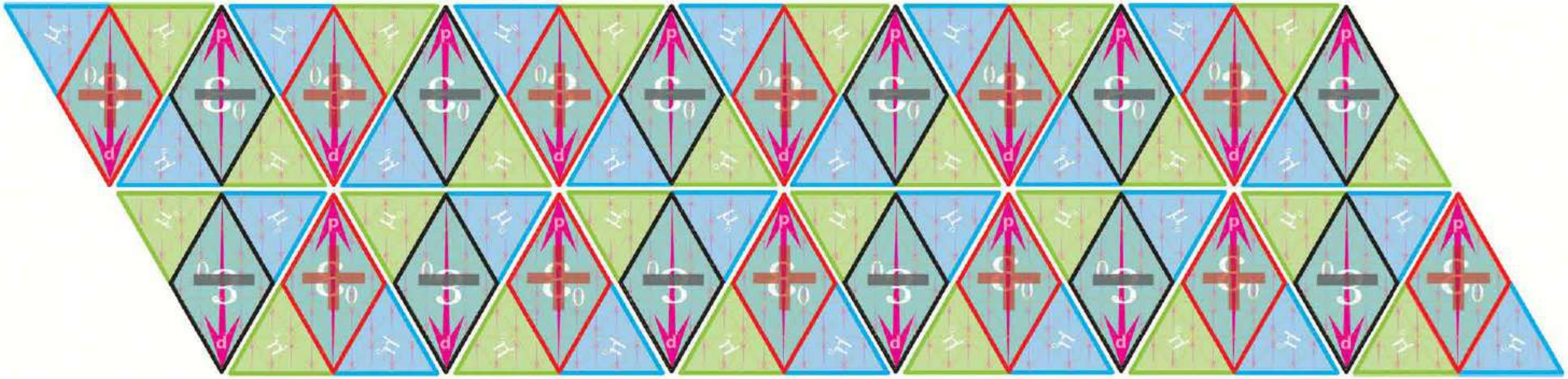
p
d
μ_0
ε_0

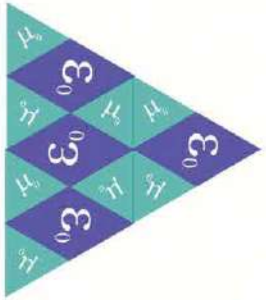
μ_0
ε_0

2D radiant EM mass-Energies

n5 - Neutral Tetryon Templates

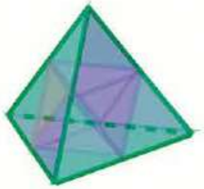

3D standing-wave Matter

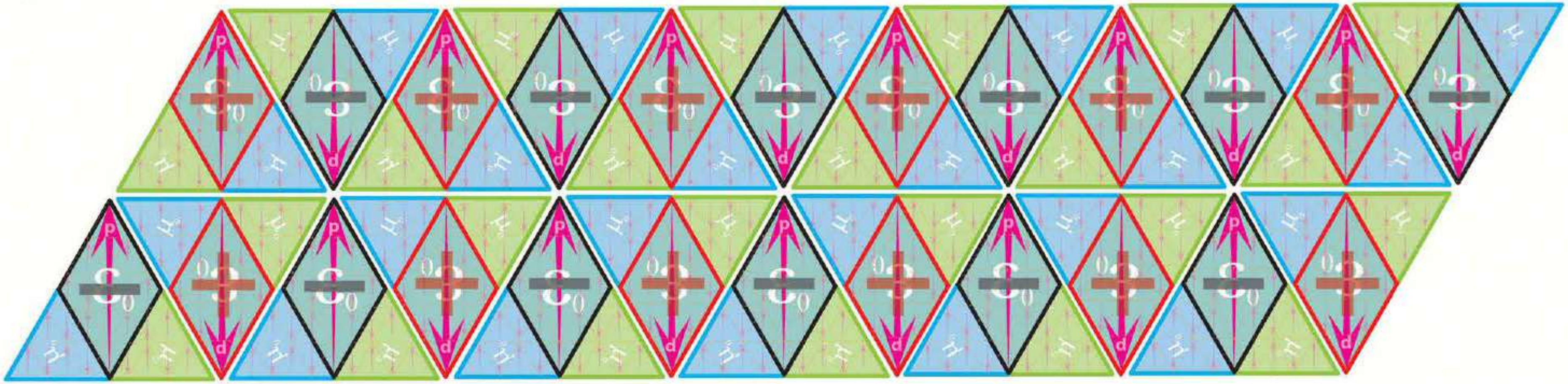
p
d
μ_0
ε_0

Copyright ABRAHAM [2008] - All rights reserved

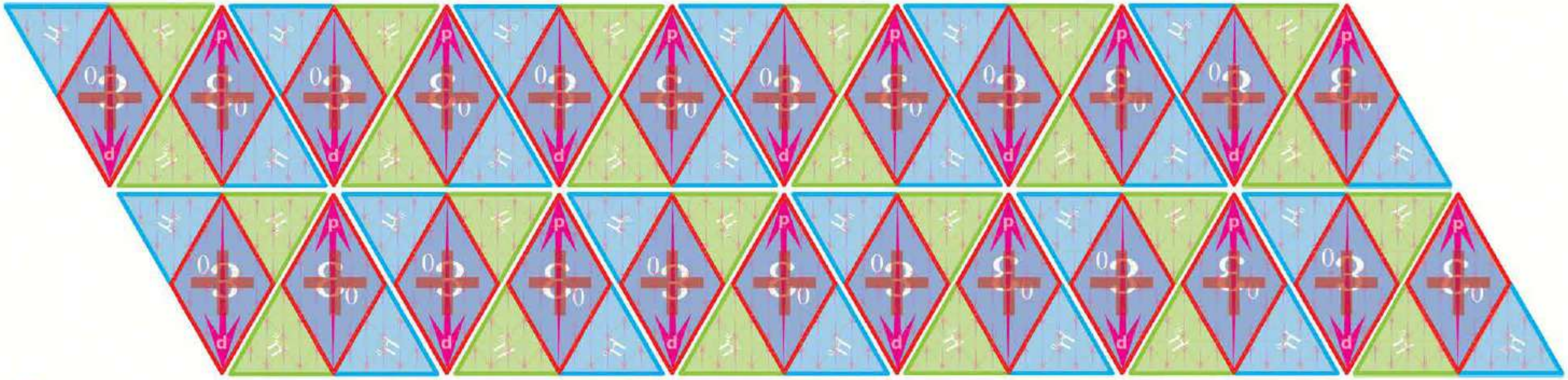

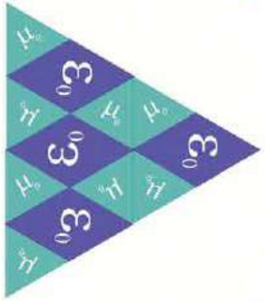

2D radiant EM mass-Energies

n6 - Charged Tetryon Templates

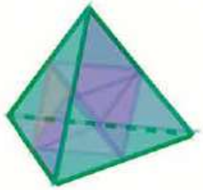

3D standing-wave Matter

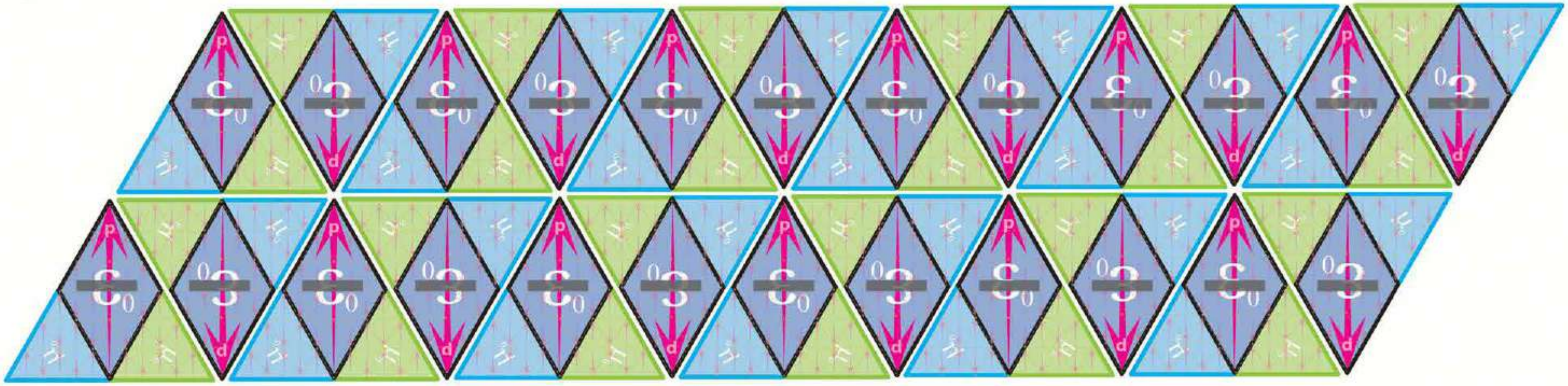

Copyright ABRAHAM [2008] - All rights reserved

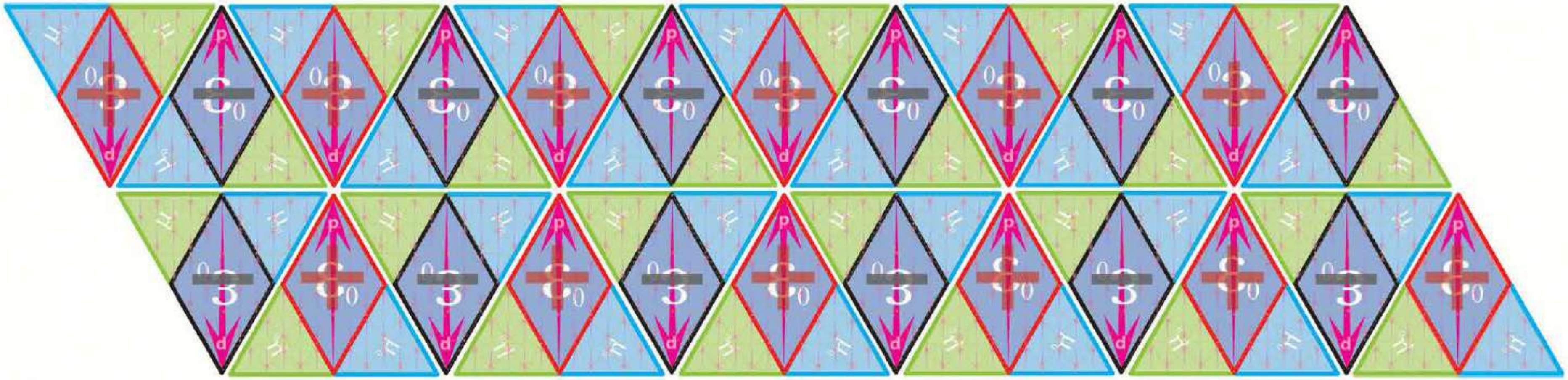
p
d
ε_0
μ_0

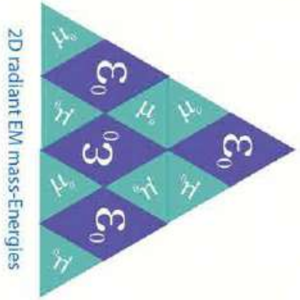
2D radiant EM mass-Energies
ε_0
μ_0

n6 - Neutral Tetryon Templates

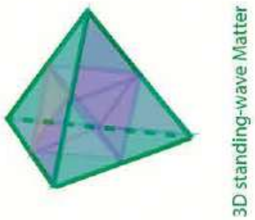
3D standing-wave Matter

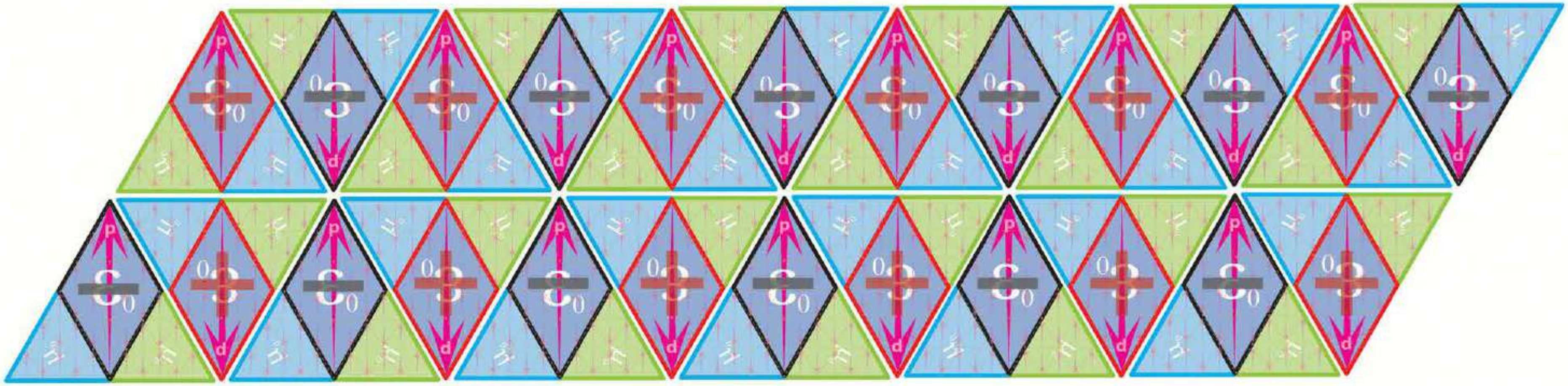
p
d
ε_0
μ_0

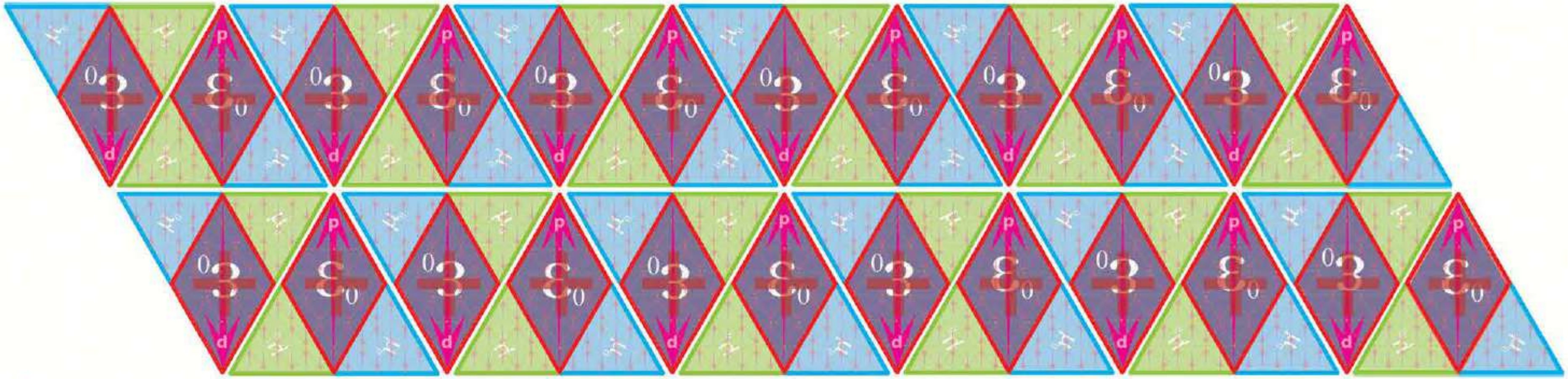

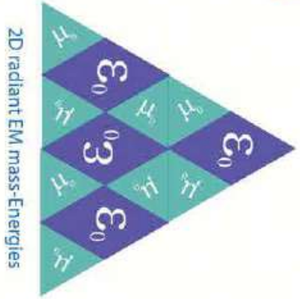
2D radiant EM mass-Energies
ε0
μ0

n7 - Charged Tetryon Templates

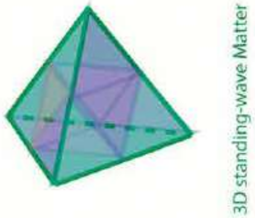
3D standing-wave Matter

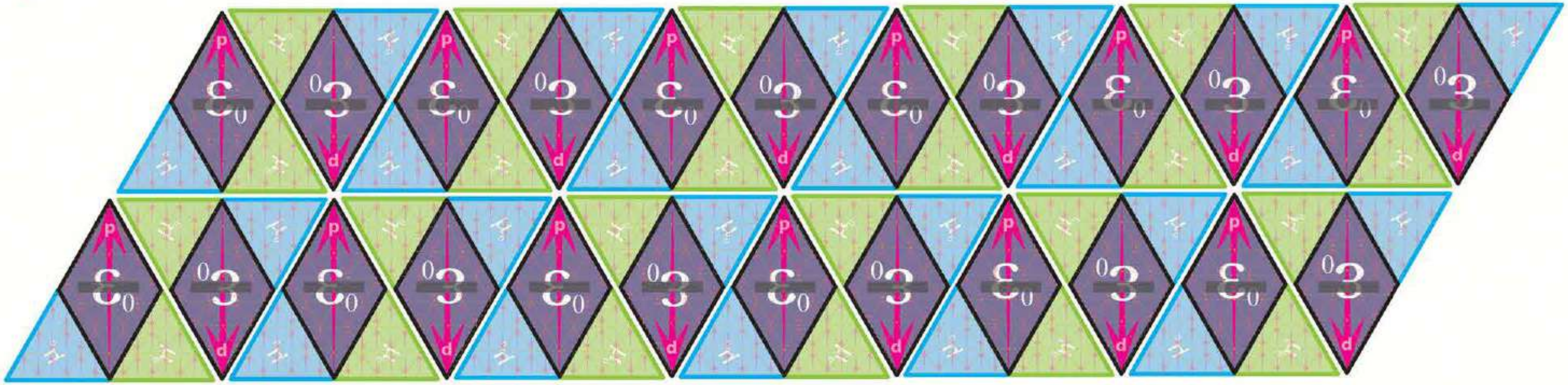

Copyright ABRAHAM [2008] - All rights reserved

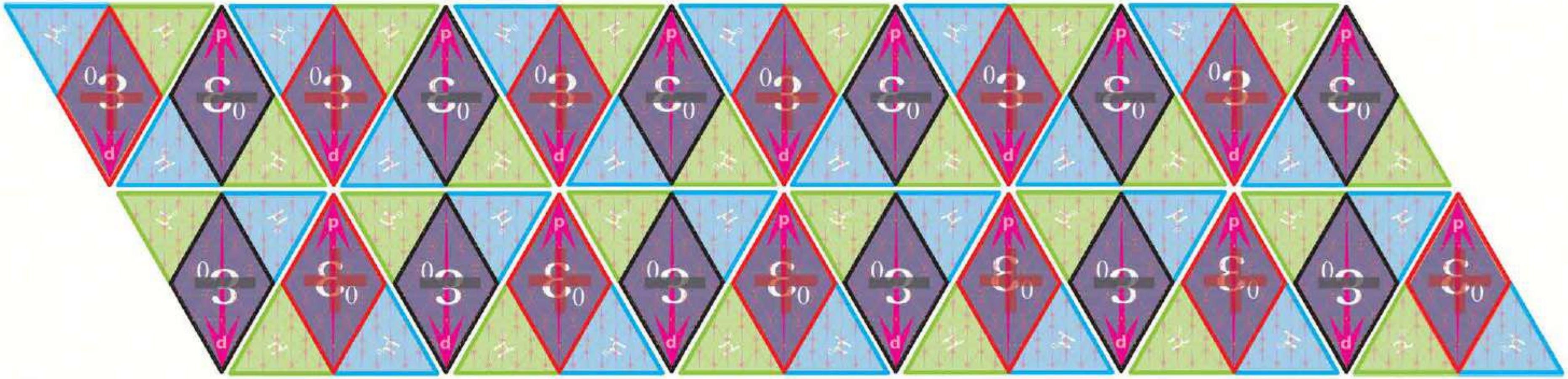

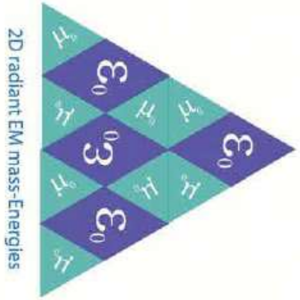
2D radiant EM mass-Energies
ε0
μ0

n7 - Neutral Tetryon Templates

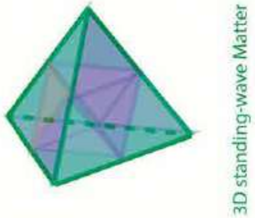
3D standing-wave Matter

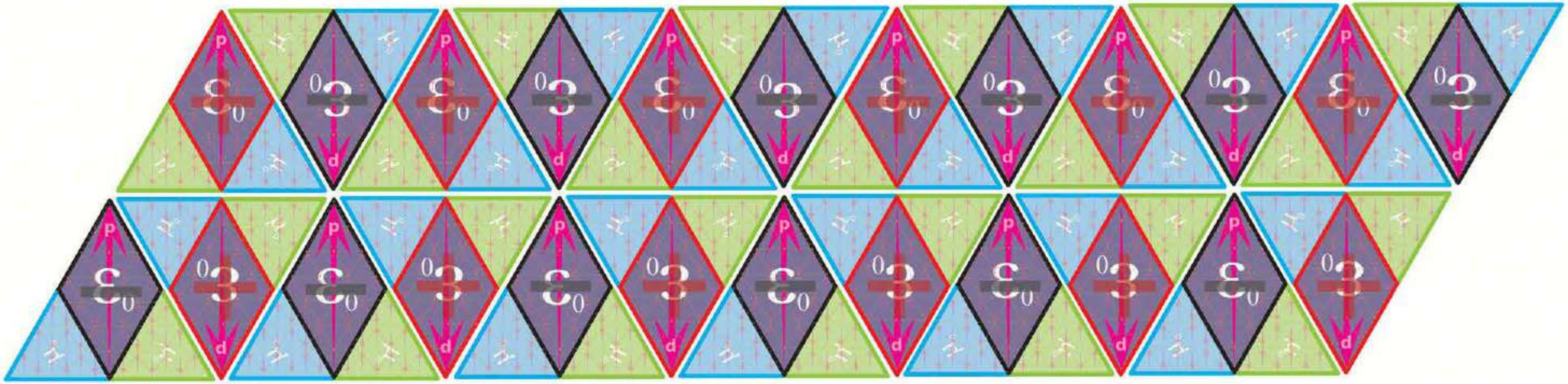

Copyright ABRAHAM [2008] - All rights reserved

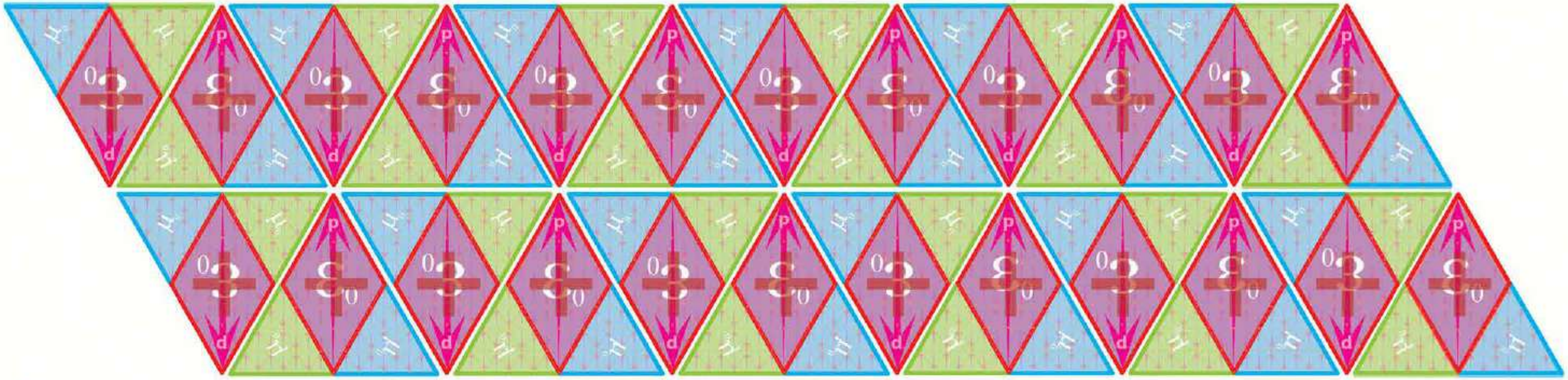

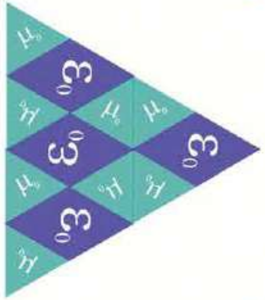

2D radiant EM mass-Energies

n8 - Charged Tetryon Templates

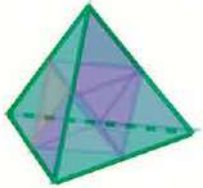

3D standing-wave Matter

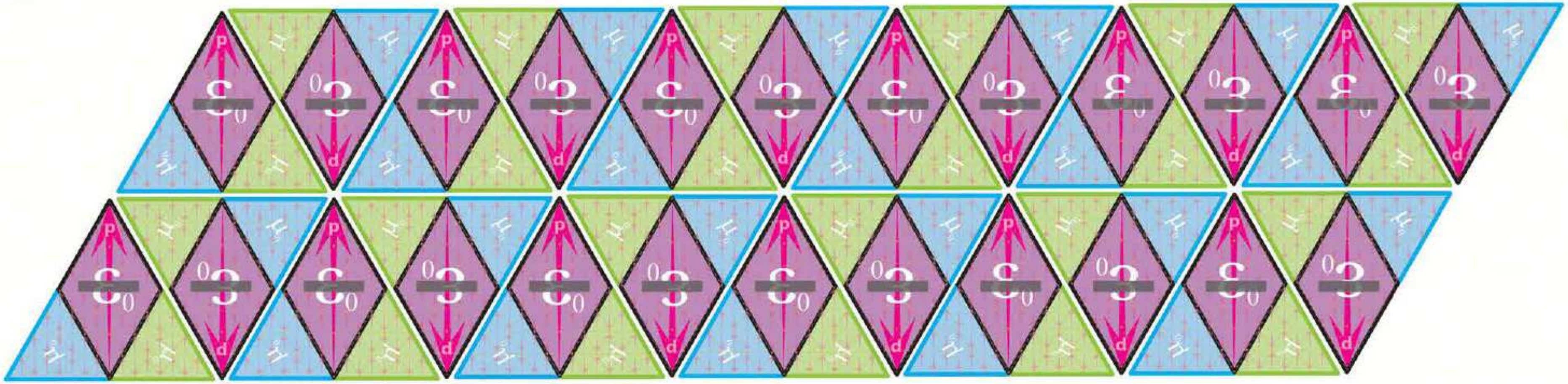

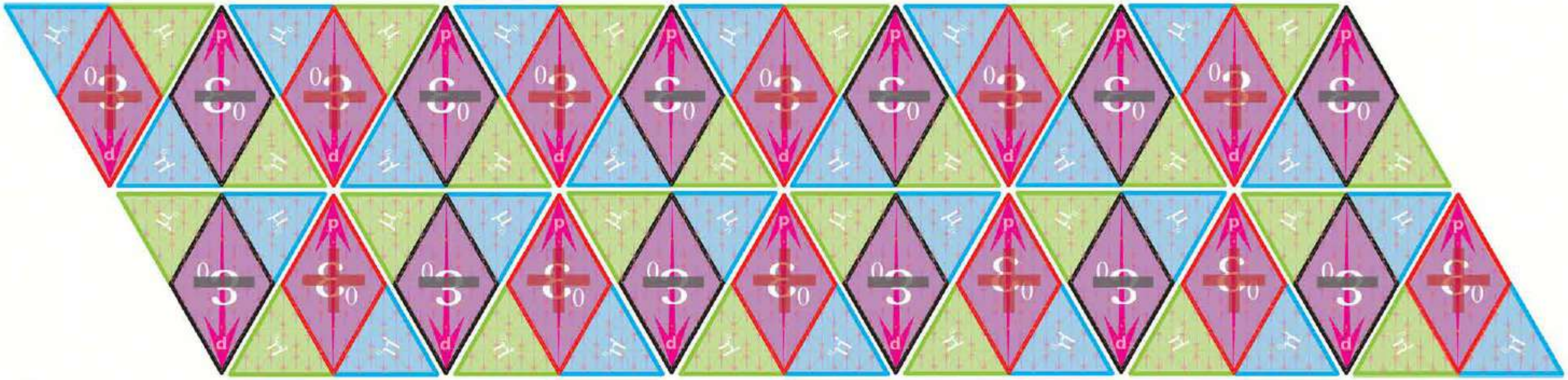
p
d
μ_0
ε_0

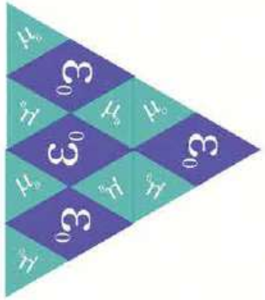
μ_0
ε_0

2D radiant EM mass-Energies

n8 - Neutral Tetryon Templates

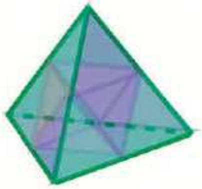

3D standing-wave Matter

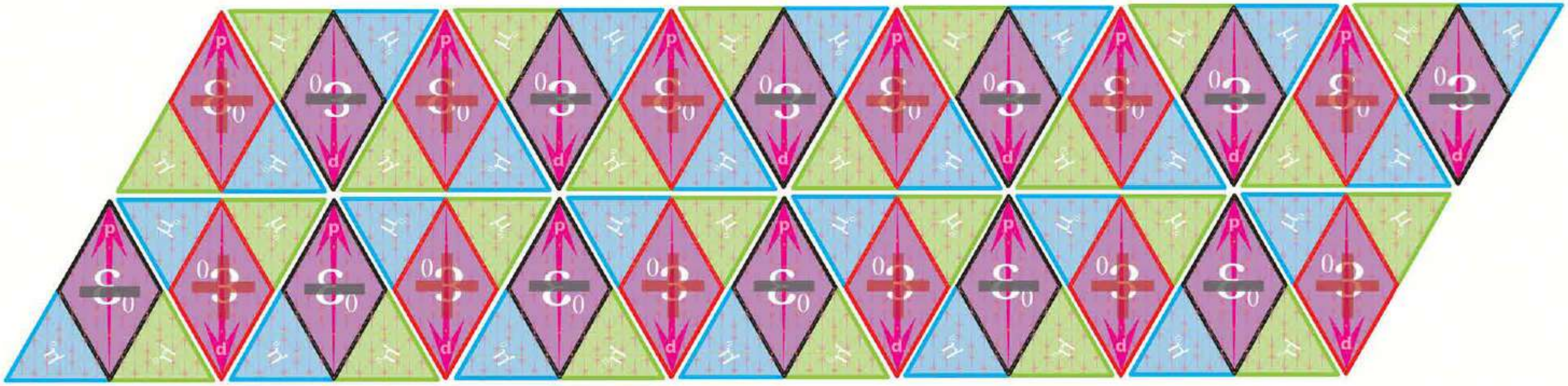
p
d
μ_0
ε_0

Copyright ABRAHAM [2008] - All rights reserved

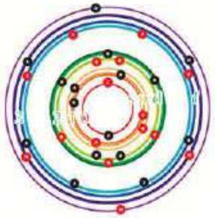

Each shell consists of one or more subshells, and each subshell consists of one or more atomic orbitals.

Periodic element geometries

An electron shell may be thought of as an orbit followed by electrons around an atom's nucleus. The closest shell to the nucleus is called the "1 shell" (also called "K shell"), followed by the "2 shell" (or "L shell"), then the "3 shell" (or "M shell"), and so on further and further from the nucleus. The shell letters K, L, M, ... are alphabetical

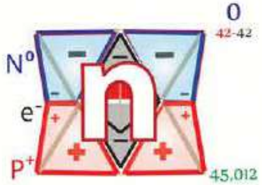
0
42-42
N°
e⁻
n
P⁺
45,012

Each shell can contain only an integer number of whole deuterium nuclei [Proton, Neutron & electron]

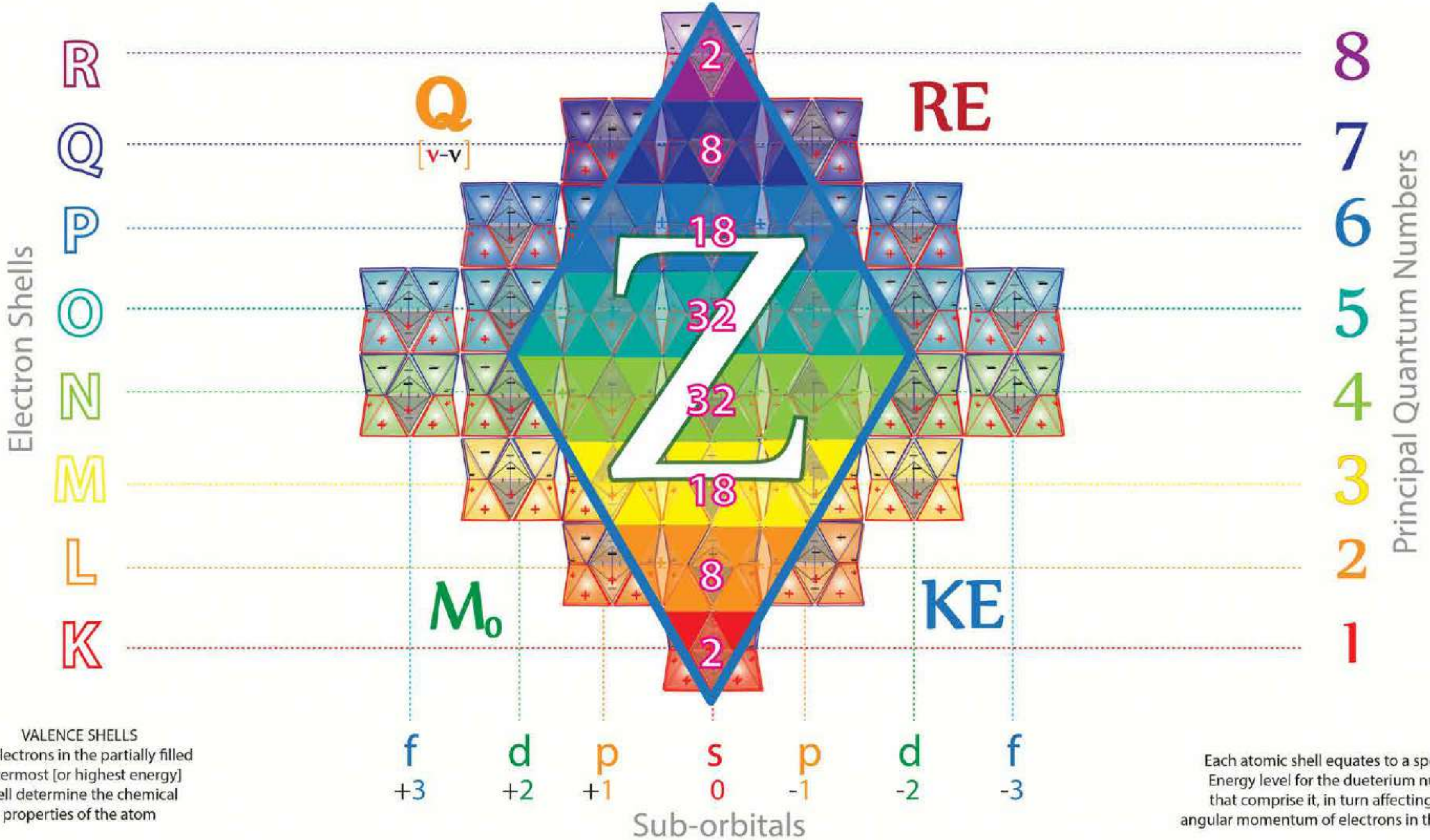
Electron Shells
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
M₀
KE
2
8
18
Z
32
32
18
8
2
8
7
6
5
4
3
2
1
Principal Quantum Numbers
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
Sub-orbitals

VALENCE SHELLS
The electrons in the partially filled outermost [or highest energy] shell determine the chemical properties of the atom

Each atomic shell equates to a specific Energy level for the dueterium nuclei that comprise it, in turn affecting the angular momentum of electrons in that shell

Periodic Harmonic motions

$x = A \cos(\omega t + \varphi)$

Circular motion

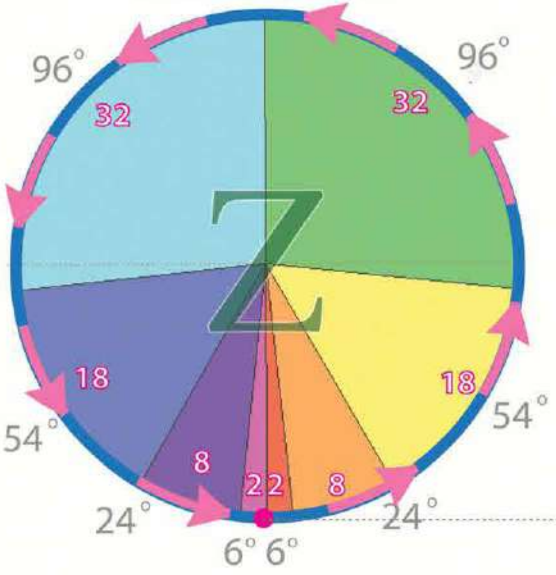
96°
32
96°
32
Z
18
18
54°
54°
8
2 2
8
24°
24°
6° 6°

circular harmonic motion

Circular motions describe the motion of a body with a changing velocity vector [the result of an acceleration force].

Much of the math in of modern physics is predicated on the assumption that π [where it appears] is related to the properties of a circle

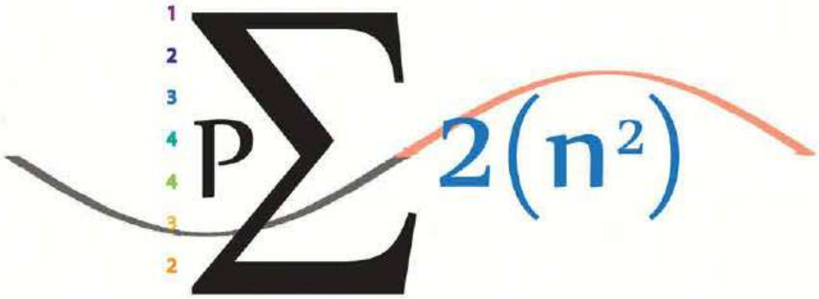
1
2
3
4
4
3
2
$P\sum 2(n^2)$

Simple harmonic motion can be visualized as the projection of uniform circular motion onto one axis

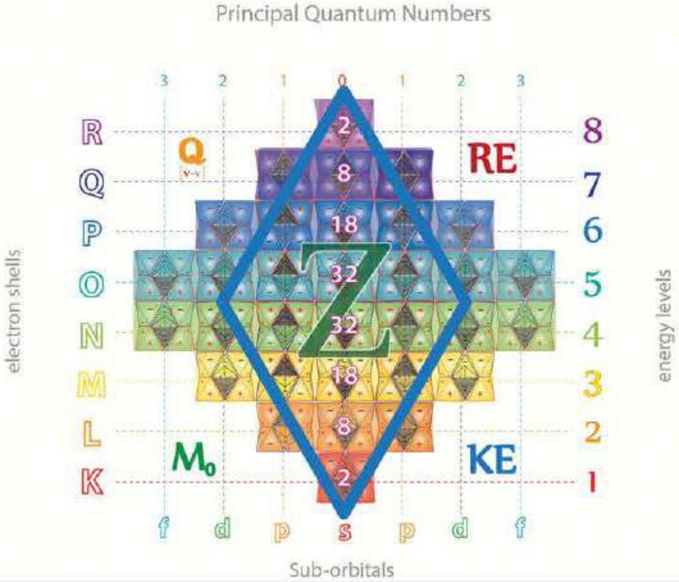
Principal Quantum Numbers
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M_0
KE
2
8
18
32
Z
32
18
8
2
8
7
6
5
4
3
2
1
electron shells
energy levels
f d p s p d f
Sub-orbitals

$F = -kx$

Linear motion

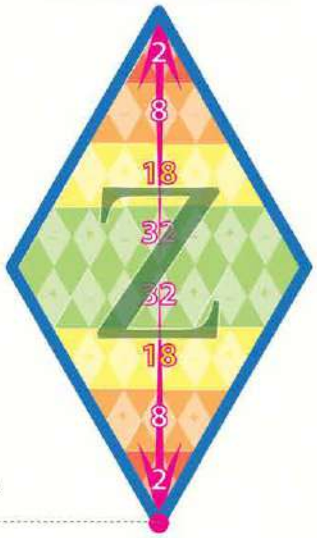
2
8
18
32
Z
32
18
8
2

simple harmonic motion

Nuclei per shell in elements follows a 'periodic summation rule' that is reflective of photonic energies

E

The quantum level mass-energies of Baryons determines the kinetic energies of electrons

1n - 1 e19v

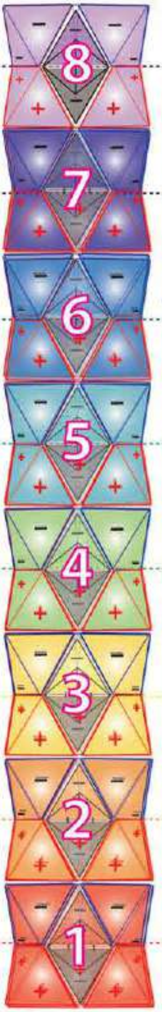

Baryon mass-energies	n	E		electron KEM field hv^2
73,728	n32	8	v8	3.268969693 e15
69,192	n31	7	v7	2.502804921 e15
64,800	n30	6	v6	1.838795452 e15
60,552	n29	5	v5	1.276941286 e15
56,448	n28	4	v4	8.172424232 e14
52,488	n27	3	v3	4.596988631 e14
48,672	n26	2	v2	2.043106058 e14
45,000	n25	1	v1	5.107765145 e13
Compton frequencies		**1.2 e20 electron rest Matter is velocity invariant**		**spectral frequencies**
930 MeV		**496 keV**		**13.6 eV**

Atomic nuclei mass-energies

Each element's weight [mass-Matter in a gravitational field] is the result of the total quanta comprising that element

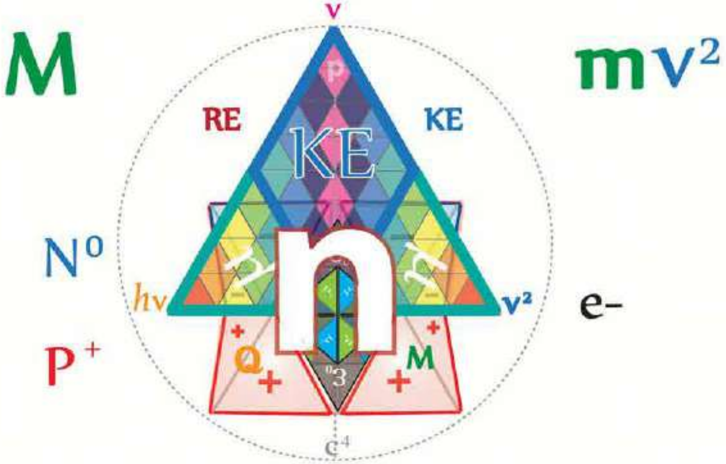

The nuclei forming each atomic shell have specific mass-energy quanta

Baryon rest masses lepton rest mass KEM

$$\overset{8}{\underset{1}{n}}\left[\,[72(n)^{2}]+[12e19]+[m_{e}v^{2}]\,\right]$$

Deuterium mass-energy per shell

Despite having differing mass-energies each Deuterium nuclei has the same velocity invariant Matter geometry [84π]

spin orbital coupling in synchronous quantum convertors

Electrons act as quantum scale rotating armatures in atomic nuclei and can only have specific energies reflective of the electron orbital energy level of the Baryons in which they are found

They acheive these energy levels by absorbing or emitting photons to acheive the specific angular momentum required

Baryons 930.947 MeV + **KEM fields** 13.525 ev + **electrons** 496.519 keV

Mapping Planck mass-energy contributions to elementary Matter and isotopes

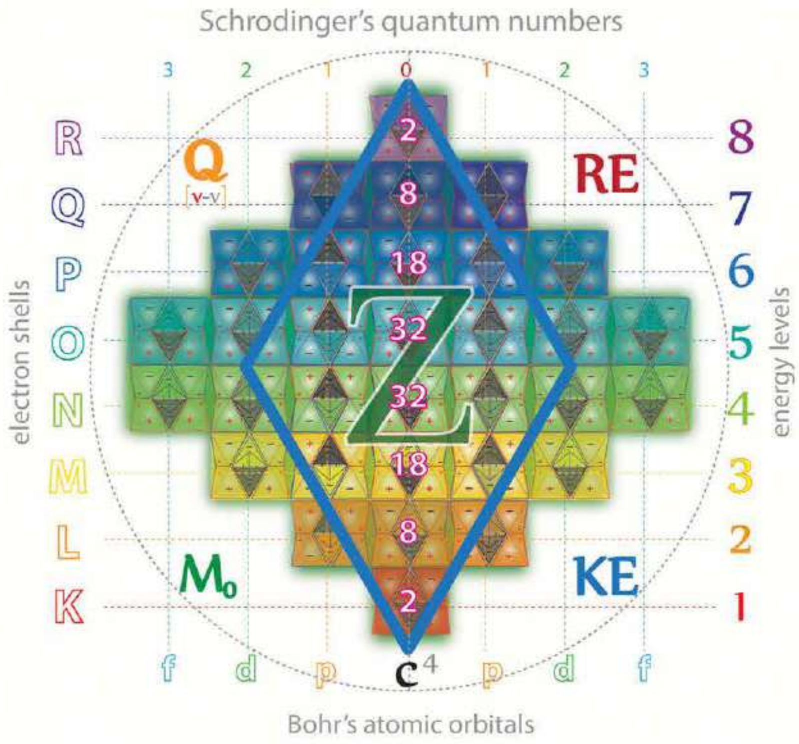
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M₀
KE
2
8
18
32
Z
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f
d
p
c⁴
p
d
f
Bohr's atomic orbitals

$$E = hv^2$$

$$n^2 + v^2 + e = Z$$

general form quatratic equation

$$ax^2 + bx + c = Z$$

$$E = nhv$$

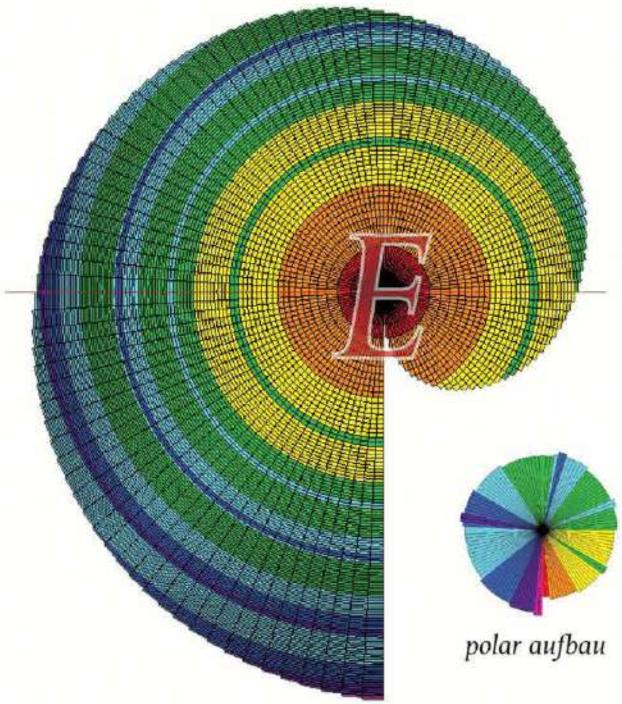
E
polar aufbau

polar energy spirals courtesy of Rene Cormier

Identifying electron rest Matter topologies as velocity invariant we can re-arrange the component Planck mass-energy geometry formulation of periodic elements to

$$h\left[72[v^2] + v + 1.20\text{e}20\ v\right]$$

Deuteron rest mass + Spectral lines + electron rest mass

reveal a quadratic formulation for all Z numbers

Copyright ABRAHAM [2008] - All rights reserved

$$\sum_{K}^{R}$$

nuclei per shell

$$P\sum_{K}^{R} Z$$

Periodic Summation

Periodic summation is a notation developed for Tetryonic theory to model the geometric series addtion of Z[n²] energy level Deuterium nuclei that form the periodic elements

$$\sum_{1}^{120} Z$$

element number

STEP ONE

Periodic summation follows the atomic shell electron config

1 R
2
3
4
4
3
2
1 K

$$P\sum_{K}^{R} 2(x^2) =$$

Each atomic shell can hold only a fixed number of deuterium nuclei

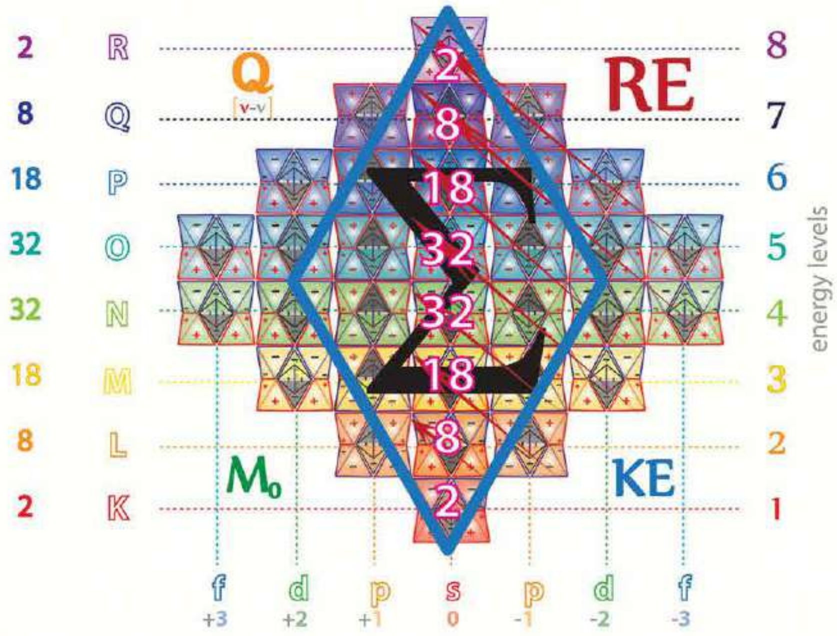

STEP TWO

Periodic elements build up following the aufbau sequence

$P\sum R = 2$	2 nuclei [74.496 ea]	120	Unbinilium
	+		
$P\sum Q = 8$	8 nuclei [69,780 ea]	118	Ununoctium
	+		
$P\sum P = 18$	18 nuclei [65,232 ea]	110	Darmstadtium
	+		
$P\sum O = 32$	32 nuclei [60,852 ea]	92	Uranium
	+		
$P\sum N = 32$	32 nuclei [56,640 ea]	60	Neodymuim
	+		
$P\sum M = 18$	18 nuclei [52,596 ea]	28	Argon
	+		
$P\sum L = 8$	8 nuclei [48,720 ea]	10	Neon
	+		
$P\sum K = 2$	2 nuclei [45,012 ea]	2	Helium
		0	Hydrogen

THe LHS of the notation determine the number of nuclei in each atomic shell, from the periodic mass-energy levels for atoms, and the RHS follows the aufbau building principle to determine the rest mass-Matter of any specific element

Each periodic element is made of Z [n² energy] deuterium nuclei

Aufbau

Z# [z Protons [24-12]
z Neutrons [18-18]
z electrons [0-12]] n1-8

Planck mass-energies form the surface integral of rest Matter topologies for each periodic element

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number.
It is used as rule to determine which nuclei are stable or unstable.

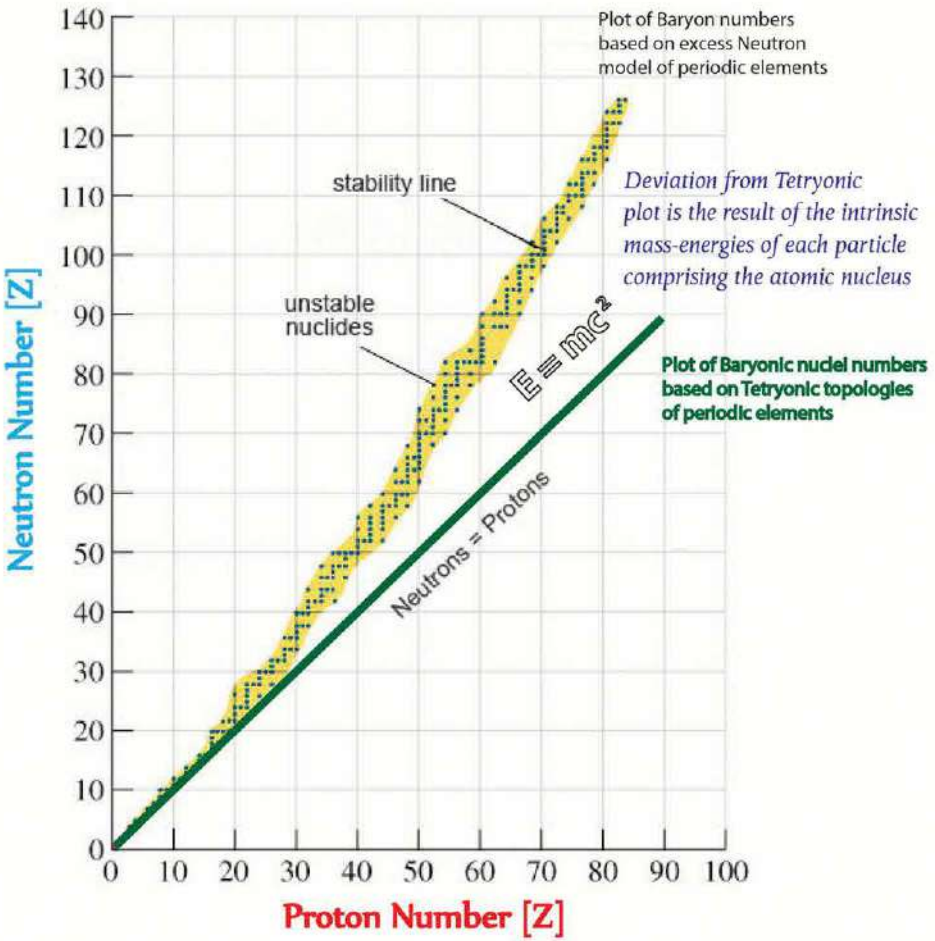

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

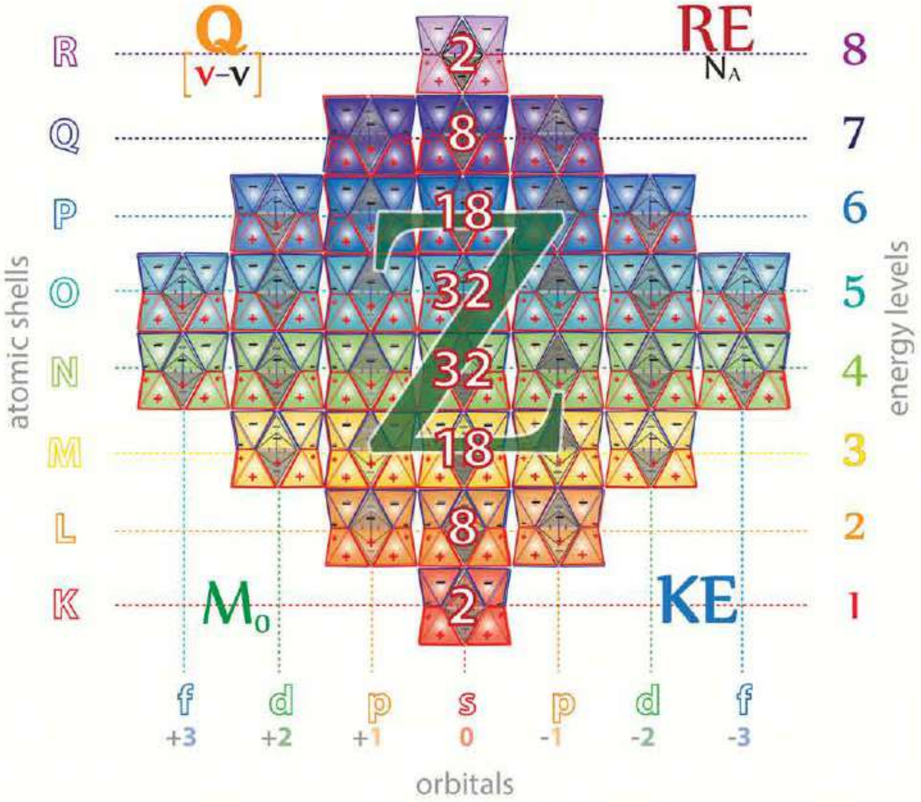
Q
[v-v]
RE
N_A
R
Q
P
O
N
M
L
K
atomic shells
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
energy levels
M_0
KE
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
orbitals

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

Planck mass-energy contributions to elementary Matter and isotopes

The mass-energy content of Deuterium nuclei creates the molar mass of elements
[not extra neutrons in excess of the elemental number]

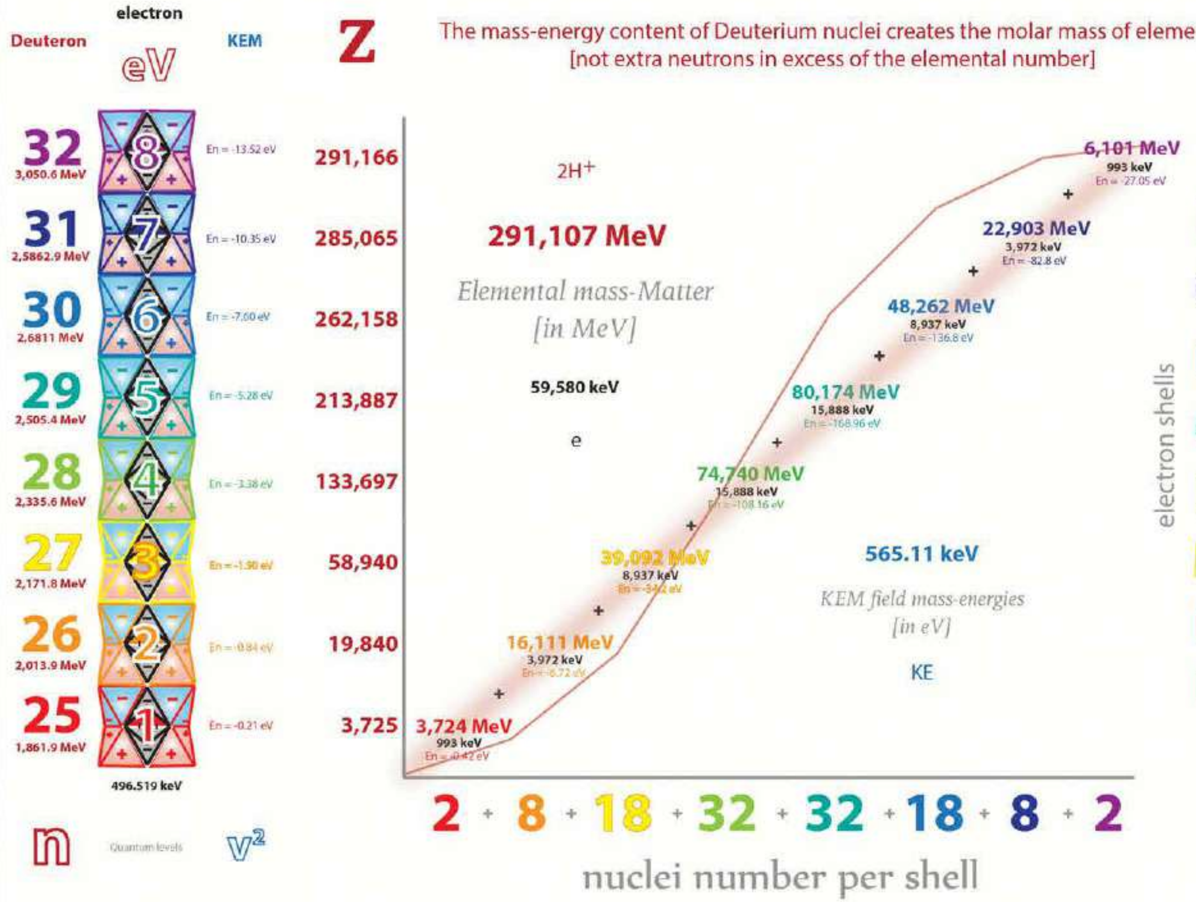
Deuteron
electron
eV
KEM
Z
32
3,050.6 MeV
8
En = -13.52 eV
291,166
31
2,5862.9 MeV
7
En = -10.35 eV
285,065
30
2,6811 MeV
6
En = -7.00 eV
262,158
29
2,505.4 MeV
5
En = -5.28 eV
213,887
28
2,335.6 MeV
4
En = -3.38 eV
133,697
27
2,171.8 MeV
3
En = -1.90 eV
58,940
26
2,013.9 MeV
2
En = -0.84 eV
19,840
25
1,861.9 MeV
1
En = -0.21 eV
3,725
496.519 keV
n
Quantum levels
v²
2H+
291,107 MeV
Elemental mass-Matter
[in MeV]
59,580 keV
e
6,101 MeV
993 keV
En = -27.05 eV
22,903 MeV
3,972 keV
En = -82.8 eV
48,262 MeV
8,937 keV
En = -136.8 eV
80,174 MeV
15,888 keV
En = -168.96 eV
74,740 MeV
15,888 keV
En = -108.16 eV
39,092 MeV
8,937 keV
En = -34.2 eV
16,111 MeV
3,972 keV
En = -6.72 eV
3,724 MeV
993 keV
En = -0.42 eV
565.11 keV
KEM field mass-energies
[in eV]
KE
2 + 8 + 18 + 32 + 32 + 18 + 8 + 2
nuclei number per shell

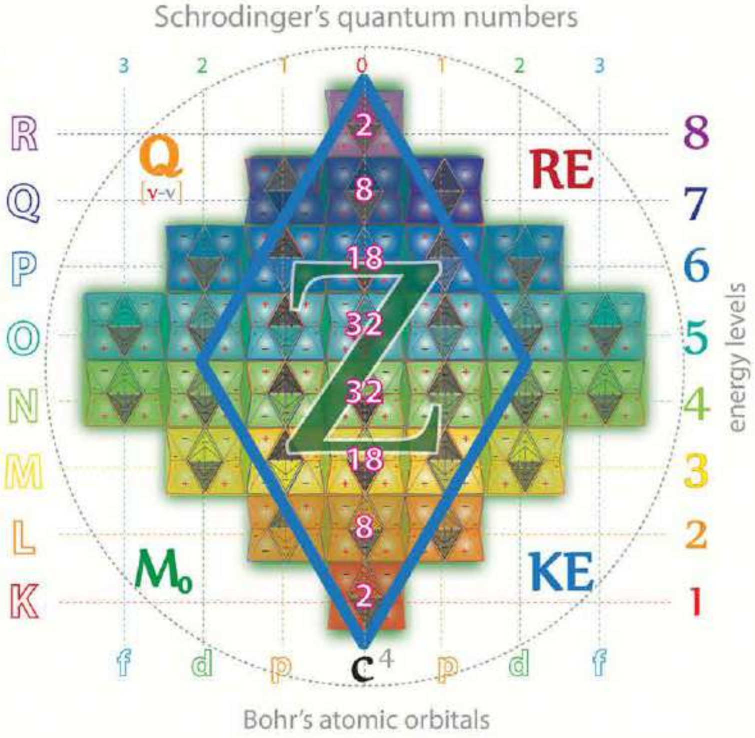
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
Z
2
8
18
32
32
18
8
2
M₀
KE
8
7
6
5
4
3
2
1
energy levels
f
d
p
c⁴
p
d
f
Bohr's atomic orbitals

n1

Elementary nuclei

are comprised of equal numbers of
Protons, Neutrons & electrons
with varying energy levels

Baryons		electrons		KEM fields
930.947 MeV	+	496.519 keV	+	13.525 ev

The mass-energy content of Matter topologies is velocity invariant

The mass-energy content of Baryons determines the KEM field of electrons

Copyright ABRAHAM [2008] - All rights reserved

All elements are comprised of n level Duetrium nuclei

Baryons	KEM fields	electrons
$Z [[72n^2]$	$+ [12v^2]$	$+ [1.2e20]]$
1,861,949 MeV	13.525 eV	496,519 keV

$2H^+$ $[72n^2]$ Z γ $[M_e v^2]$

Elemental mass-Matter
[in MeV]

The atomic shell energy levels of Deuterium nuclei in elements

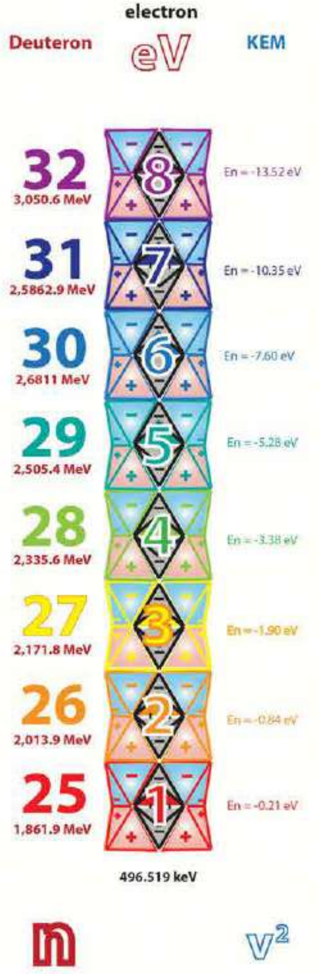
electron
Deuteron
eV
KEM
32
3,050.6 MeV
8
En = -13.52 eV
31
2,5862.9 MeV
7
En = -10.35 eV
30
2,6811 MeV
6
En = -7.60 eV
29
2,505.4 MeV
5
En = -5.28 eV
28
2,335.6 MeV
4
En = -3.38 eV
27
2,171.8 MeV
3
En = -1.90 eV
26
2,013.9 MeV
2
En = -0.84 eV
25
1,861.9 MeV
1
En = -0.21 eV
496.519 keV
m
v^2

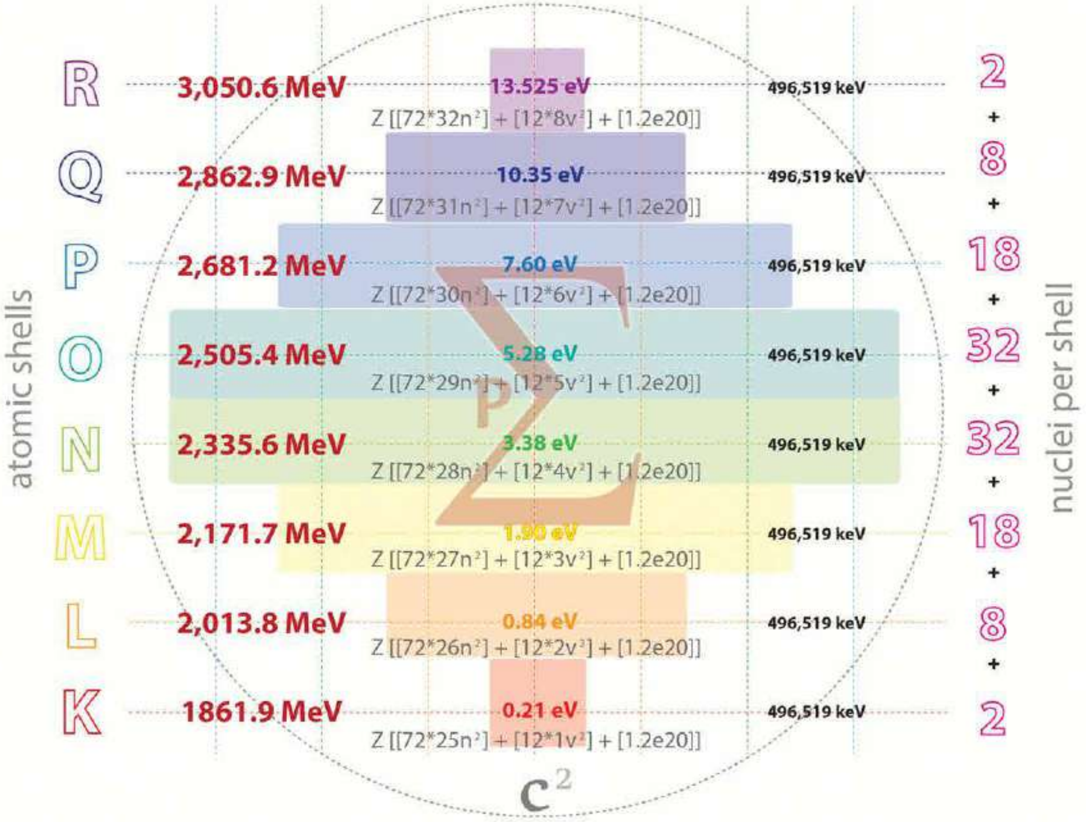
atomic shells
nuclei per shell
P
Σ
Σ
c^2

R	3,050.6 MeV	13.525 eV	$Z [[72*32n^2] + [12*8v^2] + [1.2e20]]$	496,519 keV	2
					+
Q	2,862.9 MeV	10.35 eV	$Z [[72*31n^2] + [12*7v^2] + [1.2e20]]$	496,519 keV	8
					+
P	2,681.2 MeV	7.60 eV	$Z [[72*30n^2] + [12*6v^2] + [1.2e20]]$	496,519 keV	18
					+
O	2,505.4 MeV	5.28 eV	$Z [[72*29n^2] + [12*5v^2] + [1.2e20]]$	496,519 keV	32
					+
N	2,335.6 MeV	3.38 eV	$Z [[72*28n^2] + [12*4v^2] + [1.2e20]]$	496,519 keV	32
					+
M	2,171.7 MeV	1.90 eV	$Z [[72*27n^2] + [12*3v^2] + [1.2e20]]$	496,519 keV	18
					+
L	2,013.8 MeV	0.84 eV	$Z [[72*26n^2] + [12*2v^2] + [1.2e20]]$	496,519 keV	8
					+
K	1861.9 MeV	0.21 eV	$Z [[72*25n^2] + [12*1v^2] + [1.2e20]]$	496,519 keV	2

6,101 MeV
993 keV
En = -27.05 eV
+
22,903 MeV
3,972 keV
En = -82.8 eV
+
48,262 MeV
8,937 keV
En = -316.8 eV
+
80,174 MeV
15,888 keV
En = -168.96 eV
+
74,740 MeV
15,888 keV
En = -108.16 eV
+
39,092 MeV
8,937 keV
En = -34.2 eV
+
16,111 MeV
3,972 keV
En = -6.72 eV
+
3,724 MeV
993 keV
En = -0.42 eV

[1.2e20]]

e

Determines the spectral line [KEM field energies] of electrons bound to them

The relativistic rest mass-energy-Matter of all periodic elements

is the sum of the mass-energies of all atomic nuclei and spectral lines
that comprise its mass-Matter topology as measured in
any spatial co-ordinate system per unit of time

the rest mass-Matter of bound photo-electrons is velocity invariant

Ionisation energies

Fractional quantum differentials

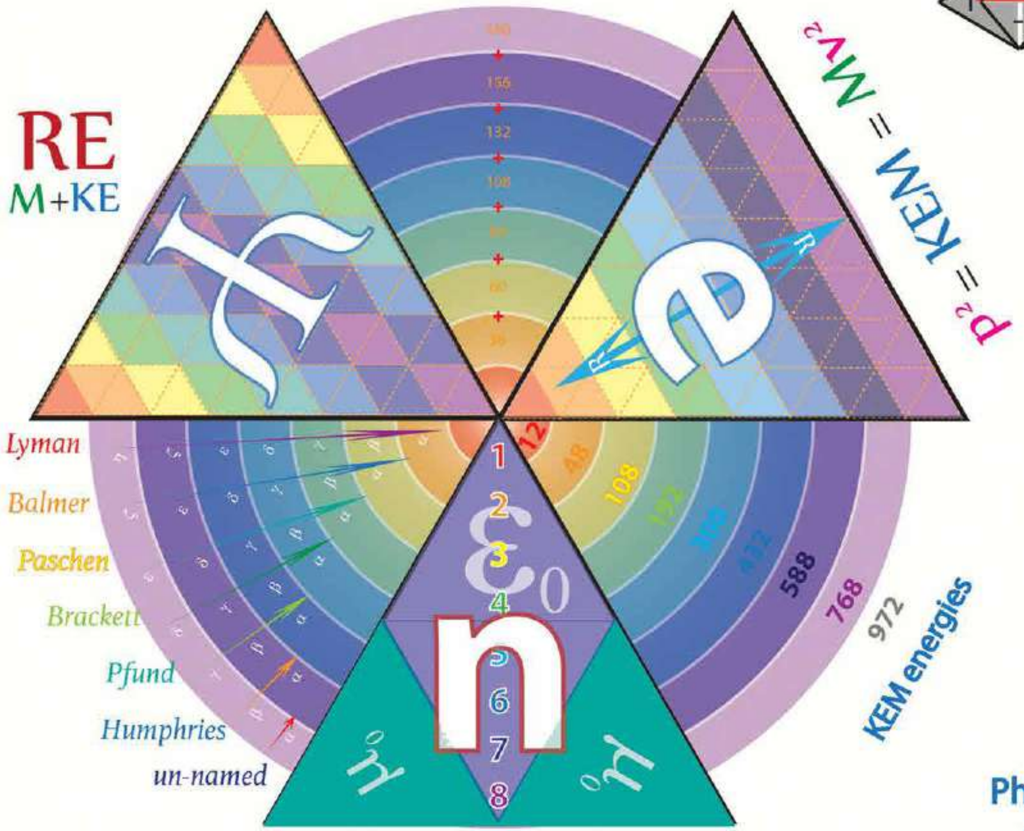
RE
M+KE
Mv^2 = KEM = p^2
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
1
2
3
4
5
6
7
8
ε_0
n
12
48
108
192
300
432
588
768
972
KEM energies

$$hf = \Delta Mv = \Delta p$$

spectral lines are produced by accelerating electrons

Note: this is an Illustrative schema for modelling KEM field energies
All KEM fields possess the same physical spatial geometry
in radial-time defined spatial co-ordinate systems

c^2 c^4

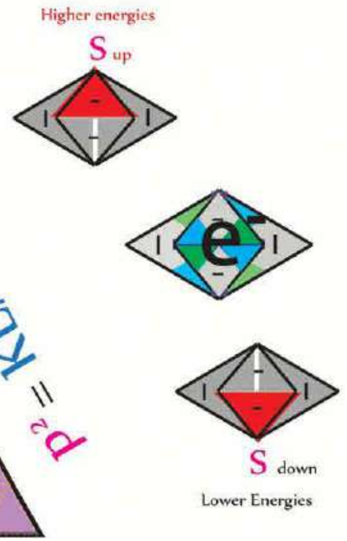
Higher energies
S up
e
S down
Lower Energies

photo-electronic energy momenta

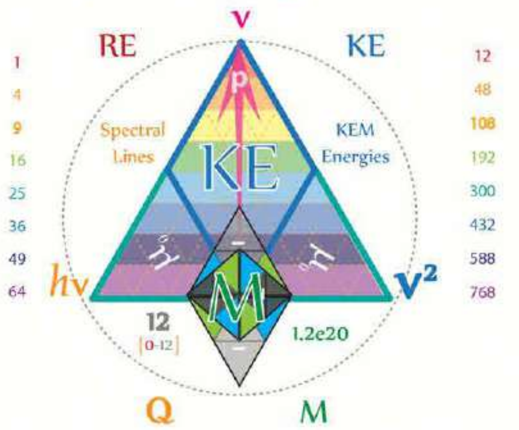
v
RE
KE
Spectral Lines
KEM Energies
KE
M
hv
v^2
12
[0-12]
1.2e20
Q
M
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768

Mapping photo-electron transition energies
to Tetryonic energy momenta geometries
reveals many key facts about the
ionisation energies of nuclei

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

The differing fractional KEM field energy momenta of electrons
that results from their transitions to specific energy nuclei
in elements results in differing QAM quanta and
produces spectral lines and fine line splitting

Photo-electrons absorb/emit spectral energies

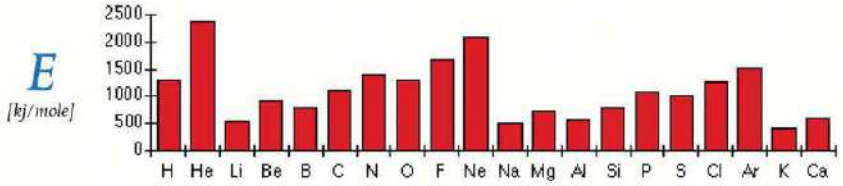

$$E = eV = \frac{1}{4\pi\epsilon_0}\frac{ne^2}{a}$$

Spectral line transitions

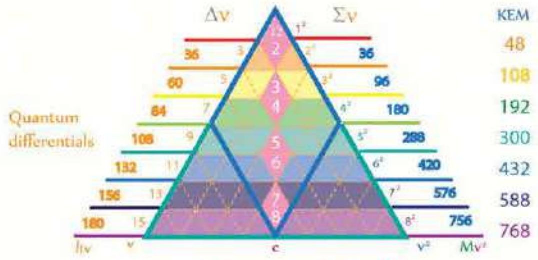
Δv
Σv
Quantum differentials
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
36
96
180
288
420
576
756
hv
v
c
v^2
Mv²

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Copyright ABRAHAM [2008] - All rights reserved

Element numbers

Nuclei per shell in elements follow
a 'periodic summation rule'
that is reflective of
photonic energies

1 R 2 3 4 4 3 2 1 K

$$P\sum_{K}^{R} 2(x^2) = 2+8+18+32+32+18+8+2$$

Z

120	Unbinilium	n8
119	Ununennium	
118	Ununoctium	n7
87	Francium	
112	Copernicium	n6
55	Caesium	
102	Nobelium	n5
37	Rubidium	
70	Ytterbuim	n4
19	Potassium	
30	Zinc	n3
11	Sodium	
10	Neon	n2
3	Lithium	
2	Helium	n1
1	Deuterium	

Hydrogen 0

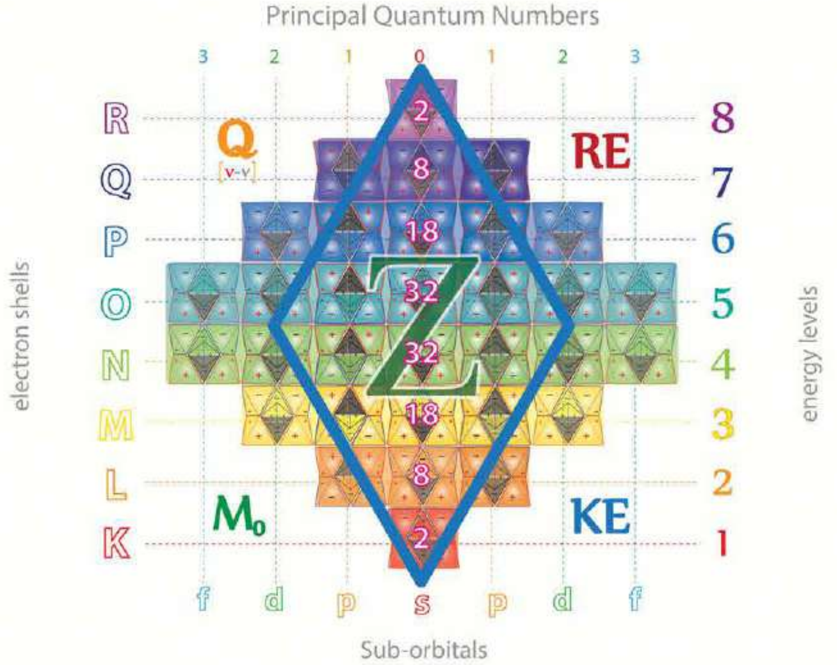
Principal Quantum Numbers
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
electron shells
energy levels
Q
[v-v]
RE
M₀
KE
2 8 18 32 32 18 8 2
Z
f d p s p d f
Sub-orbitals

Periodic mass-ENERGY-Matter

Following periodic summation rules for shell filling
n[1-8] quantum energy deuterium nuclei
combine to form elementary Matter

$$P\sum_{K}^{R}{}_{25}^{32}\Big[\overset{\text{Baryon rest masses}}{[72(n)^2]}+\overset{\text{lepton rest mass}}{[12e19]}+\overset{\text{KEM}}{[m_e v^2]}\Big]_{1}^{8}$$

Deuterium mass-energy per shell

The measured weight of Matter in gravitational fields is the result of
planar mass-energies in tetryonic standing-wave geometries

The periodicity of all the elements,
along with their exact molar rest mass-energies and
quantum wavefunctions can be described with Tetryonic geometries

Ionisation energies

γ

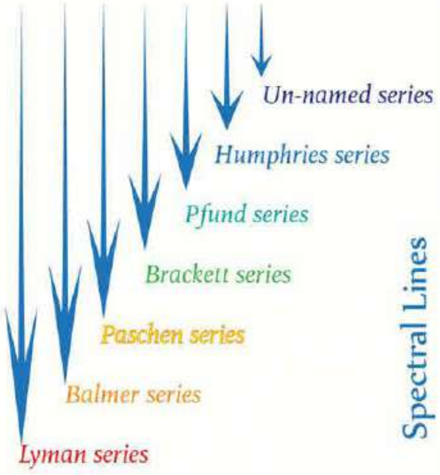
Un-named series
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
Spectral Lines

$$Mv^2 = KEM = hcR_H$$

Photon emission/absorption

Spectral line differentials

$$\Delta hv = \Delta Mv = \Delta p$$

changes to energy momenta

accelerate photo-electrons

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg's formula is
a mathematical description
of Tetryonic energy geometry

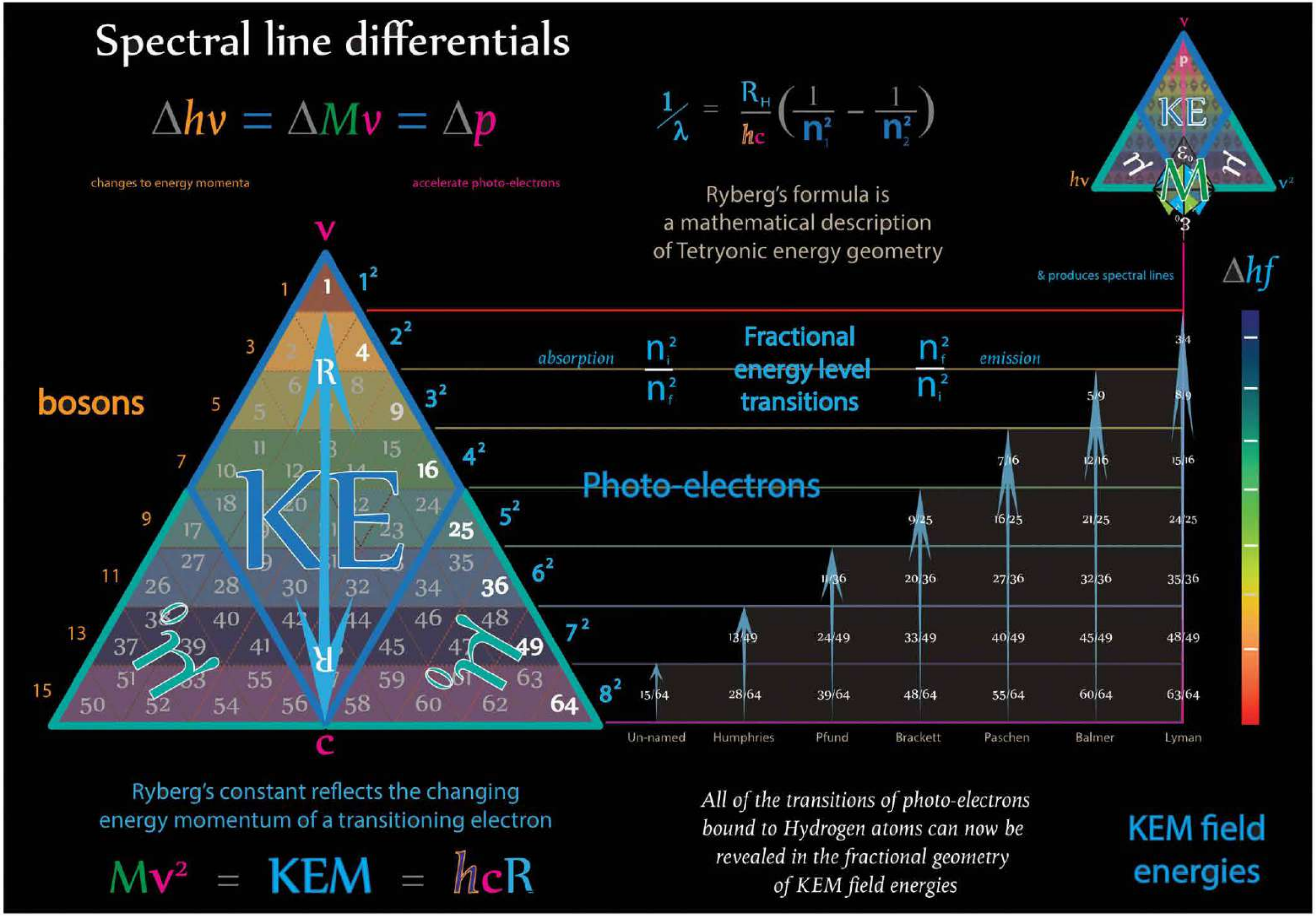
v
P
KE
ε_0
M
hv
v^2
3
& produces spectral lines
Δhf
bosons
v
1
3
5
7
9
11
13
15
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
KE
R
c
absorption
$\frac{n_i^2}{n_f^2}$
Fractional energy level transitions
$\frac{n_f^2}{n_i^2}$
emission
Photo-electrons
3/4
8/9
15/16
24/25
35/36
48/49
63/64
5/9
12/16
21/25
32/36
45/49
60/64
7/16
16/25
27/36
40/49
55/64
9/25
20/36
33/49
48/64
11/36
24/49
39/64
13/49
28/64
15/64
Un-named
Humphries
Pfund
Brackett
Paschen
Balmer
Lyman

Ryberg's constant reflects the changing
energy momentum of a transitioning electron

$$Mv^2 = KEM = hcR$$

*All of the transitions of photo-electrons
bound to Hydrogen atoms can now be
revealed in the fractional geometry
of KEM field energies*

KEM field energies

Fermat's method of Factoring

also known as 'the difference of two squares' is used to factorise large numbers

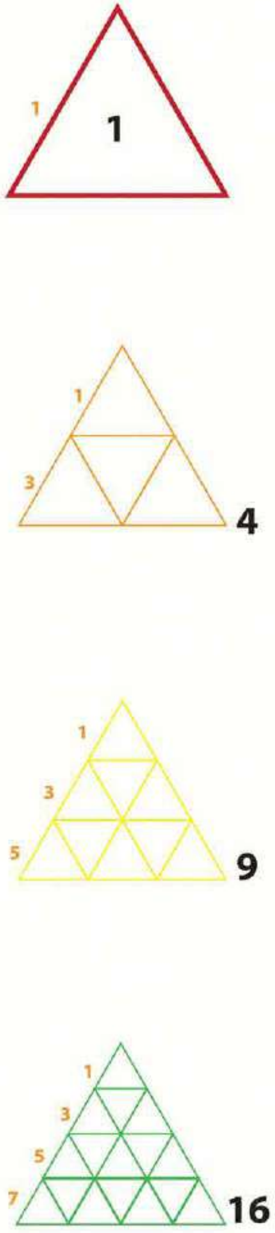

1
1
1
3
4
1
3
5
9
1
3
5
7
16

$$[x\text{-}y].[x+y]$$

$$x^2 - y^2$$

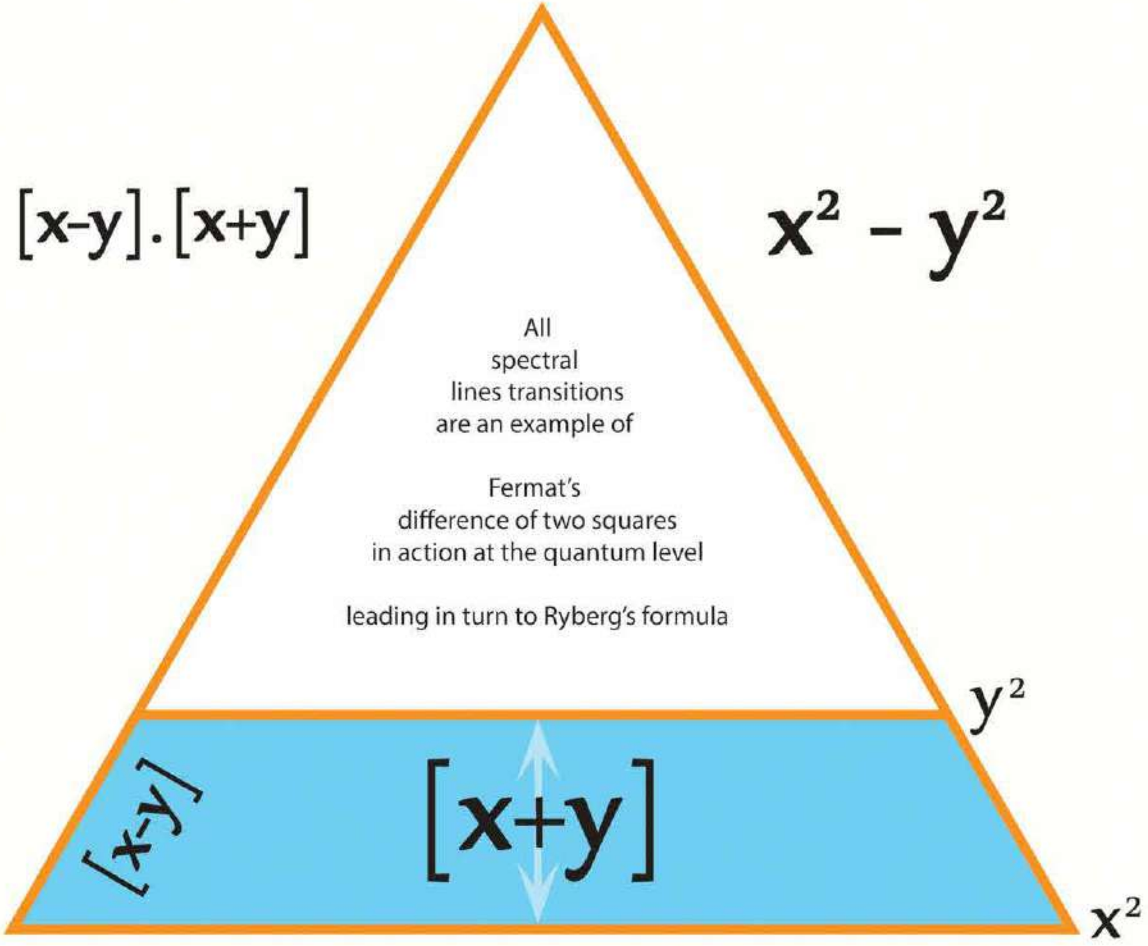

y^2
$[x\text{-}y]$
$[x+y]$
x^2

All
spectral
lines transitions
are an example of

Fermat's
difference of two squares
in action at the quantum level

leading in turn to Ryberg's formula

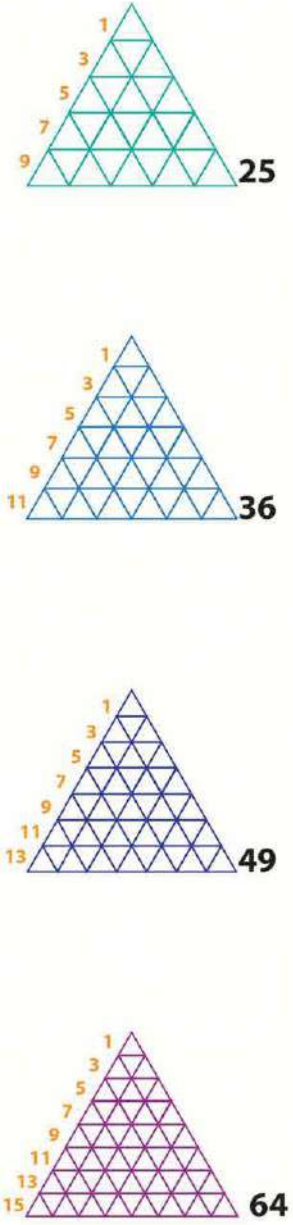

1
3
5
7
9
25
1
3
5
7
9
11
36
1
3
5
7
9
11
13
49
1
3
5
7
9
11
13
15
64

Fermat knew that every odd number could be written as the difference of two squares or as revealed geometrically through Tetryonic theory's equilateral geometry, every 'SQUARE' number is the sequential sum of ODD numbers

Tetryonics 96.02 - Fermat's method of Factoring

Fractions

A fraction is a number that shows how many equal parts there are

3/4

24/25

In quantum mechanics fractions appear in quantum steps
as a result of the equilateral geometry of Planck energy momenta

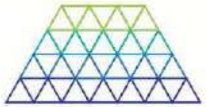

40/49

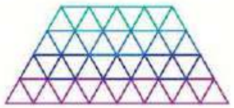

48/64

80/81

.987654321

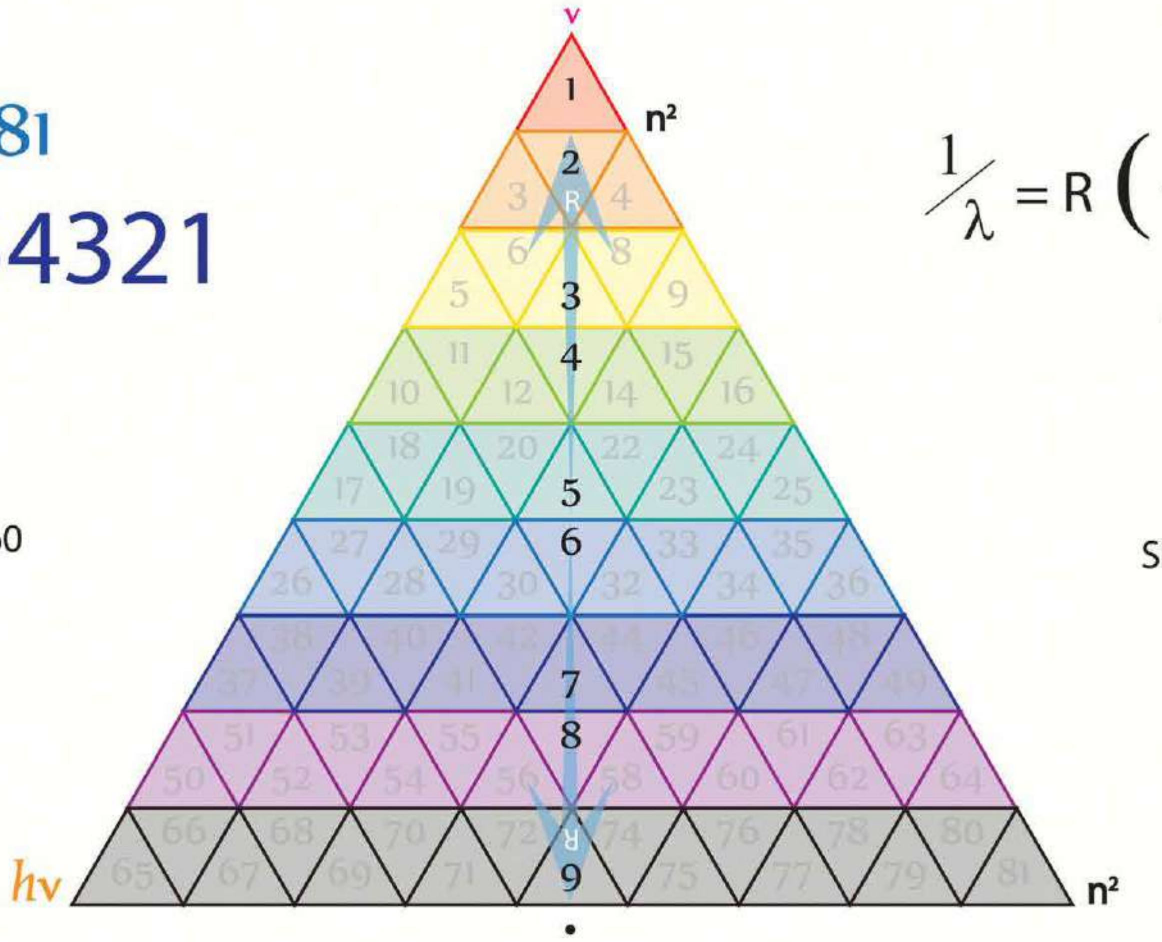
v
n²
1
2
R
3
4
5
6
7
8
9
n²
hv
3 4 6 8 5 9
10 11 12 14 15 16
17 18 19 20 22 23 24 25
26 27 28 29 30 32 33 34 35 36
37 38 39 40 41 42 44 45 46 47 48 49
50 51 52 53 54 55 56 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 74 75 76 77 78 79 80 81

$$\frac{1}{\lambda} = R\left(\frac{1}{1} - \frac{1}{81}\right)$$

.987654321

Cos 60

.5

1/2

Sin 60

.866025403

$\sqrt{3}/2$

Tetryonic geometry explains the fractional mathematics of Rydberg's formula

Lyman spectral transitions

Δhv = KEM = hcR

Planck — 13.525 eV — Rydberg

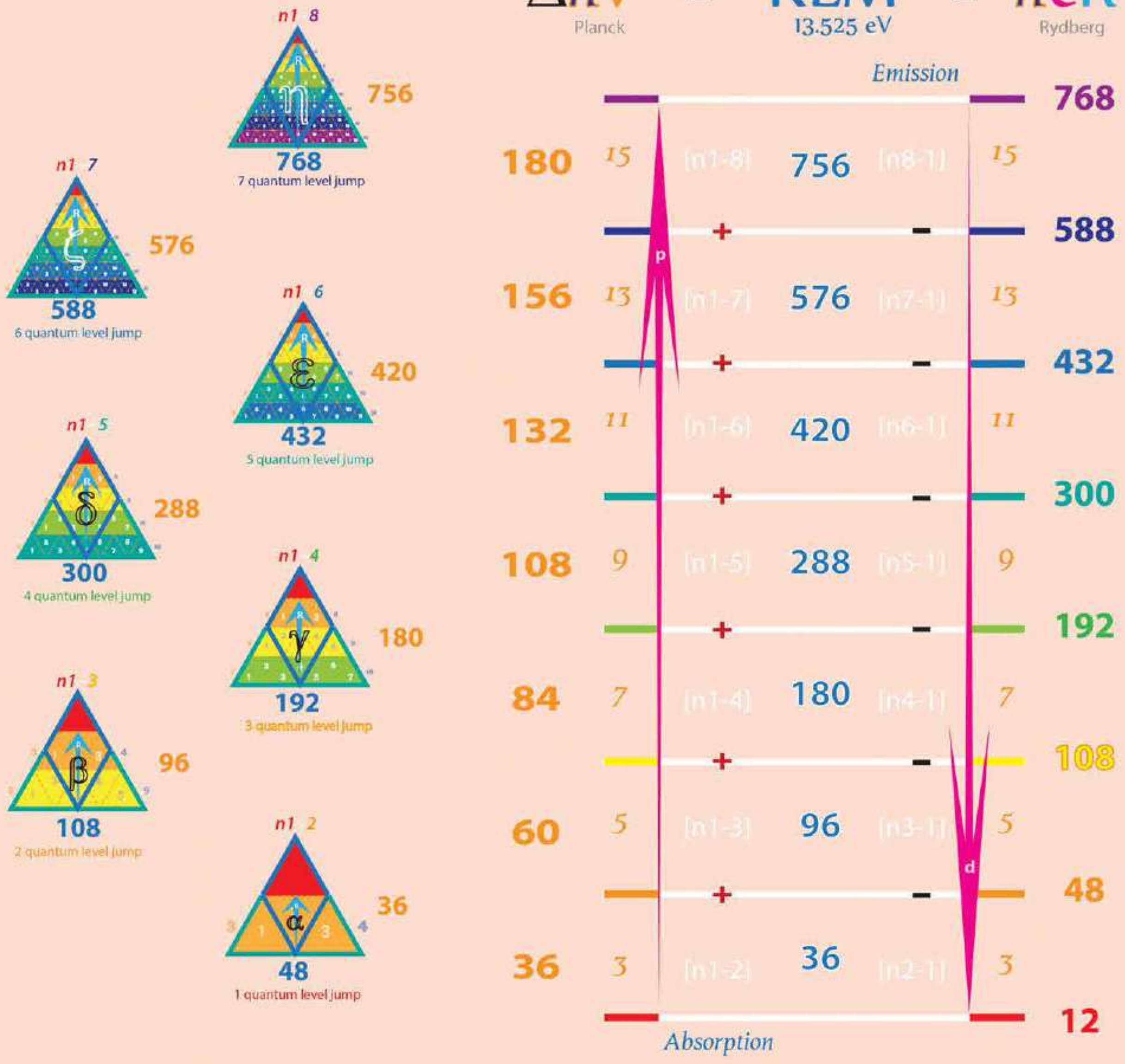
n1 8
756
768
7 quantum level jump
n1 7
576
588
6 quantum level jump
n1 6
420
432
5 quantum level jump
n1 5
288
300
4 quantum level jump
n1 4
180
192
3 quantum level jump
n1 3
96
108
2 quantum level jump
n1 2
36
48
1 quantum level jump
Emission
Absorption
p
d
+
-

						768	63/64	.9843
180	15	[n1-8]	756	[n8-1]	15			
						588	48/49	.9795
156	13	[n1-7]	576	[n7-1]	13			
						432	35/36	.9722
132	11	[n1-6]	420	[n6-1]	11			
						300	24/25	.9600
108	9	[n1-5]	288	[n5-1]	9			
						192	15/16	.9375
84	7	[n1-4]	180	[n4-1]	7			
						108	8/9	.8888
60	5	[n1-3]	96	[n3-1]	5			
						48	3/4	.7500
36	3	[n1-2]	36	[n2-1]	3			
						12		

λ	93.17075598	nm
$\tilde{v}$	10,732,981.5	m-1
f	$3.217666905 \times 10^{15}$	Hz
E	13.31395504	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{64}\right)$ .984375

λ	93.62569132	nm
$\tilde{v}$	10,680,829.01	m-1
f	$3.202031983 \times 10^{15}$	Hz
E	13.24926138	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{49}\right)$ .9795

λ	94.33539045	nm
$\tilde{v}$	10,600,475.55	m-1
f	$3.177942622 \times 10^{15}$	Hz
E	13.14958523	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{36}\right)$ .9722

λ	95.53641971	nm
$\tilde{v}$	10,467,212.43	m-1
f	$3.137991343 \times 10^{15}$	Hz
E	12.98427616	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{25}\right)$ .96

λ	97.82929378	nm
$\tilde{v}$	10,221,5587.14	m-1
f	$3.064444671 \times 10^{15}$	Hz
E	12.67995718	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{16}\right)$ .9375

λ	103.1793333	nm
$\tilde{v}$	9,691,863.362	m-1
f	$2.90554754 \times 10^{15}$	Hz
E	12.02247792	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{9}\right)$ .888

λ	122.2866172	nm
$\tilde{v}$	8,177,509.712	m-1
f	$2.451555737 \times 10^{15}$	Hz
E	10.14396575	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{4}\right)$ .75

ΔMv^2

Tetryonics 96.04 - Lyman transitions

Copyright ABRAHAM [2008] - All rights reserved

Balmer spectral transitions

$$\Delta h\nu = \text{KEM} = hcR$$

Planck — 3.381 eV — Rydberg

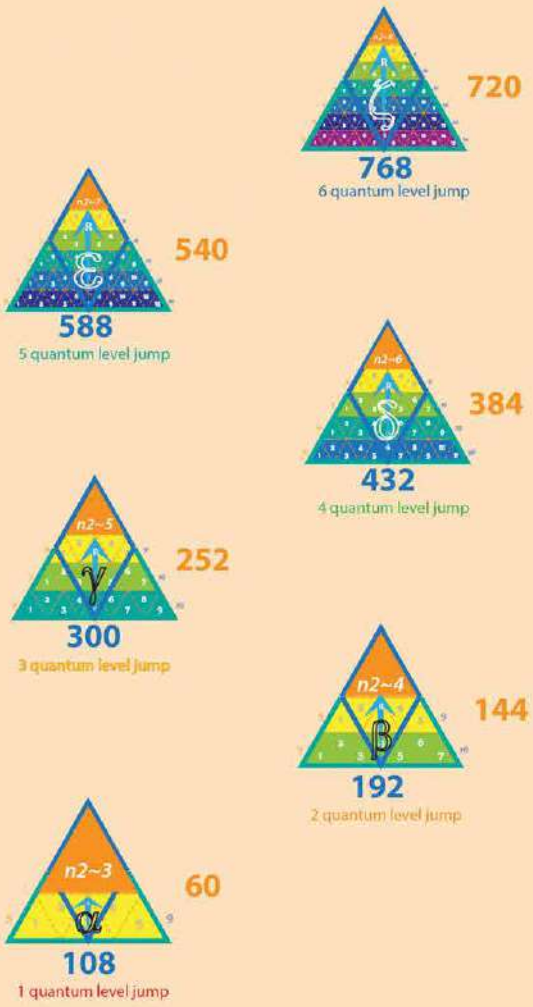

720
768
6 quantum level jump

540
588
5 quantum level jump

384
432
4 quantum level jump

252
300
3 quantum level jump

144
192
2 quantum level jump

60
108
1 quantum level jump

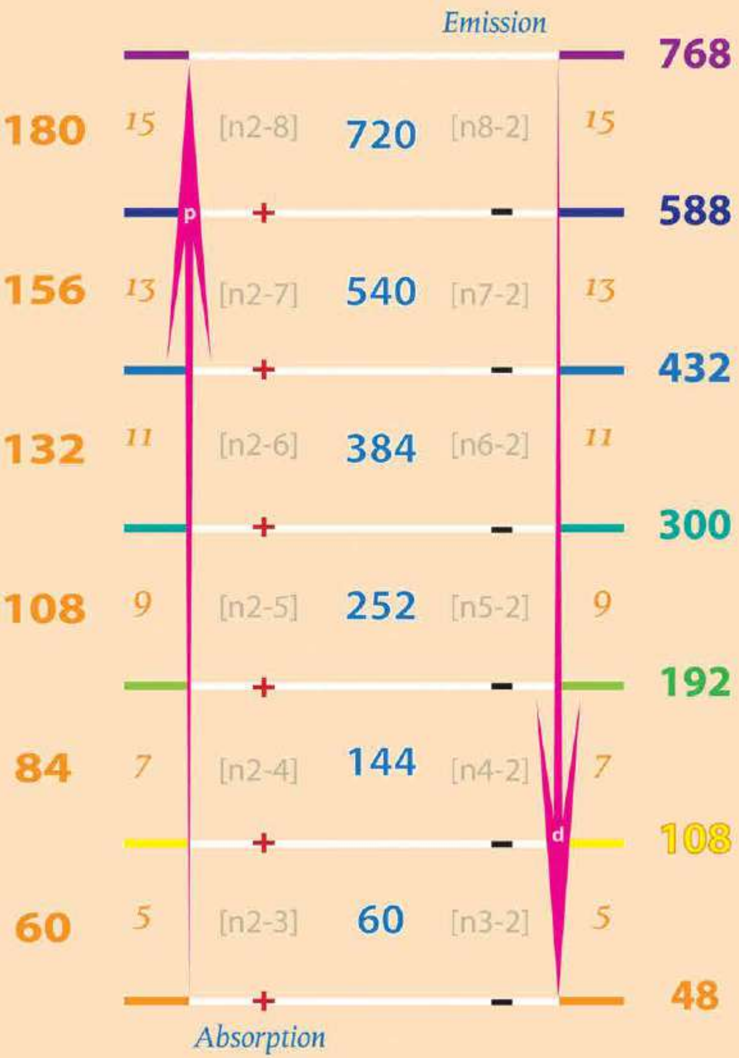

Emission

						768
180	15	[n2-8]	720	[n8-2]	15	
		+		-		588
156	13	[n2-7]	540	[n7-2]	13	
		+		-		432
132	11	[n2-6]	384	[n6-2]	11	
		+		-		300
108	9	[n2-5]	252	[n5-2]	9	
		+		-		192
84	7	[n2-4]	144	[n4-2]	7	
		+		-		108
60	5	[n2-3]	60	[n3-2]	5	
		+		-		48

Absorption

60/64
.2343

λ	391.3171751	nm
$\tilde{\nu}$	2,555,471.1785	m-1
f	7.661111677 x 10^{14}	Hz
E	3.169989296	eV
Ω	[illegible]	[illegible]

$R\left(\frac{1}{4}-\frac{1}{64}\right)$
.2343

45/49
.2295

λ	399.4696163	nm
$\tilde{\nu}$	2,5033,319.299	m-1
f	7.50476246 x 10^{14}	Hz
E	3.105295637	eV
Ω	[illegible]	[illegible]

$R\left(\frac{1}{4}-\frac{1}{49}\right)$
.2295

32/36
.2222

λ	412.7173331	nm
$\tilde{\nu}$	2,422,965.84	m-1
f	7.26386885 x 10^{14}	Hz
E	3.00561948	eV
Ω	[illegible]	[illegible]

$R\left(\frac{1}{4}-\frac{1}{36}\right)$
.2222

21/25
.2100

λ	436.7379187	nm
$\tilde{\nu}$	2,289,702.719	m-1
f	6.864356063 x 10^{14}	Hz
E	2.840310409	eV
Ω	[illegible]	[illegible]

$R\left(\frac{1}{4}-\frac{1}{25}\right)$
.21

12/16
.1875

λ	489.1464689	nm
$\tilde{\nu}$	2,044,377.428	m-1
f	6.128889342 x 10^{14}	Hz
E	2.535991437	eV
Ω	[illegible]	[illegible]

$R\left(\frac{1}{4}-\frac{1}{16}\right)$
.1875

5/9
.138

λ	660.347733	nm
$\tilde{\nu}$	1,514,353.65	m-1
f	4.539918031 x 10^{14}	Hz
E	1.878512175	eV
Ω	[illegible]	[illegible]

$R\left(\frac{1}{4}-\frac{1}{9}\right)$
.138

$\Delta M v^2$

Paschen spectral transitions

$$\Delta h\nu \quad = \quad KEM \quad = \quad hcR$$

Planck 1.502 eV Rydberg

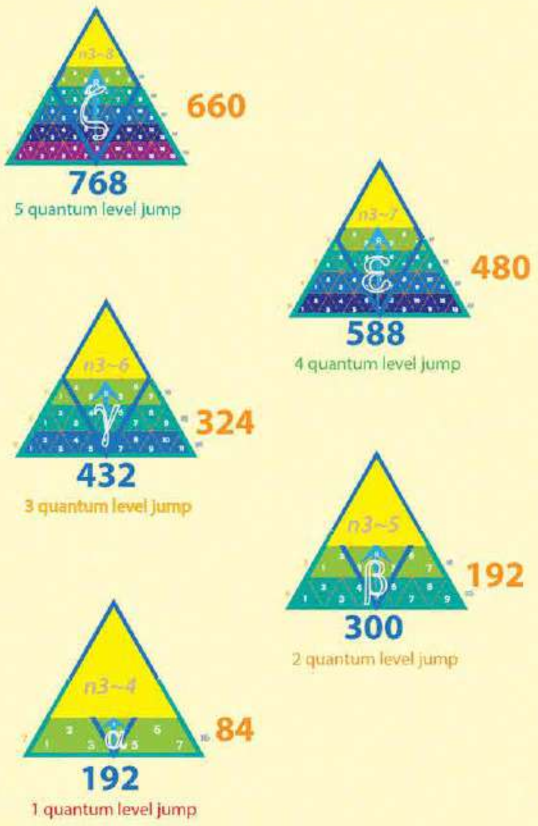

660
768
5 quantum level jump

480
588
4 quantum level jump

324
432
3 quantum level jump

192
300
2 quantum level jump

84
192
1 quantum level jump

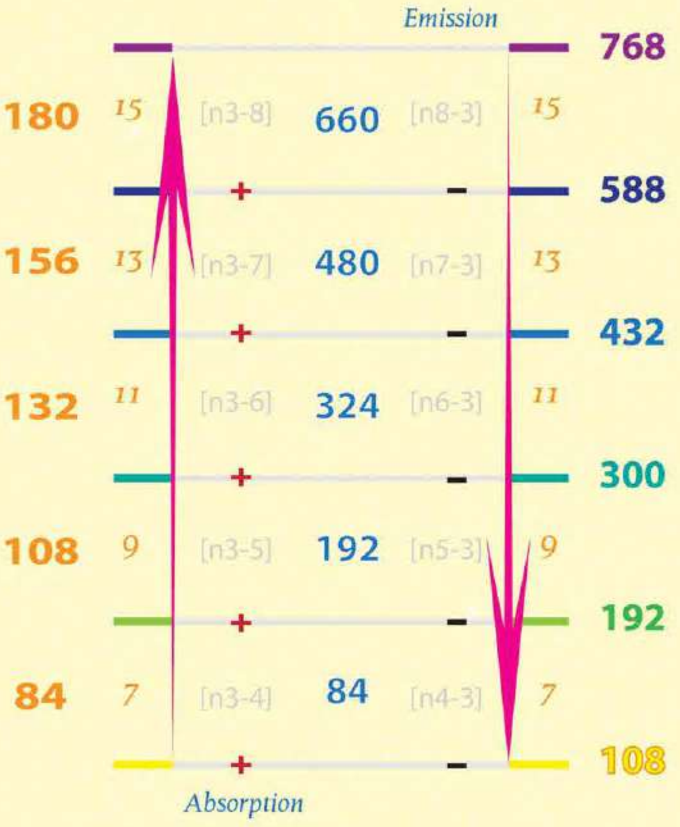

Emission

768

180 *15* [n3-8] 660 [n8-3] *15*

+ − 588

156 *13* [n3-7] 480 [n7-3] *13*

+ − 432

132 *11* [n3-6] 324 [n6-3] *11*

+ − 300

108 *9* [n3-5] 192 [n5-3] *9*

+ − 192

84 *7* [n3-4] 84 [n4-3] *7*

+ − 108

Absorption

55/64
.0954

λ	960.5057935	nm
ṽ	1,041,118.135	m-1
f	$3.121193646 \times 10^{14}$	Hz
E	**1.29147712**	**eV**
Ω	287.9523928	m^2/s

$R\left(\frac{1}{9}-\frac{1}{64}\right)$
.09548

40/49
.0907

λ	1,011.157466	nm
ṽ	988,965.6492	m-1
f	$2.9464844428 \times 10^{14}$	Hz
E	**1.226783461**	**eV**
Ω	303.1373822	m^2/s

$R\left(\frac{1}{9}-\frac{1}{49}\right)$
.0907

27/36
.8033

λ	1,100.579555	nm
ṽ	908,612.1902	m-1
f	$2.723950819 \times 10^{14}$	Hz
E	**1.127107305**	**eV**
Ω	329.94545	m^2/s

$R\left(\frac{1}{9}-\frac{1}{36}\right)$
.0833

16/25
.0711

λ	1,289.741666	nm
ṽ	775,349.069	m-1
f	$2.324438032 \times 10^{14}$	Hz
E	**0.961798233**	**eV**
Ω	360.6548243	m^2/s

$R\left(\frac{1}{9}-\frac{1}{25}\right)$
.0711

7/16
.0486

λ	1,886.707809	nm
ṽ	530,023.7776	m-1
f	$1.588971311 \times 10^{14}$	Hz
E	**0.657479261**	**eV**
Ω	565.6207715	m^2/s

$R\left(\frac{1}{9}-\frac{1}{16}\right)$
.0486

ΔMv^2

Copyright ABRAHAM [2008] - All rights reserved

Brackett spectral transitions

$$\Delta h v = \text{KEM} = hcR$$

Planck — 0.8451 eV — Rydberg

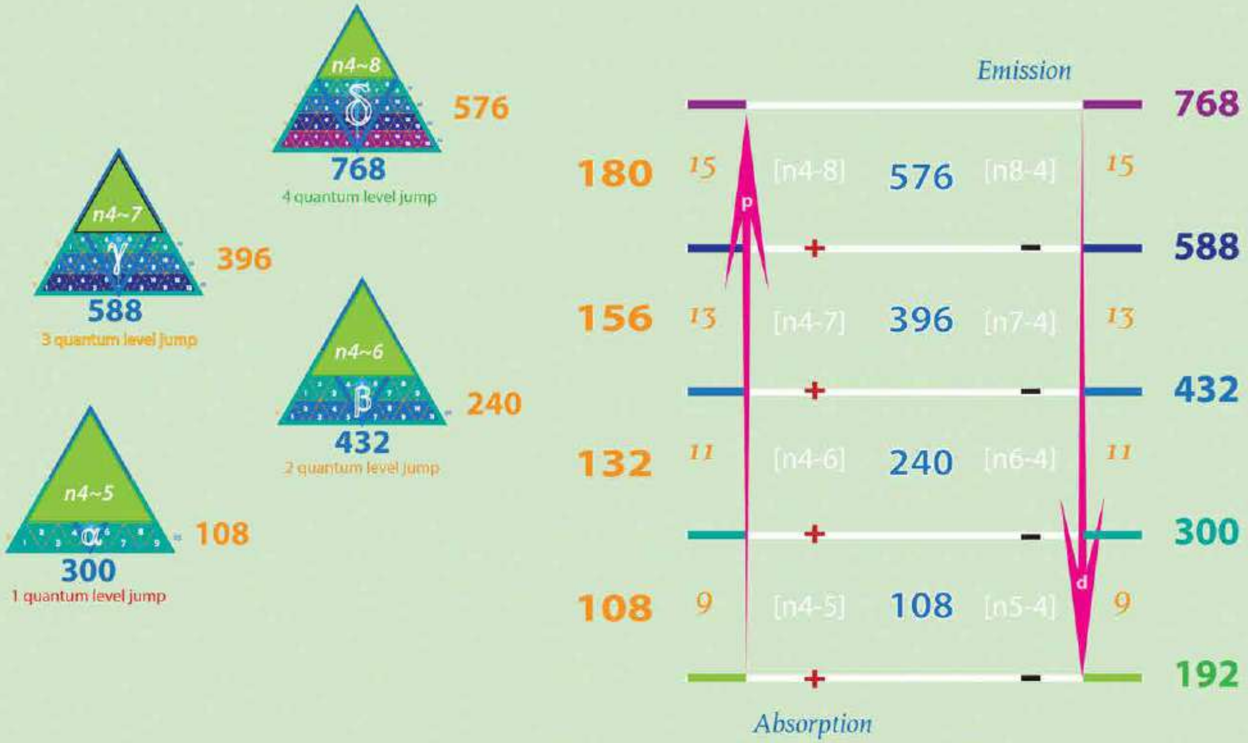
n4~8
δ
576
768
4 quantum level jump
n4~7
γ
396
588
3 quantum level jump
n4~6
β
240
432
2 quantum level jump
n4~5
α
108
300
1 quantum level jump
Emission
768
180
15
[n4-8]
576
[n8-4]
15
+
−
588
156
13
[n4-7]
396
[n7-4]
13
+
−
432
132
11
[n4-6]
240
[n6-4]
11
+
−
300
108
9
[n4-5]
108
[n5-4]
9
p
d
+
−
192
Absorption

48/64
.0468

λ	1,956.585876	nm
$\tilde{v}$	511,094.357	m-1
f	1.532222335 x 10^{14}	Hz
E	**0.634001085**	**eV**
Ω	586.569689	m^2/s

$R\left(\frac{1}{16}-\frac{1}{64}\right)$ .0468

33/49
.04209

λ	2,178.92518	nm
$\tilde{v}$	458,941.8716	m-1
f	1.375873118 x 10^{14}	Hz
E	**0.569307096**	**eV**
Ω	653.2233354	m^2/s

$R\left(\frac{1}{16}-\frac{1}{49}\right)$ .04209

20/36
.0347

λ	2,641.390932	nm
$\tilde{v}$	378,588.4126	m-1
f	1.134979508 x 10^{14}	Hz
E	**0.469630433**	**eV**
Ω	791.8690801	m^2/s

$R\left(\frac{1}{16}-\frac{1}{36}\right)$ .0347

9/25
.0225

λ	4,076.220574	nm
$\tilde{v}$	245,325.2914	m-1
f	7.35466721x 10^{13}	Hz
E	**0.30432052**	**eV**
Ω	1,222.020185	m^2/s

$R\left(\frac{1}{16}-\frac{1}{25}\right)$ .0225

$\Delta M v^2$

Copyright ABRAHAM [2008] - All rights reserved

Pfund spectral transitions

$$\Delta h\nu = \text{KEM} = hcR$$

Planck P

0.541 eV

Rydberg

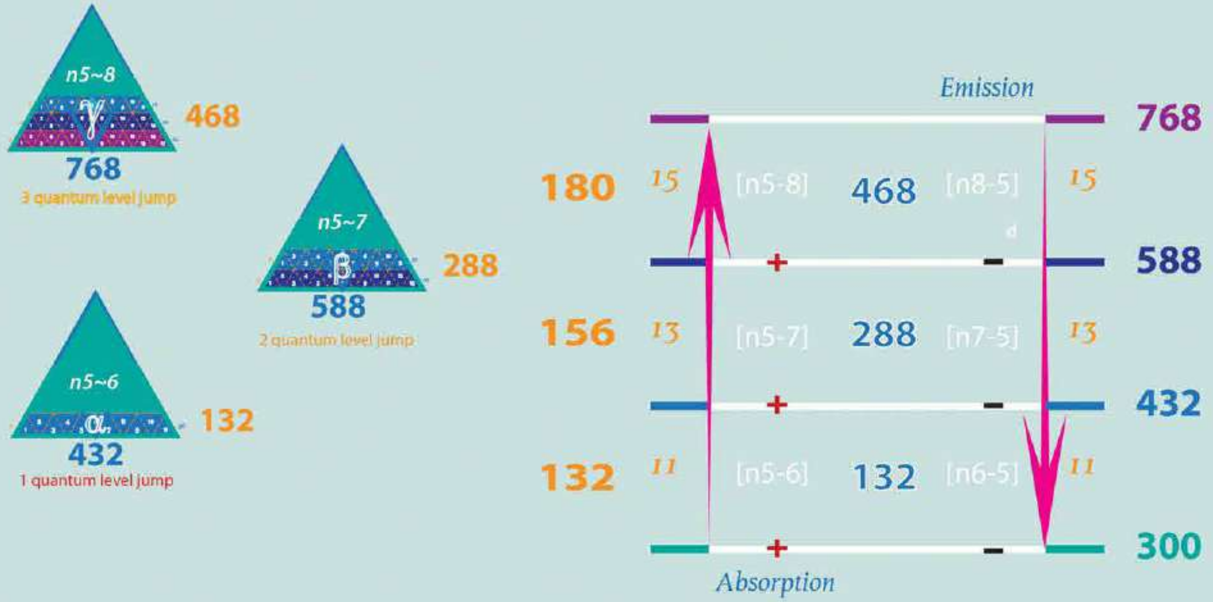

n5~8
γ
468
768
3 quantum level jump
n5~7
β
288
588
2 quantum level jump
n5~6
α
132
432
1 quantum level jump
Emission
768
180
15
[n5-8]
468
[n8-5]
15
588
156
13
[n5-7]
288
[n7-5]
13
432
132
11
[n5-6]
132
[n6-5]
11
300
Absorption
+
-

39/64
.0243

λ	3,762.665146	nm
$\tilde{\nu}$	265,769.0656	m-1
f	7.967556145 x 10^{13}	Hz
E	0.329678886	eV
Ω	1,128.018633	m^2/s

$$R\left(\frac{1}{25}-\frac{1}{64}\right)$$

.0243

24/49
.0195

λ	4,681.284566	nm
$\tilde{\nu}$	213,616.5802	m-1
f	6.404063965 x 10^{13}	Hz
E	0.264985227	eV
Ω	1,403.413807	m^2/s

$$R\left(\frac{1}{25}-\frac{1}{49}\right)$$

.01959

11/36
.0122

λ	7,503.951512	nm
$\tilde{\nu}$	133,263.1212	m-1
f	3.995127867 x 10^{13}	Hz
E	0.165309071	eV
Ω	2,249.628068	m^2/s

$$R\left(\frac{1}{25}-\frac{1}{36}\right)$$

.0122

$$\Delta M v^2$$

Humphreys spectral transitions

$\Delta h\nu$ = KEM = hcR

Planck — 0.375 eV — Rydberg

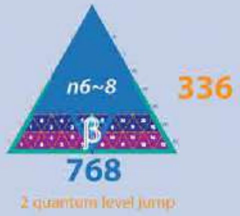
n6~8
β
336
768

2 quantum level jump

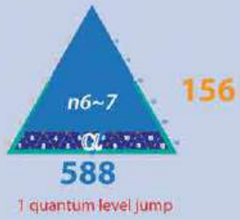
n6~7
α
156
588

1 quantum level jump

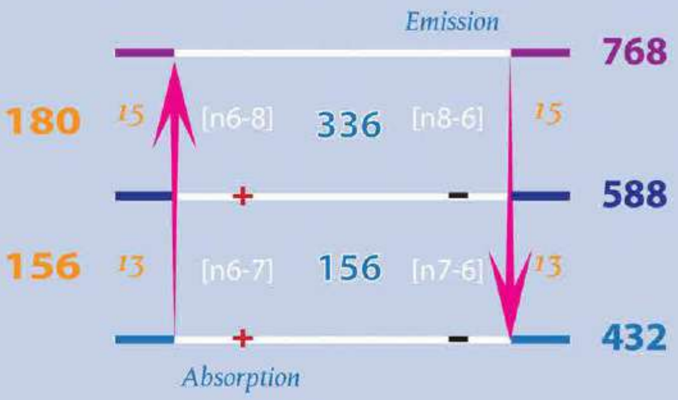
Emission
768
180
15
[n6-8]
336
[n8-6]
15
+
-
588
156
13
[n6-7]
156
[n7-6]
13
+
-
432
Absorption

28/64

.01215

λ	7,546.831235	nm
ν̃	132,505.9444	m-1
ƒ	$3.972428277 \times 10^{13}$	Hz
E	0.164369815	eV
Ω	1,128.018833	m²/s

$R\left(\frac{1}{36}-\frac{1}{64}\right)$

.01215

13/49

.00733

λ	12,445.01497	nm
ν̃	80,353.459	m-1
ƒ	$2.4089936098 \times 10^{13}$	Hz
E	0.099676156	eV
Ω	1,403.413807	m²/s

$R\left(\frac{1}{36}-\frac{1}{49}\right)$

.0073

$\Delta M v^2$

Un-named spectral transition

$$\Delta h\nu \;=\; \text{KEM} \;=\; hcR$$

Planck — 0.276 eV — Rydberg

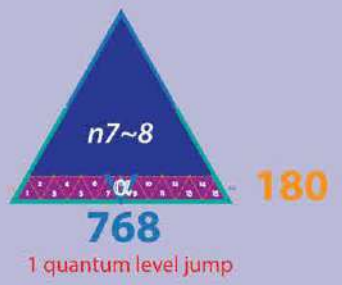

n7~8
α
180
768
1 quantum level jump

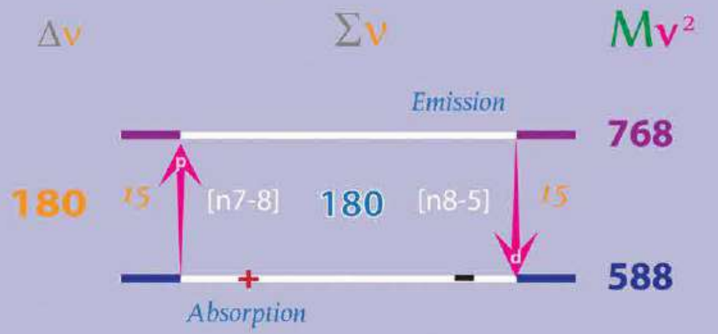

Δν
Σν
Mν²
Emission
768
180
15
[n7-8]
180
[n8-5]
15
p
d
+
−
588
Absorption

15/64

.0047

λ	19,174.54158	nm
$\tilde{\nu}$	52,152.48541	m-1
f	1.563492179 x 10^{13}	Hz
E	0.064693658	eV
Ω	5.748382952	m²/s

$$R\left(\frac{1}{49} - \frac{1}{64}\right)$$

.0047

$$\Delta M\nu^2$$

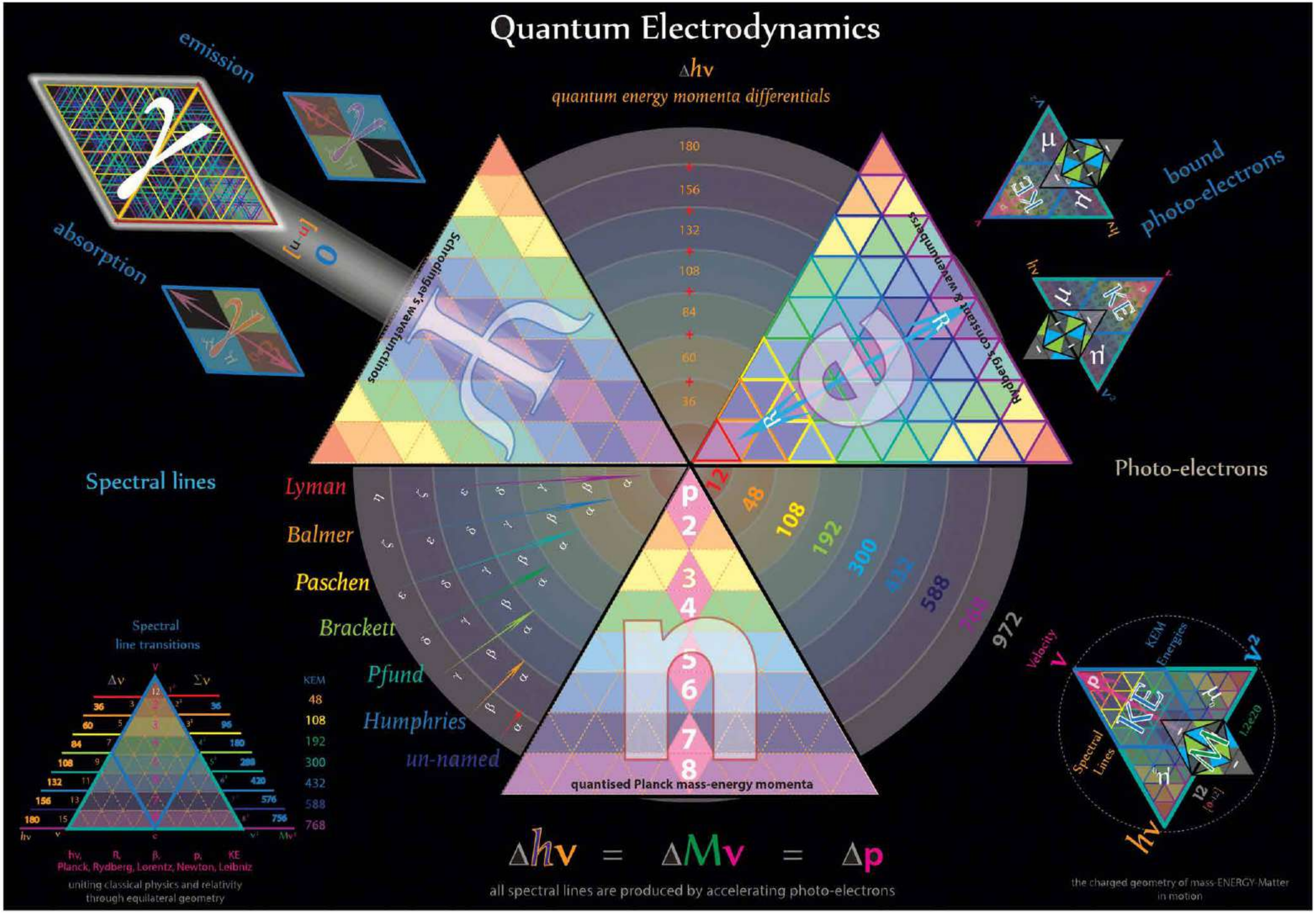
Quantum Electrodynamics
Δhv
quantum energy momenta differentials
emission
absorption
[n-n]
0
Schrödinger's wavefunctions
Rydberg's constant & wavenumbers
R
180
156
132
108
84
60
36
bound photo-electrons
μ
KE
ɳ
hv
v²
Photo-electrons
Spectral lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
α
β
γ
δ
ε
ζ
η
p
2
3
4
5
6
7
8
n
12
48
108
192
300
432
588
768
972
quantised Planck mass-energy momenta
Spectral line transitions
Δv
Σv
v
36
60
84
108
132
156
180
3
5
7
9
11
13
15
1²
2²
3²
4²
5²
6²
7²
8²
36
96
180
288
420
576
756
hv
v
c
v²
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz
uniting classical physics and relativity through equilateral geometry
KEM
48
108
192
300
432
588
768
Velocity
v
KEM Energies
v2
p
KE
μ
M
ɳ
1.2e20
Spectral Lines
12
[ϕ-12]
hv
$\Delta hv = \Delta Mv = \Delta p$
all spectral lines are produced by accelerating photo-electrons
the charged geometry of mass-ENERGY-Matter in motion

Tetryonics 96.11 - Atomic spectral series transitions

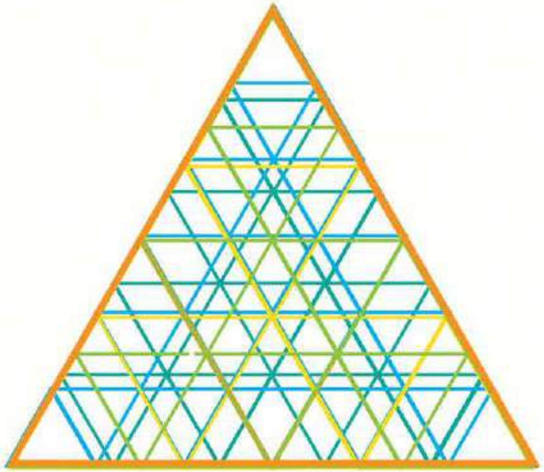

Fractionals and fractals

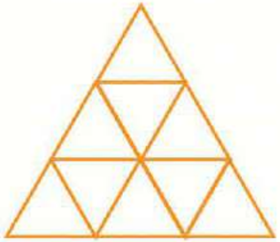

ninths

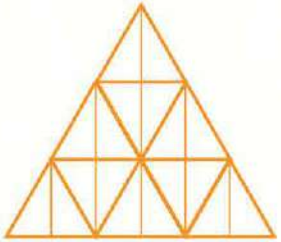

eigteenths

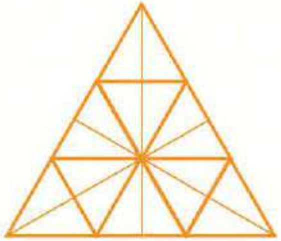

Eighteenths

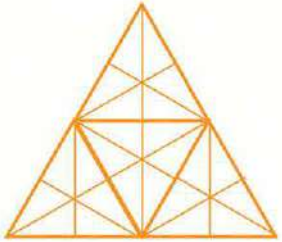

Twenty-fourths

White light
[EM radiation]
is comprised of many
superpositioned
frequencies

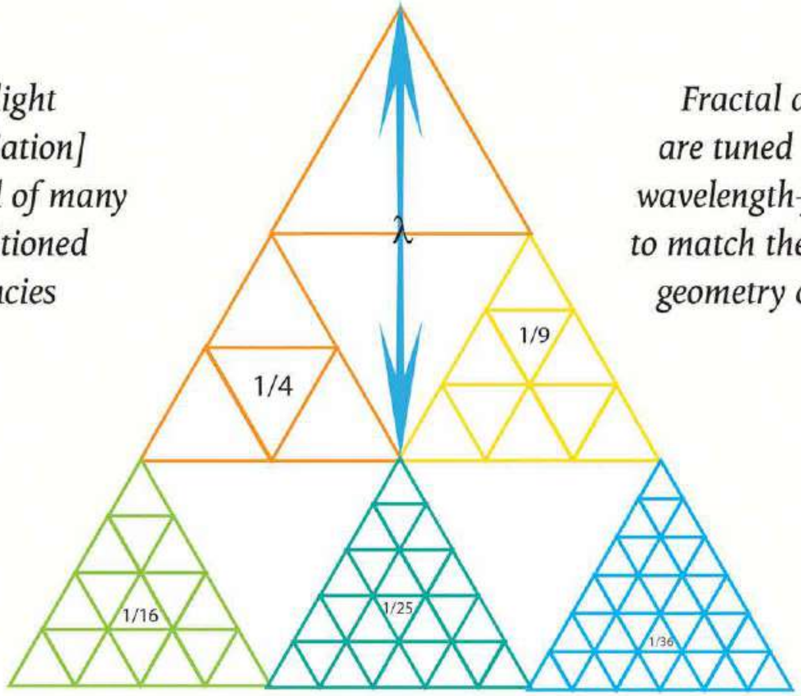

Fractal antennas
are tuned to specific
wavelength-frequencies
to match the equilateral
geometry of photons

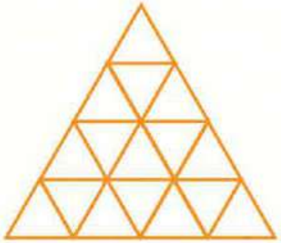

Sizteenths

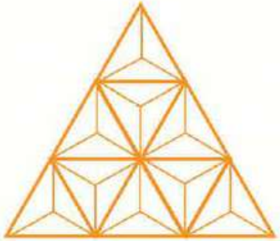

Twenty-sevenths

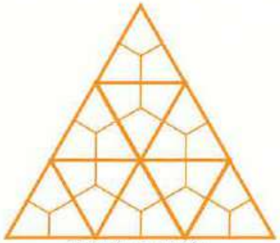

Twenty-sevenths

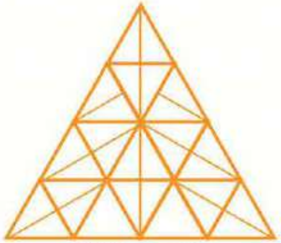

Thirty-seconds

Polygons and Angles

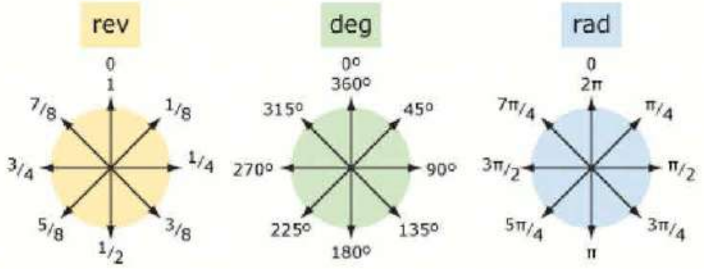
rev
0
1
1/8
1/4
3/8
1/2
5/8
3/4
7/8
deg
0°
360°
45°
90°
135°
180°
225°
270°
315°
rad
0
2π
π/4
π/2
3π/4
π
5π/4
3π/2
7π/4

polygon exterior and interior angles

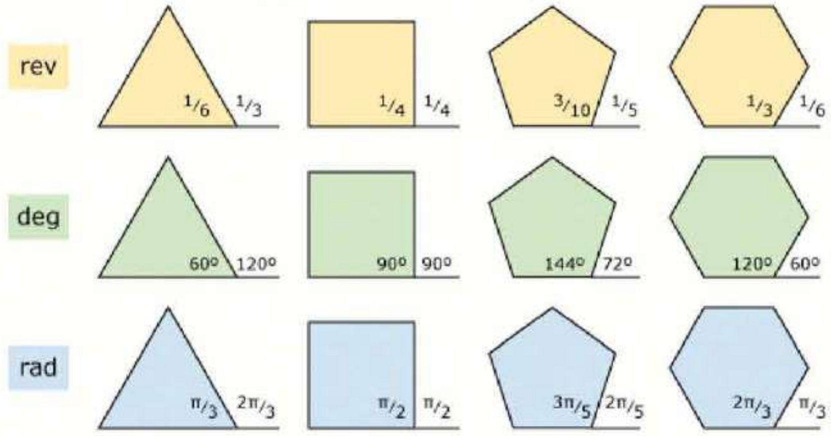
rev
1/6
1/3
1/4
1/4
3/10
1/5
1/3
1/6
deg
60°
120°
90°
90°
144°
72°
120°
60°
rad
π/3
2π/3
π/2
π/2
3π/5
2π/5
2π/3
π/3

dividing the circle

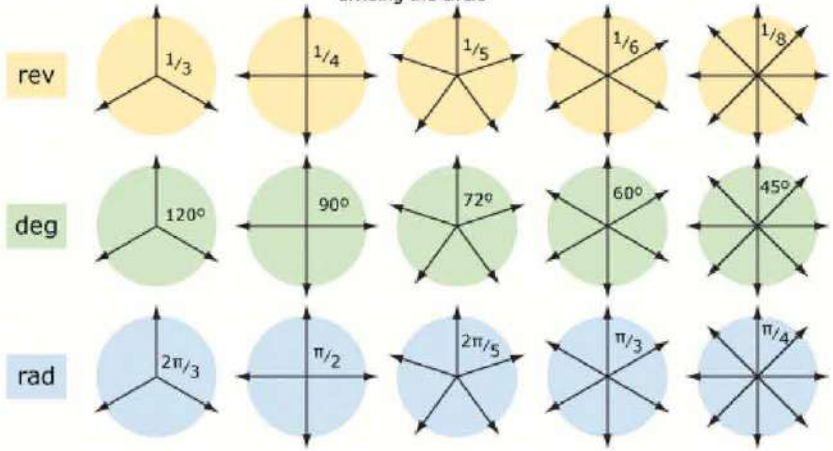
rev
1/3
1/4
1/5
1/6
1/8
deg
120°
90°
72°
60°
45°
rad
2π/3
π/2
2π/5
π/3
π/4

supplementary angles

complementary angles

triangles

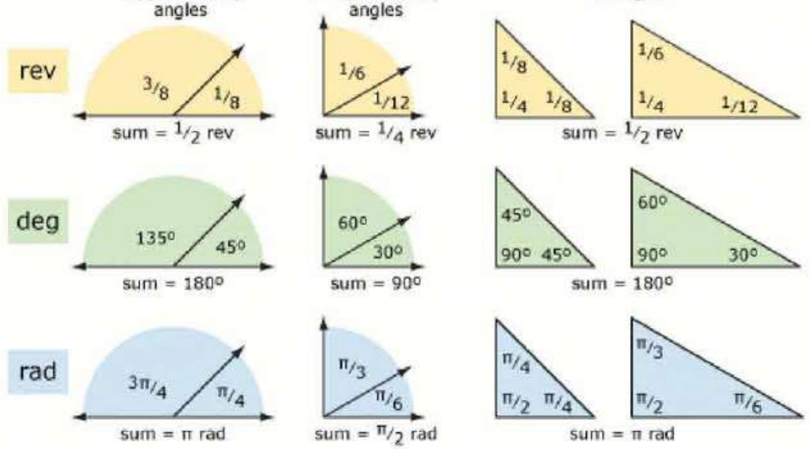
rev
3/8
1/8
sum = 1/2 rev
1/6
1/12
sum = 1/4 rev
1/8
1/4
1/8
1/6
1/4
1/12
sum = 1/2 rev
deg
135°
45°
sum = 180°
60°
30°
sum = 90°
45°
90°
45°
60°
90°
30°
sum = 180°
rad
3π/4
π/4
sum = π rad
π/3
π/6
sum = π/2 rad
π/4
π/2
π/4
π/3
π/2
π/6
sum = π rad

Copyright ABRAHAM [2008] - All rights reserved

Pi vs Tau

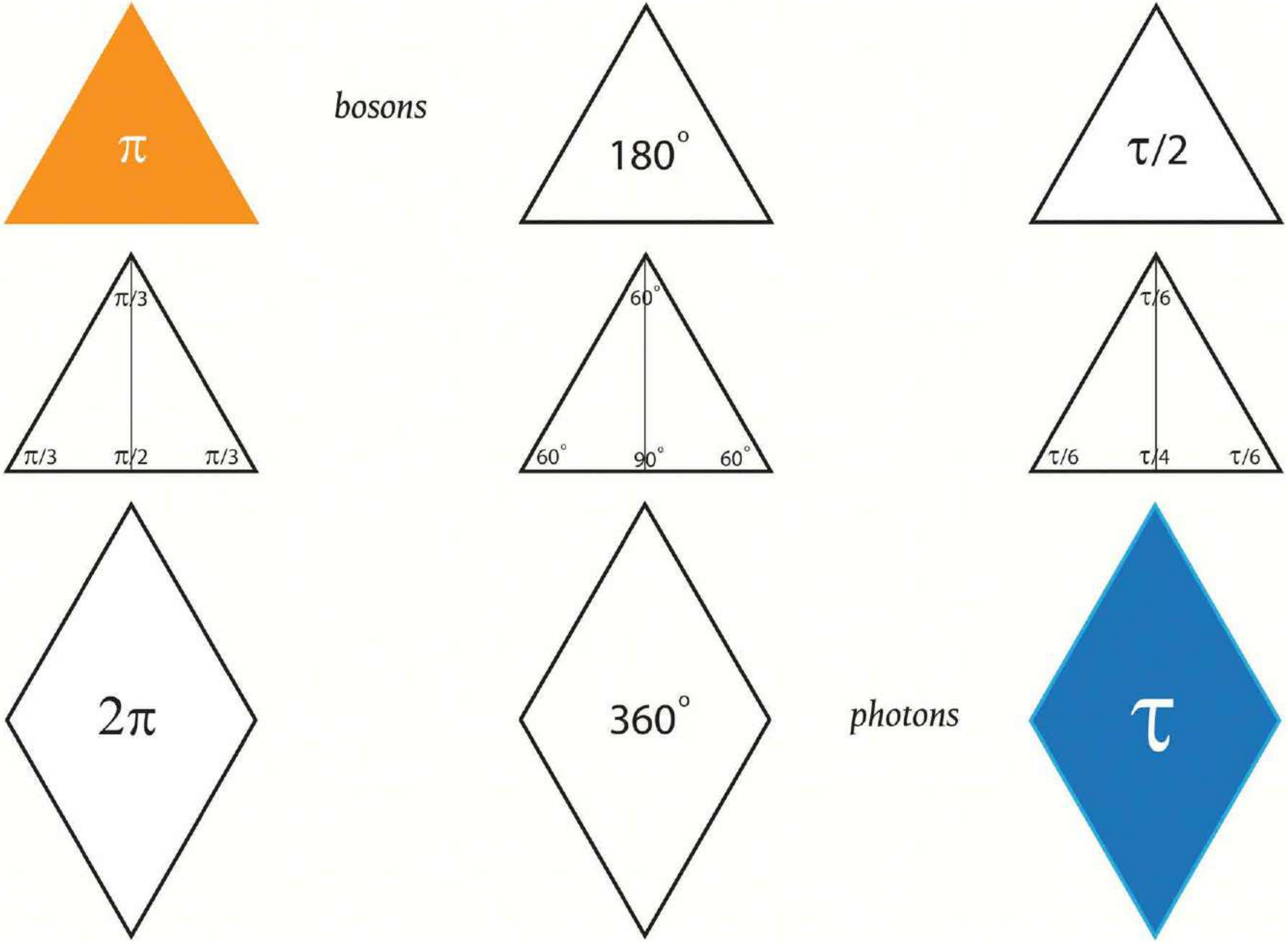
π
bosons
180°
τ/2
π/3
π/3
π/2
π/3
60°
60°
90°
60°
τ/6
τ/6
τ/4
τ/6
2π
360°
photons
τ

Triangular dissection of an equilateral triangle

is a way of dividing up a original triangle into smaller equilateral triangles,
such that none of the smaller triangles overlap

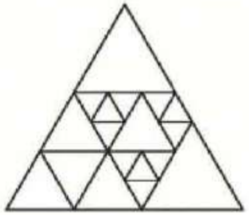

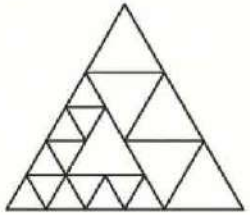

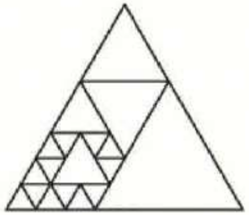

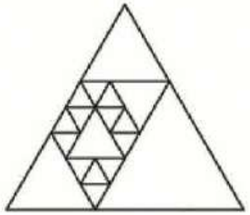

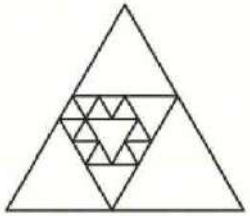

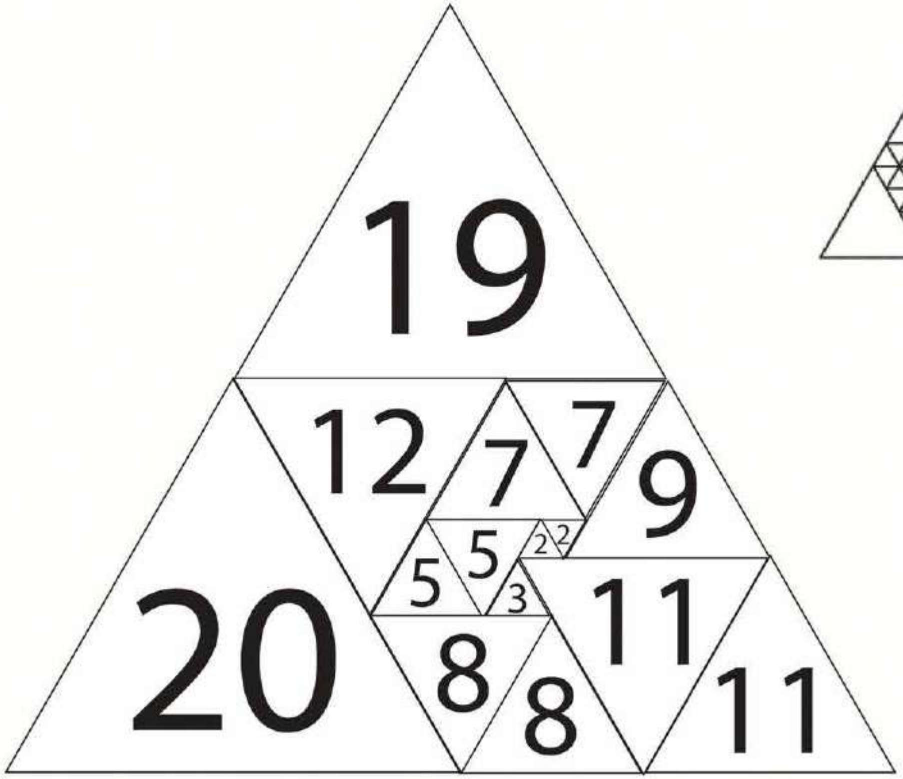

lowest order perfect equilateral triangle dissected by equilateral triangles

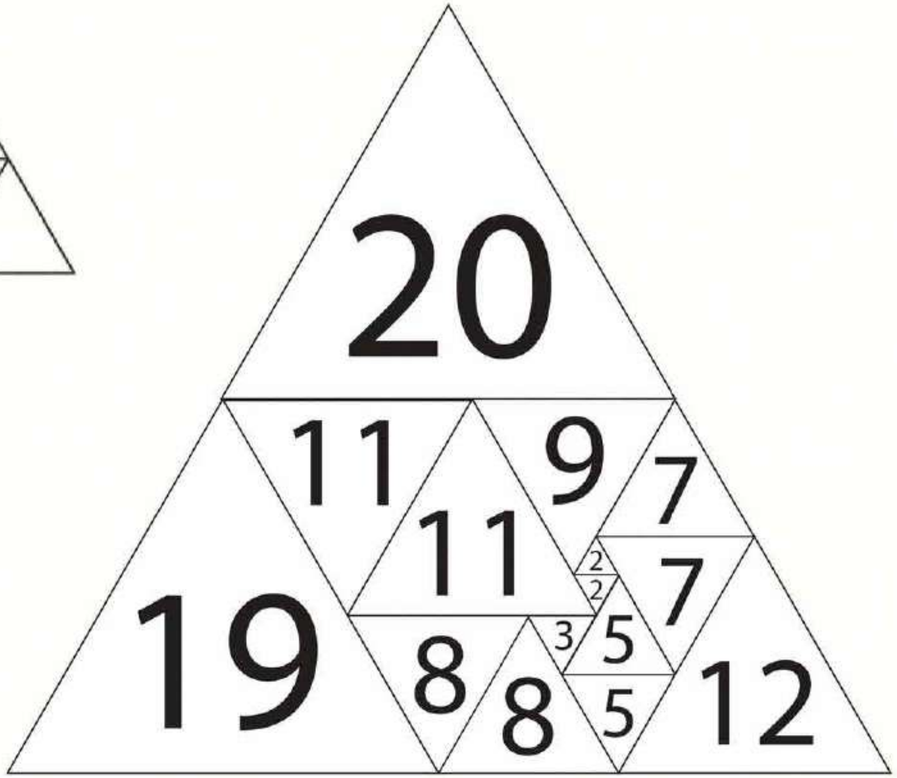

lowest order perfect dissected equilateral triangle, an isomer of the first

Golden mean Spirals

Golden Mean Spiral – This spiral is derived via the golden rectangle, a unique rectangle which has the golden ratio.
This form is found everywhere in nature: the Nautilus Shell, the face of a Sunflower, fingerprints, our DNA, and the shape of the Milky Way

Golden spiral

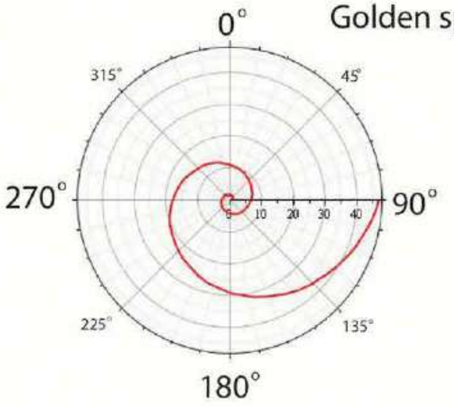

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, ...

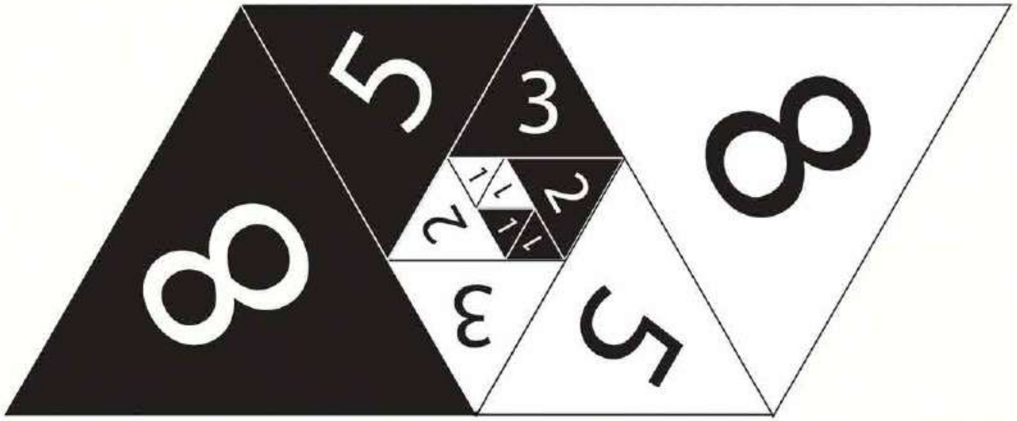

Golden spirals

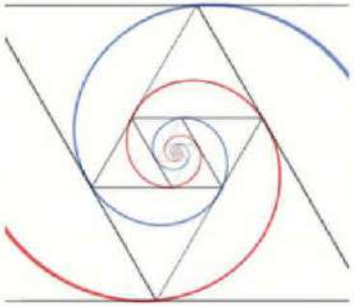

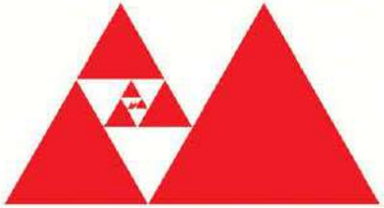

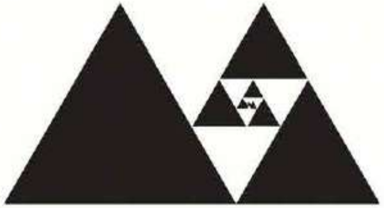

Continued fractions and the Fibonacci sequence

The convergence of the continued fractions

$$1+\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\cdots$$

are:

$$\frac{2}{1},\ \frac{3}{2},\ \frac{5}{3},\ \frac{8}{5},\ \frac{13}{8},\ \frac{21}{13},\ \frac{34}{21},\ \frac{55}{34},\ \frac{89}{55},\ \frac{144}{89},\ \frac{233}{144},\ \frac{377}{233},\ \frac{610}{377},$$

The Golden Ratio (Golden Mean, Golden Section)
is defined mathematically as:

$$\varphi=\frac{1+\sqrt{5}}{2}=1.6180339887\ldots.$$

Copyright ABRAHAM [2008] - All rights reserved

Koch fractal Curve

Niels Fabian Helge von Koch (January 25, 1870 – March 11, 1924) was a Swedish mathematician who gave his name to one of the earliest fractal curves ever known

He described the Koch curve, or Koch snowflakes as it popularly known, in a 1904 paper entitled "On a continuous curve without tangents constructible from elementary geometry"

Von Koch wrote several papers on number theory . One of his results was a 1901 theorem proving that the Riemann hypothesis is equivalent to a strengthened form of the prime number theorem.

Three Koch curves form the snowflake.

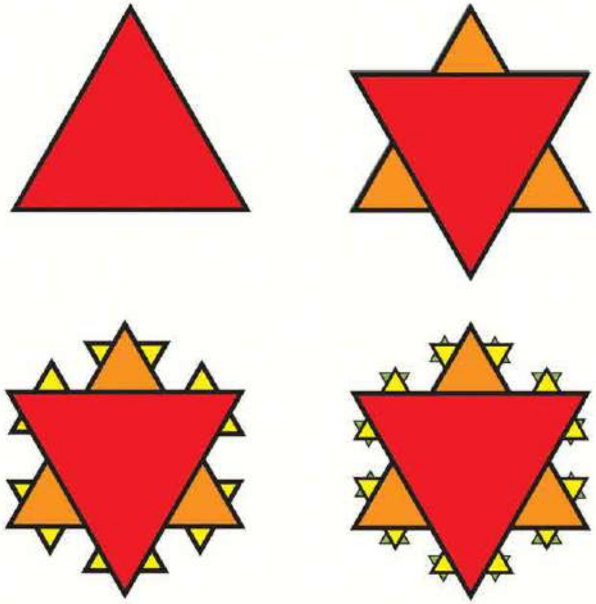

The Koch snowflake (or Koch star) is a mathematical curve and one of the earliest fractal curves to have been described.

Actually Koch described what is now known as the Koch curve, which is the same as the now popular snowflake, except it starts with a line segment instead of an equilateral triangle.

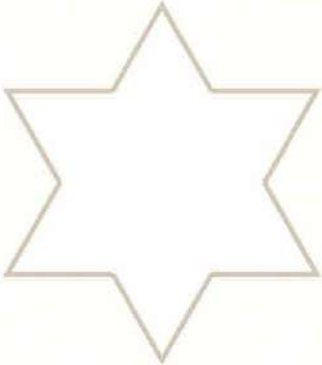

$$a = \frac{1}{2} + \frac{i}{\sqrt{12}}$$

The Koch curve is a special case of the Cesaro curve where:

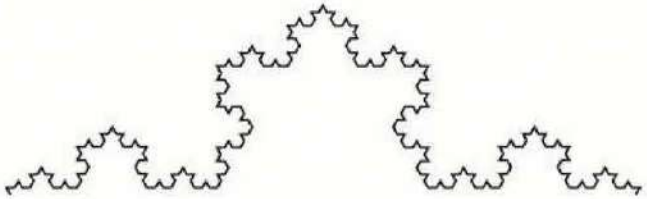

which is in turn a special case of the de Rham curve.

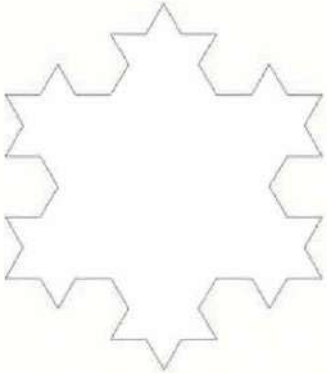

Copyright ABRAHAM [2008] - All rights reserved

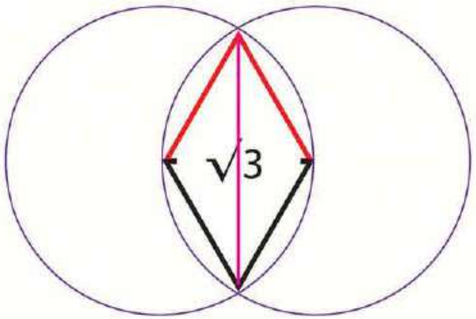

vesica piscis

$$\frac{1351}{780} > \sqrt{3} > \frac{265}{153}\,.$$

Koch snowflake

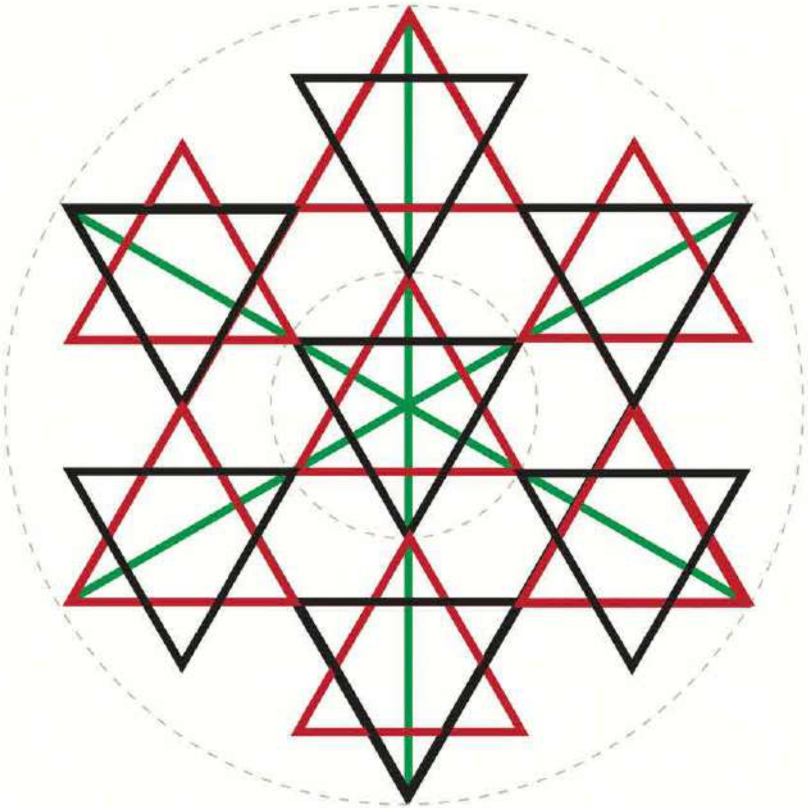

space-time

star tetrahedron

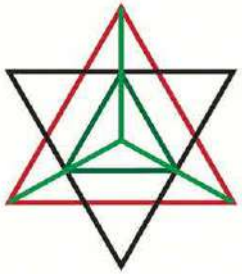

Flower of Life

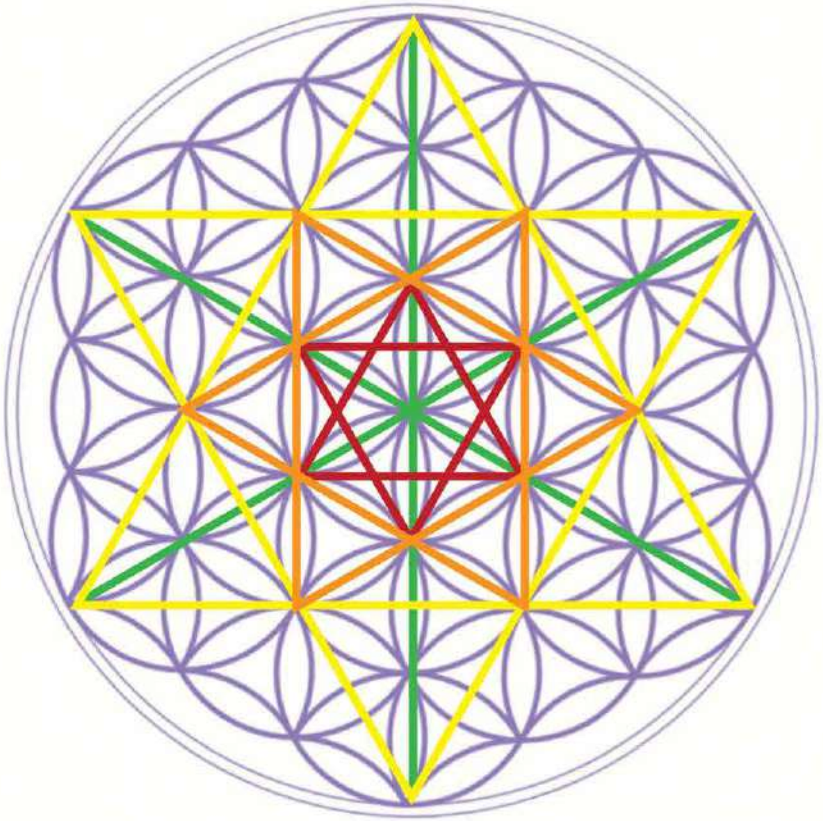

The Flower of Life is a name for a geometrical figure composed of multiple evenly-spaced, overlapping circles.

Unit circles - SINE WAVES - Photons

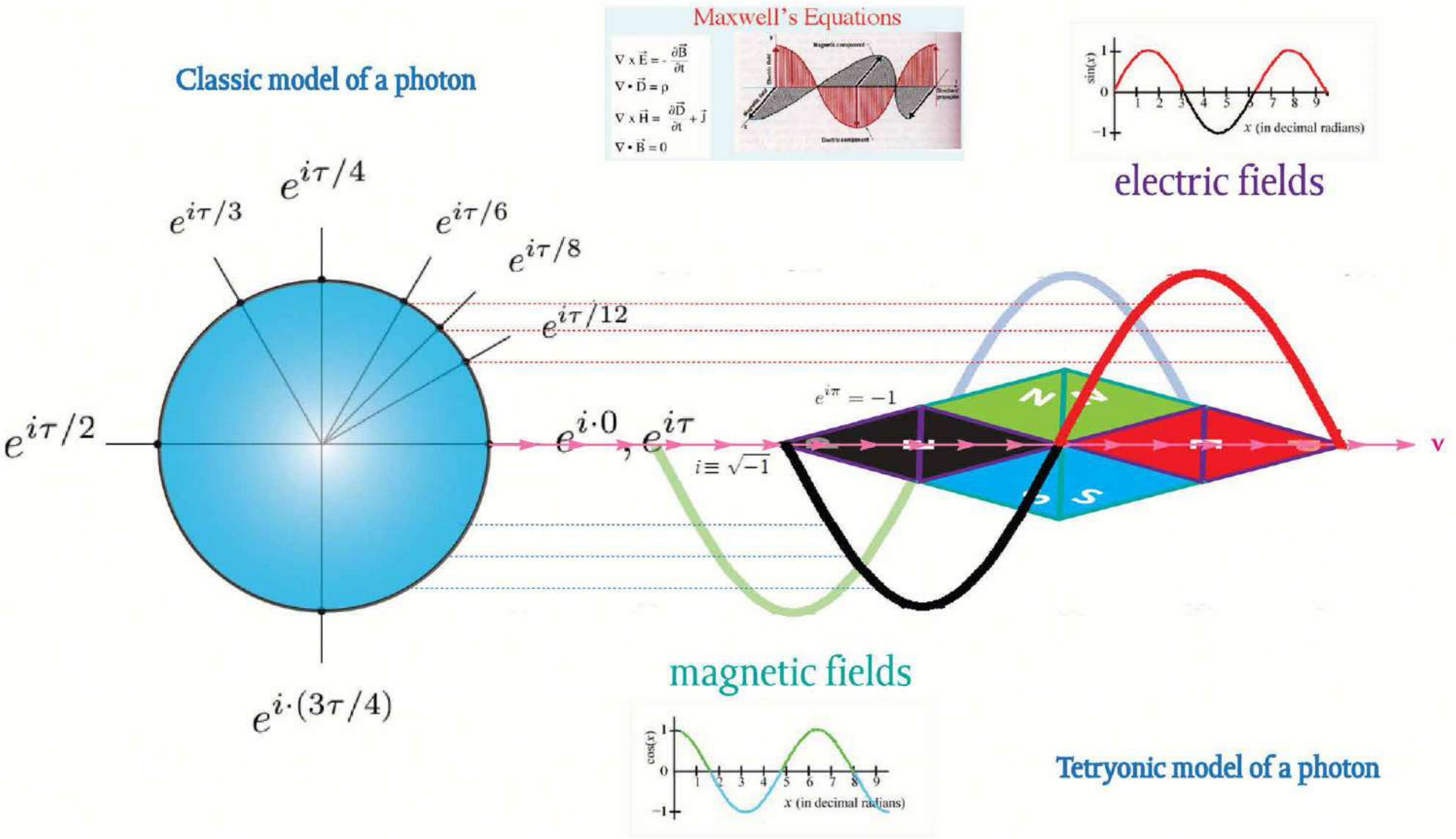
Classic model of a photon
Maxwell's Equations
electric fields
magnetic fields
Tetryonic model of a photon
$e^{i\tau/4}$
$e^{i\tau/3}$
$e^{i\tau/6}$
$e^{i\tau/8}$
$e^{i\tau/12}$
$e^{i\tau/2}$
$e^{i\cdot 0}, e^{i\tau}$
$e^{i\cdot(3\tau/4)}$
$e^{i\pi} = -1$
$i \equiv \sqrt{-1}$
v
N
S
$\sin(x)$
$\cos(x)$
1
0
−1
1 2 3 4 5 6 7 8 9
x (in decimal radians)

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$
$$\nabla \cdot \vec{D} = \rho$$
$$\nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J}$$
$$\nabla \cdot \vec{B} = 0$$

Magnetic waveforms are 90 degrees out of phase with Electric waveforms

Boson distributions in monochromatic EM waves

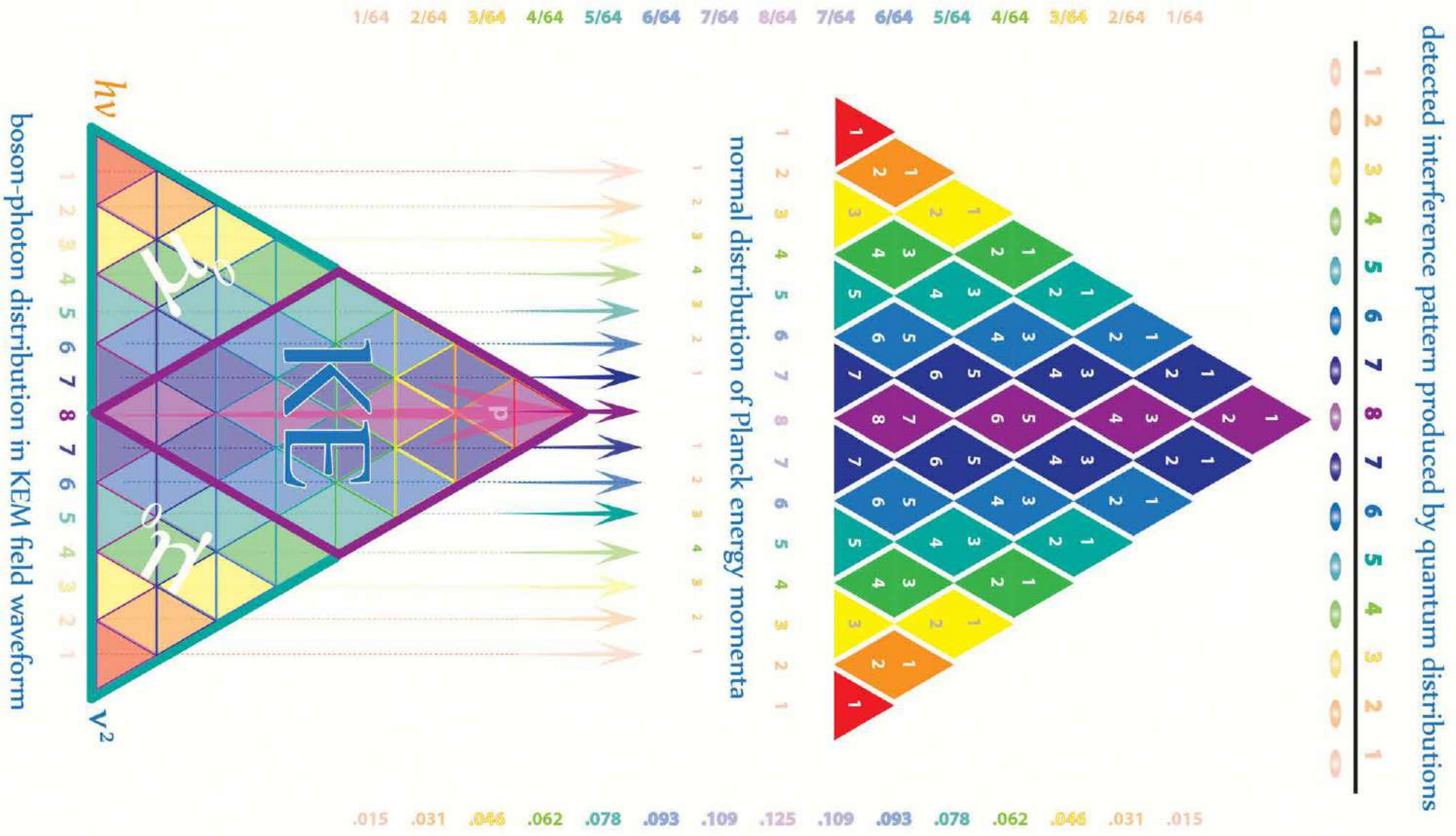
1/64 2/64 3/64 4/64 5/64 6/64 7/64 8/64 7/64 6/64 5/64 4/64 3/64 2/64 1/64
hν
μ₀
ε₀
KE
p
v²
boson-photon distribution in KEM field waveform
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
normal distribution of Planck energy momenta
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
detected interference pattern produced by quantum distributions
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
.015 .031 .046 .062 .078 .093 .109 .125 .109 .093 .078 .062 .046 .031 .015

Probability distributions of monochromatic EM waves

Copyright ABRAHAM [2008] - All rights reserved

Wave interference patterns

The double-slit experiment, sometimes called Young's experiment, is a demonstration that matter and energy can display characteristics of both waves and particles, and demonstrates the fundamentally probabilistic nature of quantum mechanical phenomena and Establishes the quantum interference principle known as wave–particle duality.

Matter particles are stopped by the barrier but the [K]EM wave passes through both slits and is diffracted by them producing weaker EM waves that then superposition with each other to produce interference patterns

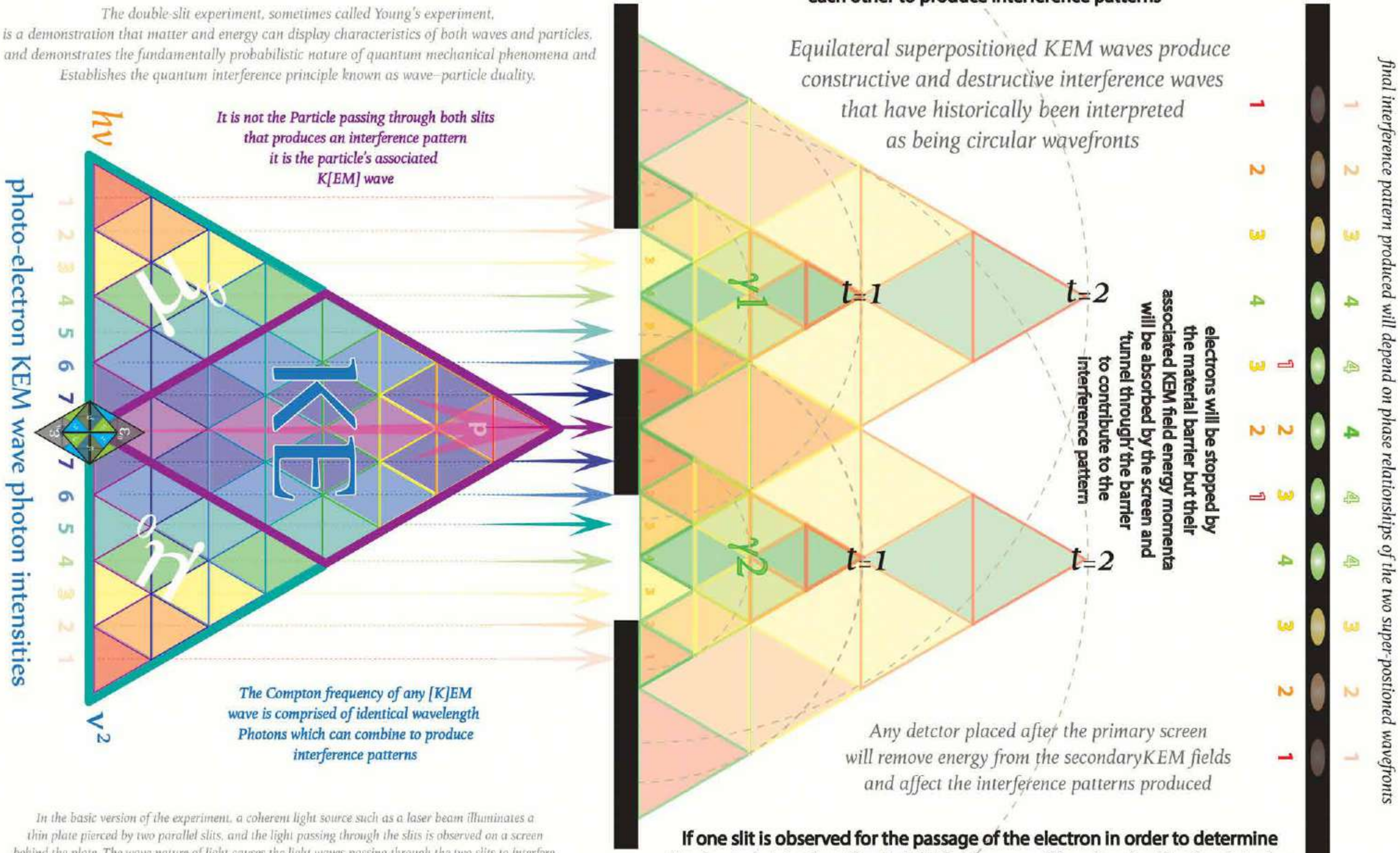

It is not the Particle passing through both slits that produces an interference pattern it is the particle's associated K[EM] wave

Equilateral superpositioned KEM waves produce constructive and destructive interference waves that have historically been interpreted as being circular wavefronts

electrons will be stopped by the material barrier but their associated KEM field energy momenta will be absorbed by the screen and 'tunnel through' the barrier to contribute to the interference pattern

final interference pattern produced will depend on phase relationships of the two super-positioned wavefronts

The Compton frequency of any [K]EM wave is comprised of identical wavelength Photons which can combine to produce interference patterns

Any detctor placed after the primary screen will remove energy from the secondaryKEM fields and affect the interference patterns produced

In the basic version of the experiment, a coherent light source such as a laser beam illuminates a thin plate pierced by two parallel slits, and the light passing through the slits is observed on a screen behind the plate. The wave nature of light causes the light waves passing through the two slits to interfere, producing bright and dark bands on the screen — a result that would not be expected if light consisted strictly of particles. However, on the screen, the light is always found to be absorbed as though it were composed of discrete particles or photons.

If one slit is observed for the passage of the electron in order to determine its physical state then the its KEM field wave will be absorbed by the detector resulting in only one wave remaining, enforcing a classical particle outcome

Copyright ABRAHAM [2008] - All rights reserved

Quantum computing via EM wave super-positioning

BY superpositioning two beams
of EM radiation the resultant 'colours'
will perform quantum level computations
that can be read via the resultant
interference patterns produced

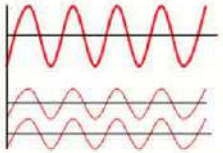

In phase

constructive interference

additive in-phase EM waves

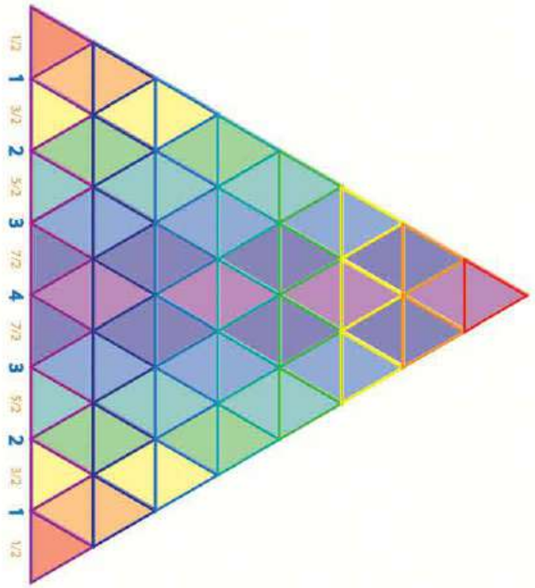
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Additive

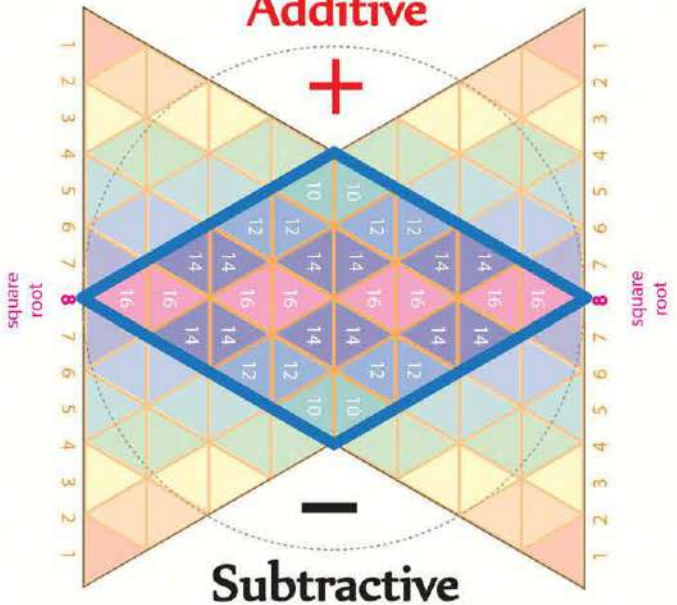
+
−
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
square root
10
12
14
16
14
12
10

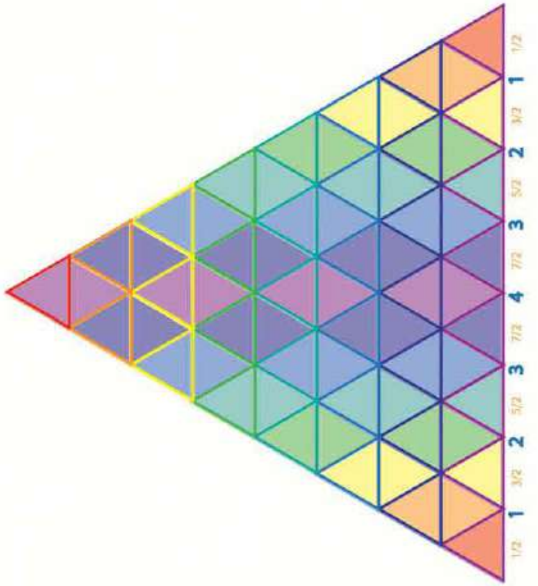
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Subtractive

Out of phase

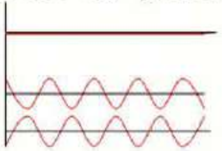

destructive interference

subtractive out-of-phase EM waves

Various basic operations, such as
ADDITION, SUBTRACTION and
SQUARE ROOTS etc
are all easily computed using
EM wave super-positioning

The lines of Force

By utilising the statiscal distribution of equilateral Planck energy momenta quanta in EM waves
Tetryonic theory provides a practical geometric solution to quantum computing problems

q

bosons

P vs. NP

*The set of all decision problems for which an algorithm exists which can be carried out by a deterministic Turing machine in **polynomial** time*

*the set of all decision problems for which an algorithm exists which can be carried out by a **non-deterministic** Turing machine in **polynomial** time*

E

photons

An algorithm of time complexity ***O****(n)* is one which increases in time linearly as the "size of the problem" (whatever n stands for) increases.

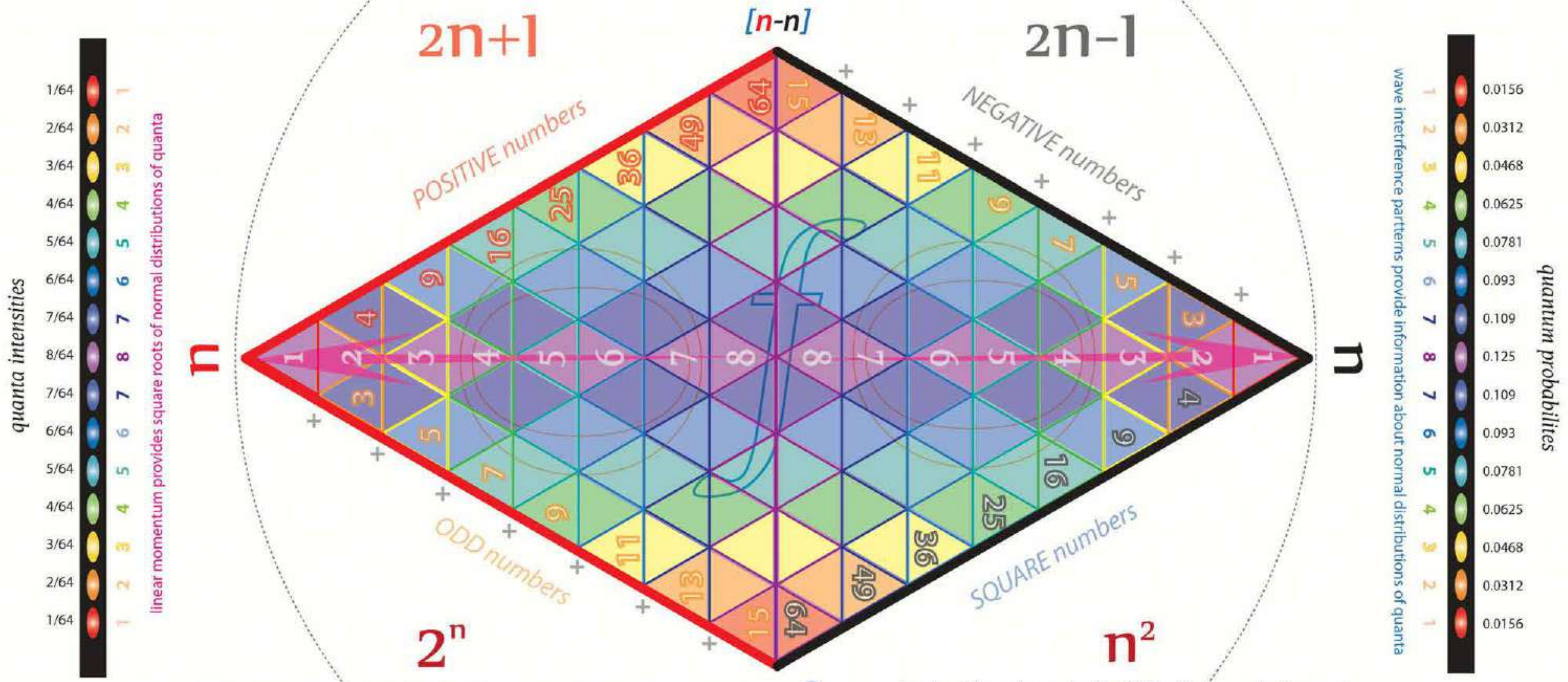
2n+1
[n-n]
2n-1
POSITIVE numbers
NEGATIVE numbers
ODD numbers
SQUARE numbers
n
n
f
quanta intensities
1/64 2/64 3/64 4/64 5/64 6/64 7/64 8/64 7/64 6/64 5/64 4/64 3/64 2/64 1/64
linear momentum provides square roots of normal distributions of quanta
wave interference parterns provide information about normal distributions of quanta
quantum probabilites
0.0156 0.0312 0.0468 0.0625 0.0781 0.093 0.109 0.125 0.109 0.093 0.0781 0.0625 0.0468 0.0312 0.0156
2ⁿ
n²

An algorithm of complexity *O(2ⁿ)* utlises exponential quanta; increasing n by 1 will double the quanta required

An algorithm of complexity *O(n²)* utilises quadratic quanta, meaning that if you double n it will use four times as many quanta.

$2n^2$

n^o

Unlike Math treatise on P vs Np that require exponential polynomial time ***O****(nᵏ)* Tetryonic geometry of EM fields utilise exponential energies per second

2hv

1 second

hf

By relating p to the number of equilateral wave fields in any EM wave very large [pn²] data sets can be modelled and processed every second

Copyright ABRAHAM [2008] - All rights reserved

Quantum Cryptography

If the numerical sequences were applied to amplitude modulation
their non-repeating numerical sequences would appear
to be purely random noise

**Using Tetryonic geometry
advanced non-repeating cyphers
of any complexity
can be easily developed**

The level of encryption can
be easily increased (without limit)
by increasing the dimensions
of the cypher geometry

eg: number of quanta
2D vs 3D geometries

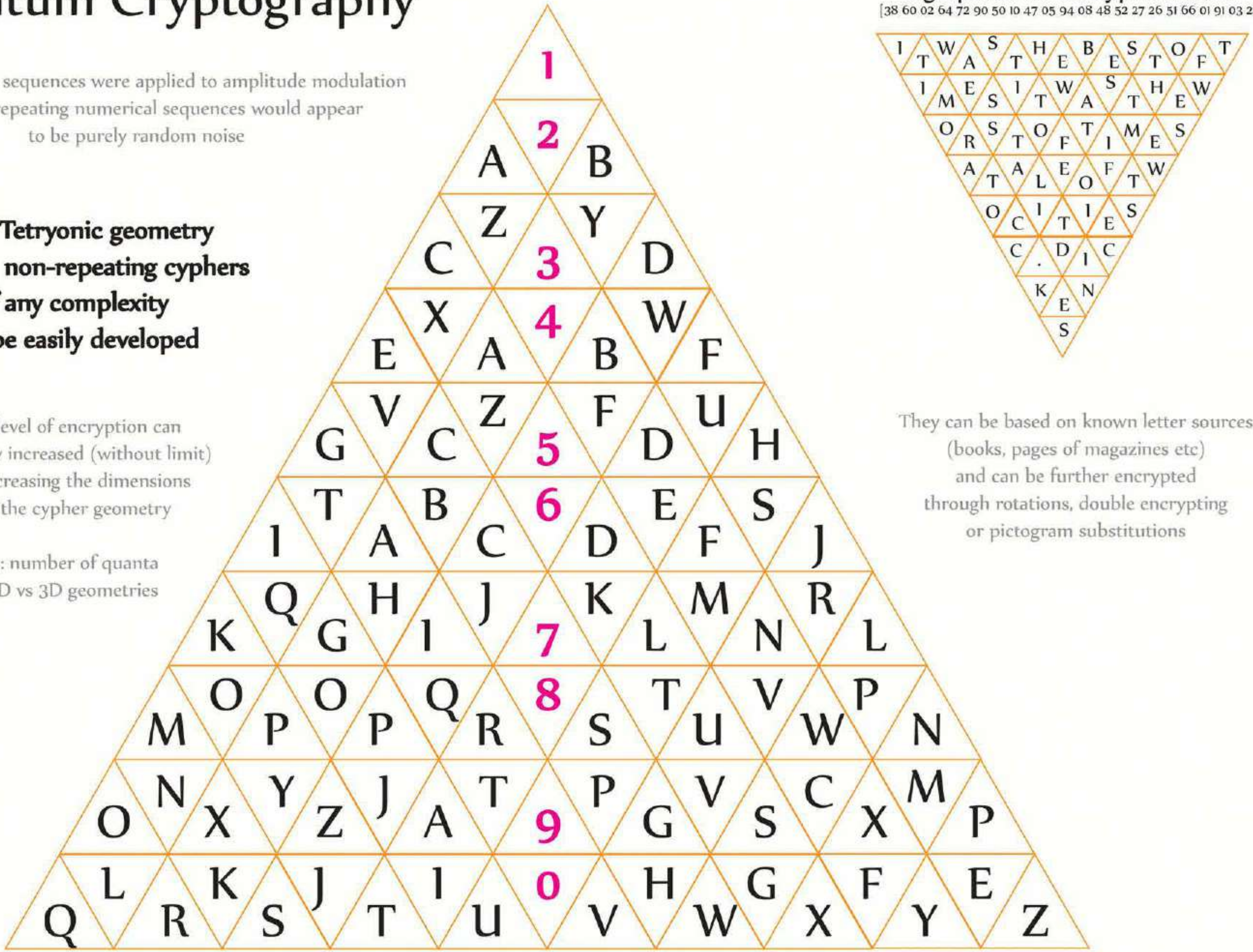
1
A 2 B
C Z 3 Y D
E X A 4 B W F
G V C Z 5 F D U H
I T A B C 6 D E F S J
K Q G H I J 7 K L M N R L
M O P O P Q R 8 S T U V W P N
O N X Y Z J A T 9 P G V S C X M P
Q L R K S J T I U 0 V H W G X F Y E Z

eg: quantum encryption 1024
[38 60 02 64 72 90 50 10 47 05 94 08 48 52 27 26 51 66 01 91 03 20]
I T W A S T H E B E S T O F T
I M E S I T W A S T H E W
O R S T O F T I M E S
A T A L E O F T W
O C I T I E S
C . D I C
K E N
S

They can be based on known letter sources
(books, pages of magazines etc)
and can be further encrypted
through rotations, double encrypting
or pictogram substitutions

Copyright ABRAHAM [2008] - All rights reserved

Quantum Computing

The Proton/Neutron geometries of atomic nuclei can be built at the quantum scale
to create an atomic nuclei that can operate as a Opto-memory-transistive computing element,
many elements can then be combined in lattices to create super computers no larger than bacterium

Spin UP

*Energy can be gated through individual nuclei
using the centre Baryon as the base transistor element,
in turn effecting the energies of bound photo-electrons*

Spin DOWN

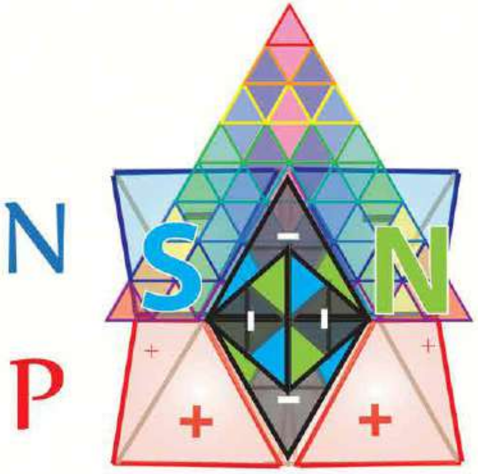

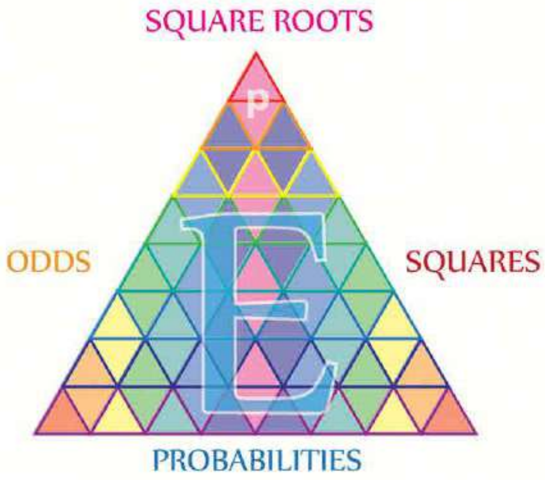

Q-bits

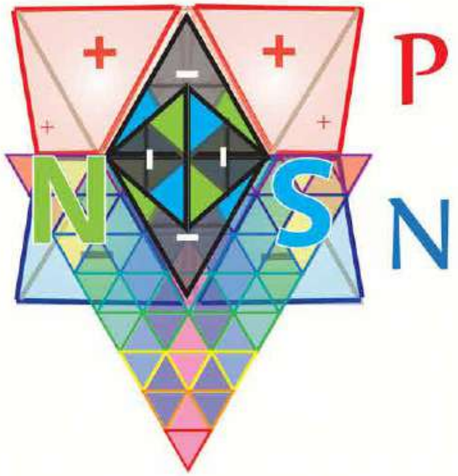

1

*photo-electronic transitions can be used to directly recieve
or emit memory states through the absorption and emission
of spectral photons of specific energy momenta*

0

Copyright ABRAHAM [2008] - All rights reserved

Scalar mass-energies

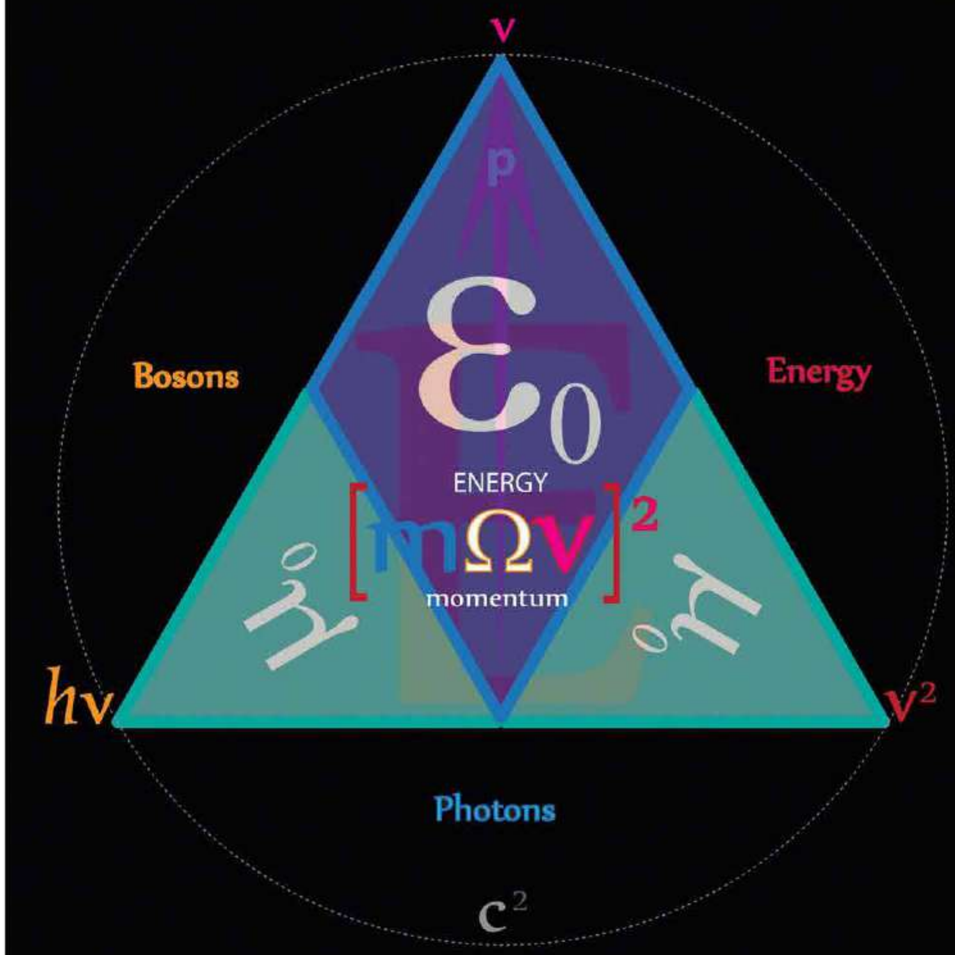
v
p
$\mathcal{E}_0$
ENERGY
Bosons
Energy
$[m\Omega v]^2$
momentum
μ_0
ε_0
hv
v^2
Photons
c^2

Quantum distributions

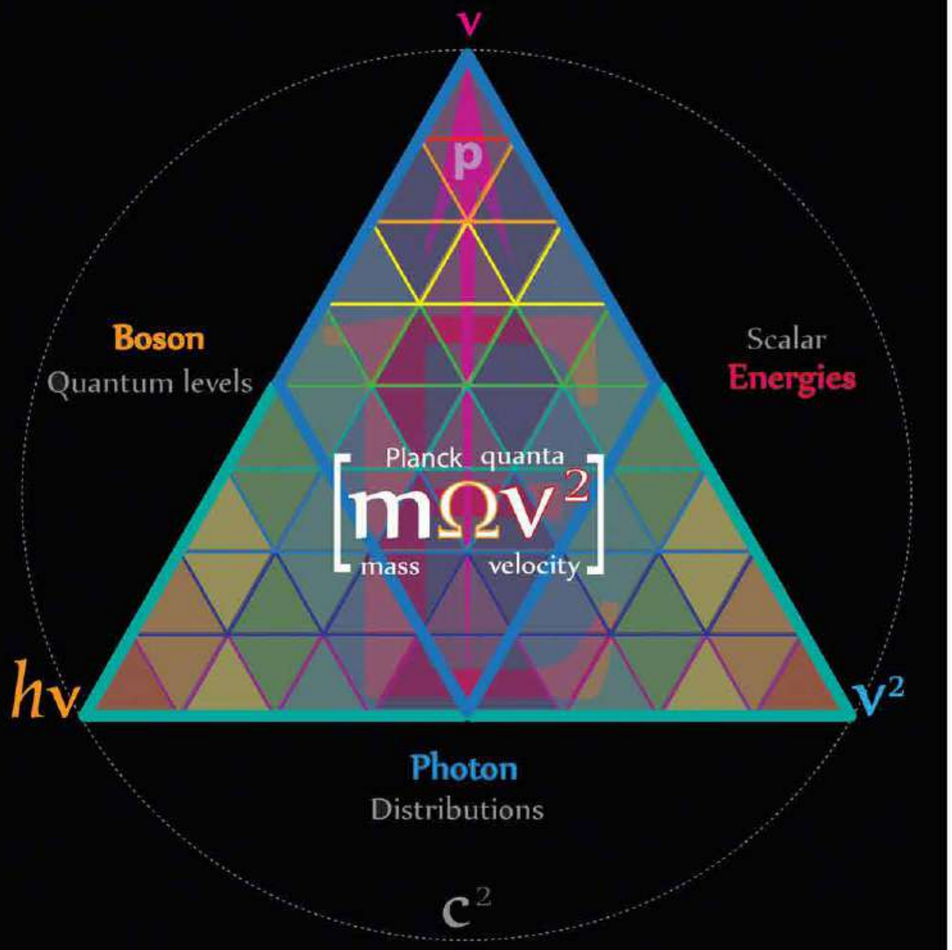
v
p
Boson
Quantum levels
Scalar
Energies
Planck quanta
$[m\Omega v^2]$
mass
velocity
hv
v^2
Photon
Distributions
c^2

EM mass

Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

ENERGY

2D mass

EM mass

$$\frac{n\pi}{c^2}\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

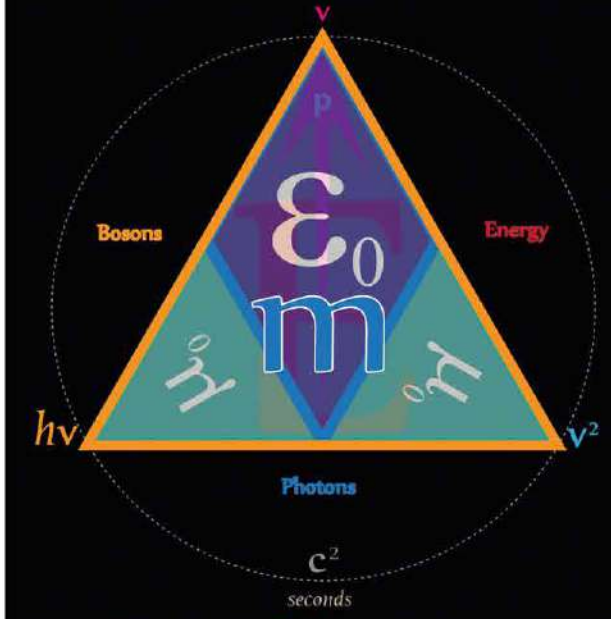
v
p
Bosons
Energy
ε_0
m
hv
v^2
Photons
c^2
seconds

Measure of planar energies
in a nπ radiant wave
geometry
[per second]

mass quantum
7.376238634 e-51 kg

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

momenta

Equilateral energy momenta
has a mass equivalence
per unit of Time

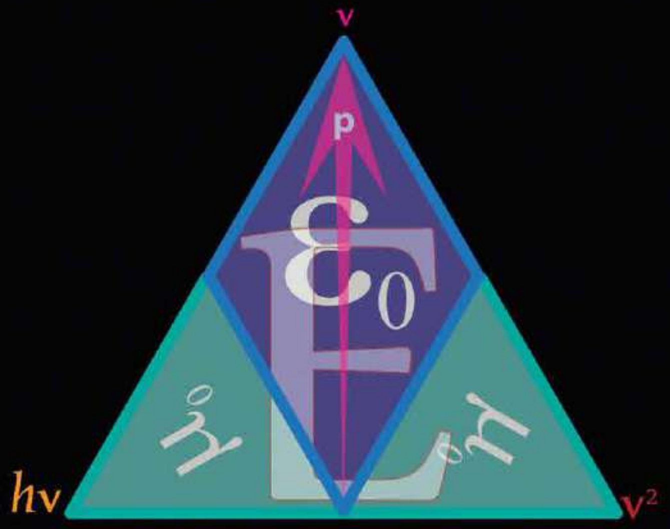
v
p
ε_0
E
hv
v^2

c^2 is the radial geometry
created by radiant energy
in 1 second

Planck Constant
6.629432672 e-34 J.s

3D Matter

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Measure of mass-energy
in a 4nπ standing wave
topology
[per second squared]

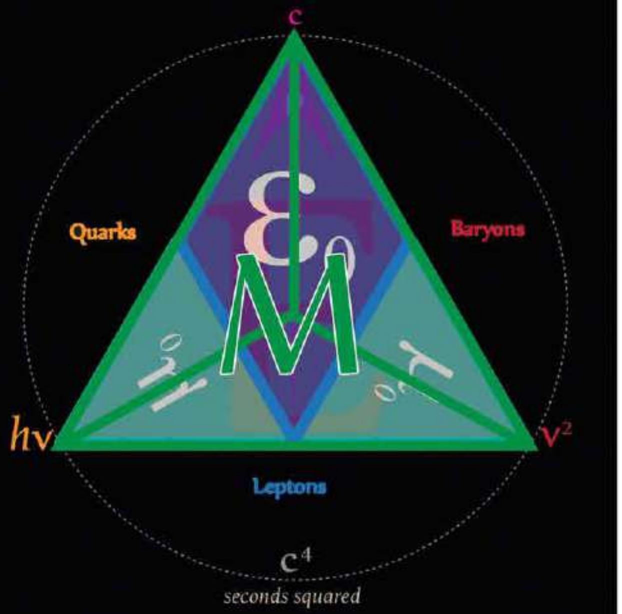
c
Quarks
Baryons
ε_0
M
hv
v^2
Leptons
c^4
seconds squared

Matter quantum
2.9504955454 e-50 kg

Copyright ABRAHAM [2008] - All rights reserved

Planck's constant - the quantum of Action

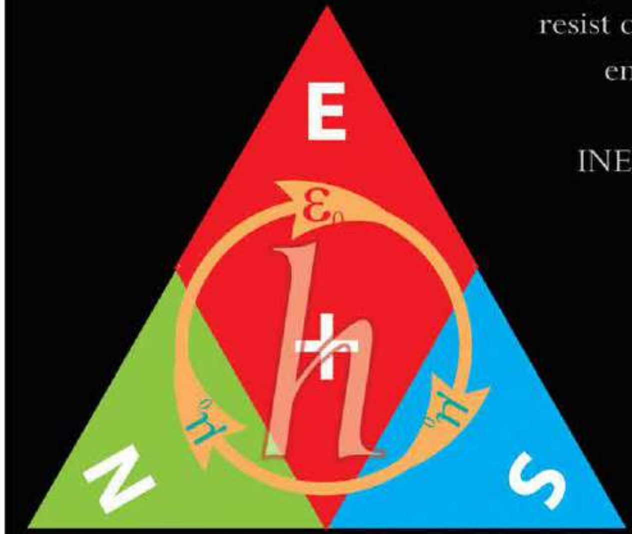
E
ε_0
$\hbar$
μ_0
μ_0
N
S

all ideal quantum inductive loops
resist changes to their
energy levels

INERTIAL mass

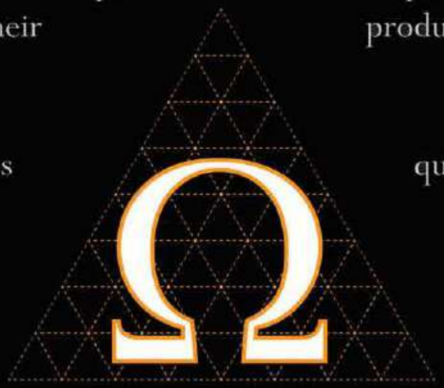

all equilateral energy momenta
produce square root vector
Forces

quantum of WORK

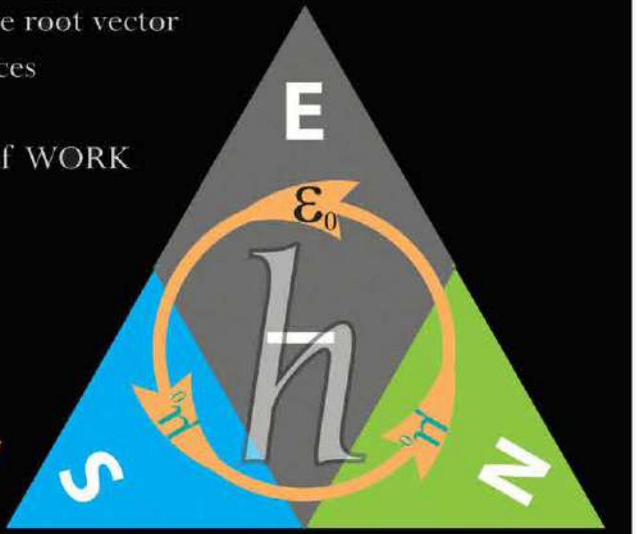
E
ε_0
$\hbar$
μ_0
μ_0
S
N

Quantised angular momentum is equilateral geometry

The two sides of Planck's Constant

The square root of all Energy is linear momentum

Equilateral
energy quanta
per second

mass

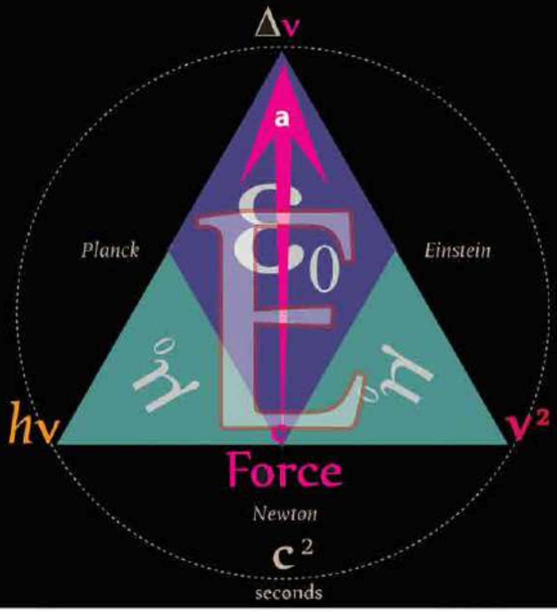
Δv
a
ε_0
E
μ_0
μ_0
Planck
Einstein
hv
v²
Force
Newton
c²
seconds

Quantised
angular momenta
per second

time

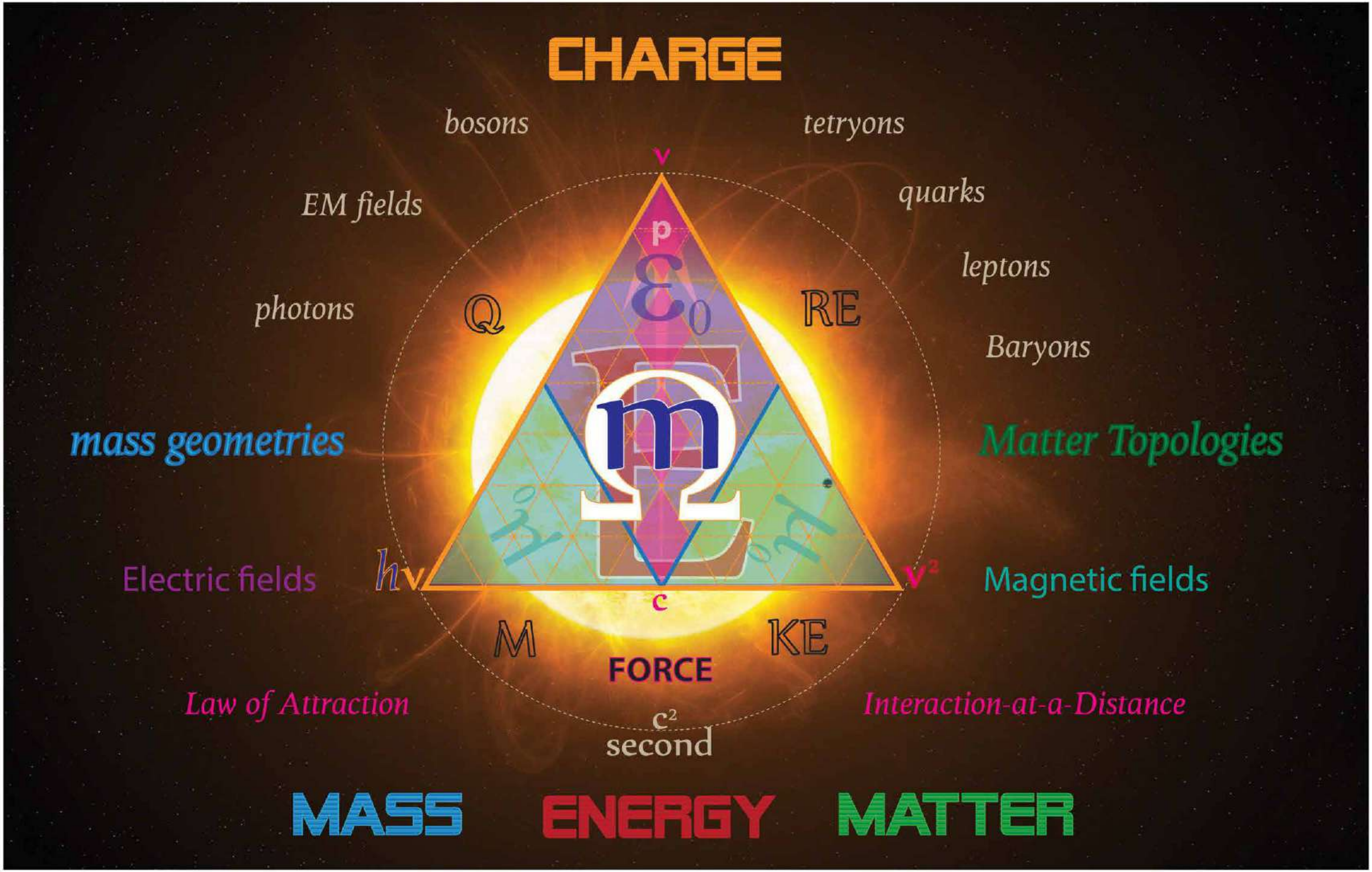
CHARGE
bosons
tetryons
v
EM fields
quarks
p
leptons
ε_0
photons
Q
RE
Baryons
E
m
Ω
mass geometries
Matter Topologies
μ_0
μ_0
Electric fields
hv
v^2
Magnetic fields
c
M
KE
FORCE
Law of Attraction
Interaction-at-a-Distance
c^2
second
MASS
ENERGY
MATTER

Tetryonics 99.04 - Charged Planck mass-ENERGY-Matter

All squared energies have equilateral geometries

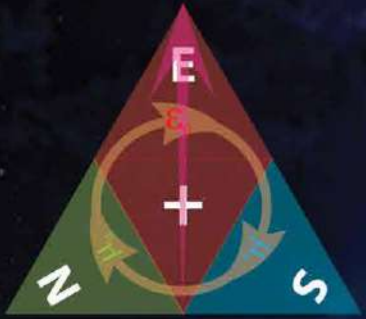

2D
radiant
mass-energy
geometries

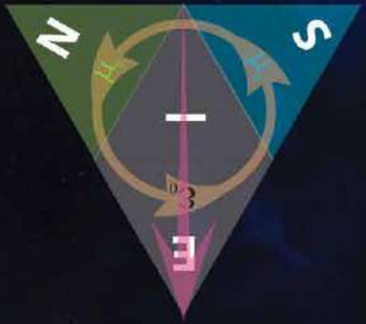

Q

RE

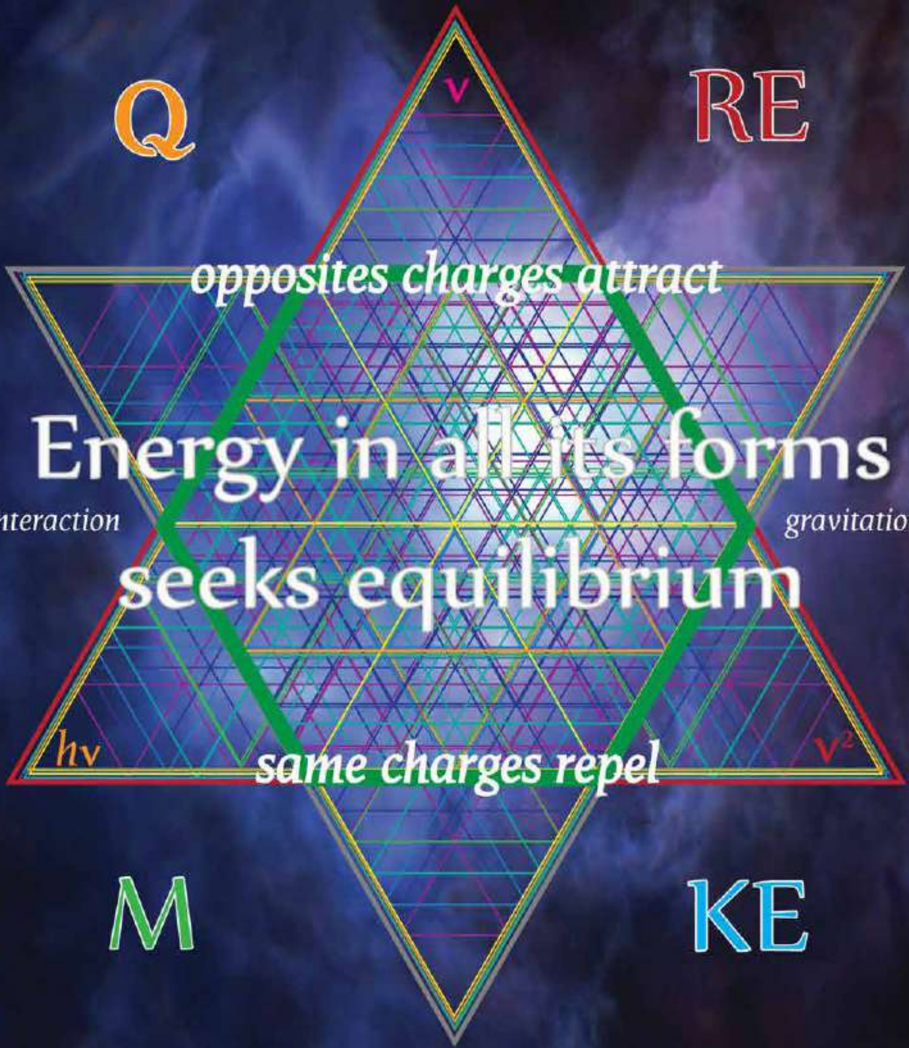
v
opposites charges attract
interaction
gravitation
hv
v²
same charges repel

Energy in all its forms seeks equilibrium

M

KE

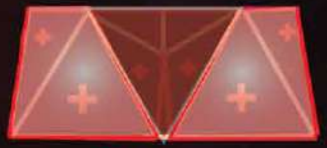

Proton

3D
standing-wave
Matter
topologies

Neutron

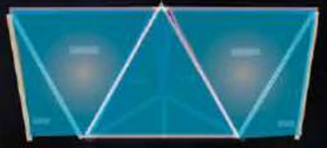

There are NO infinities

only Eternity

Copyright ABRAHAM [2008] - All rights reserved

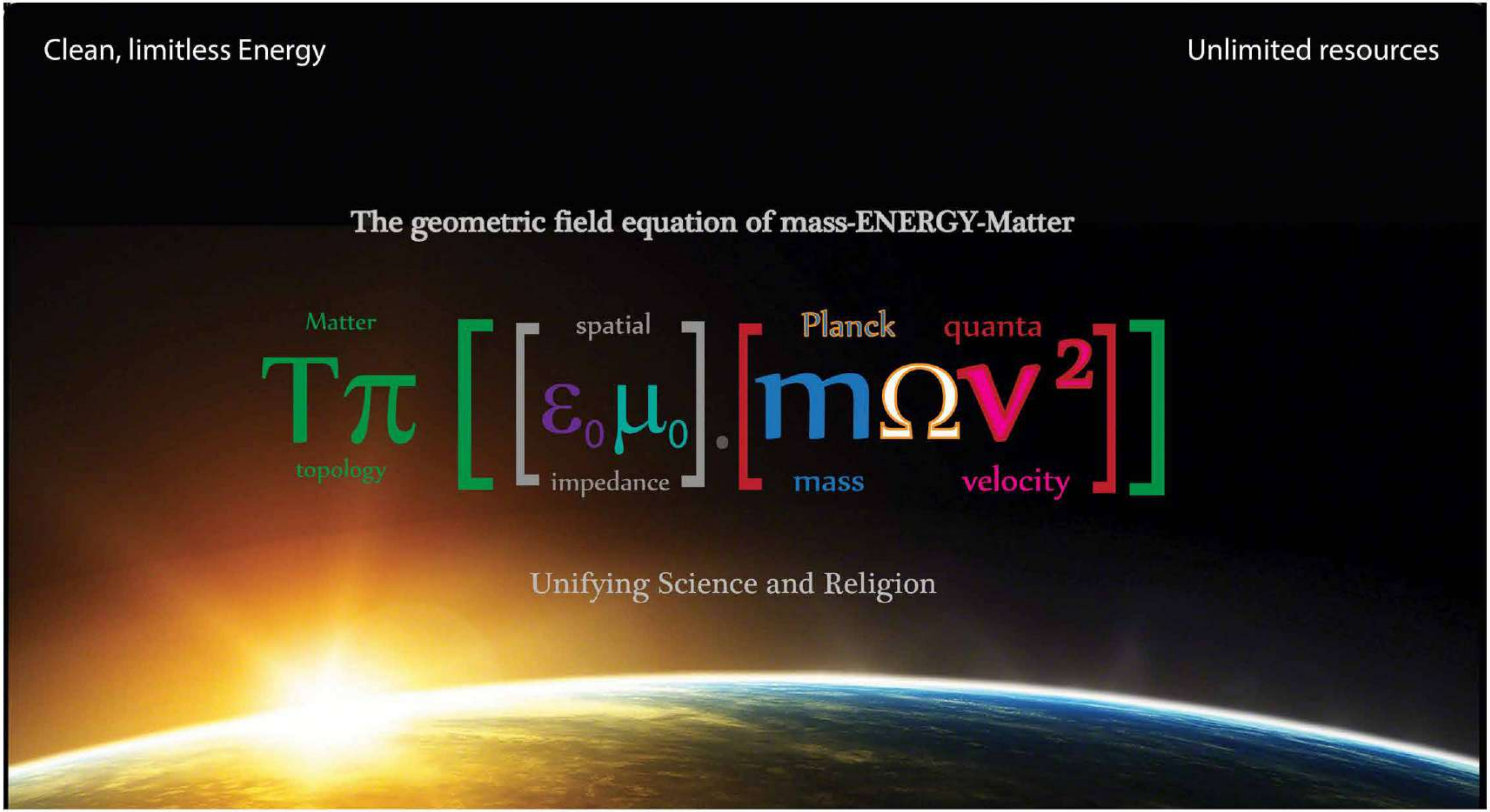

Clean, limitless Energy

Unlimited resources

The geometric field equation of mass-ENERGY-Matter

$$\underset{\text{topology}}{\overset{\text{Matter}}{T\pi}}\left[\left[\underset{\text{impedance}}{\overset{\text{spatial}}{\varepsilon_0\mu_0}}\right]\cdot\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]$$

Unifying Science and Religion

Tetryonics 99.06 - Tetryonic Unified field equation

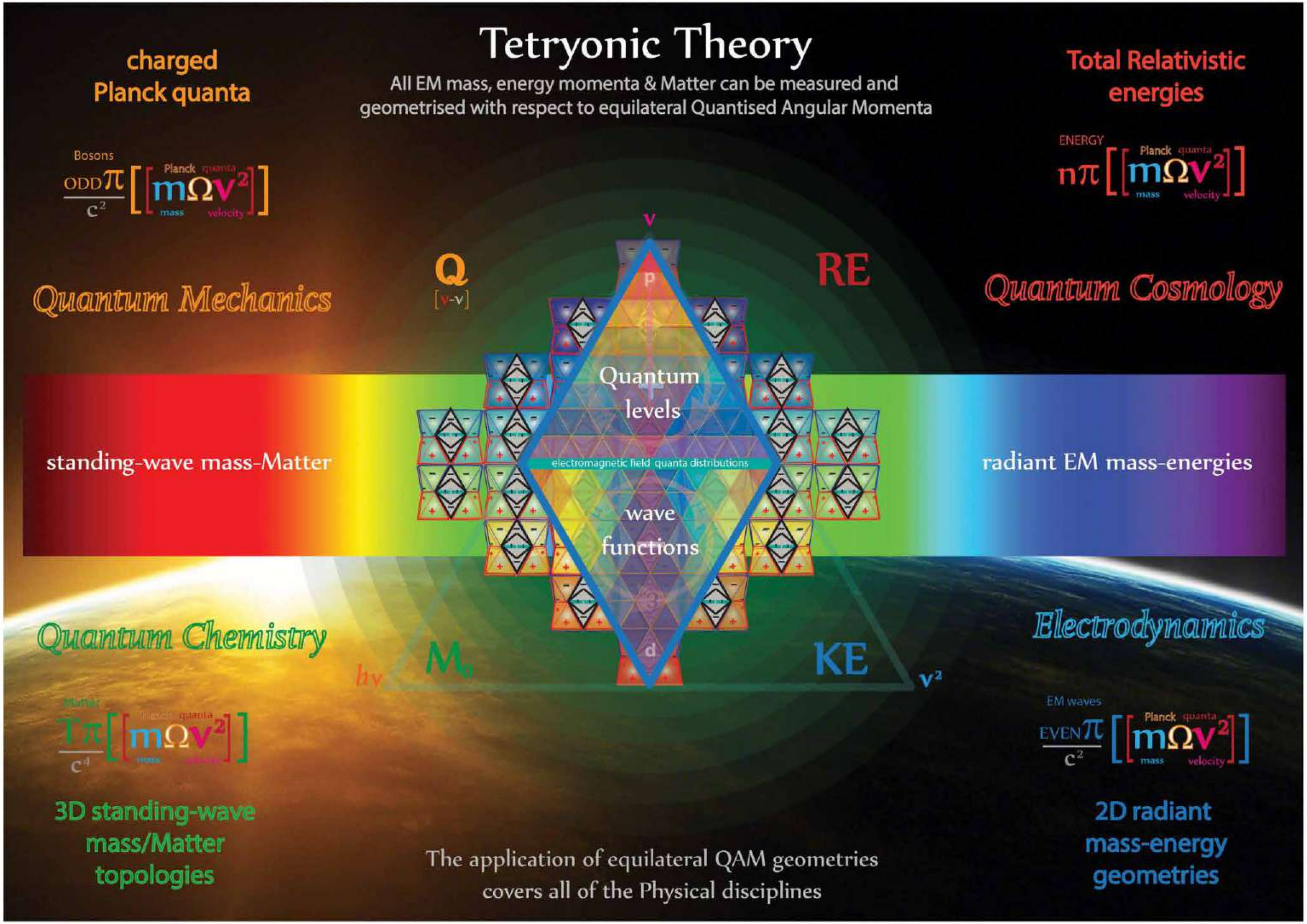

Tetryonic Theory

All EM mass, energy momenta & Matter can be measured and geometrised with respect to equilateral Quantised Angular Momenta

charged Planck quanta

Bosons

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Total Relativistic energies

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

Quantum Mechanics

Quantum Cosmology

standing-wave mass-Matter

radiant EM mass-energies

Quantum Chemistry

Matter

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

3D standing-wave mass/Matter topologies

Electrodynamics

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

2D radiant mass-energy geometries

The application of equilateral QAM geometries covers all of the Physical disciplines

Planck
quanta
mass
velocity
v
Q
[v-v]
RE
p
Quantum
levels
electromagnetic field quanta distributions
wave
functions
d
hv
M_0
KE
v^2

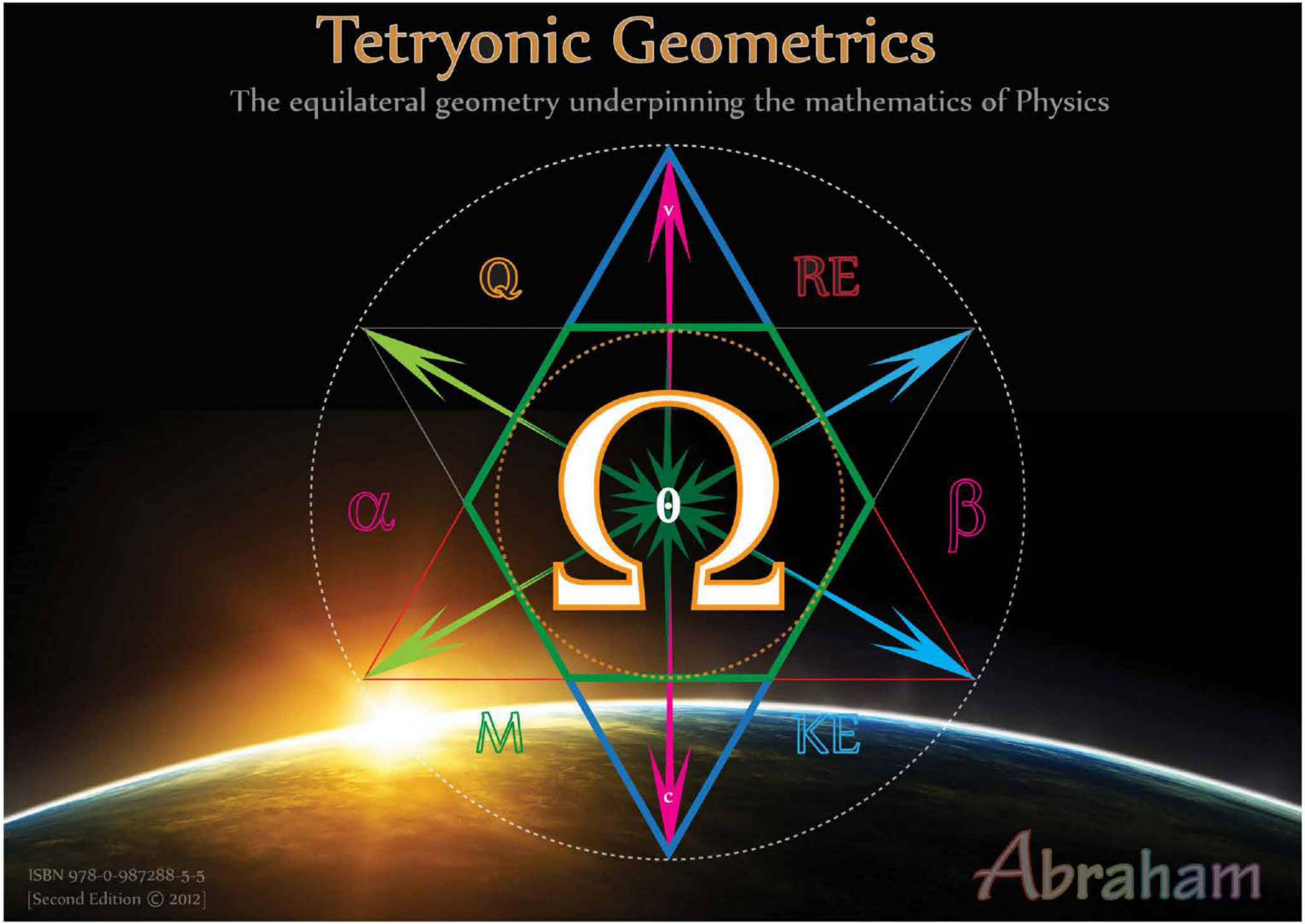

Tetryonic Geometrics

The equilateral geometry underpinning the mathematics of Physics

ISBN 978-0-987288-5-5
[Second Edition © 2012]

Abraham

Copyright ABRAHAM [2008] - All rights reserved

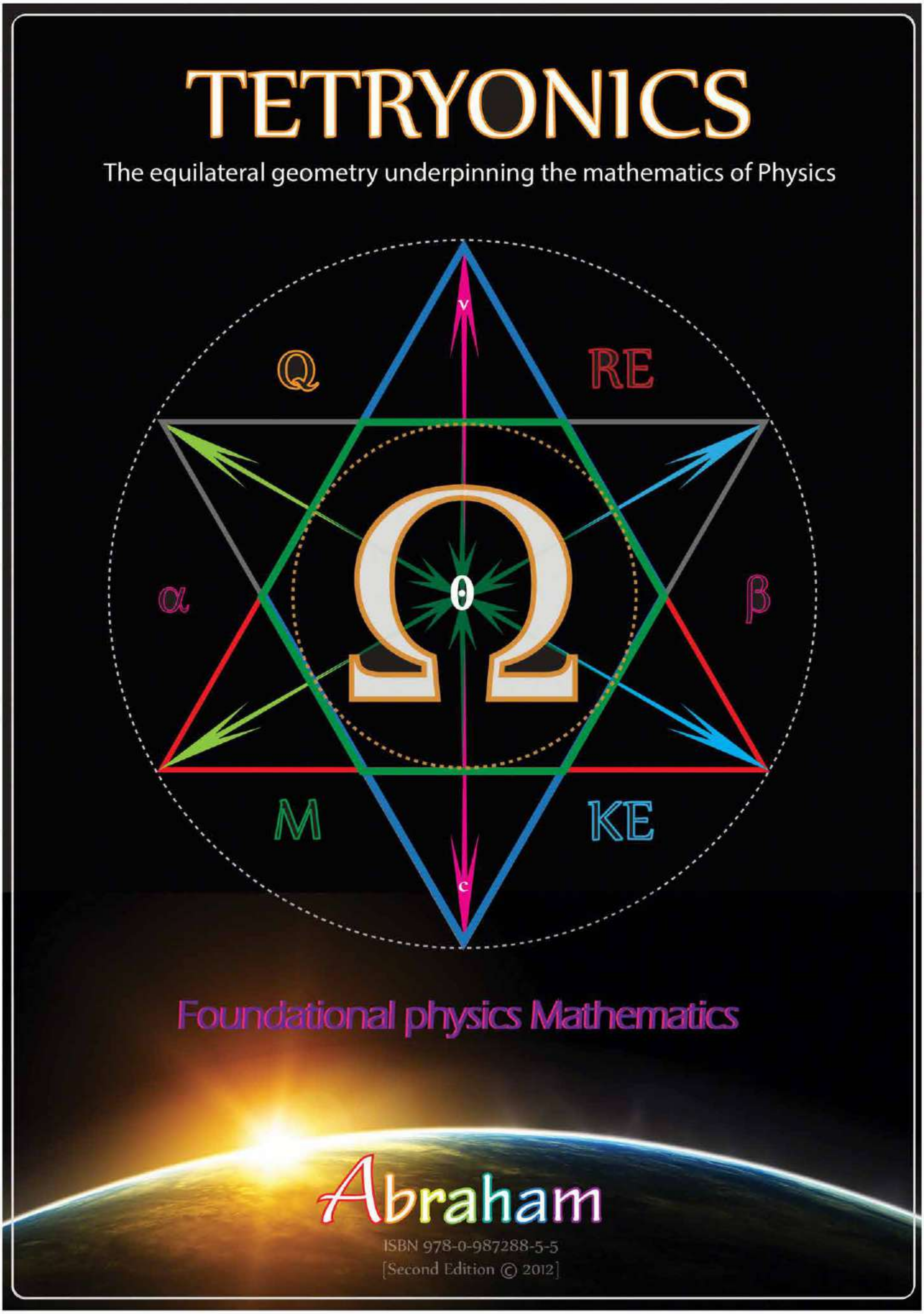

TETRYONICS

The equilateral geometry underpinning the mathematics of Physics

Foundational physics Mathematics

Abraham

ISBN 978-0-987288-5-5

[Second Edition © 2012]

Geometry and the Theory of Everything

Plato

(c.428-348 BC)

DIFFERENT TYPE OF GEOMETRIES

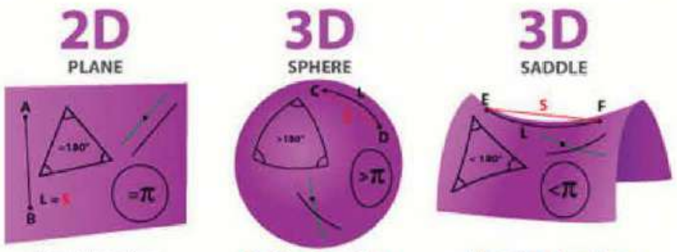

Zero Curvature
Euclidion geometry

Positive Curvature
Elliptic geometry

Negative Curvature
Hyperbolic geometry

(studied by Omar Khayyam, Girolamo Saccheri, Bernhard Riemann, ...)

Euclid

(c.330-275 BC, fl. c.300 BC)

The Socratic tradition was not particularly congenial to mathematics, as may be gathered from Socrates' inability to convince himself that 1 plus 1 equals 2, but it seems that his student Plato gained an appreciation for mathematics after a series of conversations with his friend Archytas in 388 BC.

One of the things that most caught Plato's imagination was the existence and uniqueness of what are now called the five "Platonic solids".

It's uncertain who first described all five of these shapes - it may have been the early Pythagoreans - but some sources (including Euclid) indicate that Theaetetus (another friend of Plato's) wrote the first complete account of the five regular solids.

Presumably this formed the basis of the constructions of the Platonic solids that constitute the concluding Book XIII of Euclid's Elements.

In any case, Plato was mightily impressed by these five definite shapes that constitute the only perfectly symmetrical arrangements of a set of (non-planar) points in space, and late in life he expounded a complete "theory of everything", in the treatise called Timaeus, based explicitly on these five solids.

Interestingly, almost 2000 years later, Johannes Kepler was similarly fascinated by these five shapes, and developed his own cosmology from them

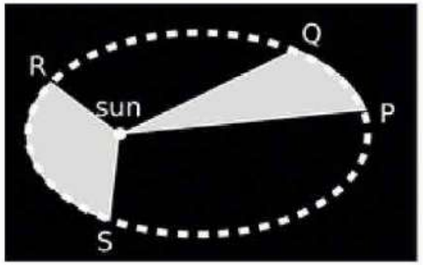

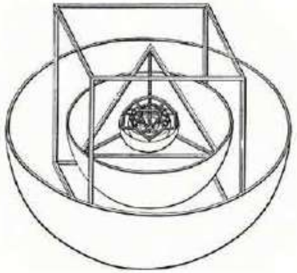

Tetractys

The Greek Tetractys is a
triangular figure consisting of ten points
arranged in four rows:

one, two, three, and four points in each row,
which is the geometrical representation of the
fourth triangular number.

As a mystical symbol, it was very important to
the secret worship of the Pythagoreans.

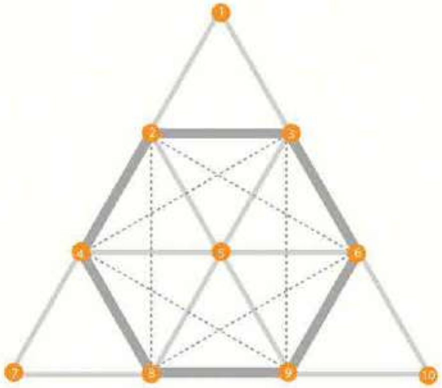

Sacred numbers

A

The Tetractys historically symbolized the four elements
[Earth, Air, Fire, and Water]
and the relationship between Humanity
and the cosmos created by GOD

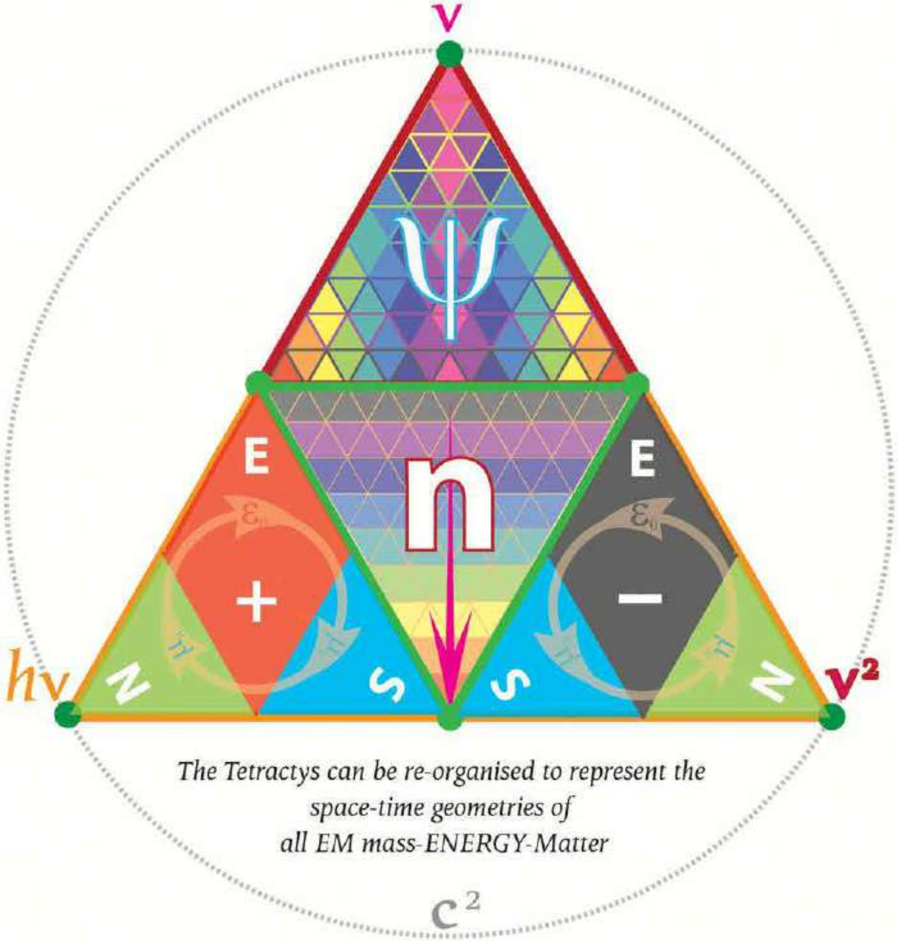

*The Tetractys can be re-organised to represent the
space-time geometries of
all EM mass-ENERGY-Matter*

The single triangle in the first row represents zero-dimensions (a point)
A vector direction in one-dimension can be represented as a line between any two points
The second row represents a Boson (two-dimensions in a plane defined by a rhombus of three triangles)
The whole figure folded represents three-dimensions (a tetrahedron defined by four apex points)
Photons of ElectroMagnetic mass-Energy quanta are represented by two opposing triangles

The tetrad was the name given to
the number four -
in Pythagorean philosophy
there were four seasons
and four elements,
and the number was also associated
with planetary motions and music

As a mystical symbol, it was very important to
the secret worship of the Pythagoreans.

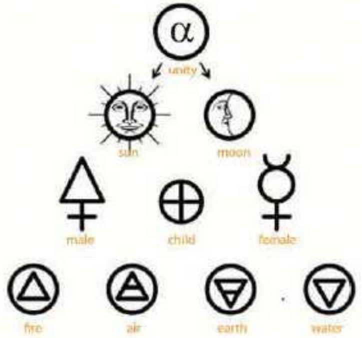

The Cosmos

Ω

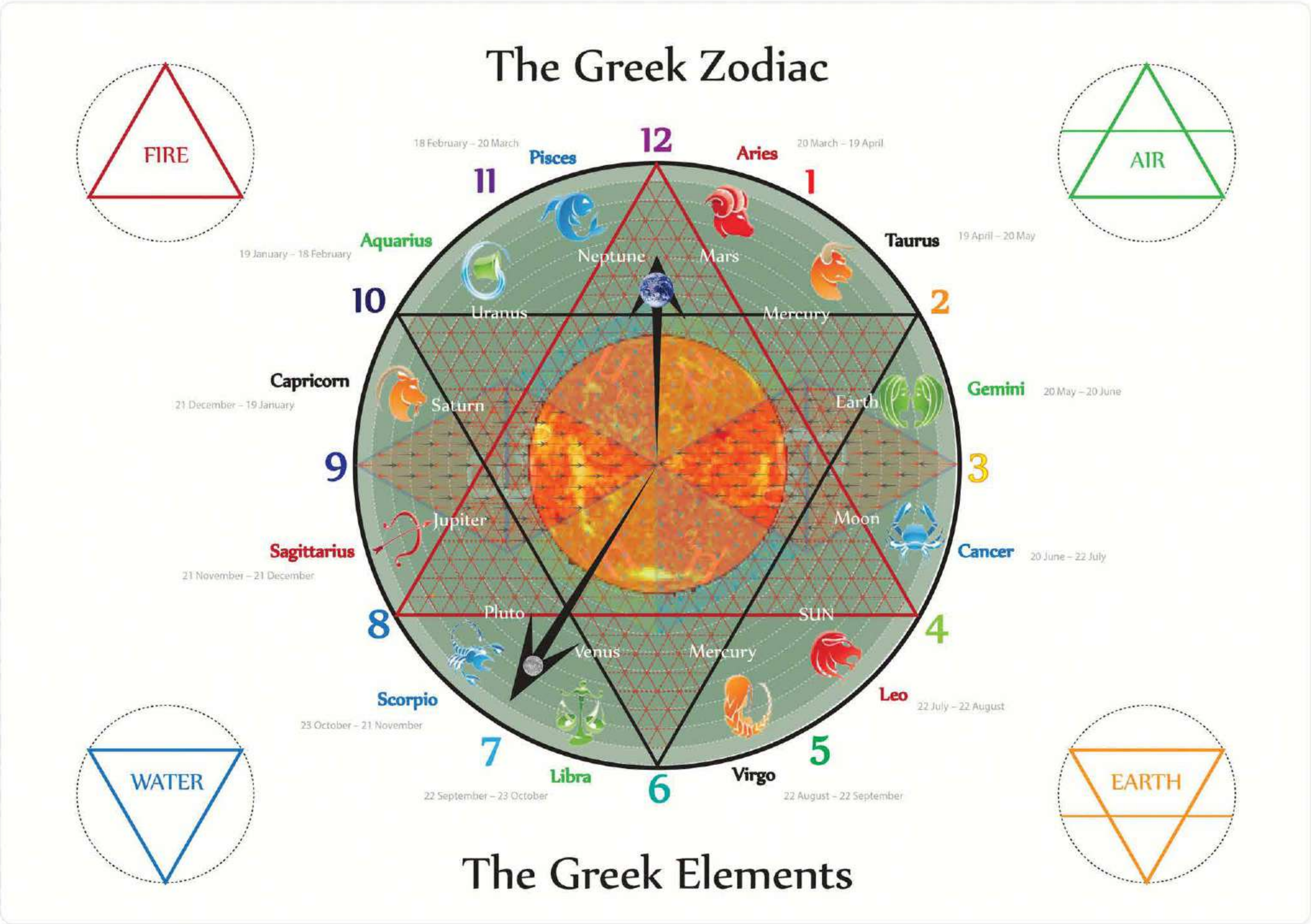
The Greek Zodiac
FIRE
AIR
12
Pisces
18 February – 20 March
11
Aries
20 March – 19 April
1
Aquarius
19 January – 18 February
10
Taurus
19 April – 20 May
2
Neptune
Mars
Uranus
Mercury
Capricorn
21 December – 19 January
Saturn
Earth
Gemini
20 May – 20 June
9
3
Jupiter
Moon
Sagittarius
21 November – 21 December
Cancer
20 June – 22 July
8
Pluto
SUN
4
Venus
Mercury
Scorpio
23 October – 21 November
Leo
22 July – 22 August
7
Libra
22 September – 23 October
6
Virgo
22 August – 22 September
5
WATER
EARTH
The Greek Elements

Copyright ABRAHAM [2008] - All rights reserved

Equilateral Triangles

An equilateral triangle is a triangle in which all three sides are equal

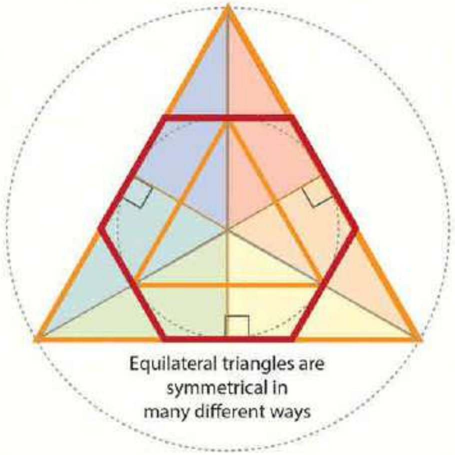

Equilateral triangles are
symmetrical in
many different ways

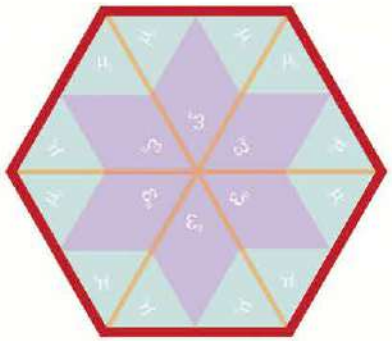

Any six equilateral triangles
joined can make a hexagon.

It is unique in that it is the only polygon
that can be tiled [or divided] and produce
only identical geometries and squares numbers

The equilateral triangle
is eminently suited for
the construction of fractals

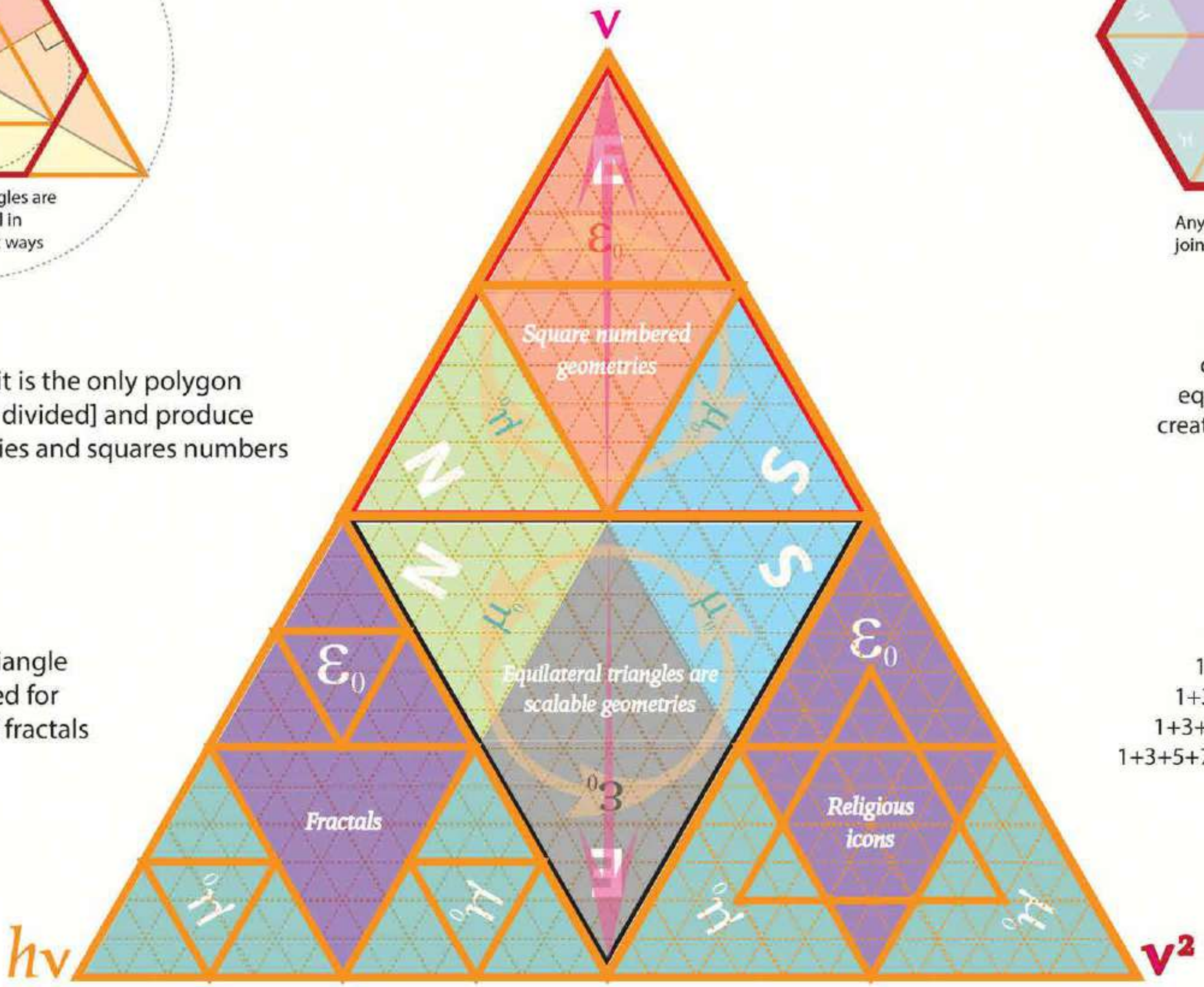
v
ε_0
Square numbered geometries
μ_0
N
S
Equilateral triangles are scalable geometries
ε_0
Fractals
Religious icons
μ_0
hv
v²

An equilateral triangle is simply a specific case of a regular polygon with 3 sides

The
tesselation of
odd numbered
equilateral triangles
creates square numbers

That is,

1
1+3 = 4
1+3+5 = 9
1+3+5+7 = 16
1+3+5+7+9 = 25
1+3+5+7+9+11 = 36
1+3+5+7+9+11+13 = 49
1+3+5+7+7+9+11+13+15 = 64
etc

Copyright ABRAHAM [2008] - All rights reserved

The Pythagorean Theorem

Though attributed to Pythagoras, it is not certain that he was the first person to prove it.

The first clear proof came from Euclid, and it is possible the concept was known 1000 years before Pythoragas by the Babylonians

$$a^2 + b^2 = c^2$$

The square of the hypotenuse of a triangle is equal to the sum of the squares of its sides.

In any right triangle,
the area of the square whose side is the hypotenuse
(the side opposite the right angle) is equal to
the sum of the areas of the squares
whose sides are the two legs
(the two sides that meet at a right angle)

Pythagoras of Samos

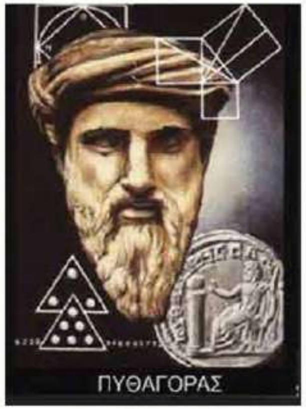

about (570 – 495 BC)

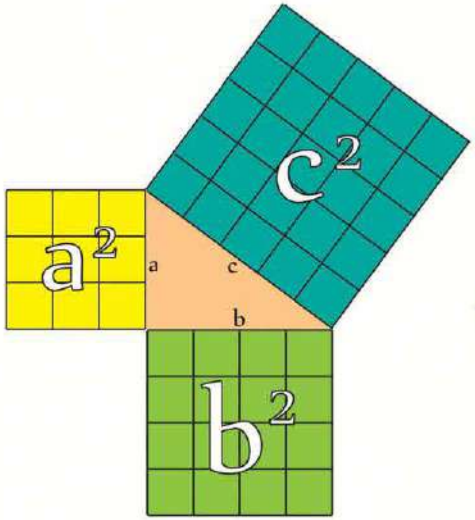

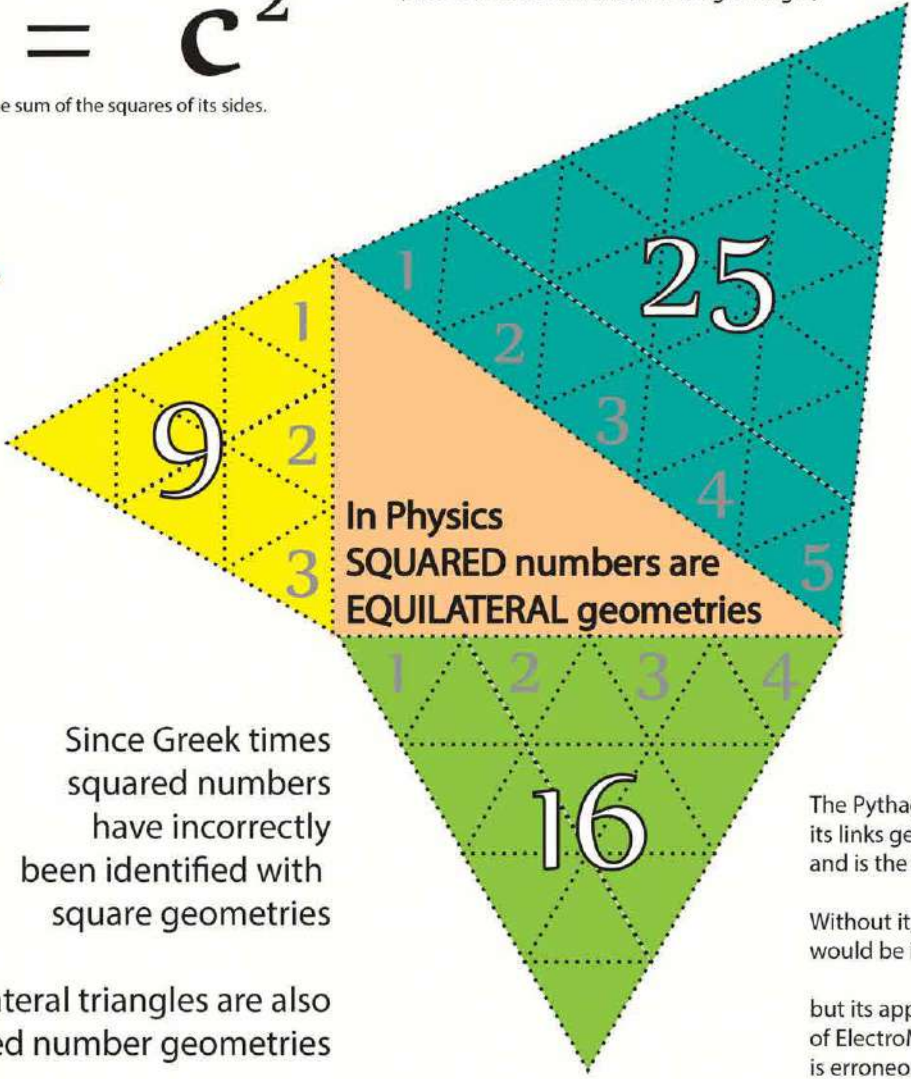

In Physics
SQUARED numbers are
EQUILATERAL geometries

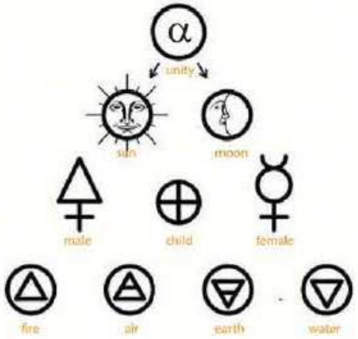

Pythagorean Tetractys

Since Greek times
squared numbers
have incorrectly
been identified with
square geometries

Equilateral triangles are also
squared number geometries

The Pythagorean equation is at the core of much of geometry,
its links geometry with algebra,
and is the foundation of trigonometry.

Without it, accurate surveying, mapmaking, and navigation
would be impossible,

but its application to the energy-momenta geometries
of ElectroMagnetic fields and Matter in motion in Physics
is erroneous and must be corrected for science to advance

Hexagons

A regular hexagon can be subdivided into six equilateral triangles

Hexagons are the only regular polygon that can be subdivided into another regular polygon

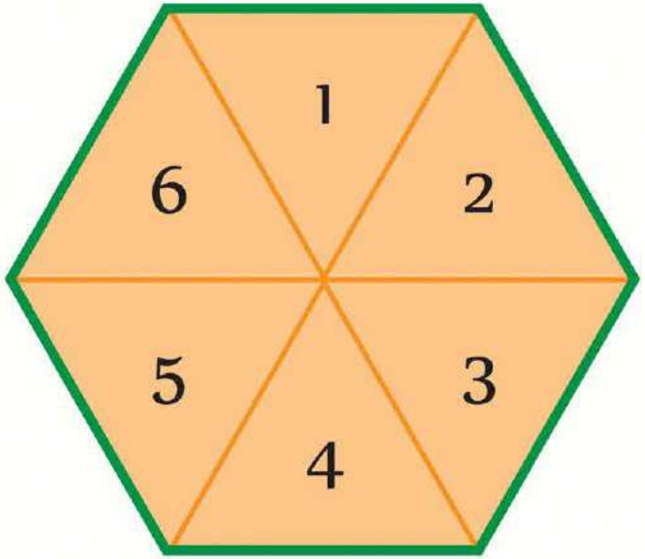

Hexagons are the unique regular polygon such that the distance between the center and each vertex is equal to the length of each side

Six is a highly composite number, the second-smallest composite number, and the first perfect number.

That is, 1*2*3=1+2+3=6

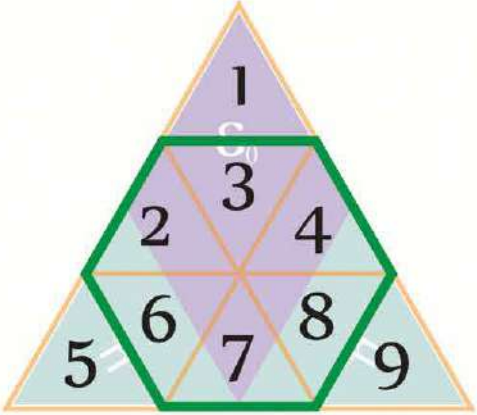

Energy geometries

Atomic nuclei geometries

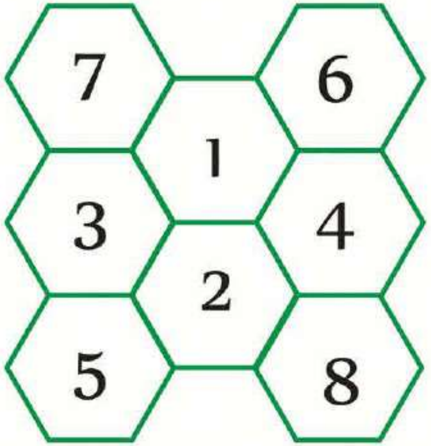

Hexagons can be tiled or tessellated in a regular pattern on a flat two-dimensional plane

An interesting relationship between circular and hexagonal geometry is that hexagonal patterns often appear spontaneously when natural forces are trying to approximate circles

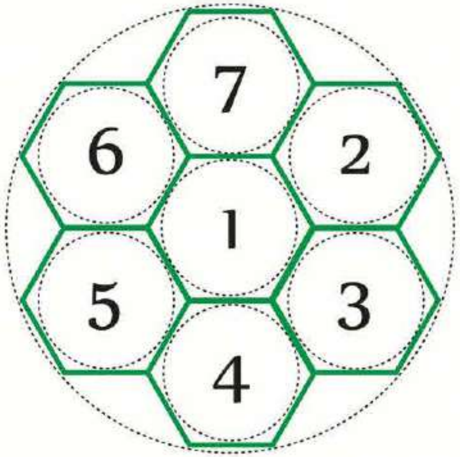

Hexagonal tessellation is topologically identical to the close packing of circles on a plane

Copyright ABRAHAM [2008] - All rights reserved

Platonic Solids

The Pythagoreans knew that there were only five regular convex solids,
the tetrahedon, cube, octahedron, icosohedron and dodecahedron
and each one could be accurately circumscribed by a sphere.

FIRE

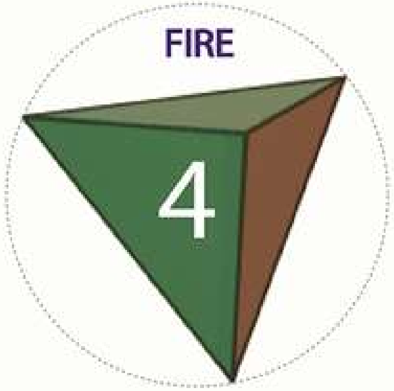

4 triangles meet to form
a Tetrahedron

EARTH

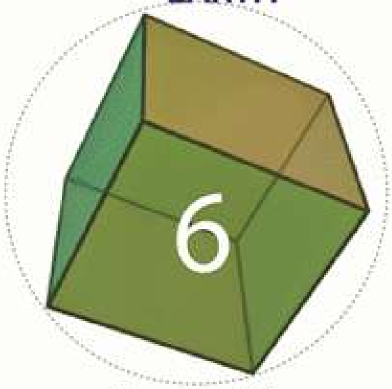

6 squares meet to form
a Cube

AIR

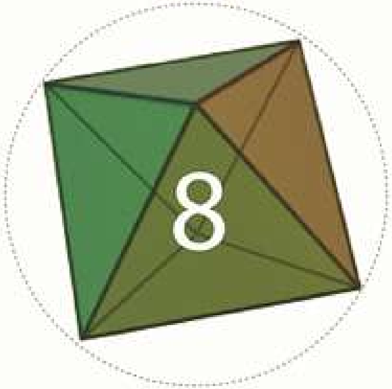

8 triangles meet to form
a Octahedron

tetryons **4**

regular deltahedrons

leptons **12**

8 ***quarks***

regular deltahedrons

20 ***Baryons***

$$F - E + V = 2$$

faces edges vertexs

12 pentagons meet to form
a Dodecahedron

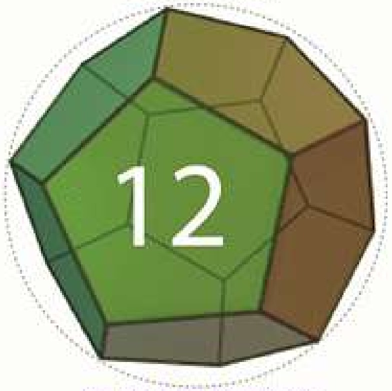

THE HEAVENS

Plato

(c.428-348 BC)

The philosopher Plato concluded that they must be the fundamental building blocks
– the atoms – of nature, and assigned to them what he believed to be
the essential elements of the universe.

20 triangles meet to form
a Icosahedron

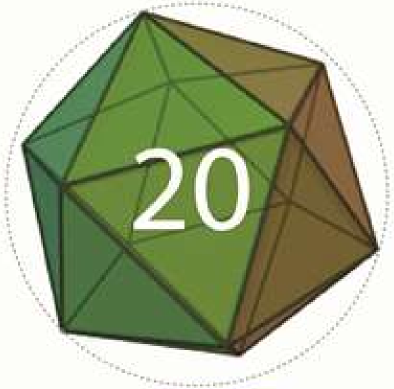

WATER

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic Solids

Despite their unique topologies Tetryonic solids are not unlike Platonic solids save that their toplogies are comprised entirely from complex hitherto undescribed **4nπ equilateral Planck mass-energy momenta geometries** that also match the **Euler numbers of Platonic solids**

+ tetryon

4 faces
6 edges
4 vertices

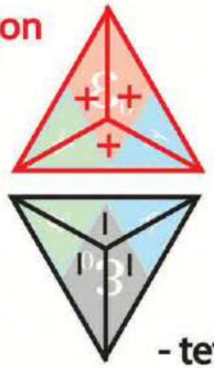

2

- tetryon

up quark

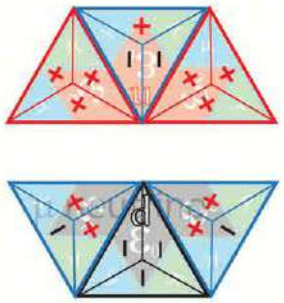

2

8 faces
12 edges
6 vertices

down quark

mass geometries

4π *tetryons* **4π**

regular deltahedrons

12π *leptons* **12π**

Matter topologies

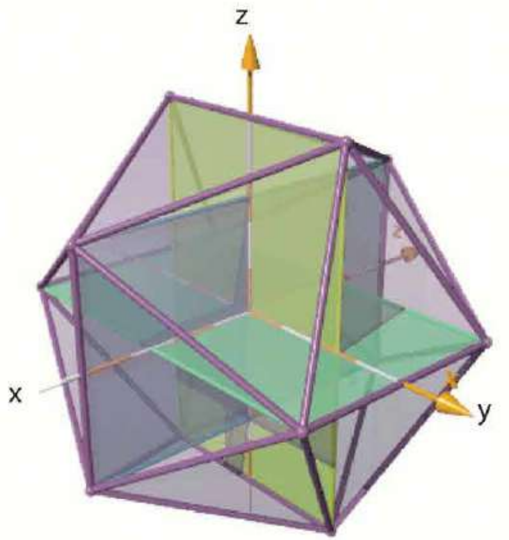

Matter topologies

8π *quarks* **12π**

regular deltahedrons

20π *Baryons* **36π**

mass geometries

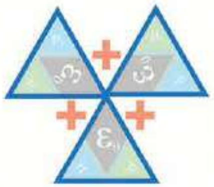

neutrino

12 faces
18 edges
8 vertices

2

electron

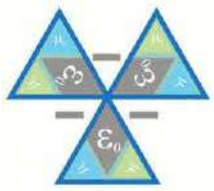

Their equilateral topologies are best described as regular topologic-deltahedrons:

tetra-delta-hedrals	**tetryons**	4π external charge fascia
octa-delta-hedrals	**quarks**	8π external charge fascia
dodeca-delta-hedrals	**leptons**	12π external charge fascia
icoso-delta-hedrals	**Baryons**	20π external charge fascia

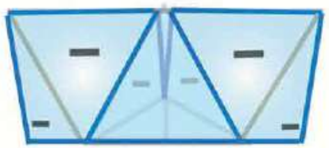

Neutron

2

20 faces
30 edges
12 vertices

Proton

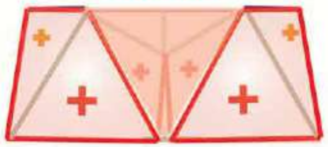

note:
Charged mass-energy fascia geometries and edges become "hidden" upon the meshing of delta-hedra to form Matter topologies

Copyright ABRAHAM [2008] - All rights reserved

Euclidean geometry

Euclid

(c.330-275 BC, fl. c.300 BC)

Arguably the most influential Mathematics book ever written is Euclid's 'The Elements'

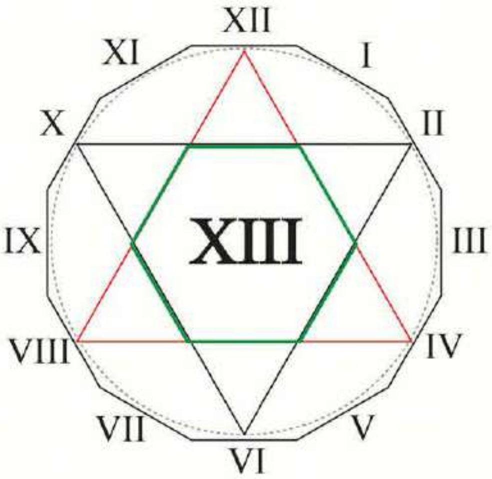

In all, it contains 465 theorems and proofs, described in a clear, logical and elegant style, and using only a compass and a straight edge.

The Elements - Book 1 - Definition 20

Of the trilateral figures, an equilateral triangle is that which has its three EQUAL sides

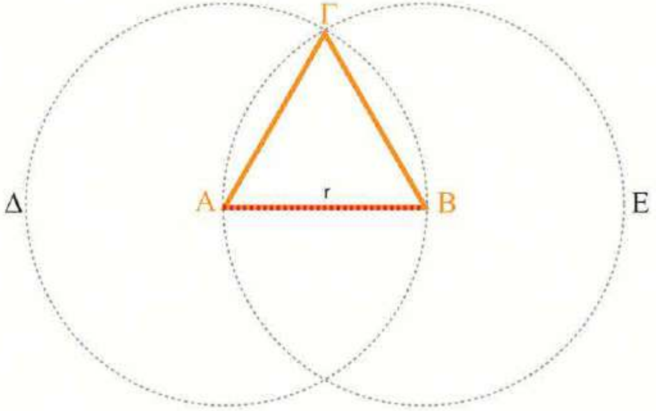

Euclid's Elements - Book 1 - Proposition 1
Method of constructing an Equialteral triangle

Euclid's five general axioms were:

Things which are equal to the same thing are equal to each other.

If equals are added to equals, the wholes (sums) are equal.

If equals are subtracted from equals, the remainders (differences) are equal.

Things that coincide with one another are equal to one another.

The whole is greater than the part

Copyright ABRAHAM [2008] - All rights reserved

Archimedes

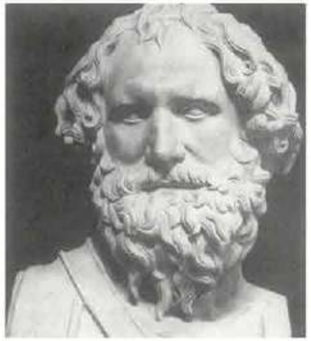

c. (287 BC – c. 212 BC)

As its definition relates to the circle,
π is found in many formulae in trigonometry and geometry,
especially those concerning circles, ellipses, or spheres.

It is also found in formulae from other branches of science,
such as cosmology, number theory, statistics, fractals,
thermodynamics, mechanics, and electromagnetism

Incorrect identification of Pi [c/d] as opposed to Pi radians
in Physics has led to the inappropriate association
of spherical particles to the physical sciences
whereas equilateral triangles & tetrahedra
form its true geometry

The number π is a mathematical constant that is
the ratio of a circle's circumference to its diameter.

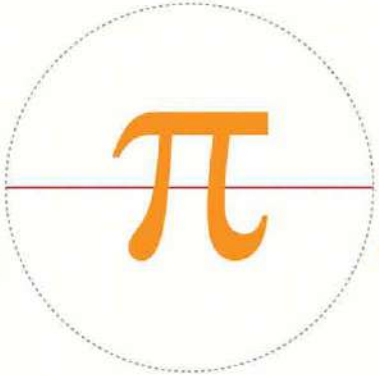

Pi

π is an irrational number
its decimal representation
never ends and never repeats.

$\pi = c/d$

The ratio C/d is constant,
regardless of the circle's size

3.141592654.......

Proof of the fact that C=2πr and how Archimedes proved it

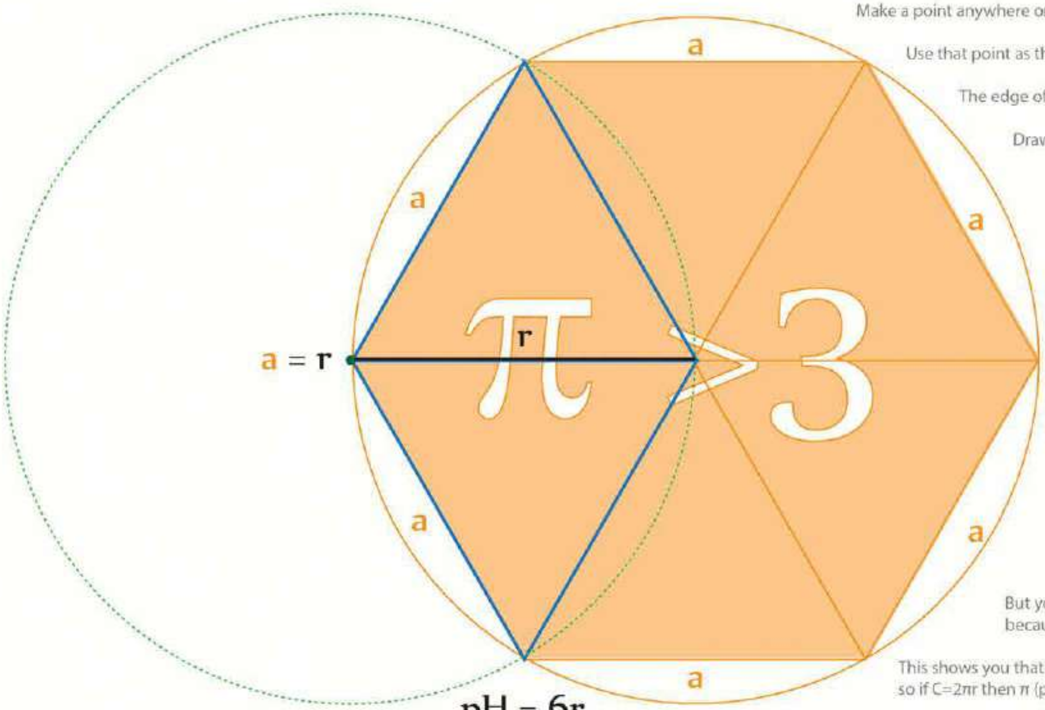

Draw any circle.

Make a point anywhere on the circumference of the [green] circle.

Use that point as the center of a second [blue] circle with the same radius as the green circle.

The edge of the blue circle should touch the center of the green circle.

Draw the line segment connecting the centers of the two circles.

That forms the radius of both of the circles.

Now draw the line connecting the center of the blue circle to where it crosses the green circle
on both sides, and complete the triangles.

You should have two equilateral triangles whose sides are equal to the radius of the green circle

Now extend all of the radius lines so they become diameter lines,
all the way across the circle, and finish drawing all of the triangles to connect them.

You've got six equilateral triangles now, that make an orange hexagon.

So the perimeter of your hexagon is the same as six times the radius of your circle.

But your circumference is a little bigger than the perimeter of your hexagon,
because the shortest distance between two points is always a straight line.

This shows you that the circumference of the blue circle has to be more than 6r,
so if C=2πr then π (pi) has to be a little bigger than 3, which it is.

The more sides we draw on our polygon, the closer we will get to the real value of pi (3.14159 etc.).
Using a polygon with 96 sides, Archimedes was able to calculate that π was a little bigger than 3.1408

The Golden Ratio

Two quantities are in the golden ratio (φ) if the ratio of the sum of the quantities to the larger quantity is equal to the ratio of the larger quantity to the smaller one.

The figure to the right illustrates the geometric relationship

[1.61803]

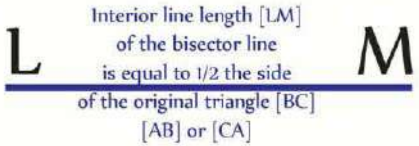

Interior line length [LM] of the bisector line is equal to 1/2 the side of the original triangle [BC] [AB] or [CA]

Φ

Golden ratio phi

$$\hbar = \frac{h}{2\pi}$$

Tetryonic geometry reveals the maximum E-field amplitude of the reduced Planck constant to be an example of the Golden ratio in physics

velocity vectors

The height of the triangle [ALM] produced by the bisecting line is 1/2 the height of the height of [ABC]

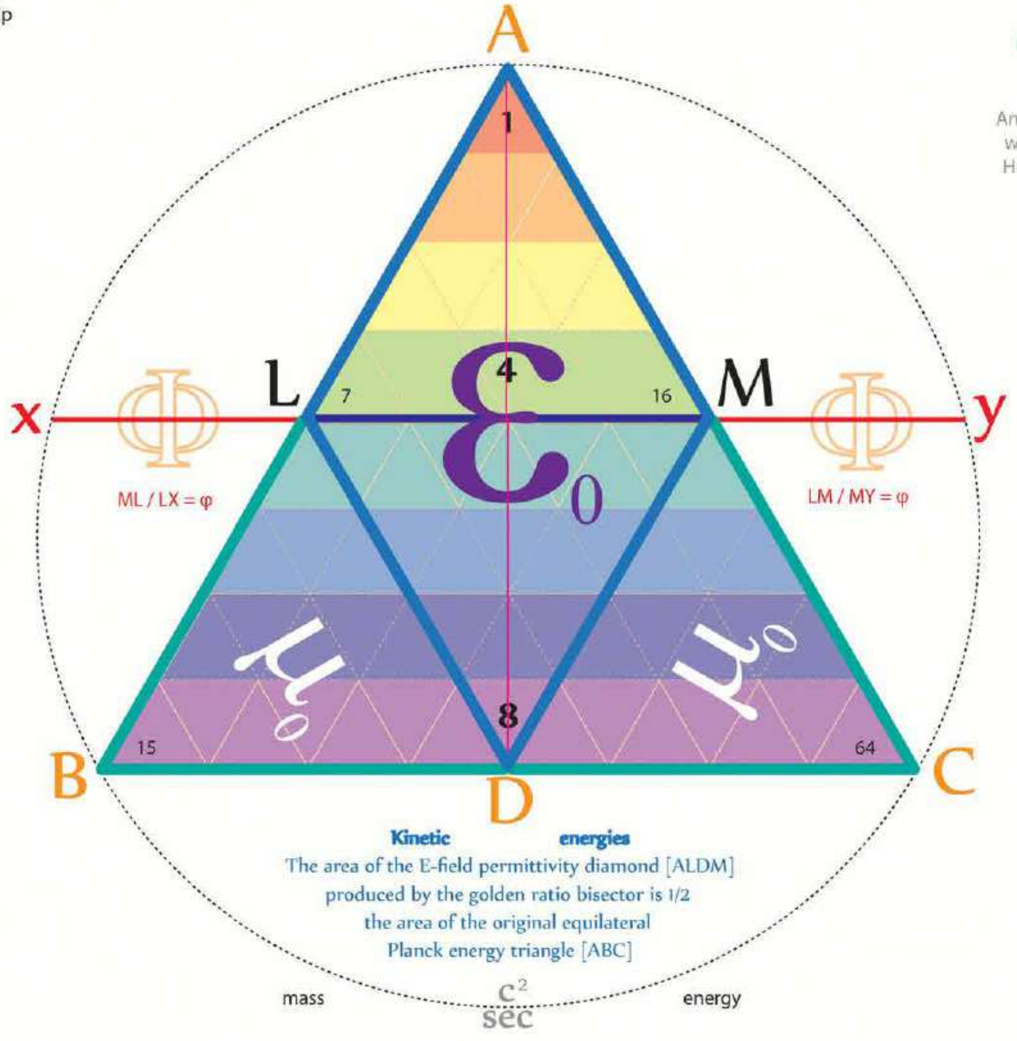

Kinetic energies

The area of the E-field permittivity diamond [ALDM] produced by the golden ratio bisector is 1/2 the area of the original equilateral Planck energy triangle [ABC]

Planck's constant [*h*]

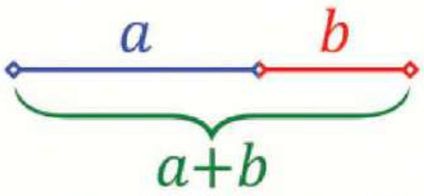

a+*b* is to *a* as *a* is to *b*

An intriguing showing of φ in an equilateral triangle was observed by George Odom, a resident of the Hudson River Psychiatric Center, in the early 1980s

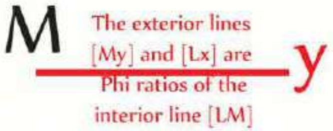

The exterior lines [My] and [Lx] are Phi ratios of the interior line [LM]

Let L and M be the midpoints of the sides AB and AC of an equilateral triangle ABC

By measurement and the Intersecting Chords Theorem

$$MX \cdot MY = AM \cdot MC.$$

which is of the form

$$(a + b) \cdot b = a \cdot a.$$

Denoting a/b = x, we see that

$$1 + x = x^2,$$

George Odom

The Golden Rhombus

Applying the golden ratio (φ) to quantum scale electrodynamic geometry we can quickly determine that the linear momentum and magnetic moment vectors of photons & EM waves can also be expressed as a golden ratio

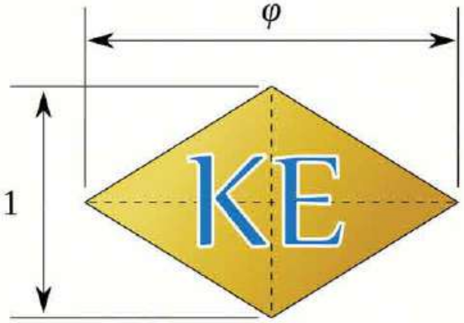

vector velocities

vs

magnetic vector

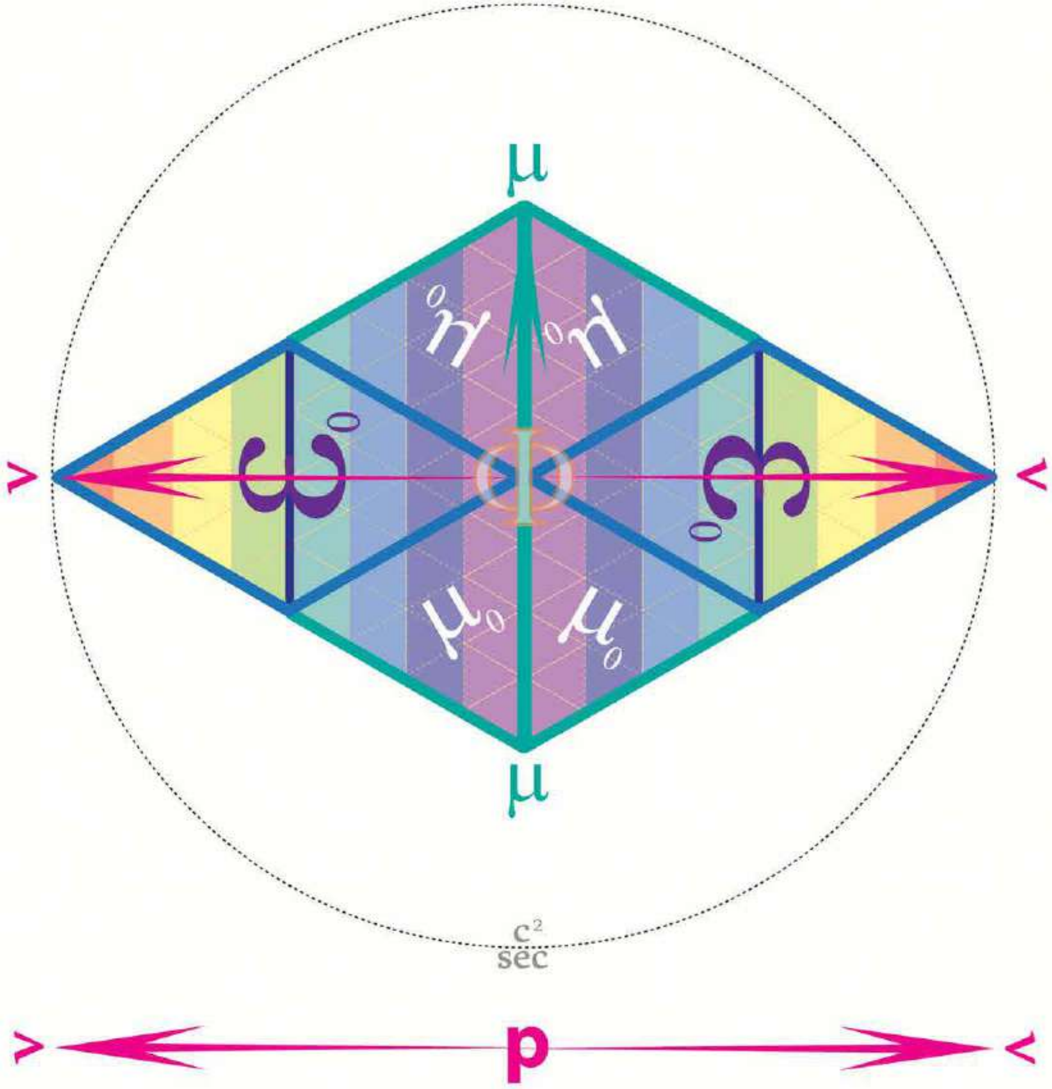

Φ

Golden ratio phi

[1.61803]

magnetic Force

vs

electric Force

photons

of

EM energy

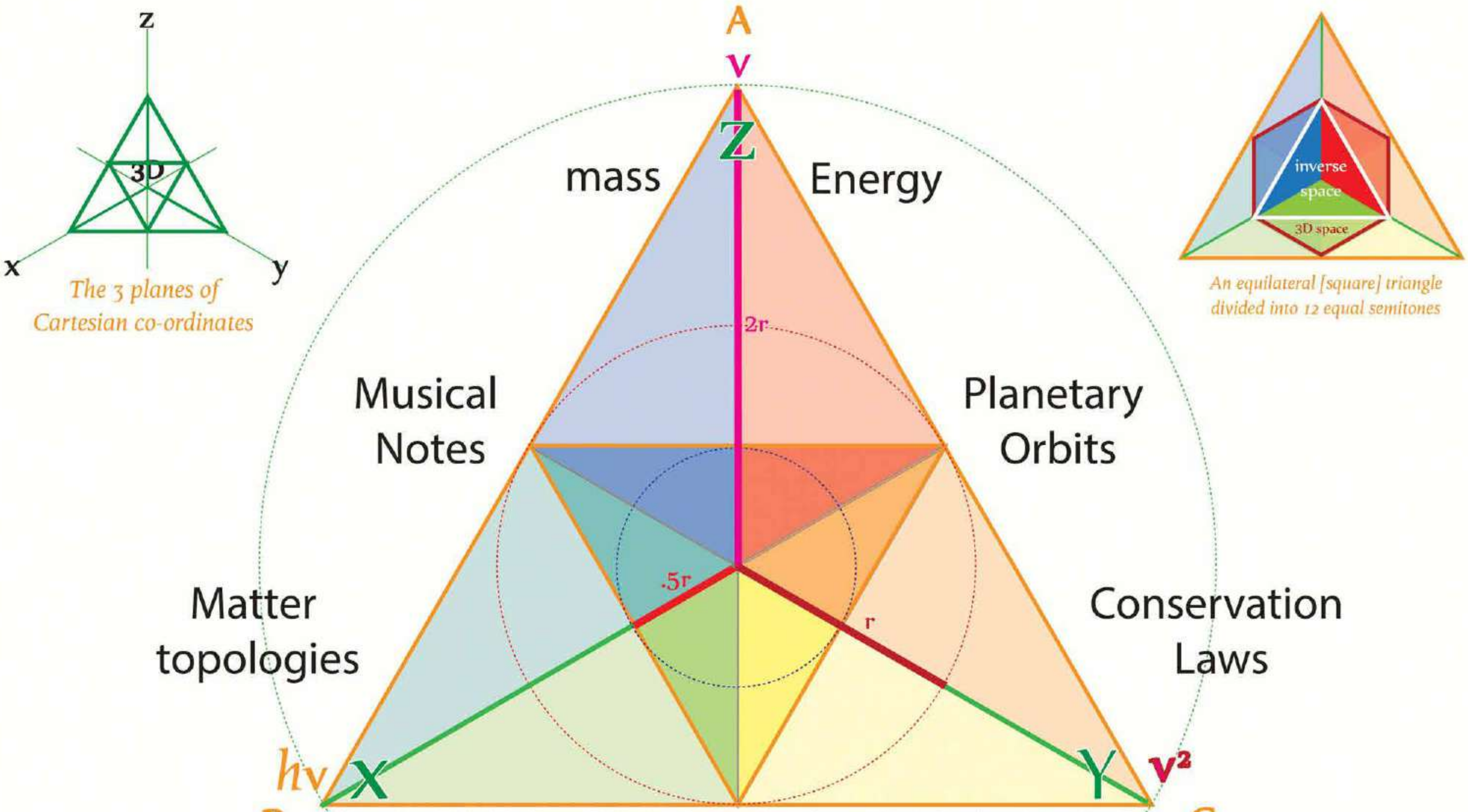
z
3D
x
y

The 3 planes of Cartesian co-ordinates

A
v
Z
mass
Energy
2r
Musical Notes
Planetary Orbits
.5r
r
Matter topologies
Conservation Laws
hv
X
Y
v²
B
C

Pythagoras is credited with the discovery that the intervals between harmonious musical notes always have whole number ratios.

An equilateral [square] triangle divided into 6 equal semitones and 6 equal quarter tones

inverse space
3D space

An equilateral [square] triangle divided into 12 equal semitones

With an Equilateral triangle, draw lines from the center of the circle to each vertex and each midpoint, creating six right triangles, as shown with six different colored triangles.
Each right triangle has a radius of the circle for one leg, and half of a side of the original triangle for another.
Any two right triangles sharing one of the radial lines must therefore be congruent, and that implies that the hypotenuses of the triangles are all equal.
This in turn shows that the six right triangles are all congruent, and so the large triangle they combine to form is equilateral

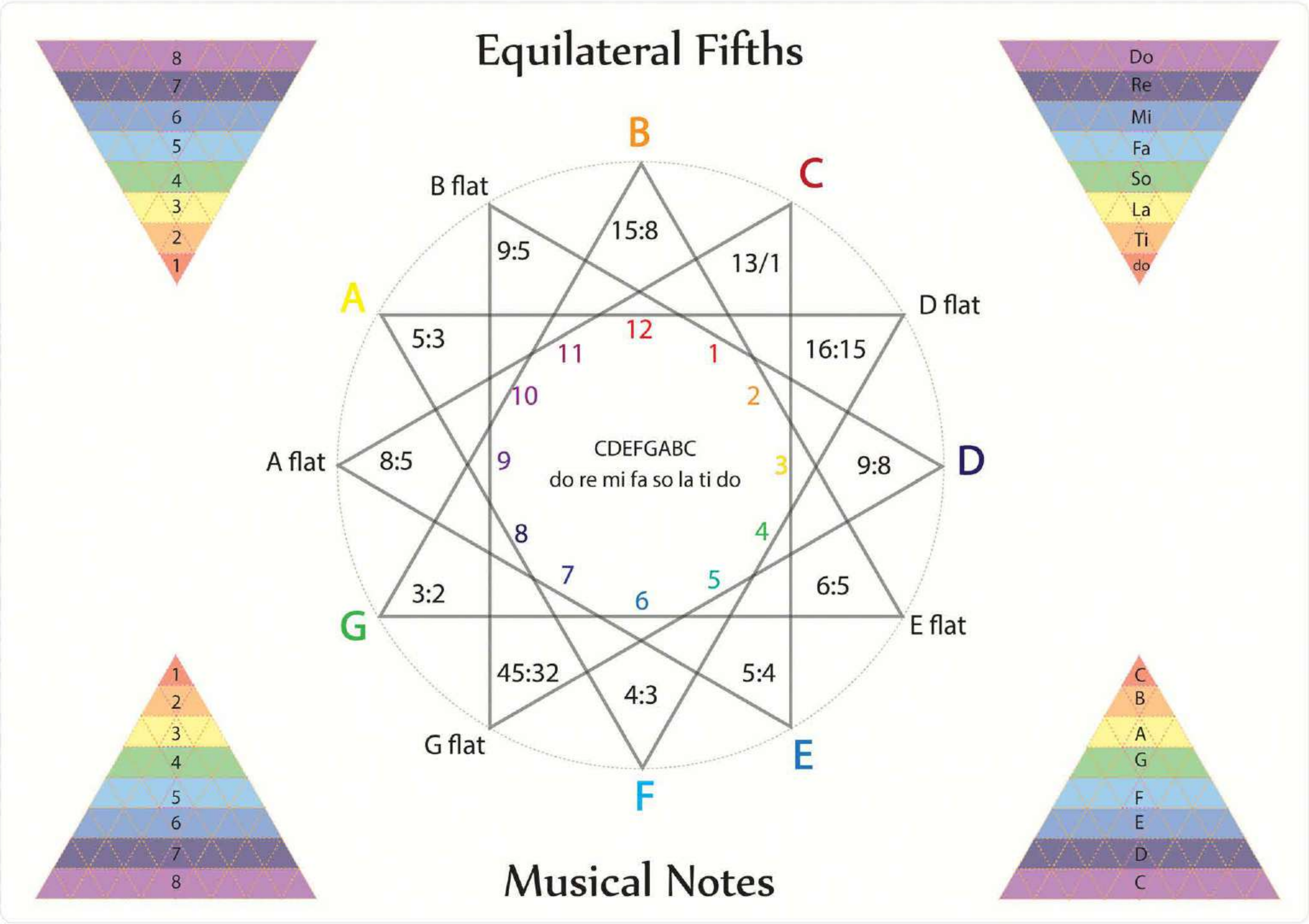
Equilateral Fifths
8
7
6
5
4
3
2
1
Do
Re
Mi
Fa
So
La
Ti
do
B
C
B flat
D flat
A
A flat
D
G
E flat
G flat
E
F
15:8
9:5
13/1
5:3
16:15
8:5
9:8
3:2
6:5
45:32
5:4
4:3
12
1
2
3
4
5
6
7
8
9
10
11
CDEFGABC
do re mi fa so la ti do
1
2
3
4
5
6
7
8
C
B
A
G
F
E
D
C
Musical Notes

Tetryonics 81.14- Equilateral Fifths

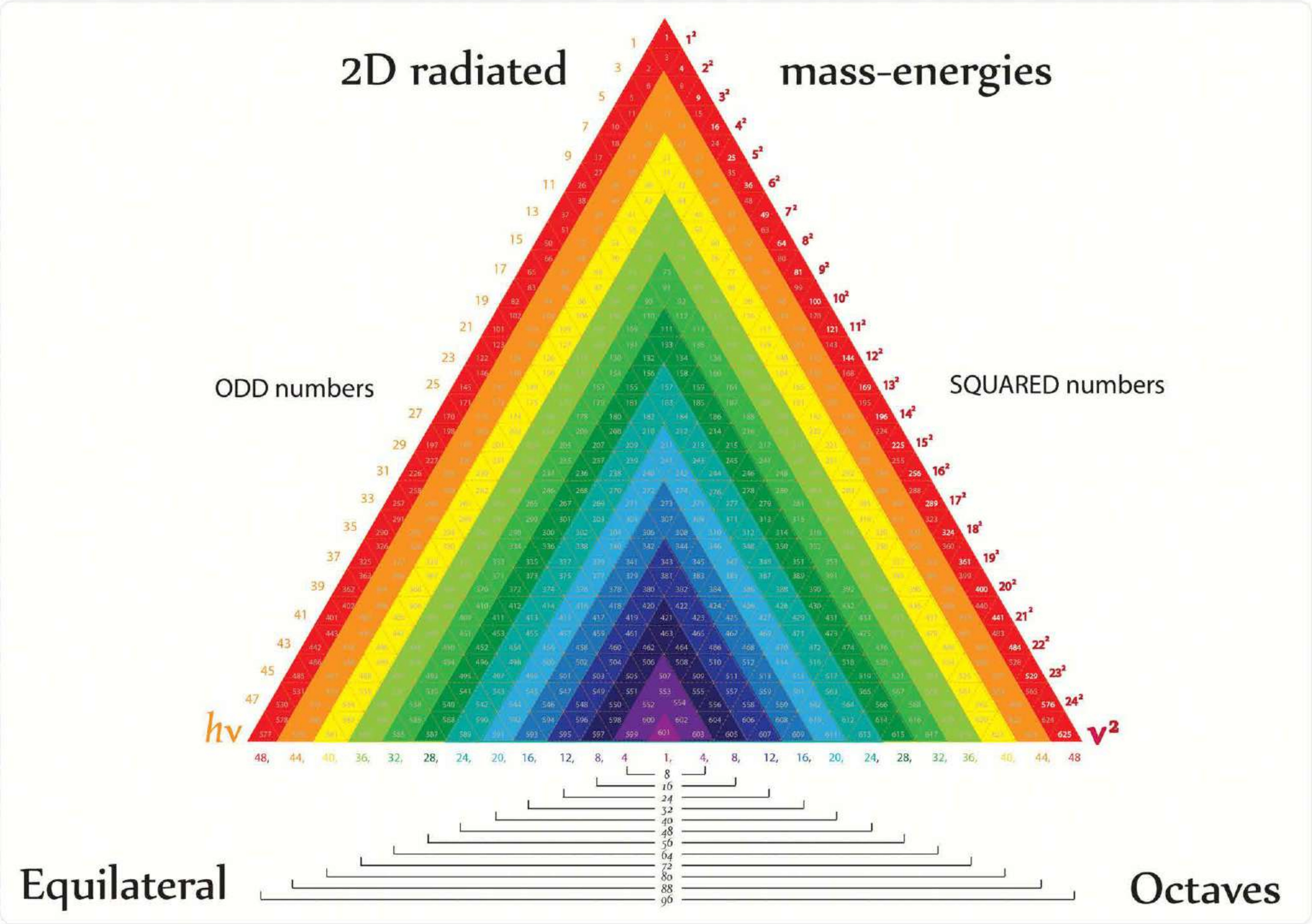
2D radiated
mass-energies
ODD numbers
SQUARED numbers
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
hv
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
9^2
10^2
11^2
12^2
13^2
14^2
15^2
16^2
17^2
18^2
19^2
20^2
21^2
22^2
23^2
24^2
v^2
48, 44, 40, 36, 32, 28, 24, 20, 16, 12, 8, 4 1, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48
8
16
24
32
40
48
56
64
72
80
88
96
Equilateral
Octaves

Tetryonics 81.15 - Equilateral Octaves

Ternary Diagrams

Viviani's Theorem implies that lines parallel to the sides of an equilateral triangle
provide (homogeneous/barycentric/areal/trilinear) coordinates for ternary diagrams
for representing three quantities A,B,C whose sum is a constant
(which can be normalized to unity).

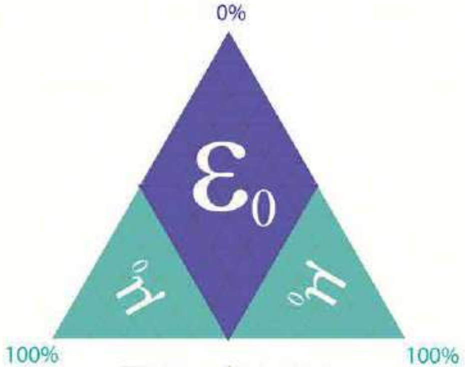
0%
ε_0
μ_0
μ_0
100%
100%

Ternary diagrams
should NOT be used
to model EM field
strengths

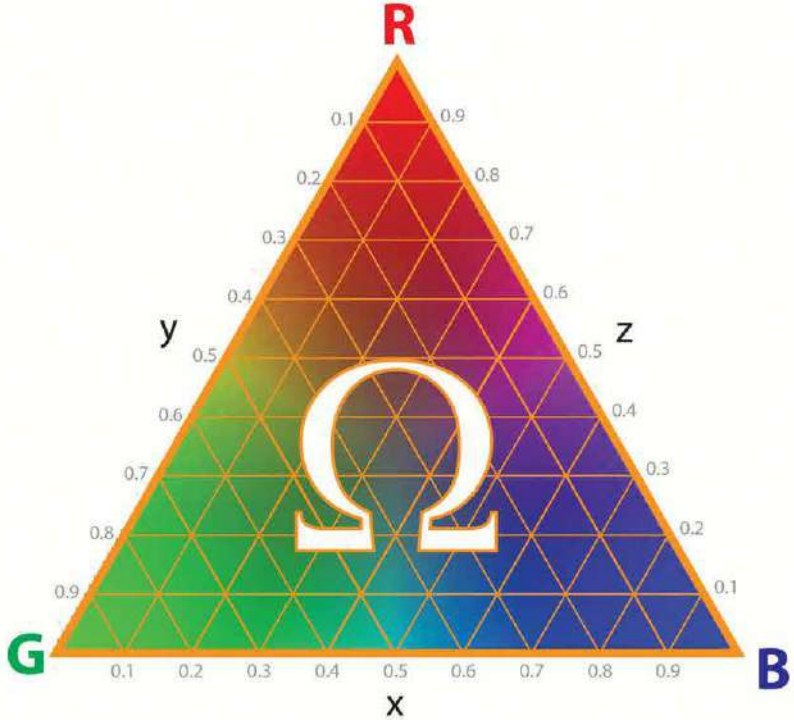
R
G
B
Ω
y
z
x
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

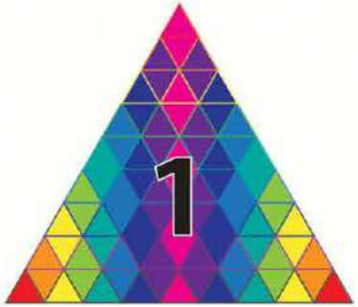

In a ternary plot, of EM energy
the Electric field [E] and the
Magnetic dipole [N-S]
must renormalise
to 1

A ternary diagram is simply a triangular coordinate system in which the 3 edges correspond to the axes.

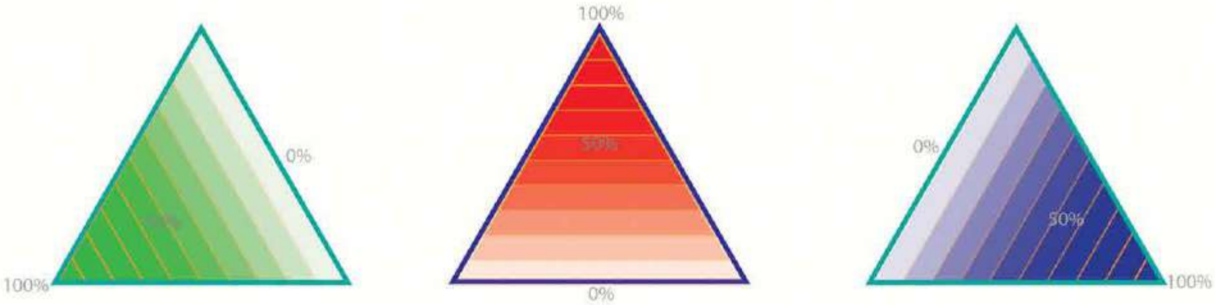
0%
50%
100%
100%
50%
0%
0%
50%
100%

Trilinear charts are commonly used for finding the result of mixing three components (such as gases, chemical compounds, soil, color, etc.) that add to 100% of a quantity.

Whilst the Pythagorean Theorem boasts a slightly greater economy of terms than the Eutrigon Theorem (Wayne Roberts 2003), the latter contains an important area not included in the former:

the area enclosed or swept out by the three points of the triangle in question

$$ab = a^2 + b^2 - c^2$$

the area of any eutrigon is equal to the combined areas of the equilateral triangles on legs 'a' and 'b', minus the area of the equilateral triangle on its hypotenuse 'c'.

The Pythagorean theorem

$$a^2 + b^2 = c^2$$

The square of the hypotenuse of a triangle is equal to the sum of the squares of its sides.

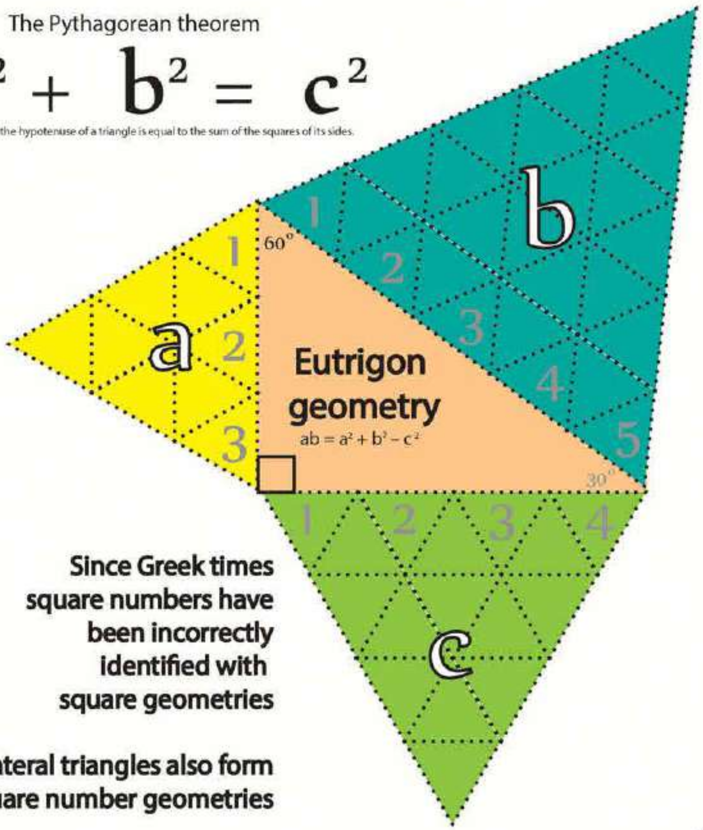

Since Greek times square numbers have been incorrectly identified with square geometries

Equilateral triangles also form square number geometries

Eutrigons

are an important new class of triangle (mathematically defined by Wayne Roberts), as the analogue of the right-triangle in orthogonal (Cartesian) coordinate geometry

Kepler's Second Law of planetary motion

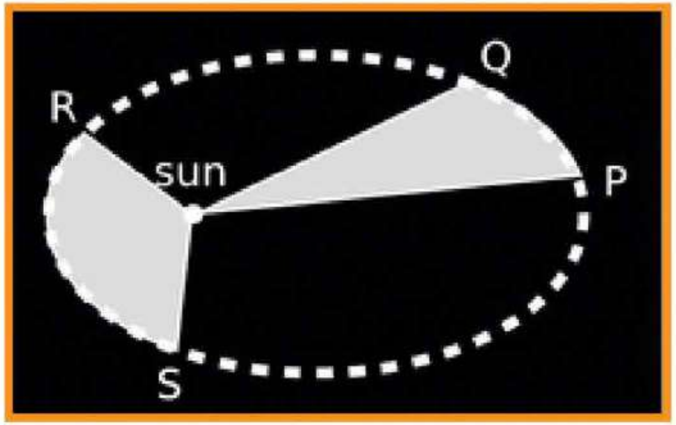

The orbit of every planet is an ellipse with the Sun at one of the two foci.

A line joining a planet and the Sun sweeps out equal areas during equal intervals of time.

The algebraic form of the Eutrigon Theorem, (like the algebraic form of Pythagoras' Theorem), is proven to be special case of the Cosine Rule...

Tetryonic theory reveals the equilateral [square] energy geometry that reveals the 'harmonics at play' in physical laws such as the second law of Kepler, and in many other phenomena in physics, chemistry, cosmology, biochemistry and number theory thus providing the foundation for the mathematics of quantum mechanics

Copyright ABRAHAM [2008] - All rights reserved

Number theory

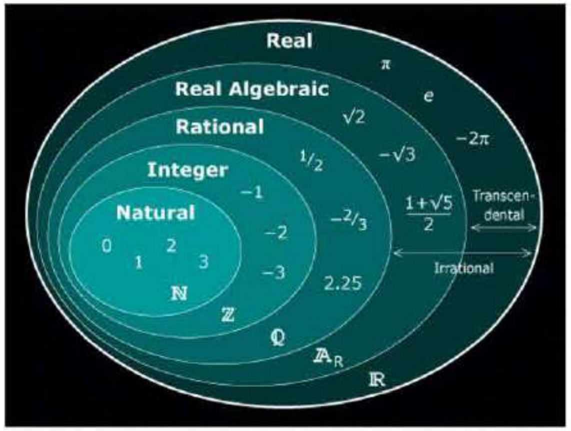
Real
Real Algebraic
Rational
Integer
Natural
0
1
2
3
ℕ
ℤ
ℚ
$\mathbb{A}_\mathbb{R}$
ℝ
π
e
√2
−2π
−√3
1/2
−1
−2
−2/3
−3
2.25
$\frac{1+\sqrt{5}}{2}$
Transcendental
Irrational

The Pythagoreans also established the foundations of number theory, with their investigations of triangular, square and also perfect numbers (numbers that are the sum of their divisors).

They discovered several new properties of square numbers, such as that the square of a number n is equal to the sum of the first n odd numbers (e.g. 1 + 3 + 5 + 7 = 16).

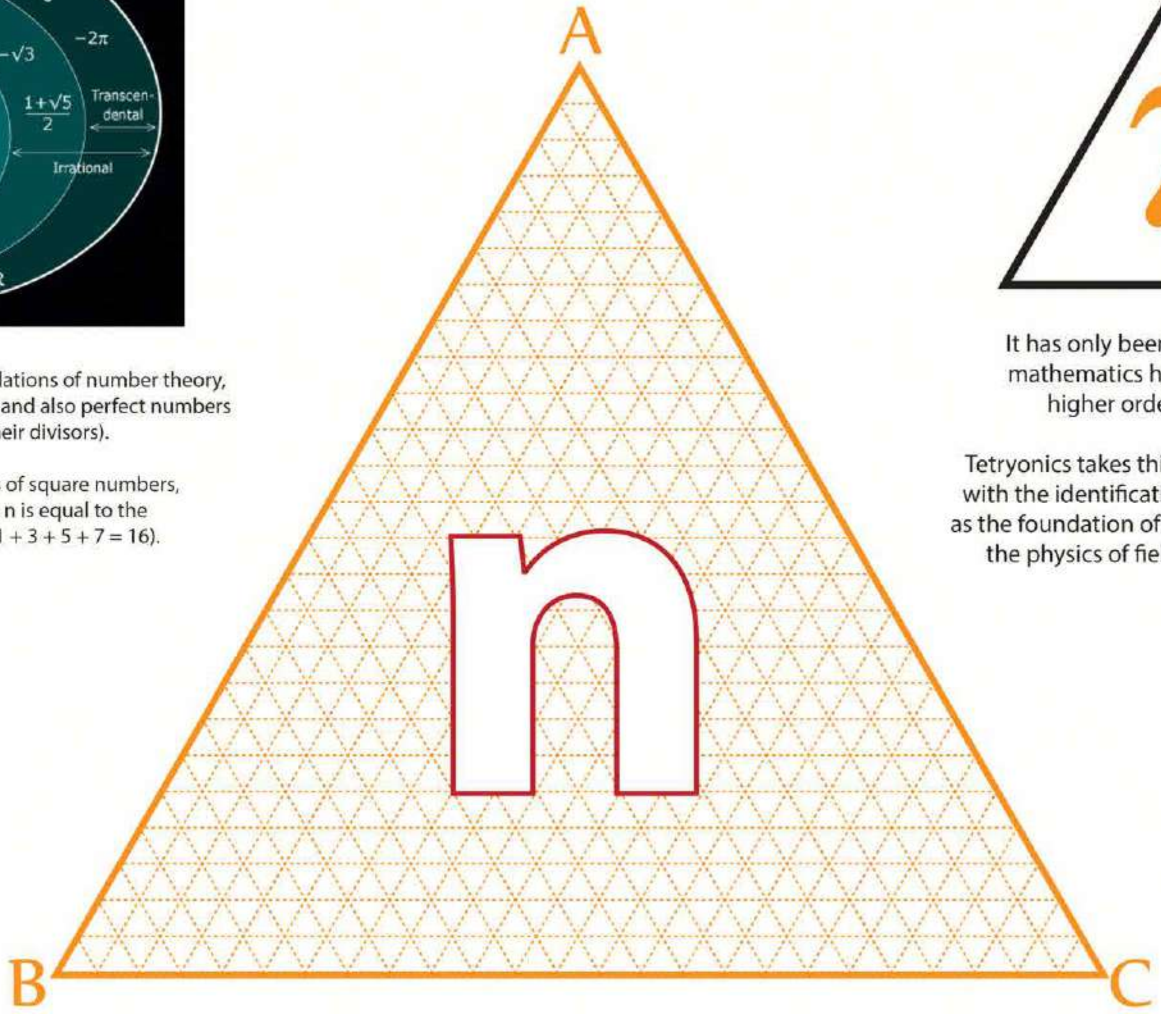

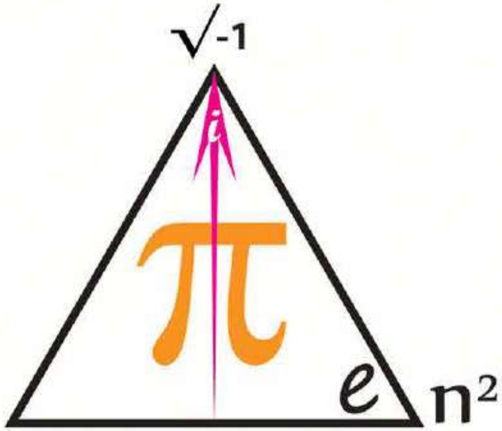

It has only been in recent centuries that mathematics has begun to explore the higher order irrational numbers

Tetryonics takes this investigation to new levels with the identification of equilateral geometries as the foundation of transcendental numbers and the physics of fields and particles in motion

What mathematics has failed to appreciate is the significance application of equilateral geometries to the 'square' numbers of physics and science in general

Copyright ABRAHAM [2008] - All rights reserved

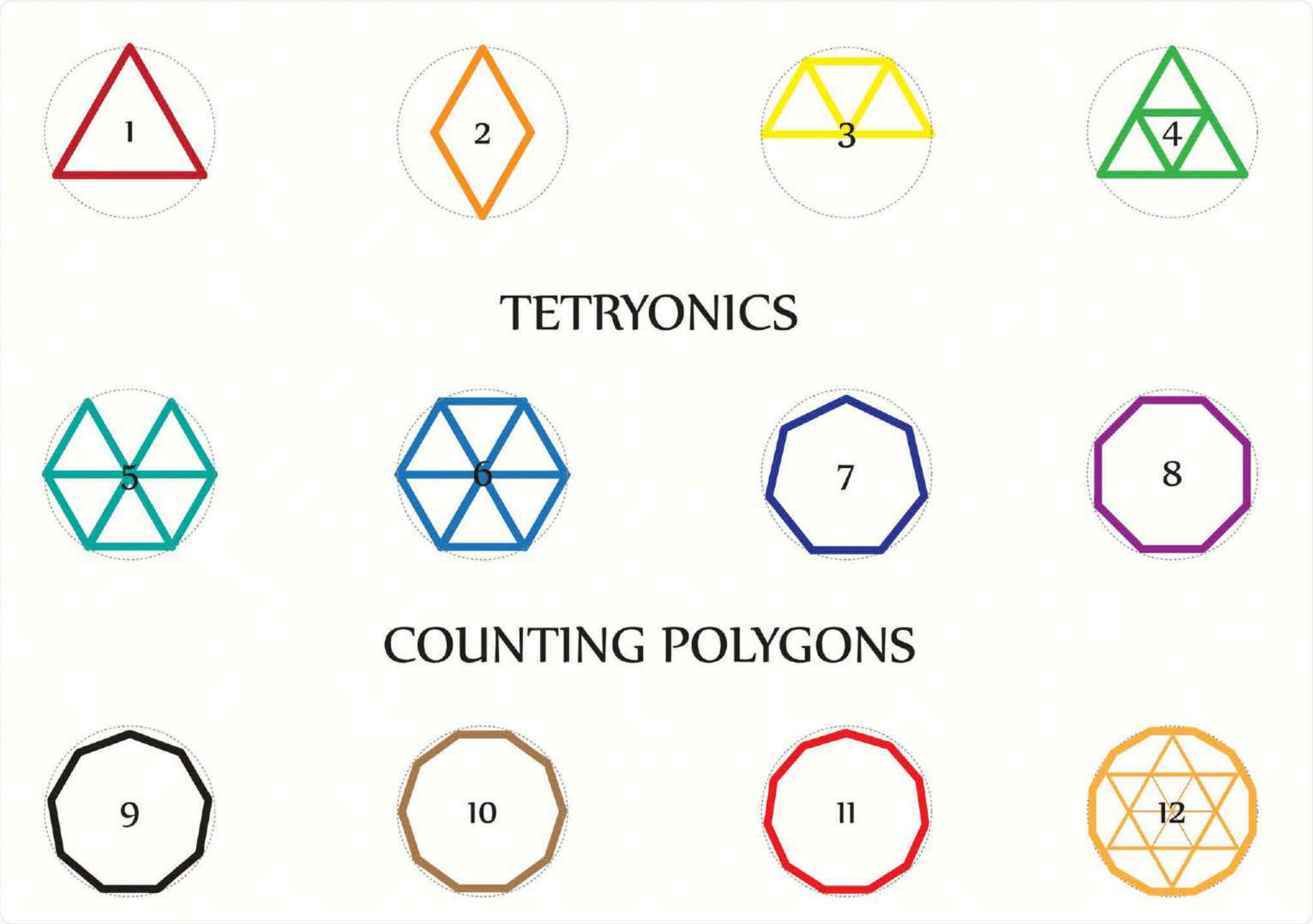

TETRYONICS

COUNTING POLYGONS

1
2
3
4
5
6
7
8
9
10
11
12

Tetryonics 82.02 - Counting Polygons

Copyright ABRAHAM [2008] - All rights reserved

SQUARED energies in quantum mechanics are EQUILATERAL geometries

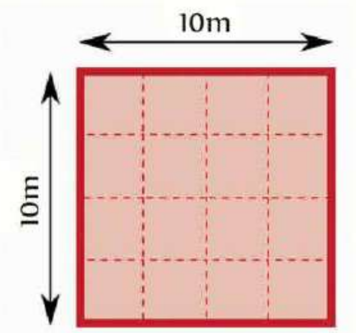

Square

area = s^2 = [100]

Circles

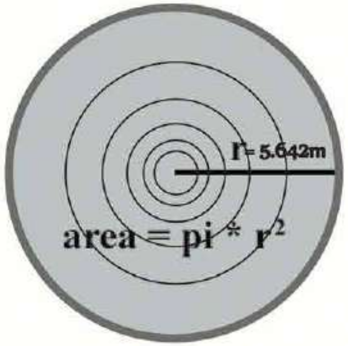
r = 5.642m
area = pi * r^2

= pi *$[5.642]^2$
= 100

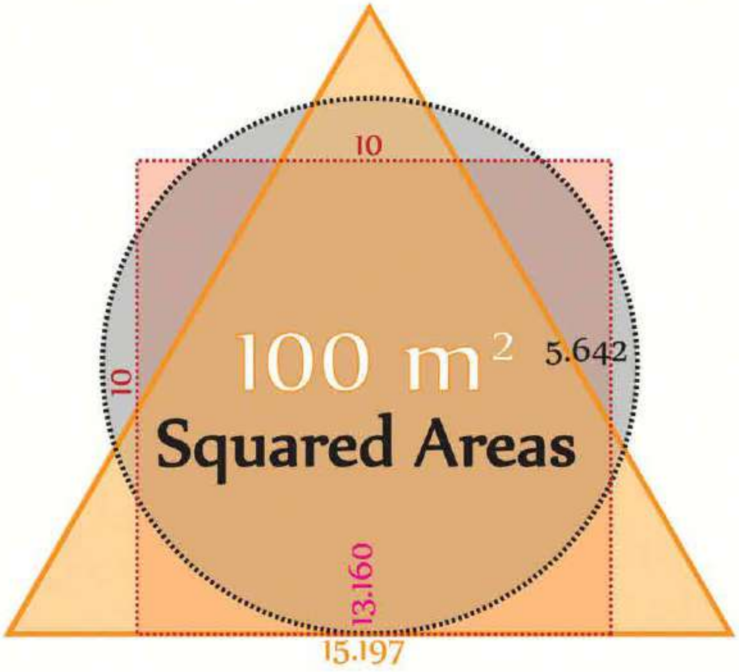
10
10
5.642
100 m²
Squared Areas
13.160
15.197

can be created by a number of planar geometries

For a long time it has been assumed by scientists (and mathematicians)
that circular [and squared] geometries are the geometric foundation of all physics,
leading to a serioulsy flawed model of particles and forces in quantum mechanics

Tetryonic theory now reveals that quantised equilateral angular momenta creates the foundational geometry of all the mass-Energy-Matter & forces of physics

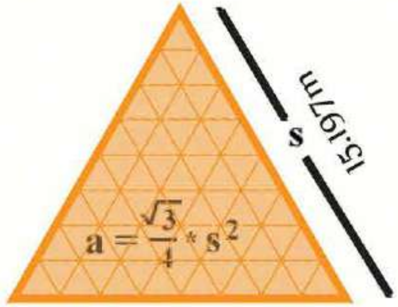
s
15.197m
$a = \frac{\sqrt{3}}{4} * s^2$

Equilateral

area = $(\frac{1}{2}*b)*h$

Triangles

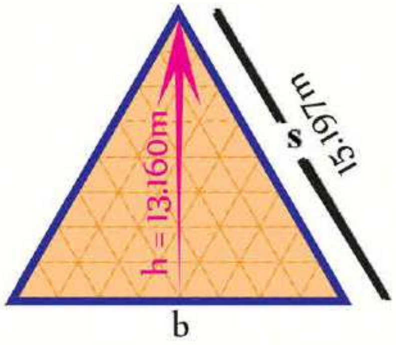
h = 13.160m
s
15.197m
b

b h
[.5x15.197] x 13.160
= 100

Copyright ABRAHAM [2008] - All rights reserved

0.987654321

SQUARE ROOTS

80/81

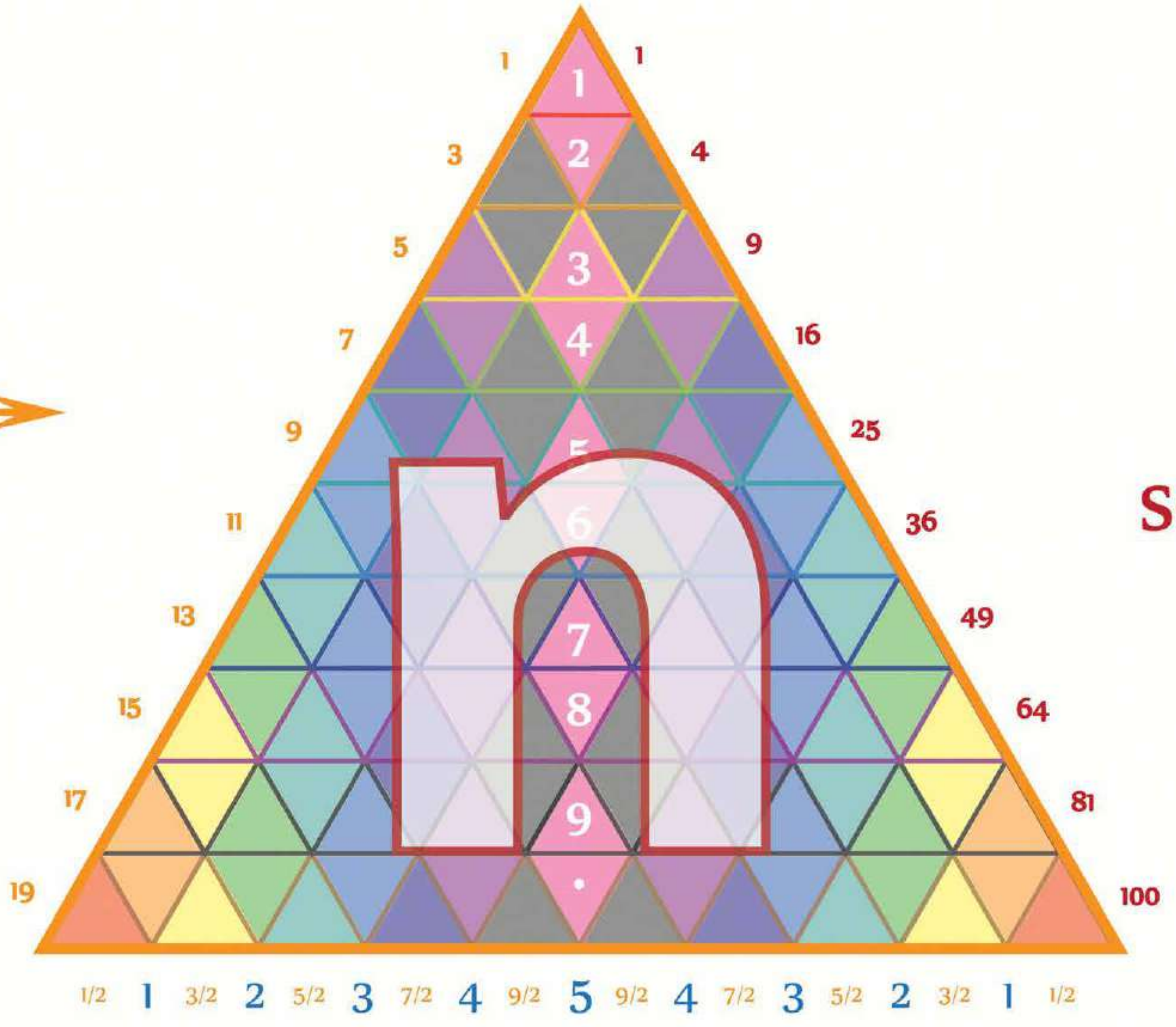
1
3
5
7
9
11
13
15
17
19
1
4
9
16
25
36
49
64
81
100
1
2
3
4
5
6
7
8
9
•
n
1/2 1 3/2 2 5/2 3 7/2 4 9/2 5 9/2 4 7/2 3 5/2 2 3/2 1 1/2

ODDS

SQUARES

EVEN DISTRIBUTIONS

Integers

The integers (from the Latin integer), literally "untouched", hence "whole"
in Tetryonics it is the basis for the Planck charge quantum

$$2n-1$$

$$[[n] + [n-1]]$$

Triangular numbered geometries are NOT equialteral geometries

0, 1, 3, 6, 10, 15, 21, 28, 36, 45, 55, ...

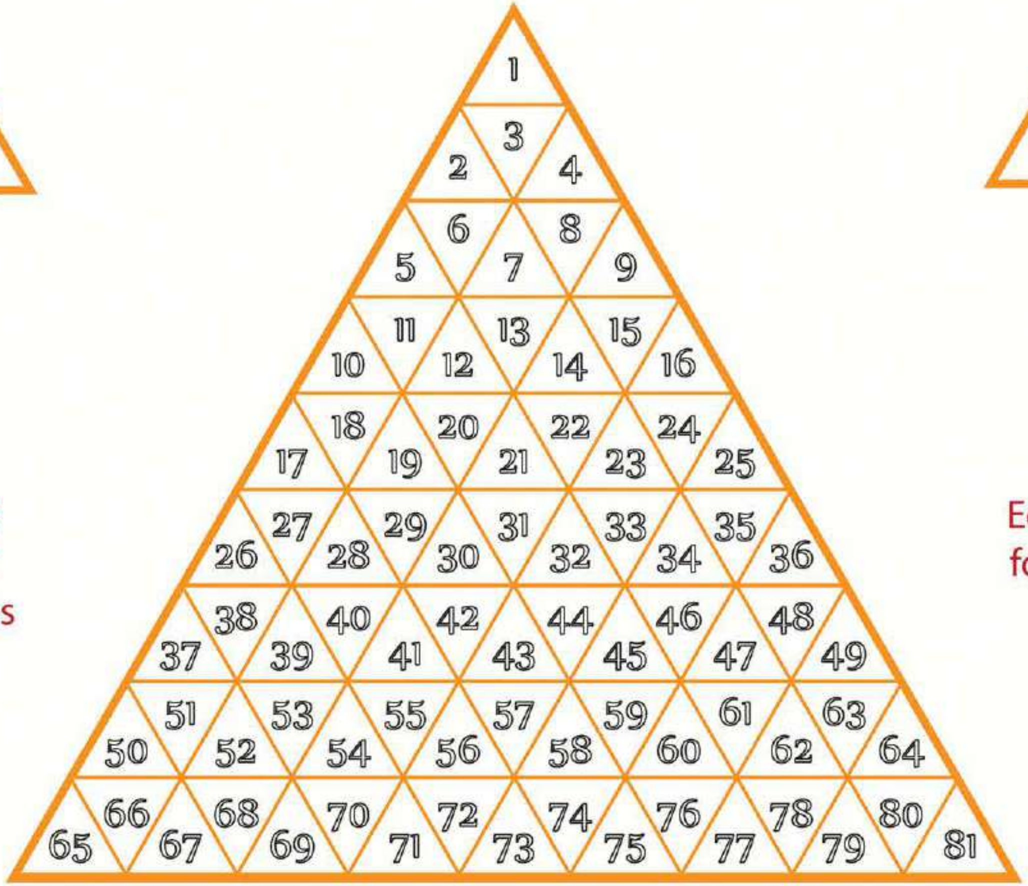
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

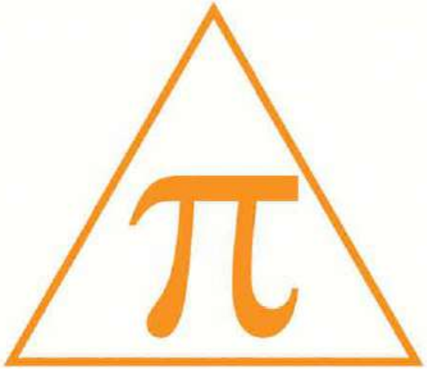

$$\sum_{1}^{n} [2n-1]$$

Equilateral geometries form **SQUARE** number geometries

1, 4, 9, 16, 25, 49, 64, 81, 100, 121, 144, ...

1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1

Equilateral energy quanta form a normal longitudinal distribution

Viewed as a subset of the real numbers, they are numbers that can be written without a fractional or decimal component

ε_0
π
μ_0
μ_0

Bosons are a transverse measure of scalar energy momenta

E
μ_0
n
3
μ_0
E
ε_0
ε_0
+
−
N
S
S
N

ODD numbers

An odd number is an integer which is not a multiple of two.

Quantum

Bosons have ODD number quanta

Quantum Levels

hv

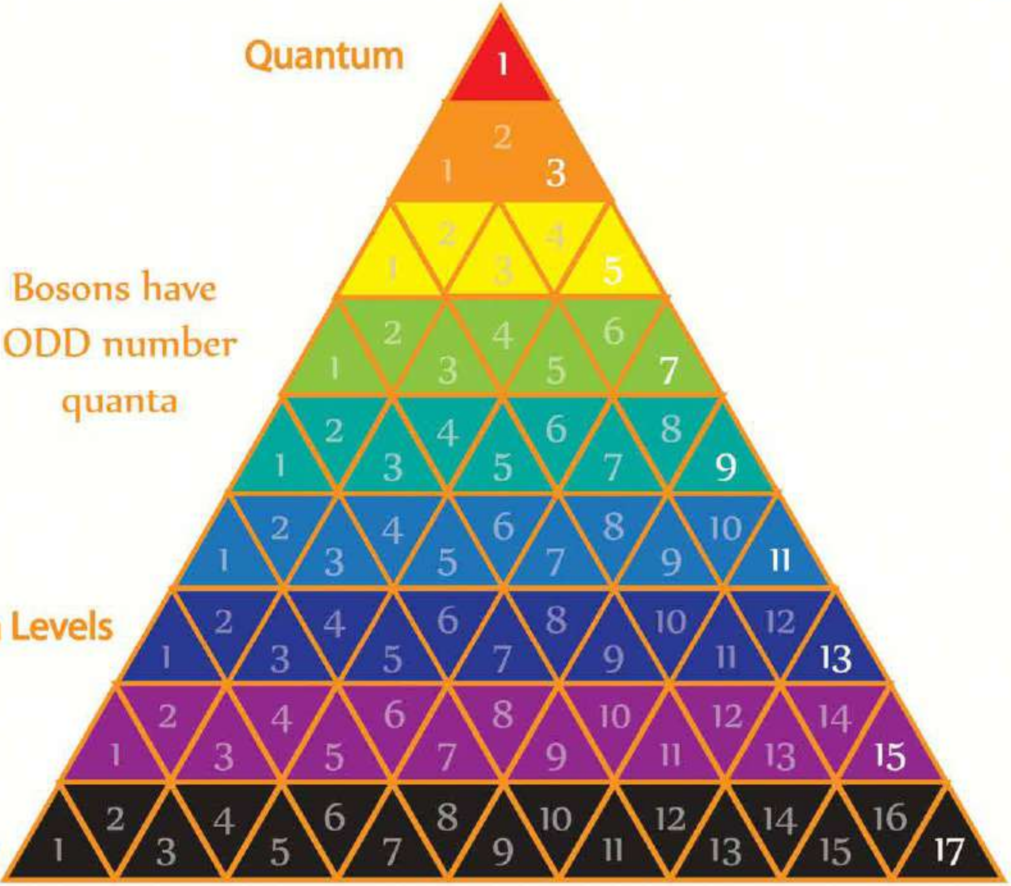

1
1 2 3
1 2 3 4 5
1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9 10 11
1 2 3 4 5 6 7 8 9 10 11 12 13
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

1

3

5

7

9

11

13

15

17

An odd number, when divided by two, will result in a fraction

Tau radians

Around the whole outside of a circle,
there are about 6.283 radians - or, Tau radians

3.141592654

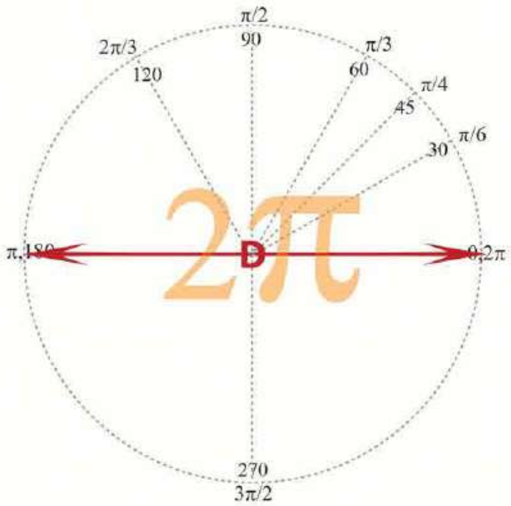

π/2
90
2π/3
120
π/3
60
π/4
45
π/6
30
π,180
0,2π
2π
D
270
3π/2

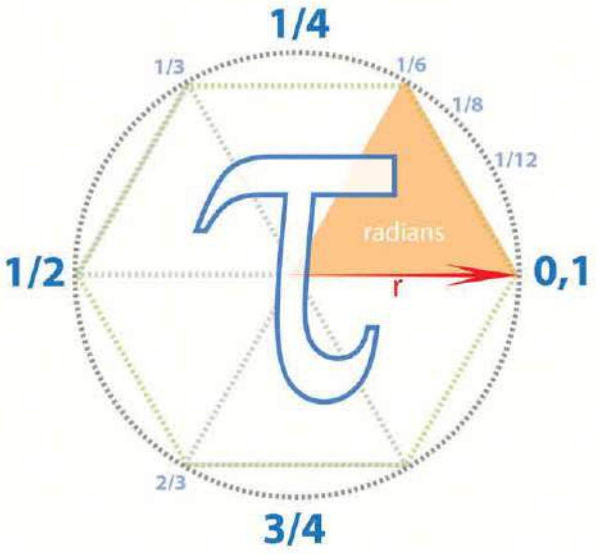

1/4
1/3
1/6
1/8
1/12
1/2
0,1
τ
radians
r
2/3
3/4

6.283185307

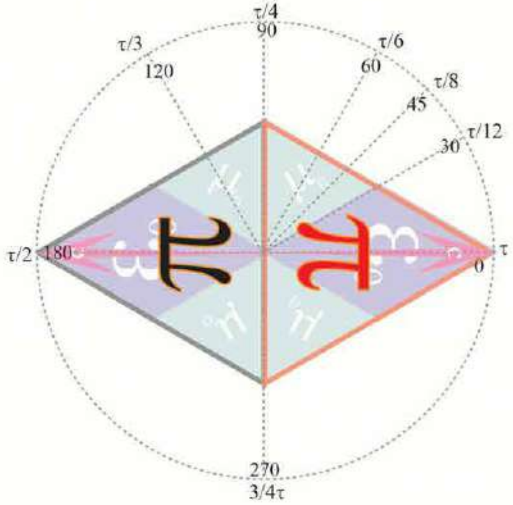

τ/4
90
τ/3
120
τ/6
60
τ/8
45
τ/12
30
τ/2
180
τ
0
270
3/4τ

τ is a more 'natural' radian system for geometric physics than π
Tau = 2π = 360 degree rotation about a point

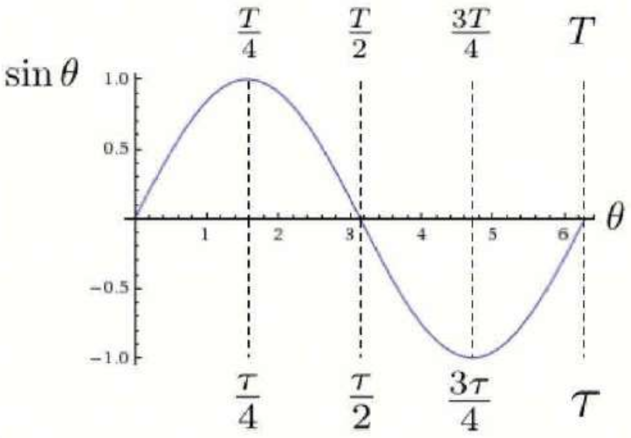

π
historically defined as the
ratio of a circle's circumference to its DIAMETER
should be redefined in physics to

τ
the ratio of its circumference to its RADIUS

in doing so many of the π/2 terms common to physics
will re automatically rationalised and will better reflect
the Tetryonic geometry of mass-ENERGY-Matter in motion

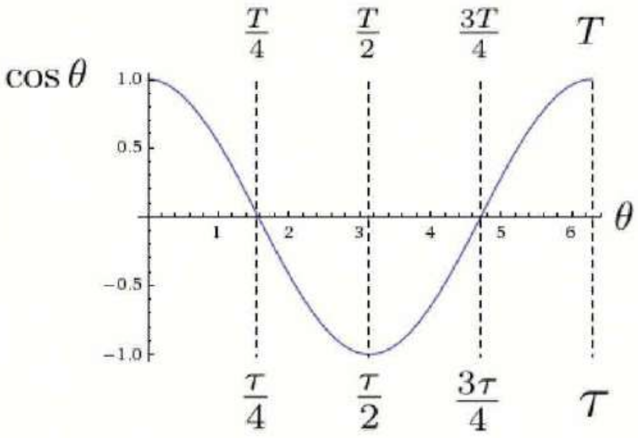

Copyright ABRAHAM [2008] - All rights reserved

Photons are a longitudinal measure of scalar energy momenta

E
ε_0
+
2
N
N
S
S
−
ε_0
E

Photons are
EVEN number
quantums

EVEN numbers

An integer that is not an odd number is an even number

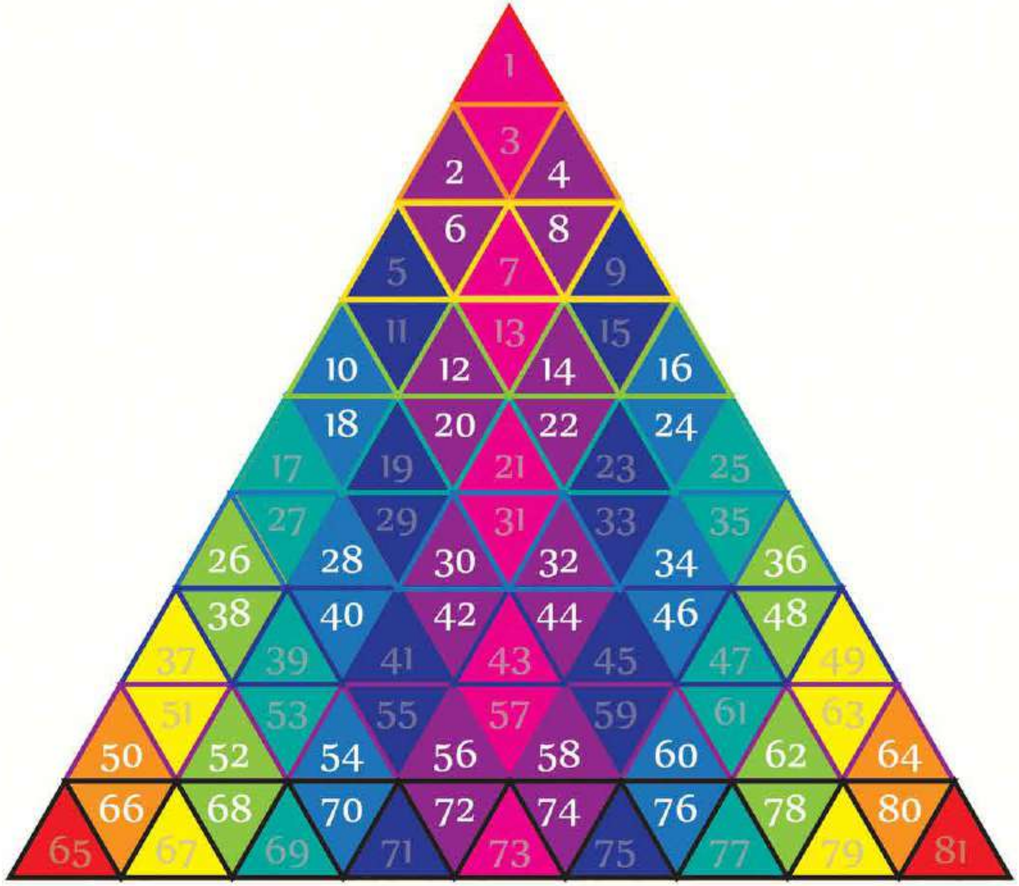
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

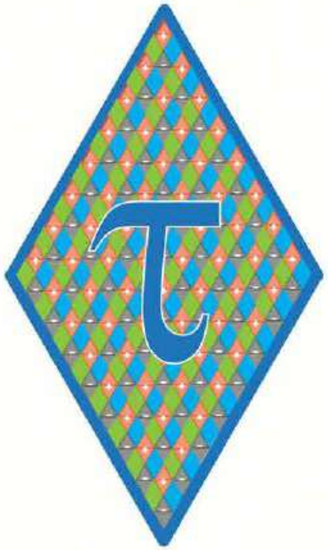

EM waves are
comprised of EVEN
numbered quanta

If an even number is divided by two, the result is another whole number

1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

If an odd number is divided by two, the result is a fractional number

Copyright ABRAHAM [2008] - All rights reserved

Triangular numbers

Historically, a triangular number counts quanta that can pack together to form an equilateral triangle

1, 3, 6, 10, 15, 21, 28, 36, 45, 55, ...

this form of geometric counting of same charges over-complicated the simpler physical reality

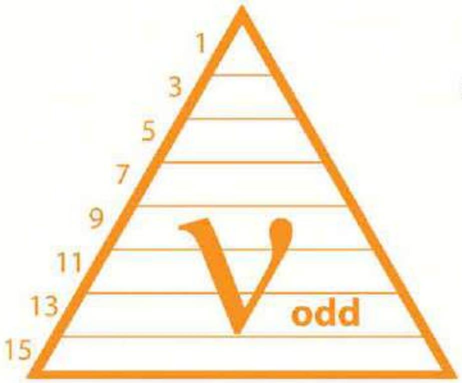
1
3
5
7
9
11
13
15
ν
odd

$$2\nu - 1$$

$$[[\nu] + [\nu-1]]$$

Equilateral chords or quantum levels are ODD numbers

1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21,....

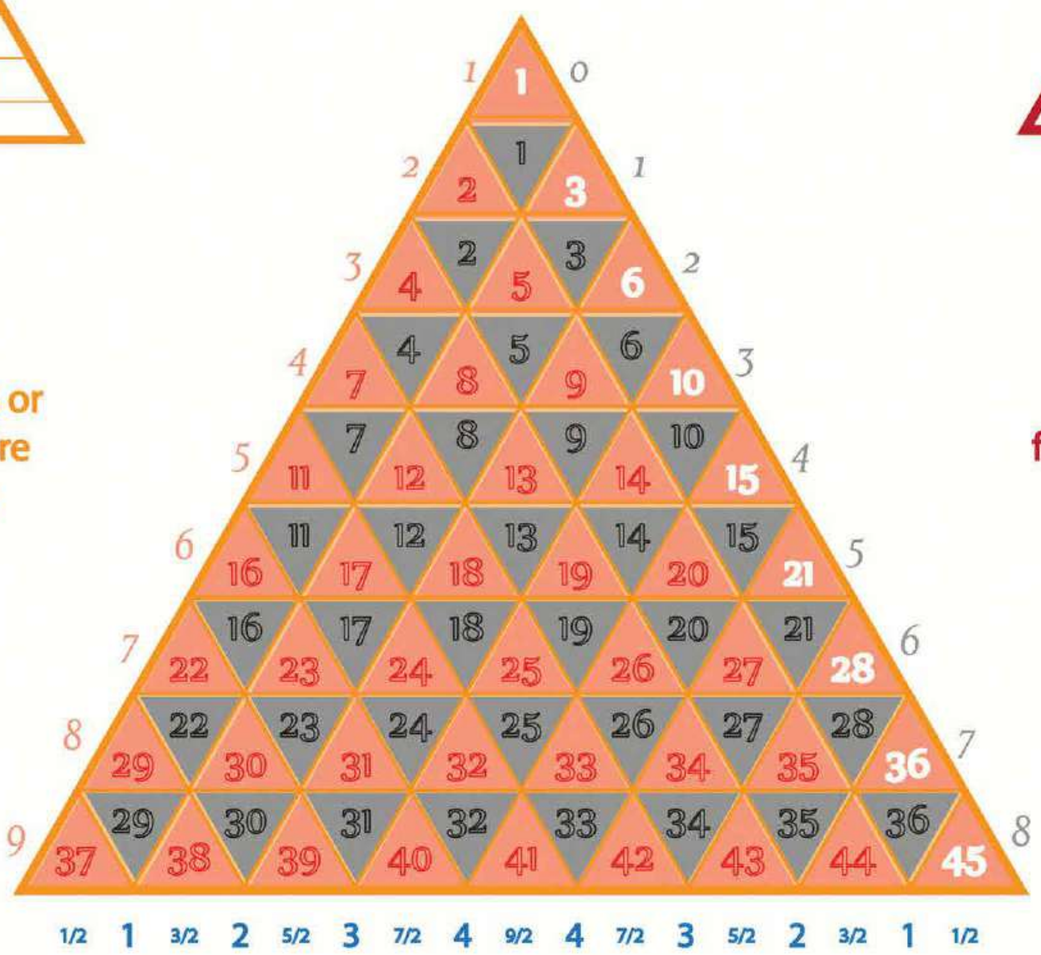
1 1 0
2 1 2 3 1
3 2 3 4 5 6 2
4 4 5 6 7 8 9 10 3
5 7 8 9 10 11 12 13 14 15 4
6 11 12 13 14 15 16 17 18 19 20 21 5
7 16 17 18 19 20 21 22 23 24 25 26 27 28 6
8 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 7
9 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 8
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

Triangular energy quanta form normal distributions

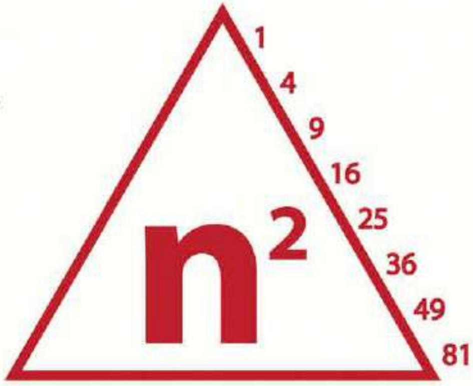
1
4
9
16
25
36
49
81
n²

$$\sum_{1}^{n} [2n-1]$$

Equilateral geometries form SQUARE numbered geometries

1, 4, 9, 16, 25, 49, 64, 81, 100, 121, 144, ...

Copyright ABRAHAM [2008] - All rights reserved

Squared numbers

A square number, sometimes also called a perfect square,
is the result of an integer multiplied by itself

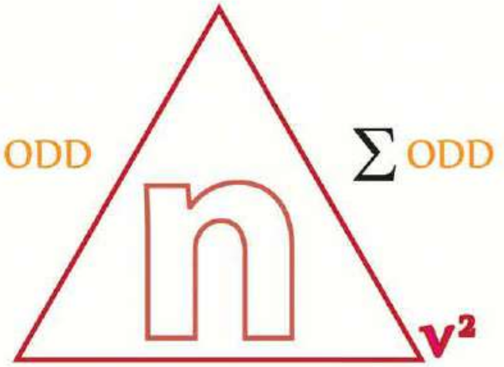
ODD
$\sum$ ODD
n
v^2

1
+
3
+
5
+
7
+
9
+
11
+
13
+
15
+
17

Square numbers result from
the summing of consecutive
ODD numbers

ODDS

2n-1

[[n] + [n-1]]

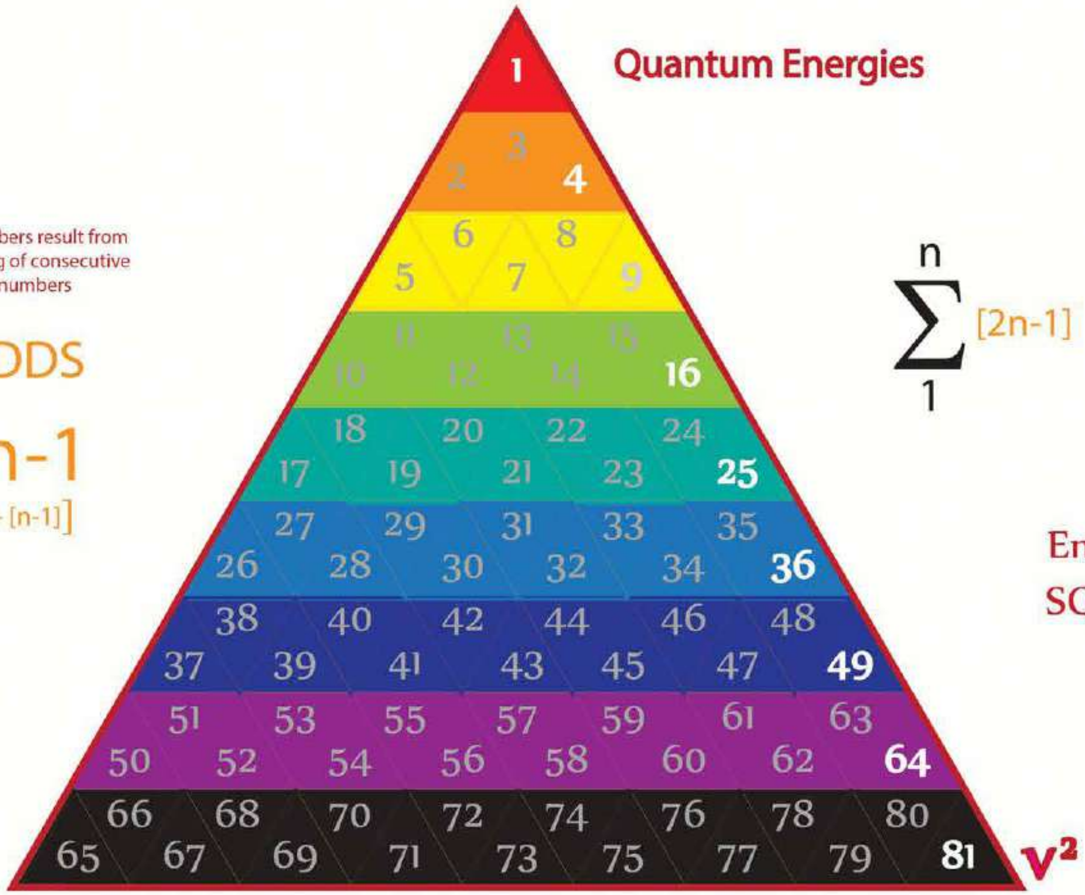
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
v^2

Quantum Energies

$$\sum_{1}^{n} [2n-1]$$

Energy levels have
SQUARE number
quanta

Compton Frequency

In Tetryonics Square numbers produce equilateral geometries

Copyright ABRAHAM [2008] - All rights reserved

Square roots

A square root of a number is a number that, when it is multiplied by itself (squared) , gives the first number again.

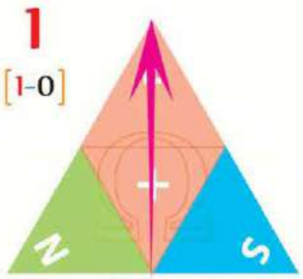

Root of positive one

-*i* and +*i*

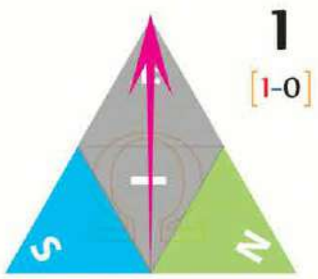

Root of negative one

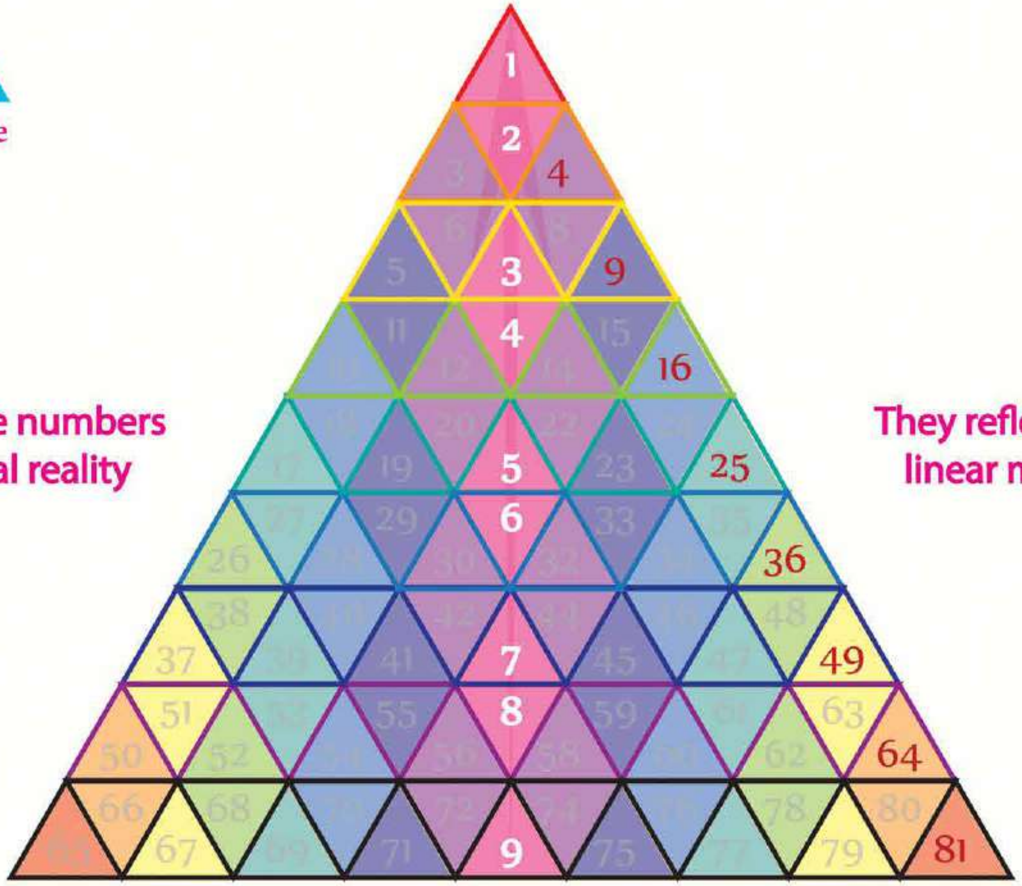

Square roots of negative numbers have a basis in physical reality

They reflect the real non-negative linear momentum of a system

A whole number with a square root that is also a whole number is called a perfect square in Tetryonic theory they are actually equilateral geometries

Copyright ABRAHAM [2008] - All rights reserved

Real Numbers

In mathematics, a real number is a value that represents a quantity along a continuous line. The real numbers include all the rational numbers,

Quantum levels

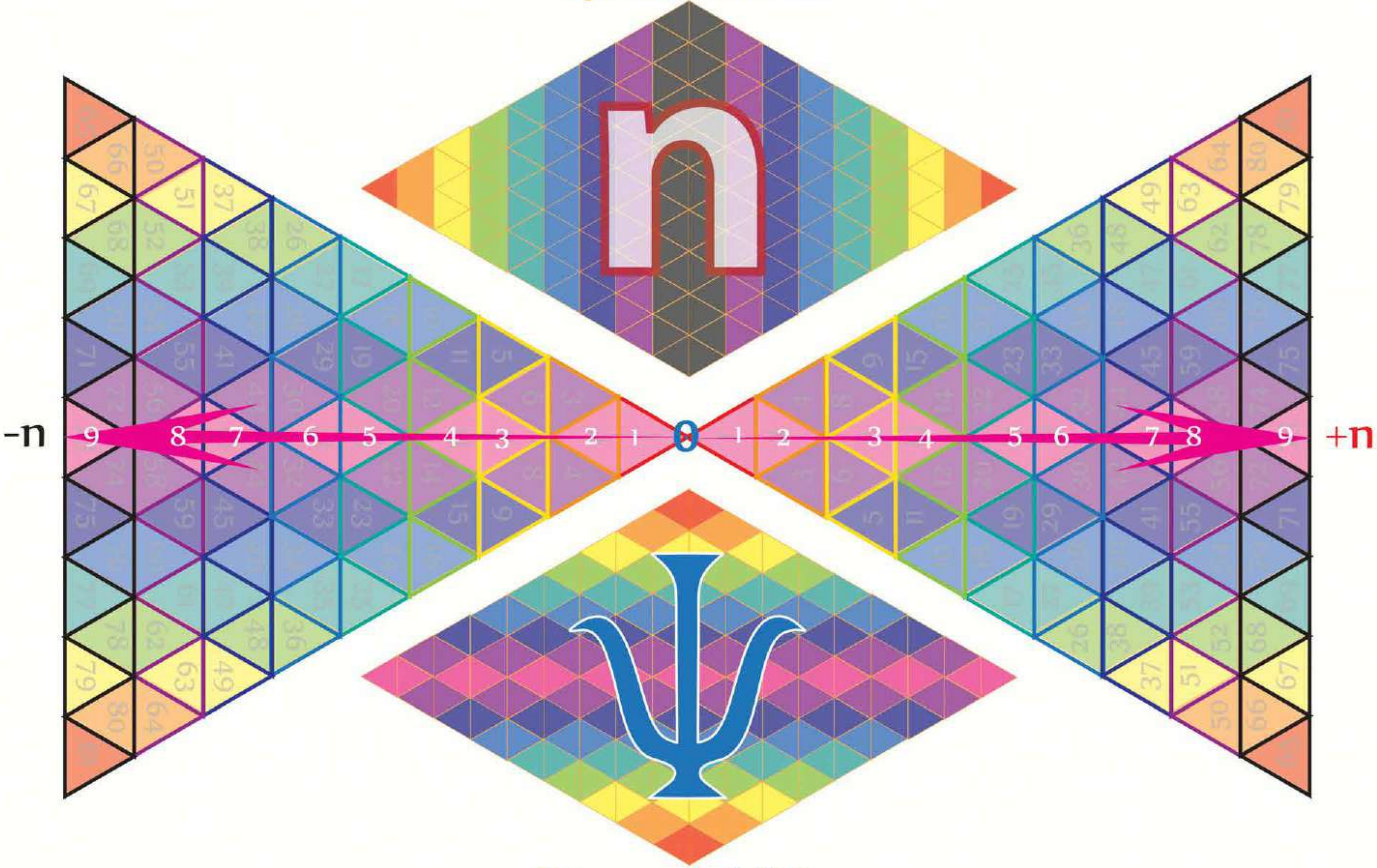

Wave probabilities

Copyright ABRAHAM [2008] - All rights reserved

ONE to ∞^2

Quantum levels

ODD numbers

[2n-1]

1, 3, 5, 7, 9......

SQUARE numbers

the sum of consecutive ODD numbers
1+3+5+7+9+......

Scalar Energies

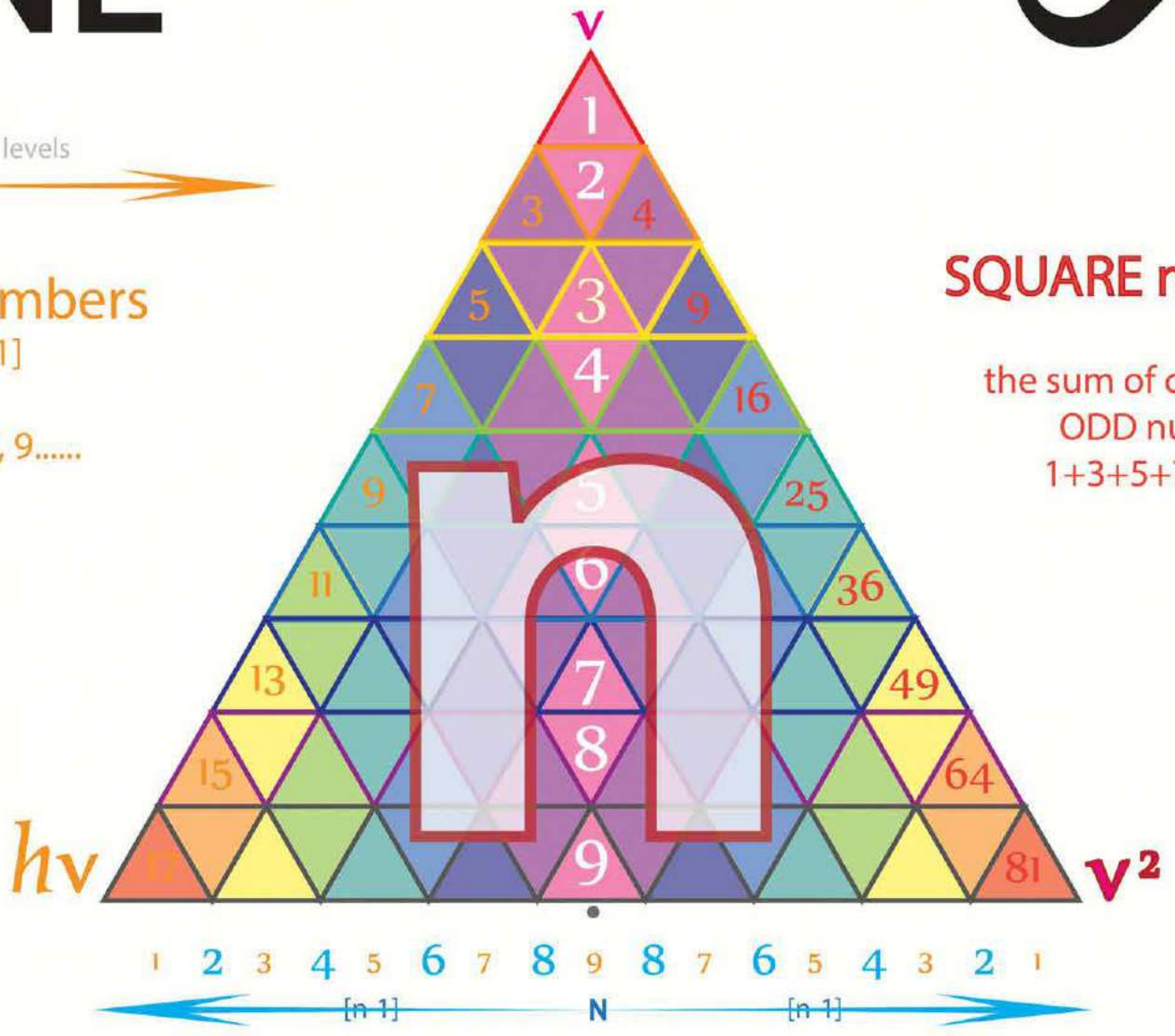
v
1
2
3
4
3
5
9
4
7
16
5
9
25
6
11
36
7
13
49
8
15
64
9
hv
17
81
v²
n
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1
[n-1]
N
[n-1]

Linear energy momentum

NORMAL DISTRIBUTION

Tetryonics 82.13 - One to Infinity squared

Copyright ABRAHAM [2008] - All rights reserved

Basic Properties of nested scribed Equilateral Triangles

Given an equilateral triangle of side s

altitude

$\sqrt{3/2}\,S$

$3S$

perimeter

$\sqrt{3/6}\,S$

in-radius

circum-radius

$\sqrt{3/3}\,S$

$\sqrt{3/4}\,S^2$

area

$\sqrt{\pi/12}\,S^2$

in-circle area

circum-circle area

$\sqrt{\pi/3}\,S^2$

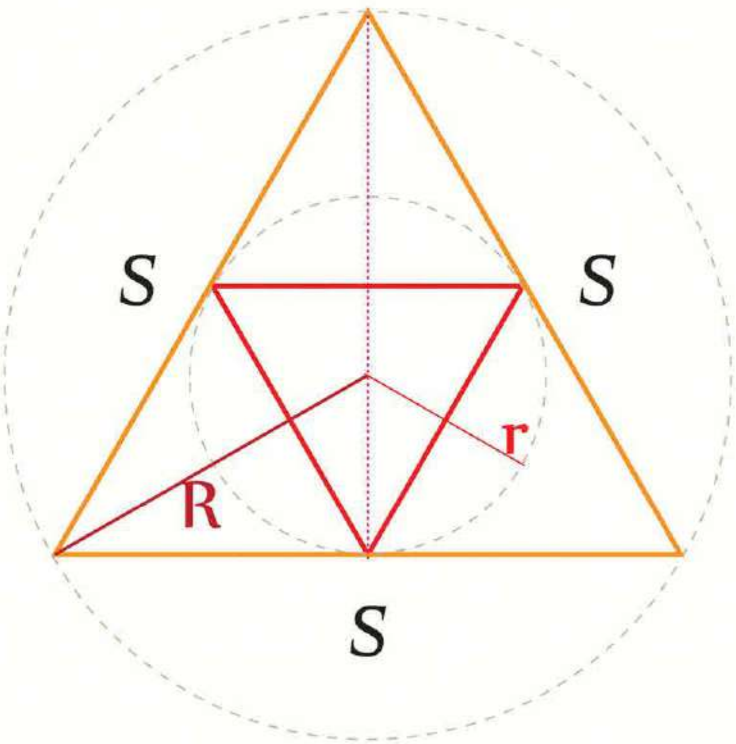

Tetryonic [equilateral] geometry

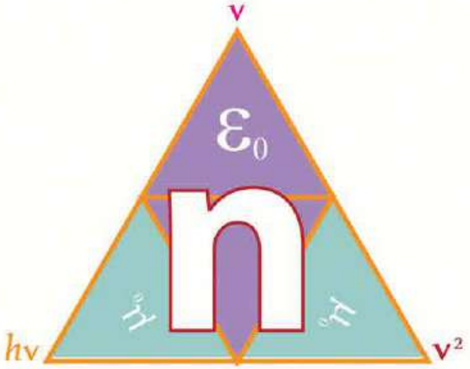

The equilateral triangle exhibits 'square symmetry'
it can always be divided into n^2 [number]
of smaller self-similar parts

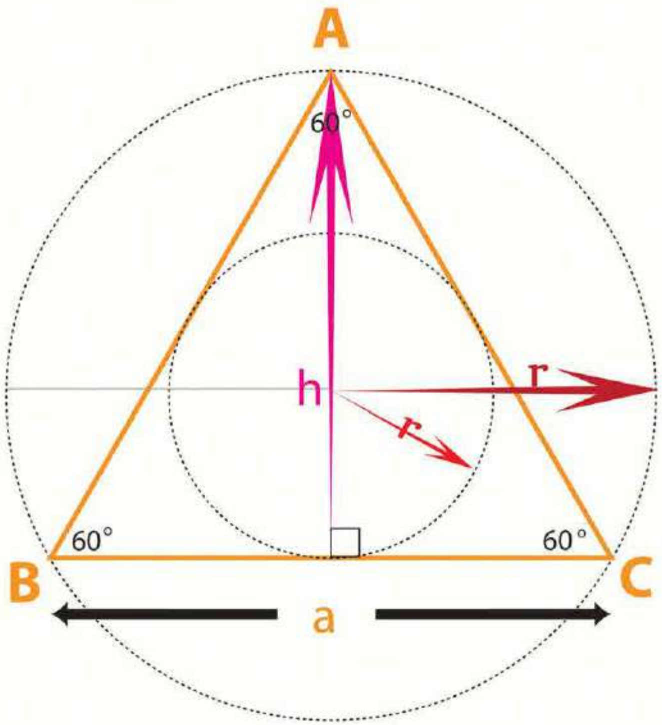

All triangles are
flat euclidean
π radian
geometries
[180°]

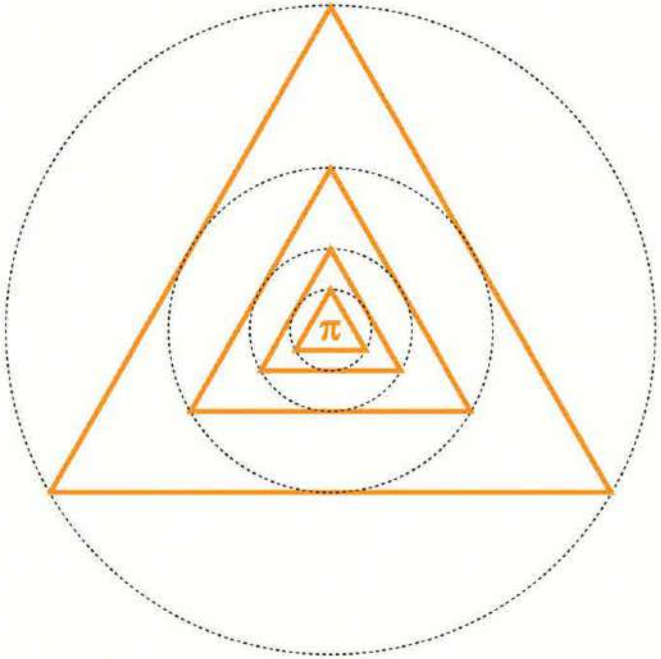

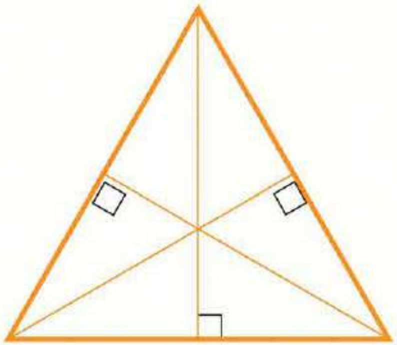

An equilateral triangle is the most symmetrical triangle,
having 3 lines of reflection and rotational symmetry of order 3 about its center

CHARGE

Scribed equilateral geometries

reflect space-time's geometric relationship with charged mass-energy

Energy

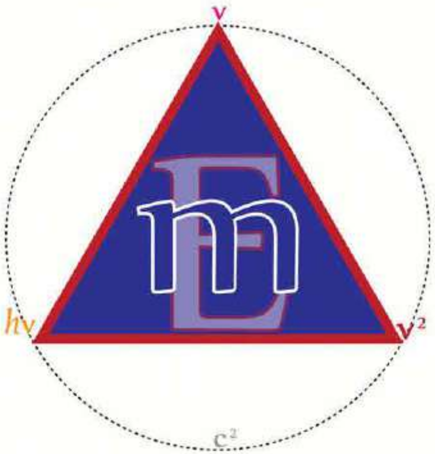

Energy per spatial unit

mass

$$m = \frac{E}{c^2}$$

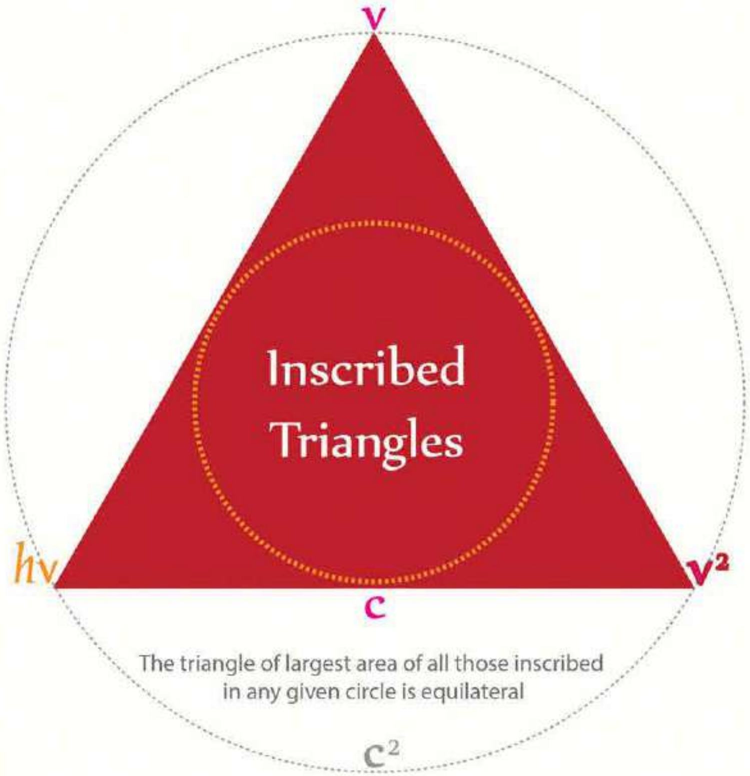
v
Inscribed Triangles
hv
v²
c
c²

The triangle of largest area of all those inscribed in any given circle is equilateral

QAM

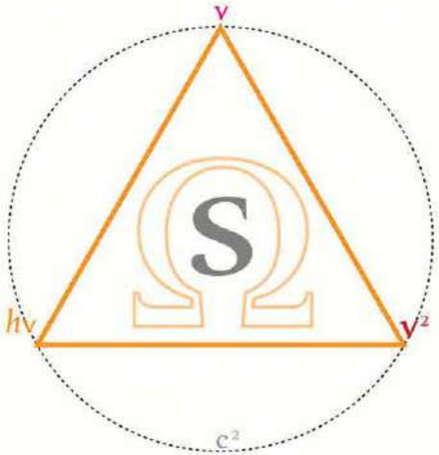

QAM per spatial unit

seconds

$$\frac{\Omega}{c^2} = s$$

Circumscribed Triangles

reflect Energy's relationship with Time

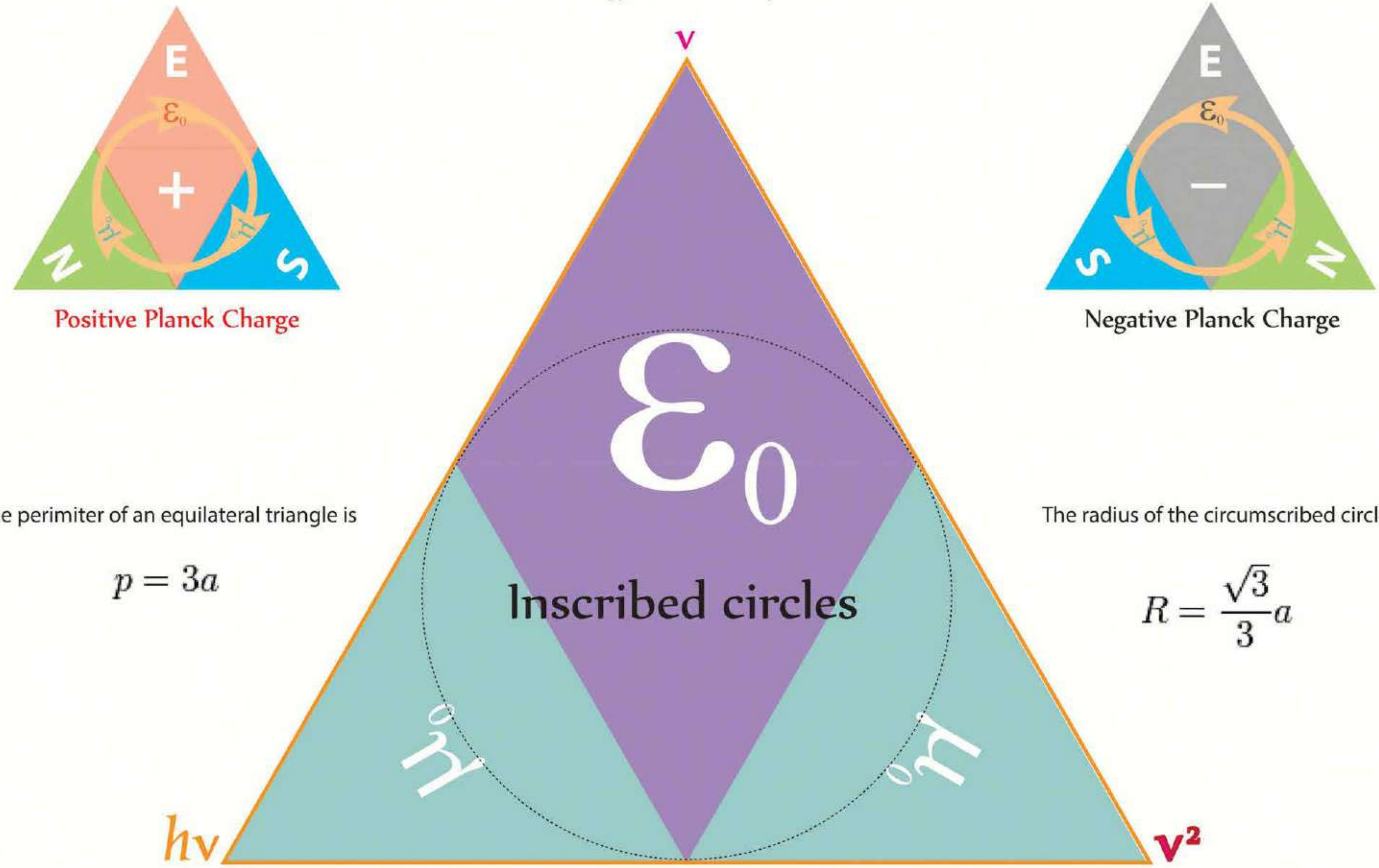
E
ε_0
+
μ_0
μ_0
N
S

Positive Planck Charge

E
ε_0
−
μ_0
μ_0
S
N

Negative Planck Charge

v
ε_0
Inscribed circles
μ_0
μ_0
hv
v²

The perimiter of an equilateral triangle is

$$p = 3a$$

The radius of the circumscribed circle is

$$R = \frac{\sqrt{3}}{3}a$$

The equilateral triangle has the smallest area of all those circumscribed around a given circle

Circumscribed circles

By Euler's inequality, the equilateral triangle has the smallest ratio R/r of the circumradius to the inradius of any triangle: specifically, R/r = 2

The ratio of the area to the square of the perimeter of an equilateral triangle

$$\frac{1}{12\sqrt{3}}$$

The ratio of the area of the incircle to the area of an equilateral triangle

$$\frac{\pi}{3\sqrt{3}}$$

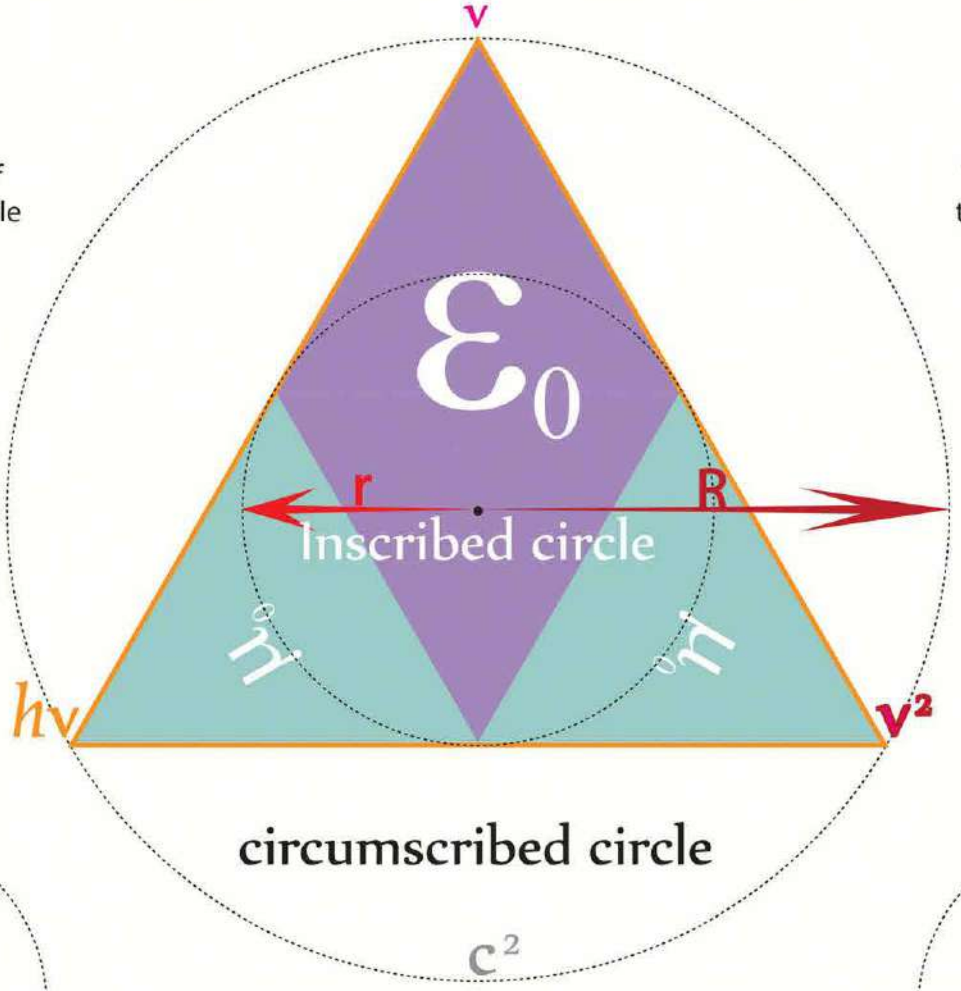
v
ε_0
r
R
Inscribed circle
μ_0
μ_0
hv
v²
circumscribed circle
c²

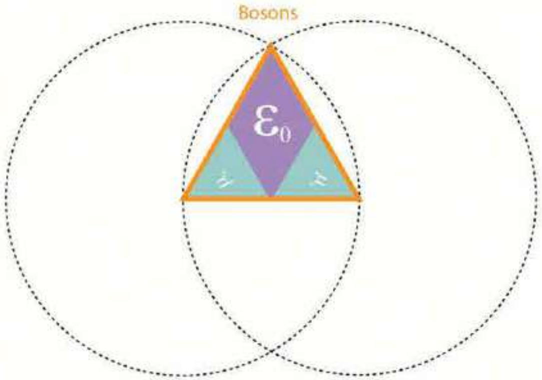
Bosons
ε_0

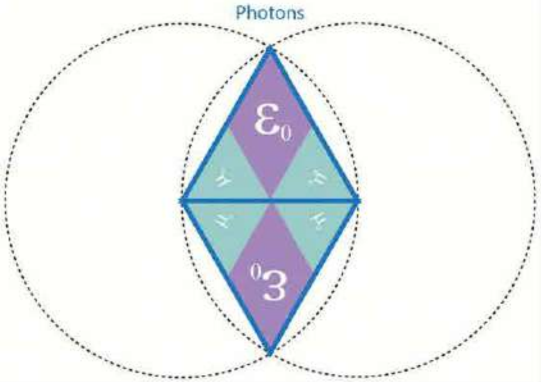
Photons
ε_0
ε_0

Equilateral triangles and Tetrahedrons will scale at exactly the same proportion as Circles and Spheres scribing them

Charged mass-ENERGIES

Charge is the result of quantised angular momenta
[the inscribed circular flux of energy in equilateral geometries]

Time is a measure of changing quantised angular momenta
[the circumscribed spatial co-ordinate of equilateral energy geometries]

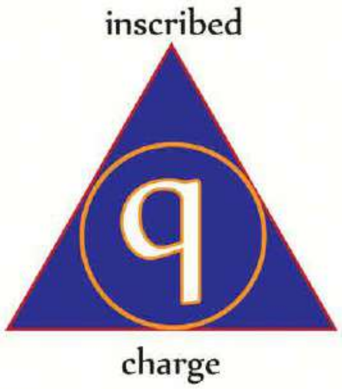

It is the equilateral geometry of energy not a classical vector rotation that creates QAM

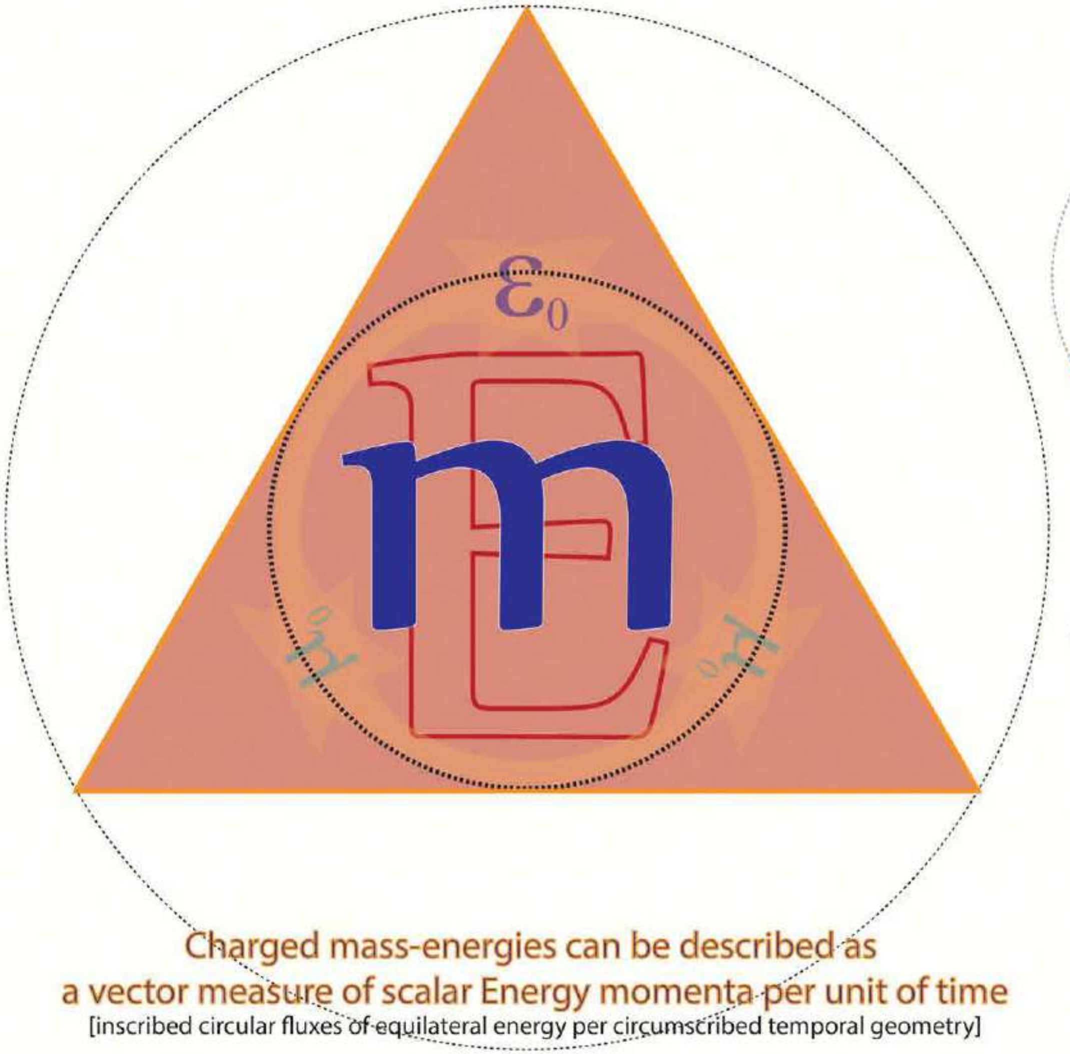

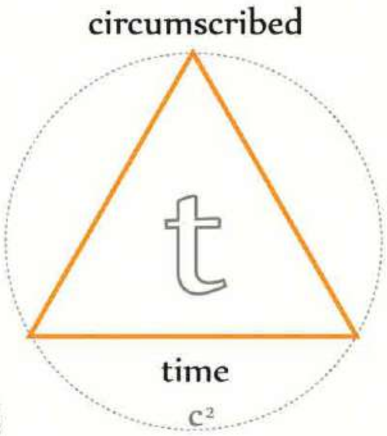

Scalar EM mass is a measure of equilateral Planck energy per spatial co-ordinate system

Charged mass-energies can be described as a vector measure of scalar Energy momenta per unit of time
[inscribed circular fluxes of equilateral energy per circumscribed temporal geometry]

Copyright ABRAHAM [2008] - All rights reserved

Trigometric functions

$$\sin\ \theta = \frac{\text{opposite}}{\text{hypotenuse}}$$

$$\cos\ \theta = \frac{\text{adjacent}}{\text{hypotenuse}}$$

$$\tan\ \theta = \frac{\text{opposite}}{\text{adjacent}}$$

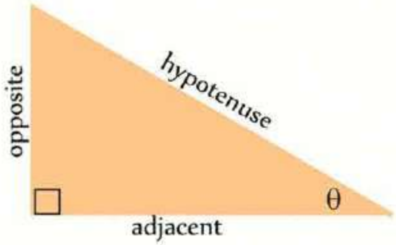
opposite
hypotenuse
θ
adjacent

Standard trigometric functions must be carefully applied to measurements of equilateral Planck mass-energy geometries in scribed circular space-time co-ordinate systems

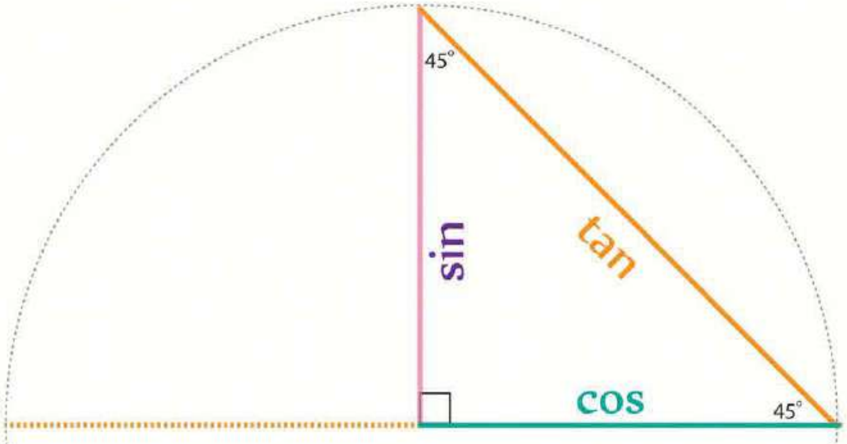
45°
sin
tan
cos
45°

The most familiar trigonometric functions are the sine, cosine, and tangent.

In the context of the standard unit circle with radius 1,
where a triangle is formed by a ray originating at the origin and
making some angle with the x-axis,

the SINE of the angle gives the length of the y-component (rise) of the triangle,
the COSINE gives the length of the x-component (run), and
the TANGENT function gives the slope
(y-component divided by the x-component)

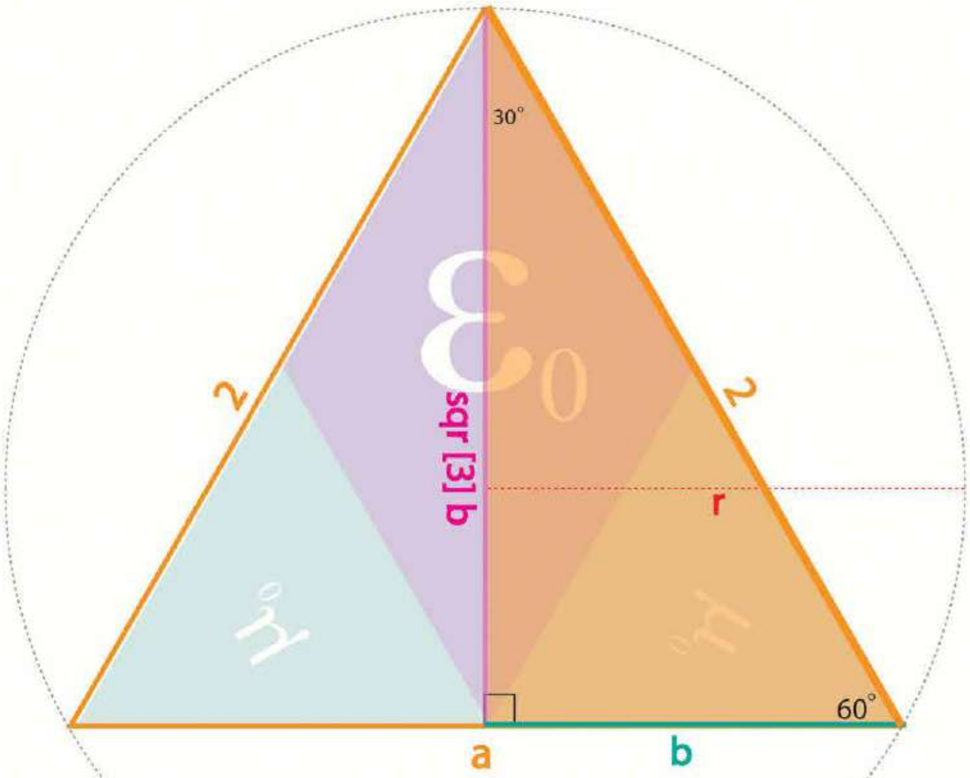
30°
ε_0
sqr [3] b
2
2
r
μ_0
60°
a
b

In most cases it is in fact easier to
ignore the trigonomteric functions and math
and simpy draw their related equilateral geometries

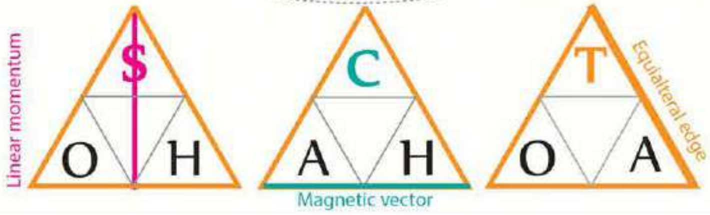
Linear momentum
S
O
H
C
A
H
Magnetic vector
T
O
A
Equilateral edge

The roots of scribed equilateral triangles

Scalar equilateral energies map directly onto circular space-time co-ordinates
through their square root linear momentum

r

The ratio of the
circumscribed circle of
an equilateral triangle
to its inscribed circle
is 2:1

d

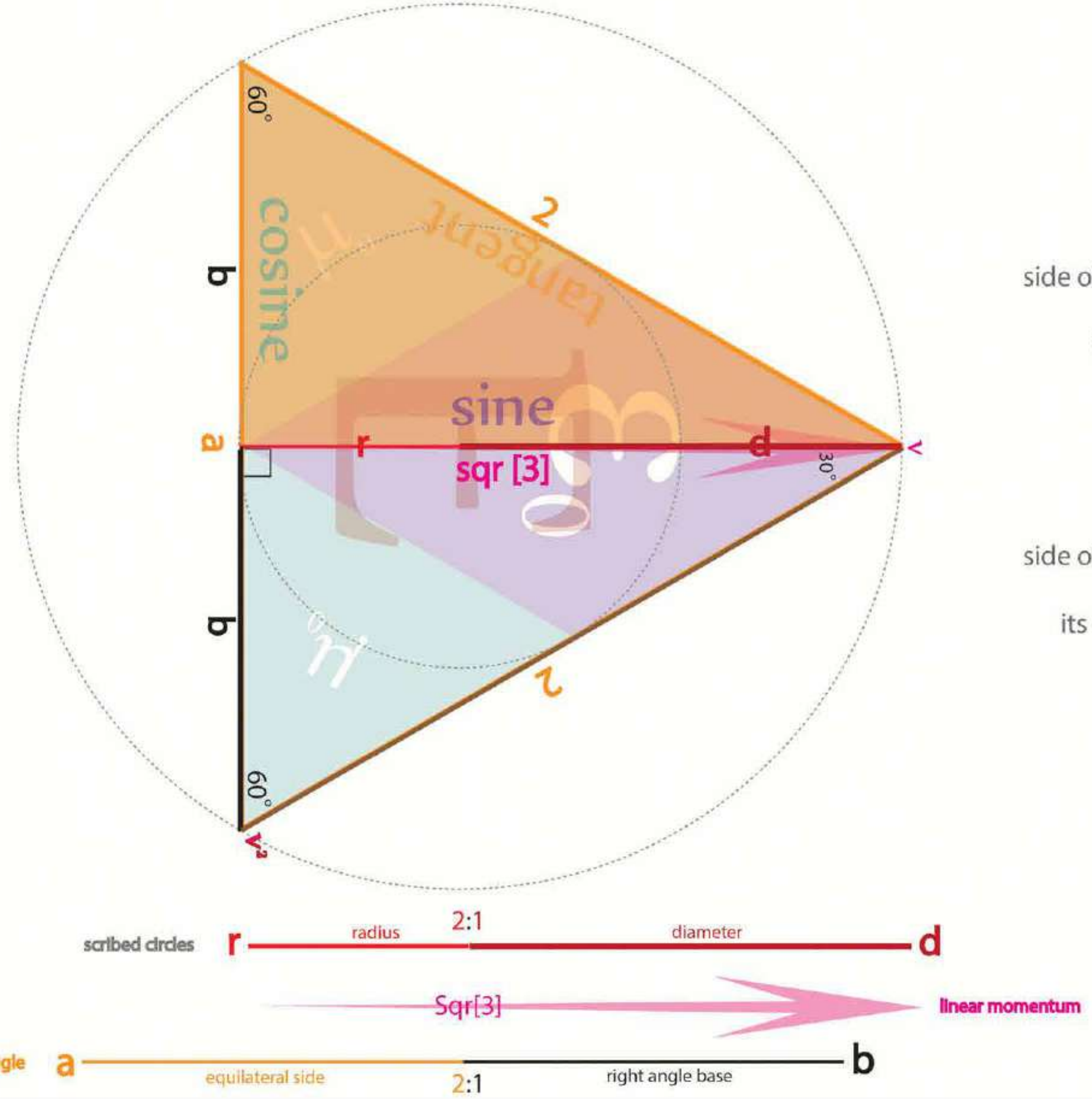
60°
2
tangent
cosine
b
sine
a
r
d
sqr [3]
v
30°
b
2
60°
v²
scribed circles
r
radius
2:1
diameter
d
Sqr[3]
linear momentum
equilateral triangle
a
equilateral side
2:1
right angle base
b

0.866 r

The ratio of the
side of an equilateral triangle
to the radius of
its incribed circle
is sqr [3]/2

a

The ratio of the
side of an equilateral triangle
to the radius of
its circumscribed circle
is sqr [3]

1.732 d

Equilateral triangles and scribed circles

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1 - \beta^2}} = \frac{dt}{d\tau}$$

$$\gamma(\beta) = 1 + \frac{1}{2}\beta^2 + \frac{3}{8}\beta^4 + \frac{5}{16}\beta^6 + \frac{35}{128}\beta^8 + \ldots$$

$$L = L'\sqrt{1 - \frac{v^2}{c^2}}$$

$$E_k = \frac{mc^2}{\sqrt{1 - (v/c)^2}} - mc^2$$

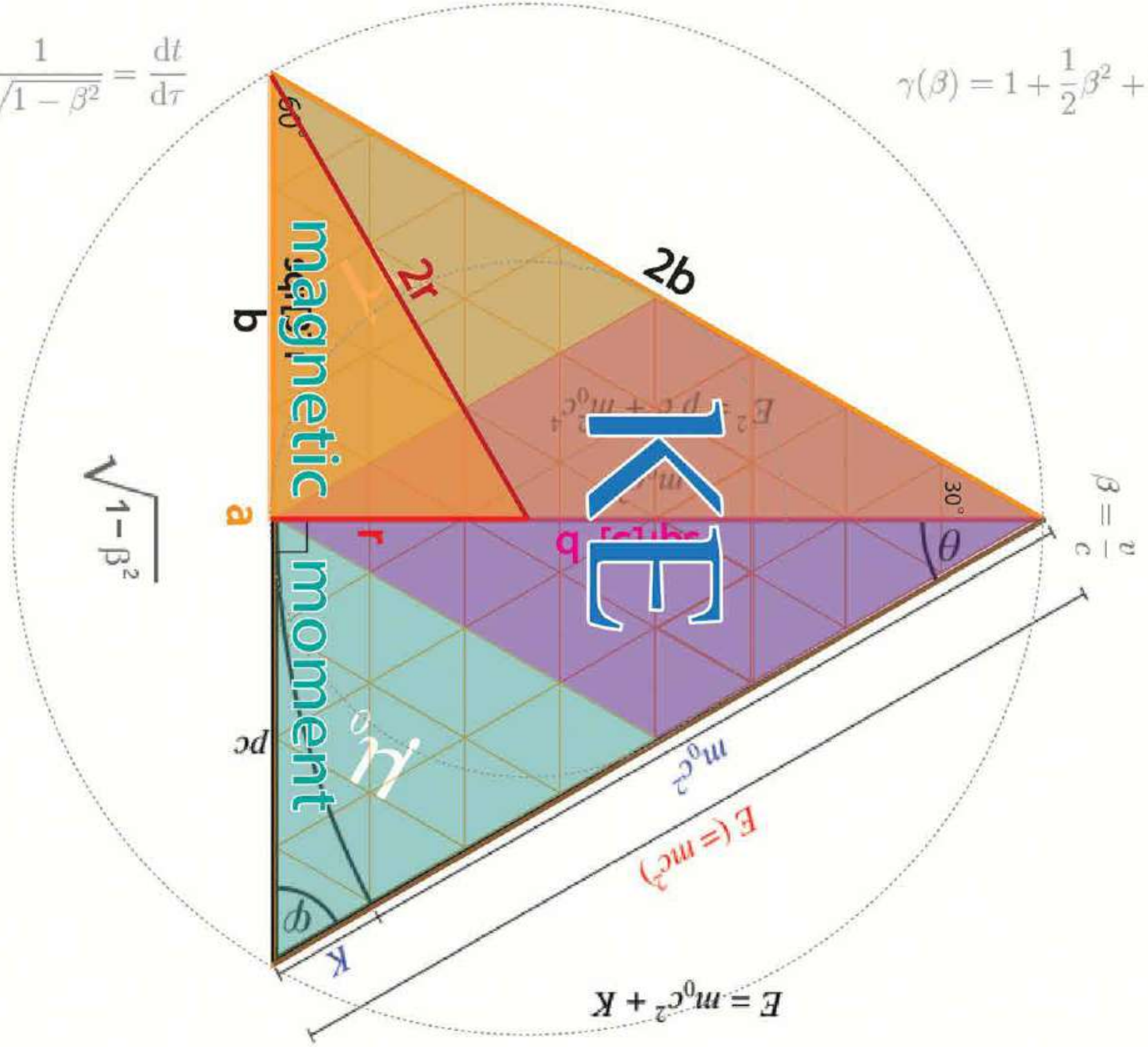
60°
2r
2b
b
magnetic
moment
μ
a
r
q
KE
30°
θ
φ
pc
K
m_0c^2
$E^2 = p^2c^2 + m_0^2c^4$
$E\ (= mc^2)$
$E = m_0c^2 + K$
$\sqrt{1-\beta^2}$
$\beta = \frac{v}{c}$
μ_0

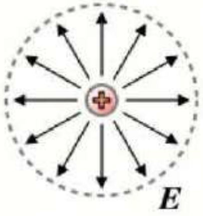

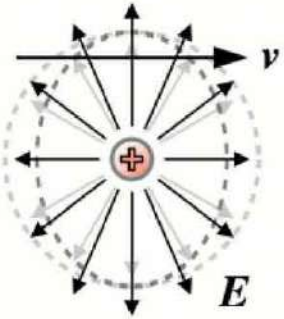

$$mc^2\left\{\frac{1}{\sqrt{1 - v^2/c^2}} - 1\right\}$$

$$p = \frac{m_0 v}{\sqrt{1 - \frac{v^2}{c^2}}} = \gamma m_0 v$$

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$\sqrt{1 - \left(\frac{v}{c}\right)^2}$$

All the relativistic relationships historically attributed to circularised energies and modelled through the use of the math of right angled triangles are in fact the result of equilateral, scalar geometries

Finite sequences and series have defined first and last terms,

$$\sum a_n = a_0 + a_1 + a_2$$

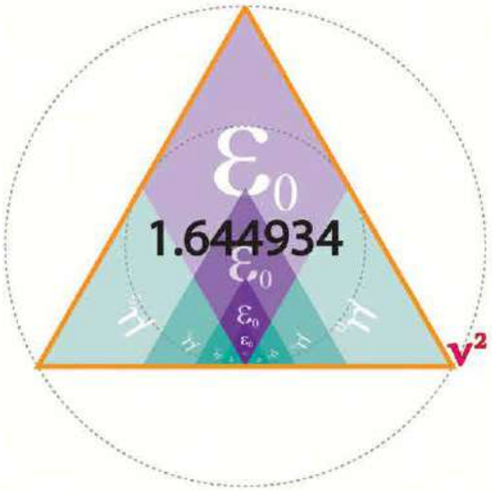

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \lim_{n \to +\infty} \left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} \right).$$

whereas infinite sequences and series continue indefinitely

The Basel problem is a famous problem in mathematical analysis with relevance to number theory, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735.

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \ldots = \frac{\pi^2}{6}$$

Tetryonics now provides a geometric solution to visualising and solving the Basel problem

$$\zeta(2) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \frac{\pi^2}{6} \approx 1.645;$$

Tetryonic Infinite Series

is a infinite sequence of square numbers,
the result of adding all those terms together
[or their geometric inverse]

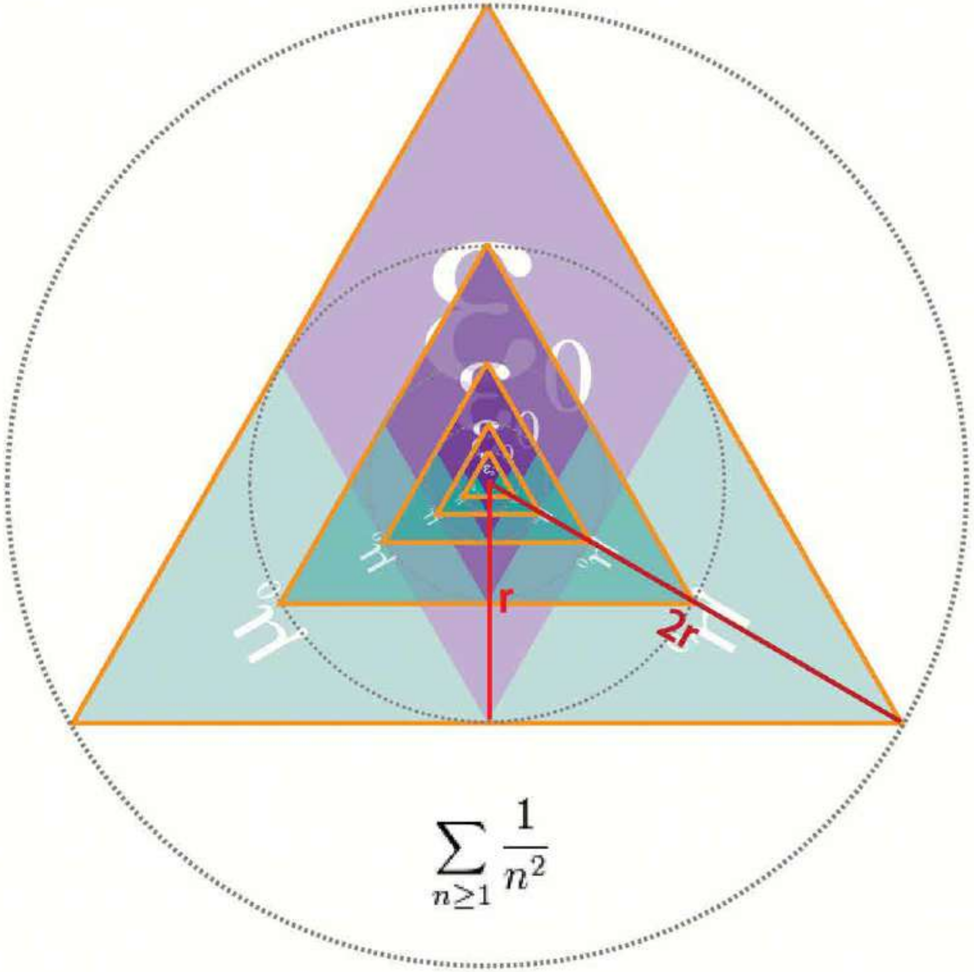

$$\sum_{n \geq 1} \frac{1}{n^2}$$

The entire sum of the series is equal to
twice the size of the radius of the largest inscribed circle
which is equal to the largest circle circumscribing the triangular series.

Copyright ABRAHAM [2008] - All rights reserved

Inverting the Circle

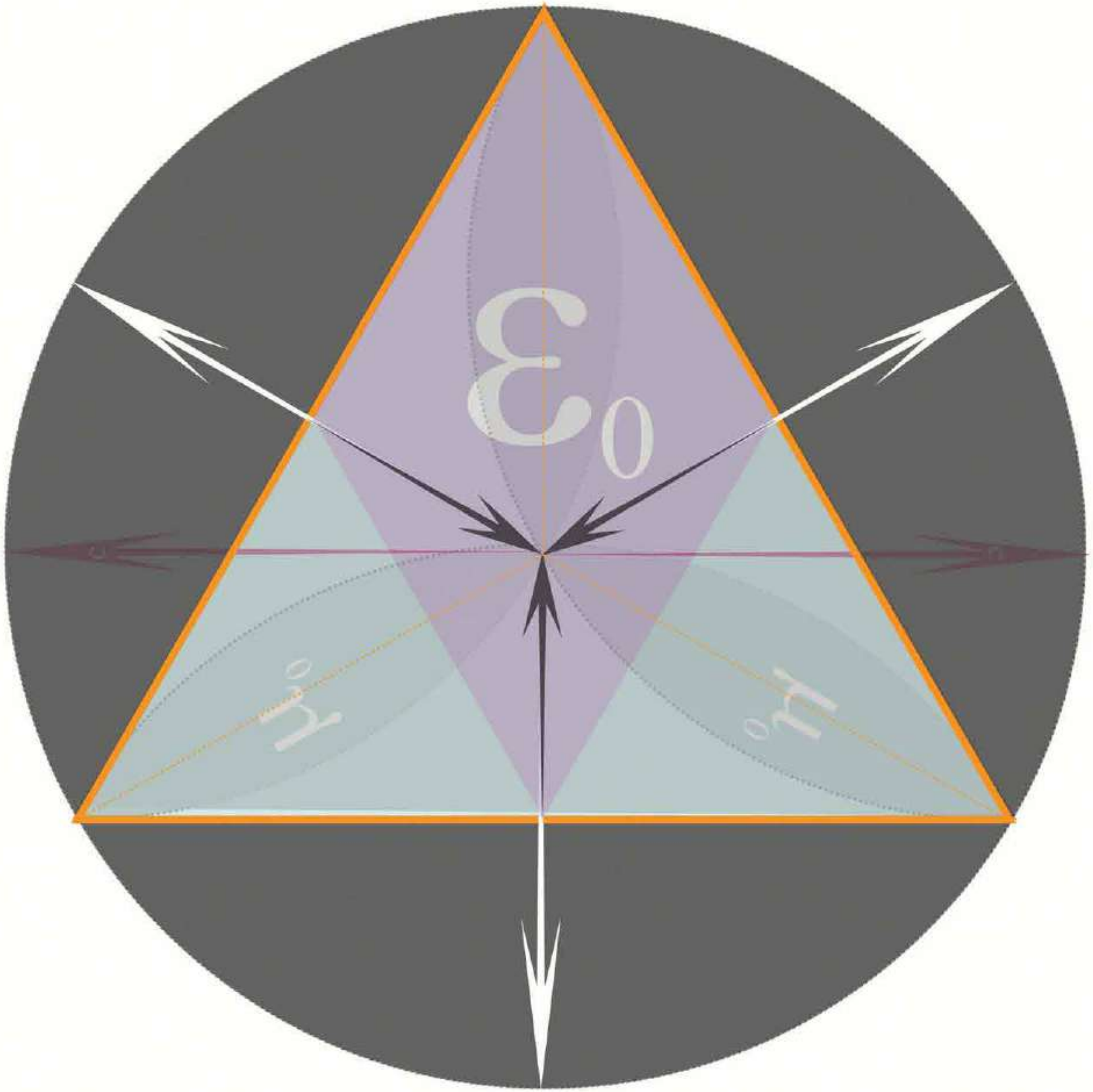

Electric
Permittivity

ε_o

Magnetic
Permeability

μ_o

c^2

$\frac{1}{c^2}$

2D spatial
co-ordinates

ElectroMagnetic
fields

Irrational numbers

an irrational number cannot be represented as a simple fraction.
Irrational numbers are those real numbers that cannot be represented as terminating or repeating decimals

sin(x)=-square root of (3)/2

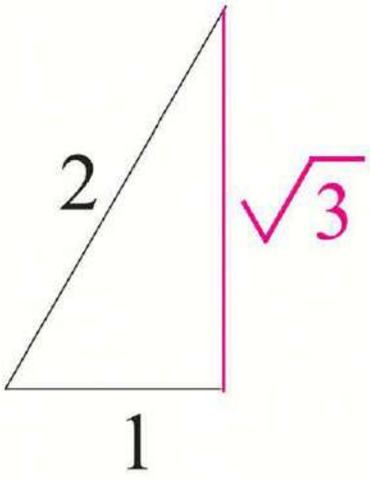

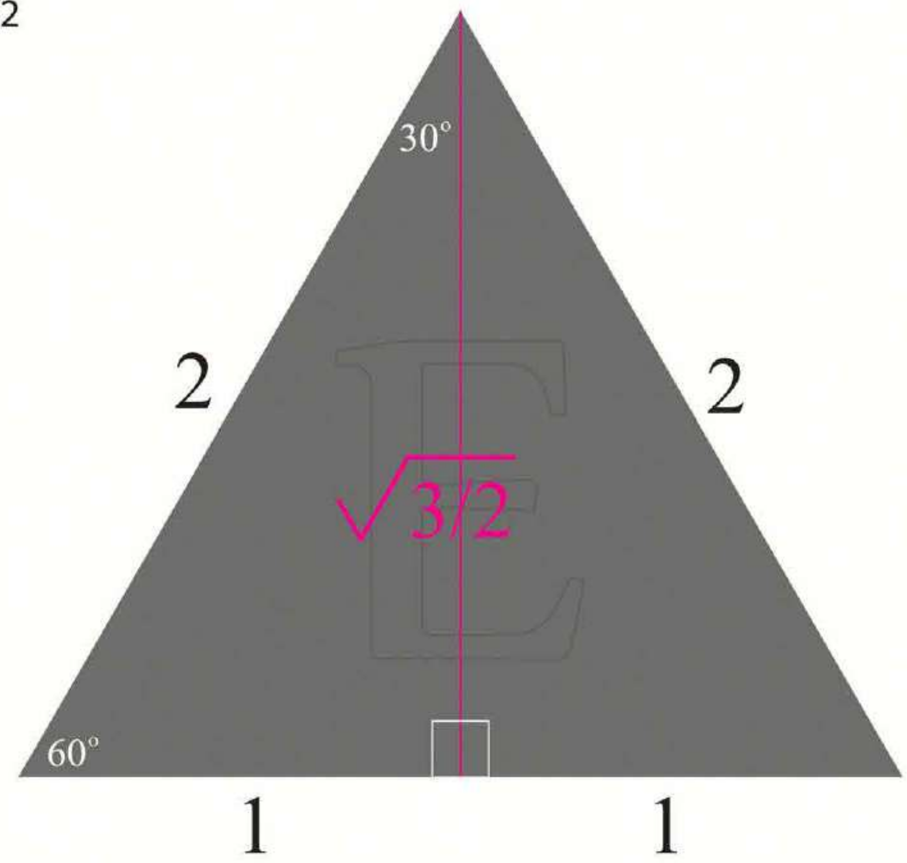
30°
2
2
$\sqrt{3/2}$
E
60°
1
1

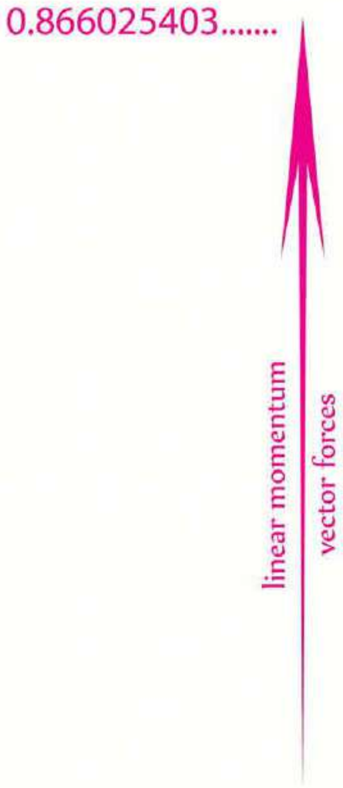
0.866025403.......
linear momentum
vector forces

Pythagφras' theoreom
and iratiφnal numbers expressed in terms of right-angled triangles in Physics
offer a 'half truth' regarding the equilateral geometry of Energy

Leibniz

Newton

linear momentum

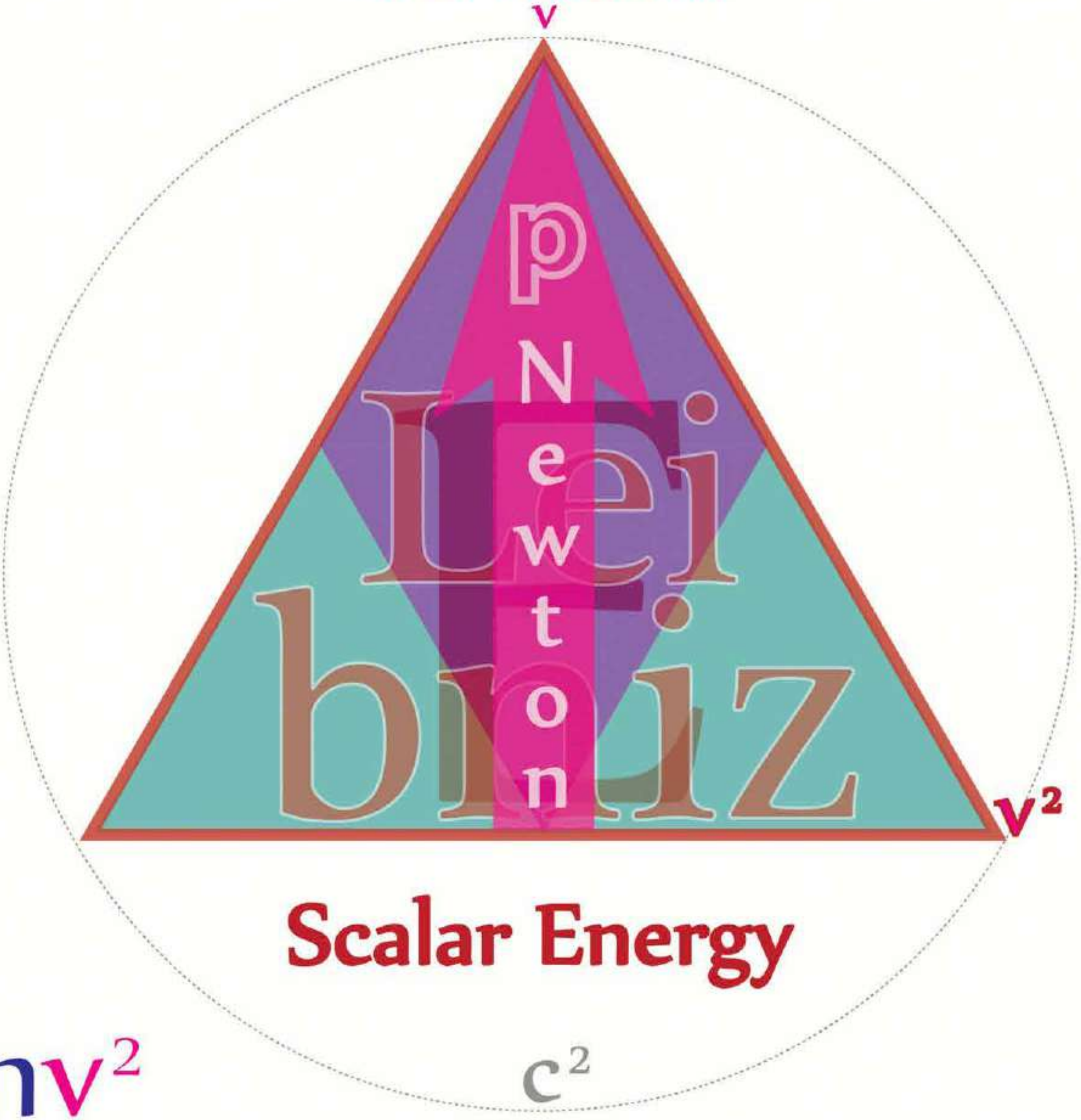

Newton focused his work on linear momentum which he developed into his famous laws of motion

Leibniz introduced the concept of vis viva (living force) which attempted to quantify life force

Newton and Leibniz disagreed about what the world is made of and how its physics shaped our scientific concepts of force, energy, and momentum

Tetryonics reveals the physical relationships they both described mathematically as geometric properties of equilateral Planck energy momenta

Scalar Energy

$E=mv^2$

$p=mv$

Geometric Square Roots

In geometrical terms, the square root function maps the area of a square to its side length.

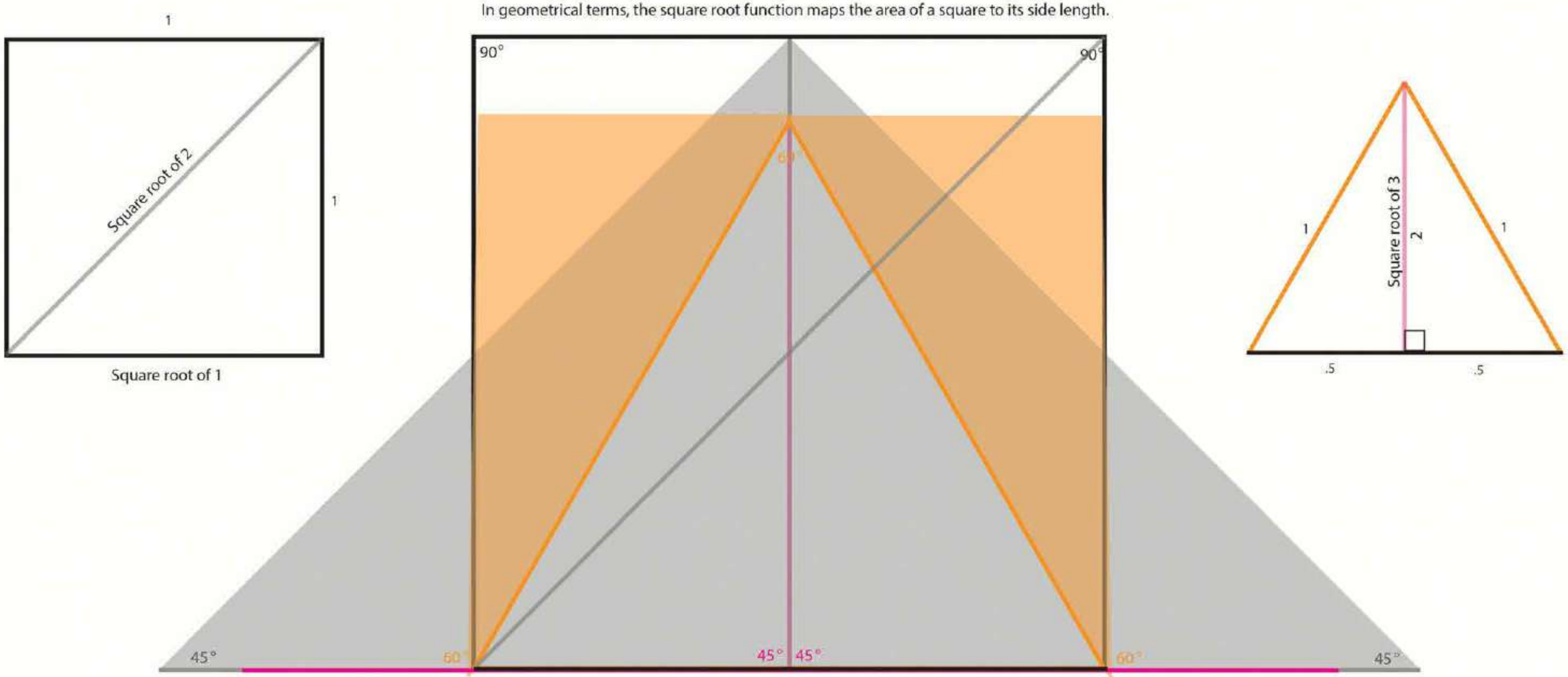

"In mathematics, as in any scientific research, we find two tendencies present. On the one hand, the tendency toward abstraction seeks to crystallize the logical relations inherent in the maze of material that is being studied, and to correlate the material in a systematic and orderly manner. On the other hand, the tendency toward intuitive understanding fosters a more immediate grasp of the objects one studies, a live rapport with them, so to speak, which stresses the concrete meaning of their relations"

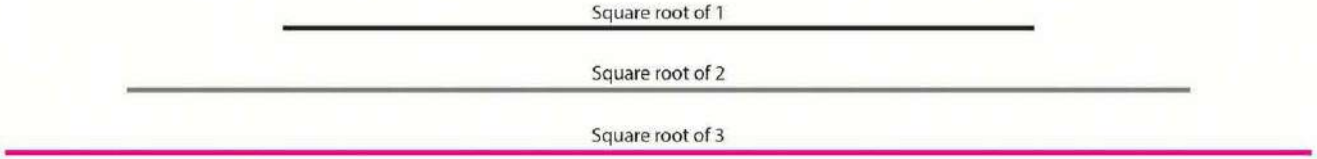

"As to geometry, in particular, the abstract tendency has here led to the magnificent systematic theories of Algebraic Geometry, of Riemannian Geometry, and of Topology; these theories make extensive use of abstract reasoning and symbolic calculation in the sense of algebra.

Notwithstanding this, it is still as true today as it ever was that intuitive understanding plays a major role in geometry. And such concrete intuition is of great value not only for the research worker, but also for anyone who wishes to study and appreciate the results of research in geometry"

David Hilbert [Geometry and the Imagination]

Square Roots in Physics

In mathematics, a square root of a number a is a number [n] such that [n]2 = x,
or, in other words, a number [n] whose square (the result of multiplying the number by itself, or [n × n]) is x.

Modern calculators use the
Square Root Algorithm
to calculate the value

It is an approximate numerical value

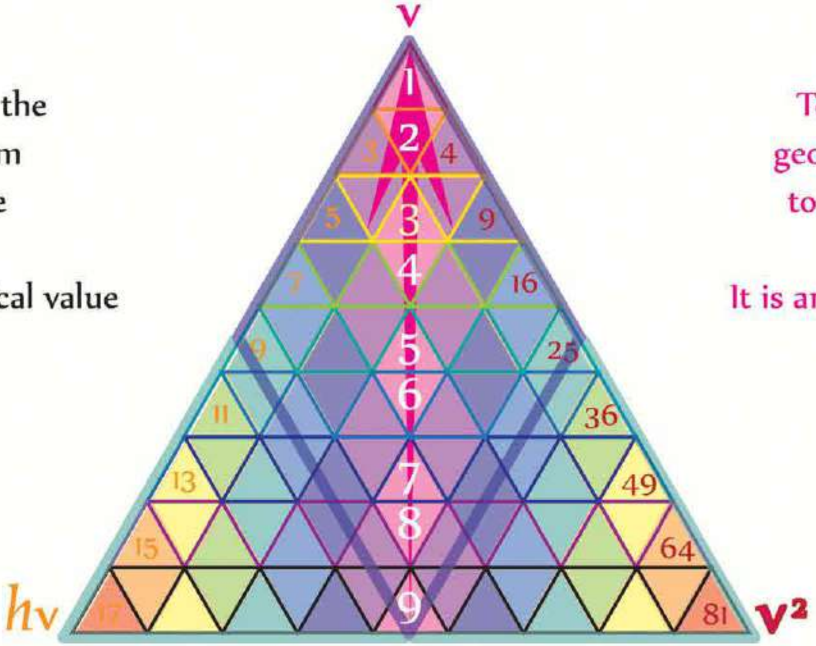
v
1
2
3
4
3
4
5
9
5
6
7
16
9
25
11
36
7
13
49
8
15
64
9
hv
17
81
v²

Tetryonics uses the
geometric Square Root
to calculate the value

It is an exact geometric value

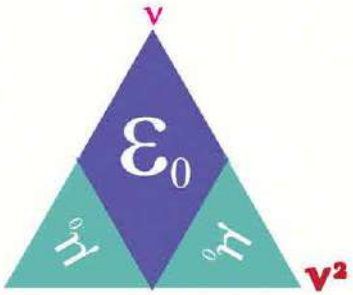
v
ε_0
μ_0
μ_0
v²

$$E = mv^2$$

*In classical geometry,
the square root function maps the
area of a square to its side length.*

$$p = \sqrt{E}$$

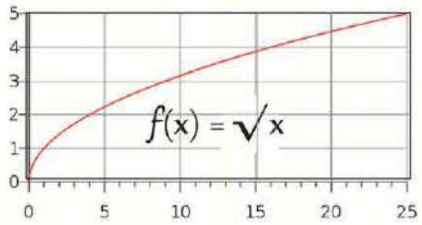

Geometry can easily map irrational numbers

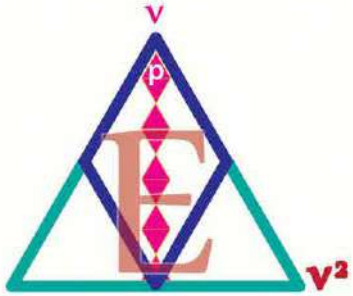

$$p^2 = E$$

*In physics,
the square root function maps
ENERGY [E] to momentum [mv]*

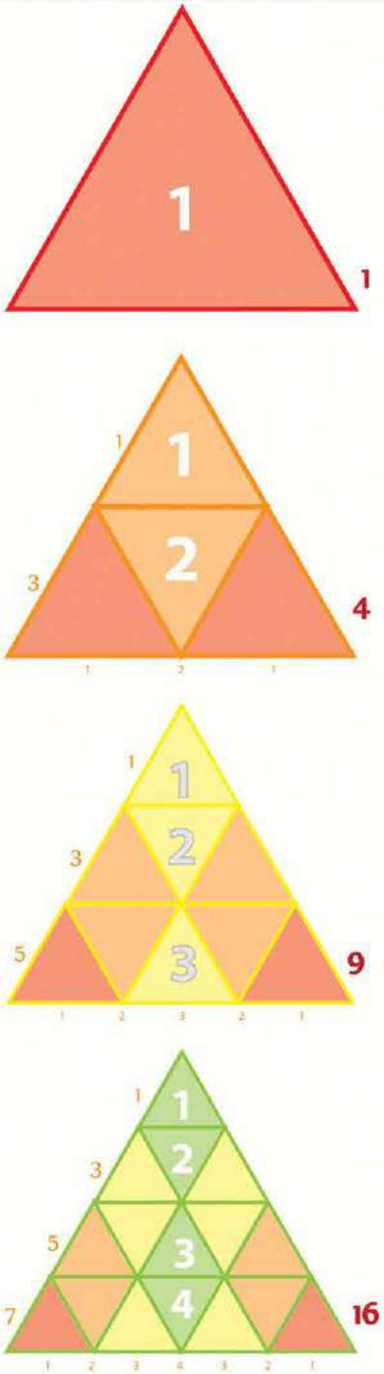

The Square roots of n

Historically, any number raised to the power of 2 has been modeled using a polygon--the square
That's why we call raising a number to the second power "squaring the number."

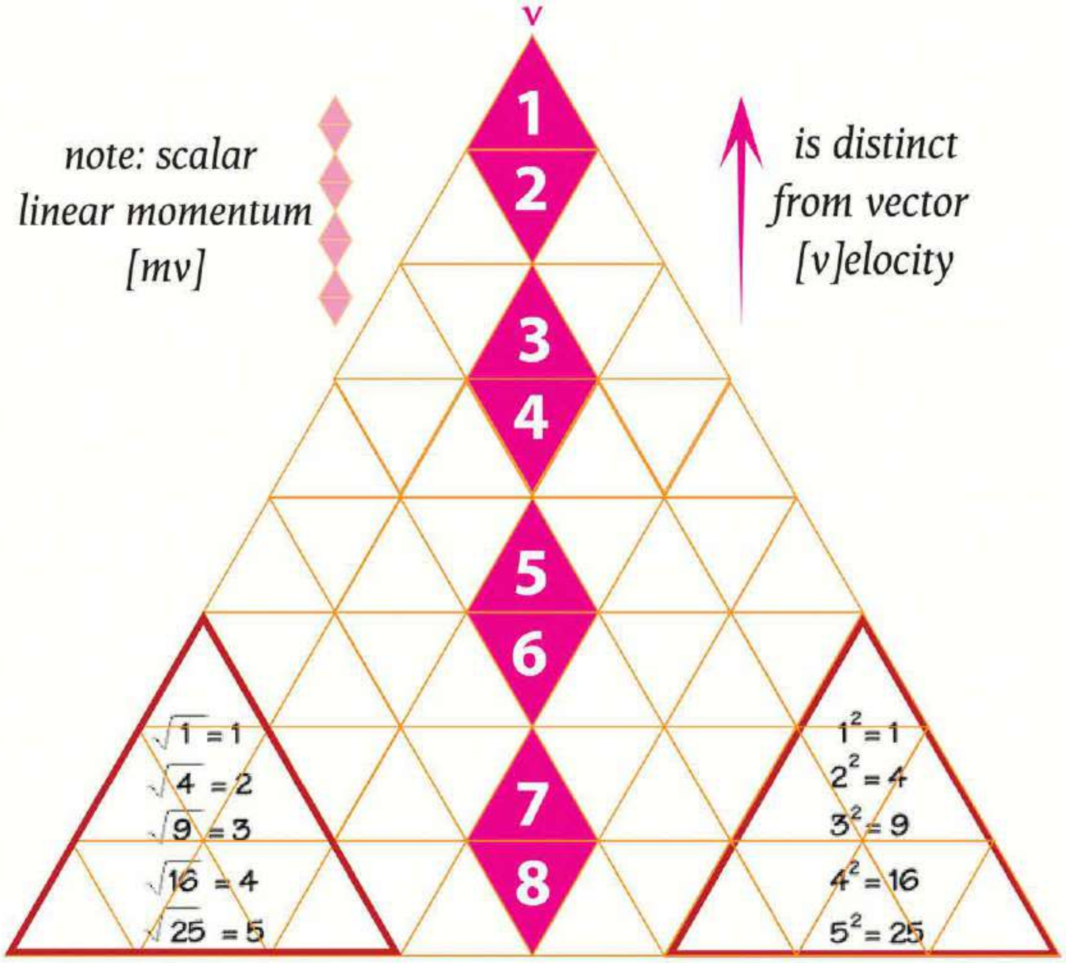

note: scalar linear momentum [mv]

is distinct from vector [v]elocity

$\sqrt{1} = 1$
$\sqrt{4} = 2$
$\sqrt{9} = 3$
$\sqrt{16} = 4$
$\sqrt{25} = 5$

$1^2 = 1$
$2^2 = 4$
$3^2 = 9$
$4^2 = 16$
$5^2 = 25$

[In physics square numbers are in fact equilateral geometries]
The perfect squares are squares of whole numbers.
Here are the first eight perfect squares

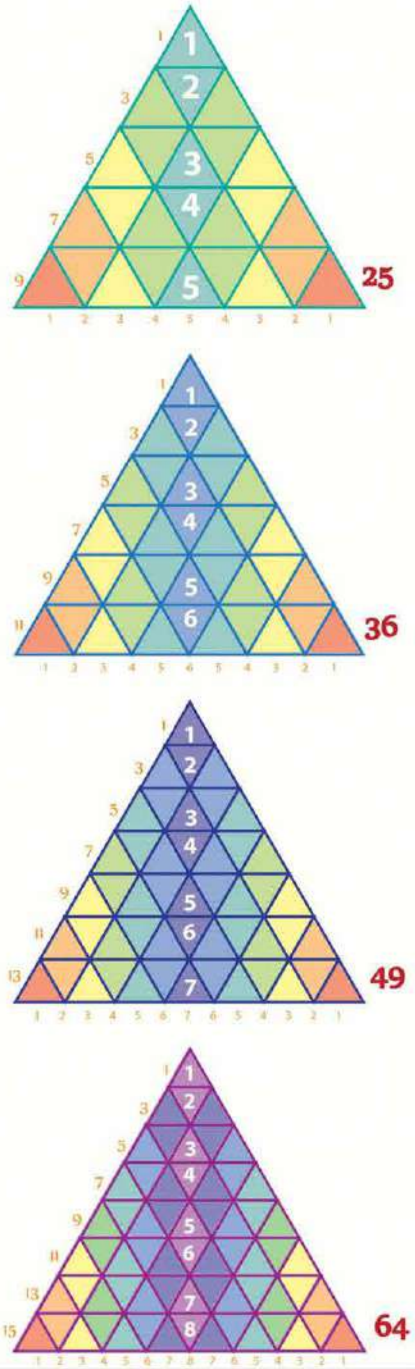

The Square root of Negative 1

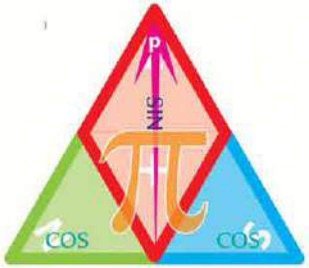
P
SIN
π
COS
COS

Positive fields are out of phase with Negative fields

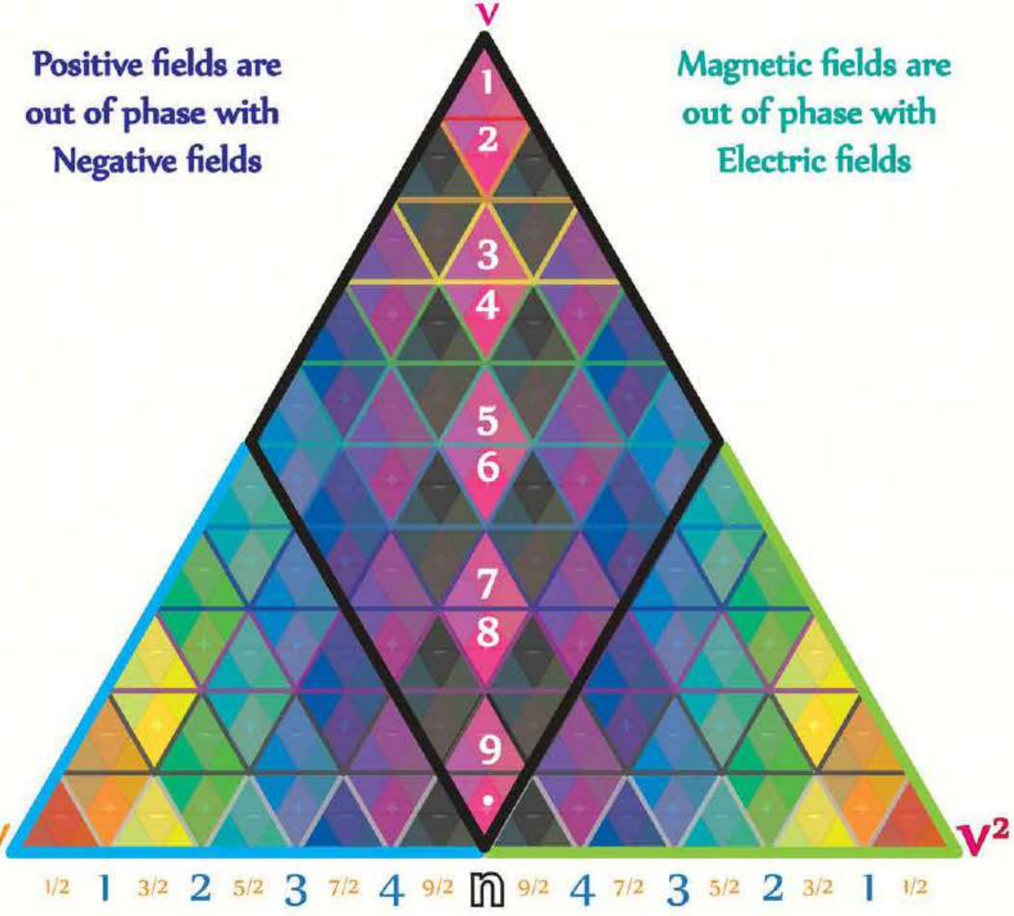
v
1
2
3
4
5
6
7
8
9
hv
v²
1/2 1 3/2 2 5/2 3 7/2 4 9/2 n 9/2 4 7/2 3 5/2 2 3/2 1 1/2

Magnetic fields are out of phase with Electric fields

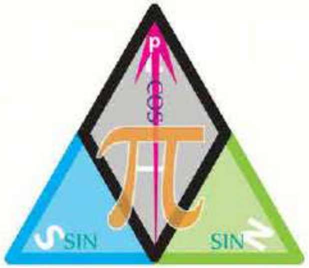
P
COS
π
SIN
SIN

Euler's Formula

Euler's formula is often considered to be the basis of the complex number system. In deriving this formula, Euler established a relationship between the trigonometric functions, sine and cosine, and e raised to a power

$$e^{ix} = \cos(x) + i\sin(x)$$

a mathematical description of EM-Energy waveforms

Leonhard Euler

(15 April 1707 – 18 September 1783)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

$$\sin x + \cos x = 1 + x - \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

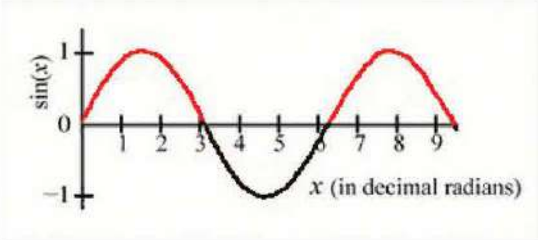

$$e^{i\pi} = -1$$

$$e^{i\pi} + 1 = 0$$

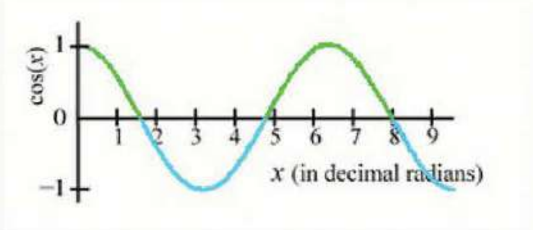

Copyright ABRAHAM [2008] - All rights reserved

Geometric means

The geometric mean of two numbers, is the square root of their product

geometric square root
of positive one

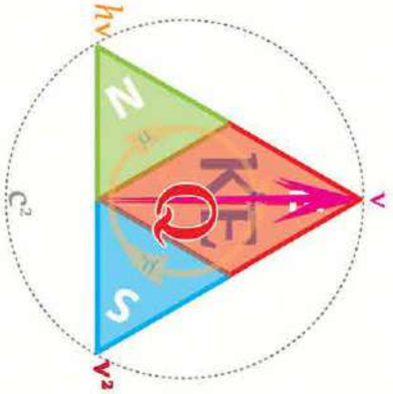

$$F = k \frac{Q_1 Q_2}{r^2}$$

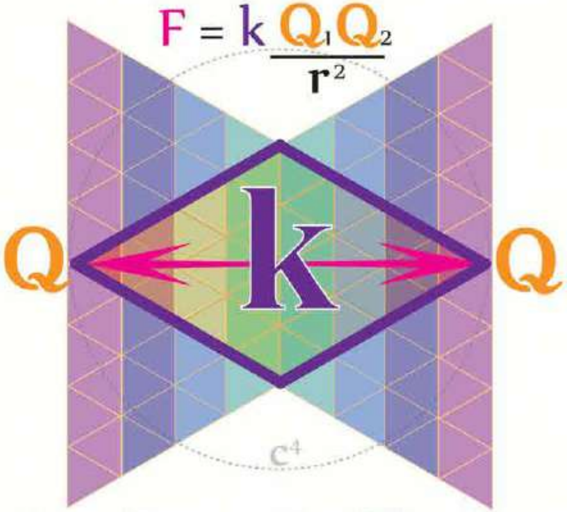

geometric square root
of negative one

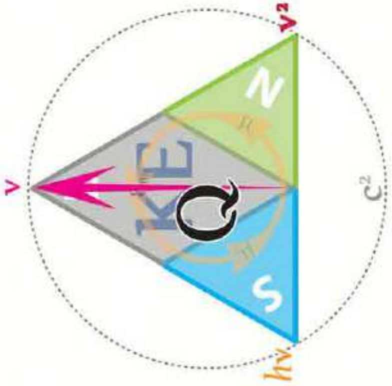

In physics, the geometric mean of two superpositioned fields produces a vector square root Force

It is generally stated that
the geometric mean applies
only to positive numbers.

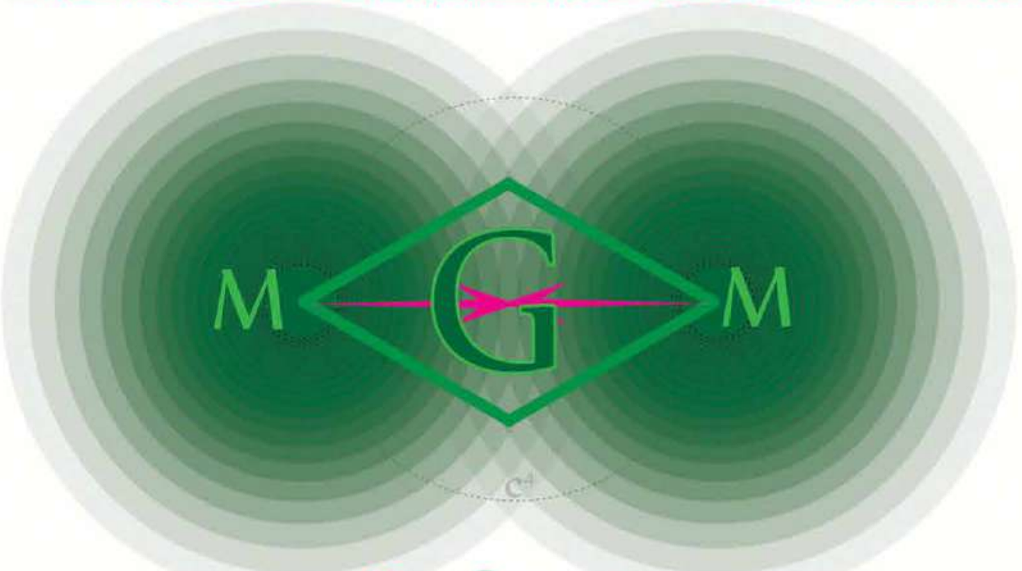

$$F = -G \frac{m_1 m_2}{r^2}$$

In Tetryonic geometry
the geometric mean applies
to positive & negative fields.

Copyright ABRAHAM [2008] - All rights reserved

Superpositioning

When two or more waves traverse the same space,
the net amplitude at each point is the
sum of the amplitudes of the individual waves.

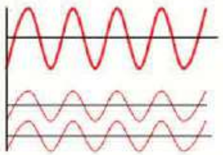

constructive interference

In phase

Additive

+

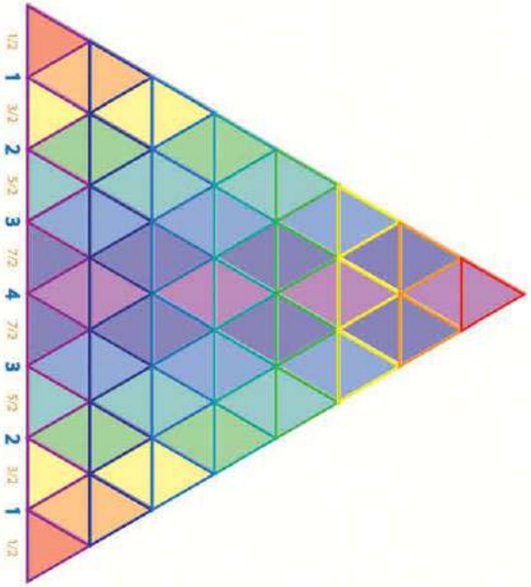
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

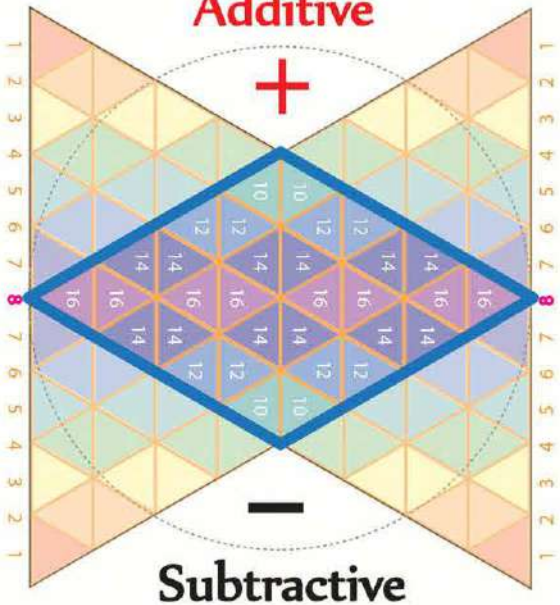
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
10
12
14
16

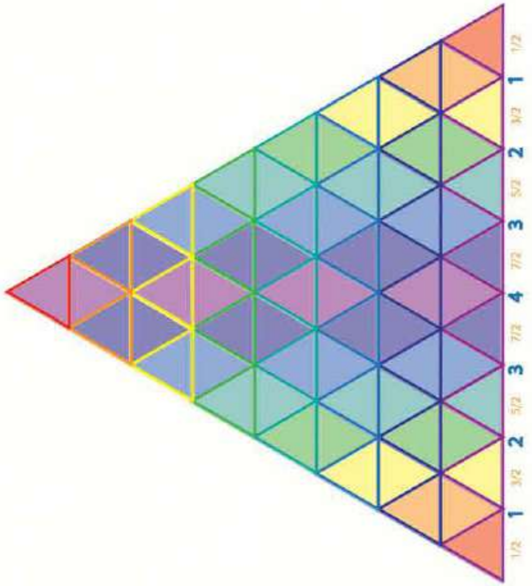
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

−

Subtractive

Out of phase

destructive interference

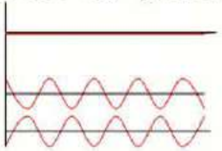

When two or more waves traverse the same space,
if the summed variation has a smaller amplitude
than the constiuent component variations.

The lines of Force

Tetryonic Multiplication table

A multiplication table is a mathematical grid used to define a multiplication operation and its results

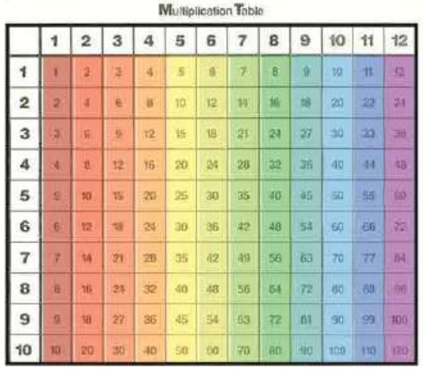

Multiplication Table

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	3	4	5	6	7	8	9	10	11	12
2	2	4	6	8	10	12	14	16	18	20	22	24
3	3	6	9	12	15	18	21	24	27	30	33	36
4	4	8	12	16	20	24	28	32	36	40	44	48
5	5	10	15	20	25	30	35	40	45	50	55	60
6	6	12	18	24	30	36	42	48	54	60	66	72
7	7	14	21	28	35	42	49	56	63	70	77	84
8	8	16	24	32	40	48	56	64	72	80	88	96
9	9	18	27	36	45	54	63	72	81	90	99	100
10	10	20	30	40	50	60	70	80	90	100	110	120

Historically Multiplication tables
have been based on Square geometries

In Tetryonics multiplication tables can
also be based on EQUILATERAL geometries

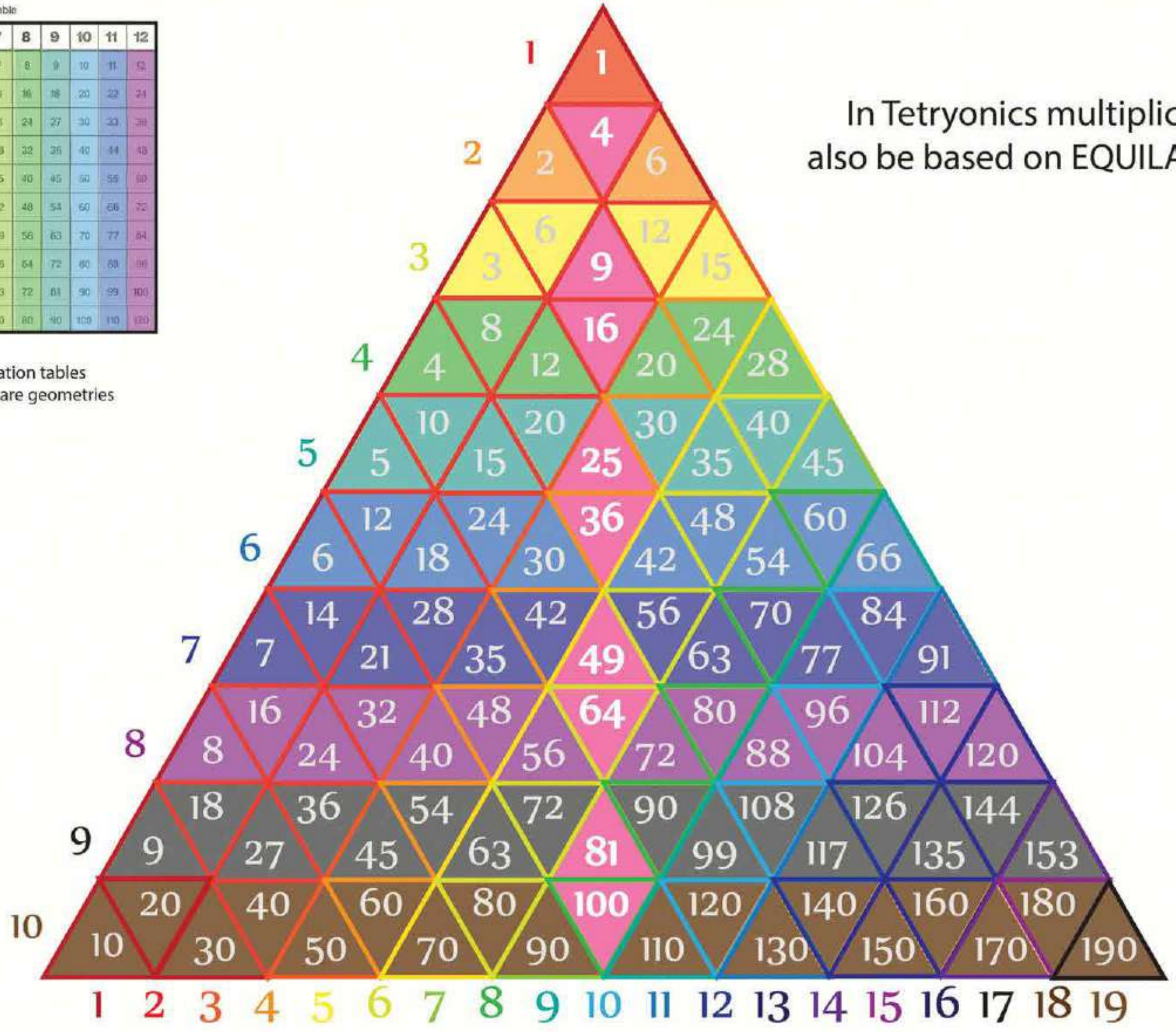

1: 1
2: 2 4 6
3: 3 6 9 12 15
4: 4 8 12 16 20 24 28
5: 5 10 15 20 25 30 35 40 45
6: 6 12 18 24 30 36 42 48 54 60 66
7: 7 14 21 28 35 42 49 56 63 70 77 84 91
8: 8 16 24 32 40 48 56 64 72 80 88 96 104 112 120
9: 9 18 27 36 45 54 63 72 81 90 99 108 117 126 135 144 153
10: 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190

The integer multiplicitors are colour coded

Copyright ABRAHAM [2008] - All rights reserved

Rhombic Multiplication Tables

1	2	3	4	5	6	7	8	9	10
2	**4**	6	8	10	12	14	16	18	20
3	6	**9**	12	15	18	21	24	27	30
4	8	12	**16**	20	24	28	32	36	40
5	10	15	20	**25**	30	35	40	45	50
6	12	18	24	30	**36**	42	48	54	60
7	14	21	28	35	42	**49**	56	63	70
8	16	24	32	40	48	56	**64**	72	80
9	18	27	36	45	54	63	72	**81**	90
10	20	30	40	50	60	70	80	90	**100**

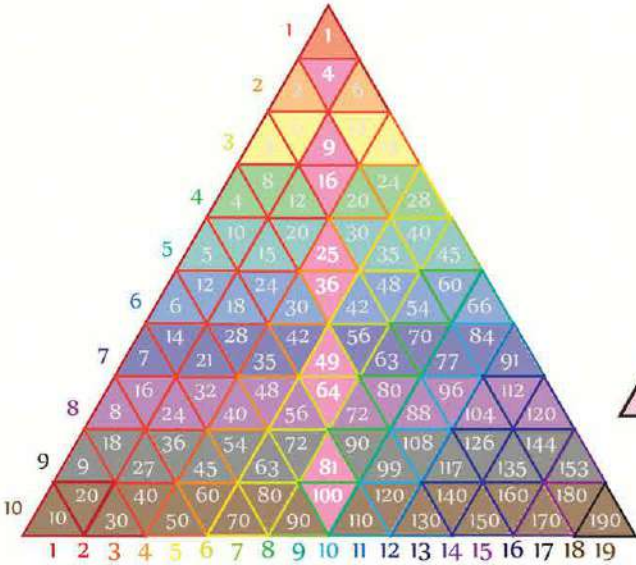

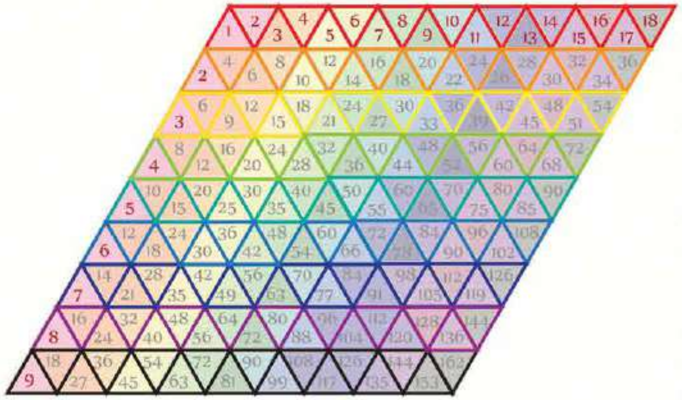

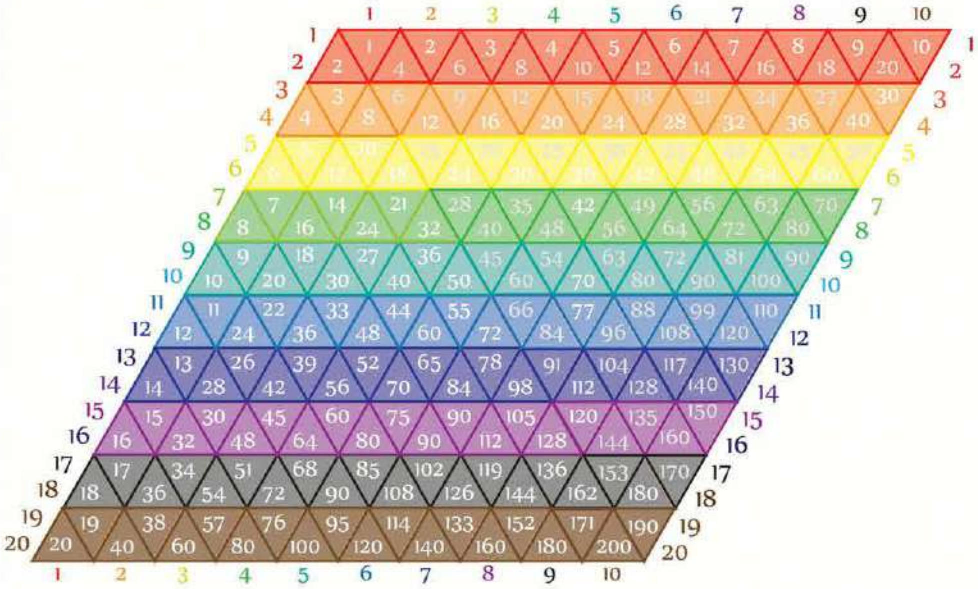

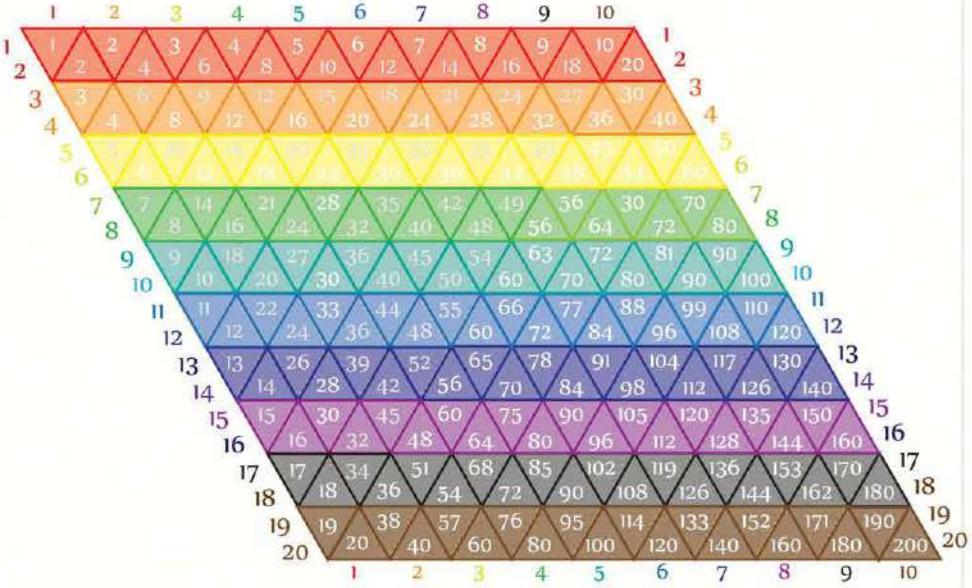

Photonic Root Tables

Tetryonic multiplication table can take a number of geometric forms

Square root median

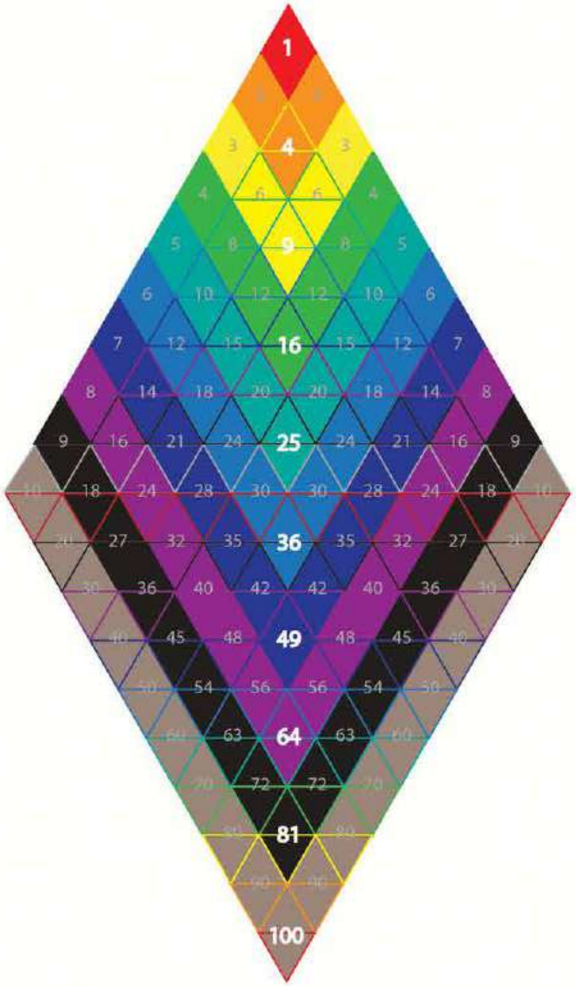

Table read diagonally

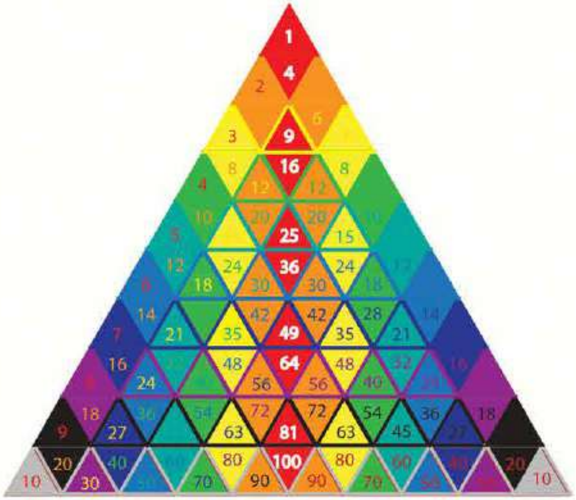

1	2	3	4	5	6	7	8	9	10
2	**4**	6	8	10	12	14	16	18	20
3	6	**9**	12	15	18	21	24	27	30
4	8	12	**16**	20	24	28	32	36	40
5	10	15	20	**25**	30	35	40	45	50
6	12	18	24	30	**36**	42	48	54	60
7	14	21	28	35	42	**49**	56	63	70
8	16	24	32	40	48	56	**64**	72	80
9	18	27	36	45	54	63	72	**81**	90
10	20	30	40	50	60	70	80	90	**100**

Integer median

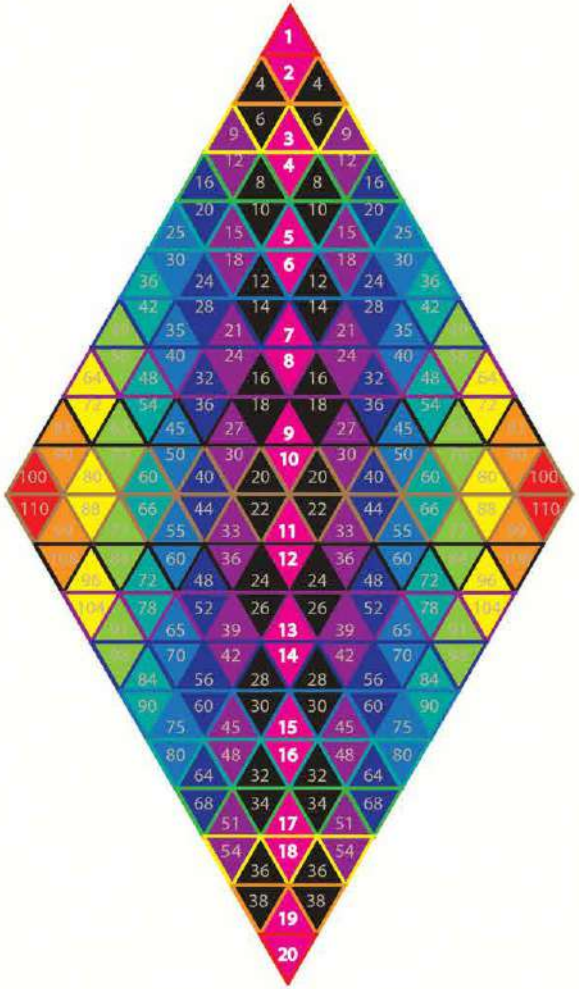

Table read from centre
to outside edge then down

Divide [division] tables

Minus [subtraction] tables

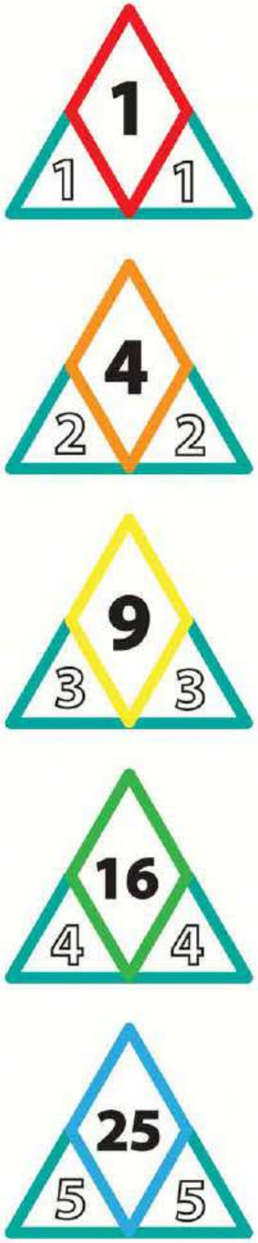
1
1
1
4
2
2
9
3
3
16
4
4
25
5
5

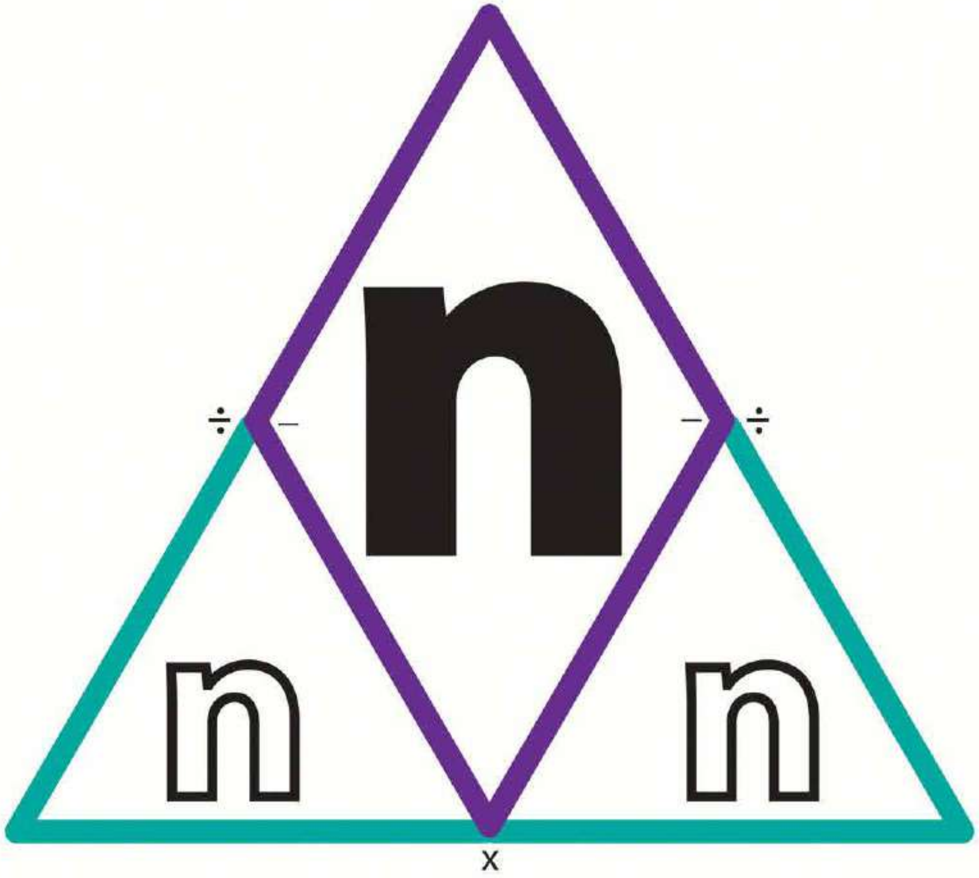

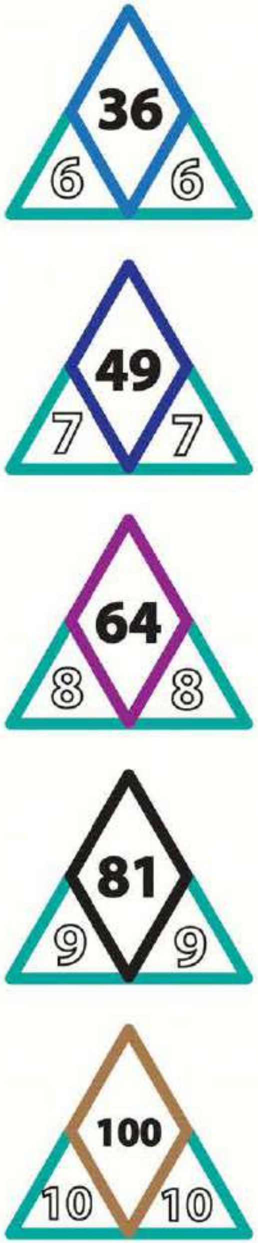
36
6
6
49
7
7
64
8
8
81
9
9
100
10
10

Times [multiplication] tables

Plus [addition] tables

Copyright ABRAHAM [2008] - All rights reserved

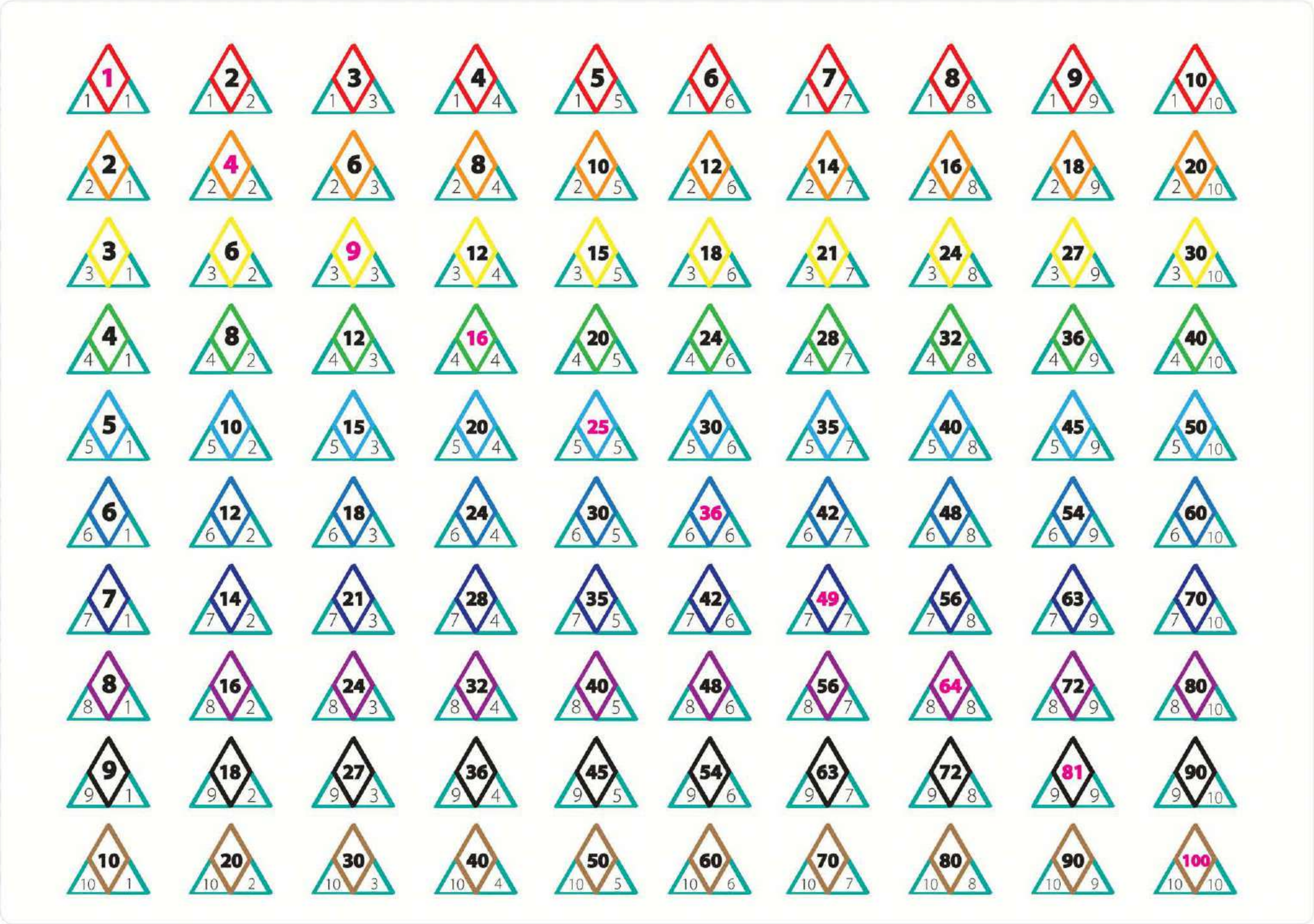

1 1 1	2 1 2	3 1 3	4 1 4	5 1 5	6 1 6	7 1 7	8 1 8	9 1 9	10 1 10
2 2 1	4 2 2	6 2 3	8 2 4	10 2 5	12 2 6	14 2 7	16 2 8	18 2 9	20 2 10
3 3 1	6 3 2	9 3 3	12 3 4	15 3 5	18 3 6	21 3 7	24 3 8	27 3 9	30 3 10
4 4 1	8 4 2	12 4 3	16 4 4	20 4 5	24 4 6	28 4 7	32 4 8	36 4 9	40 4 10
5 5 1	10 5 2	15 5 3	20 5 4	25 5 5	30 5 6	35 5 7	40 5 8	45 5 9	50 5 10
6 6 1	12 6 2	18 6 3	24 6 4	30 6 5	36 6 6	42 6 7	48 6 8	54 6 9	60 6 10
7 7 1	14 7 2	21 7 3	28 7 4	35 7 5	42 7 6	49 7 7	56 7 8	63 7 9	70 7 10
8 8 1	16 8 2	24 8 3	32 8 4	40 8 5	48 8 6	56 8 7	64 8 8	72 8 9	80 8 10
9 9 1	18 9 2	27 9 3	36 9 4	45 9 5	54 9 6	63 9 7	72 9 8	81 9 9	90 9 10
10 10 1	20 10 2	30 10 3	40 10 4	50 10 5	60 10 6	70 10 7	80 10 8	90 10 9	100 10 10

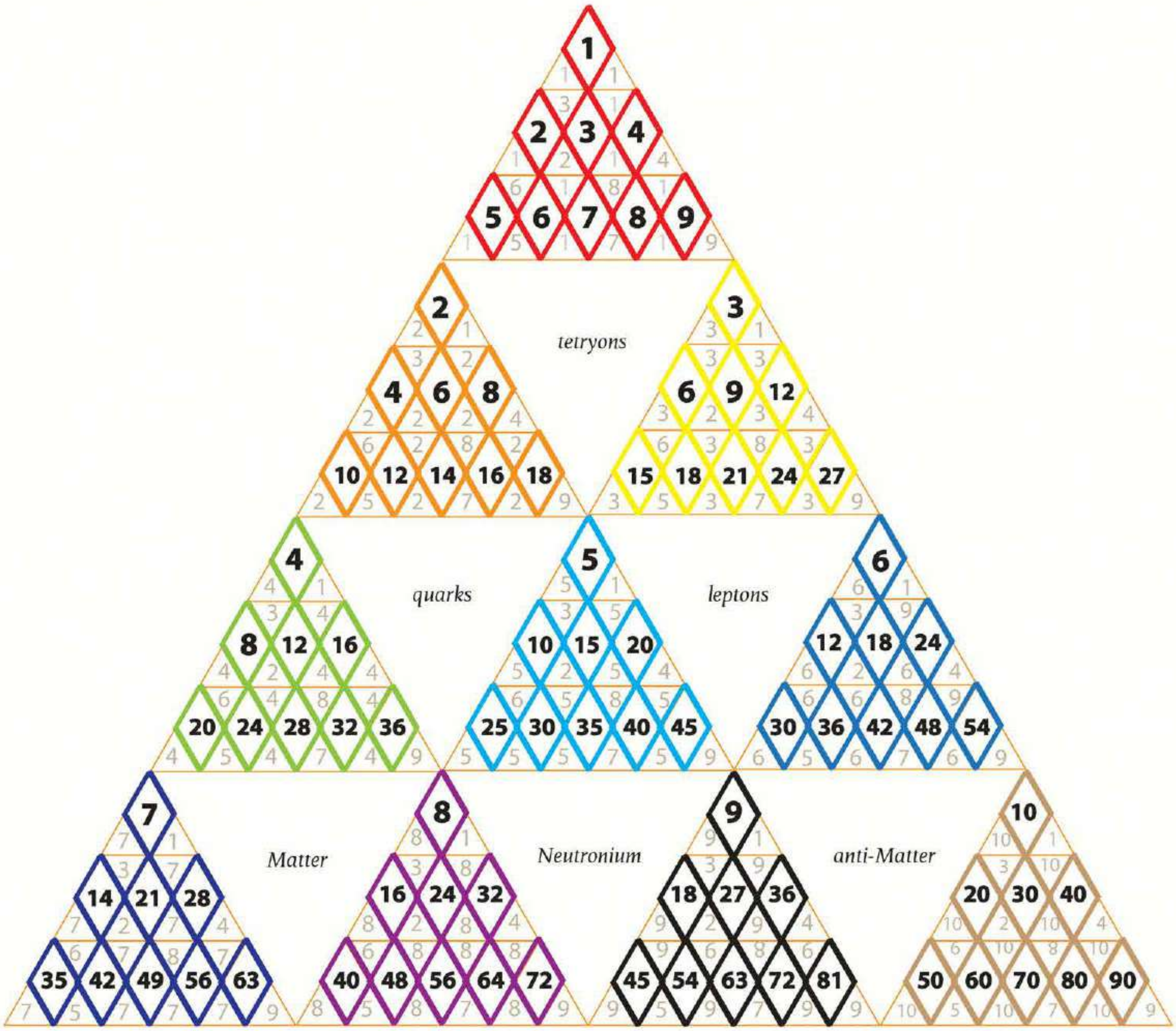

tetryons
quarks
leptons
Matter
Neutronium
anti-Matter

Tetryonics 85.07 - Tetrad Math Tables

Copyright ABRAHAM [2008] - All rights reserved

Irrational Numbers

An irrational number is defined to be any number that is the part of the real number system that cannot be written as a complete ratio of two integers

One well known irrational number is π

Irrational numbers can be easily represented geometrically

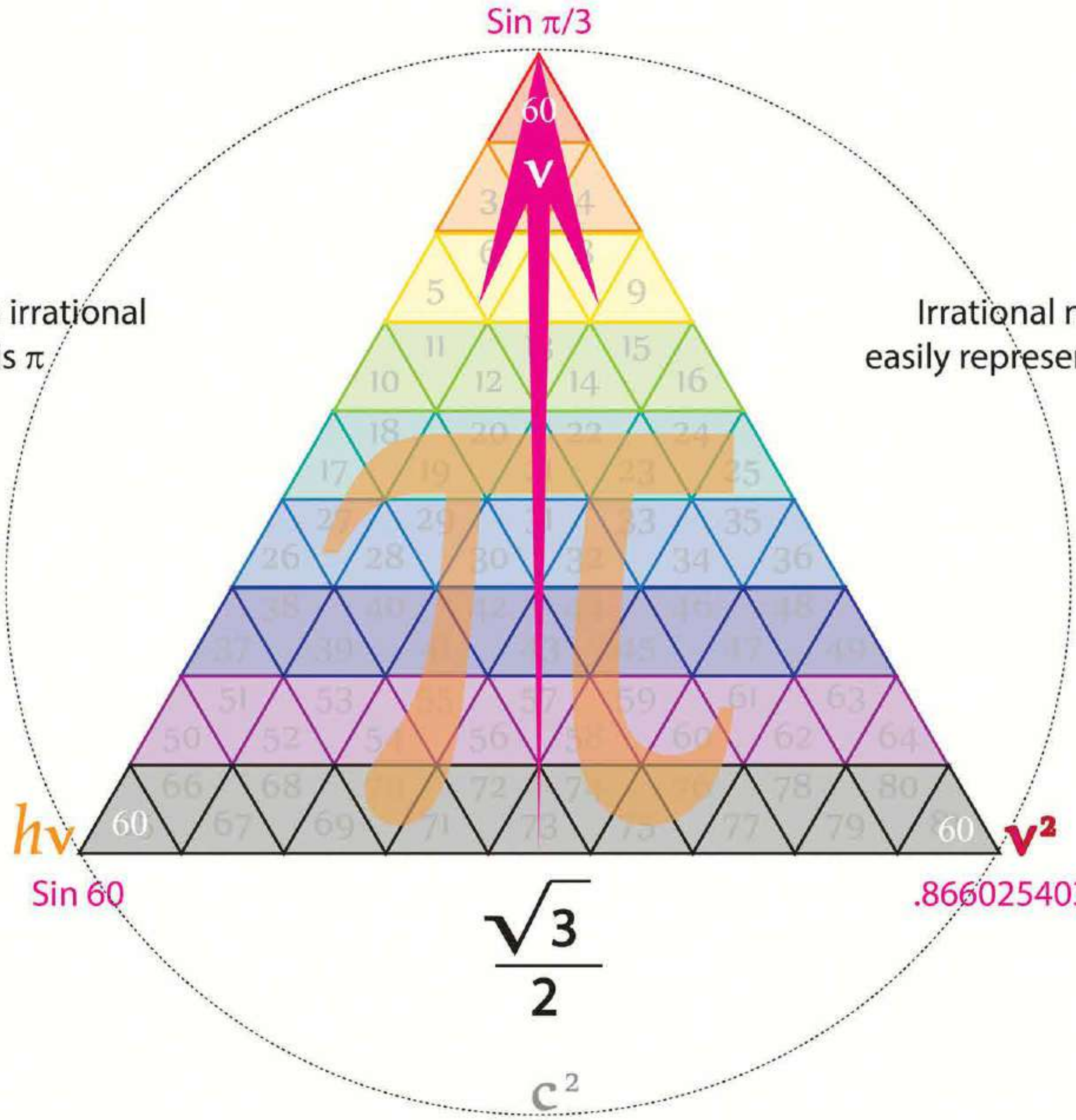
Sin π/3
60
v
π
hv
60
Sin 60
60
v²
.866025403
$\frac{\sqrt{3}}{2}$
c²

Exponentials & Logarithms

e and the Natural Log are inverse functions of each other:

e^x is the amount of continuous GROWTH after a certain amount of time.
Natural Log (ln) is the amount of TIME of continuous growth to reach a certain level

1	0
10	1
100	2
1,000	3

$10{,}000 = 10^4$

e

How much growth after x units of time
(and 100% continuous growth)

$$e^x = \lim_{n\to\infty}(1+x/n)^n,$$

GROWTH

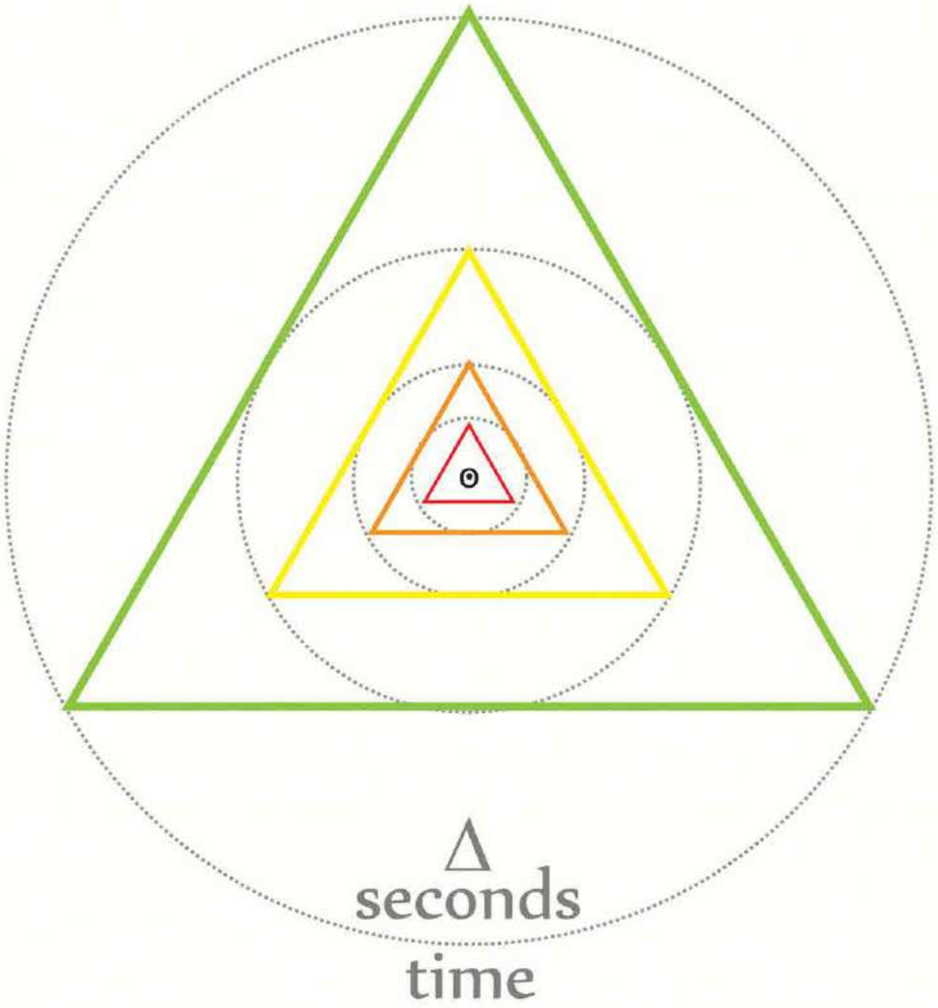

$\log(10{,}000) = 4$

ln

ln(x) lets us plug in continuous growth
and get the time it would take.

$$\ln(x) = \lim_{n\to\infty} n(x^{1/n}-1).$$

PERIOD

Exponential growth

$$\text{growth} = e = \lim_{n\to\infty}\left(1+\frac{r}{n}\right)^{t}$$

e represents the idea that all continually growing systems
are scaled versions of a common rate

GEOMETRIC

3.141592654

π

Pi is the ratio between
circumference and diameter
shared by all circles.

It is a fundamental ratio inherent in all circles
and therefore impacts any calculation of
circumference, area, volume, and surface areas

Pi radians are equally important and show
all quantised equilateral energy geometries
are related to their scribed circles

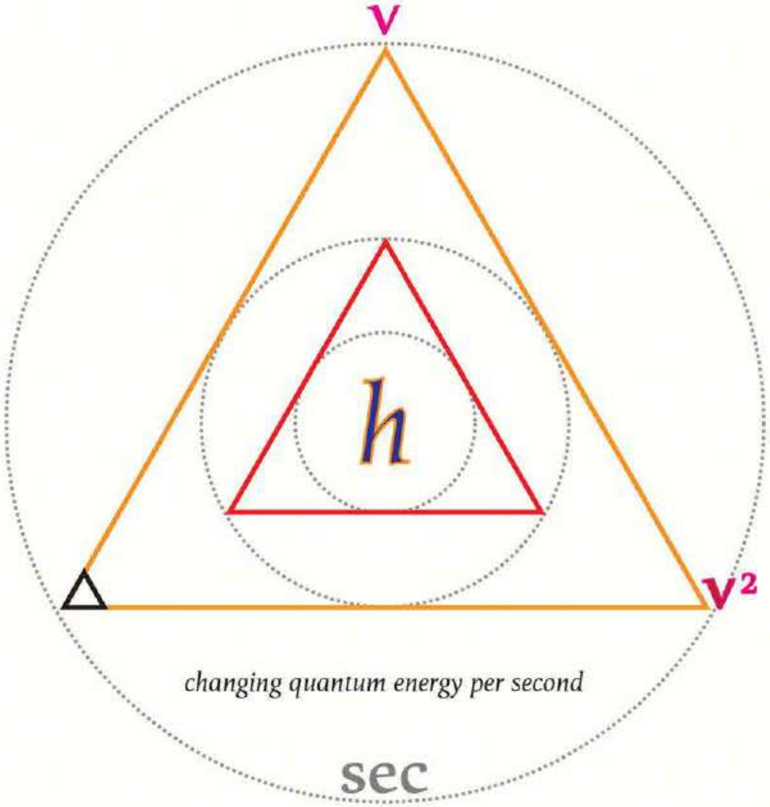

GROWTH

2.718281828

e

e is the base rate of growth
shared by all continually growing processes.

e lets you take a simple growth rate
(where all the change happens all at once
at the end of a period of time - ie quantised growth)

e shows up whenever systems grow
exponentially and continuously.....
radioactive decay, interest calculations
and populations

tetryons

EM fields

quarks

leptons

**e can be applied to the equilateral energy geometries of physical systems
only where the rate of increase is a integer factor of a squared number**

Baryons

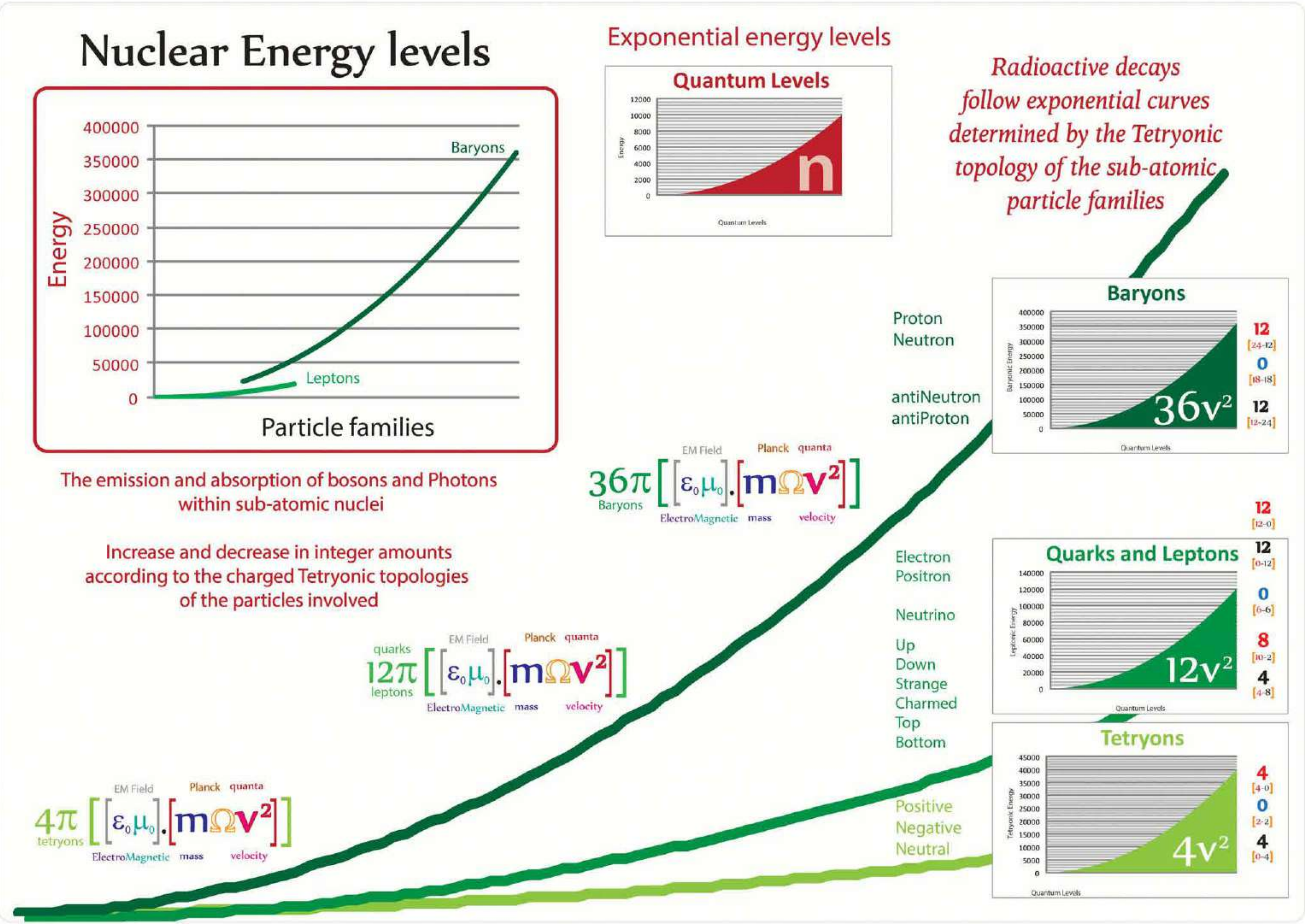

Nuclear Energy levels

The emission and absorption of bosons and Photons within sub-atomic nuclei

Increase and decrease in integer amounts according to the charged Tetryonic topologies of the particles involved

Exponential energy levels

Radioactive decays follow exponential curves determined by the Tetryonic topology of the sub-atomic particle families

Proton
Neutron
antiNeutron
antiProton
Baryons
Baryonic Energy
400000
350000
300000
250000
200000
150000
100000
50000
0
36v²
Quantum Levels
12
[24-12]
0
[18-18]
12
[12-24]
EM Field
Planck
quanta
Baryons
ElectroMagnetic
mass
velocity

$36\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

12
[12-0]
12
[0-12]
Electron
Positron
Neutrino
Up
Down
Strange
Charmed
Top
Bottom
Quarks and Leptons
Leptonic Energy
140000
120000
100000
80000
60000
40000
20000
0
12v²
Quantum Levels
0
[6-6]
8
[10-2]
4
[4-8]
quarks
leptons
EM Field
Planck
quanta
ElectroMagnetic
mass
velocity

$12\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Positive
Negative
Neutral
Tetryons
Tetryonic Energy
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
4v²
Quantum Levels
4
[4-0]
0
[2-2]
4
[0-4]
tetryons
EM Field
Planck
quanta
ElectroMagnetic
mass
velocity

$4\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Copyright ABRAHAM [2008] - All rights reserved

Series addition & the Riemann Zeta Function

The second series addition of the Reimann Zeta function is where x=2: (pi^2)/6=1+1/2^2+1/3^2+1/4^2+...
(the sum of the reciprocals of the squares)

$$\zeta(s) = \sum_{n} \frac{1}{n^s}$$

$$\sum_{n=1}^{\infty} \frac{1}{n^x}$$

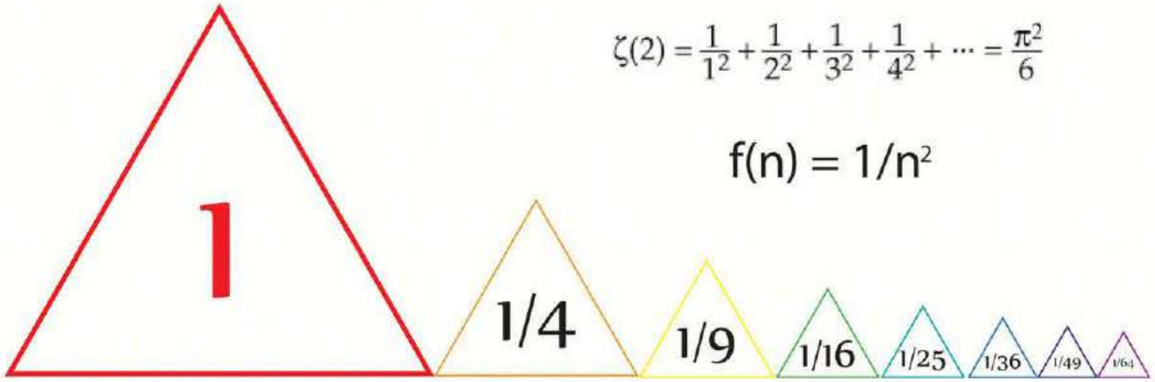

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}$$

$$f(n) = 1/n^2$$

In mathematics, the Riemann zeta function, is a prominent function of great significance in number theory. It is a named after German mathematician Bernhard Riemann. It is so important because of its relation to the distribution of prime numbers. It also has applications in other areas such as physics, probability theory, and applied statistics

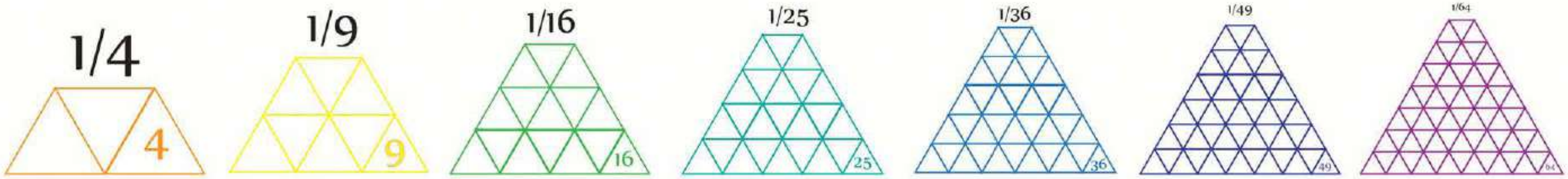

The mystery of prime numbers
Question: which natural numbers are prime? how are they distributed among natural numbers?

Primes are basic building blocks for natural numbers:

- any natural number is a product of prime numbers
- a prime number is only divisible by itself and by 1:
  (it cannot be further simplified)

We don't know how to predict where the prime numbers are:

"Prime numbers grow like weeds among the natural numbers, seeming to obey no other law than that of chance but also exhibit stunning regularity" (Don Zagier, number theorist)

Adding the odd numbers in order produces the square numbers

Apart from 2, all primes are odd numbers;
the difference between two consecutive squares being odd,
every prime can be expressed as the difference between two squares

SPECTRAL LINES

Primes can be modelled as Bosons
[ODD number energy momenta geometries]

$[2n-1]$

KEM ENERGIES

Primes can be expressed as Bosons
[the difference of two squared energies]

$n^2 - [n-1]^2$

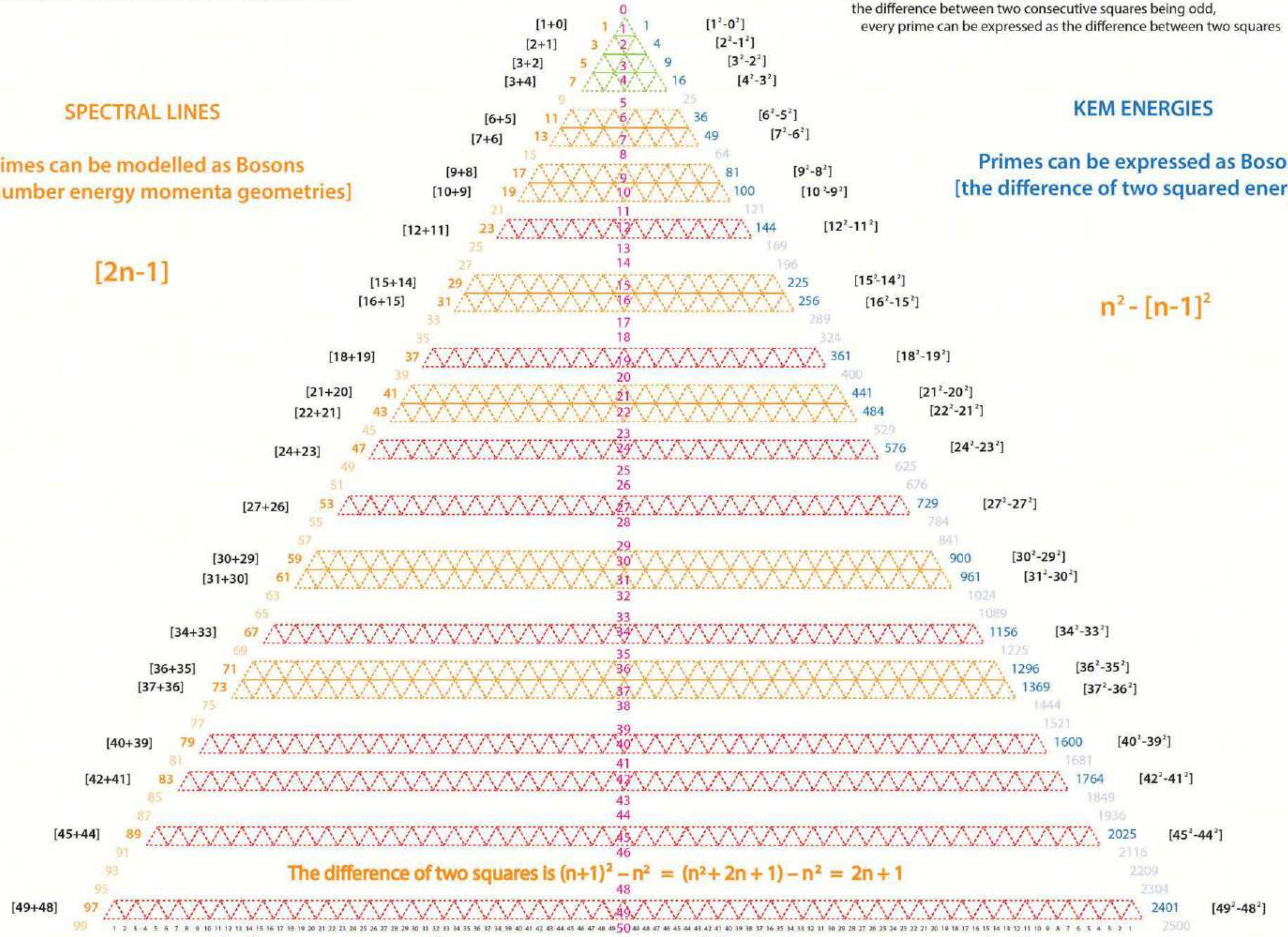

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 48 49 50
[1+0] 1
[2+1] 3
[3+2] 5
[3+4] 7
9
[6+5] 11
[7+6] 13
15
[9+8] 17
[10+9] 19
21
[12+11] 23
25
27
[15+14] 29
[16+15] 31
33
35
[18+19] 37
39
[21+20] 41
[22+21] 43
45
[24+23] 47
49
51
[27+26] 53
55
57
[30+29] 59
[31+30] 61
63
65
[34+33] 67
69
[36+35] 71
[37+36] 73
75
77
[40+39] 79
81
[42+41] 83
85
87
[45+44] 89
91
93
95
[49+48] 97
99
1 $[1^2-0^2]$
4 $[2^2-1^2]$
9 $[3^2-2^2]$
16 $[4^2-3^2]$
25
36 $[6^2-5^2]$
49 $[7^2-6^2]$
64
81 $[9^2-8^2]$
100 $[10^2-9^2]$
121
144 $[12^2-11^2]$
169
196
225 $[15^2-14^2]$
256 $[16^2-15^2]$
289
324
361 $[18^2-19^2]$
400
441 $[21^2-20^2]$
484 $[22^2-21^2]$
529
576 $[24^2-23^2]$
625
676
729 $[27^2-27^2]$
784
841
900 $[30^2-29^2]$
961 $[31^2-30^2]$
1024
1089
1156 $[34^2-33^2]$
1225
1296 $[36^2-35^2]$
1369 $[37^2-36^2]$
1444
1521
1600 $[40^2-39^2]$
1681
1764 $[42^2-41^2]$
1849
1936
2025 $[45^2-44^2]$
2116
2209
2304
2401 $[49^2-48^2]$
2500

The difference of two squares is $(n+1)^2 - n^2 = (n^2 + 2n + 1) - n^2 = 2n + 1$

$[n-1]$ all energy quanta create normal distributions $[n-1]$

Copyright ABRAHAM [2008] - All rights reserved

Prime number Distributions

A twin prime is a prime number that differs from another prime number by two, for example the twin prime pair (41, 43). Sometimes the term twin prime is used for a pair of twin primes; an alternative name for this is prime twin or prime pair.

Twin primes appear despite the general tendency of gaps between adjacent primes to become larger as the numbers themselves get larger due to the prime number theorem (the "average gap" between primes less than n is log(n)).

$$p_2 - p_1 = 2$$

Every twin prime pair except (3, 5) is of the form (6n – 1, 6n + 1) for some natural number n, and with the exception of n = 1, n must end in 0, 2, 3, 5, 7, or 8

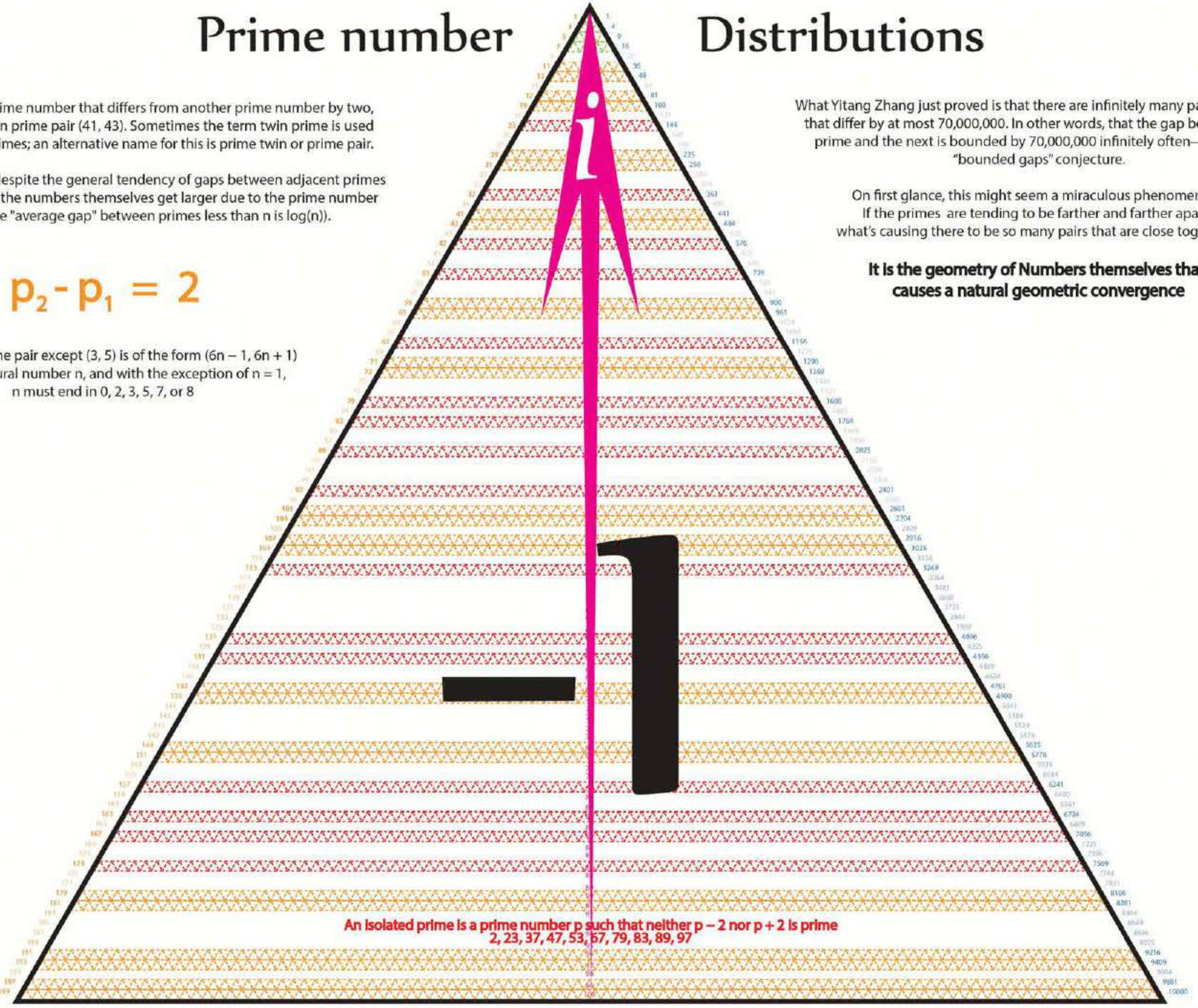

An isolated prime is a prime number p such that neither p – 2 nor p + 2 is prime
2, 23, 37, 47, 53, 67, 79, 83, 89, 97

What Yitang Zhang just proved is that there are infinitely many pairs of primes that differ by at most 70,000,000. In other words, that the gap between one prime and the next is bounded by 70,000,000 infinitely often—thus, the "bounded gaps" conjecture.

On first glance, this might seem a miraculous phenomenon. If the primes are tending to be farther and farther apart, what's causing there to be so many pairs that are close together?

It is the geometry of Numbers themselves that causes a natural geometric convergence

Copyright ABRAHAM [2008] - All rights reserved

The Digital roots of Prime numbers

The digital root (also repeated digital sum) of a number is the (single digit) value obtained by an iterative process of summing digits, on each iteration using the result from the previous iteration to compute a digit sum.

The process continues until a single-digit number is reached.

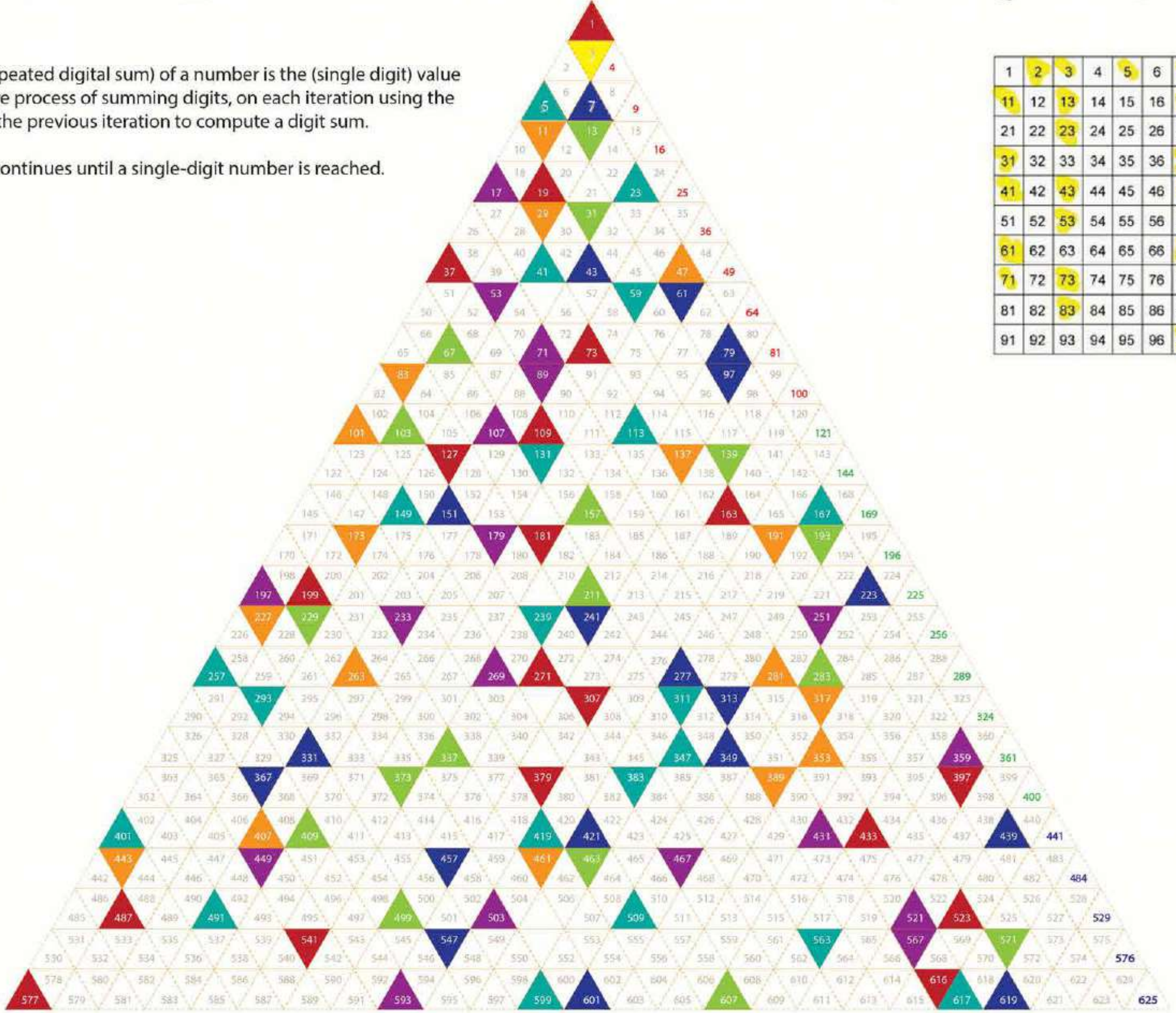

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Tetryonics 86.08 - Digital roots of Primes

Archimedes is given credit for the first calculus.

Archimedes infinite series

Today's calculus was published by Newton.

1 + 3/4 + 3/16 + 3/64 + 3/256 + 3/1024 + ...

1 + .75 + .1875 + .046875 + .01171875 + ...

$$3\left(\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3}+\frac{1}{4^4}+\cdots\right)=1.$$

$$\frac{3}{4}+\frac{3}{4^2}+\frac{3}{4^3}+\frac{3}{4^4}+\cdots=1.$$

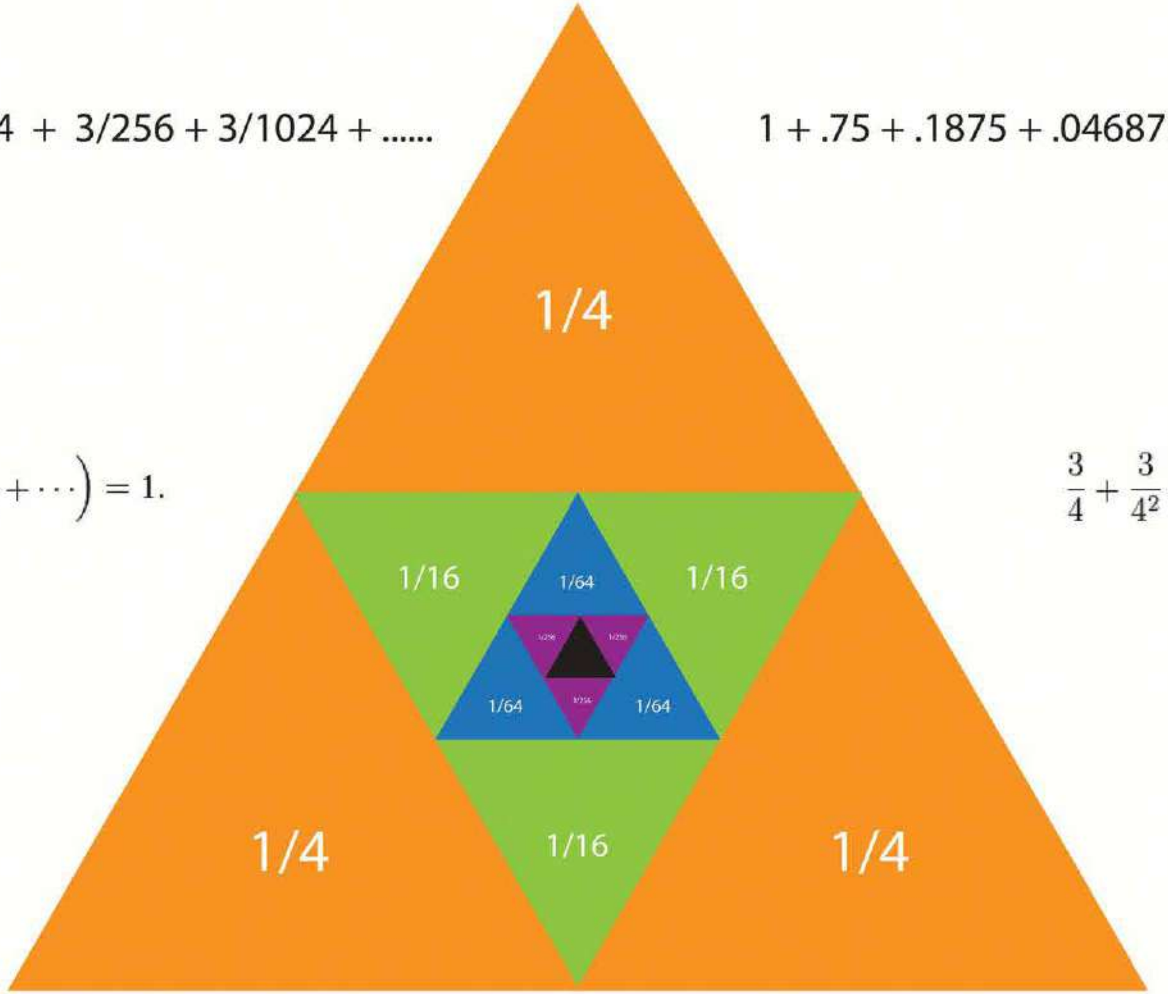
1/4
1/16
1/64
1/16
1/256
1/256
1/64
1/256
1/64
1/4
1/16
1/4

Nested convergent infinite series

Summing a convergent infinite series

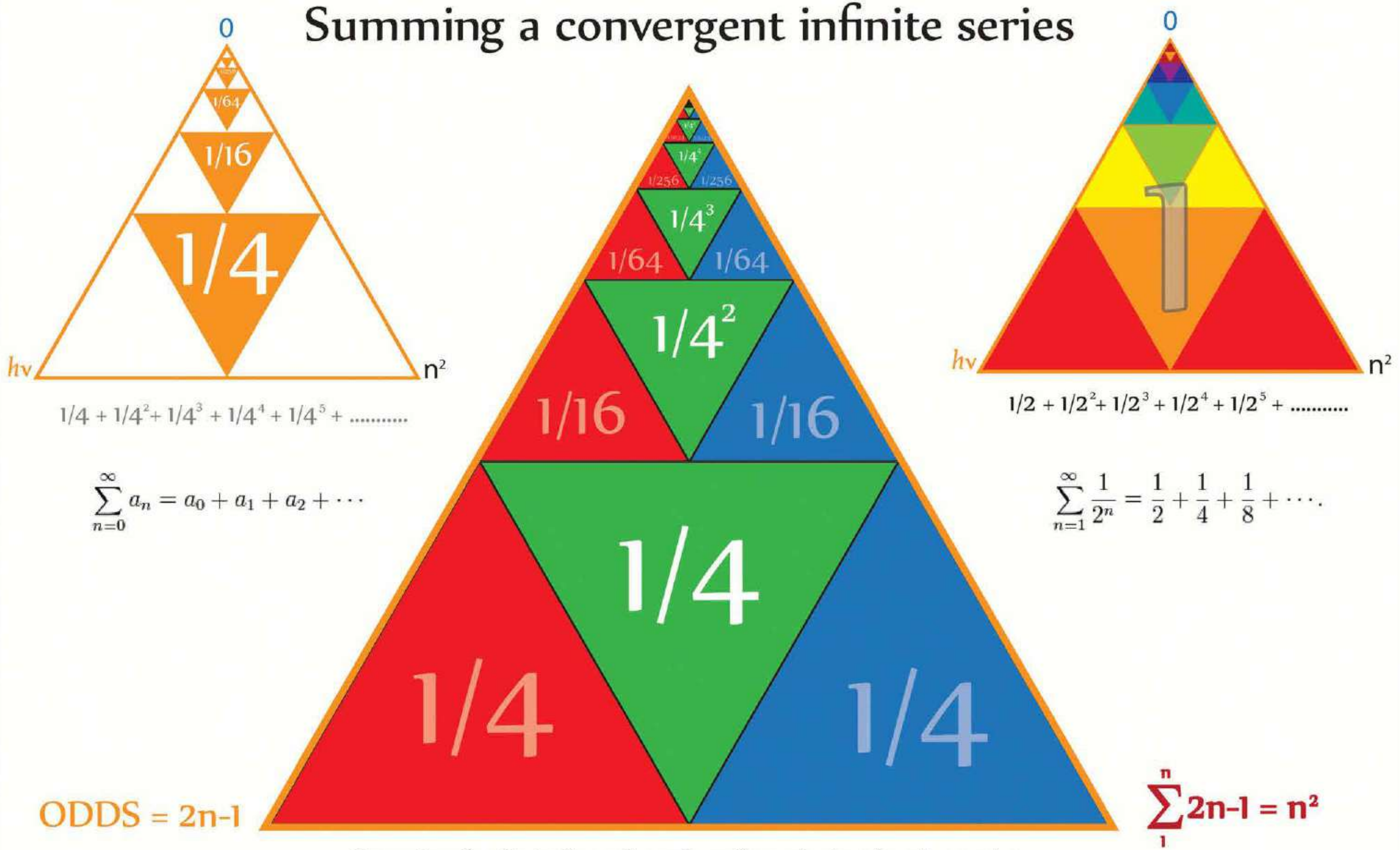
0
1/64
1/16
1/4
hv
n²
0
1/4⁴
1/256
1/256
1/4³
1/64
1/64
1/4²
1/16
1/16
1/4
1/4
1/4
0
1
hv
n²

1/4 + $1/4^2$ + $1/4^3$ + $1/4^4$ + $1/4^5$ + ...

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots$$

1/2 + $1/2^2$ + $1/2^3$ + $1/2^4$ + $1/2^5$ + ...

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

ODDS = 2n-1

$$\sum_{1}^{n} 2n-1 = n^2$$

Summing the dissimilar coloured equilateral triangles gives unity

1/3 1/3 1/3

Copyright ABRAHAM [2008] - All rights reserved

1/3 1/3 1/3

Summing the dissimilar coloured equilateral triangles gives unity

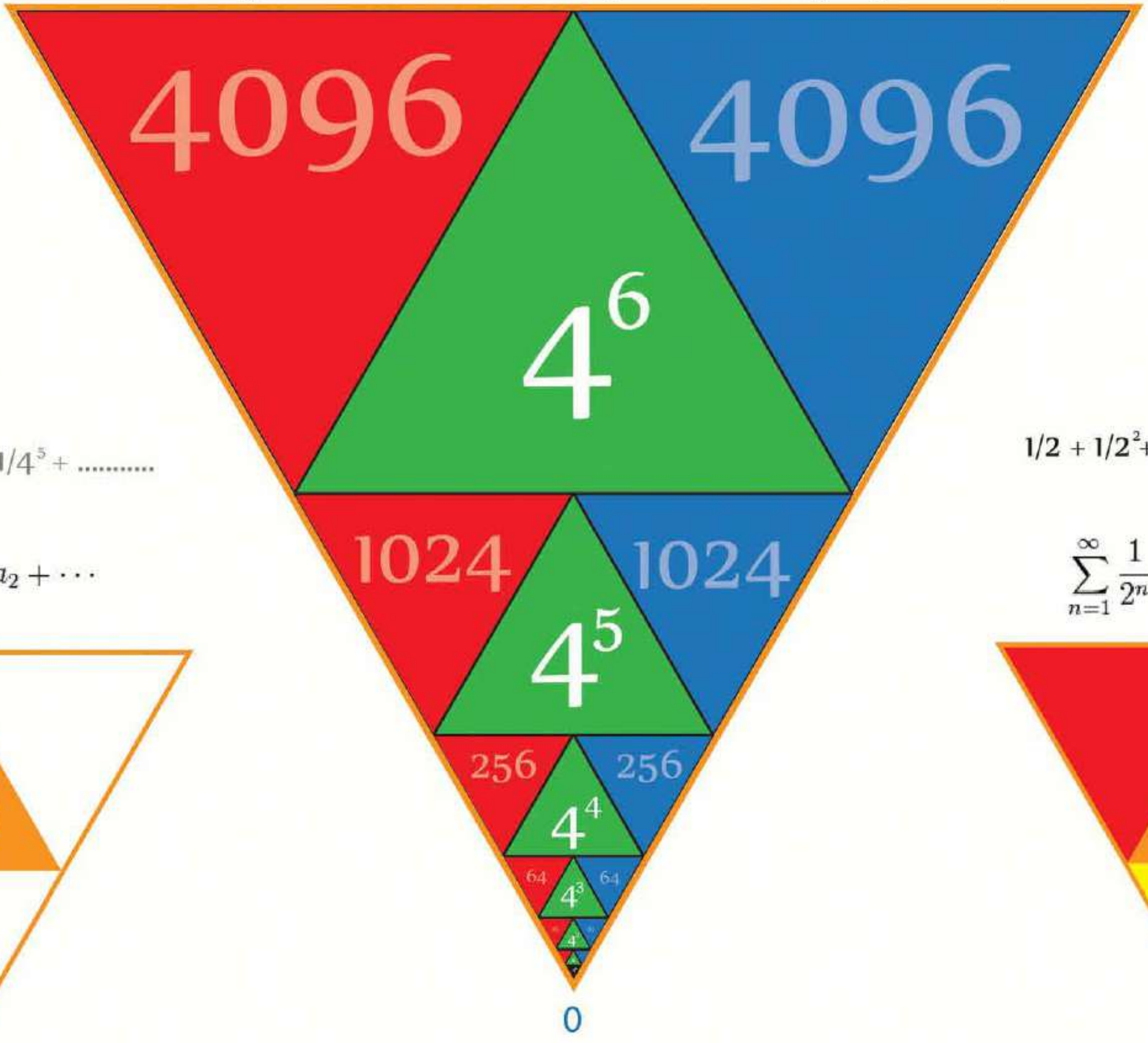
4096
4096
4^6
1024
1024
4^5
256
256
4^4
64
64
4^3
0

1/4 + 1/4^2 + 1/4^3 + 1/4^4 + 1/4^5 + ...

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots$$

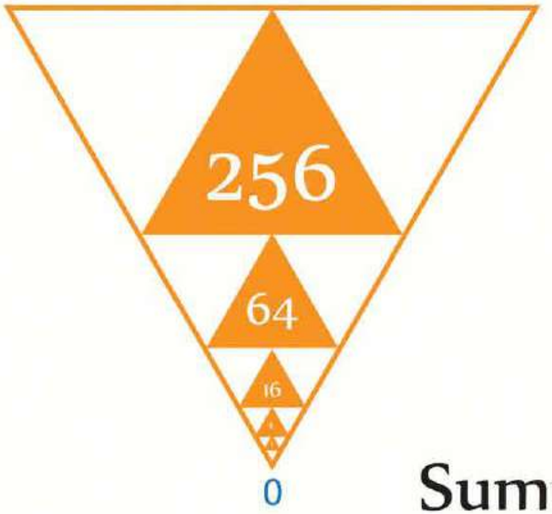

1/2 + 1/2^2 + 1/2^3 + 1/2^4 + 1/2^5 + ...

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

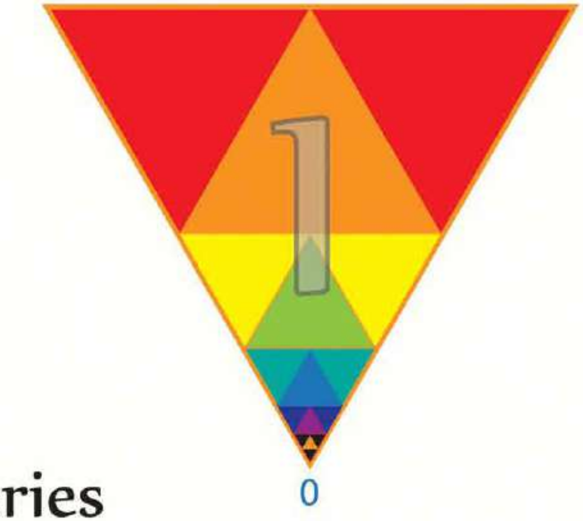

Summing a divergent infinite series

Copyright ABRAHAM [2008] - All rights reserved

Pietro Mengoli

(1626 – June 7, 1686)

The Basel Problem

The Basel problem is a famous problem in mathematical analysis with relevance to number theory, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735.

Leonhard Euler

(15 April 1707 – 18 Sept 1783)

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6} \approx 1.644934$$

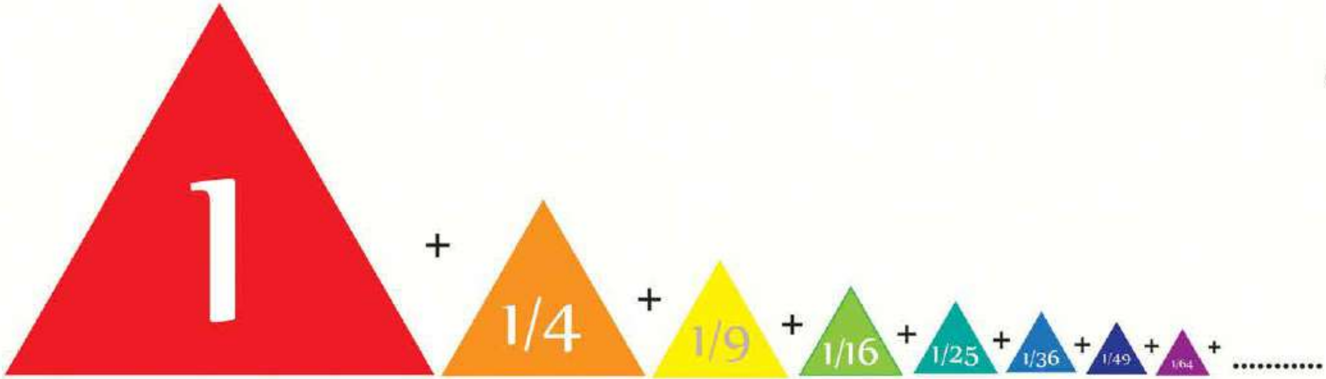

The Basel problem asks for the precise summation of the reciprocals of the squares of the natural numbers, i.e. the precise sum of the infinite series

$$f(n) = 1/n^2$$

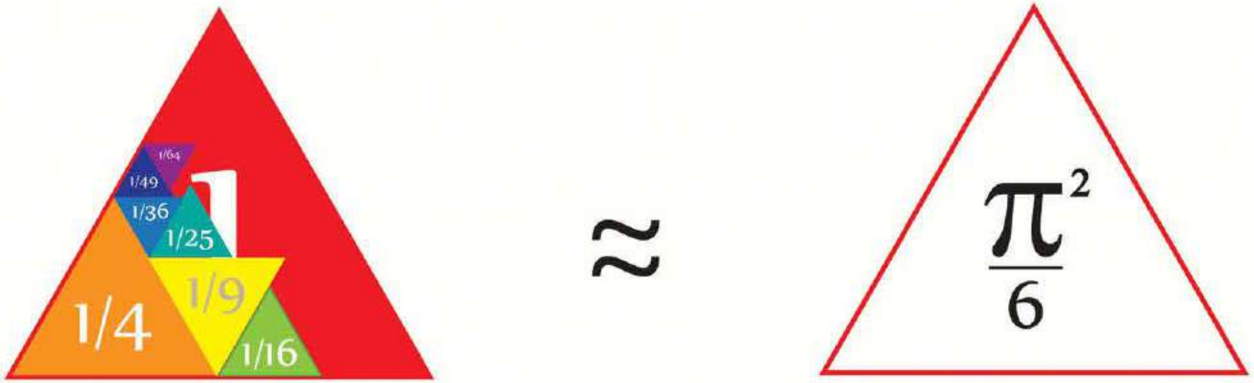

Integrals of mass-energy

is a means of finding scalar areas using summation and limits.

Integration is a micro adding of ***CONTINUOUS*** quantities.

$$\int_{n}^{\infty} f(x)$$

Integration is a special case of summation.

Integration is defined as the limit of a summation as the number of elements approches infinity while a part of their respective value approaches zero.

Summation is the finite sum of multiple, fixed values.

$$\sum_{n}^{\infty} x$$

Summation is a macro adding of ***DISCRETE*** quantities.

The summation of equilateral energy momenta quanta with respect to their linear vector components

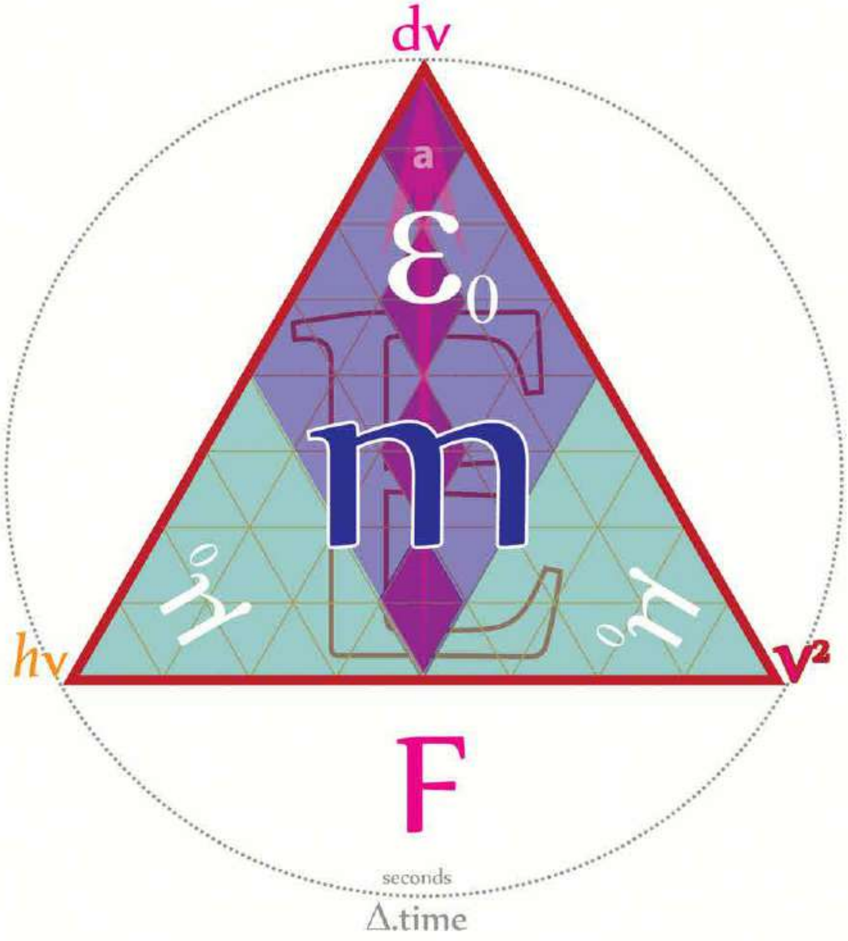

Integrating the energy quanta contained within equilateral charge geometry gives the variable Force required to acheive changes in motion [Energy, work, acceleration]

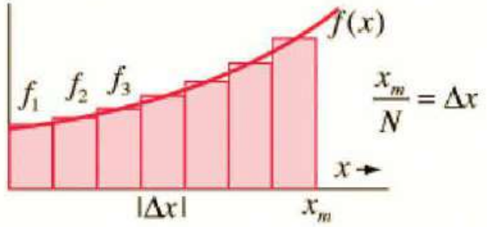

The Integral of the continuous area under the curve is the summation of an infinte number of disctrete rectangular measurements made to a specified limit

An integration isn't a simple summation, but the limit of a sequence of summations

All Planck energies are discrete equilateral geometries

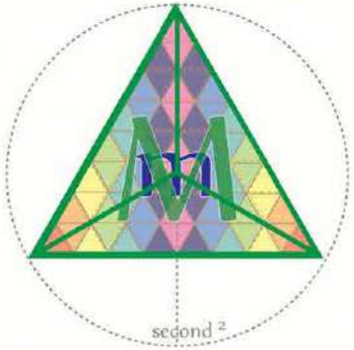

mass is the surface integral of EM energy geometries per unit of time

Copyright ABRAHAM [2008] - All rights reserved

"The calculus of infinitesimals"

The fundamental theorem of calculus
simply states that the sum of infinitesimal changes in a quantity over time
adds up to the net change in the quantity.

Scalar Energy

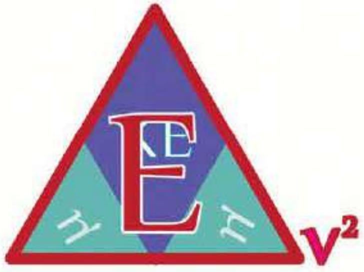

Leibniz's vis viva
(Latin for living force)
is mv2,

[2 x kinetic energy]

$E = mv^2$

Gottfried Wilhelm von Leibniz

(July 1, 1646 – November 14, 1716)

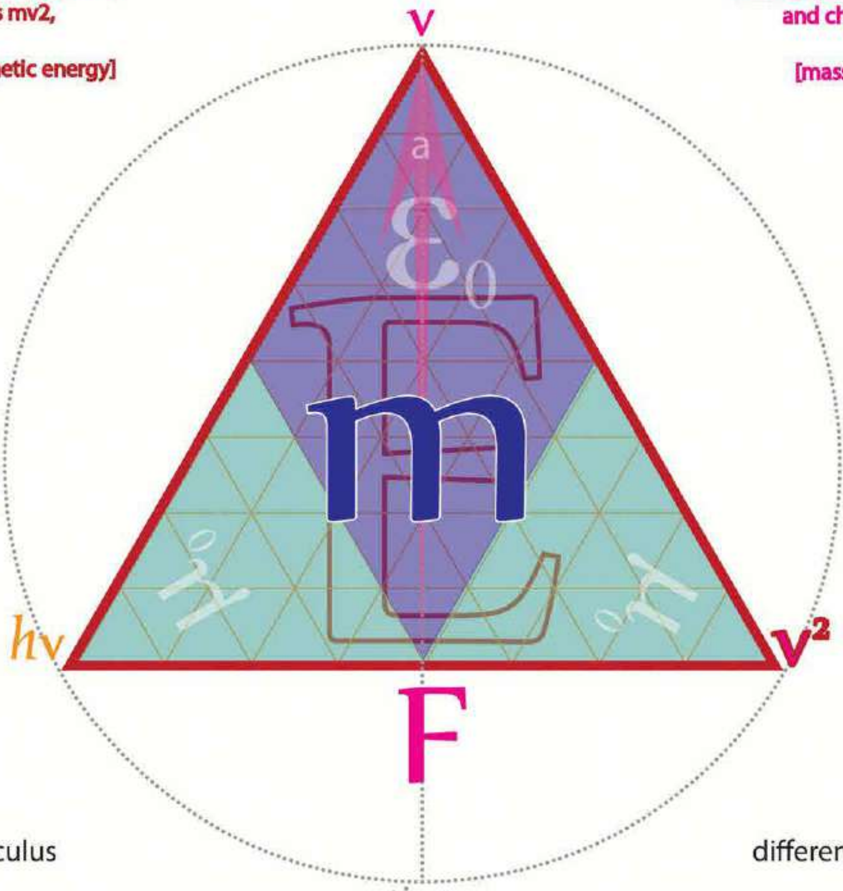

vector Force

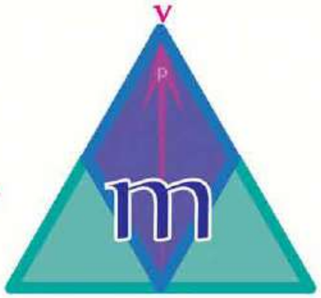

Much of Newton's work
centred around momentum
and changes to it

[mass.velocity]

$F = m[dv/dt]$

Sir Isaac Newton

(1643-1727)

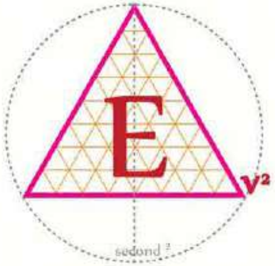

integral calculus

The founders of calculus thought
of the integral as an infinite sum of
rectangles of infinitesimal width

The fundamental theorem of calculus is a theorem that links the concept of the integral with the derivative of a function.

differential calculus

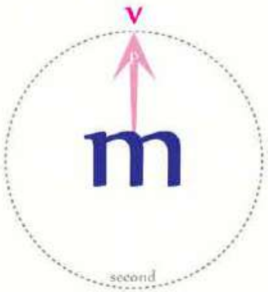

In calculus, a branch of mathematics,
the derivative is a measure of how
a function changes as its input changes

Copyright ABRAHAM [2008] - All rights reserved

Leibniz

scalar energies

linear momentum is the square root vector force of the scalar energy required to do a set amount of work

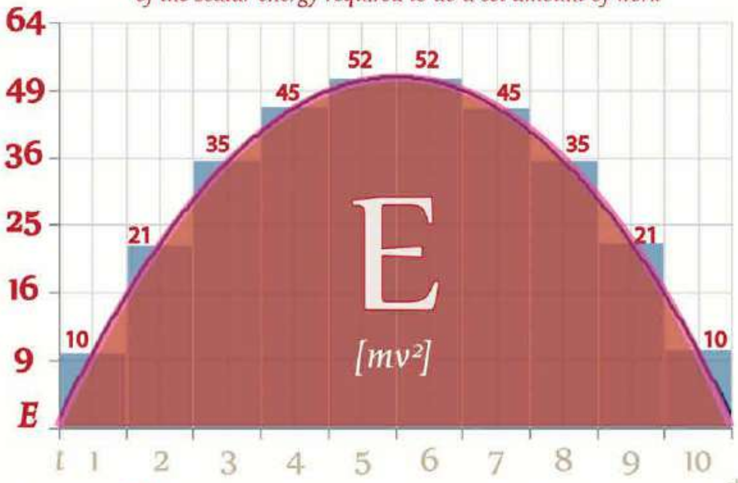

Energy momentum geometries

$$E = p^2$$

Energy momentum

$$kg\left[\frac{m}{s}\right]^2$$

print out this page

then cut out the 10 Planck energy momenta triangles shown

slice fine lines along the pink linear momentum arrows with a razor blade or similar as highlighted under the Newtonian acceleration curve

Newton

linear momentum

$$d[mv]/dt = dp/dt = ma$$

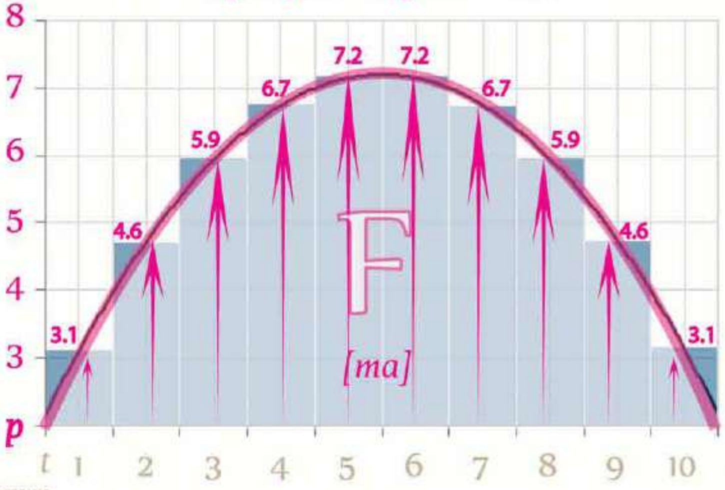

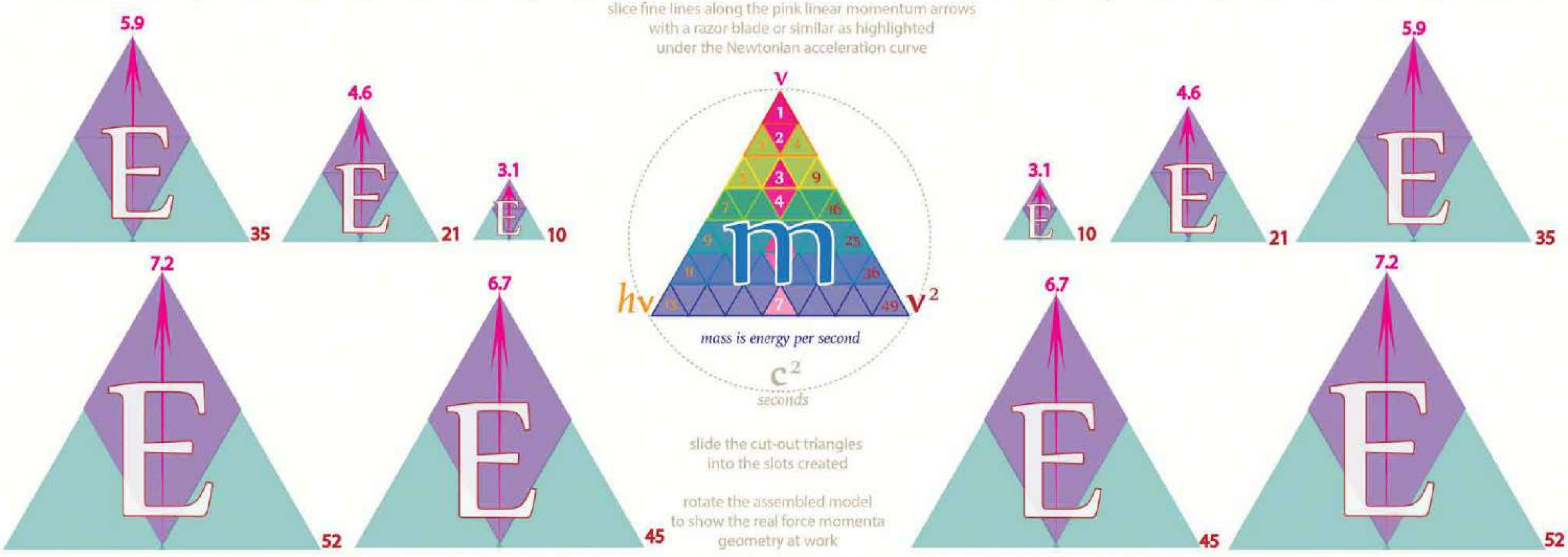
5.9
E
35
4.6
E
21
3.1
E
10
7.2
E
52
6.7
E
45
v
1
2
3
4
m
7
hv
49
v²
c²
seconds
3.1
E
10
4.6
E
21
5.9
E
35
6.7
E
45
7.2
E
52

mass is energy per second

slide the cut-out triangles into the slots created

rotate the assembled model to show the real force momenta geometry at work

Tetryonics 87.07 - Physically modeling the geometric forces of acceleration in calculus

Copyright ABRAHAM [2008] - All rights reserved

linear momentum is the square root vector force of the scalar energy required to do a set amount of work

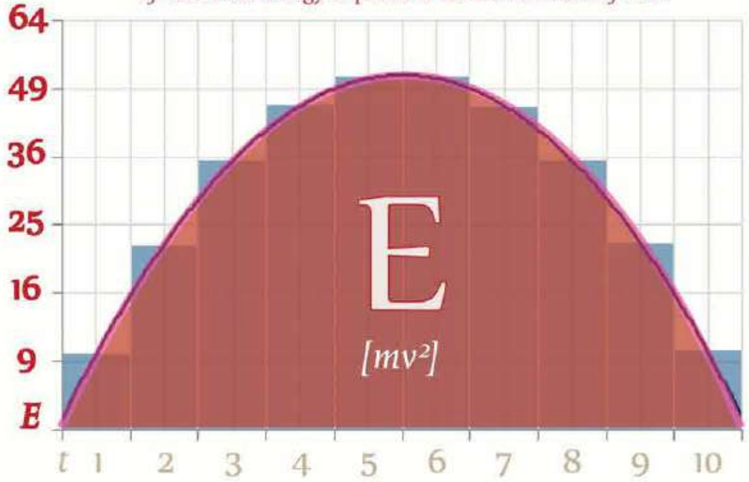

p

linear momentum

$kg\frac{m}{s}$

print out this page

$d[mv]/dt = dp/dt = ma$

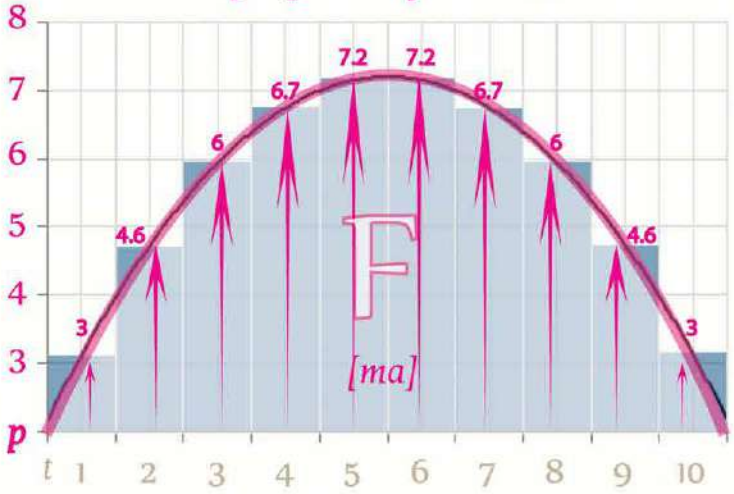

Leibniz

scalar energies

$$E = mv^2$$

mass-energy momenta relationship

square root equilateral energy is linear momentum

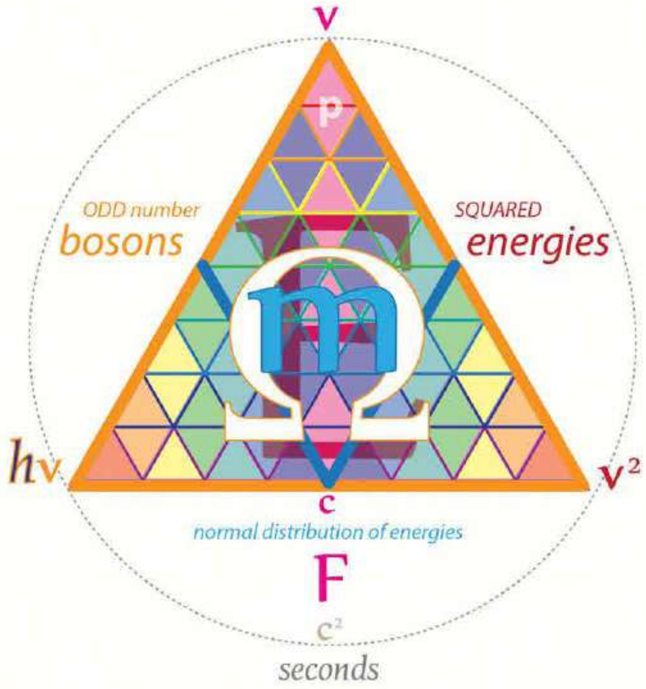
v
p
ODD number
bosons
SQUARED
energies
m
Ω
hv
v²
c
normal distribution of energies
F
c²
seconds

Newton

linear momentum

$$F = ma$$

mass is a scalar constant relating Force to acceleration

$$F = m\frac{\Delta v}{\Delta t}$$

Planck quanta and their vector linear momentum lie at 90 degrees to the angle use in the graphs of motion in calculus

Differentiation

Differentiation is concerned with things like speeds and accelerations, slopes and curves etc.
These are Rates of Change, they are things that are defined locally.

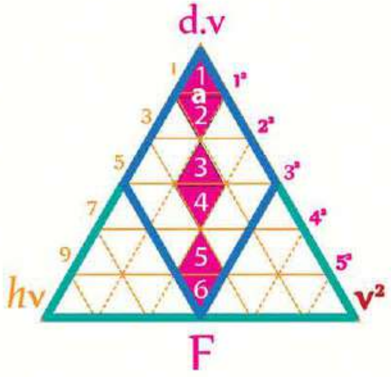
d.v
1
a
2
3
4
5
6
1
3
5
7
9
1²
2²
3²
4²
5²
hv
v²
F

An increase in a force opposing an object's vector velocity results in ***DECCELERATION***

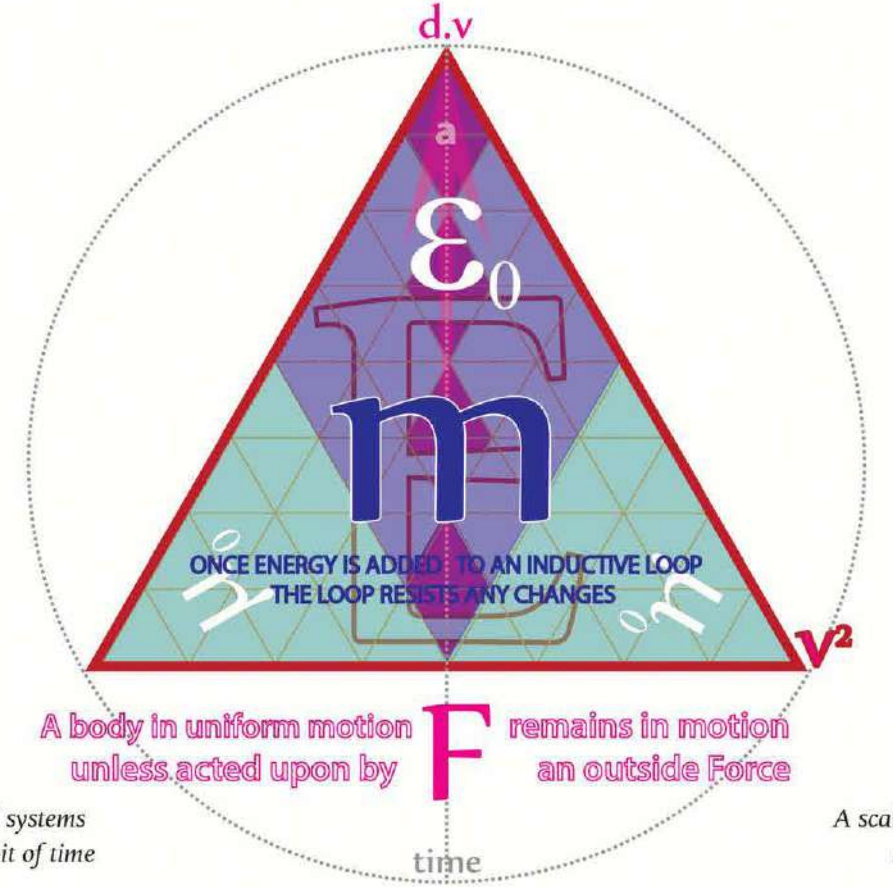
d.v
a
ε_0
m
ONCE ENERGY IS ADDED TO AN INDUCTIVE LOOP
THE LOOP RESISTS ANY CHANGES
v²
time

A body in uniform motion remains in motion unless acted upon by an outside Force F

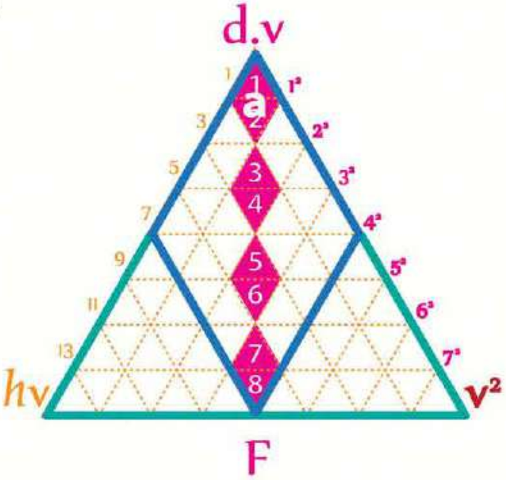
d.v
1
a
2
3
4
5
6
7
8
1
3
5
7
9
11
13
1²
2²
3²
4²
5²
6²
7²
hv
v²
F

An increase in a force in line with an object's vector velocity results in ***ACCELERATION***

A linear measure of Forces acting on physical systems resulting in changes to distance covered per unit of time

A scalar measure of Forces acting on physical systems resulting in changes to their rate of motion

$$v = \frac{\Delta d}{\Delta t}$$

$$= m/s$$

$$F = ma$$

In physics, the derivative of the displacement of a moving body with respect to time is the velocity of the body, and the derivative of velocity with respect to time is acceleration.

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \mathbf{v}\frac{\mathrm{d}m}{\mathrm{d}t}$$

Newton's second law of motion states that the derivative of the momentum of a body equals the force applied to the body.

$$a = \frac{\Delta v}{\Delta t}$$

$$= m/s^2$$

Visualising the geometric half-truths of relativistic physics

The source of all the physical relationships of mass-energy momenta and the constants in Physics is the Equilateral Triangle (and all texts must be corrected)

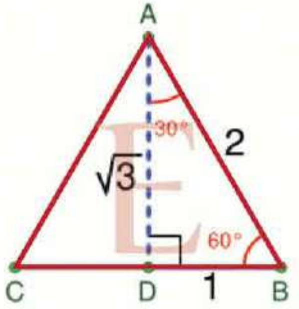

Energy geometries within Physics including Special Relativity with its Lorentz corrections have historically been incorrectly illustrated through the geometry of right angled triangles

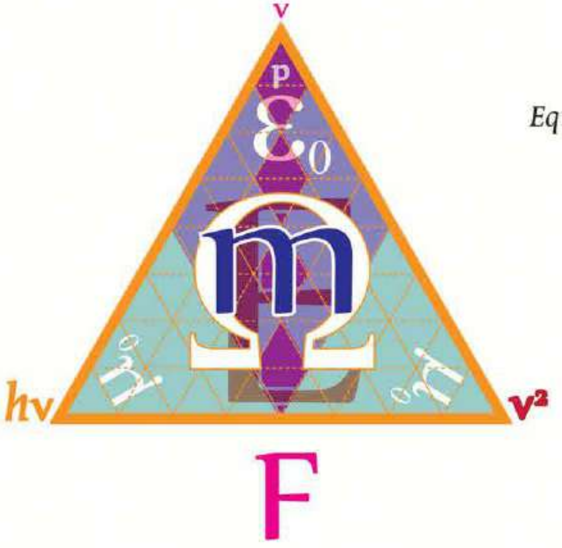

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = m\mathbf{a}$$

$$\mathbf{F} = m\mathbf{a}\,.$$

Physics is geometry,
one cannot be separated from the other

Equilateral geomtries lead to a intuitive understandings of Physics, Chemistry, Electrodynamics and Gravitation along with all the other apsects of Nature.

$$mv^2 = E = hv^2$$

ENERGY

6.629432673 e-34 J

Planck quanta

$$[m\Omega v^2]$$

mass velocity

7.376238634 e-51 kg

momentum

$$p^2 = mv^2$$

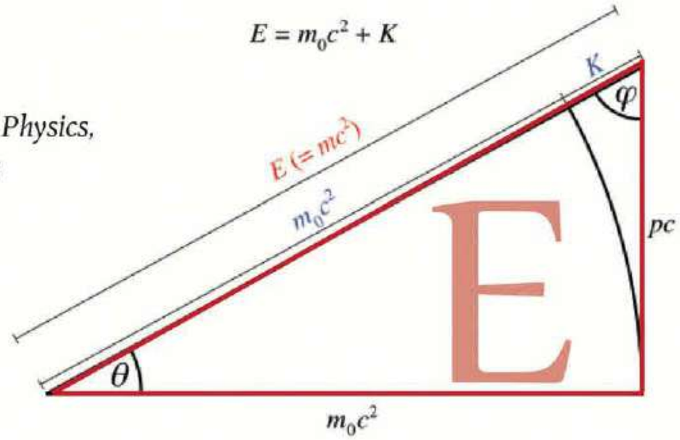

$$E^2 = p^2c^2 + m_0^2c^4$$

Generalizing, we see that the square of the total energy, mass, or distance in spacetime is the sum of the components squared.

We can see an origin of distance in spacetime relating to velocity in pc in which Energy is subject to Lorentz corrections [v/c]

$$E = pc.$$

Additionally, EM mass can be directly related tot the Energy content of a body by the velocity of Energy

$$E = mc^2$$

Velocity

In physics, velocity is the measurement of the rate and direction of change in the position of an object.

It is a vector physical quantity; both magnitude and direction are required to define it.

The scalar absolute value (magnitude) of velocity is speed, a quantity that is measured in metres per second (m/s or ms–1) when using the SI (metric) system.

$$\frac{m}{s}$$

Speed is the scalar value of the Distance traveled per unit of TIme

$$\bar{\mathbf{v}} = \frac{\Delta \mathbf{x}}{\Delta t}.$$

Velocity is the vector value of the Distance traveled per unit of Time

$$\frac{\vec{m}}{s}$$

Velocity squared is the scalar value of the Distance traveled per unit of Time squared (Energy of a given spatial volume)

$$\frac{m^2}{s^2}$$

v Velocity $\frac{m}{s}$

sec

is a 2D
radial Space-time
MEASUREMENT

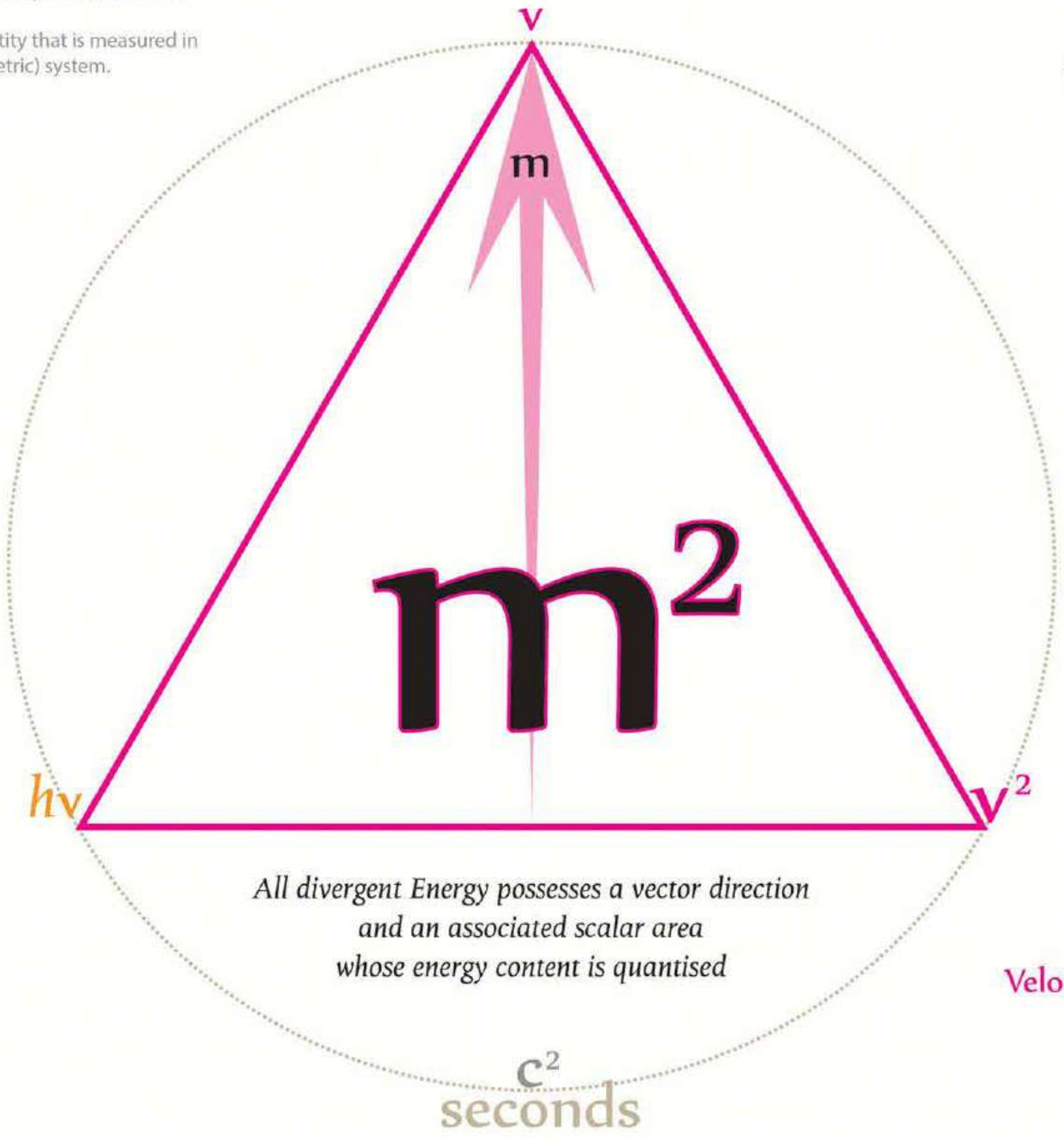

All divergent Energy possesses a vector direction
and an associated scalar area
whose energy content is quantised

Velocity squared

$$\frac{m^2}{s^2}$$

Copyright ABRAHAM [2008] - All rights reserved

is a 3D
Spherical Space-time
MEASUREMENT

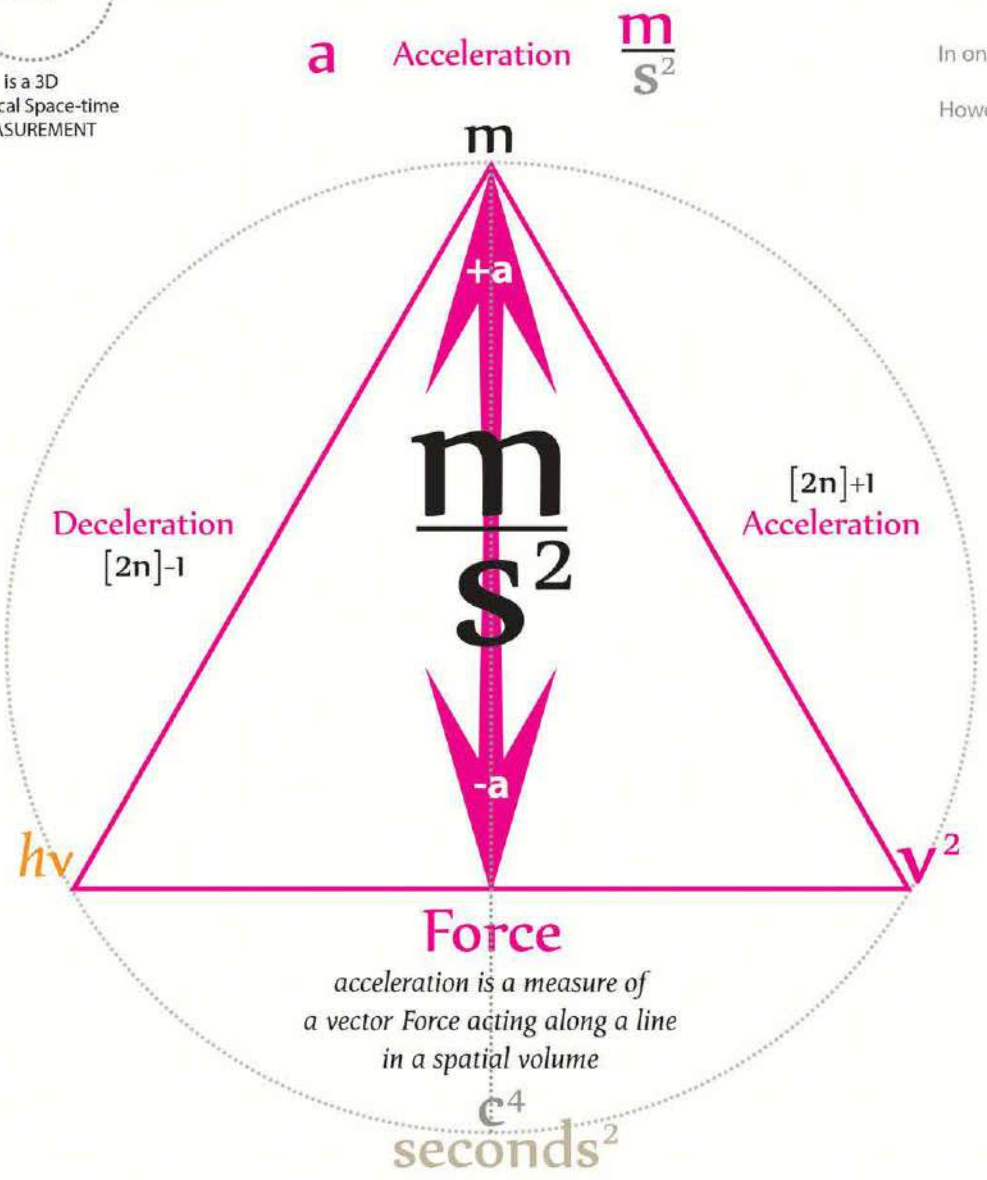

acceleration is a measure of
a vector Force acting along a line
in a spatial volume

Acceleration

In physics, acceleration is the rate of change of velocity over time [dt]

In one dimension, acceleration is the rate at which something speeds up or slows down.

However, since velocity is a vector, acceleration describes the rate of change of both the magnitude and the direction of velocity.

Acceleration has the dimensions [Length]/[Time Squared]
In SI units, acceleration is measured in meters per second squared (m/s^2).

$$a = \frac{\Delta y}{\Delta x} = \frac{\Delta v}{\Delta t}.$$

In classical mechanics, for a body with constant mass,
the acceleration of the body is proportional to the net force acting on it
(Newton's second law)

$$F=ma \longrightarrow a=F/m$$

Force

$$kg\,\frac{m}{s^2}$$

Additionally, for a mass with constant velocity,
(ie in an inertial frame)
the energy of motion is expressed as its momentum
(acceleration causes changes in Energy-momentum)

$$p = kg\,\frac{m}{s}$$

$\mathrm{kg}\,\frac{\mathrm{m}}{\mathrm{s}}$

linear
momentum

p

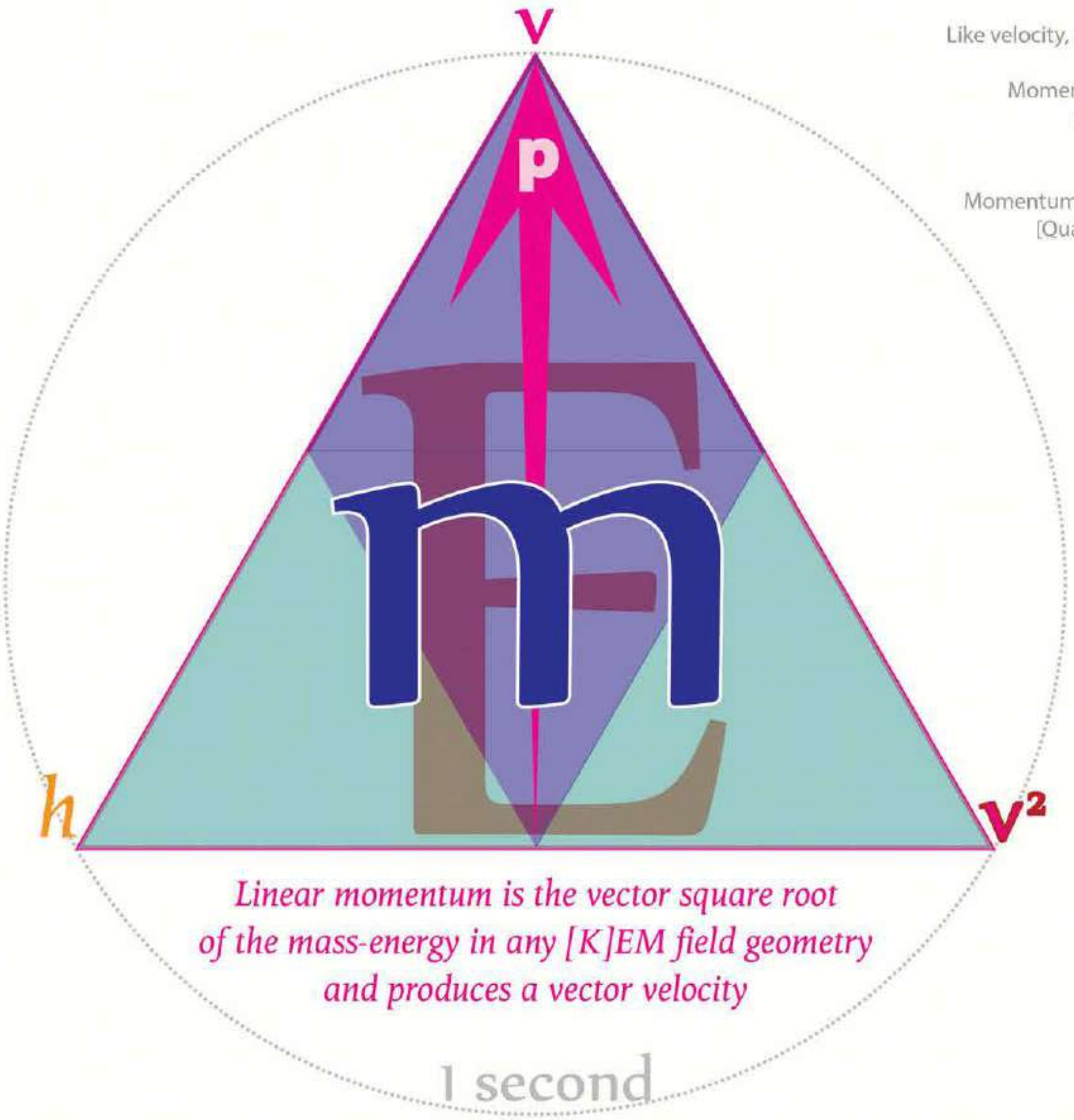

Linear momentum is the vector square root of the mass-energy in any [K]EM field geometry and produces a vector velocity

Momentum

In classical mechanics, momentum (pl. momenta; SI unit kg·m/s, or, equivalently, N·s) is the product of the mass and velocity of an object (p).

Like velocity, momentum is a vector quantity, possessing a direction as well as a magnitude.

Momentum is a conserved quantity (law of conservation of linear momentum), meaning that if a closed system is not affected by external forces, its total momentum cannot change.

Momentum should be referred to in its specific forms to distinguish it in its various forms [Quantised Angular, Linear, Rotational and quantum/nuclear momentum]

$$\boxed{p = \frac{hv^2}{v} = mv}$$

Although originally expressed in Newton's Second Law, the conservation of momentum also holds in special relativity and, with appropriate definitions, a (generalized) momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory, and general relativity.

In relativistic mechanics, non-relativistic momentum is further multiplied by the Lorentz factor.

$$p^2 = E = mv^2$$

Energy can be expressed as the square of linear momentum

$$\mathrm{kg}\,\frac{\mathrm{m}^2}{\mathrm{s}^2}$$

equilateral Planck energy momenta

$$m\Omega v^2$$

Energy-momentum relationship

The total number of equilateral Planck quanta [quantised mass-energy momenta] is directly related to the square of its linear momentum [mass-velocity]

Quantum Mechanics

$$h = \frac{E}{v^2}$$

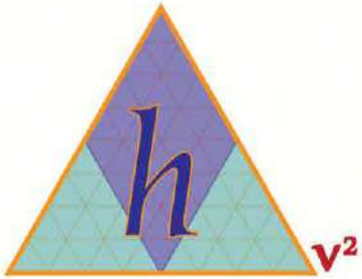

Quantised energy equilateral momenta

$$hv^2$$

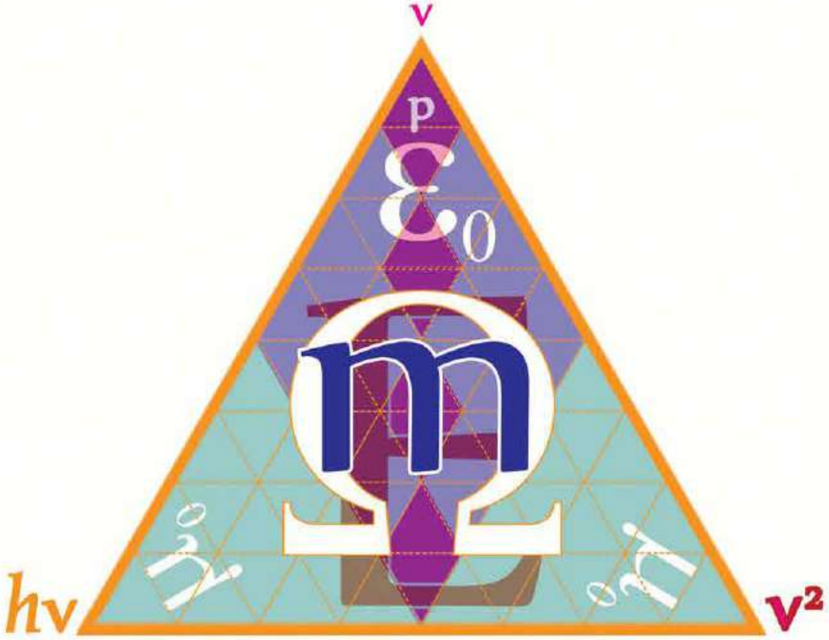

Quantised Energy momenta is related to Scalar mass energy momenta through the equilateral geometry of Planck's constant

Newtonian physics

$$m = \frac{E}{v^2}$$

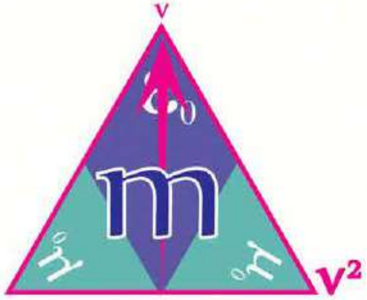

Scalar energy linear momentum

$$pv$$

$$m\Omega v^2 = E = mv^2$$

Copyright ABRAHAM [2008] - All rights reserved

Inertial resistance to Force

Inertia is the resistance of any physical object
to a change in its state of motion.

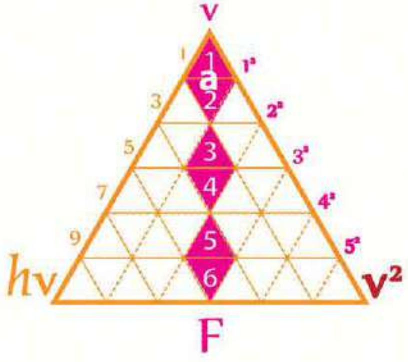

Changes to mass-velocity require a corresponding change to boson quanta

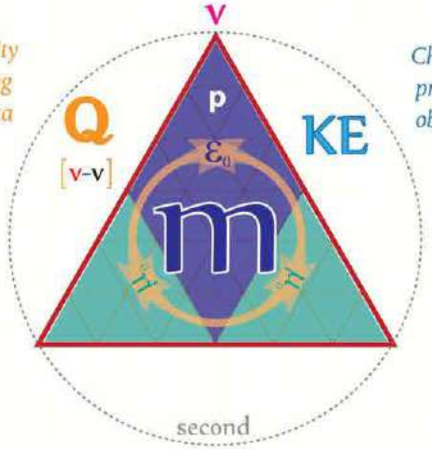

Changes to mass-velocity produces a change in an object's Kinetic Energies

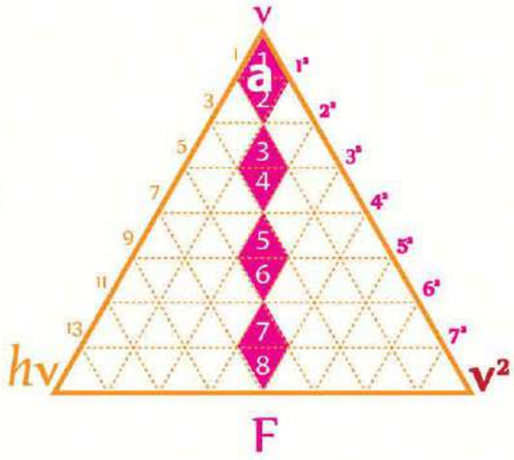

$$E = mc^2$$

The total intrinsic momenta of all energy waveforms is the sum of their constituent Quantised Angular momenta (mass-energy momenta)

$$M_0 = m/c^2$$

Any change in motion results in changes to the Charge geometries creating in turn proportional changes to KEM mass and momenta components

$$F = ma$$

The 'inductive resistance' of Charge quanta fields to changes in their mass-energy momenta content is what we term Inertia

$$KEM = Mv^2$$

Matter in motion has Kinetic Energies in addition to invariant rest mass-Energy

KE = RE - rest Matter

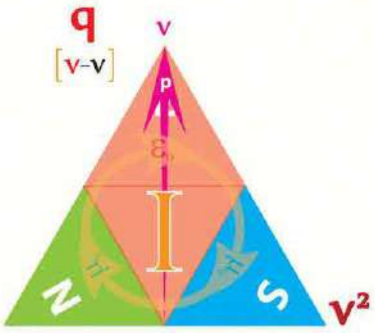

quantised moment of inertia

Any change to an object's velocity results in a corresponding change to its mass-Energy momenta which is reflected by its inertia

$$p^2 = E = mv^2$$

Inertial mass can be related to Charge through inductive Planck quanta

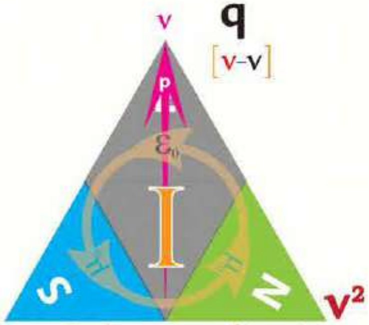

quantised moment of inertia

Quantised Angular momentum

As it is a physical [equilateral] geometry QAM is conservative in any system where there are no external Forces and serves as the foundational geometric source for all the conservation laws of physics

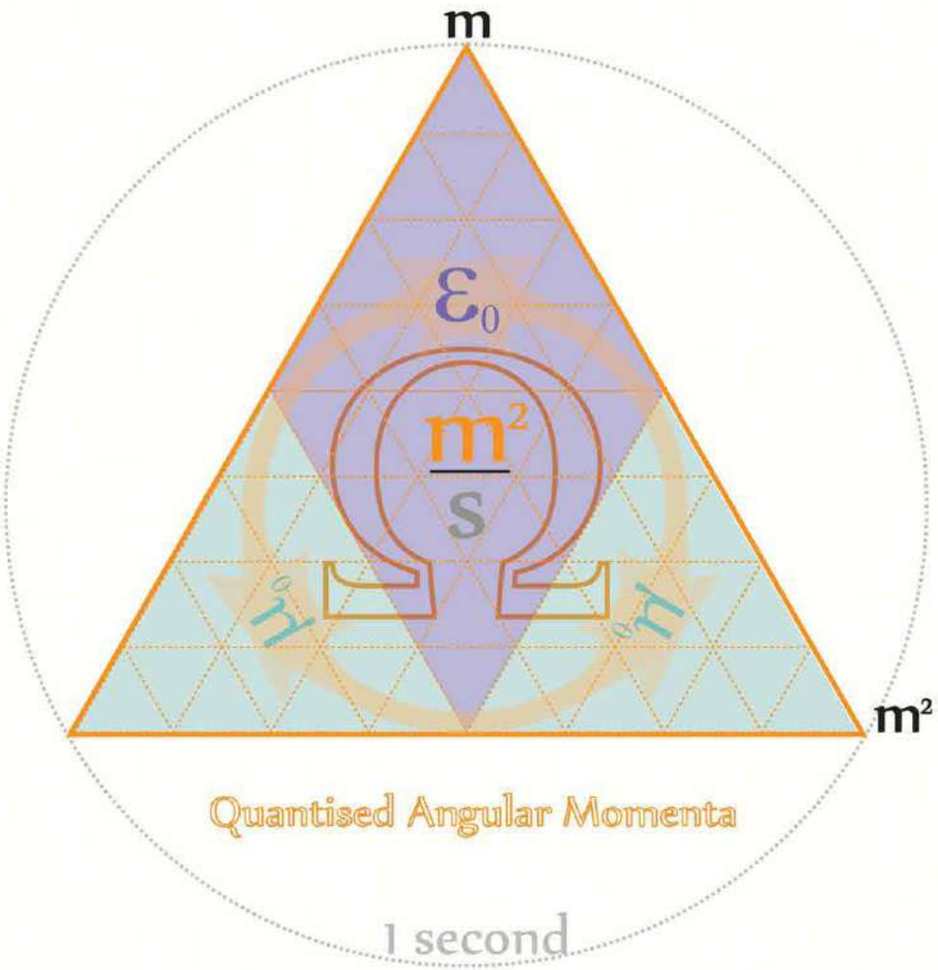
m
ε0
m²/s
Ω
μ0
μ0
m²
Quantised Angular Momenta
1 second

Angular momentum is sometimes described as the rotational analog of linear momentum, in Tetryonics it is revealed to be the equilateral geometry of quantised mass-energy momenta within any defined space-time co-ordinate system

A major re-definition of quantised angular momentum in physics is revealed

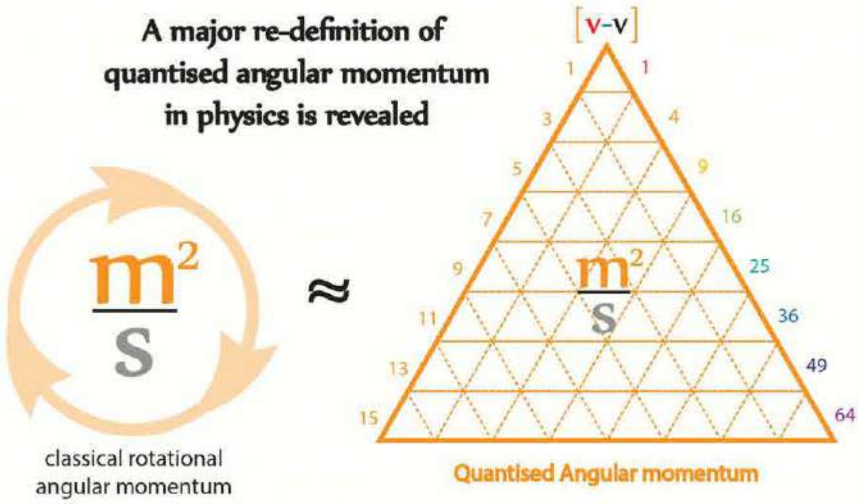
m²/s
≈
[v-v]
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
m²/s
classical rotational angular momentum
Quantised Angular momentum

In quantum mechanics, angular momentum is quantised – that is, it cannot vary continuously, but only in ODD number "quantum steps" between the allowed SQUARE nuclear Energy levels

In physics, angular momentum, moment of momentum, or rotational momentum is a conserved vector quantity that can be used to describe the overall state of a physical system.

When applied to specific mass-Energy-Matter systems QAM reveals the true quantum geometry and nature of Energy in our universe

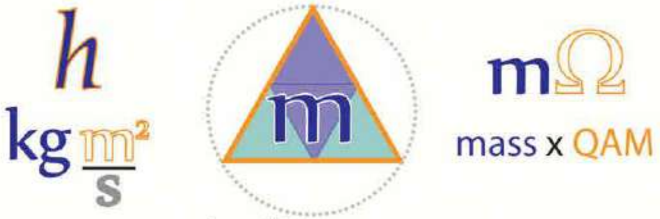
h
kg m²/s
m
mΩ
mass x QAM

Planck's Constant

Normally viewed as an expression of rotational momentum Quantised Angular Momentum [QAM] is in fact a result of the equilateral geometric quantization of mass-energy

Copyright ABRAHAM [2008] - All rights reserved

Charged geometries

All charge geometries are nett divergent

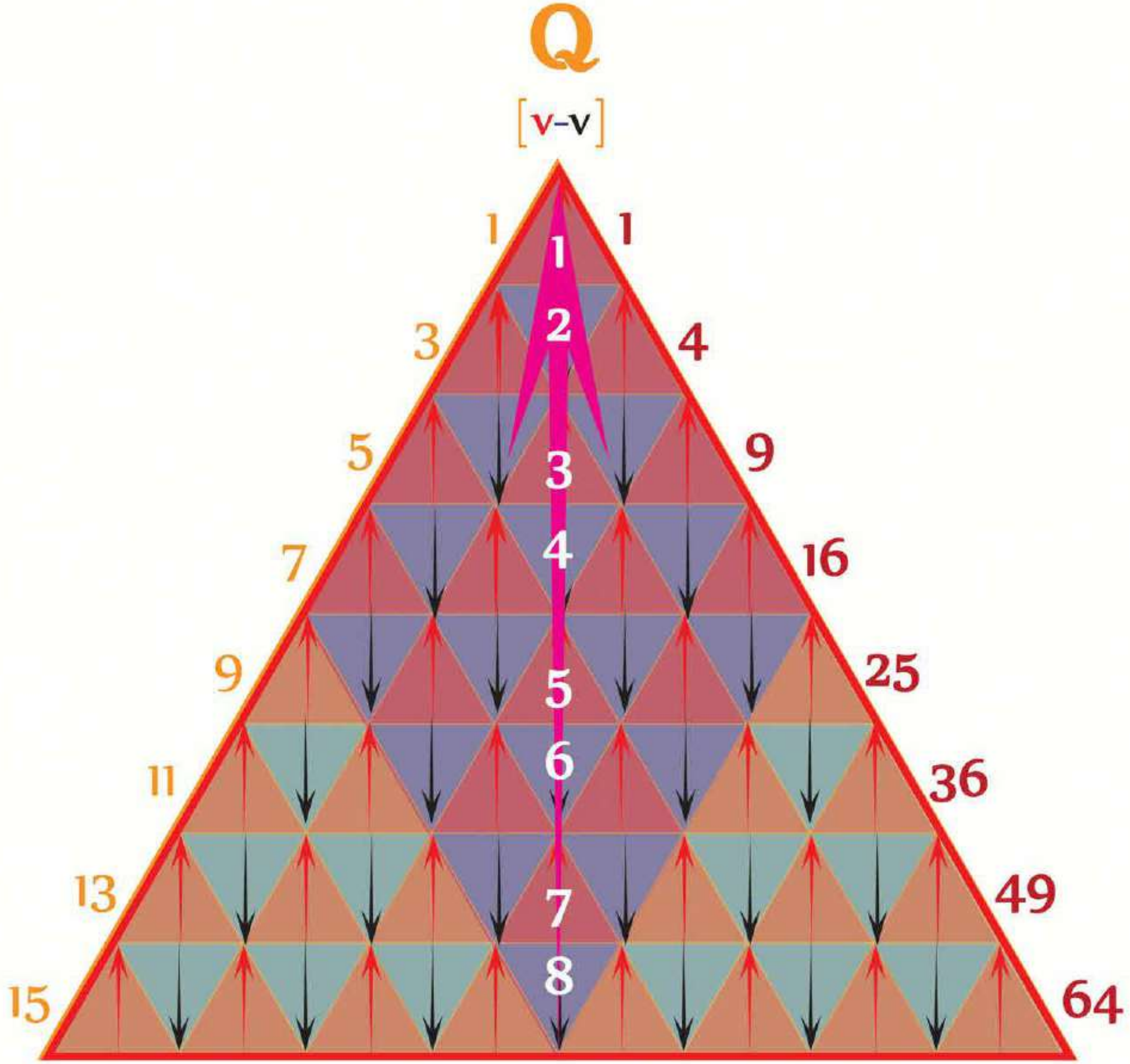
Q
[v-v]
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
1
2
3
4
5
6
7
8

Divergent energy momenta	Convergent energy momenta
1	0
3	1
6	3
10	6
15	10
21	15
28	21
36	28

All charge geometries are comprised of finite equilateral energy momenta quanta

Renormalisation

Renormalization was first developed in quantum electrodynamics (QED) to make sense of infinite integrals in perturbation theory.
The problem of infinities first arose in the classical electrodynamics of point particles in the19th and
later in the calculation of Gravitational fields in General Relativity in the early 20th century.

In QED Infinities must be cancelled

In Tetryonics Infinities do not exist

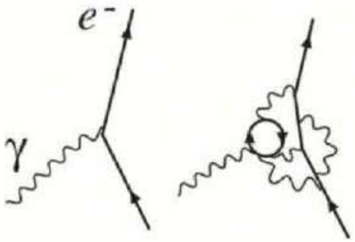

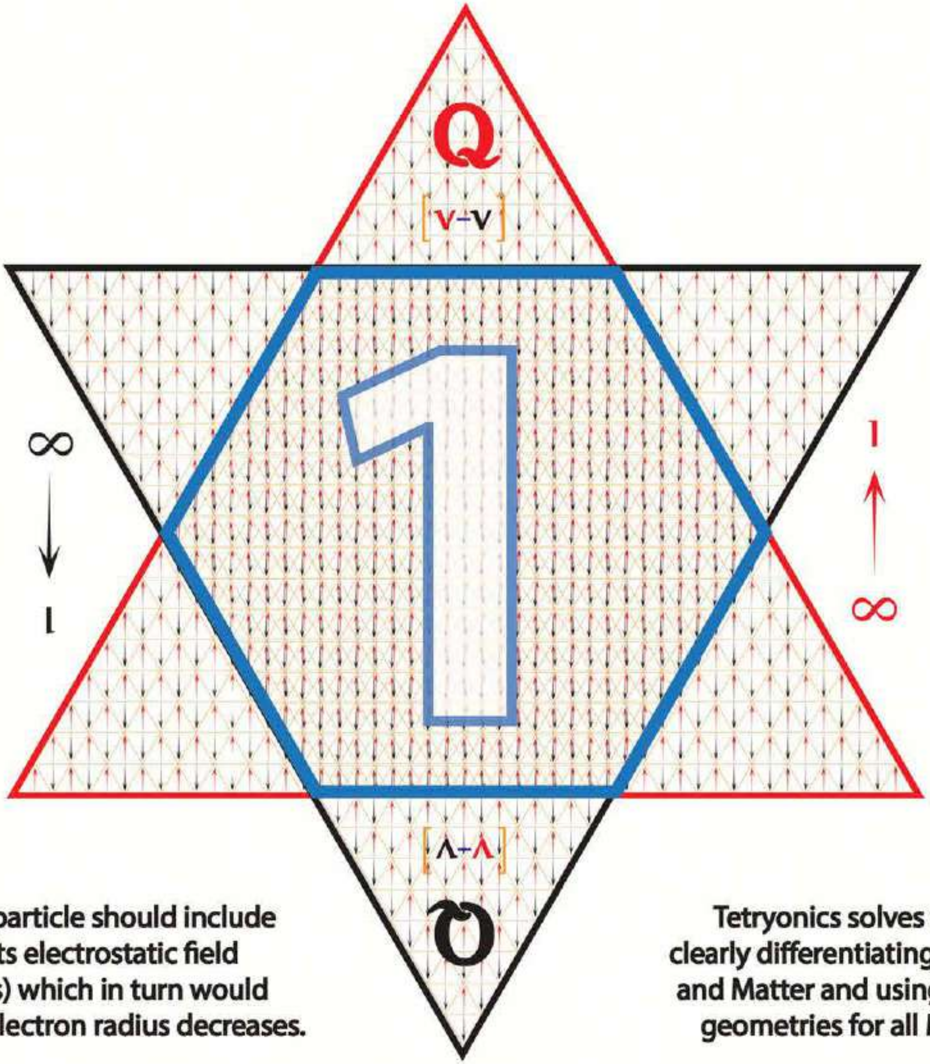

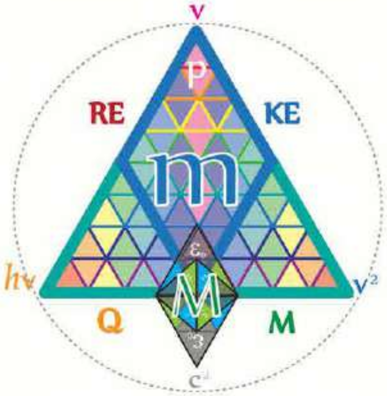

The mass of a charged particle should include the mass-energy in its electrostatic field (Electromagnetic mass) which in turn would approach infinity as the electron radius decreases.

Tetryonics solves the problem by clearly differentiating between EM mass and Matter and using finite equilateral geometries for all Matter in motion

Initially viewed as a suspicious provisional procedure by some of its originators, renormalization was eventually embraced as an important and self-consistent tool in several fields of physics and mathematics.

2D space $[c^2]$

The adjective Cartesian refers to the French mathematician and philosopher René Descartes who developed the coordinate system in 1637

Since then many other coordinate systems have been developed such as the polar coordinates for the plane, and the spherical and cylindrical coordinates for three-dimensional space.

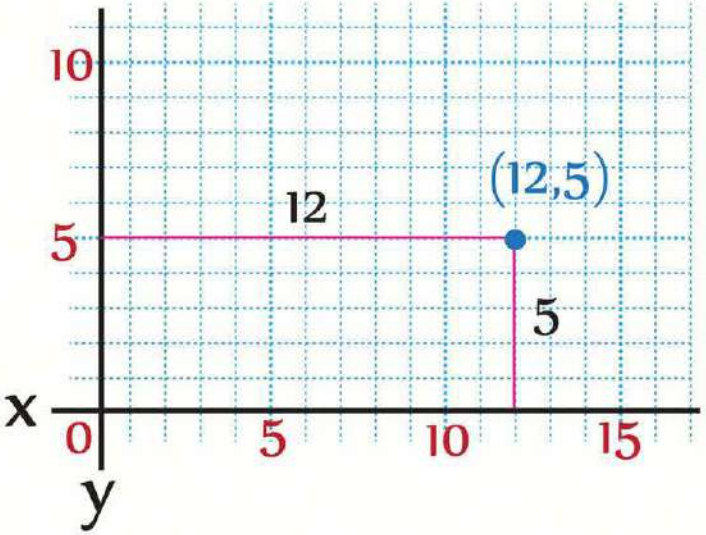

Cartesian coordinates are the foundation of analytic geometry, and provide enlightening geometric interpretations for many other branches of mathematics, such as linear algebra, complex analysis, differential geometry, multivariate calculus, group theory, and more

Mapping 3D spaces using Recti-linear co-ordinates

Cartesian coordinates can be defined as the positions of the perpendicular projections of a point onto the two or more axes, expressed as signed distances from the origin.

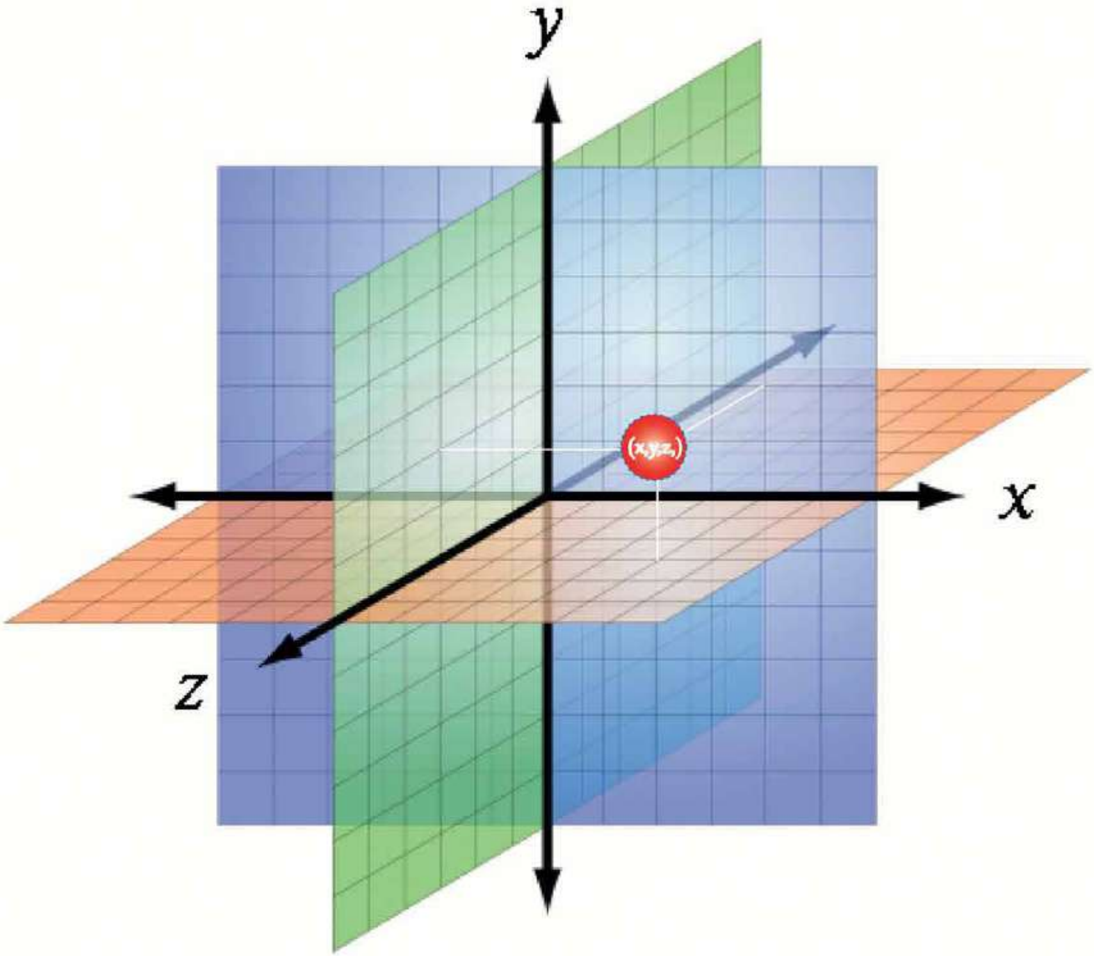

3D Cartesian co-ordinate $[c^3]$ systems are distinct from spherical co-ordinate $[c^4]$ systems

Polar co-ordinates

In mathematics, the polar coordinate system is a two-dimensional co-ordinate system in which each point on a plane is determined by a distance from a fixed point and an angle from a fixed direction.

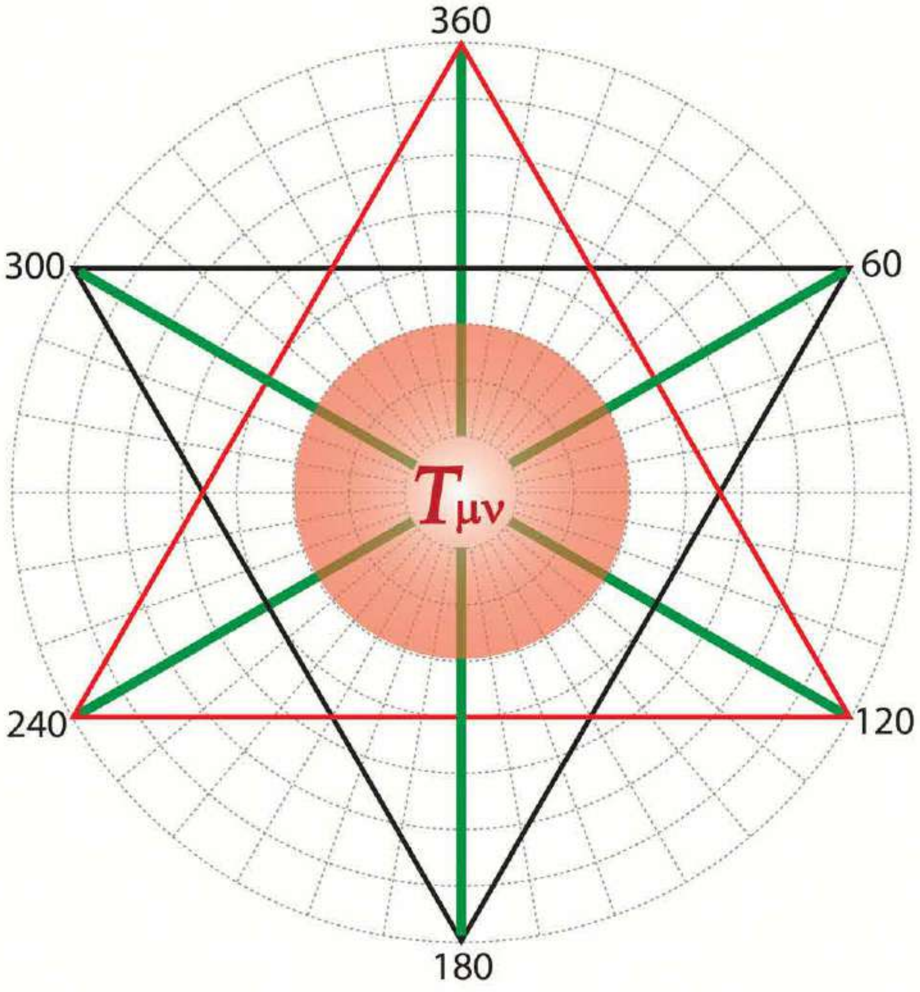

In geometry, curvilinear coordinates are a coordinate system for Euclidean space in which the co-ordinate lines may be curved.

Action Dynamics

Curvilinear co-ordinates may be derived from a set of rectilinear Cartesian coordinates by using a locally invertible transformation that maps one point to another in both systems

Metric Tensors

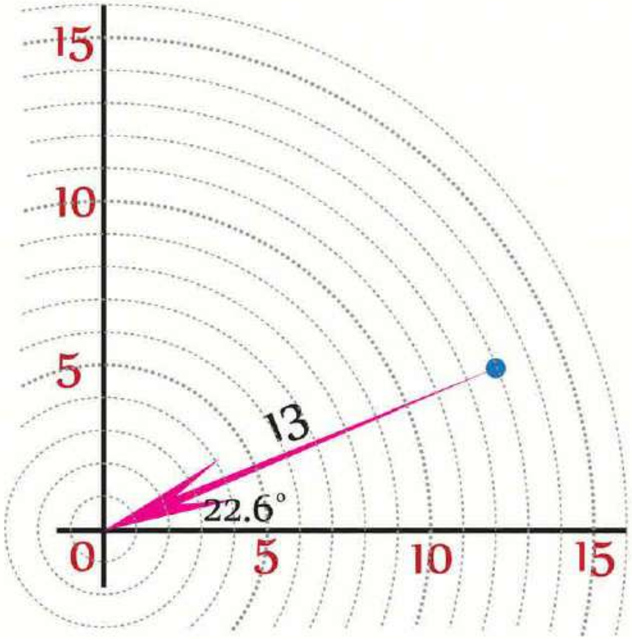

Gravitational acceleration

Polar or curvilinear co-ordinate systems are used extensively by Einstein in his theory of General Relavtivity

Reimannian curved space-time

Copyright ABRAHAM [2008] - All rights reserved

Co-ordinate transformations

There are many different possible coordinate systems for describing geometrical figures and they can all be related to one another.

Such relations are described by coordinate transformations which give formulas for the coordinates in one system in terms of the coordinates in another system

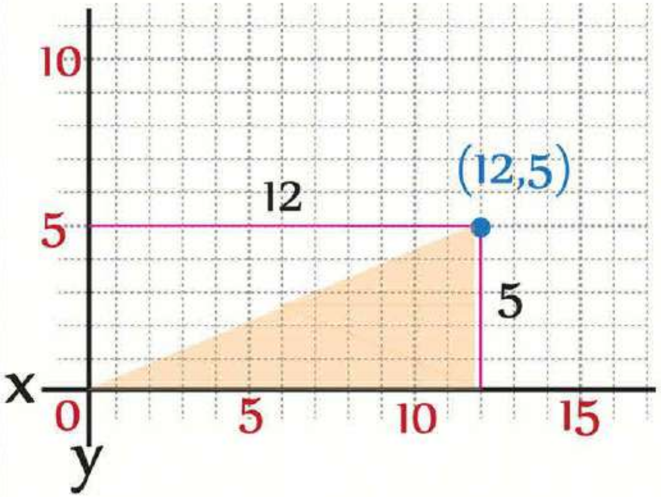

Converting between Polar and Cartesian coordinates

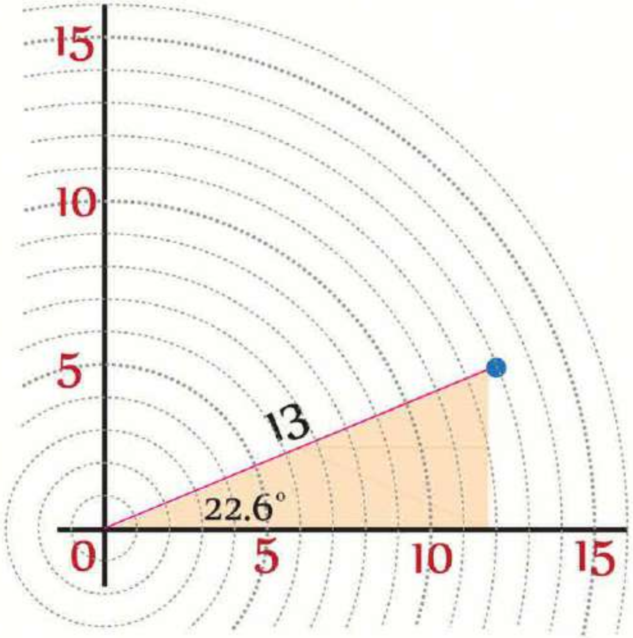

$$\tan\theta = \frac{\text{opposite}}{\text{adjacent}}$$

$$= \tan^{-1}(5/12)$$

$$= 22.619$$

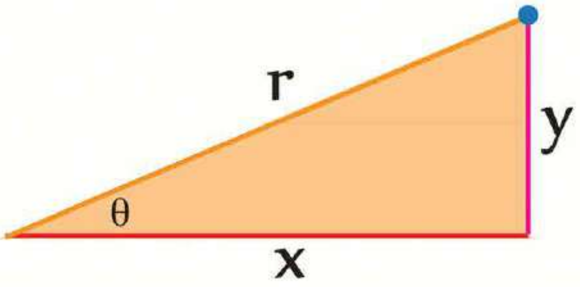

$$r^2 = 12^2 + 5^2$$

$$= \sqrt{169}$$

$$= 13$$

solving for the triangle using trigometric functions

Spatial co-ordinate systems

Spacetime is any mathematical co-ordinate system or model that combines space and time into a single continuum.

Spacetime is usually interpreted with space as being three-dimensional with time playing the role of a fourth dimension that is different from the spatial dimensions.

From a Euclidean space perspective, the universe has three spatial dimensions and one dimension of time [reflected by quantised angular momentum].

Cubic

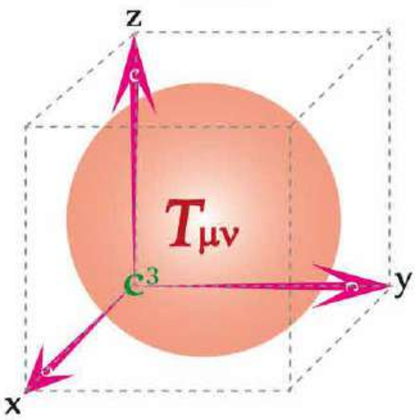

Spherical

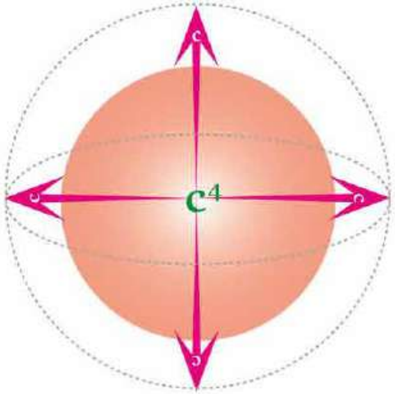

seconds²

Tetryonics maps spatial co-ordinates through the momenta vectors of equilateral Energy

Cartesian Space-Time

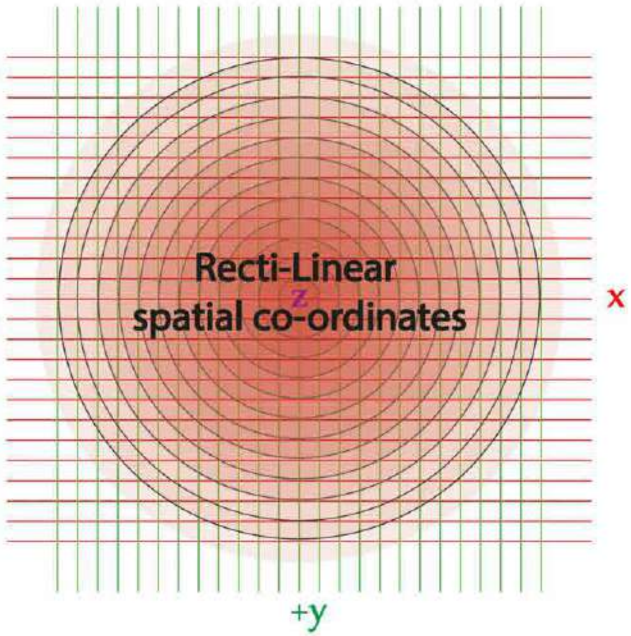

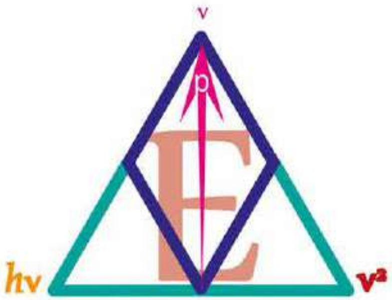

EM mass-ENERGY momenta are equilateral geometries

Mapping equilateral Energy geometries onto recti or curvi-linear spatial co-ordinate systems introduces mathematical complexity to a otherwise simplistic geometry for all EM mass-Energy-Matter interactions

Tetryonic Space-Time

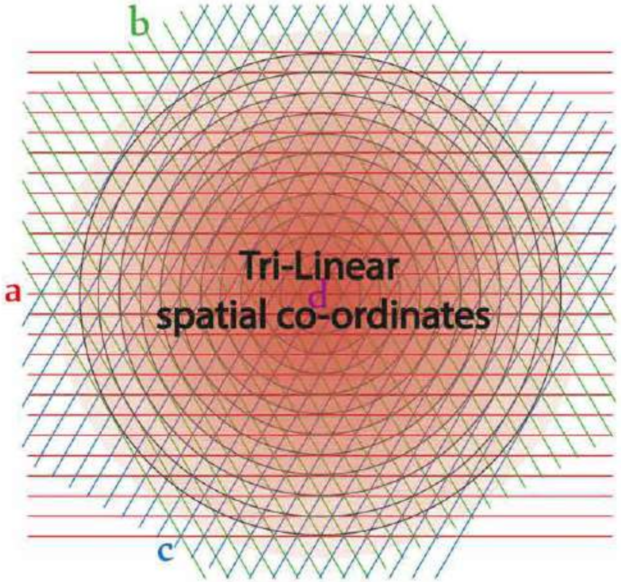

In physics spatial co-ordinates to date have been based on Cartesian co-ordinates when in fact Energy momenta follow a Tri-Linear co-ordinate geometry

Copyright ABRAHAM [2008] - All rights reserved

vectors

1 Dimensional velocity

Euclidean

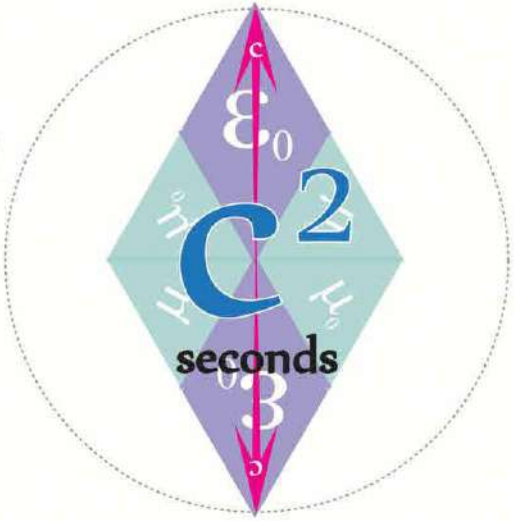

2 Dimensions velocity squared

$$m = \frac{E}{c^2}$$

mass-energy equivalence

mass-Matter

$$\frac{m}{c^2} = M$$

equivalence

$T_{\mu\nu}$

energy-Matter equivalence

$$M = \frac{E}{c^4}$$

GR fails differentiate betweem EM mass-ENERGY-Matter unlike Tetryonic theory

Cartesian co-ordinates

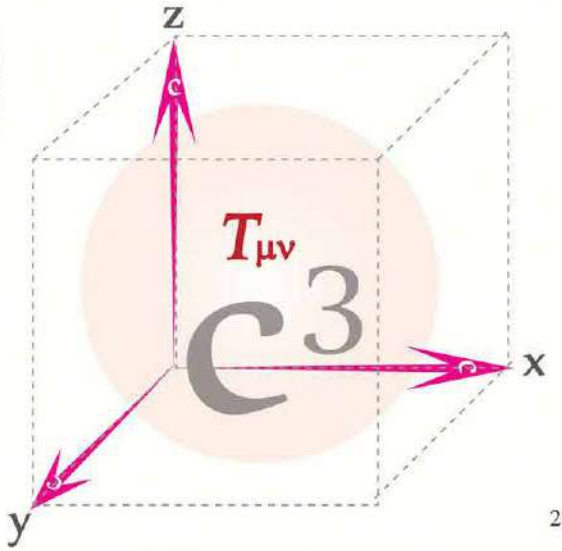

3 Dimensions velocity cubed

Spherical

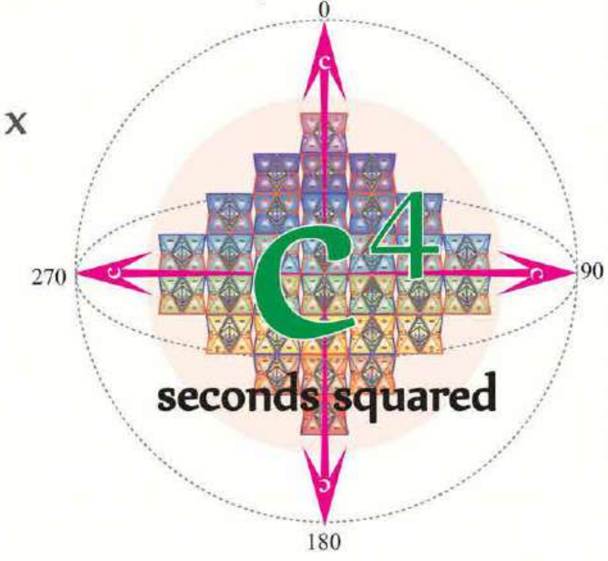

3 Dimensions quadrature velocity

Space-time co-ordinates

The propagation of Energy momenta forms distinct spatial co-ordinate systems

Where angles are typically measured in degrees (°) or radians (rad), where 360° = 2π rad = τ radians.

Tetryonic co-ordinate systems

Energy per second

m

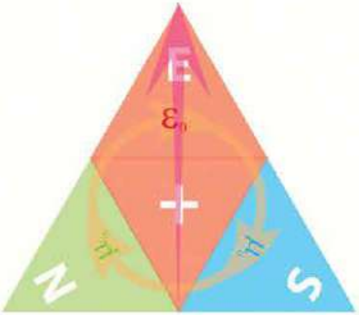
E
ε_0
+
N
S
μ_0
μ_0

2D mass-energy geometries

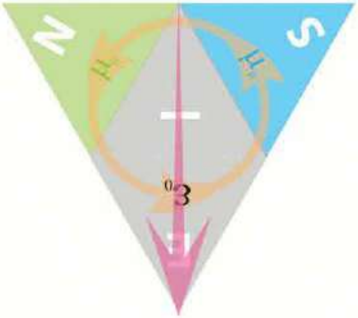
N
S
μ_0
μ_0
-
ε_0
E

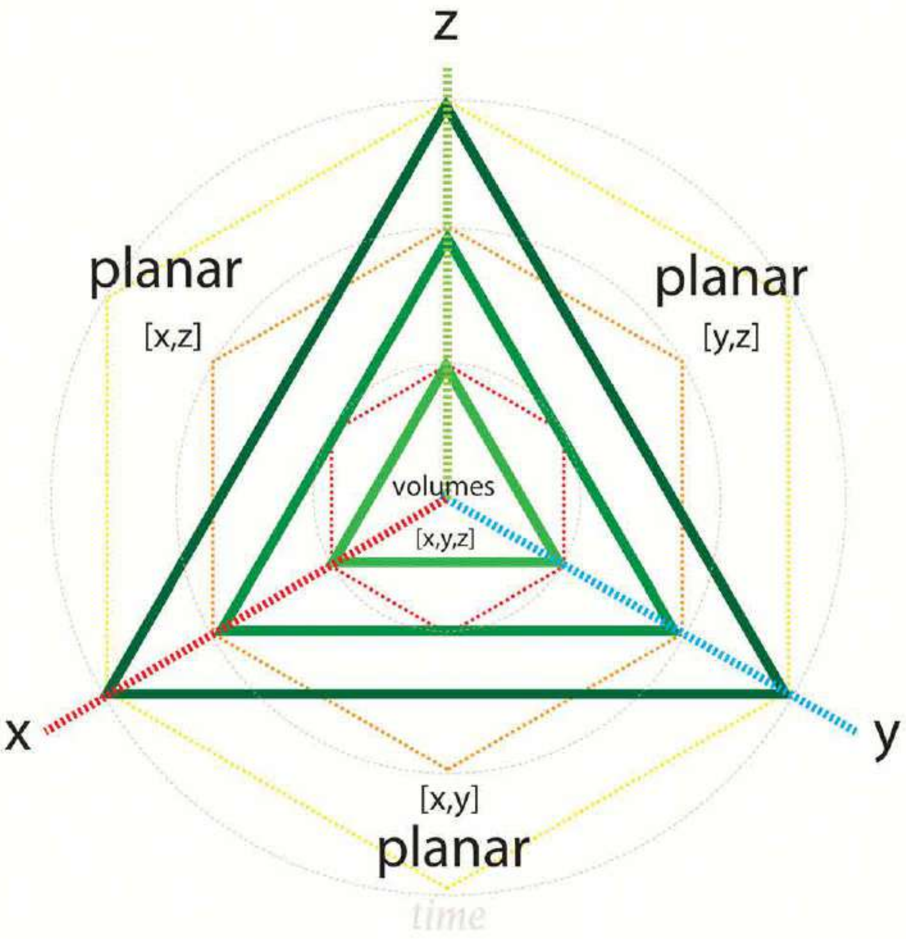
z
planar
[x,z]
planar
[y,z]
volumes
[x,y,z]
x
y
[x,y]
planar
time

Planar mass-energy geometries have no z-components

Differentiation between 2D mass-energy & 3D Matter is key to extending our understanding of physics

Matter topologies have z-components

Energy per second squared

M

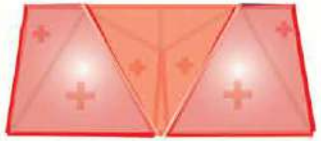

Proton

3D Matter topologies

Neutron

Copyright ABRAHAM [2008] - All rights reserved

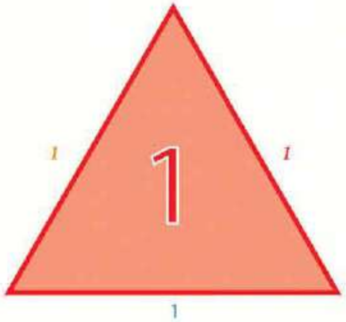

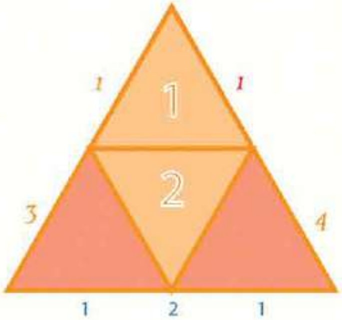

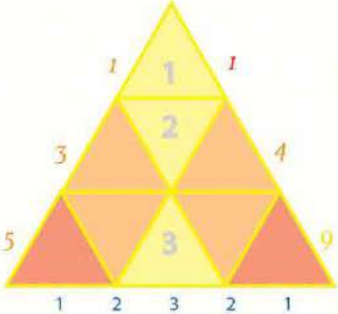
1
1
1
2
3
4
3
5
9
1 2 3 2 1

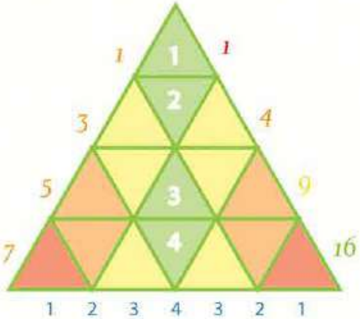
1
1
1
2
3
4
3
4
5
9
7
16
1 2 3 4 3 2 1

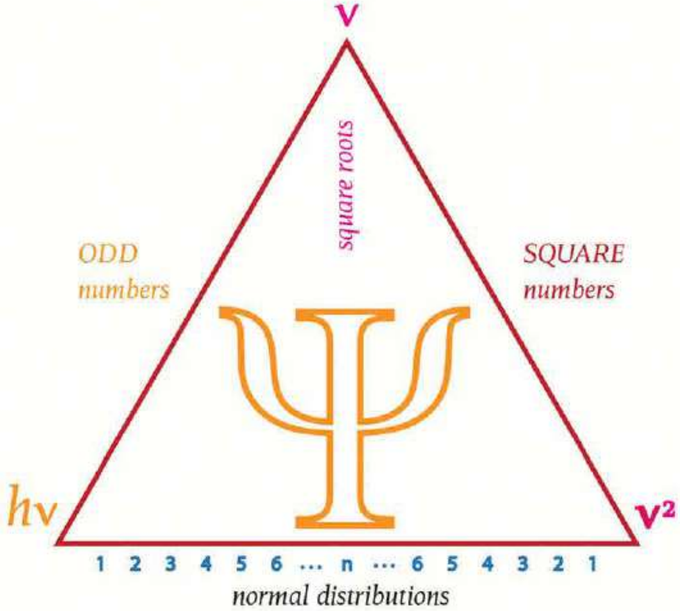
v
square roots
ODD numbers
SQUARE numbers
Ψ
hv
v²
1 2 3 4 5 6 ... n ... 6 5 4 3 2 1
normal distributions

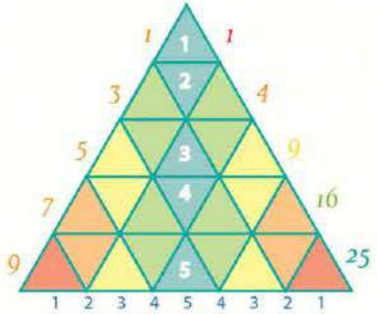
1
1
1
2
3
4
3
5
9
4
7
16
5
9
25
1 2 3 4 5 4 3 2 1

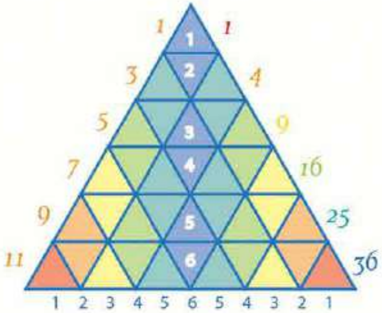
1
1
1
2
3
4
3
5
9
4
7
16
5
9
25
6
11
36
1 2 3 4 5 6 5 4 3 2 1

Quantum probability distributions

The equilateral geometry and distribution of quantised Energy momenta provides the basis for all statistical probabilities in Quantum mechanics, thermodynamic & information entropy - including a solution to Heisenberg's Uncertainty Principle thus paving the way forward for a new understanding, and manipulation of physical phenomena at the quantum level

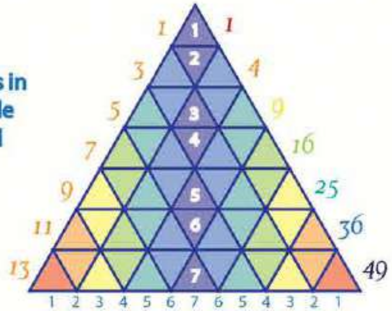
1
1
1
2
3
4
3
5
9
4
7
16
5
9
25
6
11
36
7
13
49
1 2 3 4 5 6 7 6 5 4 3 2 1

1

1 2 1

1 2 3 2 1

1 2 3 4 3 2 1

1 2 3 4 5 4 3 2 1

1 2 3 4 5 6 5 4 3 2 1

1 2 3 4 5 6 7 6 5 4 3 2 1

1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Normal distributions are extremely important in statistics, and are often used in the natural and social sciences for real-valued random variables whose distributions are not known

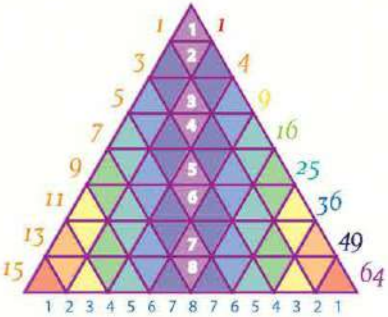
1
1
1
2
3
4
3
5
9
4
7
16
5
9
25
6
11
36
7
13
49
8
15
64
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Copyright ABRAHAM [2008] - All rights reserved

Quantum Probability Distributions

The normal distribution is a probability distribution.
It is also called Gaussian distribution because it was discovered by Carl Friedrich Gauss.

x/n^2

Probabilities are
the square of the
Amplitude

n

$\Sigma\, x/n^2$

All probabilities
are re-normalisable
and sum to Unity

1

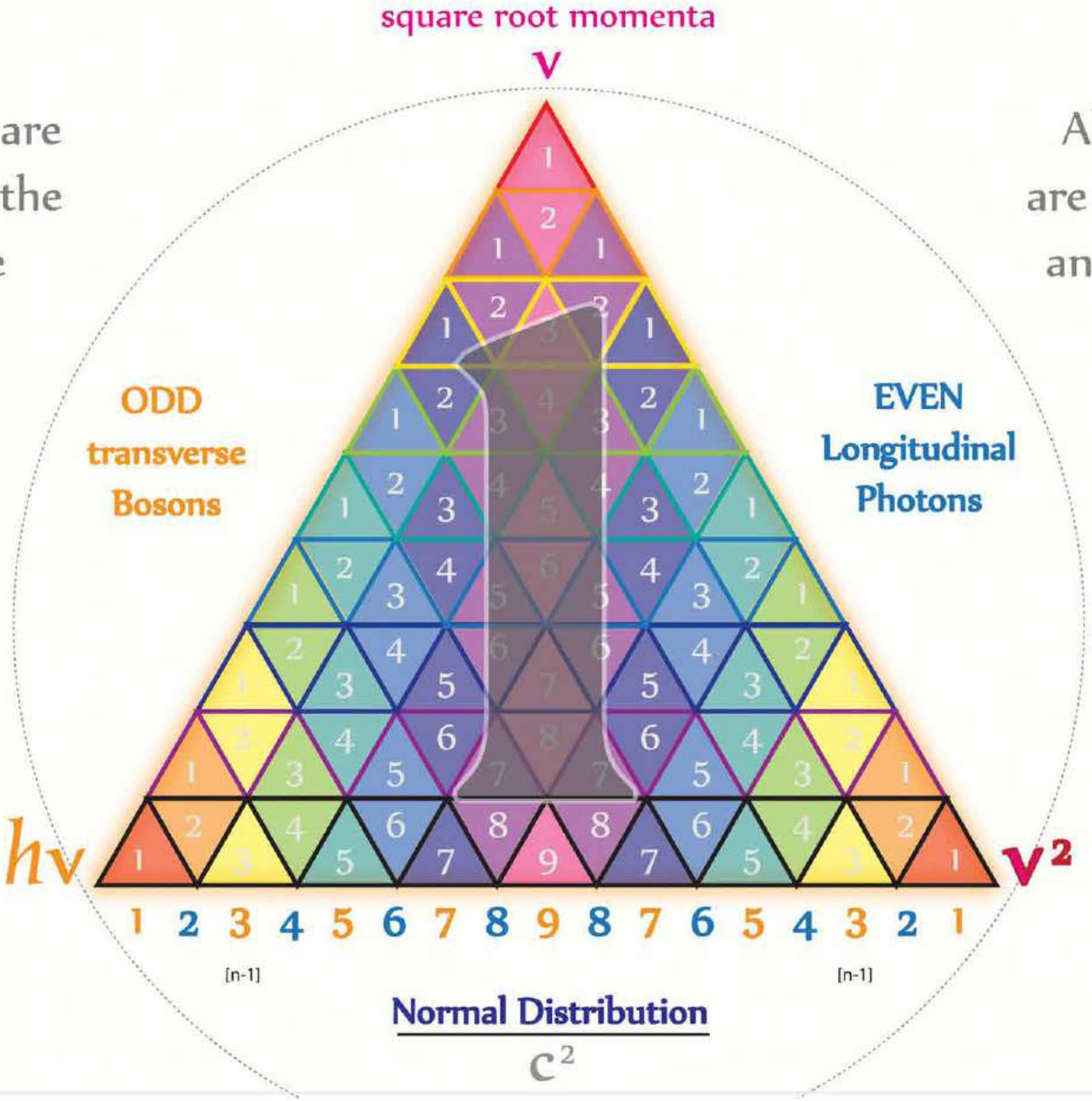
square root momenta
v
ODD
transverse
Bosons
EVEN
Longitudinal
Photons
Planck
quanta
hv
v^2
Square
energies
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1
[n-1]
[n-1]
Normal Distribution
c^2

Quanta Distributions

Quantum Levels

transverse bosons

$$2h\nu = E = hf$$

Quantum Probabilities

longitudinal photons

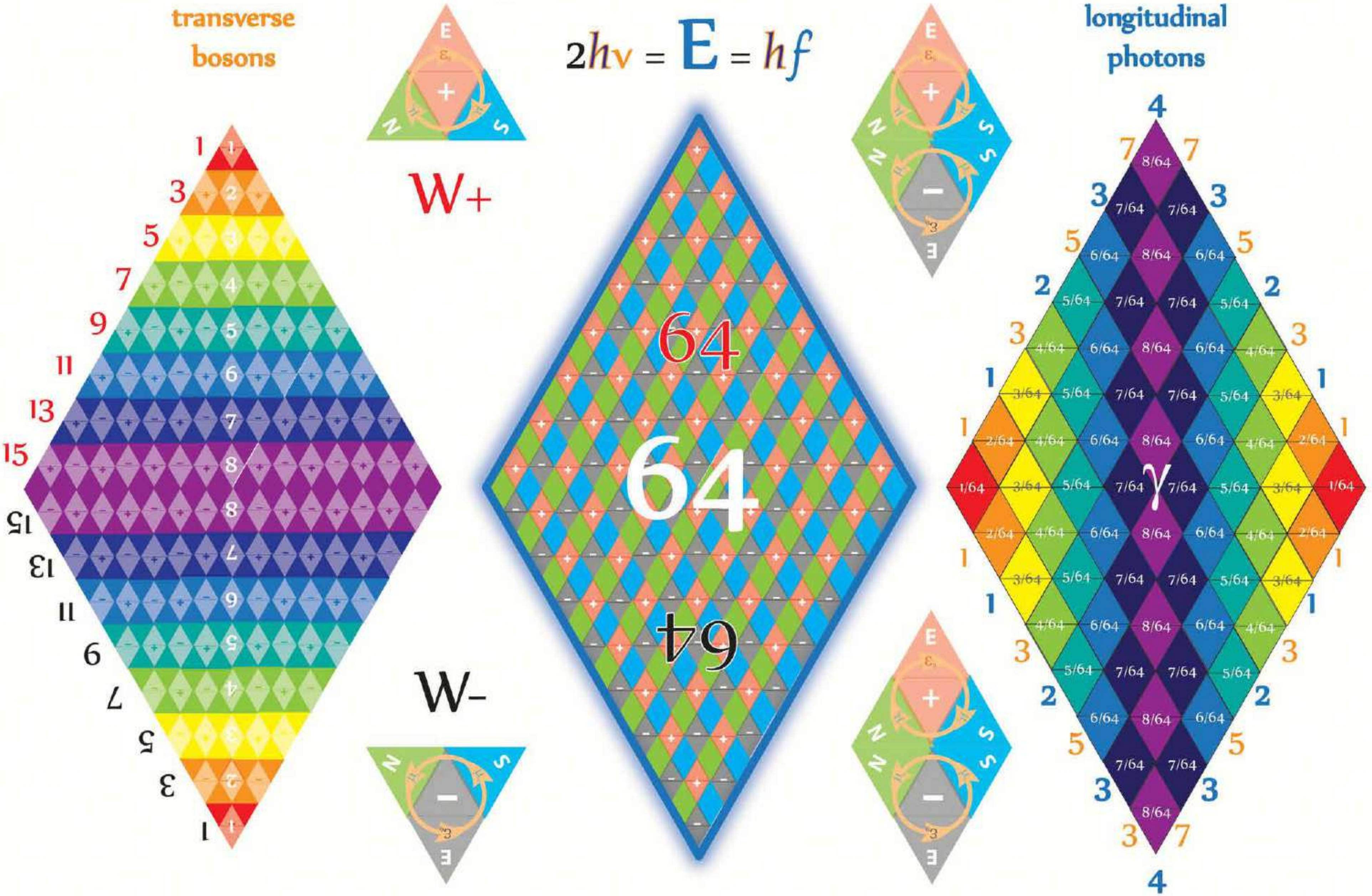
1
3
5
7
9
11
13
15
15
13
11
9
7
5
3
1
W+
W-
64
64
64
E
ε₀
+
−
N
S
4
7
3
5
2
1
γ
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64

Wavefunctions

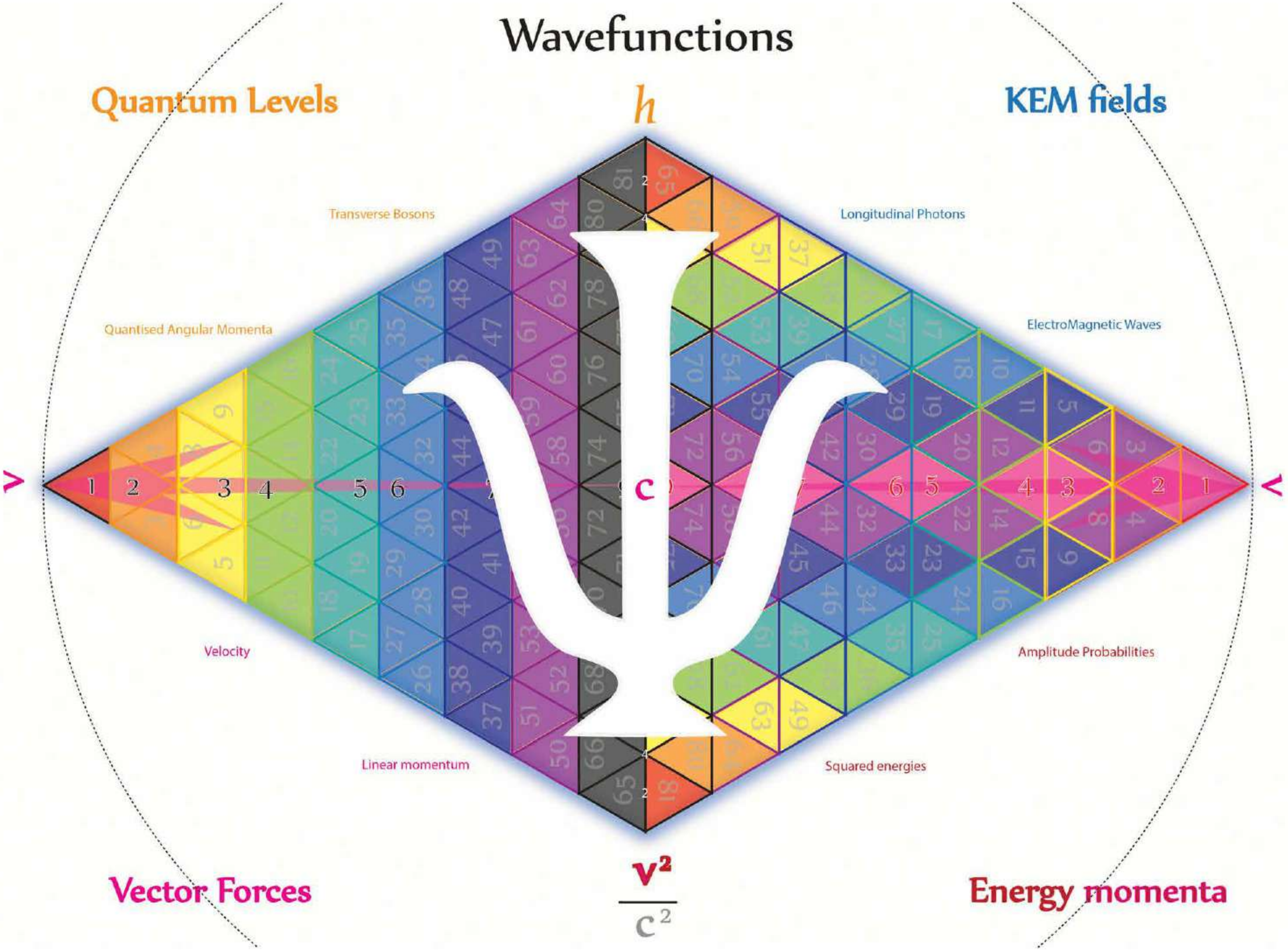
Quantum Levels
h
KEM fields
Transverse Bosons
Longitudinal Photons
Quantised Angular Momenta
ElectroMagnetic Waves
v
1 2 3 4 5 6
c
6 5 4 3 2 1
v
Velocity
Amplitude Probabilities
Linear momentum
Squared energies
Vector Forces
$\frac{v^2}{c^2}$
Energy momenta

EM waveforms

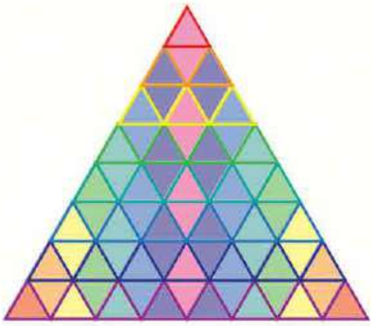

All EM waveforms can be measured by either their Transverse EM masses [Bosons] or their Longitudinal EM masses [Photons]

PHOTONS

EM waves Planck quanta

$$\frac{\text{EVEN}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

vs.

BOSONS

Bosons Planck quanta

$$\frac{\text{ODD}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

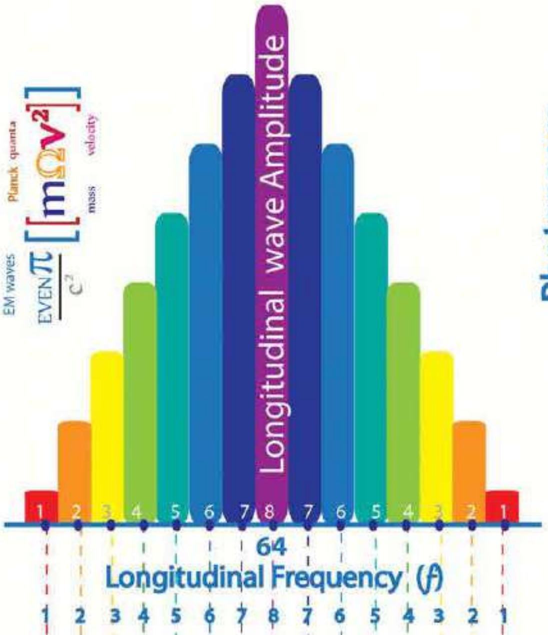
Longitudinal wave Amplitude
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
64
Longitudinal Frequency (f)
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Photons are longitudinal quanta

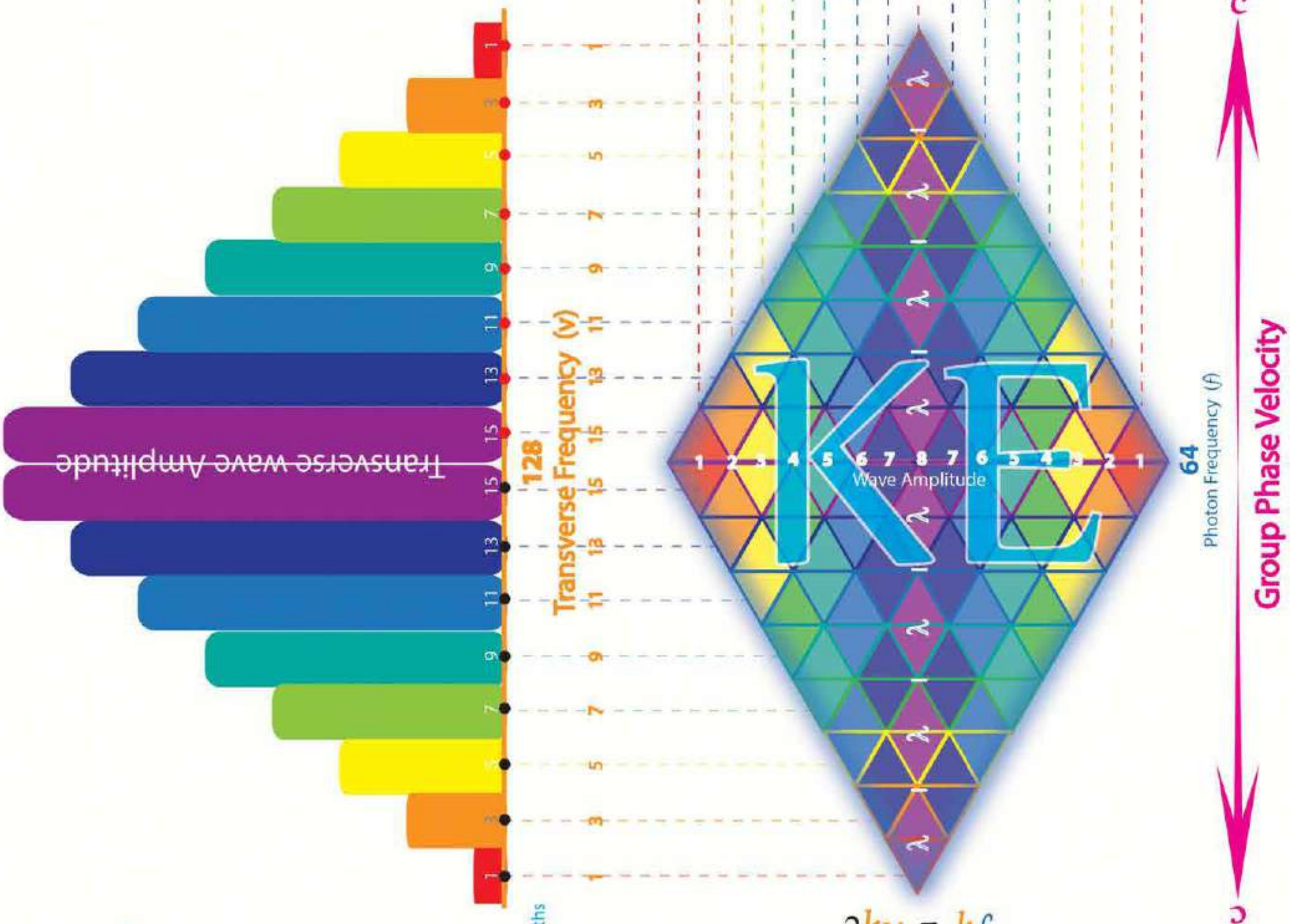
Transverse wave Amplitude
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
128
Transverse Frequency (v)
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
λ
KE
Wave Amplitude
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
64
Photon Frequency (f)
Group Phase Velocity
c
c
EM Wavelengths

$$2hv = hf$$

Bosons are transverse quanta

Copyright ABRAHAM [2008] - All rights reserved

Quantum Energy distributions

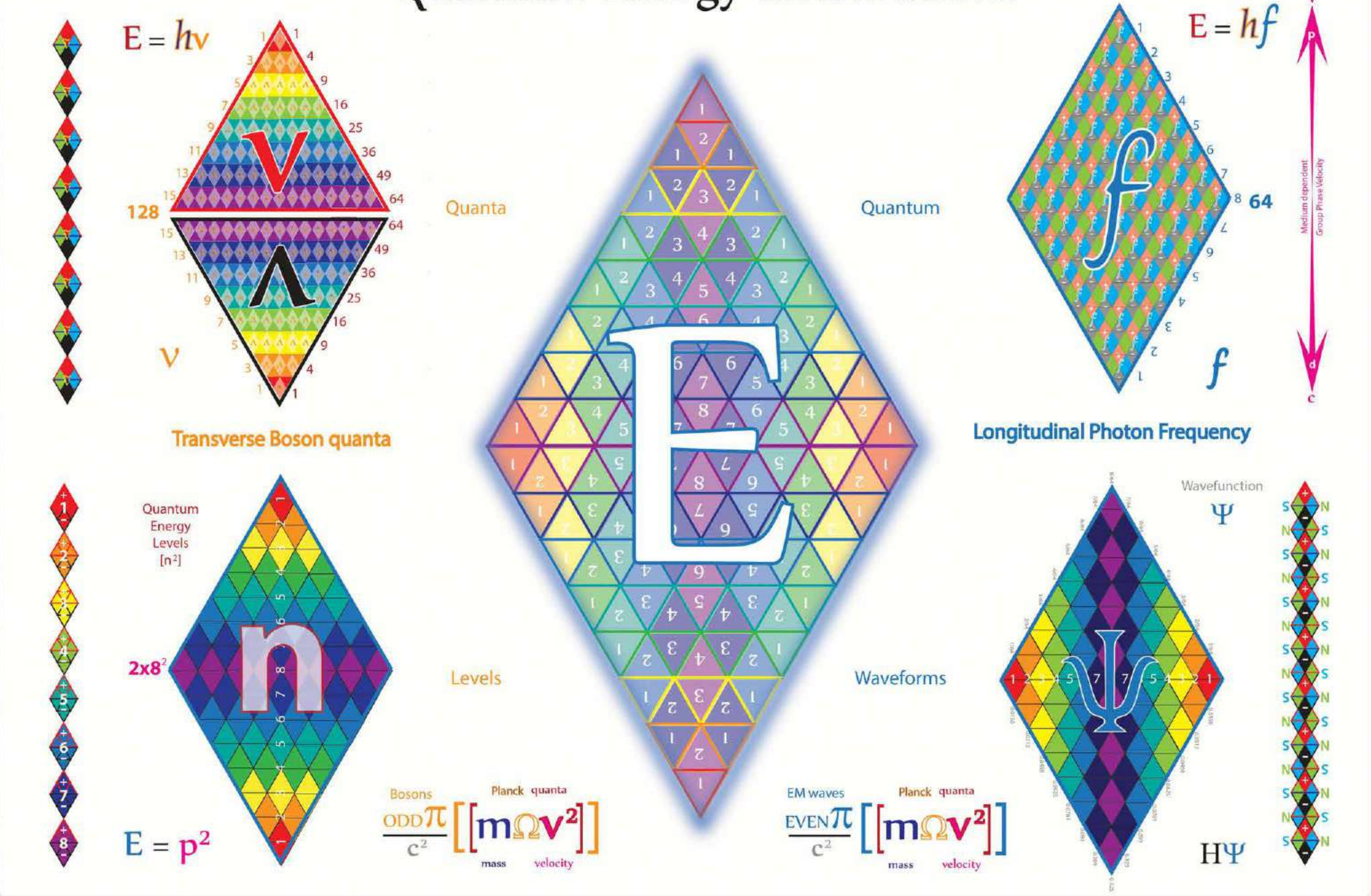
1
2
3
4
5
6
7
8
9
11
13
15
16
25
36
49
64
ν
Λ
E
f
n
Ψ
+
-
S
N
v
p
d
c
Medium dependent
Group Phase Velocity

$E = h\nu$

128

ν

Quanta

Transverse Boson quanta

$E = hf$

64

Quantum

Longitudinal Photon Frequency

f

Quantum Energy Levels $[n^2]$

$2x8^2$

Levels

$E = p^2$

Bosons

Planck quanta

$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$

mass velocity

Waveforms

EM waves

Planck quanta

$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$

mass velocity

Wavefunction

Ψ

$H\Psi$

Normal Distributions

The Gaussian distribution sometimes informally called the bell curve.

Pierre de Fermat

Pierre de Fermat is given credit
for early developments that led to
infinitesimal calculus.

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

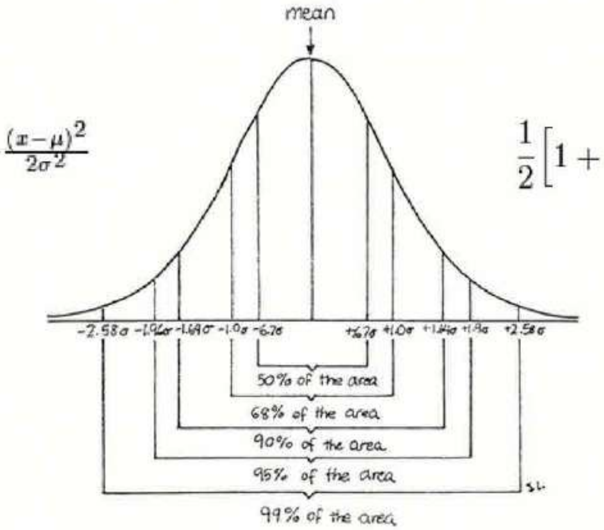

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$

Leonhard Euler

Leonhard Euler developed a formula
which links complex exponentiation
with trigonometric functions

**A bell shaped curve defines the standard normal distribution,
in which the probability of observing a point is greatest near the average,
and declines rapidly as one moves away from the mean.**

all ODD numbers are

$$a^2 - b^2 = [a-b].[a+b]$$

the difference of two squares

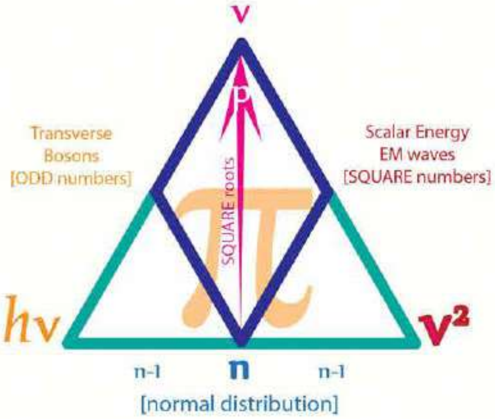
v
Transverse
Bosons
[ODD numbers]
Scalar Energy
EM waves
[SQUARE numbers]
SQUARE roots
hv
v²
n-1
n
n-1
[normal distribution]

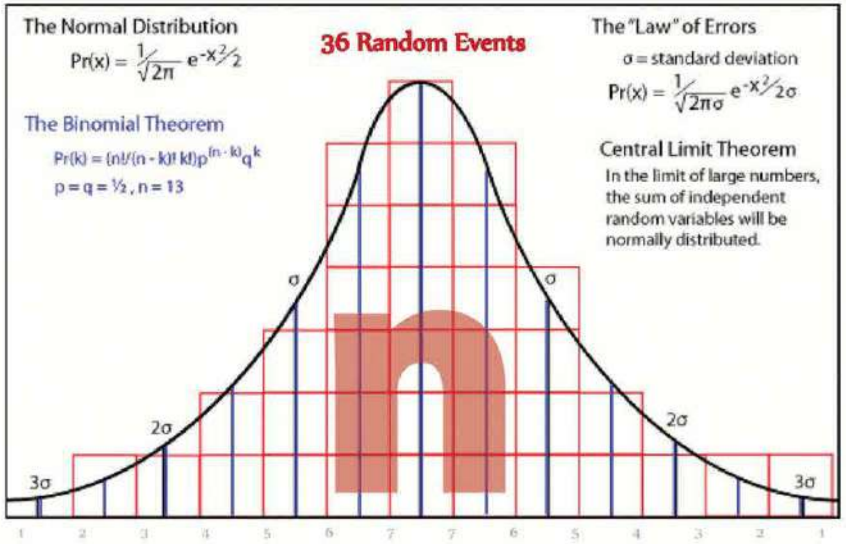

The Normal Distribution

$$\Pr(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

The Binomial Theorem

$$\Pr(k) = (n!/(n-k)!\,k!)\,p^{(n-k)}q^{k}$$

$$p = q = ½,\ n = 13$$

The "Law" of Errors

σ = standard deviation

$$\Pr(x) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{-x^2/2\sigma}$$

Central Limit Theorem

In the limit of large numbers, the sum of independent random variables will be normally distributed.

$$e^{i\pi} = -1$$

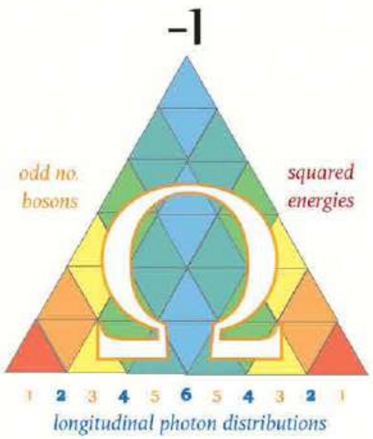
-1
odd no.
bosons
squared
energies
1
2
3
4
5
6
5
4
3
2
1
longitudinal photon distributions

In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function

Copyright ABRAHAM [2008] - All rights reserved

Fundamental theorem of Energy momenta

A nth level scalar energy momenta waveform has exactly n linear momentum in unit circle co-ordinate systems
(with Longitudinal and Transverse equilateral Planck waveforms being orthogonal to each other)

scalar energies

E

mv^2

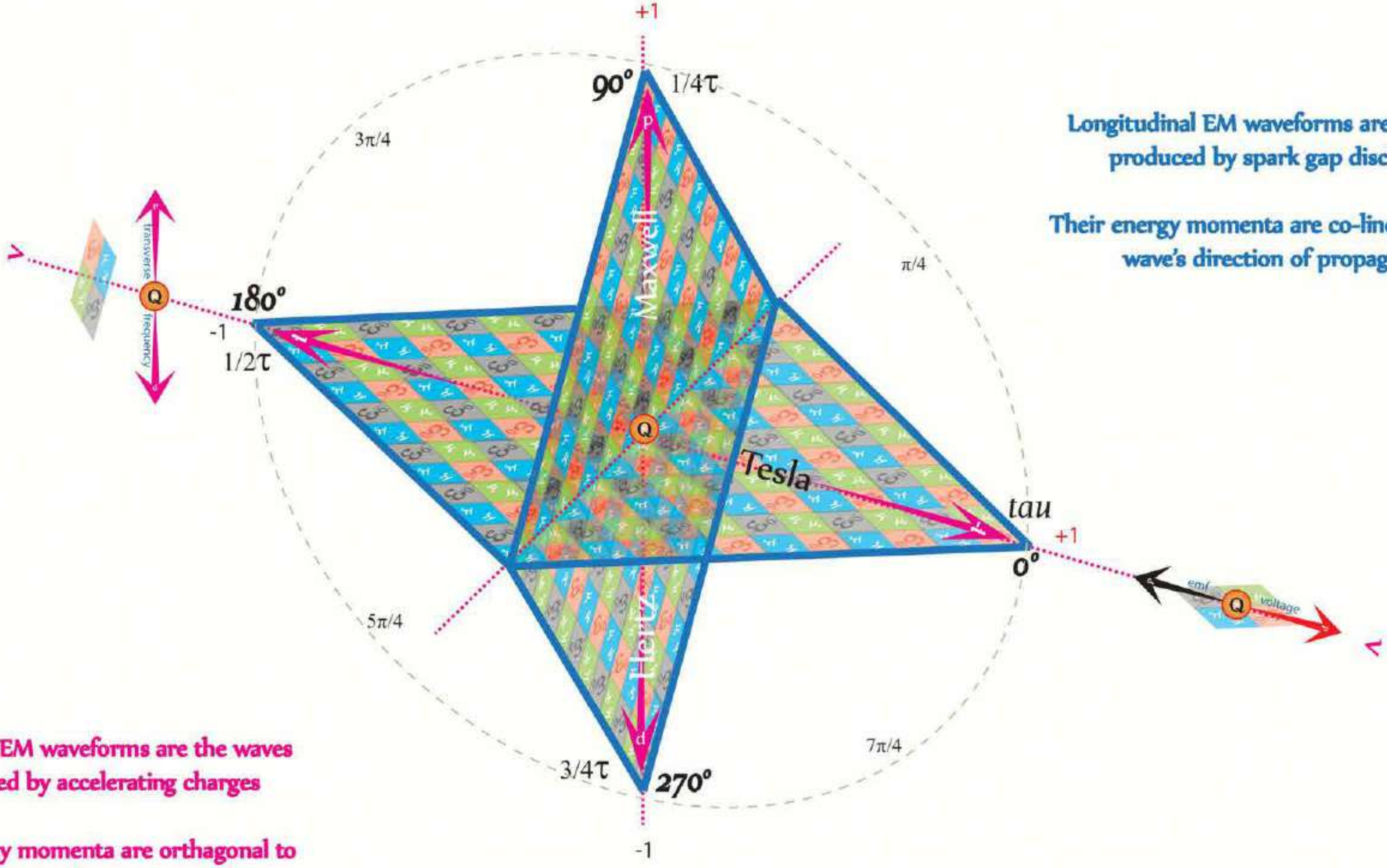
+1
90°
1/4τ
3π/4
π/4
transverse
frequency
Q
180°
-1
1/2τ
Maxwell
Q
Tesla
tau
+1
0°
emf
voltage
Q
Hertz
5π/4
7π/4
3/4τ
270°
-1

Longitudinal EM waveforms are the waves produced by spark gap discharges

Their energy momenta are co-linear with the wave's direction of propagation

p

mv

linear momentum

Transverse EM waveforms are the waves produced by accelerating charges

Their energy momenta are orthogonal to the direction of wave propagation

[Mathematics] *i* or *j* [Electrical theory]

square root of negative one

Velocity and Time dependent EM fields

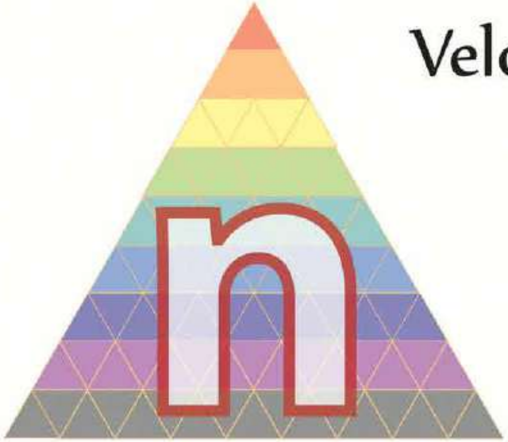

Quantum levels

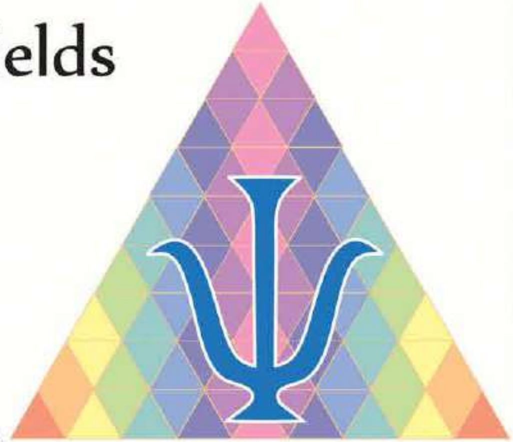

Wave probabilities

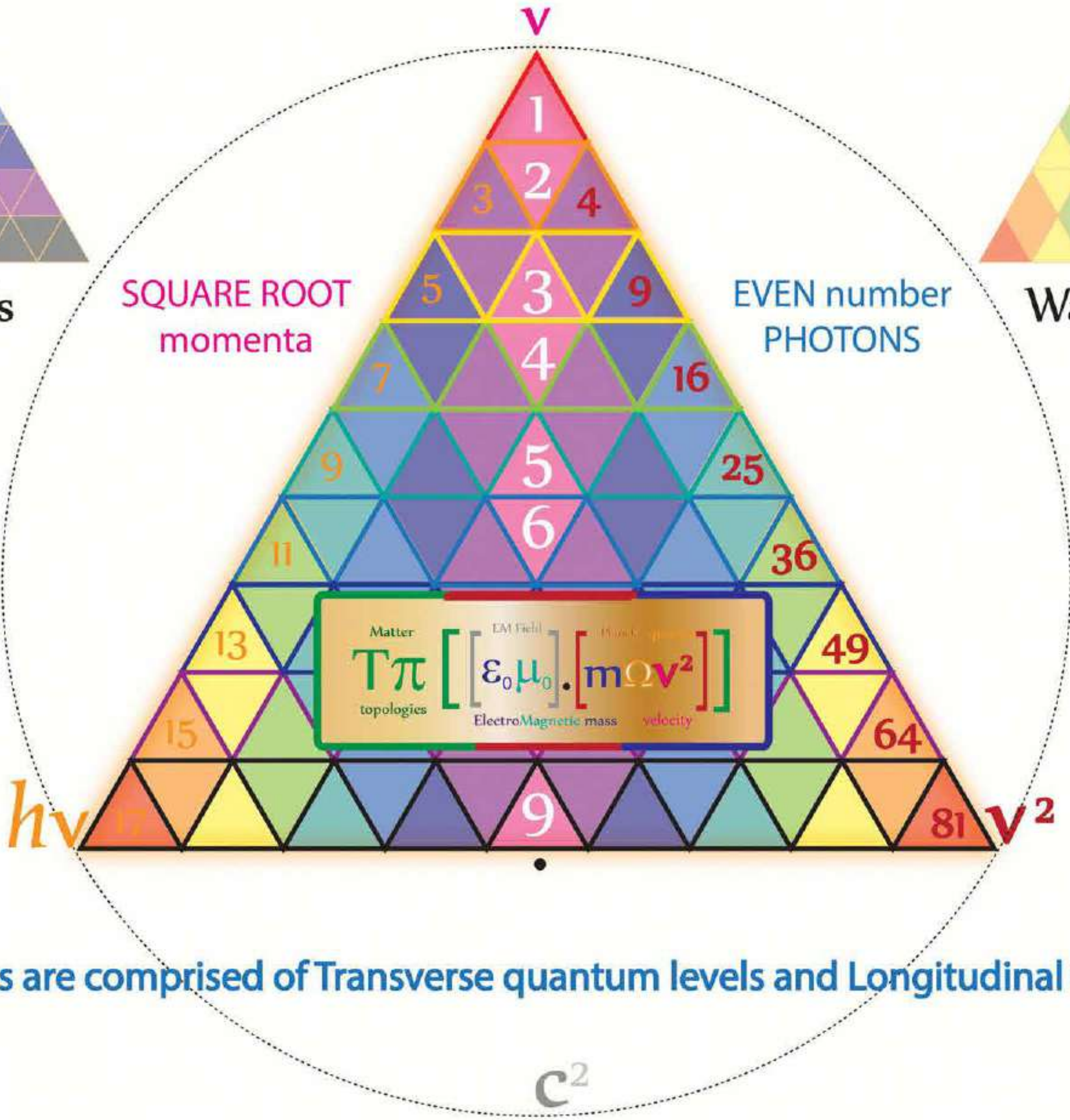
v
1
2
3
4
3
5
9
4
7
16
5
9
25
6
11
36
13
49
15
64
9
17
81
hv
v²
c²
Matter
topologies
EM Field
ElectroMagnetic mass
velocity

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

SQUARE ROOT
momenta

EVEN number
PHOTONS

ODD number
BOSONS

SQUARE number
ENERGIES

All EM geometries are comprised of Transverse quantum levels and Longitudinal wave probabilities

Copyright ABRAHAM [2008] - All rights reserved

Matrices

Matrices are a key tool in linear algebra. One use of matrices is to represent linear transformations, which are higher-dimensional analogs of linear functions of the form f(x) = cx, where c is a constant; matrix multiplication corresponds to composition of linear transformations.

Further developing equilateral Matrices and tensor mathematics to reflect the 2D geometry of EM, KEM and GEM quantum fields, along with the geometric quantisation of mass-energiy momenta and their energy distributions allows for field interactions to be accurately visualised and modelled

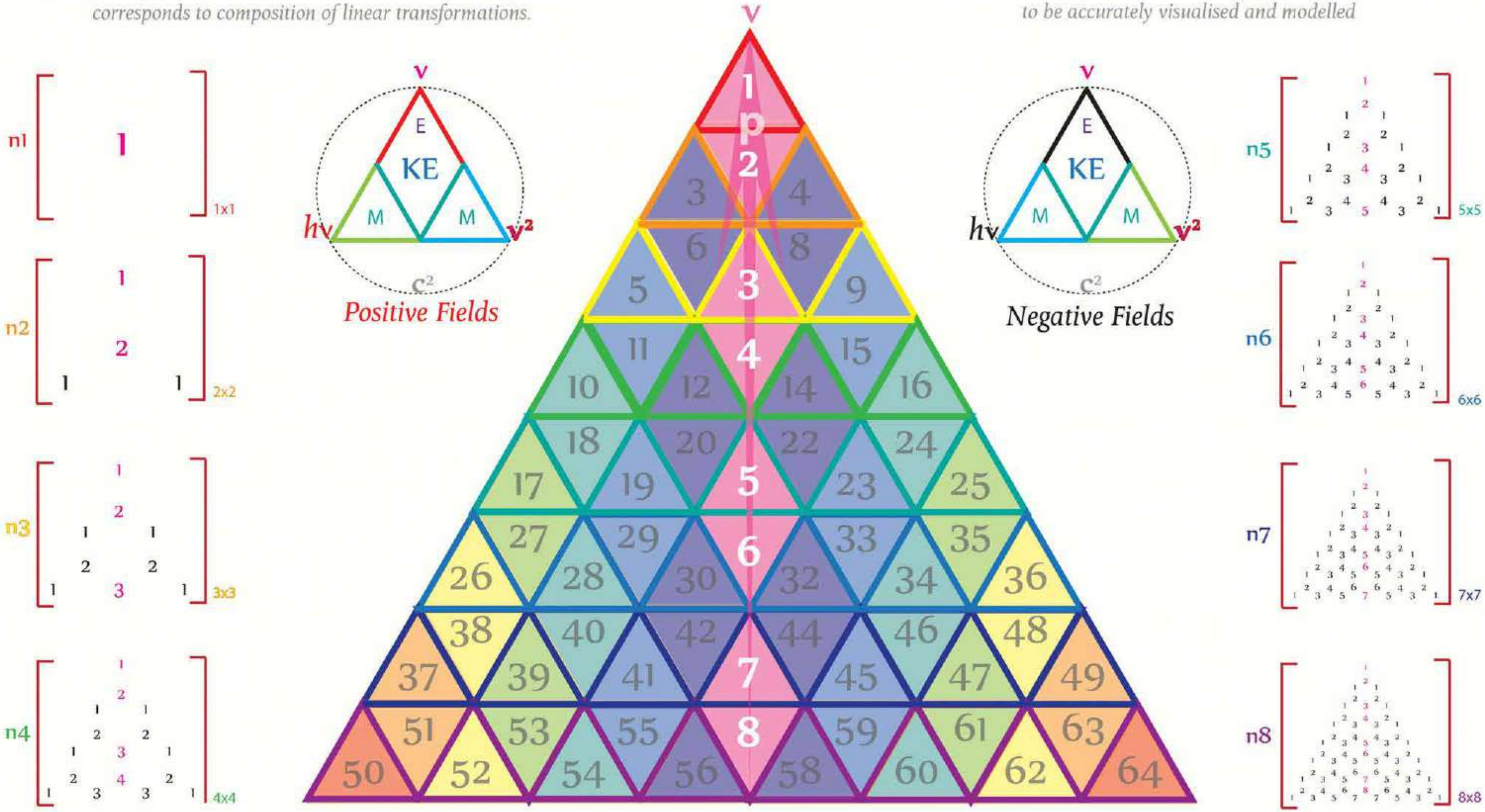

Tensors are geometric entities introduced into mathematics and physics to extend the notion of scalars, geometric vectors, and matrices to increasingly higher orders.

Modifying Square matrices to reflect the equialteral geometries of Tetryonic fields allows for the accurate geometric modelling of all Scalar & Vector fields along with their varied intrinsic quantum energies and physical properties

Copyright ABRAHAM [2008] - All rights reserved

Energy momenta Tensors

momenta
(a property of Energy)
is converative

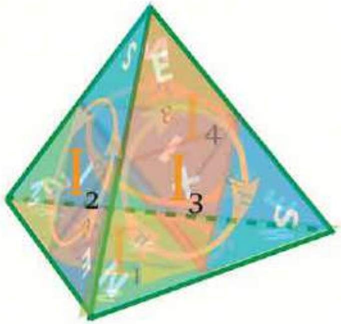

$$p_M \longrightarrow (E, p1, p2, p3, p4).$$

All standing-wave Matter topologies can be modelled using its Tetryonic charge energy momenta Tensors

with an additional Kinetic EM energy-momenta tensor required for Matter in motion

$$p_{KEM} \longrightarrow (E, p5).$$

Matter
(a geometric property)
is NOT conservative

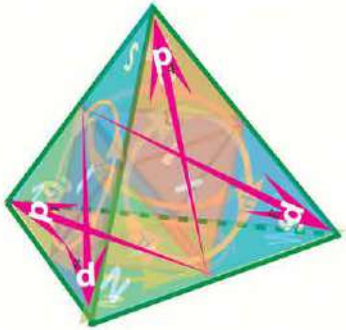

2D+1 [SR] mass-energy momenta can be folded into 3D+1 [GR] Matter that can be modelled using 4 Energy-momenta tensors

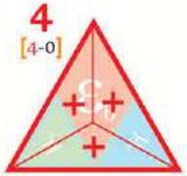

Positive standing-wave Matter

$$E = \sum_{\text{all fascia}} E$$

Energy of a massive particle is the total of all Planck quanta (compton frequency) in a 3D geometry

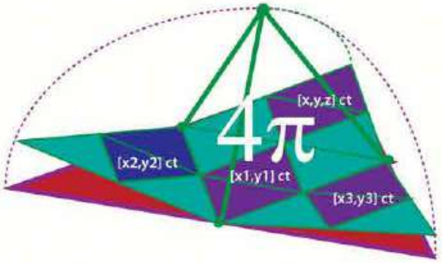

$$\vec{p} = \sum_{\text{all fascia}} \vec{p}$$

Total Momentum is the total of all quanta linear Momenta in a 3D particle

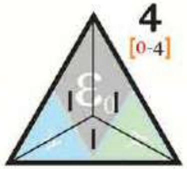

Negative standing-wave Matter

Relativistic Matter in motion

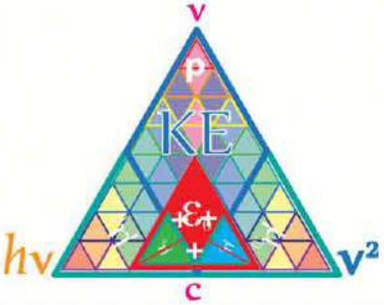

$$I = \sum_{\text{all fascia}} I$$

Inertial mass is the total of all inertia in a 3D particle

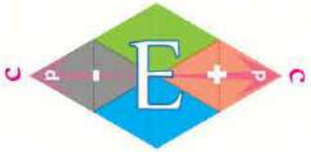

Photons have null energy-momentum tensors

Note that both 2D mass-energies [Special Relativity] and 3D Matter [General Relativity] have distinct Energy momenta

It is the 3D Tetrahedral topologies that provides a definative basis for Matter

Relativistic Matter in motion

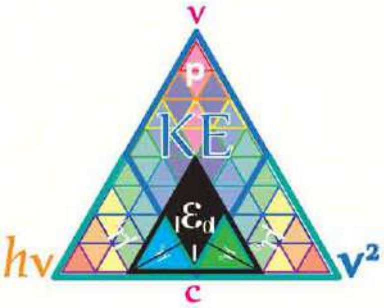

Tetryonic Energy momenta tensors should not be confused with Four vector tensors which map energy-momenta vectors in 3D spatial[cartesian] co-ordinate systems

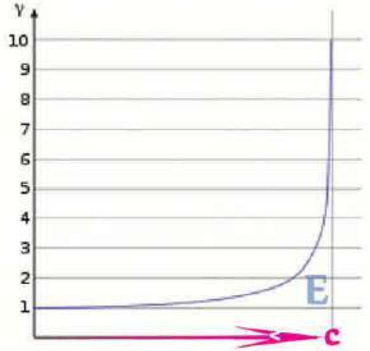

Relativistic Lorentz correction factor

The Lorentz factor or Lorentz term is an expression which appears in several equations in special relativity
It is often incorrectly graphed [as pictured to the right] as an increasing function of velocity approaching but never reaching c

**'c' is the maximum velocity acheivable through the electrical acceleration of particles
is NOT the maximum velocity possible for Matter in motion**

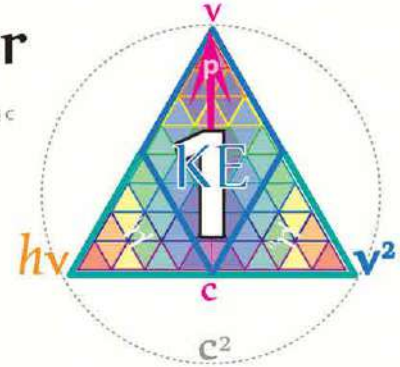

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

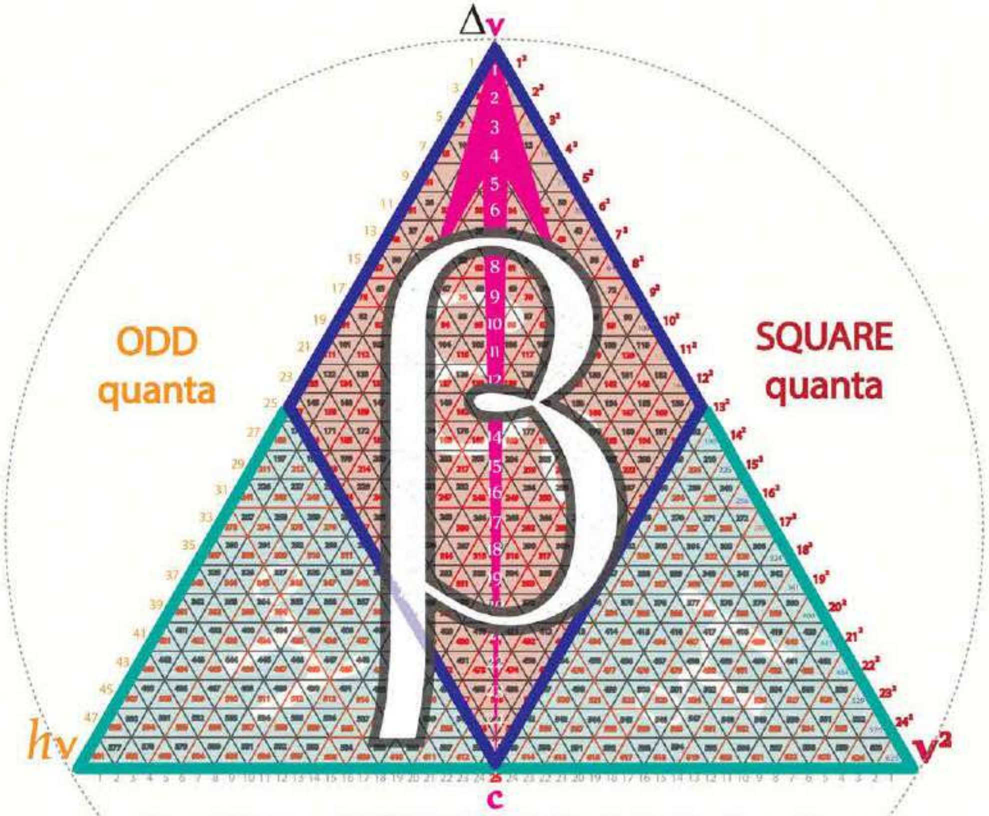

*Lorentz factors are LINEAR and SCALAR velocity related corrections
to the relativistic mass-energy momenta content of any
physical system accelerated by EM Forces*

Scalar correction factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

$$\frac{\Delta\gamma}{c^2}$$ Wavelength contraction

Time Dilation $$\frac{\Delta\Omega}{c^2}$$

$$\frac{\Delta mv}{c^2}$$ Relativistic momentum

$$\frac{\Gamma}{c^2}$$ Lorentz corrections

Relativistic mass-energy $$\frac{\Delta hv}{c^2}$$

Copyright ABRAHAM [2008] - All rights reserved

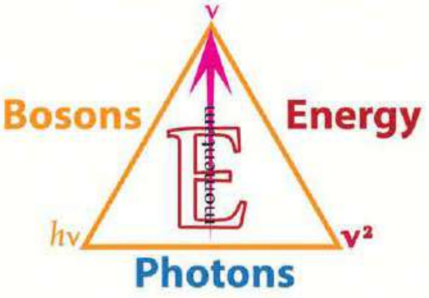
v
Bosons
Energy
momentum
E
hv
v²
Photons

form the basis for all charged energy geometries and their associated quanta distributions

Equilateral energies

Any measurements of mass without a Space-Time co-ordinate system are measurements of Energy

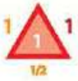

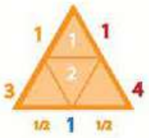
1
3
1
4
1
2
1/2 1 1/2

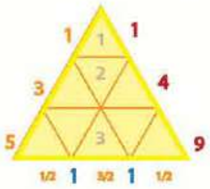
1
3
5
1
4
9
1
2
3
1/2 1 3/2 1 1/2

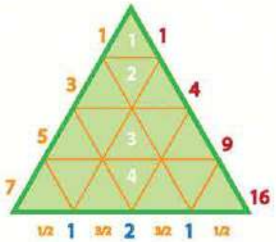
1
3
5
7
1
4
9
16
1
2
3
4
1/2 1 3/2 2 3/2 1 1/2

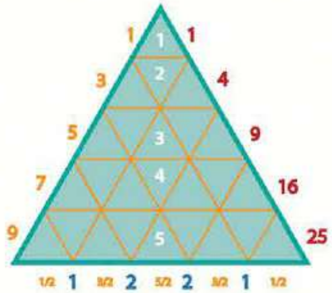
1
3
5
7
9
1
4
9
16
25
1
2
3
4
5
1/2 1 3/2 2 5/2 2 3/2 1 1/2

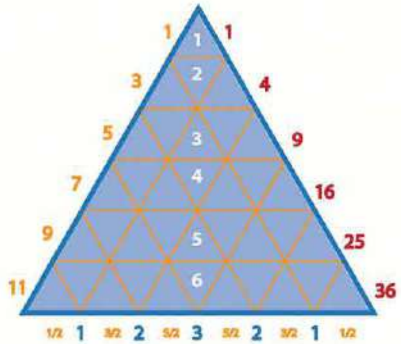
1
3
5
7
9
11
1
4
9
16
25
36
1
2
3
4
5
6
1/2 1 3/2 2 5/2 3 5/2 2 3/2 1 1/2

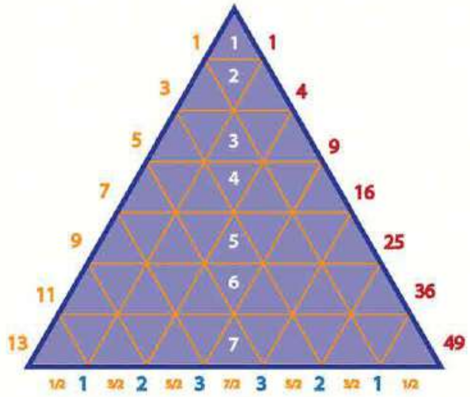
1
3
5
7
9
11
13
1
4
9
16
25
36
49
1
2
3
4
5
6
7
1/2 1 3/2 2 5/2 3 7/2 3 5/2 2 3/2 1 1/2

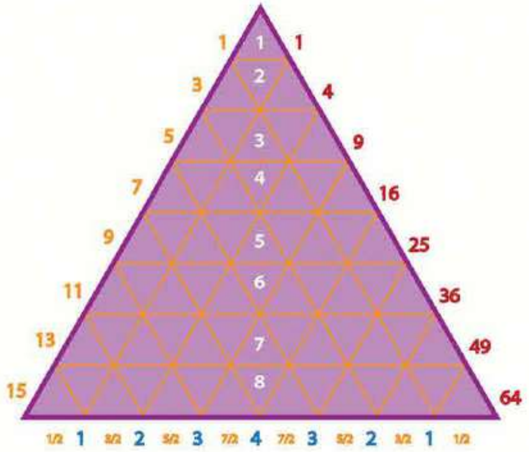
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
1
2
3
4
5
6
7
8
1/2 1 3/2 2 5/2 3 7/2 4 7/2 3 5/2 2 3/2 1 1/2

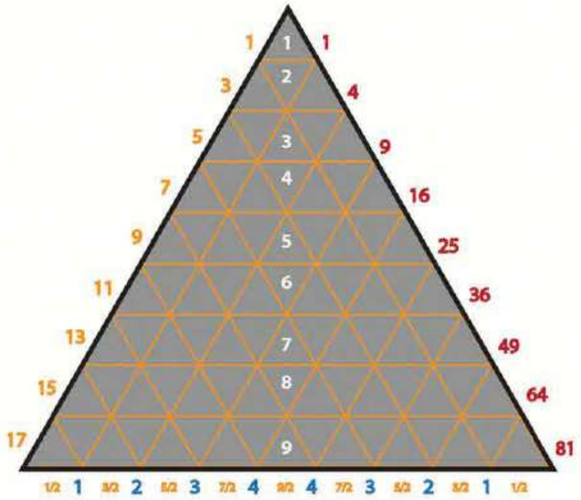
1
3
5
7
9
11
13
15
17
1
4
9
16
25
36
49
64
81
1
2
3
4
5
6
7
8
9
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

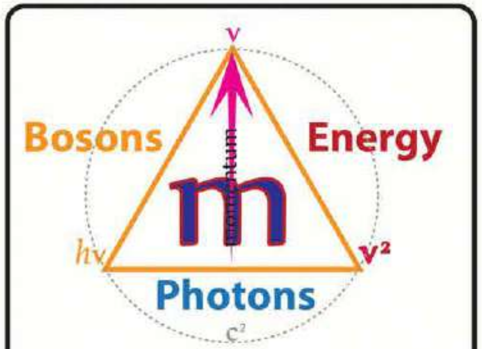
v
Bosons
Energy
momentum
m
hv
v²
Photons
c²

Time specific measurements of equilateral ENERGY momenta geometries forms the basis for EM mass

Electromagnetic mass

2D mass-energy is the surface integral of 3D Matter

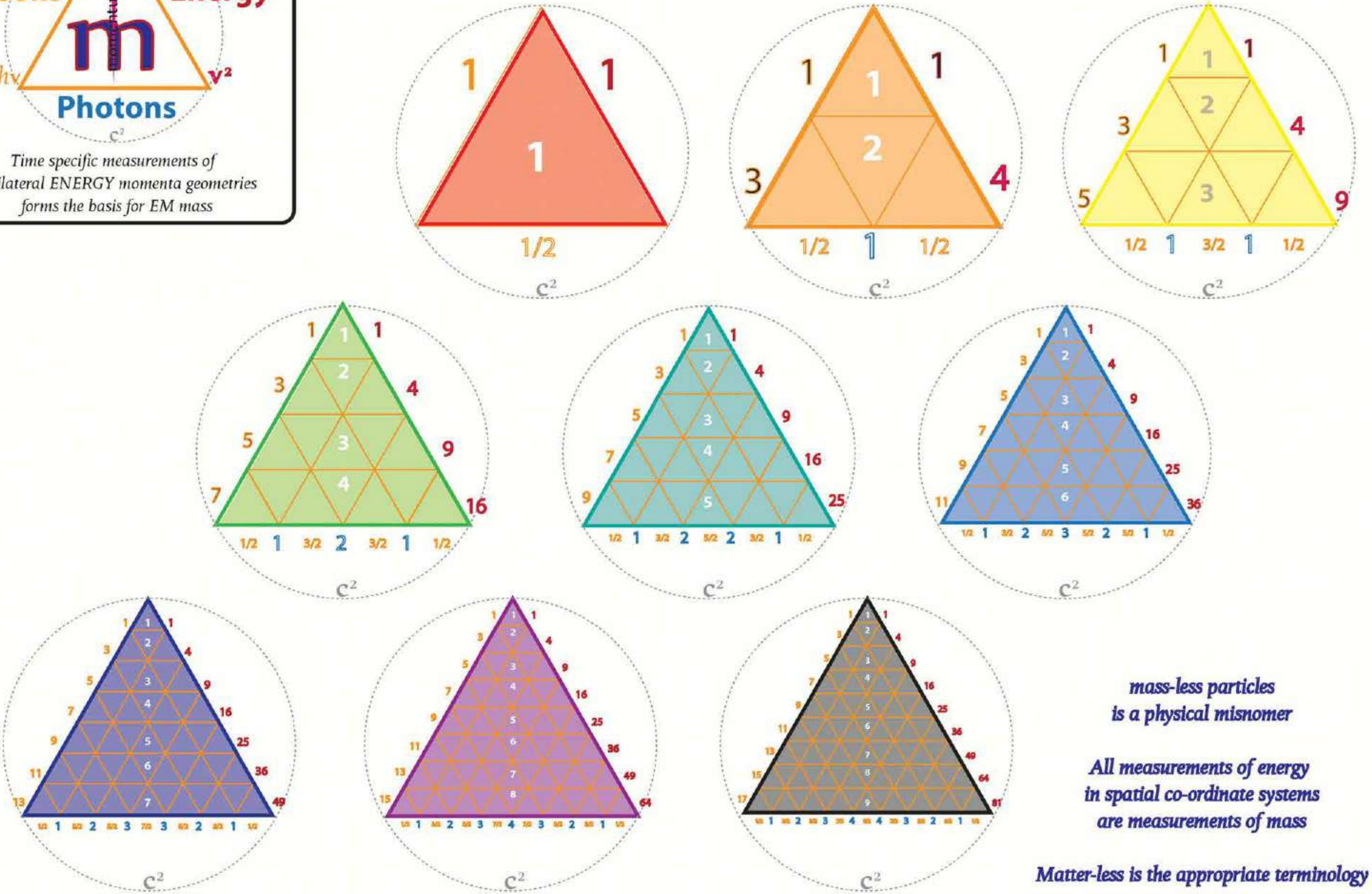
1 1 1
1/2
c²
1 1 1 2 3 4
1/2 1 1/2
c²
1 1 1 2 3 4 3 5 9
1/2 1 3/2 1 1/2
c²
1 1 1 2 3 4 3 5 9 4 7 16
1/2 1 3/2 2 3/2 1 1/2
c²
1 1 1 2 3 4 3 5 9 4 7 16 5 9 25
1/2 1 3/2 2 5/2 2 3/2 1 1/2
c²
1 1 1 2 3 4 3 5 9 4 7 16 5 9 25 6 11 36
1/2 1 3/2 2 5/2 3 5/2 2 3/2 1 1/2
c²
1 1 1 2 3 4 3 5 9 4 7 16 5 9 25 6 11 36 7 13 49
1/2 1 3/2 2 5/2 3 7/2 3 5/2 2 3/2 1 1/2
c²
1 1 1 2 3 4 3 5 9 4 7 16 5 9 25 6 11 36 7 13 49 8 15 64
1/2 1 3/2 2 5/2 3 7/2 4 7/2 3 5/2 2 3/2 1 1/2
c²
1 1 1 2 3 4 3 5 9 4 7 16 5 9 25 6 11 36 7 13 49 8 15 64 9 17 81
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2
c²

mass-less particles
is a physical misnomer

All measurements of energy
in spatial co-ordinate systems
are measurements of mass

Matter-less is the appropriate terminology

Standing-wave Matter

Matter is a higher order 3D topology created by standing wave mass-energies

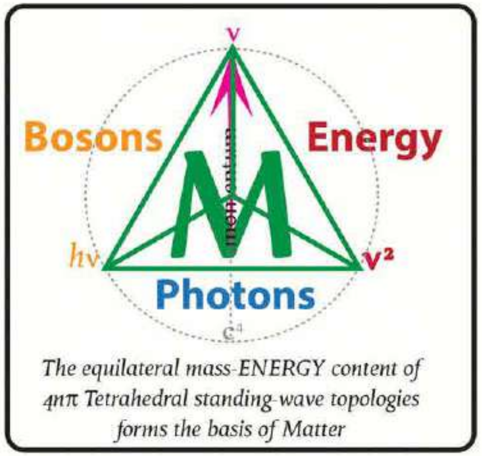
v
Bosons
Energy
M
momentum
hv
v²
Photons
c^4

The equilateral mass-ENERGY content of 4nπ Tetrahedral standing-wave topologies forms the basis of Matter

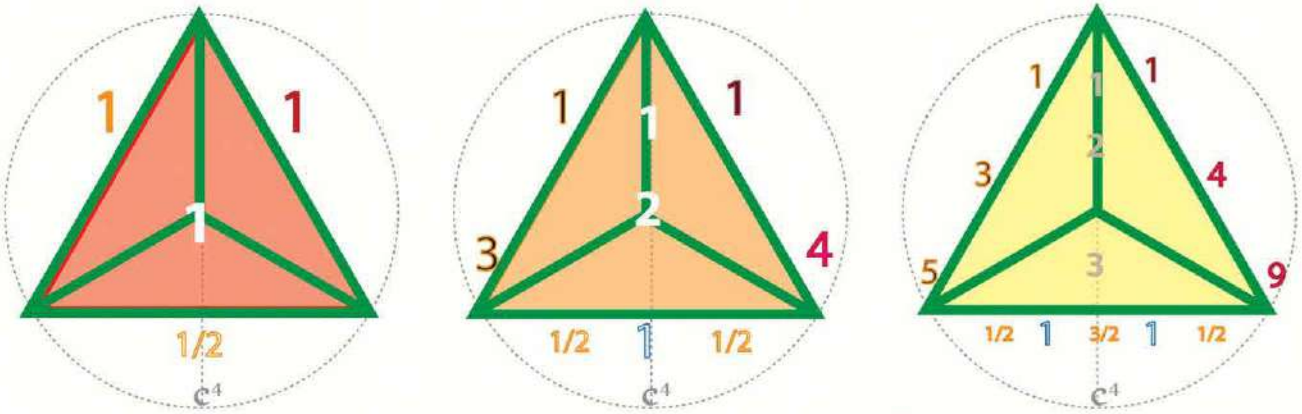
1
1
1
1/2
c^4
1
1
1
2
3
4
1/2
1
1/2
c^4
1
1
1
2
3
3
4
5
9
1/2
1
3/2
1
1/2
c^4

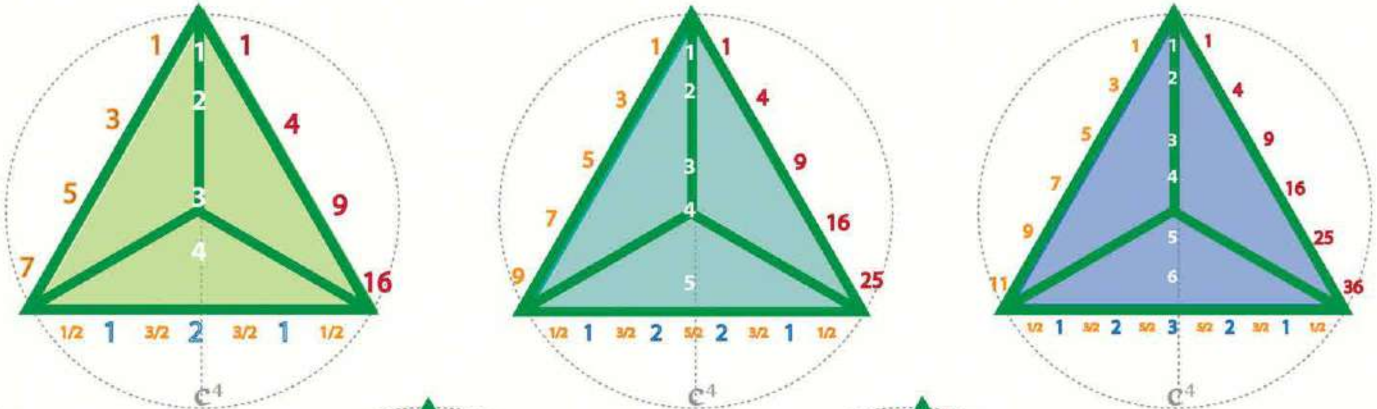
1
3
5
7
1
4
9
16
1
2
3
4
1/2
1
3/2
2
3/2
1
1/2
c^4
1
3
5
7
9
1
4
9
16
25
1
2
3
4
5
1/2
1
3/2
2
5/2
2
3/2
1
1/2
c^4
1
3
5
7
9
11
1
4
9
16
25
36
1
2
3
4
5
6
1/2
1
3/2
2
5/2
3
5/2
2
3/2
1
1/2
c^4

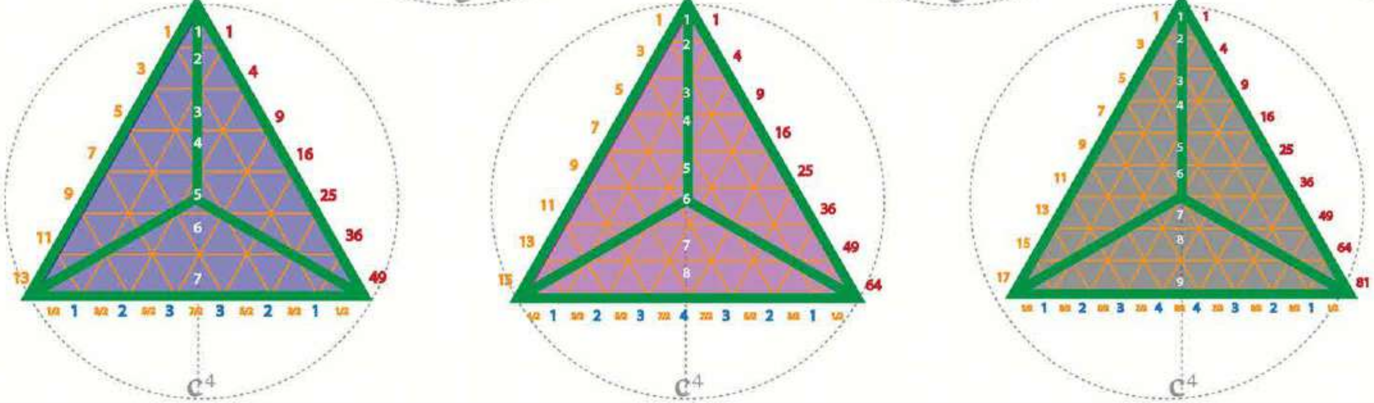
1
3
5
7
9
11
13
1
4
9
16
25
36
49
c^4
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
c^4
1
3
5
7
9
11
13
15
17
1
4
9
16
25
36
49
64
81
c^4

Only mass-ENERGIES contained in the tetryonic fascia of standing-wave topologies contribute to weight

Periodic element nuclei

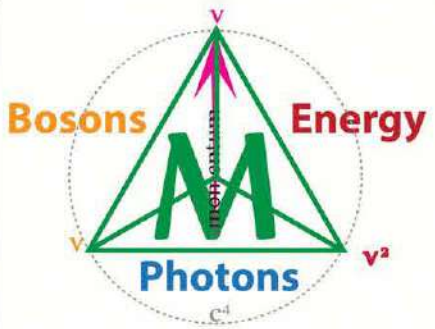
v
Bosons
Energy
M
momentum
v
v²
Photons
c4

The mass-ENERGY content of
4nπ Tetrahedral standing-wave topologies
forms the basis of all periodic elements

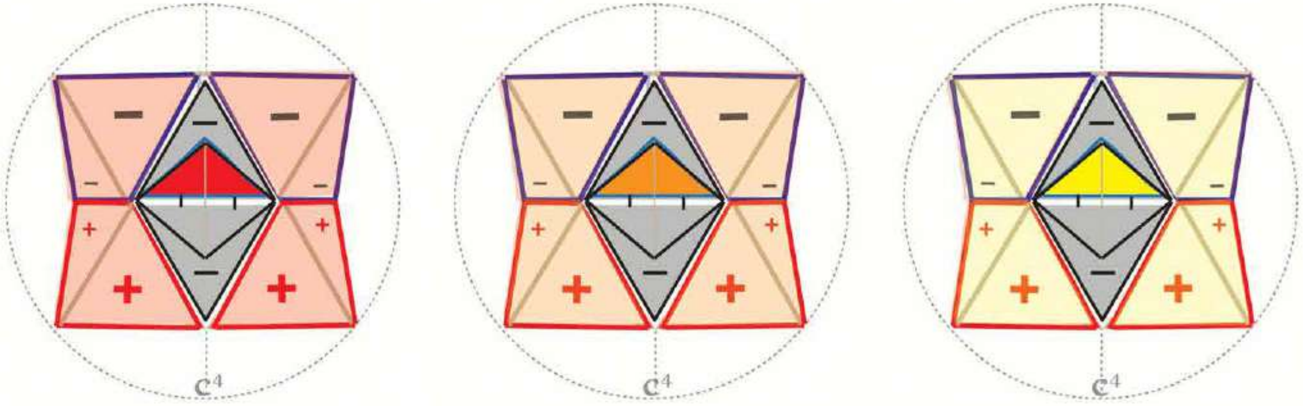
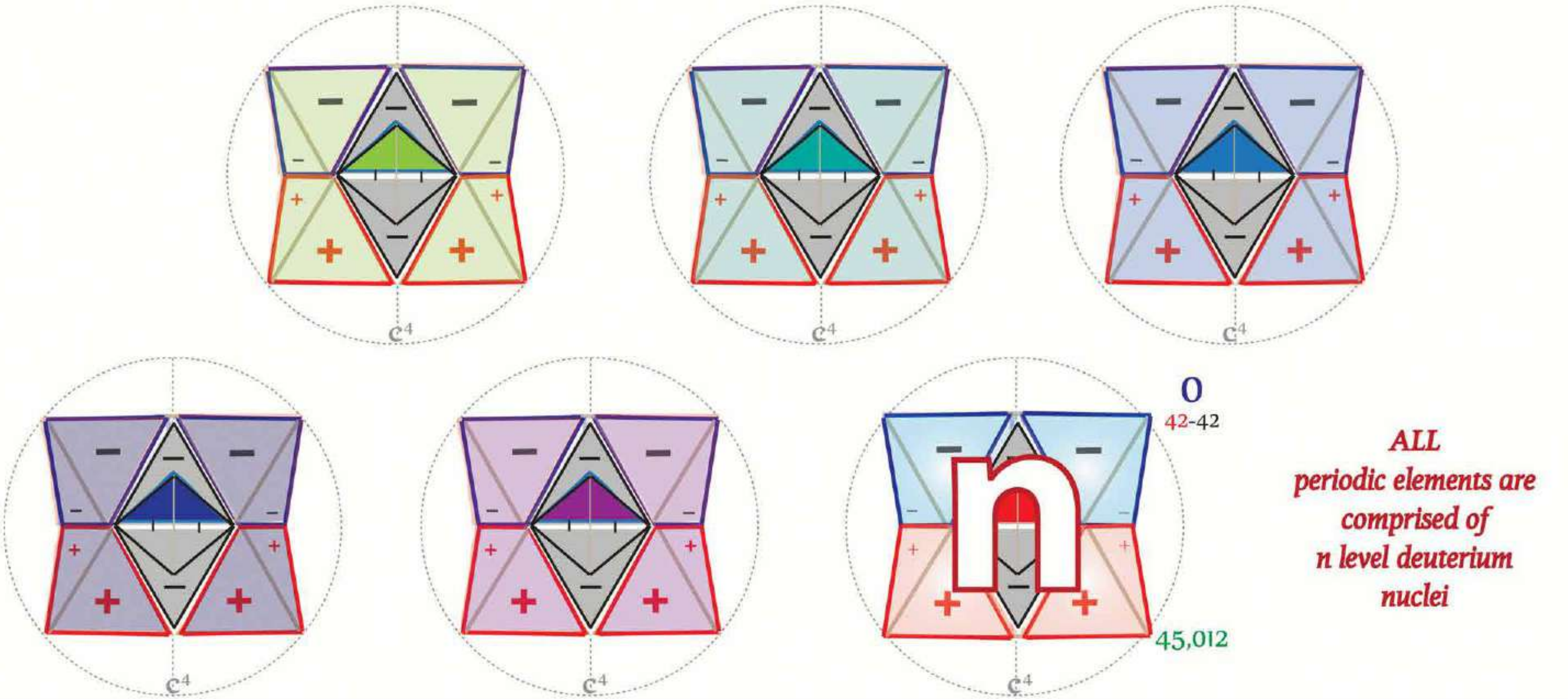
c4
n
0
42-42
45,012

ALL
periodic elements are
comprised of
n level deuterium
nuclei

Copyright ABRAHAM [2008] - All rights reserved

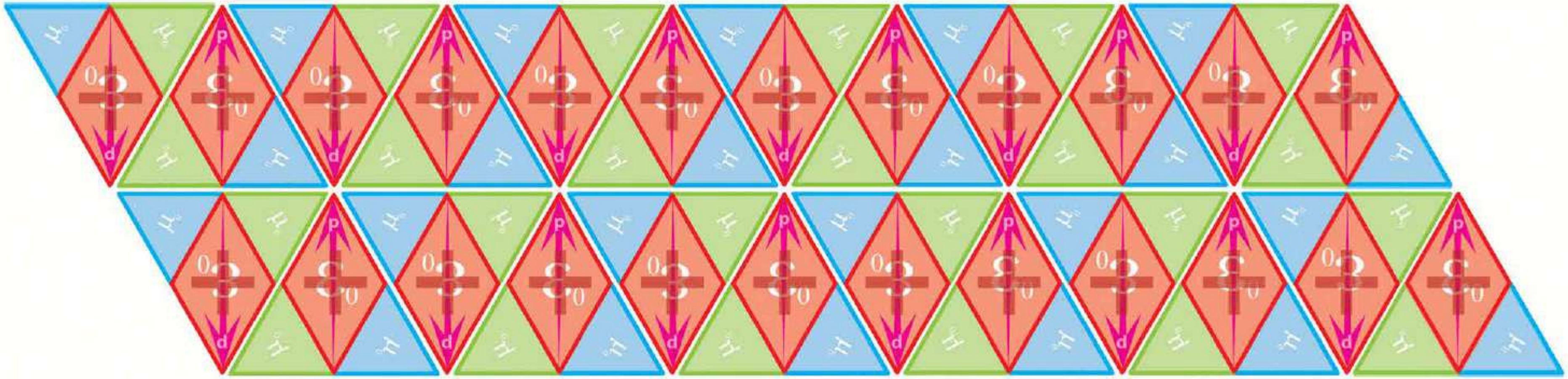

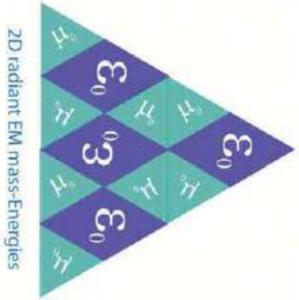
2D radiant EM mass-Energies
ε0
μ0

n1 - Charged Tetryon Templates

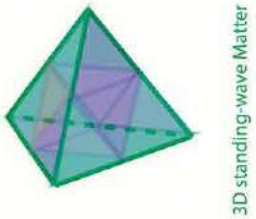
3D standing-wave Matter

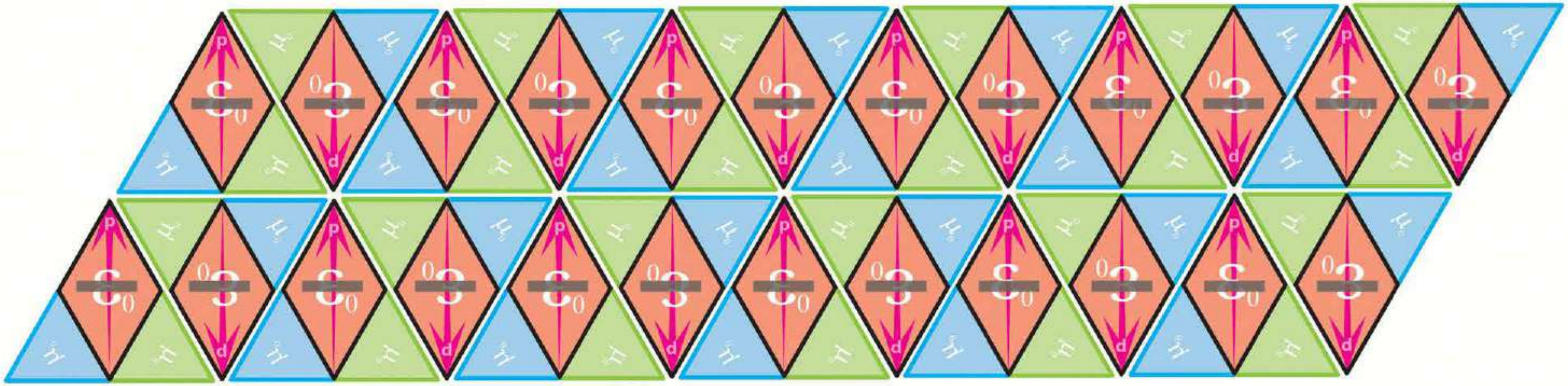

Copyright ABRAHAM [2008] - All rights reserved

n1 - Neutral Tetryon Templates

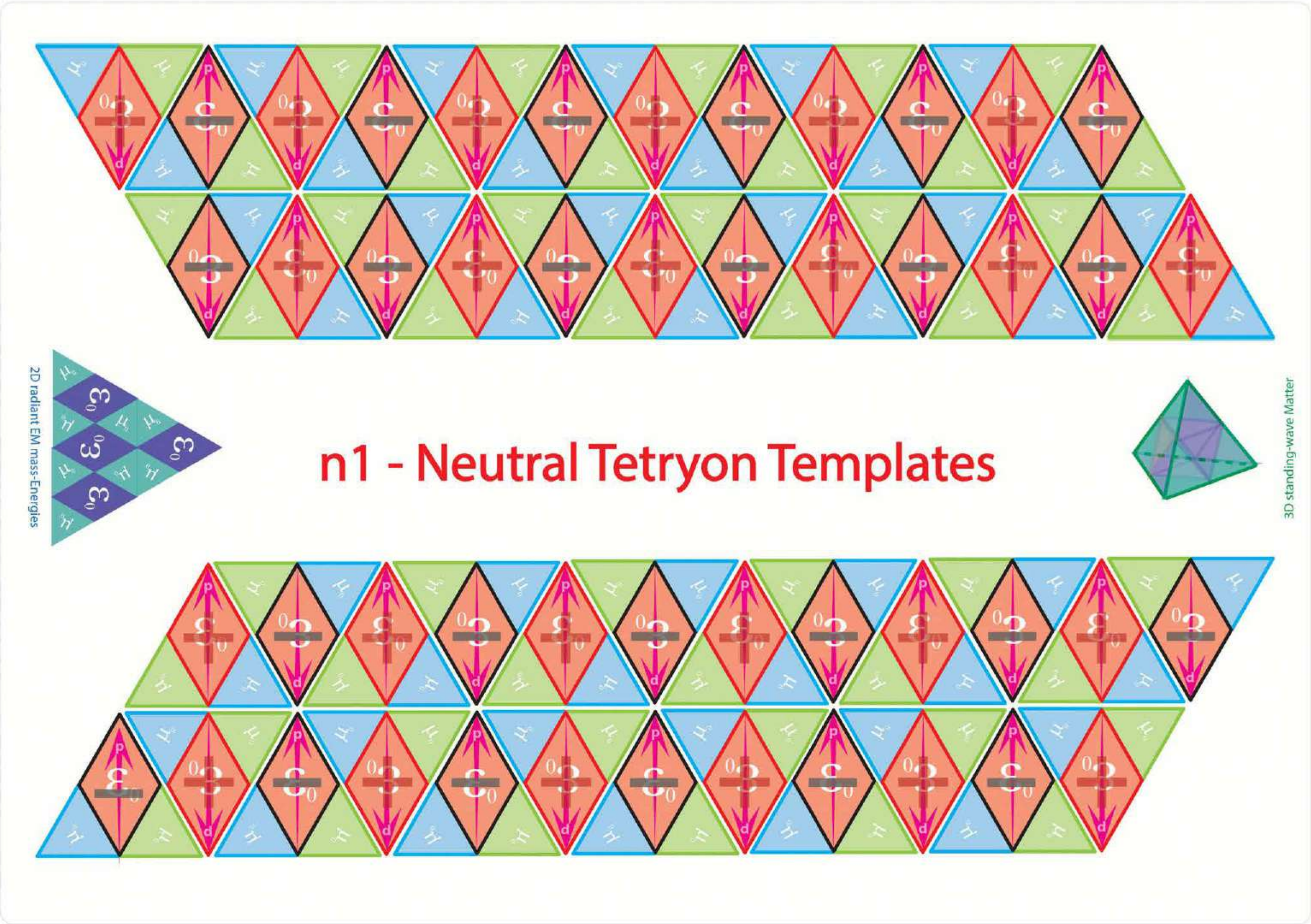
2D radiant EM mass-Energies
3D standing-wave Matter
p
d
μ_0
ε_0

Copyright ABRAHAM [2008] - All rights reserved

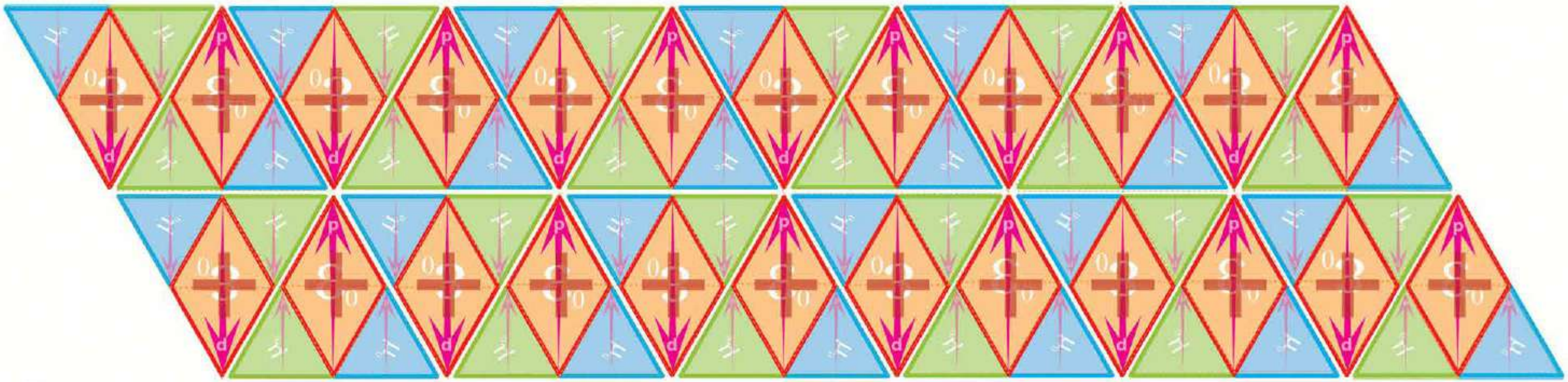

2D radiant EM mass-Energies

n2 - Charged Tetryon Templates

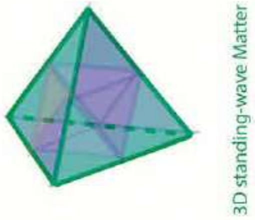
3D standing-wave Matter

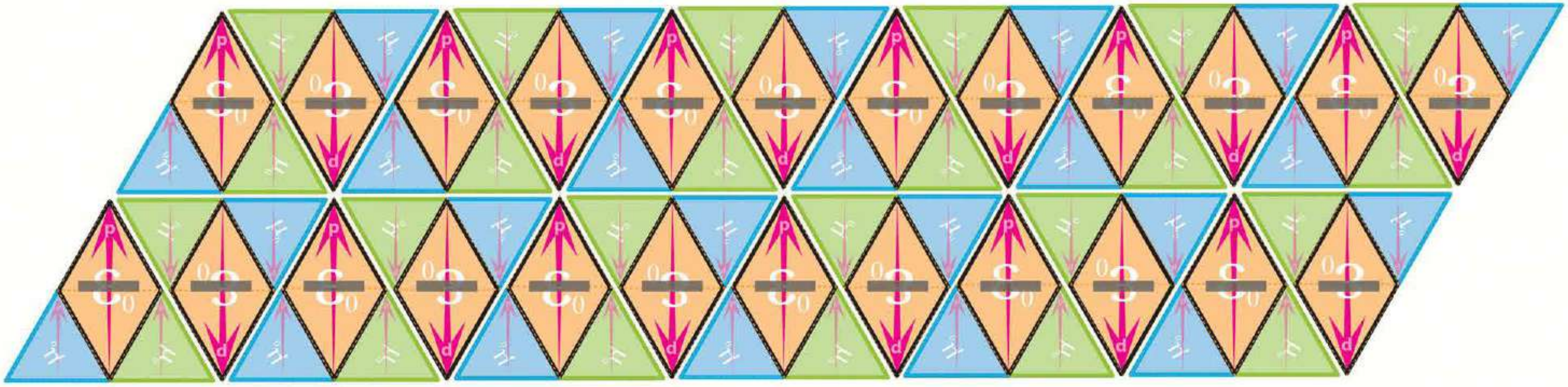

Copyright ABRAHAM [2008] - All rights reserved

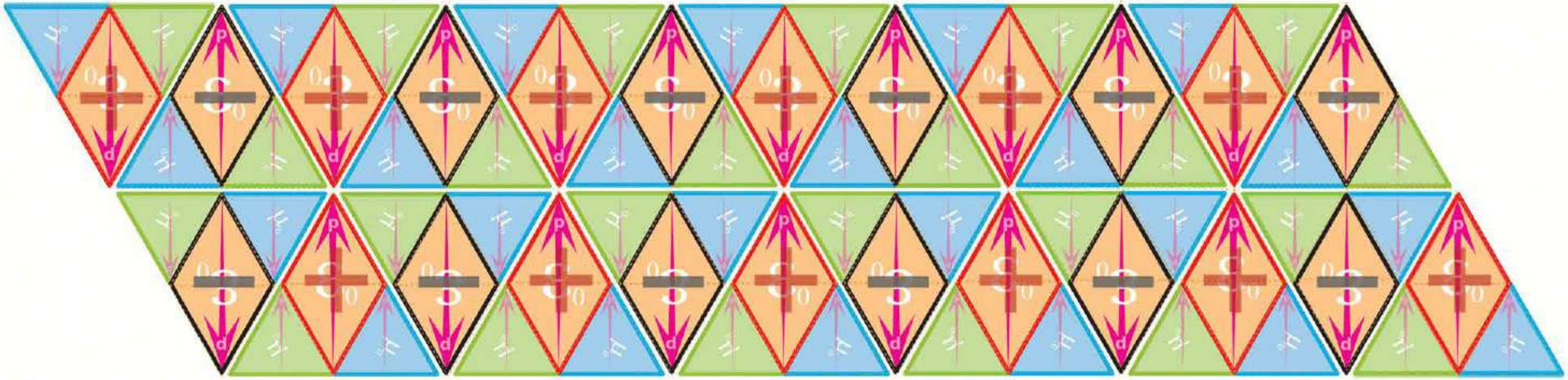

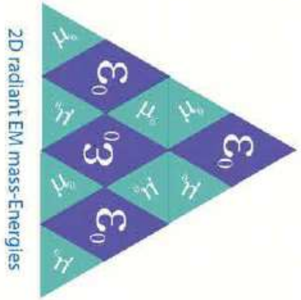
2D radiant EM mass-Energies

n2 - Neutral Tetryon Templates

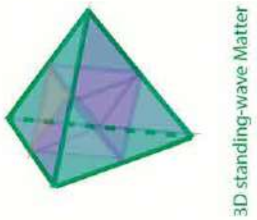
3D standing-wave Matter

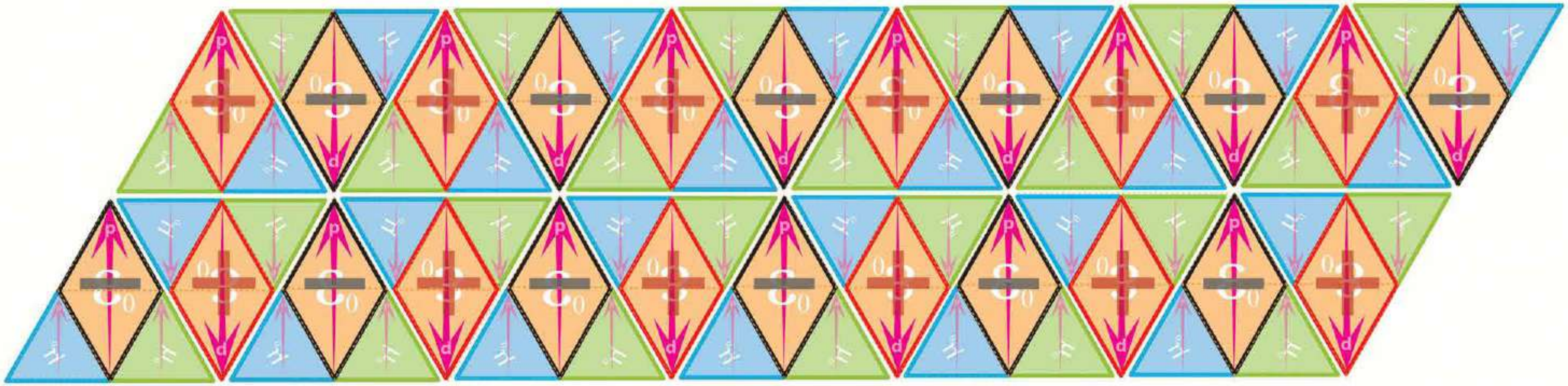

Copyright ABRAHAM [2008] - All rights reserved

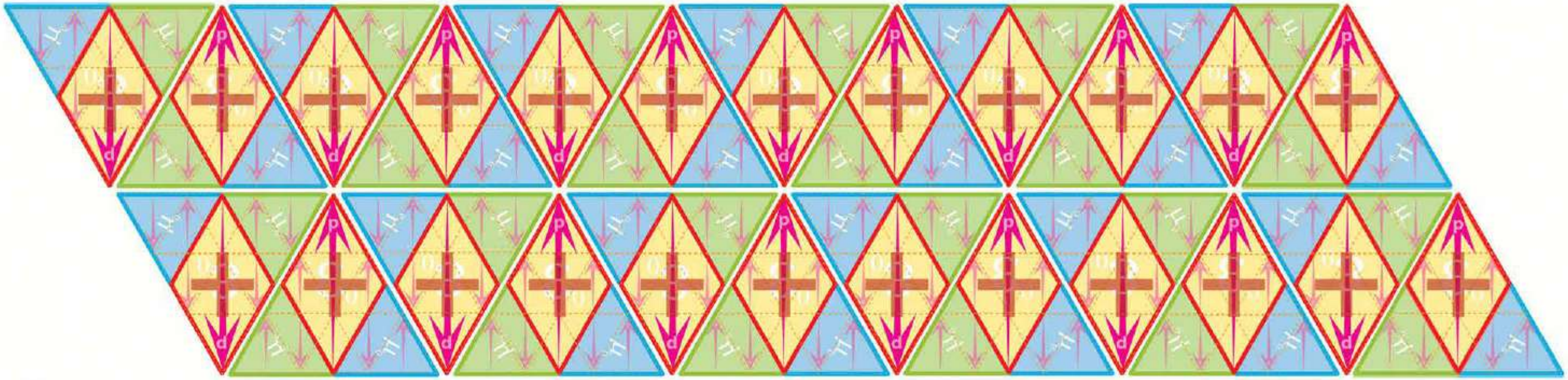
p
d
μ_0
ε_0

2D radiant EM mass-Energies
μ_0
ε_0

n3 - Charged Tetryon Templates

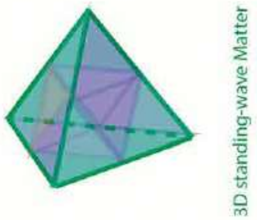
3D standing-wave Matter

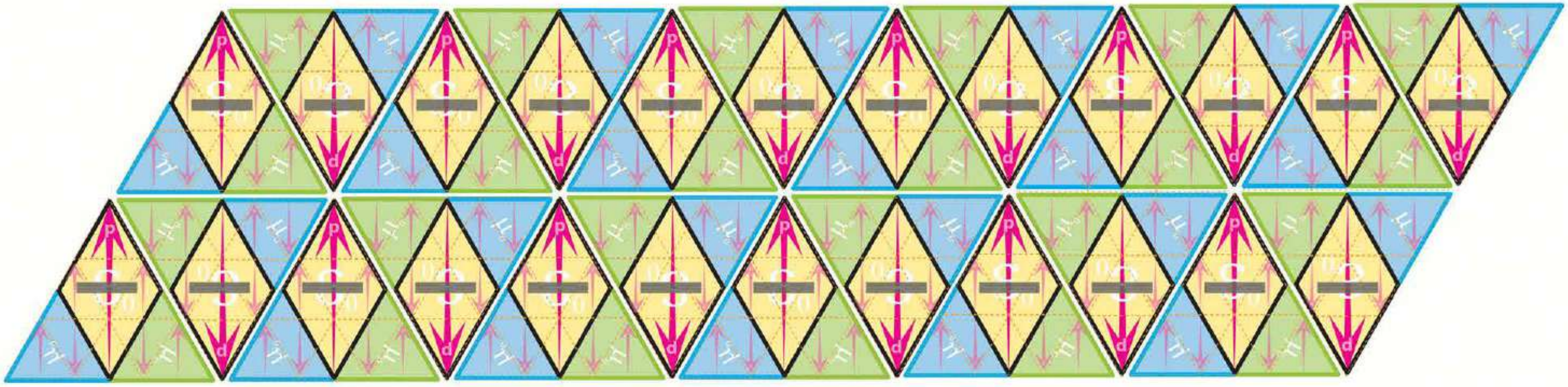
p
d
μ_0
ε_0

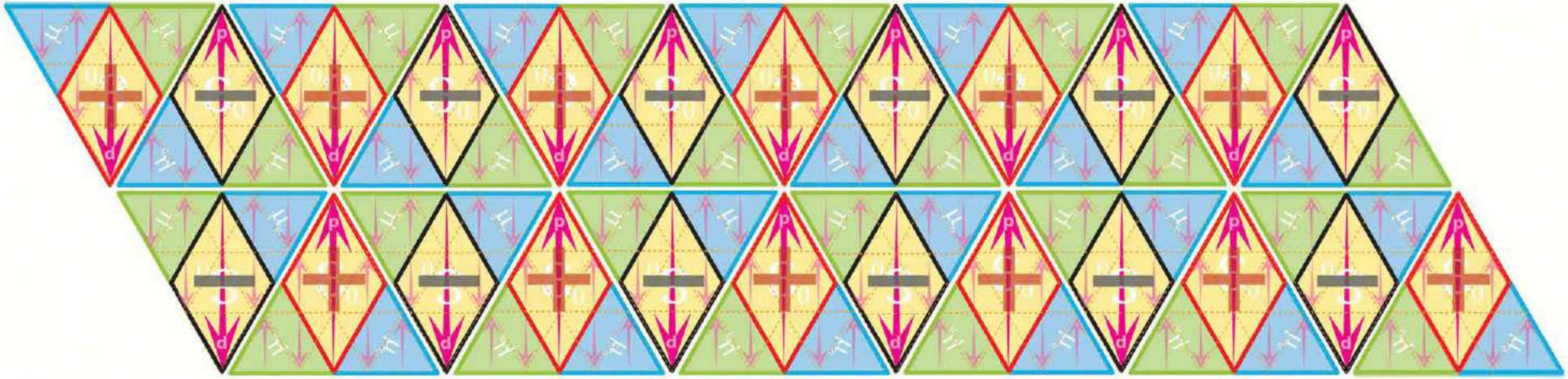

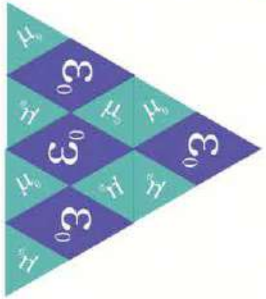

2D radiant EM mass-Energies

n3 - Neutral Tetryon Templates

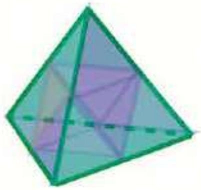

3D standing-wave Matter

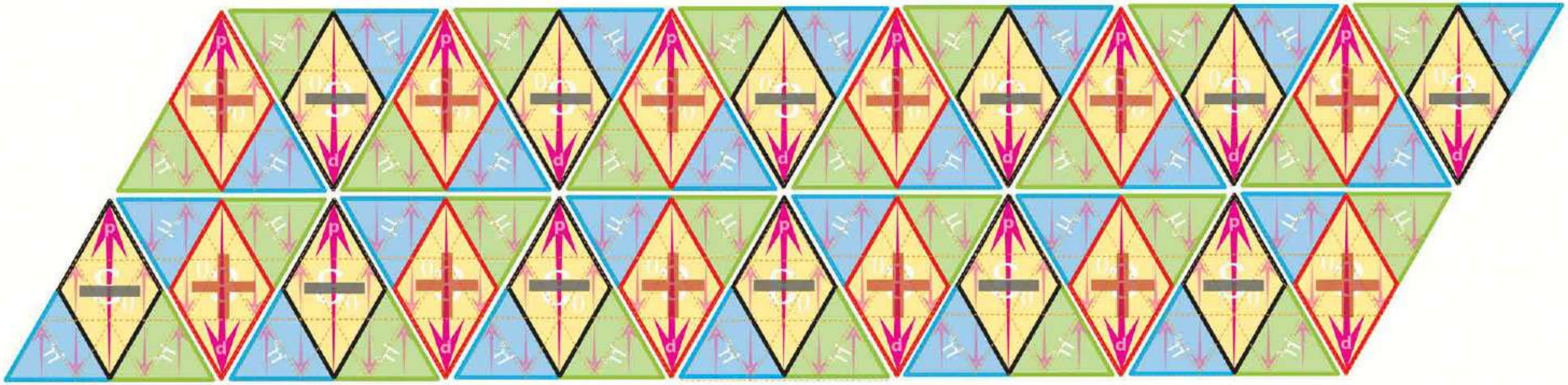

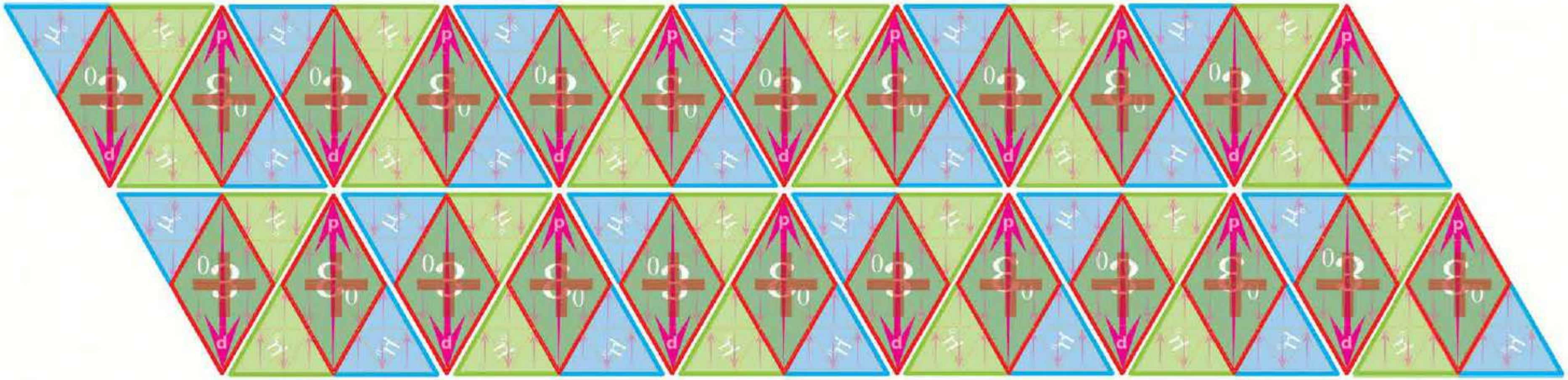

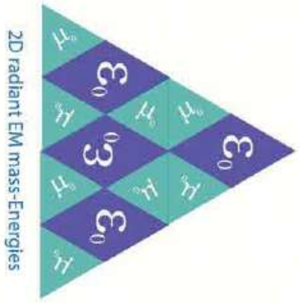
2D radiant EM mass-Energies
ε₀
μ₀

n4 - Charged Tetryon Templates

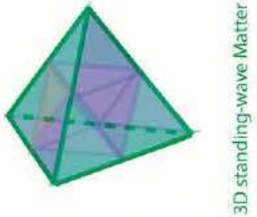
3D standing-wave Matter

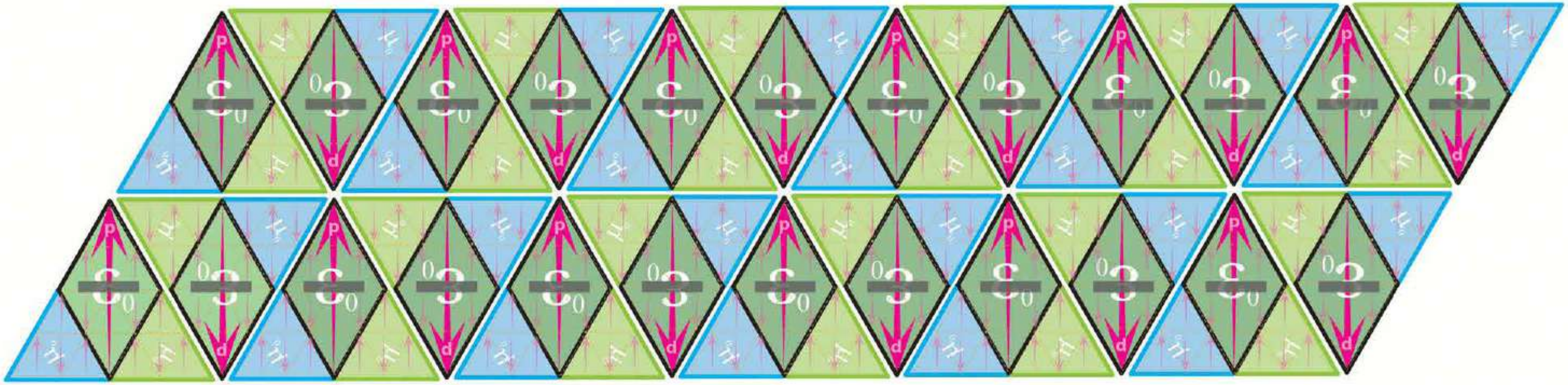

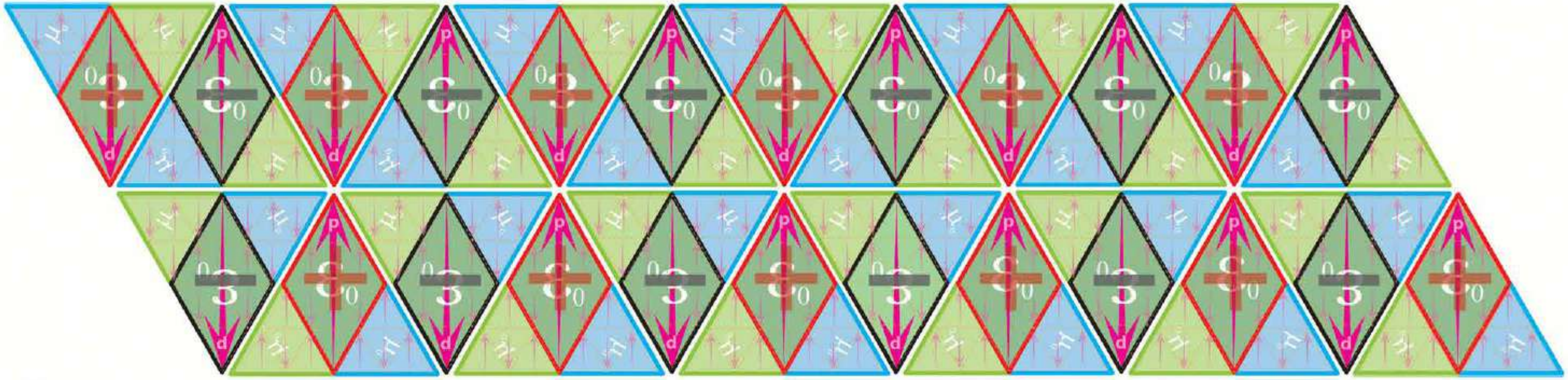

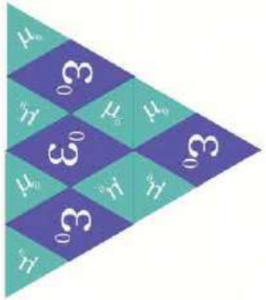

2D radiant EM mass-Energies

n4 - Neutral Tetryon Templates

3D standing-wave Matter

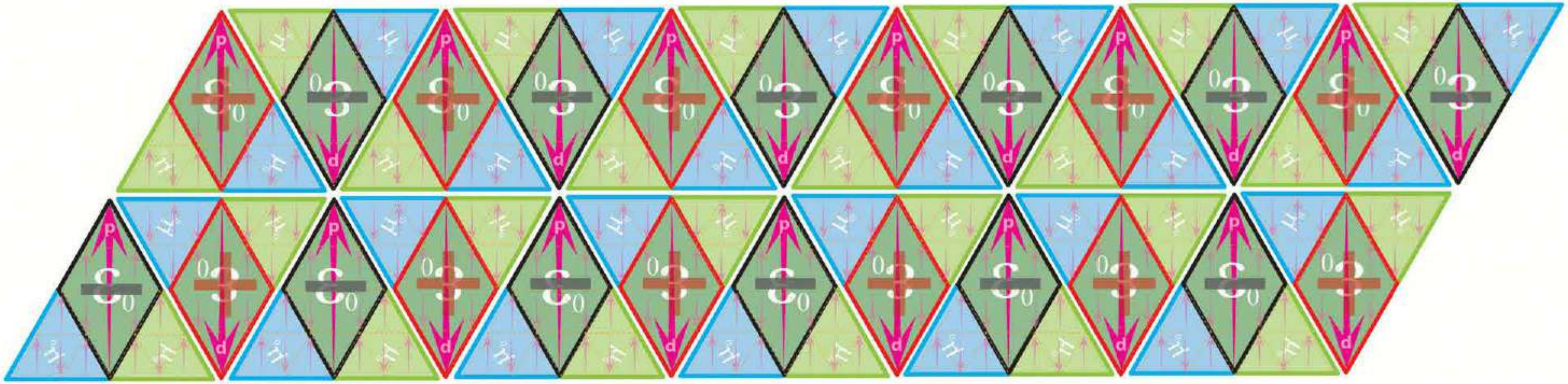

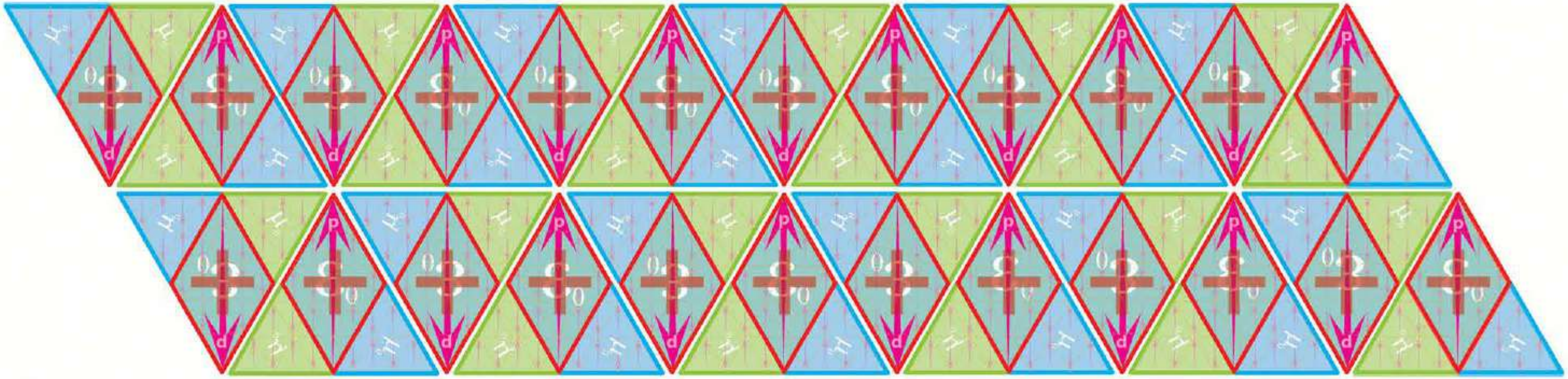

2D radiant EM mass-Energies

n5 - Charged Tetryon Templates

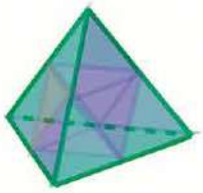

3D standing-wave Matter

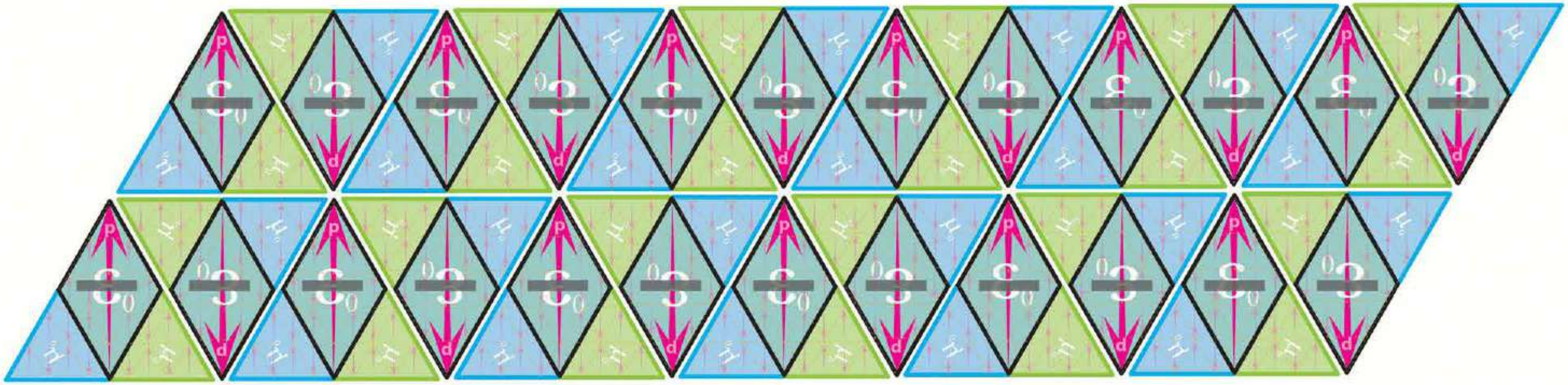

Copyright ABRAHAM [2008] - All rights reserved

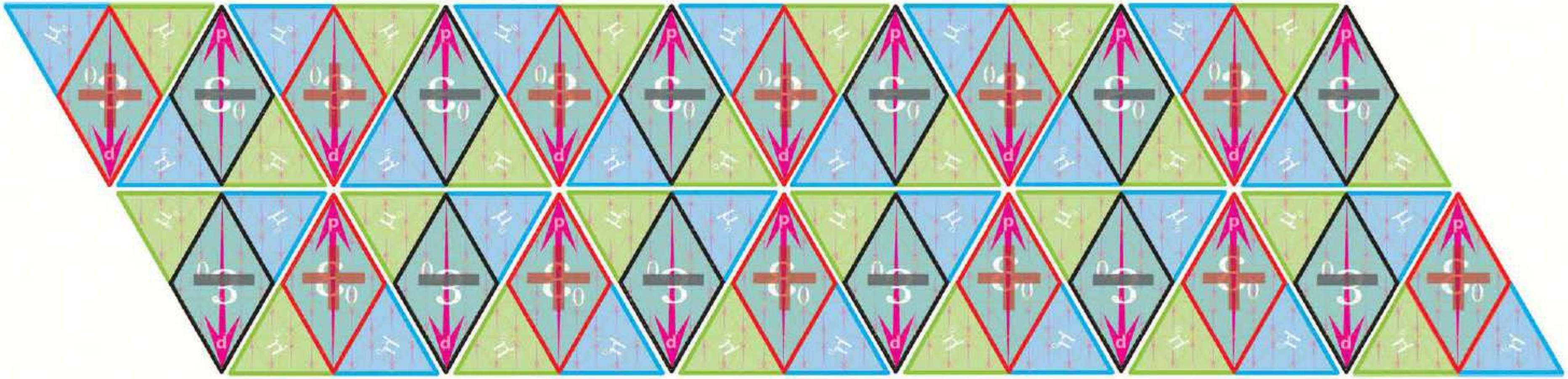

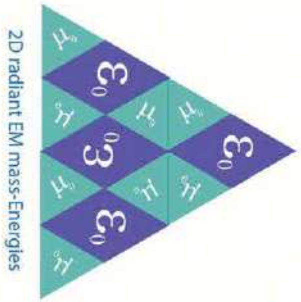
2D radiant EM mass-Energies
ε0
μ0

n5 - Neutral Tetryon Templates

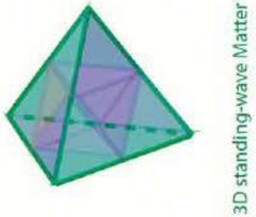
3D standing-wave Matter

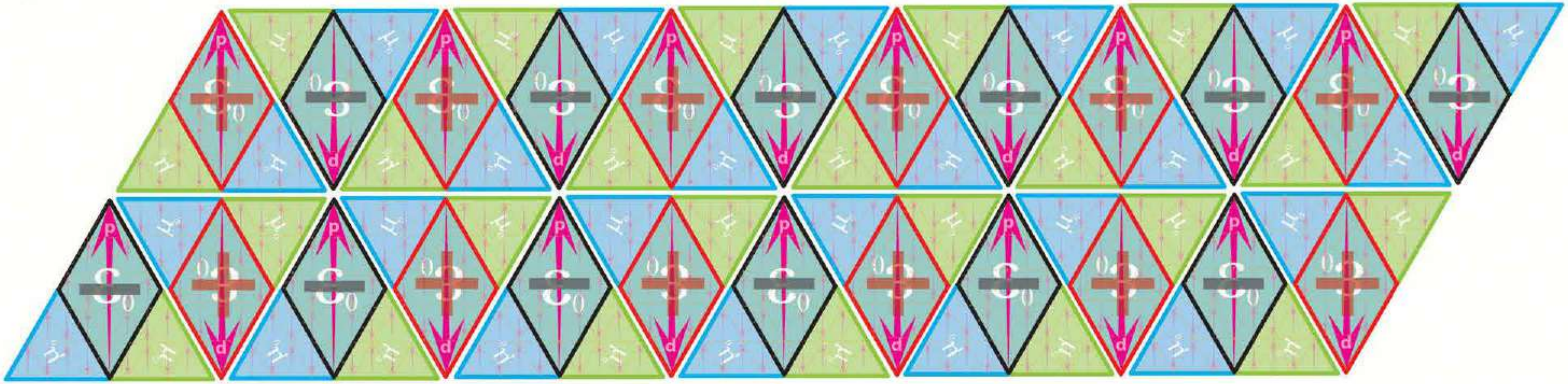

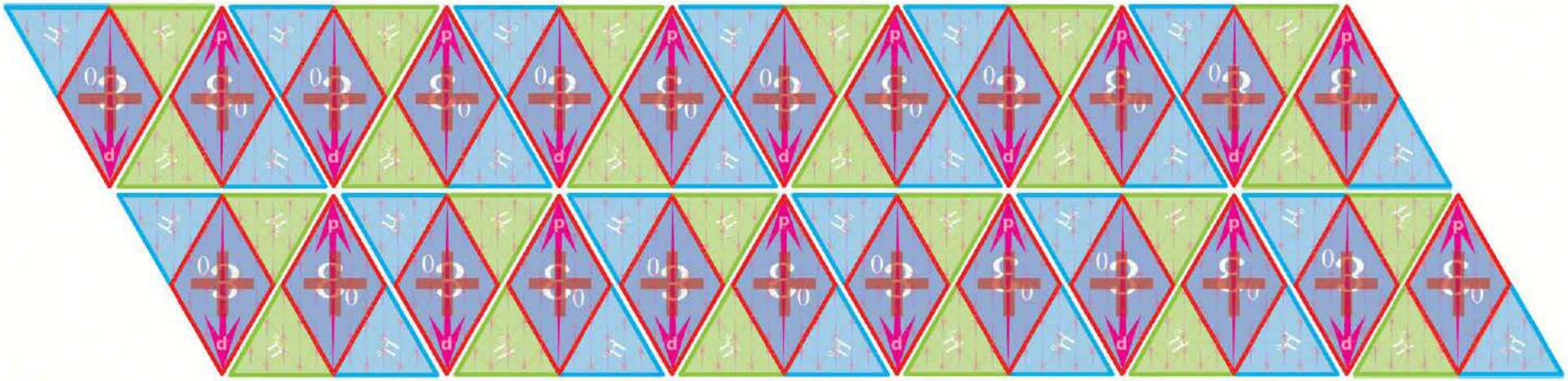

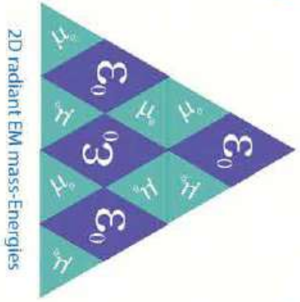
2D radiant EM mass-Energies
ε0
μ0

n6 - Charged Tetryon Templates

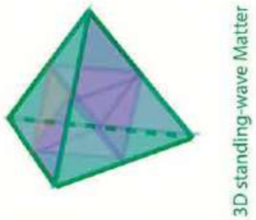
3D standing-wave Matter

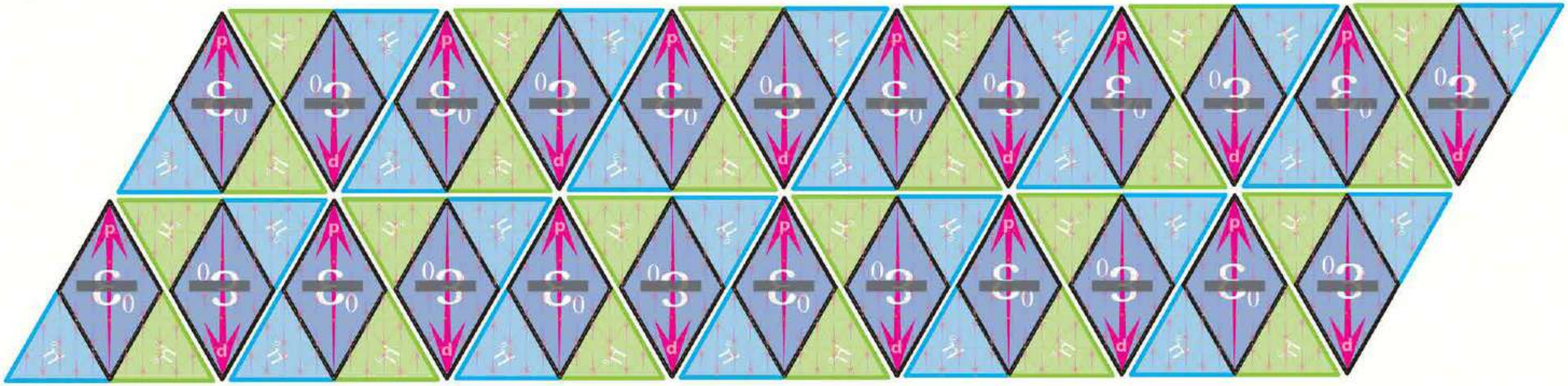

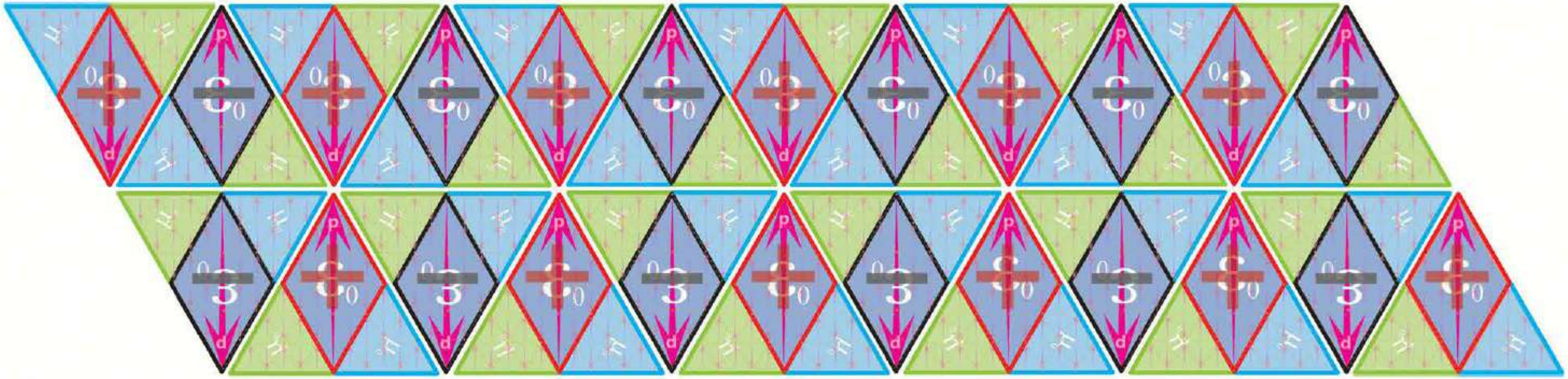

2D radiant EM mass-Energies

n6 - Neutral Tetryon Templates

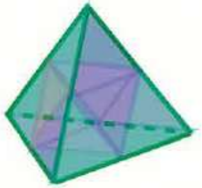

3D standing-wave Matter

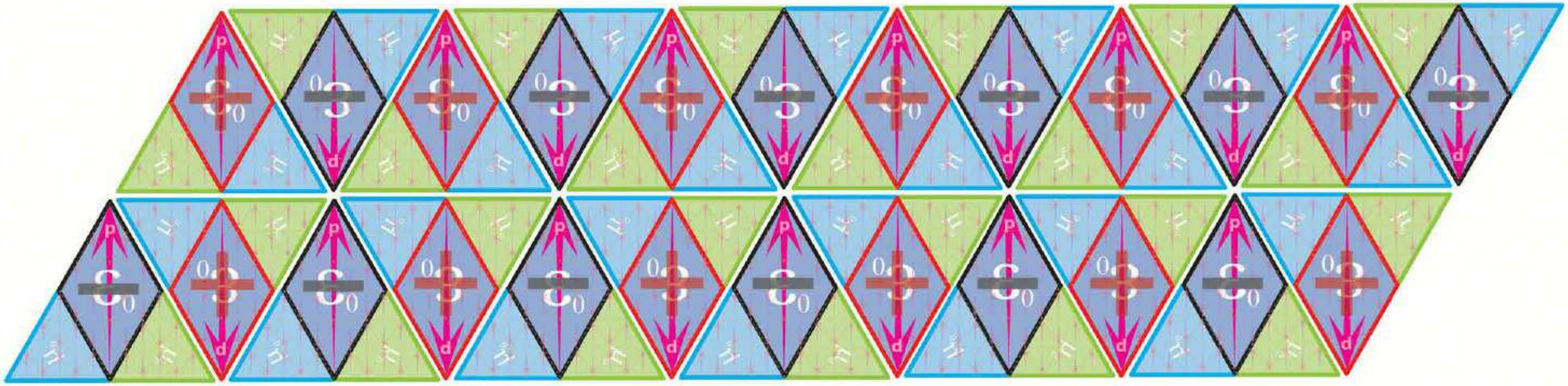

Copyright ABRAHAM [2008] - All rights reserved

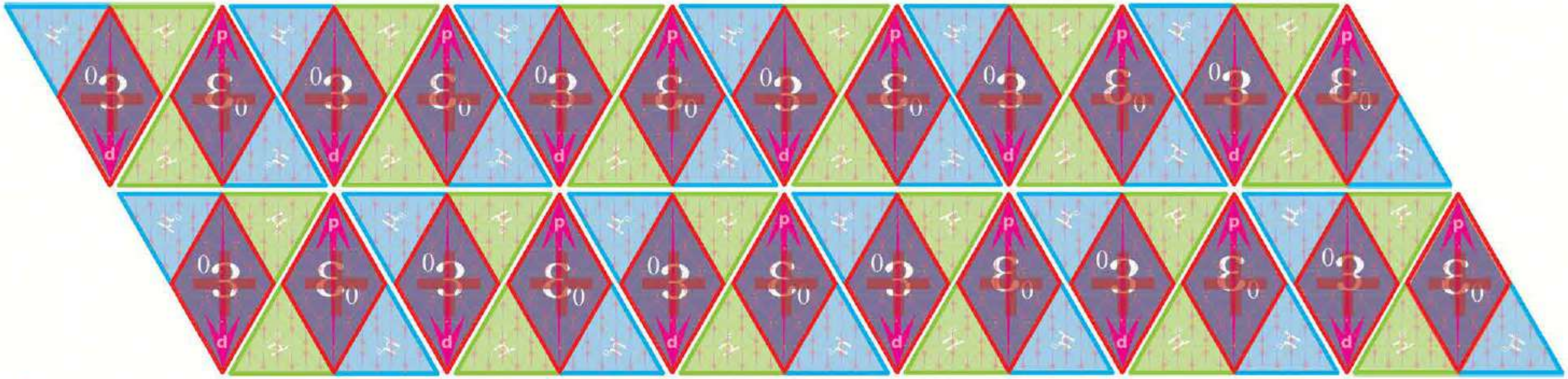
p
d
ε_0
μ_0

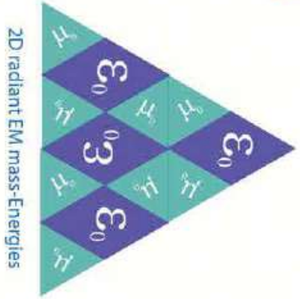
2D radiant EM mass-Energies
ε_0
μ_0

n7 - Charged Tetryon Templates

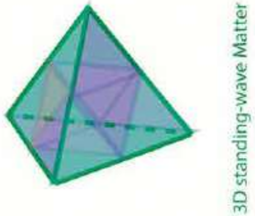
3D standing-wave Matter

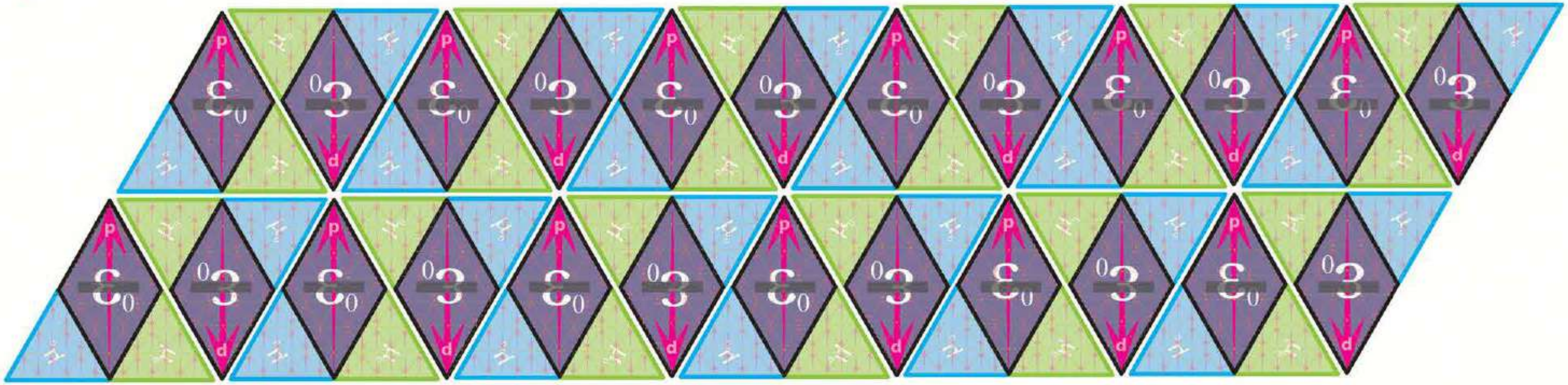
p
d
ε_0
μ_0

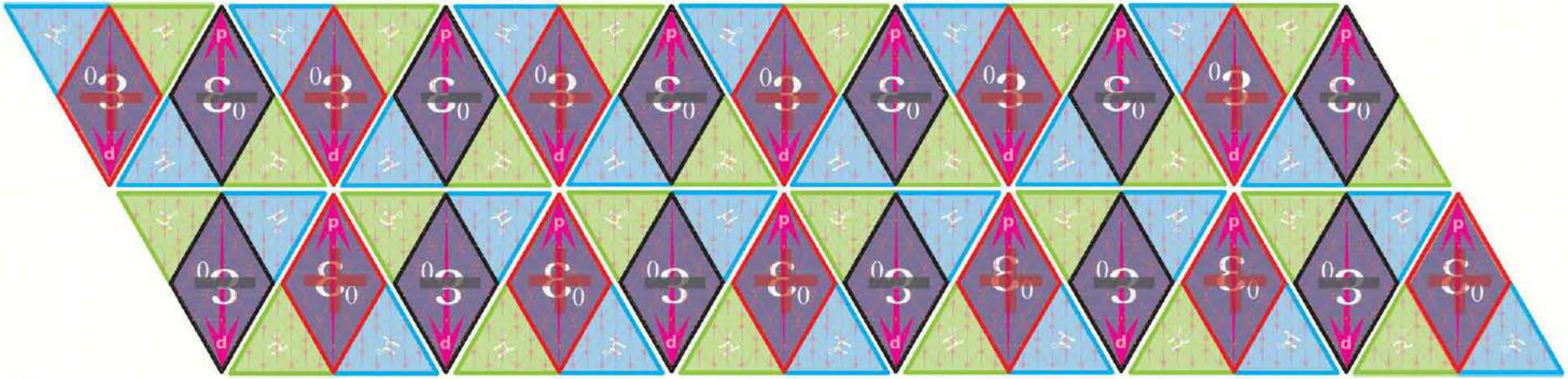

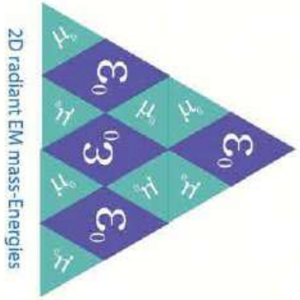
2D radiant EM mass-Energies
ε0
μ0

n7 - Neutral Tetryon Templates

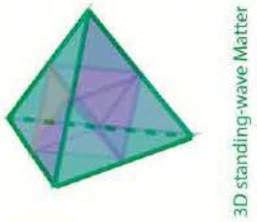
3D standing-wave Matter

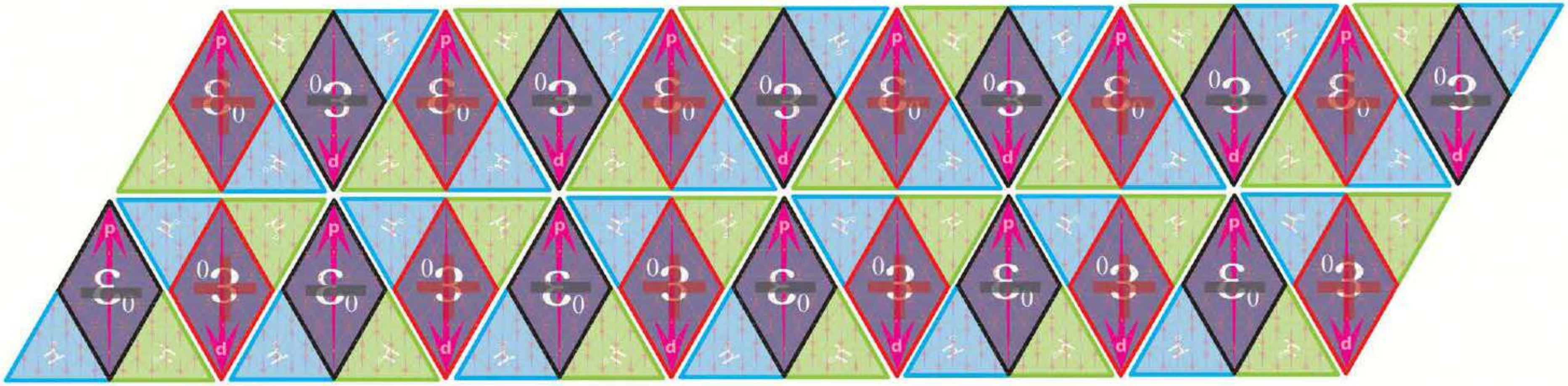

Copyright ABRAHAM [2008] - All rights reserved

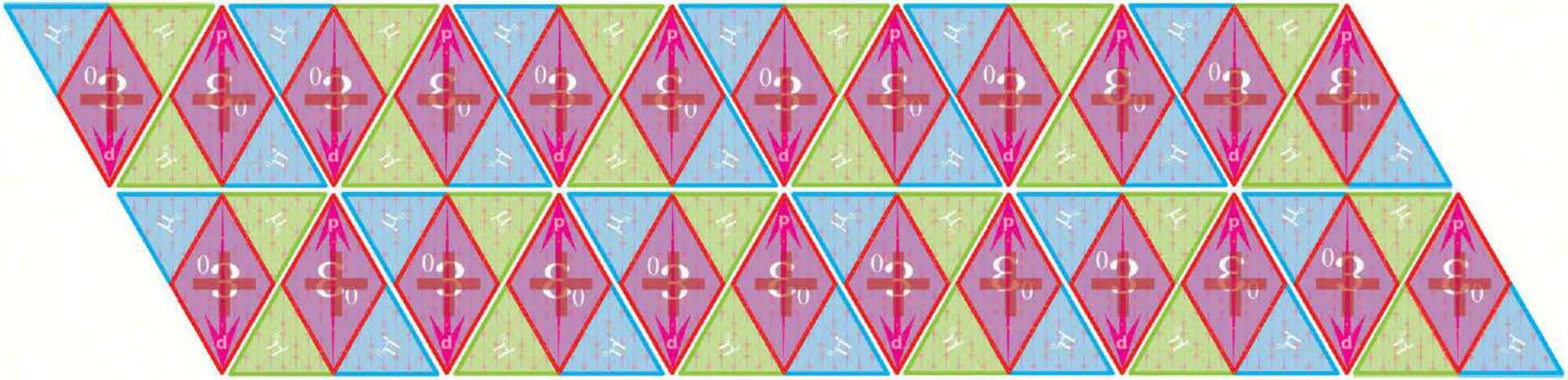
p
d
μ_0
ε_0

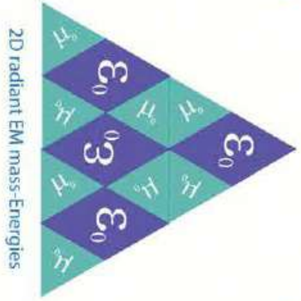
2D radiant EM mass-Energies
ε_0
μ_0

n8 - Charged Tetryon Templates

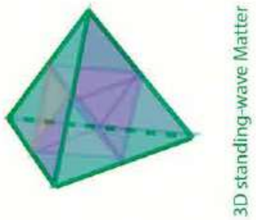
3D standing-wave Matter

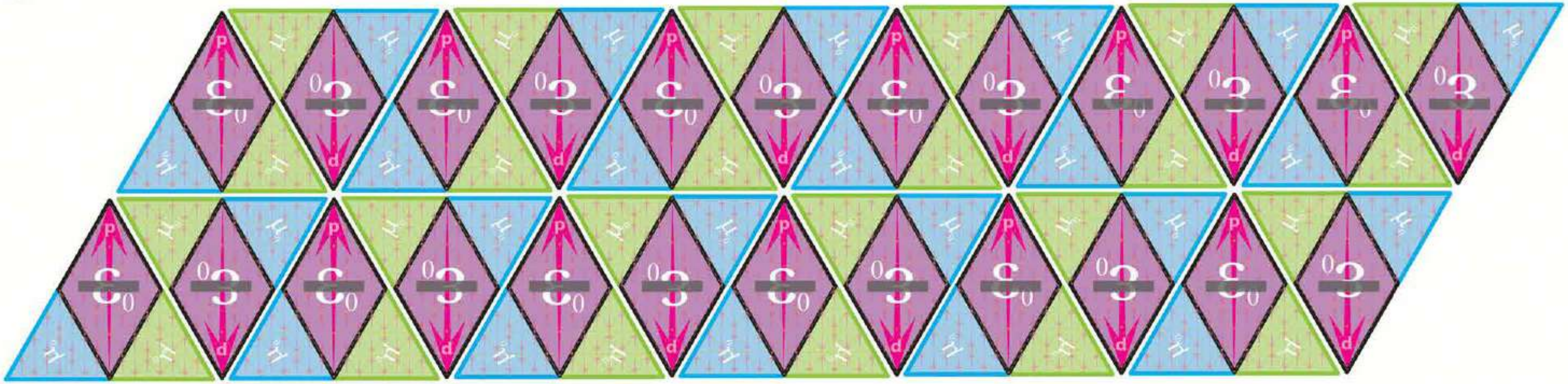
p
d
μ_0
ε_0

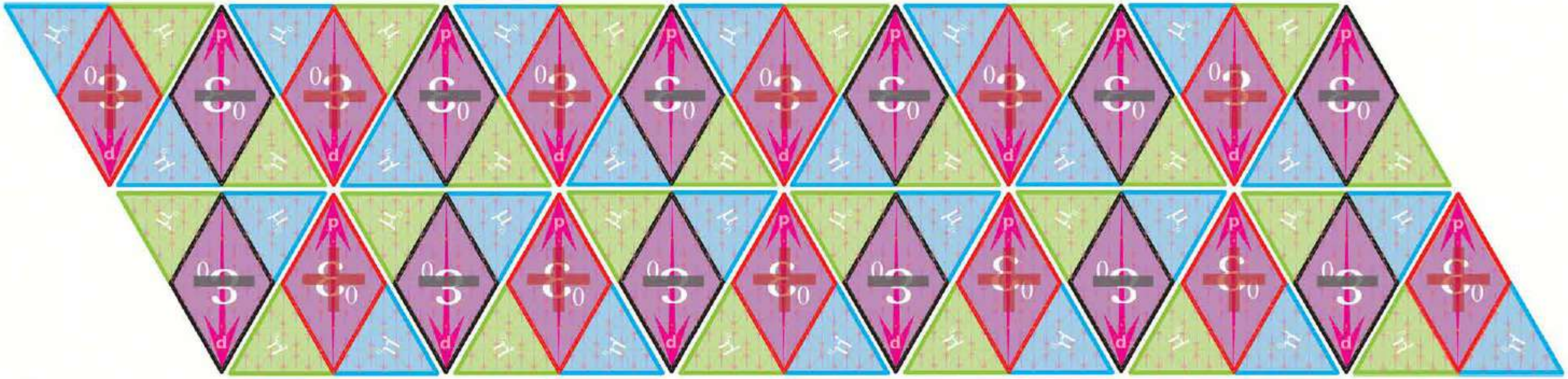
p
d
ε_0
μ_0

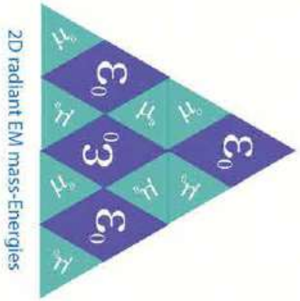
2D radiant EM mass-Energies
ε_0
μ_0

n8 - Neutral Tetryon Templates

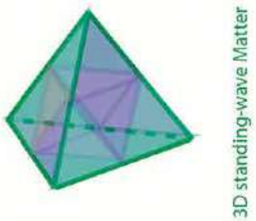
3D standing-wave Matter

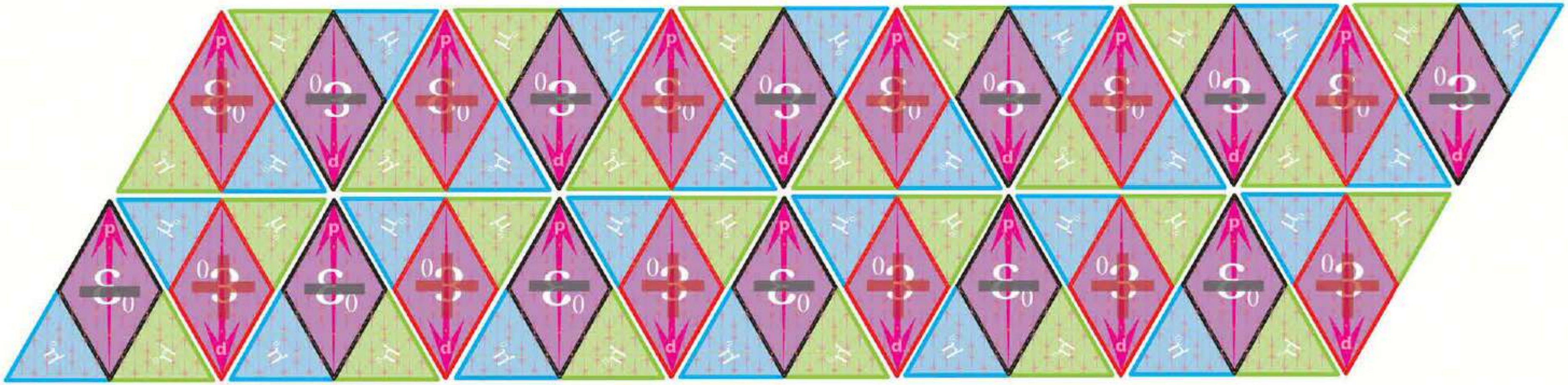
p
d
ε_0
μ_0

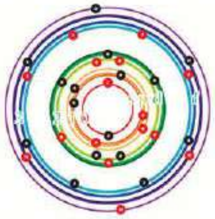

Periodic element geometries

An electron shell may be thought of as an orbit followed by electrons around an atom's nucleus. The closest shell to the nucleus is called the "1 shell" (also called "K shell"), followed by the "2 shell" (or "L shell"), then the "3 shell" (or "M shell"), and so on further and further from the nucleus. The shell letters K, L, M, ... are alphabetical

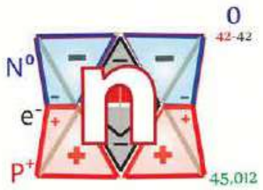
0
42-42
N^0
e^-
n
P^+
45,012

Each shell consists of one or more subshells, and each subshell consists of one or more atomic orbitals.

Each shell can contain only an integer number of whole deuterium nuclei [Proton, Neutron & electron]

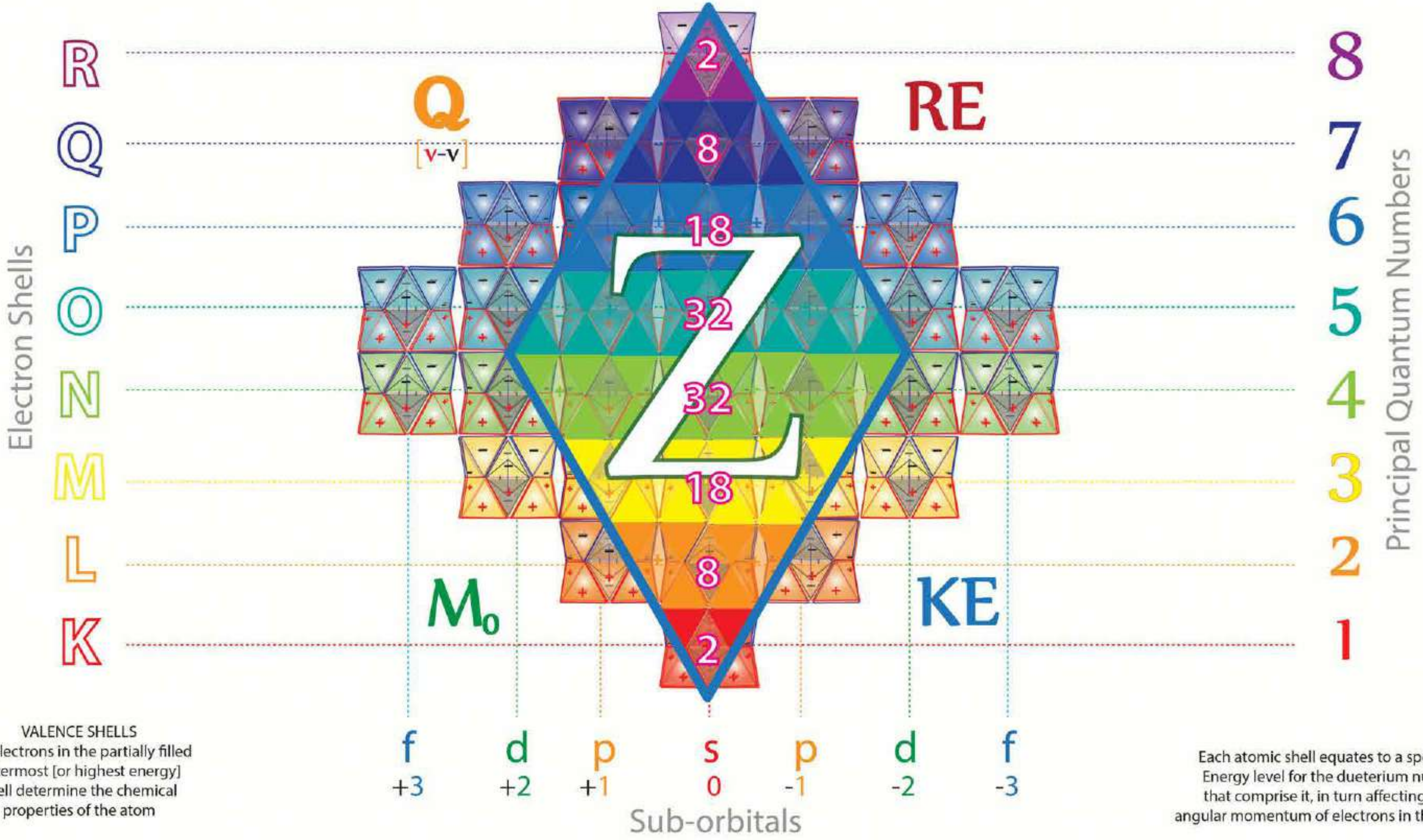
Electron Shells
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
Z
2
8
18
32
32
18
8
2
M_0
KE
8
7
6
5
4
3
2
1
Principal Quantum Numbers
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
Sub-orbitals

VALENCE SHELLS
The electrons in the partially filled outermost [or highest energy] shell determine the chemical properties of the atom

Each atomic shell equates to a specific Energy level for the dueterium nuclei that comprise it, in turn affecting the angular momentum of electrons in that shell

Periodic Harmonic motions

$x = A\cos(\omega t + \varphi)$

Circular motion

Much of the math in of modern physics
is predicated on the assumption that
π [where it appears] is related to
the properties of a circle

$F = -kx$

Linear motion

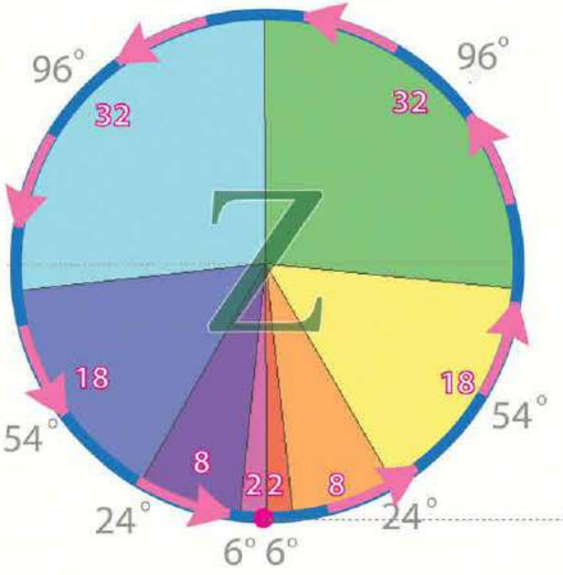
96°
96°
32
32
Z
18
18
54°
54°
8
2 2
8
24°
24°
6° 6°

$$p\sum 2(n^2)$$

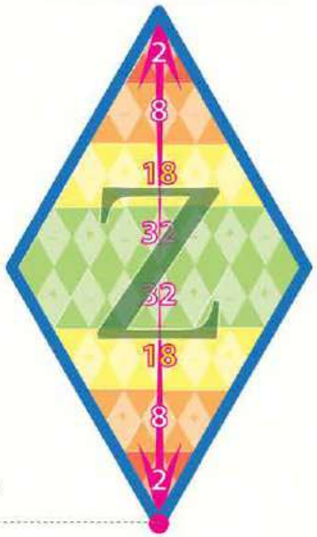
2
8
18
32
Z
32
18
8
2

Simple harmonic motion can be visualized as the projection of uniform circular motion onto one axis

circular harmonic motion

simple harmonic motion

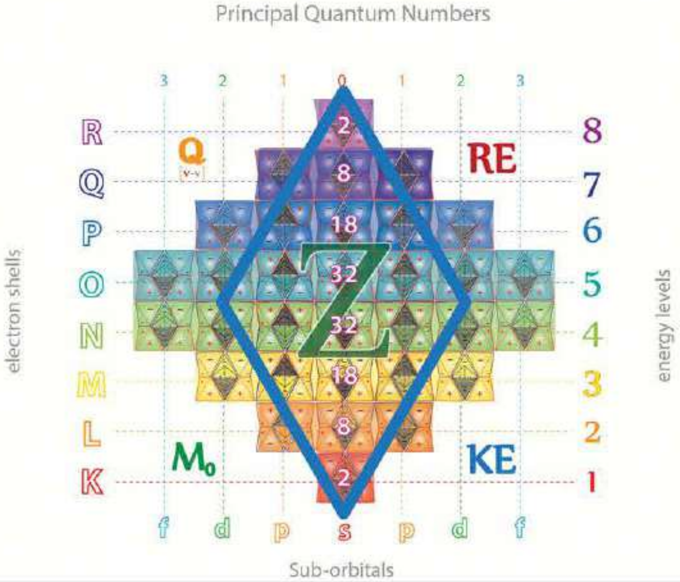
Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
2
8
18
32
Z
32
18
8
2
M₀
KE
8
7
6
5
4
3
2
1
electron shells
energy levels
f
d
p
s
p
d
f
Sub-orbitals

Circular motions describe the motion of a body with a changing velocity vector [the result of an acceleration force].

Nuclei per shell in elements follows a 'periodic summation rule' that is reflective of photonic energies

Baryon mass-energies

E

The quantum level mass-energies of Baryons determines the kinetic energies of electrons

1n - 1 e19v

electron KEM field

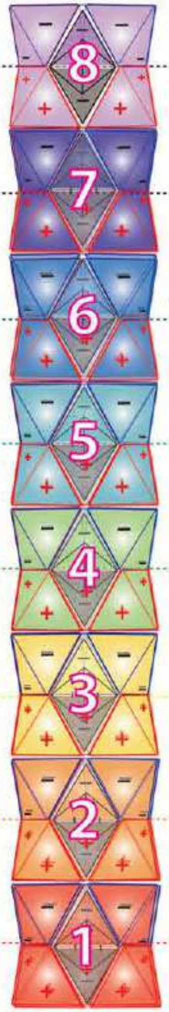

n			hv^2
73,728	n32	v8	3.268969693 e15
69,192	n31	v7	2.502804921 e15
64,800	n30	v6	1.838795452 e15
60,552	n29	v5	1.276941286 e15
56,448	n28	v4	8.172424232 e14
52,488	n27	v3	4.596988631 e14
48,672	n26	v2	2.043106058 e14
45,000	n25	v1	5.107765145 e13

Compton frequencies

930 MeV

1.2 e20
electron rest Matter is velocity invariant

496 keV

spectral frequencies

13.6 eV

Atomic nuclei mass-energies

Each element's weight [mass-Matter in a gravitational field] is the result of the total quanta comprising that element

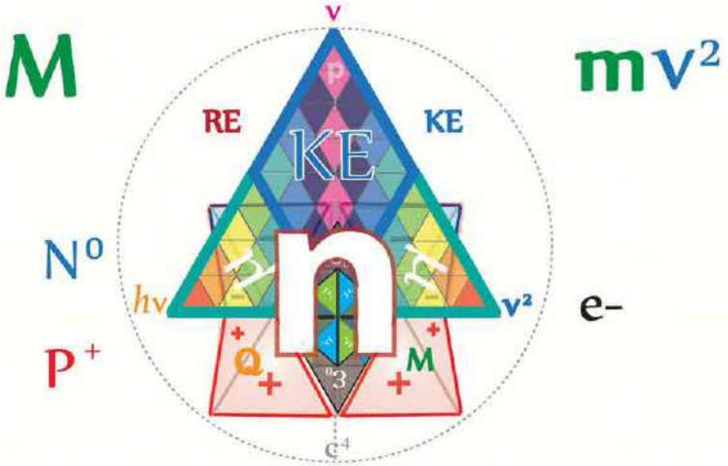

The nuclei forming each atomic shell have specific mass-energy quanta

Baryon rest masses lepton rest mass KEM

$$\overset{8}{\underset{1}{n}}\Big[\,[72(n)^2]+[12e19]+[m_e v^2]\,\Big]$$

Deuterium mass-energy per shell

Despite having differing mass-energies each Deuterium nuclei has the same velocity invariant Matter geometry [84π]

spin orbital coupling in synchronous quantum convertors

Electrons act as quantum scale rotating armatures in atomic nuclei and can only have specific energies reflective of the electron orbital energy level of the Baryons in which they are found

They acheive these energy levels by absorbing or emitting photons to acheive the specific angular momentum required

Baryons **KEM fields** **electrons**

930.947 MeV + **13.525 ev** + **496.519 keV**

Mapping Planck mass-energy contributions to elementary Matter and isotopes

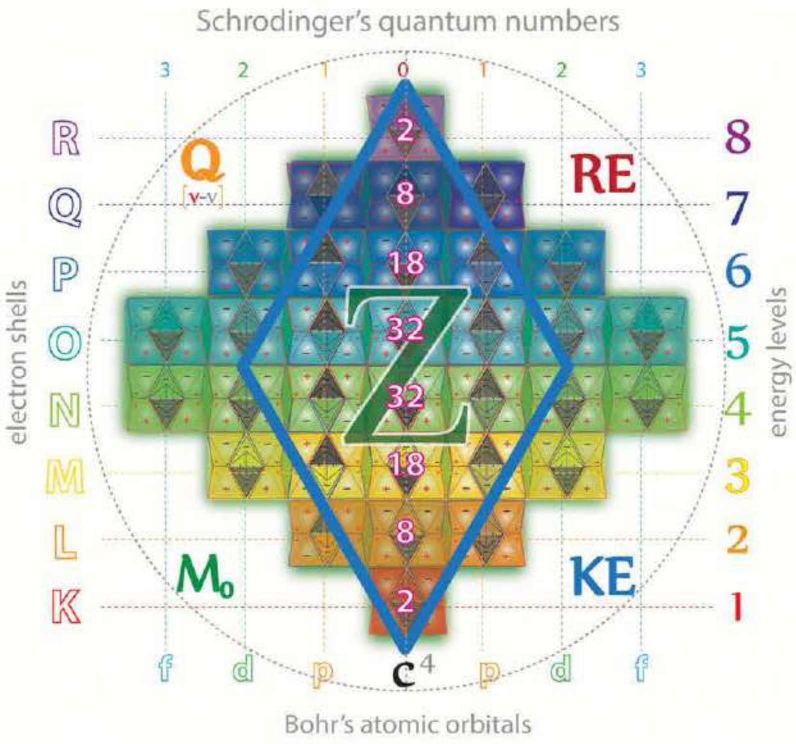
Schrodinger's quantum numbers
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
2
8
18
32
Z
32
18
8
2
M_0
KE
8
7
6
5
4
3
2
1
energy levels
f d p c^4 p d f
Bohr's atomic orbitals

$$E = hv^2$$

$$n^2 + v^2 + e = Z$$

general form quatratic equation

$$ax^2 + bx + c = Z$$

$$E = nhv$$

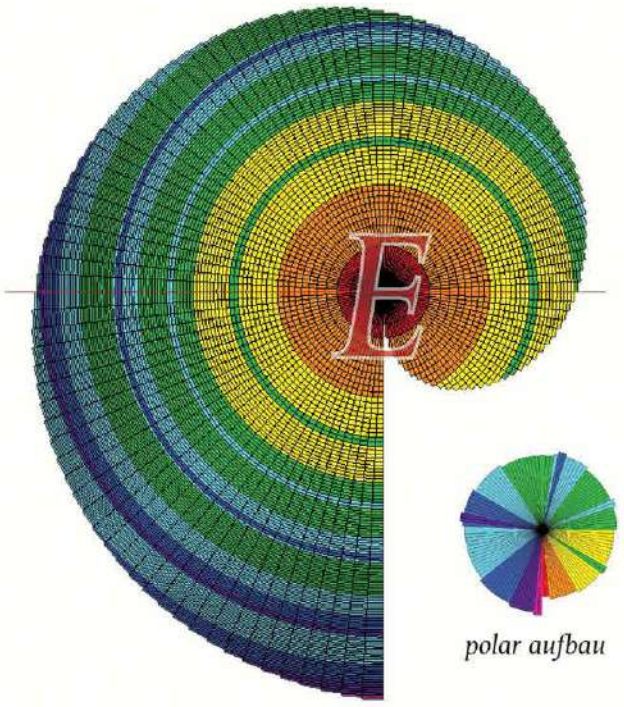
E
polar aufbau

polar energy spirals courtesy of Rene Cormier

Identifying electron rest Matter topologies as velocity invariant we can re-arrange the component Planck mass-energy geometry formulation of periodic elements to

$$h\left[\underbrace{72[v^2]}_{\text{Deuteron rest mass}} + \underbrace{v}_{\text{Spectral lines}} + \underbrace{1.20\,e20\ v}_{\text{electron rest mass}}\right]$$

reveal a quadratic formulation for all Z numbers

Copyright ABRAHAM [2008] - All rights reserved

$$P\sum_{K}^{R} Z$$

Periodic Summation

Periodic summation is a notation developed for Tetryonic theory
to model the geometric series addtion of Z[n²] energy level
Deuterium nuclei that form the periodic elements

$$\sum_{K}^{R}$$

nuclei per shell

STEP ONE

Periodic summation follows the atomic shell electron config

$$\sum_{1}^{120} Z$$

element number

STEP TWO

Periodic elements build up following the aufbau sequence

1 R
2
3
4
4
3
2
1 K

$$P\sum_{K}^{R} 2(x^2) =$$

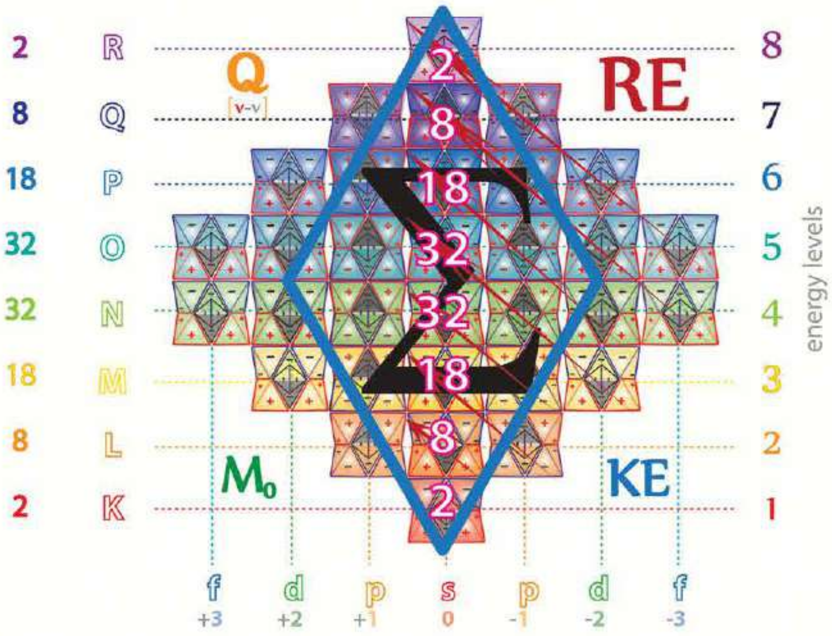

Each atomic shell can hold only a fixed number of deuterium nuclei

$P\sum R = 2$	2 nuclei [74.496 ea] +	120	Unbinilium
$P\sum Q = 8$	8 nuclei [69,780 ea] +	118	Ununoctium
$P\sum P = 18$	18 nuclei [65,232 ea] +	110	Darmstadtium
$P\sum O = 32$	32 nuclei [60,852 ea] +	92	Uranium
$P\sum N = 32$	32 nuclei [56,640 ea] +	60	Neodymuim
$P\sum M = 18$	18 nuclei [52,596 ea] +	28	Argon
$P\sum L = 8$	8 nuclei [48,720 ea] +	10	Neon
$P\sum K = 2$	2 nuclei [45,012 ea]	2	Helium
		0	Hydrogen

THe LHS of the notation determine the number of nuclei in each atomic shell, from the periodic mass-energy levels for atoms, and the RHS follows the aufbau building principle to determine the rest mass-Matter of any specific element

Aufbau

Each periodic element is made of Z [n² energy] deuterium nuclei

$$Z\#\left[\begin{array}{ll} z\ \text{Protons} & [24\text{-}12] \\ z\ \text{Neutrons} & [18\text{-}18] \\ z\ \text{electrons} & [0\text{-}12] \end{array}\right] n1\text{-}8$$

Planck mass-energies form the surface integral of rest Matter topologies for each periodic element

Copyright ABRAHAM [2008] - All rights reserved

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number.
It is used as rule to determine which nuclei are stable or unstable.

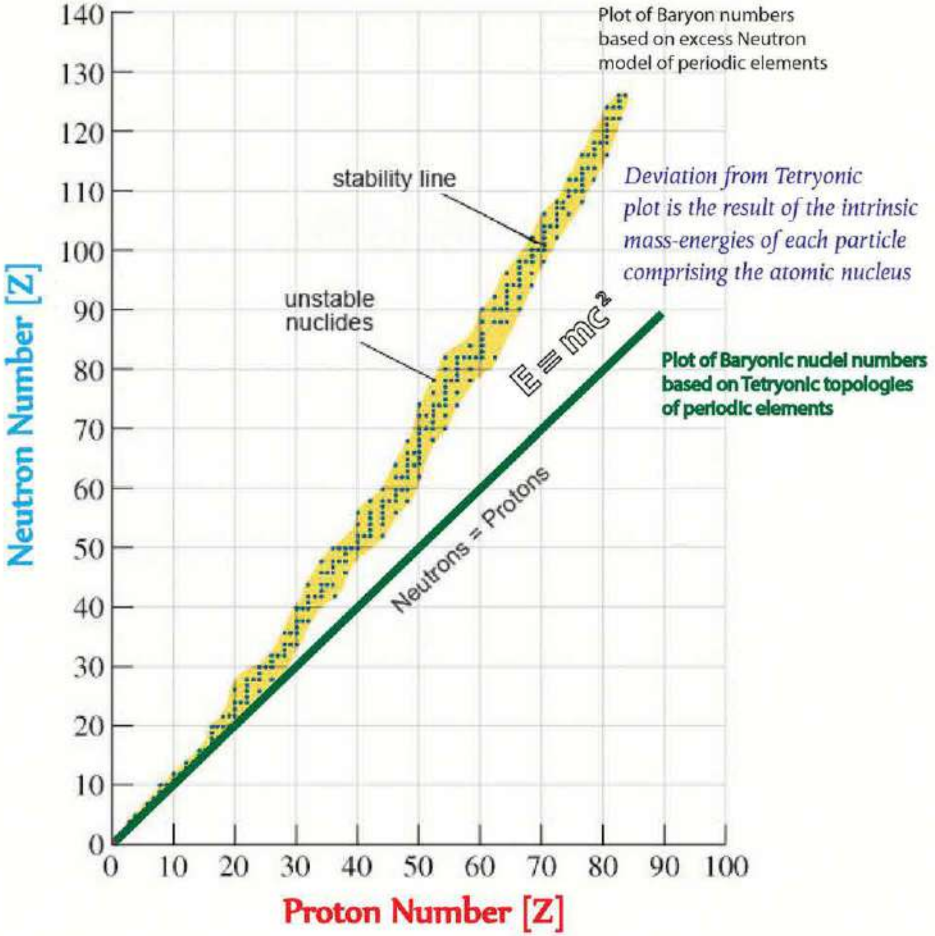

Plot of Baryon numbers based on excess Neutron model of periodic elements

Deviation from Tetryonic plot is the result of the intrinsic mass-energies of each particle comprising the atomic nucleus

Plot of Baryonic nuclei numbers based on Tetryonic topologies of periodic elements

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

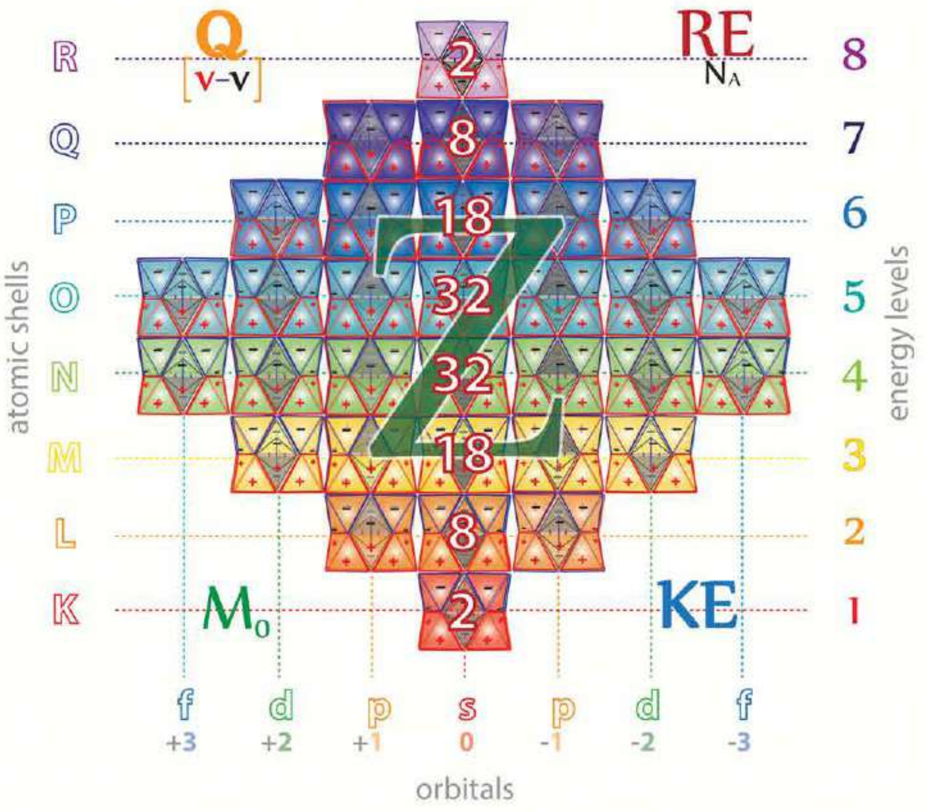

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

Copyright ABRAHAM [2008] - All rights reserved

Planck mass-energy contributions to elementary Matter and isotopes

The mass-energy content of Deuterium nuclei creates the molar mass of elements
[not extra neutrons in excess of the elemental number]

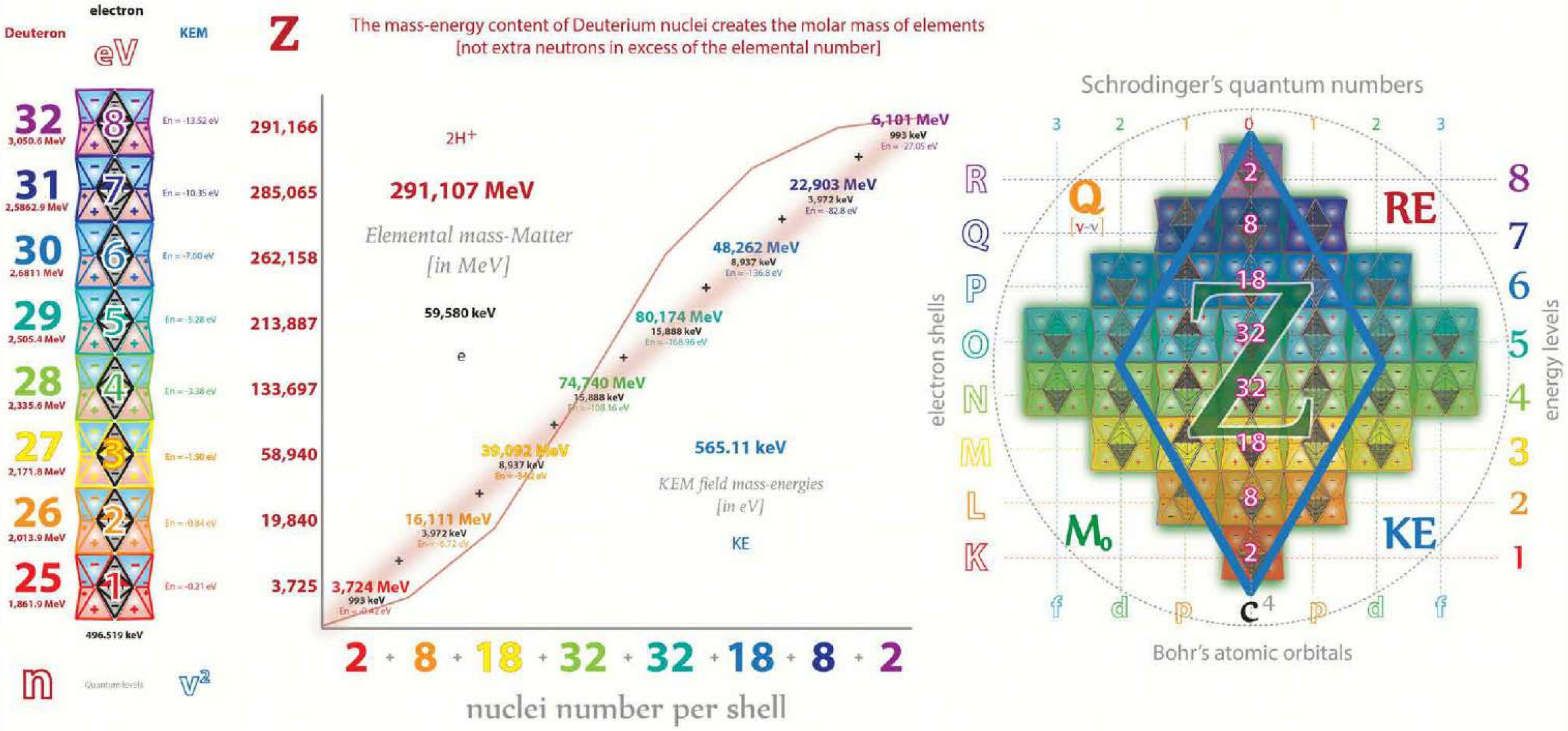

Deuteron
electron
eV
KEM
Z
32
3,050.6 MeV
31
2,5862.9 MeV
30
2,6811 MeV
29
2,505.4 MeV
28
2,335.6 MeV
27
2,171.8 MeV
26
2,013.9 MeV
25
1,861.9 MeV
8
7
6
5
4
3
2
1
En = -13.52 eV
En = -10.35 eV
En = -7.60 eV
En = -5.28 eV
En = -3.38 eV
En = -1.90 eV
En = -0.84 eV
En = -0.21 eV
291,166
285,065
262,158
213,887
133,697
58,940
19,840
3,725
496.519 keV
n
Quantum levels
V²
2H⁺
291,107 MeV
Elemental mass-Matter
[in MeV]
59,580 keV
e
565.11 keV
KEM field mass-energies
[in eV]
KE
3,724 MeV
993 keV
En = -0.42 eV
16,111 MeV
3,972 keV
En = -6.72 eV
39,092 MeV
8,937 keV
En = -34.2 eV
74,740 MeV
15,888 keV
En = -108.16 eV
80,174 MeV
15,888 keV
En = -168.96 eV
48,262 MeV
8,937 keV
En = -136.8 eV
22,903 MeV
3,972 keV
En = -82.8 eV
6,101 MeV
993 keV
En = -27.05 eV
2 + 8 + 18 + 32 + 32 + 18 + 8 + 2
nuclei number per shell
Schrodinger's quantum numbers
3
2
1
0
1
2
3
electron shells
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
Z
2
8
18
32
32
18
8
2
M₀
KE
c⁴
8
7
6
5
4
3
2
1
energy levels
f
d
p
p
d
f
Bohr's atomic orbitals

n1

Elementary nuclei

are comprised of equal numbers of
Protons, Neutrons & electrons
with varying energy levels

Baryons		electrons		KEM fields
930.947 MeV	+	**496.519 keV**	+	**13.525 ev**

The mass-energy content of Matter topologies is velocity invariant

The mass-energy content of Baryons determines the KEM field of electrons

Copyright ABRAHAM [2008] - All rights reserved

All elements are comprised of n level Duetrium nuclei

Baryons **KEM fields** **electrons**

$$Z\,[[72n^2] + [12v^2] + [1.2e20]]$$

1,861,949 MeV **13.525 eV** **496,519 keV**

$2H^+$ $[72n^2]$ **Z** γ $[M_ev^2]$

Elemental mass-Matter
[in MeV]

The atomic shell energy levels of Deuterium nuclei in elements

Deuteron **electron eV** **KEM**

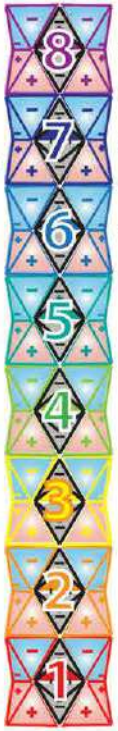

Deuteron		KEM
32 3,050.6 MeV		En = -13.52 eV
31 2,5862.9 MeV		En = -10.35 eV
30 2,6811 MeV		En = -7.60 eV
29 2,505.4 MeV		En = -5.28 eV
28 2,335.6 MeV		En = -3.38 eV
27 2,171.8 MeV		En = -1.90 eV
26 2,013.9 MeV		En = -0.84 eV
25 1,861.9 MeV		En = -0.21 eV

496.519 keV

m v^2

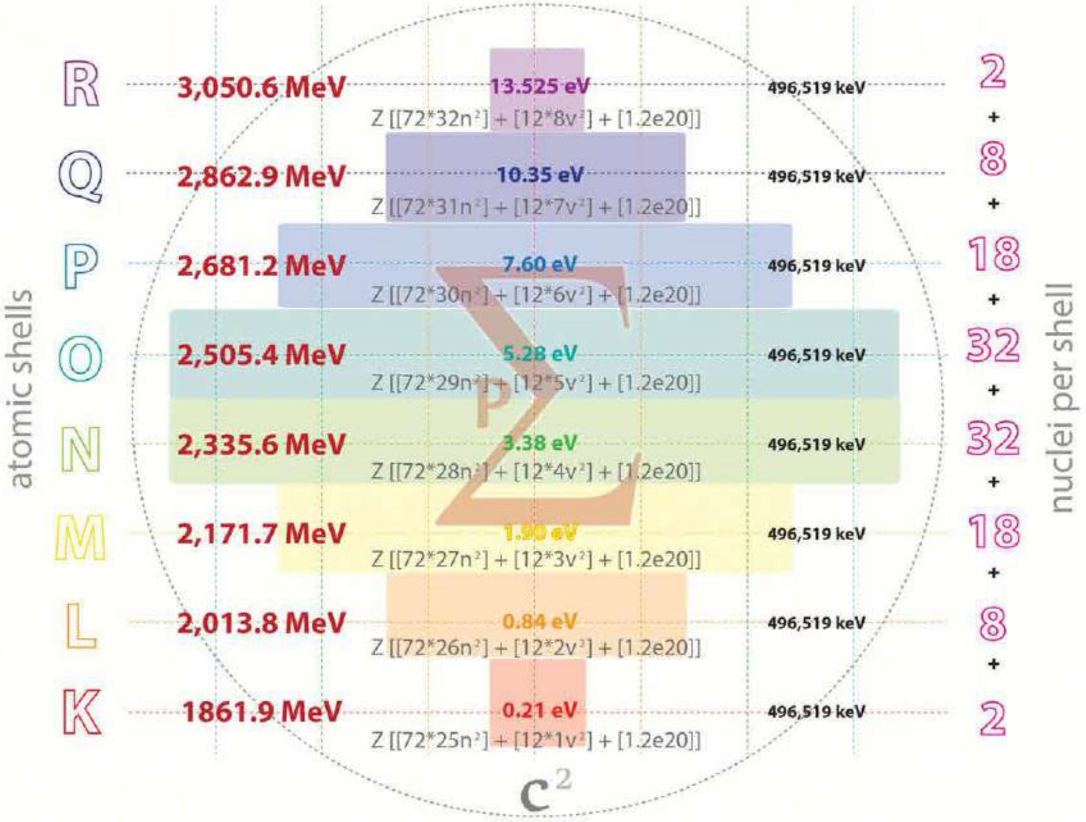

R	3,050.6 MeV	13.525 eV Z [[72*32n²] + [12*8v²] + [1.2e20]]	496,519 keV	2
				+
Q	2,862.9 MeV	10.35 eV Z [[72*31n²] + [12*7v²] + [1.2e20]]	496,519 keV	8
				+
P	2,681.2 MeV	7.60 eV Z [[72*30n²] + [12*6v²] + [1.2e20]]	496,519 keV	18
				+
O	2,505.4 MeV	5.28 eV Z [[72*29n²] + [12*5v²] + [1.2e20]]	496,519 keV	32
				+
N	2,335.6 MeV	3.38 eV Z [[72*28n²] + [12*4v²] + [1.2e20]]	496,519 keV	32
				+
M	2,171.7 MeV	1.90 eV Z [[72*27n²] + [12*3v²] + [1.2e20]]	496,519 keV	18
				+
L	2,013.8 MeV	0.84 eV Z [[72*26n²] + [12*2v²] + [1.2e20]]	496,519 keV	8
				+
K	1861.9 MeV	0.21 eV Z [[72*25n²] + [12*1v²] + [1.2e20]]	496,519 keV	2

6,101 MeV
993 keV
En = -27.05 eV
+
22,903 MeV
3,972 keV
En = -82.8 eV
+
48,262 MeV
8,937 keV
En = -316.8 eV
+
80,174 MeV
15,888 keV
En = -168.96 eV
+
74,740 MeV
15,888 keV
En = -108.16 eV
+
39,092 MeV
8,937 keV
En = -34.2 eV
+
16,111 MeV
3,972 keV
En = -6.72 eV
+
3,724 MeV
993 keV
En = -0.42 eV

[1.2e20]]

e

Determines the spectral line [KEM field energies] of electrons bound to them

The relativistic rest mass-energy-Matter of all periodic elements

is the sum of the mass-energies of all atomic nuclei and spectral lines that comprise its mass-Matter topology as measured in any spatial co-ordinate system per unit of time

the rest mass-Matter of bound photo-electrons is velocity invariant

Ionisation energies

Fractional quantum differentials

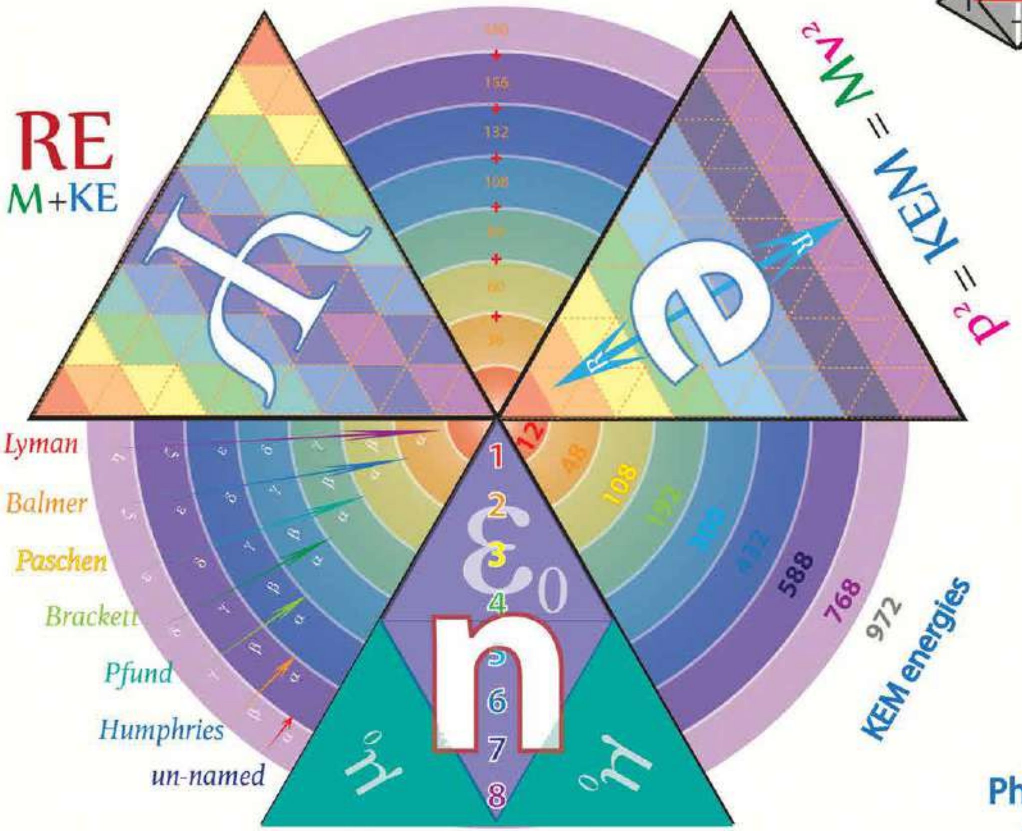
RE
M+KE
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
p^2 = KEM = Mv^2
KEM energies
1
2
3
4
5
6
7
8
12
48
108
192
300
432
588
768
972
ε_0
n

$$hf = \Delta Mv = \Delta p$$

spectral lines are produced by accelerating electrons

Note: this is an Illustrative schema for modelling KEM field energies
All KEM fields possess the same physical spatial geometry
in radial-time defined spatial co-ordinate systems

c^2 c^4

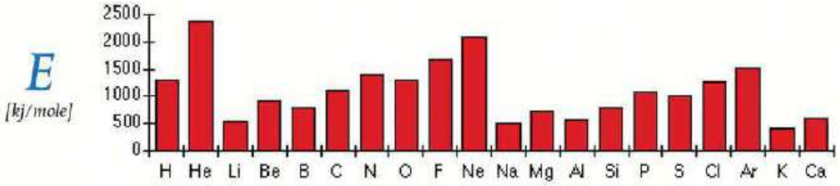

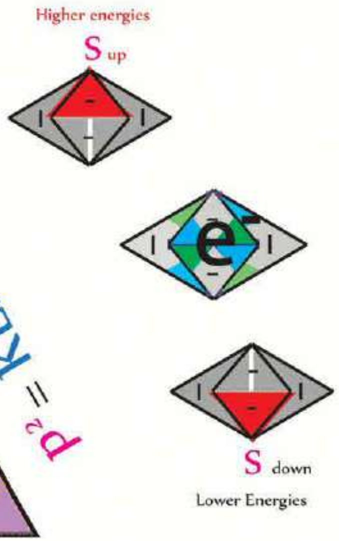
Higher energies
S up
e
S down
Lower Energies

Photo-electrons absorb/emit spectral energies

$$E = eV = \frac{1}{4\pi\epsilon_0}\frac{ne^2}{a}$$

photo-electronic energy momenta

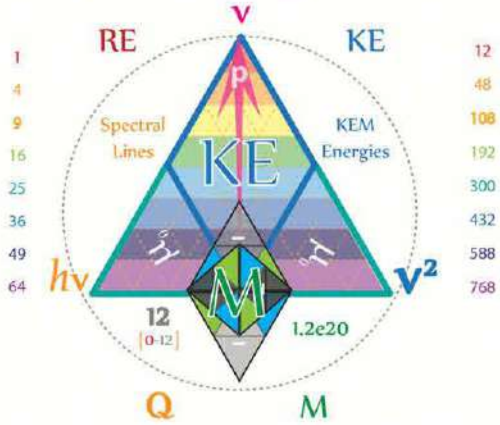
v
RE
KE
Spectral Lines
KEM Energies
KE
p
M
hv
v^2
12
[0-12]
1.2e20
Q
M
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768

Mapping photo-electron transition energies
to Tetryonic energy momenta geometries
reveals many key facts about the
ionisation energies of nuclei

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

The differing fractional KEM field energy momenta of electrons
that results from their transitions to specific energy nuclei
in elements results in differing QAM quanta and
produces spectral lines and fine line splitting

Spectral line transitions

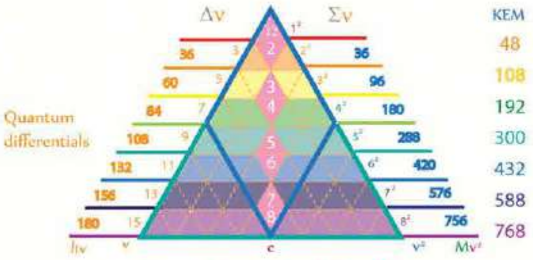
Δv
Σv
KEM
Quantum differentials
36
60
84
108
132
156
180
3
5
7
9
11
13
15
1
2
3
4
5
6
7
8
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
36
96
180
288
420
576
756
48
108
192
300
432
588
768
hv
v
c
v^2
Mv^2

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Element numbers

Nuclei per shell in elements follow
a 'periodic summation rule'
that is reflective of
photonic energies

1 R
2
3
4
4
3
2
1 K

$$ P\sum_{K}^{R} 2(x^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2 $$

Z

120 Unbinilium 119 Ununennium	n8
118 Ununoctium 87 Francium	n7
112 Copernicium 55 Caesium	n6
102 Nobelium 37 Rubidium	n5
70 Ytterbuim 19 Potassium	n4
30 Zinc 11 Sodium	n3
10 Neon 3 Lithium	n2
2 Helium 1 Deuterium	n1
Hydrogen 0	

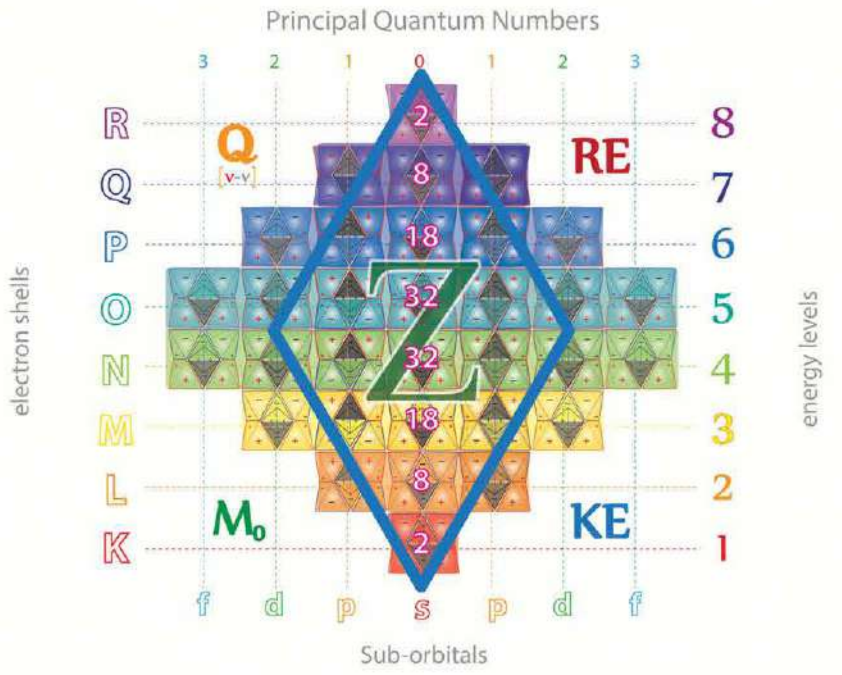
Principal Quantum Numbers
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M₀
KE
Z
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f d p s p d f
Sub-orbitals

Periodic mass-ENERGY-Matter

Following periodic summation rules for shell filling
n[1-8] quantum energy deuterium nuclei
combine to form elementary Matter

Baryon rest masses lepton rest mass KEM

$$ P\sum_{K\,25}^{R\,32}\left[[72(n)^2]+[12e19]+[m_e v^2]\right]_1^8 $$

Deuterium mass-energy per shell

The measured weight of Matter in gravitational fields is the result of planar mass-energies in tetryonic standing-wave geometries

The periodicity of all the elements,
along with their exact molar rest mass-energies and
quantum wavefunctions can be described with Tetryonic geometries

Ionisation energies

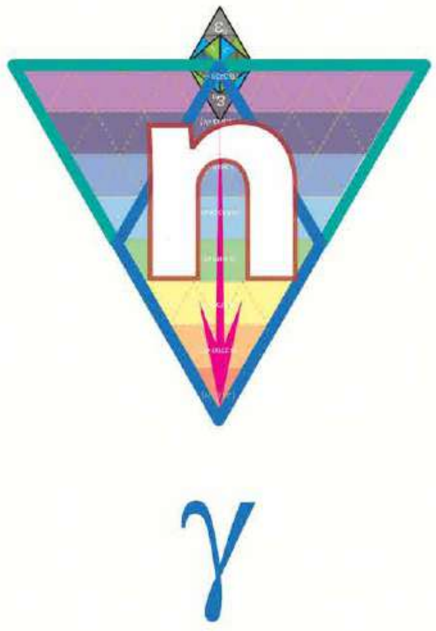

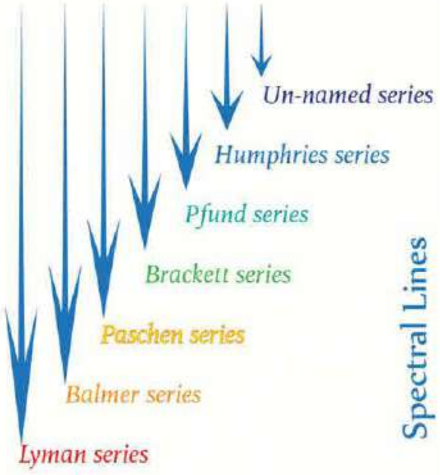
Un-named series
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
Spectral Lines

$$ Mv^2 = KEM = hcR_H $$

Photon emission/absorption

Copyright ABRAHAM [2008] - All rights reserved

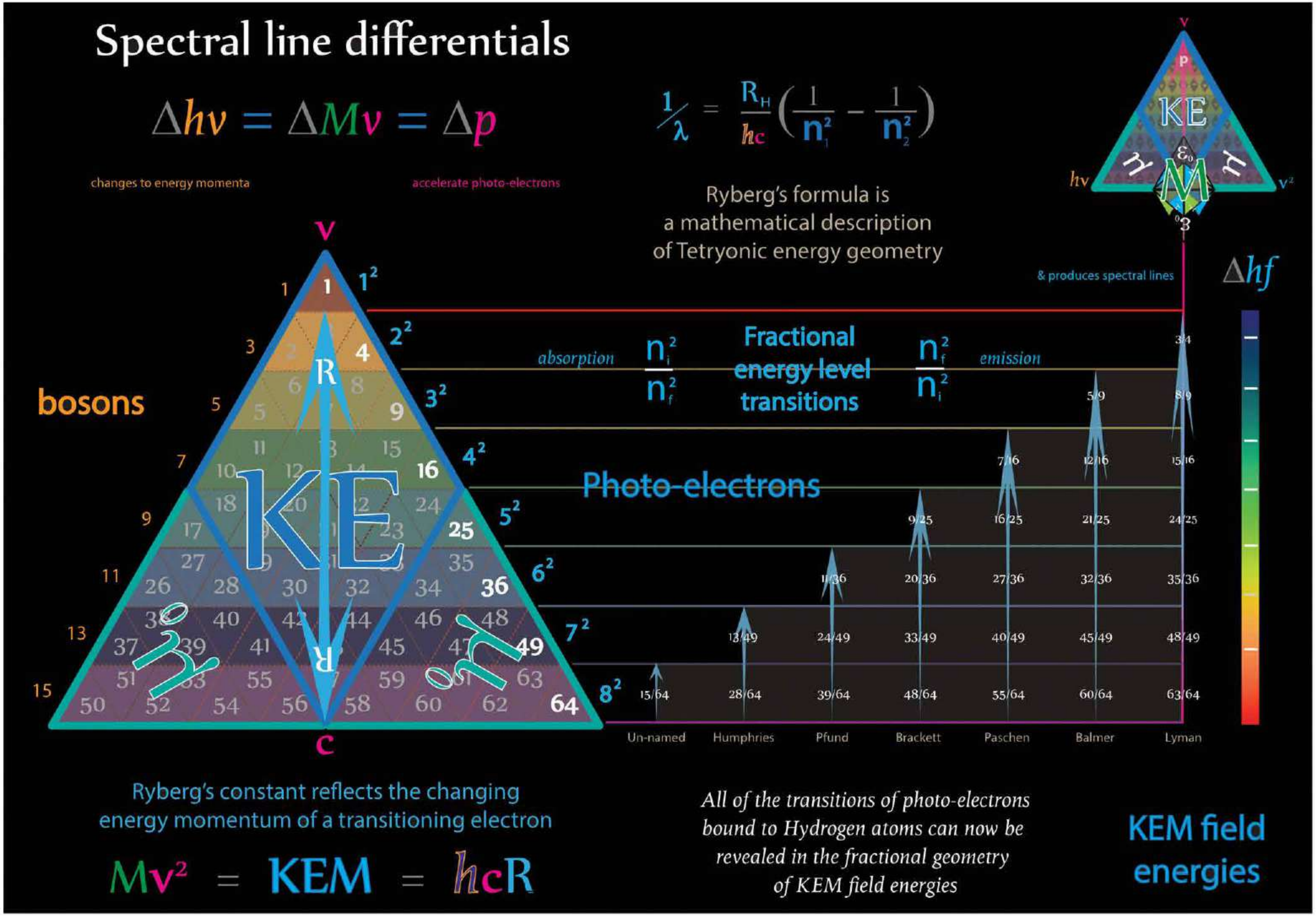

Spectral line differentials

$$\Delta hv = \Delta Mv = \Delta p$$

changes to energy momenta

accelerate photo-electrons

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg's formula is
a mathematical description
of Tetryonic energy geometry

& produces spectral lines

Fractional energy level transitions

absorption $\frac{n_i^2}{n_f^2}$

emission $\frac{n_f^2}{n_i^2}$

Photo-electrons

Ryberg's constant reflects the changing
energy momentum of a transitioning electron

$$Mv^2 = KEM = hcR$$

*All of the transitions of photo-electrons
bound to Hydrogen atoms can now be
revealed in the fractional geometry
of KEM field energies*

KEM field energies

bosons
v
c
KE
R
1
3
5
7
9
11
13
15
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
1 2 4 6 8 5 9 11 15 16 10 12 14 18 20 22 24 17 23 25 27 35 26 28 30 32 34 36 38 40 42 44 46 48 37 39 41 45 47 49 51 53 55 59 63 50 52 54 56 58 60 62 64
KE
ε_0
M
hv
v^2
3
Δhf
3/4
5/9 8/9
7/16 12/16 15/16
9/25 16/25 21/25 24/25
11/36 20/36 27/36 32/36 35/36
13/49 24/49 33/49 40/49 45/49 48/49
15/64 28/64 39/64 48/64 55/64 60/64 63/64
Un-named
Humphries
Pfund
Brackett
Paschen
Balmer
Lyman

Fermat's method of Factoring

also known as 'the difference of two squares' is used to factorise large numbers

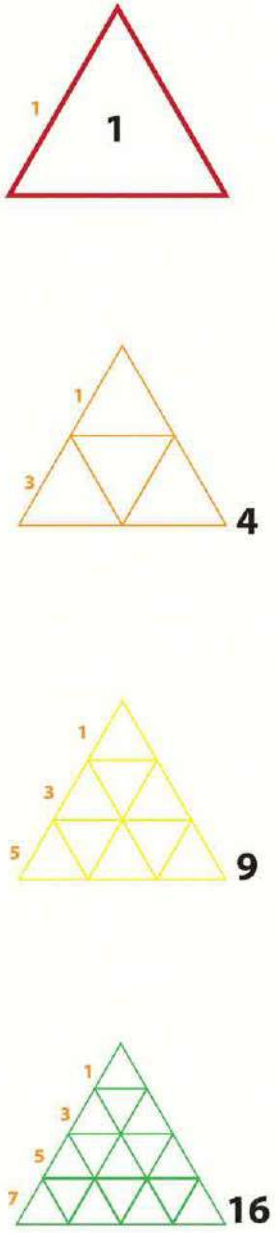
1
1
1
3
4
1
3
5
9
1
3
5
7
16

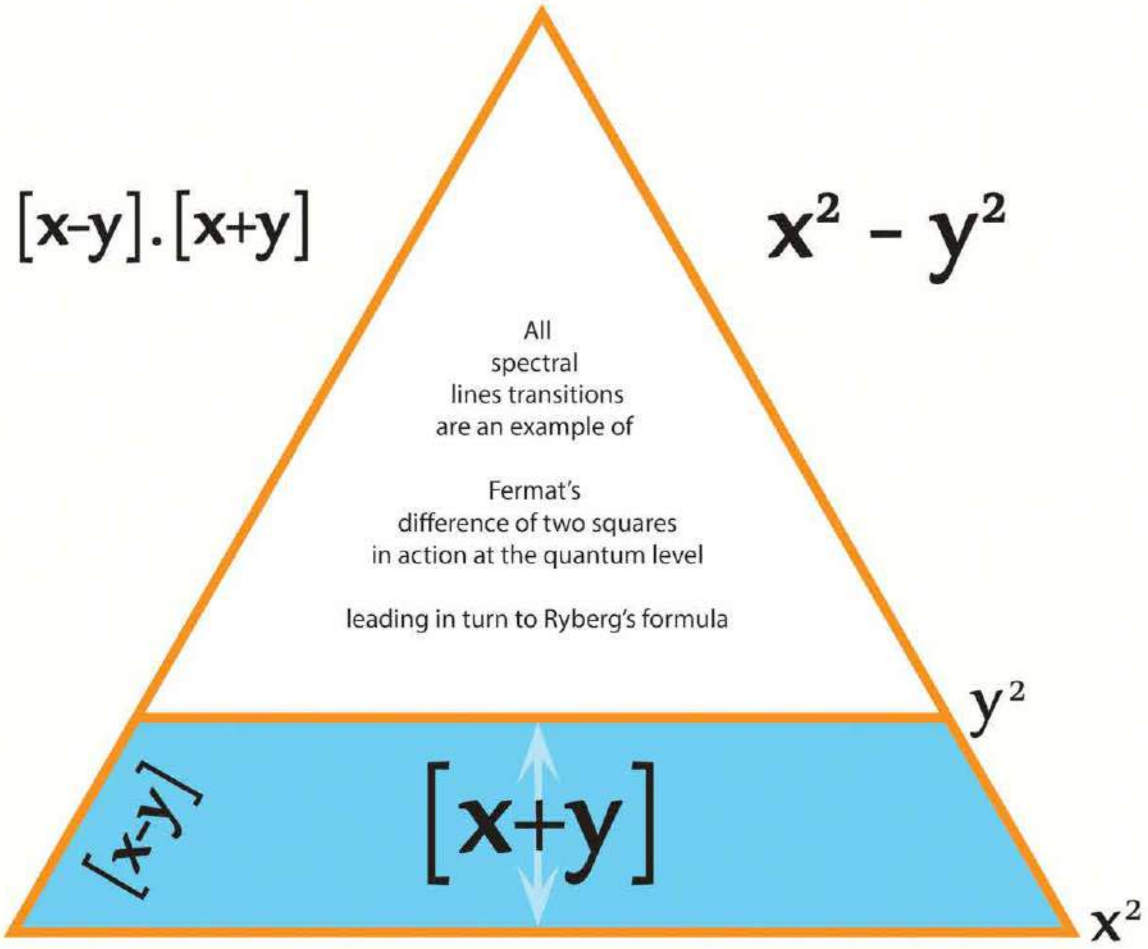

$[x-y].[x+y]$

$x^2 - y^2$

All
spectral
lines transitions
are an example of

Fermat's
difference of two squares
in action at the quantum level

leading in turn to Ryberg's formula

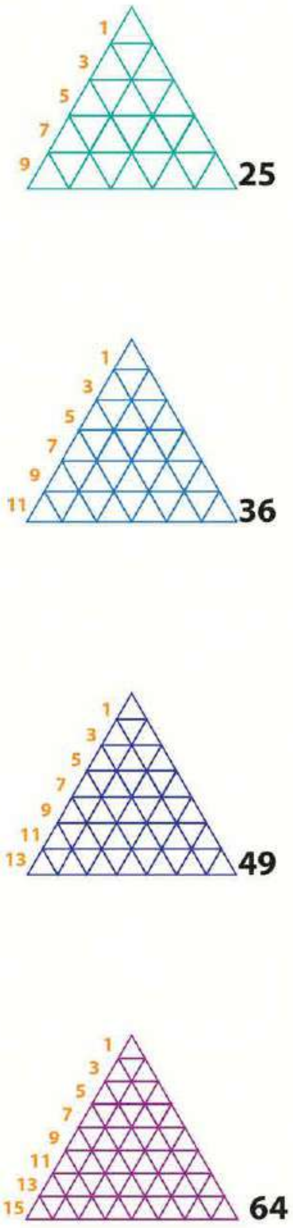
1
3
5
7
9
25
1
3
5
7
9
11
36
1
3
5
7
9
11
13
49
1
3
5
7
9
11
13
15
64

Fermat knew that every odd number could be written as the difference of two squares
or as revealed geometrically through Tetryonic theory's equilateral geometry,
every 'SQUARE' number is the sequential sum of ODD numbers

Copyright ABRAHAM [2008] - All rights reserved

Fractions

A fraction is a number that shows how many equal parts there are

3/4

24/25

In quantum mechanics fractions appear in quantum steps
as a result of the equilateral geometry of Planck energy momenta

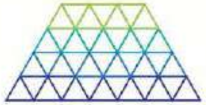

40/49

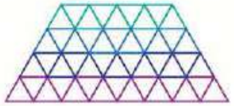

48/64

80/81

.987654321

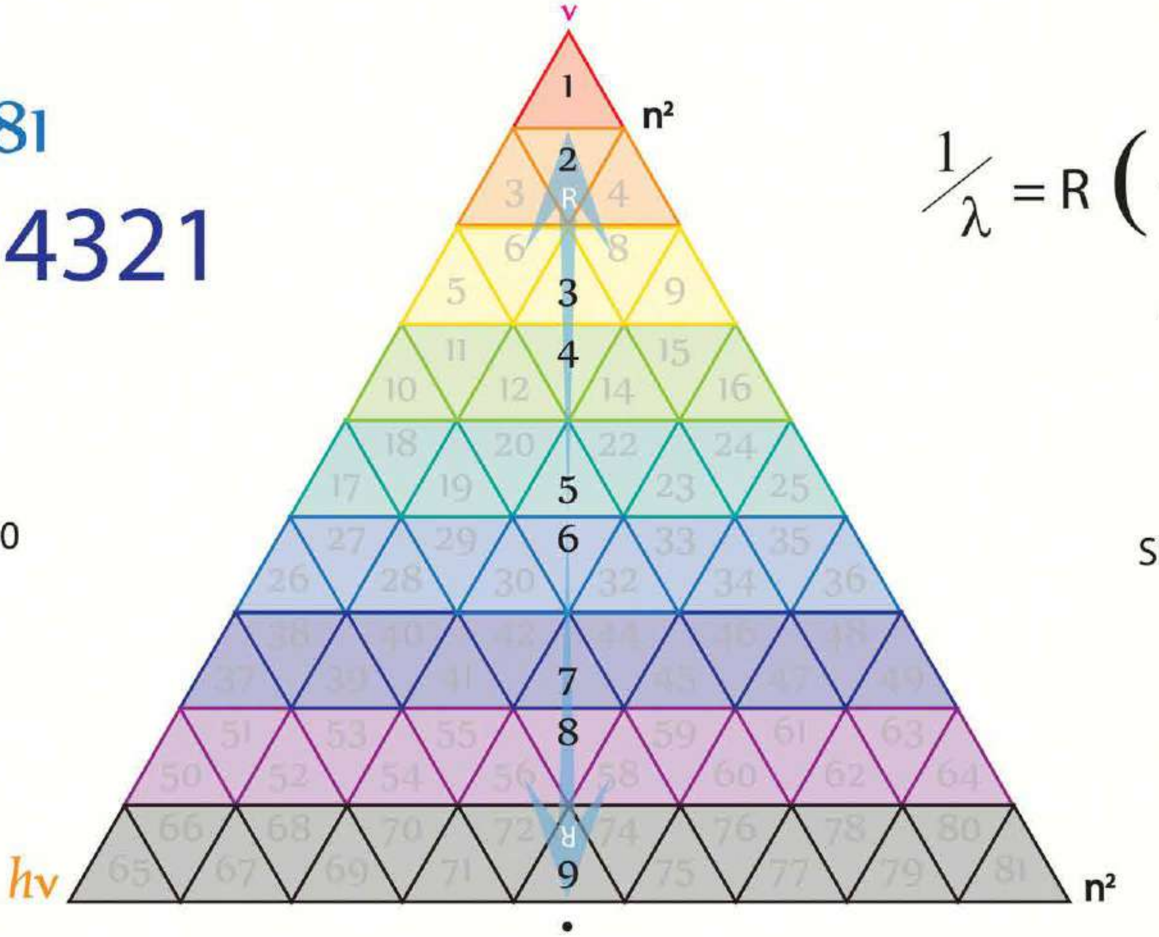
v
1
n²
2
3
R
4
6
8
5
3
9
11
4
15
10
12
14
16
18
20
22
24
17
19
5
23
25
27
29
6
33
35
26
28
30
32
34
36
38
40
42
44
46
48
37
39
41
7
43
45
47
49
51
53
55
8
59
61
63
50
52
54
56
58
60
62
64
66
68
70
72
R
74
76
78
80
hv
65
67
69
71
9
75
77
79
81
n²

$$\frac{1}{\lambda} = R\left(\frac{1}{1} - \frac{1}{81}\right)$$

.987654321

Cos 60

.5

1/2

Sin 60

.866025403

$\sqrt{3}/2$

Tetryonic geometry explains the fractional mathematics of Rydberg's formula

Lyman spectral transitions

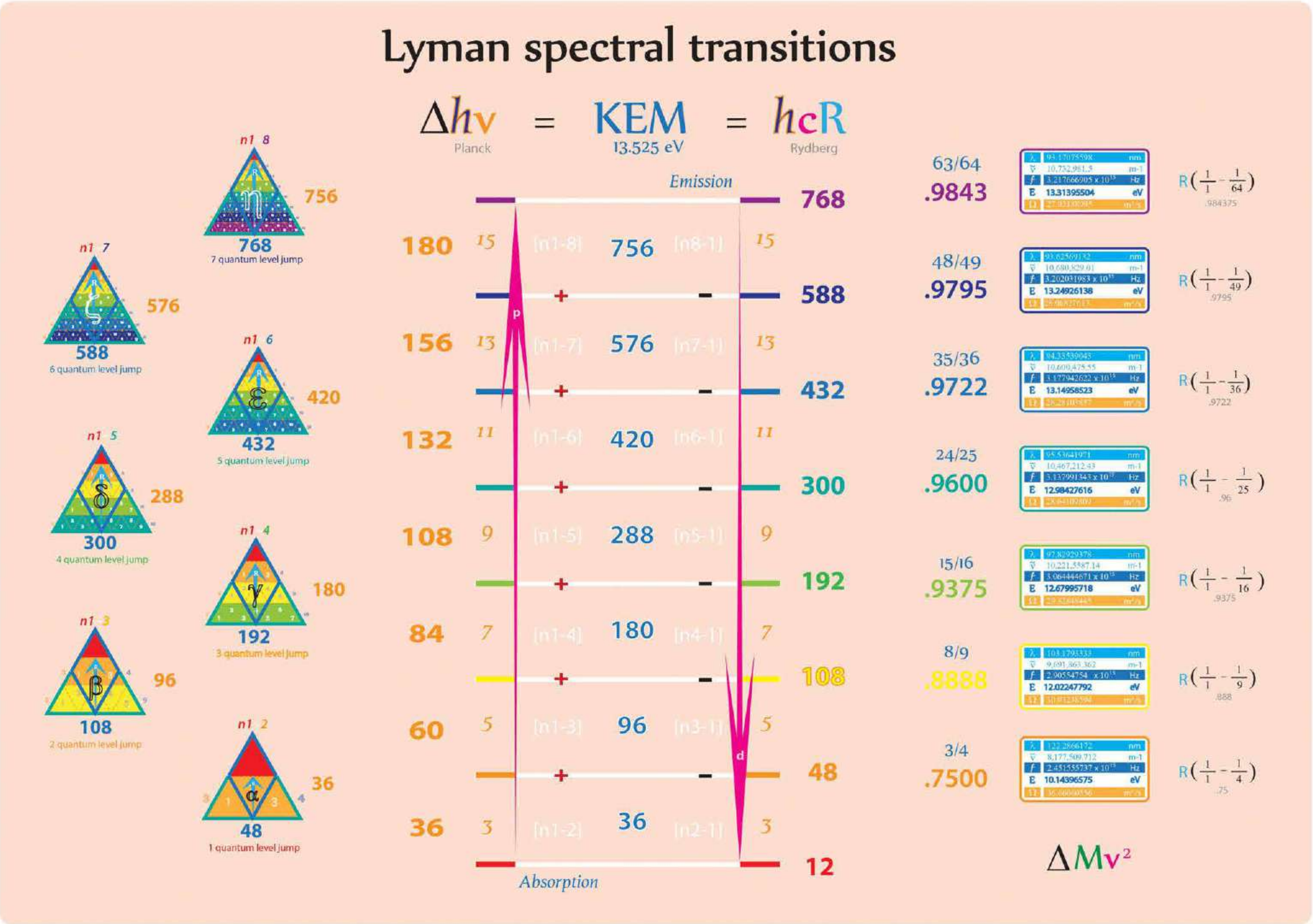

Δhv = KEM = hcR

Planck — 13.525 eV — Rydberg

n1 8
756
η
768
7 quantum level jump
n1 7
576
ζ
588
6 quantum level jump
n1 6
420
ε
432
5 quantum level jump
n1 5
288
δ
300
4 quantum level jump
n1 4
180
γ
192
3 quantum level jump
n1 3
96
β
108
2 quantum level jump
n1 2
36
α
48
1 quantum level jump
Emission
768
180 15 [n1-8] 756 [n8-1] 15
+ 588 -
156 13 [n1-7] 576 [n7-1] 13
+ 432 -
132 11 [n1-6] 420 [n6-1] 11
+ 300 -
108 9 [n1-5] 288 [n5-1] 9
+ 192 -
84 7 [n1-4] 180 [n4-1] 7
+ 108 -
60 5 [n1-3] 96 [n3-1] 5
+ 48 -
36 3 [n1-2] 36 [n2-1] 3
12
Absorption
p
d

63/64
.9843

λ	93.17075598	nm
ṽ	10,732,981.5	m-1
f	3.217666905 x 10^{15}	Hz
E	13.31395504	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{64}\right)$ .984375

48/49
.9795

λ	93.62569132	nm
ṽ	10,680,829.61	m-1
f	3.202031983 x 10^{15}	Hz
E	13.24926138	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{49}\right)$ .9795

35/36
.9722

λ	94.33539043	nm
ṽ	10,600,475.55	m-1
f	3.177942622 x 10^{15}	Hz
E	13.14958523	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{36}\right)$ .9722

24/25
.9600

λ	95.53641971	nm
ṽ	10,467,212.43	m-1
f	3.137991343 x 10^{15}	Hz
E	12.98427616	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{25}\right)$ .96

15/16
.9375

λ	97.82929378	nm
ṽ	10,221,5587.14	m-1
f	3.064444671 x 10^{15}	Hz
E	12.67995718	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{16}\right)$ .9375

8/9
.8888

λ	103.1793333	nm
ṽ	9,691,863.362	m-1
f	2.90554754 x 10^{15}	Hz
E	12.02247792	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{9}\right)$ .888

3/4
.7500

λ	122.2866172	nm
ṽ	8,177,509.712	m-1
f	2.451555737 x 10^{15}	Hz
E	10.14396575	eV
Ω	[illegible]	m²/s

$R\left(\frac{1}{1}-\frac{1}{4}\right)$ .75

ΔMv^2

Balmer spectral transitions

$$\Delta h\nu \quad = \quad \text{KEM} \quad = \quad hcR$$

Planck — 3.381 eV — Rydberg

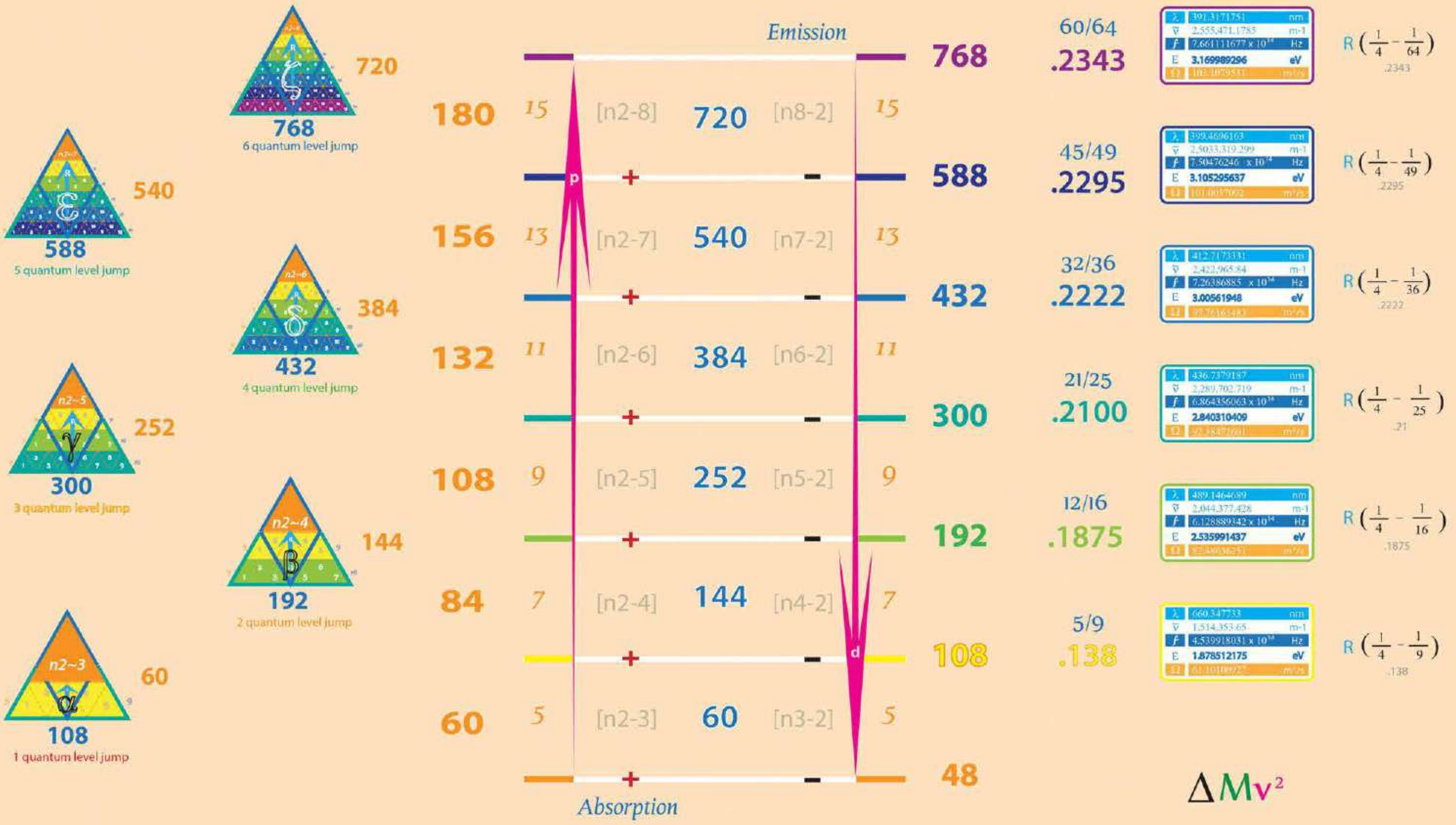

720
768
6 quantum level jump

540
588
5 quantum level jump

384
432
4 quantum level jump

252
300
3 quantum level jump

144
192
2 quantum level jump

60
108
1 quantum level jump

Emission

						768
180	15	[n2-8]	720	[n8-2]	15	
						588
156	13	[n2-7]	540	[n7-2]	13	
						432
132	11	[n2-6]	384	[n6-2]	11	
						300
108	9	[n2-5]	252	[n5-2]	9	
						192
84	7	[n2-4]	144	[n4-2]	7	
						108
60	5	[n2-3]	60	[n3-2]	5	
						48

Absorption

		λ (nm)	$\tilde{\nu}$ (m-1)	*f* (Hz)	E (eV)	Ω (m²/s)	
768	60/64 .2343	391.3171751	2,555,471.1785	7.661111677 x 10^{14}	**3.169989296**	103.1079531	$R\left(\frac{1}{4}-\frac{1}{64}\right)$ .2343
588	45/49 .2295	399.4696163	2,5033,319.299	7.50476246 x 10^{14}	**3.105295637**	101.0057092	$R\left(\frac{1}{4}-\frac{1}{49}\right)$ .2295
432	32/36 .2222	412.7173331	2,422,965.84	7.26386885 x 10^{14}	**3.00561948**	97.76161483	$R\left(\frac{1}{4}-\frac{1}{36}\right)$ .2222
300	21/25 .2100	436.7379187	2,289,702.719	6.864356063 x 10^{14}	**2.840310409**	92.38472691	$R\left(\frac{1}{4}-\frac{1}{25}\right)$ .21
192	12/16 .1875	489.1464689	2,044,377.428	6.128889342 x 10^{14}	**2.535991437**	87.48036251	$R\left(\frac{1}{4}-\frac{1}{16}\right)$ .1875
108	5/9 .138	660.347733	1,514,353.65	4.539918031 x 10^{14}	**1.878512175**	81.10108927	$R\left(\frac{1}{4}-\frac{1}{9}\right)$ .138

$\Delta M v^2$

Copyright ABRAHAM [2008] - All rights reserved

Paschen spectral transitions

$\Delta h\nu$ = KEM = hcR

Planck — 1.502 eV — Rydberg

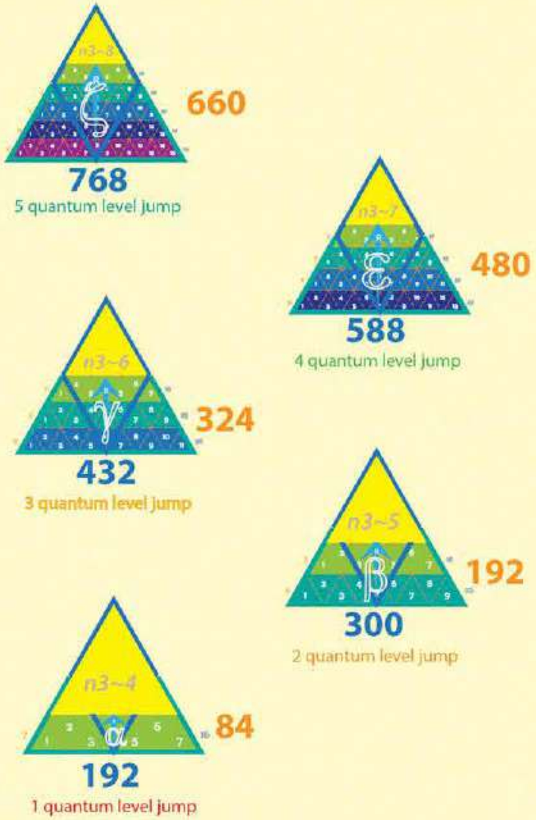
n3-8
ζ
660
768
5 quantum level jump
n3-7
ε
480
588
4 quantum level jump
n3-6
γ
324
432
3 quantum level jump
n3-5
β
192
300
2 quantum level jump
n3-4
α
84
192
1 quantum level jump

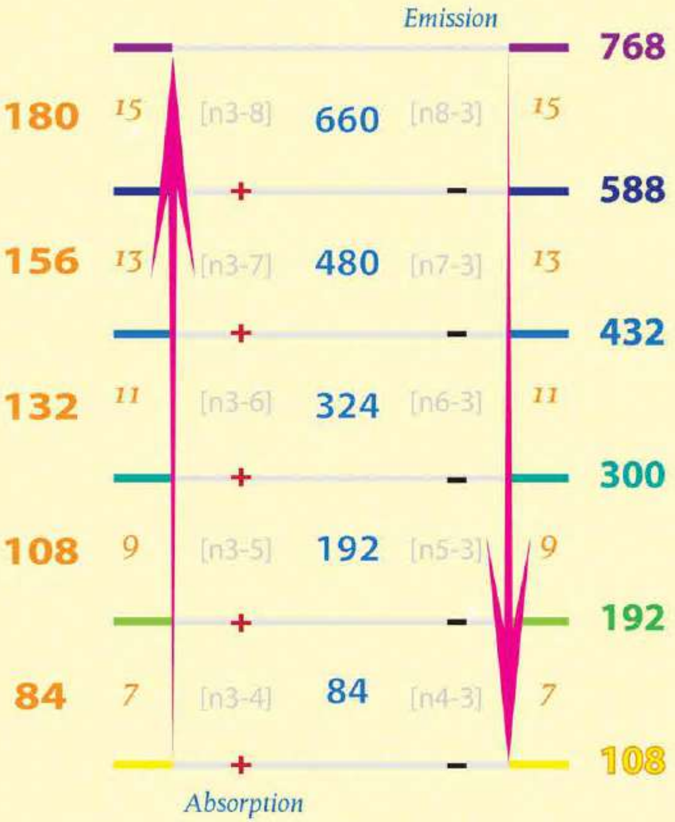
Emission
768
180
15
[n3-8]
660
[n8-3]
15
+
-
588
156
13
[n3-7]
480
[n7-3]
13
+
-
432
132
11
[n3-6]
324
[n6-3]
11
+
-
300
108
9
[n3-5]
192
[n5-3]
9
+
-
192
84
7
[n3-4]
84
[n4-3]
7
+
-
108
Absorption

55/64
.0954

λ	960.5057935	nm
$\tilde{\nu}$	1,041,118.135	m-1
f	$3.121193646 \times 10^{14}$	Hz
E	1.29147712	eV
Ω	287.9523928	m^2/s

$R\left(\frac{1}{9}-\frac{1}{64}\right)$ .09548

40/49
.0907

λ	1,011.157466	nm
$\tilde{\nu}$	988,965.6492	m-1
f	$2.9464844428 \times 10^{14}$	Hz
E	1.226783461	eV
Ω	303.1373822	m^2/s

$R\left(\frac{1}{9}-\frac{1}{49}\right)$ .0907

27/36
.8033

λ	1,100.579555	nm
$\tilde{\nu}$	908,612.1902	m-1
f	$2.723950819 \times 10^{14}$	Hz
E	1.127107305	eV
Ω	329.94545	m^2/s

$R\left(\frac{1}{9}-\frac{1}{36}\right)$ .0833

16/25
.0711

λ	1,289.741666	nm
$\tilde{\nu}$	775,349.069	m-1
f	$2.324438032 \times 10^{14}$	Hz
E	0.961798233	eV
Ω	386.6548243	m^2/s

$R\left(\frac{1}{9}-\frac{1}{25}\right)$ .0711

7/16
.0486

λ	1,886.707809	nm
$\tilde{\nu}$	530,023.7776	m-1
f	$1.588971311 \times 10^{14}$	Hz
E	0.657479261	eV
Ω	565.6207715	m^2/s

$R\left(\frac{1}{9}-\frac{1}{16}\right)$ .0486

ΔMv^2

Copyright ABRAHAM [2008] - All rights reserved

Brackett spectral transitions

$$\Delta h\nu = \text{KEM} = hcR$$

Planck 0.8451 eV Rydberg

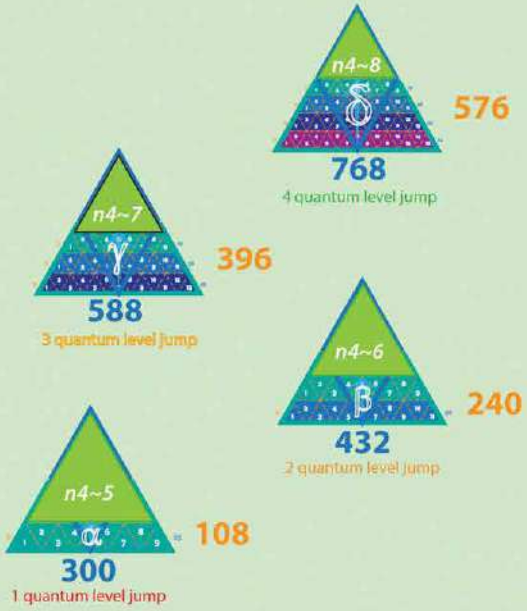
n4~8
δ
576
768
4 quantum level jump
n4~7
γ
396
588
3 quantum level jump
n4~6
β
240
432
2 quantum level jump
n4~5
α
108
300
1 quantum level jump

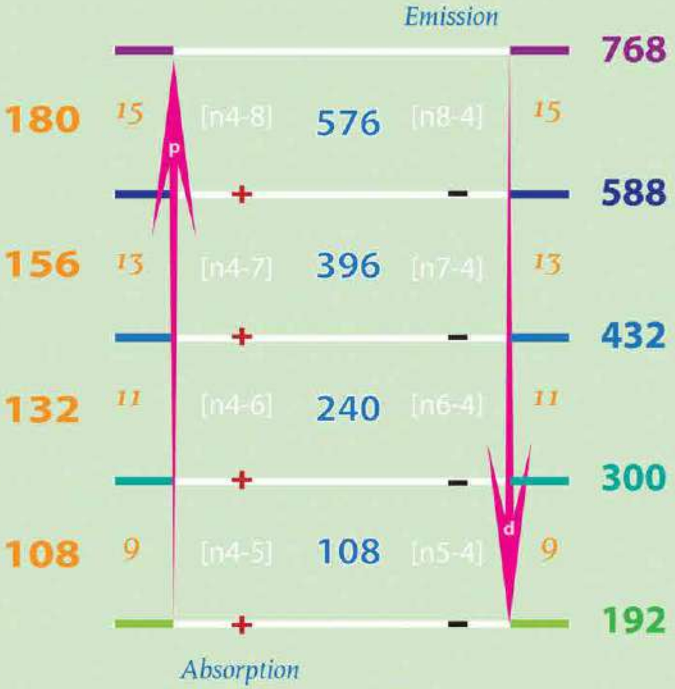
Emission
768
180
15
[n4-8]
576
[n8-4]
15
p
+
−
588
156
13
[n4-7]
396
[n7-4]
13
+
−
432
132
11
[n4-6]
240
[n6-4]
11
+
−
300
108
9
[n4-5]
108
[n5-4]
9
d
+
−
192
Absorption

48/64
.0468

λ	1,956.585876	nm
$\tilde{\nu}$	511,094.357	m-1
f	1.532222335 x 10 ^{14}	Hz
E	**0.634001085**	**eV**
Ω	586.569689	m²/s

$R\left(\frac{1}{16}-\frac{1}{64}\right)$ .0468

33/49
.04209

λ	2,178.92518	nm
$\tilde{\nu}$	458,941.8716	m-1
f	1.375873118 x 10 ^{14}	Hz
E	**0.569307096**	**eV**
Ω	653.2233354	m²/s

$R\left(\frac{1}{16}-\frac{1}{49}\right)$ .04209

20/36
.0347

λ	2,641.390932	nm
$\tilde{\nu}$	378,588.4126	m-1
f	1.134979508 x 10 ^{14}	Hz
E	**0.469630433**	**eV**
Ω	791.8690801	m²/s

$R\left(\frac{1}{16}-\frac{1}{36}\right)$ .0347

9/25
.0225

λ	4,076.220574	nm
$\tilde{\nu}$	245,325.2914	m-1
f	7.35466721x 10 ^{13}	Hz
E	**0.30432052**	**eV**
Ω	1,222.020185	m²/s

$R\left(\frac{1}{16}-\frac{1}{25}\right)$ .0225

$\Delta M v^2$

Pfund spectral transitions

$$\Delta h\nu = KEM = hcR$$

Planck P 0.541 eV Rydberg

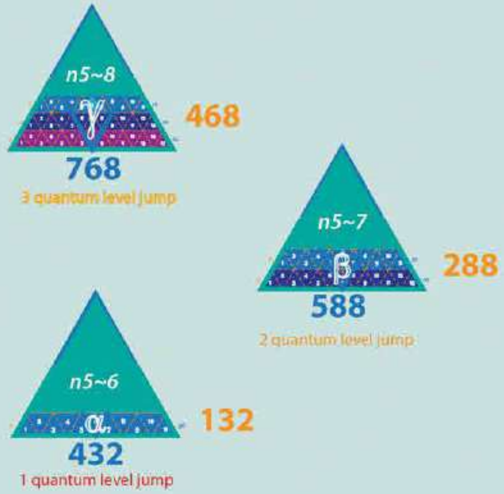
n5~8
γ
468
768
3 quantum level jump
n5~7
β
288
588
2 quantum level jump
n5~6
α
132
432
1 quantum level jump

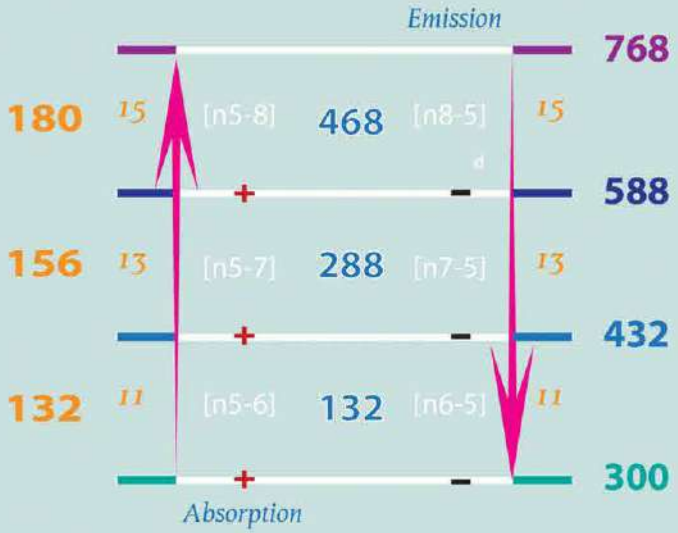
Emission
768
180
15
[n5-8]
468
[n8-5]
15
588
156
13
[n5-7]
288
[n7-5]
13
432
132
11
[n5-6]
132
[n6-5]
11
300
Absorption

39/64
.0243

λ	3,762.665146	nm
$\tilde{\nu}$	265,769.0656	m-1
f	$7.967556145 \times 10^{13}$	Hz
E	**0.329678886**	**eV**
Ω	1,128.018633	m^2/s

$R\left(\frac{1}{25}-\frac{1}{64}\right)$ .0243

24/49
.0195

λ	4,681.284566	nm
$\tilde{\nu}$	213,616.5802	m-1
f	$6.404063965 \times 10^{13}$	Hz
E	**0.264985227**	**eV**
Ω	1,403.413807	m^2/s

$R\left(\frac{1}{25}-\frac{1}{49}\right)$ .01959

11/36
.0122

λ	7,503.951512	nm
$\tilde{\nu}$	133,263.1212	m-1
f	$3.995127867 \times 10^{13}$	Hz
E	**0.165309071**	**eV**
Ω	2,249.628068	m^2/s

$R\left(\frac{1}{25}-\frac{1}{36}\right)$ .0122

$\Delta M v^2$

Humphreys spectral transitions

$$\Delta hv = \text{KEM} = hcR$$

Planck

0.375 eV

Rydberg

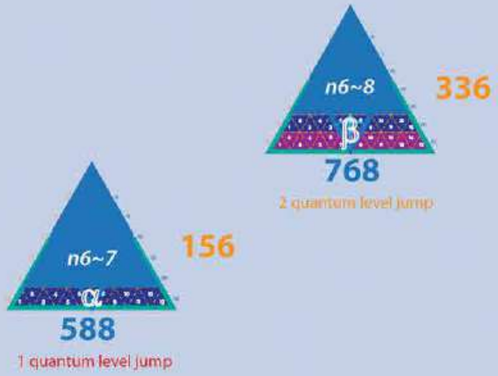
n6~8
β
336
768
2 quantum level jump
n6~7
α
156
588
1 quantum level jump

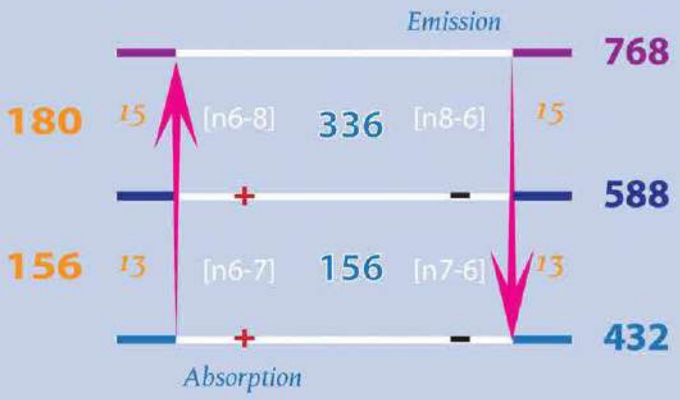
Emission
768
180
15
[n6-8]
336
[n8-6]
15
+
-
588
156
13
[n6-7]
156
[n7-6]
13
+
-
432
Absorption

28/64

.01215

λ	7,546.831235	nm
$\tilde{v}$	132,505.9444	m-1
f	$3.972428277 \times 10^{13}$	Hz
E	**0.164369815**	**eV**
Ω	1,128.018633	m^2/s

$$R\left(\frac{1}{36} - \frac{1}{64}\right)$$

.01215

13/49

.00733

λ	12,445.01497	nm
$\tilde{v}$	80,353.459	m-1
f	$2.4089936098 \times 10^{13}$	Hz
E	**0.099676156**	**eV**
Ω	1,403.413807	m^2/s

$$R\left(\frac{1}{36} - \frac{1}{49}\right)$$

.0073

$$\Delta Mv^2$$

Copyright ABRAHAM [2008] - All rights reserved

Un-named spectral transition

$\Delta h\nu$ = KEM = hcR

Planck 0.276 eV Rydberg

180

768

1 quantum level jump

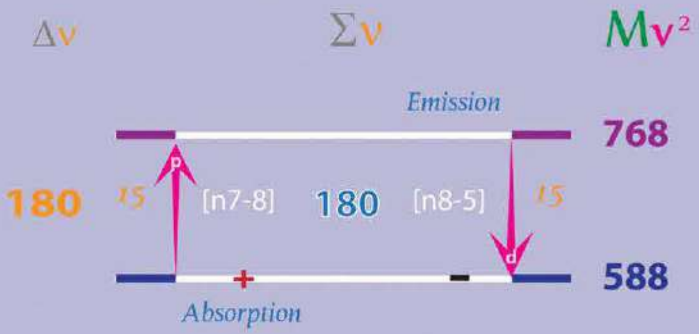

Δv
Σv
Mv²
Emission
768
180
15
p
[n7-8]
180
[n8-5]
15
d
+
-
588
Absorption

15/64

.0047

λ	19,174.54158	nm
ṽ	52,152.48541	m-1
f	$1.563492179 \times 10^{13}$	Hz
E	0.064693658	eV
Ω	5.748382952	m²/s

$R\left(\frac{1}{49}-\frac{1}{64}\right)$

.0047

$\Delta M v^2$

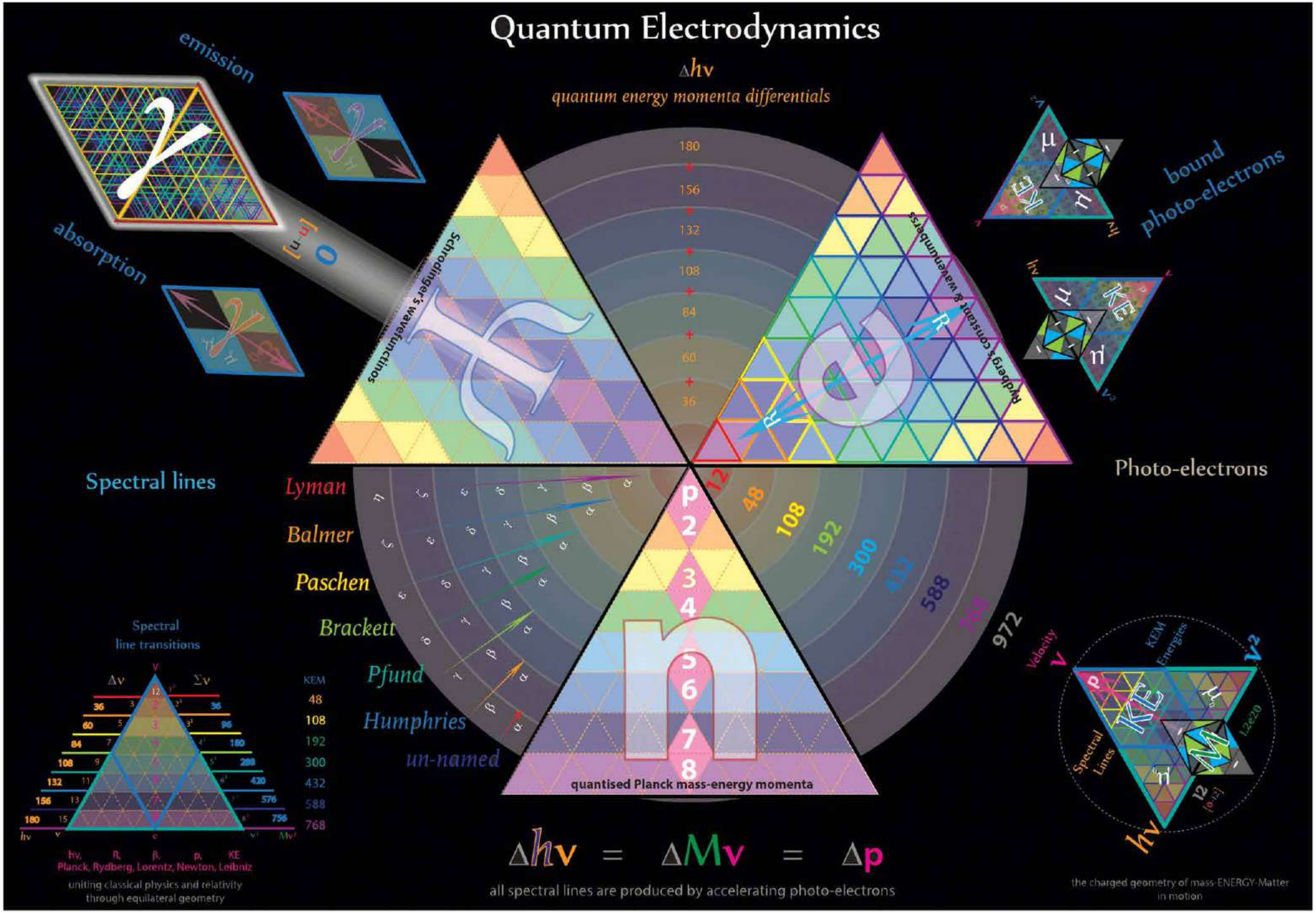
Quantum Electrodynamics
Δhv
quantum energy momenta differentials
emission
absorption
[n-n]
0
180
156
132
108
84
60
36
Schrodinger's wavefunctions
Rydberg's constant & wavenumbers
R
R
bound
photo-electrons
μ
KE
ɳ
Photo-electrons
Spectral lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
p
2
3
4
5
6
7
8
n
12
48
108
192
300
432
588
768
972
quantised Planck mass-energy momenta
Spectral
line transitions
Δv
Σv
v
12
36
60
84
108
132
156
180
3
5
7
9
11
13
15
36
96
180
288
420
576
756
hv
v
c
v²
Mv²
KEM
48
108
192
300
432
588
768
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz
uniting classical physics and relativity
through equilateral geometry
Velocity
v
KEM
Energies
v²
KE
M
Spectral
Lines
1.2e20
12
[ϕ-12]
hv
Δhv = ΔMv = Δp
all spectral lines are produced by accelerating photo-electrons
the charged geometry of mass-ENERGY-Matter
in motion

Tetryonics 96.11 - Atomic spectral series transitions

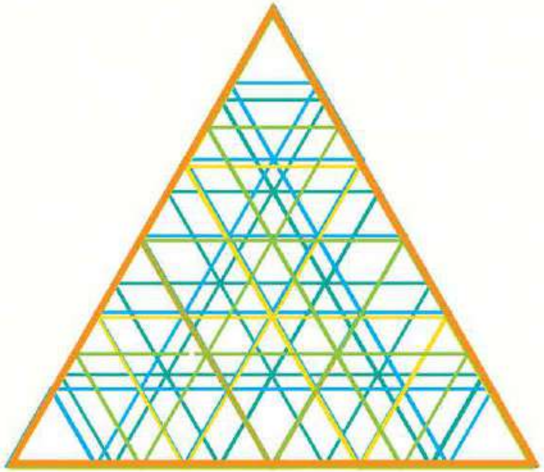

Fractionals and fractals

White light
[EM radiation]
is comprised of many
superpositioned
frequencies

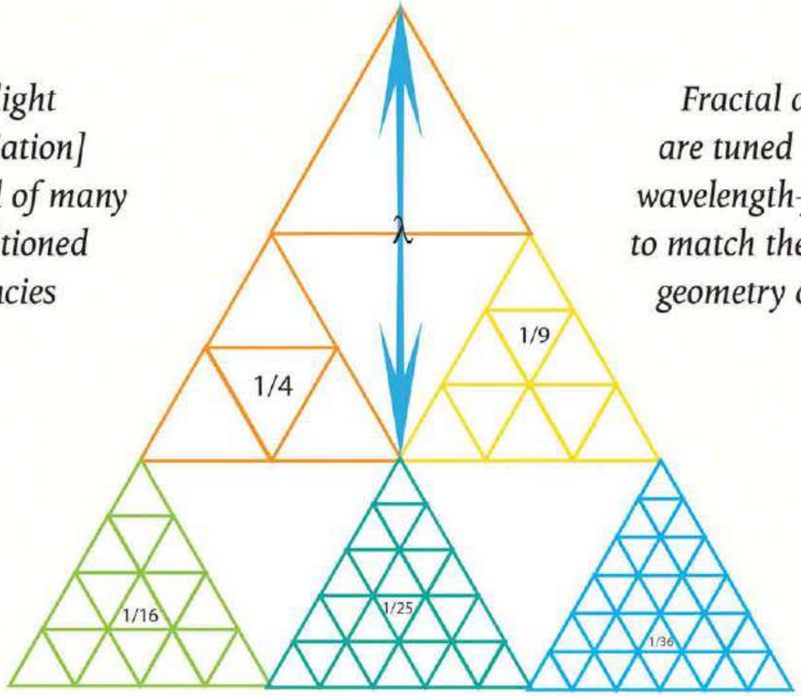

Fractal antennas
are tuned to specific
wavelength-frequencies
to match the equilateral
geometry of photons

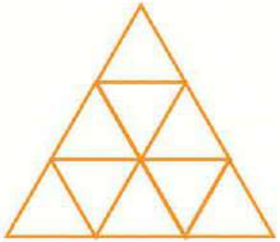

ninths

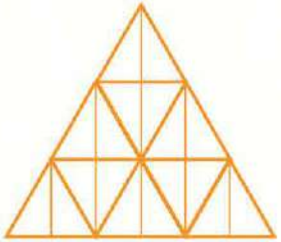

eigteenths

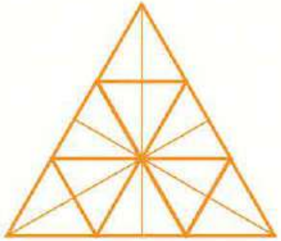

Eighteenths

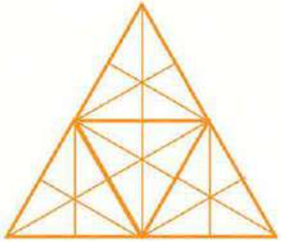

Twenty-fourths

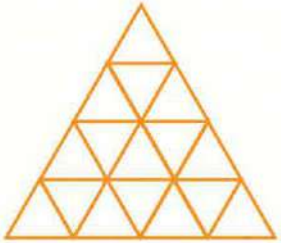

Sizteenths

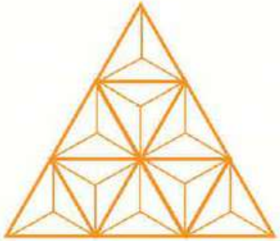

Twenty-sevenths

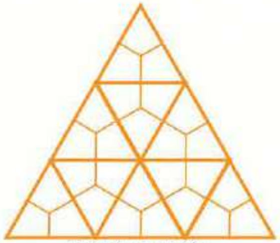

Twenty-sevenths

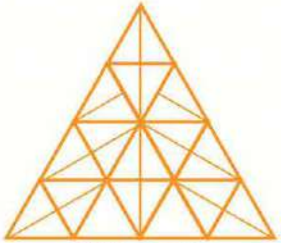

Thirty-seconds

Polygons and Angles

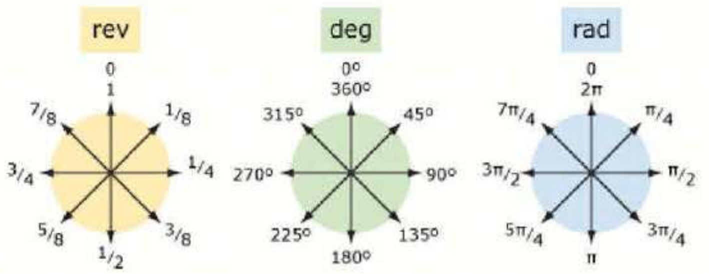
rev
0
1
1/8
1/4
3/8
1/2
5/8
3/4
7/8
deg
0°
360°
45°
90°
135°
180°
225°
270°
315°
rad
0
2π
π/4
π/2
3π/4
π
5π/4
3π/2
7π/4

polygon exterior and interior angles

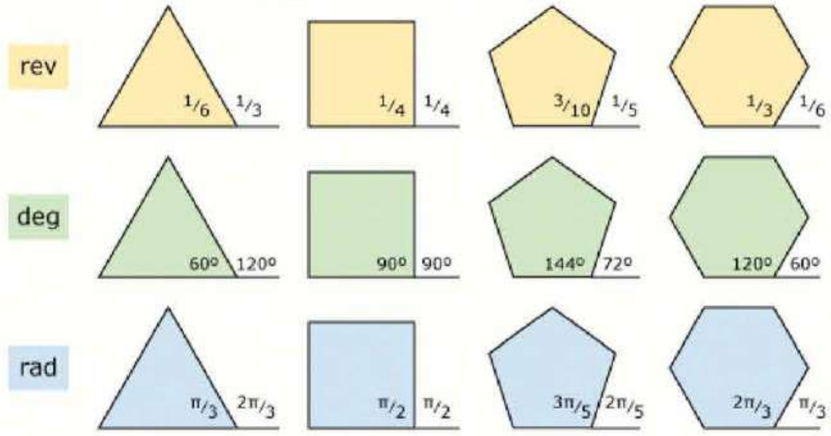
rev
1/6
1/3
1/4
1/4
3/10
1/5
1/3
1/6
deg
60°
120°
90°
90°
144°
72°
120°
60°
rad
π/3
2π/3
π/2
π/2
3π/5
2π/5
2π/3
π/3

dividing the circle

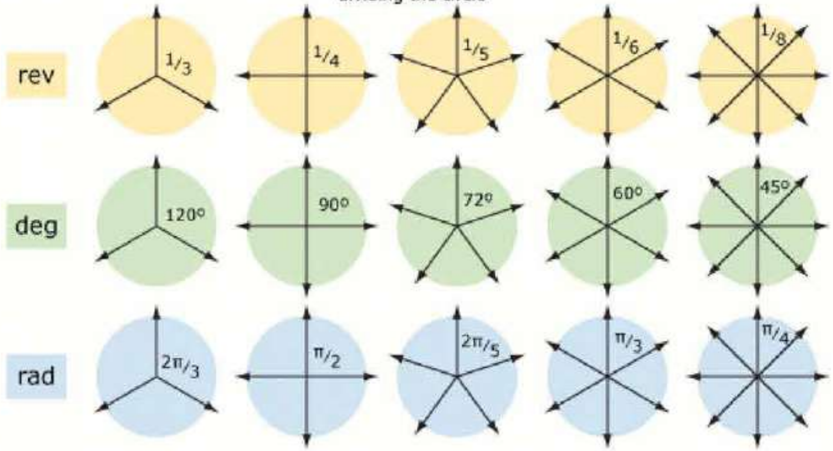
rev
1/3
1/4
1/5
1/6
1/8
deg
120°
90°
72°
60°
45°
rad
2π/3
π/2
2π/5
π/3
π/4

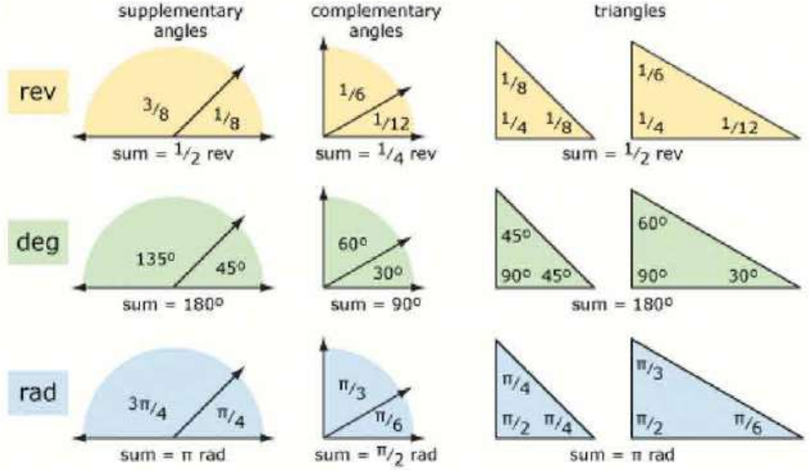
supplementary angles
complementary angles
triangles
rev
3/8
1/8
sum = 1/2 rev
1/6
1/12
sum = 1/4 rev
1/8
1/4
1/8
1/6
1/4
1/12
sum = 1/2 rev
deg
135°
45°
sum = 180°
60°
30°
sum = 90°
45°
90°
45°
60°
90°
30°
sum = 180°
rad
3π/4
π/4
sum = π rad
π/3
π/6
sum = π/2 rad
π/4
π/2
π/4
π/3
π/2
π/6
sum = π rad

Copyright ABRAHAM [2008] - All rights reserved

Pi vs Tau

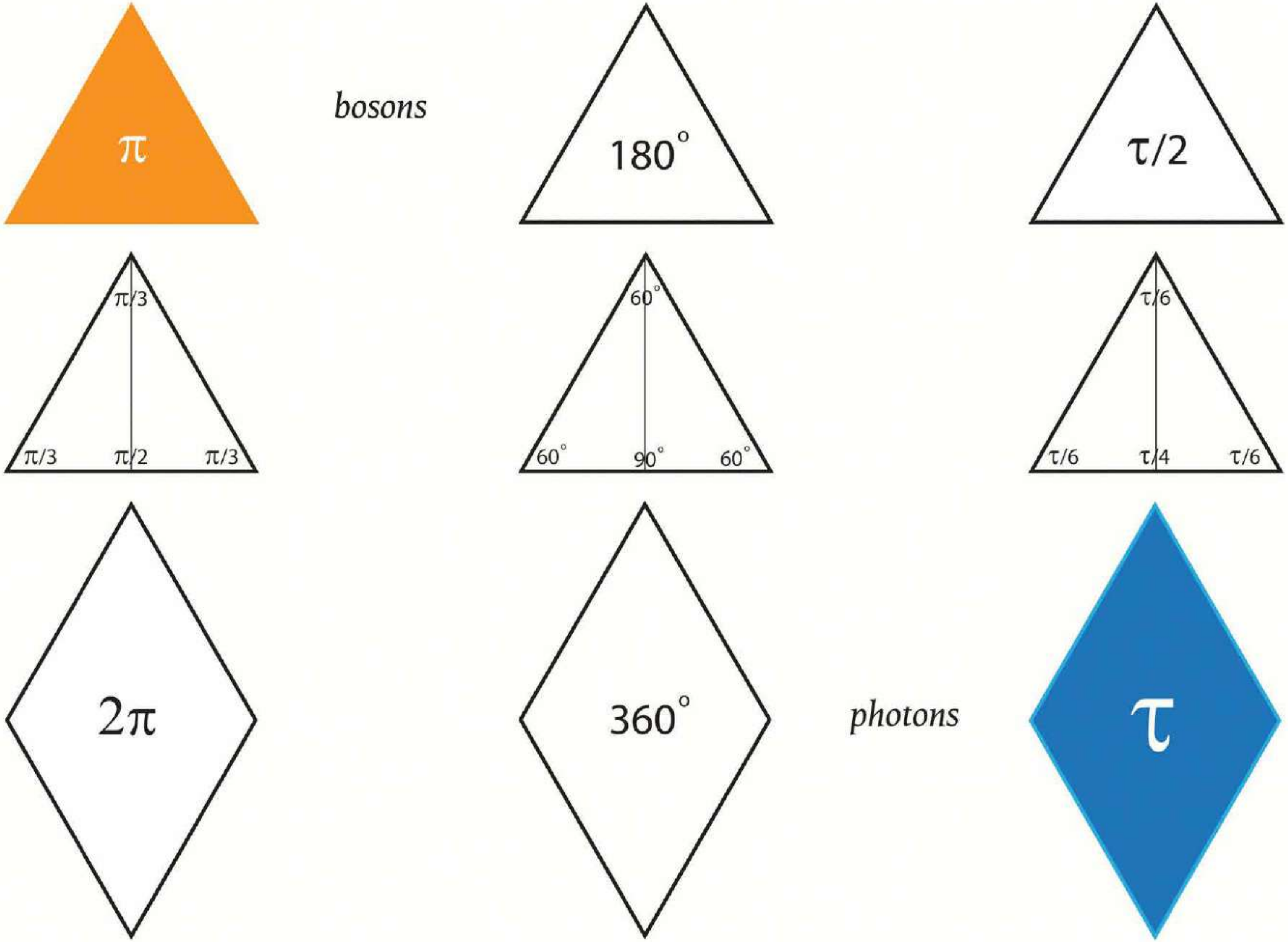
π
bosons
180°
τ/2
π/3
π/3
π/2
π/3
60°
60°
90°
60°
τ/6
τ/6
τ/4
τ/6
2π
360°
photons
τ

Triangular dissection of an equilateral triangle

is a way of dividing up a original triangle into smaller equilateral triangles,
such that none of the smaller triangles overlap

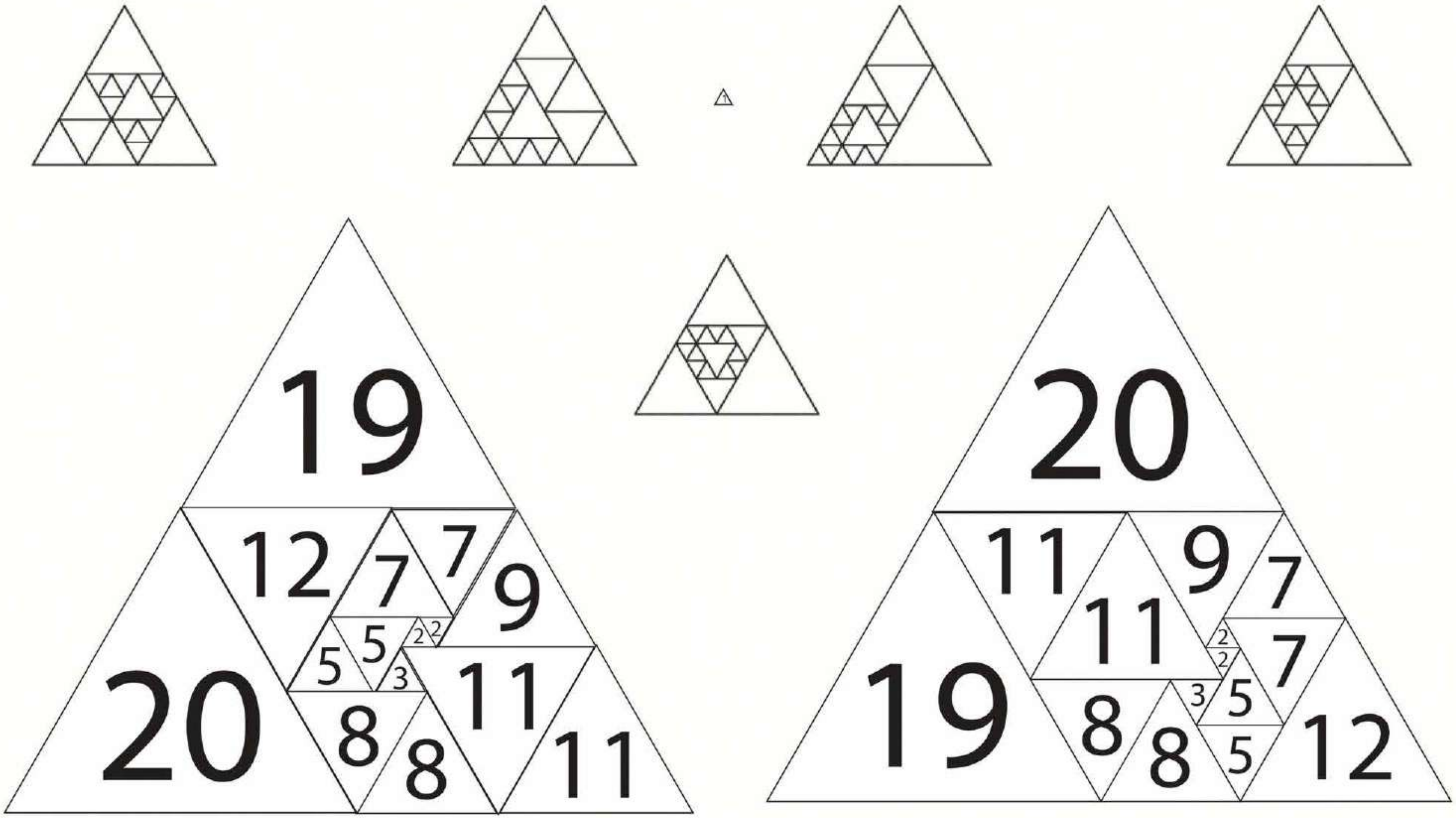

lowest order perfect equilateral triangle dissected by equilateral triangles

lowest order perfect dissected equilateral triangle, an isomer of the first

Copyright ABRAHAM [2008] - All rights reserved

Golden mean Spirals

Golden Mean Spiral – This spiral is derived via the golden rectangle, a unique rectangle which has the golden ratio.
This form is found everywhere in nature: the Nautilus Shell, the face of a Sunflower, fingerprints, our DNA, and the shape of the Milky Way

Golden spiral

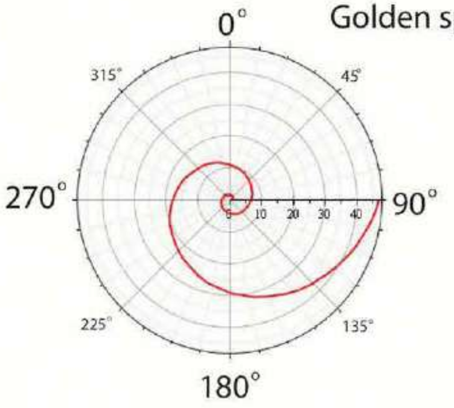

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, ...

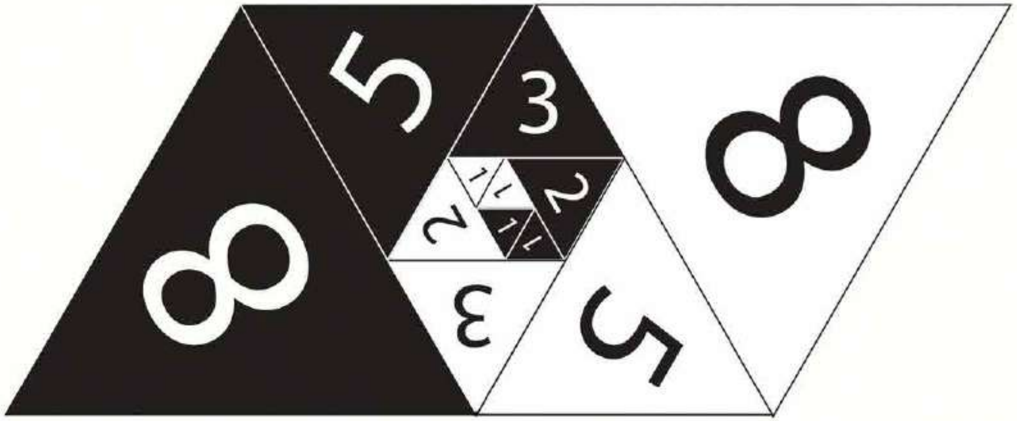

Golden spirals

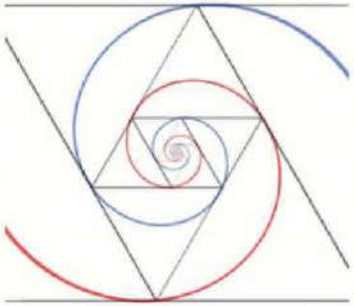

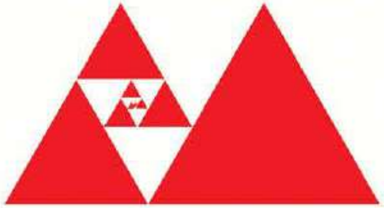

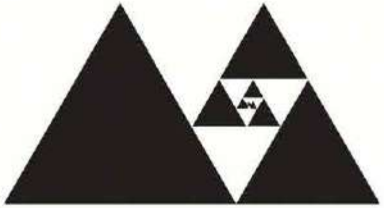

Continued fractions and the Fibonacci sequence

The convergence of the continued fractions

$$1+\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\cdots$$

are:

$$\frac{2}{1},\ \frac{3}{2},\ \frac{5}{3},\ \frac{8}{5},\ \frac{13}{8},\ \frac{21}{13},\ \frac{34}{21},\ \frac{55}{34},\ \frac{89}{55},\ \frac{144}{89},\ \frac{233}{144},\ \frac{377}{233},\ \frac{610}{377},$$

The Golden Ratio (Golden Mean, Golden Section)
is defined mathematically as:

$$\varphi=\frac{1+\sqrt{5}}{2}=1.6180339887\ldots.$$

Koch fractal Curve

Niels Fabian Helge von Koch (January 25, 1870 – March 11, 1924) was a Swedish mathematician who gave his name to one of the earliest fractal curves ever known

He described the Koch curve, or Koch snowflakes as it popularly known, in a 1904 paper entitled "On a continuous curve without tangents constructible from elementary geometry"

Von Koch wrote several papers on number theory . One of his results was a 1901 theorem proving that the Riemann hypothesis is equivalent to a strengthened form of the prime number theorem.

Three Koch curves form the snowflake.

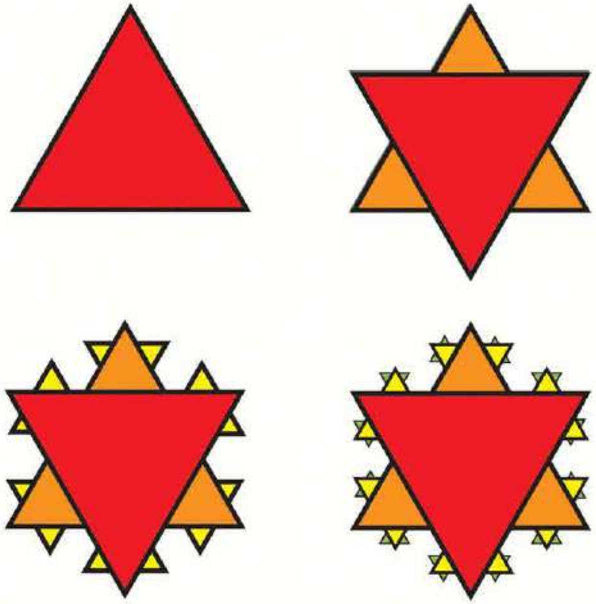

The Koch snowflake (or Koch star) is a mathematical curve and one of the earliest fractal curves to have been described.

Actually Koch described what is now known as the Koch curve, which is the same as the now popular snowflake, except it starts with a line segment instead of an equilateral triangle.

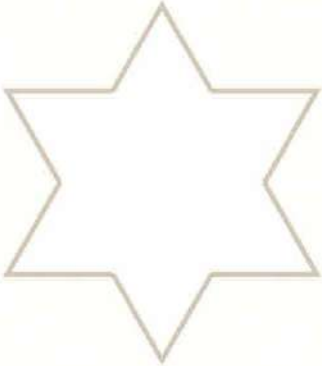

$$a = \frac{1}{2} + \frac{i}{\sqrt{12}}$$

The Koch curve is a special case of the Cesaro curve where:

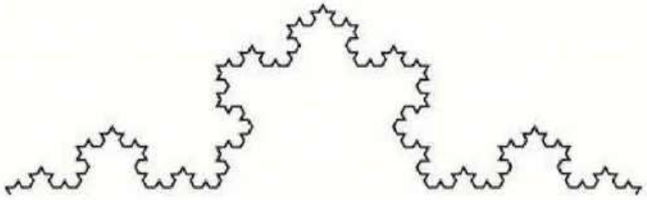

which is in turn a special case of the de Rham curve.

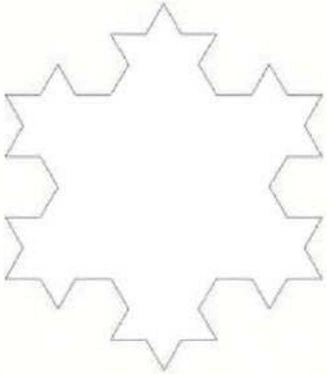

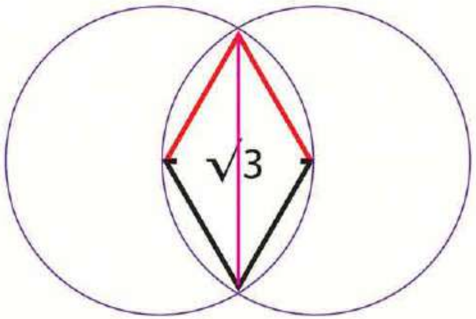

vesica piscis

$$\frac{1351}{780} > \sqrt{3} > \frac{265}{153}\,.$$

Koch snowflake

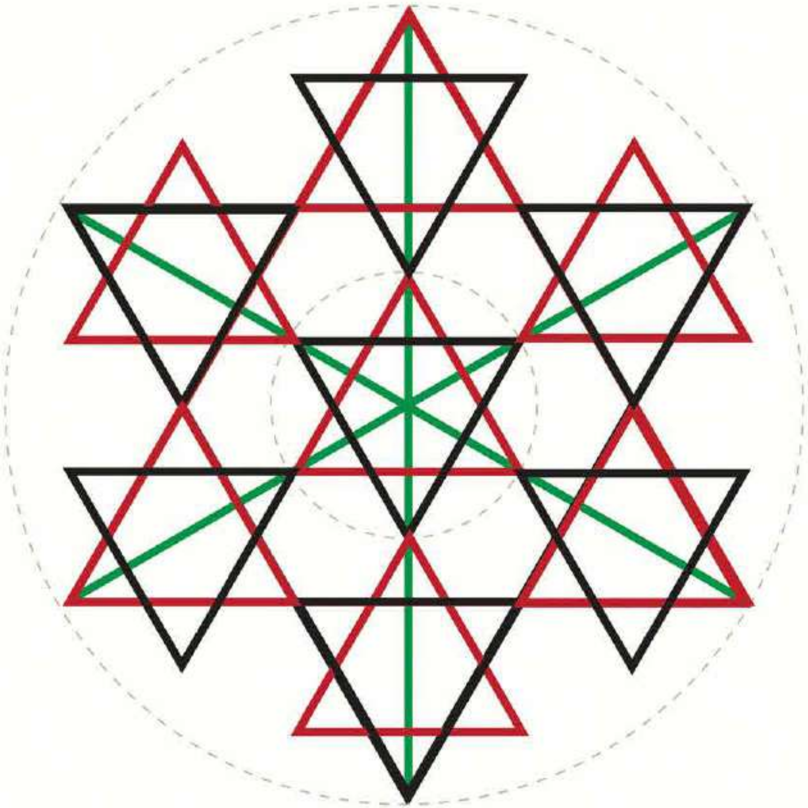

space-time

star tetrahedron

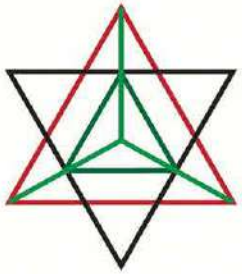

Flower of Life

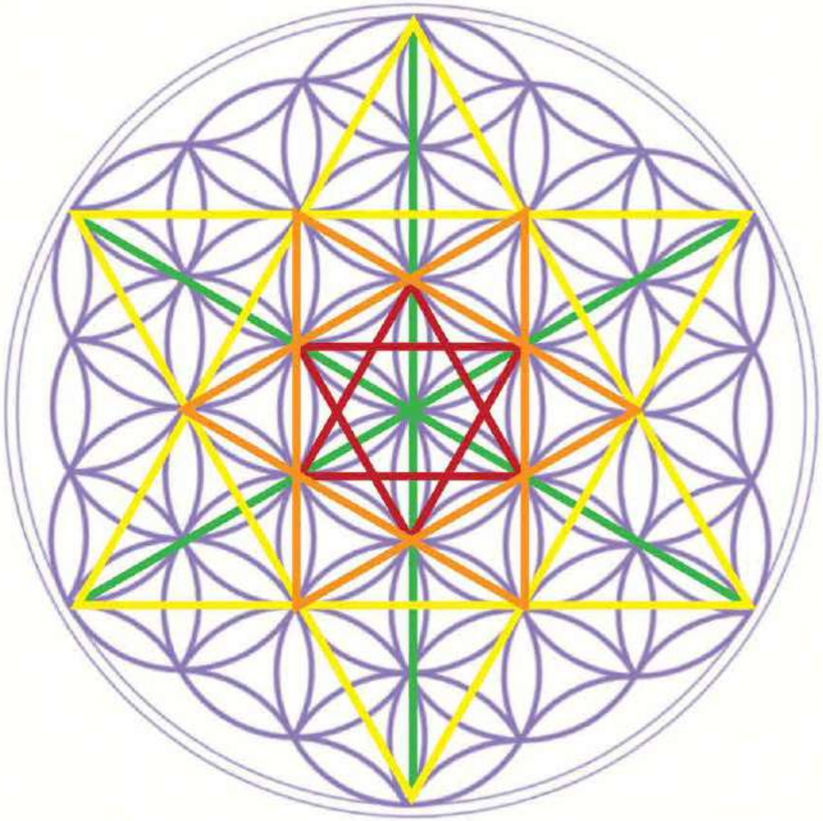

The Flower of Life is a name for a geometrical figure composed of multiple evenly-spaced, overlapping circles.

Copyright ABRAHAM [2008] - All rights reserved

Unit circles - SINE WAVES - Photons

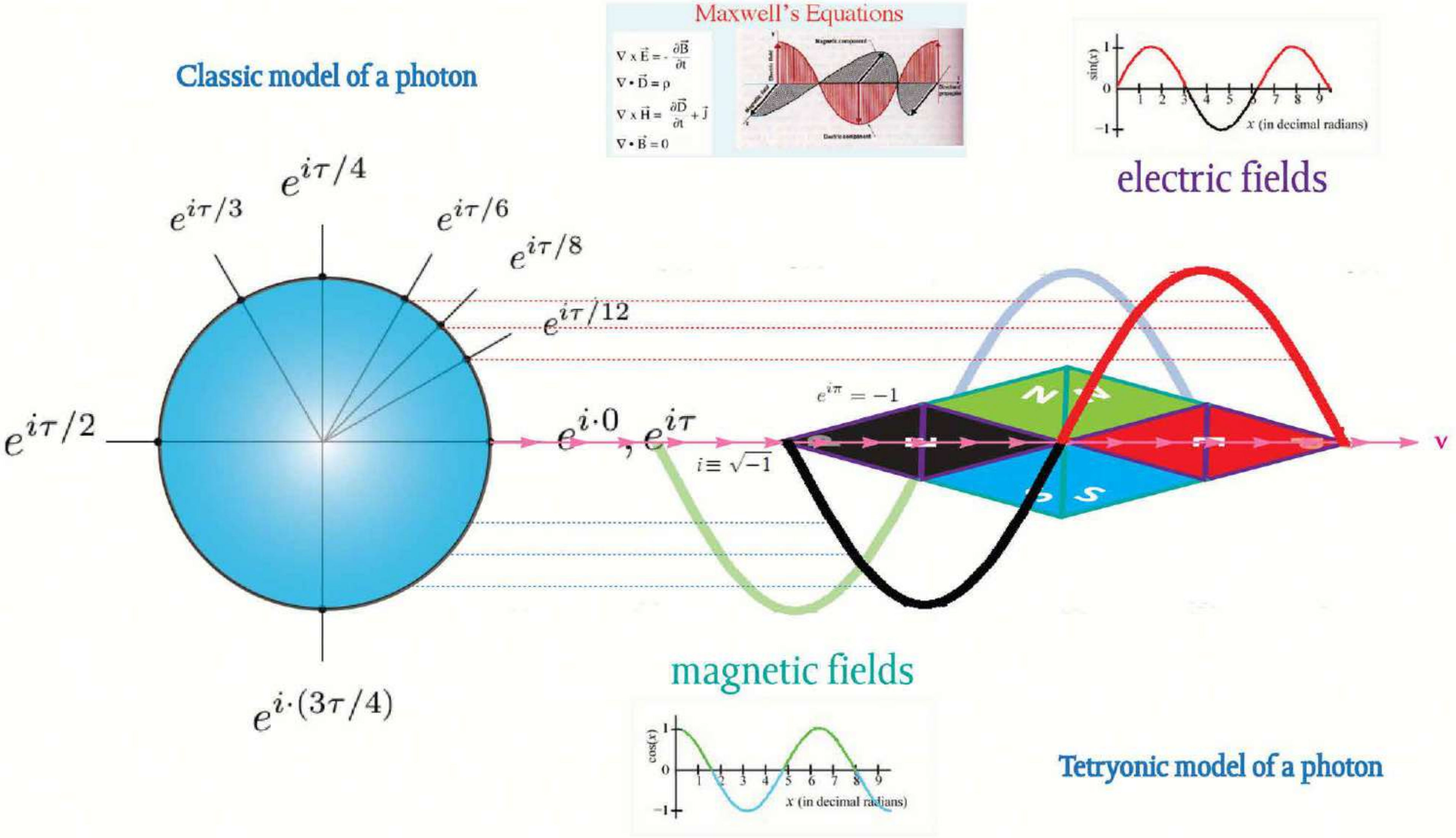

Classic model of a photon

Maxwell's Equations

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\nabla \cdot \vec{D} = \rho$$

$$\nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J}$$

$$\nabla \cdot \vec{B} = 0$$

electric fields

magnetic fields

Tetryonic model of a photon

Magnetic component
Electric component
$e^{i\tau/4}$
$e^{i\tau/3}$
$e^{i\tau/6}$
$e^{i\tau/8}$
$e^{i\tau/12}$
$e^{i\tau/2}$
$e^{i\cdot 0}, e^{i\tau}$
$i \equiv \sqrt{-1}$
$e^{i\pi} = -1$
$e^{i\cdot(3\tau/4)}$
N
N
S
S
v
sin(x)
1
0
−1
1 2 3 4 5 6 7 8 9
x (in decimal radians)
cos(x)
1
0
−1
1 2 3 4 5 6 7 8 9
x (in decimal radians)

Magnetic waveforms are 90 degrees out of phase with Electric waveforms

Copyright ABRAHAM [2008] - All rights reserved

Boson distributions in monochromatic EM waves

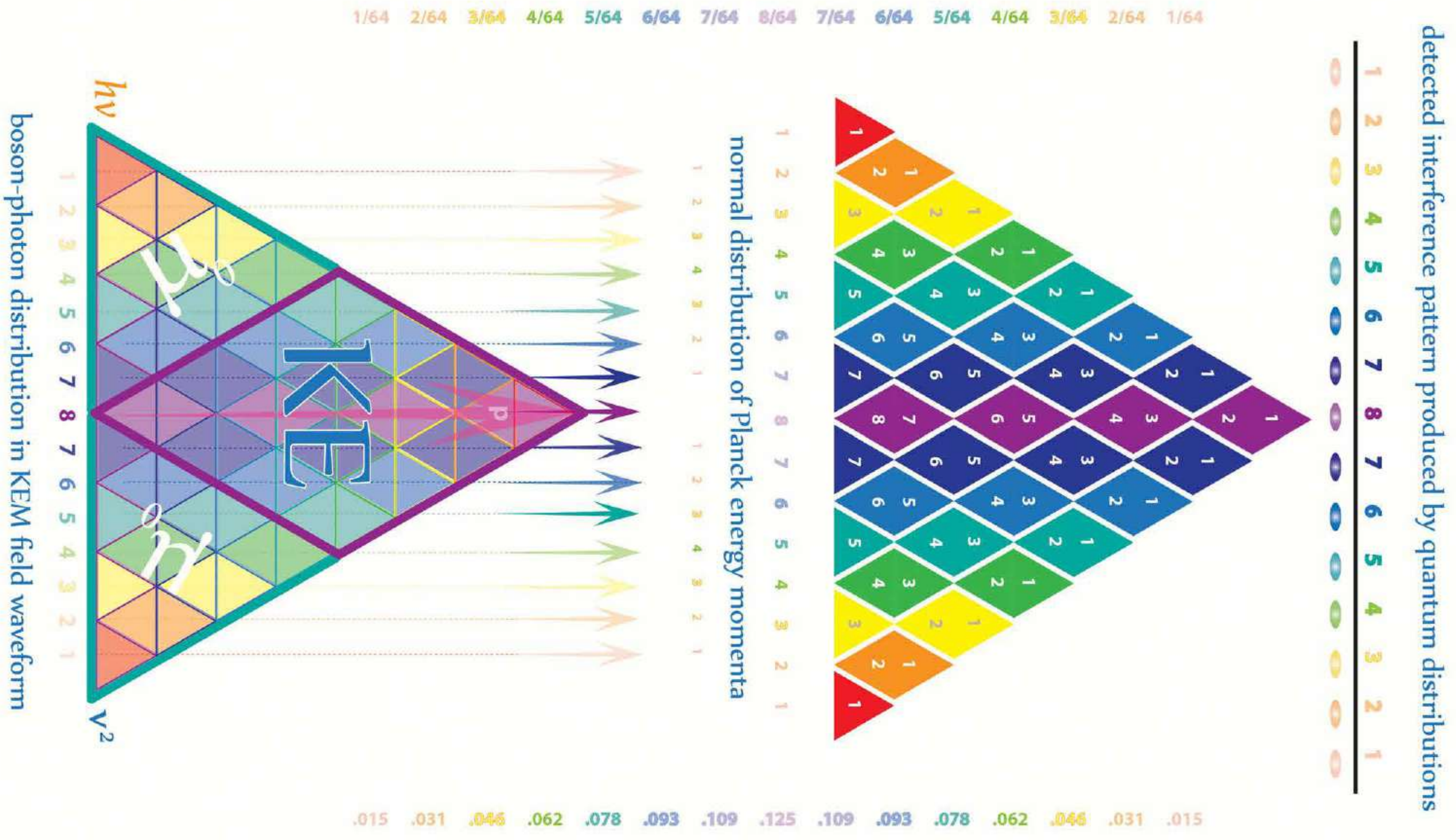
1/64 2/64 3/64 4/64 5/64 6/64 7/64 8/64 7/64 6/64 5/64 4/64 3/64 2/64 1/64
hν
v^2
boson-photon distribution in KEM field waveform
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
μ₀
ε₀
KE
p
normal distribution of Planck energy momenta
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
detected interference pattern produced by quantum distributions
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
.015 .031 .046 .062 .078 .093 .109 .125 .109 .093 .078 .062 .046 .031 .015

Probability distributions of monochromatic EM waves

Tetryonics 98.02 - Boson distributions in monochromatic EM waves

Wave interference patterns

*The double-slit experiment, sometimes called Young's experiment,
is a demonstration that matter and energy can display characteristics of both waves and particles,
and demonstrates the fundamentally probabilistic nature of quantum mechanical phenomena and
Establishes the quantum interference principle known as wave–particle duality.*

Matter particles are stopped by the barrier but the [K]EM wave passes through both slits and is diffracted by them producing weaker EM waves that then superposition with each other to produce interference patterns

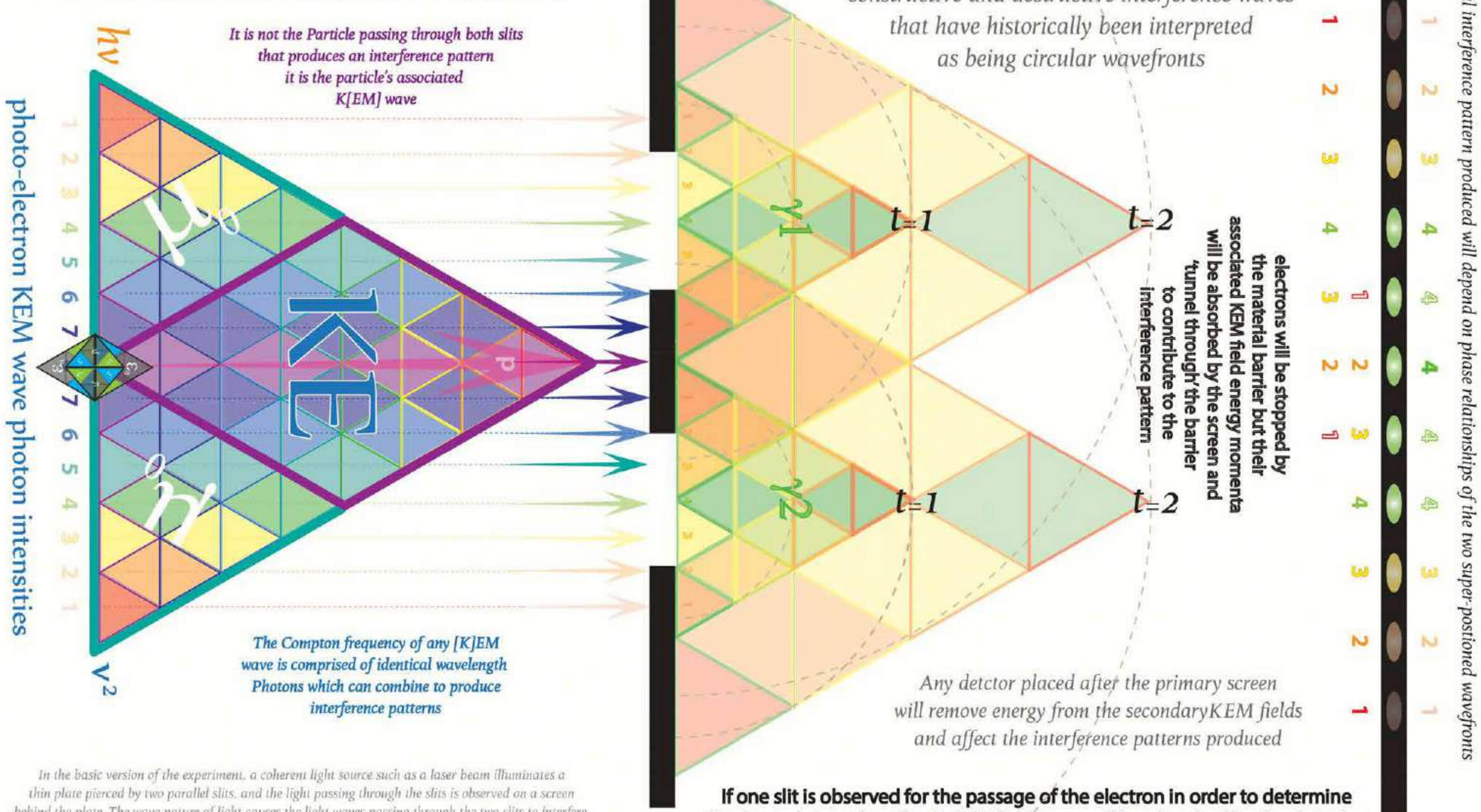

*It is not the Particle passing through both slits
that produces an interference pattern
it is the particle's associated
K[EM] wave*

*Equilateral superpositioned KEM waves produce
constructive and destructive interference waves
that have historically been interpreted
as being circular wavefronts*

**electrons will be stopped by
the material barrier but their
associated KEM field energy momenta
will be absorbed by the screen and
'tunnel through' the barrier
to contribute to the
interference pattern**

final interference pattern produced will depend on phase relationships of the two super-positioned wavefronts

*The Compton frequency of any [K]EM
wave is comprised of identical wavelength
Photons which can combine to produce
interference patterns*

*Any detctor placed after the primary screen
will remove energy from the secondaryKEM fields
and affect the interference patterns produced*

If one slit is observed for the passage of the electron in order to determine its physical state then the its KEM field wave will be absorbed by the detector resulting in only one wave remaining, enforcing a classical particle outcome

In the basic version of the experiment, a coherent light source such as a laser beam illuminates a thin plate pierced by two parallel slits, and the light passing through the slits is observed on a screen behind the plate. The wave nature of light causes the light waves passing through the two slits to interfere, producing bright and dark bands on the screen — a result that would not be expected if light consisted strictly of particles. However, on the screen, the light is always found to be absorbed as though it were composed of discrete particles or photons.

Quantum computing via EM wave super-positioning

BY superpositioning two beams
of EM radiation the resultant 'colours'
will perform quantum level computations
that can be read via the resultant
interference patterns produced

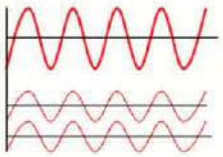

In phase

constructive interference

additive in-phase EM waves

Additive

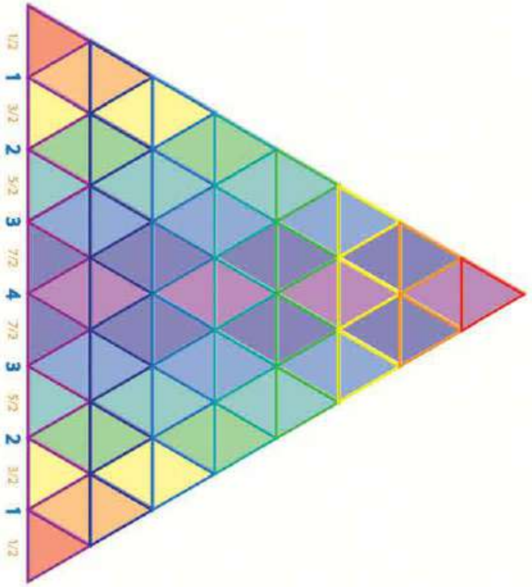

1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

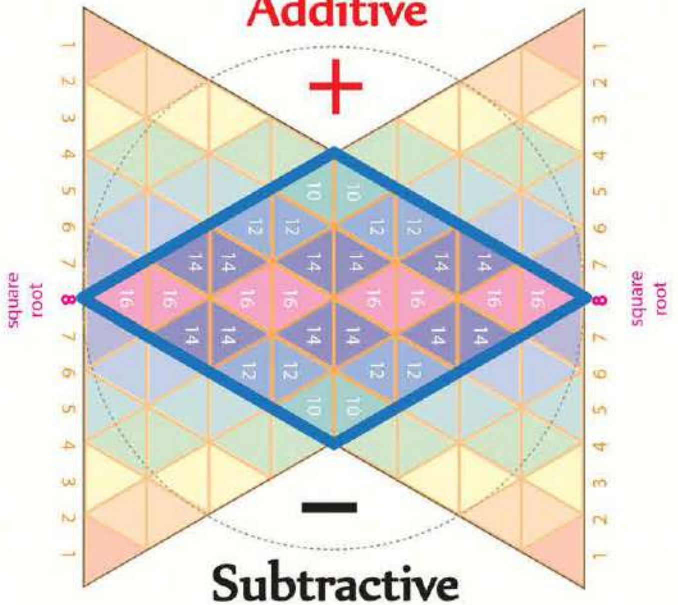

+
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
square root
10
12
14
16
−

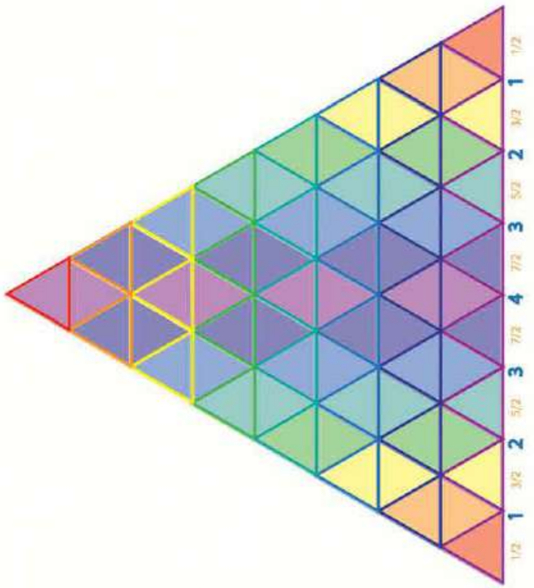

1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Subtractive

Out of phase

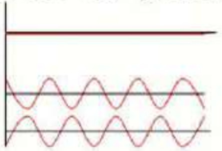

destructive interference

subtractive out-of-phase EM waves

Various basic operations, such as
ADDITION, SUBTRACTION and
SQUARE ROOTS etc
are all easily computed using
EM wave super-positioning

The lines of Force

By utilising the statiscal distribution of equilateral Planck energy momenta quanta in EM waves
Tetryonic theory provides a practical geometric solution to quantum computing problems

q

bosons

P vs. NP

*The set of all decision problems for which an algorithm exists which can be carried out by a deterministic Turing machine in **polynomial** time*

*the set of all decision problems for which an algorithm exists which can be carried out by a **non-deterministic** Turing machine in **polynomial** time*

E

photons

An algorithm of time complexity $O(n)$ is one which increases in time linearly as the "size of the problem" (whatever n stands for) increases.

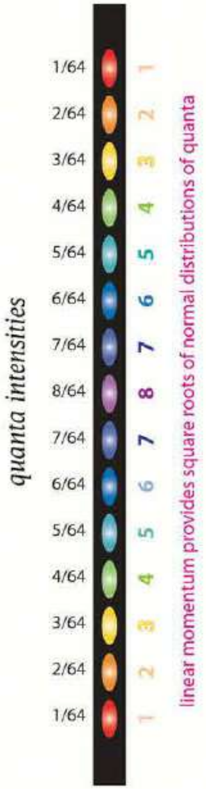
quanta intensities
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
linear momentum provides square roots of normal distributions of quanta

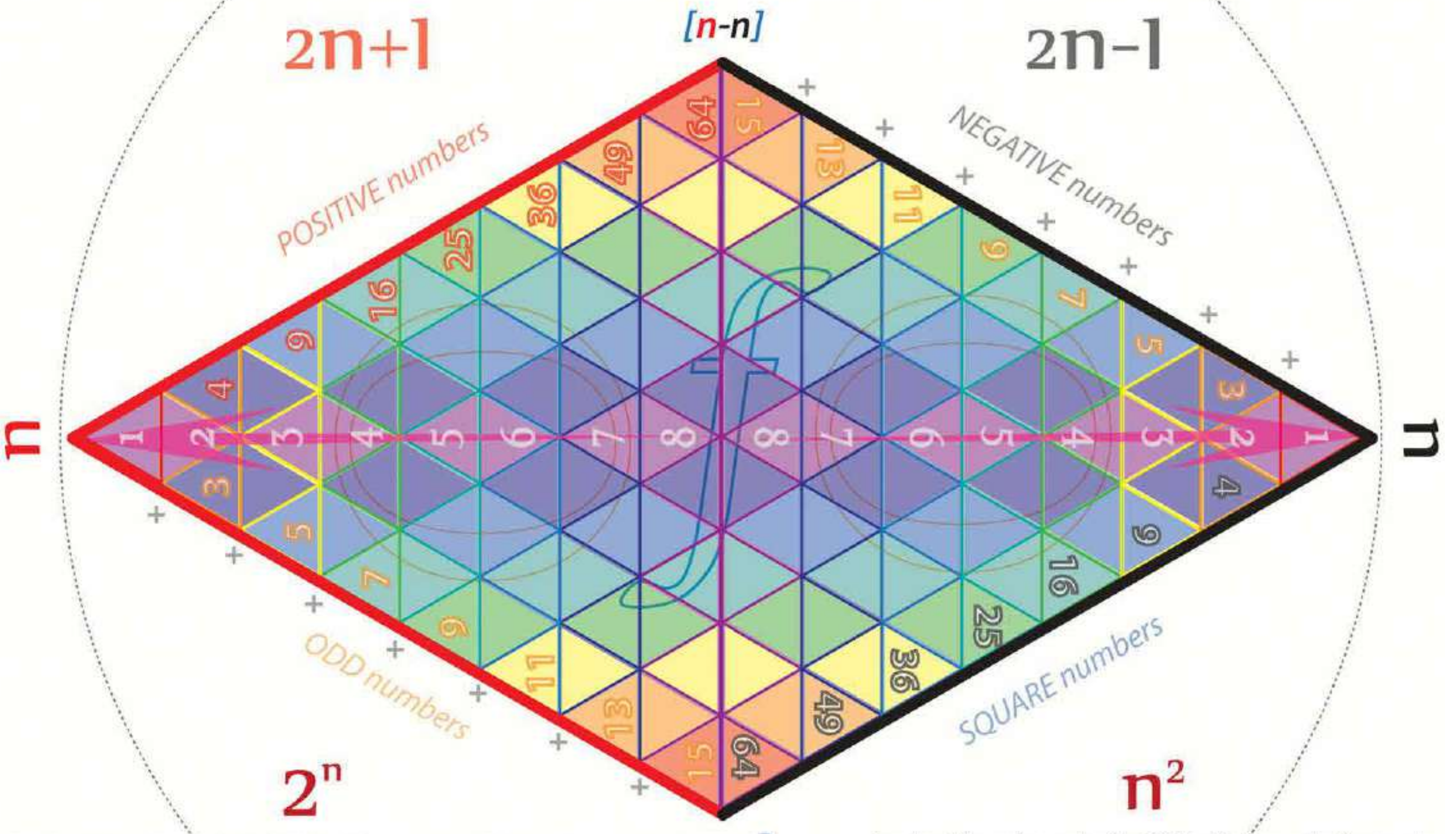
2n+1
[n-n]
2n-1
POSITIVE numbers
NEGATIVE numbers
ODD numbers
SQUARE numbers
n
n
1
2
3
4
5
6
7
8
8
7
6
5
4
3
2
1
64
49
36
25
16
9
4
15
13
11
9
7
5
3
f
+
2^n
n^2

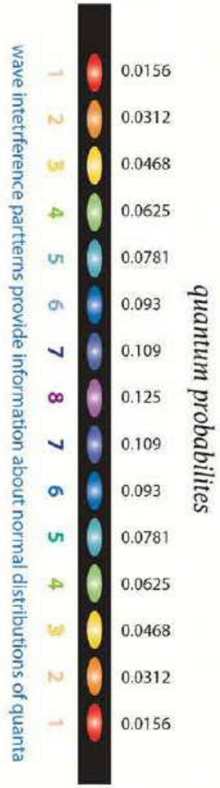
wave interference parterns provide information about normal distributions of quanta
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
0.109
0.093
0.0781
0.0625
0.0468
0.0312
0.0156
quantum probabilites

2^n

An algorithm of complexity $O(2^n)$ utlises exponential quanta; increasing n by 1 will double the quanta required

$2n^2$

n^2

An algorithm of complexity $O(n^2)$ utilises quadratic quanta, meaning that if you double n it will use four times as many quanta.

n^o

Unlike Math treatise on P vs Np that require exponential polynomial time $O(n^k)$ Tetryonic geometry of EM fields utilise exponential energies per second

2hv

1 second

hf

By relating p to the number of equilateral wave fields in any EM wave very large [pn^2] data sets can be modelled and processed every second

Copyright ABRAHAM [2008] - All rights reserved

Quantum Cryptography

If the numerical sequences were applied to amplitude modulation
their non-repeating numerical sequences would appear
to be purely random noise

**Using Tetryonic geometry
advanced non-repeating cyphers
of any complexity
can be easily developed**

The level of encryption can
be easily increased (without limit)
by increasing the dimensions
of the cypher geometry

eg: number of quanta
2D vs 3D geometries

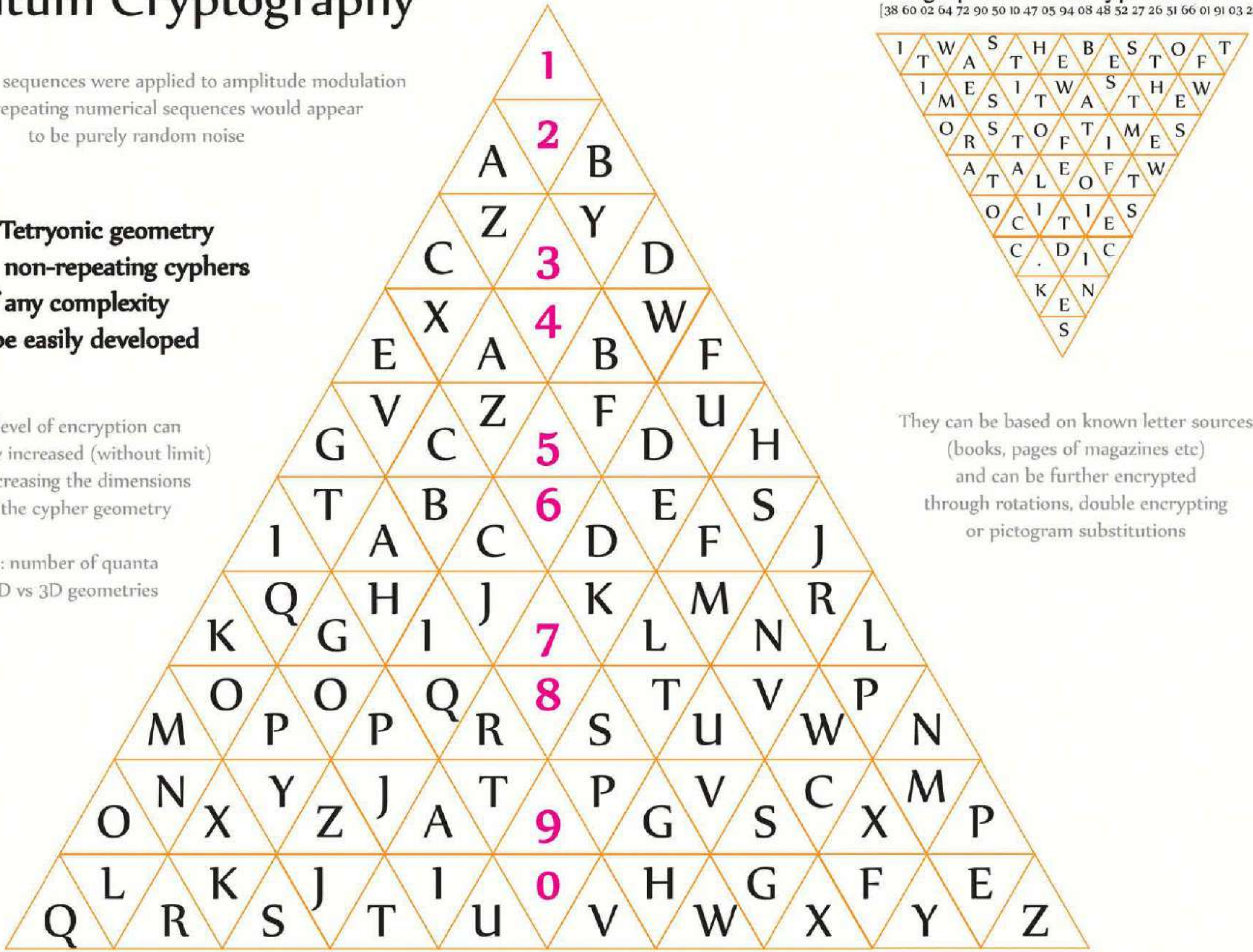
1
A 2 B
C Z 3 Y D
E X A 4 B W F
G V C Z 5 F D U H
I T A B C 6 D E F S J
K Q G H I J 7 K L M N R L
M O P O P Q R 8 S T U V W P N
O N X Y Z J A T 9 P G V S C X M P
Q L R K S J T I U 0 V H W G X F Y E Z

eg: quantum encryption 1024
[38 60 02 64 72 90 50 10 47 05 94 08 48 52 27 26 51 66 01 91 03 20]

I T W A S T H E B E S T O F T
I M E S I T W A S T H E W
O R S T O F T I M E S
A T A L E O F T W
O C I T I E S
C . D I C
K E N
S

They can be based on known letter sources
(books, pages of magazines etc)
and can be further encrypted
through rotations, double encrypting
or pictogram substitutions

Quantum Computing

The Proton/Neutron geometries of atomic nuclei can be built at the quantum scale
to create an atomic nuclei that can operate as a Opto-memory-transistive computing element,
many elements can then be combined in lattices to create super computers no larger than bacterium

Spin UP

Energy can be gated through individual nuclei
using the centre Baryon as the base transistor element,
in turn effecting the energies of bound photo-electrons

Spin DOWN

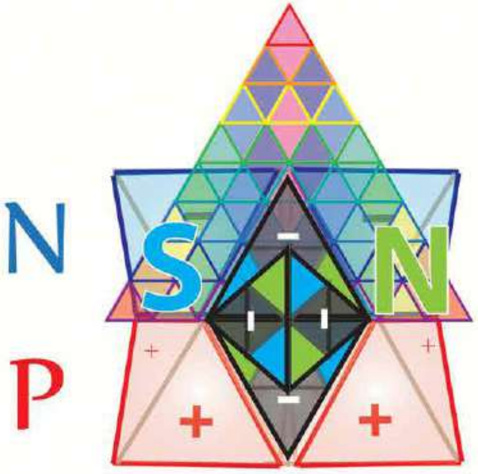

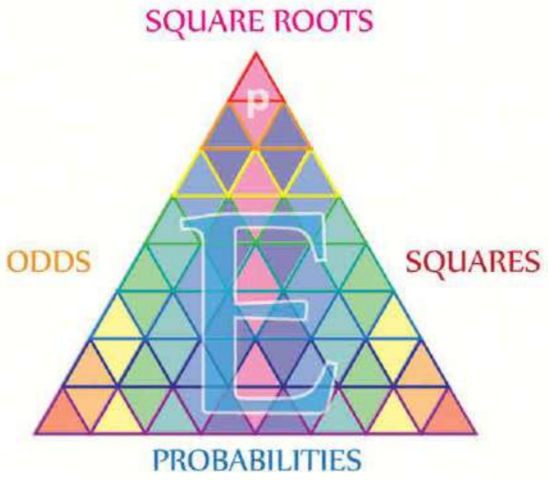

Q-bits

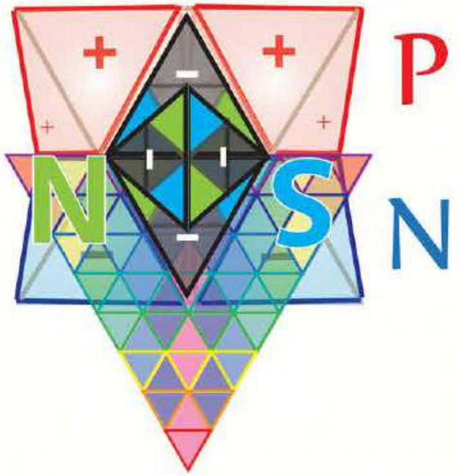

1

photo-electronic transitions can be used to directly recieve
or emit memory states through the absorption and emission
of spectral photons of specific energy momenta

0

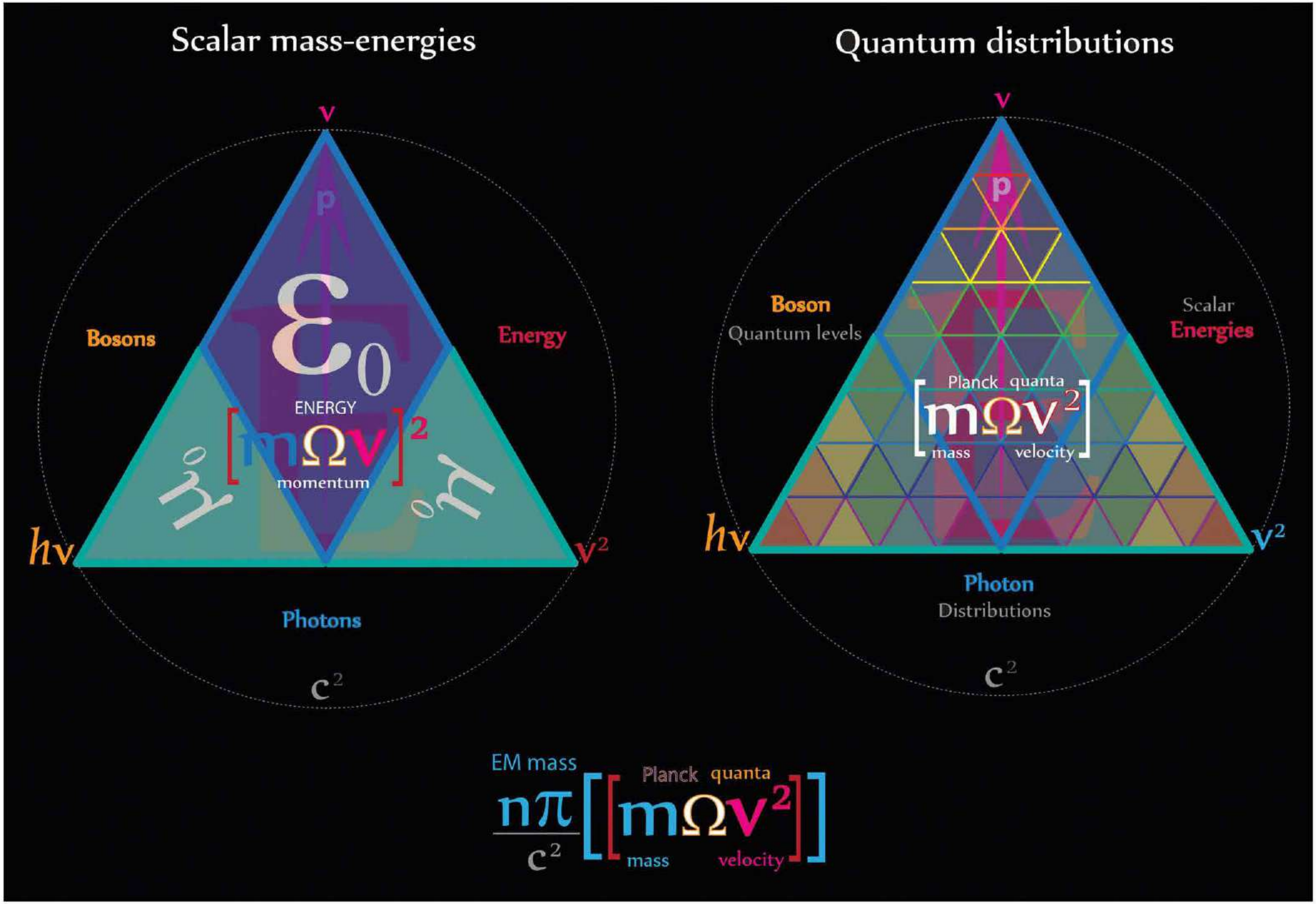
Scalar mass-energies
v
p
$\mathcal{E}_0$
Bosons
Energy
ENERGY
$[m\Omega v]^2$
momentum
μ_0
ε_0
hv
v^2
Photons
c^2
Quantum distributions
v
p
Boson
Quantum levels
Scalar
Energies
Planck quanta
$[m\Omega v^2]$
mass
velocity
hv
v^2
Photon
Distributions
c^2
EM mass
Planck quanta
$\frac{n\pi}{c^2}[[m\Omega v^2]]$
mass
velocity

2D mass

$$\overset{\text{EM mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

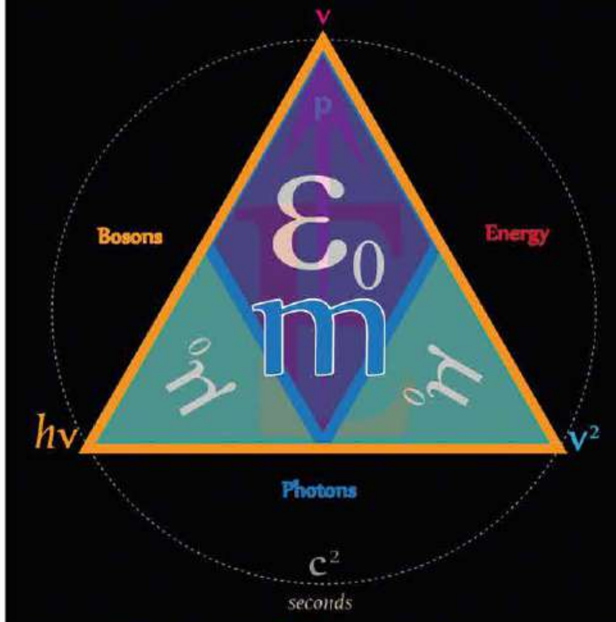
v
p
ε0
m
Bosons
Energy
hv
v²
Photons
c²
seconds

Measure of planar energies
in a nπ radiant wave
geometry
[per second]

mass quantum
7.376238634 e-51 kg

ENERGY

$$\overset{\text{ENERGY}}{n\pi}\left[\left[\underset{\text{momenta}}{m\Omega v^2}\right]\right]$$

Equilateral energy momenta
has a mass equivalence
per unit of Time

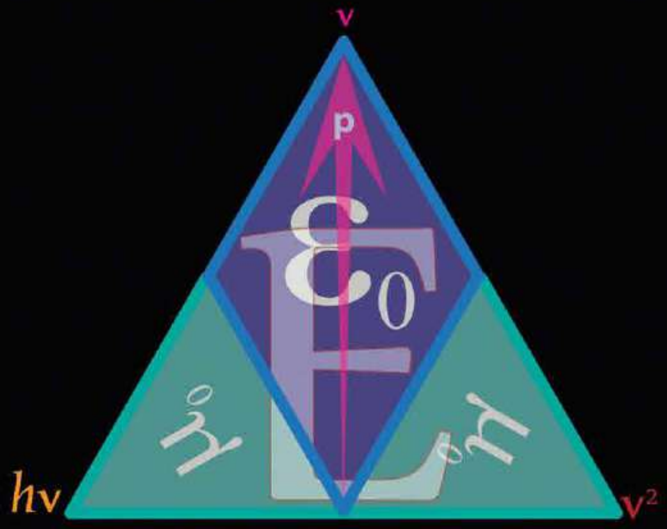

c^2 is the radial geometry
created by radiant energy
in 1 second

Planck Constant
6.629432672 e-34 J.s

3D Matter

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Measure of mass-energy
in a 4nπ standing wave
topology
[per second squared]

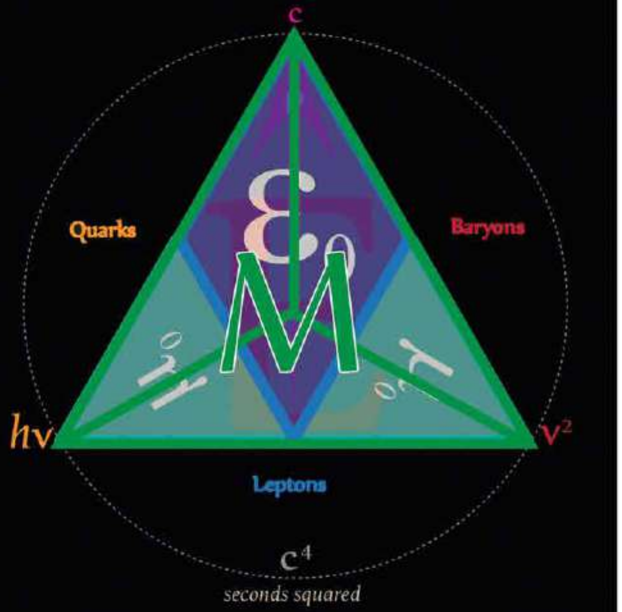
c
ε0
M
Quarks
Baryons
hv
v²
Leptons
c⁴
seconds squared

Matter quantum
2.9504955454 e-50 kg

Planck's constant - the quantum of Action

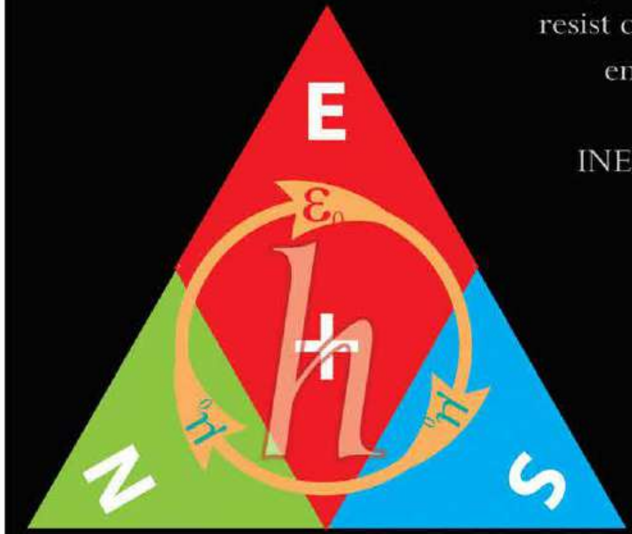
E
ε_0
$\hbar$
μ_0
μ_0
N
S

all ideal quantum inductive loops
resist changes to their
energy levels

INERTIAL mass

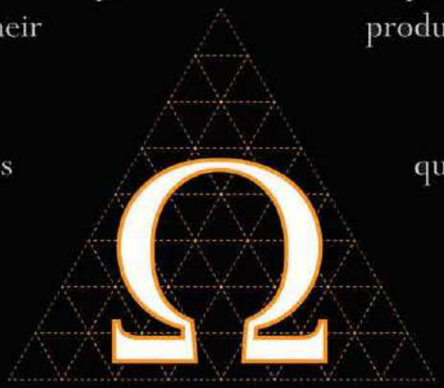

all equilateral energy momenta
produce square root vector
Forces

quantum of WORK

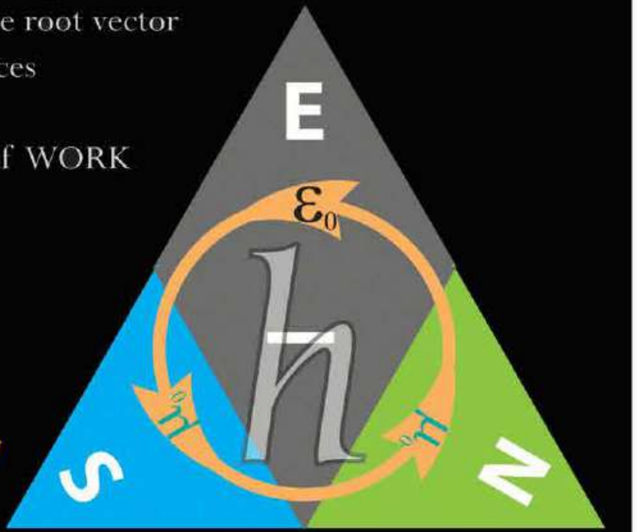
E
ε_0
h
μ_0
μ_0
S
N

Quantised angular momentum is equilateral geometry

The two sides of Planck's Constant

The square root of all Energy is linear momentum

Equilateral
energy quanta
per second

mass

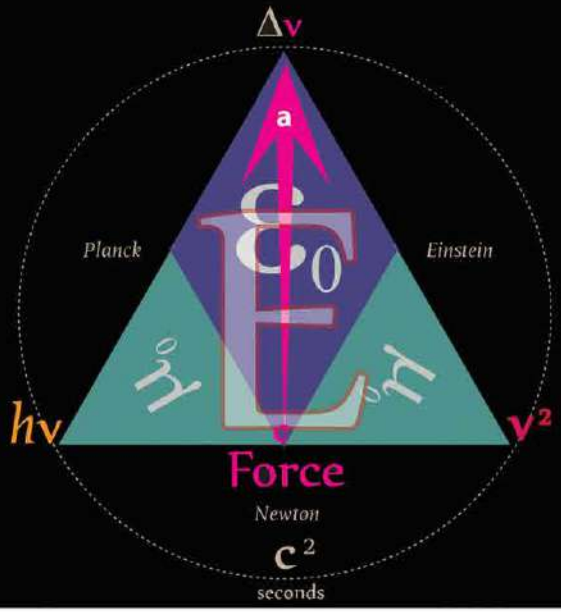
Δv
a
ε_0
E
μ_0
μ_0
Planck
Einstein
hv
v^2
Force
Newton
c^2
seconds

Quantised
angular momenta
per second

time

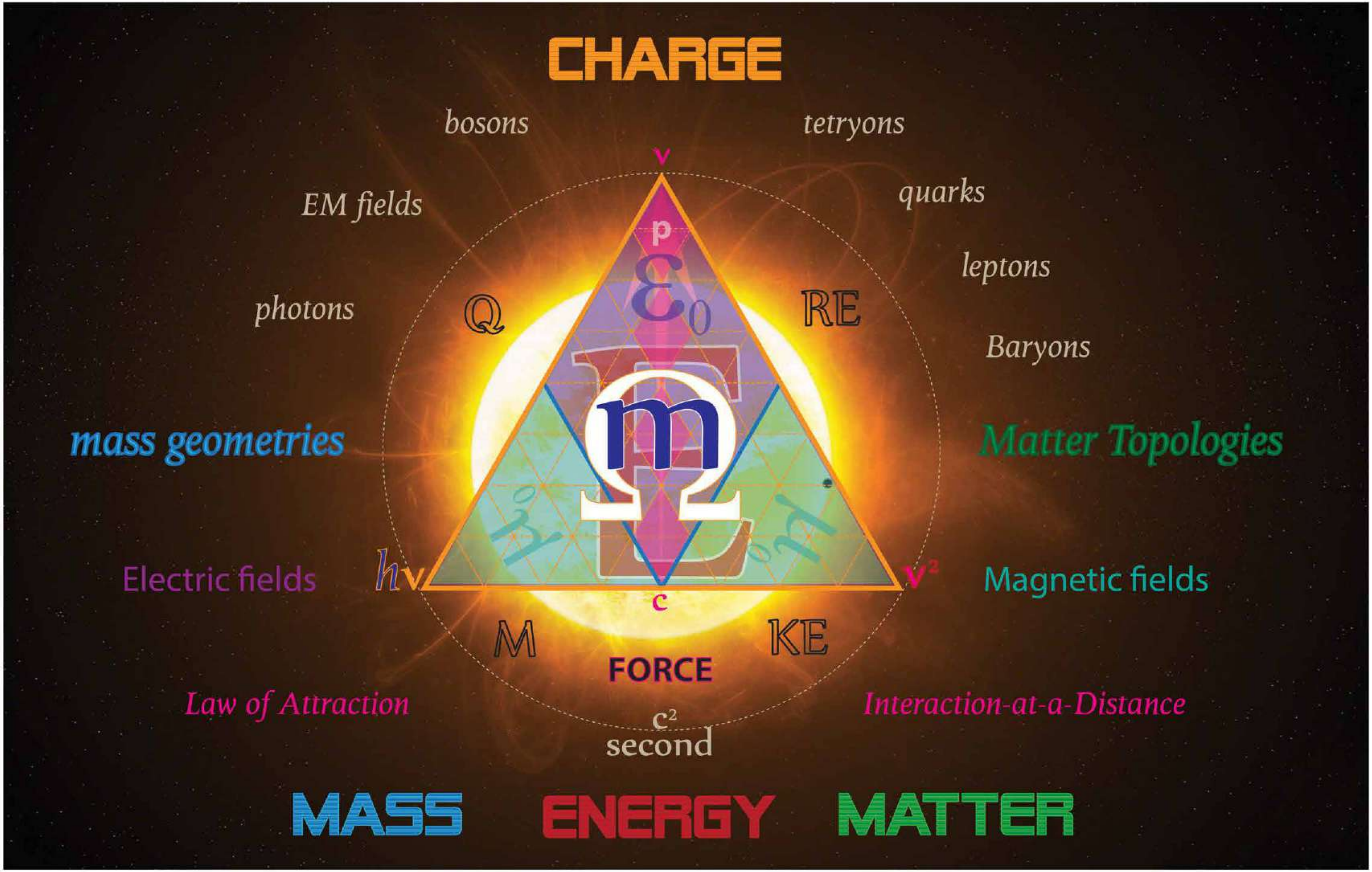
CHARGE
bosons
tetryons
v
EM fields
quarks
p
leptons
ε0
photons
Q
RE
Baryons
E
m
Ω
mass geometries
Matter Topologies
Electric fields
hv
v²
Magnetic fields
c
M
KE
FORCE
Law of Attraction
Interaction-at-a-Distance
c²
second
MASS
ENERGY
MATTER

Tetryonics 99.04 - Charged Planck mass-ENERGY-Matter

Copyright ABRAHAM [2008] - All rights reserved

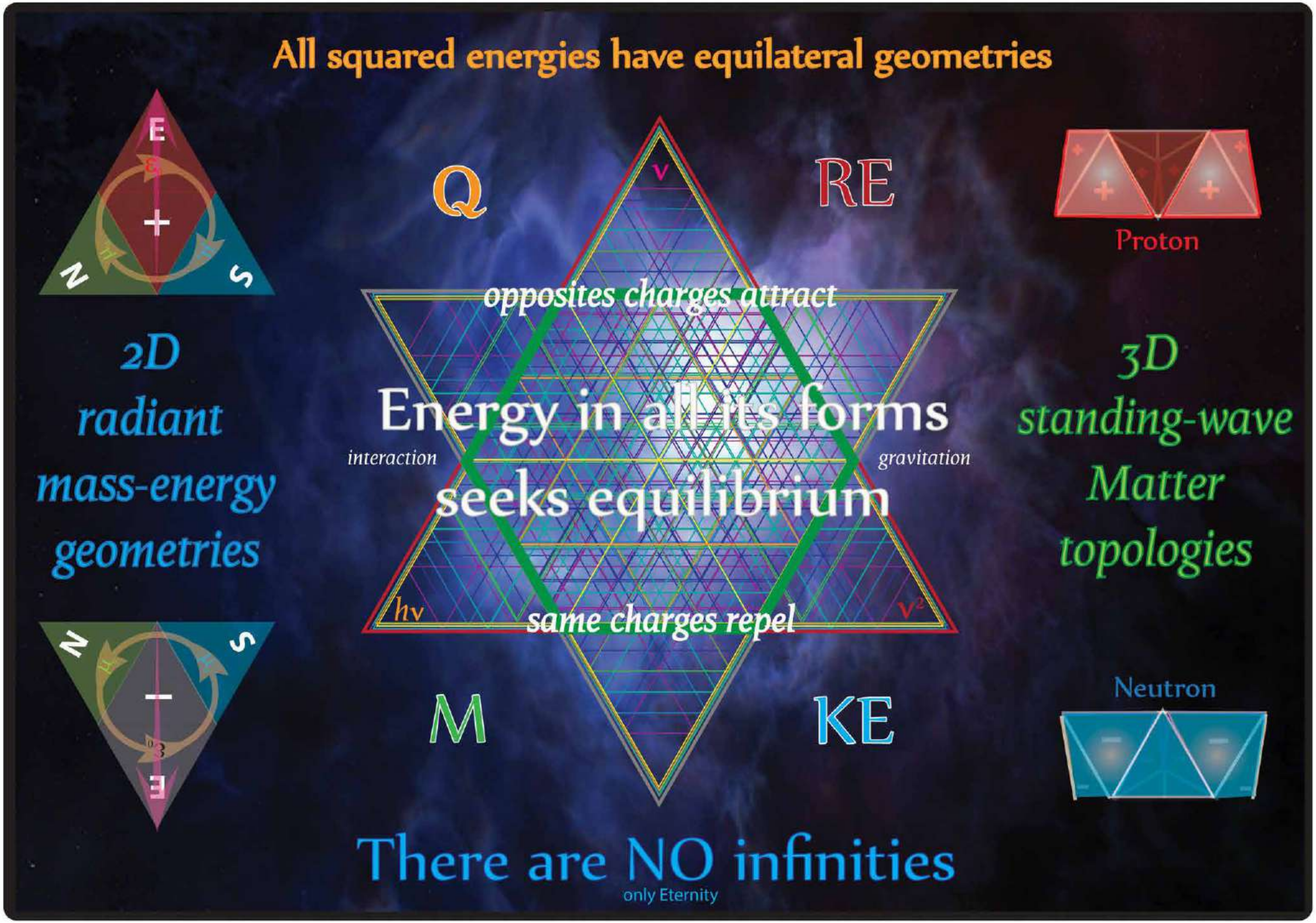

All squared energies have equilateral geometries

E
+
N
S
Q
RE
v
Proton
interaction
gravitation
hv
v²
N
S
−
E
M
KE
Neutron

opposites charges attract

Energy in all its forms seeks equilibrium

same charges repel

2D radiant mass-energy geometries

3D standing-wave Matter topologies

There are NO infinities

only Eternity

Copyright ABRAHAM [2008] - All rights reserved

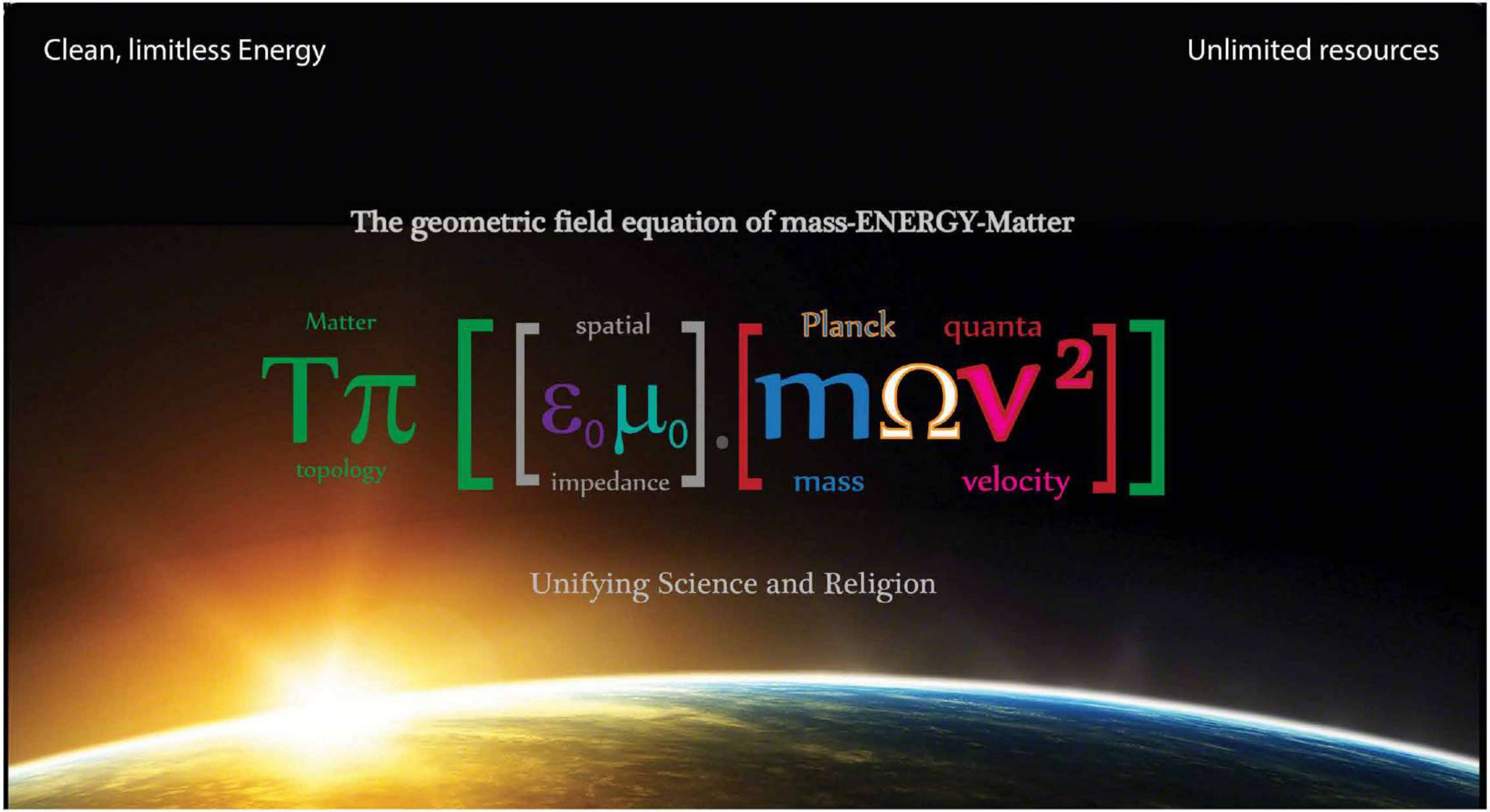

Clean, limitless Energy

Unlimited resources

The geometric field equation of mass-ENERGY-Matter

Matter spatial Planck quanta

$$T\pi \left[\left[\varepsilon_0 \mu_0 \right] \cdot \left[m \Omega v^2 \right] \right]$$

topology impedance mass velocity

Unifying Science and Religion

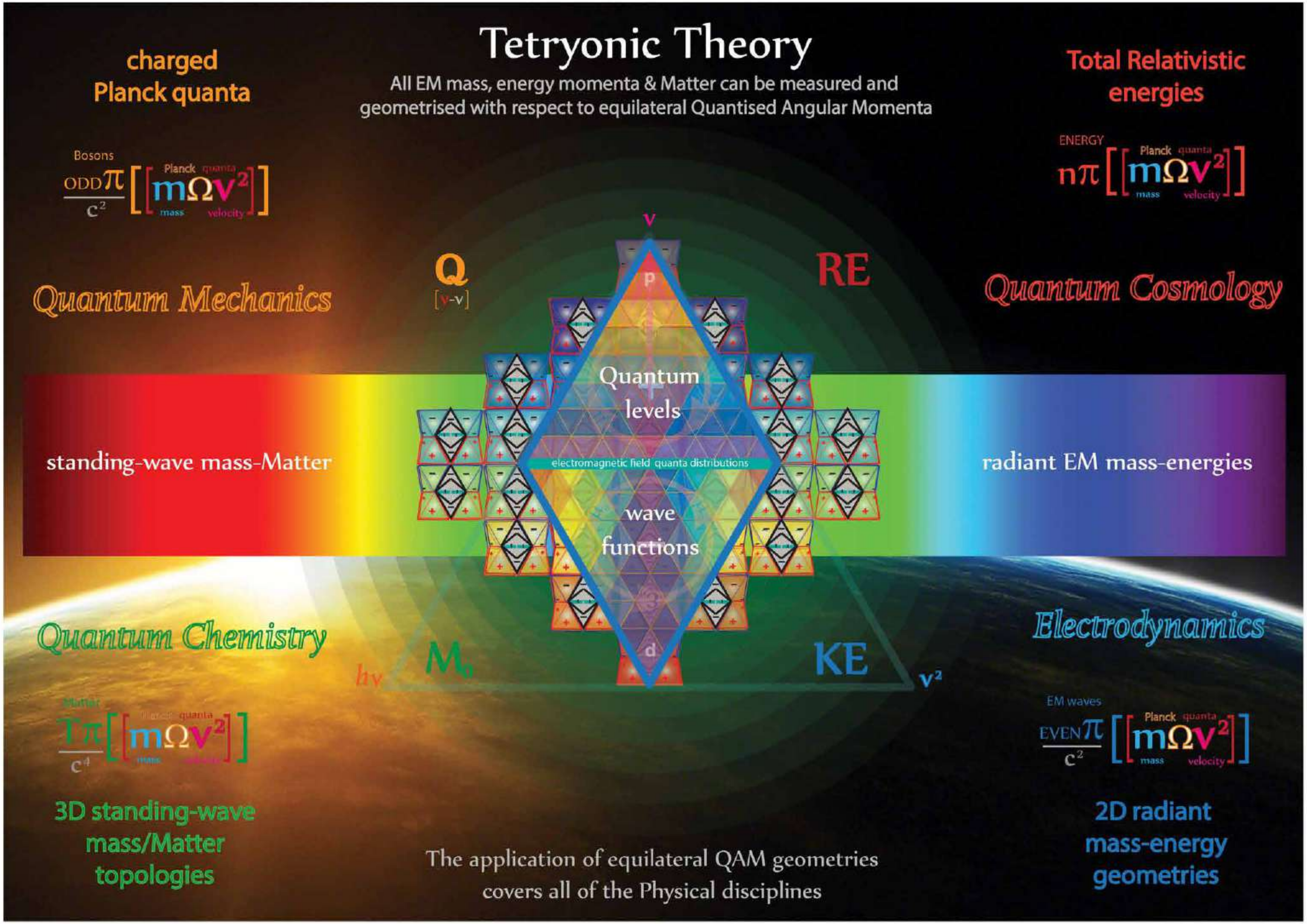

Tetryonic Theory

All EM mass, energy momenta & Matter can be measured and geometrised with respect to equilateral Quantised Angular Momenta

charged Planck quanta

Bosons

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

Total Relativistic energies

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

Quantum Mechanics

Quantum Cosmology

standing-wave mass-Matter

radiant EM mass-energies

Quantum Chemistry

Electrodynamics

Matter

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

3D standing-wave mass/Matter topologies

2D radiant mass-energy geometries

The application of equilateral QAM geometries covers all of the Physical disciplines

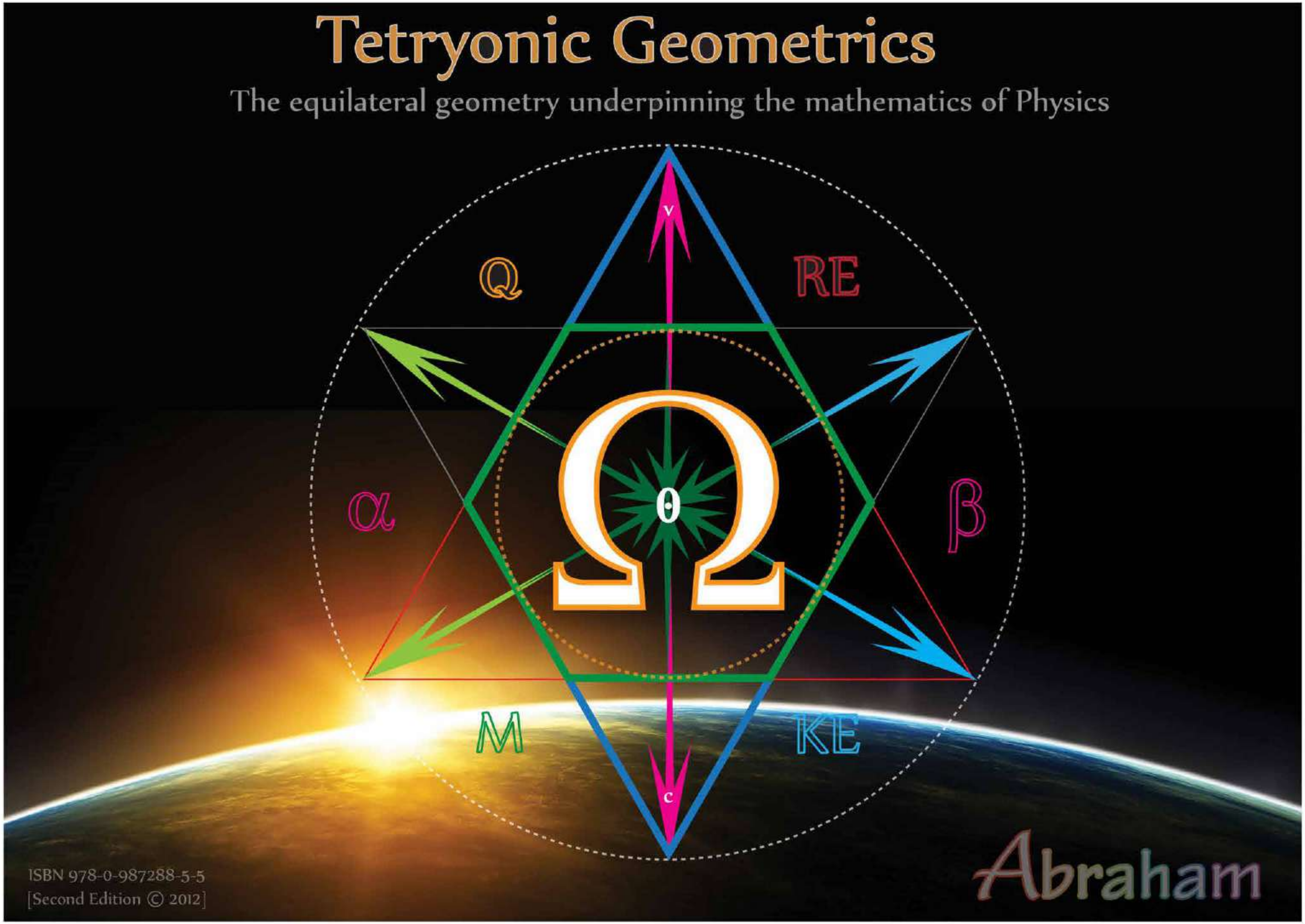

Tetryonic Geometrics

The equilateral geometry underpinning the mathematics of Physics

ISBN 978-0-987288-5-5
[Second Edition © 2012]

Abraham

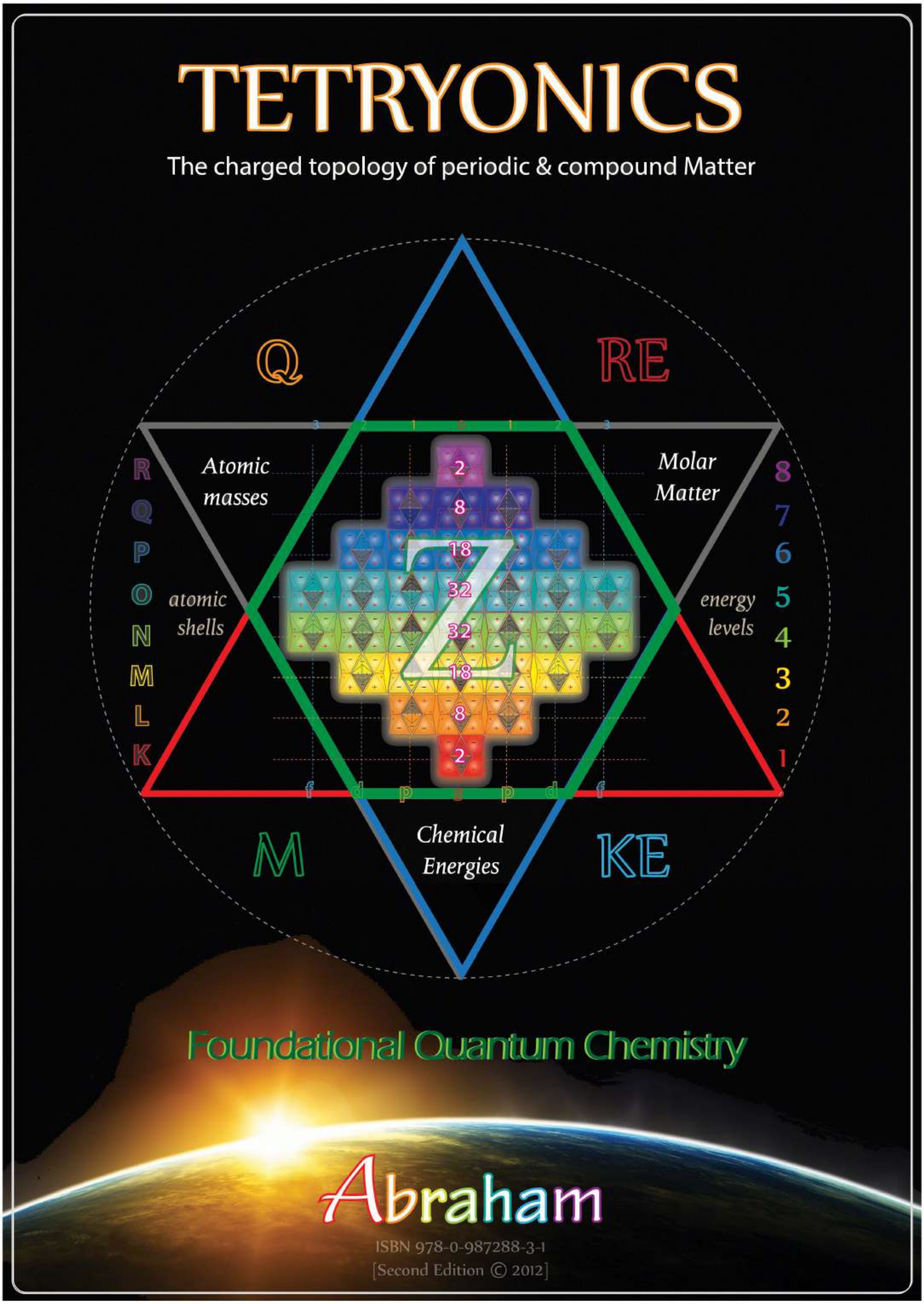

TETRYONICS

The charged topology of periodic & compound Matter

Q
RE
M
KE
Atomic masses
Molar Matter
atomic shells
energy levels
Chemical Energies
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
3
2
1
0
1
2
3
f
d
p
s
p
d
f

Foundational Quantum Chemistry

Abraham

ISBN 978-0-987288-3-1
[Second Edition © 2012]

Copyright ABRAHAM [2008] - All rights reserved

Protons **+** *Neutrons*

u+d+u d+u+d

All electrons, quarks and Baryons are made up of 4n+ (Tetryonic) standing-wave EM fields.

As well as having nett Tetryonic charged topologies ranging between [+24] ~ [-12] they all posses distinct ELECTRIC FIELDS that are concentrated in 3 apex points as indicated in the illustrations

These points result from the orientation of Electric apexes and orthagonal Magnetic dipole field edges that make up each particle's externalised EM fields.

The Positive and Negative electric apex points, obey the Law of Interaction forcing separated nuceli to combine due to their individual nett Tetryonic charges and provide a means of orienting nuclei to each other to create larger particles [elements, allotropes and compounds]

External Magnetic (H) fields can interact with the integral magnetic (B) dipoles of Tetryonic particles forcing them to orientate in specific directions to facilite chemical bonding [nuclear forces]

Additionally, external Electric fields can interact with the integral electric fields attracting or repelling them depending on the polarity of the external electric field [Electrostatics]

External energies can be induced into these integral EM fields via inductive coupling or the absorption of spectral photons in turn leading to an increase in the strengthes of the integral EM apexes in turn increasing the Strong Nuclear Force.

Residual Electro-Magnetic Forces

allow Neutrons and Protons to attract via the opposite Electric charge points created by their constituent Quarks in order to create Elementary Nuclei

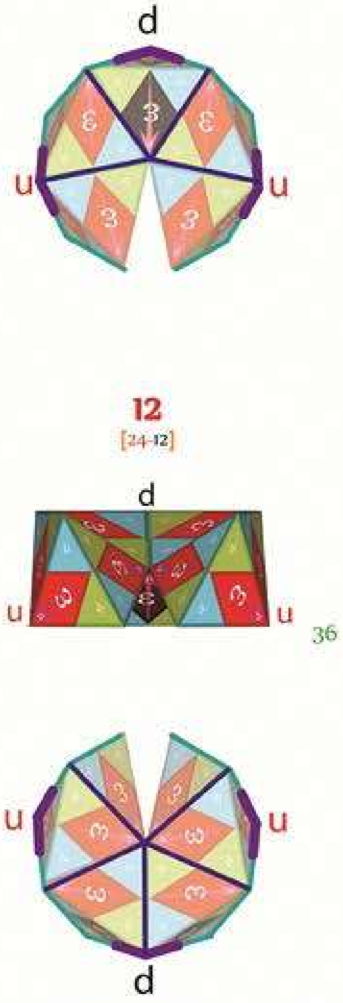

The orientation of the component Electric fields within 3D Matter creates macroscopic force apexes via externalised 'E-points'

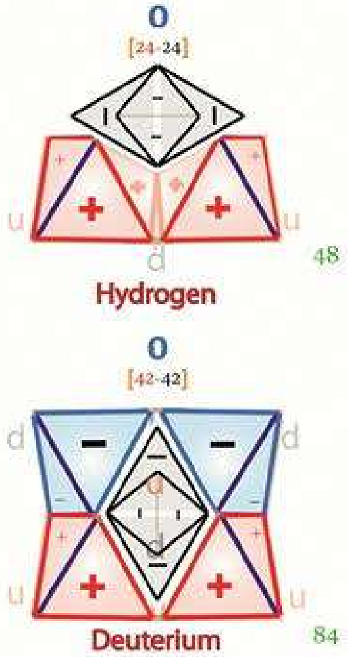

The Strong Nuclear force binds Matter together

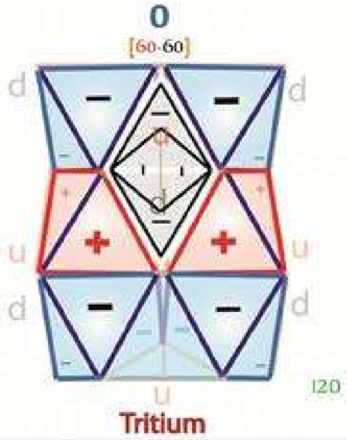

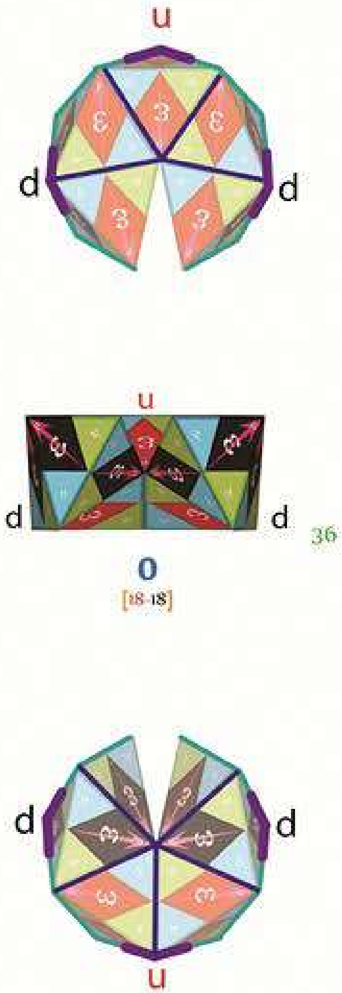

The orientation of the component Magnetic fields within 3D Matter creates macroscopic force apexes via externalised 'M-dipoles'

Copyright ABRAHAM [2008] - All rights reserved

Nucleonic residual EM force

The attraction between Baryonic E&M field apexs, a result of their component Quark arrangements, results in the formation of heavier and more complex Nuclei

The residual e-field apexes and m-field dipoles form two rings of residual EM fields around the circumference of atomic nuclei

Neutrons

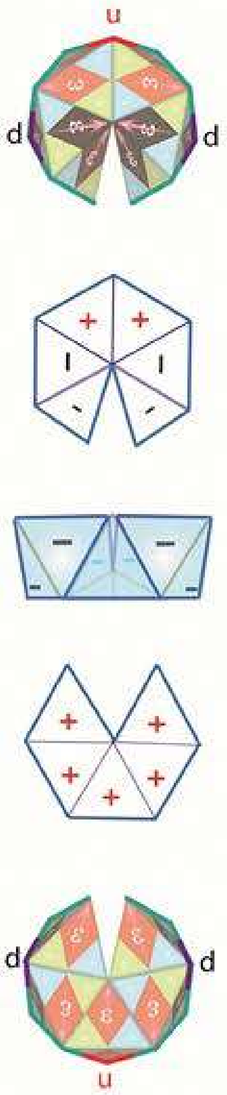

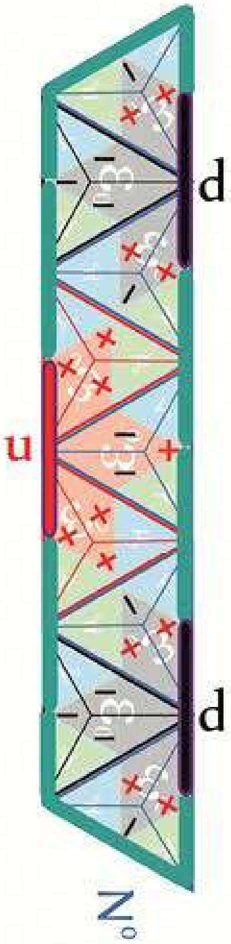

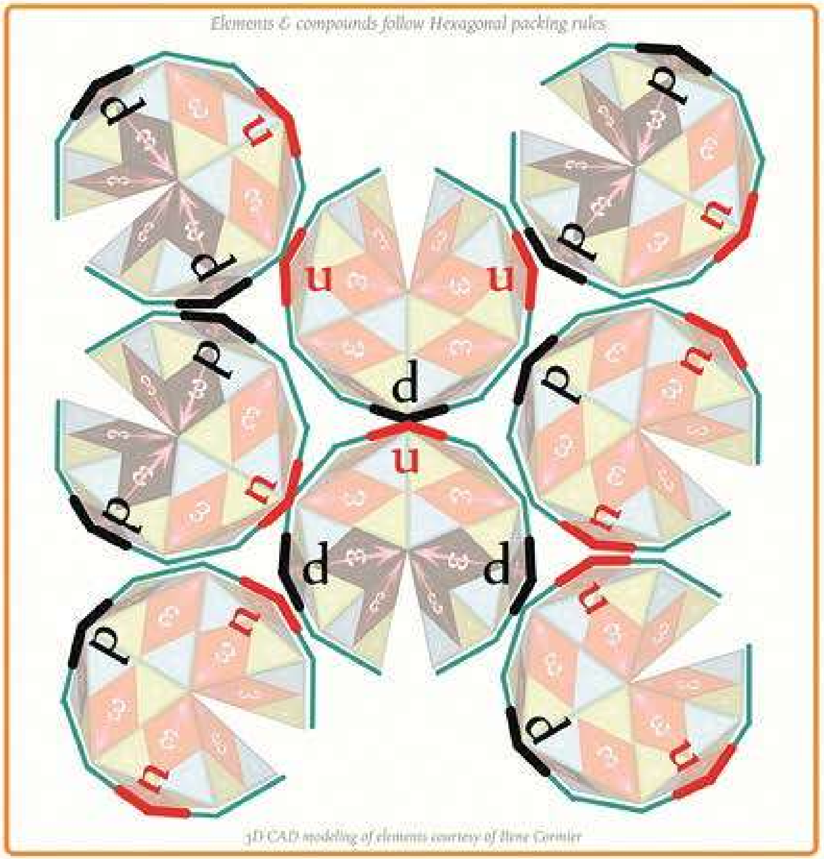

Elements & compounds follow Hexagonal packing rules

3D CAD modeling of elements courtesy of Ilene Cormier

Protons

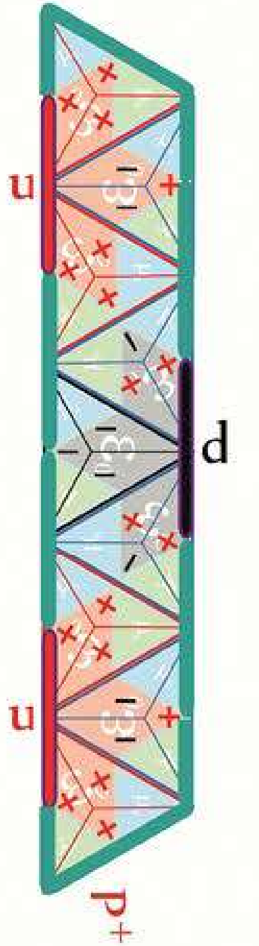

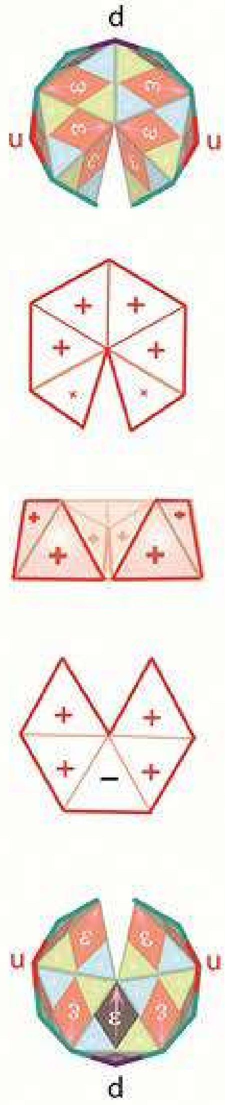

E-field apexes and their polarities highlight the quark alignment of all atomic nuclei and elements

UP Quark
Positive Electric field apex

DOWN Quark
Negative Electric field apex

Copyright ABRAHAM [2008] - All rights reserved

Insulators and Conductors

The position of electrons in Nuclei within Atomic Elements
results in the properties of Insulators or Conductors

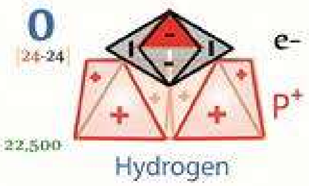
0
[24-24]
e-
P+
22,500
Hydrogen

Coloumbic forces

Electrons are attracted to the residual EM net (+12) positive charge of Protons or n[+12] unbalanced Ionic charges of nuclei

Conductor

Electrical energies move around
the material via boson exchanges
and electron movement

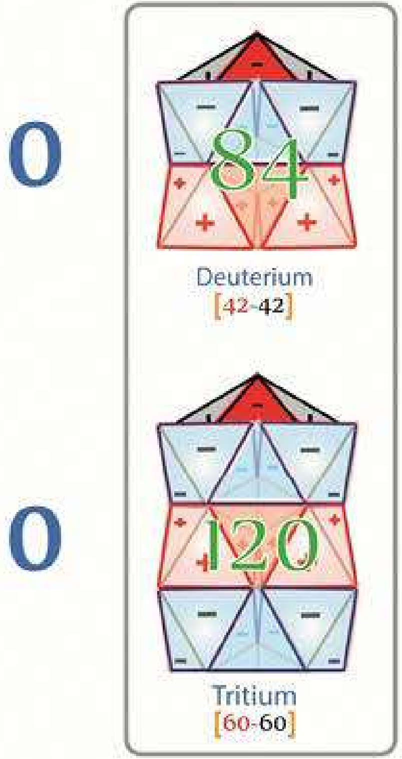
0
84
Deuterium
[42-42]
0
120
Tritium
[60-60]

≈

Insulator

Electrical energy is fixed within the nucleus
as electrostatic charges & released upon demand
via electron rotation/motion within the nucleus

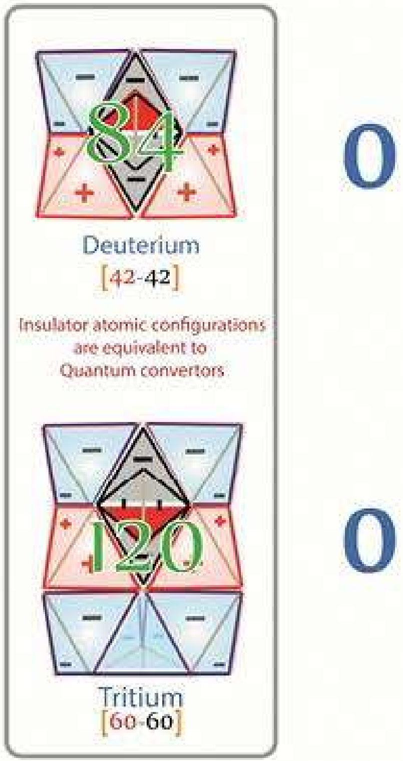
84
Deuterium
[42-42]
0
120
Tritium
[60-60]
0

Insulator atomic configurations
are equivalent to
Quantum convertors

Ions

Charge (energy) is moved around
material via electron movement

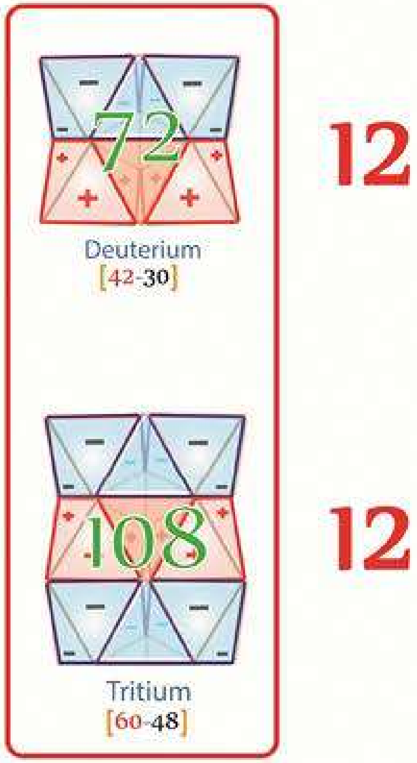
72
Deuterium
[42-30]
12
108
Tritium
[60-48]
12

**Conductive materials contain 'free' electrons that
can be readily or easily moved within the material**

**Insulator materials have electrons that are 'bound' tightly to the atoms
and store charges locally where they are applied**

**Materials that have been ionised are
more likely to become Conductors
as they easily attract and bind
free electrons to them**

Copyright ABRAHAM [2008] - All rights reserved

Nucleon Quark Arrangement

The nuclei arrangement of each atomic shell [quantum level] is the result of quark EM field interactions

Quantum levels

Atomic shells

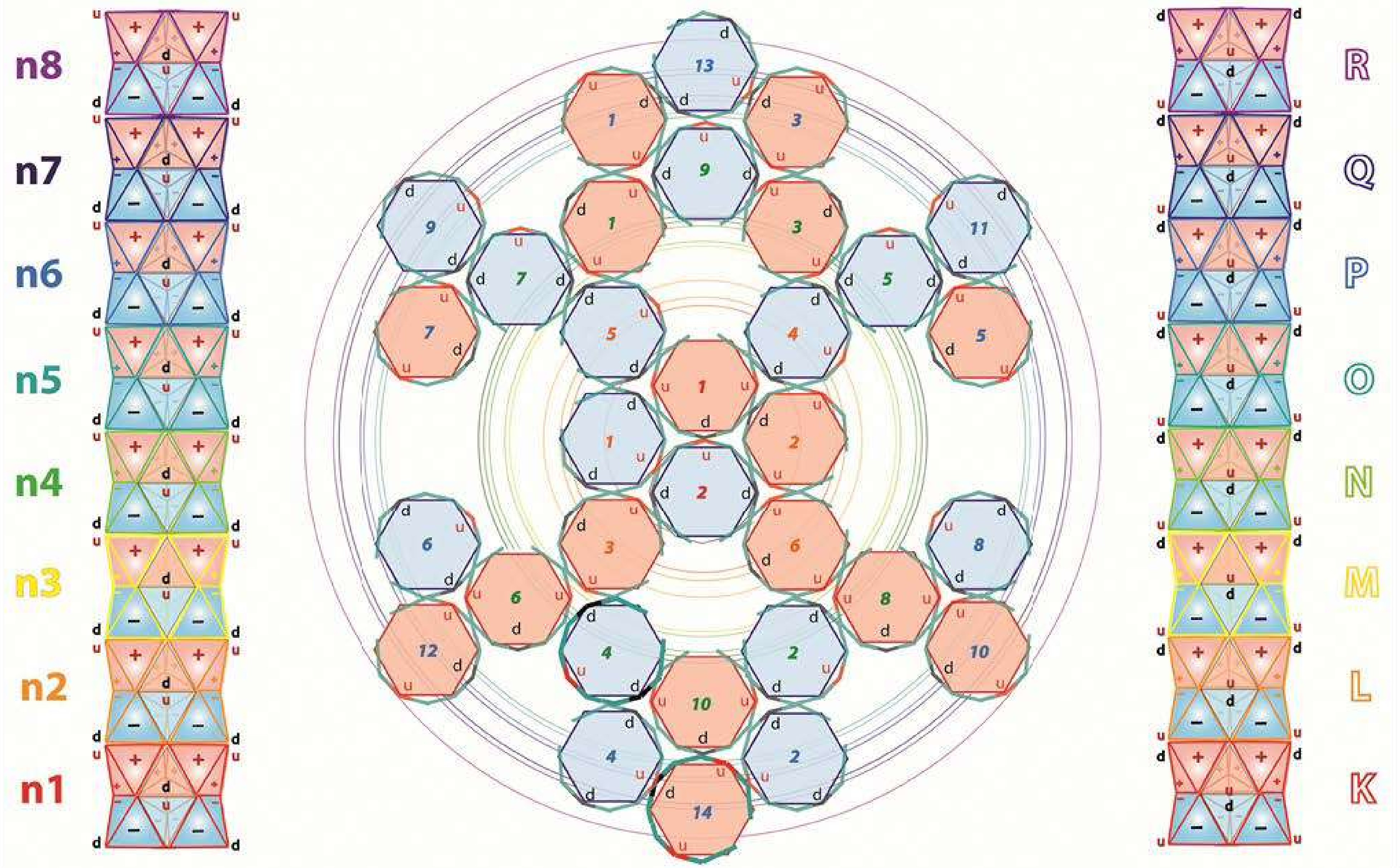
n8
n7
n6
n5
n4
n3
n2
n1
R
Q
P
O
N
M
L
K
u
d
+
−
13
1
3
9
1
3
9
7
5
11
7
5
4
5
1
1
2
2
6
3
6
8
6
8
12
4
2
10
10
4
2
14

Nucleon Charges and Bonding

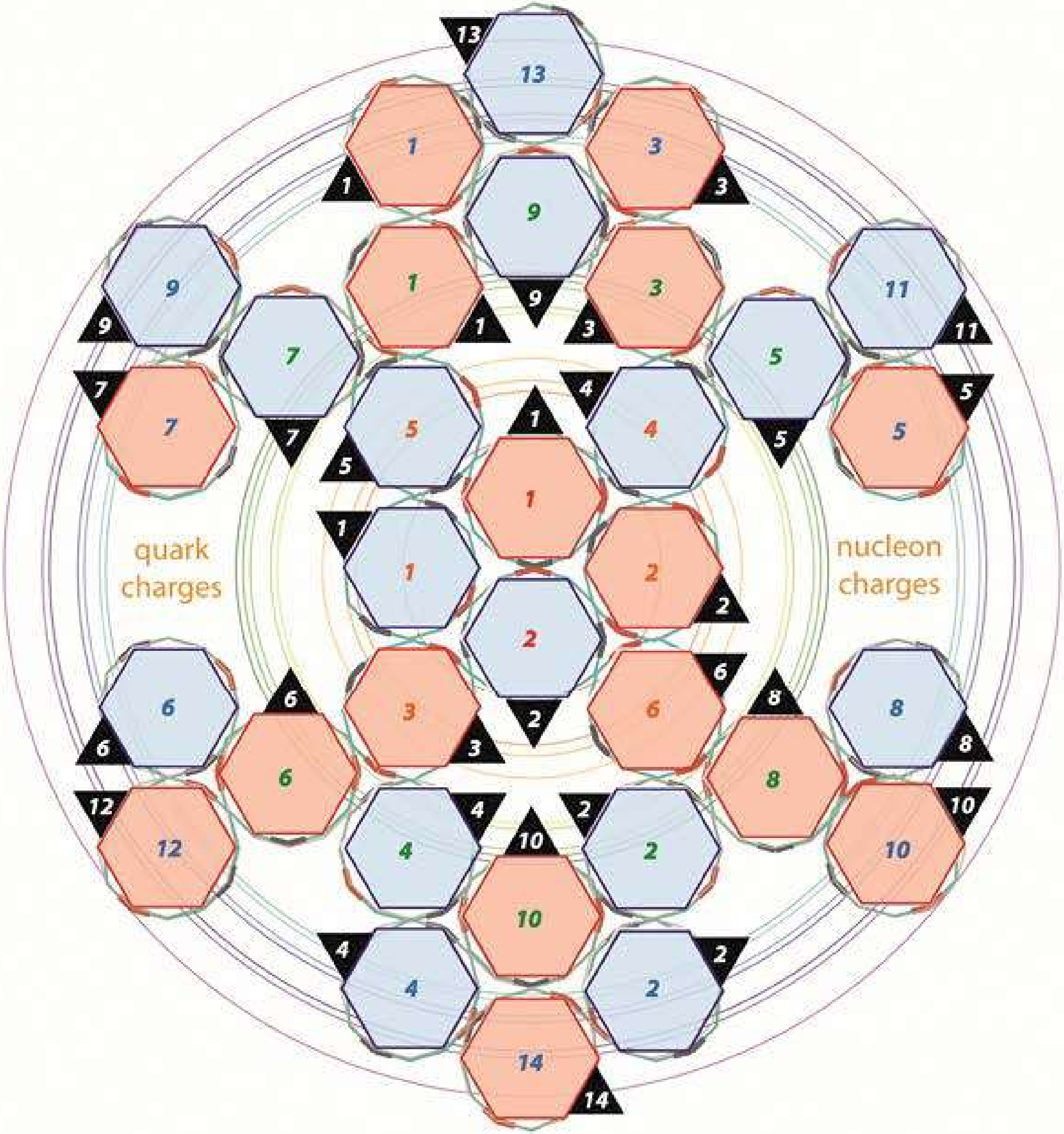
quark charges
nucleon charges

(Strong force - topological Electric Points)

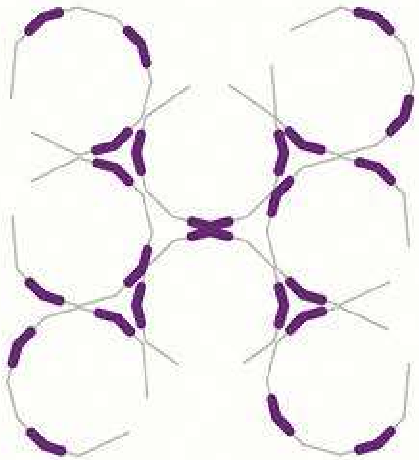

residual EM forces

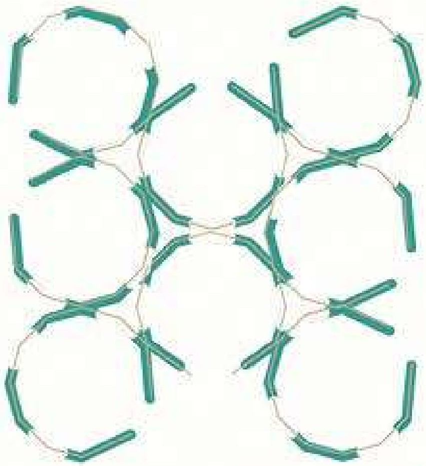

(Strong force - topological Magnetic dipoles)

Bound electron arrangements

Externally bound electrons produce sub-orbital patterns
different to the electron orbitals of internally bound electrons

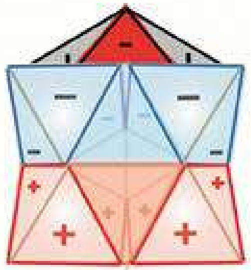

Deuterium

electrons are externally bound
to the Deuteron nuclei

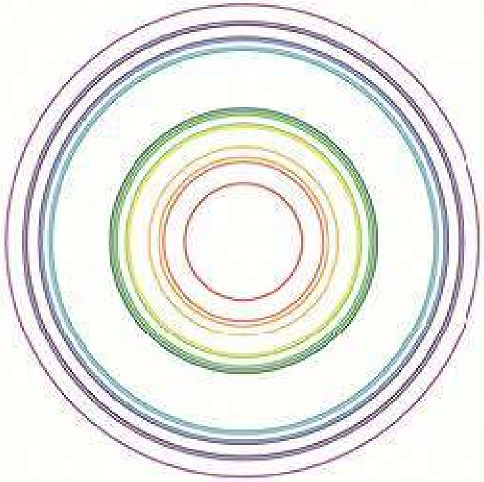

Conductors

Charges are free to move and equalise

electrons require less energy
to 'break free' from Nuclei

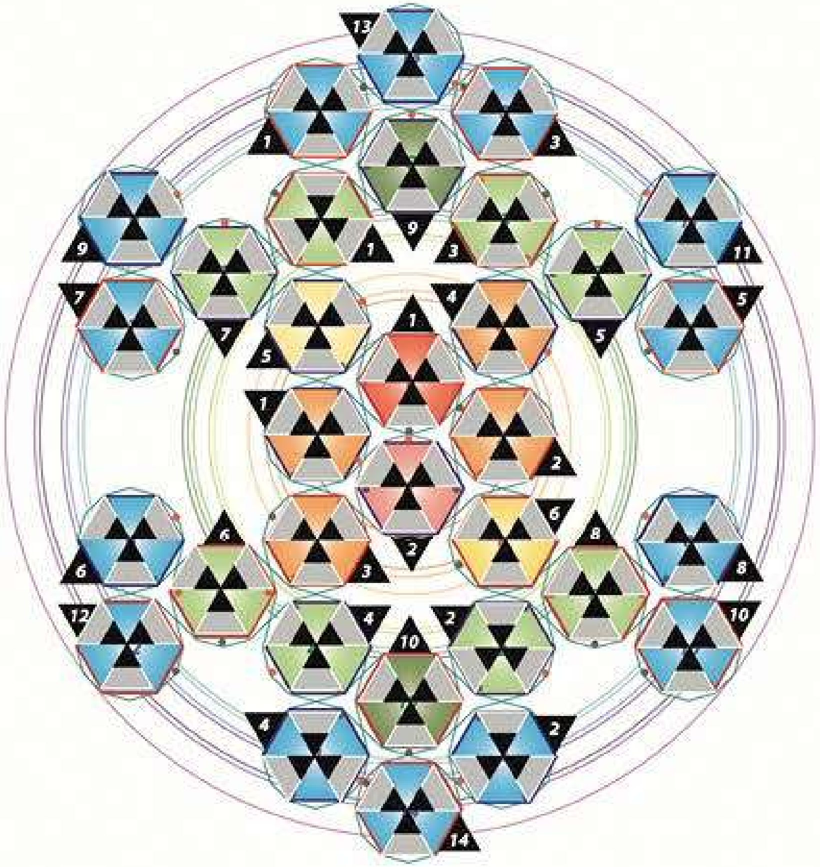

The electron orbitals of
conductors are lower energies
than those of insulators

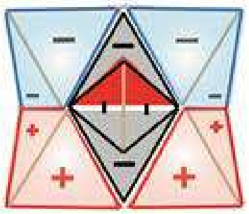

Deuterium

electrons are internally bound
in the Deuteron nuclei

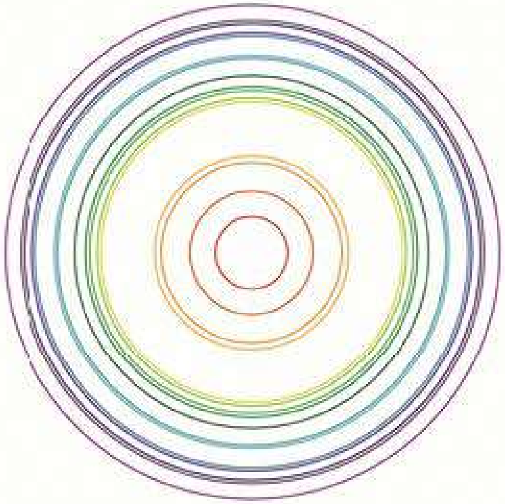

Insulators

Charges are bound to specific locations

electrons require more energy
to 'break free' from Nuclei

2D mass-energy geometries form the fabric of 3D Matter topologies

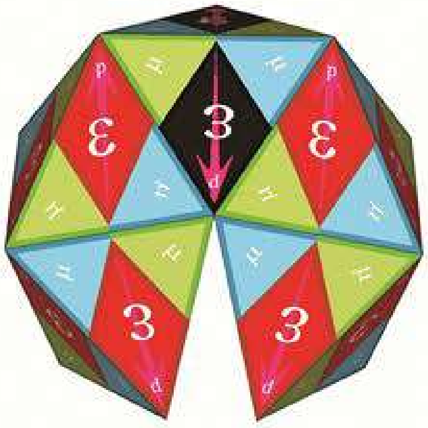

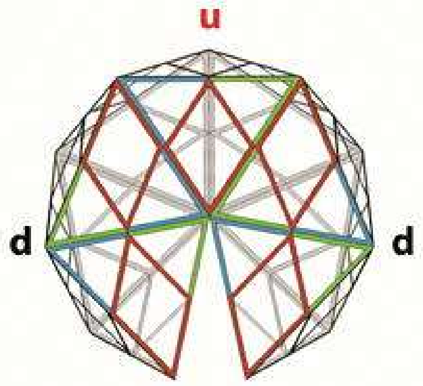

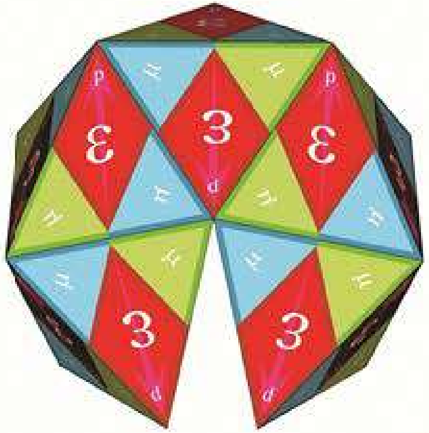

Proton

12
[24-12]
nett Charge
component charges

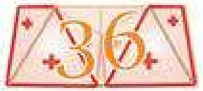

charged mass-energy geometry

20π
Matter topology

12
[42-30]

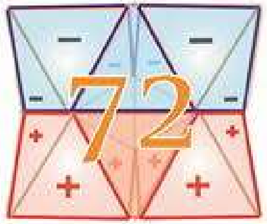

40π

Deuteron

Neutron

nett Charge
component charges
0
[18-18]

charged mass-energy geometry

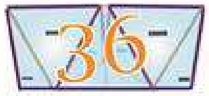

Matter topology
20π

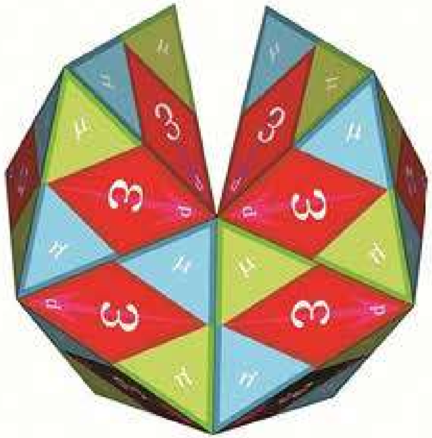

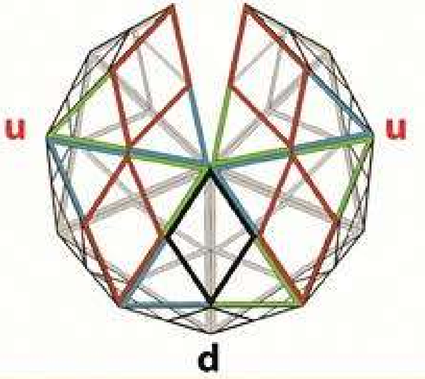

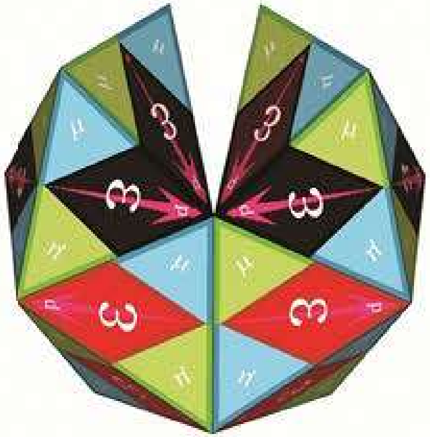

Charge provides the framework for the mass-energy geometry of Matter

Copyright ABRAHAM [2008] - All rights reserved

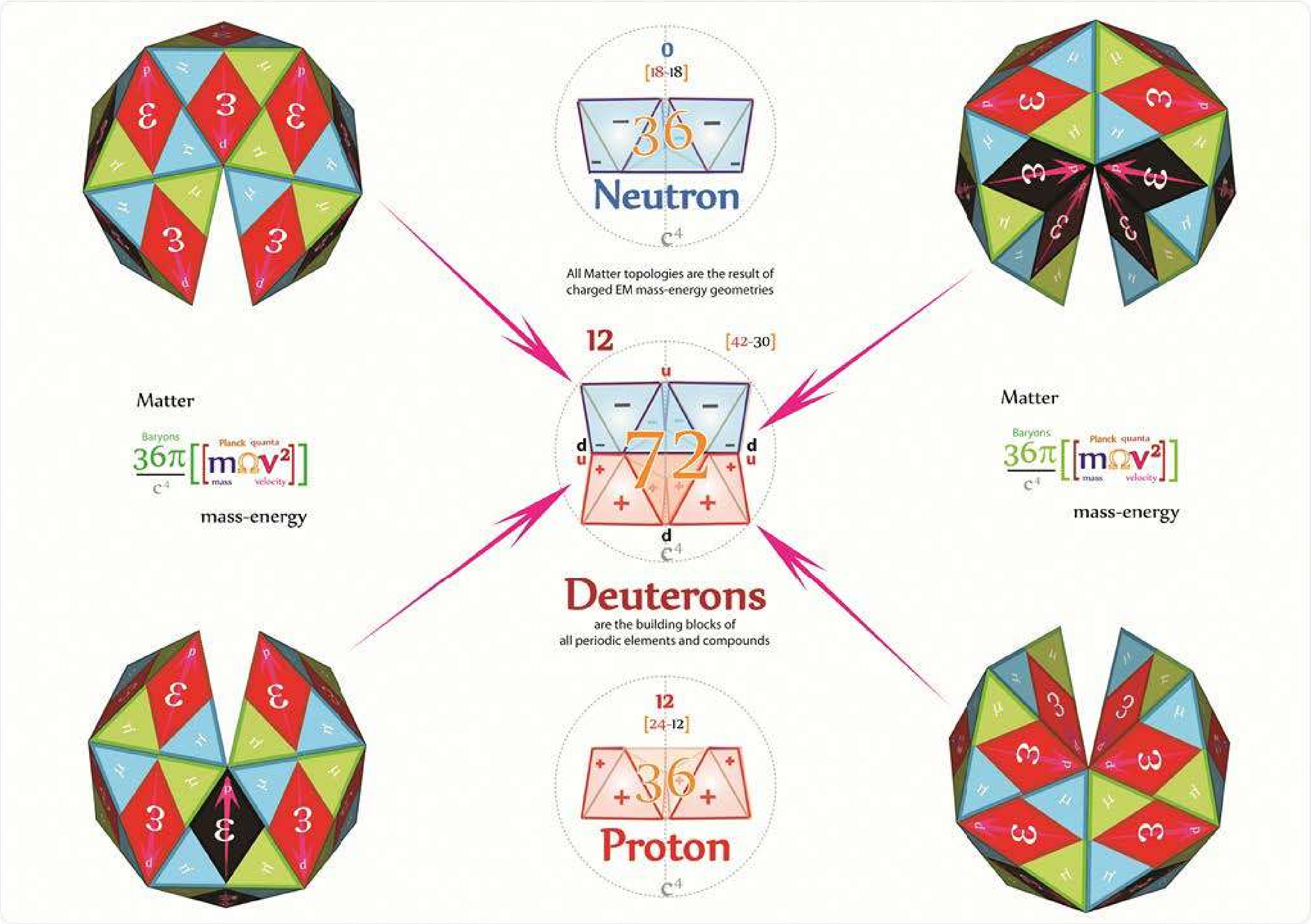
0
[18-18]
36
Neutron
c^4
12
[42-30]
u
d
u
d
u
72
d
c^4
12
[24-12]
36
Proton
c^4
ε
μ
p
d

All Matter topologies are the result of charged EM mass-energy geometries

Deuterons

are the building blocks of all periodic elements and compounds

Matter

Baryons Planck quanta

$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energy

Matter

Baryons Planck quanta

$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energy

Copyright ABRAHAM [2008] - All rights reserved

Strong Forces and Nuclear Bonding

How do Baryons with Positive and Neutral charges attract each other and bind to form stable elements ?

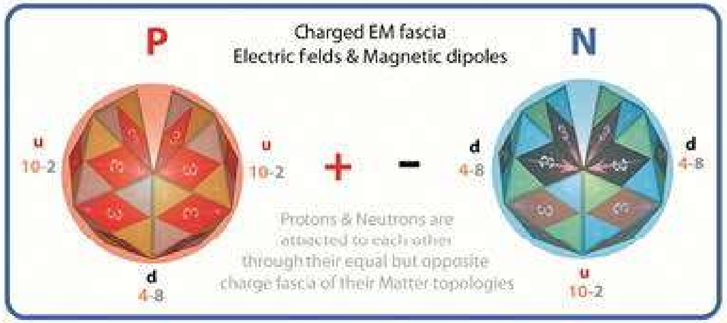
P
Charged EM fascia
Electric felds & Magnetic dipoles
N
u
10-2
u
10-2
+
-
d
4-8
d
4-8
d
4-8
u
10-2

Protons & Neutrons are attracted to each other through their equal but opposite charge fascia of their Matter topologies

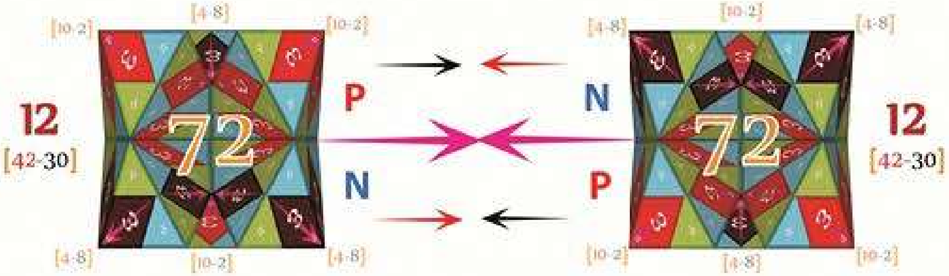
[10-2]
[4-8]
[10-2]
12
[42-30]
72
P
N
N
P
72
12
[42-30]
[4-8]
[10-2]
[4-8]
[4-8]
[10-2]
[4-8]
[10-2]
[4-8]
[10-2]

Once nuclei have been created their external electric fields & magnetic dipoles continue to attract and bind individual nuclei together via the Residual EM Force as nuclei seek charge equilibrium by combining with each other and electrons to form neutral elements

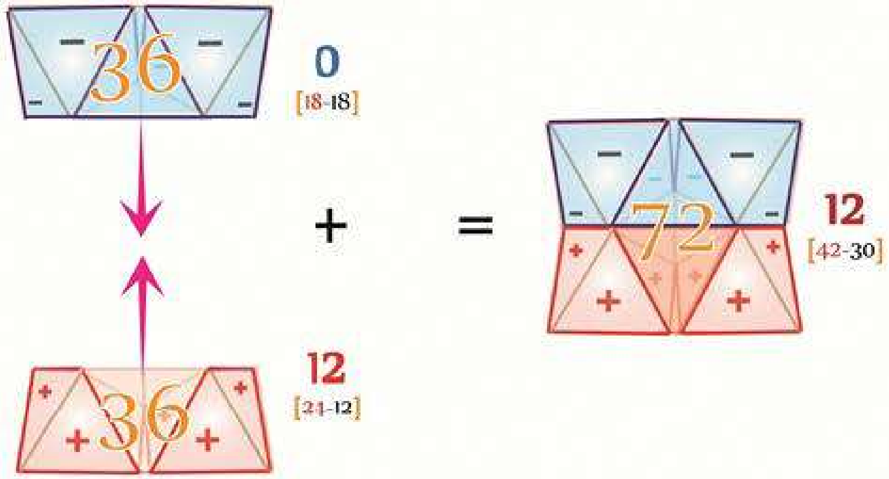
36
0
[18-18]
+
=
72
12
[42-30]
36
12
[24-12]

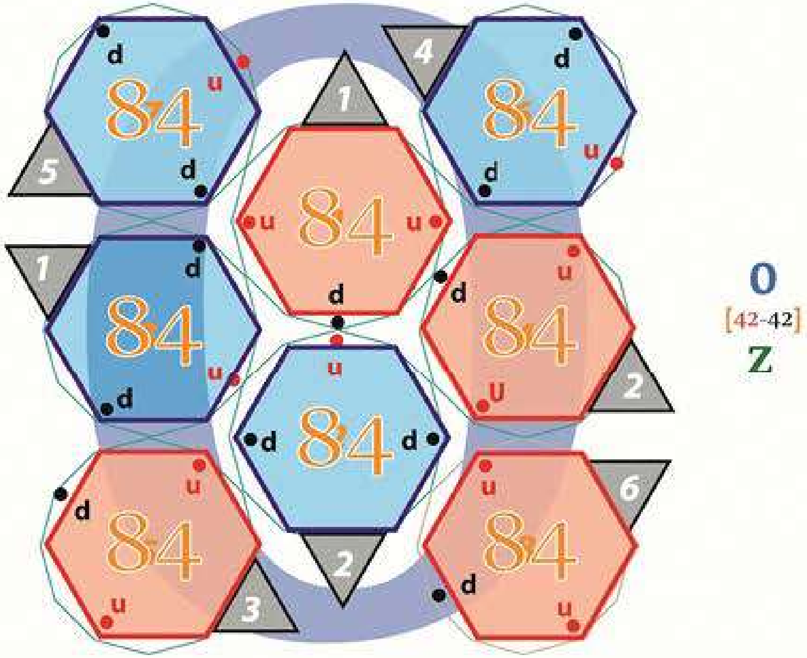
84
84
84
84
84
84
84
84
d
u
d
d
u
d
u
u
d
d
u
d
u
d
u
d
d
u
d
u
d
u
d
u
1
2
3
4
5
6
1
2
0
[42-42]
Z

The attraction and binding of Protons and Neutrons through their electric charge imbalances creates Deuterons which have +12 charges

The residual Z[+12] charge is what attracts electrons to form neutral atomic nuclei via Coulombic attraction

Strong Nucleonic Bonding

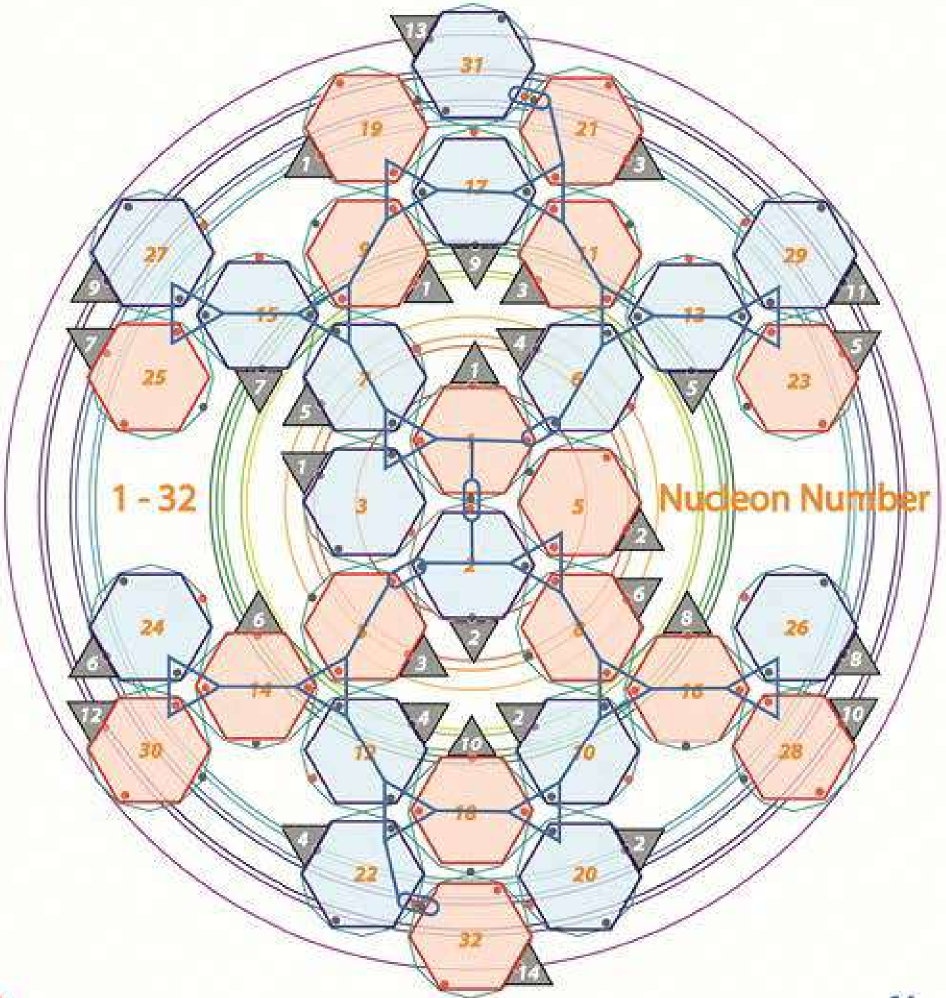
1 - 32
Nucleon Number
31
19
21
17
9
11
27
29
15
13
25
23
7
6
3
5
1
2
24
26
14
16
8
30
28
12
10
18
22
20
32

P

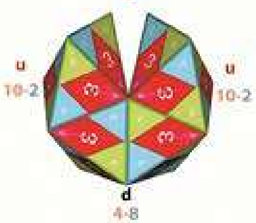
u
10-2
u
10-2
d
4-8

+

In addition to the Strong colour force
a separate residual binding force arises
from the external apexes formed by
Positive and Negative Electric points of quarks
in each Baryon [nuclear-chemical bonds]

-

N

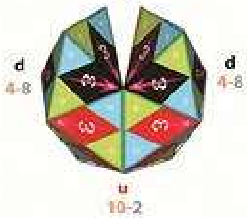
d
4-8
d
4-8
u
10-2

All energy seeks equilibrium

Strong E-force
nuclear linkages

External Electric field points
bind via Charge fascia interactions
with Plus and Minus Electric points
combining and sharing energy
throughout the resultant nuclei

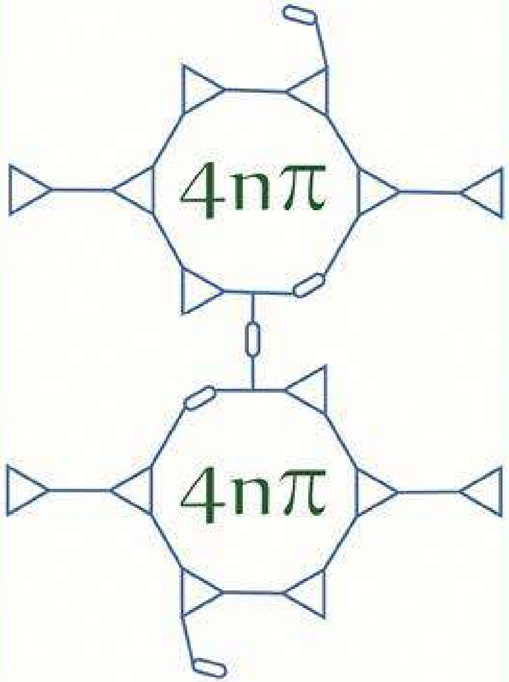

● *Positive Electric field apex*

Negative Electric field apex ●

Hydrogenic vs Nucleonic electron binding

If a unbound Proton attracts an Electron
the Electron can be bound to the nuclei
in a number of differing orientations
[each with differing spin energies]

All atomic nuclei (and elements) are Deuteron'ic nuclei with a mixture of *orthagonal*, parallel and anti-paralllel spin orientations

(this is why Rydberg is less accurate for elemental nuclei compared to Hydrogenic atoms - see QM spin)

Electrons can be bound to deuteron nuclei
in four distinct orientations [2 horizontal & 2 vertical]
with each spin coupling orientation producing
differing energy electron orbitals
[wrt to the nuclear magnetic moments]

Ejecting electrons from atomic nuclei
by adding energies to their KEM fields
[the Photoelectric effect]
creates Positive Ions

Vertically orientated electrons
within Proton-Neutron Nuclei [Deuterons]
create quantum synchronous convertor geometries

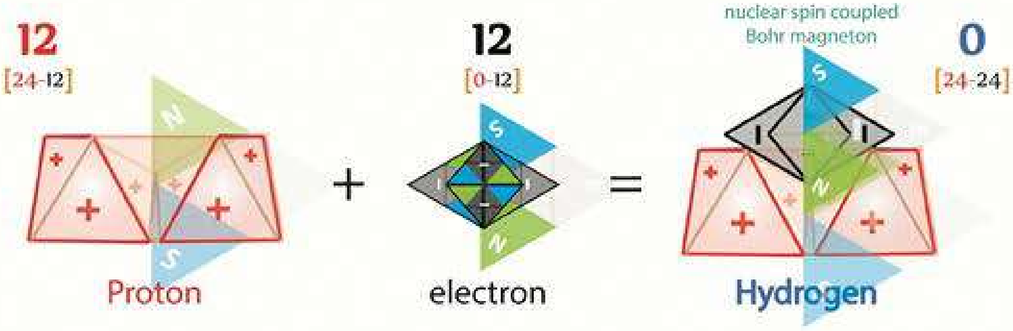
12
[24-12]
Proton
+
12
[0-12]
electron
=
nuclear spin coupled
Bohr magneton
0
[24-24]
Hydrogen

The energy levels of Baryons determines the KEM field energy of bound electrons

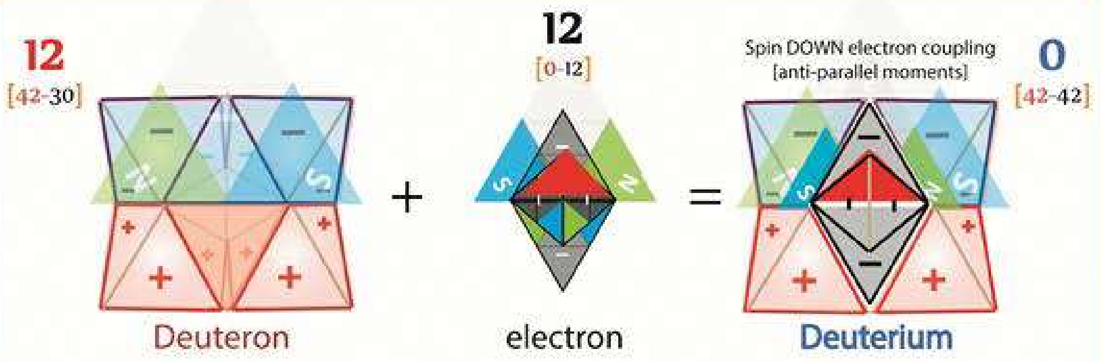
12
[42-30]
Deuteron
+
12
[0-12]
electron
=
Spin DOWN electron coupling
[anti-parallel moments]
0
[42-42]
Deuterium

Bohr magnetons are always referenced wrt the Nuclear magneton

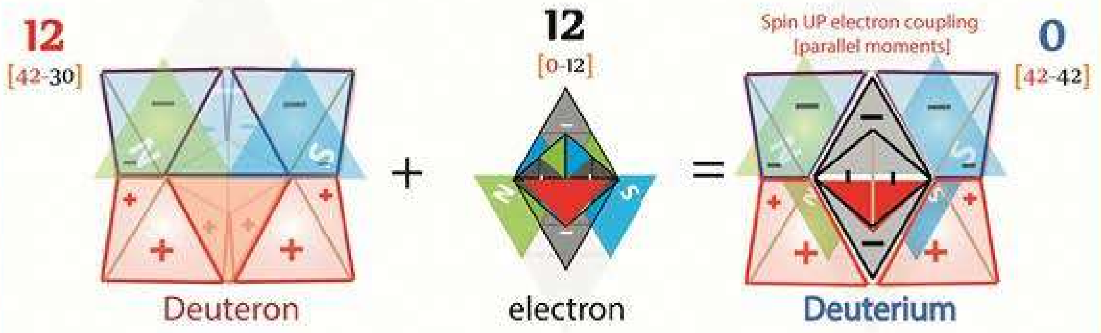
12
[42-30]
Deuteron
+
12
[0-12]
electron
=
Spin UP electron coupling
[parallel moments]
0
[42-42]
Deuterium

electrons produce stronger magnetic moments due to their mass-charge quotient

Atomic Nucleus Master Template

Nucleon Number Z

1 Proton	[24-12]
1 electron	[0-12]
1 Neutron	[18-18]

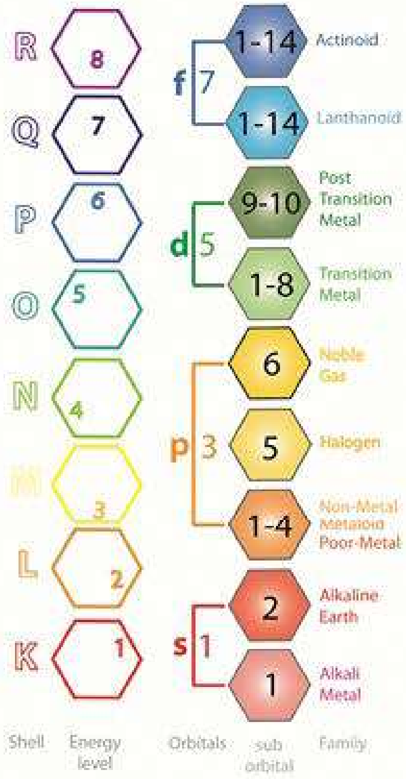
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
f 7
1-14
Actinoid
1-14
Lanthanoid
d 5
9-10
Post Transition Metal
1-8
Transition Metal
p 3
6
Noble Gas
5
Halogen
1-4
Non-Metal Metaloid Poor-Metal
s 1
2
Alkaline Earth
1
Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

○ s 1 sub-Orbital (2 electrons max)

○ p 3 sub-Orbitals (6 electrons max)

○ d 5 sub-Orbitals (10 electrons max)

○ f 7 sub-Orbitals (14 electrons max)

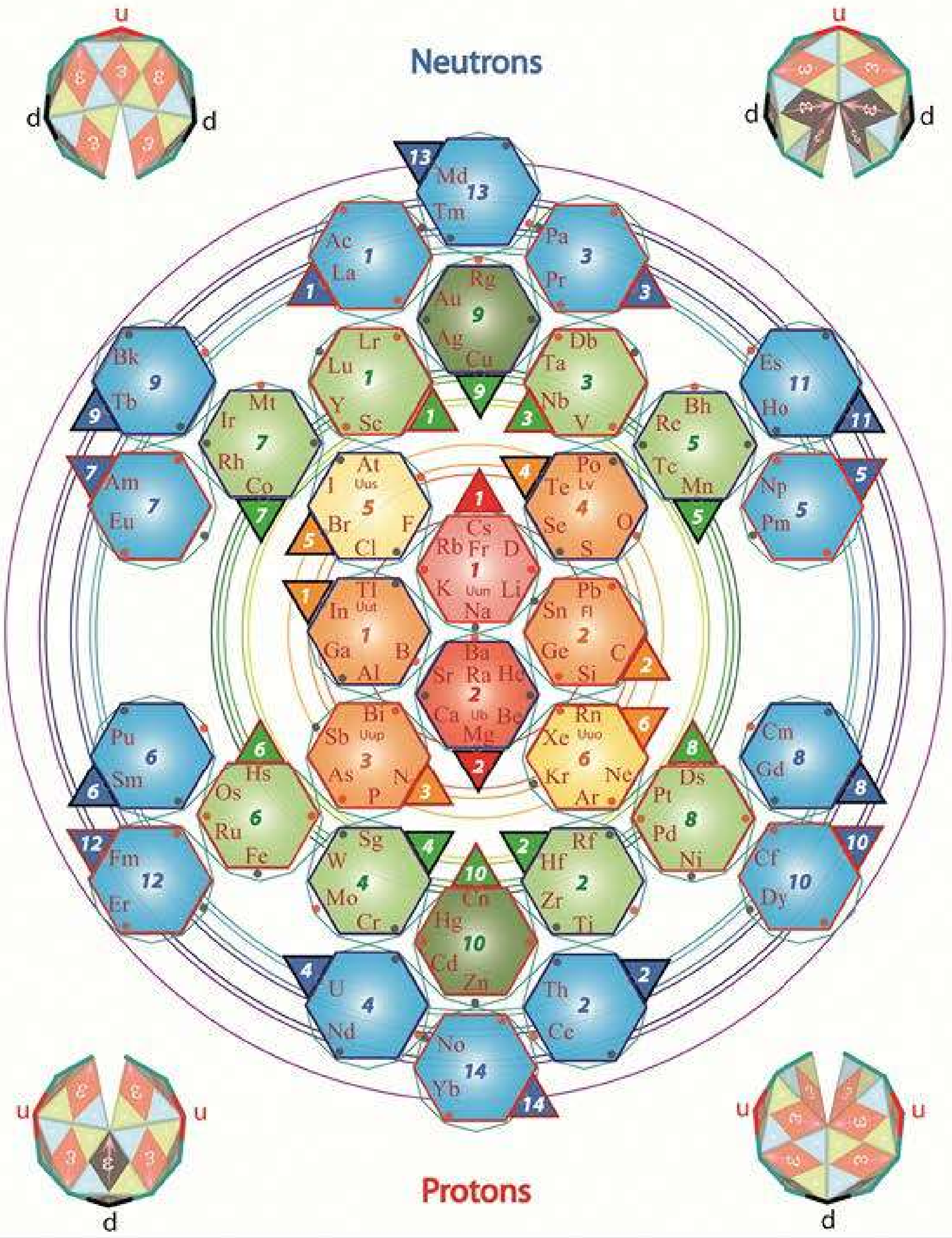
u
d
d
u
d
d
Neutrons
13
Md 13 Tm
Ac 1 La
1
Pa 3 Pr
3
Rg Au 9 Ag Cu
9
Lr Lu 1 Y Sc
1
Db Ta 3 Nb V
3
Bk 9 Tb
9
Mt Ir 7 Rh Co
7
Bh Re 5 Tc Mn
5
Es 11 Ho
11
Am 7 Eu
7
At I Uus 5 Br F Cl
5
Cs Rb Fr D 1 K Uun Li Na
1
Po Te Lv 4 Se O S
4
Np 5 Pm
5
Tl In Uut 1 Ga B Al
1
Ba Sr Ra He 2 Ca Ub Be Mg
2
Pb Sn Fl 2 Ge C Si
2
Pu 6 Sm
6
Hs Os 6 Ru Fe
6
Bi Sb Uup 3 As N P
3
Rn Xe Uuo 6 Kr Ne Ar
6
Ds Pt 8 Pd Ni
8
Cn 8 Gd
8
Fm 12 Er
12
Sg W 4 Mo Cr
4
Cn Hg 10 Cd Zn
10
Rf Hf 2 Zr Ti
2
Cf 10 Dy
10
U 4 Nd
4
Th 2 Ce
2
No 14 Yb
14
u
u
d
u
u
d
Protons

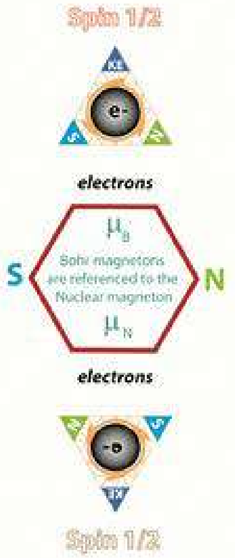
Spin 1/2
e-
electrons
μ_B
Bohr magnetons are referenced to the Nuclear magneton
μ_N
S
N
electrons
-e
Spin 1/2

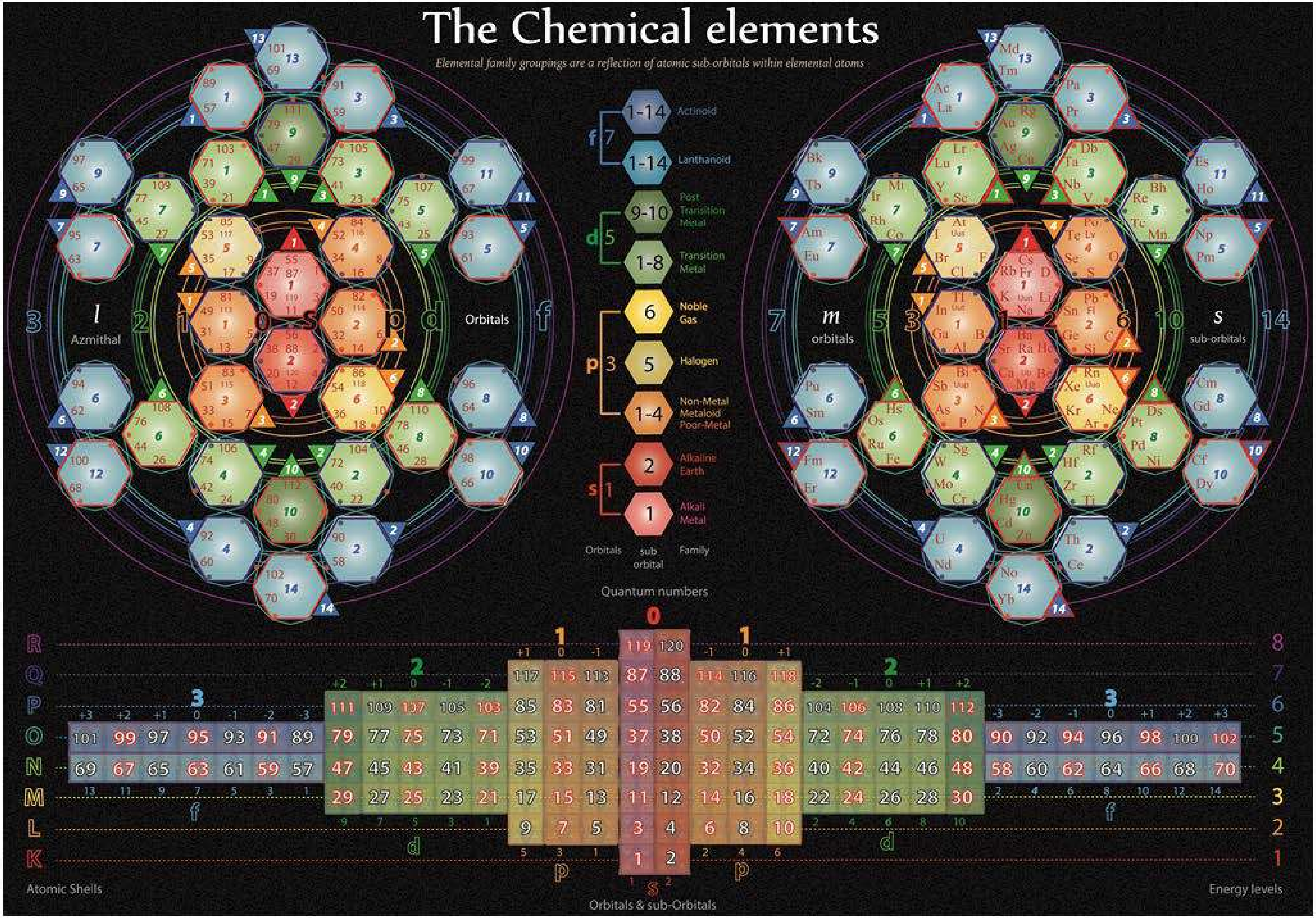

The Chemical elements

Elemental family groupings are a reflection of atomic sub-orbitals within elemental atoms

Orbitals	sub orbital	Family
f 7	1-14	Actinoid
	1-14	Lanthanoid
d 5	9-10	Post Transition Metal
	1-8	Transition Metal
p 3	6	Noble Gas
	5	Halogen
	1-4	Non-Metal Metaloid Poor-Metal
s 1	2	Alkaline Earth
	1	Alkali Metal

	3							2					1			0		1			2					3							
	+3	+2	+1	0	-1	-2	-3	+2	+1	0	-1	-2	+1	0	-1			-1	0	+1	-2	-1	0	+1	+2	-3	-2	-1	0	+1	+2	+3	
R																119	120																8
Q													117	115	113	87	88	114	116	118													7
P								111	109	107	105	103	85	83	81	55	56	82	84	86	104	106	108	110	112								6
O	101	99	97	95	93	91	89	79	77	75	73	71	53	51	49	37	38	50	52	54	72	74	76	78	80	90	92	94	96	98	100	102	5
N	69	67	65	63	61	59	57	47	45	43	41	39	35	33	31	19	20	32	34	36	40	42	44	46	48	58	60	62	64	66	68	70	4
M								29	27	25	23	21	17	15	13	11	12	14	16	18	22	24	26	28	30								3
L													9	7	5	3	4	6	8	10													2
K																1	2																1
	13	11	9	7	5	3	1	9	7	5	3	1	5	3	1	1	2	2	4	6	2	4	6	8	10	2	4	6	8	10	12	14	
				f						d				p		s			p				d						f				

Atomic Shells

Orbitals & sub-Orbitals

Energy levels

Tetryonics 43.02 - The Chemical Elements

Element Numbers

Z *

1 Proton	[24-12]
1 Neutron	[18-18]
1 electron	[0-12]

Each elemental nuclei is made from Deuterium

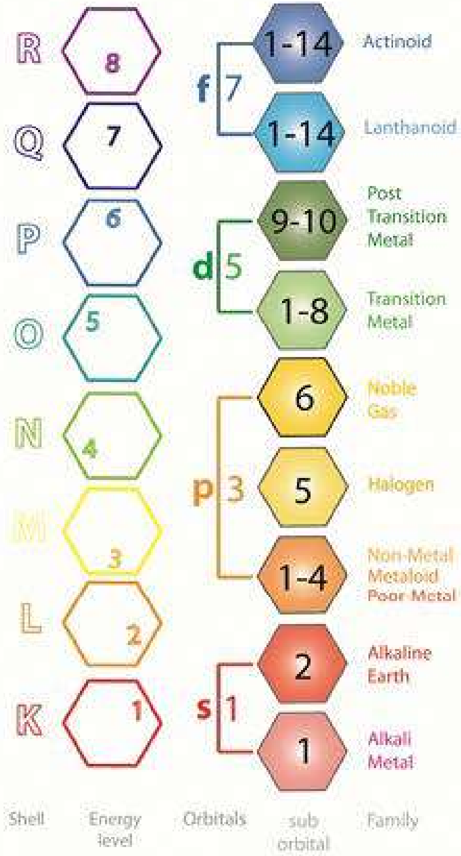
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

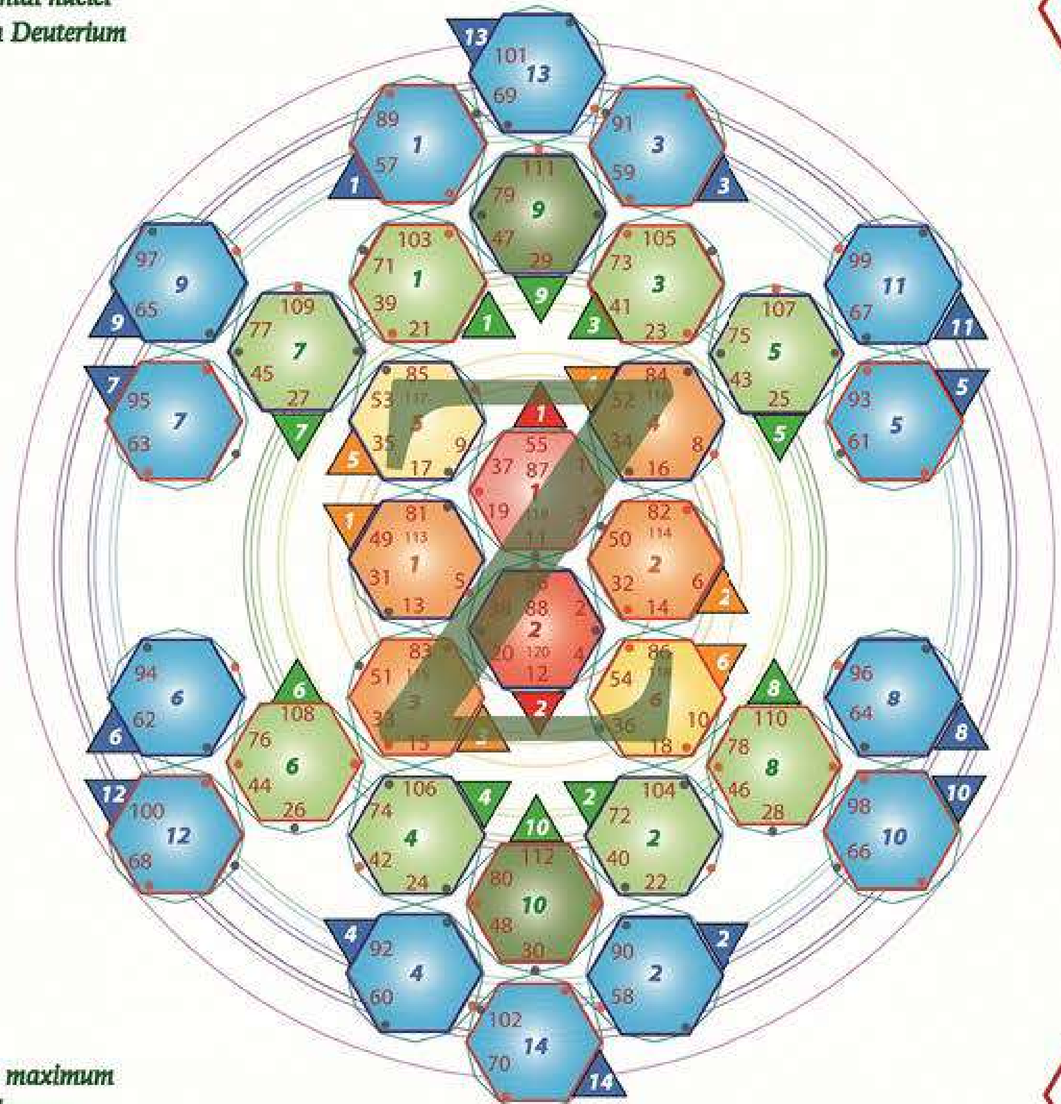
13
101
13
69
89
1
57
1
91
3
59
3
111
79
9
47
29
9
97
9
65
9
103
71
1
39
21
1
105
73
3
41
23
3
99
11
67
11
109
77
7
45
27
7
107
75
5
43
25
5
95
7
63
7
93
5
61
5
85
53
117
35
17
5
84
52
4
34
16
8
55
37
87
1
19
1
81
49
113
1
31
13
1
82
50
114
2
32
6
14
2
88
2
20
120
12
2
83
51
33
15
3
86
54
6
36
18
10
6
94
6
62
6
108
76
6
44
26
6
110
78
8
46
28
8
96
8
64
8
12
100
12
68
106
74
4
42
24
4
104
72
2
40
22
2
98
10
66
10
112
80
10
48
30
10
4
92
4
60
90
2
58
2
102
14
70
14

Atomic shells

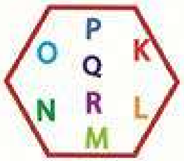

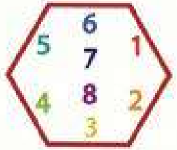

Energy levels

o	s	1 sub-Orbital	(2 electrons max)
O	p	3 sub-Orbitals	(6 electrons max)
O	d	5 sub-Orbitals	(10 electrons max)
O	f	7 sub-Orbitals	(14 electrons max)

There are a maximum of 120 elements possible

Copyright ABRAHAM [2008] - All rights reserved

Element Names

Z * 1 Proton [24-12]
1 Neutron [18-18]
1 electron [0-12]

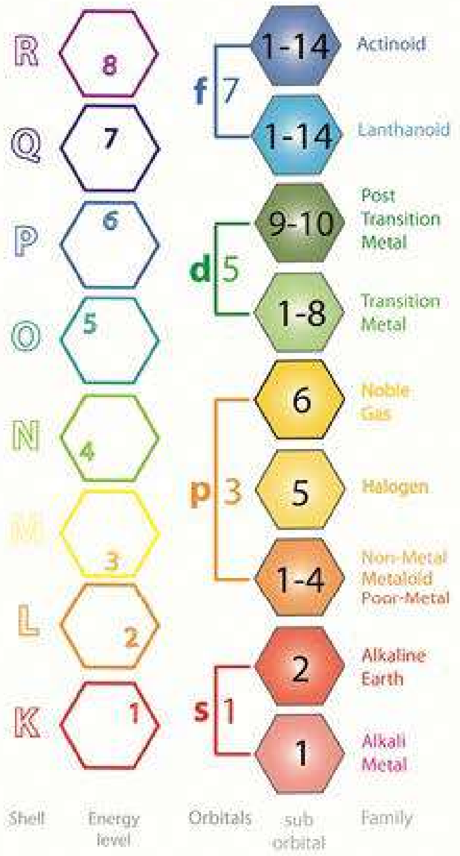
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f 7
d 5
p 3
s 1
1-14 Actinoid
1-14 Lanthanoid
9-10 Post Transition Metal
1-8 Transition Metal
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

There are a maximum of 120 elements possible

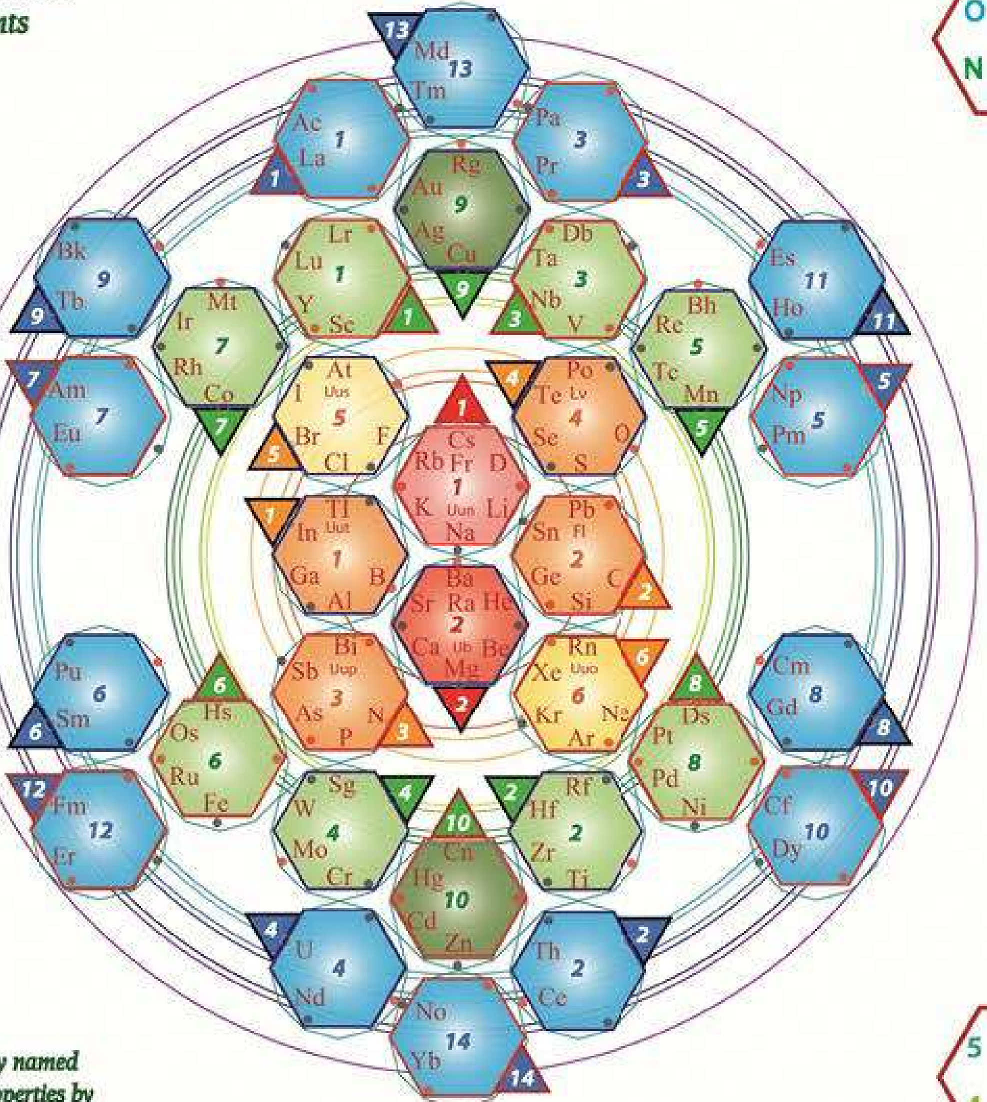
13
Md
13
Tm
Ac
1
La
1
Pa
3
Pr
3
Rg
Au
9
Ag
Cu
9
Lr
Lu
1
Y
Sc
1
Db
Ta
3
Nb
V
3
Bk
9
Tb
9
Es
11
Ho
11
Mt
Ir
7
Rh
Co
7
Bh
Re
5
Tc
Mn
5
At
I
Uus
5
Br
F
Cl
5
Po
Te
Lv
4
Se
O
S
4
1
Cs
Rb
Fr
D
1
K
Uun
Li
Na
7
Am
7
Eu
Np
5
Pm
5
Tl
In
Uut
1
Ga
B
Al
1
Pb
Sn
Fl
2
Ge
C
Si
2
Ba
Sr
Ra
He
2
Ca
Ub
Be
Mg
2
Bi
Sb
Uup
3
As
N
P
3
Rn
Xe
Uuo
6
Kr
Ne
Ar
6
Pu
6
Sm
6
6
Hs
Os
6
Ru
Fe
8
Ds
Pt
8
Pd
Ni
Cm
Gd
8
8
12
Fm
12
Er
Sg
W
4
Mo
Cr
4
2
Rf
Hf
2
Zr
Ti
10
Cn
Hg
10
Cd
Zn
Cf
10
Dy
10
4
U
4
Nd
Th
2
Ce
2
No
14
Yb
14

Atomic shells
O
P
Q
K
N
R
L
M

○ s 1 sub-Orbital (2 electrons max)

○ p 3 sub-Orbitals (6 electrons max)

○ d 5 sub-Orbitals (10 electrons max)

○ f 7 sub-Orbitals (14 electrons max)

They were originally named according to their properties by their discoverer but have recently been named after famous scientists

5
6
7
1
4
8
2
3
Energy levels

Copyright ABRAHAM [2008] - All rights reserved

Electron orbital configurations

A huge number of differing d and f orbital configurations are possible given the number of nuclei and bond points created by elemental topologies

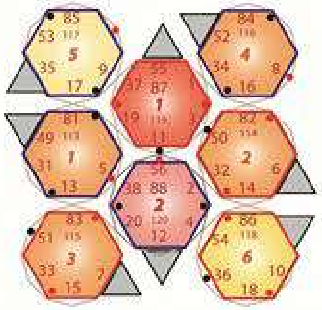
85
53
117
5
35
9
17
55
37
87
1
19
119
11
84
52
116
4
34
8
16
81
49
113
1
31
5
13
56
38
88
2
20
120
4
12
82
50
114
2
32
6
14
83
51
115
3
33
7
15
86
54
118
6
36
10
18

As the number of nucleons increases so does the complexity of the electron orbitals possible

quantum snowflakes

However all have a stable 'core' grouping of nuclei comprised of s and p electron orbitals

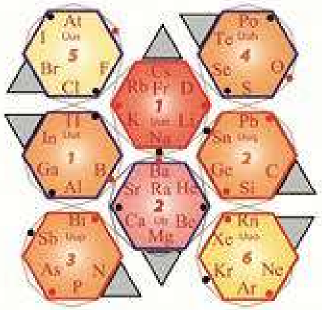
At
I
Uus
5
Br
F
Cl
Cs
Rb
Fr
D
1
K
Uue
Li
Na
Po
Te
Uuh
4
Se
O
S
Tl
In
Uut
1
Ga
B
Al
Ba
Sr
Ra
He
2
Ca
Ubn
Be
Mg
Pb
Sn
Uuq
2
Ge
C
Si
Bi
Sb
Uup
3
As
N
P
Rn
Xe
Uuo
6
Kr
Ne
Ar

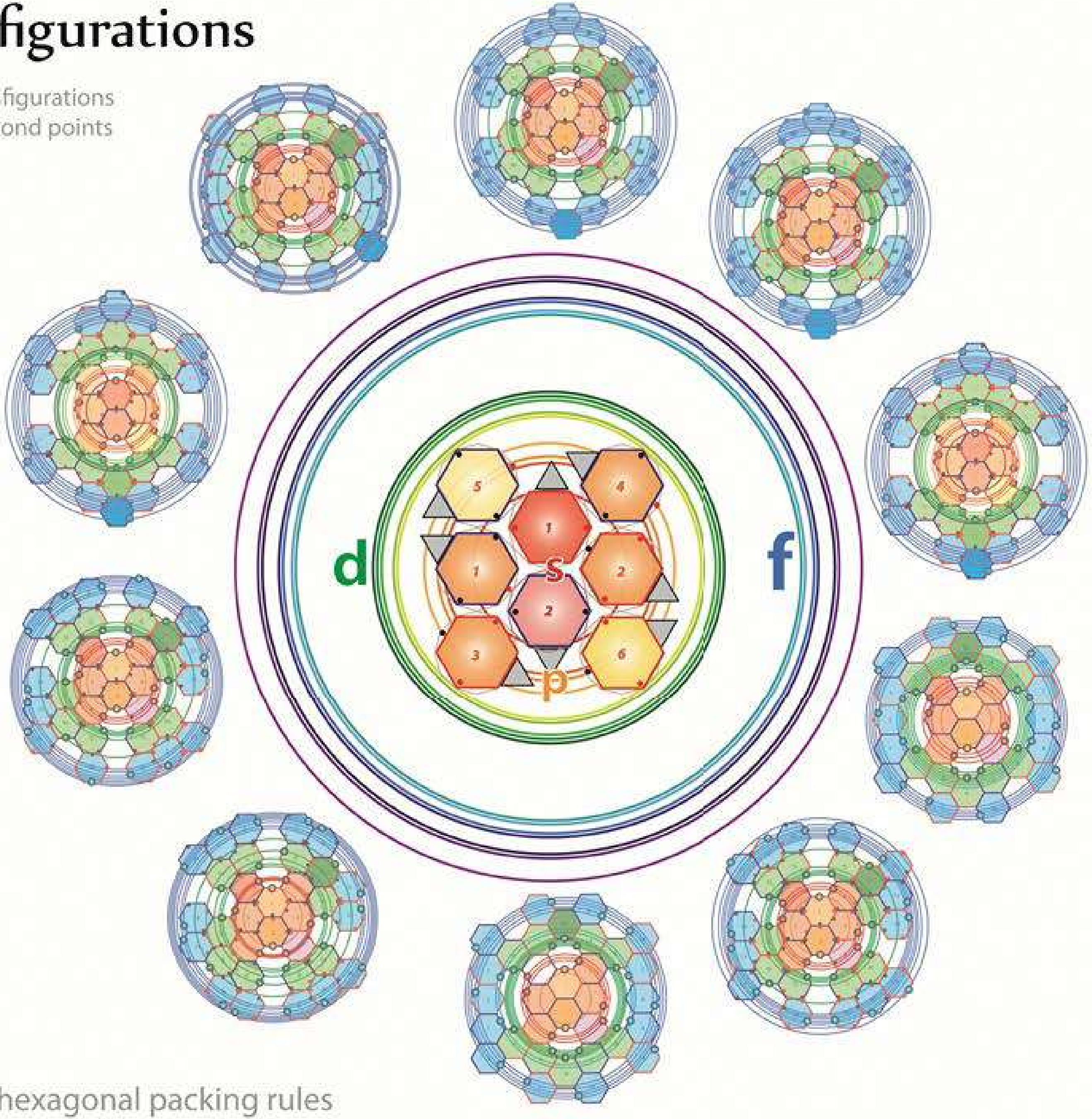
d
f
s
p
5
4
1
1
2
2
3
6

All nuclei bonding closely follows hexagonal packing rules

Atomic Nucleus
Exploded view

1 Proton [24-12]
1 electron [0-12]
1 Neutron [18-18]

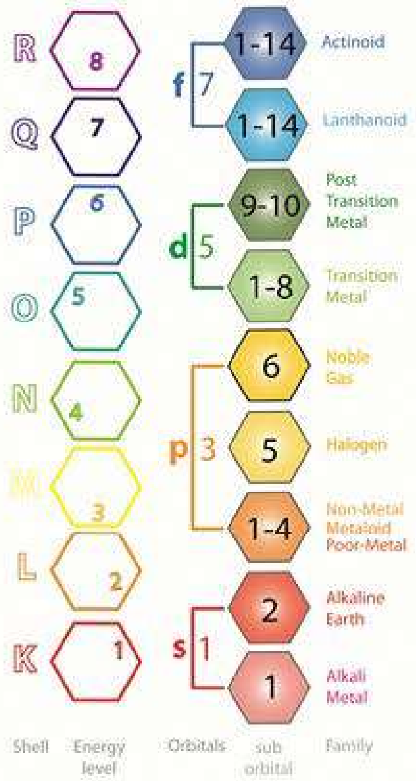
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

○ **s** *1 sub-Orbital (2 electrons max)*

○ **p** *3 sub-Orbitals (6 electrons max)*

○ **d** *5 sub-Orbitals (10 electrons max)*

○ **f** *7 sub-Orbitals (14 electrons max)*

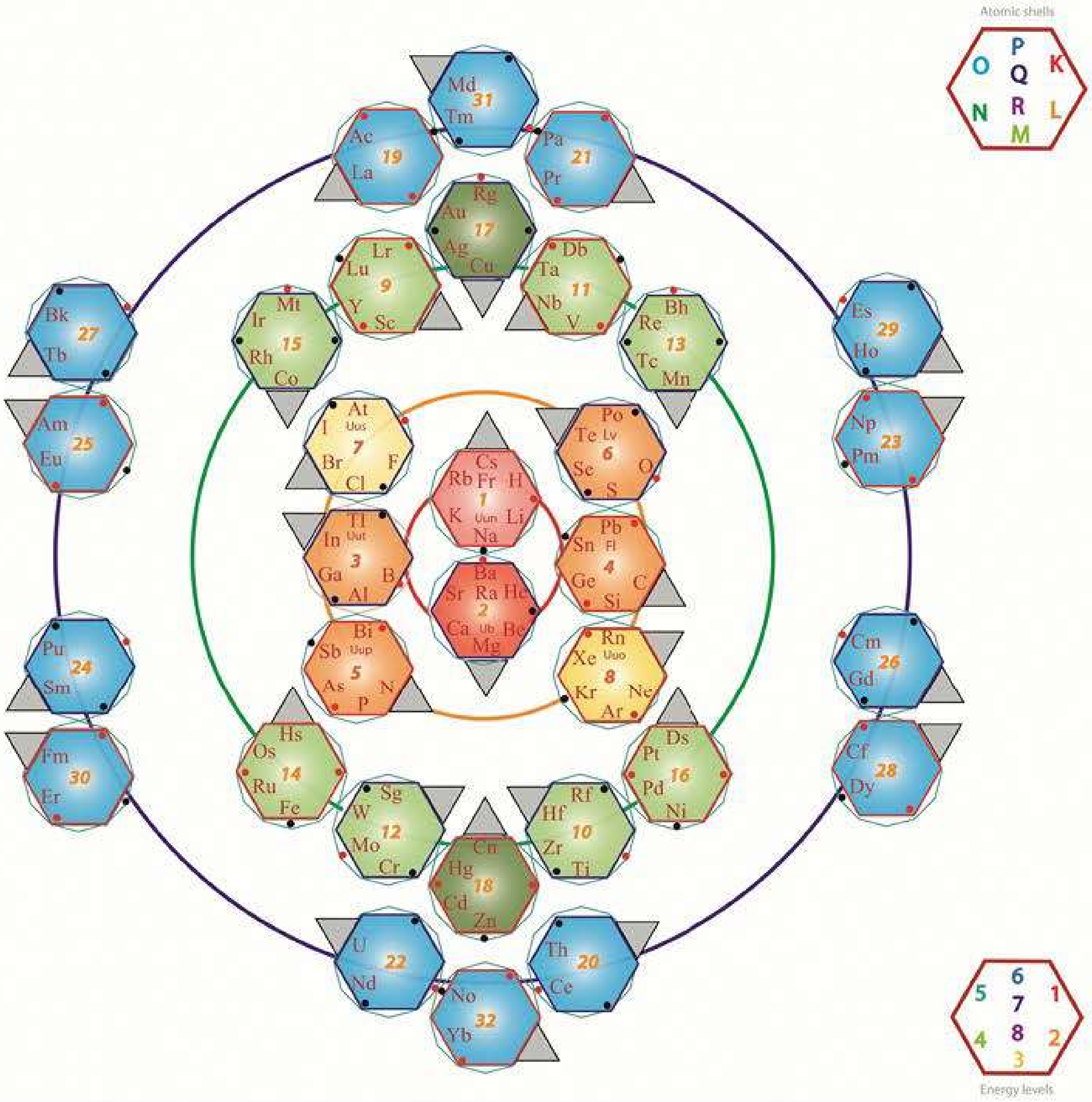
Atomic shells
P Q K O R L N M
Cs Rb Fr H 1 K Uun Li Na
Ba Sr Ra He 2 Ca Ub Be Mg
Tl In Uut 3 Ga B Al
Pb Sn Fl 4 Ge C Si
Bi Sb Uup 5 As N P
Po Te Lv 6 Se O S
At I Uus 7 Br F Cl
Rn Xe Uuo 8 Kr Ne Ar
Lr Lu 9 Y Sc
Rf Hf 10 Zr Ti
Db Ta 11 Nb V
Sg W 12 Mo Cr
Bh Re 13 Tc Mn
Hs Os 14 Ru Fe
Mt Ir 15 Rh Co
Ds Pt 16 Pd Ni
Rg Au 17 Ag Cu
Cn Hg 18 Cd Zn
Ac 19 La
Th 20 Ce
Pa 21 Pr
U 22 Nd
Np 23 Pm
Pu 24 Sm
Am 25 Eu
Cm 26 Gd
Bk 27 Tb
Cf 28 Dy
Es 29 Ho
Fm 30 Er
Md 31 Tm
No 32 Yb
6 5 1 7 8 4 2 3
Energy levels

Copyright ABRAHAM [2008] - All rights reserved

Atomic radii of elements

The Mendeleev block arrangement of periodic elements presents a number of problems in representing the true topologies and properties of all the periodic elements:

Questions are raised over Hydrogen's appropriate position in the table

The table correctly positions elements according to their elemental properties but offers no insight into the individual geometries of the periodic elements themselves

The table does not reflect the charged quantum geometries of nuclei in each atomic shell

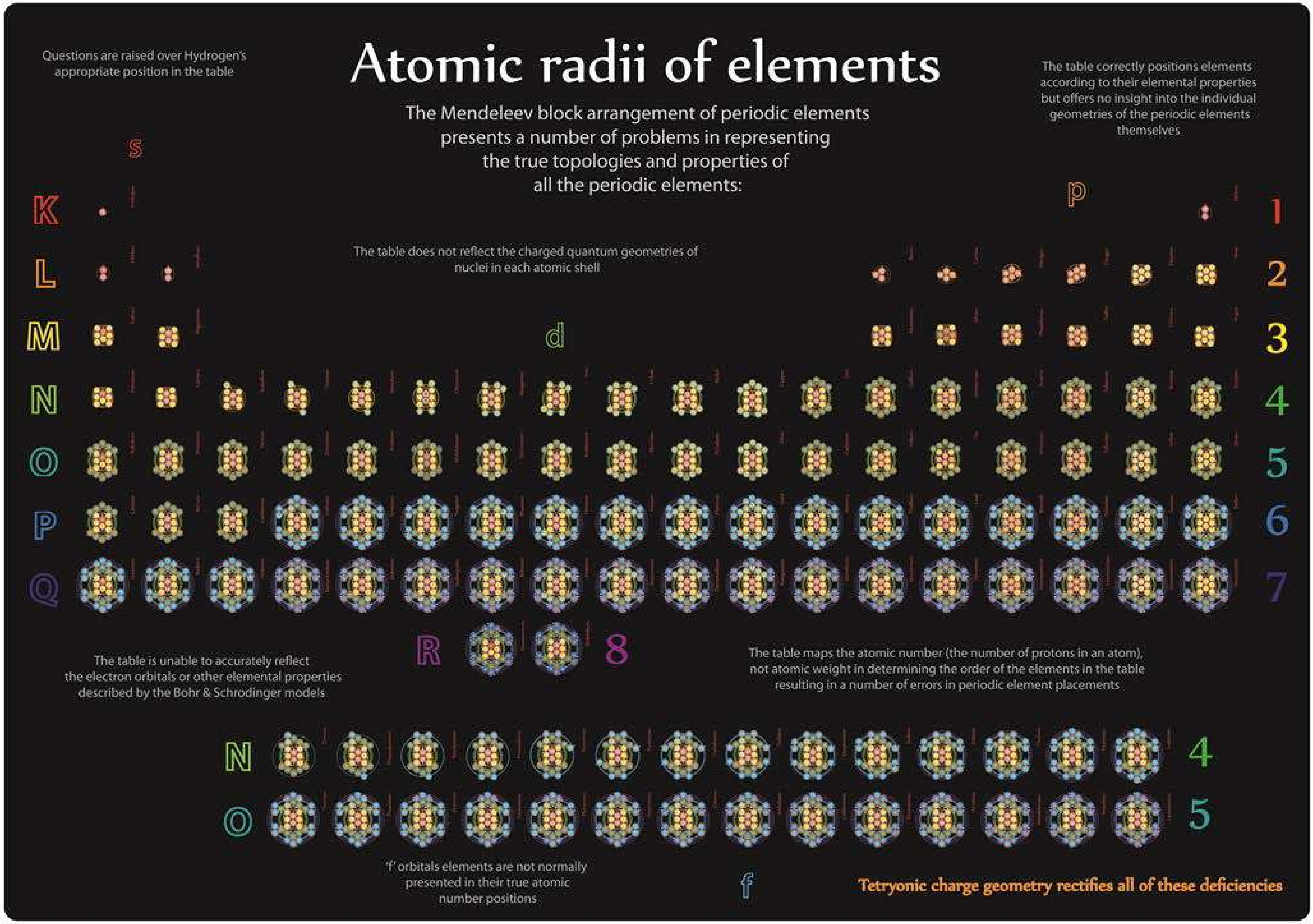

The table is unable to accurately reflect the electron orbitals or other elemental properties described by the Bohr & Schrodinger models

The table maps the atomic number (the number of protons in an atom), not atomic weight in determining the order of the elements in the table resulting in a number of errors in periodic element placements

'f' orbitals elements are not normally presented in their true atomic number positions

Tetryonic charge geometry rectifies all of these deficiencies

Tetryonics 43.07 - Atomic radii of elements

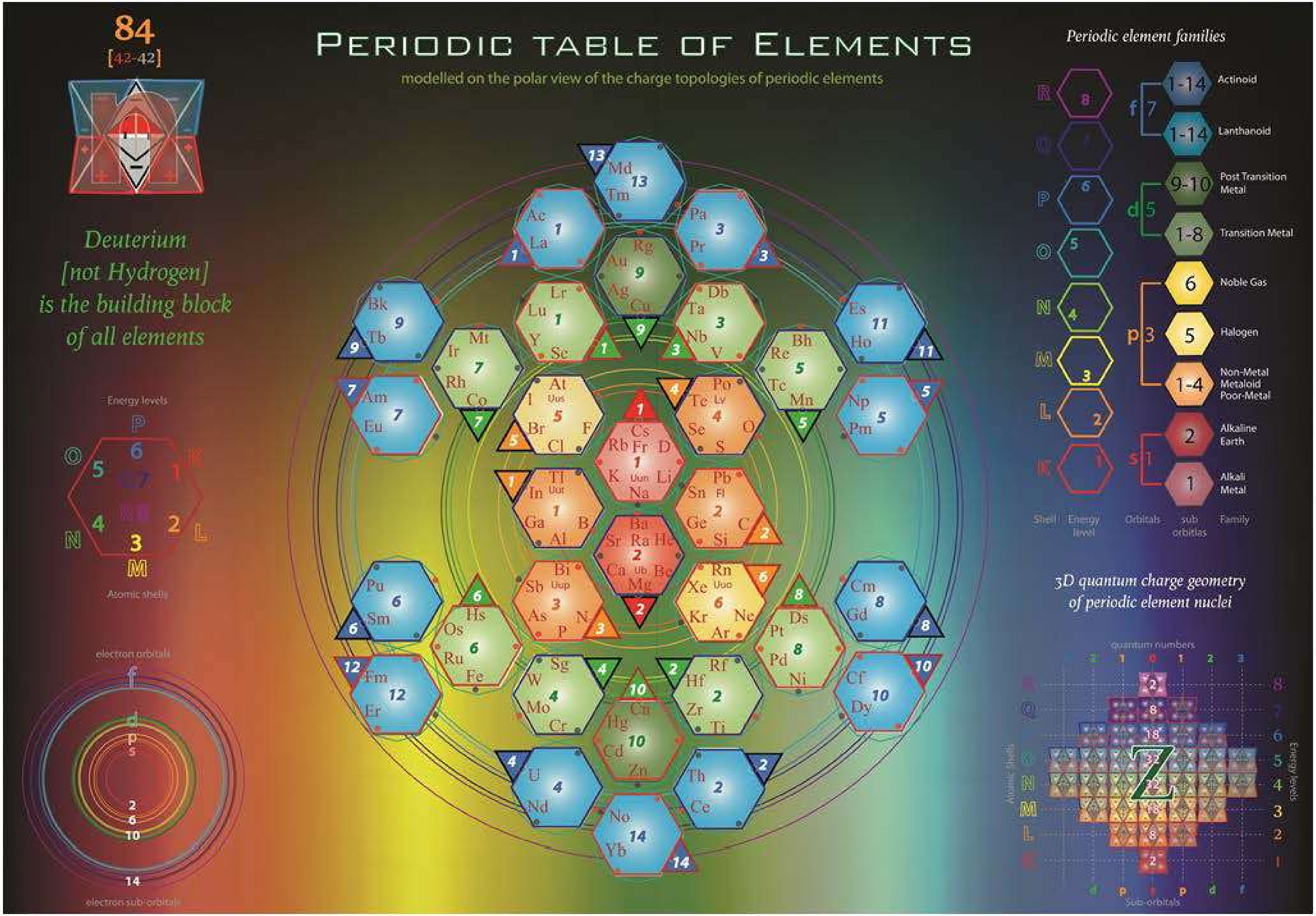

PERIODIC TABLE OF ELEMENTS

modelled on the polar view of the charge topologies of periodic elements

84
[42-42]

Deuterium [not Hydrogen] is the building block of all elements

Periodic element families

3D quantum charge geometry of periodic element nuclei

Energy levels
P 6
O 5
7
1 K
8
N 4
2 L
3
M
Atomic shells
electron orbitals
f
d
p
s
2
6
10
14
electron sub-orbitals
13 Md Tm 13
Ac La 1 1
Pa Pr 3 3
Rg Au Ag Cu 9 9
Bk Tb 9 9
Lr Lu Y Sc 1 1
Db Ta Nb V 3 3
Es Ho 11 11
Mt Ir Rh Co 7 7
Bh Re Tc Mn 5 5
Am Eu 7 7
At I Br Cl F Uus 5 5
1 Cs Rb Fr D K Uun Li Na 1
Po Te Se O S Lv 4 4
Np Pm 5 5
Tl In Ga Al B Uut 1 1
Pb Sn Ge C Si Fl 2 2
Ba Sr Ra He Ca Ub Be Mg 2 2
Pu Sm 6 6
Hs Os Ru Fe 6 6
Bi Sb As P N Uup 3 3
Rn Xe Kr Ar Ne Uuo 6 6
Ds Pt Pd Ni 8 8
Cm Gd 8 8
Fm Er 12 12
Sg W Mo Cr 4 4
Cn Hg Cd Zn 10 10
Rf Hf Zr Ti 2 2
Cf Dy 10 10
U Nd 4 4
Th Ce 2 2
No Yb 14 14
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbitals
Family
quantum numbers
2 1 0 1 2 3
2
8
18
32
32
18
8
2
Z
Atomic shells
Energy levels
8 7 6 5 4 3 2 1
d p s p d f
Sub-orbitals

Tetryonics 43.08 - Periodic Table 2.0

THE ATOMIC NUCLEUS

the charged quantum mass-energy geometry of periodic element topologies

Neils Bohr

(7 October 1885 – 18 November 1962)

NEILS BOHR

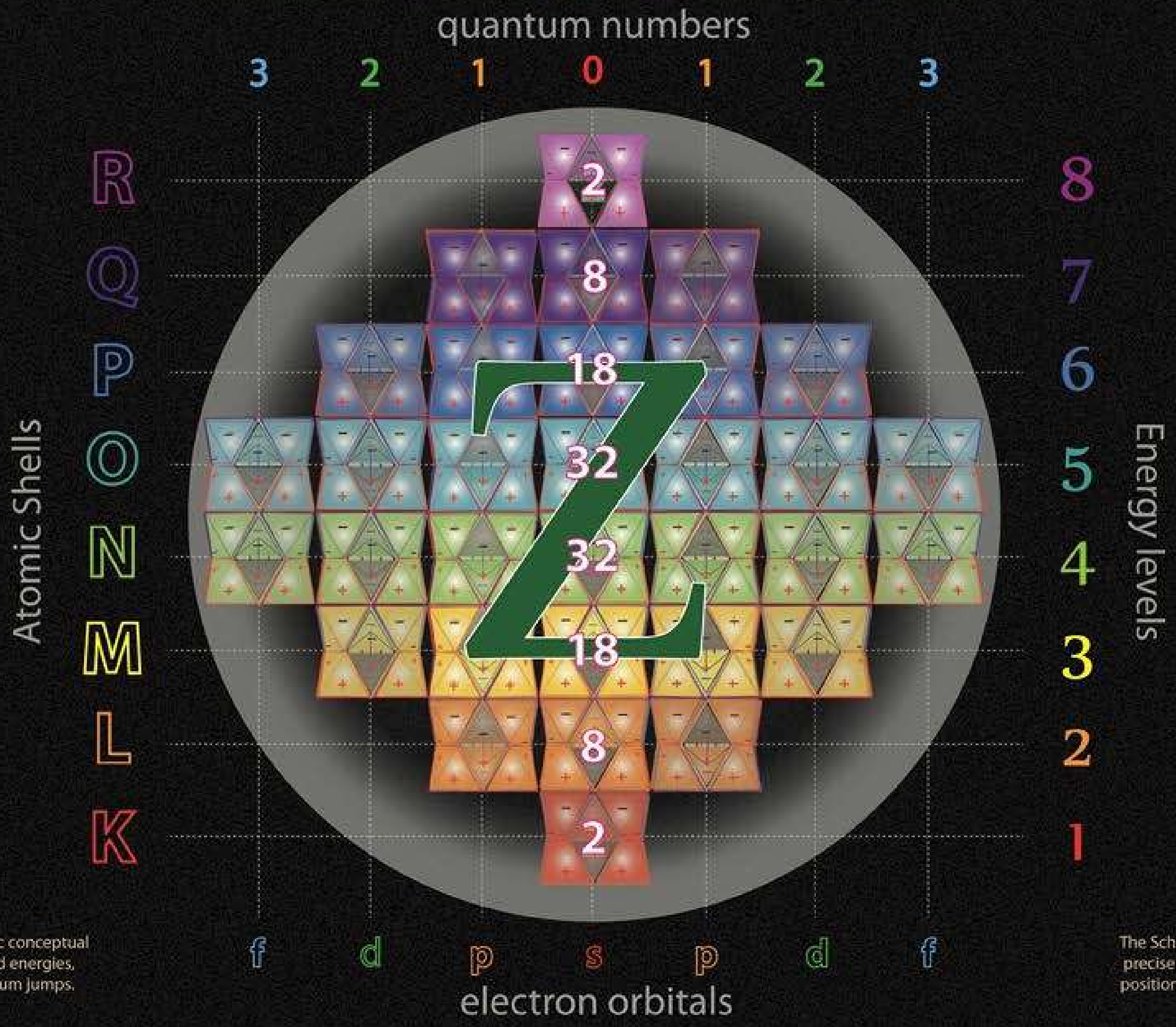
quantum numbers
3
2
1
0
1
2
3
Atomic Shells
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
Energy levels
f
d
p
s
p
d
f
electron orbitals

Erwin Schrödinger

[12 August 1887 – 4 January 1961]

SCHRÖDINGER

The Bohr model gives us a basic conceptual model of electrons orbits and energies, along with the idea of quantum jumps.

The Schrondinger wave equation gives precise details of spectra and electron position and motions within the nuclei

The uncertainty principle erroneously dictates that position and momentum cannot be simultaneously determined.

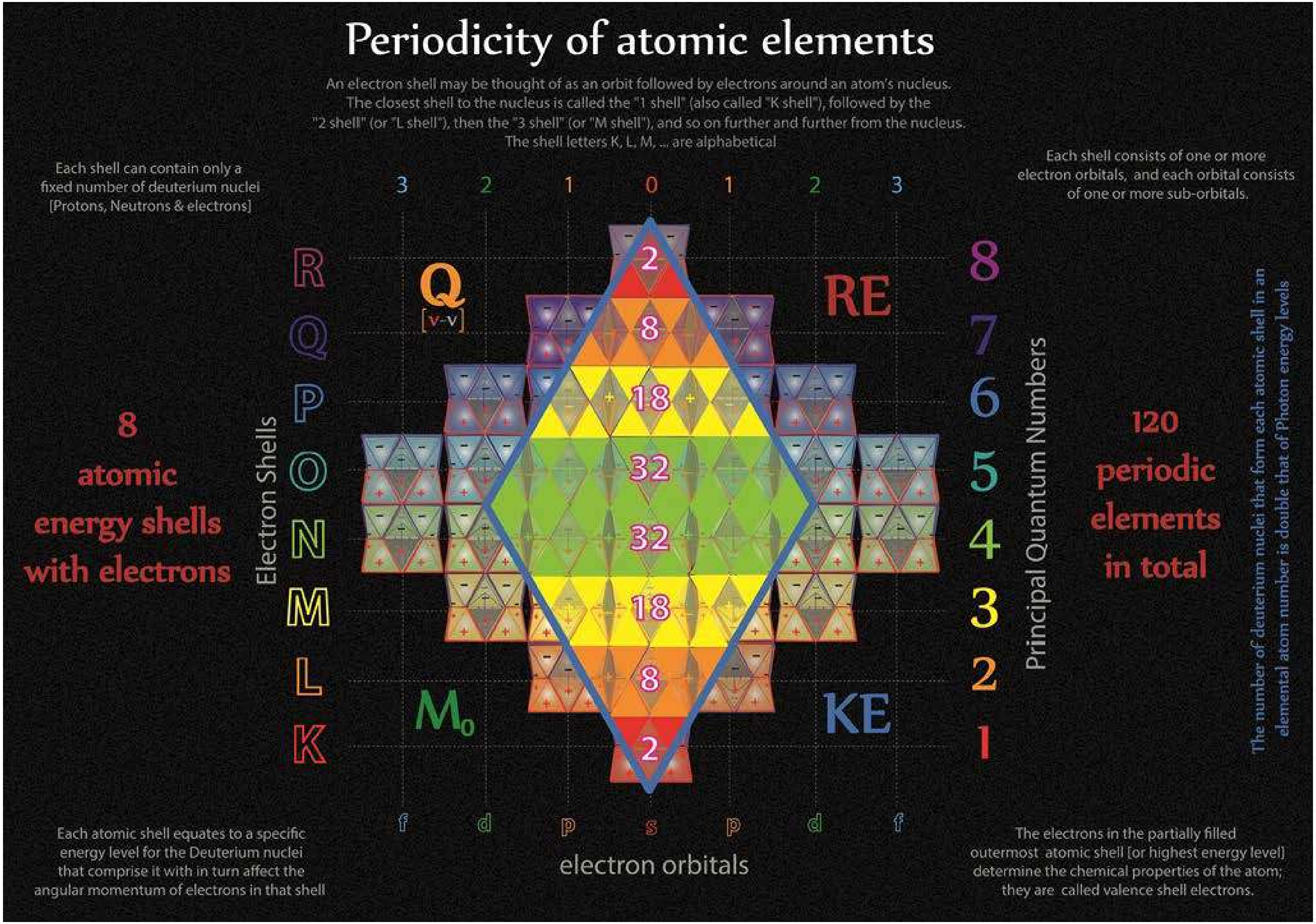

Periodicity of atomic elements

An electron shell may be thought of as an orbit followed by electrons around an atom's nucleus.
The closest shell to the nucleus is called the "1 shell" (also called "K shell"), followed by the
"2 shell" (or "L shell"), then the "3 shell" (or "M shell"), and so on further and further from the nucleus.
The shell letters K, L, M, ... are alphabetical

Each shell can contain only a
fixed number of deuterium nuclei
[Protons, Neutrons & electrons]

Each shell consists of one or more
electron orbitals, and each orbital consists
of one or more sub-orbitals.

8
atomic
energy shells
with electrons

120
periodic
elements
in total

The number of deuterium nuclei that form each atomic shell in an elemental atom number is double that of Photon energy levels

Each atomic shell equates to a specific
energy level for the Deuterium nuclei
that comprise it with in turn affect the
angular momentum of electrons in that shell

The electrons in the partially filled
outermost atomic shell [or highest energy level]
determine the chemical properties of the atom;
they are called valence shell electrons.

Tetryonics 44.02 - Periodicity of atomic elements

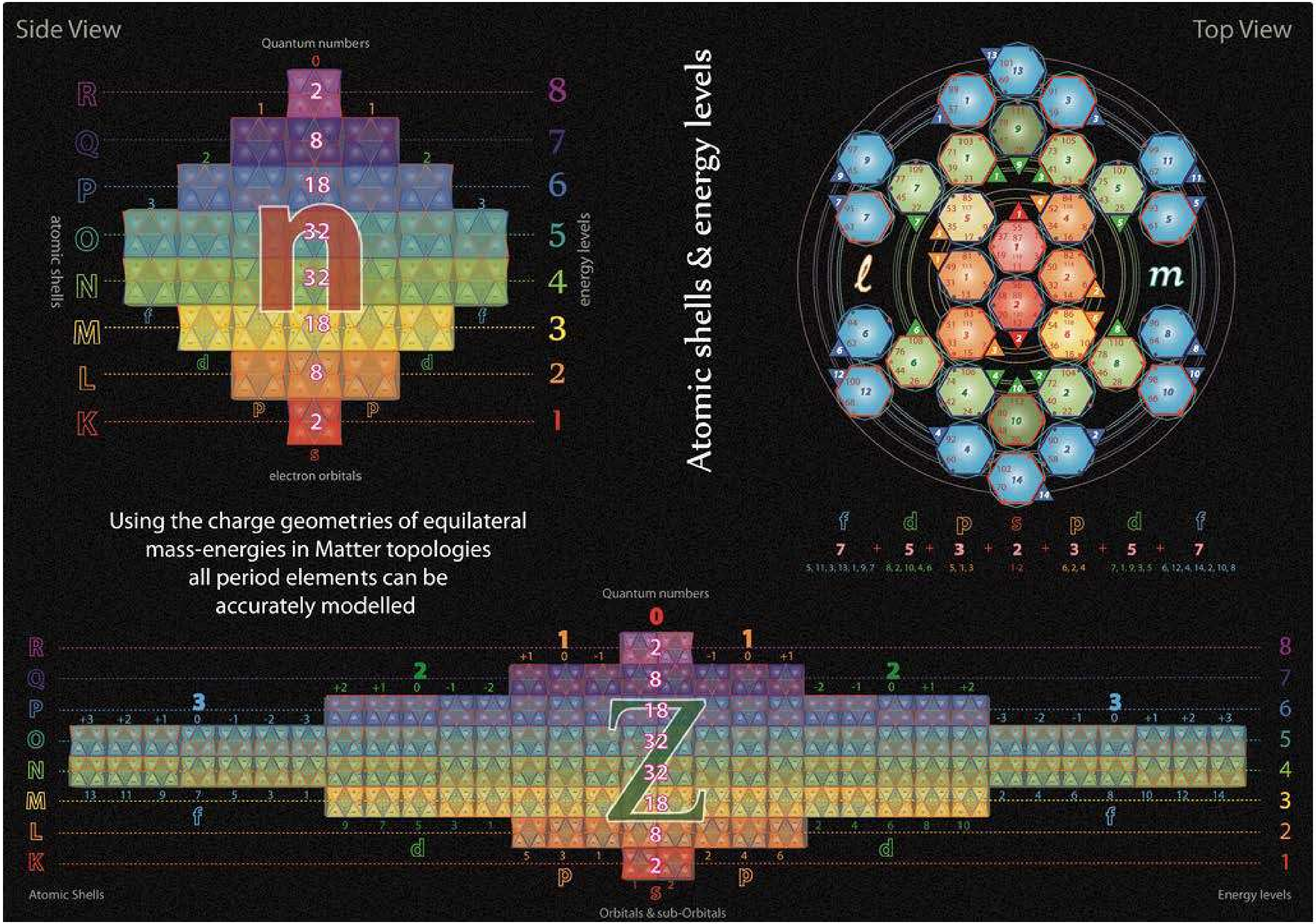

Atomic shells & energy levels

Using the charge geometries of equilateral mass-energies in Matter topologies all period elements can be accurately modelled

Side View
Quantum numbers
0
1
2
3
R
Q
P
O
N
M
L
K
atomic shells
2
8
18
32
32
18
8
2
n
8
7
6
5
4
3
2
1
energy levels
f
d
p
s
electron orbitals
Top View
l
m
f d p s p d f
7 + 5 + 3 + 2 + 3 + 5 + 7
Quantum numbers
0
1
2
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
2
8
18
32
32
18
8
2
Z
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f
d
p
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

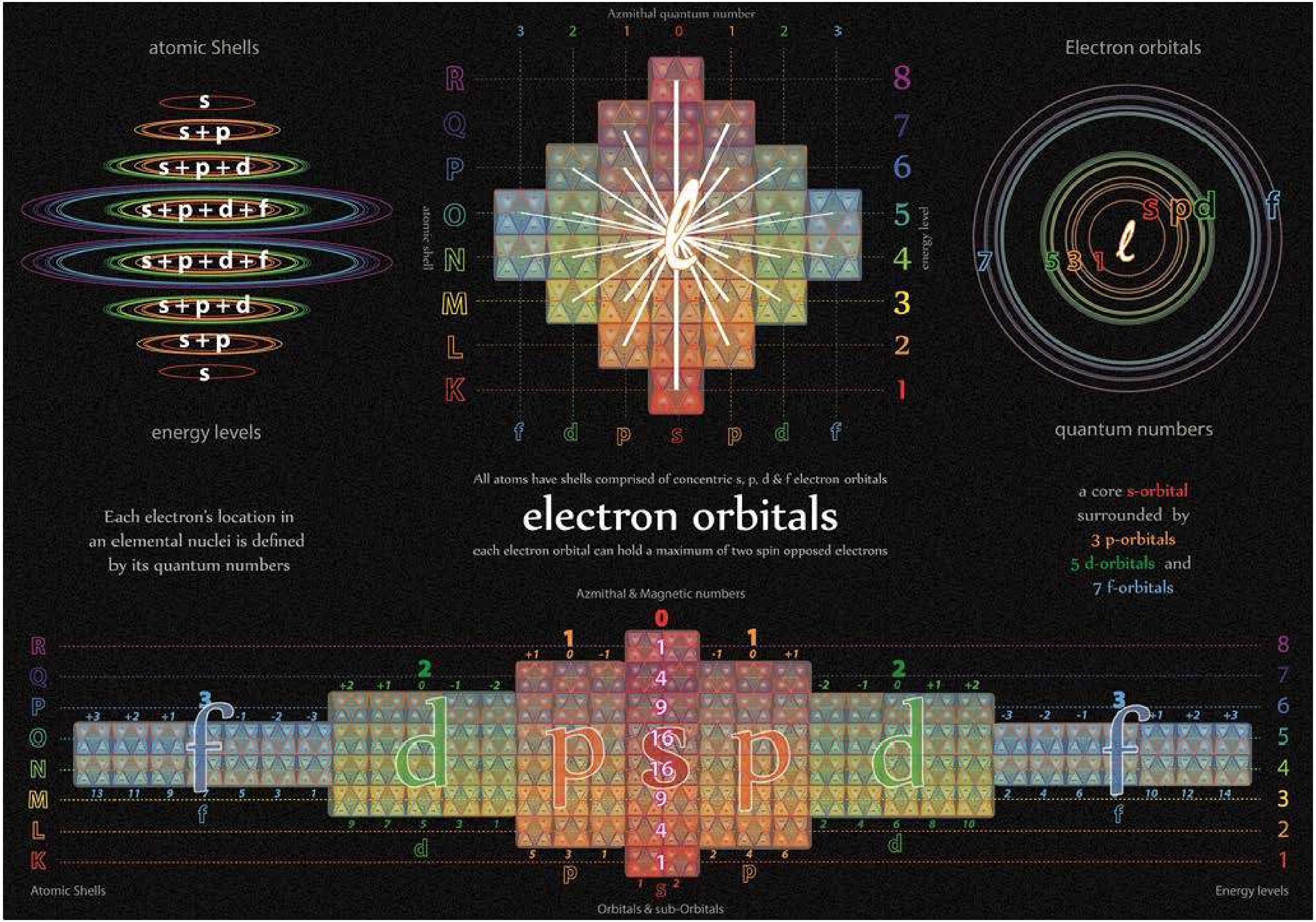

All atoms have shells comprised of concentric s, p, d & f electron orbitals

electron orbitals

each electron orbital can hold a maximum of two spin opposed electrons

Each electron's location in an elemental nuclei is defined by its quantum numbers

a core s-orbital surrounded by 3 p-orbitals 5 d-orbitals and 7 f-orbitals

atomic Shells
s
s+p
s+p+d
s+p+d+f
s+p+d+f
s+p+d
s+p
s
energy levels
Azmithal quantum number
3 2 1 0 1 2 3
R Q P O N M L K
atomic shell
8 7 6 5 4 3 2 1
energy level
ℓ
f d p s p d f
Electron orbitals
s p d f
7 5 3 1
ℓ
quantum numbers
Azmithal & Magnetic numbers
0
1 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2 -2 -1 0 +1 +2
3 3
+3 +2 +1 -1 -2 -3 -3 -2 -1 +1 +2 +3
1 4 9 16 16 9 4 1
f d p s p d f
13 11 9 7 5 3 1
2 4 6 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 s 2
R Q P O N M L K
8 7 6 5 4 3 2 1
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 44.04 - Electron Orbitals

Copyright ABRAHAM [2008] - All rights reserved

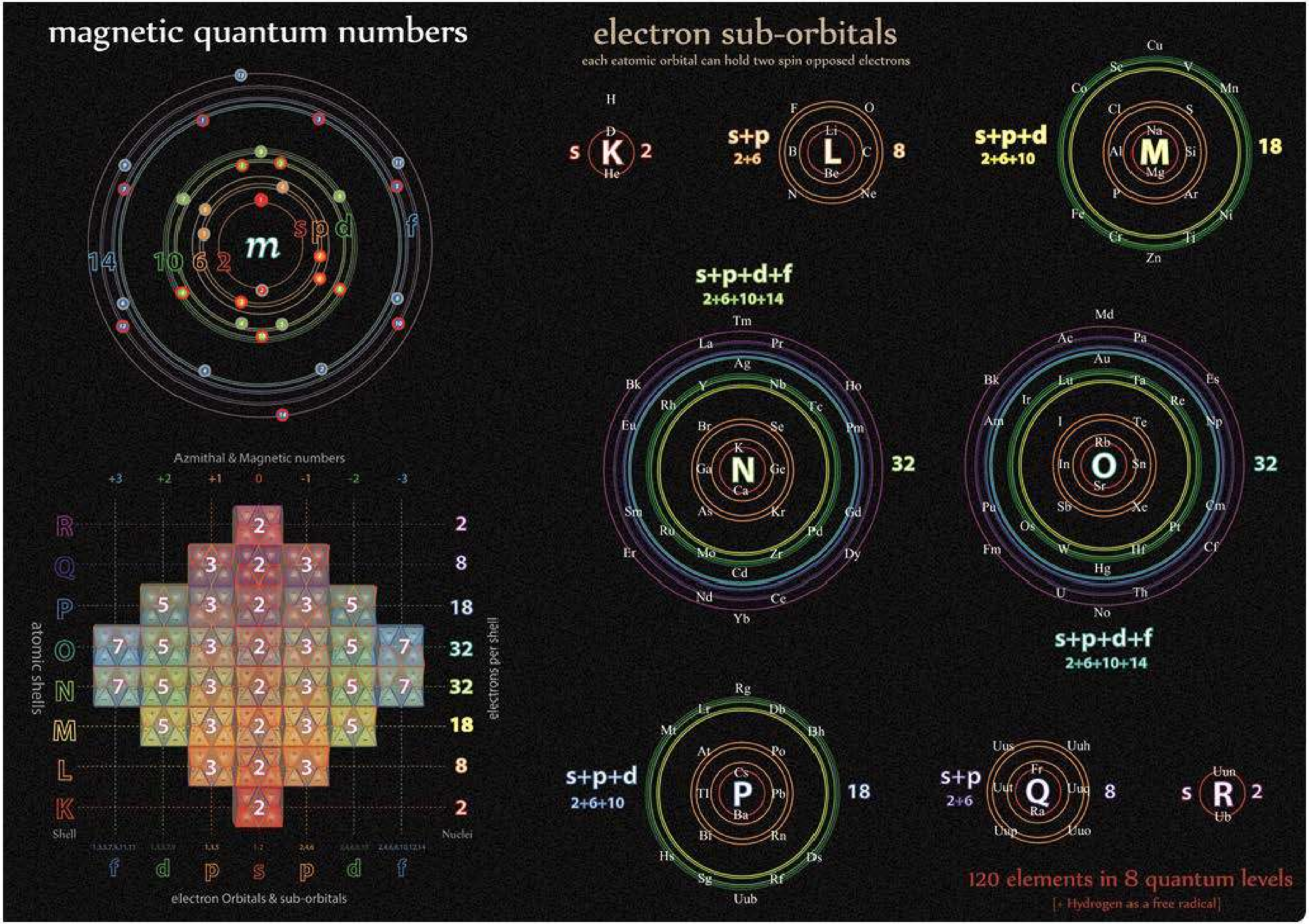
magnetic quantum numbers
m
s p d f
14 10 6 2
Azmithal & Magnetic numbers
+3 +2 +1 0 -1 -2 -3
R Q P O N M L K
atomic shells
Shell
2 8 18 32 32 18 8 2
electrons per shell
Nuclei
f d p s p d f
electron Orbitals & sub-orbitals
electron sub-orbitals
each eatomic orbital can hold two spin opposed electrons
s K 2
s+p 2+6 L 8
s+p+d 2+6+10 M 18
s+p+d+f 2+6+10+14 N 32
s+p+d+f 2+6+10+14 O 32
s+p+d 2+6+10 P 18
s+p 2+6 Q 8
s R 2
120 elements in 8 quantum levels
[+ Hydrogen as a free radical]

Tetryonics 44.05 - Electron sub-Obitals

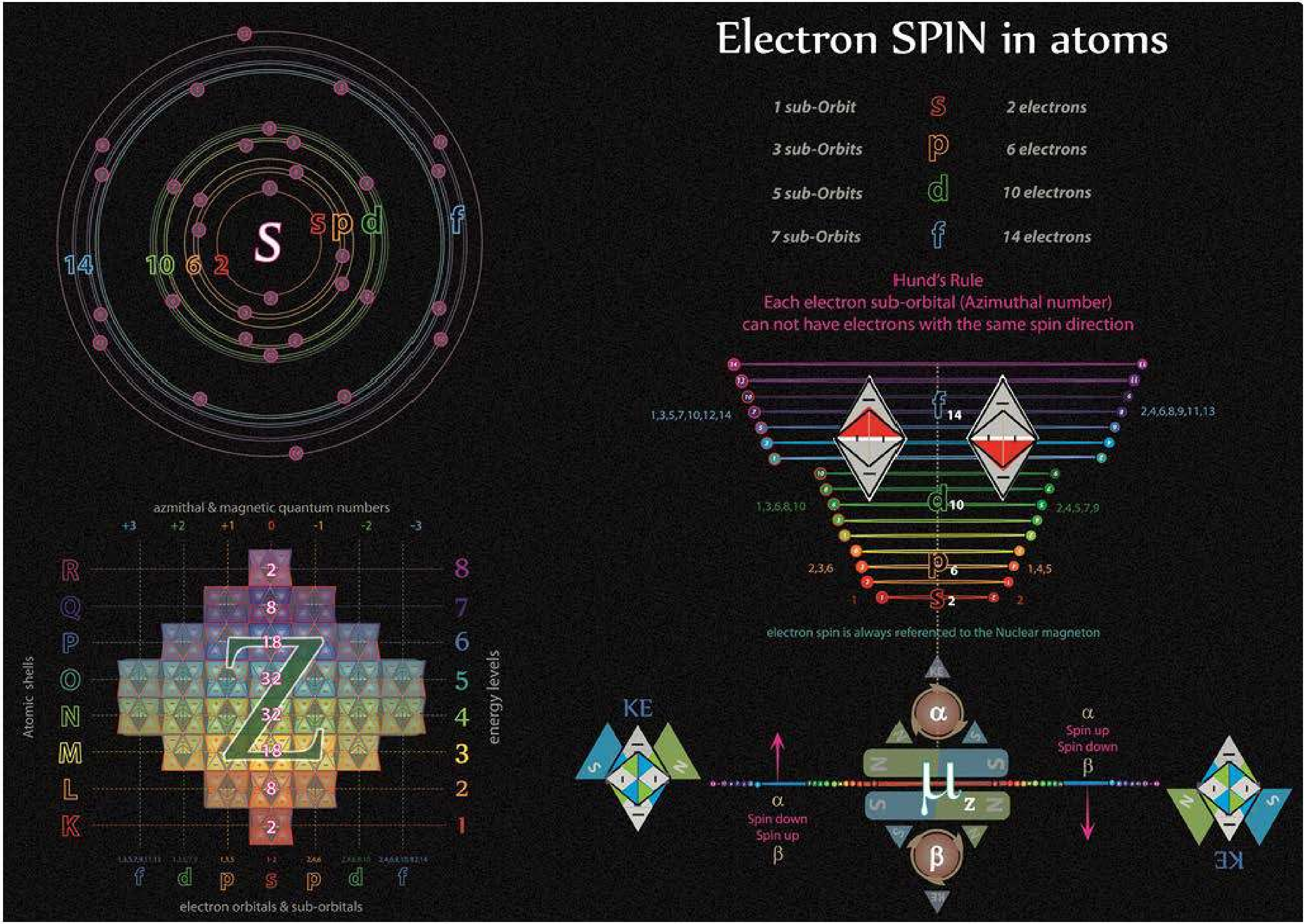

Electron SPIN in atoms

1 sub-Orbit	s	*2 electrons*
3 sub-Orbits	p	*6 electrons*
5 sub-Orbits	d	*10 electrons*
7 sub-Orbits	f	*14 electrons*

Hund's Rule
Each electron sub-orbital (Azimuthal number)
can not have electrons with the same spin direction

electron spin is always referenced to the Nuclear magneton

S
s p d f
14 10 6 2
azmithal & magnetic quantum numbers
+3 +2 +1 0 -1 -2 -3
R Q P O N M L K
Atomic shells
8 7 6 5 4 3 2 1
energy levels
2 8 18 32 32 18 8 2
Z
f d p s p d f
electron orbitals & sub-orbitals
1,3,5,7,10,12,14
f 14
2,4,6,8,9,11,13
1,3,6,8,10
d 10
2,4,5,7,9
2,3,6
p 6
1,4,5
1
s 2
2
KE
α
Spin up
Spin down
β
α
Spin down
Spin up
β
μ z
N S
KE

Tetryonics 44.06 - Electron spin

Quantum numbers

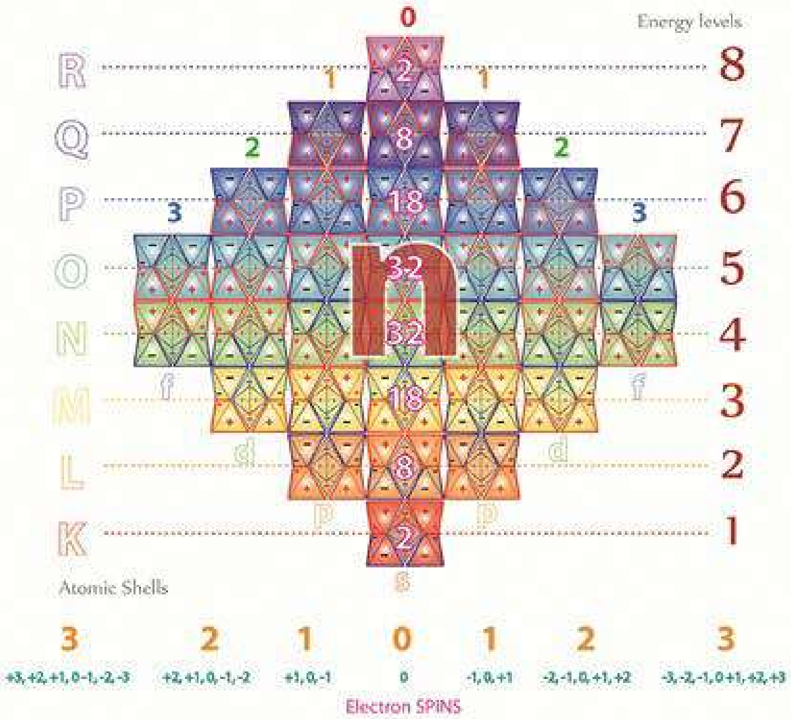
Energy levels
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
2
8
18
32
n
32
18
8
2
f
f
d
d
p
p
s
Atomic Shells
3
2
1
0
1
2
3
+3, +2, +1, 0 -1, -2, -3
+2, +1, 0, -1, -2
+1, 0, -1
0
-1, 0, +1
-2, -1, 0, +1, +2
-3, -2, -1, 0 +1, +2, +3
Electron SPINS

The properties of every atom's specific electron configuration can be described by four quantum numbers:

n (1-8)
Principal
(n = 1, 2, 3, 4 ...)

ℓ (0-3)
Azimuthal
(l = 0, 1 ... n−1)

m (2ℓ+1)
Magnetic
(ml = −l, −l+1 ... 0 ... l−1, l)

s ±½
electron Spin
(ms = −1/2 or +1/2)

The Bohr model was a one-dimensional model that used one quantum number to describe the distribution of electrons

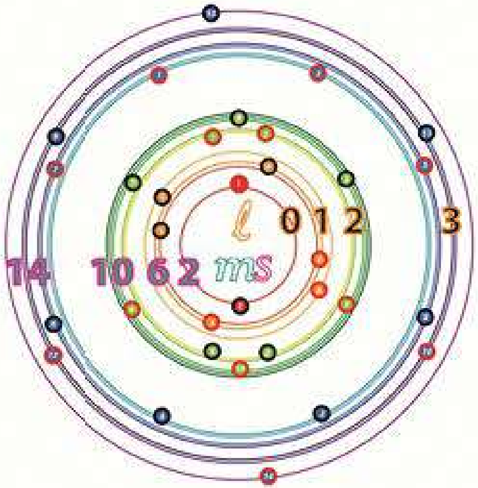
ℓ
0
1
2
3
14
10
6
2
ms

The three coordinates that come from Schrödinger's wave equations are the principal (n), angular (l), and magnetic (m) quantum numbers.

These quantum numbers describe the size, shape, and orientation in space of the orbitals of any particular atom mathematically.

Each electron's quantum numbers are unique and cannot be shared by another electron in that atom.

These four numbers, n, ℓ, m and s can be used to describe any electron in a stable atom [and can be mapped back to the classical terminology of Shells and electron orbitals].

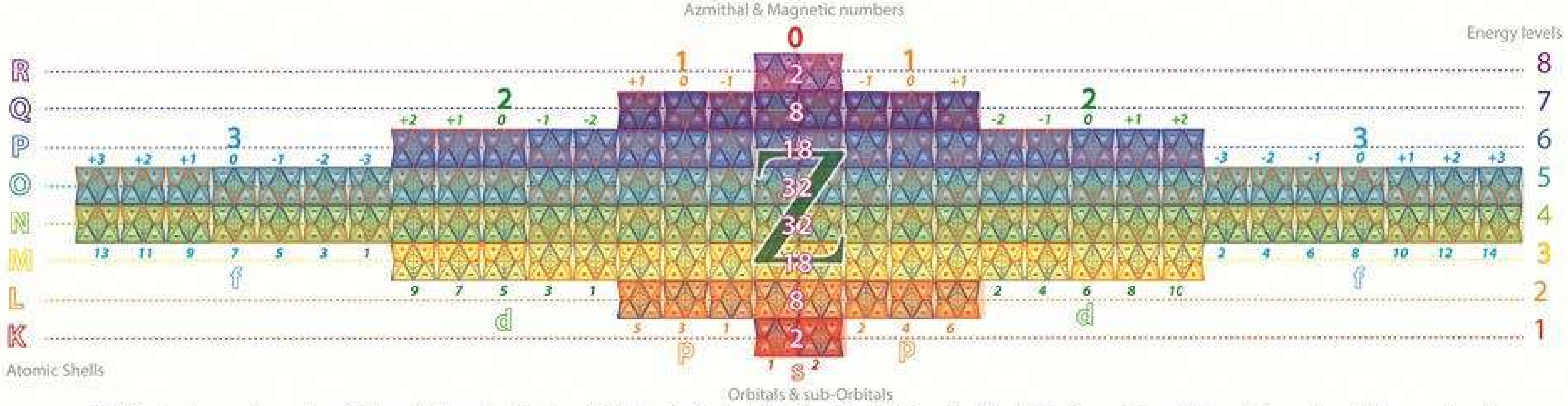
Azmithal & Magnetic numbers
Energy levels
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
2
8
18
32
Z
32
18
8
2
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f
f
d
d
p
p
s
Atomic Shells
Orbitals & sub-Orbitals

A stable atom has equal numebrs of Protons, electrons [and Neutrons], all following the Pauli Exclusion Principle thus orientating their spins so that each element has a unique electron configuration

Copyright ABRAHAM [2008] - All rights reserved

PRINCIPAL quantum number

The first describes the electron shell, or energy level, of an atom.

This is the only quantum number introduced by the Bohr model

The principal quantum number can only have positive integer values

atomic shells

n(1-8)

(n = 1, 2, 3, 4 ...)

energy levels

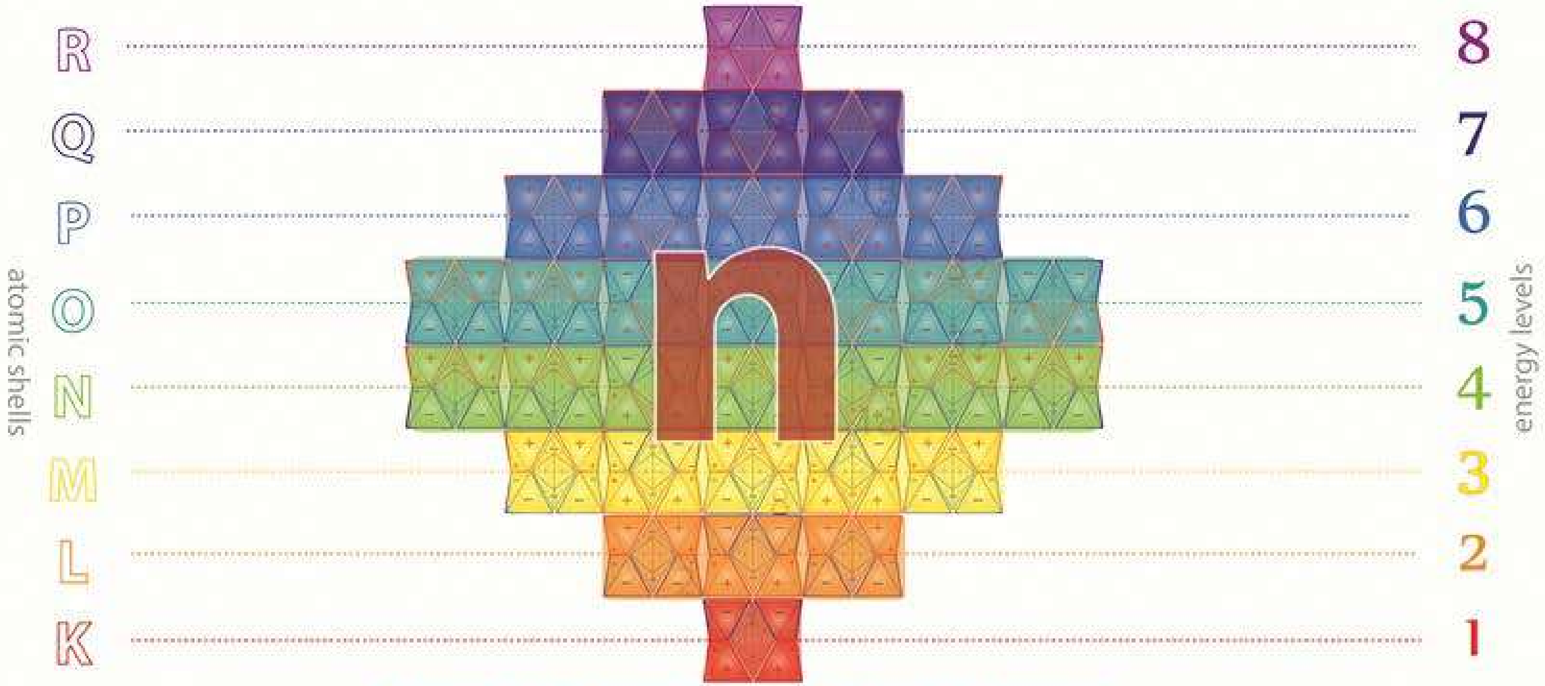

As energies of the Baryons comprising the atomic nuclei increases, the electron bound to each nuclei also possesses more KEM field energies and is therefore less tightly bound to the nucleus

Atomic Shells

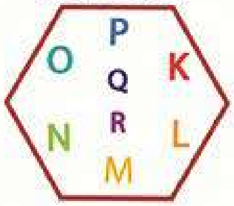

(K, L, M, N, O, P, Q, R)

Atomic shells relate directly to Principal quantum numbers

(1, 2, 3, 4, 5, 6, 7, 8)

Energy levels

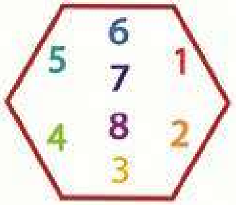

Copyright ABRAHAM [2008] - All rights reserved

AZMITHAL quantum number

The azimuthal quantum number is a quantum number assigned to any atomic orbital that describes its orbital angular momentum and determines the shape of the electron orbital

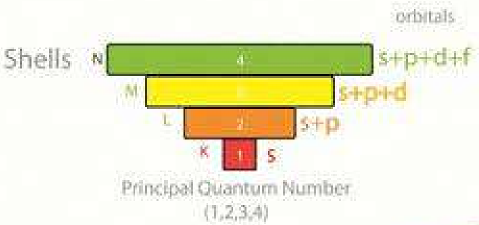
orbitals
Shells
N
4
s+p+d+f
M
3
s+p+d
L
2
s+p
K
1
s
Principal Quantum Number
(1,2,3,4)

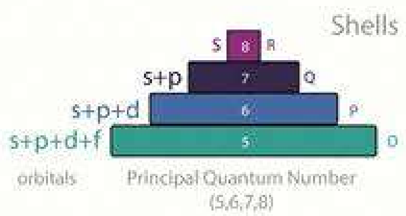
Shells
S
8
R
s+p
7
Q
s+p+d
6
P
s+p+d+f
5
O
orbitals
Principal Quantum Number
(5,6,7,8)

s p d f

electron orbitals

$(l = 0, 1 \dots n-1)$

ℓ (0-7)
Azimuthal

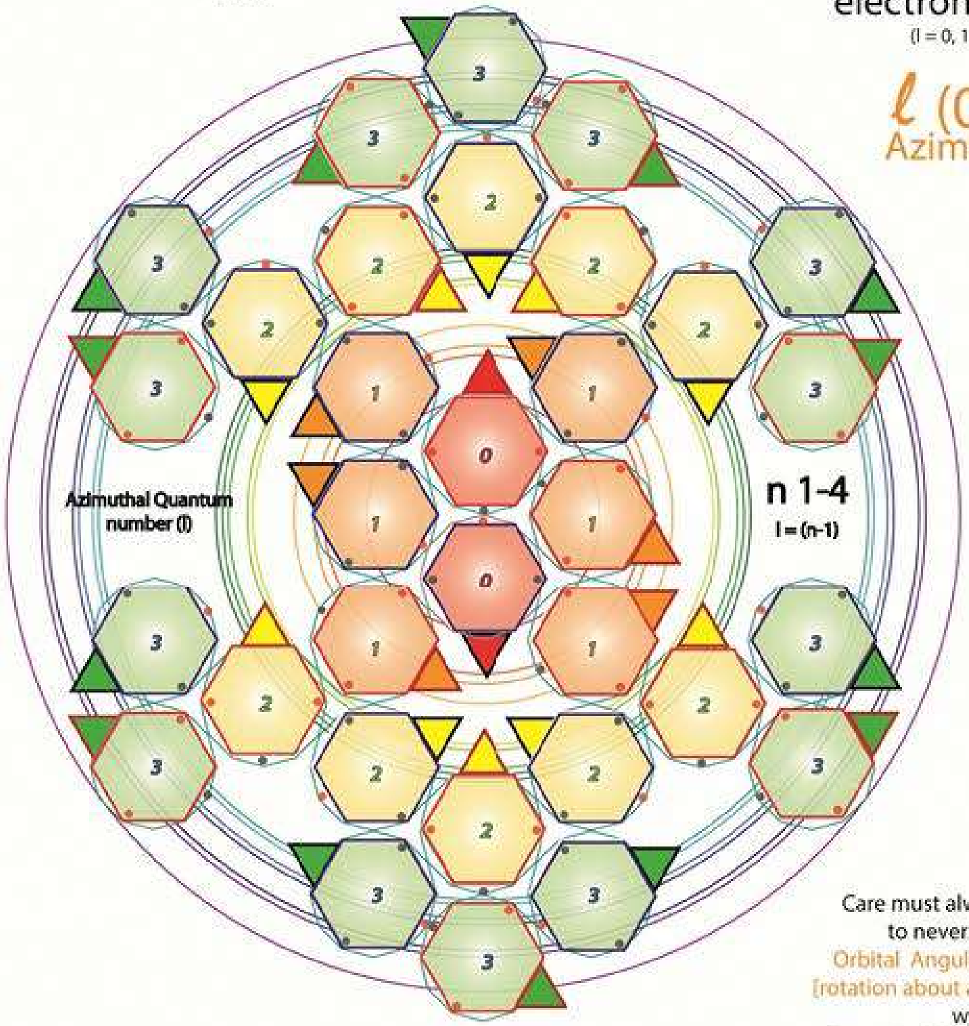
Azimuthal Quantum number (l)
n 1-4
l = (n-1)
0
1
2
3

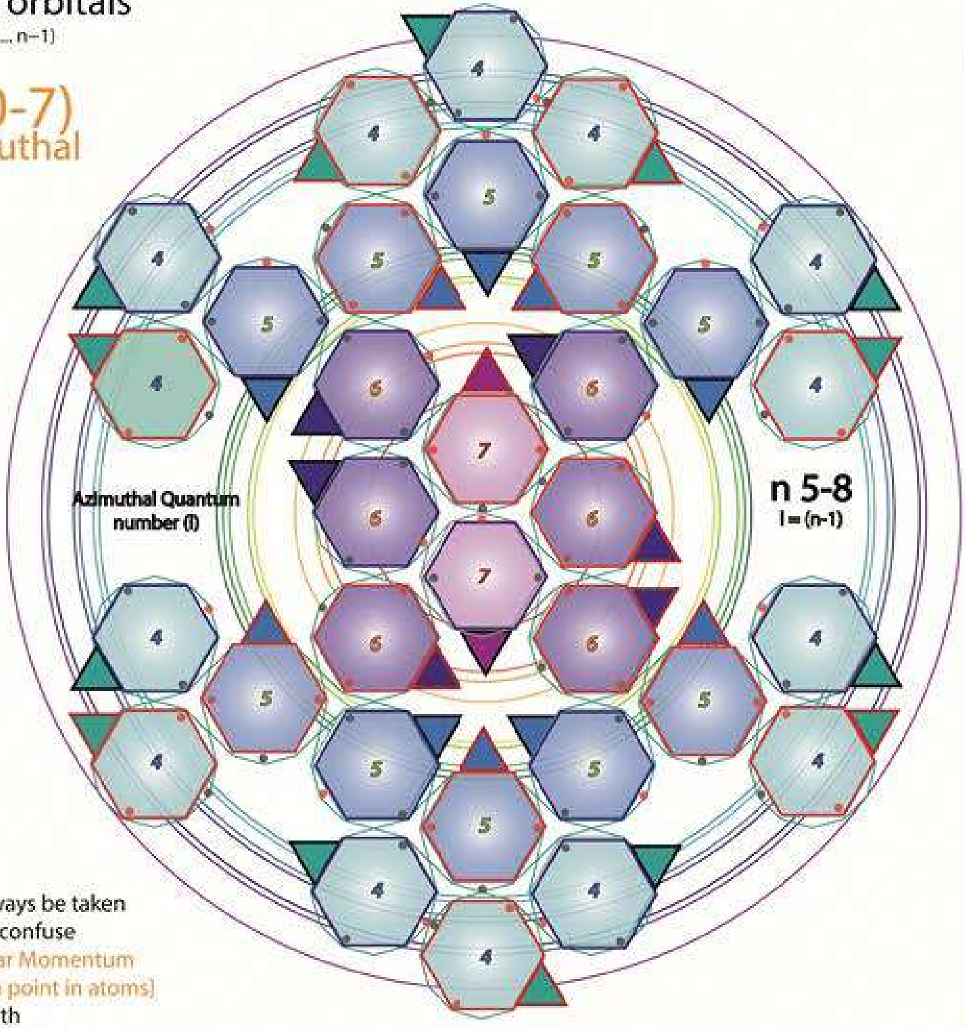
Azimuthal Quantum number (l)
n 5-8
l = (n-1)
4
5
6
7

Care must always be taken
to never confuse
Orbital Angular Momentum
[rotation about a point in atoms]
with
Quantised Angular Momenta
[equilateral Planck energy geometries]

MAGNETIC quantum number

The magnetic quantum number denotes the energy levels available within any subshell

Magnetic numbers do not continue to increase as the Principal numbers increase instead they reverse after n4 to reflect the charged quanutm geometry of Elements and do not follow the current computer models in popular use

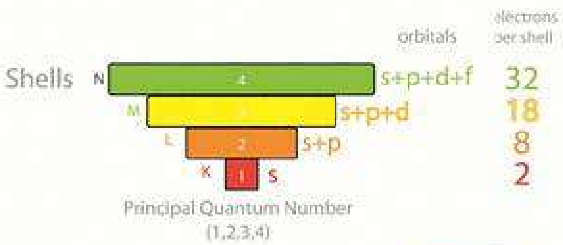
Shells
orbitals
electrons per shell
N
4
s+p+d+f
32
M
3
s+p+d
18
L
2
s+p
8
K
1
s
2
Principal Quantum Number
(1,2,3,4)

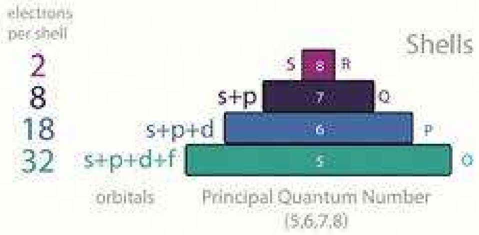
electrons per shell
Shells
2
S
8
R
8
s+p
7
Q
18
s+p+d
6
P
32
s+p+d+f
5
O
orbitals
Principal Quantum Number
(5,6,7,8)

s p d f

electron sub-orbitals

(ml = –l, –l+1 ... 0 ... l–1, l)

$m\ell(2\ell+1)$

Magnetic

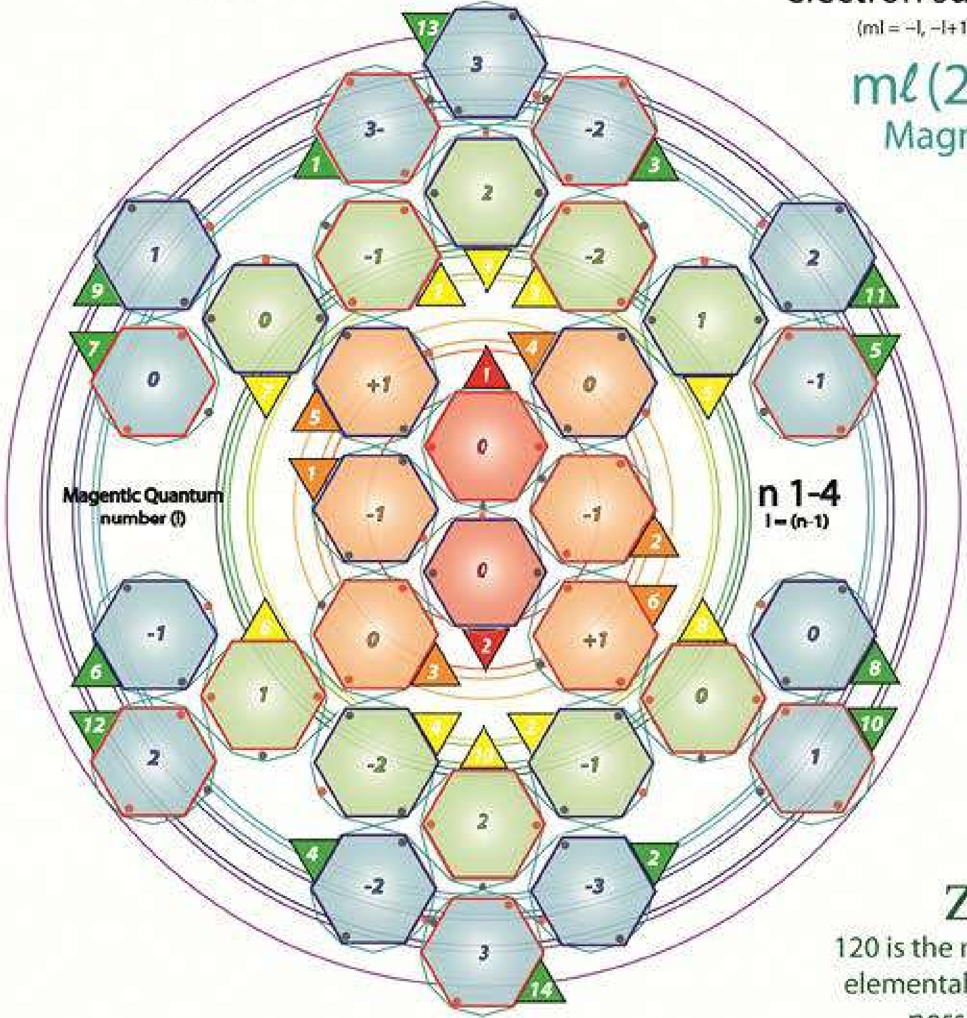
Magentic Quantum number (l)
n 1-4
l – (n-1)

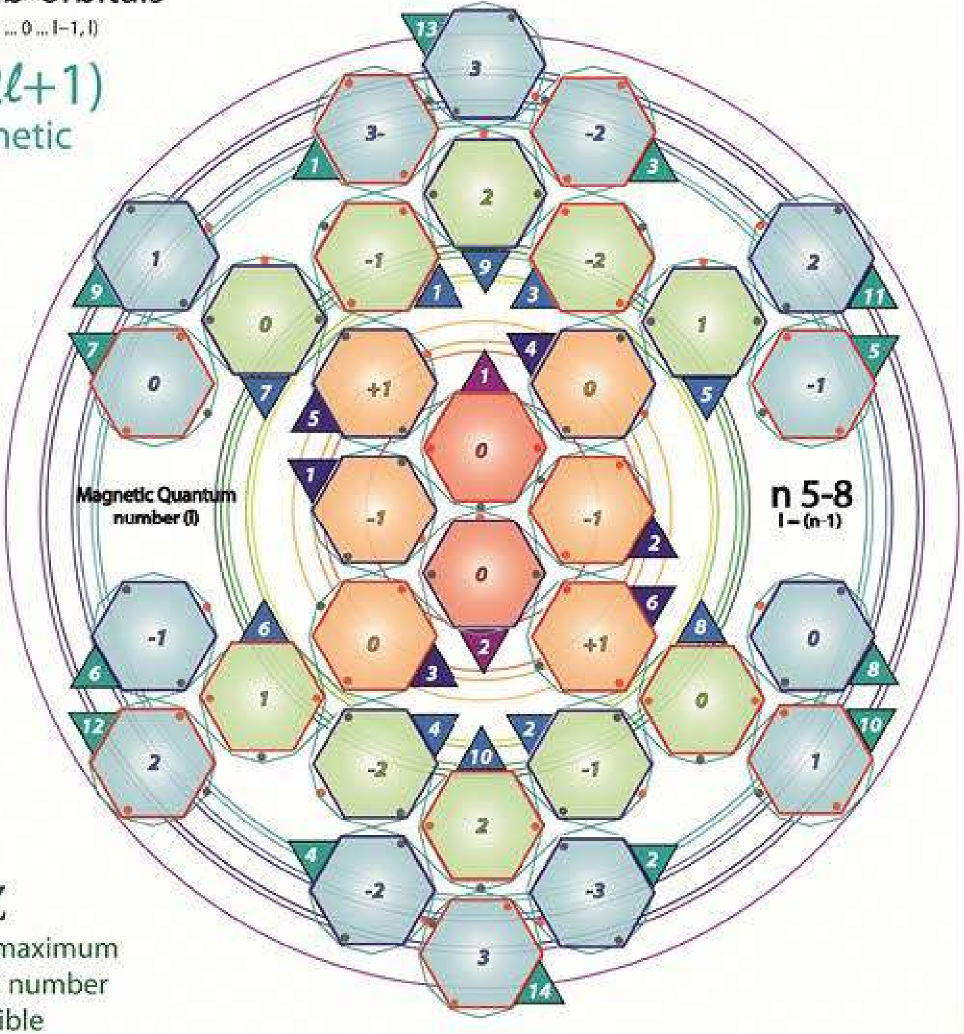
Magentic Quantum number (l)
n 5-8
l – (n-1)

Z

120 is the maximum elemental number possible

Copyright ABRAHAM [2008] - All rights reserved

SPIN quantum number

The spin quantum number is a quantum number that parameterizes
the intrinsic angular momentum (or spin angular momentum),
of any given electron anywhere in an atomic nucleus

parallel magnetic moments

Spin UP

higher coupling energies

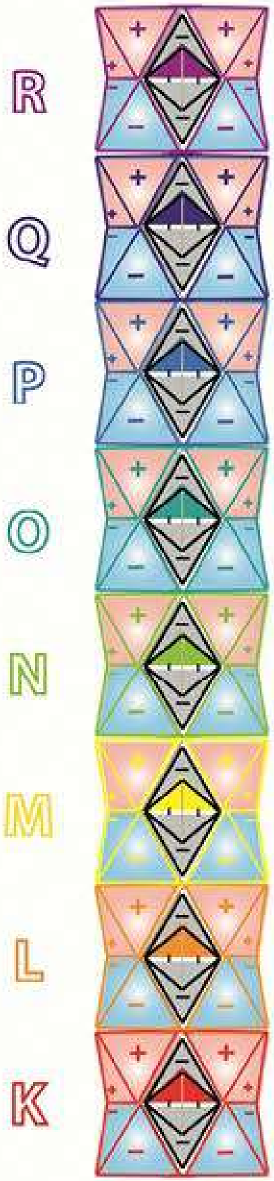

anti-parallel magnetic moments

Spin DOWN

lower coupling energies

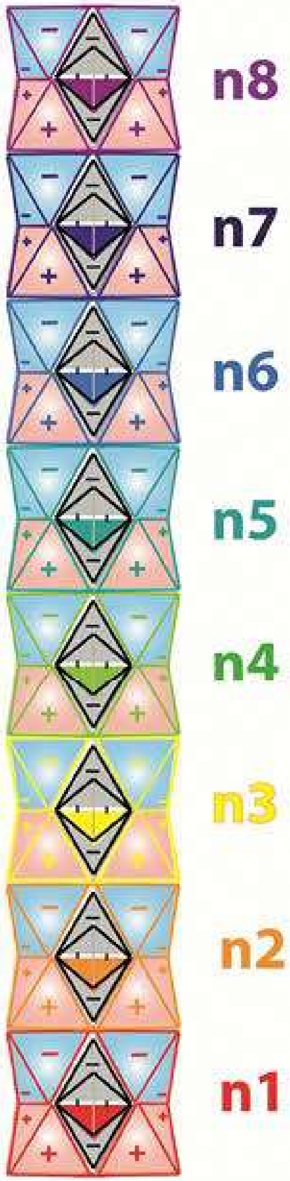
n8
n7
n6
n5
n4
n3
n2
n1
−
+

**Electron spin can orientate in either direction within Nuclei,
providing the nett spins follow the Hund rule and Pauli exclusion principle**

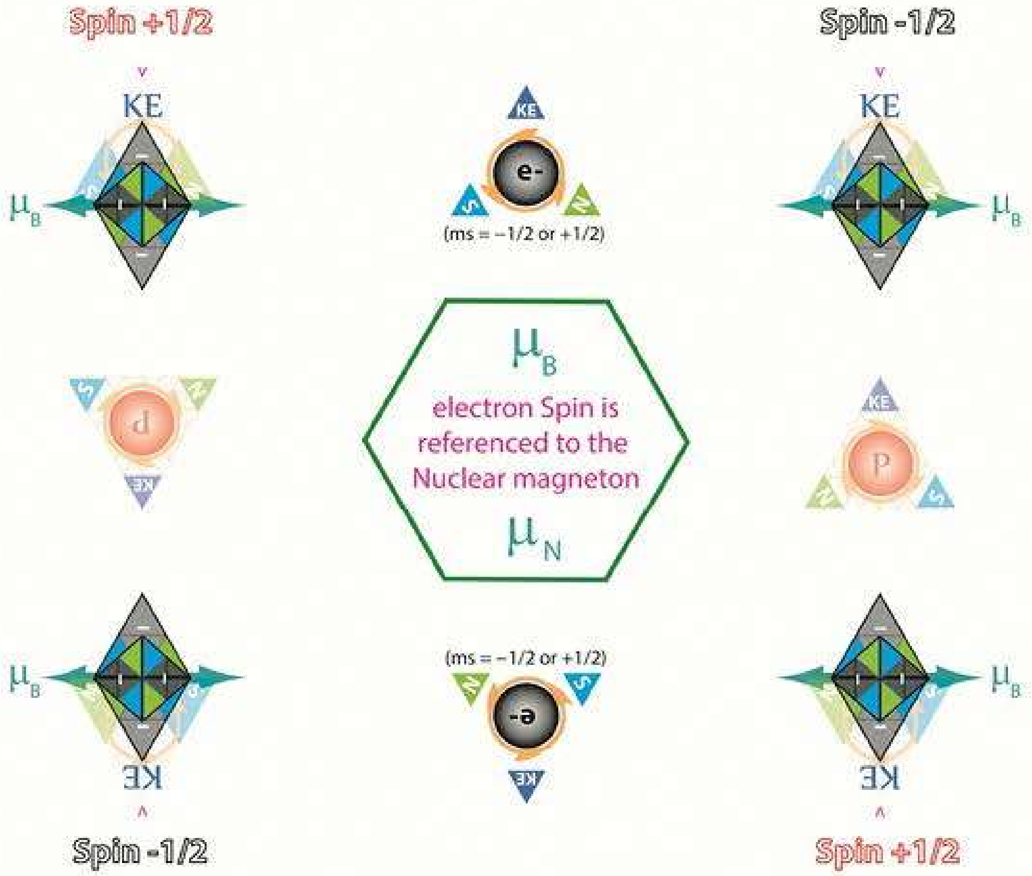
Spin +1/2
Spin -1/2
KE
μ_B
e-
(ms = −1/2 or +1/2)
d
p
KE
μ_B
μ_N
(ms = −1/2 or +1/2)
-e
Spin -1/2
Spin +1/2

electron Spin is
referenced to the
Nuclear magneton

**The nuclear energy levels of the Baryons comprising Elemental nuclei
determine the energy-momenta of electrons bound to them**

Copyright ABRAHAM [2008] - All rights reserved

Modifying Hund's Rule

Electrons fill orbitals in an alternating sequential numbering pattern due to nucleon placement creating opposed direction electron spins

The increased stability of the atom, most commonly manifested in a lower energy state, arises because the high-spin state forces the unpaired electrons to reside in different spatial orbitals.

A commonly given reason for the increased stability of high multiplicity states is that the different occupied spatial orbitals create a larger average distance between electrons, reducing electron-electron repulsion energy. In reality, it has been shown that the actual reason behind the increased stability is a decrease in the screening of electron-nuclear attractions[1].

The total spin state is calculated as the total number of unpaired electrons + 1, or twice the total spin + 1 written as 2s+1.

As a result of Hund's rule, constraints are placed on the way atomic orbitals are filled using the Aufbau principle.

Before any two electrons occupy an orbital in a subshell, other orbitals in the same subshell must first each contain one electron. Also, the electrons filling a subshell will have parallel spin before the shell starts filling up with the opposite spin electrons (after the first orbital gains a second electron).

As a result, when filling up atomic orbitals, the maximum number of unpaired electrons (and hence maximum total spin state) is assured.

s	p	d	f
▲1 1 2▼	▼1 1 2▲	▲1 1 2▼	▲ 1 1 2 ▼
	▲3 2 4▼	▲3 2 4▼	▲ 3 2 4 ▼
	▼5 3 6▲	▼5 3 6▲	▲ 5 3 6 ▼
		▼7 4 8▲	▲ 7 4 8 ▼
		▼9 5 10▲	▼ 9 5 10 ▲
			▼ 11 6 12 ▲
			▼ 13 7 14 ▲

▲

Hund's rule of orbital filling
must be modified to reflect
the true orbital filling order

▼

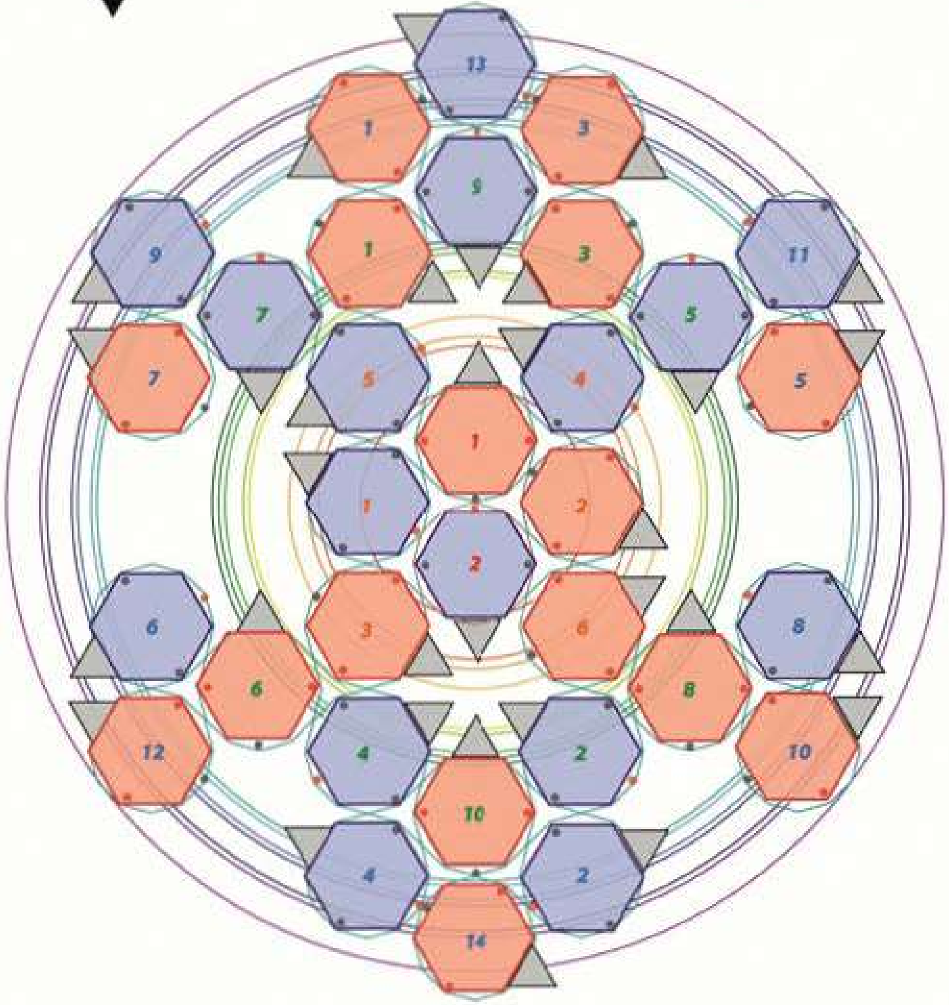

Sub-orbitals fill in order of numbering

Electrons spins pair before next orbital is filled

ie.
p1 [DOWN] and p2 [UP] fill
before
p3 [UP] and p4 [DOWN]
before
p5 [DOWN] and p6 [UP] etc

Principle quantum Energies

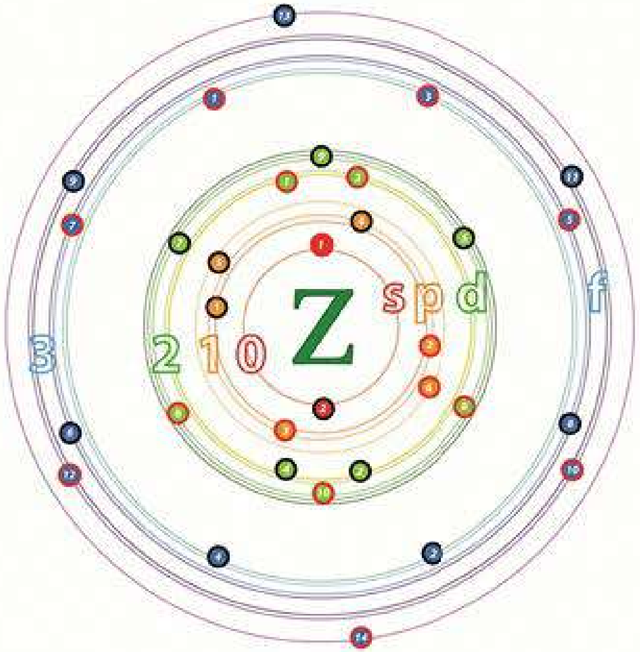

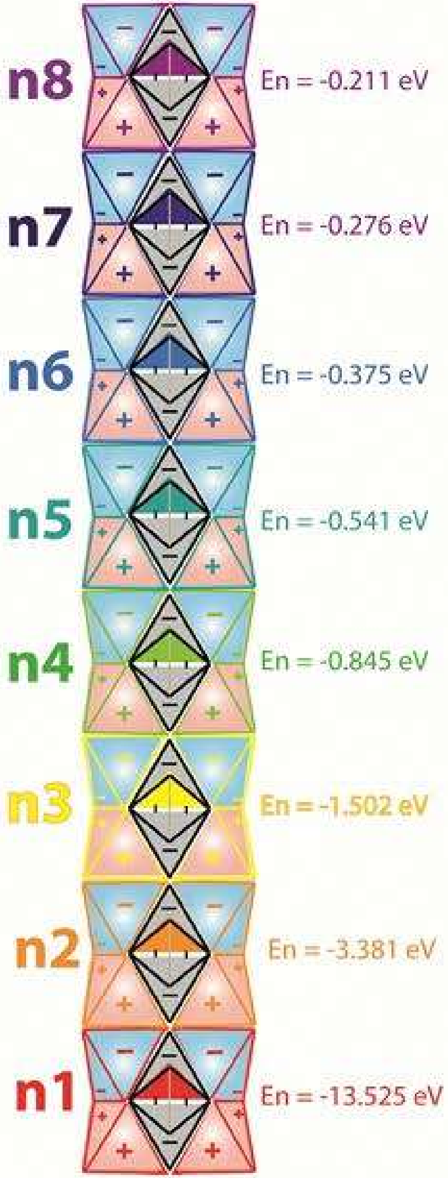

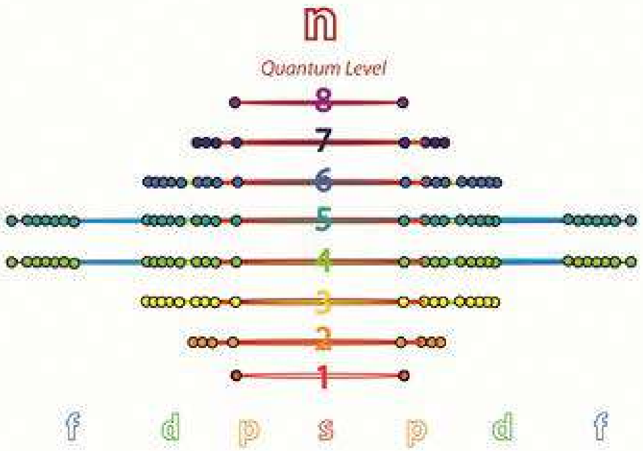

s orbits	n1-8
p orbits	n2-7
d orbits	n3-6
f orbits	n4-5

In an atom - electron energies are proportional to their intrinsic Kinetic Energies - which in turn are directly proportional to the quantum energy level of the nuclei which the electron binds to in their respective atomic shells

In a nucleus, lower energy orbits have less 'paired' nuclei supplying energy. The more energy you give a nuclei the faster it casuses the bound electron to rotate. If you give the nuclei enough energy, it will impart eneough energy to its electron for it to leave the system entirely.

The same is true for an electron orbital. Higher values of n mean more energy for the electron and the corresponding KEM field enegies of the electron is larger, resulting in increased angular momentum.

Values of n start at 1 and go up by integer amounts.

If enough energy is added to the system by incident Photons a electron will leave the atom creating a positively charged nuclei [ionisation].

Eigenstate value

KEM field energy [per n] required to exceed 13.525 eV at which point the photo-electron has sufficient KE to break free of the Nucleus

$$E_n = \frac{E_1}{n^2} = \frac{-13.6eV}{n^2}, n = 1, 2, 3...$$

The possible Kinetic Energies (quantum levels) of an electron are directly related to the energy level of the Nuclei in each Quantum Level

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 1

K shell

n1

Ground State electron

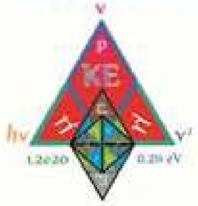
v
P
KE
hv
v'
1.2e20
0.28 eV

$iE = -13.313\ eV$

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
1	1s1	**Deuterium**	1	1	1
2	1s2	**Helium**	2	2	2

2

Deuterium [not Hydrogen] is the building block of elements

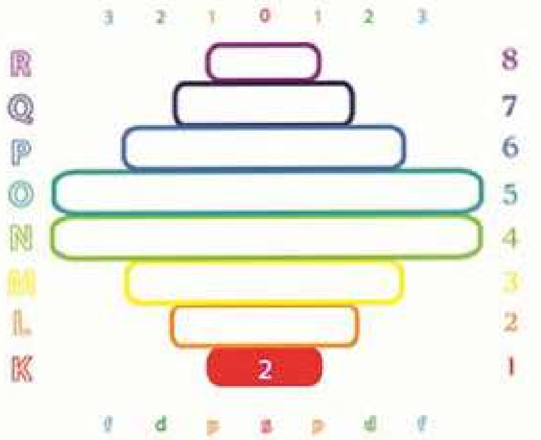
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
f d p s p d f

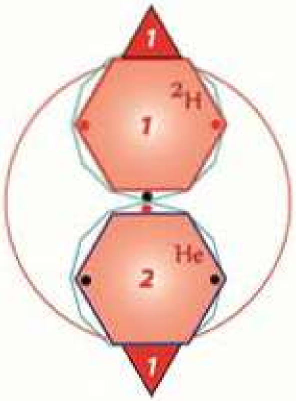

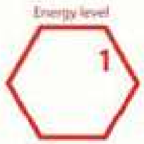
Energy level
1

Z #		Name	Protons	electrons	Neutrons
1	1s1	**Hydrogen**	1	1	0

Hydrogen is a free radical element

n1

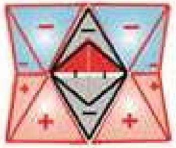

Quantum Level 2

The energy levels of bound electrons is determined by Baryons

Z#		Name	Protons	electrons	Neutrons
3	2s1	Lithium	3	3	3
4	2s2	Beryllium	4	4	4
5	1p1	Boron	5	5	5
6	1p2	Carbon	6	6	6
7	1p3	Nitrogen	7	7	7
8	1p4	Oxygen	8	8	8
9	1p5	Fluorine	9	9	9
10	1p6	Neon	10	10	10

8

Deuterium [not Hydrogen] is the building block of elements

L shell

n2

Ground State electron

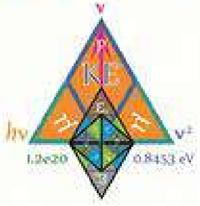
v
KE
hv
v^2
1.2e20
0.8453 eV

$iE = -12.679\ eV$

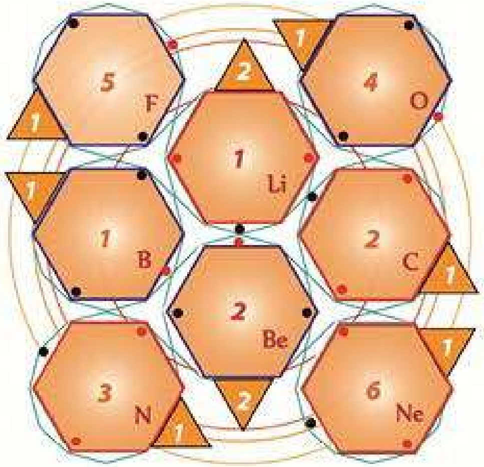
5
F
2
1
4
O
1
1
Li
1
1
B
2
C
1
2
Be
1
3
N
2
1
6
Ne

Energy level
2

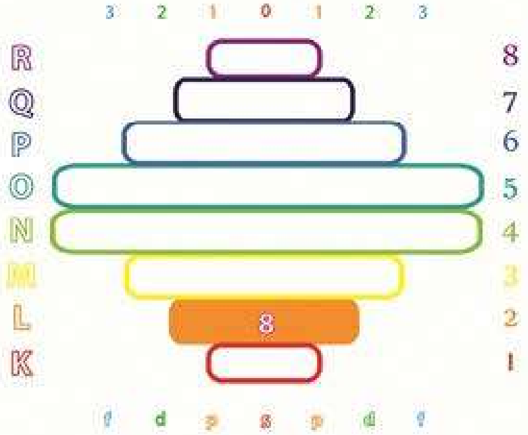
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
8
f
d
p
s
p
d
f

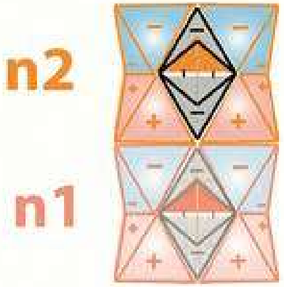

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 3

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
11	3s1	Sodium	11	11	11
12	3s2	Magnesium	12	12	12
13	2p1	Aluminium	13	13	13
14	2p2	Silicon	14	14	14
15	2p3	Phosphorus	15	15	15
16	2p4	Sulfur	16	16	16
17	2p5	Chlorine	17	17	17
18	2p6	Argon	18	18	18
21	3d1	Scandium	21	21	21
22	3d2	Titanium	22	22	22
23	3d3	Vanadium	23	23	23
24	3d4	Chromium	24	24	24
25	3d5	Manganese	25	25	25
26	3d6	Iron	26	26	26
27	3d7	Cobalt	27	27	27
28	3d8	Nickel	28	28	28
29	3d9	Copper	29	29	29
30	3d10	Zinc	30	30	30

18

Deuterium [not Hydrogen] is the building block of elements

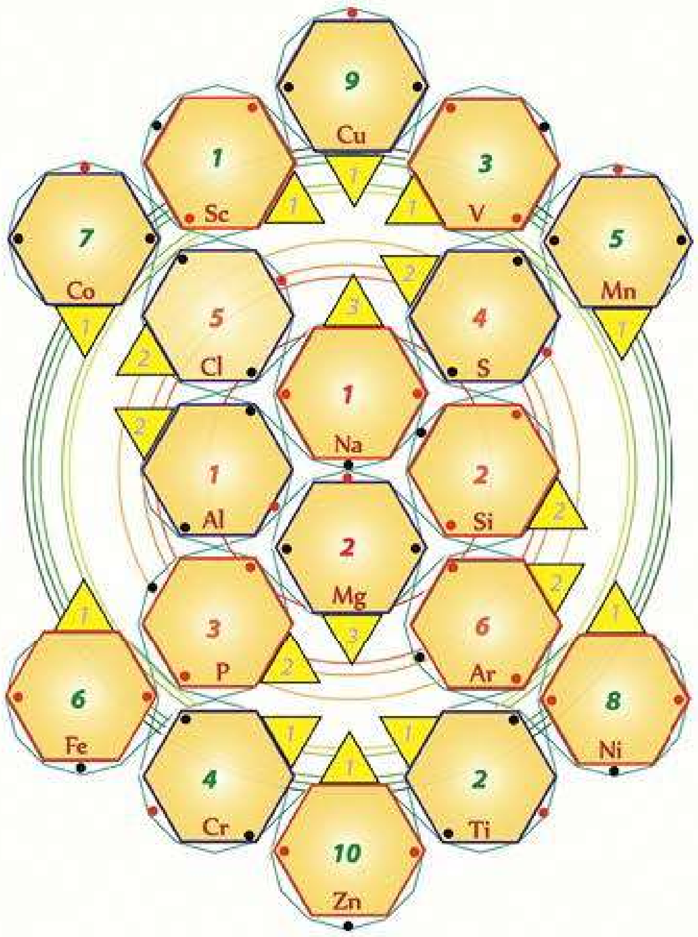
9
Cu
1
Sc
3
V
7
Co
5
Mn
5
Cl
4
S
1
Na
1
Al
2
Si
2
Mg
3
P
6
Ar
6
Fe
8
Ni
4
Cr
2
Ti
10
Zn
1
2
3

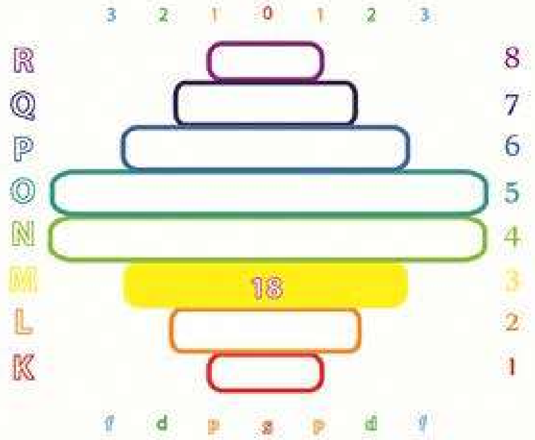
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
18
f
d
p
s
p
d
f

M shell

n3

Ground State electron

v
KE
hv
v²
1.2e20
1.9008 eV

$iE = -11.623\ eV$

Energy level

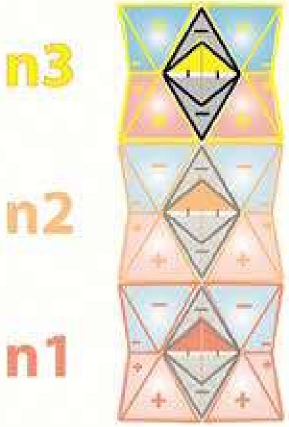

Quantum Level 4

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
19	4s1	Potassium	19	19	19
20	4s2	Calcium	20	20	20
31	4p1	Gallium	31	31	31
32	4p2	Germanium	32	32	32
33	4p3	Arsenic	33	33	33
34	4p4	Selenium	34	34	34
35	4p5	Bromine	35	35	35
36	4p6	Krypton	36	36	36
39	4d1	Yttrium	39	39	39
40	4d2	Zirconium	40	40	40
41	4d3	Niobium	41	41	41
42	4d4	Molybdenum	42	42	42
43	4d5	Technetium	43	43	43
44	4d6	Ruthenium	44	44	44
45	4d7	Rhodium	45	45	45
46	4d8	Palladium	46	46	46
47	4d9	Silver	47	47	47
48	4d10	Cadmium	48	48	48
57	4f1	Lanthanum	57	57	57
58	4f2	Cerium	58	58	58
59	4f3	Praseodymium	59	59	59
60	4f4	Neodymium	60	60	60
61	4f5	Promethium	61	61	61
62	4f6	Samarium	62	62	62
63	4f7	Europium	63	63	63
64	4f8	Gadolinium	64	64	64
65	4f9	Terbium	65	65	65
66	4f10	Dysprosium	66	66	66
67	4f11	Holium	67	67	67
68	4f12	Erbium	68	68	68
69	4f13	Thulium	69	69	69
70	4f14	Ytterbium	70	70	70

32

Deuterium [not Hydrogen] is the building block of elements

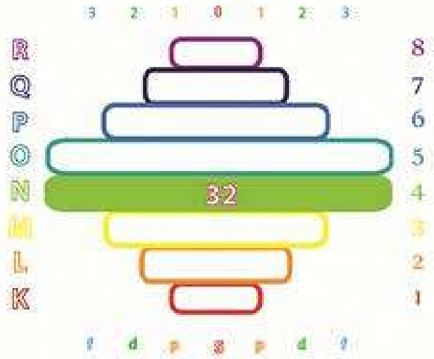
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
32
f d p s p d f

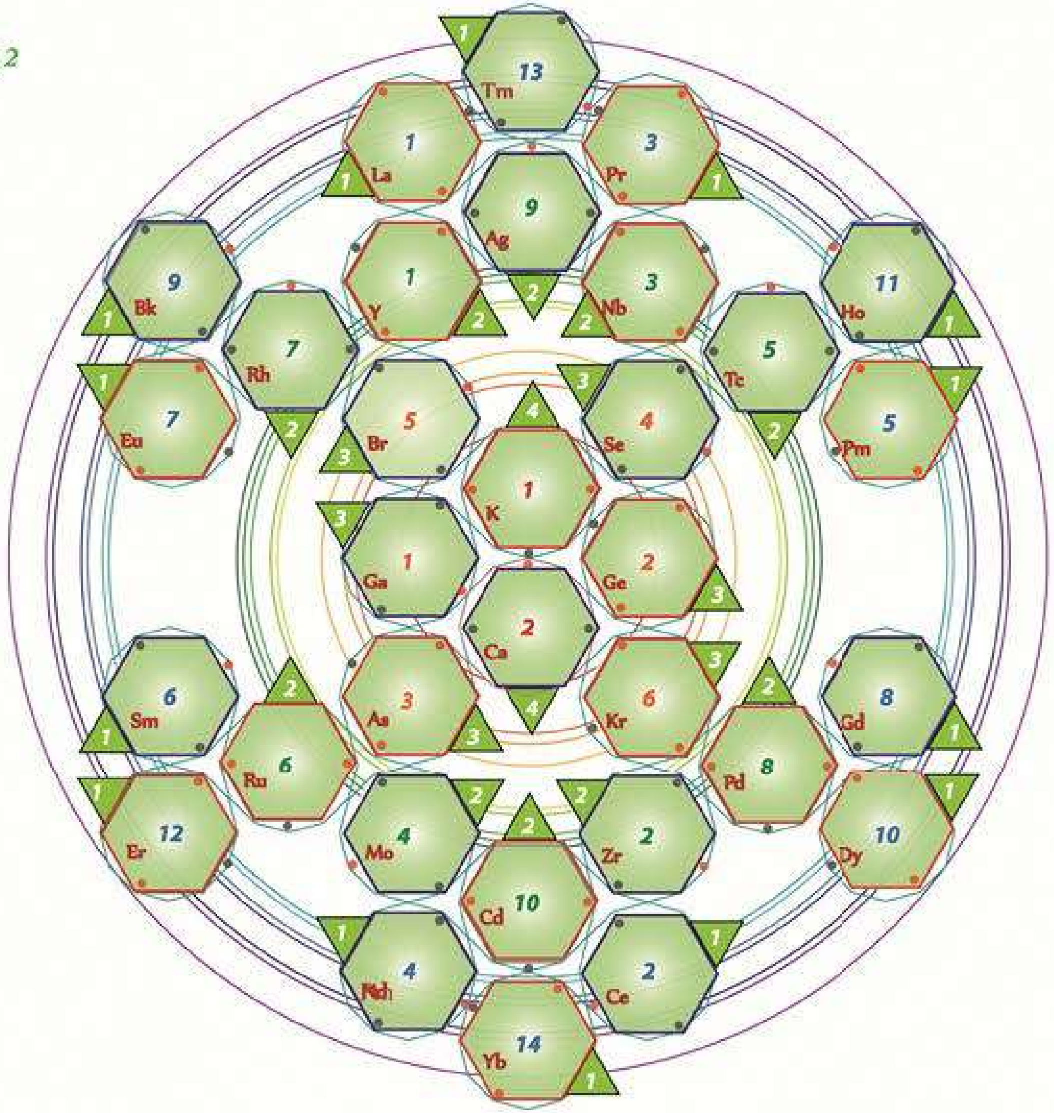
13
Tm
1
La
3
Pr
9
Ag
1
Y
3
Nb
9
Bk
11
Ho
7
Rh
5
Tc
7
Eu
5
Pm
5
Br
4
Se
1
K
1
Ga
2
Ge
2
Ca
6
Sm
8
Gd
3
As
6
Kr
6
Ru
8
Pd
12
Er
10
Dy
4
Mo
2
Zr
10
Cd
4
Nd
2
Ce
14
Yb

N shell

n4

Ground State electron

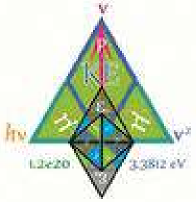
hv
v²
3.3812 eV

$iE = -10.143\ eV$

Energy level
4

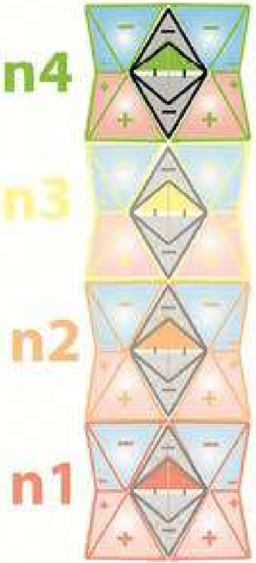

Quantum Level 5

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
37	5s1	Rubidium	37	37	37
38	5s2	Strontium	38	38	38
49	5p1	Indium	49	49	49
50	5p2	Tin	50	50	50
51	5p3	Antimony	51	51	51
52	5p4	Tellurium	52	52	52
53	5p5	Iodine	53	53	53
54	5p6	Xenon	54	54	54
71	5d1	Lutetium	71	71	71
72	5d2	Hafnium	72	72	72
73	5d3	Tantalum	73	73	73
74	5d4	Tungsten	74	74	74
75	5d5	Rhenium	75	75	75
76	5d6	Osmium	76	76	76
77	5d7	Iridium	77	77	77
78	5d8	Platinum	78	78	78
79	5d9	Gold	79	79	79
80	5d10	Mercury	80	80	80
89	5f1	Actinium	89	89	89
90	5f2	Thorium	90	90	90
91	5f3	Protactium	91	91	91
92	5f4	Uranium	92	92	92
93	5f5	Neptunium	93	93	93
94	5f6	Plutonium	94	94	94
95	5f7	Americium	95	95	95
96	5f8	Curium	96	96	96
97	5f9	Berkelium	97	97	97
98	5f10	Californium	98	98	98
99	5f11	Einsteinium	99	99	99
100	5f12	Fermium	100	100	100
101	5f13	Mendelevium	101	101	101
102	5f14	Nobelium	102	102	102

32

Deuterium [not Hydrogen] is the building block of elements

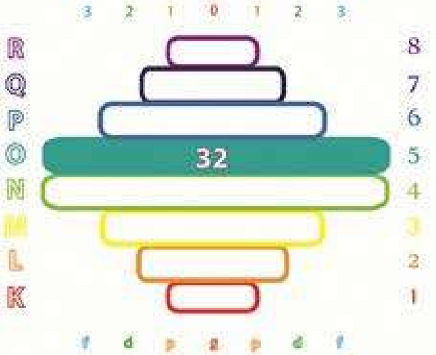

3 2 1 0 1 2 3
R 8
Q 7
P 6
O 32 5
N 4
M 3
L 2
K 1
f d p s p d f

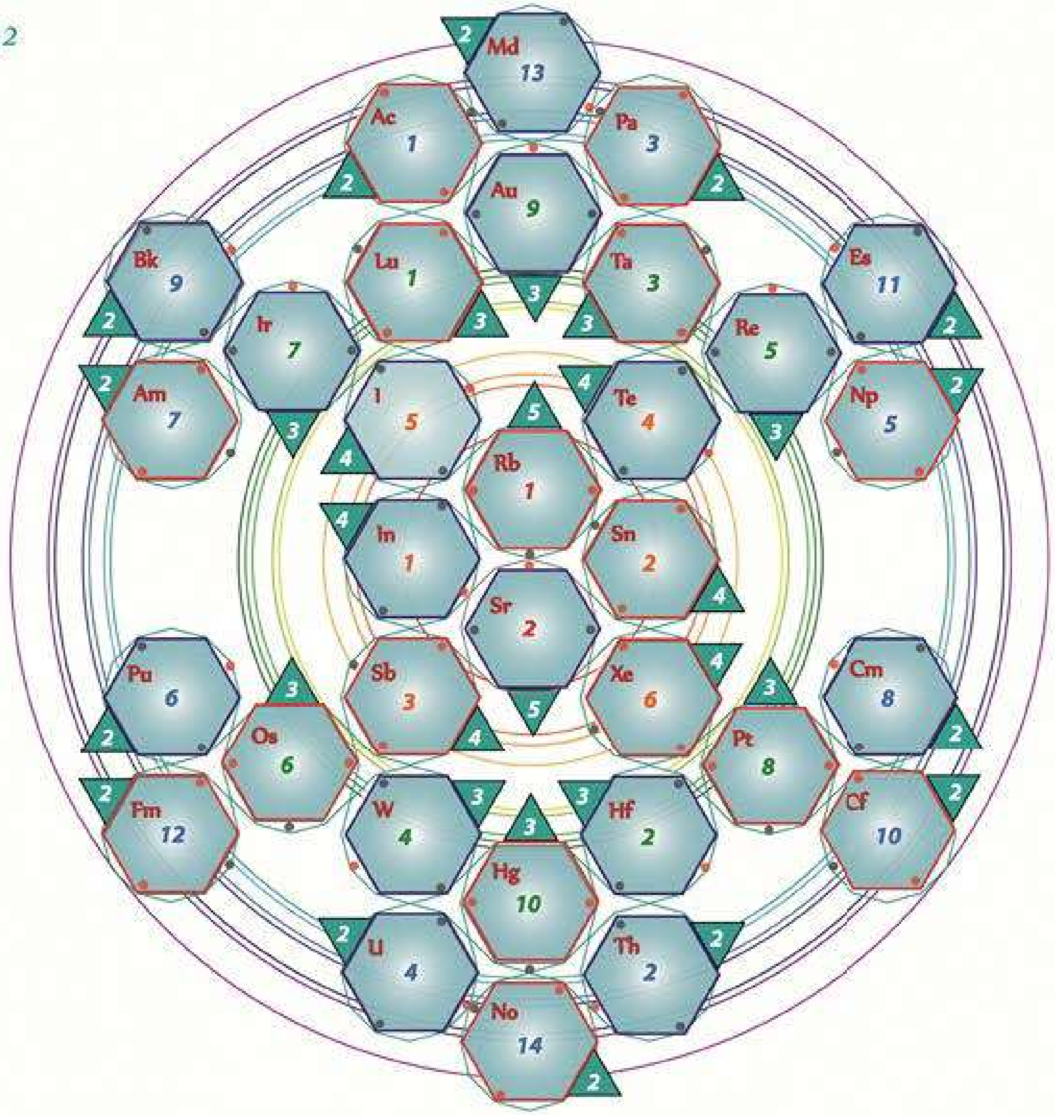

Md 13
Ac 1
Pa 3
Au 9
Bk 9
Lu 1
Ta 3
Es 11
Ir 7
Re 5
Am 7
I 5
Te 4
Np 5
Rb 1
In 1
Sn 2
Sr 2
Pu 6
Sb 3
Xe 6
Cm 8
Os 6
Pt 8
Fm 12
W 4
Hf 2
Cf 10
Hg 10
U 4
Th 2
No 14

O shell

n5

Ground State electron

v
hv
v²
1.2e20
3.5282 eV

$iE = -8.241\ eV$

Energy level

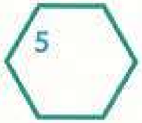

n5
n4
n3
n2
n1

Quantum Level 6

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
55	6s1	Caesium	55	55	55
56	6s2	Barium	56	56	56
81	6p1	Thallium	81	81	81
82	6p2	Lead	82	82	82
83	6p3	Bismuth	83	83	83
84	6p4	Polonium	84	84	84
85	6p5	Astatine	85	85	85
86	6p6	Radon	86	86	86
103	6d1	Lawrencium	103	103	103
104	6d2	Rutherfordium	104	104	104
105	6d3	Dubnium	105	105	105
106	6d4	Seaborgium	106	106	106
107	6d5	Bohrium	107	107	107
108	6d6	Hassium	108	108	108
109	6d7	Meitnerium	109	109	109
110	6d8	Darmstadtium	110	110	110
111	6d9	Roetgenium	111	111	111
112	6d10	Copernicium	112	112	112

18

Deuterium [not Hydrogen] is the building block of elements

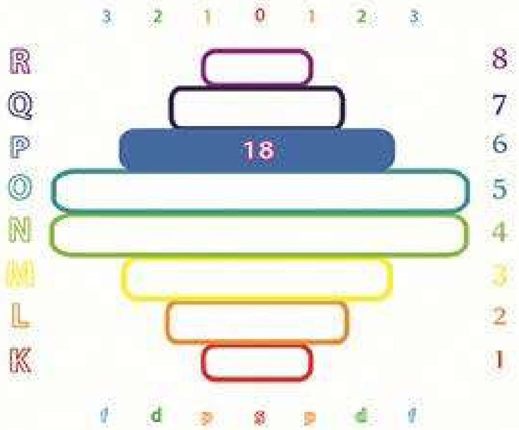
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
18
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Energy level
6

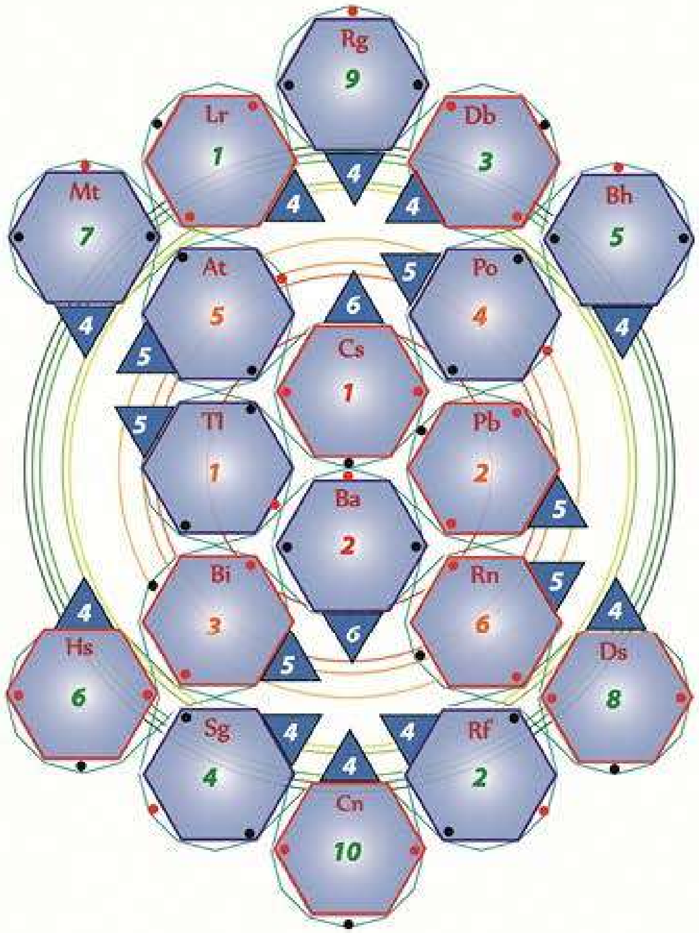
Rg 9
Lr 1
Db 3
Mt 7
Bh 5
At 5
Po 4
Cs 1
Tl 1
Pb 2
Ba 2
Bi 3
Rn 6
Hs 6
Ds 8
Sg 4
Rf 2
Cn 10
4
5
6

P shell

n6

Ground State electron

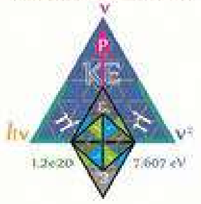
P
KE
hv
v²
1.2e20
7.607 eV

$iE = -5.917\ eV$

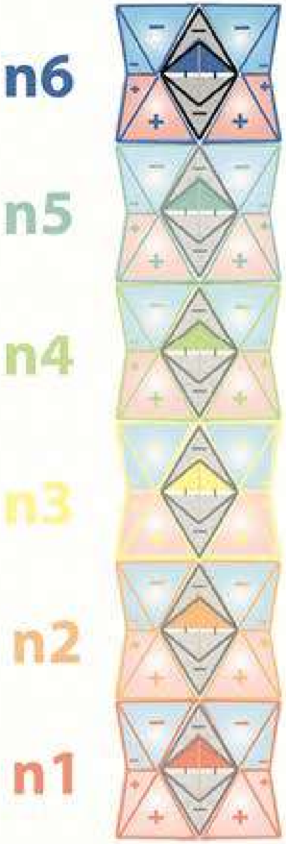
n6
n5
n4
n3
n2
n1

Quantum Level 7

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
87	7s1	Francium	87	87	87
88	7s2	Radium	88	88	88
113	7p1	Ununtrium	113	113	113
114	7p2	Flerovium	114	114	114
115	7p3	Ununpentium	115	115	115
116	7p4	Livermorium	116	116	116
117	7p5	Ununseptium	117	117	117
118	7p6	Ununoctium	118	118	118

8

Deuterium [not Hydrogen] is the building block of elements

Q shell

n7

Ground State electron

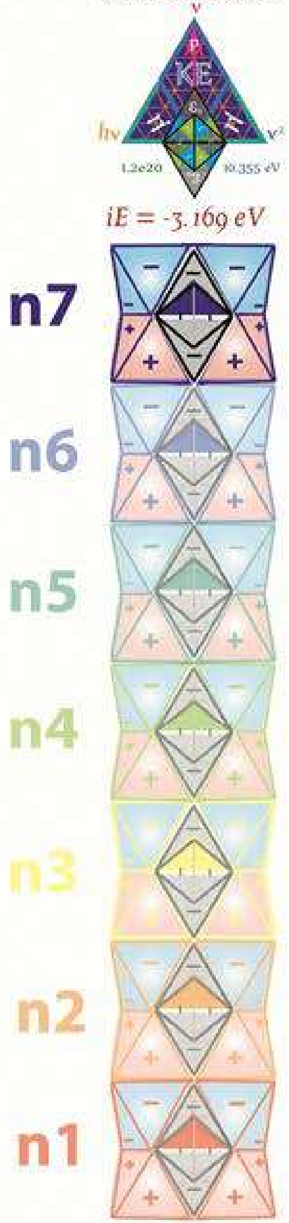
v
P
KE
hv
v²
1.2e20
10.355 eV
iE = -3.169 eV
n7
n6
n5
n4
n3
n2
n1
−
+

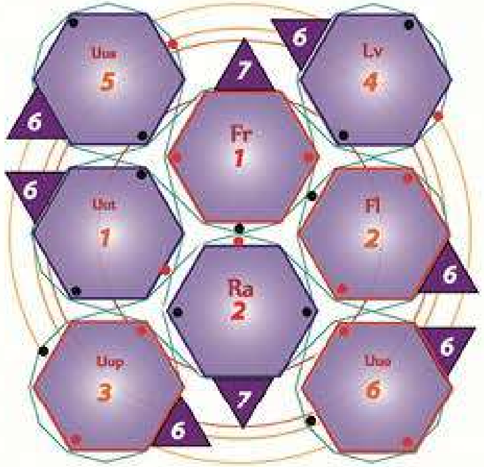
Uus
5
Lv
4
Fr
1
Uut
1
Fl
2
Ra
2
Uup
3
Uuo
6
7
6

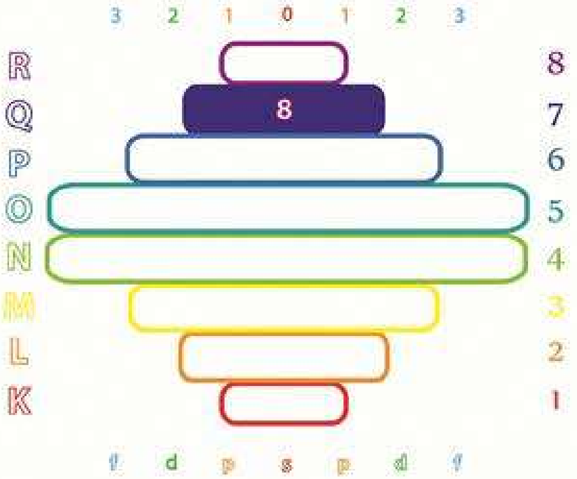
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Energy level
7

Quantum Level 8

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
119	8s1	Ununnonium	119	119	119
120	8s2	Unbinilium	120	120	120

2

Deuterium [not Hydrogen] is the building block of elements

R shell

n8

Ground State electron

v
P
KE
hv
v²
1.2e20
13.525 eV

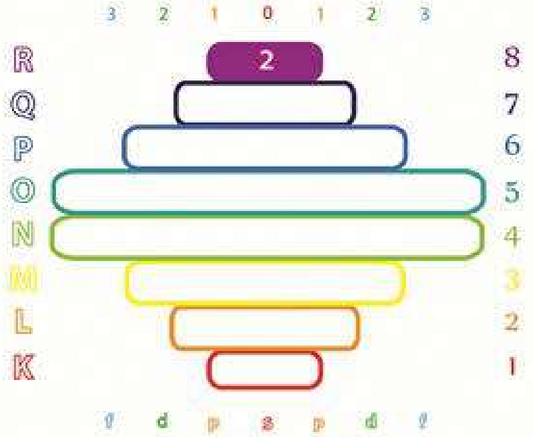
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

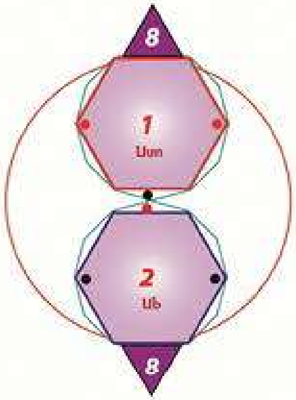

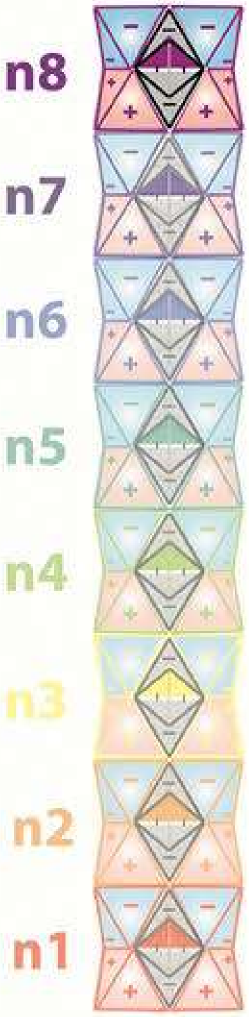
n8
n7
n6
n5
n4
n3
n2
n1
−
+

Energy level

Spectral line transitions

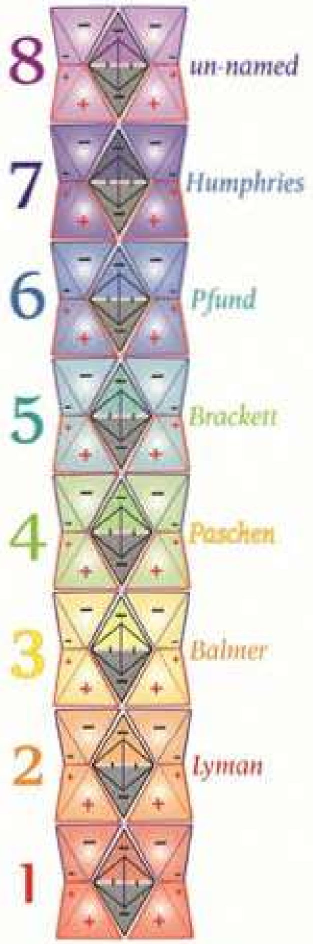
8
un-named
7
Humphries
6
Pfund
5
Brackett
4
Paschen
3
Balmer
2
Lyman
1

Quantum level jumps

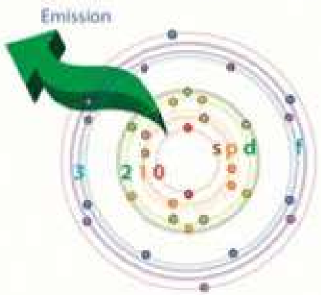
Emission
s p d
f
3 2 1 0

Quantum Level Jumps

Photon Absorption and Release

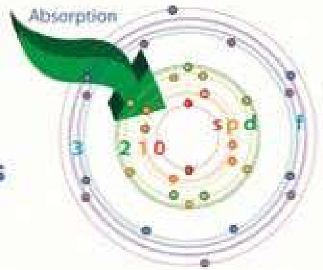
Absorption
s p d
f
3 2 1 0

Photo-electrons can only transition between principal energy Baryons in the atomic nuclei in discrete steps [or quantum jumps] because Baryons determine the KEM energy levels of electrons in nuclei

[see Tetryonic QED for full details on spectral line mechanics]

Initial Quantum level

8	768	756	720	660	576	468	336	180	0
7	588	576	540	480	396	288	156	0	
6	432	420	384	324	240	132	0		
5	300	288	252	192	108	0			
4	192	180	144	84	0				
3	108	96	60	0					
2	48	36	0						
1	12	0							
		12	48	108	192	300	432	588	768
		1	2	3	4	5	6	7	8

Final Quantum level

$$\Delta p = \Delta M v = hf$$

accelerating photo-electrons produce spectral lines

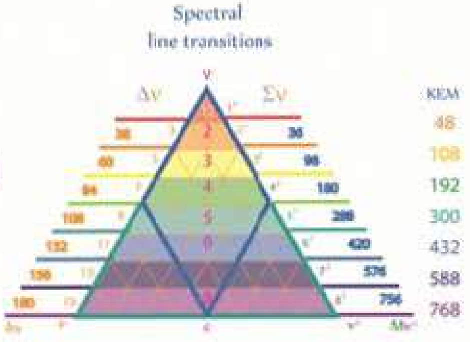
Spectral
line transitions
v
Δv
Σv
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
36
96
180
288
420
576
756

Quantum differentials

hv, R, μ, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Nuclear energy emission~absorption

If atoms are left undisturbed, their electrons usually fill the lowest available energy levels and stay there, in their "ground state."

Occasionally, however, they may also be pushed up to some higher energy ("become excited") e.g. by a collision with a fast atom or electron, one which got extra speed from an electric voltage or from some source of heat.

An atom/electron elevated to one of its higher "excited levels" soon falls back to a lower level ("undergoes a quantum jump"), emitting a photon whose energy corresponds to the difference between the levels.

That need not be the ground state: the atom/electron might descend to that state in several steps, emitting a photon at each step on the way.

Velocity
v
p
Spectral
Lines
KEM
Energies
KE
hv
v²
M
12
1.2e20

Tetryonics 46.10 - Quantum level jumps

Quantum transitions

(Orbital Shells - Bound energy states)

Quantum levels 1-8 are also referred to as Atomic Shells K-R

Photo-electrons can only transition in discrete steps [or quantum jumps] within atomic nuclei shells because Baryons determine the KEM energy levels of electrons in nuclei

Any photo-electron bound in a Deuterium nuclei will have specific quantised KEM field energies and angular momenta

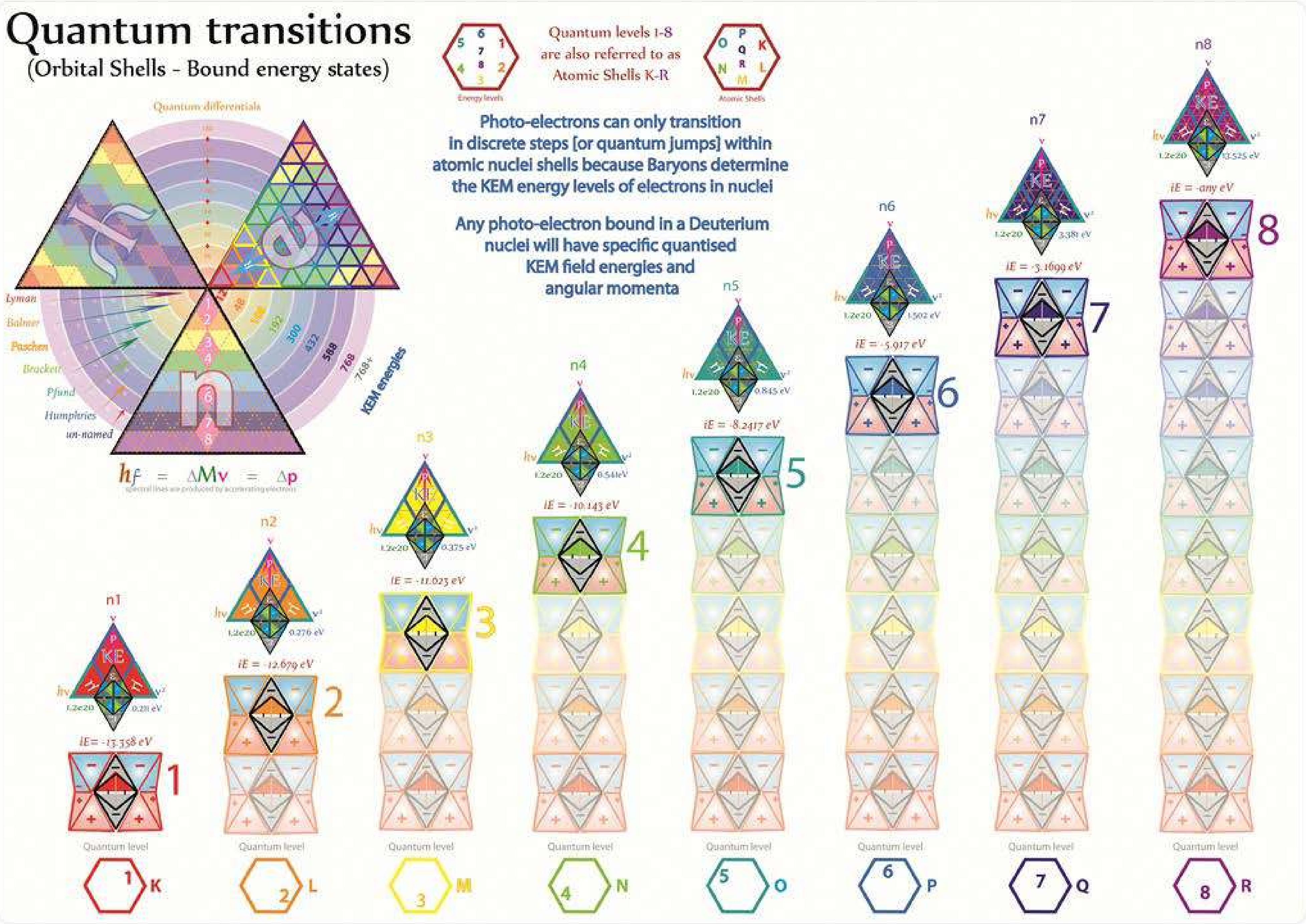
Energy levels
Atomic Shells
Quantum differentials
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
1
2
3
4
5
6
7
8
48
108
192
300
432
588
768
768+
KEM energies
hf = ΔMv = Δp
spectral lines are produced by accelerating electrons
n1
hv
KE
1.2e20
0.211 eV
iE= -13.358 eV
1
n2
1.2e20
0.276 eV
iE = -12.679 eV
2
n3
1.2e20
0.375 eV
iE = -11.623 eV
3
n4
1.2e20
0.541eV
iE = -10.143 eV
4
n5
1.2e20
0.845 eV
iE = -8.2417 eV
5
n6
1.2e20
1.502 eV
iE = -5.917 eV
6
n7
1.2e20
3.381 eV
iE = -3.1699 eV
7
n8
1.2e20
13.525 eV
iE = -any eV
8
Quantum level
1 K
2 L
3 M
4 N
5 O
6 P
7 Q
8 R

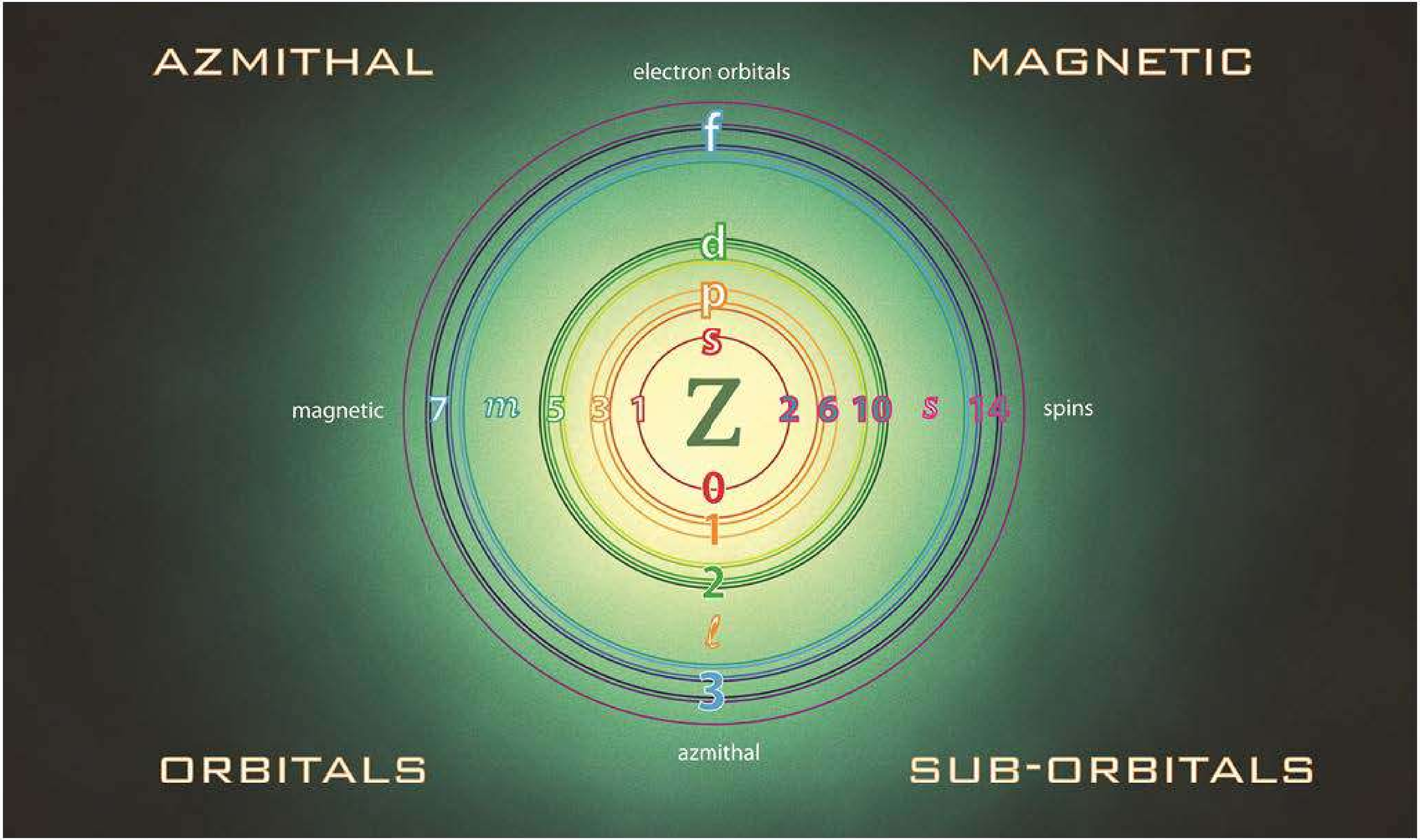
AZMITHAL
MAGNETIC
electron orbitals
f
d
p
s
magnetic
7
m
5
3
1
Z
2
6
10
s
14
spins
0
1
2
l
3
azmithal
ORBITALS
SUB-ORBITALS

Atomic Orbitals

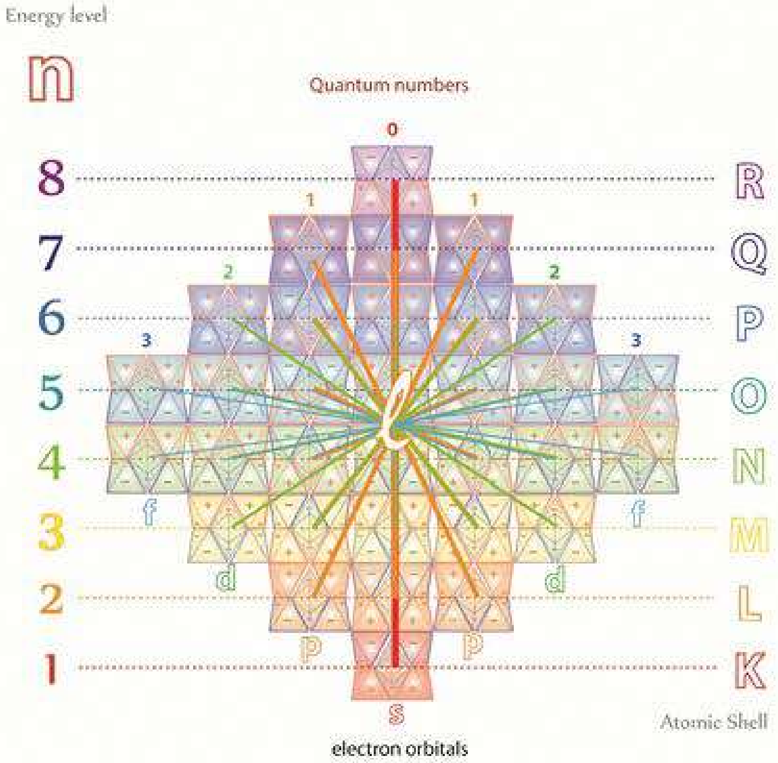
Energy level
n
Quantum numbers
0
1
1
2
2
3
3
8
7
6
5
4
3
2
1
R
Q
P
O
N
M
L
K
s
s+p
s+p+d
s+p+d+f
s+p+d+f
s+p+d
s+p
s
ℓ
f
f
d
d
p
p
s
electron orbitals
Atomic Shell

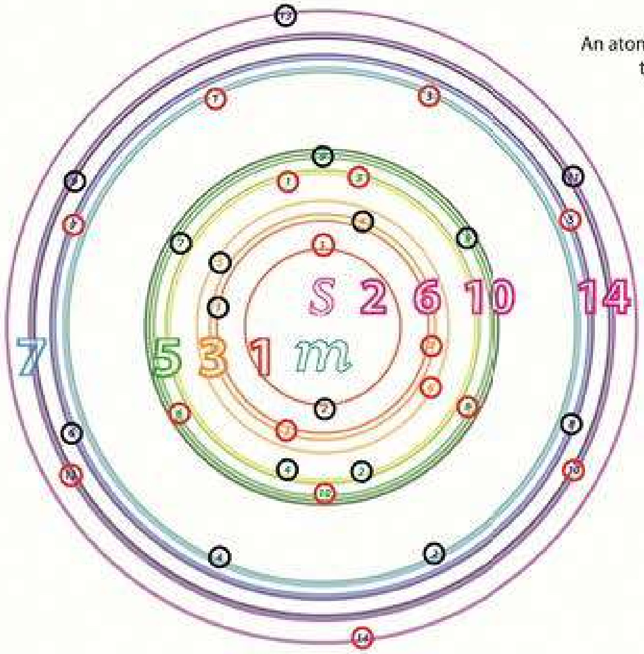
S
2
6
10
14
7
5
3
1
m

An atomic orbital is a mathematical function that describes the wave-like behavior of either one electron or a pair of electrons in an atom

Atomic orbitals are typically categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and an angular momentum vector component, respectively.

Historically used to define the pedagogical electron cloud model of an atom Tetryonics reveals the true geometry of atomic nuclei

Each orbital is defined by a different set of quantum numbers and contains a maximum of two spin opposed electrons.

Aufbau

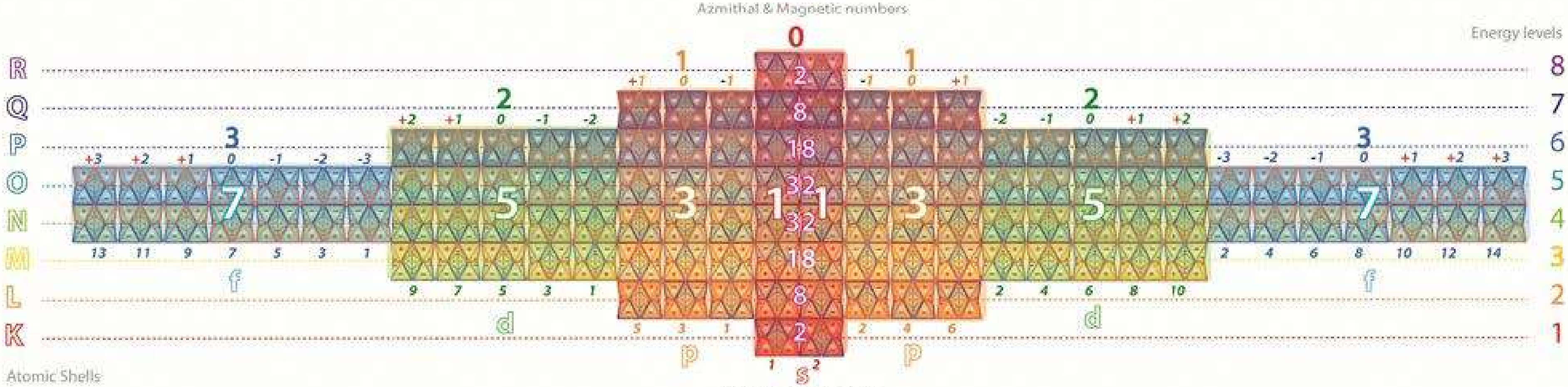
Azmithal & Magnetic numbers
Energy levels
0
1
1
2
2
3
3
+1
0
-1
-1
0
+1
+2
+1
0
-1
-2
-2
-1
0
+1
+2
+3
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+1
+2
+3
2
8
18
32
32
18
8
2
7
5
3
1
1
3
5
7
13
11
9
7
5
3
1
2
4
6
8
10
12
14
9
7
5
3
1
2
4
6
8
10
5
3
1
2
4
6
1
2
f
f
d
d
p
p
s
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Atomic Shells

Orbitals & sub-Orbitals

Electron Spins can be either up or down providing they obey the Pauli exclusion principle

Copyright ABRAHAM [2008] - All rights reserved

S Orbital

1 Orbit (2 electrons max)

Alkali Metals & Alkali Earths

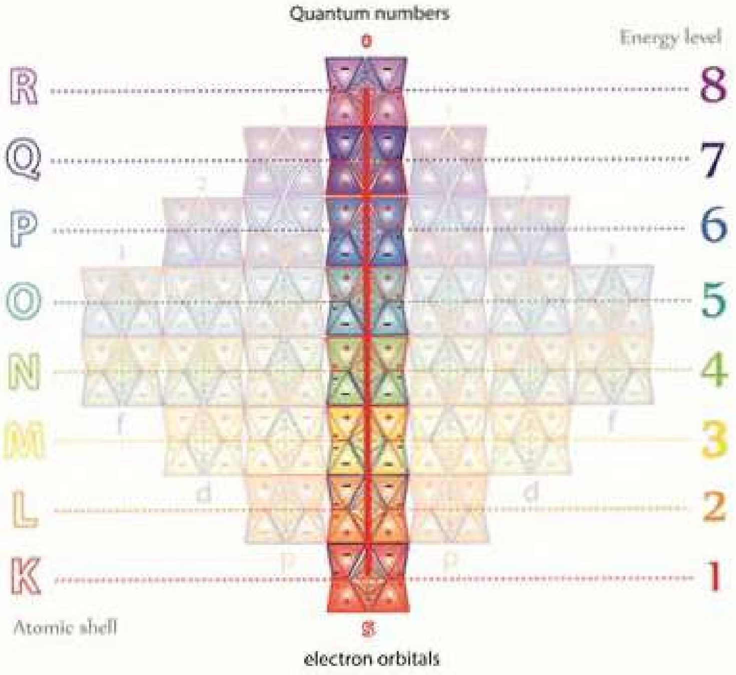
Quantum numbers
0
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
Atomic shell
s
electron orbitals

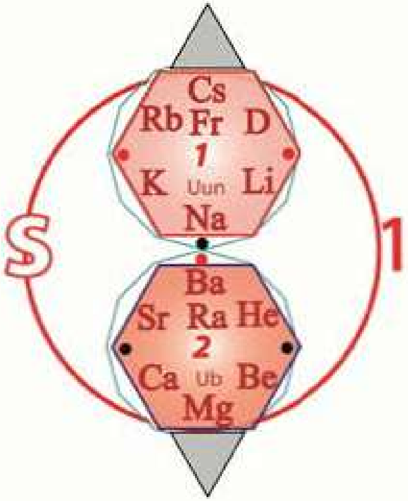
Cs
Rb
Fr
D
1
K
Uun
Li
Na
Ba
Sr
Ra
He
2
Ca
Ub
Be
Mg
S
1

		n
azmithal number	0	l
magnetic numbers	0	m
spins	-1/2 +1/2	S

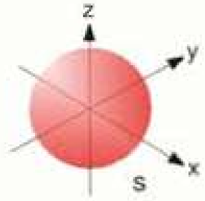

erroneous computer model of 's' electron orbitals

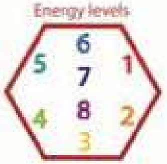
Energy levels
6
5
1
7
8
4
2
3

p Orbital

3 Orbits (6 electrons max)

Non-Metals, Halogens & Nobel gases

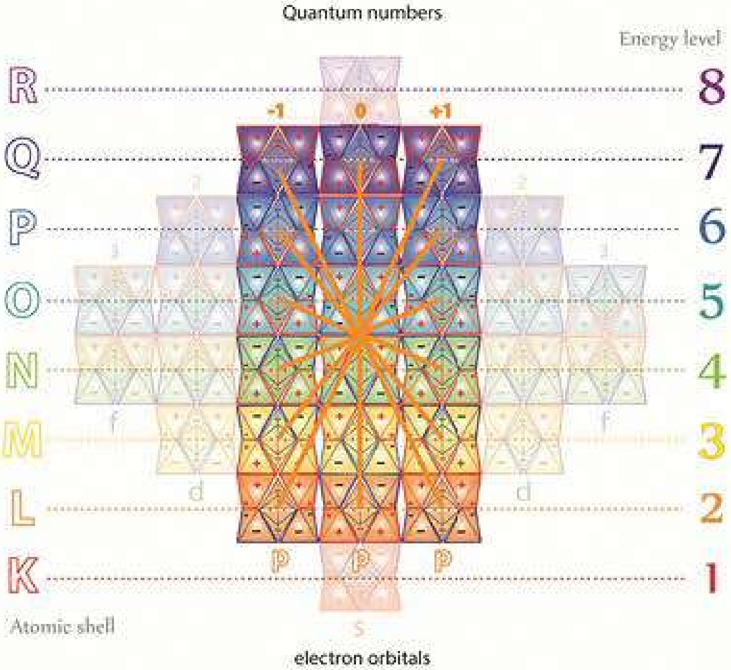
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-1
0
+1
p
p
p
s
d
d
f
f
Atomic shell
electron orbitals

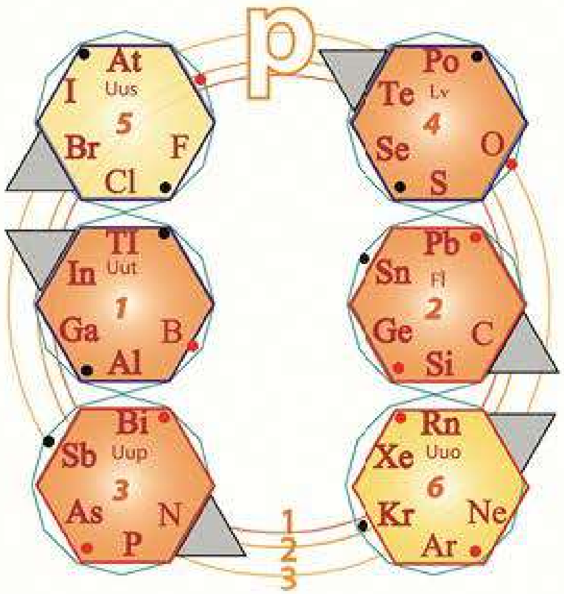
p
At
Uus
5
I
Br
F
Cl
Po
Te
Lv
4
Se
O
S
Tl
In
Uut
1
Ga
B
Al
Pb
Sn
Fl
2
Ge
C
Si
Bi
Sb
Uup
3
As
N
P
Rn
Xe
Uuo
6
Kr
Ne
Ar
1
2
3

				n
azmithal number		1		l
magnetic numbers	-1	0	+1	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	s

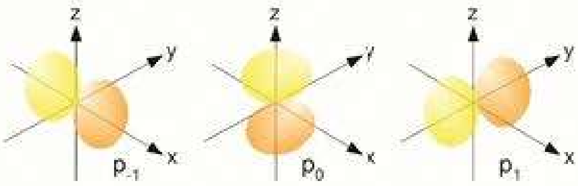
z
y
x
p_{-1}
z
y
x
p_0
z
y
x
p_1

erroneous computer model of 'p' electron orbitals

Quantum levels
6
5
7
4
2
3

d Orbital

5 Orbits (10 electrons max)

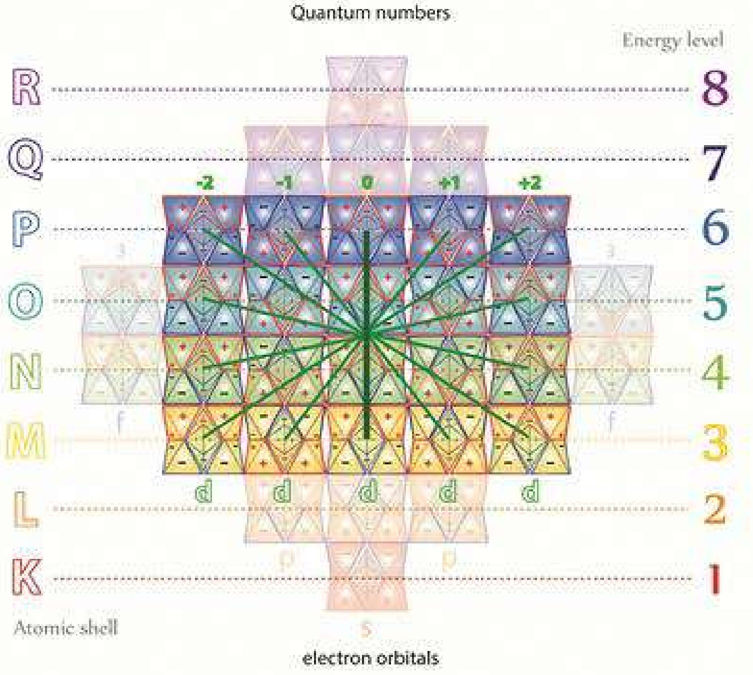
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-2
-1
0
+1
+2
d
d
d
d
d
f
f
p
p
s
Atomic shell
electron orbitals

						n
azimthal number			2			ℓ
magnetic numbers	-2	-1	0	+1	+2	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	s

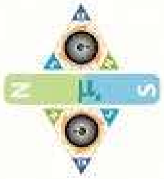

Transition & post-Transition Metals

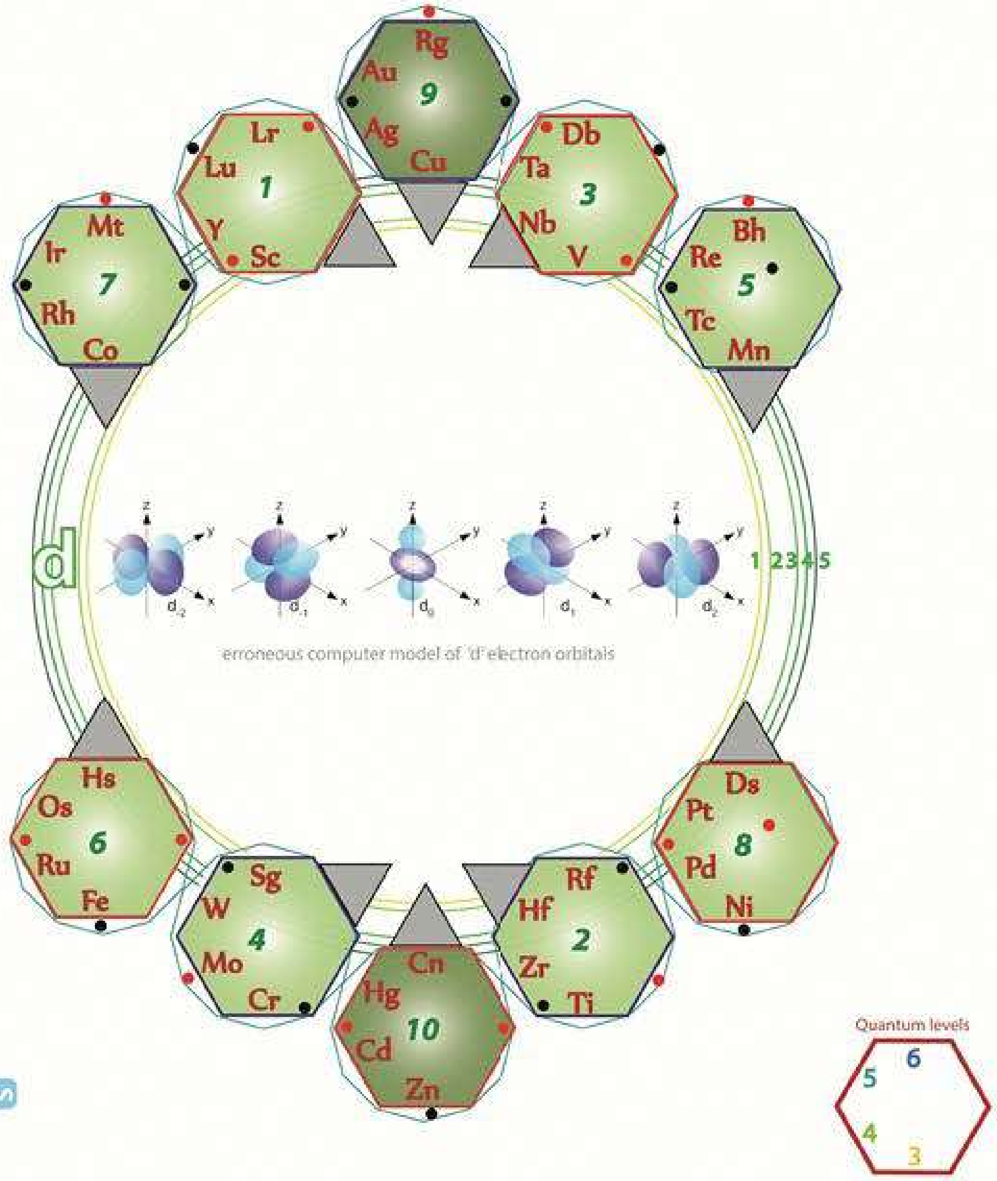
Rg
Au
9
Ag
Cu
Lr
Lu
1
Y
Sc
Db
Ta
3
Nb
V
Mt
Ir
7
Rh
Co
Bh
Re
5
Tc
Mn
d
d_{-2}
d_{-1}
d_0
d_1
d_2
1 2 3 4 5
Hs
Os
6
Ru
Fe
Sg
W
4
Mo
Cr
Cn
Hg
10
Cd
Zn
Rf
Hf
2
Zr
Ti
Ds
Pt
8
Pd
Ni
Quantum levels
6
5
4
3

erroneous computer model of 'd' electron orbitals

f Orbital

7 Orbits (14 electrons max)

Lanthanoids & Actinoids

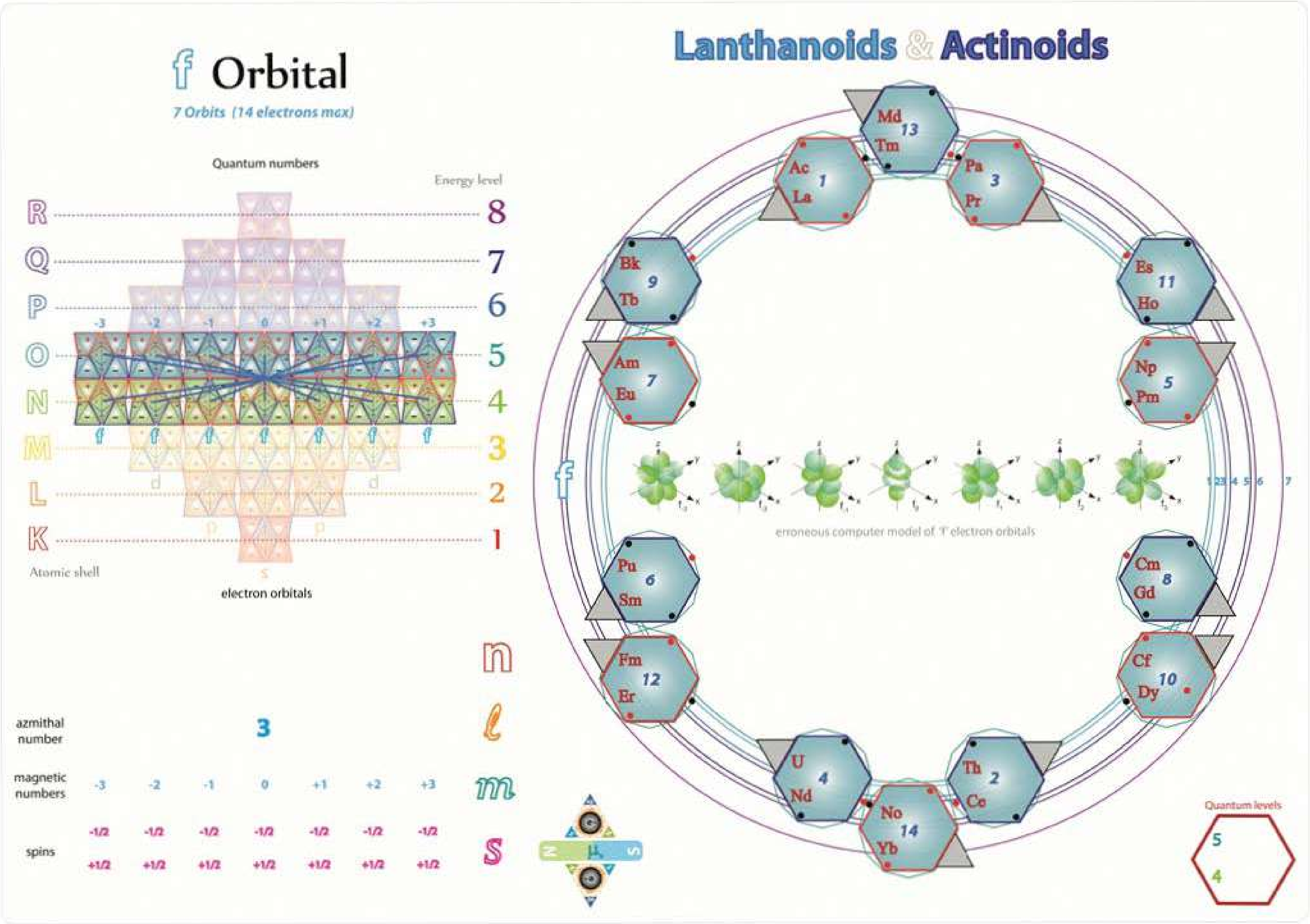
Quantum numbers
Energy level
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
-3 -2 -1 0 +1 +2 +3
f f f f f f f
d d
p p
s
Atomic shell
electron orbitals
n
azmithal number
3
ℓ
magnetic numbers
-3 -2 -1 0 +1 +2 +3
m
spins
-1/2 -1/2 -1/2 -1/2 -1/2 -1/2 -1/2
+1/2 +1/2 +1/2 +1/2 +1/2 +1/2 +1/2
s
Md 13 Tm
Ac 1 La
Pa 3 Pr
Bk 9 Tb
Es 11 Ho
Am 7 Eu
Np 5 Pm
f
1 2 3 4 5 6 7
erroneous computer model of 'f' electron orbitals
Pu 6 Sm
Cm 8 Gd
Fm 12 Er
Cf 10 Dy
U 4 Nd
Th 2 Ce
No 14 Yb
N
μ
S
Quantum levels
5
4

Tetryonics 47.06 - 'f' Orbital

Orbital energy variations

All Elements have stable core electron configurations of s & p orbitals for each energy level as revealed through diffraction studies

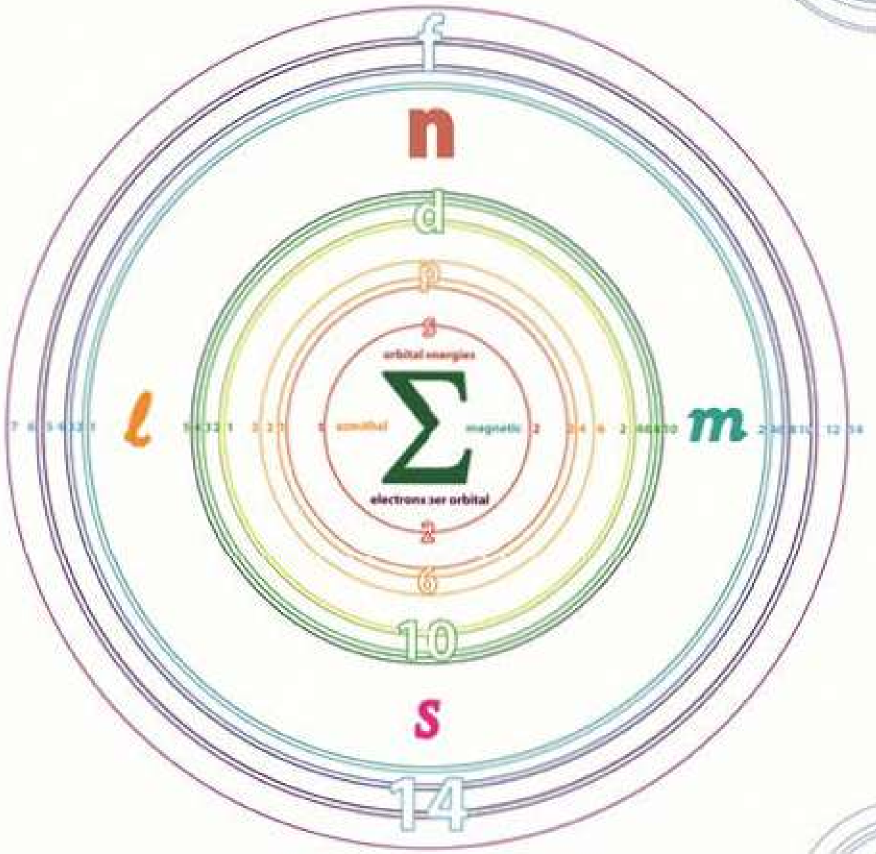
f
n
d
p
s
orbital energies
ℓ
azimuthal
Σ
magnetic
m
electrons per orbital
2
6
10
s
14

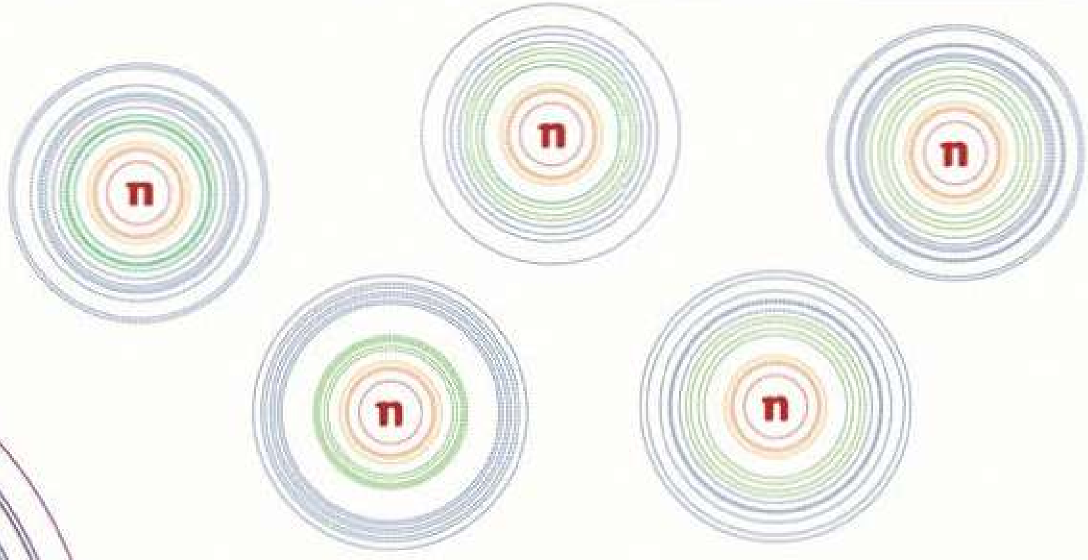

As additional nuclei bind to form d & f orbitals they can do so by bonding in many positions, with each location producing different orbital energies for each electron that binds to nuclei in that position

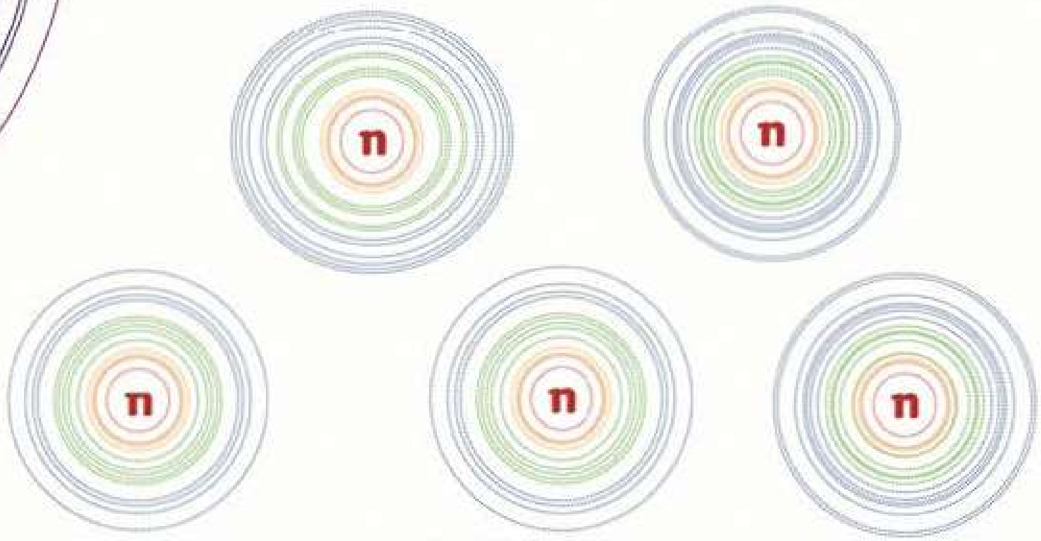

The final energy levels of each orbital is the result of the energy of the Baryons in the nuclei and the spin coupling energies of the photo-electrons bound to them

Copyright ABRAHAM [2008] - All rights reserved

Electrons per shell

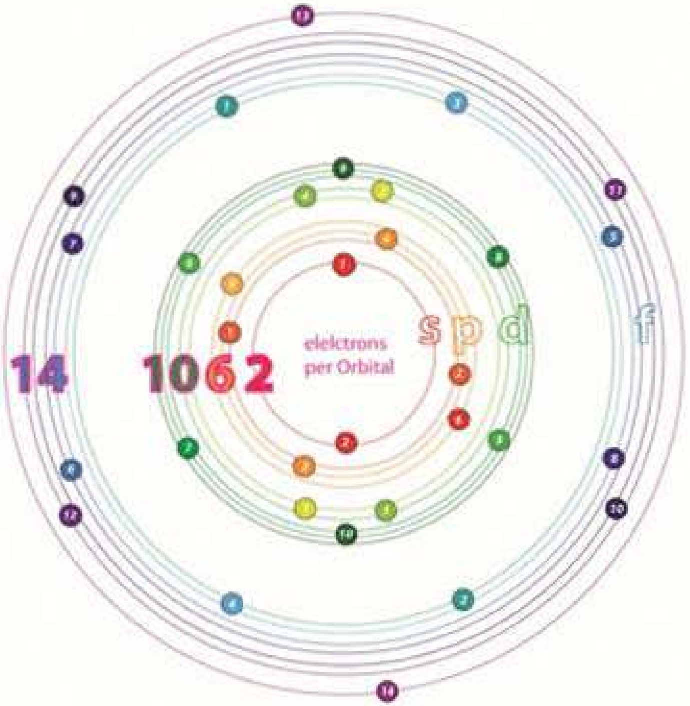

Each energy shell of a periodic element can hold only a fixed number of electrons

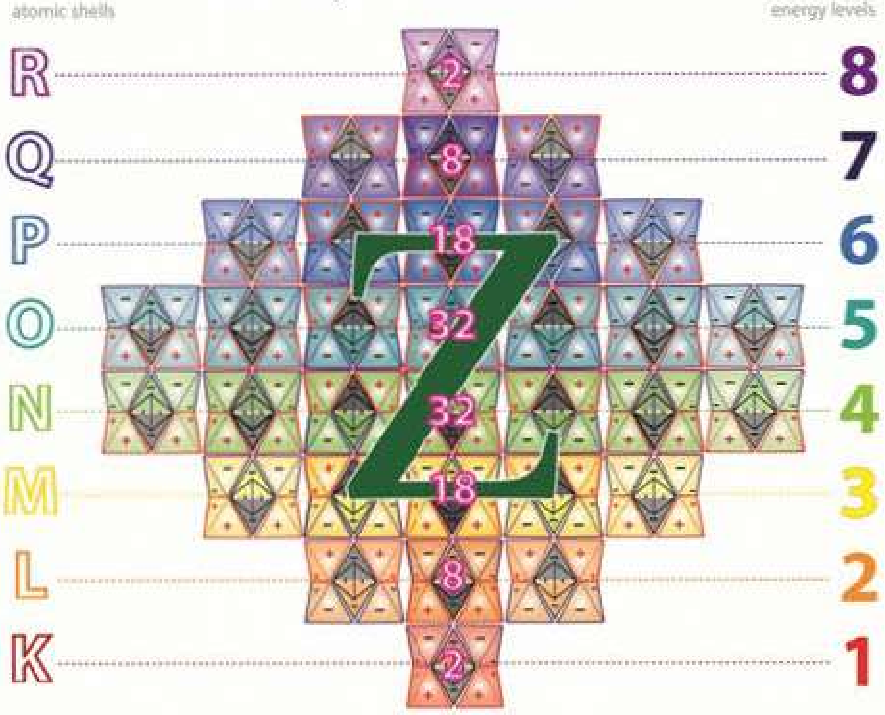

Atomic Orbitals

f	d	p	s	p	d	f
5, 11, 3, 13, 1, 9, 7	8, 2, 10, 4, 6	5, 1, 3	1-2	6, 2, 4	7, 1, 9, 3, 5	6, 12, 4, 14, 2, 10, 8

7 + 5 + 3 + 2 + 3 + 5 + 7

Rules governing the allowed combinations of Quantum Numbers

The three quantum numbers (n, l, and m)
that describe an orbital are integers: 0, 1, 2, 3, and so on

n (1-8) Principal [illegible]	The principal quantum number (n) cannot be zero.
ℓ (0-3) Azimuthal [illegible]	The angular quantum number (l) can be any integer between 0 and n - 1.
mℓ (2ℓ+1) Magnetic [illegible]	The magnetic quantum number (m) can be any integer between -l and +l.
m_s ±$\frac{1}{2}$ Spin Projection [illegible]	The Spin of electrons in any nuclear sub-orbital can only be +1/2 (Spin UP) or -1.2 (Spin DOWN).

Z

1, 2, 3, 4, 5, 6, 7, 8

s, p, d, f

1	2
1, 3, 5	2, 4, 6
1, 3, 5, 7, 9	2, 4, 6, 8, 10
1, 3, 5, 7, 9, 11, 13	2, 4, 6, 8, 10, 12, 14
down up	up down

Quantum Numbers

n	4-5	3-6	2-7	1-8	2-7	3-6	4-5
ℓ	3	2	1	0	1	2	3
m	-3	-2	-1	1	+1	+2	+3
s	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2

Erwin Schrödinger

(12 August 1887 – 4 January 1961)

Using Tetryonic charged geometries for mass-ENERGY-Matter,
an electron's position and velocity CAN be modelled simultaneously
(but any attempt to measure or interact with it, will affect its component energy-momenta)

Electron Position Uncertainty

Atomic orbitals are typically described as "hydrogen-like" (meaning one-electron) wave functions over any spatial region of measurement, categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and a vector momentum component, respectively

Werner Heisenberg

(5 December 1901 – 1 February 1976)

Lepton's are physically Spin 1 fermion particles that can easily be misconstrued as having entirely different spin numbers without the correct physical topologies to base the observed measurements on

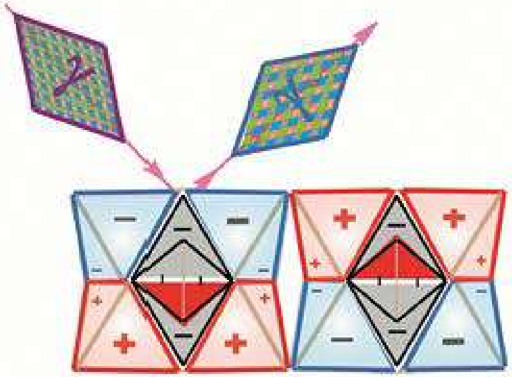

The energies of photo-electrons are determined by the Baryons they bind to & incident photons

Quantum Mechanics is a mathematical representation of equilateral energy momenta interactions and the charged geometries of mass-ENERGY-Matter

Determining the motion of electrons bound to atomic nuclei is akin to measuring the motion of variable speed electric fan blades mounted at various heights within a rotating carousel

Leptons are 12 loop quantum rotors

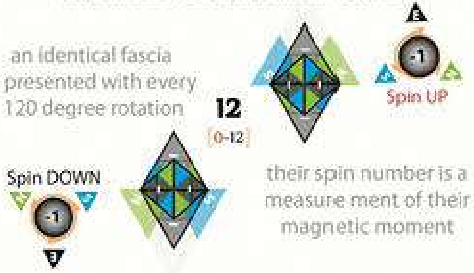

an identical fascia
is presented with every
120 degree rotation

their spin number is a
measure ment of their
magnetic moment

The unique 12 faceted topology
of leptons results in an identical EM geometry
being oberved for every 120° rotation of the particle

Leading to the interpretation that
the Lepton disappears and re-appears
when being 'observed' or measured

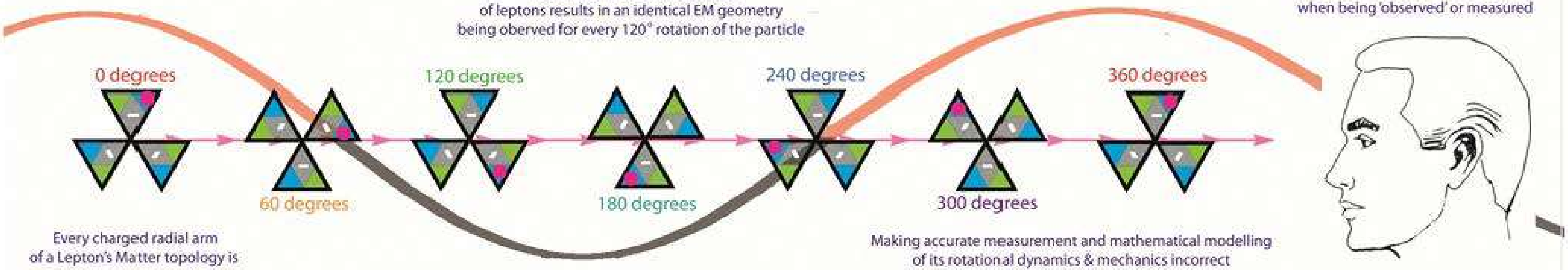

Every charged radial arm
of a Lepton's Matter topology is
identical to every other

Making accurate measurement and mathematical modelling
of its rotational dynamics & mechanics incorrect
without the correct physical topologies

Electron modelling & probability calculations

12
[0-12]
RE
KE
S

Electron positional measurements
have proven to be historically
difficult to accurately model
due to the charged 12π
rotating topologies
of leptons

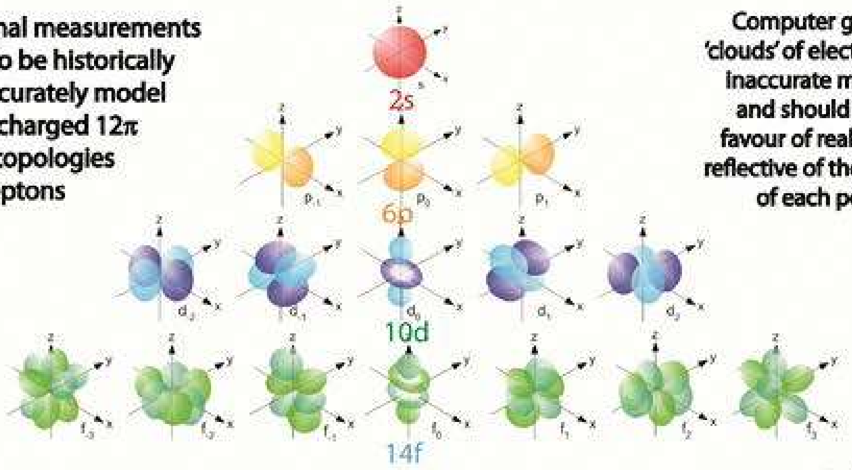

Computer generated plots of
'clouds' of electron probabilites are
inaccurate mis-representations
and should be abandoned in
favour of realistic atomic models
reflective of the charged geometry
of each periodic element

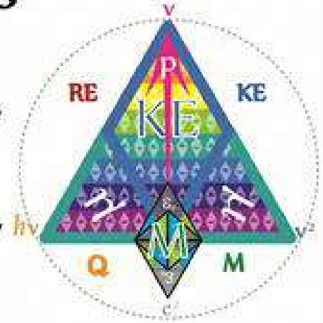

Every elemental atom can be viewed as
a quantum carousel with a unique number
of oscillating fans positioned around it.

Each fan has 3 blades and a fixed speed n[1-8]
related to its height above ground level,
AND the carousel is turning around on its axis

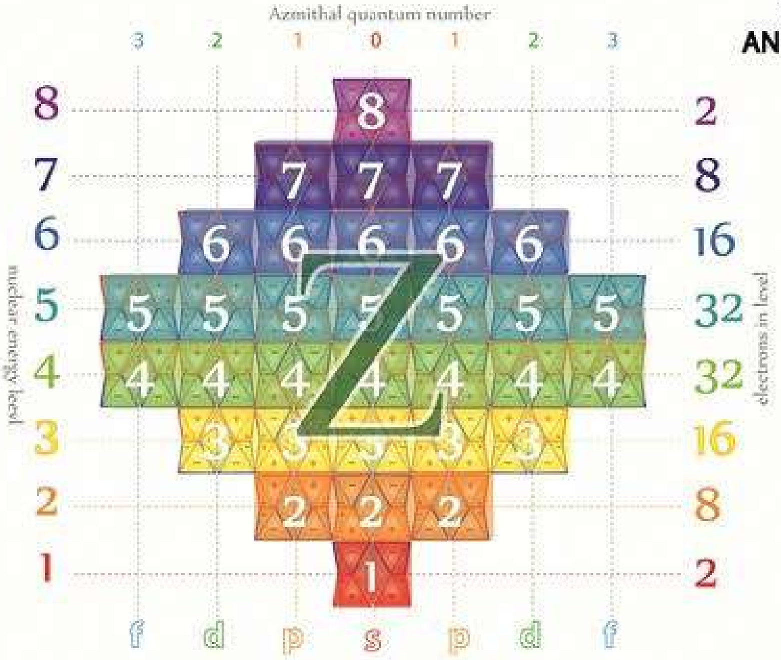

The current computer generated
electron probability diagrams in popular
use at present can now be show to be a
misrepresentative model of mathematical
modelling of electron sub-orbital energies

The nuclear quantum levels [n],
intrinsic quantised angular momentum [h] and
orbital angular momentum [l] of each electron
bound within atomic nuclei are all the direct
result of the Baryonic energies of the nuclei
they are bound to

**Each level of the quantum carousel
can contain only a limited number of
fans each running at a specific speed**

**All periodic elements
are made form
Deuterium nuclei**

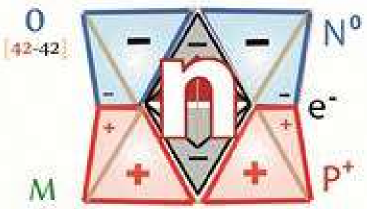

**The baryonic energies of nuclei
determines the energies of
bound photo-electrons**

Imagine trying to measure (or model) the motion of any 1 quantum scale blade while the carousel rotates

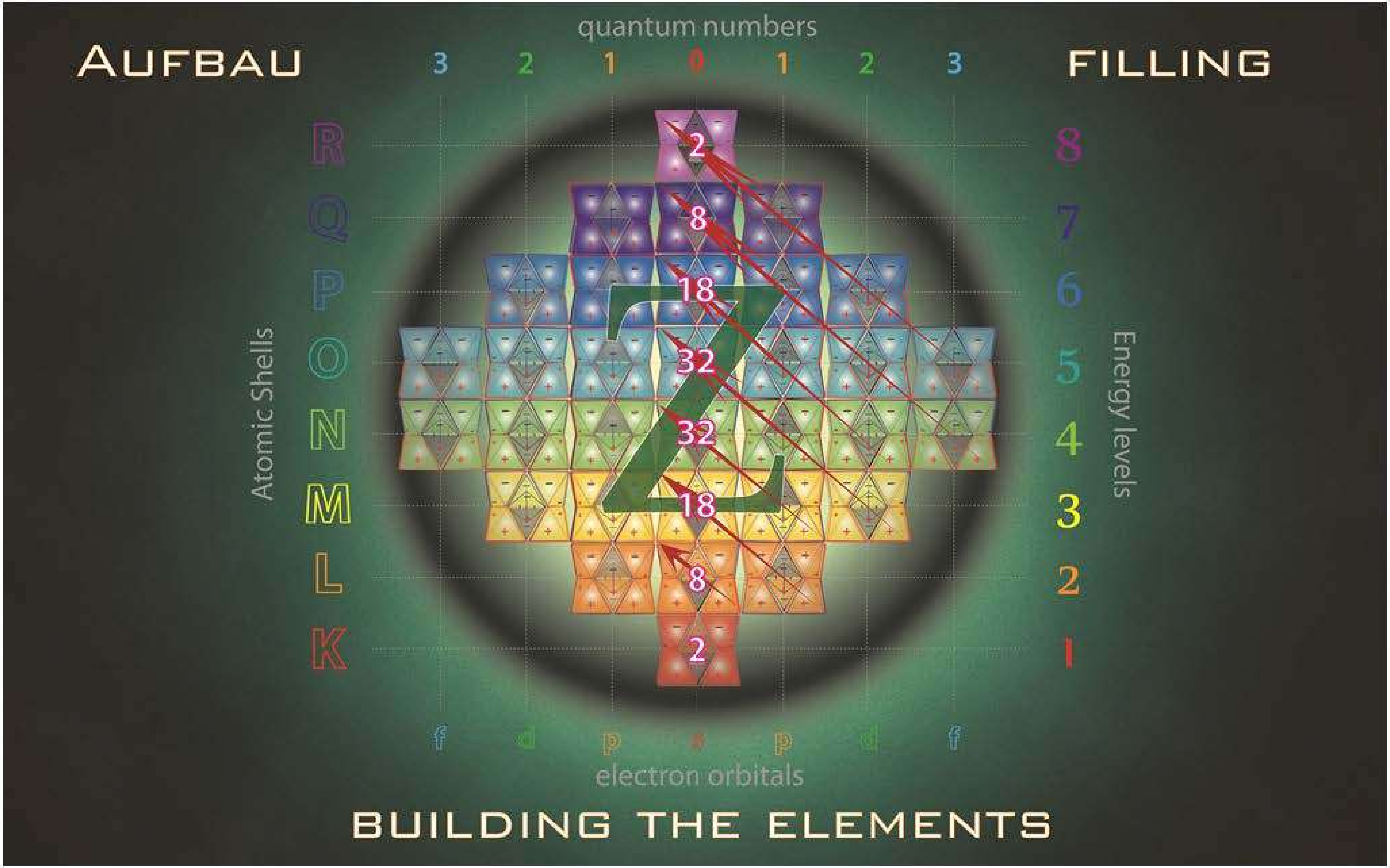
quantum numbers
AUFBAU
FILLING
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Atomic Shells
8
7
6
5
4
3
2
1
Energy levels
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
electron orbitals
BUILDING THE ELEMENTS

Tetryonics 48.01 - aufbau geometry

Copyright ABRAHAM [2008] - All rights reserved

Quantum Topologies

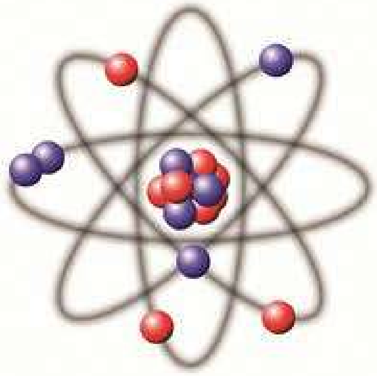

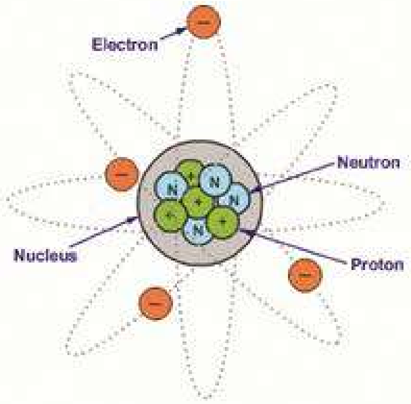
Electron
Neutron
Nucleus
Proton
N

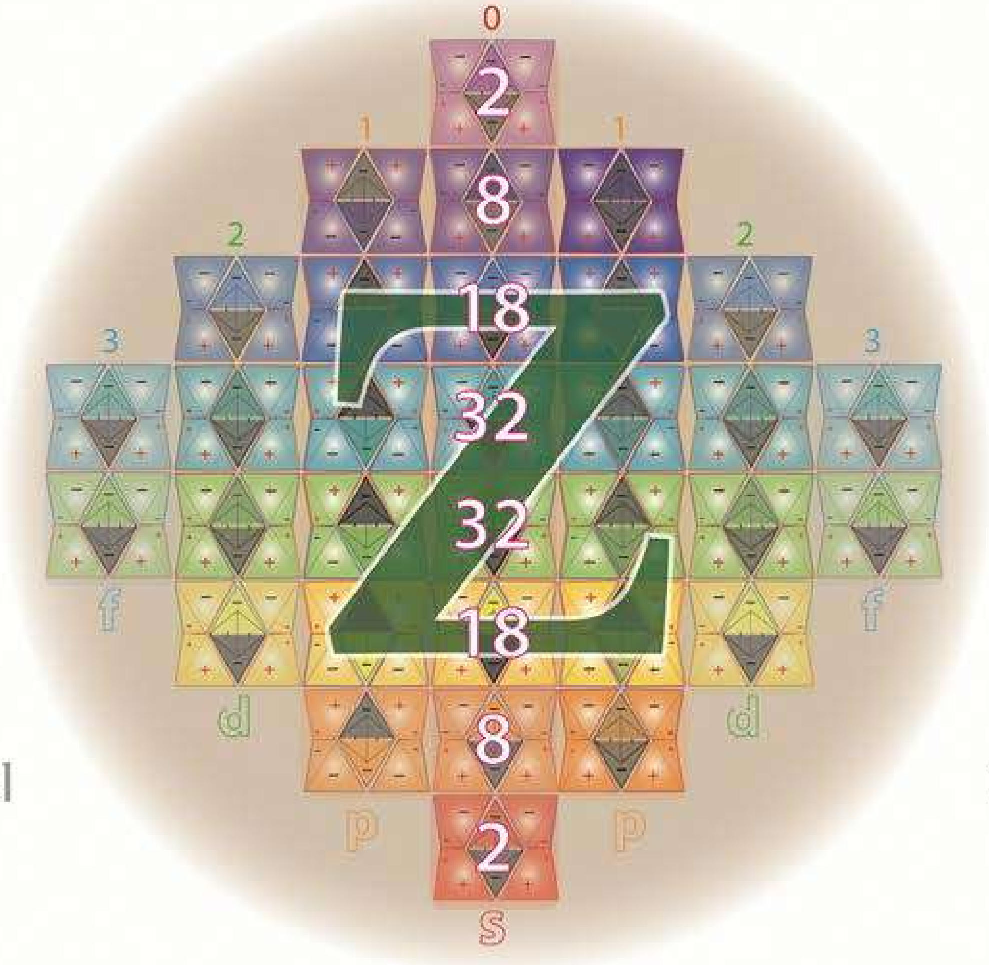
0
1
1
2
2
3
3
2
8
18
32
32
18
8
2
Z
f
f
d
d
p
p
s

Dalton Model

Quantum Model

Thomson Model

Rutherford Model

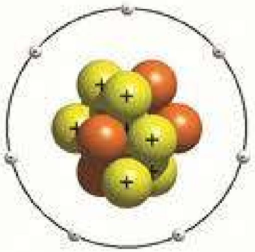

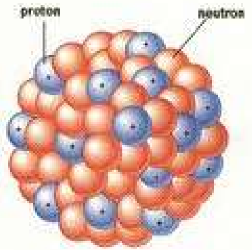
proton
neutron

Historically viewed as a spherical object
Tetryonic charge geometry has finally revealed the
true quantum topology of all atoms

Copyright ABRAHAM [2008] - All rights reserved

Element numbers

The rule dictating how many nuclei form each Atomic shell is known as the Aufbau principle.

The physical and chemical properties of elements is determined by the atomic structure.
The atomic structure is, in turn, determined by the electrons and
which shells, subshells and orbitals they reside in.

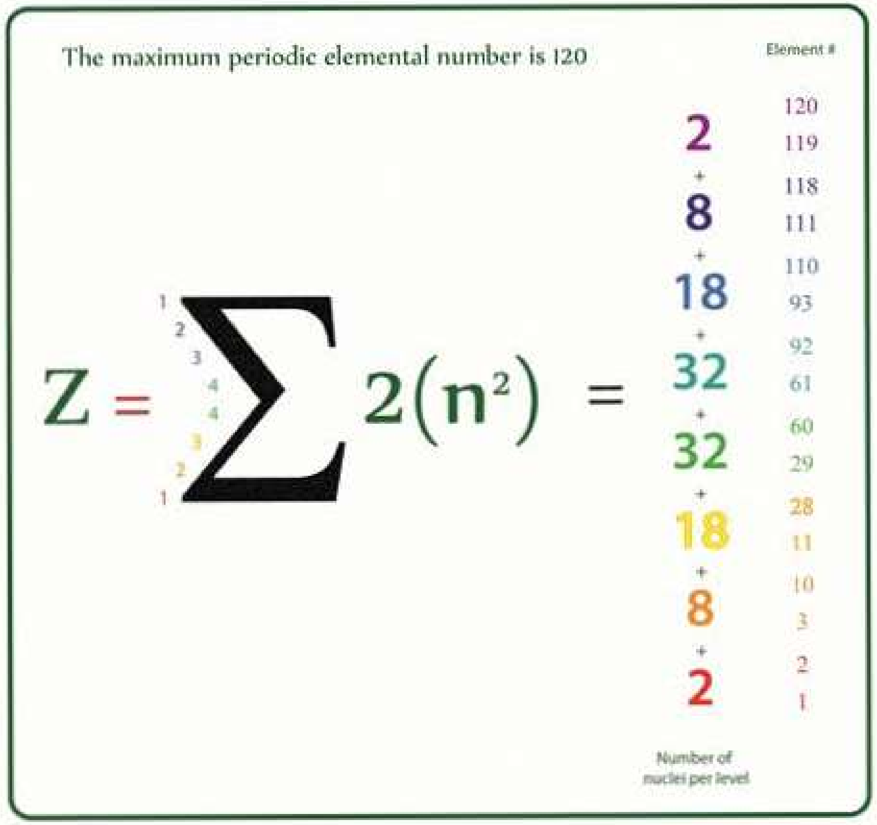

The maximum periodic elemental number is 120

$$Z = \sum 2(n^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

1
2
3
4
4
3
2
1
Element #
120
119
118
111
110
93
92
61
60
29
28
11
10
3
2
1
Number of nuclei per level

The number of nuclei per quantum level is reflective of photonic energy levels and provides the foundational geometry for all of the periodic elements

The number of possible nuclei in each Quantum level follows aufbau principle 'numbers' which can be determined using the following summation formula

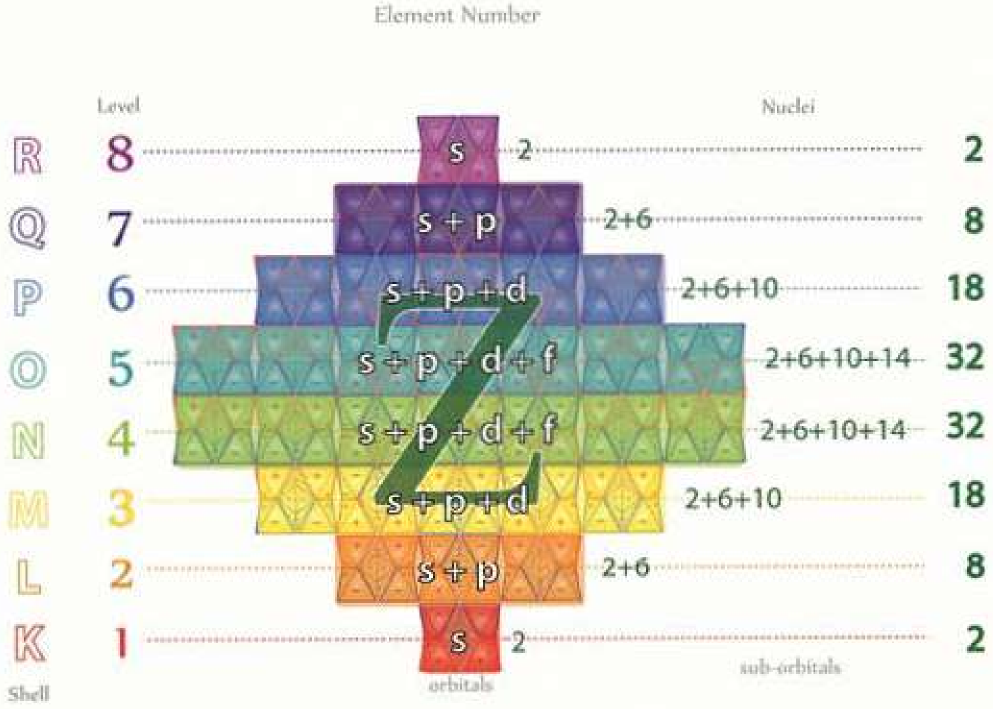

Element Number
Level
Nuclei
R
8
s
2
2
Q
7
s + p
2+6
8
P
6
s + p + d
2+6+10
18
O
5
s + p + d + f
2+6+10+14
32
N
4
s + p + d + f
2+6+10+14
32
M
3
s + p + d
2+6+10
18
L
2
s + p
2+6
8
K
1
s
2
2
Z
Shell
orbitals
sub-orbitals

Deuterium is the building block of all elements

Each element has equal numbers of Protons, electron & Neutrons with their stored mass-energies making up the molar masses of elements not excess neutron as currently modelled

Copyright ABRAHAM [2008] - All rights reserved

Aufbau Principle

(Nuclei number and position)

1 Sub-Orbit (2 electrons max) ○ s

3 Sub-Orbits (6 electrons max) ○ p

○ d *5 Sub-Orbits (10 electrons max)*

○ f *7 Sub-Orbits (14 electrons max)*

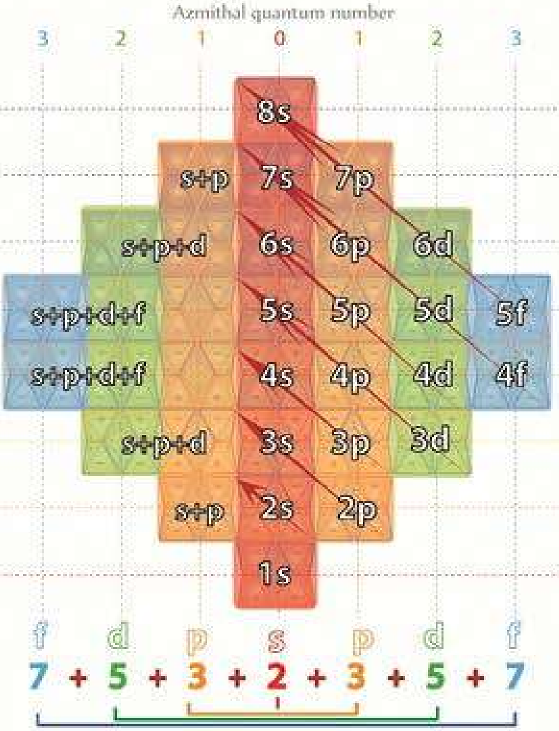

atomic shells			energy level
R	2	2	8
Q	2+6	8	7
P	2+6+10	18	6
O	2+6+10+14	32	5
N	2+6+10+14	32	4
M	2+6+10	18	3
L	2+6	8	2
K	2	2	1

2 6 10 14

Nuclei	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	Quantum number	per level
R	119	120																															8	2
Q	87	88	113	114	115	116	117	118																									7	8
P	55	56	81	82	83	84	85	86	103	104	105	106	107	108	109	110	111	112															6	18
O	37	38	49	50	51	52	53	54	71	72	73	74	75	76	77	78	79	80	89	90	91	92	93	94	95	96	97	98	99	100	101	102	5	32
N	19	20	31	32	33	34	35	36	39	40	41	42	43	44	45	46	47	48	57	58	59	60	61	62	63	64	65	66	67	68	69	70	4	32
M	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	29	30															3	18
L	3	4	5	6	7	8	9	10																									2	8
K	1	2																															1	2
Shells	s1	s2	p1	p2	p3	p4	p5	p6	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14		

sub-orbitals

A specific nuclear energy is associated with each electron configuration and, upon certain conditions, bound electrons are able to move from one orbital to another by emission or absorption of a quantum of energy, in the form of a photon.

The Aufbau principle can also be used to model the binding configuration of Protons and Neutrons in an atomic nucleus, according to their specific Baryonic Energy levels, in turn revealing the quantum geometry of all periodic Elements

Wolfgang Pauli

(25 April 1900 – 15 December 1958)

The orbitals of lower energy are filled in first with the electrons and only then are the higher energy orbitals filled.

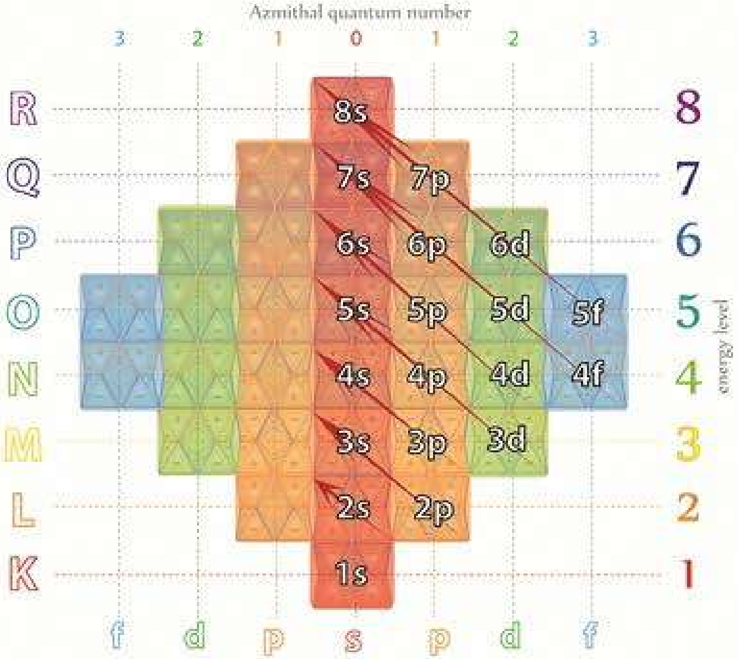
Azmithal quantum number
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8s
7s
7p
6s
6p
6d
5s
5p
5d
5f
4s
4p
4d
4f
3s
3p
3d
2s
2p
1s
8
7
6
5
4
3
2
1
energy level
f
d
p
s
p
d
f

Friedrich Hermann Hund

(4 February 1896 - 31 March 1997)

Orbitals of equal energy are each occupied by one electron before any orbital is occupied by a second electron, and all electrons in singly occupied orbitals must have the same spin state

aufbau electron orbital filling

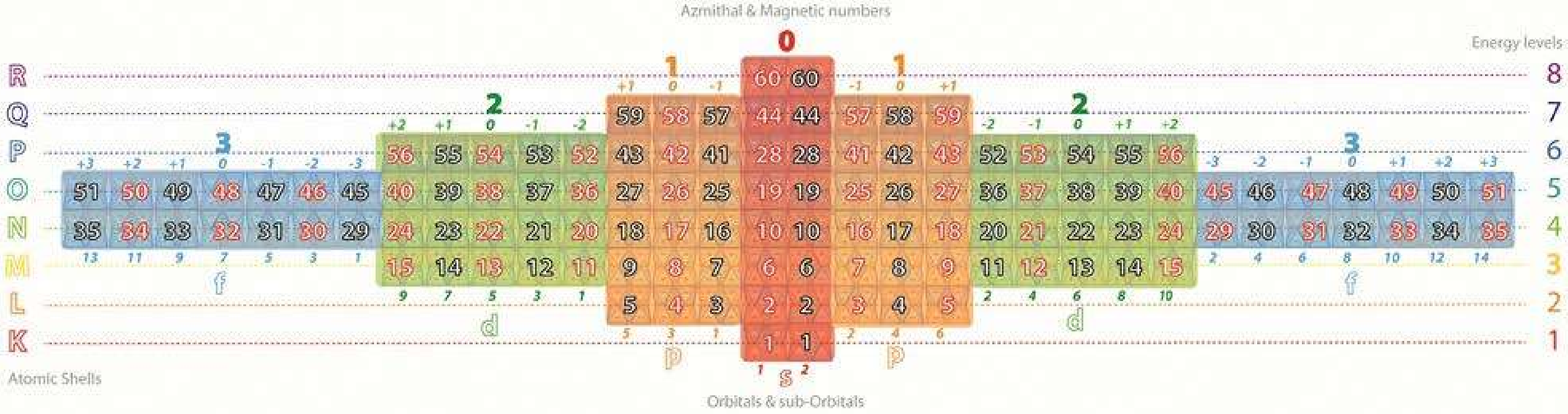
Azmithal & Magnetic numbers
Energy levels
0
1
1
2
2
3
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
R 60 60 8
Q 59 58 57 44 44 57 58 59 7
P 56 55 54 53 52 43 42 41 28 28 41 42 43 52 53 54 55 56 6
O 51 50 49 48 47 46 45 40 39 38 37 36 27 26 25 19 19 25 26 27 36 37 38 39 40 45 46 47 48 49 50 51 5
N 35 34 33 32 31 30 29 24 23 22 21 20 18 17 16 10 10 16 17 18 20 21 22 23 24 29 30 31 32 33 34 35 4
M 15 14 13 12 11 9 8 7 6 6 7 8 9 11 12 13 14 15 3
L 5 4 3 2 2 3 4 5 2
K 1 1 1
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 3 1
p
1 s 2
2 4 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
Atomic Shells
Orbitals & sub-Orbitals

Aufbau construction

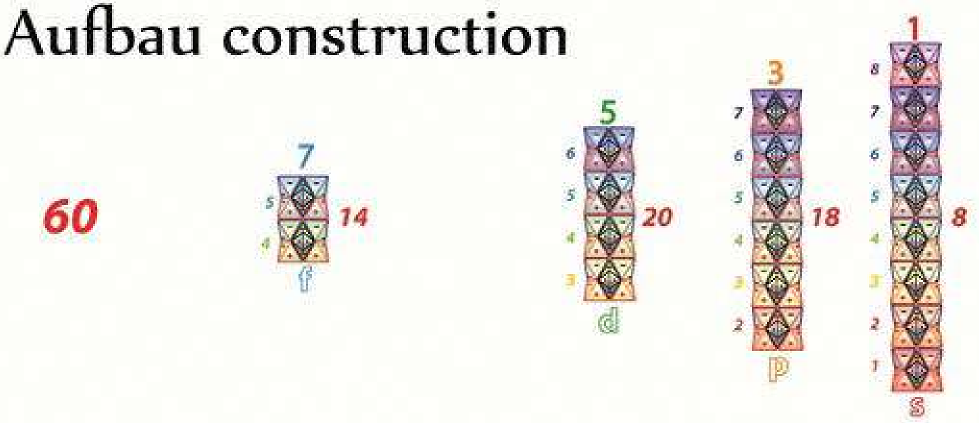
60
7
14
f
5
20
d
3
18
p
1
8
s

Proton
base
Hemisphere

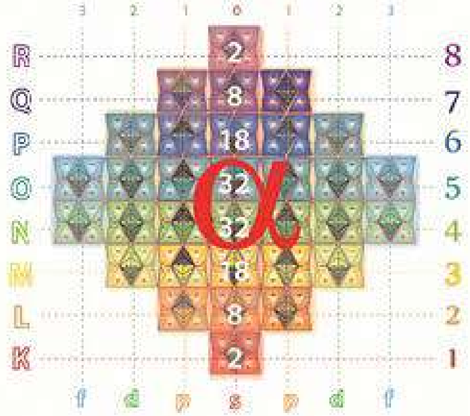
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
α
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

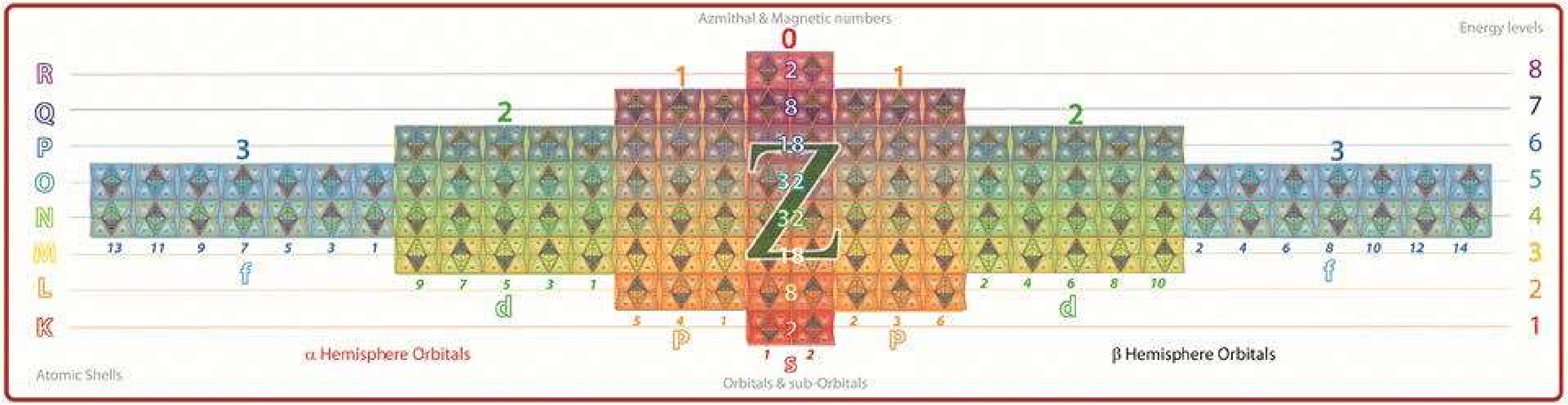
Azmithal & Magnetic numbers
Energy levels
0
1
1
2
2
3
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 4 1
p
1 s 2
2 3 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
α Hemisphere Orbitals
β Hemisphere Orbitals
Atomic Shells
Orbitals & sub-Orbitals

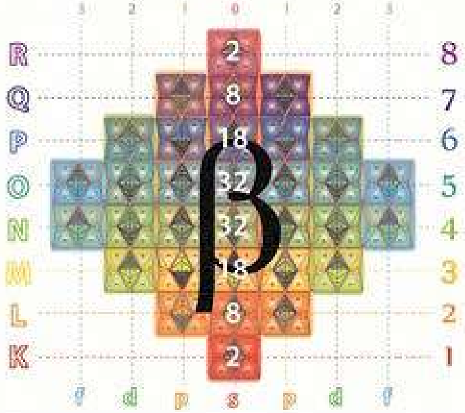
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
β
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Neutron
base
Hemisphere

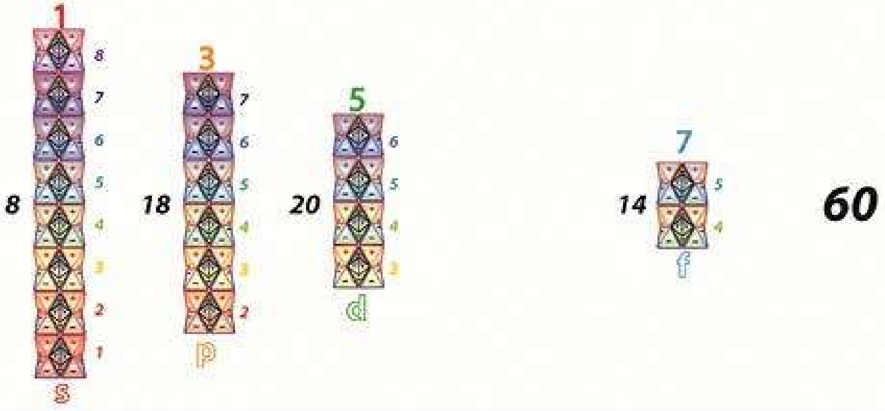
1
8
s
3
18
p
5
20
d
7
14
f
60

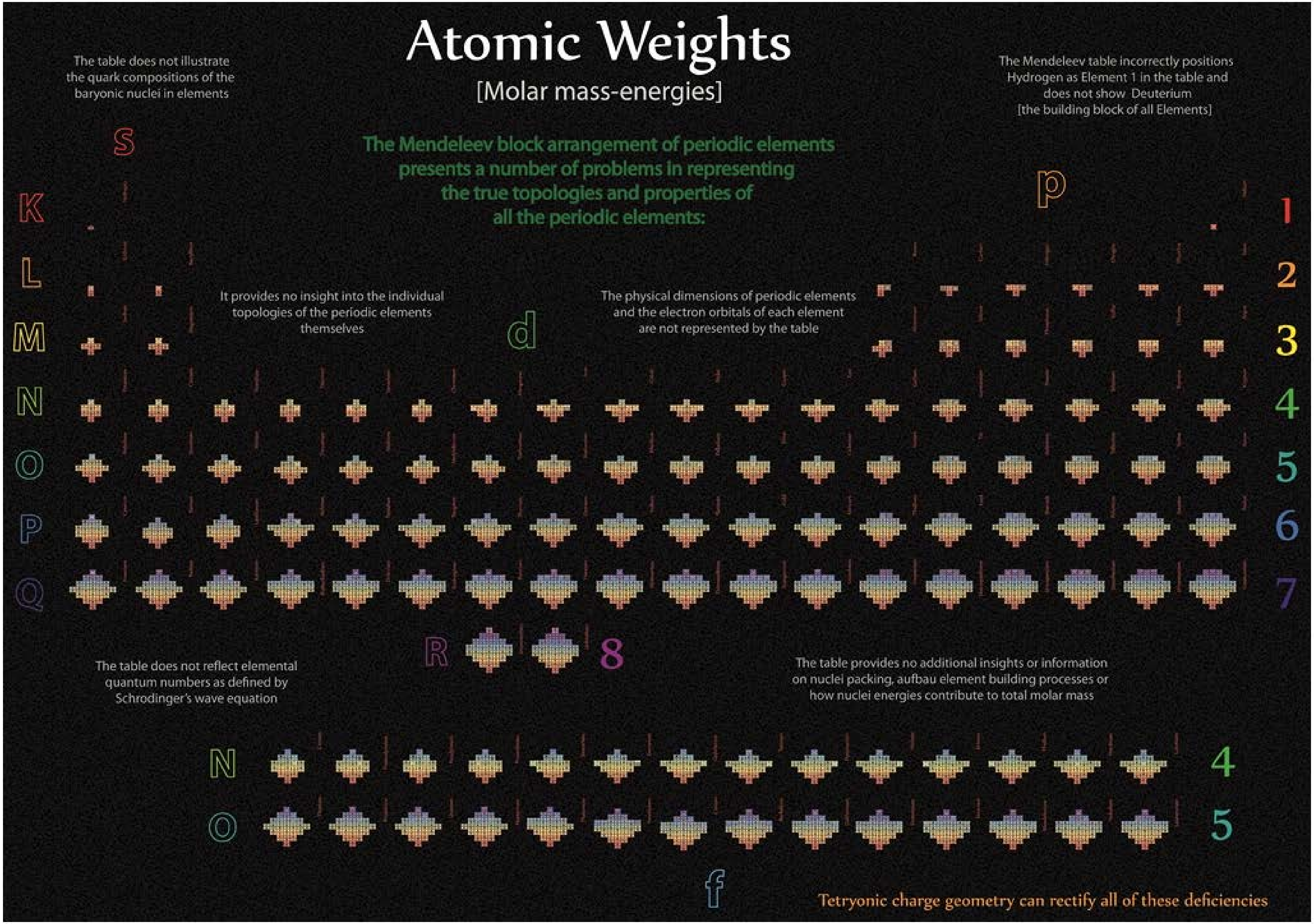

Atomic Weights

[Molar mass-energies]

The table does not illustrate the quark compositions of the baryonic nuclei in elements

The Mendeleev table incorrectly positions Hydrogen as Element 1 in the table and does not show Deuterium [the building block of all Elements]

The Mendeleev block arrangement of periodic elements presents a number of problems in representing the true topologies and properties of all the periodic elements:

It provides no insight into the individual topologies of the periodic elements themselves

The physical dimensions of periodic elements and the electron orbitals of each element are not represented by the table

The table does not reflect elemental quantum numbers as defined by Schrodinger's wave equation

The table provides no additional insights or information on nuclei packing, aufbau element building processes or how nuclei energies contribute to total molar mass

Tetryonic charge geometry can rectify all of these deficiencies

Tetryonics 49.01 - Atomic Weight

E

The quantum level mass-energies of Baryons determines the kinetic energies of electrons

1n - 1 e19v

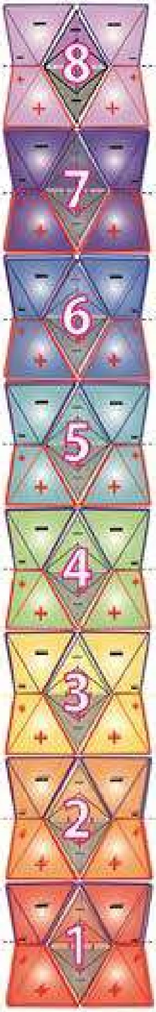

Baryon mass-energies n		E		electron KEM field hv^2
73,728	n32	8	v8	3.268969693 e15
69,192	n31	7	v7	2.502804921 e15
64,800	n30	6	v6	1.838795452 e15
60,552	n29	5	v5	1.276941286 e15
56,448	n28	4	v4	8.172424232 e14
52,488	n27	3	v3	4.596988631 e14
48,672	n26	2	v2	2.043106058 e14
45,000	n25	1	v1	5.107765145 e13

Compton frequencies

930 MeV

1.2e20
electron rest Matter is velocity invariant

496 keV

spectral frequencies

13.6 eV

Atomic nuclei mass-energies

Each element's weight [mass-Matter in a gravitational field] is the result of the total quanta comprising that element

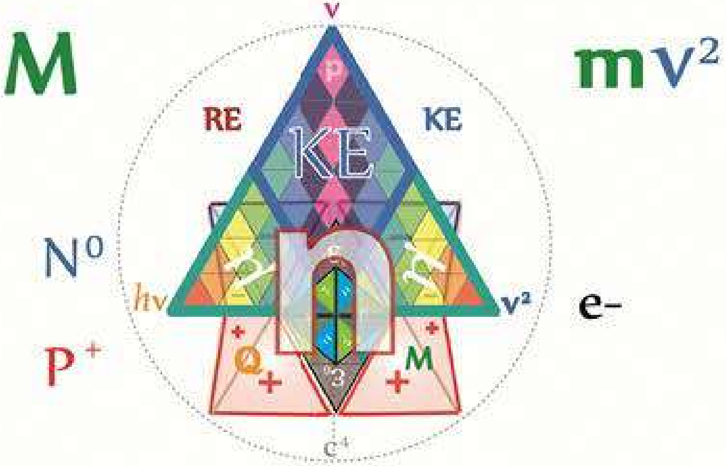

The nuclei forming each atomic shell have specific mass-energy quanta

$$\overset{8}{\underset{1}{n}}\left[\overset{\text{Baryon rest masses}}{\left[72(n)^2\right]}+\overset{\text{lepton rest mass}}{\left[12e19\right]}+\overset{\text{KEM}}{\left[m_e v^2\right]}\right]$$

Deuterium mass-energy per shell

Despite having differing mass-energies each Deuterium nuclei has the same velocity invariant Matter geometry [84π]

spin orbital coupling in synchronous quantum convertors

Electrons act as quantum scale rotating armatures in atomic nuclei and can only have specific energies reflective of the electron orbital energy level of the Baryons in which they are found

They acheive these energy levels by absorbing or emitting photons to acheive the specific angular momentum required

Hydrogen Ionisation Energies [/n²]

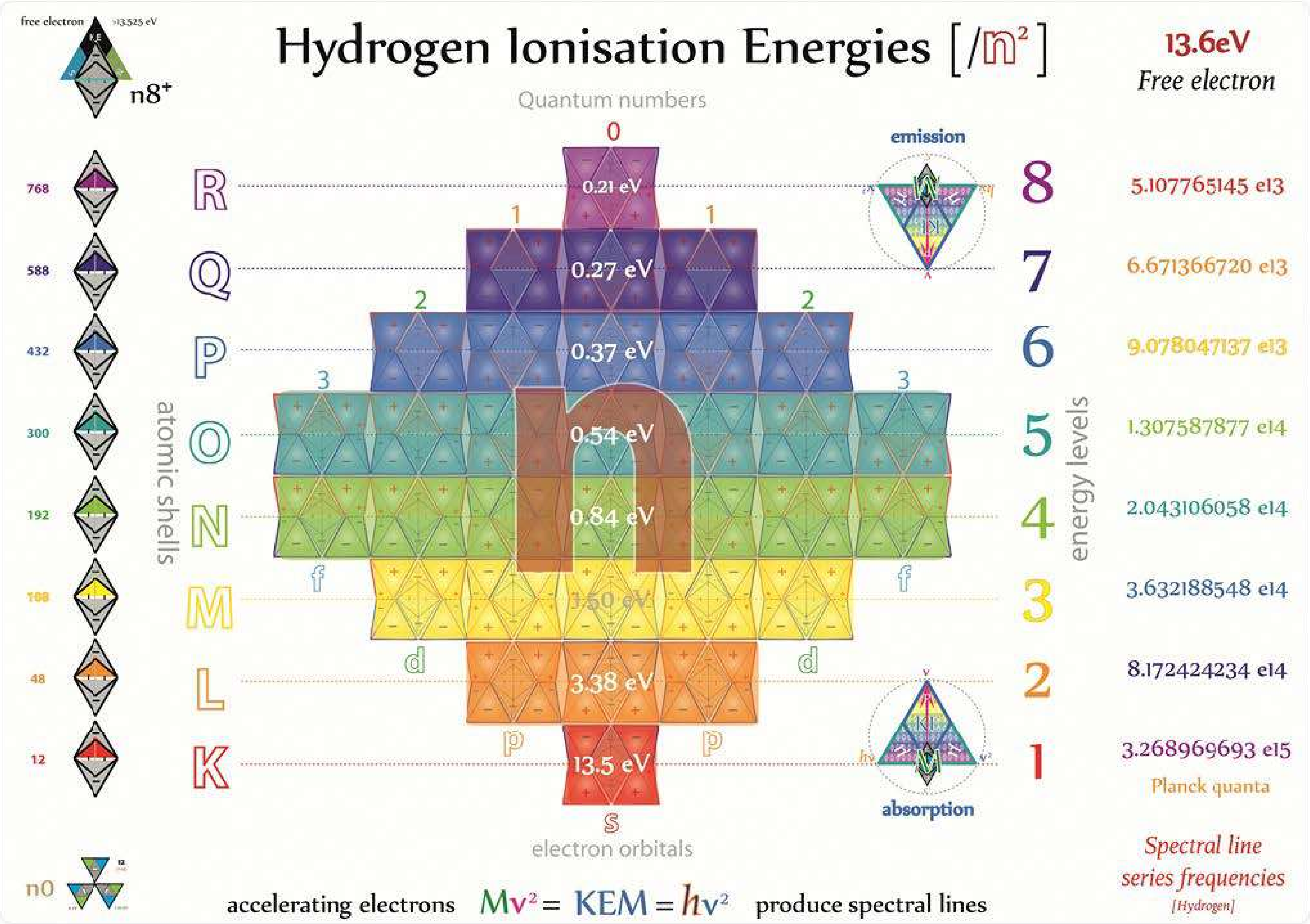
free electron
>13.525 eV
K.E.
n8⁺
13.6eV
Free electron
Quantum numbers
0
1
1
2
2
3
3
f
f
d
d
p
p
s
electron orbitals
emission
absorption
768
588
432
300
192
108
48
12
n0
atomic shells
R
Q
P
O
N
M
L
K
0.21 eV
0.27 eV
0.37 eV
0.54 eV
0.84 eV
1.50 eV
3.38 eV
13.5 eV
n
8
7
6
5
4
3
2
1
energy levels
5.107765145 e13
6.671366720 e13
9.078047137 e13
1.307587877 e14
2.043106058 e14
3.632188548 e14
8.172424234 e14
3.268969693 e15
Planck quanta
Spectral line
series frequencies
[Hydrogen]

accelerating electrons $Mv^2 = KEM = hv^2$ produce spectral lines

Redefining Atomic weights

Carbon 12

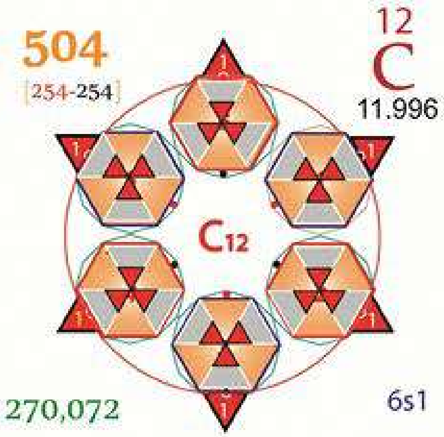

6 Protons	[24-12]	n1
6 Neutrons	[18-18]	
6 electrons	[0-12]	

Unified atomic Matter unit

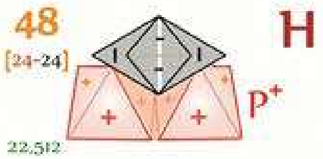

1.660538783e-27 kg

1 Proton	[24-12]	n1
1 electron	[0-12]	

Atomic weight (symbol: Ar) is a dimensionless physical quantity,
the ratio of the average mass of atoms of an element (from a given source) to
1/12 of the mass of an atom of carbon-12 (known as the unified atomic mass unit)

The 'unified atomic mass unit' currently in use is known to be inaccurate and must be corrected in order to bring clarity & increased accuracy to the atomic weights of all elements

$A_r = 22{,}512$
Hydrogen

Defining Hydrogen as having an exact atomic Planck mass of 22512n quanta provides uniformity with Tetryonics

Deuterium is the building block of all elements in the period table

$A_r = 45{,}012$
Deuterium

Defining Deuterium as having an exact atomic Planck mass of 45012n quanta reflects the true charged geometries of all Elements & their topologies

1/12 C_{12}

$A_r = 22{,}506$

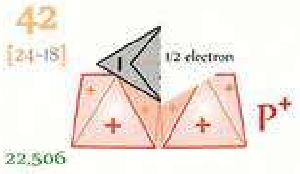

1.660096209e-27 kg

1 Proton	[24-12]	n1
.5 electron	0-6	

D

1/6 C

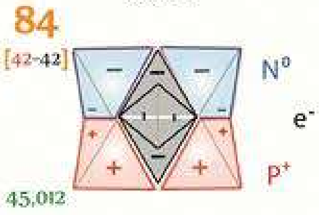

3.320192418e-27 kg

1 Proton	[24-12]	n1
1 Neutron	[18-18]	
1 electron	[0-12]	

Planck mass-energy units

NAu = 0.001 kg

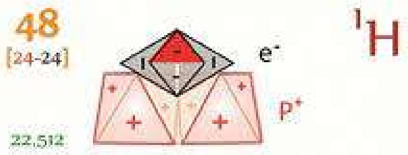
48
[24-24]
22,512
e⁻
P⁺
¹H

1.660538841 e-27 kg

One Da is approximately equal to the mass of one proton or one neutron

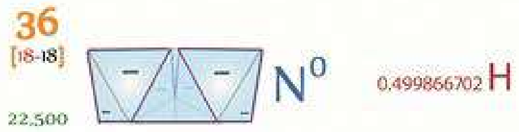
36
[18-18]
22,500
N⁰
0.499866702 H

1.659653693 e-27 kg

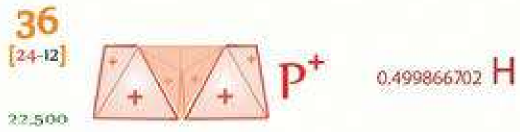
36
[24-12]
22,500
P⁺
0.499866702 H

Deuterium is the building block of all elements

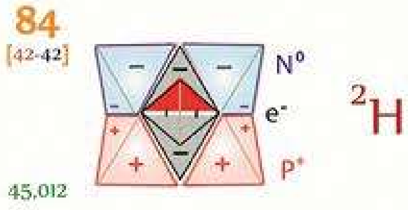
84
[42-42]
45,012
N⁰
e⁻
P⁺
²H

3.320192534 e-27 kg

The unified atomic mass unit (symbol: u) or Dalton (symbol: Da)
is a unit that is used for indicating mass on an atomic or molecular scale

270,072

1/12 the mass of a C12 graphene atom at rest in its electronic ground state

1.660538782(83)x10–27 kg

22,506

is an inaccurate means of determining the exact rest mass of a Hydrogen atom

22,512

Carbon 12 has 270,072n planck quanta
(270,072 / 12 = 22,506)

Hydrogen has a mass of 22,512n (22,500+12) requiring all mass to be calculated directly using the Planck mass-energy quantum (.001kg / N_A / 22,512) & Tetryonic charge geometries

Using Tetryonic theory to define
n Planck mass = $7.376238634 \times 10^{-32}$ Kg
(see Tetryonics QM 15.04)

exact atomic rest masses for all particles, elements and compounds can be determined directly from atomic theory

$N_A = 6.02214179 \text{ e}23$

The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12; its symbol is "mol".

NA(12u) = 0.012 kg

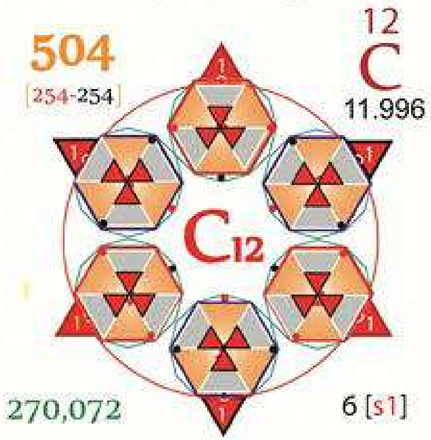
504
[254-254]
12
C
11.996
C12
270,072
6 [s1]

6 Protons [24-12]
6 Neutrons [18-18]
6 electrons [0-12]
n1

Carbon has a number of differing atomic configurations (allotropes)

6 Protons [24-12]
6 Neutrons [18-18]
6 electrons [0-12]
n1-2

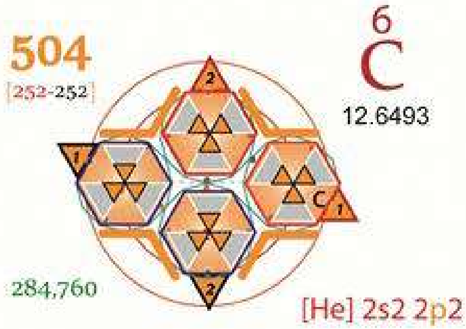
504
[252-252]
6
C
12.6493
284,760
[He] 2s2 2p2

Copyright ABRAHAM [2008] - All rights reserved

Planck mass-energy contributions to the measured weights of periodic, elementary mass-Matter topologies

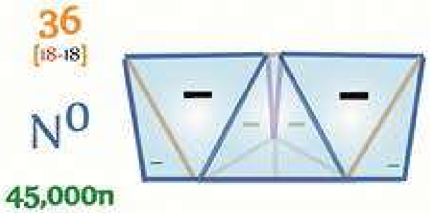

1.659653693 e-27 kg

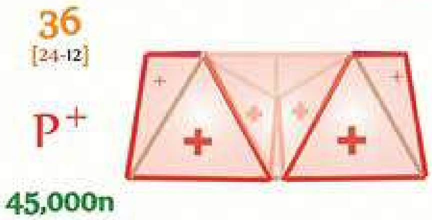

8.851486361 e-31 kg

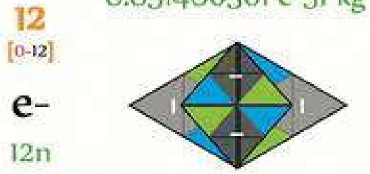

2.41110961 e-35 kg

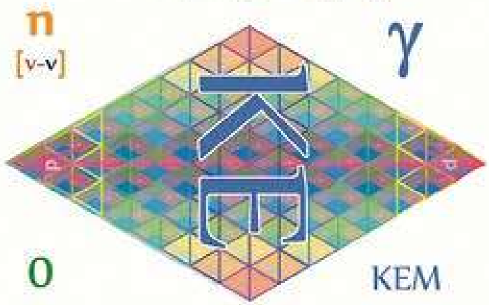

Baryons have 2,25e23 Planck quanta comprising their rest Matter topologies

[930.974 MeV]

Δ 1875 x

[496.5 keV]

Leptons have 1.2 e20 Planck quanta comprising their rest Matter topologies

[496.5 keV]

Δ 36,711 x

[13.6 eV]

Photons are planar geometries [Matter-less]
(purely Kinetic mass-Energy and momenta)

The Lyman alpha spectral line mass-energy contribution to the mass of a Deuterium nucleus is negligible

Electron quantum level energies are determined by the energy of the Nuclei they bind to in elements

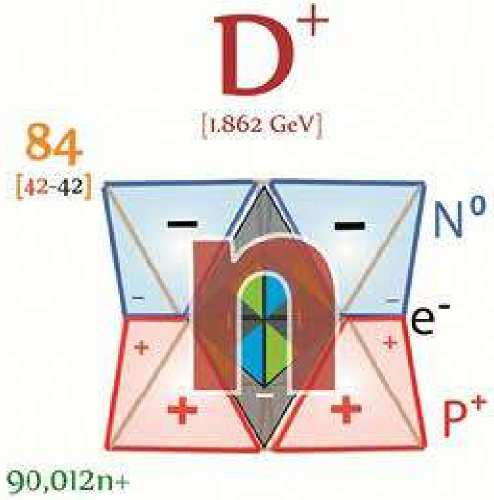

Photons contribute spectral mass-energies to the nuclei mass but are themselves Matterless [2D zero rest mass-energies]

Photons are 2π charge mass-energy geometries

Ionisation energies

Fractional quantum differentials

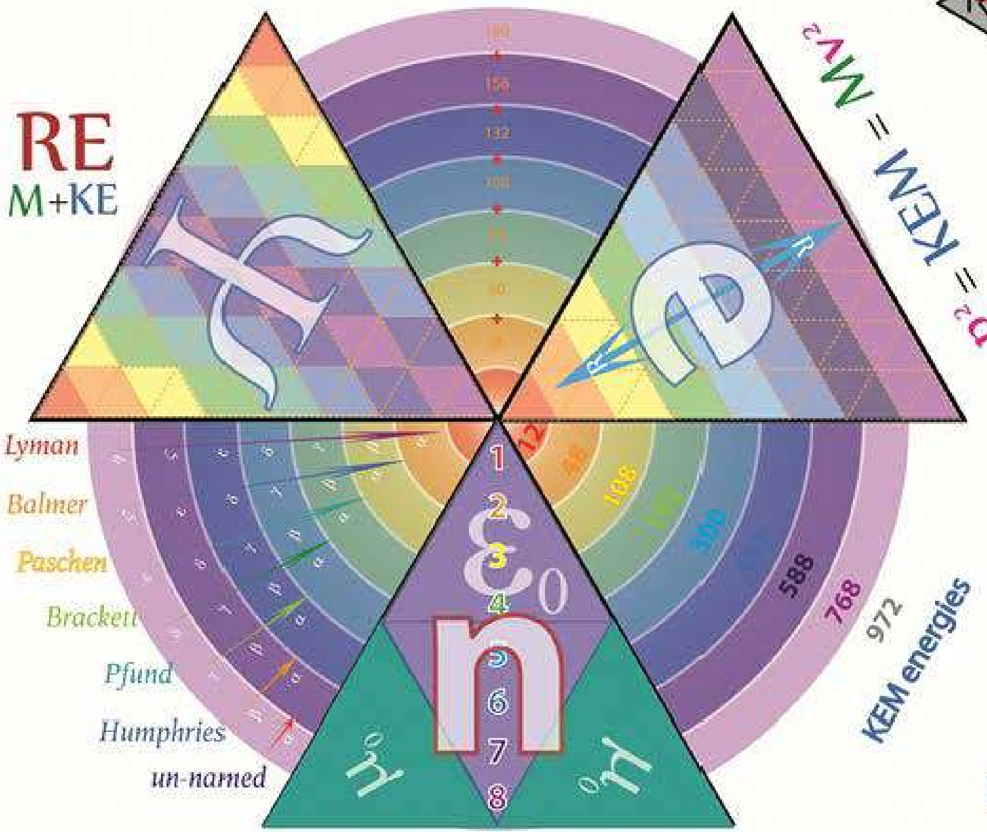
RE
M+KE
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
p² = KEM = Mv²
KEM energies
12
48
108
192
300
432
588
768
972
1
2
3
4
5
6
7
8
ε0
n

$$hf = \Delta M v = \Delta p$$

spectral lines are produced by accelerating electrons.

c^2

Note: this is an Illustrative schema for modelling KEM field energies
All KEM fields possess the same physical spatial geometry
in radial-time defined spatial co-ordinate systems

c^4

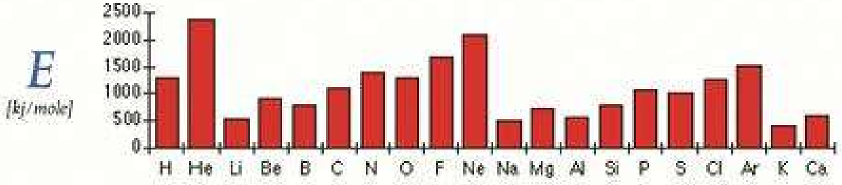

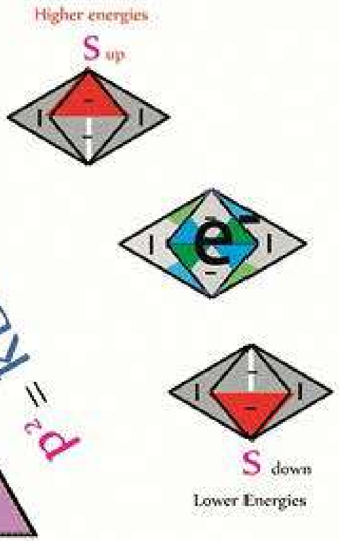
Higher energies
S up
e
S down
Lower Energies

Photo-electrons absorb/emit spectral energies

$$E = eV = \frac{1}{4\pi\epsilon_0}\frac{ne^2}{a}$$

photo-electronic energy momenta

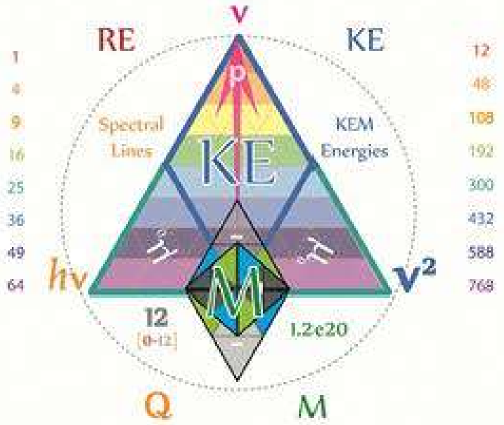
RE
KE
v
p
KE
Spectral Lines
KEM Energies
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768
hv
v²
M
12
[0-12]
1.2e20
Q
M

Mapping photo-electron transition energies
to Tetryonic energy momenta geometries
reveals many key facts about the
ionisation energies of nuclei

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

The differing fractional KEM field energy momenta of electrons
that results from their transitions to specific energy nuclei
in elements results in differing QAM quanta and
produces spectral lines and fine line splitting

Spectral line transitions

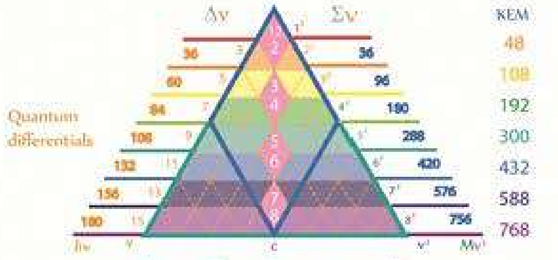
Δv
Σv
KEM
48
108
192
300
432
588
768
Quantum differentials
36
60
84
108
132
156
180
36
96
180
288
420
576
756
hv
v
c
v²
Mv²

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Elementary ionisation energies

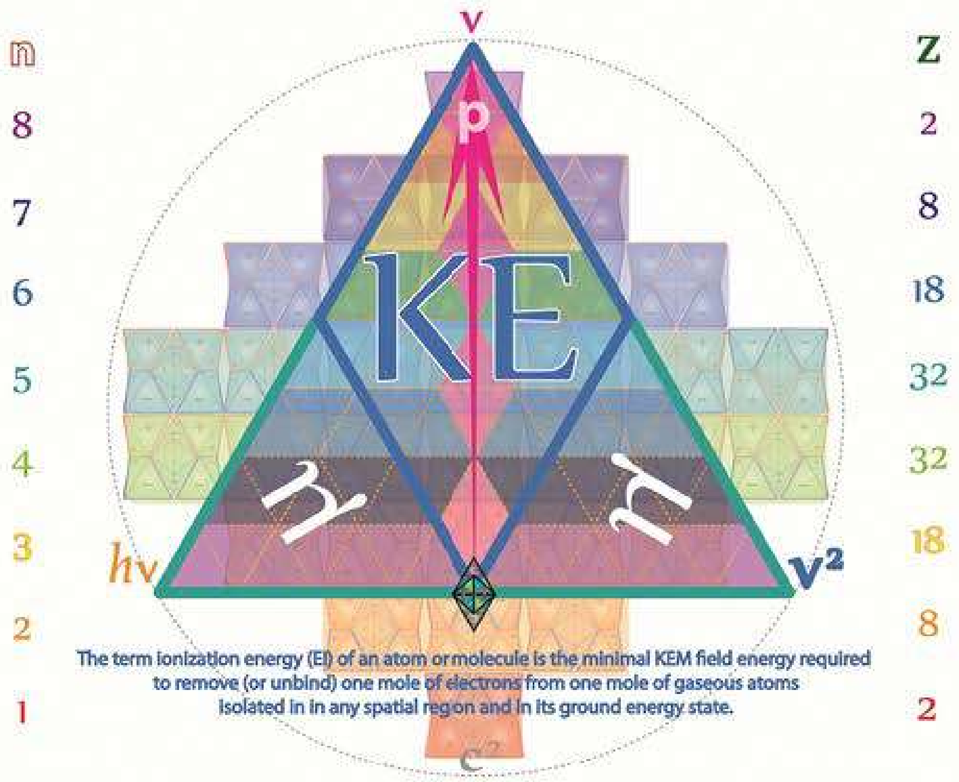

The term ionization energy (EI) of an atom or molecule is the minimal KEM field energy required to remove (or unbind) one mole of electrons from one mole of gaseous atoms isolated in in any spatial region and in its ground energy state.

The term "ionization energy" is sometimes used as a name for the work needed to remove (or un-bind) the highest energy photoelectron from an atom or molecule.

However, due to interactions with surfaces, this value differs from the ionization energy of the atom or molecule in question when it is located by itself in free space.

So, in the case of surface-adsorbed atoms and molecules, it may be better to use the more general term "electron binding energy", in order to avoid confusion.

Both these names are also sometimes used to describe the work needed to remove an electron from a "lower" orbital (i.e., not the topmost orbital) for both free and adsorbed atoms; in such cases it is necessary to specify the orbital from which the electron has been removed

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

Every electron in each elementary orbit has a unique ionisation energy

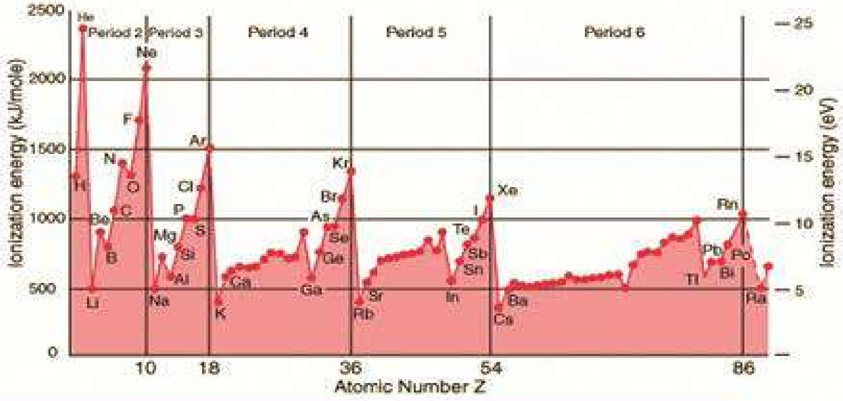

Nuclei	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	per shell	
													Element number																					
R	119	120																															8	2
Q	87	88	113	114	115	116	117	118																									7	8
P	55	56	81	82	83	84	85	86	103	104	105	106	107	108	109	110	111	112															6	18
O	37	38	49	50	51	52	53	54	71	72	73	74	75	76	77	78	79	80	89	90	91	92	93	94	95	96	97	98	99	100	101	102	5	32
N	19	20	31	32	33	34	35	36	39	40	41	42	43	44	45	46	47	48	57	58	59	60	61	62	63	64	65	66	67	68	69	70	4	32
M	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	29	30															3	18
L	3	4	5	6	7	8	9	10																									2	8
K	1	2																															1	2
Shells	s				p								d														f						Energy level	
													Orbitals																					
	s1	s2	p1	p2	p3	p4	p5	p6	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14		
													sub-orbitals																					

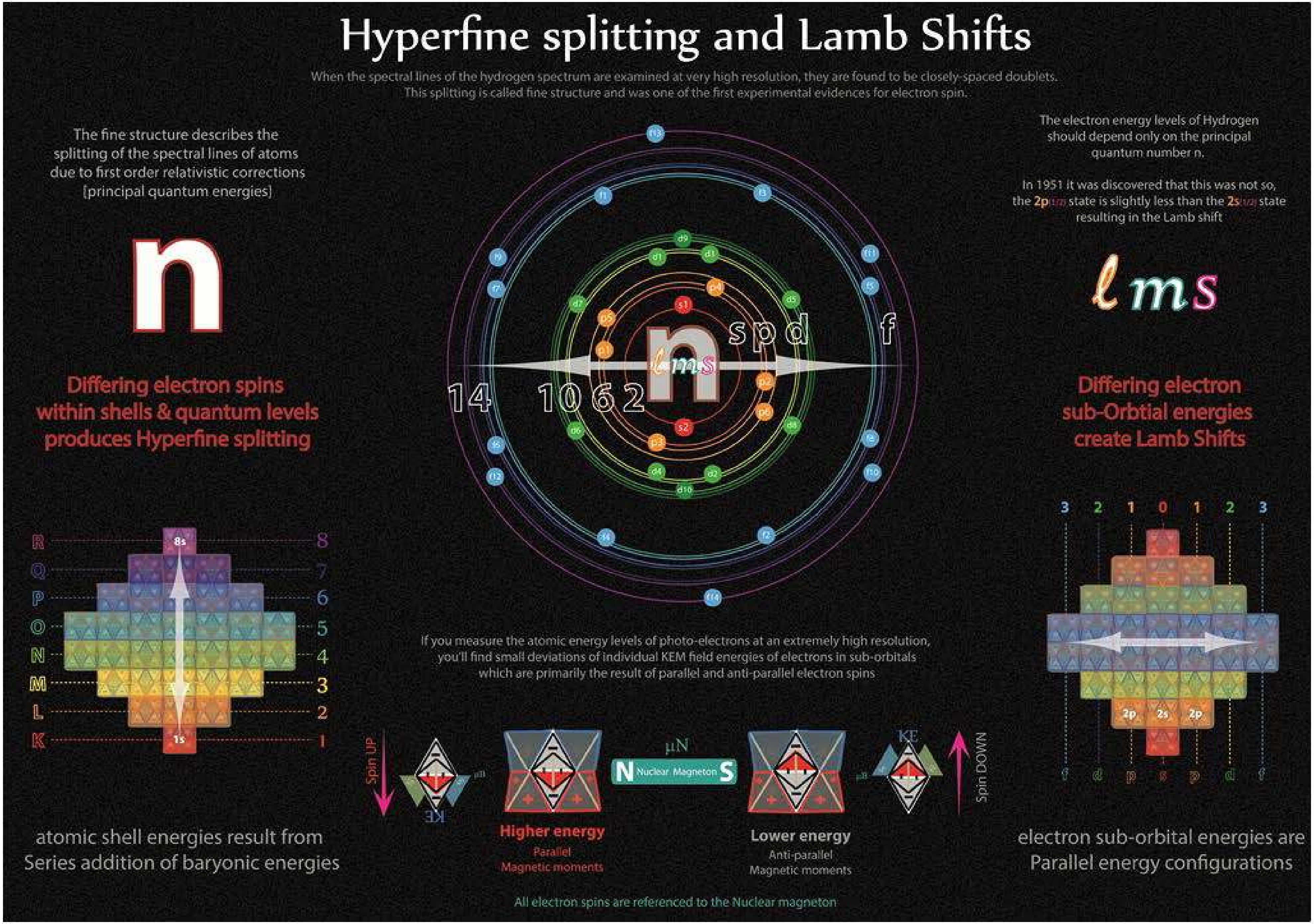

Hyperfine splitting and Lamb Shifts

When the spectral lines of the hydrogen spectrum are examined at very high resolution, they are found to be closely-spaced doublets.
This splitting is called fine structure and was one of the first experimental evidences for electron spin.

The fine structure describes the splitting of the spectral lines of atoms due to first order relativistic corrections [principal quantum energies]

n

Differing electron spins within shells & quantum levels produces Hyperfine splitting

The electron energy levels of Hydrogen should depend only on the principal quantum number n.

In 1951 it was discovered that this was not so, the $2p_{1/2}$ state is slightly less than the $2s_{1/2}$ state resulting in the Lamb shift

ℓ m s

Differing electron sub-Orbtial energies create Lamb Shifts

If you measure the atomic energy levels of photo-electrons at an extremely high resolution, you'll find small deviations of individual KEM field energies of electrons in sub-orbitals which are primarily the result of parallel and anti-parallel electron spins

atomic shell energies result from Series addition of baryonic energies

Higher energy
Parallel
Magnetic moments

Lower energy
Anti-parallel
Magnetic moments

electron sub-orbital energies are Parallel energy configurations

All electron spins are referenced to the Nuclear magneton

f13
f1
f3
f9
f11
f7
f5
d9
d1
d11
p4
d7
s1
d5
p5
n
s p d
f
p1
ℓ m s
14
10 6 2
p2
p6
d6
s2
d8
p3
f6
f8
f10
f12
d4
d2
d10
f4
f2
f14
R
Q
P
O
N
M
L
K
8s
1s
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
2p 2s 2p
f d p s p d f
Spin UP
KE
μB
μN
N Nuclear Magneton S
KE
μB
Spin DOWN

M_0

RE

KE

3D Matter topologies are comprised of charged 2D mass-energies

Relativistic mass-ENERGY-Matter

Relativity fails at the foundational level to explain and differentiate between mass-ENERGY and Matter in physical systems

2D equilateral mass-energies are euclidean geometries

Energy per second²

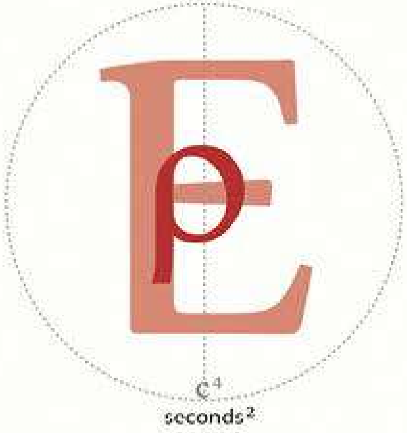
c^4
seconds²

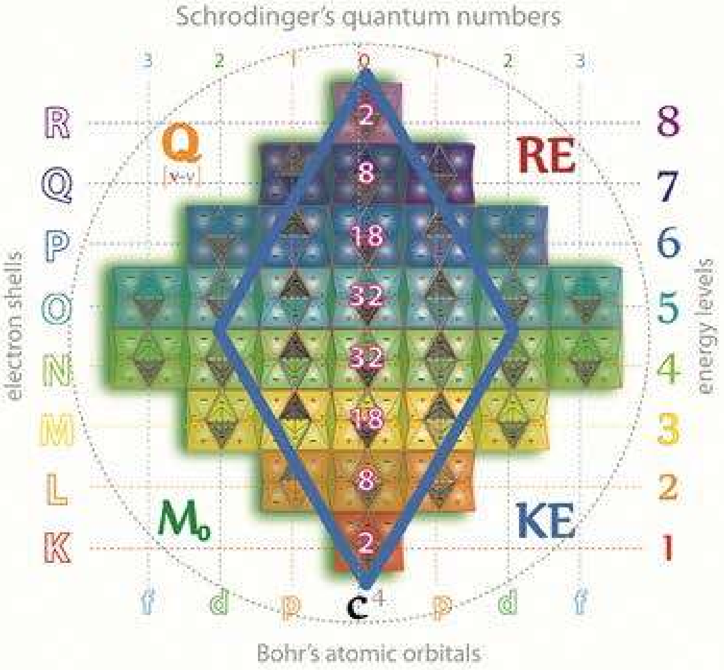
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M_0
KE
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f
d
p
c^4
p
d
f
Bohr's atomic orbitals

Energy per second

Einstein's relativistic [Lorentz corrected] stress energy tensor aggregates all forms of energy into a single energy density gradient

Matter Planck quanta

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

mass velocity

atomic energies

$$\frac{Tmn}{c^4} + \frac{mv^2}{c^2}$$

electron spins

mass Planck quanta

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

mass velocity

mass-Matter

standing wave mass-energiescreate
the material substance of all chemical elements

3D rest Matter + Lorentz corrected 2D Kinetic Energies = total Relativistic Energies

mass-energies

radiant planar mass-energies create
EM fields, spectral lines & chemical interactions

Deuterium is the building block of all Elements (save Hydrogen)

The geometry of any Element is determined by its Charge

The Molar Weight of any Element is a measure of its standing wave mass-energies

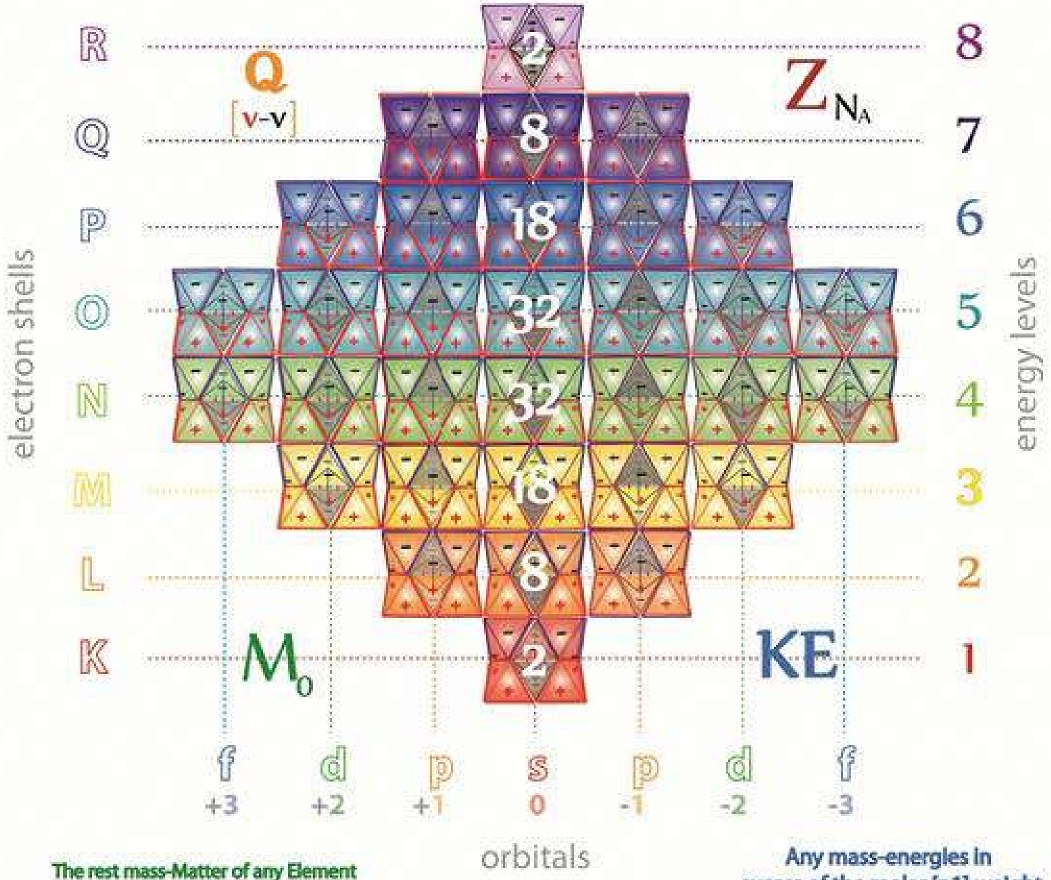

The rest mass-Matter of any Element is determined by the total number of quanta making up the Protons, Neutrons and electrons that comprise them (in their respective energy levels)

Any mass-energies in excess of the molar [n1] weight is a measurement of a element's CHEMICAL energies

Important point to note:

The Kinetic Energy difference between any Element's total [n1] Deuteron mass-energies and its Molar mass has historically been incorrectly explained as resulting from an excess number of Neutrons in the atom it is not. Z# = (number of Protons = number of electrons = number of Neutrons)

Elementary mass-Matter

$$\sum_{K}^{R}$$

	n per nuclei 1e19v = n	
2	2 nuclei [74,496 ea]	120 Unbinilium 119 Ununennium
	+	
8	8 nuclei [69,780 ea]	118 Ununoctium 87 Francium
	+	
18	18 nuclei [65,232 ea]	112 Copernicium 55 Caesium
	+	
32	32 nuclei [60,852 ea]	102 Nobelium 37 Rubidium
	+	
32	32 nuclei [56,640 ea]	70 Ytterbuim 19 Potassium
	+	
18	18 nuclei [52,596 ea]	30 Zinc 11 Sodium
	+	
8	8 nuclei [48,720 ea]	10 Neon 3 Lithium
	+	
2	2 nuclei [45,012 ea]	2 Helium 1 Deuterium

The rest mass-Matter of any Element is the sum total of its constituent Z[n²] energy level Deuterium nuclei

Aufbau

$$Z\#\left[\begin{array}{ll} z \text{ Protons} & [24\text{-}12] \\ z \text{ Neutrons} & [18\text{-}18] \\ z \text{ electrons} & [0\text{-}12] \end{array}\right] n1\text{-}8$$

(ie Calcium [20] = 2+8+10 n level Deuterium nuclei)

Copyright ABRAHAM [2008] - All rights reserved

Periodic Harmonic motions

$x = A \cos(\omega t + \varphi)$

Circular motion

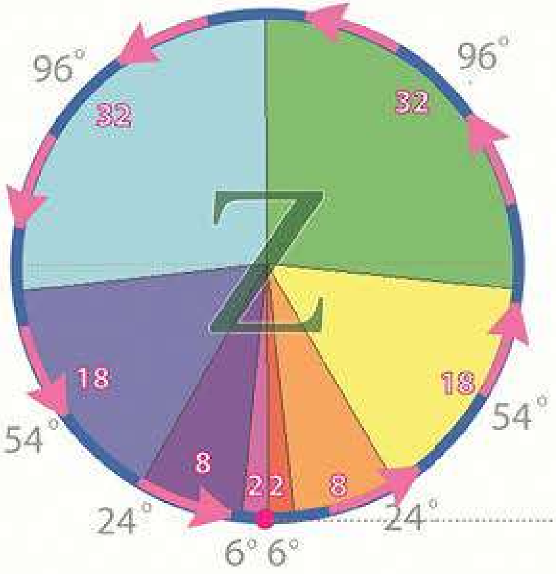
96°
96°
32
32
Z
18
18
54°
54°
8
2 2
8
24°
24°
6° 6°

circular harmonic motion

**Circular motions describe
the motion of a body
with a changing velocity vector
[the result of an acceleration force].**

Much of the math in of modern physics
is predicated on the assumption that
π [where it appears] is related to
the properties of a circle

$P\sum 2(n^2)$

Simple harmonic motion can be visualized as the projection of uniform circular motion onto one axis

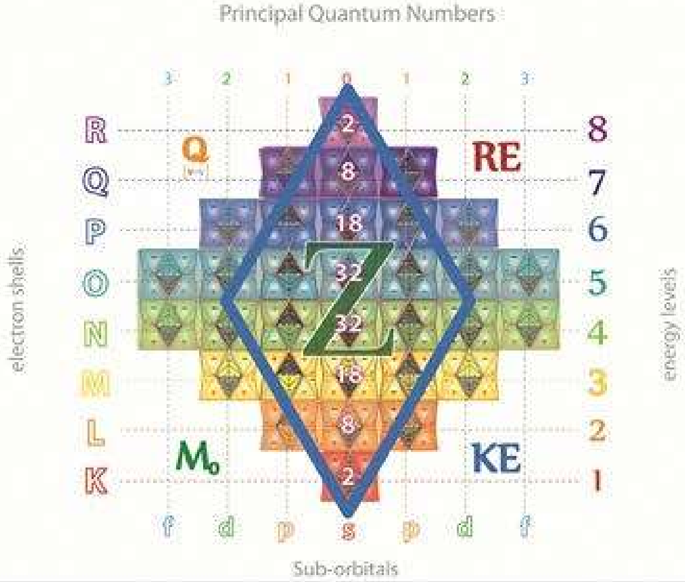
Principal Quantum Numbers
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
2
8
18
32
Z
32
18
8
2
M_0
KE
8
7
6
5
4
3
2
1
electron shells
energy levels
f d p s p d f
Sub-orbitals

$F = -kx$

Linear motion

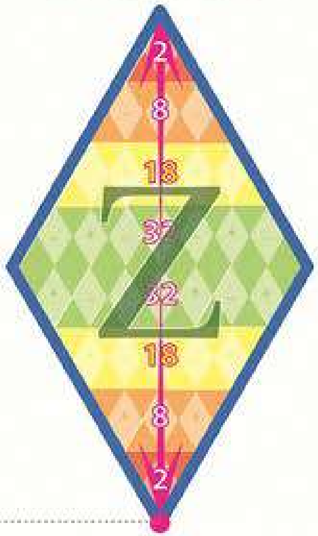
2
8
18
32
Z
32
18
8
2

simple harmonic motion

**Nuclei per shell in elements follows
a 'periodic summation rule'
that is reflective of
photonic energies**

Copyright ABRAHAM [2008] - All rights reserved

$$\sum_{K}^{R}$$

nuclei per shell

$$_{P}\sum_{K}^{R} Z$$

Periodic Summation

Periodic summation is a notation developed for Tetryonic theory to model the geometric series addtion of Z[n²] energy level Deuterium nuclei that form the periodic elements

$$\sum_{1}^{120} Z$$

element number

STEP ONE

Periodic summation follows the atomic shell electron config

STEP TWO

Periodic elements build up following the aufbau sequence

1 R
2
3
4
4
3
2
1 K

$$_{P}\sum_{K}^{R} 2(x^2) =$$

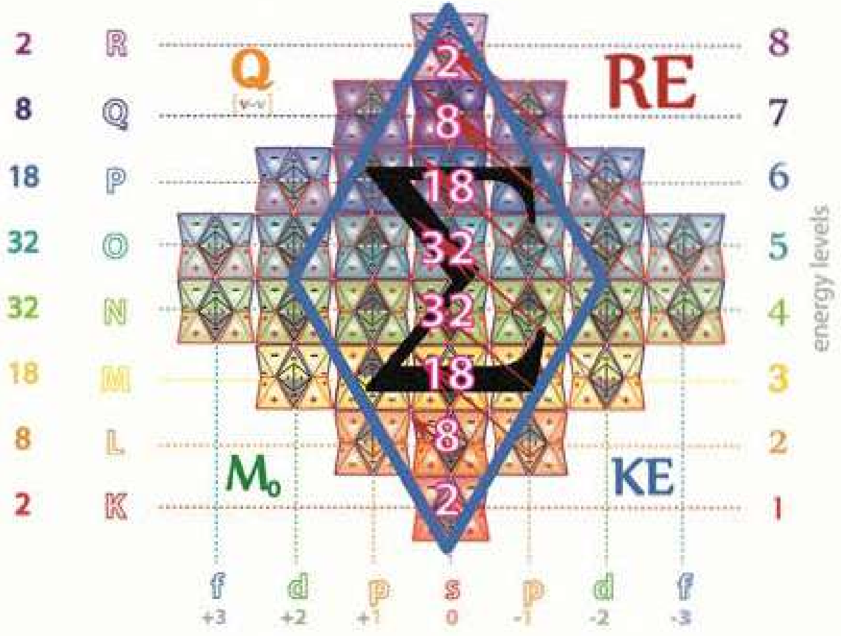

Each atomic shell can hold only a fixed number of deuterium nuclei

$_{P}\sum R = 2$	2 nuclei [74,496 ea]	120	Unbinilium
	+		
$_{P}\sum Q = 8$	8 nuclei [69,780 ea]	118	Ununoctium
	+		
$_{P}\sum P = 18$	18 nuclei [65,232 ea]	110	Darmstadtium
	+		
$_{P}\sum O = 32$	32 nuclei [60,852 ea]	92	Uranium
	+		
$_{P}\sum N = 32$	32 nuclei [56,640 ea]	60	Neodymuim
	+		
$_{P}\sum M = 18$	18 nuclei [52,596 ea]	28	Argon
	+		
$_{P}\sum L = 8$	8 nuclei [48,720 ea]	10	Neon
	+		
$_{P}\sum K = 2$	2 nuclei [45,012 ea]	2	Helium
		0	Hydrogen

THe LHS of the notation determine the number of nuclei in each atomic shell, from the periodic mass-energy levels for atoms, and the RHS follows the aufbau building principle to determine the rest mass-Matter of any specific element

Aufbau

Each periodic element is made of Z [n² energy] deuterium nuclei

$$Z\# \left[\begin{array}{ll} z \text{ Protons} & [24\text{-}12] \\ z \text{ Neutrons} & [18\text{-}18] \\ z \text{ electrons} & [0\text{-}12] \end{array} \right] n1\text{-}8$$

Planck mass-energies form the surface integral of rest Matter topologies for each periodic element

Element numbers

Nuclei per shell in elements follow
a 'periodic summation rule'
that is reflective of
photonic energies

$$\sideset{^{\mathbb{R}}_{\mathbb{K}}}{}{\;} {}^{P}\sum 2\left(x^2\right) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

1 2 3 4 4 3 2 1

Z

120	Unbinilium	n8
119	Ununennium	
118	Ununoctium	n7
87	Francium	
112	Copernicium	n6
55	Caesium	
102	Nobelium	n5
37	Rubidium	
70	Ytterbuim	n4
19	Potassium	
30	Zinc	n3
11	Sodium	
10	Neon	n2
3	Lithium	
2	Helium	n1
1	Deuterium	
Hydrogen	0	

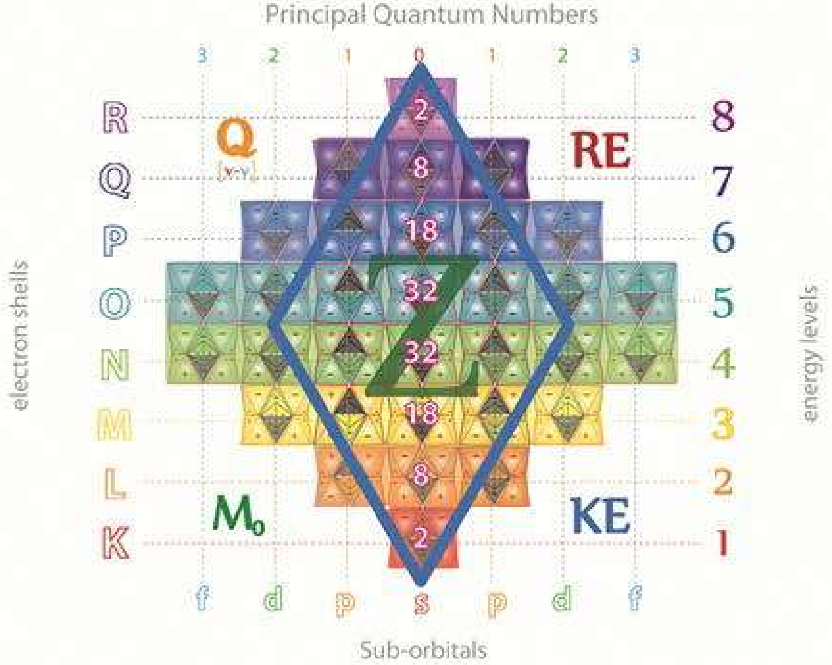

Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
M₀
KE
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
electron shells
energy levels
f
d
p
s
p
d
f
Sub-orbitals

Periodic mass-ENERGY-Matter

Following periodic summation rules for shell filling
n[1-8] quantum energy deuterium nuclei
combine to form elementary Matter

$$\sideset{^{\mathbb{R}}_{\mathbb{K}}}{}{\;} {}^{P}\sum_{25}^{32}\left[\underbrace{\overbrace{\left[72(n)^2\right]}^{\text{Baryon rest masses}}+\overbrace{\left[12e19\right]}^{\text{lepton rest mass}}+\overbrace{\left[m_e v^2\right]}^{\text{KEM}}}_{\text{Deuterium mass-energy per shell}}\right]_1^8$$

The measured weight of Matter in gravitational fields is the result of
planar mass-energies in tetryonic standing-wave geometries

The periodicity of all the elements,
along with their exact molar rest mass-energies and
quantum wavefunctions can be described with Tetryonic geometries

Ionisation energies

γ

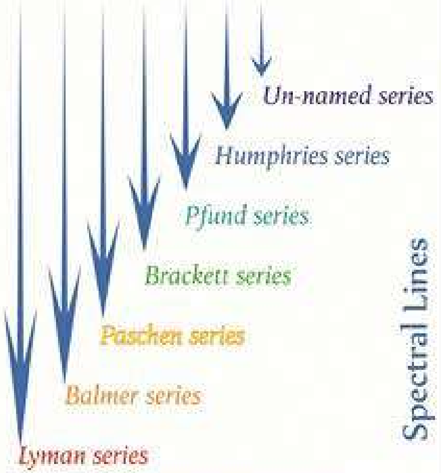

Un-named series
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
Spectral Lines

$$\boxed{Mv^2 = \text{KEM} = \hbar c R_H}$$

Photon emission/absorption

Copyright ABRAHAM [2008] - All rights reserved

Photo-electron ionisation energies

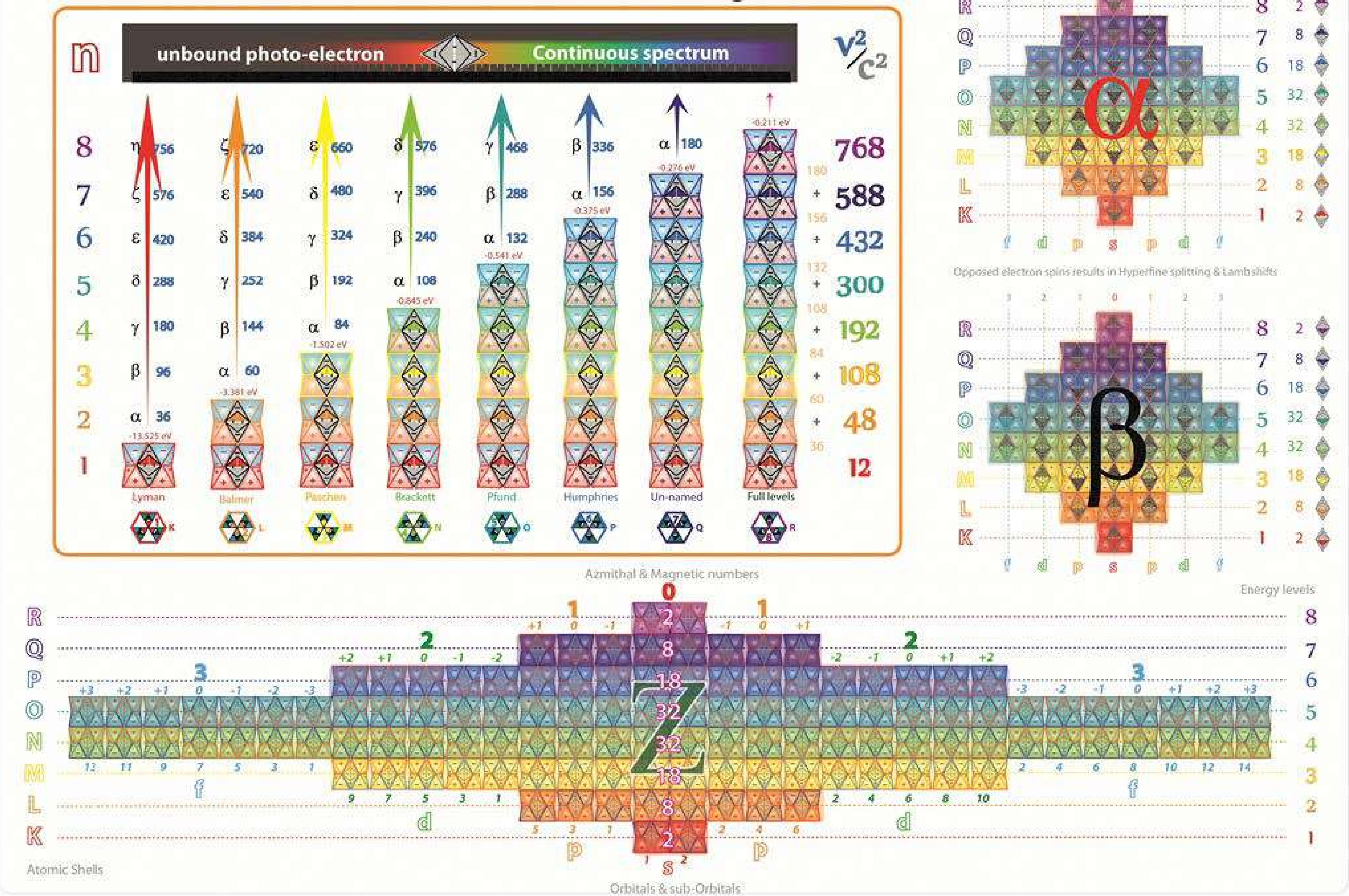
n
unbound photo-electron
Continuous spectrum
v²/c²
8
7
6
5
4
3
2
1
η 756
ζ 576
ε 420
δ 288
γ 180
β 96
α 36
-13.525 eV
ζ 720
ε 540
δ 384
γ 252
β 144
α 60
-3.381 eV
ε 660
δ 480
γ 324
β 192
α 84
-1.502 eV
δ 576
γ 396
β 240
α 108
-0.845 eV
γ 468
β 288
α 132
-0.541 eV
β 336
α 156
-0.375 eV
α 180
-0.276 eV
-0.211 eV
180
156
132
108
84
60
36
768
588
432
300
192
108
48
12
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
Full levels
K
L
M
N
O
P
Q
R
α
β
f d p s p d f
8 2
7 8
6 18
5 32
4 32
3 18
2 8
1 2
Opposed electron spins results in Hyperfine splitting & Lamb shifts
Azmithal & Magnetic numbers
Energy levels
0
1
2
3
+1 0 -1
+2 +1 0 -1 -2
+3 +2 +1 0 -1 -2 -3
2
8
18
32
32
18
8
2
Z
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f
d
p
s
Atomic Shells
Orbitals & sub-Orbitals

Tetryonics 50.05 - Photo-electron ionisation energies

Eigenstate - Ionisation energies

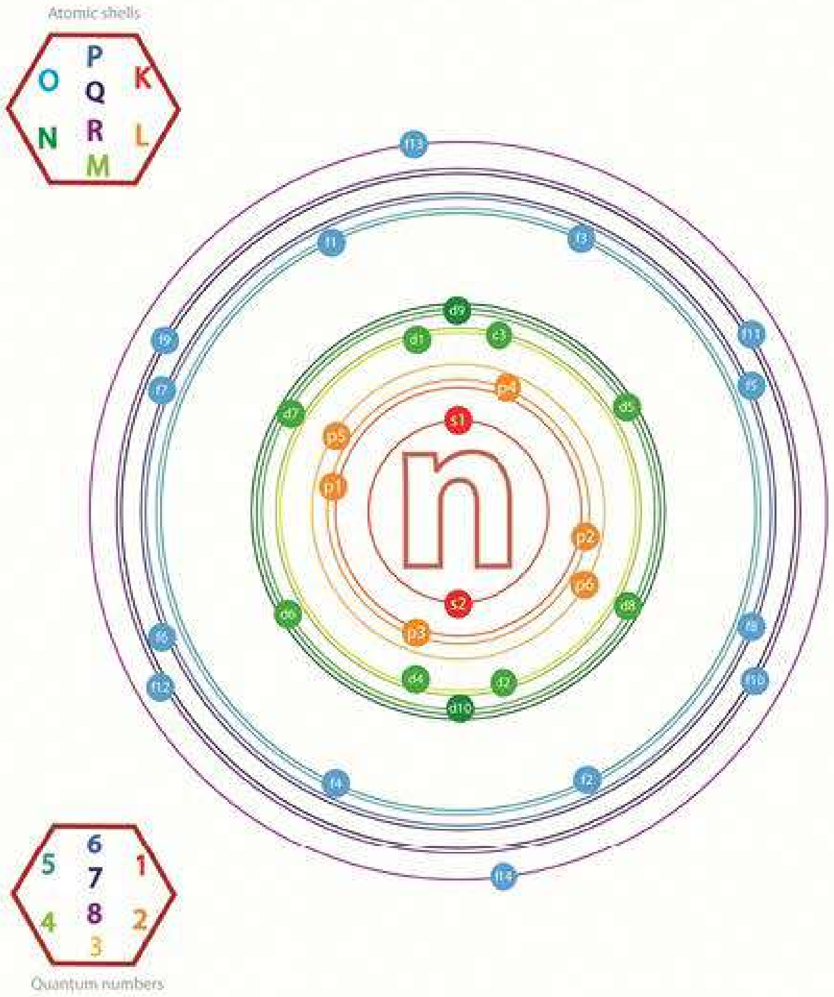
Atomic shells
P
O Q K
N R L
M
n
s1
s2
p1 p2 p3 p4 p5 p6
d1 d2 d3 d4 d5 d6 d7 d8 d9 d10
f1 f2 f3 f4 f5 f6 f7 f8 f9 f10 f11 f12 f13 f14
6
5 7 1
4 8 2
3
Quantum numbers

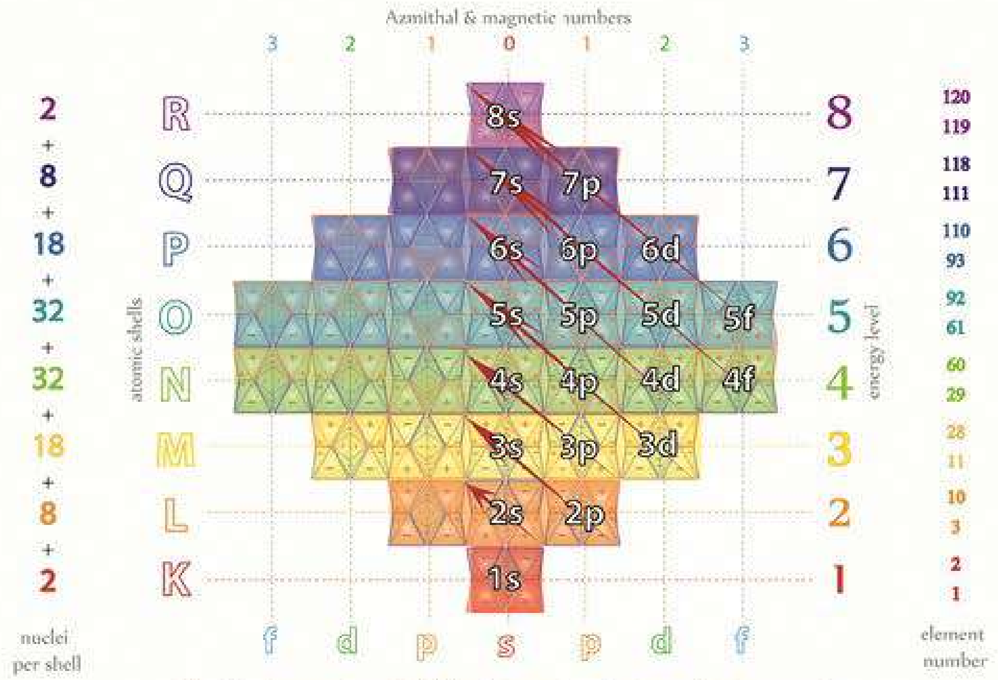
Azmithal & magnetic numbers
3 2 1 0 1 2 3
2 + 8 + 18 + 32 + 32 + 18 + 8 + 2
nuclei per shell
atomic shells
R Q P O N M L K
8s
7s 7p
6s 6p 6d
5s 5p 5d 5f
4s 4p 4d 4f
3s 3p 3d
2s 2p
1s
f d p s p d f
8 7 6 5 4 3 2 1
energy level
120 119 118 111 110 93 92 61 60 29 28 11 10 3 2 1
element number

The ionisation energies of individual atoms varies due to many factors, namely:
electron spin-orbital coupling with Baryons of specific energies,
the relativistic energies of photo-electrons bound in nuclei
and Zitterbewegung effects on bound electrons

First ionisation energies for all periodic elements

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

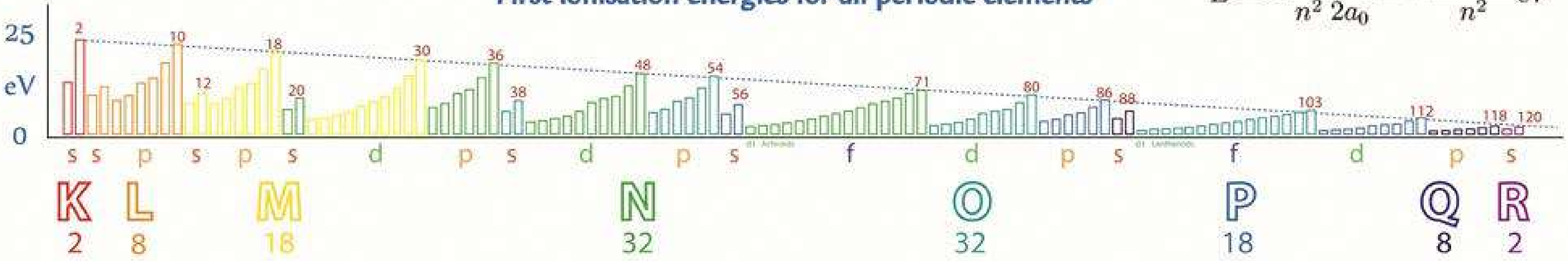

Copyright ABRAHAM [2008] - All rights reserved

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number.
It is used as rule to determine which nuclei are stable or unstable.

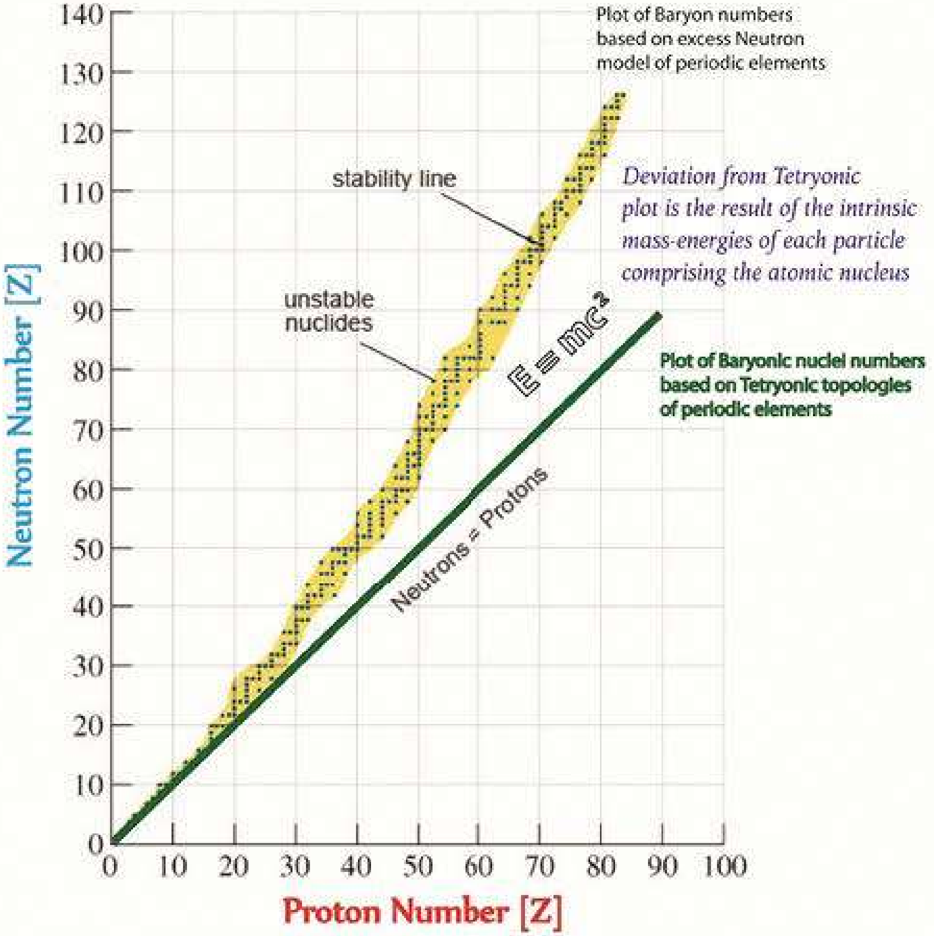

Historically, Proton-electron numbers are viewed as being equivalent
in neutral elementary matter with the excess molar mass measured
being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of
Protons, Neutrons & Electrons with their molar mass-Matter
being determined by their quantum level mass-energies

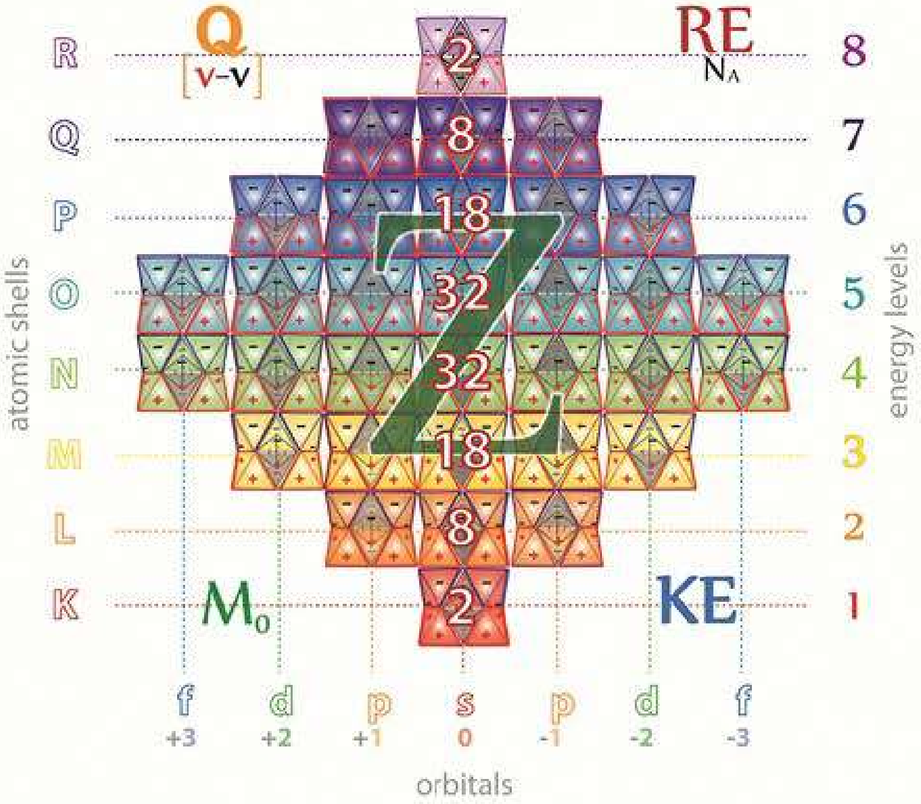
Q
[v-v]
RE
N_A
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
M_0
KE
atomic shells
energy levels
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
orbitals

Tetryonic modelling of the charged mass-ENERGY-Matter topologies
of elementary atoms and the nuclei that comprise them, reveals a DIRECT
LINEAR relationship for the number of Protons-electrons-Neutrons
in all periodic elements and nuclear isotopes

Planck mass-energy contributions to elementary Matter and isotopes

The mass-energy content of Deuterium nuclei creates the molar mass of elements
[not extra neutrons in excess of the elemental number]

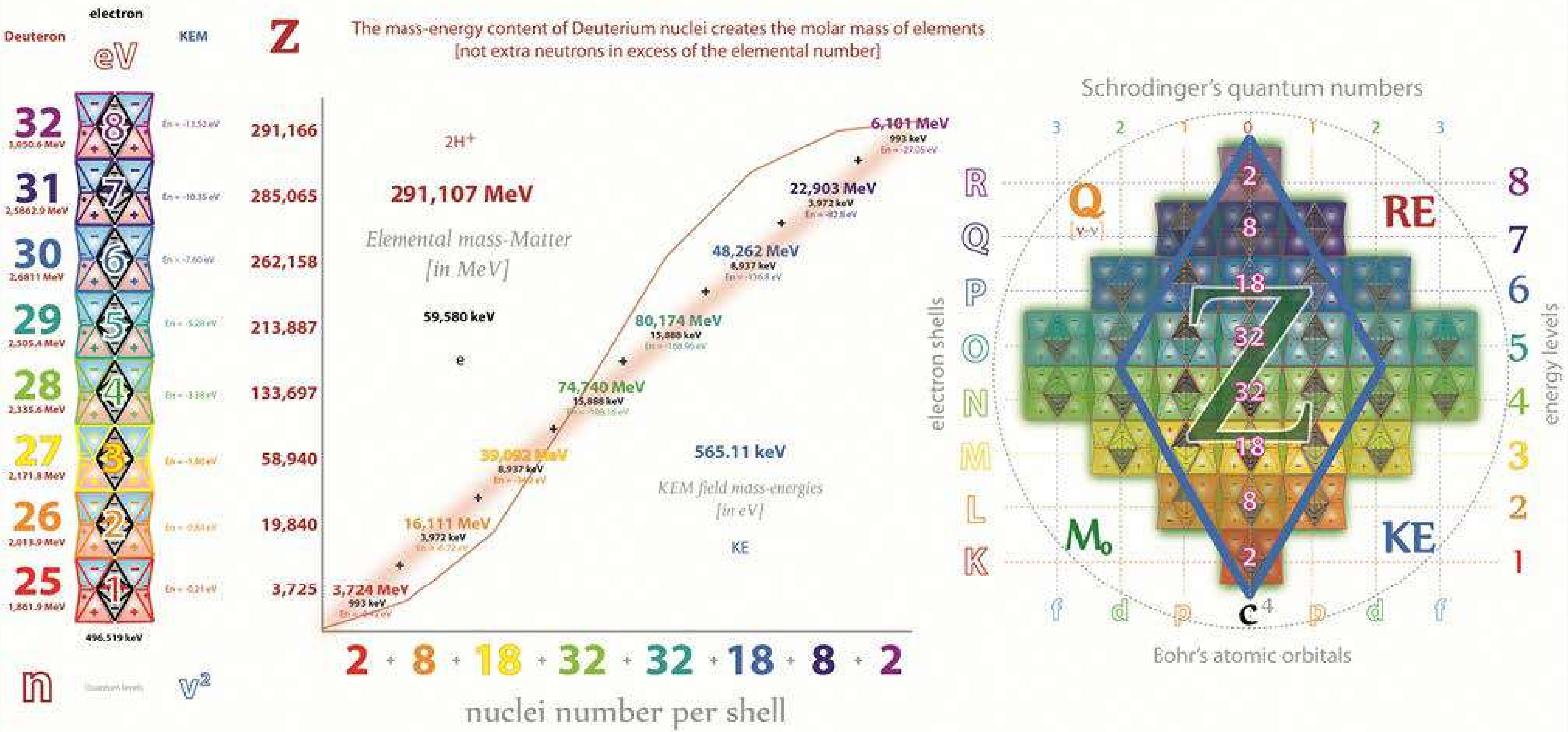
Deuteron
electron
eV
KEM
Z
32
3,050.6 MeV
31
2,5862.9 MeV
30
2,6811 MeV
29
2,505.4 MeV
28
2,335.6 MeV
27
2,171.8 MeV
26
2,013.9 MeV
25
1,861.9 MeV
8
7
6
5
4
3
2
1
496.519 keV
En = -13.52 eV
En = -10.35 eV
En = -7.60 eV
En = -5.28 eV
En = -3.38 eV
En = -1.90 eV
En = -0.84 eV
En = -0.21 eV
n
Quantum levels
v²
291,166
285,065
262,158
213,887
133,697
58,940
19,840
3,725
2H⁺
291,107 MeV
Elemental mass-Matter
[in MeV]
59,580 keV
e
6,101 MeV
993 keV
En = -27.06 eV
22,903 MeV
3,972 keV
48,262 MeV
8,937 keV
80,174 MeV
15,888 keV
74,740 MeV
15,888 keV
39,092 MeV
8,937 keV
16,111 MeV
3,972 keV
3,724 MeV
993 keV
565.11 keV
KEM field mass-energies
[in eV]
KE
2 + 8 + 18 + 32 + 32 + 18 + 8 + 2
nuclei number per shell
electron shells
R
Q
P
O
N
M
L
K
Schrodinger's quantum numbers
3
2
1
0
1
2
3
Q
[v·v]
RE
M₀
KE
Z
2
8
18
32
32
18
8
2
c
4
8
7
6
5
4
3
2
1
energy levels
f
d
p
p
d
f
Bohr's atomic orbitals

n1
Elementary nuclei

are comprised of equal numbers of
Protons, Neutrons & electrons
with varying energy levels

Baryons		electrons		KEM fields
930.947 MeV	+	**496.519 keV**	+	**13.525 ev**

The mass-energy content of Matter topologies is velocity invariant

The mass-energy content of Baryons determines the KEM field of electrons

Copyright ABRAHAM [2008] - All rights reserved

Baryons 930.947 MeV + **KEM fields** 13.525 ev + **electrons** 496.519 keV

Mapping Planck mass-energy contributions to elementary Matter and isotopes

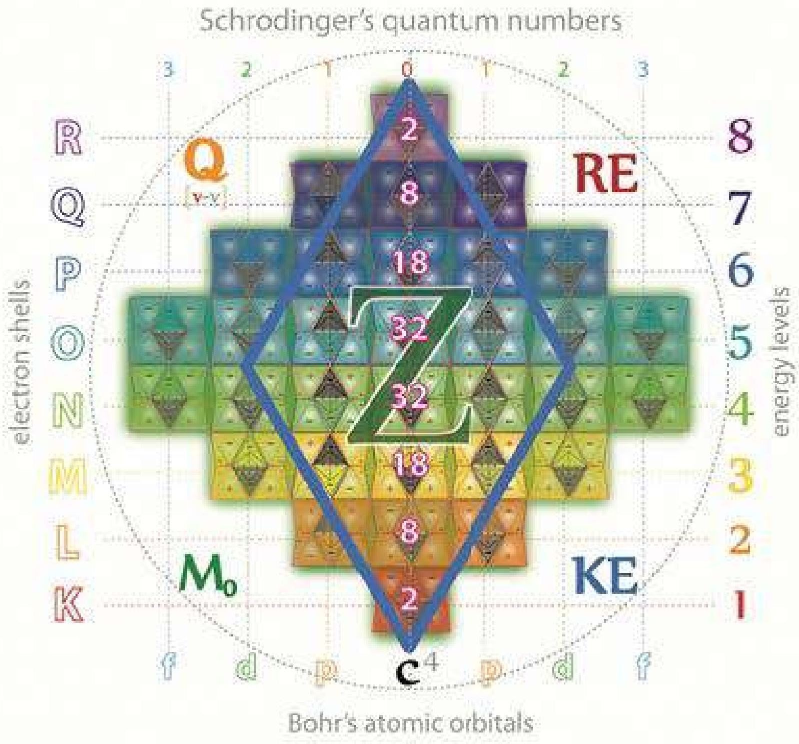
Schrodinger's quantum numbers
3
2
1
0
1
2
3
electron shells
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
Z
2
8
18
32
32
18
8
2
M₀
KE
8
7
6
5
4
3
2
1
energy levels
f
d
p
c⁴
p
d
f
Bohr's atomic orbitals

$$E = h\nu^2$$

$$n^2 + v^2 + e = Z$$

general form quatratic equation

$$ax^2 + bx + c = Z$$

$$E = nhv$$

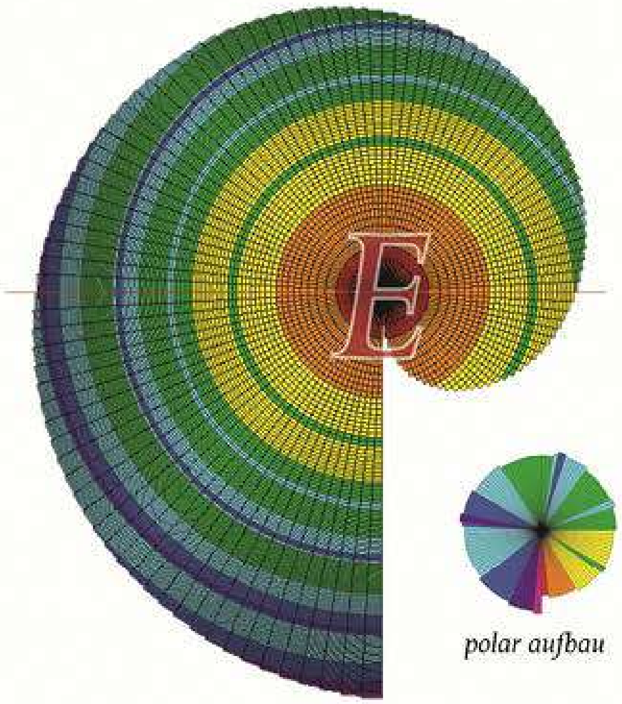
E
polar aufbau

polar energy spirals courtesy of Rene Cormier

Identifying electron rest Matter topologies as velocity invariant we can re-arrange the component Planck mass-energy geometry formulation of periodic elements to

$$h\left[\underset{\text{Deuteron rest mass}}{72[v^2]} + \underset{\text{Spectral lines}}{v} + \underset{\text{electron rest mass}}{1.20\text{ e}20\ v}\right]$$

reveal a quadratic formulation for all Z numbers

All elements are comprised of n level Duetrium nuclei

Baryons **KEM fields** **electrons**

$$Z\,[[72n^2] + [12v^2] + [1.2e20]]$$

1,861,949 MeV 13.525 eV 496,519 keV

$2H^+$

$[72n^2]$

Z

γ

$[M_e v^2]$

Elemental mass-Matter [in MeV]

The atomic shell energy levels of Deuterium nuclei in elements

Deuteron **electron eV** **KEM**

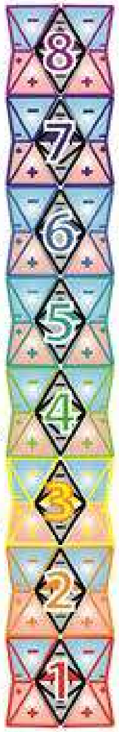

Deuteron		KEM
32 3,050.6 MeV	8	En = -13.52 eV
31 2,5862.9 MeV	7	En = -10.35 eV
30 2,6811 MeV	6	En = -7.60 eV
29 2,505.4 MeV	5	En = -5.28 eV
28 2,335.6 MeV	4	En = -3.38 eV
27 2,171.8 MeV	3	En = -1.90 eV
26 2,013.9 MeV	2	En = -0.84 eV
25 1,861.9 MeV	1	En = -0.21 eV

496.519 keV

n v²

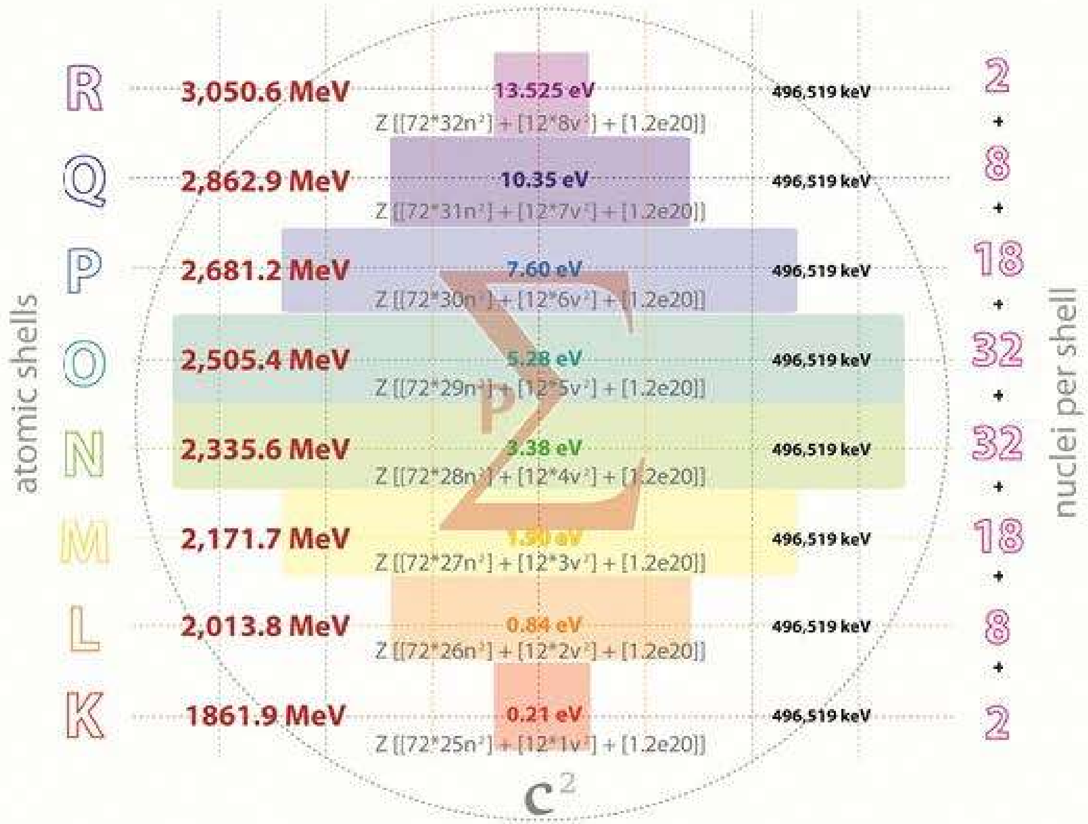

atomic shells				nuclei per shell
R	3,050.6 MeV	13.525 eV Z [[72*32n²] + [12*8v²] + [1.2e20]]	496,519 keV	2
Q	2,862.9 MeV	10.35 eV Z [[72*31n²] + [12*7v²] + [1.2e20]]	496,519 keV	+ 8
P	2,681.2 MeV	7.60 eV Z [[72*30n²] + [12*6v²] + [1.2e20]]	496,519 keV	+ 18
O	2,505.4 MeV	5.28 eV Z [[72*29n²] + [12*5v²] + [1.2e20]]	496,519 keV	+ 32
N	2,335.6 MeV	3.38 eV Z [[72*28n²] + [12*4v²] + [1.2e20]]	496,519 keV	+ 32
M	2,171.7 MeV	1.90 eV Z [[72*27n²] + [12*3v²] + [1.2e20]]	496,519 keV	+ 18
L	2,013.8 MeV	0.84 eV Z [[72*26n²] + [12*2v²] + [1.2e20]]	496,519 keV	+ 8
K	1861.9 MeV	0.21 eV Z [[72*25n²] + [12*1v²] + [1.2e20]]	496,519 keV	+ 2

6,101 MeV
993 keV
En = -27.05 eV
+
22,903 MeV
3,972 keV
En = -82.8 eV
+
48,262 MeV
8,937 keV
En = -316.8 eV
+
80,174 MeV
15,888 keV
En = -168.96 eV
+
74,740 MeV
15,888 keV
En = -108.16 eV
+
39,092 MeV
8,937 keV
En = -34.2 eV
+
16,111 MeV
3,972 keV
En = -6.72 eV
+
3,724 MeV
993 keV
En = -0.42 eV

[1.2e20]]

e

Determines the spectral line [KEM field energies] of electrons bound to them

The relativistic rest mass-energy-Matter of all periodic elements

is the sum of the mass-energies of all atomic nuclei and spectral lines that comprise its mass-Matter topology as measured in any spatial co-ordinate system per unit of time

the rest mass-Matter of bound photo-electrons is velocity invariant

Copyright ABRAHAM [2008] - All rights reserved

48
H
1.0000
22,512

Avagadro's number

6.022141579 e26
atoms in 1KG of Matter

1 KG mass
[of Matter]

1.660538841 e-27 kg
atomic rest mass-Matter

Hydrogen

01

using SI units Avagadro's number can be expressed exactly as the inverse rest mass of Hydrogen

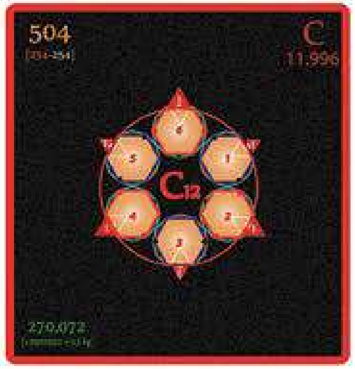

Weighted atomic mass

Carbon

12

5.019789213 e25
atoms in 1KG of Matter

6.02214078 e 23

1.99211552 e-26 kg
atomic rest mass-Matter

Tetryonic charge geometries make weighted atomic mass measurements and calculations obsolete

22,506 1/12 of Carbon 12 [Graphene] is not equal to 1 Hydrogen atom 22,512
(Deuterium is the building block of all atomic elements)

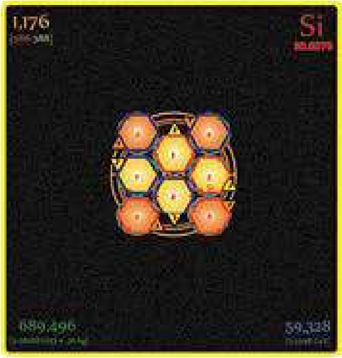

International Avagadro project

1.966225348 e25
atoms in 1KG of Matter

5.085887033 e-26 kg
atomic rest mass-Matter

Silicon

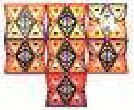

28

The gram was originally defined in 1795 as the mass of one cubic centimeter of water at 4°C,
making the kilogram equal to the mass of one liter of water.

The prototype kilogram, manufactured in 1799 and from which the current kilogram is based
has a mass equal to the mass of 1.000025 liters of water

In recent years two major experiments, namely the Watt balance & Avagadro projects, have been attempting to
measure and define 1KG of mass-Matter in terms of electrical force and the number of atoms respectively
in order to better define 1KG of mass-Matter precisely for all future physical references

Platinum

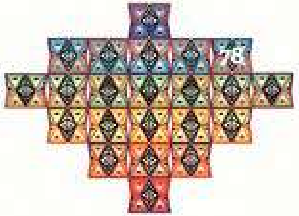

191

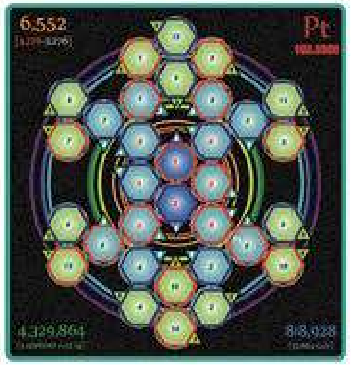

La Grande K

2.817950081 e24
atoms in 1KG of Matter
3.181804449 e23

90% Pt
10% Ir

*The La grande K is an alloy of 90% Platinum & 10% Iridium
that has been slowly losing mass since its manufacture*

3.1893811012 e-25 kg
atomic rest mass-Matter
3.142870708 e-25 kg

***All atomic rest masses are for atoms at absolute zero and any deviation is a measure of
the topological Matter's Kinetic energy content [chemical energy, KEM fields and/or spectral lines]***

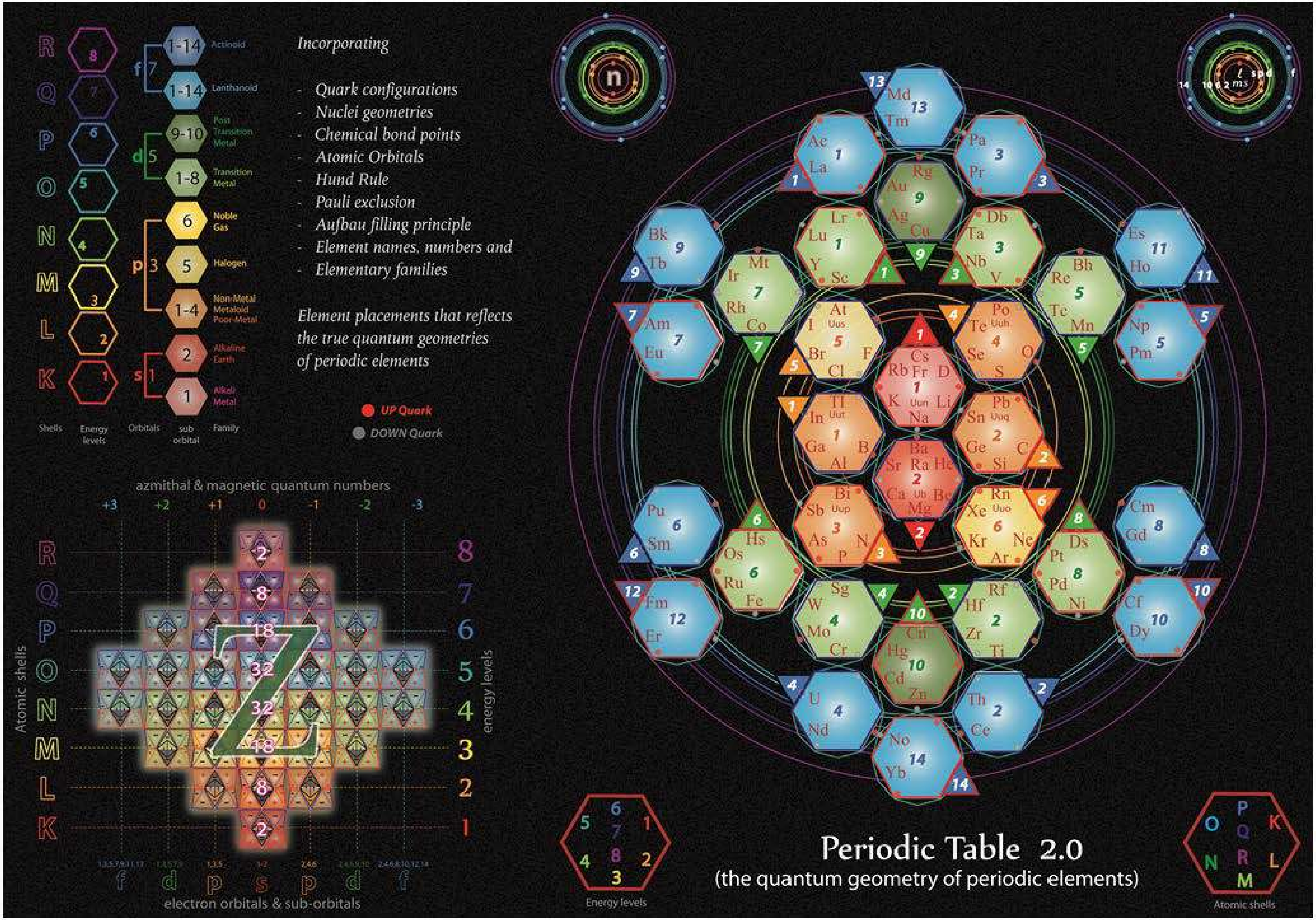
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shells
Energy levels
Orbitals
sub orbital
Family
UP Quark
DOWN Quark
azmithal & magnetic quantum numbers
+3
+2
+1
0
-1
-2
-3
Atomic shells
energy levels
Z
2
8
18
32
32
18
8
2
f
d
p
s
p
d
f
electron orbitals & sub-orbitals
n
Energy levels
Atomic shells

Incorporating

- Quark configurations
- Nuclei geometries
- Chemical bond points
- Atomic Orbitals
- Hund Rule
- Pauli exclusion
- Aufbau filling principle
- Element names, numbers and
- Elementary families

Element placements that reflects the true quantum geometries of periodic elements

Periodic Table 2.0

(the quantum geometry of periodic elements)

Copyright ABRAHAM [2008] - All rights reserved

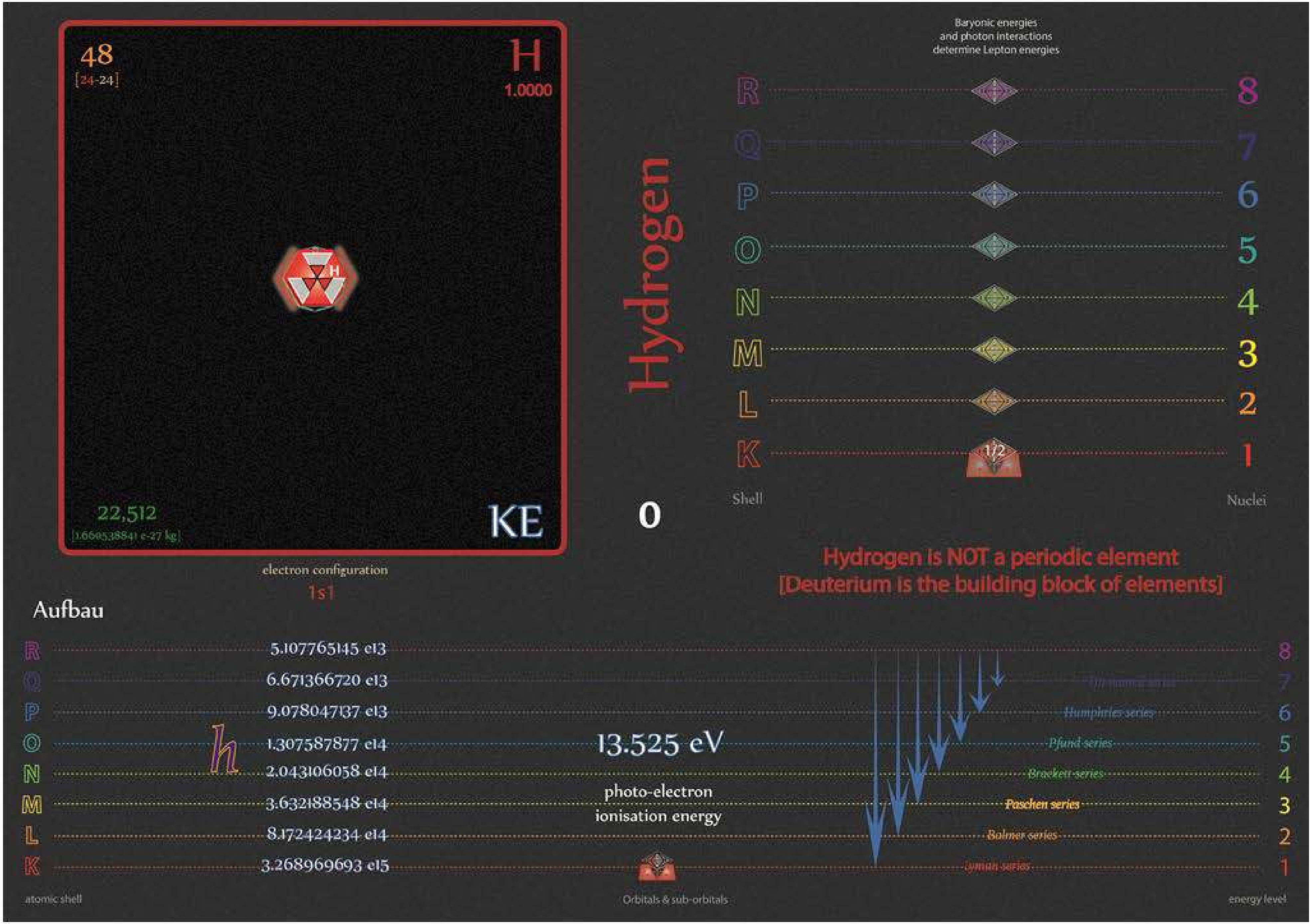

48
[24-24]

H
1.0000

22,512
[1.660538841 e-27 kg]

KE

Hydrogen

0

Baryonic energies
and photon interactions
determine Lepton energies

Shell	Nuclei
R	8
Q	7
P	6
O	5
N	4
M	3
L	2
K	1

1/2

Hydrogen is NOT a periodic element
[Deuterium is the building block of elements]

electron configuration
1s1

Aufbau

atomic shell			energy level
R	5.107765145 e13		8
Q	6.671366720 e13		7
P	9.078047137 e13	Humphries series	6
O	1.307587877 e14	Pfund series	5
N	2.043106058 e14	Brackett series	4
M	3.632188548 e14	Paschen series	3
L	8.172424234 e14	Balmer series	2
K	3.268969693 e15	Lyman series	1

h

13.525 eV

photo-electron
ionisation energy

Orbitals & sub-orbitals

Tetryonics 51.00 - Hydrogen atom

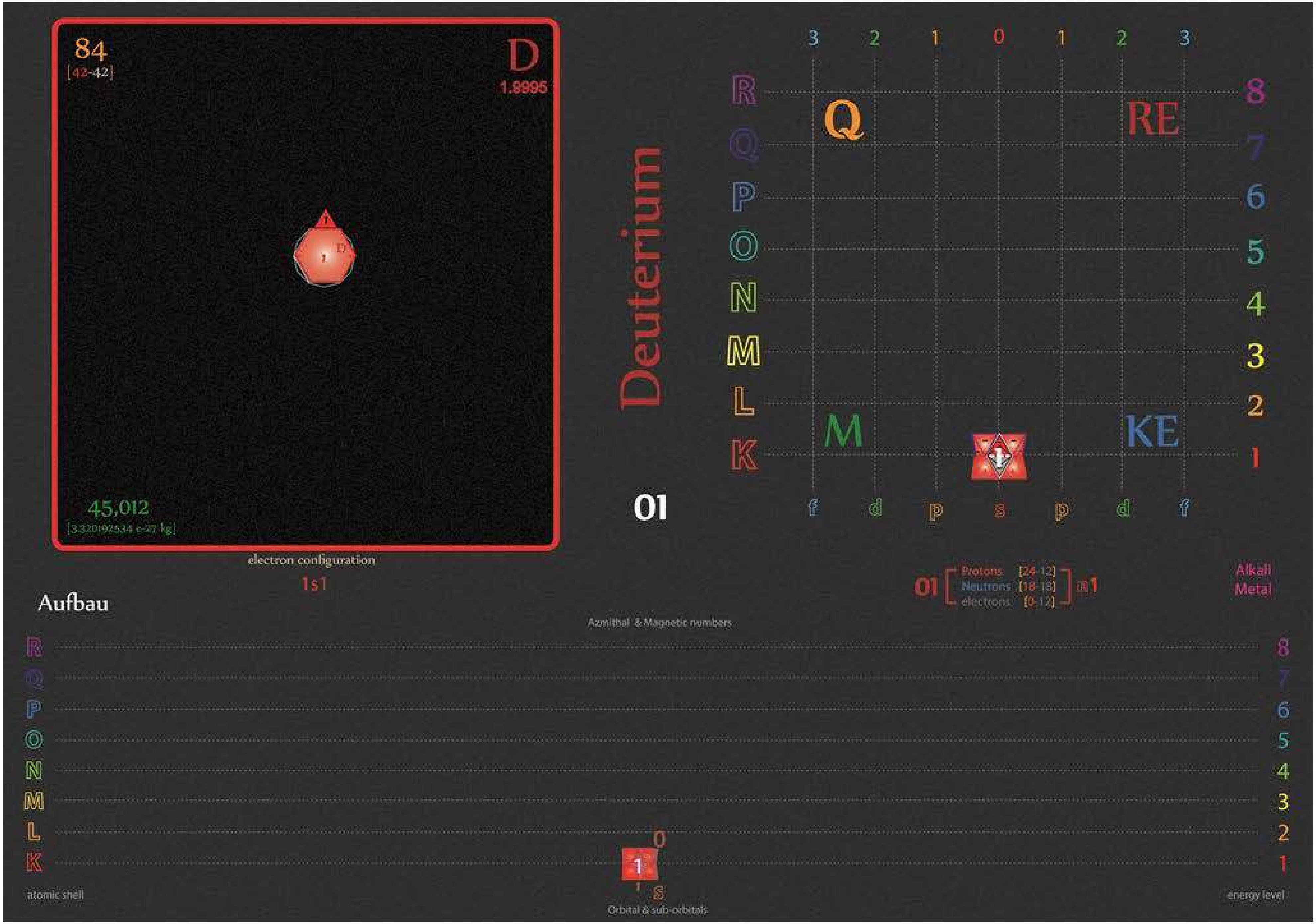
84
[42-42]
D
1.9995
D
45,012
[3.320192534 e-27 kg]
electron configuration
1s1
Deuterium
01
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
1
f
d
p
s
p
d
f
01
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1
Alkali Metal
Azmithal & Magnetic numbers
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
s
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.01 - Deuterium atom

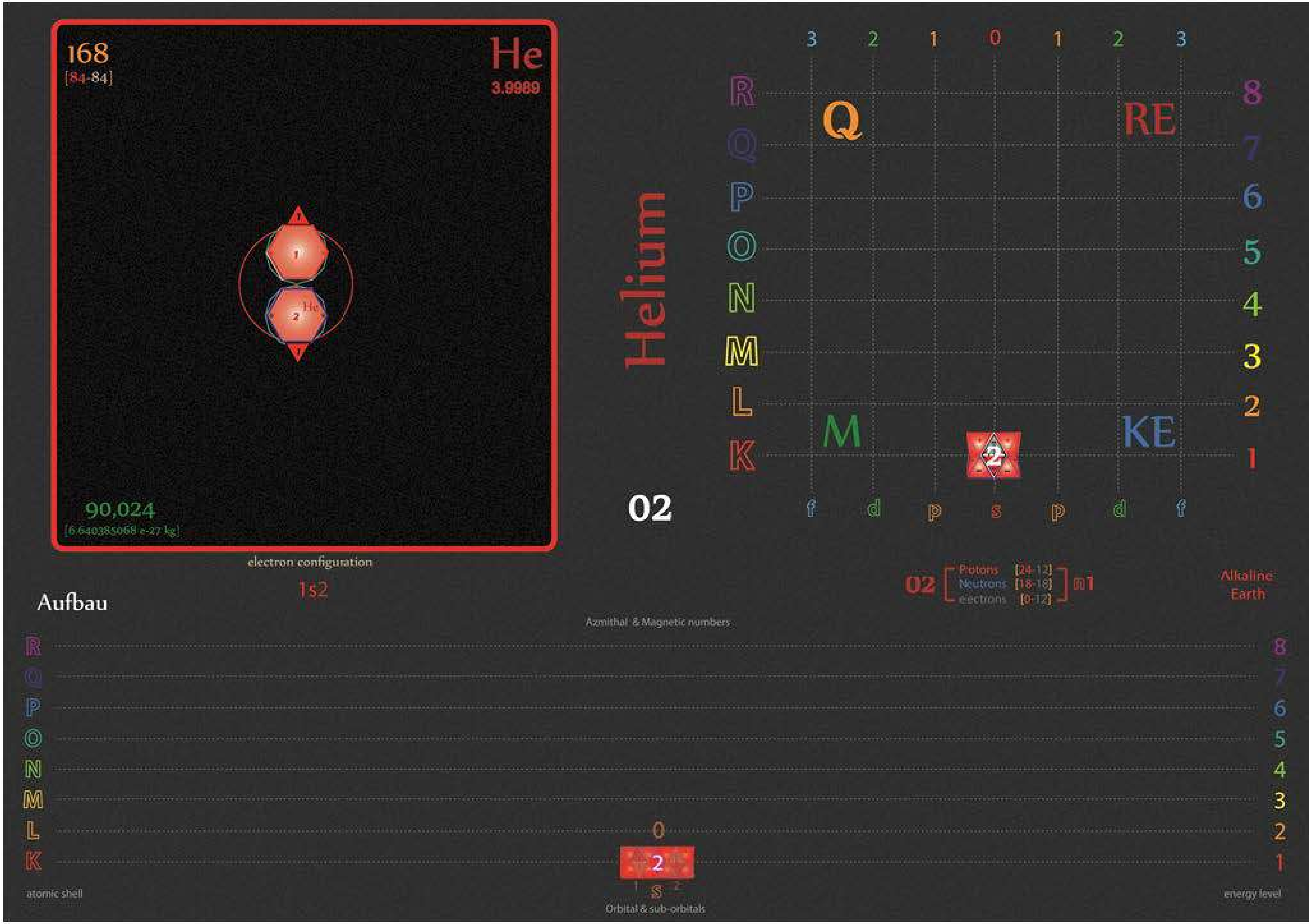
168
[84-84]
He
3.9989
1
1
2
He
1
90,024
[6.640385068 e-27 kg]
electron configuration
1s2
Aufbau
Helium
02
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
2
f d p s p d f
02
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1
Alkaline Earth
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
1 s 2
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.02 - Helium atom

Copyright ABRAHAM [2008] - All rights reserved

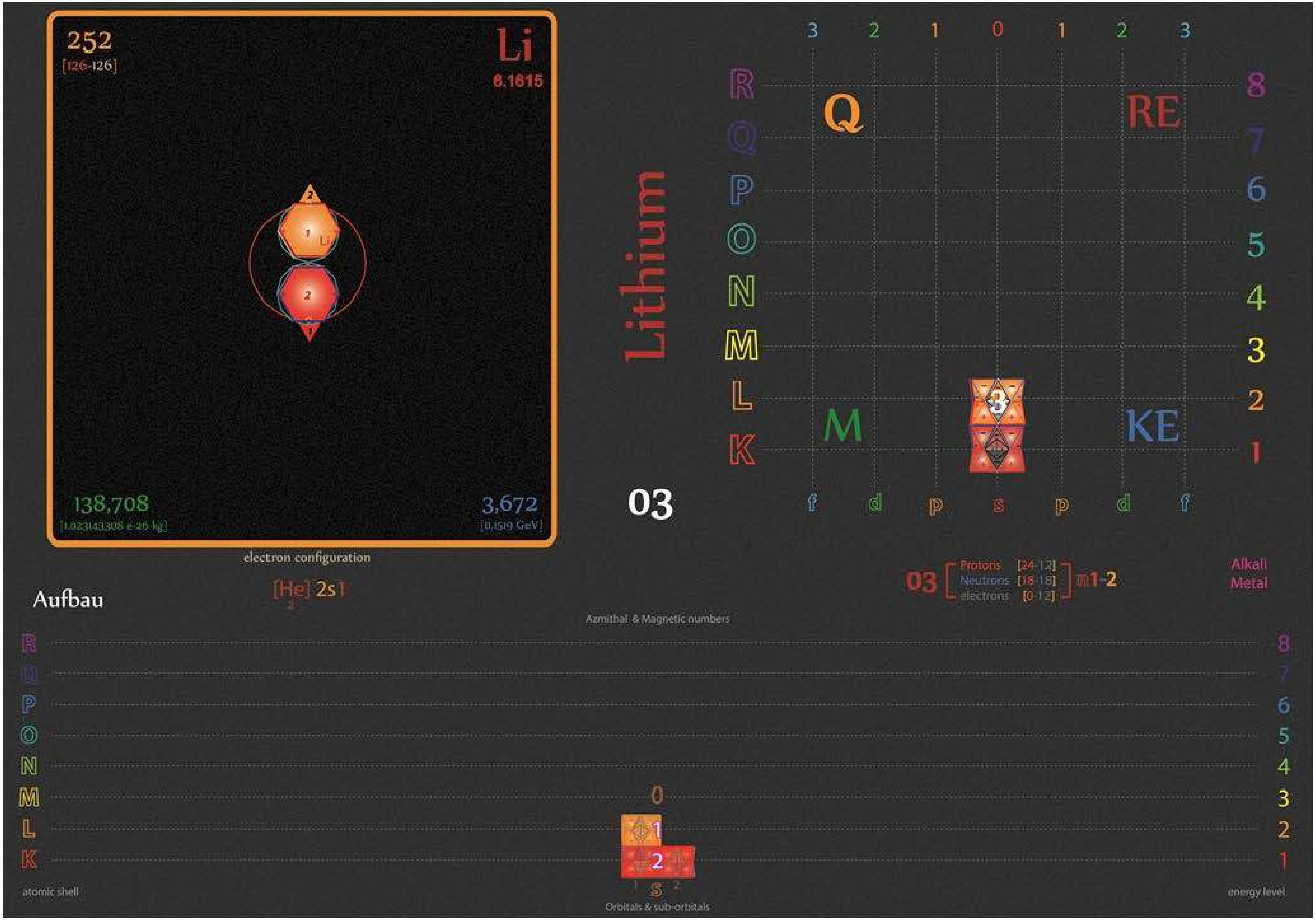
252
[126-126]
Li
6.1615
2
1
Li
2
1
138,708
[1.023143308 e-26 kg]
3,672
[0.1519 GeV]
electron configuration
[He] 2s1
2
Aufbau
Lithium
03
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
3
f
d
p
s
p
d
f
03
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Alkali
Metal
Azimthal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
2
1
s
2
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.03 - Lithium atom

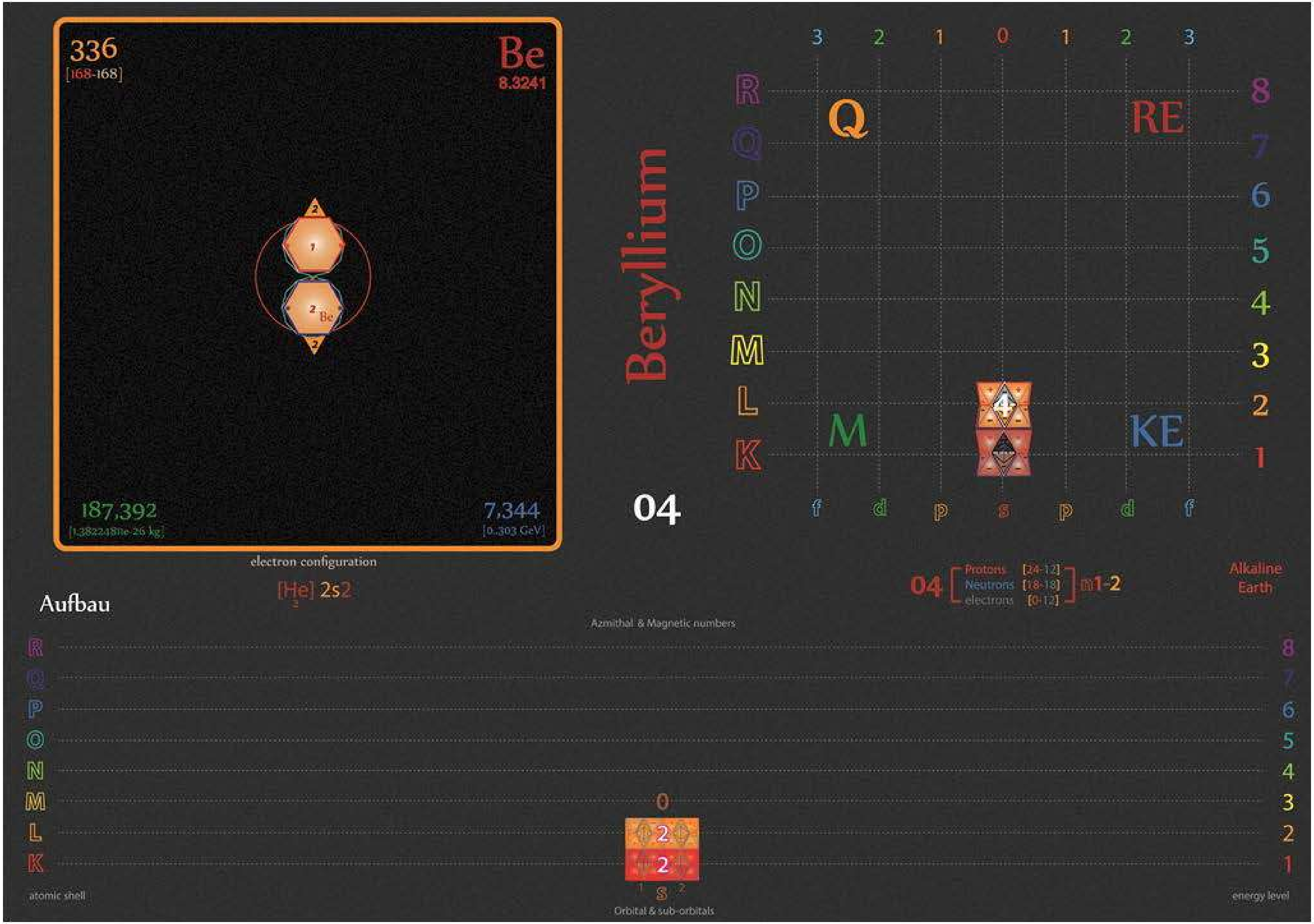
336
[168-168]
Be
8.3241
2
1
2
Be
2
187,392
[1.38224818e-26 kg]
7,344
[0.303 GeV]
electron configuration
[He] 2s2
Beryllium
04
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
4
f d p s p d f
04
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Alkaline Earth
Azmithal & Magnetic numbers
0
2
2
1 s 2
Orbital & sub-orbitals
atomic shell
energy level

Tetryonics 51.04 - Beryllium atom

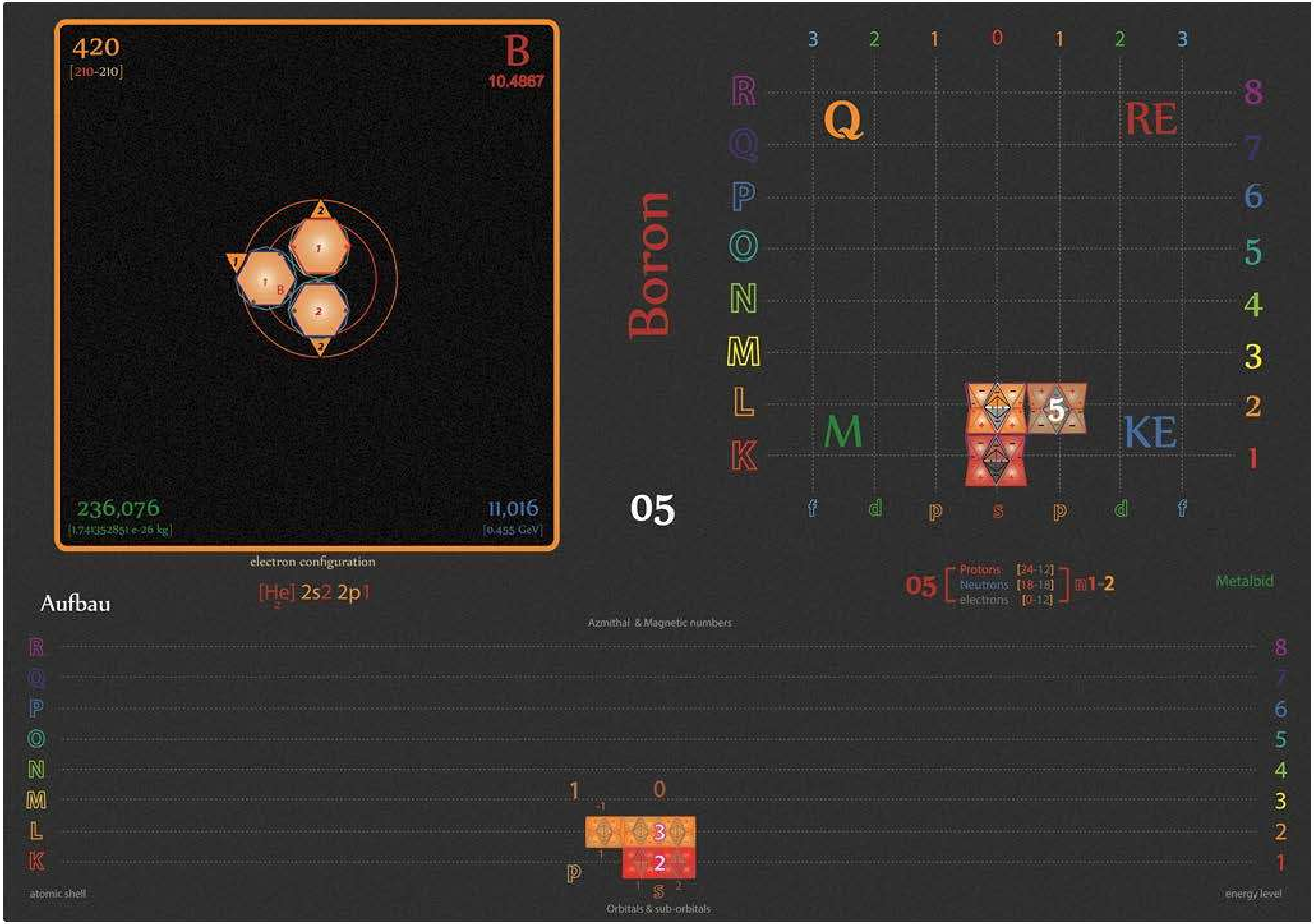
420
[210-210]
B
10.4867
236,076
[1.741352851 e-26 kg]
11,016
[0.455 GeV]
electron configuration
[He] 2s2 2p1
Boron
05
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
5
f d p s p d f
05 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-2
Metaloid
Azmithal & Magnetic numbers
1 0
-1
3
2
p
s
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.05 - Boron atom

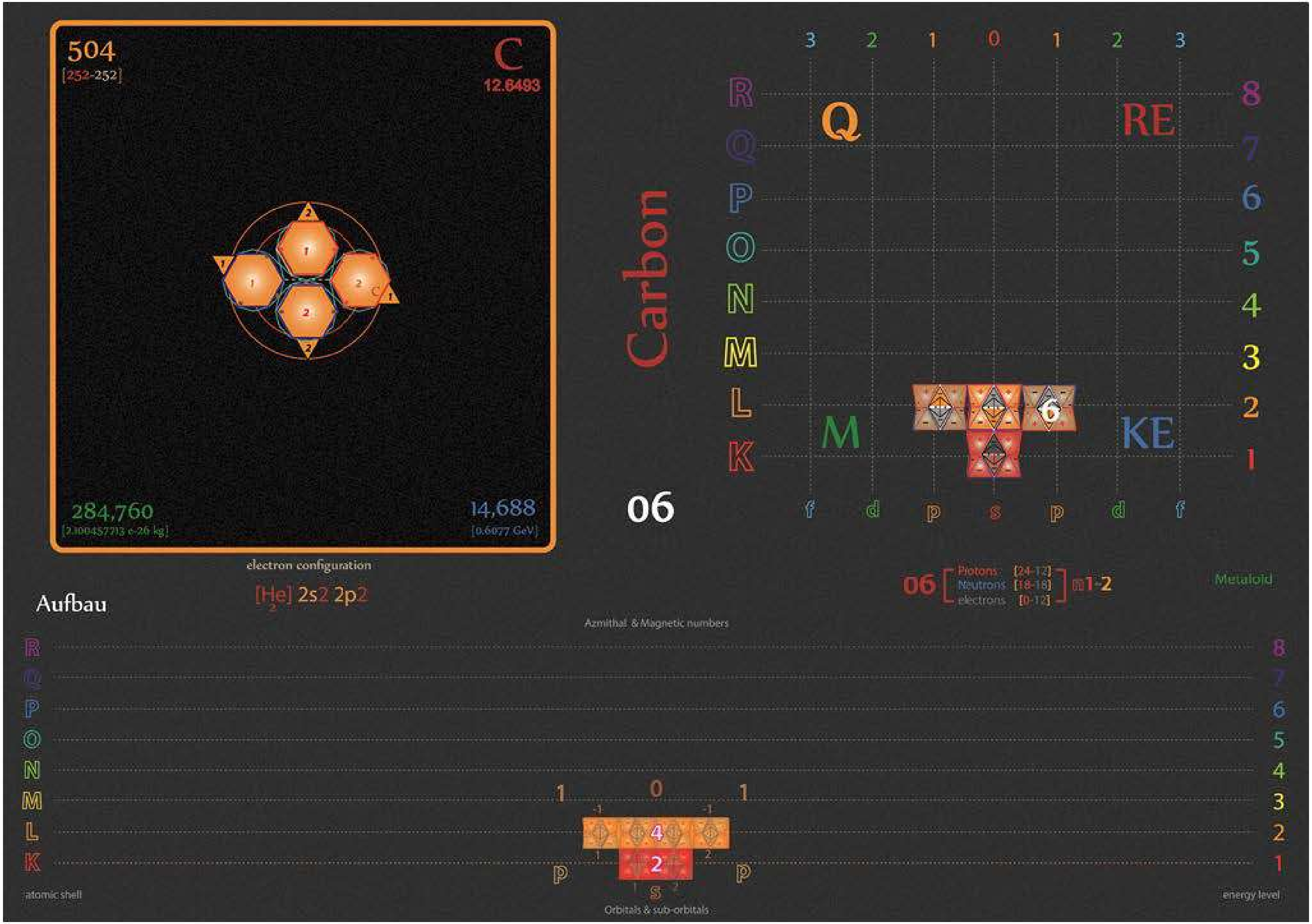
504
[252-252]
C
12.6493
1
2
2
1
1
2
1
2
284,760
[2.100457713 e-26 kg]
14,688
[0.6077 GeV]
Carbon
06
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
6
f
d
p
s
p
d
f
electron configuration
[He] 2s2 2p2
2
Aufbau
06
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metalloid
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
-1
-1
4
1
2
2
p
p
1
s
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.06 - Carbon atom

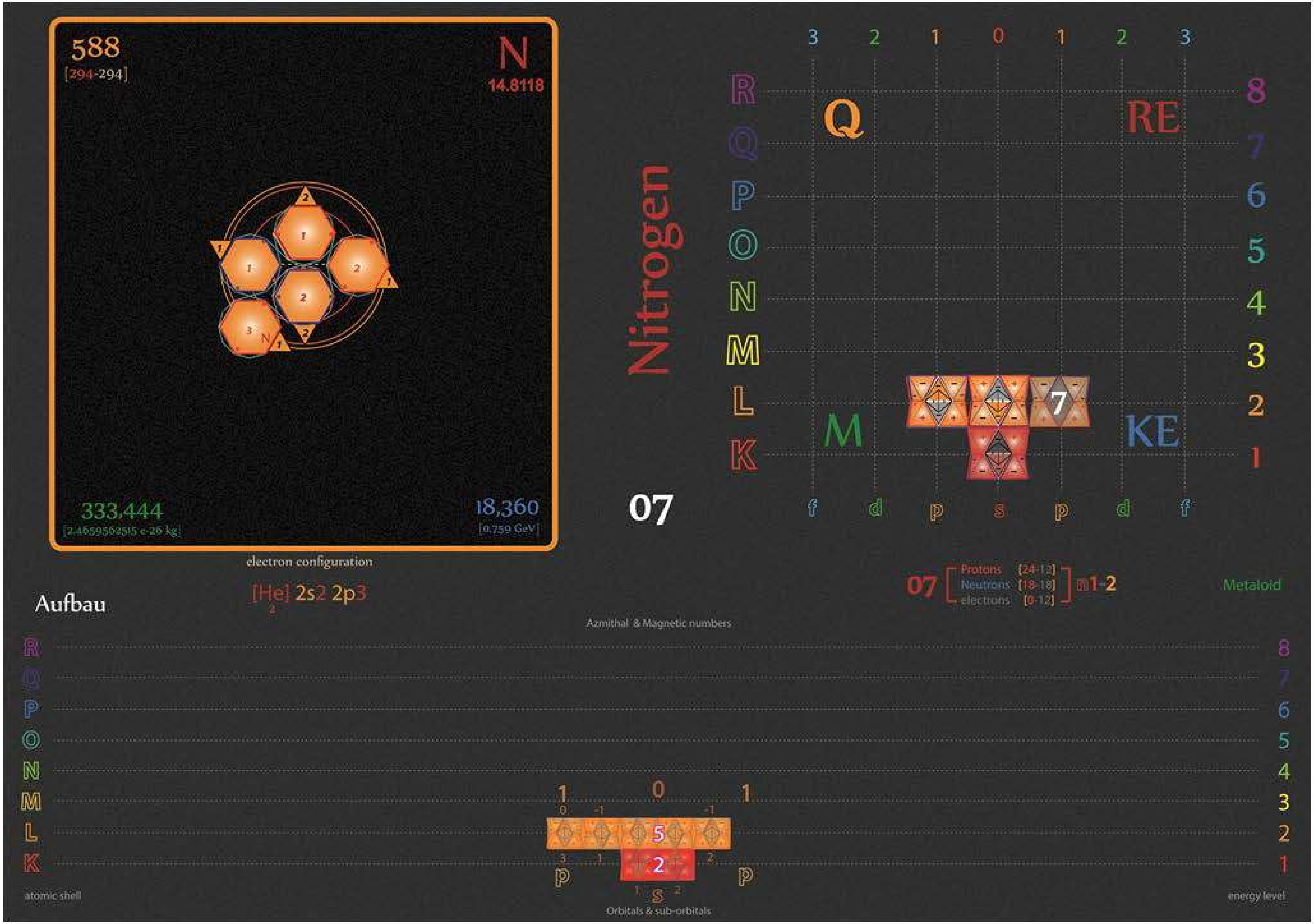
588
[294-294]
N
14.8118
333,444
[2.4659562515 e-26 kg]
18,360
[0.759 GeV]
electron configuration
[He] 2s2 2p3
2
Aufbau
Nitrogen
07
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
7
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
07
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
1
0
1
0
-1
-1
5
2
3
1
2
p
s
p
1
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.07 - Nitrogen atom

Copyright ABRAHAM [2008] - All rights reserved

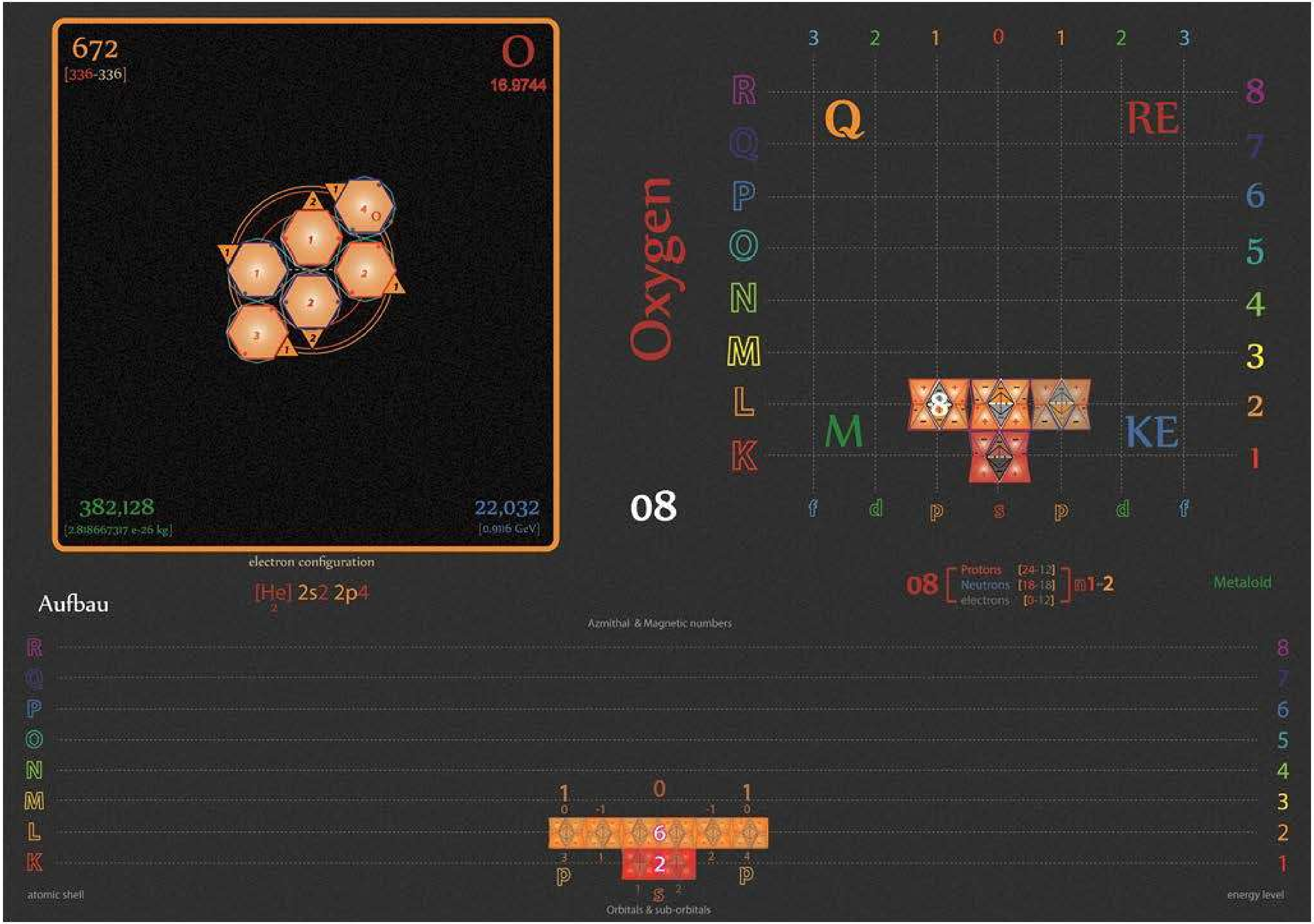
672
[336-336]
O
16.9744
382,128
[2.818667317 e-26 kg]
22,032
[0.9116 GeV]
electron configuration
[He] 2s2 2p4
2
Aufbau
Oxygen
08
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
08 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Metaloid
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
6
2
3 1 2 4
p s p
1 2
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.08 - Oxygen atom

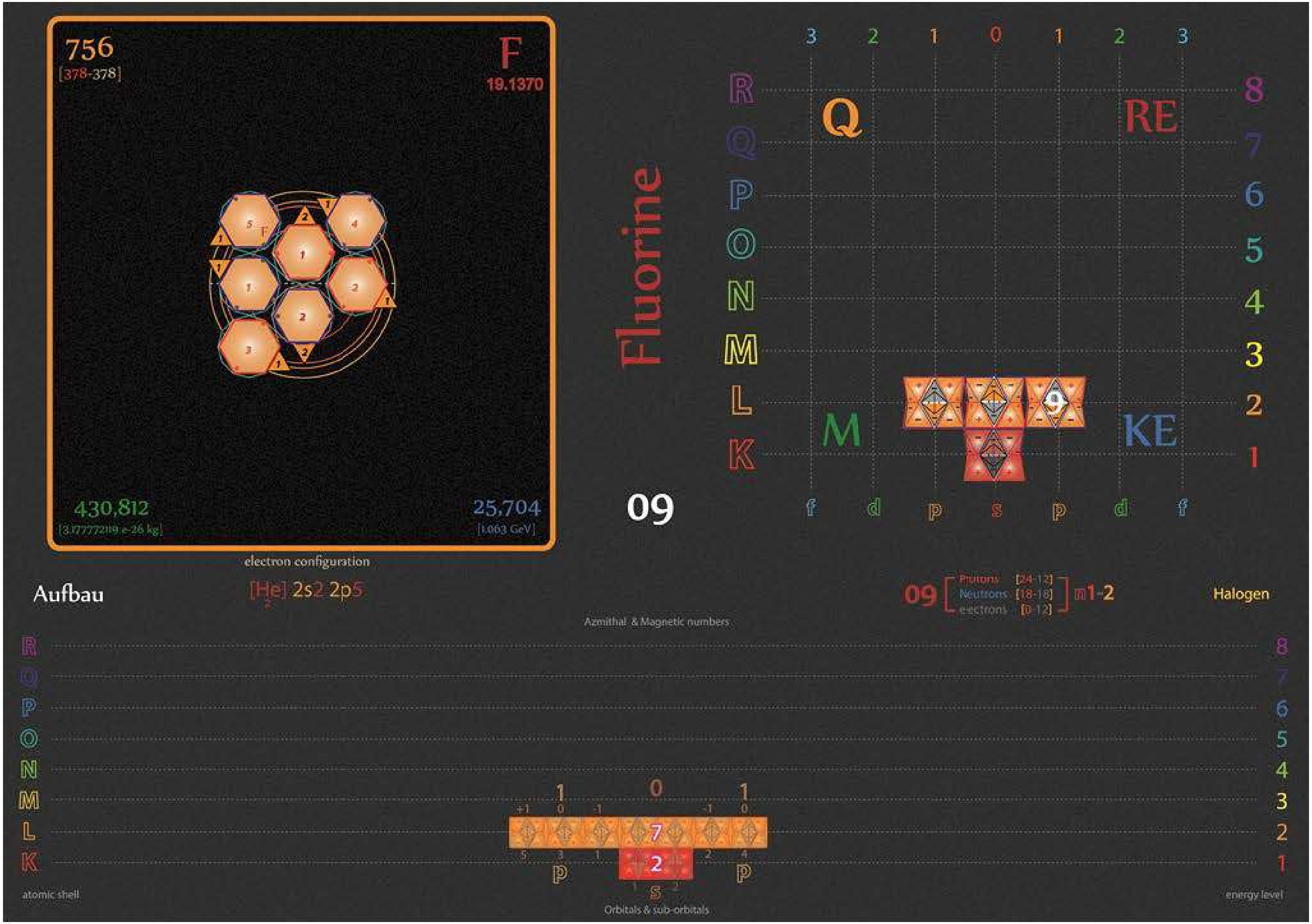
756
[378-378]
F
19.1370
430,812
[3.17777219 e-26 kg]
25,704
[1.063 GeV]
Fluorine
09
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
9
f d p s p d f
electron configuration
Aufbau
[He] 2s2 2p5
2
09
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Halogen
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0
7
2
5 3 1 2 4
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.09 - Fluorine atom

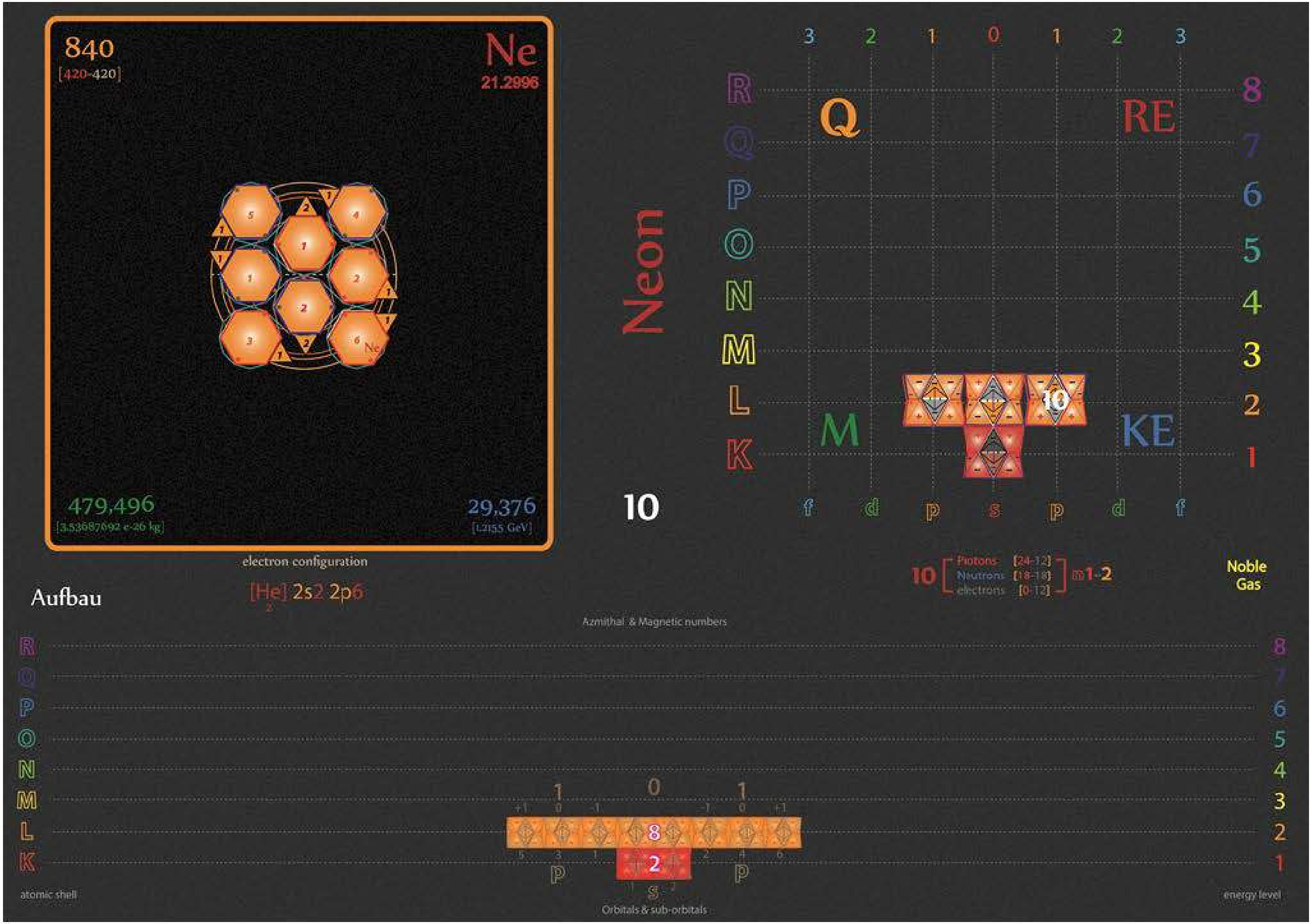
840
[420-420]
Ne
21.2996
479,496
[3.53687692 e-26 kg]
29,376
[1.2155 GeV]
Neon
10
electron configuration
[He] 2s2 2p6
2
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
10 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Noble Gas
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.10 - Neon atom

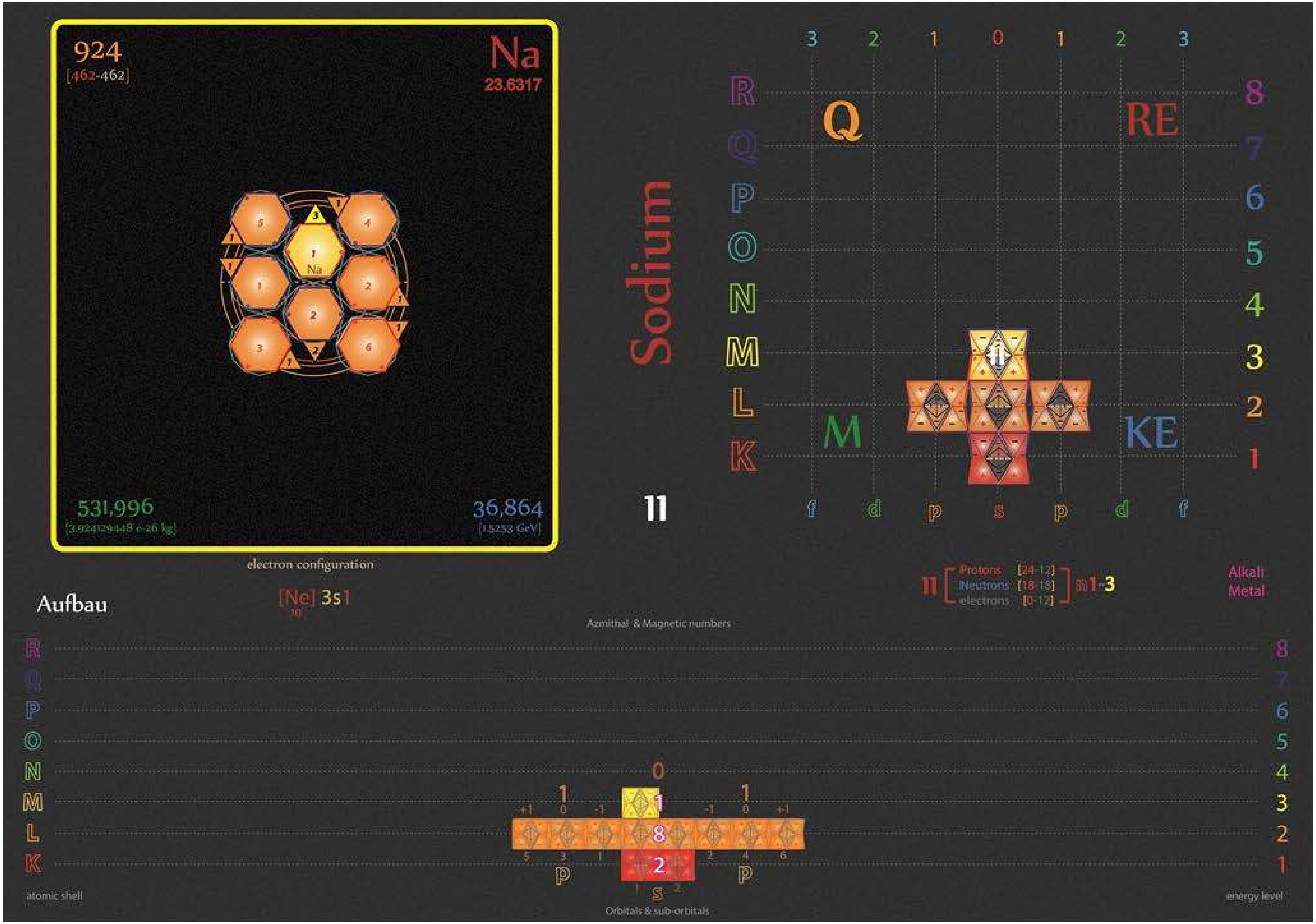
924
[462-462]
Na
23.6317
531,996
[3.924129448 e-26 kg]
36,864
[15253 GeV]
Sodium
11
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
11
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-3
Alkali
Metal
electron configuration
Aufbau
[Ne] 3s1
10
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1 1
+1 0 -1 -1 0 +1
1
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.11 - Sodium atom

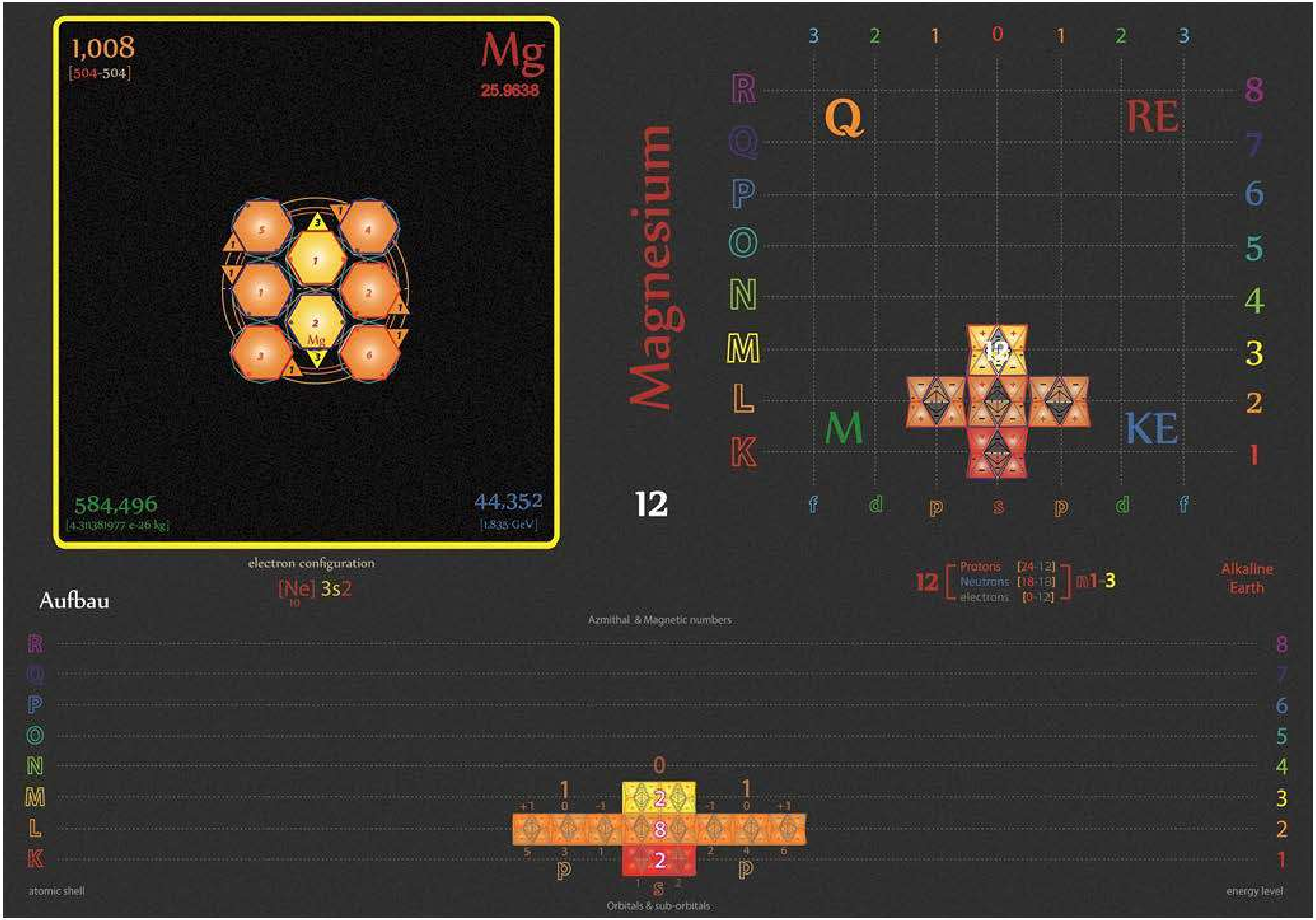
1,008
[504-504]
Mg
25.9838
Mg
584,496
[4.31381977 e-26 kg]
44,352
[1.835 GeV]
electron configuration
[Ne] 3s2
10
Aufbau
Magnesium
12
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
12
12 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Alkaline Earth
Azmithal & Magnetic numbers
0
1
1
+1 0 -1
-1 0 +1
2
8
2
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.12 - Magnesium atom

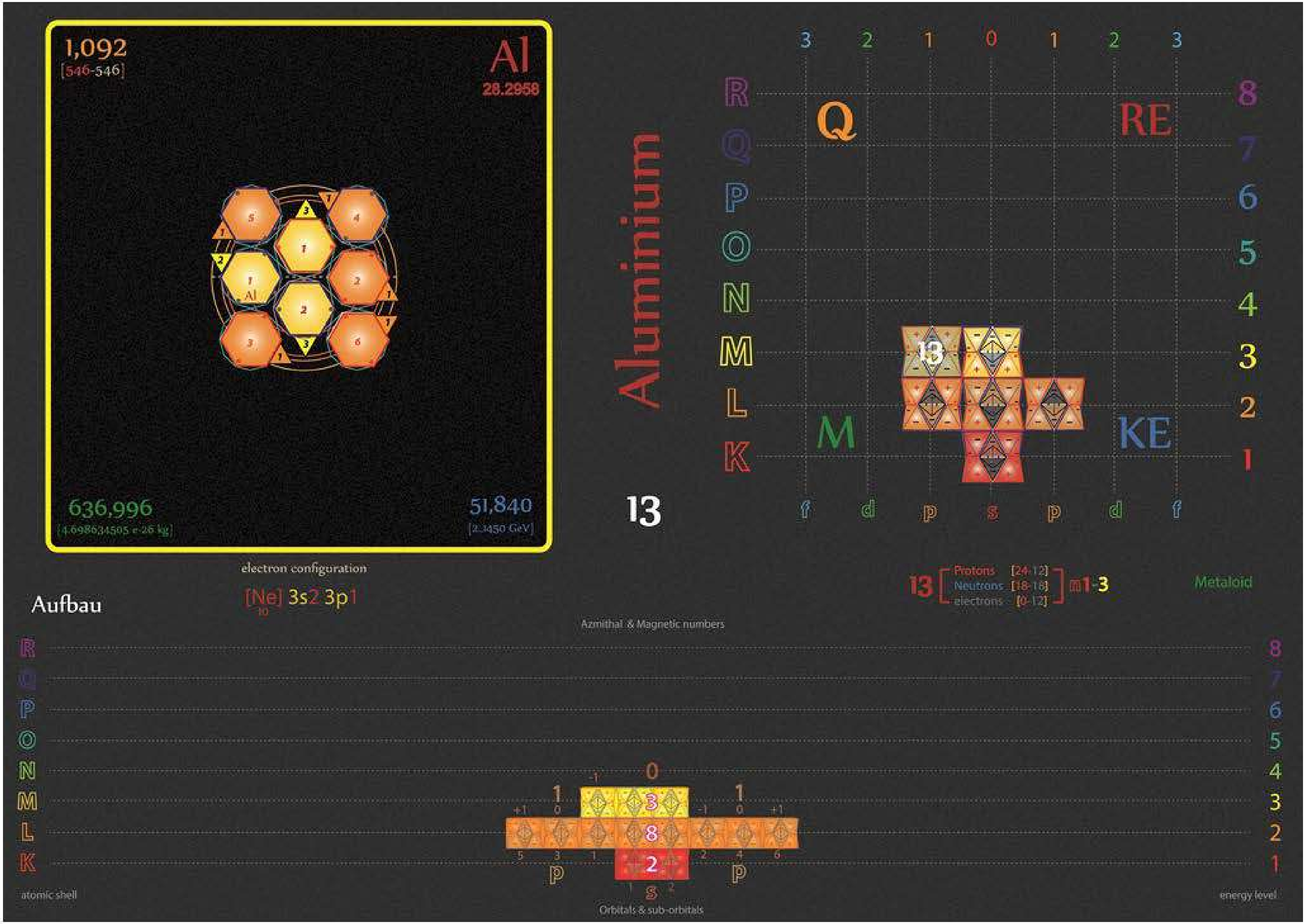
1,092
[546-546]
Al
28.2958
1
Al
2
3
4
5
6
636,996
[4.698634505 e-26 kg]
51,840
[2.3450 GeV]
Aluminium
13
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
13
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
[Ne] 3s2 3p1
10
Aufbau
13
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Metaloid
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
-1
+1
0
-1
0
+1
3
8
2
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.13 - Aluminium atom

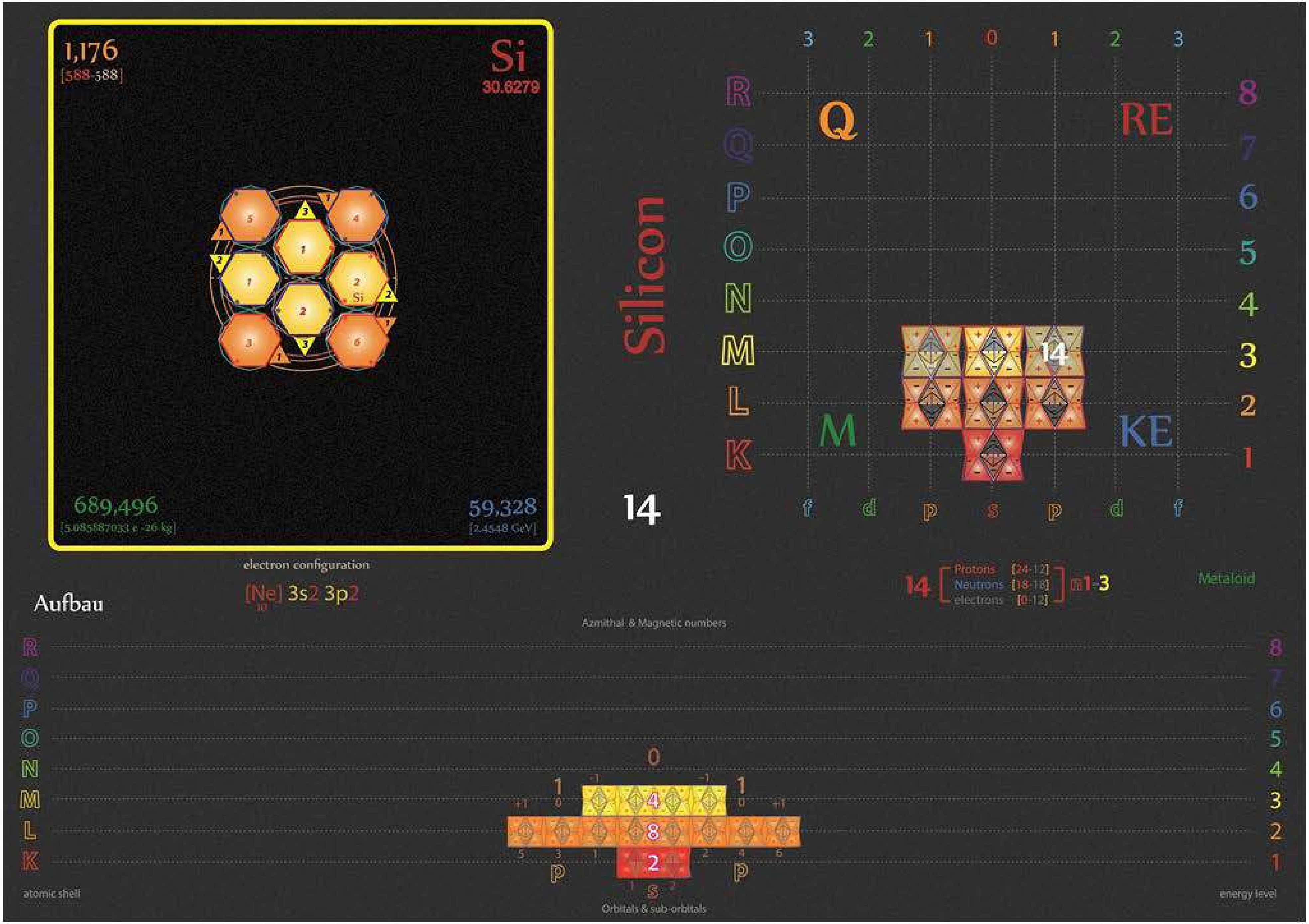
1,176
[588-588]
Si
30.6279
5
4
1
1
2
Si
2
3
6
3
689,496
[5.085887033 e -26 kg]
59,328
[2.4548 GeV]
Silicon
14
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
14
f d p s p d f
electron configuration
[Ne] 3s2 3p2
10
Aufbau
14
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
1
-1
-1
+1
0
0
+1
4
8
2
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.14 - Silicon atom

Copyright ABRAHAM [2008] - All rights reserved

1,260
[630-630]

P
32.9600

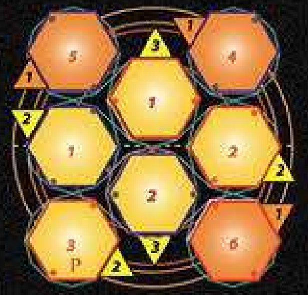
5
4
3
1
1
2
1
2
2
2
1
3
P
2
3
6

741,996
[5.473139562 e-26 kg]

66,816
[2.7646 GeV]

Phosphorus

15

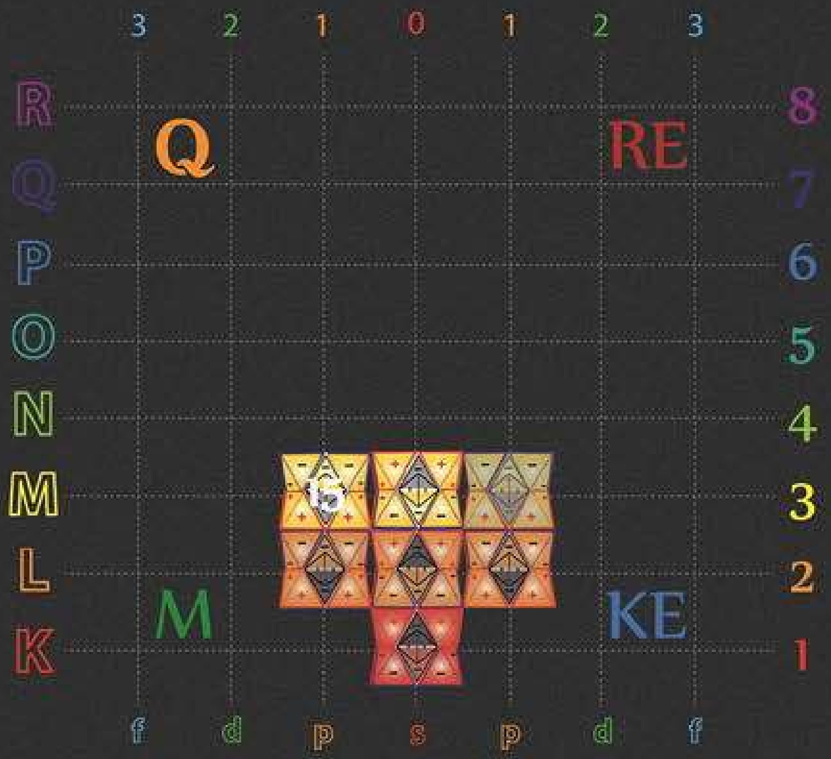
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
15
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

[Ne] 3s2 3p3
10

Aufbau

15 [Protons [24-12] / Neutrons [18-18] / electrons [0-12]] m1-3

Metaloid

Azmithal & Magnetic numbers

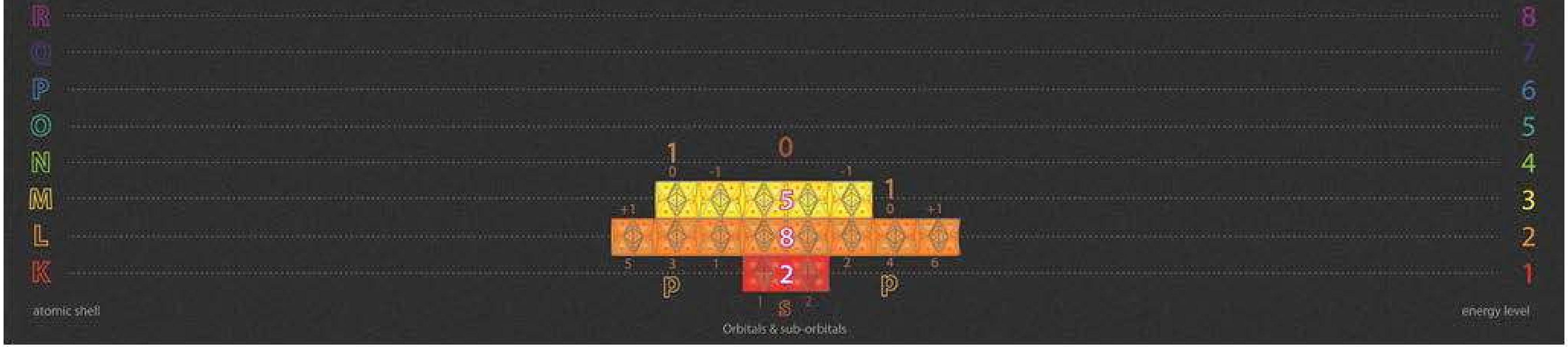
R
Q
P
O
N
M
L
K
1
0
0
-1
-1
5
1
0
+1
+1
8
5
3
1
2
4
6
2
p
p
1
s
2
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbitals & sub-orbitals

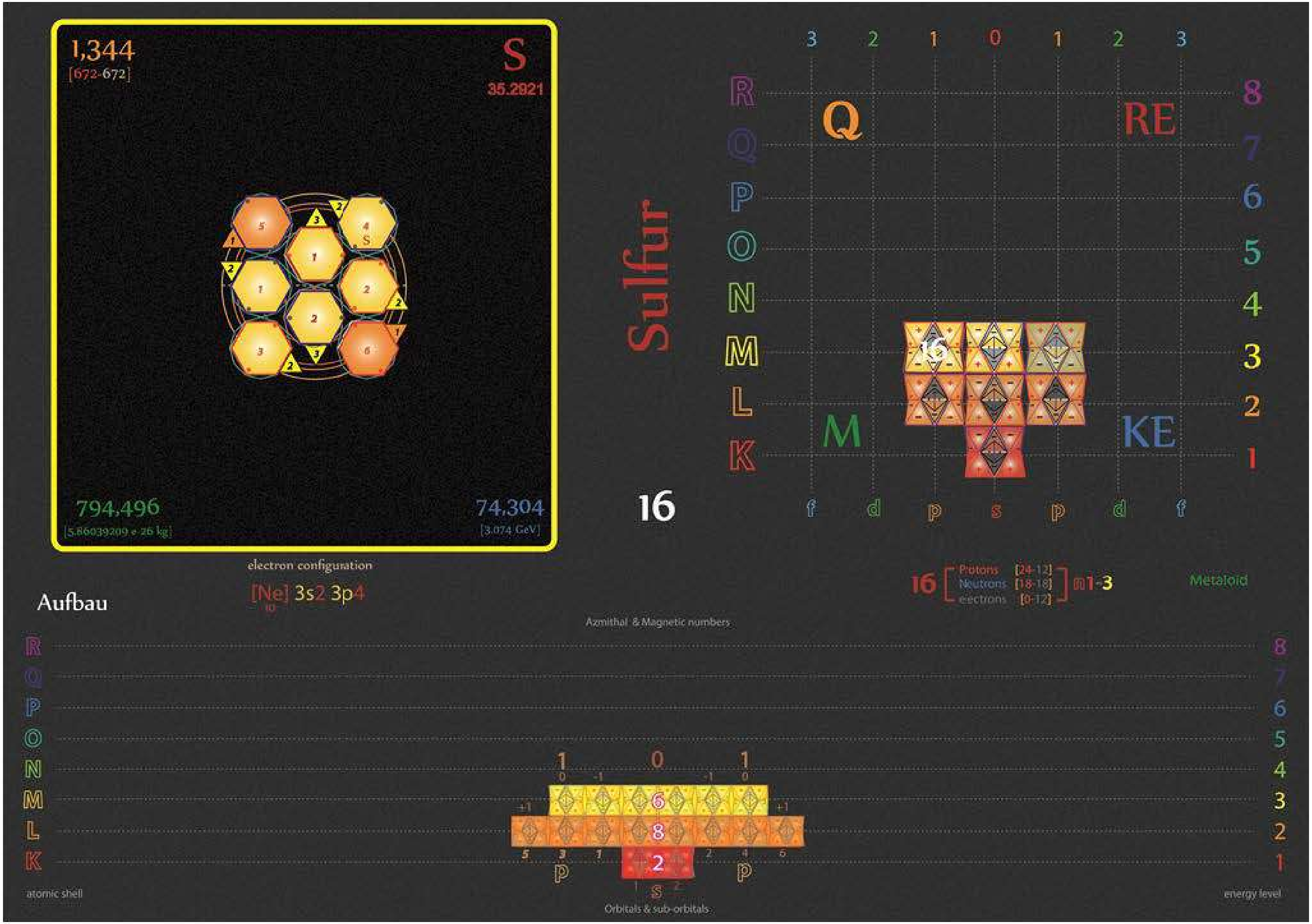
1,344
[672-672]
S
35.2921
794,496
[5.86039209 e-26 kg]
74,304
[3.074 GeV]
electron configuration
[Ne] 3s2 3p4
10
Aufbau
Sulfur
16
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
16 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Metaloid
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
+1 +1
6
8
2
5 3 1 2 4 6
p s p
1 2
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.16 - Sulfur atom

Copyright ABRAHAM [2008] - All rights reserved

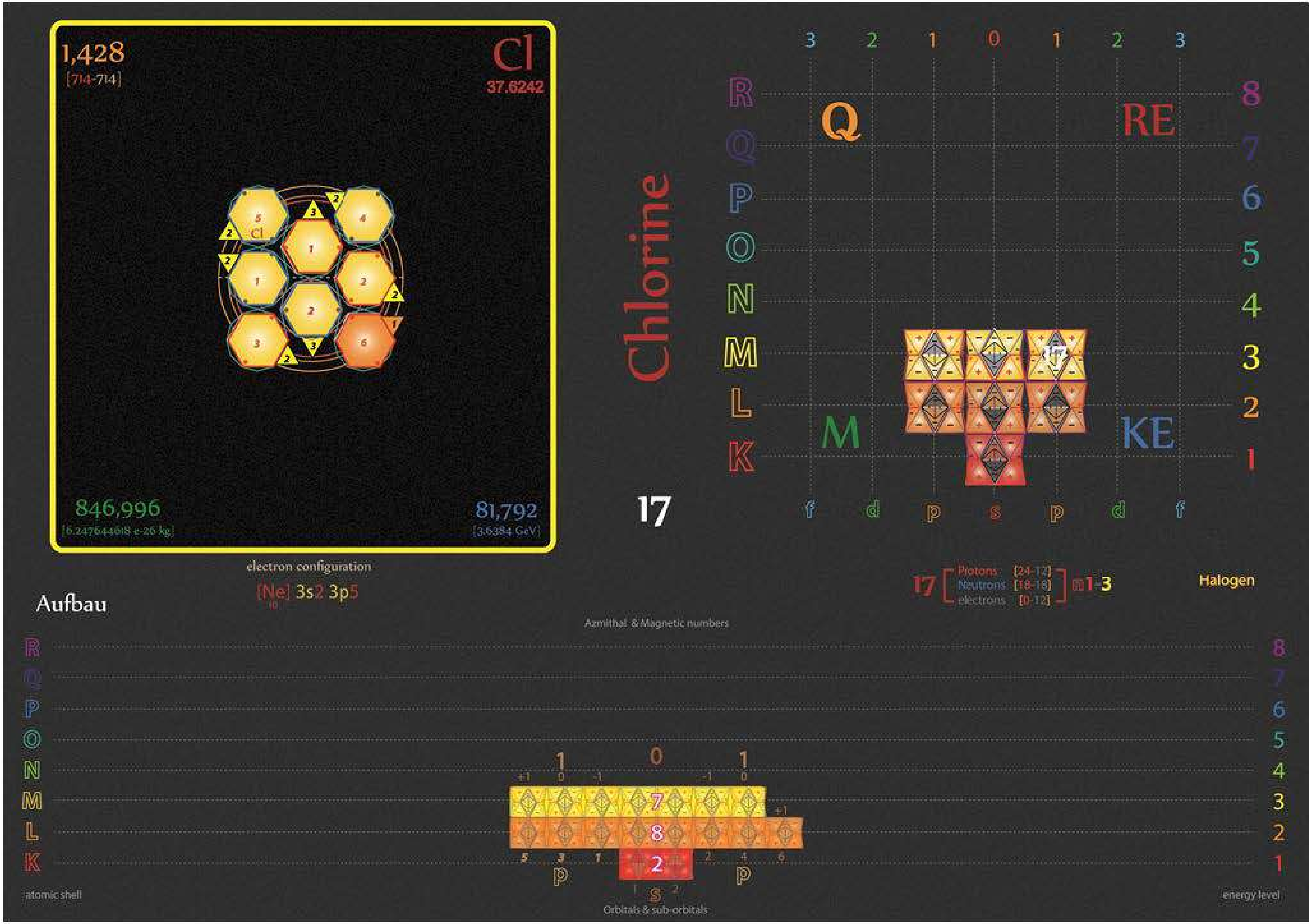
1,428
[714-714]
Cl
37.6242
846,996
[6.247644618 e-26 kg]
81,792
[3.6384 GeV]
Chlorine
17
electron configuration
[Ne] 3s2 3p5
10
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
17
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Halogen
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
7
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.17 - Chlorine atom

Copyright ABRAHAM [2008] - All rights reserved

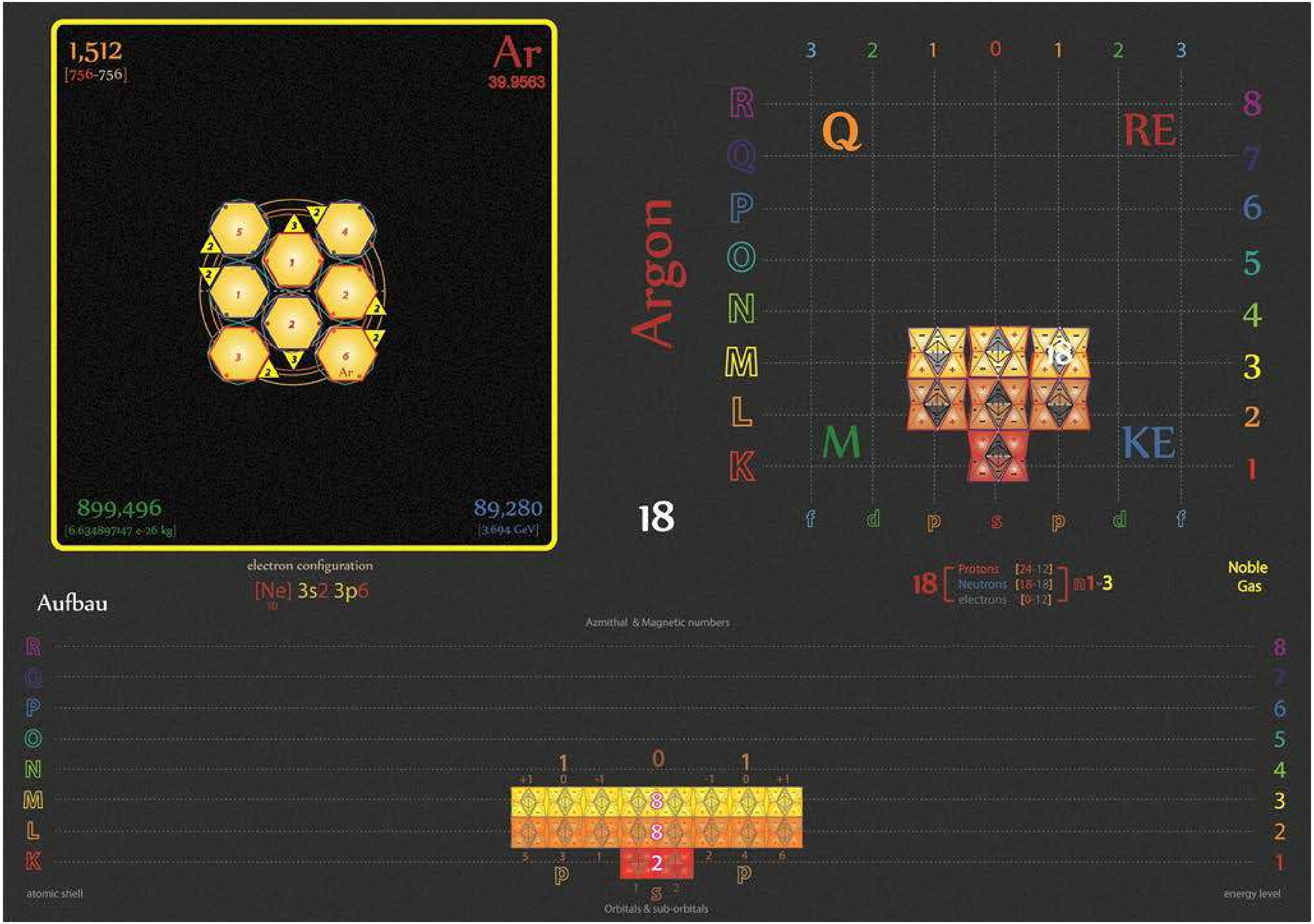
1,512
[756-756]
Ar
39.9563
5
4
3
2
1
1
2
2
3
6
Ar
899,496
[6.634897147 e-26 kg]
89,280
[3.694 GeV]
electron configuration
[Ne] 3s2 3p6
10
Aufbau
Argon
18
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
18
f d p s p d f
18
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-3
Noble Gas
Azimuthal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0 +1
8
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.18 - Argon atom

Copyright ABRAHAM [2008] - All rights reserved

1,596
[798-798]

K
42.4643

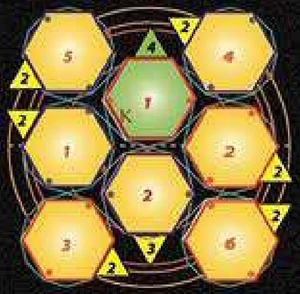
5
4
1
4
2
2
2
1
2
2
2
3
6
3
2

955,956
[7.05135958 e-26 kg]

100,728
[4.167 GeV]

Potassium

19

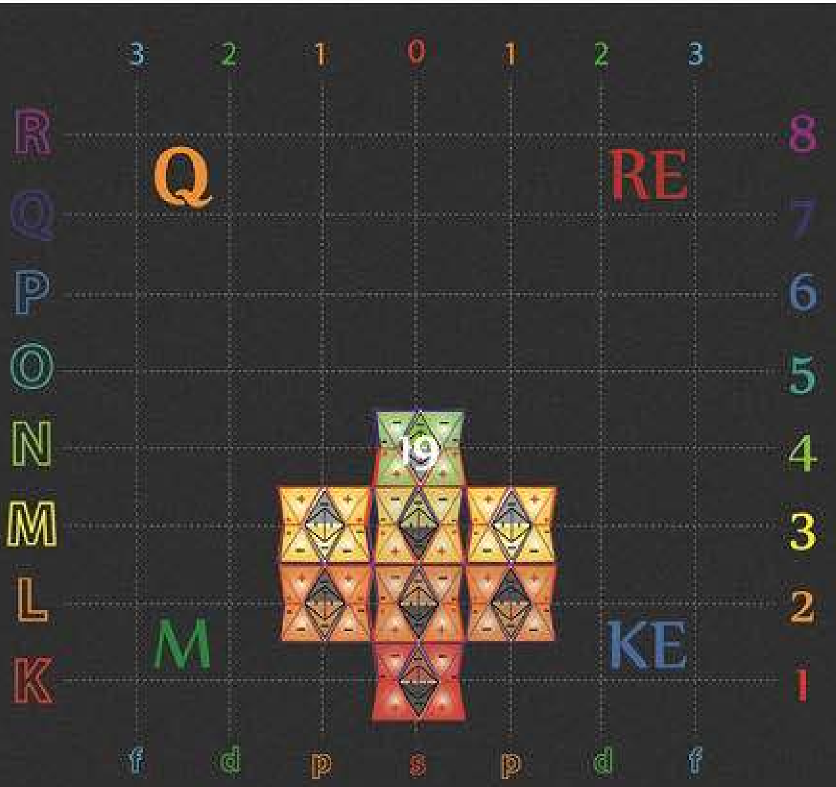
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
19
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

19 [Protons [24-12] / Neutrons [18-18] / electrons [0-12]] n1-4

Alkali
Metal

electron configuration

[Ar] 4s 1
18

Aufbau

Azimithal & Magnetic numbers

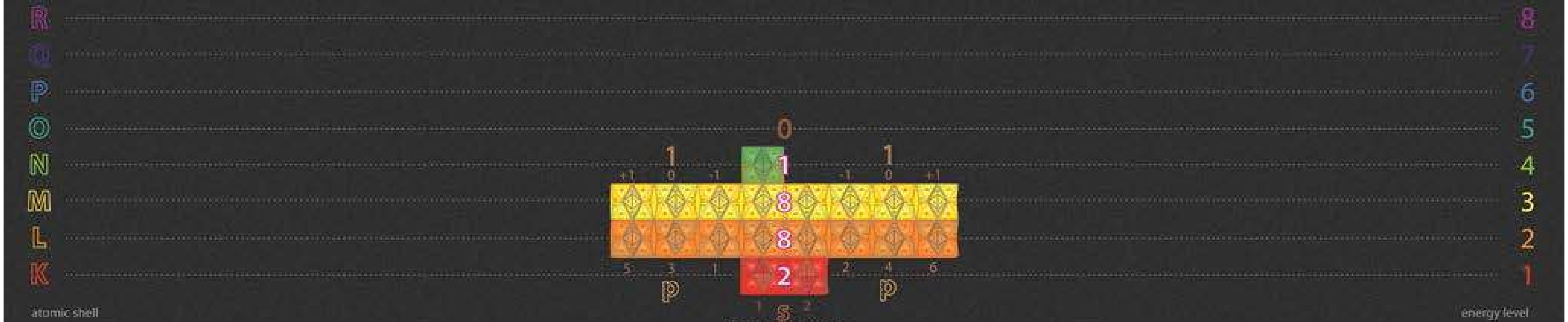
R
Q
P
O
N
M
L
K
0
1
1
+1
0
-1
1
-1
0
+1
1
8
8
2
5
3
1
2
4
6
p
p
1
s
2
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.19 - Potassium atom

Copyright ABRAHAM [2008] - All rights reserved

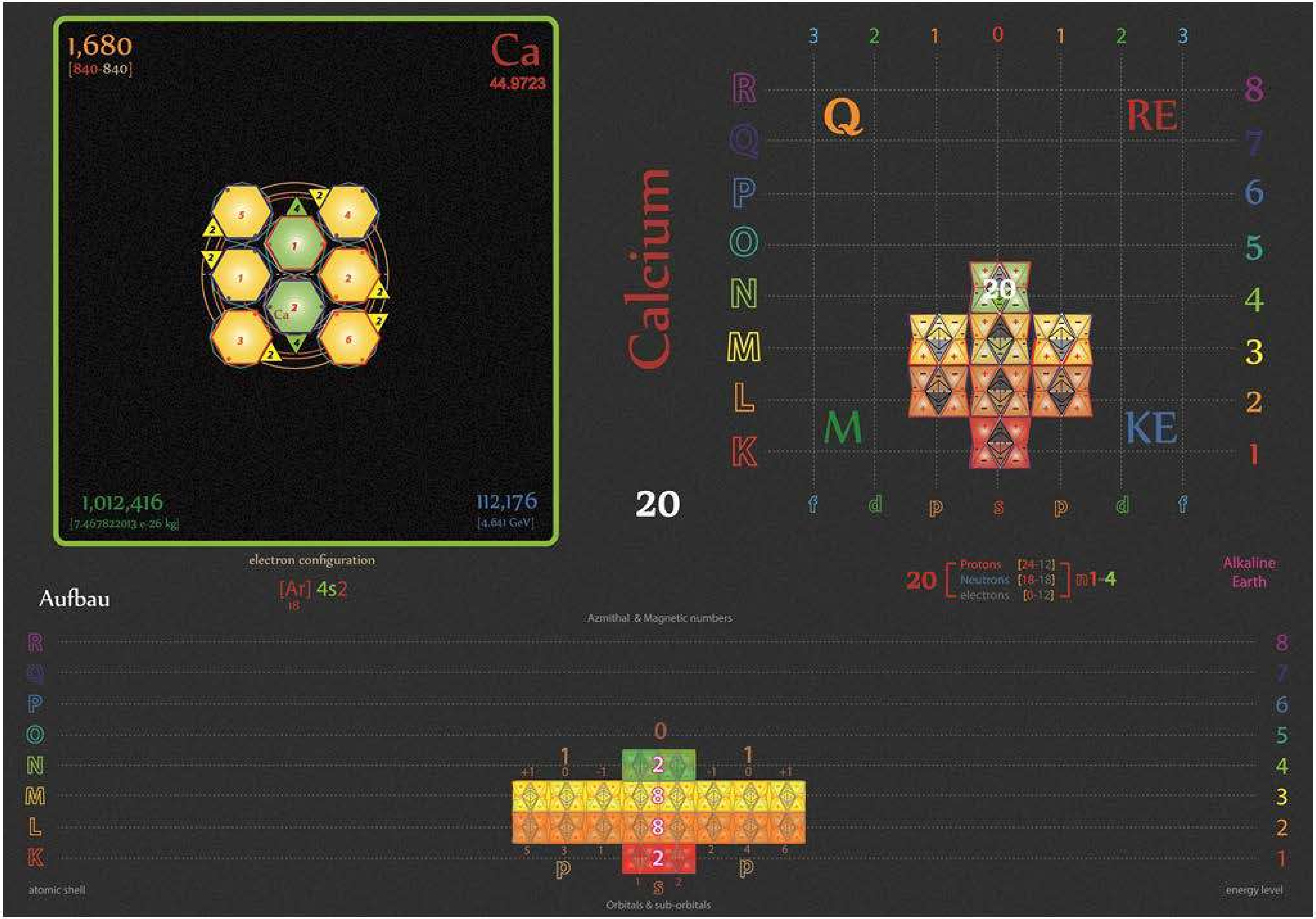
1,680
[840-840]
Ca
44.9723
1,012,416
[7.467822013 e-26 kg]
112,176
[4.641 GeV]
Calcium
20
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Ar] 4s2
18
Aufbau
20 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Alkaline Earth
Azmithal & Magnetic numbers
0
1
1
+1 0 -1 -1 0 +1
2
8
8
2
5 3 1 2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.20 - Calcium atom

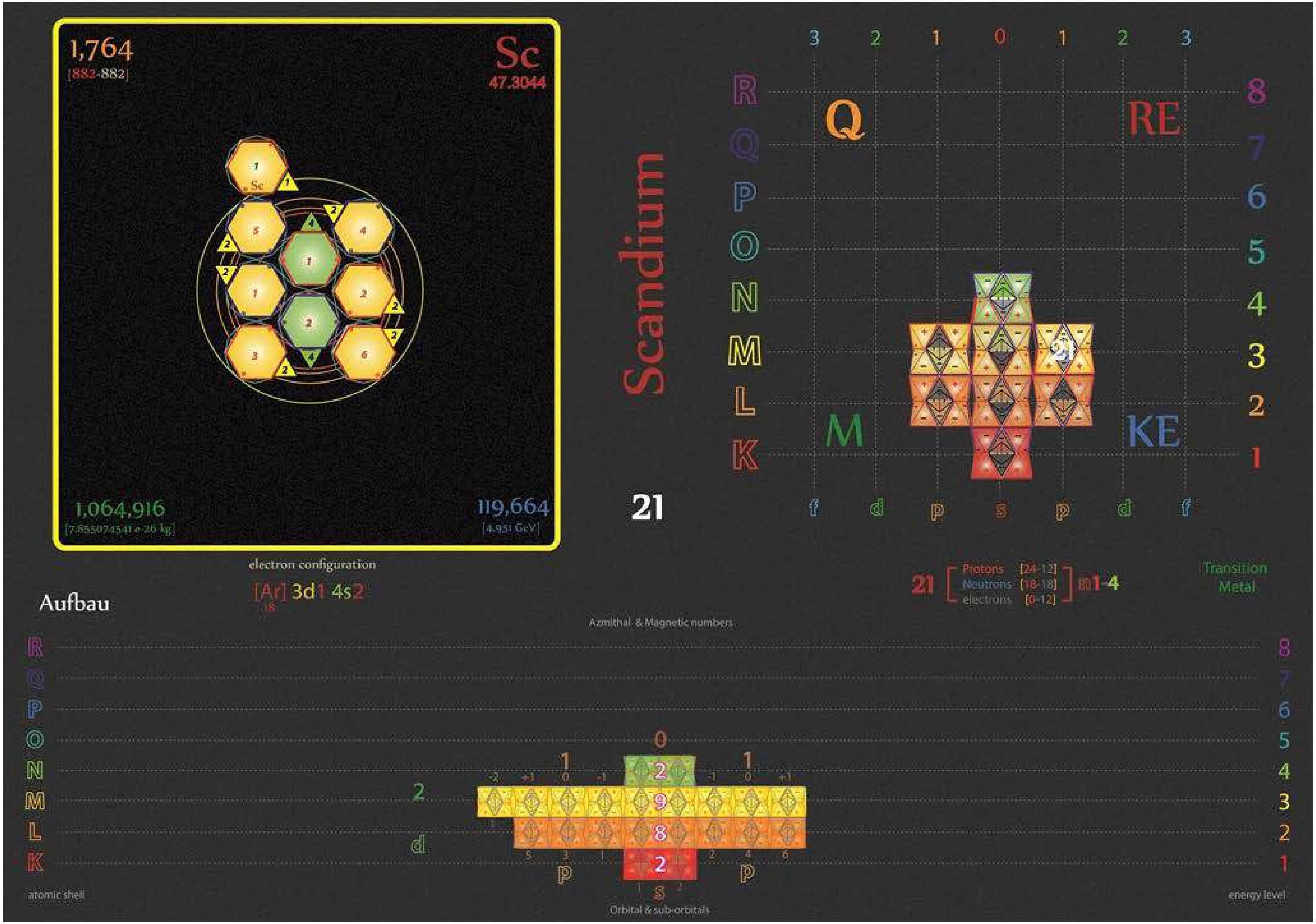
1,764
[882-882]
Sc
47.3044
1
Sc
5
4
1
2
3
6
1,064,916
[7.855074541 e-26 kg]
119,664
[4.951 GeV]
electron configuration
[Ar] 3d1 4s2
18
Scandium
21
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
21
21 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Aufbau
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
-2 +1 0 -1 -1 0 +1
2
9
8
2
2
d
5 3 1 2 4 6
p
p
1 2
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.21 - Scandium atom

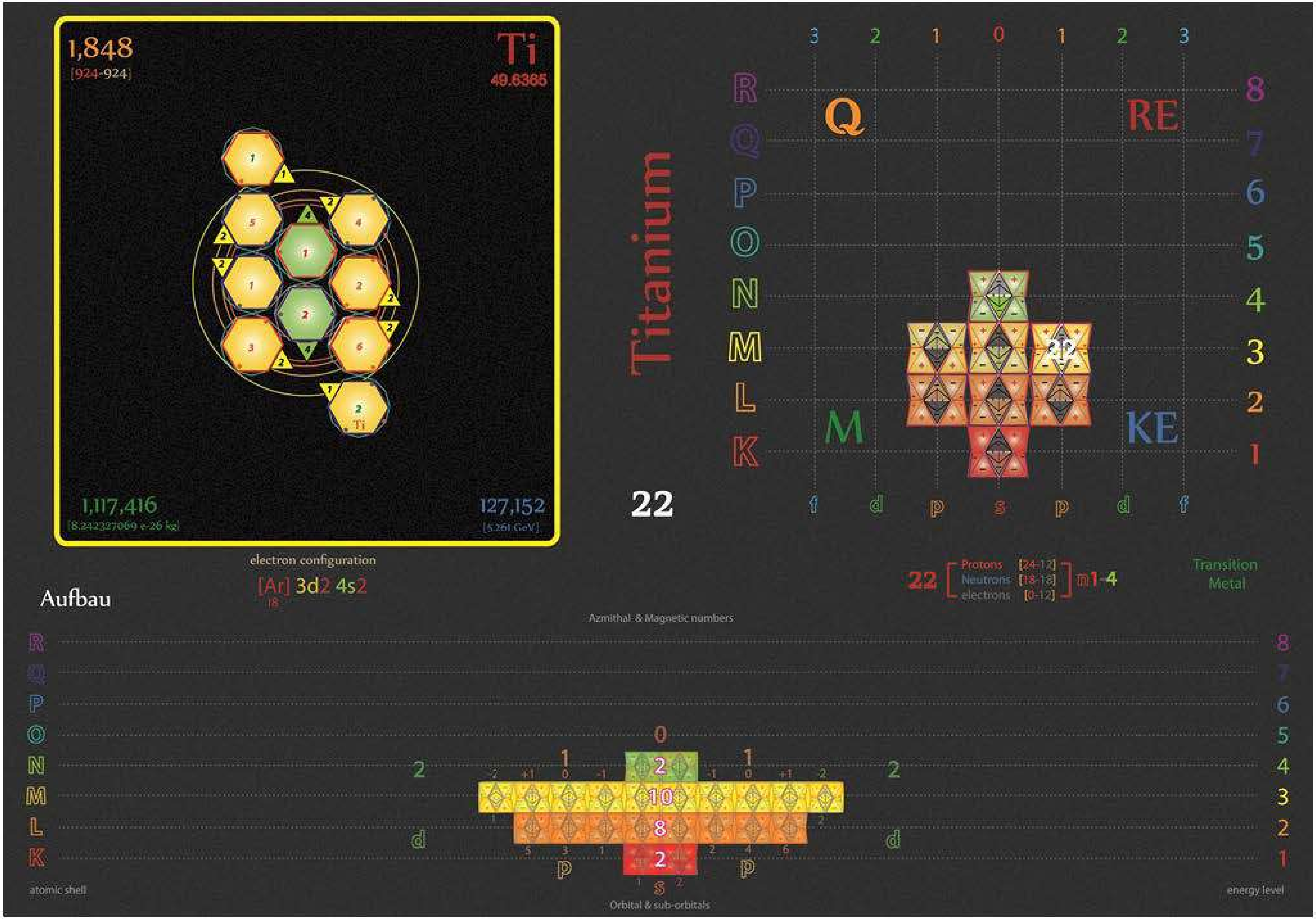
1,848
[924-924]
Ti
49.6365
1,117,416
[8.242327069 e-26 kg]
127,152
[5.261 GeV]
Titanium
22
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
22
f d p s p d f
electron configuration
[Ar] 3d2 4s2
18
Aufbau
22 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal- & Magnetic numbers
0
1 1
2 2
-2 +1 0 -1 -1 0 +1 2
2
10
8
2
d d
p p
s
Orbital & sub-orbitals
atomic shell
energy level

Tetryonics 51.22 - Titanium atom

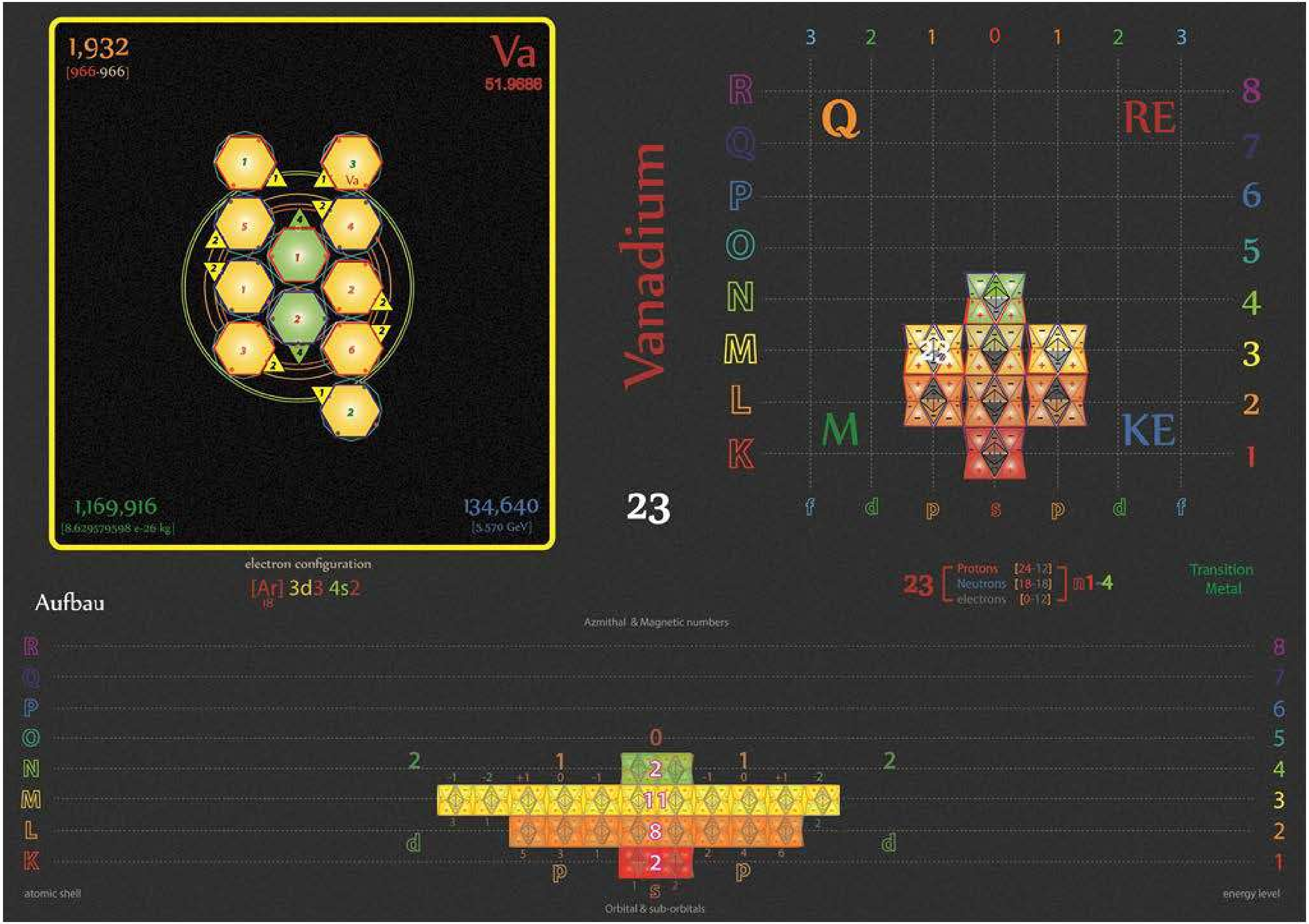
1,932
[966-966]
Va
51.9886
1,169,916
[8.629579598 e-26 kg]
134,640
[5.570 GeV]
Vanadium
23
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Ar] 3d3 4s2
18
23 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Aufbau
Azmithal & Magnetic numbers
0
2 1 1 2
-1 -2 +1 0 -1 -1 0 +1 -2
2
11
8
2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.23 - Vanadium atom

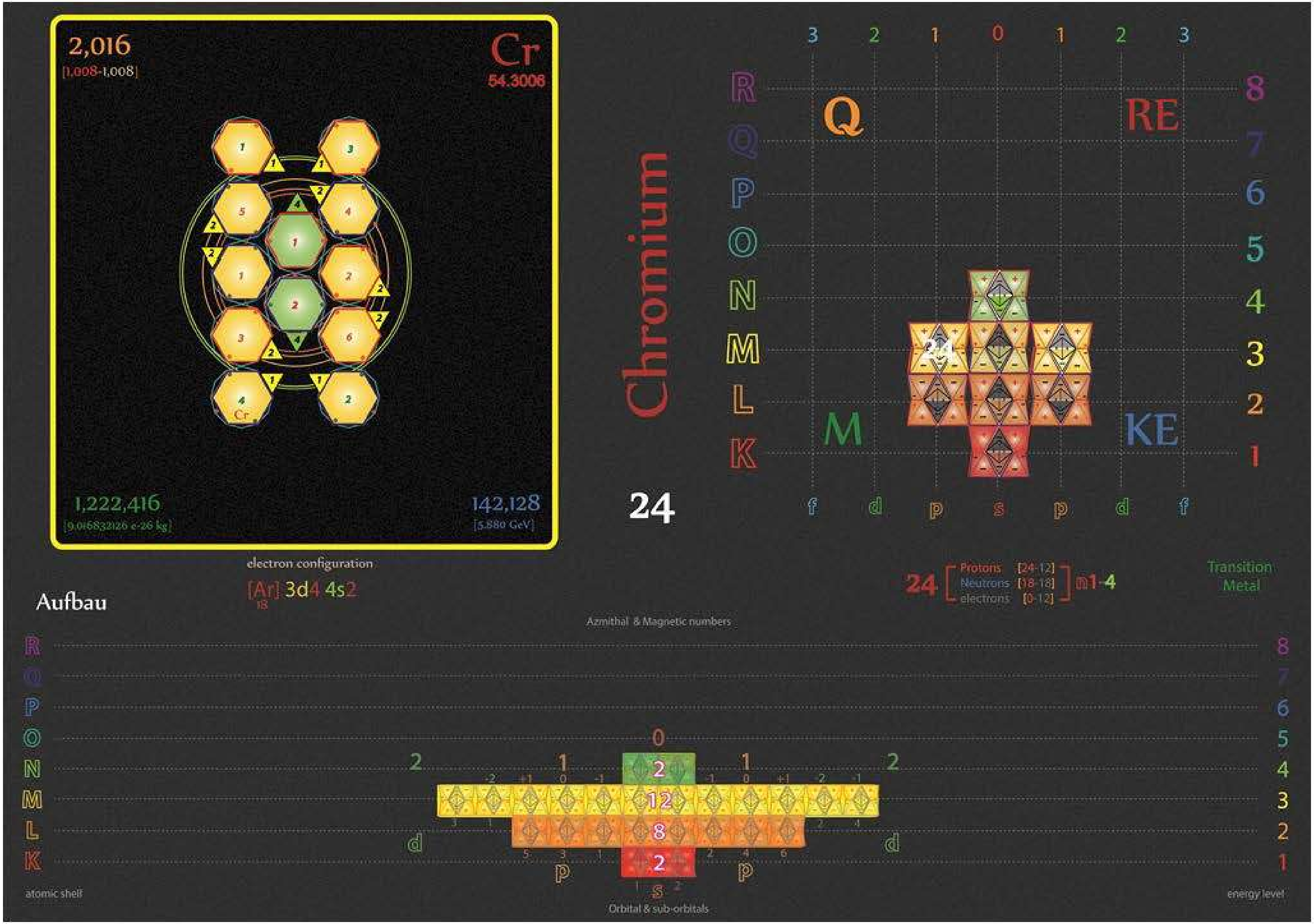
2,016
[1,008-1,008]
Cr
54.3008
1,222,416
[9.016832126 e-26 kg]
142,128
[5.880 GeV]
Chromium
24
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
[Ar] 3d4 4s2
18
Aufbau
24 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
12
8
2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.24 - Chromium atom

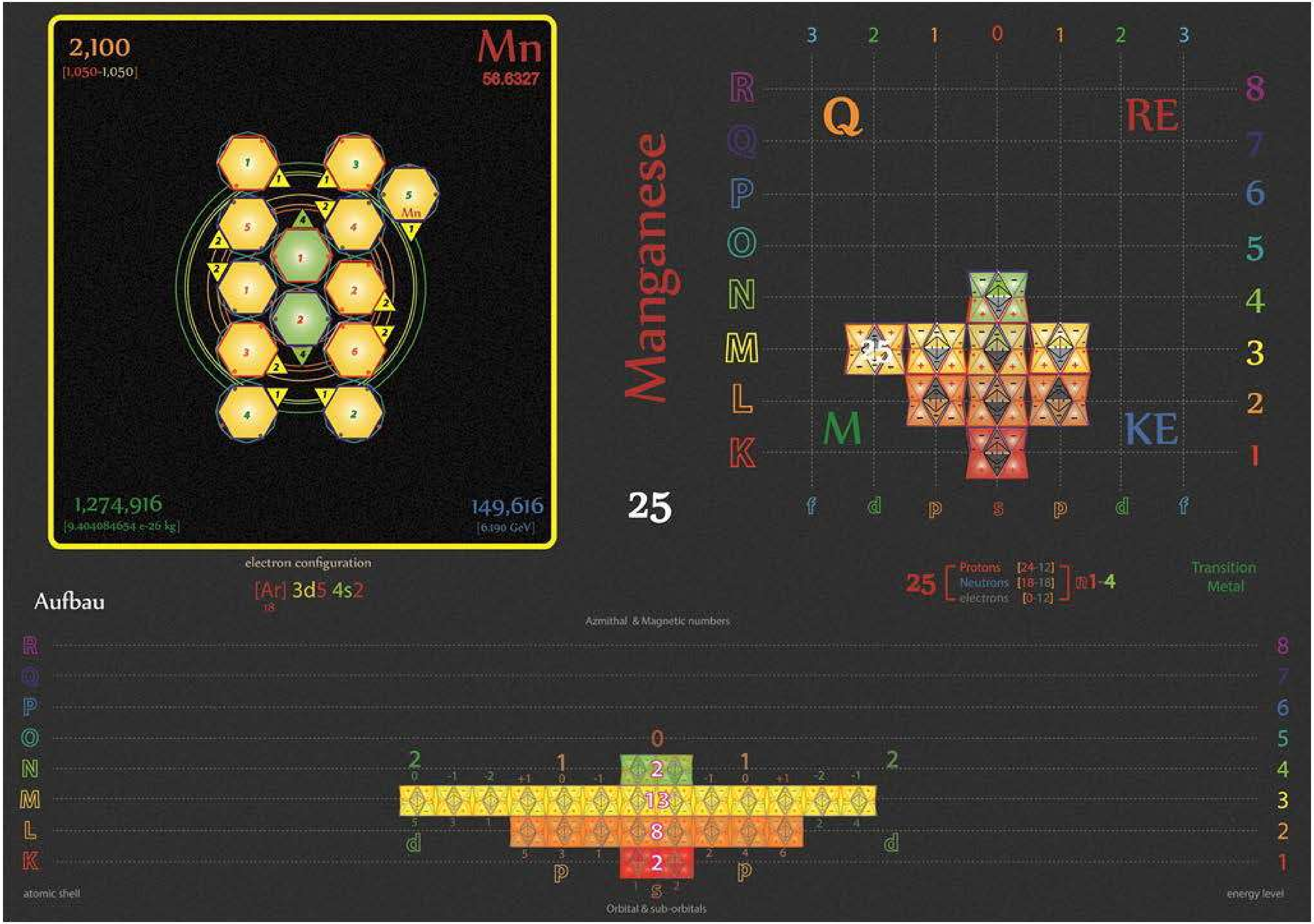
2,100
[1,050-1,050]
Mn
56.6327
5
Mn
1,274,916
[9.404084654 e-26 kg]
149,616
[6.190 GeV]
Manganese
25
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
25
f d p s p d f
electron configuration
[Ar] 3d5 4s2
18
Aufbau
25
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-4
Transition Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
1
1
2
2
13
8
2
d
p
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.25 - Manganese atom

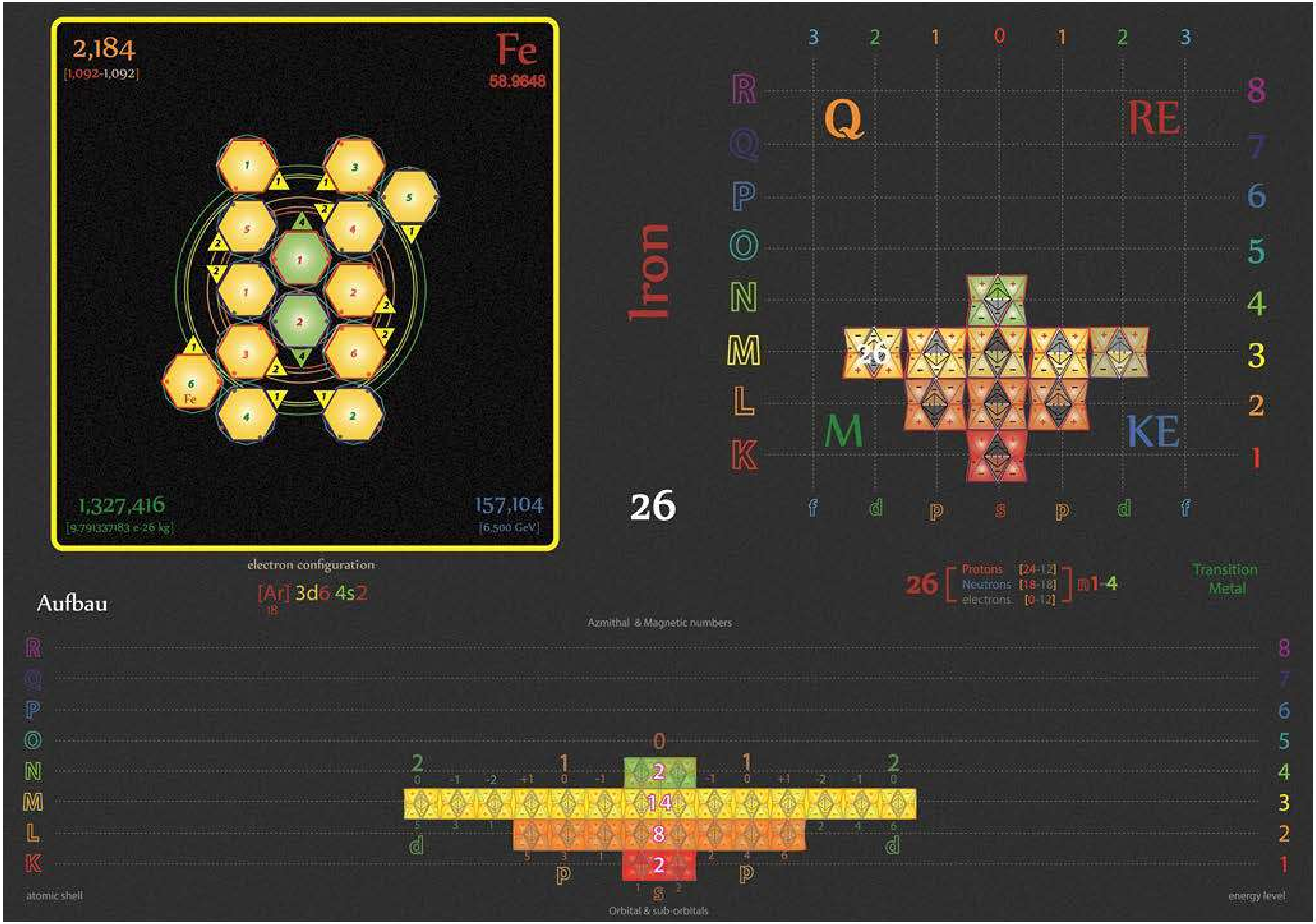
2,184
[1,092-1,092]
Fe
58.9848
6
Fe
1,327,416
[9.791337183 e-26 kg]
157,104
[6,500 GeV]
Iron
26
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
26
f d p s p d f
electron configuration
[Ar] 3d6 4s2
18
Aufbau
26 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2
14
8
2
d
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.26 - Iron atom

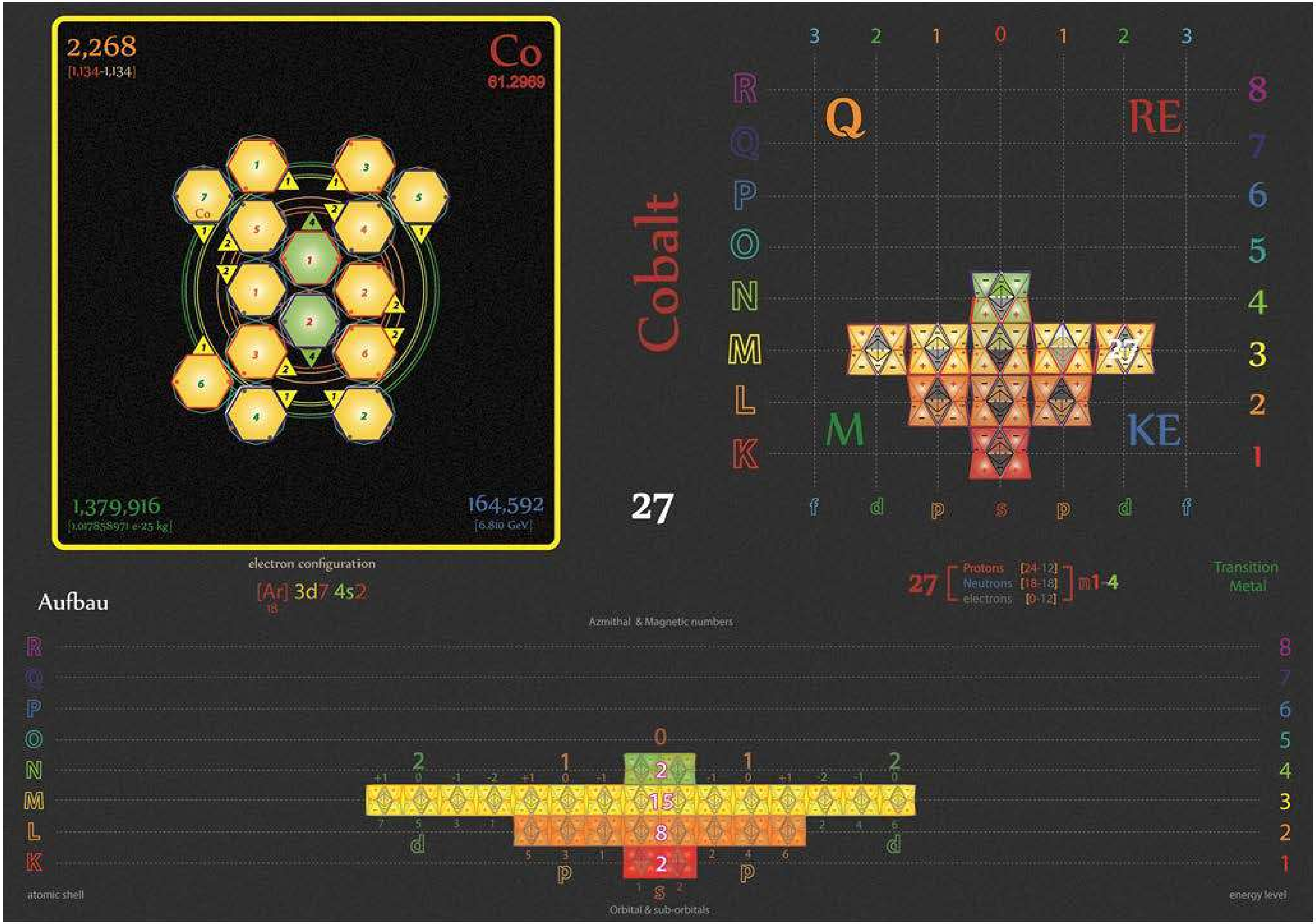
2,268
[1,134-1,134]
Co
61.2969
Co
1,379,916
[1.017858971 e-25 kg]
164,592
[6.810 GeV]
Cobalt
27
electron configuration
[Ar] 3d7 4s2
18
Aufbau
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
27
f
d
p
s
p
d
f
27
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2
1
1
2
+1
0
-1
-2
+1
0
+1
-1
0
+1
-2
-1
0
2
15
8
2
7
5
3
1
2
4
6
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.27 - Cobalt atom

Copyright ABRAHAM [2008] - All rights reserved

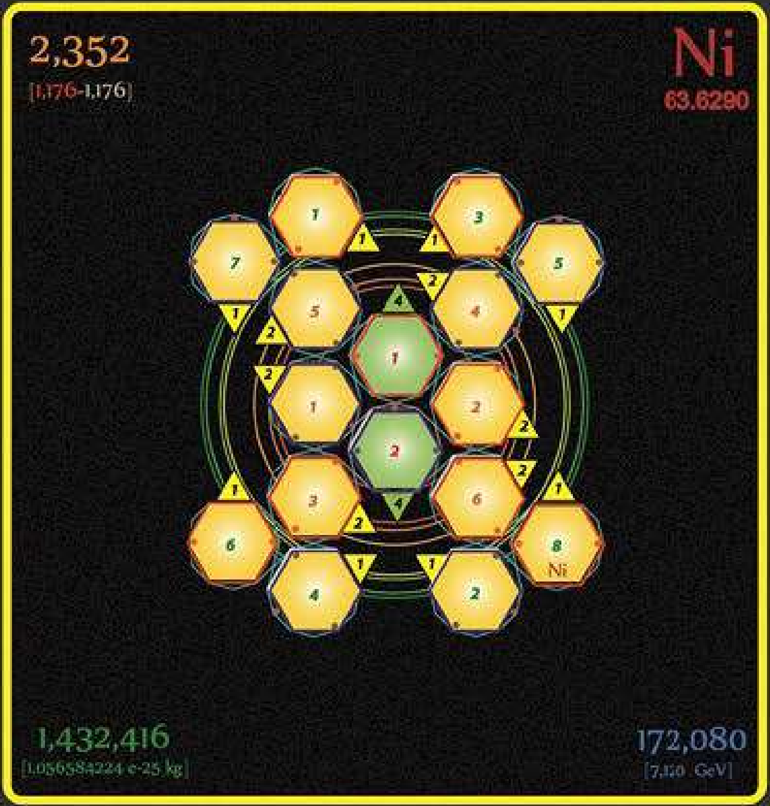

2,352
[1,176-1,176]

Ni
63.6290

1,432,416
[1.056584224 e-25 kg]

172,080
[7,110 GeV]

Nickel

28

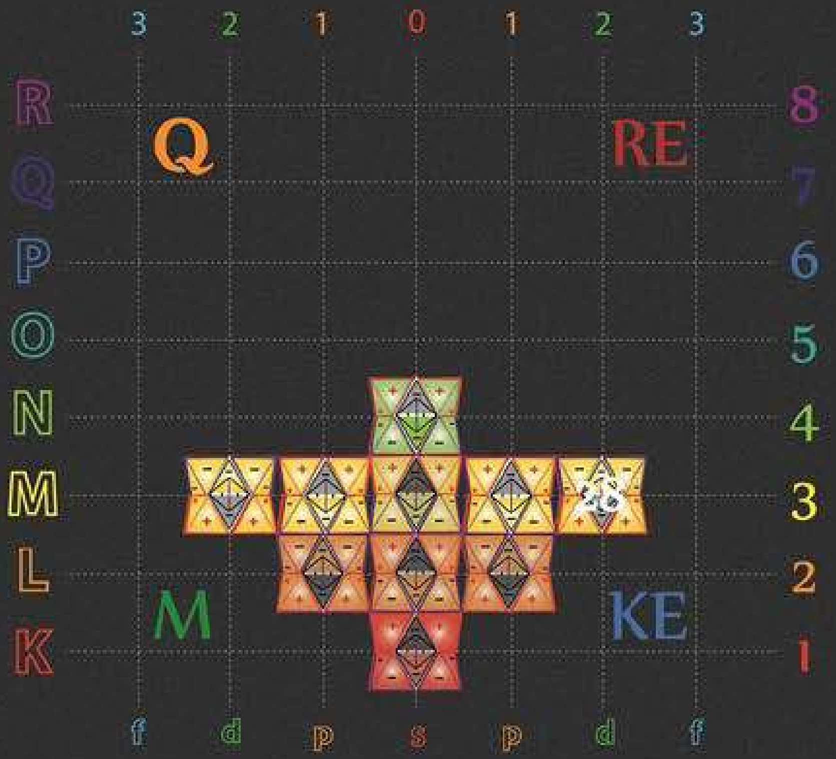

3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
28
f
d
p
s
p
d
f

electron configuration

[Ar] 3d8 4s2
18

28 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4

Transition Metal

Aufbau

Azmithal & Magnetic numbers

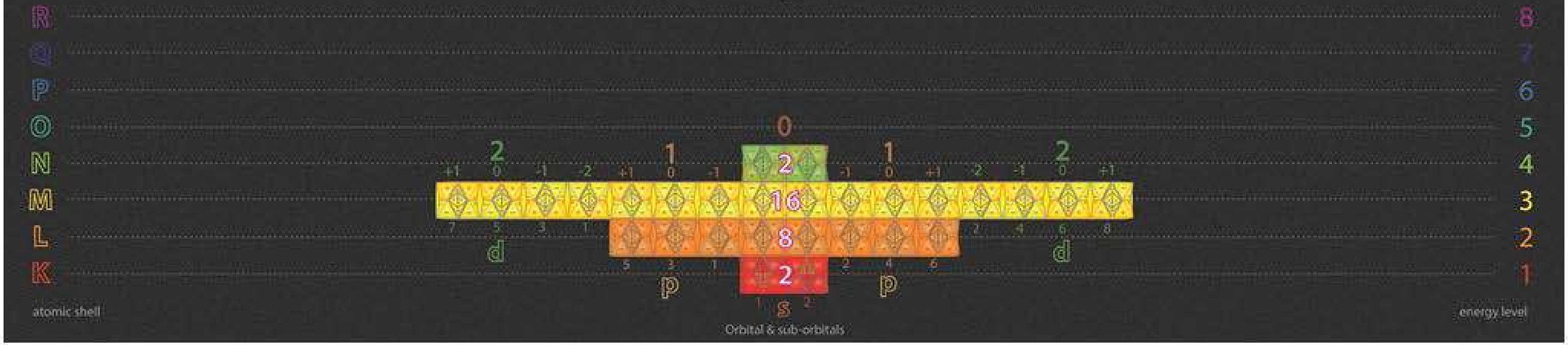

R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
+1
2
16
8
2
7
5
3
1
2
4
6
8
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level

Orbital & sub-orbitals

Tetryonics 51.28 - Nickel atom

Copyright ABRAHAM [2008] - All rights reserved

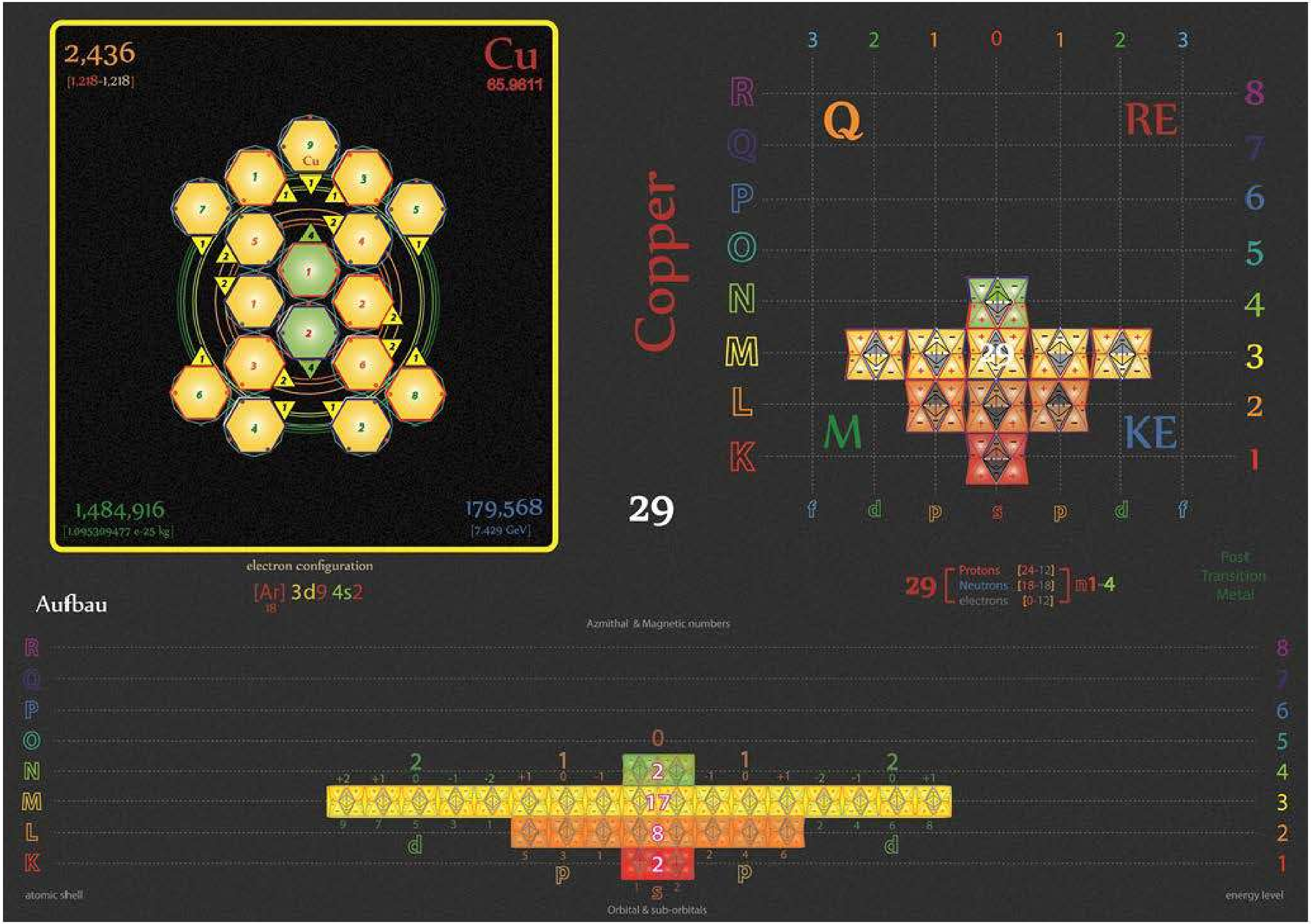
2,436
[1,218-1,218]
Cu
65.9811
Cu
1,484,916
[1.095309477 e-25 kg]
179,568
[7.429 GeV]
electron configuration
[Ar] 3d9 4s2
18
Aufbau
Copper
29
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
29
f d p s p d f
29 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Post Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
2
17
8
2
9 7 5 3 1 2 4 6 8
5 3 1 2 4 6
1 2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.29 - Copper atom

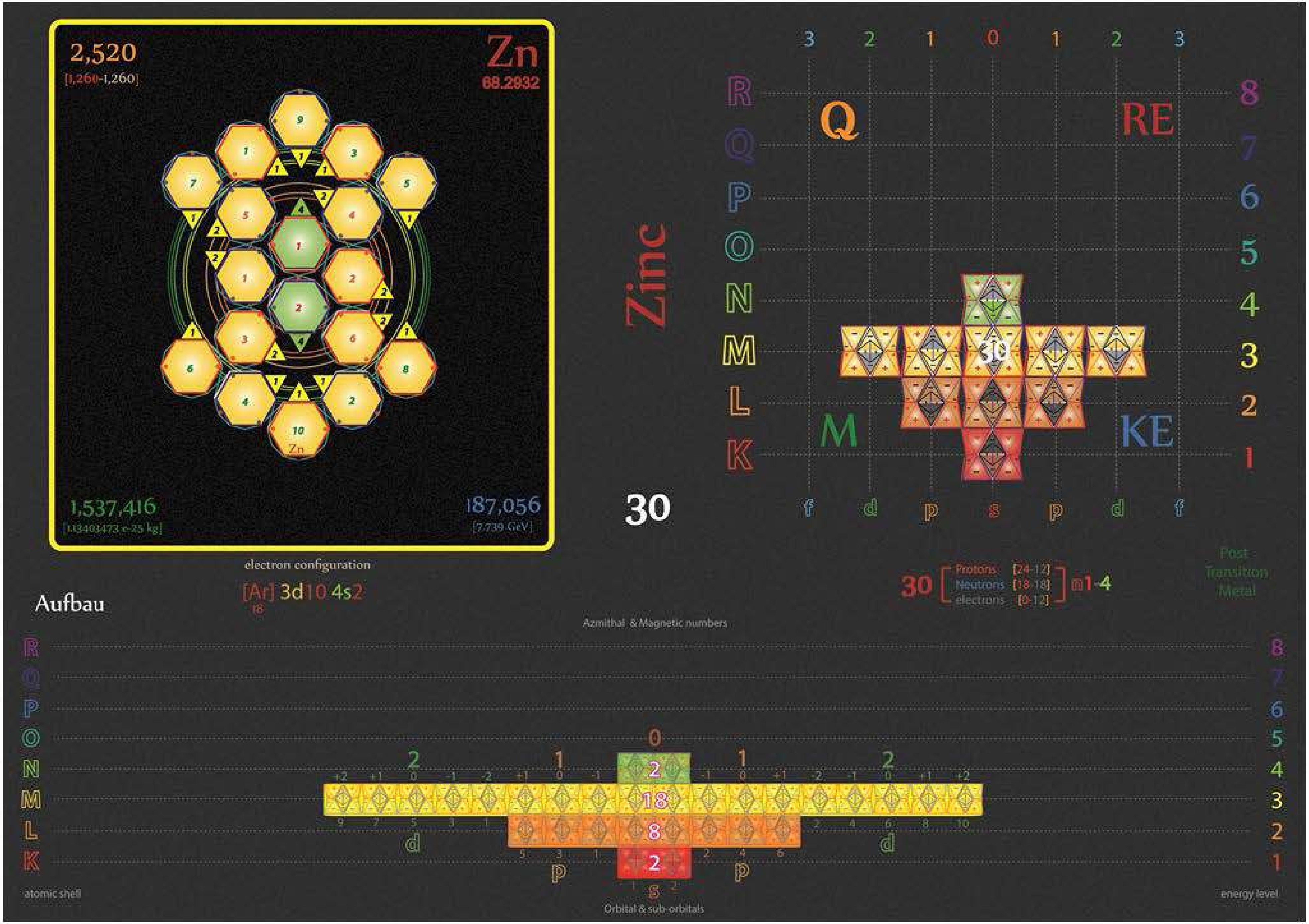
2,520
[1,260-1,260]
Zn
68.2932
9
1
3
7
5
5
4
1
1
2
2
3
6
6
8
4
2
10
Zn
1,537,416
[1.13403473 e-25 kg]
187,056
[7.739 GeV]
electron configuration
[Ar] 3d10 4s2
18
Aufbau
Zinc
30
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
30
f d p s p d f
30
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Post-Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
18
8
2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d d p p s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.30 - Zinc atom

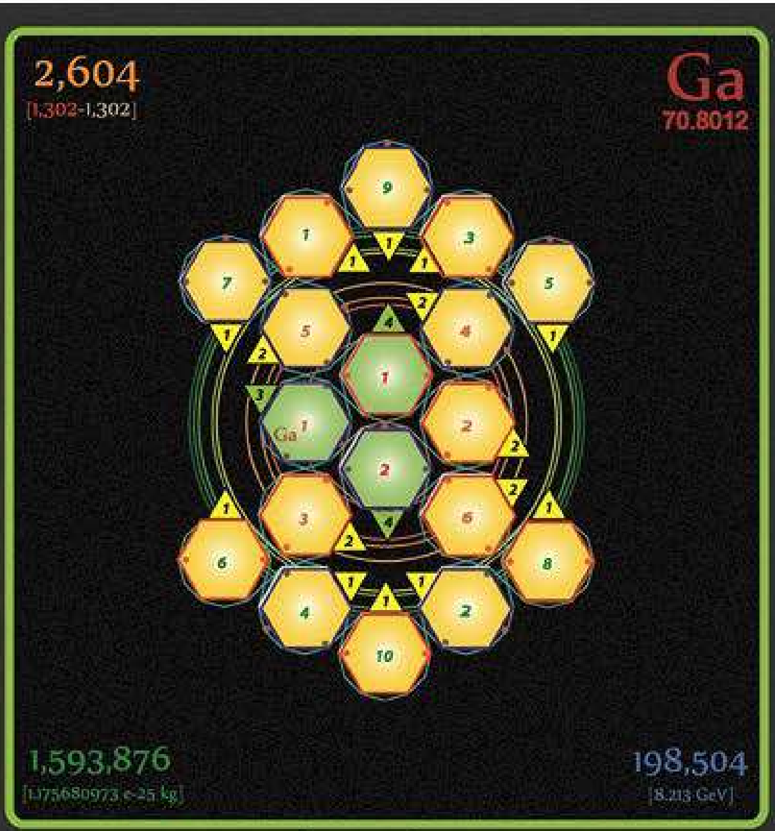
2,604
[1,302-1,302]
Ga
70.8012
Ga
1,593,876
[1.175680973 e-25 kg]
198,504
[8.213 GeV]

Gallium

31

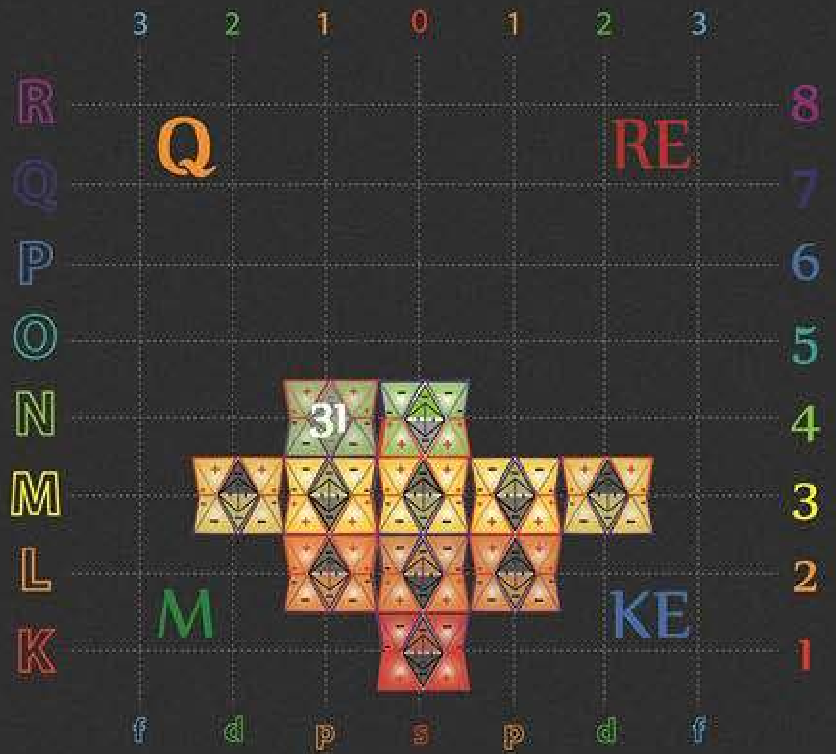
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
31
8
7
6
5
4
3
2
1
f d p s p d f

31 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4

Metaloid

electron configuration

[Ar] 3d10 4s2 4p1
18

Aufbau

Azmithal & Magnetic numbers

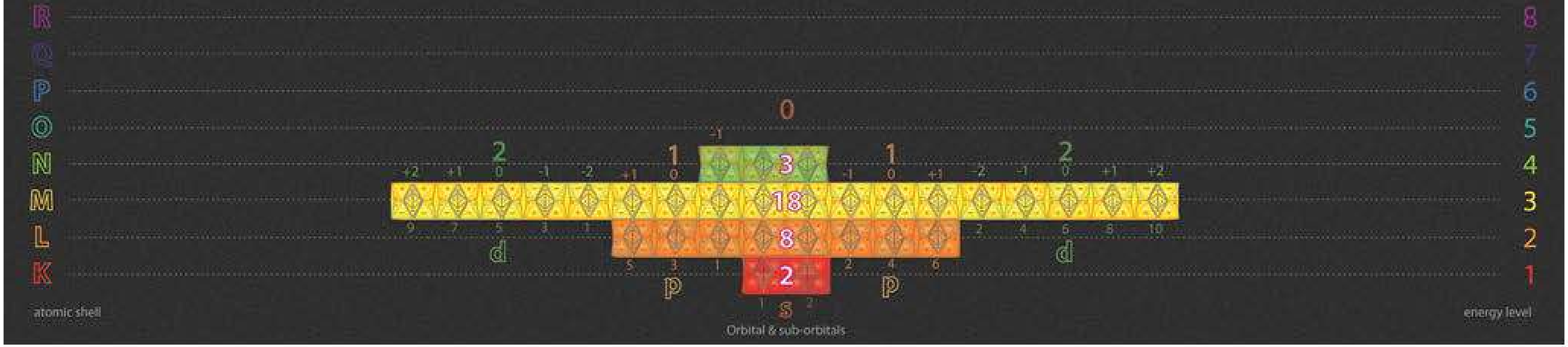
R
Q
P
O
N
M
L
K
0
-1
2
1
1
2
+2 +1 0 -1 -2 +1 0 -1 0 +1 -2 -1 0 +1 +2
3
18
8
2
9 7 5 3 1 2 4 6 8 10
d
d
5 3 1 2 4 6
p
p
1 2
s
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbital & sub-orbitals

Tetryonics 51.31 - Gallium atom

Copyright ABRAHAM [2008] - All rights reserved

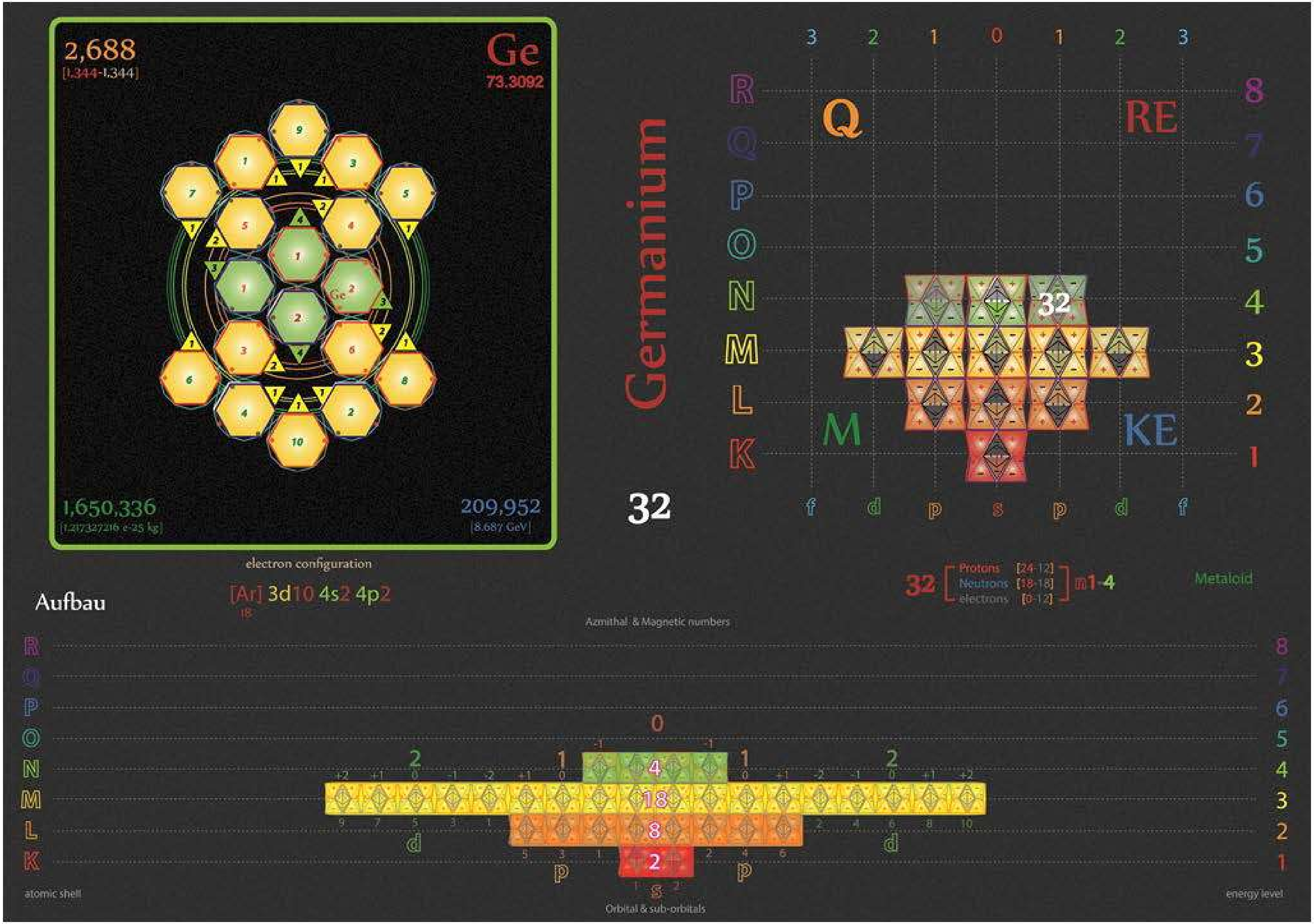
2,688
[1.344-1.344]
Ge
73.3092
1,650,336
[1.217327216 e-25 kg]
209,952
[8.687 GeV]
Germanium
32
electron configuration
[Ar] 3d10 4s2 4p2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
32 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Metaloid
Azmithal & Magnetic numbers
0
1 1
2 2
4
18
8
2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.32 - Germanium atom

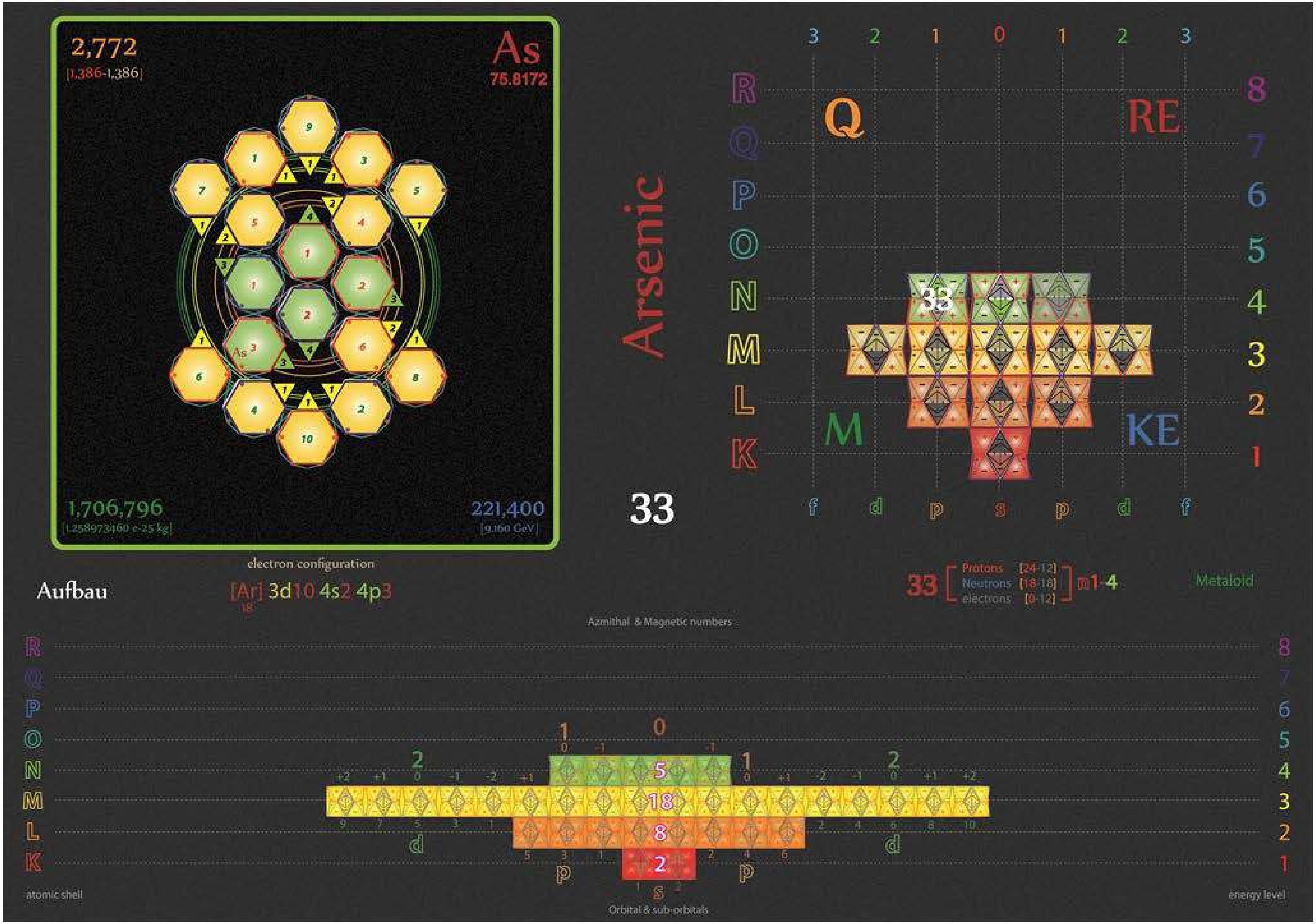
2,772
[1,386-1,386]
As
75.8172
1,706,796
[1.258973460 e-25 kg]
221,400
[9.160 GeV]
Arsenic
33
electron configuration
Aufbau
[Ar] 3d10 4s2 4p3
18
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
33
f d p s p d f
33 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Metaloid
Azmithal & Magnetic numbers
5
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.33 - Arsenic atom

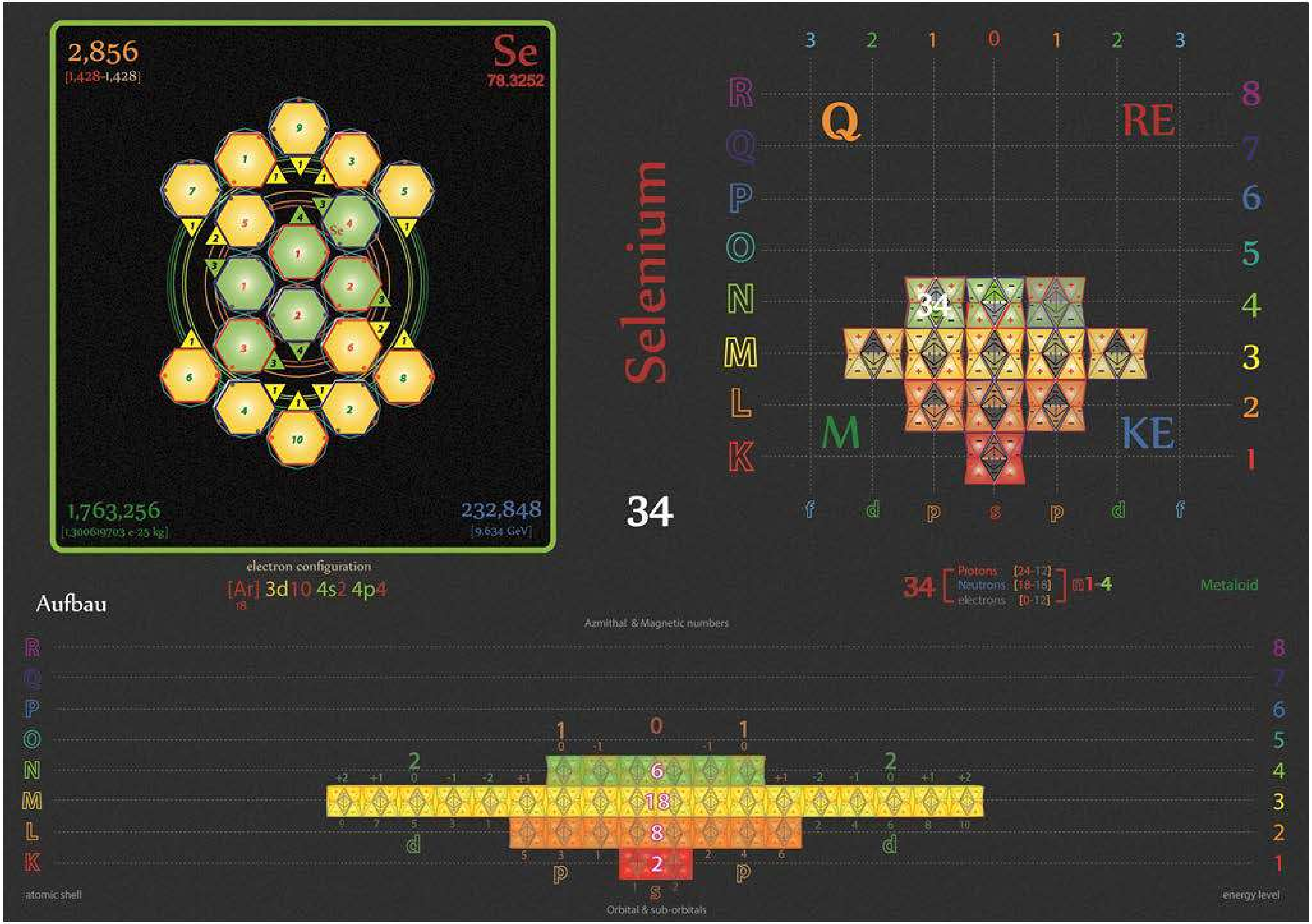
2,856
[1,428-1,428]
Se
78.3252
1,763,256
[1.30069703 e-25 kg]
232,848
[9.634 GeV]
electron configuration
[Ar] 3d10 4s2 4p4
18
Selenium
34
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
34 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Metaloid
Azmithal & Magnetic numbers
6
18
8
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.34 - Selenium atom

Copyright ABRAHAM [2008] - All rights reserved

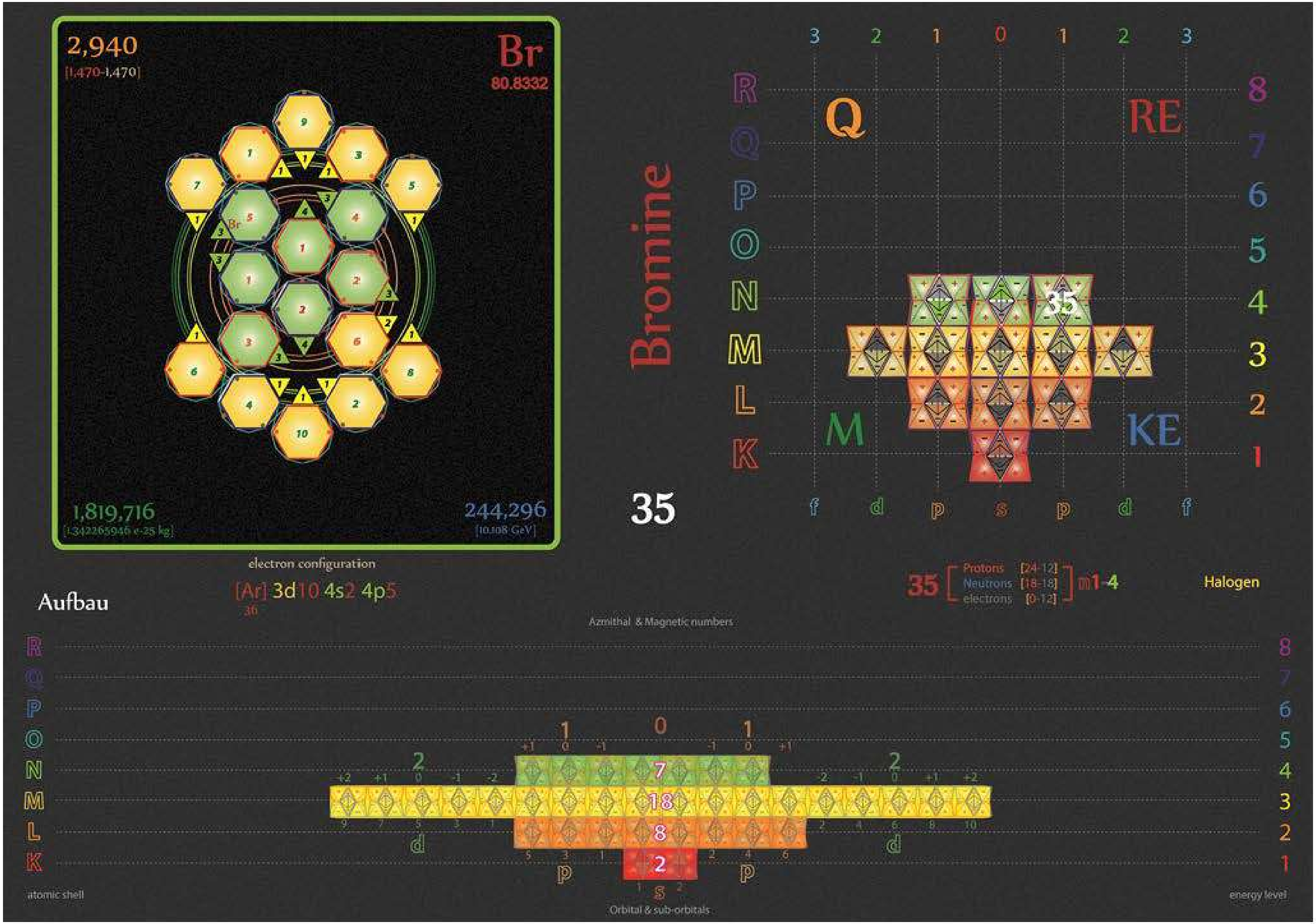
2,940
[1.470-1.470]
Br
80.8332
1,819,716
[1.342265946 e-25 kg]
244,296
[10.108 GeV]
Bromine
35
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
35
f d p s p d f
electron configuration
[Ar] 3d10 4s2 4p5
36
Aufbau
35
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Halogen
Azmithal & Magnetic numbers
7
18
8
2
d
p
s
d
p
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.35 - Bromine atom

Copyright ABRAHAM [2008] - All rights reserved

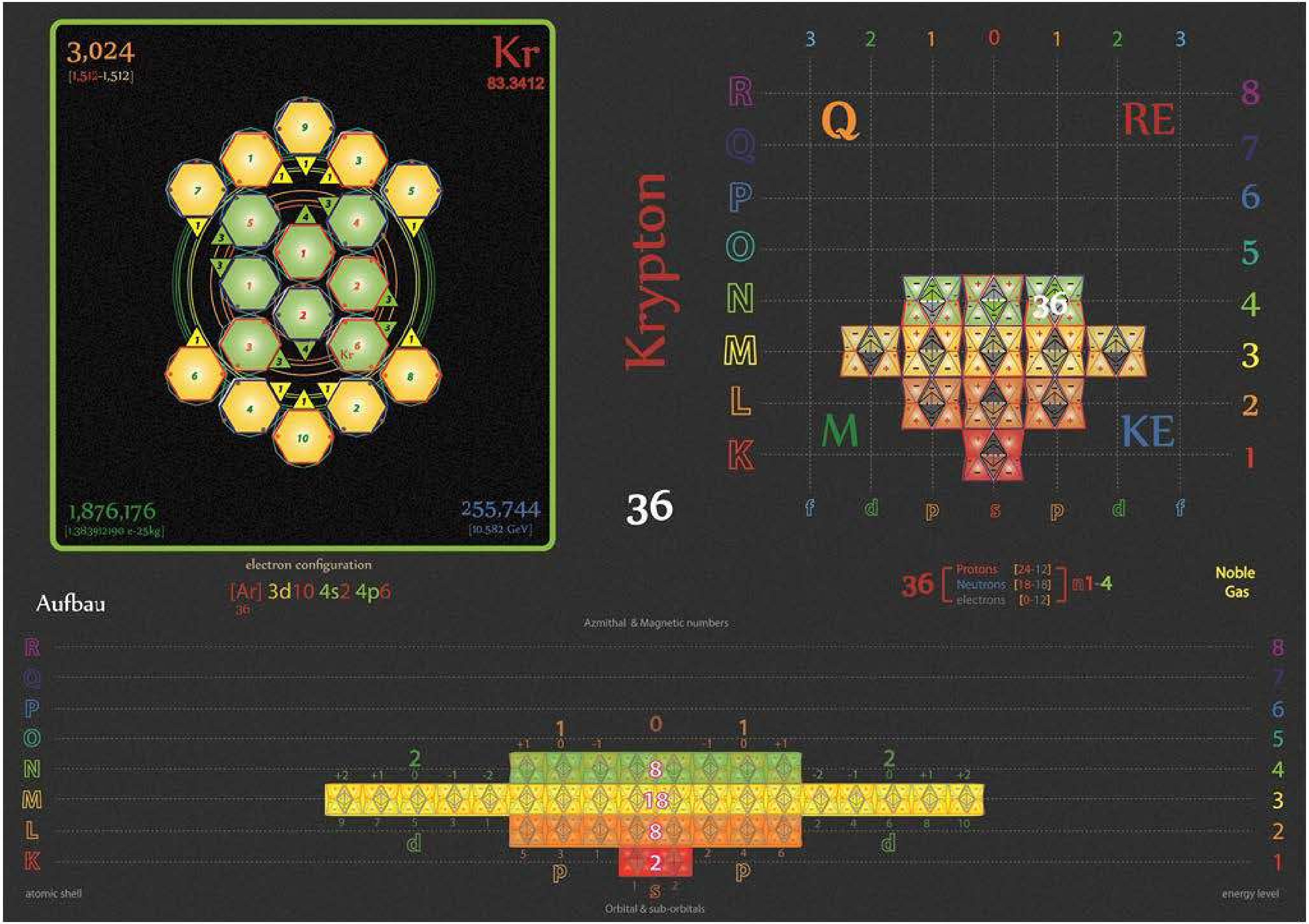
3,024
[1,512-1,512]
Kr
83.3412
1,876,176
[1.38391290 e-25kg]
255,744
[10.582 GeV]
Krypton
36
electron configuration
[Ar] 3d10 4s2 4p6
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
36 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Noble Gas
Azmithal & Magnetic numbers
8
18
8
2
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.36 - Krypton atom

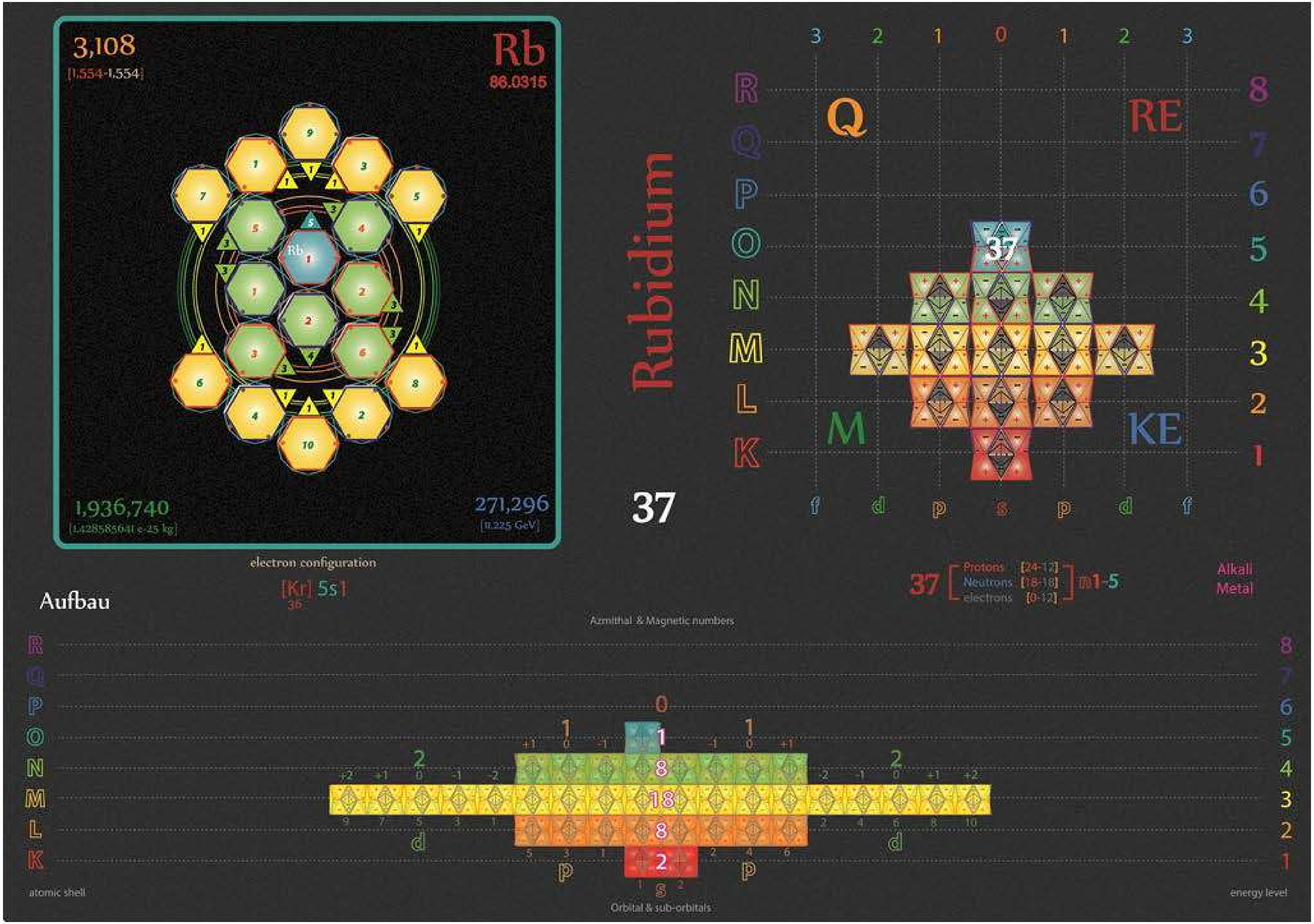
3,108
[1,554-1,554]
Rb
86.0315
Rb
1,936,740
[1.428585641 e-25 kg]
271,296
[11.225 GeV]
electron configuration
[Kr] 5s1
36
Rubidium
37
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
37
f d p s p d f
37
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-5
Alkali
Metal
Aufbau
Azmithal & Magnetic numbers
0
1
1
2
2
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
1
8
18
8
2
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.37 - Rubidium atom

Copyright ABRAHAM [2008] - All rights reserved

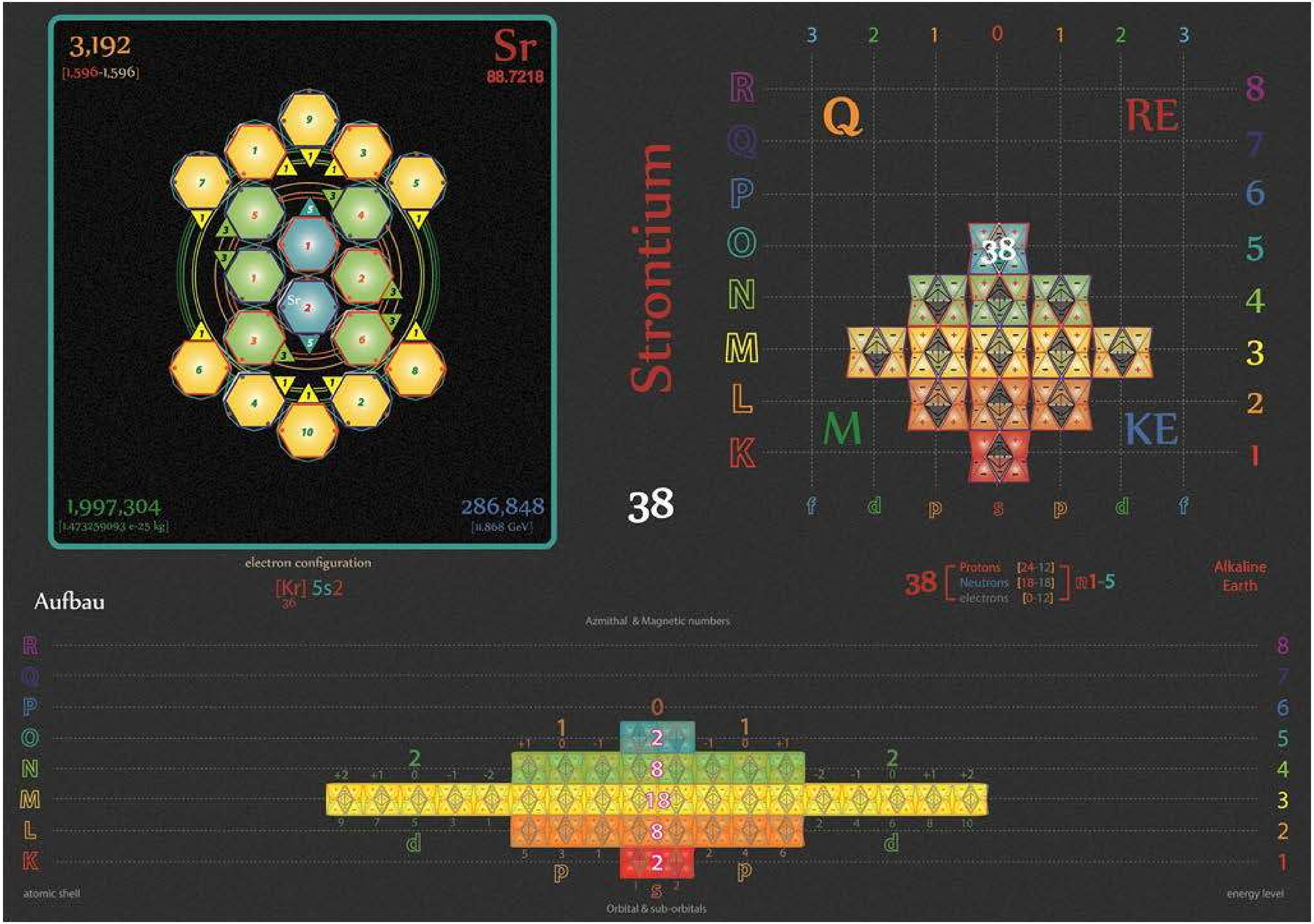
3,192
[1,596-1,596]
Sr
88.7218
1,997,304
[1.473259093 e-25 kg]
286,848
[11.868 GeV]
Strontium
38
electron configuration
[Kr] 5s2
36
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
38
38
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-5
Alkaline
Earth
Azmithal & Magnetic numbers
0
1
1
2
2
2
8
18
8
2
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.38 - Strontium atom

Copyright ABRAHAM [2008] - All rights reserved

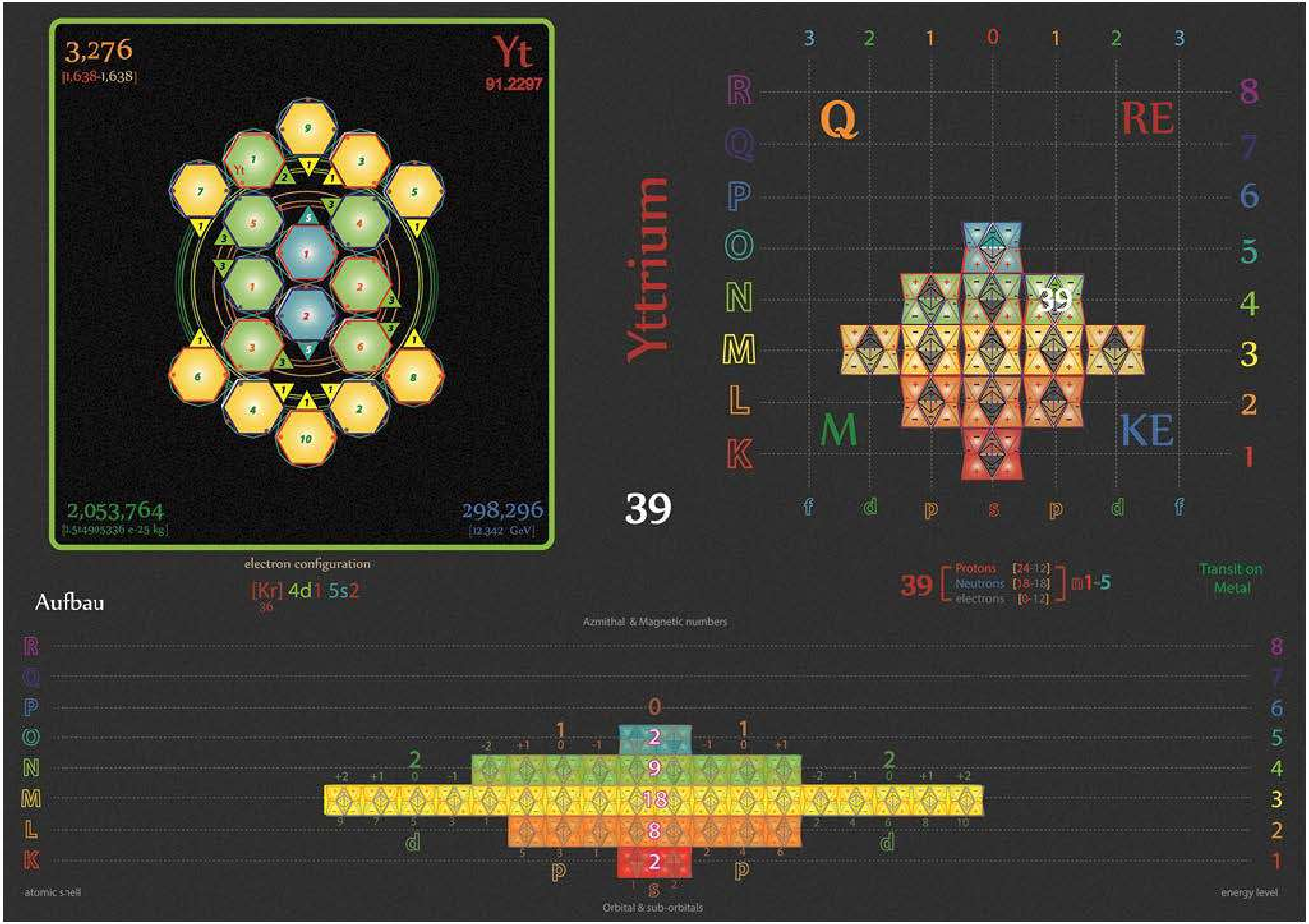
3,276
[1,638-1,638]
Yt
91.2297
2,053,764
[1.514905236 e-25 kg]
298,296
[12.342 GeV]
Yttrium
39
electron configuration
[Kr] 4d1 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
39
f d p s p d f
39 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
-2 +1 0 -1
-1 0 +1
+2 +1 0 -1
-2 -1 0 +1 +2
2
9
18
8
2
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.39 - Yttrium atom

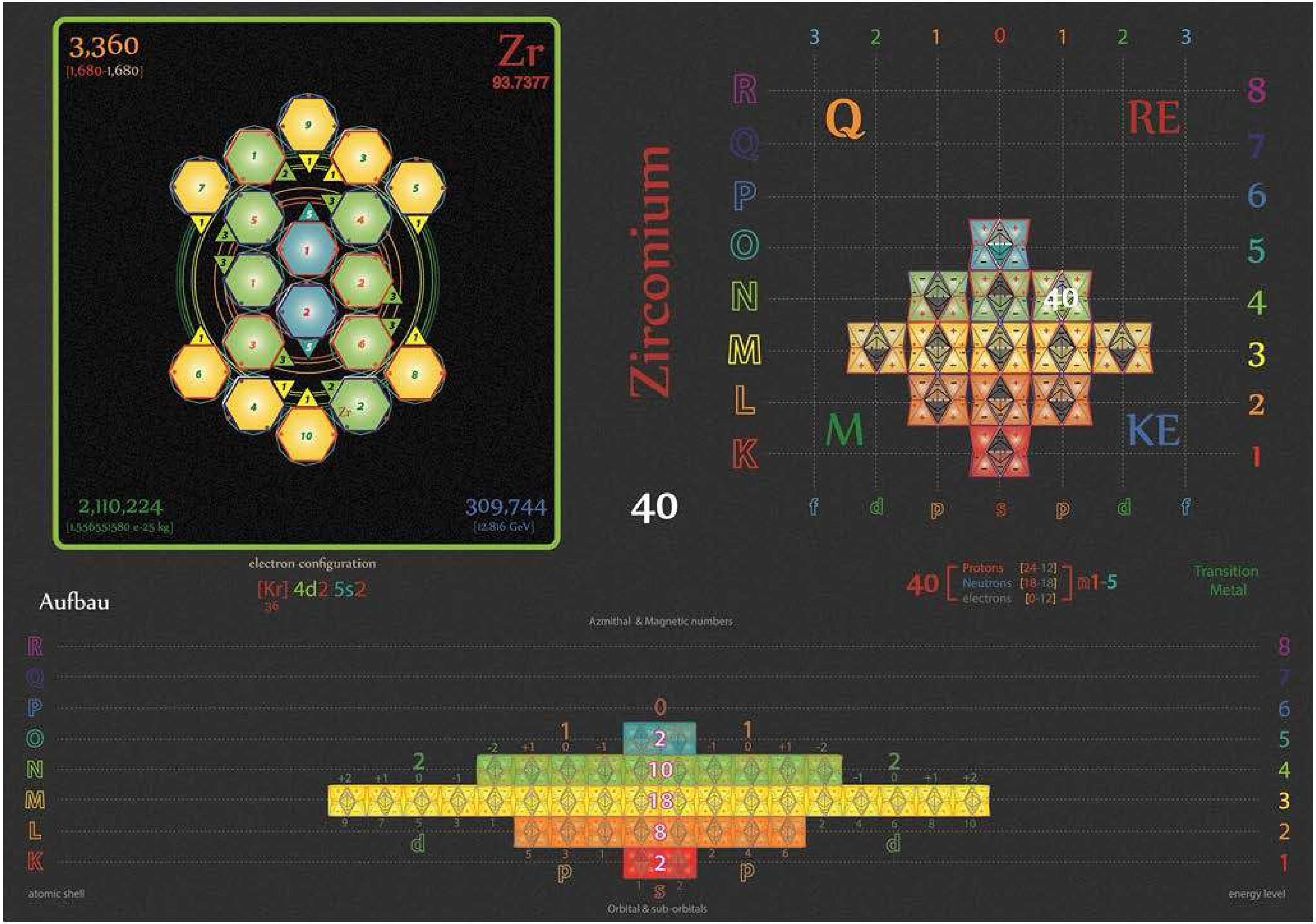
3,360
[1,680-1,680]
Zr
93.7377
2,110,224
[1.5565551580 e-25 kg]
309,744
[12.816 GeV]
Zr
electron configuration
[Kr] 4d2 5s2
36
Zirconium
40
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
40
40 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Aufbau
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
10
18
8
2
-2 +1 0 -1
1
-1 0 +1 -2
2
+2 +1 0 -1
-1 0 +1 +2
9 7 5 3 1
2 4 6 8 10
d
5 3 1
2 4 6
p
1 2
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.40 - Zirconium atom

Copyright ABRAHAM [2008] - All rights reserved

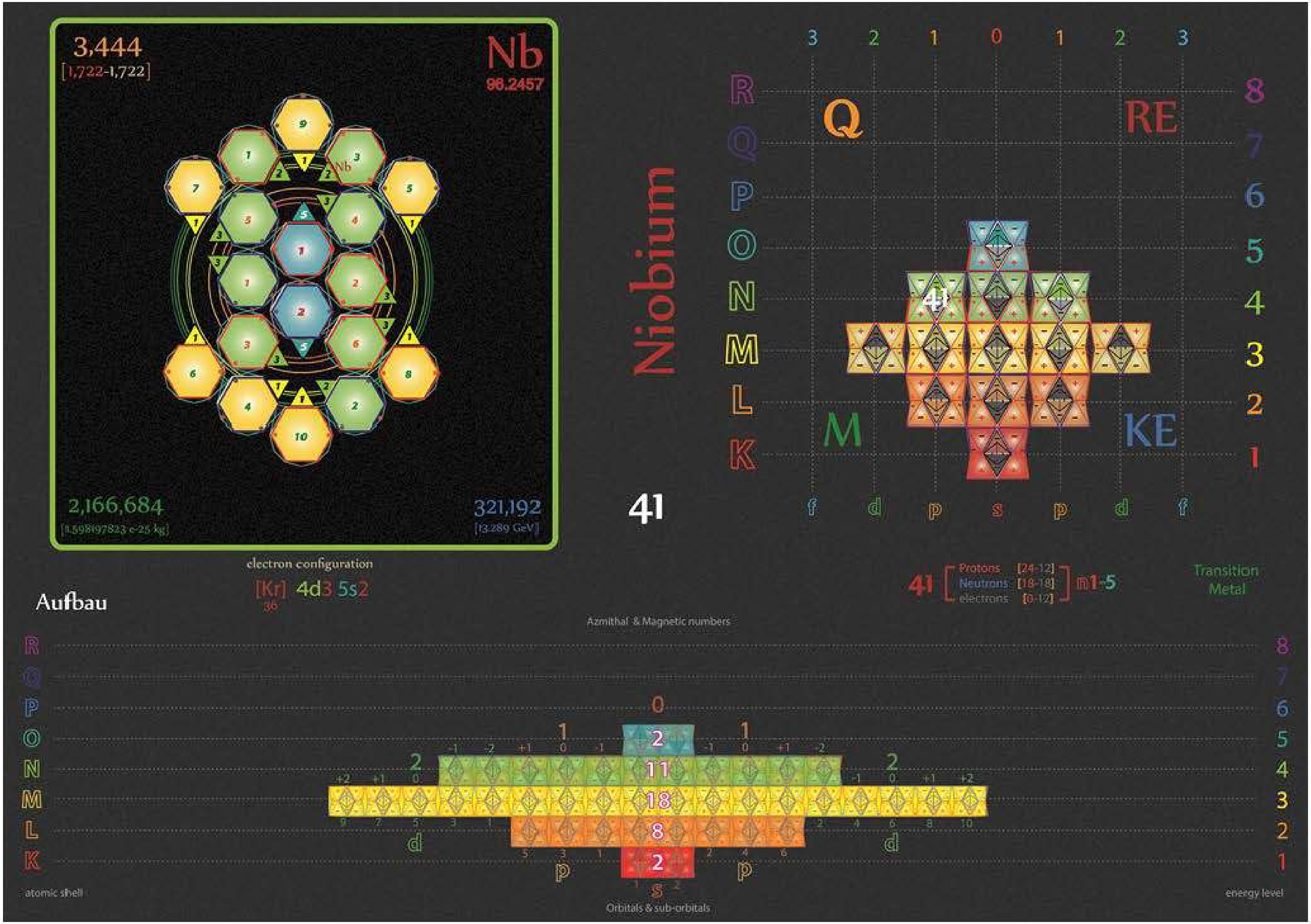
3,444
[1,722-1,722]
Nb
93.2457
2,166,684
[1.598197823 e-25 kg]
321,192
[13.289 GeV]
Niobium
41
electron configuration
[Kr] 4d3 5s2
36
Aufbau
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
41 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
2
11
18
8
2
d
p
s
d
p
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.41 - Niobium atom

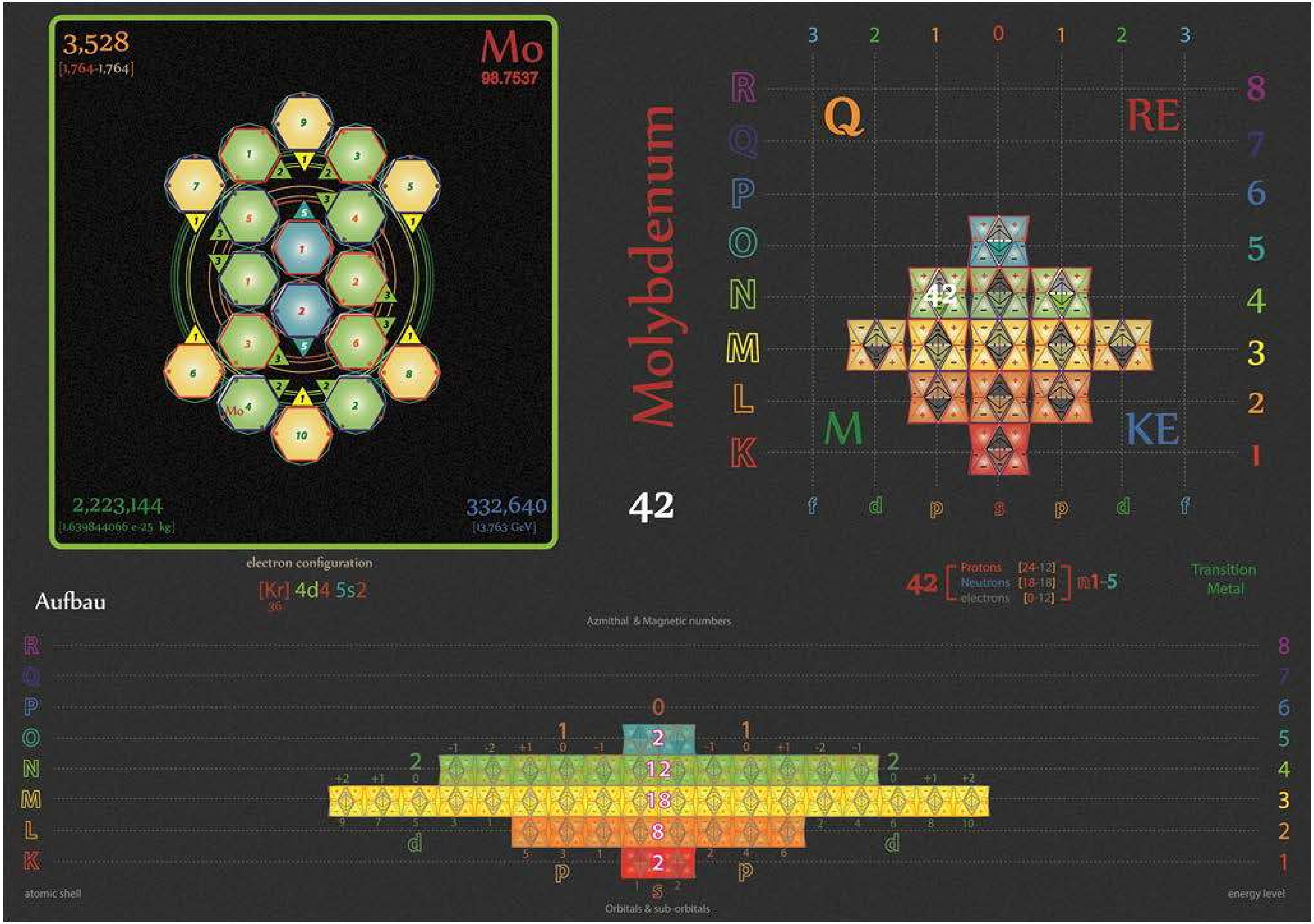
3,528
[1,764-1,764]
Mo
98.7537
2,223,144
[1.639844066 e-25 kg]
332,640
[13.763 GeV]
Molybdenum
42
electron configuration
[Kr] 4d4 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
42
f d p s p d f
42 Protons [24-12] Neutrons [18-18] electrons [0-12] n1-5
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
-1 -2 +1 0 -1 -1 0 +1 -2 -1
+2 +1 0 0 +1 +2
2
12
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.42 - Molybdenum atom

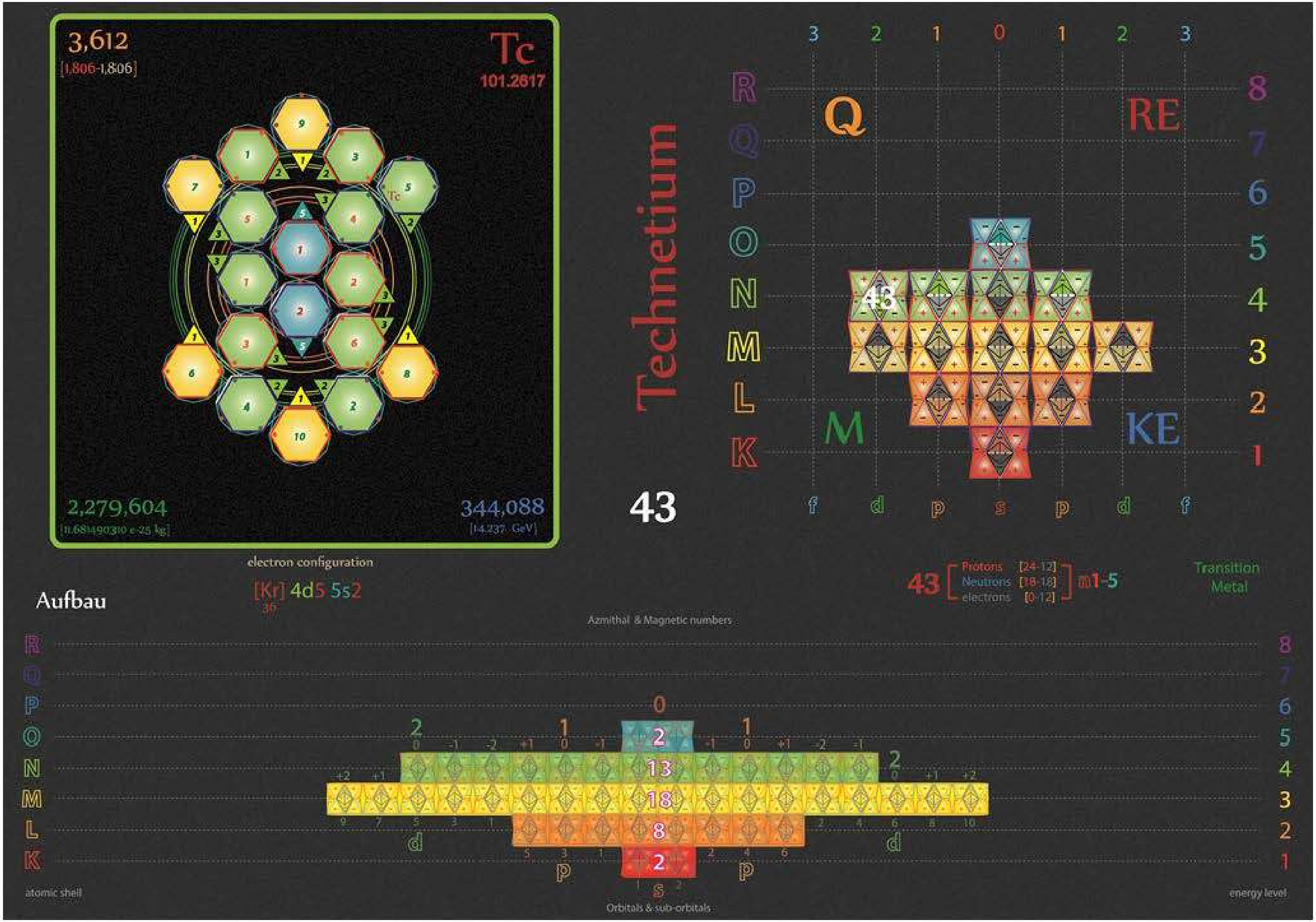
3,612
[1,806-1,806]
Tc
101.2617
2,279,604
[11.6814903I0 e-25 kg]
344,088
[14.237 GeV]
Technetium
43
electron configuration
[Kr] 4d5 5s2
36
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
43
43 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2
13
18
8
2
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.43 - Technetium atom

Copyright ABRAHAM [2008] - All rights reserved

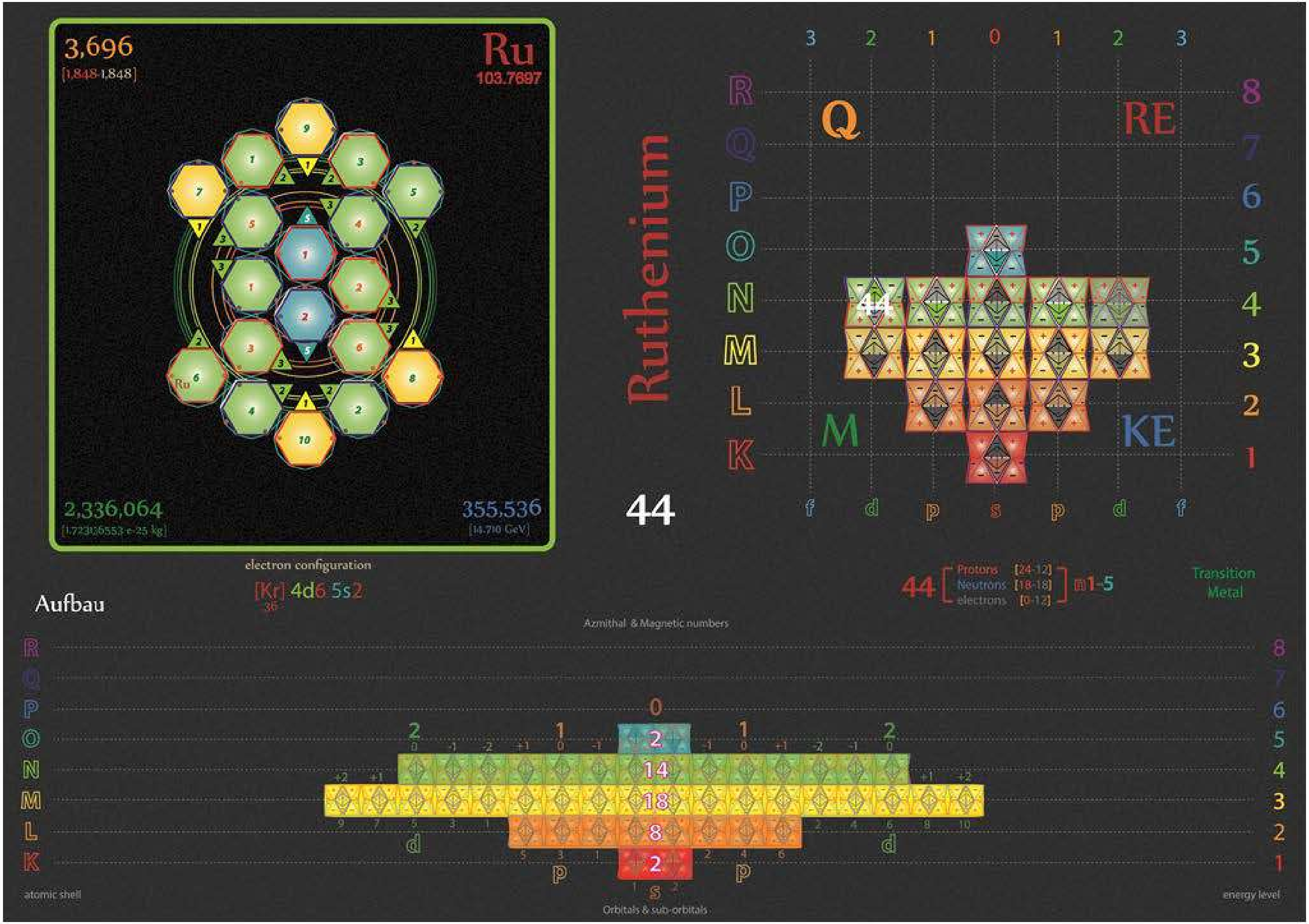
3,696
[1,848-1,848]
Ru
103.7697
2,336,064
[1.7231.6553 e-25 kg]
355,536
[14.710 GeV]
Ruthenium
44
electron configuration
[Kr] 4d6 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
44
f d p s p d f
44 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5
Transition Metal
Azmithal & Magnetic numbers
0
2 1 2 1 2
14
18
8
2
d p s p d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.44 - Ruthenium atom

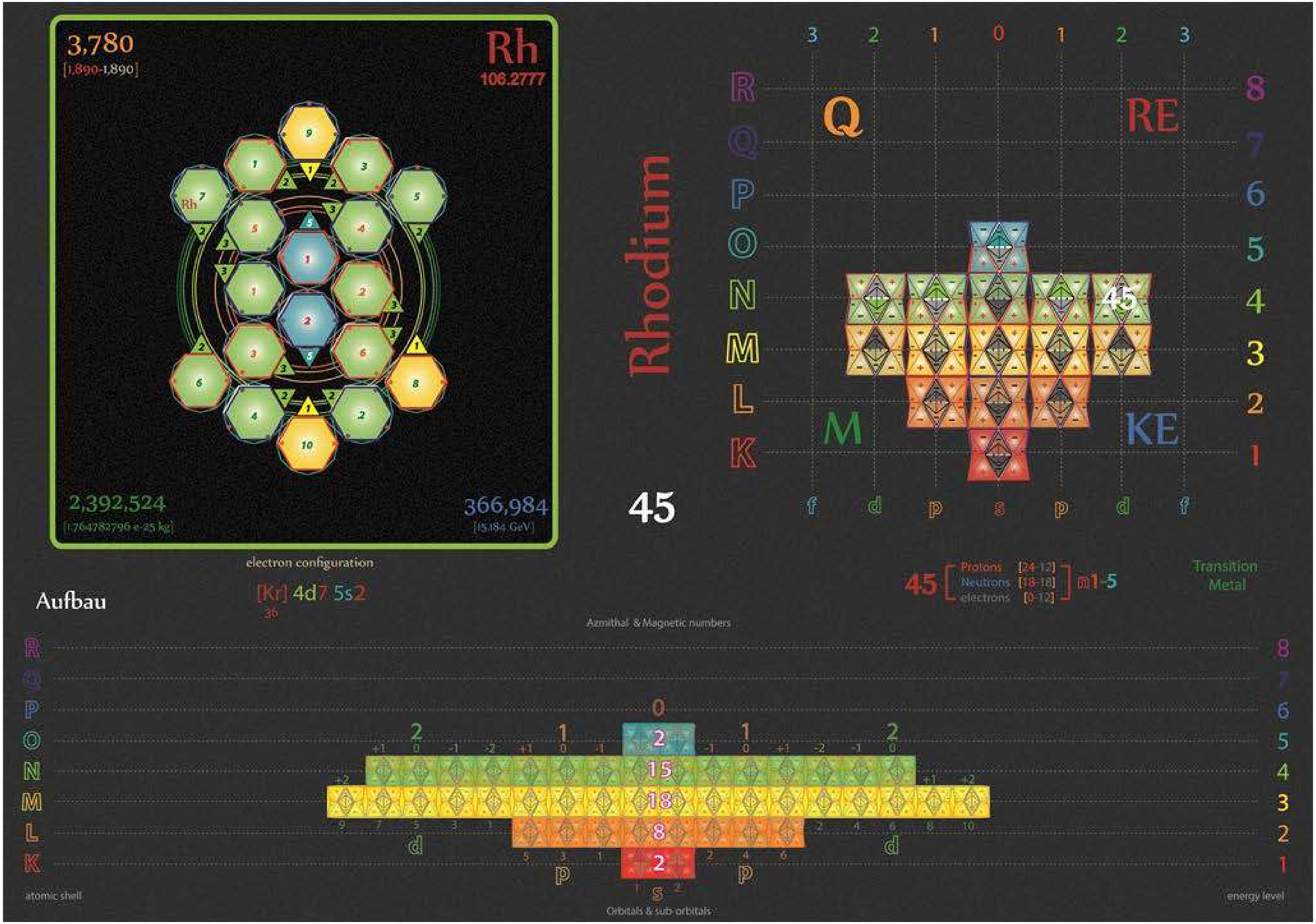
3,780
[1,890-1,890]
Rh
106.2777
2,392,524
[1.764782796 e-25 kg]
366,984
[15.184 GeV]
Rhodium
45
electron configuration
[Kr] 4d7 5s2
36
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
45 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5
Transition Metal
Azmithal & Magnetic numbers
2
15
18
8
2
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.45 - Rhodium atom

Copyright ABRAHAM [2008] - All rights reserved

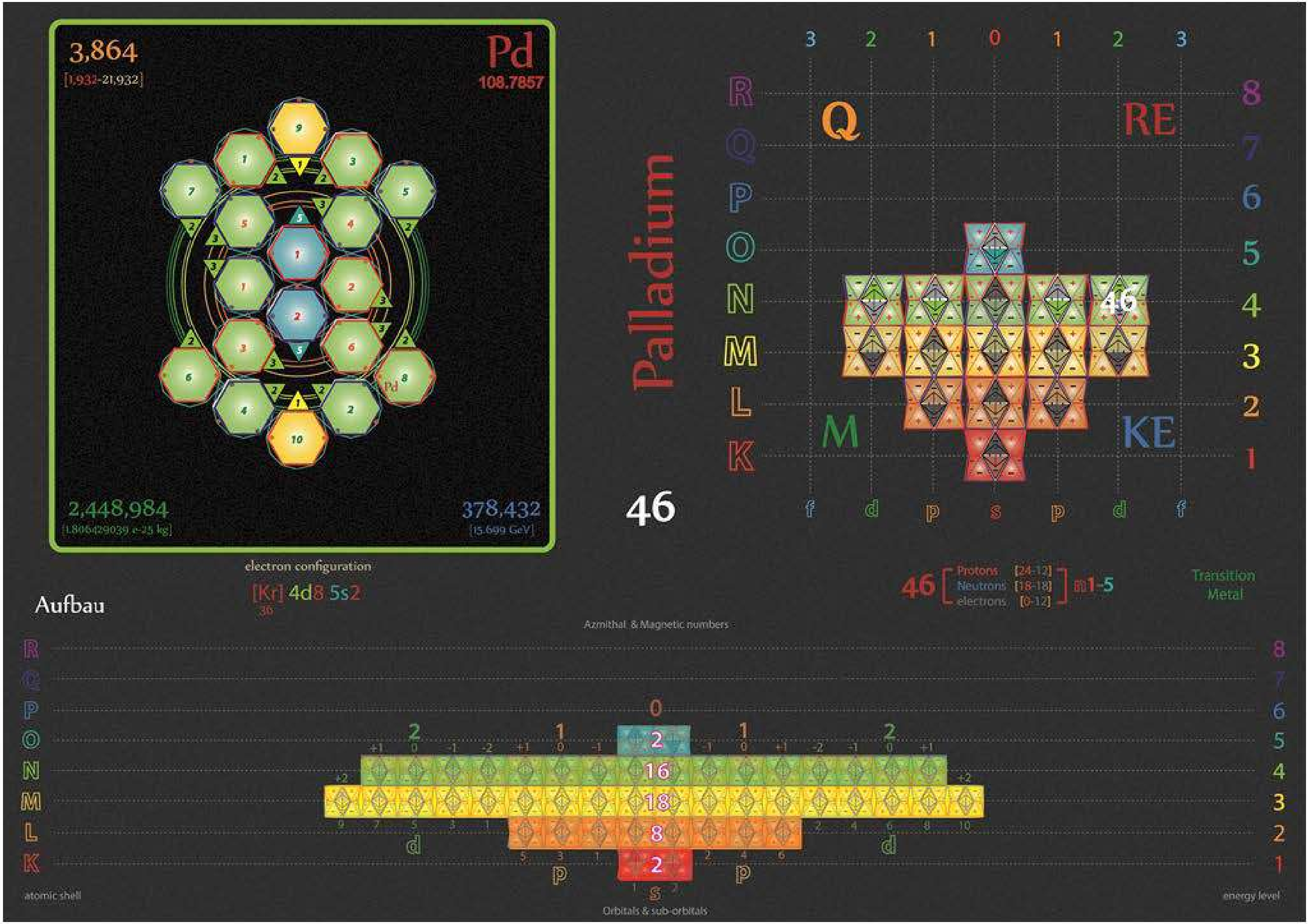
3,864
[1,932-21,932]
Pd
108.7857
2,448,984
[1.806429039 e-25 kg]
378,432
[15.699 GeV]
Palladium
46
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d8 5s2
36
Aufbau
46 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
16
18
8
2
d p s p d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.46 - Palladium atom

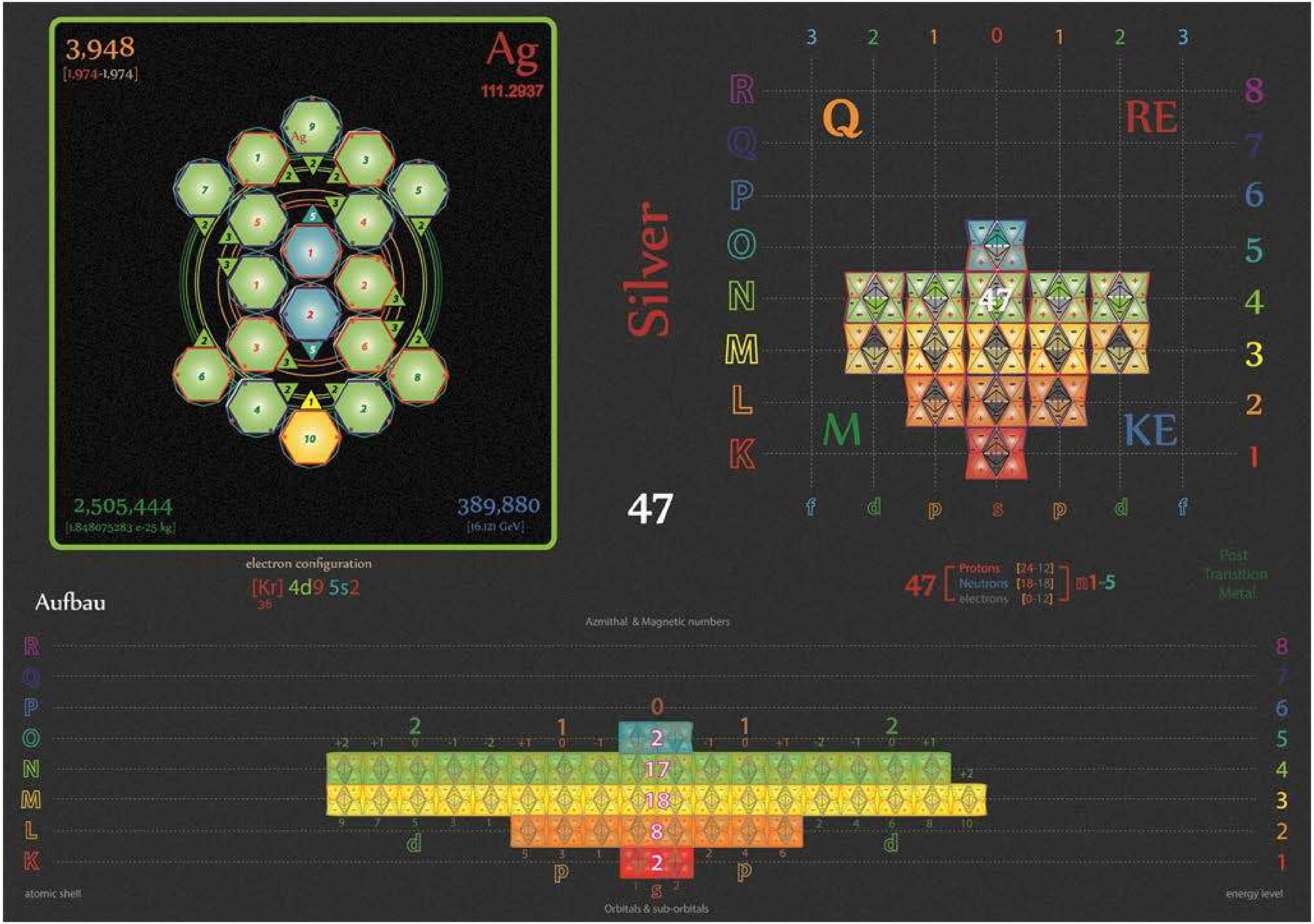
3,948
[1,974-1,974]
Ag
111.2937
2,505,444
[1.848075283 e-25 kg]
389,880
[16.121 GeV]
Silver
47
electron configuration
[Kr] 4d9 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
47 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Post Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
17
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.47 - Silver atom

Copyright ABRAHAM [2008] - All rights reserved

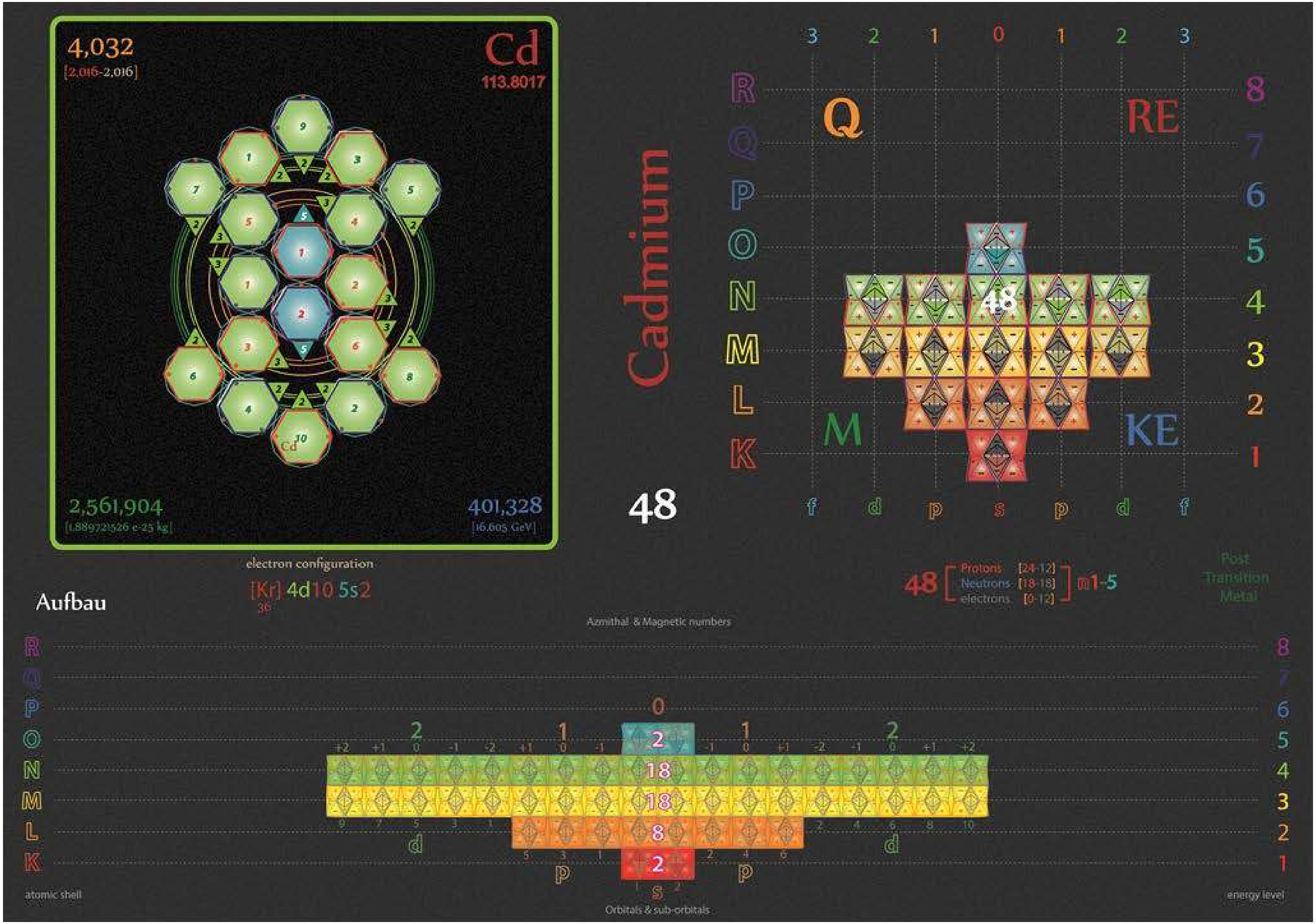
4,032
[2,016-2,016]
Cd
113.8017
Cd
10
2,561,904
[1.889721526 e-25 kg]
401,328
[16.605 GeV]
Cadmium
48
electron configuration
[Kr] 4d10 5s2
36
Aufbau
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
48
48
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-5
Post Transition Metal
Azmithal & Magnetic numbers
0
2
1
1
2
+2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
+1
+2
2
18
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.48 - Cadmium atom

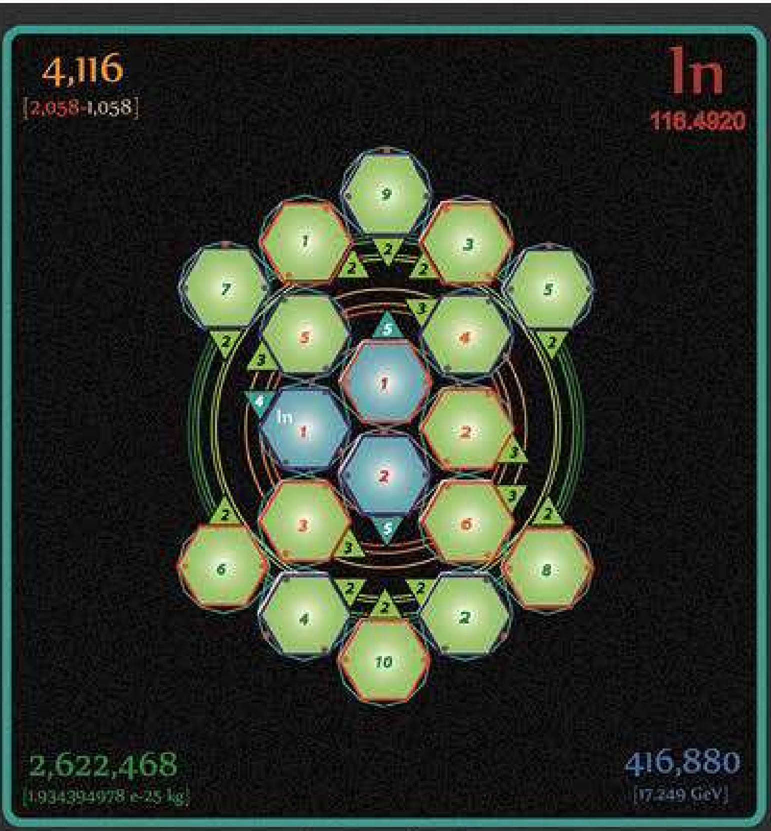
4,116
[2,058-1,058]
In
116.4920
In
1
2
3
4
5
6
7
8
9
10
2,622,468
[1.934394978 e-25 kg]
416,880
[17.249 GeV]

Indium

49

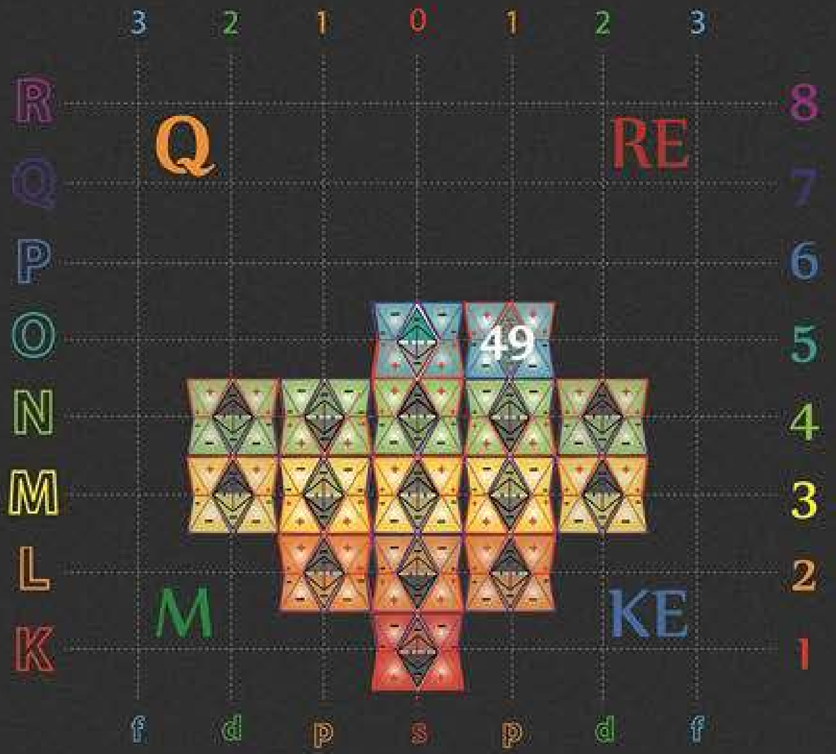
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
49
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

[Kr] 4d10 5s2 5p1
36

Aufbau

49 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5

Azmithal & Magnetic numbers

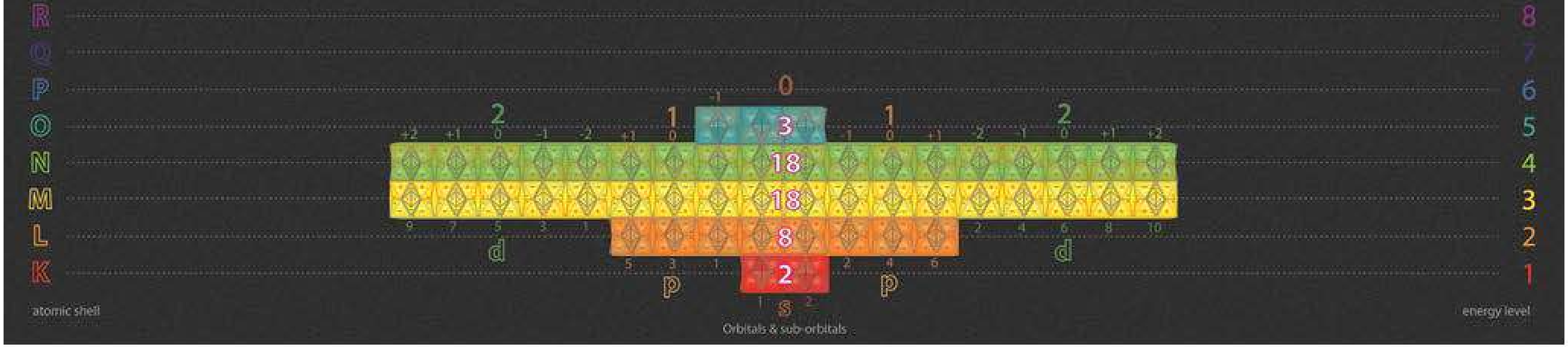
R
Q
P
O
N
M
L
K
0
-1
2
1
1
2
+2
+1
0
-1
-2
+1
0
-1
0
+1
-2
-1
0
+1
+2
3
18
18
8
2
9
7
5
3
1
2
4
6
8
10
d
5
3
1
2
4
6
p
p
d
1
2
s
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.49 - Indium atom

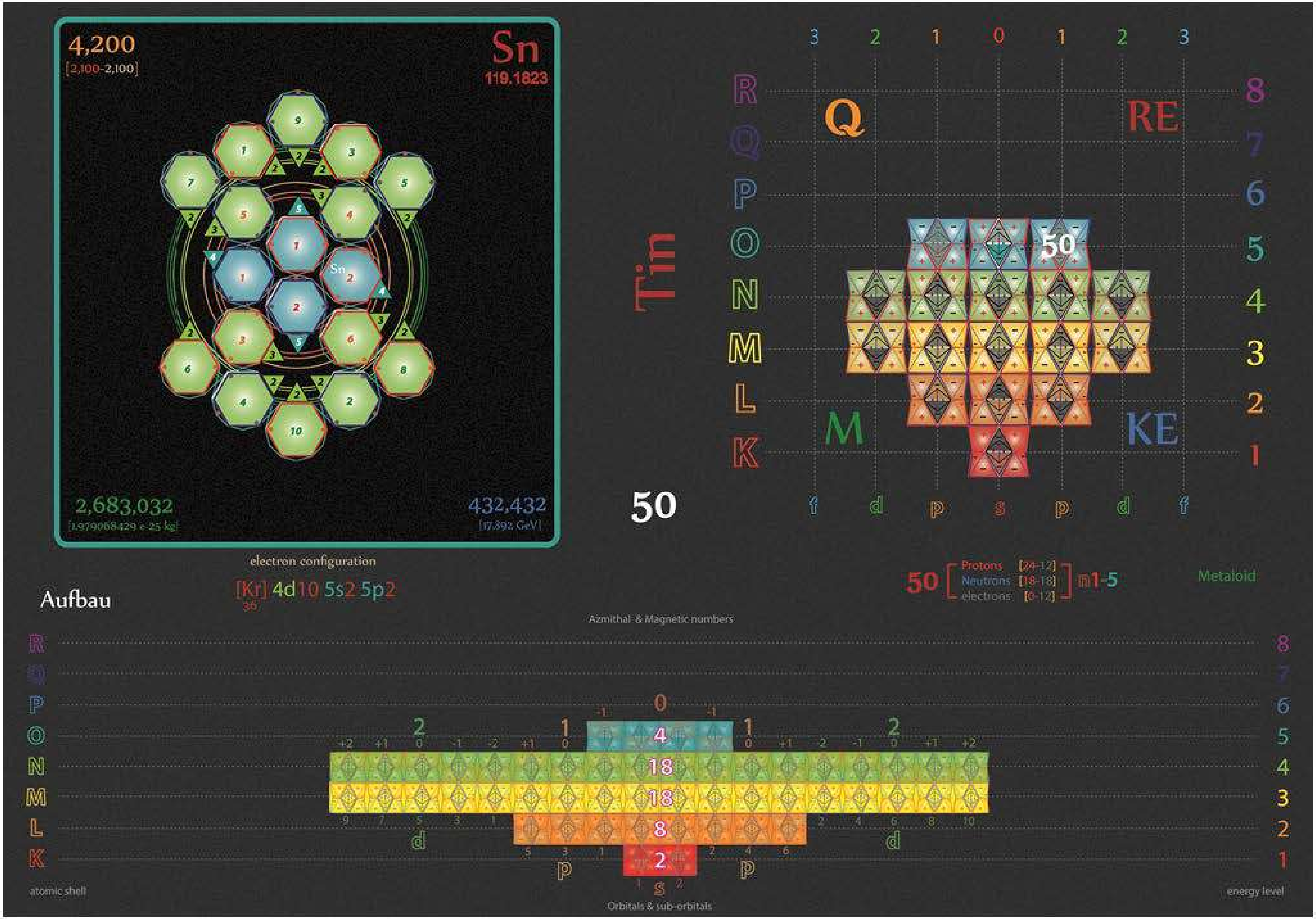
4,200
[2,100-2,100]
Sn
119.1823
2,683,032
[1.979068429 e-25 kg]
432,432
[17.392 GeV]
electron configuration
[Kr] 4d10 5s2 5p2
36
Aufbau
Tin
50
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
50 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Azmithal & Magnetic numbers
4
18
18
8
2
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.50 - Tin atom

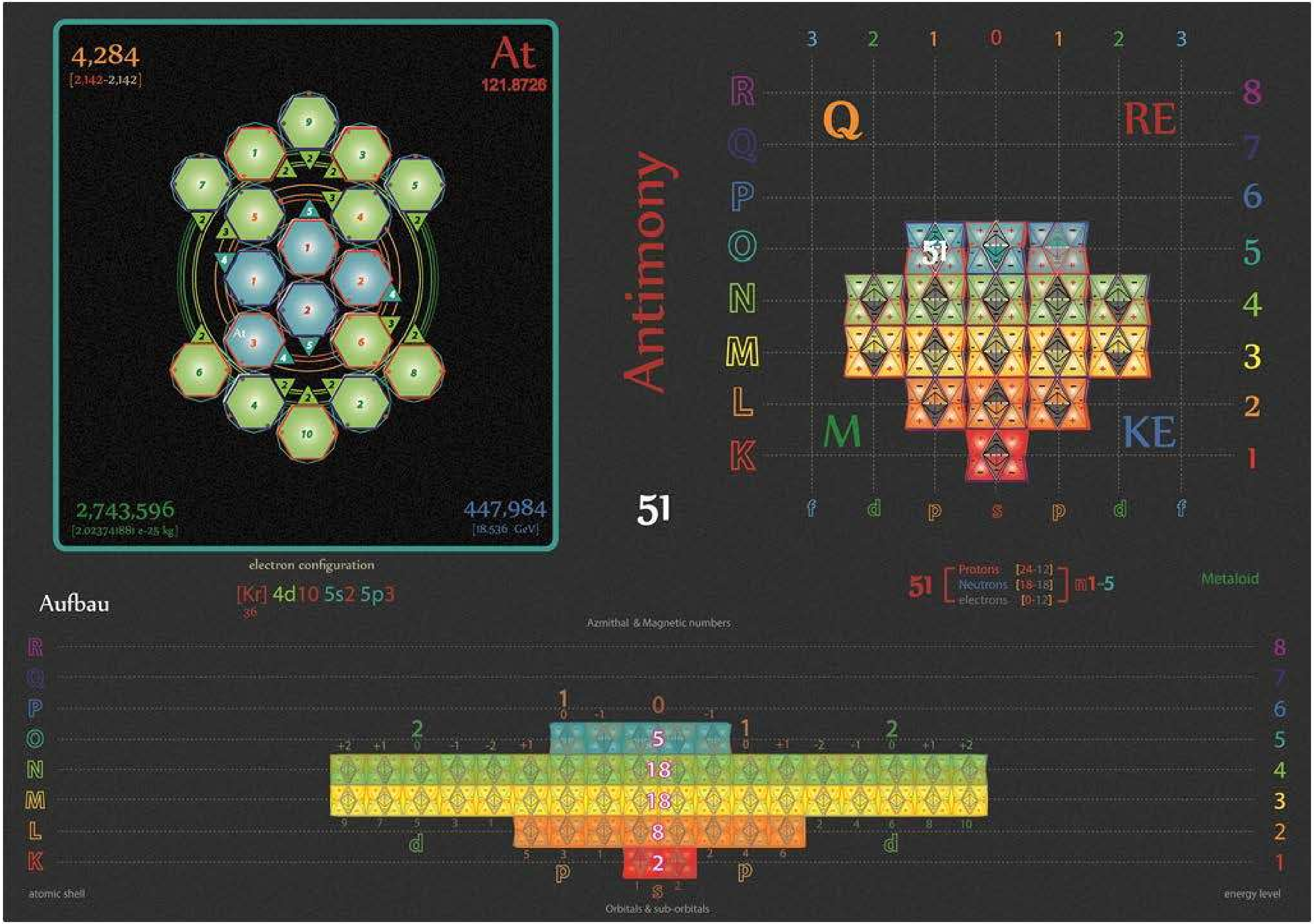
4,284
[2,142-2,142]
At
121.8726
2,743,596
[2.02374188i e-25 kg]
447,984
[18.536 GeV]
Antimony
51
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d10 5s2 5p3
36
Aufbau
51 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Azmithal & Magnetic numbers
5
18
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.51 - Antimony atom

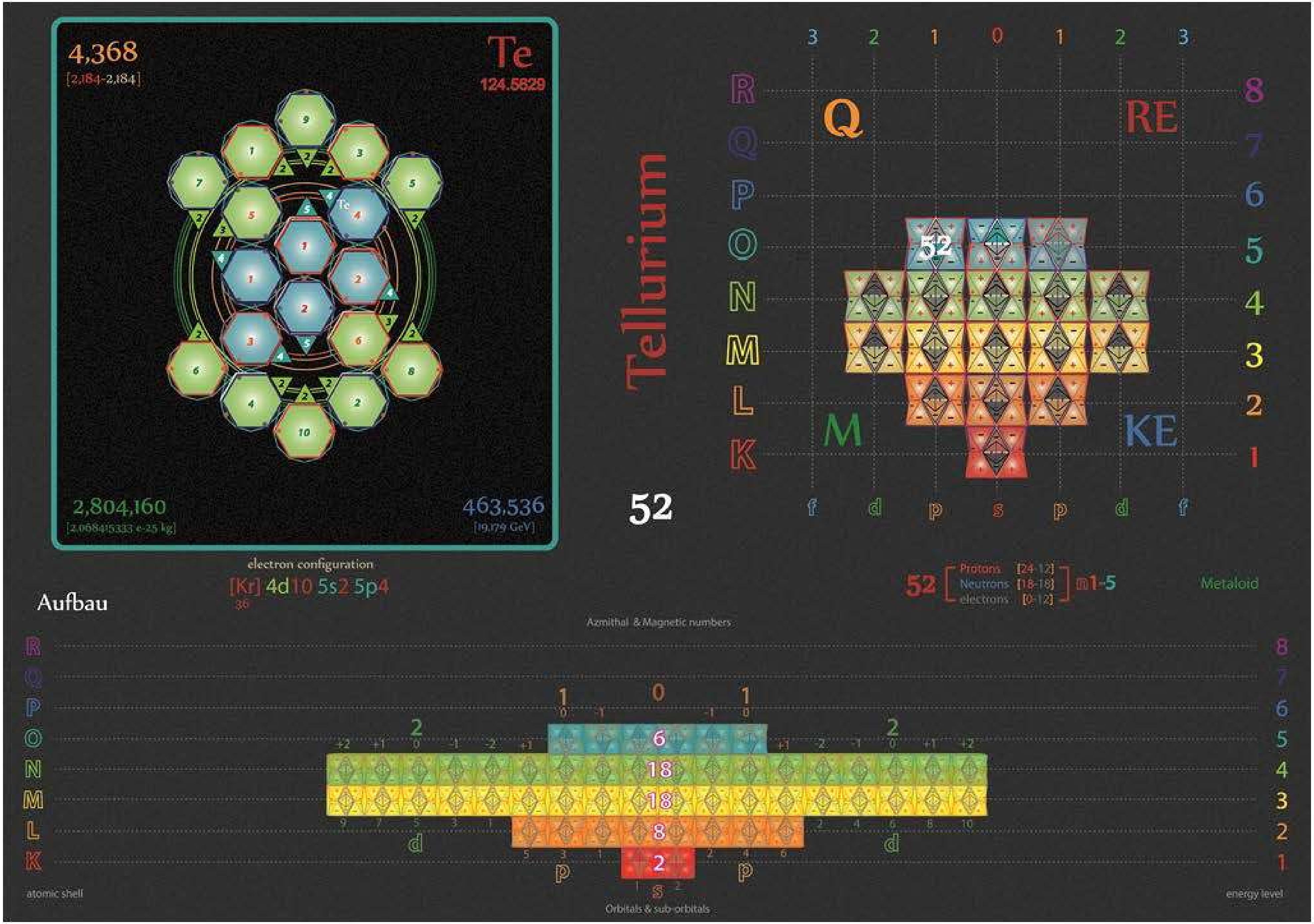
4,368
[2,184-2,184]
Te
124.5829
2,804,160
[2.0684!5333 e-25 kg]
463,536
[19,179 GeV]
Tellurium
52
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d10 5s2 5p4
36
52 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Aufbau
Azmithal & Magnetic numbers
6
18
18
8
2
d p s p d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.52 - Tellurium atom

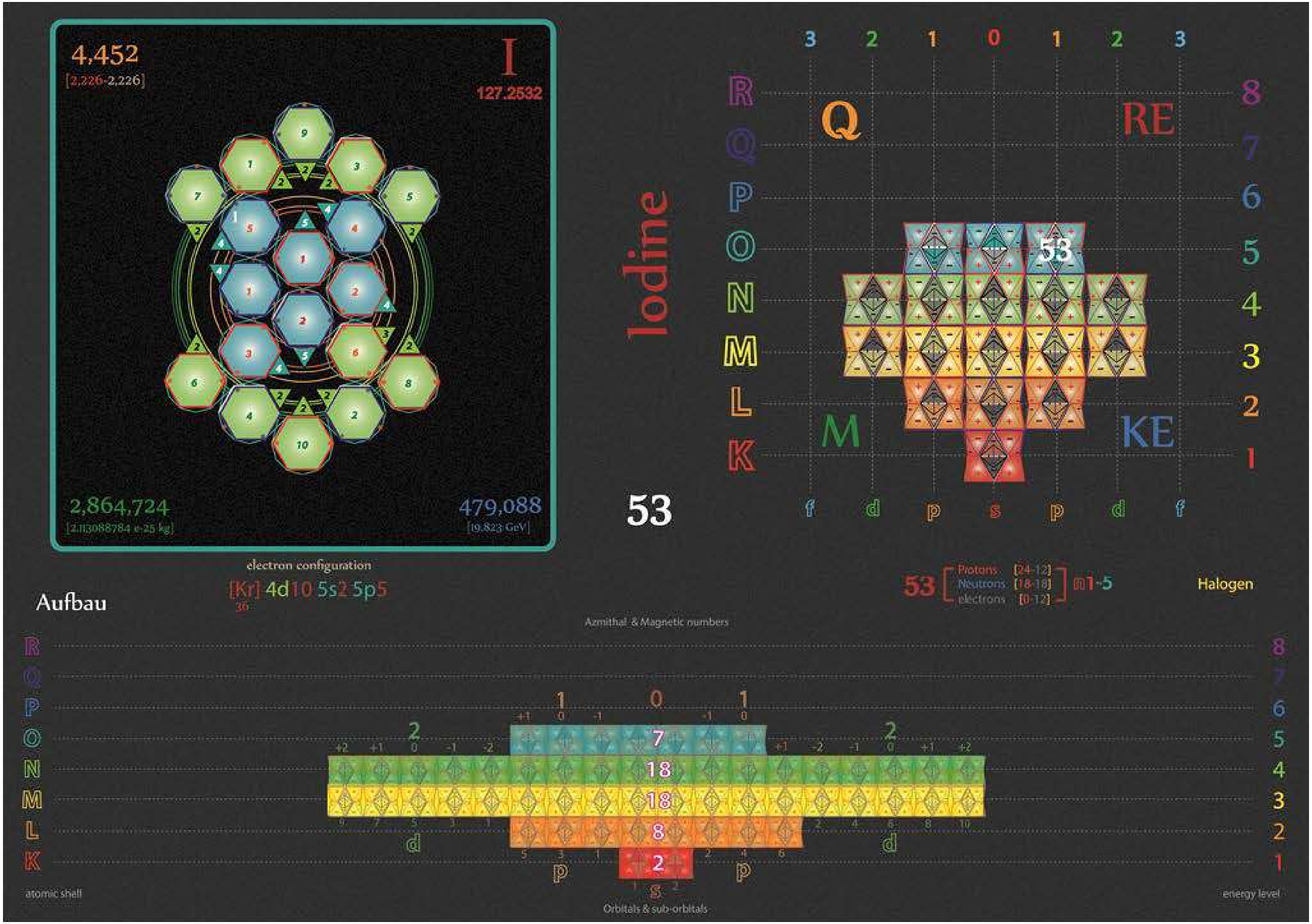
4,452
[2,226-2,226]
I
127.2532
2,864,724
[2.113088784 e-25 kg]
479,088
[19.823 GeV]
Iodine
53
electron configuration
[Kr] 4d10 5s2 5p5
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
53
f d p s p d f
53 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Halogen
Azmithal & Magnetic numbers
7
18
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.53 - Iodine atom

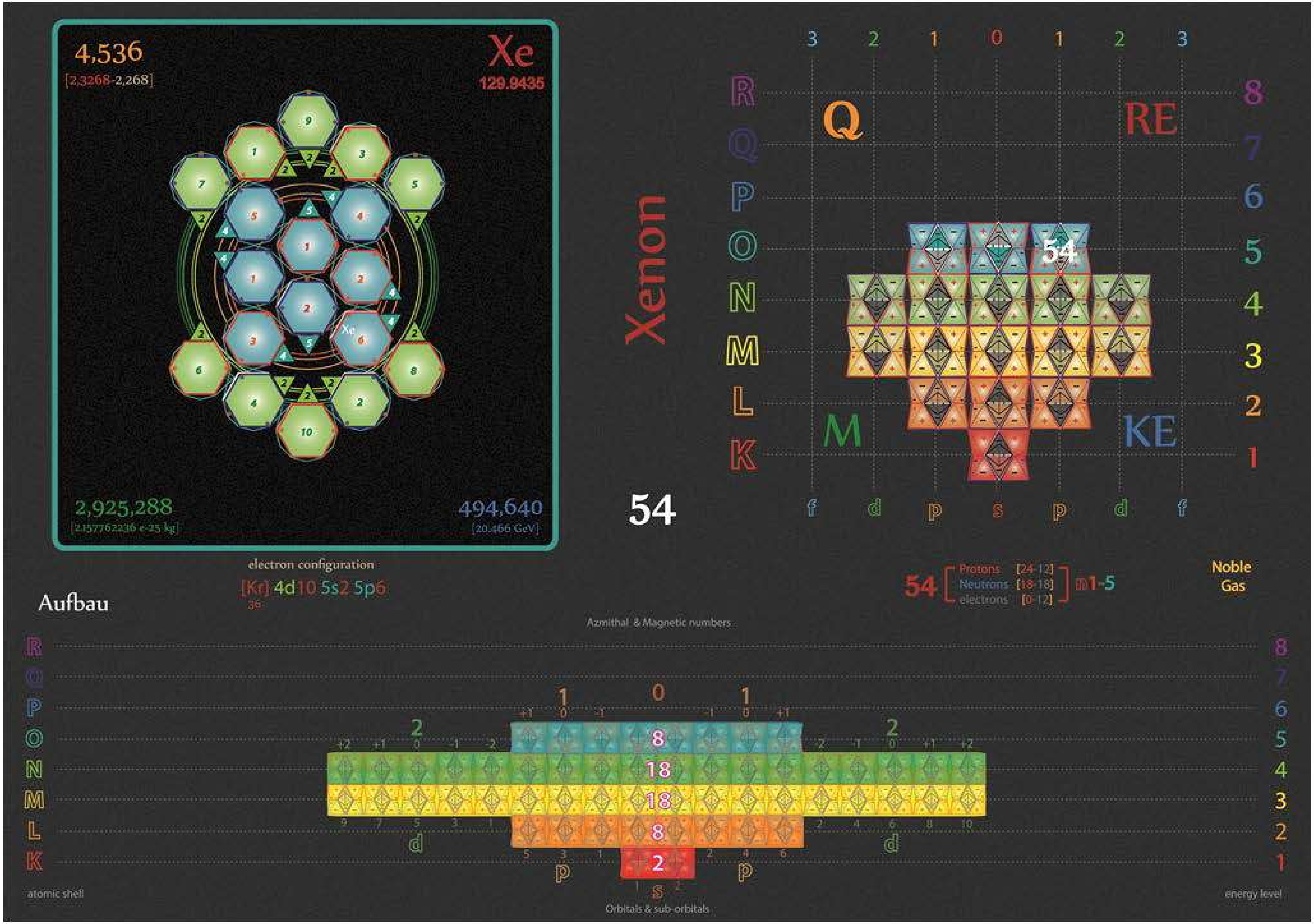
4,536
[2,3268-2,268]
Xe
129.9435
2,925,288
[2.157762236 e-25 kg]
494,640
[20.466 GeV]
Xenon
54
electron configuration
[Kr] 4d10 5s2 5p6
36
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
54
54 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Noble Gas
Azmithal & Magnetic numbers
8
18
18
8
2
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.54 - Xenon atom

Copyright ABRAHAM [2008] - All rights reserved

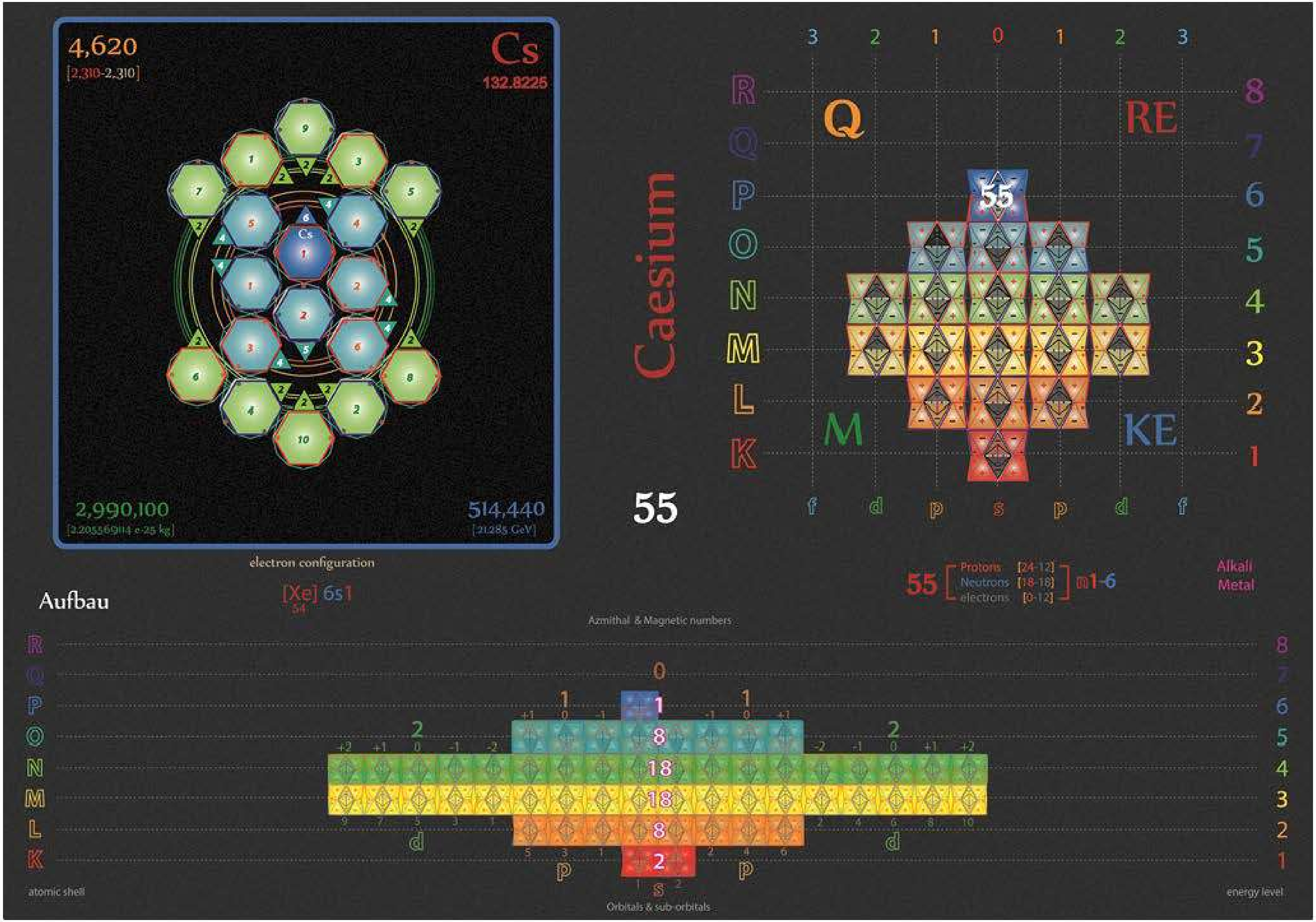

4,620
[2,310-2,310]
Cs
132.8225
Cs
2,990,100
[2.205569114 e-25 kg]
514,440
[21,285 GeV]
Caesium
55
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
55
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
[Xe] 6s1
54
Aufbau
55
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Alkali
Metal
Azmithal & Magnetic numbers
0
1
1
2
2
+1
0
-1
-1
0
+1
+2
+1
0
-1
-2
-2
-1
0
+1
+2
1
8
18
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.55 - Caesium atom

Copyright ABRAHAM [2008] - All rights reserved

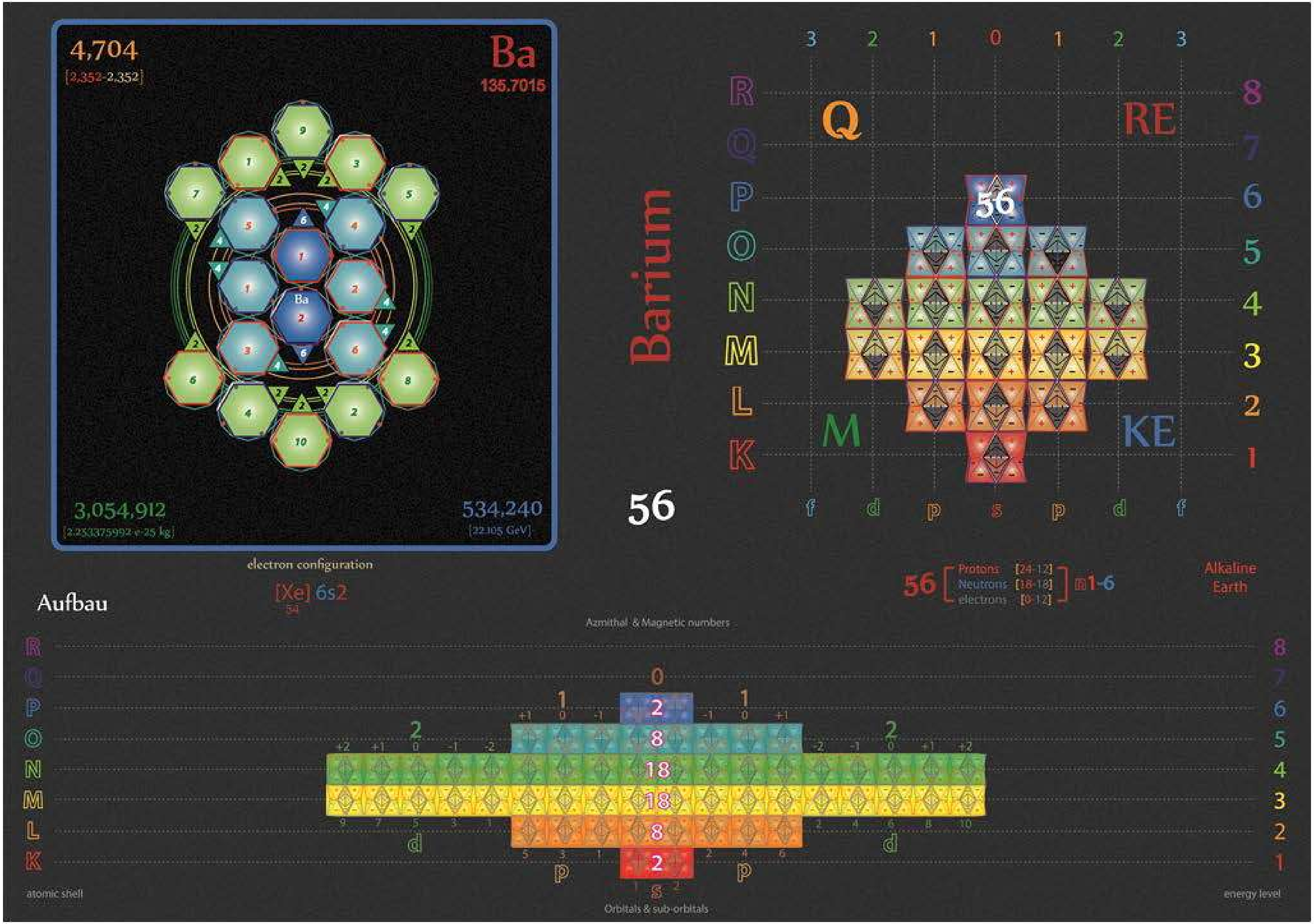
4,704
[2,352-2,352]
Ba
135.7015
Ba
3,054,912
[2.233375992 e-25 kg]
534,240
[22.105 GeV]
electron configuration
[Xe] 6s2
54
Barium
56
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
56
8 7 6 5 4 3 2 1
f d p s p d f
56
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-6
Alkaline Earth
Aufbau
Azmithal & Magnetic numbers
0
1 1
2 2
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -2 -1 0 +1 +2
2
8
18
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.56 - Barium atom

Copyright ABRAHAM [2008] - All rights reserved

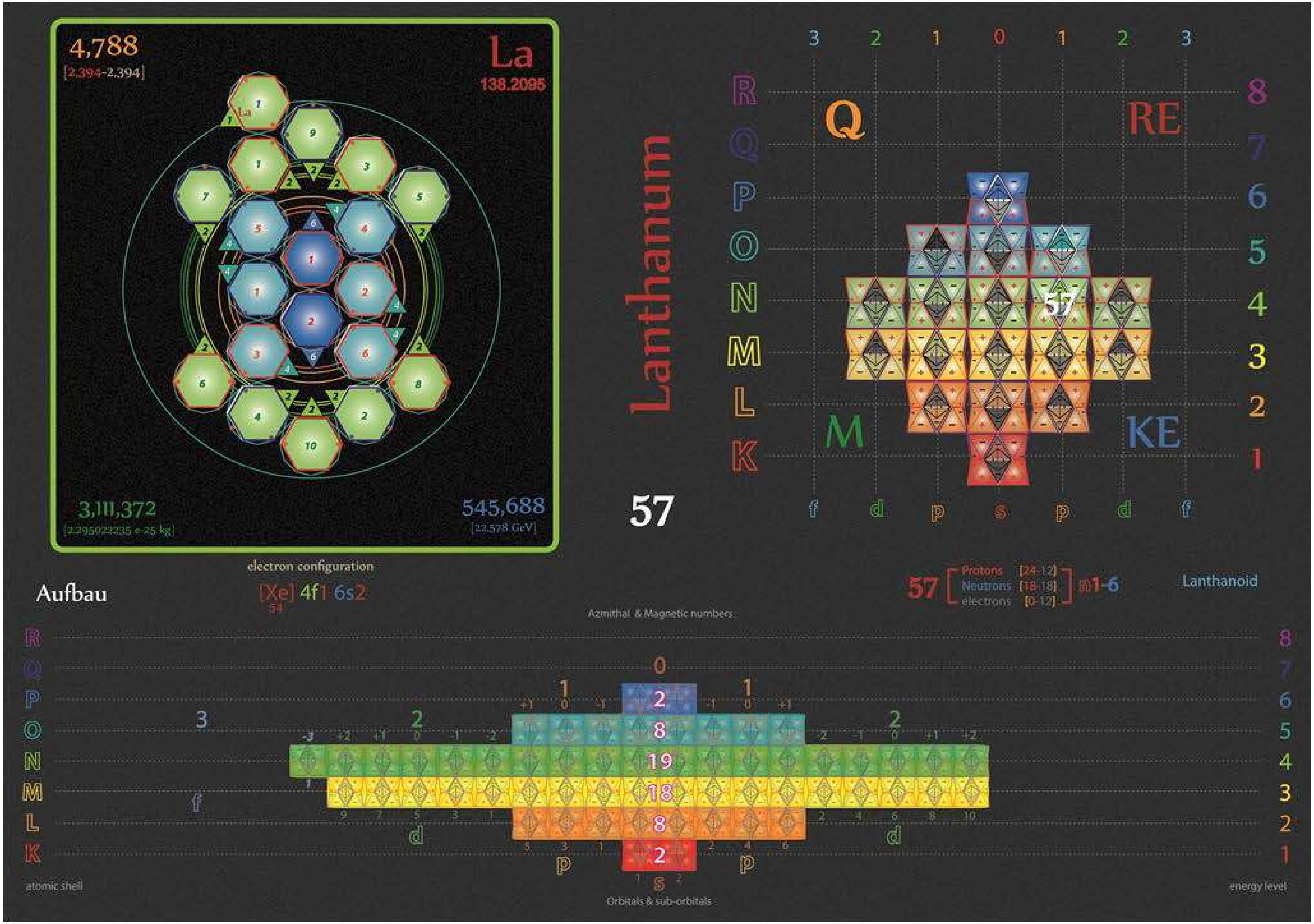
4,788
[2.394-2.394]
La
138.2095
3,111,372
[2.295022235 e-25 kg]
545,688
[22.578 GeV]
Lanthanum
57
R Q P O N M L K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
57
electron configuration
[Xe] 4f1 6s2
54
Aufbau
57 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2 2
3
+1 0 -1 -1 0 +1
-3 +2 +1 0 -1 -2 -2 -1 0 +1 +2
2
8
19
18
8
2
f
d d
p p
s
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.57 - Lanthanum atom

Copyright ABRAHAM [2008] - All rights reserved

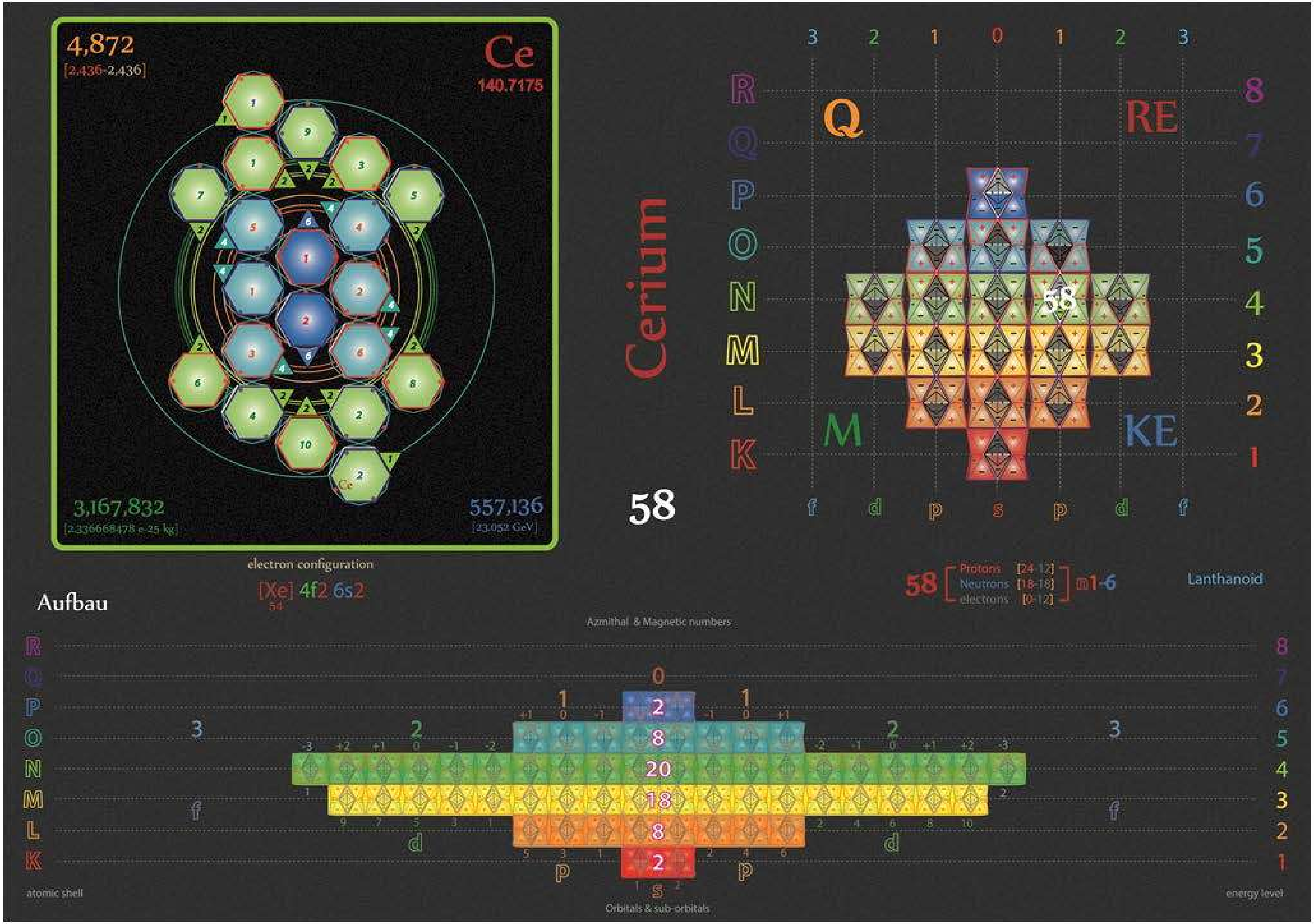
4,872
[2,436-2,436]
Ce
140.7175
3,167,832
[2.336668478 e-25 kg]
557,136
[23,052 GeV]
Cerium
58
electron configuration
[Xe] 4f2 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
58 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
atomic shell
energy level
Orbitals & sub-orbitals
2
8
20
18
8
2

Tetryonics 51.58 - Cerium atom

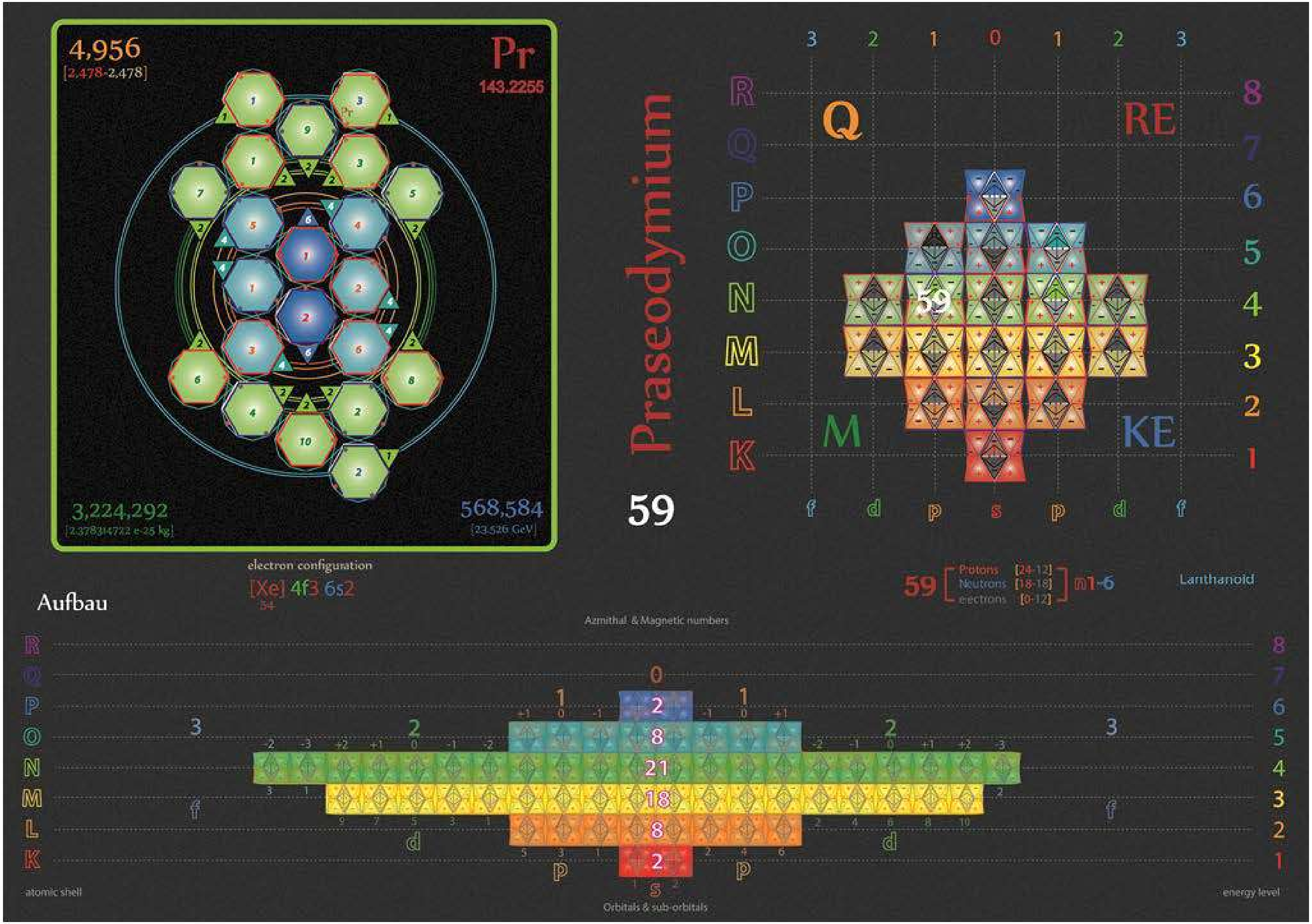
4,956
[2,478-2,478]
Pr
143.2255
3,224,292
[2.378314722 e-25 kg]
568,584
[23.526 GeV]
Praseodymium
59
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Xe] 4f3 6s2
54
59 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
2
8
21
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.59 - Praseodymium atom

Copyright ABRAHAM [2008] - All rights reserved

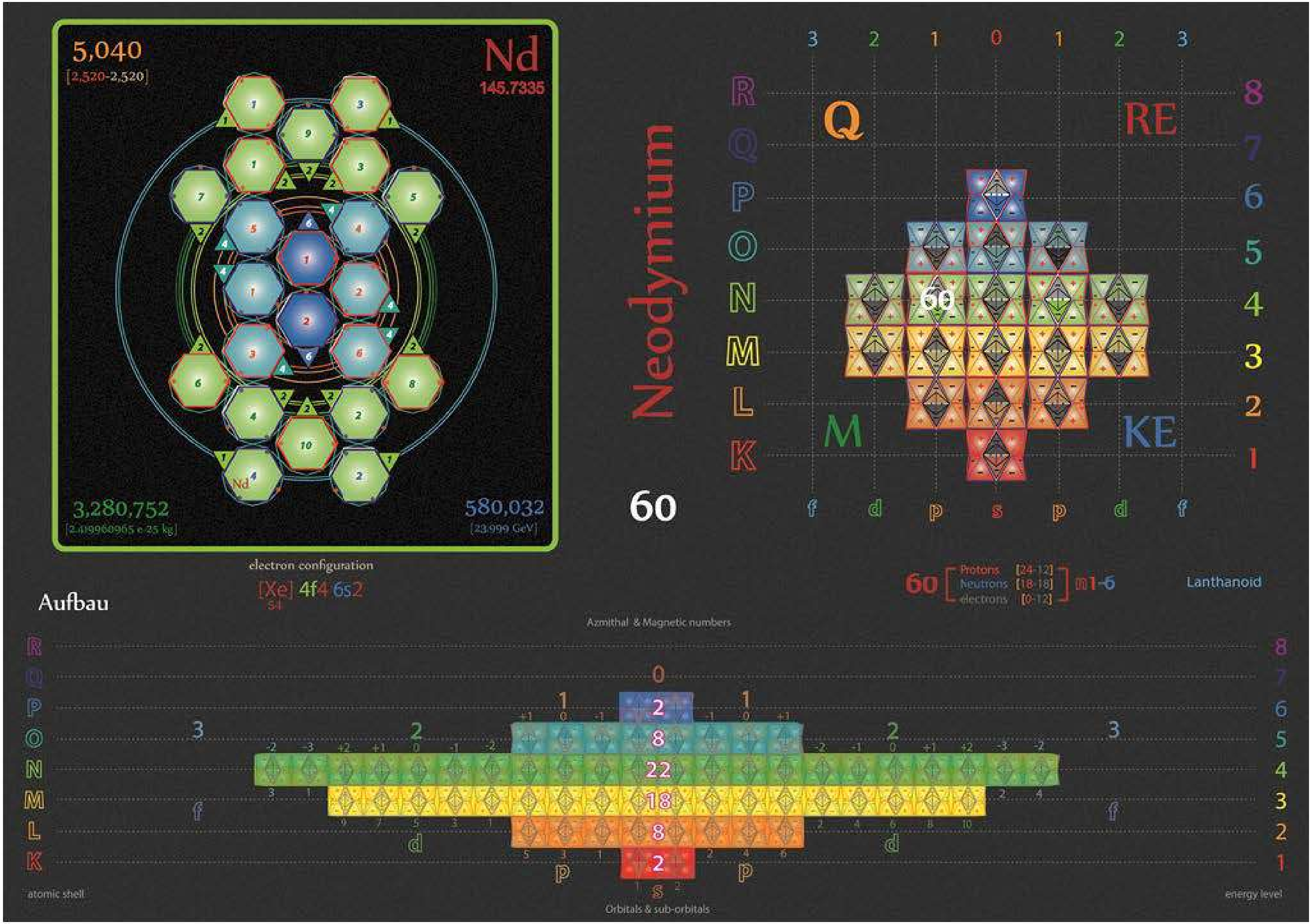
5,040
[2,520-2,520]
Nd
145.7335
Neodymium
60
3,280,752
[2.49960965 e-25 kg]
580,032
[23.999 GeV]
electron configuration
[Xe] 4f4 6s2
54
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
60
60 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
2
8
22
18
8
2
f
f
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.60 - Neodymium atom

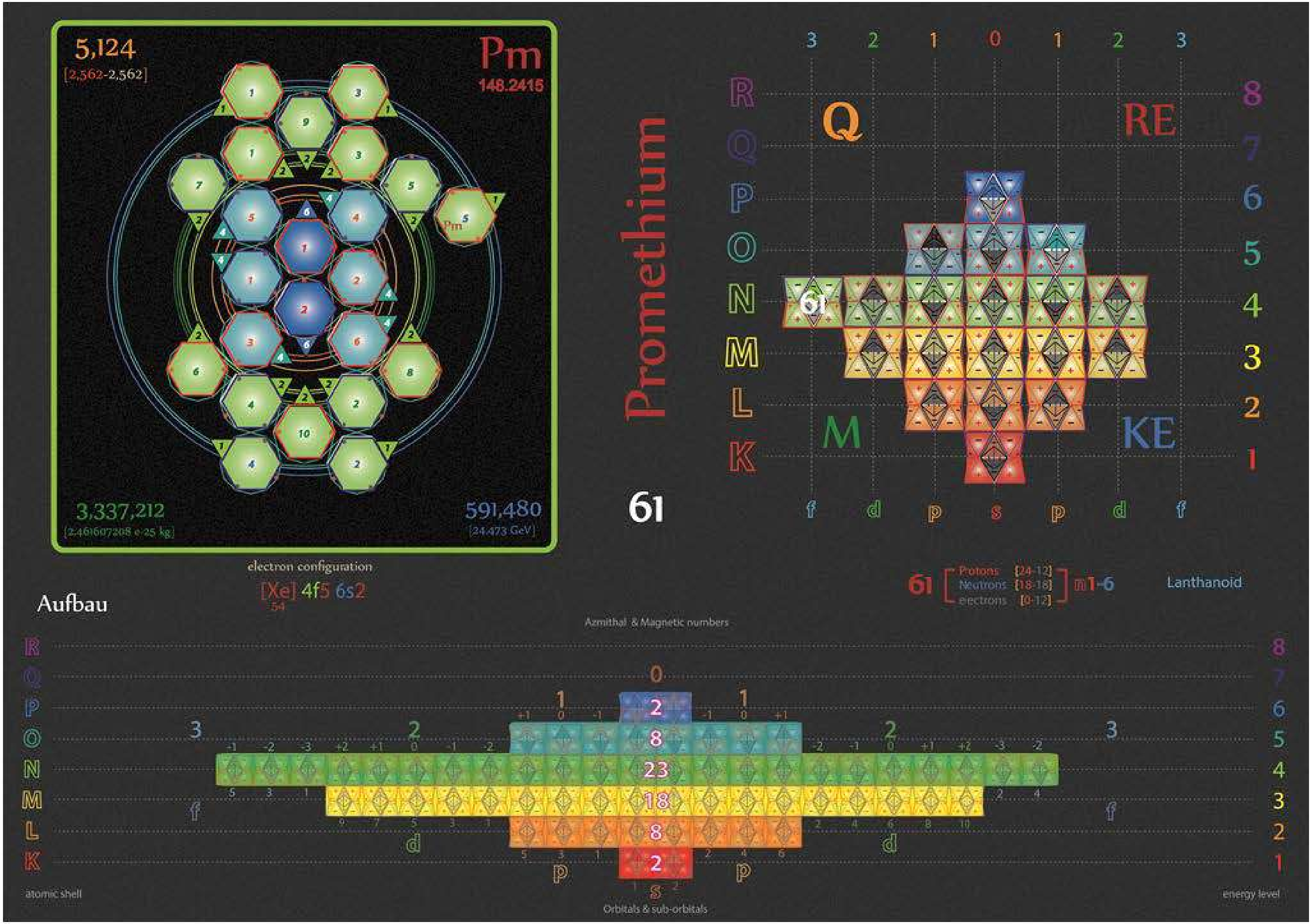

5,124
[2,562-2,562]
Pm
148.2415
3,337,212
[2.461607208 e-25 kg]
591,480
[24.473 GeV]
Promethium
61
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f5 6s2
54
61 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
8
23
18
8
2
f f
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.61 - Promethium atom

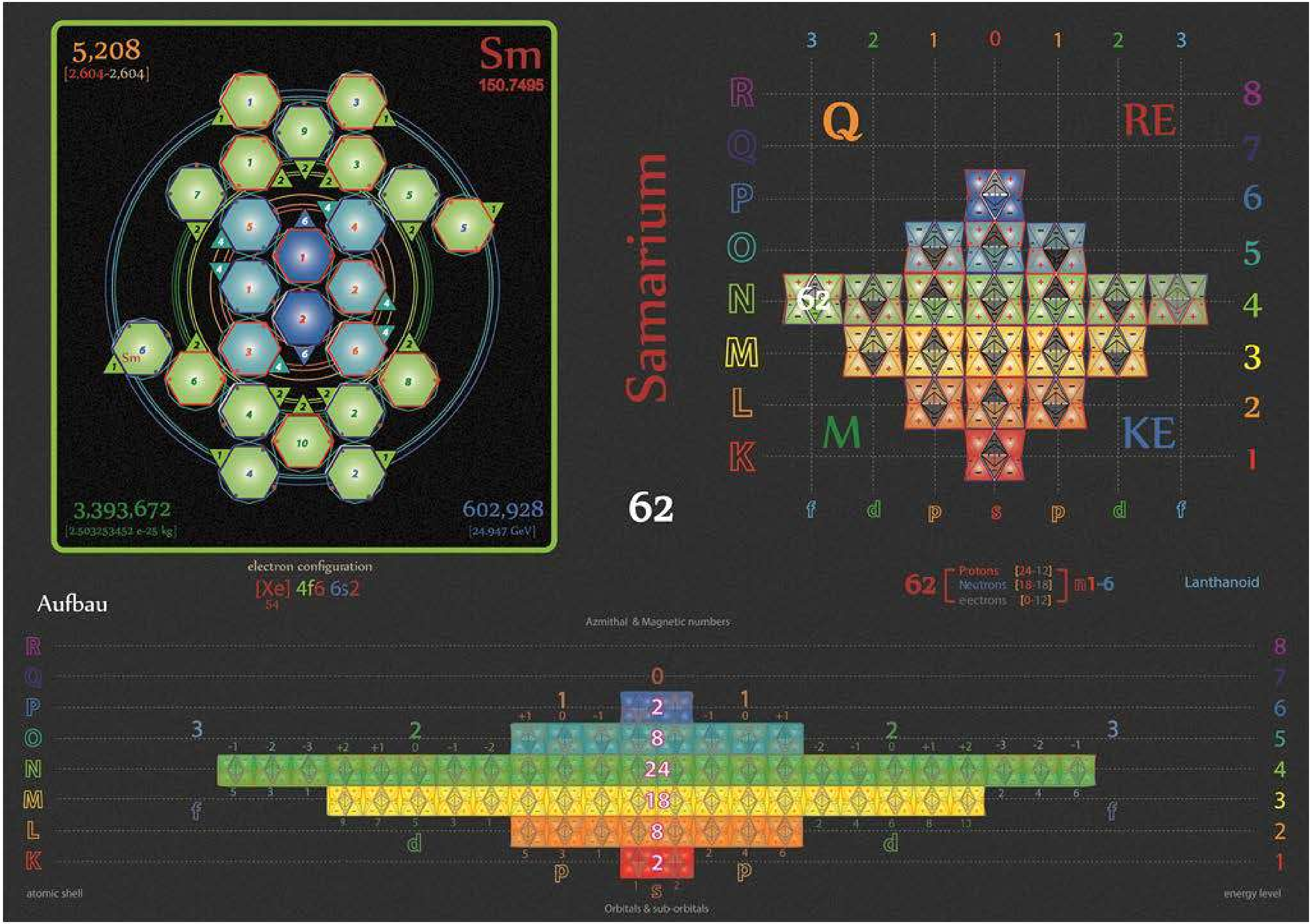
5,208
[2,604-2,604]
Sm
150.7495
3,393,672
[2.503233452 e-25 kg]
602,928
[24.947 GeV]
Samarium
62
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f6 6s2
54
62
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
2
8
24
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.62 - Samarium atom

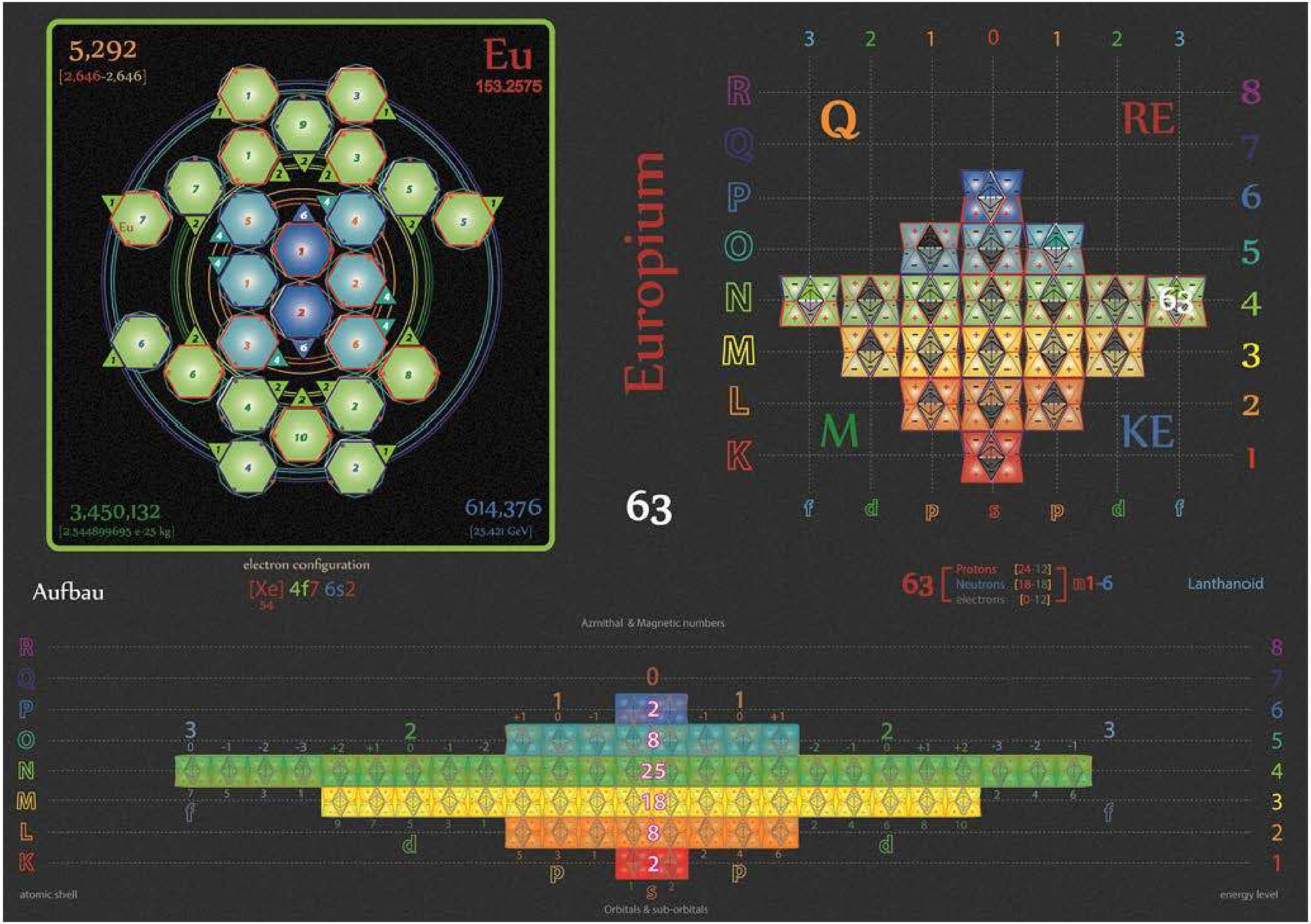
5,292
[2,646-2,646]
Eu
153.2575
Europium
63
3,450,132
[2.544899695 e-25 kg]
614,376
[25.421 GeV]
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Xe] 4f7 6s2
54
63
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Azimuthal & Magnetic numbers
2
8
25
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.63 - Europium atom

Copyright ABRAHAM [2008] - All rights reserved

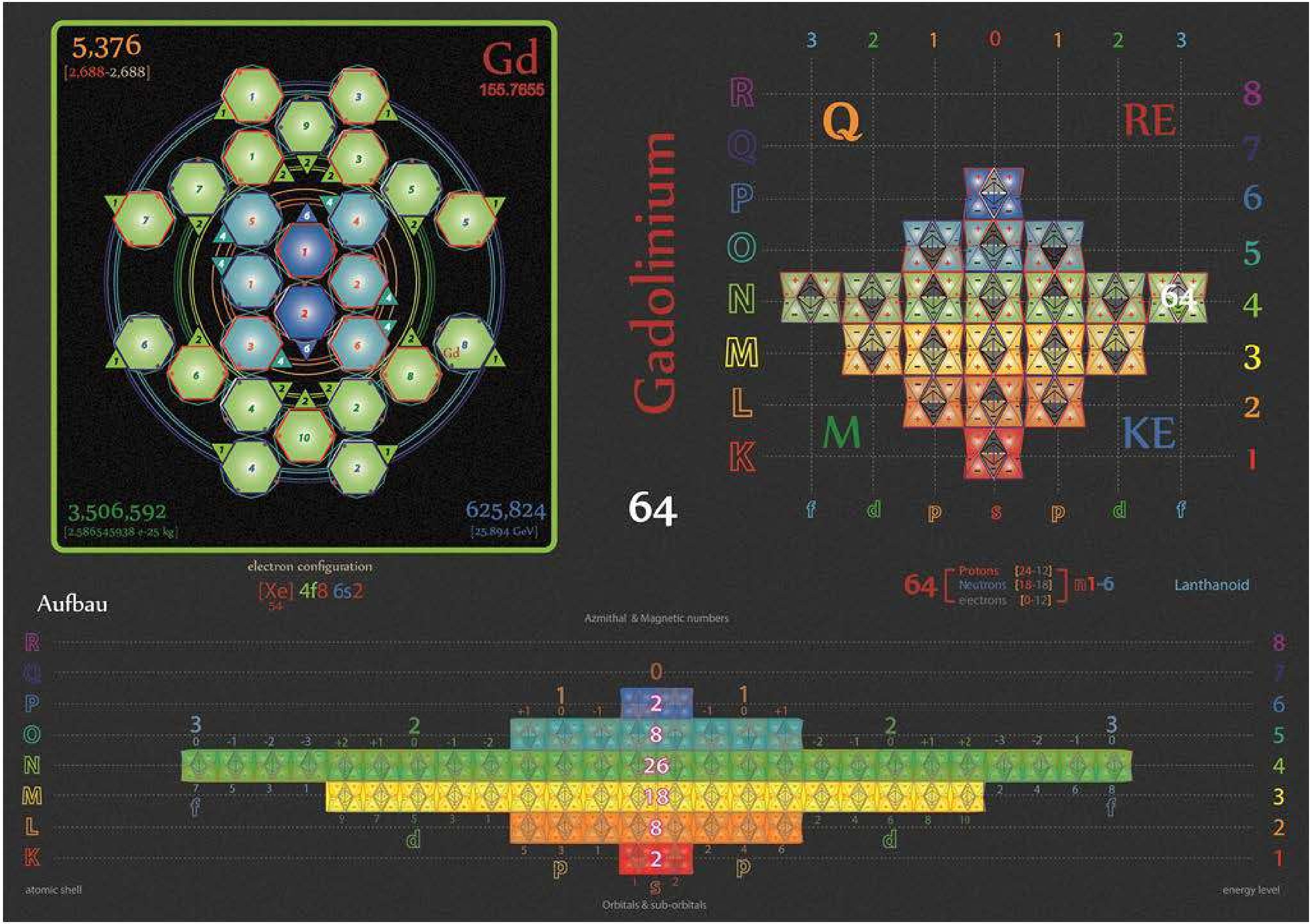
5,376
[2,688-2,688]
Gd
155.7655
3,506,592
[2.586545938 e-25 kg]
625,824
[25.894 GeV]
Gadolinium
64
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
64
f d p s p d f
electron configuration
[Xe] 4f8 6s2
54
Aufbau
64 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
2
8
26
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.64 - Gadolinium atom

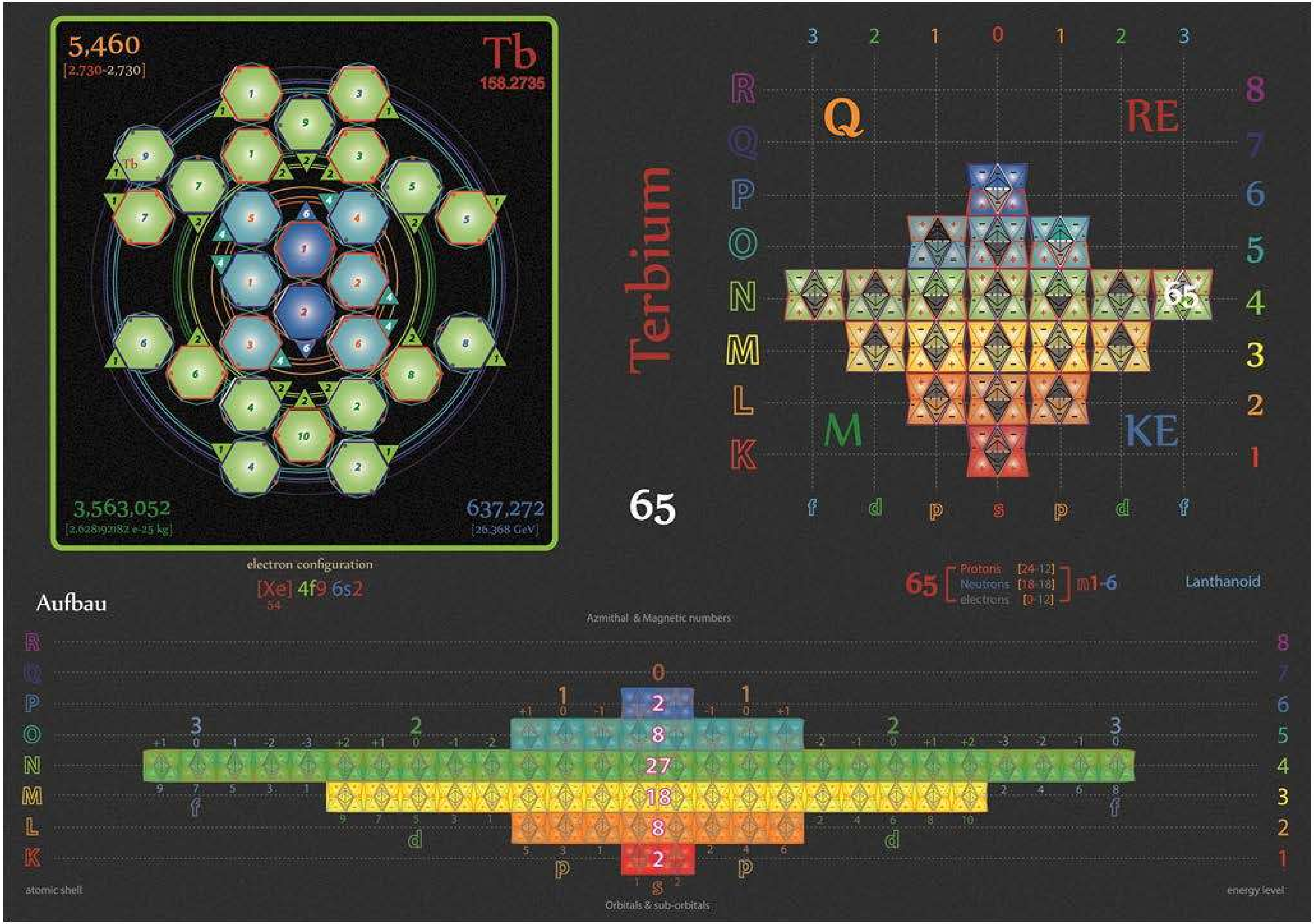
5,460
[2,730-2,730]
Tb
158.2735
3,563,052
[2.628192182 e-25 kg]
637,272
[26.368 GeV]
Terbium
65
R
Q
P
O
N
M
L
K
3
2
1
0
1
2
3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
65
electron configuration
[Xe] 4f9 6s2
54
65 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
2
8
27
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.65 -Terbium atom

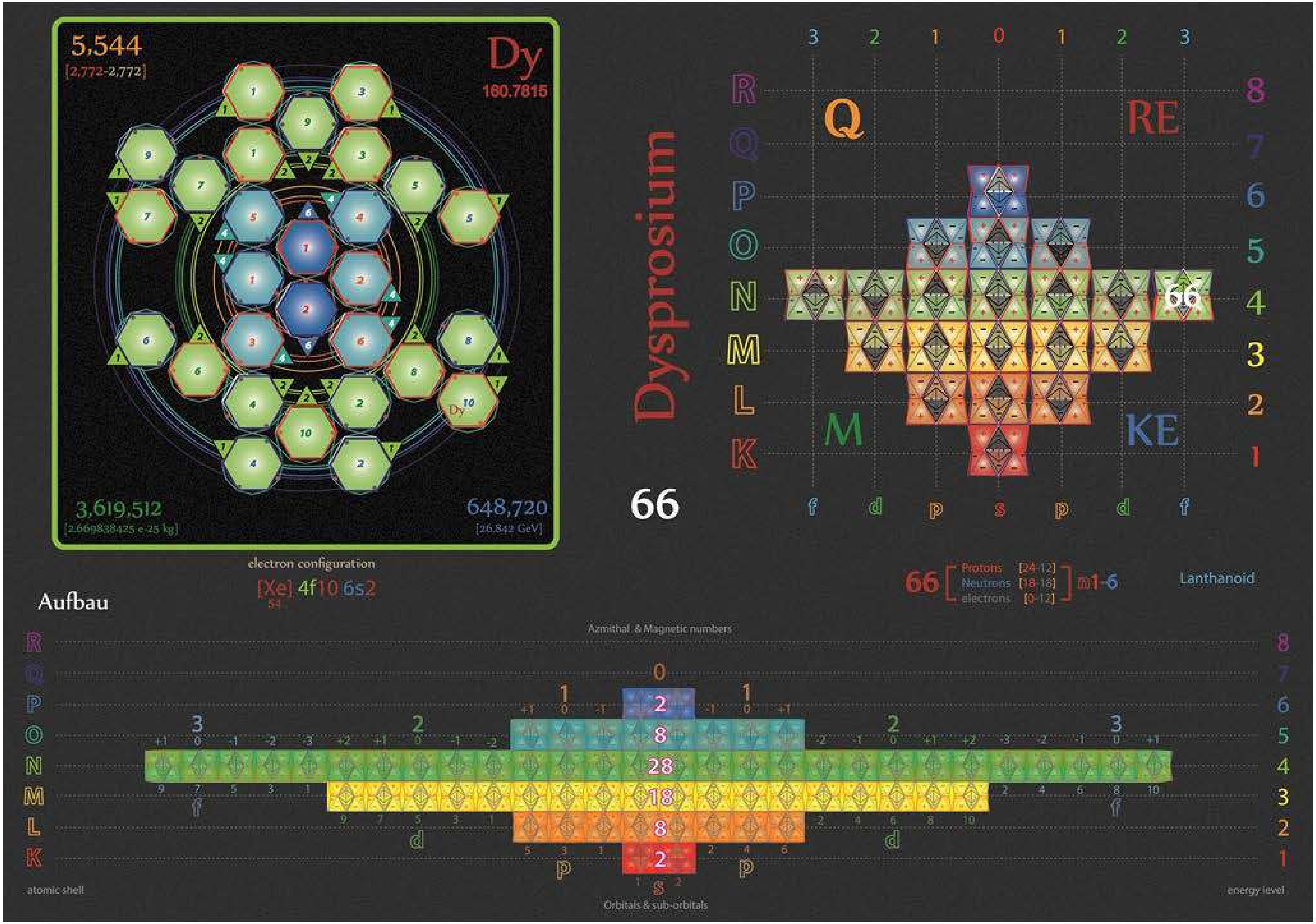
5,544
[2,772-2,772]
Dy
160.7815
3,619,512
[2.669838425 e-25 kg]
648,720
[26.842 GeV]
electron configuration
[Xe] 4f10 6s2
54
Dysprosium
66
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
66
f d p s p d f
66 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
8
28
18
8
2
f f
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.66 - Dysprosium atom

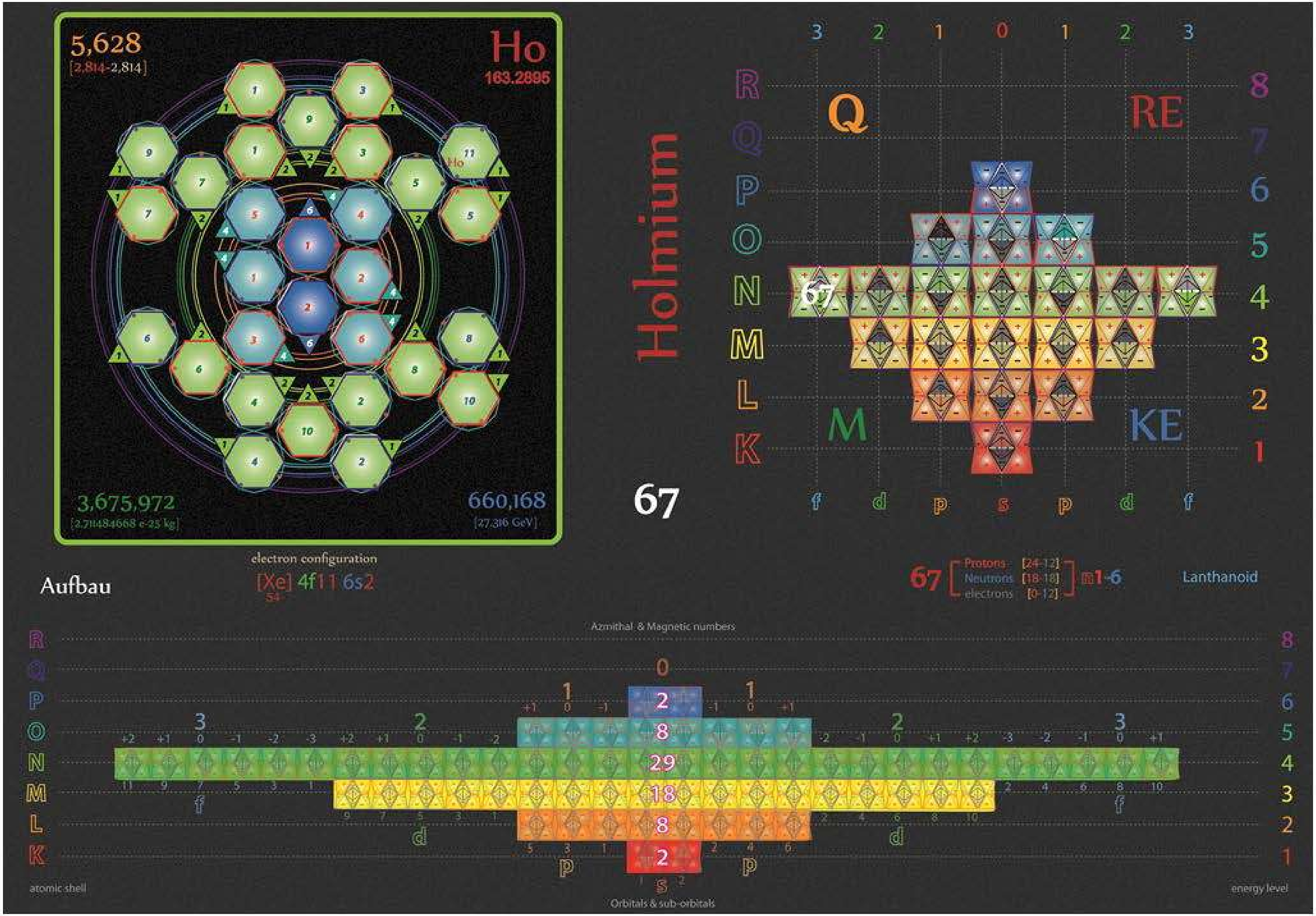
5,628
[2,814-2,814]
Ho
163.2895
3,675,972
[2,711484668 e-25 kg]
660,168
[27,316 GeV]
Holmium
67
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
[Xe] 4f11 6s2
54
Aufbau
67
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
29
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.67 - Holmium atom

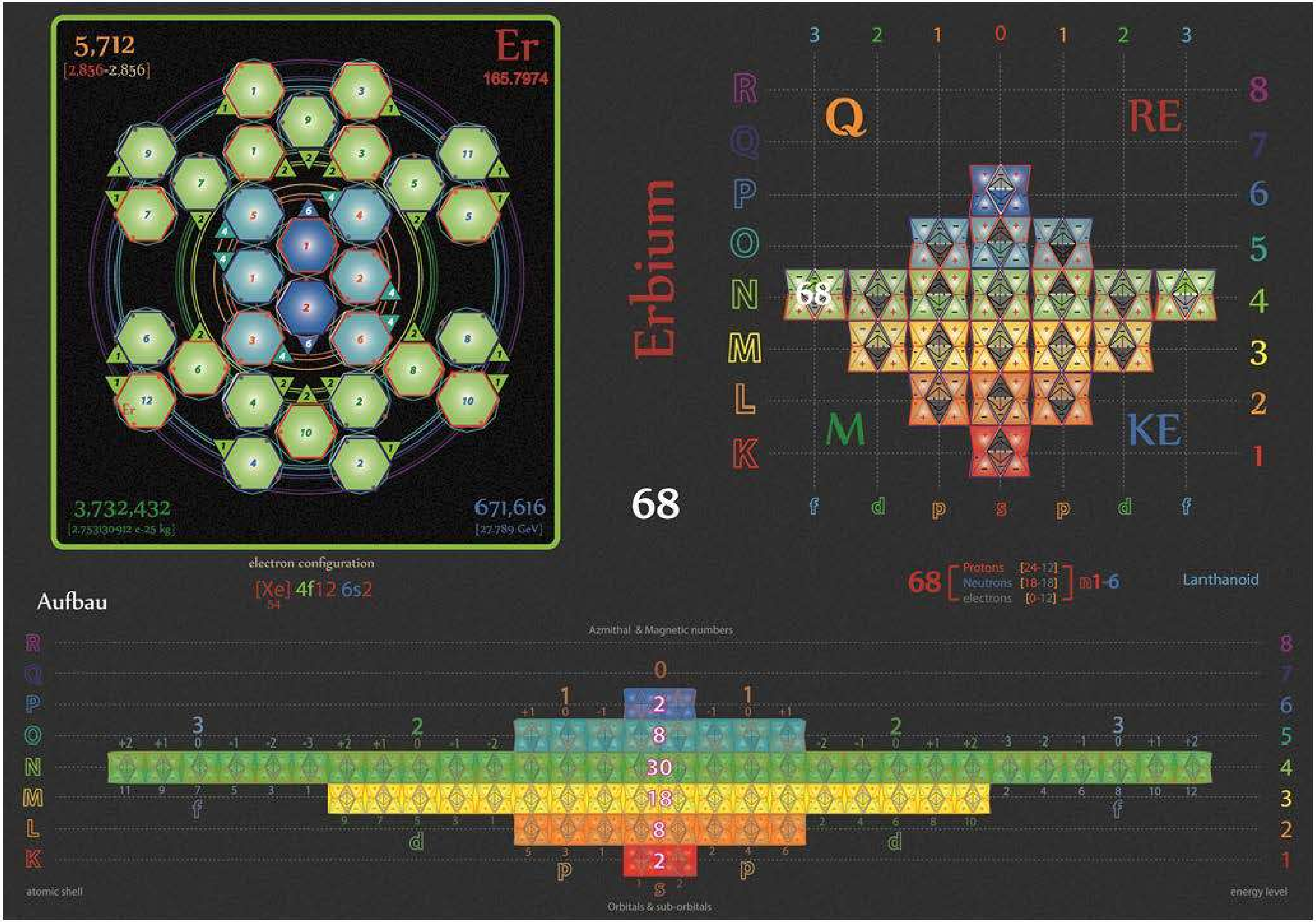
5,712
[2,856-2,856]
Er
165.7974
3,732,432
[2.753130912 e-25 kg]
671,616
[27.789 GeV]
electron configuration
[Xe] 4f12 6s2
54
Erbium
68
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
f
d
p
s
p
d
f
68
68
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
2
8
30
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.68 - Erbium atom

Copyright ABRAHAM [2008] - All rights reserved

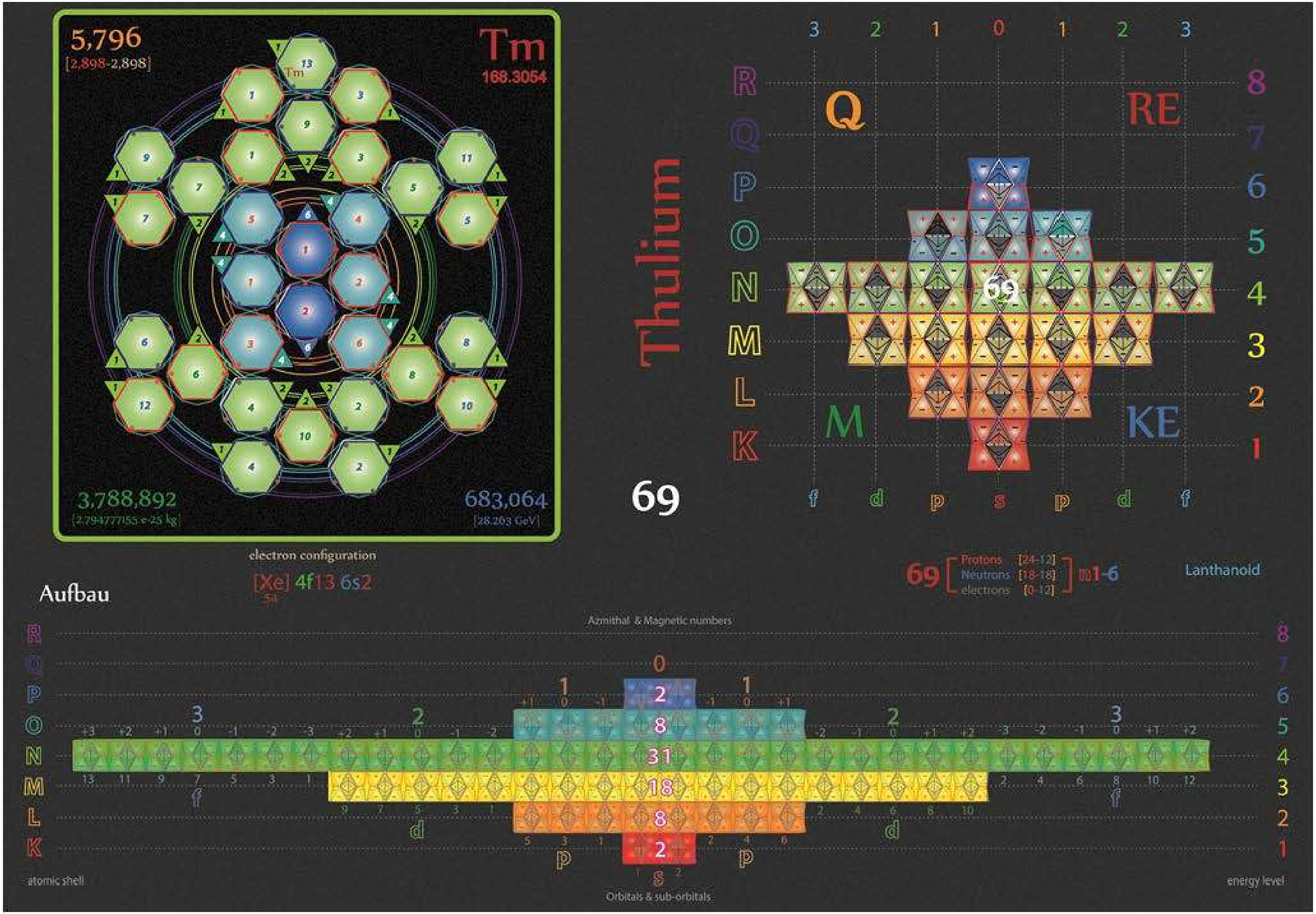
5,796
[2,898-2,898]
Tm
168.3054
3,788,892
[2.794777155 e-25 kg]
683,064
[28.263 GeV]
Thulium
69
electron configuration
[Xe] 4f13 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
69
f d p s p d f
69 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1
2
8
31
18
8
2
3
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.69 - Thulium atom

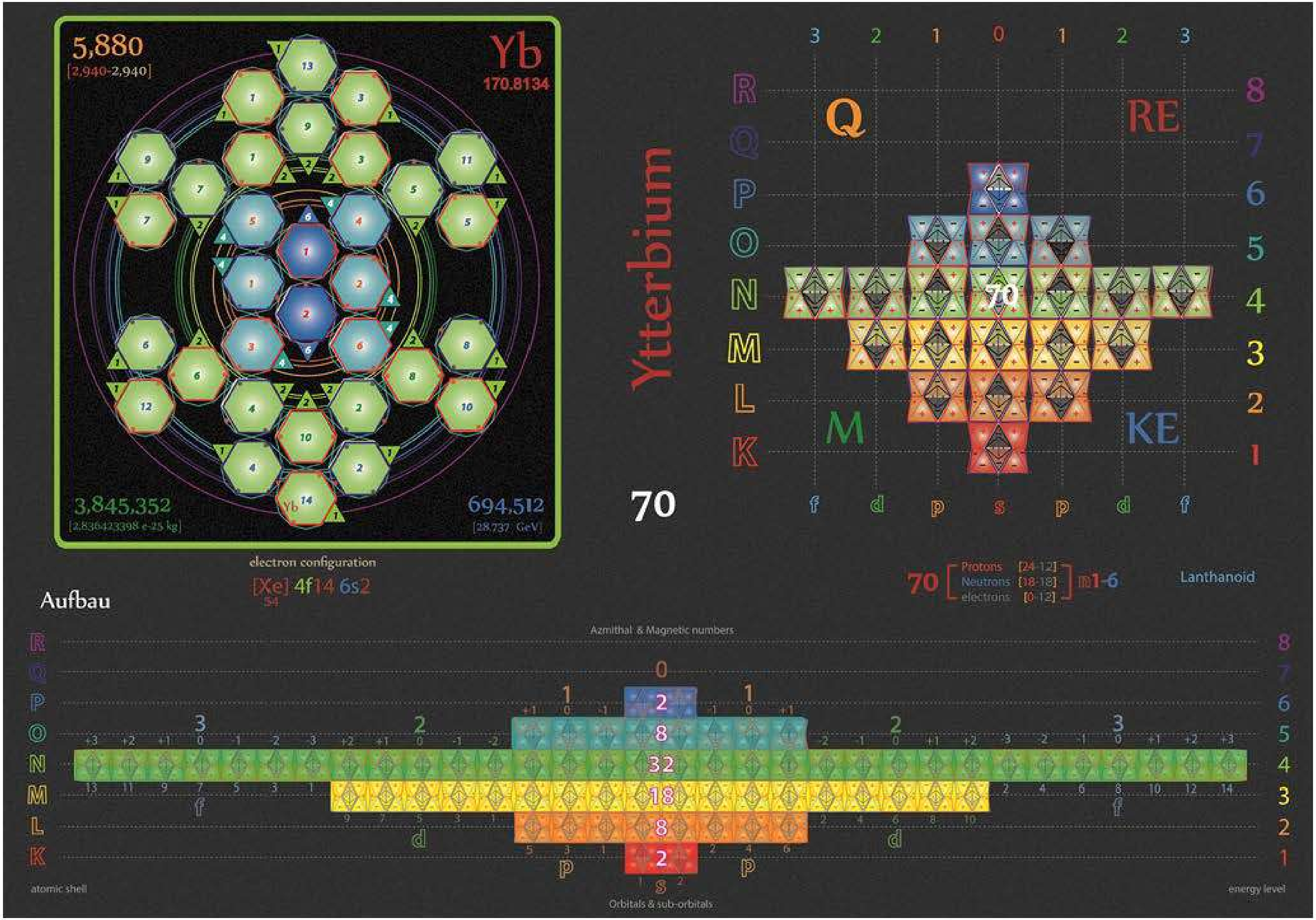
5,880
[2,940-2,940]
Yb
170.8134
3,845,352
[2,836423398 e-25 kg]
694,512
[28,737 GeV]
Ytterbium
70
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f14 6s2
54
Aufbau
70 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.70 - Ytterbium atom

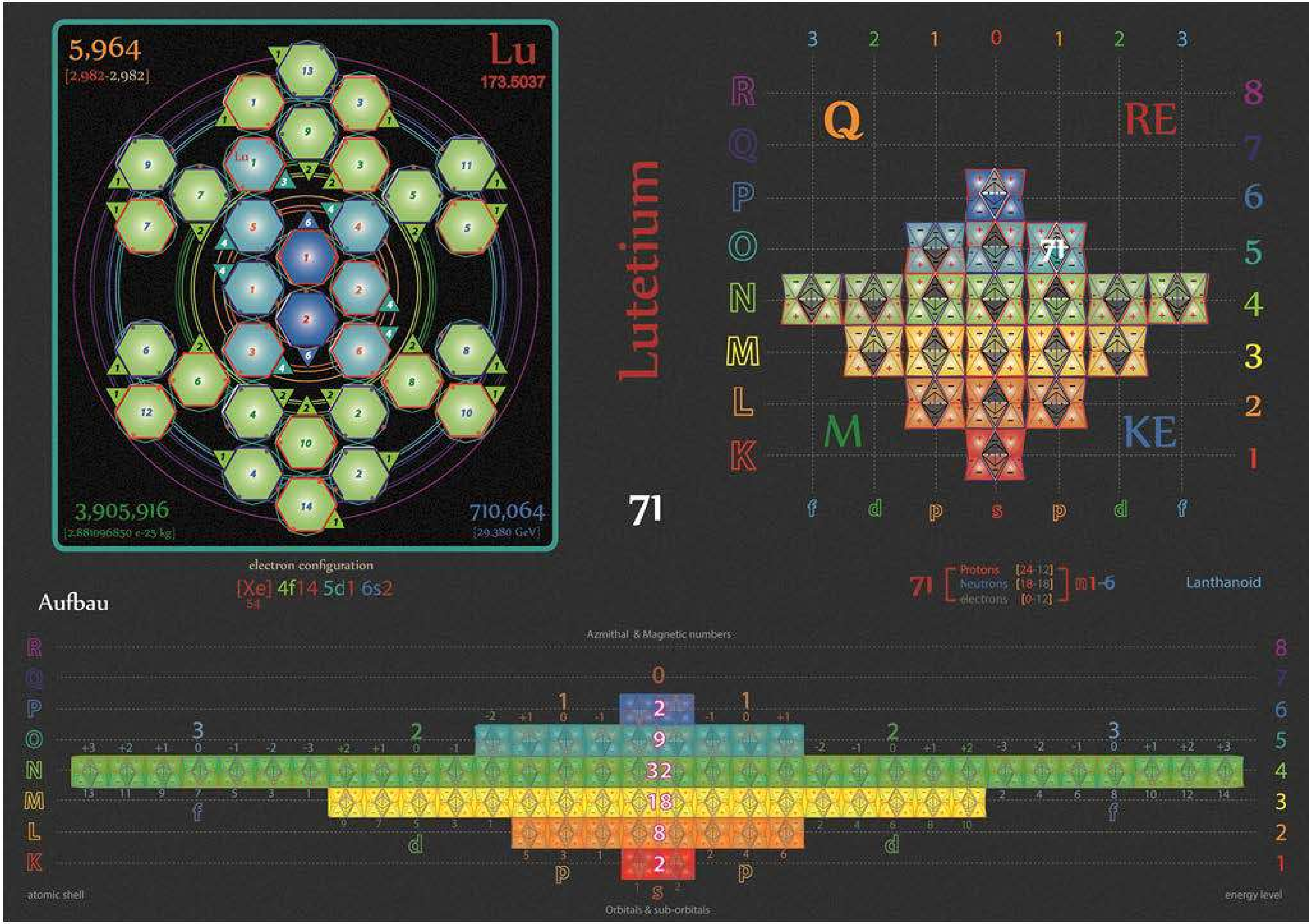
5,964
[2,982-2,982]
Lu
173.5037
3,905,916
[2.881096850 e-25 kg]
710,064
[29.380 GeV]
Lutetium
71
electron configuration
[Xe] 4f14 5d1 6s2
54
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
71 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Lanthanoid
Azmithal & Magnetic numbers
0
2
9
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.71 - Lutetium atom

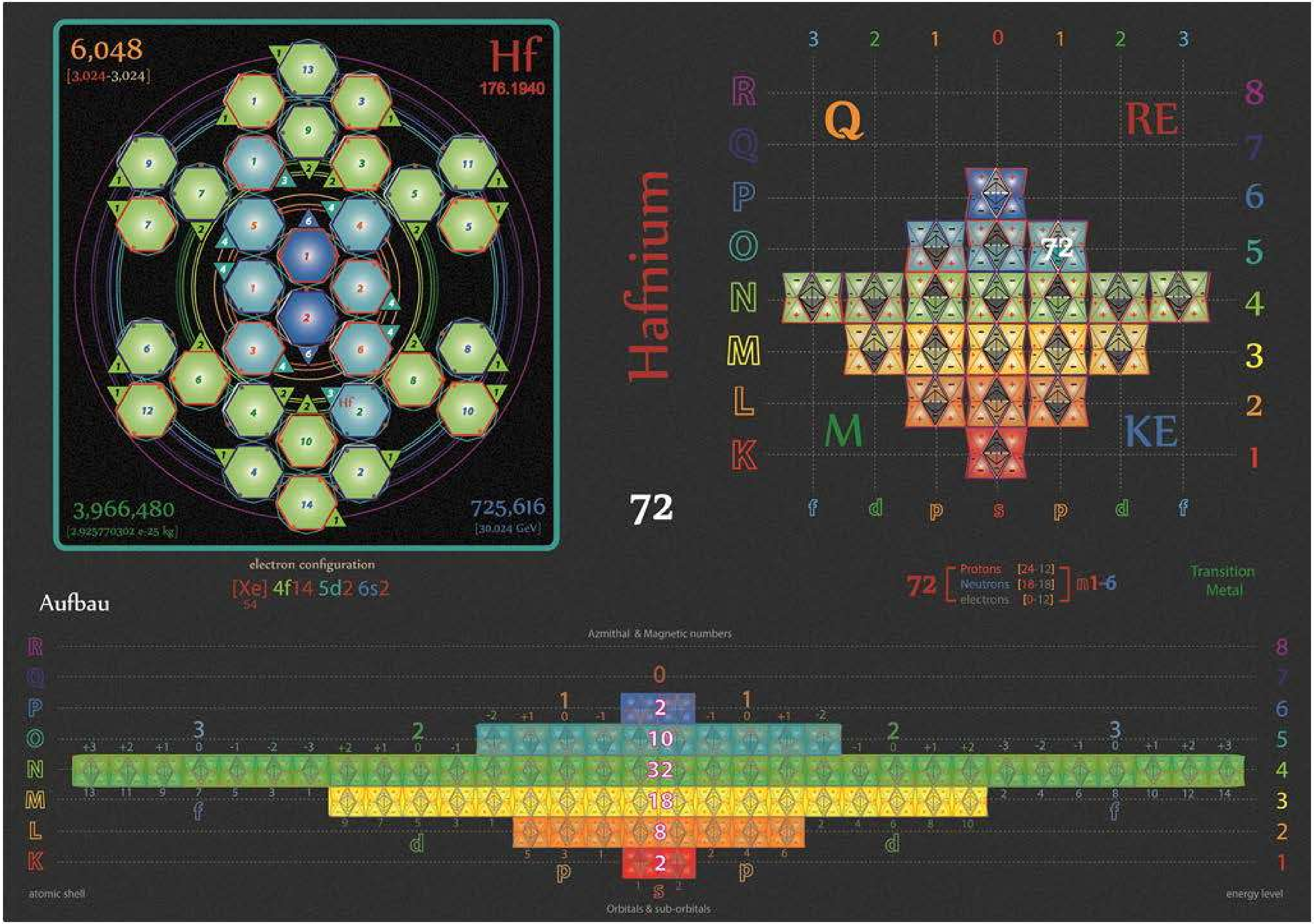
6,048
[3,024-3,024]
Hf
176.1940
3,966,480
[2.925770302 e-25 kg]
725,616
[30.024 GeV]
Hafnium
72
electron configuration
[Xe] 4f14 5d2 6s2
54
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
72
72 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
2
10
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.72 - Hafnium atom

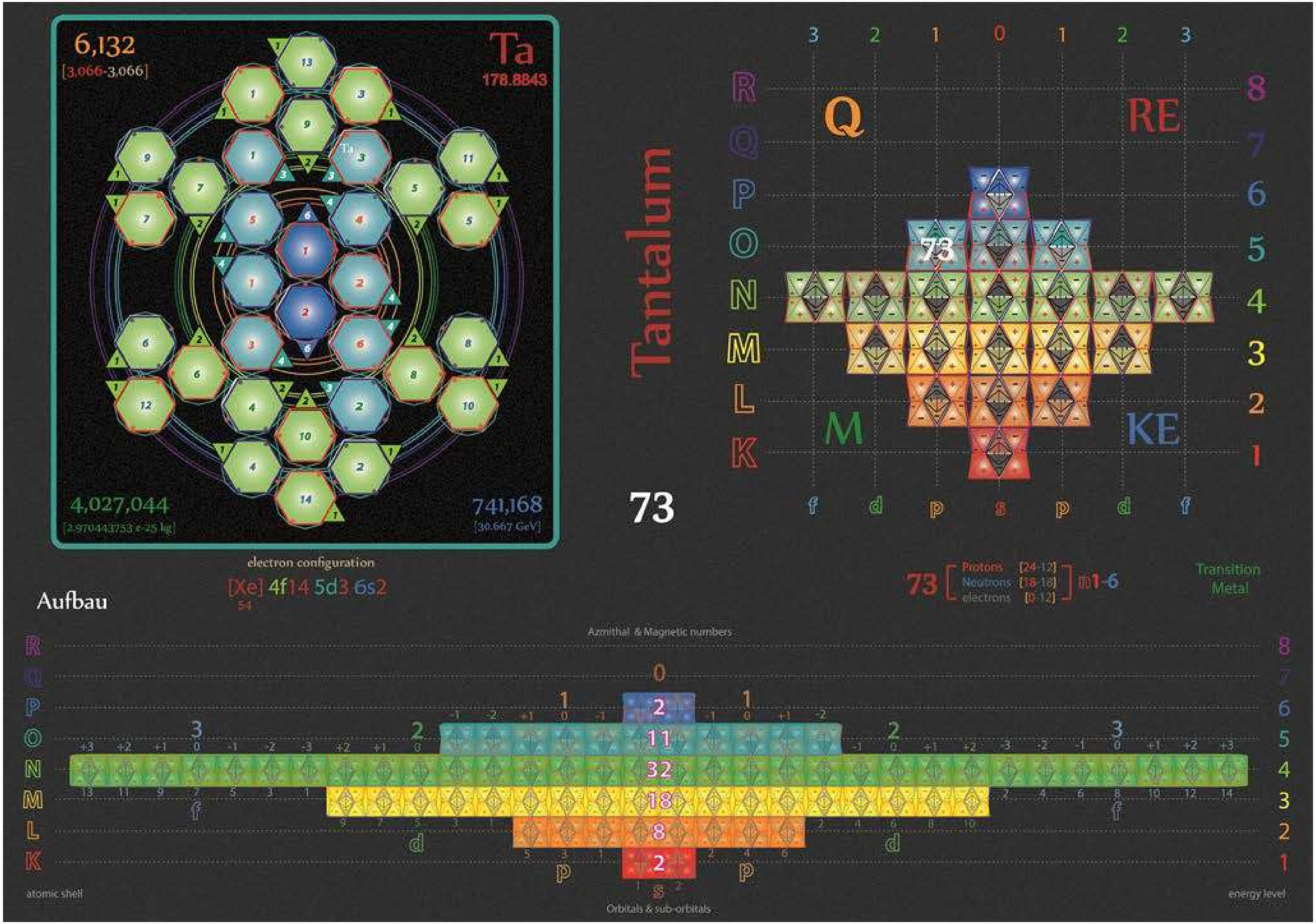
6,132
[3,066-3,066]
Ta
178.8843
4,027,044
[2.970443753 e-25 kg]
741,168
[30.667 GeV]
Tantalum
73
electron configuration
[Xe] 4f14 5d3 6s2
54
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
Q
RE
M
KE
f
d
p
s
p
d
f
73
73 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
1
1
3
2
2
3
2
11
32
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.73 - Tantalum atom

Copyright ABRAHAM [2008] - All rights reserved

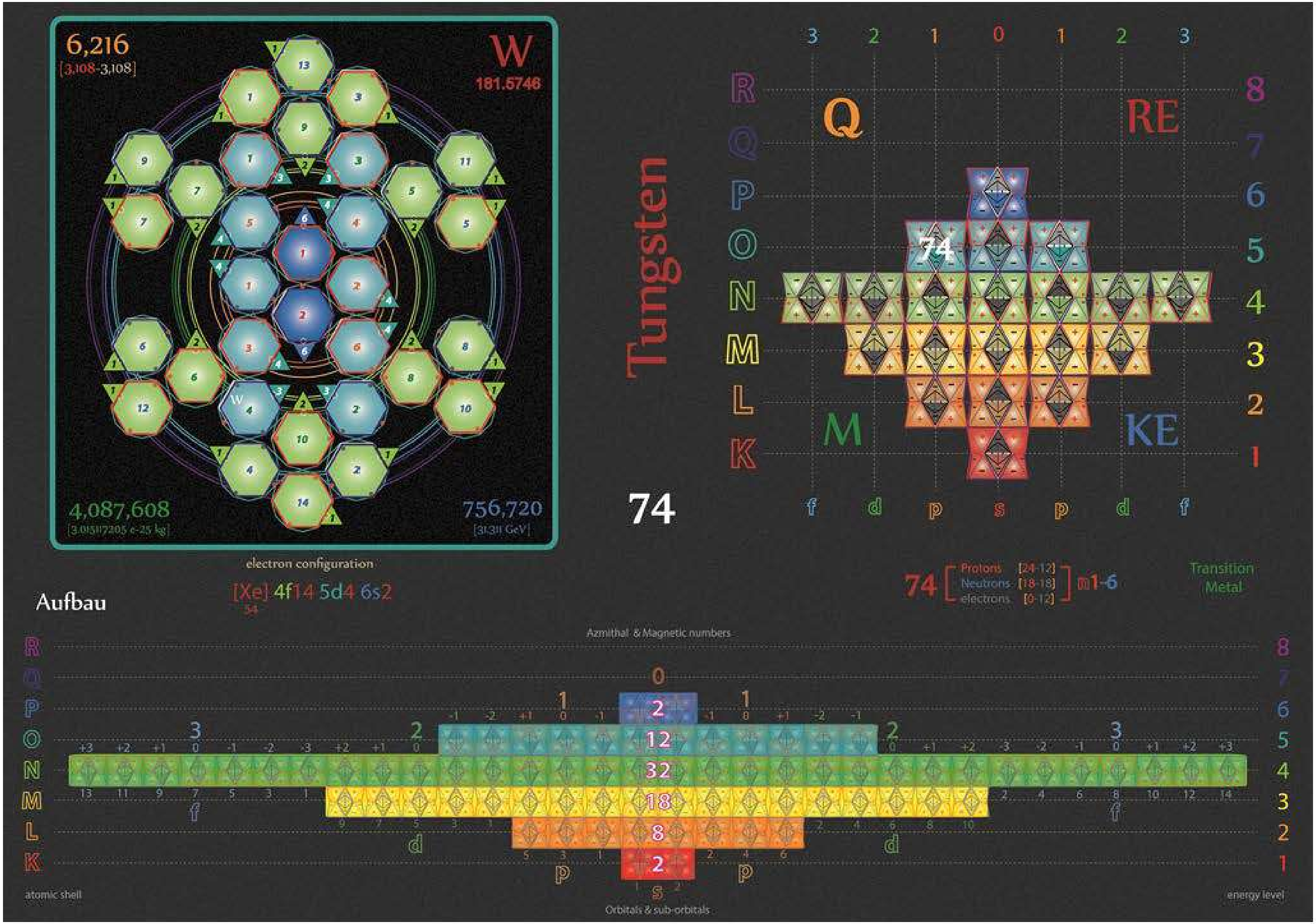
6,216
[3,108-3,108]
W
181.5746
4,087,608
[3.015117205 e-25 kg]
756,720
[31311 GeV]
Tungsten
74
electron configuration
[Xe] 4f14 5d4 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
74
f d p s p d f
74 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
2
12
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.74 - Tungsten atom

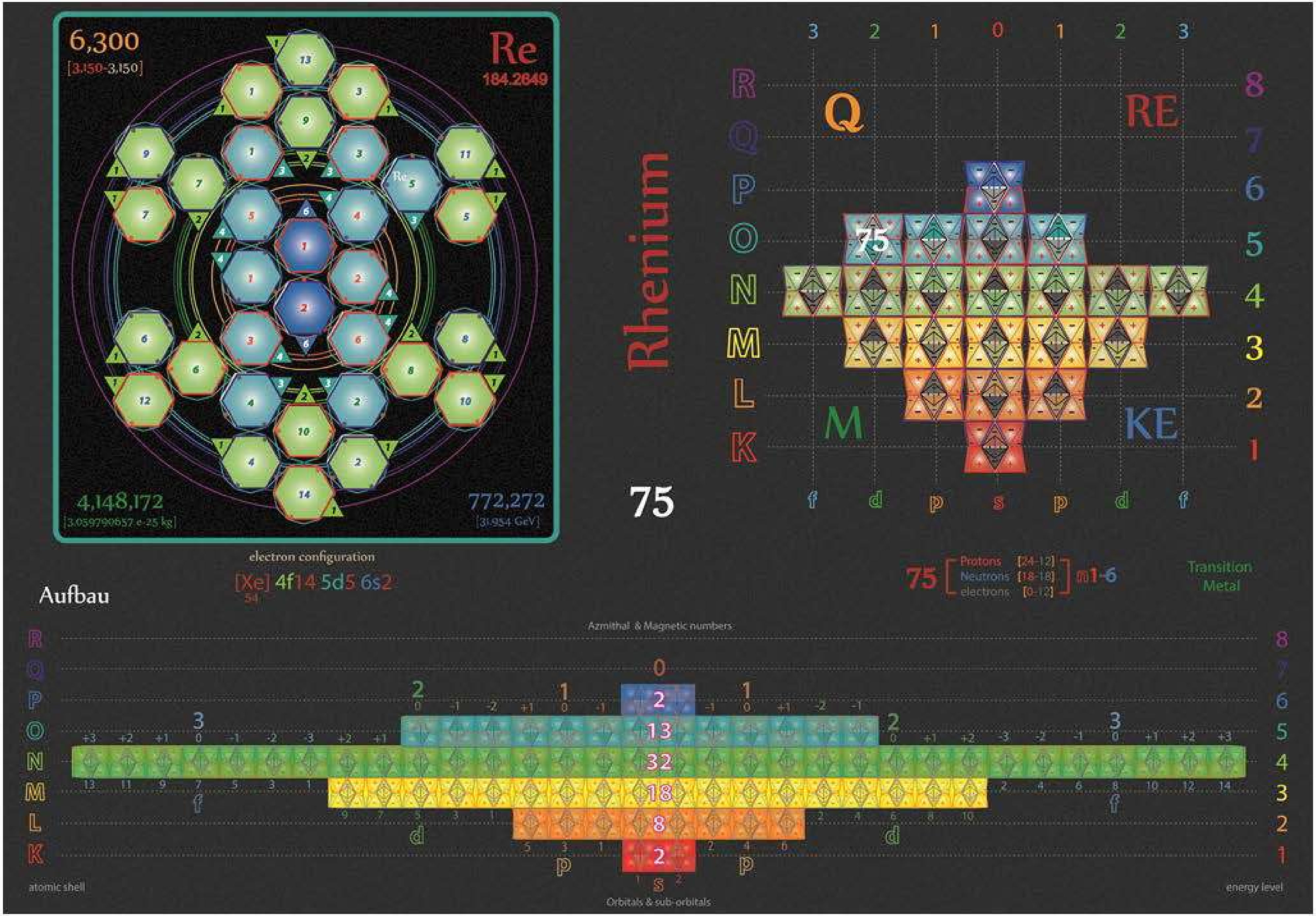
6,300
[3,150-3,150]
Re
184.2849
4,148,172
[3.059790657 e-25 kg]
772,272
[31.954 GeV]
Rhenium
75
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f14 5d5 6s2
54
75 [Protons [24-12]
Neutrons [18-18]
electrons [0-12]] m1-6
Transition
Metal
Aufbau
Azmithal & Magnetic numbers
2
13
32
18
8
2
f
d
p
s
p
d
f
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.75 - Rhenium atom

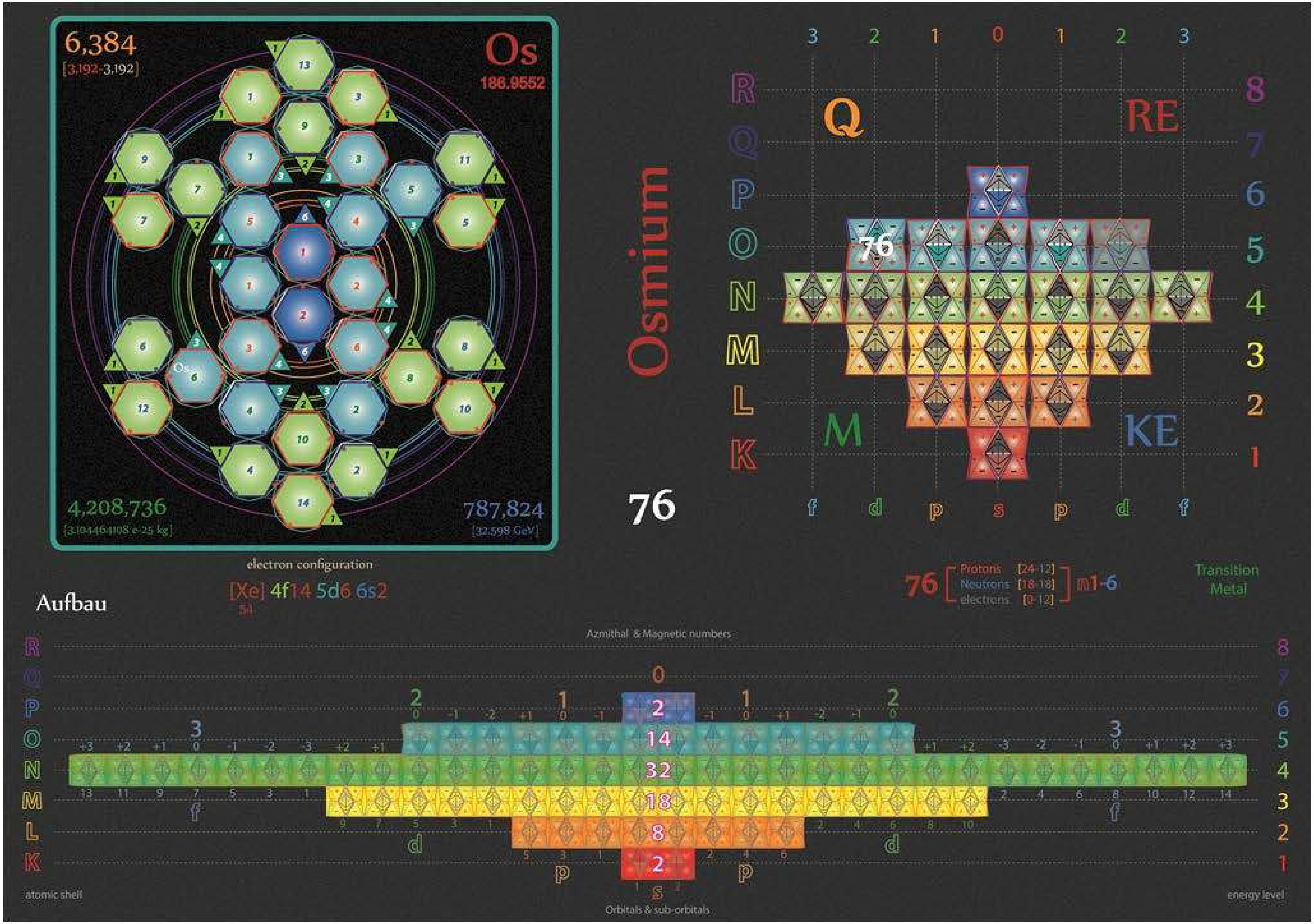
6,384
[3,192-3,192]
Os
186.9552
4,208,736
[3.104464108 e-25 kg]
787,824
[32.598 GeV]
Osmium
76
electron configuration
[Xe] 4f14 5d6 6s2
5d
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
76
f d p s p d f
76 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
2
14
32
18
8
2
3 2 1 1 2 3
+3 +2 +1 0 -1 -2 -3 +2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2 -3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1 2 4 6 8 10 12 14
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.76 - Osmium atom

Copyright ABRAHAM [2008] - All rights reserved

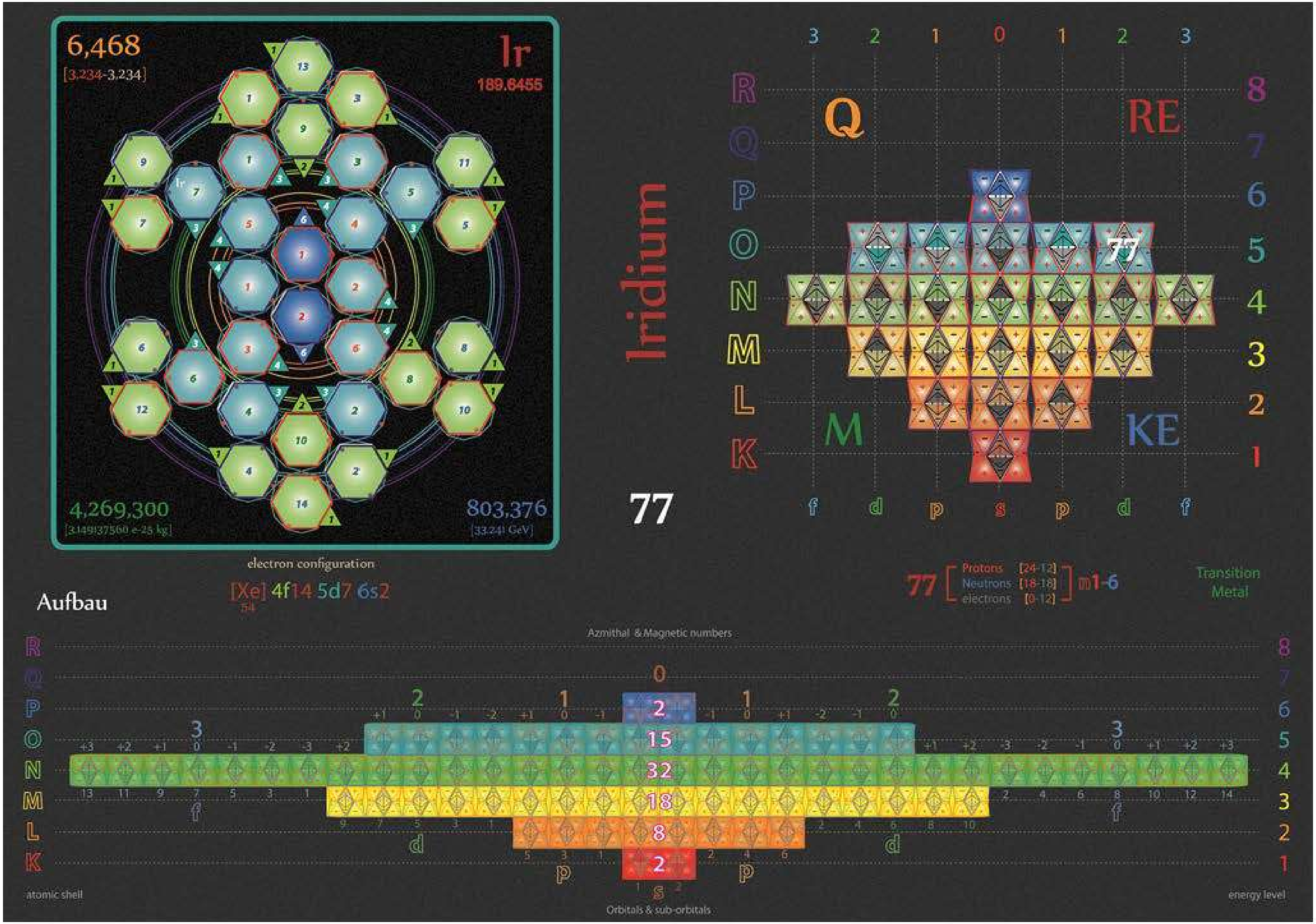
6,468
[3.234-3.234]
Ir
189.8455
4,269,300
[3.149137560 e-25 kg]
803,376
[33.241 GeV]
Iridium
77
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
Q
RE
M
KE
77
f
d
p
s
p
d
f
electron configuration
[Xe] 4f14 5d7 6s2
54
Aufbau
77
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Transition
Metal
Azmithal & Magnetic numbers
0
2
15
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.77 - Iridium atom

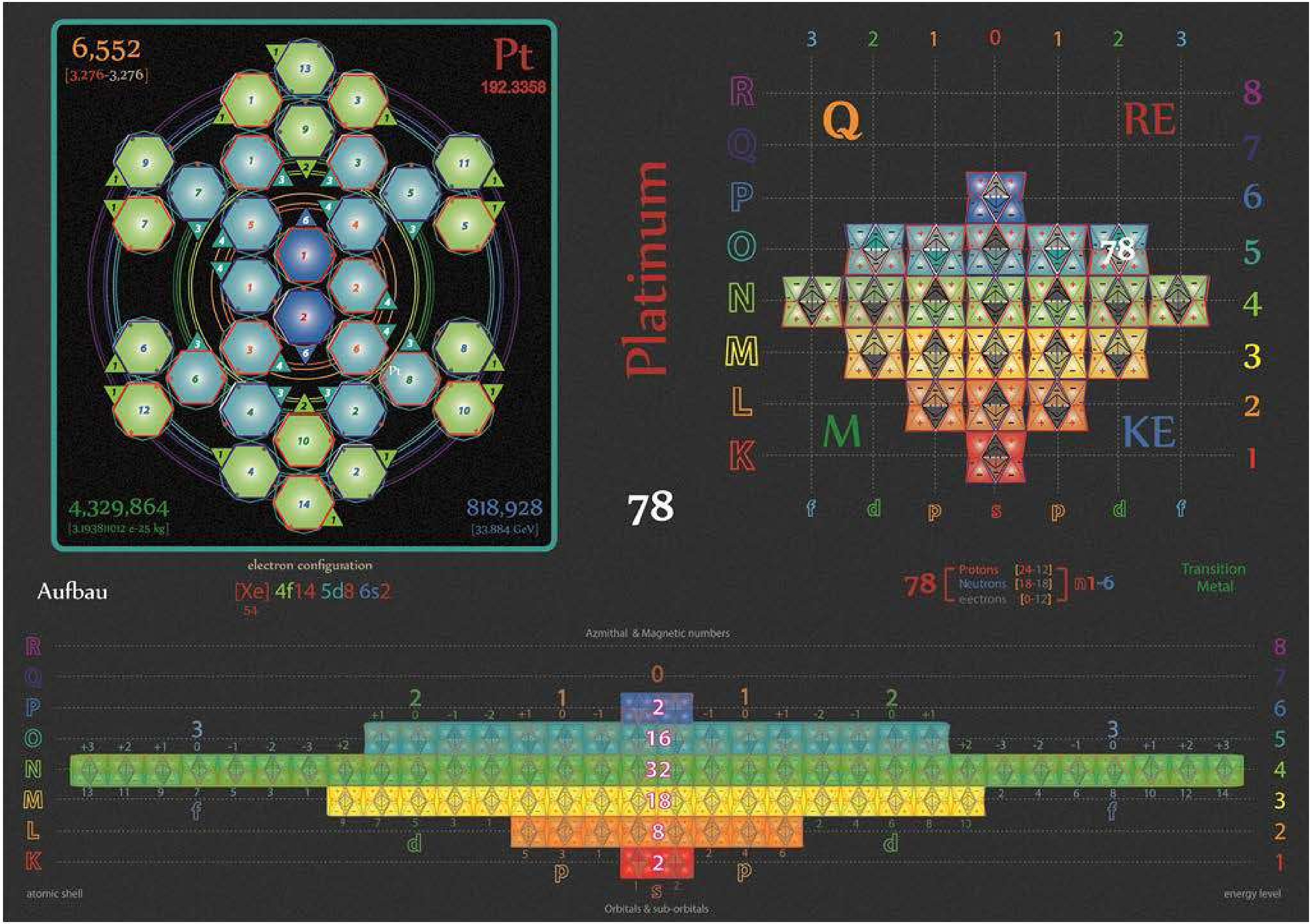
6,552
[3,276-3,276]
Pt
192.3358
4,329,864
[3.19381012 e-25 kg]
818,928
[33.884 GeV]
Platinum
78
R
Q
P
O
N
M
L
K
Q
RE
M
KE
3
2
1
0
1
2
3
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
Aufbau
[Xe] 4f14 5d8 6s2
54
78 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Transition Metal
Azmithal & Magnetic numbers
2
16
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.78 - Platinum atom

Copyright ABRAHAM [2008] - All rights reserved

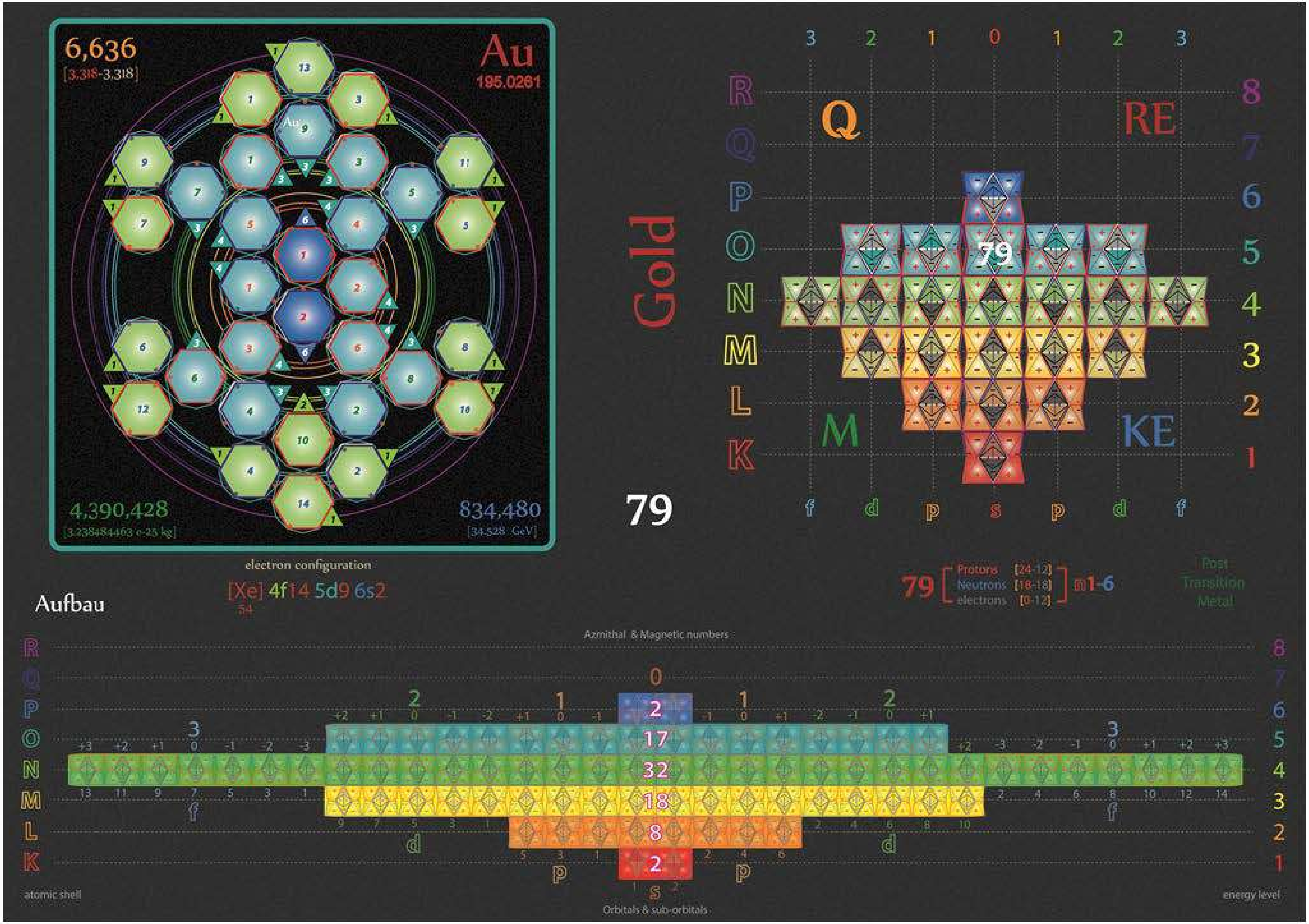
6,636
[3.318-3.318]
Au
195.0261
4,390,428
[3.238484463 e-25 kg]
834,480
[34.528 GeV]
electron configuration
[Xe] 4f14 5d9 6s2
54
Gold
79
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
79
79 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Post Transition Metal
Aufbau
Azmithal & Magnetic numbers
2
17
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.79 - Gold atom

Mercury

80

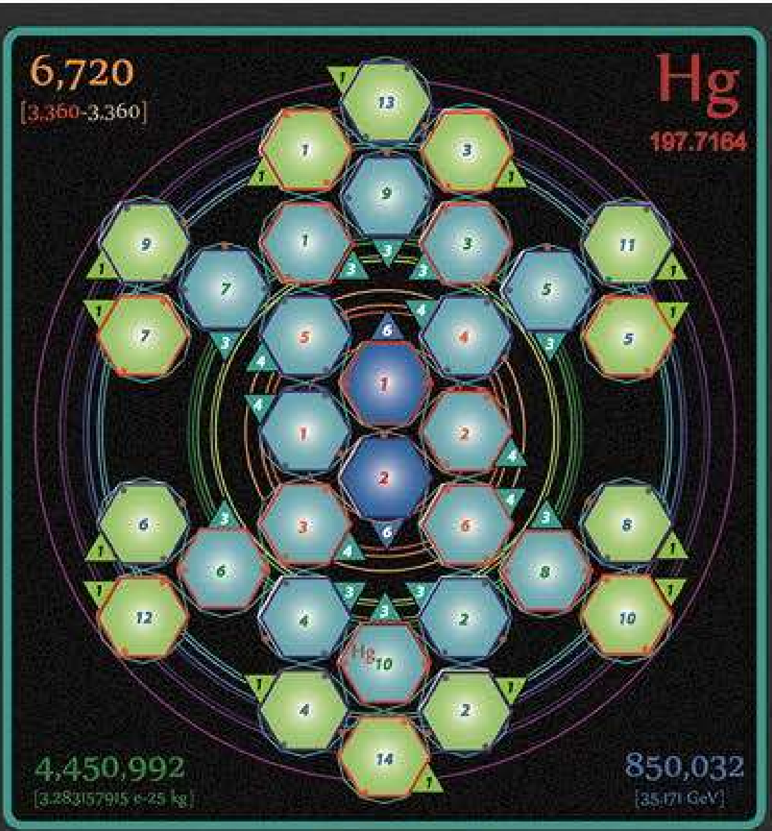
6,720
[3,360-3,360]
Hg
197.7164
4,450,992
[3.283157915 e-25 kg]
850,032
[35.171 GeV]

electron configuration

[Xe] 4f14 5d10 6s2
54

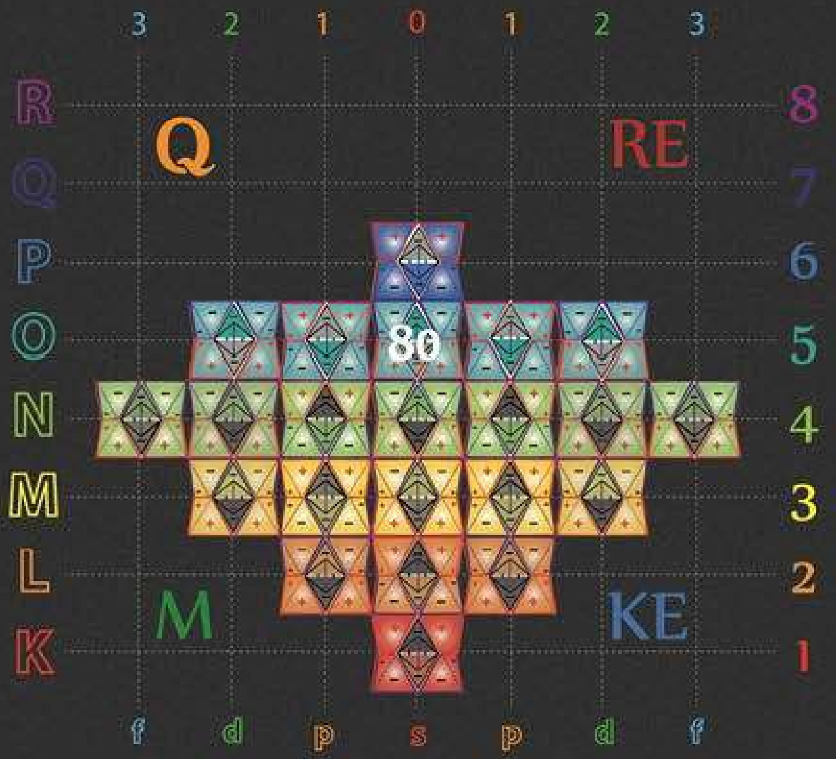
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
80
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

80 [Protons [24-12]
Neutrons [18-18]
electrons [0-12]] @1-6

Post
Transition
Metal

Aufbau

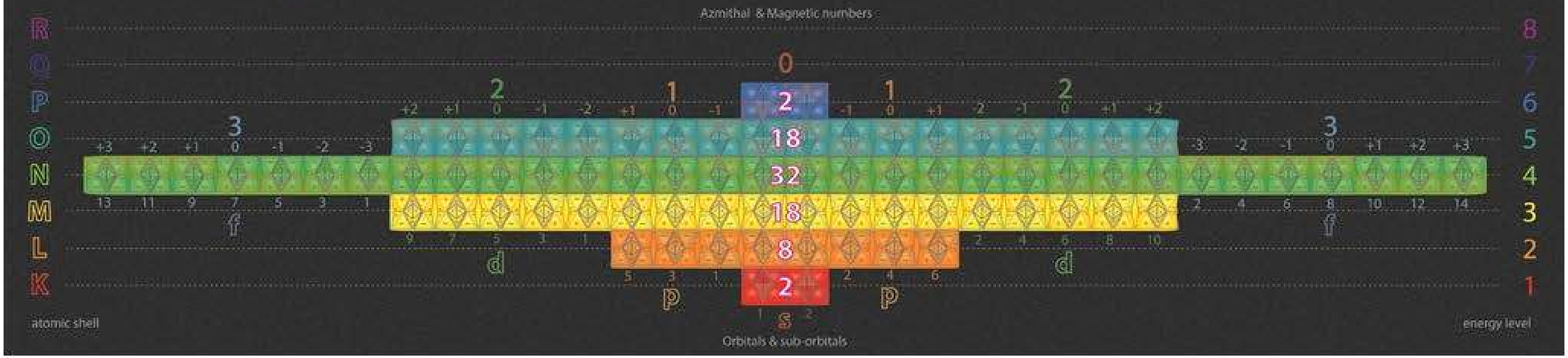
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
2
18
32
18
8
2
3
2
1
1
2
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 3 1
p
1
s
2
2 4 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
8
7
6
5
4
3
2
1
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.80 - Mercury atom

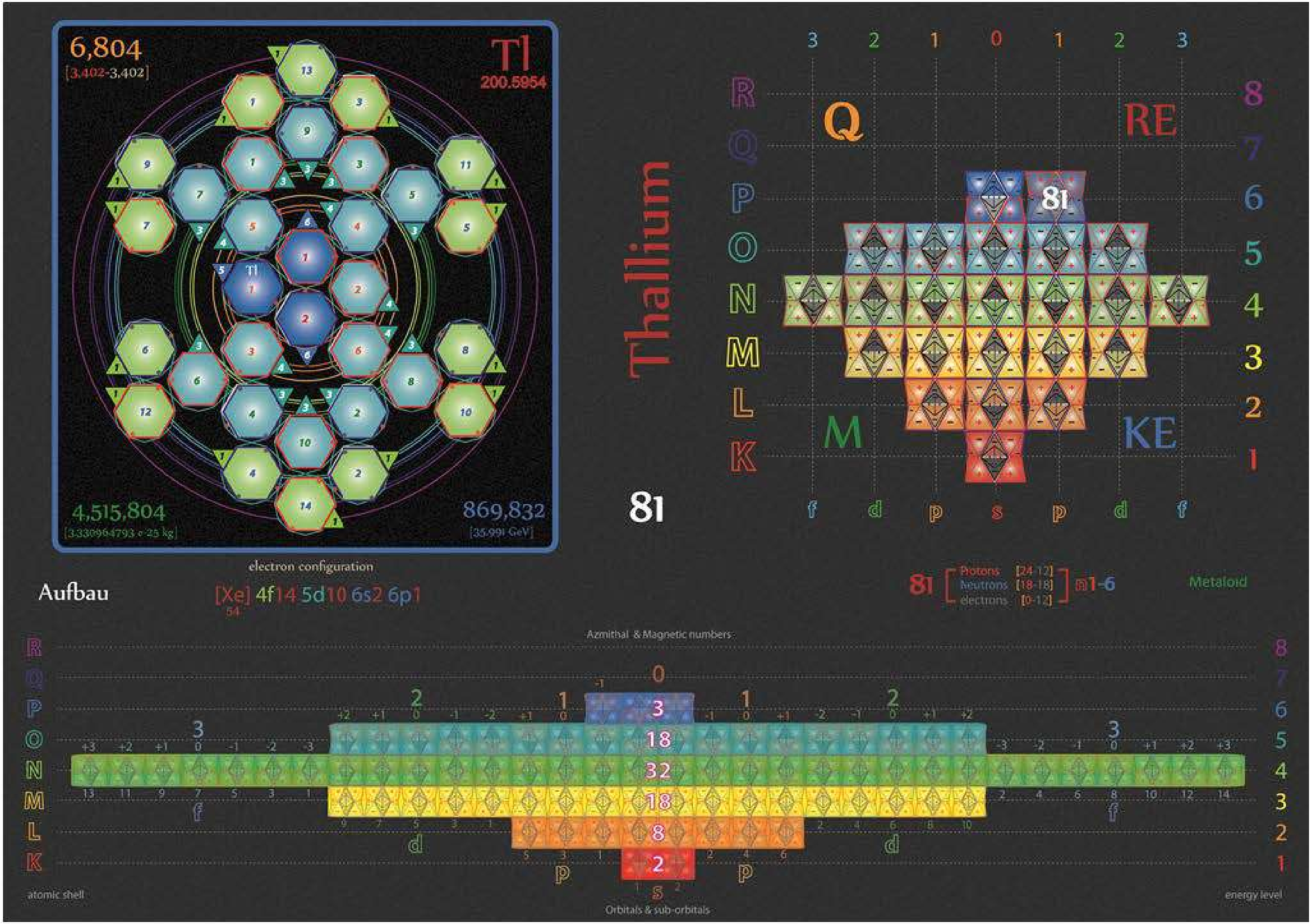
6,804
[3,402-3,402]
Tl
200.5954
4,515,804
[3.330964793 e-25 kg]
869,832
[35.991 GeV]
Thallium
81
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
f
d
p
s
p
d
f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p1
54
81
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-6
Metaloid
Azmithal & Magnetic numbers
0
3
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.81 - Thallium atom

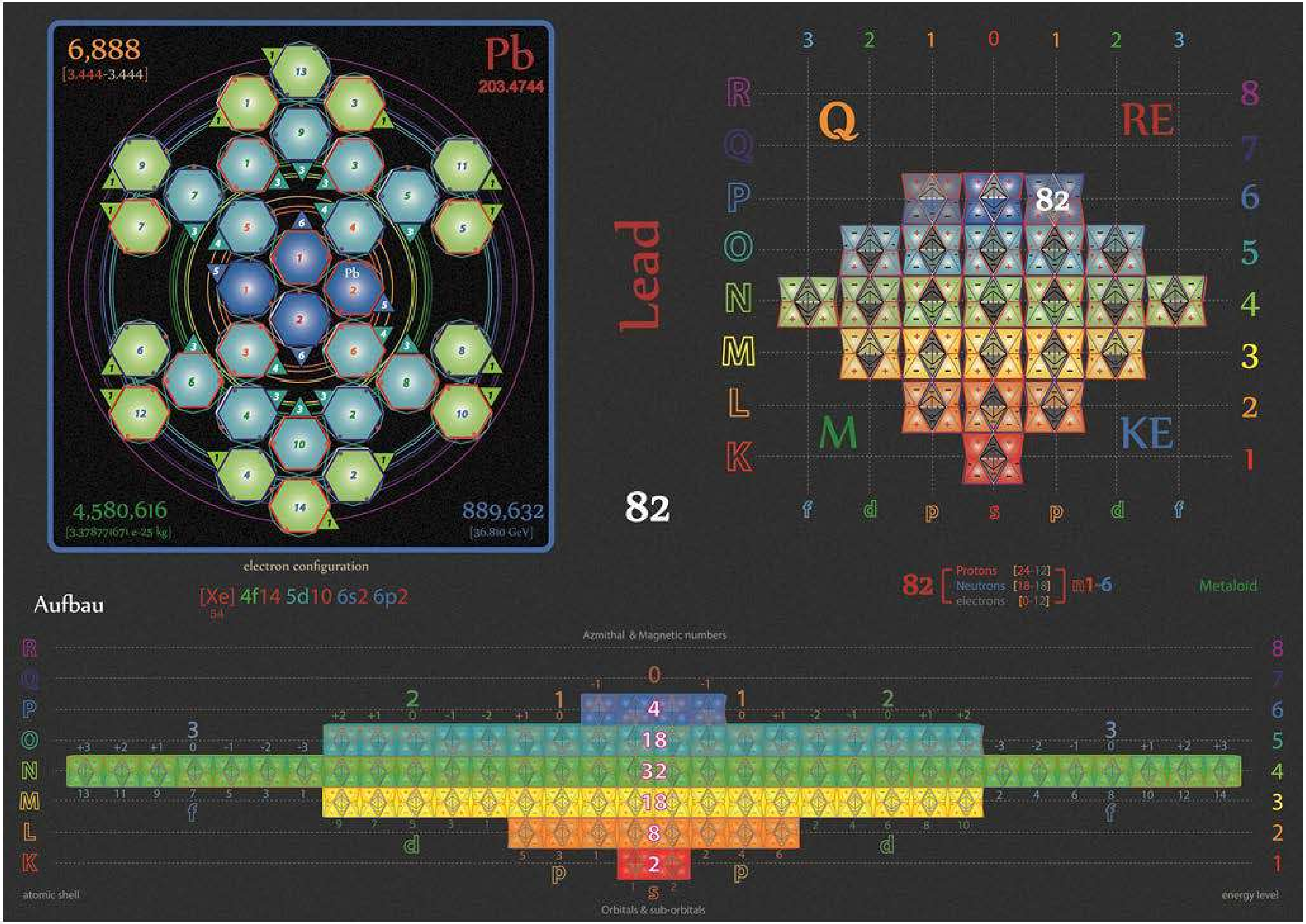
6,888
[3.444-3.444]
Pb
203.4744
4,580,616
[3.37877167 e-25 kg]
889,632
[36.810 GeV]
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p2
54
Lead
82
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
82
f d p s p d f
82
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Metaloid
Azmithal & Magnetic numbers
0
4
18
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.82 - Lead atom

Copyright ABRAHAM [2008] - All rights reserved

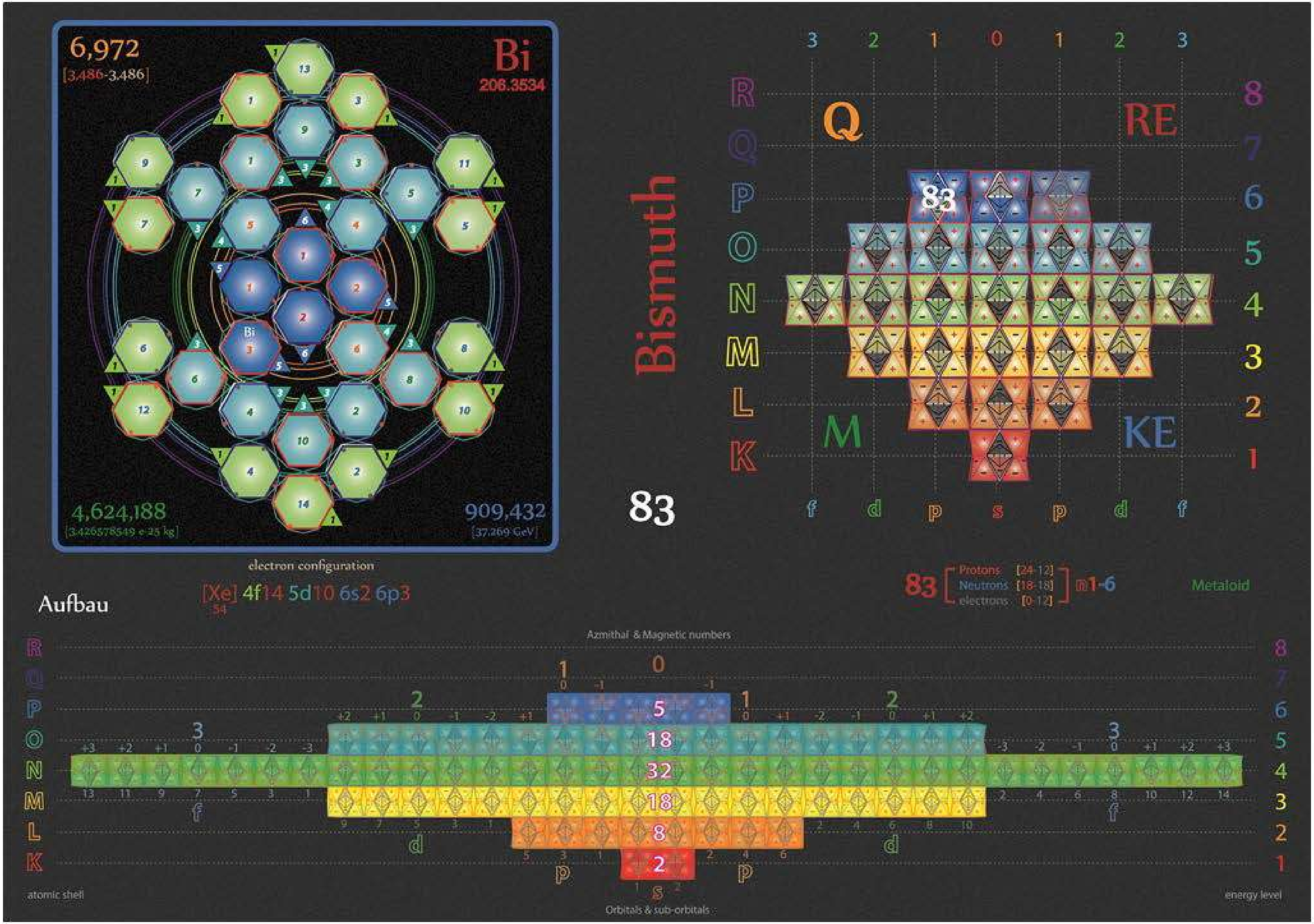
6,972
[3,486-3,486]
Bi
208.3534
4,624,188
[3.426578549 e-25 kg]
909,432
[37.269 GeV]
Bismuth
83
electron configuration
[Xe] 4f14 5d10 6s2 6p3
54
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
3 2 1 0 1 2 3
f d p s p d f
83 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Metaloid
Azmithal & Magnetic numbers
5
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.83 - Bismuth atom

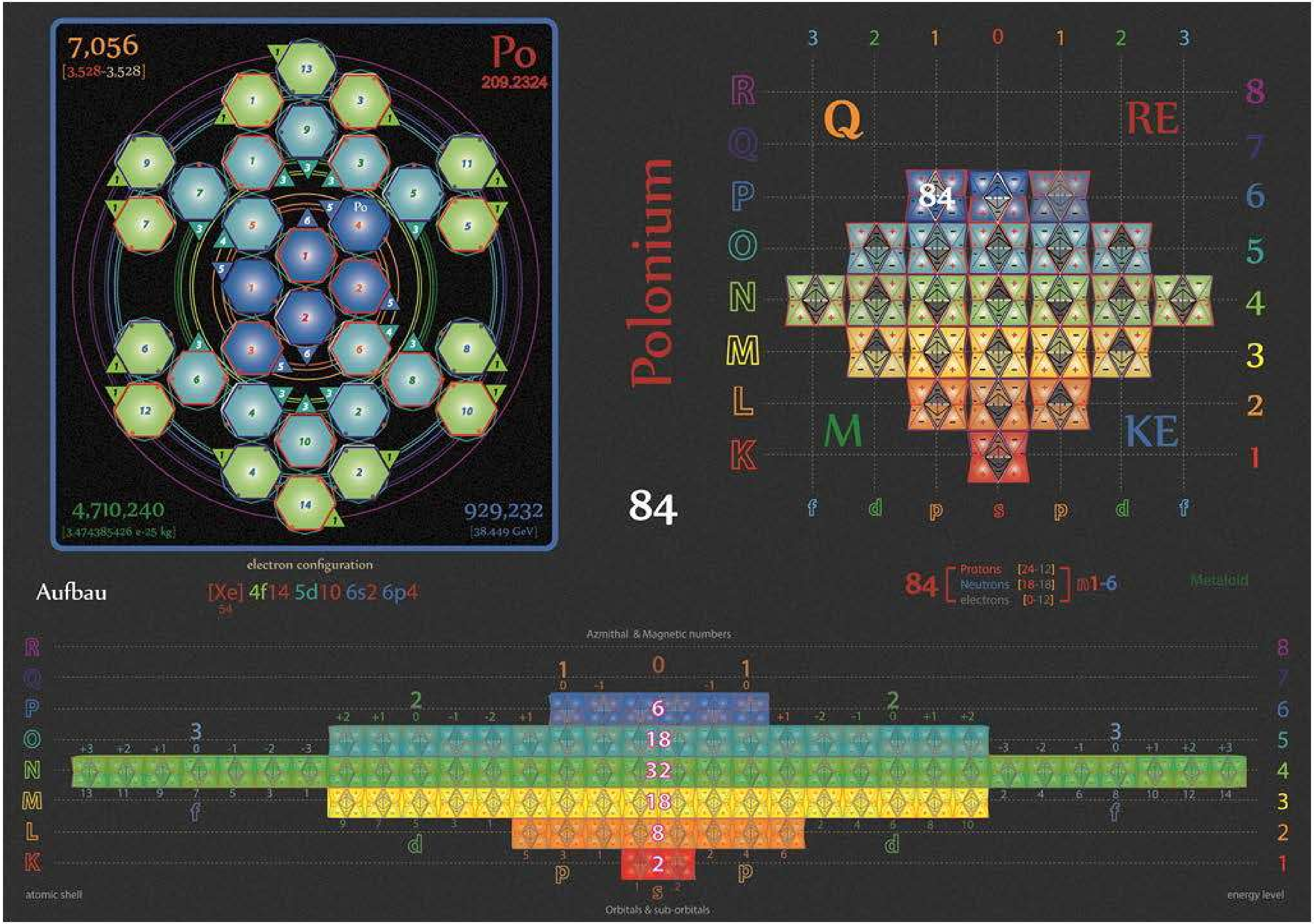
7,056
[3,528-3,528]
Po
209.2324
4,710,240
[3.474385426 e-25 kg]
929,232
[38.449 GeV]
Polonium
84
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p4
54
84 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Metalloid
Azmithal & Magnetic numbers
6
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.84 - Polonium atom

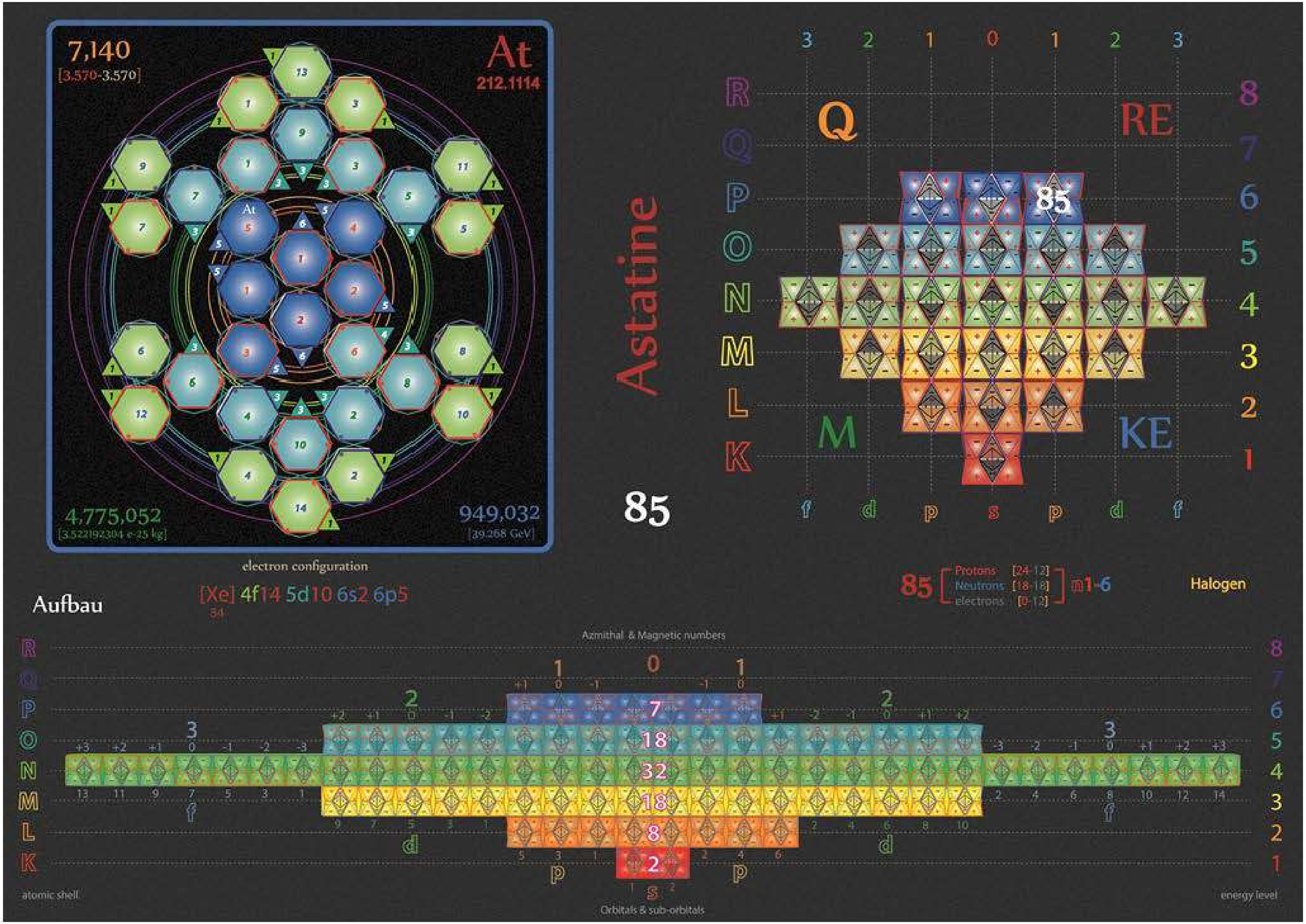
7,140
[3.570-3.570]
At
212.1114
4,775,052
[3.522192304 e-25 kg]
949,032
[39.268 GeV]
Astatine
85
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p5
54
85 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Halogen
Azmithal & Magnetic numbers
7
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.85 - Astatine atom

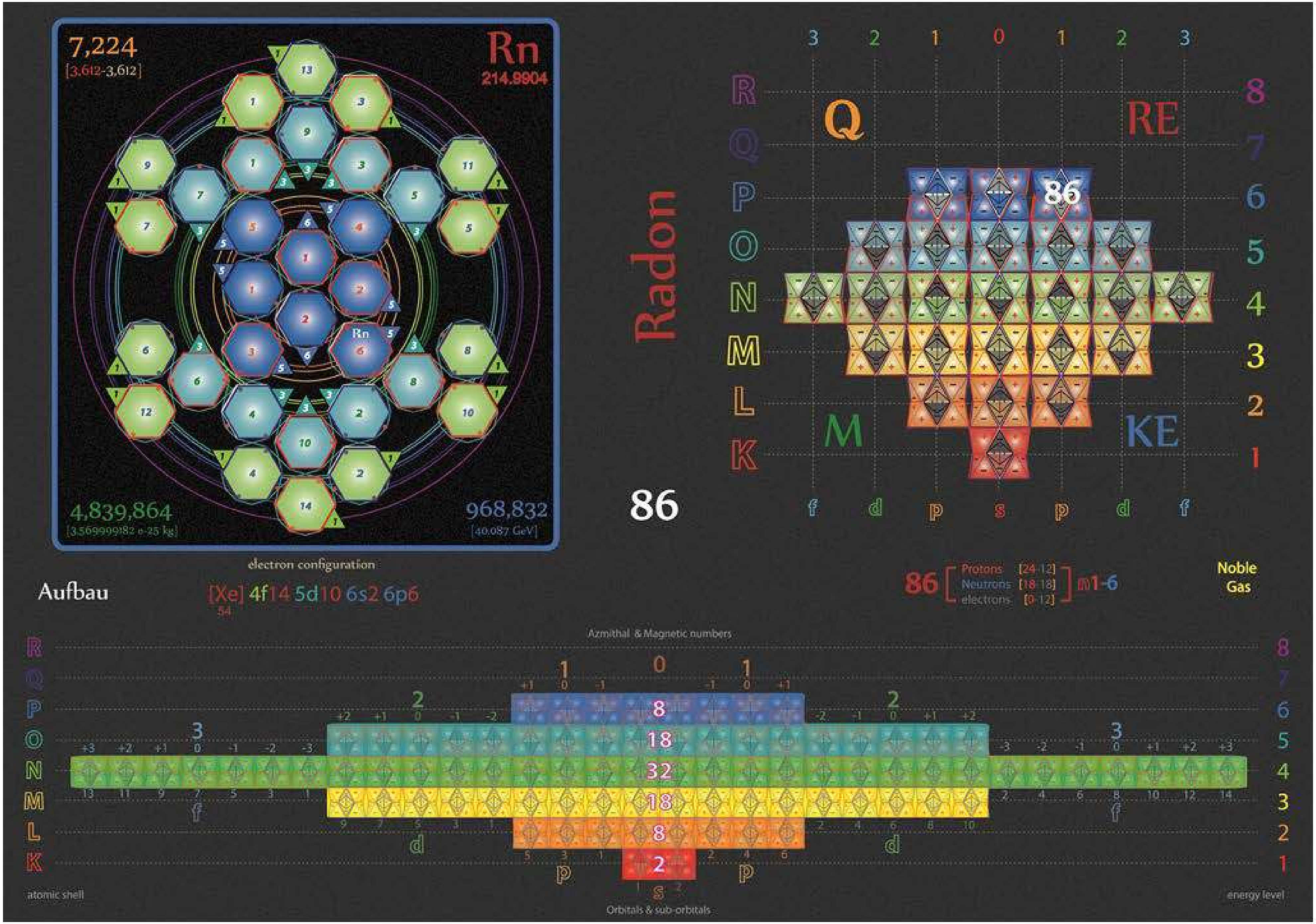
7,224
[3,612-3,612]
Rn
214.9904
4,839,864
[3.569999182 e-25 kg]
968,832
[40.087 GeV]
Radon
86
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p6
54
86 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Noble Gas
Azmithal & Magnetic numbers
8
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.86 - Radon atom

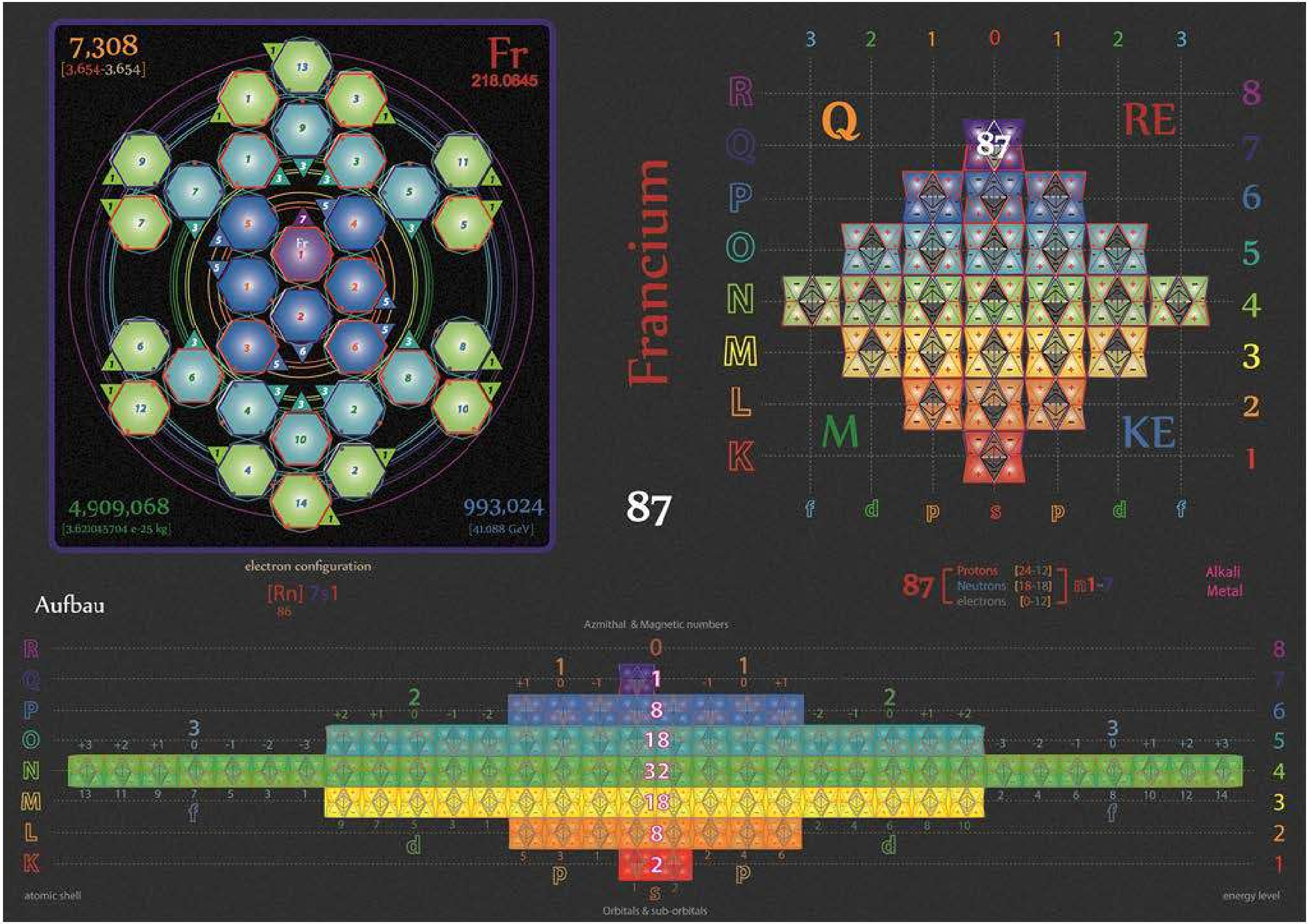
7,308
[3,654-3,654]
Fr
218.0845
4,909,068
[3.62045704 e-25 kg]
993,024
[41.088 GeV]
Francium
87
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Rn] 7s1
86
87 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Alkali Metal
Aufbau
Azmithal & Magnetic numbers
0
1
8
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.87 - Francium atom

Copyright ABRAHAM [2008] - All rights reserved

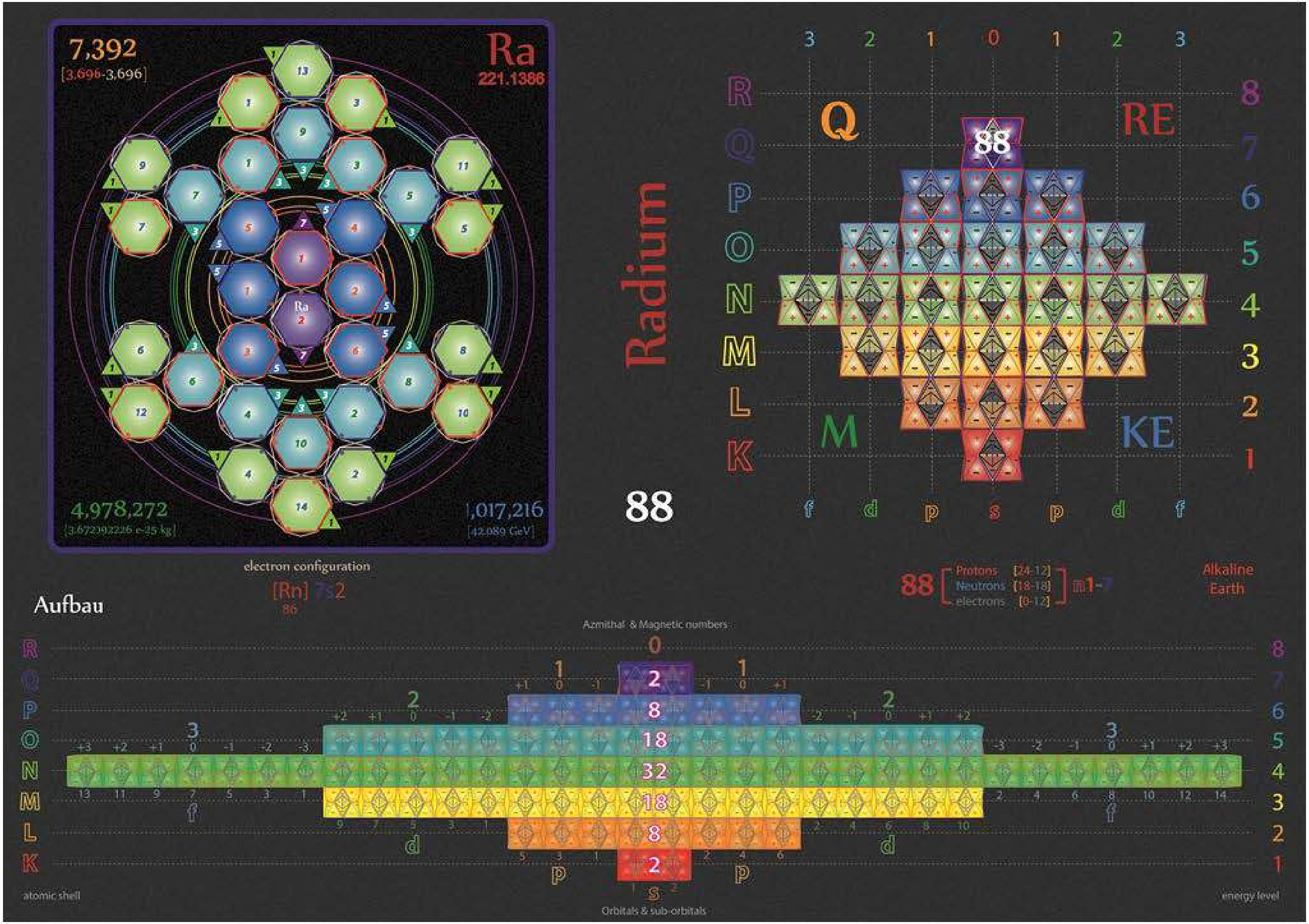
7,392
[3,696-3,696]
Ra
221.1386
4,978,272
[3.672392226 e-25 kg]
1,017,216
[42.089 GeV]
Radium
88
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Rn] 7s2
86
88 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Alkaline Earth
Aufbau
Azimthal & Magnetic numbers
0
1 1
2 2
3 3
2
8
18
32
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.88 - Radium atom

Copyright ABRAHAM [2008] - All rights reserved

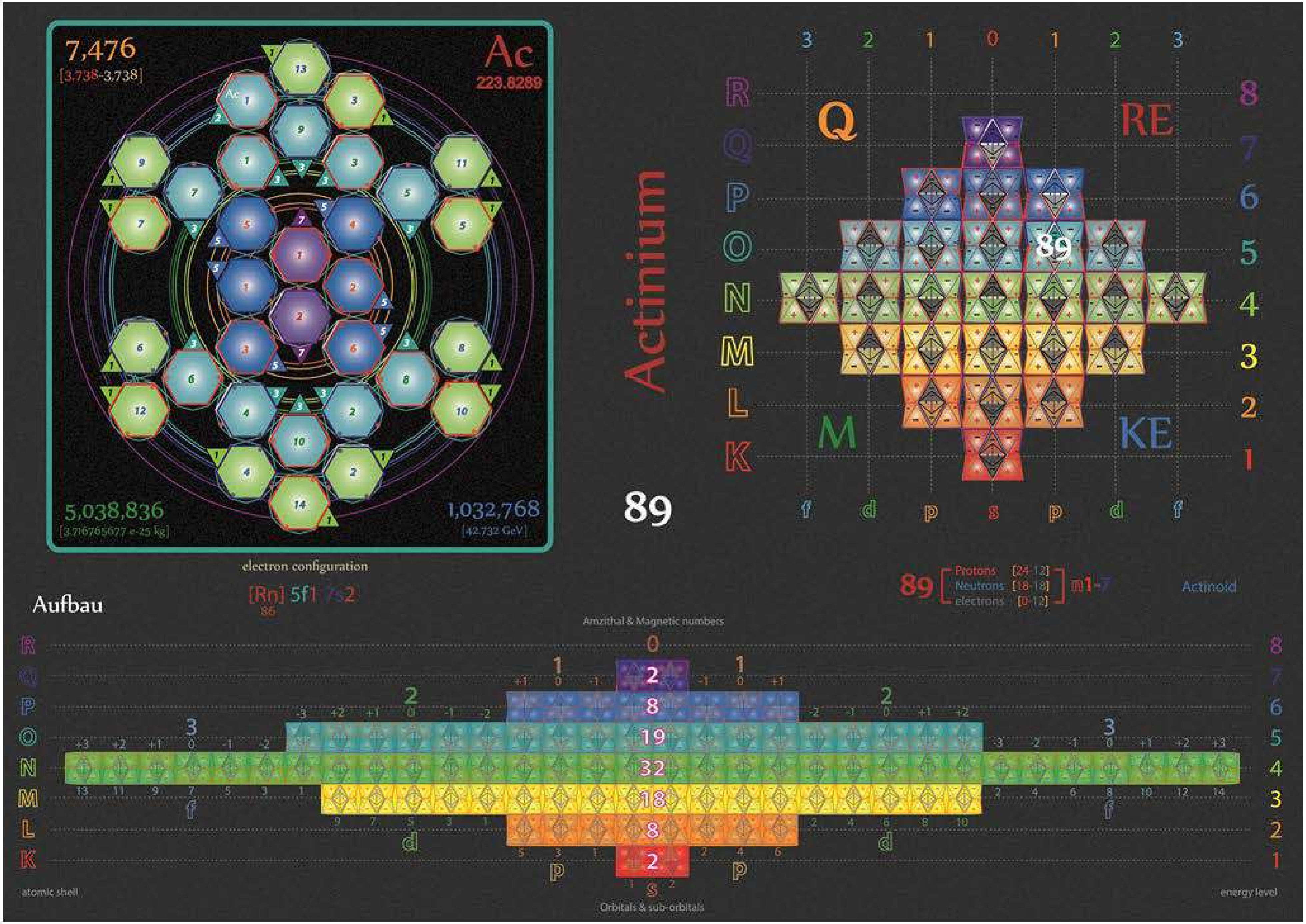
7,476
[3,738-3,738]
Ac
223.8289
5,038,836
[3.716765677 e-25 kg]
1,032,768
[42.732 GeV]
Actinium
89
R Q P O N M L K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Rn] 5f1 7s2
86
Aufbau
89
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Amzithal & Magnetic numbers
2
8
19
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.89 - Actinium atom

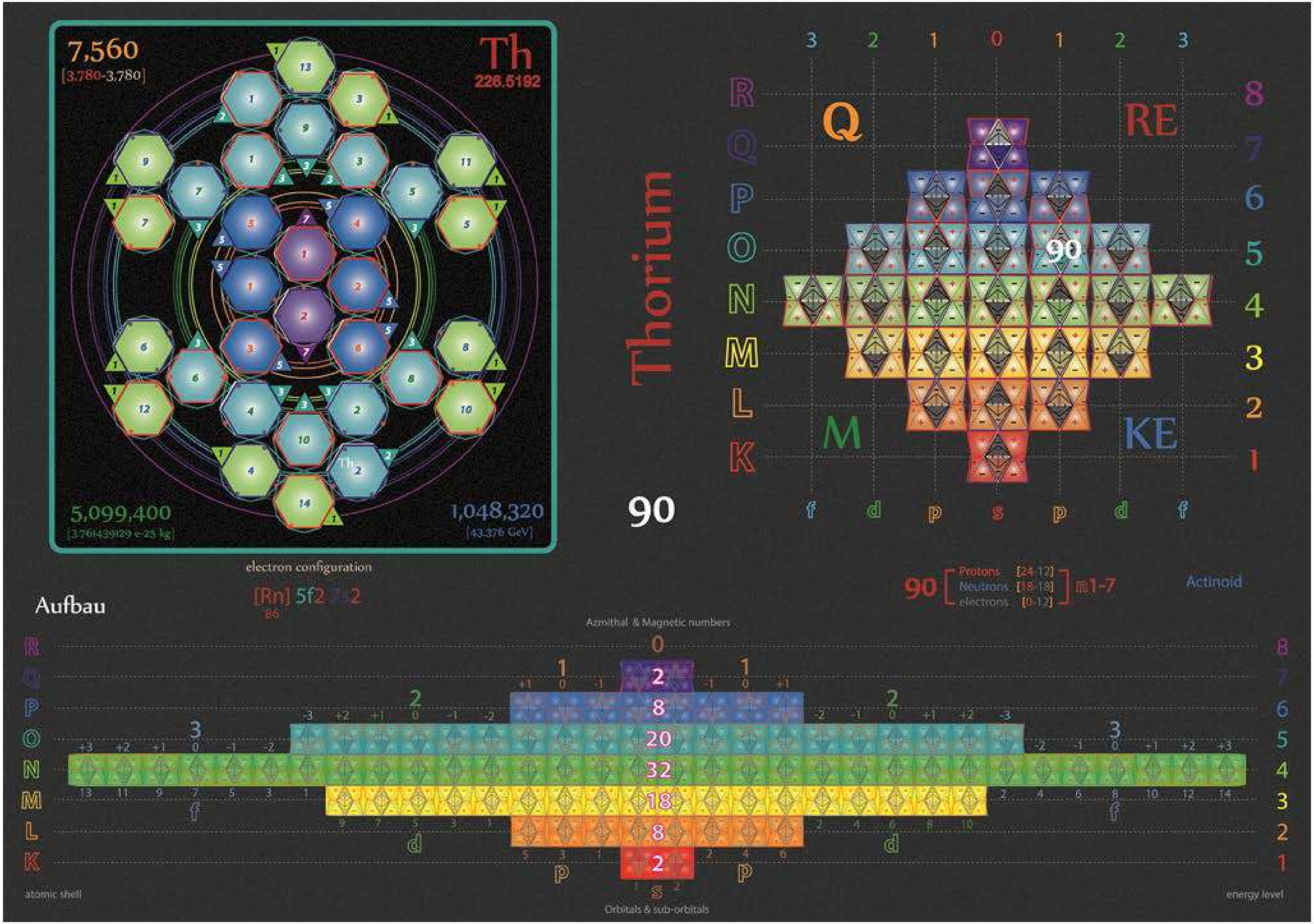
7,560
[3,780-3,780]
Th
226.5192
5,099,400
[3.761439129 e-25 kg]
1,048,320
[43.376 GeV]
Thorium
90
electron configuration
[Rn] 5f2 7s2
86
Aufbau
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
90
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azimuthal & Magnetic numbers
0
1
1
2
2
3
3
2
8
20
32
18
8
2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
2 4 6 8 10 12 14
f
f
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.90 - Thorium atom

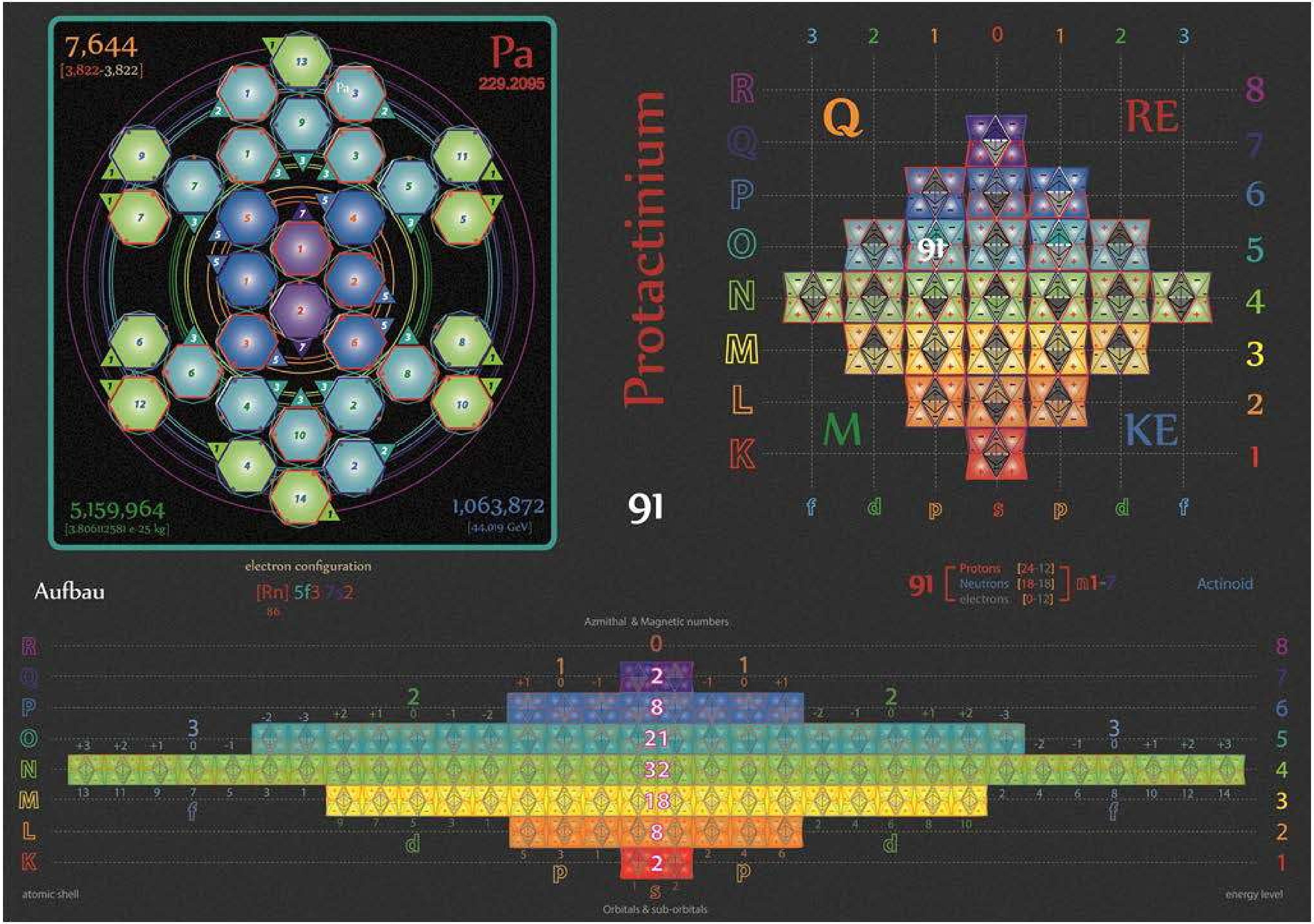
7,644
[3,822-3,822]
Pa
229.2095
5,159,964
[3.806112581 e-25 kg]
1,063,872
[44.019 GeV]
Protactinium
91
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
91 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-7
Actinoid
Aufbau
electron configuration
[Rn] 5f3 7s2
86
Azmithal & Magnetic numbers
0
2
8
21
32
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.91 - Protactinium atom

Copyright ABRAHAM [2008] - All rights reserved

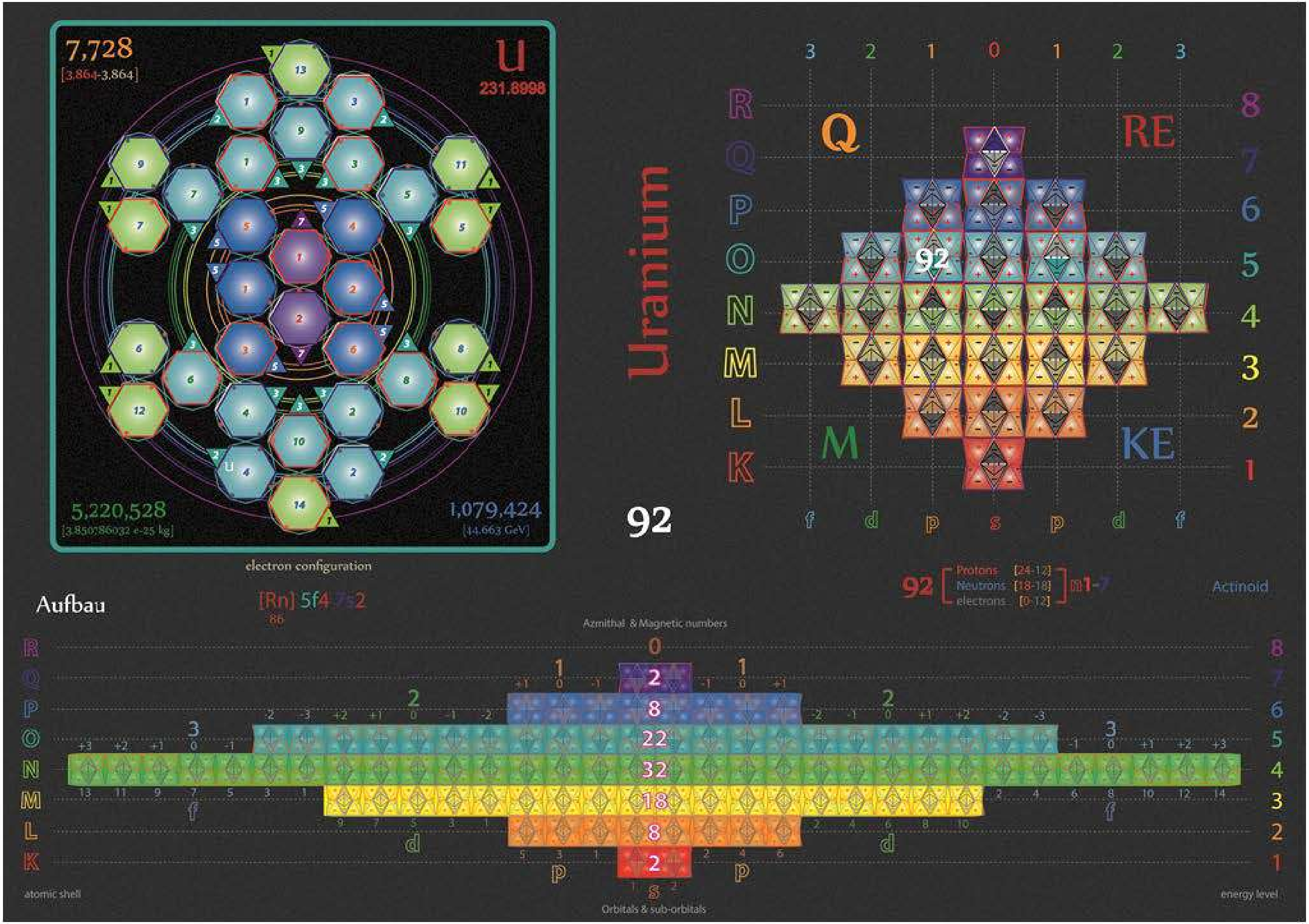
7,728
[3,864-3,864]
U
231.8998
5,220,528
[3.850786032 e-25 kg]
1,079,424
[14.663 GeV]
electron configuration
Uranium
92
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
92 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Aufbau
[Rn] 5f4 7s2
86
Azmithal & Magnetic numbers
2
8
22
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.92 - Uranium atom

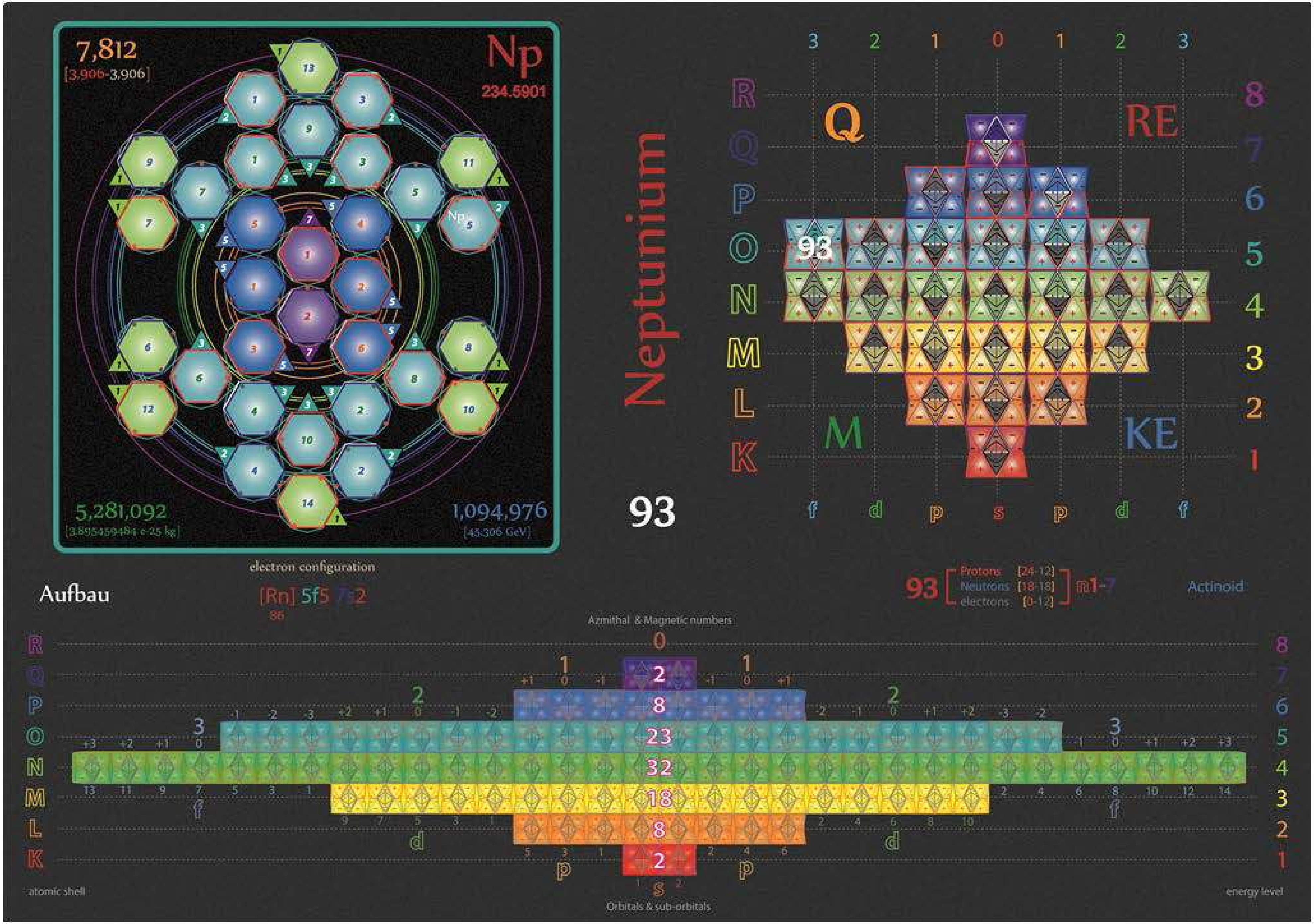
7,812
[3,906-3,906]
Np
234.5901
5,281,092
[3.895459484 e-25 kg]
1,094,976
[45.306 GeV]
Neptunium
93
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f5 7s2
86
93 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
23
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.93 - Neptunium atom

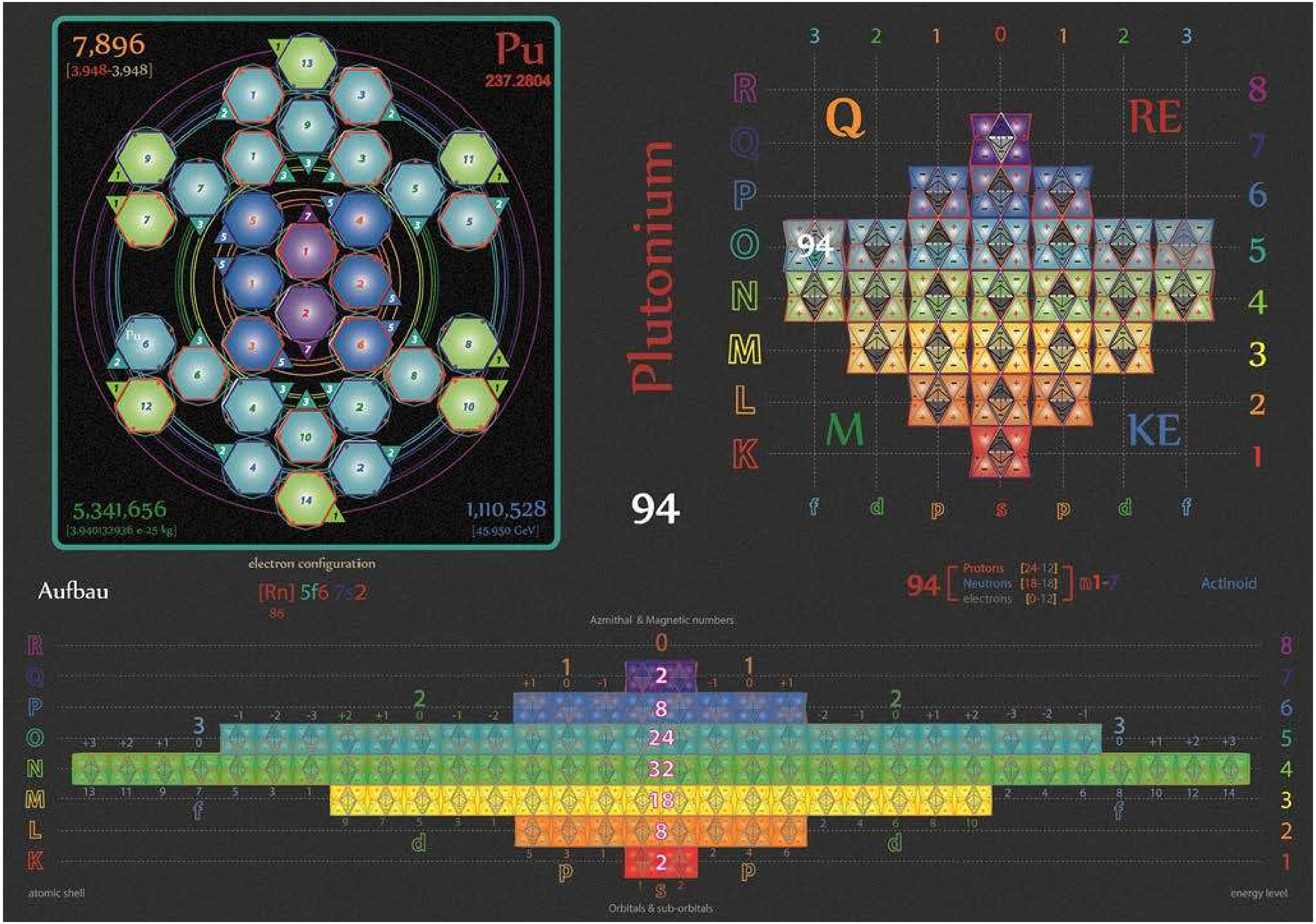
7,896
[3,948-3,948]
Pu
237.2804
5,341,656
[3.940132936 e-25 kg]
1,110,528
[45.950 GeV]
Plutonium
94
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f6 7s2
86
94 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
24
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.94 - Plutonium atom

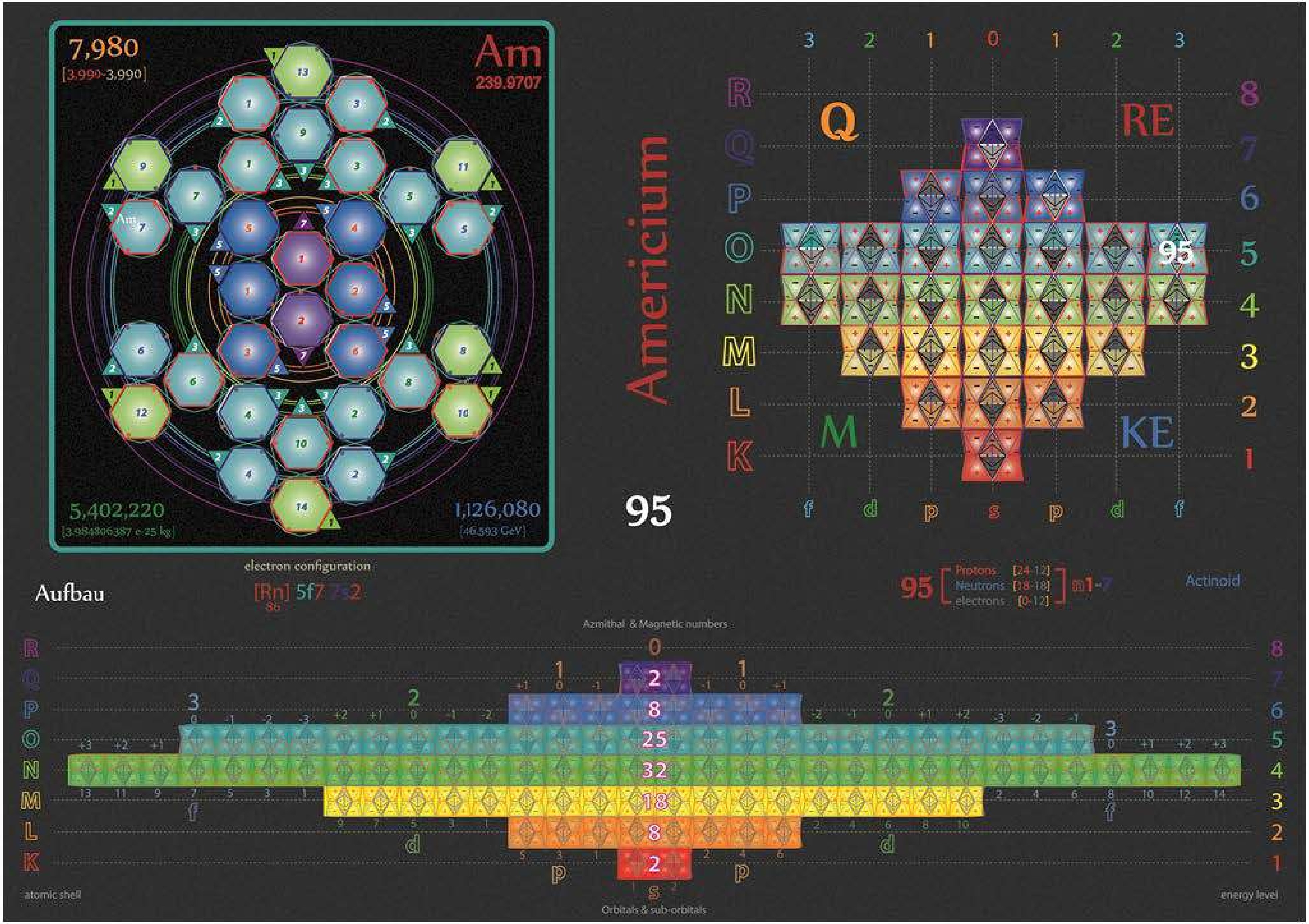
7,980
[3,990-3,990]
Am
239.9707
5,402,220
[3.9848063387 e-25 kg]
1,126,080
[46.593 GeV]
Americium
95
R Q P O N M L K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
95 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Actinoid
Aufbau
electron configuration
[Rn] 5f7 7s2
86
Azmithal & Magnetic numbers
2
8
25
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.95 - Americium atom

Copyright ABRAHAM [2008] - All rights reserved

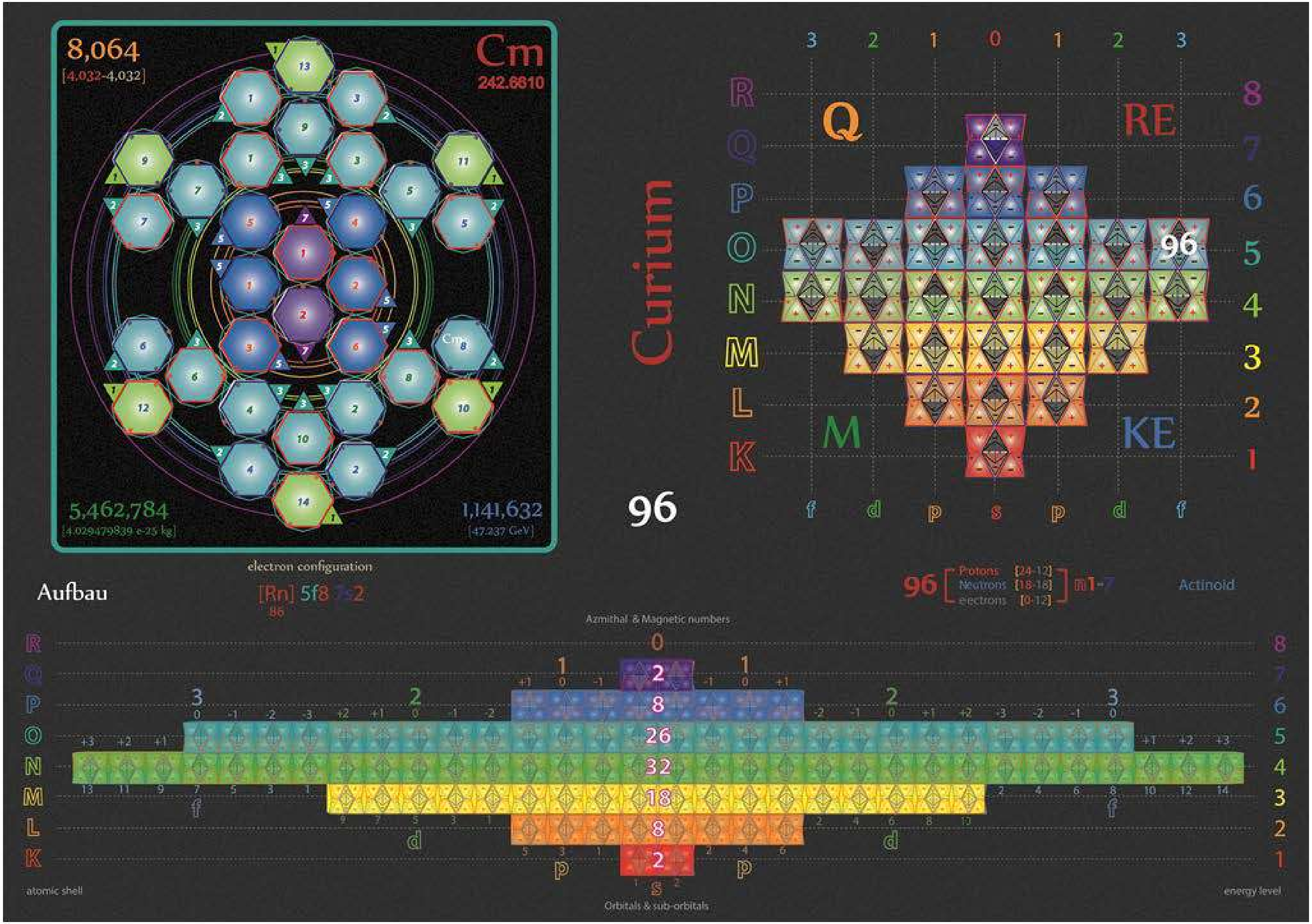
8,064
[4,032-4,032]
Cm
242.6810
5,462,784
[4.029479839 e-25 kg]
1,141,632
[47.237 GeV]
Curium
96
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
Aufbau
electron configuration
[Rn] 5f8 7s2
86
96
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-7
Actinoid
Azmithal & Magnetic numbers
2
8
26
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.96 - Curium atom

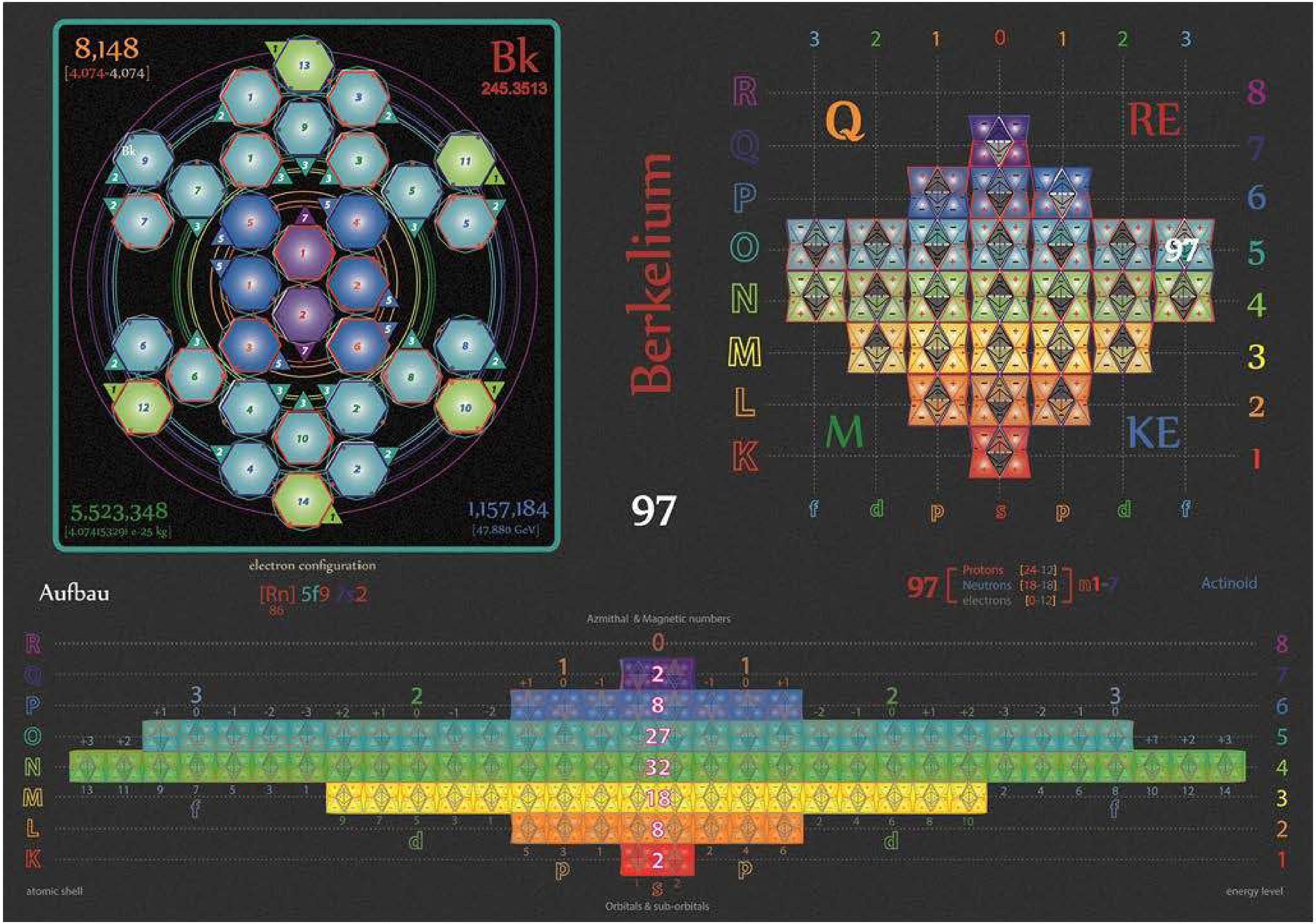
8,148
[4,074-4,074]
Bk
245.3513
5,523,348
[4.07415329I e-25 kg]
1,157,184
[47.880 GeV]
Berkelium
97
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f9 7s2
86
97
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
27
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.97 - Berkelium atom

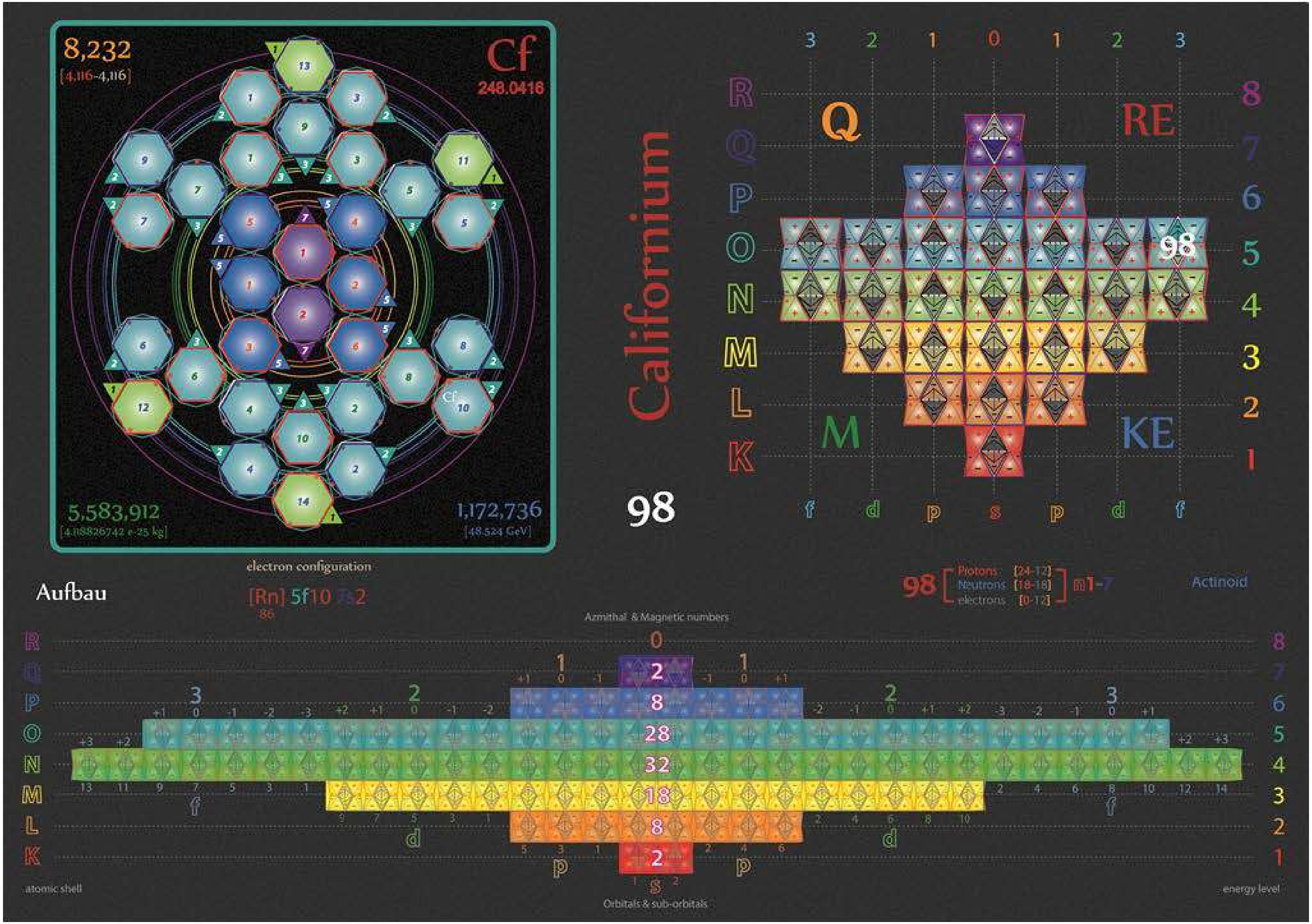
8,232
[4,116-4,116]
Cf
248.0418
5,583,912
[4.118826742 e-25 kg]
1,172,736
[48.524 GeV]
Californium
98
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f10 7s2
86
98
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
28
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.98 - Californium atom

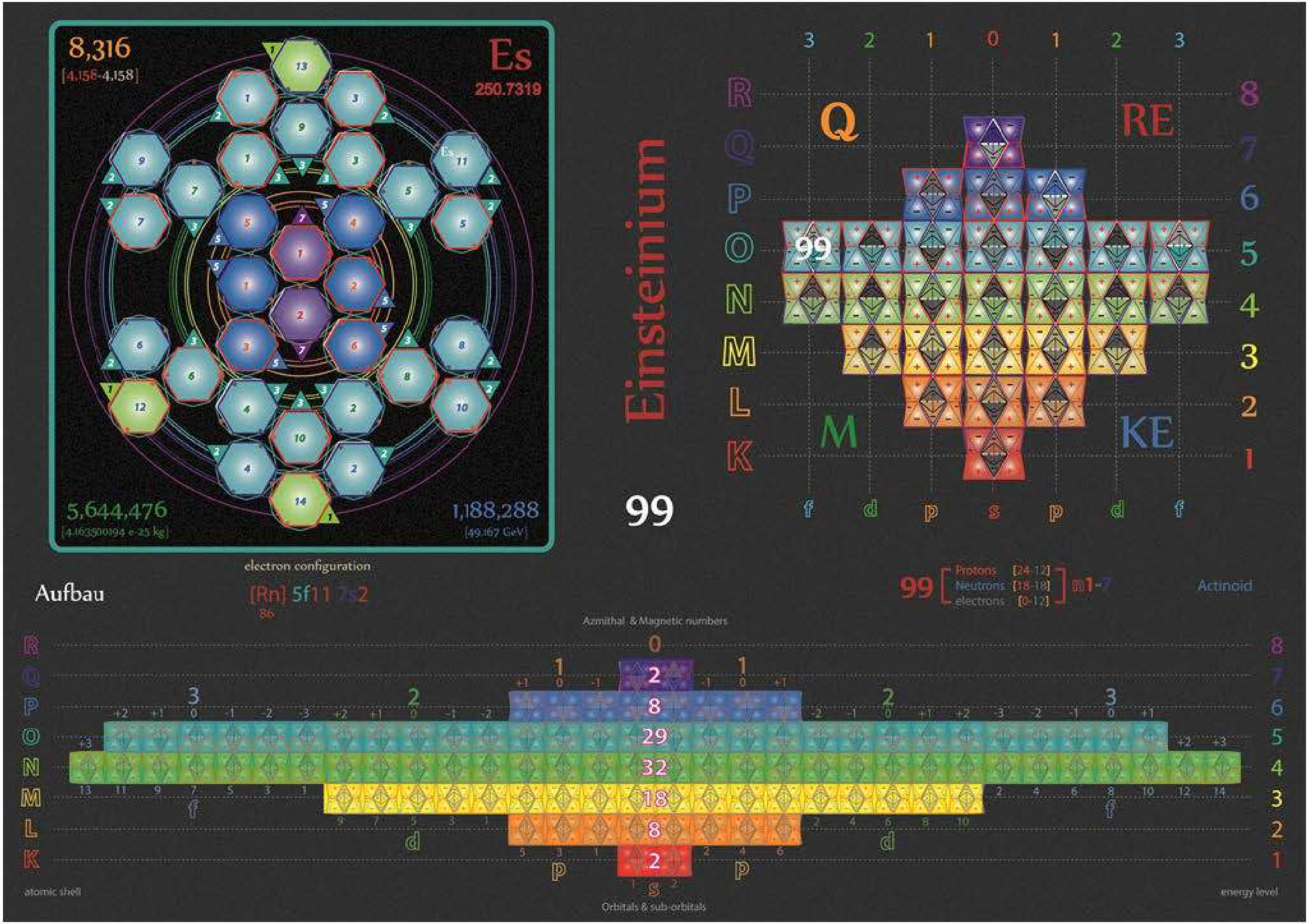
8,316
[4,158-4,158]
Es
250.7319
5,644,476
[4.163500194 e-25 kg]
1,188,288
[49.167 GeV]
Einsteinium
99
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
f
d
p
s
p
d
f
electron configuration
[Rn] 5f11 7s2
86
Aufbau
99 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Actinoid
Azmithal & Magnetic numbers
2
8
29
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.99 - Einsteinium atom

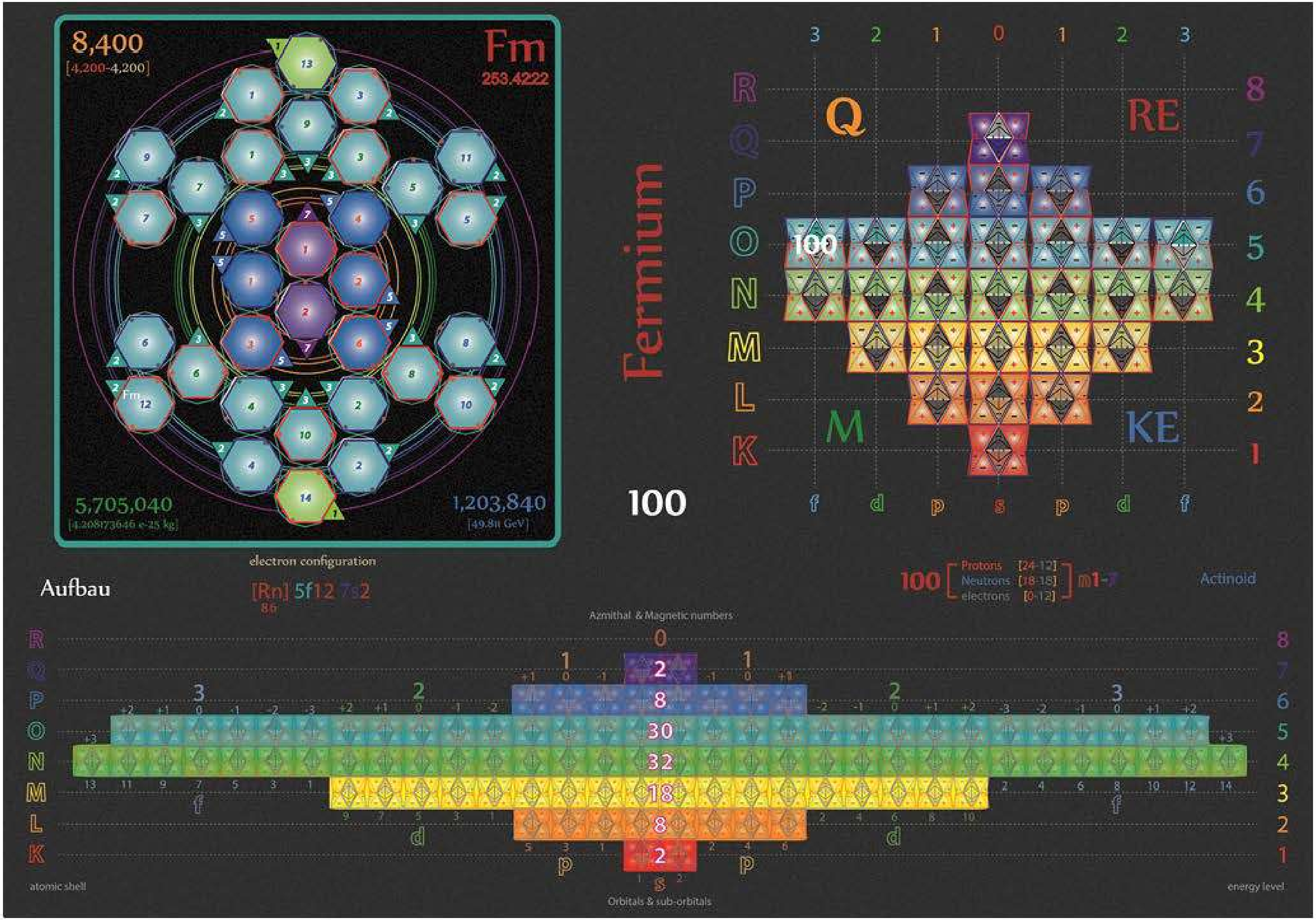
8,400
[4,200-4,200]
Fm
253.4222
5,705,040
[4.208173646 e-25 kg]
1,203,840
[49.811 GeV]
Fermium
100
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f12 7s2
86
100 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
30
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.100 - Fermium atom

Copyright ABRAHAM [2008] - All rights reserved

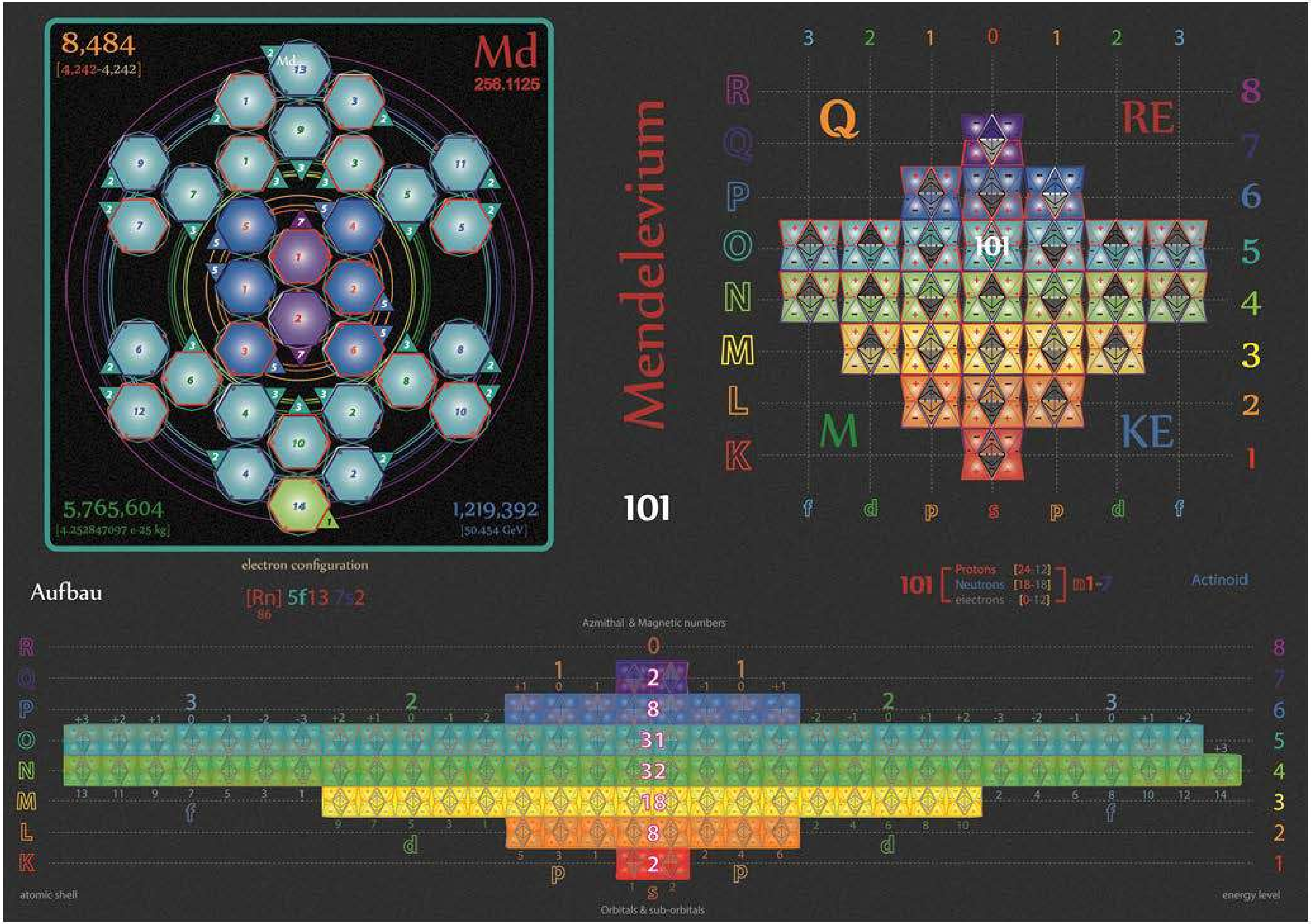
8,484
[4,242-4,242]
Md
258.1125
5,765,604
[4.252847097 e-25 kg]
1,219,392
[50.454 GeV]
Mendelevium
101
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f13 7s2
86
101
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
31
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.101 - Mendelevium atom

Copyright ABRAHAM [2008] - All rights reserved

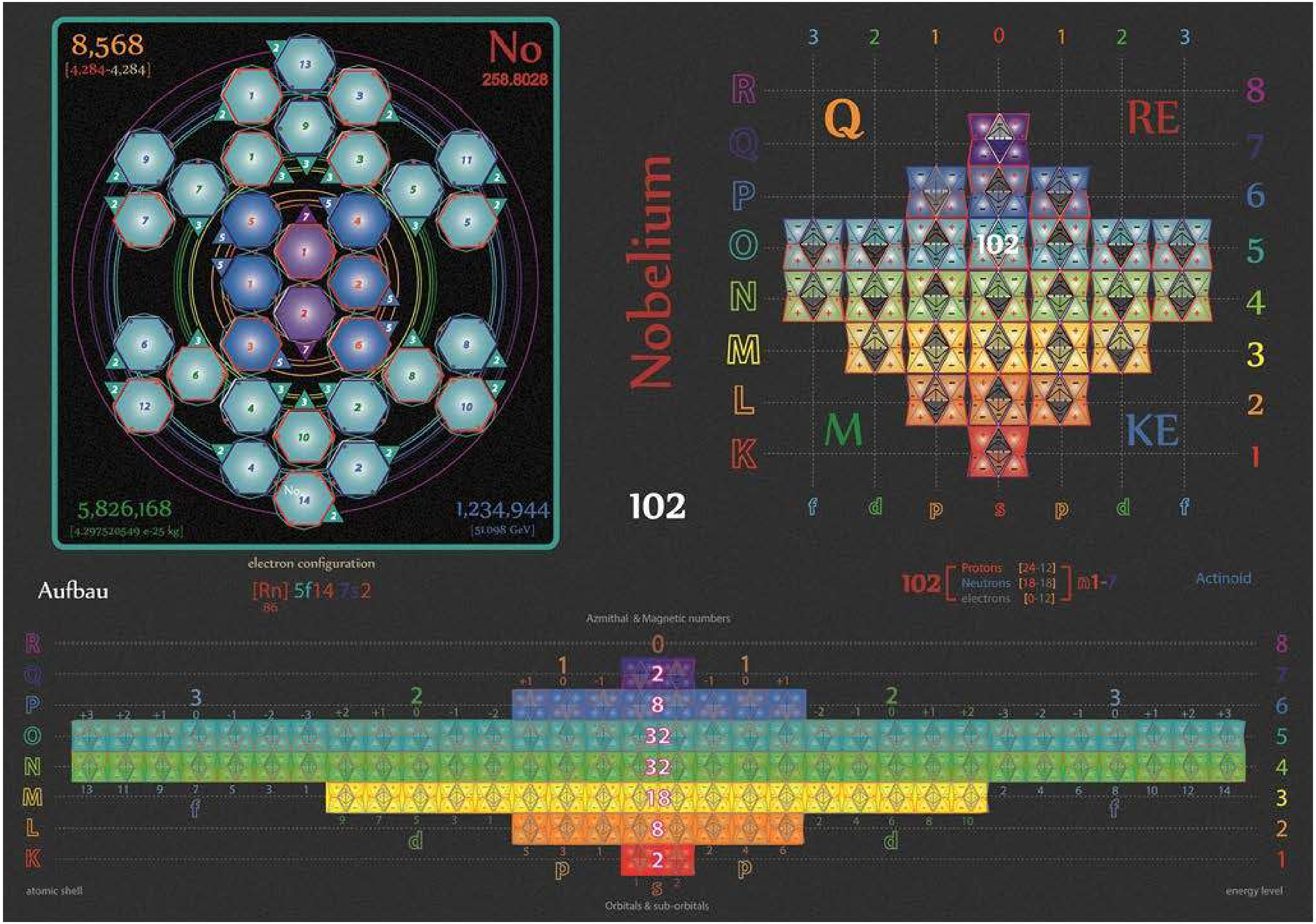
8,568
[4,284-4,284]
No
258.8028
5,826,168
[4.297520549 e-25 kg]
1,234,944
[51.098 GeV]
Nobelium
102
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
102
102 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-7
Actinoid
electron configuration
Aufbau
[Rn] 5f14 7s2
86
Azmithal & Magnetic numbers
2
8
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.102 - Nobelium atom

Copyright ABRAHAM [2008] - All rights reserved

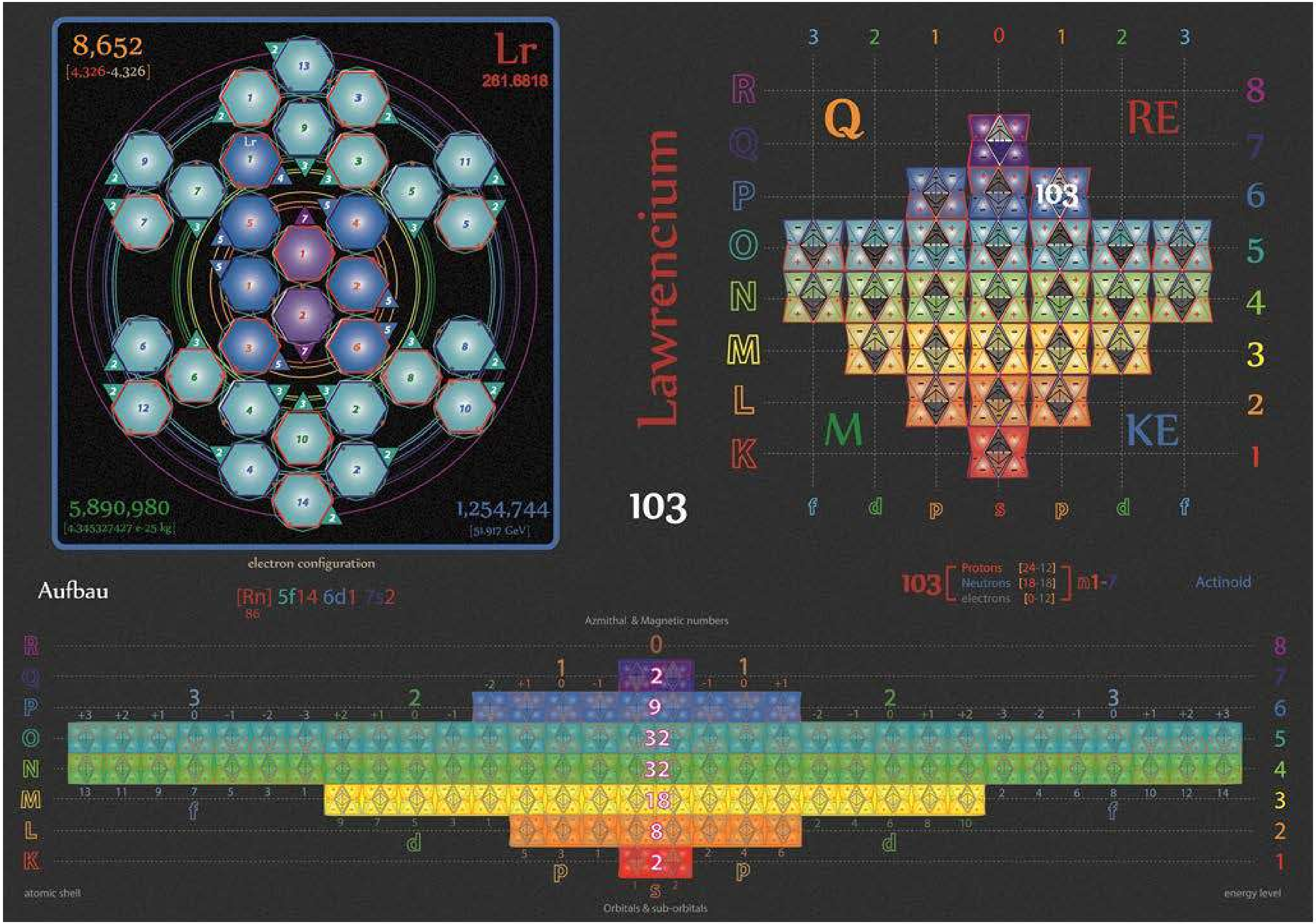
8,652
[4,326-4,326]
Lr
261.6818
5,890,980
[4.345327427 e-25 kg]
1,254,744
[51.917 GeV]
electron configuration
Lawrencium
103
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
103 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Aufbau
[Rn] 5f14 6d1 7s2
86
Azmithal & Magnetic numbers
0
1
2
3
2
9
32
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.103 - Lawrencium atom

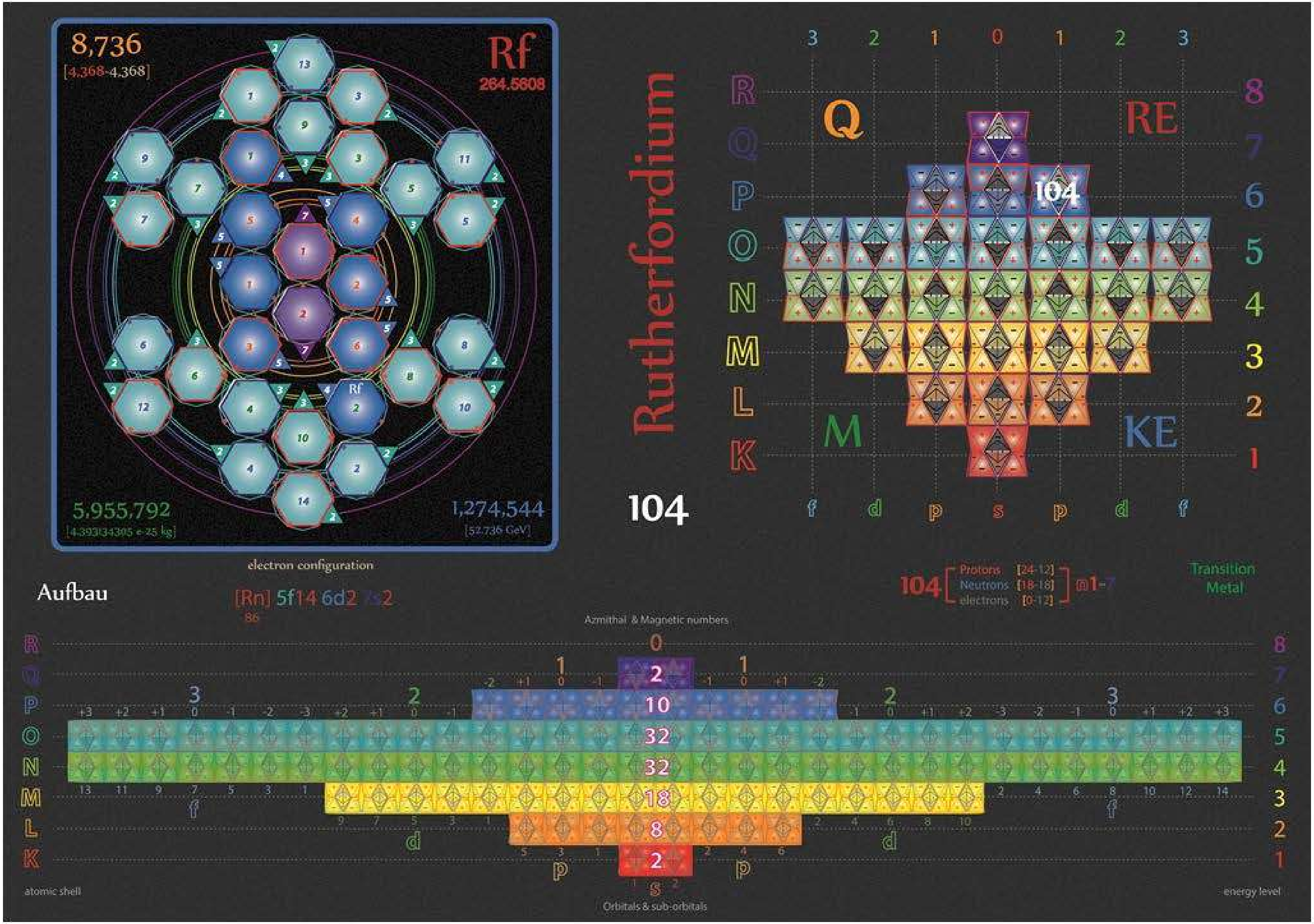
8,736
[4,368-4,368]
Rf
264.5608
5,955,792
[4.393134305 e-25 kg]
1,274,544
[52.736 GeV]
Rutherfordium
104
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
Aufbau
[Rn] 5f14 6d2 7s2
86
104 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Transition Metal
Azimthal & Magnetic numbers
2
10
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.104 - Rutherfordium atom

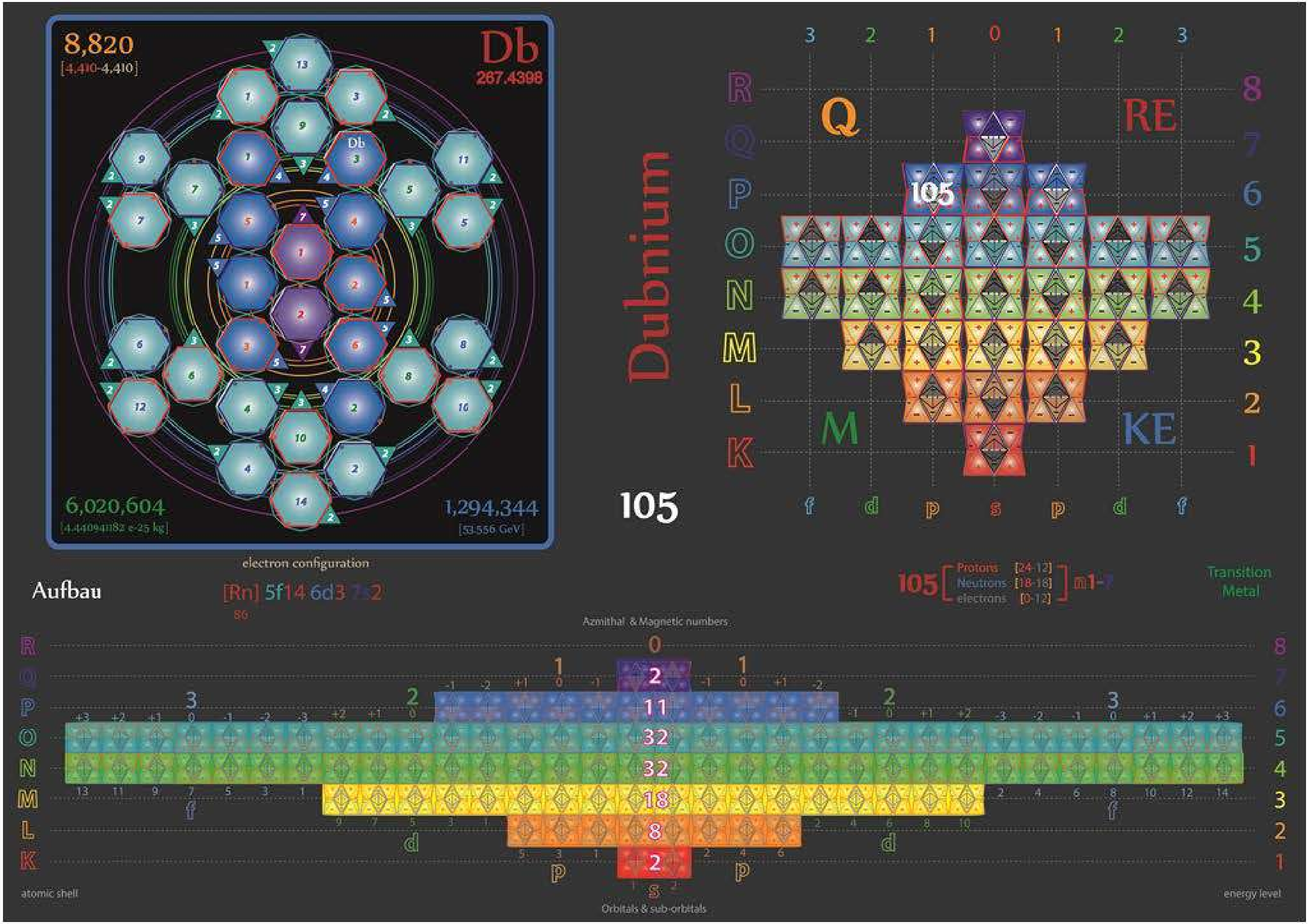
8,820
[4,410-4,410]
Db
287.4398
6,020,604
[4.44094182 e-25 kg]
1,294,344
[53.556 GeV]
Dubnium
105
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
Aufbau
[Rn] 5f14 6d3 7s2
86
105 [Protons [24-12] Neutons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
0
2
11
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.105 - Dubnium atom

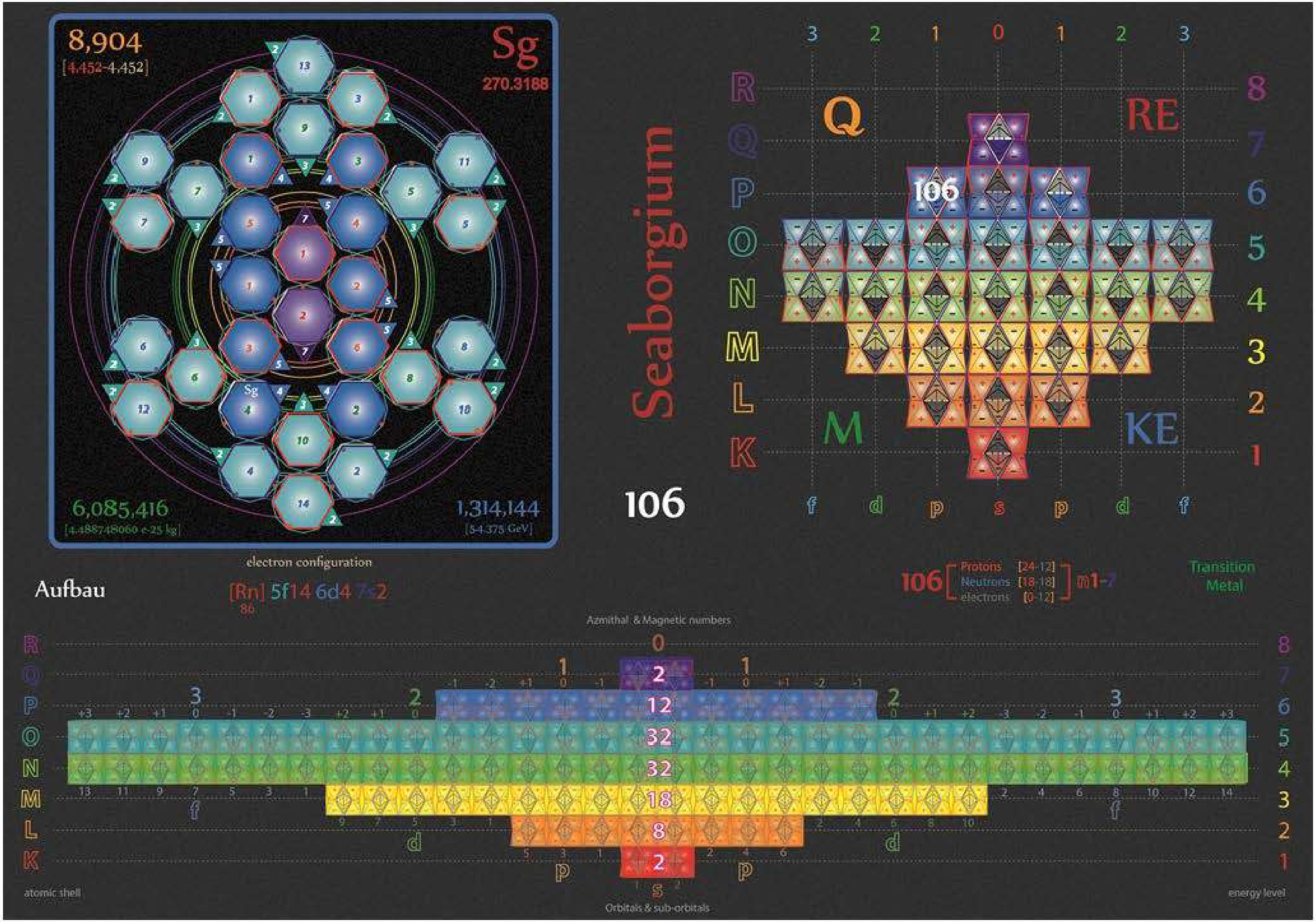
8,904
[4,452-4,452]
Sg
270.3188
6,085,416
[4.488748060 e-25 kg]
1,314,144
[54.378 GeV]
Seaborgium
106
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d4 7s2
86
106
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Transition Metal
Azimthal & Magnetic numbers
2
12
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.106 - Seaborgium atom

Copyright ABRAHAM [2008] - All rights reserved

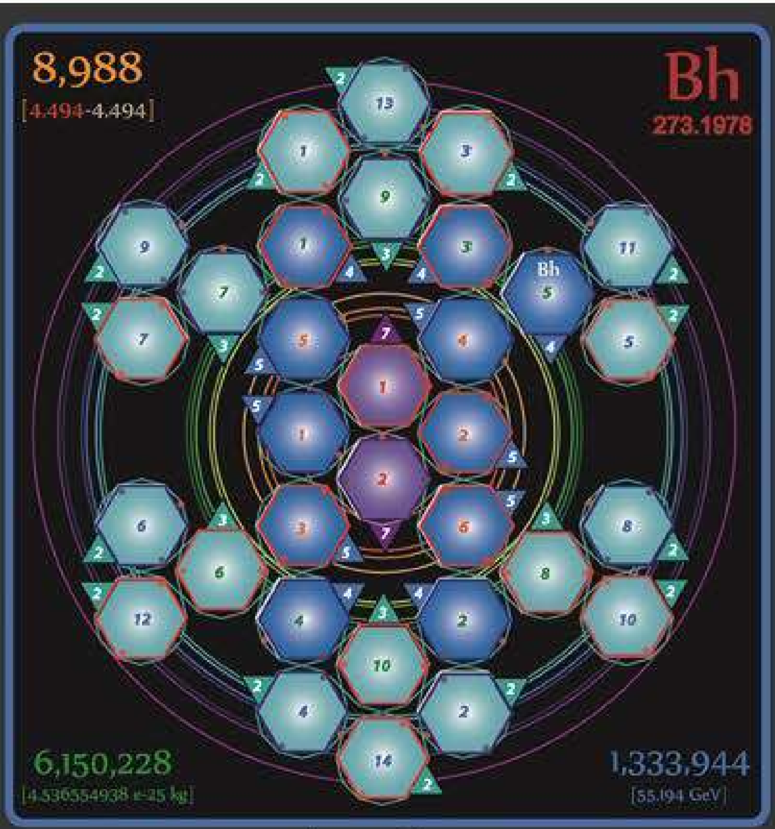
8,988
[4.494-4.494]
Bh
273.1978
6,150,228
[4.536554938 e-25 kg]
1,333,944
[55.194 GeV]

Bohrium

107

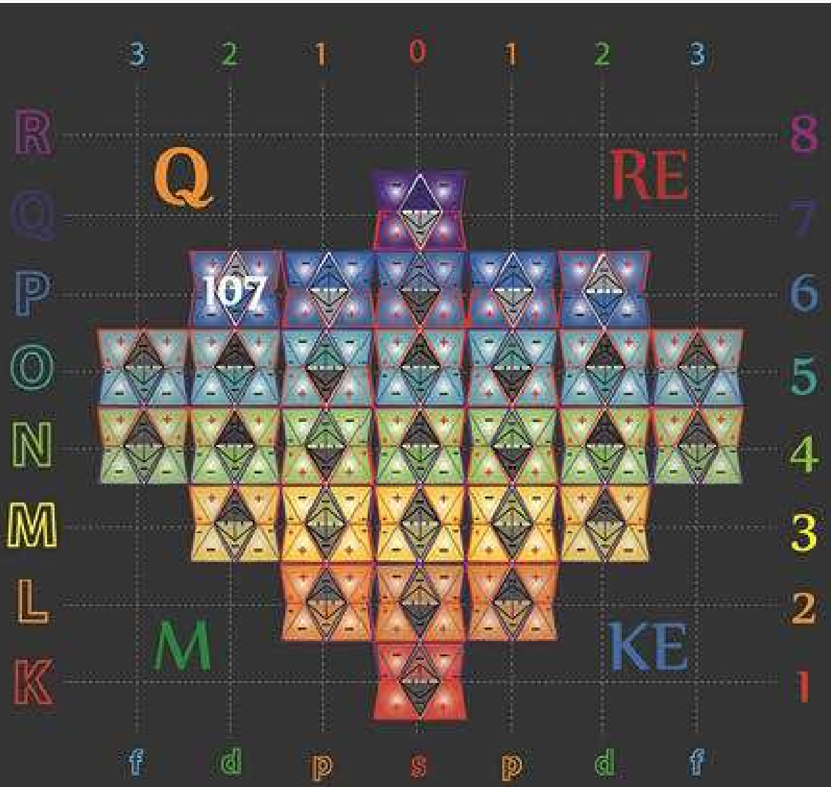
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
107
f d p s p d f

electron configuration

Aufbau

[Rn] 5f14 6d5 7s2
86

107 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7

Transition Metal

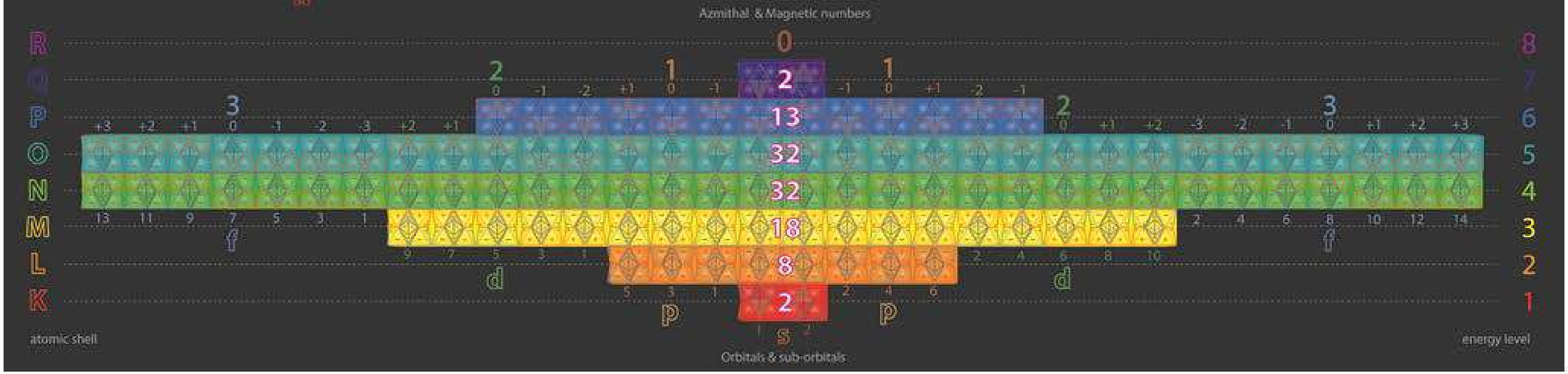
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
13
32
32
18
8
2
3 2 1 1 2 3
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.107 - Bohrium atom

Copyright ABRAHAM [2008] - All rights reserved

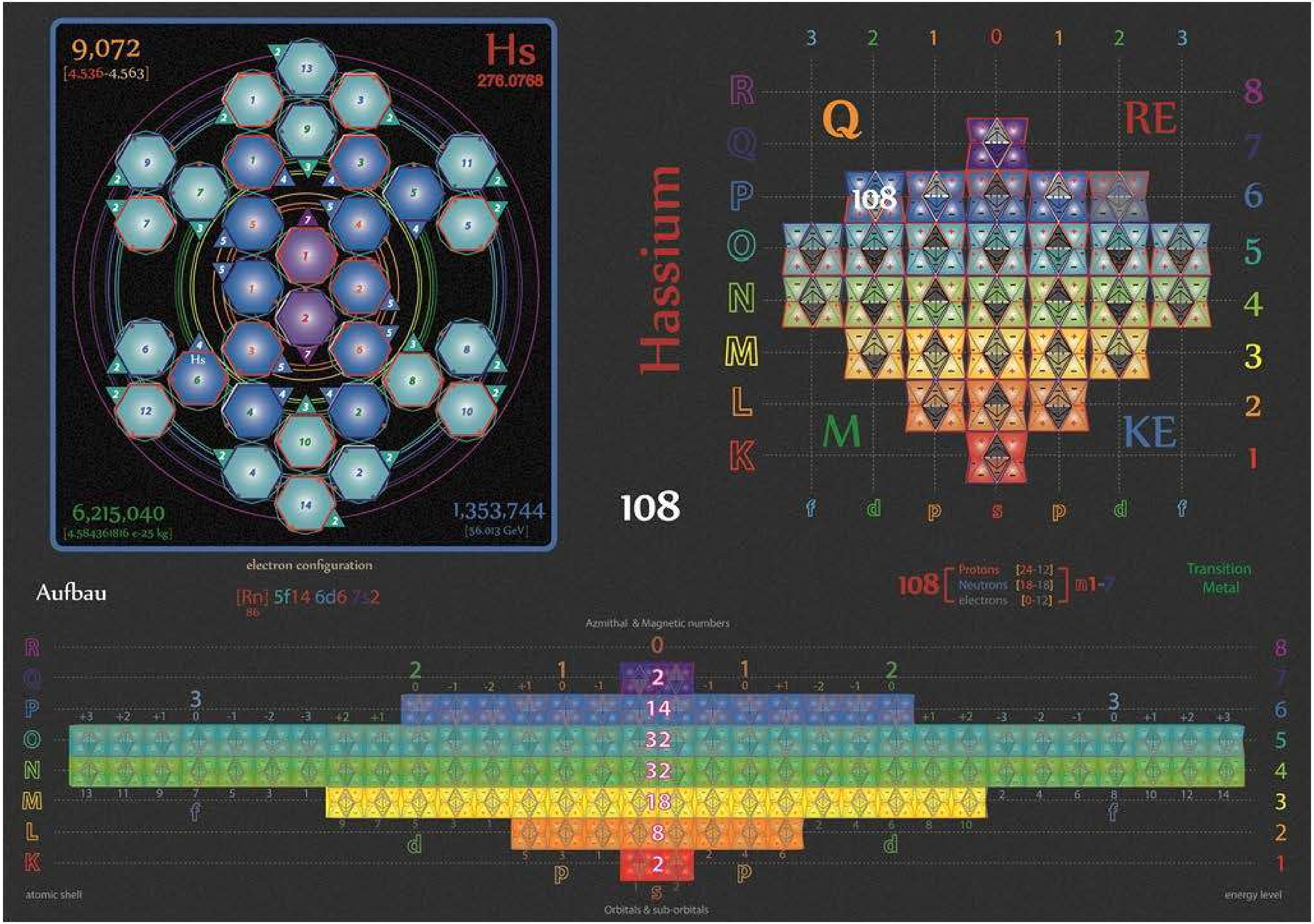
9,072
[4.536-4.563]
Hs
276.0768
6,215,040
[4.584361816 e-25 kg]
1,353,744
[56.013 GeV]
Hassium
108
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d6 7s2
86
108 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
0
2 14 32 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.108 - Hassium atom

Copyright ABRAHAM [2008] - All rights reserved

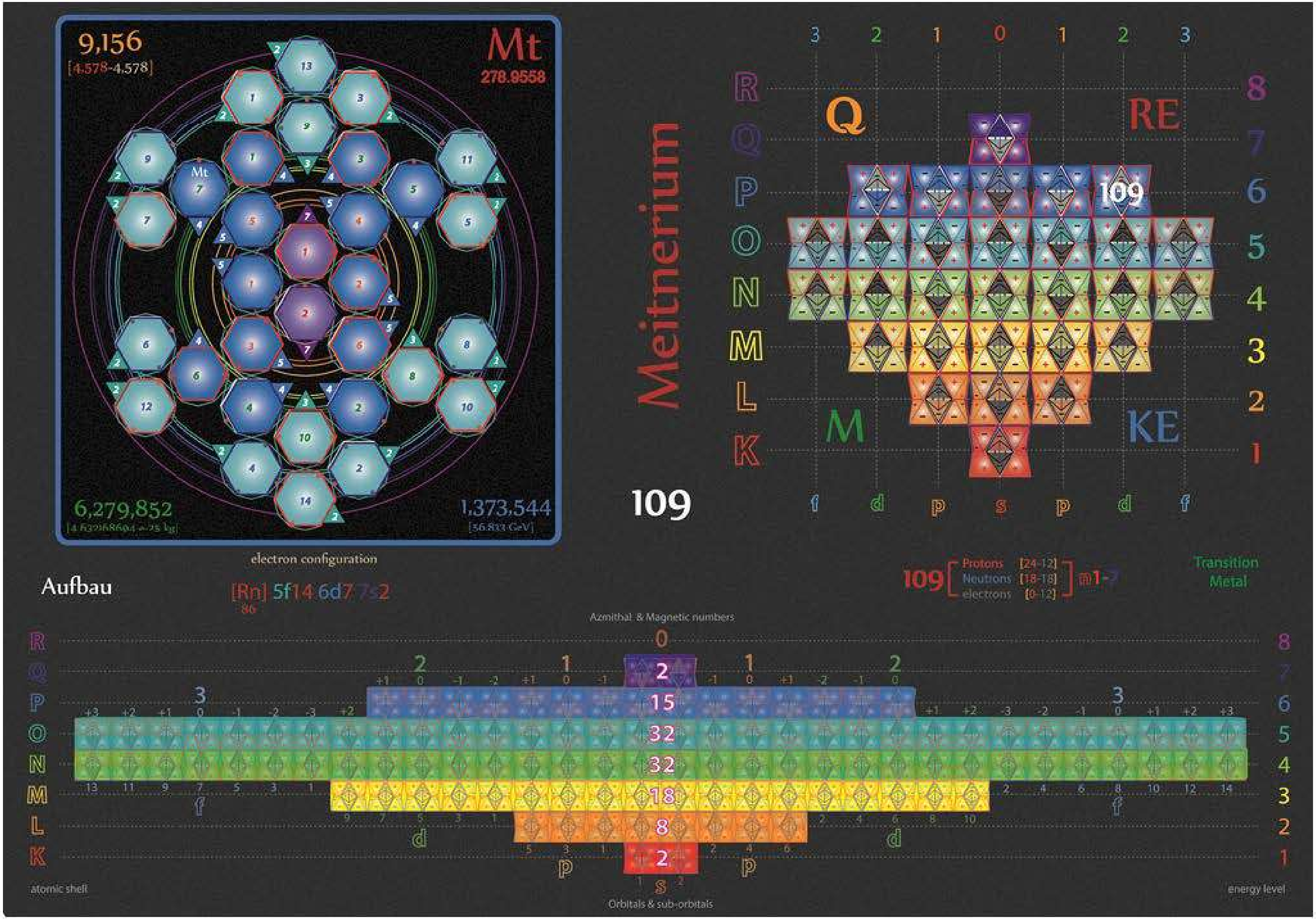
9,156
[4.578-4.578]
Mt
278.9558
6,279,852
[4.632968604 e-25 kg]
1,373,544
[56.833 GeV]
electron configuration
Aufbau
[Rn] 5f14 6d7 7s2
86
Meitnerium
109
R
Q
P
O
N
M
L
K
Q
RE
M
KE
109
3 2 1 0 1 2 3
f d p s p d f
8
7
6
5
4
3
2
1
109
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
@1-7
Transition Metal
Azmithal & Magnetic numbers
2
15
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.109 - Meitnerium atom

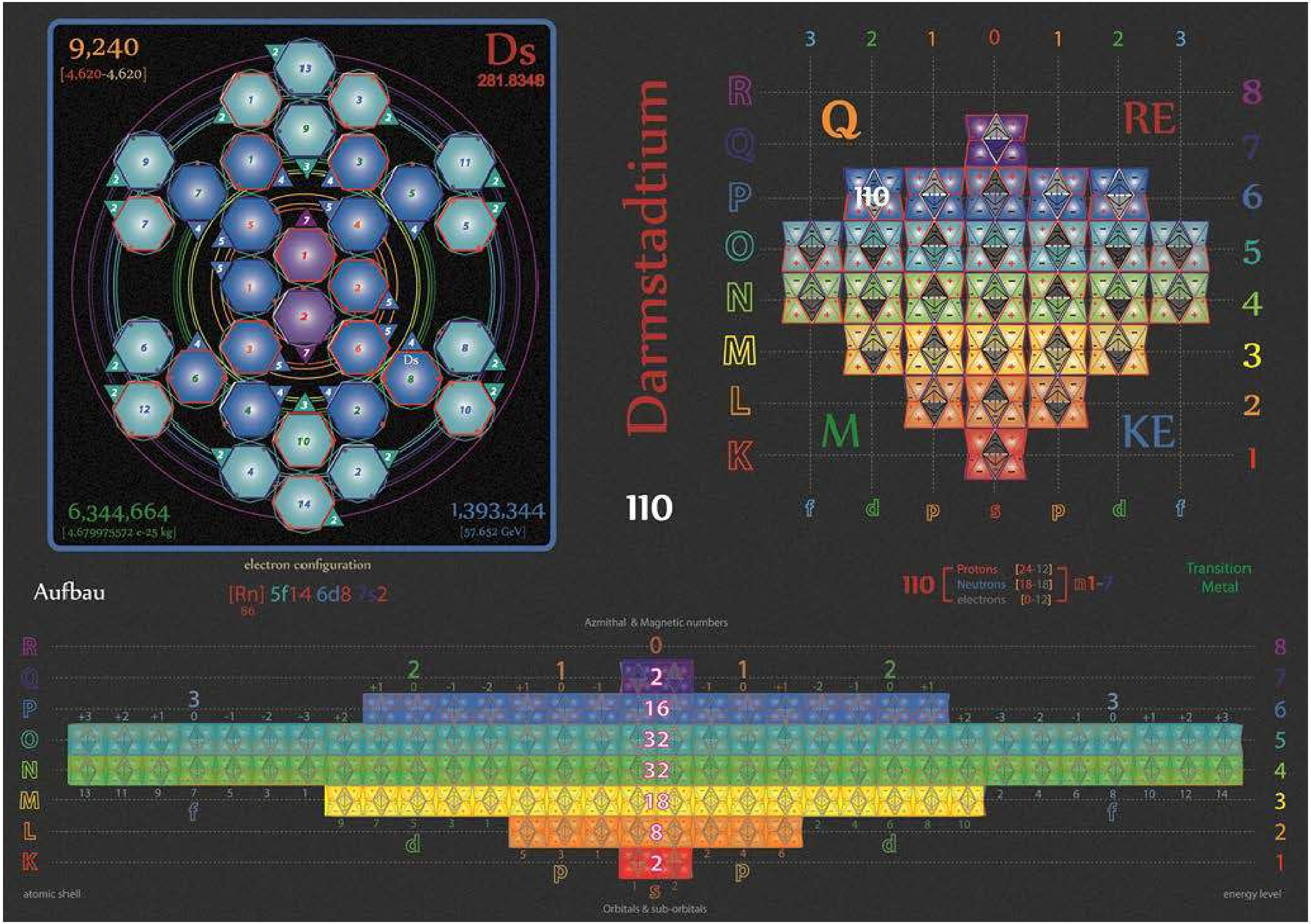
9,240
[4,620-4,620]
Ds
281.8348
6,344,664
[4.679975572 e-25 kg]
1,393,344
[57.652 GeV]
Darmstadtium
110
R
Q
P
O
N
M
L
K
3
2
1
0
1
2
3
8
7
6
5
4
3
2
1
Q
RE
M
KE
f
d
p
s
p
d
f
electron configuration
Aufbau
[Rn] 5f14 6d8 7s2
86
110
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Transition Metal
Azmithal & Magnetic numbers
2
16
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.110 - Darmstadtium atom

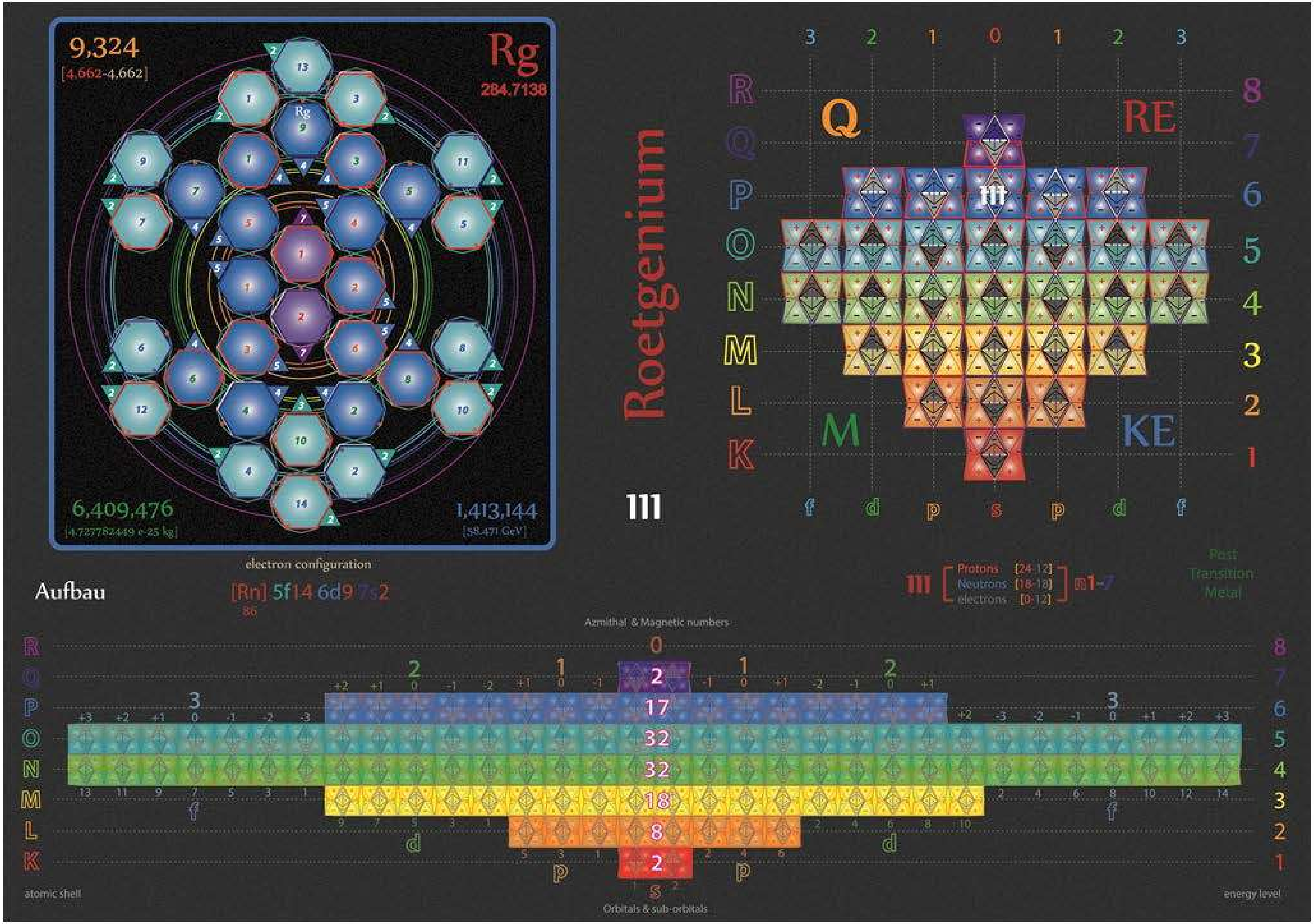
9,324
[4,662-4,662]
Rg
284.7138
6,409,476
[4.727782449 e-25 kg]
1,413,144
[58.471 GeV]
Roetgenium
111
electron configuration
Aufbau
[Rn] 5f14 6d9 7s2
86
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
111
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
1-7
Post Transition Metal
Azmithal & Magnetic numbers
2
17
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.111 - Roetgenium atom

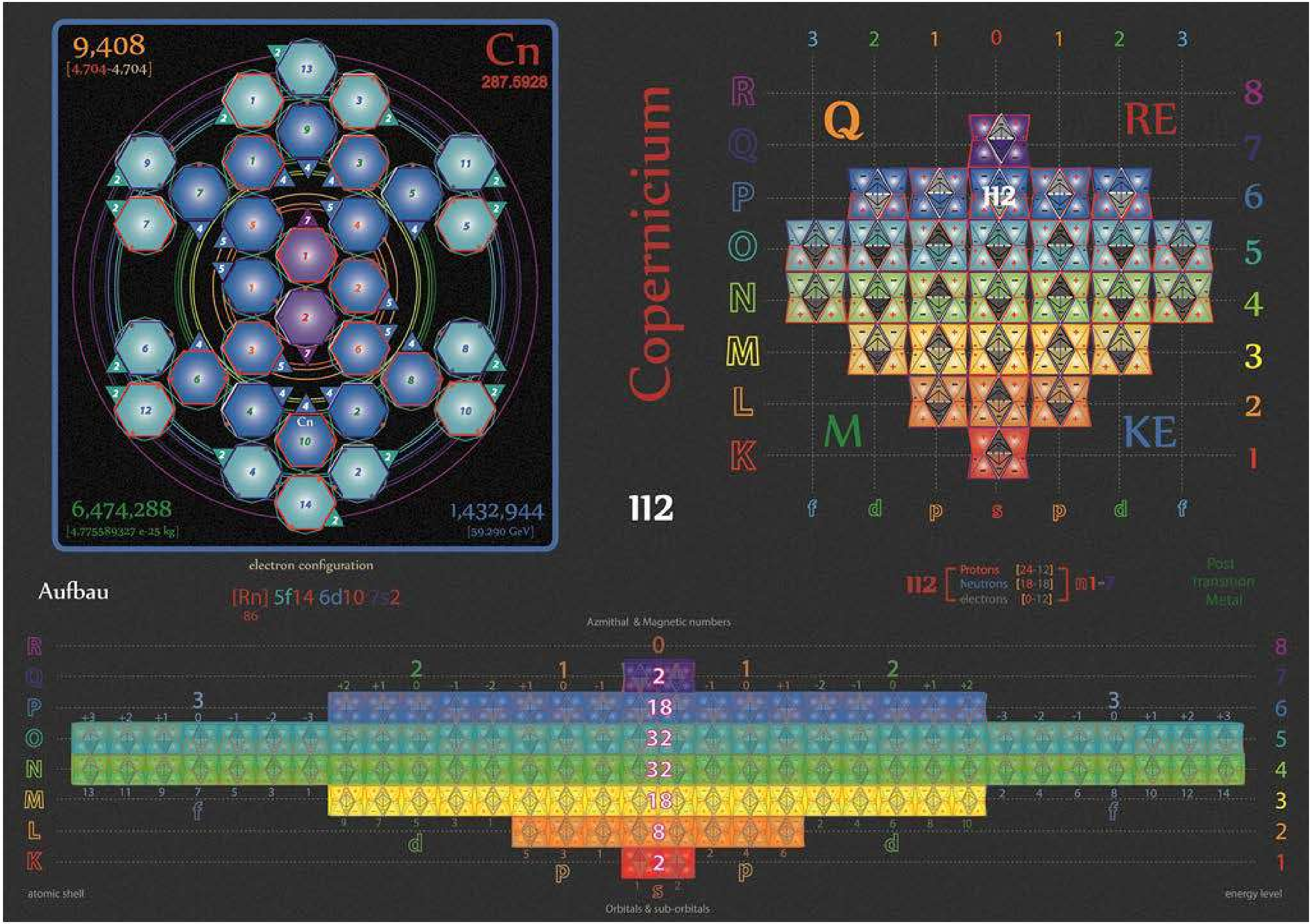
9,408
[4,704-4,704]
Cn
287.5928
6,474,288
[4.775589327 e-25 kg]
1,432,944
[59.290 GeV]
Copernicium
112
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2
86
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
112 Protons [24-12] Neutrons [18-18] electrons [0-12] n1-7
Post Transition Metal
Azmithal & Magnetic numbers
2
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.112 - Copernicium atom

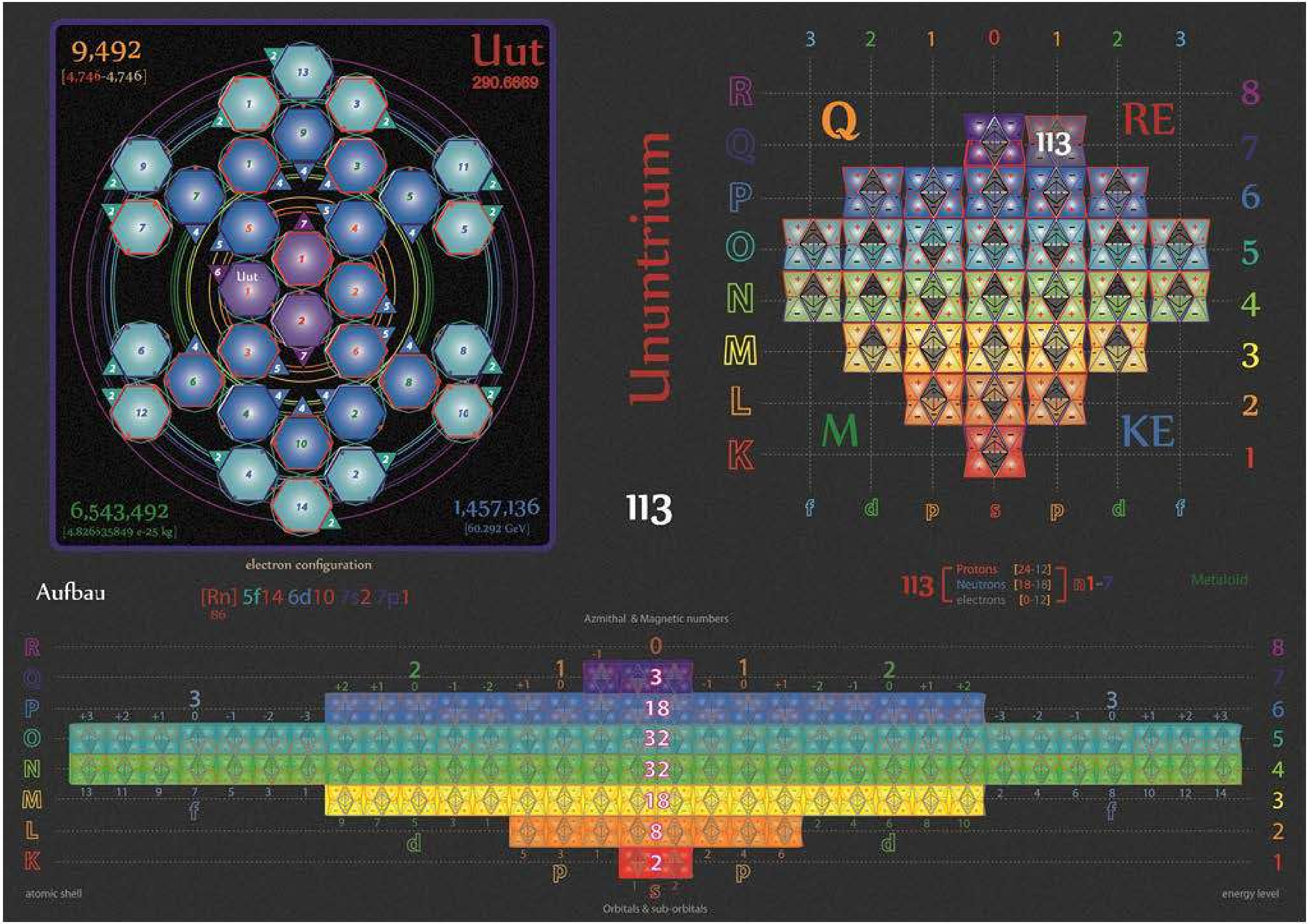
9,492
[4,746-4,746]
Uut
280.6669
Uut
6,543,492
[4.82653849 e-25 kg]
1,457,136
[60.292 GeV]
Ununtrium
113
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
113
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p1
86
113
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-7
Metalloid
Azmithal & Magnetic numbers
3
18
32
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.113 - Ununtrium atom

Copyright ABRAHAM [2008] - All rights reserved

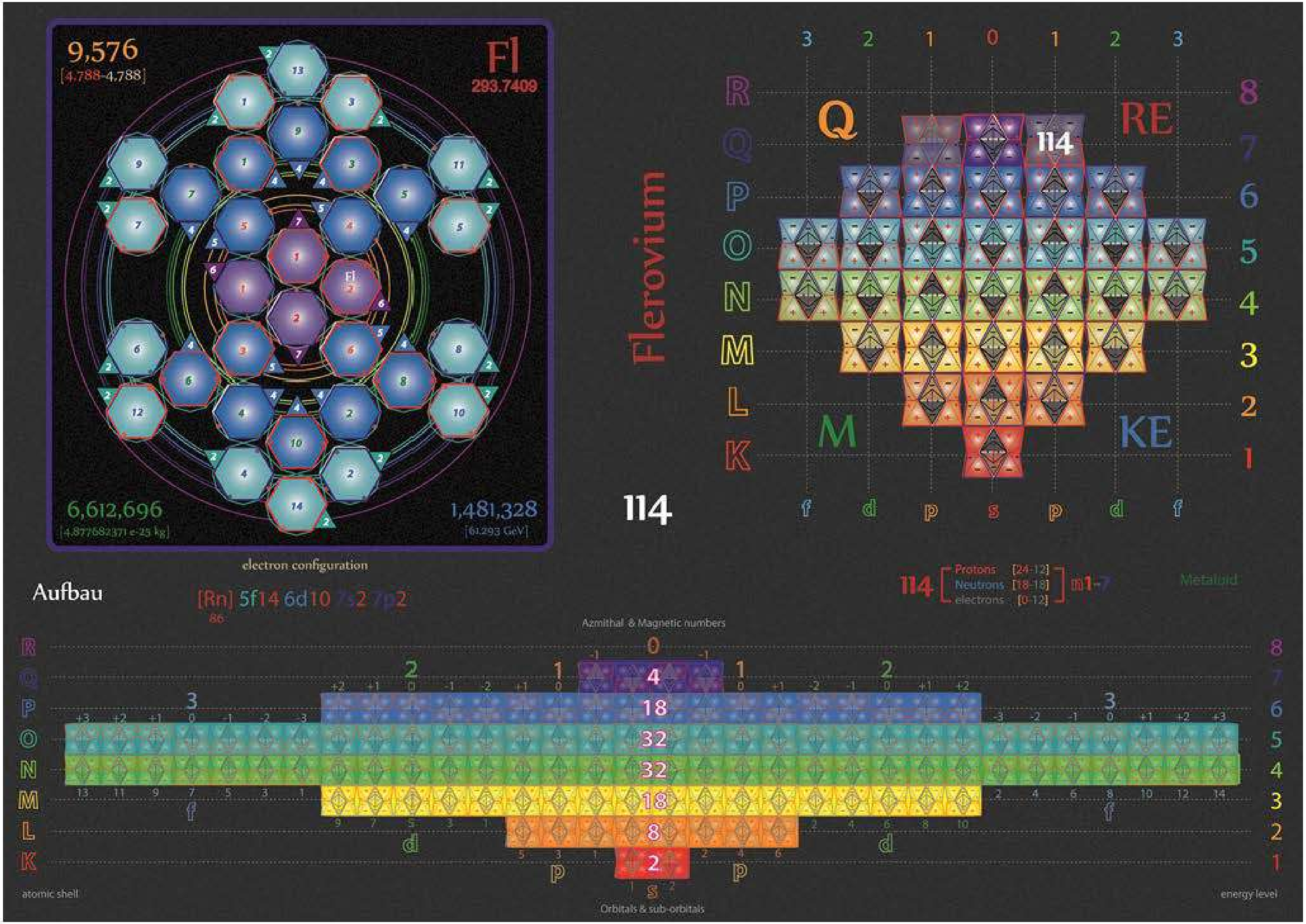
9,576
[4,788-4,788]
Fl
293.7409
6,612,696
[4.877682371 e-25 kg]
1,481,328
[61.293 GeV]
Flerovium
114
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p2
86
114
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Metalloid
Azmithal & Magnetic numbers
0
4
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.114 - Flerovium atom

Copyright ABRAHAM [2008] - All rights reserved

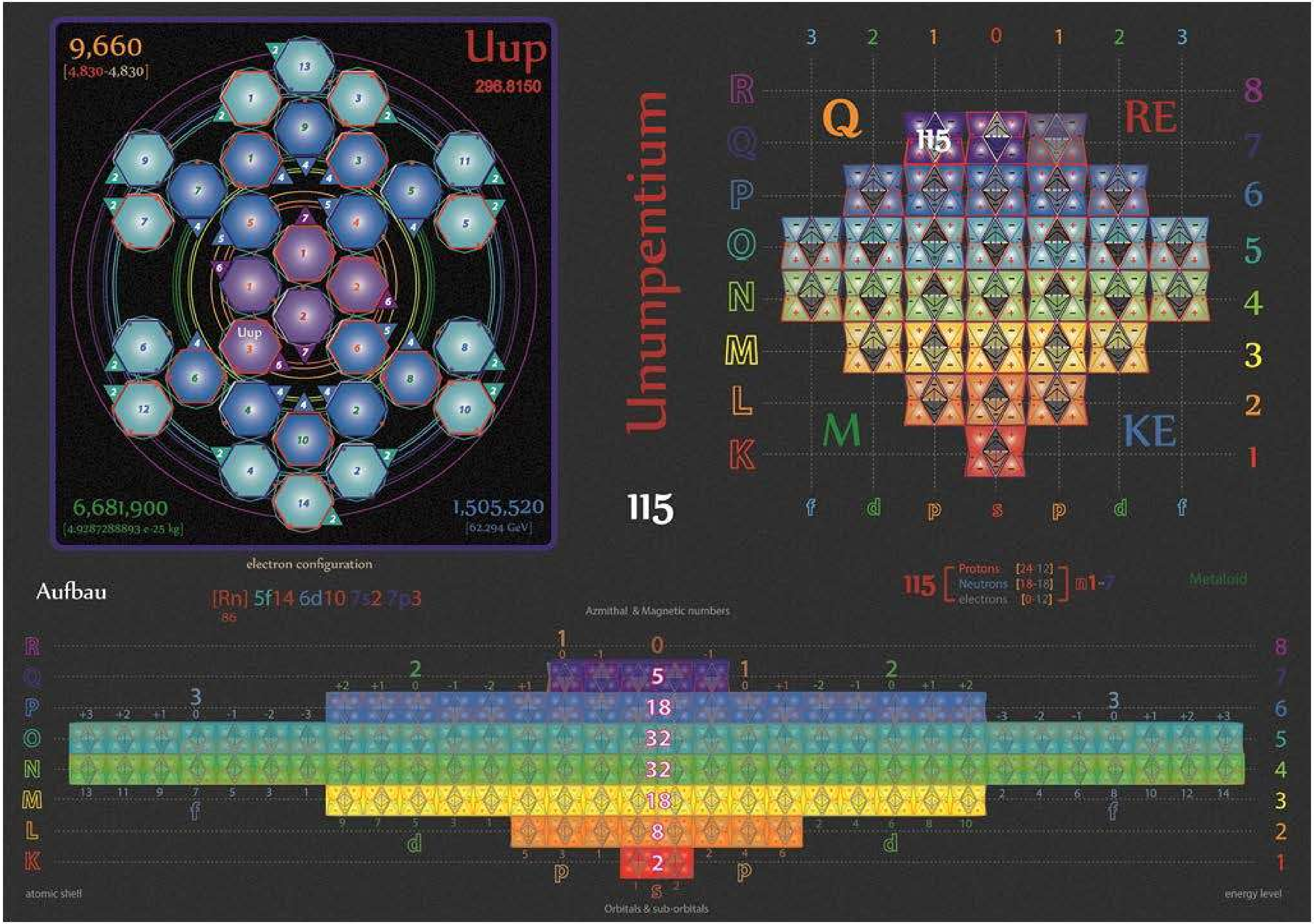
9,660
[4,830-4,830]
Uup
296.8150
6,681,900
[4.9287288893 e-25 kg]
1,505,520
[62.294 GeV]
Ununpentium
115
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p3
86
Q
RE
M
KE
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
115
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-7
Metalloid
Azmithal & Magnetic numbers
5
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.115 - Ununpentium atom

Copyright ABRAHAM [2008] - All rights reserved

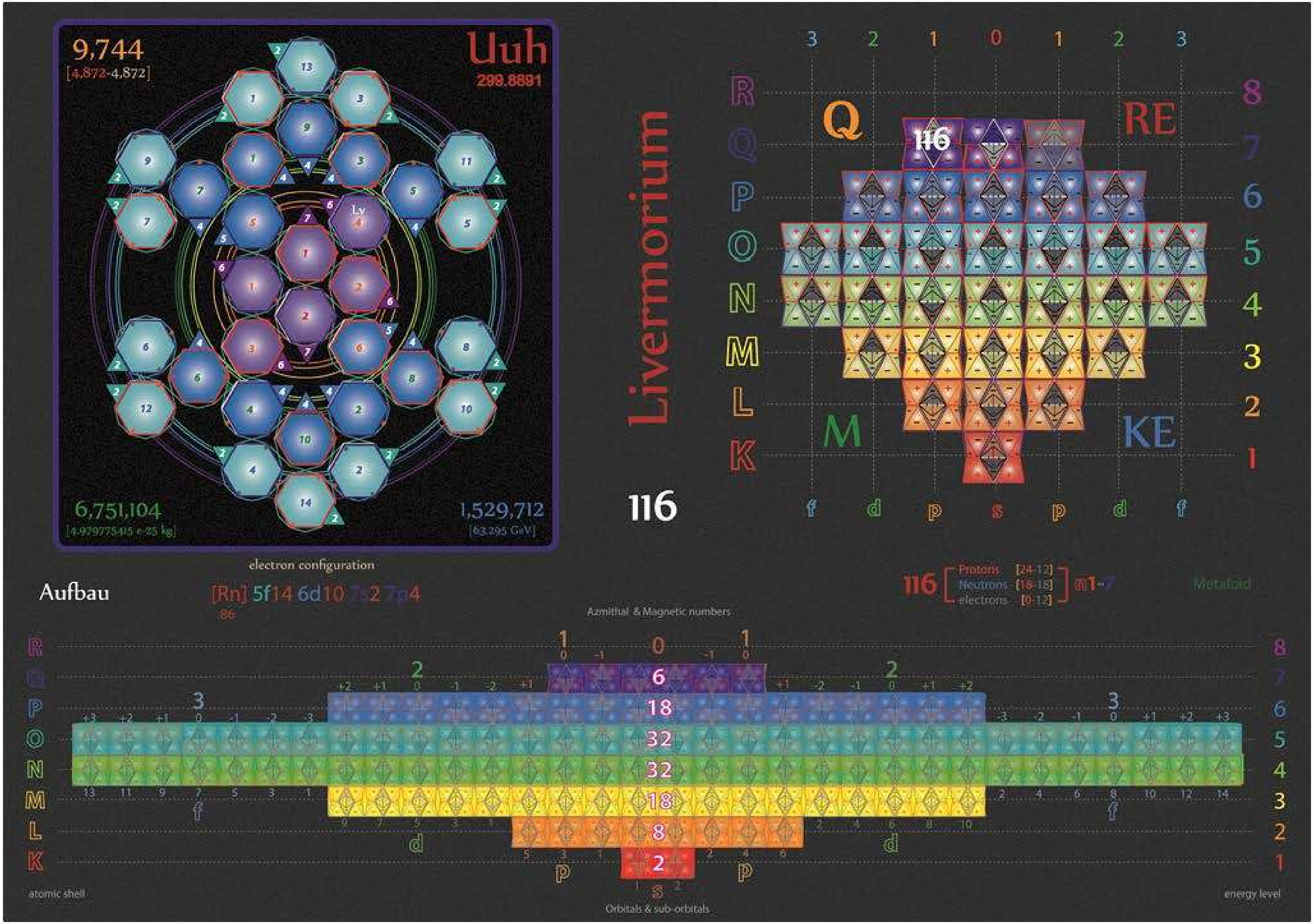
9,744
[4,872-4,872]
Uuh
289.8891
Lv
6,751,104
[4.979775415 e-25 kg]
1,529,712
[63.295 GeV]
Livermorium
116
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8
7
6
5
4
3
2
1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p4
86
116 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Metaloid
Azmithal & Magnetic numbers
6
18
32
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.116 - Livermorium atom

Copyright ABRAHAM [2008] - All rights reserved

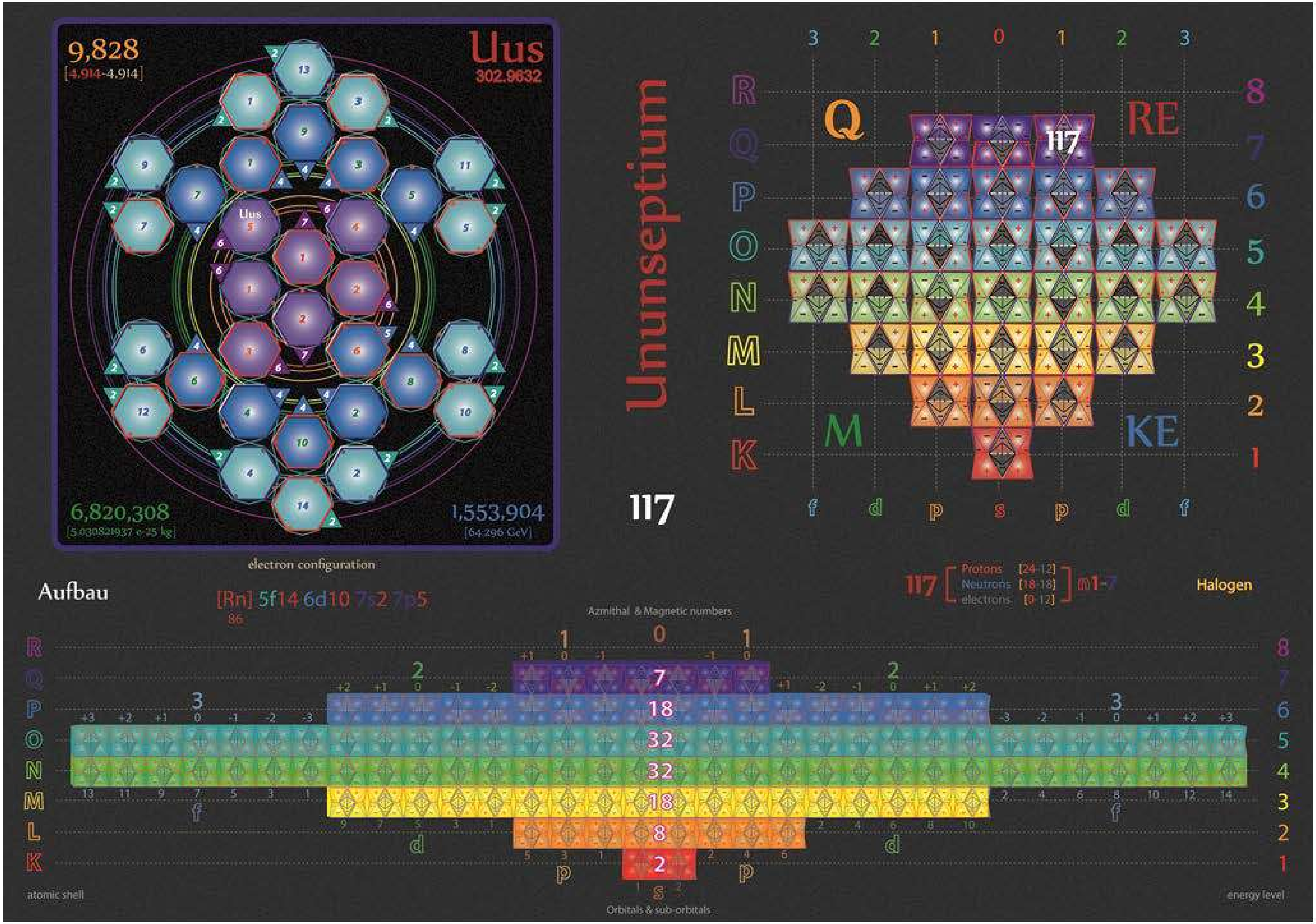
9,828
[4.914-4.914]
Uus
302.9832
Uus
6,820,308
[5.030821937 e-25 kg]
1,553,904
[64.296 GeV]
electron configuration
Ununseptium
117
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
117
f d p s p d f
117
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Halogen
Aufbau
[Rn] 5f14 6d10 7s2 7p5
86
Azmithal & Magnetic numbers
7
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.117 - Ununseptium atom

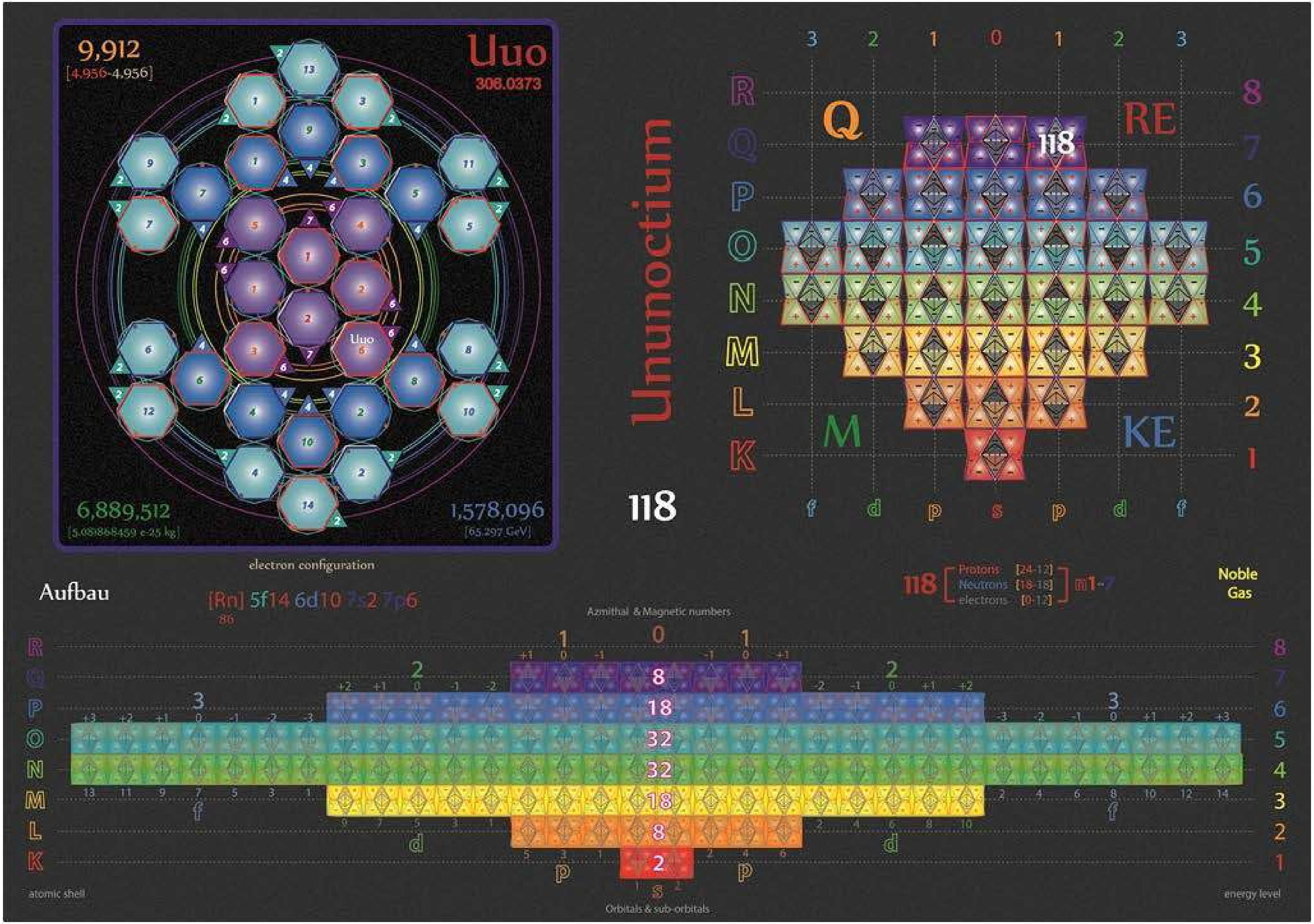
9,912
[4,956-4,956]
Uuo
308.0373
6,889,512
[5.09868459 e-25 kg]
1,578,096
[65.297 GeV]
Uuo
electron configuration
Ununoctium
118
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
118
118 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Noble Gas
Aufbau
[Rn] 5f14 6d10 7s2 7p6
86
Azmithal & Magnetic numbers
1 0 1
2 2
3 3
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 0 +1 +2 +3
8
18
32
32
18
8
2
13 11 9 7 5 3 1 2 4 6 8 10 12 14
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 2
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.118 - Ununoctium atom

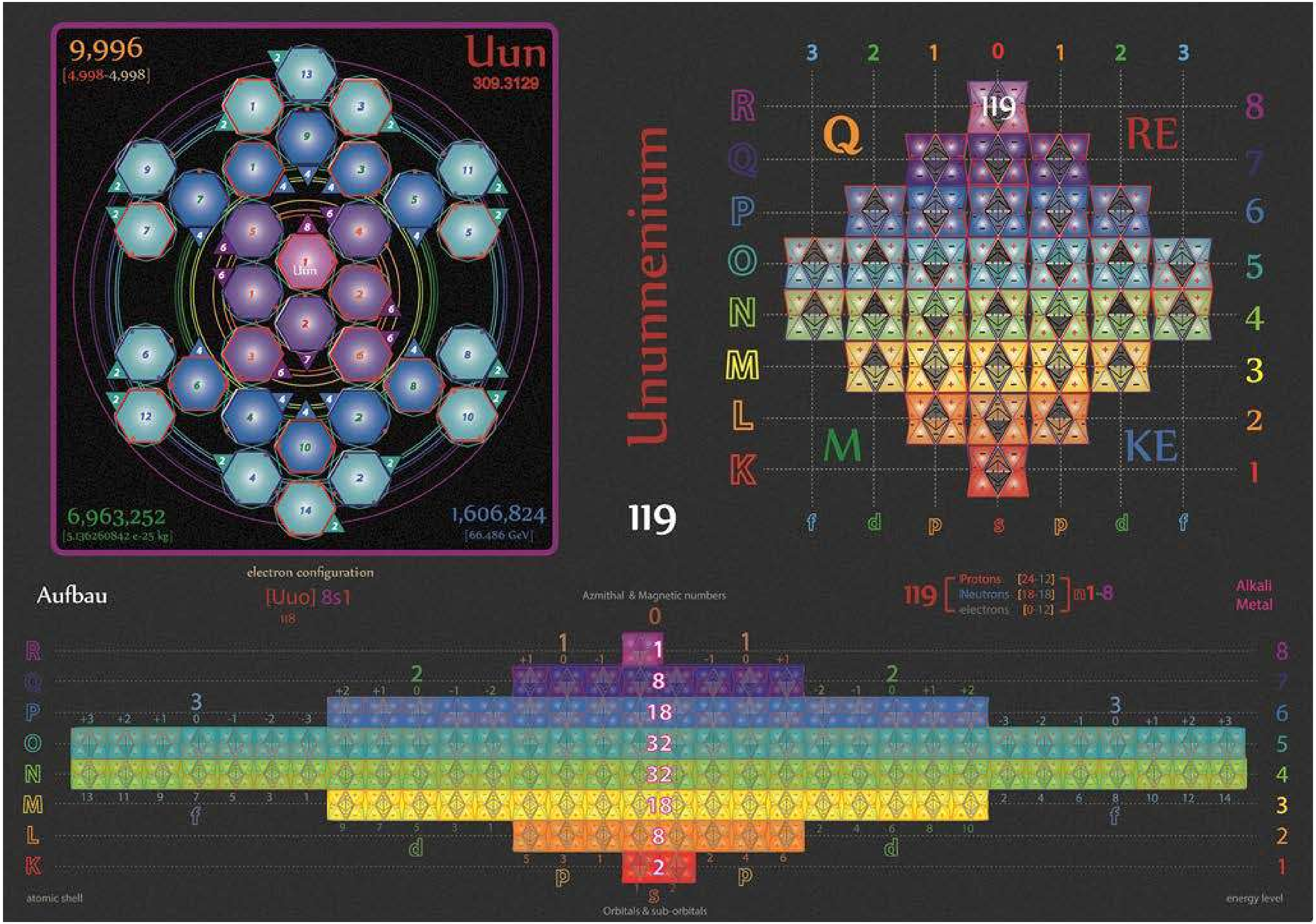
9,996
[4.998-4.998]
Uun
309.3129
6,963,252
[5.136260842 e-25 kg]
1,606,824
[66.486 GeV]
Ununnenium
119
R
Q
P
O
N
M
L
K
Q
RE
M
KE
3
2
1
0
1
2
3
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
Aufbau
electron configuration
[Uuo] 8s1
118
Azmithal & Magnetic numbers
119
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-8
Alkali
Metal
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.119 - Ununnenium atom

Copyright ABRAHAM [2008] - All rights reserved

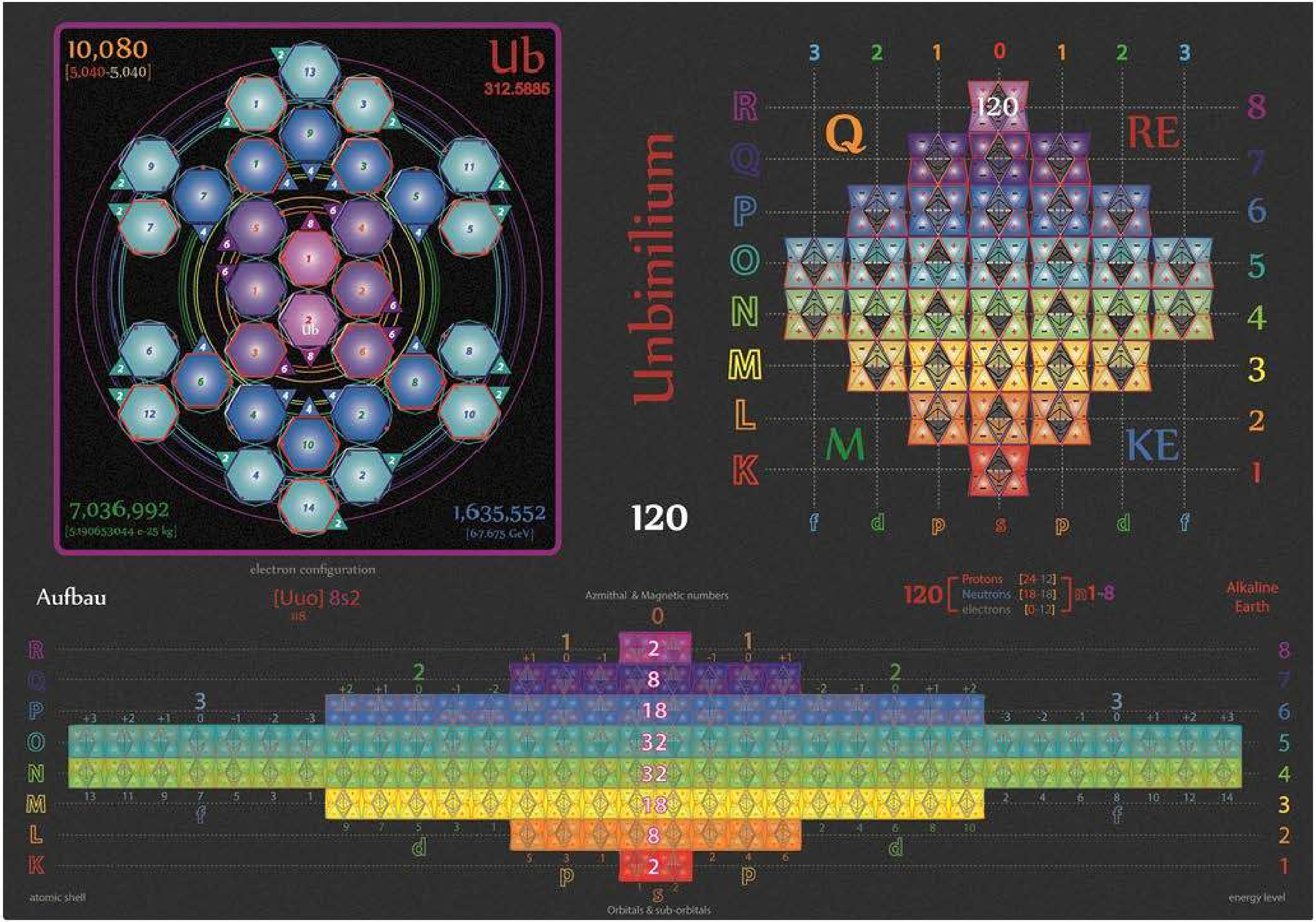
10,080
[5,040-5,040]
Ub
312.5886
7,036,992
[5.190653044 e-25 kg]
1,635,552
[67.675 GeV]
electron configuration
Unbinilium
120
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Aufbau
[Uuo] 8s2
118
Azmithal & Magnetic numbers
120 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-8
Alkaline Earth
2
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.120 - Unbinilium atom

Copyright ABRAHAM [2008] - All rights reserved

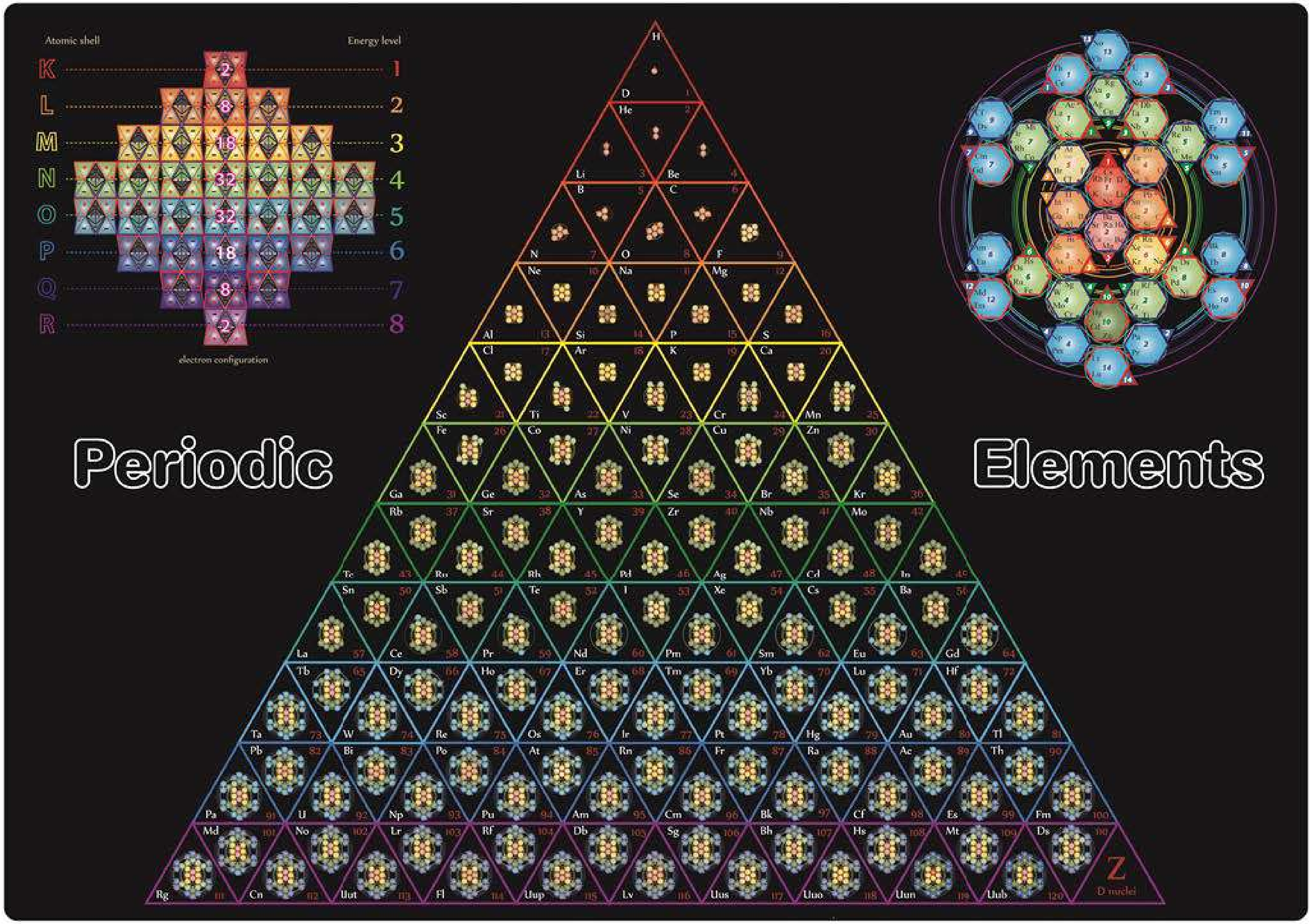
Atomic shell
Energy level
K 1
L 2
M 3
N 4
O 5
P 6
Q 7
R 8
2
8
18
32
32
18
8
2
electron configuration
Periodic
Elements
H
D 1
He 2
Li 3
Be 4
B 5
C 6
N 7
O 8
F 9
Ne 10
Na 11
Mg 12
Al 13
Si 14
P 15
S 16
Cl 17
Ar 18
K 19
Ca 20
Sc 21
Ti 22
V 23
Cr 24
Mn 25
Fe 26
Co 27
Ni 28
Cu 29
Zn 30
Ga 31
Ge 32
As 33
Se 34
Br 35
Kr 36
Rb 37
Sr 38
Y 39
Zr 40
Nb 41
Mo 42
Tc 43
Ru 44
Rh 45
Pd 46
Ag 47
Cd 48
In 49
Sn 50
Sb 51
Te 52
I 53
Xe 54
Cs 55
Ba 56
La 57
Ce 58
Pr 59
Nd 60
Pm 61
Sm 62
Eu 63
Gd 64
Tb 65
Dy 66
Ho 67
Er 68
Tm 69
Yb 70
Lu 71
Hf 72
Ta 73
W 74
Re 75
Os 76
Ir 77
Pt 78
Hg 79
Au 80
Tl 81
Pb 82
Bi 83
Po 84
At 85
Rn 86
Fr 87
Ra 88
Ac 89
Th 90
Pa 91
U 92
Np 93
Pu 94
Am 95
Cm 96
Bk 97
Cf 98
Es 99
Fm 100
Md 101
No 102
Lr 103
Rf 104
Db 105
Sg 106
Bh 107
Hs 108
Mt 109
Ds 110
Rg 111
Cn 112
Uut 113
Fl 114
Uup 115
Lv 116
Uus 117
Uuo 118
Uun 119
Uub 120
Z
D nuclei

Tetryonics 52.01 - Tetryonic Element Table

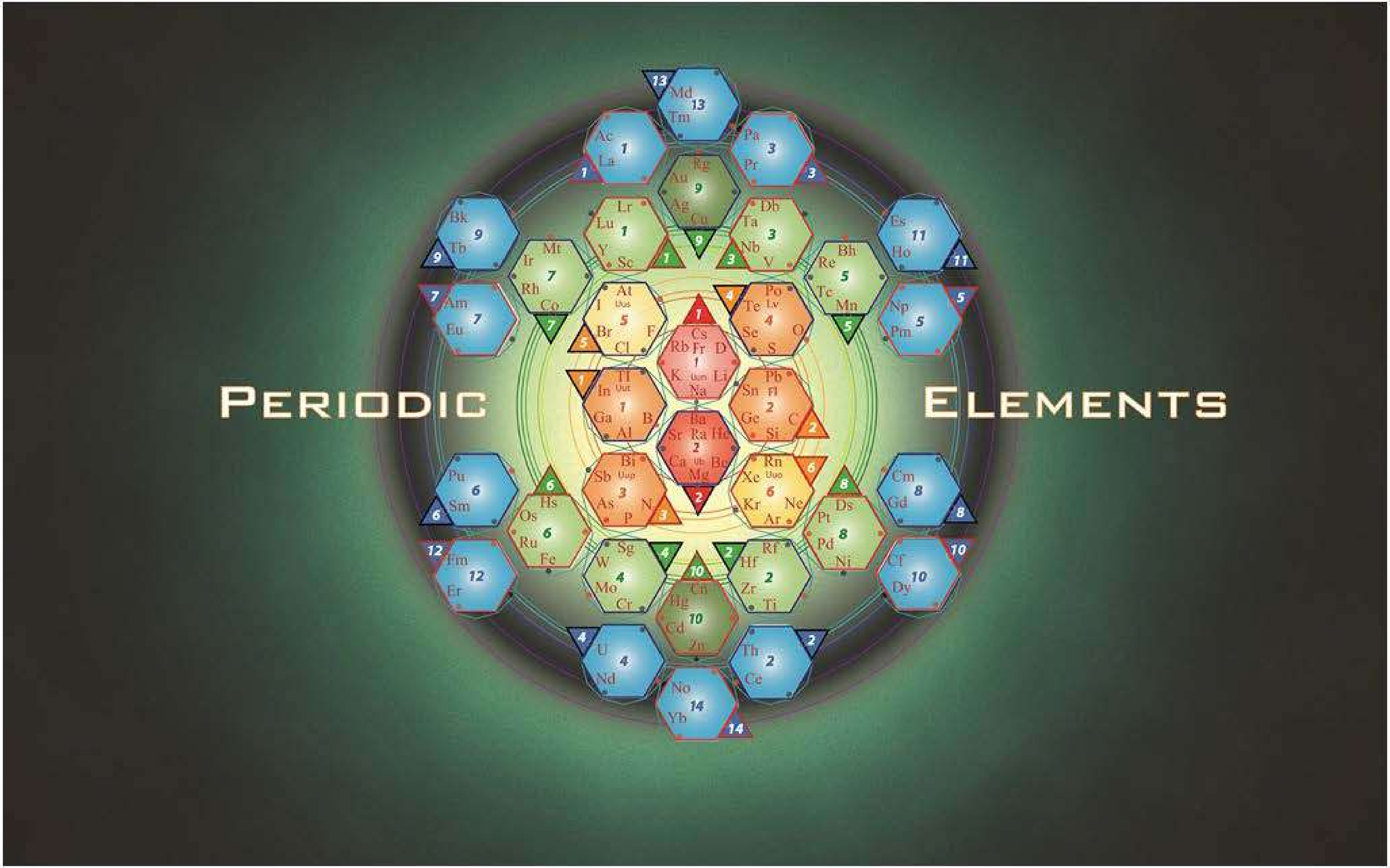
PERIODIC
ELEMENTS
13 Md 13 Tm
1 Ac 1 La
Pa 3 Pr 3
Rg Au 9 Ag Cu 9
Bk 9 Tb 9
Lr Lu 1 Y Sc 1
Db Ta 3 Nb V 3
Es 11 Ho 11
Mt Ir 7 Rh Co 7
Bh Re 5 Tc Mn 5
7 Am 7 Eu
At I Uus 5 Br F Cl 5
1 Cs Rb Fr D 1 K Uun Li Na
4 Po Te Lv 4 Se O S
Np 5 Pm 5
1 Tl In Uut 1 Ga B Al
Pb Sn Fl 2 Ge C Si 2
Ba Sr Ra He 2 Ca Ub Be Mg 2
Pu 6 Sm 6
6 Hs Os 6 Ru Fe
Bi Sb Uup 3 As N P 3
Rn Xe Uuo 6 Kr Ne Ar 6
8 Ds Pt 8 Pd Ni
Cm 8 Gd 8
12 Fm 12 Er
Sg W 4 Mo Cr 4
10 Cn Hg 10 Cd Zn
2 Rf Hf 2 Zr Ti
Cf 10 Dy 10
4 U 4 Nd
Th 2 Ce 2
No 14 Yb 14

Tetryonics 52.02 - Periodic Table [Elements]

Copyright ABRAHAM [2008] - all rights reserved

TETRYONICS

The charged geometry of mass-ENERGY-Matter

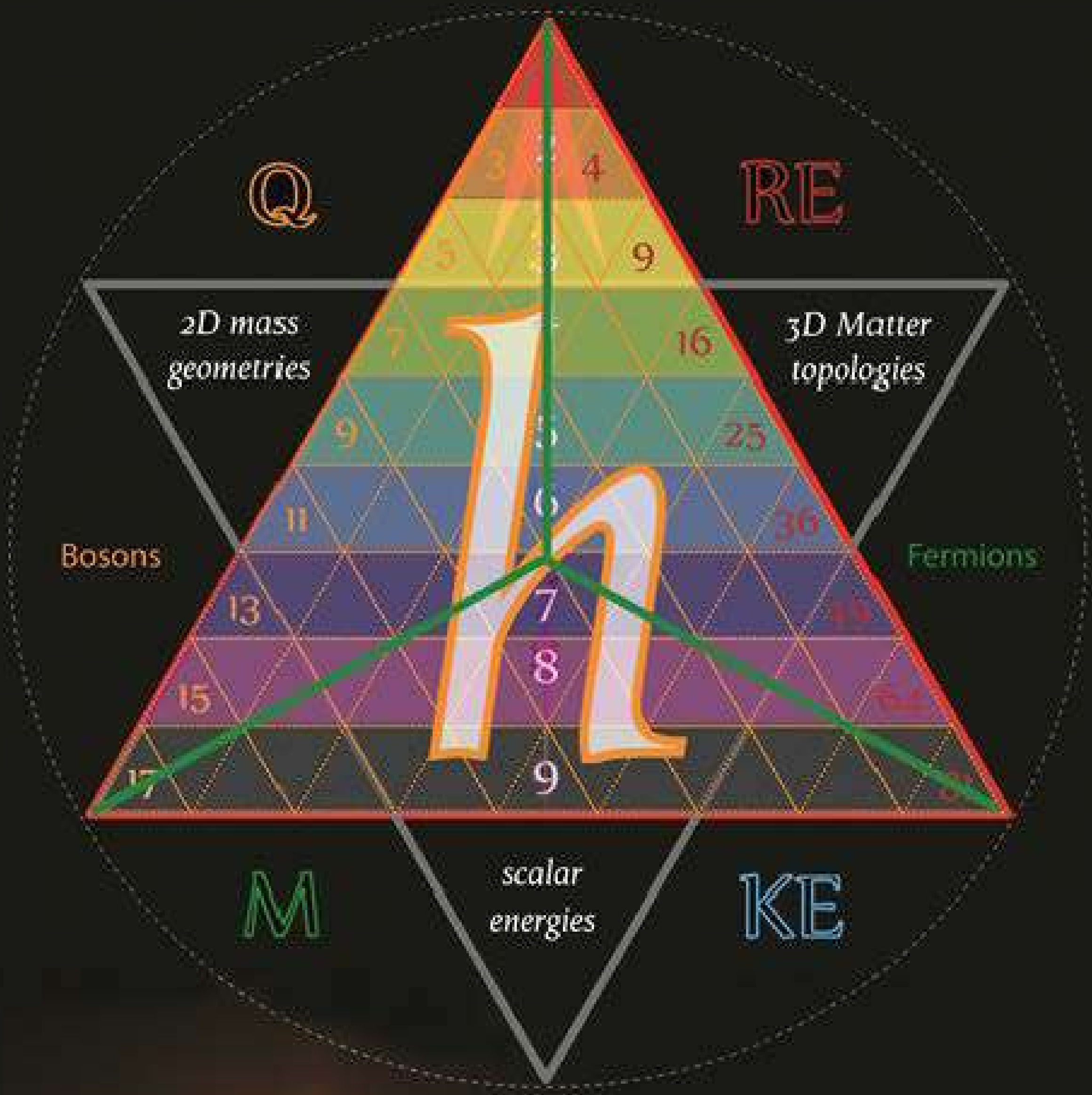
Q
RE
2D mass geometries
3D Matter topologies
Bosons
Fermions
h
3
4
5
9
7
16
9
25
11
36
13
15
17
6
7
8
9
M
scalar energies
KE

Foundational Quantum Mechanics

Abraham

ISBN 978-0-9872884-4-7
[Second Edition © 2012]

Copyright ABRAHAM [2008] - all rights reserved

FUNDAMENTUM QUANTUM MECHANICA

questions **Q**

A answers

[TETRYONICS]

A fundamental re-interpretation of the geometry of quantised angular momentum is required to complete the physics of 'The Standard model'

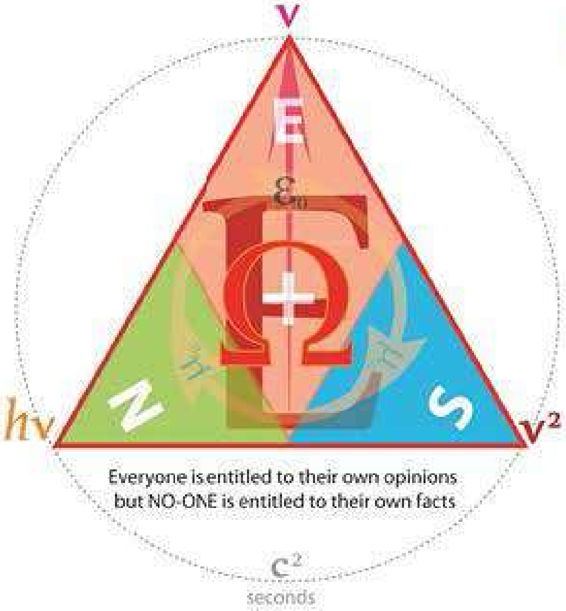

Everyone is entitled to their own opinions
but NO-ONE is entitled to their own facts

Having removed the impossible,
anything that remains, however improbable,
must be the truth

Mathematics is the language of Physics,
and Geometry is its grammar

Philosophiae Naturalis Principia geometricae

> *"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."*
>
> *[Albert Einstein]*

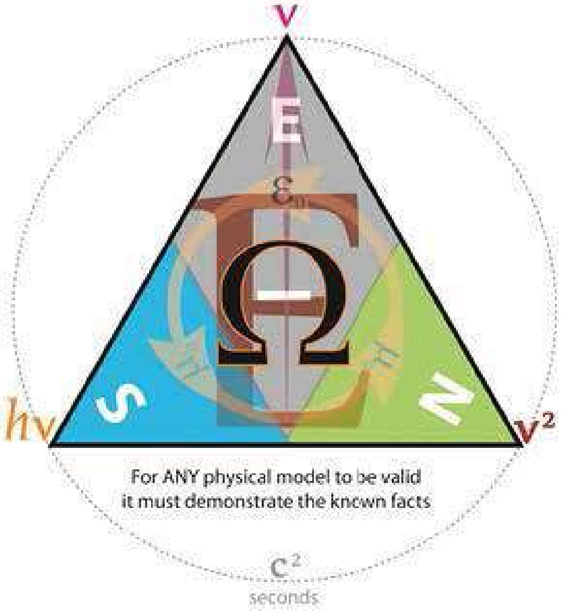

For ANY physical model to be valid
it must demonstrate the known facts

Science is born from observation,
and the reasoning of known facts
in search of underlying truths

In the following pages the true geometry of quantum mechanics is revealed, leading scientific endeavour into new realms of understanding

Copyright ABRAHAM [2008] - all rights reserved

mass-ENERGY-Matter

The a-priori revelation of Tetryonic theory is that all square mass-energies possess equilateral momenta geometries

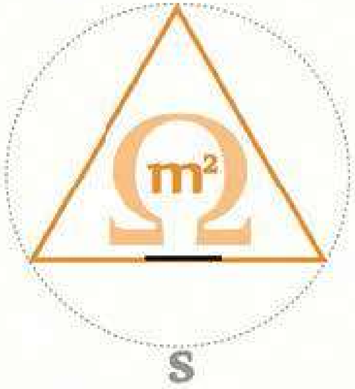

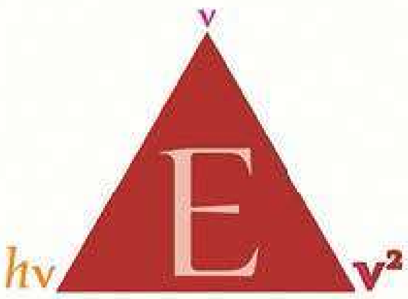

The equilateral Quantised Angular Momentum intrinsic to Planck mass-energy momenta produces charged geometries

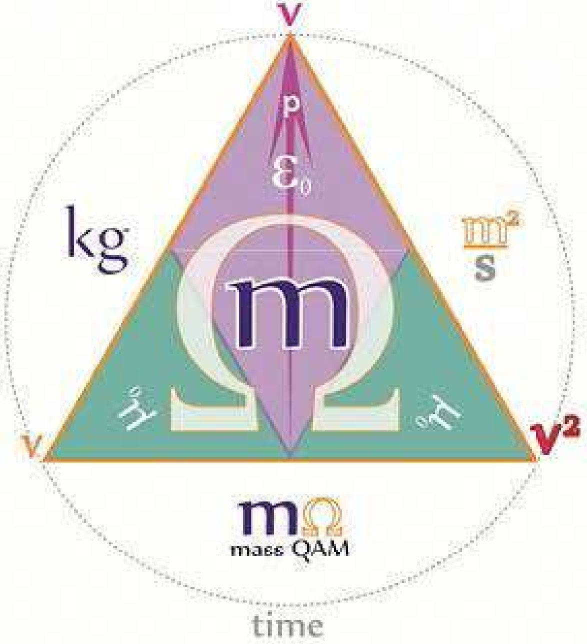
v
p
ε₀
kg
m²
s
Ω
m
v
v²
mΩ
mass QAM
time

The quantum mechanics of velocity, quanta, EM fields and mass-Energy-Matter can be fully revealed through their equilateral geometries

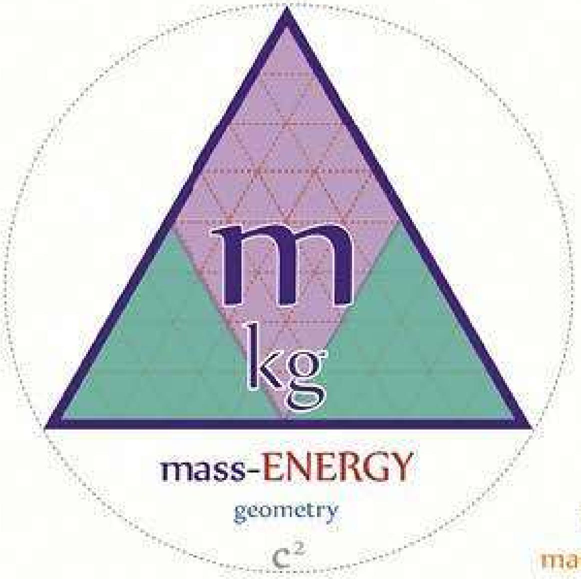
m
kg
mass-ENERGY
geometry
c²

A long hidden topology is revealed

Equilateral triangles are the foundational geometry for all mass-ENERGY-Matter topologies and physical Force interactions

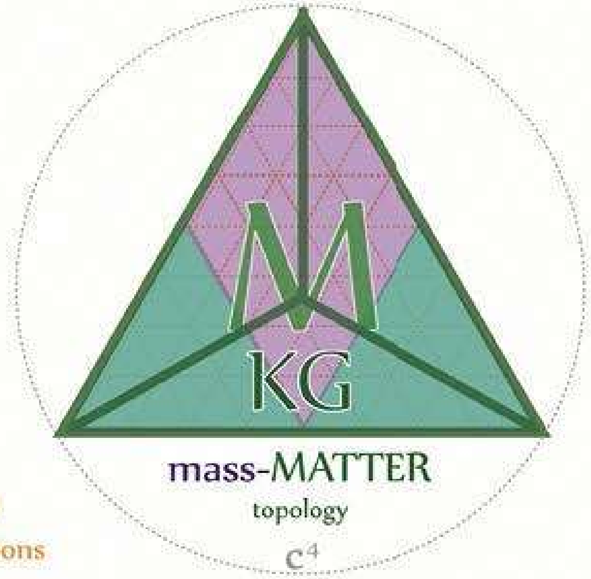
M
KG
mass-MATTER
topology
c⁴

SQUARED energies in quantum mechanics are EQUILATERAL geometries

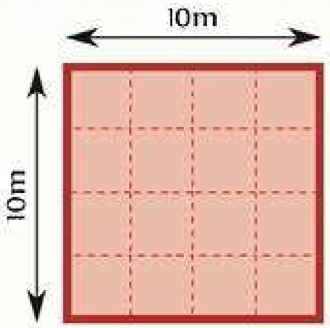

Square

area = s^2 = [100]

Circles

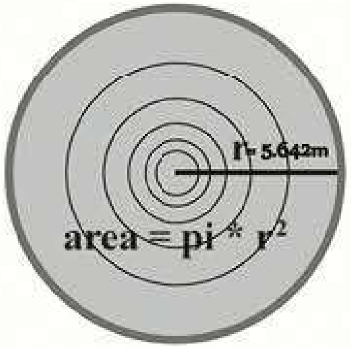
r=5.642m
area = pi * r^2

= pi *$[5.642]^2$
= 100

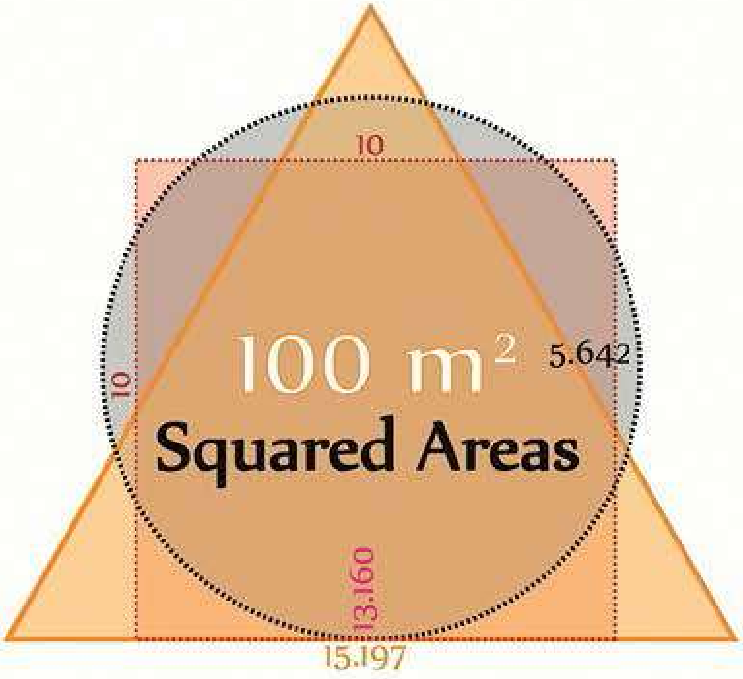
10
10
5.642
100 m²
Squared Areas
13.160
15.197

can be created by a number of planar geometries

*For a long time it has been assumed by scientists (and mathematicians)
that circular [and squared] geometries are the geometric foundation of all physics.
leading to a serioulsy flawed model of particles and forces in quantum mechanics*

Tetryonic theory now reveals that quantised equilateral angular momenta creates the foundational geometry of all the mass-Energy-Matter & forces of physics

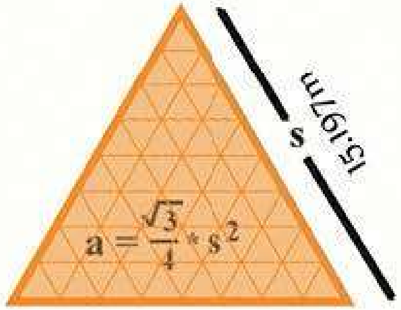
15.197m
s
$a = \frac{\sqrt{3}}{4} * s^2$

Equilateral

area = $(\frac{1}{2}*b)*h$

Triangles

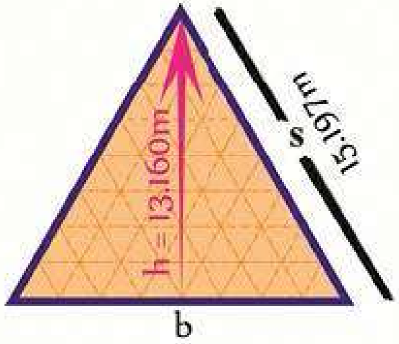
15.197m
s
h = 13.160m
b

b h
[.5x15.197] x 13.160
= 100

Copyright ABRAHAM [2008] - all rights reserved

Integers

The integers (from the Latin integer), literally "untouched", hence "whole"
in Tetryonics it is the basis for the quantum

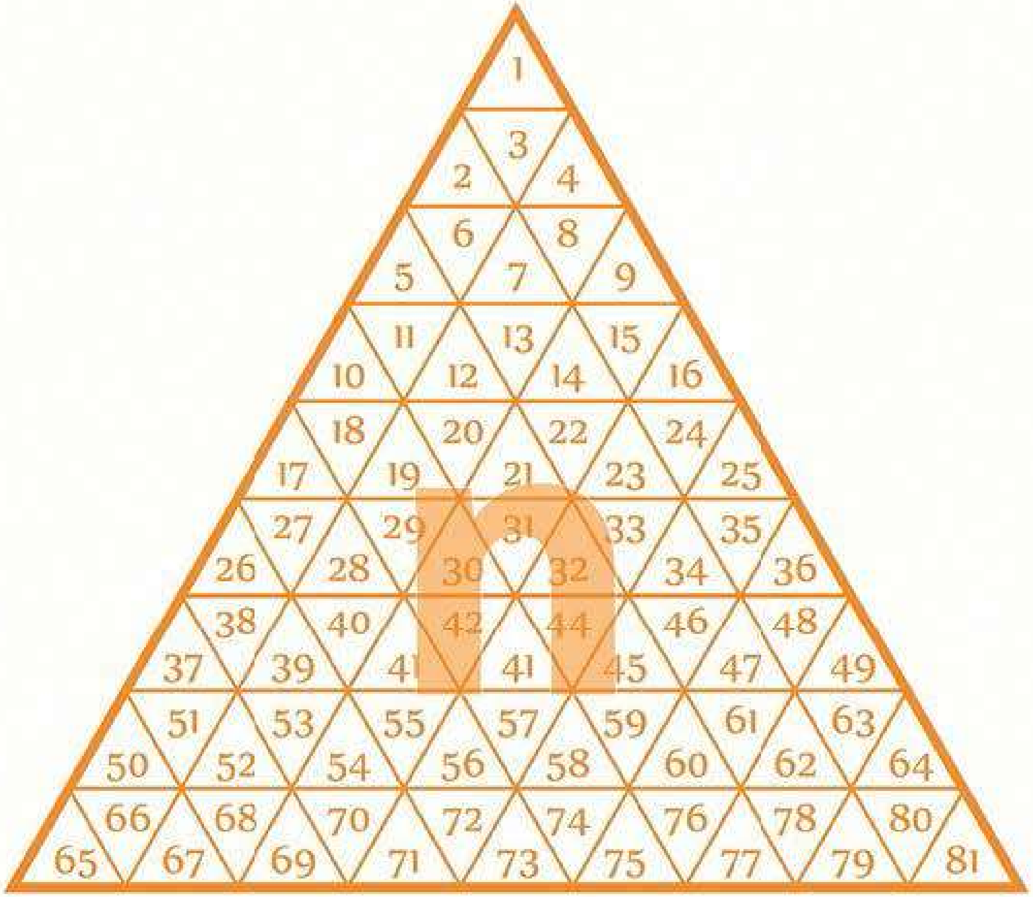
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
n

Viewed as a subset of the real numbers, they are numbers that can be written without a fractional or decimal component

Copyright ABRAHAM [2008] - all rights reserved

ODD numbers

An odd number is an integer which is not a multiple of two.

2n-1

2n+1

ODD numbers
in each level

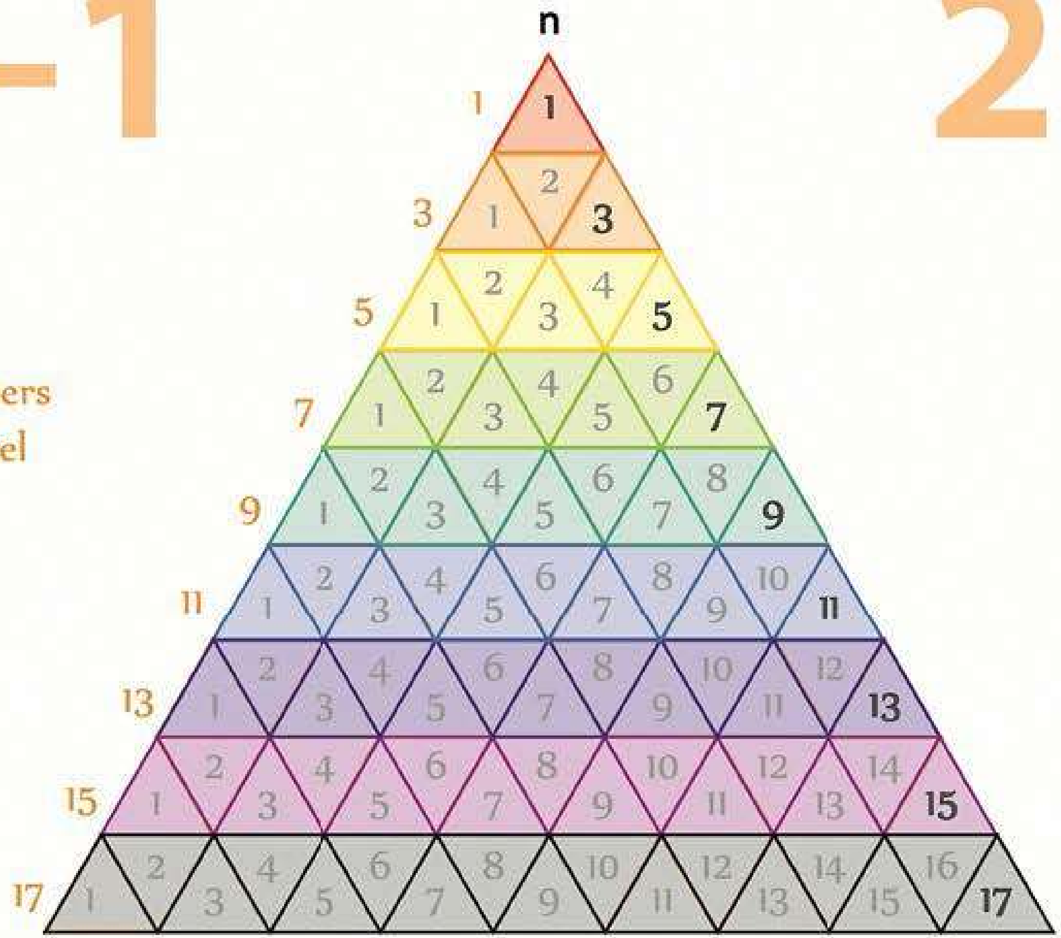

Bosons have
ODD number
quanta

An odd number, when divided by two, will result in a fraction

Copyright ABRAHAM [2008] - all rights reserved

EVEN numbers

An integer that is not an odd number is an even number.

2

Photons have
EVEN number
quanta

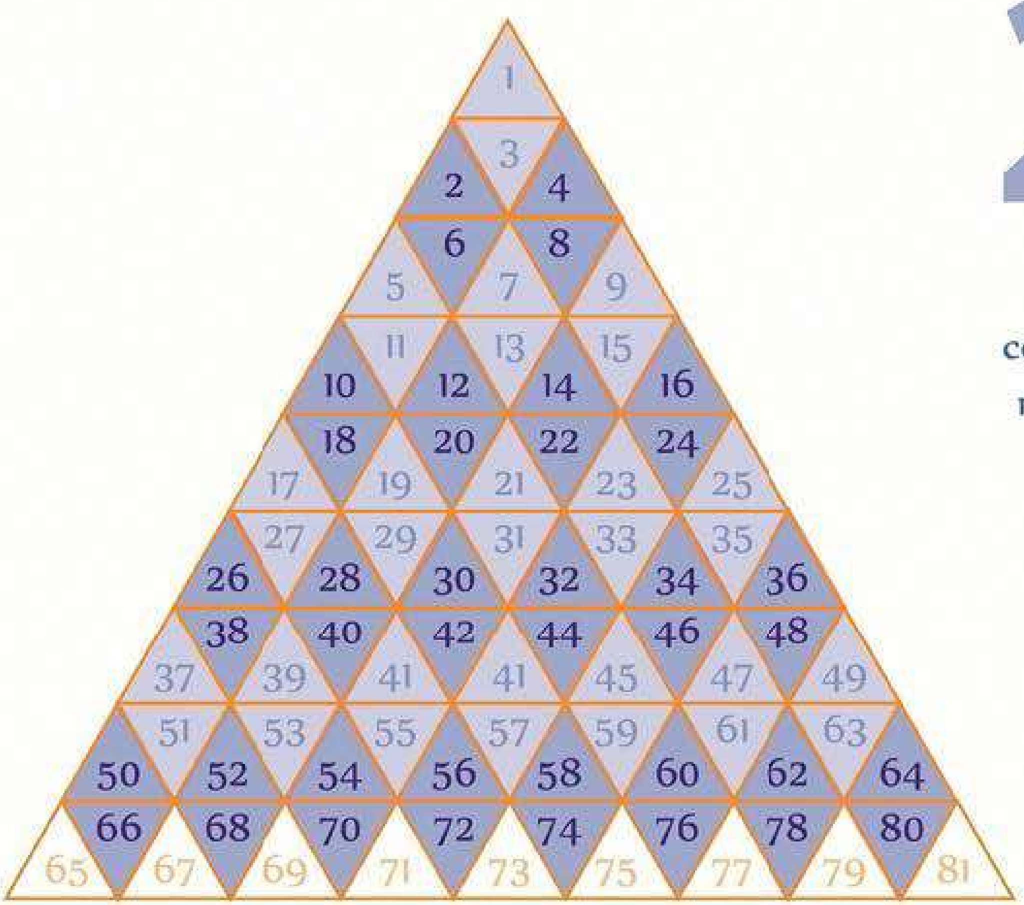
1
3
2
4
6
8
5
7
9
11
13
15
10
12
14
16
18
20
22
24
17
19
21
23
25
27
29
31
33
35
26
28
30
32
34
36
38
40
42
44
46
48
37
39
41
41
45
47
49
51
53
55
57
59
61
63
50
52
54
56
58
60
62
64
66
68
70
72
74
76
78
80
65
67
69
71
73
75
77
79
81

2n

EM waves are
comprised of EVEN
numbered quanta

An even number is defined as a whole number that is a multiple of two.
If an even number is divided by two, the result is another whole number.

Square numbers

A square number, sometimes also called a perfect square,
is the result of an integer multiplied by itself

$2n-1$

Quantum levels
have ODD number
geometries

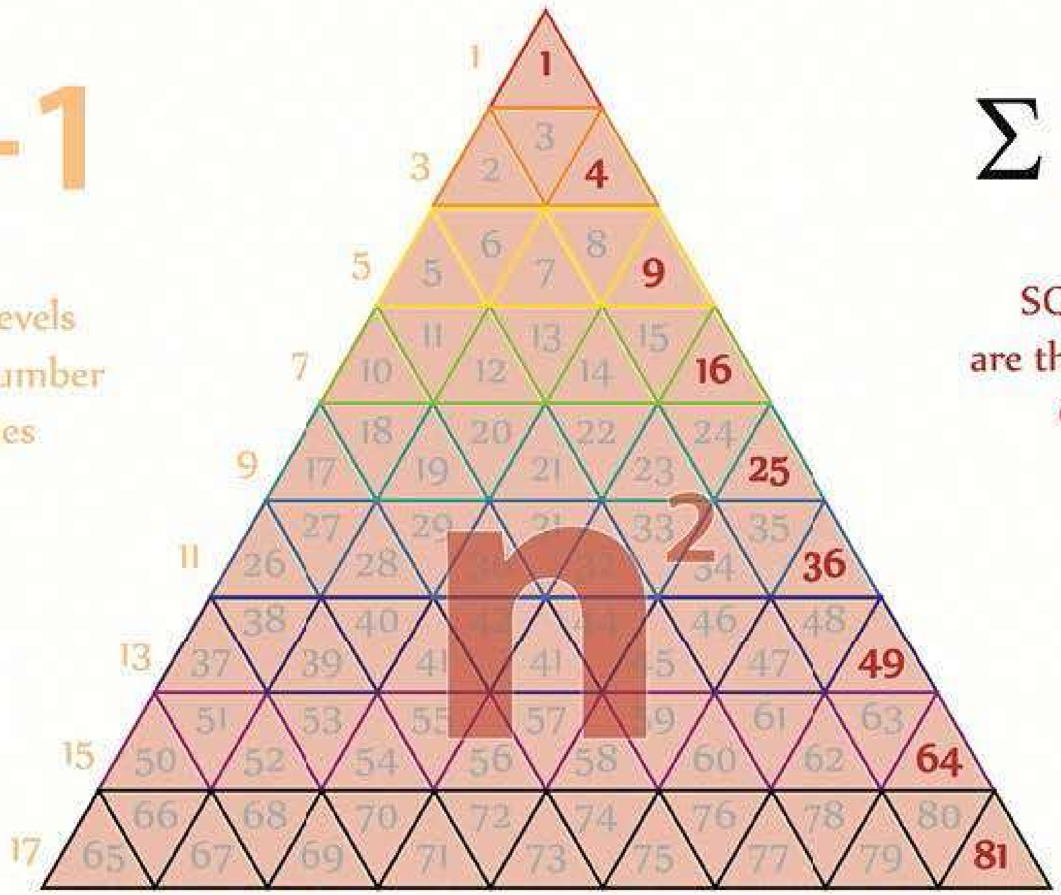
1
3
5
7
9
11
13
15
17
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
n^2

$\Sigma\ 2n-1$

SQUARE numbers
are the sum of successive
ODD numbers

Square
numbers

In Tetryonics SQUARE numbers are EQUILATERAL geometries

Copyright ABRAHAM [2008] - all rights reserved

Square roots

A square root of a number is a number that, when it is multiplied by itself (squared) , gives the first number again.

-*i* and +*i*

Against Mathematical convention,
square roots of negative numbers
are real numbers

In Physics
every complex number
except 0 has 2 square roots.

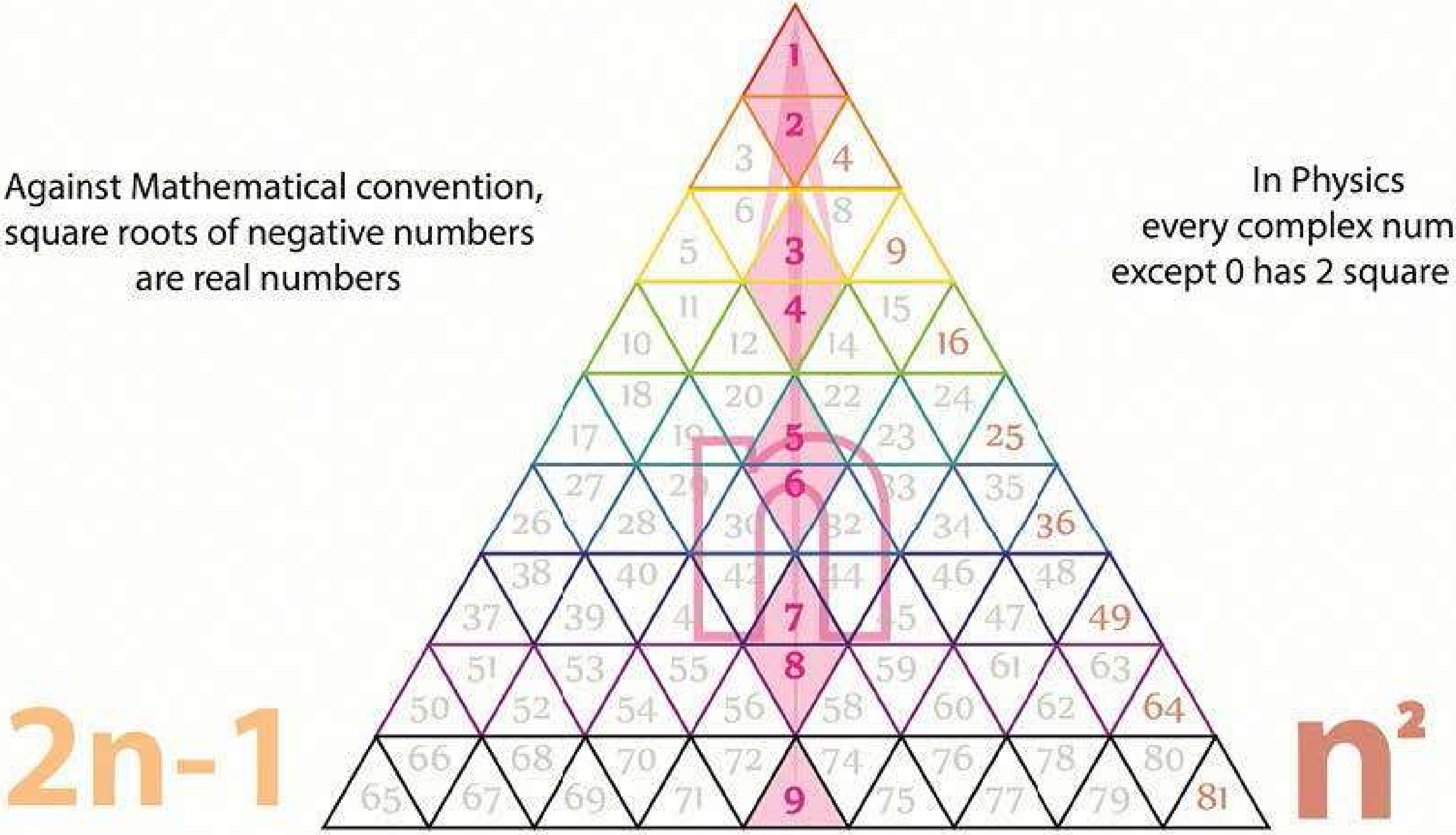

A whole number with a square root that is also a whole number is called a perfect square

Copyright ABRAHAM [2008] - all rights reserved

Real Numbers

A real number is a value that represents a quantity along a continuous line.

The real numbers include all the rational numbers,

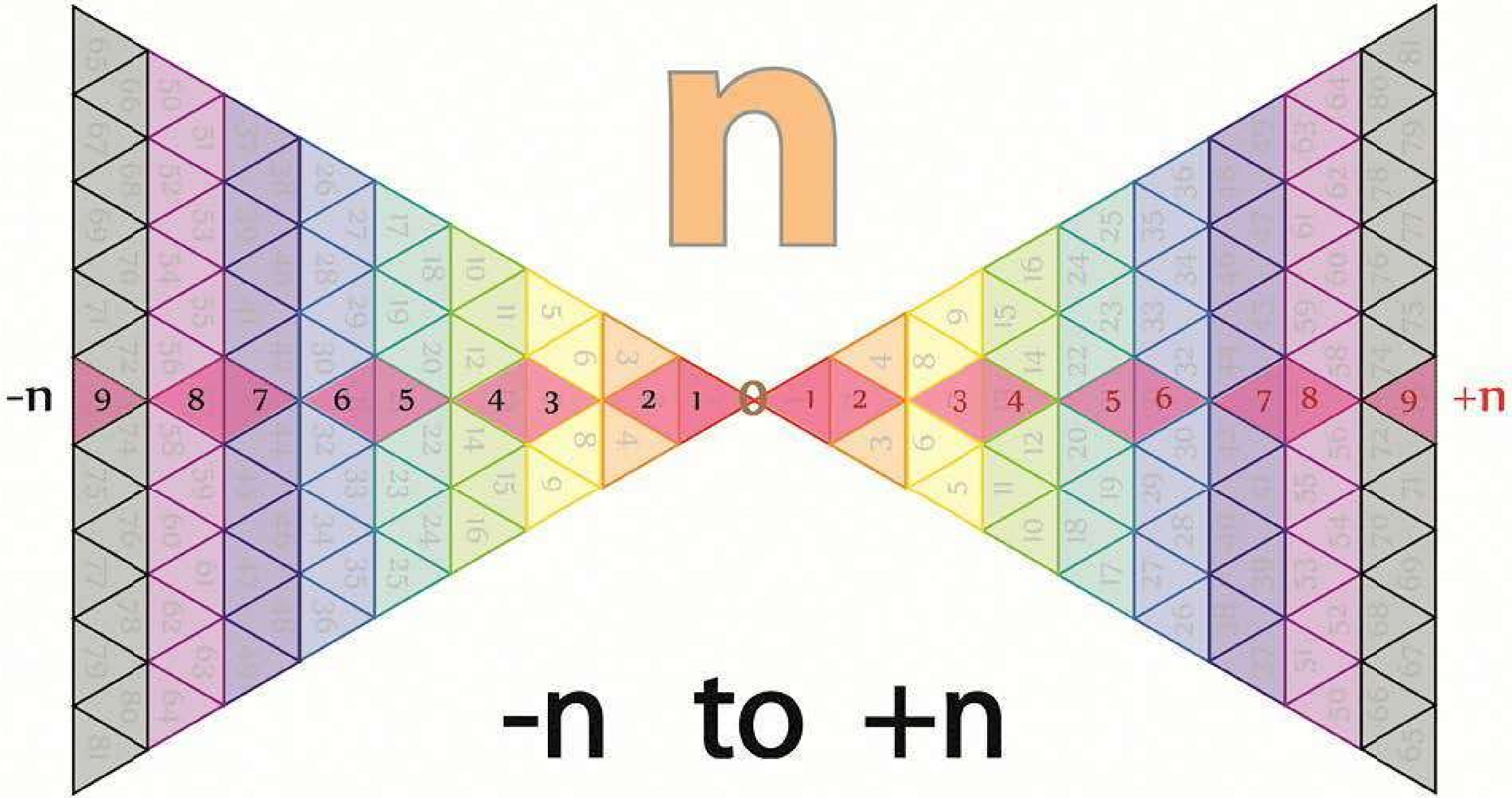
n
-n
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
+n
-n to +n

Irrational Numbers

An irrational number is defined to be any real number that cannot be written as a complete ratio of two integers

$$i^2 = -1$$

$+i$ and $-i$

Well known irrational & imanginary numbers in Math are π and i

π

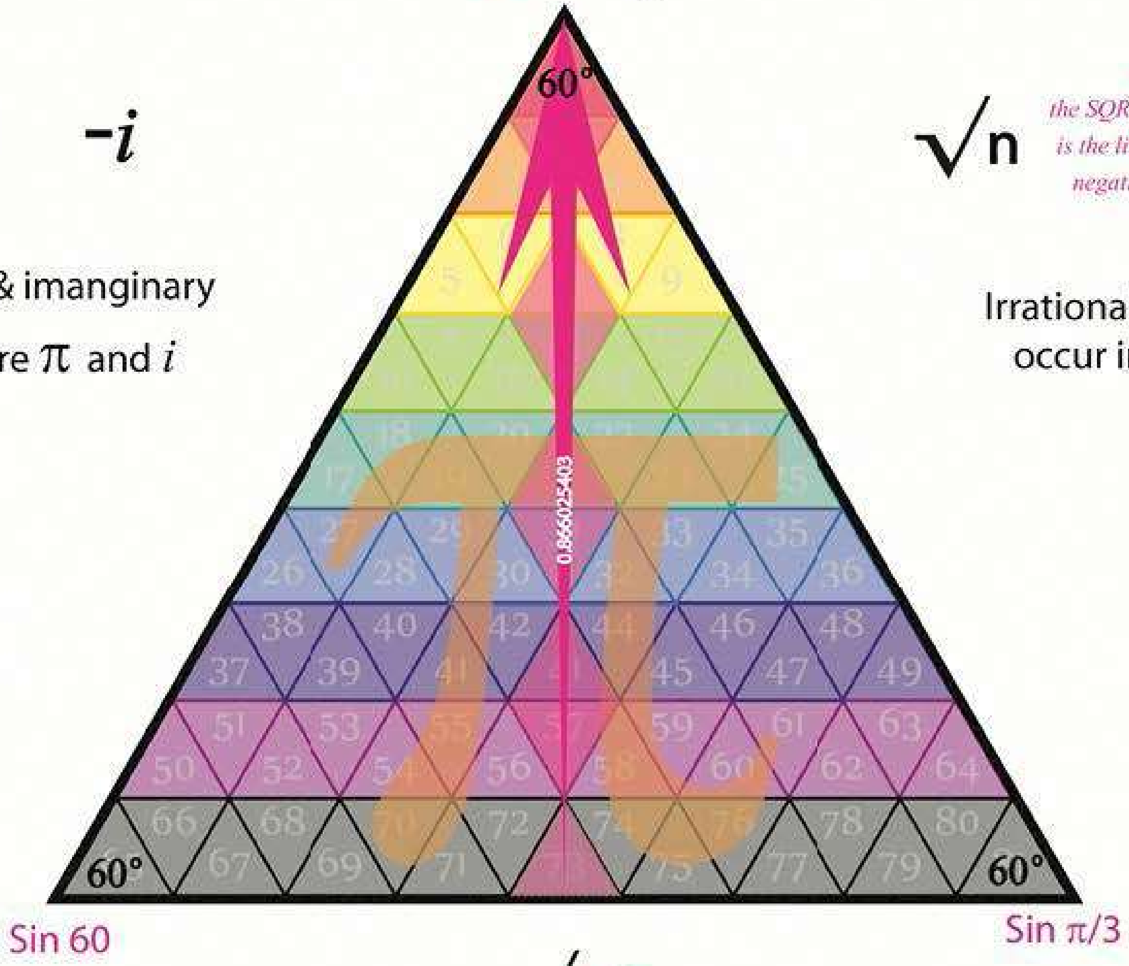
60°
0.866025403
26
27
28
30
33
34
35
36
37
38
39
40
42
45
46
47
48
49
50
51
52
53
56
59
60
61
62
63
64
66
67
68
69
71
72
77
78
79
80
60°
60°
Sin 60
Sin π/3

$\sqrt{3/2}$

$\sqrt{n}$ *in Tetryonics the SQR of a negative number is the linear momentum of a negative charge EM field* $\sqrt{-n}$

Irrational numbers often occur in mathematics

$\sqrt{-n}$

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic Colour Code

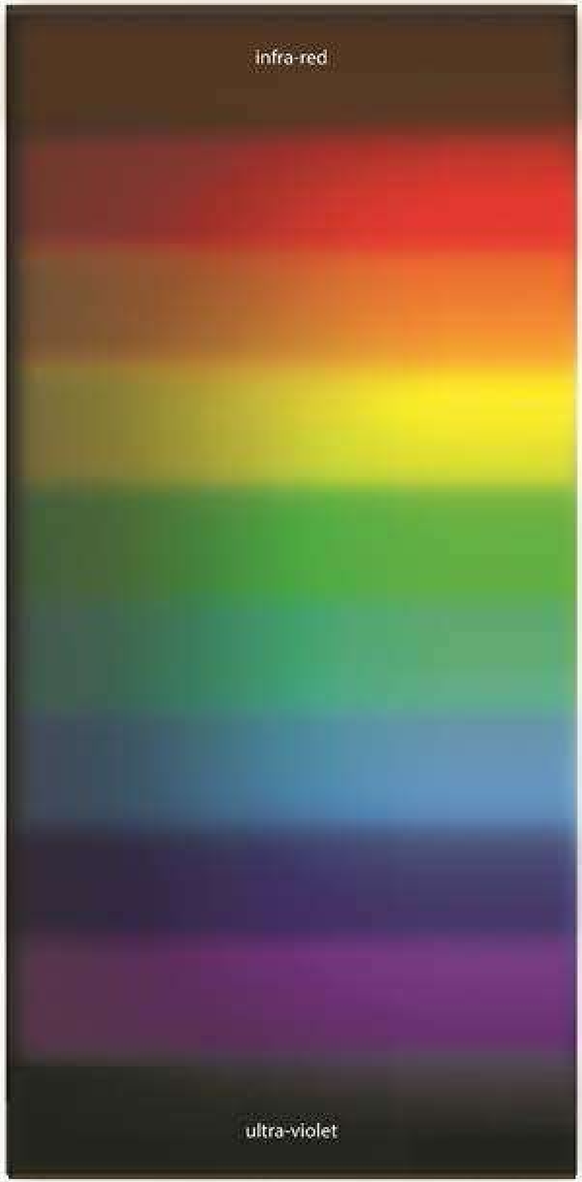

0	Brown
1	Red
2	Orange
3	Yellow
4	Green
5	Aqua
6	Blue
7	Indigo
8	Violet
9	Black

A colour code is used to indicate the varying quantum levels of the numerous forms of mass-ENERGY-Matter and serves to illustrate relationships between various Physical properties.

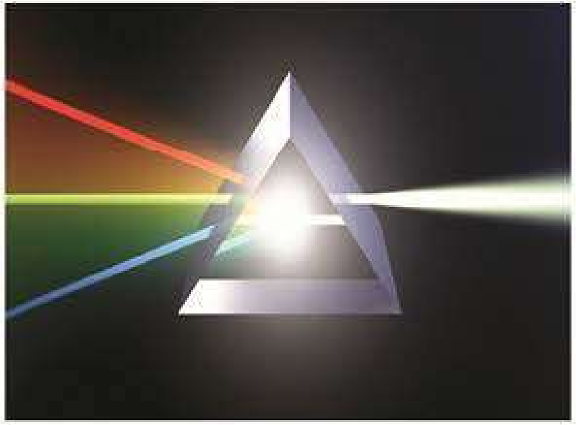

Tetryonics uses a colour code that is based on the spectral colours of dispersed White Light

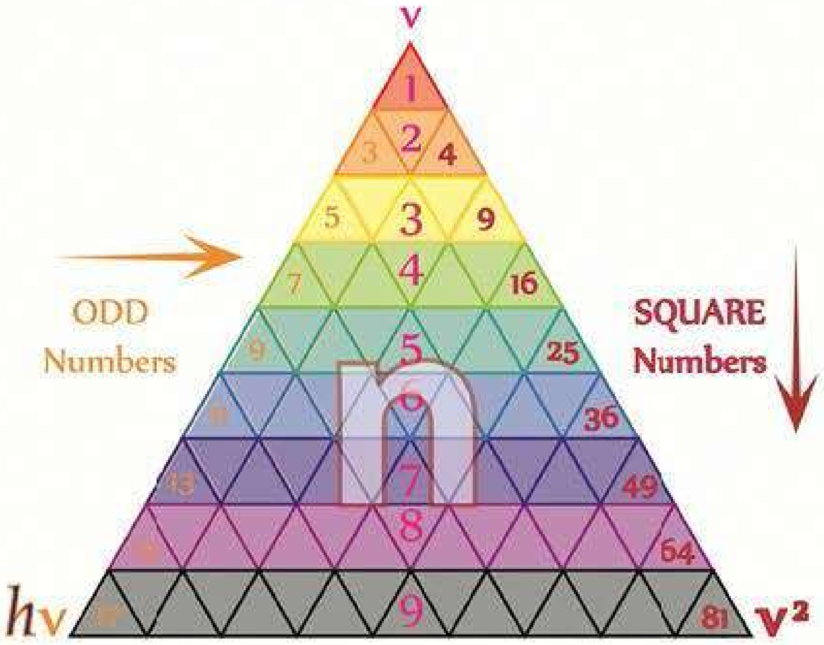
v
1
2
3
4
3
5
9
4
7
16
5
9
25
6
36
7
13
49
8
64
9
81
n
ODD Numbers
SQUARE Numbers
hv
v²

Copyright ABRAHAM [2008] - all rights reserved

Free Space

A contiguous volume or area
of any regular geometry
that is free, available, or unoccupied

$$E = 0$$

in any form

A Spatial region
is defined so as to
measure the physics
of mass-ENERGY-Matter
within its confines

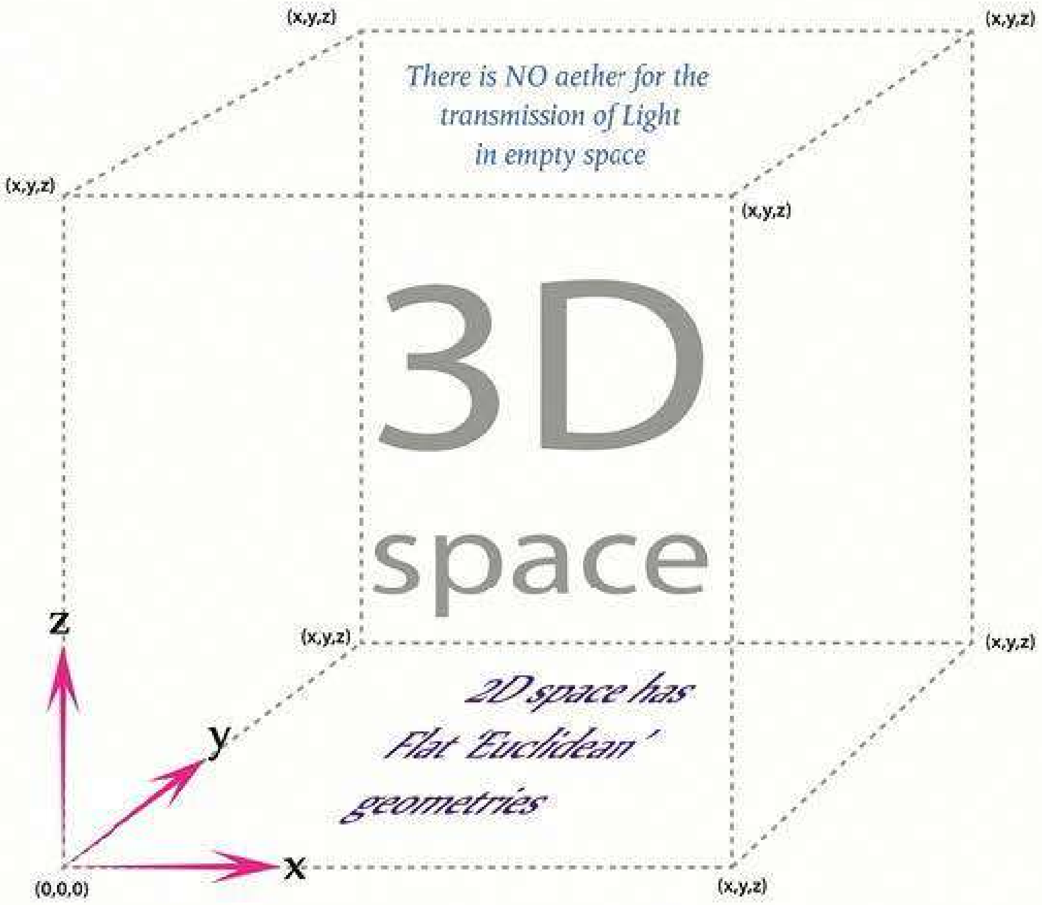

*There is NO aether for the
transmission of Light
in empty space*

3D
space

*2D space has
Flat Euclidean'
geometries*

Energy moves through Space
in various forms:
radiant energies, Matter etc

Space can be
Cubic, Spherical or
Tetrahedral as defined by
the spatial co-ordinates
used to define the region

Empty Space is defined as a topology whose volume is devoid of Energy

Copyright ABRAHAM [2008] - all rights reserved

Space-Time co-ordinates

Charge radiates unilaterally
in line with its
vector linear momentum

Vector forces

The linear vector of divergent Energy
forms various co-ordinate systems
based on its vector directions
per unit of Time

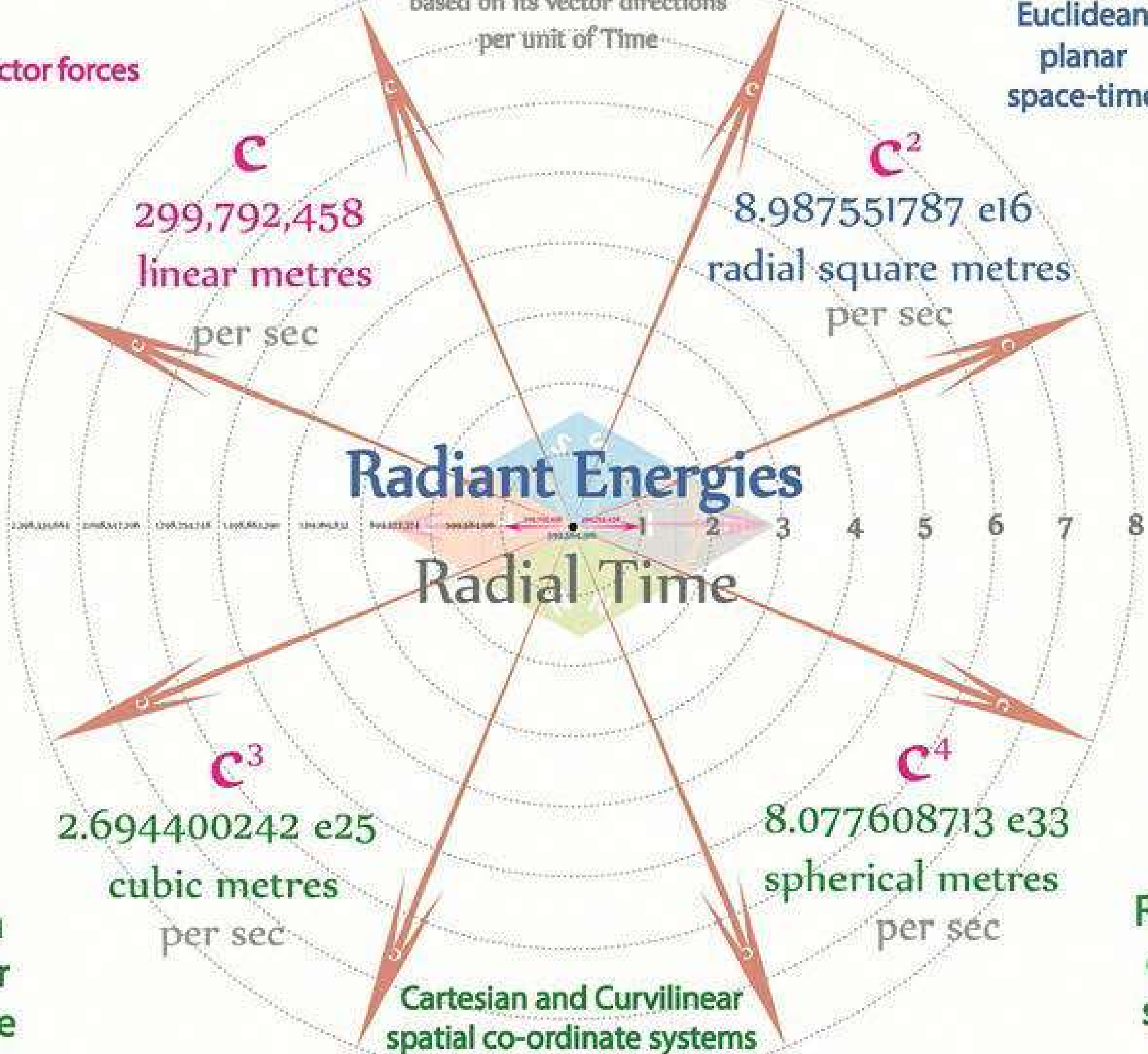

Radiant Energies

Radial Time

c
299,792,458
linear metres
per sec

c^2
8.987551787 e16
radial square metres
per sec

c^3
2.694400242 e25
cubic metres
per sec

c^4
8.077608713 e33
spherical metres
per sec

Euclidean
planar
space-time

EM Light radiates bi-laterally
co-linear with its two
linear momentum vectors

Time in Physics is a measure
of how long it takes for light to
travel 299,792,458 metres
from its source in a vacuum

$$\frac{299{,}792{,}458 \text{ m}}{\text{second}}$$

$$\frac{\text{metre}}{299{,}792{,}458 \text{ sec}}$$

A metre in Physics is
the distance light travels in
1/ 299,792,458 of a second
from its source

Cartesian
rectilinear
space time

Cartesian and Curvilinear
spatial co-ordinate systems
both measure 3D Matter

Riemannian
curvilinear
space-time

Spatial geometries

based on the vector speed of Light
form distinct spatial co-ordinate systems
for the measurement of physics

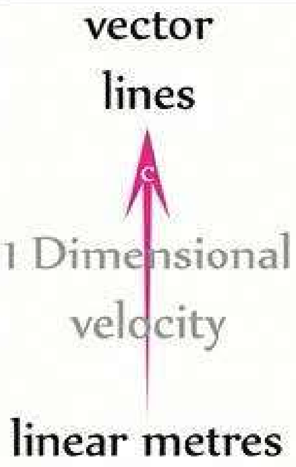
vector
lines

1 Dimensional
velocity

linear metres

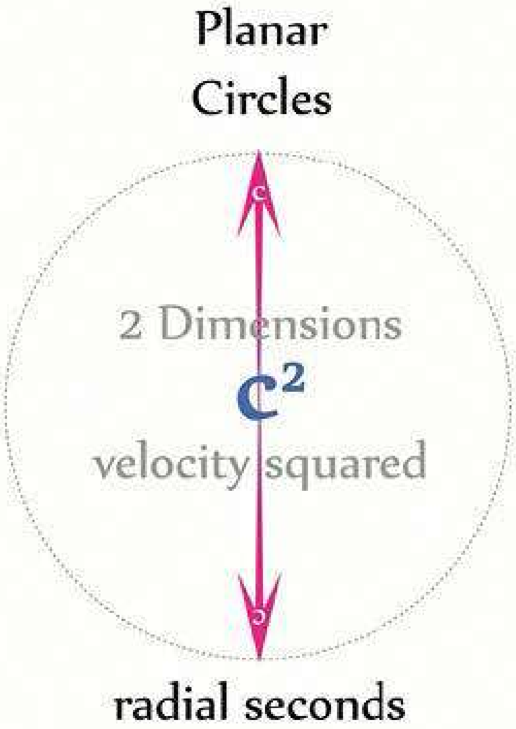
Planar
Circles

2 Dimensions
c^2
velocity squared

radial seconds

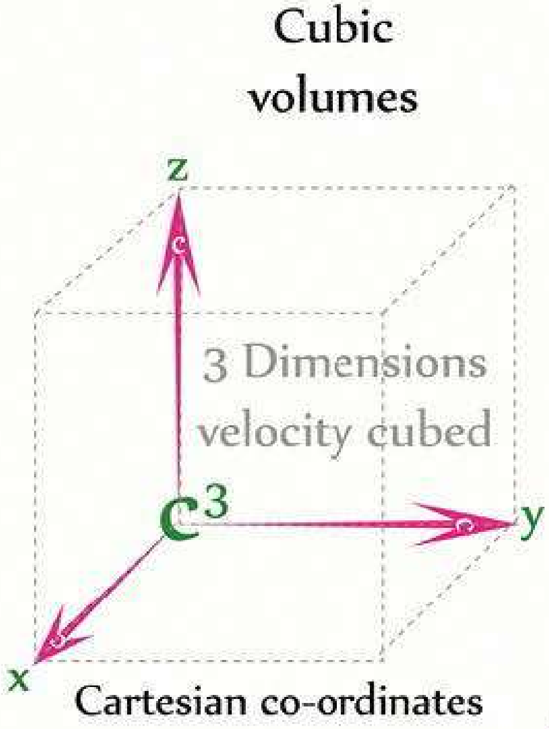
Cubic
volumes

3 Dimensions
velocity cubed
c^3

Cartesian co-ordinates

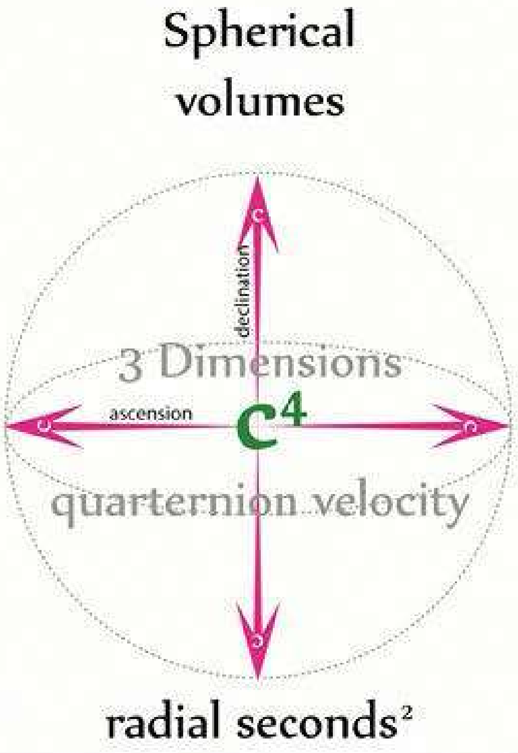
Spherical
volumes

3 Dimensions
c^4
quarternion velocity

radial seconds²

Forces

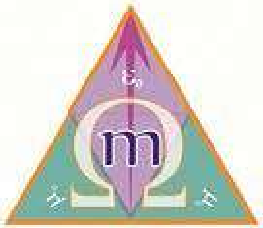

2D planar radiant geometry

mass-energies

Energy has an equilateral geometry and forms
Tetrahedral topologies within any
spatial region

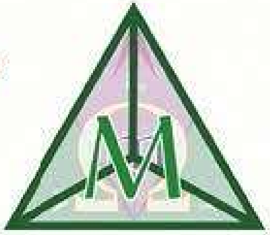

3D standing-wave topology

mass-Energy-Matter
and all forces

Copyright ABRAHAM [2008] - all rights reserved

c is the natural velocity of light

Energy has a 2D
equilateral geometry
[Euclidian]

Time - as a measure
of divergent energies
has a radial geometry

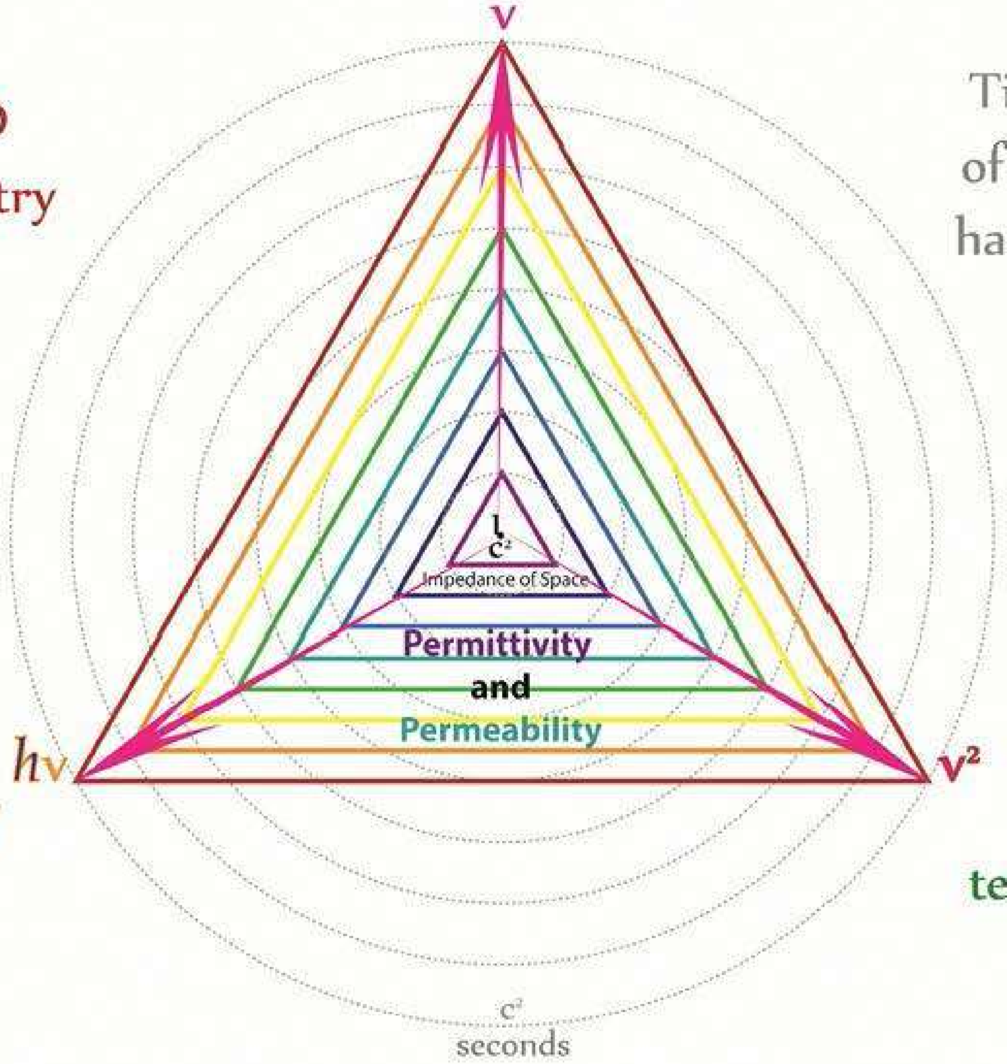
v
hv
v²
$\frac{1}{c^2}$
Impedance of Space
Permittivity
and
Permeability
c²
seconds

Equilateral geometry
is the inverse of
radial geometry

3D Matter has a
tetrahedral mass-energy
topology

The scalar spatial geometry of Energy

Copyright ABRAHAM [2008] - all rights reserved

Velocity

Velocity is the measurement of the rate and direction of change in the position of an object.

It is a vector physical quantity; both magnitude and direction are required to define it.

The absolute scalar value (magnitude) of velocity is speed, a quantity that is measured in metres per second (m/s or ms--1) when using the SI (metric) system.

v Velocity $\frac{m}{s}$

sec
c^2
is a 2D RADIAL
SPACE-TIME
co-ordinate system

$$\frac{m}{s}$$

Speed is the scalar value of the
Distance traveled per unit of TIme

$$\bar{\mathbf{v}} = \frac{\Delta x}{\Delta t}.$$

Velocity is the vector value of the
Distance traveled per unit of Time

$$\frac{\overrightarrow{m}}{s}$$

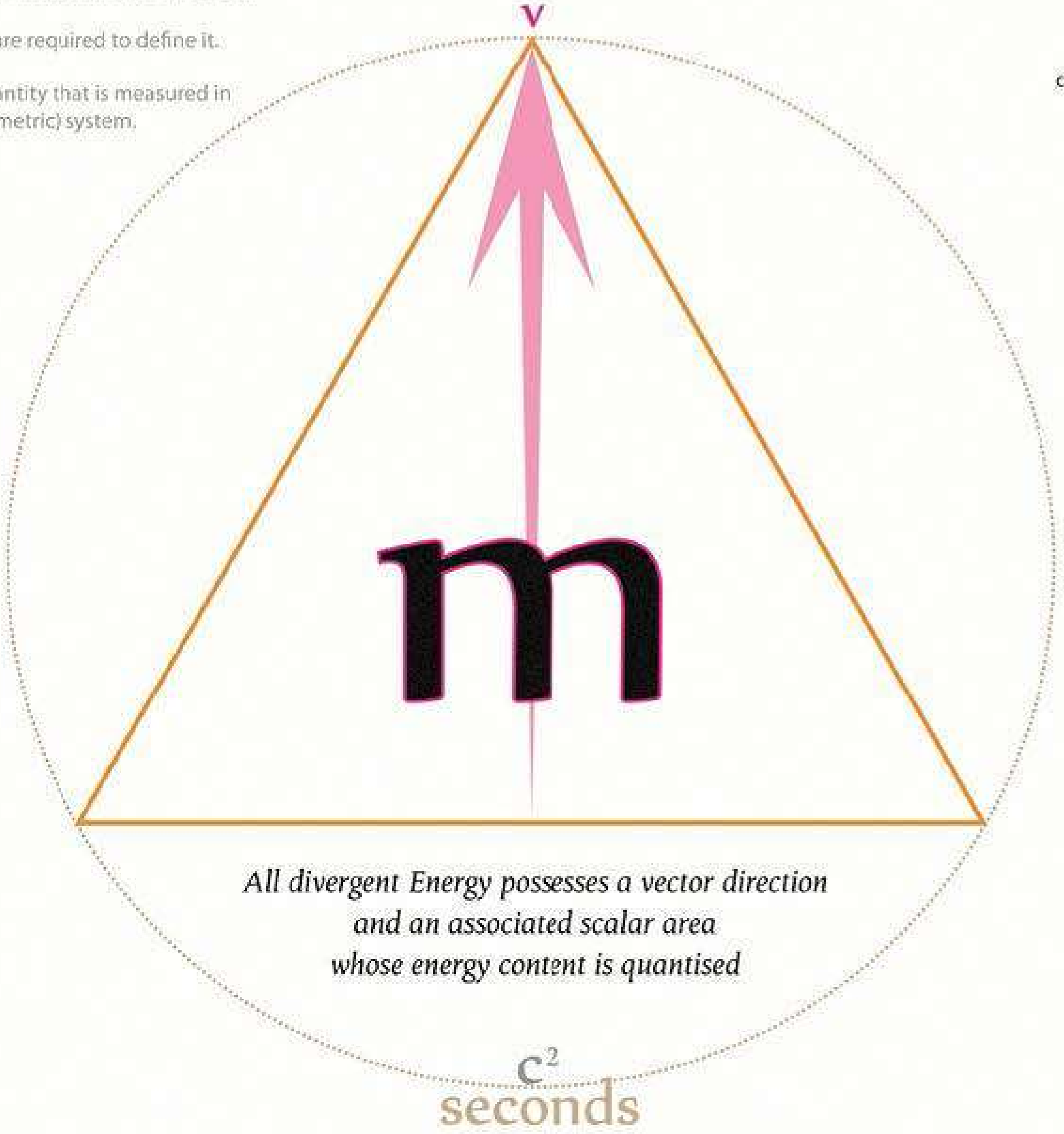

All divergent Energy possesses a vector direction
and an associated scalar area
whose energy content is quantised

Acceleration

is a 3D SPHERICAL
SPACE-TIME
co-ordinate system

a acceleration $\frac{m}{s^2}$

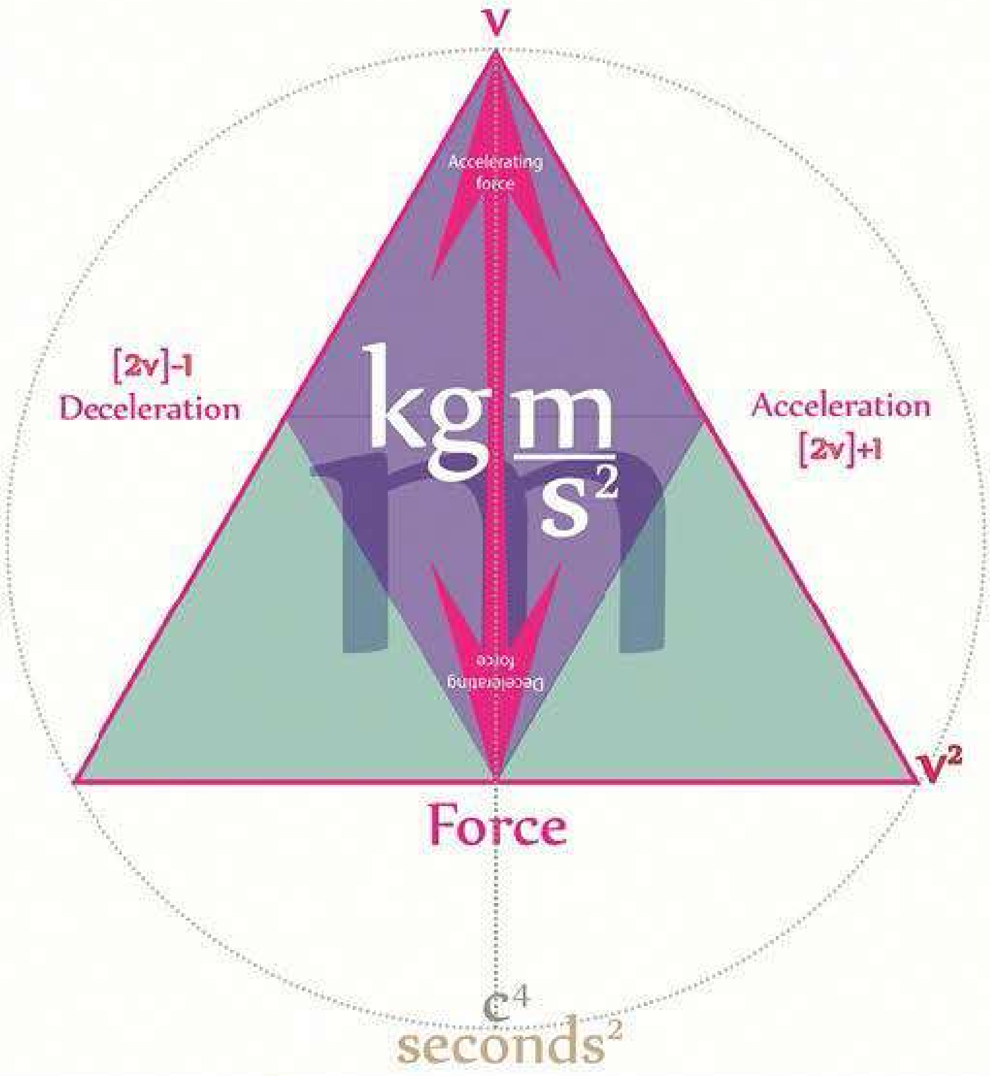
v
Accelerating force
kg m/s²
Decelerating force
[2v]-1
Deceleration
Acceleration
[2v]+1
v²
Force
c⁴
seconds²

In physics, acceleration is the rate of change of velocity [dv] over time [dt]

In one dimension, acceleration is the rate at which something speeds up or slows down.

However, since velocity is a vector, acceleration describes the rate of change of both the magnitude and the direction of velocity.

Acceleration has the dimensions [Length]/[Time Squared]
In SI units, acceleration is measured in meters per second squared (m/s^2).

$$a = \frac{\Delta y}{\Delta x} = \frac{\Delta v}{\Delta t}.$$

In classical mechanics, for a body with constant mass,
the acceleration of the body is proportional to the net force acting on it
(Newton's second law)

$$F=ma \longrightarrow a=F/m$$

Force
$$kg\,\frac{m}{s^2}$$

Additionally, for a mass with constant velocity,
(ie in an inertial frame)
the energy of motion is expressed as its momentum
(acceleration causes changes in Energy-momentum)

$$p = kg\,\frac{m}{s}$$

Quantised Angular momentum

As it is a physical [equilateral] geometry QAM is conservative in any system where there are no external Forces and serves as the foundational geometric source for all the conservation laws of physics

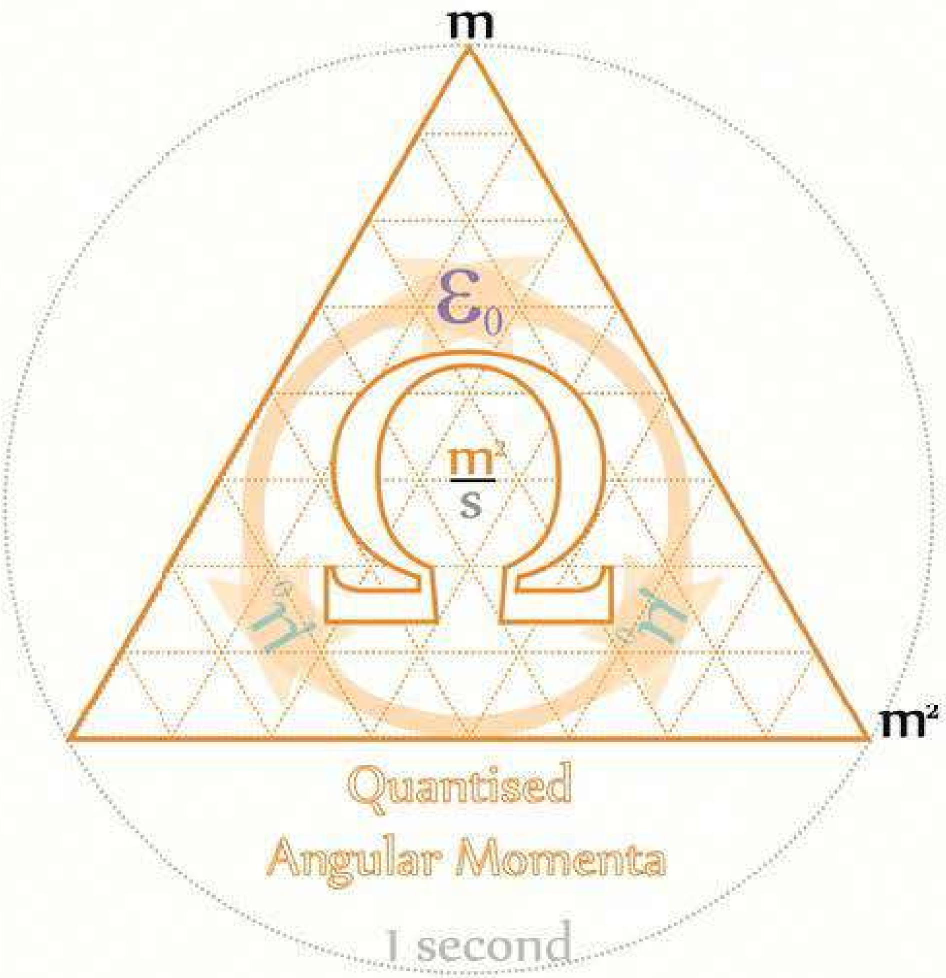
m
ε_0
Ω
$\frac{m^2}{s}$
μ_0
μ_0
m^2
Quantised
Angular Momenta
1 second

Conservation of Quantised Angular momenta

In QFT, angular momentum is is considered to be the rotational analog of linear momentum, in Tetryonics it is revealed to be the equilateral geometry of mass-energy within any defined spatial co-ordinate system

A major re-definition of
quantised angular momenta
in physics is revealed

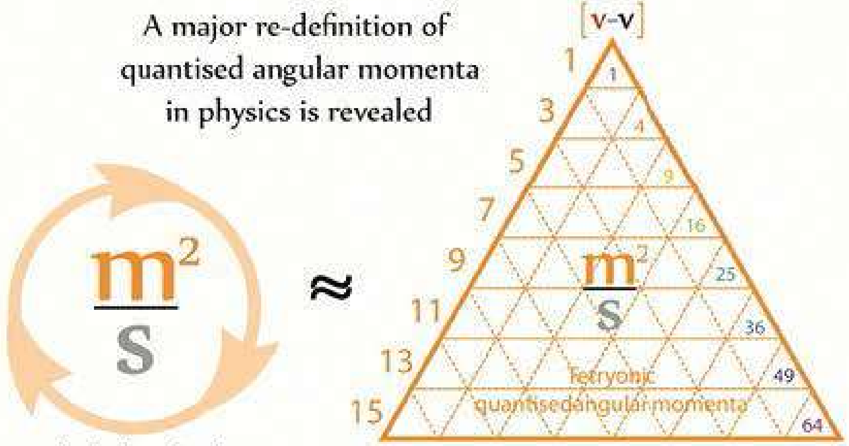
$\frac{m^2}{s}$
≈
[v-v]
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
$\frac{m^2}{s}$
Tetryonic
quantised angular momenta

classical rotational
angular momenta

Tetryonic quantum mechanics

In quantum mechanics, angular momenta is quantised – that is, it cannot vary continuously, but only in ODD number "quantum steps" between the allowed SQUARE nuclear Energy levels

In physics, angular momentum, moment of momentum,
or rotational momentum is a conserved vector quantity
that can be used to describe the overall state of a physical system.

When applied to specific mass-energy-Matter systems
QAM reveals the true quantum geometry and
nature of Energy in our universe

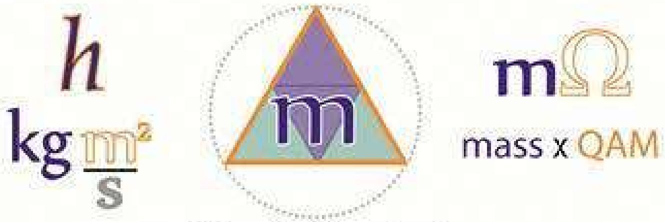
h
$kg\frac{m^2}{s}$
m
$m\Omega$
mass x QAM

Planck's Constant

Normally viewed as an expression of rotational momentum
Quantised Angular Momentum [QAM] is in fact a result of
the equilateral geometric quantization of mass-energy

mass - Energy geometry

Newton developed his
Laws using Force
$p = mv$
(linear momentum)

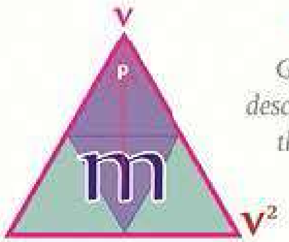

Gottfried Leibniz first described Scalar Energy as the square of velocity

Planck developed his Heat
Law using quantised mass
$h = m\Omega$
(Planck Constant)

Planck's equation for heat energy describes transverse masses [Bosons]

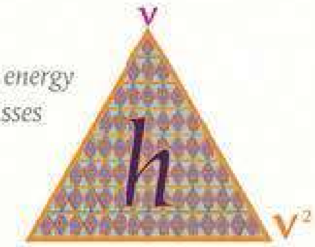

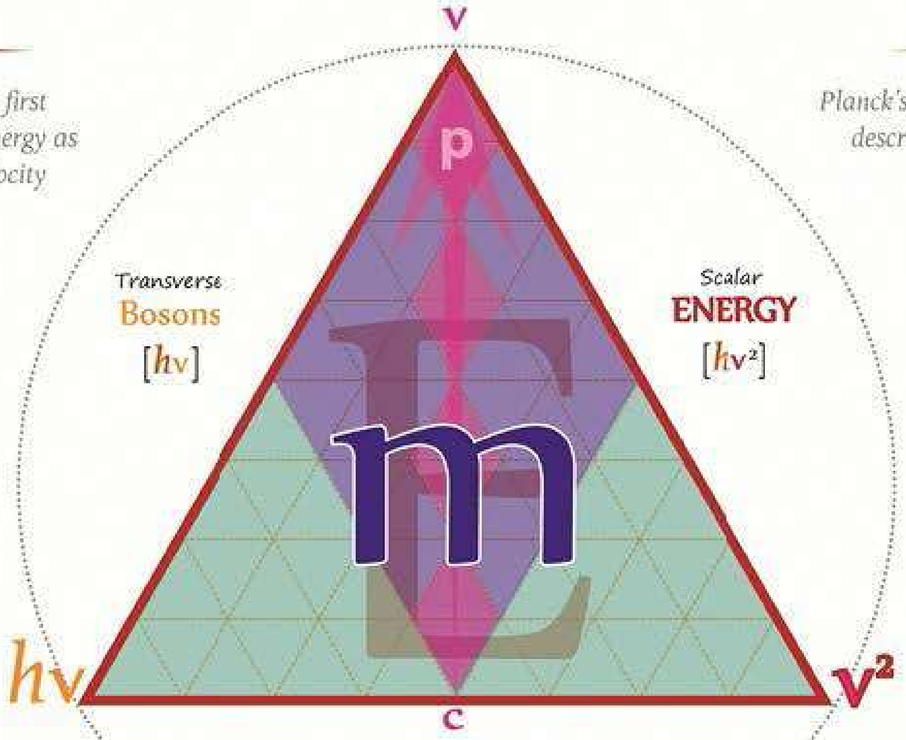

Transverse
Bosons
[hv]

Scalar
ENERGY
[hv^2]

$E = mv^2$

Leibnitz mass-energy equivalence

$E = nhv$

Planck mass-energy equivalence

mass velocity
squared

quantised mass
per second

$$mv^2 = E = hv^2$$

Leibniz-Planck mass-Energy equivalence

$$kg\,\frac{m^2}{s^2}$$

Planck quanta per second

$$\left[kg\,\frac{m^2}{s}\,\frac{1}{s}\right]$$

scalar energies can be related to velocity-momenta
through mass geometries

mv^2	*scalar* Leibniz-Newton mass-Energy equivalence	$m\Omega v^2$	*quantised* Tetryonic-Planck mass-Energy equivalence	hv^2

mass-Energy Forms

Scalar/Linear forms

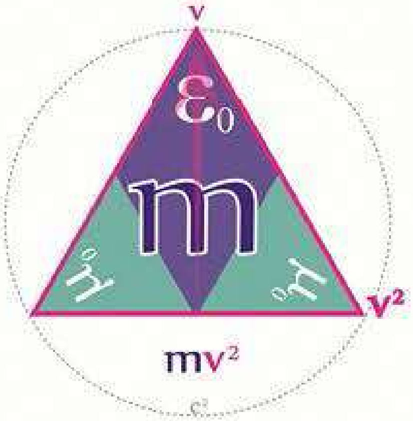

mass x velocity squared

$$\mathrm{kg}.\left[\frac{m}{s}\right]^2$$

Energy momentum

Leibniz (and Newton) showed that the Energy of a system can be viewed as a product of its mass x velocity squared or (equally as linear momentum squared), forever linking Energy to velocity through the scalar property of mass

$$mv^2 = E = p^2$$

Energy is the ability to do work in varying forms such as potential, kinetic, & mechanical energies, work, heat, and chemical or electrical energies.

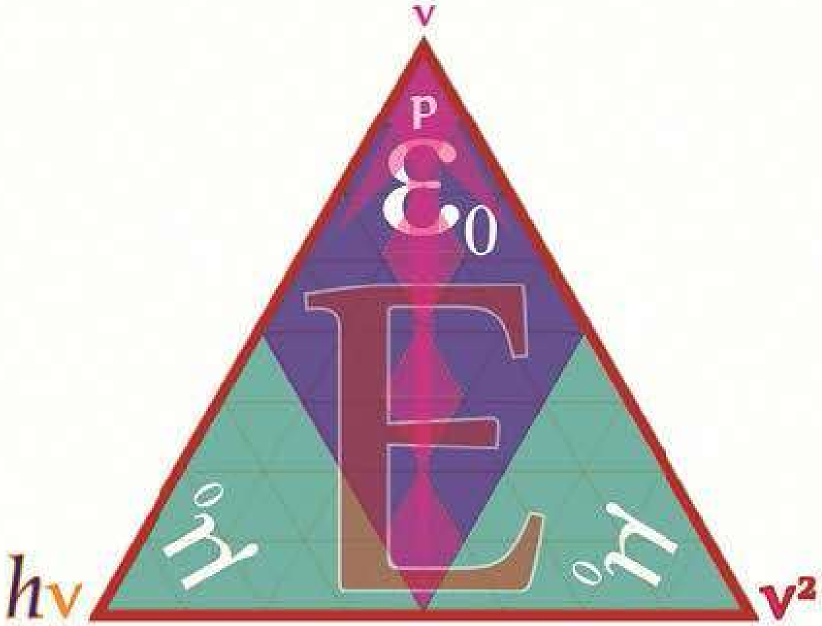

$$\boxed{mv^2 = E = hv^2}$$

The total energy contained in an object is identified with its EM mass, and Energy (like mass), cannot be created or destroyed

Tetryonic reveals mass to be a scalar measurement of quantised [equilateral] energy per unit of Time

$$m[\Omega]v^2$$

Energy is subject to the law of conservation

Quantised form

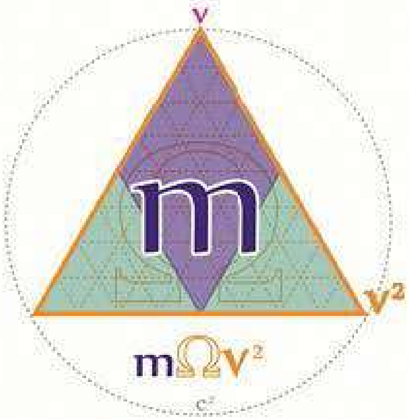

Planck quanta per second

$$\left[\mathrm{kg}.\frac{m^2}{s}\right].s^{-1}$$

mass-Energy quanta

Max Planck reveeated that energy was not continuous, it was quantized – only certain energies are allowed. Continuous energy is a scalar propertyo of mass-energy and its quantisation is the result of its equilateral geometry

$$E = hv^2$$

In quantum mechanics energy is defined in terms of the energy operator as a time derivative of the wave function

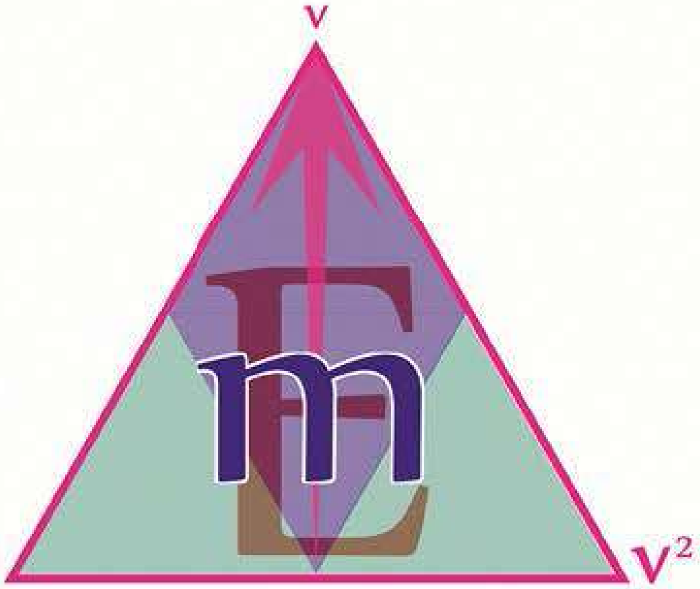

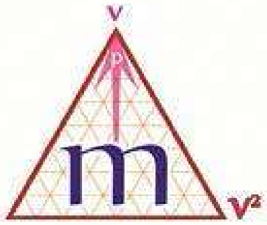

Planck quanta

$$\left[\left[m\Omega v\right]^2\right]$$

mass velocity

linear momentum squared

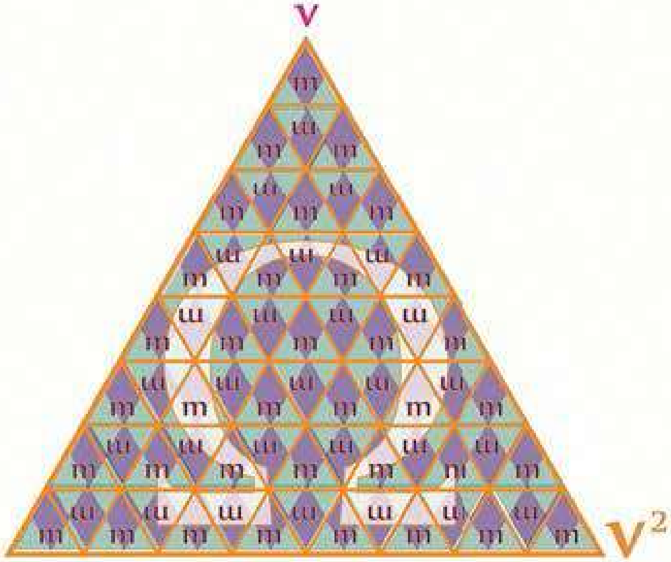

Energy is mass-velocity squared

Everything in our Universe results
from the equilateral geometry
of quantised mass-energies

Energy is Planck-quanta squared

$$E = mv^2$$

Quantised Energy

$$E = hv^2$$

scalar mass

kg

Planck quanta per second

$$\left[\left[kg\,\frac{m^2}{s}\right]\frac{1}{s}\right]$$

mass velocity squared

quantised mass

$$kg\,\frac{m^2}{s}$$

$$m = \frac{E}{v^2}$$

Scalar Energy

$$h = \frac{E}{v^2}$$

Note: There is a direct correspondence between Velocity and Planck quanta numbers
(ie as velocity varies, the energy quanta varies as well by the square of the linear change)

Copyright ABRAHAM [2008] - all rights reserved

Linear Momentum

p Momentum $\mathrm{kg}\frac{\mathrm{m}}{\mathrm{s}}$

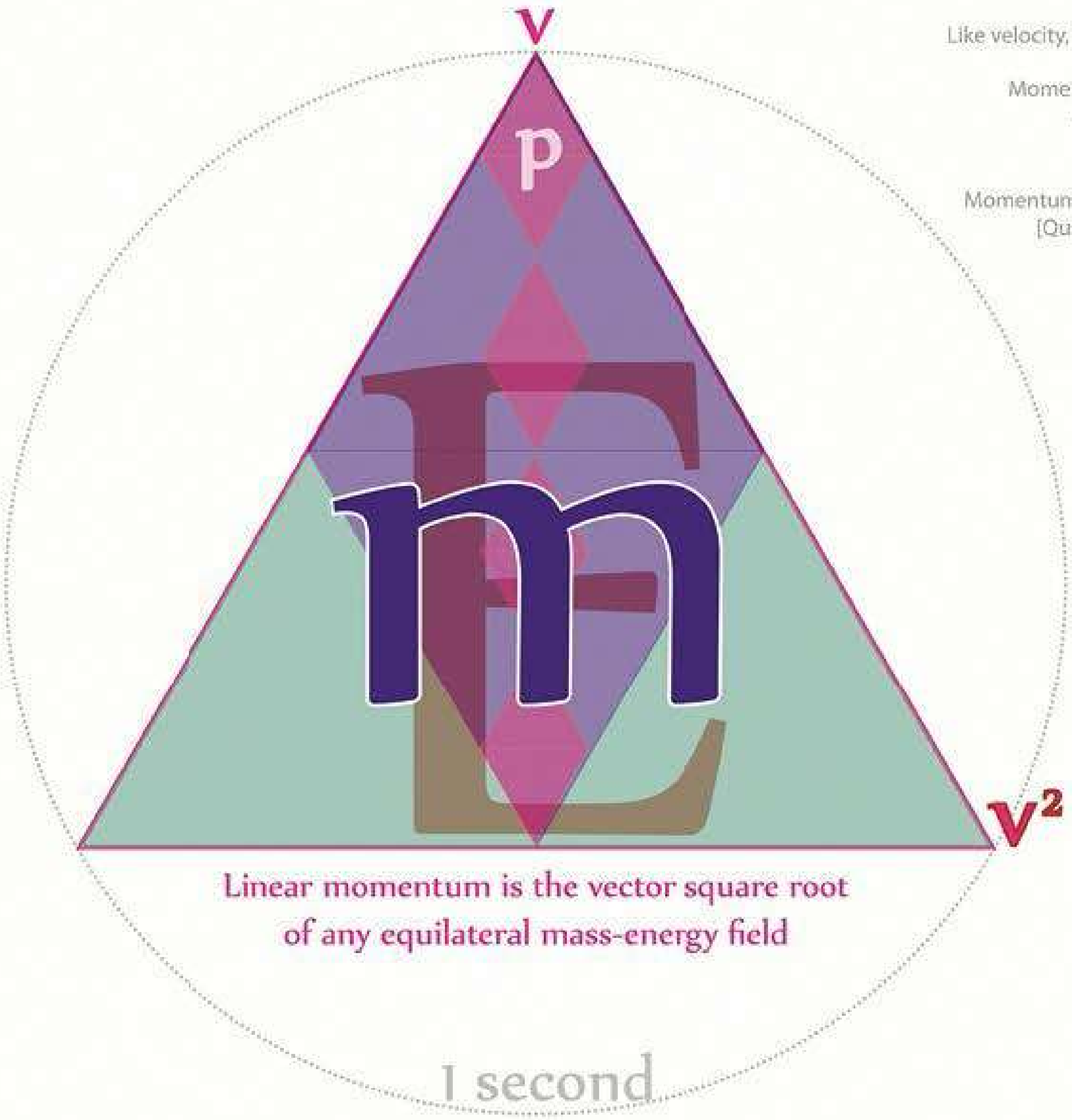

Linear momentum is the vector square root
of any equilateral mass-energy field

In classical mechanics, momentum (pl. momenta; SI unit kg·m/s, or, equivalently, N·s)
is the product of the mass and velocity of an object (p).

Like velocity, momentum is a vector quantity, possessing a direction as well as a magnitude.

Momentum is a conserved quantity (law of conservation of linear momentum),
meaning that if a closed system is not affected by external forces,
its total momentum cannot change.

Momentum should be referred to in its specific forms to distinguish it in its various forms
[Quantised Angular, Linear, Rotational and quantum/nuclear momentum]

$$\boxed{p = \frac{hv^2}{v} = mv}$$

Although originally expressed in Newton's Second Law, the conservation of momentum also holds in special relativity and, with appropriate definitions, a (generalized) momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory, and general relativity.

In relativistic mechanics, non-relativistic momentum is further multiplied by the Lorentz factor.

$$p^2 = E = mv^2$$

Energy can be expressed as
the square of linear momentum

$$\mathrm{kg}\frac{\mathrm{m}^2}{\mathrm{s}^2}$$

Copyright ABRAHAM [2008] - all rights reserved

Ω

Just as Tetyonic geometry distingushes between angular momenta and linear momentum it also distingushes between linear momentum and the vector velocities it produces

p

Linear Momentum

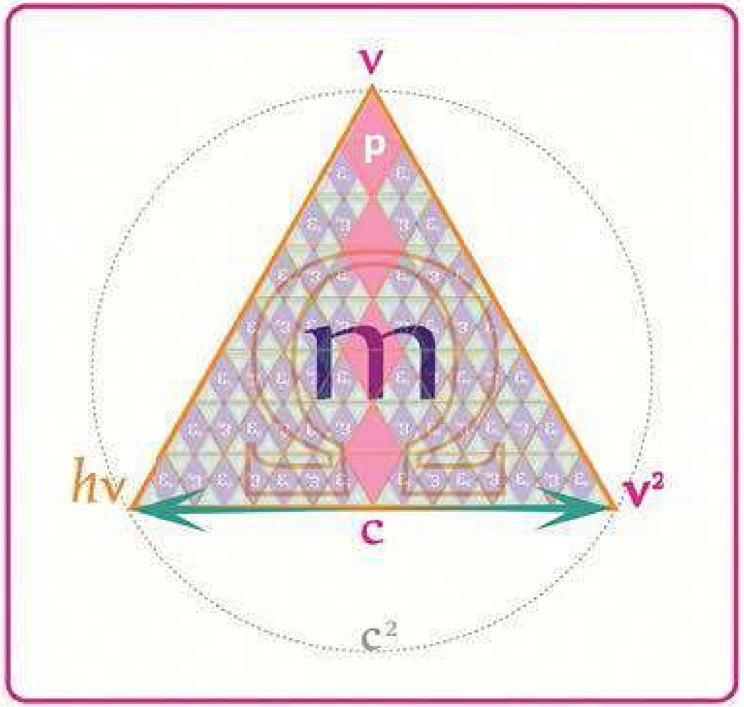

SCALAR square root

Scalars are quantities that are fully described by only their magnitude

$p = mv$

$E = pv$

$E = p^2$

$E = mv^2$

Velocity

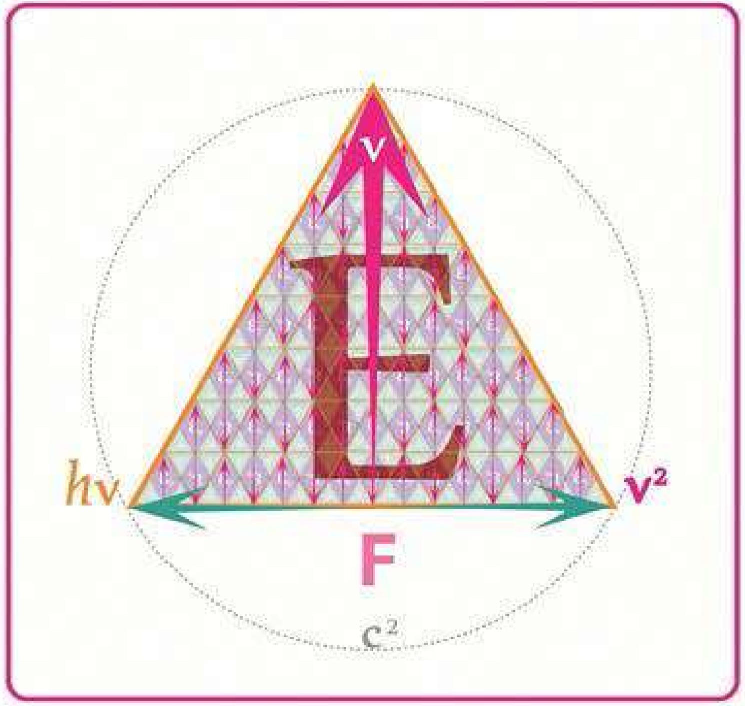

VECTOR square root

Vectors fully describe both the magnitude and direction.

p

linear momentum is a scalar component of all equilateral mass-energies that produce vector velocities

v

Tetryonics 01.12 - Momentum-velocity relationship

Copyright ABRAHAM [2008] - all rights reserved

Velocity-Quanta equivalence

Classically, the Energy of massive bodies was determined using the Newtonian mass-velocity relationship

but most recently Quantum mechanics was developed utilising the Planck's quantised Energy relationship

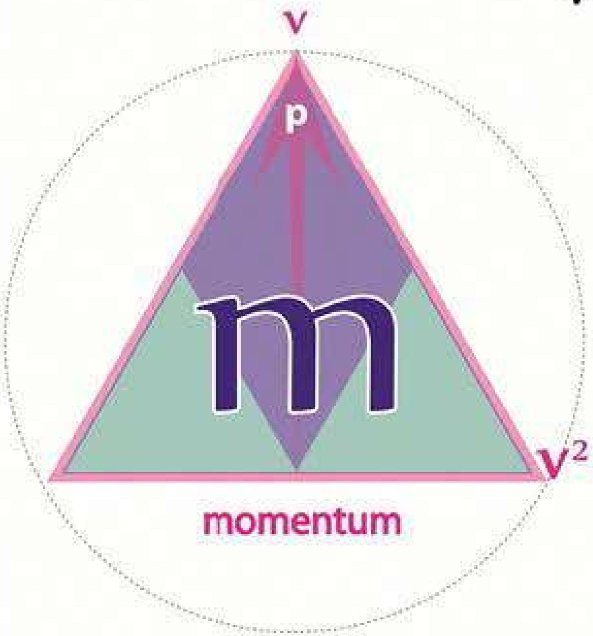

momentum

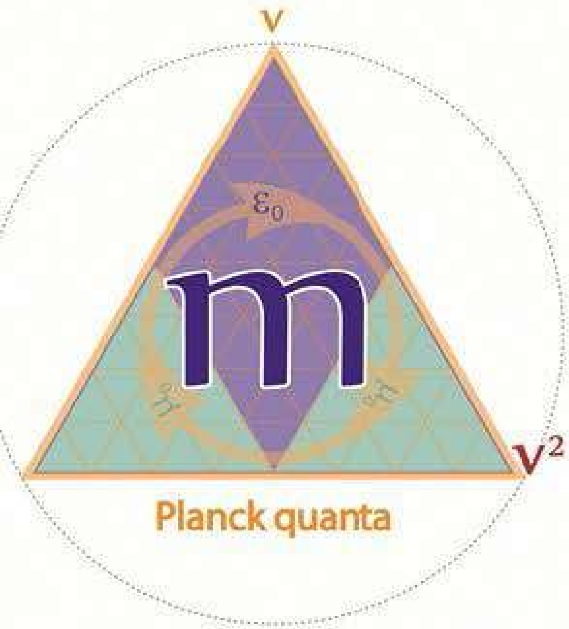

Planck quanta

classical vector force

mv^2

mass
linear velocity
squared

$kg\left[\frac{m}{s}\right]^2$

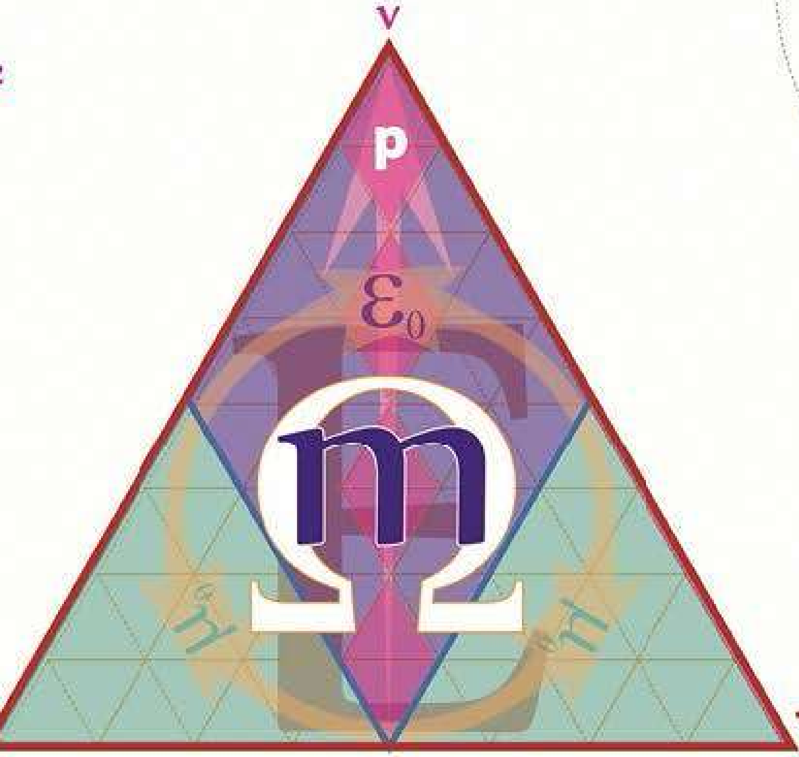

quantised energy momenta

hv^2

mass
angular momenta
per second

$kg\left[\frac{m^2}{s}\frac{1}{s}\right]$

The EM mass-Energy relationship can be revealed either by linear or angular momentum analysis

Energy-momentum relationship

The total number of Planck quanta [mass-angular momenta] in any physical system is directly related to the square of its linear momentum [mass-velocity]

$$h = \frac{E}{v^2}$$

$$m = \frac{E}{v^2}$$

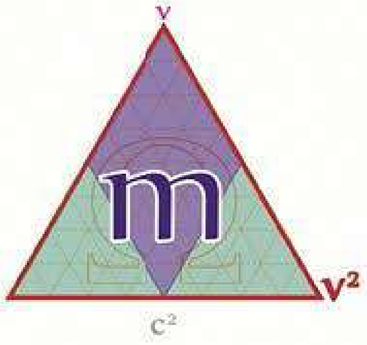

Quantised Energy-momentum

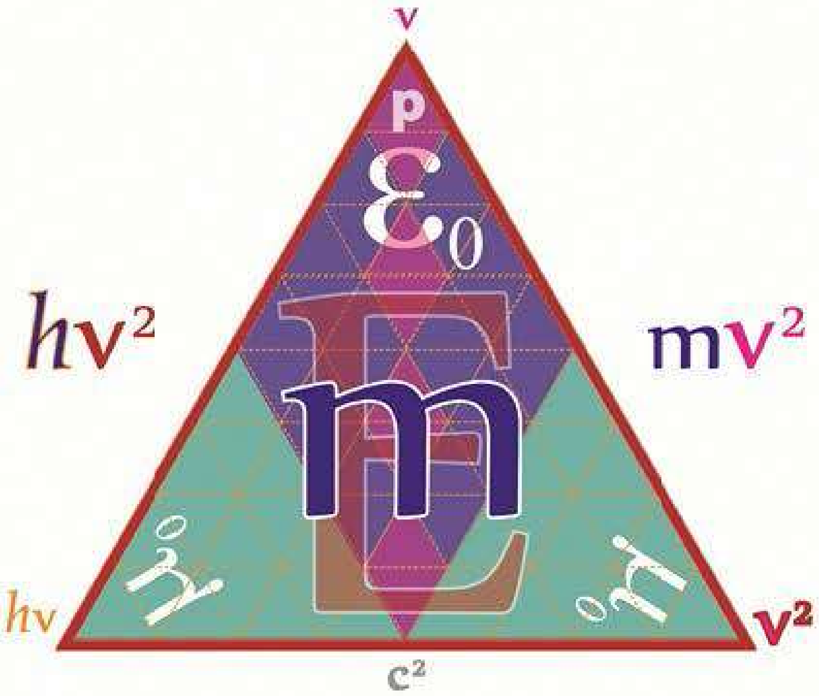

mass-energy momenta

The omega geometry of Energy produces the direct relationships between Planck's constant-quanta and mass-Energy-momentum of any spatio-temporal co-ordinate system

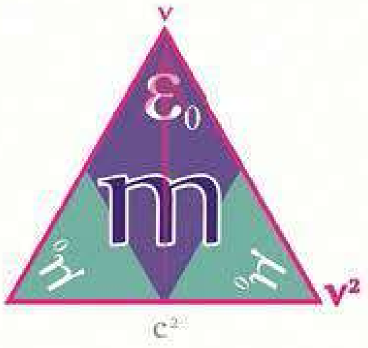

Linear Energy-momentum

$$E = m\Omega v^2$$

Quantised Angular Momenta is an equilateral geometry

$$E = pc$$

Linear momentum is a vector Force

$$m\Omega v^2 = E = mv^2$$

mass is a derived physical property relating Energy momenta to Velocity

Copyright ABRAHAM [2008] - all rights reserved

CHARGE

Charge is a measure of mass.QAM/second
[the equilateral geometry of Energy]
that gives form to all physics

ElectroMagnetic Charge
is a quantum property
resulting from the equilateral
QAM geometry of mass-Energy

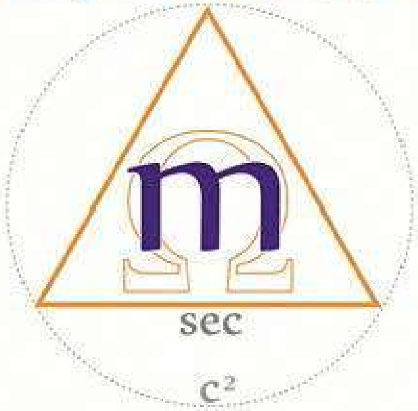

The two ElectroMagnetic charge
geometries possible can be modelled
by the flux of electrical energy
in ideal inductive loops

It is a measure of the arrangment of Planck quanta
geometries/topologies within any specific
space-time co-ordinate system

q

Clockwise inductive
energy flux

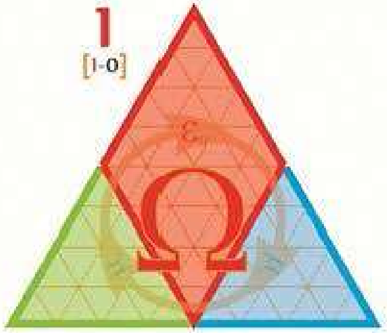

Positive charged
mass-energy momenta

$$\frac{hv}{c^2}$$

$$kg\,\frac{m^2}{s}\,\frac{s^2}{m^2}$$

$$kg.s$$

1.33518 e-20 s

Counter clockwise
energy flux

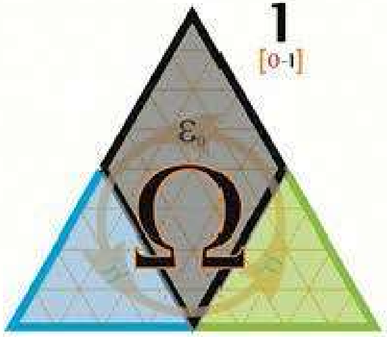

Negative charged
mass-energy momenta

q

Measuring charge geometries

Charge comes in two types, called Negative and Positive
which create the Law of Interaction (historically, the Law of Attraction)

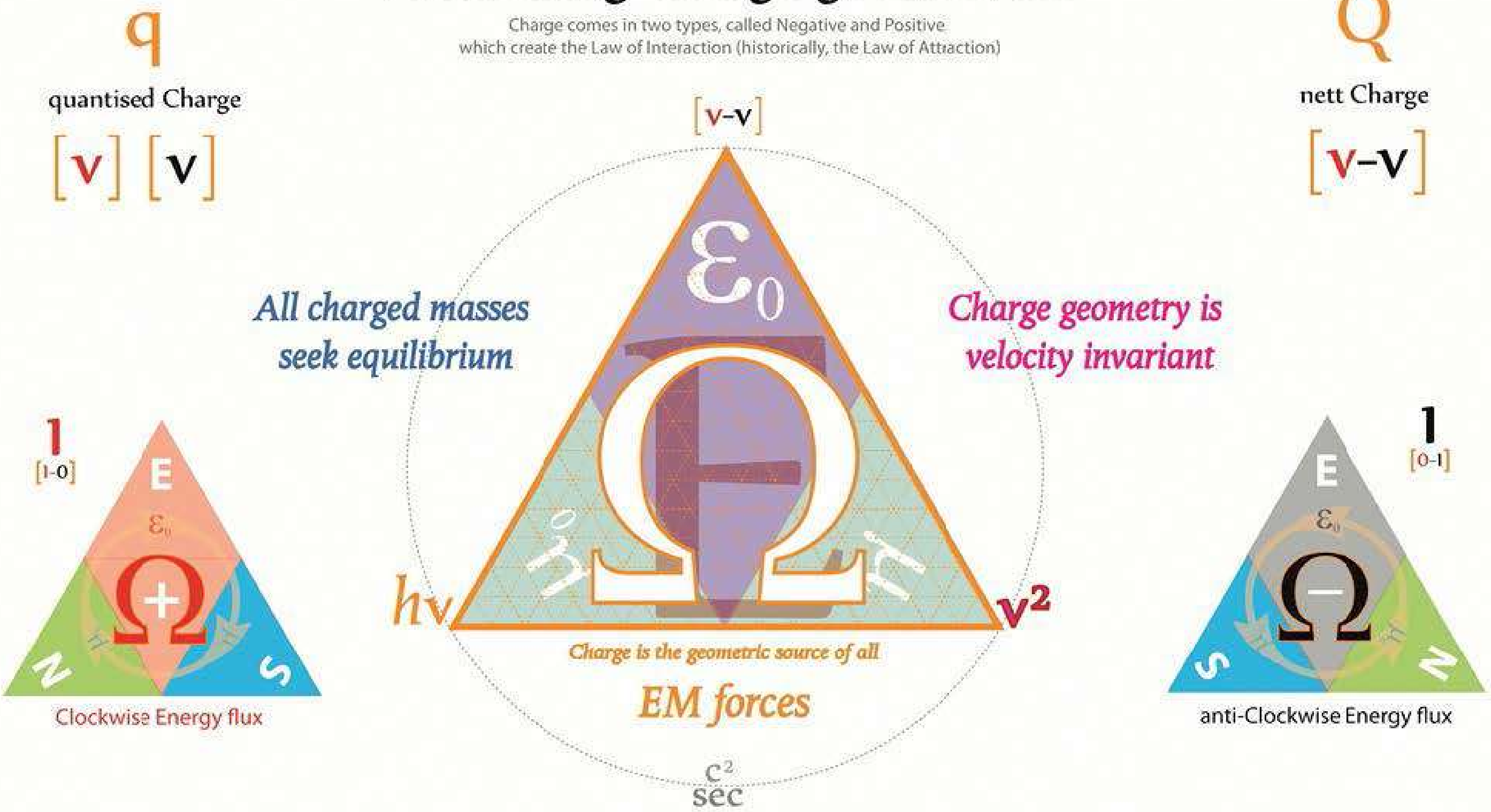
q
quantised Charge
[v] [v]
Q
nett Charge
[v-v]
[v-v]
ε_0
Ω
hv
v^2
c^2
sec
1
[1-0]
E
ε_0
Ω +
N
S
1
[0-1]
E
ε_0
Ω –
S
N

All charged masses seek equilibrium

Charge geometry is velocity invariant

Charge is the geometric source of all EM forces

Clockwise Energy flux

anti-Clockwise Energy flux

Historically defined as a physical property of mass-Matter that causes it to experience a force charge is actually the result of the equilateral geometry of qantised mass angular momenta which provides the physical mechanics of mass-ENERGY-Matter differentiation and interaction

The SI unit of quantity of charge is the Coulomb, $C \approx$ mass seconds $\approx$ kg.s which can also be expressed as Amp-seconds

EM Field Geometry

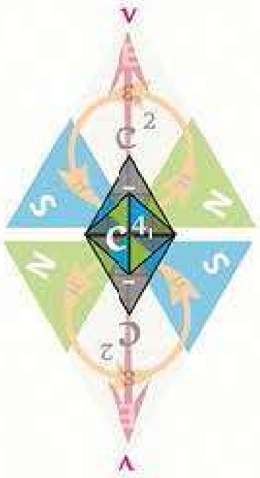

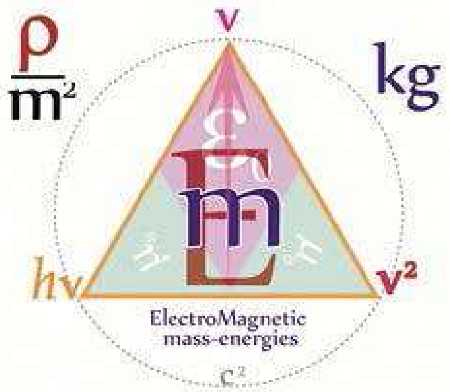
ρ
m²
v
kg
ε
m
E
hν
v²
ElectroMagnetic
mass-energies
c²

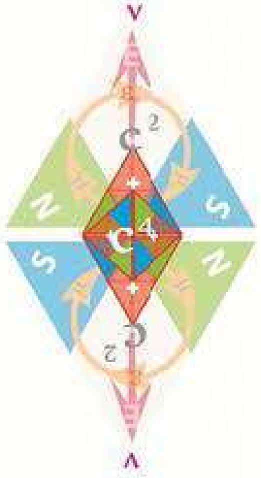

mass-energies are 2D radiant EM field geometries

[299,792,458 m/s]

'c' forms a CONSTANT of proportionality for different spatio-temporal co-ordinate systems used to measure mass-ENERGY-Matter

All Matter are 3D EM standing wave topologies

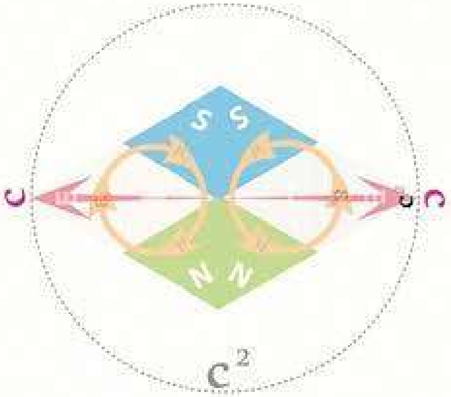

EM fields create 'interaction-at-a-distance'

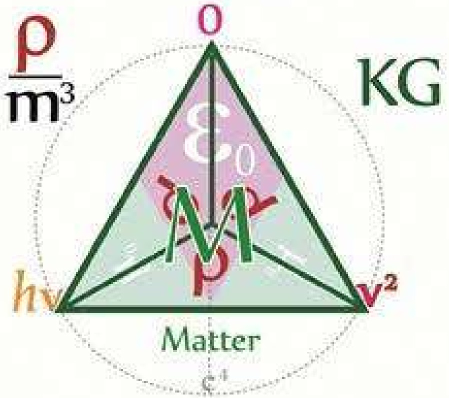
ρ
m³
0
KG
ε0
M
hν
v²
Matter
c⁴

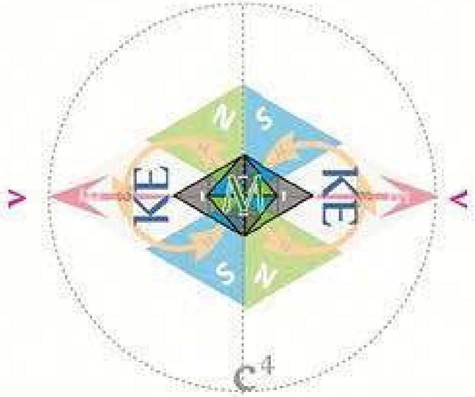
v
KE
N S
M
S N
KE
v
c⁴

Electrostatic Matter has opposing 2D KEM fields

Copyright ABRAHAM [2008] - all rights reserved

Energy quantisation

All mass-Energy-Matter can only have certain integer Energy-momenta,
[mass-Energy in all its forms is QUANTISED]

q

Net quanitised angular momenta
[inductive energy flux]
determines Charge

$$m\left[\frac{\Omega}{c^2}\right]$$

$$kg\,\frac{m^2}{s}\cdot\frac{s^2}{m^2}$$

seconds

Transverse quanta
create Quantum Levels

The equilateral geometry of
quantised angular momentum
creates charged masses

Planck's Constant is in fact
mass x equilateral QAM

$$kg.\frac{m^2}{s}$$

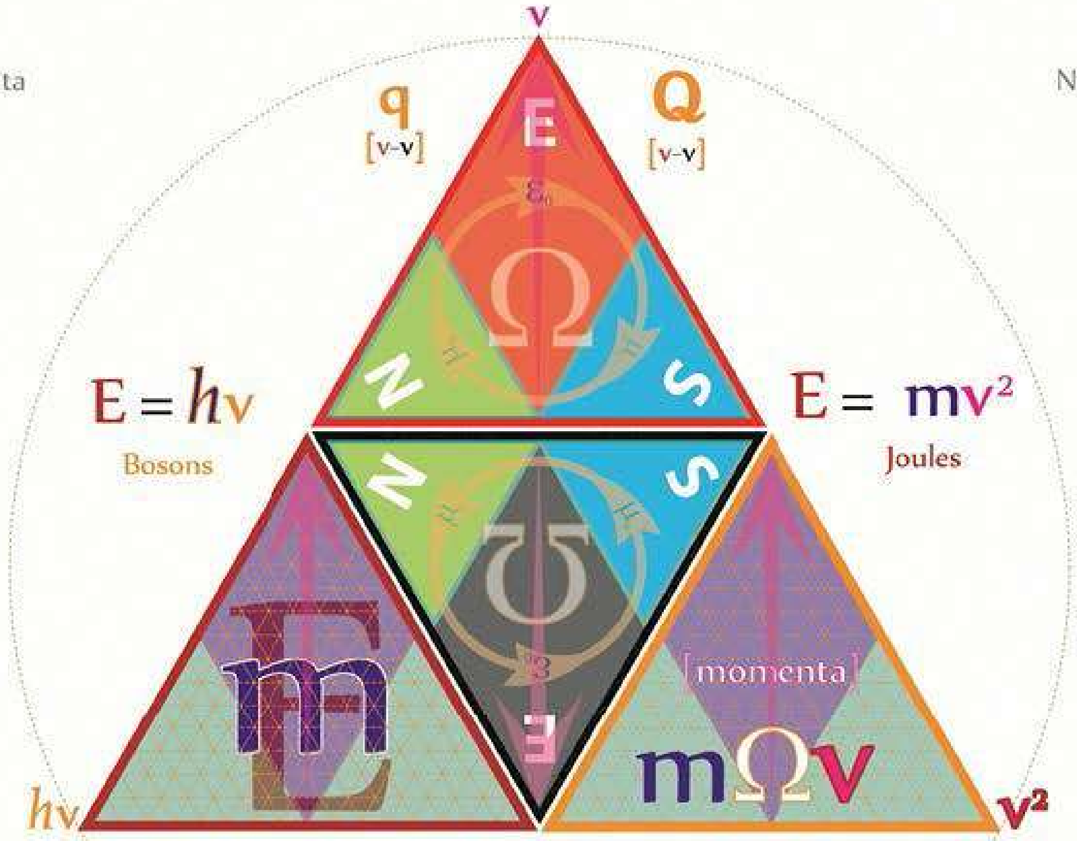

$$kg.\frac{m^2}{s}$$

Planck's Constant

n.Planck quanta per second is Energy

q

Net quanitised angular momenta
[inductive energy flux]
determines Charge

$$m\left[\frac{\Omega}{c^2}\right]$$

$$kg\,\frac{m^2}{s}\cdot\frac{s^2}{m^2}$$

seconds

Scalar (nett) quanta
create Square Energies

Planck's Constant
can be described in
a number of differing ways

Scalar Energy x Time	$kg.\frac{m^2}{s^2}.\,s$
Linear momentum x Distance	$kg.\frac{m}{s}.\,m$

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic Mnemonics

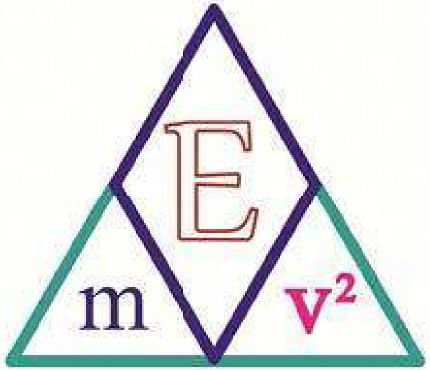

Energy momentum

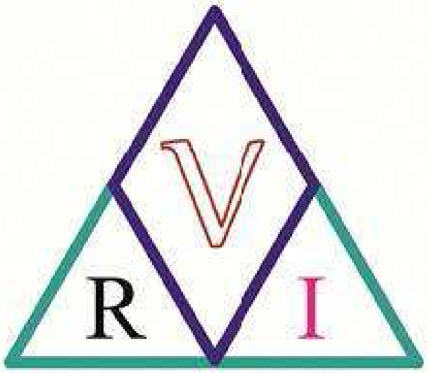

Electrical flow

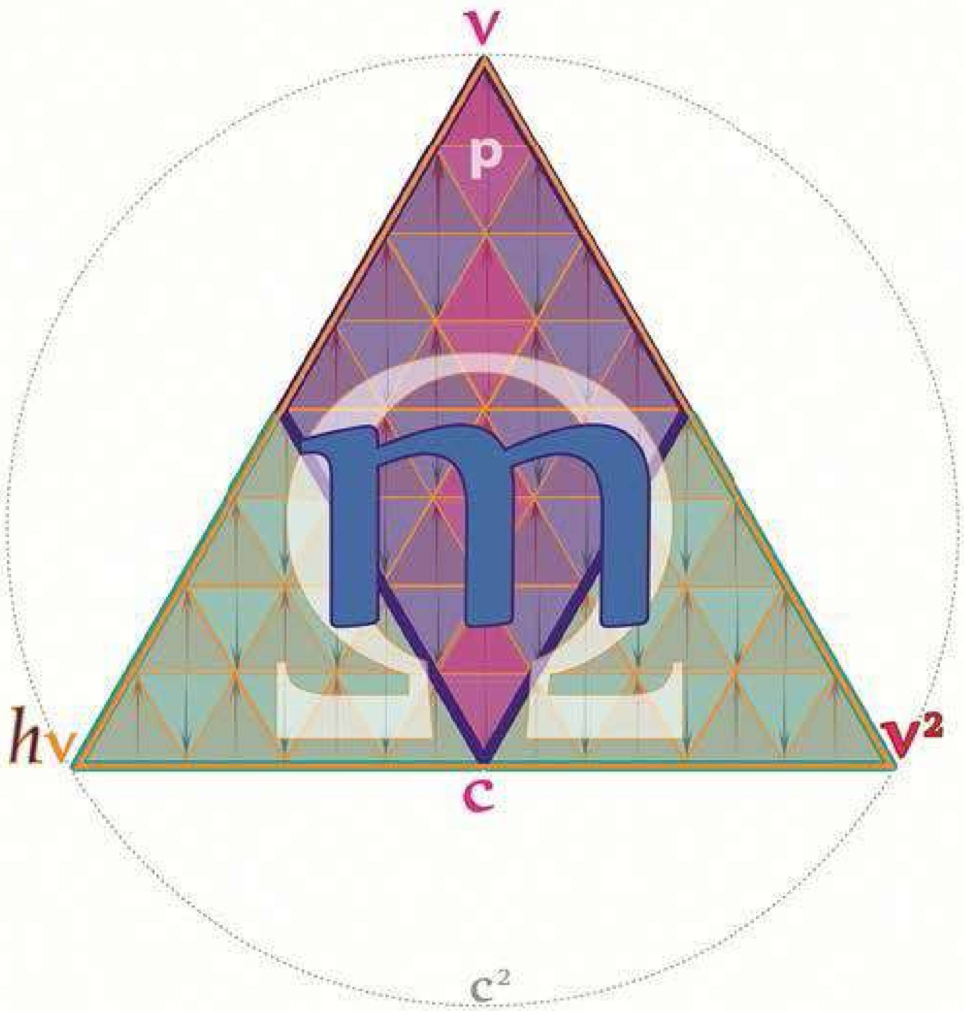
v
p
m
Ω
hv
v²
c
c²

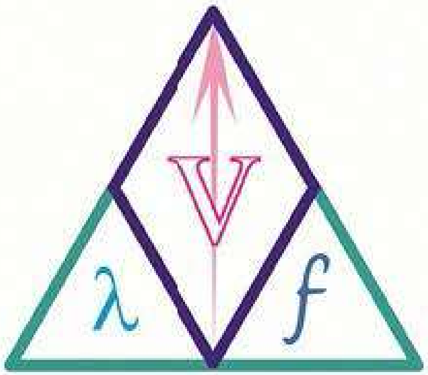

velocity of propagation

Many physical relationships are can be represented with DELTA mnemonics

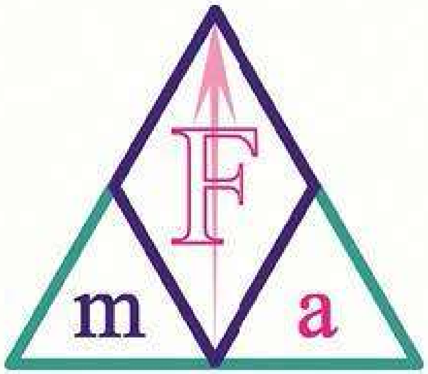

Inertial Force

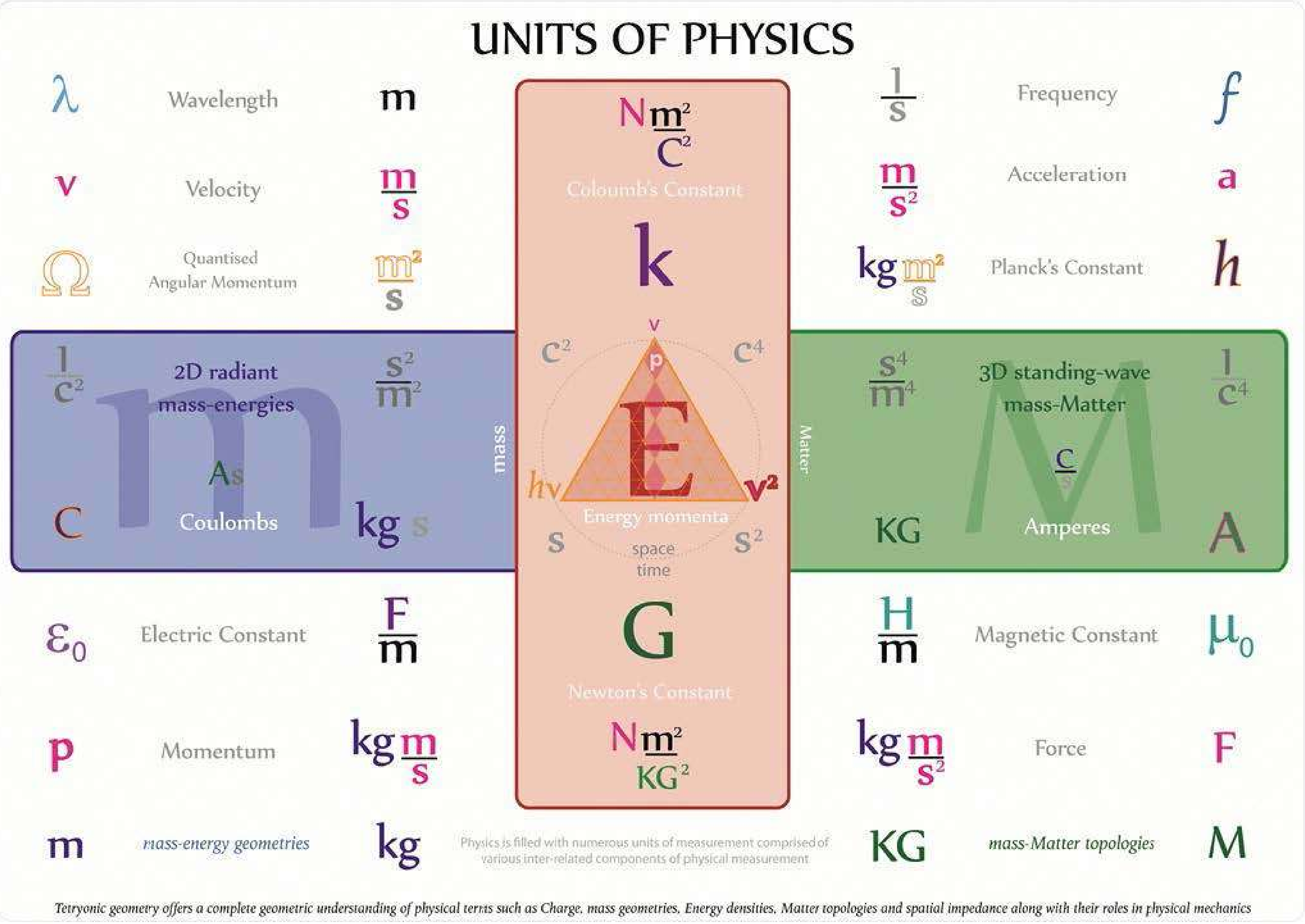

UNITS OF PHYSICS

λ	Wavelength	m	$\frac{Nm^2}{C^2}$ Coloumb's Constant	$\frac{1}{s}$	Frequency	f
v	Velocity	$\frac{m}{s}$		$\frac{m}{s^2}$	Acceleration	a
Ω	Quantised Angular Momentum	$\frac{m^2}{s}$	k	$kg\,\frac{m^2}{s}$	Planck's Constant	h
$\frac{1}{c^2}$	2D radiant mass-energies	$\frac{s^2}{m^2}$		$\frac{s^4}{m^4}$	3D standing-wave mass-Matter	$\frac{1}{c^4}$
C	As Coulombs	$kg\ s$		KG	$\frac{C}{s}$ Amperes	A
ε_0	Electric Constant	$\frac{F}{m}$	G Newton's Constant	$\frac{H}{m}$	Magnetic Constant	μ_0
p	Momentum	$kg\,\frac{m}{s}$	$\frac{Nm^2}{KG^2}$	$kg\,\frac{m}{s^2}$	Force	F
m	*mass-energy geometries*	kg	Physics is filled with numerous units of measurement comprised of various inter-related components of physical measurement	KG	*mass-Matter topologies*	M

Tetryonic geometry offers a complete geometric understanding of physical terms such as Charge, mass geometries, Energy densities, Matter topologies and spatial impedance along with their roles in physical mechanics

Tetryonics 01.20- Physical Units

Copyright ABRAHAM [2008] - all rights reserved

Scalar mass-energies

have an EQUILATERAL geometry
(60°- 60°- 60°)

$$\frac{\sqrt{3}}{2}\,s$$

The Area of an Equilateral triangle

½ base x height

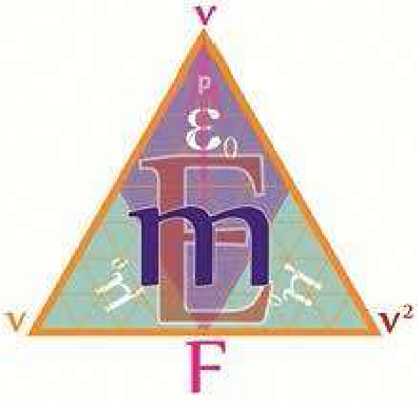

v
p
ε_0
E
m
v
v^2
F

$$\sin\frac{\pi}{3} = \sin 60^\circ = \frac{\sqrt{3}}{2}$$

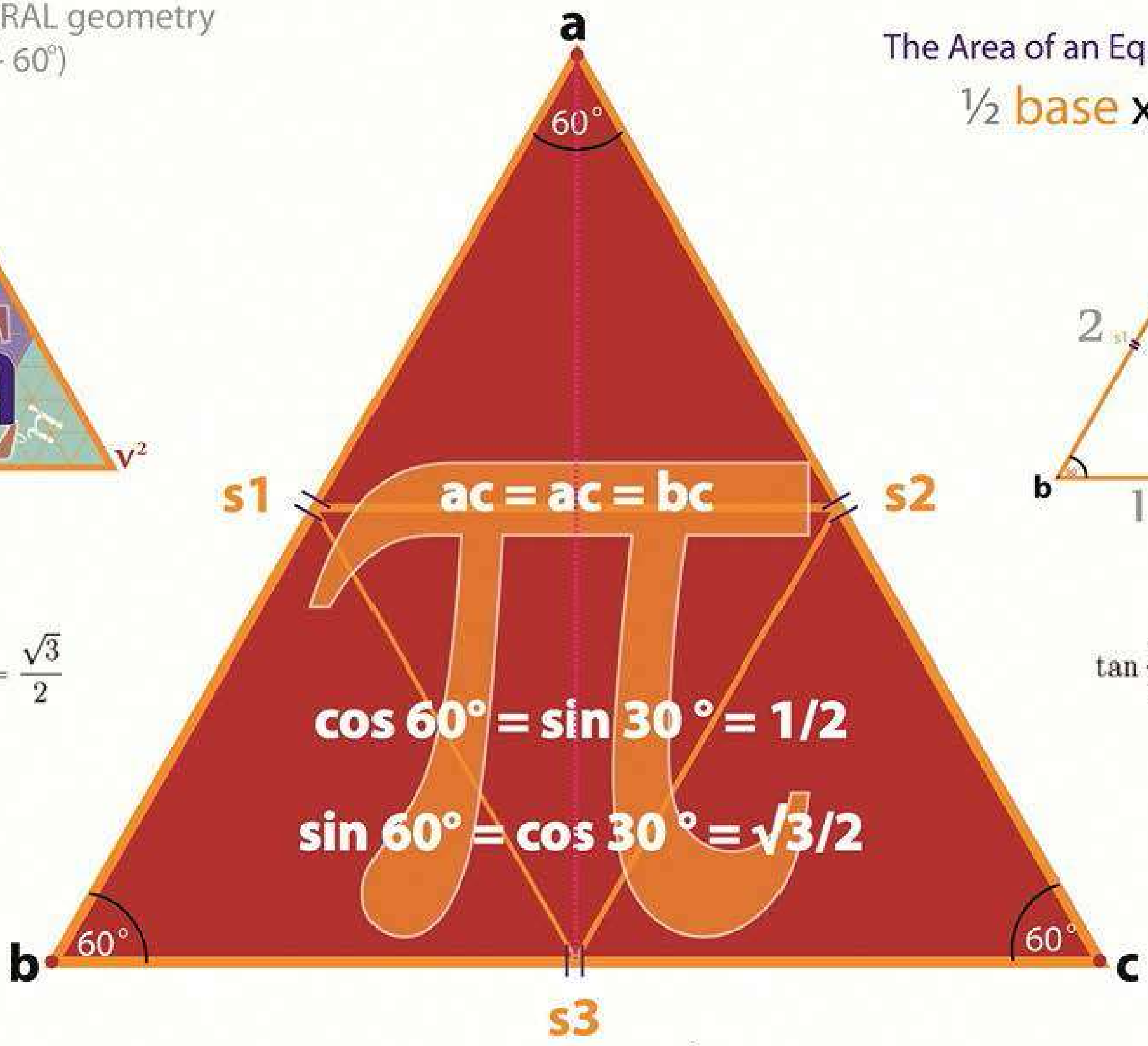

a
60°
s1
s2
b
60°
c
60°
s3

ac = ac = bc

cos 60° = sin 30 ° = 1/2

sin 60° = cos 30 ° = √3/2

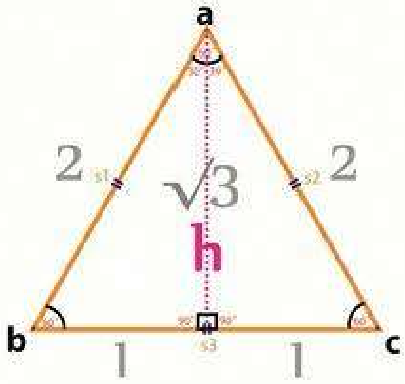

a
2
s1
√3
h
2
s2
b
c
s3
1
1

$$\tan\frac{\pi}{3} = \tan 60^\circ = \sqrt{3}$$

$$\cos\frac{\pi}{3} = \cos 60^\circ = \frac{1}{2}$$

Copyright ABRAHAM [2008] - all rights reserved

Pythagorian geometry

Energy geometries within Physics including Special Relativity and Lorentz corrections have historically been incorrectly illustrated as having the geometry of right angled triangles

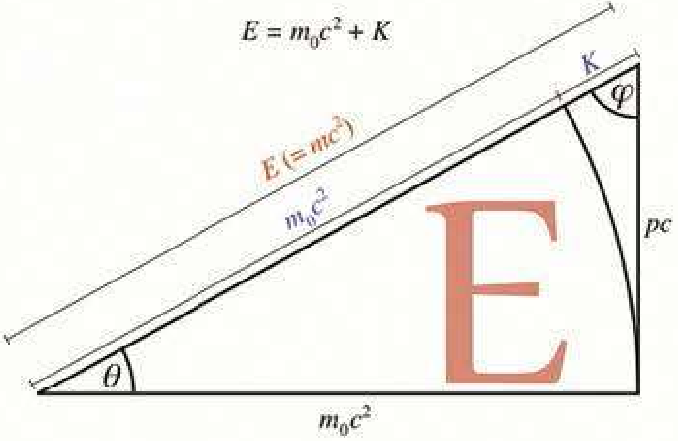

$$E^2 = p^2c^2 + m_0^2c^4$$

Generalizing, we see that the square of the total mass-energies is the sum of the components squared.
[shown incrorectly formulated in this above equation]

We can see an origin of distance in spacetime relating to velocity in pc in which Energy is subject to Lorentz corrections [v/c]

$$E = pc.$$

Additionally, EM mass can be directly related to the energy content of a body by the velocity of Energy

$$E = mc^2$$

Physics is geometry,
one cannot be separated from the other

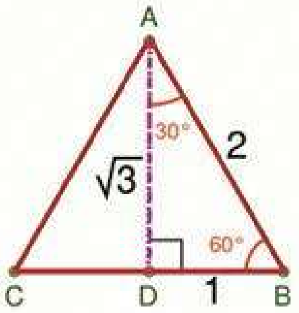

$$mv^2 = E = hv^2$$

There are three ways to look at geometry
– mathematically, verbally, and visually,

Of the three, Visually will be shown to be superior leading to intuitive understandings of Physics, Chemistry, Electrodynamics and Gravitation along with all their related physical attributes

6.629432672 e-34 J

Planck quanta

$$[m\Omega v]^2$$

mass velocity

7.376238634 e-51 kg

$$mv = p$$

Tetryonic geometry

The source of all the physical relationships of mass-Energy momenta & their constants is the geometry of equilateral Planck Triangles (and all texts must be corrected)

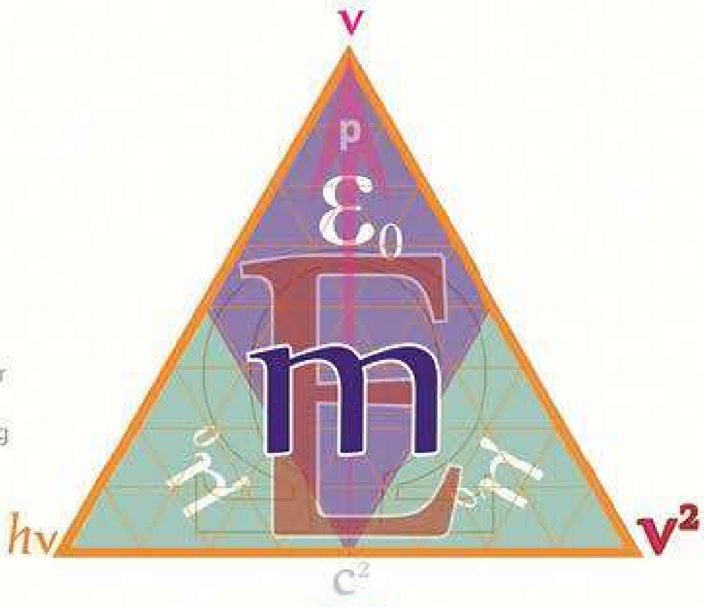

F

$$\sum \mathbf{F} = \frac{d\mathbf{P}}{dt} = m\frac{d\mathbf{v}}{dt} = m\mathbf{a}$$

Newton's Second law of Motion
is based on changes to linear momentum

$$\mathbf{F} = m\mathbf{a}\,.$$

Tetryonics and Pi radians

Although not historically considered a physical constant, π appears routinely in equations describing fundamental principles of the Universe, due in no small part to its relationship to the nature of the circle and, correspondingly, spherical coordinate systems.

The quantised equilateral geometry of mass-energy momenta is measured in π radians

Using units such as Planck units can sometimes eliminate π from formulae.

Heisenberg's uncertainty principle, which shows that the uncertainty in the measurement of a particle's position (Δx) and momentum (Δp) can not both be arbitrarily small at the same time:

$$\Delta x\,\Delta p \geq \frac{h}{4\pi} = \frac{\hbar}{2}$$

Einstein's field equation of general relativity:

$$R_{ik} - \frac{g_{ik}R}{2} + \Lambda g_{ik} = \frac{8\pi G}{c^4}T_{ik}$$

The cosmological constant Λ from Einstein's field equation is related to the intrinsic energy density of the vacuum ρvac via the gravitational constant G as follows:

$$\Lambda = 8\pi G\ \rho_{vac}$$

Coulomb's law for the electric force, describing the force between two electric charges (q1 and q2) separated by distance r:

$$F = \frac{|q_1 q_2|}{4\pi\varepsilon_0 r^2}$$

Magnetic permeability of free space relates the production of a magnetic field in a vacuum by an electric current in units of Newtons (N) and Amperes (A):

$$\mu_0 = 4\pi \cdot 10^{-7}\,\mathrm{N/A^2}$$

Kepler's third law constant, relating the orbital period (P) and the semimajor axis (a) to the masses (M and m) of two co-orbiting bodies:

$$\left(\frac{2\pi}{P}\right)^2 a^3 = \omega^2 a^3 = G(M+m)$$

and the Gaussian formula for a Normal Distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-(x-\mu)^2/(2\sigma^2)}$$

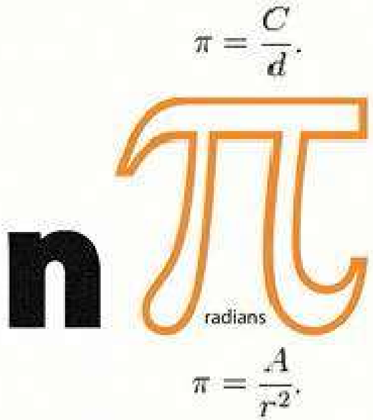

$$\pi = \frac{C}{d}.$$

$$\pi = \frac{A}{r^2}.$$

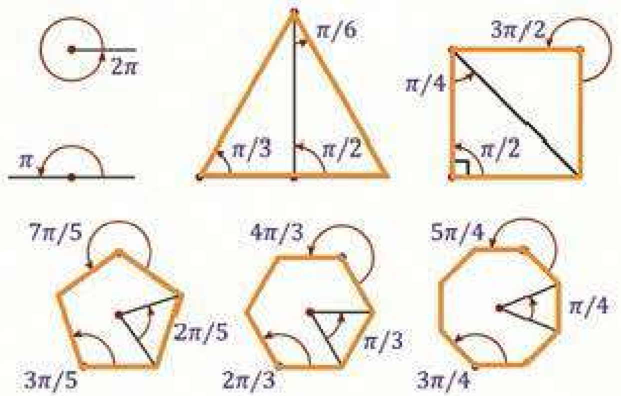

Pi radian mass-energy fascia geometries

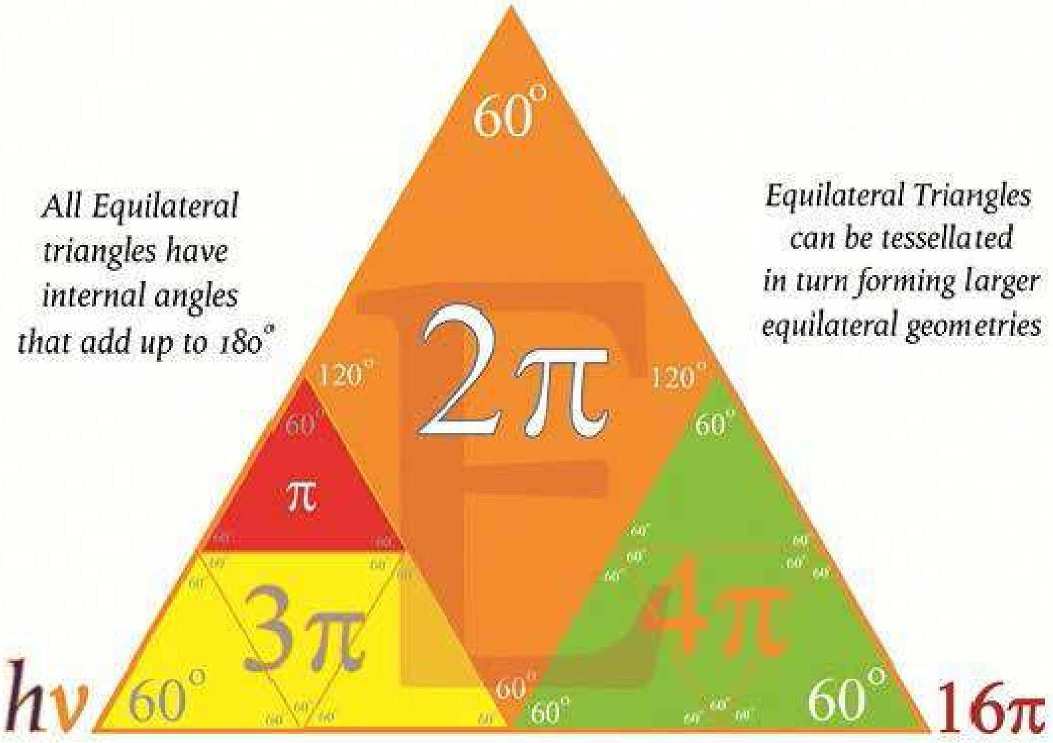

All Equilateral triangles have internal angles that add up to 180°

Equilateral Triangles can be tessellated in turn forming larger equilateral geometries

SQUARED numbers in physics are EQUILATERAL geometries

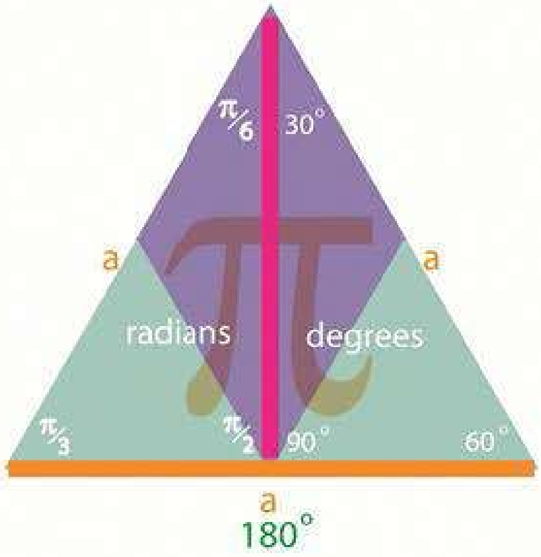
π/6
30°
a
a
π
radians
degrees
π/3
π/2
90°
60°
a
180°

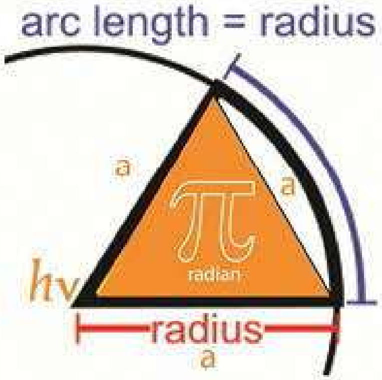
arc length = radius
a
a
π
radian
hv
radius
a

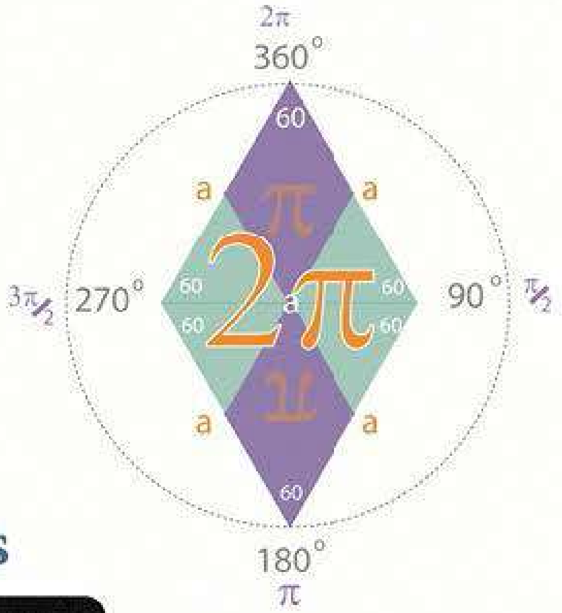
2π
360°
60
a
a
π
3π/2
270°
60
60
2π
a
60
60
90°
π/2
a
a
60
180°
π

nπ mass-energy geometries

180° π An equilateral triangle of mass-energy momenta has a geometry of π radians 2π 360°

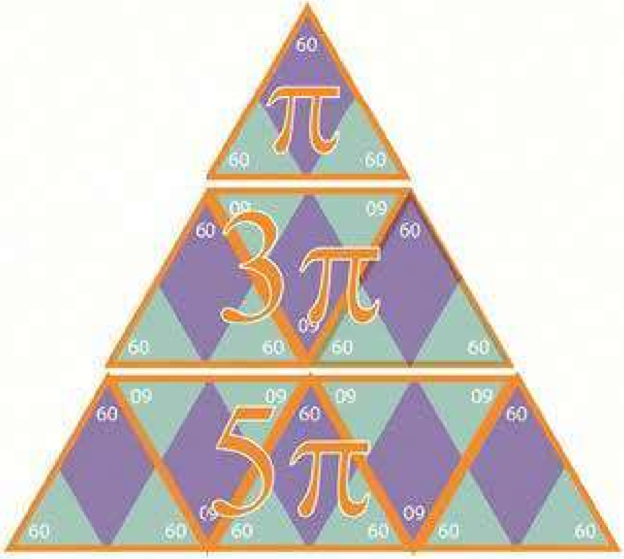

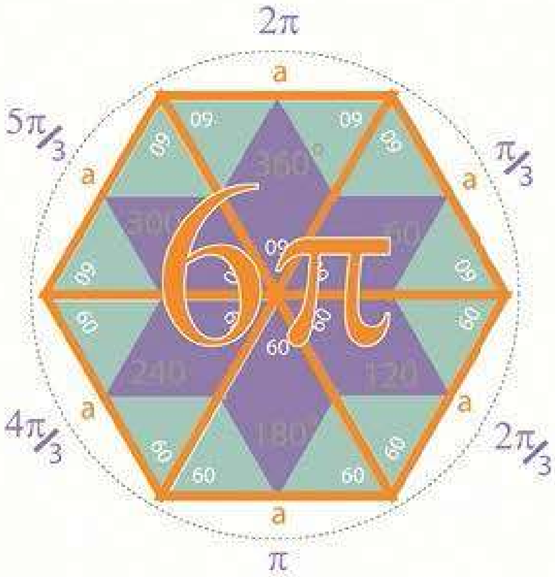
2π
a
5π/3
π/3
a
a
360°
300
60
6π
240
120
180
4π/3
2π/3
a
a
a
π

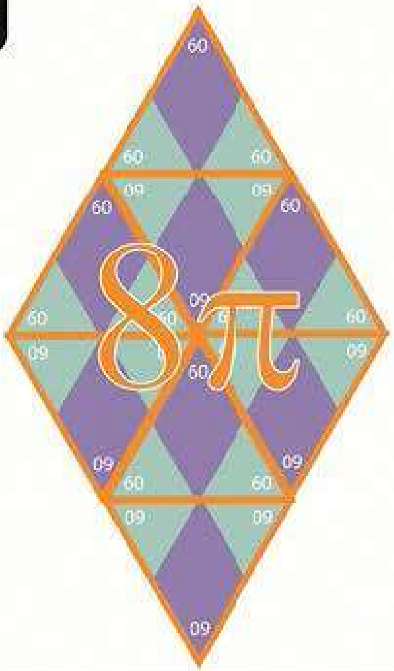

Tetryonic Cardinal Angles

equilateral mass-energy geometries form tetrahedral mass-Matter topologies

EM Charge

Equilateral energies

Q

E

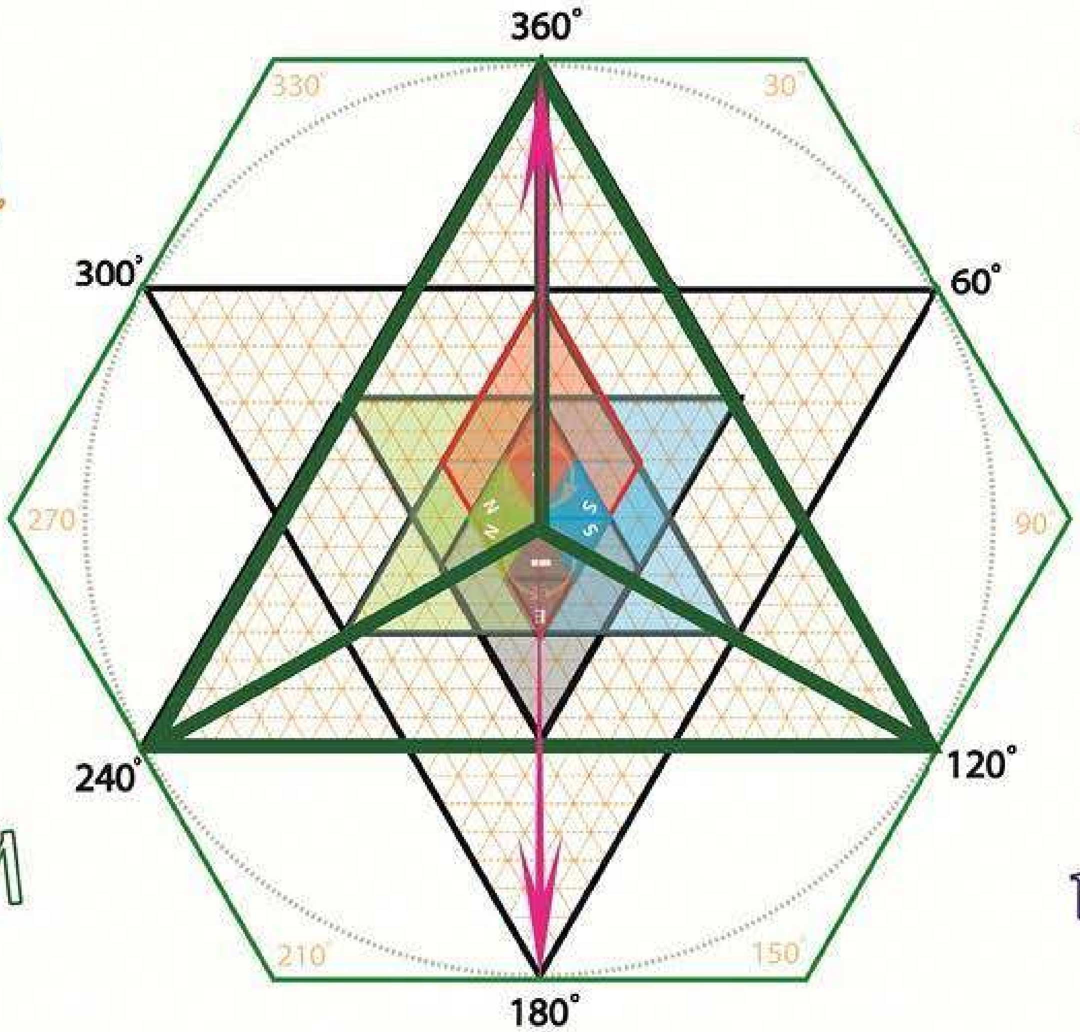
360°
330
30
300°
60°
270
90
N
N
S
S
240°
120°
210
150
180°

Space

Time

M

m

Matter topologies

mass geometries

radiant light

in radial spatial co-ordinate systems defined by the speed of light

Charge EM Field Planck quanta

$$\text{ODD}\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Bosons & quantum levels ElectroMagnetic mass velocity

Kinetic Energies EM Field Planck quanta

$$\text{EVEN}\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Photons & EM waves ElectroMagnetic mass velocity

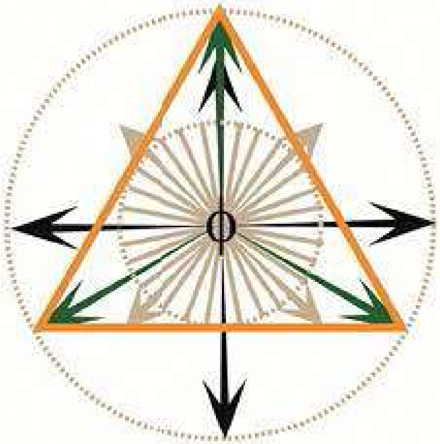

ALL Matter topologies stem from tessellated equilateral mass-energies

Matter EM Field Planck quanta

$$4n\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Fermions ElectroMagnetic mass velocity

Matter's 4π mass-Energies ares Lorentz invariant to accelerations

mass-energy geometries

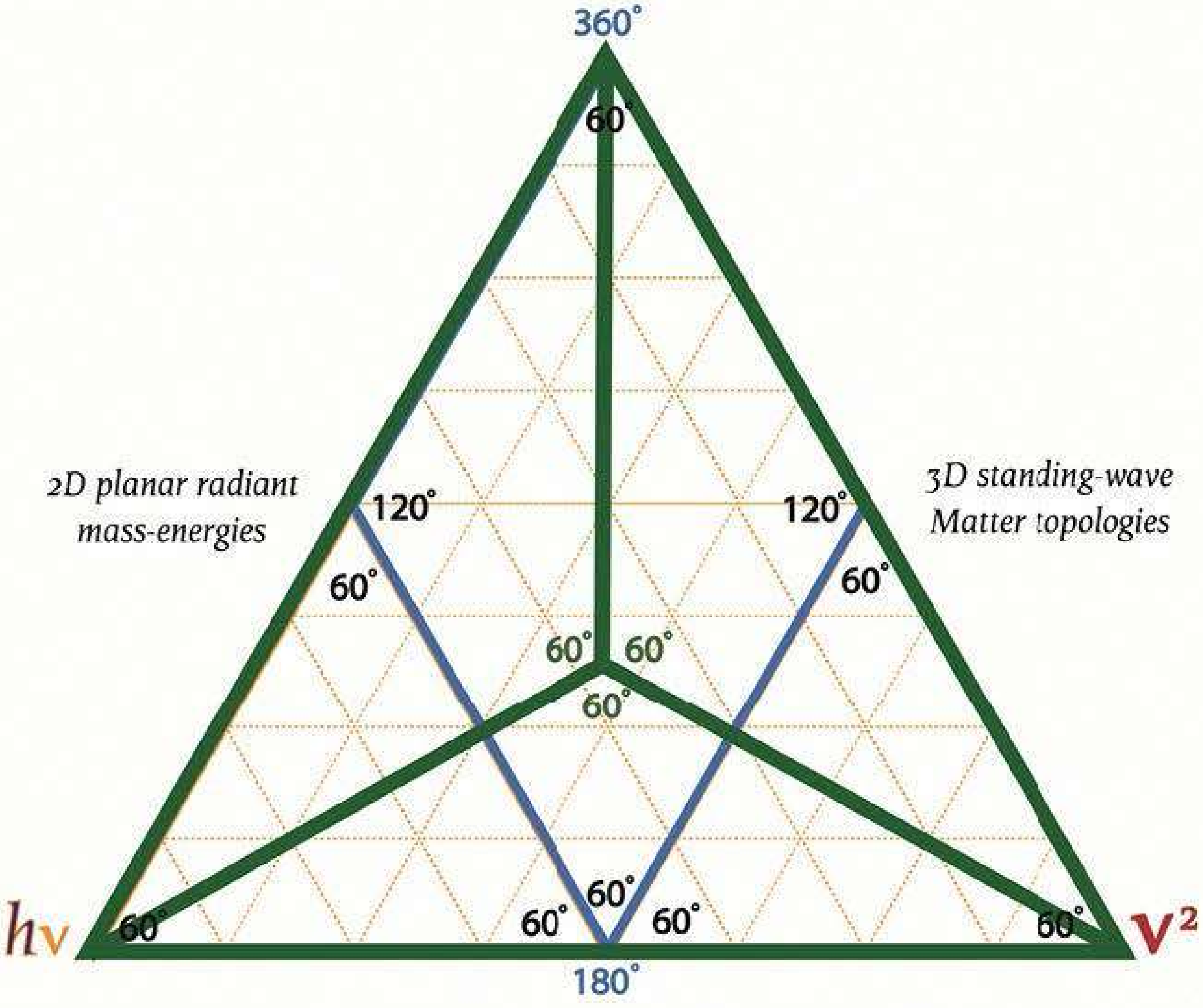
360°
60°
2D planar radiant mass-energies
120°
120°
3D standing-wave Matter topologies
60°
60°
60°
60°
60°
60°
60°
60°
60°
hν
60°
v²
180°

Matter topologies

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic geometry

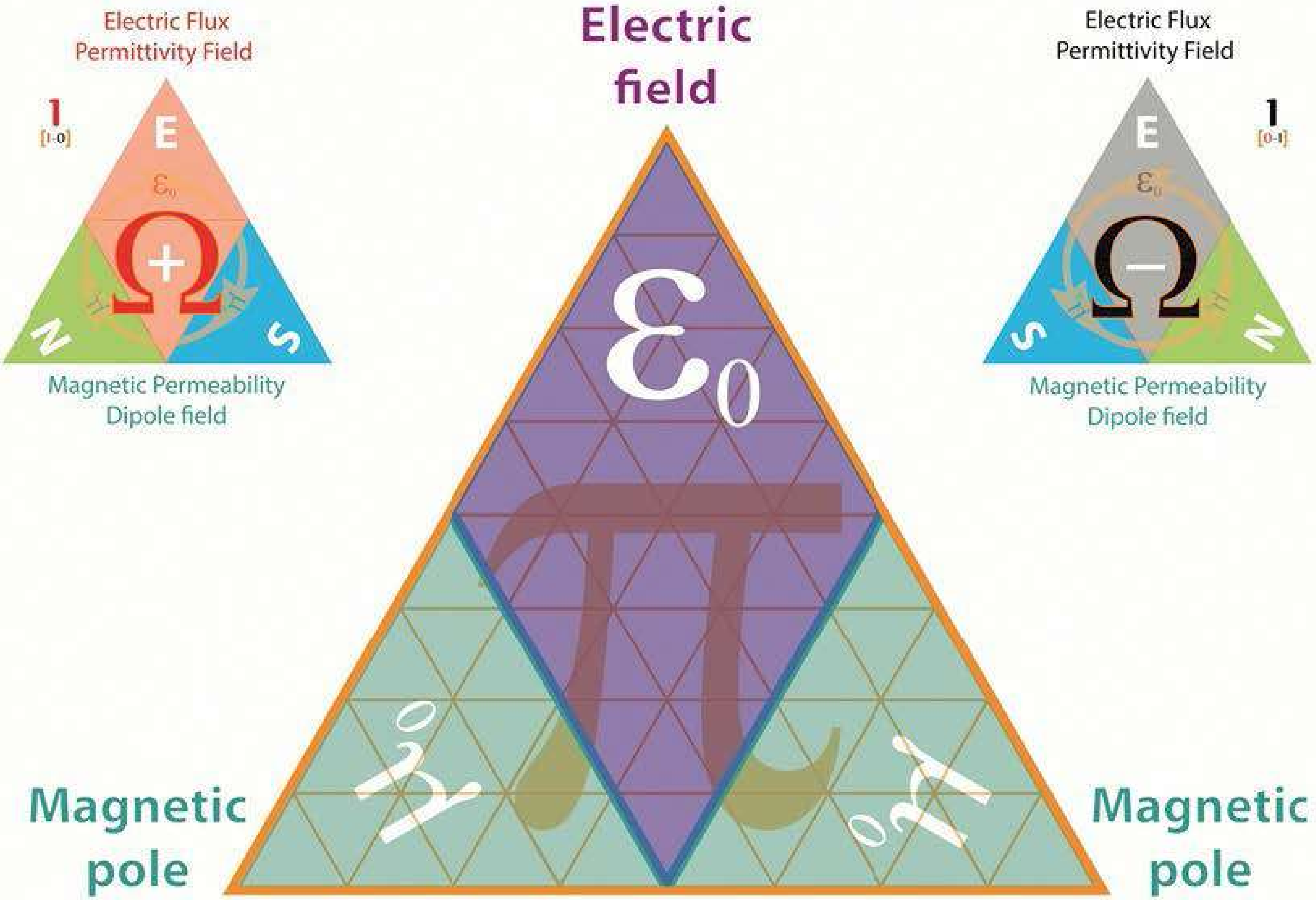
Electric Flux
Permittivity Field
1
[1-0]
E
ε_0
Ω
+
N
S
Magnetic Permeability
Dipole field
Electric
field
ε_0
π
μ_0
μ_0
Magnetic
pole
Magnetic
pole
Electric Flux
Permittivity Field
1
[0-1]
E
ε_0
Ω
−
S
N
Magnetic Permeability
Dipole field

equilateral quantised angular momenta
is the foundational geometry of all mass-ENERGY-Matter

The Golden Triangle

Planck's formulation for Energy is imprecise for use in Tetryonics
and does not reflect the velocity-momenta relationship
inherent in the equilateral geometry of Energy

Planck quantum levels

Energy is gained or lost in
whole number multiples of the quantity hv

$$E = n[hv]$$

ODD

Charge EM Field Planck quanta

$$ODD\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Bosons & quantum levels ElectroMagnetic mass velocity

$$3 = 3.[1]$$

$$9 = 9.[1]$$

$$25 = 25.[1]$$

$$hv$$

The generalise formulation of Plank's heat law
E = nhv is now changed to a specific formulation of
E = [ODD] hv for transverse quantum levels [Bosons]

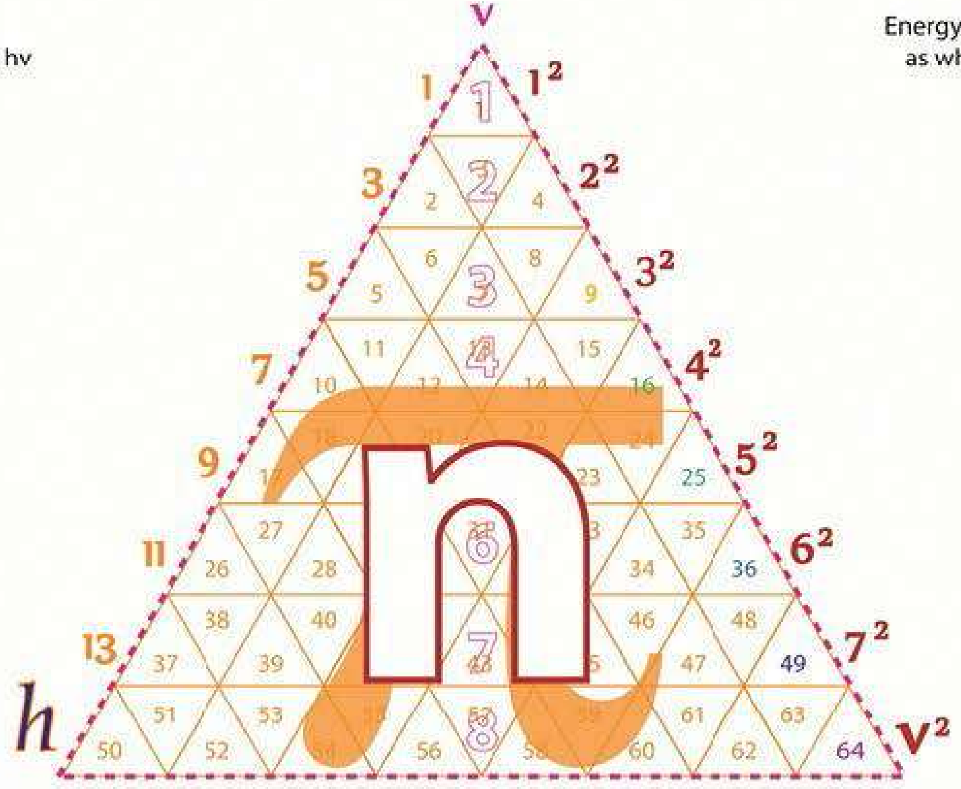

mass-energy momenta are geometrically related to velocity

Tetryonic geometry [nπ] redefines Planck's quantum formulation for heat energies from a generalised equation for 2D energy momenta into a geometric formulation for all mass-energy momenta in Matter

[all equilateral [π] geometries contain square number quanta]

Scalar energy quanta

Energy is gained or lost in equilateral geometries
as whole number multiples of the quantity hv²

$$E = n[hv]$$

SQUARED

Planck quanta

$$E = n\pi[[m\Omega v^2]]$$

mass velocity

$$3 = 3.[1^2]$$

$$9 = 1.[3^2]$$

$$25 = 1.[5^2]$$

$$hv^2$$

The general formulation of Plank's heat law
is also changed to a specific formulation of
E = hv² for scalar EM waveforms [ENERGY]

Scalar field geometries

All scalar fields are comprised of Transverse and Longitudinal mass-energy momenta all of which are formed from equilateral quantum geometries

Charge is a conserved force

Energy is a conserved quantity

Bosons are transverse ODD planck geometries

$$E = n\pi[[hv]]$$

Planck quanta

Energy is scalar SQUARE planck quanta

$$E = n\pi[[hv^2]]$$

Planck quanta squared

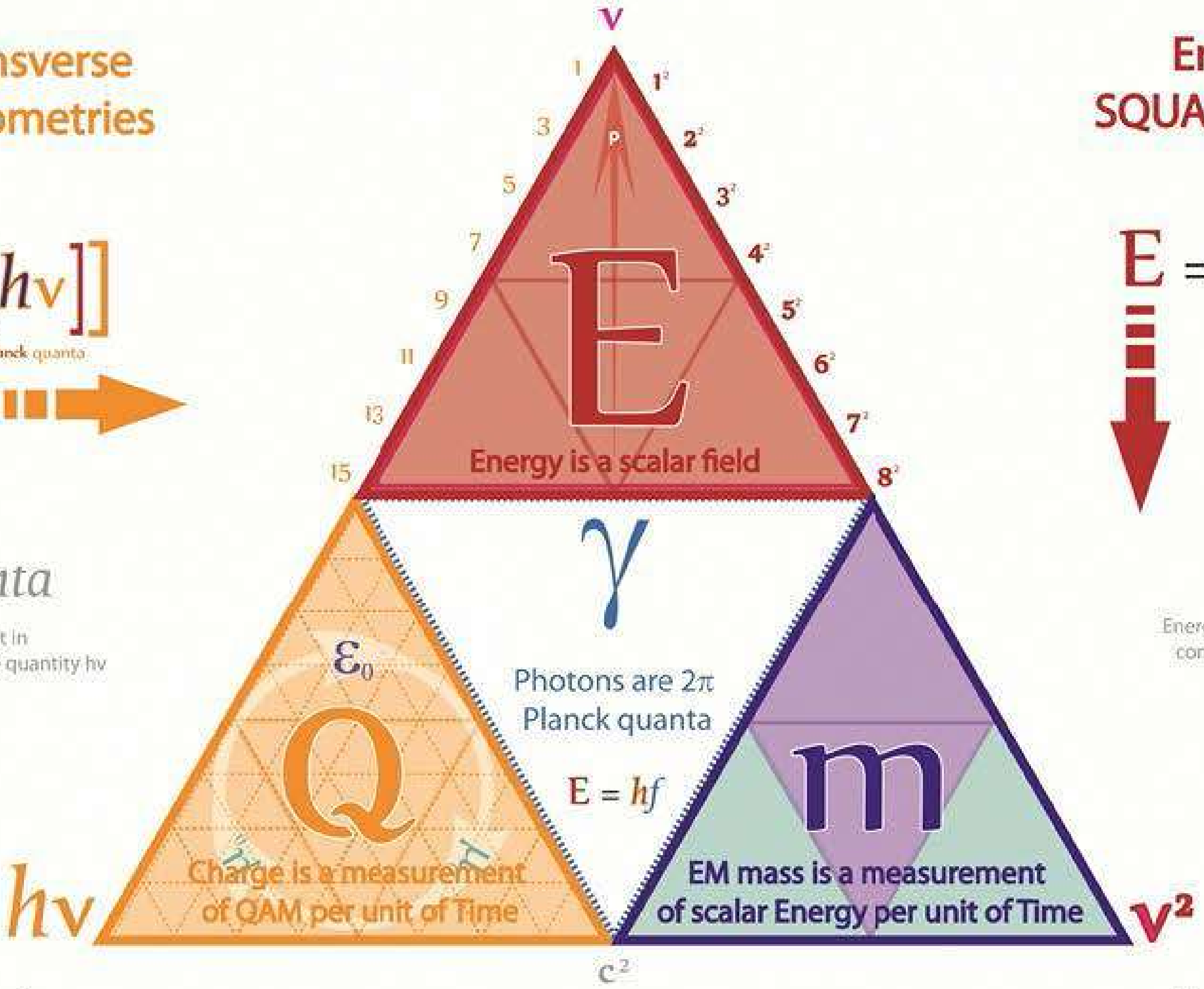
v
1
3
5
7
9
11
13
15
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
P
E
Energy is a scalar field
γ
Photons are 2π Planck quanta
E = hf
ε_0
Q
Charge is a measurement of QAM per unit of Time
m
EM mass is a measurement of scalar Energy per unit of Time
hv
v^2
c^2

Planck quanta

Energy is gained or lost in whole number multiples of the quantity hv

Scalar energies

Energy is gained or lost in equilateral quanta containing whole number multiples of hv^2

All energy is comprised of EM fields

All squared energy geometries have component quantum levels

Squared energy levels in quantum physics

are in fact equilateral Planck mass-energy geometries

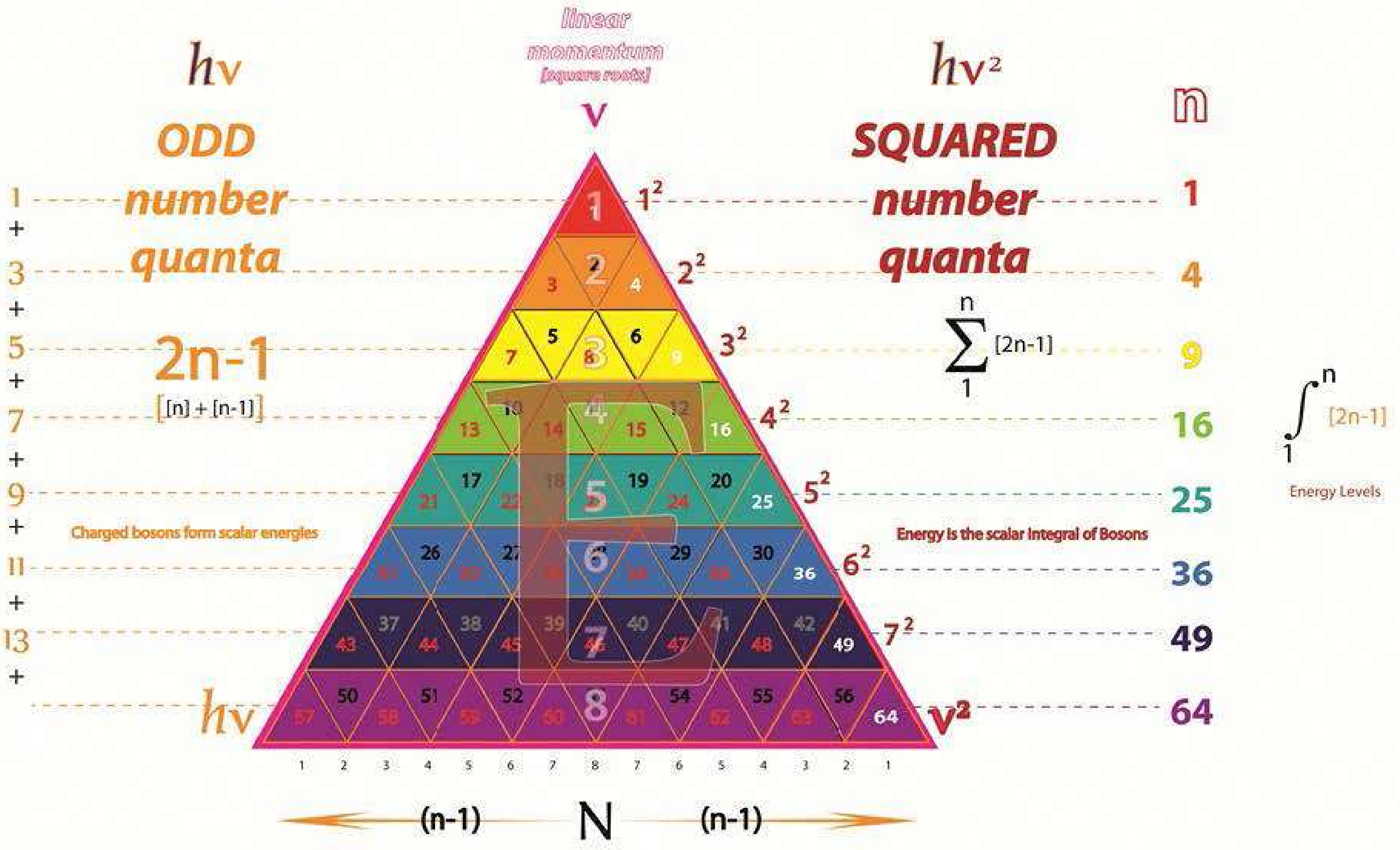
linear
momentum
[square roots]
v
hv
ODD
number
quanta
2n-1
[n] + [n-1]
Charged bosons form scalar energies
1 + 3 + 5 + 7 + 9 + 11 + 13 +
hv
hv²
SQUARED
number
quanta
n
$\sum_{1}^{n}$ [2n-1]
$\int_{1}^{n}$ [2n-1]
Energy Levels
Energy is the scalar integral of Bosons
1²
2²
3²
4²
5²
6²
7²
v²
1
4
9
16
25
36
49
64
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
(n-1)
N
(n-1)

Cumulative Distribution

NORMAL DISTRIBUTION

['Bell Curve']

Probability Distribution

Energy quanta defined

Quantum levels

Charge is gained or lost in
odd number multiples of the quantity hv

$$E = n\pi[[hv]]$$

Planck quanta

transverse

ODD quanta

Scalar Energies

Energy is gained or lost in equilateral geometries
containing whole number multiples of the quantity hv²

$$E = n\pi[[hv^2]]$$

Planck quanta squared

scalar

SQUARE quanta

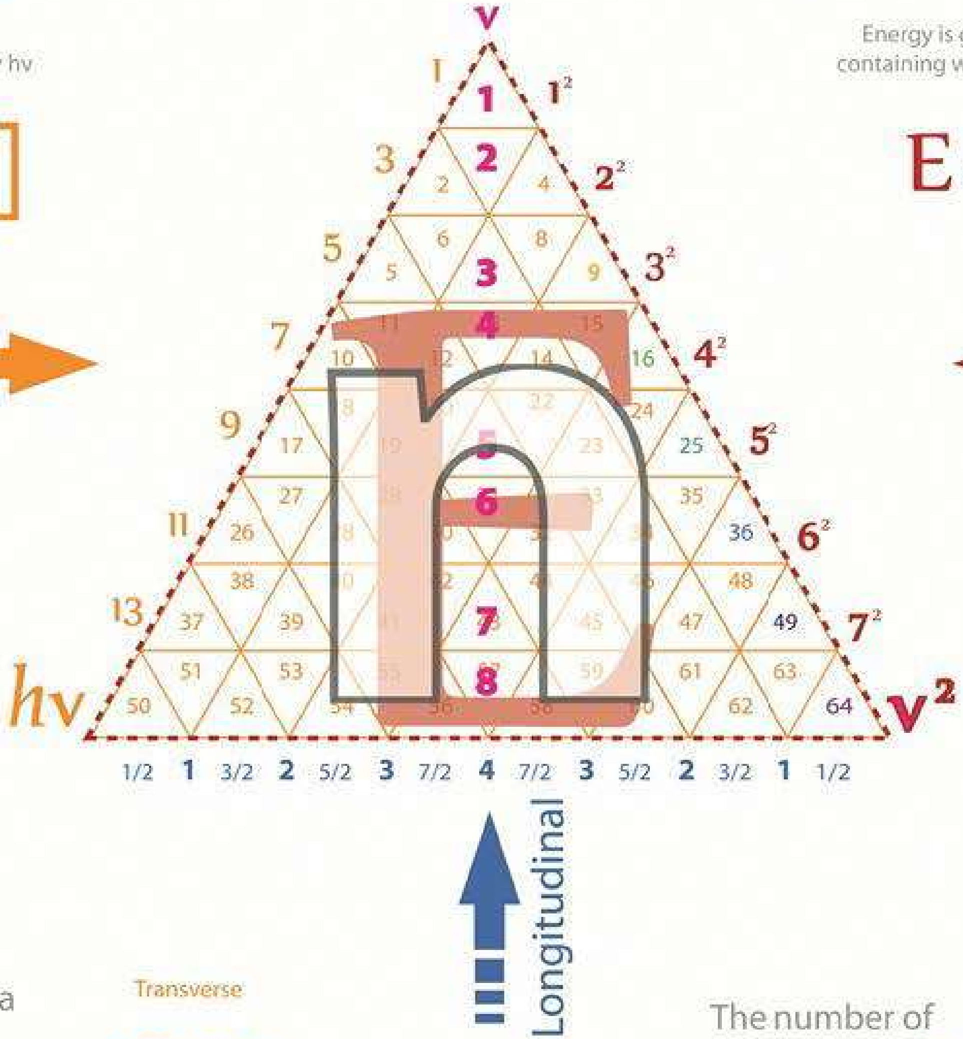
v
1
3
5
7
9
11
13
hv
1²
2²
3²
4²
5²
6²
7²
v²
1
2
3
4
5
6
7
8
1/2 1 3/2 2 5/2 3 7/2 4 7/2 3 5/2 2 3/2 1 1/2
Longitudinal

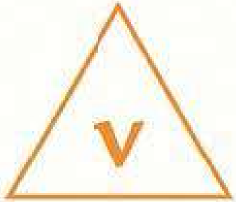

Quantum

The **number** of Planck quanta
in any physical system

$$v = f/2$$

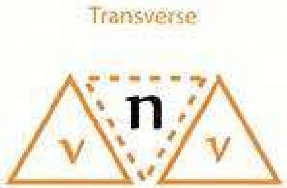
Transverse
n
v
v

Quantum Levels

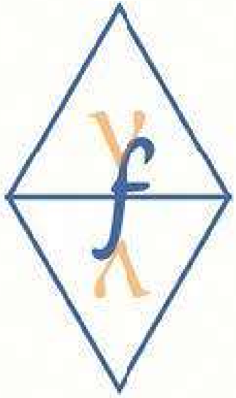

The number of
repeating waveform **cycles**
in a system

$$2v = f$$

Frequency

EVEN quanta

Tetryonic field equation

E

$$E = n\pi\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

electromagnetic energy is a scalar

$$E$$

m

$$\underset{\text{mass}}{n\pi}\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

mass is an inertial constant
that relates Force to acceleration

$$\frac{E}{c^2}$$

[v-v]

Q

$$\underset{\text{Bosons \& quantum levels}}{\overset{\text{Charge}}{ODD\pi}}\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

equilateral charge geometry
is Lorentz invariant

$$\frac{\Omega}{c^2}$$

KE

$$\underset{\text{Photons \& EM waves}}{\overset{\text{Kinetic Energies}}{EVEN\pi}}\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

KEM fields mass-energy geometries
are Lorentz variant to accelerations

$$\frac{1}{2}m\left[\frac{v^2}{c^2}\right]$$

M

$$\underset{\text{Fermions}}{\overset{\text{Matter}}{4n\pi}}\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

Matter's 4π mass-energy topologies
are Lorentz invariant to accelerations

$$4n\pi\left[\frac{E}{c^4}\right]$$

Energy momenta geometry

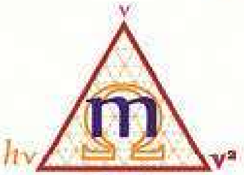

Energy is the total Planck quanta per second

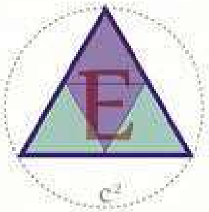

EM mass is scalar energy per unit of Time

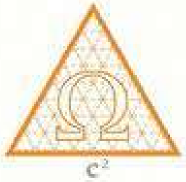

Charge is the nett Angular momentum per spatial co-ordinate system

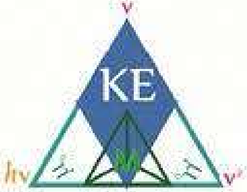

Kinetic Energy is the Electric field
mass-energy of Matter moving at v

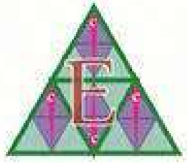

mass-energy geometry

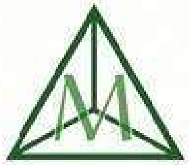

Matter topology

mass-Energy equivalence

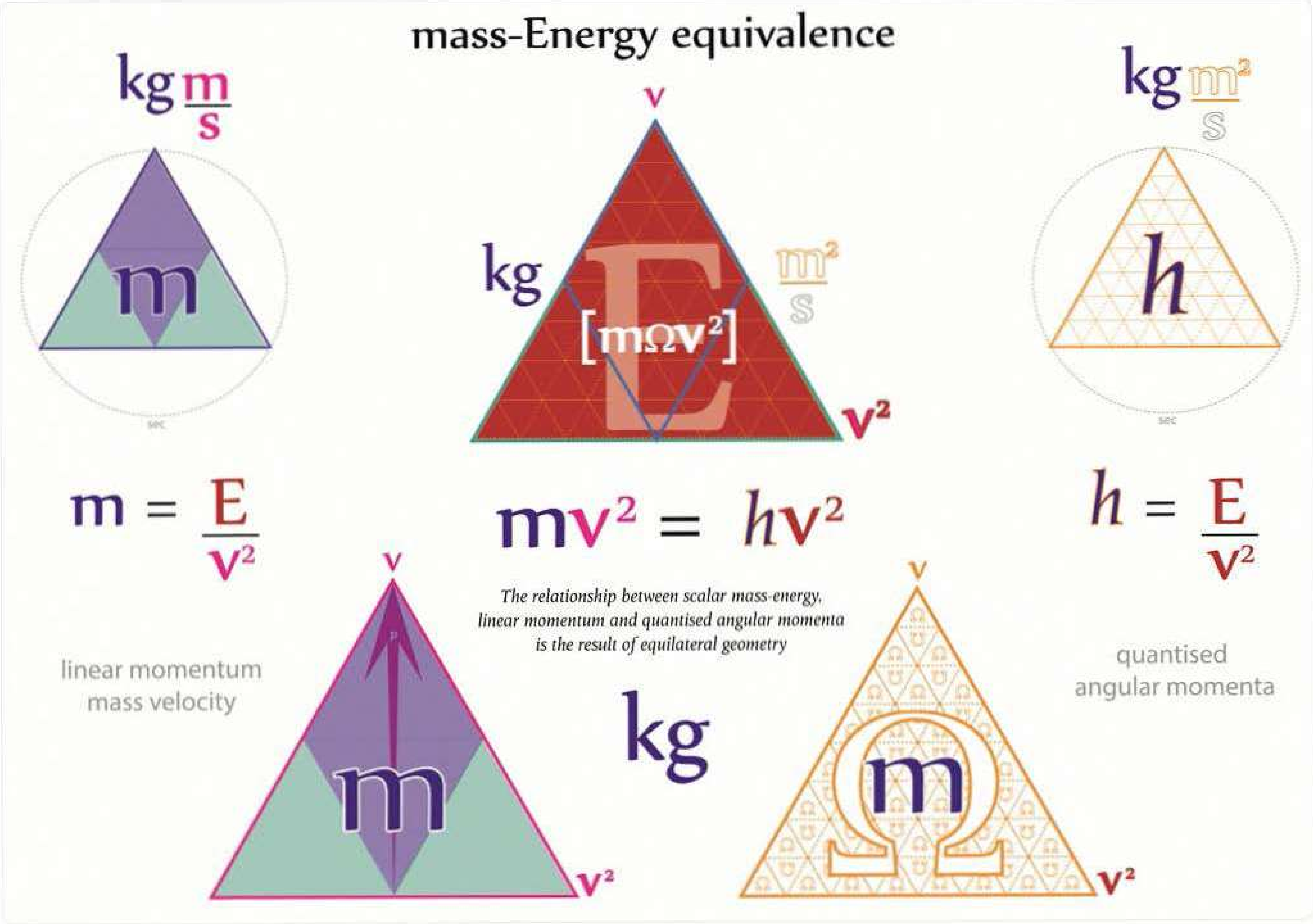
kg m/s
m
sec
v
kg
m²/s
E
[mΩv²]
v²
kg m²/s
h
sec
v
p
m
v²
kg
v
Ω
m
v²

$$m = \frac{E}{v^2}$$

$$mv^2 = hv^2$$

$$h = \frac{E}{v^2}$$

The relationship between scalar mass-energy, linear momentum and quantised angular momenta is the result of equilateral geometry

linear momentum
mass velocity

quantised
angular momenta

Tetryonics 02.13 - mass-Energy equivalance

mass-Energy geometries

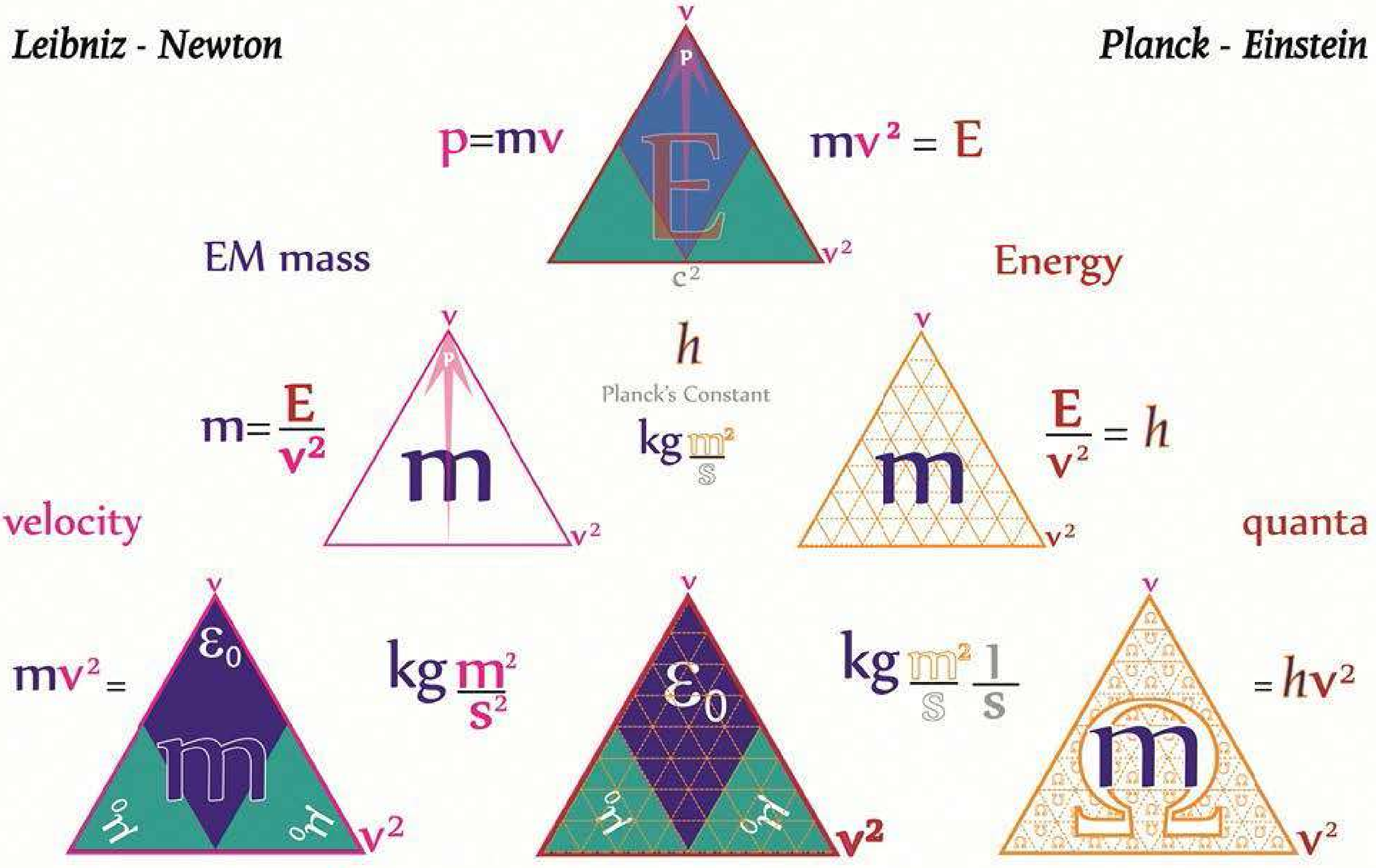
v
P
E
v²
c²
v
P
m
v²
v
m
v²
v
ε0
m
μ0
μ0
v²
v
ε0
μ0
μ0
v²
v
Ω
m
v²

Leibniz - Newton

Planck - Einstein

$p=mv$

$mv^2 = E$

EM mass

Energy

$m=\frac{E}{v^2}$

h

Planck's Constant

$kg\frac{m^2}{s}$

$\frac{E}{v^2} = h$

velocity

quanta

$mv^2 =$

$kg\frac{m^2}{s^2}$

$kg\frac{m^2}{s}\frac{1}{s}$

$= hv^2$

velocity squared

quanta squared

There exists an intrinsic geometric relationship between mass-velocity and the quantisation of Energy momenta

EM mass-energy

$$E = mc^2$$

EM mass is related to Matter throught the square of the velocity of light of 8.987551787 e16 m^2/s^2 [c^2]

EM mass-Matter

$$E = Mc^4$$

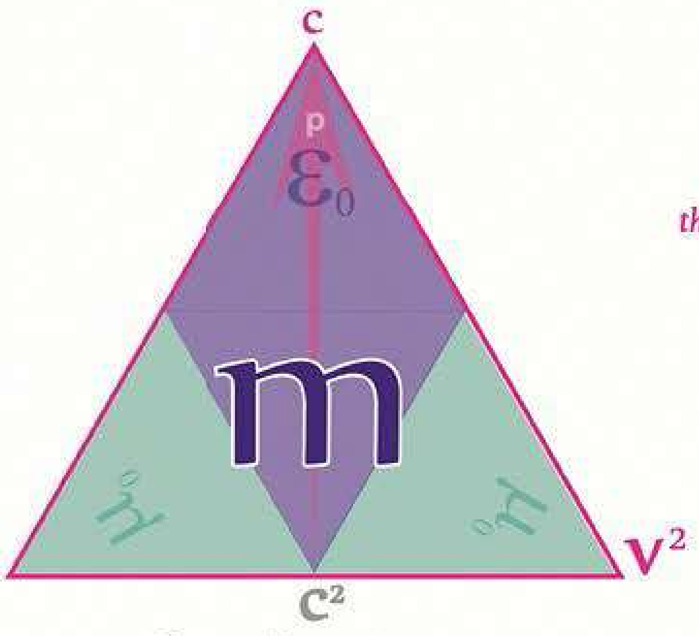
c
p
ε_0
m
μ_0
μ_0
v^2
c^2

Electro-Magnetic mass

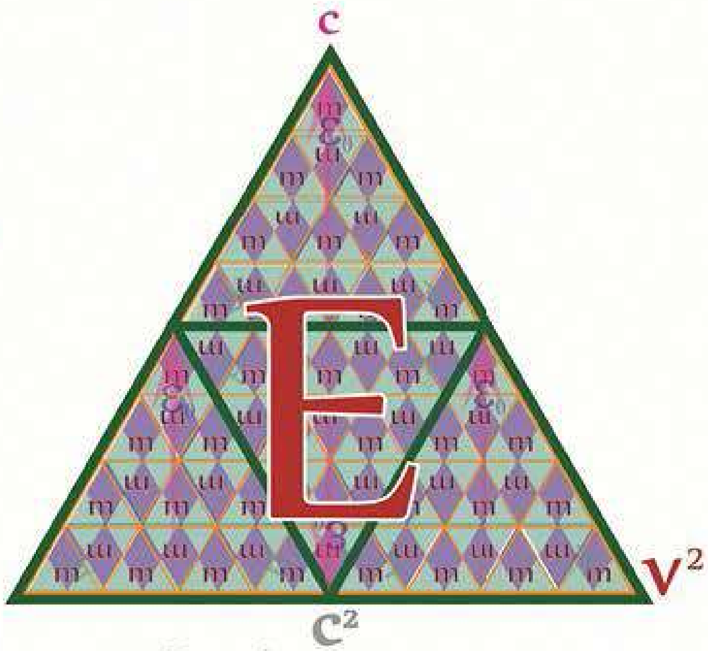
c
ε_0
E
m
v^2
c^2

Matter's energy quanta

scalar EM mass

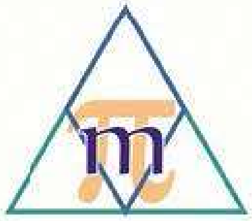

$$M = \frac{E}{c^4}$$

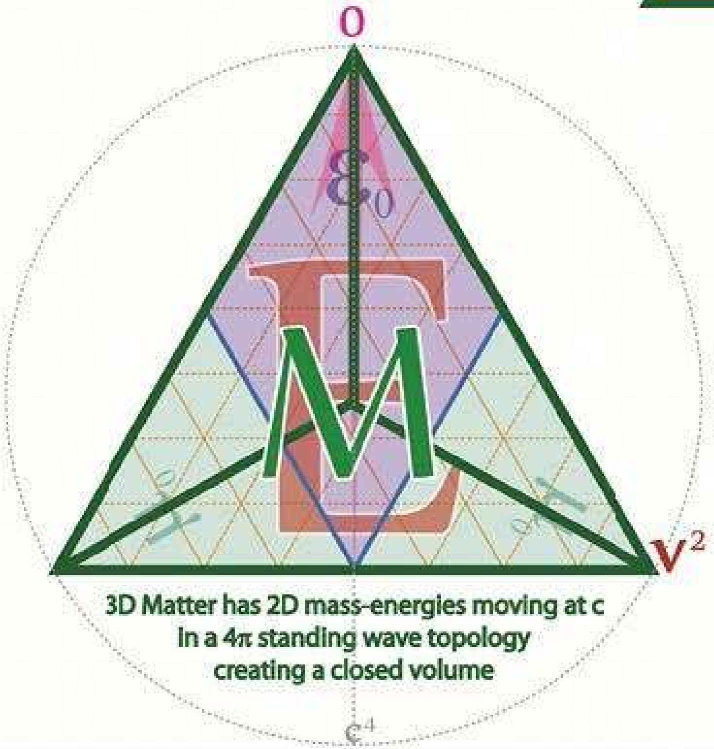
0
ε_0
E
M
μ_0
μ_0
v^2
c^4

3D Matter has 2D mass-energies moving at c in a 4π standing wave topology creating a closed volume

quantised mass

$$h = \frac{E}{v^2}$$

Kinetic Electro-Magnetic fields

[The energies of Motion]

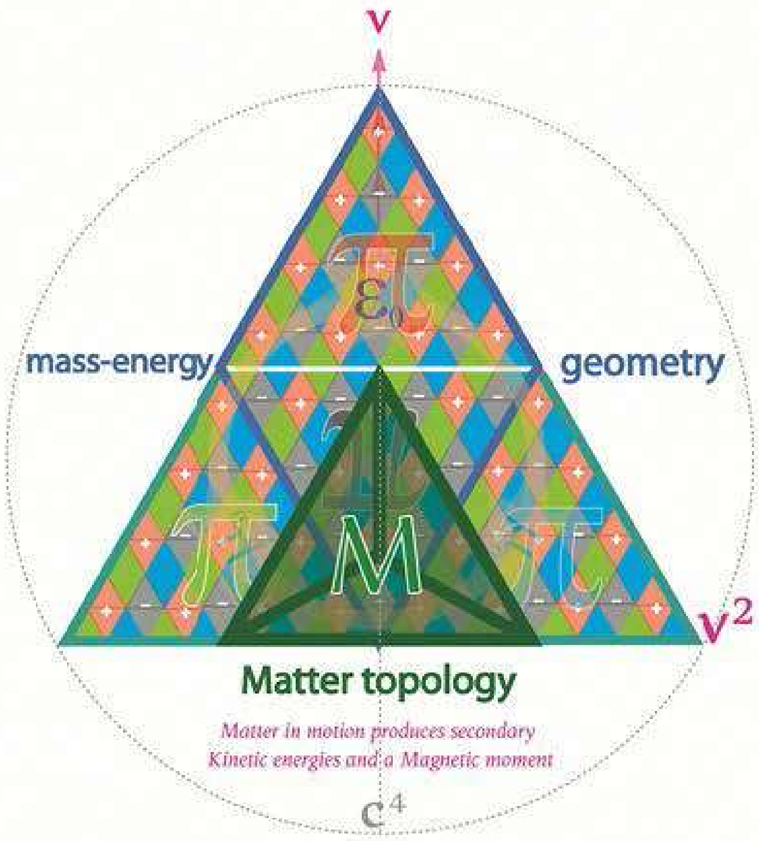
v
mass-energy
geometry
ε_0
M
v^2
Matter topology
Matter in motion produces secondary
Kinetic energies and a Magnetic moment
c^4

$$p^2 = E = Mv^2$$

The Electric field energy in any EM field is equal and orthogonal to the Magnetic field energy

$$p^2 = KEM = Mv^2$$

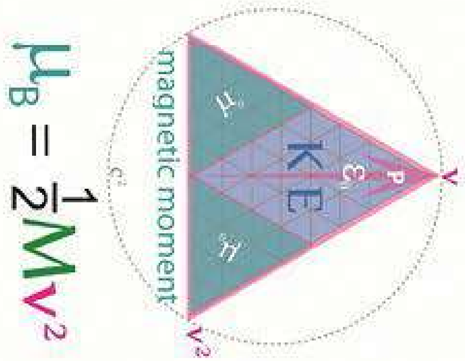
magnetic moment
μ_0
ε_0
KE
p
v
v^2
c

$$\mu_B = \frac{1}{2}Mv^2$$

$$KE = \frac{1}{2}Mv^2$$

All Matter in motion possess momenta and kinetic mass-energies in extrinsic KEM fields

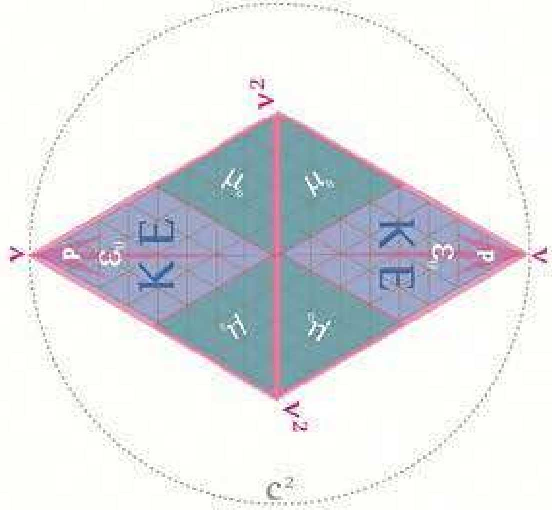
v^2
μ_0
ε_0
KE
p
v
v^2
c^2

$$E_\gamma = \frac{2mv^2}{c^2}$$

These 2D planar fields are subject to Lorentz factor corrections that apply to the KEM field mass-energies of Matter in motion

Charge & Kinetic EM fields

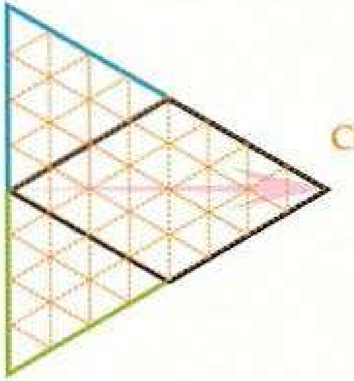

Charges are the result
of quantised
angular momenta

Kinetic Energies
result from
Matter in motion

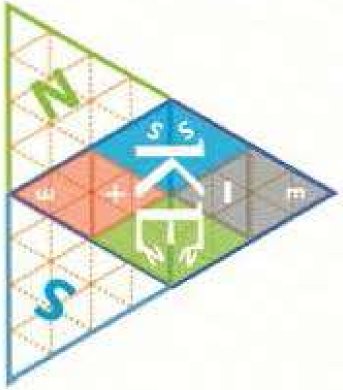

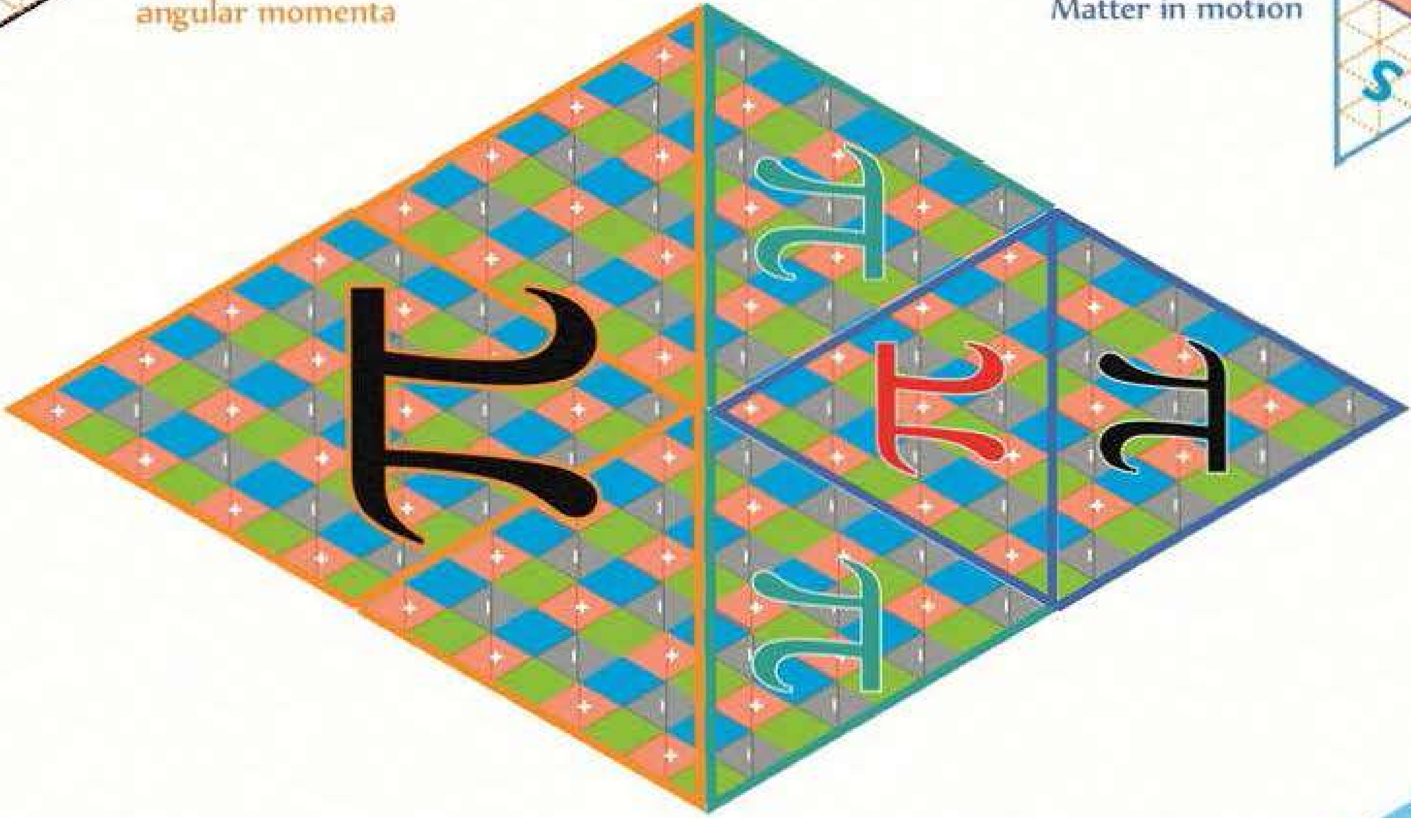

$$mv^2 = E = hv^2$$

$$p^2 = E = m\Omega v^2$$

$$KEM = Mv^2$$

$$KE = \frac{1}{2}Mv^2$$

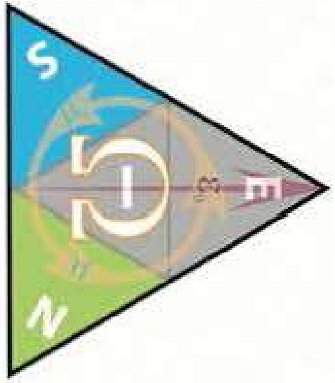

EM field energies externalised on the fascia of 3D Matter geometries
form the physical property known as elementary Charge

[k]EM energies in amy region of free space are viewed as either
a Kinetic energy field with an associated Magnetic moment
or
a neutral charge Photon of ElectroMagnetic mass-energy

Copyright ABRAHAM [2008] - all rights reserved

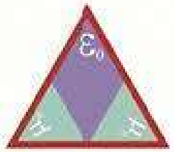

$$m_0 = \frac{E}{c^2}$$

Rest mass is equivalent to the total quantity of Energy In a body or system (divided by c²)

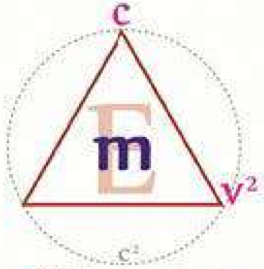

2D fields propagate at c
3D Matter moves at v

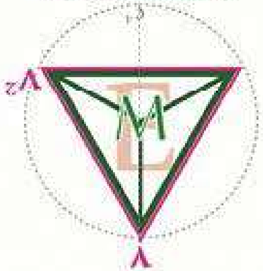

inertial mass

$$m = \frac{F}{a}$$

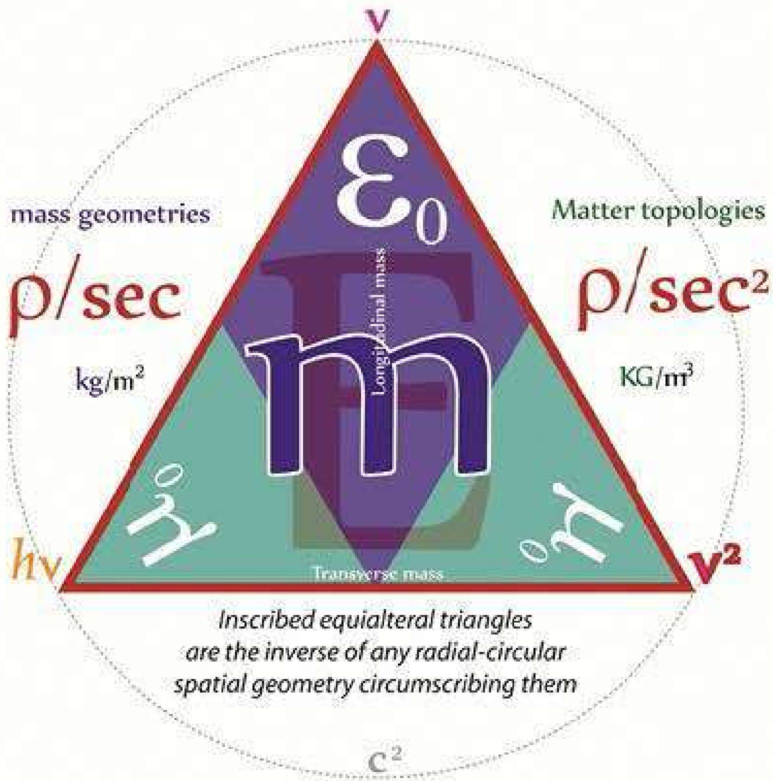

Inscribed equialteral triangles are the inverse of any radial-circular spatial geometry circumscribing them

mass must be distinguished from Matter in physics, because Matter is a poorly-defined concept in modern science, and although all types of agreed-upon Matter exhibit properties of mass, it is also the case that there many types of Energy that posses NO Matter topology, such as potential energy, kinetic energies and electromagnetic radiation (photons)

EM Field Planck quanta

$$\underset{\text{mass}}{n\pi}\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

ElectroMagnetic mass velocity

The term 'massless' must be re-termed 'Matterless' to reflect true physical attributes of mass-energy-momenta

EM mass is a planar measurement of 2D energy per unit of time

Thus, all 3D Matter topologies have charged fascia comprised of 2D scalar of mass-energies, but closed volume 3D topology is not a property of 2D EM mass-energy geometries

mass

ρ

EM mass is a measure of planar energy density per second

$$\frac{mv^2}{c^2} \quad \boxed{\frac{m\Omega v^2}{c^2}} \quad \frac{hv^2}{c^2}$$

per unit of time

EM mass geometries are subject to Lorentz corrections

velocity QAM quanta

$$\frac{mv^2}{c^2} = \frac{m\Omega v^2}{c^2}$$

per unit of time

Gravitational Matter

$$g = G\frac{M}{r^2}$$

Copyright ABRAHAM [2008] - all rights reserved

Electromagnetic mass-energy

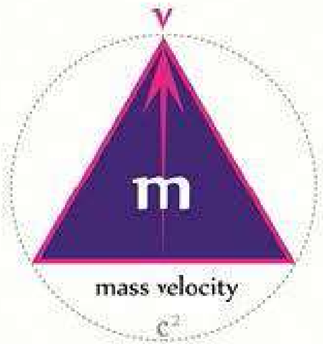

In physics, EM mass–energy equivalence is the concept that the EM mass of a body is a measure of its energy content

Using this concept, EM mass is a property of all Energy, and Energy is a property of all EM mass, and the two properties are connected by a constant.

Using Tetryonic geometry it can be shown that the constant is the equilateral geometry of QAM thus unifying Classical mechanics and Relativisic mechanics

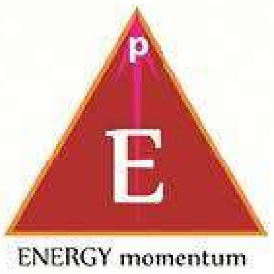

2D mass-ENERGY geometry is NOT **3D Matter topology**

$$m = \frac{E}{v^2}$$

Classical mass

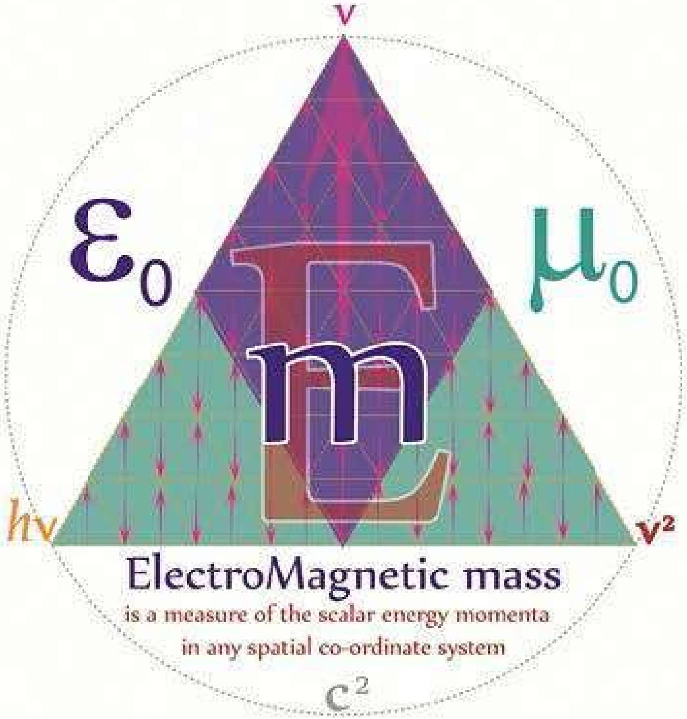

ElectroMagnetic mass

is a measure of the scalar energy momenta in any spatial co-ordinate system

$$m = \frac{E}{c^2}$$

Relativistic mass

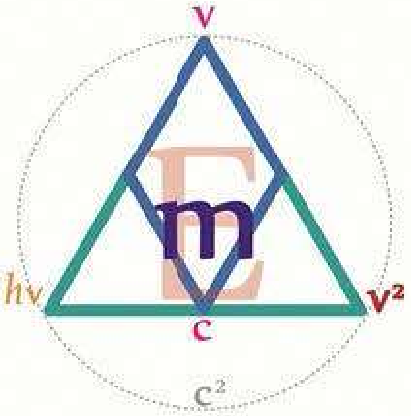

radiant 2D mass-energies
are planar equilateral
energy geometries

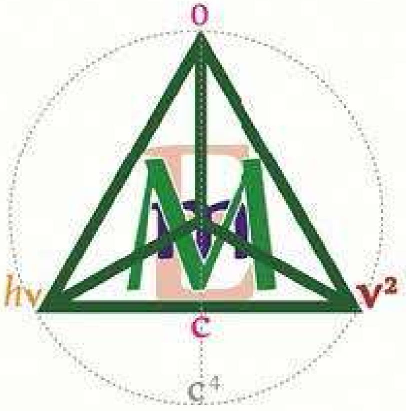

standing-wave 3d mass-Matter
are tetrahedral energy
momenta topologies

Relativity shows that rest mass and rest energy are essentially equivalent, via the well-known relationship ($E=mc^2$)

EM mass-Energy-Matter

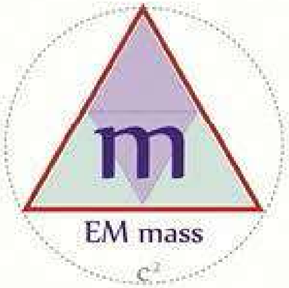

$$\left[\frac{\Omega}{c^{2}}\right]$$

Charge is a measure of the quantised angular momentum of any physical system

Tetryonic Matter

3D topology

planar spatial Impedance

Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^{2}\right]\right]$$

ElectroMagnetic 2D mass velocity

Planck quanta

$$\left[m\Omega v\right]^{2}$$

mass velocity

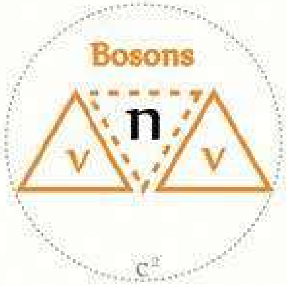
Bosons
n
v
v
c²

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^{2}\right]\right]$$

ODD quanta

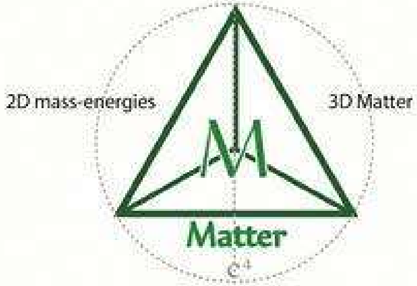
2D mass-energies
3D Matter
M
Matter
c⁴

$$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^{2}\right]\right]$$

TETRYONS

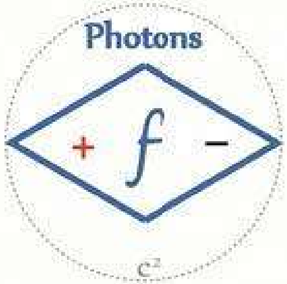
Photons
+
f
−
c²

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^{2}\right]\right]$$

EVEN quanta

Copyright ABRAHAM [2008] - all rights reserved

EM mass-Energy momenta

Planck's Constant

is the 'quantum of Action'

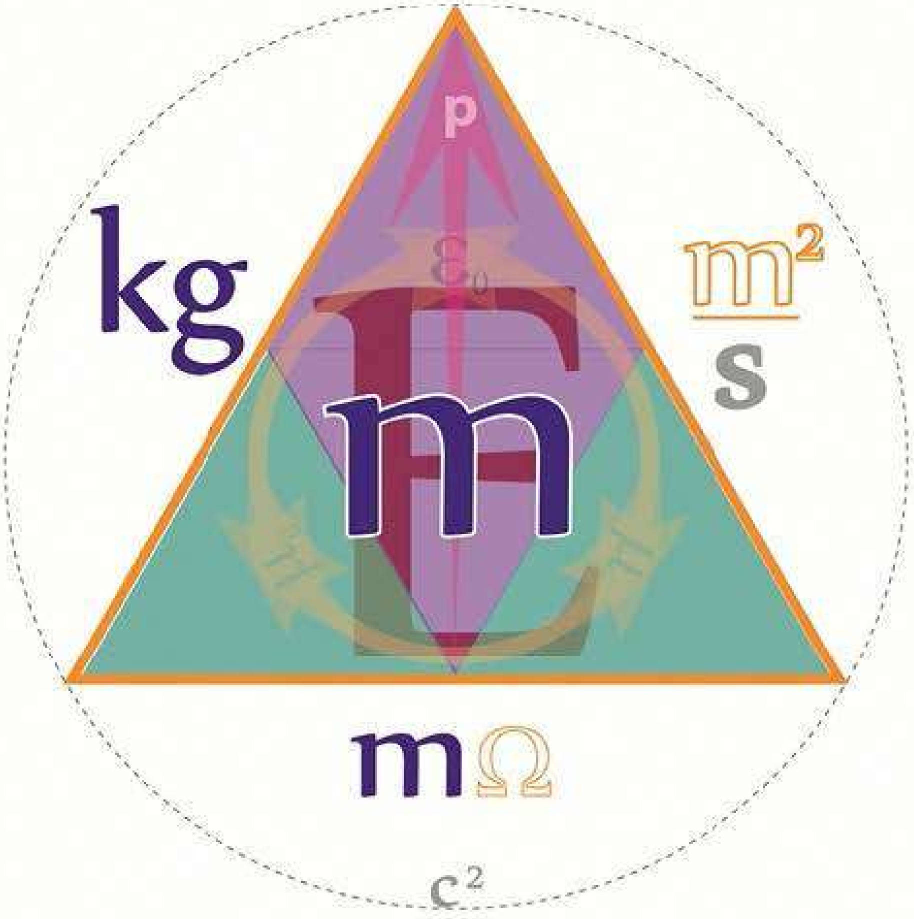

$$6.629432672 \times 10^{-34}\ J.s$$

mass momenta

Energy

$$\left[\frac{m\Omega v^2}{s}\right]$$

E

6.629432672 e-34 J

mass-Energy

7.376238634 e-51 kg

m

$$\left[\frac{E}{c^2}\right]$$

mass

quantised angular momenta

$$\left[\frac{m^2}{s}\right]$$

Ω

0.0012 m2/s

Charge-momenta

1.33518 e-20 s

s

$$\left[\frac{\Omega}{c^2}\right]$$

charge

Tetryonics 02.21 - EM mass-Energy momenta

Copyright ABRAHAM [2008] - all rights reserved

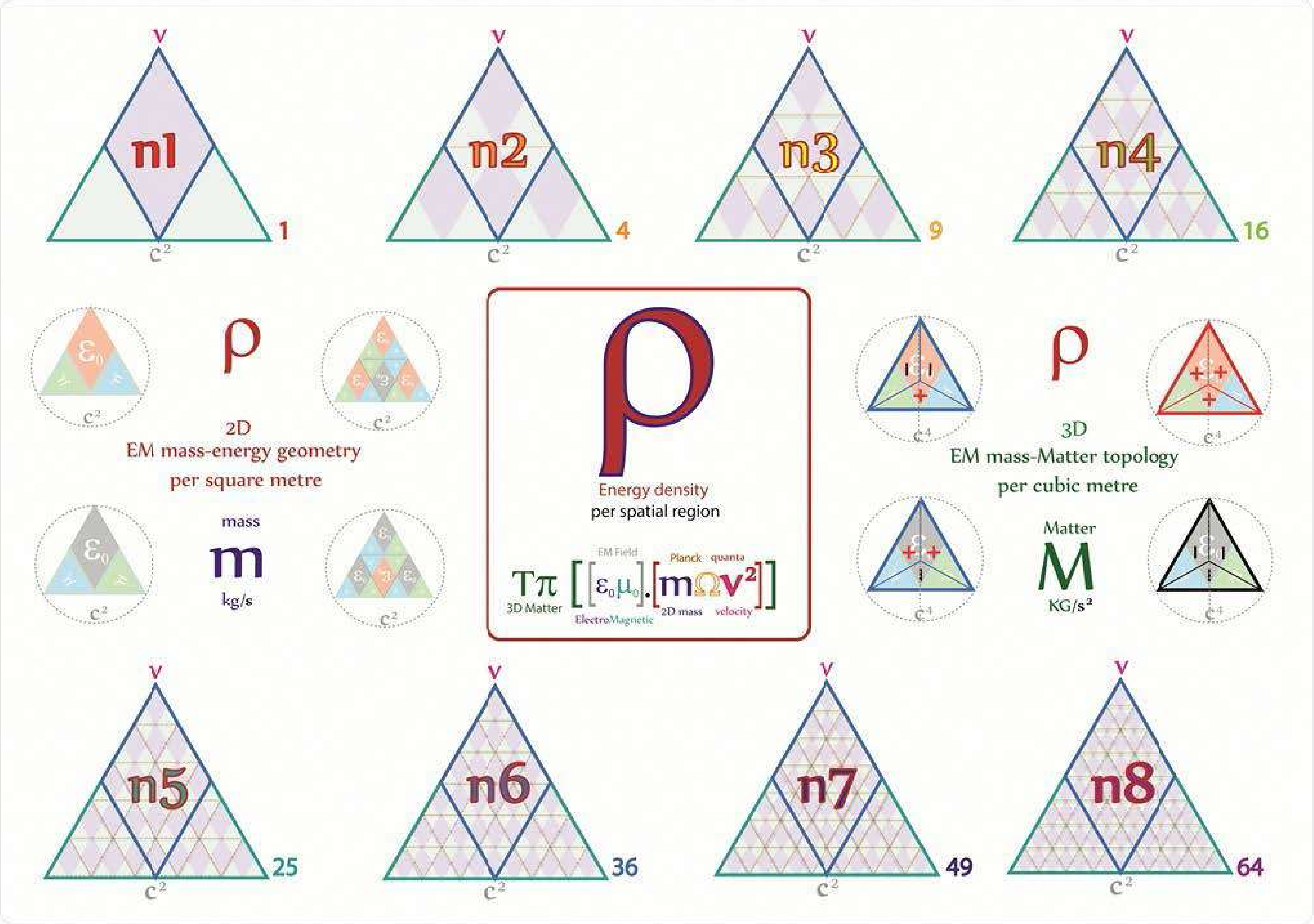
v
n1
1
c²
v
n2
4
c²
v
n3
9
c²
v
n4
16
c²
ρ
2D
EM mass-energy geometry
per square metre
mass
m
kg/s
ε0
c²
ρ
Energy density
per spatial region
Tπ [[ε0μ0].[mΩv²]]
EM Field
Planck
quanta
3D Matter
ElectroMagnetic
2D mass
velocity
ρ
3D
EM mass-Matter topology
per cubic metre
Matter
M
KG/s²
c⁴
v
n5
25
c²
v
n6
36
c²
v
n7
49
c²
v
n8
64
c²

Zero Point Fields [ZPFs]

IDEAL QUANTUM INDUCTORS
(equilateral triangle Energy geometry)

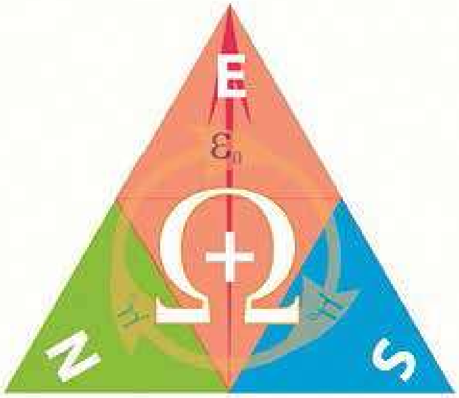

ZPFs are quantum inductive tank circuits
(Short-circuited 'IDEAL' inductors with energy)
ZPFs charge energies do NOT oscillate
[The magnetic dipole vector determines charge]

The EM FIELD

Electric flux fields can propagate in any direction
Magnetic fields are always at 90 degrees to Electric fields

Magnetic dipole fields propagate in 2 directions at
180 degrees to each other (bi-directionally)
forming North and South poles

Electric flux field energy is directly proportional to the resultant Magnetic dipole field energy and vice versa

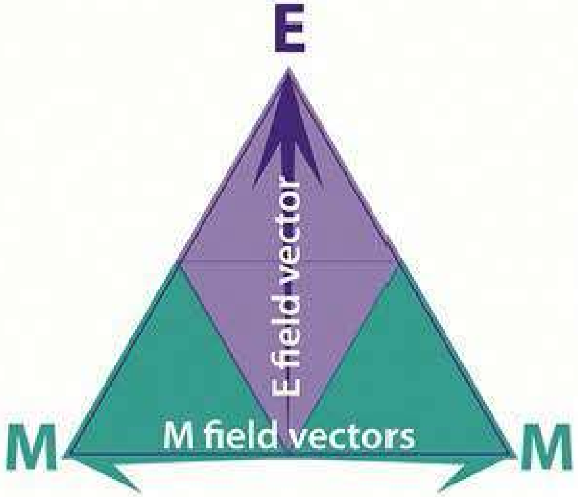

The linear Electric field strength is directly
proportional to its associated transverse
Magnetic field which propagates bi-directionally
from & into the bloch wall of the Zero point field

Zero Point Fields consist of
Electric and Magnetic (EM) fields
propagating at 90 degrees
to each other

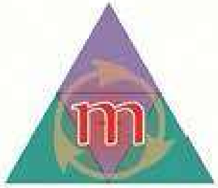

h

Positive Charge ZPF

Nett positive Planck quanta
with North-South m-dipole vector

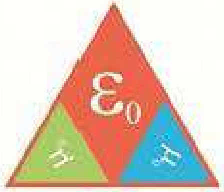

Magnetic MONOPLES do NOT exist

Energy quanta always form charged Electric fields and dipole Magnetic fields

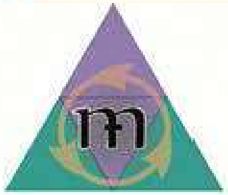

h

Negative Charge ZPF

Nett negative Planck quanta
with South-North m-dipole vector

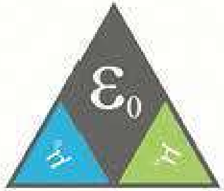

As localised energy quanta increases
(number of ZPFs per time unit)
the charge geometry remains the same

$mv^2 = E$

Quantised Angular Momentum

is the equilateral geometry of scalar energy momenta per unit of time

$E = hv^2$

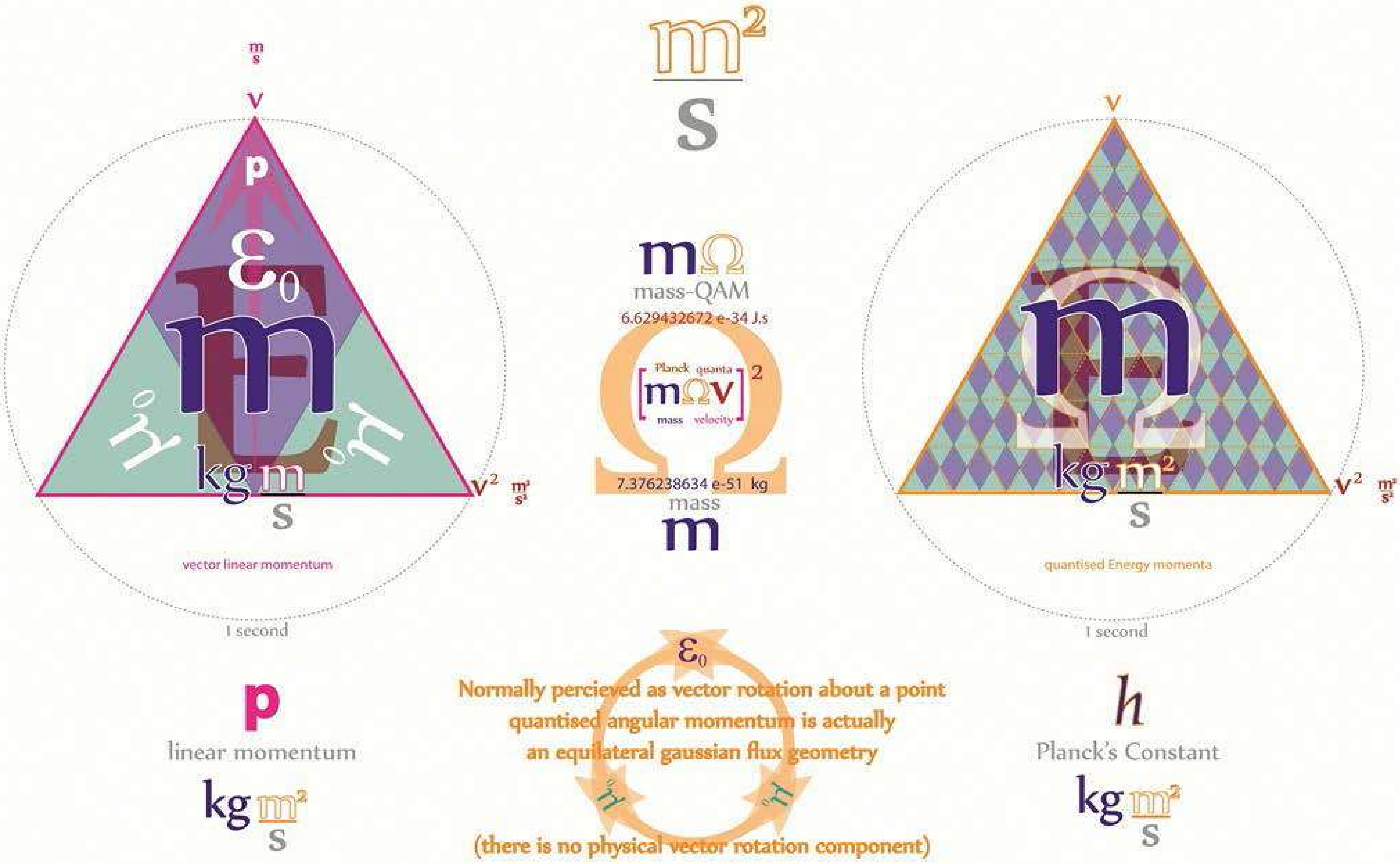

$\frac{m^2}{s}$
$\frac{m}{s}$
v
p
ε_0
E
m
μ_0
kg $\frac{m}{s}$
v^2 $\frac{m^2}{s^2}$
vector linear momentum
1 second
mΩ
mass-QAM
6.629432672 e-34 J.s
Planck quanta
$[m\Omega v]^2$
mass velocity
7.376238634 e-51 kg
mass
m
v
m
E
Ω
kg $\frac{m^2}{s}$
v^2 $\frac{m^2}{s^2}$
quantised Energy momenta
1 second
p
linear momentum
kg $\frac{m^2}{s}$
ε_0
μ_0
μ_0
h
Planck's Constant
kg $\frac{m^2}{s}$

Normally percieved as vector rotation about a point quantised angular momentum is actually an equilateral gaussian flux geometry

(there is no physical vector rotation component)

Charged mass & Matter in motion

produce differing but related measurments of EM force

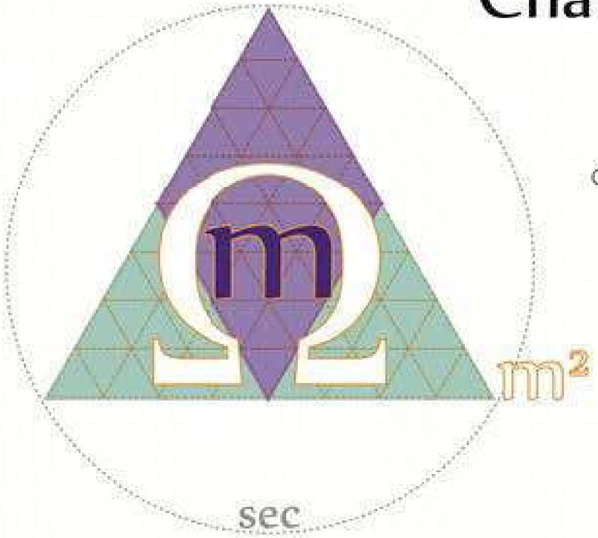

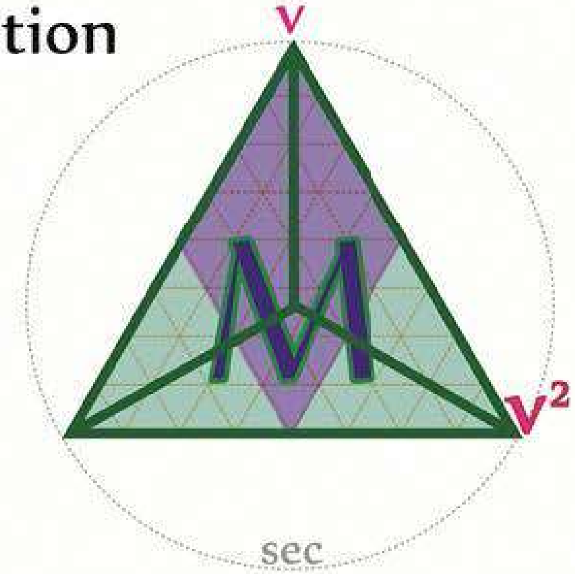

Charged masses per second $\frac{C}{s} \approx kg \approx A$ Charged Matter in motion

Amperes

By definition in Tetryonics the quantised angular momentum of mass in any defined spatial co-ordinate system provides the quantum basis for

$$\boxed{kg\,\frac{m^2}{s}}$$

charged mass-energies

Quantised angular momenta creates CHARGE

Vector linear momentum creates CURRENT

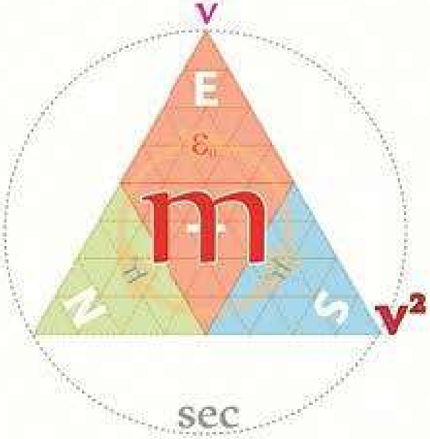

charged mass

$$C \approx kg.s \approx As$$

Coulombs Amp.sec

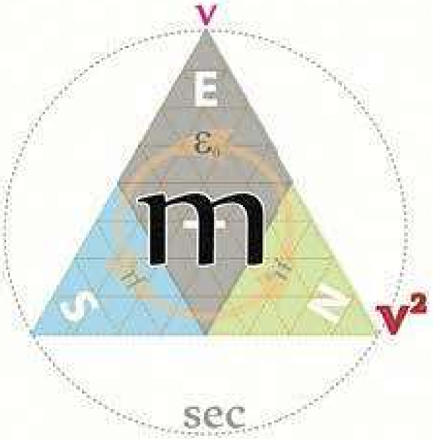

Charge, through its energy interactions, creates the geometric scaffolding for ALL 3D EM mass-Energy-Matter geometries

Charged mass geometry

The symmetry of Charge geometries provides a geometric foundation for all mass-Energy-Matter particles and physical forces

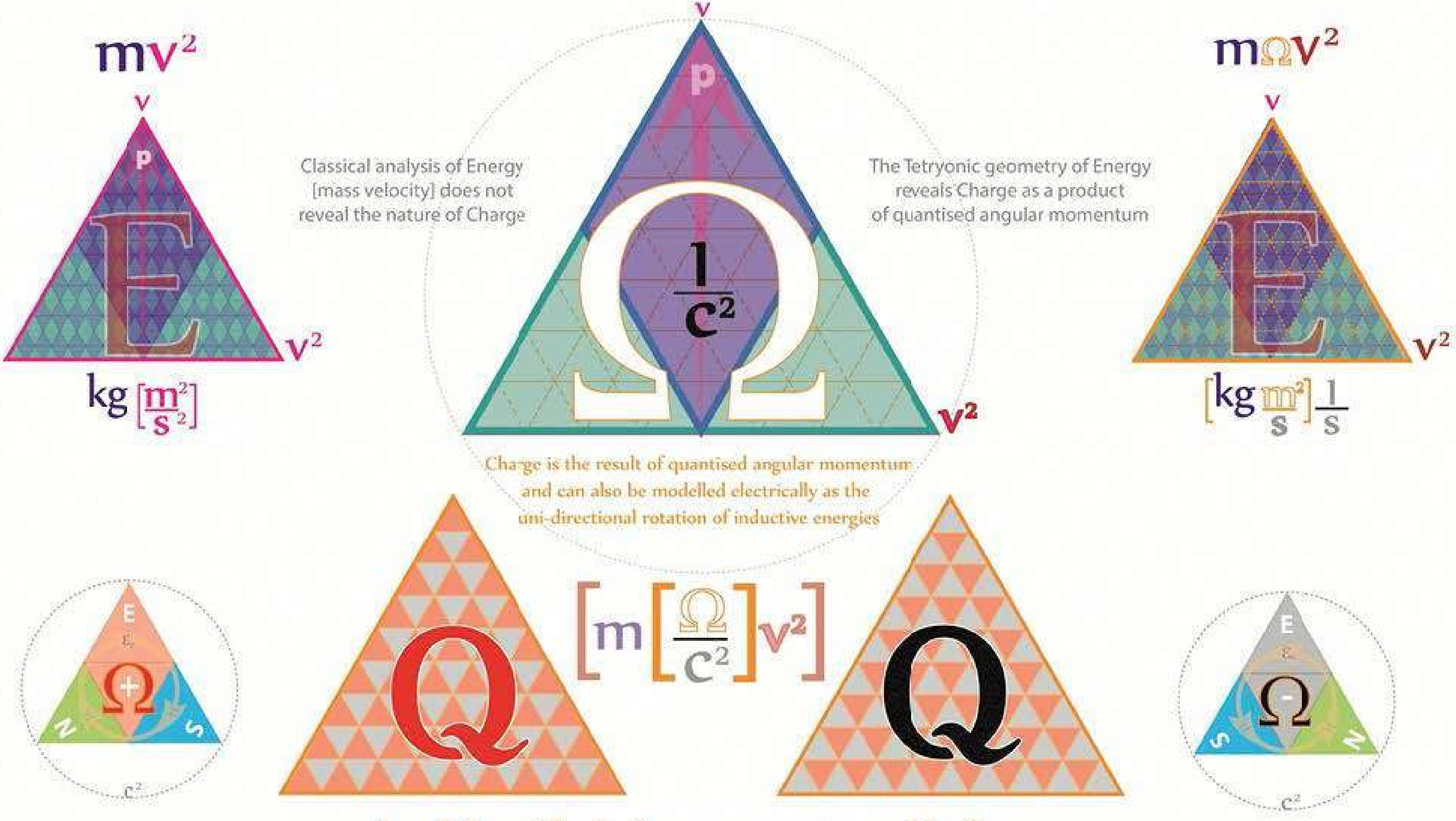

mv^2

$kg\left[\frac{m^2}{s^2}\right]$

Classical analysis of Energy [mass velocity] does not reveal the nature of Charge

$\frac{1}{c^2}$

The Tetryonic geometry of Energy reveals Charge as a product of quantised angular momentum

$m\Omega v^2$

$\left[kg\,\frac{m^2}{s}\right]\frac{1}{s}$

Charge is the result of quantised angular momentum and can also be modelled electrically as the uni-directional rotation of inductive energies

$\left[m\left[\frac{\Omega}{c^2}\right]v^2\right]$

'a un-balanced (broken) energy symmetry resulting from the nett geometry of equilateral EM mass-energy quanta'

Copyright ABRAHAM [2008] - all rights reserved

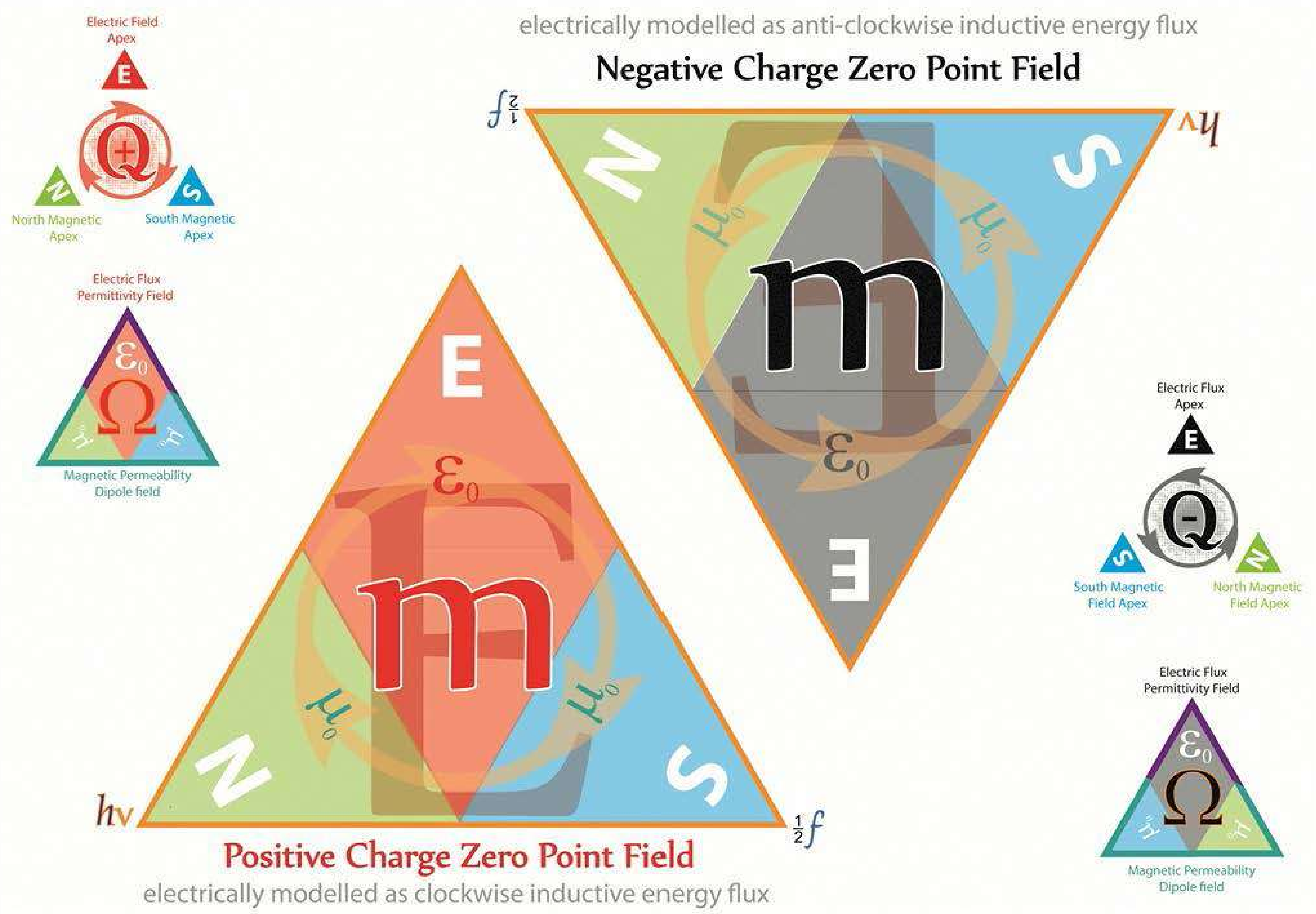

electrically modelled as anti-clockwise inductive energy flux

Negative Charge Zero Point Field

Positive Charge Zero Point Field

electrically modelled as clockwise inductive energy flux

Electric Field Apex
E
Q
+
N
S
North Magnetic Apex
South Magnetic Apex
Electric Flux Permittivity Field
ε_0
Ω
μ_0
μ_0
Magnetic Permeability Dipole field
$\frac{1}{2}f$
hv
N
S
m
μ_0
μ_0
ε_0
E
Electric Flux Apex
E
Q
-
S
N
South Magnetic Field Apex
North Magnetic Field Apex
E
ε_0
m
μ_0
μ_0
N
S
hv
$\frac{1}{2}f$
Electric Flux Permittivity Field
ε_0
Ω
μ_0
μ_0
Magnetic Permeability Dipole field

Elementary and quantum Charges

"All known Fermions have charge topologies that are integer 1/3 multiples of the elementary charge"

ε_0
Ω

It can be shown that the "quantum of charge" is the intrinsic angular momentum of a Planck mass

It then follows that Quasi-particles [Tetryons and Quarks] have charge topologies with nett charges of 0, 4 and 8

Further, it can be shown that the "elementary charge" applied to Leptons and Baryons (12 times quantum charge) is reflective of their nett charged Matter topologies

topology *charge*

$T\pi[\pm q]$

geometry

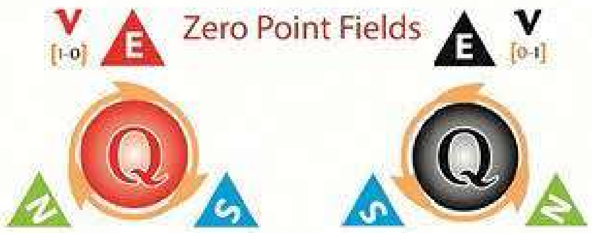

All Charged particles and their respective KEM fields can be modelled with ZPF field geometries reflective of their nett Charge topologies

Opposites attract
Similars repel

Quarks

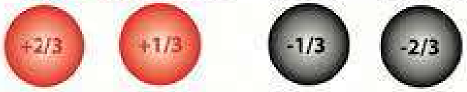

up	anti-down	down	anti-up
Charmed Top	Strange Bottom	Strange Bottom	Charmed Top

Leptons

Positron	neutrino	Muon
Tau anti-electron	Tau neutrino	Tau

Baryons

Proton	**Neutron**	**Negatron**

quantised Charges
(EM energy flux rotation within a ZPF geometry)

q
quantum charges

[v] [v]

The elementary charge, usually denoted as e, is the electric charge carried by a single proton, or equivalently, the absolute value of the electric charge carried by a single electron

[v-v]

Q
Elementary charges

Elementary Charges
are the nett charge created by mass-energy-Matter topologies

Tetryons

4	0	4
[4-0]	[2-2]	[0-4]

Quarks

up	down	down	up
8	4	4	8
[10-2]	[8-4]	[4-8]	[2-10]

12	0	0	12
[12-0]	[6-6]	[6-6]	[0-12]
e	ve	ve	e

Leptons

Baryons

Proton	Neutron	Negatron
12	0	12
[24-12]	[18-18]	[12-24]

All Charges seek Equilibrium

Sir Issac Newton

(25 December 1642 – 20 March 1727)

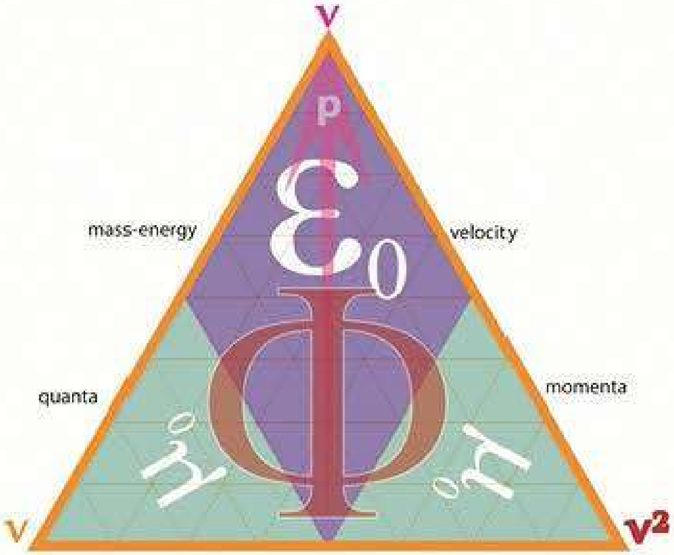

Any changes to velocity
result in changes to Energy-momenta
within a charged geometry

$$\sum \mathbf{F} = 0 \Rightarrow \frac{d\mathbf{v}}{dt} = 0$$

Newton's first law of motion says:
"A body maintains the current state of motion unless acted upon by an external force."

$$F = ma$$

Inertia is the resistance of any physical object to a change in its state of motion or rest, or the tendency of an object to resist any change in its motion.

It is proportional to an object's mass.

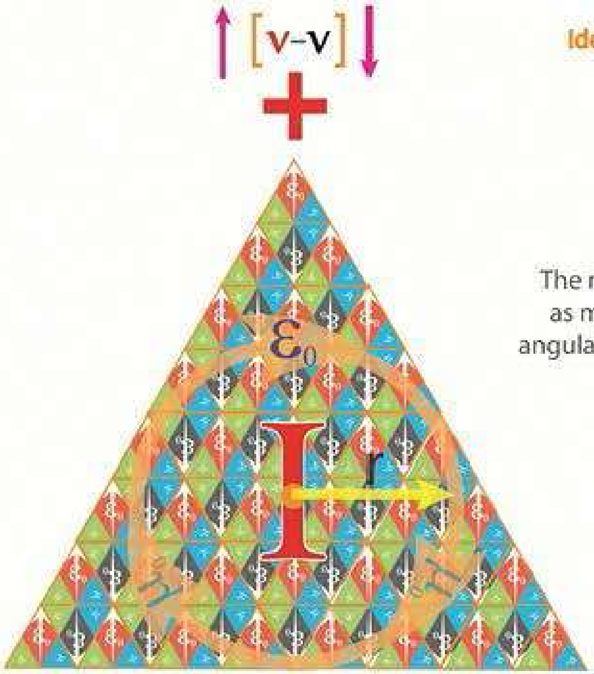

Ideal Inductive loops (ZPF quanta) will oppose ANY changes to their energy levels and consequently, Inertia can be viewed as an outcome of quantum ZPF self-inductance

Inertia

The moment of inertia plays much the same role in rotational dynamics as mass does in linear dynamics, describing the relationship between angular momentum and angular velocity, torque and angular acceleration, and several other quantities

The difference between the impeding (inertial) linear momenta and the co-linear (supportive) momenta is a result of the nett Planck mass-energy momenta within the charged geometries of Matter & its KEM fields

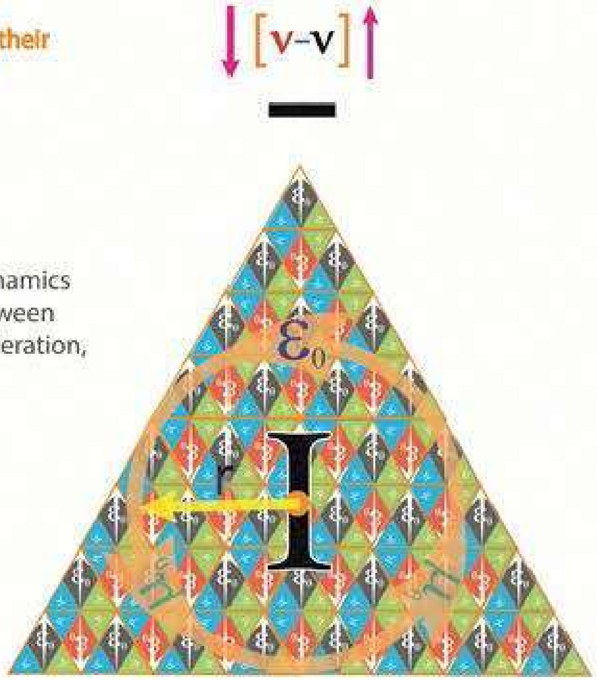

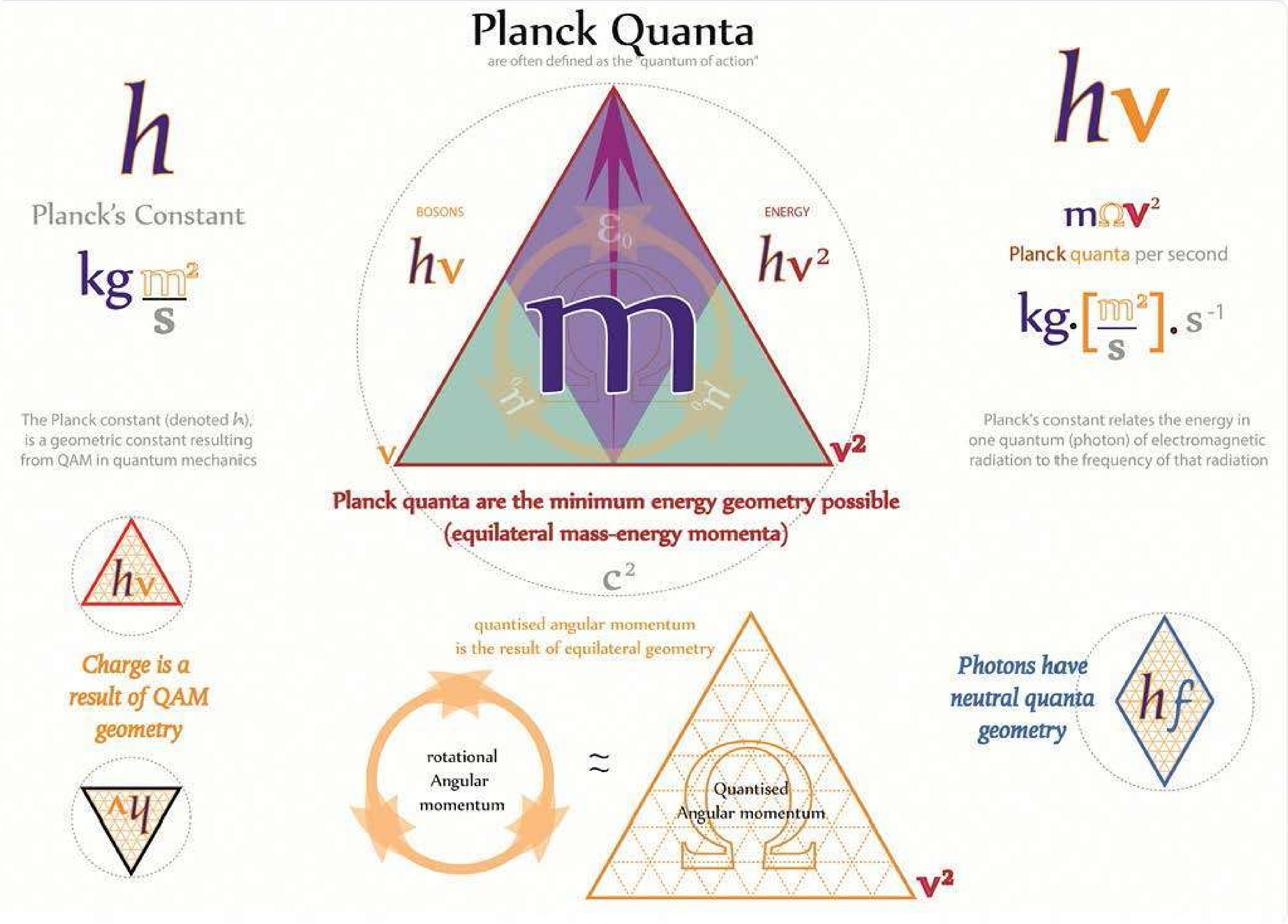

Planck Quanta

are often defined as the "quantum of action"

h

Planck's Constant

$\mathrm{kg}\,\frac{\mathrm{m}^2}{\mathrm{s}}$

The Planck constant (denoted h), is a geometric constant resulting from QAM in quantum mechanics

BOSONS
$h\nu$
ENERGY
$h\nu^2$
ε_0
m
μ_0
μ_0
v
v^2

Planck quanta are the minimum energy geometry possible (equilateral mass-energy momenta)

c^2

$h\nu$

$m\Omega v^2$

Planck quanta per second

$\mathrm{kg}.\left[\frac{\mathrm{m}^2}{\mathrm{s}}\right].\mathrm{s}^{-1}$

Planck's constant relates the energy in one quantum (photon) of electromagnetic radiation to the frequency of that radiation

Charge is a result of QAM geometry

quantised angular momentum is the result of equilateral geometry

rotational Angular momentum
≈
Quantised Angular momentum
v^2

Photons have neutral quanta geometry

Copyright ABRAHAM [2008] - all rights reserved

$E = n.h\nu$

Planck Number

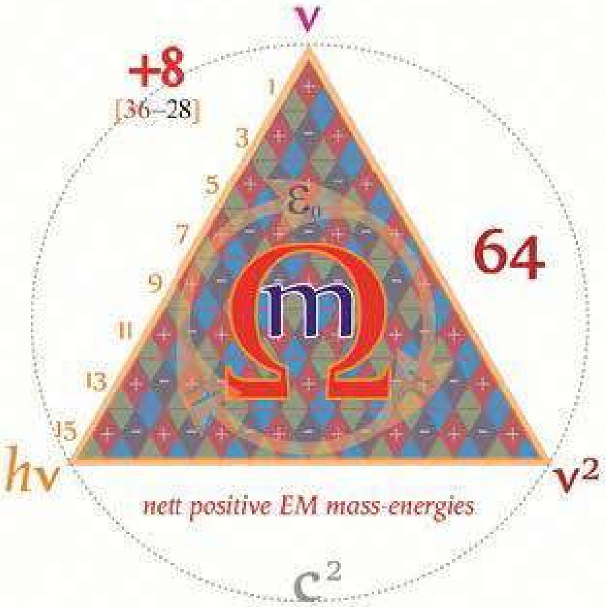

v
+8
[36–28]
1
3
5
7
9
11
13
15
Ɛn
m
Ω
64
hv
v²
nett positive EM mass-energies
c²

$h\nu$

tranverse
Bosons

Max Planck

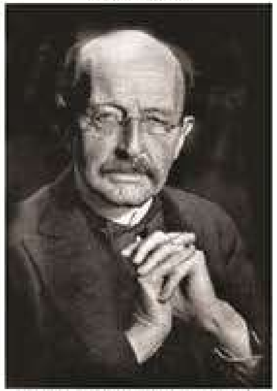

(April 23, 1858 – October 4, 1947)

hf

longitudinal
Photons

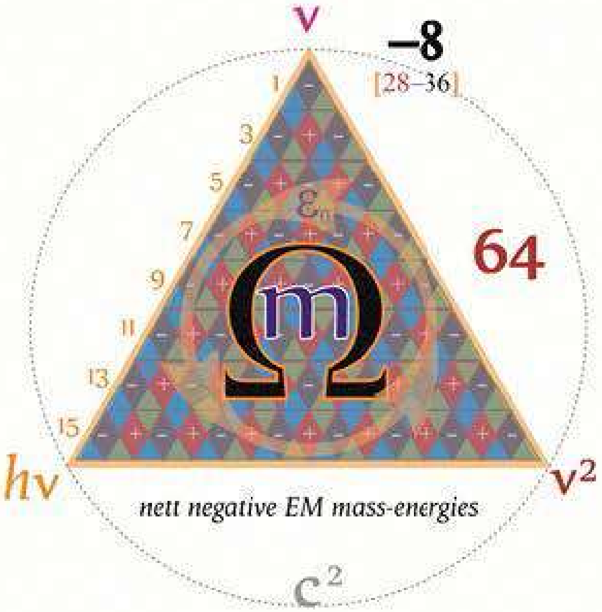

v
–8
[28–36]
1
3
5
7
9
11
13
15
Ɛn
m
Ω
64
hv
v²
nett negative EM mass-energies
c²

$$1e19\nu = \underset{\text{quantum number}}{n} = 5e18f$$

BOSONS
are the quanta of Charge

$$\underset{\text{Boson}}{1\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck quanta}}{\Big[\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}\Big]}\Big]$$

Equating **nu**mber of quanta
is the source of a number

$$h\nu \neq hf$$

Planck Einstein

$$\underset{\text{Photon}}{2\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck quanta}}{\Big[\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}\Big]}\Big]$$

to photon **f**requency
of quantum misconceptions

PHOTONS
are the neutral quanta of EM waves

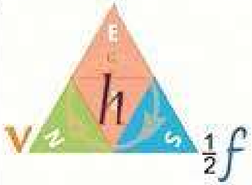

E
h
v
N
S
$\frac{1}{2}f$

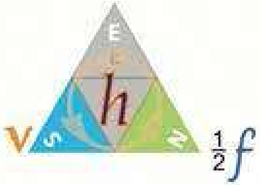

E
h
v
S
N
$\frac{1}{2}f$

The Planck constant was first described as the proportionality constant between
the energy (E) of a photon and the frequency (ν) of its associated electromagnetic wave.
[it in fact describes the energy of bosons]

The zero-point energy for a
simple harmonic oscillator of frequency f
is ½ hf

The relation between the energy and frequency is called the Planck–Einstein relationship.

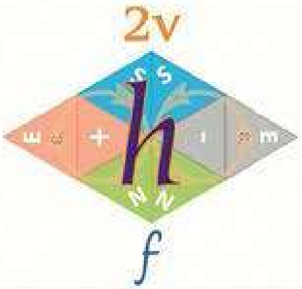

2v
S
S
E
+
h
-
m
N
N
f

Copyright ABRAHAM [2008] - all rights reserved

Quantum Inductors

+ ZPF

*A positive ZPF can
be viewed as a quantun quoin
or an quantum 'ideal' inductor
with an internalised energy flux
that is the opposite of
a negative ZPF*

**Clockwise
current flow**

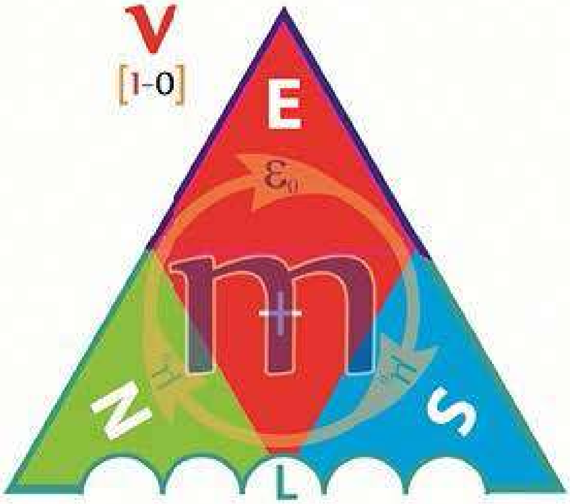

POSITIVE

Quantum L circuit
[quantum inductive circuit]

*An "ideal inductor" has inductance,
but no resistance or capacitance, and will not dissipate energy
(until it interacts with other ZPFs or Matter) and forms the basis for
all Charge-Parity-Time [CPT] interactions*

A ZPF is fixed in either a + or - state
[Quantum Inductor circuit]

Its energy flux direction as modelled electrically] is relative
to the observer's view or the direction of measurement

***Charge polarity is opposed on opposite faces
of the same quoin [quantum coin]***

The direction of inherent energy flux
from the presective of the observer
determines ZPF charge polarity

***The Quantum Inductor (L) circuit stores
energy as EM mass in π geometries,
it does not oscillate***

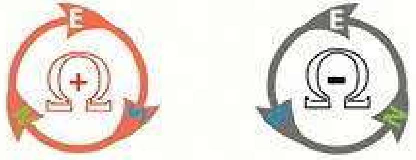

Energy received is stored indefinitely
until its release via weak interaction

[Inductive Magnetic coupling]

***The quantum Inductive circuit
is a SINGLE charge tri-field inductive energy loop***

*It does not oscillate energy between two opposing charges
its differing energy fields are the reuslt of its equilateral QAM geometry
storing Electric energy in its E field, and Magnetic energy in its M field*

The direction of the QAM flux that models inertia is relative
to the observer

- ZPF

*A negative ZPF can
be viewed as a quantum quoin
or an quantum 'ideal' inductor
with an internalised energy flux
that is the opposite of
a positive ZPF*

**Counter-Clockwise
Current Flow**

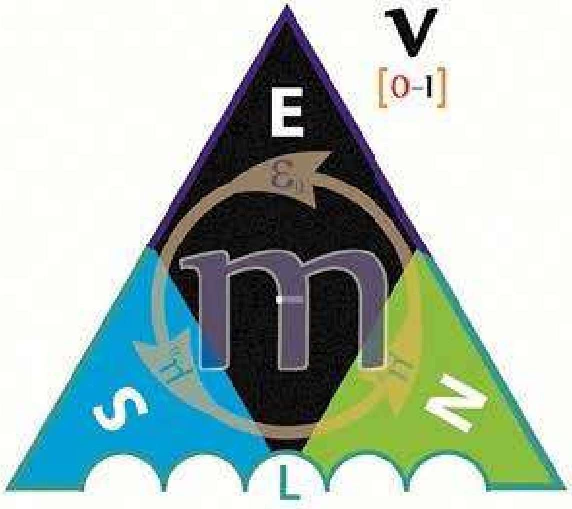

NEGATIVE

Quantum L circuit
[quantum inductive circuit]

Quantum Inductors and EM energy Levels

quantum Inductors

[ZPFs]

[Bosons]

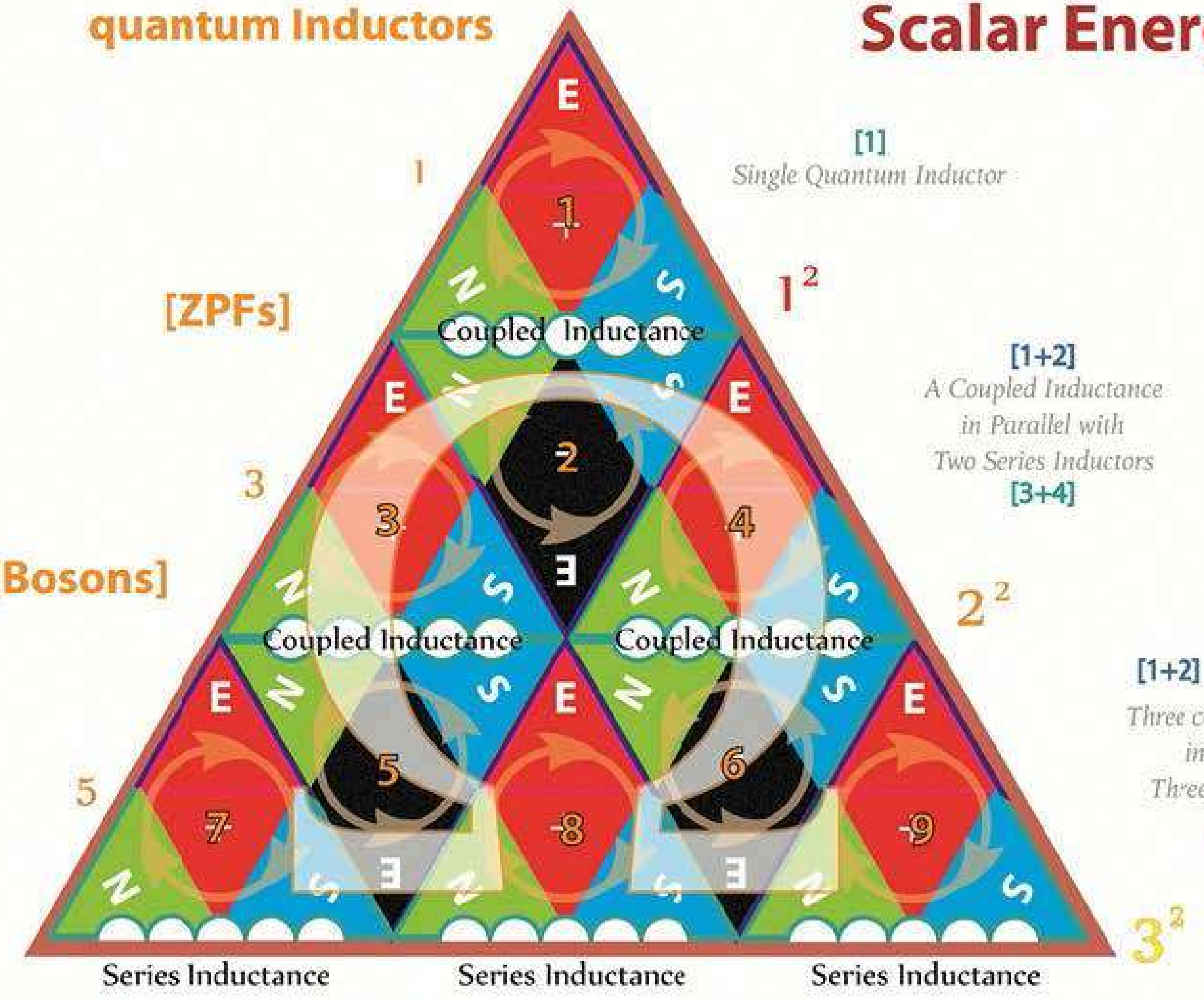
E
1
1
Coupled Inductance
N
S
E
E
2
3
4
3
Coupled Inductance
Coupled Inductance
E
E
E
5
6
5
7
8
9
1^2
2^2
3^2
Series Inductance
Series Inductance
Series Inductance

Scalar Energies

[1]
Single Quantum Inductor

[1+2]
Coupled
Quantum Oscillator

[1+2]
A Coupled Inductance
in Parallel with
Two Series Inductors
[3+4]

E

[1+2] [3+5] [4+6]
Three coupled Inductances
in Parallel with
Three Series Inductors
[7+8+9]

All quantum levels form inductive magnetic dipole bases
[Weak Force]

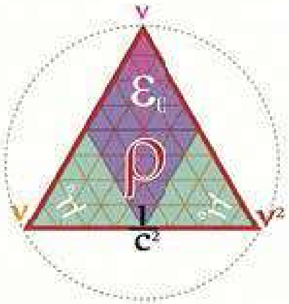
v
ε_0
ρ
v
v^2
1
c^2

coupled ODD# mass-energies
form SQUARE energy geometries

Quantum energy levels can be viewed as various combinations of:

Coupled Inductances

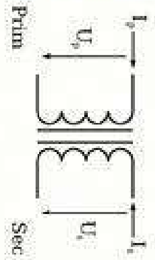
Prim
U_p
I_p
Sec
U_s
I_s

Parallel Inductances

$$L_{total} = \frac{1}{\frac{1}{L_1} + \frac{1}{L_2} + \ldots \frac{1}{L_n}}$$

Series Inductances

$$L_{total} = L_1 + L_2 + \ldots L_n$$

The Golden Triangle

Charged mass-energy geometries

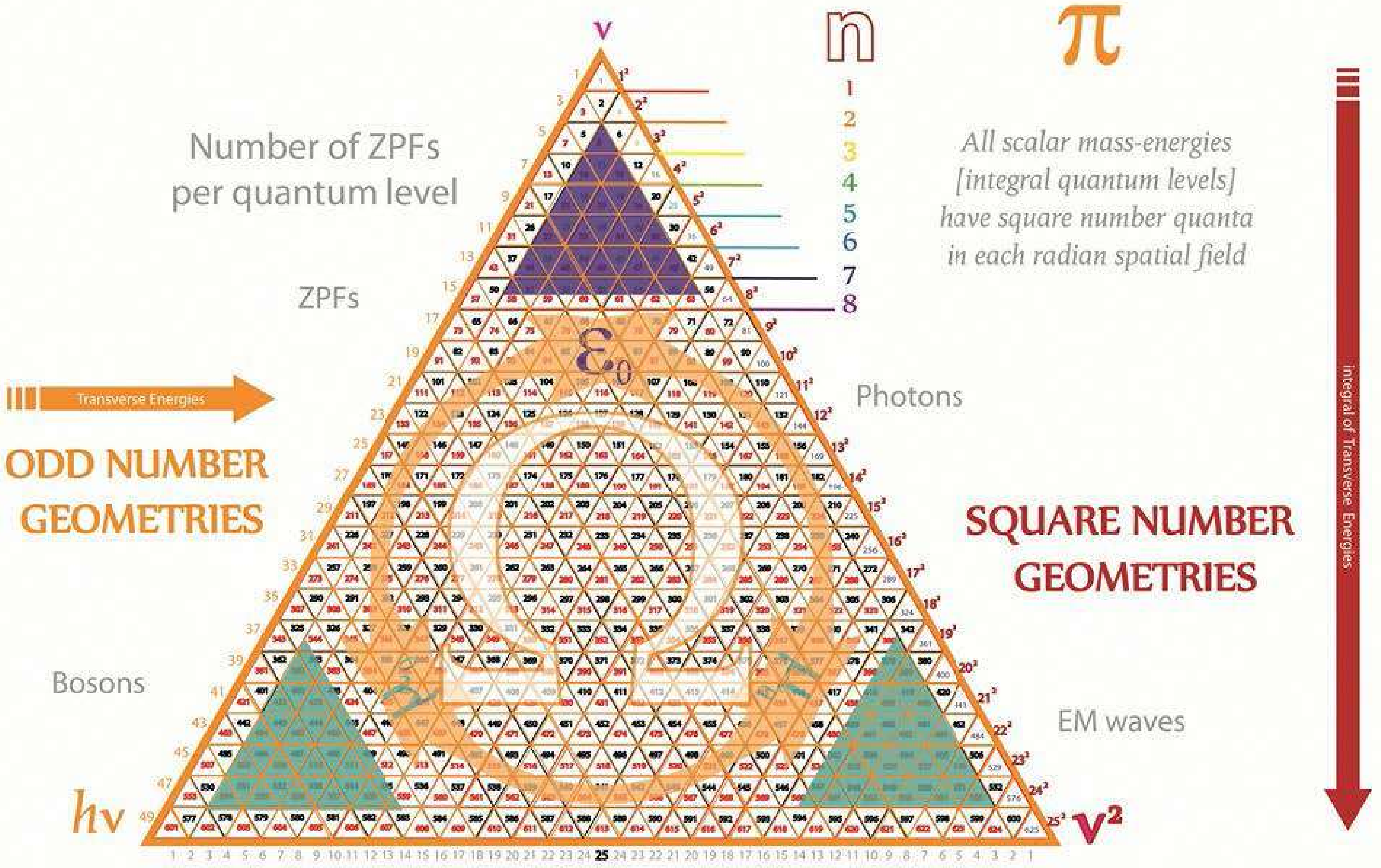
v
n
1
2
3
4
5
6
7
8
π
ZPFs
ε0
Photons
Transverse Energies
Bosons
EM waves
hv
v²
Integral of Transverse Energies
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Number of ZPFs per quantum level

All scalar mass-energies [integral quantum levels] have square number quanta in each radian spatial field

ODD NUMBER GEOMETRIES

SQUARE NUMBER GEOMETRIES

Probabilities **NORMAL DISTRIBUTIONS** Wavefunctions

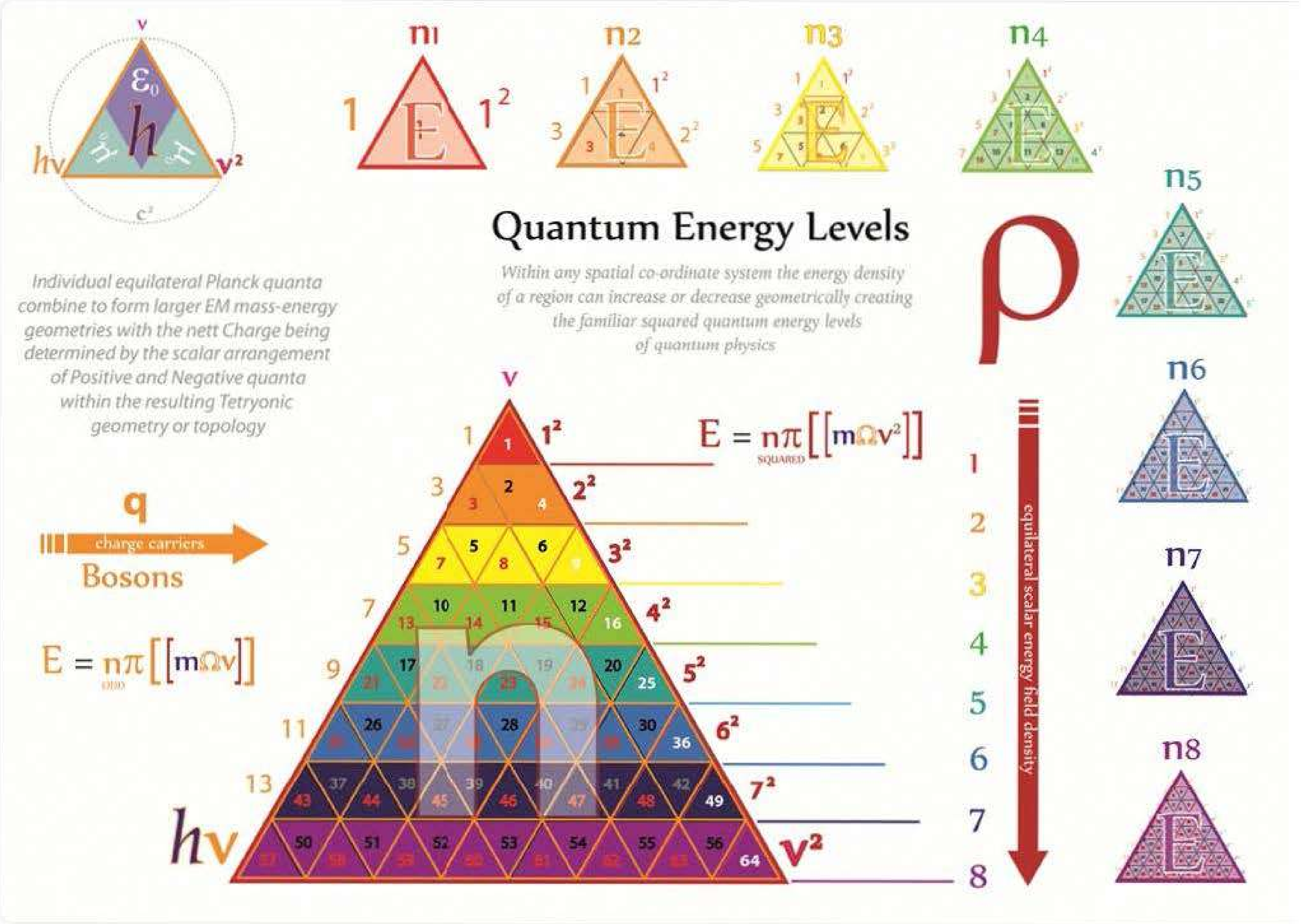

Quantum Energy Levels

Within any spatial co-ordinate system the energy density of a region can increase or decrease geometrically creating the familiar squared quantum energy levels of quantum physics

Individual equilateral Planck quanta combine to form larger EM mass-energy geometries with the nett Charge being determined by the scalar arrangement of Positive and Negative quanta within the resulting Tetryonic geometry or topology

v
ε_0
h
hv
v²
c²
n1
n2
n3
n4
n5
n6
n7
n8
E
1
1²
3
2²
5
3²
7
4²
9
5²
11
6²
13
7²
8²
ρ
q
charge carriers
Bosons
$E = n\pi[[m\Omega v]]$
$E = n\pi_{\text{SQUARED}}[[m\Omega v^2]]$
equilateral scalar energy field density
1
2
3
4
5
6
7
8
hv
v²
n

Tetryonics 03.13 - Quantum Energy levels

Copyright ABRAHAM [2008] - all rights reserved

Magnetic Vectors

Intrinsic Magnetic vectors are transverse to E fields

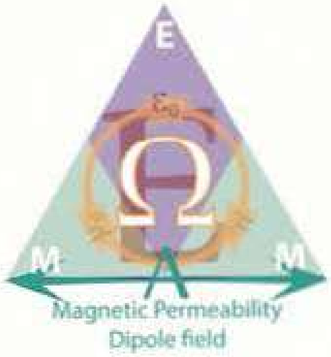
E
ε0
Ω
M
M
A
Magnetic Permeability
Dipole field

External Magnetic fields are termed B fields

Magnetic vectors can be modelled geometrically or electrically through energy field fluxes [quantised angular momenta]

Opposite charges moving in opposing directions produce co-directional Magnetic vectors

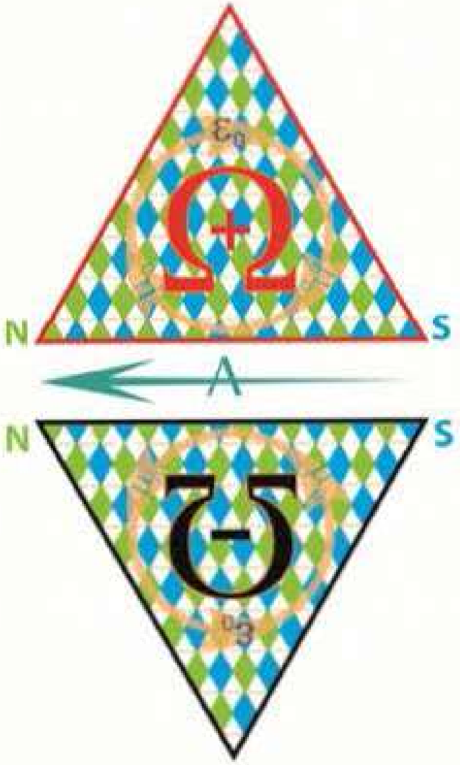
ε0
+
Ω
N
S
A
N
S
-
Ω
ε0

Within a Magnetic dipole

Magnetic Vector A

energy 'flows' South to North

External to a Magnetic dipole energy 'flows' North to South

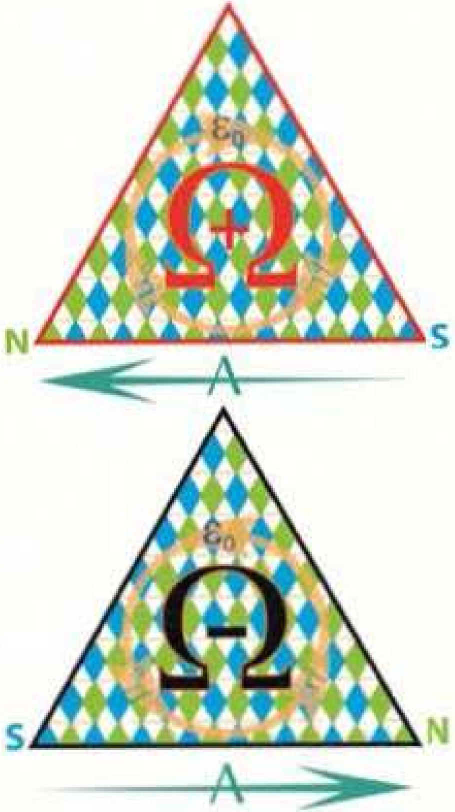
ε0
+
Ω
N
S
A
ε0
-
Ω
S
N
A

Opposite charges moving in the same direction produce Magnetic vectors in opposition

Charge Fields

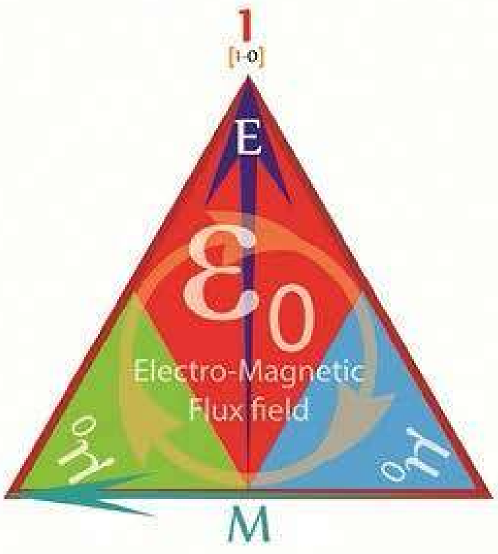
1
[1-0]
E
ε0
Electro-Magnetic
Flux field
μ0
μ0
M

EM energy fluxes in a Positive ZPF flows
are electrically modelled as Clockwise
(from North to South)

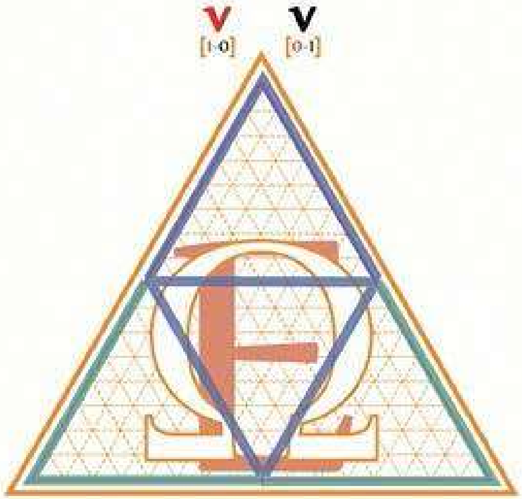

The EM flux directions of Charges
can be modelled vectorially with
Electric and Magnetic vectors

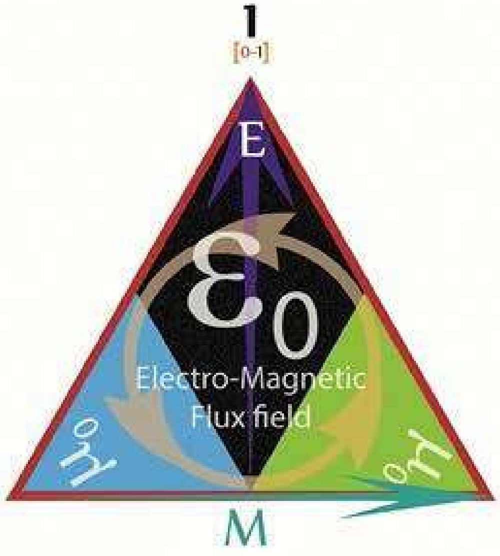
1
[0-1]
E
ε0
Electro-Magnetic
Flux field
μ0
μ0
M

EM energy fluxes in a Negative ZPF
are electrically modelled as counter-Clockwise
(from North to South)

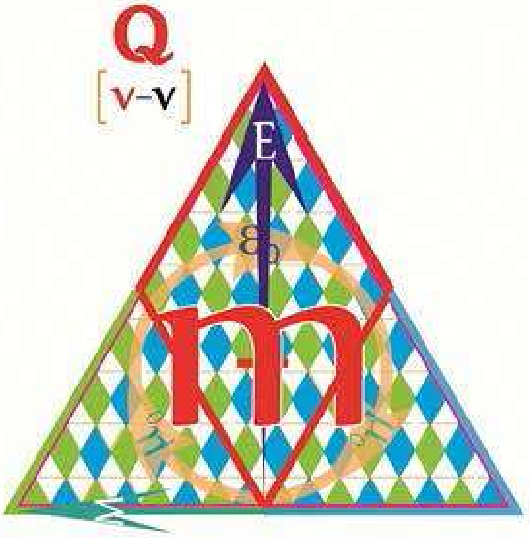
Q
[v-v]
E
ε0
m
μ0
μ0

*All rotational planck energy fluxes
can serve as models for the nett
quantised angular momenta of
any mass-energy geometry*

Polarised
Electric and Magnetic fluxes
in ElectroMagnetic fields arise from
intrinsic quantised angular momentum

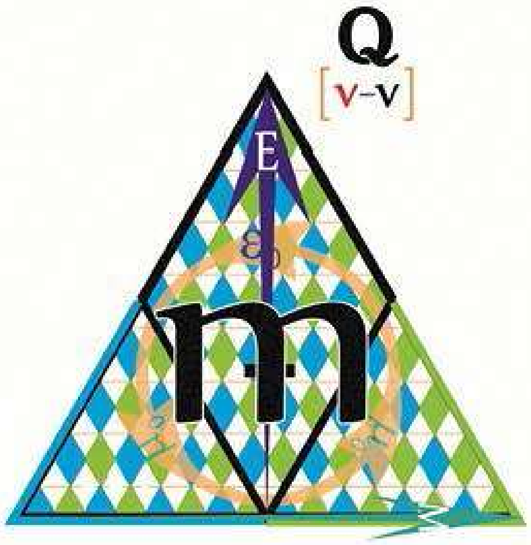
Q
[v-v]
E
ε0
m
μ0
μ0

EM fields

are the combined Electric & Magnetic fields
resulting from the quantised angular momentum
of mass-energy in any region of free space

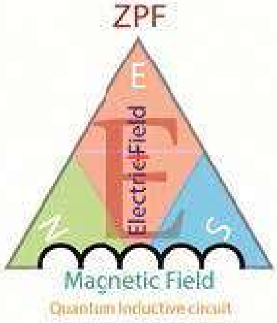

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

EM Permittivity-Permeability is a measure of how much resistance
is encountered when the quantised angular momenta of
EM energies form an electro-magnetic field in a vacuum

$$\text{Celeritas} = 299{,}792{,}458\ \frac{m}{s}$$

$$c_0 = \frac{1}{\sqrt{\mu_0\varepsilon_0}}.$$

EM field Permittivity-Permeability

$$\varepsilon_0\mu_0 = 1.112650056\text{e-}17\ \frac{s^2}{m^2}$$

$$\left[\frac{A^2\,s^4}{kg\,m^3}\cdot\frac{kg\,m}{A^2\,s^2}\right]$$

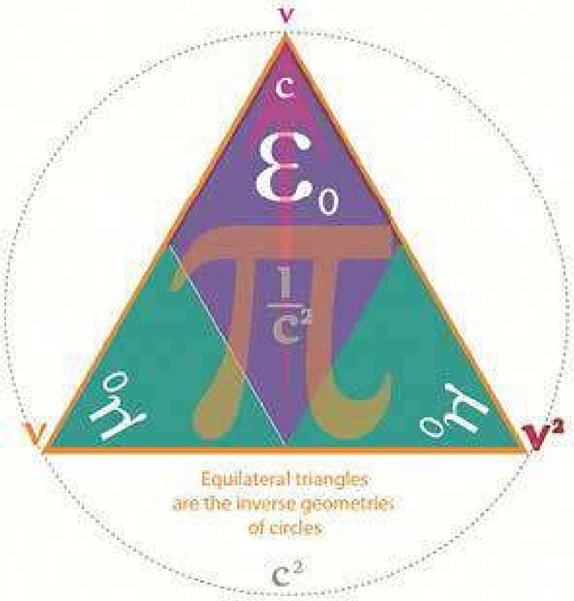

Equilateral triangles
are the inverse geometries
of circles

m		$n\pi$	$[\varepsilon_0\mu_0]$	$[m\Omega v^2]$
mass	=	charged	Spatial geometry	Energy density

Electric Constant $= 8.85418785\text{ e-}12\ \frac{F}{m}$ $\quad\frac{A^2\,s^4}{kg\,m^3}$

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

The permittivity of empty space, equal to 1 in centimeter-gram-second electrostatic units and numerically, to 8.854 × 10-12 farad per meter in International System units where c is the speed of light in meters per second. Symbolized ε0.

Magnetic Constant $= 1.25663706\text{ e-}6\ \frac{H}{m}$ $\quad\frac{kg\,m}{A^2\,s^2}$

$$k_0 = \frac{1}{4\pi\varepsilon_0}$$

A measure of the degree to which molecules of some material polarize (align) under the influence of an electric field; symbol k0, units F/m (farads per metre).

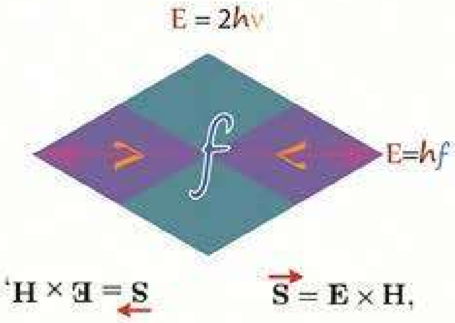

The Energy-momenta
of ZPFs form
natural Poynting vectors

Copyright ABRAHAM [2008] - all rights reserved

EM field Permittivity

The Electric constant, commonly called
the vacuum permittivity, or permittivity of free space,
relates the units for electric charge to
mechanical quantities such as
length and force.

The name Vacuum Permittivty is a misnomer and should be replaced with the correct term EM field Permittivity

The strength of Electric fields is determined by the Electrical Permittivity Constant

$$E = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}$$

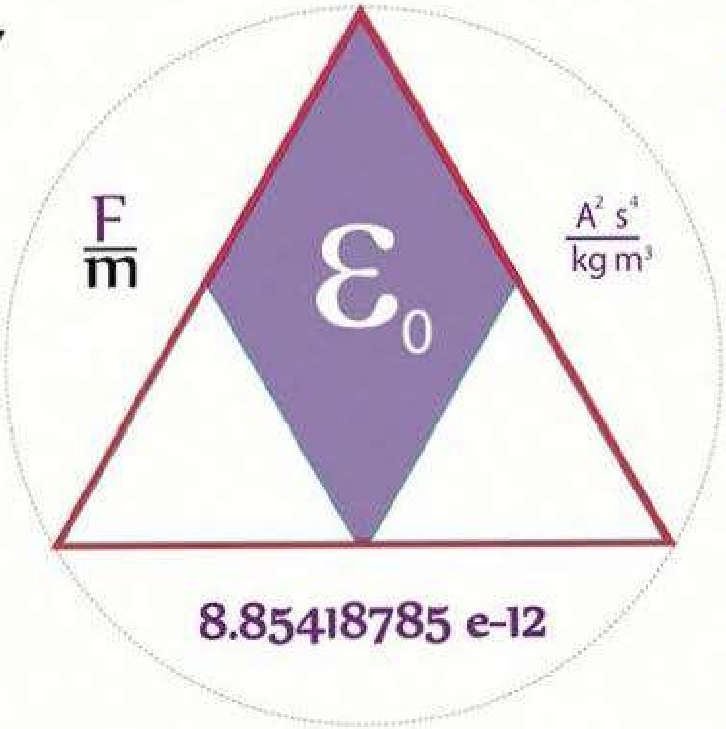

8.85418785 e-12

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

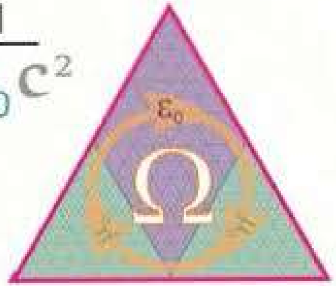

"Ampere's Law states that for any closed loop path, the sum of the quantities (B.ds) for all path elements into which the complete loop has been divided is equal to the product of μ0 and the total current enclosed by the loop."

$$k = \frac{1}{4\pi\varepsilon_0}$$

The permittivity of empty space, equal to 1 in centimeter-gram-second electrostatic units and to 107/4πc2 farads per meter or, numerically, to 8.854 × 10-12 farad per meter in International System units, where c is the speed of light in meters per second.

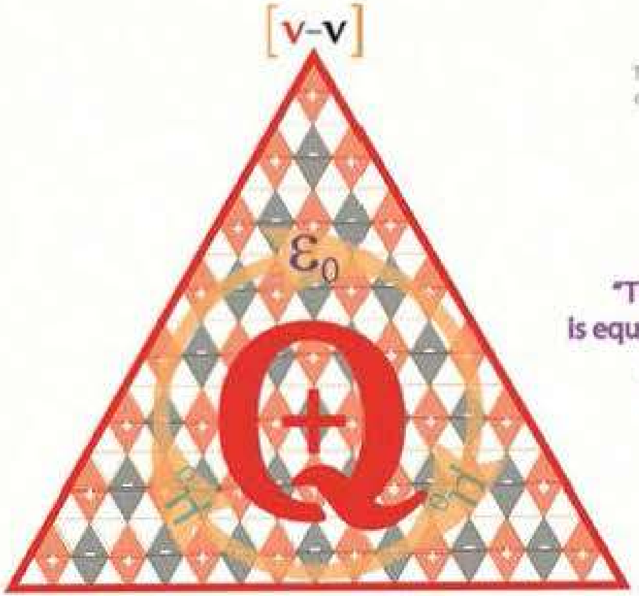

Positive Charge Electric Field

Gauss' Law:
"The total of the electric flux out of a closed surface is equal to the charge enclosed divided by the permittivity"

This applies equally to any geometry chosen to tessellate a surface area

Superpostioned E fields
gives rise to Coulomb Forces

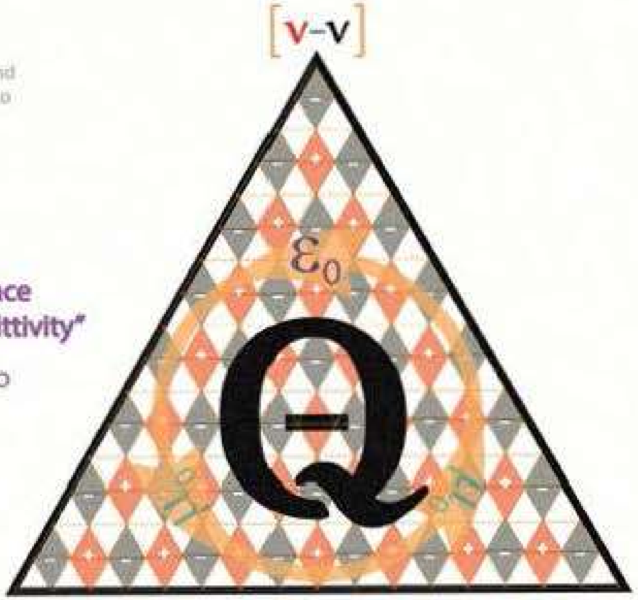

Negative Charge Electric Field

Copyright ABRAHAM [2008] - all rights reserved

Electric permittivity Fields

Negative externalised Planck quanta
(Counter-clockwise energy fluxs)

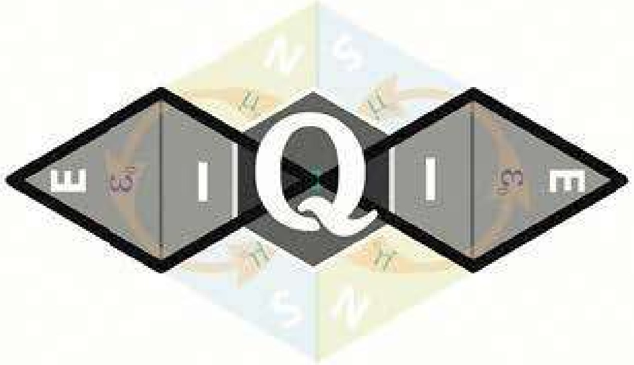

Positive externalised Planck quanta
(Clockwise energy fluxs)

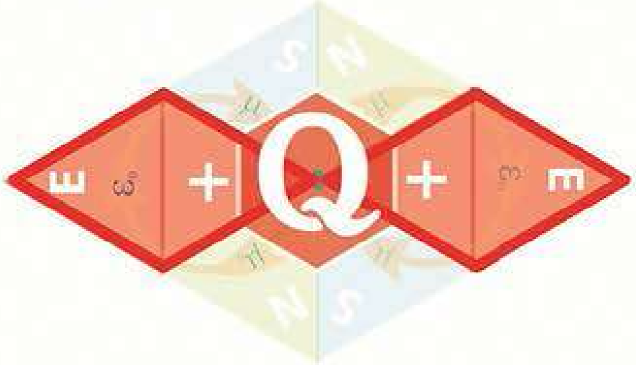

Coupled same charge ZPFs
have neutralised Magnetic fields

In Electro-statics superpositioned E fields
with interactive energy momenta
are the interactive mechanism
for Coulombic forces

Electrostatic Particles
in motion have Kinetic energies
resulting in Magnetic moments

Polar view

Negative E-fields

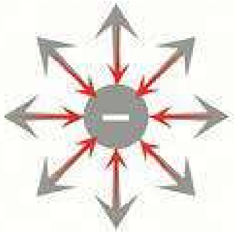

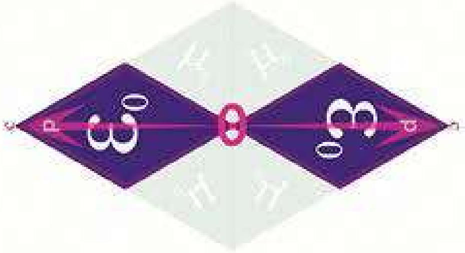

Polar view

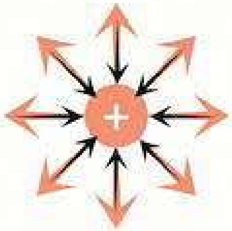

Positive E-fields

NEGATIVE electrostatic E fields
attract positive charges and
repel negative charges

ε_0

POSITIVE electrostatic E fields
attract negative charges and
repel positive charges

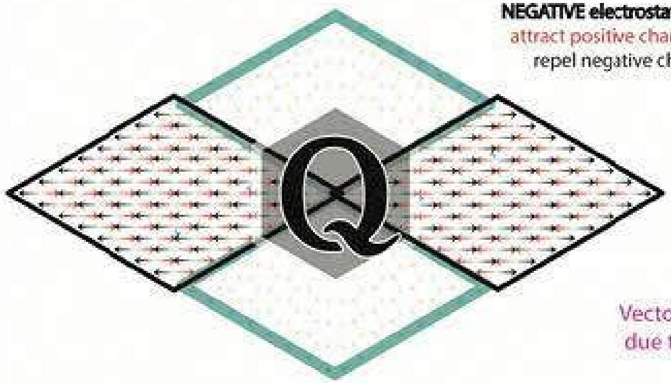

Opposites Attract
Similars Repel

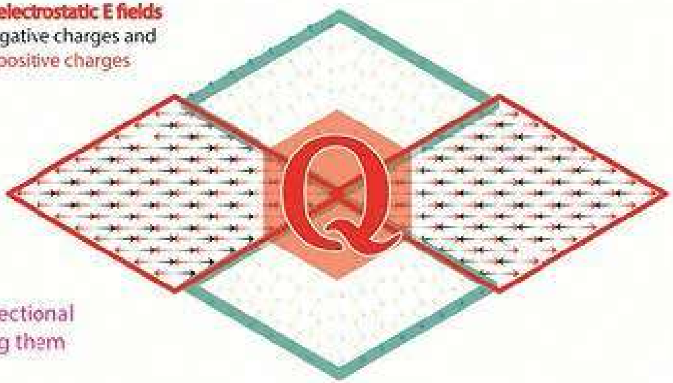

Vectorial momenta forces in EM fields are bi-directional
due to the energy-momenta quanta comprising them

The currently stated 'standard' premise of Electrical Energy flowing from Positive to Negative is misleading (as Energy also flows from Negative to Positive at the same time)

EM field Permeability

The permeability of free space,
also called absolute permeability.

The name Vacuum Permeability is a misnomer and should be replaced with the correct term EM field Permeability

The magnetic constant has the value of
4π x 10-7 henry per meter.

The strength of Magnetic fields is determined by the Magnetic Permeability Constant

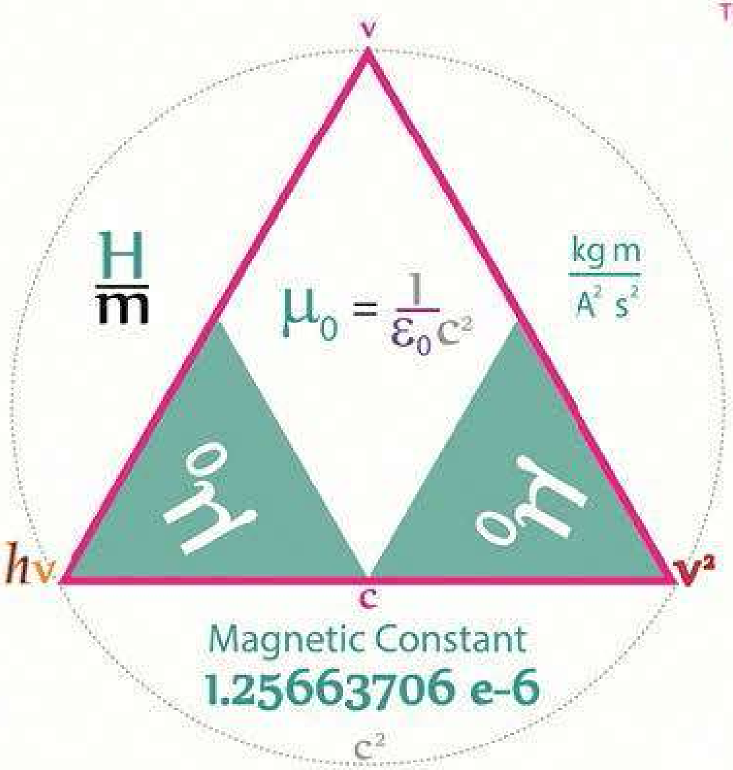
v
$\frac{H}{m}$
$\mu_0 = \frac{1}{\varepsilon_0 c^2}$
$\frac{kg\ m}{A^2\ s^2}$
μ_0
μ_0
hv
c
v²
Magnetic Constant
1.25663706 e-6
c²

F

The magnetic field is most commonly defined in terms of the Lorentz force it exerts on moving electric charges.

The magnetic field generated by a steady current (a constant flow of electric charges in which charge is neither accumulating nor depleting at any point) is described by the Biot-Savart law

$$\mu_0$$

$$B = \mu_0 H$$

A measure of the degree to which molecules of some material polarize (align) under the influence of an electric field.

Measured in units of units H/m (Henries per metre).

There are
NO magnetic monopoles
[not under any condition]

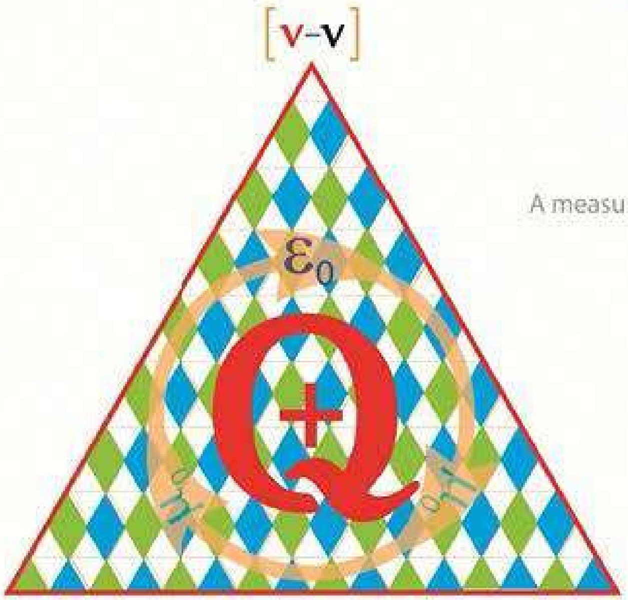
[v-v]
ε_0
+
Q
μ_0
μ_0

Positive Magnetic Moment

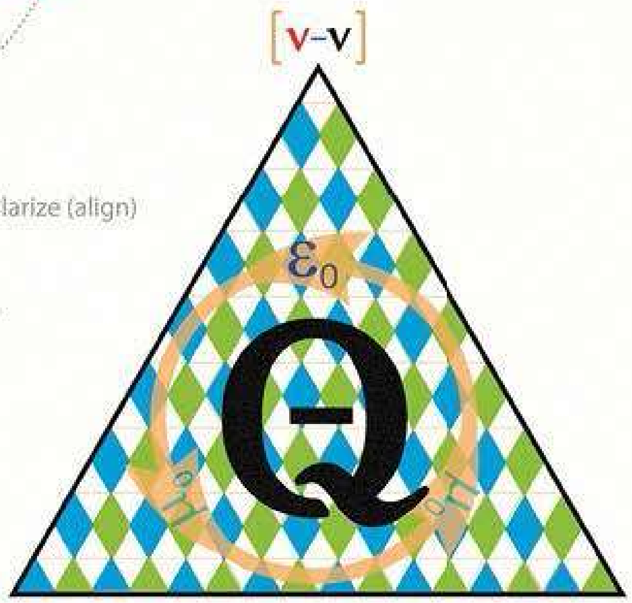
[v-v]
ε_0
-
Q
μ_0
μ_0

Negative Magnetic Moment

Magnetic permeability Fields

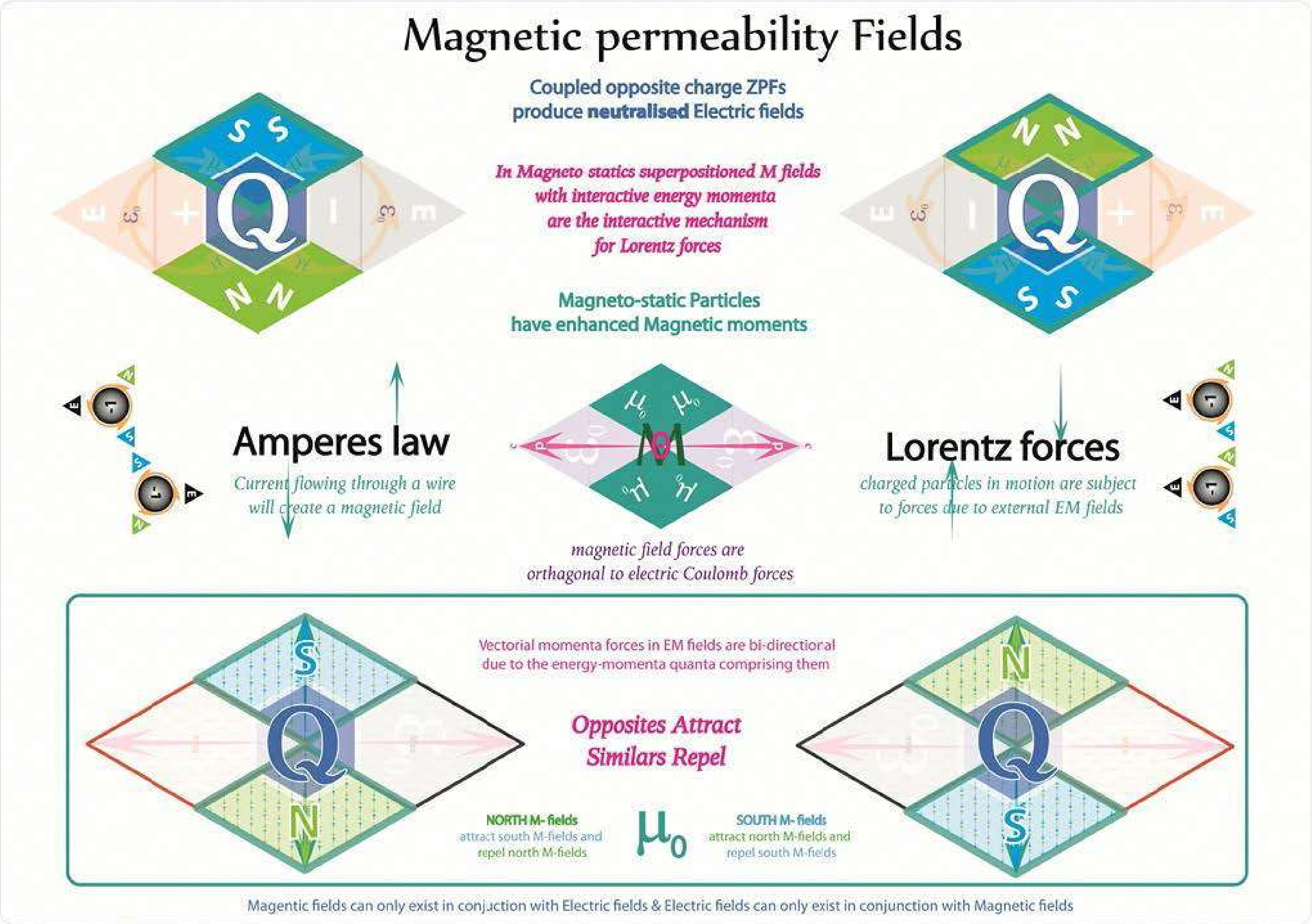

Coupled opposite charge ZPFs
produce **neutralised** Electric fields

In Magneto statics superpositioned M fields
with interactive energy momenta
are the interactive mechanism
for Lorentz forces

Magneto-static Particles
have enhanced Magnetic moments

Amperes law

Current flowing through a wire
will create a magnetic field

Lorentz forces

charged particles in motion are subject
to forces due to external EM fields

magnetic field forces are
orthagonal to electric Coulomb forces

Vectorial momenta forces in EM fields are bi-directional
due to the energy-momenta quanta comprising them

Opposites Attract
Similars Repel

NORTH M- fields
attract south M-fields and
repel north M-fields

SOUTH M- fields
attract north M-fields and
repel south M-fields

Magentic fields can only exist in conjuction with Electric fields & Electric fields can only exist in conjunction with Magnetic fields

Tetryonics 04.07 - Magnetic Permeability fields

Copyright ABRAHAM [2008] - all rights reserved

Electro-static fields

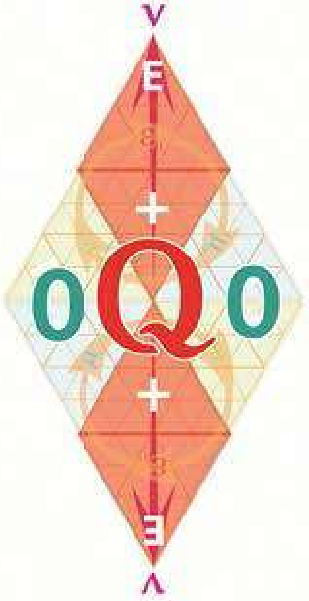

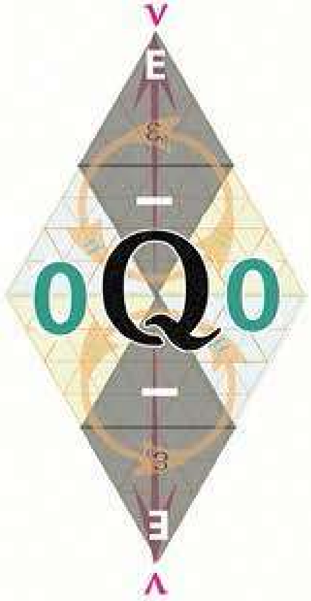

Similar charge electric dipole pairings
create 'neutralised' Magnetic dipoles

*There are no such things as purely
Electric or Magnetic fields*

v=0

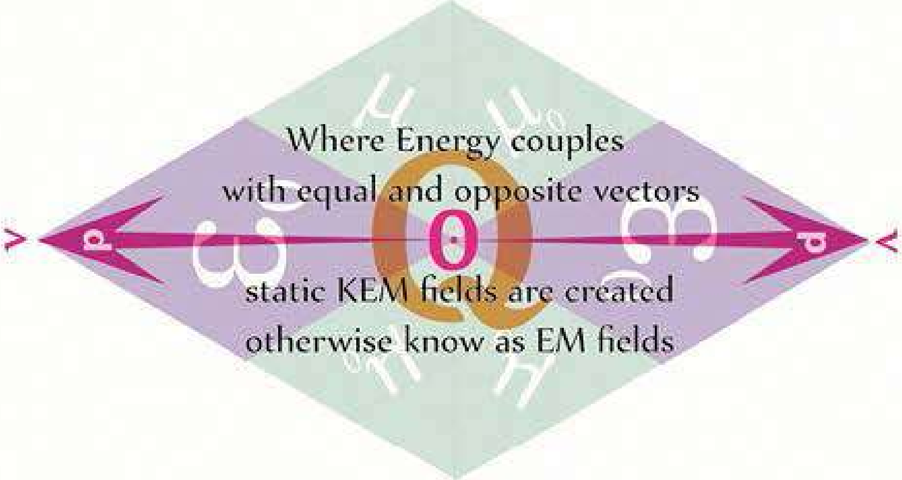

Where Energy couples
with equal and opposite vectors
static KEM fields are created
otherwise know as EM fields

**Fields
of
Force**

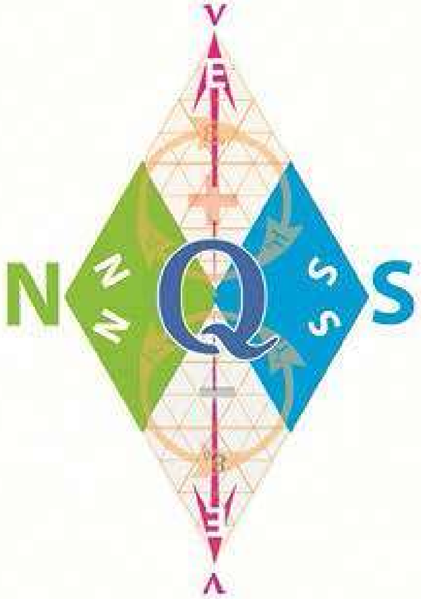

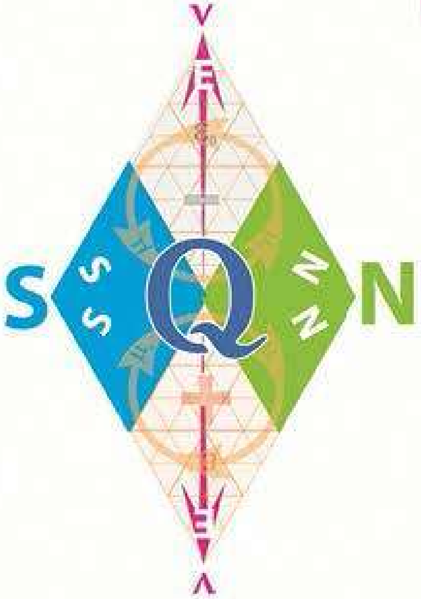

*All energy fields are Electro-Magnetic in nature
a direct product of their equilateral geometry*

'Neutralised' electric dipole pairings
create Magnetic dipole fields

Magneto-static fields

Charged EM field geometries

Opposite charge fields
can produce neutral E-fields
(with magnetic moments)

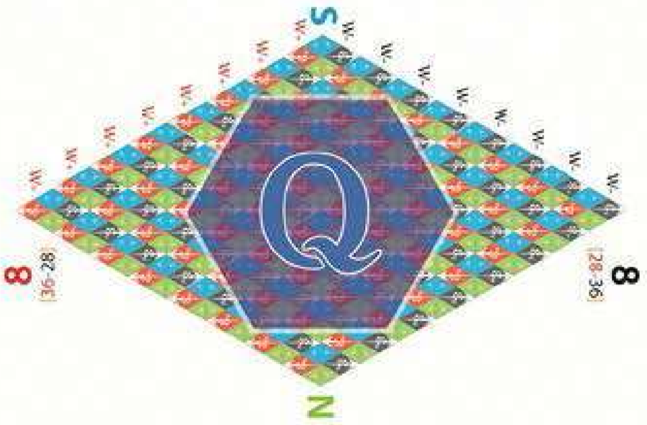

0
[64-64]

Electrostatic charged matter generate
charged energy fields around them

Moving charged particles generate
Kinetic energy & Magnetic moments

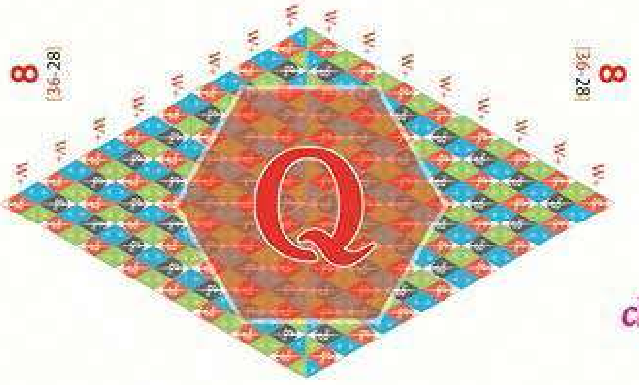

16
[72-56]

Charged electrostatic fields

*accelerate charged particles
vectorally dependent on their quantum
charge mass-energy momenta field geometries*

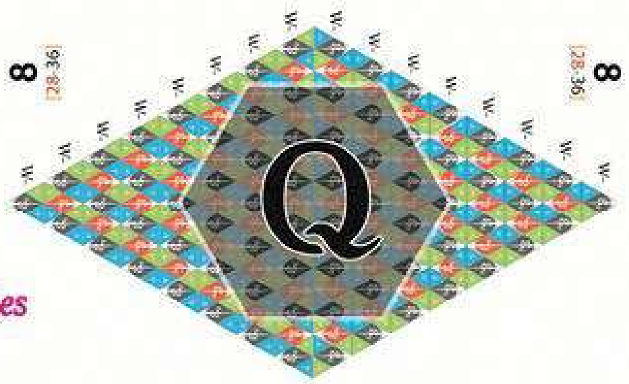

16
[56-72]

Positive charge electrostatic fields
attract Negative charges
repel Positive charges

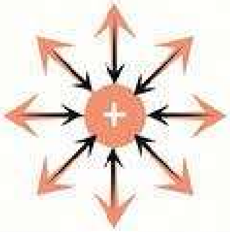

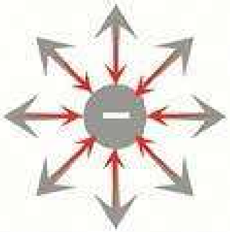

Negative charge electrostatic fields
attract Positive charges
repel Negative charges

Copyright ABRAHAM [2008] - all rights reserved

E field acceleration of charged particles

Electric fields can accelerate charged particles within their field geometry
dependent on the particle's nett charged mass-Matter topology

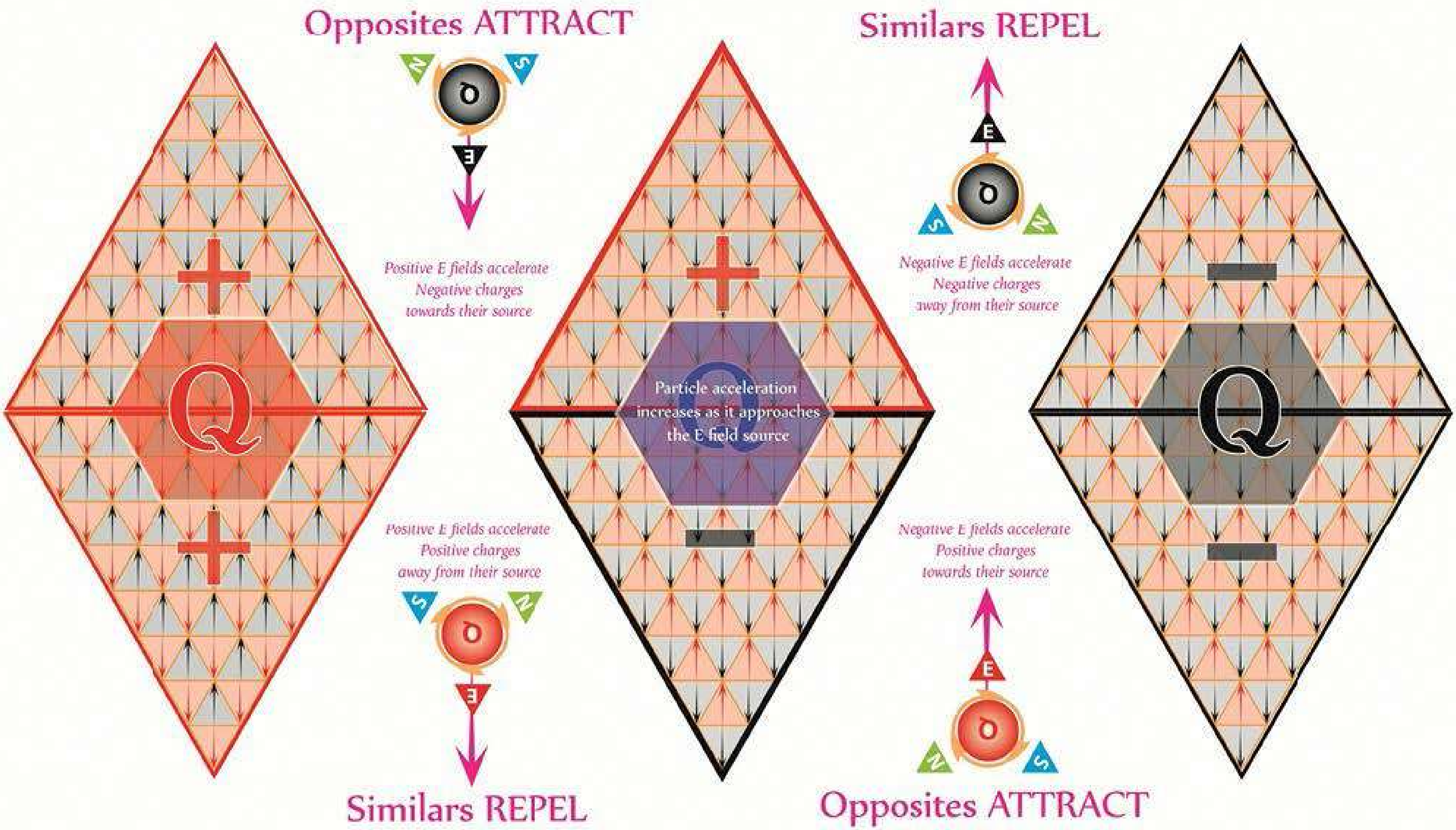

Opposites ATTRACT

Positive E fields accelerate
Negative charges
towards their source

Similars REPEL

Negative E fields accelerate
Negative charges
away from their source

Particle acceleration
increases as it approaches
the E field source

Positive E fields accelerate
Positive charges
away from their source

Similars REPEL

Negative E fields accelerate
Positive charges
towards their source

Opposites ATTRACT

The charge quanta within Neutral particles are affected equally by Electric fields

Tetryonics 04.10 - E field acceleration of charged particles

Copyright ABRAHAM [2008] - all rights reserved

ZPFs can combine to form
4 distinct static EM fields

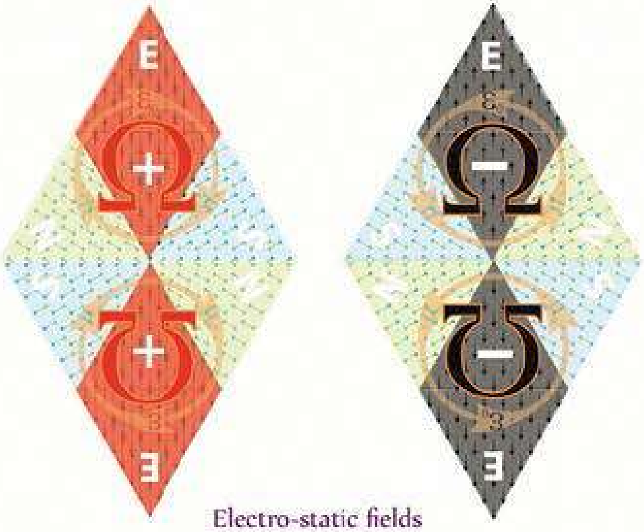

Electro-static fields

ElectroMagnetic fields

Magneto-static fields

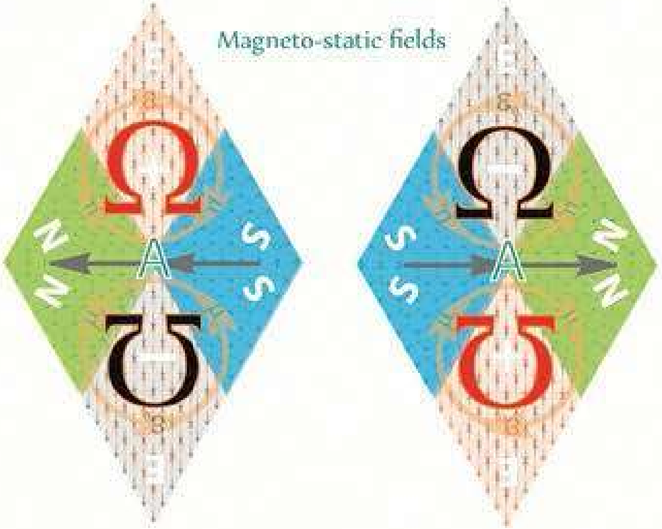

Similars REPEL
[QAM momenta vectors in opposition]

Opposites ATTRACT
[QAM momenta vectors in same direction]

'THE LAW OF INTERACTION'

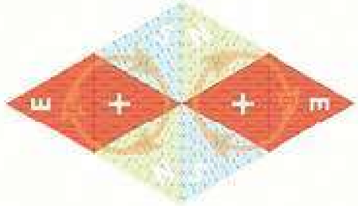

Flux momenta in
opposition

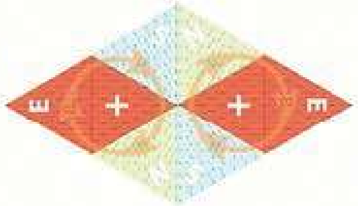

Electro-static fields
$1/r^2$

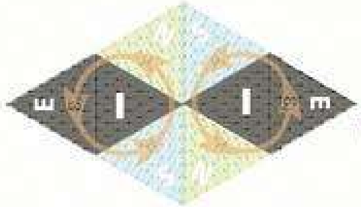

Flux momenta in
the same direction

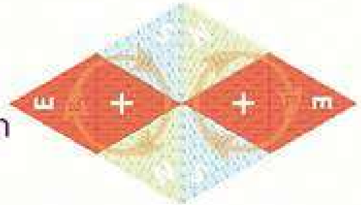

'FORCES OF INTERACTION'

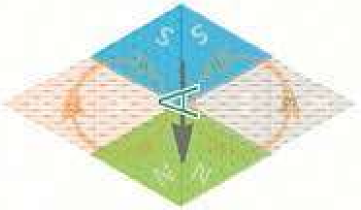

Parallel

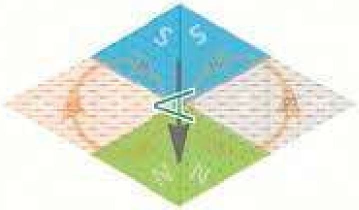

Flux momenta in
the same direction

Magneto-static fields
$1/r^2$

Flux momenta in
opposition

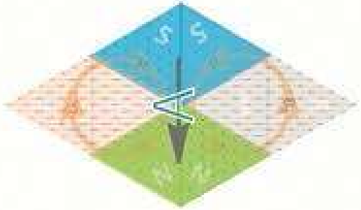

Anti-parallel

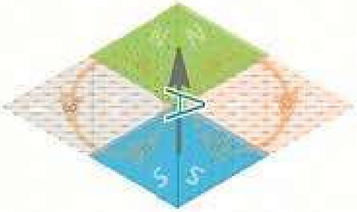

Positive electrostatic fields

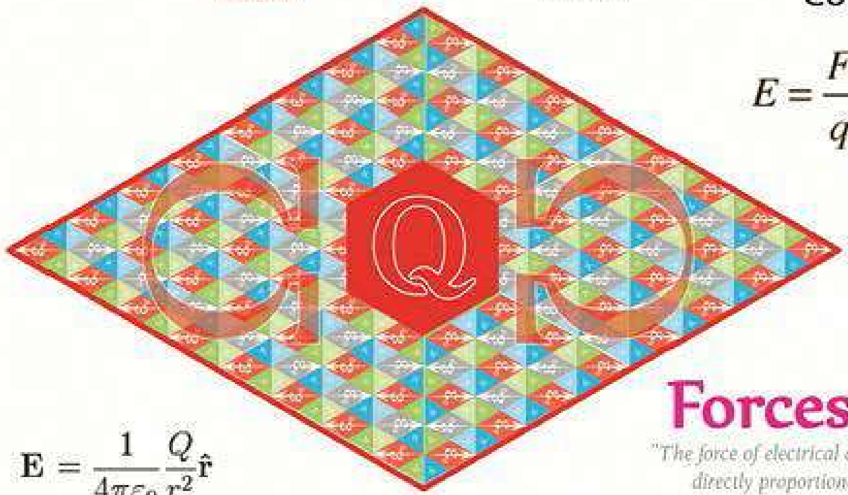
divergent
positive
momenta
convergent
negative
momenta
Q

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}$$

Coulomb's Force Law

$$E = \frac{F}{q} = \frac{kQ_{source}q}{qr^2} = \frac{kQ_{source}}{r^2}$$

Negative electrostatic fields

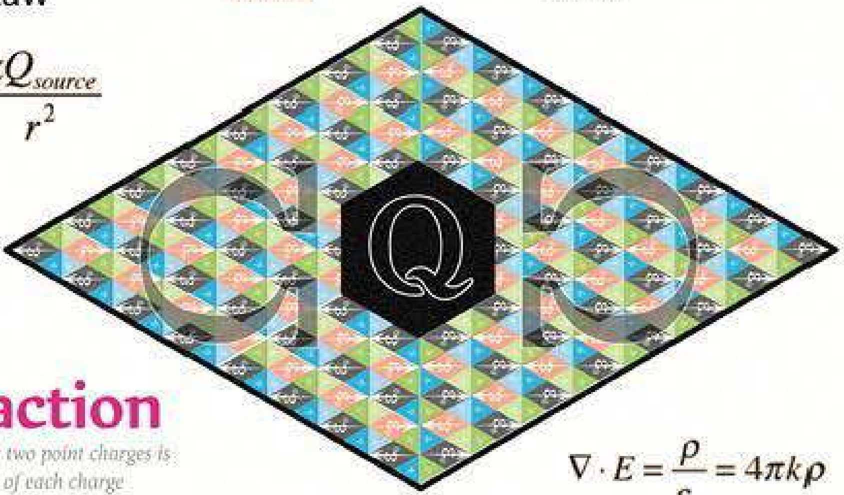
convergent
positive
momenta
divergent
negative
momenta
Q

$$\nabla \cdot E = \frac{\rho}{\varepsilon_0} = 4\pi k\rho$$

Forces of Interaction

"The force of electrical attraction or repulsion between two point charges is directly proportional to the product of magnitude of each charge and inversely proportional to the square of distance between them"

attract Negative charged masses
repel Positive charged masses

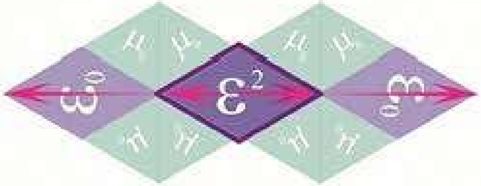
ε_0
ε^2
ε_0
μ

attract Positive charged masses
repel Negative charged masses

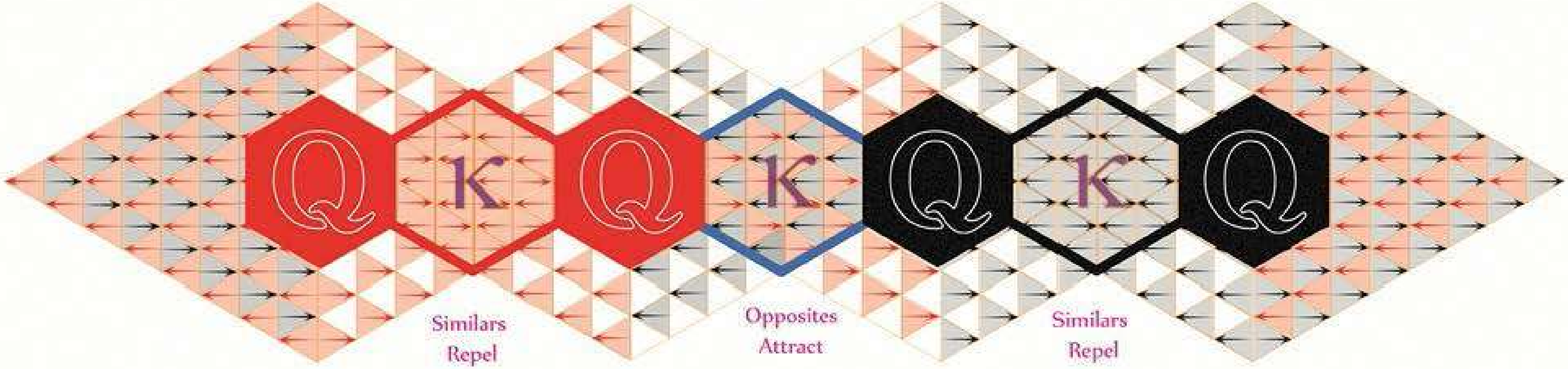
Q
K
Q
K
Q
K
Q
Similars
Repel
Opposites
Attract
Similars
Repel

Copyright ABRAHAM [2008] - all rights reserved

Electro-Magnetic field Lines

The magnetic field at any given point is specified by both its direction and magnitude

Magnetic dipoles can be produced by the coupling of opposite electric charge fields

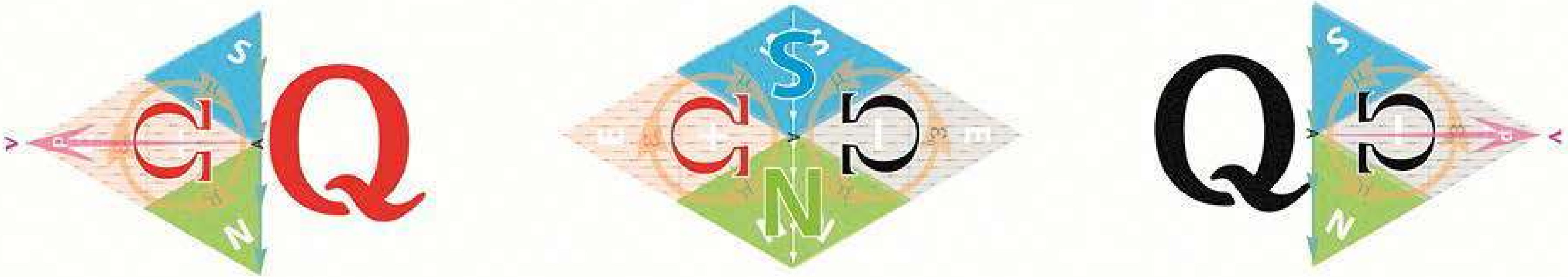

or

by accelerating charged bodies of Matter which in turn produce associated Kinetic energies
[Neutral Electric field and equal strength Magnetic moment]

Magnetic South Poles have divergent south vectors and convergent north vectors

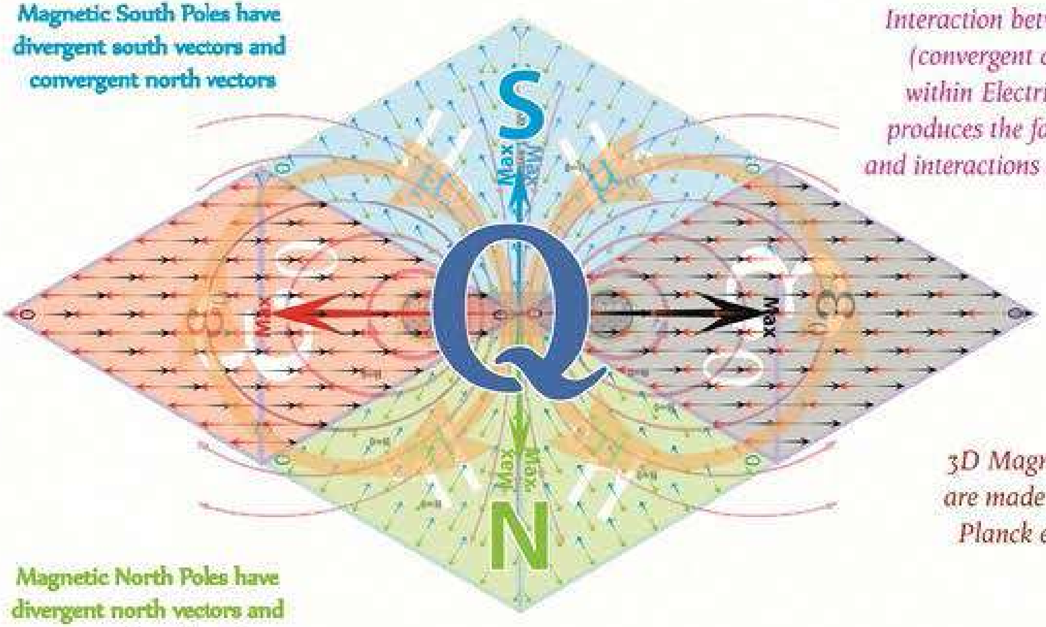

Magnetic North Poles have divergent north vectors and convergent south vectors

Interaction between vectorally opposed (convergent and divergent) vectors within Electric and Magnetic fields produces the familar lines of force and and interactions of Electro-Magnetic fields

3D Magnetic lines of Force are made up of 2D fields of Planck energy momenta

Every particle of Matter in motion possesses intrinsic Planck M-field dipoles becasue of their nett charged quanta which combine to produce a nuclear magneton.

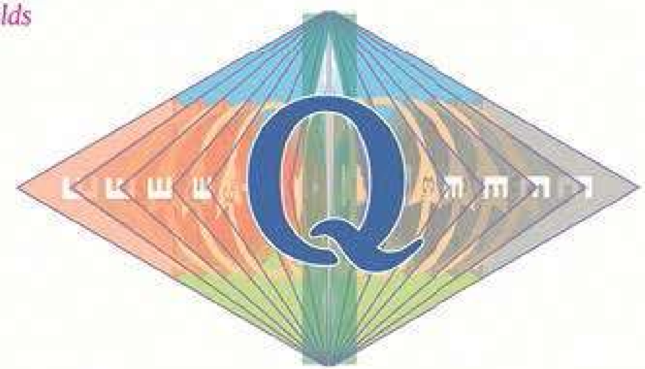

The magnetic field lines of permanent bar magnets are the result of the equilateral charge [QAM] geometry of neutralised quanta that form their electrostatic fields

Magnetic field Forces

$$\frac{m^2\cdot kg}{s^2\cdot A^2} = \frac{J}{A^2} = \frac{Wb}{A} = \frac{s^2}{F} = \frac{V\cdot s}{A} = \frac{J/C\cdot s}{C/s} = \frac{J\cdot s^2}{C^2} = \frac{m^2\cdot kg}{C^2}$$

In physics superpositioned M fields with interactive energy momenta produce Lorentz forces

attractive magnetic forces

Magnetic moments of same charges moving in the same direction

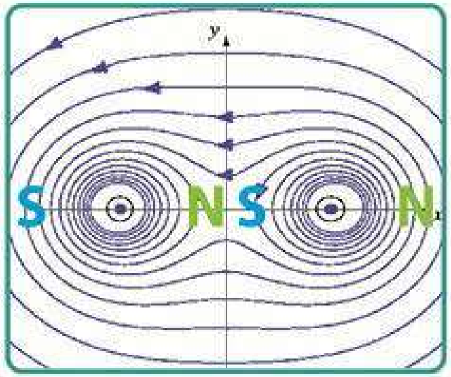

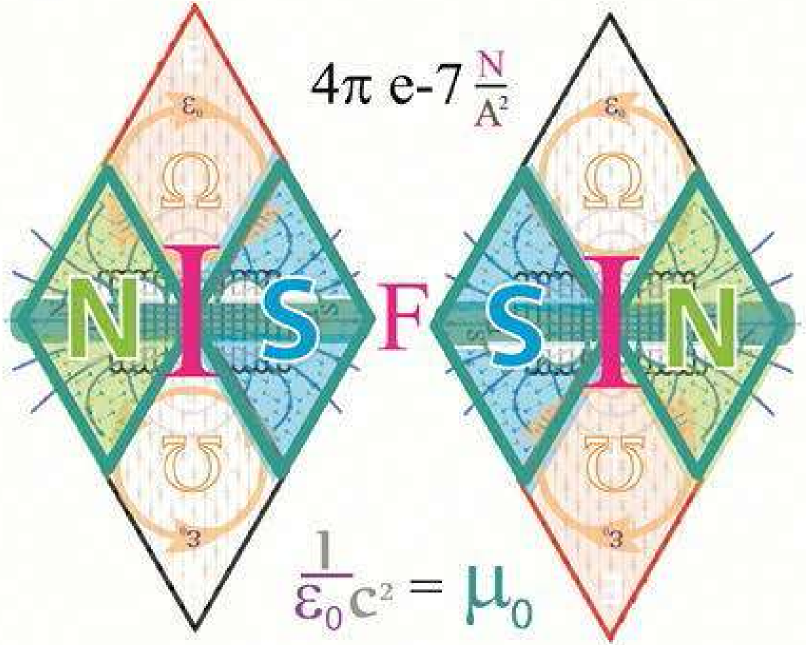

$$4\pi\ e\text{-}7\ \frac{N}{A^2}$$

$$\frac{1}{\varepsilon_0 c^2} = \mu_0$$

repulsive magnetic forces

Magnetic moments of same charges moving in opposite directions

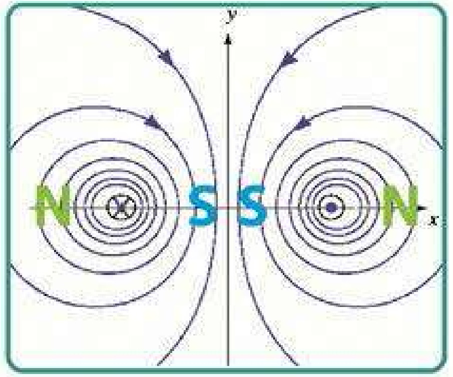

magnetostatic fields

magnetic dipole

$$H = \frac{B}{\mu_0}$$

Coupled opposite charge EM fields produce Magnetic dipole fields

Joseph Henry

(17 December 1797 – 13 May 1878)

$$\mu_B$$

Electrostatic charged particles in motion create magnetic moments

KE
magnetic moments
KE

kinetic EM fields

Magnetic field lines

Magnetostatic Dipole

Magnetic field lines were introduced by
Michael Faraday (1791-1867) who named them
"lines of force"

Michael Faraday

(22 September 1791 – 25 August 1867)

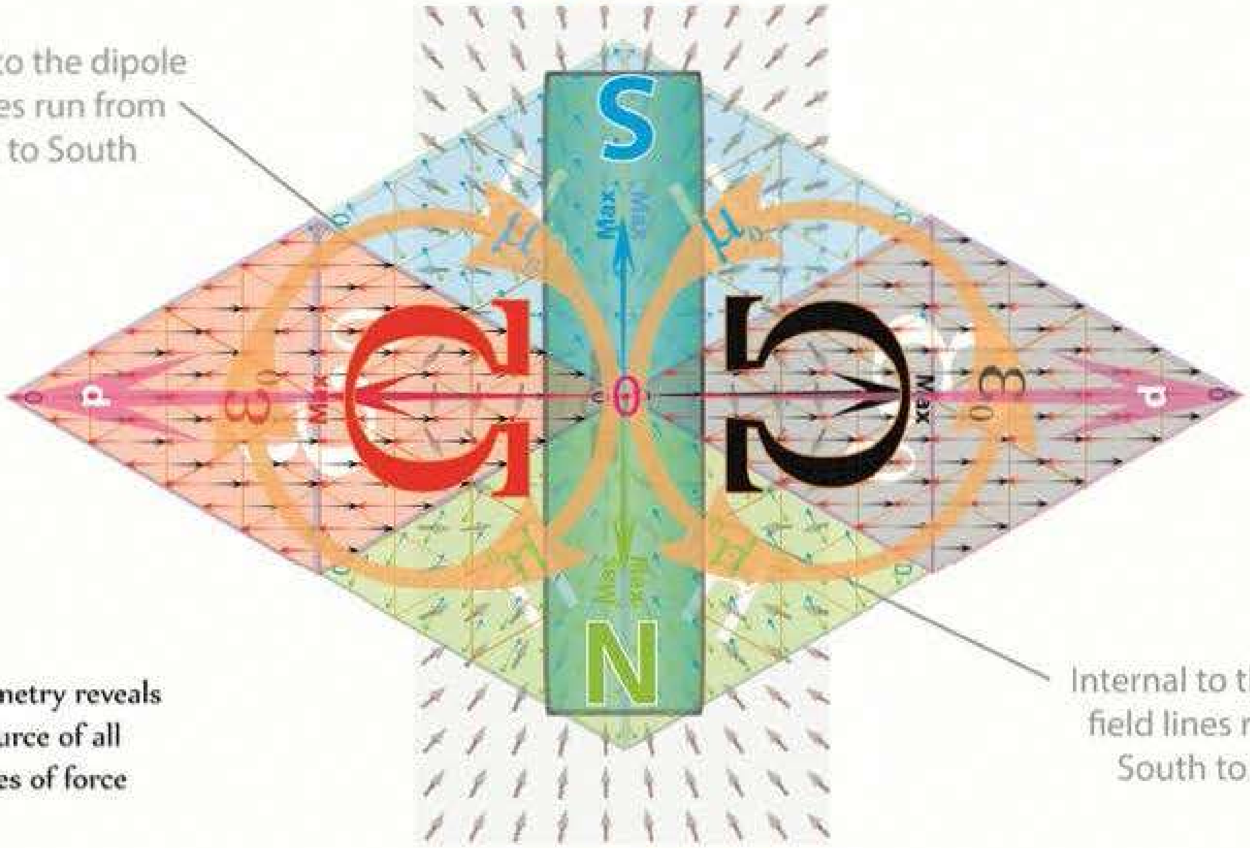

Magnetic South Poles have
divergent south vectors and
convergent north vectors

External to the dipole
field lines run from
North to South

Positive Electric fields have
divergent positive E vectors and
convergent negative E vectors

Negative Electric fields have
divergent negative E vectors and
convergent positive E vectors

Tetryonic geometry reveals
the true source of all
EM field lines of force

Internal to the dipole
field lines run from
South to North

Magnetic North Poles have
divergent north vectors and
convergent south vectors

Permanent Magnets can be viewed as
neutral charge Electrostatic
EM fields

Similar to conventional vs electron current flows
Magnetic field vectors can be modelled with
two directions of 'magnetic flux'

Parallel Magnetic Dipoles

Michael Faraday

(22 September 1791 – 25 August 1867)

James Clerk Maxwell

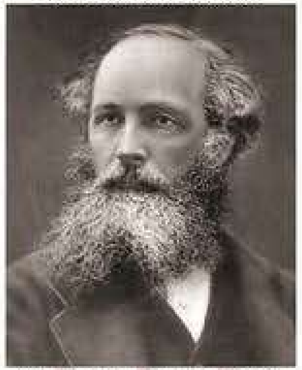

(13 June 1831 – 5 November 1879)

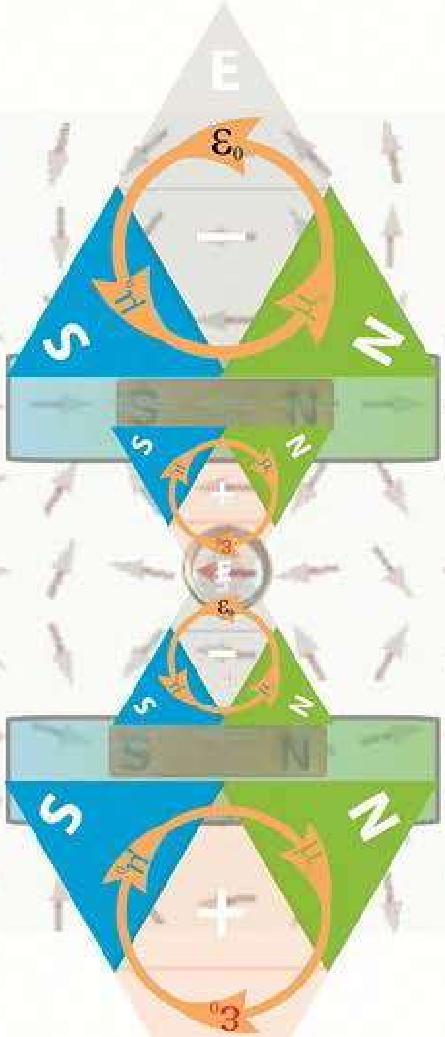

Magnetic lines of force are continuous and will always form closed loops.

Magnetic lines of force will never cross one another.

Parallel magnetic lines of force traveling in the same direction repel one another.

Parallel magnetic lines of force traveling in opposite directions tend to unite with each other and form into single lines traveling in a direction determined by the magnetic poles creating the lines of force.

Magnetic lines of force tend to shorten themselves.

Therefore, the magnetic lines of force existing between two unlike poles cause the poles to be pulled together.

Magnetic lines of force pass through all materials, both magnetic and nonmagnetic.

Magnetic lines of force always enter or leave a magnetic material at right angles to the surface

Maxwell had studied and commented on the field of electricity and magnetism as early as 1855/6 when "On Faraday's lines of force" was read to the Cambridge Philosophical Society.

The paper presented a simplified model of Faraday's work, and how the two phenomena were related. He reduced all of the current knowledge into a linked set of differential equations with 20 equations in 20 variables. [Quarterions]

This work was later published as "On physical lines of force" in March 1861.

In his 1864 paper "A dynamical theory of the electromagnetic field", Maxwell wrote, "The agreement of the results seems to show that light and magnetism are affections of the same substance, and that light is an electromagnetic disturbance propagated through the field according to electromagnetic laws

Maxwell showed that the equations predict the existence of waves of oscillating electric and magnetic fields that travel through empty space at a speed of 310,740,000m/s.

His famous equations, in their modern form of four partial differential equations, first appeared in fully developed form in his textbook A Treatise on Electricity and Magnetism in 1873.

The specific features of Faraday's field concept, in its 'favourite' and most complete form, are that force is a substance, that it is the only substance and that all forces are interconvertible through various motions of the lines of force. These features of Faraday's 'favourite notion' were not carried on by Maxwell in his approach to the problem of finding a mathematical representation for the continuous transmission of electric and magnetic forces. Maxwell considered these electric and magnetic forces to be states of stress and strain in a mechanical aether, a notion further advanced by relativity theory with its 'stress energy' tensor math.

Tetryonics reveals lines of Force to be a direct result of the various superpositioned EM field geometries of equilateral mass-energy momenta

Copyright ABRAHAM [2008] - all rights reserved

Anti-Parallel Magnetic Dipoles

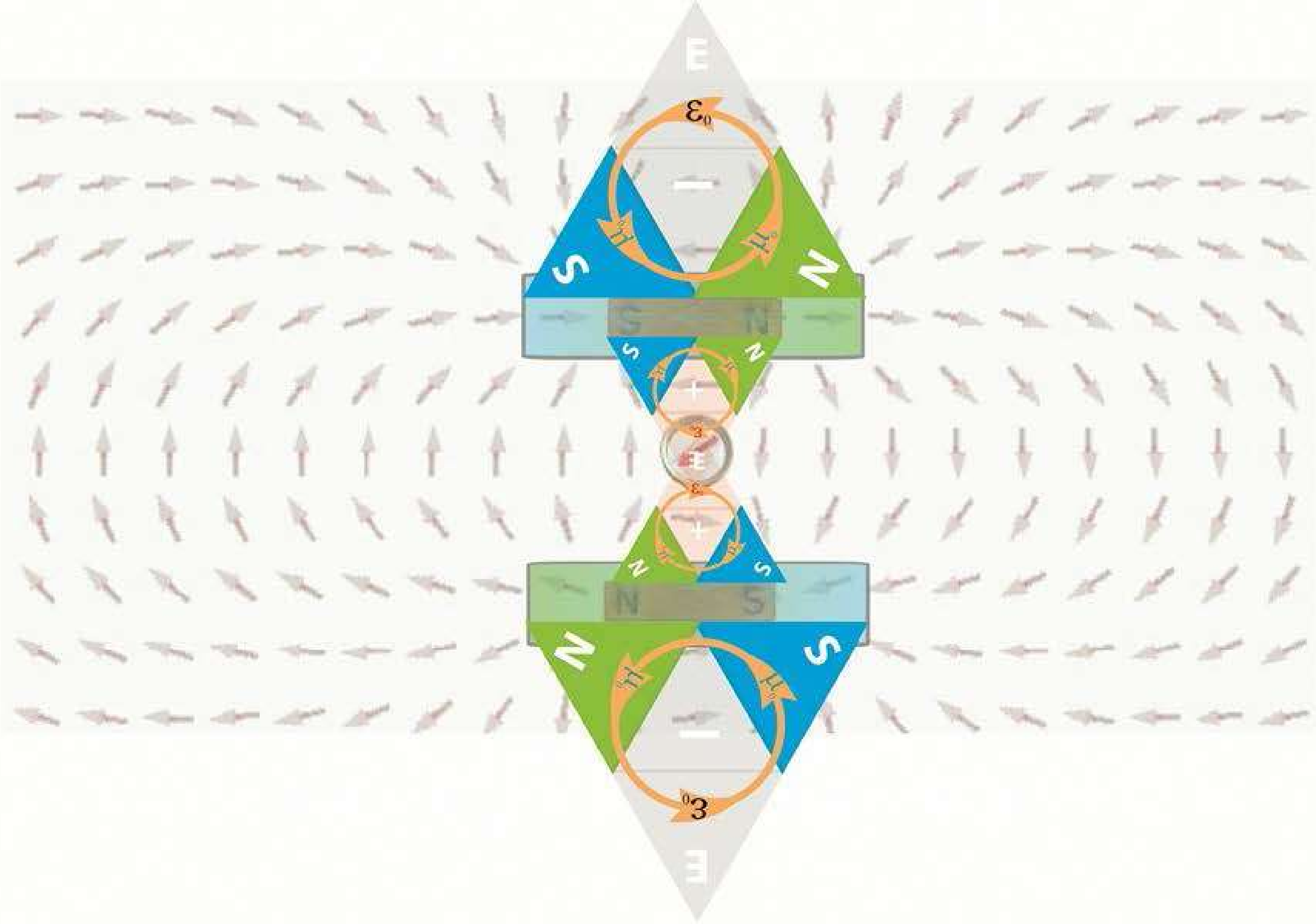
E
ε₀
μ₀
S
N
−
+
S
N
N
S
+
−
ε₀
E

Tetryonics 04.17 - Anti-Parallel Magnetic dipoles

Copyright ABRAHAM [2008] - all rights reserved

Magnetic Moments

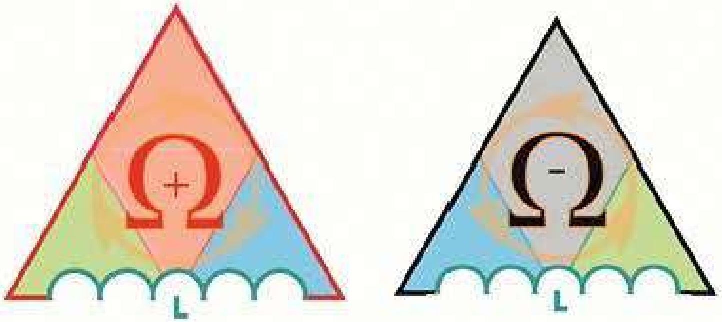

Single ZPFs are
'ideal quantum inductor elements'

1

Electro-static Energies
Charges

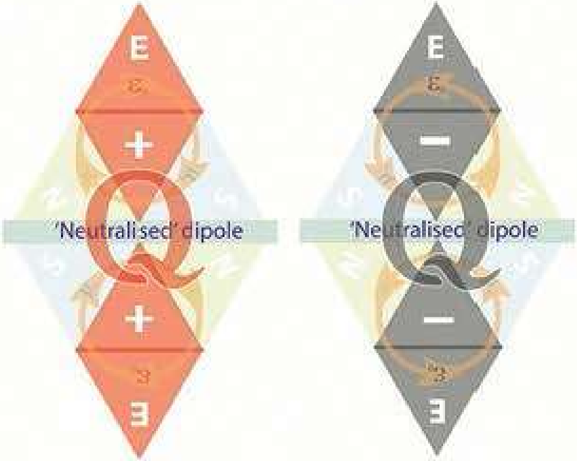
E
ε
+
N
S
'Neutralised' dipole
+
E
E
ε
–
S
N
'Neutralised' dipole
–
E

All Matter in motion possesses kinetic energies
which are stored as Planck quanta in their KEM fields

The charge geometry of KEM fields are reflective
of the interactive component of the charged topology
of the particle in motion

Each charge geometry
has distinct Magnetic dipole
alignments

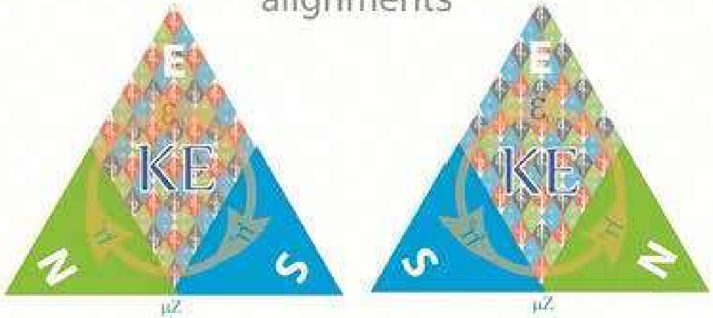
E
ε
KE
N
S
μZ
E
ε
KE
S
N
μZ

which in turn can only create
2 distinct orientations [spins] of
magnetic moments

2

Magneto-static Energies
Kinetic Energies

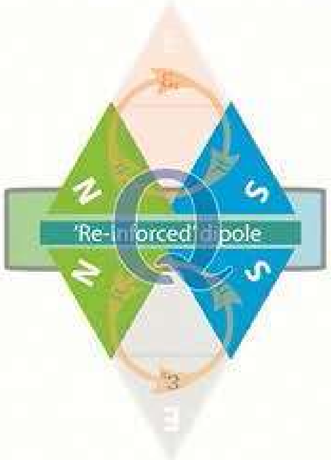
N
S
'Re-inforced' dipole
N
S

ZPF sets can form inductively coupled
quantum Harmonic Oscillators

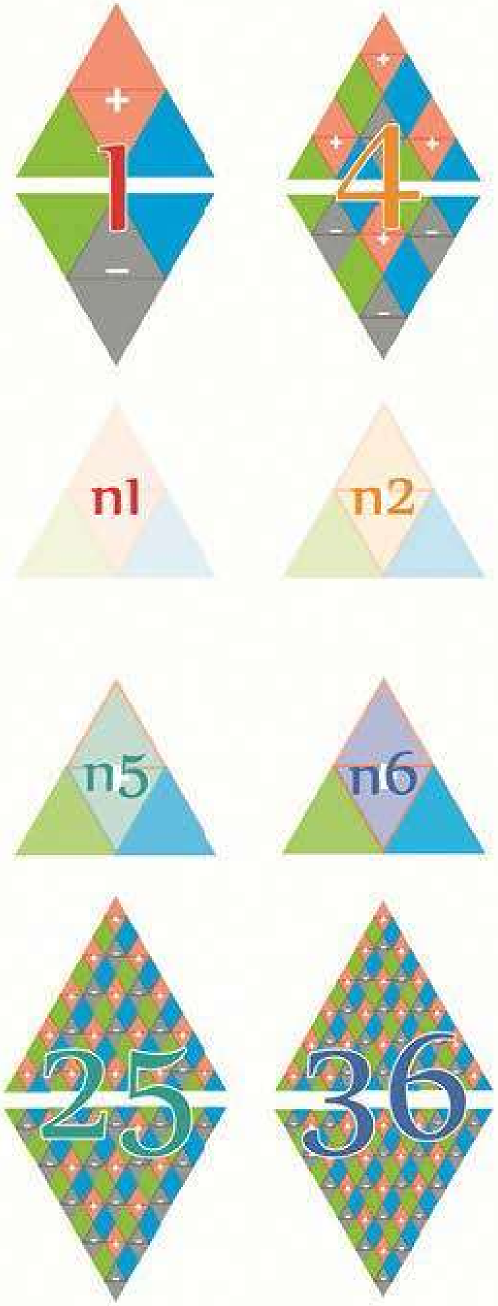
1
+
−
4
n1
n2
n5
n6
25
36

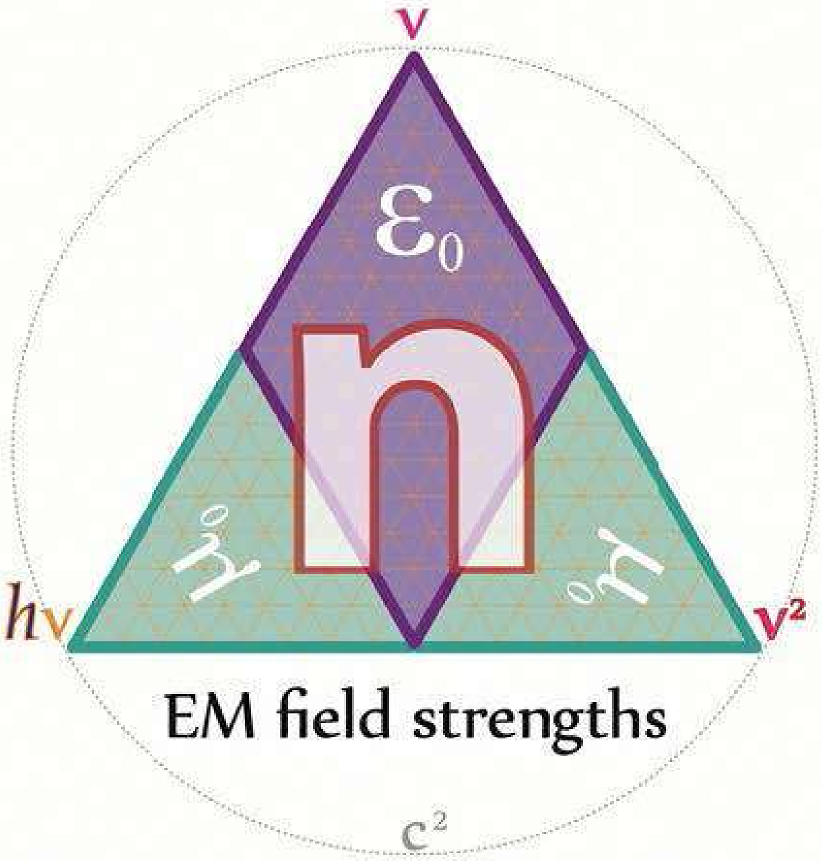
v
ε_0
n
μ_0
μ_0
hv
v²
c²

EM field strengths

All Energies are 'square' quanta scalar fields made up of quantised Electric & Magnetic fields

All EM fields have the same equilateral geometries [v^2]
Electric fields are inverse squared fields [1/r^2]
Magnetic dipoles are inverse cubed fields [1/r^3]

As the mass-energy quanta in EM fields change their geometries are subject to Lorentz contractions

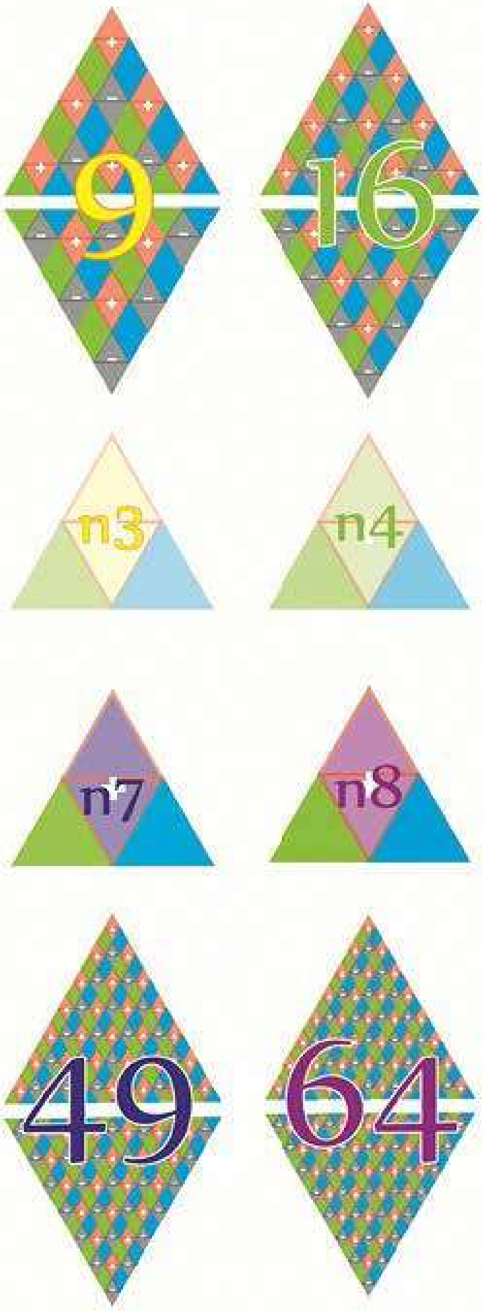
9
16
n3
n4
n7
n8
49
64

Electrostatic particle modeling

m

EM mass-energy
Tetryonic field geometry

Tetryonic [4πp] standing wave charge fields
form electrostatic Particle topologies
[Charged and Neutral Matter]

Matter
tetryonic Matter topology

M

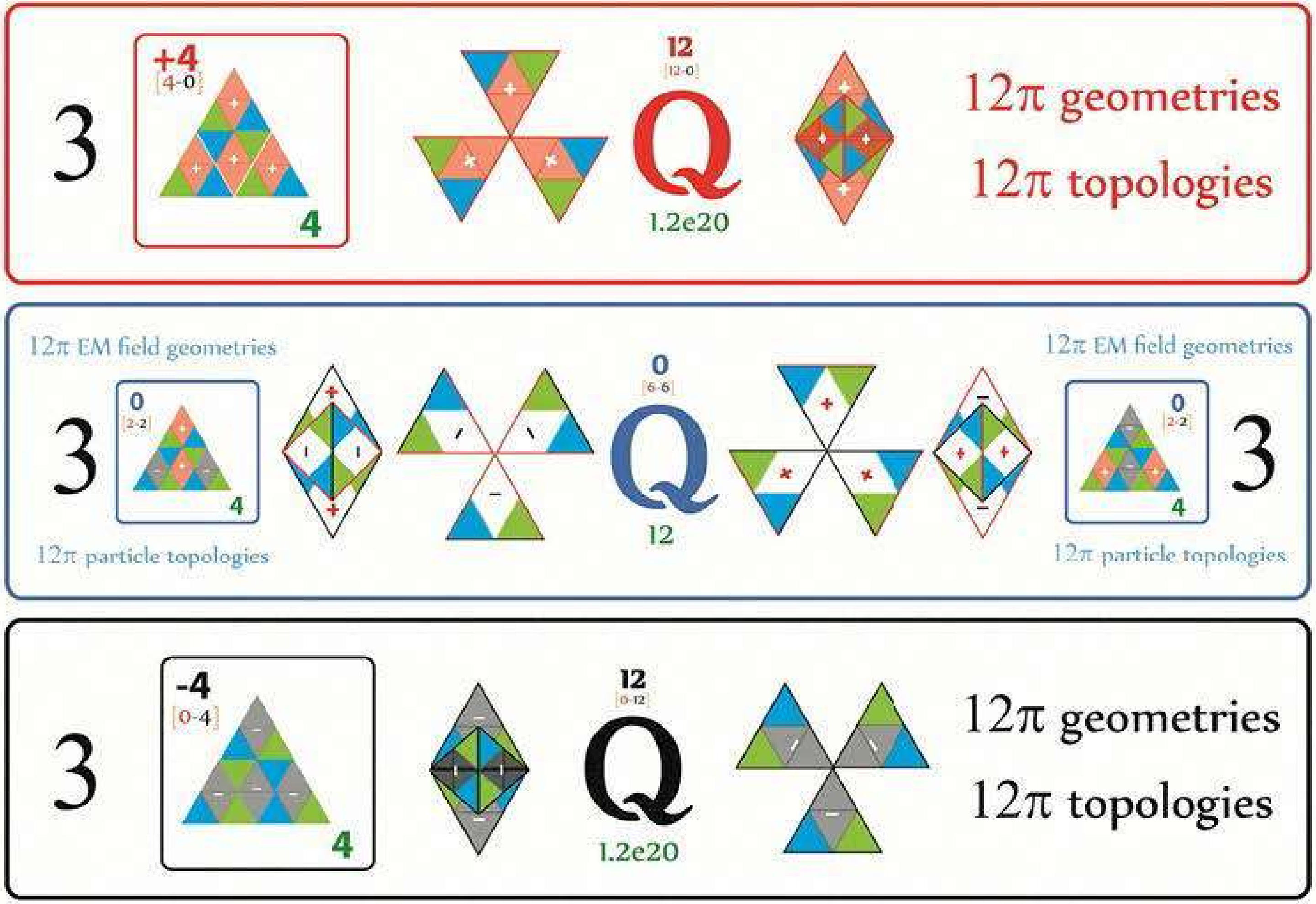
3
+4
[4-0]
4
12
[12-0]
Q
1.2e20
12π geometries
12π topologies
12π EM field geometries
3
0
[2-2]
4
0
[6-6]
Q
12
12π particle topologies
12π EM field geometries
0
[2-2]
4
3
12π particle topologies
3
-4
[0-4]
4
12
[0-12]
Q
1.2e20
12π geometries
12π topologies

Non-neutral
nett
Tetryonic quanta

Positive
Charge
Particles

Equal numbers
of opposite
Tetryonic quanta

Neutral
Charge
Particles

Non-neutral
nett
Tetryonic quanta

Negative
Charge
Particles

All particles in motion create secondary KEM field geometries
[Kinetic energies and Magnetic moments]

Tetryonics 04.20 - Electrostatic Particle modelling

Electro-static Fields

Kinetic EM Fields

Momenta
acting in oppose directions
results in zero velocity

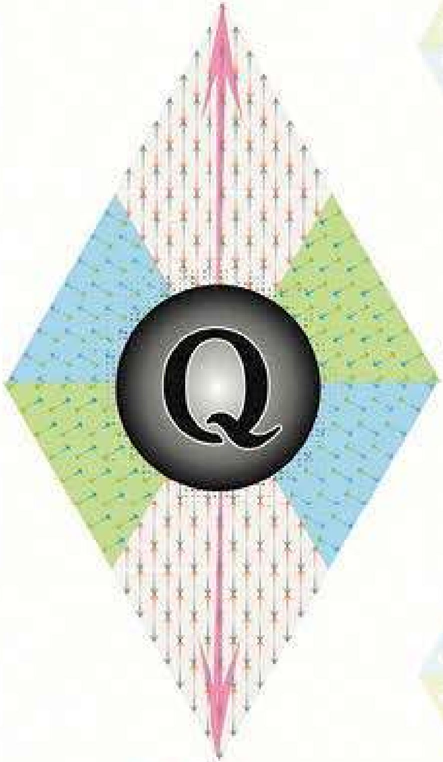

Zero Velocity
equates to
Zero nett Momentum

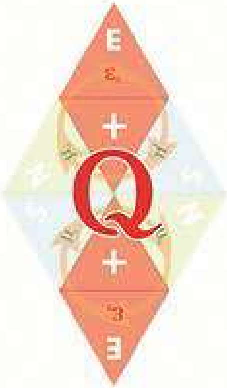

Stationary Charges have neutralised magnetic dipoles

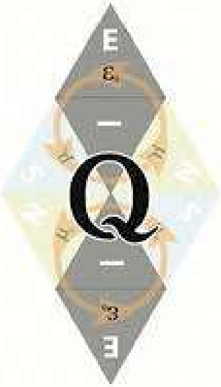

The Kinetic Electric & Magnetic energies are contained in an KEM field extending from a charged particle in motion

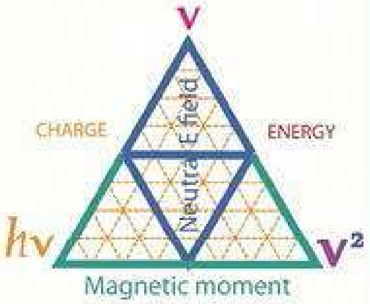
v
CHARGE
ENERGY
Neutral E field
hv
v²
Magnetic moment

Charged particles in motion produce a Magnetic moment

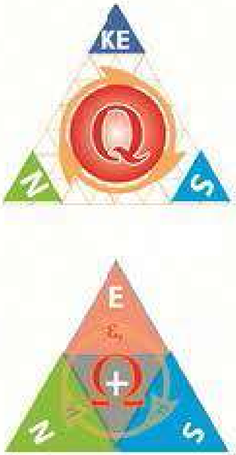
KE
Q
N
S
E
ε₀
+
N
S

Moving Charges have neutral KE fields and Magnetic moments

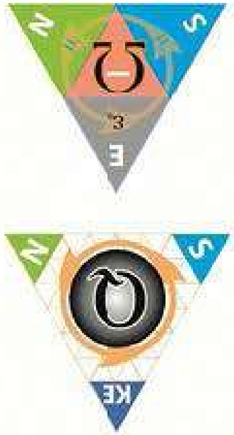
N
S
–
ε₀
E
N
S
Q
KE

Motion in any direction
produces
Kinetic & Magnetic energies

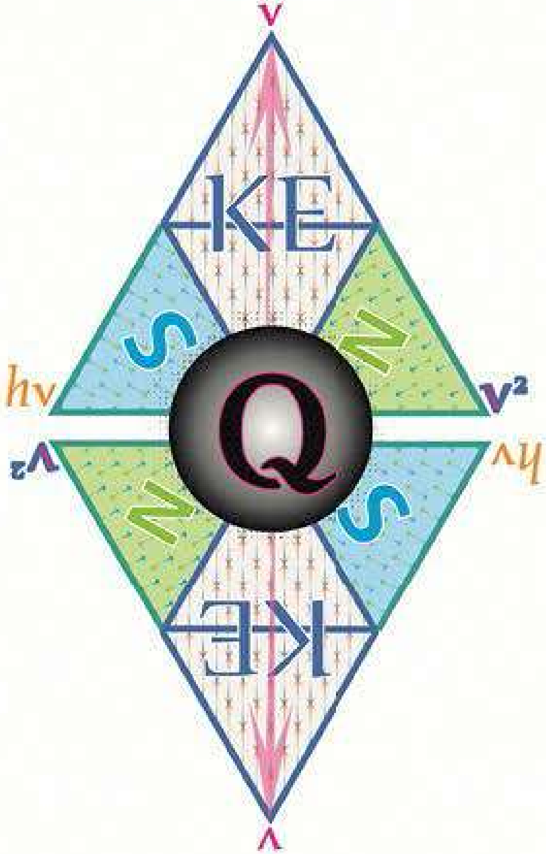
v
KE
S
N
hv
v²
Q
v²
hv
N
S
KE
v

The Strength of the
Kinetic Electric field & Magnetic moment
is directly proportional to
the square of the particle's Velocity

Electric & Kinetic Fields

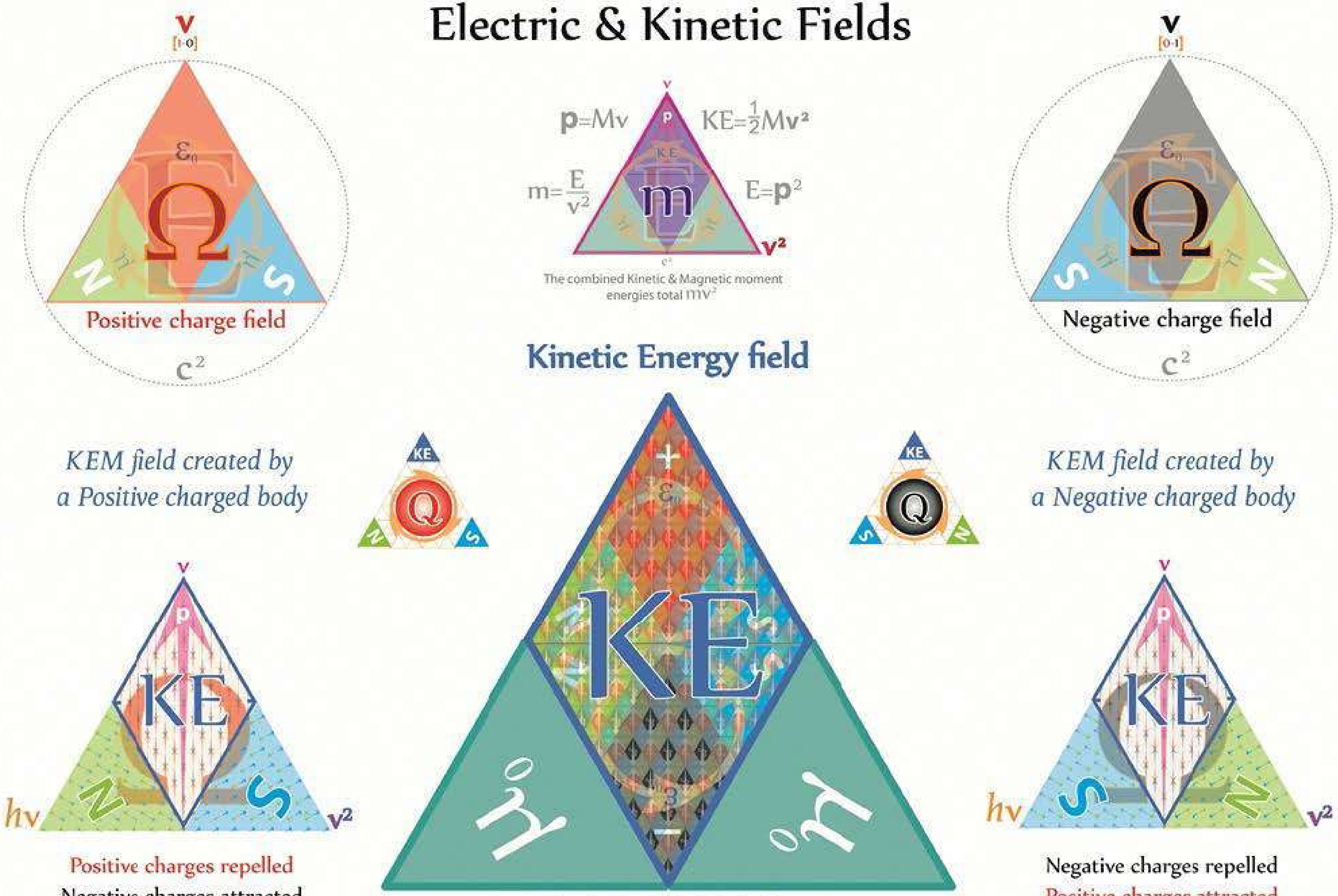
v
[1-0]
ε_0
Ω
N
S
c²
v
[0-1]
ε_0
Ω
S
N
c²
v
p
KE
m
v^2
c^2
KE
Q
N
S
KE
Q
S
N
v
p
KE
hv
N
S
v²
+
ε_0
KE
μ_0
μ_0
−
v
p
KE
hv
S
N
v²

Positive charge field

Negative charge field

$p = Mv$ $KE = \frac{1}{2}Mv^2$

$m = \frac{E}{v^2}$ $E = p^2$

The combined Kinetic & Magnetic moment energies total mv^2

Kinetic Energy field

KEM field created by a Positive charged body

KEM field created by a Negative charged body

Positive charges repelled
Negative charges attracted

Neutralised Electric fields
Enhanced Magnetic moments

Negative charges repelled
Positive charges attracted

Copyright ABRAHAM [2008] - all rights reserved

EM & KEM force vectors

All mass-energy quanta are ideal quantum Inductors

The E&M force vectors create orthogonal equilateral EM Fields

Electric fields propogate orthogonally to the Magnetic dipole field

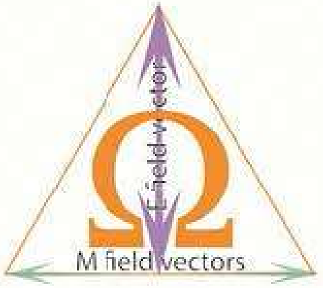

Magnetic flux external to the Magnetic dipole flows from North to South
&
Internal to a Magnetic dipole it flows South to North

When ZPFs combine to form a Magnetostatic dipole they form orthogonal magnetic vectors

Positive charge can be modelled electrically with clockwise EM fluxes [omegas]

Negative charge can be modelled electrically with anti-clockwise EM fluxes [omegas]

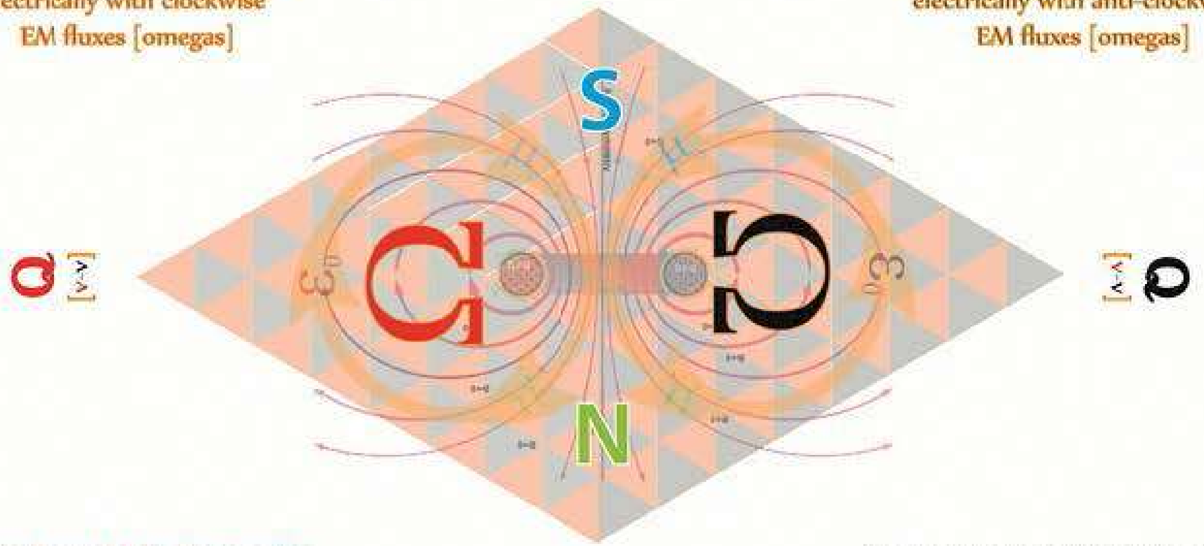

Positive charge fields of interaction result from an excess of divergent positive energy momenta

Equal energy-momenta in opposition create static fields

Negative charge fields of interaction result from an excess of divergent negative energy momenta

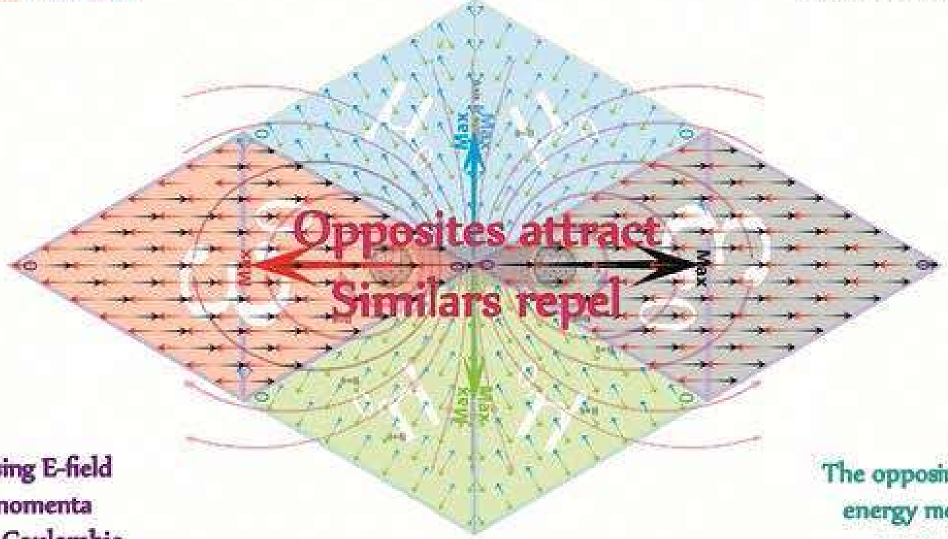

The opposing E-field energy momenta creates the Coulombic Law of Interaction

When E field is at Maximum - B field is at Minimum
when B field is at Maximum - E field is at Minimum

The opposing M-field energy momenta creates the Lorentz Force

Copyright ABRAHAM [2008] - all rights reserved

EM Forces and ZPFs

ZPF quanta can combine in differing combinations
to produce 3 distinct charged sets

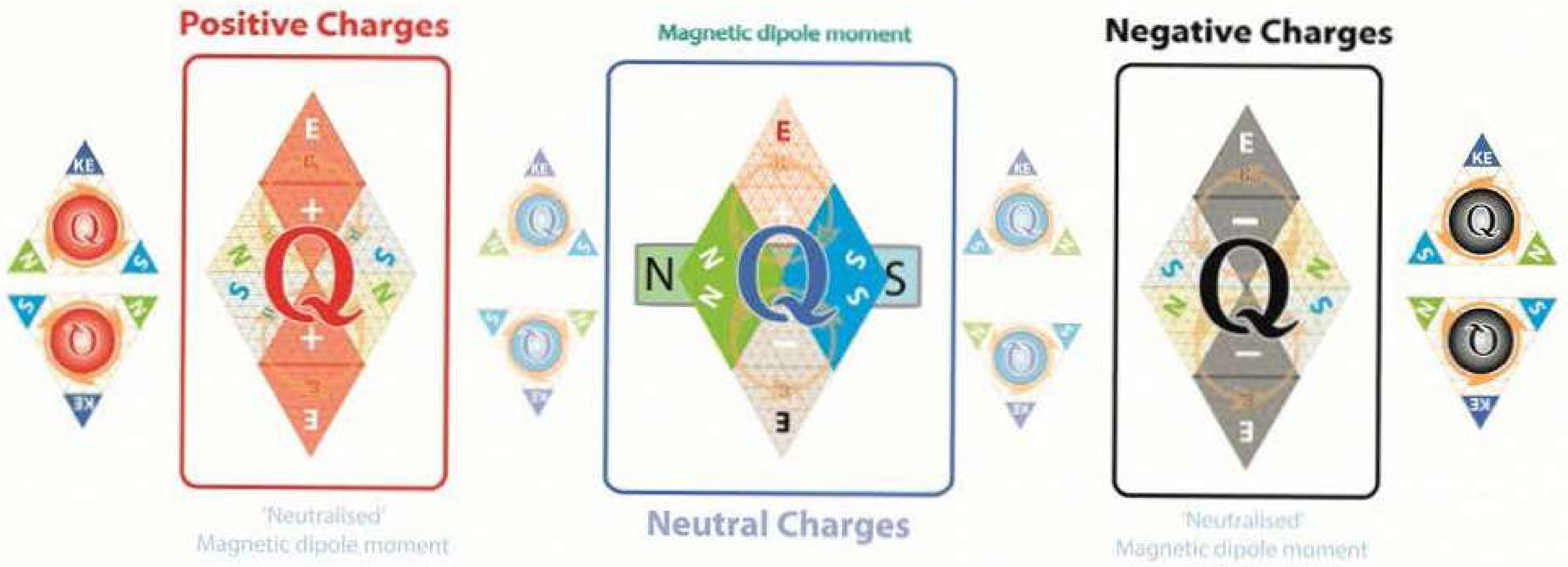

Positive Charges

Magnetic dipole moment

Negative Charges

'Neutralised'
Magnetic dipole moment

Neutral Charges

'Neutralised'
Magnetic dipole moment

All EM mass-Energy-Matter & forces can be modelled using Tetryonic geometries

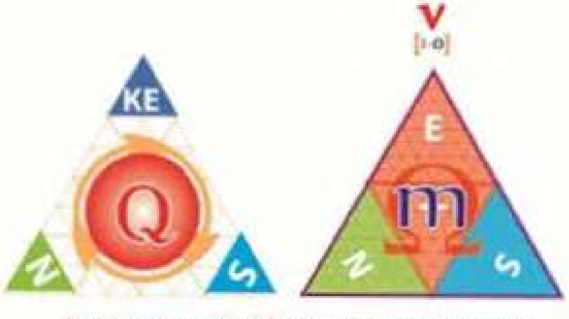

Positive charge particle KE field and magnetic moment

Zero Point Field EM geometry is the foundation for all the EM forces comprising, and acting between particles of Matter

Lorentz force, Lenz's Law, Right/Left Hand rules, etc can all be easily replaced with this simple geometric model

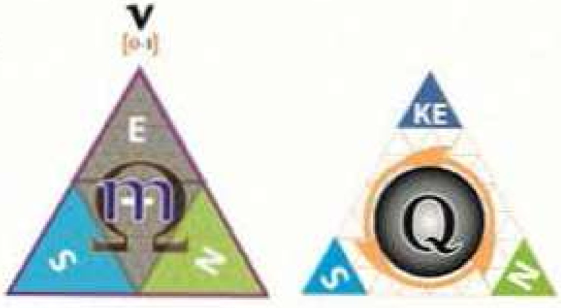

Negative charge particle KE field and magnetic moment

Kinetic EM fields

All Kinetic EM fields resulting from motion have charge field geometries resulting from the charged Matter geometry of the particle in motion

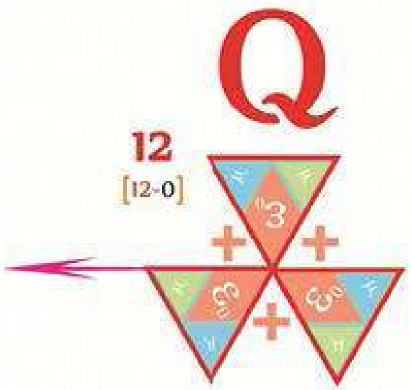

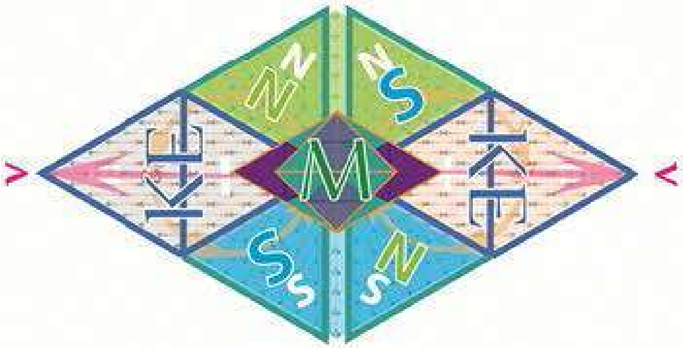

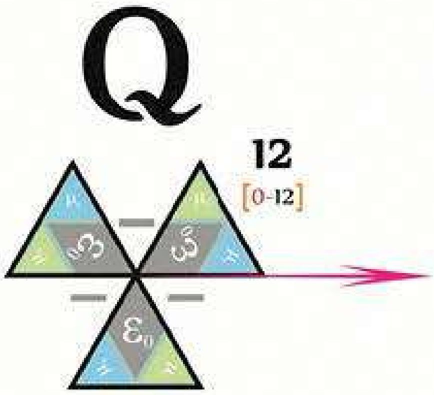

Positive charge particle topologies produce positive KEM fields

KEM field charge geometries do not contribute to the nett charge

Negative charge particle topologies produce negative KEM fields

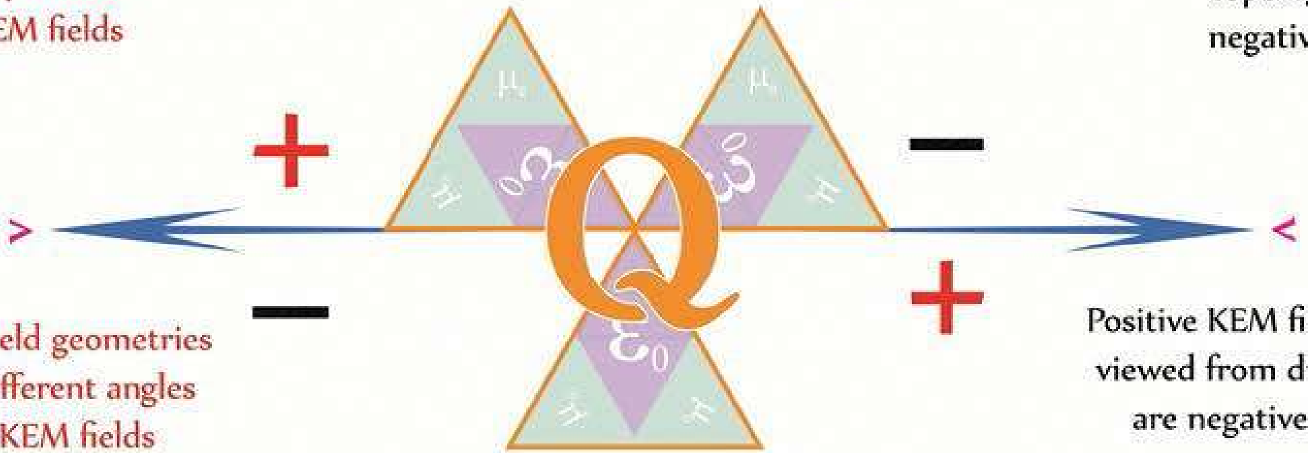

Negative KEM field geometries viewed from different angles are positive KEM fields

Positive KEM field geometries viewed from different angles are negative KEM fields

S
KE
N
Q
Magnetic moment

As a direct result of the KEM field being a EM field permeating free space the symmetry of EM fields results in KEM field geometries being viewed as having neutral KE fields with a magnetic moment

S
KE
N
Q
Magnetic moment

Copyright ABRAHAM [2008] - all rights reserved

Point Particles and KEM fields

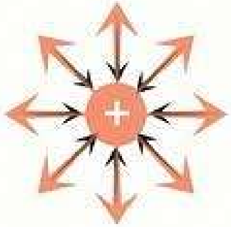

e+

12

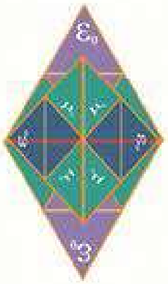

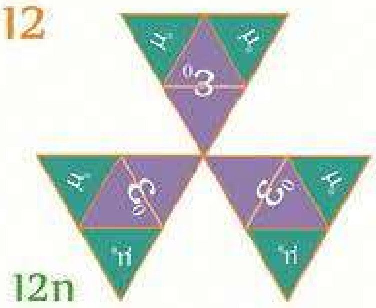
12
ε0
μ0
μ0
ε0
ε0
μ0
μ0
12n

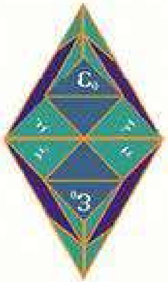

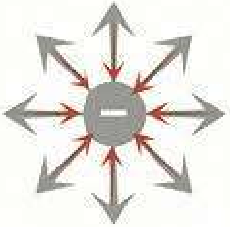

e-

12

Charged rest mass-Matter topology

Quarks T[8π]
Leptons T[12π]

velocity invariant rest mass-Matter has a standing wave topology

Baryons T[20π]

Charged Leptons at rest are Electric field standing waves
(with neutral Magnetic poles)
KE from motion generates a Magnetic Moment

Bosons nπ [ODD]

Kinetic EM field geometry is divergent from a particle's rest Matter topology

Photons nπ [EVEN]

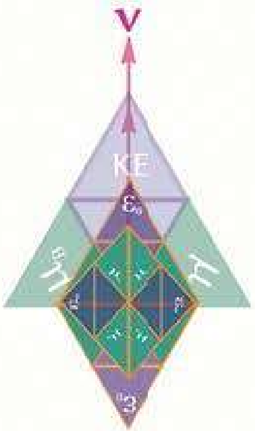
v
KE
ε0
μ0
μ0
ε0

Charge

Q

[v-v]

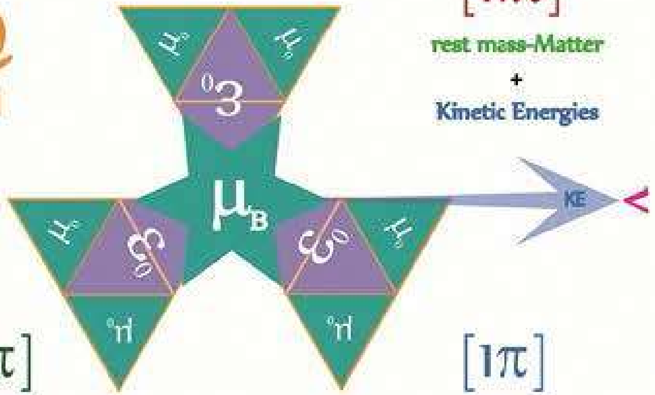
μ0
μ0
ε0
μB
μ0
ε0
μ0
ε0
μ0
μ0
KE
v

[nπ]

rest mass-Matter
+
Kinetic Energies

[12π]

rest Matter topology

[1π]

Kinetic Energy geometry

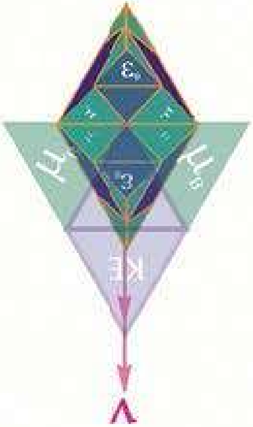
ε0
μ0
μ0
ε0
KE
v

Copyright ABRAHAM [2008] - all rights reserved

KEM fields of Matter in motion

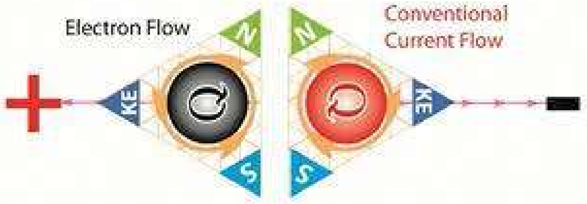
Electron Flow
Conventional Current Flow
N
N
KE
KE
S
S
+
-

ZPFs geometries can be used to model the KEM fields of charged particles

2D charged EM field geometries create Matter topologies

tetryons	[4π]	tetryons	[4π]
quarks	[12π]	quarks	[8π]
leptons	[12π]	leptons	[12π]
Baryons	[36π]	Baryons	[20π]
Elements	[84π]	Elements	[54π]

The mass-energy content of all charged fascia constituting massive particles have momenta that is proportional to the intrinsic velocity [c] of the standing wave

rest Matter

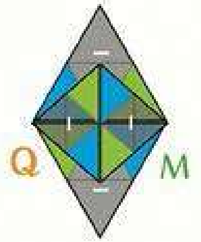

Electro-static particles have neutralised magnetic dipoles

Relativistic mass

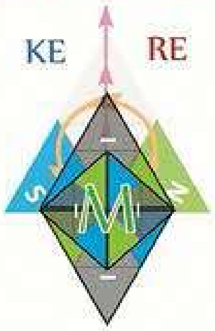

Kinetic motion produces Magnetic moments

Spherical point charges do NOT exist

positron

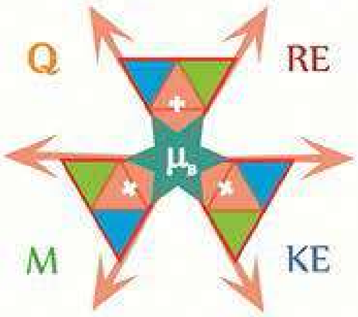

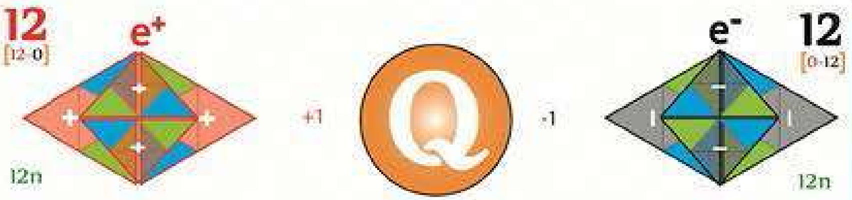
12
[12-0]
e+
12n
+1
Q
-1
e-
12
[0-12]
12n

3D Matter is a standing-wave topology resulting from radiant 2D EM geometries
All leptons and quarks both have 12 charged fasica geometries.
[but differing mass-Matter-particle topologies]

electron

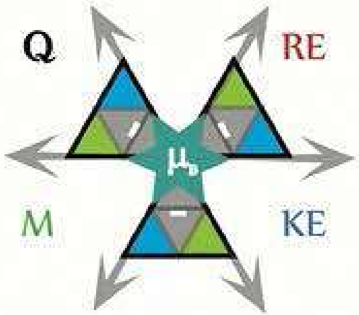

All EM fields resulting from Kinetic Energy (motion) radiate outwards
(the intrinsic KEM fields contain both Negative and Positive Energy momenta quanta)

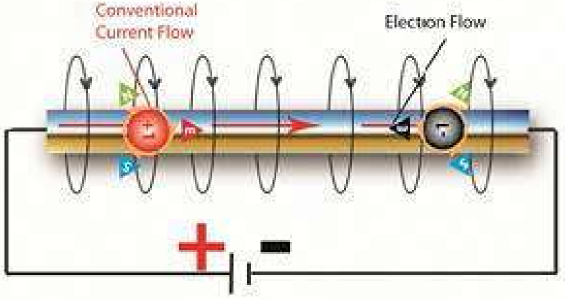
Conventional Current Flow
Electron Flow
+
-

Magnetic field around a current carrying conductor

Magnetic field produced by a Solenoid

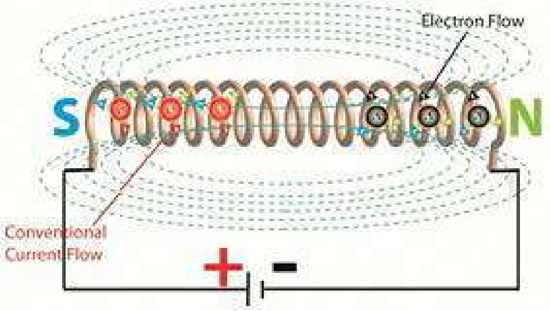
Electron Flow
S
N
Conventional Current Flow
+
-

μ_0 There is an inverse cube relationship between magnetic field strength and magnetic field force with respect to distance from the magnet. ε_0

Biot-Savart Law

is an equation in electromagnetism that describes the magnetic field B generated by an electric current.

The interaction between Magnetostatic fields for which a Magentic dipole exists obeys the inverse CUBED law of attraction and repulsion

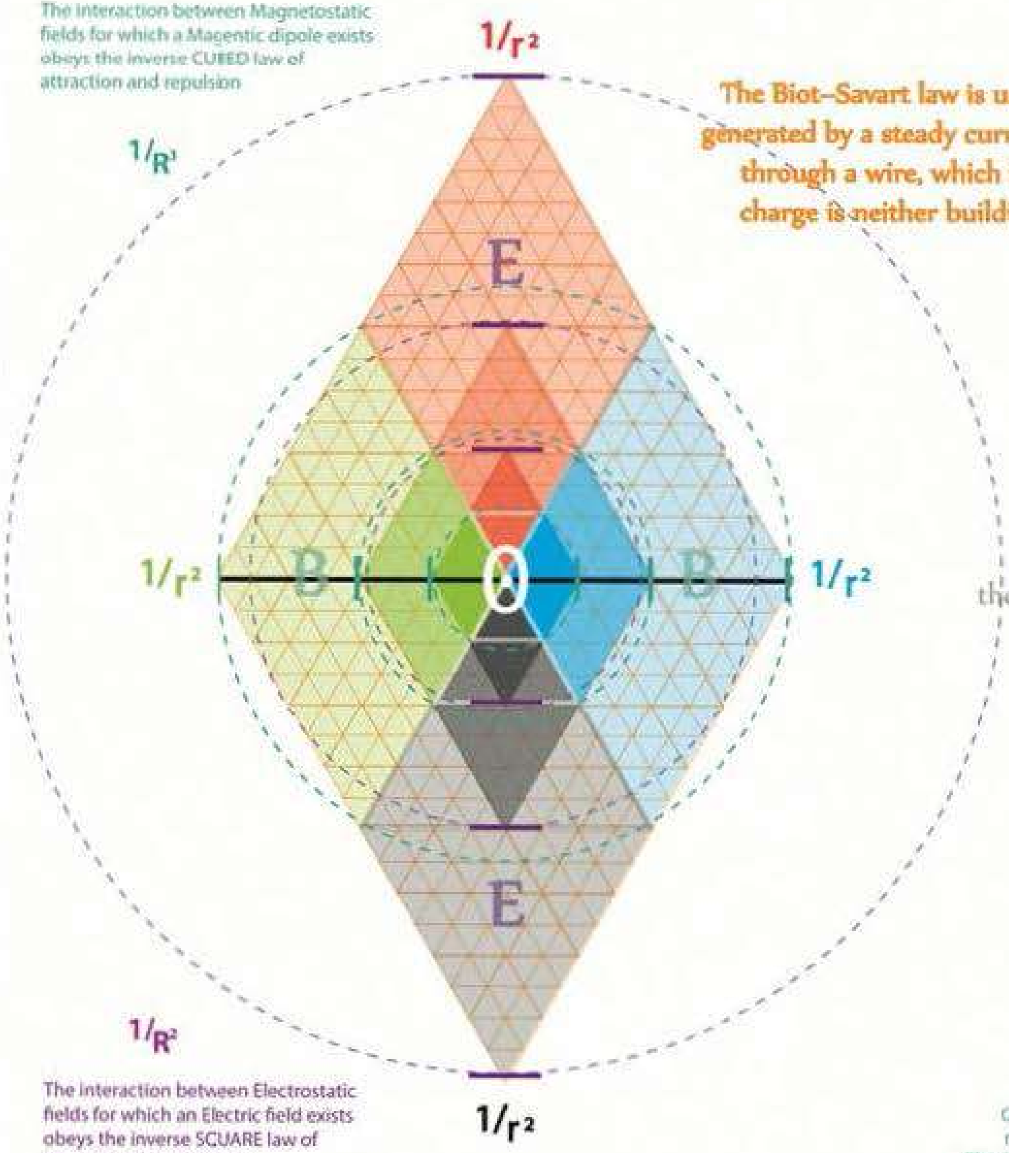

The interaction between Electrostatic fields for which an Electric field exists obeys the inverse SQUARE law of attraction and repulsion

The Biot–Savart law is used to compute the magnetic field generated by a steady current, i.e. a continual flow of charges, through a wire, which is constant in time and in which charge is neither building up nor depleting at any point

$$\mathbf{B} = \int \frac{\mu_0}{4\pi} \frac{I d\mathbf{l} \times \mathbf{r}}{|r|^3},$$

The radial distance between Magnetic dipoles is less than the distance between Electric dipoles

$$\mathbf{B} = \int \frac{\mu_0}{4\pi} \frac{I d\mathbf{l} \times \hat{\mathbf{r}}}{|r|^2},$$

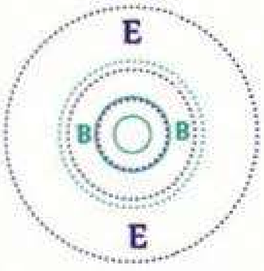

Close to one pole of a magnet, B field strength resembles the inverse square of Electric force. This is because it behaves as a "unipolar magnetic field"

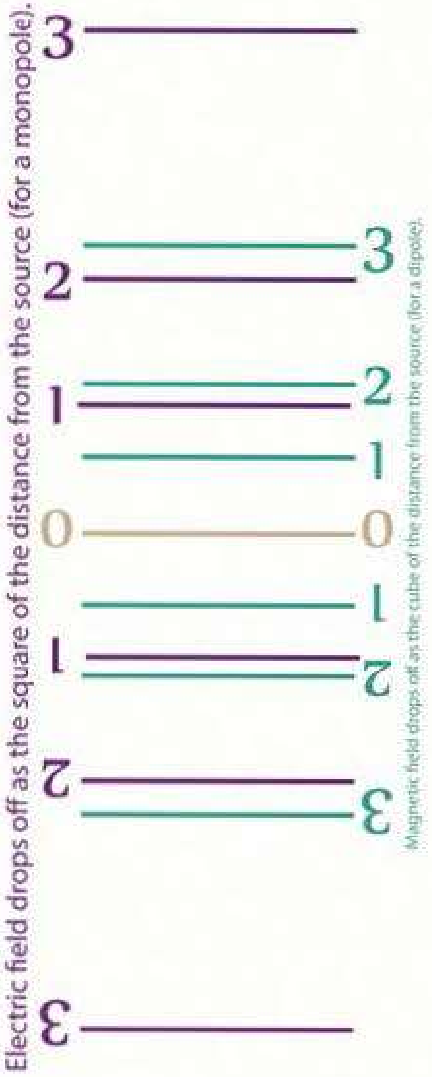

Electric field drops off as the square of the distance from the source (for a monopole).

Magnetic field drops off as the cube of the distance from the source (for a dipole).

Quantum Chirality

Its the equilateral geometry of quantised angular momenta that creates chirality in physics

*Mirror imaged Planck quanta
are NOT
identical to each other*

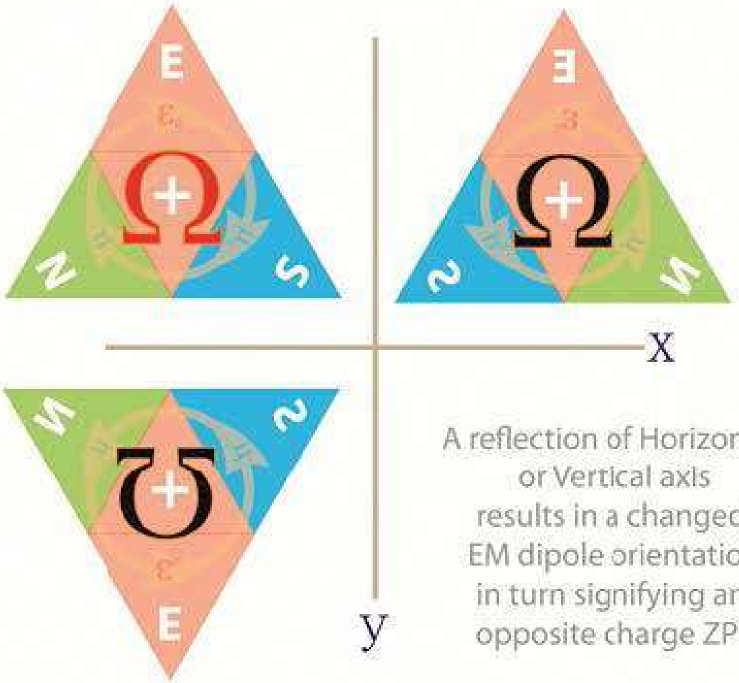

A reflection of Horizontal
or Vertical axis
results in a changed
EM dipole orientation
in turn signifying an
opposite charge ZPF

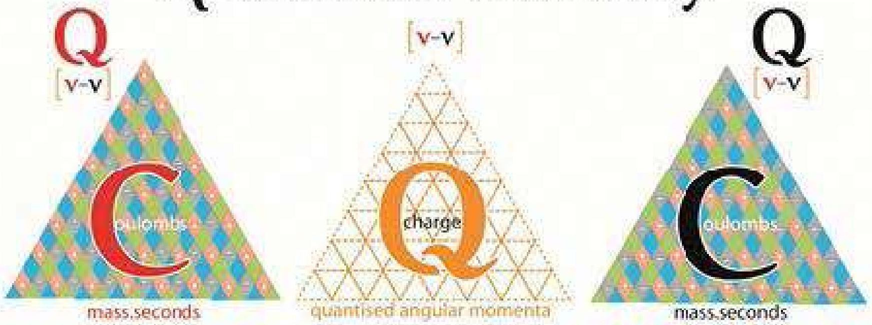
Q
[v-v]
C
coulombs
mass.seconds
[v-v]
Q
charge
quantised angular momenta
Q
[v-v]
C
coulombs
mass.seconds

*The equilateral geometry of any EM field or Matter particle
is determined by its nett Coulombic charge*

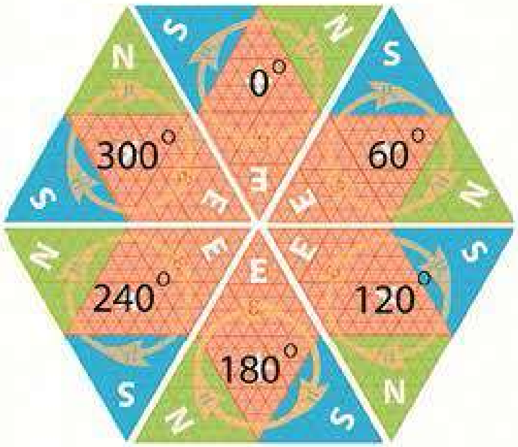
0°
60°
120°
180°
240°
300°
N
S
E

Any nominal rotation about an axis
results in a re-orientation of the
electromagnetic vectors but does not
affect any change to charge etc.

Irrespective of orientation or rotation:

Ω Positive charge fields have clockwise inductive flux geometries

Negative charge fields have counter-clockwise inductive flux geometries Ω

Pi radian - energy momenta geometries

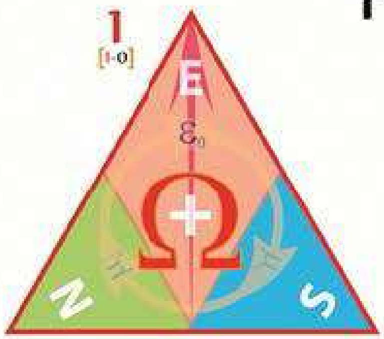
1
[1-0]
E
ε_0
Ω
+
N
S

mass-energy momenta

Q

nett charge

[v-v]

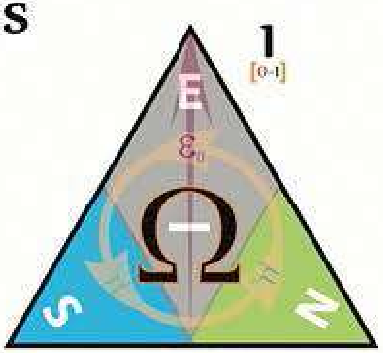
1
[0-1]
E
ε_0
Ω
−
S
N

π

pi radian energy geometries

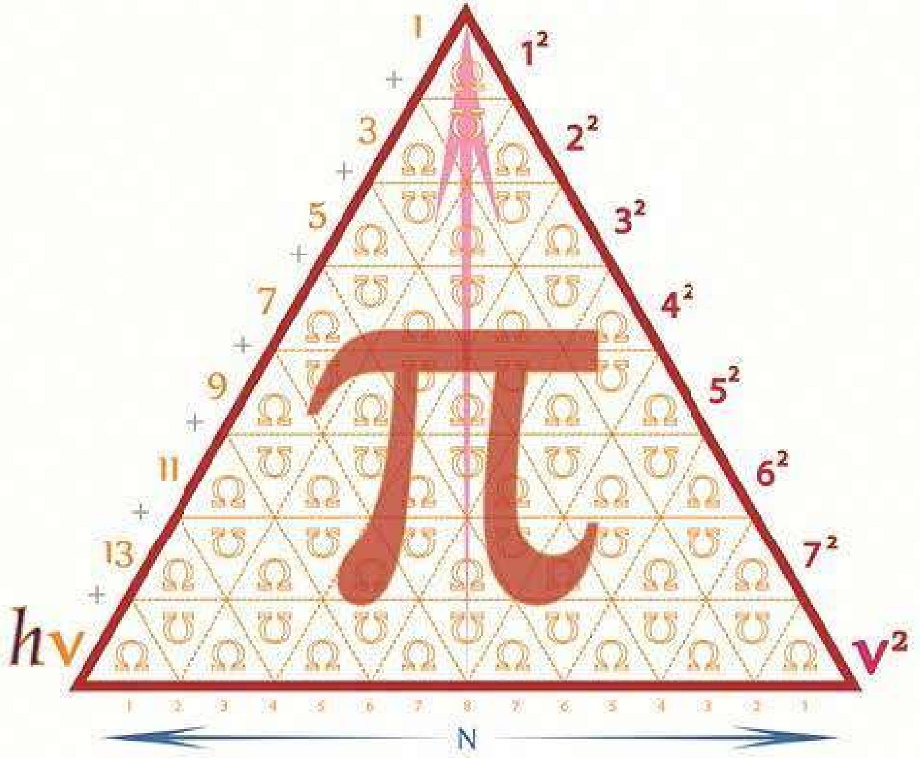
1
+
3
+
5
+
7
+
9
+
11
+
13
+
hv
1^2
2^2
3^2
4^2
5^2
6^2
7^2
v^2
π
Ω
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
N

Ω

quantised angular momentum

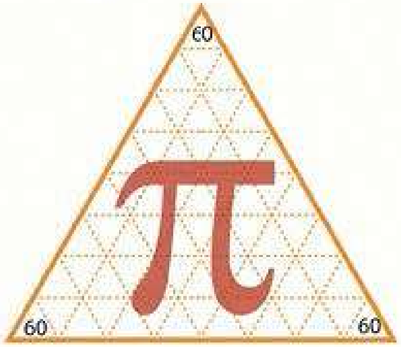

Planck quanta

$$n\pi[m\Omega v]^2$$

energy mass velocity

π radian geometries are comprised of equilateral mass momenta Ω

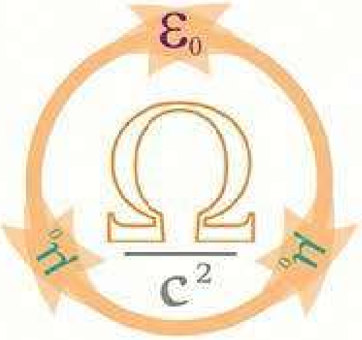
ε_0
Ω
c^2
μ_0
μ_0

Copyright ABRAHAM [2008] - all rights reserved

ZPFs and Bosons

ZPFs are Planck energy quantum elements

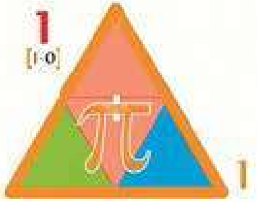

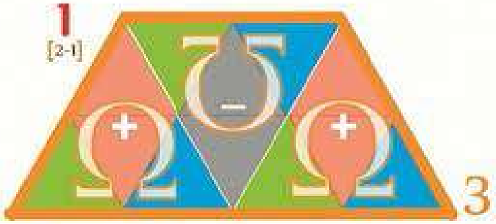

Bosons are transverse Quantum levels

Quantised angular momentum is the foundation for all Tetryonic mass-ENERGY-Matter geometries

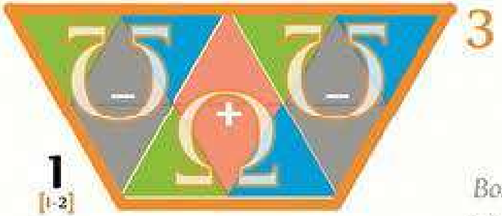

Bosons are the exchange particles for the EM force

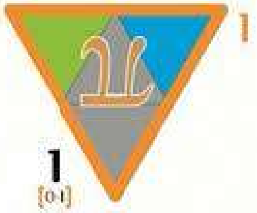

ZPFs are the quanta geometry for Charge, Bosons and Energy

Each Quantum level is an ODD number of quanta (Bosons)

Each 'squared' energy field is the sum of the preceeding Bosons

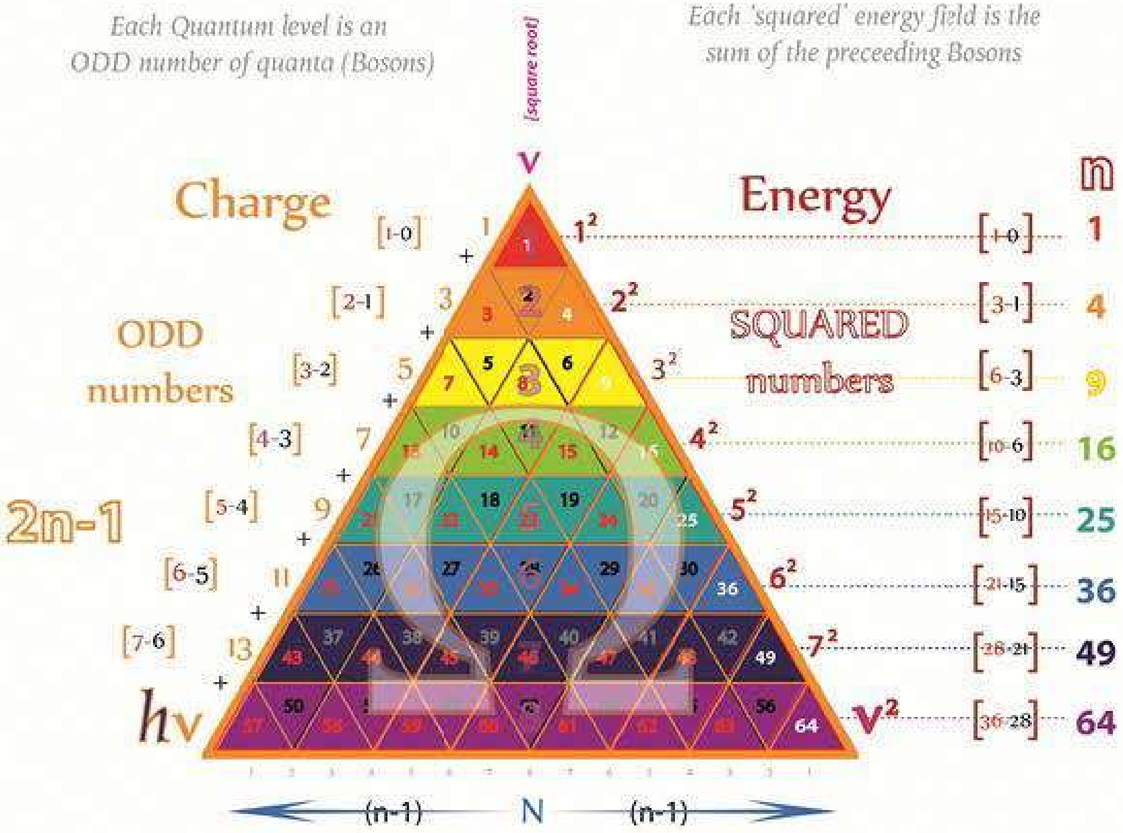
[square root]
v
Charge
Energy
n
ODD numbers
SQUARED numbers
2n-1
[1-0] 1
[2-1] 3
[3-2] 5
[4-3] 7
[5-4] 9
[6-5] 11
[7-6] 13
hv
1^2
2^2
3^2
4^2
5^2
6^2
7^2
v^2
[1-0] 1
[3-1] 4
[6-3] 9
[10-6] 16
[15-10] 25
[21-15] 36
[28-21] 49
[36-28] 64
(n-1)
N
(n-1)

The charged quanta in all mass-Energy geometries create a NORMAL Distributions

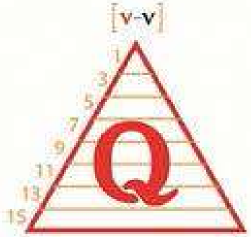
[v-v]
1
3
5
7
9
11
13
15
Q

EM Field Planck quanta

ODD $\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Bosons ElectroMagnetic mass velocity

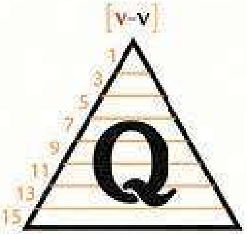
[v-v]
1
3
5
7
9
11
13
15
Q

Bosons form the geometry of Quantum Levels

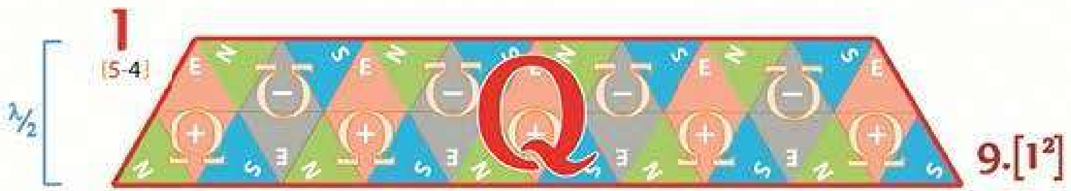

$$W^{+} = \underset{\text{ODD}}{n\pi}\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Force carrier for Positive charge particles

W^{+} Boson

EM field Planck quanta

$$\underset{\text{Bosons}}{\text{ODD}\,\pi}\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

ElectroMagnetic mass velocity

CHARGE carrier Bosons
are ODD number quanta

Bosons

form unit charge Quantum levels
that facilitate EM induction between
mass-Energy-Matter

Neutral Z Bosons and Photons
are EVEN quanta Bosons

EMField Planck quanta

$$\underset{\text{Photons}}{\text{EVEN}\,\pi}\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

ElectroMagnetic mass velocity

Neutral charge parallelogram geometry EM force carriers

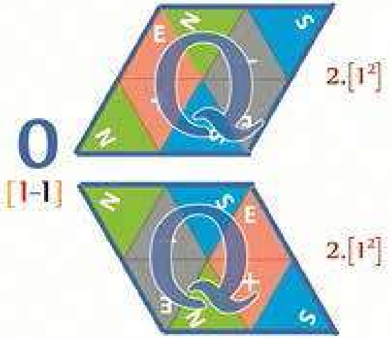

Z^{0} Boson

*Neutral Z Bosons
can be formed by
combining EVEN
numbers of W Bosons*

**They are EVEN numbered
alternate geometry Photons**

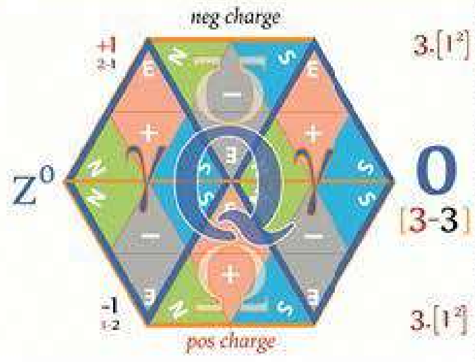

Photons are alternate (diamond) EM geometries formed from Z Bosons

W^{-} Boson

Force carrier for Negative charge particles

$$W^{-} = \underset{\text{ODD}}{n\pi}\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

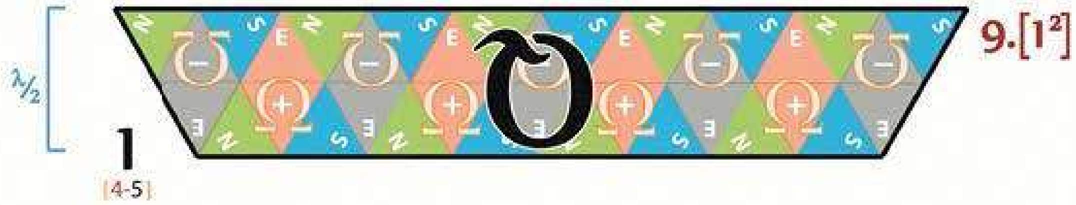

Bosons are charge carriers

[the geometric foundation of quantum levels]

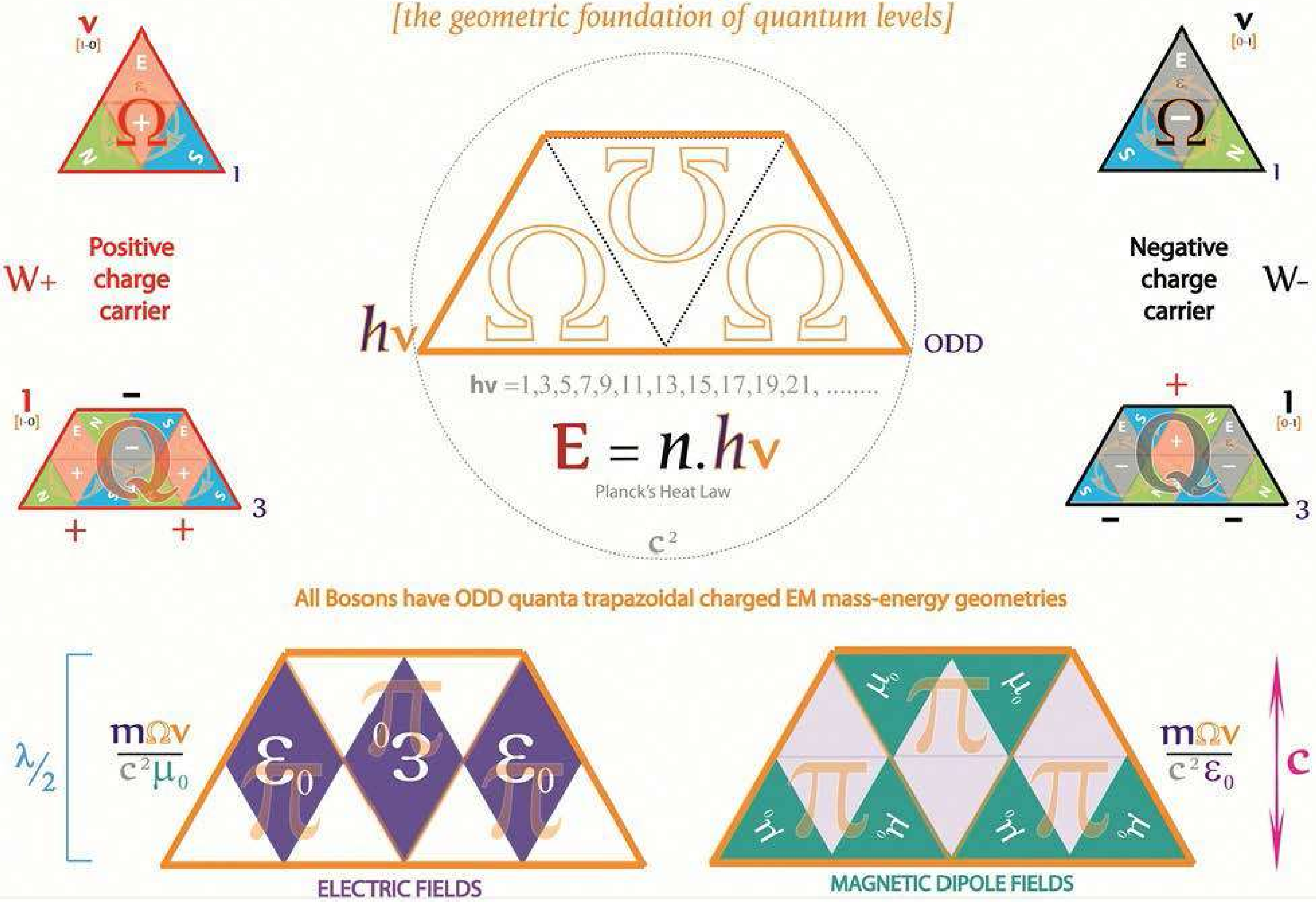
v
[1-0]
E
Ω
+
N
S
1
W+
hv
ODD
v
[0-1]
E
Ω
−
S
N
1
W-
1
[1-0]
−
+
+
3
+
1
[0-1]
−
−
3
c²
λ/2
$\frac{m\Omega v}{c^2\mu_0}$
ε_0
π
3
μ_0
$\frac{m\Omega v}{c^2\varepsilon_0}$
c

Positive charge carrier

Negative charge carrier

hv = 1,3,5,7,9,11,13,15,17,19,21, ...

$$E = n.hv$$

Planck's Heat Law

All Bosons have ODD quanta trapazoidal charged EM mass-energy geometries

ELECTRIC FIELDS

MAGNETIC DIPOLE FIELDS

Boson Frequencies

All Bosons are 1/2 wavelength EM fields with ODD number quanta
They are the geometric basis for transverse EM field Quantum levels

EM Field Planck quanta

$\text{ODD}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Bosons ElectroMagnetic mass velocity

W Bosons are comprised of
ODD number quanta

Bosons are TRANSVERSE Charge carriers

ODD #

Q

Planck quanta

$n\pi\left[[m\Omega v^2]\right]$

mass velocity

W Bosons

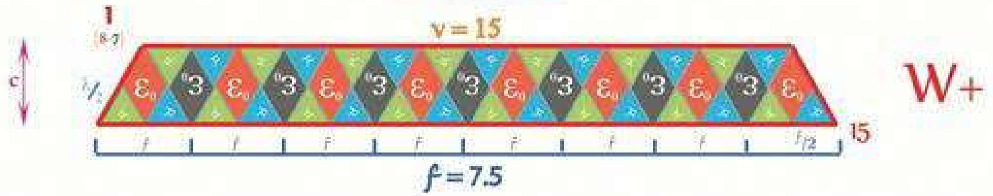
1
[8-7]
v = 15
c
½λ
f
f/2
15
f = 7.5
W+

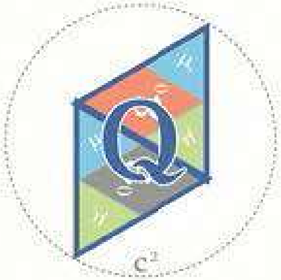

Neutral Z BOSONS and PHOTONS have differing EM geometries

EVEN #

Q

Planck quanta

$2n\pi\left[[m\Omega v^2]\right]$

mass velocity

Photons

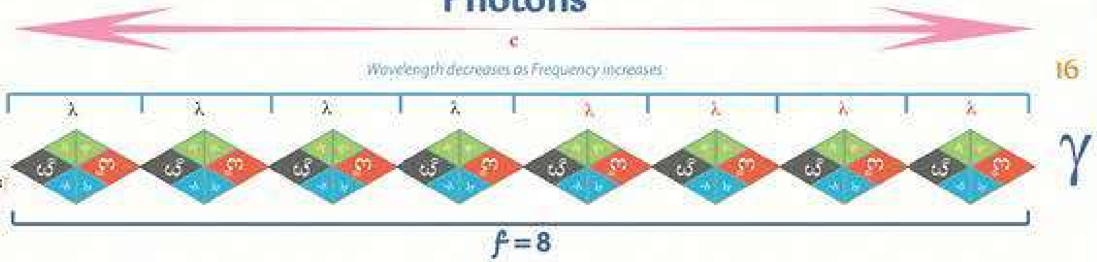
c
Wavelength decreases as Frequency increases
16
λ
0
[8-8]
f = 8
γ

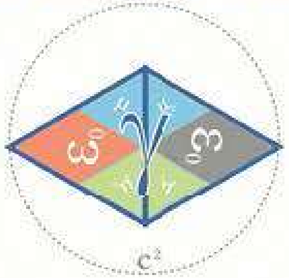

Photons are LONGITUDINAL
neutral EM force carriers
resulting from Bosons

ODD #

Q

Planck quanta

$n\pi\left[[m\Omega v^2]\right]$

mass velocity

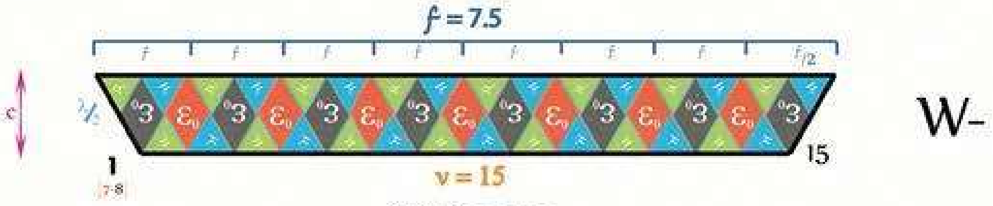
f = 7.5
f
f/2
c
½λ
1
[7-8]
15
v = 15
W-

W Bosons

EM waves are comprised of
EVEN number quanta [Photons]

EM Field Planck quanta

$\text{EVEN}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

EM waves ElectroMagnetic mass velocity

Boson Waveforms

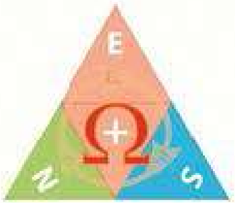

W The Electro-weak force is the result of Bosons interacting along their edge of Permeability

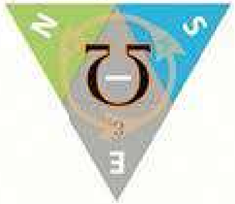

Boson exchange is the basis of Electro-Magnetic Induction & Charge transfer

S

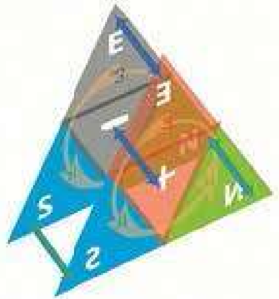

The Strong Colour Force is the result of Bosons interacting via their their Electric charge fascia

All ODD Ω geometries [bosons] create a quantum of charge

Positive charge carrier

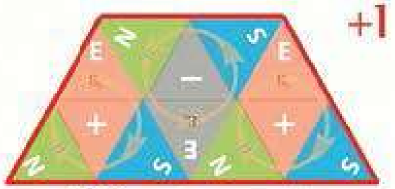

W+

ODD number quanta

All charges are comprised of odd numbered Bosons

[each Boson is a Quantum level]

ODD number quanta

W-

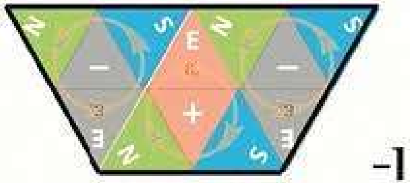
N
S
E
N
S
–
–
+
E
N
S
E
-1

Negative charge carrier

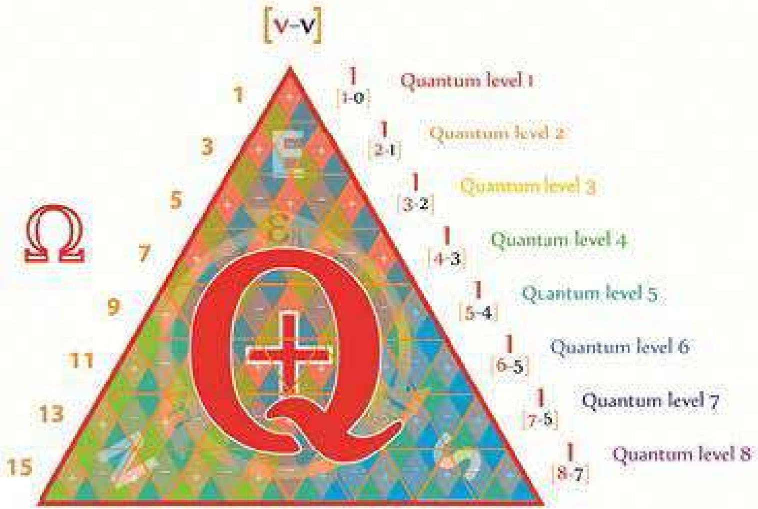
[v-v]
Ω
1
3
5
7
9
11
13
15
1 [1-0] Quantum level 1
1 [2-1] Quantum level 2
1 [3-2] Quantum level 3
1 [4-3] Quantum level 4
1 [5-4] Quantum level 5
1 [6-5] Quantum level 6
1 [7-5] Quantum level 7
1 [8-7] Quantum level 8
E
N
S
+

Integral W Bosons form square Energy geometries
Separated energy geometries create an electromotive force

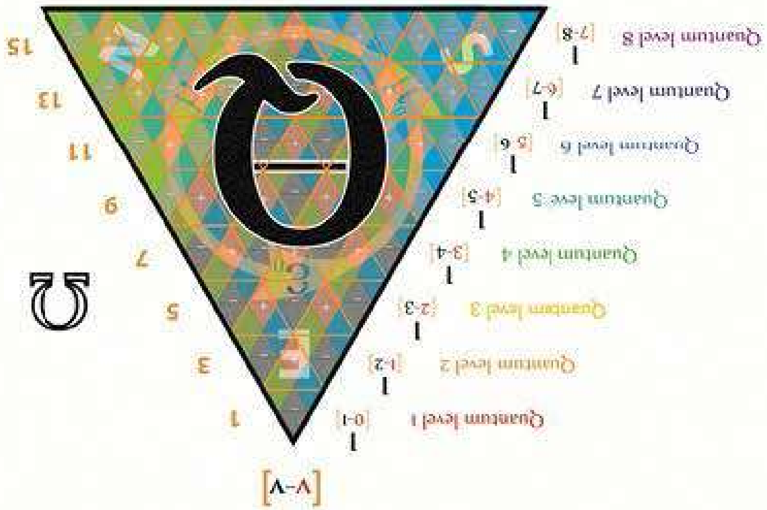
℧
15
13
11
9
7
5
3
1
1 [8-7] Quantum level 8
1 [7-6] Quantum level 7
1 [6-5] Quantum level 6
1 [5-4] Quantum level 5
1 [4-3] Quantum level 4
1 [3-2] Quantum level 3
1 [2-1] Quantum level 2
1 [1-0] Quantum level 1
[v-v]
–

Copyright ABRAHAM [2008] - all rights reserved

Bosons

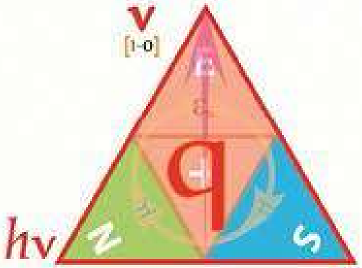

Positive Charges have nett
positive quantised angular momenta

Positive Bosons

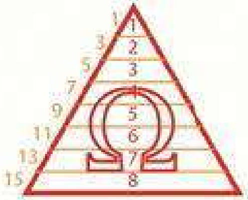
1
3
5
7
9
11
13
15
1
2
3
5
6
7
8
Ω

EM Field Planck quanta

$\text{ODD}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Bosons ElectroMagnetic mass velocity

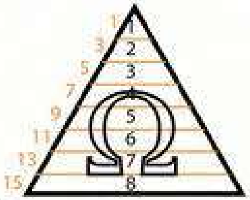
1
3
5
7
9
11
13
15
1
2
3
5
6
7
8
Ω

Bosons are the geometry of Quantum Levels

Negative Bosons

Negative Charges have nett
negative quantised angular momenta

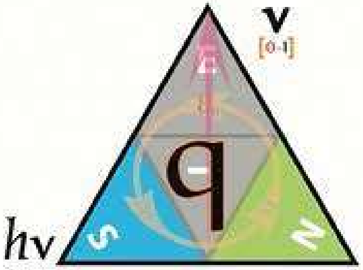

**Bosons are
TRANSVERSE
EM fields [levels]**

Each quantum Level is
a unit of Charge

**nett Charge
is a SCALAR
EM field property**

Charge

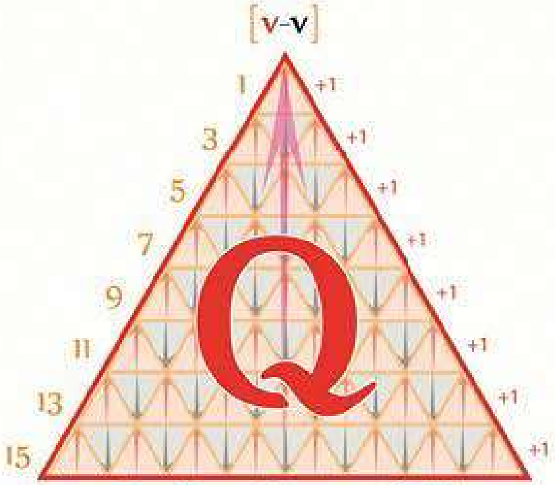
[v-v]
1
3
5
7
9
11
13
15
+1
+1
+1
+1
+1
+1
+1
+1
Q

Positive Charge field

**Charge is a measure of the nett
quantised angular momenta - energy geometry
of any given region of Space-time**

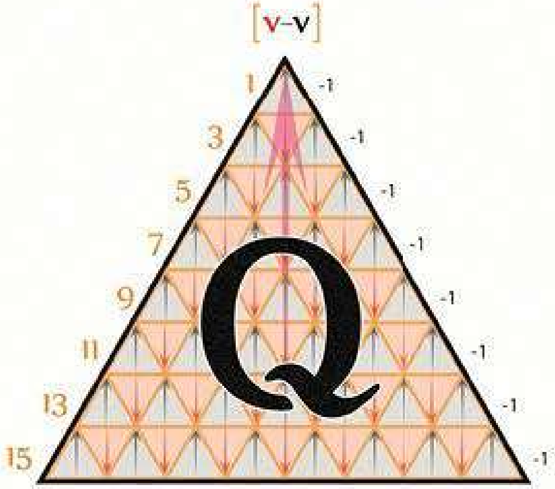
[v-v]
1
3
5
7
9
11
13
15
-1
-1
-1
-1
-1
-1
-1
-1
Q

Negative Charge field

bosons

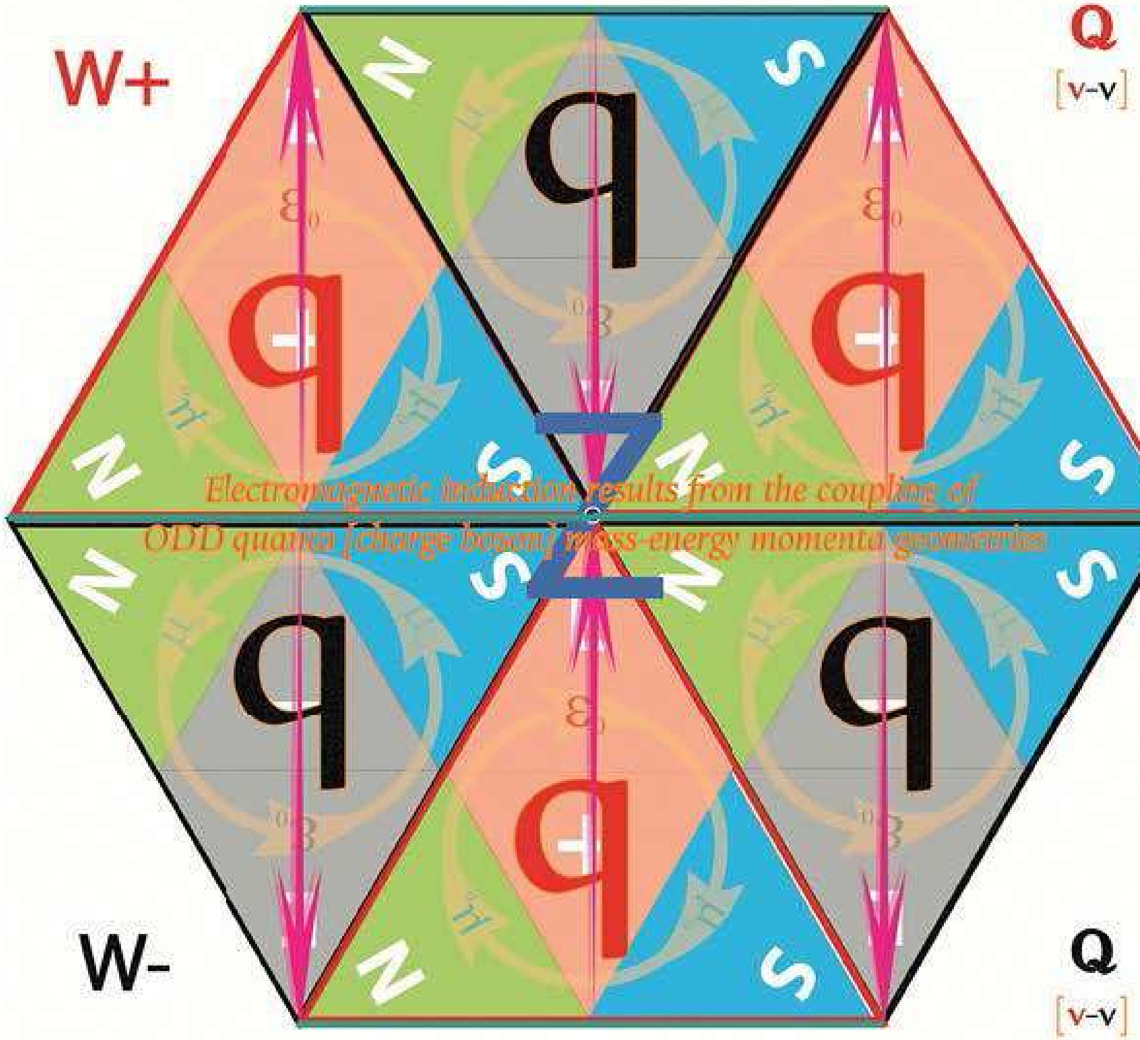

Electromagnetic induction results from the coupling of ODD quanta [charge bosons] mass-energy momenta geometries

ODD π

The vector linear momentum of all bosons & photons determines their directions of propagation

photon

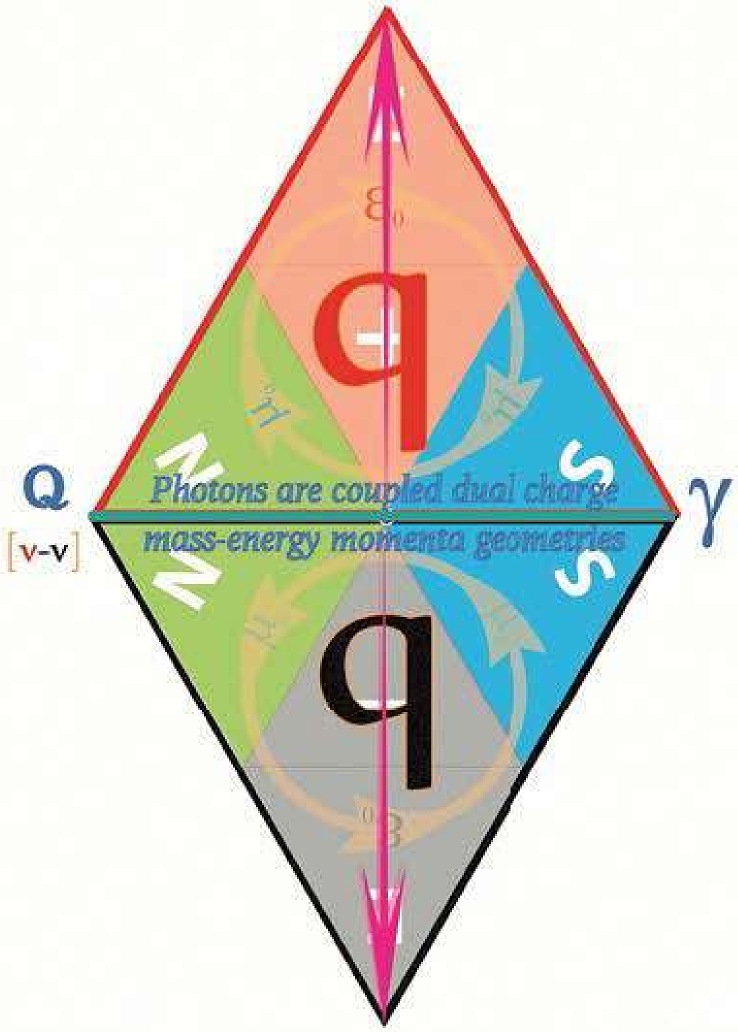

Photons are coupled dual charge mass-energy momenta geometries

2nπ

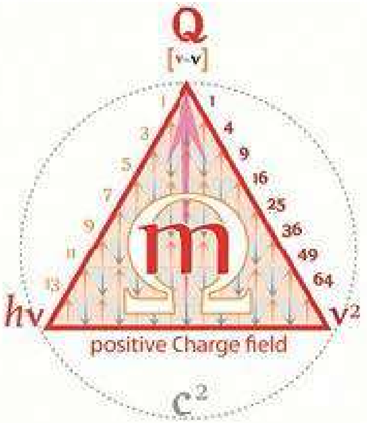
Q
[v-v]
1
3
4
5
7
9
11
13
16
25
36
49
64
m
Ω
hv
v²
positive Charge field
c²

mass geometries

m

mass
energy
spatial field
impedance
Planck
quanta
mass
velocity

$$n\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

E

mass-energy

the distribution of energy momenta in EM fields creates charged fascia

charged planar mass-energy momenta form

radiant 2D equilateral geometries

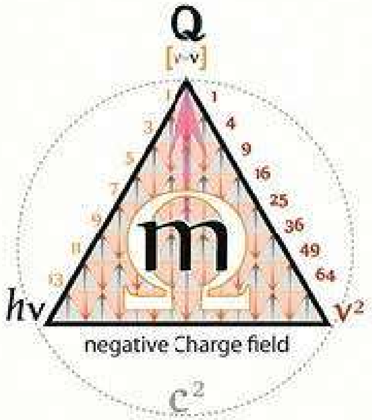
Q
[v-v]
1
3
4
5
7
9
11
13
16
25
36
49
64
m
Ω
hv
v²
negative Charge field
c²

ρ

mass-energy geometries

m

kg/m^2

energy momenta per second

$$\frac{E}{c^2} = m$$

mass

E

ENERGY

energy momenta per second squared

$$M = \frac{E}{c^4}$$

Matter

ρ

mass-Matter topologies

M

KG/m^3

charged mass-energy geometries form

standing-wave 3D tetrahedral topologies

Matter displacement topologies have internalised strong force fascia partitions

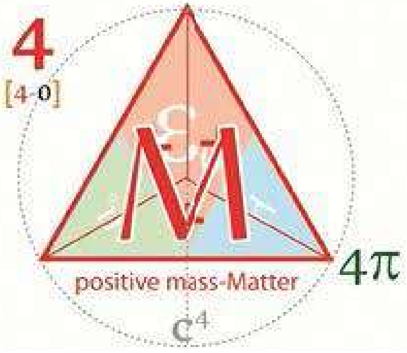
4
[4-0]
ε
M
positive mass-Matter
4π
c⁴

mass-Matter

m

Matter
spatial field
impedance
Planck
quanta
mass
velocity

$$4n\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

M

Matter topologies

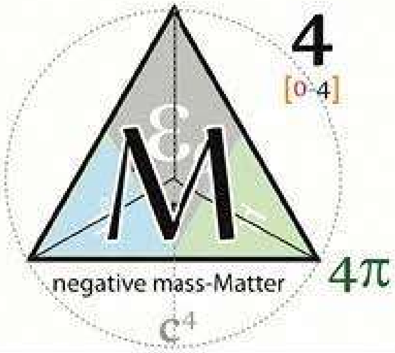
4
[0-4]
ε
M
negative mass-Matter
4π
c⁴

Copyright ABRAHAM [2008] - all rights reserved

Matter

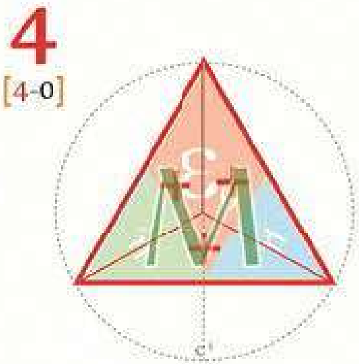

2D

ρ/sec

EM mass-energy per square metre

EM mass geometries

m

kg/m²

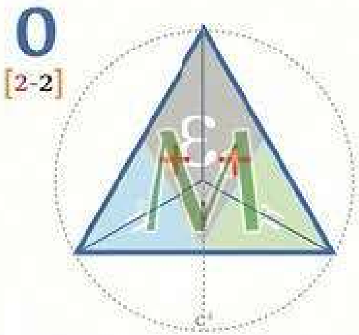

EM mass-energies have 2D planar geometries

$E = mv^2$

2D EM mass-Energy forms 3D Matter

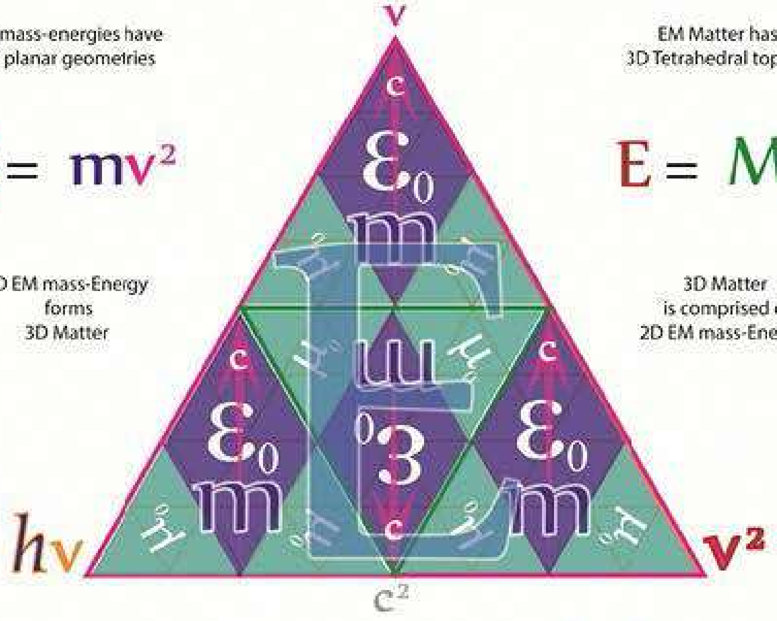

EM Matter has a 3D Tetrahedral topology

$E = Mc^4$

3D Matter is comprised of 2D EM mass-Energies

Matter is comprised of 4nπ mass-energy geometries forming a 3D standing wave topology

Tetryons — EM Field — Planck — quanta

$$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter quanta — ElectroMagnetic — mass — velocity

Matter is anything that has a closed mass-energy topology and displaces a volume
(the 3D massive building blocks of quantum particles and atomic elements)

The EM mass-energy quanta of Matter can be measured transversely and longitudinally allowing them to exhibit a number of wave-like and particle-like properties
(de Broglie wavelength & Compton frequency)

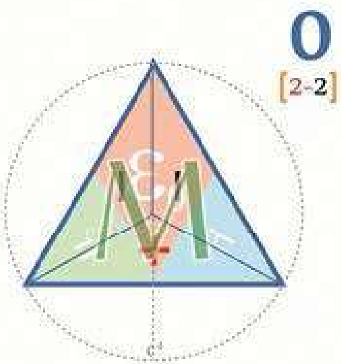

3D

ρ/sec^2

EM mass-energy per cubic metre

Matter topologies

M

kg/m³

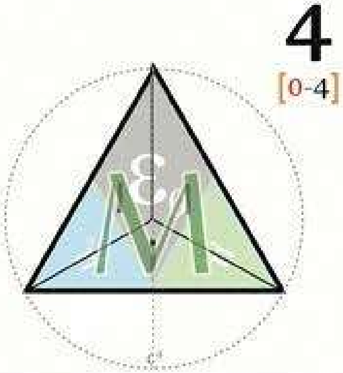

Tetryons - the quanta of Matter

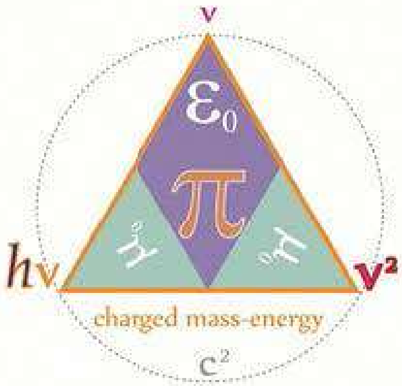
v
ε_0
π
hv
v²
charged mass-energy
c²

'massless' is a physics mis-nomer as all energy exhibits mass equivalence

ZPF

mass-Energy quanta

m

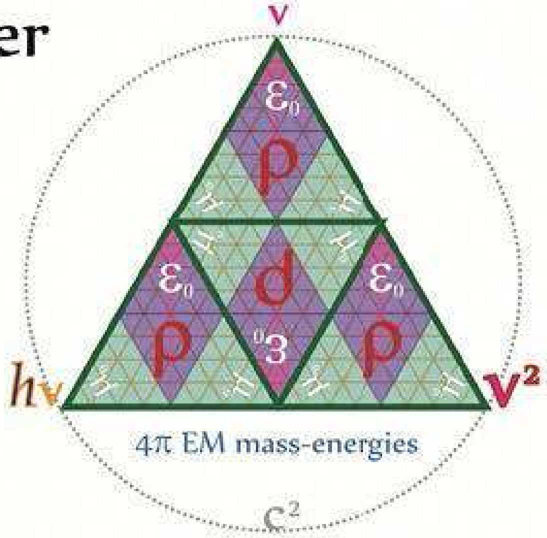
v
ε_0
ρ
d
3
hv
v²
4π EM mass-energies
c²

$$m = n\pi\left[\left[\underset{\text{impedance}}{\overset{\text{spatial}}{1/c^2}}\right].\overset{\text{Energy}}{\underset{\text{momenta}}{\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\Omega\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]}}\right]$$

TETRYONS

Platonic tetrahedrons are the foundational topologies of all 3D Matter

$$T\pi\left[\left[\underset{\text{impedance}}{\overset{\text{spatial}}{1/c^4}}\right].\overset{\text{Energy}}{\underset{\text{momenta}}{\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\Omega\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]}}\right] = \underset{\text{Matter}}{M}$$

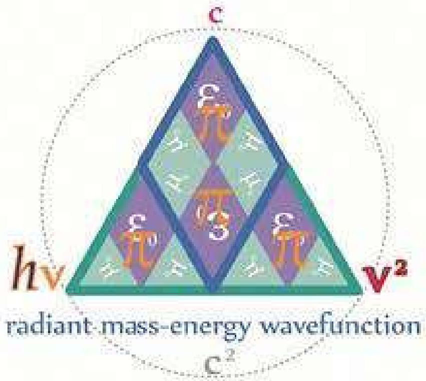
c
ε_0
π
3
hv
v²
radiant mass-energy wavefunction
c²

Tetryon

Matter Quantum

M

2D mass-energy geometries can be combined to form 3D mass-ENERGY-Matter particles

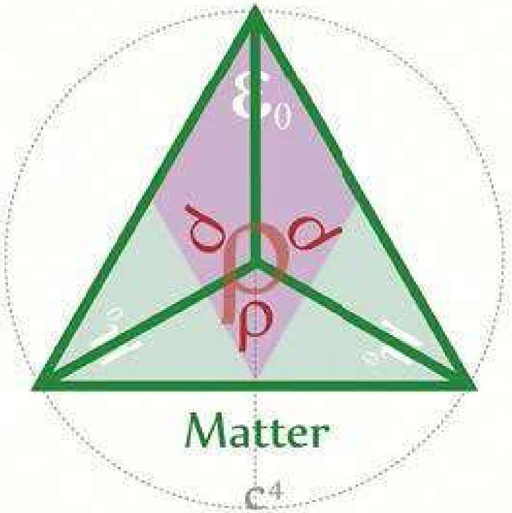
ε_0
ρ
Matter
c⁴

Standing wave mass-energy geometries and Matter topologies

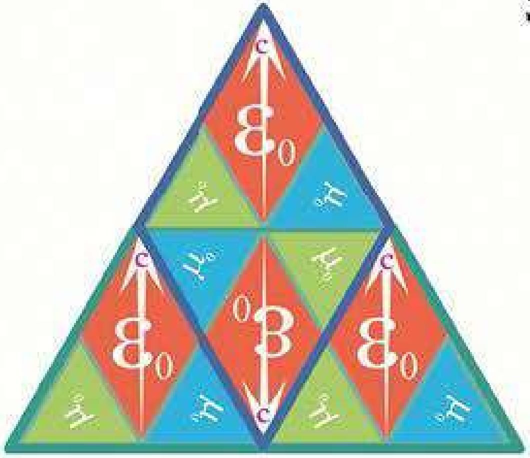

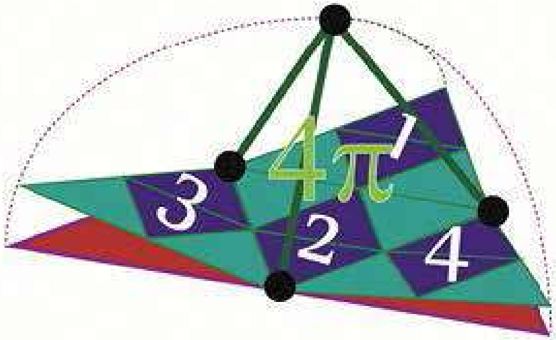

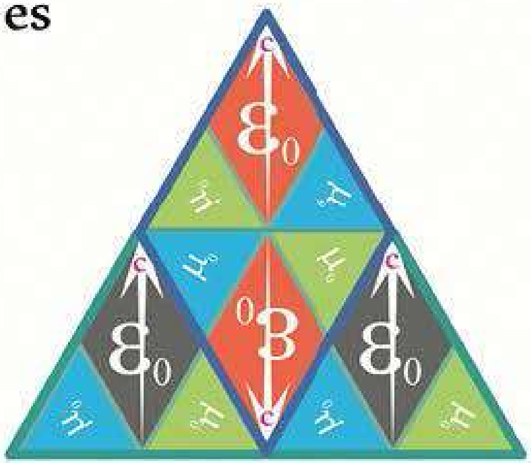

As each charged fascia's E field tries to propagate outwards it interacts with a M field dipole at each apex

$$M = T\pi\left[\left[1/c^4\right].\left[m\Omega v^2\right]\right]$$

Matter

spatial

impedance

Energy

Planck

quanta

mass

velocity

momenta

In turn energised M fields supply energy to the E Fields creating a EM standing wave [the EM topology of Matter]

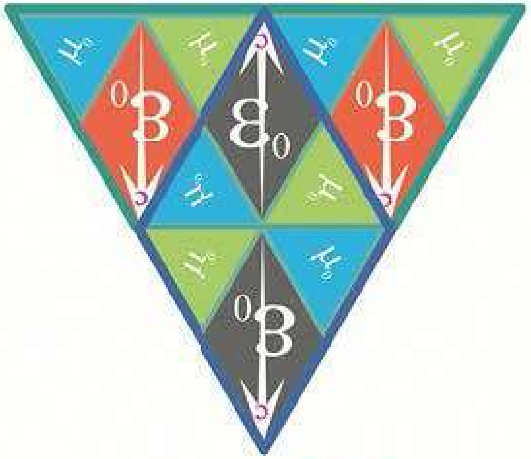

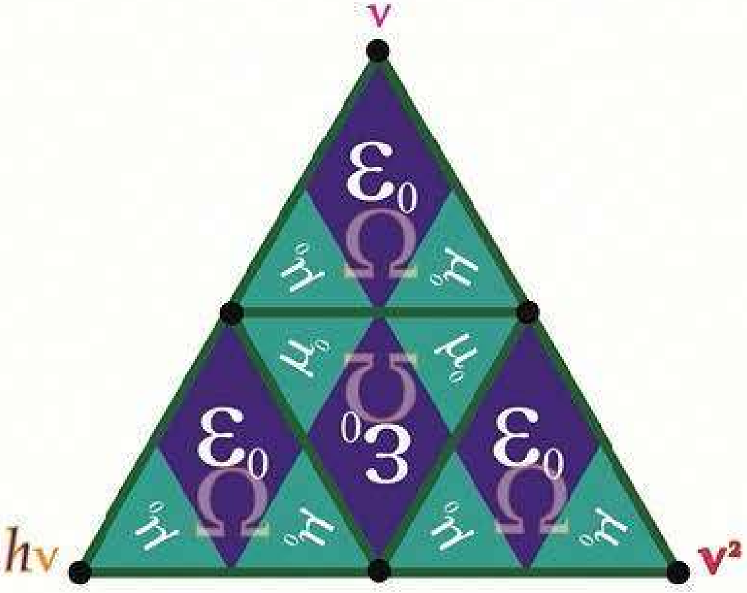

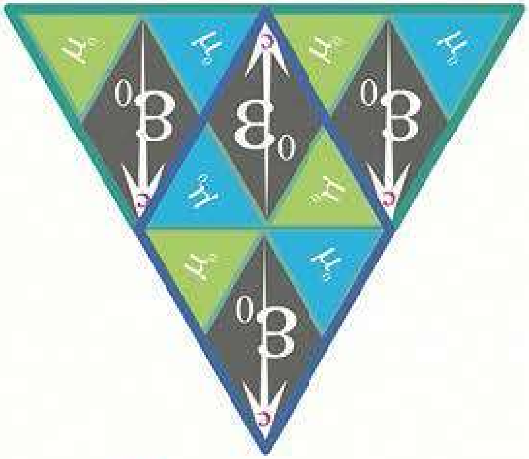

Matter stores EM mass-energies geometries in its 3D Tetrahedral topology

All mass-energy propagate at the speed of light creating linear energy momentum

Tetryonic charge topologies

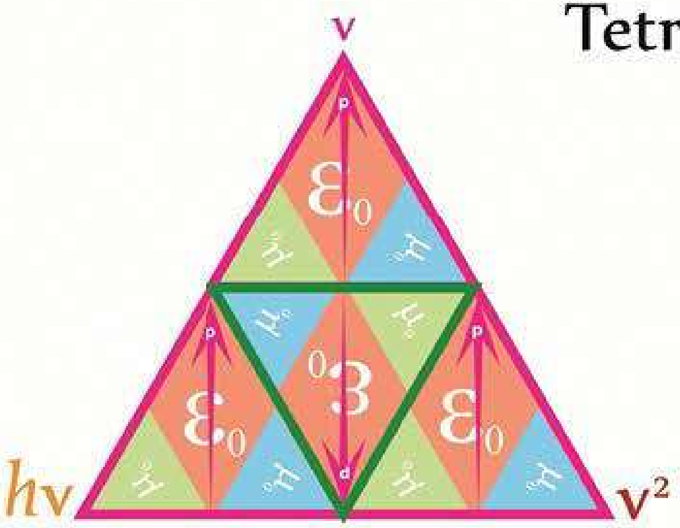
v
p
ε0
μ0
ε0
ε0
d
hv
v²

Positive Tetryon

The 3D Tetryonic volume of Matter is what distingushes it from 2D EM mass-energies

Negative Tetryon

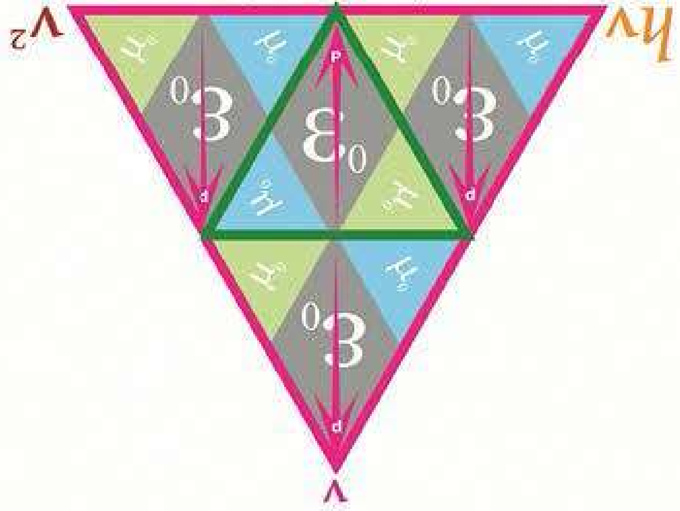
v²
hv
ε0
ε0
ε0
μ0
p
d
v

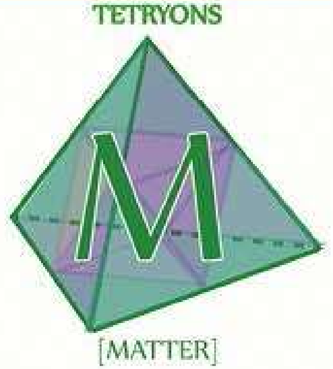
TETRYONS
M
[MATTER]

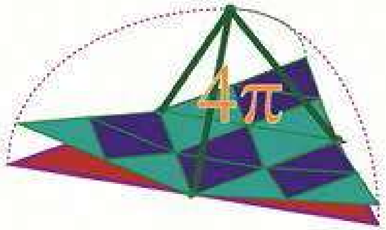

The four EM mass-energy momenta produced by Tetryonic collapse must not be confused with the classical 4 Energy-momenta (which refers to x,y,z co-ordinates & v)

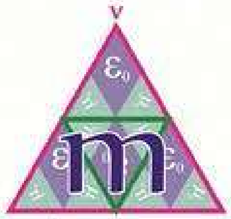

4pi radiant mass-energies

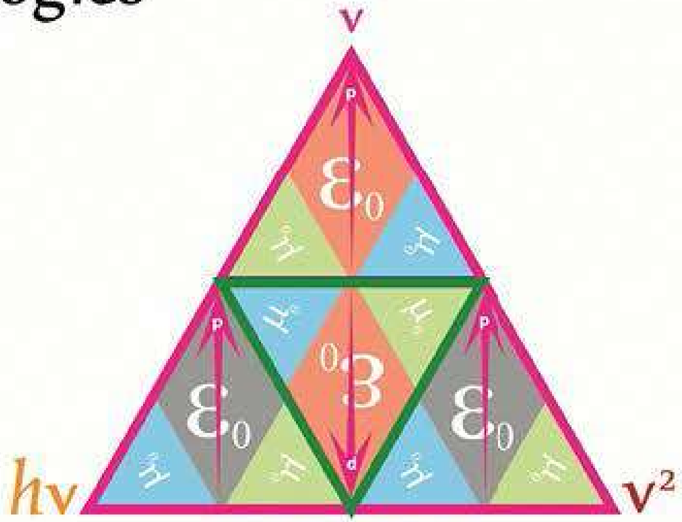
v
p
ε0
μ0
ε0
ε0
d
hv
v²

Neutral Tetryon

A 3D Tetryon can be 'collapsed' into a 2D waveform radiant four energy momenta EM waveform that conserves all of its 2D mass-energy-momenta

Neutral Tetryon

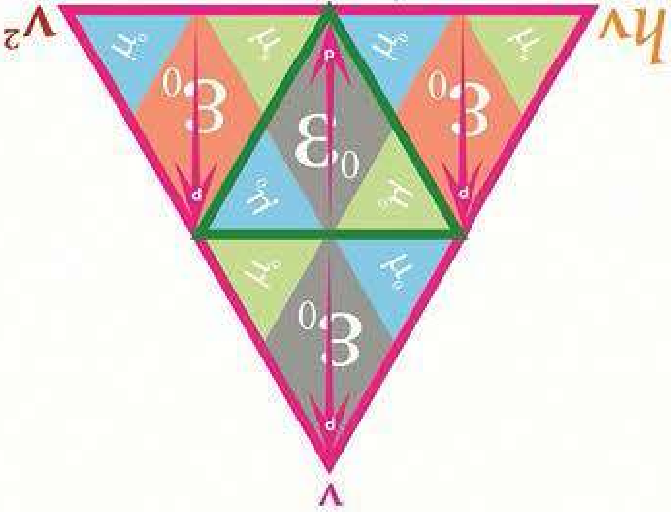
v²
hv
ε0
ε0
ε0
μ0
p
d
v

Four 2D mass-energy geometries can combine to form via charge interactions 3D tetrahedral Matter topologies

Tetryon Genesis

1. **WEAK interaction**
Mutual Inductive Coupling
[Magnetic dipoles interact]

ODDπ
charged mass-energies

Non-Zero
nett mass-Energy-momenta
result in Linear momentum

EVENπ
radiant mass-energies

2. *2D EM mass-energies*
interact to form
3D Matter topologies

EM wave momenta
can form standing waves
[Matter geometries]

3. **STRONG interaction**
Electrostatic Matter attract
via Electric charges and
Magnetic dipoles

4nπ
mass-Matter geometries

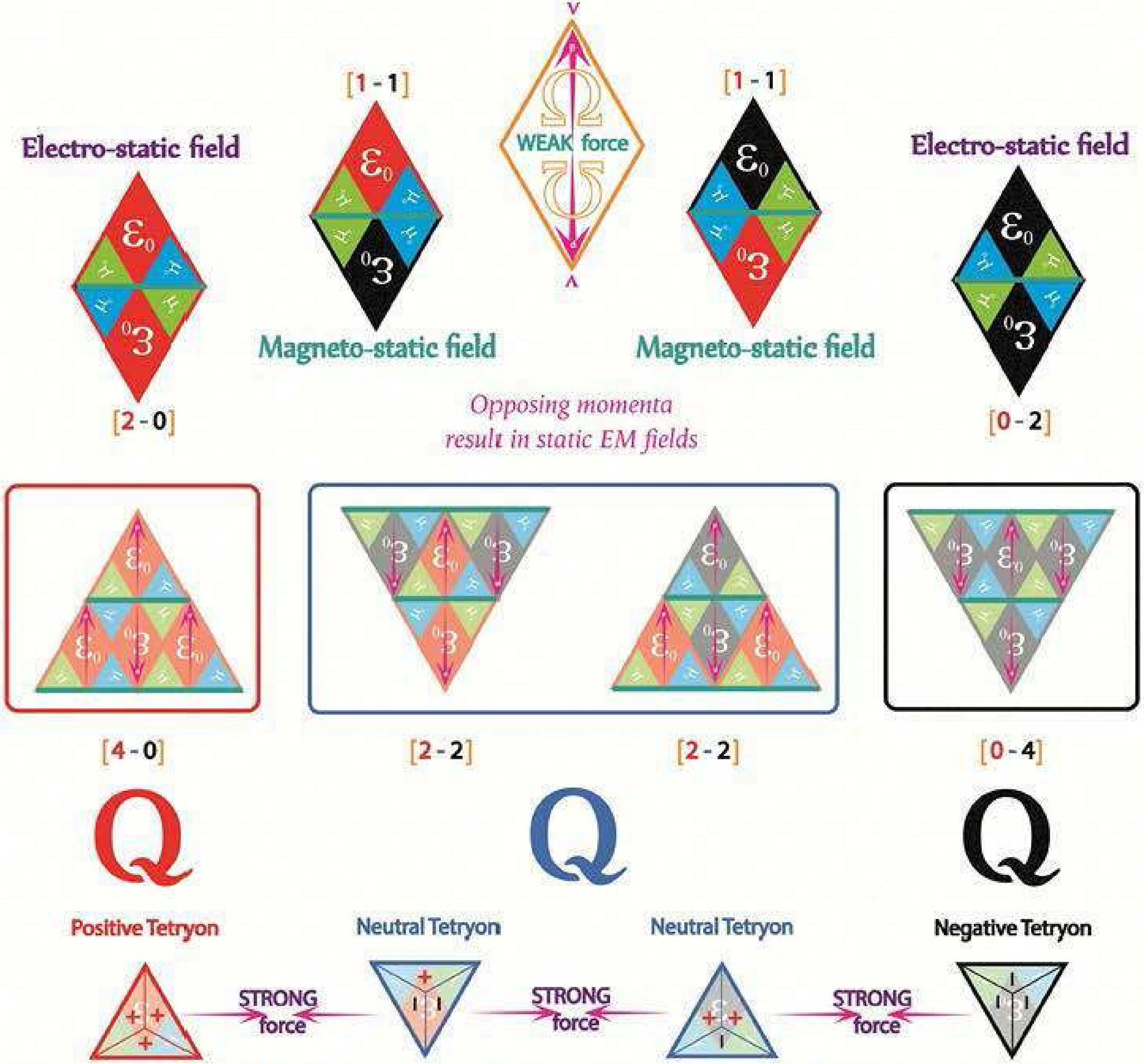
Electro-static field
[2 - 0]
[1 - 1]
Magneto-static field
WEAK force
Ω
[1 - 1]
Magneto-static field
Electro-static field
[0 - 2]
ε₀
μ₀
[4 - 0]
[2 - 2]
[2 - 2]
[0 - 4]
Q
Q
Q
Positive Tetryon
Neutral Tetryon
Neutral Tetryon
Negative Tetryon
STRONG force
STRONG force
STRONG force
+
-

Opposing momenta
result in static EM fields

All Tetryonic charges seek equilibrium

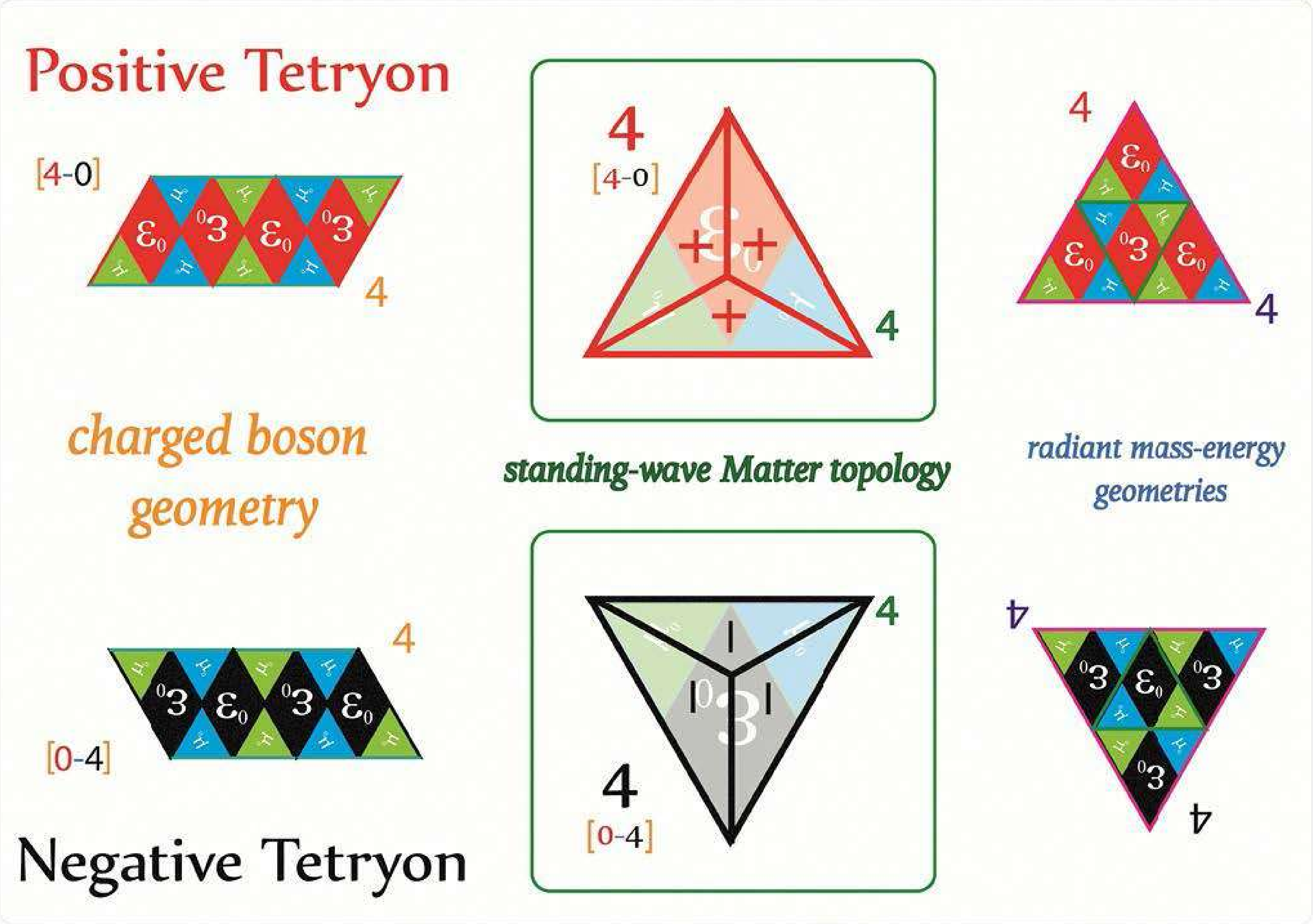

Positive Tetryon

charged boson geometry

standing-wave Matter topology

radiant mass-energy geometries

Negative Tetryon

[4-0]
4
[4-0]
4
4
4
4
ε0
μ0
[0-4]
4
4
[0-4]
4
4
4

Neutral Tetryons

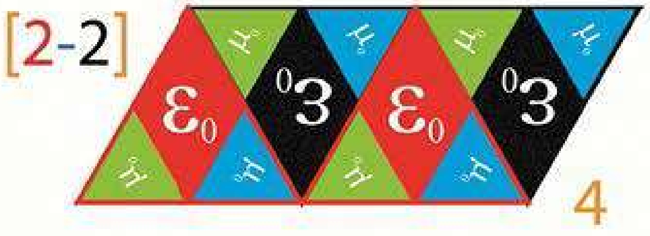
[2-2]
ε0
03
ε0
03
4

charged boson geometry

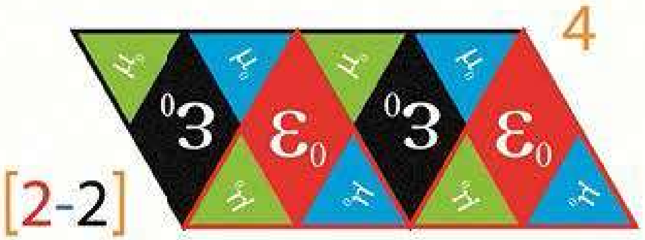
4
03
ε0
03
ε0
[2-2]

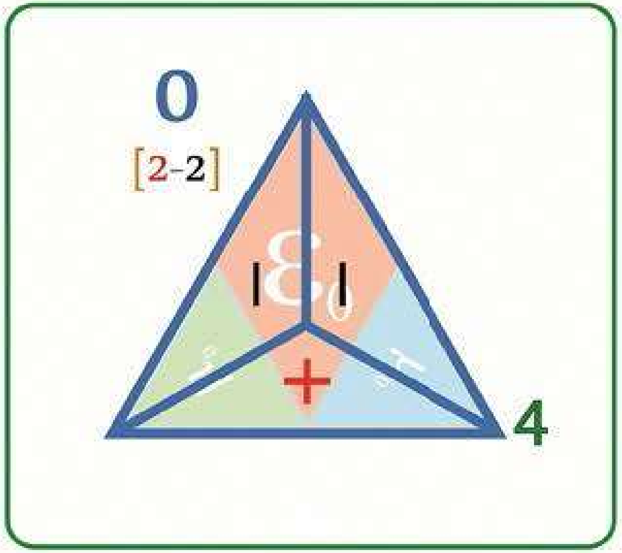

standing-wave Matter topology

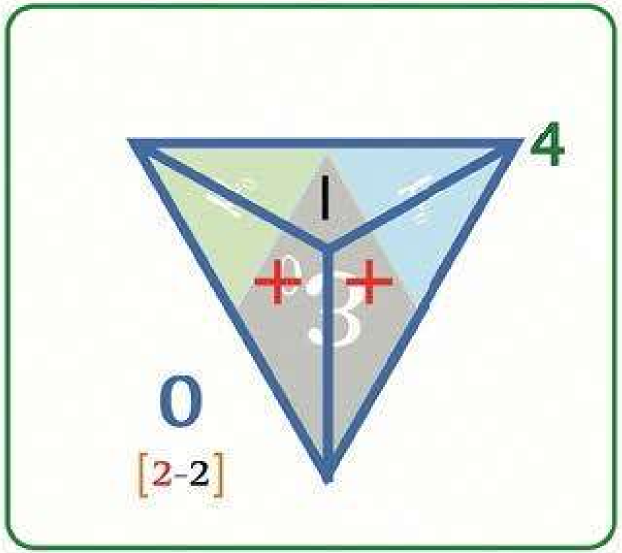

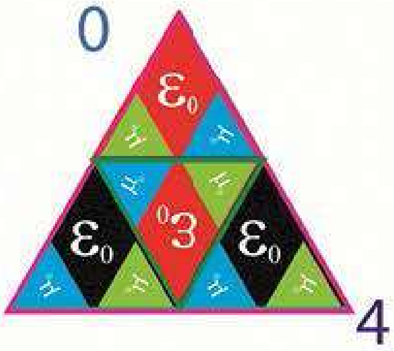
0
ε0
ε0
03
ε0
4

radiant mass-energy geometries

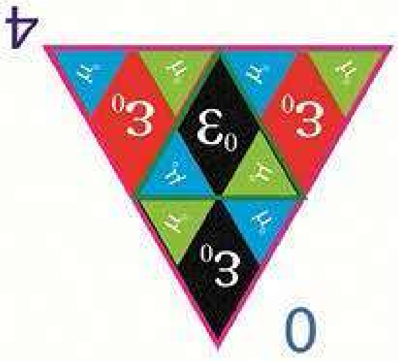
4
03
ε0
03
03
0

Copyright ABRAHAM [2008] - all rights reserved

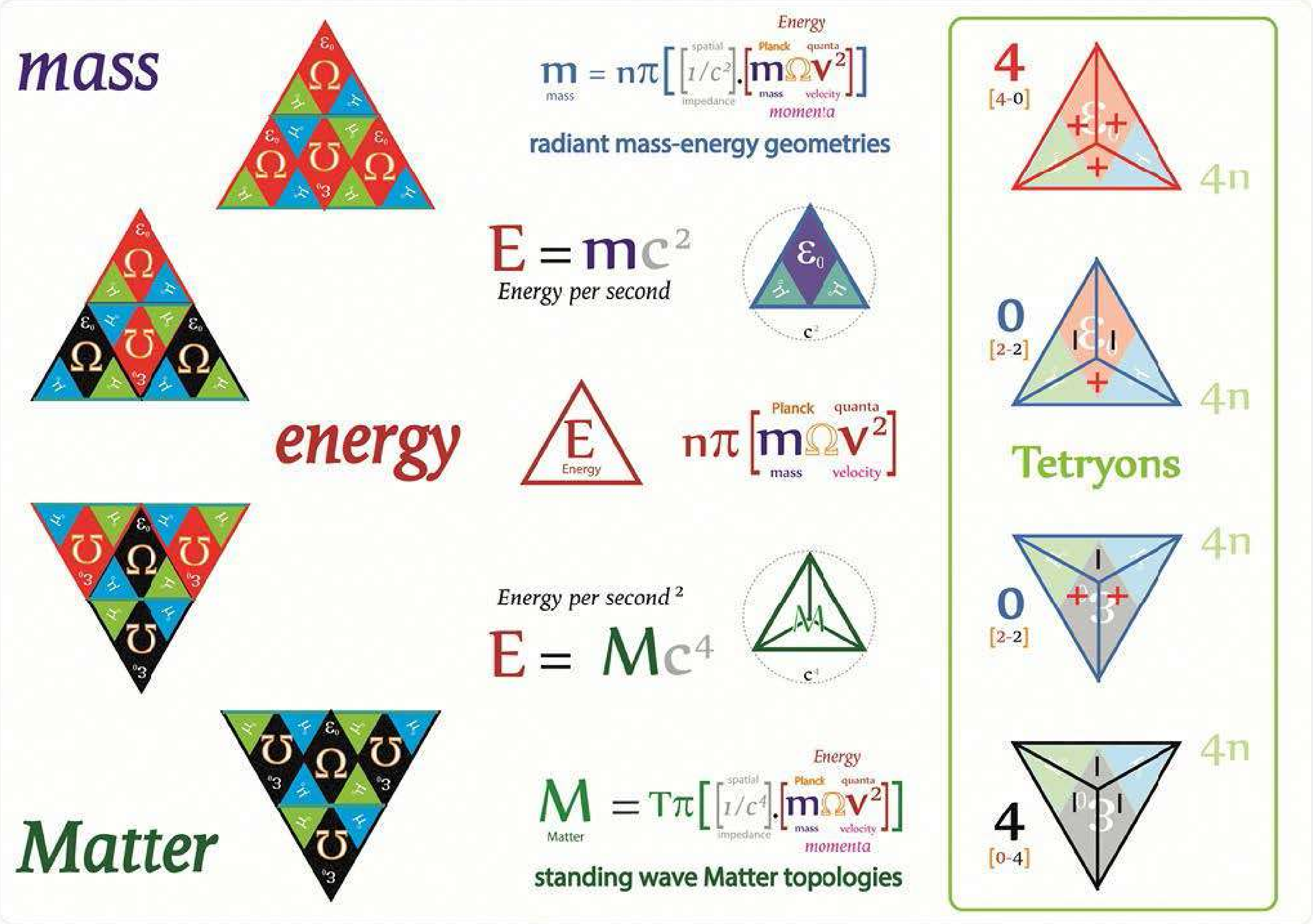
mass
Energy
spatial
Planck
quanta
impedance
mass
velocity
momenta
Energy
c²
energy
E
Energy
Planck
quanta
mass
velocity
c⁴
Energy
spatial
Planck
quanta
impedance
mass
velocity
momenta
Matter
Tetryons
4
[4-0]
4n
0
[2-2]
4n
4n
0
[2-2]
4n
4
[0-4]

mass

$$m = n\pi\left[\left[1/c^2\right].\left[m\Omega v^2\right]\right]$$

radiant mass-energy geometries

$$E = mc^2$$

Energy per second

energy

$$n\pi\left[m\Omega v^2\right]$$

Energy per second²

$$E = Mc^4$$

Matter

$$M = T\pi\left[\left[1/c^4\right].\left[m\Omega v^2\right]\right]$$

standing wave Matter topologies

Tetryons

Tetryonics 06.08 - Tetryon family

mass-ENERGY-Matter

geometries topologies

mass EM Field Planck quanta

$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

geometry ElectroMagnetic mass velocity

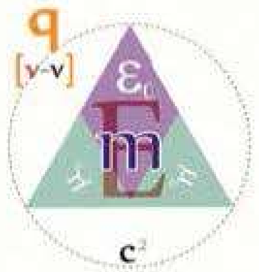

2D mass-energy geometries
are a property of
3D Matter topologies

Tetryonic Matter
is the building block
of all Fermions,
elements and
compounds

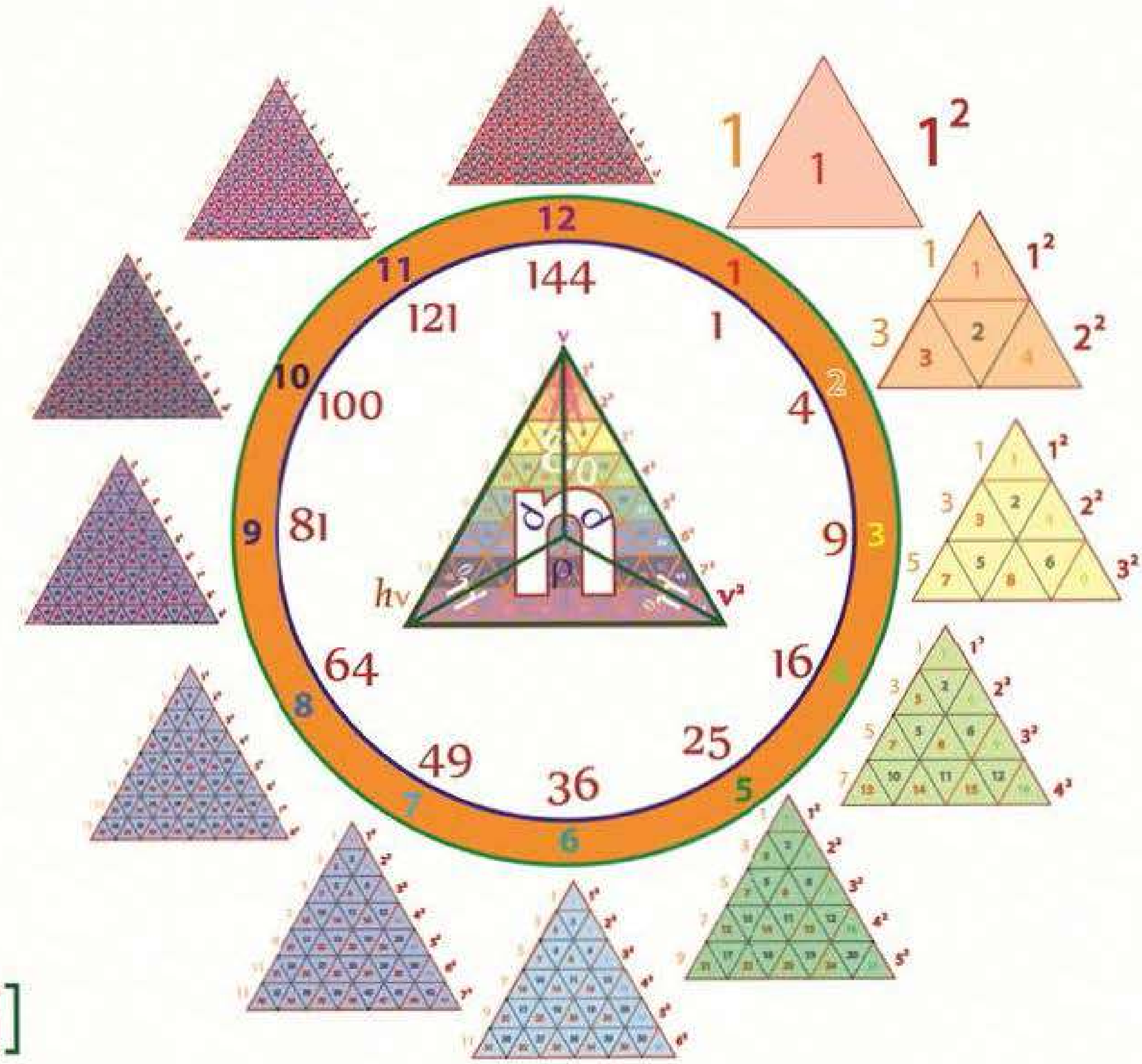
1
1
1²
1
1²
3
2
3
4
2²
1
1²
3
2
3
2²
5
5
7
6
8
3²
1²
2²
3²
4²
12
144
1
1
2
4
3
9
4
16
5
25
6
36
7
49
8
64
9
81
10
100
11
121
v
hv
v²
ρ

ALL EM masss-energies
have equilateral geometry

Transverse
EM mass-energies
form bosons
[Quantum levels]

Longitudinal
EM mass-Energies
form Photons

Bosons combine
to form SQUARED
scalar EM energies

massless is a scientific
misnomer

2D [non-Topolgical]
EM mass-ENERGY is
'Matterless'

ALL Matter
has a 3D topology

Matter EM Field Planck quanta

$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

topology ElectroMagnetic mass velocity

Tetryonic Matter

All quantum Matter has a foundational tetrahedral topology
as a result of their equilateral mass-energy geometries
[not spherical as has been assumed from the math]

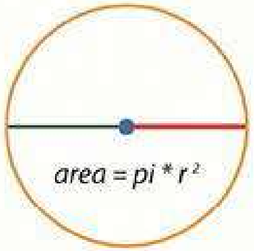

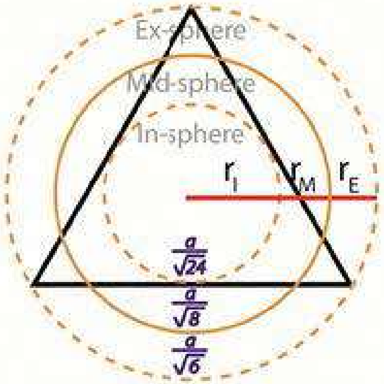
Ex-sphere
Mid-sphere
In-sphere
r_I
r_M
r_E
$\frac{a}{\sqrt{24}}$
$\frac{a}{\sqrt{8}}$
$\frac{a}{\sqrt{6}}$

Sphere

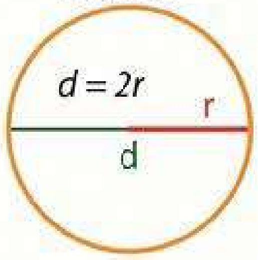

Area sphere= $4\pi r^2$
Surface curvature = $1/r^2$
Total curvature = 4π

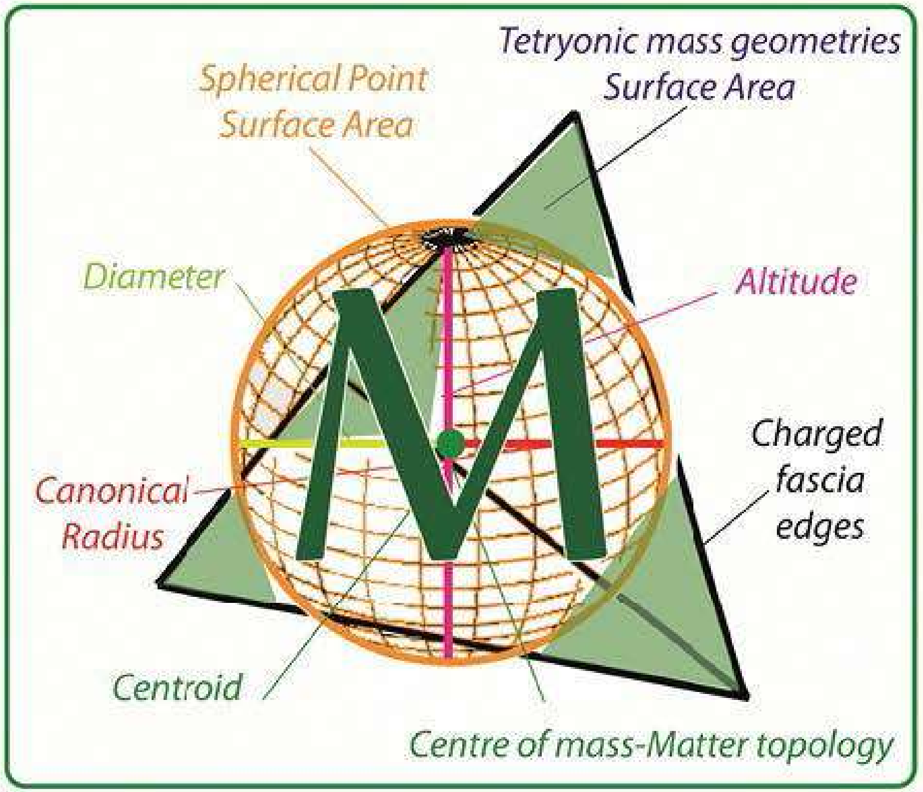
Tetryonic mass geometries
Surface Area
Spherical Point
Surface Area
Diameter
Altitude
Charged
fascia
edges
Canonical
Radius
Centroid
Centre of mass-Matter topology
M

Spheres & Tetrahedra are both 3D Platonic solids
with 4π scalar integral Gaussian topologies
and physical displacement volumes
[Gauss-Bonnet theorem]

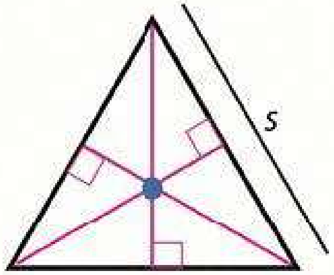

area = $\sqrt{3/4} * a^2$

ZPF fascia

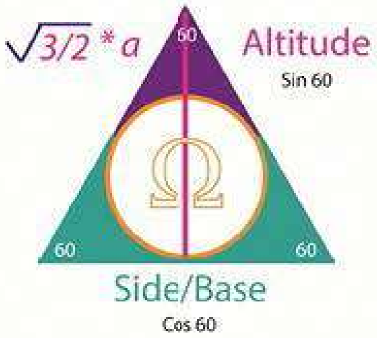
$\sqrt{3/2} * a$
Altitude
Sin 60
60
60
60
Ω
Side/Base
Cos 60

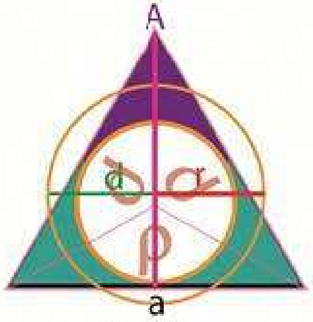

Area Tetryon

$$A = 4\,A_0 = \sqrt{3}a^2$$

Vertex curvature = π
Total curvature = 4π

Dodecyons

while quarks & leptons are comprised of
12π mass-energy geometries
they have differing final Matter topologies

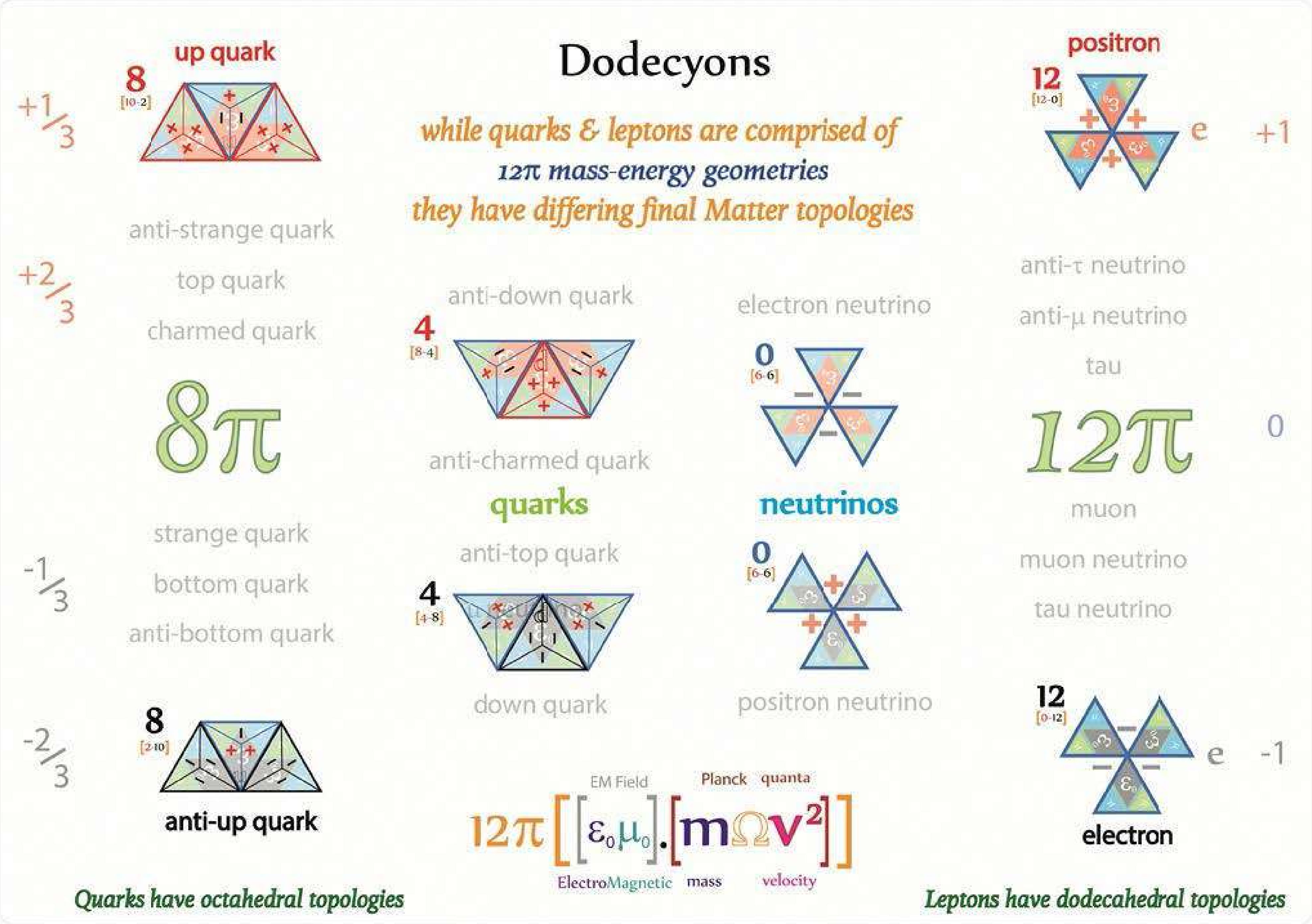
+1/3
8
[10-2]
up quark
+2/3
-1/3
-2/3
8
[2-10]
anti-up quark
4
[8-4]
4
[4-8]
0
[6-6]
0
[6-6]
12
[12-0]
positron
e
+1
0
12
[0-12]
electron
e
-1

anti-strange quark
top quark
charmed quark

8π

strange quark
bottom quark
anti-bottom quark

anti-down quark

anti-charmed quark

quarks

anti-top quark

down quark

electron neutrino

neutrinos

positron neutrino

anti-τ neutrino
anti-μ neutrino
tau

12π

muon
muon neutrino
tau neutrino

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

ElectroMagnetic mass velocity

Quarks have octahedral topologies

Leptons have dodecahedral topologies

Copyright ABRAHAM [2008] - all rights reserved

-2/3

quark octahedral topology

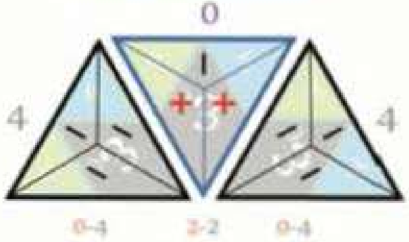

attractive dodecyon geometry

anti-up quark

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

dodecyon ElectroMagnetic mass velocity

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

1. **anti-up** quark
2. **anti-charmed** quark
3. **anti-top** quark

nett charge

8

[2-10]

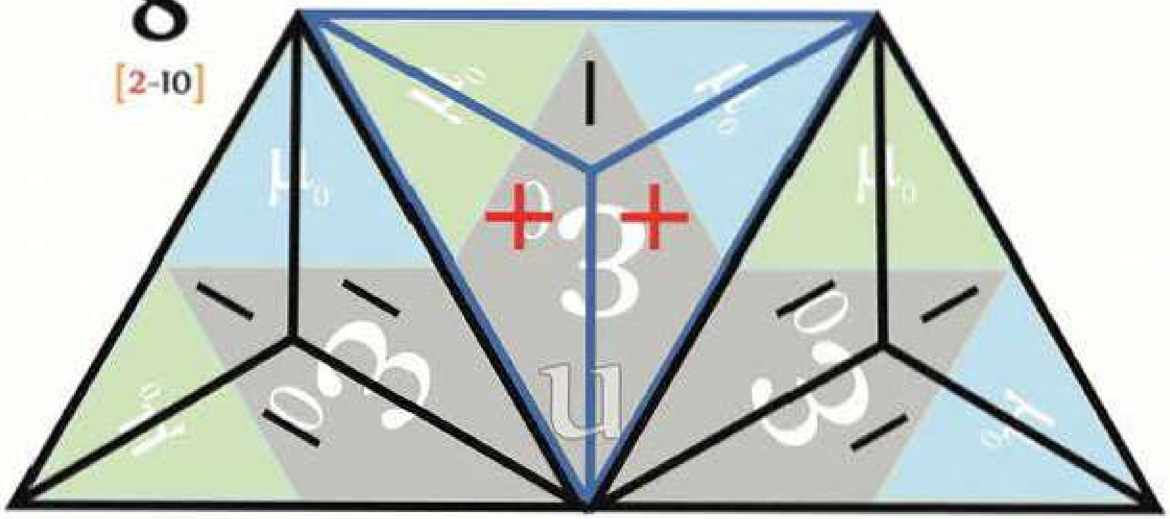

charged
mass-Matter
topology

8π

Quarks have 12 charged fascia - *mass-energy geometries*

2 Positive 10 Negative

Quarks have octahedral mass-Matter topologies

[-2/3 elementary charge]

down quark

-1/3

quark octahedral topology

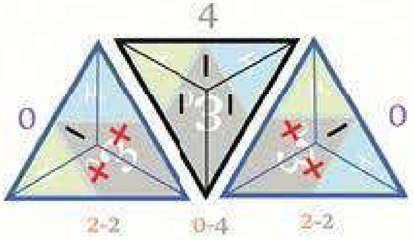

attractive dodecyon geometry

EM Field · Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

dodecyon · ElectroMagnetic · mass · velocity

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

1. **down** quark
2. **strange** quark
3. **bottom** quark

d

nett charge

4

[4-8]

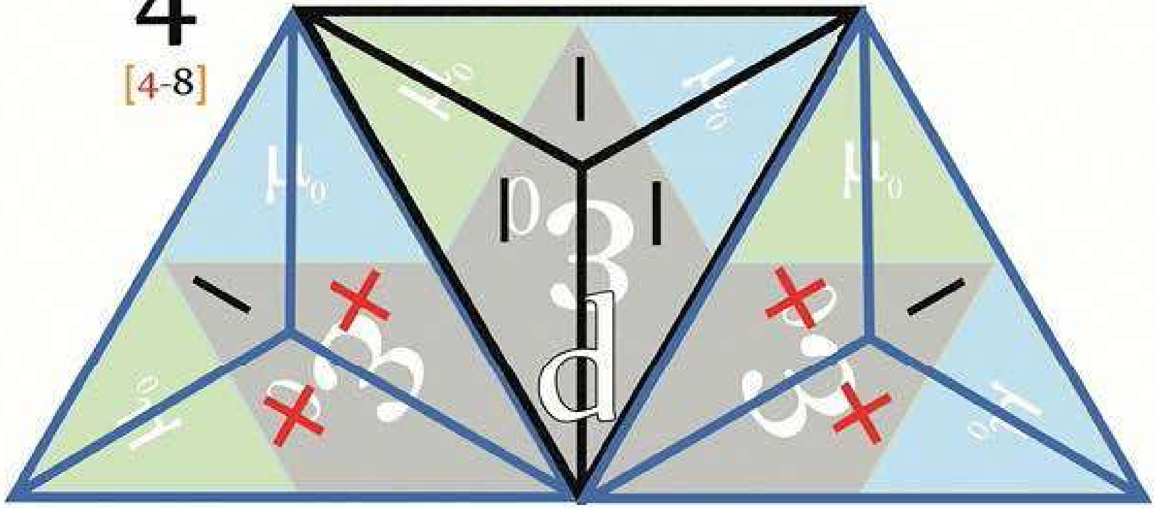

charged
mass-Matter
topology

8π

Quarks have 12 charged fascia - mass-energy geometries

4 Positive   8 Negative

Quarks have octahedral mass-Matter topologies

[-1/3 elementary charge]

anti-down quark

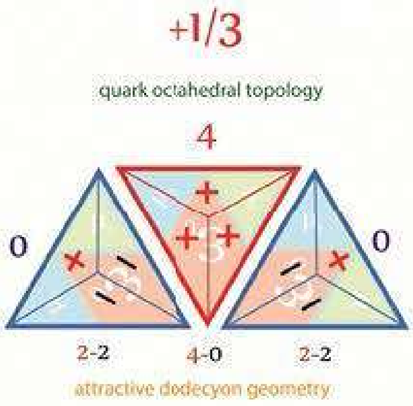
+1/3
quark octahedral topology
4
0
0
2-2
4-0
2-2
attractive dodecyon geometry

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

dodecyon ElectroMagnetic mass velocity

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

1. **anti-down** quark
2. **anti-strange** quark
3. **anti-bottom** quark

d

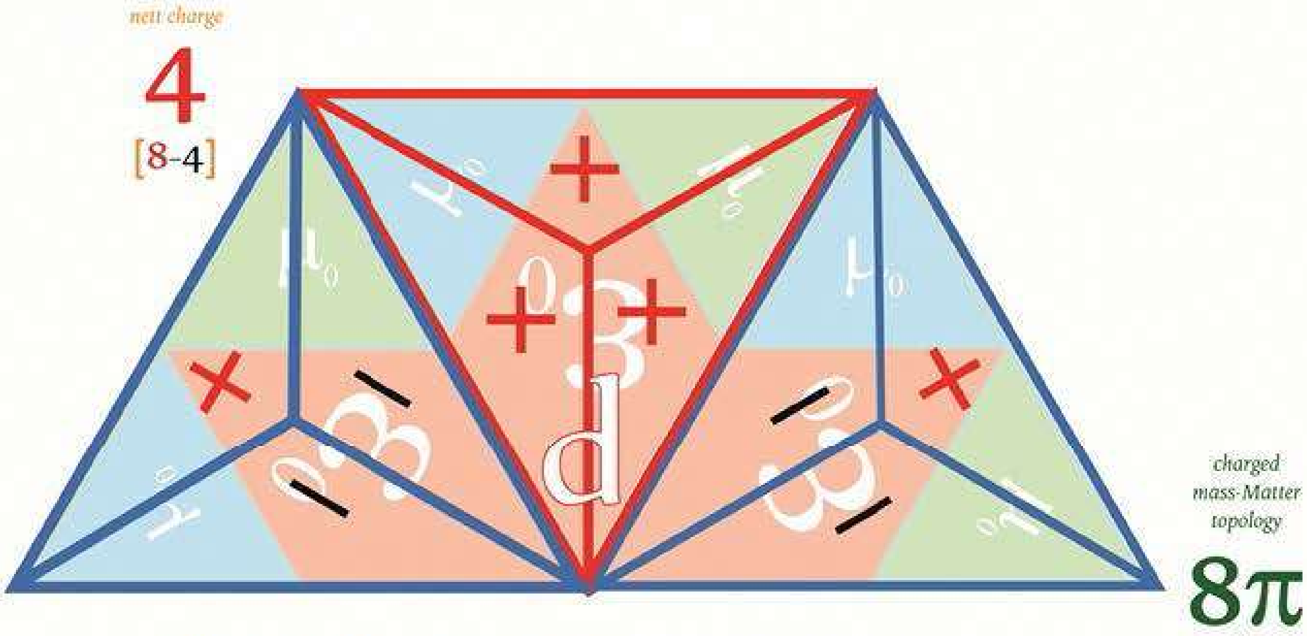
nett charge
4
[8-4]
d
charged mass-Matter topology
8π

Quarks have 12 charged fascia - ***mass-energy geometries***

8 Positive 4 Negative

Quarks have octahedral mass-Matter topologies

[+1/3 elementary charge]

Copyright ABRAHAM [2008] - all rights reserved

up quark

+2/3

quark octahedral topology

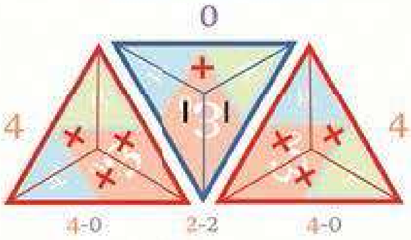

attractive dodecyon geometry

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

EM Field · Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

dodecyon · ElectroMagnetic · mass · velocity

1. **up** quark
2. **charmed** quark
3. **top** quark

u

nett charge

8

[10-2]

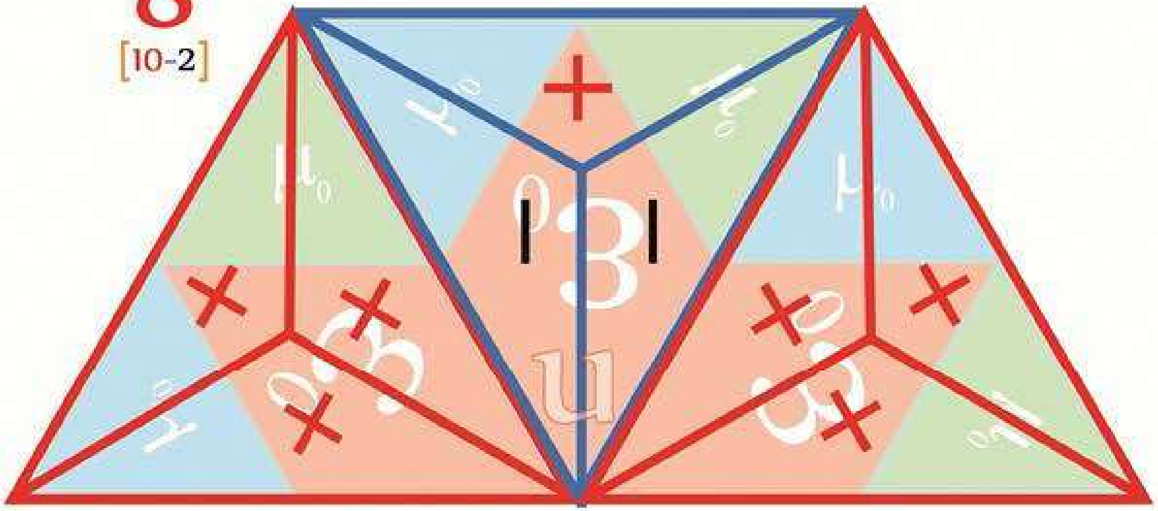

charged mass-Matter topology

8π

Quarks have 12 charged fascia - ***mass-energy geometries***

10 Positive 2 Negative

Quarks have octahedral mass-Matter topologies

[+2/3 elementary charge]

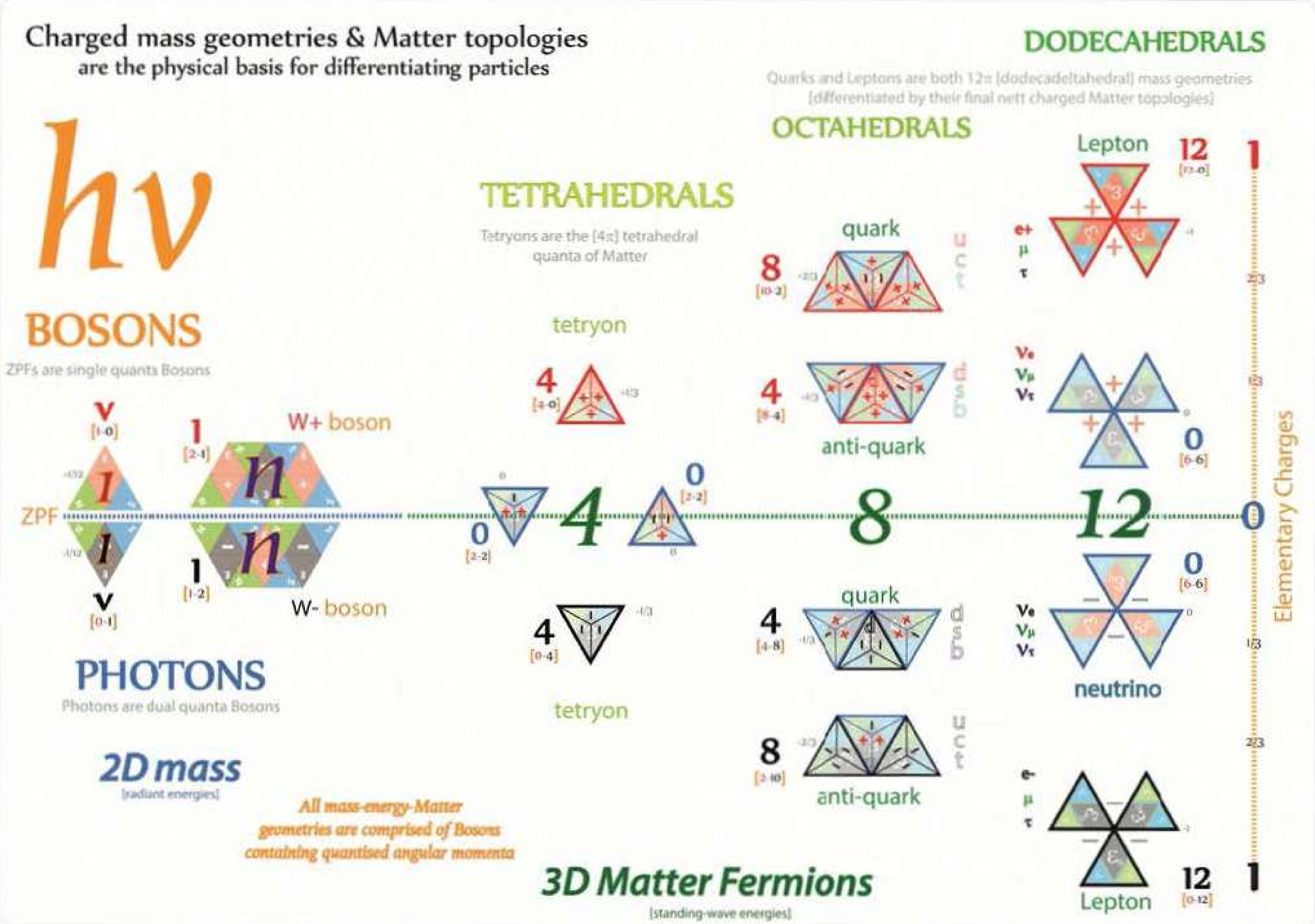

Charged mass geometries & Matter topologies are the physical basis for differentiating particles

hv

BOSONS

ZPFs are single quanta Bosons

PHOTONS

Photons are dual quanta Bosons

2D mass

[radiant energies]

All mass-energy-Matter geometries are comprised of Bosons containing quantised angular momenta

TETRAHEDRALS

Tetryons are the [4π] tetrahedral quanta of Matter

OCTAHEDRALS

DODECAHEDRALS

Quarks and Leptons are both 12π [dodecadeltahedral] mass geometries [differentiated by their final nett charged Matter topologies]

3D Matter Fermions

[standing-wave energies]

ν
[1-0]
ν
[0-1]
ZPF
1
[2-1]
W+ boson
1
[1-2]
W- boson
tetryon
4
[4-0]
+1/3
0
[2-2]
4
0
[2-2]
4
[0-4]
-1/3
tetryon
quark
8
[10-2]
+2/3
u
c
t
4
[8-4]
+1/3
d
s
b
anti-quark
8
quark
4
[4-8]
-1/3
d
s
b
8
[2-10]
-2/3
u
c
t
anti-quark
Lepton
12
[12-0]
1
e+
μ
τ
2/3
νe
νμ
ντ
1/3
0
[6-6]
12
0
0
[6-6]
νe
νμ
ντ
neutrino
1/3
2/3
e-
μ
τ
12
[0-12]
Lepton
1
Elementary Charges

divergent negative E-fields

electron

convergent negative E-fields

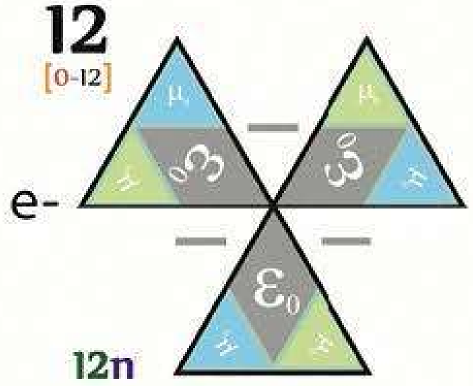

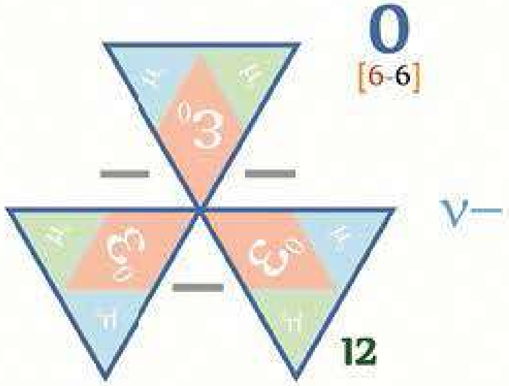

divergent negative E-fields

electron neutrino

convergent positive E-fields

Leptons have 12 charged mass-energy fascia geometries

Repulsive STRONG force creates Lepton topologies

12π

Tetryonic charged fascias creates the Matter topologies of all Leptons

Repulsive STRONG force creates Neutrino topologies

Leptons have dodecadeltahedral mass-Matter topologies

divergent positive E-fields

positron

convergent positive E-fields

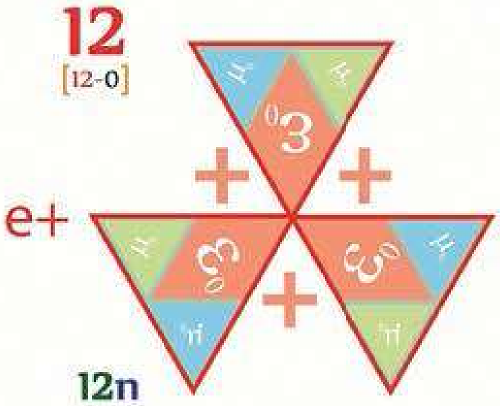

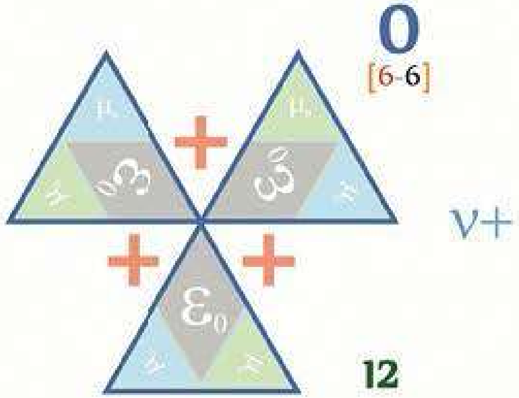

divergent positive E-fields

positron neutrino

convergent negative E-fields

Charged Leptons

EM Field Planck quanta

$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

Dodecyon ElectroMagnetic mass velocity

positron

net charge
12
[12-0]

e+
12

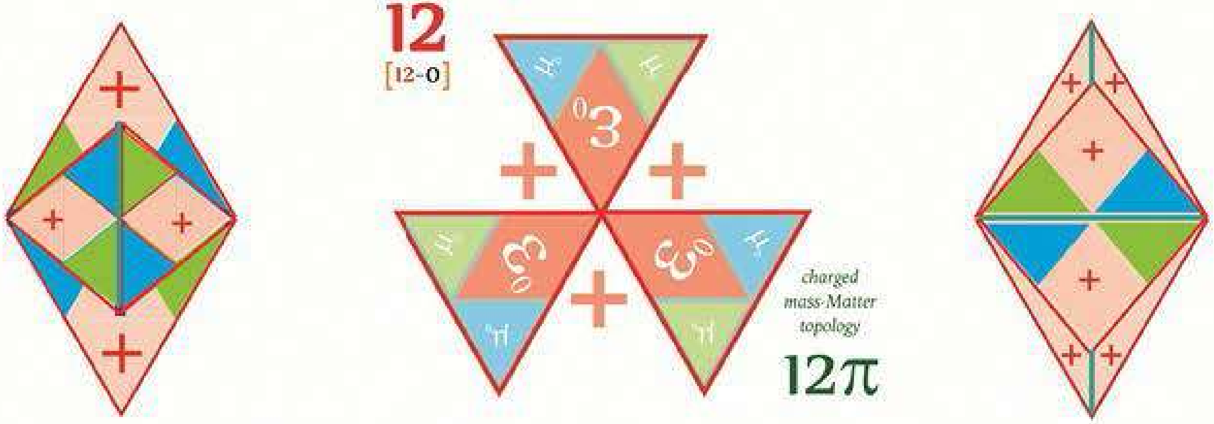

+
+
+
+
+
+
+
+
+
+
+
+
+
ε_0
ε_0
ε_0
μ_0

charged
mass-Matter
topology
12π

1. **positron**
2. **anti-muon**
3. **anti-tau**

Leptons have 12 charged mass-energy fascia geometries

12 positive charge fascia or 12 Negative charge fascia

repulsive dodecyon geometries

lepton topologies

Charged Leptons have dodecadeltahedral mass-Matter topologies

[1 elementary charge]

net charge
12
[0-12]

e-
12

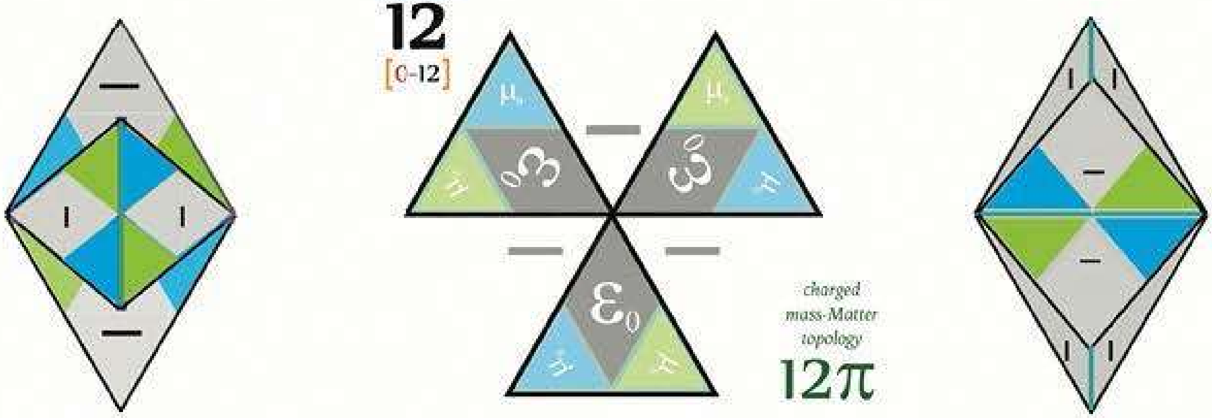

–
–
–
–
–
–
–
–
–
–
–
–
–
ε_0
ε_0
ε_0
μ_0

charged
mass-Matter
topology
12π

electron

1. **electron**
2. **muon**
3. **tau**

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

positron

repulsive charge dodecyon geometry

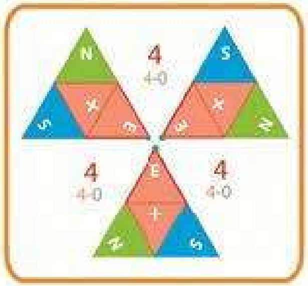
N
S
4
4-0
4
4-0
4
4-0

lepton mass-Matter topology

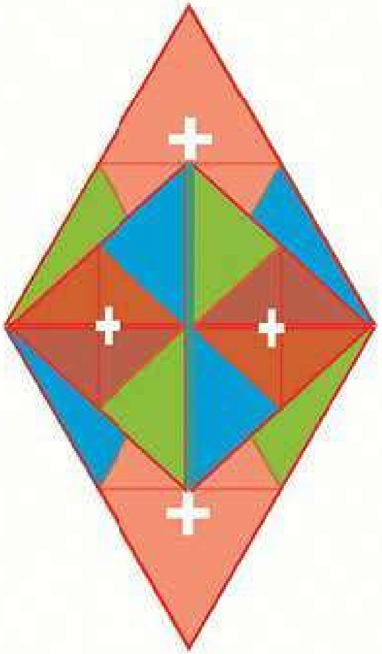

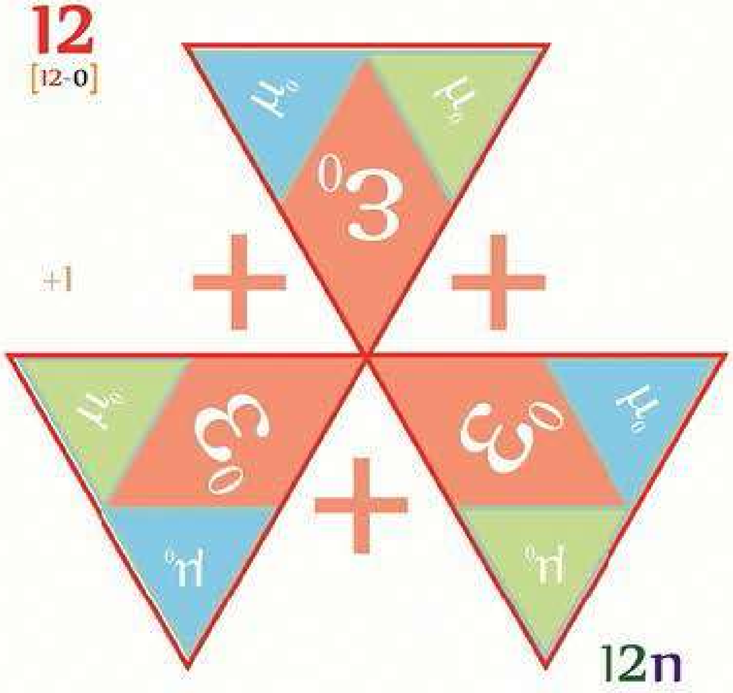
12
[12-0]
+1
μ0
μ0
ε0
μ0
ε0
μ0
ε0
μ0
μ0
+
+
+
12n

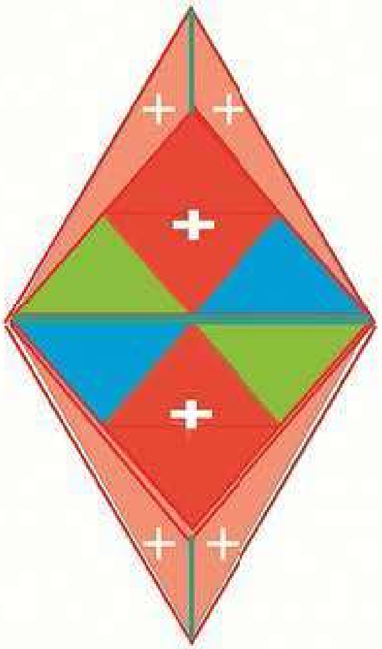

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Dodecyon ElectroMagnetic mass velocity

1. positron
2. anti-muon
3. anti-tau

Copyright ABRAHAM [2008] - all rights reserved

Neutral Leptons

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Dodecyon ElectroMagnetic mass velocity

ν-
0

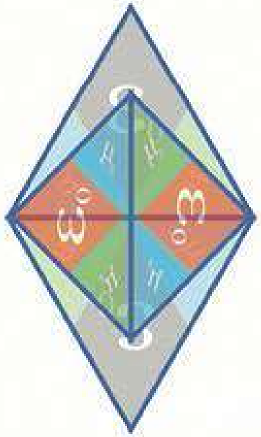

nett charge
0
[6 6]

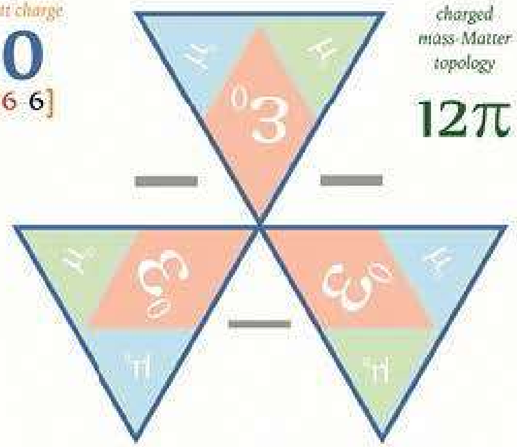

charged
mass-Matter
topology
12π

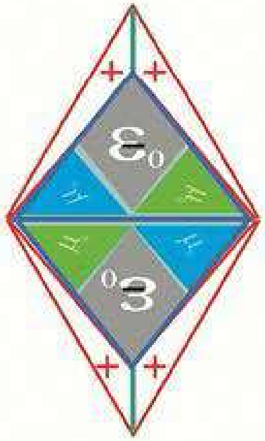

1. electron neutrino
2. muon neutrino
3. tau neutrino

Neutrinos have 12 charged mass-energy fascia geometries

6 Positive & 6 Negative

repulsive dodecyon geometries

lepton topologies

Neutrinos have neutral dodecadeltahedral mass-Matter topologies

[0 elementary charge]

ν+
0

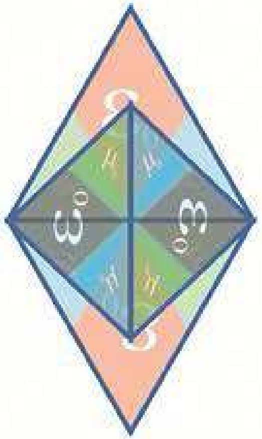

nett charge
0
[6-6]

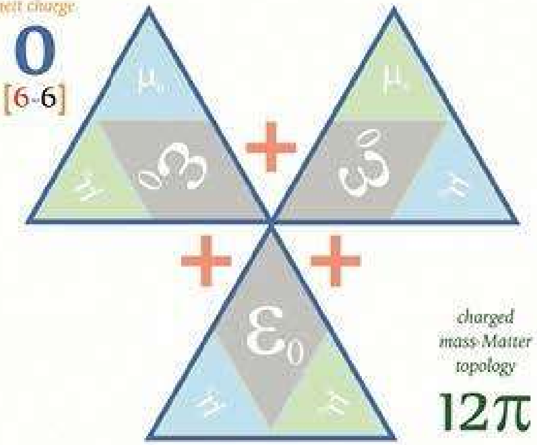

charged
mass-Matter
topology
12π

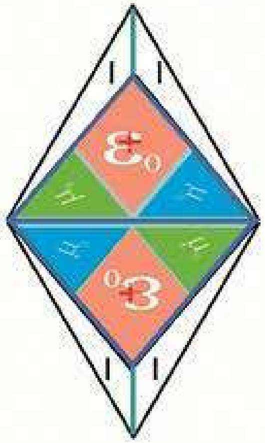

1. positron neutrino
2. anti- muon neutrino
3. anti-tau neutrino

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

Copyright ABRAHAM [2008] - all rights reserved

electron

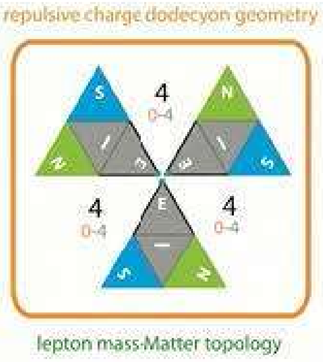
repulsive charge dodecyon geometry
4
0-4
4
0-4
4
0-4
S
N
E
lepton mass-Matter topology

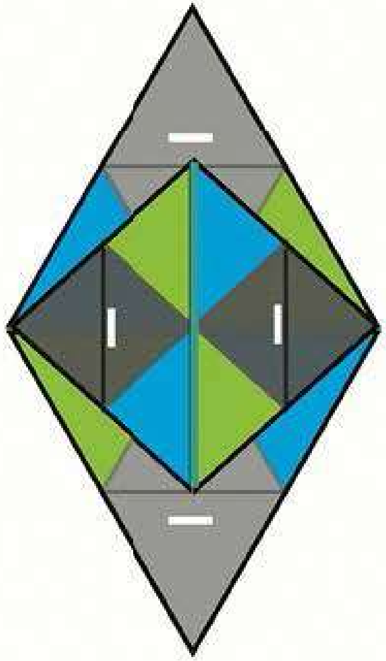

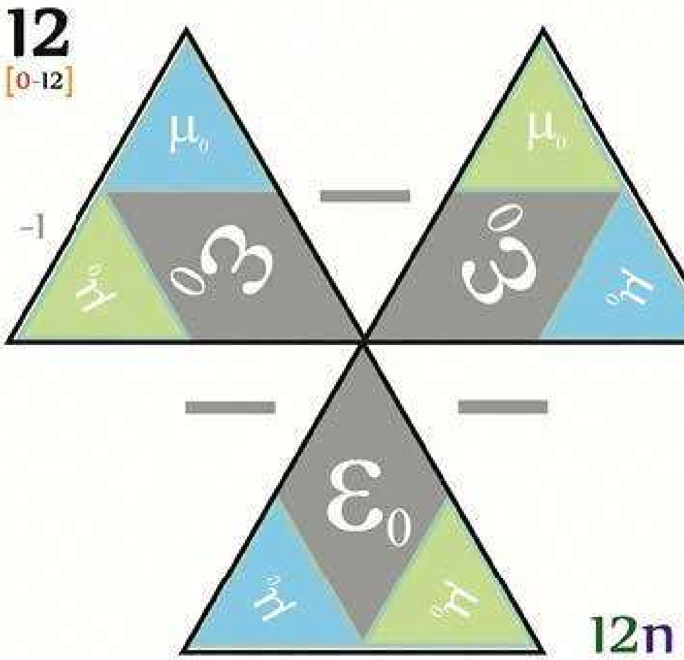
12
[0-12]
-1
μ_0
ε_0
12n

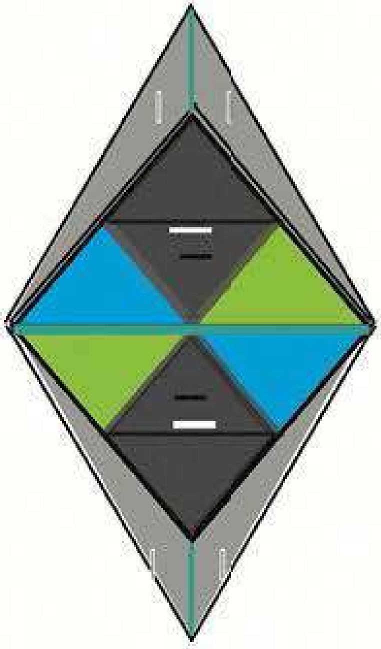

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Dodecyon ElectroMagnetic mass velocity

1. electron
2. muon
3. tau

Copyright ABRAHAM [2008] - all rights reserved

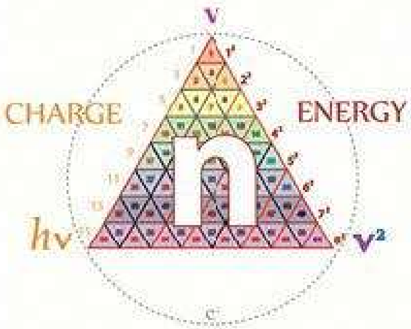
v
CHARGE
ENERGY
n
hv
v²
c

Leptronic Quantum levels

[KEM field mass-energy geometry]

v^2 Total quanta		n Electron quantum level
1.2e20	1e19	1
4.8e20		4
1.08e21		9
1.92e21		16
3.00e21		25
4.32e21		36
5.88e21		49
7.68e21		64

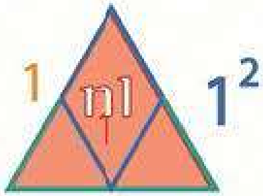

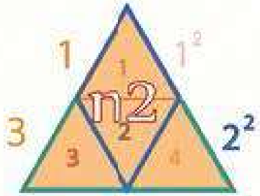

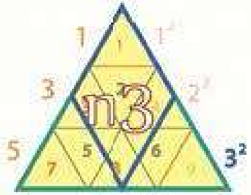
1
n3
1²
3
2²
5
3²

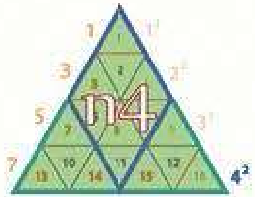
1
n4
1²
3
2²
5
3²
7
4²

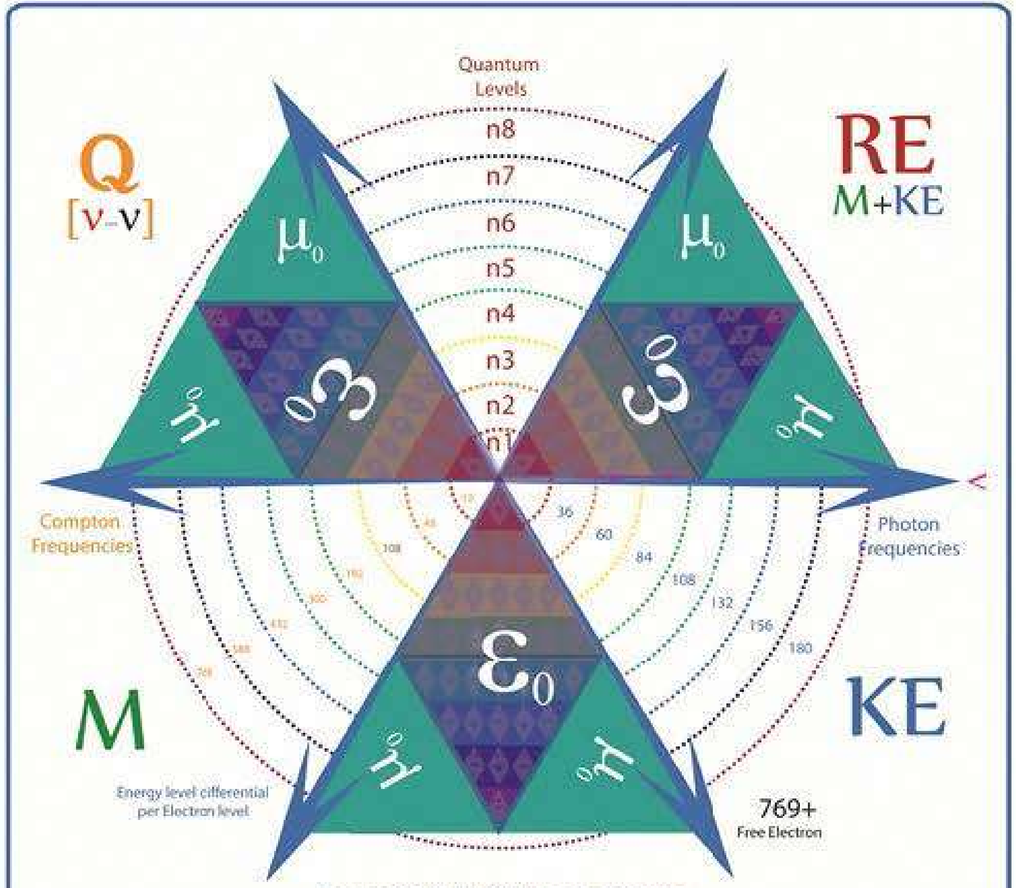
Quantum Levels
n8
n7
n6
n5
n4
n3
n2
n1
Q
[ν-ν]
RE
M+KE
μ₀
ε₀
Compton Frequencies
Photon Frequencies
v
36
60
84
108
132
156
180
M
KE
Energy level differential per Electron level
769+
Free Electron

Leptronic KEM field mass-energies

$$p^2 = KEM = Mv^2$$

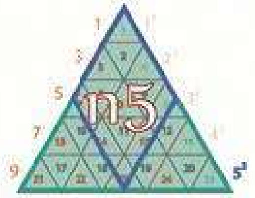

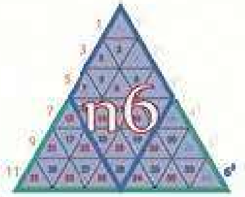

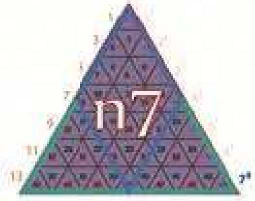

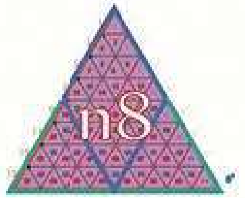

Illustrative schema only: All Leptonic quantum levels have the same equilateral KEM geometry as the compton frequency of the KEM field increases the wavelengths of the quanta decrease

Tetryonics 08.06 - Leptronic quantum levels

Copyright ABRAHAM [2008] - all rights reserved

$P^2 = KEM = Mv^2$

Electron energy levels

As the total Energy content of particles increase
the Quantum levels of their KEM fields increase as well,
in turn this is reflected in their associated
Compton Frequencies and deBroglie wavelengths

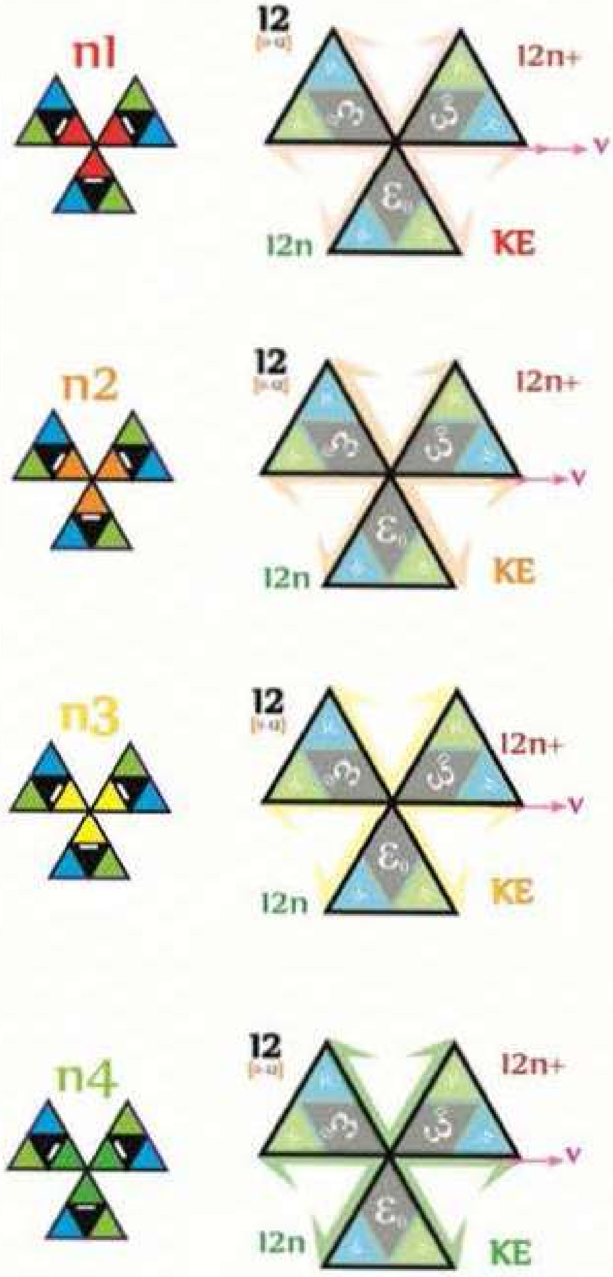
n1
12
[0-12]
12n+
v
12n
KE
n2
12
[0-12]
12n+
v
12n
KE
n3
12
[0-12]
12n+
v
12n
KE
n4
12
[0-12]
12n+
v
12n
KE

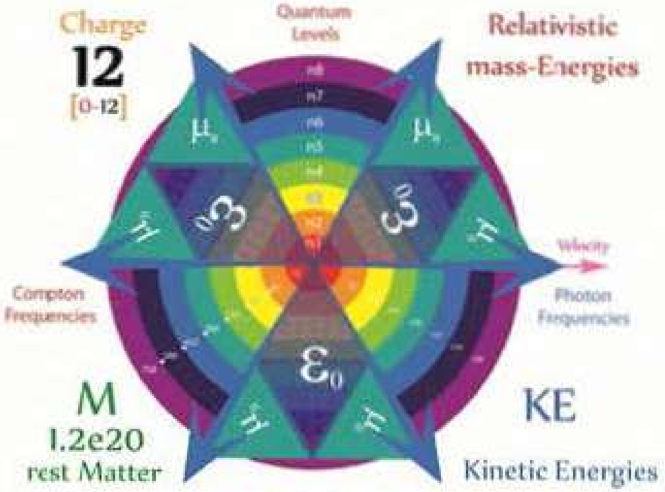
Charge
12
[0-12]
Quantum Levels
Relativistic mass-Energies
n8
n7
n6
n5
n4
n3
n2
n1
μ_0
μ_0
ε_0
ε_0
ε_0
Velocity
Compton Frequencies
Photon Frequencies
M
1.2e20
rest Matter
KE
Kinetic Energies

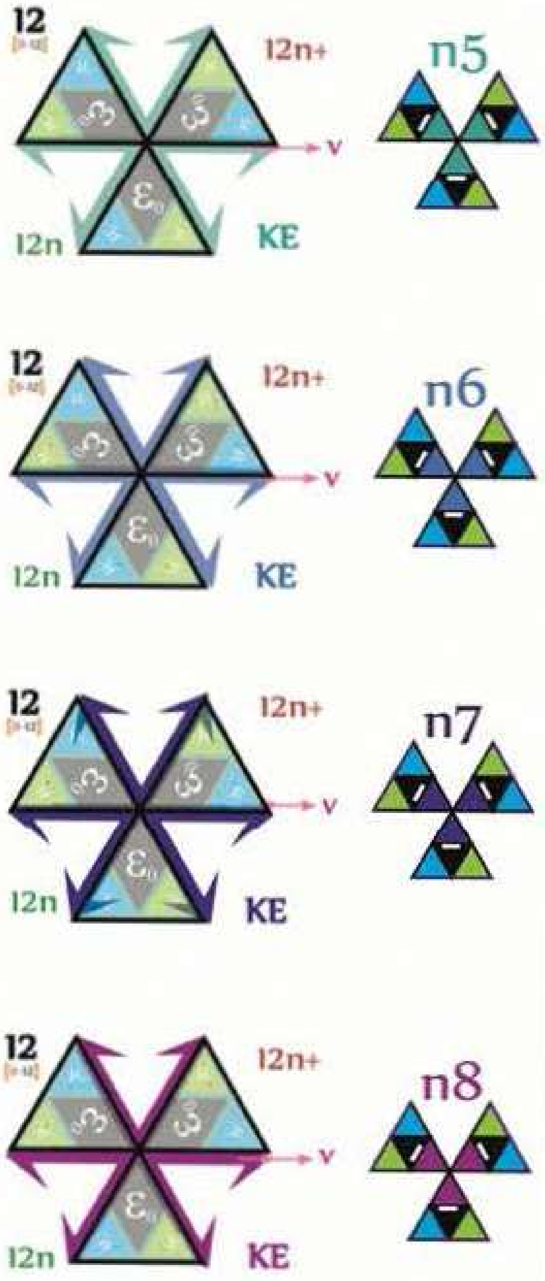
12
[0-12]
12n+
n5
v
12n
KE
12
[0-12]
12n+
n6
v
12n
KE
12
[0-12]
12n+
n7
v
12n
KE
12
[0-12]
12n+
n8
v
12n
KE

All Tetryonic geometries [EM fields & Matter]
absorb and release energy exponentially
in equilateral Quantum steps

The ratio of the area of a sphere

$$A = 4\pi r^2 = \pi d^2.$$

to

the area of a 3 Tetryhedron Lepton shape

$$A = 4\,A_0 = \sqrt{3}a^2$$

is

2.418399152

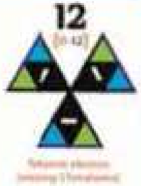

The spherical model of an electron's topology is wrong

Tetryonics 08.07 - Electron energy levels

Copyright ABRAHAM [2008] - all rights reserved

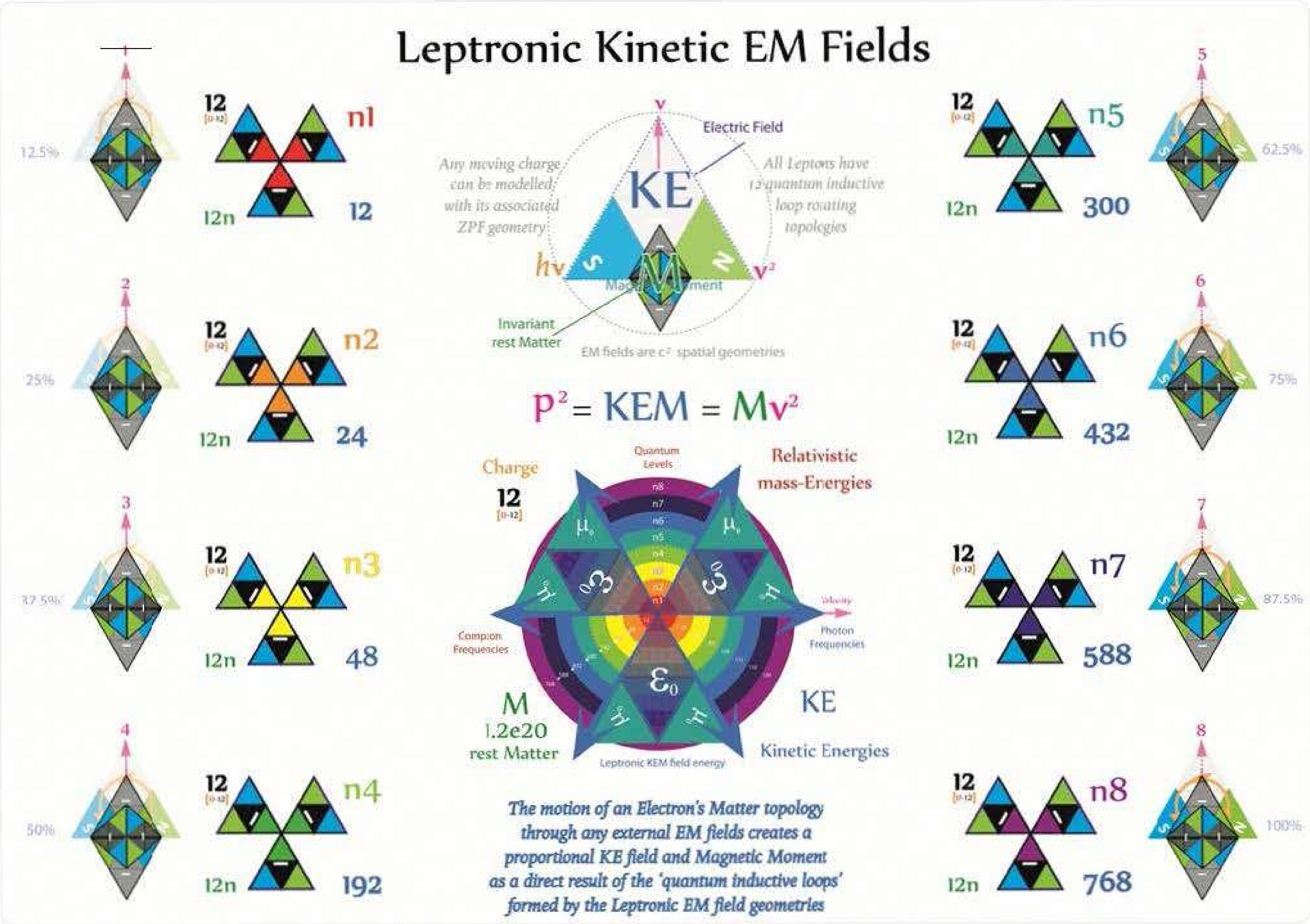

Leptronic Kinetic EM Fields

Any moving charge can be modelled with its associated ZPF geometry

All Leptons have 12 quantum inductive loop rotating topologies

EM fields are c^2 spatial geometries

$$P^2 = KEM = Mv^2$$

The motion of an Electron's Matter topology through any external EM fields creates a proportional KE field and Magnetic Moment as a direct result of the 'quantum inductive loops' formed by the Leptronic EM field geometries

1
12.5%
12
[0-12]
n1
12n
12
2
25%
12
[0-12]
n2
12n
24
3
37.5%
12
[0-12]
n3
12n
48
4
50%
12
[0-12]
n4
12n
192
v
Electric Field
KE
hv
S
M
N
v²
Magnetic Moment
Invariant rest Matter
5
62.5%
12
[0-12]
n5
12n
300
6
75%
12
[0-12]
n6
12n
432
7
87.5%
12
[0-12]
n7
12n
588
8
100%
12
[0-12]
n8
12n
768
Charge
12
[0-12]
Quantum Levels
n8
n7
n6
n5
n4
n3
n2
n1
Relativistic mass-Energies
μ₀
ε₀
Velocity
Photon Frequencies
Compton Frequencies
M
1.2e20
rest Matter
KE
Kinetic Energies
Leptronic KEM field energy

Tetryonics 08.08 - Leptronic KEM fields

Copyright ABRAHAM [2008] - all rights reserved

Kinetic field modelling

Tetryonic
Positive Charge
KEM field geometry

RE

Mv^2
2D
Kinetic Energy field

$E = mc^2$
3D
rest mass-Matter topology

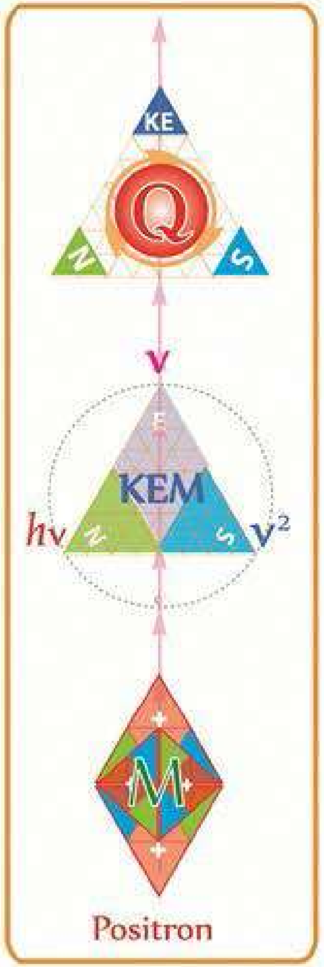

Positron

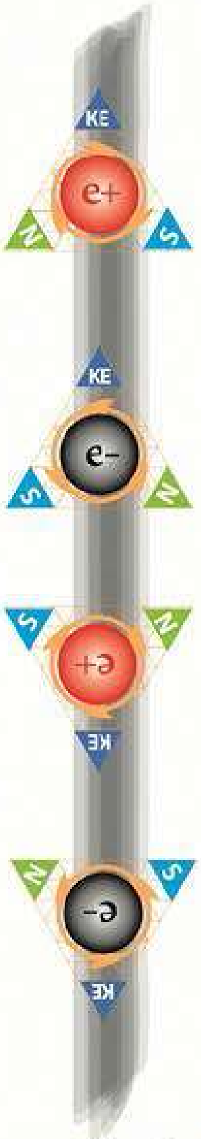

Similar charges moving
in the same direction
produce attractive
magnetic dipoles

Opposite charges moving
in the same direction
produce repulsive
magnetic dipoles

F_A

Opposite charges moving
in different directions
produce attractive
magnetic dipoles

Similar charges moving
in opposing directions
produce repulsive
magnetic dipoles

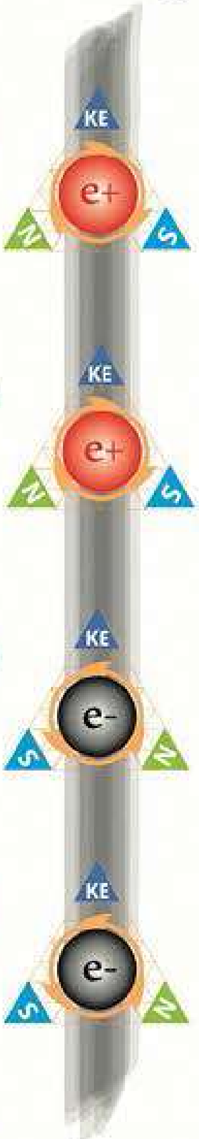

Tetryonic
Negative Charge
KEM field geometry

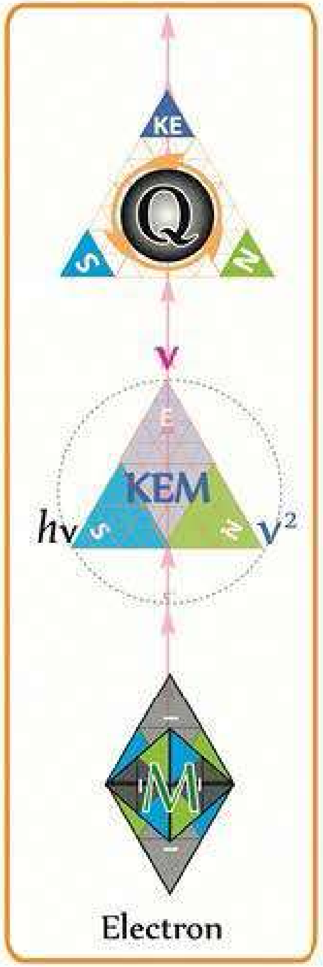

Electron

RE

Mv^2
2D
Kinetic Energy field

$E = mc^2$
3D
rest mass-Matter topology

The KEM fields of all
Matter particles in motion
can be modelled using
ZPF EM geometries

Kinetic EM field geometry determines the Magnetic Moments of all particles in motion

The KEM E-field is reflective
of the nett particle charge
and forms a directional
vector of motion

Lepton Families & Generations

ν
KE
M
hν
ν²
rest Matter is invariant
c

Charged Matter topologies in motion creates Kinetic EM field geometries

Charged mass-energy geometries form elementary particle topologies

12π

$$KEM = Mv^2$$

Total relativistic mass-energies [rest Matter + KEM]

All high mass-energy particles 'decay' into low mass-energy particles by emitting photons

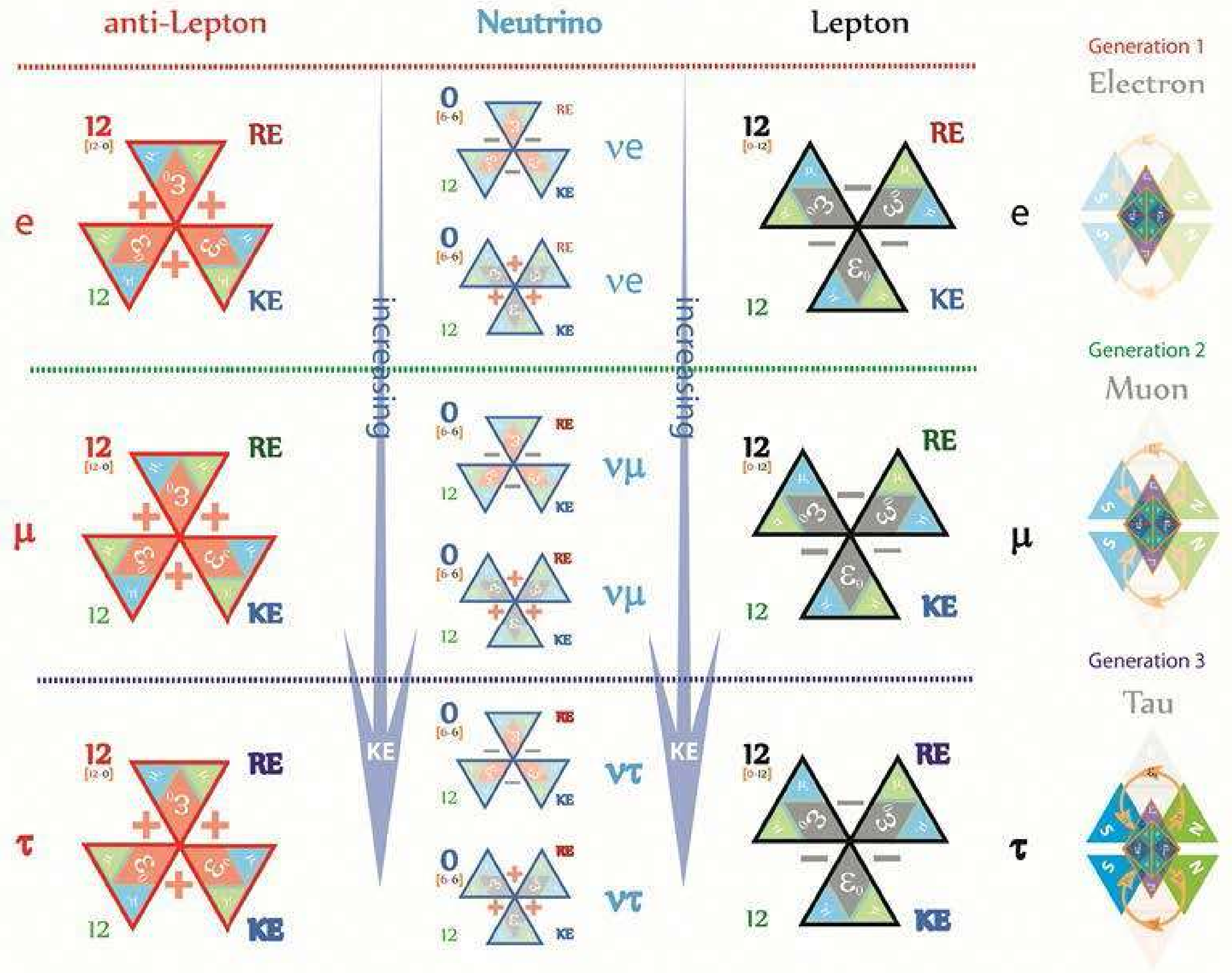
anti-Lepton
Neutrino
Lepton
Generation 1
Electron
e
12
[12-0]
RE
12
KE
0
[6-6]
RE
12
KE
νe
0
[6-6]
RE
12
KE
νe
12
[0-12]
RE
12
KE
e
Generation 2
Muon
μ
12
[12-0]
RE
12
KE
0
[6-6]
RE
12
KE
νμ
0
[6-6]
RE
12
KE
νμ
12
[0-12]
RE
12
KE
μ
Generation 3
Tau
τ
12
[12-0]
RE
12
KE
0
[6-6]
RE
12
KE
ντ
0
[6-6]
RE
12
KE
ντ
12
[0-12]
RE
12
KE
τ
increasing
KE
increasing
KE

Note: All Lepton geometries are the same size - only the energy density of the EM field changes

Lepton Helicity and Chirality

Particle Spin and Handedness is always referenced with respect to the direction of nett total system linear Momentum

Particle Spin

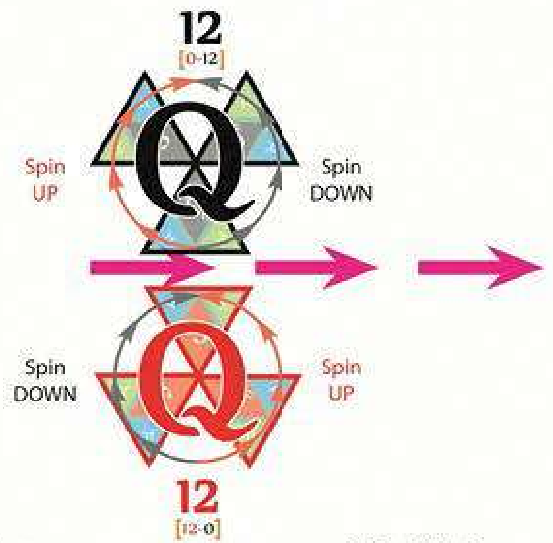
12
[0-12]
Spin
UP
Spin
DOWN
Spin
DOWN
Spin
UP
12
[12-0]

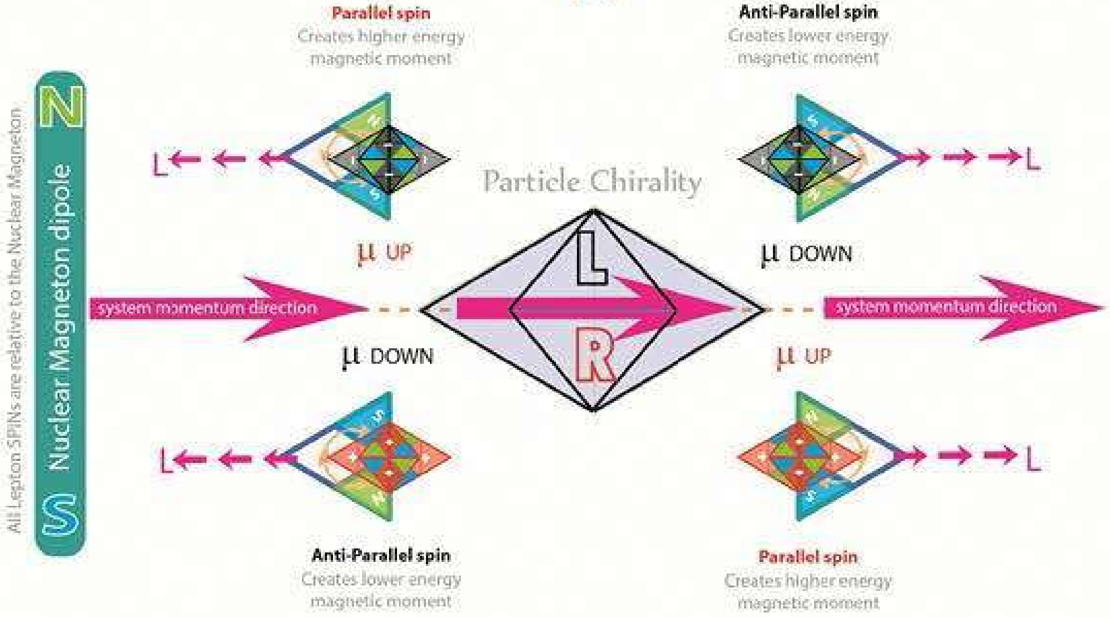
Parallel spin
Creates higher energy magnetic moment
Anti-Parallel spin
Creates lower energy magnetic moment
All Lepton SPINs are relative to the Nuclear Magneton
N
Nuclear Magneton dipole
S
L
L
μ UP
Particle Chirality
μ DOWN
system momentum direction
system momentum direction
μ DOWN
μ UP
L
R
L
L
Anti-Parallel spin
Creates lower energy magnetic moment
Parallel spin
Creates higher energy magnetic moment

Particle Chirality

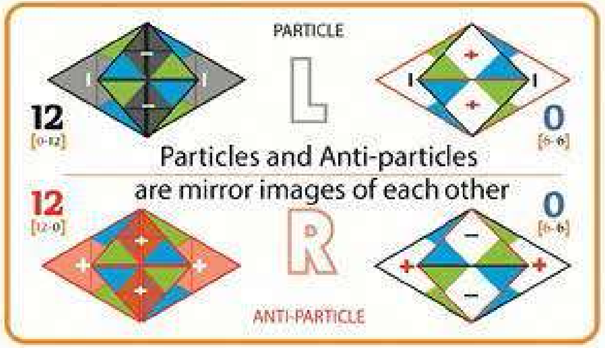
PARTICLE
12
[0-12]
L
0
[6-6]
12
[12-0]
R
0
[6-6]
ANTI-PARTICLE

Particles and Anti-particles are mirror images of each other

Particle Helicity

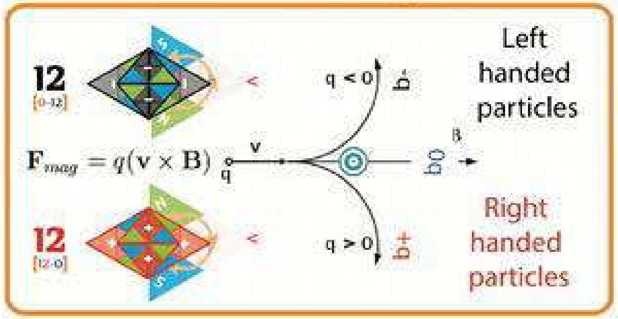
12
[0-12]
12
[12-0]
$F_{mag} = q(v \times B)$
v
q
q < 0
q > 0
β-
β+
B
Left handed particles
Right handed particles

Lorentz Force and Transformation

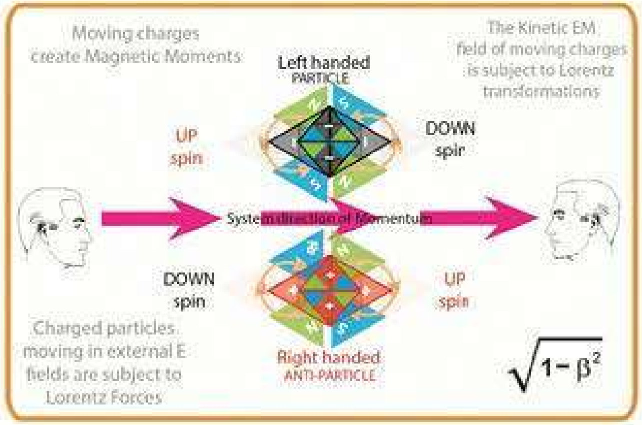
Moving charges create Magnetic Moments
The Kinetic EM field of moving charges is subject to Lorentz transformations
Left handed PARTICLE
UP spin
DOWN spin
System direction of Momentum
DOWN spin
UP spin
Charged particles moving in external E fields are subject to Lorentz Forces
Right handed ANTI-PARTICLE
$\sqrt{1-\beta^2}$

'Point Charges'

The only true 'point charges' are Zero Point Fields

Positive Particles

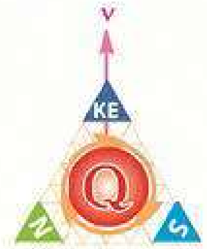

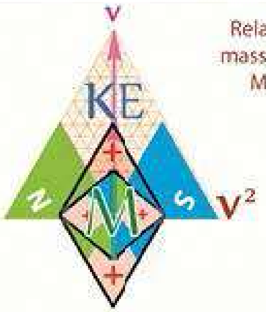

Relativistic mass-energy Model

Kinetic Energies

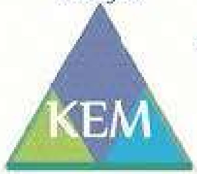

Magnetic moment

velocity related mass-energy geometry

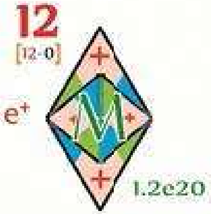
12
[12-0]
e⁺
M
1.2e20

Invariant rest Matter topology

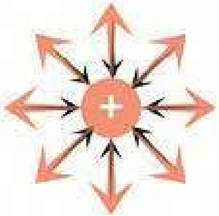

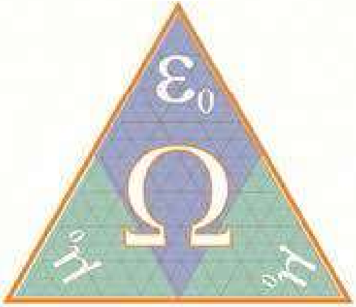

Static charge particles have neutral M Fields
Charges in motion have magnetic moments

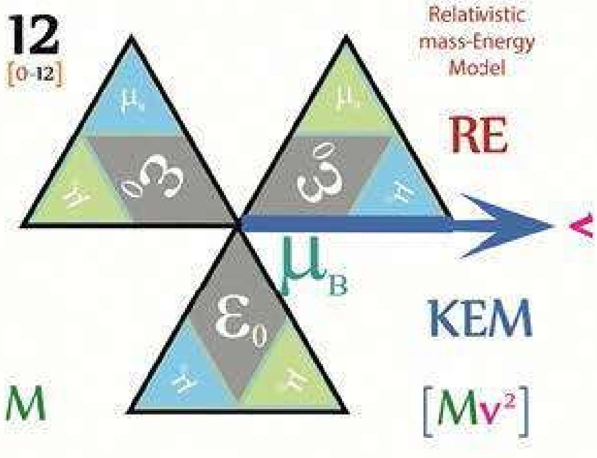
12
[0-12]
μ₀
ε₀
μ₀
ε₀
μ₀
RE
v
μB
ε₀
KEM
M
[Mv²]

Relativistic mass-Energy Model

It is the Kinetic EM field geometry of charged Matter topologies in motion that produces Magnetic Moments

[not a relativistic distortion of speherical charge topologies]

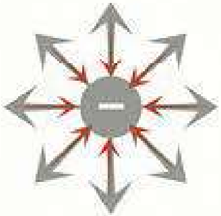

Negative Particles

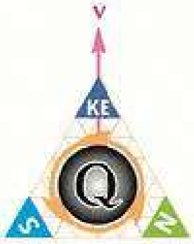

Relativistic mass-energy Model

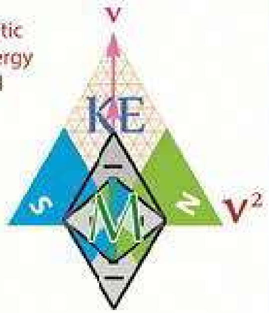

Knetic Energies

velocity related mass-energy geometry

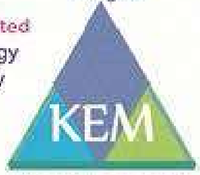

Magnetic moment

Invariant rest Matter topology

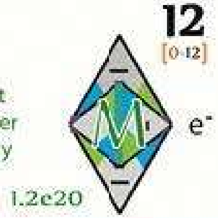
12
[0-12]
M
e⁻
1.2e20

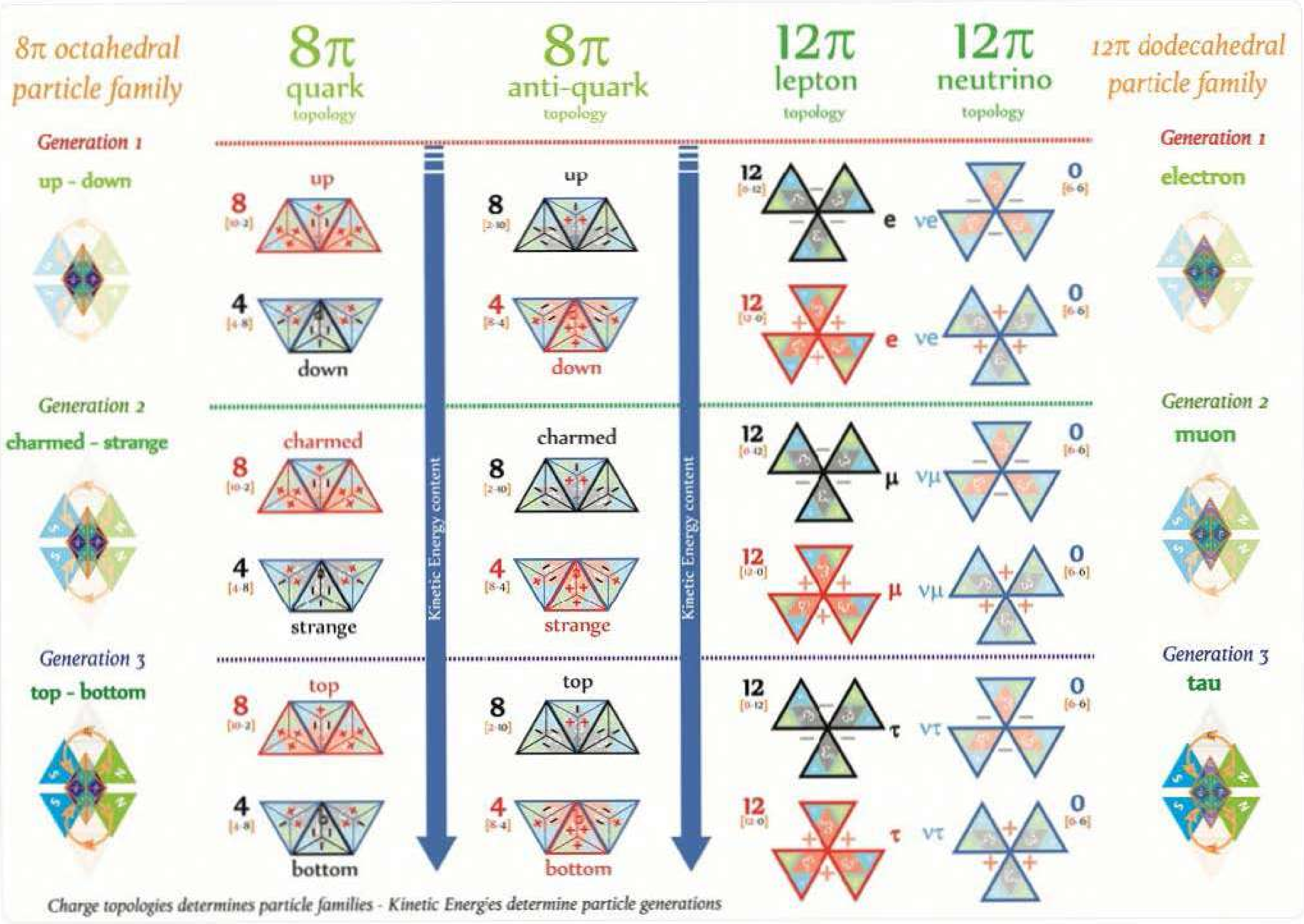
8π octahedral particle family
8π quark topology
8π anti-quark topology
12π lepton topology
12π neutrino topology
12π dodecahedral particle family
Generation 1
up - down
up
8 [10-2]
4 [4-8]
down
up
8 [2-10]
4 [8-4]
down
12 [0-12]
e
ve
0 [6-6]
12 [12-0]
e
ve
0 [6-6]
Generation 1
electron
Generation 2
charmed - strange
charmed
8 [10-2]
4 [4-8]
strange
charmed
8 [2-10]
4 [8-4]
strange
12 [0-12]
μ
vμ
0 [6-6]
12 [12-0]
μ
vμ
0 [6-6]
Generation 2
muon
Generation 3
top - bottom
top
8 [10-2]
4 [4-8]
bottom
top
8 [2-10]
4 [8-4]
bottom
12 [0-12]
τ
vτ
0 [6-6]
12 [12-0]
τ
vτ
0 [6-6]
Generation 3
tau
Kinetic Energy content
Kinetic Energy content
Charge topologies determines particle families - Kinetic Energies determine particle generations

Tetryonics 08.13 - Quark & Lepton families

Copyright ABRAHAM [2008] - all rights reserved

Mesons

are 24π subatomic mass-Matter particles composed of one quark and one antiquark

$$24\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM Field
Planck quanta
Mesons
ElectroMagnetic
mass
velocity

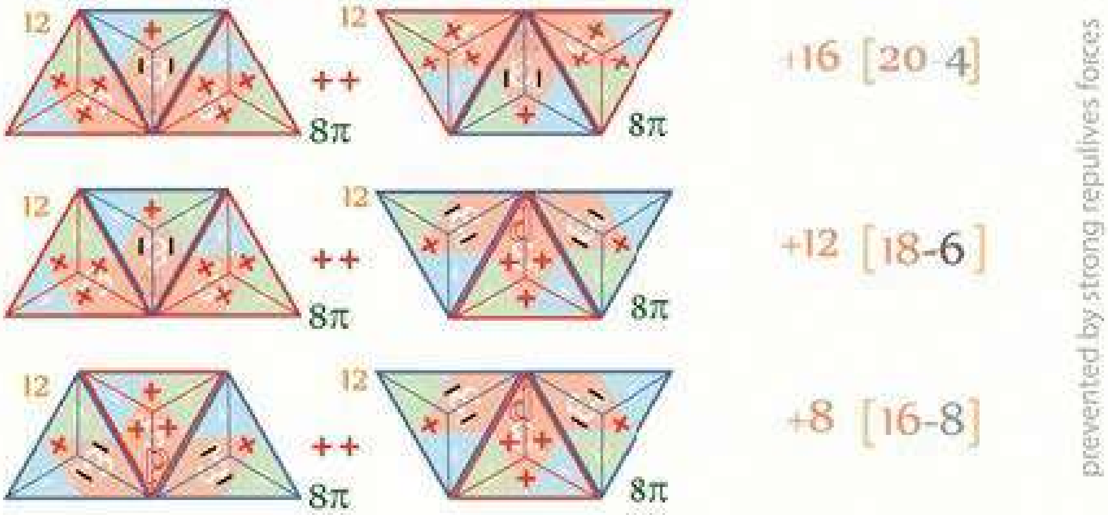

12
12
++
8π
8π
+16 [20-4]
12
12
++
8π
8π
+12 [18-6]
12
12
++
8π
8π
+8 [16-8]
prevented by strong repulives forces

24π charged mass-energy fascia geometries

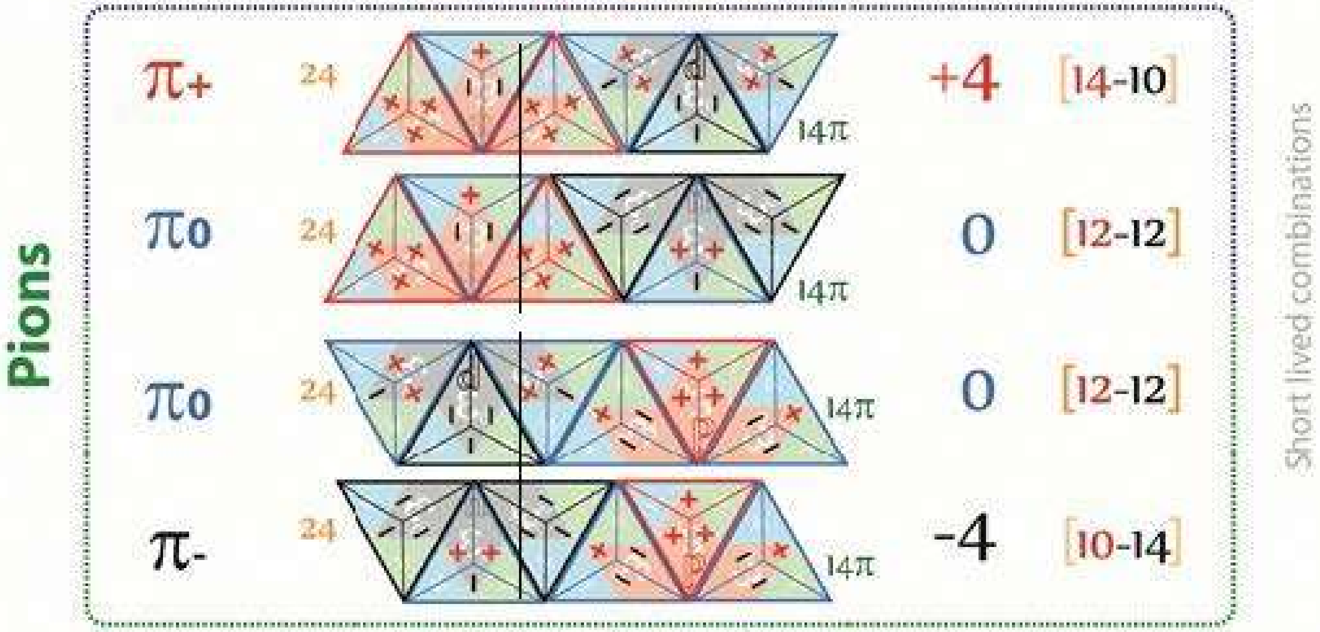

Pions
π+
24
14π
+4
[14-10]
πo
24
14π
0
[12-12]
πo
24
14π
0
[12-12]
π-
24
14π
-4
[10-14]
Short lived combinations

14π final mass-Matter topologies

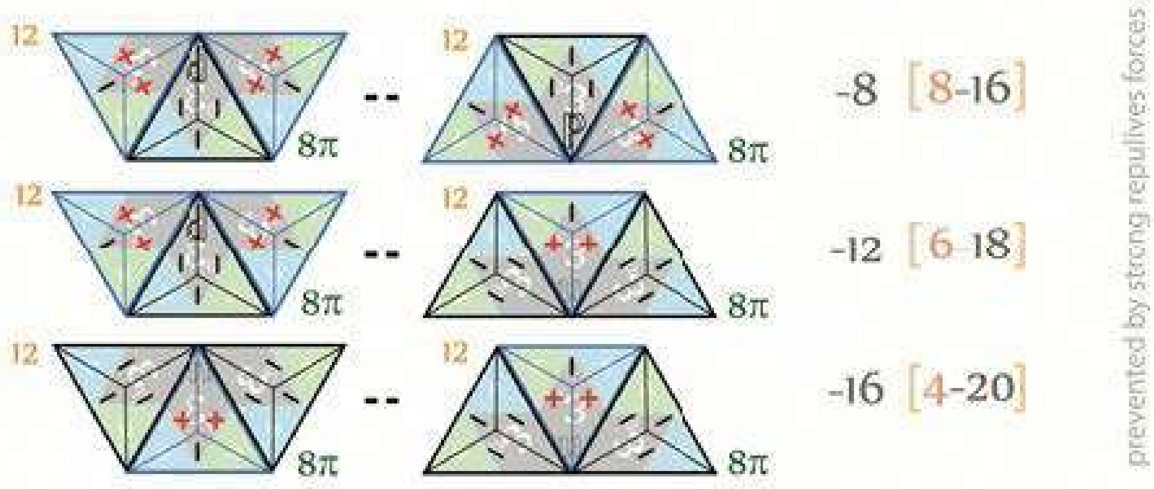

12
12
--
8π
8π
-8 [8-16]
12
12
--
8π
8π
-12 [6-18]
12
12
--
8π
8π
-16 [4-20]
prevented by strong repulives forces

Copyright ABRAHAM [2008] - all rights reserved

PION decay

24π dual -quark mass-energy geometries

12π lepton mass-energy geometries

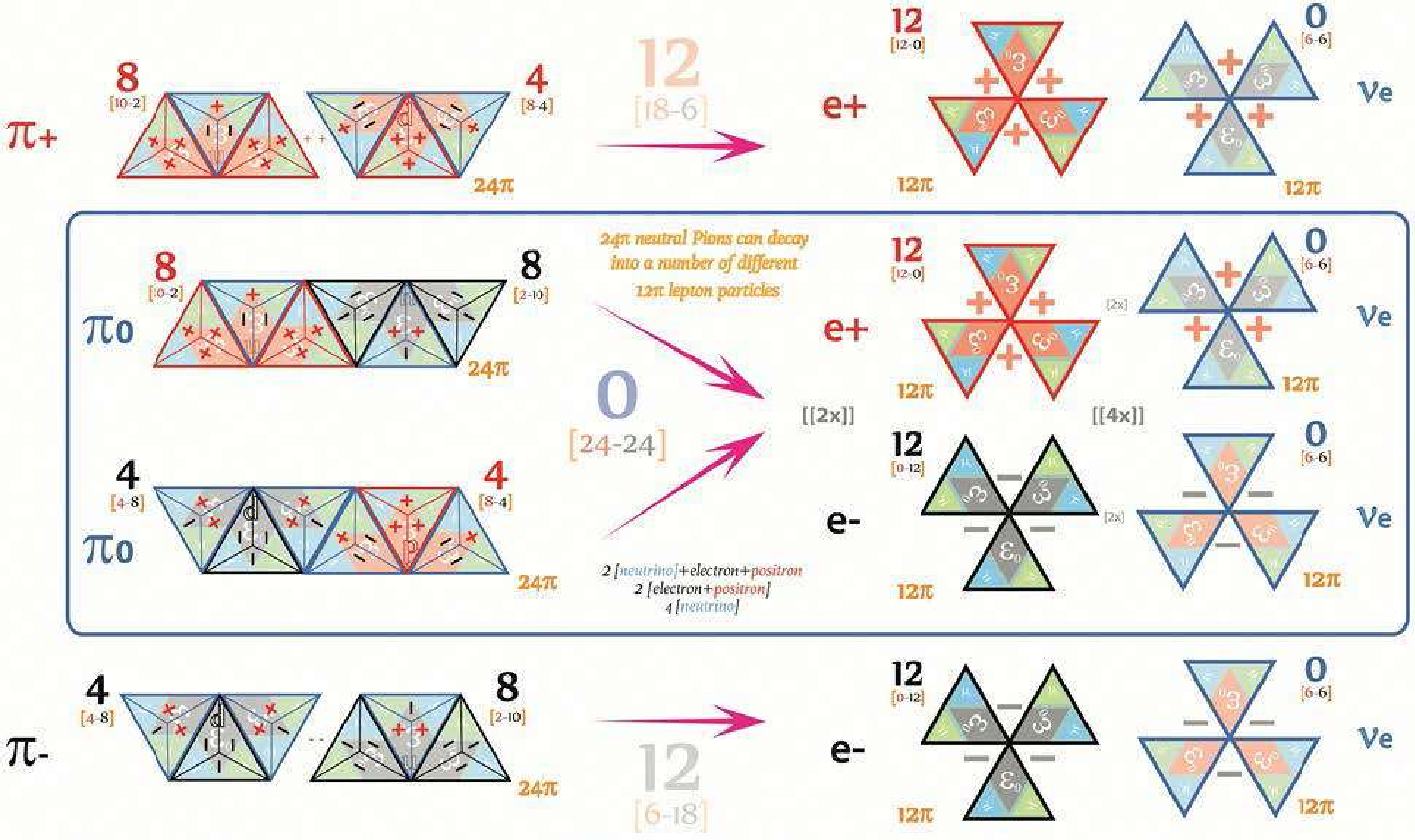
π+
8
[10-2]
4
[8-4]
24π
12
[18-6]
e+
12
[12-0]
12π
0
[6-6]
νe
12π
πo
8
[0-2]
8
[2-10]
24π
24π neutral Pions can decay
into a number of different
12π lepton particles
0
[24-24]
[[2x]]
[[4x]]
[2x]
12π
4
[4-8]
4
[8-4]
πo
24π
e-
12
[0-12]
[2x]
12π
2 [neutrino]+electron+positron
2 [electron+positron]
4 [neutrino]
π-
4
[4-8]
8
[2-10]
24π
12
[6-18]
12π

$24\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

EM Field Planck quanta

Mesons ElectroMagnetic mass velocity

24π dual dodecyon geometries

Despite having 24π geometries Mesons have same nett Charges as 36π Baryons

36π tri-dodecyon geometries

$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$ **+** $12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

Lepton - Neutrino generational pairing produced is dependent on the mass-energy levels of the interacting Pions

Copyright ABRAHAM [2008] - all rights reserved

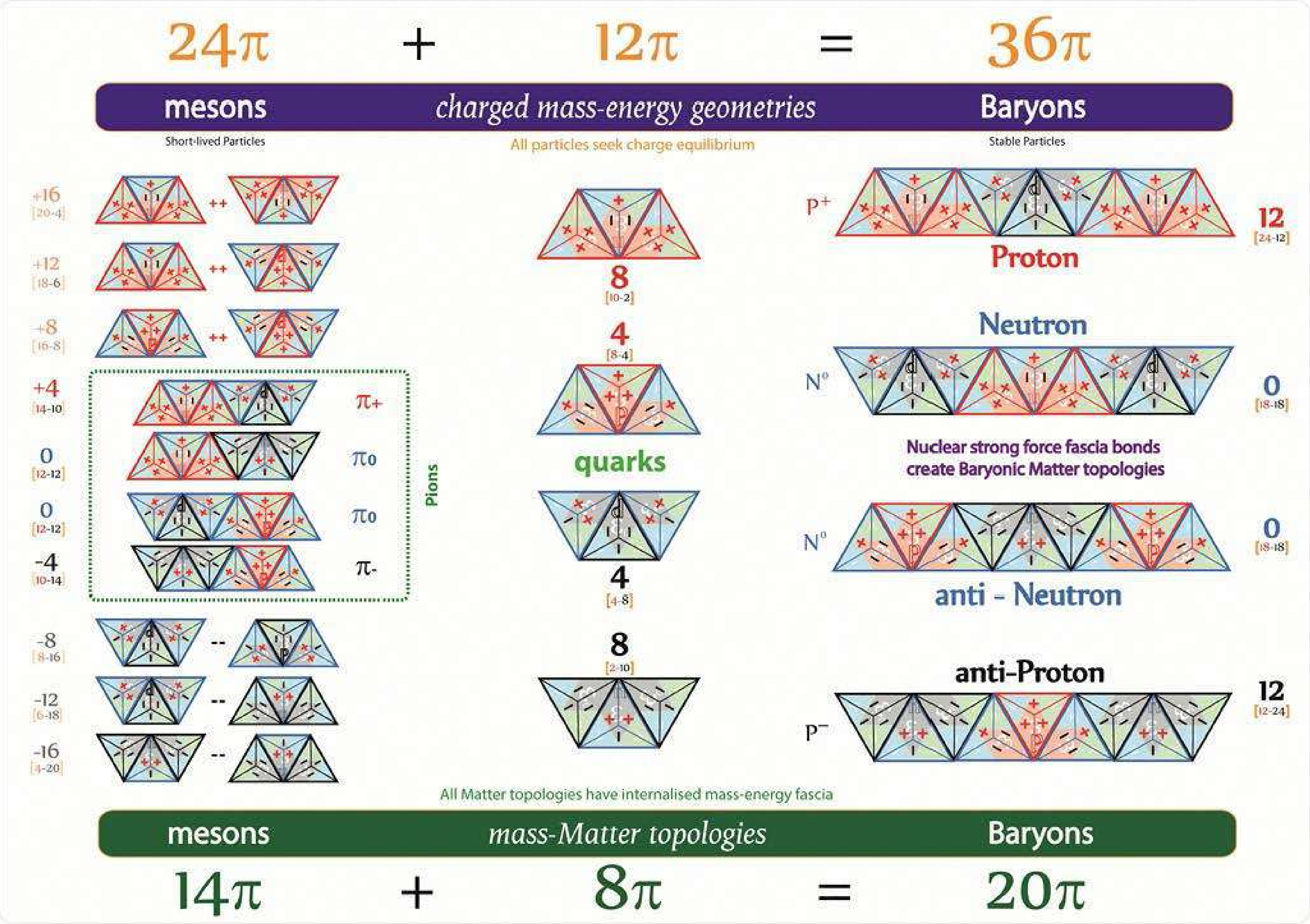

$24\pi + 12\pi = 36\pi$

mesons *charged mass-energy geometries* **Baryons**

Short-lived Particles

All particles seek charge equilibrium

Stable Particles

+16 [20-4]
++
+12 [18-6]
++
+8 [16-8]
++
+4 [14-10]
π+
0 [12-12]
πo
0 [12-12]
πo
-4 [10-14]
π-
Pions
-8 [8-16]
--
-12 [6-18]
--
-16 [4-20]
--
8 [10-2]
4 [8-4]
quarks
4 [4-8]
8 [2-10]
P+
Proton
12 [24-12]
Neutron
N°
0 [18-18]
N°
0 [18-18]
anti - Neutron
anti-Proton
P−
12 [12-24]

Nuclear strong force fascia bonds create Baryonic Matter topologies

All Matter topologies have internalised mass-energy fascia

mesons *mass-Matter topologies* **Baryons**

$14\pi + 8\pi = 20\pi$

Tetryonics 09.03 - Baryon Formation [Meson-Quark interactions]

Copyright ABRAHAM [2008] - all rights reserved

Baryons

Baryons are 36π mass-energy geometries

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Double Charmed Xi

Charmed Sigma

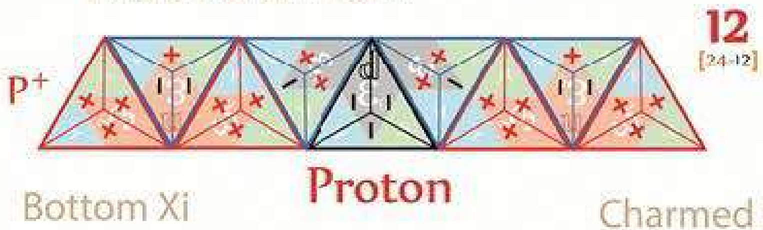
p+
12
[24-12]
Proton

Lambda

Double Bottom Xi

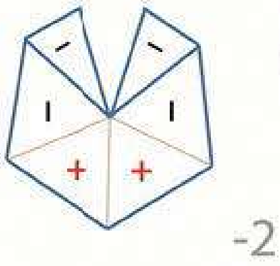

Bottom Xi

Charmed Bottom Omega

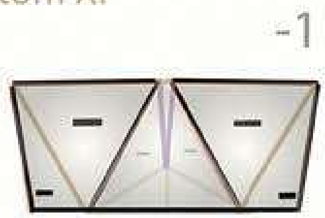

Bottom Omega

Bottom Lambda

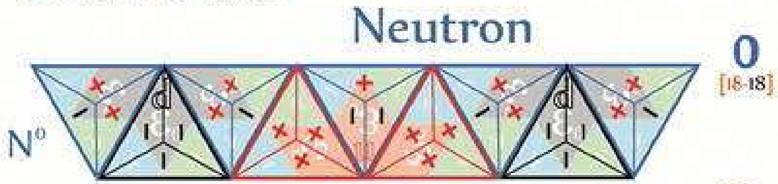
Neutron
0
[18-18]
N°

Bottom Sigma

Charmed Omega

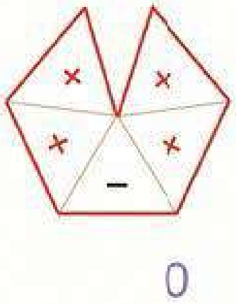

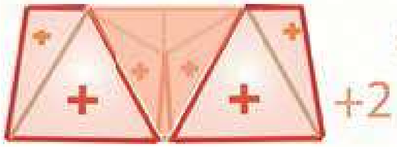

Charmed Bottom Omega Prime

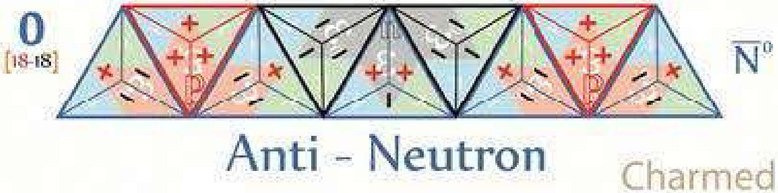
0
[18-18]
$\overline{N}^0$
Anti - Neutron

Charmed Xi Prime

Charmed Lambda

Double Charmed Omega

Xi

Bottom Xi Prime

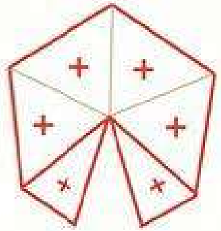

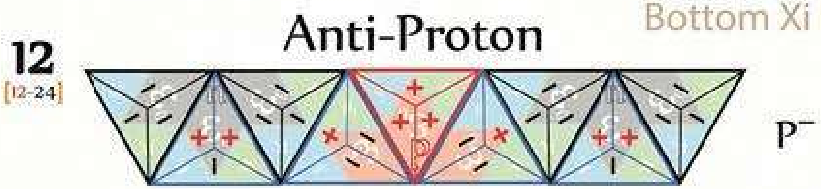
12
[12-24]
Anti-Proton
P⁻

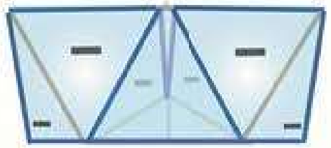

Sigma

Double Bottom Omega

that result in 20π mass-Matter topologies

Proton

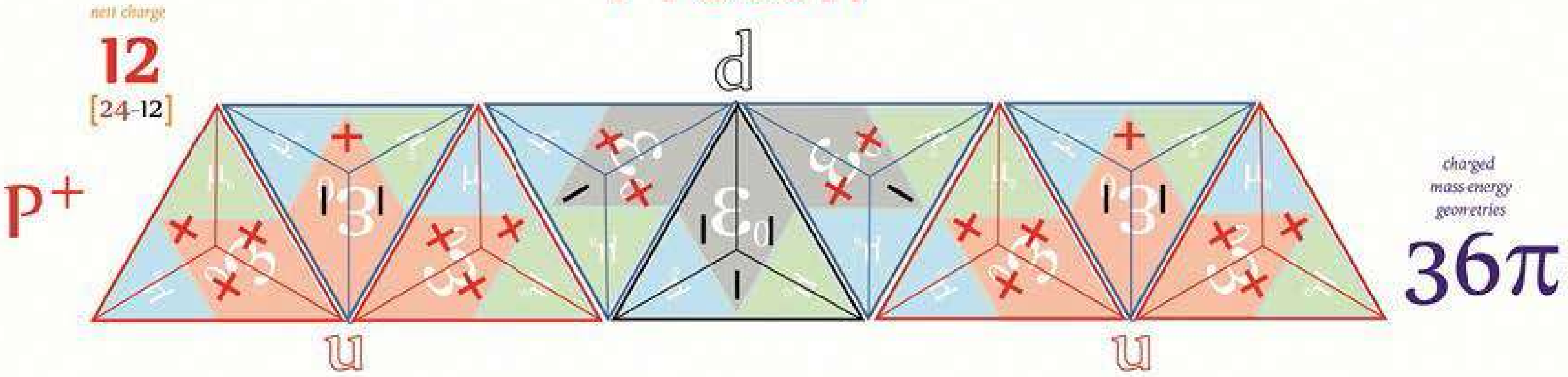
nett charge
12
[24-12]
p+
d
u
u
charged mass-energy geometries
36π

Tri-quark mass-energy geometry

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Baryon mass-Matter topology

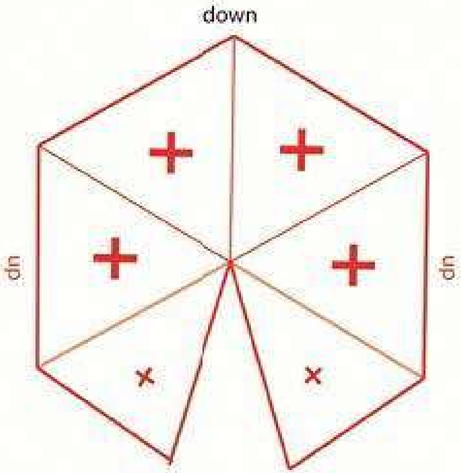

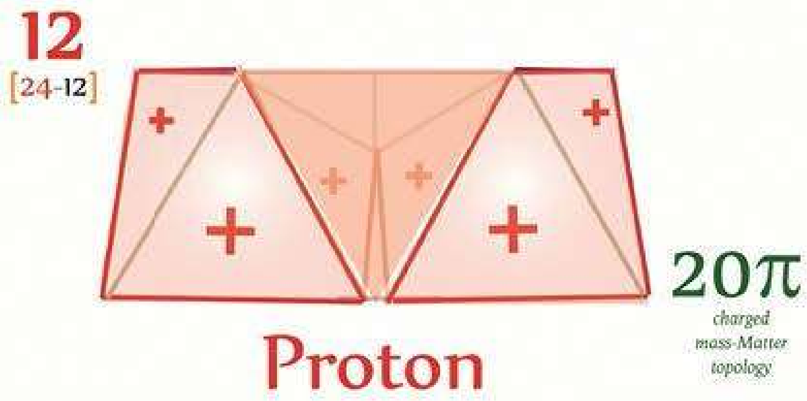
12
[24-12]
Proton
20π
charged mass-Matter topology

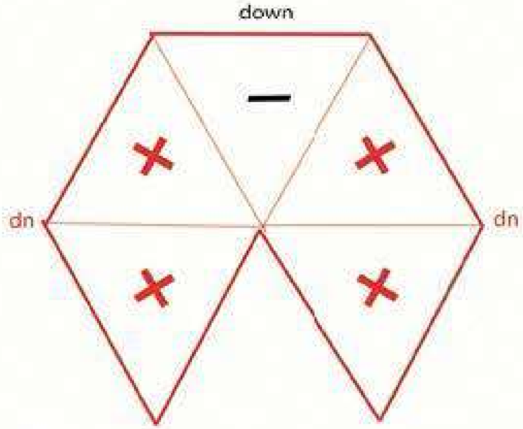

Neutron

nett charge

0

[18-18]

N^0

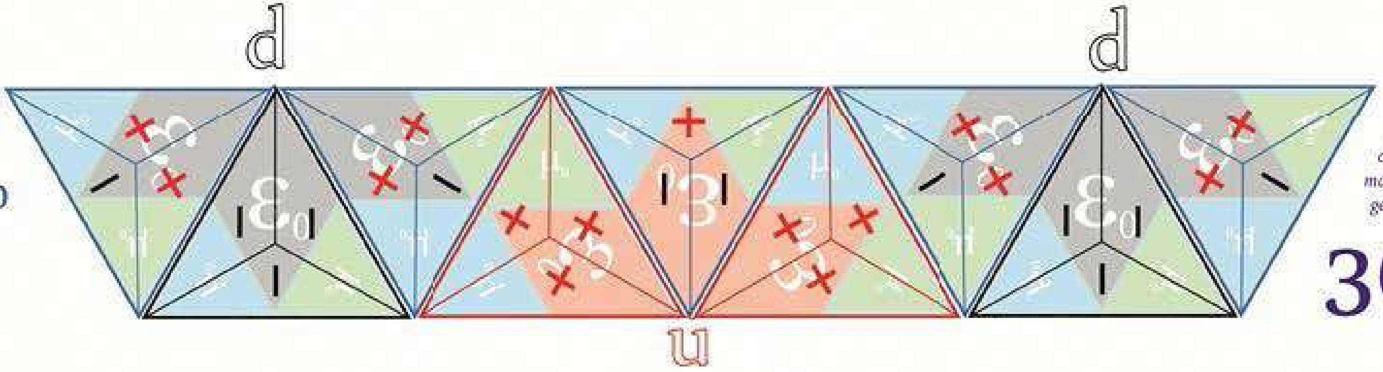
d
d
u
ε_0
ε_0
ε
+
×
−

charged mass-energy geometries

36π

Tri-quark mass-energy geometry

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Baryon mass-Matter topology

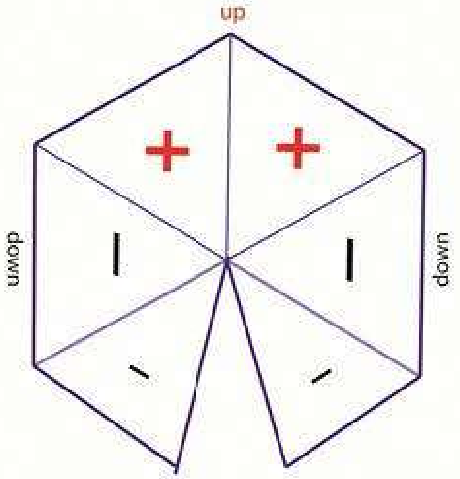
up
down
down
+
+
−
−
−
−

0

[18-18]

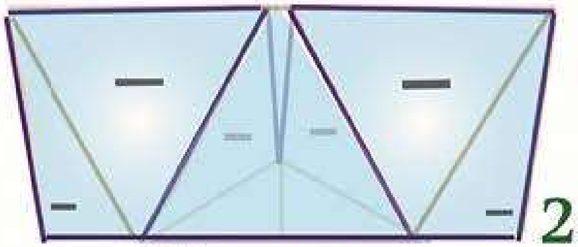

Neutron

20π

charged mass-Matter topology

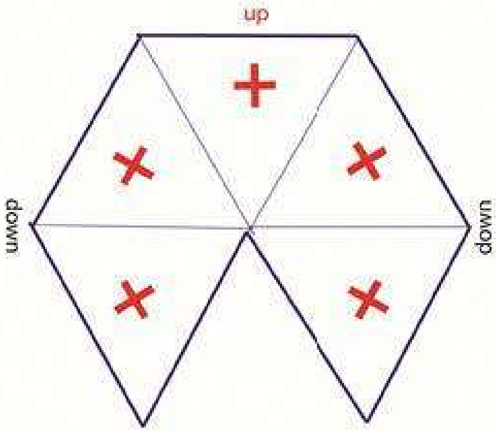
up
down
down
+
×
×
×
×

Copyright ABRAHAM [2008] - all rights reserved

anti-Neutron

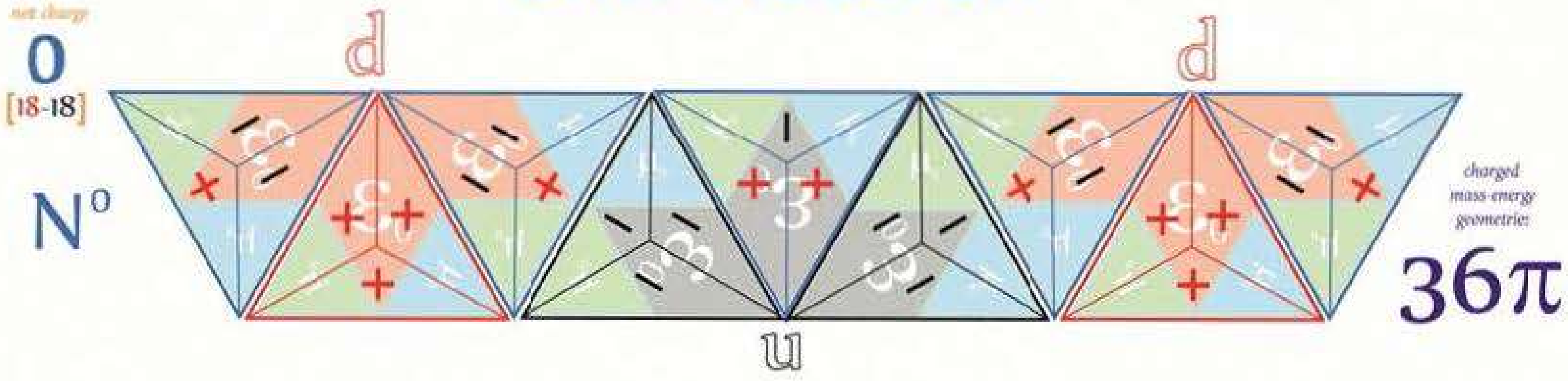
net charge
0
[18-18]
N^0
d
d
u
charged mass-energy geometries
36π

Tri-quark mass-energy geometry

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Baryon mass-Matter topology

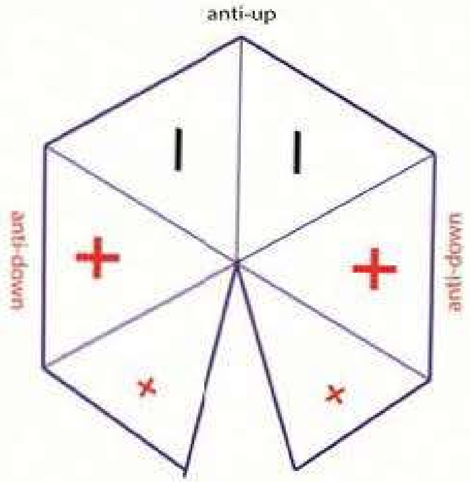
anti-up
anti-down
anti-down

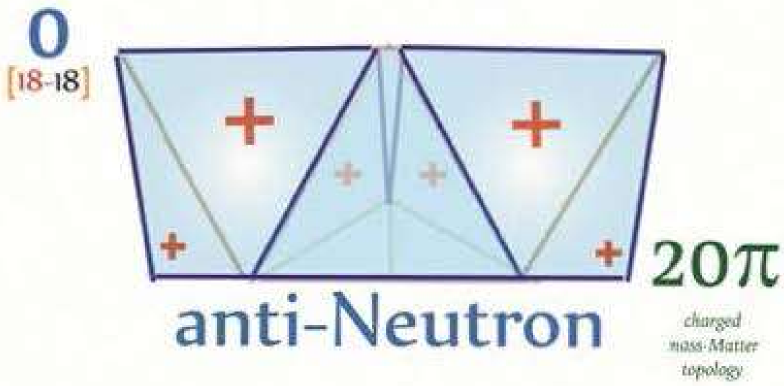
0
[18-18]
anti-Neutron
20π
charged mass-Matter topology

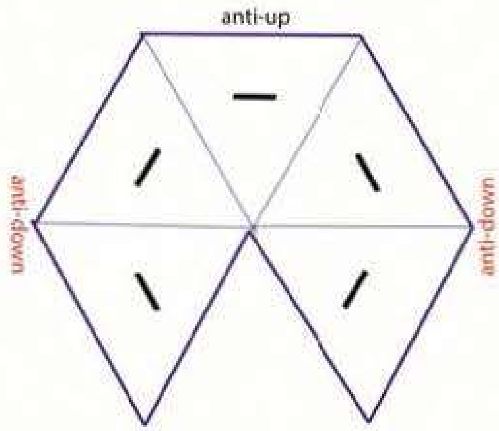
anti-up
anti-down
anti-down

anti-Proton

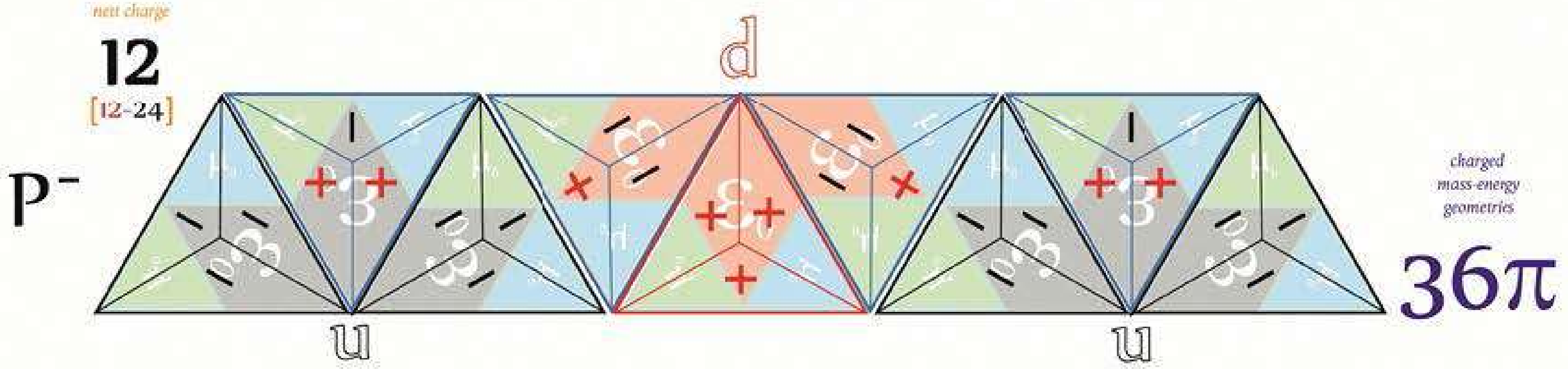
nett charge
12
[12-24]
P⁻
d
u
u
charged mass-energy geometries
36π

Tri-quark mass-energy geometry

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Baryon mass-Matter topology

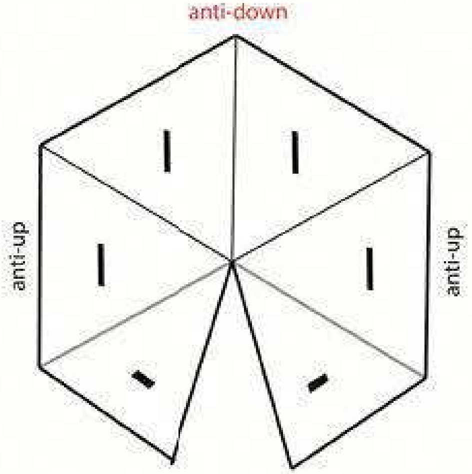
anti-down
anti-up
anti-up

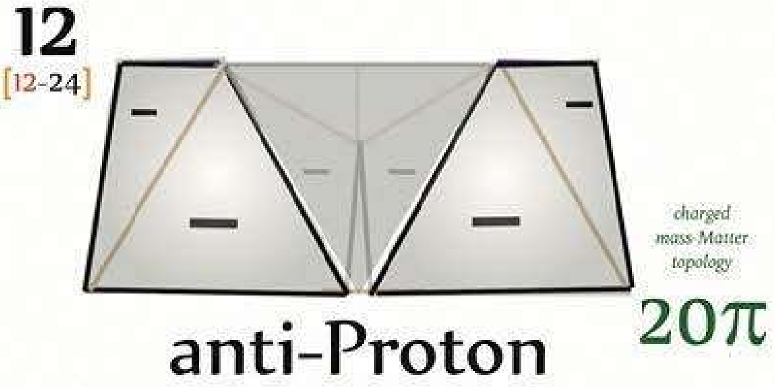
12
[12-24]
charged mass-Matter topology
20π
anti-Proton

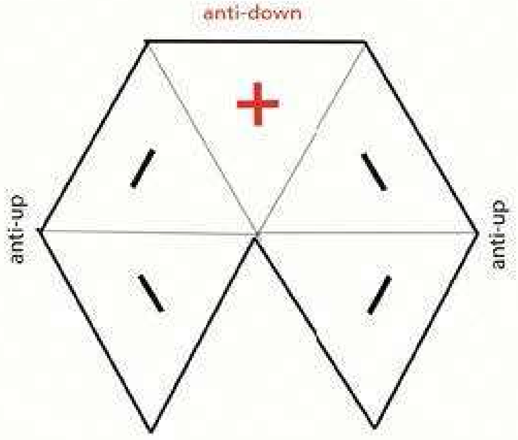
anti-down
anti-up
anti-up

Copyright ABRAHAM [2008] - all rights reserved

Baryon geometry

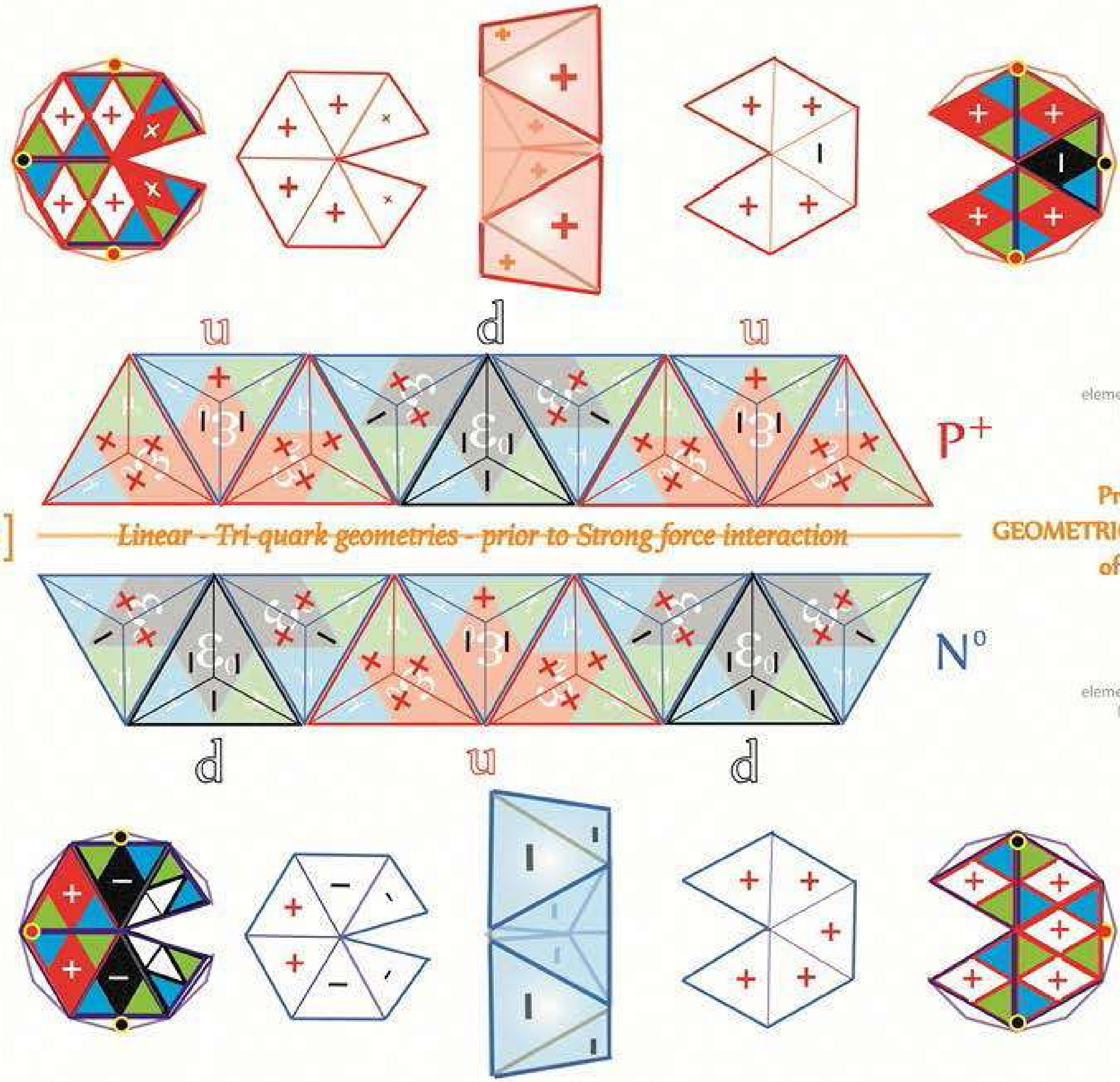

Proton

12

Tetryonic Charge

[24-12]

[36] *Linear - Tri-quark geometries - prior to Strong force interaction*

+1

elementary charge
Proton

Protons are
GEOMETRIC MIRROR IMAGES
of Neutrons

[18-18]

Tetryonic Charge

0

Neutron

0

elementary charge
Neutron

Copyright ABRAHAM [2008] - all rights reserved

Chared particle families

[charged mass geometries & Matter topologies]

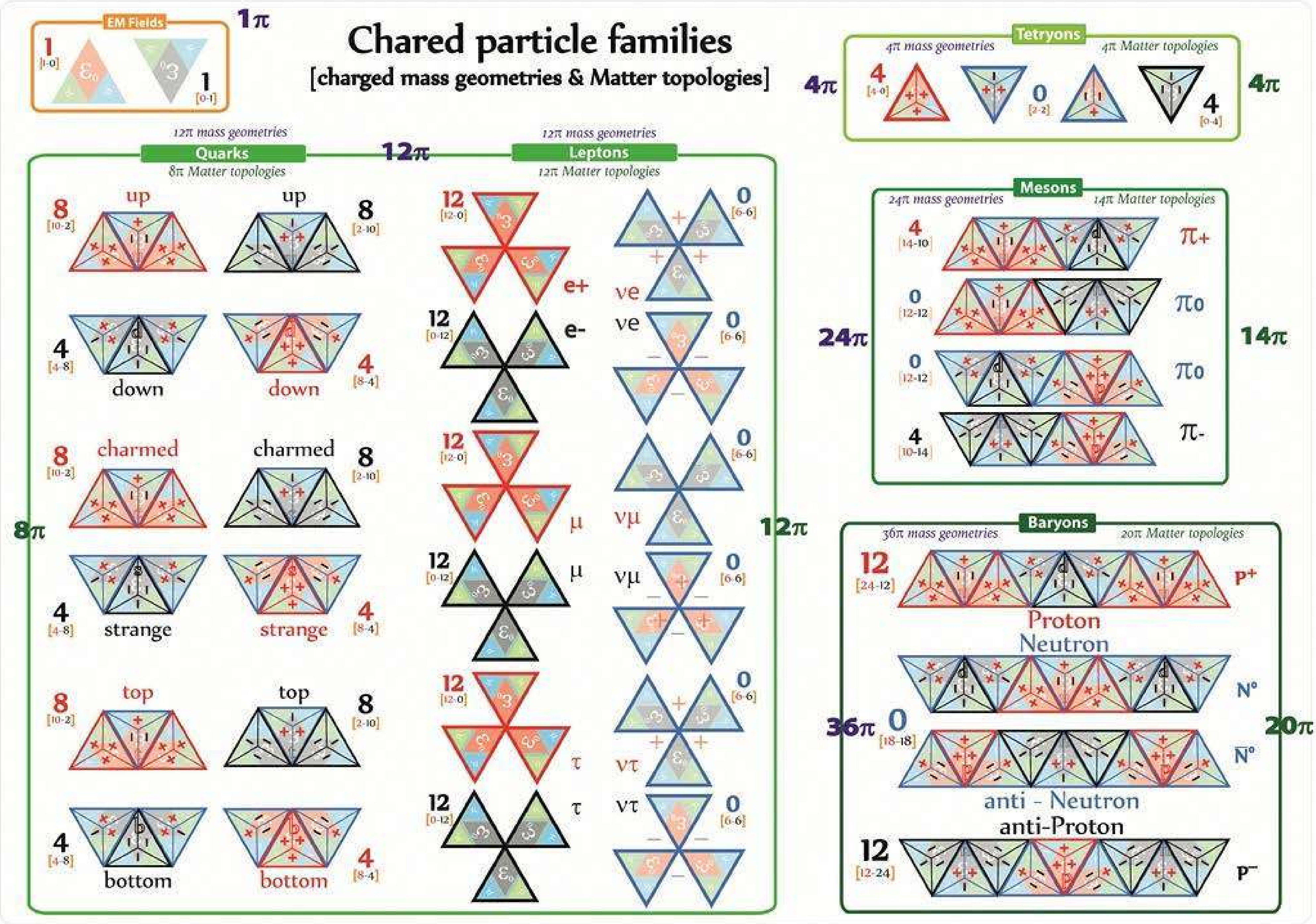
EM Fields
1
[1-0]
1
[0-1]
1π
Tetryons
4π mass geometries
4π Matter topologies
4π
4
[4-0]
0
[2-2]
4
[0-4]
4π
12π mass geometries
Quarks
8π Matter topologies
12π
12π mass geometries
Leptons
12π Matter topologies
8
[10-2]
up
up
8
[2-10]
4
[4-8]
down
down
4
[8-4]
8
[10-2]
charmed
charmed
8
[2-10]
8π
4
[4-8]
strange
strange
4
[8-4]
8
[10-2]
top
top
8
[2-10]
4
[4-8]
bottom
bottom
4
[8-4]
12
[12-0]
e+
12
[0-12]
e-
0
[6-6]
νe
νe
0
[6-6]
12
[12-0]
μ
12
[0-12]
μ
0
[6-6]
νμ
νμ
0
[6-6]
12π
12
[12-0]
τ
12
[0-12]
τ
0
[6-6]
ντ
ντ
0
[6-6]
Mesons
24π mass geometries
14π Matter topologies
24π
4
[14-10]
π+
0
[12-12]
π0
0
[12-12]
π0
4
[10-14]
π-
14π
Baryons
36π mass geometries
20π Matter topologies
12
[24-12]
P+
Proton
Neutron
N°
36π
0
[18-18]
N°
20π
anti - Neutron
anti-Proton
12
[12-24]
P−

Tetryonics 10.07 - Tetryonic Charged particle families

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic Energy levels

Energy
Planck quanta

$$E = n\pi\left[\left[m\Omega v^2\right]\right]$$

mass velocity
momentum

Energy levels

𝔪

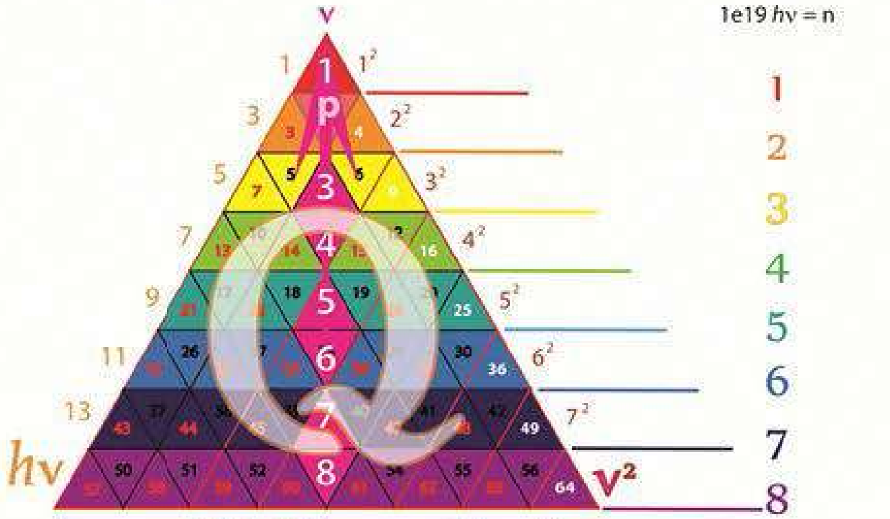
v
1e19 hv = n
1
1²
3
2²
5
3²
7
4²
9
5²
11
6²
13
7²
hv
v²
p
1
2
3
4
5
6
7
8

charged field quantisation

Every charged fascia of a Field or Particle holds squared energy levels
[equilateral EM mass-energy geometries & Matter topologies]

rest mass-Matter

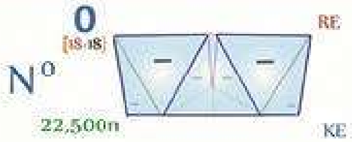
0
[18-18]
N⁰
22,500n
RE
KE

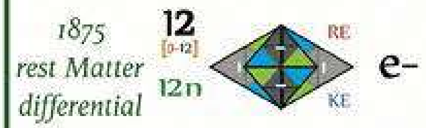
1875
rest Matter
differential
12
[0-12]
12n
RE
KE
e-

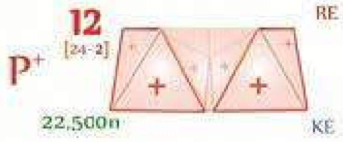
12
[24-2]
P⁺
22,500n
RE
KE

The total mass-energy for any particle is its rest mass-Matter + Kinetic Energies

The mass-energies of its charge geometry determines its rest mass-Matter

An electron's nuclear kinetic energy level is determined by the Baryons it binds to (or by incident photons)

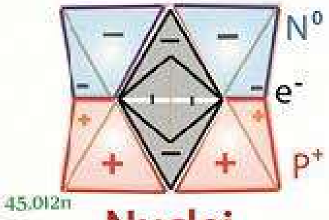
N⁰
e⁻
P⁺
45,012n

Nuclei

Nuclear Quantum Level | *electron Kinetic Energy Level*

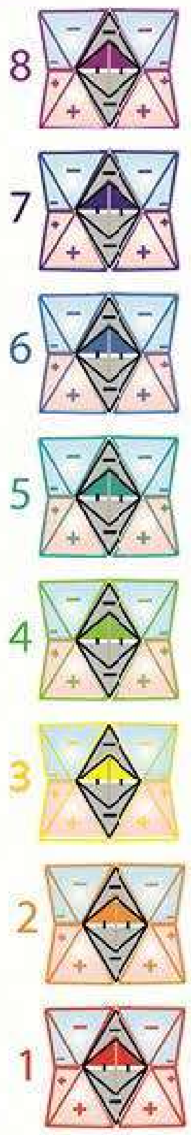

Nuclear Quantum Level	electron Kinetic Energy Level
8	768
7	588
6	432
5	300
4	192
3	108
2	48
1	12

As the Nuclei quantum level increases - so does the electron's energy and tmomenta

The Tetryonic KEM Colour code

As the energy levels of atomic nuclei increase/decrease the KEM fields of leptons bound to them change accordingly

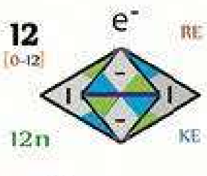
12
[0-12]
e⁻
RE
12n
KE

Electron

1	2	3	4	5	6	7	8
▲	▲	▲	▲	▲	▲	▲	▲
1	4	9	16	25	36	49	64

All energy levels are Tetryonic square number quantas
ie 1, 4, 9, 16, 25 etc (1,2,3,4,5)

Ek-Field Planck quanta

$$KEM = 4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Kinetic energies Electromagnetic mass velocity

tri-quark charged toplogy combinations

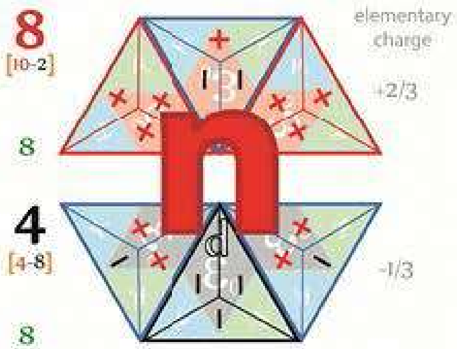

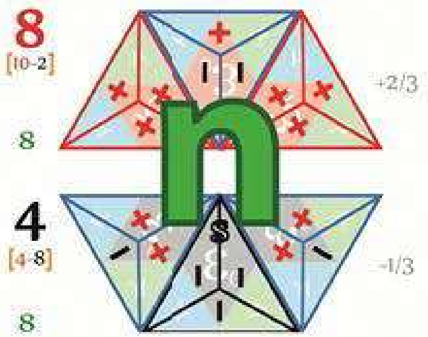

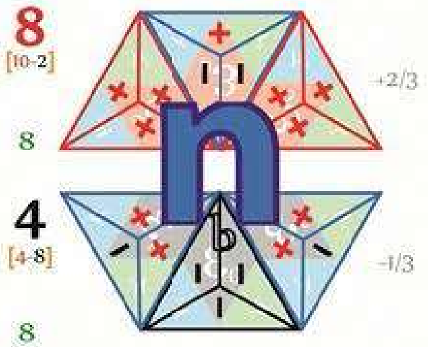

Quarks seek charge equilibrium by
forming Baryonic Matter topologies
as a result of charge interactions
[STRONG nuclear Force]

UP
DOWN] gen 1
CHARMED
STRANGE] gen 2
TOP
BOTTOM] gen 3

[+8 +8 +8] [10-2] [10-2] [10-2] **24** (+) 2 2/3 2/3 2/3

[+8 -4 +8] [10-2] [4-8] [10-2] **12** (+) 1 2/3 1/3 2/3

The particle Zoo is the result of tri-quark mass-energy topologies

[-4 +8 -4] [4-8] [10-2] [4-8] **0** (0) 0 1/3 2/3 1/3

[-4 -4 -4] [4-8] [4-8] [4-8] **12** (-) 1 1/3 1/3 1/3

Twice the number of combinations are possible with the inclusion of anti-matter topologies

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Any combinations of inter-generational quarks
results in differing final mass-energies
in the created Baryonic Matter
due to energy equalisation

The Particle Zoo

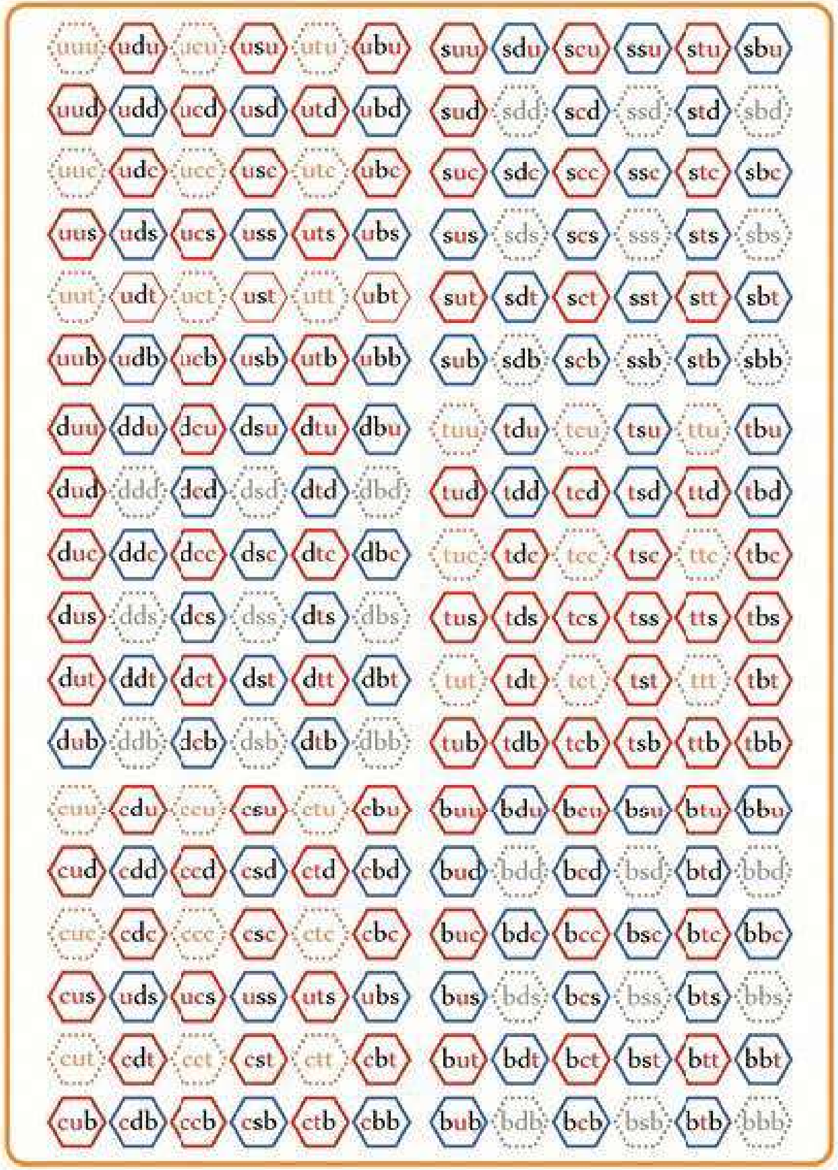

uuu	udu	ucu	usu	utu	ubu	suu	sdu	scu	ssu	stu	sbu
uud	udd	ucd	usd	utd	ubd	sud	sdd	scd	ssd	std	sbd
uuc	udc	ucc	usc	utc	ubc	suc	sdc	scc	ssc	stc	sbc
uus	uds	ucs	uss	uts	ubs	sus	sds	scs	sss	sts	sbs
uut	udt	uct	ust	utt	ubt	sut	sdt	sct	sst	stt	sbt
uub	udb	ucb	usb	utb	ubb	sub	sdb	scb	ssb	stb	sbb
duu	ddu	dcu	dsu	dtu	dbu	tuu	tdu	tcu	tsu	ttu	tbu
dud	ddd	dcd	dsd	dtd	dbd	tud	tdd	tcd	tsd	ttd	tbd
duc	ddc	dcc	dsc	dtc	dbc	tuc	tdc	tcc	tsc	ttc	tbc
dus	dds	dcs	dss	dts	dbs	tus	tds	tcs	tss	tts	tbs
dut	ddt	dct	dst	dtt	dbt	tut	tdt	tct	tst	ttt	tbt
dub	ddb	dcb	dsb	dtb	dbb	tub	tdb	tcb	tsb	ttb	tbb
cuu	cdu	ccu	csu	ctu	cbu	buu	bdu	bcu	bsu	btu	bbu
cud	cdd	ccd	csd	ctd	cbd	bud	bdd	bcd	bsd	btd	bbd
cuc	cdc	ccc	csc	ctc	cbc	buc	bdc	bcc	bsc	btc	bbc
cus	uds	ucs	uss	uts	ubs	bus	bds	bcs	bss	bts	bbs
cut	cdt	cct	cst	ctt	cbt	but	bdt	bct	bst	btt	bbt
cub	cdb	ccb	csb	ctb	cbb	bub	bdb	bcb	bsb	btb	bbb

216 possible Baryon topologies [excluding anti-matter] with each combination producing differing Baryonic nett energy levels

Quarks always form symmetrical [udu] or [dud] charged Baryonic topologies
[same charges never combine except under high energy conditions]
ie. Proton is a [udu] not [uud] as is commonly stated

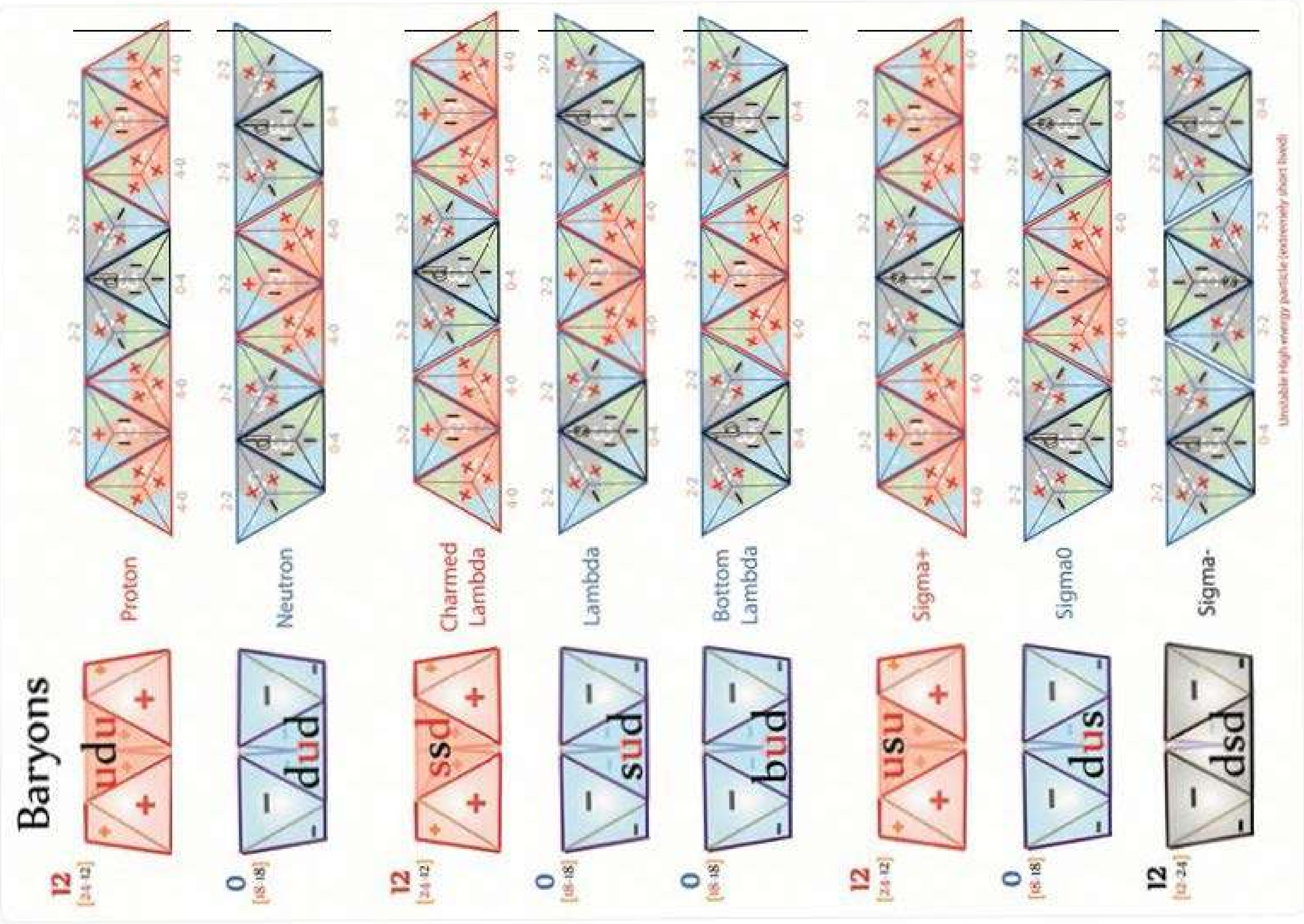
Baryons
12 [24·12]
udu
Proton
0 [48·18]
dud
Neutron
12 [24·12]
ssd
Charmed Lambda
0 [48·18]
sud
Lambda
0 [48·18]
bud
Bottom Lambda
12 [24·12]
usu
Sigma+
0 [48·18]
dus
Sigma0
12 [12·24]
dsd
Sigma-
Unstable High energy particle (extremely short lived)

Tetryonics 10.10 - Proton to Sigma-

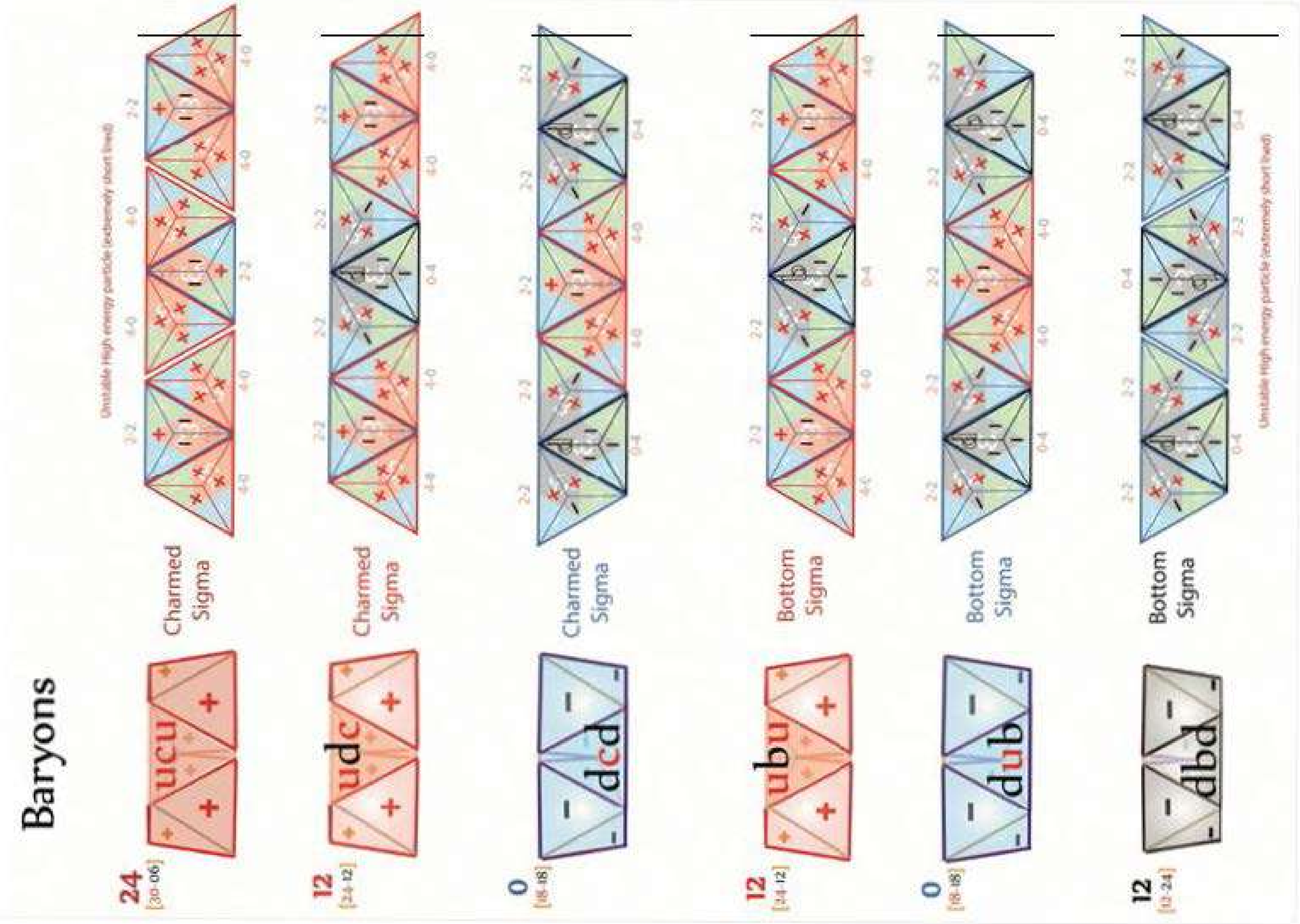
Baryons
24
[30-06]
ucu
Charmed Sigma
12
[24-12]
udc
Charmed Sigma
0
[18-18]
dcd
Charmed Sigma
12
[24-12]
ubu
Bottom Sigma
0
[18-18]
dub
Bottom Sigma
12
[12-24]
dbd
Bottom Sigma
Unstable High energy particle (extremely short lived)
Unstable High energy particle (extremely short lived)

Tetryonics 10.11 - Charmed to Bottom Sigma

Copyright ABRAHAM [2008] - all rights reserved

Baryons

0
[18-18]

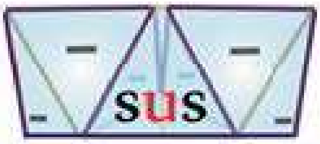

Xi

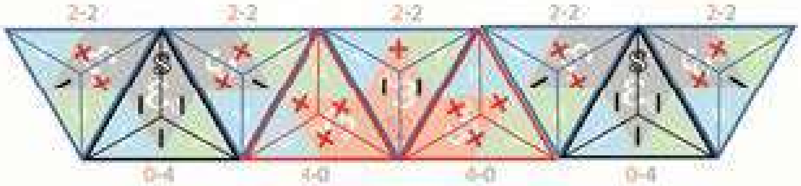
2-2
2-2
2-2
2-2
2-2
s
s
0-4
4-0
4-0
0-4

12
[12-24]

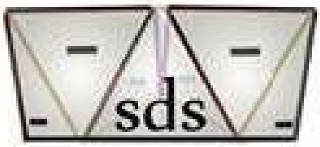

Xi

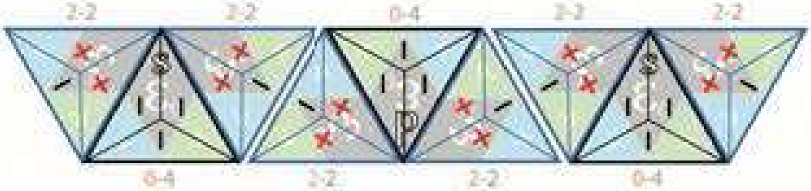
2-2
2-2
0-4
2-2
2-2
s
p
s
0-4
2-2
2-2
0-4

Unstable High energy particle (extremely short lived)

Charmed Xi Prime

12
[24-12]

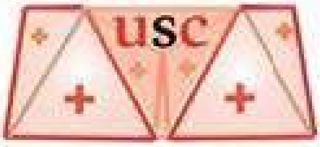

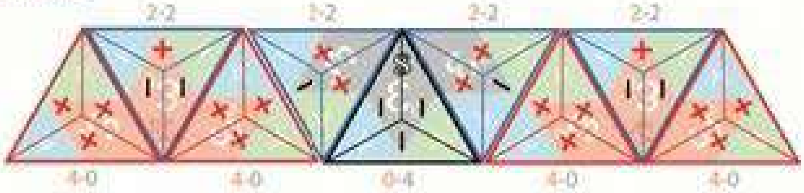
2-2
2-2
2-2
2-2
s
4-0
4-0
0-4
4-0
4-0

Charmed Xi Prime

0
[18-18]

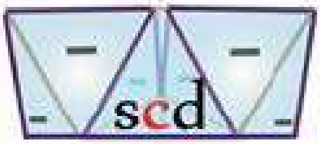

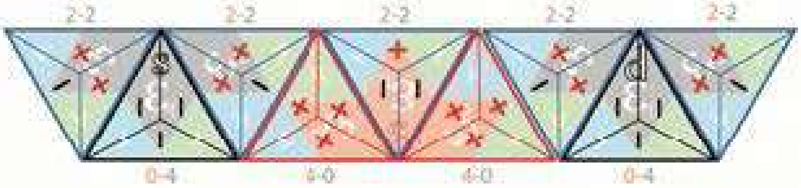
2-2
2-2
2-2
2-2
2-2
s
d
0-4
4-0
4-0
0-4

Double Charmed Xi

24
[30-06]

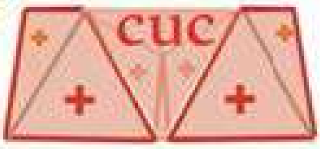

Unstable High energy particle (extremely short lived)

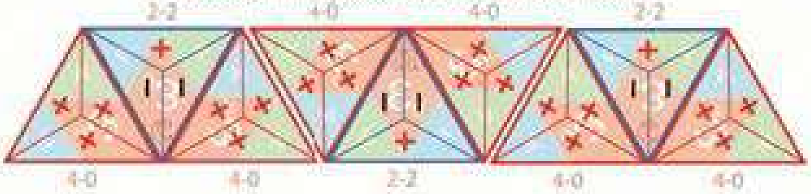
2-2
4-0
4-0
2-2
4-0
4-0
2-2
4-0
4-0

Double Charmed Xi

12
[24-12]

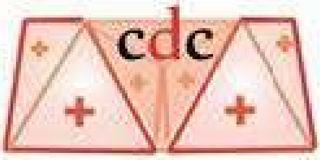

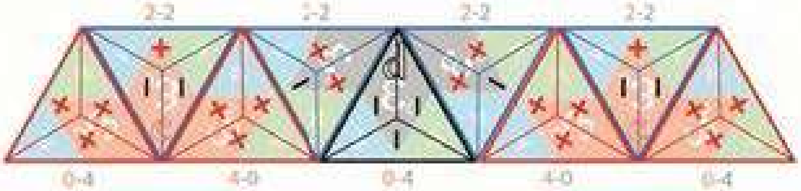
2-2
2-2
2-2
2-2
d
0-4
4-0
0-4
4-0
0-4

Copyright ABRAHAM [2008] - all rights reserved

Baryons

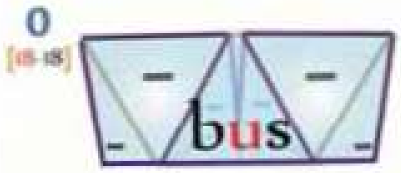

Bottom Xi
(Cascade B)

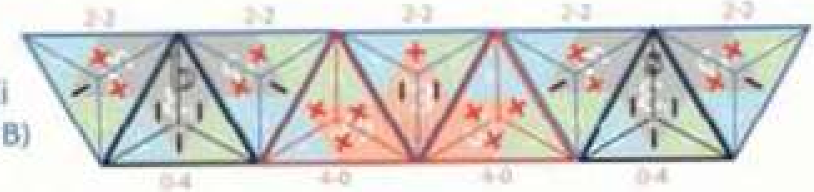
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

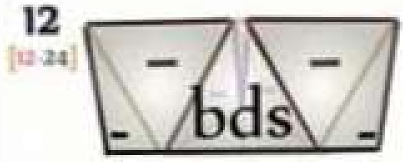

Bottom Xi
(Cascade B)

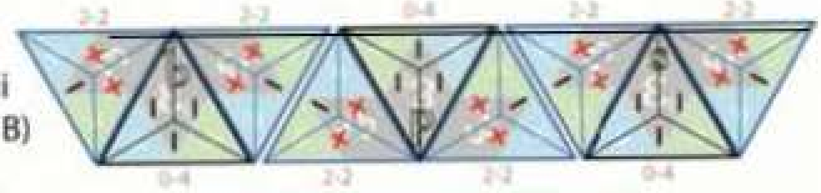
2-2
2-2
0-4
2-2
2-2
0-4
2-2
2-2
0-4

Unstable High energy particle (extremely short lived)

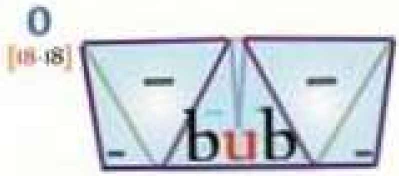

Double
Bottom Xi

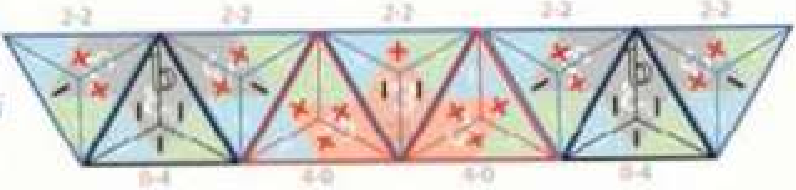
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

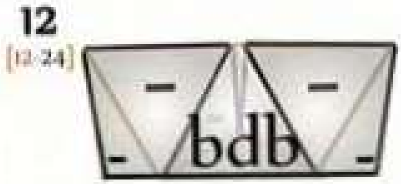

Double
Bottom Xi

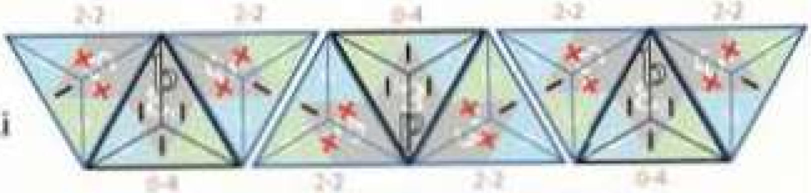
2-2
2-2
0-4
2-2
2-2
0-4
2-2
2-2
0-4

Unstable High energy particle (extremely short lived)

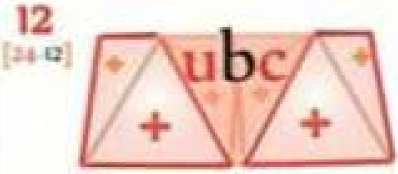

Charmed
Bottom Xi

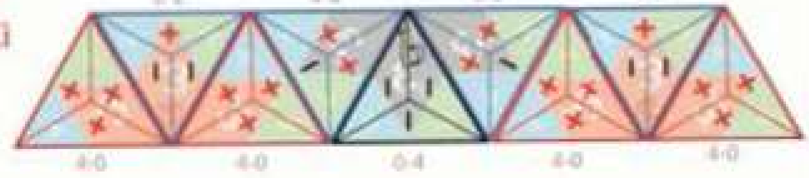
2-2
2-2
2-2
2-2
4-0
4-0
0-4
4-0
4-0

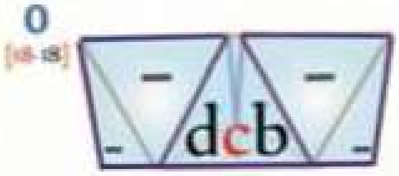

Charmed
Bottom Xi

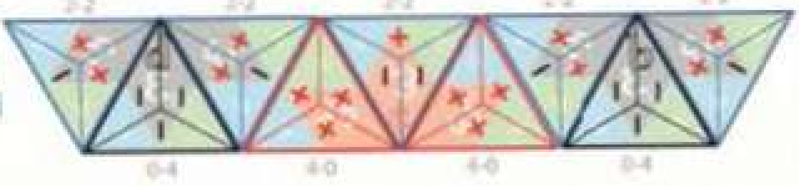
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

Copyright ABRAHAM [2008] - all rights reserved

Baryons

0
[18-18]

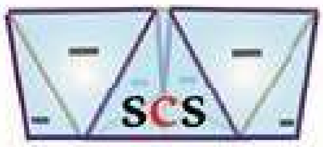

Charmed
Omega

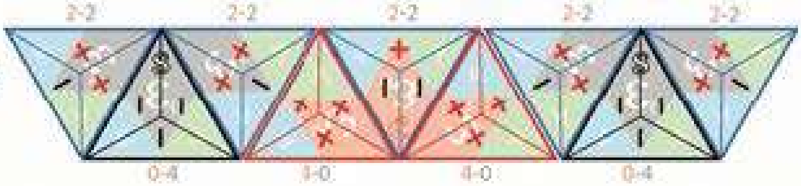
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

12
[12-24]

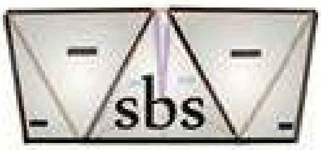

Bottom
Omega

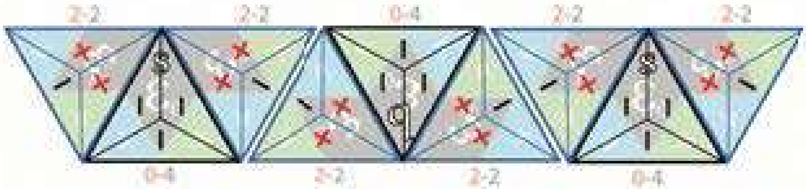
2-2
2-2
0-4
2-2
2-2
0-4
2-2
2-2
0-4

Unstable Highenergy particle (extremely short lived)

12
[24-12]

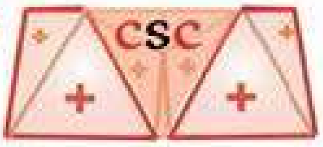

Double
Charmed
Omega

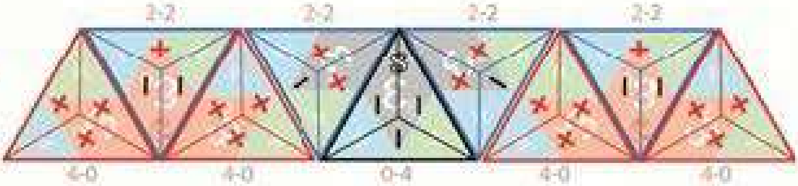
2-2
2-2
2-2
2-2
4-0
4-0
0-4
4-0
4-0

0
[18-18]

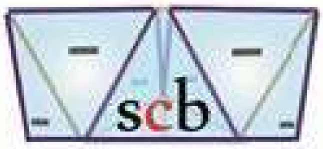

Charmed
Bottom
Omega

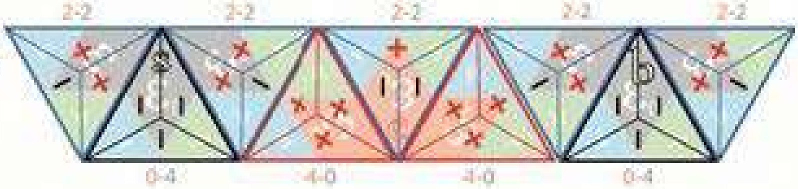
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

12
[12-24]

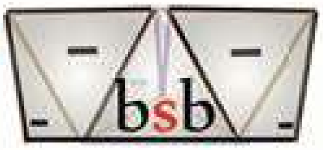

Double
Bottom
Omega

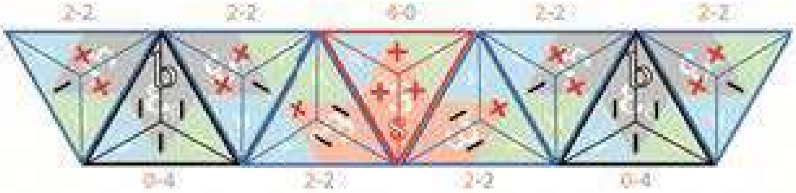
2-2
2-2
4-0
2-2
2-2
0-4
2-2
2-2
0-4

12
[24-12]

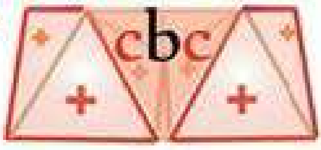

Double
Charmed
Bottom
Omega

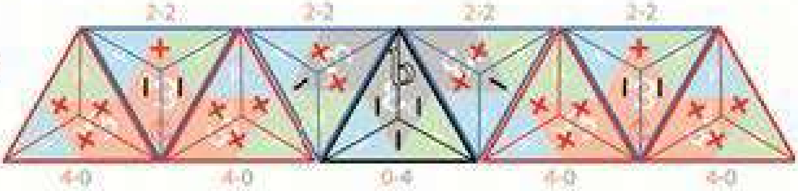
2-2
2-2
2-2
2-2
4-0
4-0
0-4
4-0
4-0

0
[18-18]

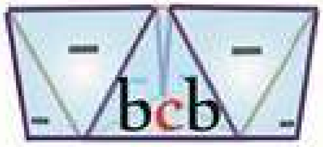

Charmed
Double
Bottom
Omega

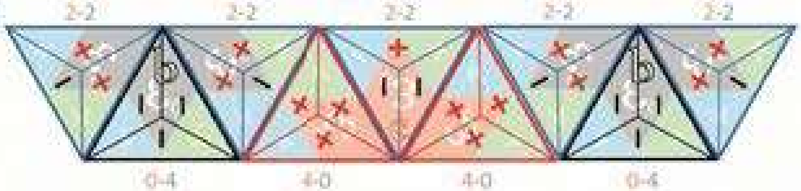
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

ZPF **Photons** **Bosons** **Tetryons**

$n\pi$ charged mass-Energy-Matter geometries

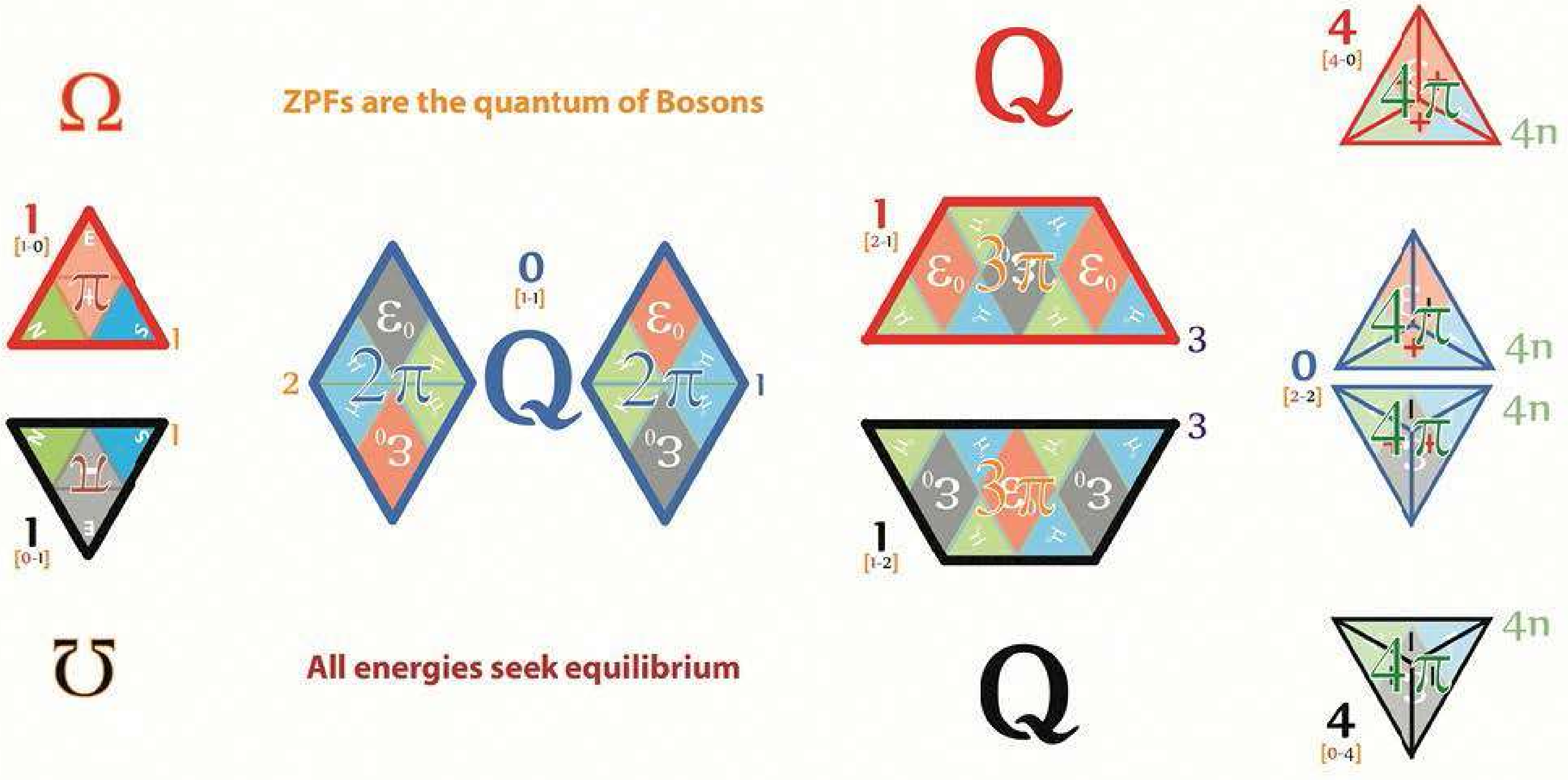
Ω
1
[1-0]
π
1
1
1
[0-1]
℧
ZPFs are the quantum of Bosons
0
[1-1]
2
ε0
2π
3
Q
1
All energies seek equilibrium
Q
1
[2-1]
ε0 3π ε0
3
3
3 3π 3
1
[1-2]
Q
4
[4-0]
4π
4n
0
[2-2]
4π
4n
4π
4n
4π
4n
4
[0-4]

Energy momenta

$n\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v}\right]^2\right]$ Energy

EM mass-Energies

$\text{EVEN}\pi\left[\left[\overset{\text{EM Fields}}{\varepsilon_0\mu_0}\right]_{\text{ElectroMagnetic}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$ mass-energy

Charges

$\text{ODD}\pi\left[\left[\overset{\text{EM Fields}}{\varepsilon_0\mu_0}\right]_{\text{ElectroMagnetic}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$ Charge

Matter

$4n\pi\left[\left[\overset{\text{EM Fields}}{\varepsilon_0\mu_0}\right]_{\text{ElectroMagnetic}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$ Matter

Tetryonics 11.01 - Charged Energy-Matter geometries

Gluons

The gluon can be considered to be the fundamental exchange particle facilitating the strong interaction between protons and neutrons in a nucleus.

Gluons are the exchange particles for the 'colour' strong force between quarks, analogous to the exchange of photons in the electromagnetic force between two charged particles.

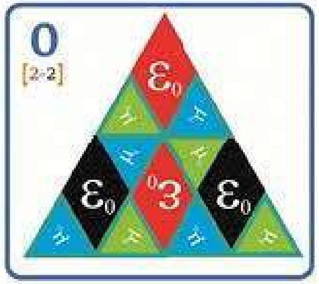
0
[2-2]
ε_0
$^0 3$
ε_0
ε_0

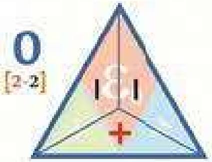

Ω

Gluons are considered to be bi-coloured, carrying a unit of colour and a unit of anti-colour

℧

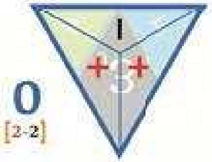

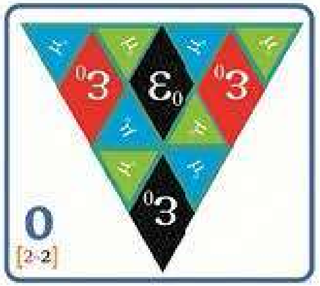
$^0 3$
ε_0
$^0 3$
$^0 3$
0
[2-2]

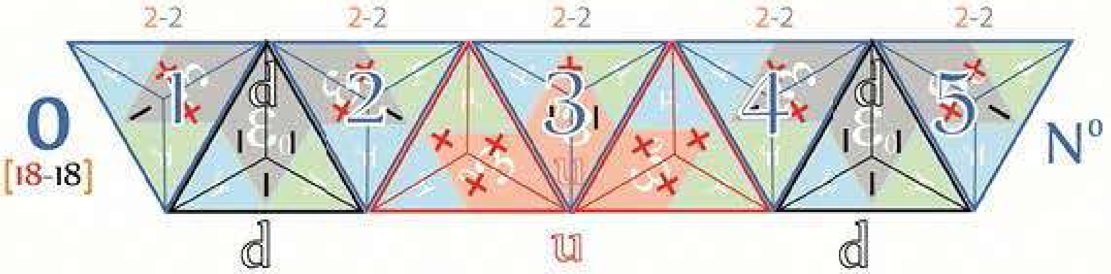
2-2
2-2
2-2
2-2
2-2
0
[18-18]
1
2
3
4
5
N°
d
u
d

Q

Gluons are the neutral Tetryons inside all fermionic Matter topologies

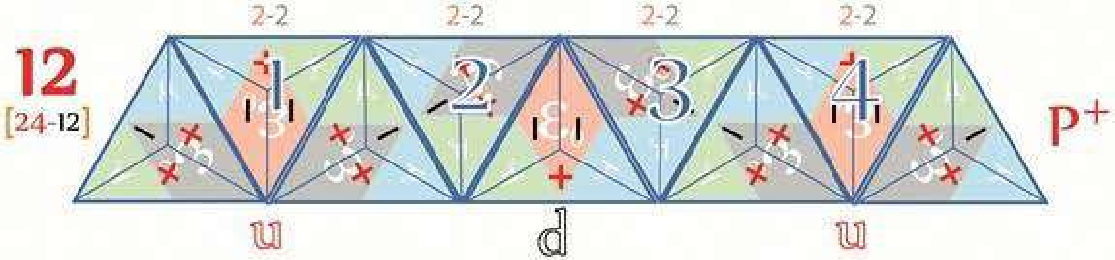
2-2
2-2
2-2
2-2
12
[24-12]
1
2
3
4
P+
u
d
u

They are the neutral charge [di-electric] Tetryons located between charged Tetryons in Matter topologies

[attractive coulombic forces between their opposite charged fascia create the 'colour force]

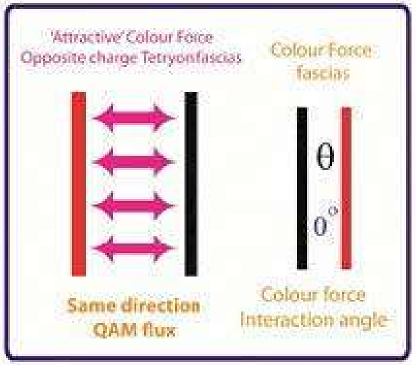
'Attractive' Colour Force
Opposite charge Tetryonfascias
Colour Force fascias
θ
0°
Same direction QAM flux
Colour force Interaction angle

The interaction between neutral Tetryons (Gluons) and all charged Tetryons is the Strong Colour force

nett charge of Gluons is same as a Neutron

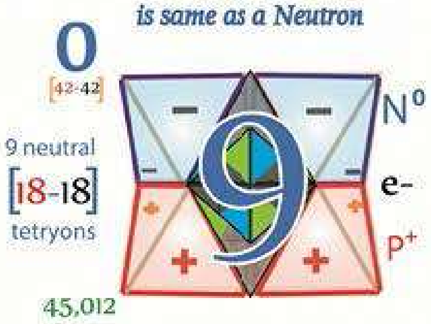
0
[42-42]
9 neutral
[18-18]
tetryons
45,012
9
N°
e-
p+

Electrons contain no Gluons

d d
e- p- u p+
ν^0 u N^0 e+

Neutral Tetryons combine with charged Tetryons to form all Particles

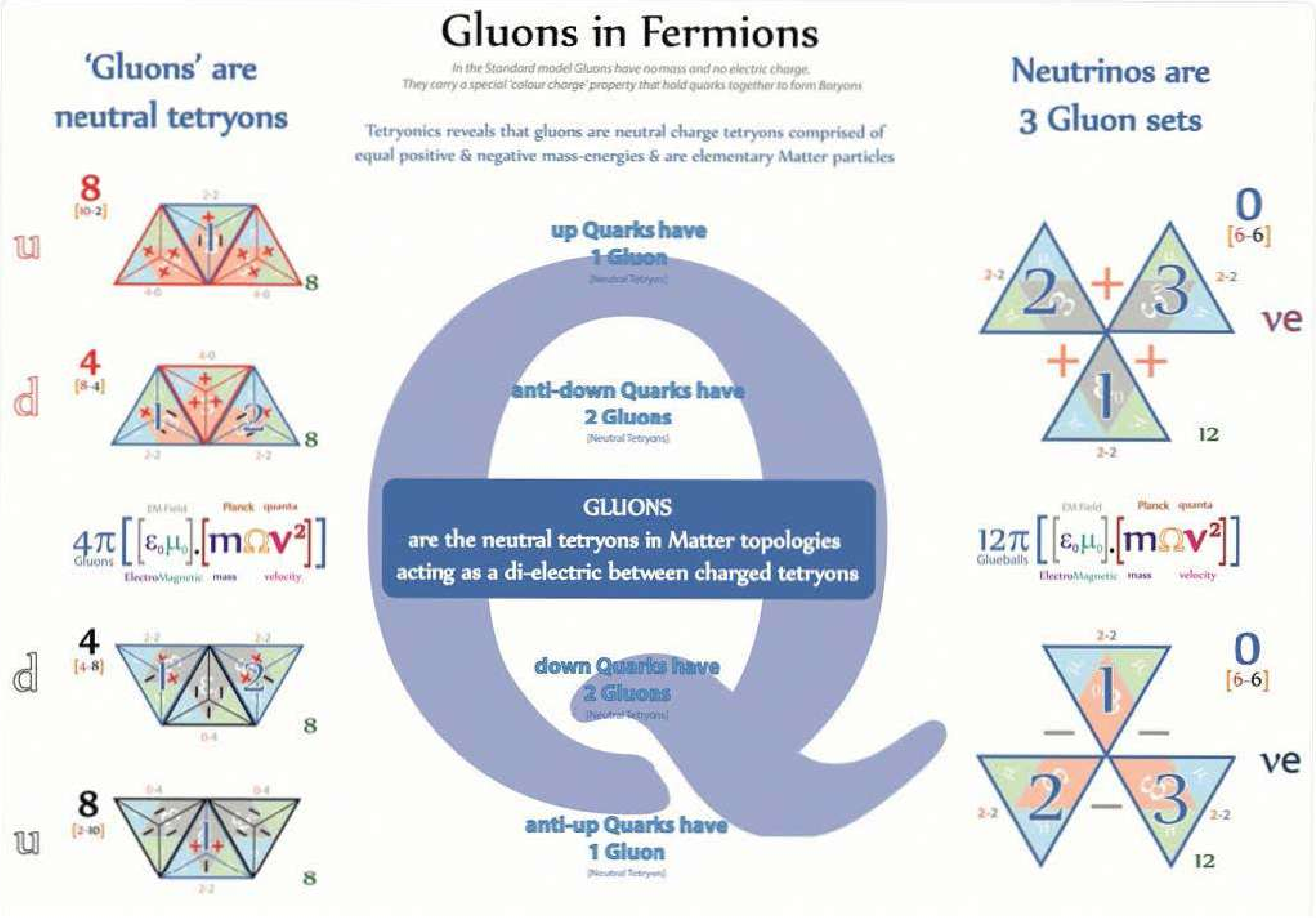

Gluons in Fermions

In the Standard model Gluons have no mass and no electric charge.
They carry a special 'colour charge' property that hold quarks together to form Baryons

Tetryonics reveals that gluons are neutral charge tetryons comprised of equal positive & negative mass-energies & are elementary Matter particles

'Gluons' are neutral tetryons

Neutrinos are 3 Gluon sets

up Quarks have 1 Gluon
[Neutral Tetryon]

anti-down Quarks have 2 Gluons
[Neutral Tetryons]

GLUONS
are the neutral tetryons in Matter topologies acting as a di-electric between charged tetryons

down Quarks have 2 Gluons
[Neutral Tetryons]

anti-up Quarks have 1 Gluon
[Neutral Tetryon]

$$4\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Gluons; EM Field; Planck quanta; ElectroMagnetic; mass; velocity

$$12\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Glueballs; EM Field; Planck quanta; ElectroMagnetic; mass; velocity

Tetryonics 11.03 - Gluons in Fermions

Copyright ABRAHAM [2008] - all rights reserved

Baryonic Gluons

In 'the Standard Model,' Gluons are vector gauge bosons that mediate strong interactions of quarks in quantum chromodynamics (QCD). Unlike the electrically neutral photon of quantum electrodynamics (QED), gluons themselves carry colour charge and therefore participate in the strong interaction in addition to mediating it, making QCD significantly harder to analyze than QED.

They are considered to be elementary particles which act as the exchange particles (or gauge bosons) for the colour force between quarks, analogous to the exchange of photons in the electromagnetic force between two charged particles

Tetryonics simplifies the current definition of Gluons and clearly identifies their geometric properties, along with their role in particle genesis

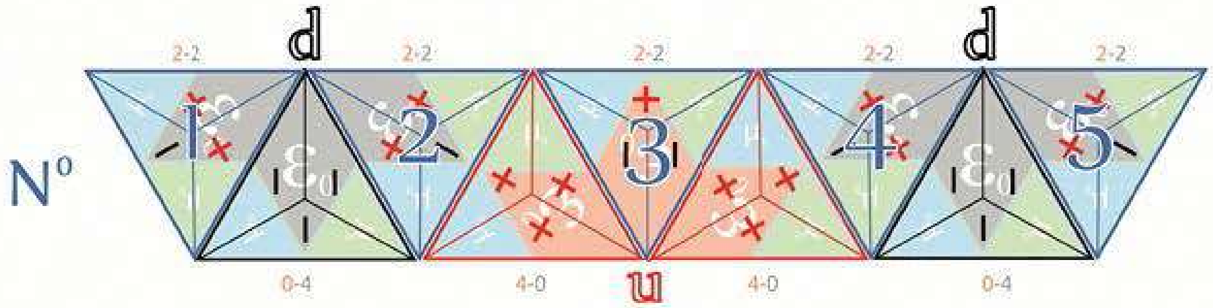

Neutrons have 5 Gluons

[Neutral Tetryons]

0

[18-18]

Deuterium nuclei (being the constituent quanta of all Elements) have 9 neutral Tetryons (Gluons) which in turn contribute to their gravitational mass along with the charged Tetryons

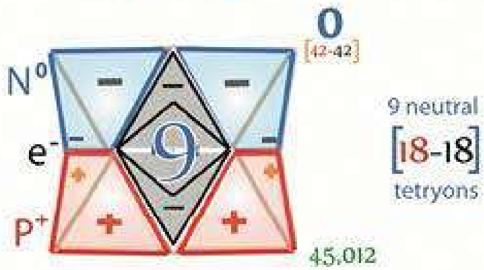

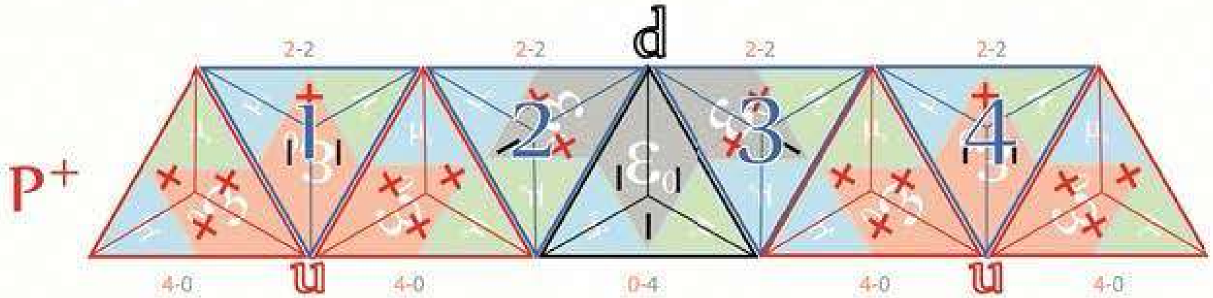

12

[24-12]

Protons have 4 Gluons

[Neutral Tetryons]

'Glueballs'

In particle physics, a glueball is a hypothetical composite particle. It solely consists of gluon particles, without valence quarks.

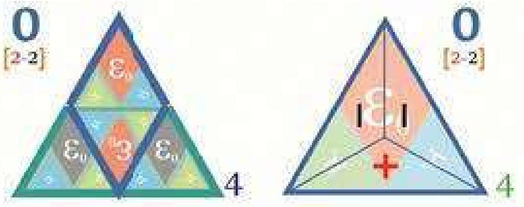

Neutral Tetryons are the result of equal number charged EM fields combining to form neutral charge Tetryonic topologies

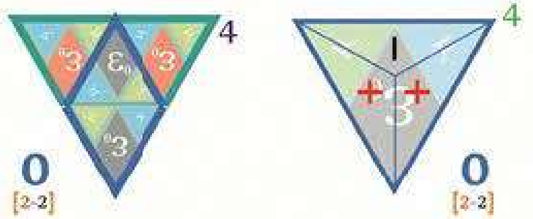

It is conceivable that given the right conditions (ie a cloud of neutral Tetryons), that in the absence of charged Tetryons to interact with, a Glueball topology can be formed entirely from neutral Tetryons

Note: Despite their total neutral charge Glueball topologies are polarised

Neutrinos and some exotic neutral charge Baryons can be considered to be 'Gluon topologies'

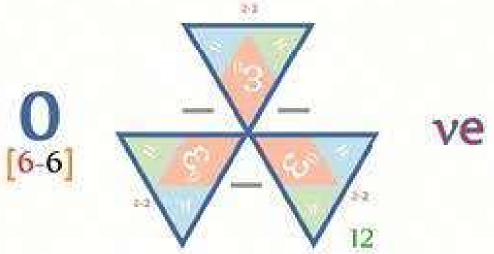

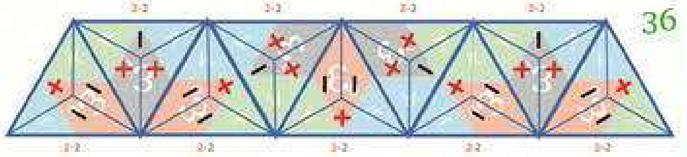

It is possible to create a family of Baryonic particles comprised entirely of neutral Tetryons

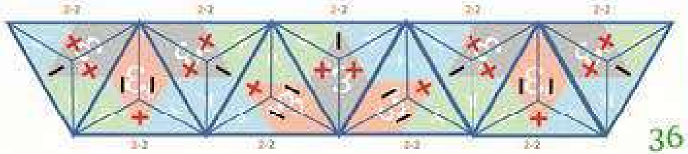

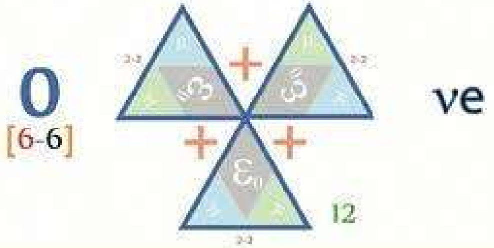

Neutrons are not considered to be Glueballs as they contain charged Tetryons

OMEGA Particle
Icosahedron Glueball

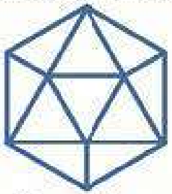

20 Neutral Teryons

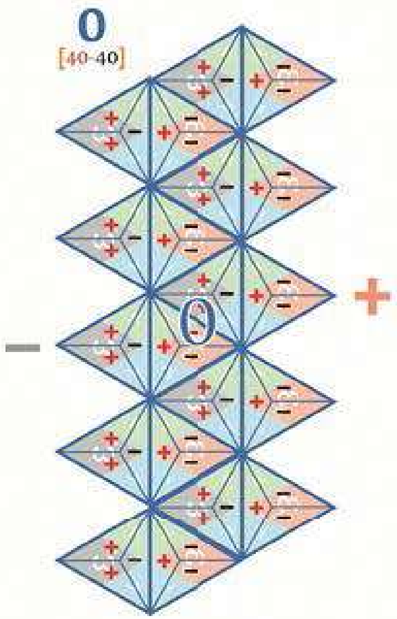

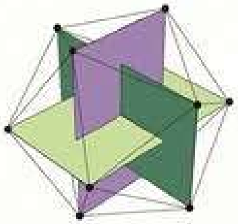

The 3 internalised planes formed by the external apex points correspond with the 3 spatial dimensions of Cartesian geometry

Copyright ABRAHAM [2008] - all rights reserved

Non-gluonic Baryon formation

Non-neutral 'UP' quark
has +4 charge as opposed
to +8 charge of UP quark

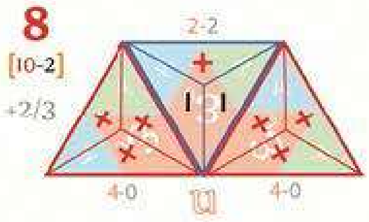
8
[10-2]
+2/3
2-2
4-0
u
4-0

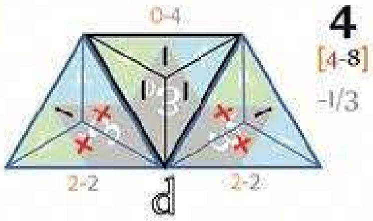
0-4
4
[4-8]
-1/3
2-2
d
2-2

Non-neutral 'DOWN' quark
has +4 charge as opposed
to -4 charge of DOWN quark

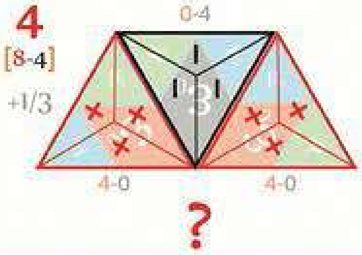
4
[8-4]
+1/3
0-4
4-0
4-0
?

Baryonic Particles can
be formed with non-neutral dielectric Tetryons
in lieu of the usual neutral Tetyrons [Gluons] -
altering the nett charges of the quarks formed

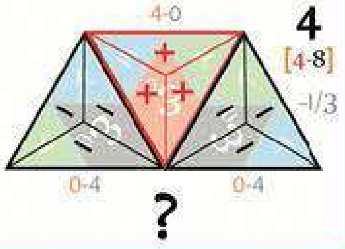
4-0
4
[4-8]
-1/3
0-4
0-4
?

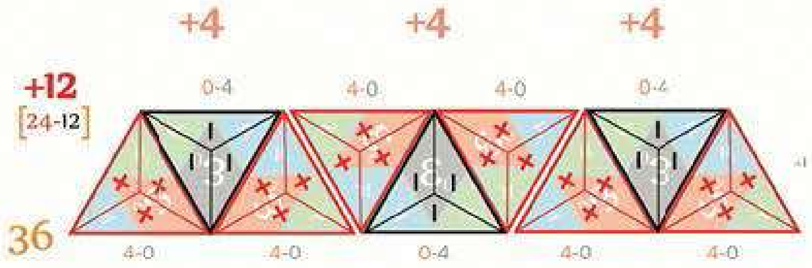
+4
+4
+4
+12
[24-12]
0-4
4-0
4-0
0-4
-1
36
4-0
4-0
0-4
4-0
4-0

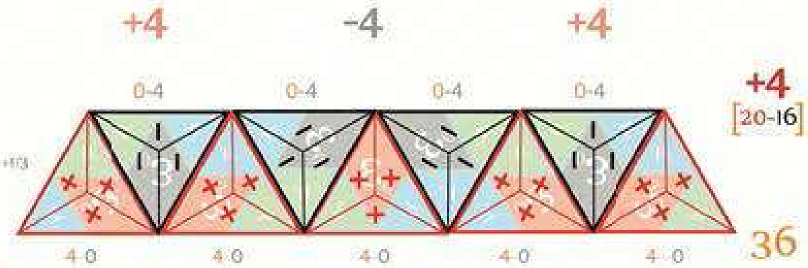
+4
-4
+4
0-4
0-4
0-4
0-4
+4
[20-16]
+1/3
36
4-0
4-0
4-0
4-0
4-0

Un-stable form

[rapidly decays into constituent Quarks]

Resulting in another family of Baryons [CHARGEBALLS]

Stable Form

[possibly mistakenly viewed as Bottom Quarks]

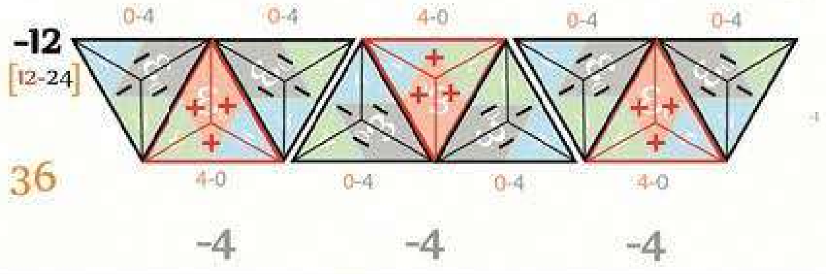
0-4
0-4
4-0
0-4
0-4
-12
[12-24]
-1
36
4-0
0-4
0-4
4-0
-4
-4
-4

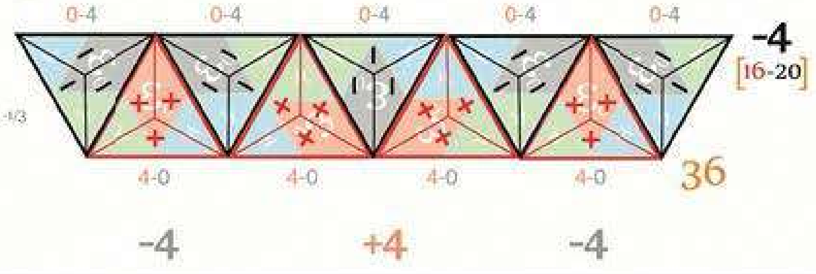
0-4
0-4
0-4
0-4
0-4
-4
[16-20]
-1/3
36
4-0
4-0
4-0
4-0
-4
+4
-4

Electro-Magnetic fields

Quantised Charge
is the nett geometry of mass-Energy-Matter formed by equilateral ElectroMagnetic fields and can be modelled with classical vector flux rotations

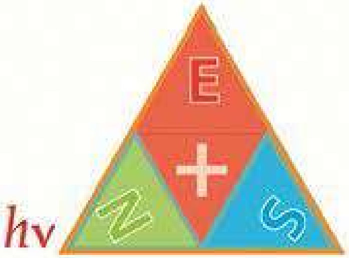

Positive Charge ZPF

Magnetic Monopoles do NOT exist

Negative Charge ZPF

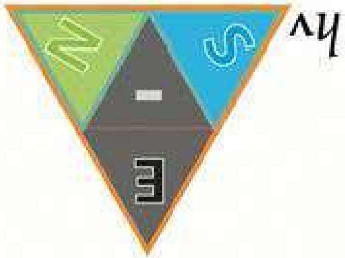

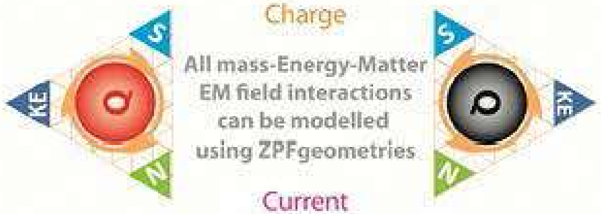

Charge

All mass-Energy-Matter EM field interactions can be modelled using ZPFgeometries

Current

ε_0 The charge geometry of all EM fields determines their Electric Permittivity and Magnetic Permeability μ_0

The 'zero point' field has scalar EM energies

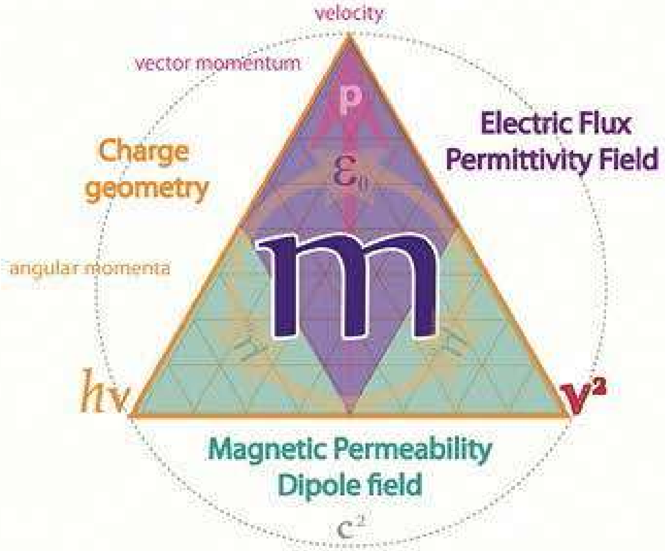
velocity
vector momentum
Electric Flux Permittivity Field
Charge geometry
angular momenta
p
ε0
m
hv
v²
Magnetic Permeability Dipole field
c²

The EM field is a equilateral waveform with the Magnetic field always orthogonal to the Electric

The Magnetic fields propogate bi-directionally and the Electric field is responsible for producing linear momentum

The Electric field and Magnetic fields are equal to each other and directly proportional to the velocity of propagation

Photons
are the neutral quanta of EM waves they are comprised of two opposite charge EM field quanta

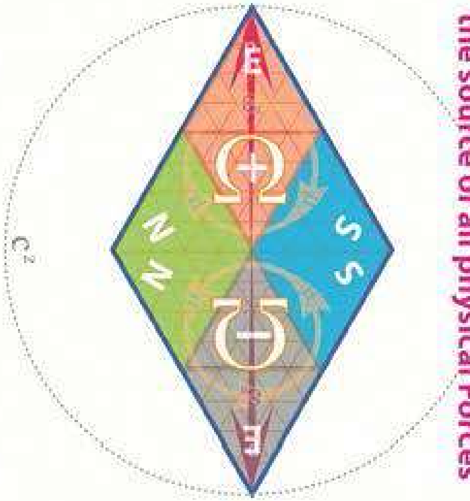
E
Ω
+
N
N
S
S
Ω
-
E
c²

Electro-Magnetic fields are the source of all physical Forces

Zero Point fields are polarised and are the sources of

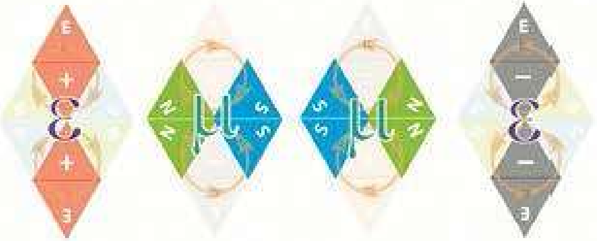
E
+
ε
+
E
N
μ
S
S
μ
N
E
-
ε
-
E

Electro-static and **Magneto-static** fields and particles

The Weak Force

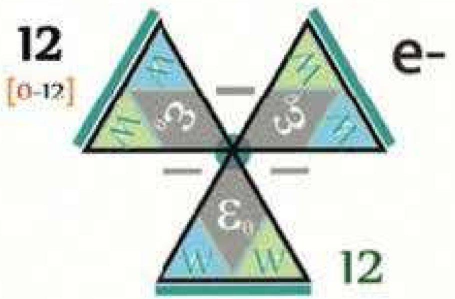
12
[0-12]
e-
ε₀
12

6 Inductive 'Weak' edges
(3 internalised)
(3 externalised)

Quantum Energy is stored or released, and distributed throughout Tetryonic geometries in ODD number quanta [Bosons]

Quantum Inductive loops

Quantum Tank circuit

Energy received is stored in quantum inductive loops
(until its release via weak interaction - Boson exchange)

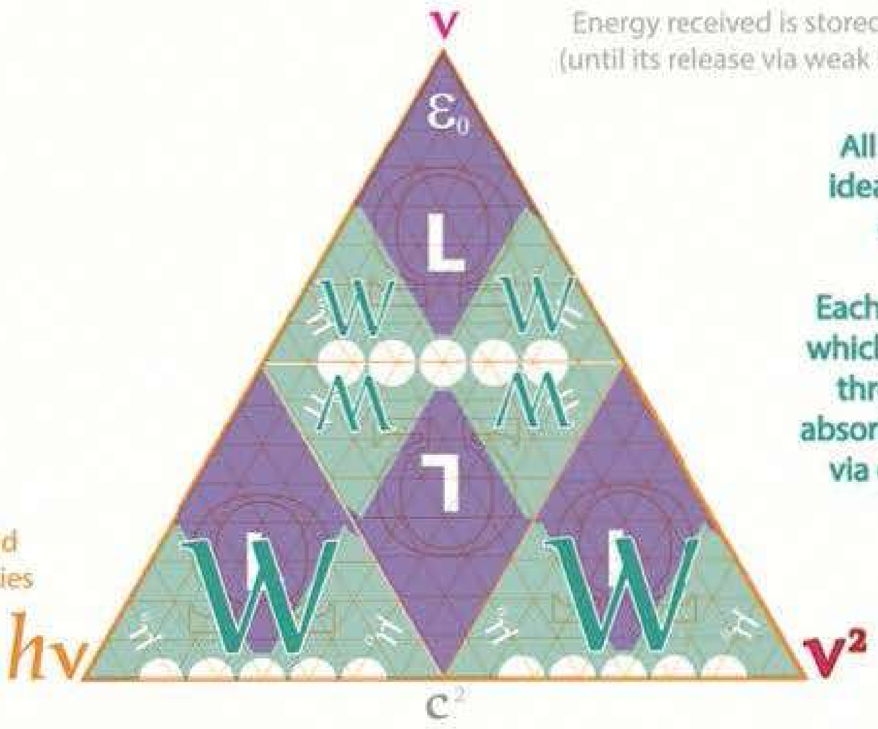
v
ε₀
L
W
W
W
W
L
W
W
hv
v²
c²

All ElectroMagnetic fields are ideal quantum inductive loops of EM energy momenta

Each has a Magnetic edge (base) which acts as a quantum inductor through which energy can be absorbed or released in EM quanta via electroMagnetic induction [exchange Bosons]

Leptons and Quarks
have the same Tetryonic numbers
but differing Matter topologies

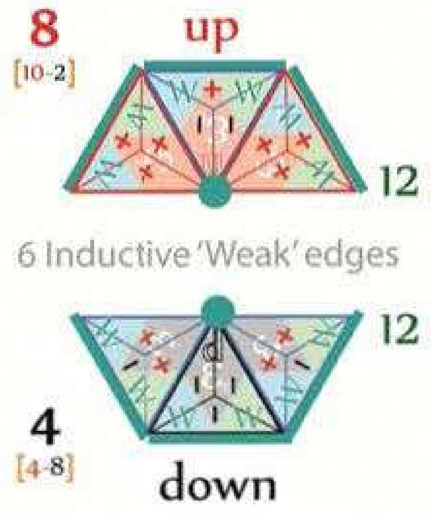
8
[10-2]
up
12
12
4
[4-8]
down

6 Inductive 'Weak' edges

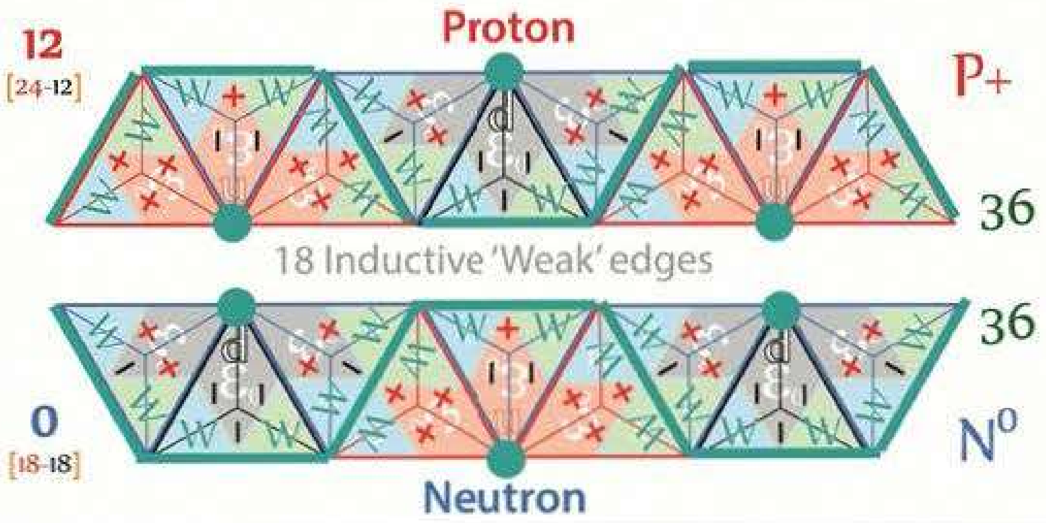
12
[24-12]
Proton
P+
36
36
0
[18-18]
N⁰
Neutron

18 Inductive 'Weak' edges

The Strong Force

The force resulting from the interaction of 2 Tetryons along their parallel planar (Electric charge) surfaces

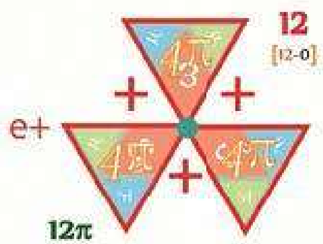

12
[12-0]
e+
+
+
+
4π
12π

Leptons Repulsive colour Force

Leptons are 'held together' by a radial Weak Force axis

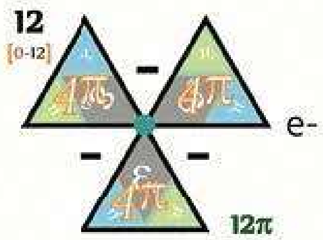

12
[0-12]
-
-
-
e-
4π
12π

Leptons and Quarks result from the Strong 'colour' force interactions in dodeca-deltahedrons

(In Quarks the attractive strong force produces Octahedral topologies)
(In Leptons the repulsive strong force produces Dodecahedral topologies)

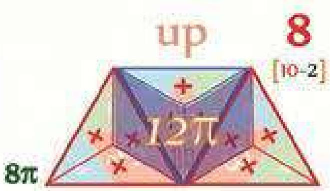

up
8
[10-2]
12π
8π

Quarks
4 Fascia 'Colour' Attraction

Quarks Attractive colour Force

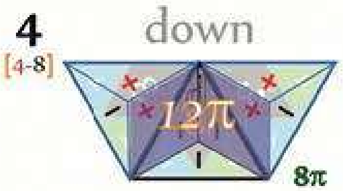

4
[4-8]
down
12π
8π

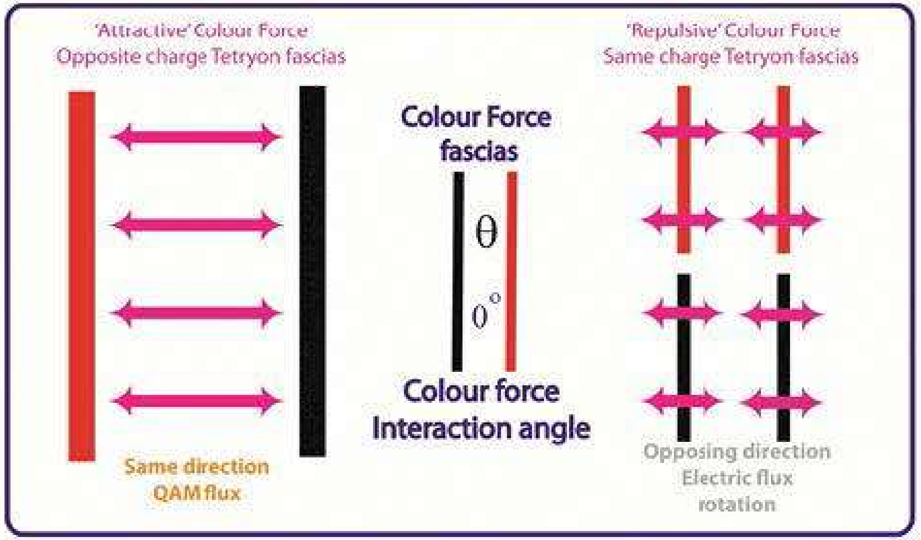

'Attractive' Colour Force
Opposite charge Tetryon fascias
'Repulsive' Colour Force
Same charge Tetryon fascias
Colour Force fascias
θ
0°
Colour force Interaction angle
Same direction QAM flux
Opposing direction Electric flux rotation

Strong Force strength is directly proportional to the nett mass-energy quanta involved
(Increases/decreases as total Tetryon mass-energy quanta increases/decreases)

Proton

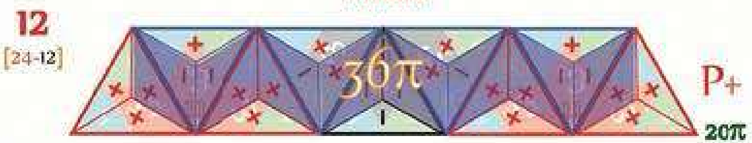

12
[24-12]
36π
P+
20π

Baryons Attractive colour Forces

Baryon formation
16 Fascia 'Colour' Attraction

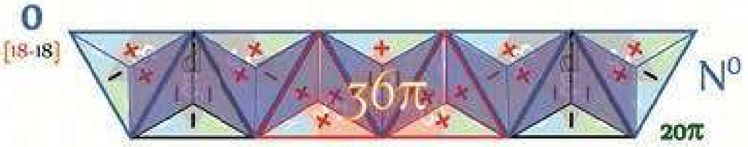

0
[18-18]
36π
N0
20π

Neutron

The Strong Colour Force is the binding force that holds Nuclei together

Tetryonic Fascia (Planar Electric Charge) interaction

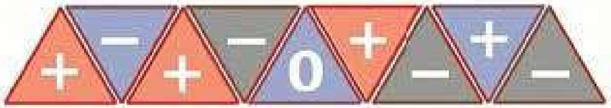

It interacts through oppositely charged Tetryon Fascias in contact with each other in all Matter (save Leptons)

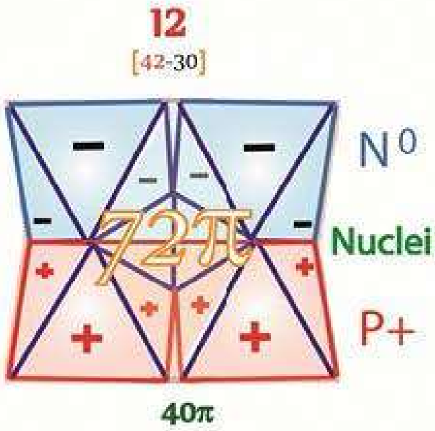

12
[42-30]
72π
N0
Nuclei
P+
40π

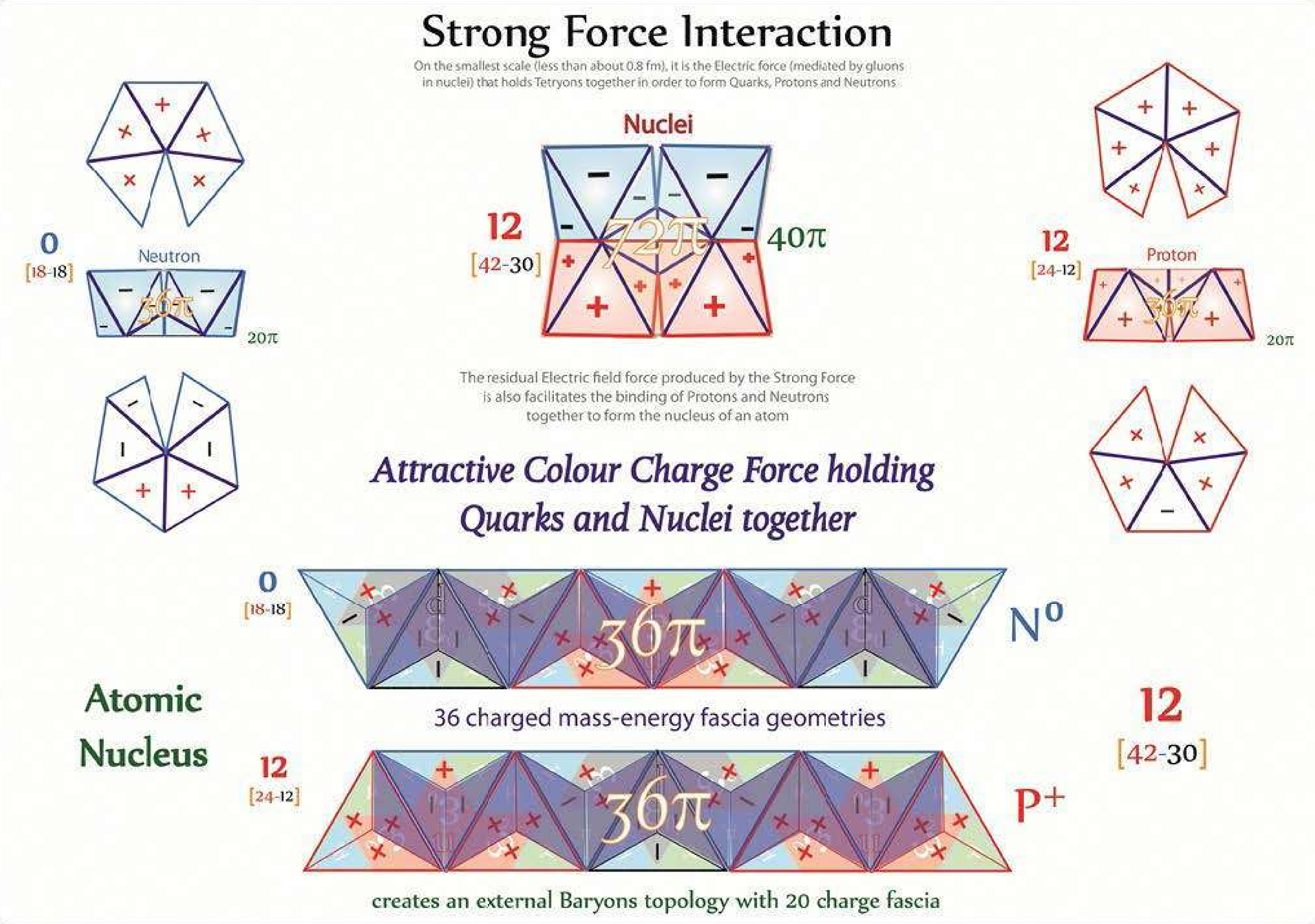

Strong Force Interaction

On the smallest scale (less than about 0.8 fm), it is the Electric force (mediated by gluons in nuclei) that holds Tetryons together in order to form Quarks, Protons and Neutrons

The residual Electric field force produced by the Strong Force is also facilitates the binding of Protons and Neutrons together to form the nucleus of an atom

Attractive Colour Charge Force holding Quarks and Nuclei together

36 charged mass-energy fascia geometries

creates an external Baryons topology with 20 charge fascia

Nuclei
12
[42-30]
72π
40π
0
[18-18]
Neutron
36π
20π
12
[24-12]
Proton
36π
20π
0
[18-18]
36π
N⁰
Atomic Nucleus
12
[42-30]
12
[24-12]
36π
P+

Tetryonics 12.04 - Strong Force interaction

Copyright ABRAHAM [2008] - all rights reserved

Proton Hydrogen ion

36

1 Proton	24-12	
Total Charge	24-12	[+12]

H^+

F+ n=25

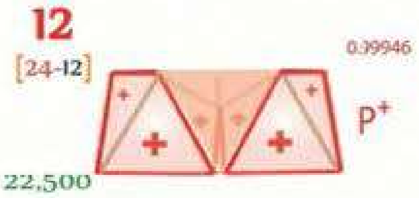

Proton - Hydrogen ion

The Proton is a Baryon with a Positive Tetryonic charge geometry of +12 [24-12] [and a final Matter topology of 20π]

It attracts electrons [0-12] and Neutrons [18-18] through their opposite charged facia in order to reach charge equilibrium

The Proton's opposite charge particle is the electron
If it binds with one or more Neutrons it will still attract electrons in order to acheive Tetryonic charge equilibrium.

As a Hydrogen nuclei it is highly reactive

Neutral Hydrogen

48

1 Proton	24-12	
1 electron	0-12	
Total Charge	24-24	[0]

H^0

F+ n=25
e- n=1

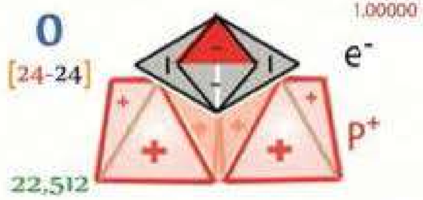

Neutral Hydrogen

Hydrogen is the Second Neutral Tetryonic geometry [Baryon] formed in the creation of Matter

It is a [24-24] NEUTRAL [balanced] Tetryonic charge geometry

The Neutral Hydrogen Atom interacts with Deuterium and other Elements via its external nucleon fascia charges (residual EM Interactions)

Deuterium

84

1 Proton	24-12	
1 Neutron	18-18	
1 electron	0-12	
Total Charge	42-42	[0]

D

N+ n=25
F+ n=25
e- n=1

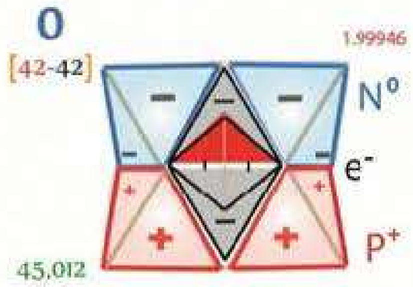

Deuterium

Deuterium has the same nett charge as Neutral Hydrogen but has an increased Tetryonic mass- Matter topology

It has an additional 9 Tetryons in its nucleus (due to the Neutron) and consequently is larger than the Hydrogen nucleus

It has a [42-42] NEUTRAL Tetryonic charge geometry
Deuterium nuclei combine to form all other elements

Tritium

84

1 Proton	24-12	
1 Neutron	18-18	
1 electron	0-12	
Total Charge	42-42	[0]

D

N+ n=31
F+ n=31
e- n=7
Radioactive Isotope

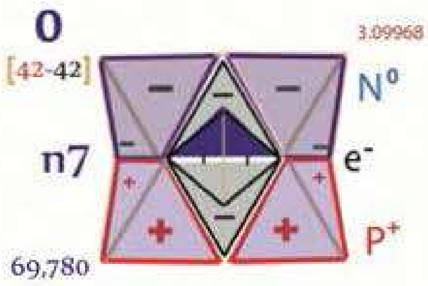

Tritium

(Radioactive Deuterium)

Tritium has the same nett Tetryonic charge as Neutral Hydrogen and Deuterium (D) but due to its higher quantum energy levels it has additional mass-energy quanta in its charged geometry

It has the same 84 charged fascia in its nucleus as Deuterium but its increased mass-energy is equivalent to that of 1 Neutron.

This extra nuclear energy is the source of its radioactivity and ongoing confusion in chemistry where it is released as a variety of decay particles over time

Copyright ABRAHAM [2008] - all rights reserved

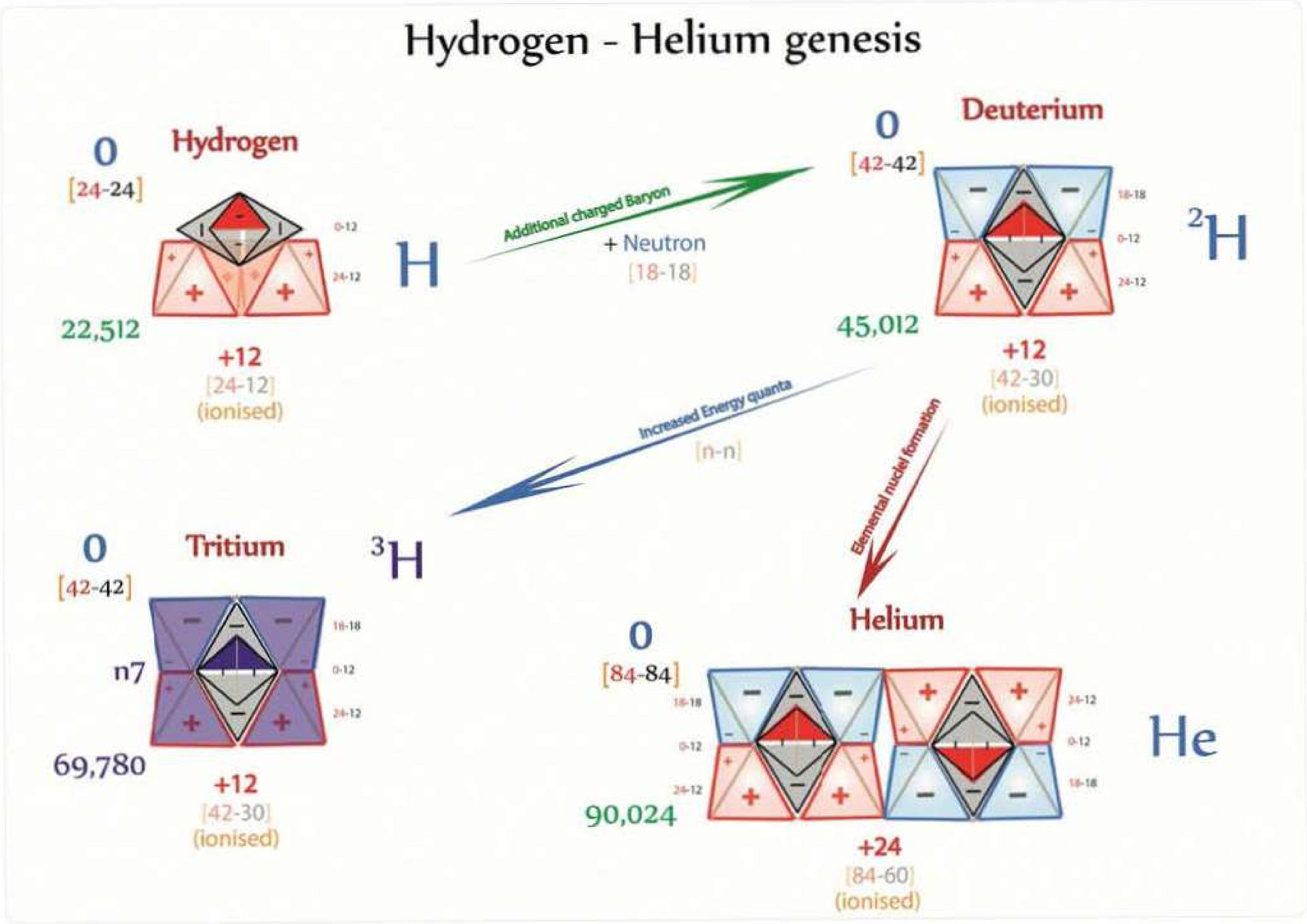
Hydrogen - Helium genesis
0
Hydrogen
[24-24]
0-12
24-12
H
22,512
+12
[24-12]
(ionised)
Additional charged Baryon
+ Neutron
[18-18]
0
Deuterium
[42-42]
18-18
0-12
24-12
²H
45,012
+12
[42-30]
(ionised)
Increased Energy quanta
[n-n]
Elemental nuclei formation
0
Tritium
³H
[42-42]
18-18
0-12
24-12
n7
69,780
+12
[42-30]
(ionised)
0
Helium
[84-84]
18-18
0-12
24-12
24-12
0-12
18-18
He
90,024
+24
[84-60]
(ionised)

Tetryonics 13.02 - Hydrogen-Helium genesis

Hydrogen Family

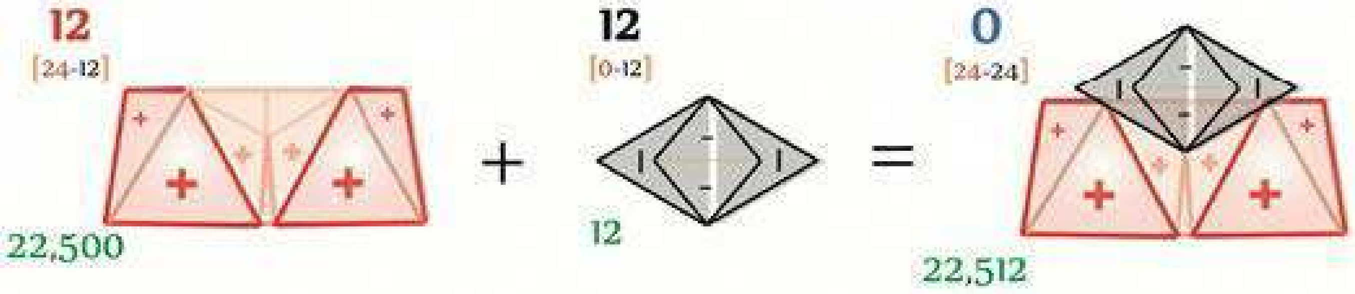
12
[24-12]
22,500
+
12
[0-12]
12
=
0
[24-24]
22,512

Proton 36π

electron 12π

Hydrogen 48π

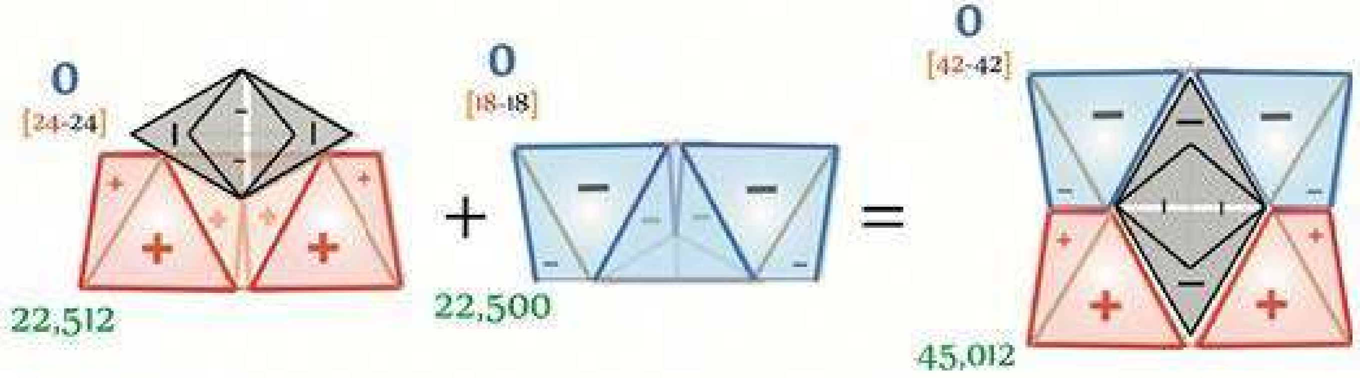
0
[24-24]
22,512
+
0
[18-18]
22,500
=
0
[42-42]
45,012

Hydrogen 48π

Neutron 36π

Deuterium 84π

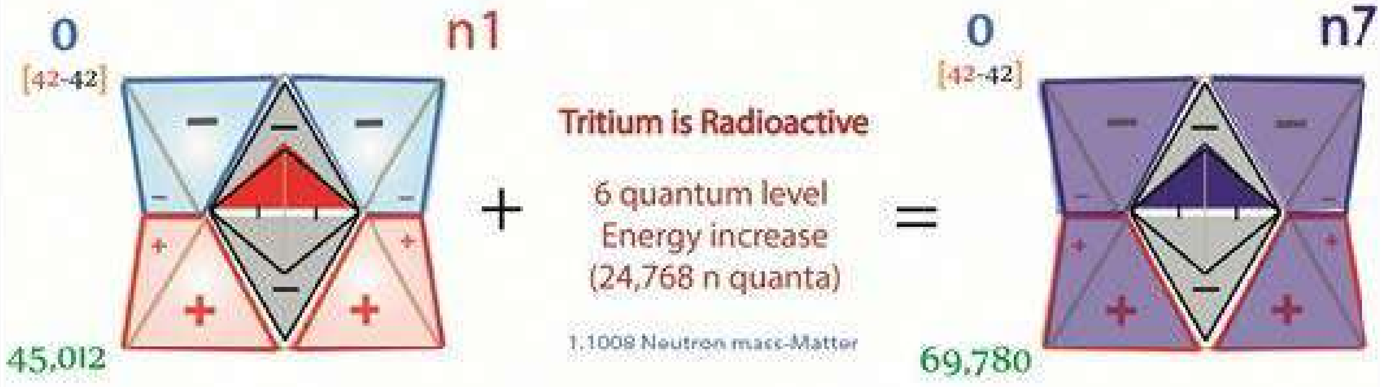
0
[42-42]
n1
45,012
+
=
0
[42-42]
n7
69,780

Tritium is Radioactive

6 quantum level
Energy increase
(24,768 n quanta)

1.1008 Neutron mass-Matter

Deuterium 84π

Tritium 84π

Copyright ABRAHAM [2008] - all rights reserved

Copyright ABRAHAM [2008] - all rights reserved

Hydrogen-Helium3 genesis

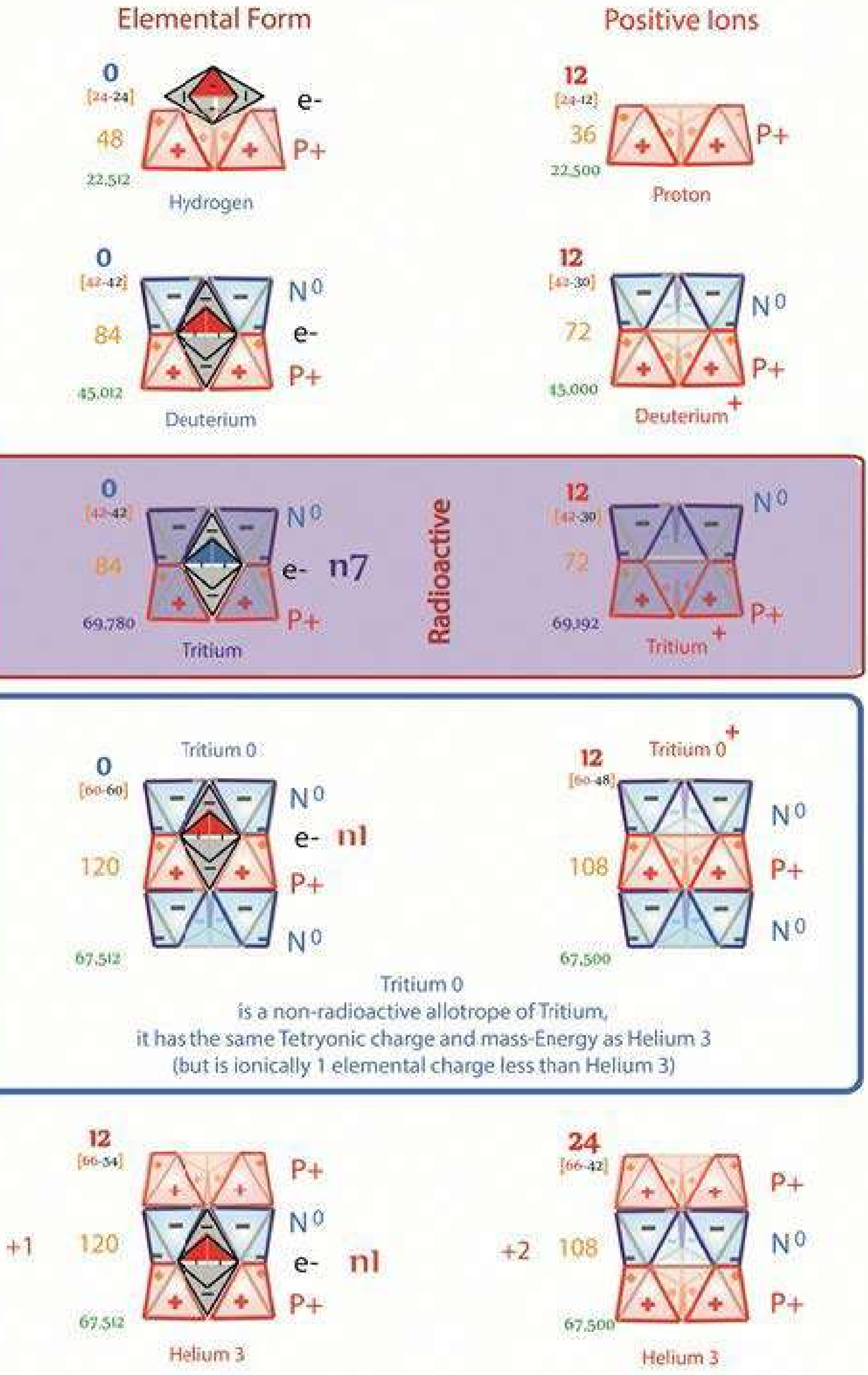
Elemental Form
Positive Ions
0
[24-24]
48
22.512
e-
P+
Hydrogen
12
[24-12]
36
22.500
P+
Proton
0
[42-42]
84
45.012
N0
e-
P+
Deuterium
12
[42-30]
72
45.000
N0
P+
Deuterium+
0
[42-42]
84
69.780
N0
e-
n7
P+
Tritium
Radioactive
12
[42-30]
72
69.192
N0
P+
Tritium+
Tritium 0
0
[60-60]
120
67.512
N0
e-
n1
P+
N0
Tritium 0+
12
[60-48]
108
67.500
N0
P+
N0
12
[66-54]
+1
120
67.512
P+
N0
e-
n1
P+
Helium 3
24
[66-42]
+2
108
67.500
P+
N0
P+
Helium 3

Tritium 0
is a non-radioactive allotrope of Tritium,
it has the same Tetryonic charge and mass-Energy as Helium 3
(but is ionically 1 elemental charge less than Helium 3)

Quantum Batteries

Atomic nuclei can be easily scaled to non-quantum sizes to offer clean, safe and portable long term energy storage devices that can store energy indefinitely and release it on demand anywhere in the World

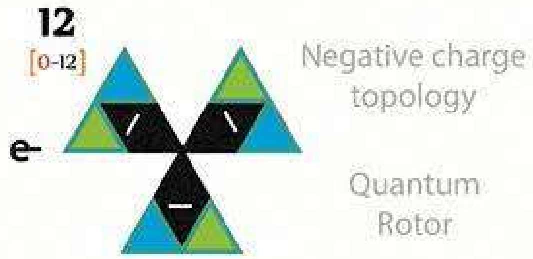
12
[0-12]
e-
Negative charge topology
Quantum Rotor

12 loop quantum inductive rotor

The quantum battery is unique in that in addition to storing energy indefinitely, when the nuclei combine with a lepton it has the ability to release specific energies [photons] by way of its synchronous quantum convertor topology

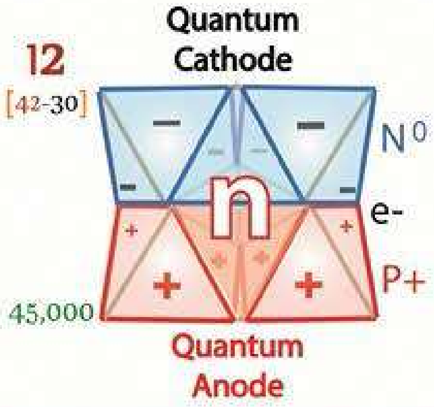
Quantum Cathode
12
[42-30]
n
N 0
e-
P+
45,000
Quantum Anode

Energy stored in quantum batteries is measured as mass

(Atomic Nuclei)

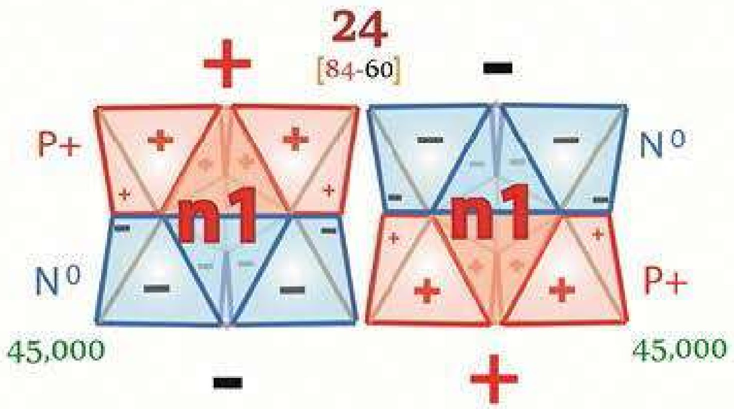
24
[84-60]
+
-
P+
N 0
n1
n1
N 0
P+
45,000
45,000
-
+

anti-Parallel Configuration

+E90
03-2013
Energizer

The non-neutral charge of atomic nuclei attract free leptons into 'bound' states within the various **n** levels of atomic shells releasing energy as spectral photons

Series Configuration

+E90
03-2013
Energizer

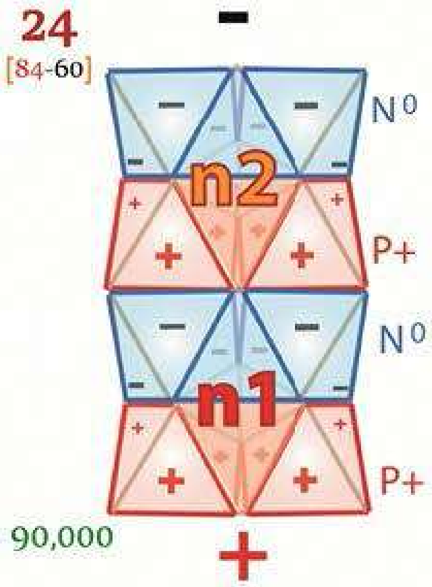
24
[84-60]
-
N 0
n2
P+
N 0
n1
P+
90,000
+

Copyright ABRAHAM [2008] - all rights reserved

Quantum synchronous Converters

With the addition of a quantum rotor (lepton) a quantum battery can be converted from a storage device into a energy distribution device.

And just like the quantum battery, the quantum convertor can be scaled to any size in order to provide tailor-made energy efficient delivery devices

The electron can be viewed as a rotating inductor consisting of 3 negative Tetryons

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]_{n1}^{n8}$$

Leptons ElectroMagnetic mass velocity

The electron has a charged Matter topology that is electrically equivalent to a quantum 6 loop inductive rotor

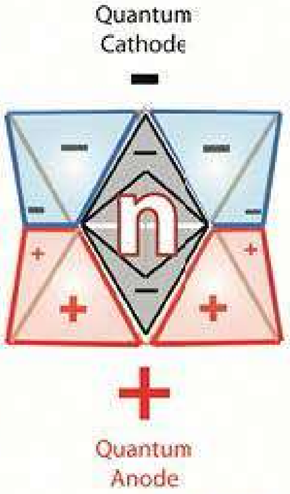

EM Field Planck quanta

$$72\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]_{n25}^{n32}$$

Nuclei ElectroMagnetic mass velocity

EM Field Planck quanta

$$84\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]_{n25}^{n32}$$

Deuterons ElectroMagnetic mass velocity

Where varying levels and frequencies of Energy are transmitted long distances and need to be stored for later release on demand the 'ideal' mechanical device is the rotating (or synchronous) convertor

Changes in energy-momenta results in photon emission/absorption lines photon emission/absorption produces changes in leptronic energy-QAM and results in the quantum transition of electrons in atomic orbitals

Changes in Baryonic energy levels induces a directly proportional change in electron energy levels

3 forms of mass-energy momenta are possessed by quantum convertors

Angular Momentum (motional energy)
linear momentum of Baryons & leptons

Photonic spectral energies (emission/absorption)
KEM fields geometries of photo-electrons

Nuclear mass-energies (stored energy quanta)
standing wave mass-energies of Matter topologies

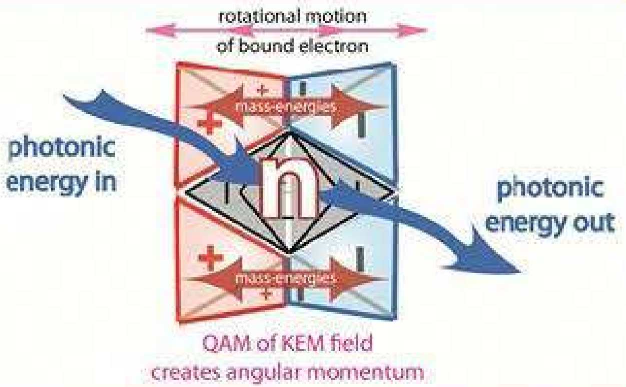

A change in any 1 of the 3 types of energy in a atom results in a proportional change in the other 2

External EM fields and incident photons can all affect the quantum energy levels of the atomic nuclei

Copyright ABRAHAM [2008] - all rights reserved

$$Mv^2 = KEM = hv^2$$

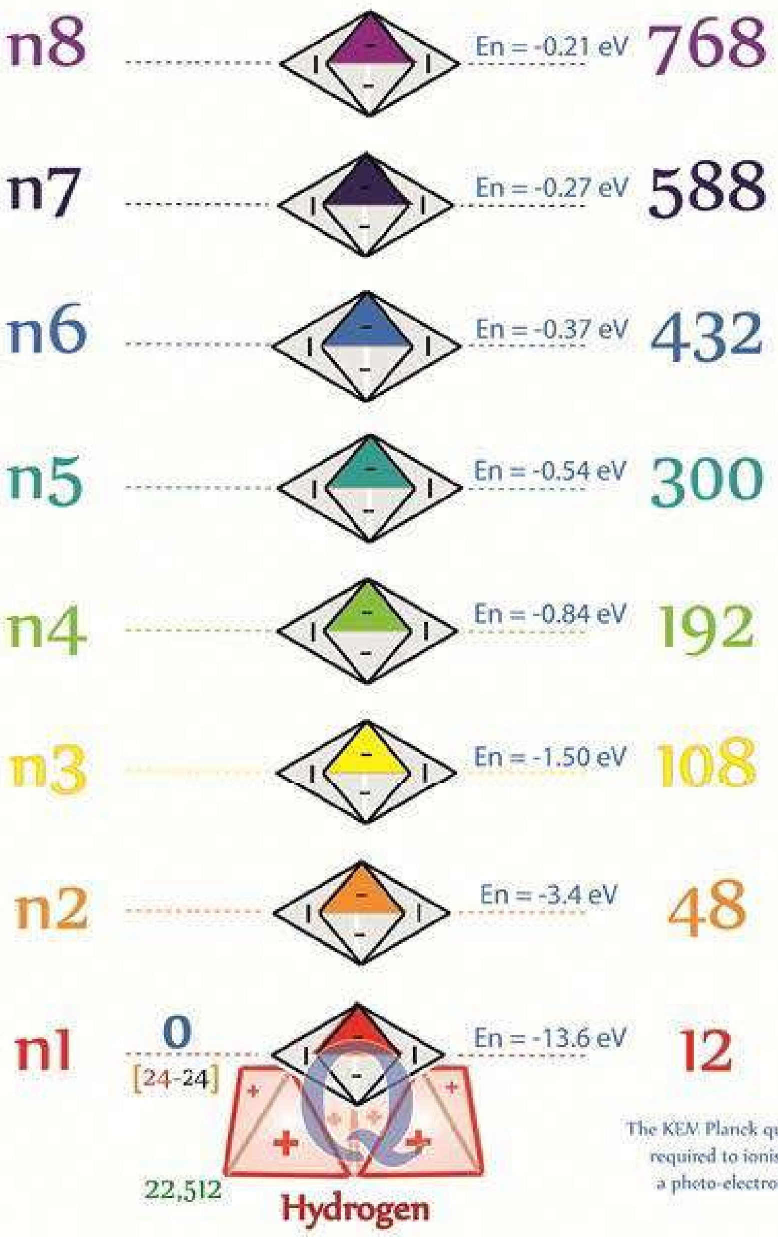
n8
En = -0.21 eV
768
n7
En = -0.27 eV
588
n6
En = -0.37 eV
432
n5
En = -0.54 eV
300
n4
En = -0.84 eV
192
n3
En = -1.50 eV
108
n2
En = -3.4 eV
48
n1
0
[24-24]
En = -13.6 eV
12
22,512
Hydrogen

The KEM Planck quanta
required to ionise
a photo-electron

Hydrogen Ionisation Energies

The ionisation energy level for each quantum level is proportional to the square of the quantum number

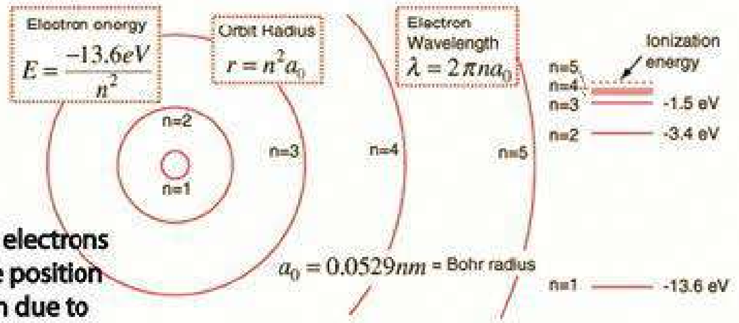
Electron energy
$E = \frac{-13.6eV}{n^2}$
Orbit Radius
$r = n^2 a_0$
Electron Wavelength
$\lambda = 2\pi n a_0$
n=1
n=2
n=3
n=4
n=5
$a_0 = 0.0529nm$ = Bohr radius
Ionization energy
n=5
n=4
n=3
-1.5 eV
n=2
-3.4 eV
n=1
-13.6 eV

In Hydrogen nuclei electrons remain in the same position close to the Proton due to Coulombic attraction.

The established model of electron orbitals having circular radii of increasing size in proportion to their energies is incorrect

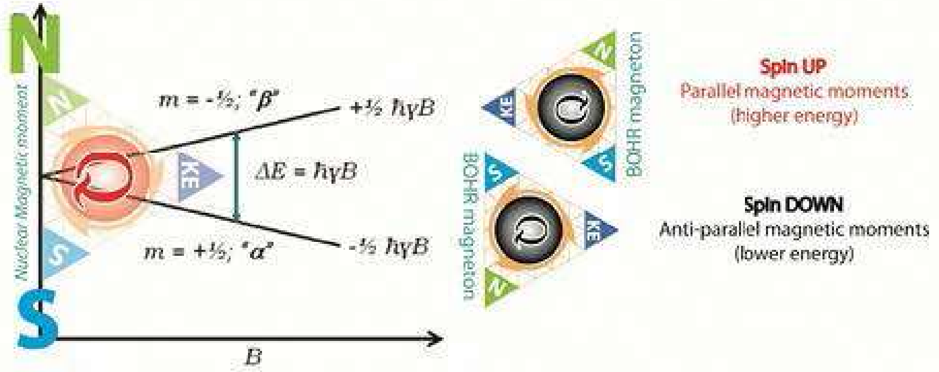
N
Nuclear Magnetic moment
S
N
S
KE
$m = -½; "\beta"$
$+½ \hbar\gamma B$
$\Delta E = \hbar\gamma B$
$m = +½; "\alpha"$
$-½ \hbar\gamma B$
B
BOHR magneton
BOHR magneton
Spin UP
Parallel magnetic moments
(higher energy)
Spin DOWN
Anti-parallel magnetic moments
(lower energy)

The intrinsic magnetic moment of the rotating electron couples wrt the nuclear magnetic moment to produce a split in the resulting spectral lines produced by quantum level transitions

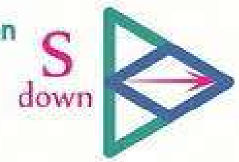

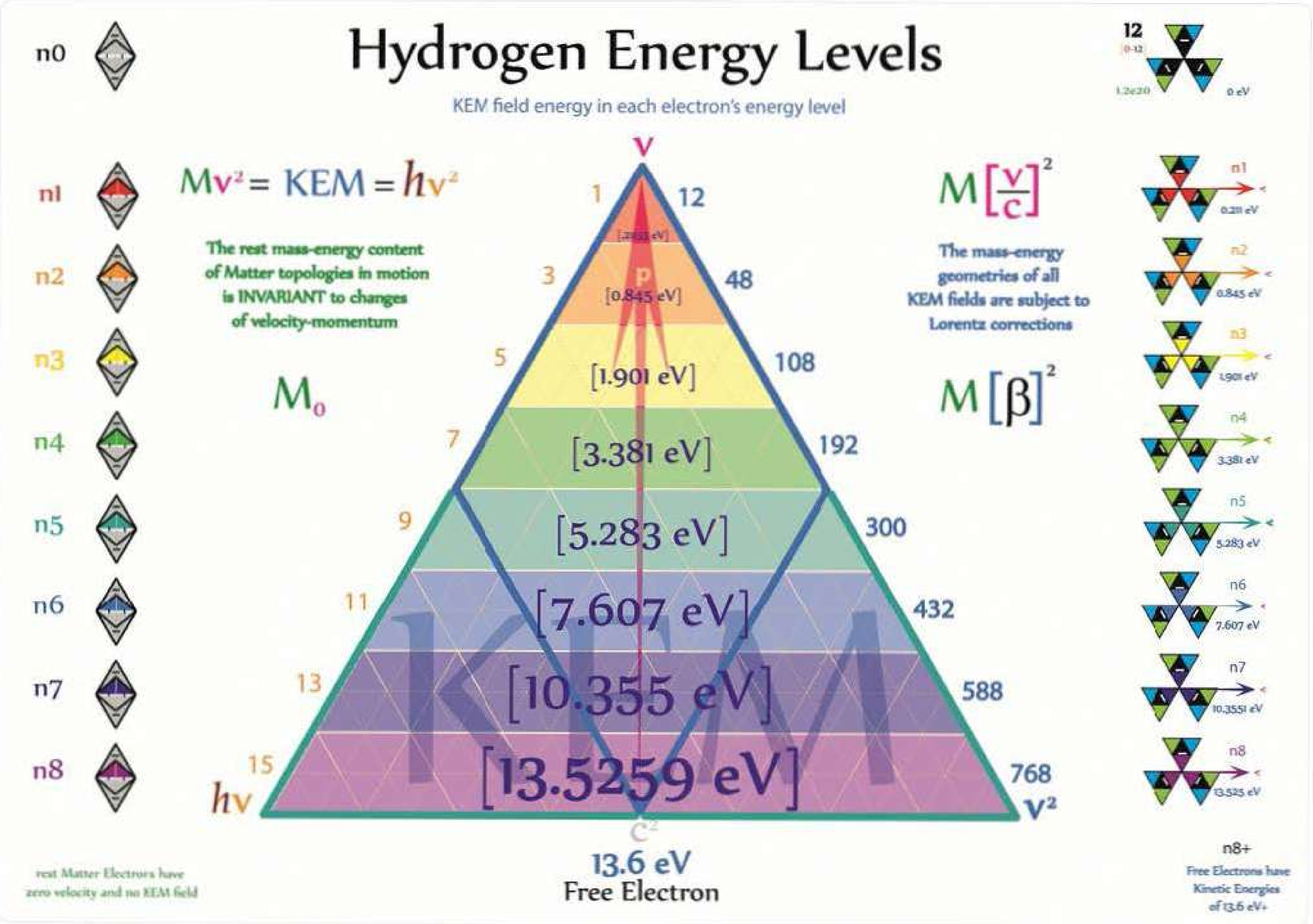

Hydrogen Energy Levels

KEM field energy in each electron's energy level

$Mv^2 = KEM = hv^2$

The rest mass-energy content
of Matter topologies in motion
is INVARIANT to changes
of velocity-momentum

M_0

$M\left[\frac{v}{c}\right]^2$

The mass-energy
geometries of all
KEM fields are subject to
Lorentz corrections

$M[\beta]^2$

13.6 eV
Free Electron

rest Matter Electrons have
zero velocity and no KEM field

n8+
Free Electrons have
Kinetic Energies
of 13.6 eV+

n0
n1
n2
n3
n4
n5
n6
n7
n8
12
0-12
1.2e20
0 eV
v
1
3
5
7
9
11
13
15
12
48
108
192
300
432
588
768
P
[.2113 eV]
[0.845 eV]
[1.901 eV]
[3.381 eV]
[5.283 eV]
[7.607 eV]
[10.355 eV]
[13.5259 eV]
KEM
hv
c²
v²
n1
0.211 eV
n2
0.845 eV
n3
1.901 eV
n4
3.381 eV
n5
5.283 eV
n6
7.607 eV
n7
10.3551 eV
n8
13.525 eV

Kinematics

An inertial frame of reference is one in which the motion of a particle not subject to forces and results in motion in a straight line at constant velocity

Deceleration

[Inertial Frame MINUS odd quanta]

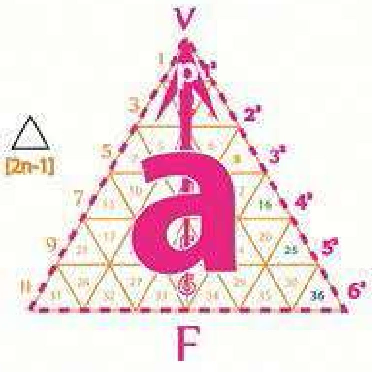

Newton's First Law

$$\sum \mathbf{F} = 0 \Rightarrow \frac{d\mathbf{v}}{dt} = 0$$

Every body persists in its state of being at rest or of moving uniformly straight forward, except insofar as it is compelled to change its state by force impressed

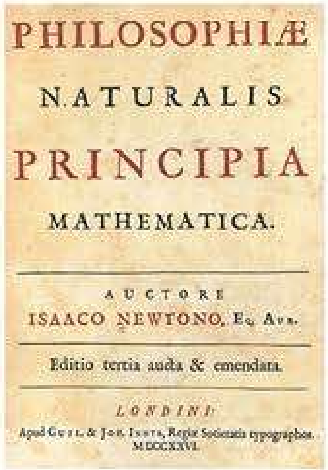

PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

AUCTORE
ISAACO NEWTONO, Eq. Aur.

Editio tertia aucta & emendata.

LONDINI:
Apud Guil. & Joh. Innys, Regiæ Societatis typographos.
MDCCXXVI.

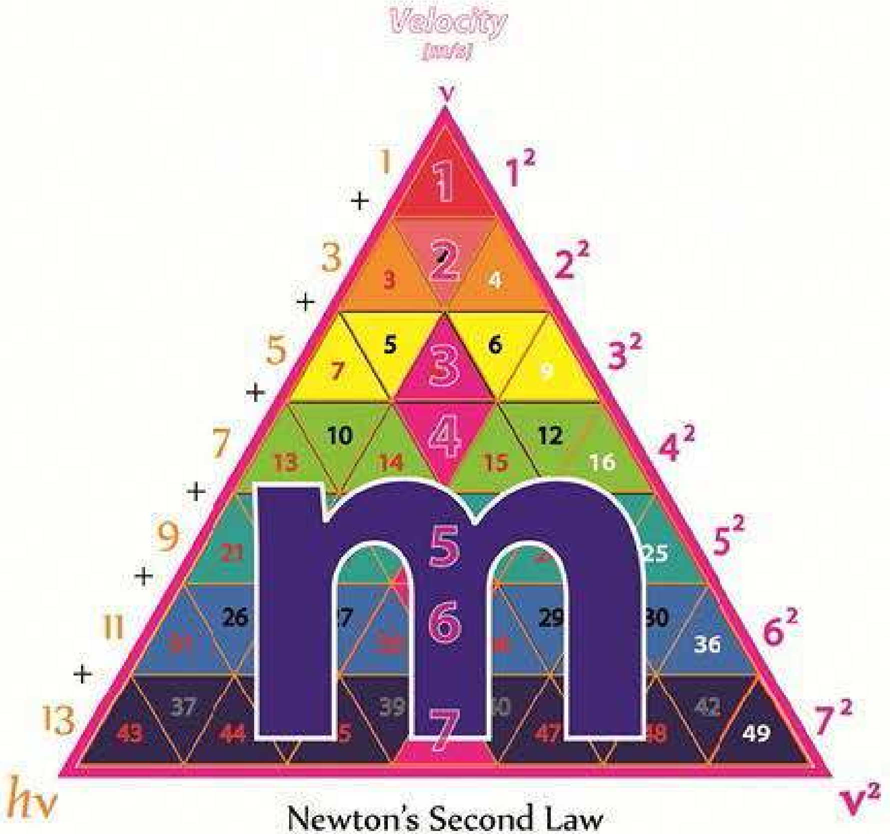

Newton's Second Law

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}_{cm}}{\mathrm{d}t} = m\mathbf{a}_{cm}$$

Force creates a change in Momentum over Time

A body of [m]ass subject to a net [F]orce undergoes an [a]cceleration that has the same direction as the force and a magnitude that is directly proportional to the force and inversely proportional to the mass, i.e., F = ma.

Alternatively, the total force applied on a body is equal to the time derivative of linear momentum of the body.

Sir Issac Newton

(25 December 1642 – 20 March 1727)

$$a = \frac{\Delta y}{\Delta x} = \frac{\Delta v}{\Delta t}.$$

$$\mathbf{F} = m\mathbf{a} \quad \longrightarrow \quad \mathbf{a} = \mathbf{F}/m$$

Acceleration

[Inertial Frame PLUS odd quanta]

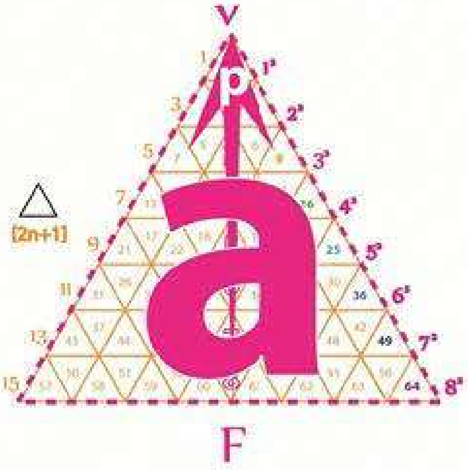

Copyright ABRAHAM [2008] - all rights reserved

Inertia and Force

Any change to the energy-momenta content of a closed inductive loop requires a proportional change to the loop's energy density

momentum

The classical definition of Momentum relates the mass of a material body at given velocity (v) to its Momentum (p); it is a proportionality factor in the formula

$$p = mv$$

Force

This meaning of a body's inertia therefore is altered from the classsical definition of "a tendency to maintain momentum" to a description of the measure of how difficult it is to change the momentum of a body

$$F = ma$$

linear quanta forces

divergent ↑ $[v - v]$ ↓ convergent

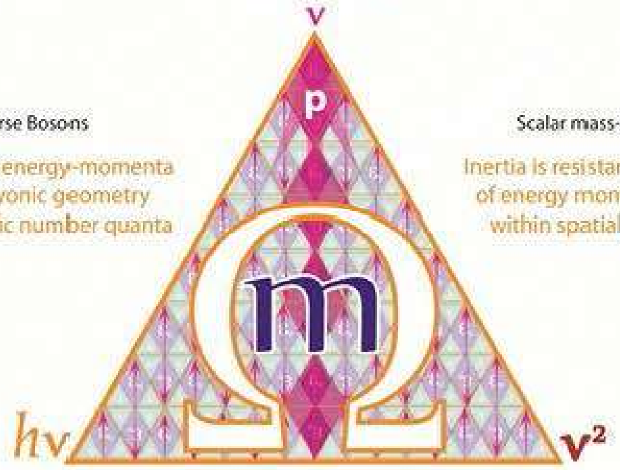

Transverse Bosons

Any change to energy-momenta levels in Tetryonic geometry requires specific number quanta

Scalar mass-energy

Inertia is resistance to change of energy momenta vectors within spatial geometries

[all inductive loops resist changes to their energy levels]

Inertia is the resistance of any physical object to a change in its state of motion or rest, or the tendency of an object to resist any change in its motion.

The principle of inertia is one of the fundamental principles of classical physics which is used to describe the motion of matter and how it is affected by applied forces

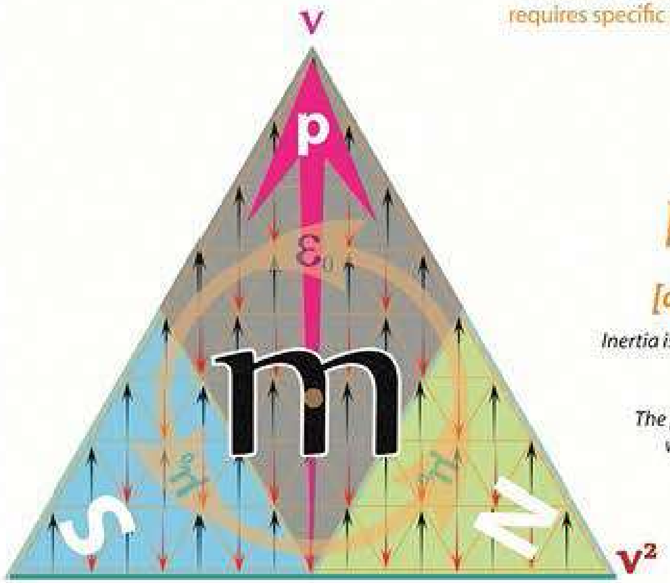

Linear momentum is the nett square root of mass-energy quanta

p

$$E = mv^2$$

Any change in the energy-momenta content of a body of mass-Matter results in a proportional change in its momentum-velocity

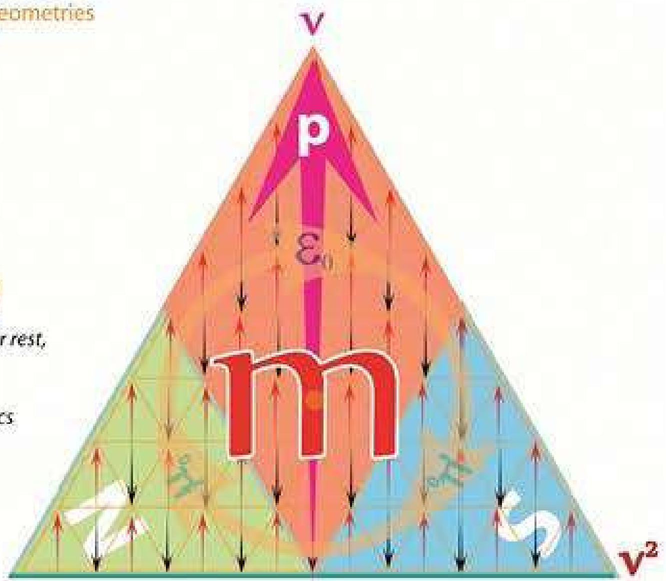

Force is the sum of the linear mass-energy momenta quanta

F

Tetryonics 14.02 - Inertia and Force

Copyright ABRAHAM [2008] - all rights reserved

Gottfried Wilhelm von Leibniz

(July 1, 1646 – November 14, 1716)

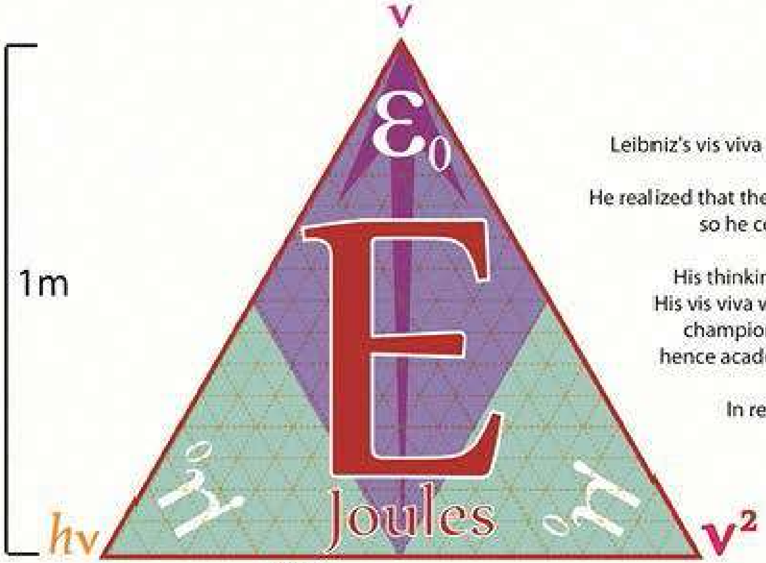

Scalar Energy

Leibniz's vis viva (Latin for living force) is mv², twice the modern kinetic energy.

He realized that the total energy would be conserved in certain mechanical systems, so he considered it an innate motive characteristic of Matter.

His thinking gave rise to another regrettable nationalistic dispute. His vis viva was seen as rivaling the conservation of linear momentum championed by Newton in England and by Descartes in France; hence academics in those countries tended to neglect Leibniz's idea.

In reality, both energy and momentum are conserved, so the two approaches are equally valid.

Newtons

of force

The nett direction of Force within energy geometry is UNIDIRECTIONAL
ie the force exerted is the result of the nett linear momenta irrespective of charge

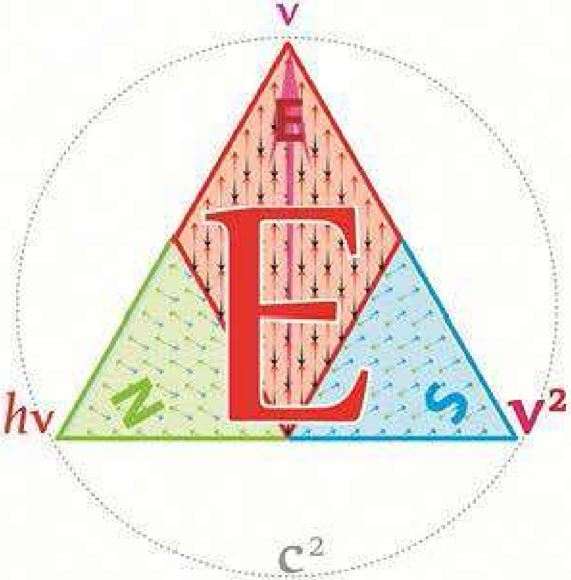

[Joule seconds] per second

mass x velocity squared $\text{kg}.\left[\frac{\text{m}}{\text{s}}\right]^2$

Energy $\text{kg}.\frac{\text{m}^2}{\text{s}^2}$

$$\mathrm{J} = \frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^2} = \mathrm{N}\cdot\mathrm{m} = \mathrm{Pa}\cdot\mathrm{m}^3 = \mathrm{W}\cdot\mathrm{s}$$

A Joule is equal to the energy expended (or work done) in applying
a force of one newton through a distance of one metre
(1 newton metre or N·m)

momentum x velocity $\left[\text{kg}.\frac{\text{m}}{\text{s}}\right].\frac{\text{m}}{\text{s}}$

Planck's quanta per second $\left[\text{kg}.\frac{\text{m}^2}{\text{s}}\right].\text{s}^{-1}$

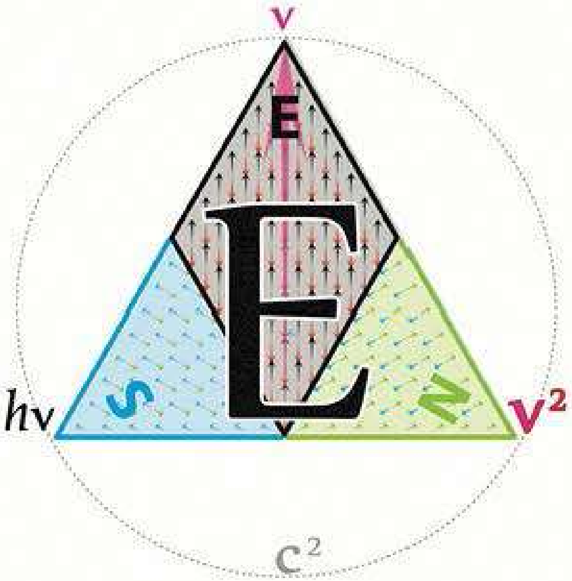

[Joule seconds] per second

Copyright ABRAHAM [2008] - all rights reserved

Linear Momentum

The Energy-momentum relationship is a fundamental physical property used to determine the mass of a body

Using the formula for mass-Energy equivalence as it relates to Photons moving at 'c'

$$E = h\nu = \frac{hc}{\lambda} = mc^2$$

Noting that the rest mass in the case of EM fields (Eosons and Photons) is to equal Zero

we can derive a relationship for Momentum - Energy - Wavelength showing that

$$p = E/c$$

Thus momentum in Photons is directly related to the EM energy content of the photon and the mass-energy content of any massive 3D body

$$p = \hbar k = \frac{h\nu}{c} = \frac{h}{\lambda}.$$

and momentum in Particles is related to the total EM Energies of an object (its rest Matter + KE) and the wavelengths associated with those distinct energy levels

$$p = \frac{E}{c} = \frac{h\nu}{c} = \frac{h}{\lambda}$$

Linear momentum is the SQUARE ROOT of equilateral Planck mass-energy geometries and produces a undirectional vector force

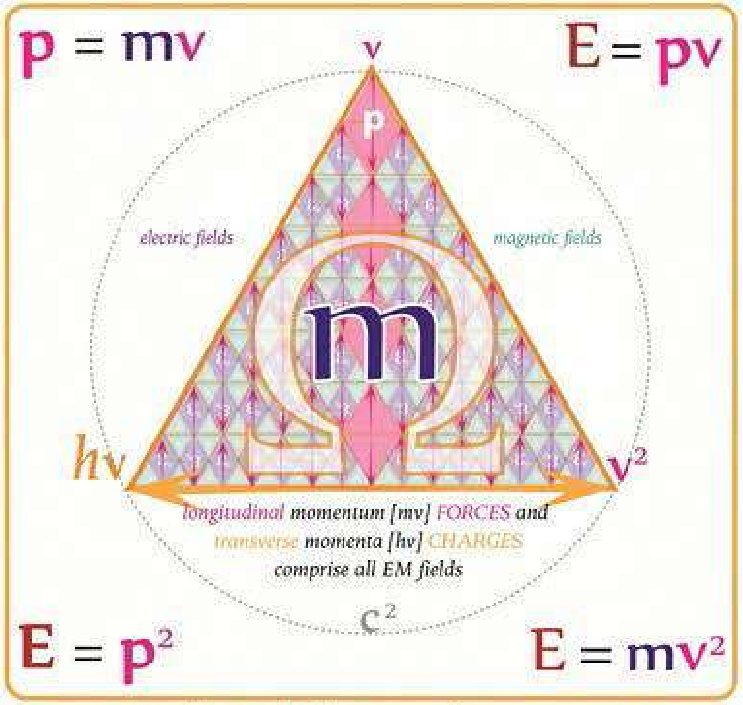

$$p = mv \qquad E = pv$$

longitudinal momentum [mv] FORCES and transverse momenta [hv] CHARGES comprise all EM fields

$$E = p^2 \qquad E = mv^2$$

Longitudinal [linear] momentum

$$p = n\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v}\right]\right]$$

Energy

momentum

Momentum is a conserved quantity, meaning that the total momentum of any closed system (one not affected by external forces) cannot change.

Linear Momentum is the intrinsic sqare root vector component of Force

$$p = mv$$

In classical mechanics, momentum is the product of the mass and velocity of an object.

$$\sum \mathbf{F} = \frac{d\mathbf{p}}{dt} = m\frac{d\mathbf{v}}{dt} + \mathbf{v}\frac{dm}{dt}$$

In relativistic mechanics, this quantity is multiplied by the Lorentz factor.

$$p = \frac{n\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]}{v} = mv$$

EM field momentum is a function of its energy density, and is directly proportional to the group velocity

$$\left[\sqrt{\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v}\right]^2}\right]$$

Vector linear momentum

is the square root of the KEM field energies produced by Matter in motion

Energy

$$n\pi\left[\left[m\Omega v\right]\right]$$

momenta

Classical formulation

$$p = mv$$

Linear Momentum
is the vector force of
mass times its velocity

Quantum formulation

$$p = \frac{E}{v}$$

Linear Momentum is
Energy dividec by the
square root of its quanta

Energy required to change
the velocity of a mass

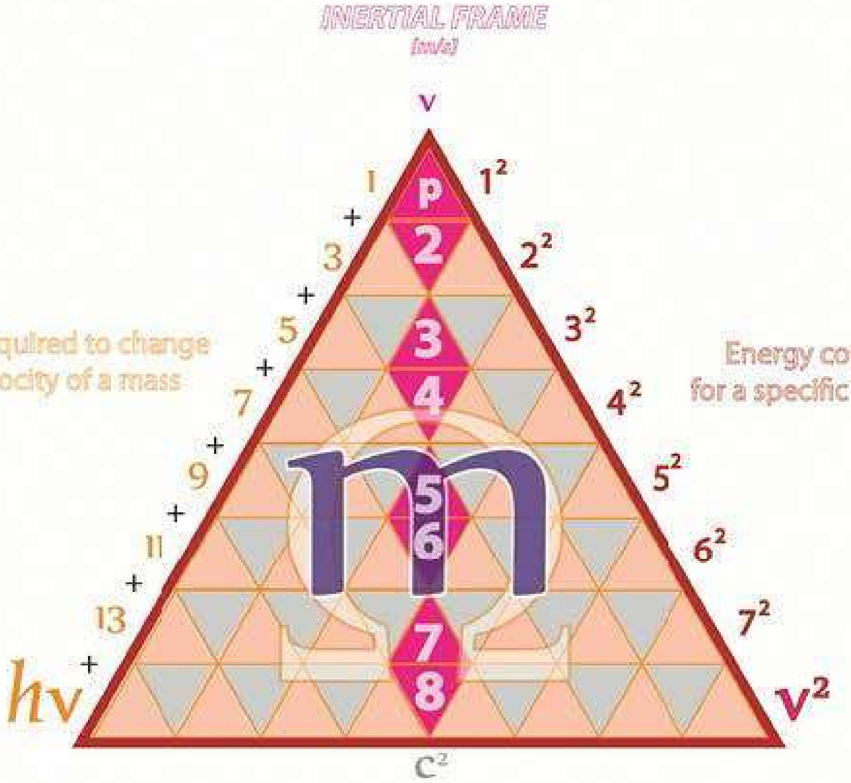

Energy content
for a specific velocity

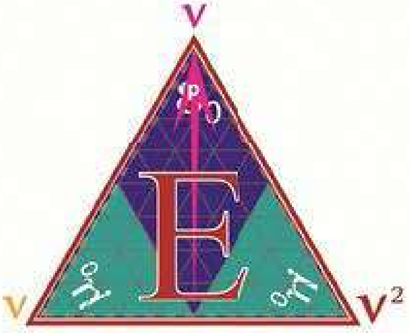

$$\frac{E}{c^2}.v$$

$$\mathbf{p} = \sum_{i=1}^{n} m_i \mathbf{v}_i = m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + m_3\mathbf{v}_3 + \cdots + m_n\mathbf{v}_n\,,$$

$$\frac{hv^2}{v}$$

The linear momentum of a EM field or system of particles is
the vector sum of the linear momenta in the KEM fields
of individual particles in any spatial co-ordinate system

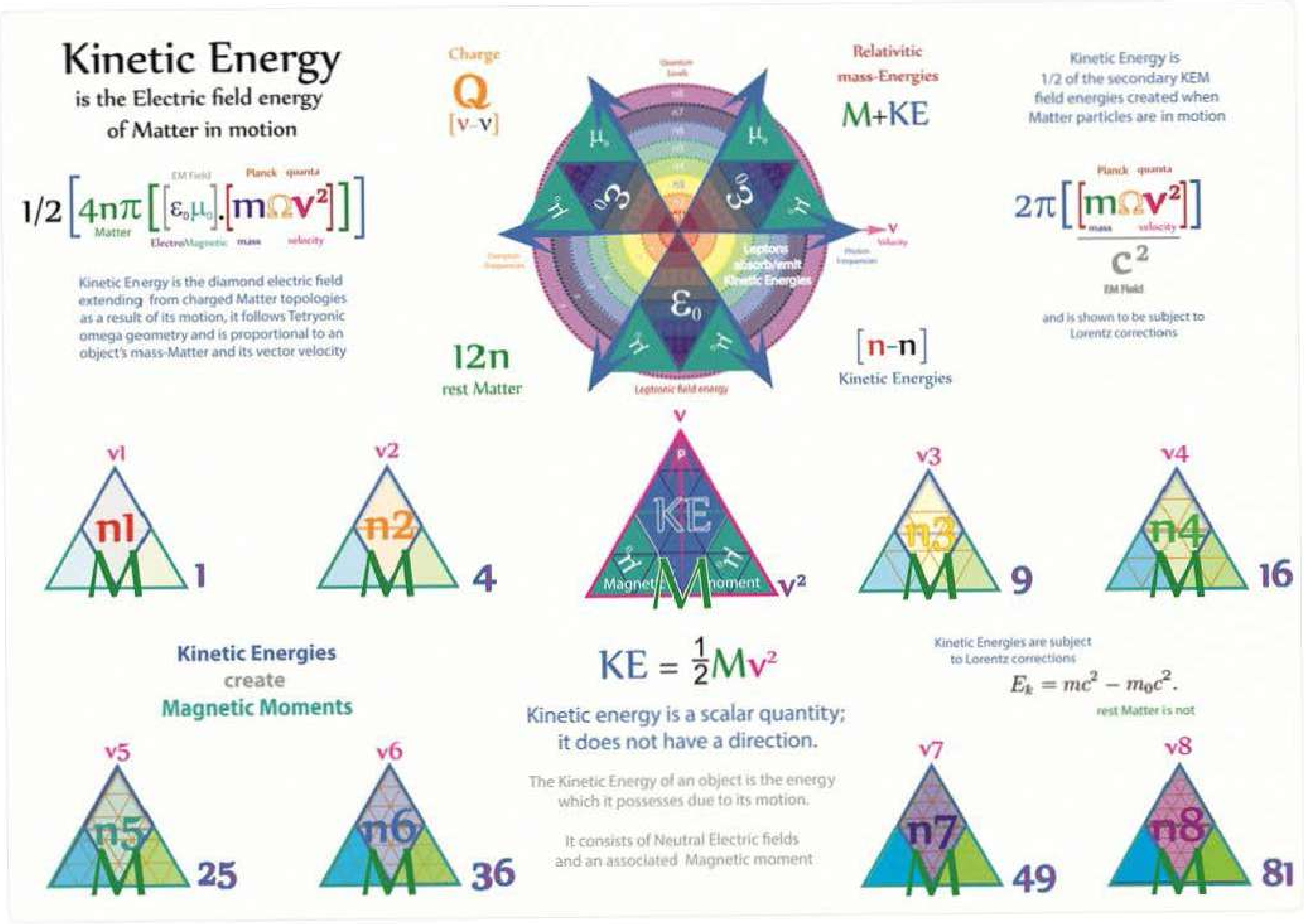

Kinetic Energy

is the Electric field energy of Matter in motion

$$\frac{1}{2}\left[\underset{\text{Matter}}{4n\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck quanta}}{[\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}]}\right]\right]$$

Kinetic Energy is the diamond electric field extending from charged Matter topologies as a result of its motion, it follows Tetryonic omega geometry and is proportional to an object's mass-Matter and its vector velocity

Charge

Q

[v-v]

12n

rest Matter

Relativitic mass-Energies

M+KE

[n-n]

Kinetic Energies

Kinetic Energy is 1/2 of the secondary KEM field energies created when Matter particles are in motion

$$\frac{2\pi\left[\overset{\text{Planck quanta}}{[\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}]}\right]}{\underset{\text{EM Field}}{c^2}}$$

and is shown to be subject to Lorentz corrections

Quantum Levels
μ_0
ω_0
ε_0
Leptons absorb/emit Kinetic Energies
Leptonic field energy
v
Velocity
Photon frequencies
v1
n1
M
1
v2
n2
M
4
v
KE
Magnetic Moment
M
v²
v3
n3
M
9
v4
n4
M
16
v5
n5
M
25
v6
n6
M
36
v7
n7
M
49
v8
n8
M
81

Kinetic Energies

create

Magnetic Moments

$$KE = \frac{1}{2}Mv^2$$

Kinetic energy is a scalar quantity; it does not have a direction.

The Kinetic Energy of an object is the energy which it possesses due to its motion.

It consists of Neutral Electric fields and an associated Magnetic moment

Kinetic Energies are subject to Lorentz corrections

$$E_k = mc^2 - m_0c^2.$$

rest Matter is not

Kinetic Energy and Magnetic moments

$$Mv^2 = KEM = p^2$$

Unlike charged fields KEM fields have neutralised Electric Fields

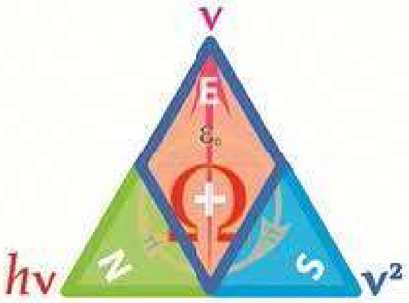
v
E
ε_0
Ω +
N
S
hv
v²

KEM field geometry of positive charged Matter in motion

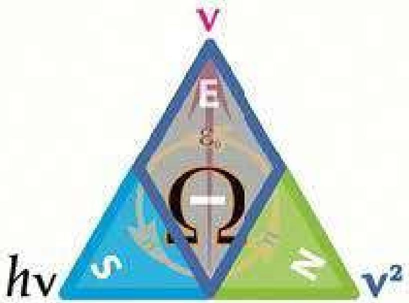
v
E
ε_0
Ω −
S
N
hv
v²

KEM field geometry of negative charged Matter in motion

$$KE = \frac{1}{2}Mv^2$$

Often noted as being two distinct EM energies [Kinetic Energy and Magnetic moments] are shown to be orthogonal aspects of the same KEM field of motion

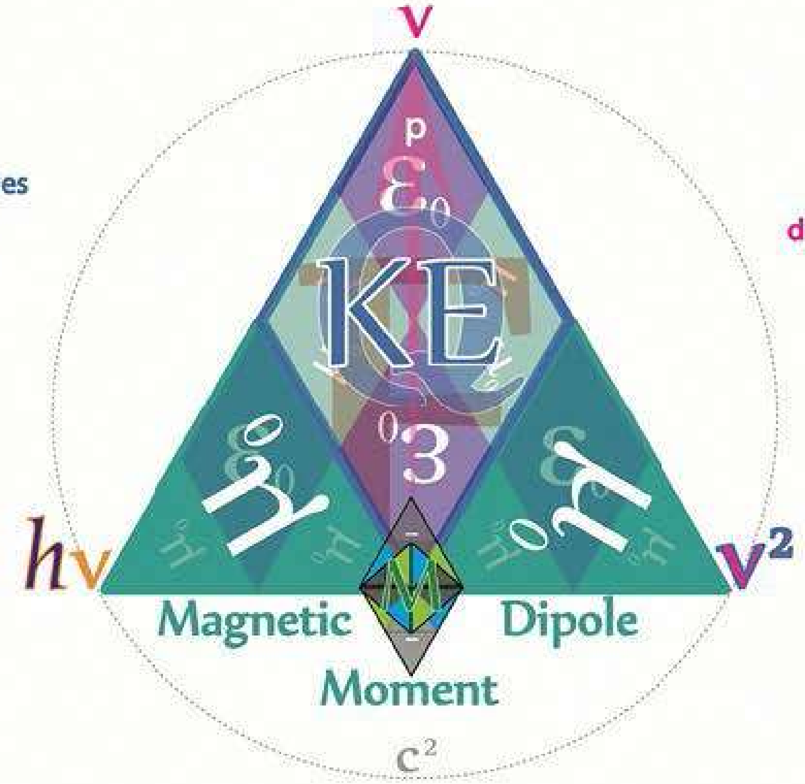
v
p
ε_0
KE
ε_0
ε_0
ε_0
hv
v²
Magnetic
Dipole
Moment
c²

As the velocity of a particle increases so does its Kinetic Energy and Magnetic moment creating Lorentz variable KEM fields

$$\mu_B = \frac{1}{2}Mv^2$$

$$\mu_B = \frac{1}{2}Mv^2$$

Copyright ABRAHAM [2008] - all rights reserved

Types of Momentum

Ω

quantised angular momenta

[m²/s]

The quantised 'angular momentum' of each Planck mass-energy geometry, gives rise to the two quantum Charges

p

Linear momentum

[kg·m/s]

The square root (v) of each Planck quanta's mass-energy geometry (v^2) is vector Linear Momentum

ℓ

Angular momentum

[kg·m²/s]

The orbital angular momentum of electrons in atoms associated with a given quantum state

... [n-1] [n] [n+1] ...

There are 3 forms of momenta in physics

equilateral charge geometry

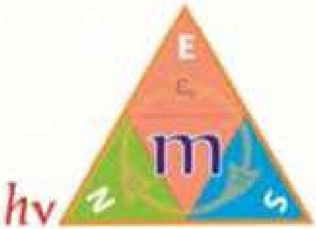

h

[kg·m²/s]

$m\Omega$

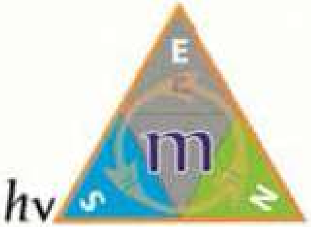

vector momentum

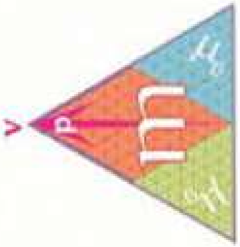

mv

[kg·m/s]

p

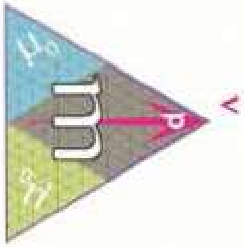

vector rotation about a point

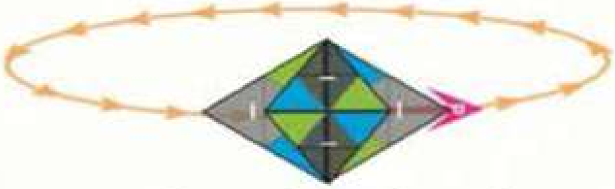

$$L^2 = L_x^2 + L_y^2 + L_z^2$$

Kinetic Energy vs Momentum

An important difference is that Kinetic energy is a scalar quantity - it has no direction in space
momentum is a vector quantity - it has a direction in space, momenta combine like forces do.

$$p = Mv$$

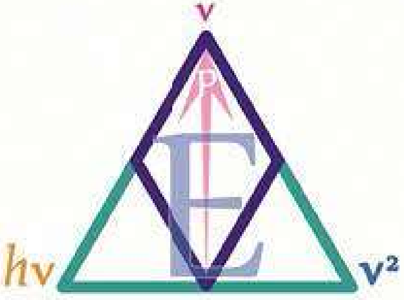

Energy momenta

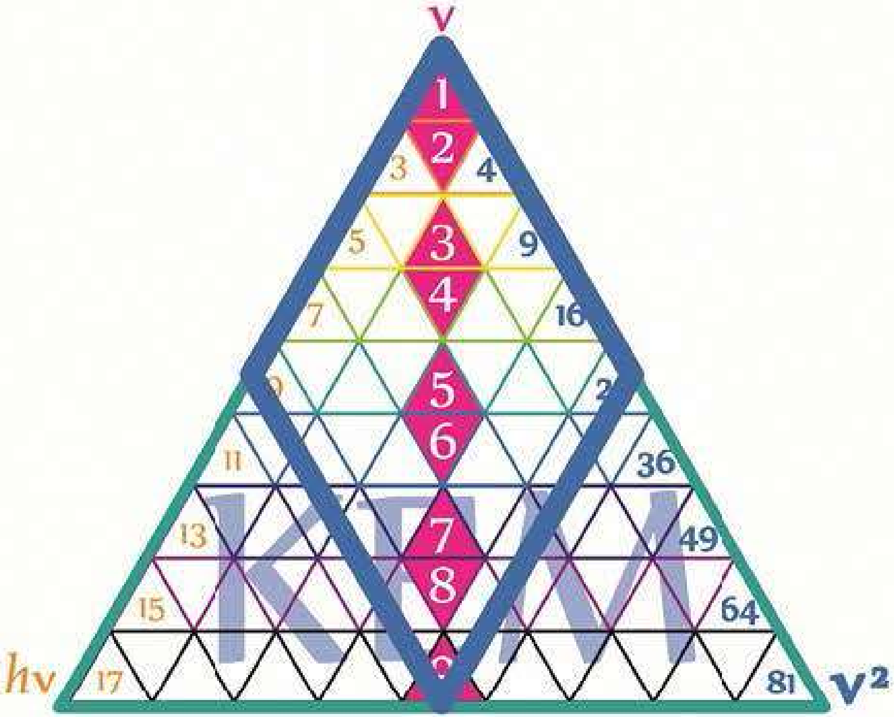

$$KE = 1/2[Mv^2]$$

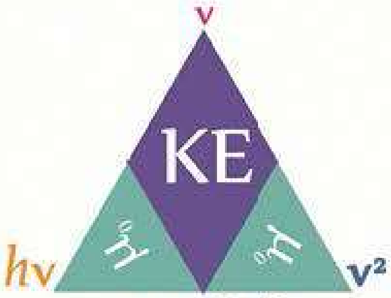

KEM field

$$p = \sqrt{E}$$

$$KEM = Mv^2$$

In Tetryonic geometry, the square root maps the linear momenta [mv] of a field to its ENERGY

$$p^2/2m = KE$$

In Tetryonic geometry, E-field geometry maps the kinetic energy [1/2 mv] of a field to its ENERGY

Linear momentum has a different geometric ENERGY relationship to that of Electric fields

mass-Energy-momentum

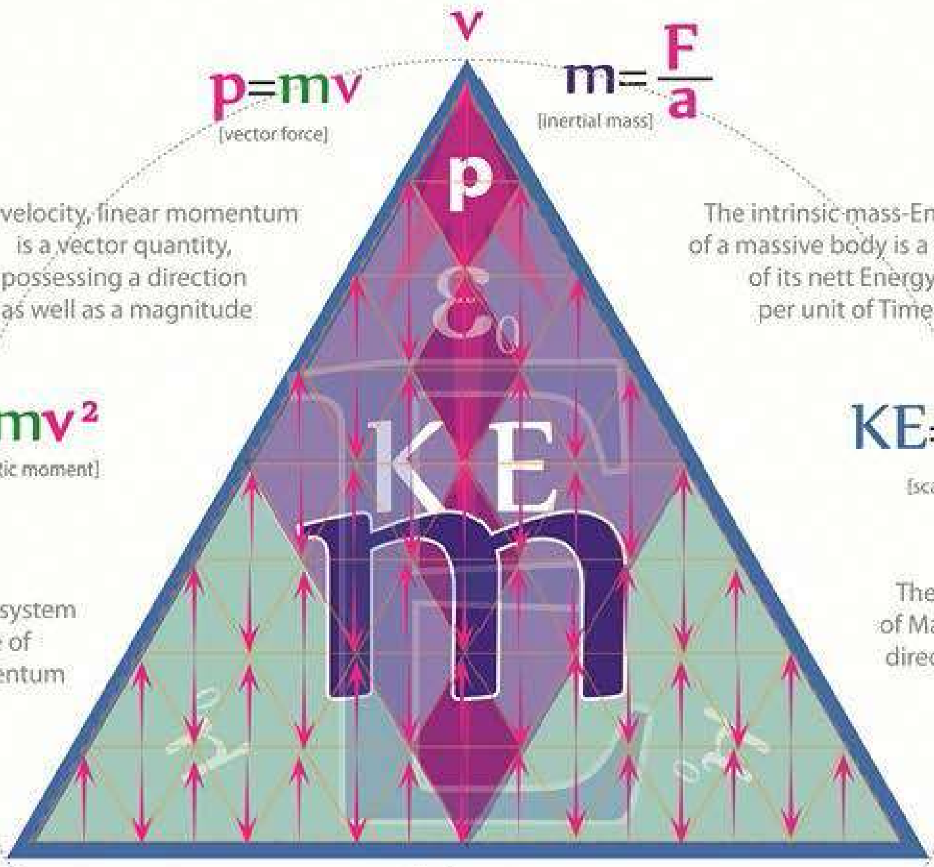

$p=mv$

[vector force]

$m=\frac{F}{a}$

[inertial mass]

Like velocity, linear momentum
is a vector quantity,
possessing a direction
as well as a magnitude

The intrinsic mass-Energy
of a massive body is a result
of its nett Energy
per unit of Time

$KEM=mv^2$

[Kinetic Energy & magnetic moment]

$KE=\frac{1}{2}mv^2$

[scalar E-field]

The Energy of a system
is the square of
its linear momentum

The Kinetic energy
of Matter in motion is
directly related to its
velocity

$E=mv^2$

[scalar KEM field energies
of Matter in motion]

$E=p^2$

[Energy momenta]

Planck quanta

$[m\Omega v]$

mass velocity

acceleration results in
changes to momenta

$F=ma$

Newton's
Second law of Motion

Linear momentum is the square root
of the Energy of a system.

It is the nett linear Force
resulting from Matter in motion
and was used by Newton
as the foundation for
his Laws of motion

Linear momentum is also a conserved quantity, meaning that if a closed system is not affected by external forces, its total linear momentum cannot change. Although originally expressed in Newton's second law, the conservation of linear momentum also holds in special relativity and, with appropriate definitions, a (generalized) linear momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory, and general relativity. In relativistic mechanics, non-relativistic linear momentum is further multiplied by the Lorentz factor

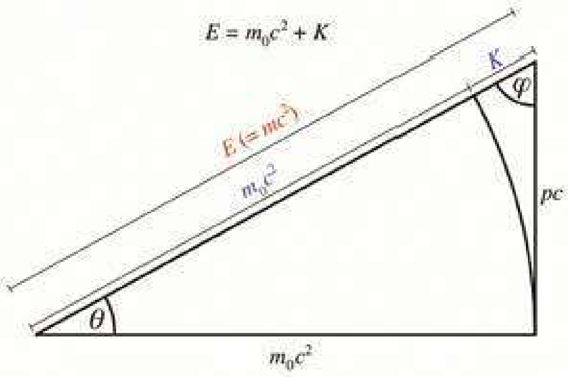

$$E^2 = p^2c^2 + m_0^2c^4$$

Often generalised as having the geometry of
right angled triangles the mass-energy momenta
relationship is fully revealed using
Equilateral geometry

E^2

**is the energy content
of a superpositioned
EM field**

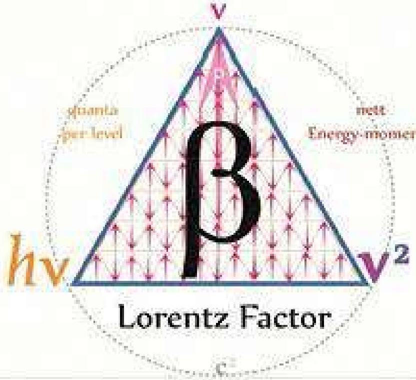

Tetryonics 14.10 - mass-Energy momentum

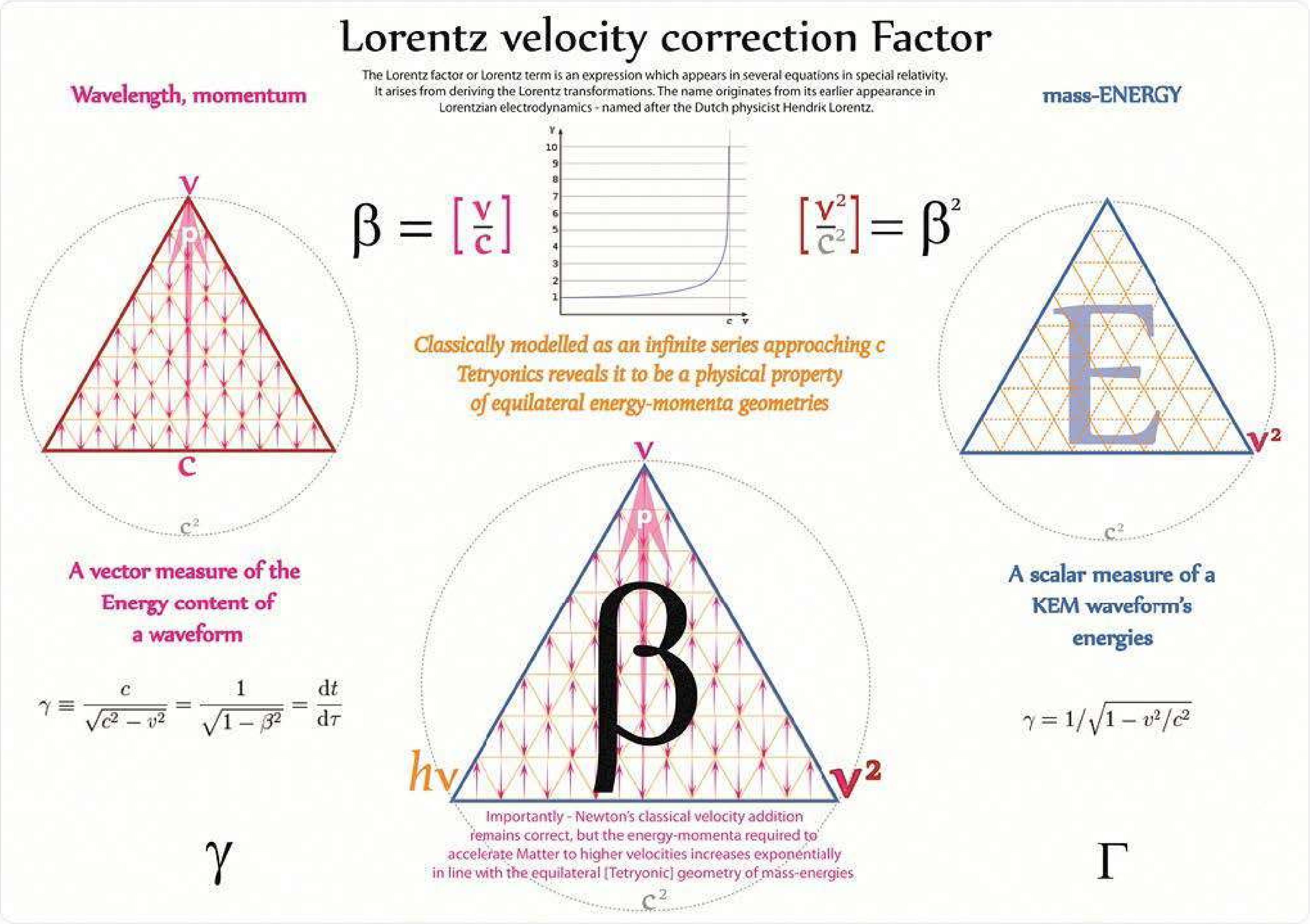

Lorentz velocity correction Factor

The Lorentz factor or Lorentz term is an expression which appears in several equations in special relativity. It arises from deriving the Lorentz transformations. The name originates from its earlier appearance in Lorentzian electrodynamics - named after the Dutch physicist Hendrik Lorentz.

Wavelength, momentum

mass-ENERGY

$$\beta = \left[\frac{v}{c}\right]$$

$$\left[\frac{v^2}{c^2}\right] = \beta^2$$

Classically modelled as an infinite series approaching c
Tetryonics reveals it to be a physical property
of equilateral energy-momenta geometries

A vector measure of the Energy content of a waveform

$$\gamma \equiv \frac{c}{\sqrt{c^2 - v^2}} = \frac{1}{\sqrt{1-\beta^2}} = \frac{dt}{d\tau}$$

A scalar measure of a KEM waveform's energies

$$\gamma = 1/\sqrt{1 - v^2/c^2}$$

Importantly - Newton's classical velocity addition remains correct, but the energy-momenta required to accelerate Matter to higher velocities increases exponentially in line with the equilateral [Tetryonic] geometry of mass-energies

γ

Γ

Tetryonics 14.11 - Lorentz velocity correction Factor

Unified Energy momenta geometry

$$m = n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM field · Planck quanta · mass · Electromagnetic · mass · velocity

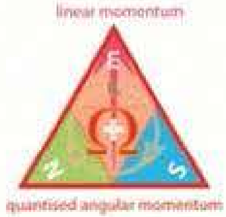
linear momentum
E
Ω
N
S
quantised angular momentum

$$p = n\pi\left[m\Omega v\right]$$

Planck quanta · energy · mass · velocity

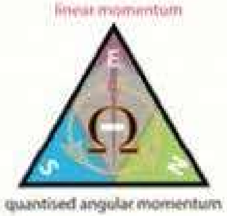
linear momentum
E
Ω
S
N
quantised angular momentum

$$E = n\pi\left[\left[m\Omega v\right]^2\right]$$

Planck quanta · mass · velocity

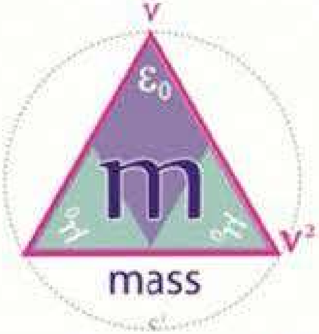
v
ε0
m
μ0
v²
mass
c²

$$mv = p$$

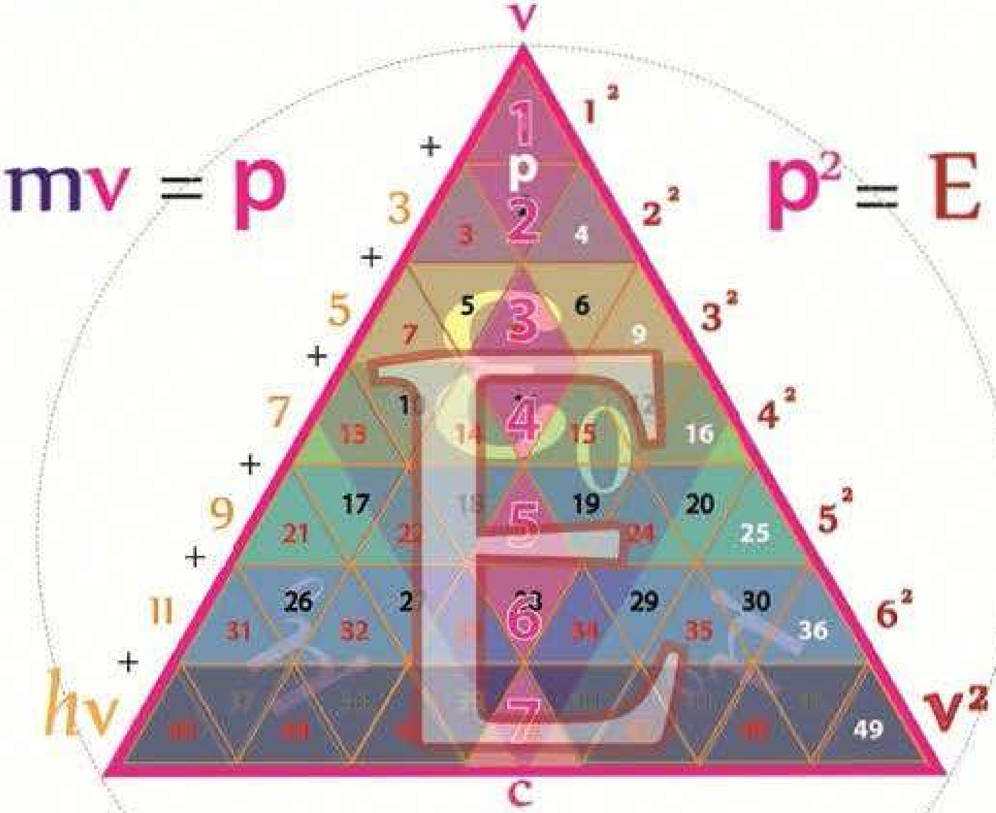
v
1²
2²
3²
4²
5²
6²
v²
1
p
2
3
4
5
6
7
+
3
5
7
9
11
hν
E
0
c
1
2
3
4
5
6
7
9
10
12
13
14
15
16
17
18
19
20
21
22
24
25
26
27
28
29
30
31
32
34
35
36
49

$$p^2 = E$$

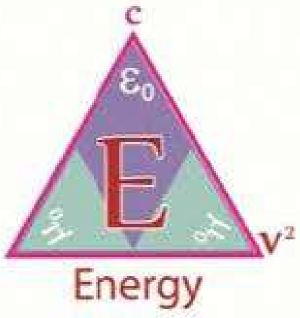
c
ε0
E
μ0
v²
Energy

$$\beta = \left[\frac{v}{c}\right]$$

Linear

Lorentz factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Scalar

Lorentz factor

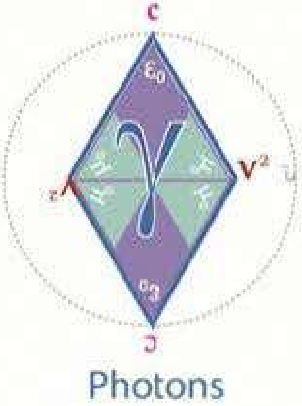
c
ε0
γ
μ0
v²
ε0
c
Photons

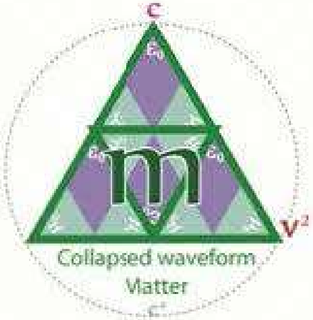
c
ε0
m
v²
Collapsed waveform
Matter
c²

Tetryonic's equilateral geometries model
EM mass-ENERGY momenta
in all its forms

c^2

$$E_\gamma = 2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM field · Planck quanta · Photons · Electromagnetic · mass · velocity

$$m = 4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM field · Planck quanta · mass · Electromagnetic · mass · velocity

Tetryonics 15.01 - Unified Energy-momenta geometry

Squared Energy distributions

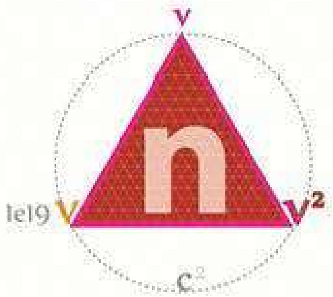
v
n
Ie19 V
v²
c²

Bosons

Transverse EM mass-Energies

Odd number sequence

Number of quanta per geometry

n = (2n-1)

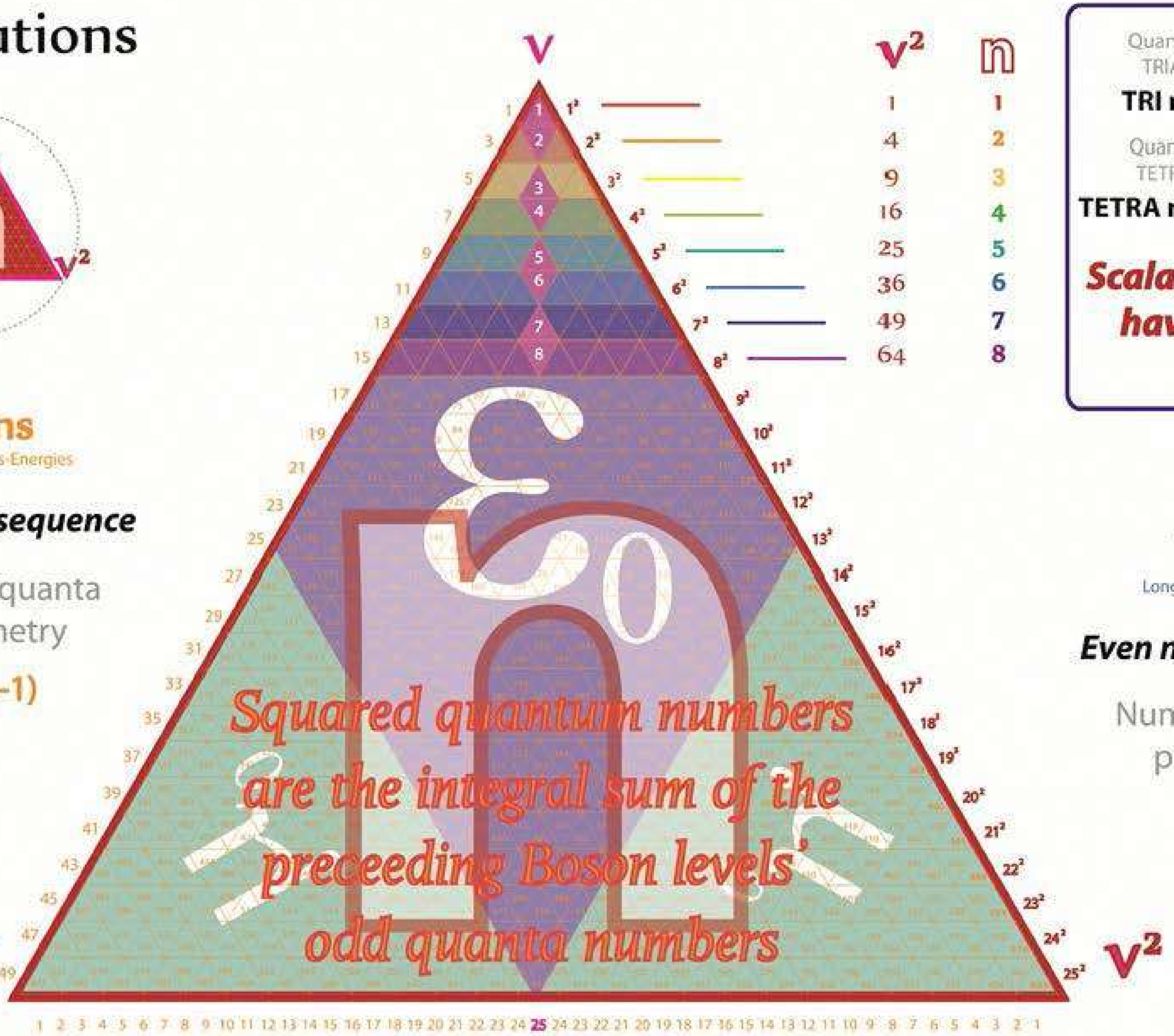
v
v²
n
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
1² 2² 3² 4² 5² 6² 7² 8² 9² 10² 11² 12² 13² 14² 15² 16² 17² 18² 19² 20² 21² 22² 23² 24² 25²
1 4 9 16 25 36 49 64
1 2 3 4 5 6 7 8
ε0
n
hv
v²
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Squared quantum numbers are the integral sum of the preceeding Boson levels' odd quanta numbers

[n-1] Transverse Quanta Distributions [n-1]

EM mass-energy

Quantum numbers are *not* TRIANGULAR NUMBERS

TRI n = (n/2) x (n+1)

Quantum numbers are *not* TETRAHEDRAL NUMBERS

TETRA n = (n/6) (n+1) (n+2)

Scalar energy levels have SQUARED quanta

Photons

Longitudinal EM mass-Energies

Even number sequence

Number of quanta per geometry

n = (2n)

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic Energy and Charge relationships

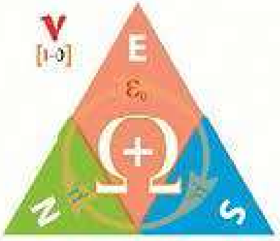

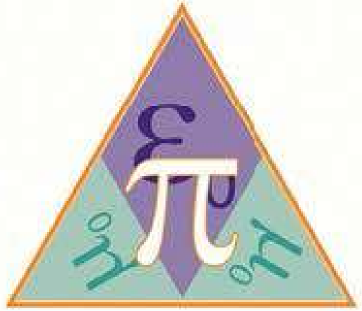

Quantised angular momenta is the geometric source of all mass-Energy-Matter relationships

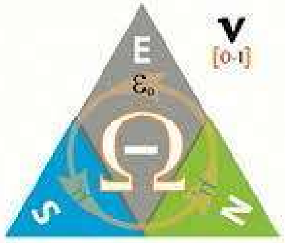

Scalar energies (are 'squared' numbers)

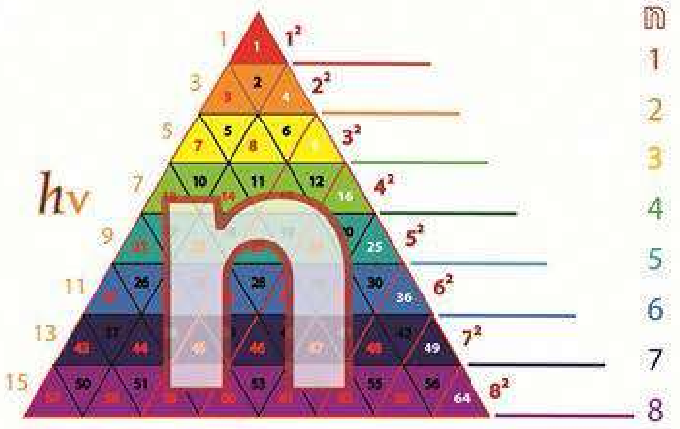
hv
1
3
5
7
9
11
13
15
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
m
1
2
3
4
5
6
7
8
n

The Energy density of a particle's charged fascia geometry determines its mass

m Charged mass-energy geometry & Matter topology determines a particle's physical characteristics (Type, Family, generation, mass etc) **M**

Charge

-12

electron

12
[0-12]

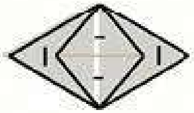

1.2 e20

mass-Matter

e-

8.851486 e-31 KG

Opposite nett elementary charge geometries

1875x
(differential)

differing Matter topologies

+12

Proton

12
[24-12]

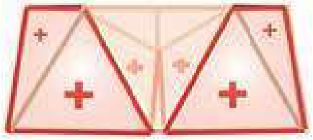

2.25 e23

1.65965 e-27 KG

P+

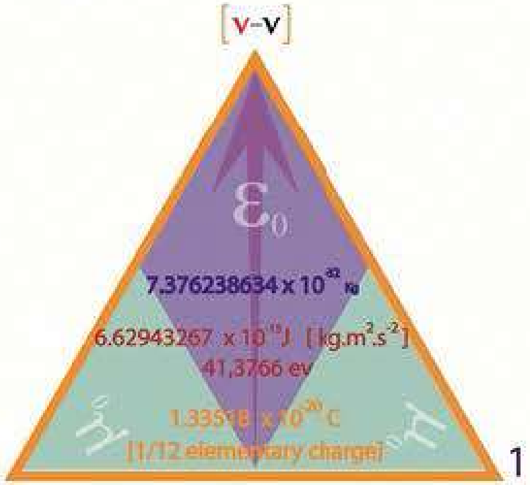
[v-v]
ε₀
7.376238634 x 10^{-51} Kg
6.62943267 x 10^{-34}J [kg.m^2.s^{-2}]
41.3766 ev
1.33518 x 10^{-20}C
[1/12 elementary charge]
μ₀
μ₀
1

Planck mass-energy

Copyright ABRAHAM [2008] - all rights reserved

Planck's energy momenta quantum

$\frac{E}{c^2}$

m

planck quanta
$[m\Omega v^2]$
mass velocity

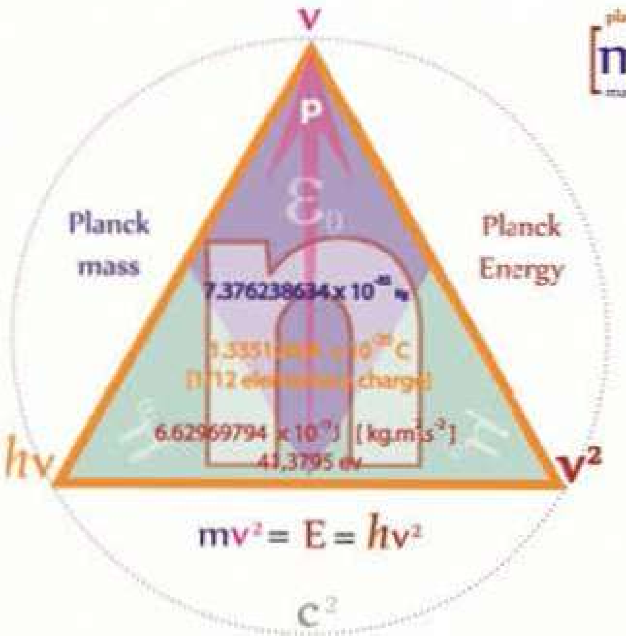
v
p
ε_Ω
Planck
mass
Planck
Energy
7.376238634 x 10[illegible] kg
1.3351[illegible] x 10[illegible] C
[1/12 elemental charge]
6.62969794 x 10[illegible] J [kg.m²s⁻²]
41,3795 ev
hv
v²
c²

E

$mv^2 = E = hv^2$

	Planck mass	**m**	m - mass quantum
q	Planck Charge quantised angular momentum / spatial viz velocity system	Ω	v - quanta of Charge $\frac{\text{elemental charge}}{12}$
	Planck Energy	**E**	E - Energy quantum

v^2 Compton Frequency	$\frac{mc^2}{h}$	1×10^{19} quanta per second
λ de Broglie wavelength	$\frac{h}{mc}$	$2.99792458 \times 10^{-11}$ meters
p Momentum	mc	2.21134×10^{-23} Ns

Planck's constant is the equilateral geometry of rest mass-energy momenta in standing wave Matter topologies

Currently the Atomic masses are calculated using 1/12 of a Molar Carbon12 atom's mass as the reference mass.

Tq [Carbon] - [252-252] **C** 12

Tm [Carbon] - 270,072

[270,072] / 12
22,506

If we calculated for the molar mass of C12 in 12 grams we would get

(.012/N/270072) $7.378205107 \times 10^{-32}$ grams

this is in error by 1/2 of the mass of an electron

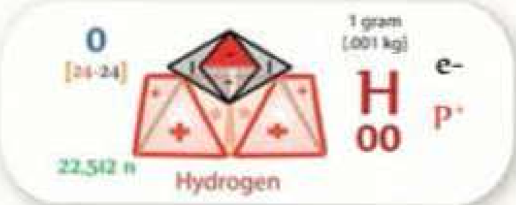

Tetryonic mass [Hydrogen] - 2.2512 e23 v

Using Tetryonic geometry we can solve for an exact Compton frequency of any mass-Matter topology and account exactly for all quantum energy contributions to the rest Matter of all electrons and Baryons in any element thus determining exact molar masses exclusive of blackbody radiation, kinetic energies and energies of measurement and avoiding 'weighted' masses

Defining [N] as the number of rest mass Hydrogen atoms in 1gram exactly and makes Avagadro's number the inverse of this number

$$\text{Molar mass} = \frac{H_1\ \text{Atomic mass}}{\text{Avagadro's No.}}$$

$$\text{mass } H_{mol}/Av = \frac{.001}{6.022141579 \times 10^{23}}$$

rest mass Hydrogen = $1.660538841 \times 10^{-27}$ g/mol

$$H_{mole}/m[H] = \frac{1.660738412 \times 10^{-27}}{22{,}512}$$

n Planck mass = $7.376238634 \times 10^{-32}$ Kg

This is an exact Atomic rest mass as opposed to the current weighted molar Atomic mass estimates (which is incorrect by **1/2** the mass of an electron)

Planck's constant is the 'quantum of action' and also contributes to the mass-energies of radiant EM wave geometries

Tetryonics 15.04 - Planck's energy momenta quantum

Copyright ABRAHAM [2008] - all rights reserved

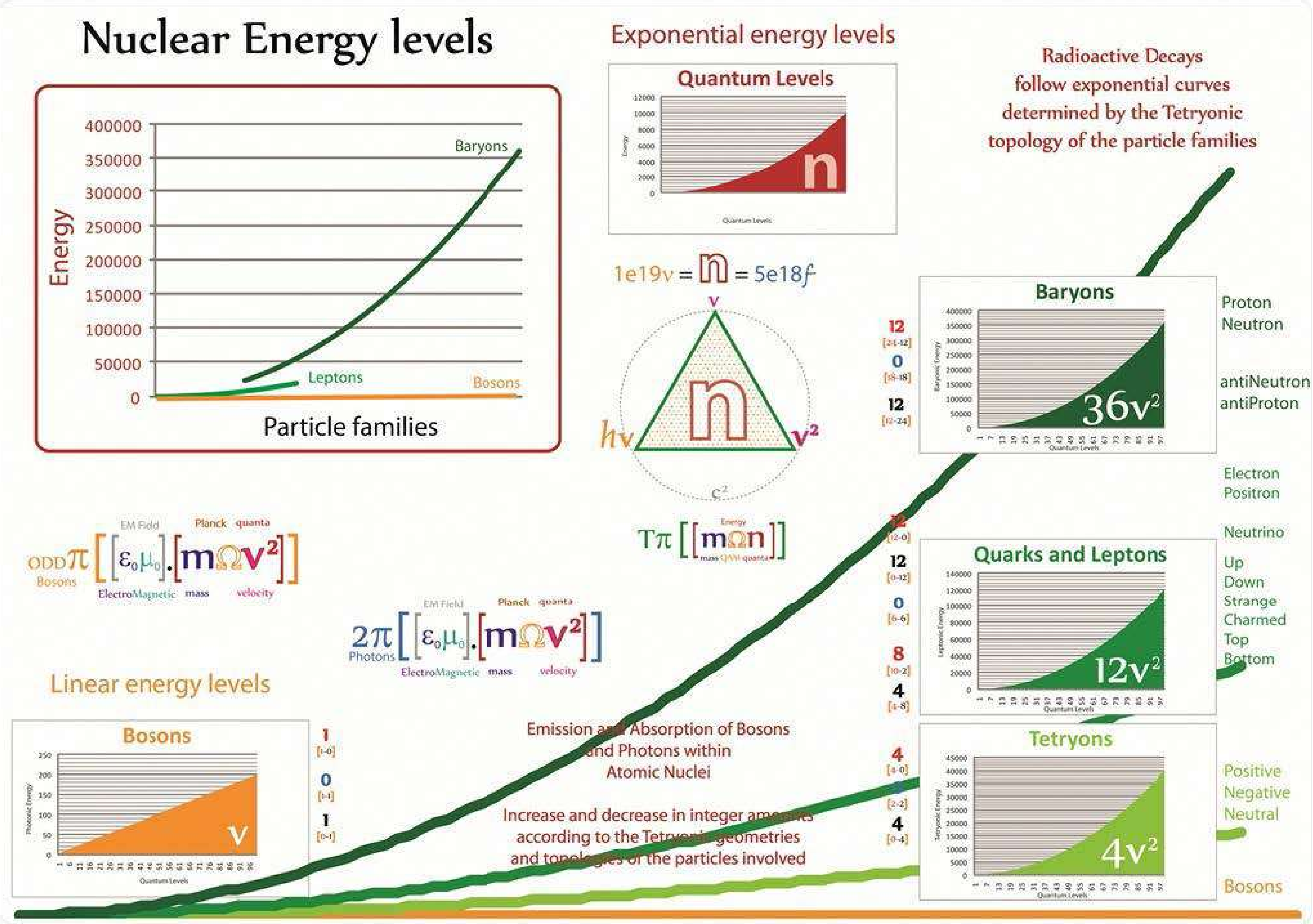

Nuclear Energy levels

Exponential energy levels

Radioactive Decays follow exponential curves determined by the Tetryonic topology of the particle families

$1e19\nu = n = 5e18f$

$T\pi\,[[m\Omega n]]$

Proton
Neutron

antiNeutron
antiProton

Electron
Positron

Neutrino

Up
Down
Strange
Charmed
Top
Bottom

$\text{ODD}\,\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$

$2\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Linear energy levels

Bosons
250
200
150
100
50
0
Photonic Energy
v
Quantum Levels
1
[1-0]
0
[1-1]
1
[0-1]

Emission and Absorption of Bosons and Photons within Atomic Nuclei

Increase and decrease in integer amounts according to the Tetryonic geometries and topologies of the particles involved

4
[4-0]
[2-2]
4
[0-4]
Tetryons
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
Tetryonic Energy
4v²
Quantum Levels

Positive
Negative
Neutral

Bosons

Tetryonics 15.05 - Nuclear energy levels

Copyright ABRAHAM [2008] - all rights reserved

Energy momenta geometry

m

All mass-energies have equilateral geometries

E

mass-Energy

$kg\frac{m^2}{s^2}$

Ω

Quantised angular momenta facilitates EM interactions

h

Planck's Constant

$kg\frac{m^2}{s}$

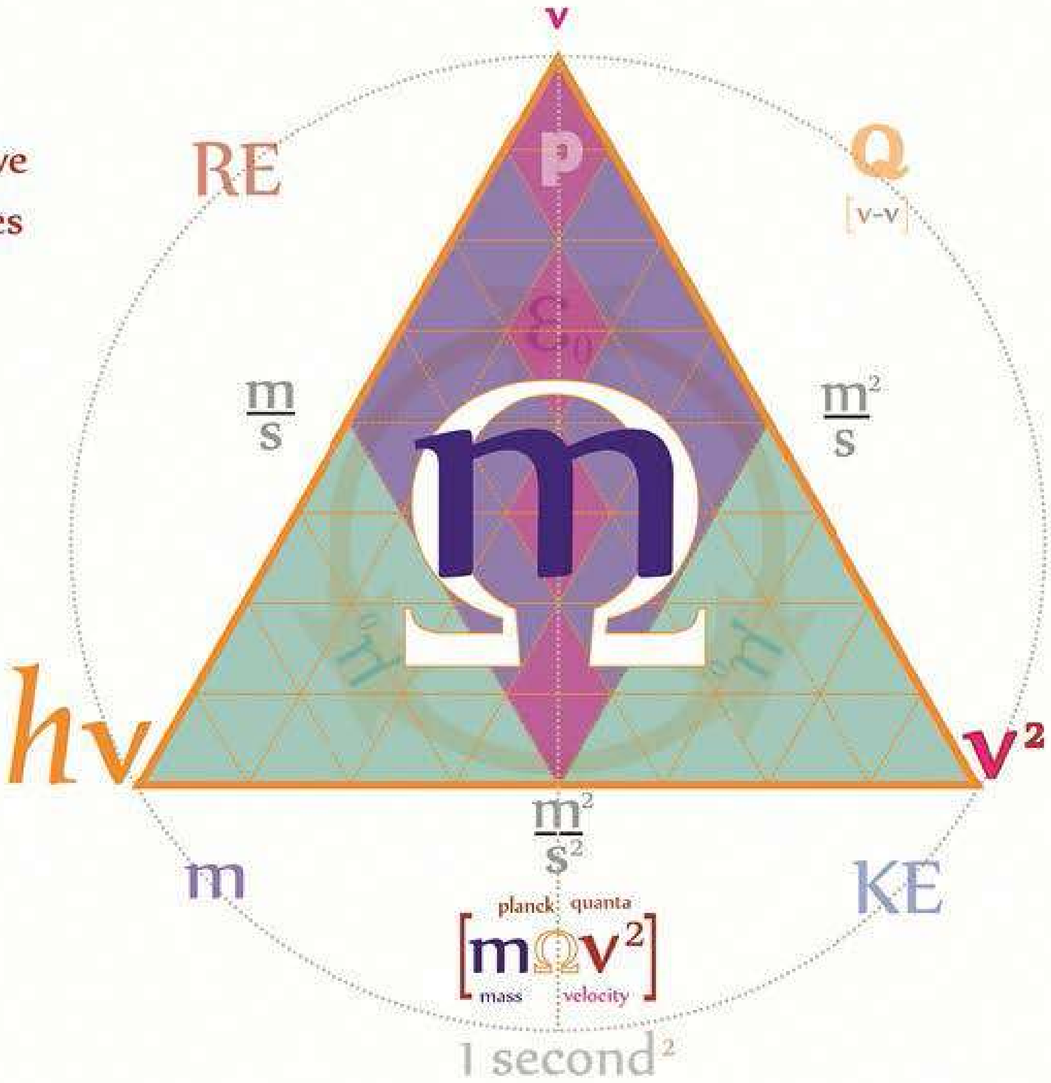
v
p
RE
Q
[v-v]
ε0
$\frac{m}{s}$
$\frac{m^2}{s}$
m
Ω
μ0
μ0
hv
v²
$\frac{m^2}{s^2}$
m
KE
planck
quanta
[mΩv²]
mass
velocity
1 second²

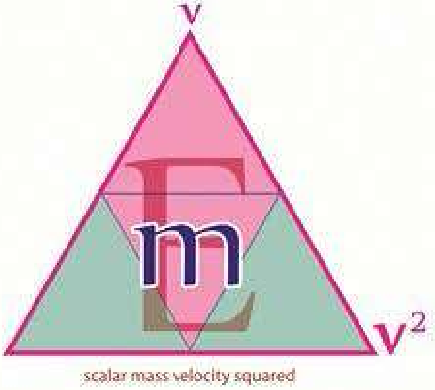

scalar mass velocity squared

$$mv^2 = E = hv^2$$

v
KE
E
hv
μ0
μ0
v²

quantised angular momenta per second

Tetryonics 15.06 - Energy momenta geometry

m

mass

E

Energy

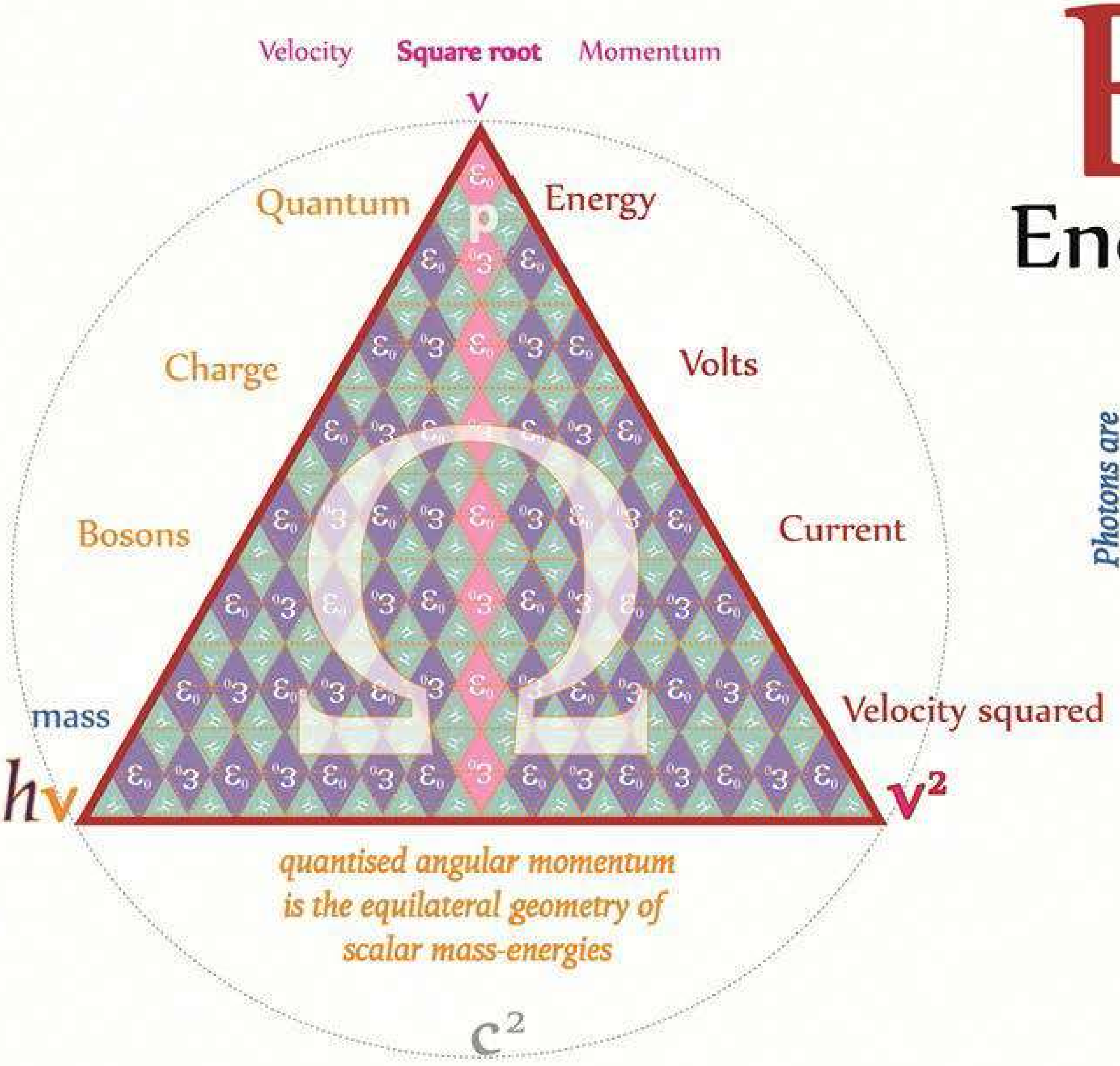

Velocity **Square root** Momentum

v

Quantum

Energy

Charge

Volts

Bosons

Current

mass

Velocity squared

hv

v²

quantised angular momentum is the equilateral geometry of scalar mass-energies

c²

Bosons are transverse EM masses

Photons are longitudinal EM masses

mass is a measure of Energy per spatial co-ordinate system

mass-Energy momenta

Electromagnetic energy is a scalar measurement of mass

Tetryonics 15.07 - EM mass-energy relationships

Copyright ABRAHAM [2008] - all rights reserved

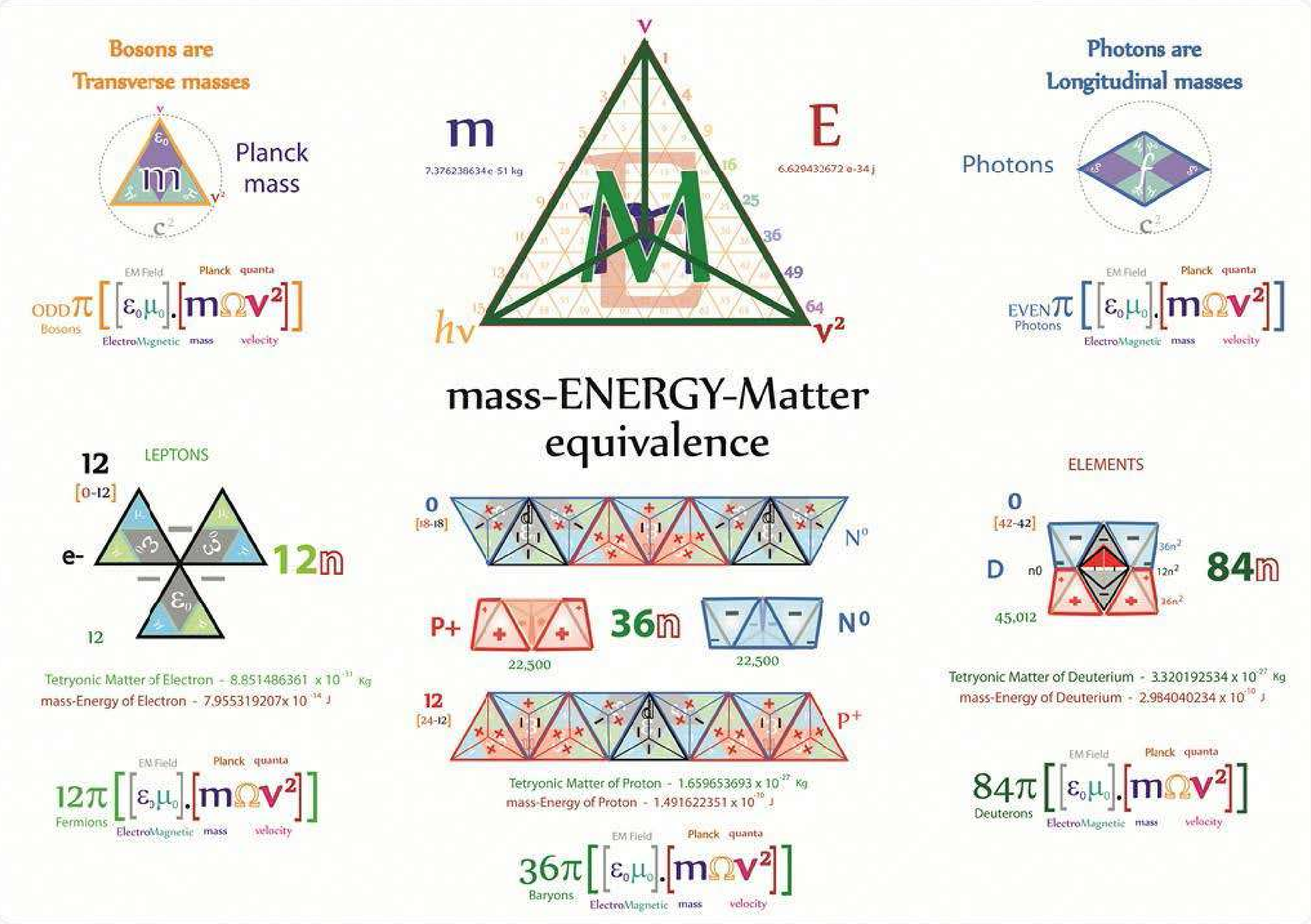

Bosons are
Transverse masses
Planck mass
Photons are
Longitudinal masses
Photons
m
7.376238634e-51 kg
E
6.629432672 e-34 j
v
hv
v²
LEPTONS
12
[0-12]
e-
12n
12
0
[18-18]
N°
P+
36n
N0
22,500
22,500
12
[24-12]
P+
ELEMENTS
0
[42-42]
D
n0
84n
45,012
36n²
12n²
36n²
EM Field
Planck
quanta
ElectroMagnetic
mass
velocity

$$\text{ODD}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons

mass-ENERGY-Matter equivalence

Tetryonic Matter of Electron - 8.851486361 x 10^{-31} Kg

mass-Energy of Electron - 7.955319207x 10^{-14} J

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Fermions

Tetryonic Matter of Proton - 1.659653693 x 10^{-27} Kg

mass-Energy of Proton - 1.491622351 x 10^{-10} J

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons

Tetryonic Matter of Deuterium - 3.320192534 x 10^{-27} Kg

mass-Energy of Deuterium - 2.984040234 x 10^{-10} J

$$84\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Deuterons

Tetryonics 15.08 - mass-Energy-Matter equivalence

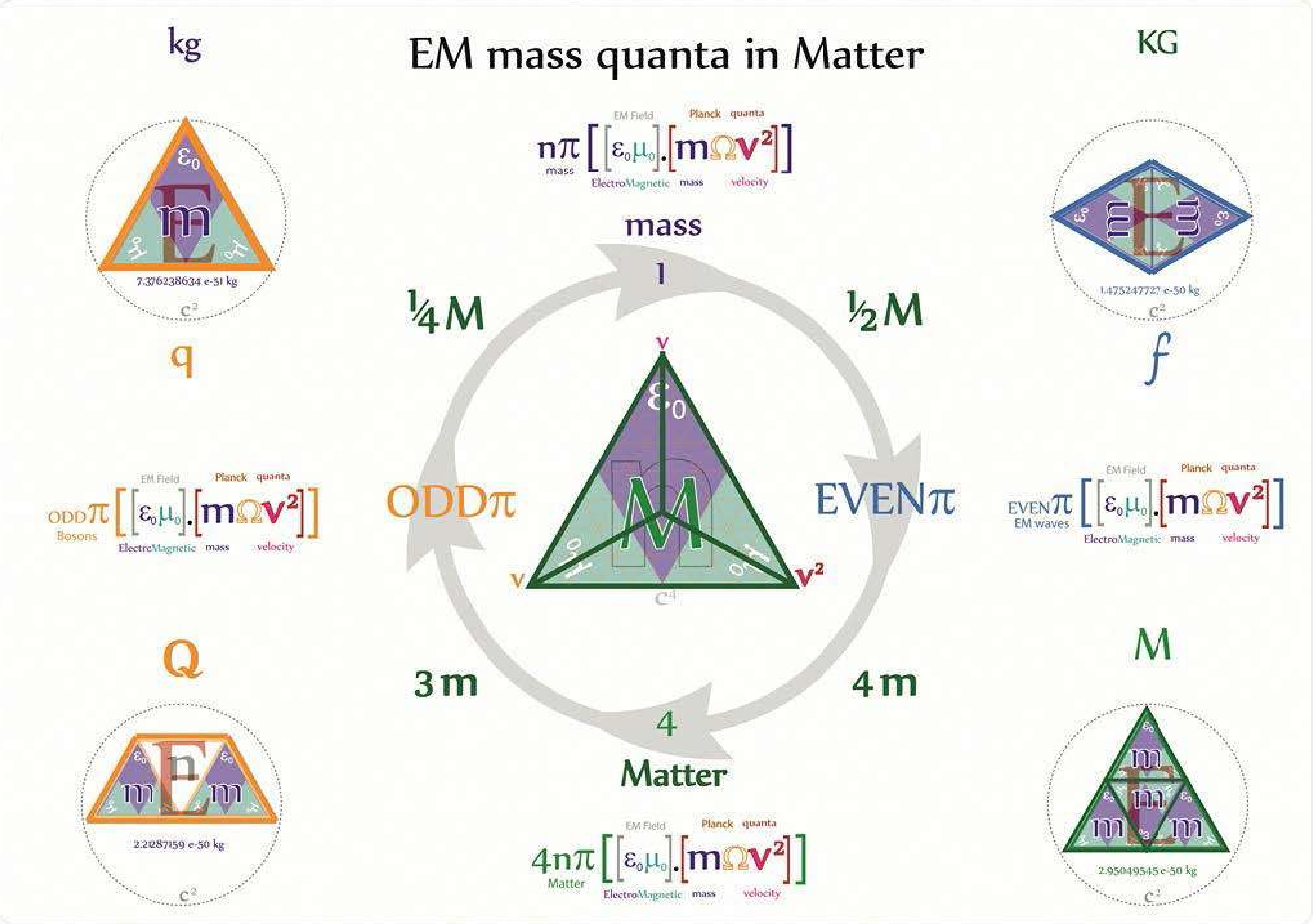
EM mass quanta in Matter
kg
KG
EM Field
Planck
quanta
nπ [[ε₀μ₀].[mΩv²]]
mass
ElectroMagnetic
mass
velocity
mass
1
¼ M
½ M
7.376238634 e-51 kg
c²
q
1.475247727 e-50 kg
c²
f
ODDπ [[ε₀μ₀].[mΩv²]]
Bosons
ODDπ
EVENπ
EVENπ [[ε₀μ₀].[mΩv²]]
EM waves
v
ε₀
M
v
v²
c⁴
Q
M
3 m
4 m
4
Matter
2.2287159 e-50 kg
c²
2.95049545 e-50 kg
c²
4nπ [[ε₀μ₀].[mΩv²]]
Matter

Tetryonics 15.09 - EM mass quanta in Matter

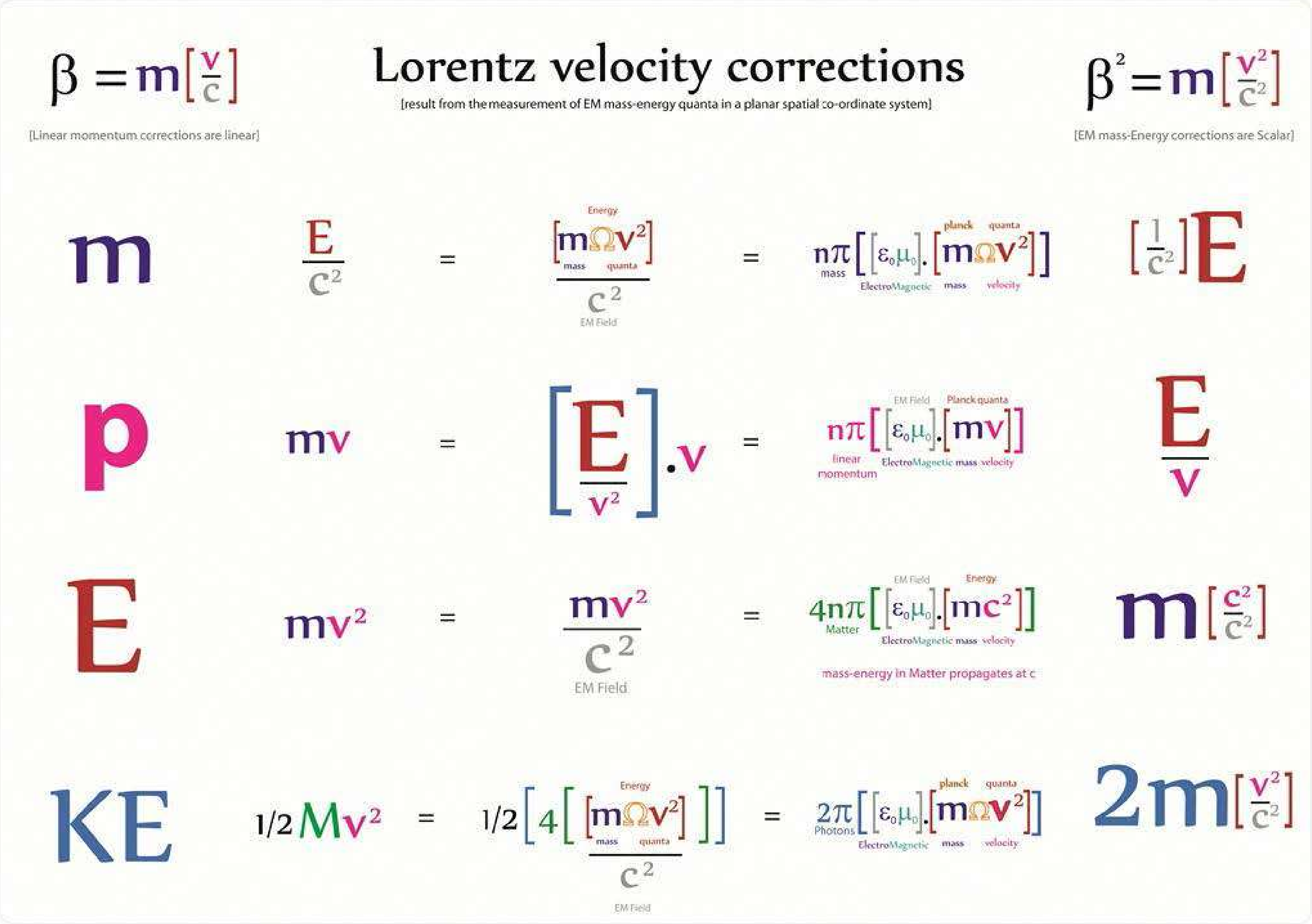

$$\beta = m\left[\frac{v}{c}\right]$$

[Linear momentum corrections are linear]

Lorentz velocity corrections

[result from the measurement of EM mass-energy quanta in a planar spatial co-ordinate system]

$$\beta^2 = m\left[\frac{v^2}{c^2}\right]$$

[EM mass-Energy corrections are Scalar]

$$m \qquad \frac{E}{c^2} = \frac{\overset{\text{Energy}}{\left[\underset{\text{mass}}{m}\,\Omega\,\underset{\text{quanta}}{v^2}\right]}}{\underset{\text{EM Field}}{c^2}} = \underset{\text{mass}}{n\pi}\left[\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}.\left[\overset{\text{planck}}{\underset{\text{mass}}{m}\,\Omega}\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right] \qquad \left[\frac{1}{c^2}\right]E$$

$$p \qquad mv = \left[\frac{E}{v^2}\right].v = \underset{\substack{\text{linear}\\ \text{momentum}}}{n\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck quanta}}{\left[\underset{\text{mass}}{m}\,\underset{\text{velocity}}{v}\right]}\right] \qquad \frac{E}{v}$$

$$E \qquad mv^2 = \frac{mv^2}{\underset{\text{EM Field}}{c^2}} = \underset{\text{Matter}}{4n\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Energy}}{\left[\underset{\text{mass}}{m}\,\underset{\text{velocity}}{c^2}\right]}\right] \qquad m\left[\frac{c^2}{c^2}\right]$$

mass-energy in Matter propagates at c

$$KE \qquad 1/2\,Mv^2 = 1/2\left[4\left[\frac{\overset{\text{Energy}}{\left[\underset{\text{mass}}{m}\,\Omega\,\underset{\text{quanta}}{v^2}\right]}}{\underset{\text{EM Field}}{c^2}}\right]\right] = \underset{\text{Photons}}{2\pi}\left[\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}.\left[\overset{\text{planck}}{\underset{\text{mass}}{m}\,\Omega}\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right] \qquad 2m\left[\frac{v^2}{c^2}\right]$$

Tetryonics 15.10 - Lorentz velocity corrections

The geometry of Constants

Quantised angular momentum is the source of all physical Constants

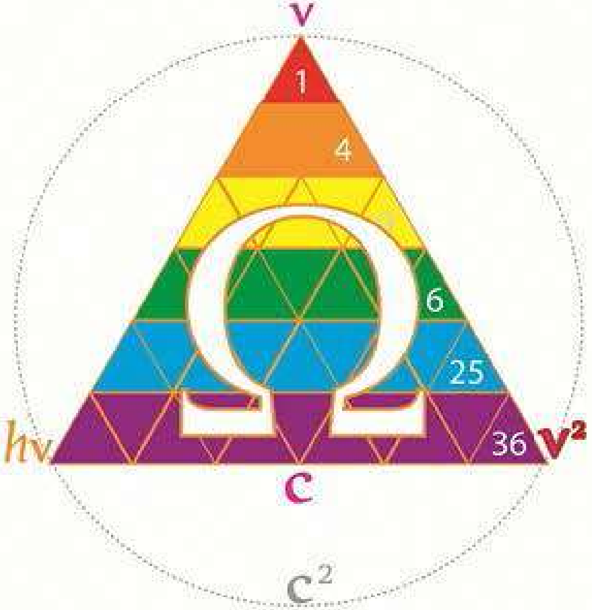
v
1
4
6
25
36
Ω
hv
v²
c
c²

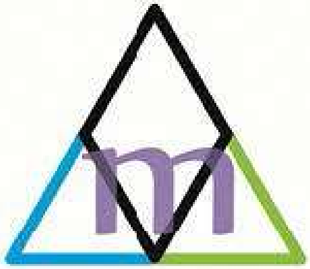

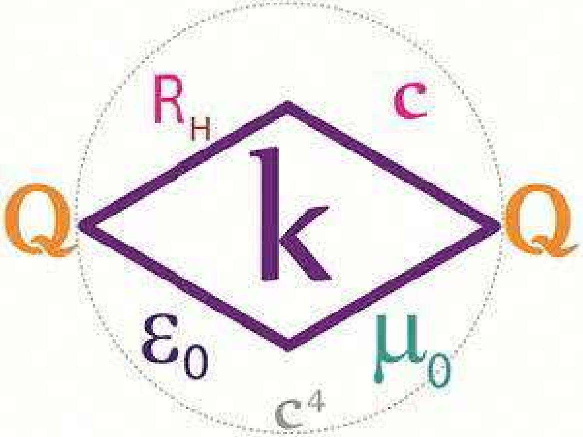
R_H
c
Q
k
Q
ε₀
μ₀
c⁴

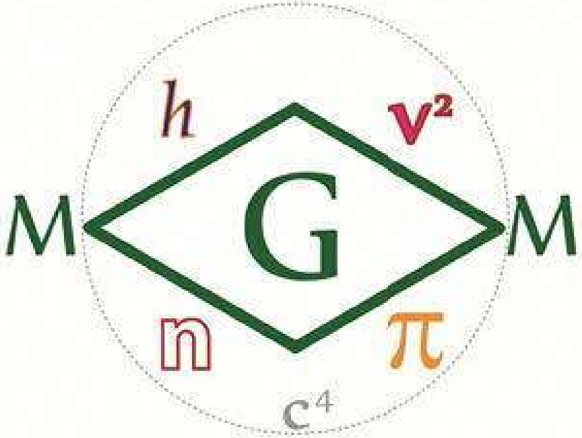
h
v²
M
G
M
n
π
c⁴

A physical constant is a physical quantity that is generally believed to be both universal in nature and constant in time.

It can be contrasted with a mathematical constant, which is a fixed numerical value but does not directly involve any physical measurement

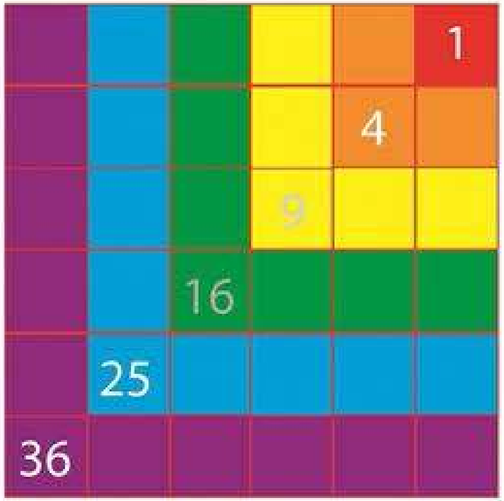

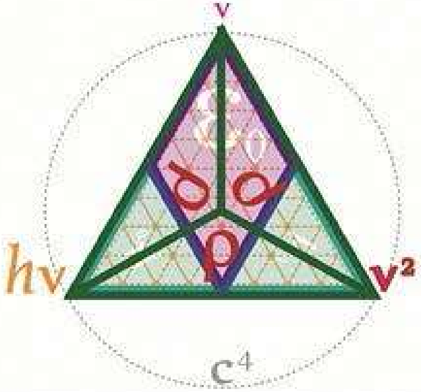

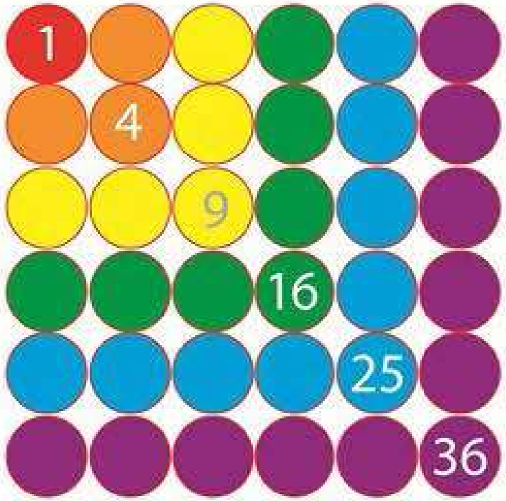

Tetryonics 16.01 - The geometry of Constants

Copyright ABRAHAM [2008] - all rights reserved

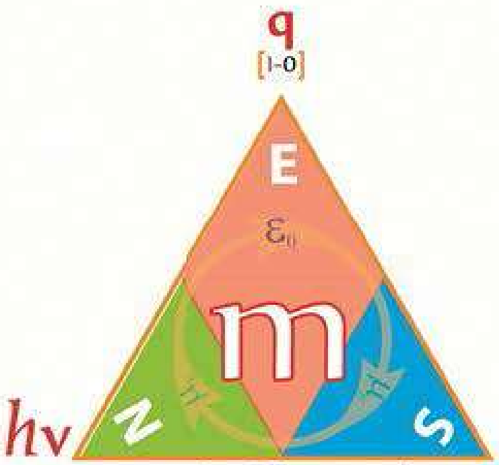

7.376238634 e-51 kg

"To interpret the vibrational
energy of N oscillators
not as a continuous,
infinitely divisible quantity,
but as a discrete quantity
composed of an integral number of
finite equal parts.

Max Planck

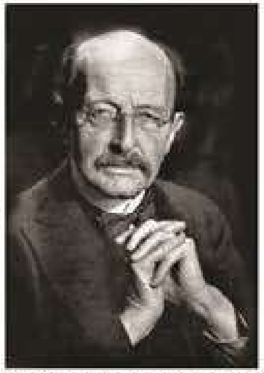

(April 23, 1858 – October 4, 1947)

$$\mathcal{E} = nh\nu$$

"Let us call each such part
of the energy element h"

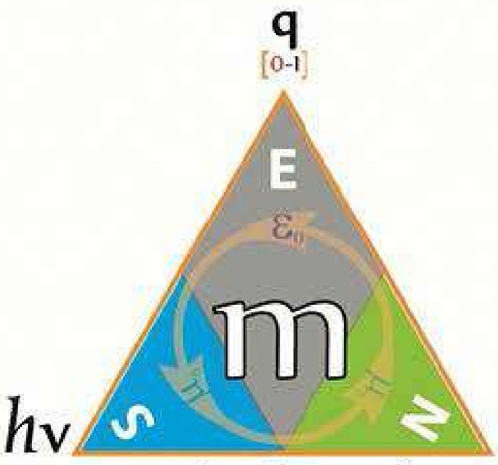

7.376238634 e-51 kg

Planck's Constant

[quantised mass-energy angular momenta]

N_A

1 mole of Hydrogen atoms
has a rest-mass of 1 gram

Solving for Planck's Constant using
the inverse of Avogadros number & Tetryonic geometry
we obtain an exact corrected value of :

$$\pm\ 6.629432672 \times 10^{-34}\ \text{J.s}$$

$$4.137664546 \times 10^{-15}\ \text{eV.s}$$

$$mv^2 = E = h\nu^2$$

A rest mass Hydrogen atom
has a Compton frequency of
2.2512 e23 Planck quanta

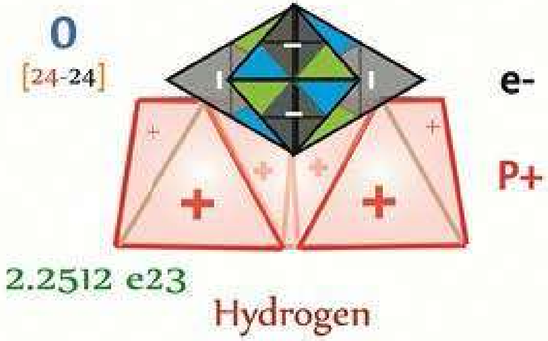

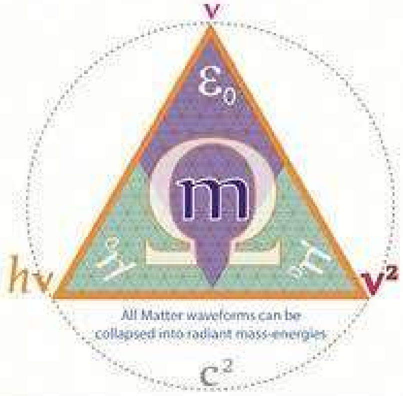

All Matter waveforms can be
collapsed into radiant mass-energies

EM Field Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter ElectroMagnetic mass velocity

*Planck's constant is the relationship between
EM mass-energy and quantised angular momenta
that provides the basis for EM charge in
Tetryonic geometries*

Copyright ABRAHAM [2008] - all rights reserved

Amedeo Avogadro

(9 August 1776 – 9 July 1856)

Avogadro's Number

The number of rest mass Hydrogen atoms in 1 gram
(and the rest molar mass of any element or compound)
can be determined directly from tetryonic theory

[exclusive of any measurement, blackbody or kinetic energies]

Using a Compton frequency of 2.2512 e23 Planck quanta
for a rest mass-Matter Hydrogen atom

atomic mass unit

$$1\ \mathrm{u} = \frac{M_\mathrm{u}}{N_\mathrm{A}} = 1.660\,538\,782(83)\times 10^{-24}\ \mathrm{g}$$

Avogadro N = $6.022141579 \times 10^{23}$

$$N^{-1} = \frac{2.2512\ \text{e}23\ v\ \text{[Hydrogen mass]}}{1.660538841 \times 10^{-24}\ \text{g}}$$

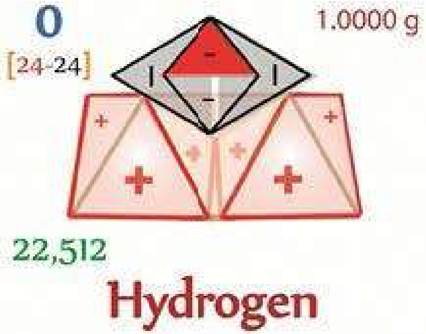

n0 *1 mol = 1 g*

Hydrogen = $1.660538841 \times 10^{-27}$ Kg

n1 *1 mol = 1.000533 g*

n0 *1 mol = 11.996801 g*

Carbon 12 = $1.99211552 \times 10^{-26}$ Kg

n1 *1 mol = 12 g*

Hydrogen's rest Tetryonic mass is 22,512 n Planck quanta [Proton - 22,500n + electron 12n]

$$\left[22{,}512\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{\left[\varepsilon_0\mu_0\right]}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]\right]^{-1}$$

The inverse mass of Hydrogen is equal to Avogadro's number

Coulomb's Constant

$$k = \frac{1}{4\pi\varepsilon_0}$$

The proportionality constant ke, called the Coulomb constant (sometimes called the Coulomb force constant), is related to defined properties of EM Energy-momenta and is used to define Electric field forces

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$$

$$8.987\ \text{e}9\ \frac{\text{Nm}^2}{\text{C}^2}$$

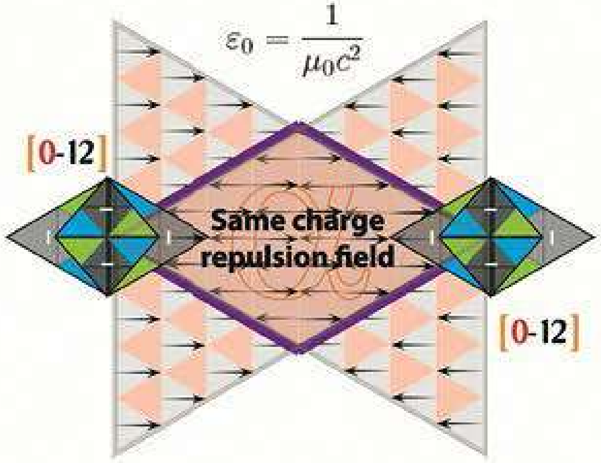

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

[0-12]
Same charge repulsion field
[0-12]

Similars repel

Linear Coulombic force interactions
are a result of charged E field
linear momenta

$$\mathbf{E} = \frac{\mathbf{F}}{q_t}$$

The Electric field can be defined as the longitudinal Force exerted by charged masses

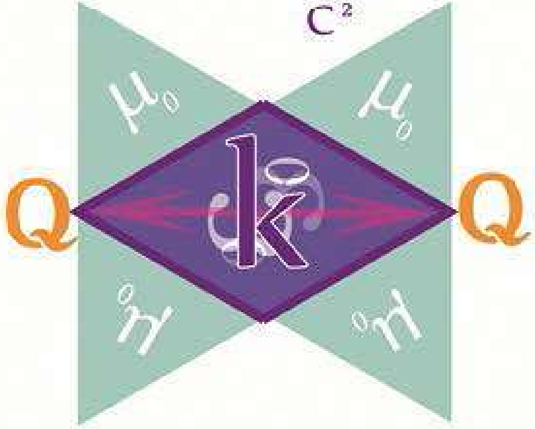

Q
μ_0
μ_0
k
ε_0
ε_0
Q

Opposites attract

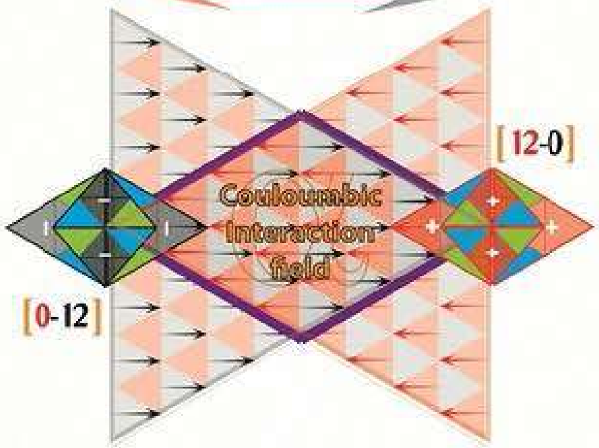

[12-0]
Couloumbic Interaction field
[0-12]

It is a measure of the interactive force
produced by the Electric field energy-momenta
of two superpositioned charge KEM fields

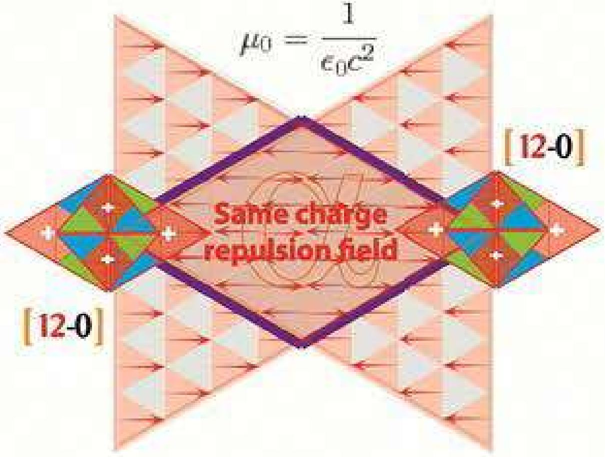

$$\mu_0 = \frac{1}{\epsilon_0 c^2}$$

[12-0]
Same charge repulsion field
[12-0]

Similars repel

Longitudinal E field forces
between Charged particles
are mediated by Photons

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}$$

The Electric field can also be derived from Coulomb's Law

The Speed of ElectroMagnetic energies

The classical behaviour of the electromagnetic field is described by Maxwell's equations, which predict that the speed c with which electromagnetic waves (such as light) propagate through the vacuum is related to the electric constant ε0 and the magnetic constant μ0 by the equation c = 1/√ε0μ0.

C

Celeritas is a Latin word for
"swiftness" or "speed".
It is often given as the origin
of the symbol c, the universal
notation for the
speed of light in a vacuum

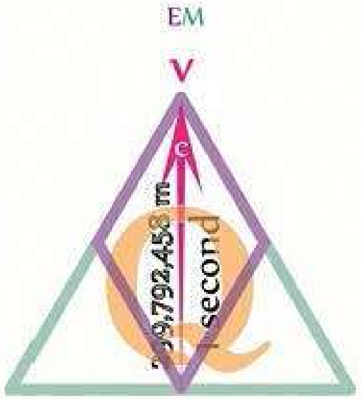

The distance energy travels
in 1 second from its source

EM waves and energy momenta
can be measured as either
Transverse or Longitudinal
waveforms with respect to
their velocity vector

All EM waves and energy
are symmetrical waveforms
whose quanta contain
ElectroMagnetic
mass-energy and momenta

The equilateral geometry of EM energy is the inverse of the spatial geometry of 1 second

The square root of equilateral energies [scalar speed] is a vector velocity

velocity is a vector
direction of the
energies of motion

speed is a scalar
property of the
energies of motion

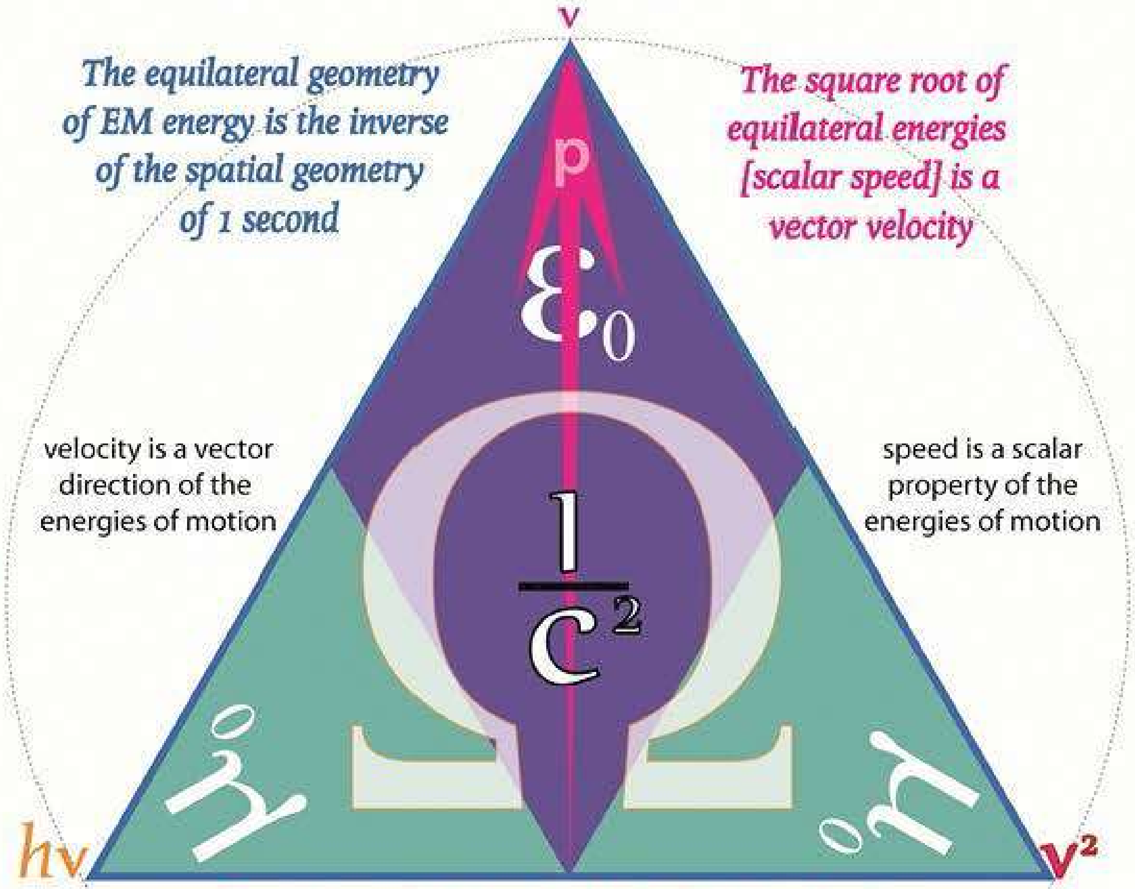

The maximum velocity possible due to electrical accelerations or the impedance of any spatial region that energy propagates through it is often referenced in physics as 'the speed of Light'
IT IS NOT AN ABSOLUTE LIMIT or BARRIER

All EM waves radiate
outwards in accord with
Biot-Savart's law

C²

The velocity vectors of
EM energies form the
geometries of
spatial co-ordinate systems

The speed of EM energies
in a vacuum is defined as
299,792,458 meters per second
(1,079,252,848.8 km/h).

$$\frac{K_E}{K_M} = 9e16 \, \frac{\frac{N\,m^2}{c^2}}{\frac{N\,s^2}{c^2}}$$

$$C^2 = 9e16 \, \frac{m^2}{s^2}$$

$$c = 299{,}792{,}458 \, \frac{m}{s}$$

The natural velocity of EM energies
can be calculated from the field's
EM permittivity & permeability
and is affected by the medium
it is propagating through

EM waveforms are
bidirectional radiant
emissions of 2D energy

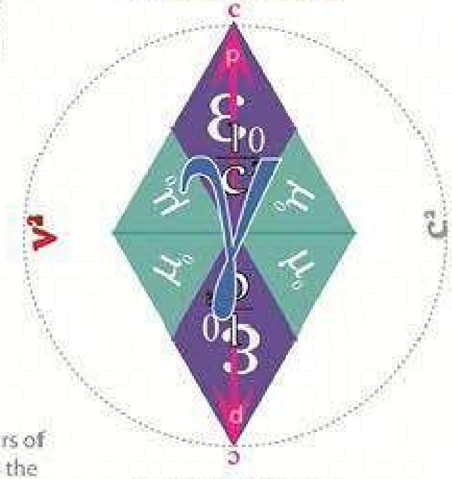

with velocity [Lorentz]
variable energy
contents

Copyright ABRAHAM [2008] - all rights reserved

mass quanta

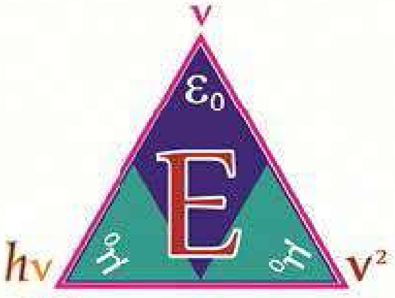

EM mass-energy quanta

Matter is a KEM standing wave propagating at c

EM Field planck quanta

$$[\varepsilon_0\mu_0]\,[m\Omega v^2]$$

mass velocity

1.112650056 e-17 s2 m-2

6.629432672 e-34 kg m2 s-2

Permittivity x Energy density

$$\frac{E}{c^2} = m$$

v
ε_0
m
v^2
mass
c

mass = 7.376238634 x 10^{-51} Kg

Planck-Einstein

Quantum masses

$$hv^2 = E = mv^2$$

$$kg\,\frac{m^2}{s}\,\frac{1}{s} \qquad kg\,\frac{m^2}{s^2} \qquad kg\,\frac{m^2}{s^2}$$

The quantum of mass-energy can be derived with several methodolgies using Tetryonic Geometry

Planck mass-energy geometries

quantised energy momenta

$$h$$

$$kg\,\frac{m^2}{s}$$

6.629432672 e-34 J

Planck quanta

$$[m\Omega v^2]$$

mass velocity

7.376238634 e-51 kg

$$kg\,\frac{m^2}{s^2}$$

$$m$$

scalar mass-energy

Charged mass-Matter topologies

The mass EM Field Planck quanta

$$T\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

of Matter ElectroMagnetic mass velocity

Matter Quanta

Tetryonic molar mass [Hydrogen] - 1 g

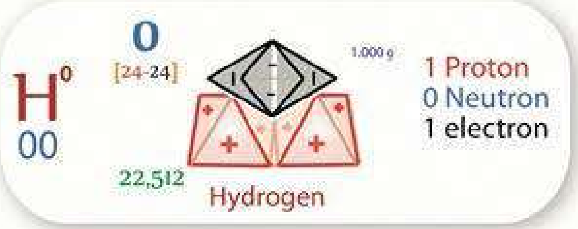

$$\text{Molar mass} = \frac{H_1\ \text{Atomic mass}}{\text{Avagadro's No.}}$$

$$\text{mass}\ H_{mol}/Av = \frac{.001}{6.022141579 \times 10^{23}}$$

$$= 1.660538841 \times 10^{-27}\ \text{g/mol}$$

$$H_{mole}/\triangle m[H] = \frac{1.660538841 \times 10^{-27}}{22{,}512}$$

mass = 7.376238634 x 10^{-32} KG

Avogadro - Mandeleev

Copyright ABRAHAM [2008] - all rights reserved

2D mass-Energy equivalence

'The Speed of Light' is the Natural velocity of Energy propagation

Energy

$$mv^2 = hv^2$$

EM mass

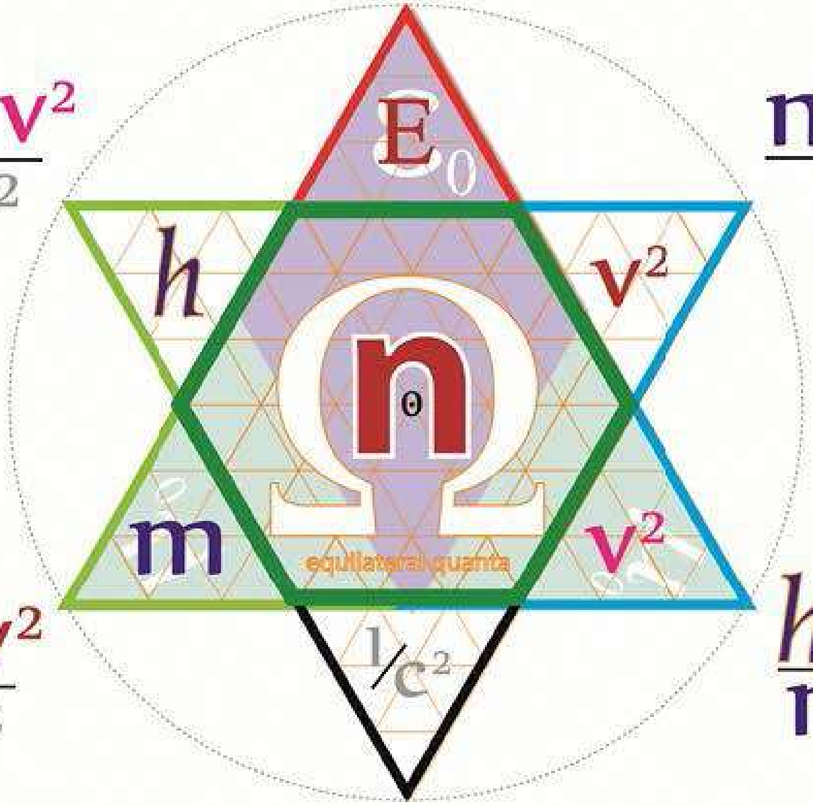
E_0
h
v^2
n_0
m
v^2
equilateral quanta
$^1/_{c^2}$

$$\frac{mv^2}{c^2}$$

$$\frac{mv^2}{h}$$

$$\frac{c^2}{v^2}$$

8.987551787 e16

1.112650056 e-17

$$\frac{m}{h}$$

$$\frac{hv^2}{c^2}$$

$$\frac{hv^2}{m}$$

$$\frac{m}{hv^2}$$

Photons

Matter

Tetryonics 16.07 - 2D mass-Energy equivalence

Zero Point Fields

EM Field Planck quanta

$n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Quanta ElectroMagnetic mass velocity

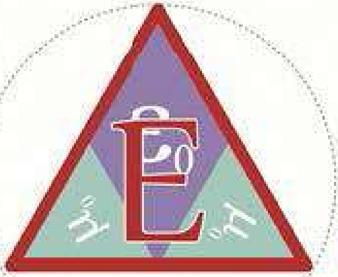

mass-energy

m=1π

radiant 4π mass-energy geometries

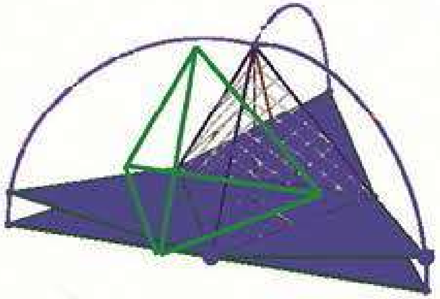

in standing wave topologies are 3D MATTER quanta

Photons

EM Field Planck quanta

$2n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Photons ElectroMagnetic mass velocity

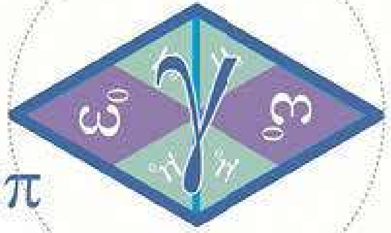

m=2π

EM energy carriers

[1, 2, 3, 4, 5, 6, 7, 8, 9,10......]

[2D EM masses are planar energy momenta]

[2, 8,18, 32, 50, 72, 98,128......]

The geometry of mass and Matter

[1, 3, 5, 7, 9, 11, 13,15, 17, 19........]

[3D Matter has mass-energy momenta and volume]

[4, 8,12, 24, 36, 48, 72, 84, 168......]

Charge Carriers

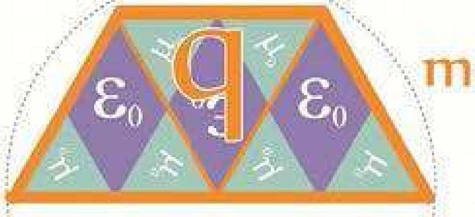

m=3π

EM Field Planck quanta

$ODDn\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Bosons ElectroMagnetic mass velocity

Bosons

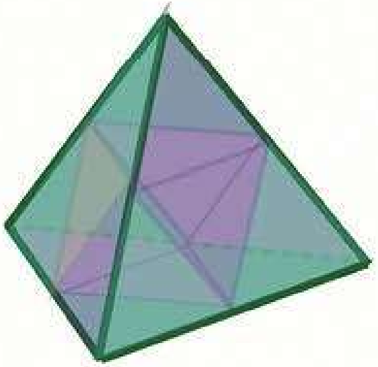

3D Matter topologies are standing waves of 2D mass-energies propagating at c

m=4π

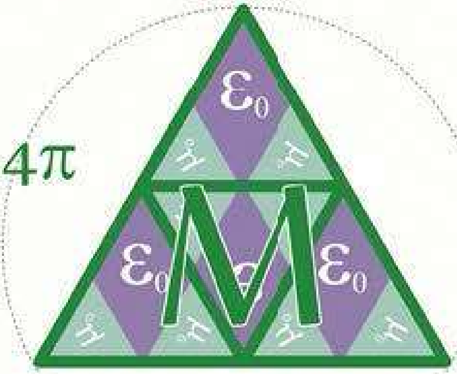

EM Field Planck quanta

$4n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Matter ElectroMagnetic mass velocity

mass-Matter

Tetryonics 17.01 - The geometry of mass and Matter

mass geometries & rest Matter topology

2D

m

mass

kg

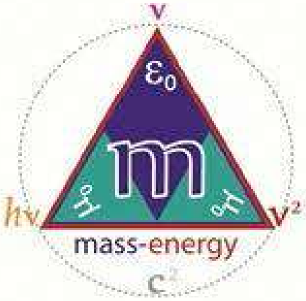
v
ε₀
m
hv
v²
mass-energy
c²

M

Matter

KG

3D

Quanta number

1e19 v

Compton frequency

Matter is a three dimensional charged mass-energy topology

EM Field Planck quanta

$$M = 4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter ElectroMagnetic mass velocity

Energy density is mass - the term 'massless particle' is a misnomer

Quantum number

n

Nuclear Level

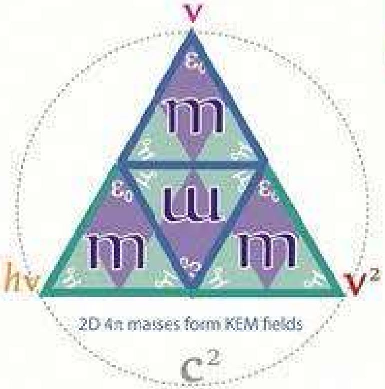
v
ε₀
m
ε₀
m
ε₀
m
m
hv
v²
2D 4π masses form KEM fields
c²

Kinetic Energies

RE = Matter + KE

Relativistic mass is the total EM energy content of a massive body (or system) in motion

The relativistic rest mass-energies of Matter are velocity invariant as 3D Matter is a EM standing wave with 2D mass- energy fascia whose velocity of propagation is the speed of Light

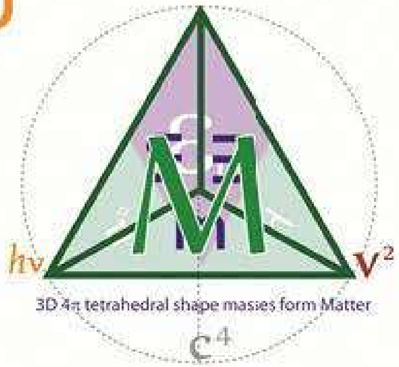
ε₀
M
m
hv
v²
3D 4π tetrahedral shape masses form Matter
c⁴

Matter

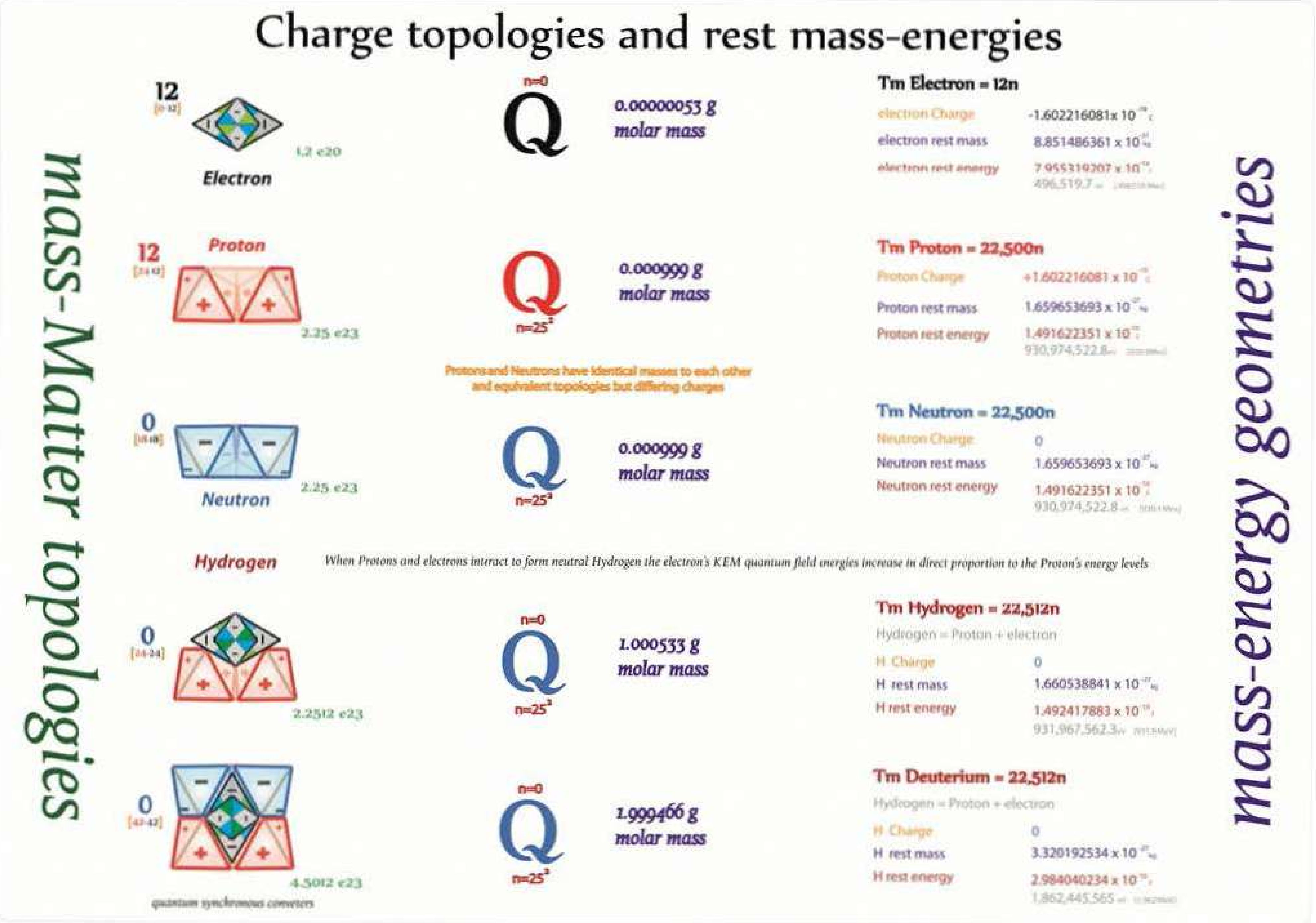
mass-Matter topologies
mass-energy geometries
12
[0-12]
1.2 e20
Electron
n=0
Q
12
[24-12]
Proton
2.25 e23
n=25²
Q
0
[18-18]
Neutron
2.25 e23
n=25²
Q
Hydrogen
0
[24-24]
2.2512 e23
n=0
Q
n=25²
0
[42-42]
4.5012 e23
n=0
Q
n=25²
quantum synchronous converters

Charge topologies and rest mass-energies

Tm Electron = 12n

0.00000053 g
molar mass

electron Charge	$-1.602216081 \times 10^{-19}$ C
electron rest mass	$8.851486361 \times 10^{-31}$ kg
electron rest energy	$7.955319207 \times 10^{-14}$ J 496,519.7 eV [illegible]

Tm Proton = 22,500n

0.000999 g
molar mass

Proton Charge	$+1.602216081 \times 10^{-19}$ C
Proton rest mass	$1.659653693 \times 10^{-27}$ kg
Proton rest energy	$1.491622351 \times 10^{-10}$ J 930,974,522.8 eV [illegible]

Protons and Neutrons have identical masses to each other and equivalent topologies but differing charges

Tm Neutron = 22,500n

0.000999 g
molar mass

Neutron Charge	0
Neutron rest mass	$1.659653693 \times 10^{-27}$ kg
Neutron rest energy	$1.491622351 \times 10^{-10}$ J 930,974,522.8 eV [illegible]

When Protons and electrons interact to form neutral Hydrogen the electron's KEM quantum field energies increase in direct proportion to the Proton's energy levels

Tm Hydrogen = 22,512n

1.000533 g
molar mass

Hydrogen = Proton + electron

H Charge	0
H rest mass	$1.660538841 \times 10^{-27}$ kg
H rest energy	$1.492417883 \times 10^{-10}$ J 931,967,562.3 eV [illegible]

Tm Deuterium = 22,512n

1.999466 g
molar mass

Hydrogen = Proton + electron

H Charge	0
H rest mass	$3.320192534 \times 10^{-27}$ kg
H rest energy	$2.984040234 \times 10^{-10}$ J 1,862,445,565 eV [illegible]

Tetryonics 17.03 - Charge topologies and rest mass-energies

Copyright ABRAHAM [2008] - all rights reserved

In order to make an exact 1kg reference mass-Matter topology

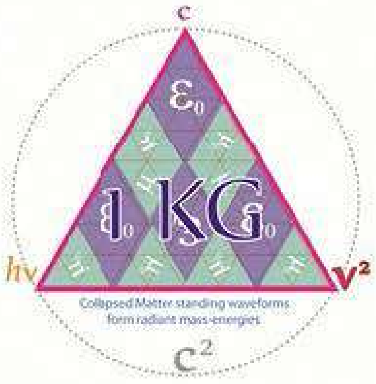
c
ε0
1 KG
ε0
hv
v²
Collapsed Matter standing waveforms form radiant mass-energies
c²

mass

$$\frac{n\pi}{c^2}\Big[\Big[m\Omega v^2\Big]\Big]$$

Planck quanta mass velocity

mass-energy

Energy per second

$$\boxed{mc^2 = E = hv^2}$$

$$m = E/c^2$$

mass is the scalar integral surface area of Matter topologies

which is equivalent to exactly 1,000 molar Hydrogen masses

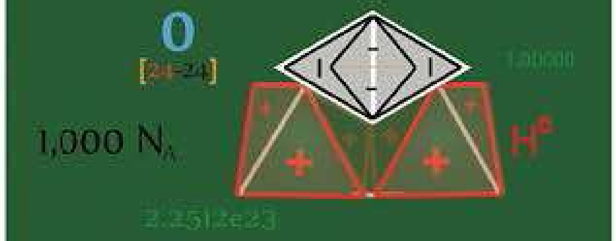
0
[24-24]
1,000 N_A
1.00000
H⁰
2.2512e23

whose Tetryonic Matter topology can be modelled using charged Planck mass-energy geometries

$$M = 1{,}000 \times N.\,T\pi\Big[\Big[m\Omega v^2\Big]\Big]$$

1kg

grams per kg

Energy

6.629432672 e-34 J

Avogadro's number

revealing that there are exactly

1.355704512 e50

Planck quanta in standing wave Matter

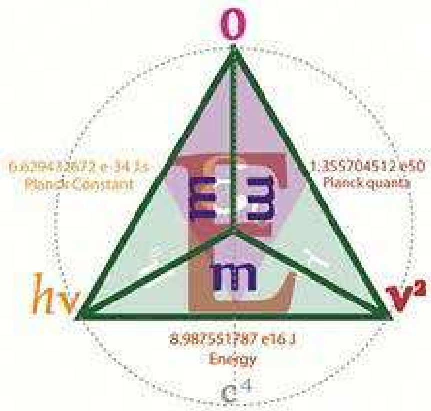
0
6.629432672 e-34 Js
Planck Constant
1.355704512 e50
Planck quanta
E
m
hv
v²
8.987551787 e16 J
Energy
c⁴

Matter

$$\frac{T\pi}{c^4}\Big[\Big[m\Omega v^2\Big]\Big]$$

Planck quanta mass velocity

Matter-energy

Energy per second squared

$$\boxed{hv^2 = E = Mc^4}$$

$$M = E/c^4$$

Matter topologies are standing-waves of mass-energy

you need to create a standing-EM wave with a specific compton frequency

Tetryonics 17.04 - Defining 1kg of Matter

EM mass and Matter defined

There remains a lot of confusion over the exact definition of EM mass and Matter resulting in the frequent interchanging of one term for the other in physical processes

This must be clarified and the two terms must be properly defined in a manner that explains their derivation and physical properties in detail.

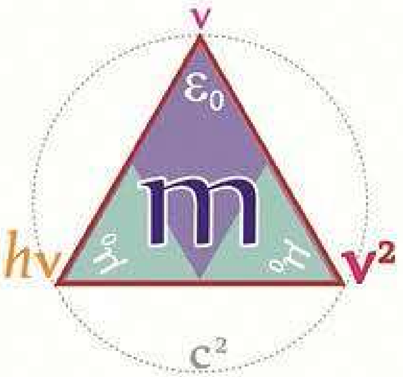

$$m = \frac{E}{c^2}$$

EM mass

3D Matter is comprised of 2D mass-energies
2D mass-energy cannot contain 3D Matter

Matter

$$M = \frac{4\pi E}{c^4}$$

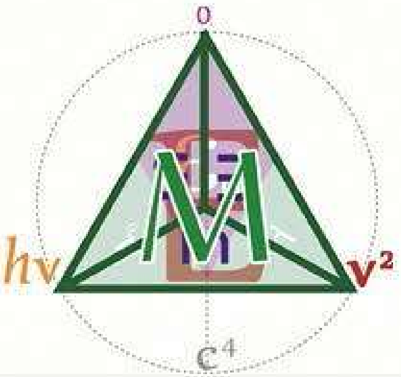

$n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM Field Planck quanta
mass
ElectroMagnetic mass velocity

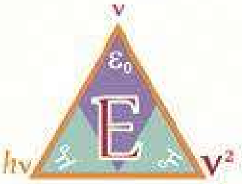

$n\pi[[m\Omega v^2]]$

Planck quanta
Energy
mass velocity

2D mass-Energies

EM mass is a measure of equilateral scalar energy per unit of Time

$$\beta = \left[\frac{v}{c}\right]$$

mass is the Lorentzian velocity corrected energy content of Tetryonic geometries

All forms of EM mass-Energy are subject to Lorentz factor relationships

$$\beta^2 = \left[\frac{v}{c}\right]^2$$

Matter is Lorentz invariant as its mass-energies propagate in a 4π standing wave geometry

Bosons and Photons are not 'massless' they are 'Matterless' [2D waveforms]

Bosons

$$m_p = \frac{hv}{c^2}$$

Planck masses

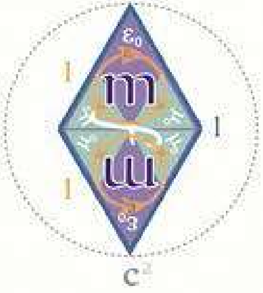

Photons

$$m_\gamma = \frac{hf}{c^2}$$

Photon mass

$ODD\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM Field Planck quanta
Bosons
ElectroMagnetic mass velocity

$EVEN\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM Field Planck quanta
EM waves
ElectroMagnetic mass velocity

$4n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM Field Planck quanta
Matter
ElectroMagnetic mass velocity

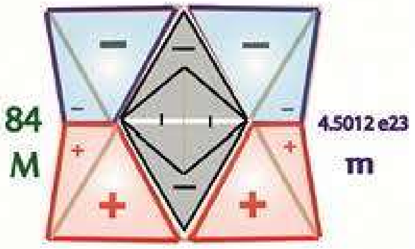

$84\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM Field Planck quanta
Deuterium
ElectroMagnetic mass velocity

Matter is the 3D topology of standing wave 2D mass-energies

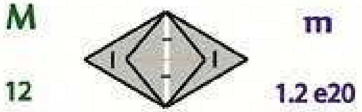

$12\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$ e-

EM Field Planck quanta
electrons

+

$72\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$ Deuteron

Nuclei
ElectroMagnetic mass velocity

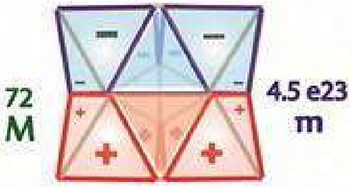

Copyright ABRAHAM [2008] - all rights reserved

Charged mass-Matter geometries

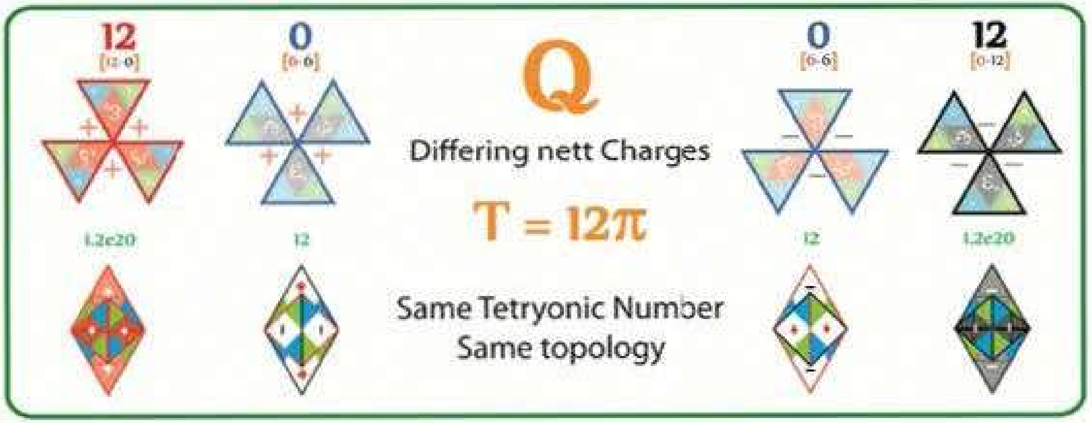

Differing nett Charges

$T = 12\pi$

Same Tetryonic Number
Same topology

Matter Topology is determined by the geometry of charged Planck T[q] quanta

Any measurement of a system's mass is subject to velocity corrections

mass = EM energy density

$$\underset{\text{mass quanta}}{n\pi}\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right] \over \underset{\text{Spatial geometry}}{c^2}$$

n [7.376238634 e-51] kg

EM mass is a measure of 2D planar energies comprising the fascia of charged topologies

$$\textbf{m} \qquad \overset{\text{ENERGY}}{n\pi}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right] \qquad \textbf{M}$$

c^2
2D space

The mass - ENERGY - Matter contents of any physical system are all related through the spatial co-ordinate system used [which in turn is determined by the speed of light]

c^4
3D space

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right] \qquad \textbf{E} \qquad \overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

The Energy content of any physical system remains the same irrespective of the spatial co-ordinates used

Matter and Charge are velocity invariant properties

Differing Tetryonic charge numbers produce differing particle topologies

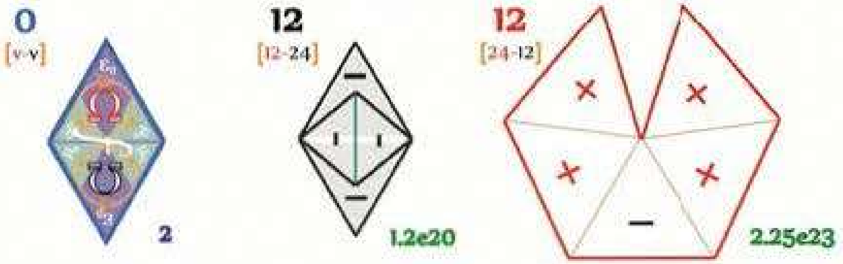

Photons $\frac{2\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$

Dodecons $\frac{12\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$

Baryons $\frac{36\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$

As the energy content [levels] of particles and fields increase their intrinsic Planck quanta and EM mass increases but their charge geometries remain the same

Tetryonics 17.06 - Charged mass-Matter geometries

Copyright ABRAHAM [2008] - all rights reserved

Charged mass-Energies

All 3D Matter particles are comprised of charged fascia whose energy content determines their 2D mass

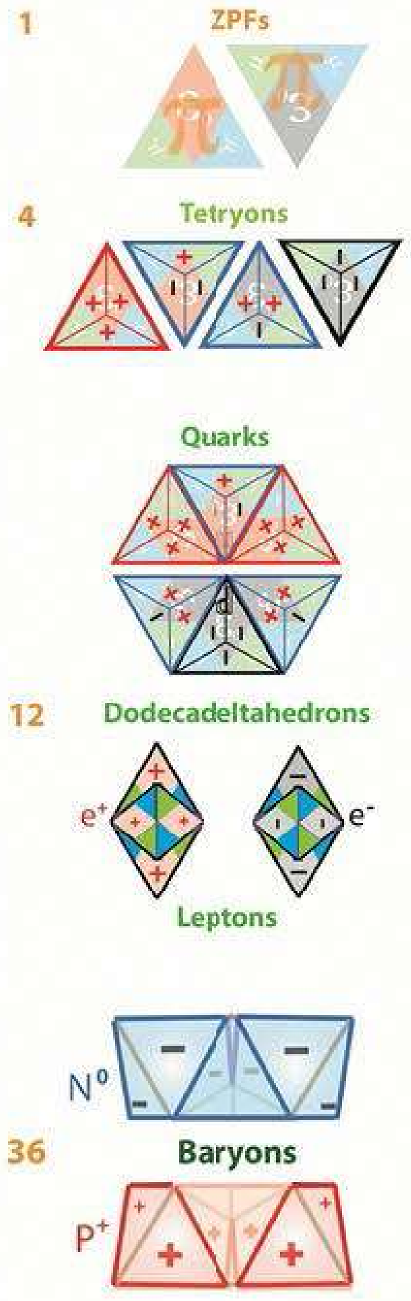

1
ZPFs
4
Tetryons
Quarks
12
Dodecadeltahedrons
e+
e-
Leptons
N0
36
Baryons
P+

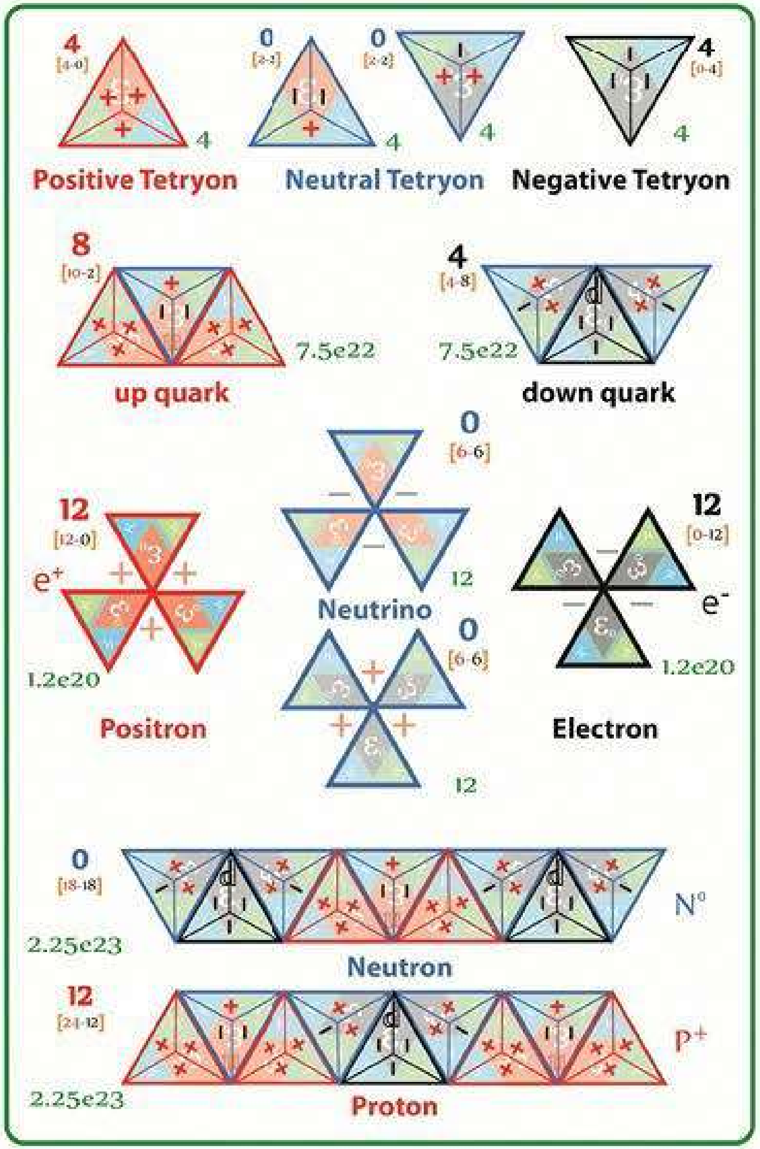

4
[4-0]
4
Positive Tetryon
0
[2-2]
4
0
[2-2]
4
Neutral Tetryon
4
[0-4]
4
Negative Tetryon
8
[10-2]
7.5e22
up quark
4
[4-8]
7.5e22
down quark
12
[12-0]
e+
1.2e20
Positron
0
[6-6]
12
Neutrino
0
[6-6]
12
12
[0-12]
e-
1.2e20
Electron
0
[18-18]
2.25e23
N0
Neutron
12
[24-12]
2.25e23
P+
Proton

3D Matter geometries

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Matter — Planck quanta — mass — velocity

Tetryons 4 charge fascia	4π	$2.651773069 \times 10^{-33}$ J $2.950495454 \times 10^{-50}$ kg	4 quanta
Quarks 12 charge fascia	12π	$4.97207450 \times 10^{-11}$ J $5.532178976 \times 10^{-28}$ kg	7.5×10^{22} quanta
Neutrinos 12 charge fascia	12π	$7.955319207 \times 10^{-33}$ J $8.851486361 \times 10^{-50}$ kg	12 quanta
Leptons 12 charge fascia	12π	$7.955319207 \times 10^{-14}$ J $8.851486361 \times 10^{-31}$ kg	1.2×10^{20} quanta
Baryons 36 charge fascia	36π	$1.491622351 \times 10^{-10}$ J $1.659653693 \times 10^{-27}$ kg	2.250×10^{23} quanta
Deuterium ion 72 charge fascia	72π	$2.984040234 \times 10^{-10}$ J $3.320192534 \times 10^{-27}$ kg	4.5012×10^{23} quanta

Copyright ABRAHAM [2008] - all rights reserved

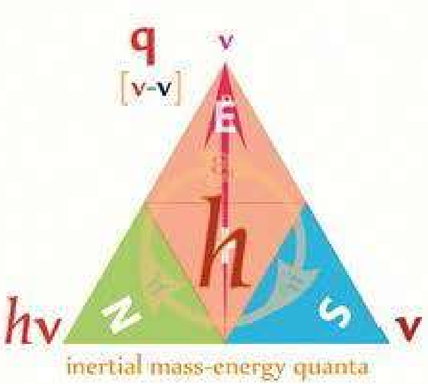
q
[v-v]
v
E
h
hv
N
S
v
inertial mass-energy quanta

In physics there exists
2 forms of energy momenta
[linear momenta &
quantised angular]

MOMENTA

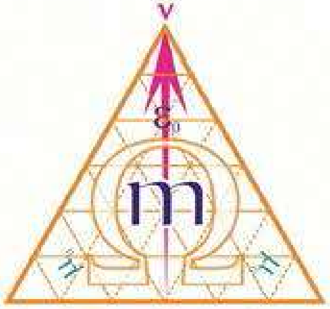
v
ε_0
m

The nett scalar
energy momenta
required to effect
changes in velocity

ENERGY

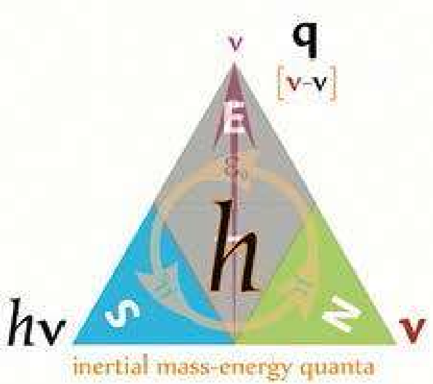
v
q
[v-v]
E
h
hv
S
N
v
inertial mass-energy quanta

Net Charge is reflective of the total energy momenta
of any EM field or Matter subjected to acceleration
[inertial mass]

Inertial resistance to Force

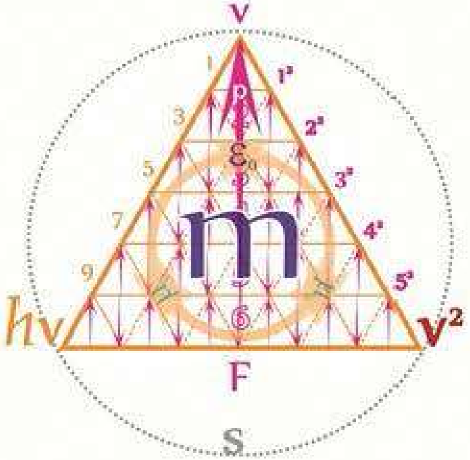
v
p
ε_0
m
1
3
5
7
9
1^2
2^2
3^2
4^2
5^2
hv
v^2
F
s

Any change in motion results in changes to
the Charge geometries creating in turn proportional
changes to the nett KEM mass-energy momenta components

$$F=ma$$

The inductive resistance of quanta in charged EM fields
to any changes to their nett mass-energy geometry
is the source ofwhat we term inertial mass

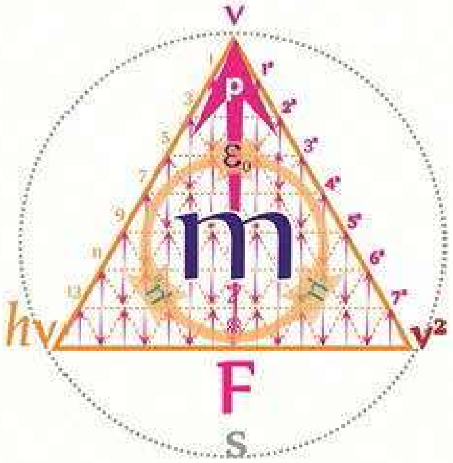
v
p
ε_0
m
1
3
5
7
9
11
13
1^2
2^2
3^2
4^2
5^2
6^2
7^2
hv
v^2
F
s

**Inertia is the resistance of any physical object
to a change in its state of motion.**

Force does work when it results in movement

$$\boxed{W = F.d = ma.d = E}$$

$$kg\,\frac{m}{s^2}\,m$$

classical mechanics

scalar

quantum mechanics

$$\Delta mv = \Delta E = \Delta m\Omega v$$

energies

linear momenta

angular momenta

$$kg\,\frac{m^2}{s^2}$$

The Planck constant is the quantum of Action

$$\boxed{E = m\Omega v^2 = hv^2}$$

$$kg\,\frac{m^2}{s}\,\frac{1}{s}$$

Changing an object's velocity results in a corresponding energy momenta change which relate to each other through its inertial mass

Tetryonics 17.08 - Inertial resistance to Force

EM field densities

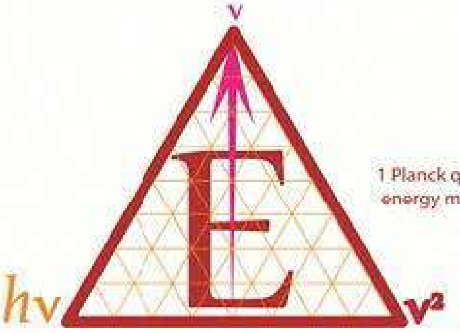

1 Planck quanta of energy momenta

EM mass is a measure of the energy content of any spatial co-ordinate system

$$\beta^2 = \left[\frac{v}{c}\right]^2$$

and is subject to Lorentz velocity corrections in 2D EM fields

6.629432672 e-34 J
1 planck quantum has a EM mass-Energy of
7.376238634 e-51 kg
and
Quantised Angular Momentum which creates Charge
1.33518 e-20 C

mass-Energy equivalence

mass velocity

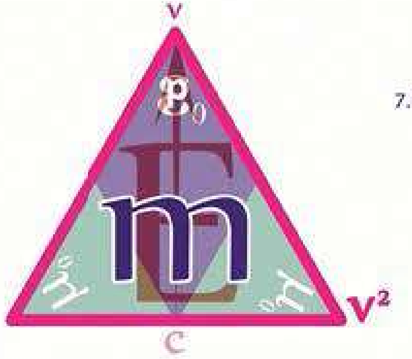

7.376238634 e-51 kg

6.629432572e-34 j.s

$$mv^2 = E = hv^2$$

8.987551787 e16 [m/s]²

6.629432672 e-34 j

quanta/sec

Matter is only mass-energy in Tetrahedral topologies [T4π+]
[else it is nπ EM mass-energies that propagate away at c]

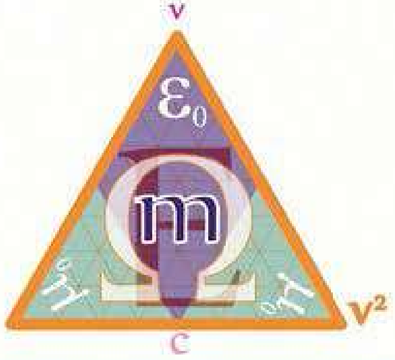

planck quanta

6.629432672 e-34 J

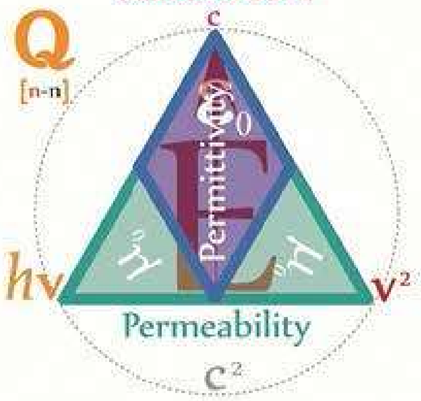

7.376238634 e-51 kg
radiant mass-energies

If reduced to a flat Euclidean space geometry
Matter topologies become radiant mass-energies

m
mass

E
ENERGY

M
Matter

$$\left[\frac{n\pi}{c^2}\rho\right]$$

Planck quanta

$$\left[m\Omega v^2\right]$$

mass velocity

$$\left[\frac{4n\pi}{c^4}\rho\right]$$

The inertial properties of electromagnetic mass - ENERGY & Matter
[can all be differentiated as energy densities per unit of time]
in any spatial co-ordinate system

2.984040234 e-10 J

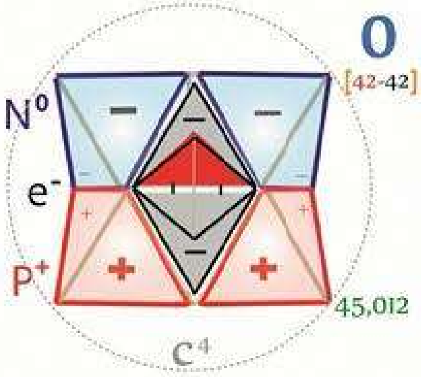

3.320192534 e-27 kg
rest mass-Matter

m

$[odd\pi]hv$

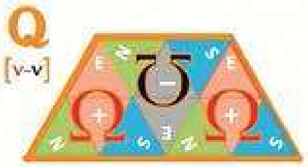

EM Field Planck quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Charge ElectroMagnetic mass velocity

Bosons

2D

Photons

m

$[even\pi]hv$

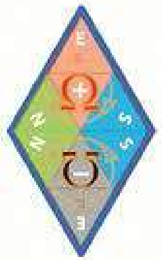

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

EM Field geometries & Matter topologies

M

$[4n\pi]hv$

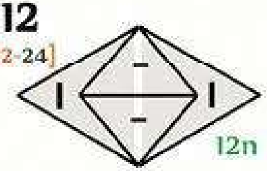

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Dodecyon ElectroMagnetic mass velocity

Leptons

3D

Baryons

M

$[4n\pi]hv$

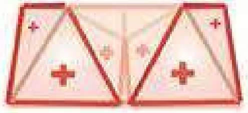

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

2D planar Euclidean electromagnetic mass-energies propagate through the vacuum energy aether without interaction

EM Field Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

mass ElectroMagnetic mass velocity

2D Fields

3D Matter

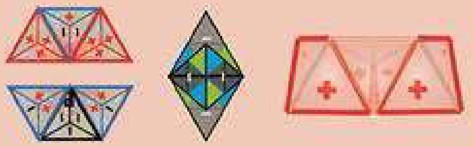

Tetryonic charged geometries provides a clear mechanism for providing all particles with their distinct properties of 2D inertial EM mass-energy and 3D Matter.

EM Field Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter ElectroMagnetic mass velocity

3D standing wave topologies of electromagnetic mass-energies interact with the vacuum energy aether at various angles through their charged [inductive] fascia

Tetryonics 17.10 - EM Field geometries and Matter topologies

Velocity invariant rest Matter

The Relativistic EM mass-energies of a system in motion is the sum of its invariant rest Matter and Kinetic Energies

Electrostatic charges have no Magnetic Moments

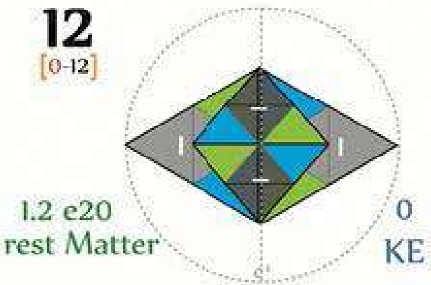

All Matter are 3D standing wave topologies

Electron rest Matter

Kinetic Energy field

$$RE = \underset{\text{Fermions}}{12\pi}\frac{[[m\Omega v^2]]}{c^4} + \underset{\text{Photons}}{2\pi}\frac{[[m\Omega v^2]]}{c^2}$$

EM fields resulting from motion are subject to Lorentz correction

Wavelengths are proportional to EM energy content

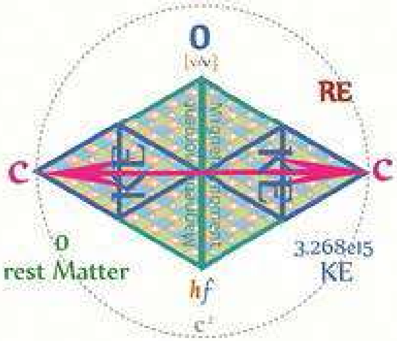

Photons are bidirectional Kinetic EM Fields

$$M = \frac{E}{c^4}$$

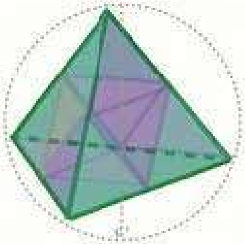

$$E = Mc^4$$

rest mass-Matter is composed of 4nπ standing wave topologies and is INVARIANT to velociiy changes

Kinetic mass-energies are divergent from invariant rest Matter topologies as a result of a particle's motion

At zero velocity the relativistic mass is equal to the invariant mass.

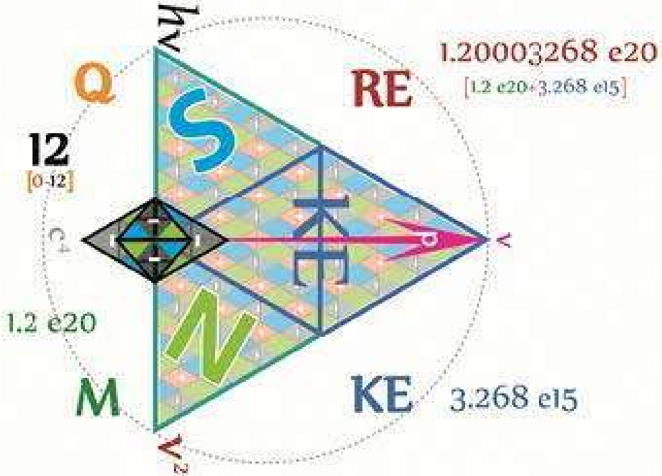

The energy which an object has due to its motion will not add mass into the invariant rest Matter of the particle in motion (it increases the total Planck quanta [EM mass-enegies] of its extended KEM field)

$$\frac{E}{c^2} = m$$

Matter in motion has a resultant velocity related Energy field that posseses the physical properties of Kinetic energy and Magnetic moments

$$KEM = Mv^2$$

KEM field Energy is directly relaied to the Velocity changes of massive particles

All 2D EM fields and 3D Matter particles have ElectroMagnetic fields, inertial mass-energies & momenta resulting from their constituent equilateral Planck quanta which possess the additional physical properties of Compton Frequency and De Broglie Wavelength

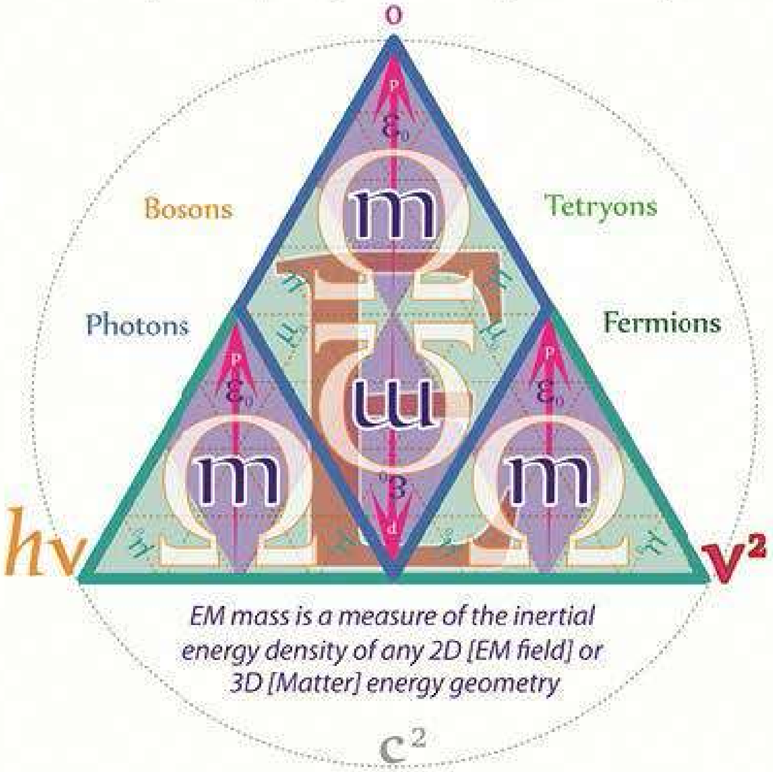

EM mass is a measure of the inertial energy density of any 2D [EM field] or 3D [Matter] energy geometry

$$m\Omega v^2 = E = mv^2$$

EM mass-energy momenta

The EM mass of an object is a fundamental property of the object; a numerical measure of its inertia; a fundamental measure of the energy density of an object.

mass geometries and Matter Topologies

All 3D Matter topologies contain 2D EM mass-energy geometries not all 2D EM mass-energy geometries form 3D Matter topologies

EM mass can be clearly defined as a measure of the energy density of any charged geometry

EM Field Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter ElectroMagnetic mass velocity

3D Matter is any mass-energy geometry that creates a closed volume Topology (4nπ Tetryonic geometry)

EM waves [FIELDS] are distinguishable from Material Particles [MATTER] through their non-Tetrahedral topologies

$$\frac{E}{c^2} = m$$

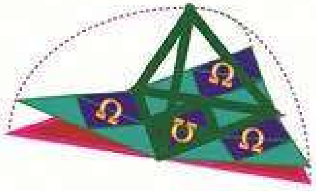

mass-energy is a conserved property
Matter is not conservative

$$M = \frac{E}{c^4}$$

The term 'massless' is a misnomer and should be discontinued in its use as all EM fields and Particles have EM mass [Energy quanta per second] geometries [other alternatives could be 2D, EM field, or Matterless]

mass-ENERGY-Matter equivalence

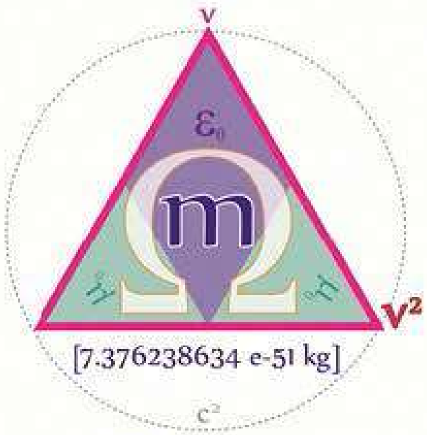
v
ε_0
m
Ω
v^2
[7.376238634 e-51 kg]
c^2

EM mass

$$\frac{E}{c^2} = m$$

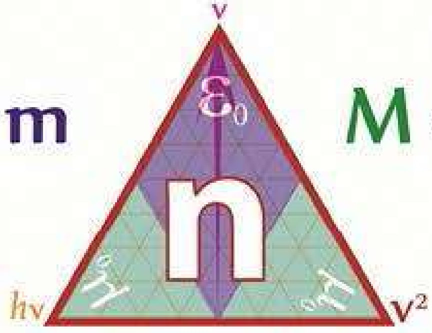
v
ε_0
n
hv
v^2

[6.629432672 e-34 kg.m²/s²]

ENERGY

All attempts to clearly differentiate between planar EM mass-energy & 3D Matter in physics have met with limited success until the introduction of charged geometries

$$M = \frac{E}{c^4}$$

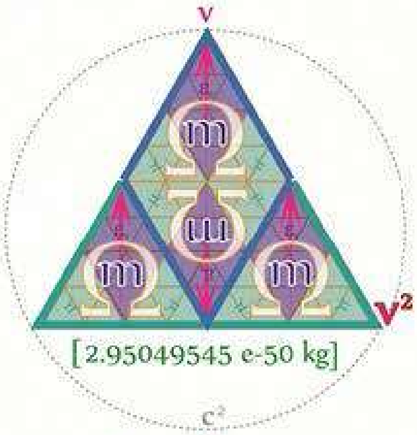
v
m
Ω
v^2
[2.95049545 e-50 kg]
c^2

Collapsed Matter

mv mass-velocity squared is Energy ← nπ → Planck quanta is Energy content **mΩ**

Newton and Leibniz measured Energy as a velocity related function of mass

Planck and Einstein viewed mass-energies as quantised particles with momenta

Energy momenta

$$mv^2 = E = hv^2$$

EM masses are subject to velocity related Lorentz corrections

Matter is a standing-wave of electromagentic energies and is Lorentz invariant

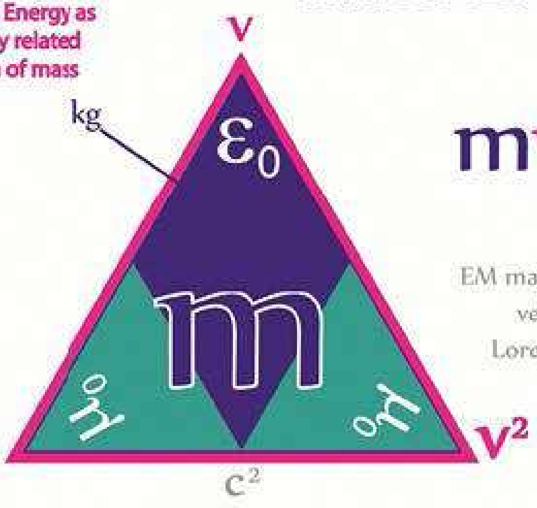
v
kg
ε_0
m
v^2
c^2

EM mass quantum

Planck quanta

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

mass velocity

Planck quanta

$$n\pi[[m\Omega v^2]]$$

mass velocity

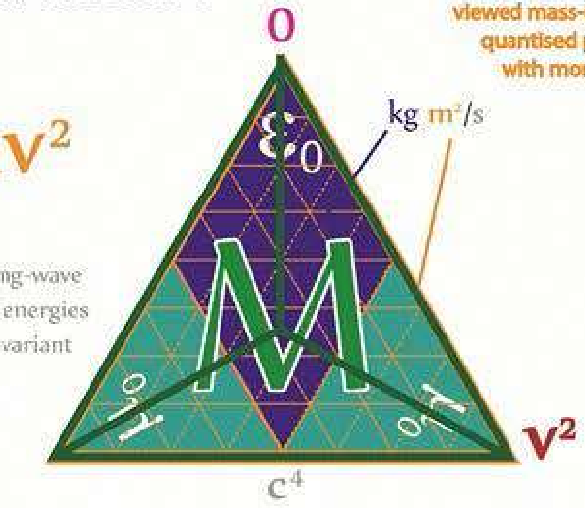
0
ε_0
kg m²/s
M
v^2
c^4

Matter quantum

Planck quanta

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

mass velocity

Tetryonics 17.13 - EM mass-Energy-Matter equivalence

Copyright ABRAHAM [2008] - all rights reserved

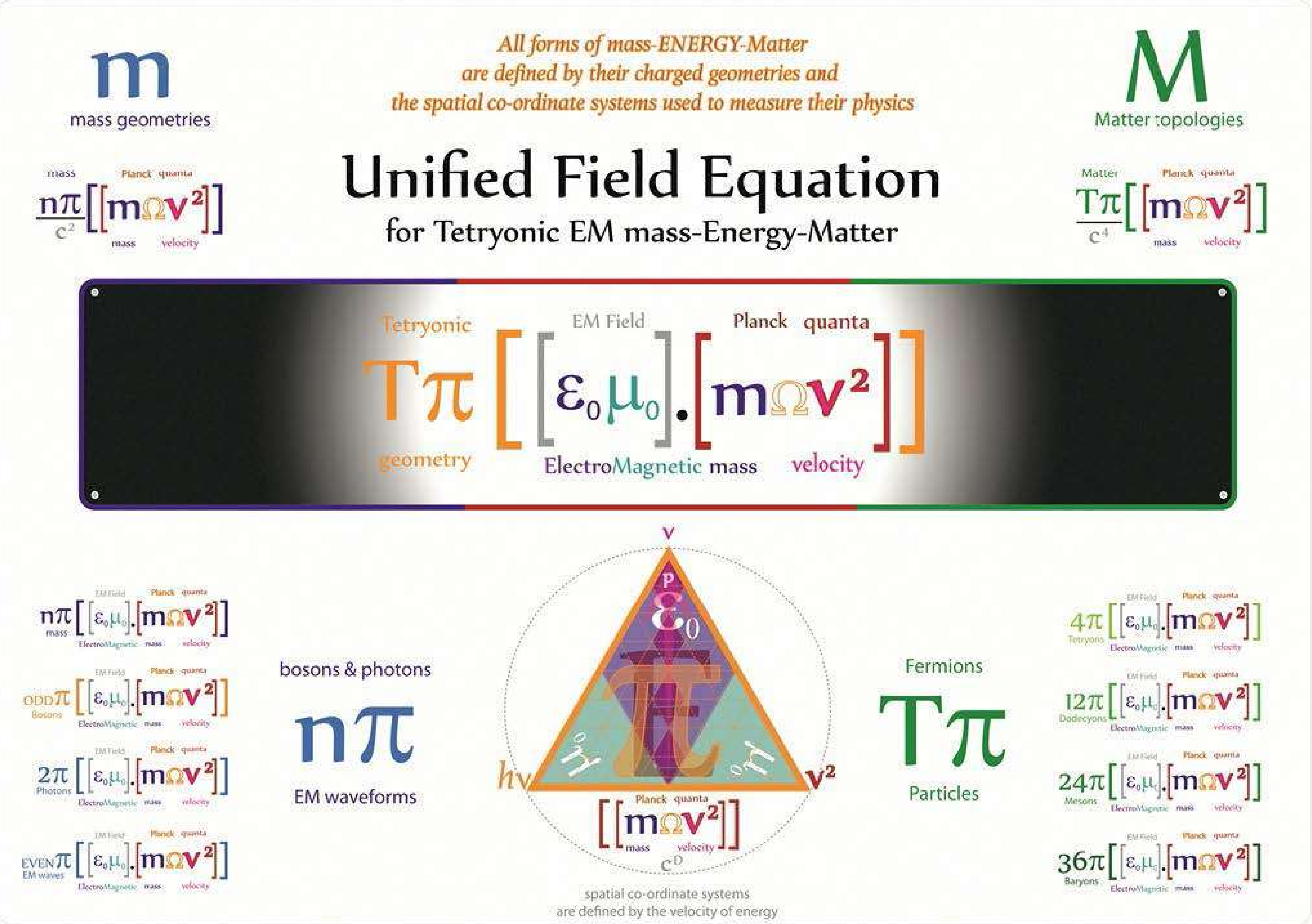

m

mass geometries

mass · Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass · velocity

All forms of mass-ENERGY-Matter
are defined by their charged geometries and
the spatial co-ordinate systems used to measure their physics

Unified Field Equation

for Tetryonic EM mass-Energy-Matter

M

Matter topologies

Matter · Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass · velocity

Tetryonic · EM Field · Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

geometry · ElectroMagnetic mass · velocity

EM Field · Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

mass · ElectroMagnetic mass · velocity

EM Field · Planck quanta

$$\text{ODD}\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons · ElectroMagnetic mass · velocity

EM Field · Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons · ElectroMagnetic mass · velocity

EM Field · Planck quanta

$$\text{EVEN}\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM waves · ElectroMagnetic mass · velocity

bosons & photons

nπ

EM waveforms

spatial co-ordinate systems
are defined by the velocity of energy

Fermions

Tπ

Particles

EM Field · Planck quanta

$$4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Tetryons · ElectroMagnetic mass · velocity

EM Field · Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Dodecyons · ElectroMagnetic mass · velocity

EM Field · Planck quanta

$$24\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Mesons · ElectroMagnetic mass · velocity

EM Field · Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons · ElectroMagnetic mass · velocity

Tetryonics 17.14 - Unified Field Equation

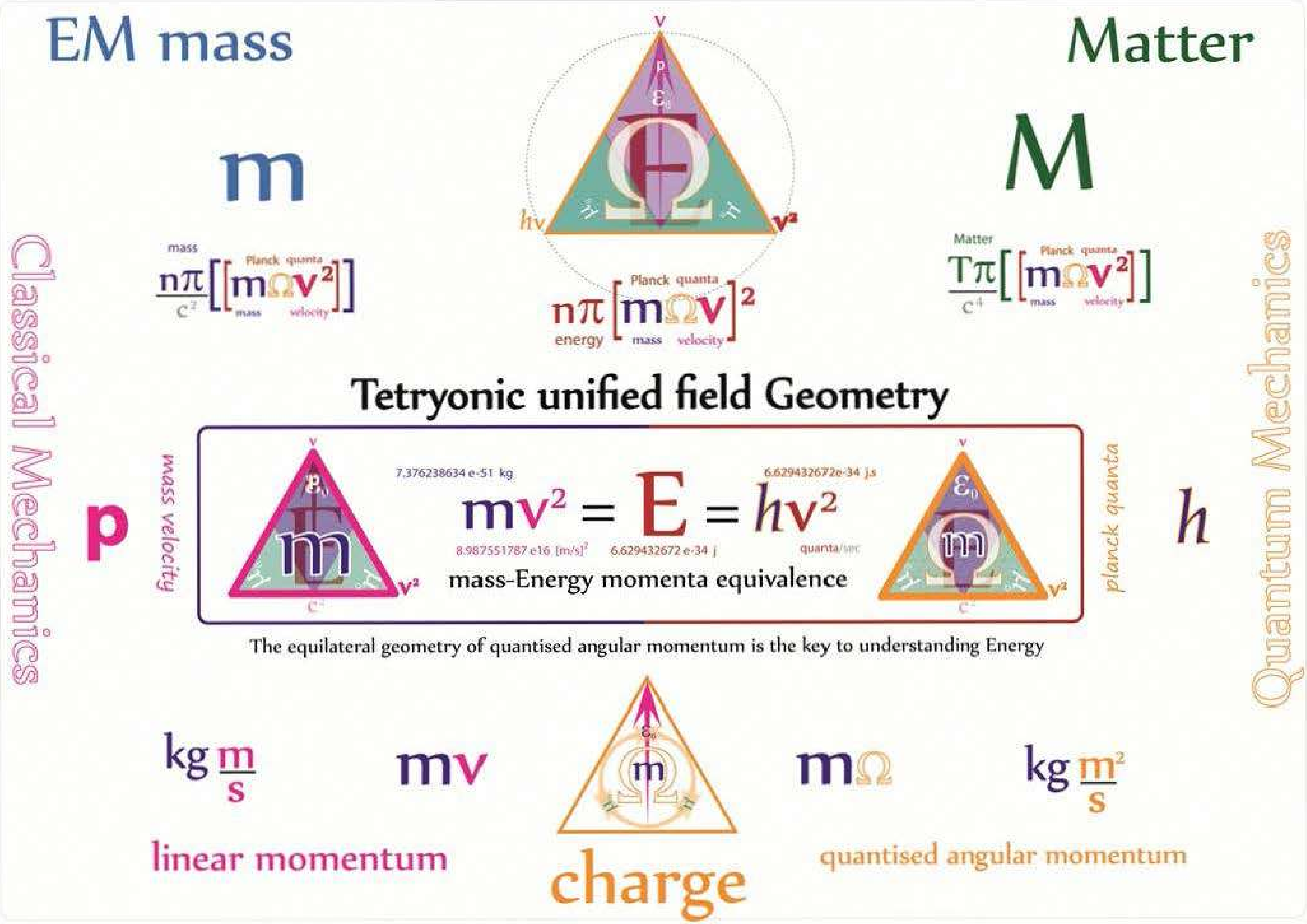

EM mass

m

mass

$\frac{n\pi}{c^2}[[m\Omega v^2]]$

Planck quanta

mass velocity

Matter

M

Matter

$\frac{T\pi}{c^4}[[m\Omega v^2]]$

Planck quanta

mass velocity

$n\pi[m\Omega v]^2$

energy Planck quanta mass velocity

Tetryonic unified field Geometry

Classical Mechanics

p

mass velocity

7.376238634 e-51 kg

6.629432672e-34 J.s

$$mv^2 = E = hv^2$$

8.987551787 e16 [m/s]²

6.629432672 e-34 J

quanta/sec

mass-Energy momenta equivalence

planck quanta

h

Quantum Mechanics

The equilateral geometry of quantised angular momentum is the key to understanding Energy

$kg\,\frac{m}{s}$

mv

$m\Omega$

$kg\,\frac{m^2}{s}$

linear momentum

charge

quantised angular momentum

$kg\,\frac{m^2}{s}\,\frac{1}{s}$

Planck quanta

$[m\Omega v^2]$

mass velocity

$kg\,\frac{m^2}{s^2}$

m

EM Field Planck quanta

$n\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$

mass ElectroMagnetic mass velocity

kg

ElectroMagnetic field

$[\varepsilon_0\mu_0]$

$\frac{1}{c^2}$

$\frac{s^2}{m^2}$

E

Energy

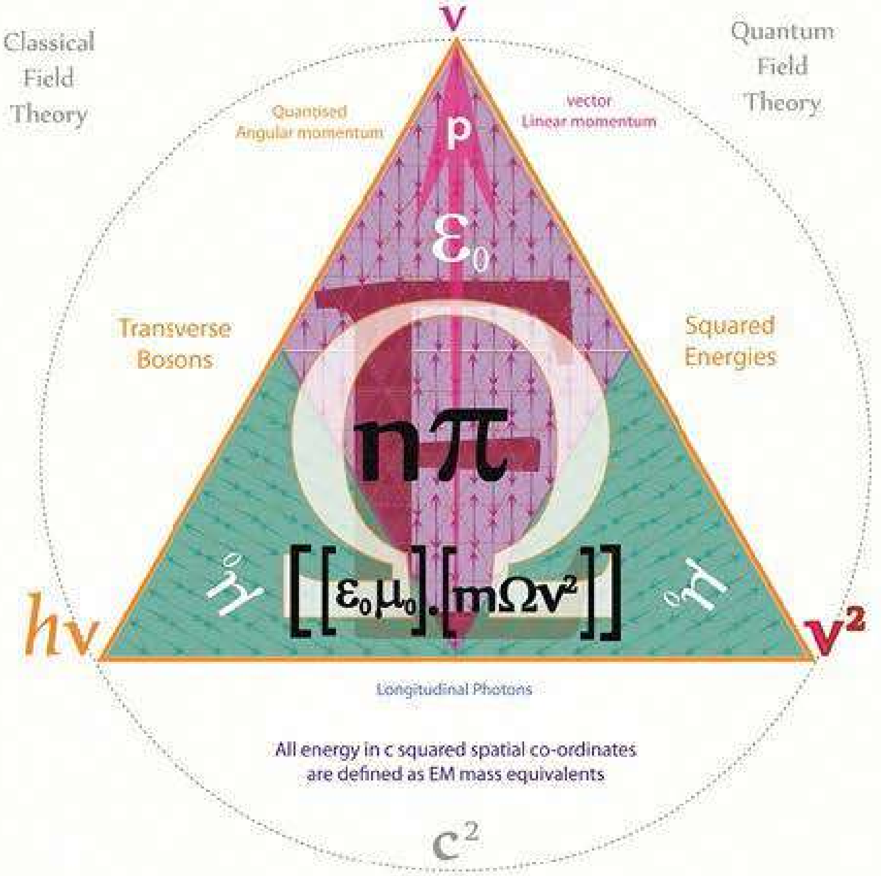
Classical
Field
Theory
Quantum
Field
Theory
v
p
ε_0
Quantised Angular momentum
vector Linear momentum
Transverse Bosons
Squared Energies
$n\pi$
$[[\varepsilon_0\mu_0].[m\Omega v^2]]$
hv
v^2
Longitudinal Photons
c^2

All energy in c squared spatial co-ordinates are defined as EM mass equivalents

Q

Charge

$\frac{s^2}{m^2}$ $\frac{m}{s^2}$

Charge

$[\varepsilon_0\mu_0\Omega]$

ElectroMagnetic QAM

s

h

$[m\Omega]$

mass QAM

$kg\,\frac{m^2}{s}$

QAM

$\frac{m^2}{s}$ Ω

Unified Field Components

The Unified field equation can be re-arranged to reveal a multitude of physical properties and relationships previously poorly understood.

Highlighting the fact that all of the constants and properties of both Classical and Quantum Mechanics are in fact geometric properties of Energy

v^2

velocity squared

Velocity - quanta

v^2

quanta/sec

Tetryonics 17.16 - UFE components

Copyright ABRAHAM [2008] - all rights reserved

EM mass Relationships

v	Velocity	$\frac{m}{s}$
a	Acceleration	$\frac{m}{s^2}$
m ↓		
p	Momentum	$kg\frac{m}{s}$
F	Force	$kg\frac{m}{s^2}$

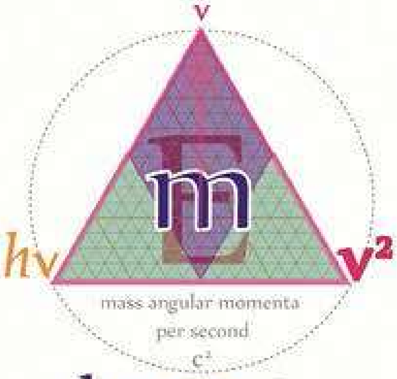

$$kg\ \frac{m^2}{s}\ \frac{1}{s}$$

EM mass is revealed to be the scalar property of 2D Energy waveforms that is at the core of many important physical processes and measurements

Ω	Quantum Angular Momentum	$\frac{m^2}{s}$
v^2	Celeritas squared	$\frac{m^2}{s^2}$
m ↓		
h	Planck's Constant	$kg\frac{m^2}{s}$
E	Energy	$kg\frac{m^2}{s^2}$

E/second

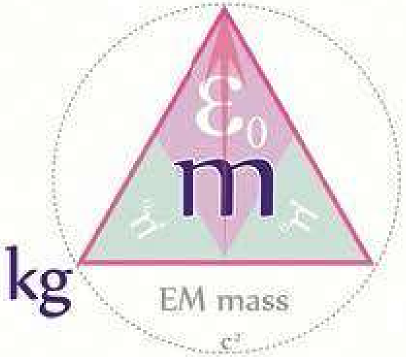

f	Frequency	$\frac{1}{s}$
T	Period	s
m ↓		
I	Current	$\frac{kg}{s}$
C	Charge	$\frac{s}{kg}$

E/second²

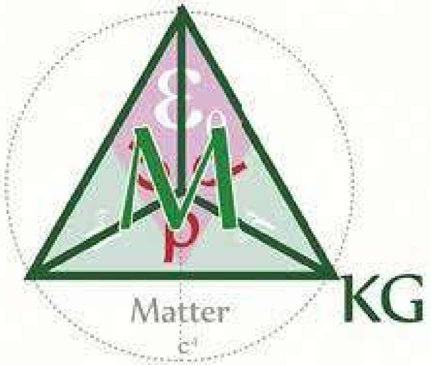

2D mass geometries should never be confused with 3D Matter topoplogies; nor should the terms be used in exchange for each other

Tetryonics 17.17 - EM mass relationships

Copyright ABRAHAM [2008] - all rights reserved

Relativistic mass-Matter

Rest Matter

$$T\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

mass-Matter

3D rest mass-Matter topolgy =
closed volume of 2D mass-energies

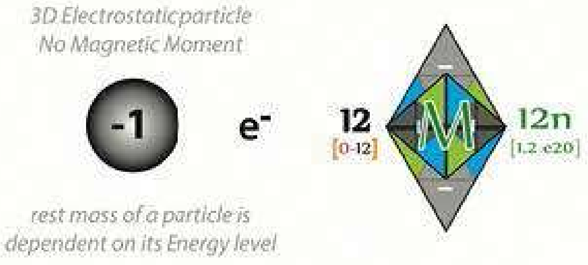

3D Electrostatic particle
No Magnetic Moment

rest mass of a particle is
dependent on its Energy level

All Matter is a Tetryonic
standing-wave charged geometry
occupying a volume in 3D spherical space

Matter

Planck quanta

$$\frac{T\pi}{c^4}\left[\underset{mass}{m}\Omega \underset{velocity}{v^2}\right]$$

↓

Dodecyon

Planck quanta

$$\frac{12\pi}{c^4}\left[\underset{mass}{m}\Omega \underset{velocity}{v^2}\right]$$

rest mass-Matter is velocity invariant
(not subject to Lorentz corrections)

The property of Matter cannot be measured using
a planar [c squared] spatial co-ordinate system

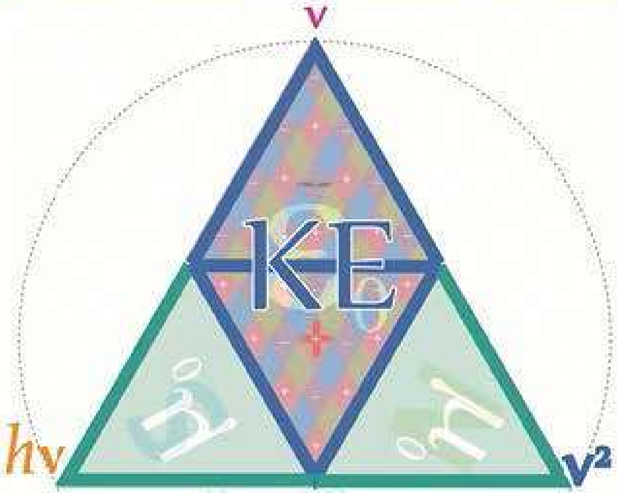

Magnetic Moment

$$\frac{KE+\mu_B}{c^2}$$

A KEM field has Electric and Magnetic quanta
whose total energy quanta is directly related
to the square of the particle's velocity

$$4\pi\left[\frac{\overset{mass}{m}\Omega \overset{velocity}{v^2}}{c^2}\right]$$

All KEM fields are subject to
Lorentz corrections

Kinetic Energies

$$4n\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

KEM field

Matter-Energy

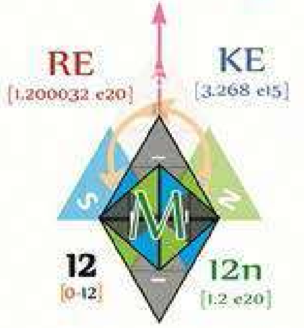

electron KEM model

Relativistic mass Energy =
rest Matter + Kinetic EM Energies

Electron rest Matter

Lepton

Planck quanta

$$\frac{12\pi}{c^4}\left[\underset{mass}{m}\Omega \underset{velocity}{v^2}\right]$$

$$+$$

KEM wave

Planck quanta

$$\frac{4\pi}{c^2}\left[\underset{mass}{m}\Omega \underset{velocity}{v^2}\right]$$

Kinetic Energies

KEM mass-energies are velocity dependent
(subject to Lorentz corrections)

Tetryonics 17.18 - Relativistic mass-Matter

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic mass & Matter

Historically interchanged due to the lack of proper definitions the physical properties of EM mass & Matter can now be firmly defined with respect to their energy equivalence and spatial geometries

m

mass is a measure of the 2D planar energy content of any physical system

mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

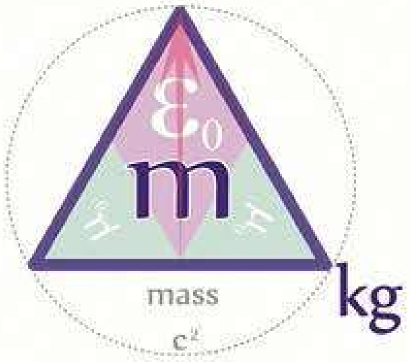

E/second

2D
radiant equilateral geometry
of EM mass-ENERGY momenta
EM mass

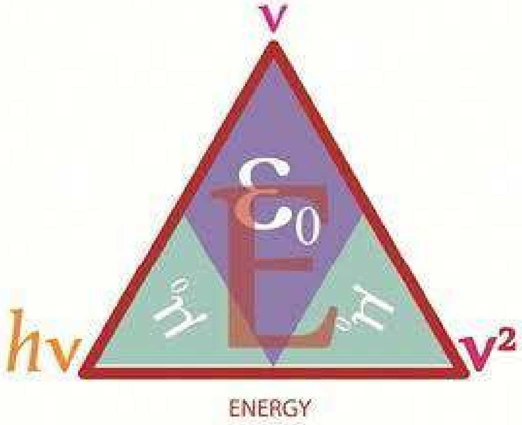

EM mass should replace the generic term mass with reference to ElectroMagnetic energy densities

MATTER is a geometric 4nπ standing wave topology of EM mass-energy geometries

The electromagnetic energies of rest Matter is never 'at rest' as the electromagnetic field energies creating mass-Matter topologies alway propagate at c

M

Matter is a measure of the 3D volumetric energy content of any physical system

Matter Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

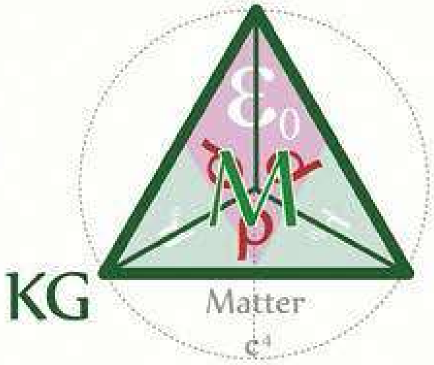

E/second²

3D
standing-wave topology
of EM mass-ENERGY momenta
rest Matter

Tetryonics 17.19 - Tetryonic mass & Matter

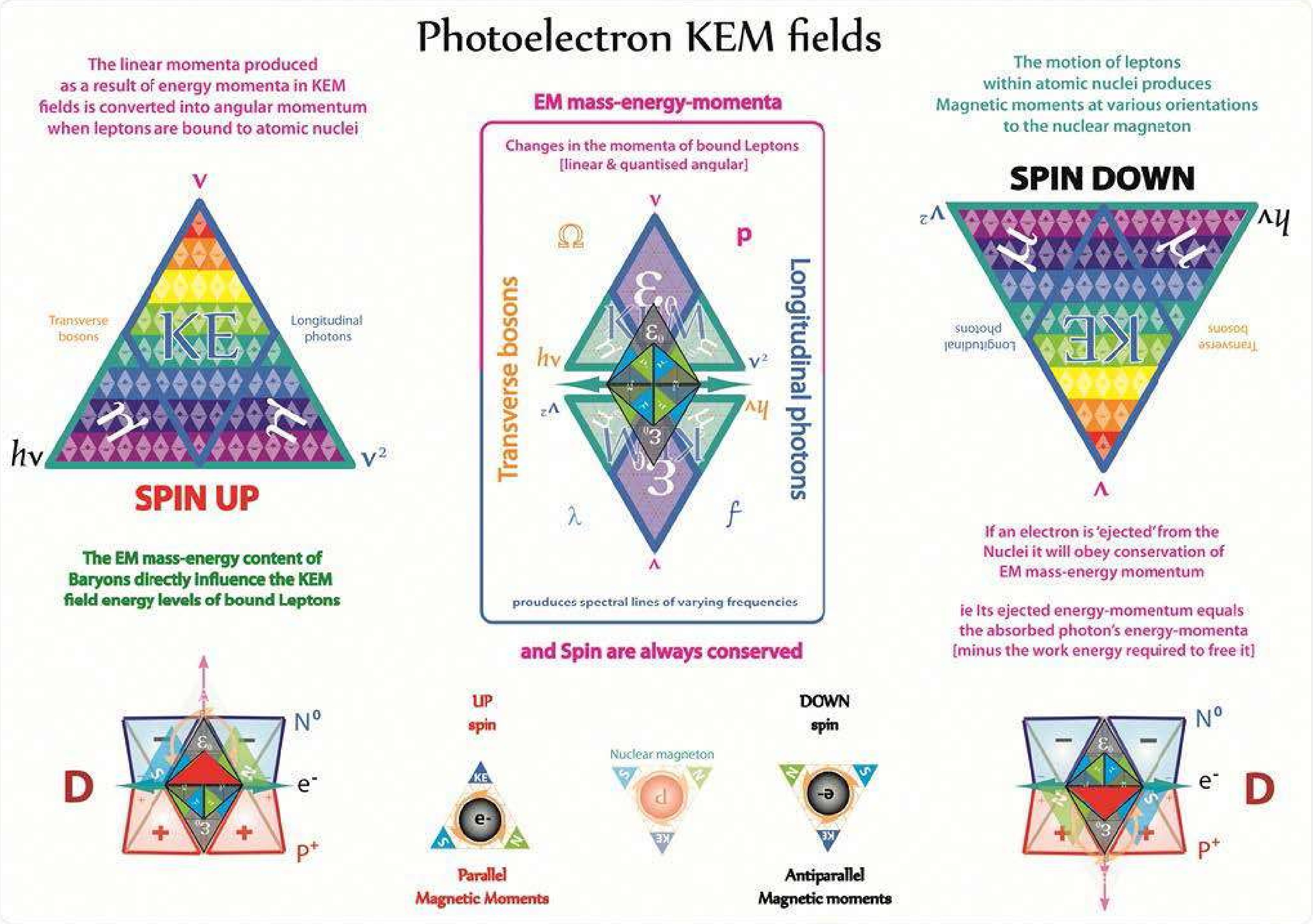
v
KE
Transverse bosons
Longitudinal photons
hν
v^2
μ
SPIN UP
Ω
p
ε
$ε_0$
$μ_0$
KEM
Transverse bosons
Longitudinal photons
λ
f
SPIN DOWN
D
N^0
e^-
P^+
UP
spin
e-
Parallel
Magnetic Moments
Nuclear magneton
p
DOWN
spin
e-
Antiparallel
Magnetic moments

Photoelectron KEM fields

The linear momenta produced as a result of energy momenta in KEM fields is converted into angular momentum when leptons are bound to atomic nuclei

The EM mass-energy content of Baryons directly influence the KEM field energy levels of bound Leptons

EM mass-energy-momenta

Changes in the momenta of bound Leptons [linear & quantised angular]

prouduces spectral lines of varying frequencies

and Spin are always conserved

The motion of leptons within atomic nuclei produces Magnetic moments at various orientations to the nuclear magneton

If an electron is 'ejected' from the Nuclei it will obey conservation of EM mass-energy momentum

ie Its ejected energy-momentum equals the absorbed photon's energy-momenta [minus the work energy required to free it]

Tetryonics 18.01- Photoelectron KEM fields

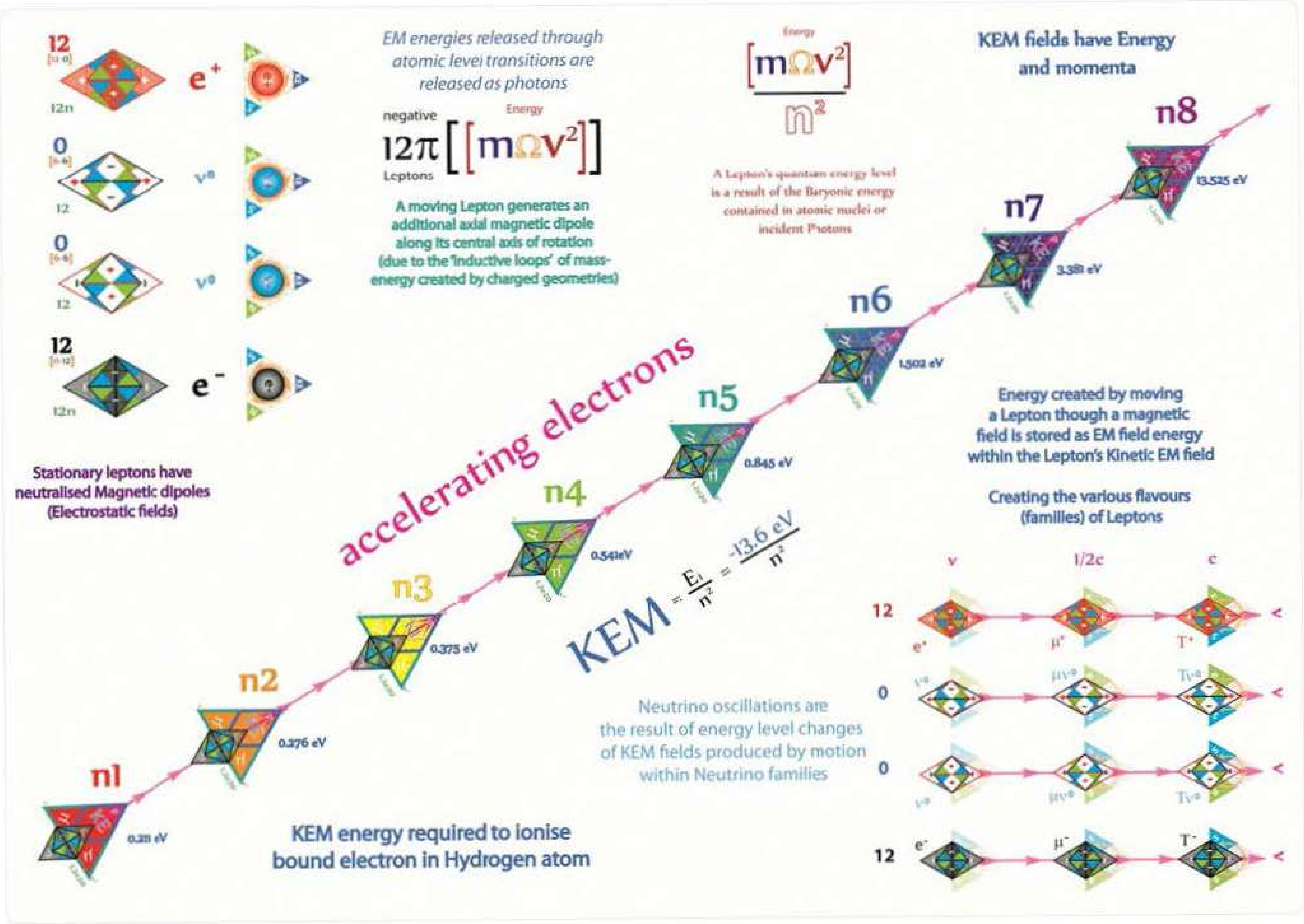

EM energies released through atomic level transitions are released as photons

negative Energy

$$12\pi\left[\left[m\Omega v^2\right]\right]$$

Leptons

A moving Lepton generates an additional axial magnetic dipole along its central axis of rotation (due to the 'inductive loops' of mass-energy created by charged geometries)

Energy

$$\frac{\left[m\Omega v^2\right]}{n^2}$$

A Lepton's quantum energy level is a result of the Baryonic energy contained in atomic nuclei or incident Photons

KEM fields have Energy and momenta

Stationary leptons have neutralised Magnetic dipoles (Electrostatic fields)

accelerating electrons

$$KEM = \frac{E_1}{n^2} = \frac{-13.6\ eV}{n^2}$$

Energy created by moving a Lepton though a magnetic field is stored as EM field energy within the Lepton's Kinetic EM field

Creating the various flavours (families) of Leptons

Neutrino oscillations are the result of energy level changes of KEM fields produced by motion within Neutrino families

KEM energy required to ionise bound electron in Hydrogen atom

12
12n
e+
0
12
v0
0
12
v0
12
12n
e-
n1
0.211 eV
n2
0.276 eV
n3
0.375 eV
n4
0.541eV
n5
0.845 eV
n6
1.502 eV
n7
3.381 eV
n8
13.525 eV
v
1/2c
c
12
e+
μ+
T+
0
v0
μv0
Tv0
0
v0
μv0
Tv0
12
e-
μ-
T-

Tetryonics 18.02 - accelerating electrons

Copyright ABRAHAM [2008] - all rights reserved

Leptronic Oscillations

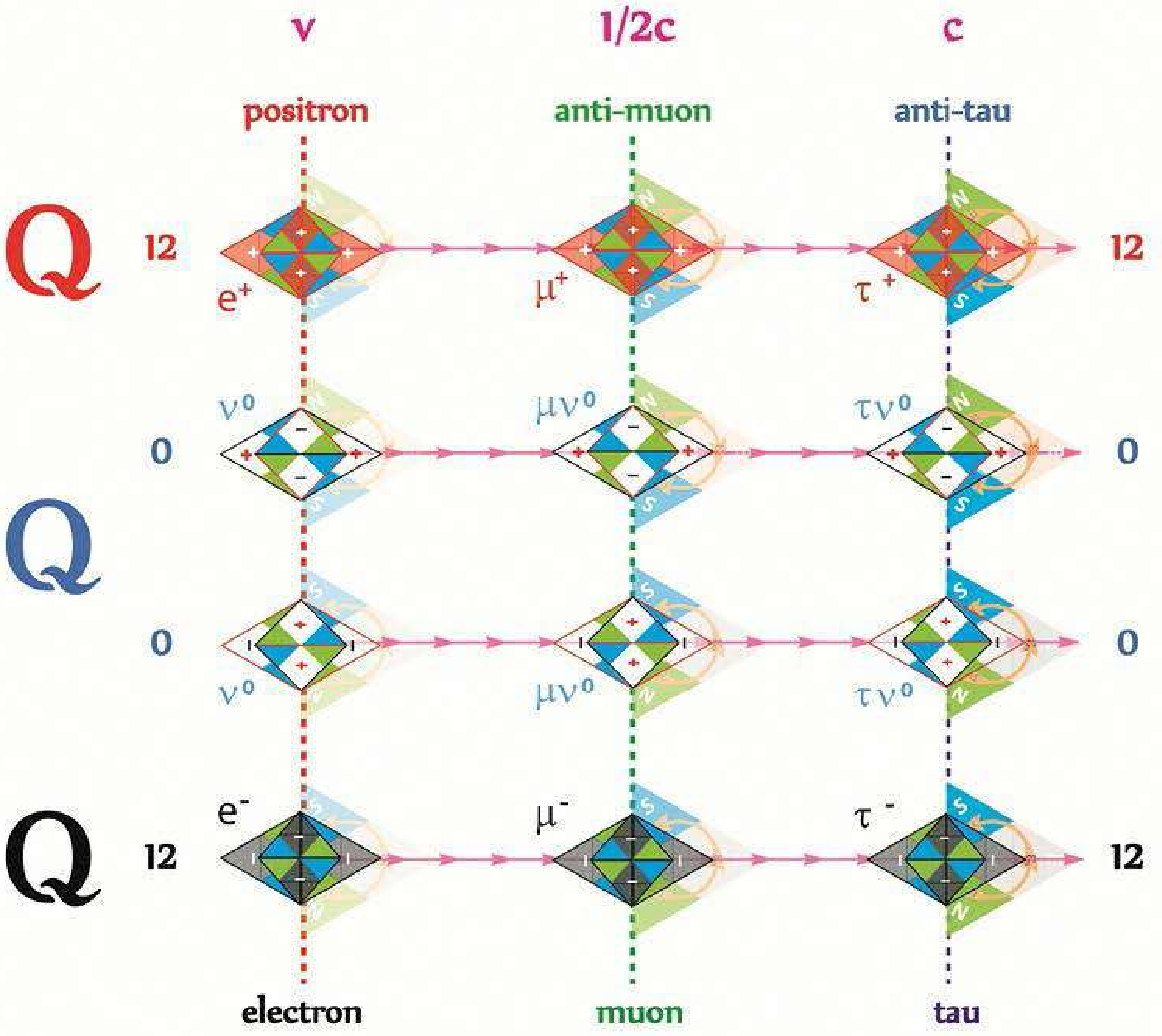
v
1/2c
c
positron
anti-muon
anti-tau
Q
12
e+
μ+
τ +
12
ν0
μν0
τν0
Q
0
0
0
0
ν0
μν0
τν0
e-
μ-
τ -
Q
12
12
N
S
electron
muon
tau

Antimatter

EM Field · Planck · quanta

$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

Leptons · ElectroMagnetic · mass · velocity

All Leptronic generations,
oscillations and types can be
accounted for through
Tetryonic geometries
& energy levels

Matter

Tetryonics 18.03 - Leptronic oscillations

Principal Quantum levels

$$Mv^2 = \text{KEM} = hv^2$$

Bound electron Eigenstate values

Eigenstate bound energy states

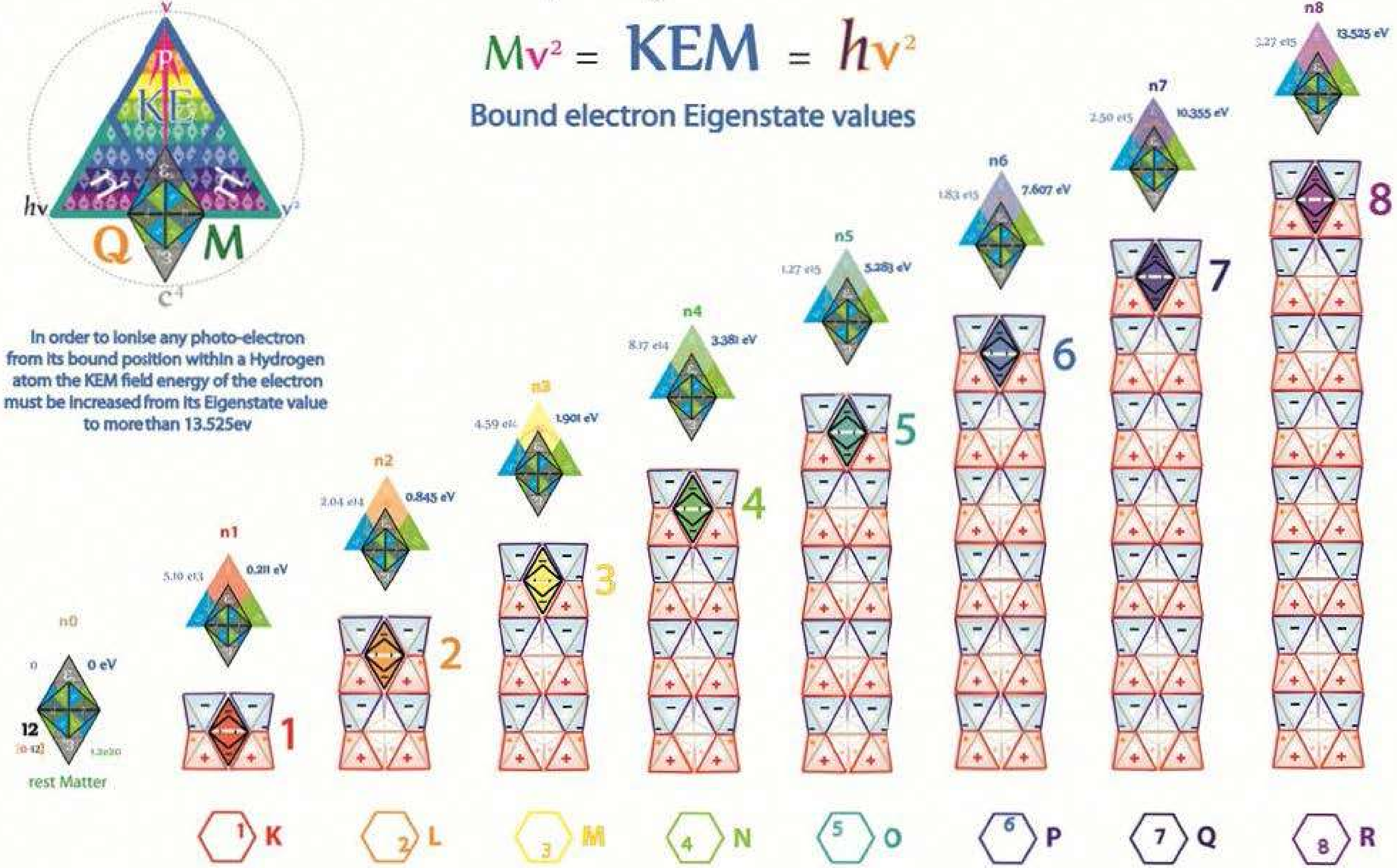
v
P
KE
hv
v^2
Q
M
c^4
n0
0
0 eV
12
[0-12]
1.2e20
rest Matter
n1
5.10 e13
0.211 eV
1
n2
2.04 e14
0.845 eV
2
n3
4.59 e14
1.901 eV
3
n4
8.17 e14
3.381 eV
4
n5
1.27 e15
5.283 eV
5
n6
1.83 e15
7.607 eV
6
n7
2.50 e15
10.355 eV
7
n8
3.27 e15
13.525 eV
8
1 K
2 L
3 M
4 N
5 O
6 P
7 Q
8 R

In order to ionise any photo-electron from its bound position within a Hydrogen atom the KEM field energy of the electron must be increased from its Eigenstate value to more than 13.525ev

The mass-energy content of Baryons comprising atomic nuclei directly determines the Kinetic energy levels of bound Leptons [forming quantum synchronous convertors]
(incident Bosons/Photons can increase these levels if they contain the exact quanta required to increase the level to the next square number)

Tetryonics 18.04 - Principal Quantum Levels

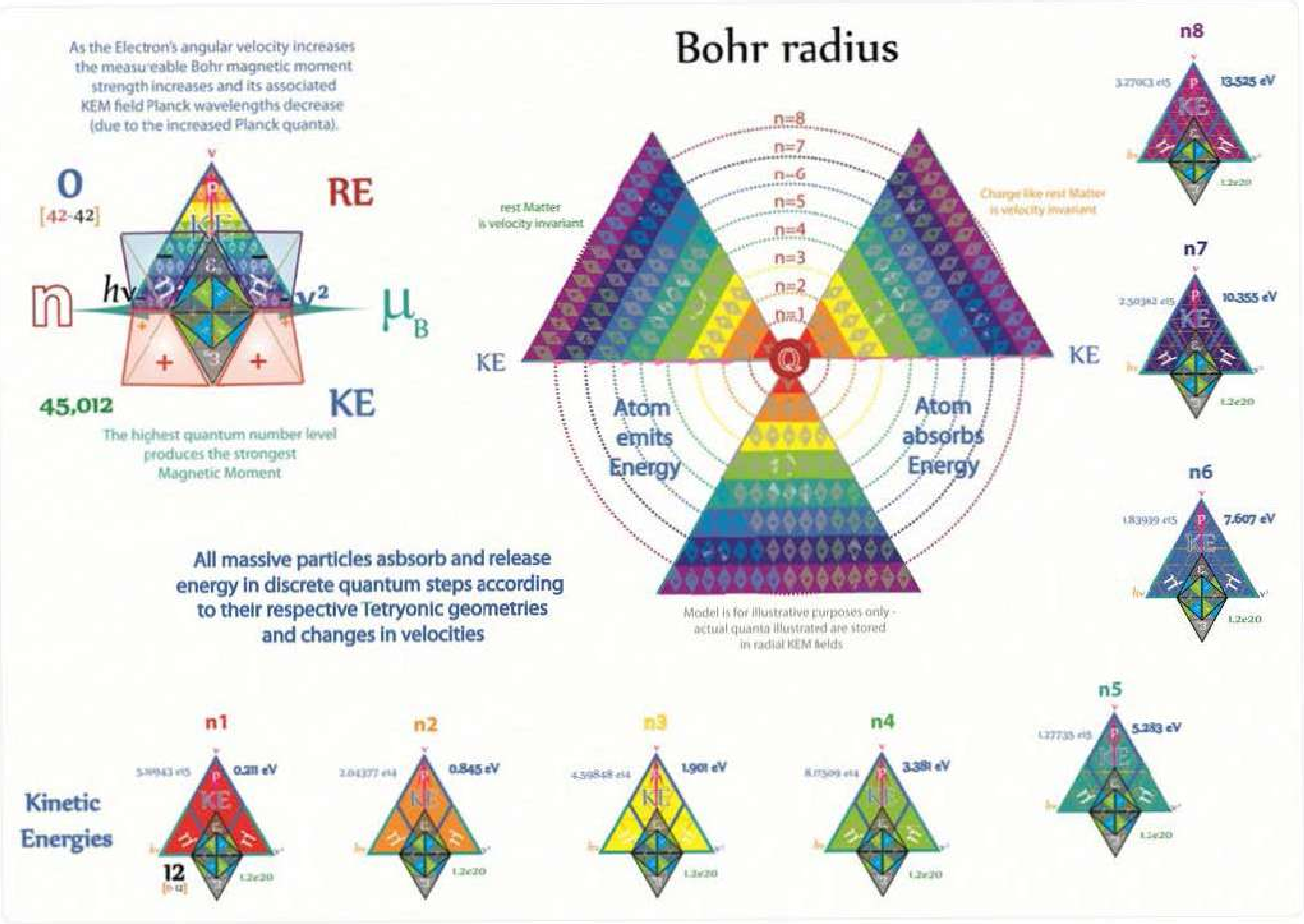

Bohr radius

As the Electron's angular velocity increases the measureable Bohr magnetic moment strength increases and its associated KEM field Planck wavelengths decrease (due to the increased Planck quanta).

The highest quantum number level produces the strongest Magnetic Moment

All massive particles asbsorb and release energy in discrete quantum steps according to their respective Tetryonic geometries and changes in velocities

Model is for illustrative purposes only - actual quanta illustrated are stored in radial KEM fields

O
[42-42]
RE
n
hv
v²
μB
P
KE
+
+
45,012
KE
rest Matter is velocity invariant
Charge like rest Matter is velocity invariant
n=8
n=7
n=6
n=5
n=4
n=3
n=2
n=1
KE
KE
Q
Atom emits Energy
Atom absorbs Energy
n8
3.27063 e15
13.525 eV
n7
2.50382 e15
10.355 eV
n6
1.83939 e15
7.607 eV
n5
1.27735 e15
5.283 eV
Kinetic Energies
n1
5.10943 e13
0.211 eV
12
[n-12]
n2
2.04377 e14
0.845 eV
n3
4.59848 e14
1.901 eV
n4
8.17509 e14
3.381 eV
1.2e20

Tetryonics 18.05 - Bohr radius

Nuclear Magnetons

are weaker than Bohr magnetons
due to differing mass-charge ratios
and non-parallel Quark magnetons

Baryons

are tri-Dodecyon geometries

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

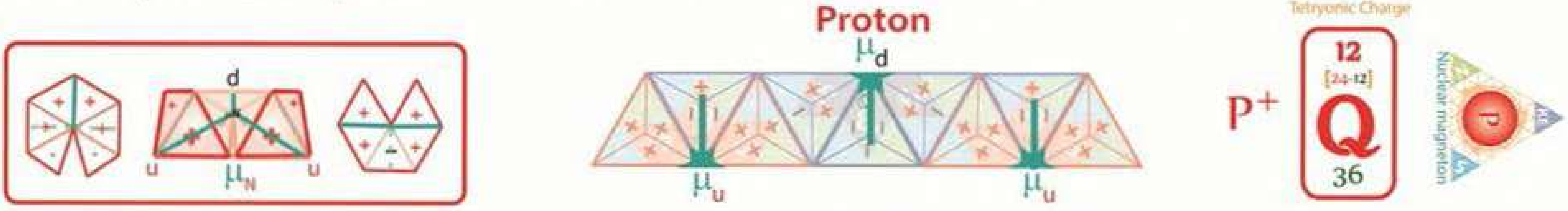
d
u
μ_N
u
Proton
μ_d
μ_u
μ_u
Tetryonic Charge
p^+
12
[24-12]
Q
36
Nuclear magneton
p

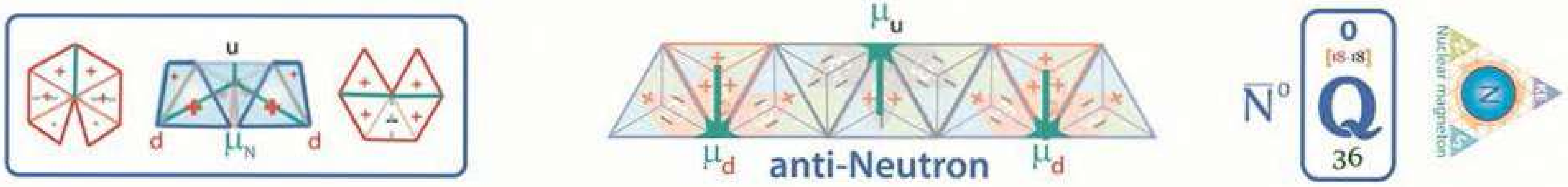
u
d
μ_N
d
μ_u
μ_d
anti-Neutron
μ_d
$\bar{N}^0$
0
[18-18]
Q
36
Nuclear magneton
N

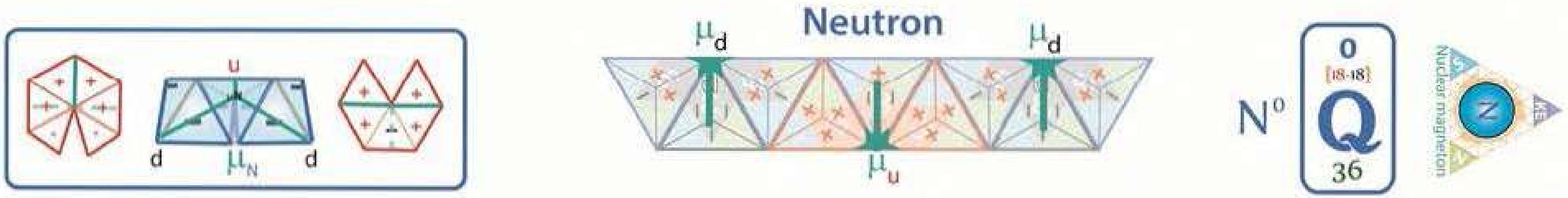
u
d
μ_N
d
μ_d
Neutron
μ_d
μ_u
N^0
0
[18-18]
Q
36
Nuclear magneton
N

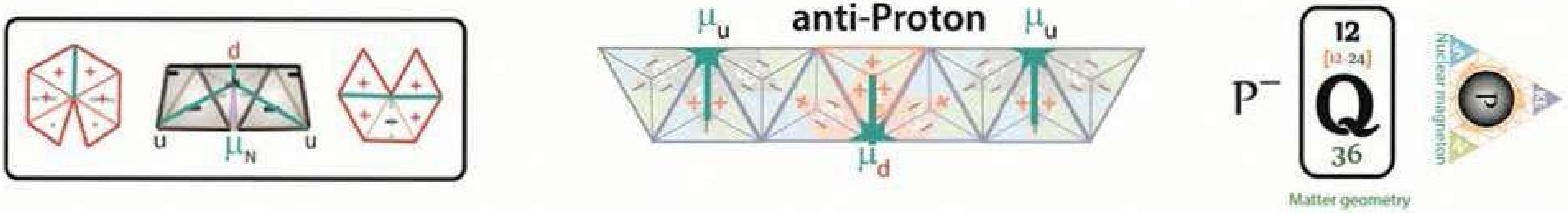
d
u
μ_N
u
μ_u
anti-Proton
μ_u
μ_d
p^-
12
[12-24]
Q
36
Nuclear magneton
p
Matter geometry

Baryonic Magnetic Moments

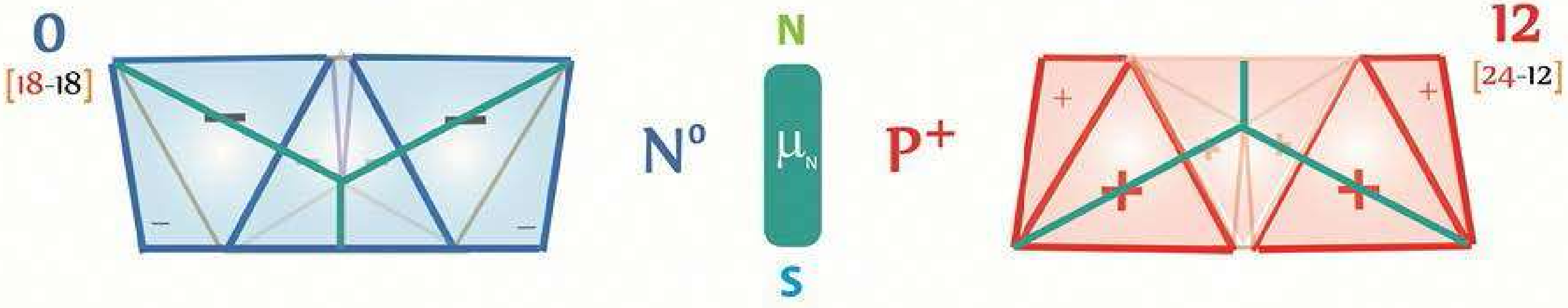
0
[18-18]
N°
N
μN
S
P+
12
[24-12]

A Baryon's Magnetic Moment
is a result of the combined non-parallel
Magnetic moments of its Quarks

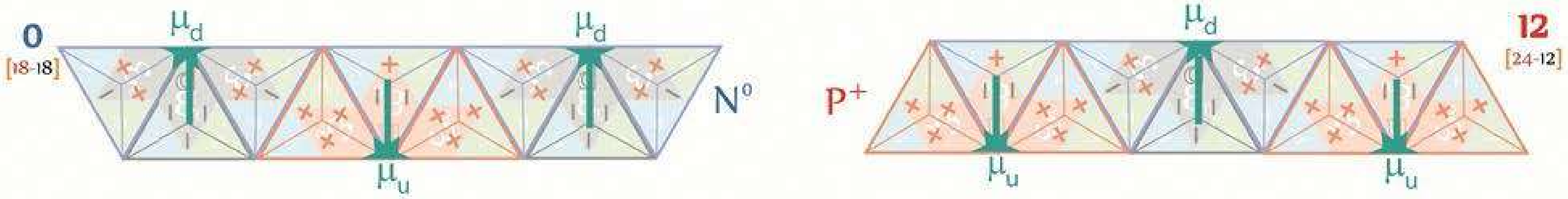
0
[18-18]
μd
μd
μu
N°
P+
μd
12
[24-12]
μu
μu

As a result of the 3 non-aligned Quark magnetic moments
resulting from their Tetryonic geometry,
the Nuclear Magneton is considerably weaker than the Bohr Magneton

Nuclear Quantum Energies

[Principal Quantum Numbers]

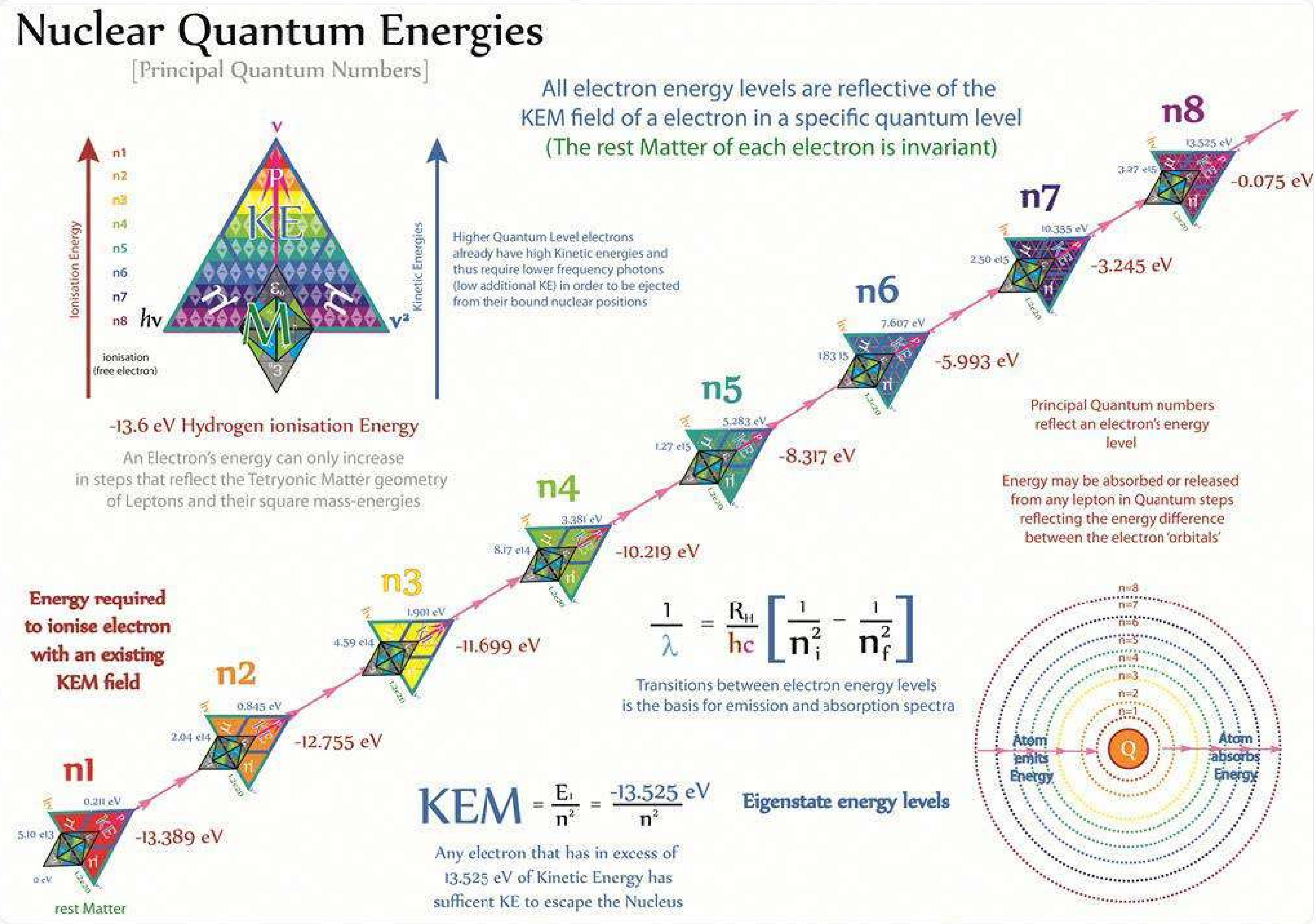
v
n1
n2
n3
n4
n5
n6
n7
n8
Ionisation Energy
ionisation
(free electron)
P
KE
M
hv
v²
Kinetic Energies
n1
0.211 eV
5.10 e13
0 eV
-13.389 eV
rest Matter
n2
0.845 eV
2.04 e14
-12.755 eV
n3
1.901 eV
4.59 e14
-11.699 eV
n4
3.381 eV
8.17 e14
-10.219 eV
n5
5.283 eV
1.27 e15
-8.317 eV
n6
7.607 eV
183 15
-5.993 eV
n7
10.355 eV
2.50 e15
-3.245 eV
n8
13.525 eV
3.37 e15
-0.075 eV
n=8
n=7
n=6
n=5
n=4
n=3
n=2
n=1
Q
Atom emits Energy
Atom absorbs Energy

All electron energy levels are reflective of the KEM field of a electron in a specific quantum level (The rest Matter of each electron is invariant)

Higher Quantum Level electrons already have high Kinetic energies and thus require lower frequency photons (low additional KE) in order to be ejected from their bound nuclear positions

-13.6 eV Hydrogen ionisation Energy

An Electron's energy can only increase in steps that reflect the Tetryonic Matter geometry of Leptons and their square mass-energies

Principal Quantum numbers reflect an electron's energy level

Energy may be absorbed or released from any lepton in Quantum steps reflecting the energy difference between the electron 'orbitals'

Energy required to ionise electron with an existing KEM field

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left[\frac{1}{n_i^2} - \frac{1}{n_f^2}\right]$$

Transitions between electron energy levels is the basis for emission and absorption spectra

$$KEM = \frac{E_i}{n^2} = \frac{-13.525\ \text{eV}}{n^2}$$

Eigenstate energy levels

Any electron that has in excess of 13.525 eV of Kinetic Energy has sufficent KE to escape the Nucleus

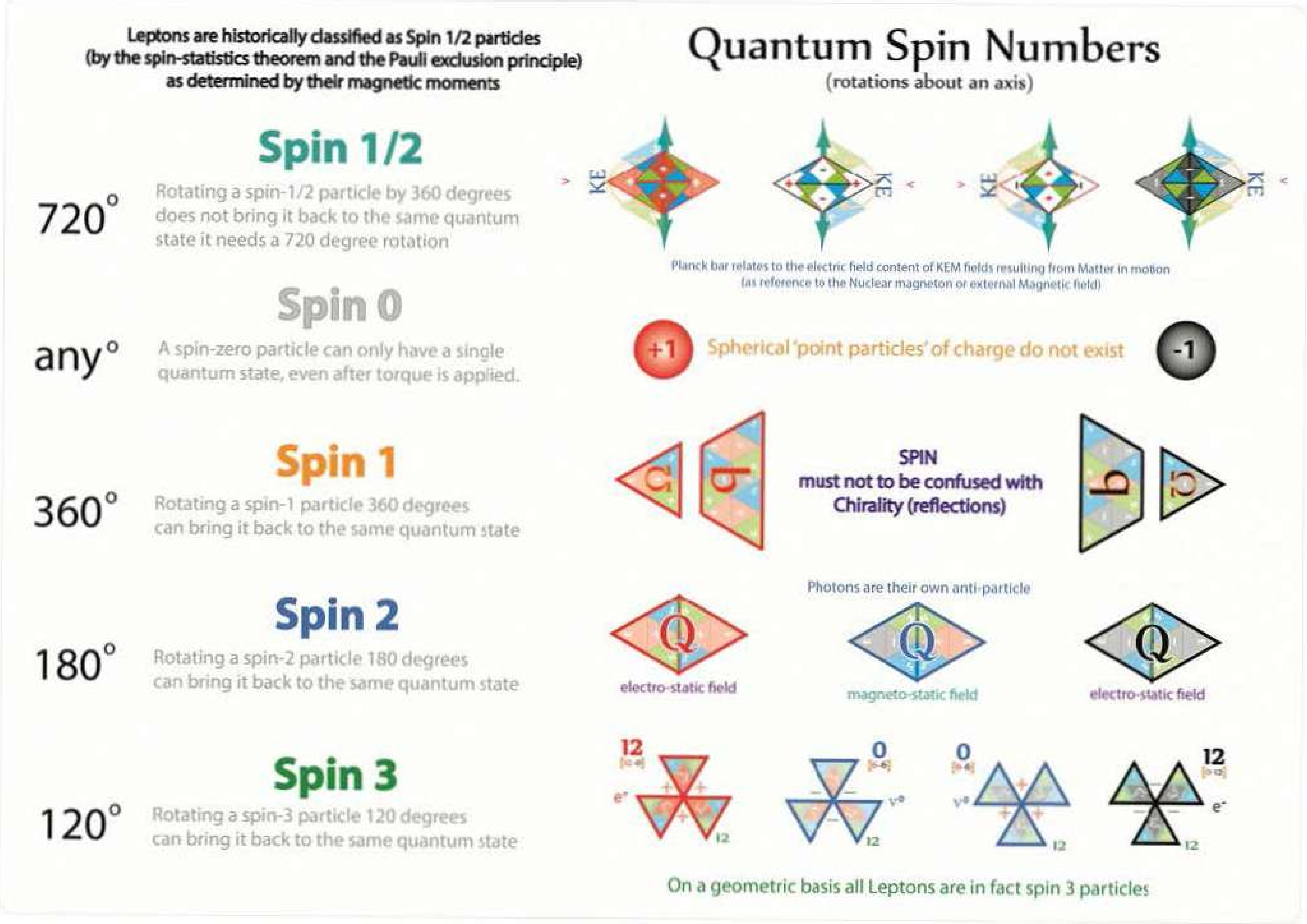

**Leptons are historically classified as Spin 1/2 particles
(by the spin-statistics theorem and the Pauli exclusion principle)
as determined by their magnetic moments**

Quantum Spin Numbers

(rotations about an axis)

Spin 1/2

720°

Rotating a spin-1/2 particle by 360 degrees does not bring it back to the same quantum state it needs a 720 degree rotation

Planck bar relates to the electric field content of KEM fields resulting from Matter in motion
(as reference to the Nuclear magneton or external Magnetic field)

Spin 0

any°

A spin-zero particle can only have a single quantum state, even after torque is applied.

Spherical 'point particles' of charge do not exist

Spin 1

360°

Rotating a spin-1 particle 360 degrees can bring it back to the same quantum state

**SPIN
must not to be confused with
Chirality (reflections)**

Spin 2

180°

Rotating a spin-2 particle 180 degrees can bring it back to the same quantum state

Photons are their own anti-particle

Spin 3

120°

Rotating a spin-3 particle 120 degrees can bring it back to the same quantum state

On a geometric basis all Leptons are in fact spin 3 particles

Tetryonics 19.01 - Quantum spin numbers

Magnetic moments are determined by the KEM fields created by vector linear momentum

Nuclear Magnetic moments are complicated by the tri-quark magnetons within all Matter topologies

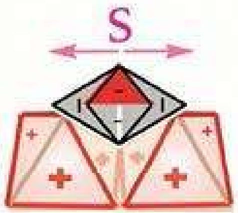

The KEM field of any charged particle in motion is reflective of the particle's nett charge topology

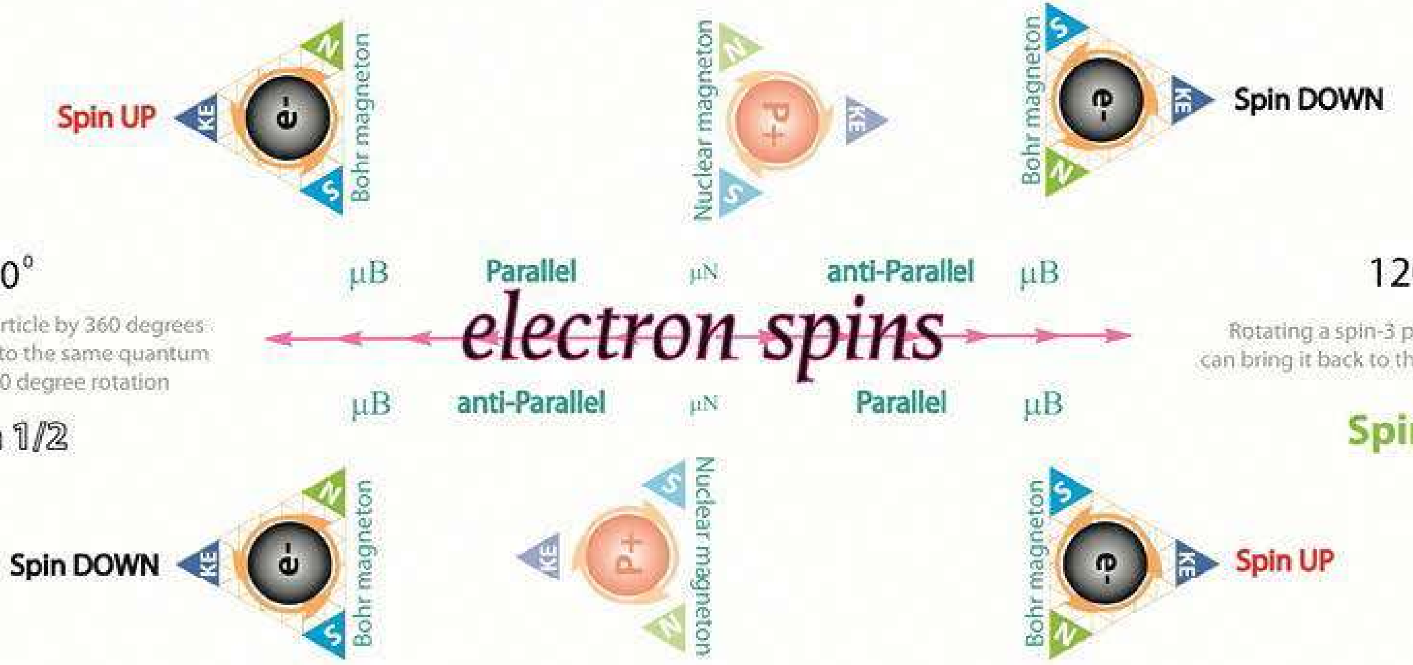
Spin UP
KE
e-
N
S
Bohr magneton
Nuclear magneton
P+
KE
N
S
Bohr magneton
e-
KE
Spin DOWN
µB
Parallel
µN
anti-Parallel
µB
electron spins
µB
anti-Parallel
µN
Parallel
µB
Spin DOWN
KE
e-
Bohr magneton
KE
P+
Nuclear magneton
Bohr magneton
e-
KE
Spin UP

720°

Rotating a spin-1/2 particle by 360 degrees does not bring it back to the same quantum state it needs a 720 degree rotation

Spin 1/2

120°

Rotating a spin-3 particle 120 degrees can bring it back to the same quantum state

Spin 3

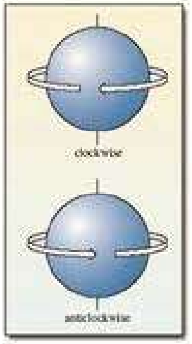
clockwise
anticlockwise

All atomic particles have a particular "spin" analogous to the Earth's rotation on its axis.

$$\mu_B = \left[\frac{Q}{2mv}\right]/\ell$$

An isolated electron has an angular momentum and a magnetic moment resulting from its spin. While an electron's spin is sometimes visualized as a literal rotation about an axis, it is in fact a fundamentally different, quantum-mechanical phenomenon with no true analogue in classical physics.

The quantum mechanical reality underlying spin is complex and still poorly understood Consequently, there is no reason to expect the above classical relation to hold.

$$\mu_B = g\left[\frac{e-}{2mv}\right]S$$

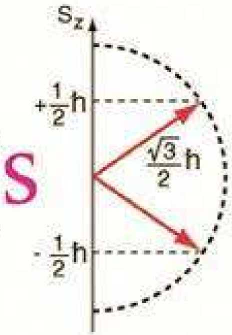
S_z
$+\frac{1}{2}\hbar$
$\frac{\sqrt{3}}{2}\hbar$
$-\frac{1}{2}\hbar$

Tetryonic KEM field geometries reveals the source and orientation of all atomic magnetic moments

Generating Magnetons

A static Electron has a negative Tetryonic charge [0-12] topology with neutralised magnetic dipoles

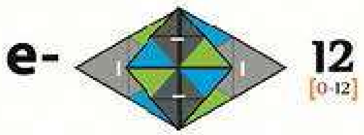

Leptons are
12 loop inductive
charge rotors

Tetryonic geometry
fully explains Leptronic 'spins'

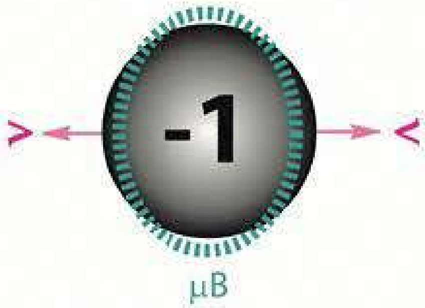

Einstein's Special Relativity model of distorted moving charges producing magnetic moments is incorrect

*BOHR Magneton
produced by Lorentzian
distortion of charges due to
relativistic velocities*

$$\mu_B = \frac{e\hbar}{2m_e}$$

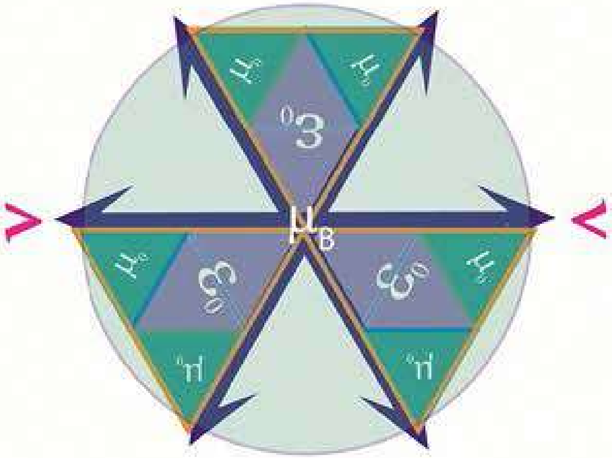

***The term "electron spin" can now be taken literally
(when modelled with Tetryonic geometries)
as an accurate description of the origin of Magnetic moments
for all Leptonic [BOHR] magnetons.***

***The previously held model of the electron as
a spinning sphere of charge must be abandoned
in favour of the true Tetryonic charge geometries
of EM mass-Energy-Matter***

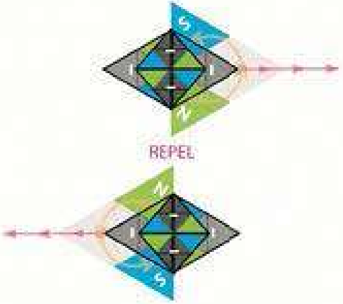

Ampere Force

$$\mathbf{B} = \frac{\mu_0 I}{4\pi}\int \frac{d\boldsymbol{\ell} \times \hat{\mathbf{r}}}{r^2},$$

$$F = 2k_A \frac{I_1 I_2}{r}$$

$$\mathbf{F} = q[\mathbf{E} + (\mathbf{v} \times \mathbf{B})].$$

Lorentz Force

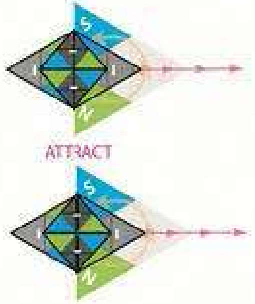

A moving Electron has a KEM field with an Electric field and a Magnetic Moment

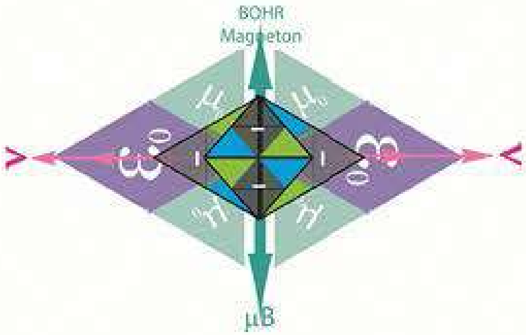

The gyromagnetic ratio of a particle or system is the ratio of its magnetic dipole moment to its angular momentum

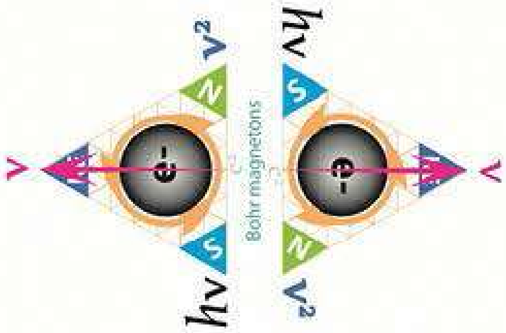

The KEM field energy of an Electron in motion is subject to relativistic corrections due to energy changes resulting from its acceleration

$$\lambda = L'\sqrt{1 - \frac{v^2}{c^2}}$$

WAVE-length contraction
of mass-energy quanta of KEM field

Stern Gerlach Experiment

In 1922 O. Stern and W. Gerlach measured the intrinsic spin angular momentum
of silver atoms and found it to take only two discrete values.
+ h / 2 and - h / 2 commonly called "spin up" and "spin down

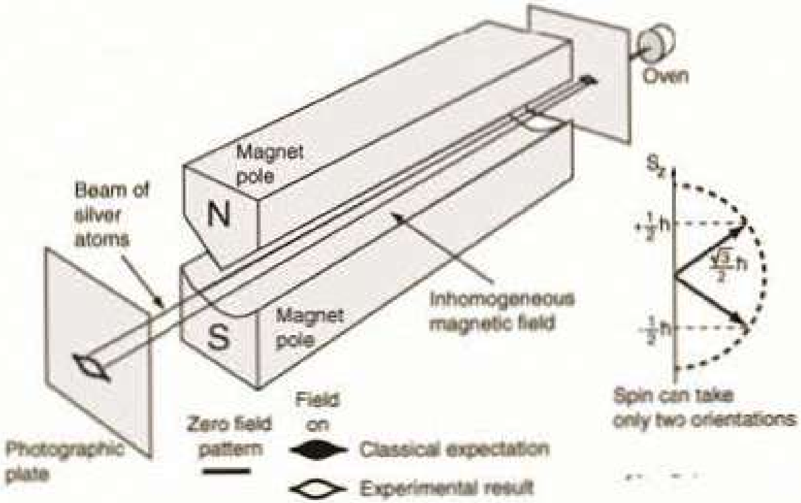
Oven
Magnet pole
N
Beam of silver atoms
S
Magnet pole
Inhomogeneous magnetic field
Photographic plate
Zero field pattern
Field on
Classical expectation
Experimental result
S_z
$+\frac{1}{2}h$
$\frac{\sqrt{3}}{2}h$
$-\frac{1}{2}h$
Spin can take only two orientations

The Stern-Gerlach experiment to determine electron SPIN

The results were interpreted to show that particles
possess an intrinsic angular momentum that is most
closely analogous to the angular momentum of a
classically spinning charged sphere, that takes
on only certain quantized values of angular momentum

Tetryonics shows the results are a product of
the KEM field produced by charge particles in motion,
with all charged particles being able to produce 2 distinct
magneton orientations as a result of the real motion of
their intrinsic quantum inductive loop [Matter]geometries
through external E fields & as referenced to external M moments.

This along with the Lorentz force produced by an external
Magnetic field acting on the KEM magnetons produces
the two results obtained

Any increase in electron energies results in
an increase in quantised angular momenta
and its vector linear momentum

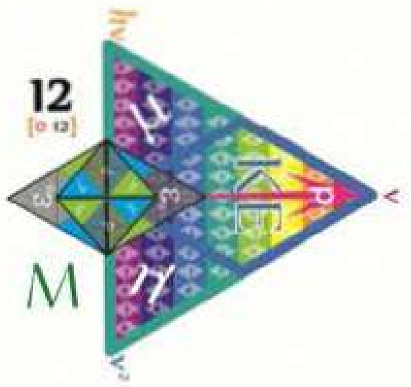
hv
12
[o 12]
M
KE
p
v
v^2

Negative charged in motion produce
negative charge KEM field geometries

Any change in the KEM energy level
of an electron results in Photon
emission or absorption

Electron Spins based on the measured Bohr Magnetons of moving electrons are reflected with KEM field geometries

**Electrons
bound in atomic nuclei
increase of decrease their
energy levels dependent
on the energy level of the nuclei
in which they are bound**

**Bound and unbound electrons
can rotate in one of two
directions influenced only by
external magnetic and electric fields
or incident Photons**

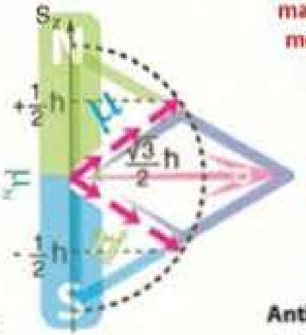
S_z
N
$+\frac{1}{2}h$
$\frac{\sqrt{3}}{2}h$
$-\frac{1}{2}h$
S

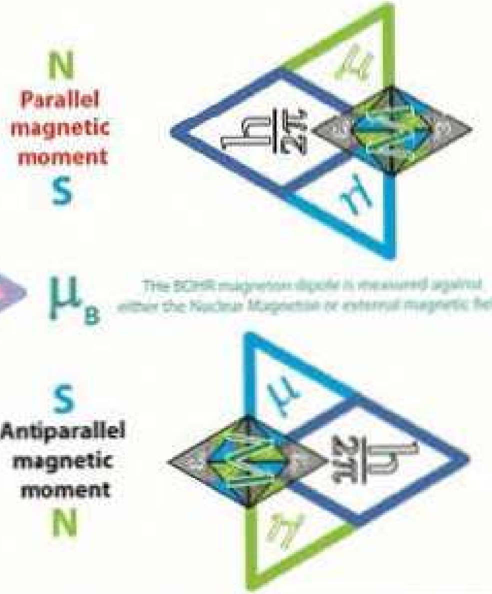
N
Parallel magnetic moment
S
μ_B
S
Antiparallel magnetic moment
N
$\frac{h}{2\pi}$
$\frac{h}{2\pi}$

The BOHR magneton dipole is measured against either the Nuclear Magnetons or external magnetic fields

The two separated beams of electrons produced
are defined as having differing SPINS
[UP or **DOWN**]

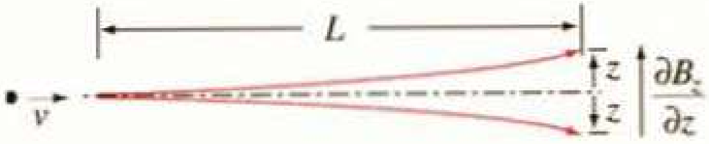
L
v
z
z
$\frac{\partial B_z}{\partial z}$

$$z = \frac{1}{2}at^2 = \frac{1}{2}\frac{F}{m}\left[\frac{L}{v}\right]^2 = \pm\frac{\mu_B L^2}{4KE}\frac{\partial B_z}{\partial z}$$

The orientation of the BOHR magneton wrt the NUCLEAR magneton
determines the 'direction' of electron SPIN

Tetryonics 19.04 - Stern Gerlach experiment

Bohr Magnetons

A moving electron is a 12 loop rotating inductor

Energy created by moving a Lepton though an external EM field is stored as Planck quanta within the Lepton's extended KEM field

Leptronic 'spin' is always determined by the Leptronic Magnetic moment as referenced against the Nuclear Magnetic Moment

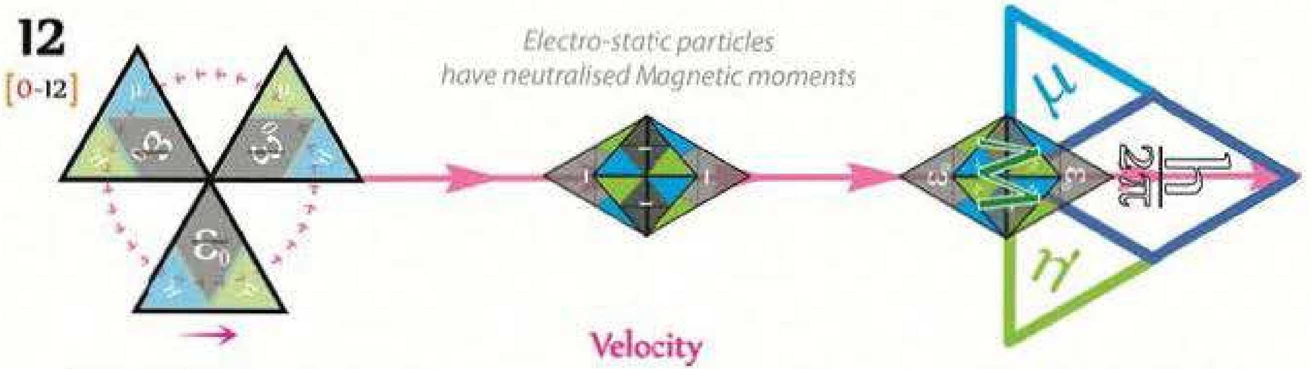
12
[0-12]
μ
ν
h/2π

Electro-static particles have neutralised Magnetic moments

A moving Lepton creates a secondary stronger intrinsic magnetic dipole moment within its KEM field which interacts with external magnetic fields

Velocity creates Kinetic Energy and Magnetic moments

Reversing the vector direction of the particle's linear momentum creates a reversed dipole Magnetic moment

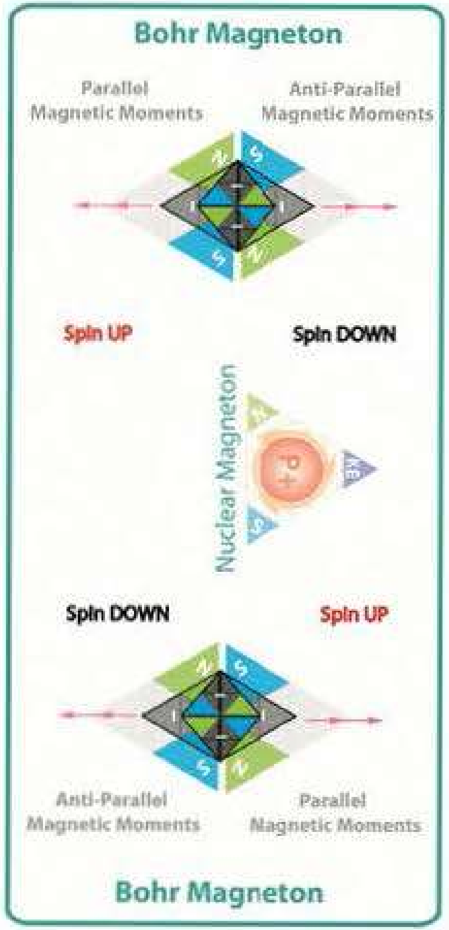
Bohr Magneton
Parallel Magnetic Moments
Anti-Parallel Magnetic Moments
Spin UP
Spin DOWN
Nuclear Magneton
p+
Spin DOWN
Spin UP
Anti-Parallel Magnetic Moments
Parallel Magnetic Moments
Bohr Magneton

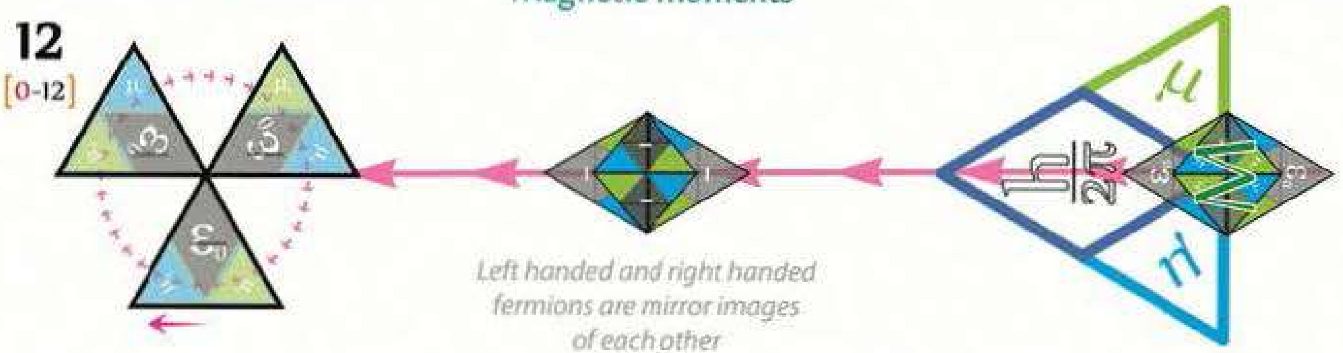
12
[0-12]
μ
ν
h/2π

Left handed and right handed fermions are mirror images of each other

Leptons are not point particles

All Leptronic macro-KEM fields and interactions with external fields can be modelled using Tetryonic geometries

Magnetic moment 'spin orientations' are reversed for opposite charge particles

Electron Spin orientation

The Bohr magneton dipole produced by Kinetic Energies
is located axially about the centre of rotation

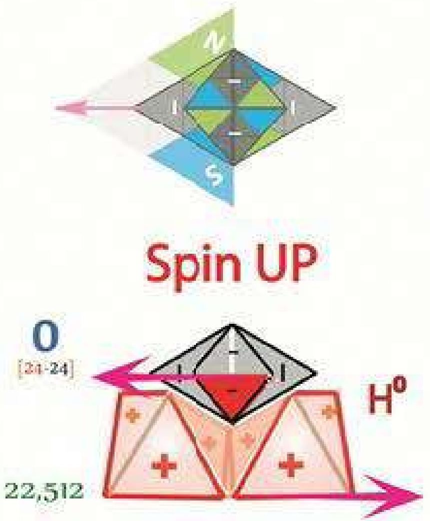
N
S
Spin UP
0
[24-24]
H⁰
22,512

Higher energy

Parallel spin

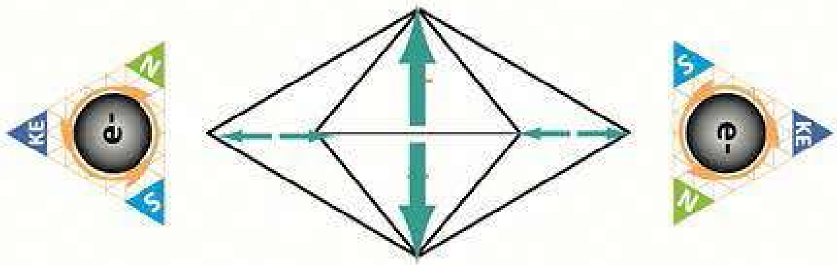
KE
e-
N
S
S
e-
KE
N

All Leptons have 12 intrinsic neutralised dipole moments
and a polarised KEM field Magnetic moment
created by the energies of its motion

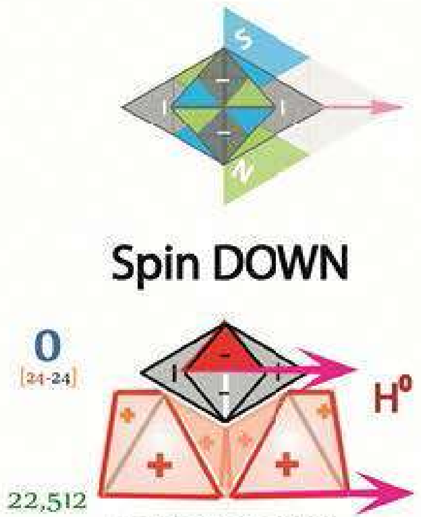
S
N
Spin DOWN
0
[24-24]
H⁰
22,512

Lower energy

Anti-parallel spin

N
External H field
S

Spin UP

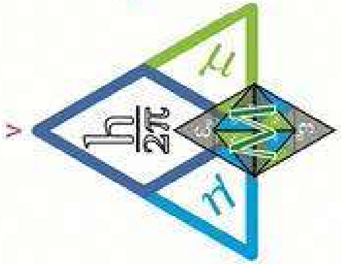

Magnetic field is Parallel to
Nuclear Magnetic moment
or external Magentic H field

Lepton
Magnetic moment

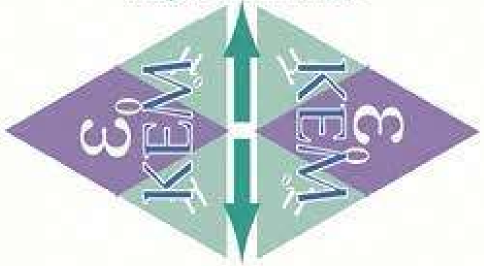

[Bohr Magneton]

$$\mu_B = \frac{e\hbar}{2m_e}$$

Spin DOWN

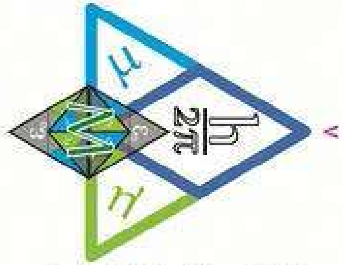

Magnetic field is Antiparallel to
Nuclear Magnetic moment
or external Magnetic H field

Nuclear Magneton
N
p+
KE
S

Nuclear magnetons
are weaker than
Bohr magnetons

All Leptronic spin directions are referenced to external Magnetic fields [either Nuclear Magnetons or H fields]

Nuclear Spins

are determined by the orientation of Bohr Magneton
with respect to the Nuclear Magneton
[or an external Magnetic field]

*election Matter geoemtries
are negative charged fascia
which create neutral intrinsic
magnetic dipole configurations*

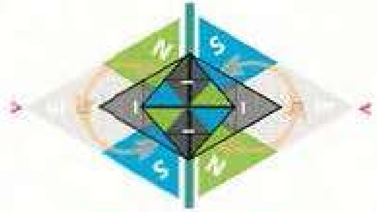

MAGNETIC MOMENT

*In a moving electron
Kinetic energy produces
an axial Magnetic moment*

*In a static electron
all intrinsic dipoes are
neutralised through
their orientations*

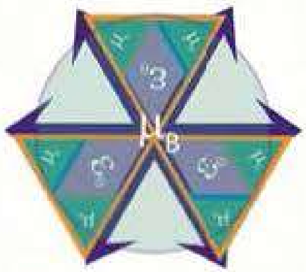

***Quantum Inductive Loops
In motion produce Kinetic Energies
In turn creating Magnetic Moments***

Electron

PARALLEL
Magnetic moments

BOHR
Magneton

ANTIPARALLEL
Magnetic moments

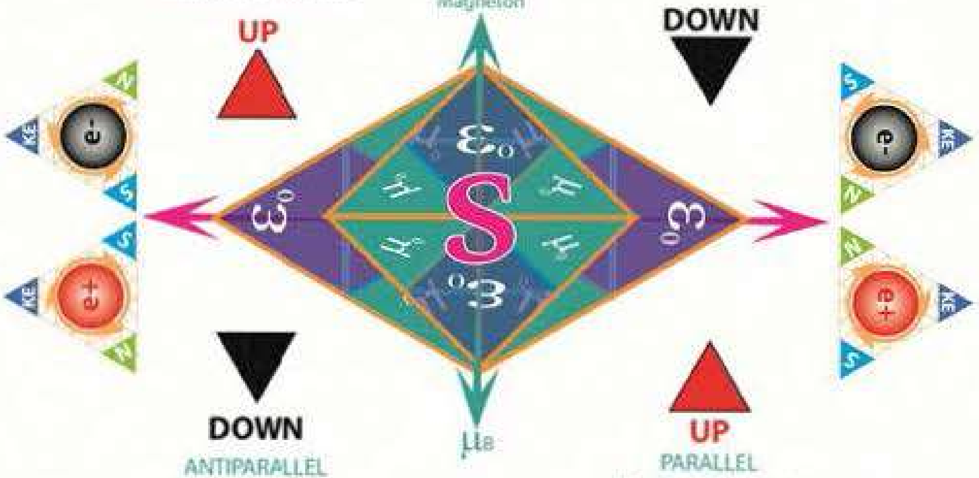

ANTIPARALLEL
Magnetic moments

Positron

PARALLEL
Magnetic moments

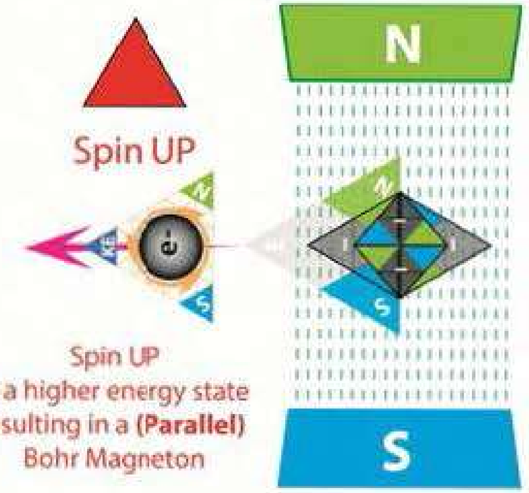

Spin UP
is a higher energy state
resulting in a **(Parallel)**
Bohr Magneton

Nuclear magneton or external Magnetic field

Spin DOWN
is a lower energy state
resulting in a **(antiParallel)**
Bohr Magneton

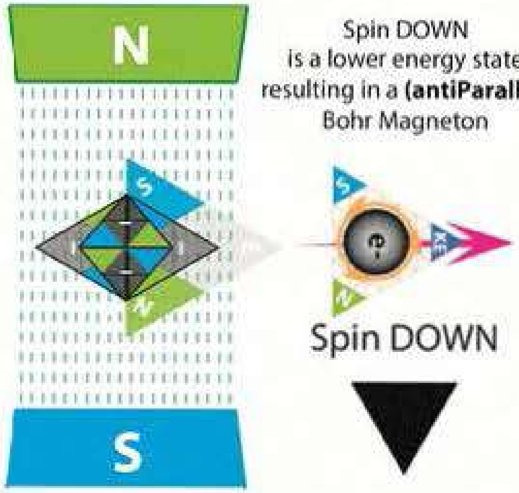

Lorentz Force

**When moving in an
external magnetic field**

The axial dipole
moment of an
electron will experience
torque force
proportional to
its velocity

**Opposites attract
Similars repel**

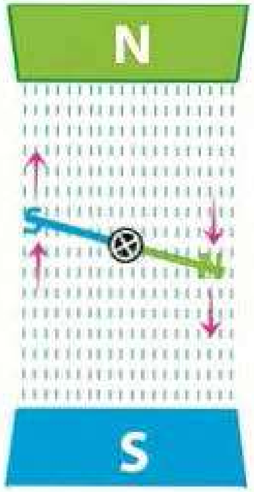

Nuclear magnetic Moment

[Nuclear magneton]

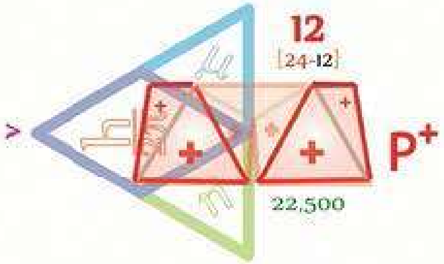
>
12
[24-12]
P+
22,500

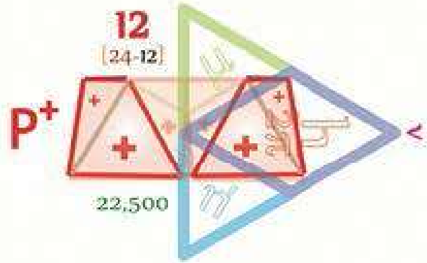
12
[24-12]
P+
22,500
<

NUCLEAR magnetons are much weaker than BOHR magnetons due to the higher charge to mass ratio of electrons

$$\frac{M_p}{M_e} = \frac{1}{1875}$$

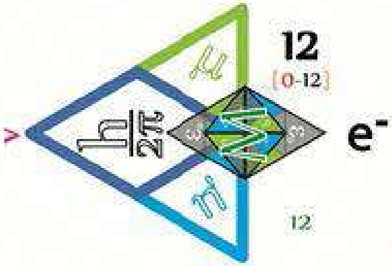

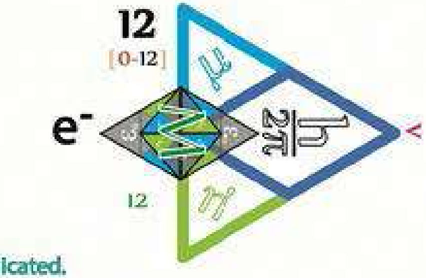

The nuclear spins for individual protons and neutrons parallels the treatment of electron spin, with spin 1/2 and an associated magnetic moment.

For the combination of electron, neutrons and protons in periodic elements, the situation is even more complicated.

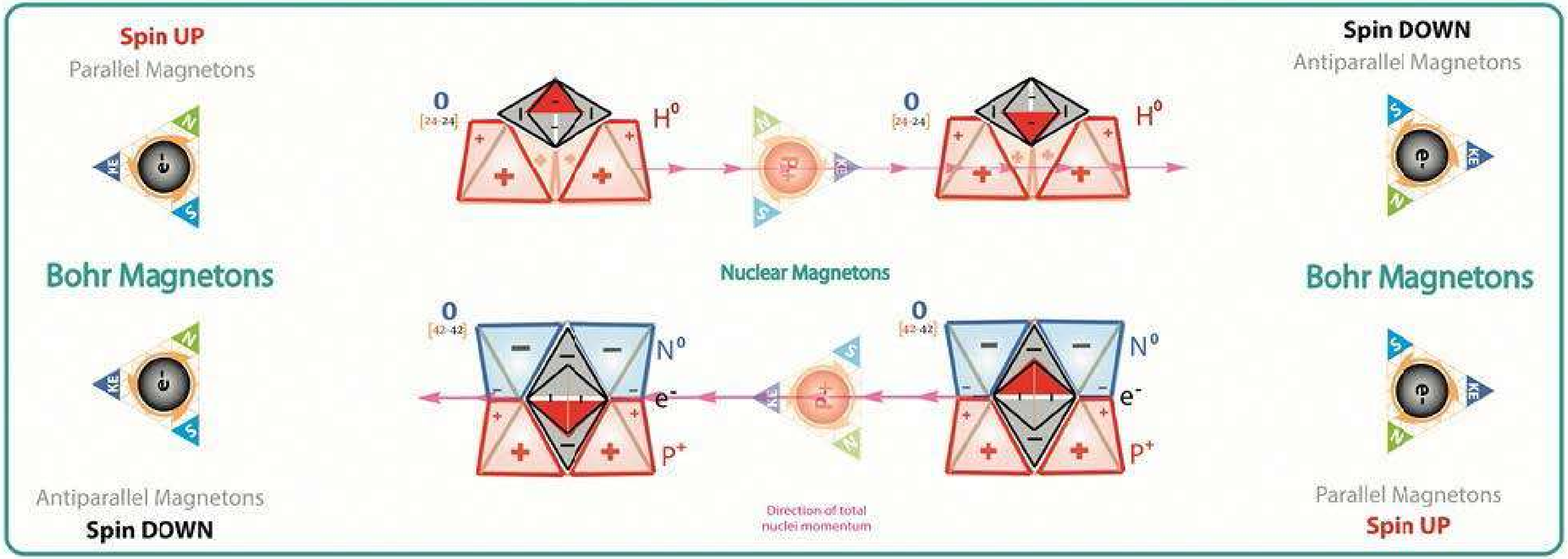
e-
KE
N
S
0
[24-24]
H0
0
[24-24]
H0
0
[42-42]
N0
e-
P+
0
[42-42]
N0
e-
P+

Spin UP
Parallel Magnetons

Spin DOWN
Antiparallel Magnetons

Bohr Magnetons

Nuclear Magnetons

Bohr Magnetons

Antiparallel Magnetons
Spin DOWN

Direction of total nuclei momentum

Parallel Magnetons
Spin UP

Tetryonics 19.08 - Nuclear Magnetic moments

Spin orbital coupling mechanics

The induced magnetic moment of electrons in atomic nuclei combine vectorally with the magnetic moment of the nuclei

anti-parallel moments

Spin DOWN

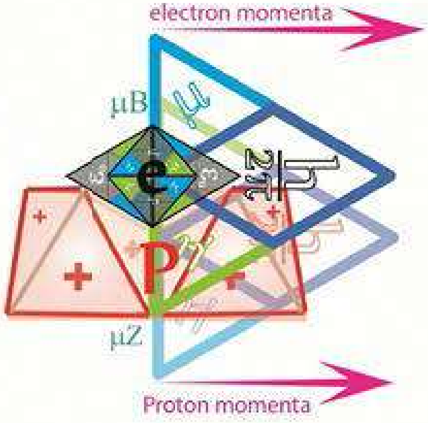

electron momenta
μB
μ
e
P
μZ
Proton momenta

μB

Bohr magnetons

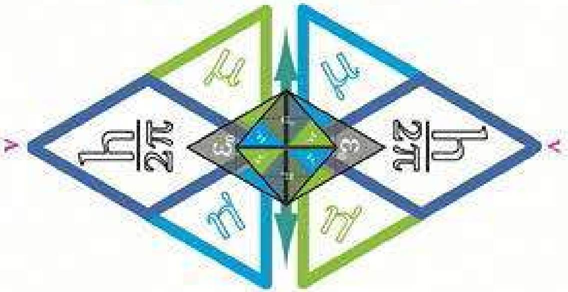

parallel moments

Spin UP

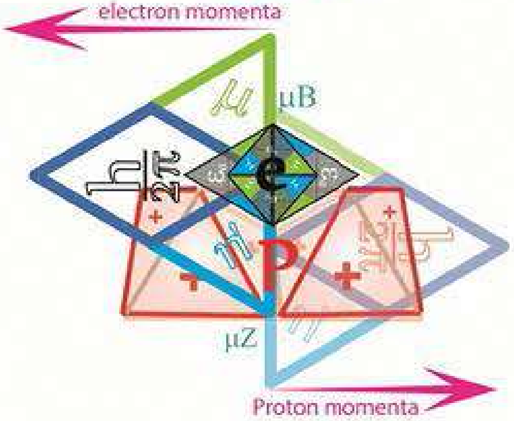

electron momenta
μB
μ
e
P
μZ
Proton momenta

The energy level differences created are manifested in Hyperfine-line splitting Zeeman effects etc.

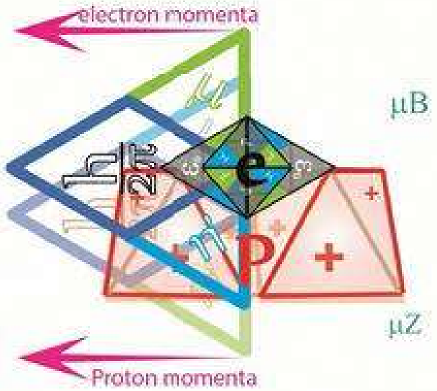

electron momenta
μB
μ
e
P
μZ
Proton momenta

Spin DOWN

anti-parallel moments

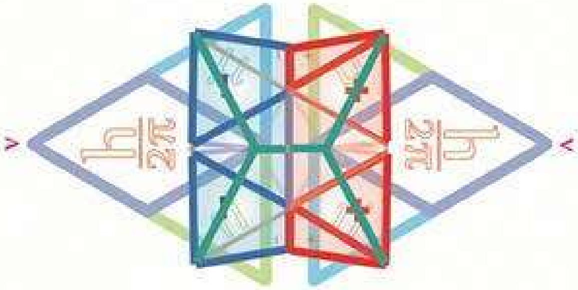

Nuclear magnetons

μZ

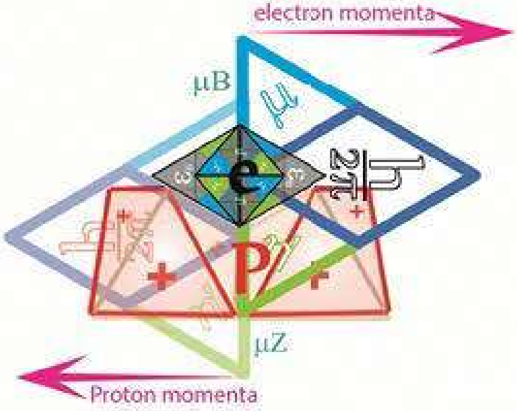

electron momenta
μB
μ
e
P
μZ
Proton momenta

Spin UP

parallel moments

The alignment of electron spins in nuclei results in diamagnetic, para-magnetic Matter

Copyright ABRAHAM [2008] - all rights reserved

Gyromagnetic Ratio

The electron is a 12 charge quantum rotor
with a uniform charge to mass density ratio,
the ratio of its magnetic moment to
its orbital angular momentum,
also known as gyromagnetic ratio

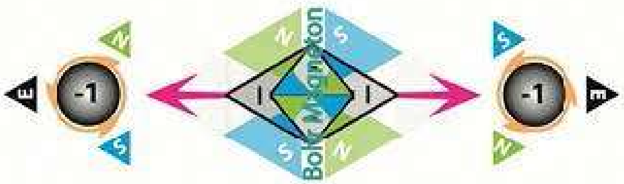

This implies that a more massive
assembly of charges spinning with the
same angular momentum will have a
proportionately weaker magnetic moment,
compared to its lighter counterpart.

12
[0-12]
1.2 e20

1
electron
12π

The Bohr Magneton is determined by the
charged KEM field geometry of Leptons

1875
charged mass-Matter differential

The combined Kinetic energy of Motion [KEM field]
and Electron Spin coupling with Nuclear Magnetons
will effect any measured Gyromagnetic ratios

22,550
Proton
36π

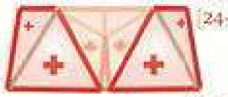

12
[24-12]
2.25 e23

-1

9.1 e-31 kg

Classical electron model
[rotating sphere]

4π

An electron is
NOT
a spherical particle

12π

12
[0-12]

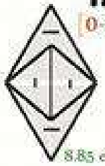

8.85 e-31 kg

Tetryonic electron
[rotating tri-tetryon topology]

In physics, the gyromagnetic ratio
(also sometimes known as the magnetogyric ratio in other disciplines)
of a particle or system is the ratio of its magnetic dipole moment to
its angular momentum, and it is often denoted by the symbol γ, gamma.

An isolated electron has an angular momentum and
a magnetic moment resulting from its spin.

Its SI units are radian per second per tesla ($s^{-1} \cdot T^{-1}$) or,
equivalently, coulomb per kilogram ($C \cdot kg^{-1}$).

Tetryonic quantum mass to Charge ratio

$$1.810109642 \times 10^{11}$$

The 2006 CODATA
– e / me = -1.758 820 150(44) x 10[11]

Nuclear Magnetons

$$\frac{e-}{e_m} \quad \frac{\text{Tetryonic elementary charge } 1.602216081\text{ e-19 coulombs}}{\text{Tetryonic electron mass } 8.851486361\text{ e-31}} \quad \frac{12q}{1.2\text{ e20}}$$

[q/m]

181,010,964,200 C/kg

Spin UP
K.E.

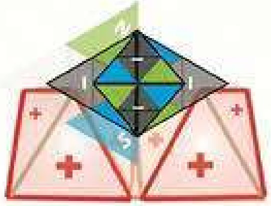

Lower energy

Electron Spin
[Orbital Angular Momentum]

N
Nuclear Magneton
S

Spin DOWN
K.E.

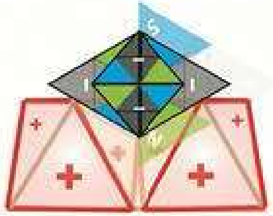

Higher energy

$$\frac{12q}{2.25\text{ e23}} \quad \frac{\text{Tetryonic elementary charge } 1.602216081\text{ e-19 coulombs}}{\text{Tetryonic Proton mass } 1.659653693\text{ e-27}} \quad \frac{e+}{P_m}$$

[q/m]

96,539,180.9 C/kg

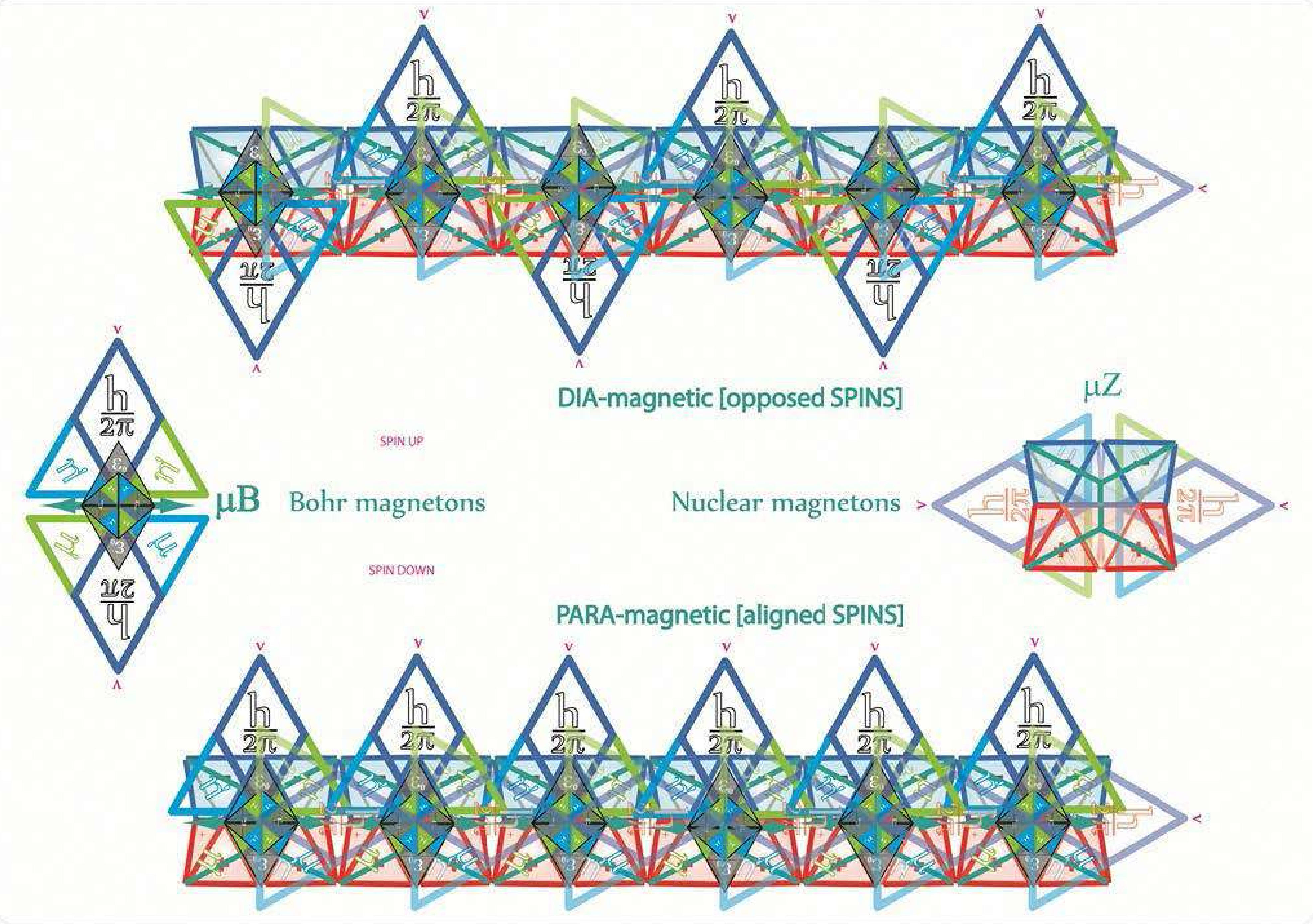
DIA-magnetic [opposed SPINS]
μZ
SPIN UP
μB
Bohr magnetons
Nuclear magnetons
SPIN DOWN
PARA-magnetic [aligned SPINS]
h/2π

Tetryonics 19.11 - DIA & PARA-magnetic fields

mass-charge ratios

12
[12-0]
1.2e20

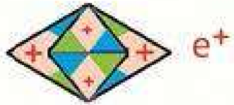

Positron

Leptons

181,010,964,200 C/kg

12
[0-12]
1.2e20

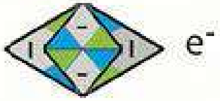

Electron

12
[24-12]
2.25e23

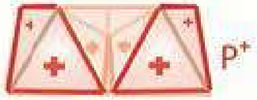

Proton

Baryons

96,539,180.9 C/kg

0
[18-18]
2.25e23

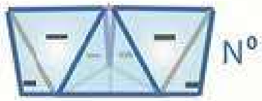

Neutron

All mass-Matter have distinct charge geometries
and once in motion are subject to EM forces
as a result of their geometries & KEM field

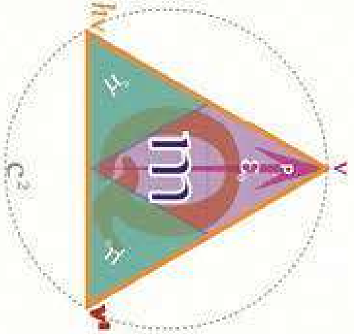

Lorentz force Law

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$

LORENTZ force | Electric force | charge velocity | Magnetic force

all Matter in motion is subject to
NEWTON'S SECOND LAW OF MOTION

$$F = ma$$

Combining
Lorentz's force law & Newton's Third Law
we obtain a EM mass-charge quotient
for any particle of mass-Matter in motion

$$(M/Q)a = E + v \times B$$

1.335180067 e-20 C

1.810109642 e11 $\left[\frac{q}{m}\right]$m **1e19 V**

7.376238634 e-51 kg

are the basis for mass spectroscopy

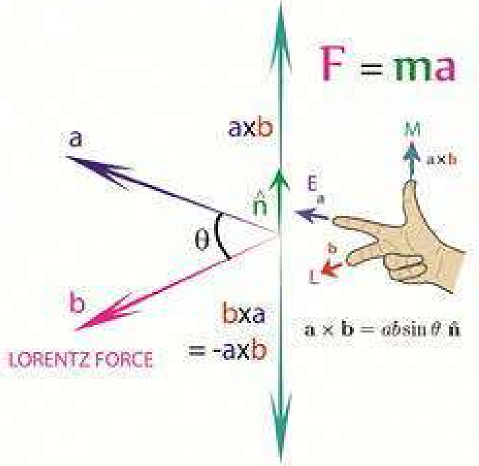

For every action there is an equal and opposite reaction
(NEWTON'S THIRD LAW OF MOTION)

Q *Any two particles with the same EM mass-to-charge ratio follow the same path in a vacuum when subjected to the same external electric field* **M**

Dependent on their generations

Charge/mass

$\left(\frac{q_t}{m_t}\right)$ **181,010,964,200 C/kg** $\left(\frac{q_e}{m_e}\right)$

4:4 12:12

Tetryons and Leptons can have identical mass-charge ratios

96,487,720.78 C/kg

Elements

48,256,721.99 C/kg

0
[24-24]
2.2512e23

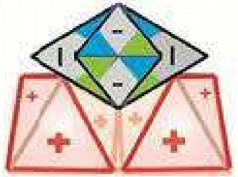

Hydrogen

Deuterium

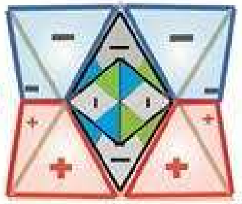

0
[42-42]
4.5012e23

Tetryonic theory shows mass-charge ratios are a measure of mass-energy geometries in Matter topologies

Tetryonics 20.01 - mass-Charge ratios

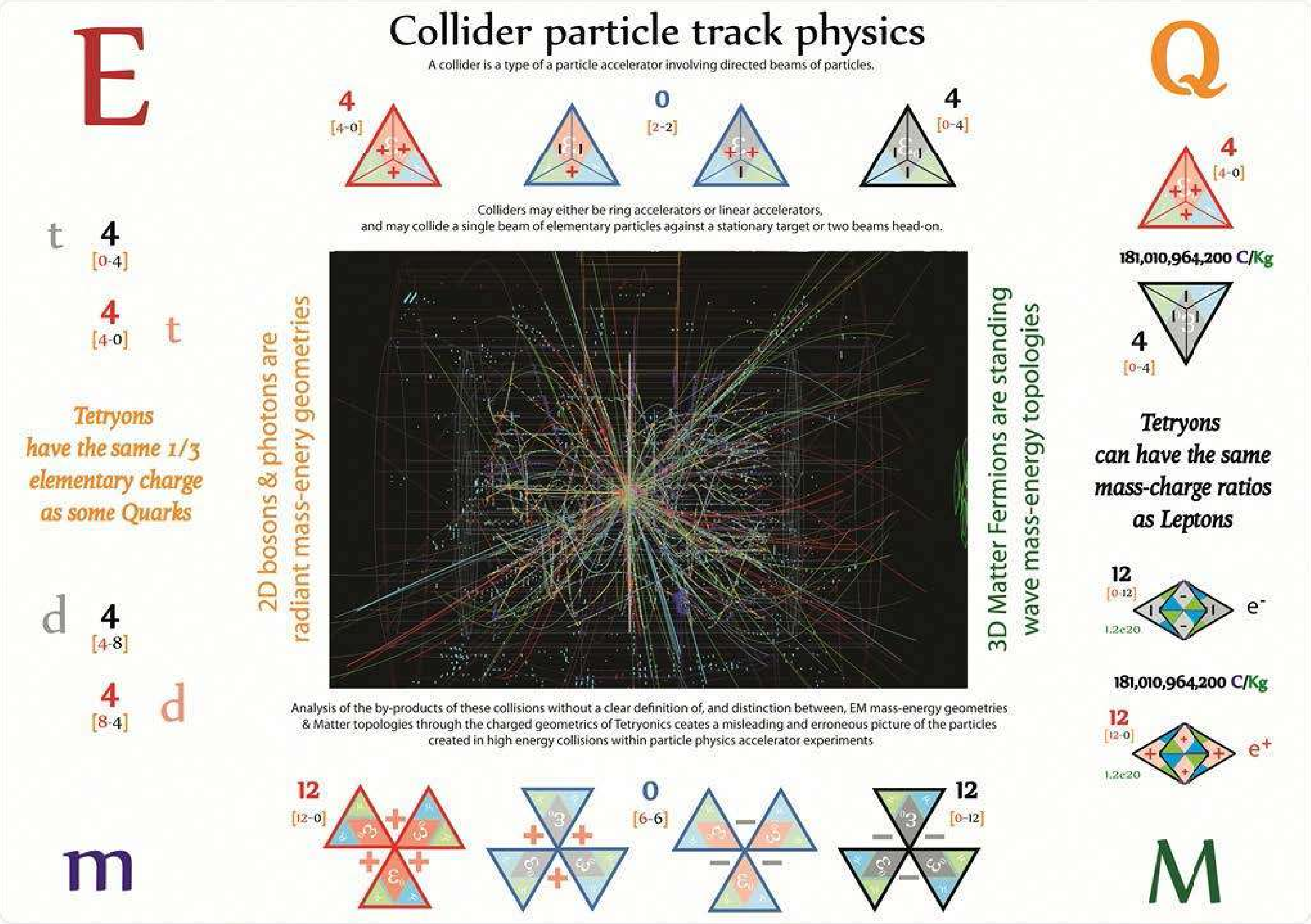

Collider particle track physics

A collider is a type of a particle accelerator involving directed beams of particles.

Colliders may either be ring accelerators or linear accelerators,
and may collide a single beam of elementary particles against a stationary target or two beams head-on.

2D bosons & photons are radiant mass-enery geometries

3D Matter Fermions are standing wave mass-energy topologies

Analysis of the by-products of these collisions without a clear definition of, and distinction between, EM mass-energy geometries & Matter topologies through the charged geometrics of Tetryonics ceates a misleading and erroneous picture of the particles created in high energy collisions within particle physics accelerator experiments

Tetryons have the same 1/3 elementary charge as some Quarks

181,010,964,200 C/Kg

Tetryons can have the same mass-charge ratios as Leptons

181,010,964,200 C/Kg

Tetryonics 20.02 - Collider particle track physics

Radiant EM mass-energy geometries

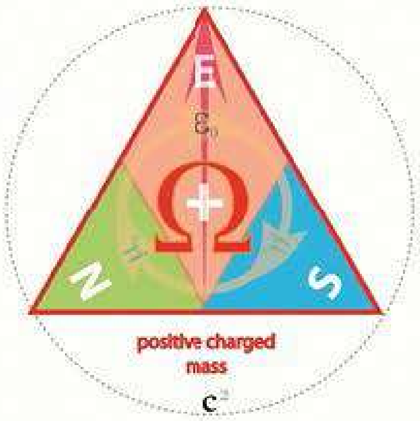

positive charged mass

mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

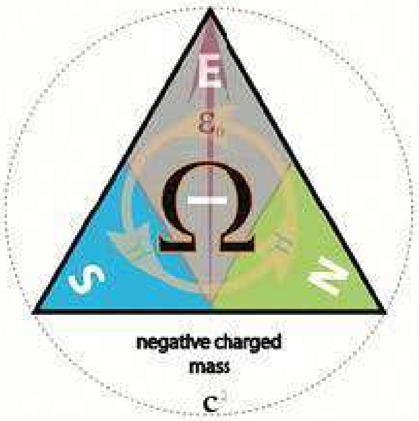

negative charged mass

Charge

Planck quanta

$$\frac{1\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Bosons

Planck quanta

$$\frac{\text{ODD}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

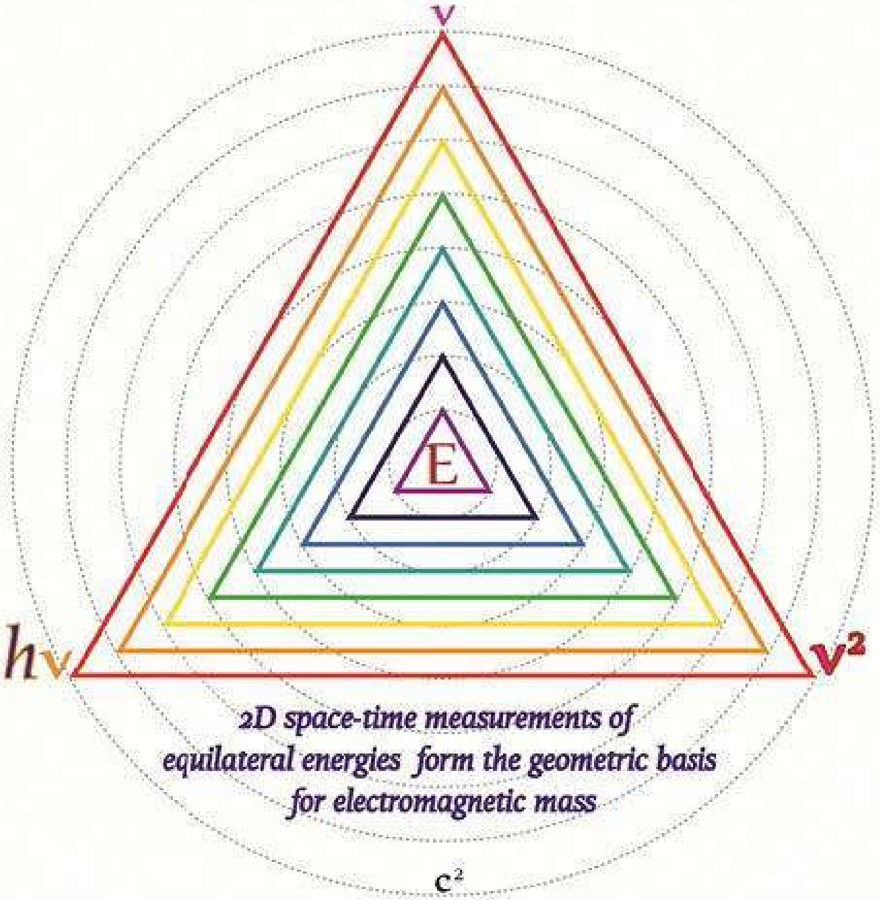

2D space-time measurements of equilateral energies form the geometric basis for electromagnetic mass

Photons

Planck quanta

$$\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

EM waves

Planck quanta

$$\frac{\text{EVEN}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

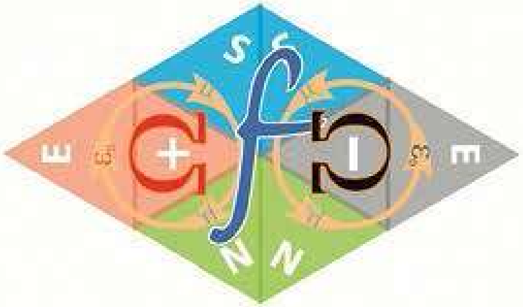

2D Planar divergent EM mass-energies form radiant Electromagnetic waves

ElectroMagnetic mass is a property of Matter
[being the energy content of its charged geometry per unit time]

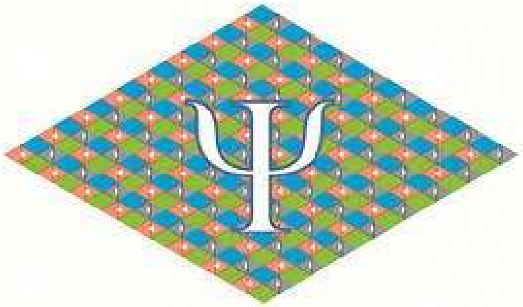

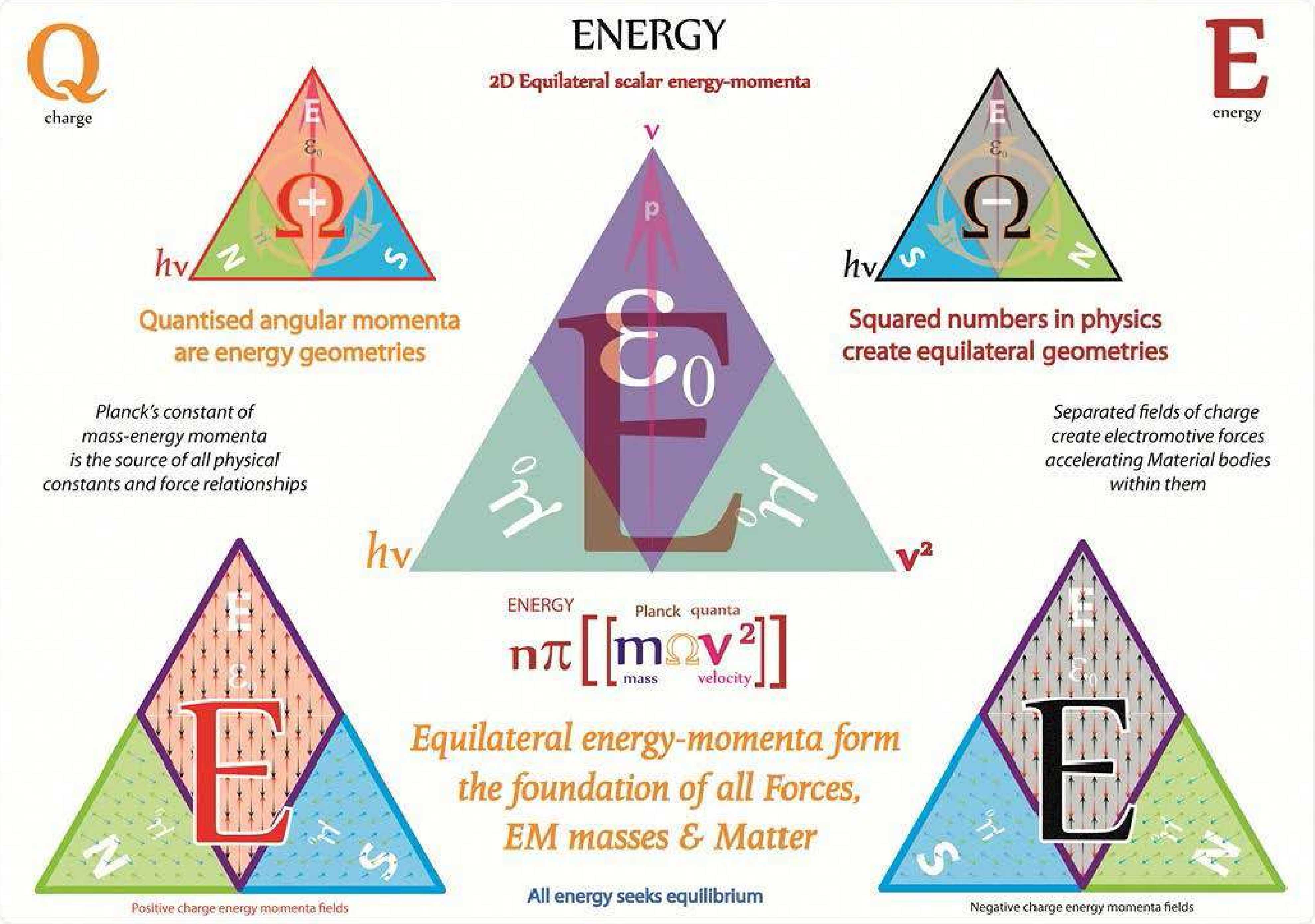

ENERGY

2D Equilateral scalar energy-momenta

Quantised angular momenta are energy geometries

Squared numbers in physics create equilateral geometries

*Planck's constant of
mass-energy momenta
is the source of all physical
constants and force relationships*

*Separated fields of charge
create electromotive forces
accelerating Material bodies
within them*

ENERGY

Planck quanta

$$n\pi\left[\left[m\Omega v^{2}\right]\right]$$

mass velocity

*Equilateral energy-momenta form
the foundation of all Forces,
EM masses & Matter*

All energy seeks equilibrium

Positive charge energy momenta fields

Negative charge energy momenta fields

3D Matter topology is NOT a property of 2D ElectroMagnetic mass-energies

it is a measure of the closed 3D standing-wave spatial topology of all fermionic particles created by their charged equilateral mass-energy momenta geometries

Matter displaces vacuum energies to create convergent gravity fields around it

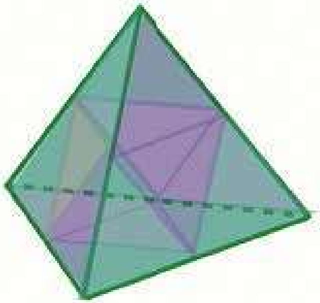

Tetryons are the quantum of Matter

Tetryons

$$\frac{4\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Fermions

$$\frac{12\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Mesons

$$\frac{24\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

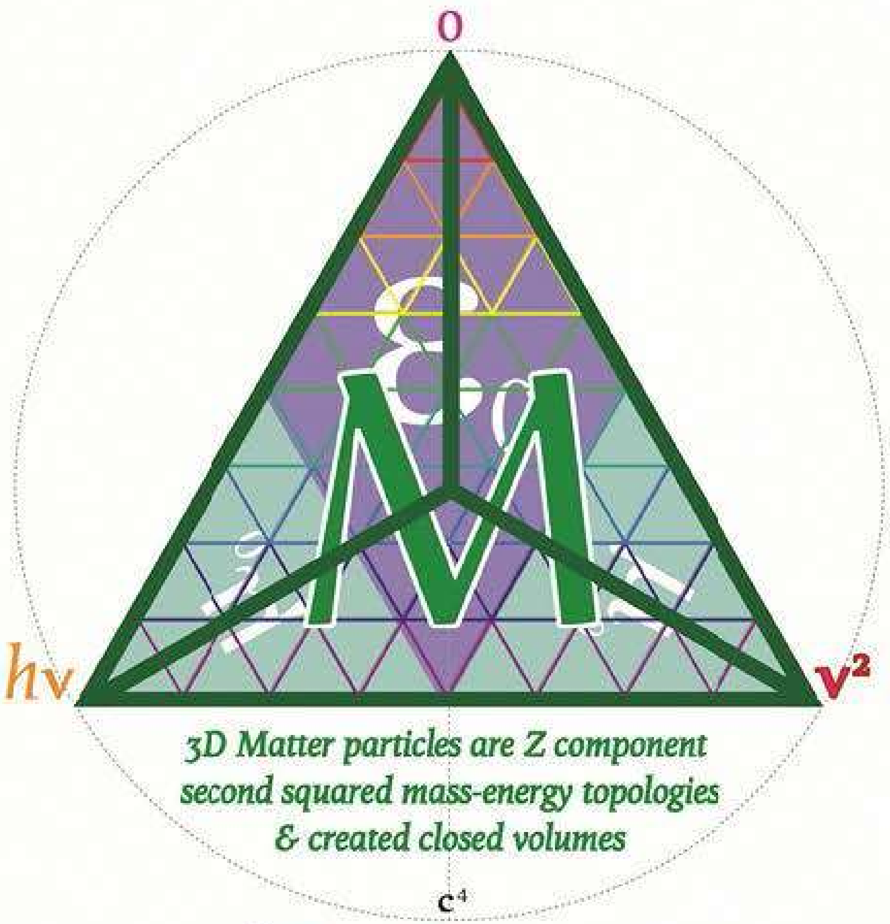

3D Matter particles are Z component second squared mass-energy topologies & created closed volumes

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

The mass-energies of Matter are Lorentz invariant to velocity changes

All Matter is comprised of [and radiates divergent] kEM mass-energies

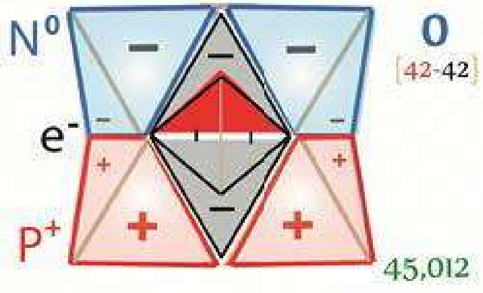

Deuterium is the quantum of all Elements

Baryons

$$\frac{36\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Nuclei

$$\frac{72\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Deuterons

$$\frac{84\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

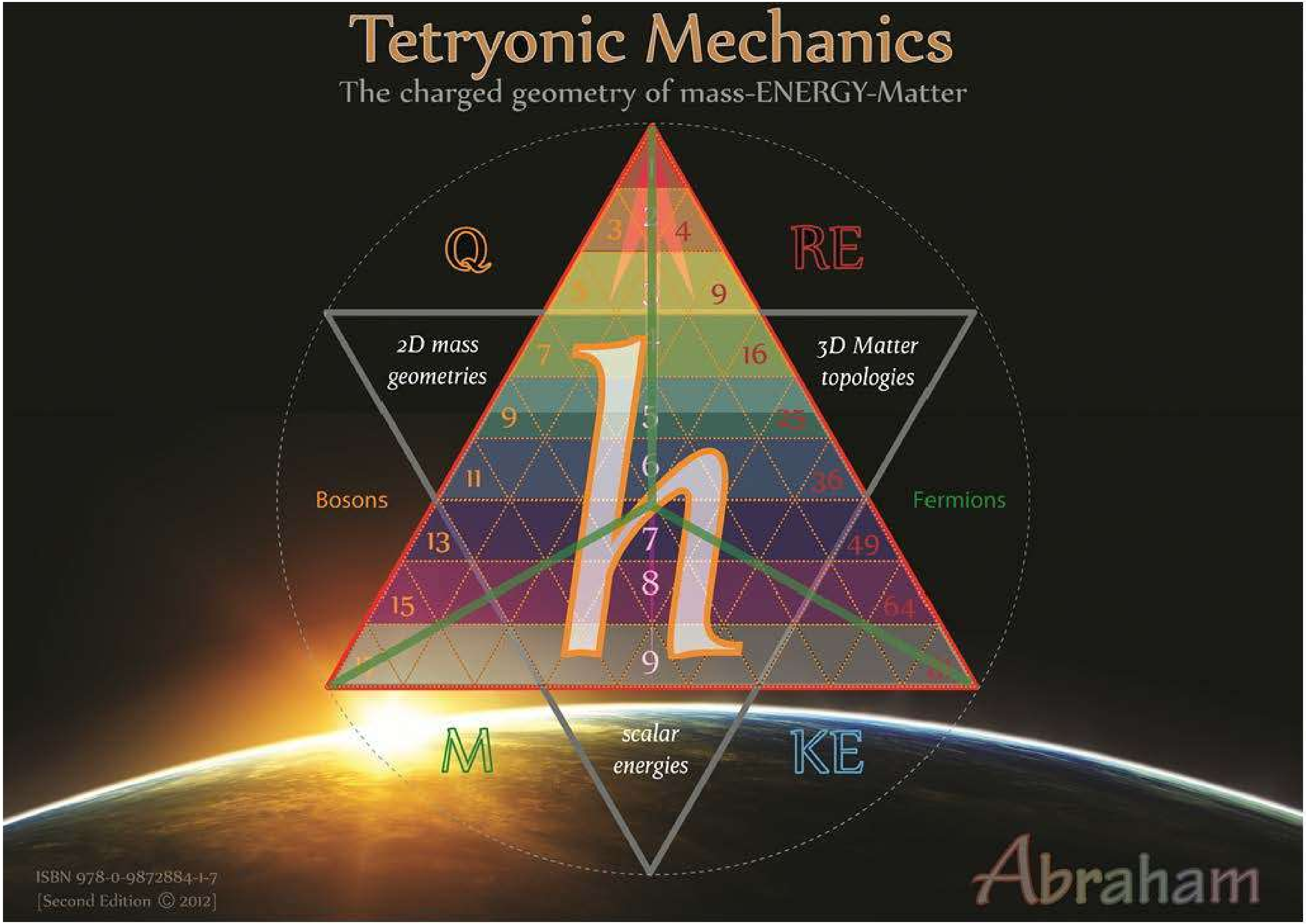

Tetryonic Mechanics

The charged geometry of mass-ENERGY-Matter

Q
RE
2D mass geometries
3D Matter topologies
Bosons
Fermions
h
M
KE
scalar energies
3
4
5
9
7
16
9
25
11
36
13
49
15
64
5
6
7
8
9
Abraham

ISBN 978-0-9872884-1-7
[Second Edition © 2012]

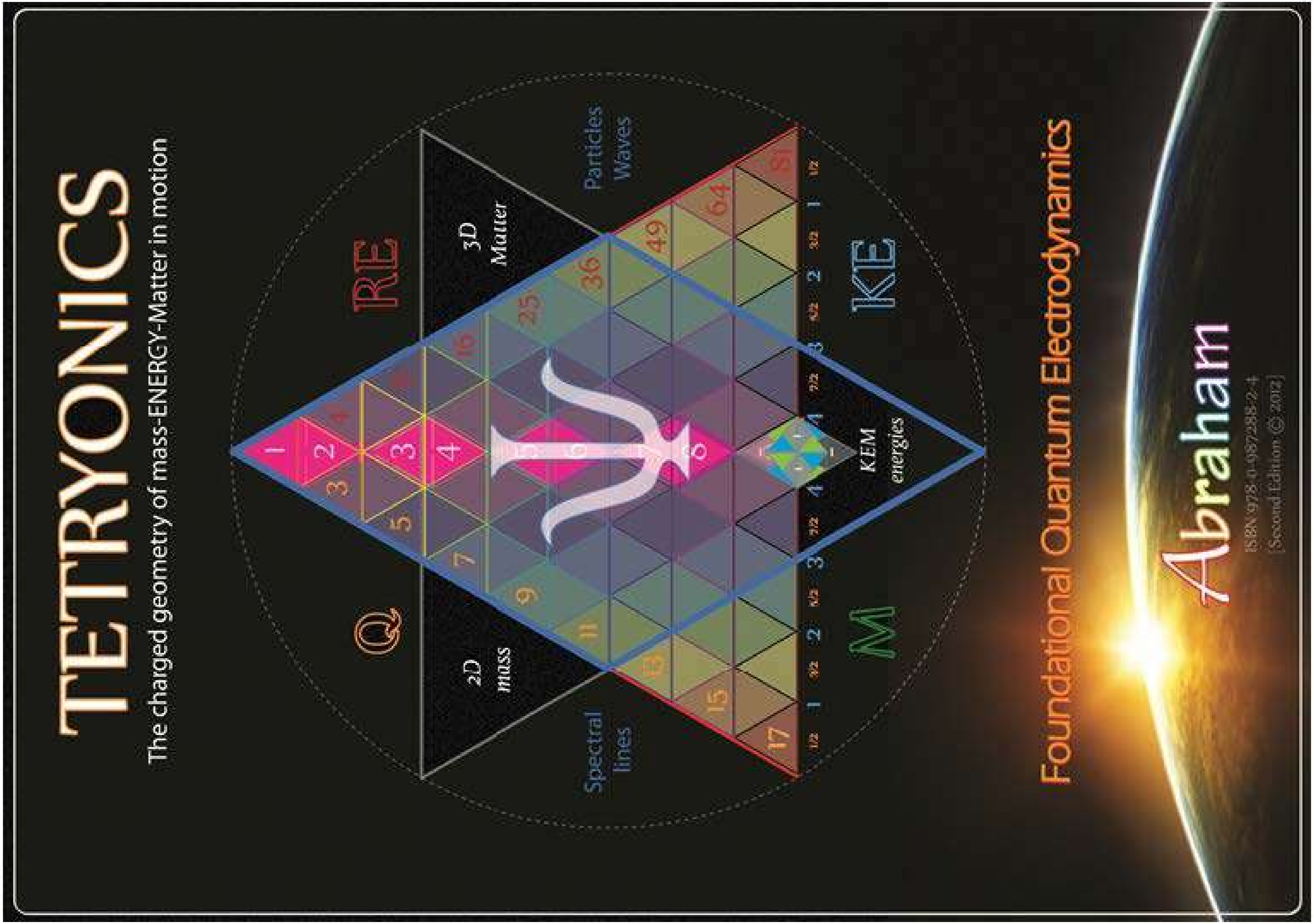

TETRYONICS

The charged geometry of mass-ENERGY-Matter in motion

Foundational Quantum Electrodynamics

Abraham

ISBN 978-0-9872882-2-4
[Second Edition © 2012]

RE
KE
Q
M
3D
Matter
2D
mass
Particles
Waves
Spectral
lines
KEM
energies
1
2
3
4
5
6
7
8
9
11
13
15
16
17
25
36
49
64
81

Tetryonics 21.00 - Tetryonic Electrodynamics title page

Copyright ABRAHAM [2008] - all rights reserved

James Clerk Maxwell

(13 June 1831 – 5 November 1879)

Q
[v-v]

Net Electric charge is the result of quanta imbalances in equilateral geometry

EM Planck Energy

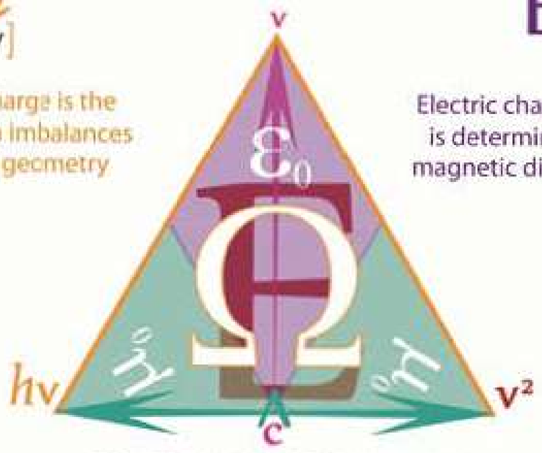

ORTHOGONAL MAGNETIC DIPOLE FIELD

E

Electric charge polarity is determined by the magnetic dipole vector

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

Gauss's Law for Electricity

$$\oint \vec{B}\cdot d\vec{A} = 0$$

Gauss's Law for Magnetism

$$\oint \vec{E}\cdot d\vec{s} = \frac{-d\Phi_B}{dt}$$

Faraday's Law

$$\oint \vec{B}\cdot d\vec{s} = \mu_0 i + \frac{\mu_0 \varepsilon_0 d\Phi_E}{dt}$$

Ampere's Law

E = electric field
A = area
q = charge
ε_0 = permittivity of free space (a constant)
B = magnetic field intensity
s = path length
Φ_B = magnetic flux
Φ_E = electric flux
t = time
μ_0 = permeability of free space (a constant)
i = current electric charge
$\mu_0\varepsilon_0$ = $1/c^2$, where c=speed of light

Maxwell's equations describe how Electric and Magnetic fields are generated and altered by each other and by charges and currents.

Maxwell's Equations

Maxwell's equations are a set of partial differential equations that, together with the Lorentz force law, form the foundation of classical electrodynamics, classical optics, and electric circuits.

ElectroMagnetic fields in turn provide the foundation of our modern electrical and communications technologies.

q
[1-0]

POSITIVE [Ω] CHARGE

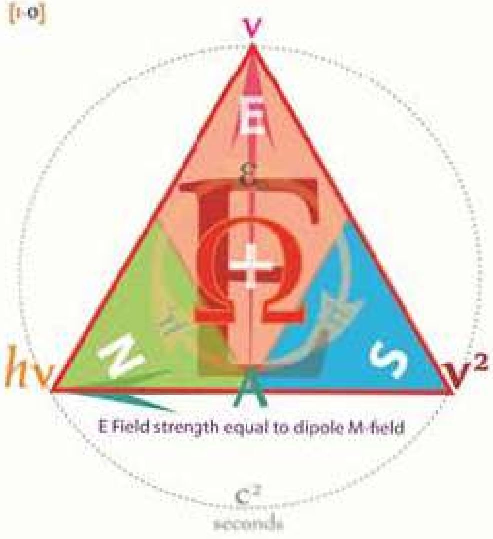

Q **Gauss' Law of Electric Flux** **Q**

Electric fields diverge from Electric charge, and produce the Coulomb force.

N **Gauss' Law of Zero Nett Magnetism** **S**

there are no isolated Magnetic poles, but the Henry force acts between the poles of a magnet.

ε_0 **Faraday's Law of Inductance** μ_0

Electric fields are produced by changing Magnetic fields.

Ampere's Circuital Law

circulating Magnetic fields are produced by moving Electric fields and by electric currents.

NEGATIVE [Ω] CHARGE

q
[0-1]

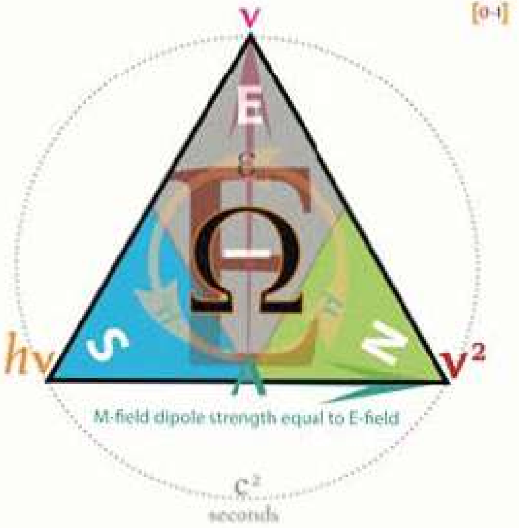

Copyright ABRAHAM [2008] - all rights reserved

Gauss' Laws

- Gauss' Law of Electric Flux and
- Gauss' Law of Magnetic dipoles

The laws were formulated by Carl Friedrich Gauss in 1835, but was not published until 1867.

They form two of Maxwell's equations which are the basis of classical electrodynamics.

The other two being:
- Faraday's law of induction; and
- Ampere's law with Maxwell's correction.

Gauss's law can be used to derive Coulomb's law, and vice versa.

Equilateral triangles are the foundational charge geometry of EM mass-Energy and Matter

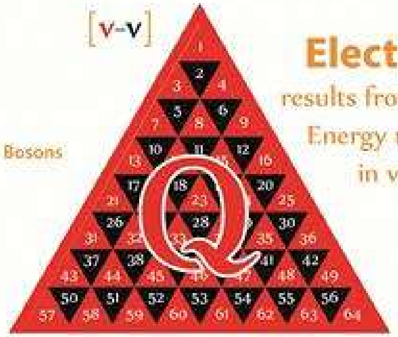

Electric Charge

results from an imbalance of Energy momenta quanta in v² geometries

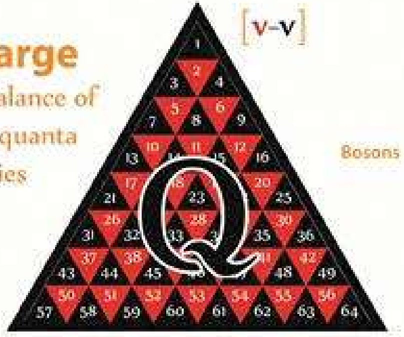

Carl Friedrich Gauss

(30 April 1777 – 23 February 1855)

Integral form

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0} = 4\pi kq$$

Differential form

$$\nabla\cdot E = \frac{\rho}{\varepsilon_0} = 4\pi k\rho$$

Electric fields are divergent from their source

Although originally envisaged as a property applied to spherical geometries and usually illustrated with square sectional boundaries integrating a surface area

$$\Phi = EA = E4\pi r^2 = \frac{Q}{\varepsilon_0}$$

Charge is shown to be the result of the symmetry of equilateral quantised angular momenta

Integral form

$$\oint \vec{B}\cdot d\vec{A} = 0$$

Differential form

$$\nabla\cdot B = 0$$

There are NO Magnetic Monopoles

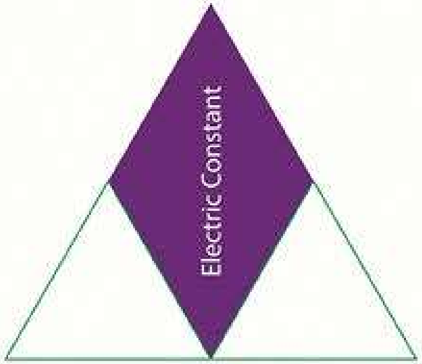

$$\varepsilon_0 = 8.85418785\text{e-}12 \ \frac{\mathrm{F}}{\mathrm{m}}$$

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

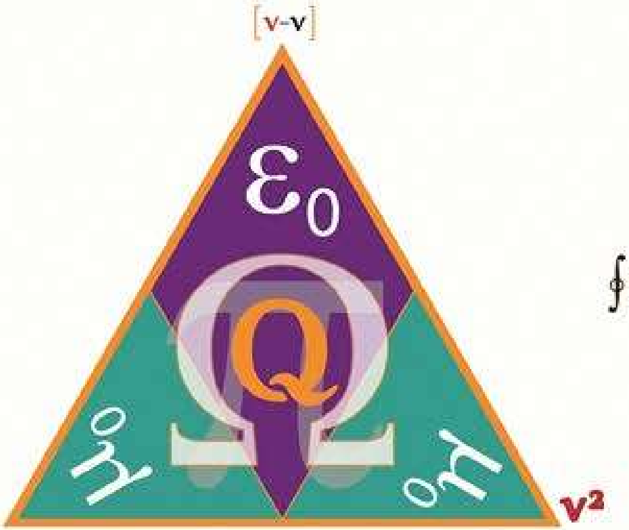

The electric flux through any closed surface is proportional to the enclosed electric charge

$$\oint \vec{B}\cdot d\vec{A} = 0$$

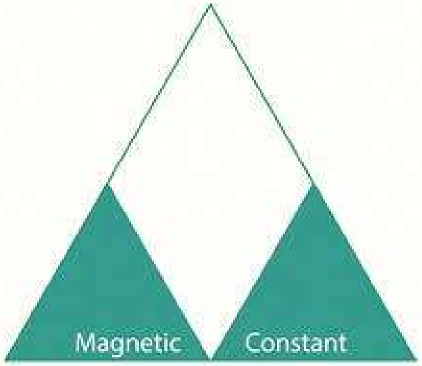

$$\mu_0 = 1.25663706\text{e-}6 \ \frac{\mathrm{H}}{\mathrm{m}}$$

The Divergent force of Charge

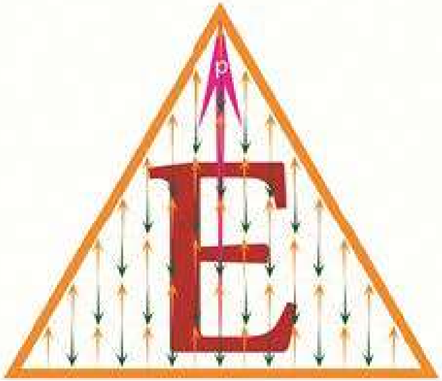

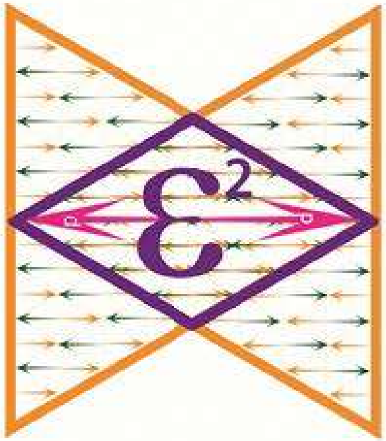

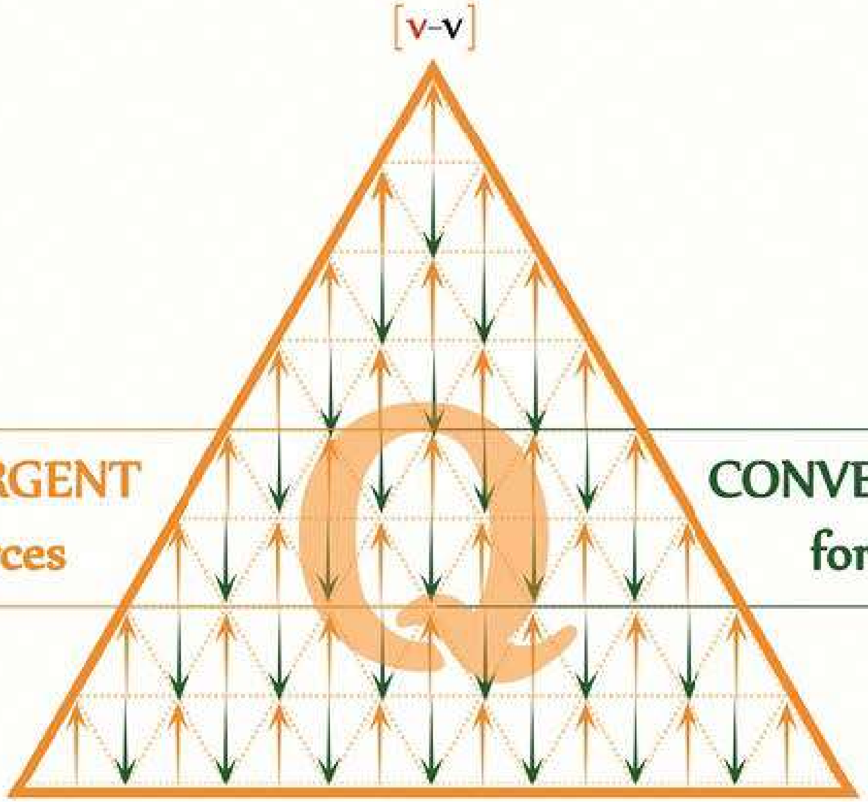

DIVERGENT forces

CONVERGENT forces

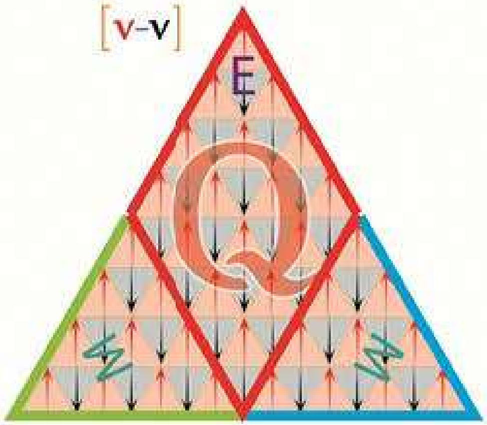

Charged EM fields create a net DIVERGENT force
as a result of their outward radiant energy momenta

but still possess CONVERGENT forces as a result of
their inherent equilateral energy momentum symmetries

Where these radiant charged fields diverge from their source
and overlap their superpositional energy momenta
create the forces of the Law of Interaction

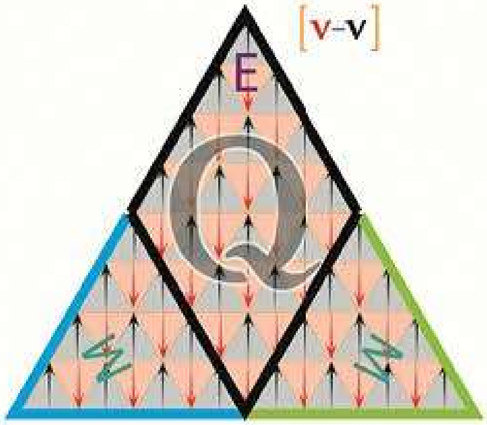

Coulomb's Law

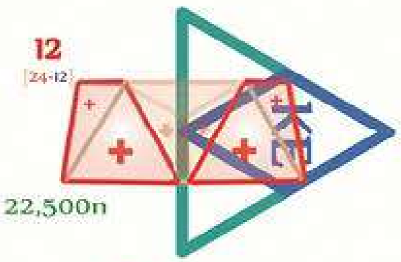
12
[24-12]
22,500n
KE

Proton

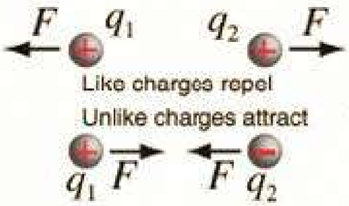
F
q_1
q_2
F
Like charges repel
Unlike charges attract
q_1
F
F
q_2

$$F = \frac{kq_1q_2}{r^2} = \frac{q_1q_2}{4\pi\varepsilon_0 r^2}$$

Coulomb's Law

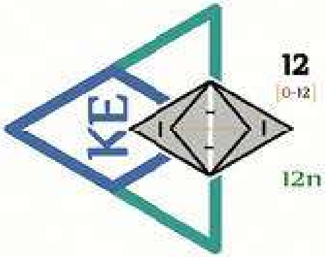
12
[0-12]
12n
KE

Electron

The magnitude of the electrostatic force between two point electric charges is directly proportional to the product of the magnitudes of each of the charges and inversely proportional to the square of the distance between the two charges.

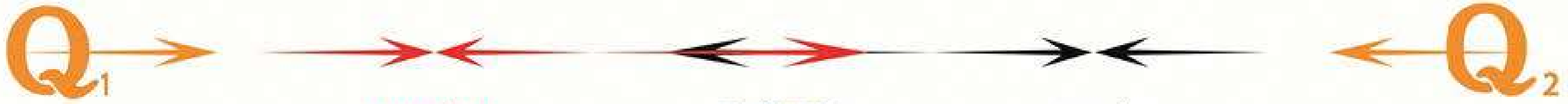

opposing energy momenta directions

co-linear energy momenta directions

opposing energy momenta directions

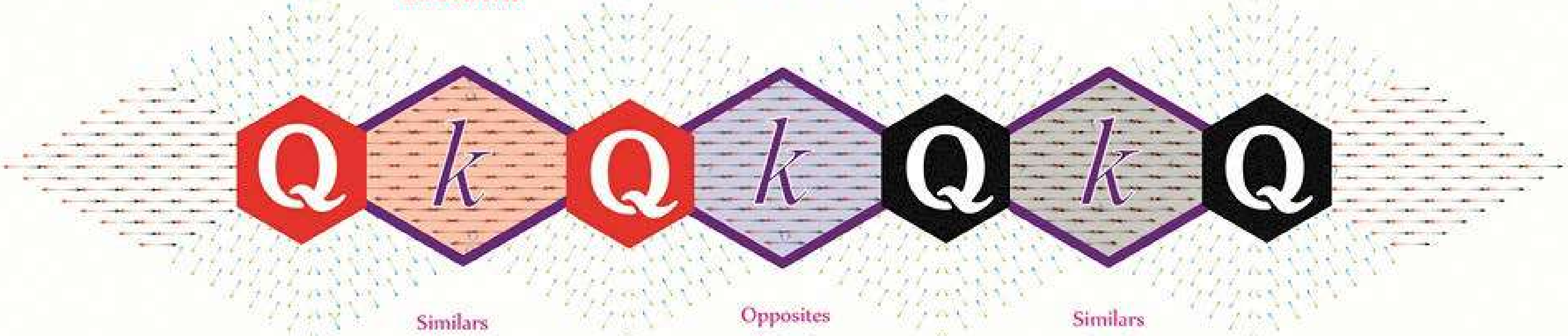

Similars Repel

Opposites Attract

Similars Repel

The law of Interaction

Changing Electric Fields produce changing Magnetic fields

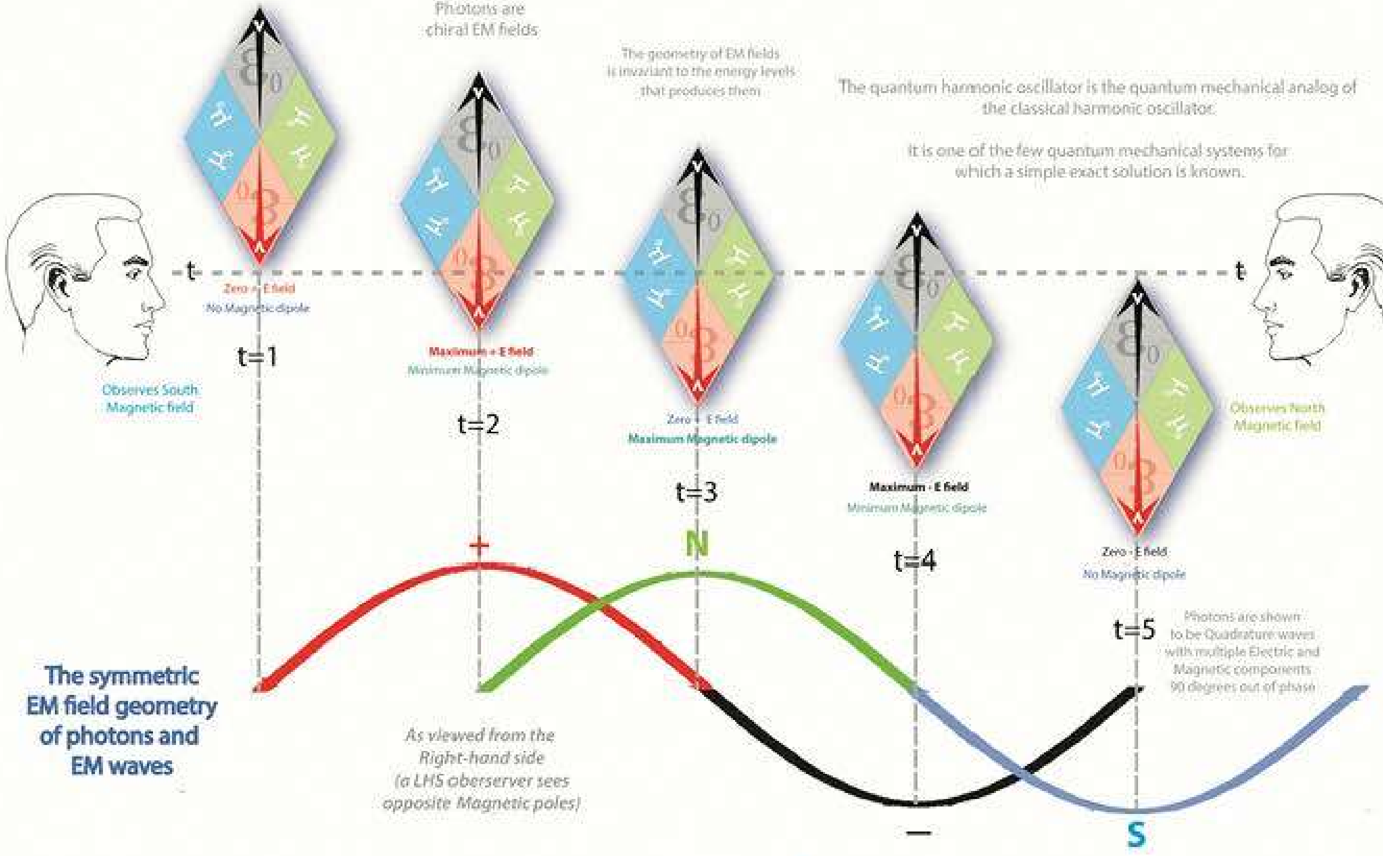
t
Zero + E field
No Magnetic dipole
t=1
Observes South Magnetic field
Maximum + E field
Minimum Magnetic dipole
t=2
Zero + E field
Maximum Magnetic dipole
t=3
Maximum - E field
Minimum Magnetic dipole
t=4
Zero - E field
No Magnetic dipole
t=5
Observes North Magnetic field
t
+
N
−
S

Photons are chiral EM fields

The geometry of EM fields is invariant to the energy levels that produces them

The quantum harmonic oscillator is the quantum mechanical analog of the classical harmonic oscillator.

It is one of the few quantum mechanical systems for which a simple exact solution is known.

Photons are shown to be Quadrature waves with multiple Electric and Magnetic components 90 degrees out of phase

The symmetric EM field geometry of photons and EM waves

As viewed from the Right-hand side (a LHS oberserver sees opposite Magnetic poles)

Changing Magnetic fields produce changing Electric fields

The symmetric ElectroMagnetic field strengths measured

Michael Faraday

(22 September 1791 – 25 August 1867)

EM field symmetry

" It is proposed that if a changing magnetic field can make an oscillating electric field, then

a changing electric field (from a oscillating electric charge) should make a changing magnetic field "

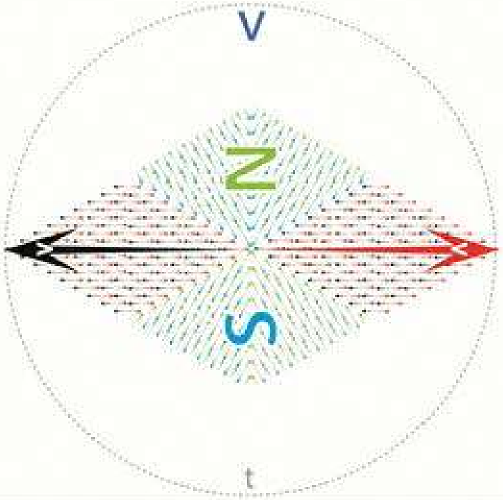

Any change in the magnetic environment of a coil of wire will cause a voltage (emf) to be "induced" in the coil.

No matter how the motional change is produced, a changing electromotive force will be generated.

The change could be produced by changing the magnetic field strength, moving a magnet toward or away from the coil, moving the coil into or out of the magnetic field, rotating the coil relative to the magnet.

All the resulting changes to the E & M fields and the vector strengths etc are the result of the inherent equilateral geometries of EM fields

Electromagnetic Induction

All ElectroMagnetic circuits are comprised of quantum inductive fields [ZPFs] and energies that obey Newton's third law and the conservation of energy

Electromagnetic induction is the production of an electric current across a conductor moving through a magnetic field.
It underlies the operation of generators, transformers, induction motors, electric motors, synchronous motors, and solenoids.

EM field coupling

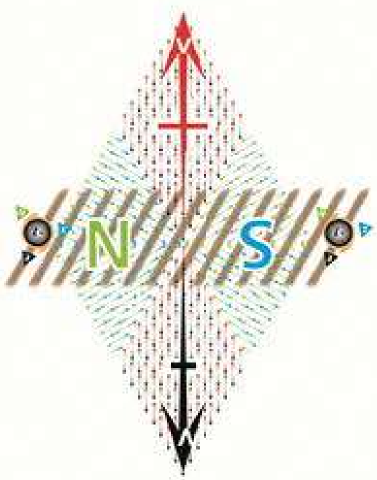

$$|\mathcal{E}| = \left|\frac{d\Phi_B}{dt}\right|$$

Michael Faraday formulated that electromotive force (EMF) produced around a closed path is proportional to the rate of change of the magnetic flux through any surface bounded by that path.

$$\nabla \times \mathbf{E} = -\frac{\partial \mathrm{B}}{\partial t}$$

$$V_\mathrm{p} = N_\mathrm{p}\frac{\mathrm{d}\Phi}{\mathrm{d}t}.$$

For the special case of a coil of wire, or a transformer circuit composed of N loops with the same area, Faraday's general equation becomes

$$V_\mathrm{s} = N_\mathrm{s}\frac{\mathrm{d}\Phi}{\mathrm{d}t},$$

Inductive Coupling

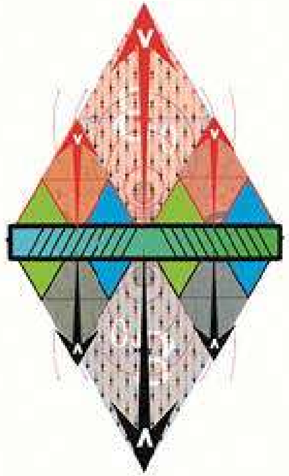

A corollary of Faraday's Law, together with Ampère's law and Ohm's law is Lenz's law:
The EMF induced in an electric circuit always acts in such a direction that the current it drives around the circuit opposes the change in magnetic flux which produces the EMF

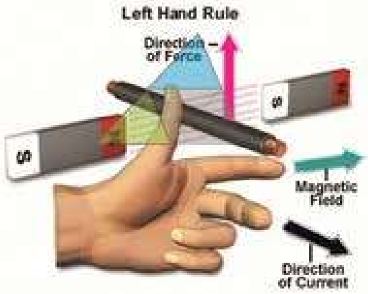

Flemings left hand rule
(for electric motors)

Coil moves UP

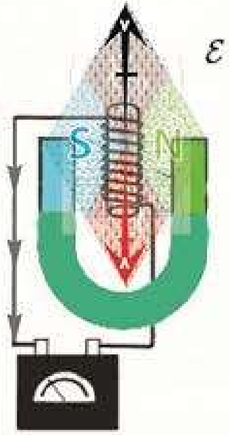

$$\mathcal{E} = -N\frac{d\Phi_B}{dt}.$$

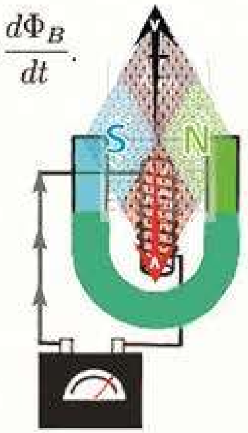

Coil moves DOWN

An electric current passes through a solenoid, resulting in a magnetic field. When you wrap your right hand around the solenoid with your fingers in the direction of the conventional current, your thumb points in the direction of the magnetic north pole.

An electric current passes through a straight wire. Here, the thumb points in the direction of the conventional current (from positive to negative), and the fingers point in the direction of the magnetic lines of flux.

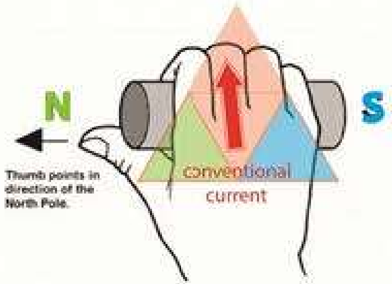

Fleming's right hand rule
(for EM induction)

Faraday's Law of Induction

Michael Faraday

(22 September 1791 – 25 August 1867)

Electromagnetic induction is the production of a potential difference (voltage) across a conductor when it is exposed to a varying magnetic field.

In the frame of a conductor moving relative to the magnet, charged particles in the conductor experience a coulombic force due to the neutralised electric fields of the permanent magnet.

Faraday's law states that:
The induced electromotive force (emf) in any closed circuit is equal to the time rate of change of the magnetic flux through the circuit.

$$\boxed{\oint \vec{E} \cdot \vec{ds} = -\frac{d\Phi_B}{dt}}$$

The motion of EM fields creates an induced motion in external fields and particles

The linear energy momenta of the neutralised E-fields produces differing directions of induced emf (current flow)

A permanent magnet has a static Magnetic dipole field and neutralised Electric field geometries

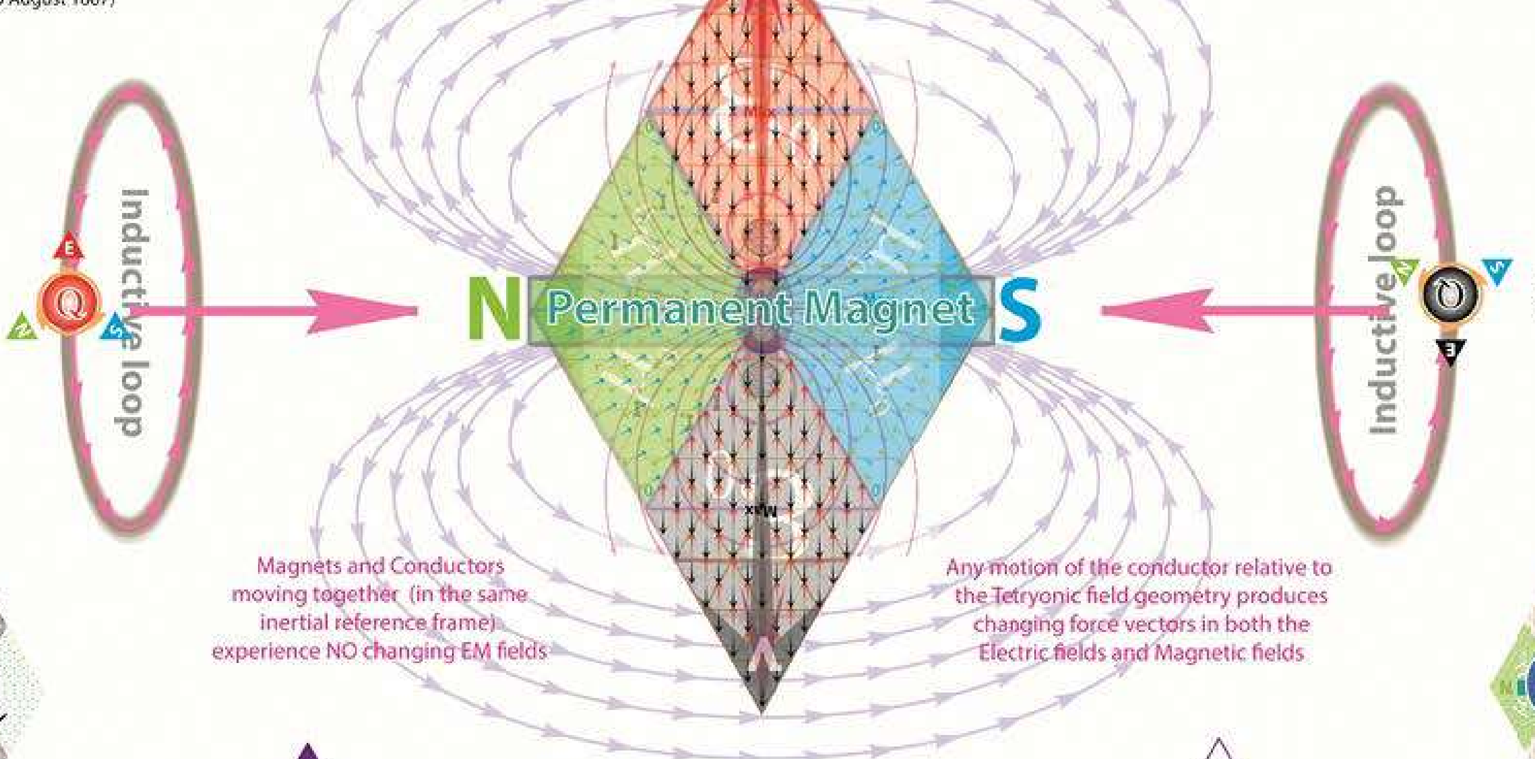

Charged particles in the conductor are subject to voltage emfs

Charged particles in the conductor are subject to voltage emfs

Magnets and Conductors moving together (in the same inertial reference frame) experience NO changing EM fields

Any motion of the conductor relative to the Tetryonic field geometry produces changing force vectors in both the Electric fields and Magnetic fields

An Electrostatic (charged) particle has Electric fields of a particular polarity and neutralised Magnetic fields

Electric Constant

It is impossible (due to the geometry of EM fields) to move either the permanent Magnet or the Conductor without producing variations in the Electric and Magnetic fields

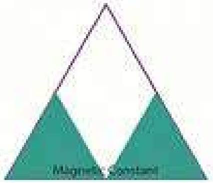

A Magneto-static body of Matter has a dipolar Magnetic field and neutralised Electric fields

Ampere's Law

In classical electromagnetism, Ampère's circuital law, relates the integrated magnetic field around a closed loop to the electric current passing through the loop.

Q

[v-v]

All energy momenta quanta have equilateral geometries, [quantised angular momenta [Ω]] that can be modelled as a continuous steady current within a closed inductive loop

$$\boxed{\oint \vec{B}\cdot d\vec{s} = \mu_0 i + \frac{1}{c^2}\frac{\partial}{\partial t}\int \vec{E}\cdot d\vec{A}}$$

The nett Charge of any particle or surface is the result of the quantised equilateral energy momenta symmetries of its integral surface area

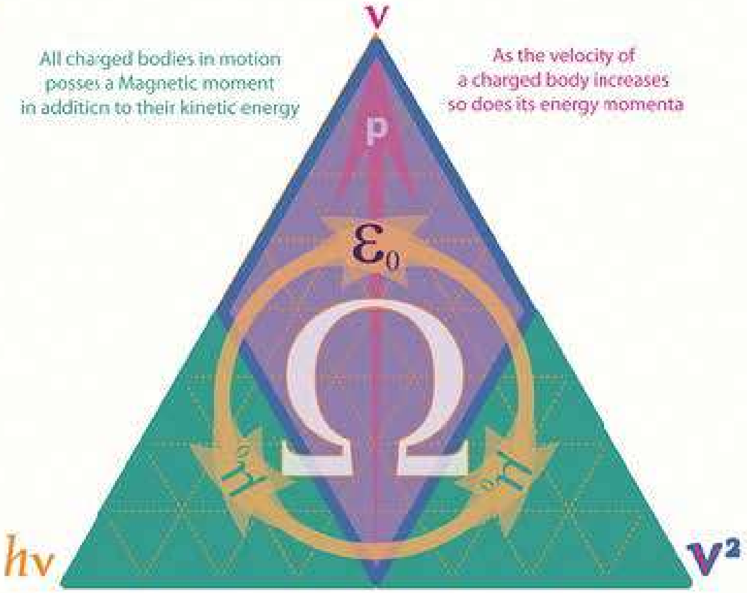
v
p
ε_0
Ω
μ_0
μ_0
hv
V²

All charged bodies in motion posses a Magnetic moment in addition to their kinetic energy

As the velocity of a charged body increases so does its energy momenta

$$\oint_C \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$$

$$\mathbf{J}_f + \mathbf{J}_D + \mathbf{J}_M = \mathbf{J}_f + \mathbf{J}_P + \mathbf{J}_M + \bigcirc\!\!\!\!\!\varepsilon_0\frac{\partial \mathbf{E}}{\partial t} = \mathbf{J} + \varepsilon_0\frac{\partial \mathbf{E}}{\partial t},$$

With the addition of Maxwell's Displacement current to account for time varying electric fields without a physical flow of charged Matter the way forward is paved for Planck, Lorentz and Tetryonics

The Closed circuit or loop can be any geometric shape with Tetryonics dictating equilateral [triangular] Planck energies & quantum charge geometries with Tetrahedral geometries for Matter quanta

Andre Ampere

(20 January 1775 – 10 June 1836)

Ampere's law relates magnetic fields to the electric currents that produce them.

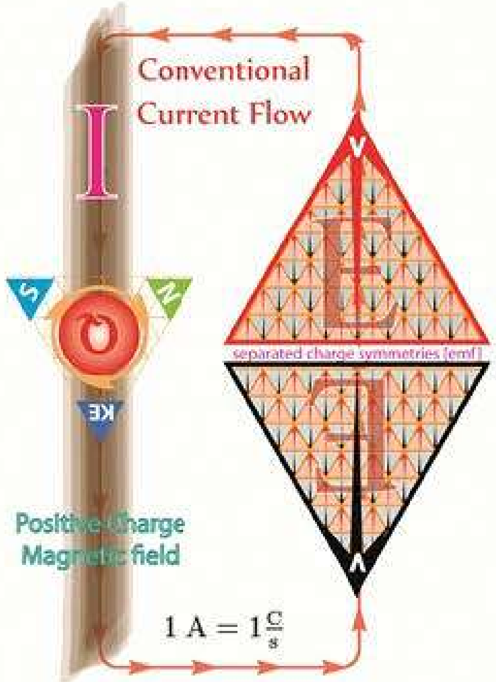
Conventional
Current Flow
I
S
N
KE
v
E
separated charge symmetries [emf]
E
v
Positive Charge
Magnetic field
$1\ \mathrm{A} = 1\frac{\mathrm{C}}{\mathrm{s}}$

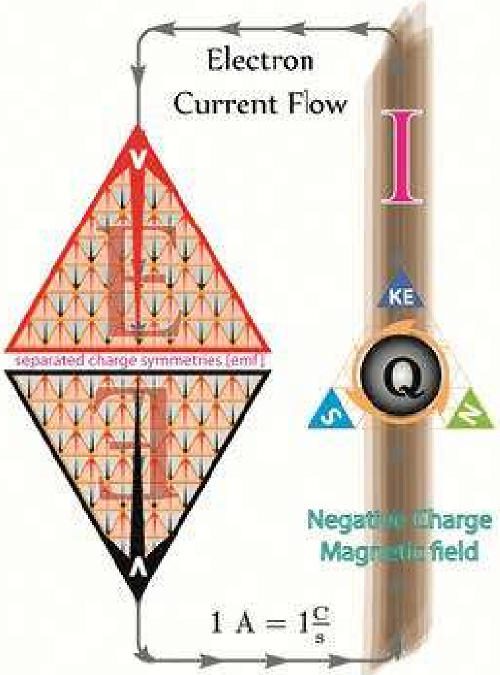
Electron
Current Flow
I
v
E
separated charge symmetries [emf]
E
v
KE
Q
S
N
Negative Charge
Magnetic field
$1\ \mathrm{A} = 1\frac{\mathrm{C}}{\mathrm{s}}$

$$F = 2k_A \frac{I_1 I_2}{r}$$

Ampère's force law states that there is an interactive force between two parallel wires carrying an electric current.

$$\mu_0 \overset{\text{def}}{=} 4\pi\,e{-}7\,\frac{N}{A^2}$$

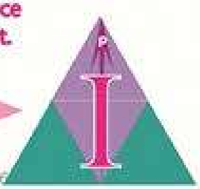

This force is used in the formal definition of the ampere which states that it is "the constant current which will produce an attractive force of 2 × 10–7 newtons per metre of length between two straight, parallel conductors of infinite length and negligible circular cross section placed one metre apart in a vacuum"

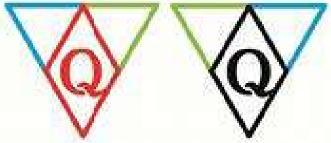

These attractions and repulsions between electric currents differ fundamentally from the effects produced by electricity in repose. First, they cease, as chemical decompositions do, as soon as we break the circuit. Second, in ordinary electric attractions and repulsions, opposite charges attract, and like charges repel; in the attractions and repulsions of electric currents, we have precisely the contrary; it is when the two conducting wires are placed parallel in such a way that their ends of the same sign are next to each other that there is attraction, and there is repulsion when the ends of the same sign are as far apart as possible. Third, in the case of attraction, when it is sufficiently strong to bring the movable conductor into contact with the fixed conductor, they remain attached to one another like two magnets, and do not separate after a while, as happens when two conducting bodies, oppositely electrified, come to touch.

André Marie Ampère (1775 - 1836)

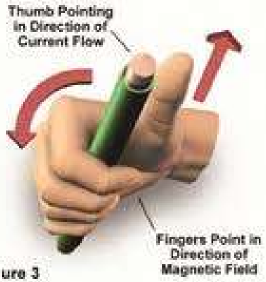
Thumb Pointing in Direction of Current Flow
Fingers Point in Direction of Magnetic Field

Figure 3

$$1\ \text{A} = 1\,\frac{\text{C}}{\text{s}}$$

Interactive force between Conductors

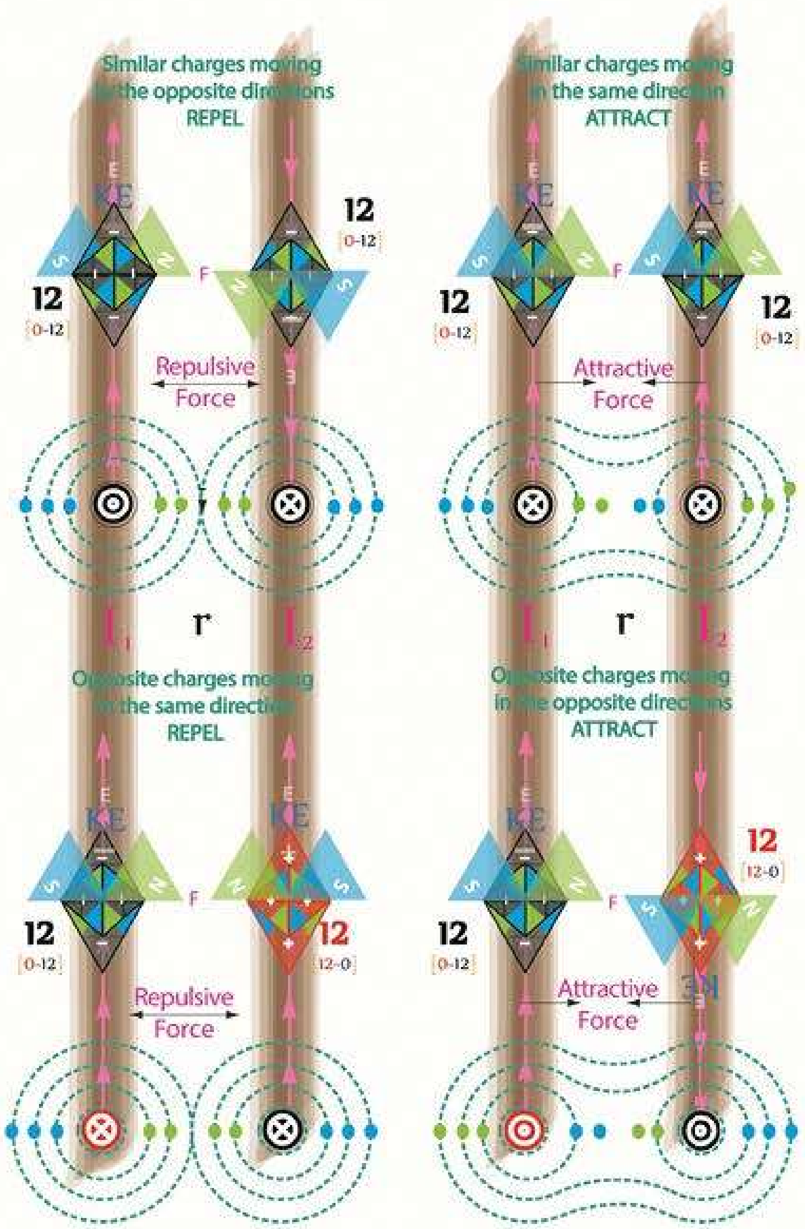
Similar charges moving in the opposite directions REPEL
Similar charges moving in the same direction ATTRACT
E
KE
S
N
F
12
[0-12]
12
[0-12]
Repulsive Force
12
[0-12]
12
[0-12]
Attractive Force
I1
r
I2
I1
r
I2
Opposite charges moving in the same direction REPEL
Opposite charges moving in the opposite directions ATTRACT
E
KE
12
[0-12]
12
[12-0]
Repulsive Force
12
[0-12]
12
[12-0]
Attractive Force

ElectroMotive exchange particles

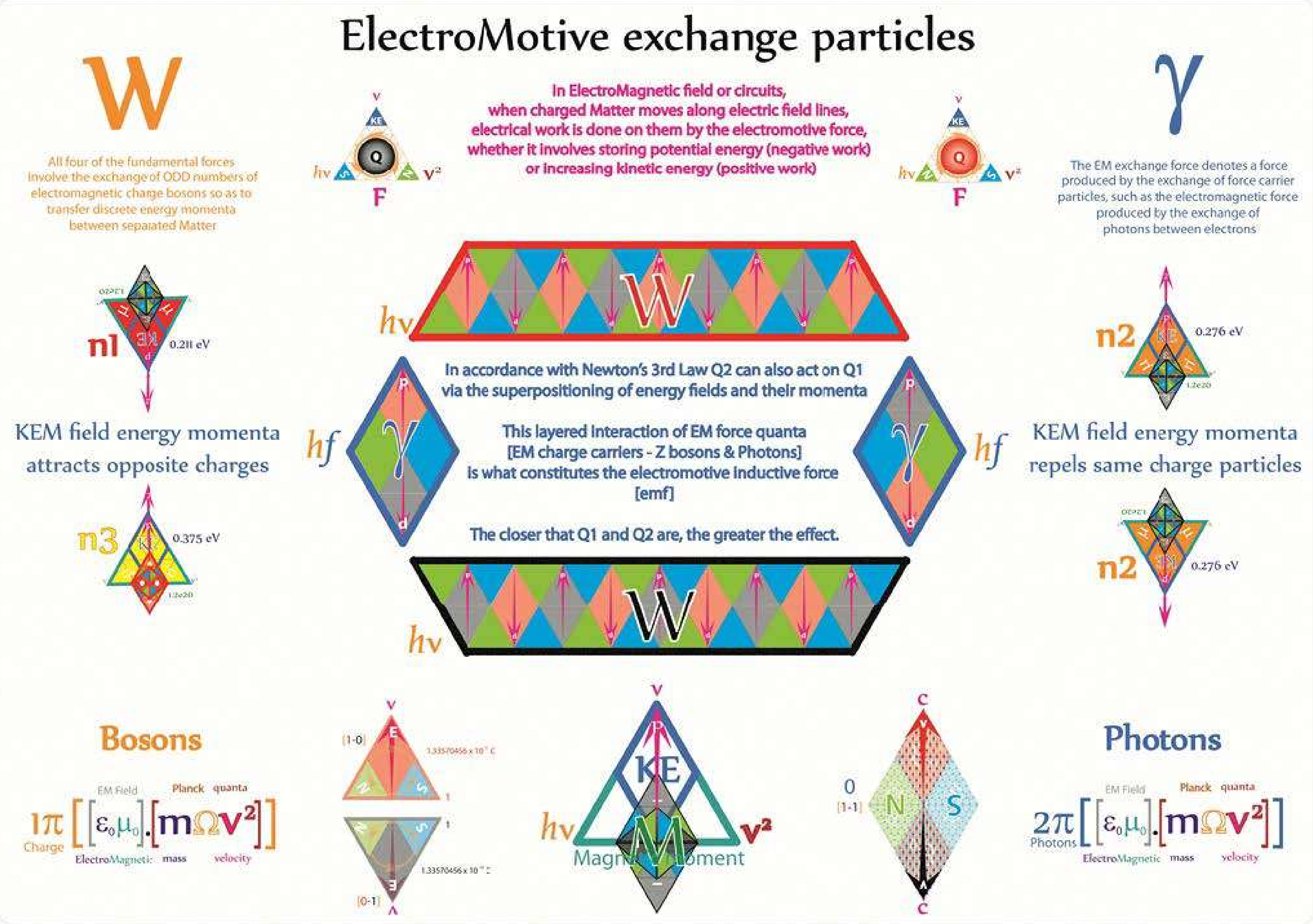

All four of the fundamental forces involve the exchange of ODD numbers of electromagnetic charge bosons so as to transfer discrete energy momenta between separated Matter

In ElectroMagnetic field or circuits, when charged Matter moves along electric field lines, electrical work is done on them by the electromotive force, whether it involves storing potential energy (negative work) or increasing kinetic energy (positive work)

The EM exchange force denotes a force produced by the exchange of force carrier particles, such as the electromagnetic force produced by the exchange of photons between electrons

In accordance with Newton's 3rd Law Q2 can also act on Q1 via the superpositioning of energy fields and their momenta

This layered interaction of EM force quanta [EM charge carriers - Z bosons & Photons] is what constitutes the electromotive inductive force [emf]

The closer that Q1 and Q2 are, the greater the effect.

KEM field energy momenta attracts opposite charges

KEM field energy momenta repels same charge particles

Bosons

EM Field Planck quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Charge ElectroMagnetic mass velocity

Photons

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

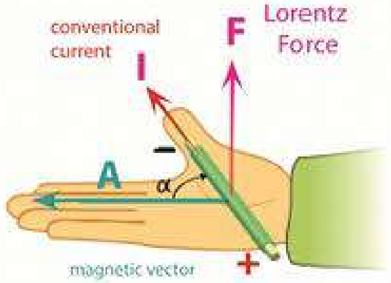
conventional current
I
F
Lorentz Force
A
α
−
+
magnetic vector

Lorentz Force

The magnetic force component of the Lorentz force manifests itself as the force that acts on a current-carrying wire in a magnetic field.

In that context, it is also called the Laplace force.

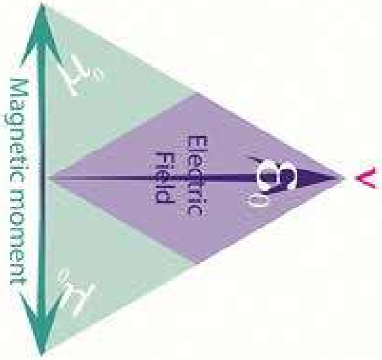
Magnetic moment
μ_0
μ_0
Electric Field
ε_0
v

Hendrik Lorentz

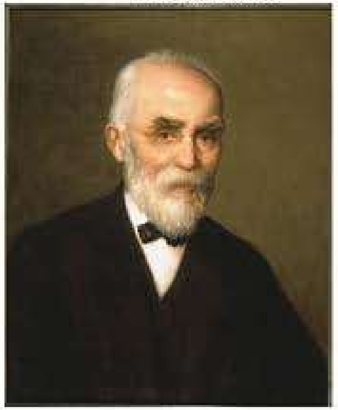

(18 July 1853 – 4 February 1928)

In physics, the Lorentz force is the force on a point charge due to electromagnetic fields.

It is given by the following equation in terms of the electric and magnetic fields

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$

LORENTZ force | *Electric force* | *charge velocity* | *Magnetic force*

Magnetic fields created by moving charges are perpendicular to the direction of motion (and can do NO work)

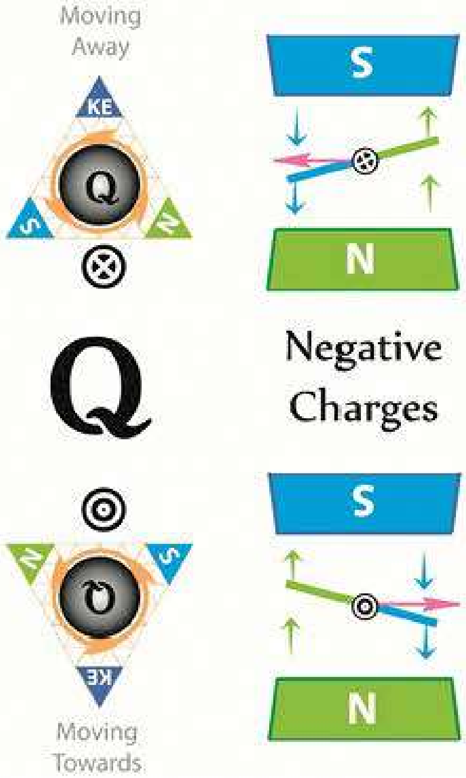
Moving Away
KE
Q
S
N
⊗
Q
◎
N
S
Q
KE
Moving Towards
S
N
S
N

Negative Charges

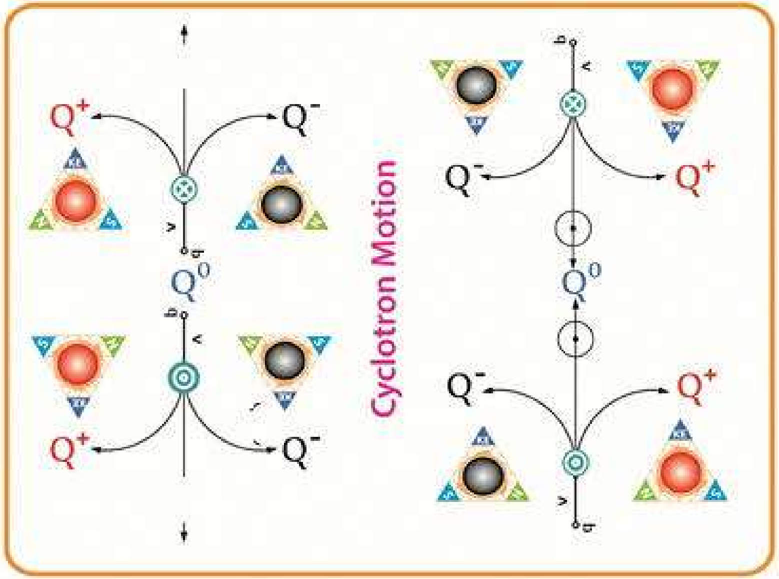
Q+
Q-
Q0
Q+
Q-
Cyclotron Motion
Q-
Q+
Q0
Q-
Q+

Positive Charges

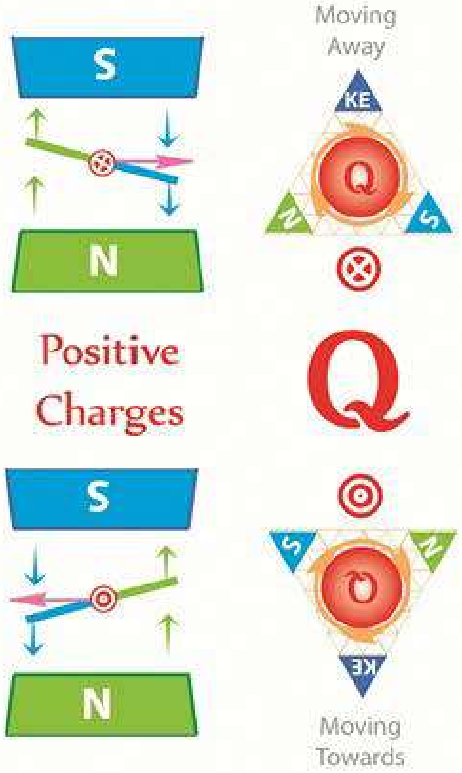
Moving Away
S
N
KE
Q
⊗
Q
◎
S
N
KE
Moving Towards

Up out of page ◎ Magnetic Field ⊗ Down into page

Copyright ABRAHAM [2008] - all rights reserved

Work - Force - Energy

Work is the result of energy-momenta transfer (by means other than Matter-transfer), and produces changes in the second system's nett energy-momenta

Work

refers to an activity involving a force and movement in the directon of the force

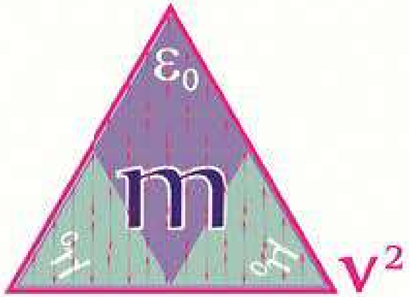

Work

$$\boxed{W = F.d = ma.d = E}$$

The Planck constant is the quantum of Action

$$kg\,\frac{m}{s^2}\,m$$

Mechanical work is a scalar quantity that can be described as the product of a force times the distance through which it acts

[Force per meter]

Force
$$\boxed{ma = \frac{\Delta p}{\Delta t} = \Delta\left[\frac{mv}{t}\right]}\quad kg\,\frac{m}{s^2}$$

Momentum is the quantum of Force

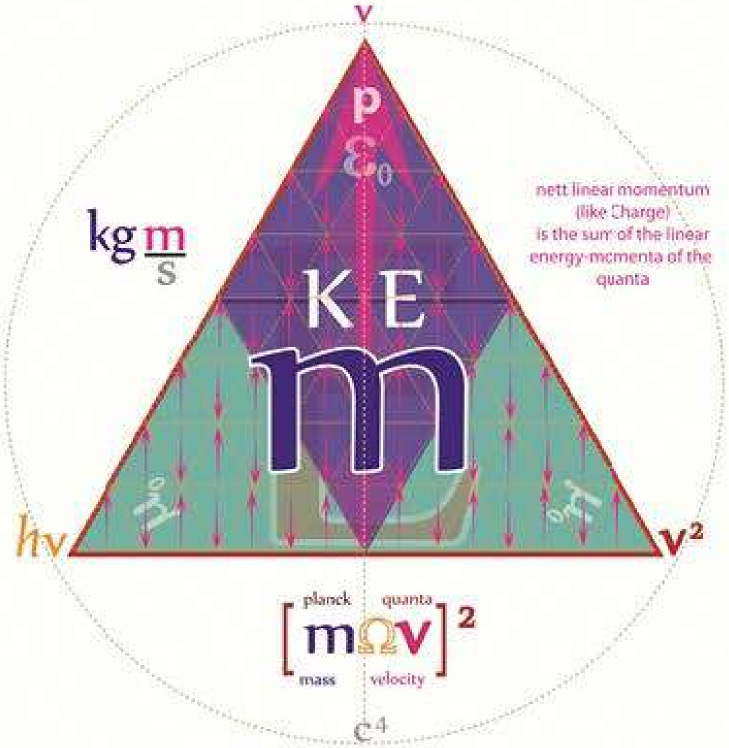

nett linear momentum (like Charge) is the sum of the linear energy-momenta of the quanta

Power is defined as the rate of using Energy or the rate of doing work

Energy

is the scalar capacity to do work
You must have energy to accomplish work

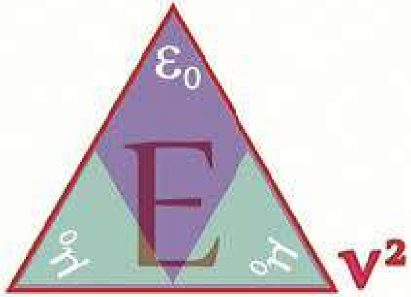

Energy

$$\boxed{E = mv^2 = hv^2}$$

Energy is the square of Momentum

$$kg\,\frac{m^2}{s^2}$$

Energy is always equivalent to the ability to exert pulls or pushes against the basic forces of nature, along a path of a certain length.

[mass-energy momentum squared]

Tetryonics 22.01 - Work, Force & Energy

Copyright ABRAHAM [2008] - all rights reserved

Watts

[Joules per second]

WORK

Force per second

In Classical Mechanics

One watt is the rate at which work is done when an object's velocity is held constant at one meter per second against constant opposing force of one newton.

Work in physics is measured in Joules

Nm/s

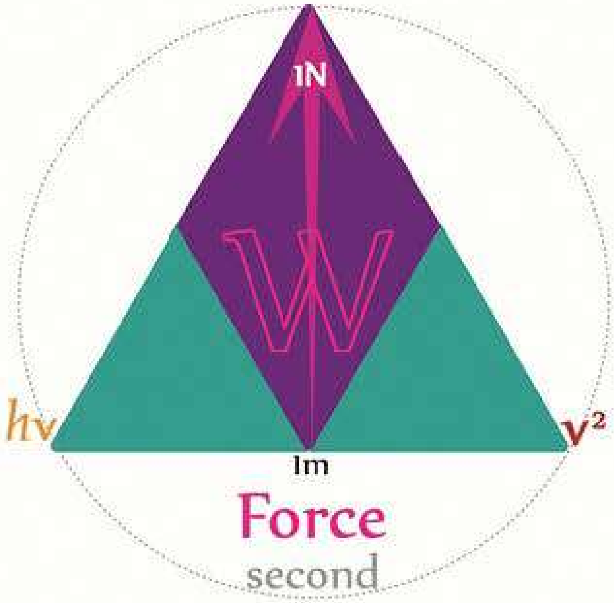

James Watt

(19 January 1736 – 25 August 1819)

The watt second is a unit of energy, equal to the joule

$$W = \frac{J}{s} = \frac{N \cdot m}{s} = \frac{kg \cdot m^2}{s^3}$$

One Watt, defined as one joule per second, measures the rate of energy conversion.

W=V.A

Energy / second
[Joules/sec]

POWER

Energy per second

In ElectroDynamics

One watt is the rate at which work is done when one ampere (A) of current flows through an electrical potential difference of one volt (V).

Electrical power is measured in Watts

J/s

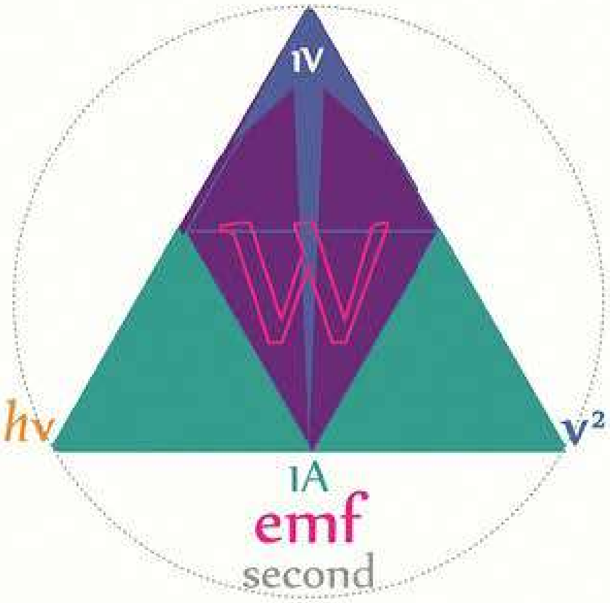

The electromotive Force

The electromotive force, or most commonly emf (seldom capitalized), or (occasionally) electromotance is "that which tends to cause current (actual electrons and ions) to flow.

Electric Fields contain bi-directional energy-momenta

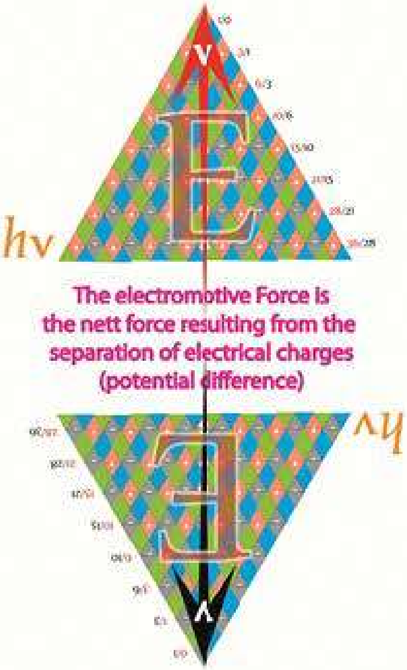

The electromotive Force is the nett force resulting from the separation of electrical charges (potential difference)

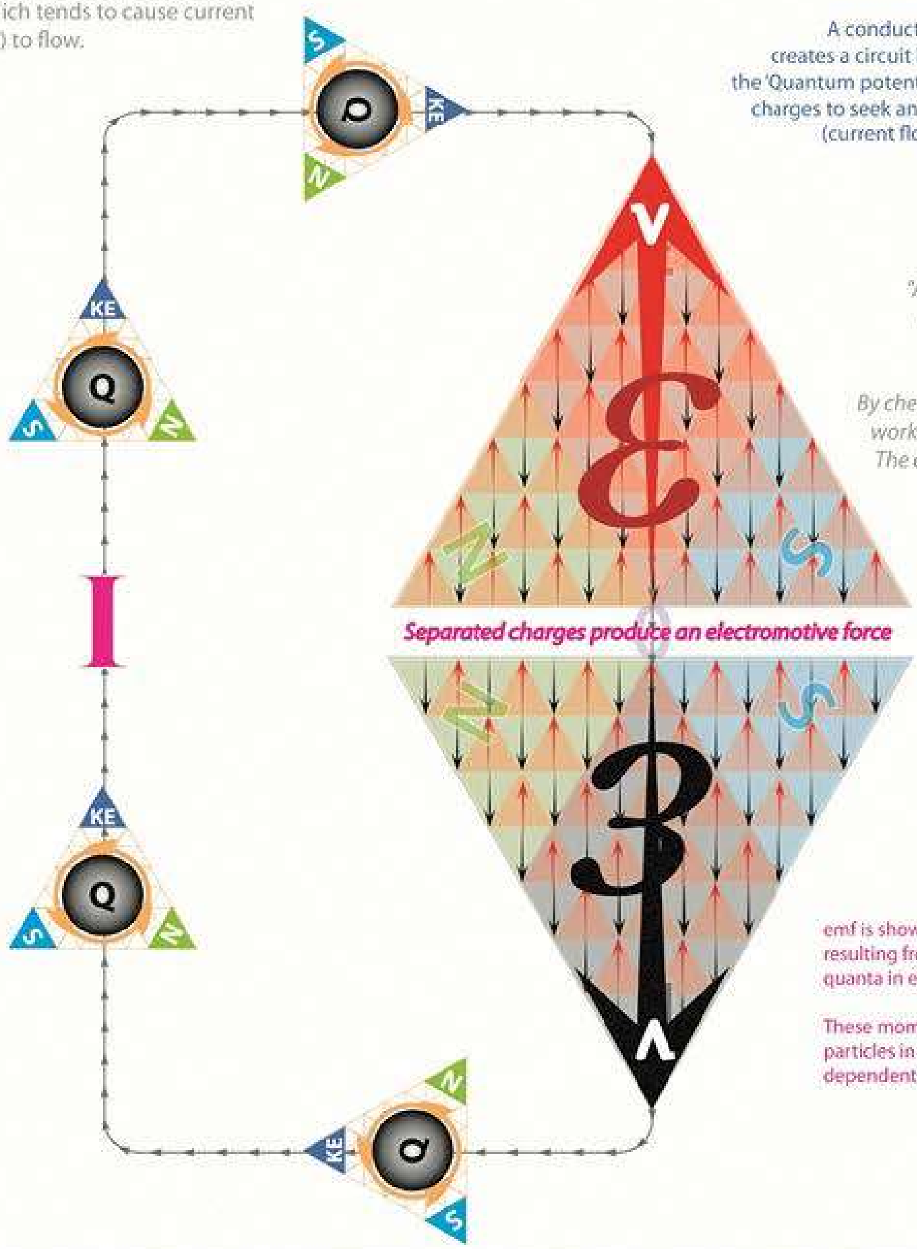

Separated charges produce an electromotive force

A conductor creates a circuit between the 'Quantum potentials' allowing charges to seek an eqilibrium (current flow)

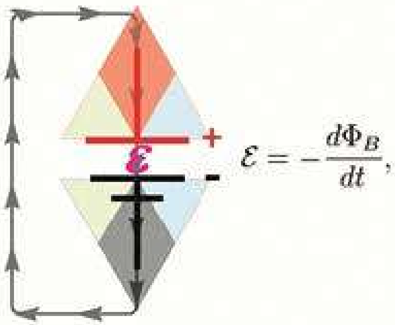

$$\mathcal{E} = -\frac{d\Phi_B}{dt},$$

"A source emf can be thought of as a kind of charge pump that acts to move charge from a point of high potential through its interior to a point of opposite potential. …

By chemical, mechanical or other means, the source of emf performs work dW on that charge to move it to the high potential terminal. The emf of the source is defined as the work dW done per charge dq: = dW/dq."

A Magnet stores charge quanta (bosons) in static Electro-Magnetic fields that in turn produce a Magnetic dipole

emf is shown to be the nett force resulting from the linear momenta quanta in every Electric 'voltage' field

These momenta accelerate charged particles in one of two directions dependent of the particle's nett charge

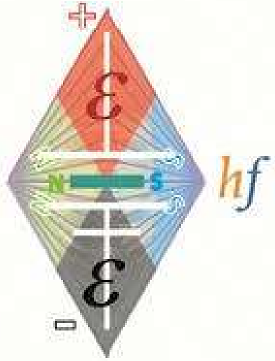

A Magnet can be viewed as a neutralised quantum emf source

All Charges seek Equilibrium

Copyright ABRAHAM [2008] - all rights reserved

Poynting Vectors

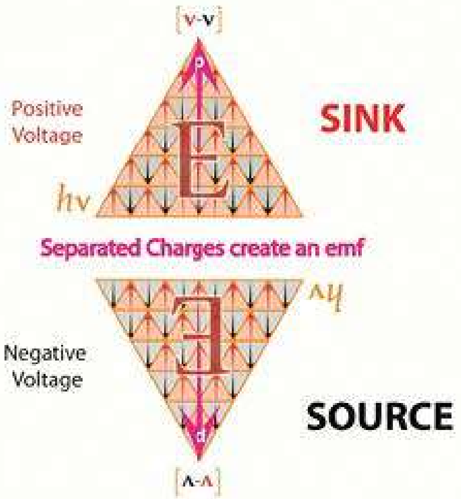
[v-v]
Positive Voltage
SINK
hv
E
Separated Charges create an emf
hv
Negative Voltage
SOURCE
[^-^]

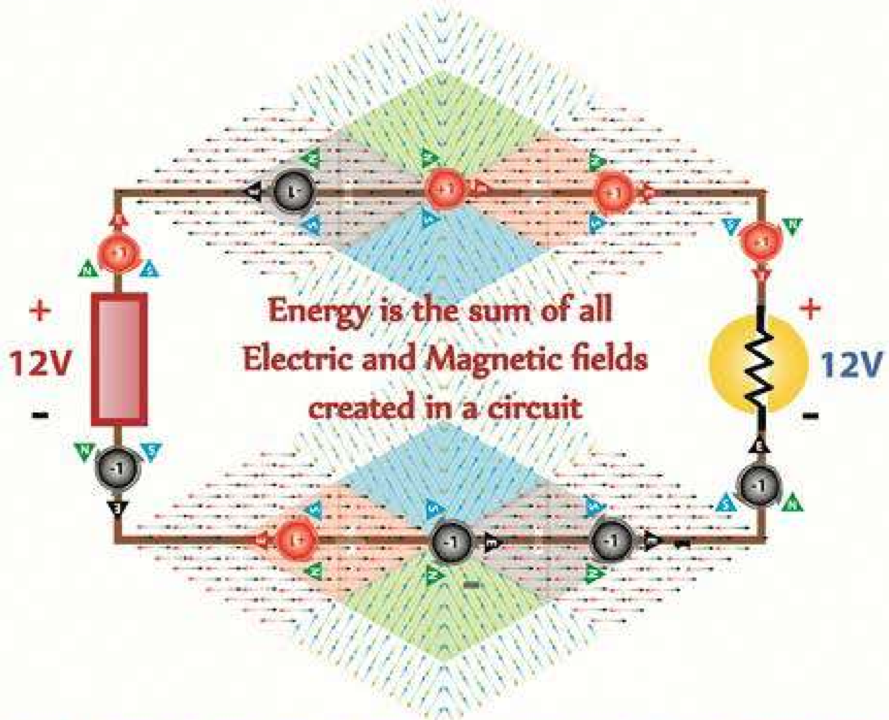

Energy is the sum of all Electric and Magnetic fields created in a circuit

Poynting Field

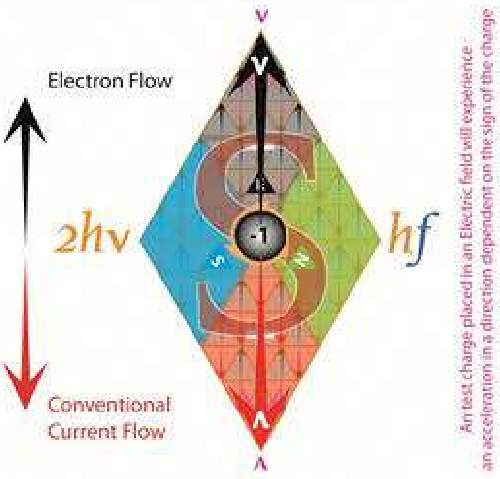
V
Electron Flow
2hv
hf
-1
Conventional Current Flow
^
An test charge placed in an Electric field will experience an acceleration in a direction dependent on the sign of the charge

John Henry Poynting

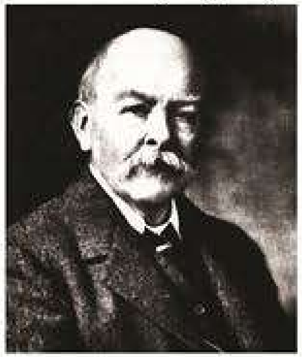

(9 September 1852 – 30 March 1914)

Electrical energy is transferred along a conductive path in an electrical circuit accelerating charged particles and ions in turn imbuing them with Kinetic Energy [voltage leads current]

There is an electric field running down the middle of the wire, which extends to just beyond its surface.

This electric field pushes the charges along against the resistance and adds to the electric field caused by the surface charges. The resultant electric field changes its direction around the circuit as the wires form a loop back to the battery.

Moving charged topologies inside the wire create a secondary magnetic field geometry around the wire

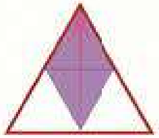

Electric fields

Energy

$$S = E \times H$$

momenta

Magnetic fields

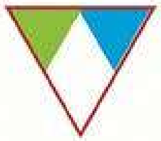

Electrons form a conductive path for Energy to flow along

The Electron drift velocity is much less than the speed of electricity(c)

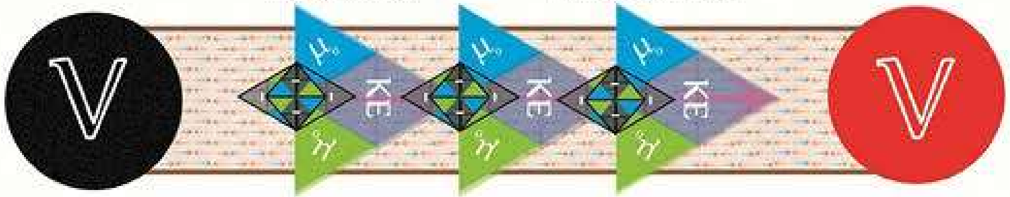
V
μ_0
KE
ϵ_0
V

Despite the very low 'drift' velocity of the Electrons they generate a large Magnetic field as a result of their motion

Energy is transferred through empty space around (and NOT in) the wires of an electric circuit via an electromagnetic field called the Poynting field

It's important to remember that the current doesn't flow in both directions, only the energy does

Copyright ABRAHAM [2008] - all rights reserved

Coulomb's Electric Interactions

$$k = \frac{1}{4\pi\varepsilon_0}$$

The proportionality constant ke, called the Coulomb constant (sometimes called the Coulomb force constant), is related to defined properties of linear EM energy momentum and is used to define Electric field forces

$$c = \frac{1}{\sqrt{\mu_0\varepsilon_0}}$$

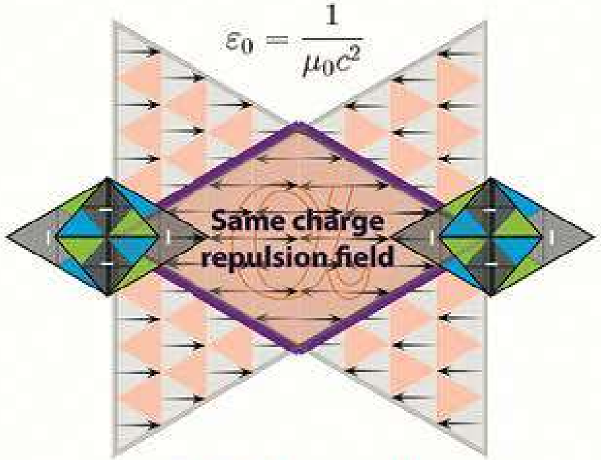

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

Similars repel

Linear Coulombic force interactions are a result of charged E field linear momenta

$$\mathbf{E} = \frac{\mathbf{F}}{q_t}$$

The Electric field can be defined by the Force exerted by a Charge

$$8.987\text{ e}9\ \frac{\text{Nm}^2}{\text{C}^2}$$

Opposites attract

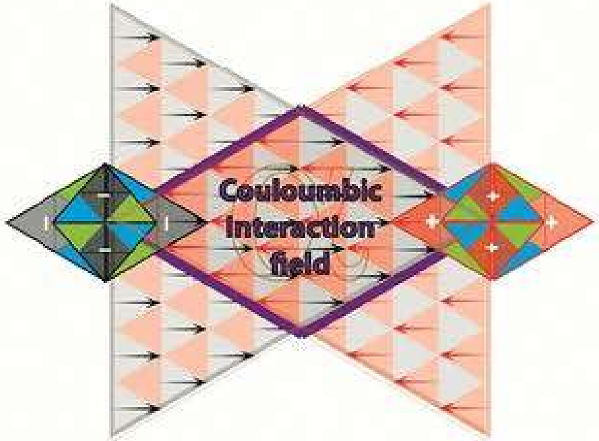

It is a measure of the interactive force produced by the Electric field energy-momenta of two superpositioned charge KEM fields

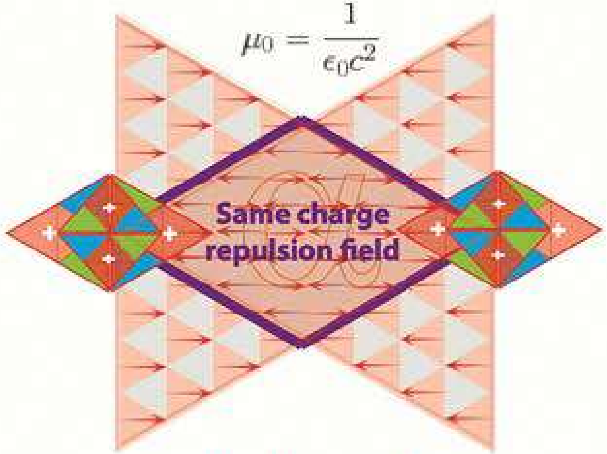

$$\mu_0 = \frac{1}{\epsilon_0 c^2}$$

Similars repel

Longitudinal E field forces between Charged particles are mediated by Photons

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}$$

The Electric field can also be derived from Coulomb's Law

Copyright ABRAHAM [2008] - all rights reserved

Coulomb Force

Electric fields produce Coulombic forces

$$F = k\frac{Q_1 Q_2}{r^2}$$

It is defined as the nett Charge transported by a steady current of one Ampere in one second.

C

1 Coulomb of nett charge is comprised of

$(12 \times 1.335180067\ e\text{-}20)^{-1}$

6.240355408 e18 electrons

kg.s

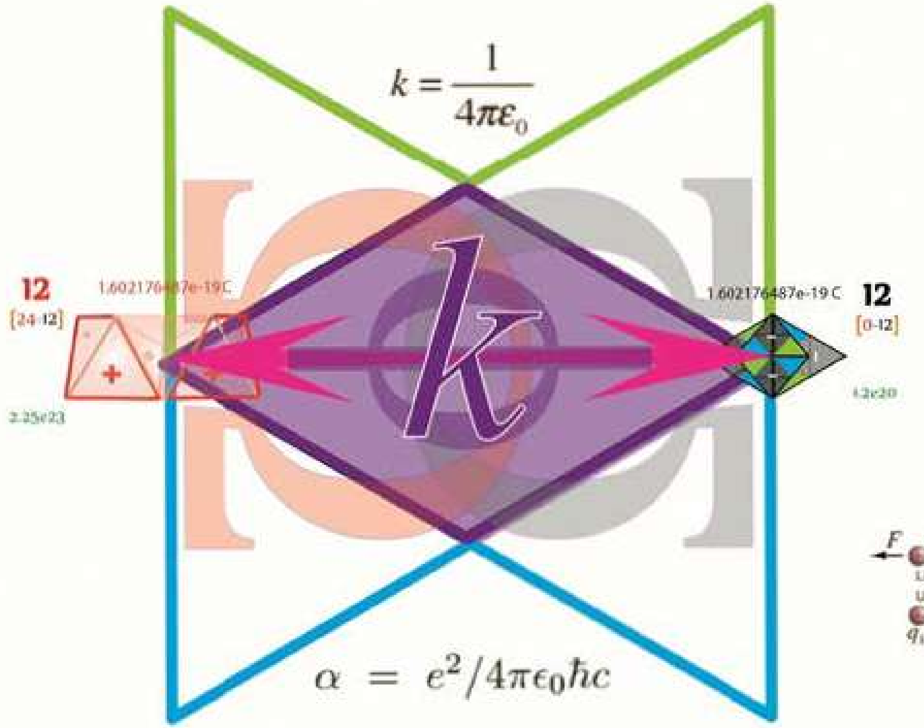
12
[24-12]
1.602176487e-19C
2.25e23
1.602176487e-19 C
12
[0-12]
1.2e20
k

$$k = \frac{1}{4\pi\varepsilon_0}$$

$$\alpha = e^2/4\pi\epsilon_0\hbar c$$

Charles-Augustin de Coulomb

(14 September 1736 – 23 August 1806)

Coulomb's Law is a law of physics describing the electric interaction between any two charged particles (and forms the basis for Ampere's Law)

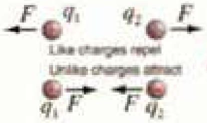
F
q_1
q_2
F
Like charges repel
Unlike charges attract
q_1
F
F
q_2

$$F = \frac{kq_1q_2}{r^2} = \frac{q_1q_2}{4\pi\varepsilon_0 r^2}$$ *Coulomb's Law*

The interactive linear momenta of super-positioned E-fields creates Coulombic forces

Coulombs

As

$1\text{C} = 1\text{A} \cdot 1\text{s}$

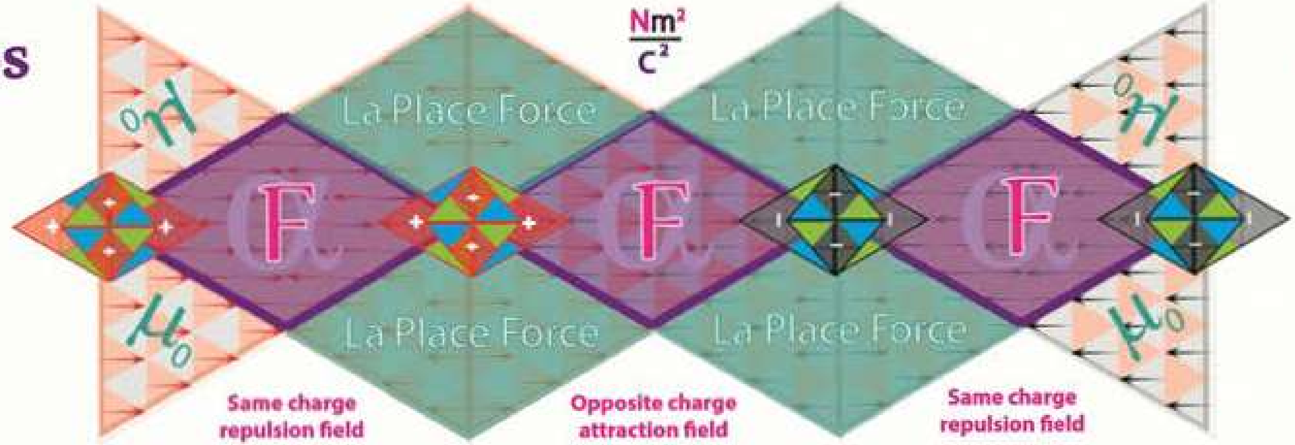
$\frac{Nm^2}{C^2}$
La Place Force
La Place Force
ε_0
ε_0
F
F
F
La Place Force
La Place Force
μ_0
μ_0

Same charge repulsion field

Opposite charge attraction field

Same charge repulsion field

Amperes

$$\frac{C}{s}$$

$$1\,\text{A} = 1\frac{\text{C}}{\text{s}}$$

Charles-Augustin de Coulomb

(14 September 1736 – 23 August 1806)

Coulomb force

$$k_c \frac{Q_1 Q_2}{r^2}$$

As

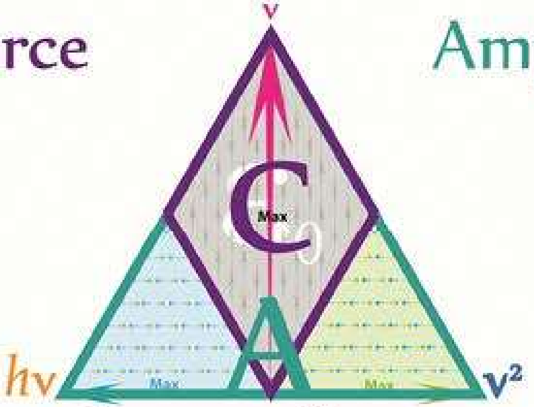
v
C
Max
A
hv
Max
Max
v²

Lorentz Forces

are the forces on a point charge due to external electromagnetic fields

$$\vec{F} = q[\vec{E} + (\vec{v} \times \vec{B})]$$

electric force

velocity

magnetic forces do not accelerate charged particles

Ampere force

$$2k_A \frac{I_1 I_2}{r}$$

$$\frac{C}{s}$$

Andre Ampere

(20 January 1775 – 10 June 1836)

Linear acceleration Electric force between charged particles

Q

Same charges

Q

Opposite charges

Q

Same charges

Q

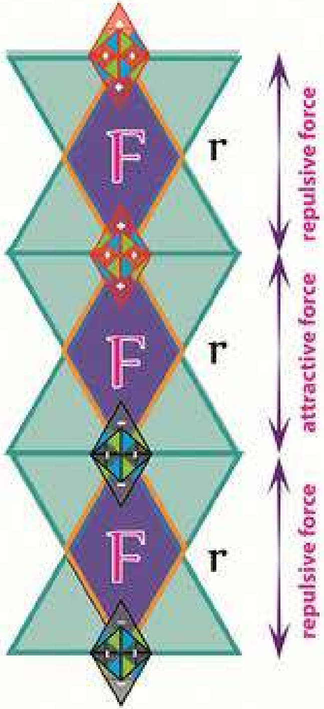
F
r
repulsive force
F
r
attractive force
F
r
repulsive force

I

The amount of charged mass-Matter topologies moving passing a point per unit time

Current

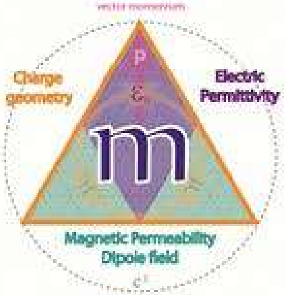
vector momentum
P
Charge geometry
Electric Permittivity
ε
m
Magnetic Permeability Dipole field
c²

Charge

The angular mass momenta possessed by a steady current of one ampere in one second.

Q

Transverse perturbative Magnetic force between charged particles

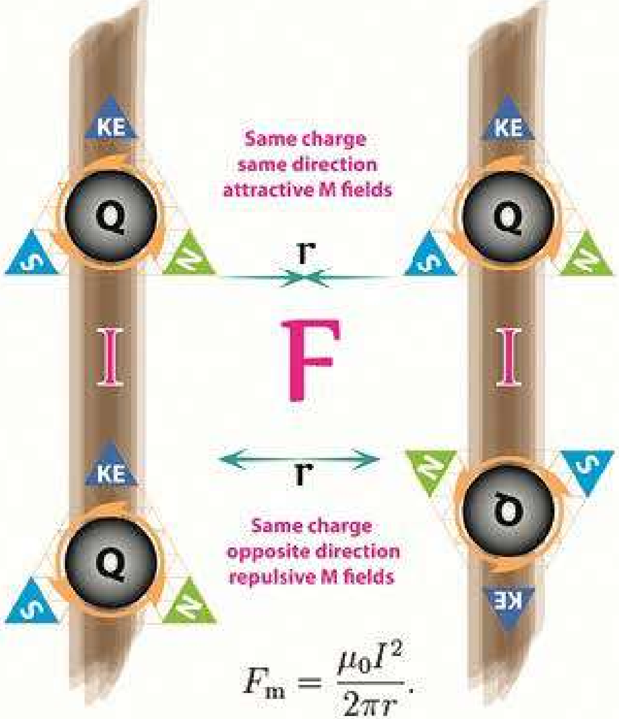
KE
Q
S
N
I
r
F
KE
Q
S
N
I
r
KE
Q
S
N
I
KE
Q
N
S
I

Same charge same direction attractive M fields

Same charge opposite direction repulsive M fields

$$F_m = \frac{\mu_0 I^2}{2\pi r}.$$

Capacitance

Any two electrical conductors separated by a non-conducting (or very high resistance) medium is a capacitor
[these can be plates, conductive wires or coaxial cables etc]

Capacitance is a function only of the physical dimensions (geometry) of conductors and the permittivity of the dielectric separating them

Plate capacitors

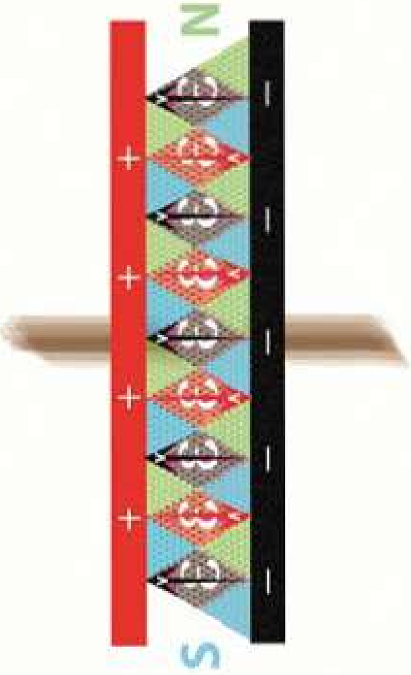

separated conductive cables

In general any separated charges will create a emf voltage

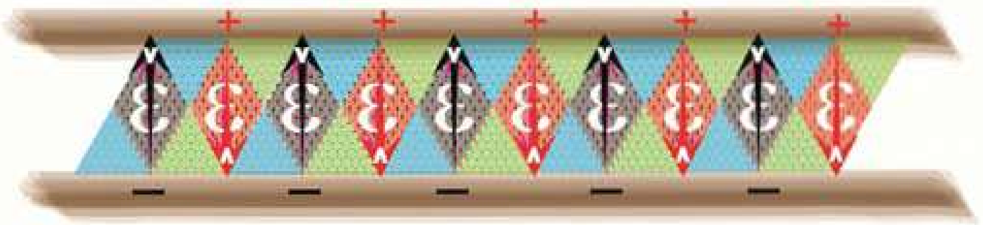

should a current flow in the conductors the capacitive field will be dominated by perpendicular, superpositioned amperian M-fields of greater strength

Coaxial cables

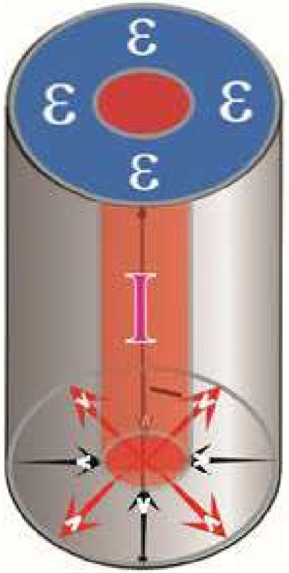

The SI unit of capacitance is the farad (symbol: F)
named after the English physicist Michael Faraday;

Capacitance is the ability of a body to store an electrical charge.

$$C = \frac{Q}{V}$$

Any object that can be electrically charged exhibits capacitance

A 1 farad capacitor when charged with1 coulomb of electrical charge will have a potential difference of 1 volt between its plates.

Copyright ABRAHAM [2008] - all rights reserved

Amperian Forces

Currents produce magnetic field forces

$$F = 2k_A \frac{I_1 I_2}{r}$$

6.241335 e18 electrons passing a given point per second constitutes one Ampere.

Coulombs are charged masses

m

that produce accelerations

Andre Ampere

(20 January 1775 – 10 June 1836)

Current is charged Matter in motion

M

that produce magnetic fields

A

1 Amp of electrons has a EM mass of

6.241355408 e18 x 8.851486361 e-31

5.524527227 e-12 kg

kg

An Ampere is a Coulomb per second.
a measure of the rate at which charged Matter moves.

A coulomb is the quantity of charge transported by one amp in one second

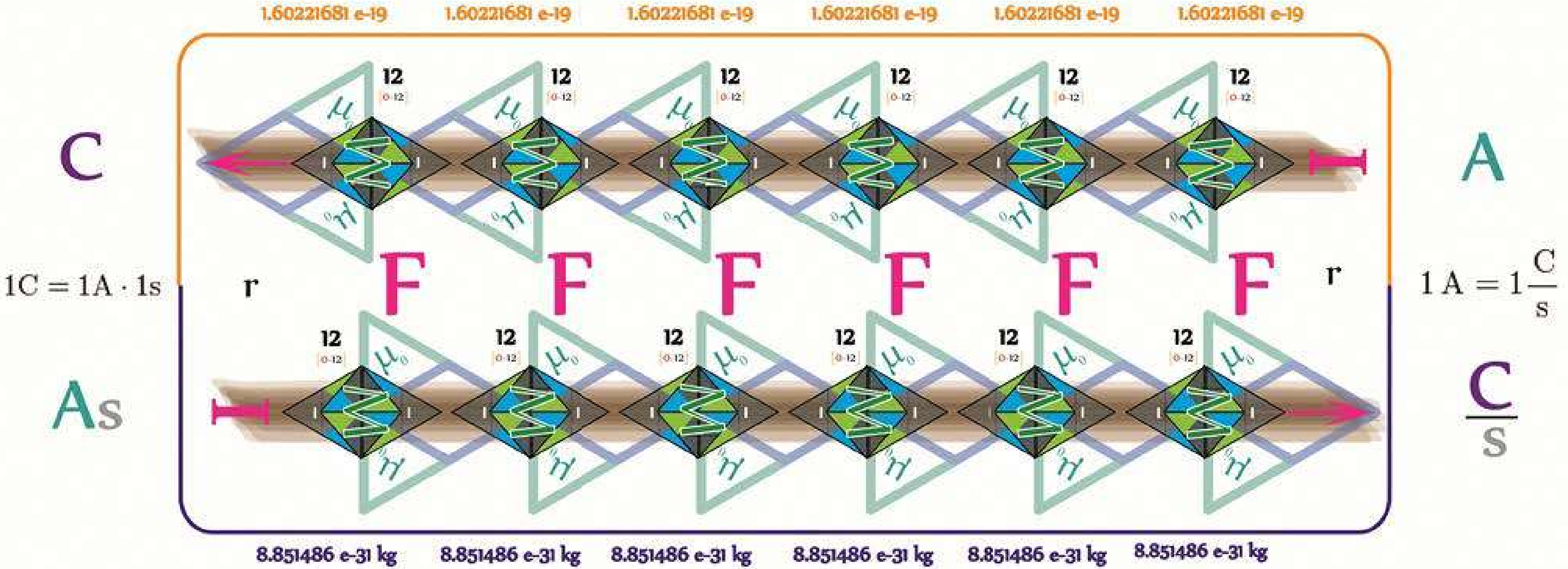
1.60221681 e-19
1.60221681 e-19
1.60221681 e-19
1.60221681 e-19
1.60221681 e-19
1.60221681 e-19
12
0-12
μ0
ε0
C
A
1C = 1A · 1s
r
F
F
F
F
F
F
r
1 A = 1 C/s
As
C/s
8.851486 e-31 kg
8.851486 e-31 kg
8.851486 e-31 kg
8.851486 e-31 kg
8.851486 e-31 kg
8.851486 e-31 kg

An ampere is one coulomb per second

Copyright ABRAHAM [2008] - all rights reserved

Ampere's Magnetic Interactions

the force of attraction or repulsion between two current-carrying wires (see first figure below) is often called Ampère's force law.

The best-known and simplest example of Ampère's force law, which underlies the definition of the ampere, the SI unit of current, states that the force per unit length between two straight parallel conductors is

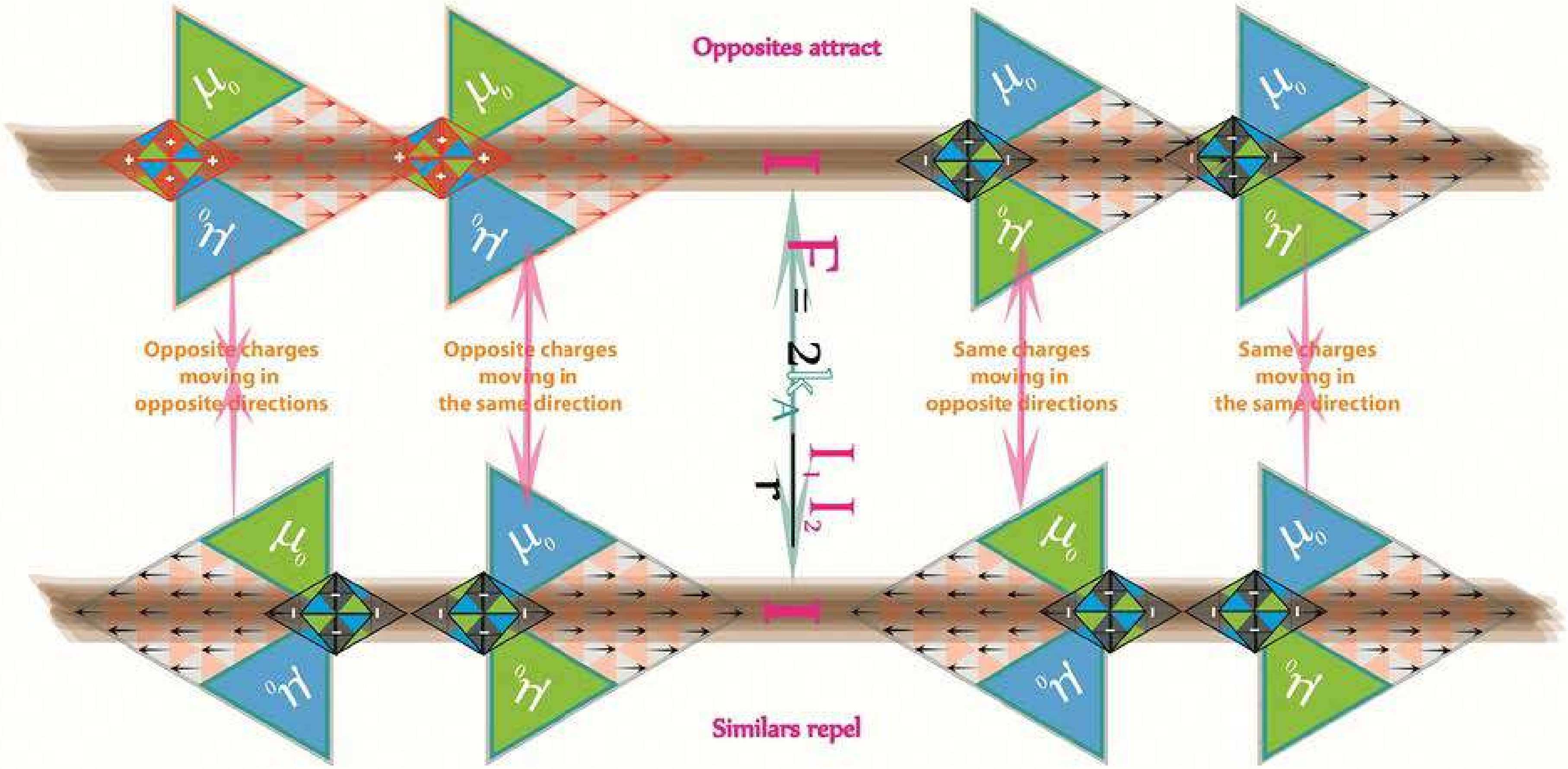

$$F = 2k_A \frac{I_1 I_2}{r}$$

Opposites attract

Opposite charges
moving in
opposite directions

Opposite charges
moving in
the same direction

Same charges
moving in
opposite directions

Same charges
moving in
the same direction

Similars repel

It is a measure of the interactive force
produced by the Magnetic field energy-momenta
of two superpositioned KEM field geometries

2D Electric field geometries accelerate charged particles

Lorentz Forces

If a particle of charge q moves with velocity v in the presence of
an electric field E and a magnetic field B, then it will experience a force

3D Matter topologies in motion create magnetic moments

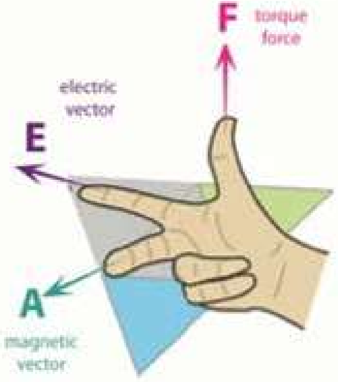
F
torque force
electric vector
E
A
magnetic vector

charged particles
in motion produce KEM fields
with magnetic moments

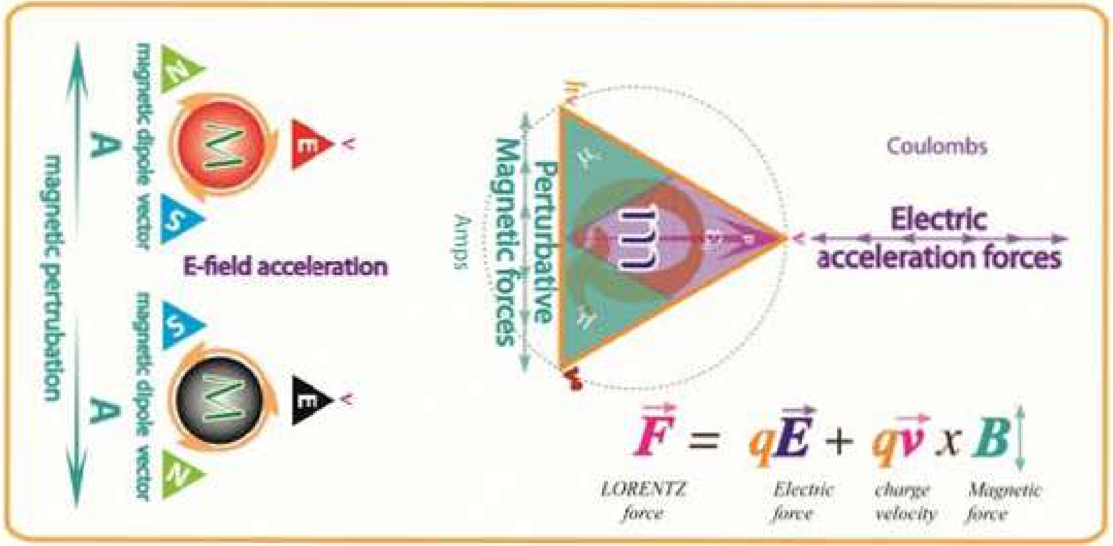
magnetic pertrubation
A
A
magnetic dipole vector
magnetic dipole vector
N
S
S
N
M
M
E
E
v
v
E-field acceleration
Amps
Perturbative
Magnetic forces
Coulombs
Electric
acceleration forces
LORENTZ force
Electric force
charge velocity
Magnetic force

$$\vec{F} = q\vec{E} + q\vec{v} \; x \; B$$

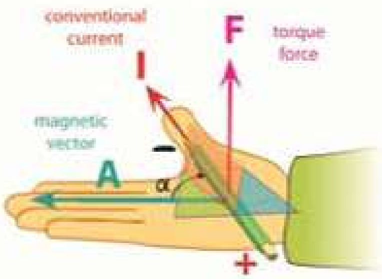
conventional current
I
F
torque force
magnetic vector
A
α
−
+

external magnetic fields
create torque forces on the
M dipoles of KEM fields

The Lorentz force is the force on a charged particle due to external electromagnetic fields.

Charles-Augustin de Coulomb

(14 September 1736 – 23 August 1806)

Magnetic field upwards through paper ⊙

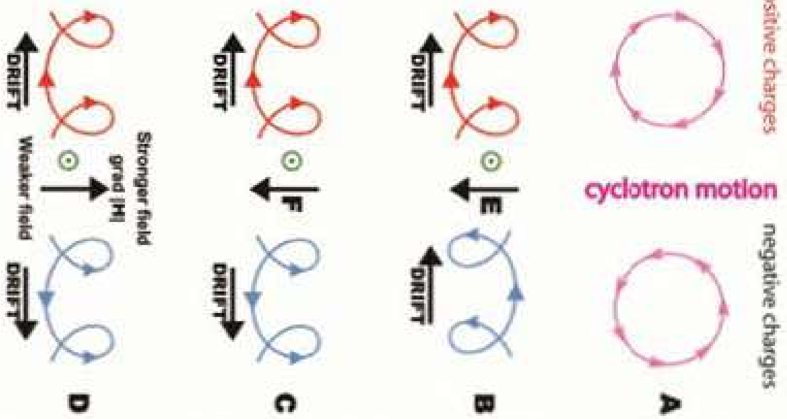
DRIFT
DRIFT
DRIFT
DRIFT
DRIFT
DRIFT
Weaker field
Stronger field
grad |H|
F
E
positive charges
cyclotron motion
negative charges
D
C
B
A

All charges in motion are subject to external EM forces or their KEM fields

Andre Ampere

(20 January 1775 – 10 June 1836)

Neutral Particle interactions

Opposites Attract

Classical Forces of interaction
are unable to explain
how Neutral particles
are attracted to other particles

Similars Repel

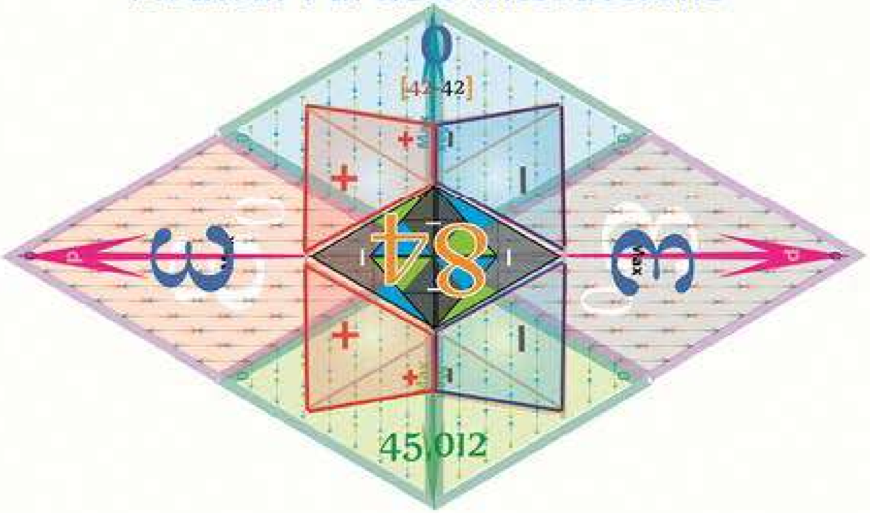
0
[42-42]
84
ε
ε
+
-
45,012

Tetryonic charge geometry

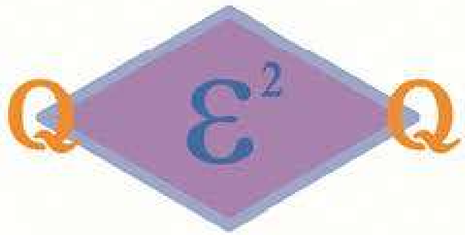
Q
ε^2
Q

reveals the equilateral field mechanics behind
EM induction and all the physical forces

All Matter topologies are comprised of charged mass-energy quanta
that effect atomic interactions via their KEM field energy momenta

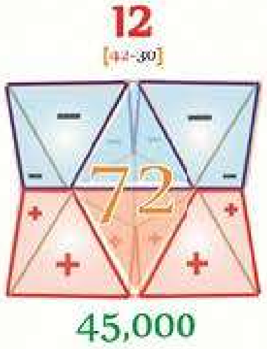
12
[42-30]
72
-
+
45,000

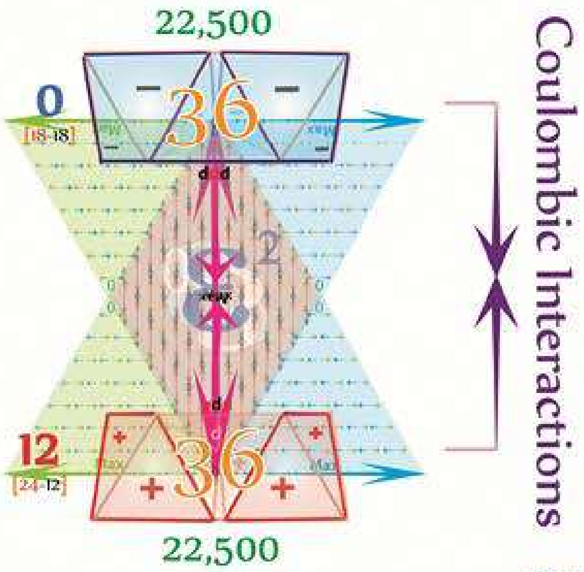
22,500
0
[18-18]
36
-
12
[24-12]
36
+
22,500
Coulombic Interactions

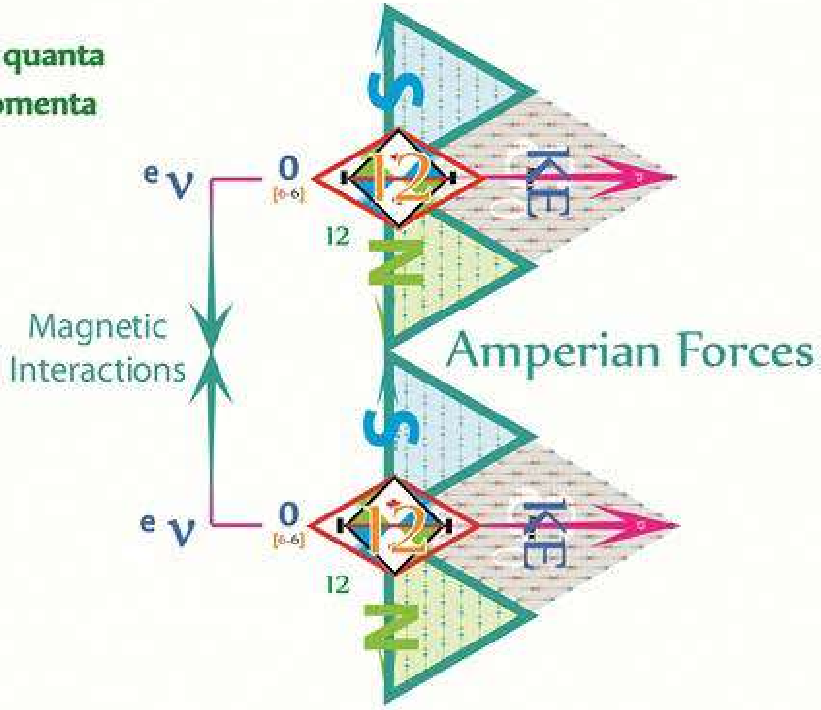
$^e\nu$
0
[6-6]
12
12
KE
S
N
Magnetic
Interactions
Amperian Forces
$^e\nu$
0
[6-6]
12
12
KE
S
N

All Matter in motion produces a secondary interactive KEM field

Copyright ABRAHAM [2008] - all rights reserved

Charged mass-Matter

All particles and fields have charged mass-energy geometries

Coulombs

C

kg.s

mass seconds

$$1\text{C} = 1\text{A} \cdot 1\text{s}$$

An elementary charge is 1.602216081 e-19 C

Amperes

A

KG

mass-Matter per second

$$1\,\text{A} = 1\frac{\text{C}}{\text{s}}$$

6.241355408 e18 electrons /sec constitutes one ampere.

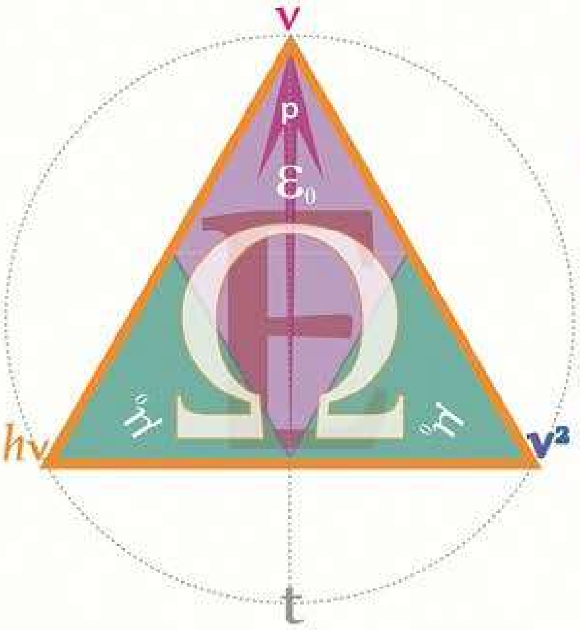

Charge is measured by the linear forces produced by the 2D equilateral geometry of mass-energy momenta

2D kEM angular mass-energy momenta

It is defined as the charge provided by a steady current of one ampere in one second.

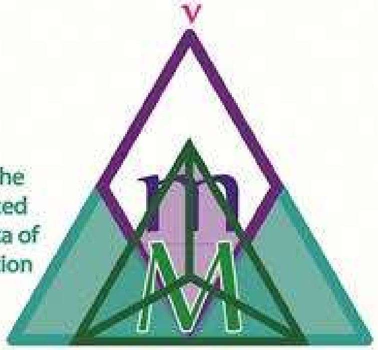

Current is measured by the transverse forces produced by mass-energy momenta of Matter topologies in motion per unit of time

3D mass-Matter momentum

The amount of charged mass- Matter passing a point per unit time

CHARGE

One coulomb of charge flowing per second equals one ampere of current.

CURRENT

$$Q = I \,.\, t$$

Charge and Current are related to each other through time

Charge

Q

Current

I

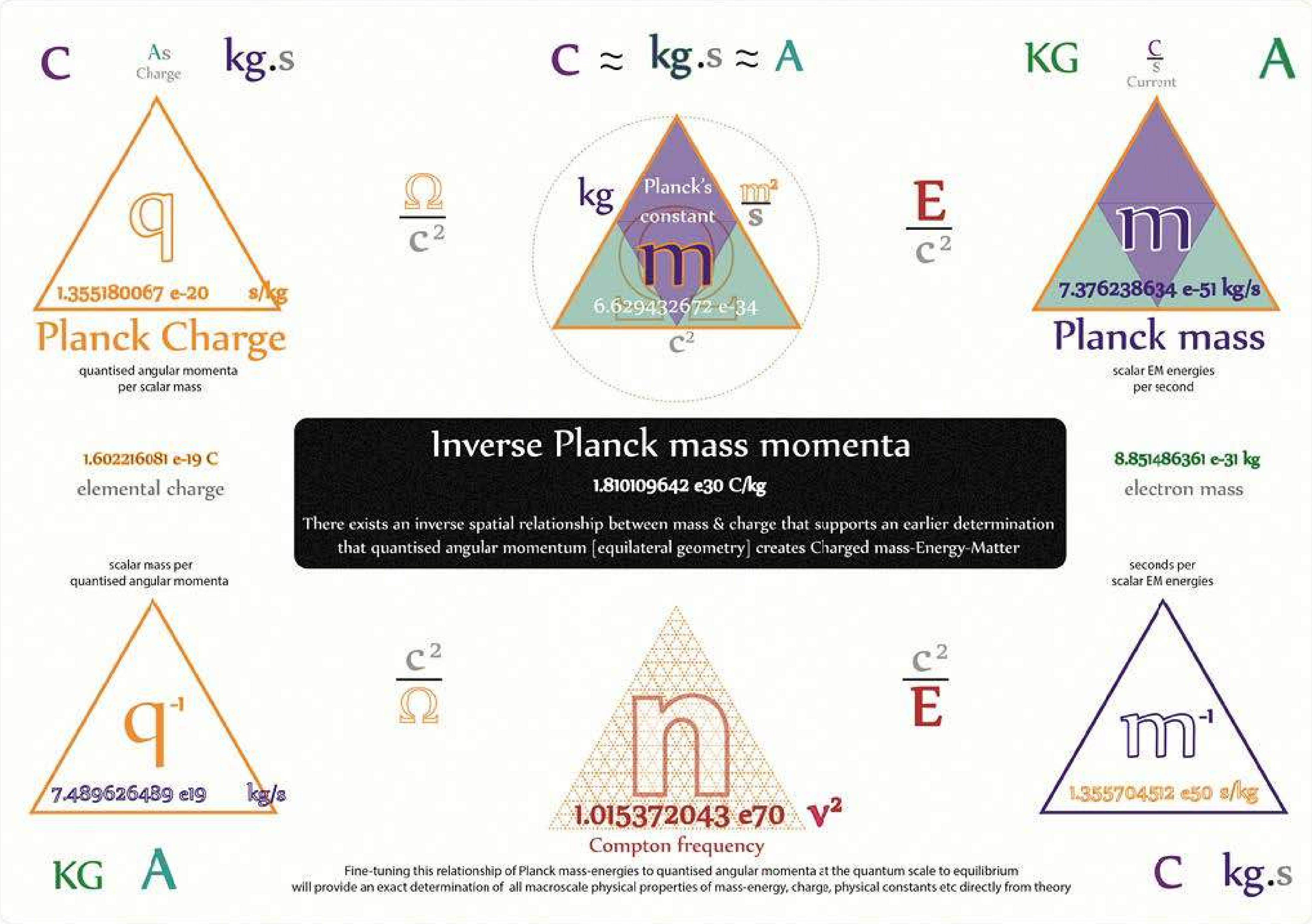
C
As
Charge
kg.s
C ≈ kg.s ≈ A
KG
C
s
Current
A
q
1.355180067 e-20 s/kg
Planck Charge
quantised angular momenta
per scalar mass
Ω
c²
kg
Planck's
constant
m²
s
m
6.629432672 e-34
c²
E
c²
m
7.376238634 e-51 kg/s
Planck mass
scalar EM energies
per second
1.602216081 e-19 C
elemental charge
8.851486361 e-31 kg
electron mass
scalar mass per
quantised angular momenta
seconds per
scalar EM energies
q⁻¹
7.489626489 e19 kg/s
c²
Ω
n
1.015372043 e70 v²
Compton frequency
c²
E
m⁻¹
1.355704512 e50 s/kg
KG A
C kg.s

Inverse Planck mass momenta

1.810109642 e30 C/kg

There exists an inverse spatial relationship between mass & charge that supports an earlier determination that quantised angular momentum [equilateral geometry] creates Charged mass-Energy-Matter

Fine-tuning this relationship of Planck mass-energies to quantised angular momenta at the quantum scale to equilibrium will provide an exact determination of all macroscale physical properties of mass-energy, charge, physical constants etc directly from theory

Copyright ABRAHAM [2008] - all rights reserved

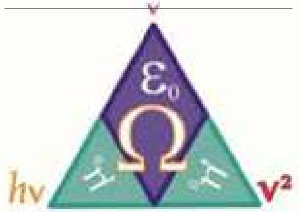

"There exists Negative and Positive electrical quanta"

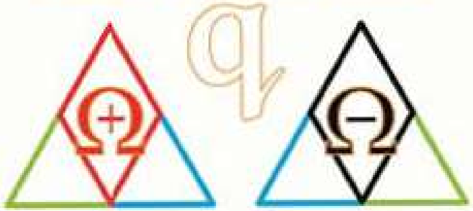

Electromagnetism: a fundamental interaction between the magnetic field and the presence and motion of an electric charge topology.

Electric charge: the geometry of EM energy momenta, also determines their electromagnetic interactions.

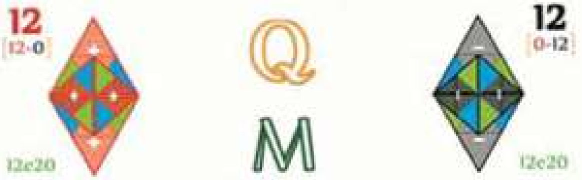

Electrically charged Matter topologies are influenced by, and produce, electromagnetic fields when in motion

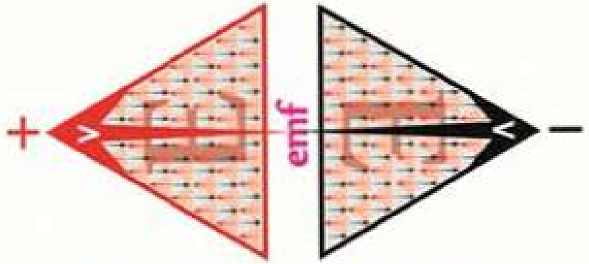

Electric potential: the capacity of an electric field to do work on an electric charge, typically measured in volts.

Electricity

Benjamin Franklin

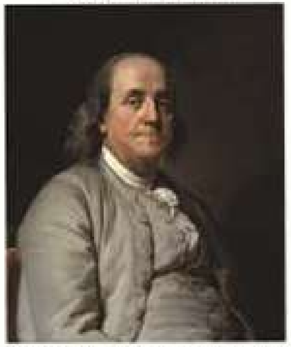

(January 17, 1706 – April 17, 1790)

Charles-Augustin_de_Coulomb
(14 September 1736 – 23 August 1806)

Andre Ampere
(20 January 1775 – 10 June 1836)

Georg Simon Ohm
(16 March 1789 – 6 July 1854)

Carl Friedrich Gauss
(30 April 1777 – 23 February 1855)

Heinrich Lenz
(February 12, 1804 – February 10, 1865)

Michael Faraday
(22 September 1791 – 25 August 1867)

James Clerk Maxwell
(13 June 1831 – 5 November 1879)

James Prescott Joule
(24 December 1818 – 11 October 1889)

John Henry Poynting
(9 September 1852 – 30 March 1914)

Nikola Tesla
(10 July 1856 – 7 January 1943)

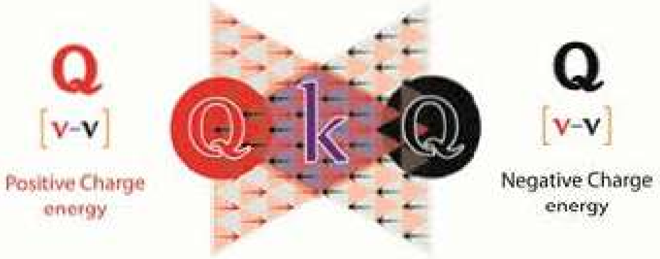

Electric field: an influence produced by one electric charge on other charges in its vicinity.

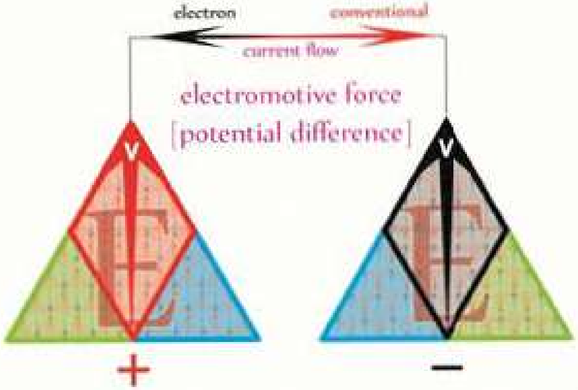

Electricity is the flow of Energy between separated Charge potentials when an electric circuit is formed measured in Volts (Joules/C)

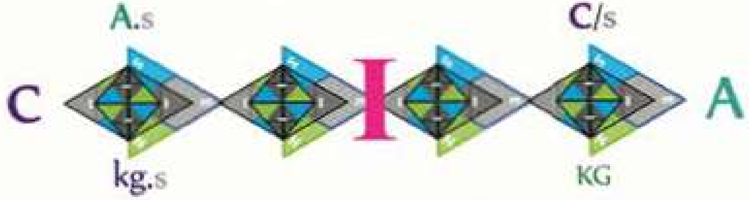

Electric current: is a secondary effect resulting from electrical energy in a circuit It is the movement or flow of electrically charged particles, typically measured in amperes.

Electricity is a general term encompassing a variety of phenomena resulting from the presence and flow of electric charge. These include many easily recognizable phenomena, such as lightning, static electricity, and the flow of electrical current in an electrical wire. In addition, electricity encompasses less familiar concepts such as the electromagnetic field and electromagnetic induction.

Copyright ABRAHAM [2008] - all rights reserved

EM field Permittivity

The Electric constant, commonly called the vacuum permittivity, or permittivity of free space, relates the units for electric charge to mechanical quantities such as length and force.

The name Vacuum Permittivty is a misnomer and should be replaced with the correct term EM field Permittivity

The strength of Electric fields is determined by the Electrical Permittivity Constant

$$E = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}$$

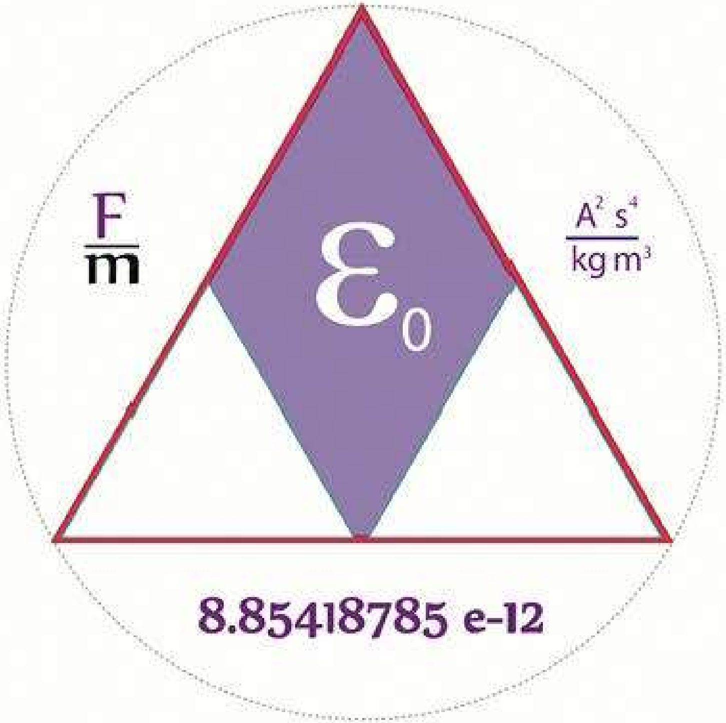

8.85418785 e-12

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

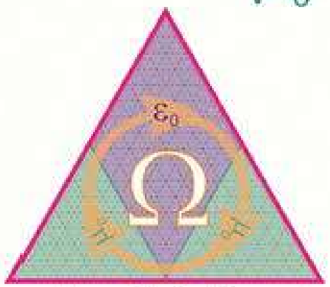

"Ampere's Law states that for any closed loop path, the sum of the quantities (B.ds) for all path elements into which the complete loop has been divided is equal to the product of µ0 and the total current enclosed by the loop."

$$k = \frac{1}{4\pi\varepsilon_0}$$

The permittivity of empty space, equal to 1 in centimeter-gram-second electrostatic units and to 107/4πc2 farads per meter or, numerically, to 8.854 × 10-12 farad per meter in International System units, where c is the speed of light in meters per second.

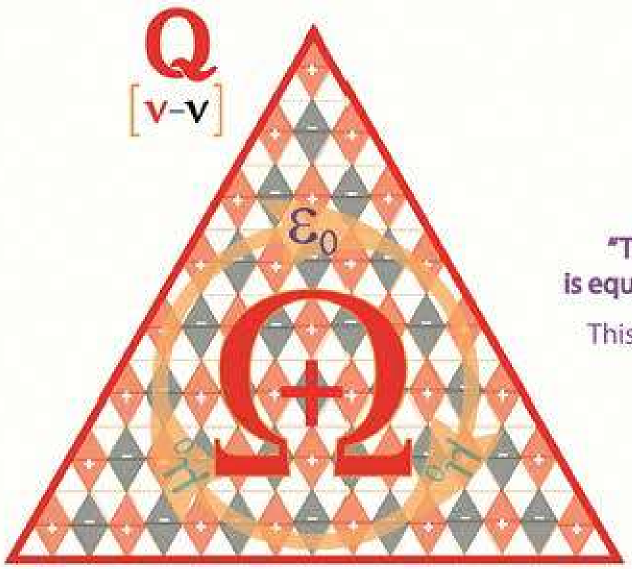

Positive Charge Electric Field

Gauss' Law:

"The total of the electric flux out of a closed surface is equal to the charge enclosed divided by the permittivity"

This applies equally to any integral geometry chosen to tessellate a surface area

Superpostioned E fields gives rise to Coulomb Forces

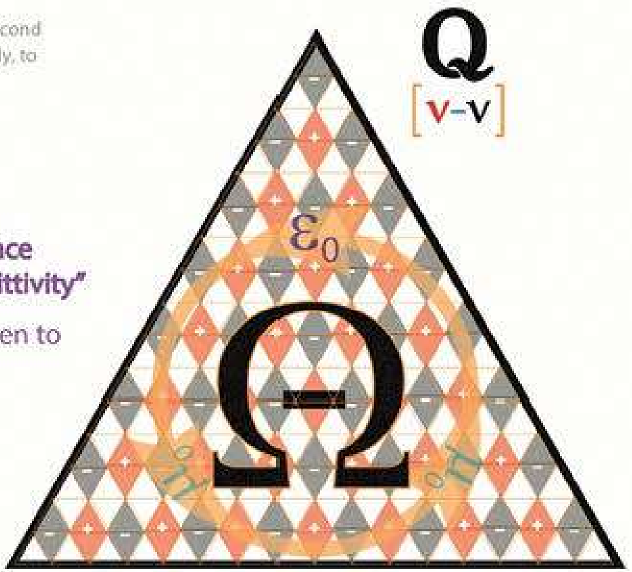

Negative Charge Electric Field

EM field Permeability

The permeability of free space,
also called absolute permeability.

The name Vacuum Permeability is a misnomer and should be replaced with the correct term EM field Permeability

The magnetic constant has the value of
4π × 10-7 henry per meter.

The strength of Magnetic fields is determined by the Magnetic Permeability Constant

μ_0

The magnetic field is most commonly defined in terms of the Lorentz force it exerts on moving electric charges.

F

The magnetic field generated by a steady current (a constant flow of electric charges in which charge is neither accumulating nor depleting at any point) as described by the Biot–Savart law

$$B = \mu_0 H$$

$\frac{H}{m}$

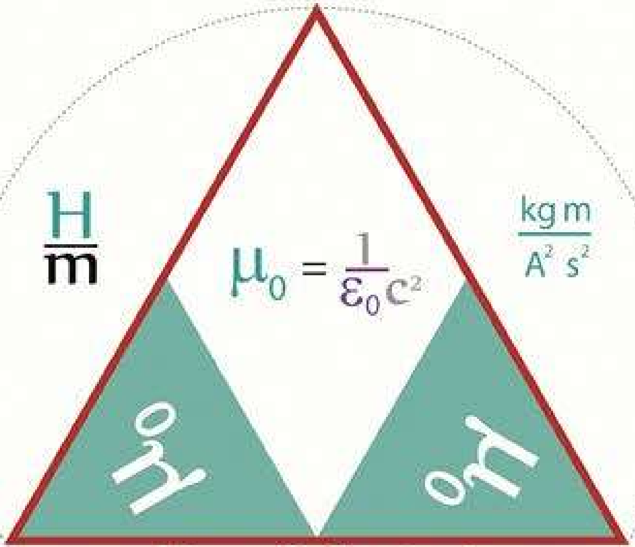

$$\mu_0 = \frac{1}{\varepsilon_0 c^2}$$

$\frac{kg\,m}{A^2\,s^2}$

Magnetic Constant

1.25663706 e-6

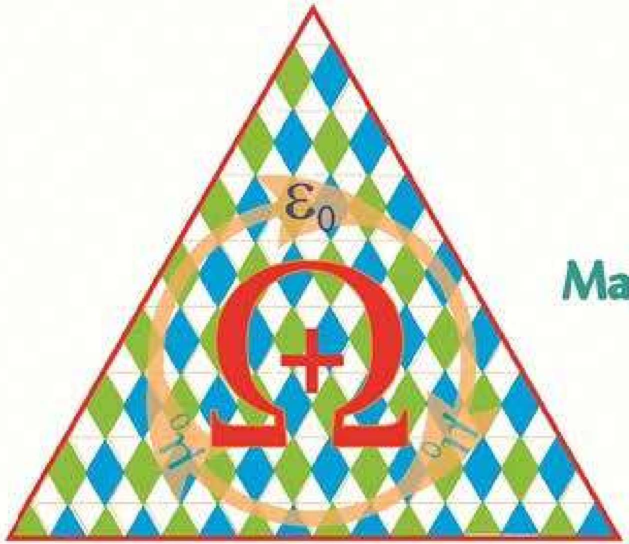

Positive Magnetic Moment

Magnetic monopoles do NOT exist

they are a mathematical identity
and not physically possible due to
the equilateral geometry
of Planck energies

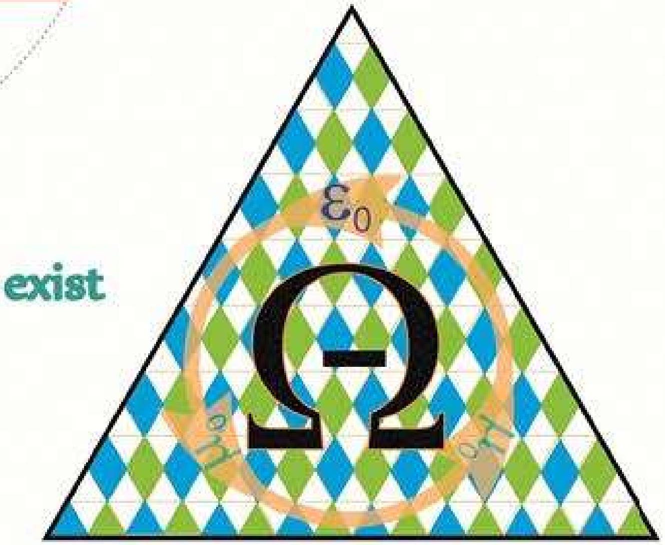

Negative Magnetic Moment

Copyright ABRAHAM [2008] - all rights reserved

James Prescott Joule

(24 December 1818 – 11 October 1889)

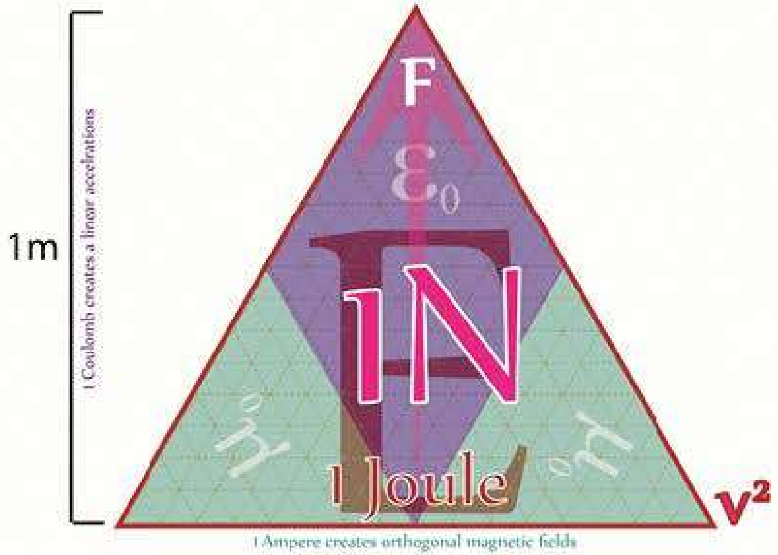
1m
1 Coulomb creates a linear accelerations
F
ε_0
E
1N
μ_0
μ_0
1 Joule
v^2
1 Ampere creates orthogonal magnetic fields

Joule

A measure of the equilateral charged
mass-energy momenta that can to do work

... the mechanical power exerted in turning a magneto-electric machine is converted into the heat evolved by the passage of the currents of induction through its coils; and, on the other hand, that the motive power of the electro-magnetic engine is obtained at the expense of the heat due to the chemical reactions of the battery by which it is worked (1845)

EM Field Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

mass ElectroMagnetic mass velocity

It is equal to the energy expended (or work done) in applying a force of one newton through a distance of one meter (1 newton metre or N·m), or in passing an electric current of one ampere through a resistance of one ohm for one second

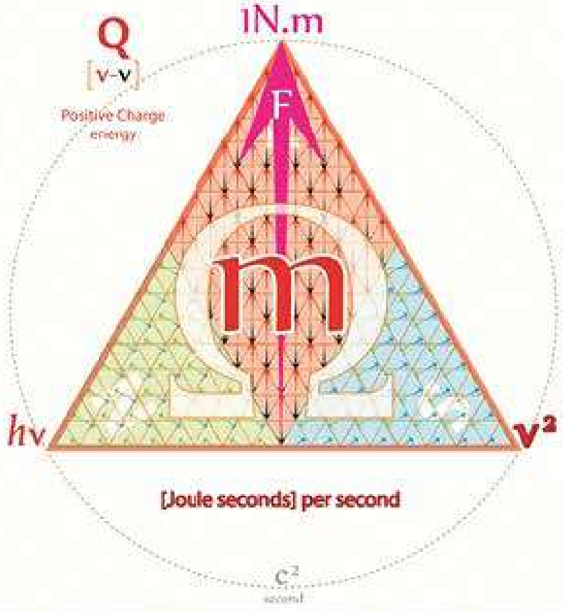
Q
[v-v]
Positive Charge energy
1N.m
F
m
hv
v^2
[Joule seconds] per second
c^2
second

Energy $\mathrm{kg}.\dfrac{\mathrm{m}^2}{\mathrm{s}^2}$

$$J = \frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^2} = \mathrm{N}\cdot\mathrm{m} = \mathrm{Pa}\cdot\mathrm{m}^3 = \mathrm{W}\cdot\mathrm{s}$$

Planck's Constant x frequency $\left[\mathrm{kg}.\frac{\mathrm{m}^2}{\mathrm{s}}\right].\mathrm{s}^{-1}$

mass x velocity squared $\mathrm{kg}.\left[\frac{\mathrm{m}}{\mathrm{s}}\right]^2$

Momentum x velocity $\left[\mathrm{kg}.\frac{\mathrm{m}}{\mathrm{s}}\right].\frac{\mathrm{m}}{\mathrm{s}}$

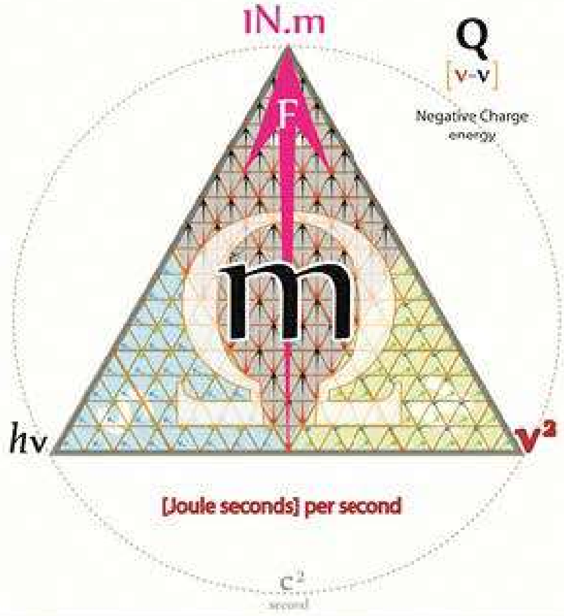
1N.m
Q
[v-v]
Negative Charge energy
F
m
hv
v^2
[Joule seconds] per second
c^2
second

Copyright ABRAHAM [2008] - all rights reserved

Volts

[Joules per Coulomb]

The volt is the unit for electric potential (voltage), electric potential difference, and electromotive force.

$$V = \frac{W}{A} = \frac{J}{A \cdot s} = \frac{N \cdot m}{A \cdot s} = \frac{kg\, m^2}{A\, s^3} = \frac{kg \cdot m^2}{C \cdot s^2} = \frac{N \cdot m}{C} = \frac{J}{C}$$

A single volt is defined as the difference in electric potential across a wire when an electric current of one ampere dissipates one watt of power.
It is also equal to the potential difference between two parallel, infinite planes spaced 1 meter apart that create an magnetic field of 1 newton per coulomb.
Additionally, it is the potential difference between two points that will impart one joule of energy per coulomb of charge that passes through it

Positive charge

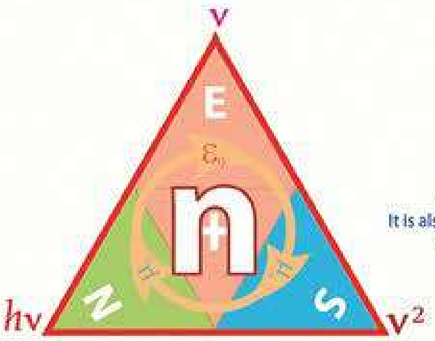
v
E
ε_0
n
+
N
S
hv
v^2

1V = 1J/C

One Volt is defined as energy consumption of
one joule per electric charge of one coulomb

Negative charge

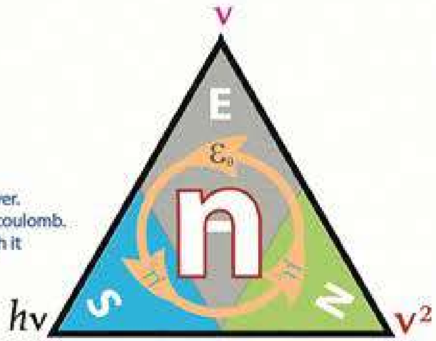
v
E
ε_0
n
-
S
N
hv
v^2

1V = 1A · 1ohm

One volt is equal to current of 1 amp
times resistance of 1 ohm

Positive voltage

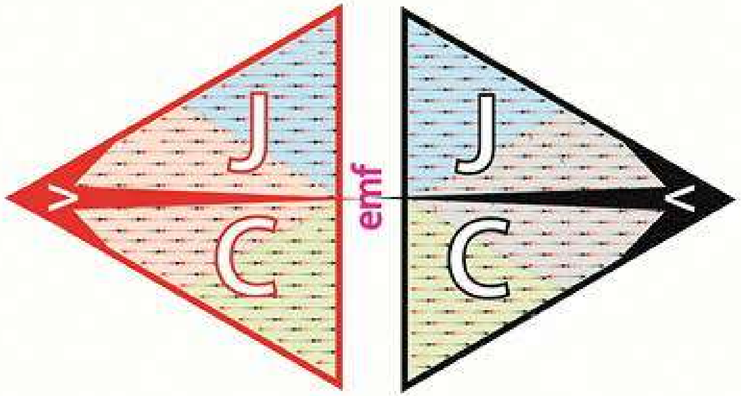

Negative voltage

Any separation of electrostatic charges, results in a voltage.
[Potential Difference]

Voltage is a scalar measure of linear electromotive forces of separated charges seeking potential equilibrium

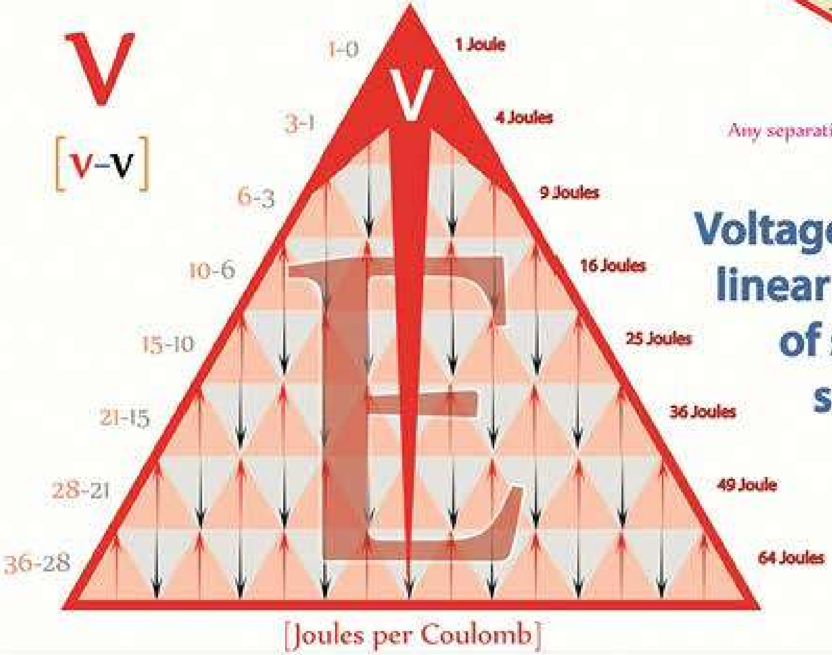
V
[v-v]
1-0
3-1
6-3
10-6
15-10
21-15
28-21
36-28
V
E
1 Joule
4 Joules
9 Joules
16 Joules
25 Joules
36 Joules
49 Joule
64 Joules
[Joules per Coulomb]

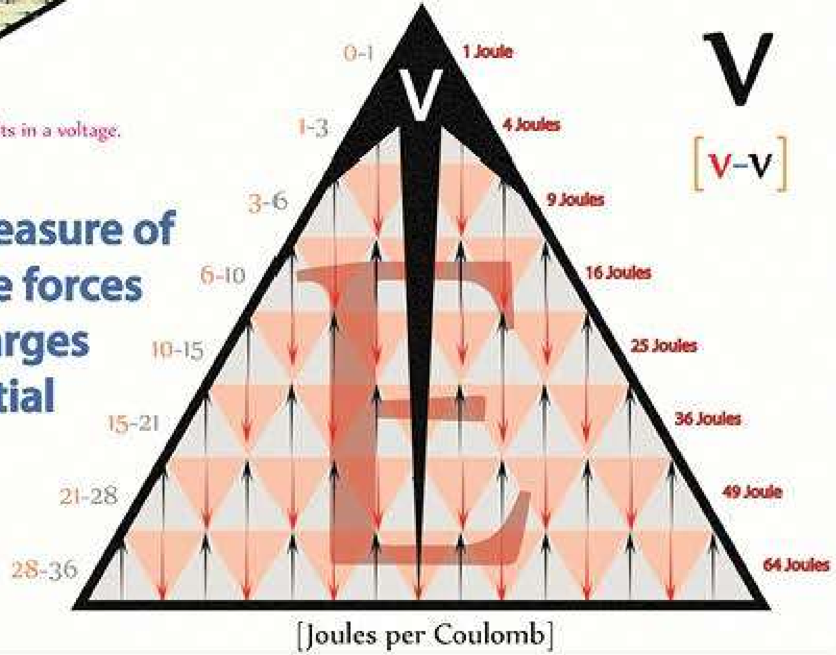
V
[v-v]
0-1
1-3
3-6
6-10
10-15
15-21
21-28
28-36
V
E
1 Joule
4 Joules
9 Joules
16 Joules
25 Joules
36 Joules
49 Joule
64 Joules
[Joules per Coulomb]

Tetryonics 23.05 - Volts and emf

Charge is the equilateral geometry of Energy that gives form to all mass & Matter

Positive charged mass-energy

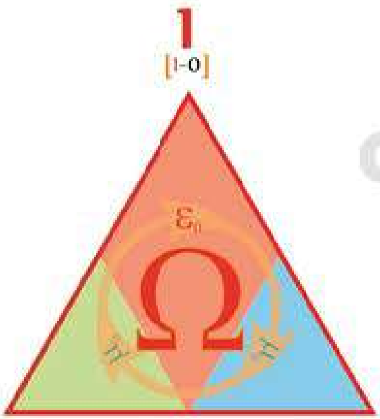

Clockwise inductive energy flux

charge

q

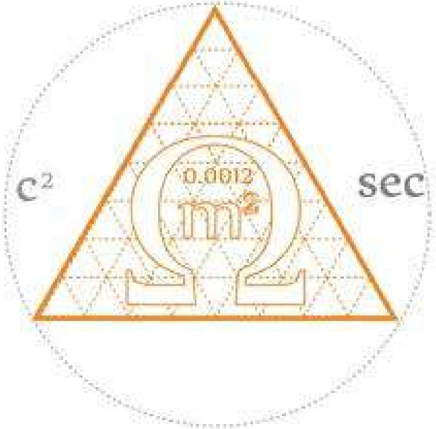

time

Negative charged mass-energy

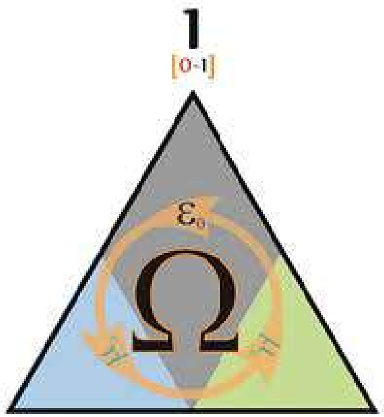

Counter clockwise energy flux

q

t

$$\frac{\Omega}{c^2}$$

$$\frac{m^2}{s}\;\frac{s^2}{m^2}$$

$$s$$

seconds

t

1.33518 e-20 s

Changes to nett quantised angular momenta [charged mass-energies] within any space-time co-ordinate system forms the basis of time

ElectroMagnetic Charge is a quantum property resulting from the equilateral QAM geometry of mass-Energy

The two ElectroMagnetic charge geometries possible can be created through tesselation of electrical energy in ideal inductive loops

Copyright ABRAHAM [2008] - all rights reserved

Coulombs

are a measure of 2D charged mass geometries
and the electromagnetic forces they produce per unit time

W+ boson integrals
nett positive charged
mass momenta

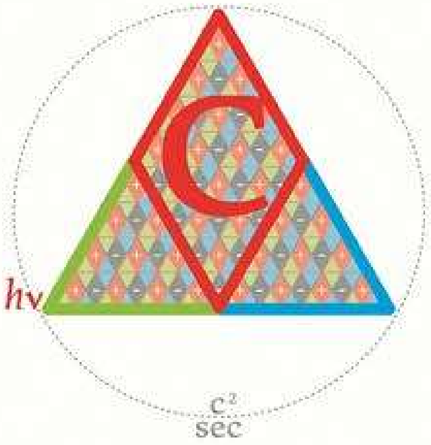

Q
[v-v]

Charles-Augustin de Coulomb

(14 September 1736 – 23 August 1806)

Derived in 1785 Coulomb's Law is a law of physics describing the electrostatic interaction between any two charged particles (and forms the basis for Ampere's Law)

Q
[v-v]

W- boson integrals
nett negative charged
mass momenta

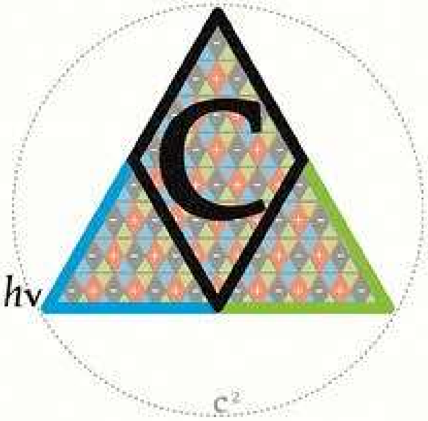

Electric forces

$$m\frac{\Omega}{c^2}$$

Magnetic forces

Coulombs

The coulomb (unit symbol: C) is the SI derived unit of electric charge (symbol: Q or q).
It is defined as the charge transported by a steady current of one ampere in one second:

Amperes

As

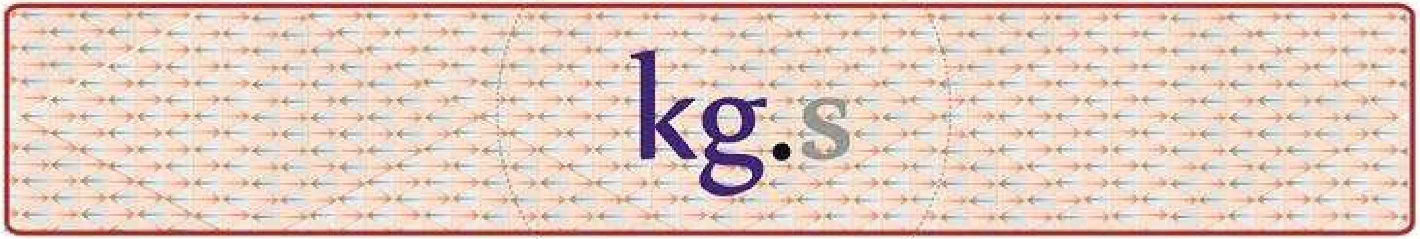

$\frac{C}{s}$

1A.1s

Interactive linear momenta of super-positioned E-fields creates Coulomb's force law

1F.1V

c²
sec

Copyright ABRAHAM [2008] - all rights reserved

$1\text{C} = 1\text{A} \cdot 1\text{s}$

1 COULOMB is a measure of linear 2D electric field energy momenta in EM field geometries

1 AMPERE is a measure of orthogonal 2D magnetic field forces produced by Matter topologies in motion

$1\,\text{A} = 1\dfrac{\text{C}}{\text{s}}$

quantised electrical energy momenta

1C is the quantity of charged masses which pass any point in a circuit in which a current of 1A flows for 1s

C

Coulombs

kg.s

A

Amperes

KG

linear acceleration energy momenta

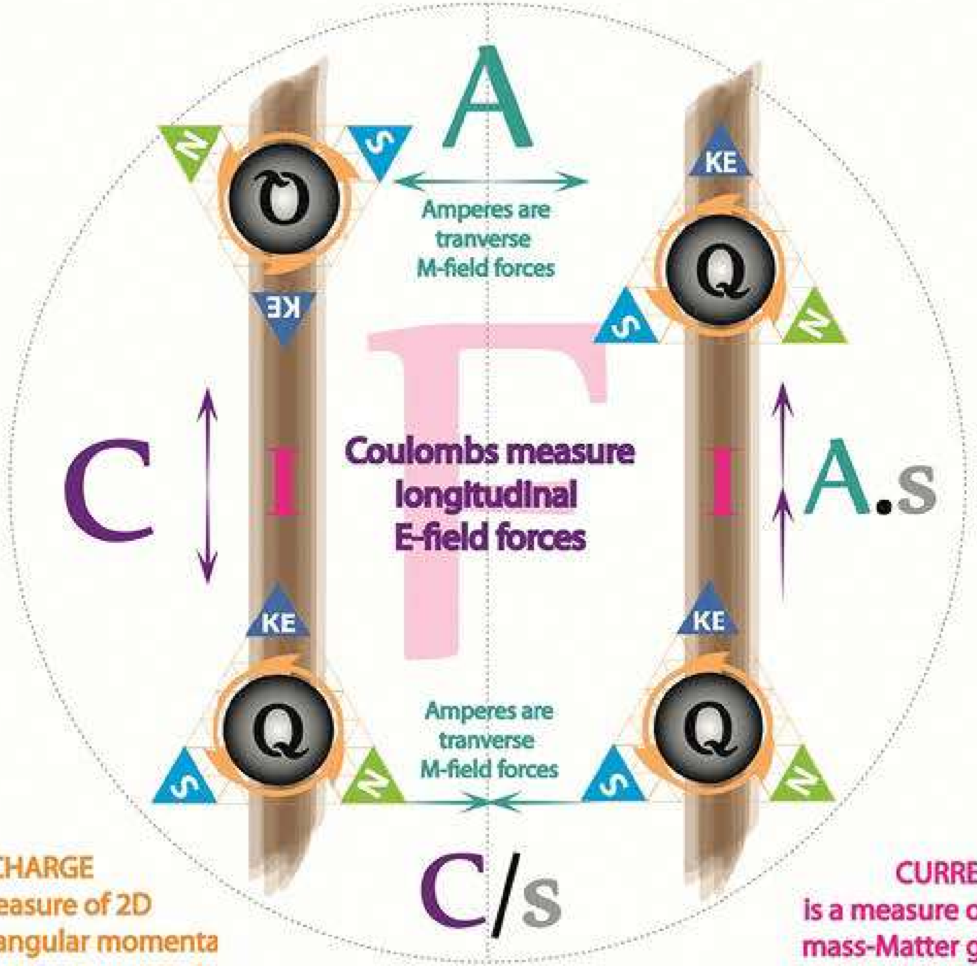
A
Amperes are tranverse M-field forces
N
S
KE
Q
C
I
F
Coulombs measure longitudinal E-field forces
A.s
Amperes are tranverse M-field forces
C/s

magnetic perturbation energy momenta

Q

Charge

CHARGE is a measure of 2D quantised angular momenta [mass -Matter geometries]

$$Q = I.t$$

quantity of charge = rate of flow of charge × time

CURRENT is a measure of charged mass-Matter geometries in motion

I

Current

Current

is a measure of charged Matter in motion in a circuit per unit of Time

An elementary charge is
1.602216081 e-19 C

C

It is defined as the charge transported by a steady current of one ampere in one second.

Coulombs

Amperes

The amount of electric charge passing a point per unit time

A

6.2413354 e18 electrons passing a given point each second constitutes one ampere.

charged **masses**

kg.s

1.33518 e-20 kg.s

7.489626 e19 kg

Matter

KG

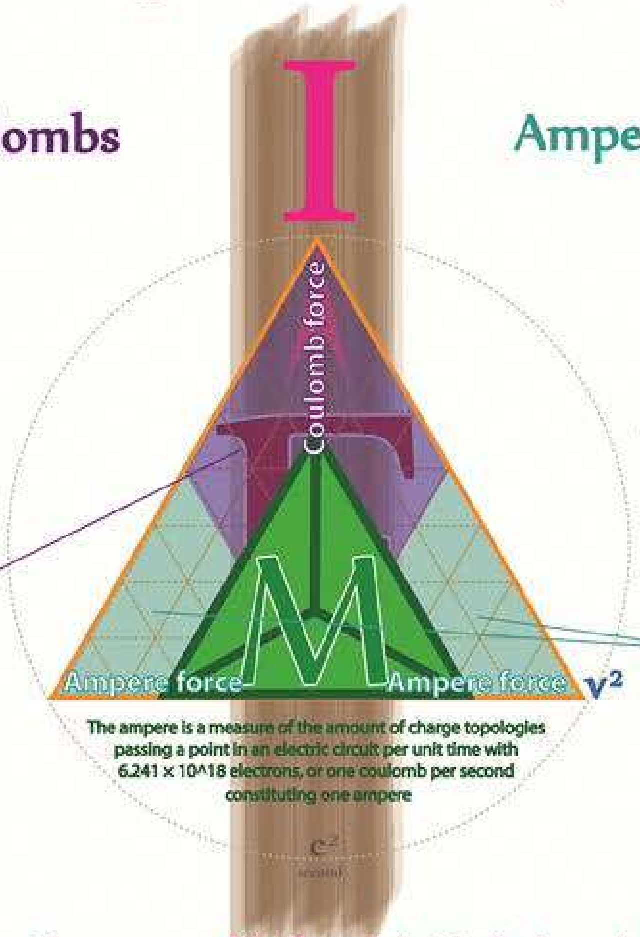

The 2D linear mass-energy momenta capable of accelerating charged particles

The transverse magnetic forces produced by Matter in motion

The ampere is a measure of the amount of charge topologies passing a point in an electric circuit per unit time with 6.241 x 10^18 electrons, or one coulomb per second constituting one ampere

$$1\text{C} = 1\text{A} \cdot 1\text{s}$$

$$1\,\text{A} = 1\frac{\text{C}}{\text{s}}$$

Electric current is a measure of the flow of electrically charged mass-Matter in a circuit as a result of an electrical Voltage potential

As

	5.524527227 e-12 kg	
Charged mass	1 second	Matter in motion
	6.241355408 e18 electrons	

$$\frac{\text{C}}{\text{S}}$$

Voltage [emf] is a measure of the energy momenta per coulomb available to exert a force on charged Matter in a circuit

Tetryonics 23.09 - Amps [mass-Matter]

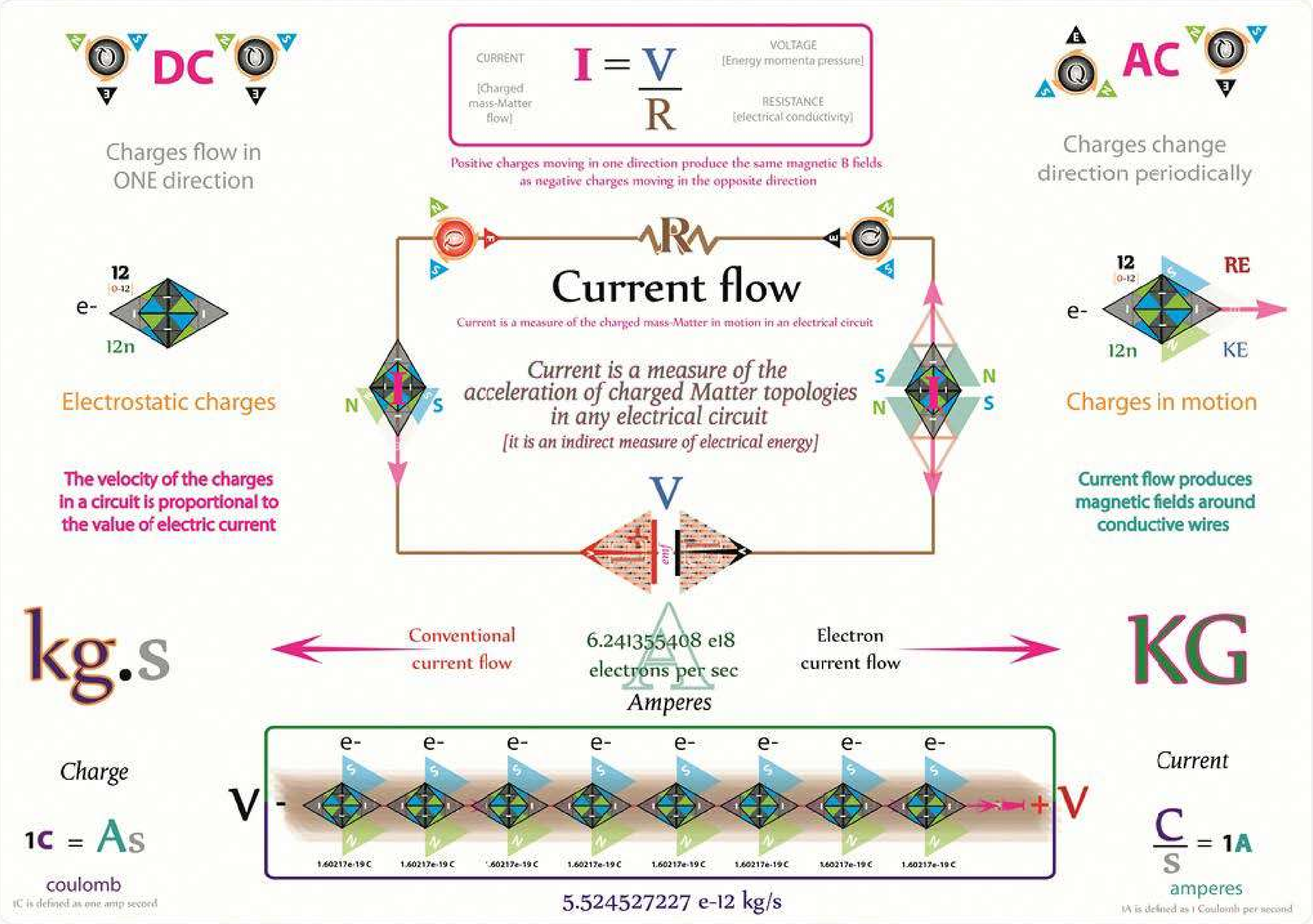

DC

Charges flow in ONE direction

$$I = \frac{V}{R}$$

CURRENT [Charged mass-Matter flow]

VOLTAGE [Energy momenta pressure]

RESISTANCE [electrical conductivity]

Positive charges moving in one direction produce the same magnetic B fields as negative charges moving in the opposite direction

AC

Charges change direction periodically

Current flow

Current is a measure of the charged mass-Matter in motion in an electrical circuit

Current is a measure of the acceleration of charged Matter topologies in any electrical circuit
[it is an indirect measure of electrical energy]

Electrostatic charges

Charges in motion

The velocity of the charges in a circuit is proportional to the value of electric current

Current flow produces magnetic fields around conductive wires

kg.s

Conventional current flow

6.241355408 e18 electrons per sec

Amperes

Electron current flow

KG

Charge

1C = As

coulomb

1C is defined as one amp second

5.524527227 e-12 kg/s

Current

$\frac{C}{s} = 1A$

amperes

1A is defined as 1 Coulomb per second

Copyright ABRAHAM [2008] - all rights reserved

The electrical resistance of a conductor is a measure of how much it opposes the passage of an electric current through it

OHM's Law

Georg Simon Ohm

(16 March 1789 – 6 July 1854)

resistivity

$$\varepsilon = \rho J$$

Electric field

Current density

General vector equation

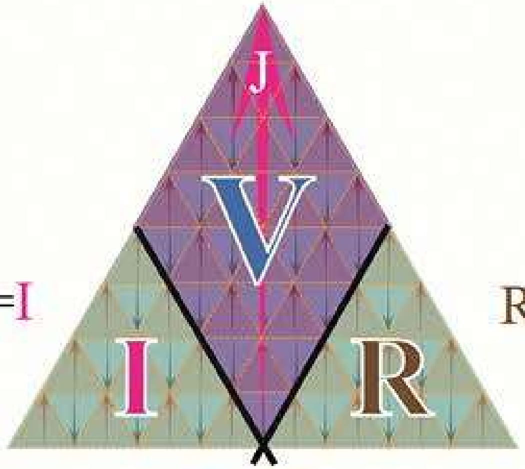

$$\frac{V}{R} = I$$

$$R = \frac{V}{I}$$

$$V = I.R$$

$$P = I.V = I^2.R$$

Power is the amount of current times the voltage level at a given point measured in watts.

Current is the flow of charged mass-Matter resulting from an electromotive force

Resistance determines how much current will flow through a conductor

Current [I.V] Voltage
POWER
Current squared [I²R] Resistance

Ohm's law states that the current through a conductor between two points is directly proportional to the potential difference across the two points.

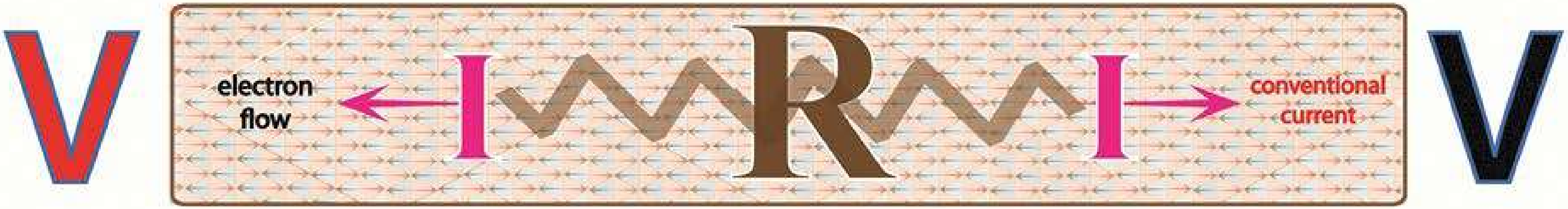

By introducing the constant of proportionality, Resistance, one arrives at the usual mathematical equation that describes this relationship

Voltage is the force motivating charges to "flow" in a circuit, it is measured as the difference in electrical potential between two points in a circuit

Copyright ABRAHAM [2008] - all rights reserved

Electrical Resistance

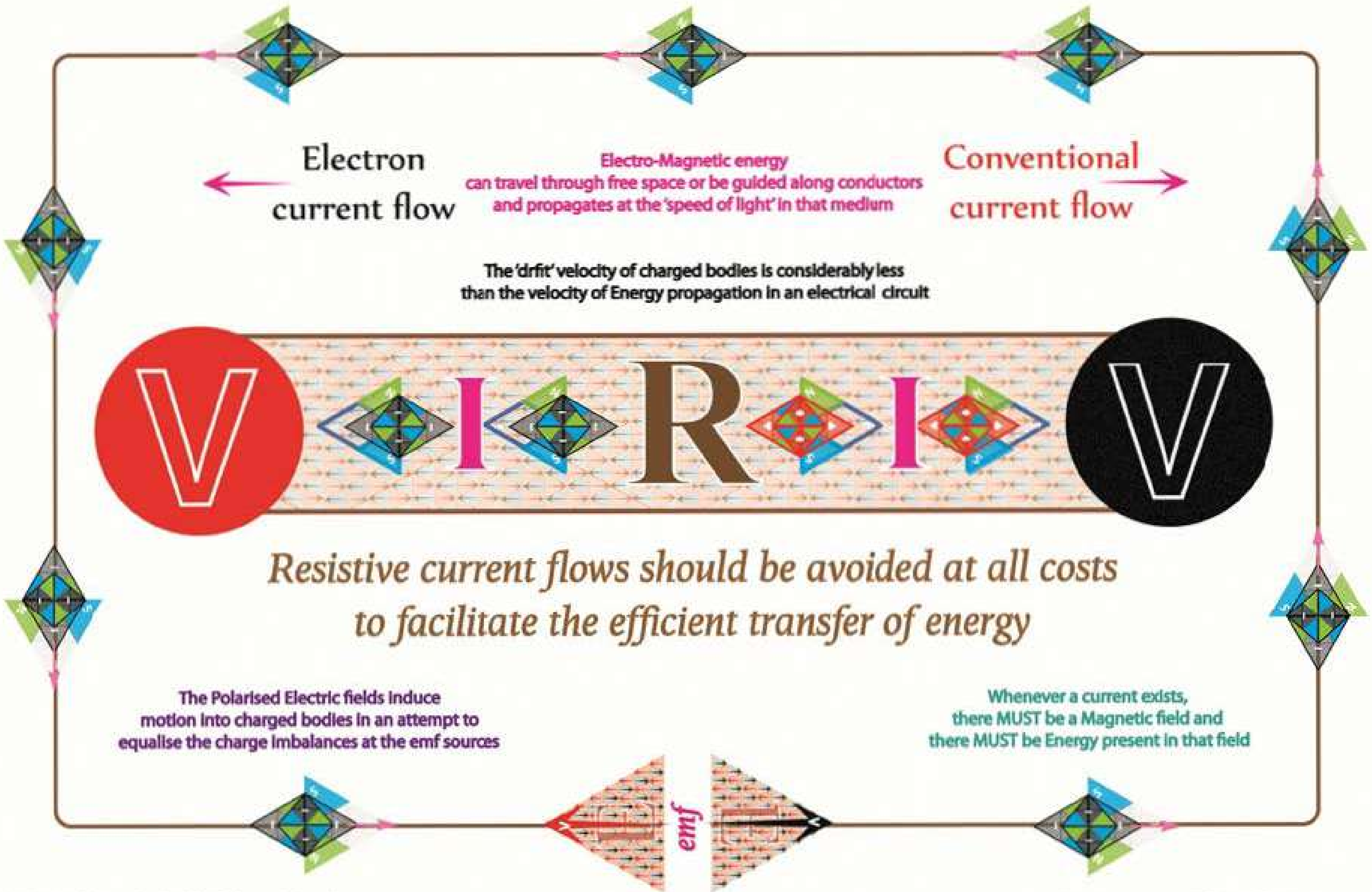

Electron current flow

Electro-Magnetic energy
can travel through free space or be guided along conductors
and propagates at the 'speed of light' in that medium

Conventional current flow

The 'drfit' velocity of charged bodies is considerably less
than the velocity of Energy propagation in an electrical circuit

*Resistive current flows should be avoided at all costs
to facilitate the efficient transfer of energy*

The Polarised Electric fields induce
motion into charged bodies in an attempt to
equalise the charge imbalances at the emf sources

Whenever a current exists,
there MUST be a Magnetic field and
there MUST be Energy present in that field

The acceleration [change in velocity] of electrons in motion
requires the release [or absorption] of Bosons of energy
from [or to] the equilateral KEM field geometry
[[spectriaaline emission~absorption]

Voltage produces a linear force that motivates
charged Matter topologies into motion in a circuit
as electrical energy geometries seek equilibrium

Power is a measure of the total charge geometry
(electrons) moving past a point along with the energy each
charge geometry (electron) possesses (KE expressed as eV)
V.I = P = I.R.I

James Prescott Joule verified Ohm's Law and determined that the heat delivered by a conductor is directly proportional to its resistance and to the square of the current through it.

He defined Power physically to be the rate at which electrical energies are created or transferred in an electric circuit

Electrical Power

Power is the integral of the cross-product of all electrical and magnetic field vectors produced by scalar energy momenta within a specific area defined by time

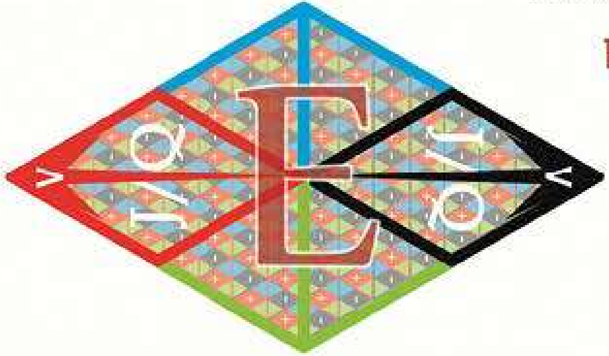

All energy is comprised of 2D charged masses

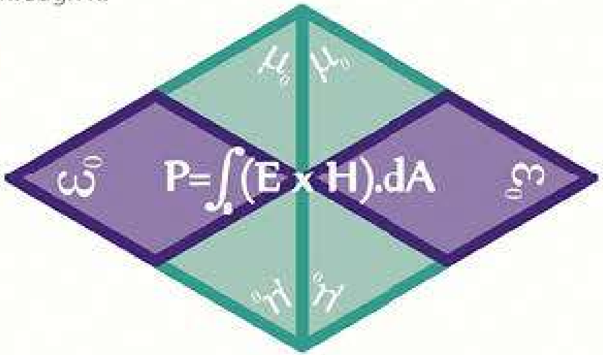
μ_0
μ_0
ε_0
$P=\int(E \times H).dA$
ε_0
μ_0
μ_0

The scalar surface integral of the Poynting vector.

The SI unit of power is the watt, one joule per second.

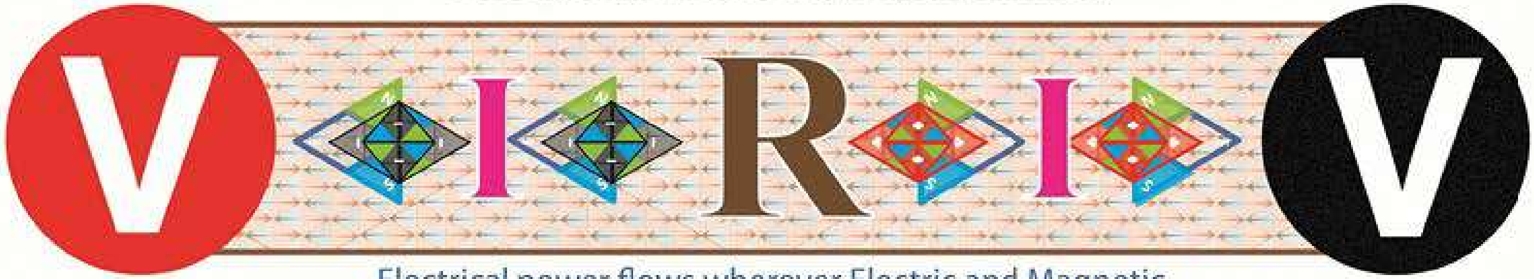

Electrical power flows wherever Electric and Magnetic fields exist together and change over time

emf Voltage

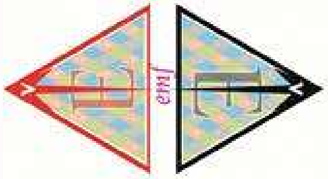

The potential to do work

$$P = V.I$$

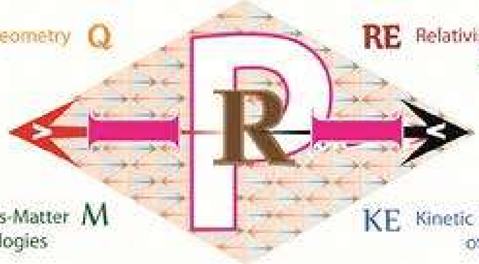
Charge geometry Q
RE Relativistic mass-energies densities
P
R
3D mass-Matter M topologies
KE Kinetic mass-energies of motion

$$I^2.R = \frac{V^2}{R}$$

Resistance

The Inertial mass of Matter

Electrical power propagates at the speed of light in an electrical circuit [dependent on the voltage source] electrons have a vastly slower 'drift' velocity resulting from their inertial interaction with energy momenta of the Voltage [emf] field

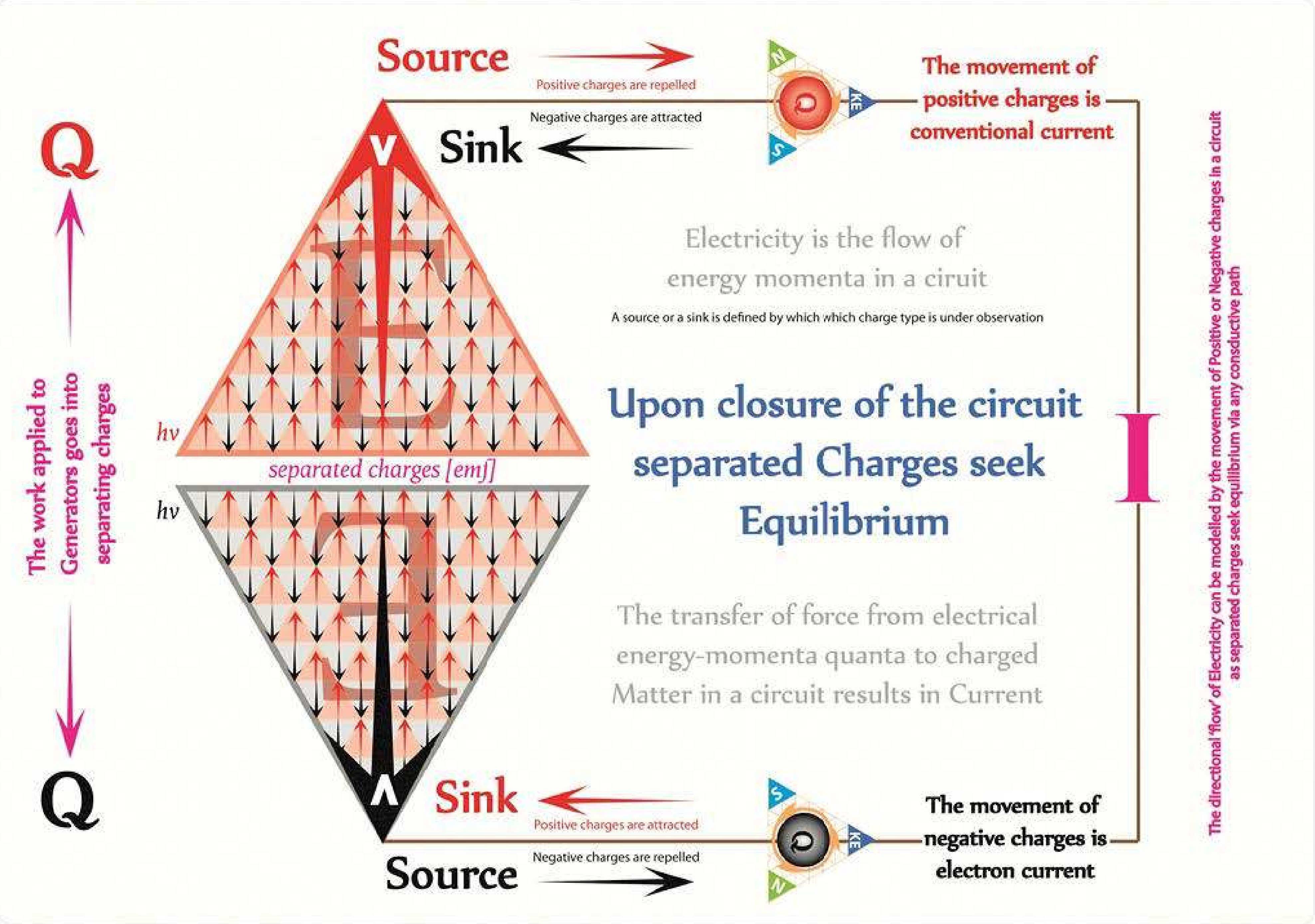
Source
Positive charges are repelled
Negative charges are attracted
Sink
Q
V
E
hv
separated charges [emf]
hv
E
V
Q
Sink
Positive charges are attracted
Negative charges are repelled
Source
N
S
KE
S
N
KE
I
The work applied to Generators goes into separating charges

The movement of positive charges is conventional current

Electricity is the flow of energy momenta in a ciruit

A source or a sink is defined by which which charge type is under observation

Upon closure of the circuit separated Charges seek Equilibrium

The transfer of force from electrical energy-momenta quanta to charged Matter in a circuit results in Current

The movement of negative charges is electron current

The directional 'flow' of Electricity can be modelled by the movement of Positive or Negative charges in a circuit as separated charges seek equilibrium via any conscductive path

Copyright ABRAHAM [2008] - all rights reserved

EM Inductive circuits

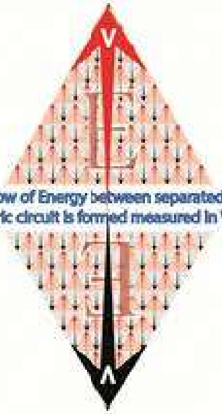

Electricity is the flow of Energy between separated Charge potentials when an electric circuit is formed measured in Volts [Joules/C]

The Electrical emf created by the energy momenta comprising the neutral EM field will cause electrons to flow in a conductor

Neutral Electric field with Magnetic dipole

Current is produced in a conductor when it is moved through a magnetic field because the neutralised Electric field lines are applying a emf on the free electrons in the conductor and causing them to move.

The process of generating current in a conductor by placing the conductor in a changing magnetic field is called induction

The stronger the Magnetic dipole the greater the potential emf

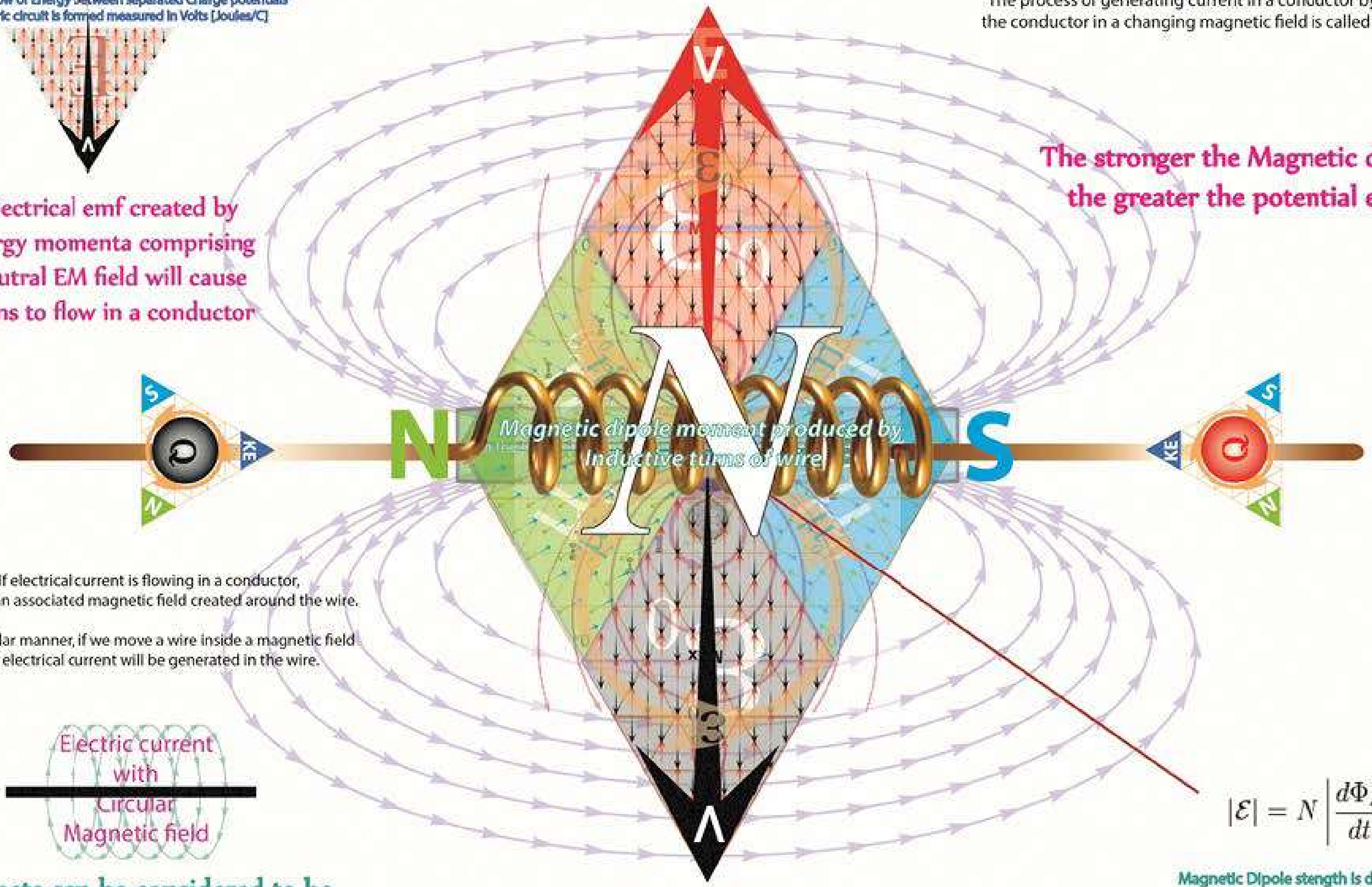

If electrical current is flowing in a conductor, there is an associated magnetic field created around the wire.

In a similar manner, if we move a wire inside a magnetic field an electrical current will be generated in the wire.

Electric current
with
Circular
Magnetic field

Magnets can be considered to be static stores of potential [emf fields]

Magnetic Dipole with Neutral Electric field

$$|\mathcal{E}| = N\left|\frac{d\Phi_B}{dt}\right|$$

Magnetic Dipole stength is dependent on:

The total Number of turns in the inductor
& the current flowing through the circuit

Copyright ABRAHAM [2008] - all rights reserved

Electron Volts

The electron volt can also be used
as a unit of mass-energy by applying
Einstein's relation $E = mc^2$.

For example, the rest mass-energy of electron
topologies if 496,532 eV (496.532 MeV).

Chemically, for 1 mole of electrons
1 eV ~ 100 kJ mol-1 (96.49 kJ mol-1)

eV

A unit of energy
equal to the work required to move
one electron through a potential
difference of 1 volt.

An Electron Volt is also
a measure of Electrical Kinetic Energy

EM Field Planck quanta

$$\frac{1}{2}\left[4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]\right]$$

Tetryons ElectroMagnetic mass velocity

As distinct from Matter's ENERGY
Kinetic Energy is 'extended from the
Tetryonic geometry of all Matter in motion
in a separate 2D KEM field

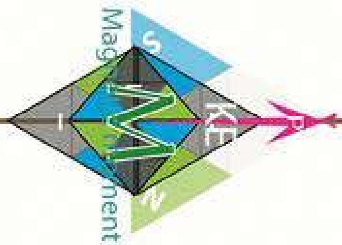

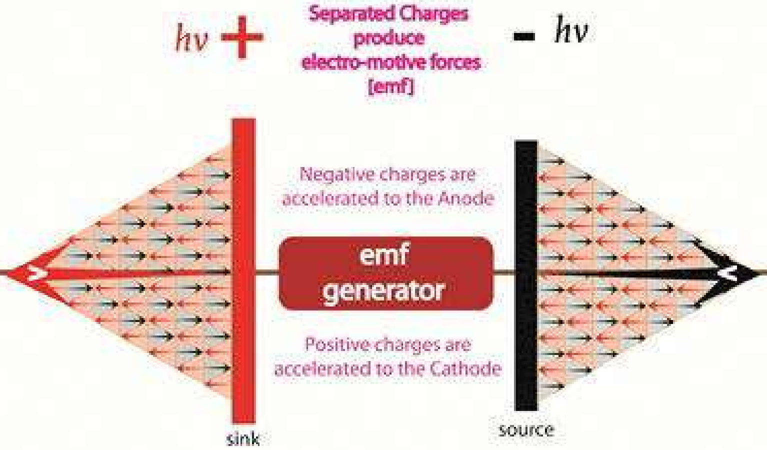

Separated Charges produce electro-motive forces [emf]

Negative charges are accelerated to the Anode

Positive charges are accelerated to the Cathode

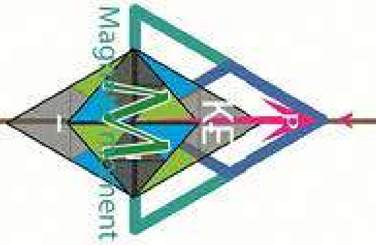

A Free electron is attracted to Positive anode of an emf source

The application of UFE to Kinetic energy calculations reveals the underlying processes of the Photo-electric effect and Light-Energy interactions with respect to Kinetic energy

Accelerated electrons have increased KE, Magnetic moments & momentum

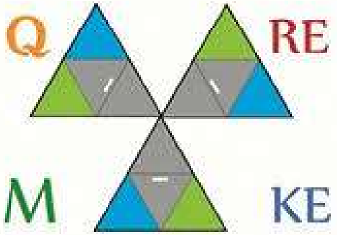

$$\frac{p^2}{2m} = KE = \frac{1}{2}Mv^2$$

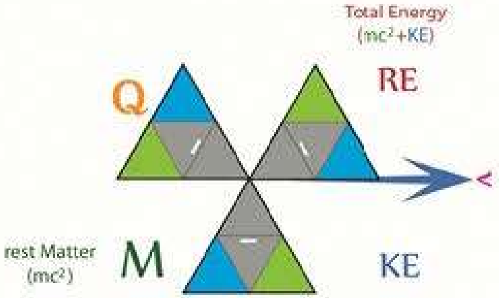

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Fermions ElectroMagnetic mass velocity

By definition, it is equal to the amount of kinetic energy momenta gained by a single unbound electron when it accelerates through an electric potential difference of one volt.

(Total Energy = Absolute rest Matter + Kinetic energy)
additionally the Lorentz relativistic correction factor (β)
presents itself naturally from the UFE when velocity is applied

Copyright ABRAHAM [2008] - all rights reserved

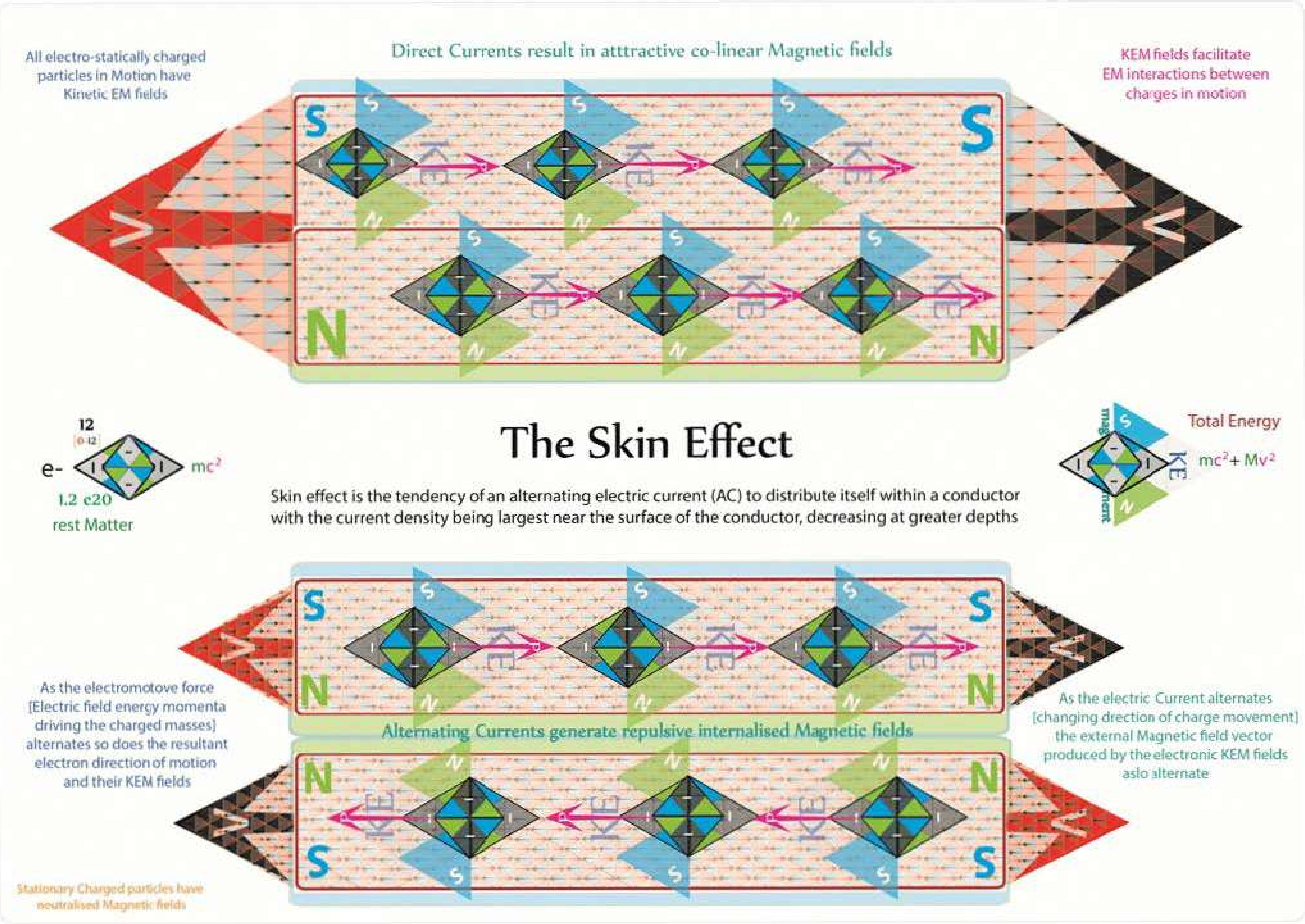

Direct Currents result in atttractive co-linear Magnetic fields

All electro-statically charged particles in Motion have Kinetic EM fields

KEM fields facilitate EM interactions between charges in motion

The Skin Effect

Skin effect is the tendency of an alternating electric current (AC) to distribute itself within a conductor with the current density being largest near the surface of the conductor, decreasing at greater depths

Alternating Currents generate repulsive internalised Magnetic fields

As the electromotove force [Electric field energy momenta driving the charged masses] alternates so does the resultant electron direction of motion and their KEM fields

As the electric Current alternates [changing direction of charge movement] the external Magnetic field vector produced by the electronic KEM fields aslo alternate

Stationary Charged particles have neutralised Magnetic fields

EM fields of particles in Motion

Positive changed quanta traveling anti-clockwise
in a solonoid creates a North-South magnetic field
orientated in the direction shown

positive ion flow

Negative charged quanta create a
reversed Magnetic dipole field and
reversing the direction of particle motion
also reverses the magnetic vector

electron flow

Solenoidal Motion

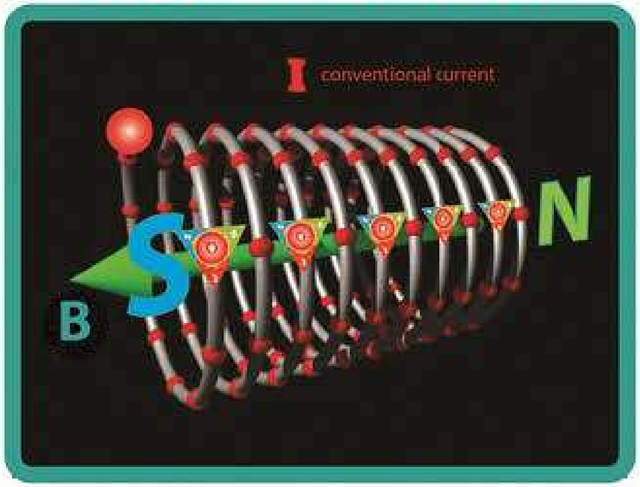

Charged particles moving in a spiral
produce a Circular Magnetic field

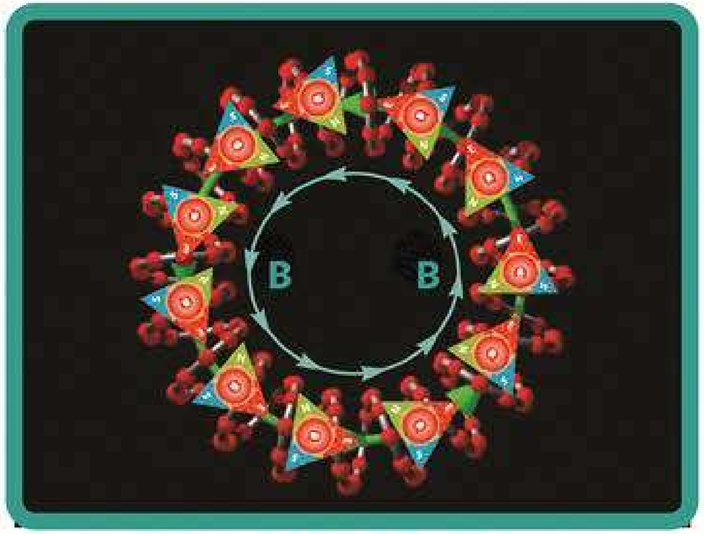

Toroidal Motion

A toroidally wound conductor produces
a Circular Magnetic field

Helical Motion

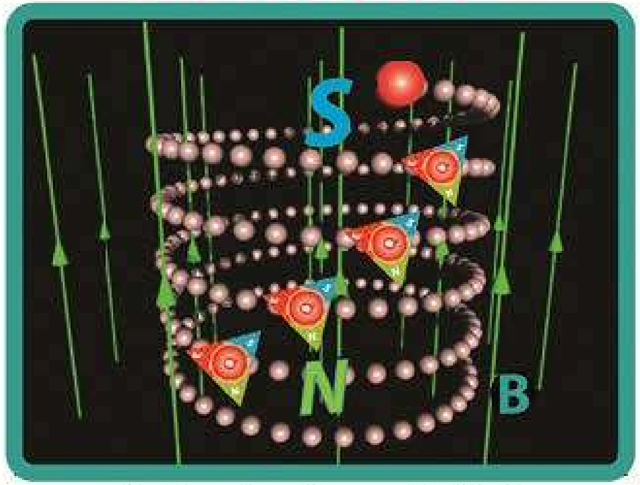

Motion in a Magnetic field

Copyright ABRAHAM [2008] - all rights reserved

Inductors

An inductor (also choke, coil, or reactor) is a passive two-terminal electrical component that stores energy in its magnetic field.

For comparison, a capacitor stores energy in an electric field, and a resistor does not store energy but rather dissipates energy as heat

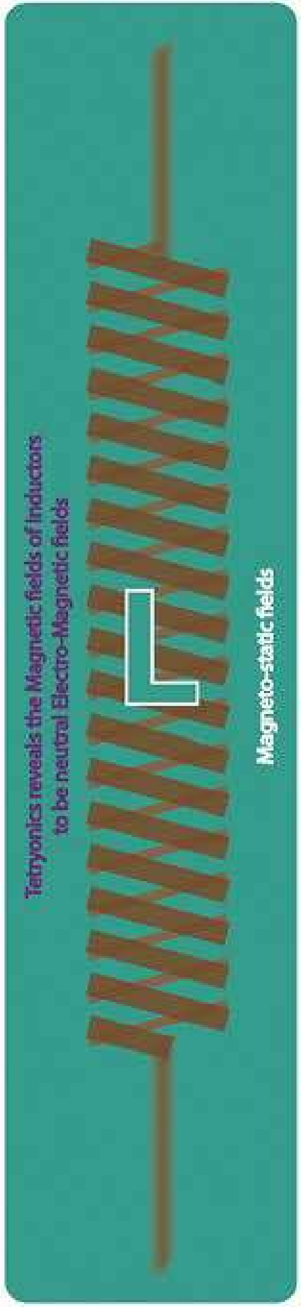

Tetryonics reveals the Magnetic fields of inductors
to be neutral Electro-Magnetic fields

Magneto-static fields

When the current flowing through an inductor changes,
a time-varying electromagnetic field is created around the coil,
and a changing emf voltage is formed

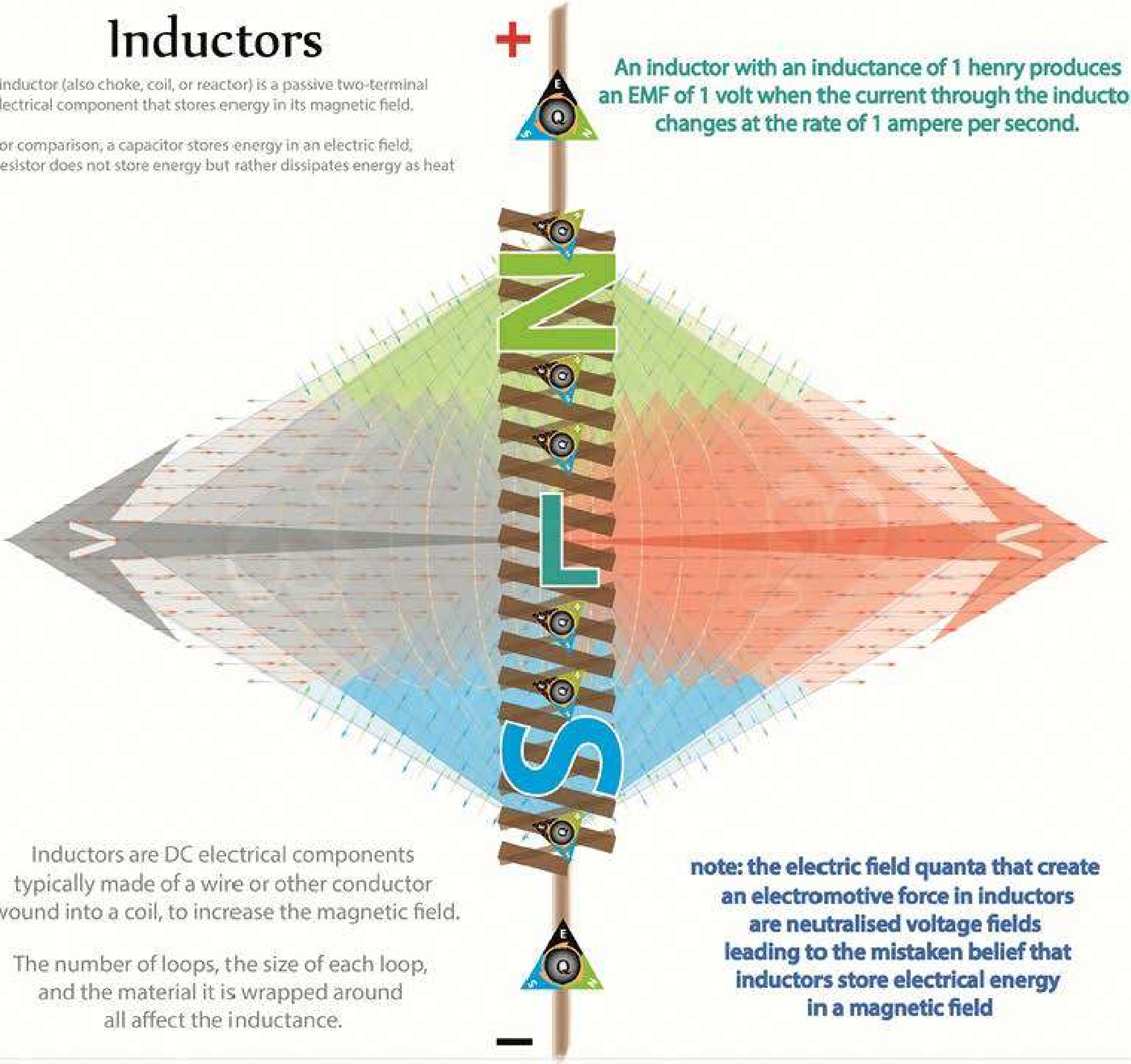

An inductor with an inductance of 1 henry produces an EMF of 1 volt when the current through the inductor changes at the rate of 1 ampere per second.

Inductors are DC electrical components typically made of a wire or other conductor wound into a coil, to increase the magnetic field.

The number of loops, the size of each loop, and the material it is wrapped around all affect the inductance.

note: the electric field quanta that create an electromotive force in inductors are neutralised voltage fields leading to the mistaken belief that inductors store electrical energy in a magnetic field

Charges moving in an Inductive winding

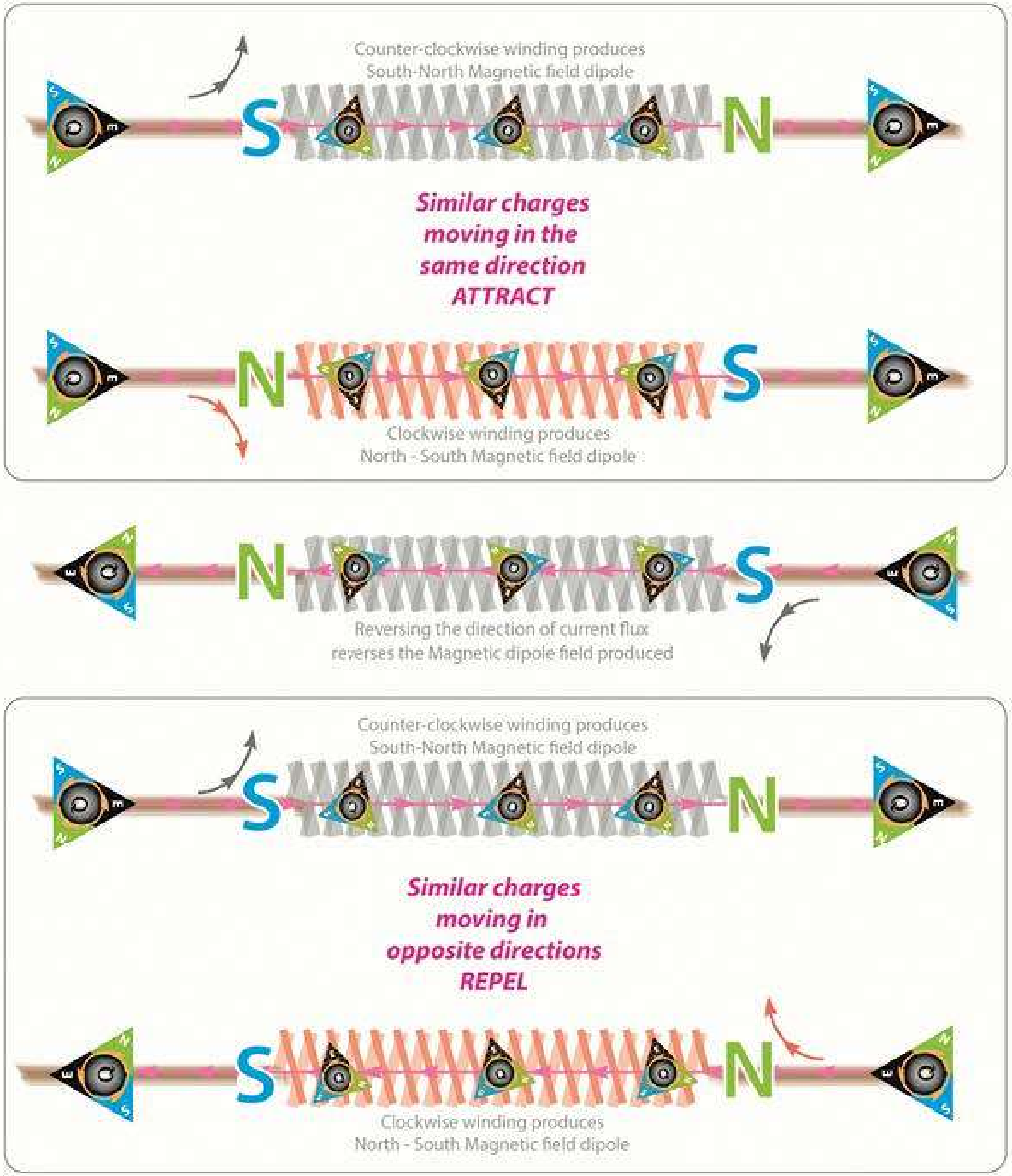

Counter-clockwise winding produces South-North Magnetic field dipole

Similar charges moving in the same direction ATTRACT

Clockwise winding produces North - South Magnetic field dipole

Reversing the direction of current flux reverses the Magnetic dipole field produced

Counter-clockwise winding produces South-North Magnetic field dipole

Similar charges moving in opposite directions REPEL

Clockwise winding produces North - South Magnetic field dipole

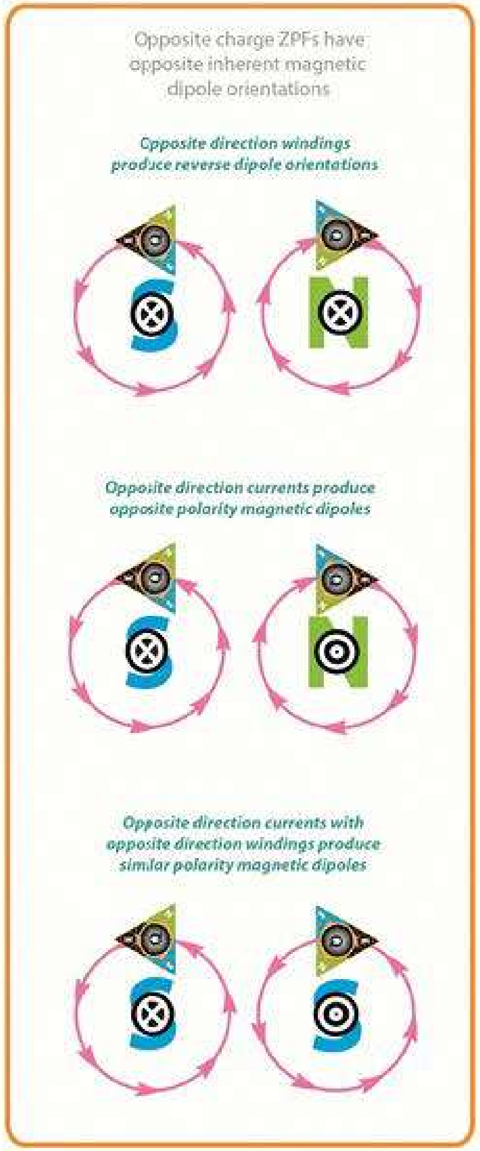

Opposite charge ZPFs have opposite inherent magnetic dipole orientations

Opposite direction windings produce reverse dipole orientations

Opposite direction currents produce opposite polarity magnetic dipoles

Opposite direction currents with opposite direction windings produce similar polarity magnetic dipoles

This is best viewed using physical models

Tetryonics 24.07 - Moving charges in an Inductor

Capacitors

Changing Electric fields
create
changing Magnetic fields

A capacitor is a passive two-terminal AC electrical component
used to store alternating electrical energy in an electric field.
It blocks the block of direct currents.

Changing Magnetic fields
create
changing Electric fields

Capacitors store & pass electrical energy as charged 2D masses

Capacitors store EM Energy
as transverse Bosons

Capacitors store charged
masses between their plates

E

Capacitors store
Energy as Charges

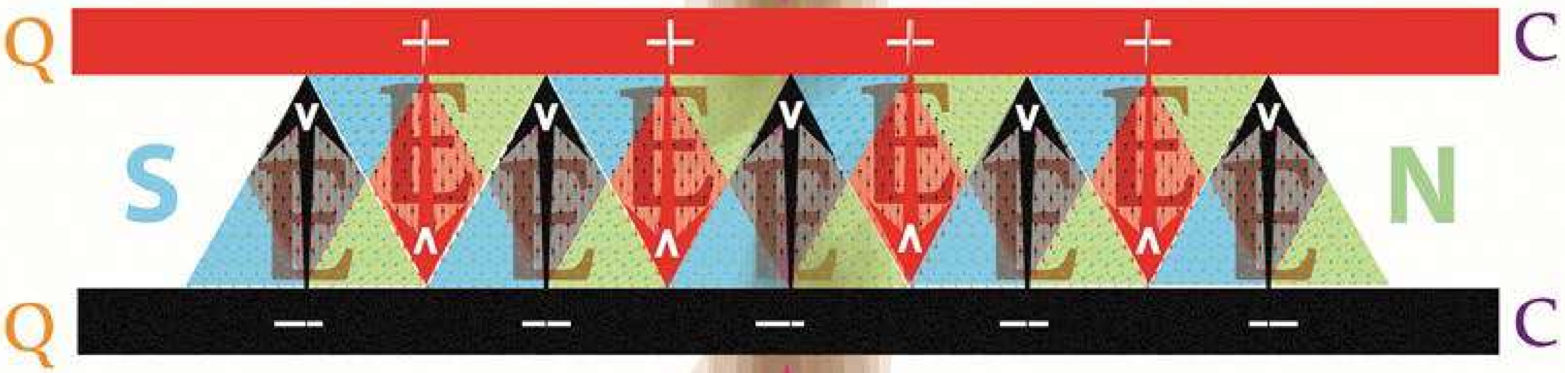

Voltage is
Joules per Coulomb

V

Charged Matter [current] does not
flow between the plates
of a Capacitor

ODD.hv

charge

A common form of energy storage device is a parallel-plate capacitor whose
capacitance is directly proportional to the surface area of the conductor plates
and inversely proportional to the separation distance between the plates.

volts

hv^2

Capacitance

Copyright ABRAHAM [2008] - all rights reserved

Capacitance

The SI unit of capacitance is the farad (symbol: F), named after the English physicist Michael Faraday

In general any separated charges will create a emf voltage

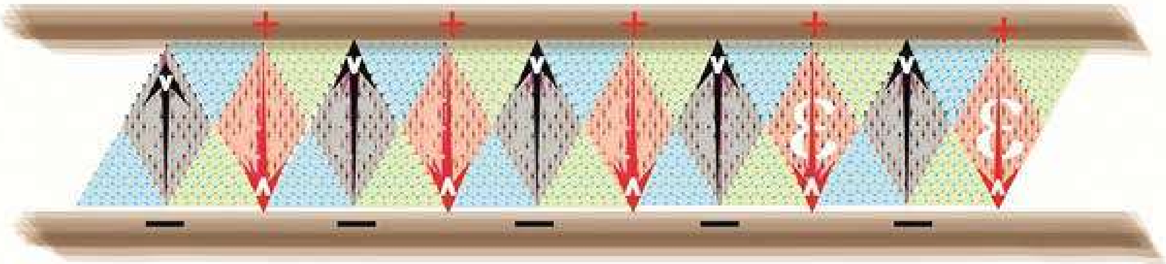

Any two electrical conductors separated by a non-conducting (or very high resistance) medium is a capacitor
[these can be plates, conductive wires or coaxial cables etc]

Plate capacitors

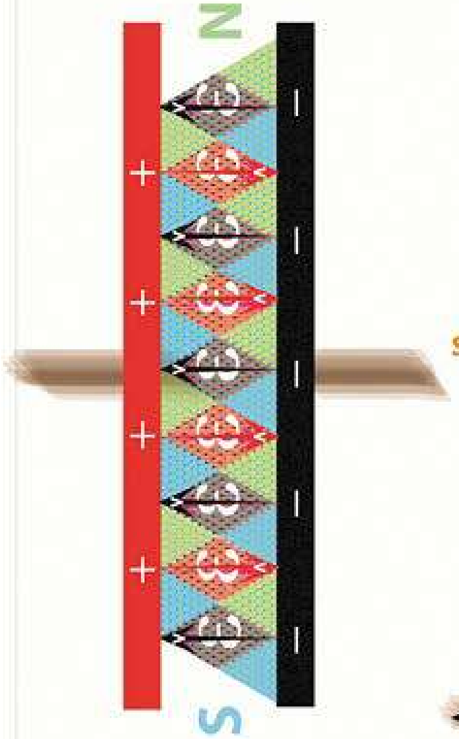

Coaxial cables

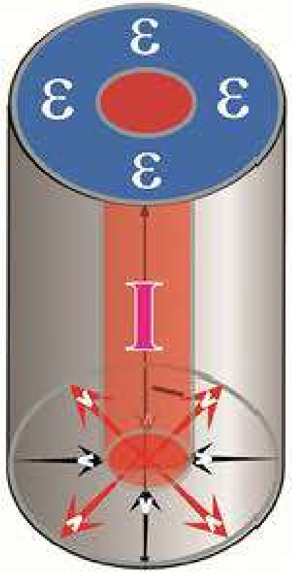

Capacitance is the ability of a body to store an electrical charge.

$$C = \frac{Q}{V}$$

Any object that can be electrically charged exhibits capacitance

A 1 farad capacitor when charged with1 coulomb of electrical charge will have a potential difference of 1 volt between its plates.

should a current flow in the conductors the capacitive field will be dominated by orthogonal, superpositioned Amperian M-fields of greater strength

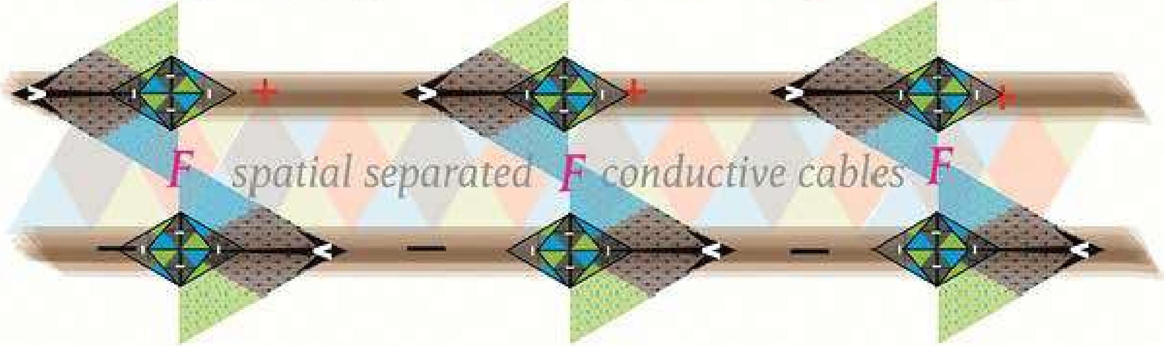

Capacitance is a function only of the physical dimensions (spatial geometry) of conductors and the permittivity of the dielectric separating them

Inductive and Capacitive energy storage

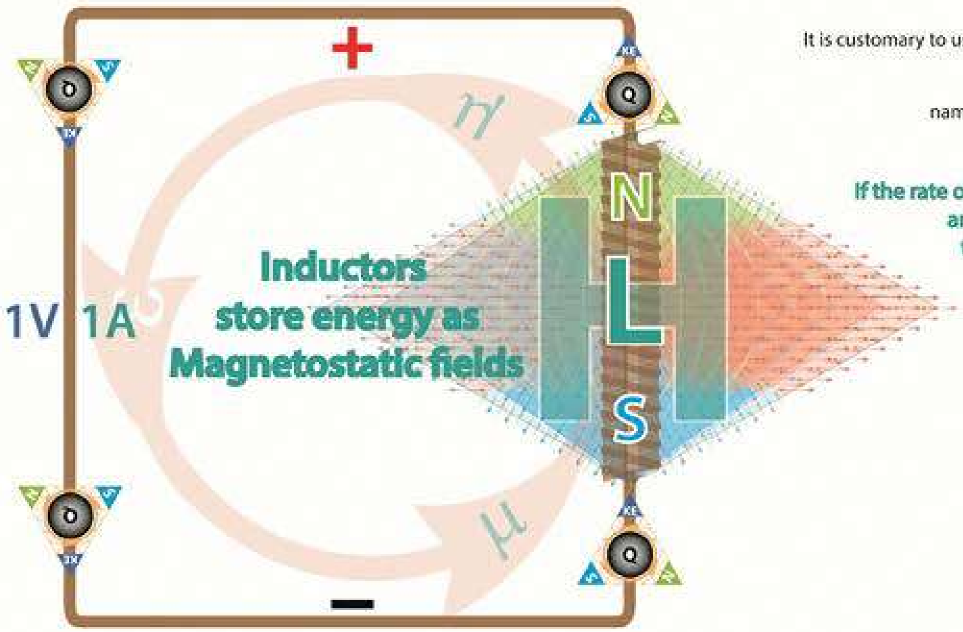

Inductors store energy as Magnetostatic fields

It is customary to use the symbol L for inductance, in honour of the physicist Heinrich Lenz.

In the SI system the unit of inductance is the Henry,
named in honor of the scientist who discovered inductance,
Joseph Henry.

If the rate of change of current in a circuit is one ampere per second and the resulting electromotive force is one volt, then the inductance of the circuit is one henry

$$v_m = \sum_{n=1}^{K} L_{m,n} \frac{di_n}{dt}.$$

The voltage across an inductor
is equal to the product of its
inductance and the time
rate of change of the current through it

Joseph Henry

[17 December 1797 – 13 May 1878]

Michael Faraday

[22 September 1791 – 25 August 1867]

The SI unit of capacitance is the farad (symbol: F),
named after the English physicist Michael Faraday

a 1 farad capacitor when charged with 1 coulomb of electrical charge will have a potential difference of 1 volt between its plates.

$$W_{\text{stored}} = \frac{1}{2}CV^2 = \frac{1}{2}\varepsilon_r\varepsilon_0\frac{A}{d}V^2.$$

Capacitance is the ability of a body to store an electrical charge.
Any object that can be electrically charged exhibits capacitance.

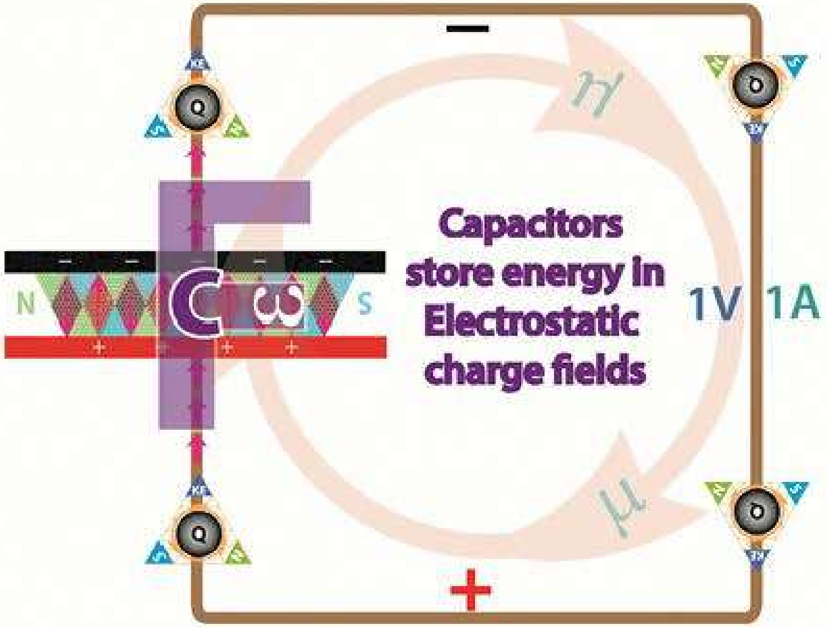

Capacitors store energy in Electrostatic charge fields

Inductive - Capacitive circuits

An LC circuit, also called a resonant circuit, tank circuit, or tuned circuit,
consists of an inductor, represented by the letter L,
and a capacitor, represented by the letter C.

LC

When connected together, they can act as an electrical resonator,
an electrical analogue of a tuning fork, storing energy
oscillating at the circuit's resonant frequency

These LC circuits are idealised models
assuming there is no dissipation of energy due to resistance over time,
with the circular EM vectors representing electron current flow in the circuit
and its translation between inductive and capacitive forms of energy storage.

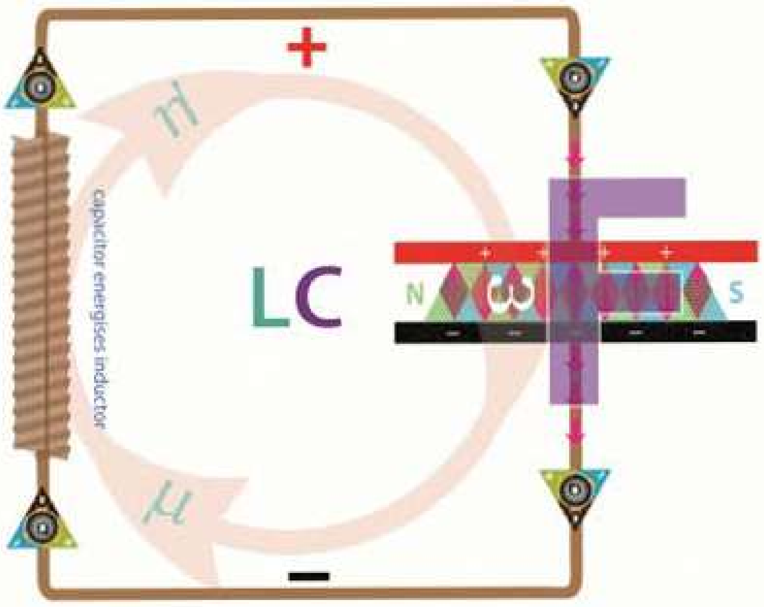

If a charged capacitor is connected across an inductor,
charge will start to flow through the inductor,
building up a magnetic field around it and
reducing the voltage on the capacitor.

Eventually all the charge on the capacitor will be gone
and the voltage across it will reach zero.

However, the current will continue,
because inductors resist changes in current.

The energy to keep it flowing is extracted from the magnetic field,
which will begin to decline.

The current will begin to charge the capacitor
with a voltage of opposite polarity to its original charge.

When the magnetic field is completely dissipated the current will stop
and the charge will again be stored in the capacitor,
with the opposite polarity as before.

Then the cycle will begin again,
with the current flowing in the
opposite direction through the inductor.

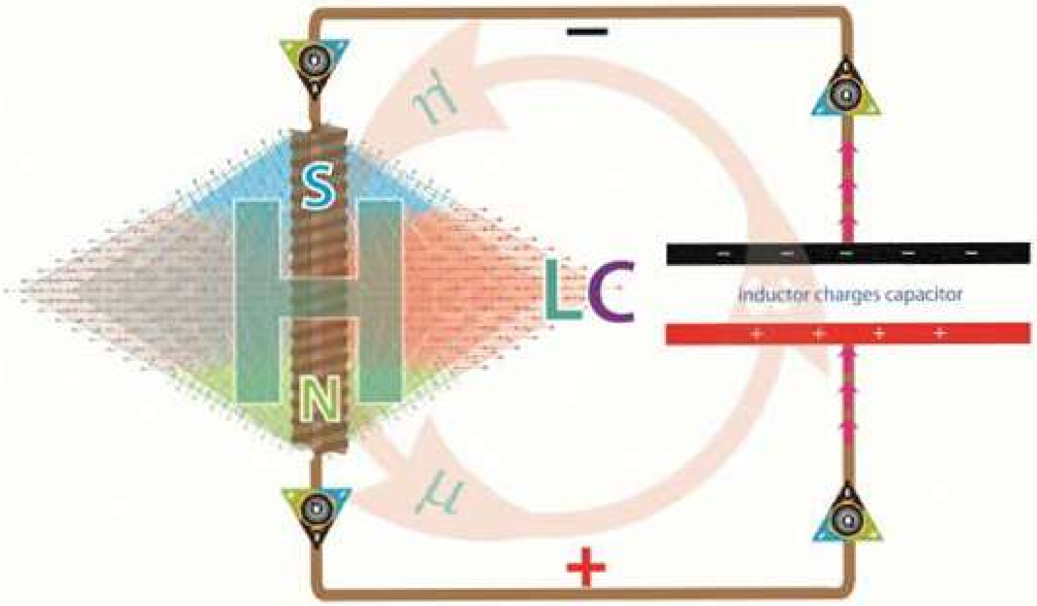

Capacitors store EM Energy as transverse Bosons

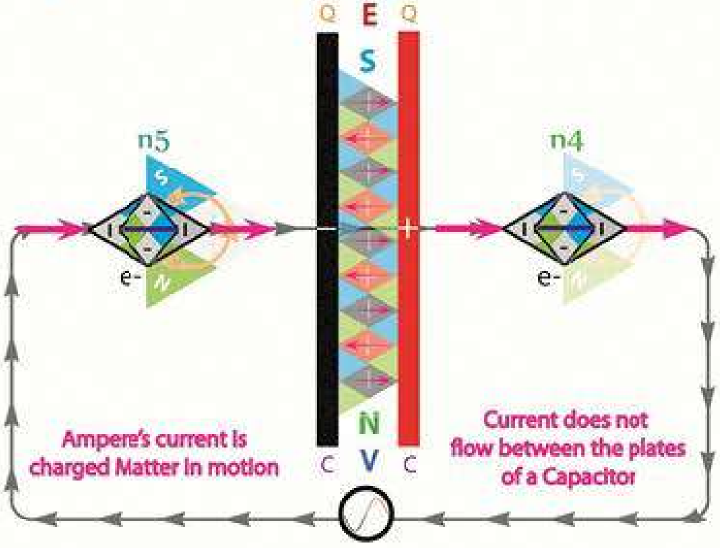

Ampere's current is charged Matter in motion

Current does not flow between the plates of a Capacitor

Capacitors store Energy as Joules per charge

Capacitors store & pass energy as charged masses

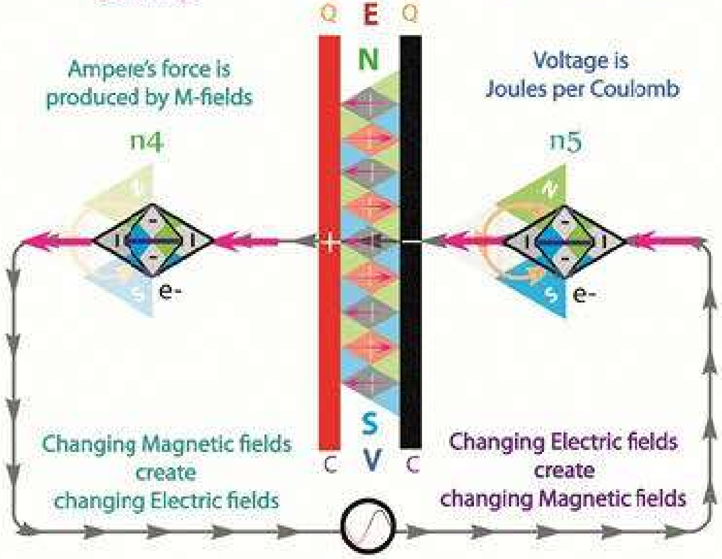

Ampere's force is produced by M-fields

Voltage is Joules per Coulomb

Changing Magnetic fields create changing Electric fields

Changing Electric fields create changing Magnetic fields

Capacitors store charged masses between their plates

Displacement Current

Maxwell's displacement current is better termed a 'displacement voltage' in light of the fact the charged Matter does not move/flow between the charged plates of a capacitor only the charged mass-energies.

Changing Electric fields produce changing Magnetic fields and vice versa as a direct result of their respective E&M field geometries within the Planck quanta of EM fields

$$\nabla \times \mathbf{B} = \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} + \mu_0 \mathbf{J},$$

As opposed to Maxwell's view of a 'stretching ether' that stores energy between the capacitive plates

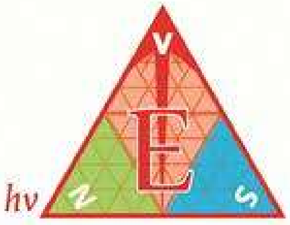

As the energy content of a charge increases so does its voltage

Equilateral energy momenta is stored compressively in the inductive quantum fields [ZPF] as it increases, and is released via charged exchange W-Bosons

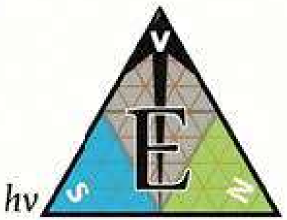

Bosons are transverse stores of energy momenta

Producing a capacitive region of ElectroMagnetic fields and stress energies

$$\mathbf{J}_D = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}$$

'Displacement Currents' between charged surfaces are produced by a increasing Electric Field component that seeks equilibrium

> *It must be carefully borne in mind that we have made only one step in the theory of the action of the medium. We have supposed it to be in a state of stress, but we have not in any way accounted for this stress, or explained how it is maintained. This step, however, seems to me to be an important one, as it explains, by the action of the consecutive parts of the medium, phenomena which were formerly supposed to be explicable only by direct action at a distance.*
>
> *Maxwell - On Physical Lines of Force (1861)*

mass

$$mc^2 = E = Mc^4$$

Matter

A clear geometric distinction can now be made between **EM mass geometries** and standing wave **Matter topologies**

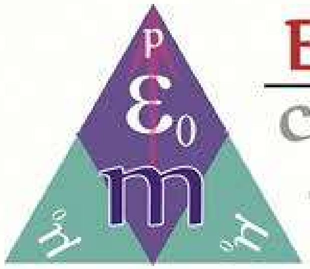

$$\frac{E}{c^2} = m$$

mass-energy momenta in a 2D equilateral geometry

Radiant waveform
EM mass-energies

$$M = \frac{E}{c^4}$$

Four mass-energy momenta in a 3D tetrahedral topology

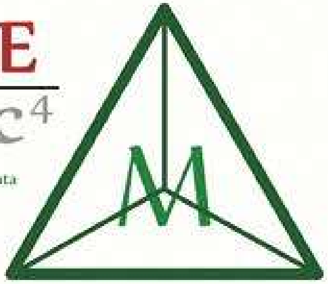

Standing waveform
EM energies

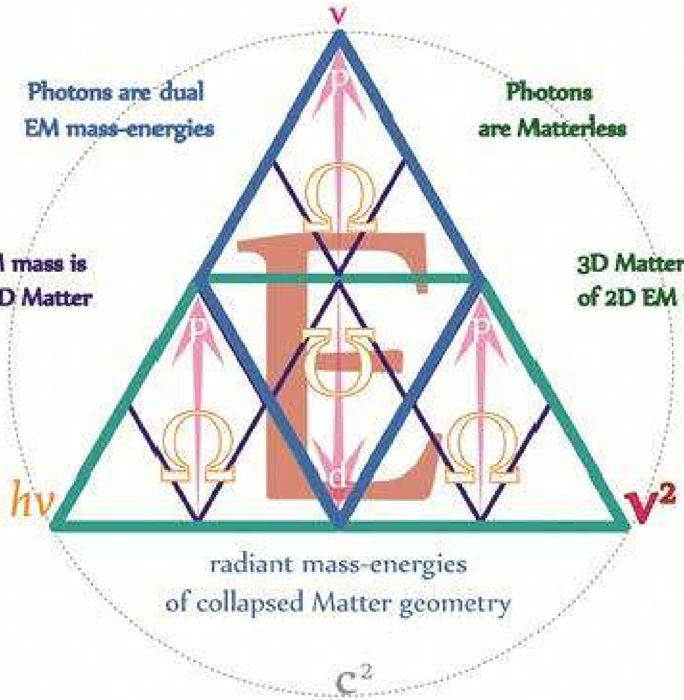

Photons are dual EM mass-energies

Photons are Matterless

2D EM mass is NOT 3D Matter

3D Matter is comprised of 2D EM mass-energies

radiant mass-energies
of collapsed Matter geometry

$$n\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

mass — EM Field — ElectroMagnetic — Planck — mass — quanta — velocity

$$T\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Matter — EM Field — ElectroMagnetic — Planck — mass — quanta — velocity

ρ/second

ρ/second^2

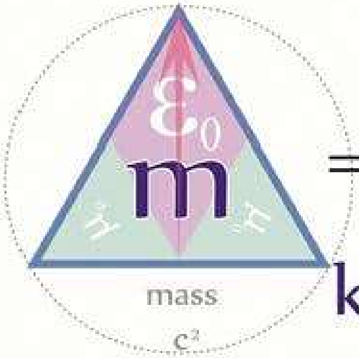

$$= m\Omega\left[\frac{v^2}{c^2}\right]$$

kg

Radiant EM masses

$$\frac{4\pi m}{c^2} = M$$

Standing-wave mass-Matter

$$m\Omega\left[\frac{v^2}{c^4}\right] =$$

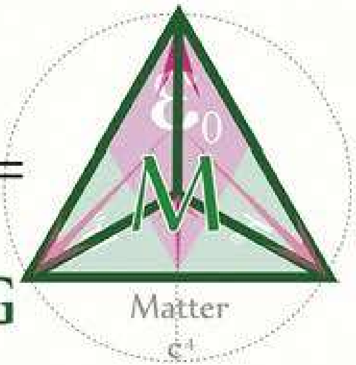

KG

massless particles are a scientific misnomer and massless must now be re-termed Matterless

de Broglie Matter-waves

In quantum mechanics, a Matter wave or de Broglie wave is the probabilistic electromagnetic quantum wave-form created by the Planck elements constituting Matter

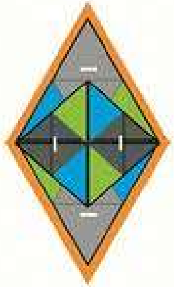

EM CHARGE provides the foundational interaction geometry for all mass & Matter

Relativistic mass-energy is the total sum of all Planck energy momenta in any spatial co-ordinate system

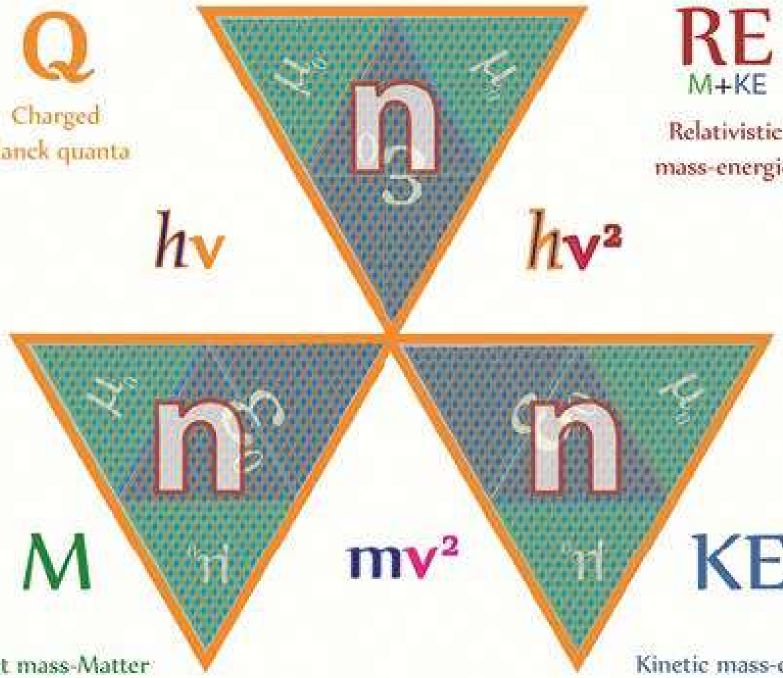

Q

Charged Planck quanta

RE

M+KE

Relativistic mass-energies

***h*v**

***h*v²**

mv²

All EM waves and Matter are made up of integral Planck quanta [any imbalance results in Charge] as the number of Planck quanta increases their associated wavelengths decrease

The relativistic stress tensor mass-energies of Matter in motion is the sum of its standing-wave mass-Matter topology and its Lorentz corrected kEM field mass-energy geometries

M

rest mass-Matter topologies

KE

Kinetic mass-energy geometries

3D MATTER is the charged topology created by standing-waves of 2D mass-energies

Kinetic EM mass-energy levels depend on the velocity of the Matter particle in motion

[in Matter energy travels in a standing wave at the speed of Light]

Kinetic EM fields are the mass-energy fields created by Matter in motion

P
KE
N
S
magnetic moment

The rest mass-energy geometries comprising all Matter topologies are velocity invariant

Copyright ABRAHAM [2008] - all rights reserved

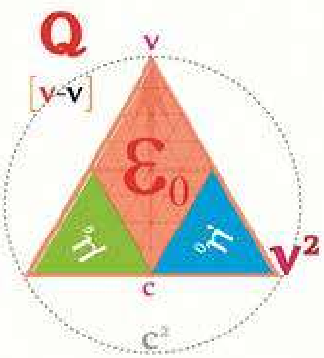

Material EM masses

deBroglie wavelength & Compton frequency

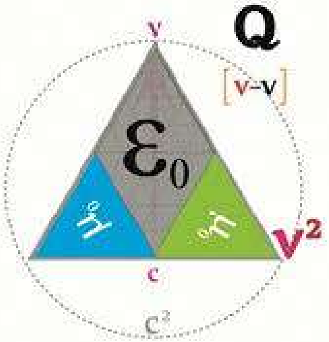

λ

Any increase in Energy quanta results in
a decrease in EM mass quanta wavelengths

v^2

In Matter all energy propagates at c

$$mv^2 = E = hv^2$$

in Standing wave Matter
v=c

Number of Quanta comprising rest Matter

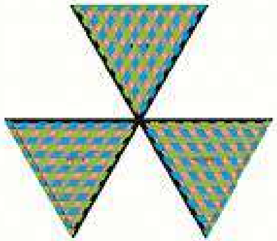

Compton Frequency

$$\frac{mc^2}{h} = 1\text{x}10^{19}$$

Electron Compton Frequency
$= 1.2 \text{ x } 10^{20}$ quanta

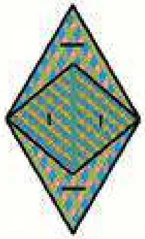

v^2

Wavelength of Quanta comprising rest Matter

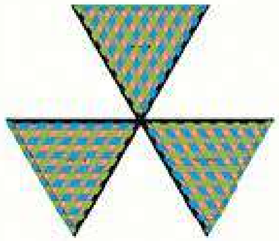

de Broglie wavelength

$$\frac{h}{mc} = 2.99792456 \text{ x } 10^{-33} \text{ m}$$

Electron de Broglie wavelength
$= 4.002769142 \text{ x } 10^{-11} \text{ m}$

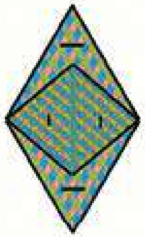

λ

Linear momentum of standing wave Energy comprising Matter

Intrinsic Particle Momenta

$$mc = 2.211340633 \text{ x } 10^{23} \text{ Ns}$$

Intrinsic Electron momenta
$= 2.95253793 \text{ x } 10^{00} \text{ Ns}$

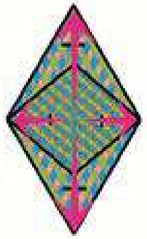

p

The Compton frequency, deBroglie wavelength
and Energy momenta of any physical system
are all related through the speed of Energy [c]

The examples above are for stationary Electrons
any motion will NOT affect the results for the electron itself
(however a extended KEM field will be produced by the motion
and its properties will be affected by changes in velocity)

In Lorentz invariant Matter the standing wave
Energy always propagates at the speed of Light [c]
(with the KEM field subject to Lorentz corrections)

KEM Quantum levels

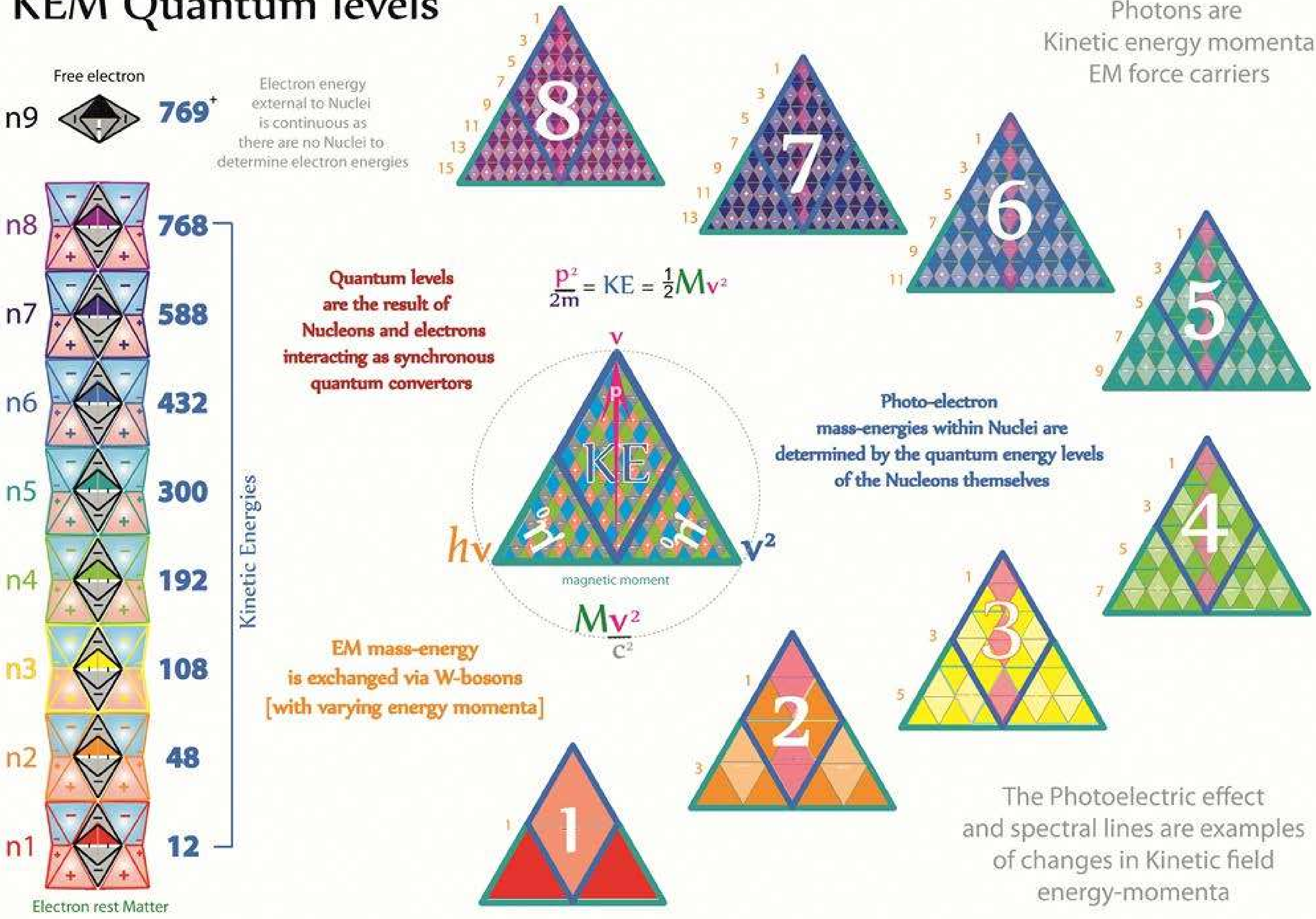
Free electron
n9
769+
n8
768
n7
588
n6
432
n5
300
n4
192
n3
108
n2
48
n1
12
Kinetic Energies
Electron rest Matter
8
1 3 5 7 9 11 13 15
7
1 3 5 7 9 11 13
6
1 3 5 7 9 11
5
1 3 5 7 9
4
1 3 5 7
3
1 3 5
2
1 3
1
1
v
p
KE
hv
v²
magnetic moment
Mv²/c²

Electron energy
external to Nuclei
is continuous as
there are no Nuclei to
determine electron energies

Photons are
Kinetic energy momenta
EM force carriers

$$\frac{p^2}{2m} = KE = \frac{1}{2}Mv^2$$

Quantum levels
are the result of
Nucleons and electrons
interacting as synchronous
quantum convertors

Photo-electron
mass-energies within Nuclei are
determined by the quantum energy levels
of the Nucleons themselves

EM mass-energy
is exchanged via W-bosons
[with varying energy momenta]

The Photoelectric effect
and spectral lines are examples
of changes in Kinetic field
energy-momenta

Copyright ABRAHAM [2008] - all rights reserved

deBroglie relationships

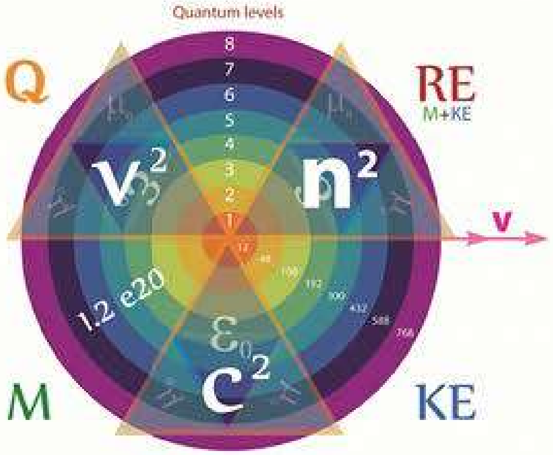
Quantum levels
8
7
6
5
4
3
2
1
Q
RE
M+KE
v²
n²
v
1.2 e20
ε₀
c²
M
KE
12
48
108
192
300
432
588
768

rest mass-Matter is
velocity invariant

*" The electron which is moving in a sine wave circular path,
will repeat the same sine wave path in each successive orbit.
The sine wave paths in consecutive orbits will exactly overlap.
The electron wave reconnects with itself and is in phase with itself. "*

An electron bound in a atomic nuclei will trace
a toroidal path around the nuclei mapped
by the sine wave motion of any point
on its charged geometric surface

$$\lambda = \frac{h}{p}$$

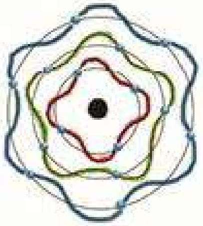

$$f = \frac{E}{h}$$

Louis de Broglie

(15 August 1892 – 19 March 1987)

***The relativistic energy of Matter in motion
is a measure of its standing wave mass-energies,
and the object's velocity related kinetic energy and
magnetic moment in the form of a KEM wave geometry***

$$p = \frac{h}{\lambda} = \frac{h}{2\pi}\frac{2\pi}{\lambda} = \hbar k$$

Although Mathematically correct
the illustrative model commonly used to
depict the deBroglie relationships
is incorrect

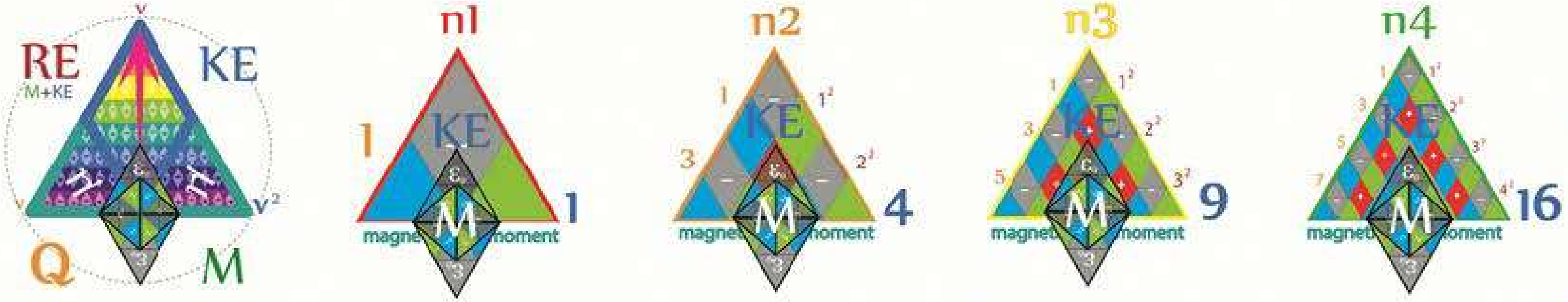
RE
M+KE
KE
v
v²
Q
M
n1
1
KE
M
magnetic moment
1
n2
4
n3
9
n4
16

The greater the
KEM field energy level
the smaller the
quantum wavelength
of the KEM field's EM mass

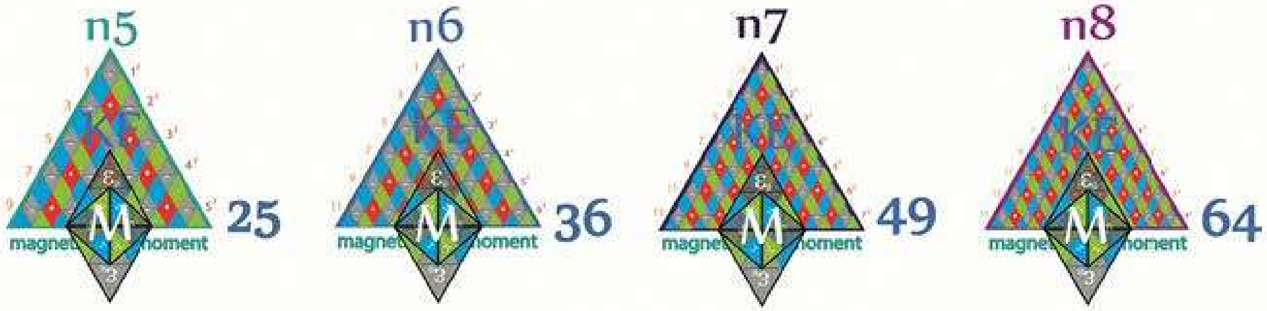
n5
25
n6
36
n7
49
n8
64
KE
M
magnetic moment

**All mass-Matter topologies are
Lorentz velocity invariant,
only the EM mass-energy content
and quantised angular momentum
of the Kinetic EM fields vary**

Copyright ABRAHAM [2008] - all rights reserved

Electron Positions in Atomic Orbitals

Werner Heisenberg

(5 December 1901 – 1 February 1976)

Erwin Schrödinger

(12 August 1887 – 4 January 1961)

Atomic orbitals are typically described as "hydrogen-like" (meaning one-electron) wave functions over space,
categorized by n, l, and m quantum numbers, [as covered in Tetryonic Chemistry] which correspond to
the electron's energy, angular momentum, and an angular momentum vector component, respectively

Using Tetryonics as a EM field model, an electron's position and velocity CAN be modelled simultaneously (but any attempt to measure or interact with it, will affect its energy levels)

Quantum Mechanics is a statistical [mathematical] representation of equilateral, charged geometries and EM energy interactions

The unique 12 facet topology
of Leptons results in the same geometry
being presented every 120° rotation of the particle

Leacing to the interpretation that
the Lepton disappears and re-appears
when being measured

Making accurate measurement of
its rotational dynamics incorrect
without correct models

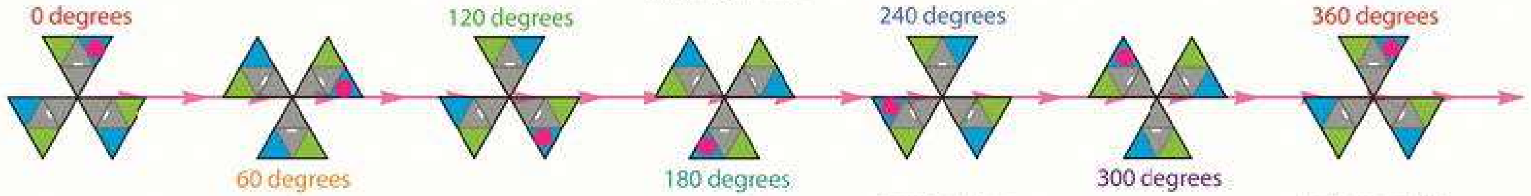

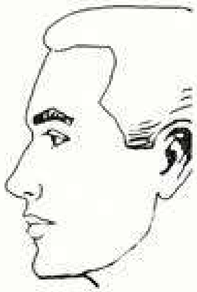

Additionally, every radial arm
of the Lepton's topology is
identical to every other

Including Electric fields and
Magnetic dipole orientations

Tetryonics resolves
these issues with
an accurate 3D-model
of Leptons

And Quantum Levels
thus providing a solid
foundation for explaining
Electro-Dynamics

Leptons are 6 loop quantum loop inductive rotors

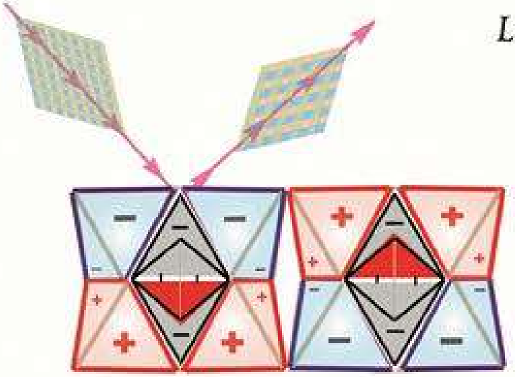

Lepton's are Spin 1 particles
that can easily be misinterpreted
as 1/2 or 3 quantum spin numbers
without the correct physical models
to interpret physcical observations on

The charged facia presented
changes continously, with
a complete fascia change
occuring in every
30 degrees of rotation

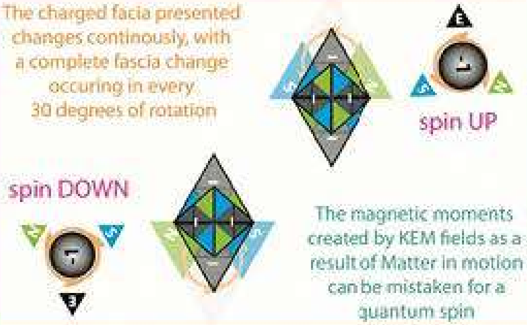

The magnetic moments
created by KEM fields as a
result of Matter in motion
can be mistaken for a
quantum spin

Distributions and Uncertainty

Probabilities are the square of the Amplitude

All probabilities are re-normalisable and sum to Unity

A Bell Curve (Normal Distribution) is a mathematical reflection of the integral Planck mass-energy momenta quanta distributions found in all equilateral charge geometries

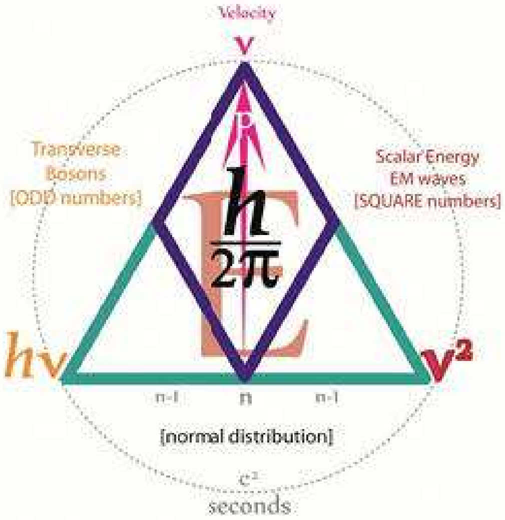
Velocity
v
Transverse
Bosons
[ODD numbers]
Scalar Energy
EM waves
[SQUARE numbers]
$\frac{h}{2\pi}$
E
hv
v²
n-1
n
n-1
[normal distribution]
c²
seconds

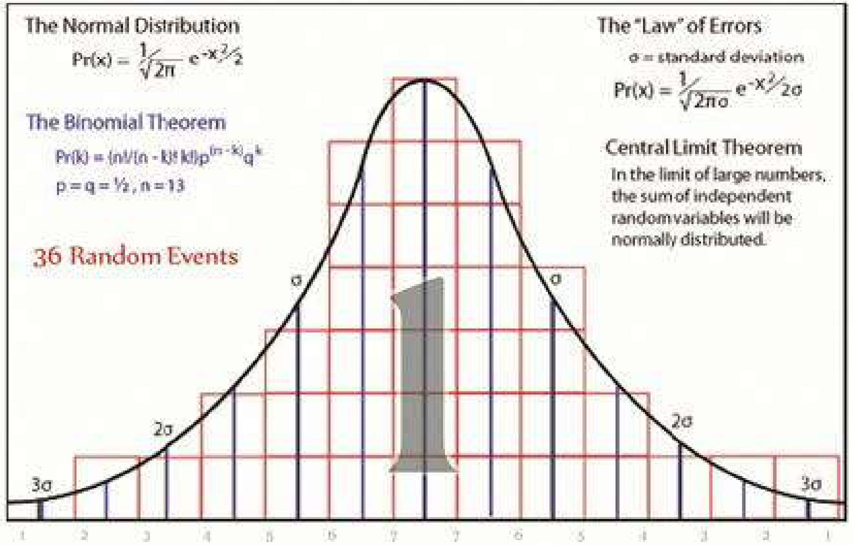

In the limit of large numbers, the sum of independent random variables will be normally distributed.

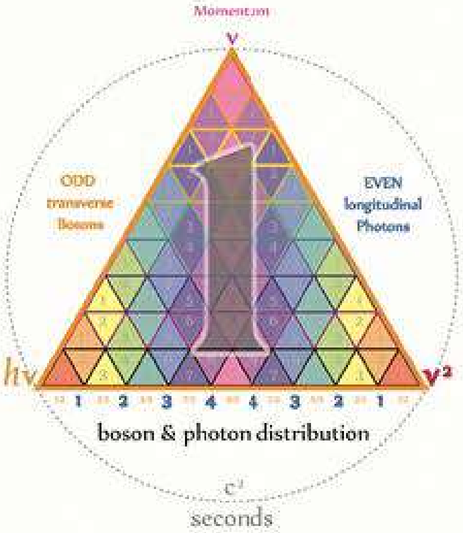
Momentum
v
ODD
transverse
Bosons
EVEN
longitudinal
Photons
hv
v²
1 2 3 4 4 3 2 1
boson & photon distribution
c²
seconds

All Matter and EM fields are comprised of equilateral mass-energy momenta and any attempt to measure the system involves the introduction of additional Planck EM quanta into the systems

John Stewart Bell

(28 June 1928 – 1 October 1990)

No physical theory of local hidden variables can reproduce all of the predictions of quantum mechanics.

WRONG

$$\Delta p\,\Delta x \geq \frac{1}{2}\hbar$$

Quantum uncertainty formulation

$$\Delta E\,\Delta t \geq \frac{1}{2}\hbar$$

The more precisely the position is determined, the less precisely the momentum is known in this instant, and vice versa.

Heisenberg, uncertainty paper, 1927

WRONG

Werner Heisenberg

(5 December 1901 – 1 February 1976)

Chance is closely related to the ideas of uncertainty and indeterminacy.
Uncertainty today is best known from Werner Heisenberg's principle in quantum mechanics.
It states that the exact position and momentum of an atomic particle can only be known within certain limits.
The product of the position error and the momentum error is equal to a multiple of Planck's constant of action.
This irreducible randomness in physical processes established the existence of chance and indeterminism in physics

Leptronic Self-energies

As the velocity of a charged particle increases
the energy level of its KEM field increases

KEM field
Energy
Planck quanta
mass
velocity
momenta

$$\frac{4n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass

radiant energy per second

$$E/c^2$$

KEM field geometries generated as a product of Matter in motion are Lorentz variant

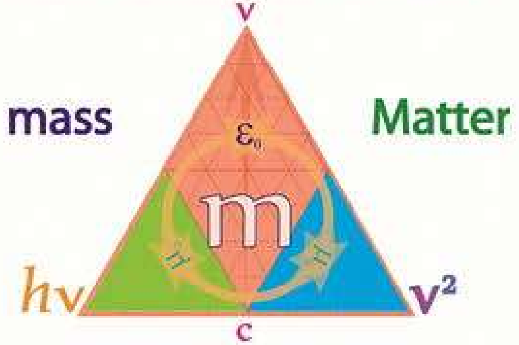
v
mass
Matter
ε₀
m
hv
v²
c

Quantum inductive loops resist changes to energies contained in their equilateral energy fields [inertial mass]

Matter
Energy
Planck quanta
mass
velocity
momenta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Matter

standing-wave energy per second squared

$$E/c^4$$

rest mass-Matter topologies are not affect by changes in velocity [Matter is Lorentz invariant]

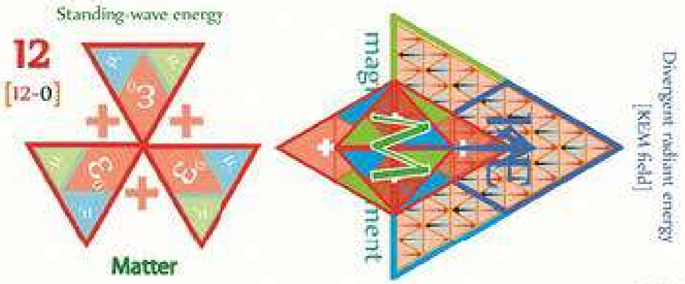
Standing-wave energy
12
[12-0]
Matter
magnet
Divergent radiant energy
[KEM field]

The Law of Interaction

The Energy momenta of an electron's KEM field polarises the region surrounding its Matter and creates fields of interaction through the super-positioning of these fields

a particle's self-energy represents the contribution to the particle's energy, or effective mass, due to interactions between the particle and the system it is part of.

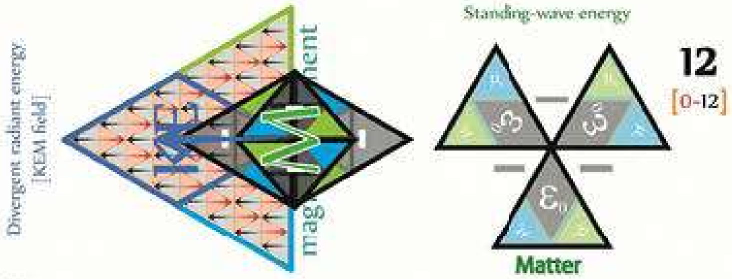
Divergent radiant energy
[KEM field]
magnet
Standing-wave energy
12
[0-12]
Matter

$$KEM = Mv^2$$

All Matter topologies are made up of quantum inductive loops which extract energy from any EM field they move through

Matter stores energy in its 3D planar facsia as charged mass-energy

additional to creating the familiar laws of interaction
[opposite attract - similars repel]
the KEM fields of leptons can act as tuned antennas
extracting energy from their environment

KEM fields store energy in their 2D planar EM field as neutral divergent mass-energies

$$E = Mc^4$$

All Matter contains EM energy propagating in a standing wave topology at the speed of light (the source of inertial mass)

Resonant EM fields

Transmitting in order to efficiently recieve
via resonant inductive coupling
of EM fields

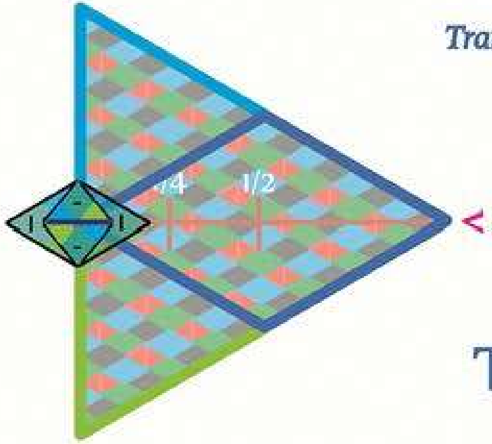

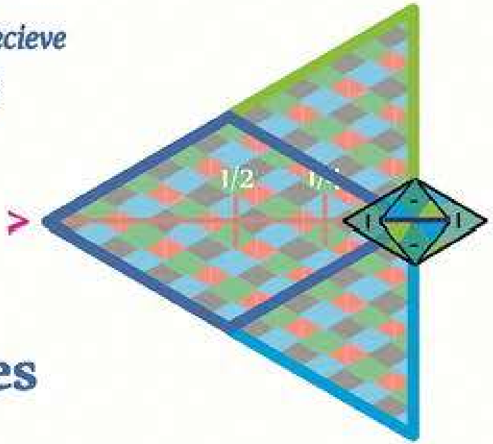

1:1

[or other harmonics]

Tuned KEM waves

Quantum inductive antenna fields

Historically, all recieiving antennas have been mechanically tuned to match the sought EM wavelength
But if power is applied to a small wavelength electrostatic topology it creates a KEM field
that produces a much larger, tuned rhombic antenna field

EM waves are hf geometries

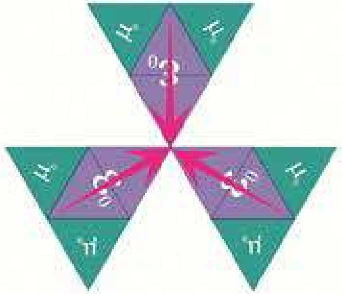

Electron's rest Matter
stores EM mass-energy
in a standing wave

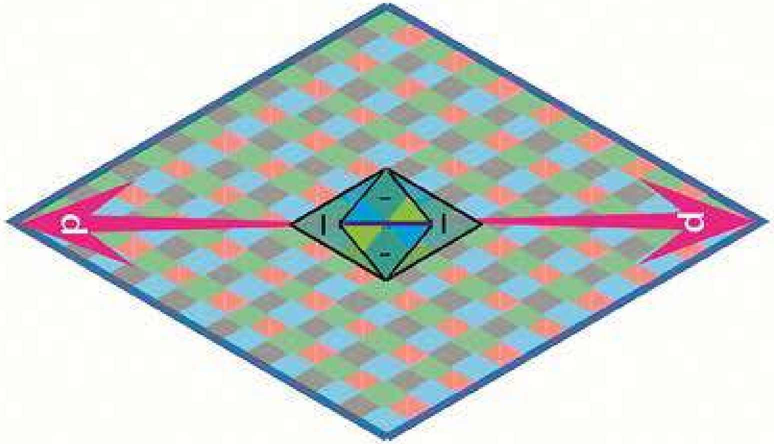

EM bosons have $h\nu$ geometries

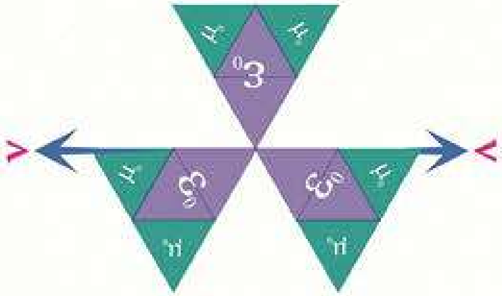

Electrostatic particles of Matter
have dual KEM energy fields that
can act as resonant antennas

Bosons

All Light is made of transverse EM energy Quanta

EM Field Planck quanta

$$1\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Bosons ElectroMagnetic mass velocity

Energy is quantised

Max Planck

$$E = n.hv$$

All Photons have mass-Energy and momentum

$$\lambda = \frac{h}{p} = \frac{h}{mv}\sqrt{1-\frac{v^2}{c^2}}$$

James Clerk Maxwell

$$\nabla\cdot\mathbf{D} = 4\pi\rho$$

$$\nabla\times\mathbf{H} = \frac{4\pi}{c}\mathbf{J} + \frac{1}{c}\frac{\partial\mathbf{D}}{\partial t}$$

$$\nabla\times\mathbf{E} + \frac{1}{c}\frac{\partial\mathbf{B}}{\partial t} = 0$$

$$\nabla\cdot\mathbf{B} = 0$$

ElectroMagnetic waveforms propagate at the speed of light

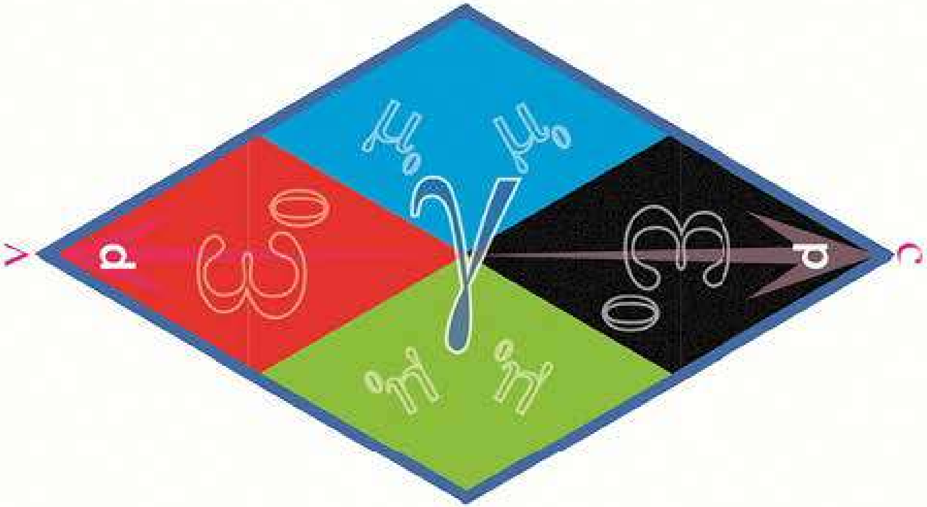

Matter is a probabilistic wave of EM particles

p

Louis de Broglie

Erwin Schrodinger

Ψ

Photons

All Light consists of longitudinal harmonic oscillators

EM Field Planck quanta

$$2\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Photons ElectroMagnetic mass velocity

Photons are particles

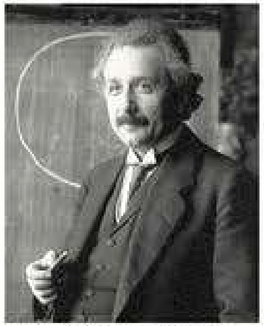
Albert Einstein

$$E = hf$$

All EM waves and Matter exhibit a Wave-Particle duality

$$i\hbar\frac{\partial}{\partial t}\Psi = \hat{H}\Psi$$

Copyright ABRAHAM [2008] - all rights reserved

Photonic EM fields

In 1865, James Clerk Maxwell's prediction that light was an electromagnetic wave, [which was confirmed experimentally in 1888 by Heinrich Hertz's detection of radio waves], seemed to be the final blow to particle models of light.

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

James Clerk Maxwell

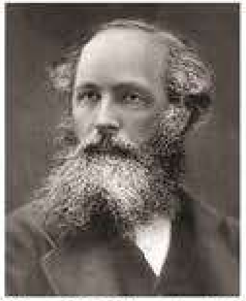

(13 June 1831 – 5 November 1879)

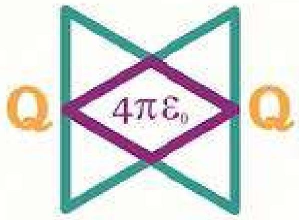

$$E = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}$$

$$\frac{K_E}{K_M} = 9e16\ \frac{\frac{N\,m^2}{C^2}}{\frac{N\,s^2}{C^2}}$$

$$c^2 = 9e16\ \frac{m^2}{s^2}$$

$$c = 299{,}792{,}458\ \frac{m}{s}$$

The speed of light can be calculated using the Electric and Magnetic constants

8.8541e-12

$$\frac{F}{m}\quad \varepsilon_0 \quad \frac{A^2\,s^4}{kg\,m^3}$$

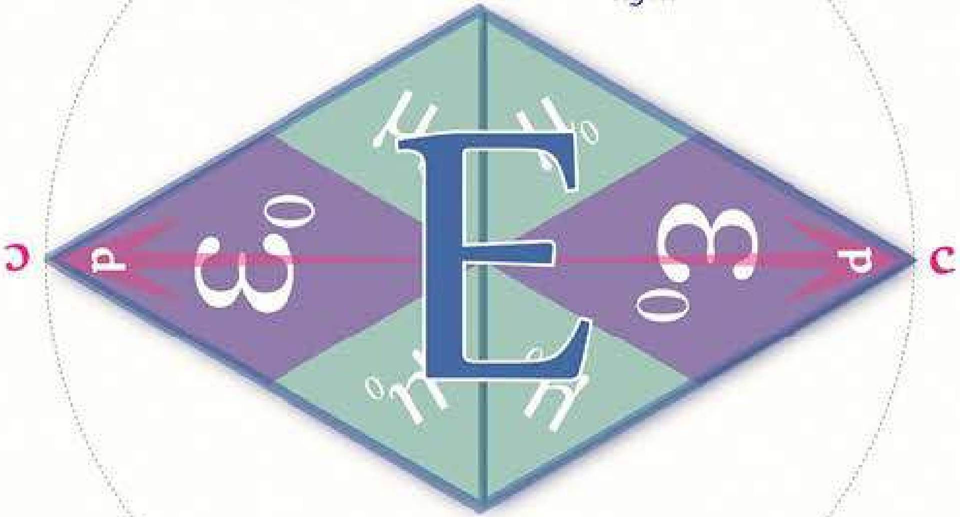

$$\frac{H}{m}\quad \mu_0 \quad \frac{kg\,m}{A^2\,s^2}$$

[4π]e-7

All EM mass-energy momenta propagates bi-directionally at c [the speed of light] in a vacuum

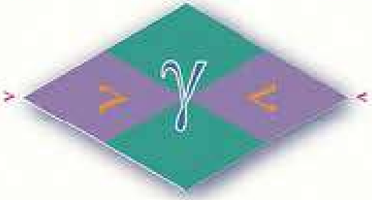

Electric field strength

The strength of Electric fields is determined by the Electrical Permittivity Constant

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

The strength of Magnetic fields is determined by the Magnetic Permeability Constant

$$B = \mu_0 H$$

Copyright ABRAHAM [2008] - all rights reserved

The ElectroMagnetic field geometry of Photons

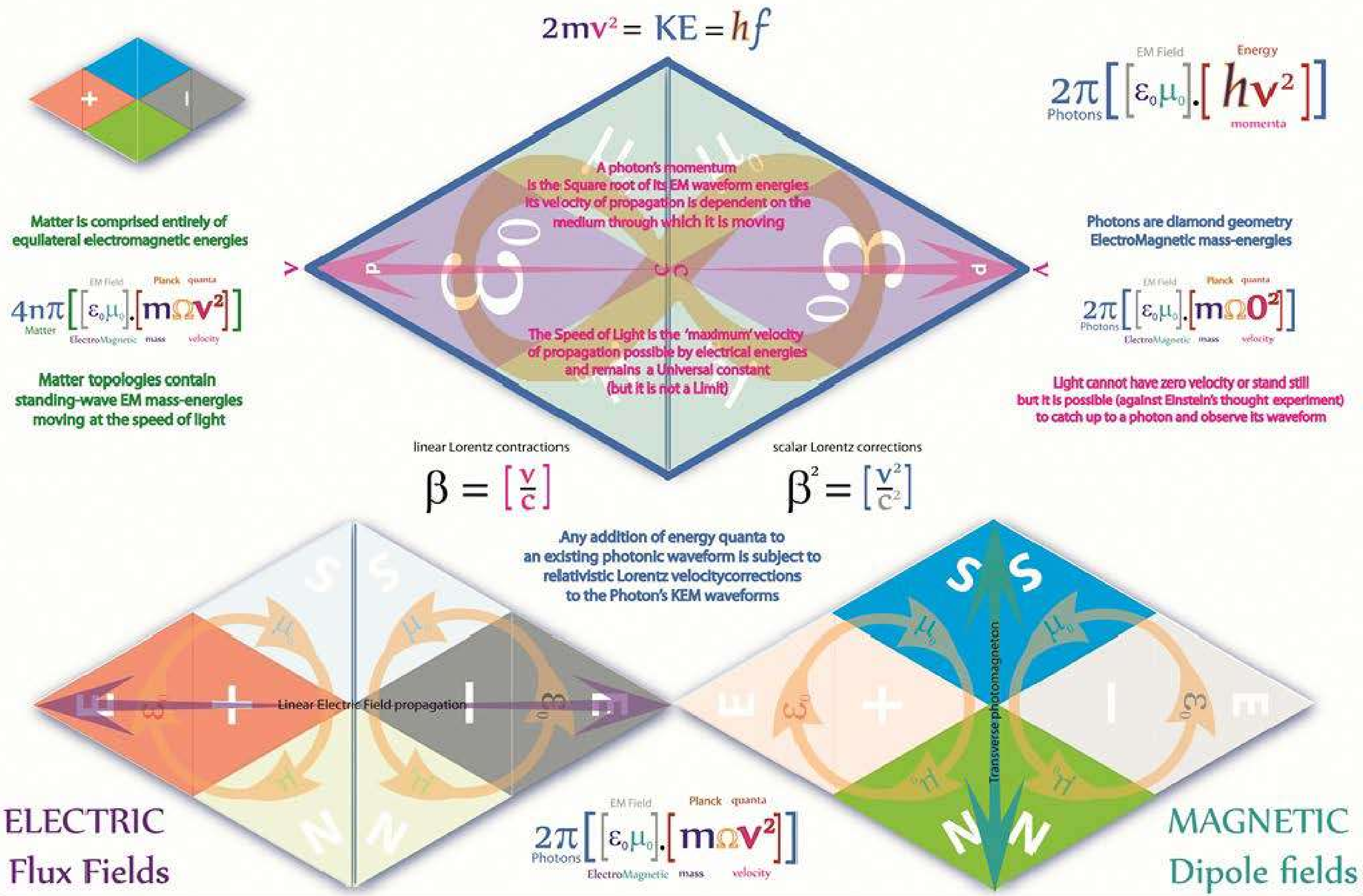

$$2mv^2 = KE = hf$$

$$\underset{\text{Photons}}{2\pi}\left[\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}\cdot\overset{\text{Energy}}{\underset{\text{momenta}}{[hv^2]}}\right]$$

A photon's momentum
is the Square root of its EM waveform energies
its velocity of propagation is dependent on the
medium through which it is moving

The Speed of Light is the 'maximum' velocity
of propagation possible by electrical energies
and remains a Universal constant
(but it is not a Limit)

Matter is comprised entirely of
equilateral electromagnetic energies

$$\underset{\text{Matter}}{4n\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}\cdot\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Matter topologies contain
standing-wave EM mass-energies
moving at the speed of light

Photons are diamond geometry
ElectroMagnetic mass-energies

$$\underset{\text{Photons}}{2\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}\cdot\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{0^2}\right]\right]$$

Light cannot have zero velocity or stand still
but it is possible (against Einstein's thought experiment)
to catch up to a photon and observe its waveform

linear Lorentz contractions

$$\beta = \left[\frac{v}{c}\right]$$

scalar Lorentz corrections

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Any addition of energy quanta to
an existing photonic waveform is subject to
relativistic Lorentz velocitycorrections
to the Photon's KEM waveforms

ELECTRIC
Flux Fields

$$\underset{\text{Photons}}{2\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}\cdot\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

MAGNETIC
Dipole fields

Tetryonics 26.03 - Photon EM field geometry

Copyright ABRAHAM [2008] - all rights reserved

Photons and Charge bosons

Photons are dual [2π]
neutral charge quanta

$$\underset{\text{Photons}}{2\pi}\left[\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{\varepsilon_0\mu_0}}\right].\left[\underset{\text{mass}}{m}\overset{\text{Planck}\quad\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Their Electric and Magnetic
fields are orthogonal to each other

**Bosons are transverse charge carriers
they combine to create ElectroMagnetic
photons which are longitudinal charge carriers**

They propagate
in differing mediums
at v/c

$$\underset{\text{EM waves}}{\text{EVEN}\,\pi}\left[\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{\varepsilon_0\mu_0}}\right].\left[\underset{\text{mass}}{m}\overset{\text{Planck}\quad\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

nπ electromagnetic waves are comprised of
numerous 2π photons of the same wavelength
(save in superpositioned states - White light)

Photons are dual Boson waveforms

$$E = 2h\nu$$

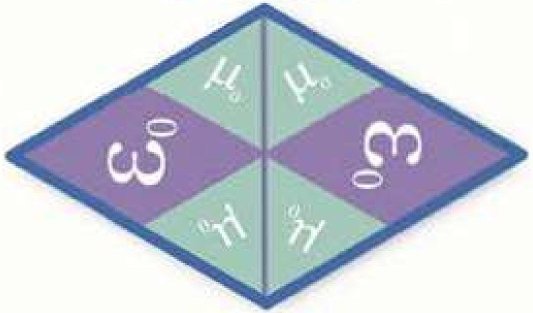

$$E = hf$$

**A moving charge creates a magnetic field throughout space
that is perpendicular to the direction of motion.**

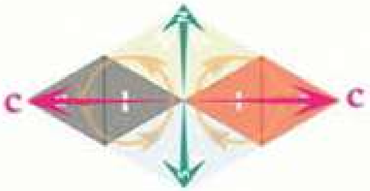

**Photons require NO ether to propagate
they are discrete bundles of EM energy-momenta**

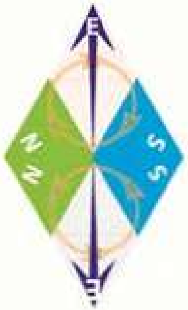

**Similarly, a magnet has an intrinsic neutral electric field
that is perpendicular to its Magnetic Dipole.**

Photon EM strength vectors

Electric field force vectors

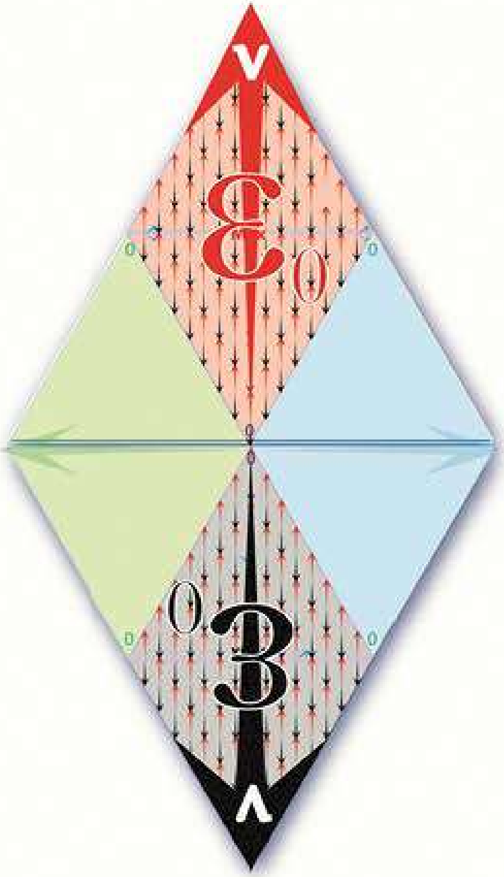
ε_0
0
0
0
0
$^0\varepsilon$

When E field flux is at Maximum
B field flux is at Minimum

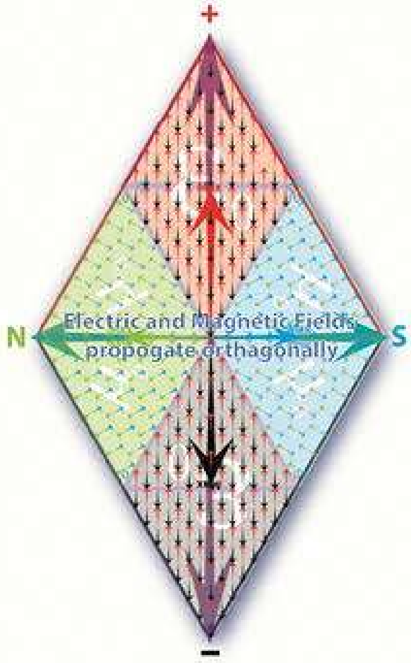

Electric and Magnetic Fields propogate orthagonally

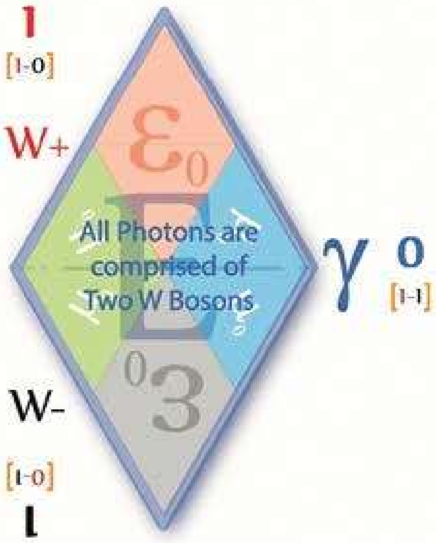
1
[1-0]
W+
ε_0
γ 0
[1-1]
$^0\varepsilon$
W-
[1-0]
1

All Photons are comprised of Two W Bosons

Magnetic field force vectors

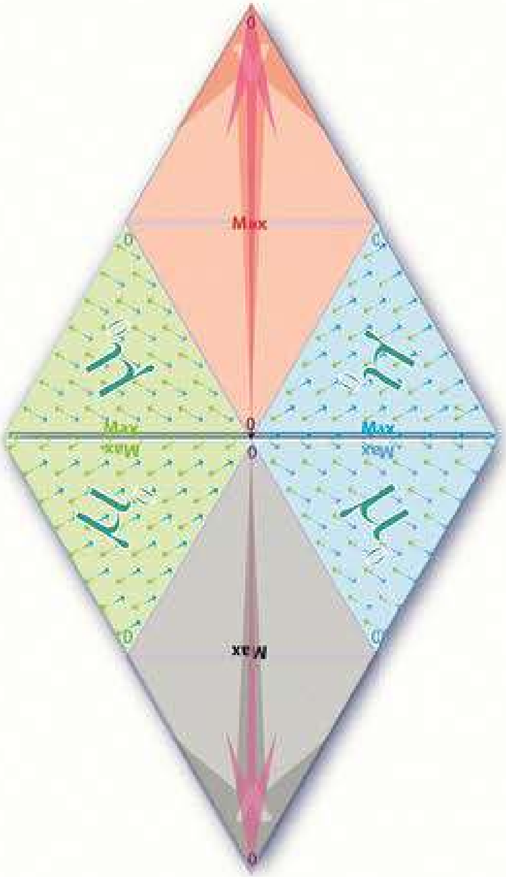
0
Max
Max
Max
Max
Max
μ_0
μ_0
μ_0
μ_0
0

When B field flux is at Maximum
E field flux is at Minimum

Photons are
2π electromagentic masses
possessing:

Frequency
Wavelength
Energy
Momentum
Kinetic Energy
Magnetic Moment
Probabilistic Properties

and can

Refract
Deflect and
Disperse

ZPFs [Bosons] are
1/2 wavelength photons

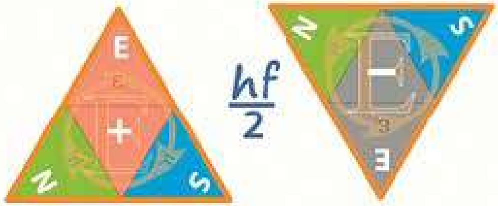

$$\frac{hf}{2}$$

$$2v = f$$

2π Planck Constants
Photon energy density

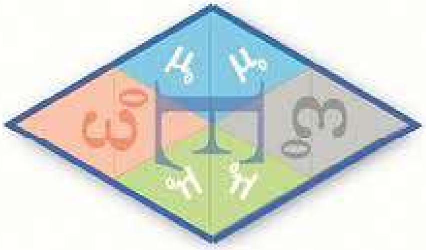

$$hf$$

Photons

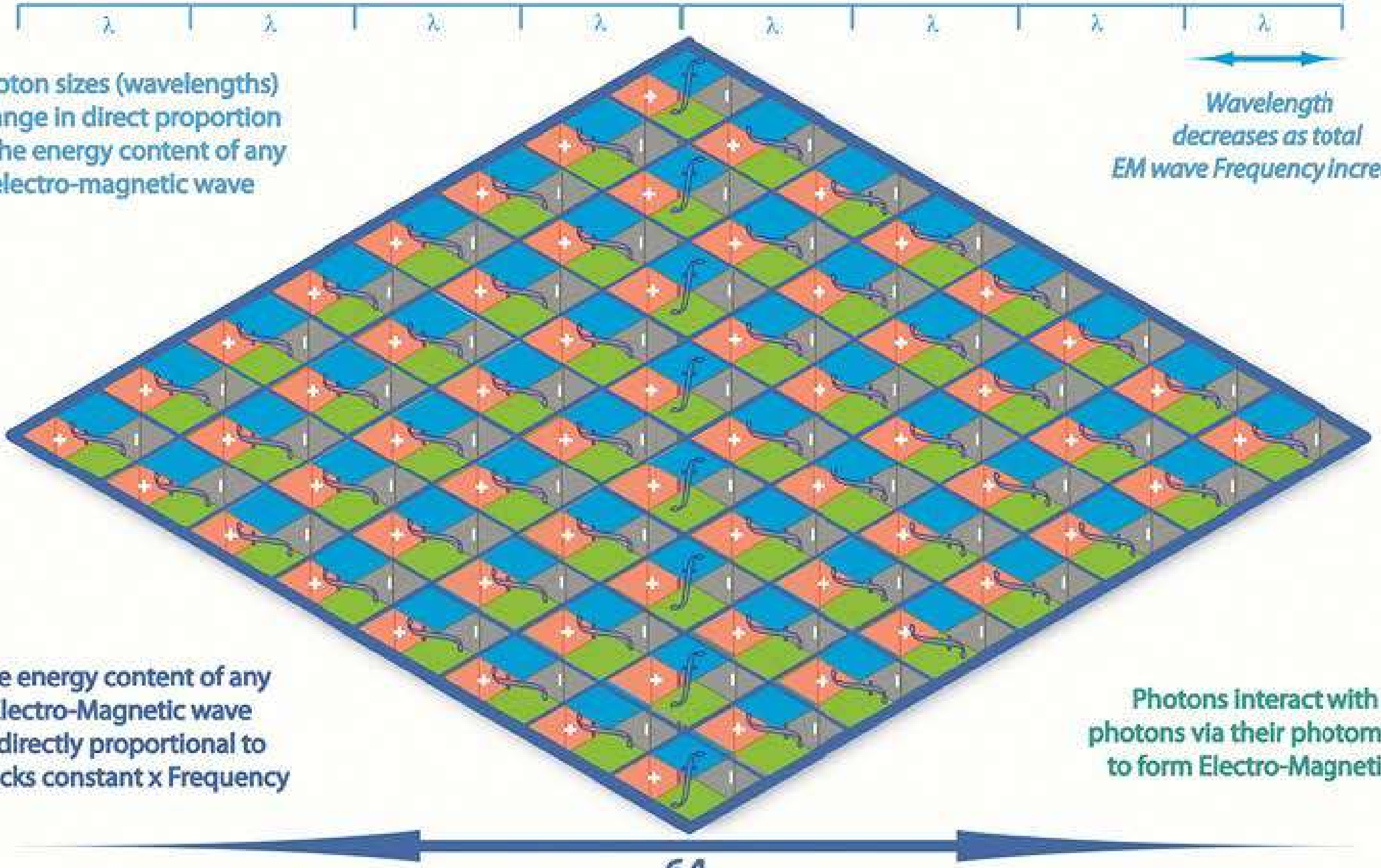

Photon sizes (wavelengths) change in direct proportion to the energy content of any electro-magnetic wave

Wavelength decreases as total EM wave Frequency increases

$$E = hf$$

The energy content of any Electro-Magnetic wave is directly proportional to Plancks constant x Frequency

Photons interact with other photons via their photomagnetons to form Electro-Magnetic waves

64
Frequency
Number of Cycles per time unit

Copyright ABRAHAM [2008] - all rights reserved

Bosons in a EM wave

EM waves are comprised of transverse W Bosons

Odd numbers
totaling to a
Square number

$$E = hv$$

$$\boxed{2v = f}$$

$$E = hf$$

Even numbers
totaling to a
Square number

EM waves are also comprised of longitudinal Photons

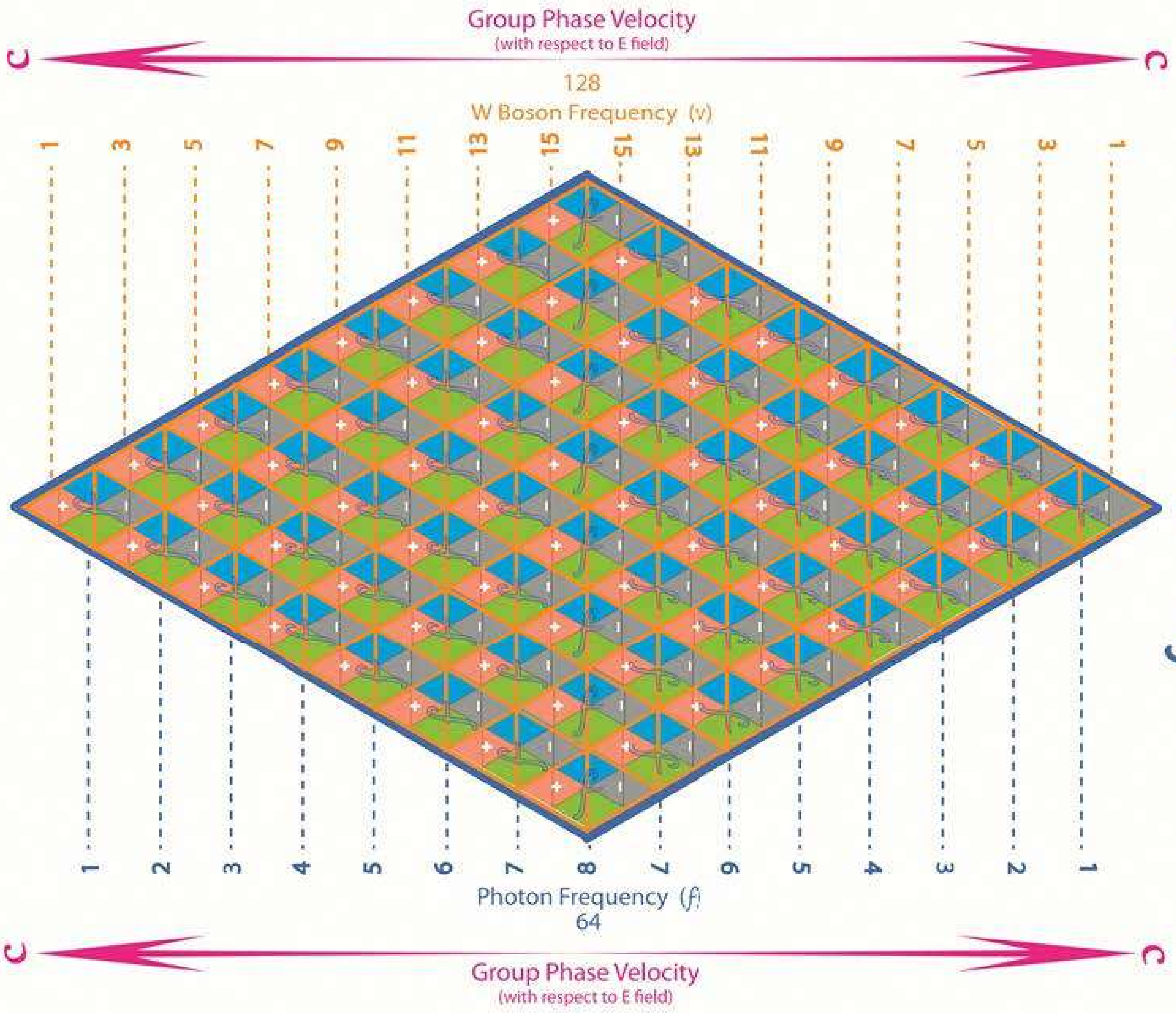
Group Phase Velocity
(with respect to E field)
c
c
128
W Boson Frequency (v)
1
3
5
7
9
11
13
15
15
13
11
9
7
5
3
1
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
Photon Frequency (f)
64
c
c
Group Phase Velocity
(with respect to E field)

$$2hv = E = hf$$

Tetryonics 26.07 - Bosons in a EM wave

Copyright ABRAHAM [2008] - all rights reserved

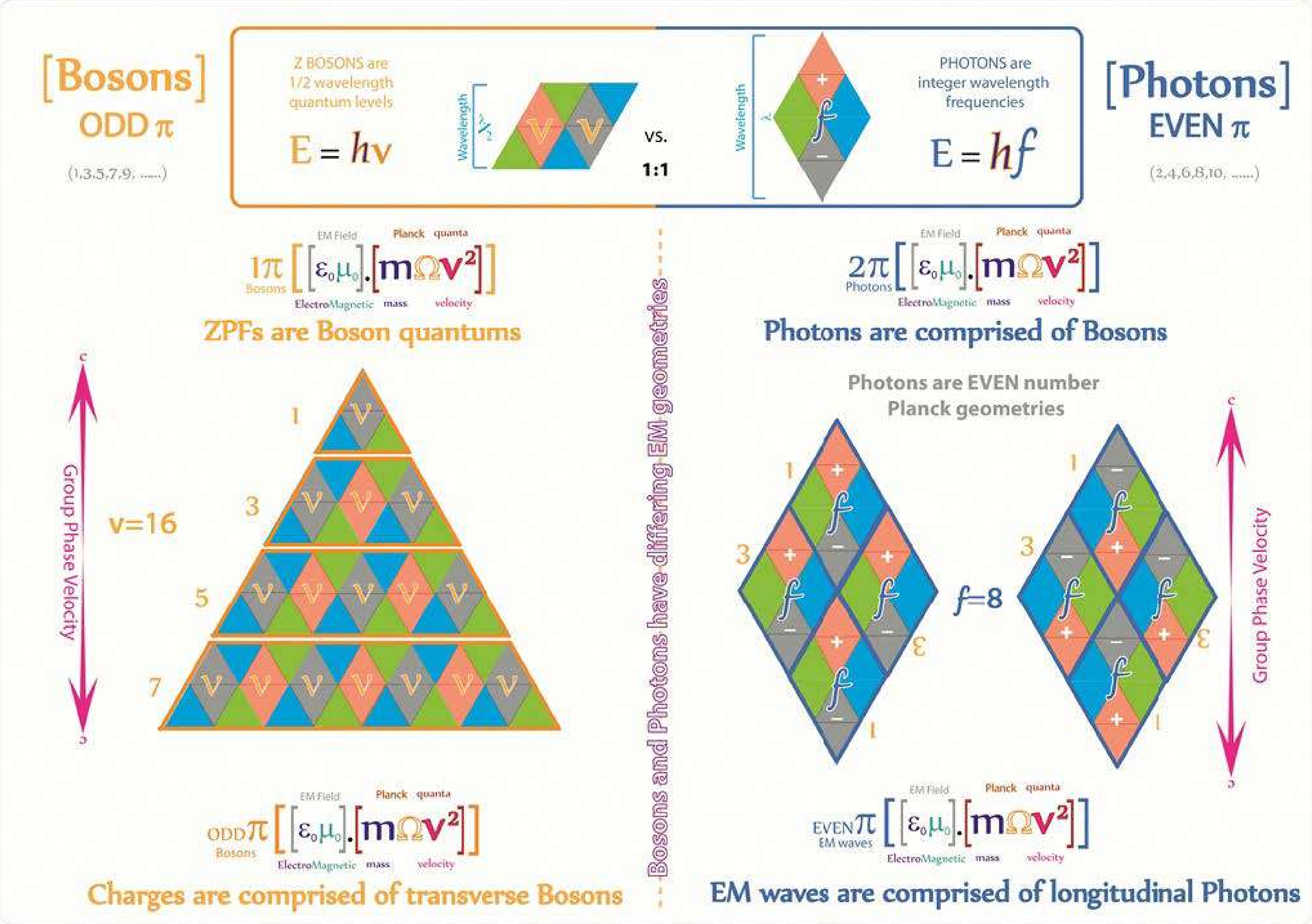

[Bosons]

ODD π

(1,3,5,7,9, ...)

Z BOSONS are
1/2 wavelength
quantum levels

$E = h\nu$

vs.

1:1

PHOTONS are
integer wavelength
frequencies

$E = hf$

[Photons]

EVEN π

(2,4,6,8,10, ...)

$$\underset{\text{Bosons}}{1\pi}\left[\left[\overset{\text{EM Field}}{\varepsilon_0\mu_0}\right].\left[\overset{\text{Planck quanta}}{m\Omega v^2}\right]\right]$$

ElectroMagnetic mass velocity

ZPFs are Boson quantums

$$\underset{\text{Photons}}{2\pi}\left[\left[\overset{\text{EM Field}}{\varepsilon_0\mu_0}\right].\left[\overset{\text{Planck quanta}}{m\Omega v^2}\right]\right]$$

ElectroMagnetic mass velocity

Photons are comprised of Bosons

Photons are EVEN number Planck geometries

$\nu=16$

$f=8$

Bosons and Photons have differing EM geometries

$$\underset{\text{Bosons}}{\text{ODD}\,\pi}\left[\left[\overset{\text{EM Field}}{\varepsilon_0\mu_0}\right].\left[\overset{\text{Planck quanta}}{m\Omega v^2}\right]\right]$$

ElectroMagnetic mass velocity

Charges are comprised of transverse Bosons

$$\underset{\text{EM waves}}{\text{EVEN}\,\pi}\left[\left[\overset{\text{EM Field}}{\varepsilon_0\mu_0}\right].\left[\overset{\text{Planck quanta}}{m\Omega v^2}\right]\right]$$

ElectroMagnetic mass velocity

EM waves are comprised of longitudinal Photons

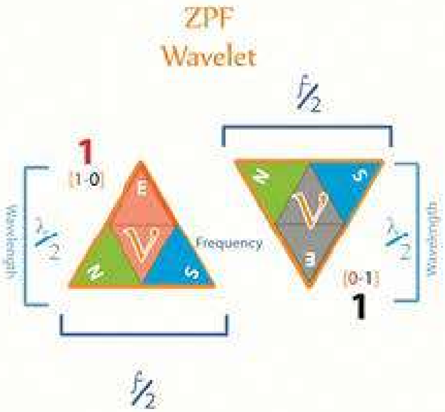
ZPF
Wavelet
f/2
1
[1-0]
E
N
S
ν
Wavelength
λ/2
Frequency
N
S
ν
E
λ/2
Wavelength
[0-1]
1
f/2

Bosons

Charge Carriers
$E = h\nu$

W+ boson

Bosons are half wavelength - non-Integer frequency waveforms due to their odd ZPF numbers

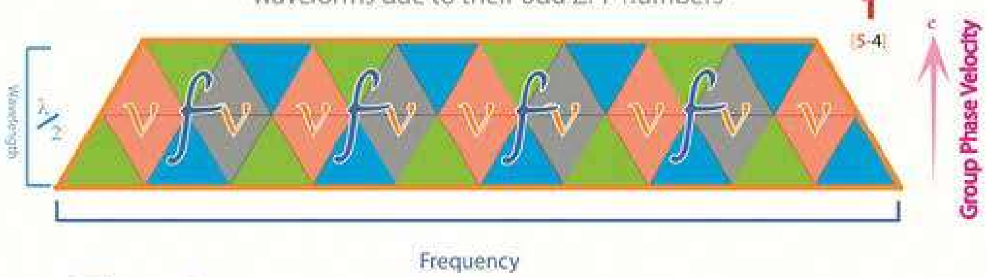
1
[5-4]
c
Group Phase Velocity
Wavelength
λ/2
ν f ν ν f ν ν f ν ν f ν ν
Frequency

$$mv^2 = E = h\nu^2$$

All Bosons are transverse ODD quanta EM masses

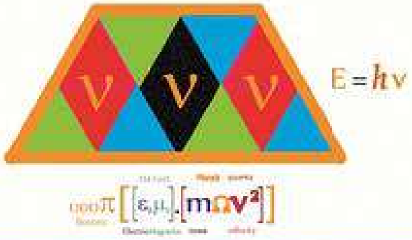
ν ν ν
E = hν
ODDπ [[ε₀μ₀].[mΩv²]]

The interchanging of Quantum number [ν]

$$2\nu = f$$

with Photon Frequency [f] is a cause of confusion

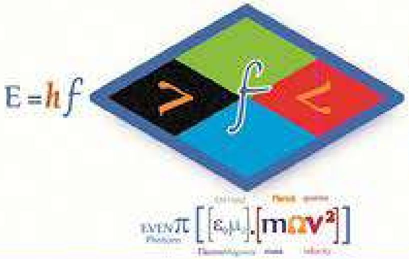
E = hf
ν f ν
EVENπ [[ε₀μ₀].[mΩv²]]

All Photons are longitudinal EVEN quanta EM masses

Transverse Bosons

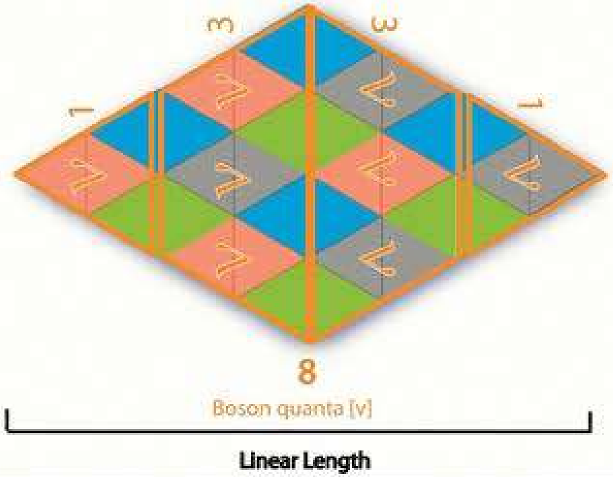
3
3
1
1
ν
8
Boson quanta [ν]
Linear Length

$$2h\nu = E = hf$$

EM Force Carriers
$E = hf$

Photons

Longitudinal Photons

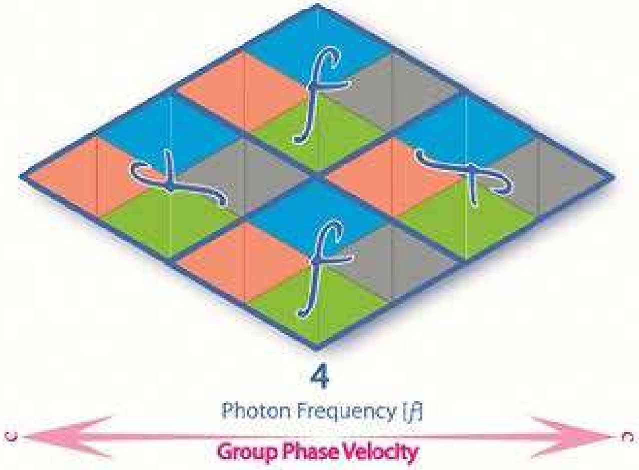
f
4
Photon Frequency [f]
c
c
Group Phase Velocity

Photonic Energies

$$2p^2 = KE = 2mv^2$$

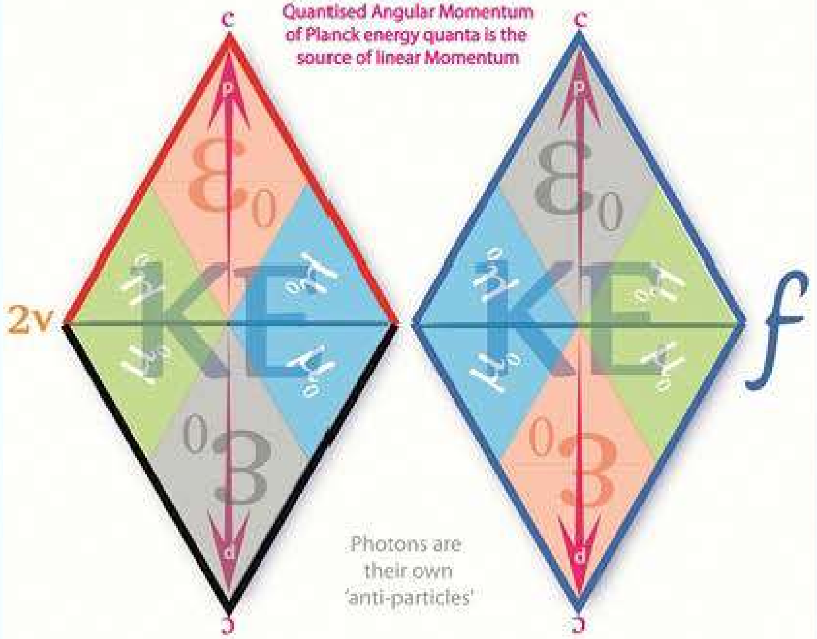

The equilateral geometries of Quantised Angular Momentum of Planck energy quanta is the source of linear Momentum

Photons are their own 'anti-particles'

$$2hv = E = hf$$

$$E_\gamma = 2mc^2$$

As Photons are 2D EM energy waveforms they should always be referred to as having EM mass-energy equivalence

γ

$$\underset{\text{Photons}}{2\pi}\Big[\big[\overset{\text{EM Field}}{\varepsilon_0\mu_0}\big]_{\text{ElectroMagnetic}}.\big[\overset{\text{Planck quanta}}{\underset{\text{mass velocity}}{m\Omega v^2}}\big]\Big]$$

$$2\pi\Big[\overset{\text{Planck quanta}}{\underset{\text{mass velocity}}{m\Omega v^2}}\Big]$$

Photons are kinetic energy [KE] wave packets:

$KE = (1/2) * m_\gamma * v^2$
$= 1/2 * (E/c^2) * c^2$
$= E/2$

KE

Frequency Number of repeating cycles of energy momenta per time unit

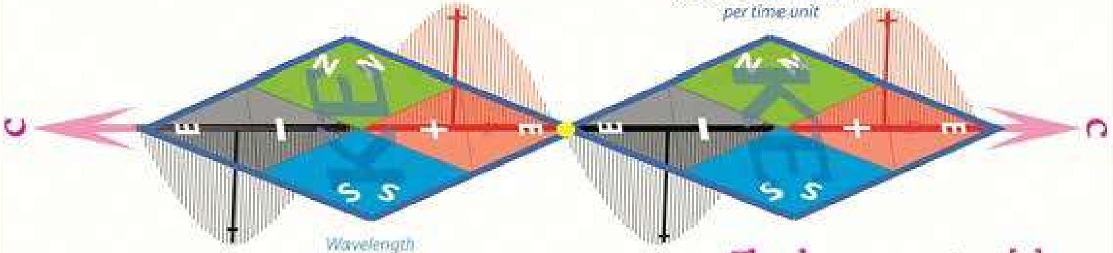

Wavelength decreases as Photon internal Frequency increases

They have momentum [p]:

$p = m_\gamma * v$
$= (E/c^2) * c$
$= E * c$

0
[v-v]

Photons are 2n neutral charge waveforms

Photons are divergent kinetic waveforms

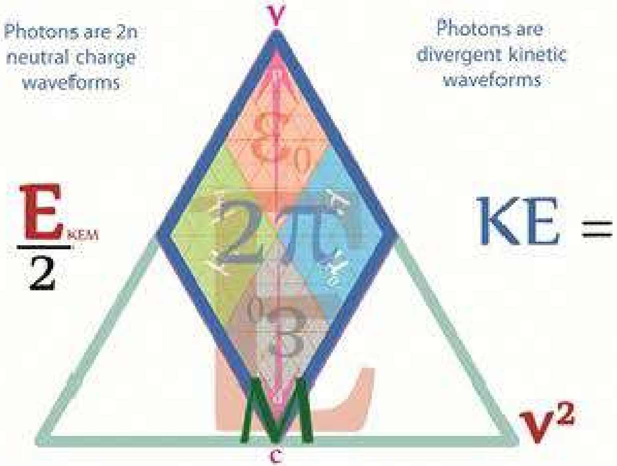

$$KE = \frac{E_{KEM}}{2}$$

$$KE = \frac{1}{2}Mv^2$$

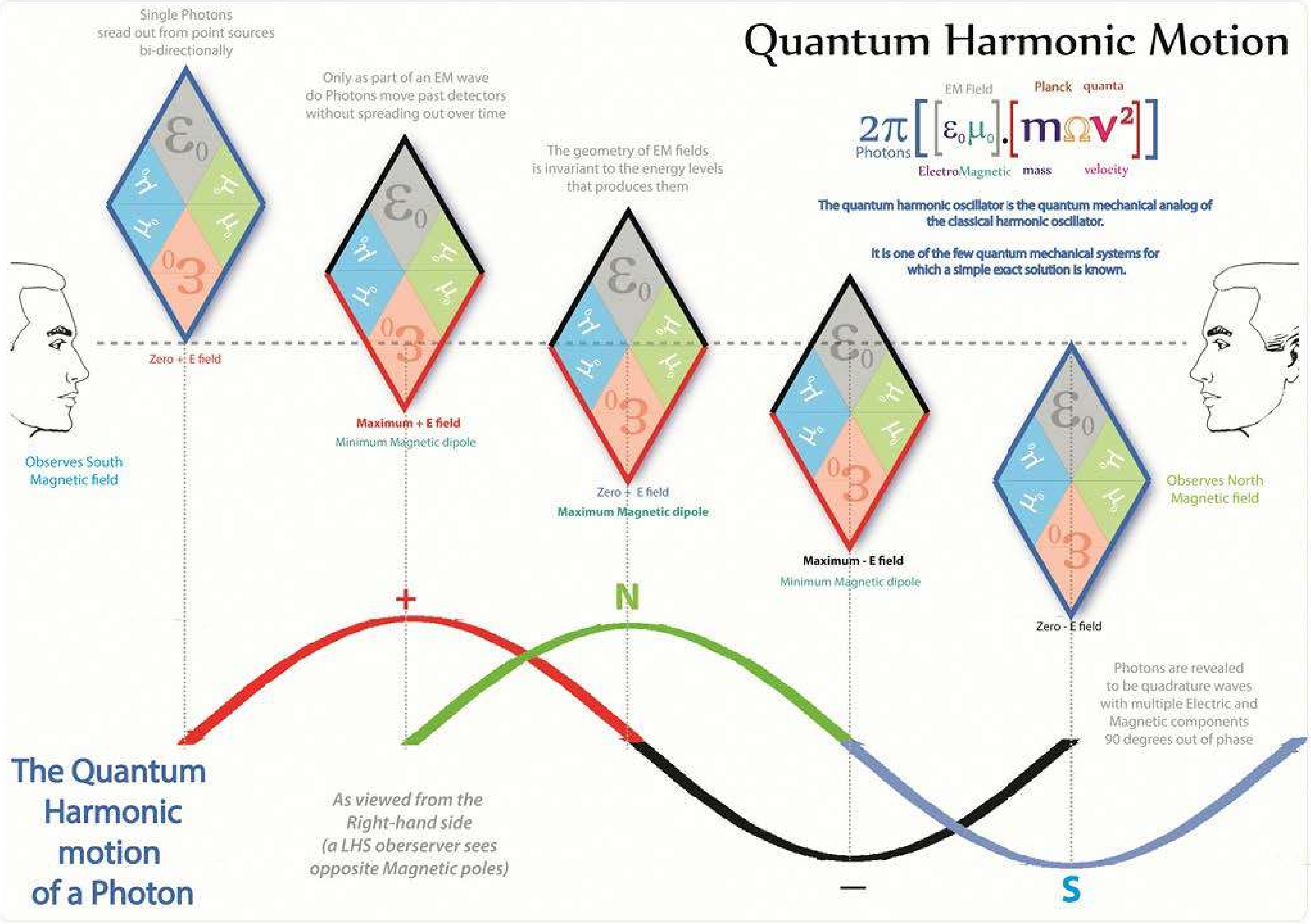

Quantum Harmonic Motion

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM Field
Planck
quanta
Photons
ElectroMagnetic
mass
velocity

The quantum harmonic oscillator is the quantum mechanical analog of the classical harmonic oscillator.

It is one of the few quantum mechanical systems for which a simple exact solution is known.

Single Photons sread out from point sources bi-directionally

Only as part of an EM wave do Photons move past detectors without spreading out over time

The geometry of EM fields is invariant to the energy levels that produces them

Photons are revealed to be quadrature waves with multiple Electric and Magnetic components 90 degrees out of phase

As viewed from the Right-hand side (a LHS oberserver sees opposite Magnetic poles)

The Quantum Harmonic motion of a Photon

ε_0
μ_0
3
Zero + E field
Maximum + E field
Minimum Magnetic dipole
Zero + E field
Maximum Magnetic dipole
Maximum - E field
Minimum Magnetic dipole
Zero - E field
Observes South Magnetic field
Observes North Magnetic field
+
N
−
S

Tetryonics 26.11 - Quantum Harmonic Motion

Copyright ABRAHAM [2008] - all rights reserved

Photonic EM fields

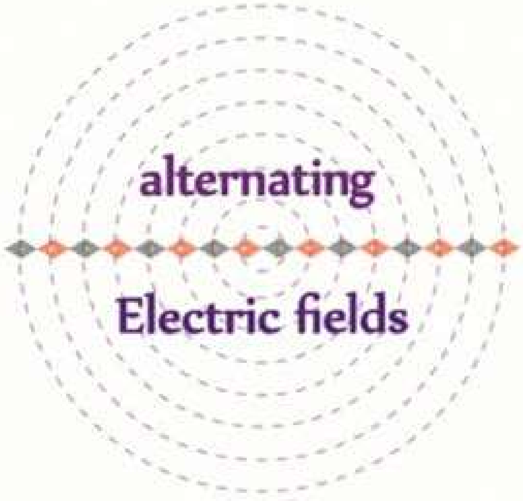
alternating
Electric fields

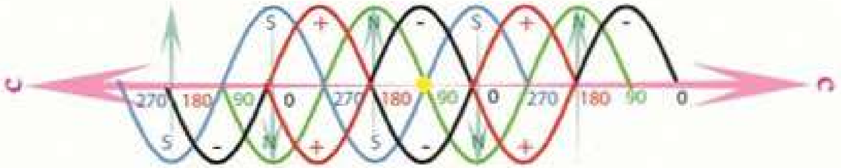
c
S
+
N
-
270
180
90
0
c

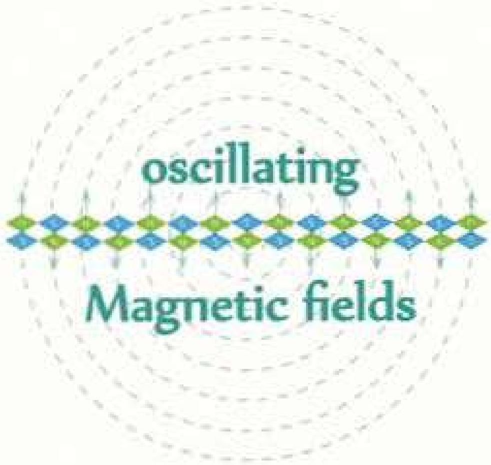
oscillating
Magnetic fields

Euler's formula is a natural geometric expression of electromagnetic waves

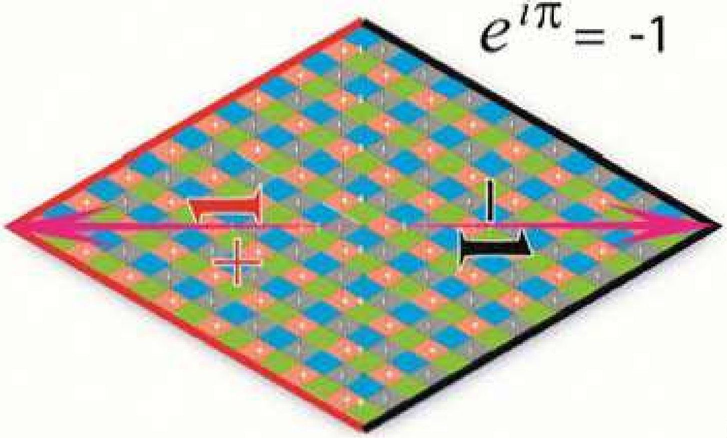

$e^{i\pi} = -1$

$e^{i\pi} + 1 = 0$

$e^{ix} = \cos x + i \sin x$

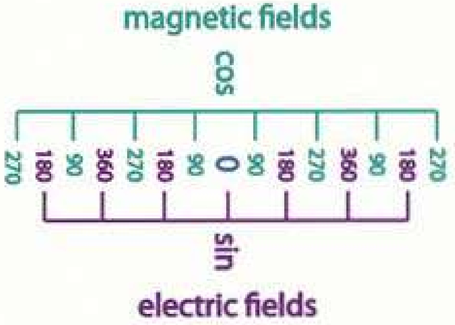
magnetic fields
cos
270
180
90
360
270
180
90
0
90
180
270
360
90
180
270
sin
electric fields

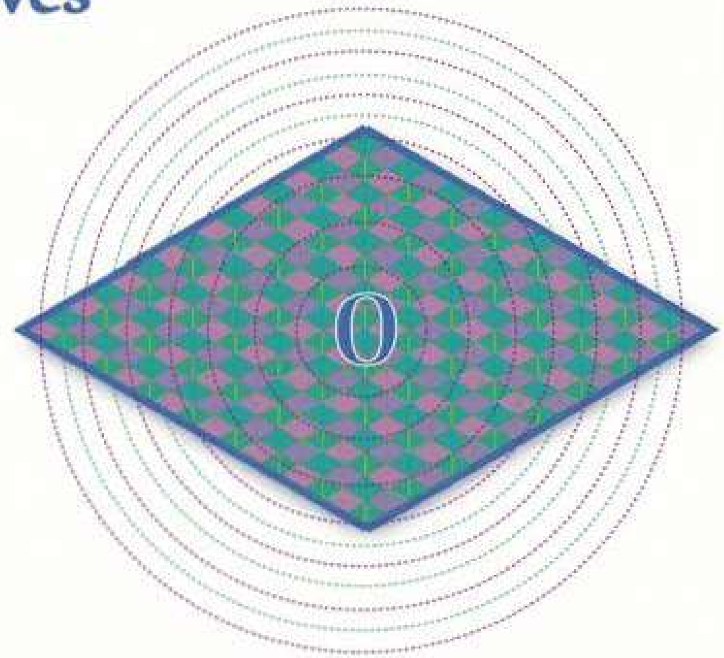

Every field geometry in a EM wave is 90 degrees out of phase with the E or M field adjacent to it

Tetryonics 26.12 - Photonic EM field phases

$$e^{i\pi} = -1$$

$$e^{i\pi} + 1 = 0$$

Euler's Formula

Euler's formula is often considered to be the basis of the complex number system. In deriving this formula, Euler established a relationship between the trigonometric functions, sine and cosine, and e raised to a power

$$e^{ix} = \cos(x) + i\sin(x)$$

a mathematical description of EM-Energy waveforms

adjacent EM fields are 90-180° out of phase

Leonhard Euler

(15 April 1707 – 18 September 1783)

Positive E-fields are out of phase with Negative fields

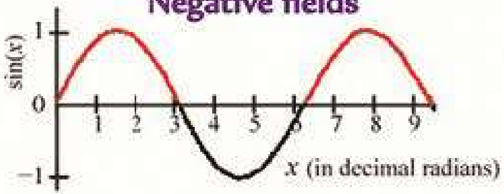

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

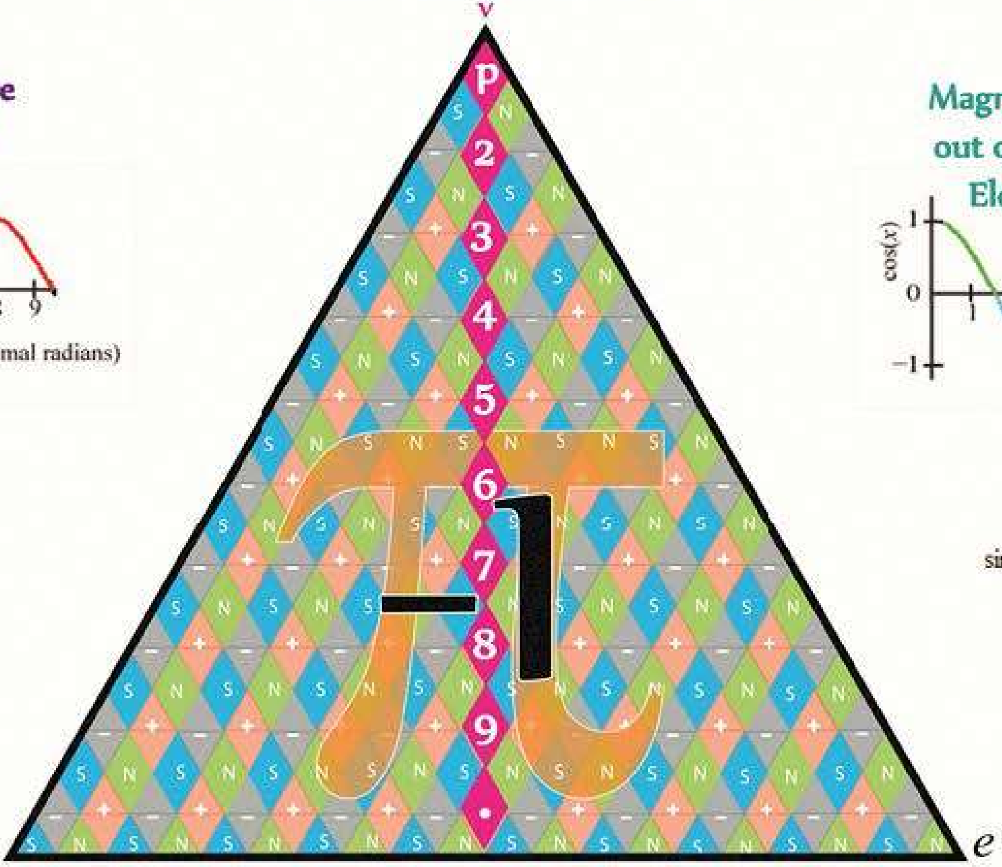

1 2 3 4 5 6 7 8 9 x 9 8 7 6 5 4 3 2 1

The above geometry is reflective of negative charge energy momenta Planck geometries

the momentum of the nett charged geometry is the Square root of Negative 1

Magnetic fields are out of phase with Electric fields

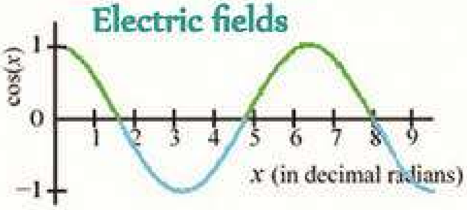

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots$$

$$\sin x + \cos x = 1 + x - \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

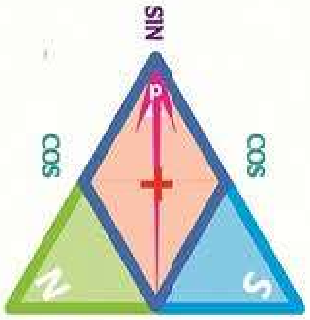

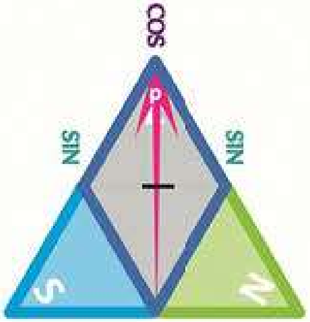

Inverse Square Law

Squared numbers are equilateral geometries

Any point source which spreads its influence equally in all directions without a limit to its range will obey the inverse square law

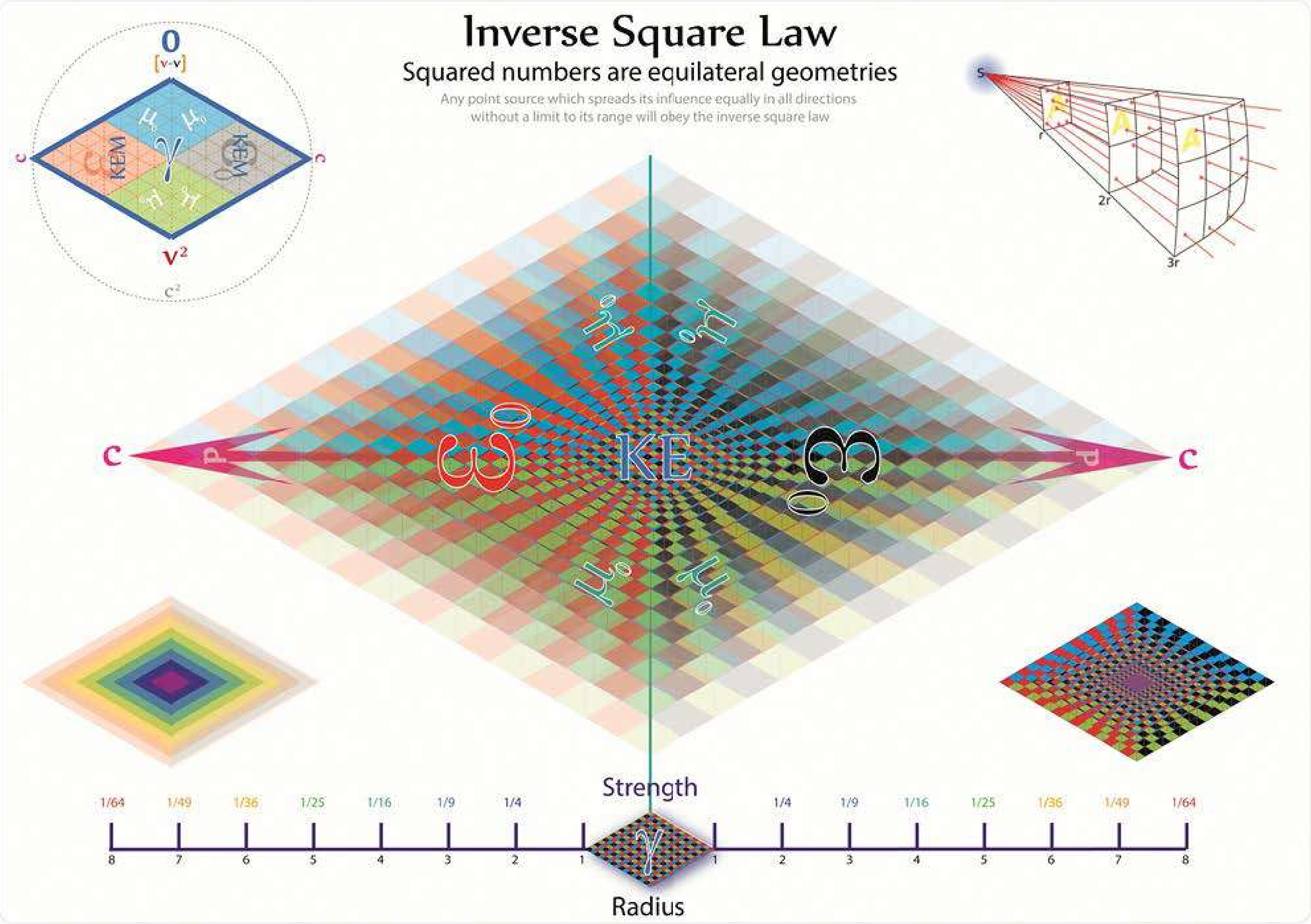
0
[v-v]
μ0
γ
KEM
ε0
c
c
v²
c²
S
r
2r
3r
A
c
d
KE
ε0
μ0
d
c
Strength
1/64
1/49
1/36
1/25
1/16
1/9
1/4
1/4
1/9
1/16
1/25
1/36
1/49
1/64
8
7
6
5
4
3
2
1
γ
1
2
3
4
5
6
7
8
Radius

mass-Energy momenta

EM Field Planck quanta

$$n\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

mass ElectroMagnetic mass velocity

Throughout history Physicists have sought relationships between EM mass-Energy and momenta in an attempt to discern the true Nature of these physical properties

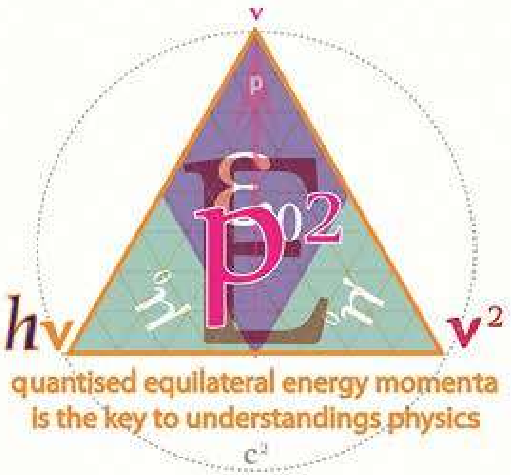

quantised equilateral energy momenta is the key to understandings physics

E

Energy

$$kg\,\frac{m^2}{s^2}$$

$E = mv$	$E = mv^2$	$v = \frac{1}{\sqrt{\varepsilon_0\mu_0}}$
Newton	*Leibnitz*	*Maxwell*

Velocity of Light

$$v = \left[\left[\frac{c^2}{\Omega}\right]\cdot\left[\frac{\Omega}{c}\right]\right]$$

f Scalar Frequency

λ Linear Wavelength

Planck	*Einstein*	*de Broglie*
$E = hv$	$E = hf$	h/p

Planck's Constant

$$kg\,\frac{m^2}{s}$$

Compton Frequency

$$\left[\frac{E}{h}\right]$$

de Broglie wavelength

$$\left[\frac{E}{p}\right]$$

Wavelength

p

Linear momentum

$$kg\,\frac{m}{s}$$

Ryberg wavenumber

$$\left[\frac{p}{E}\right]$$

Tetryonics 26.15 - Photonic mass-ENERGY momenta

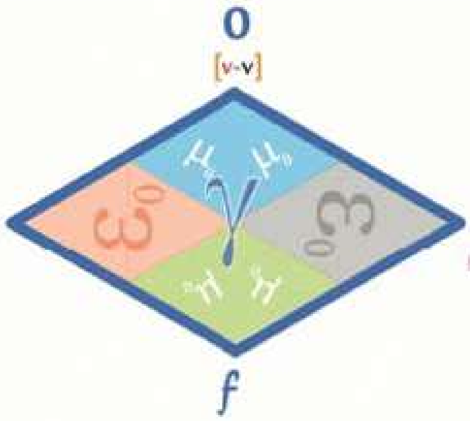

$$mv^2 = KE = hv^2$$

All Photons contain Energy momenta
and therefore have electromagnetic mass
but they are 'Matterless' geometries

Wavenumber

$$\frac{m}{s}\cdot\frac{s}{m^2}\left[\frac{c}{\Omega}\right]\frac{1}{m}$$

Light cannot have zero velocity
[stand still] but it is possible
to catch up to a photon
and observe its waveform

The Speed of Light is
the maximum speed of Electrical energy
and remains a Universal constant
(but it is not a Limit)

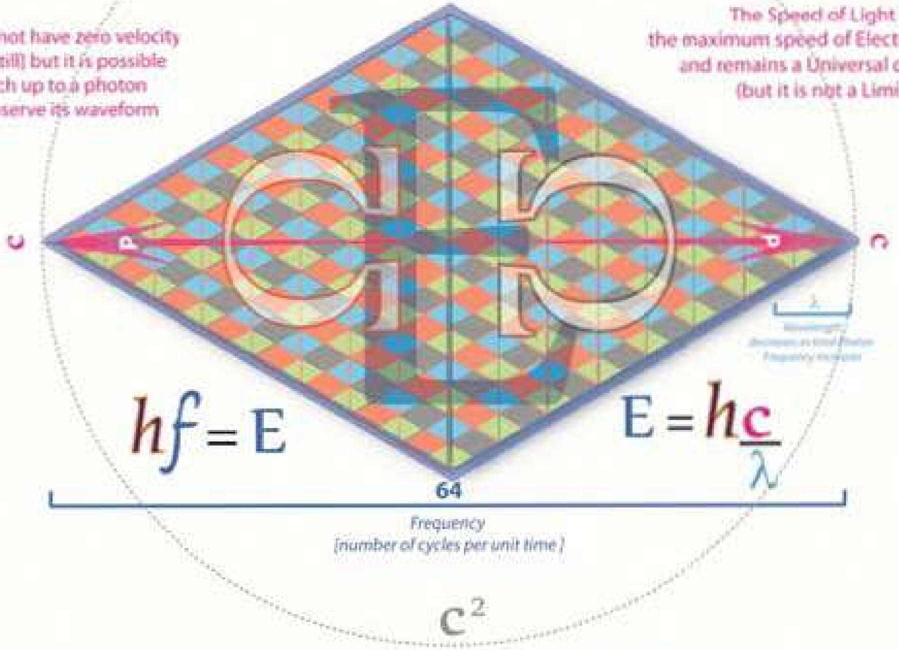

Celeritas

[Frequency x Wavelength]

EM waves are typically described
through any of their following
three physical properties:

frequency f,
wavelength λ, or
photon energy E

ElectroMagnetic Fields

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

$$1.112650056\text{ e-}17\ \frac{s^2}{m^2}$$

Frequency, momentum, and Wavelength
are all related through the geometries of c & c²

λ

$$m\left[\frac{\Omega}{c}\right]\frac{m^2}{s}\cdot\frac{s}{m}$$

Wavelength

Velocity of Light

$$c = \left[\overset{f}{\left[\frac{c^2}{\Omega}\right]}\cdot\overset{\lambda}{\left[\frac{\Omega}{c}\right]}\right]$$

$$v = \frac{1}{\sqrt{\varepsilon_0\mu_0}}$$

Frequency Wavelength

f

$$\frac{m^2}{s^2}\cdot\frac{s}{m^2}\left[\frac{c^2}{\Omega}\right]\frac{1}{s}$$

Frequency

Frequency is the number of neutral quanta sets repeating per unit time in a EM wave

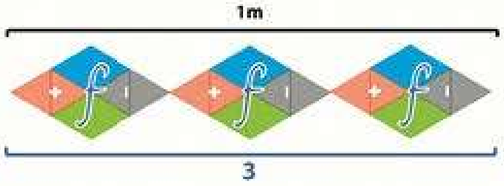

EM waves frequencies [photon number] change in direct proportion to the energy content of any electromagnetic wave

The specific wavelengths and frequencies of energy momenta in KEM fields form the basis for Spectral emission/absorption lines [care must be always taken to specify if quanta or photons are being measured]

$$2v = f$$

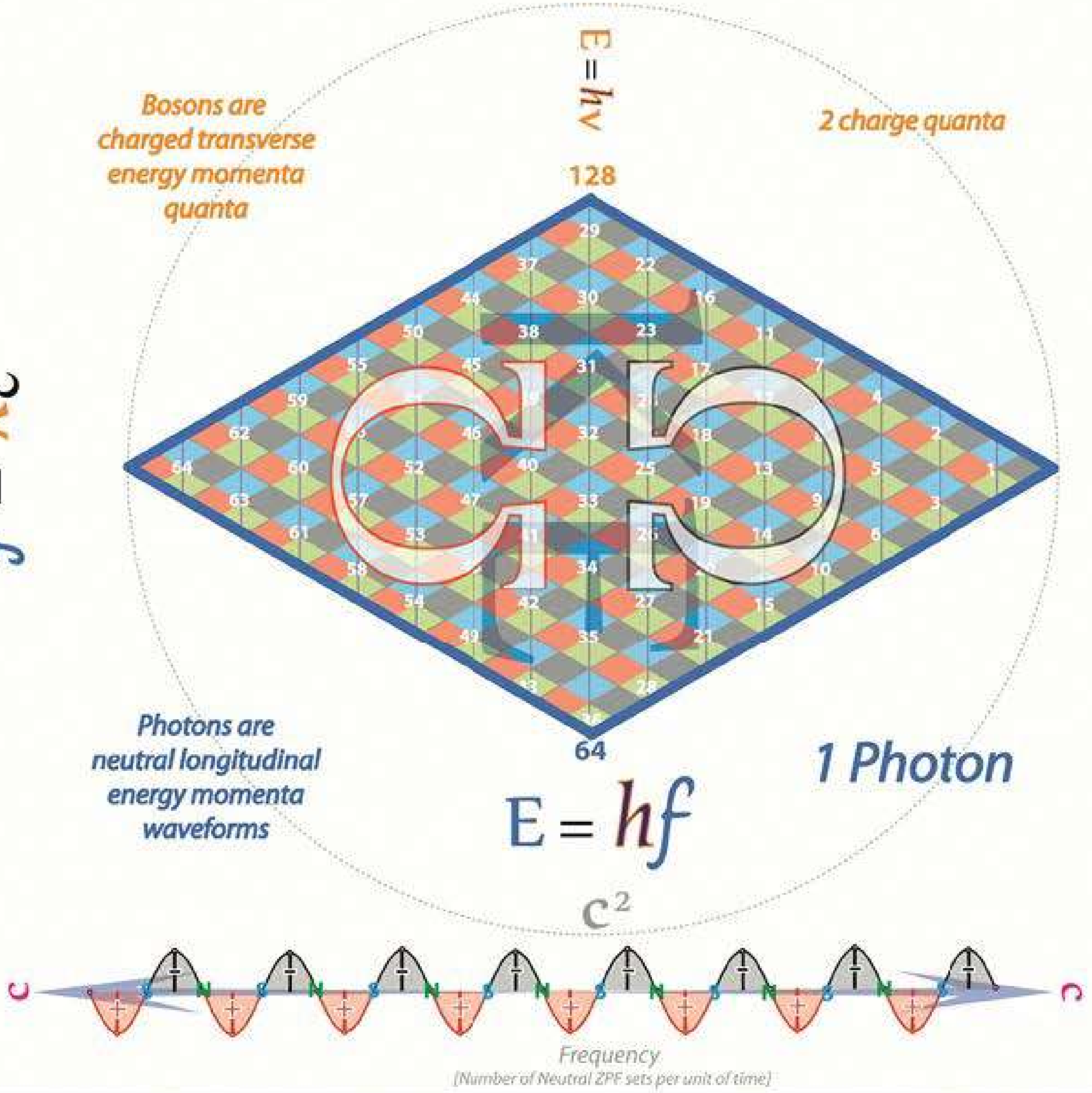

Bosons are charged transverse energy momenta quanta

2 charge quanta

Photons are neutral longitudinal energy momenta waveforms

$$E = hf$$

1 Photon

Irrespective of whether its measured as a transverse or longitudinal electromagnetic wave

Frequency is the inverse of Charge

$$f$$

$$\left[\frac{c^2}{\Omega}\right]$$

Frequency

is directly related to the nett quantised angular momentum of any space time geometry

Tetryonics 26.17 - Frequency

λ

Wavelength

Wavelength is a measure of the distance between repetitions of geometric features such as maximums, minimums, or zero-points of Electric and Magnetic fields

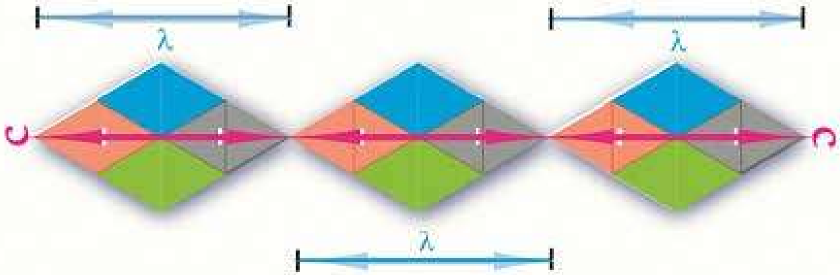

$\tilde{\nu}$

Wavelengths are the inverse of Wavenumbers

λ

For additional quantised energy momenta to be added to any spatial geometry their physical wavelengths must decrease

Lorentz velocity dependent length contractions must be correctly retermed electromagnetic WAVE-length contractions

λ

$$\left[\frac{\Omega}{c}\right]$$

Wavelength

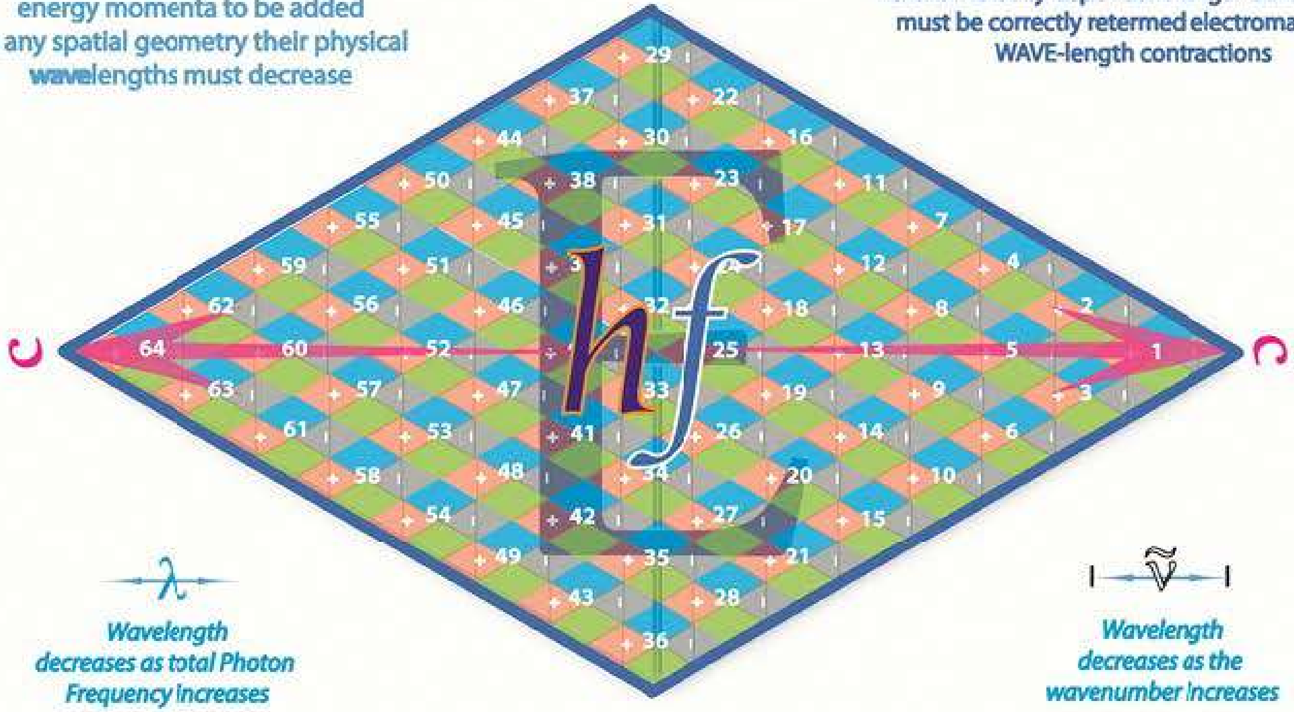

λ

Wavelength decreases as total Photon Frequency Increases

$\tilde{\nu}$

Wavelength decreases as the wavenumber Increases

In physics, the wavelength of a sinusoidal wave is the spatial period of the wave – the distance over which the wave's shape repeats. It is usually determined by considering the distance between consecutive corresponding points of the same phase, such as crests, troughs, or zero crossings, and is a characteristic of both traveling waves and standing waves, as well as other spatial wave patterns.

The concept can also be applied to periodic complex or non-sinusoidal waves

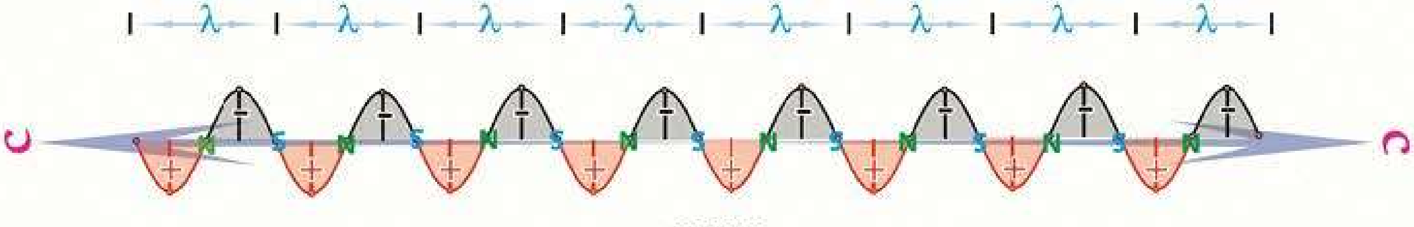

299,792,458
linear metres per second

Wavenumbers

$\tilde{\nu}$

Wavenumbers are the inverse of Wavelengths

λ

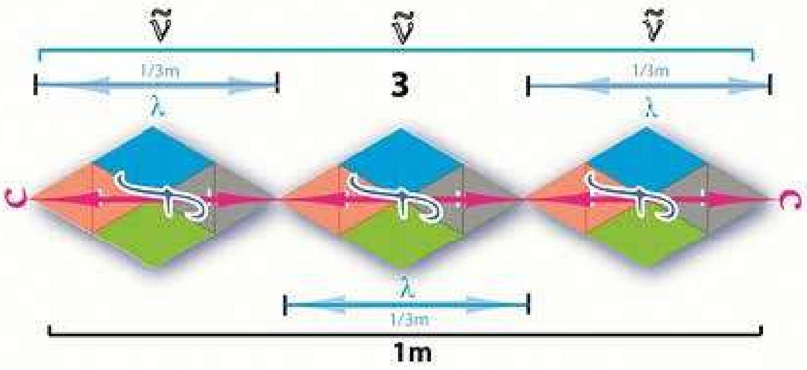

$\tilde{\nu}$
$\tilde{\nu}$
$\tilde{\nu}$
1/3m
3
1/3m
λ
λ
c
c
λ
1/3m
1m

Wavenumbers are the spatial equivalent of Frequency

$$\left[\frac{c}{\Omega}\right]$$

λ

$$\left[\frac{\Omega}{c}\right]$$

Wavelength and Frequency are related to the group velocity of quanta in an EM wave

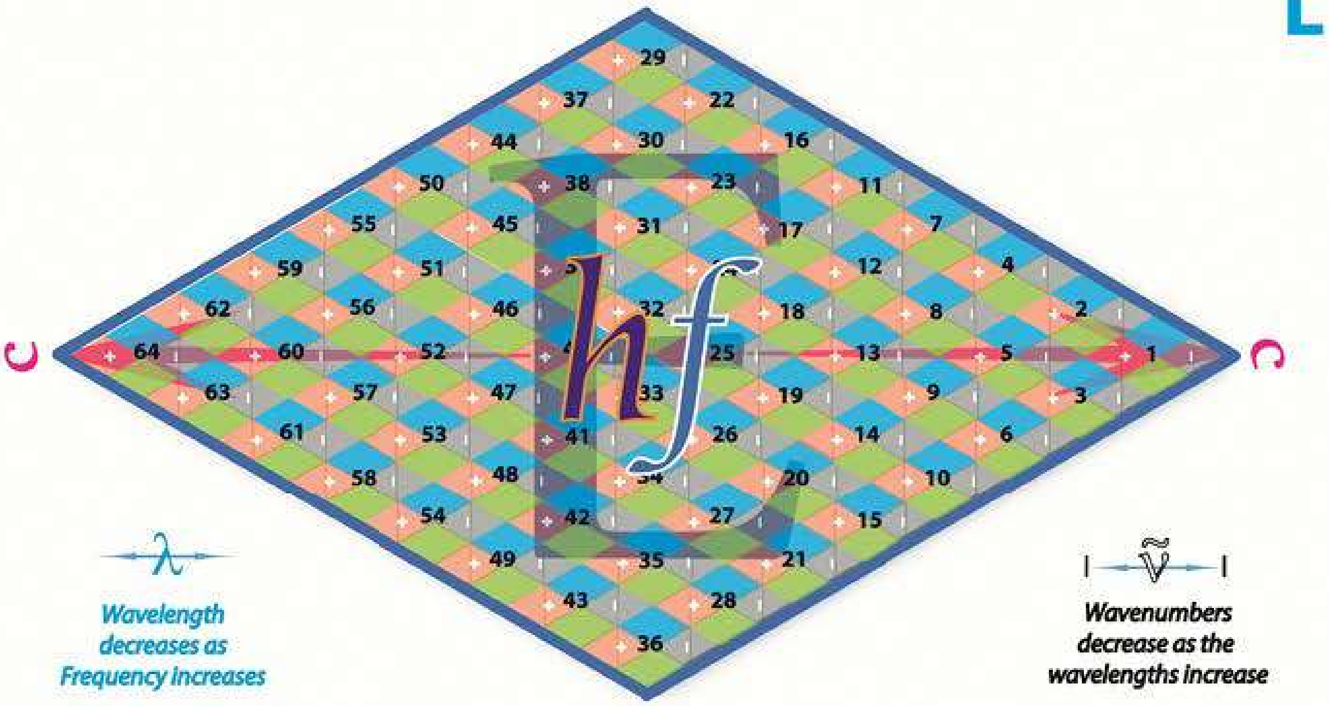

c
c
hf
29
37
22
44
30
16
50
38
23
11
55
45
31
17
7
59
51
12
4
62
56
46
32
18
8
2
64
60
52
25
13
5
1
63
57
47
33
19
9
3
61
53
41
26
14
6
58
48
20
10
54
42
27
15
49
35
21
43
28
36

$\tilde{\nu} = 1/\lambda$

Wavenumbers are inversely proportional to Wavelengths

λ

Wavelength decreases as Frequency increases

$\tilde{\nu}$

Wavenumbers decrease as the wavelengths increase

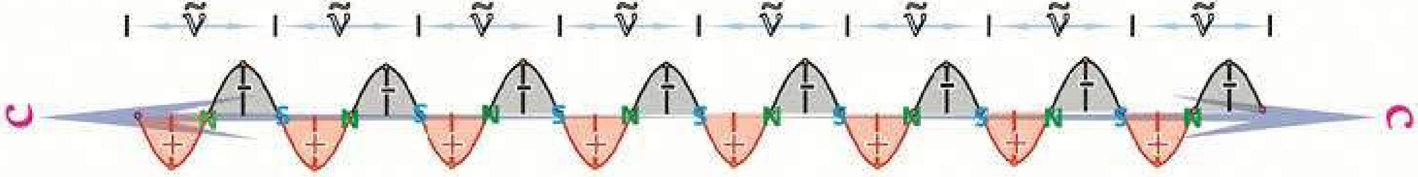

$\tilde{\nu}$
c
c
N
S
+

299,792,458
linear metres per second

Copyright ABRAHAM [2008] - all rights reserved

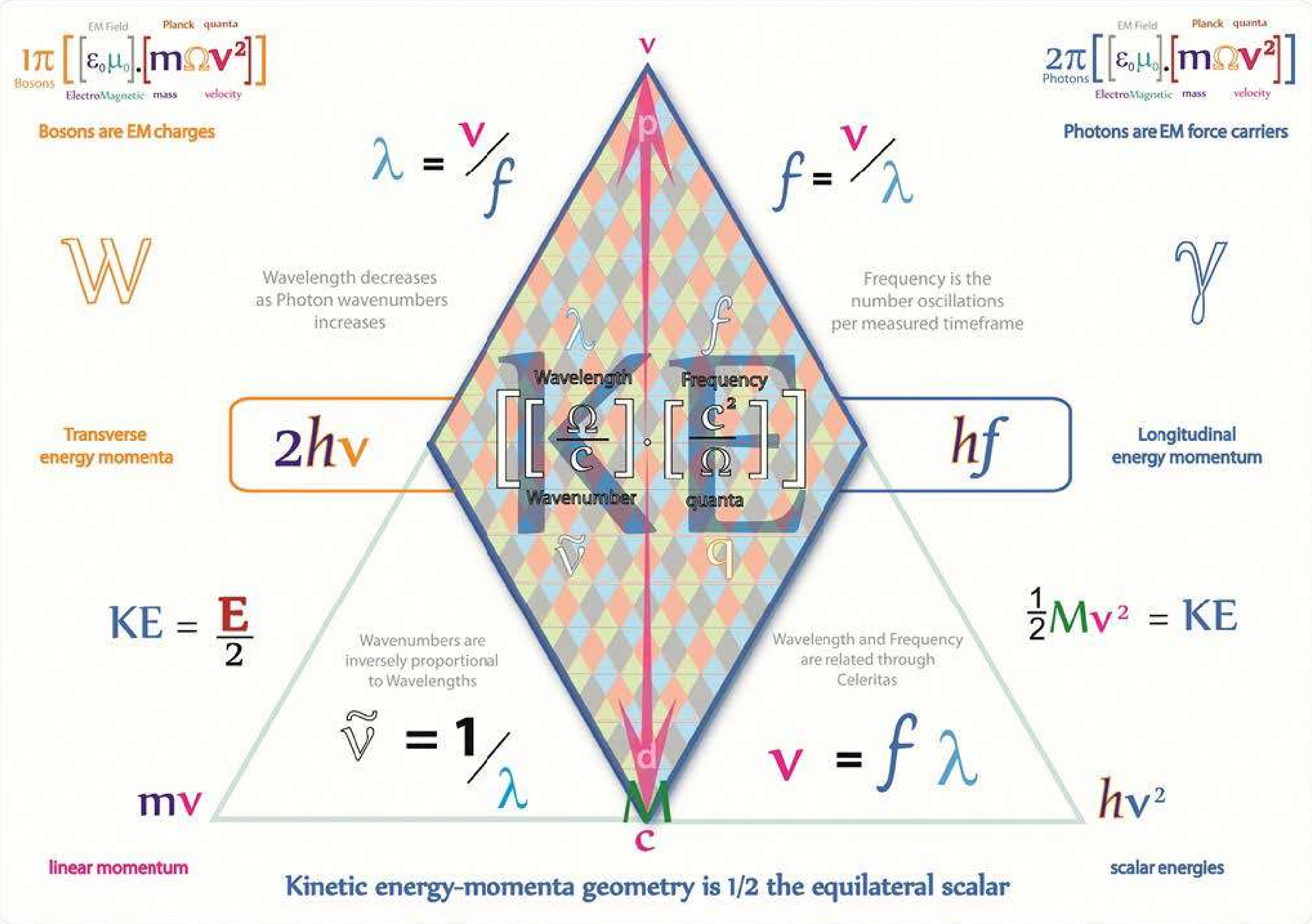
EM Field
Planck
quanta
Bosons
ElectroMagnetic
mass
velocity
EM Field
Planck
quanta
Photons
ElectroMagnetic
mass
velocity
v
p
λ
f
$\tilde{v}$
q
d
M
c

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons are EM charges

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons are EM force carriers

$\lambda = {}^{v}/_{f}$

$f = {}^{v}/_{\lambda}$

W

γ

Wavelength decreases
as Photon wavenumbers
increases

Frequency is the
number oscillations
per measured timeframe

Transverse
energy momenta

$2hv$

hf

Longitudinal
energy momentum

KE

Wavelength

Frequency

$$\left[\left[\frac{\Omega}{c}\right].\left[\frac{c^2}{\Omega}\right]\right]$$

Wavenumber

quanta

$KE = \frac{E}{2}$

$\frac{1}{2}Mv^2 = KE$

Wavenumbers are
inversely proportional
to Wavelengths

$\tilde{v} = 1/_{\lambda}$

Wavelength and Frequency
are related through
Celeritas

$v = f\lambda$

mv

hv^2

linear momentum

scalar energies

Kinetic energy-momenta geometry is 1/2 the equilateral scalar

Tetryonics 26.20 - Kinetic energy-momenta

Copyright ABRAHAM [2008] - all rights reserved

Faraday Rotation

The Faraday effect or Faraday rotation is a Magneto-optical phenomenon, that is, an interaction between light and a magnetic field in a medium. The Faraday effect causes a rotation of the plane of polarization which is linearly proportional to the component of the magnetic field in the direction of propagation.

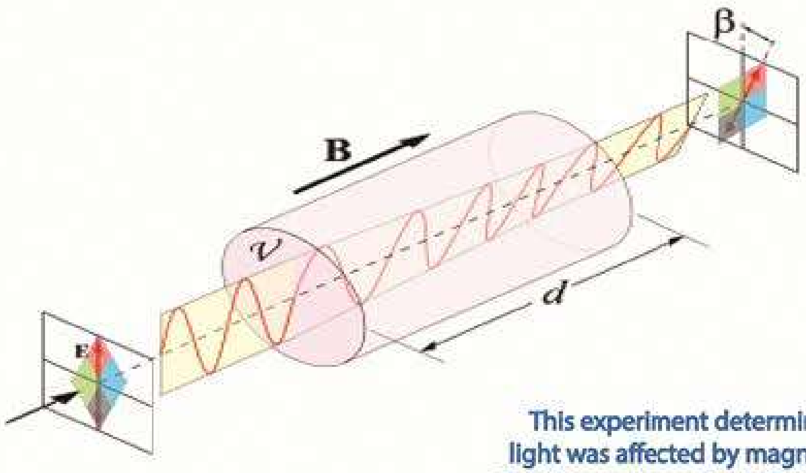

Discovered by Michael Faraday in 1845,
the Faraday effect was the first experimental
evidence that light and electromagnetism are related

Faraday summarized the entire effect as follows:

"Magnetic lines, then, in passing through silicated borate of lead, and a great number of other substances, cause these bodies to act upon a polarized ray of light when the lines are parallel to the ray, or in proportion as they are parallel to it: if they are perpendicular to the ray, they have no action upon it.

They give the diamagnetic the power of rotating the ray; and the law of this action on light is, that if a magnetic line of force be going from a north pole, or coming from a south pole, along the path of a polarized ray coming to the observer, it will rotate that ray to the right-hand; or, that if such a line of force be coming from a north pole, or going from a south pole, it will rotate such a ray to the left hand."

Michael Faraday

(22 September 1791 – 25 August 1867)

Evan's Photomagneton

A photon has a magnetic dipole.
It is an elementary magnet.
Evans discovery of the photon's
longitudinal magnetic field in
1992 is as significant, as Einstein's
discovery of relativity.

It helps in giving a physical
interpretation of wave mechanics,
two-slit interference
and the Faraday effect,

all accounted for in Tetryonic Theory.

This experiment determined that light was affected by magnetic force. This "magneto-optical effect" was later termed the Faraday effect.

Faraday experimented with other substances that yielded similar results.

The resultant effect he termed "diamagnetism" concluding that magnetism was an inherent property of all EM mass-ENERGY-Matter

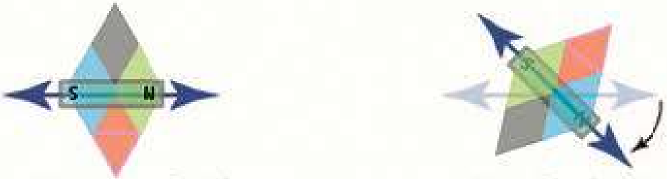

Before the
Magnetic field

After the
Magnetic field

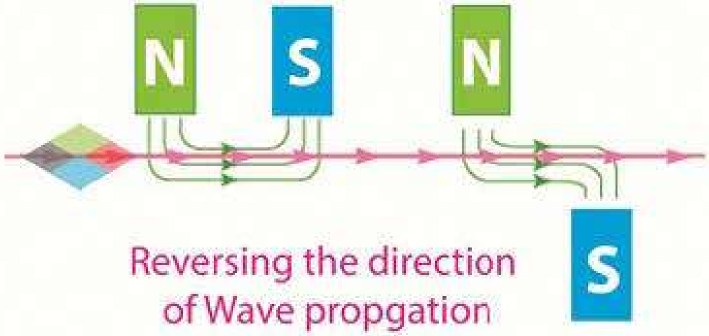

Reversing the direction
of Wave propgation
reverses the rotation
effected by the external
magnetic field

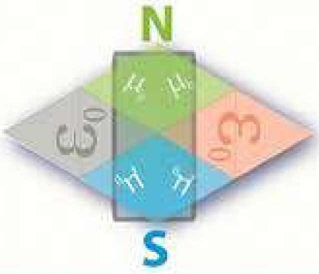

Every Photon and EM wave has polarised magnetic apexes as a result of the Planck quanta constituing them (creating Magnetic moments)

These Photo-Magnetic Moments allow for the interaction of Photons with external magnetic fields resulting in Faraday rotation of Photons and EM fields

Tetryonics 26.21 - Faraday rotation of Photons

Copyright ABRAHAM [2008] - all rights reserved

Isaac Newton allowed sunlight from a small, circular hole to fall on a prism, producing a rainbow of color. Although the production of a rainbow by a clear crystal was known to the ancients, it was Newton who showed that the colors did not originate in the crystal, but rather were components of sunlight.

Spectral colours

William Herschel:
The Discovery of Infrared Light

White Light is comprised of many coloured corpuscles

Refracting Prism

White Light

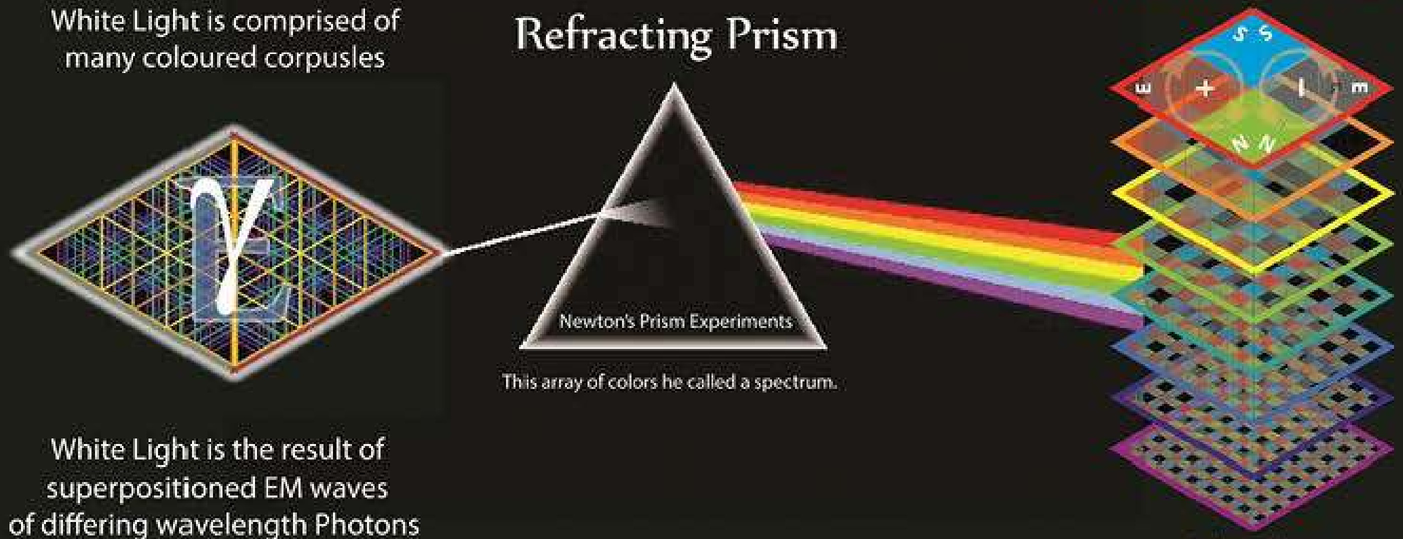

Red
Orange
Yellow
Green
Aqua
Blue
Indigo
Violet

This array of colors he called a spectrum.

White Light is the result of superpositioned EM waves of differing wavelength Photons

Johann Ritter:
The Discovery of Ultraviolet Light

Thomas Young discovered interference patterns caused by light passing through a narrow slit. Up until this point, the wave theory of light (Huygens) had had nothing convincing to offer in opposition to the particle theory of light (Newton). Now it did, because you cannot explain interference using particles.

The term "light" is often extended to wavelength ranges that the eye cannot detect - to infrared radiation, which has a frequency less than that of visible light, and to ultraviolet radiation, which have a frequency greater than that of visible light.

Light has:

Velocity
Angular momenta
Frequency
Wavelength
Relatvistic Energy
Linear momentum
Kinetic Energies
Magnetic Moment

and can

Refract
Reflect
Defract
and
Disperse

γ

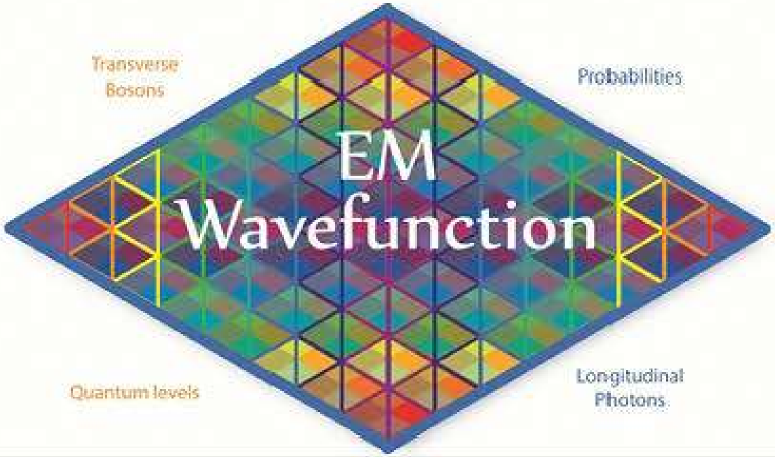

ψ

All EM waves possess Wavefunctions describing their probabilistic energy properties

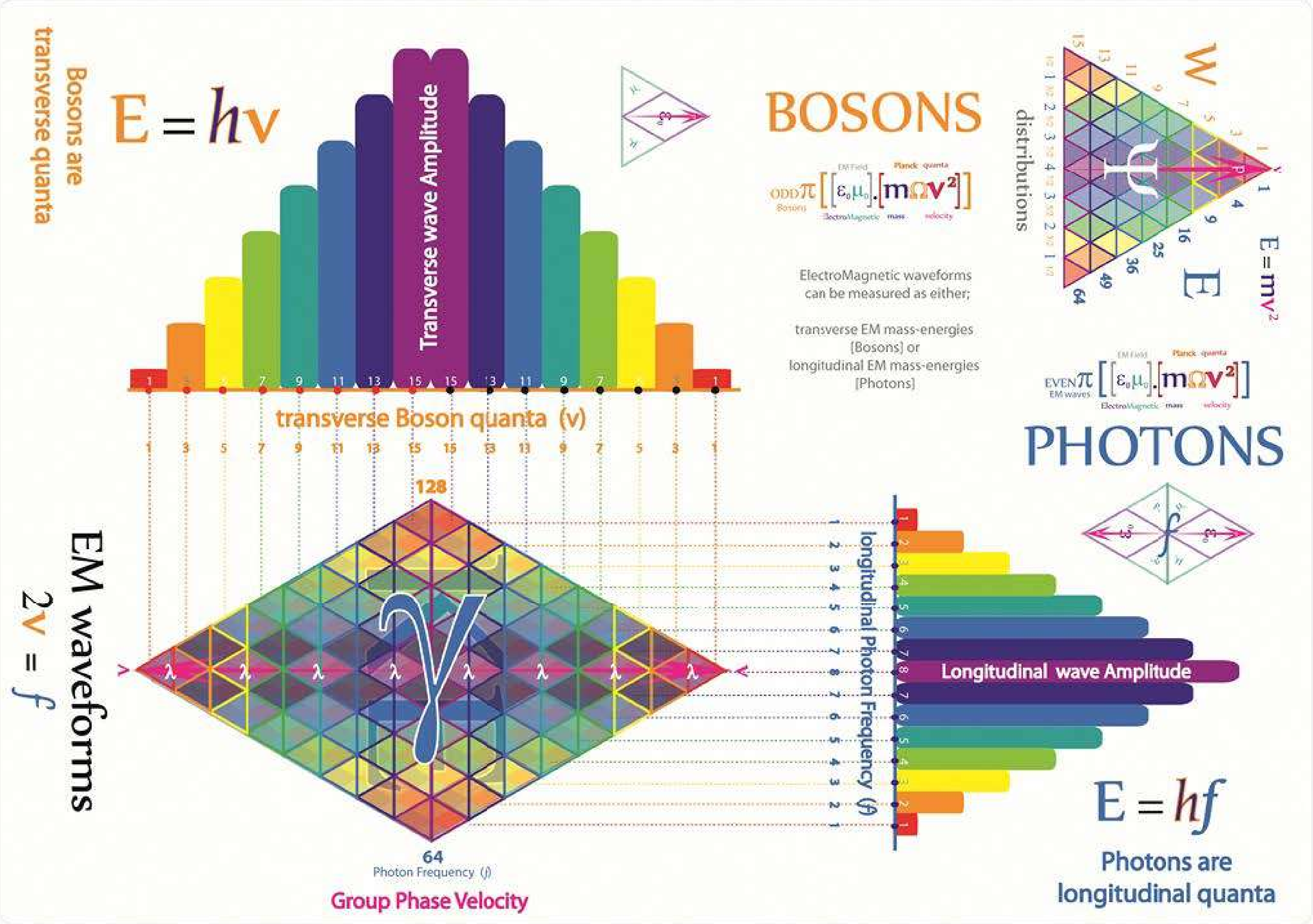
Bosons are transverse quanta
E = hv
Transverse wave Amplitude
transverse Boson quanta (v)
1 3 5 7 9 11 13 15 15 13 13 9 7 5 3 1
BOSONS
ODD π [[ε₀μ₀].[mΩv²]]
Bosons
EM Field
Planck quanta
ElectroMagnetic mass velocity
distributions
W
E
E = mv²
EVEN π [[ε₀μ₀].[mΩv²]]
EM waves
PHOTONS
128
EM waveforms
2v = f
γ
λ
64
Photon Frequency (f)
Group Phase Velocity
longitudinal Photon Frequency (f)
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
Longitudinal wave Amplitude
E = hf
Photons are longitudinal quanta

ElectroMagnetic waveforms can be measured as either;

transverse EM mass-energies [Bosons] or longitudinal EM mass-energies [Photons]

Copyright ABRAHAM [2008] - all rights reserved

Wavefunction probabilites

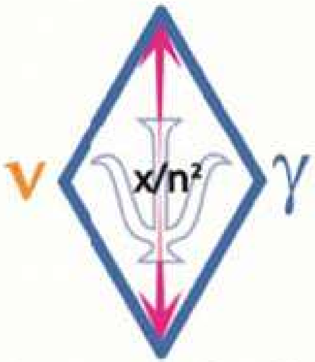

Wave Amplitude

Amplitude = bi-directional Momentum

Probability = [Amplitude]²

In his 1926 paper, Max Born suggested that the wave function of Schrödinger's wave equation represents the probability density of finding a particle

Probability of finding a Photon in a electromagnetic wave is the Square of its Amplitude

The de Broglie-Schrodinger wave fields should not to be interpreted as a mathematical description of how an event actually takes place in time and space, though, of course, they have reference to such an event.

Rather they are a mathematical description of what we can actually know about the system. They serve only to make statistical statements and predictions of the results of all measurements which we can carry out upon the system.

(Albert Einstein, on Quantum Physics, 1940)

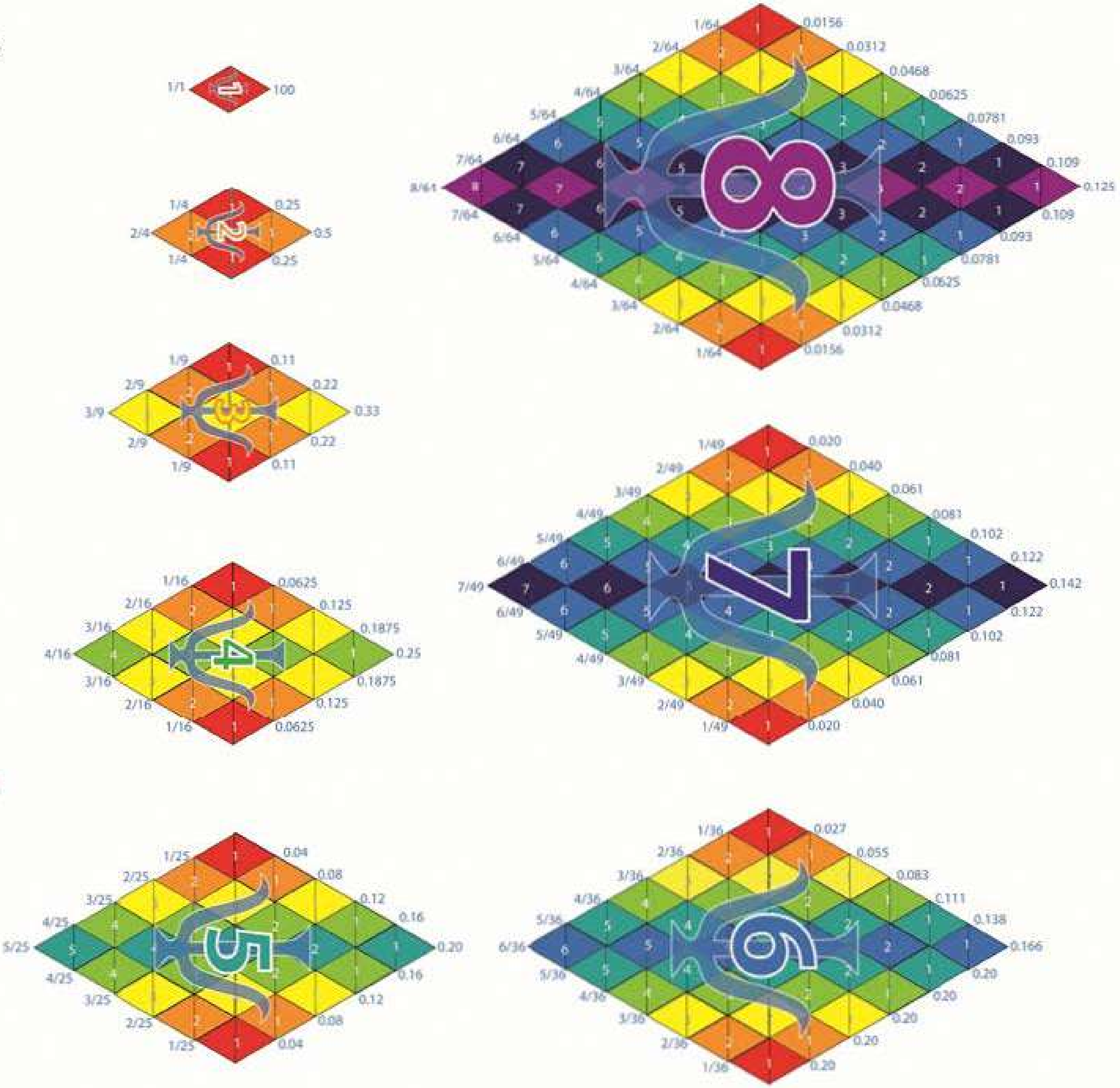
1/1
100
1/4
0.25
2/4
0.5
1/4
0.25
1/9
0.11
2/9
0.22
3/9
0.33
2/9
0.22
1/9
0.11
1/16
0.0625
2/16
0.125
3/16
0.1875
4/16
0.25
3/16
0.1875
2/16
0.125
1/16
0.0625
1/25
0.04
2/25
0.08
3/25
0.12
4/25
0.16
5/25
0.20
4/25
0.16
3/25
0.12
2/25
0.08
1/25
0.04
1/64
0.0156
2/64
0.0312
3/64
0.0468
4/64
0.0625
5/64
0.0781
6/64
0.093
7/64
0.109
8/64
0.125
7/64
0.109
6/64
0.093
5/64
0.0781
4/64
0.0625
3/64
0.0468
2/64
0.0312
1/64
0.0156
1/49
0.020
2/49
0.040
3/49
0.061
4/49
0.081
5/49
0.102
6/49
0.122
7/49
0.142
6/49
0.122
5/49
0.102
4/49
0.081
3/49
0.061
2/49
0.040
1/49
0.020
1/36
0.027
2/36
0.055
3/36
0.083
4/36
0.111
5/36
0.138
6/36
0.166
5/36
0.20
4/36
0.20
3/36
0.20
2/36
0.20
1/36
0.20

Tetryonics 27.03 - EM wave wavefunction

EM radiation patterns

Quantum wavefunctions mathematically model the statistical probability distributions of electromagnetic mass-energy within EM waves

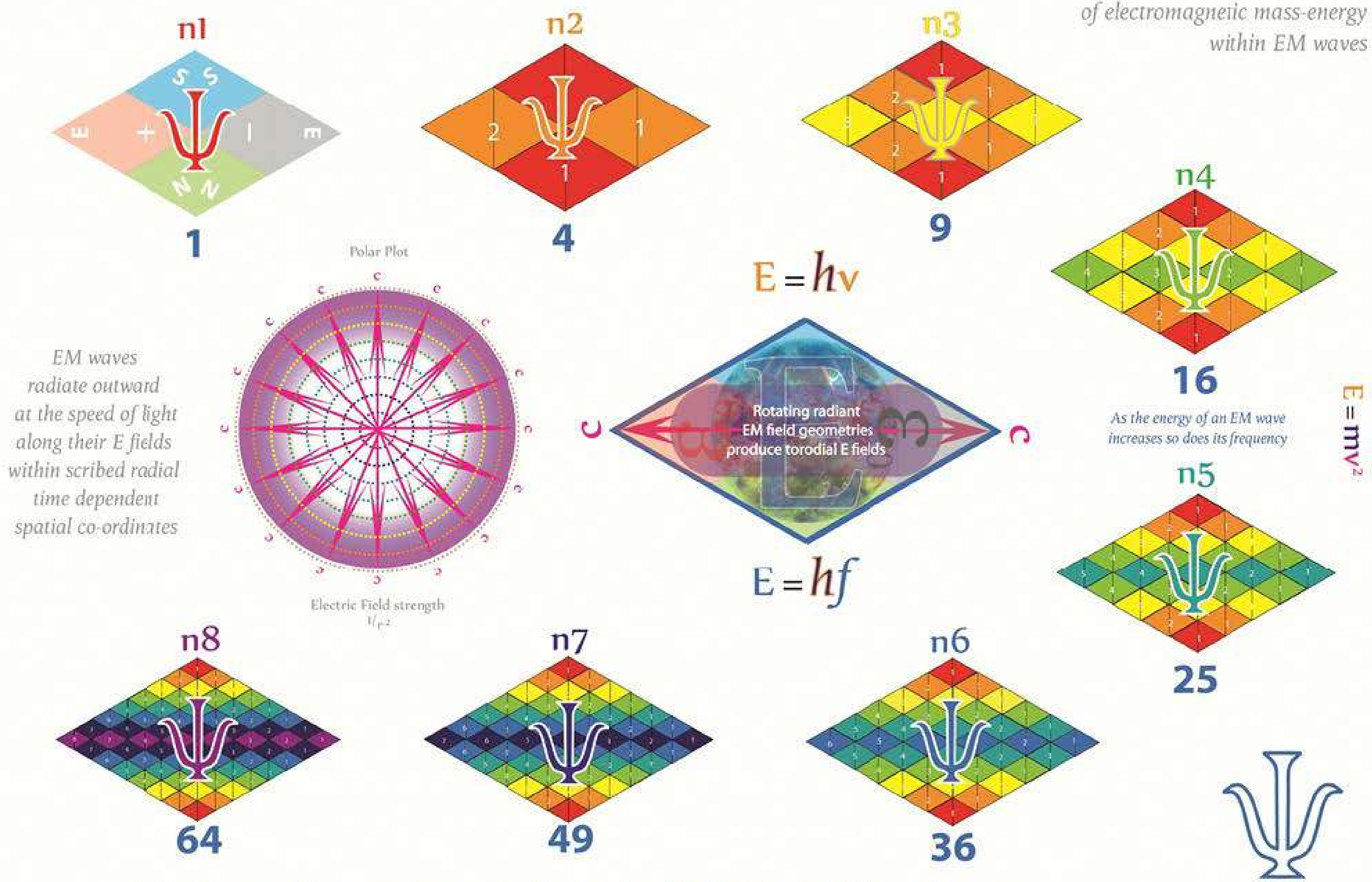
n1
S S
E
+
-
E
N N
1
n2
2
1
1
4
n3
9
n4
16
n5
25
n6
36
n7
49
n8
64
Polar Plot
c
Electric Field strength
1/r²
c
c

EM waves radiate outward at the speed of light along their E fields within scribed radial time dependent spatial co-ordinates

$E = h\nu$

Rotating radiant EM field geometries produce torodial E fields

$E = hf$

As the energy of an EM wave increases so does its frequency

$E = mv^2$

As the Energy and frequency increase, the Wavelengths of the Photons in a EM wave decreases

Copyright ABRAHAM [2008] - all rights reserved

EM wave geometry

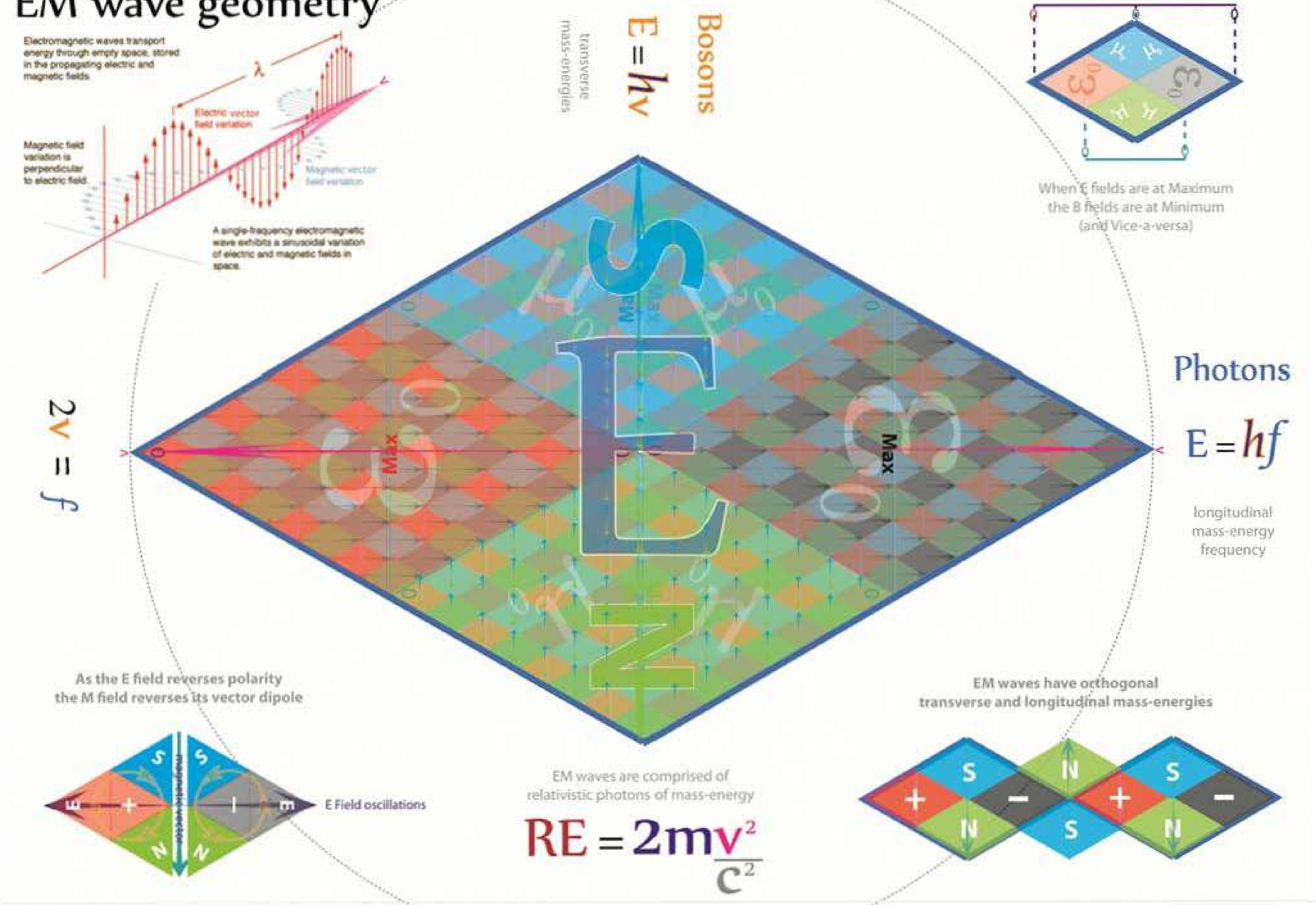

Electromagnetic waves transport energy through empty space, stored in the propagating electric and magnetic fields.

Magnetic field variation is perpendicular to electric field.

A single-frequency electromagnetic wave exhibits a sinusoidal variation of electric and magnetic fields in space.

λ
Electric vector field variation
Magnetic vector field variation
Bosons
$E = \hbar v$
transverse mass-energies
Photons
$E = hf$
longitudinal mass-energy frequency
$2v = f$
Max
Max
Max
S
E
N
+
–
E Field oscillations
magnetic vector

When E fields are at Maximum the B fields are at Minimum (and Vice-a-versa)

As the E field reverses polarity the M field reverses its vector dipole

EM waves are comprised of relativistic photons of mass-energy

$$RE = 2m\frac{v^2}{c^2}$$

EM waves have orthogonal transverse and longitudinal mass-energies

Tetryonics 27.05 - EM wave geometry

Copyright ABRAHAM [2008] - all rights reserved

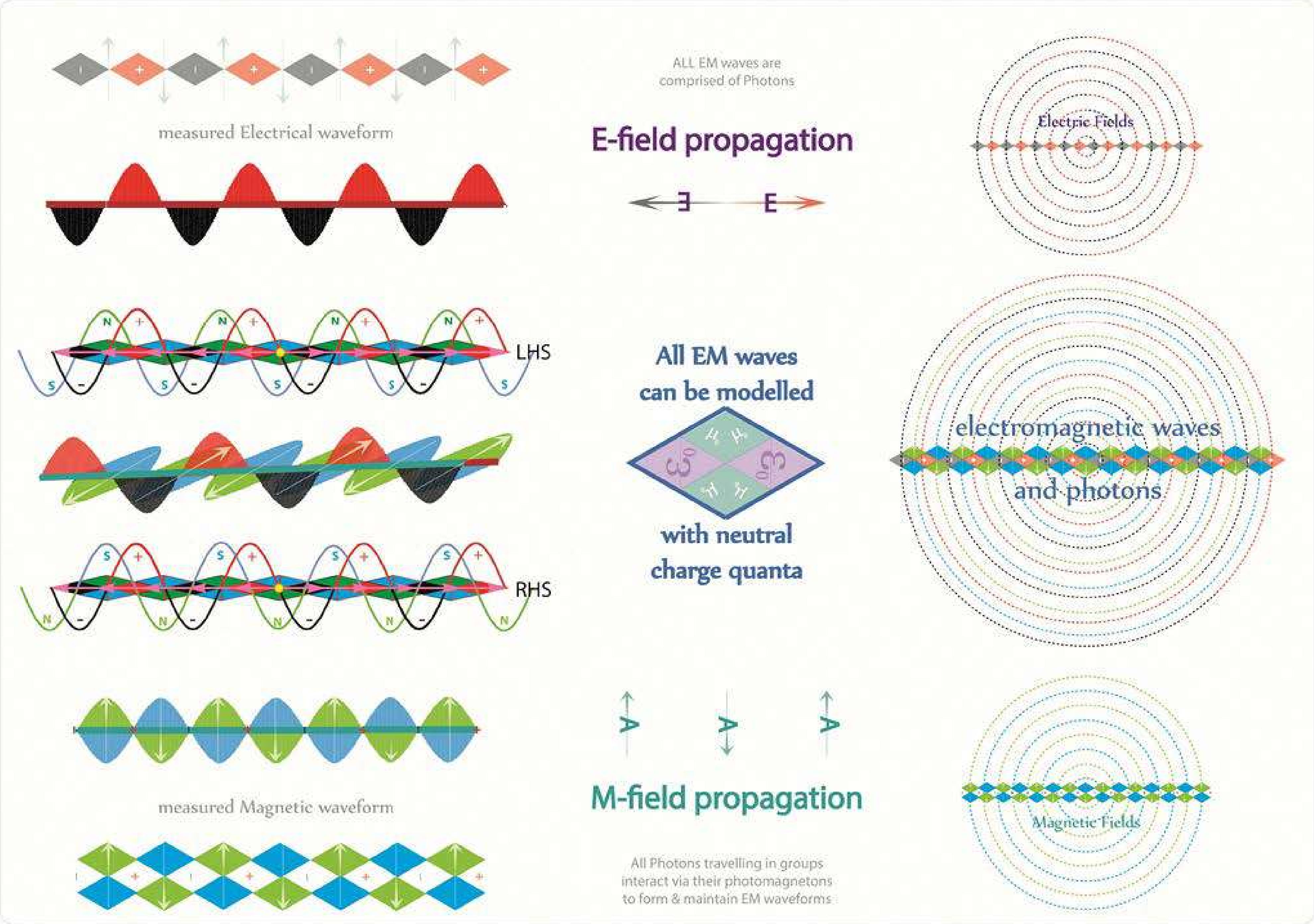
measured Electrical waveform
Electric Fields
N
S
+
-
LHS
RHS
electromagnetic waves
and photons
measured Magnetic waveform
Magnetic Fields

ALL EM waves are comprised of Photons

E-field propagation

All EM waves can be modelled with neutral charge quanta

M-field propagation

All Photons travelling in groups interact via their photomagnetons to form & maintain EM waveforms

Tetryonics 27.06 - E&M wave geometries

Copyright ABRAHAM [2008] - all rights reserved

Energy momenta in EM waveforms

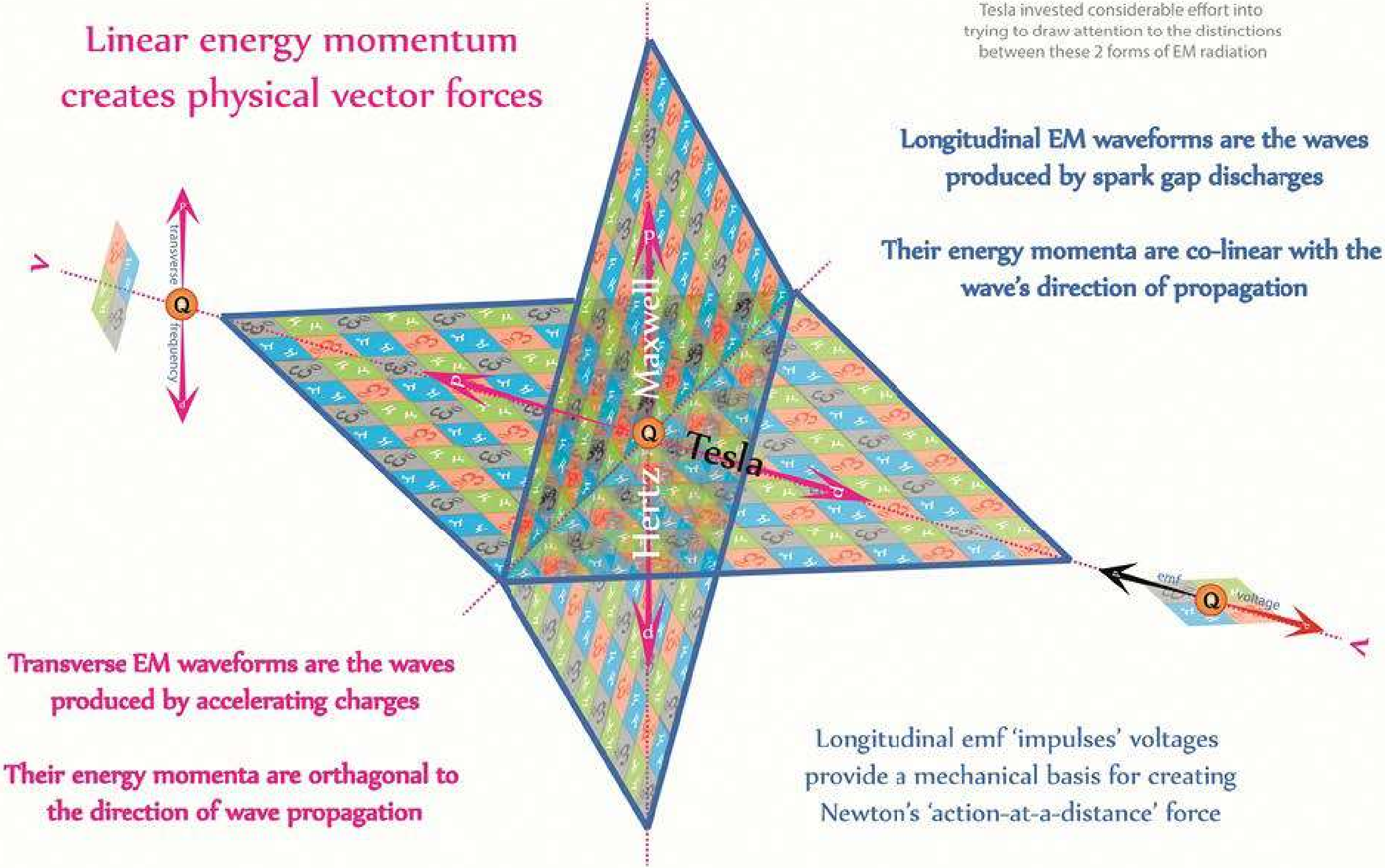

Linear energy momentum creates physical vector forces

Tesla invested considerable effort into trying to draw attention to the distinctions between these 2 forms of EM radiation

Longitudinal EM waveforms are the waves produced by spark gap discharges

Their energy momenta are co-linear with the wave's direction of propagation

Transverse EM waveforms are the waves produced by accelerating charges

Their energy momenta are orthagonal to the direction of wave propagation

Longitudinal emf 'impulses' voltages provide a mechanical basis for creating Newton's 'action-at-a-distance' force

Producing Longitudinal EM waves

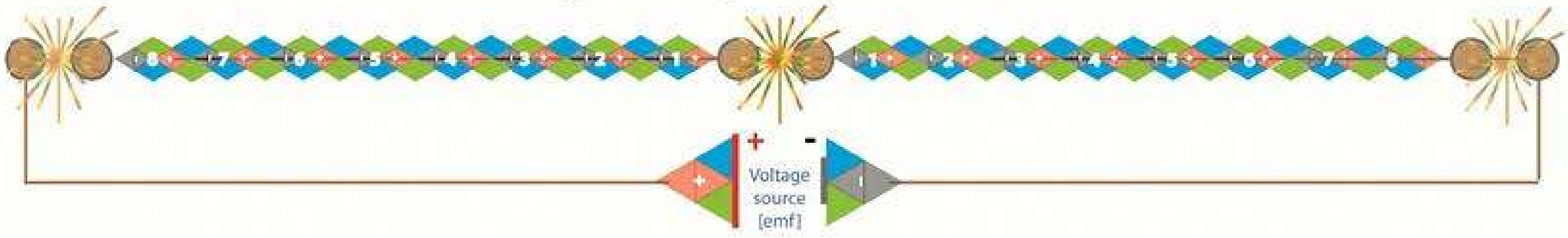

The short circuiting of voltage potentials or spark gap discharges of electrical energy are two ways to produce longitudinal EM waves

After inventing AC electricity Nikola Tesla began to investigate disruptive DC discharges

He noted the deadly power of the spark discharges and how they 'stung his face'

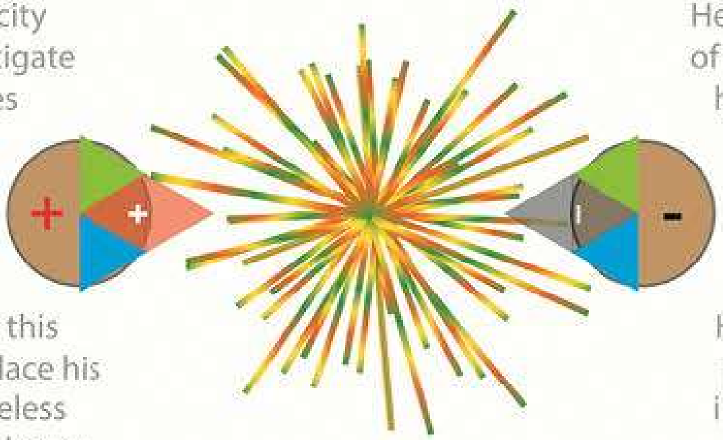

Orthagonal Magnetic fields produce reactive inductance fields

Longitudinal Electric fields produce lines of force

and he was convinced that this form of EM energy would replace his AC system and provide wireless energy to the World without losses

He strove to point out that the load circuits must be inductively coupled to the transmission circuits

or a combination of the two

Very large Voltage

High frequency

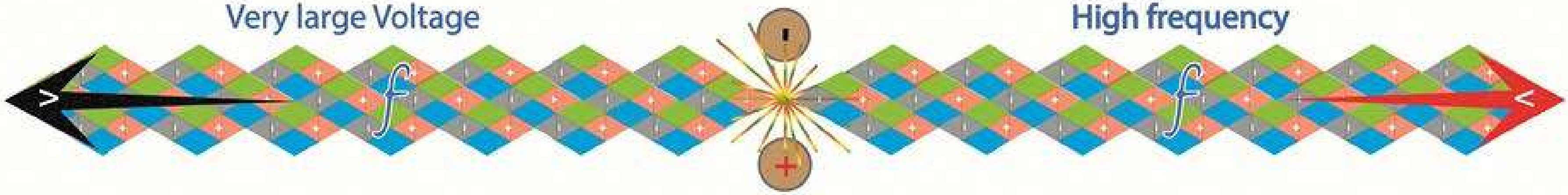

disruptive discharges

Longitudinal energy momenta

Maxwell's equations lead to the prediction of transverse electromagnetic waves as a means of transefering electrical power (where the wave's Electric fields and Magnetic fields vary perpendicularly to their direction of propagation).

Longitudinal Photons & EM waves are 2D electromagnetic waveforms with co-planar energy momenta that is co-linear with their directions of propagation and that can 'cut like a knife'

Longitudinal energy momenta waves [created by the use of spark gap technologies] are the EM energy forms created and radiated by Stars

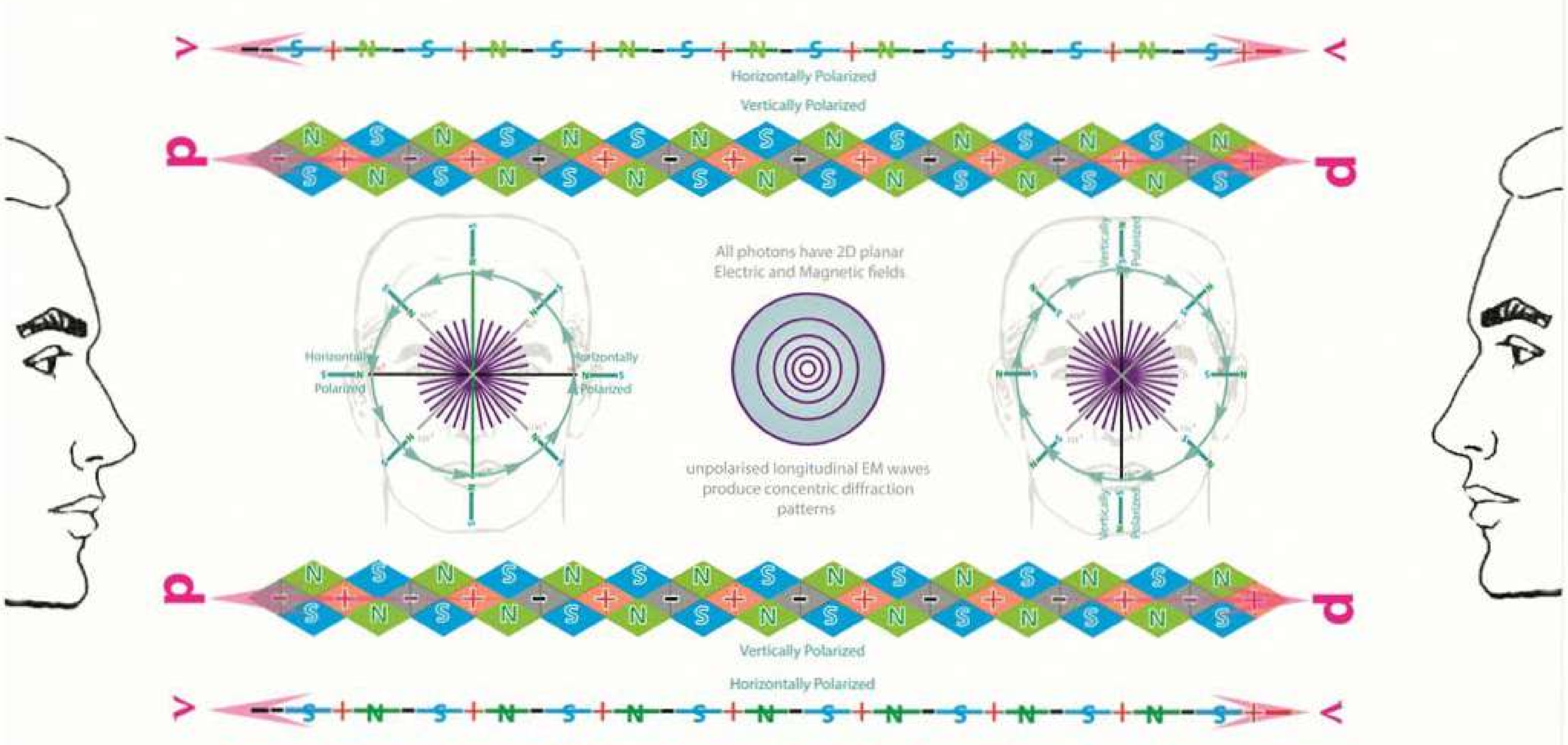
Horizontally Polarized
Vertically Polarized
Horizontally Polarized
Horizontally Polarized
Vertically Polarized
Vertically Polarized
Vertically Polarized
Horizontally Polarized

All photons have 2D planar Electric and Magnetic fields

unpolarised longitudinal EM waves produce concentric diffraction patterns

The longitudinal alignment of E field momenta with their direction of propagation produces 'action-at-a-distance' and can produce material 'faster than light' velocities, if sufficient energies are inputted

Tetryonics 27.09 - Longitudinal energy momenta

The velocity of Electrical energy momenta

The velocity of light is the limit for electrical energy acceleration

(the induced velocity of charged particles resulting from electric field energy-momenta interactions)

c

Transverse EM waves
propagate at energy at
the speed of light

Γ

Longitudinal EM waves
can propagate energies & information
in excess of the speed of light

470,912,891 m/s

463,491,072 km/second >

< 288,000 miles/second

$v = [\pi/2]c$

Longitudinal electrical fields
can create near-instantaneous
'action-at-a-distance' EM fields

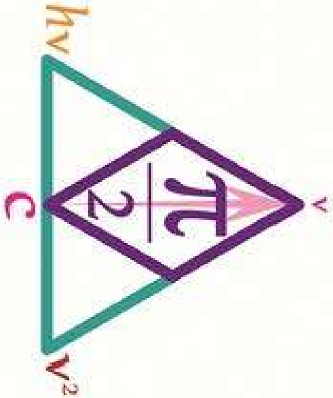

once established, these fields can
transfer momentum and information
at speeds faster than light through
co-linear impulses of momentum

Wheatstone achieved renown by a great experiment
The measurement of the velocity of electrical Energy in a wire.

He cut the wire at the middle, to form a gap which a spark might leap across, and connected its ends to the poles of a Leyden jar filled with electricity. Three sparks were thus produced, one at either end of the wire, and another at the middle. He mounted a tiny mirror on the works of a watch, so that it revolved at a high velocity, and observed the reflections of his three sparks in it.

The points of the wire were so arranged that if the sparks were instantaneous, their reflections would appear in one straight line; but the middle one was seen to lag behind the others, because it was an instant later. The electricity had taken a certain time to travel from the ends of the wire to the middle.

This time was found by measuring the amount of lag, and comparing it with the known velocity of the mirror. Having got the time, he had only to compare that with the length of half the wire, and he could find the velocity of electricity.

His results gave a calculated velocity of 288,000 miles per second,
i.e. faster than what we now know to be the speed of light

Charles Wheatstone

(6 February 1802 – 19 October 1875)

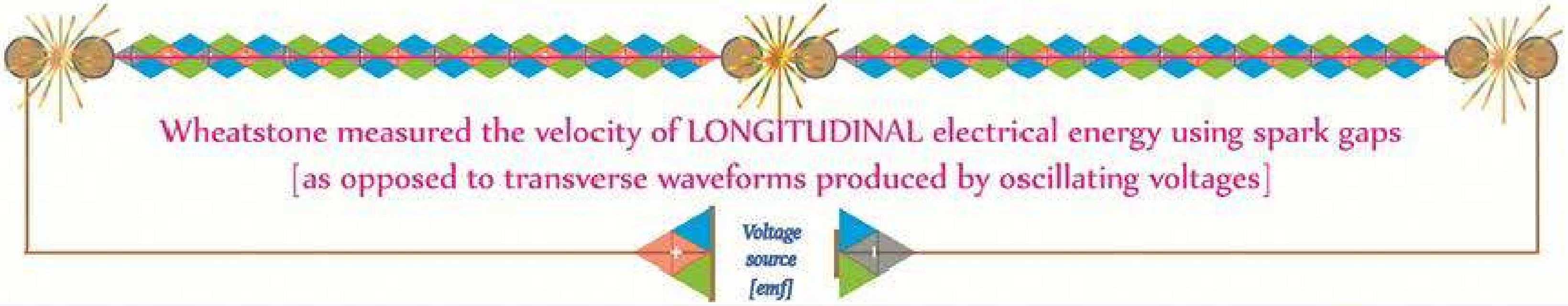

Wheatstone measured the velocity of LONGITUDINAL electrical energy using spark gaps
[as opposed to transverse waveforms produced by oscillating voltages]

Copyright ABRAHAM [2008] - all rights reserved

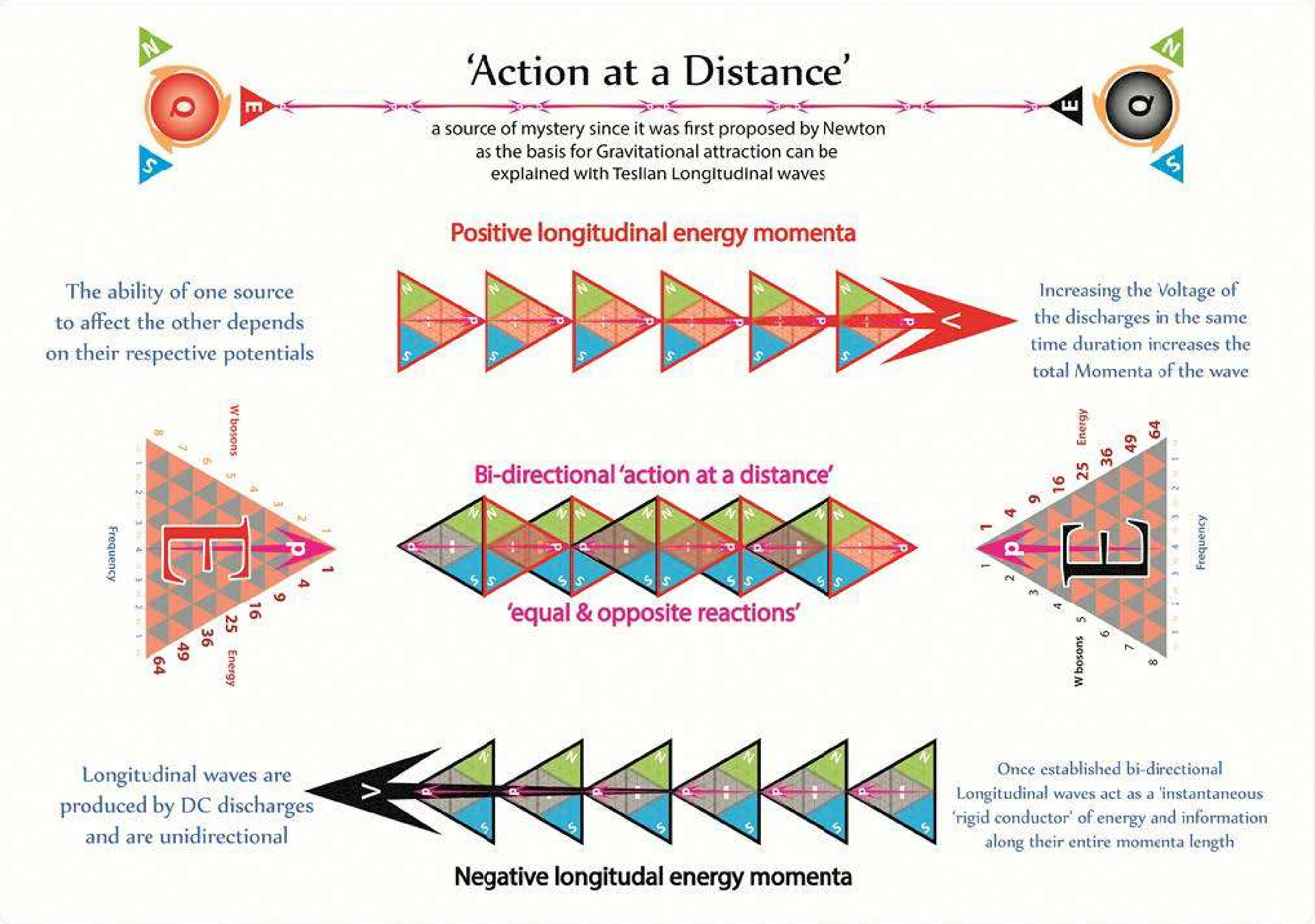

'Action at a Distance'

a source of mystery since it was first proposed by Newton
as the basis for Gravitational attraction can be
explained with Teslian Longitudinal waves

Positive longitudinal energy momenta

The ability of one source to affect the other depends on their respective potentials

Increasing the Voltage of the discharges in the same time duration increases the total Momenta of the wave

Bi-directional 'action at a distance'

'equal & opposite reactions'

Longitudinal waves are produced by DC discharges and are unidirectional

Once established bi-directional Longitudinal waves act as a 'instantaneous 'rigid conductor' of energy and information along their entire momenta length

Negative longitudal energy momenta

Tetryonics 27.11 - Longitudinal 'action at a distance'

Copyright ABRAHAM [2008] - all rights reserved

Longitudinal EM waves have their E-fields co-linear with their direction of propagation

Longitudinal Photons & EM waves are produced by disruptive spark gap discharges

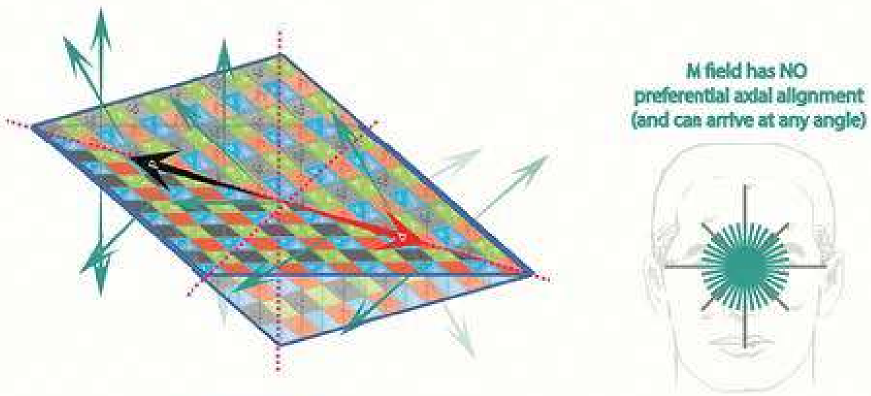
M field has NO preferential axial alignment (and can arrive at any angle)

Un-Polarised

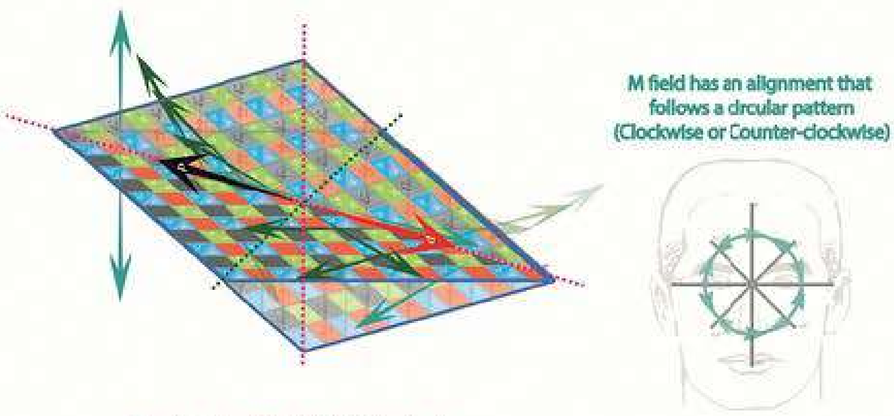
M field has an alignment that follows a circular pattern (Clockwise or Counter-clockwise)

Circularly (or Elliptically) Polarised

Longitudinal wave polarisations

emf
Q
voltage

As a convention, the polarization of longitudinal EM waves must be described by specifying the orientation of the wave's magnetic field at a point in space over one period of the oscillation

Longitudinal waves have their linear momenta aligned along their direction of propagation and are extremely efficient at accelerating charges they interact with, creating co-linear accelerations [parallel to their direction of field propagation] - they do not produce charge oscillations along their vector direction of propagation like transverse waves.

Examples of transverse waves include seismic P (primary) waves, and Tesla's longitudinal 'stinging' rays, whose electric fields and energy momenta are both co-linear with each other and parallel to the direction of energy transfer.

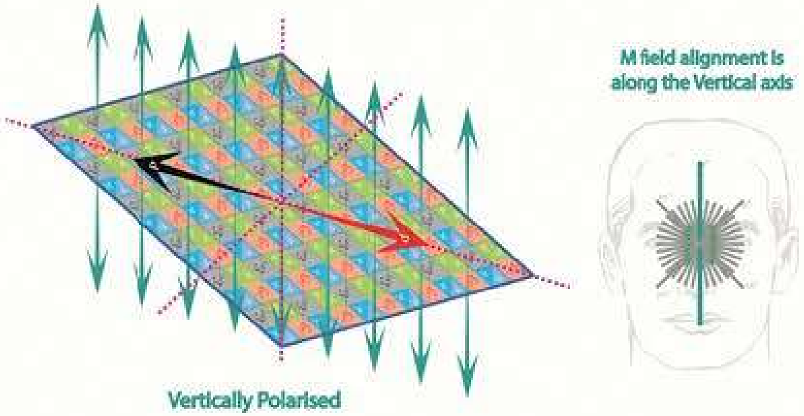
M field alignment is along the Vertical axis

Vertically Polarised

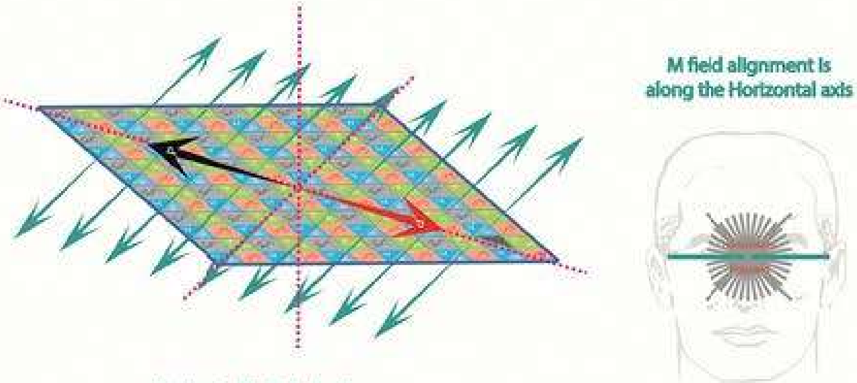
M field alignment is along the Horizontal axis

Horizontally Polarised

Tetryonics 27.12 - Longitudinal wave polarisation

ElectroMagnetic waveforms

In Teslian longitudinal waves, produced by disruptive EM discharges, the E fields are co-linear with the direction of propagation

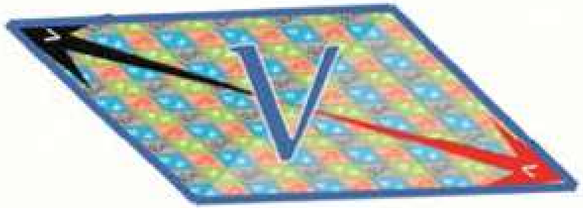

After inventing AC electricity Tesla investigated transferring energy through the air via electrostatic longitudinal waves to receiving devices. and noted the differing 'lethality' of the forces produced by these waves

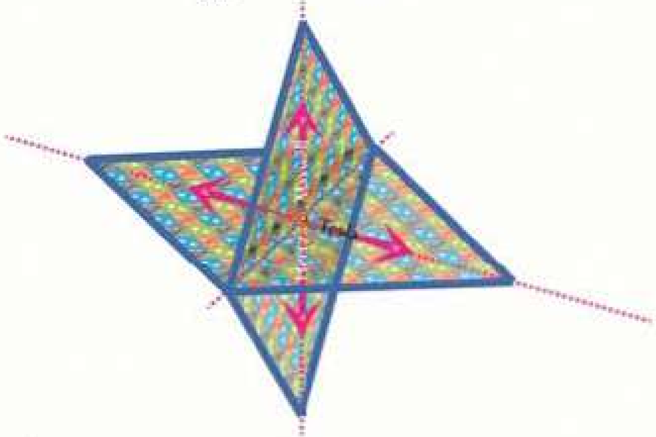

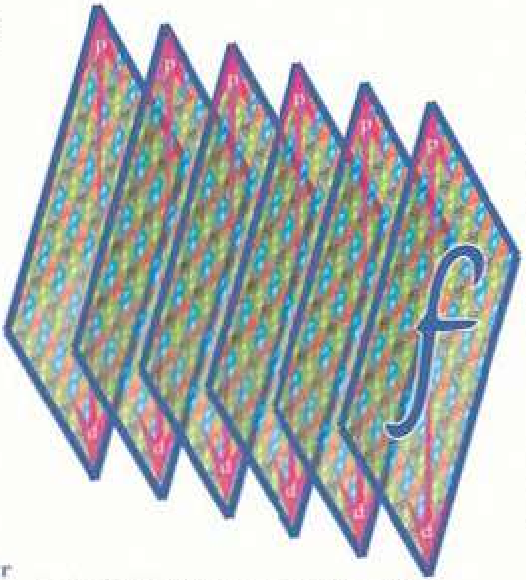

Hertzian transverse waveforms have their EM fields perpendicular to the direction of propagation

Although comprised of the same energy momenta quanta
EM waves can be formed with Electric fields perpendicular to each other
thus leading to conflicting theories of EM wave propagation

In 1887, Heinrich Hertz demonstrated the reality of Maxwell's electromagnetic waves by experimentally generating radio waves in his laboratory

T

Teslian waves
transmit Energy
in a LONGITUDINAL
waveform
producing
'action-at-a-distance'

Nikola Tesla

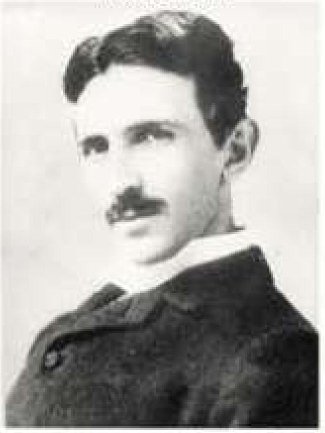

(10 July 1856 - 7 January 1943)

Volt Seconds [per metre^2]

W/m^2

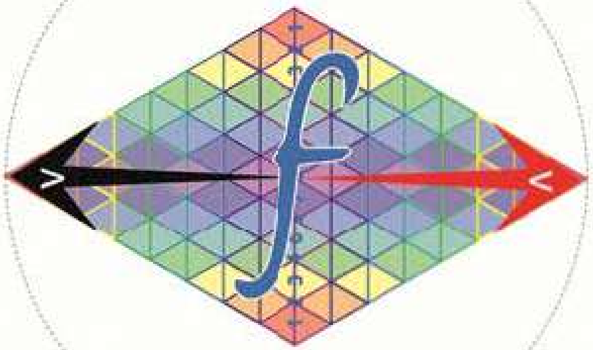

108 Mx = 1 J/A = 1 V·s = 1 Wb = 1 T·m2

c^2
sec

Heinrich Hertz

(22 Feburary 1857 - January 1 1894)

Hz

Hertzian waves
transmit Energy
in a TRANSVERSE
waveform
producing
'radio waves'

Cycles per Second

Soon after Hertz's claim of discovering Maxwell's transverse EM waves Tesla visited him and personally demonstrated the experimental error to him.
Hertz agreed with Tesla and had planned to withdraw his claim, but varying agendas intervened and set the stage for a major rift in the 'accepted' theories

Transverse EM waves have their E&M fields orthagonal with their direction of propagation

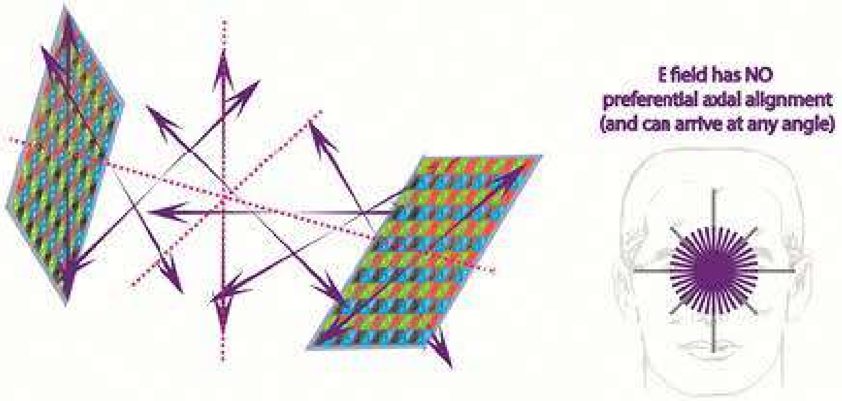
E field has NO
preferential axial alignment
(and can arrive at any angle)

Un-Polarised

Transverse Photons & EM waves are produced by accelerating/oscillating charges

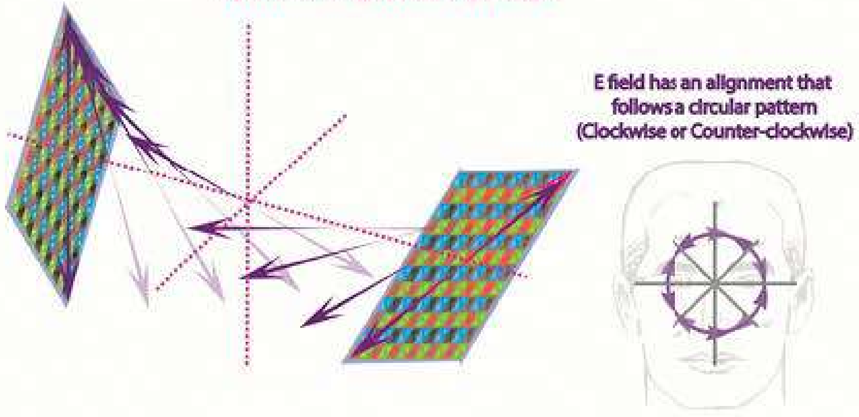
E field has an alignment that
follows a circular pattern
(Clockwise or Counter-clockwise)

Circularly (or Elliptically) Polarised

transverse
Q
frequency

Transverse wave polarisations

By convention, the polarization of EM waves is described by specifying the orientation of the wave's electric field at a point in space over one period of the oscillation

Transverse waves are inefficient propagators of EM energy as they apply their linear momenta tangentially to the charges they interact with, creating lateral accelerations [orthogonal to their direction of field propagation] - they do not accelerate charges along their vector direction of propagation like longitudinal waves.

Examples of transverse waves include seismic S (secondary) waves, and electromagnetic plane wave, whose electric (E) and magnetic (M) fields both oscillate perpendicular to each other and orthogonal to the direction of energy transfer.

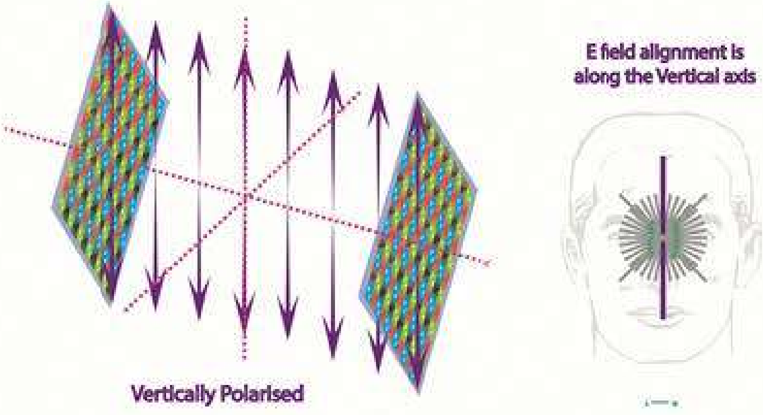
E field alignment is
along the Vertical axis

Vertically Polarised

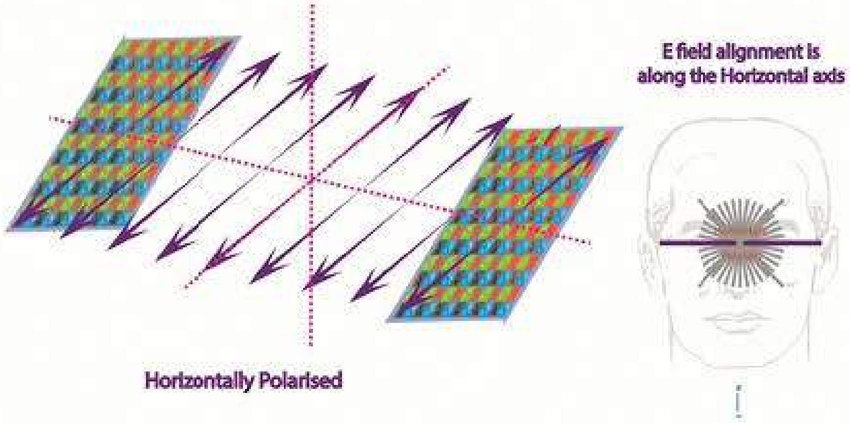
E field alignment is
along the Horizontal axis

Horizontally Polarised

Tetryonics 27.14 - Transverse wave polarisation

EM wave modulations

In telecommunications, modulation is the process of varying one or more properties of a periodic waveform, called the carrier signal, with a modulating signal which typically contains information to be transmitted

Amplitude modulation

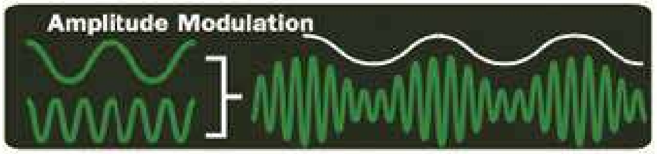
Amplitude Modulation

the amplitude of the carrier signal is varied

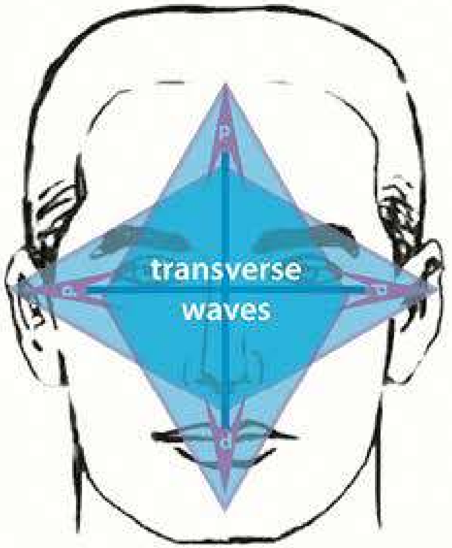
transverse
waves

the frequency of the carrier signal is varied

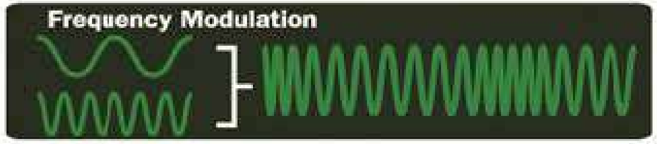
Frequency Modulation

Frequency modulation

In addition to planar Hertzian waves
there exists Longitudinal EM waves

Impulse propagation

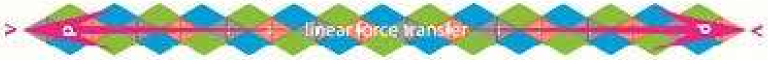
linear force transfer

As the voltage [energy/coulomb] of longitudinal waves increases their co-linear energy-momenta increase and these forces provide a mechanical basis for near instantaneous 'action-at-a-distance'

Impulse propagation is distinct from pulse modulations

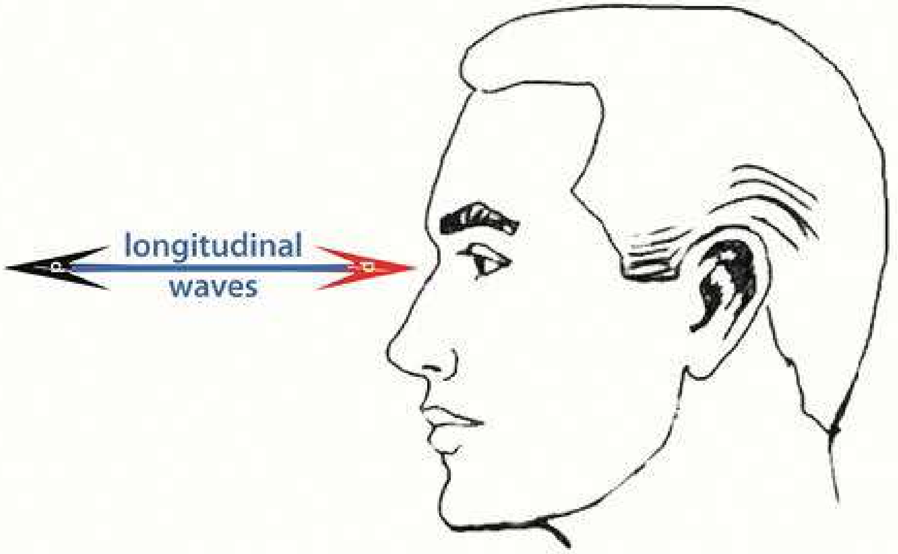
longitudinal
waves

a narrowband analog signal is pulsed over an analog baseband channel

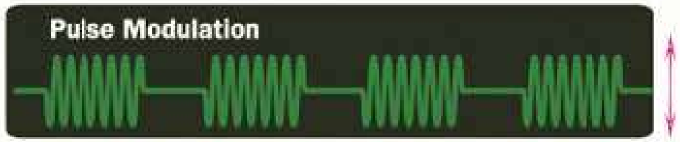
Pulse Modulation

PCM Pulsed modulation **TDM**

Transverse EM wave production

Alternating the Voltage potentials in an electrical circuit produces an Alternating Current
of electrons with Kinetic EM energies reflective of the Time-Energy duration of the AC circuit

**Alternating Voltages provide an electromotive force to Electrically charged Particles
producing tranverse Kinetic Energy fields with orthagonal Magnetic fields**

Kinetic Energies from motion
(diverge from rest Matter)

Kinetic
Matter

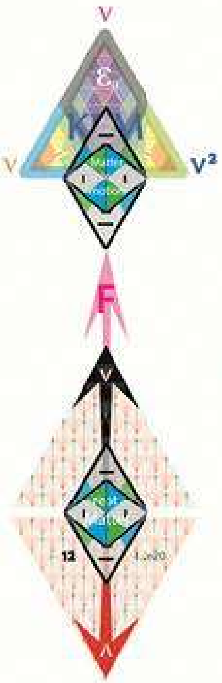

All Stationary charged particles have
NO extrinsic Magnetic moment
ie. They are Electro-static

F=ma

$$\sum \mathbf{F} = \frac{\mathrm{dp}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \mathbf{v}\frac{\mathrm{d}m}{\mathrm{d}t}$$

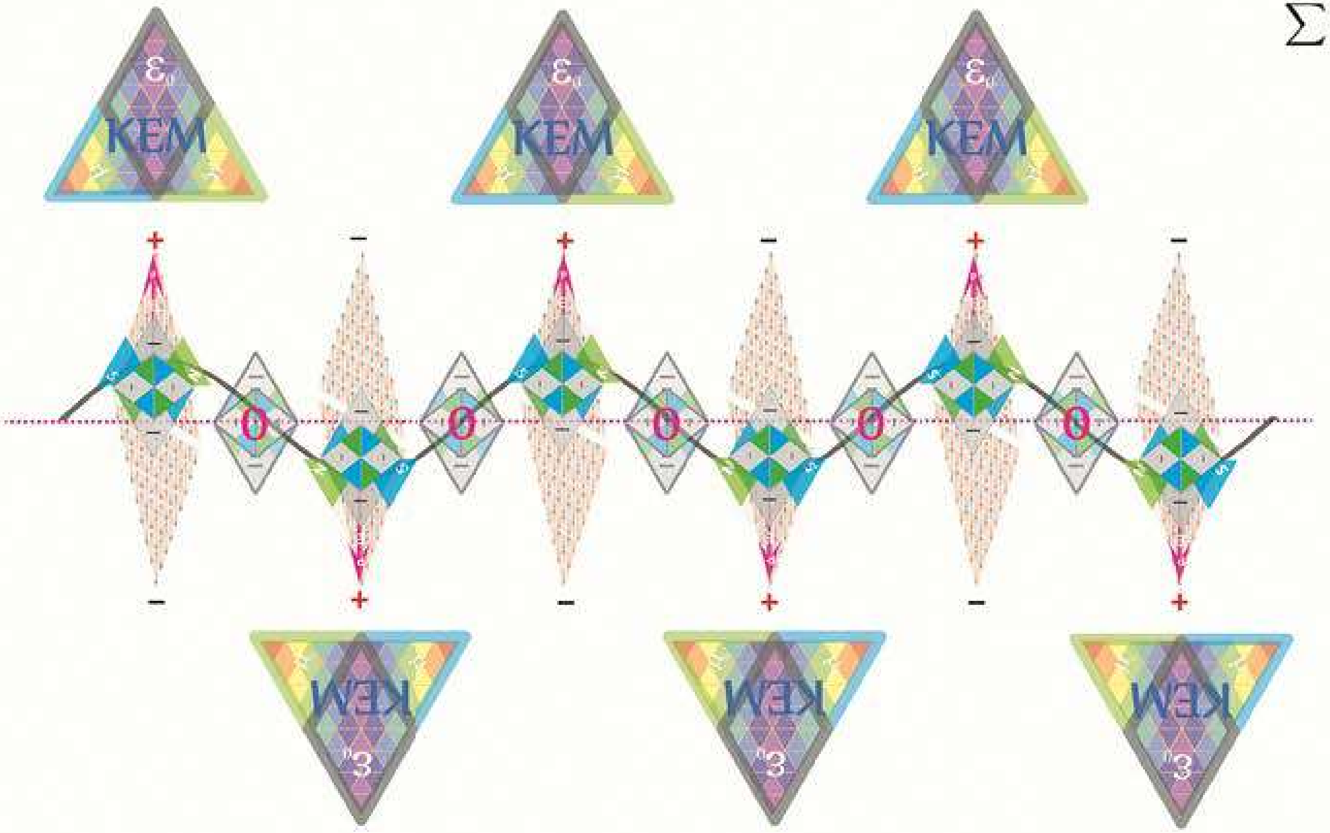

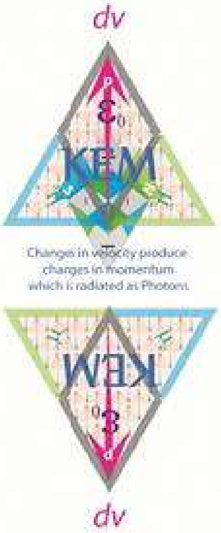

Changes in velocity produce
changes in momentum
which is radiated as Photons

Charges with constant velocities
produce
REACTIVE [K]EM FIELDS

**Any change to the momentum of a charged particle in motion
requires a corresponding quantum level energy-momentum change through the
emission/absorption of quantised EM mass-ENERGY momenta in the form of W bosons**

Accelerating Charges
produce
RADIANT EM WAVES

Modelling of planar transverse waves
(of varying amplitudes)

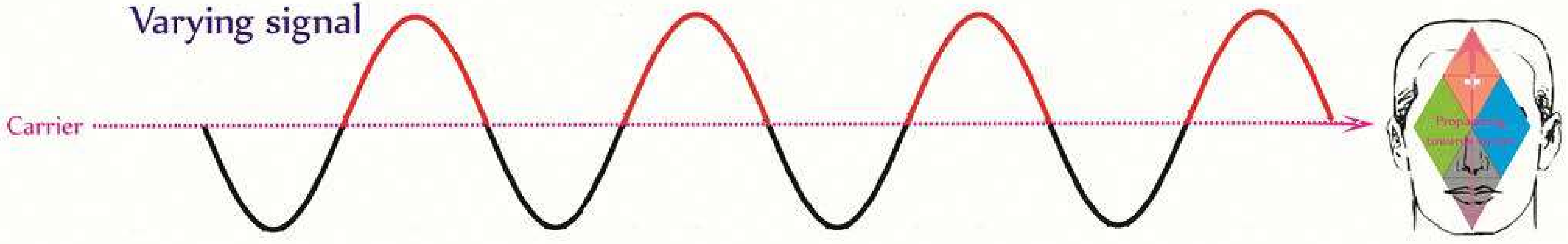
Varying signal
Carrier

AM waveform

Amplitude Modulation (AM) works by varying the strength of the transmitted signal in relation to the information being sent.

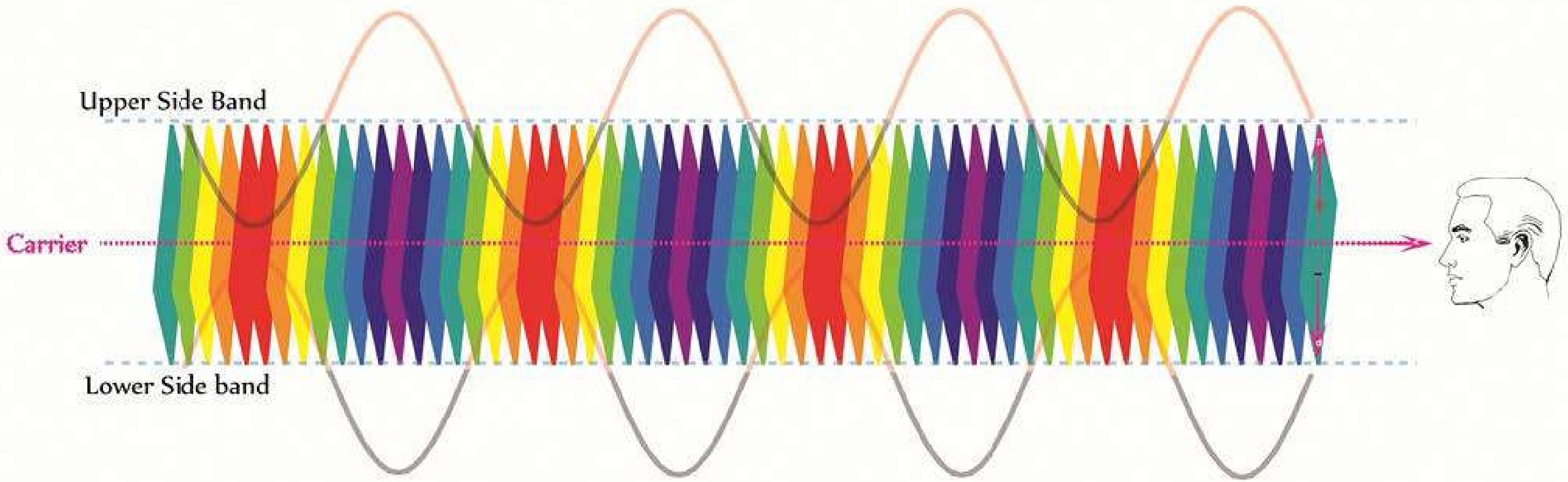
Upper Side Band
Carrier
Lower Side band

Note: In amplitude modulation each colour planar wave contains differing energy momenta [n1-red to n8-violet]

Modelling of planar transverse waves
(of the same frequencies)

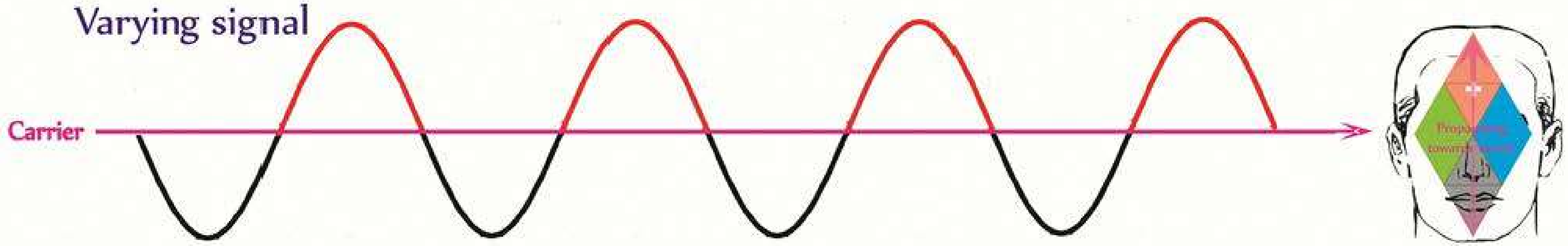

FM waveform

Frequency modulation (FM) is a form of modulation which conveys information over a carrier wave by varying its frequency (in contrast with amplitude modulation, in which the amplitude of the carrier is varied while its frequency remains constant).

In analog applications, the instantaneous frequency of the carrier is directly proportional to the instantaneous value of the input signal

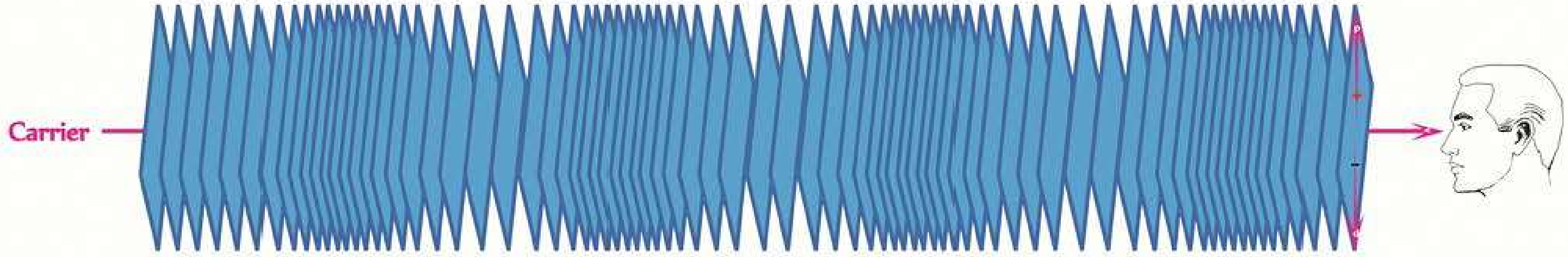

Note: In Frequency modulation all the planar waves have the same energy momenta [only the number of waves per sec varies]

Tetryonics 27.18 - Frequency Modulation

ElectroMagnetic waves

Mathematics alone [without a correct foundational geometry] is insufficent to describe EM radiation
EM waves can be Longitudinal or Transverse with respect to their direction of propagation

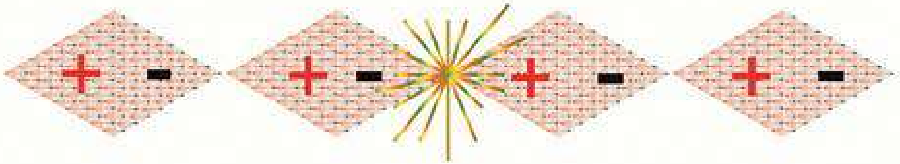

Tesla

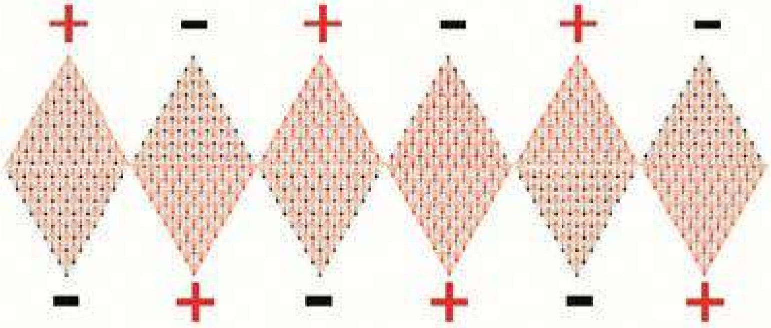

Maxwell - Hertz

It is now clear [using Tetryonic geometry] that both Tela and Maxwell were correct in their opposing theories of EM waveforms with Tesla proposing Longitudinal waves [from experiment] and Maxwell proposing Transverse waves [from his equations].

The differing opinions on the EM waveforms is shown to be a direct result of the sources of EM wave generation employed in their creation. After the discovery of the photo-electric effect Hertz and others moved to oscillating charges to produce Transverse [planar] waves while Tesla continued to investigate and utilise the older spark gap technology to produce Longitudinal waves

Spark Gaps & Plasma discharges produce longitudinal EM radiation

Accelerating charges produce transverse EM radiation

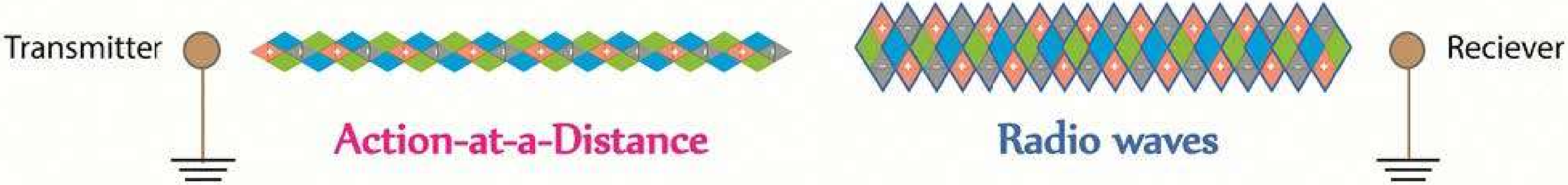

In order to explain and advance our understanding of all EM phenomena we must return to, and utilise, the older discharge technologies

Planar electromagnetic waves

In the physics of wave propagation, a plane wave (also spelled planewave) is a wave whose wavefronts are infinite parallel planes of peak-to-peak amplitude normal to the phase velocity vector.

transverse Radio waves **longitudinal Spectral lines**

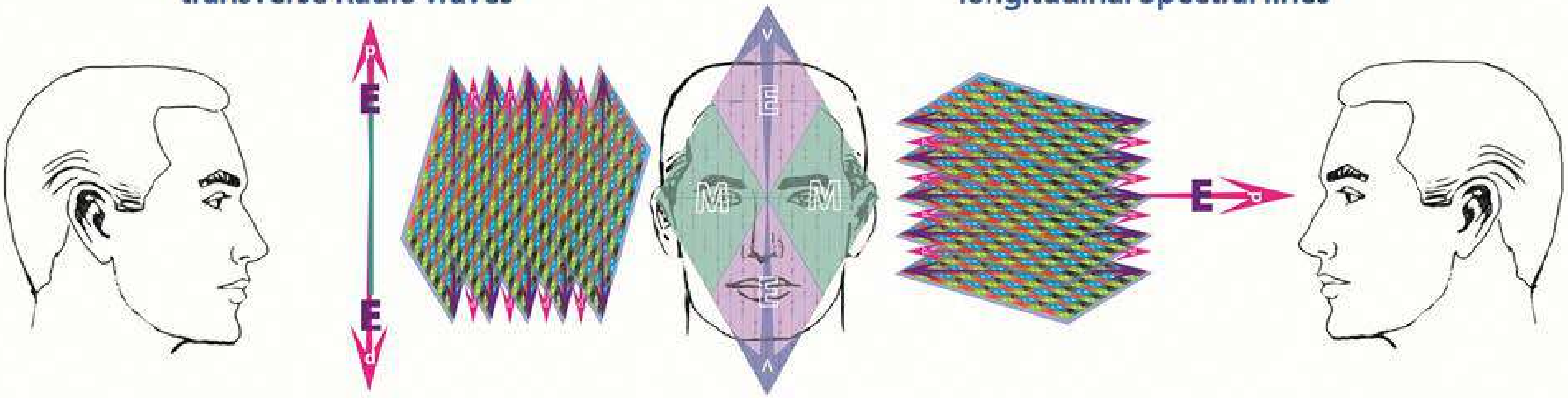
p
E
E
p
v
E
M
M
E
v
E
p

The phase relationships of their EM energy momenta can all be finally visualised and drawn through Tetryonic theory geometries

[with appropriate radian units]

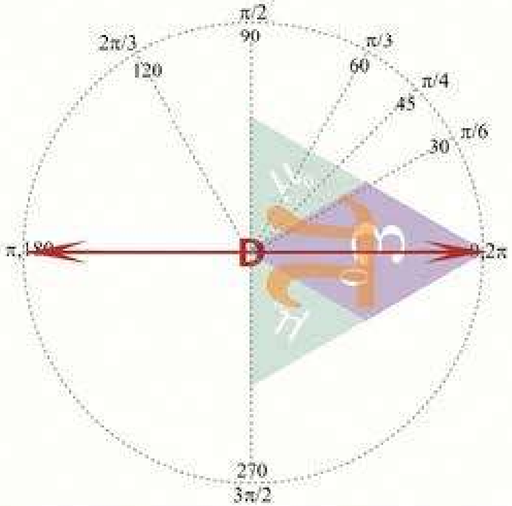
π/2
90
2π/3
120
π/3
60
π/4
45
π/6
30
π,180
0,2π
D
270
3π/2

bosons $2\pi[h\nu] = hf$ photons

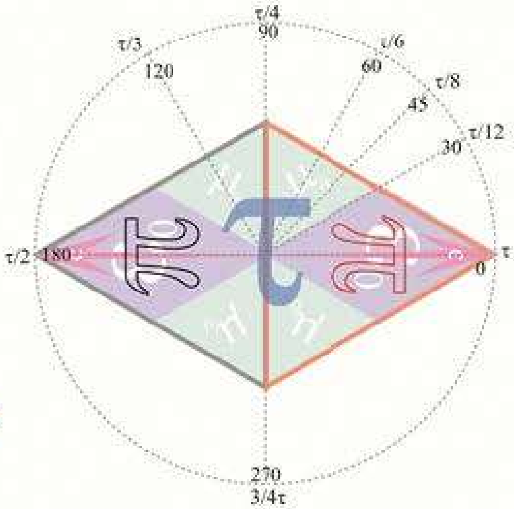
τ/4
90
τ/3
120
τ/6
60
τ/8
45
τ/12
30
τ/2
180
τ
0
270
3/4τ

τ is a more 'natural' radian unit for describing EM waves than π
tau = 2π = 360 degree rotation about a point

Lorentz co-ordinate transforms

In physics, the Lorentz transformation (or transformations) are named after the Dutch physicist Hendrik Lorentz.

The transformations describe how measurements of mass-energies
In space and time by two observers are related through mathematics

It was the result of attempts by Lorentz and others to explain how the speed of light was observed to be independent of the reference frame, and to understand the symmetries of the laws of electromagnetism.

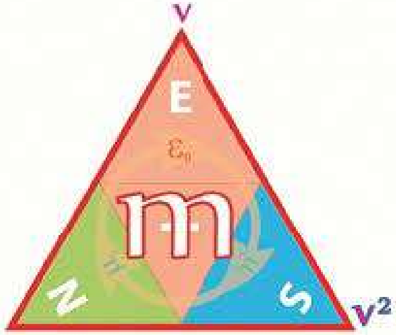

All photons are comprised of two dual-sided equilateral Planck quanta

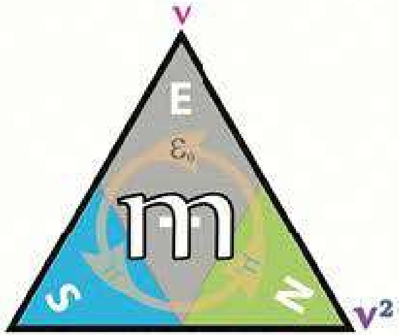

All EM waves are comprised of neutral charge quanta called Photons

Photons are comprised of mass-energies and the term 'mass-less' is not appropriate

Forces add mass-energies to KEM fields resulting in WAVElength changes

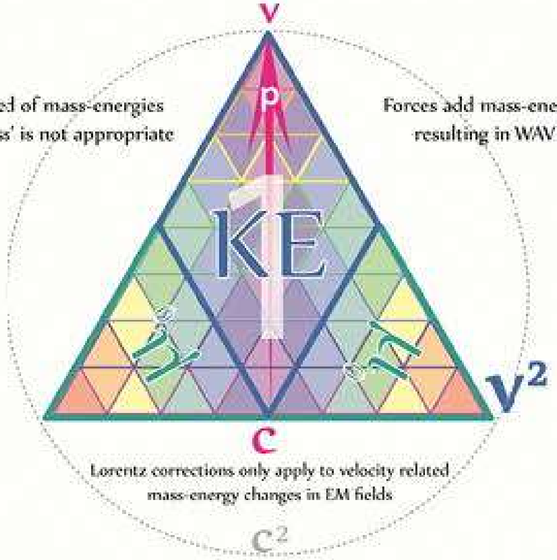

Lorentz corrections only apply to velocity related mass-energy changes in EM fields

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

Linear Lorentz boosts produce changes to the velocity and momenta of a system

Scalar correction factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

As the velocity and momenta of a system changes so does its scalar energies

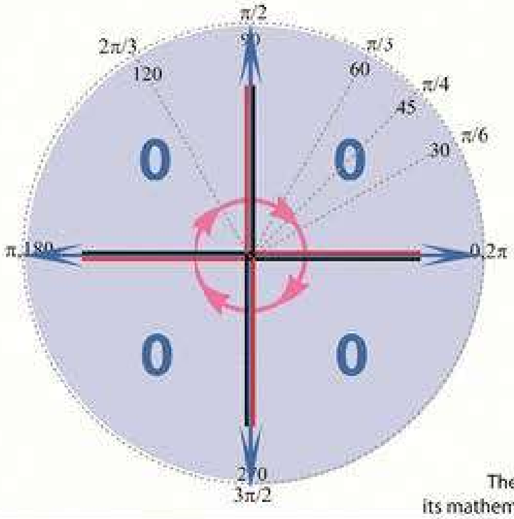

The Lorentz transformation is a linear transformation

It may include a rotation of space;
a rotation-free Lorentz transformation is called a Lorentz boost.

bosons $$2\pi[hv] = hf$$ photons

π

As photons are neutral charge mass-energy geometries care must alway be taken in identifying and specifying the planar wavefront polarities and direction of wave propagation when developing co-ordinate systems to describe their motions and resulting electromagnetic properites

τ

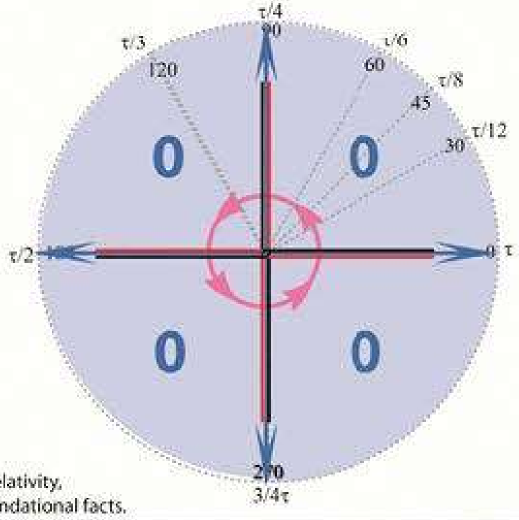

The Lorentz transformation is in accordance with special relativity, but was derived well before special relativity, its mathematical development in relativity theory has resulted in erroneous assumptions being accepted as foundational facts.

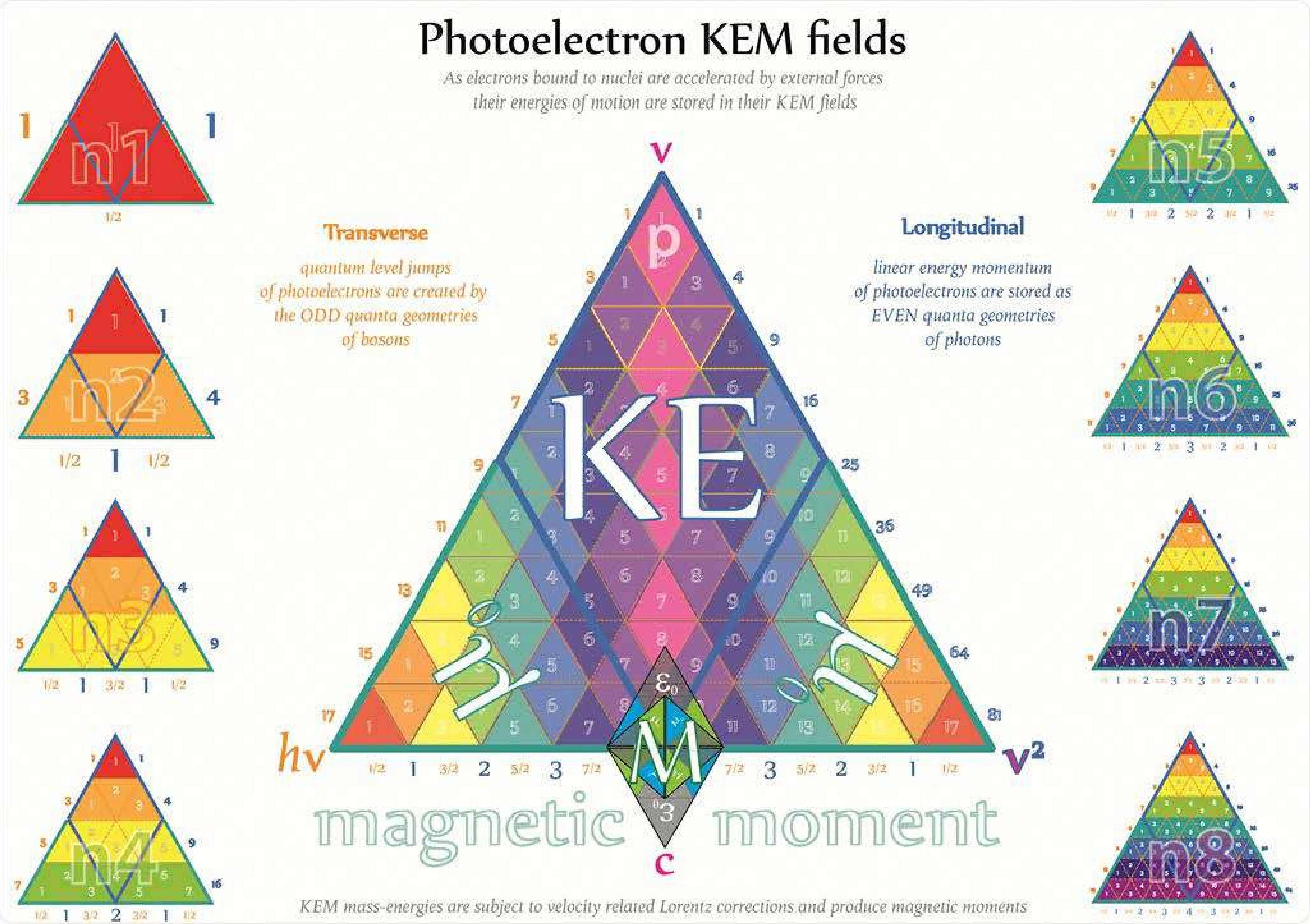

Photoelectron KEM fields

As electrons bound to nuclei are accelerated by external forces their energies of motion are stored in their KEM fields

Transverse

quantum level jumps of photoelectrons are created by the ODD quanta geometries of bosons

Longitudinal

linear energy momentum of photoelectrons are stored as EVEN quanta geometries of photons

KEM mass-energies are subject to velocity related Lorentz corrections and produce magnetic moments

n1
n2
n3
n4
n5
n6
n7
n8
v
p
KE
hv
v^2
ε_0
M
μ_0
c
magnetic
moment
1/2
1
3/2
2
5/2
3
7/2
1
3
5
7
9
11
13
15
17
1
4
9
16
25
36
49
64
81

Tetryonics 28.01 - Photoelectron KEM fields

Kinetics & Photons

The photon has several properties that distinguish it from all other subatomic particles.

It is the only elementary particle wherein a high-energy photon can transform/split into two or more low-energy photons.

A photon is a magnetic dipole. It is an elementary magnet. Evans discovery of the photon's longitudinal magnetic field in 1992 is considered as significant as Einstein's discovery of Relativity.

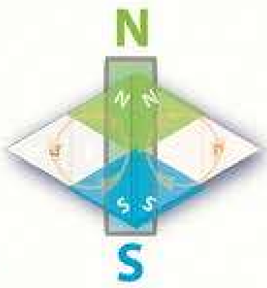

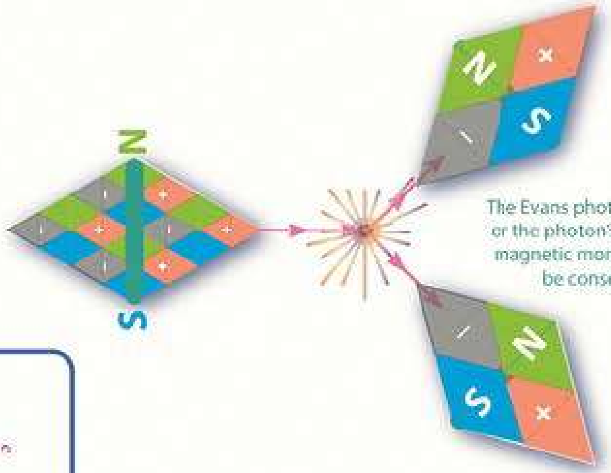

The Evans photomagneton or the photon's transverse magnetic moment has to be conserved.

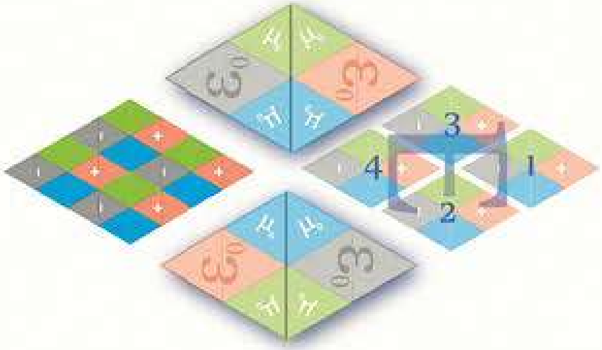

Photons can form super-position EM waves

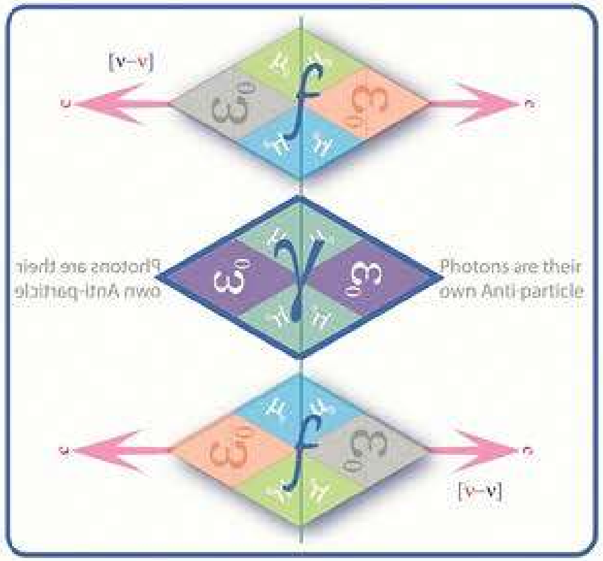

Photons are their own Anti-particle

PAIR PRODUCTION
A photon with sufficient energy can be transformed into a real electron-positron pair.

The EM mass-energy is conserved as per Einstein's equivalence of mass and energy.

The photon's magnetic charge is conserved in the magnetic moments of the created electron-positron pair.

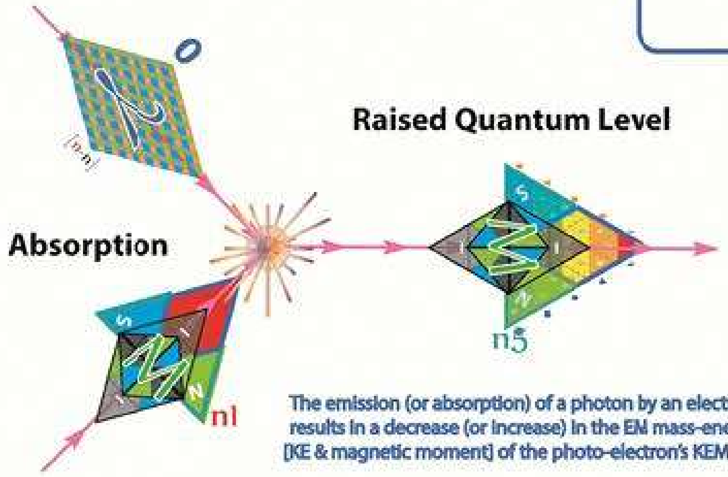

The emission (or absorption) of a photon by an electron, results in a decrease (or increase) in the EM mass-energy [KE & magnetic moment] of the photo-electron's KEM field

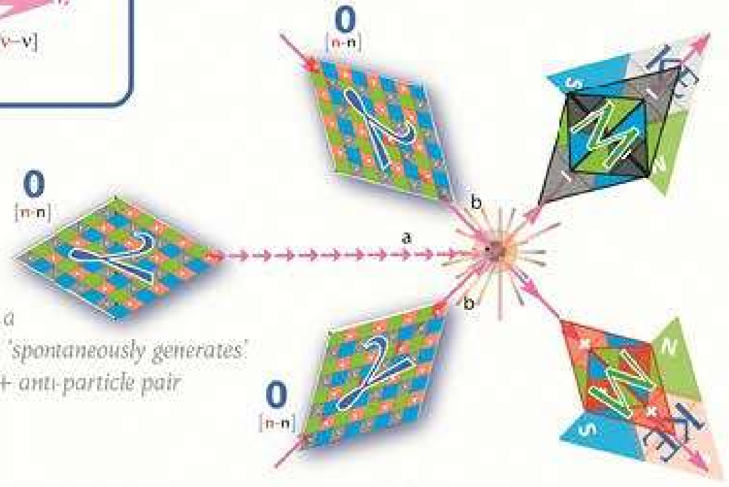

Option a
Photon 'spontaneously generates' a pair + anti-particle pair

Option b
Two Photons combine

Arthur Compton

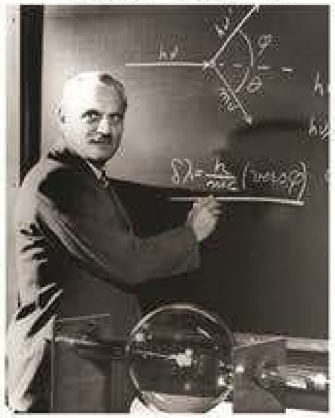

(September 10, 1892 - March 15, 1962)

Compton Scattering

The in-elastic scattering of photons in matter results in a decrease in energy (increase in wavelength) of an X-ray or gamma ray photon and is called the Compton effect.

Part of the energy of the X/gamma ray is transferred to a scattering electron, which recoils and is ejected from its atom (which becomes ionized), and the rest of the energy is taken by the scattered, "degraded" photon.

Compton frequency of photons

$$E = hf = \frac{hc}{\lambda} = mc^2$$

where v = c

Lorentz corrected energies of EM masses in motion

$$n\pi\frac{\left[m\Omega v^2\right]}{c^2}$$

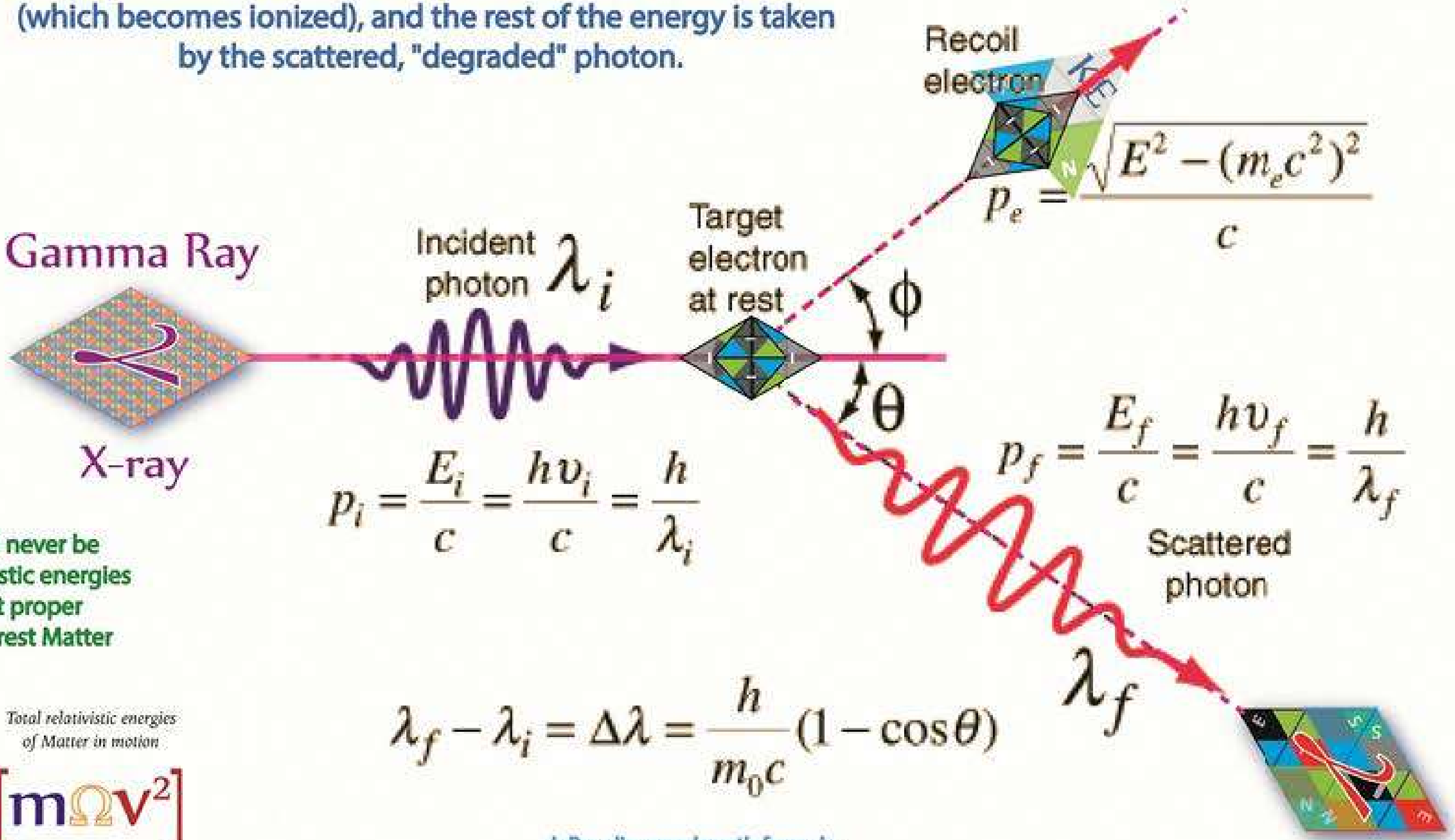

$$p_e = \frac{\sqrt{E^2 - (m_e c^2)^2}}{c}$$

$$p_i = \frac{E_i}{c} = \frac{h\upsilon_i}{c} = \frac{h}{\lambda_i}$$

$$p_f = \frac{E_f}{c} = \frac{h\upsilon_f}{c} = \frac{h}{\lambda_f}$$

Compton frequencies should never be calculated from the total relativistic energies of Matter in motion without proper consideration of the invariant rest Matter

$$\lambda_f - \lambda_i = \Delta\lambda = \frac{h}{m_0 c}(1 - \cos\theta)$$

deBroglie wavelength formula

Degraded Photon

Total relativistic energies of Matter in motion

$$n\pi\frac{\left[m\Omega v^2\right]}{c^4}$$

Pair Production

$$mv^2 = E = \tfrac{1}{2}hf$$

A photon with sufficient energy can be transformed into a real, material electron-positron pair.

The EM mass-energies of the system are conserved as per the equivalence of mass and Energy.

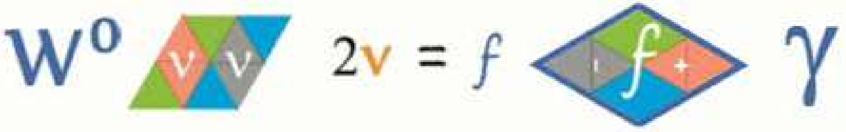
W^0
$2v = f$
γ

The photon's magnetic dipole and charges are also conserved (but in differing charge geometries)

Option a

Photon produces 2 particles

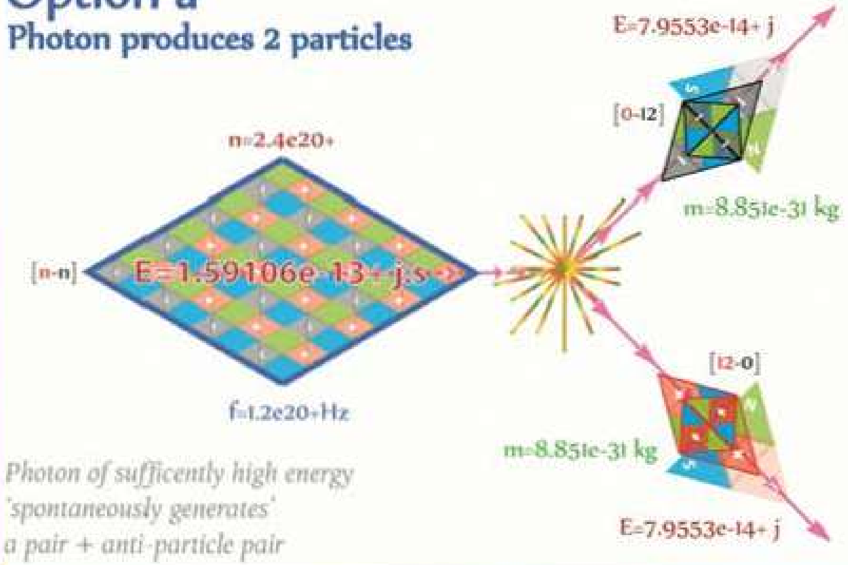
E=7.9553e-14+ j
[0-12]
n=2.4e20+
m=8.851e-31 kg
[n-n]
E=1.59106e-13+ j.s
f=1.2e20+Hz
[12-0]
m=8.851e-31 kg
E=7.9553e-14+ j

Photon of sufficently high energy 'spontaneously generates' a pair + anti-particle pair

Option b

Two Photons combine

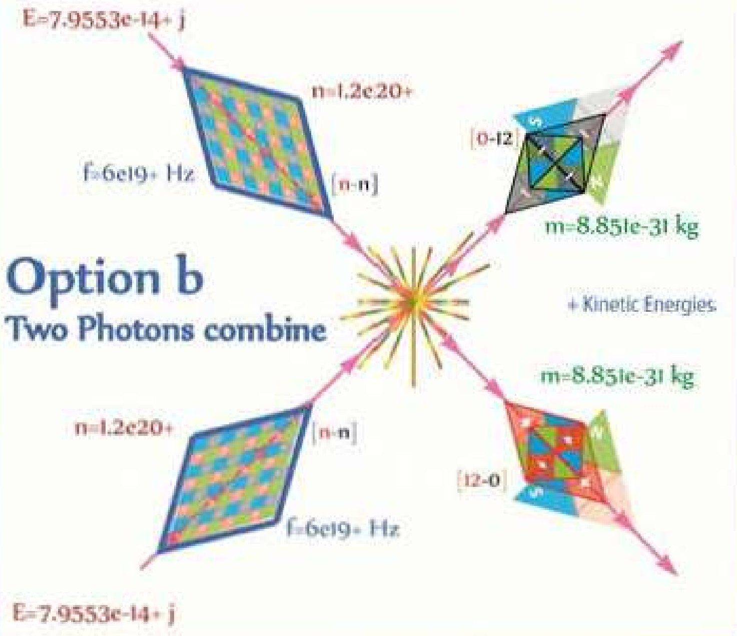
E=7.9553e-14+ j
n=1.2e20+
f=6e19+ Hz
[n-n]
[0-12]
m=8.851e-31 kg
+ Kinetic Energies
m=8.851e-31 kg
n=1.2e20+
[n-n]
[12-0]
f=6e19+ Hz
E=7.9553e-14+ j

Planck's constant	Energy	electron rest mass-Matter
6.62432672 e-34 kgm^2/s	7.9553 e-14 j	8.8514860 e-31 kg

$$\boxed{n\pi\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right] = E = T\pi\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]}$$

6 e19
photon frequency

1.2 e20
Compton frequency

All pair production - annihilations must follow the mass-Energy-Matter equivalency formula

mass $c^2 \xleftarrow{\text{per second}} E \xrightarrow{\text{per second squared}} c^4$ Matter

Kinetic Energies

$$KE = \frac{1}{2}Mv^2$$

As velocity increases so does the number of energy momenta quanta per spatial geometry

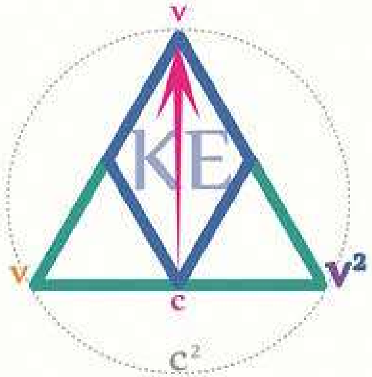

Photons contain Kinetic energies and momentum

Photonic Energy

$$E_\gamma = 2\pi\Big[\big[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\big]\Big]$$

Kinetic Energy and momenta are all related through Tetryonic geometry

Photons

$$p = \frac{h}{\lambda} = \frac{E}{c}$$

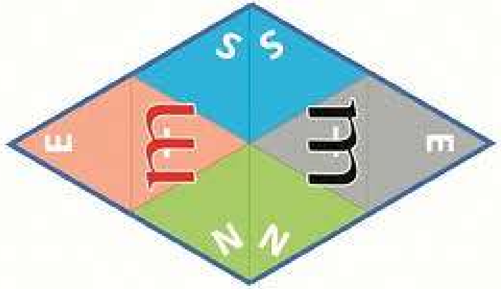

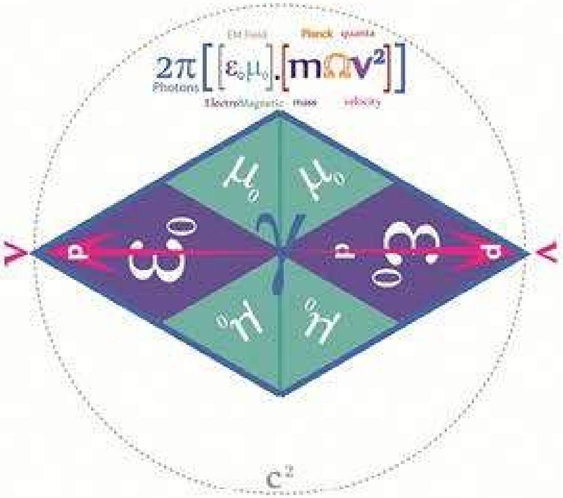

EM field
Planck quanta
2π[[ε₀μ₀].[mΩv²]]
Photons
ElectroMagnetic
mass
velocity
μ₀ μ₀
v
p
ε₀
γ
p
ε₀
p
v
μ₀ μ₀
c²

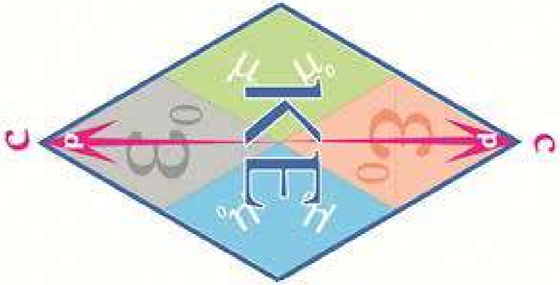

μ₀ μ₀
c
p
ε₀
KE
ε₀
p
c
μ₀ μ₀

$$E = hf$$

Momentum

$$\frac{p^2}{2m} = KE = \frac{1}{2}Mv^2$$

As the number of Photons increases so does the Momentum

$$p = \frac{n\pi\Big[\big[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\big]\Big]}{v} = E/v$$

All Photons travel at the 'speed of light' in their medium of transmission

$$p = \frac{E}{c} = \frac{h\nu}{c} = \frac{h}{\lambda}$$

$$p = n\pi\Big[\big[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v}\big]\Big]$$

linear energy-momentum

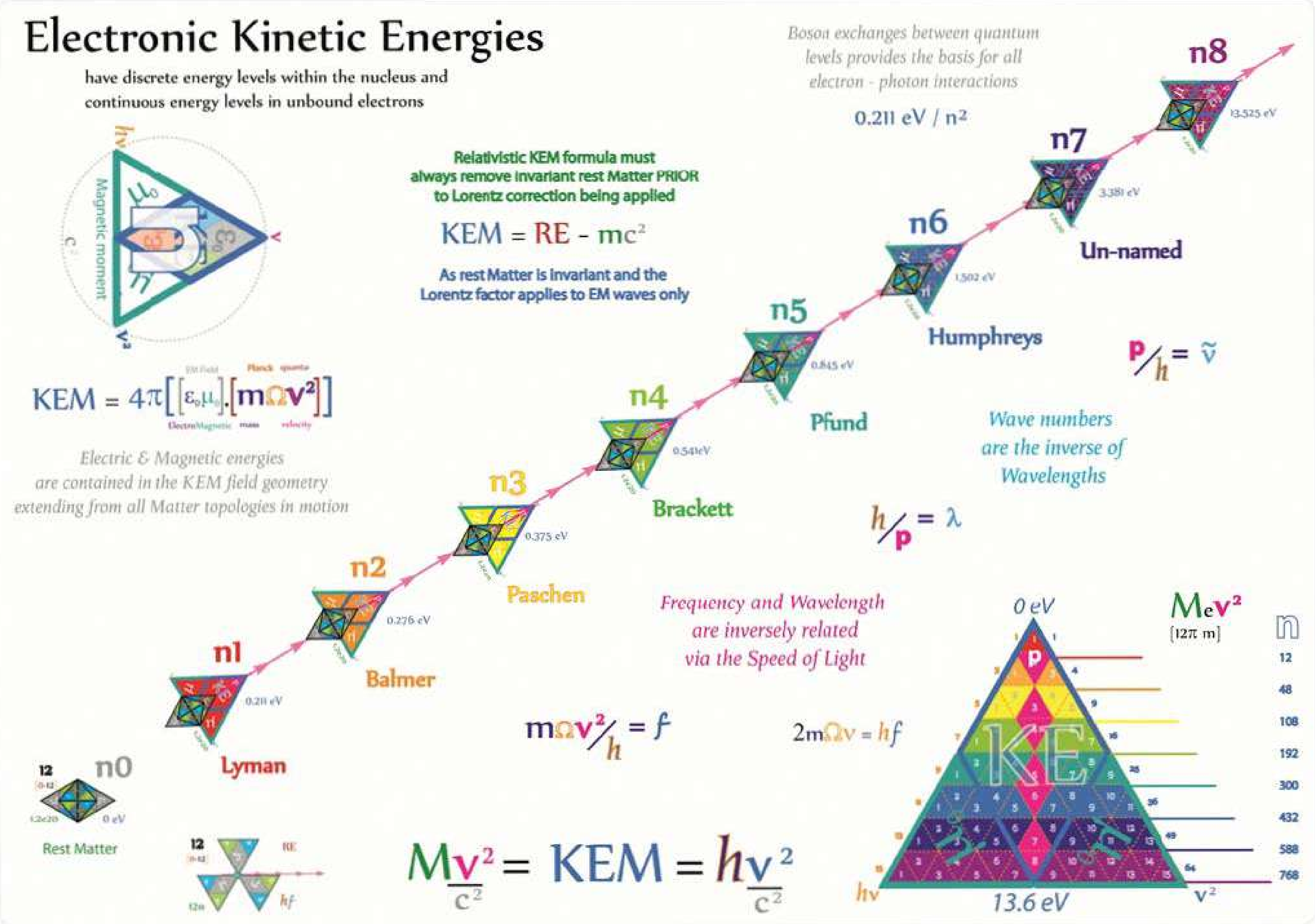

Electronic Kinetic Energies

have discrete energy levels within the nucleus and
continuous energy levels in unbound electrons

Boson exchanges between quantum
levels provides the basis for all
electron - photon interactions

$0.211\ \text{eV} / n^2$

Relativistic KEM formula must
always remove invariant rest Matter PRIOR
to Lorentz correction being applied

$$\text{KEM} = \text{RE} - mc^2$$

As rest Matter is invariant and the
Lorentz factor applies to EM waves only

$$\text{KEM} = 4\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Electric & Magnetic energies
are contained in the KEM field geometry
extending from all Matter topologies in motion

$$p/h = \tilde{\nu}$$

Wave numbers
are the inverse of
Wavelengths

$$h/p = \lambda$$

Frequency and Wavelength
are inversely related
via the Speed of Light

$$m\Omega v^2/h = f$$

$$2m\Omega v = hf$$

$$Mv^2 = \text{KEM} = hv^2$$

$$\frac{v^2}{c^2} \qquad \frac{v^2}{c^2}$$

hv
Magnetic moment
μ_0
ε_0
c^2
v^2
v
EM Field
Planck
quanta
ElectroMagnetic
mass
velocity
n8
13.525 eV
n7
3.381 eV
Un-named
n6
1.502 eV
Humphreys
n5
0.845 eV
Pfund
n4
0.541eV
Brackett
n3
0.375 eV
Paschen
n2
0.276 eV
Balmer
n1
0.211 eV
Lyman
12
n0
0 eV
Rest Matter
12
RE
hf
0 eV
13.6 eV
hv
v^2
KE
p
$Me v^2$
[12π m]
n
12
48
108
192
300
432
588
768

Copyright ABRAHAM [2008] - all rights reserved

Photonic energy

momenta of KEM fields

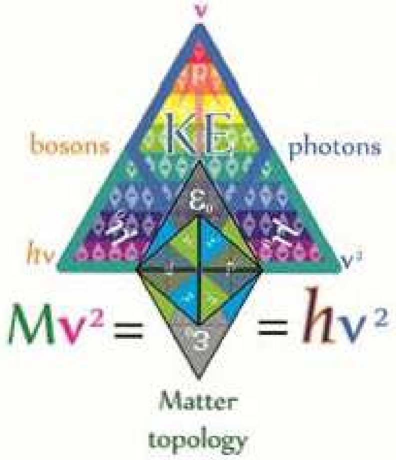
ν
KE
bosons
photons
ε₀
hν
ν²
Mv² =
= hν²
3
Matter
topology

Rydberg's formula is a wavenumber equivalent re-expression of Lorentz's velocity correction formula as it applies to the quantised energy momentum of Photo-electrons bound in atomic nuclei

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

The difference of two inverse squared KEM energies

Rydberg | Tetryonic

$$R\left(\frac{1}{1} - \frac{1}{4}\right) = R\left(36 \diagup 48\right)$$

.75 | .75

n1 and n2 are integers greater than or equal to 1 such that n1 < n2

$$\frac{1}{2}Mv^2 = KE = hcR$$

Whereas Rydberg calculated for wavenumbers using the difference of two inverse squared KEM energies.
(Inverse wavelengths)

$$\left(\frac{1}{1} - \frac{1}{4}\right)$$

.75

A mathematical solution that works but whose physical mechanics has eluded explanation

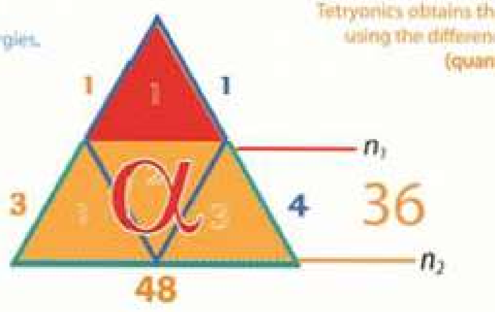
1
1
1
3
3
3
4
α
n₁
n₂
36
48

Tetryonics obtains the same answers geometrically using the difference of the two KEM energies.
(quanta differentials)

(36/48)

.75

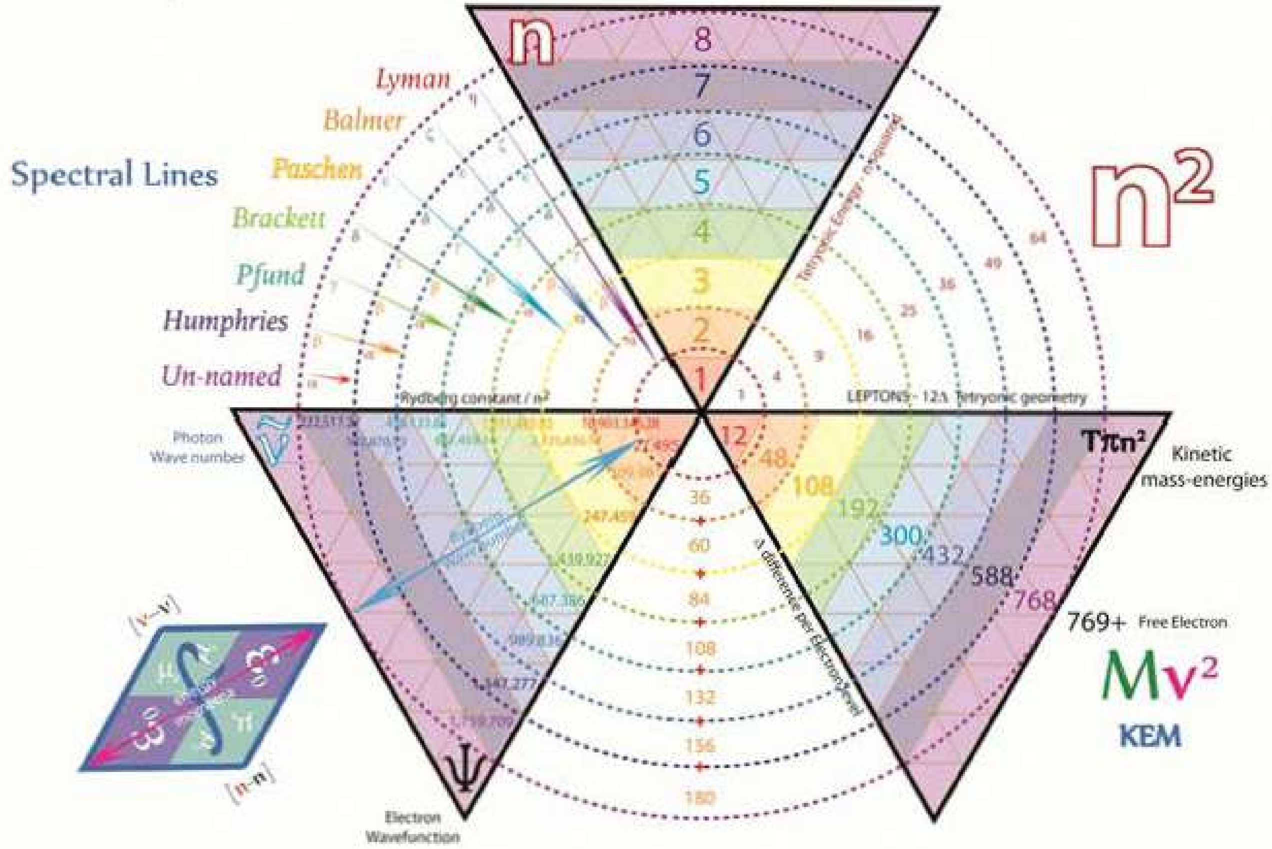
Spectral Lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
n
8
7
6
5
4
3
2
1
n²
Rydberg constant / n²
LEPTONS - 12∆ Tetryonic geometry
Photon
Wave number
ν̃
Tπn²
Kinetic
mass-energies
12
48
108
192
300
432
588
768
769+ Free Electron
Mv²
KEM
36
60
84
108
132
156
180
Δ difference per electron level
[ν—ν]
[n-n]
Ψ
Electron
Wavefunction

Any changes to the Energy momenta of the electron's KEM field must be reflected in the Energy momenta of the Photo-electrons emitted [absorbed]

$$\pm\Delta h\nu = KEM_i - KEM_f$$

The difference of two KEM energies

Tetryonics 28.07 - Photonic energy momenta of KEM fields

Kinetic Energies of Photoelectrons

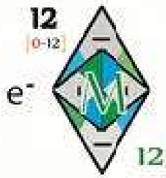

rest Matter

Kinetic energies are DIVERGENT from Matter topologies

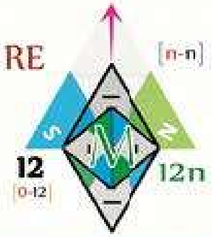
RE
[n-n]
12
[0-12]
12n

Relativistic mass-energies

rest Matter + Kinetic Energies

The total energy of a photon is absorbed into the electron's KEM field geometry

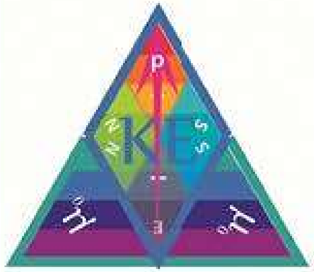

All velocity changes produce changes to mass-energies, momenta and wavefunction relationships

$$KEM = Mv^2$$

rest Matter is velocity invariant

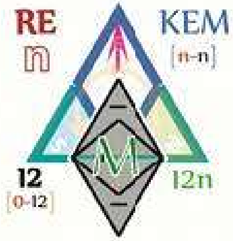
RE
n
KEM
[n-n]
12
[0-12]
12n

$1e19v = n$

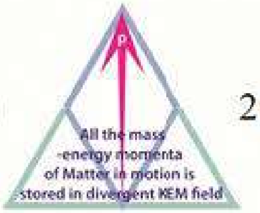

All the mass-energy momenta of Matter in motion is stored in divergent KEM field

$2v = f$

6.62943 e -34 J.s

13.5252 eV
2.1669 e -18 J

$$E = hv^2$$

3.26874 e15 quanta

KEM fields are velocity variant

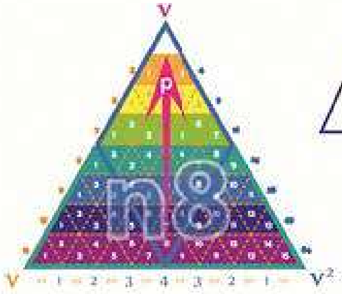
v
P
n8
v 1 2 3 4 3 2 1 v²

Δp

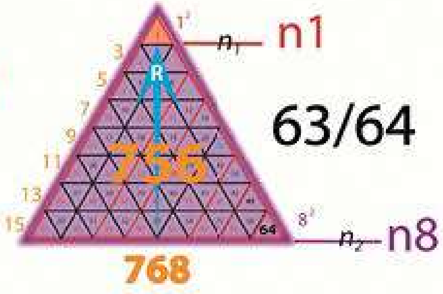
1²
n₁ n1
3
5
7
9
11
13
15
R
756
64
8²
n₂ n8
768

63/64

EM Field   Planck quanta

$$4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

KEM field   ElectroMagnetic   mass   velocity

Electron KEM Quantum levels

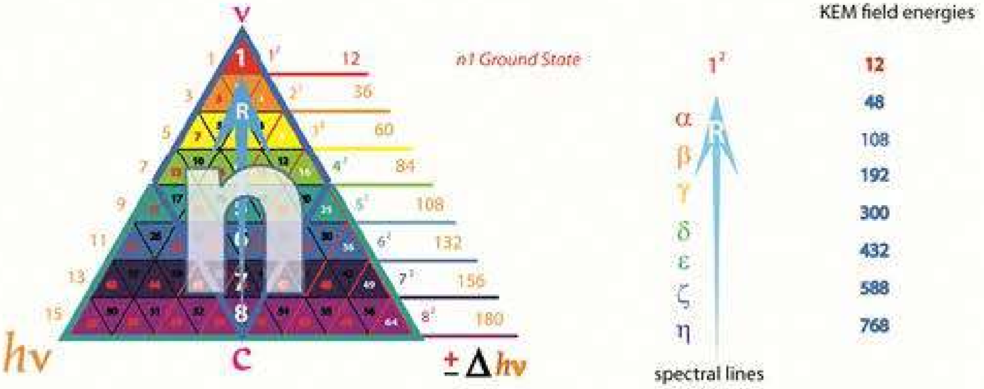
v
1 1² 12
3 2² 36
5 3² 60
7 4² 84
9 5² 108
11 6² 132
13 7² 156
15 8² 180
n1 Ground State
R
n
1
7
8
hv
c
± Δ hv
KEM field energies
1²
α
β
γ
δ
ε
ζ
η
R
12
48
108
192
300
432
588
768
spectral lines

Rydberg spectral transition jumps

Rydberg   Tetryonic

$$R_H\left(\frac{1}{1}-\frac{1}{64}\right) = \frac{1}{\lambda} = KEM\left(\frac{63 \diagup 64}{1}\right)$$

.984375   .984375

Spectral line energies

EM Field Planck quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Charge ElectroMagnetic mass velocity

ZPF

ZPF

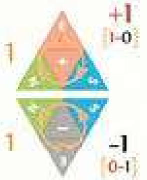

Each quantum level transition requires kinetic energies with an ODD number of EM mass-energy momenta quanta [bosons - charge carriers]

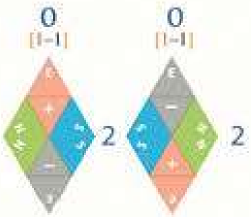

γ

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

n.*hv*

Bosons are

Odd number Charge carriers

In order to transition between levels for each emission and absorption leptons must release or absorb specific W bosons (neutral Z bosons - photons) for each KEM field change

Photons are

Even number EM force carriers

n.*hf*

EM Field Planck quanta

$$\text{ODD}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons ElectroMagnetic mass velocity

W+

W-

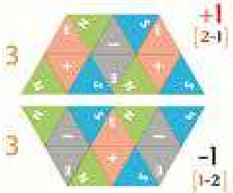

Pairs of charged W bosons combine to form neutral bosons / photons [EM induction]

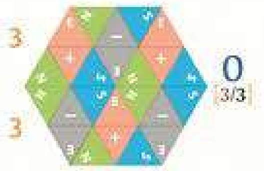

W^0

EM Field Planck quanta

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM waves ElectroMagnetic mass velocity

KEM field Differentials

Bosonic quantum level changes

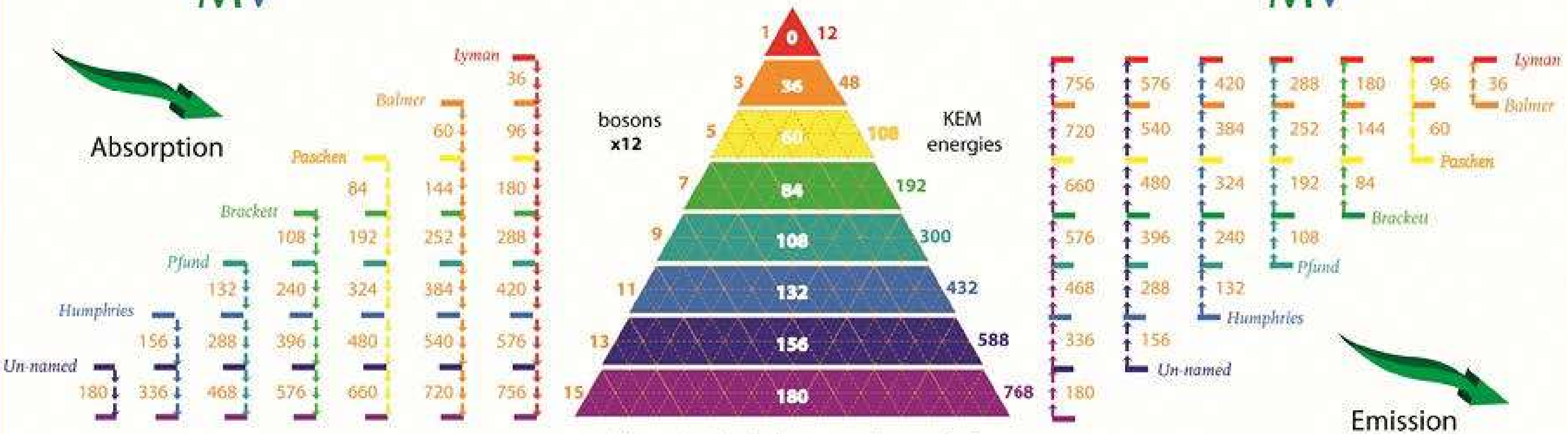
Mv²
Absorption
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
36
60
96
84
144
180
108
192
252
288
132
240
324
384
420
156
288
396
480
540
576
180
336
468
576
660
720
756
bosons
x12
1
0
12
3
36
48
5
60
108
7
84
192
9
108
300
11
132
432
13
156
588
15
180
768
KEM
energies
Mv²
756
720
660
576
468
336
180
576
540
480
396
288
156
420
384
324
240
132
288
252
192
108
180
144
84
96
60
36
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
Emission

All quantum level changes are the result of boson increases or decreases

Copyright ABRAHAM [2008] - all rights reserved

Quantum level differentials

The exact number of Kinetic energy quanta (ZPEs)
required for each bound electron transition within a nucleus

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg calculated this as the wavenumber differerential of two inverse squared energies

$$n_f^2 \qquad \overset{\text{Rydberg}}{\underset{.75}{R_H\left(\frac{1}{1} - \frac{1}{4}\right)}} = \overset{\text{Tetryonic}}{\underset{.75}{R_H\left(36 \diagup 48\right)}} \qquad n_i^2$$

Tetryonic geometry reveals it to be sum of ODD integer bosons between two squared KEM energies

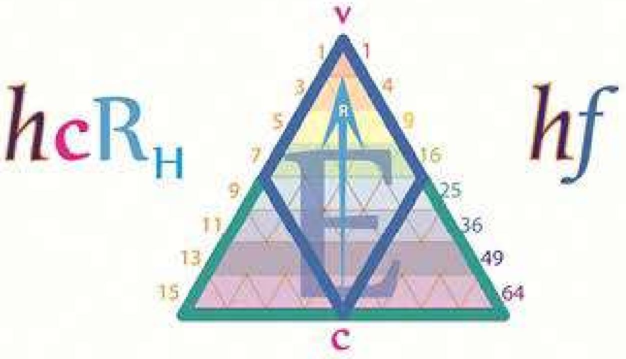
hcR_H
hf
v
c
R
E
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64

Rydberg's constant is a LINEAR measurement of wavenumbers
which can be related to squared KEM energies through
the velocity-frequency-wavelength relationship

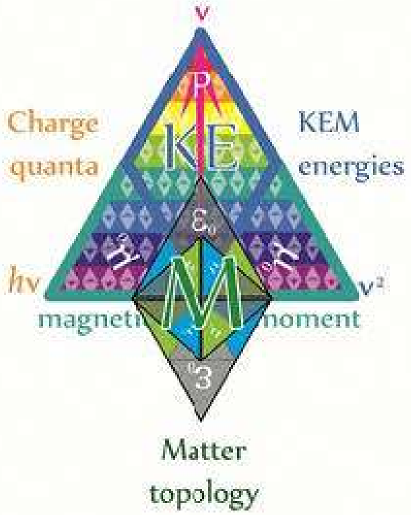
v
Charge
quanta
KEM
energies
hv
v²
magnetic moment
R
KE
ε₀
μ₀
M
3
Matter
topology

Each Quantum level jump within
an Atomic Nuclei is discrete
and requires an exact number
of EM mass-energy quanta

The calculation of the quanta
required is historically done
utilising Rydberg's Constant and
its associated formula

The quantum level [boson] differential
equates it to the release or absorption of
consecutive ODD number Planck quanta
from squared KEM field energy geometries
Δ[hv]

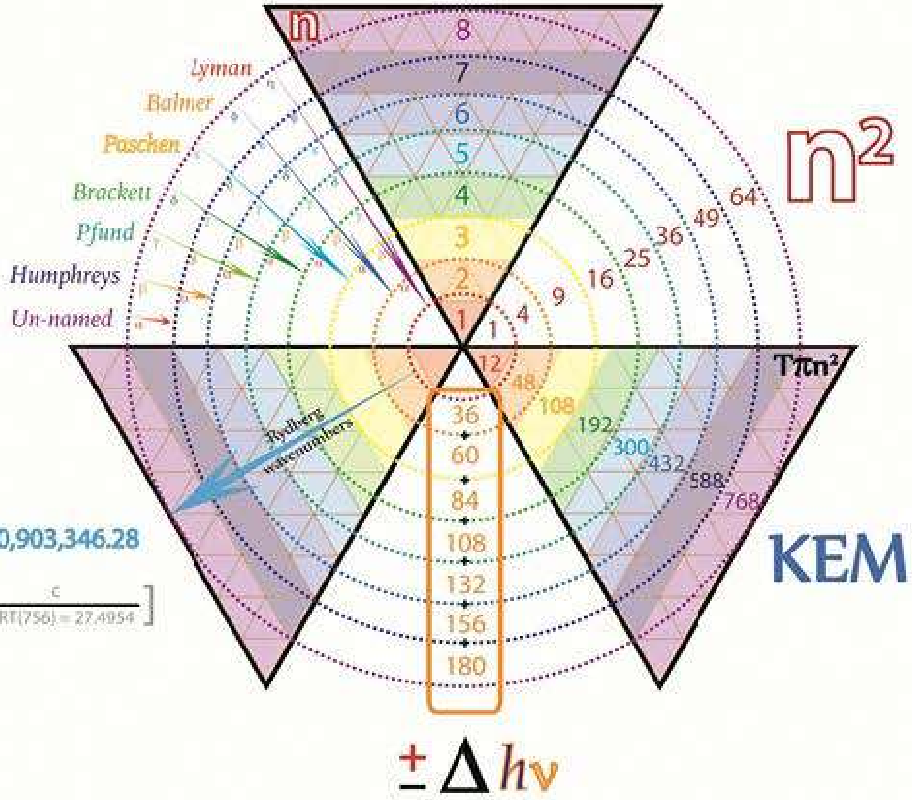
n
8
7
6
5
4
3
2
1
Lyman
Balmer
Paschen
Brackett
Pfund
Humphreys
Un-named
n²
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768
Tπn²
36
60
84
108
132
156
180
Rydberg wavenumbers
KEM
±Δhν

R_H = 10,903,346.28

$\left[\frac{c}{\text{SQRT}(756) = 27.4954}\right]$

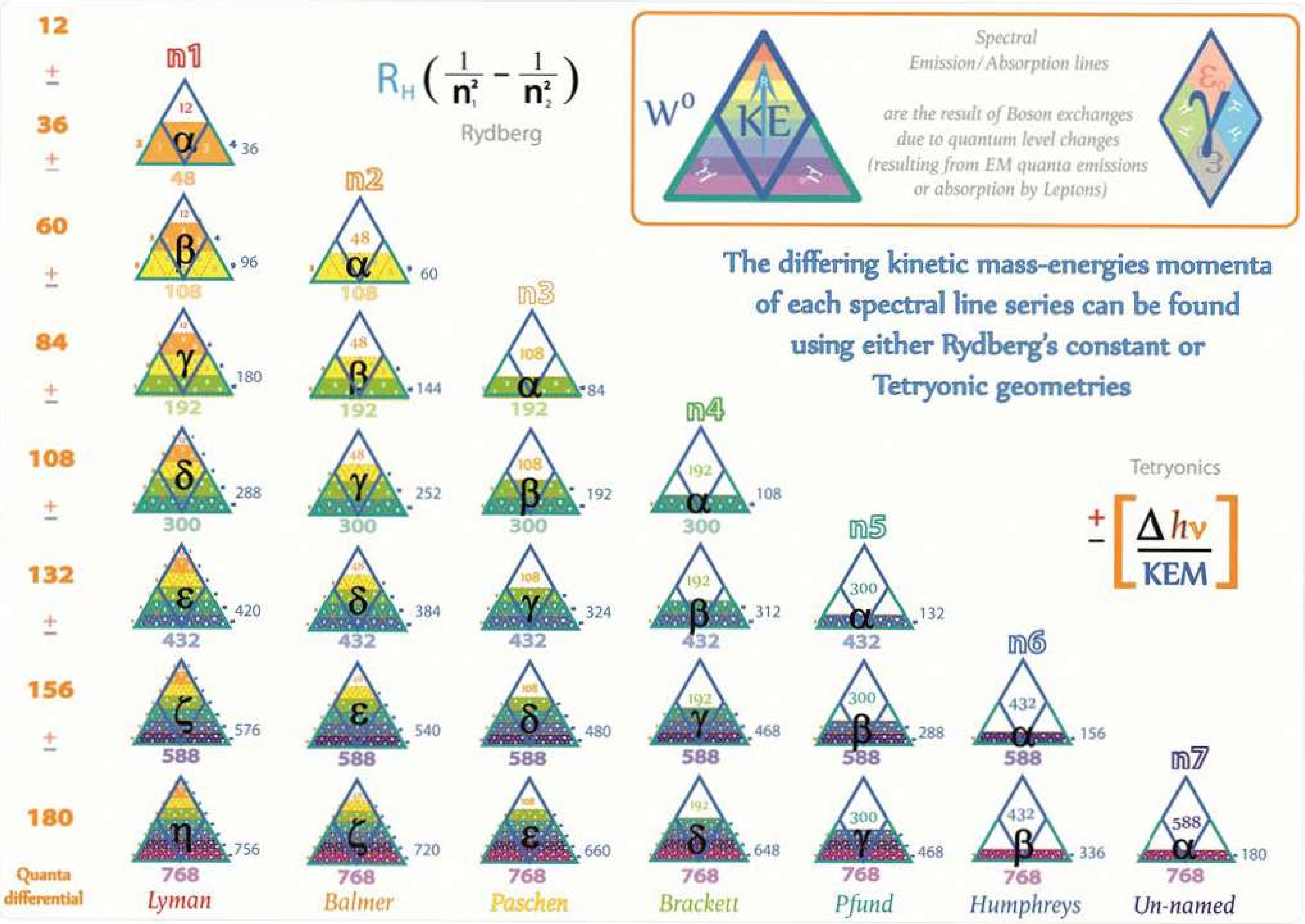

Spectral Emission/Absorption lines are the result of Boson exchanges due to quantum level changes (resulting from EM quanta emissions or absorption by Leptons)

The differing kinetic mass-energies momenta of each spectral line series can be found using either Rydberg's constant or Tetryonic geometries

12
36
60
84
108
132
156
180
Quanta differential
$R_H\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)$
Rydberg
W⁰
KE
ε₀
γ
μ
Tetryonics
$\pm\left[\frac{\Delta h\nu}{KEM}\right]$
n1
n2
n3
n4
n5
n6
n7
α
β
γ
δ
ε
ζ
η
36
96
180
288
420
576
756
60
144
252
384
540
720
84
192
324
480
660
108
312
468
648
132
288
468
156
336
180
48
108
192
300
432
588
768
Lyman
Balmer
Paschen
Brackett
Pfund
Humphreys
Un-named

Tetryonics 28.11 - Boson differentials

Absorption~Emission lines

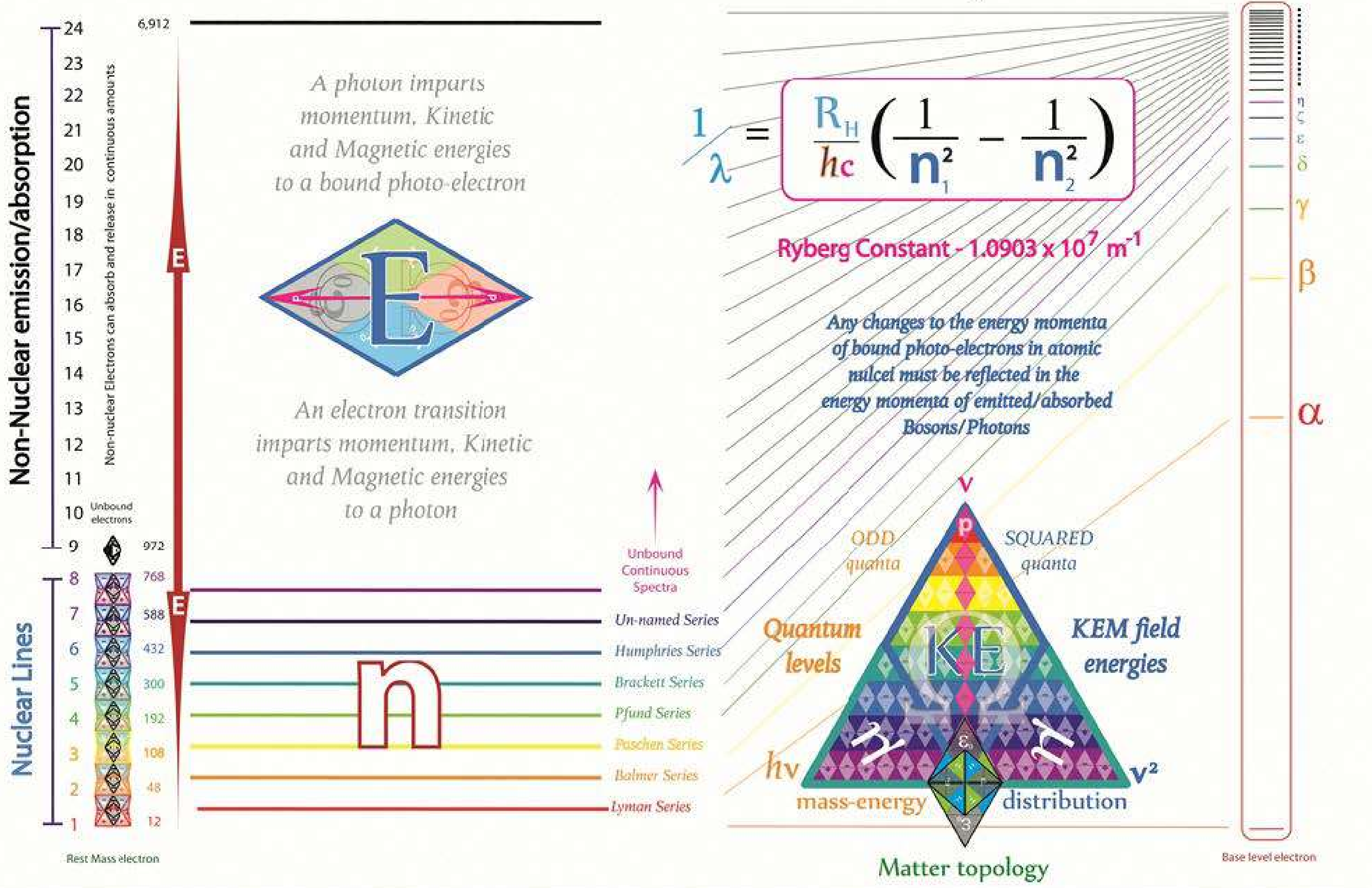

$$E = hf = \frac{hc}{\lambda} = mc^2$$

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg Constant - $1.0903 \times 10^7 \ m^{-1}$

A photon imparts momentum, Kinetic and Magnetic energies to a bound photo-electron

An electron transition imparts momentum, Kinetic and Magnetic energies to a photon

Any changes to the energy momenta of bound photo-electrons in atomic nulcei must be reflected in the energy momenta of emitted/absorbed Bosons/Photons

Spectral line series
Non-Nuclear emission/absorption
Non-nuclear Electrons can absorb and release in continuous amounts
Nuclear Lines
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
6,912
Unbound electrons
972
768
588
432
300
192
108
48
12
Rest Mass electron
E
E
n
Unbound Continuous Spectra
Un-named Series
Humphries Series
Brackett Series
Pfund Series
Paschen Series
Balmer Series
Lyman Series
η
ζ
ε
δ
γ
β
α
Base level electron
v
p
ODD quanta
SQUARED quanta
Quantum levels
KEM field energies
KE
hv
v²
mass-energy
distribution
Matter topology

Copyright ABRAHAM [2008] - all rights reserved

Spectral Lines

Spectral lines are the result of interaction between a quantum system (usually atoms, but sometimes molecules or atomic nuclei) and a single photon. When a photon has exactly the right amount of energy to allow a change in the energy state of the system (in the case of an atom this is usually an electron changing orbitals), the photon is absorbed. Then it will be spontaneously re-emitted, either in the same frequency as the original or in a cascade, where the sum of the energies of the photons emitted will be equal to the energy of the one absorbed (assuming the system returns to its original state). The direction and polarization of the new photons will, in general, correlate with those of the original photon.

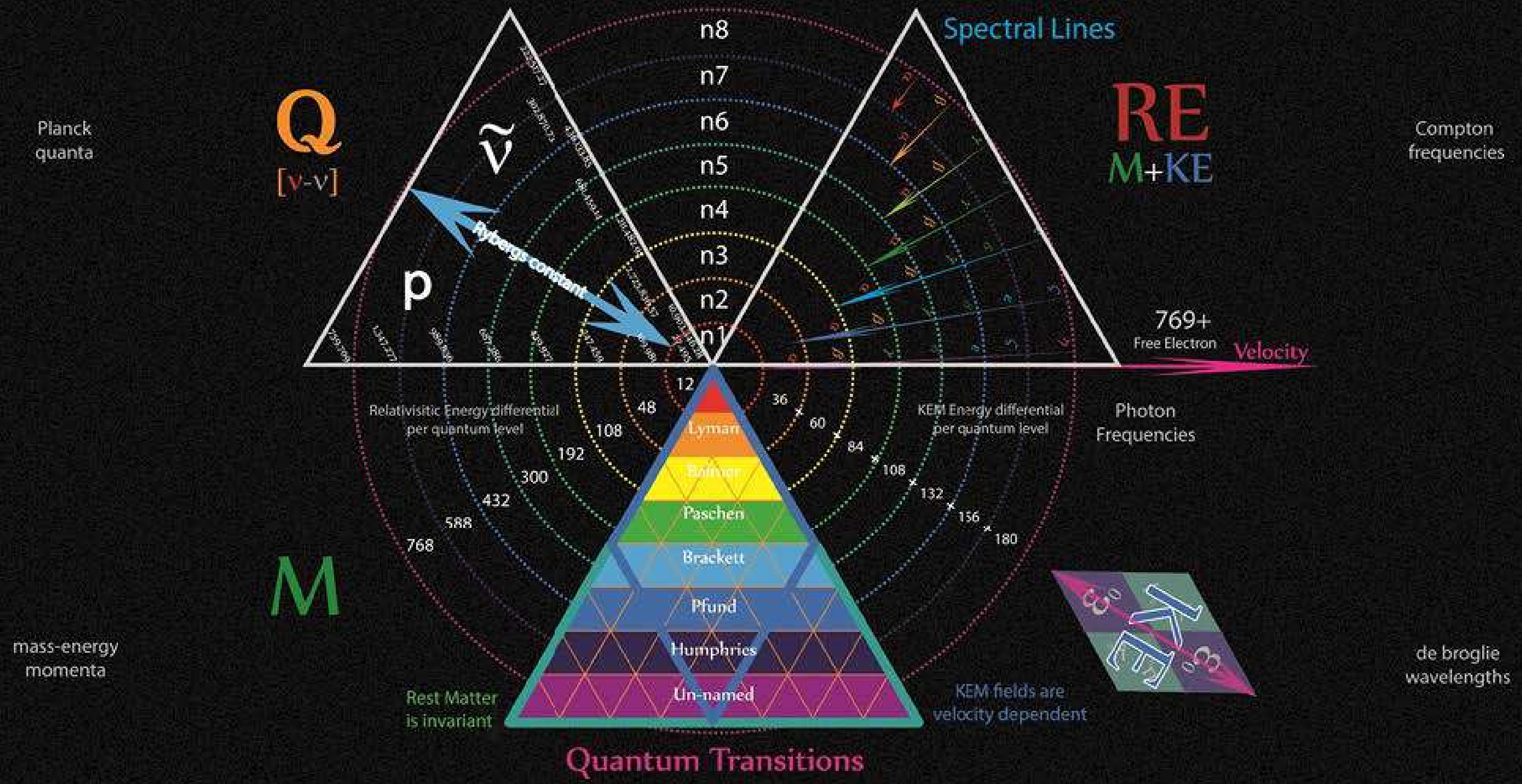
Planck
quanta
Q
[ν-ν]
ν̃
p
Rybergs constant
n8
n7
n6
n5
n4
n3
n2
n1
Spectral Lines
RE
M+KE
Compton
frequencies
769+
Free Electron
Velocity
12
48
108
192
300
432
588
768
36
60
84
108
132
156
180
Relativisitic Energy differential
per quantum level
KEM Energy differential
per quantum level
Photon
Frequencies
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
M
mass-energy
momenta
Rest Matter
is invariant
KEM fields are
velocity dependent
KE
c²
de broglie
wavelengths
Quantum Transitions

Spectral lines are highly atom-specific, and can be used to identify the chemical composition of any medium capable of letting light pass through it (typically gas is used). Several elements were discovered by spectroscopic means, such as Helium [04], Rubidum [37], Cerium [58] and Thallium [81]

Spectral lines also depend on the physical conditions of the gas, so they are widely used to determine the chemical composition of stars and other celestial bodies that cannot be analyzed by other means, as well as many other of their physical properties.

Tetryonics 29.02 - Spectral line production

Copyright ABRAHAM [2008] - all rights reserved

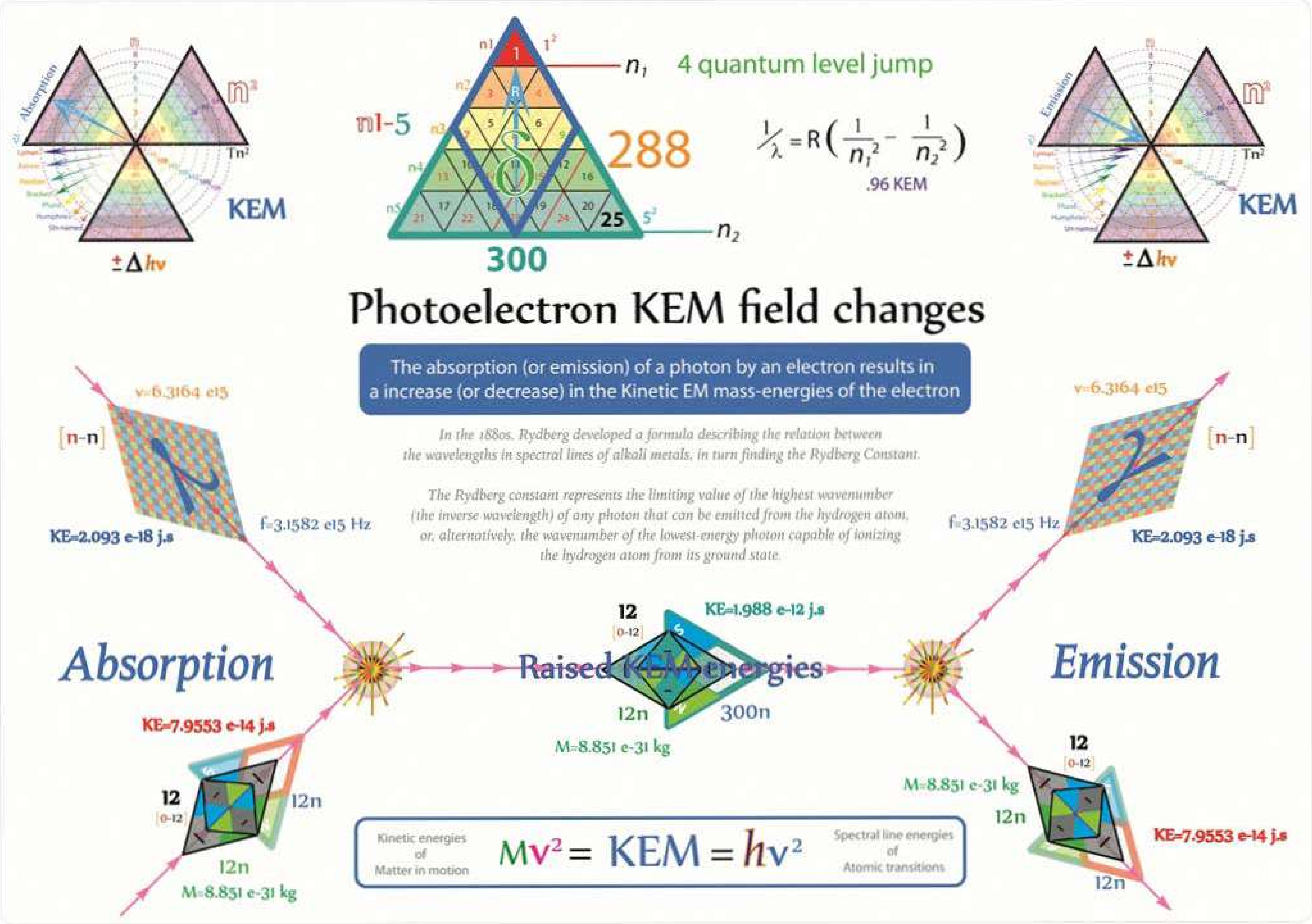
Absorption
m²
Tn²
KEM
±Δhv
n1
1²
1
n2
nl-5
n3
n4
n5
n₁
4 quantum level jump
288
25
5²
n₂
300
Emission
m²
Tn²
KEM
±Δhv
v=6.3164 e15
[n-n]
KE=2.093 e-18 j.s
f=3.1582 e15 Hz
Absorption
12
[0-12]
KE=1.988 e-12 j.s
Raised KEM energies
12n
300n
M=8.851 e-31 kg
Emission
KE=7.9553 e-14 j.s
12
[0-12]
12n
12n
M=8.851 e-31 kg
v=6.3164 e15
[n-n]
f=3.1582 e15 Hz
KE=2.093 e-18 j.s
M=8.851 e-31 kg
12
[0-12]
12n
KE=7.9553 e-14 j.s
12n

$$\frac{1}{\lambda} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

.96 KEM

Photoelectron KEM field changes

The absorption (or emission) of a photon by an electron results in a increase (or decrease) in the Kinetic EM mass-energies of the electron

In the 1880s, Rydberg developed a formula describing the relation between the wavelengths in spectral lines of alkali metals, in turn finding the Rydberg Constant.

The Rydberg constant represents the limiting value of the highest wavenumber (the inverse wavelength) of any photon that can be emitted from the hydrogen atom, or, alternatively, the wavenumber of the lowest-energy photon capable of ionizing the hydrogen atom from its ground state.

Kinetic energies of Matter in motion $Mv^2 = KEM = hv^2$ Spectral line energies of Atomic transitions

Tetryonics 29.03 - Photoelectron KEM changes

Copyright ABRAHAM [2008] - all rights reserved

Revealing Rydberg Formula's geometry

The Rydberg formula is used in atomic physics to determine the wavenumber of spectral lines of many chemical elements.

$$\frac{h}{Mv} = \lambda$$

Wavelength is inversely related to Frequency

$$1/\lambda = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

$$p/_h = \tilde{\nu}$$

Wavenumbers are a inverse measure of Wavelengths

$$p^2 = E = Mv^2$$

Tetryonic geometry can be applied to Kinetic EM field variations produced by the emission and absorption of Photons by electrons in Nuclear orbits to reveal the geometry behind Rydberg's formula

$$p = \frac{E}{v} = h\nu$$

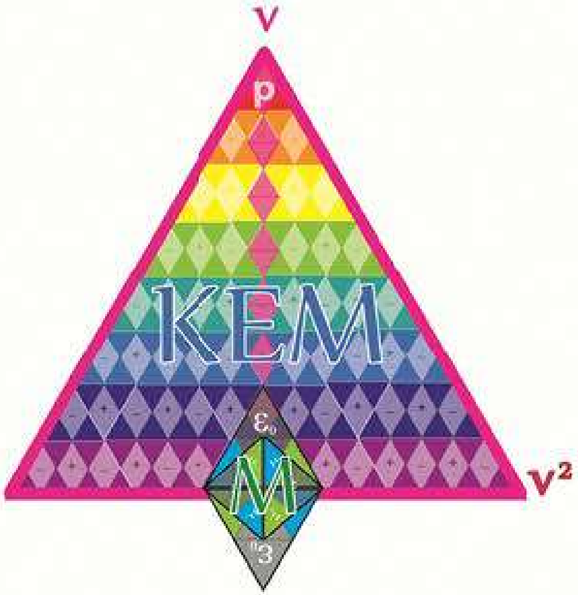

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM Field

Planck quanta

Photoelectron KEM fields

ElectroMagnetic

mass

velocity

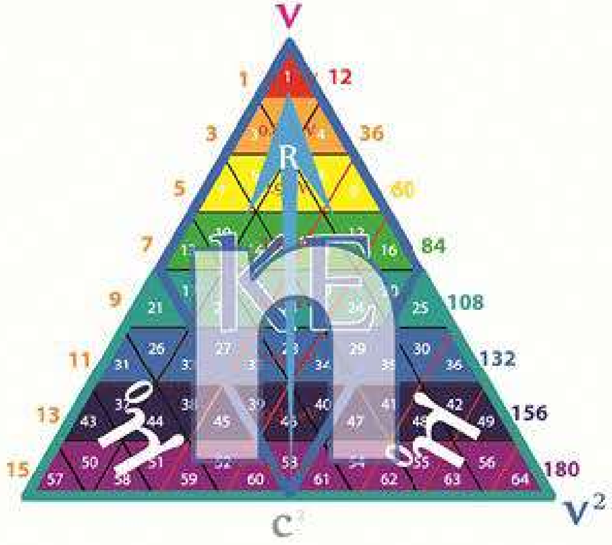

13.605 eV

$$\boxed{Mv = \frac{KEM}{v} = h\nu}$$

linear momentum

longitudinal quanta

Wavelengths and Frequency are related through the velocity of propagation

Compton frequency

$$\frac{Mv^2}{h} = f$$

Spectral line emissions reveals Rydberg's formula to be a measure of longitudinal KEM momenta

$$KEM = hcR$$

$$\lambda_{KEM} = \frac{f}{v}$$

Hydrogen spectral lines

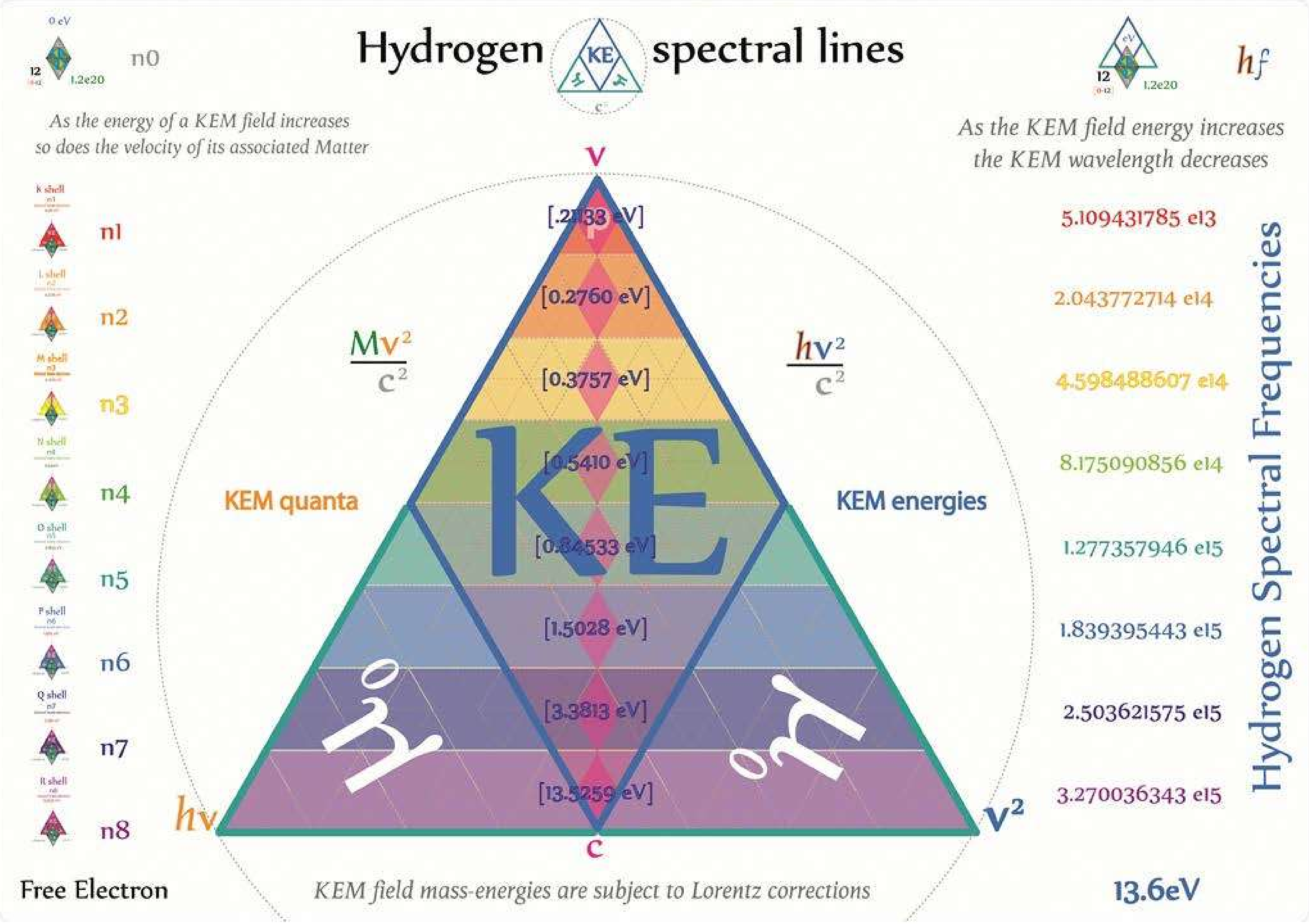

As the energy of a KEM field increases so does the velocity of its associated Matter

As the KEM field energy increases the KEM wavelength decreases

KEM quanta

KEM energies

n1	[.2133 eV]	5.109431785 e13
n2	[0.2760 eV]	2.043772714 e14
n3	[0.3757 eV]	4.598488607 e14
n4	[0.5410 eV]	8.175090856 e14
n5	[0.84533 eV]	1.277357946 e15
n6	[1.5028 eV]	1.839395443 e15
n7	[3.3813 eV]	2.503621575 e15
n8	[13.5259 eV]	3.270036343 e15

Hydrogen Spectral Frequencies

Free Electron

KEM field mass-energies are subject to Lorentz corrections

13.6eV

Tetryonics 29.05 - Hydrogen Spectral Frequencies

Copyright ABRAHAM [2008] - all rights reserved

Absorption lines

Increase in
photo-electronic
energies

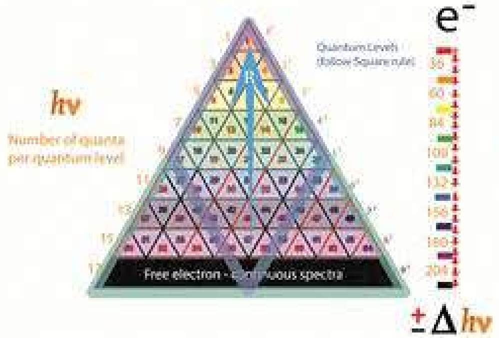
hv
Number of quanta
per quantum level
Quantum Levels
(follow Square rule)
e-
36
60
84
108
132
156
180
204
Free electron - continuous spectra
±Δhv

$$\Delta 2hv = E = hf$$

When an electron 'absorbs' a photon of energy momenta it 'jumps' from one energy level to another, dependent on the energy and frequency of the incident photon

$$f = R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

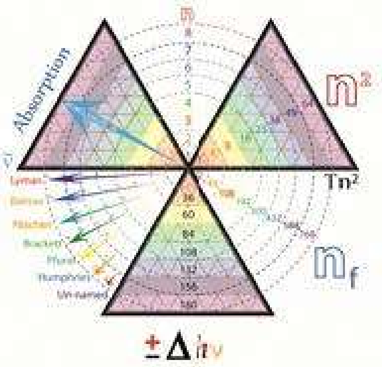
Absorption
n²
Tn²
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
36
60
84
108
132
156
180
nf
±Δhv

The quantum level of the nuclei determines the ground quantum level of electrons within the nucleus
All transitions within the nulceus are discrete quantum jumps - outside the nucleus all spectra are continuous

KEM A Hydrogen photo-electron can only exist in specific energy levels below 13.525eV within the nucleus (any excess KE results in a unbound electron with KE - velocity)

KE in excess of 13.525 eV results in unbound electron

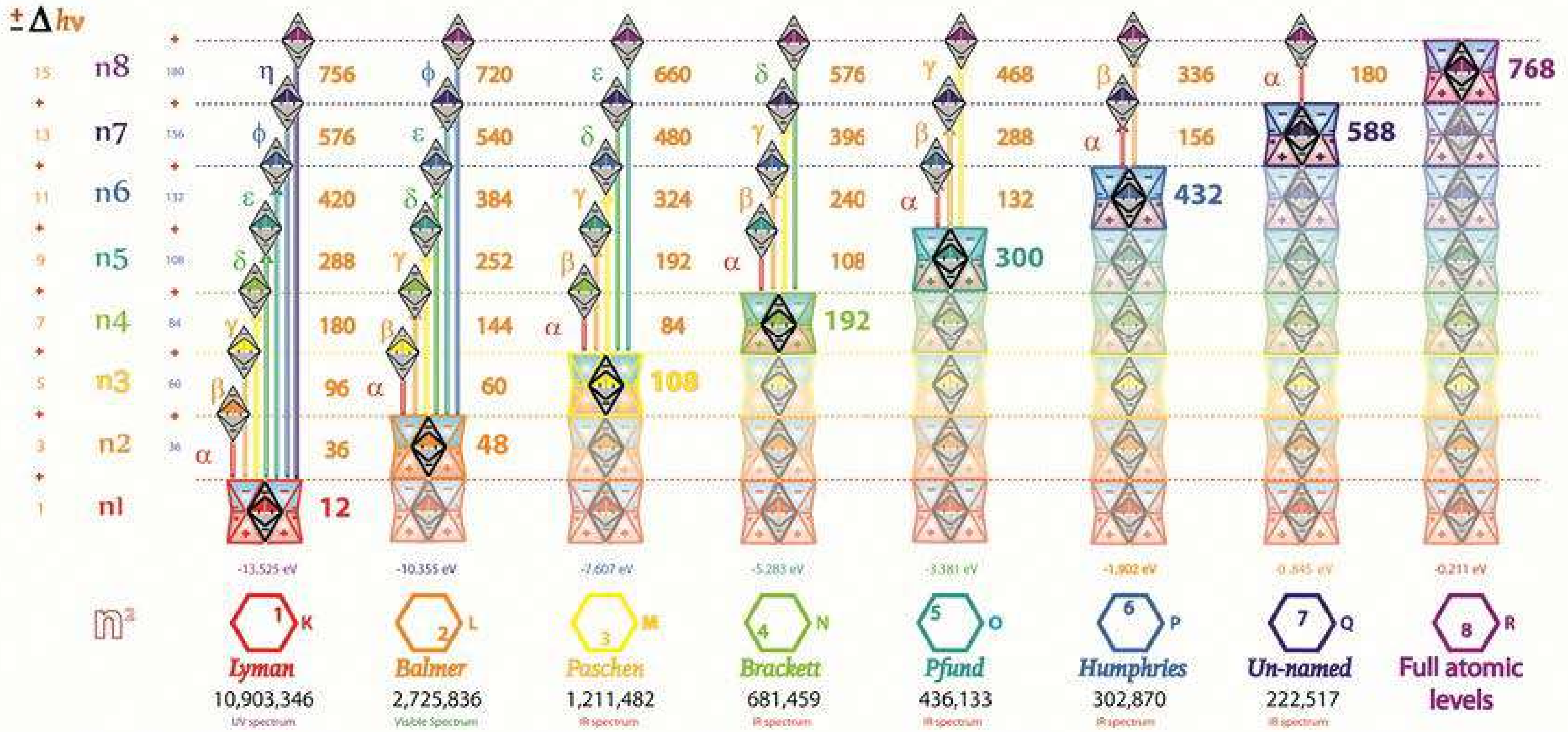
±Δhv
15
n8
180
13
n7
156
11
n6
132
9
n5
108
7
n4
84
5
n3
60
3
n2
36
1
n1
η
756
φ
576
ε
420
δ
288
γ
180
β
96
α
36
12
φ
720
ε
540
δ
384
γ
252
β
144
α
60
48
ε
660
δ
480
γ
324
β
192
α
84
108
δ
576
γ
396
β
240
α
108
192
γ
468
β
288
α
132
300
β
336
α
156
432
α
180
588
768
-13,525 eV
-10,355 eV
-7,607 eV
-5,283 eV
-3,381 eV
-1,902 eV
-0,845 eV
-0,211 eV
n²
1 K
2 L
3 M
4 N
5 O
6 P
7 Q
8 R

Lyman	Balmer	Paschen	Brackett	Pfund	Humphries	Un-named	Full atomic levels
10,903,346	2,725,836	1,211,482	681,459	436,133	302,870	222,517	
UV spectrum	Visible Spectrum	IR spectrum	IR spectrum	IR spectrum	IR spectrum	IR spectrum	

KEM fields & quantum jumps

$$Mv^2 = KEM = hv^2$$

$$f = R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

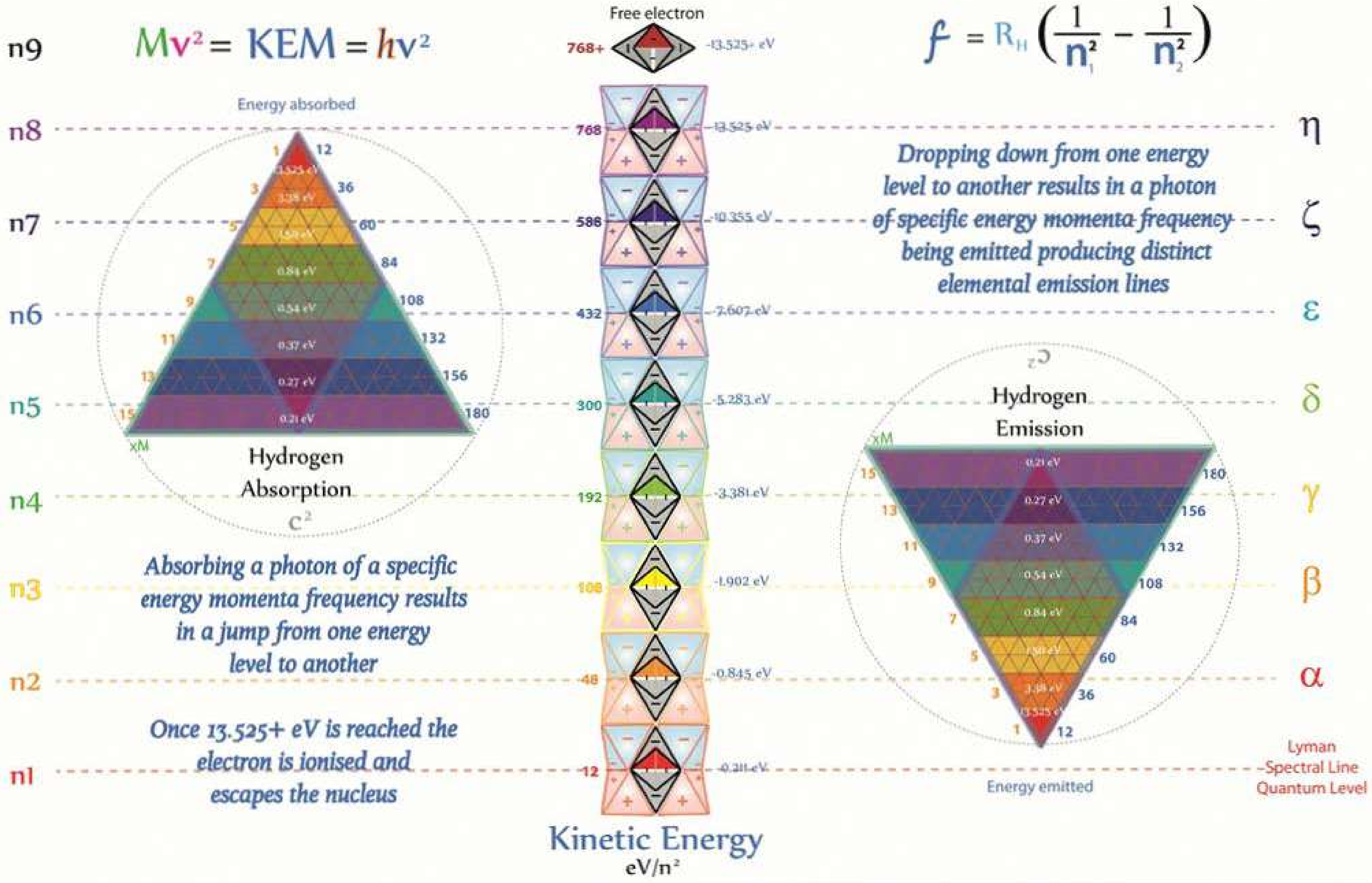
Free electron
768+
-13.525+ eV
n9
n8
n7
n6
n5
n4
n3
n2
n1
Energy absorbed
1
12
13.525 eV
3
36
3.38 eV
5
60
1.50 eV
7
84
0.84 eV
9
108
0.54 eV
11
132
0.37 eV
13
156
0.27 eV
15
180
0.21 eV
xM
Hydrogen Absorption
c²
768
-13.525 eV
588
-10.355 eV
432
-7.607 eV
300
-5.283 eV
192
-3.381 eV
108
-1.902 eV
48
-0.845 eV
12
-0.211 eV
η
ζ
ε
δ
γ
β
α
c²
Hydrogen Emission
xM
Energy emitted
Lyman Spectral Line Quantum Level
Kinetic Energy
eV/n²

Dropping down from one energy level to another results in a photon of specific energy momenta frequency being emitted producing distinct elemental emission lines

Absorbing a photon of a specific energy momenta frequency results in a jump from one energy level to another

Once 13.525+ eV is reached the electron is ionised and escapes the nucleus

Tetryonics 29.07 - KEM field & quantum jumps

Emission lines

Decrease in photo-electronic energies

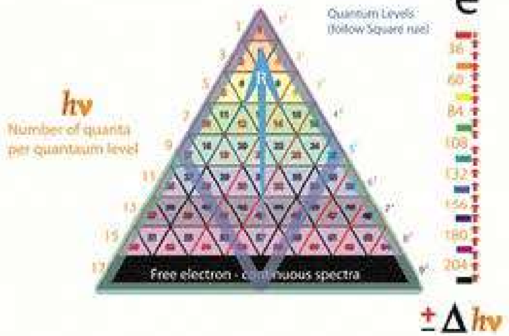
hv
Number of quanta per quantum level
Quantum Levels (follow Square rule)
e⁻
36
60
84
108
132
156
180
204
1
3
5
7
9
11
13
15
17
Free electron - continuous spectra
±Δhv

$$\Delta 2hv = E = hf$$

When an electron 'emits' a photon of energy momenta it 'drops' from one energy level to another dependent on the energy and frequency of the ejected photon

$$f = R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

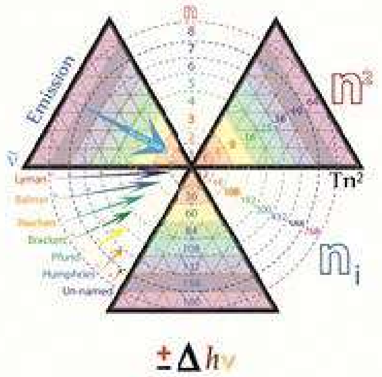
Emission
n
n²
Tn²
nᵢ
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
±Δhv

The quantum level of the nuclei determines the ground quantum level of electrons within the nucleus
All transitions within the nulceus are discrete quantum jumps - outside the nucleus all spectra are continuous

KEM

A Hydrogen electron can only exist in specific energy levels below 13.525eV within the nucleus (any excess KE must be shed as Photons)

KE in excess of 13.525 eV results in unbound electron

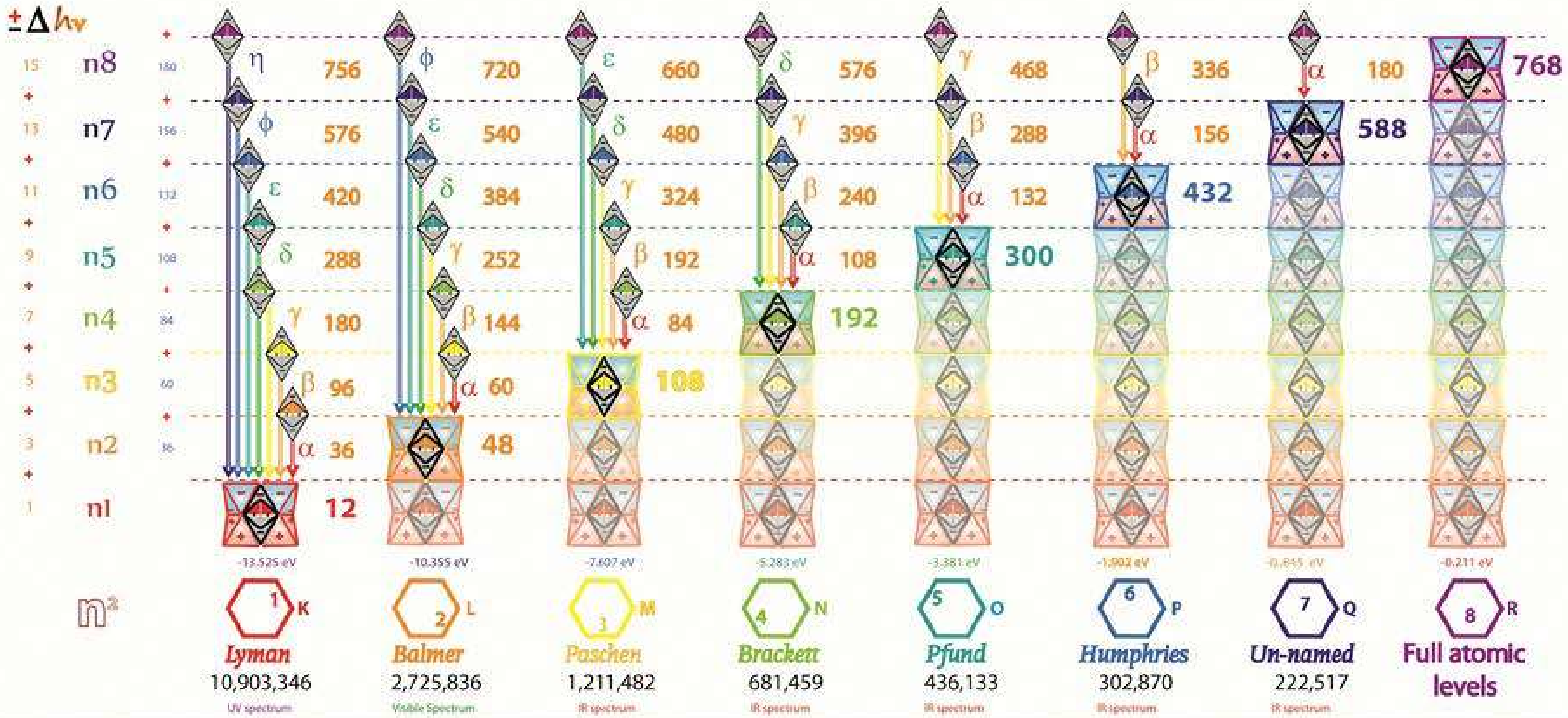
±Δhv
15 n8 180
13 n7 156
11 n6 132
9 n5 108
7 n4 84
5 n3 60
3 n2 36
1 n1
η 756
φ 576
ε 420
δ 288
γ 180
β 96
α 36
12
φ 720
ε 540
δ 384
γ 252
β 144
α 60
48
ε 660
δ 480
γ 324
β 192
α 84
108
δ 576
γ 396
β 240
α 108
192
γ 468
β 288
α 132
300
β 336
α 156
432
α 180
588
768
-13.525 eV
-10.355 eV
-7.607 eV
-5.283 eV
-3.381 eV
-1.902 eV
-0.845 eV
-0.211 eV
n²
1 K
2 L
3 M
4 N
5 O
6 P
7 Q
8 R
Lyman
10,903,346
UV spectrum
Balmer
2,725,836
Visible Spectrum
Paschen
1,211,482
IR spectrum
Brackett
681,459
IR spectrum
Pfund
436,133
IR spectrum
Humphries
302,870
IR spectrum
Un-named
222,517
IR spectrum
Full atomic levels

Spectral line series

(are produced by changes in linear momentum in photo-electric KEM fields due to atomic transitions of electrons between squared energy levels)

Lyman	27.49545417
Balmer	109.9818167
Paschen	247.4590875
Brackett	439.9272667
Pfund	687.3863542
Humphries	989.8363501
Un-named	1,347.277254

Square root of Energy required to transition nuclear quantum levels

$$\left[\sqrt{756}\right].n^2$$

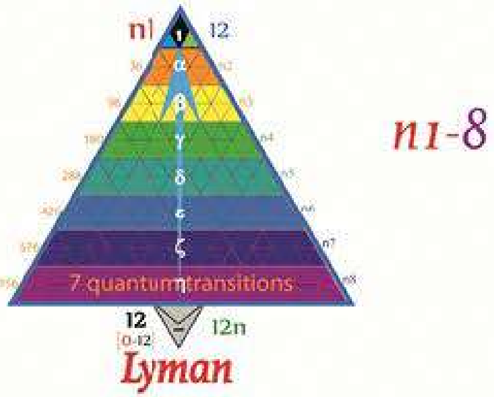
n1
12
1
α
β
γ
δ
ε
ζ
7 quantum transitions
12
0-12
12n
Lyman
n1-8

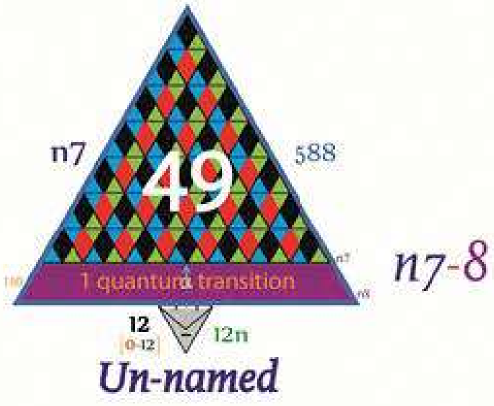
n7
49
588
1 quantum transition
12
0-12
12n
Un-named
n7-8

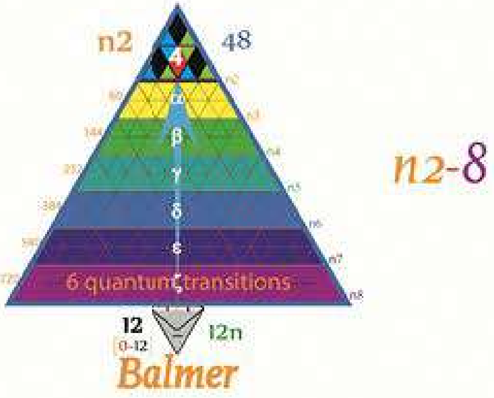
n2
4
48
α
β
γ
δ
ε
6 quantum transitions
12
0-12
12n
Balmer
n2-8

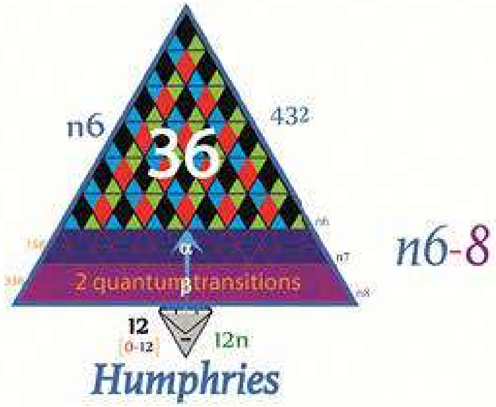
n6
36
432
α
2 quantum transitions
12
0-12
12n
Humphries
n6-8

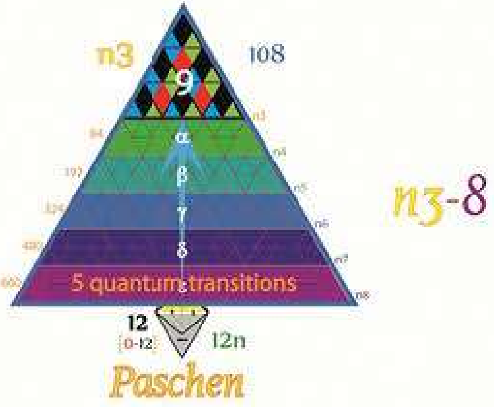
n3
9
108
α
β
γ
δ
5 quantum transitions
12
0-12
12n
Paschen
n3-8

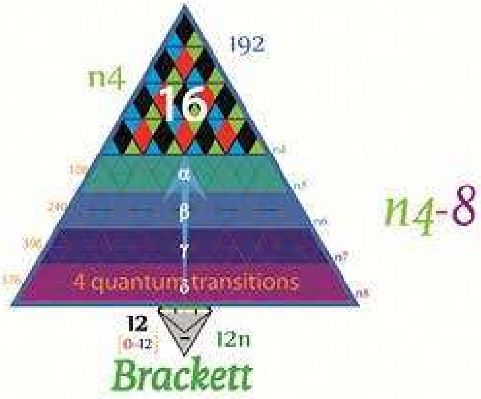
n4
16
192
α
β
γ
4 quantum transitions
12
0-12
12n
Brackett
n4-8

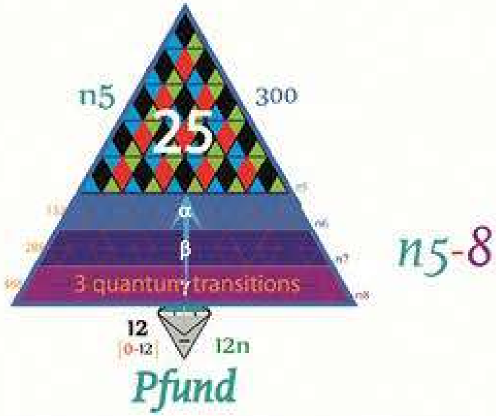
n5
25
300
α
β
3 quantum transitions
12
0-12
12n
Pfund
n5-8

Tetryonics 29.09 - Spectral line series

Johannes Robert Rydberg

(November 8, 1854 – December 28, 1919)

Rydberg's Constant

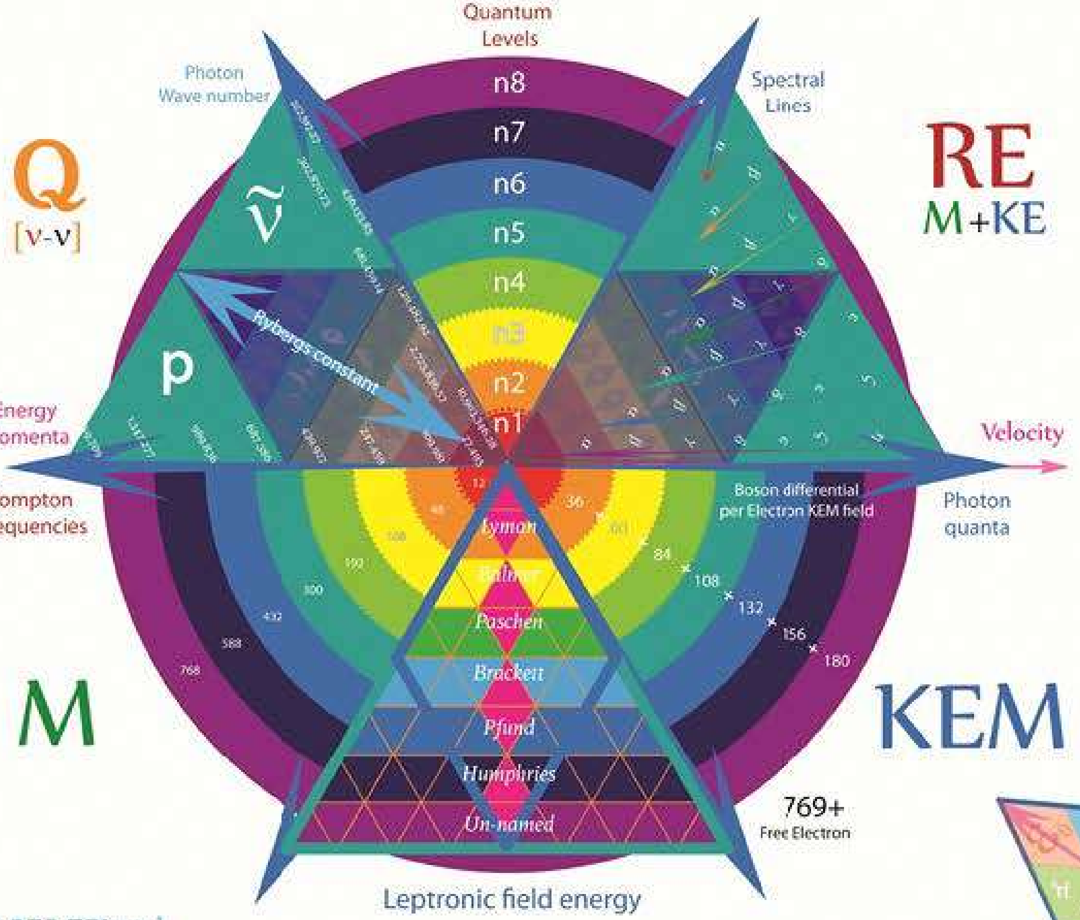
Quantum
Levels
Photon
Wave number
Spectral
Lines
Q
[ν-ν]
RE
M+KE
n8
n7
n6
n5
n4
n3
n2
n1
ν̃
p
Rydbergs constant
Energy
momenta
Compton
Frequencies
Velocity
Photon
quanta
Boson differential
per Electron KEM field
84
108
132
156
180
192
300
432
588
768
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
M
KEM
769+
Free Electron
Leptronic field energy

$$\tilde{\nu} = \frac{c}{27.49545}$$

$$f = \frac{c^2}{27.49545}$$

$$\lambda = \frac{27.49545}{c}$$

$$p = 27.49545$$

$$\lambda = v/f$$

Using the Tetryonic model of a photo-electron and its associated quantum KEM energy levels we see that Rydberg's constant is a linear measure of the photo-electronic KEM field - and in turn directly related to its scalar energy-momenta through 'c'

$$R_H = \frac{c}{27.49545}$$

$1.0903346.28 \times 10^7$ m-1

$$R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Currently accepted value - 10,973,731 m^{-1}

Rydberg's Constant can be shown to be the square root of the total number of quanta required to transition a n1 electron to n8
(the highest quantum level possible in elemental nuclei)
as per Planck's heat law

Tetryonic theory value - 10,903,346 m^{-1}

$$p^2 = KEM = Mv^2$$

Quantum Jump

The energy momenta of emitted photons is a function of the photo-electron's KEM wavefunction

27.49545

KE
Photons

$$\boxed{Mv^2 = \underset{13.525\ eV}{KEM} = hcR_H}$$

Copyright ABRAHAM [2008] - all rights reserved

Wavenumbers and Planck's Law

Wavenumbers are revealed to be the number of quanta [nhv] required for a photo-electron to transition between specific quantum levels divided by the speed of light [c]

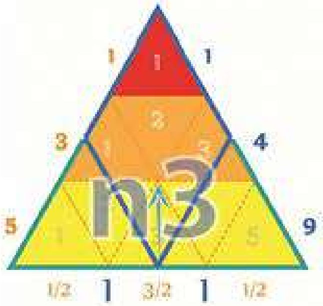

Balmer spectral lines [absorption]

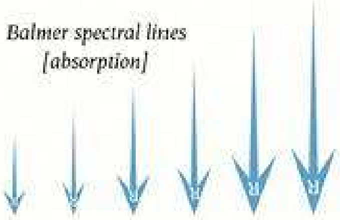

Max Karl Ernst Ludwig Planck

(April 23, 1858 – October 4, 1947)

Balmer spectral lines [emission]

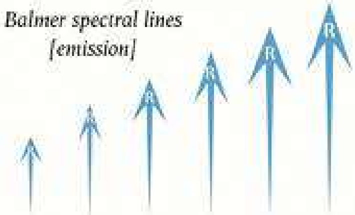

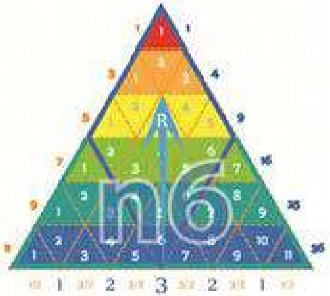

$$E = n[h\nu]$$

Wavenumbers are the number of Planck quanta required to reach a specific energy level

$$E = hc\tilde{\nu}$$

Photo-electrons emit and absorb photons to create blackbody radiation

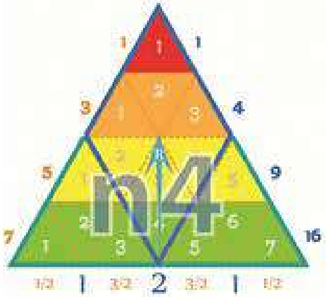

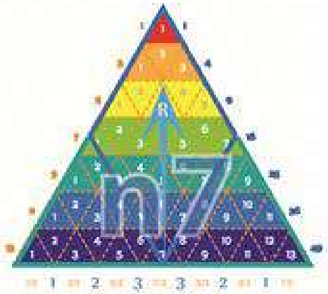

Planck's law describes the electromagnetic radiation emitted by a black body in thermal equilibrium at a specific temperature

$$\Delta[nh\nu]$$

SQUARED energy levels have ODD numbered quanta in each energy level [bosons]

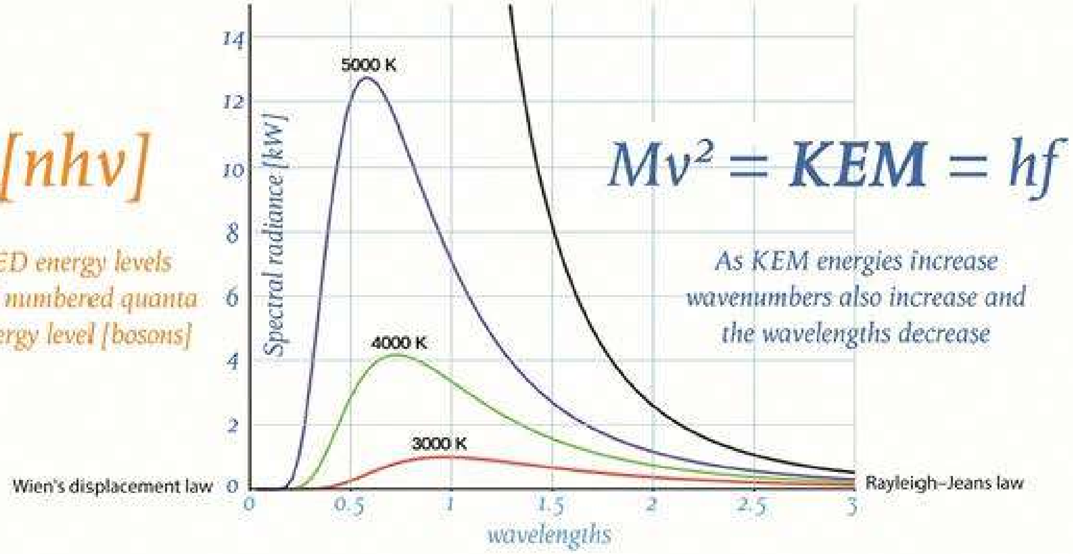

$$Mv^2 = \mathbf{KEM} = hf$$

As KEM energies increase wavenumbers also increase and the wavelengths decrease

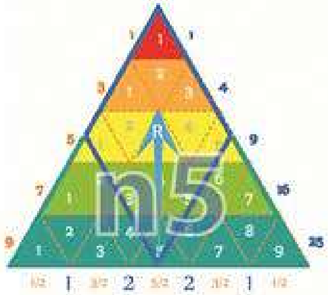

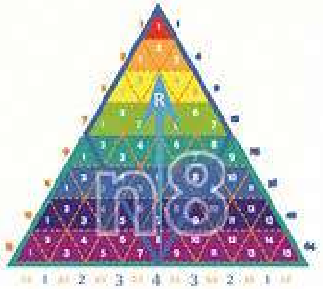

Spectral lines can be understood in quantum theory as differences between energy levels and proportional to wavenumber, frequency or wavelengths

Kinetic mass-energy momenta relationships

Photons are comprised of dual EM masses

Photons are KEM energy-momenta propagating at the speed of Light

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

kg **m**

$\frac{1}{s}$ f

Frequency and Wavelength are related by the natural speed of Energy

c

m λ

KG **0**

Q

Photons have neutral charge

[ν-ν]

f

RE

Photons are Lorentz variant

p kg $\frac{m}{s}$

v $\frac{m}{s}$

c

'c' is the natural speed of Energy in free space

c $\frac{m}{s}$

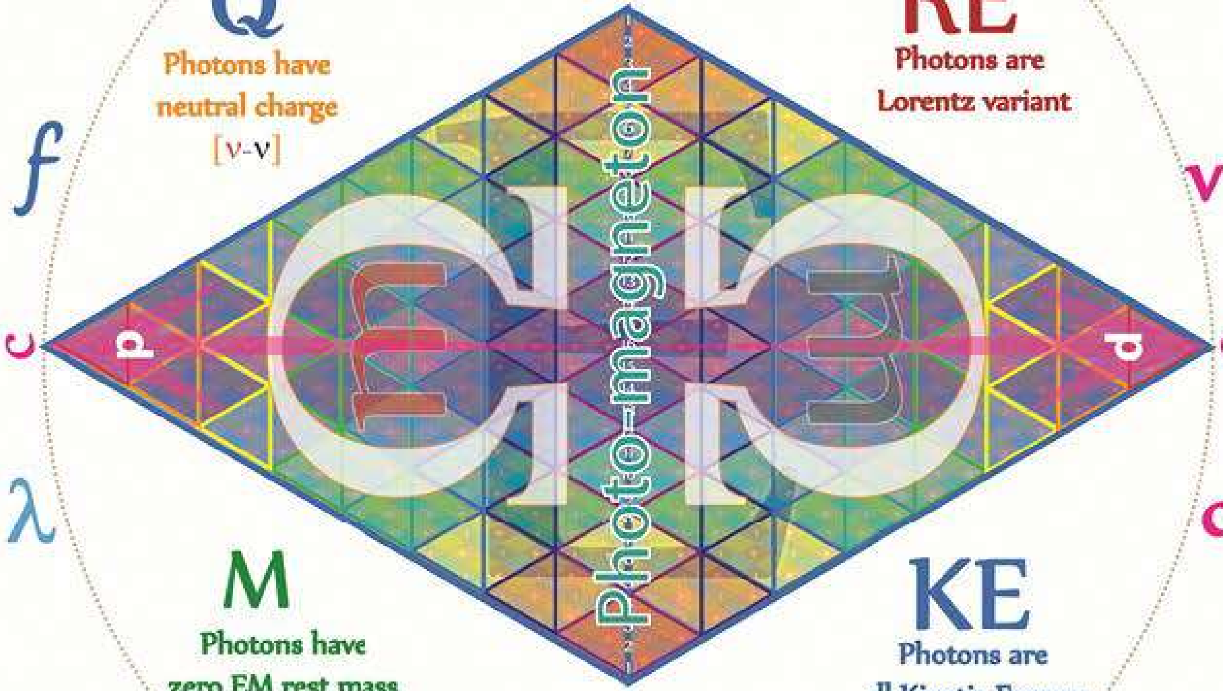

M

Photons have zero EM rest mass

KE

Photons are all Kinetic Energy

E kg $\frac{m^2}{s^2}$

$$\Delta 2hv = E = hf$$

Planck Einstein

c^2

1 second

Photons are 'Matterless' 2D neutral charge mass-energy geometries

Kinetic Energy is 1/2 the energy created by Matter topologies in motion

Tetryonics 29.12 - Kinetic mass-energy relationships

Copyright ABRAHAM [2008] - all rights reserved

Relativistic Lorentz corrections

Hendrik Lorentz

$$\sqrt{1-\beta^2}$$

$$E=\frac{mc^2}{\sqrt{1-\frac{v^2}{c^2}}}$$

$$\sqrt{1-\left(\frac{v}{c}\right)^2}$$

(18 July 1853 – 4 February 1928)

The Lorentz transformation was originally the result of attempts by Lorentz and others to explain how the speed of light was observed to be independent of the reference frame and later expanded upon in an attempt to understand the creation of emfs in magnets and symmetries in electromagnetic forces

Einstein inlcuded the transformation in his theory of Special relativity [SR].

The Lorentz transformation supersedes the Galilean transformation of Newtonian physics, which assumes an absolute space and time.

According to special relativity, the Galilean transformation is a good approximation only at relative speeds much smaller than the speed of light.

Tetryonics expands upon all of this revealing this relationship as a result of the geometry of electromagnetic mass-energies in motion

rest mass-Matter

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Energy

Planck quanta

mass velocity

momentum

spatial co-ordinate system

Matter is EM energy propagating at the 'speed of Light' In a standing wave energy topology

One of the greatest mistakes in relativistic mechanics is the application of Lorentz corrections to Matter.

It stems from there being no definition and enforced differentiation between EM mass and Matter

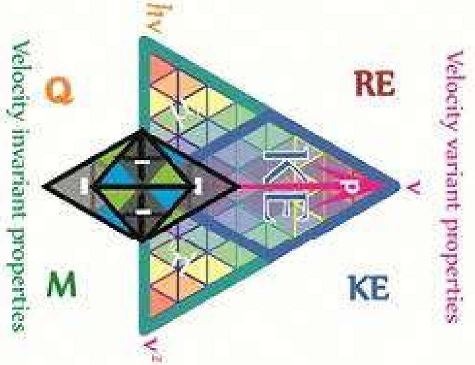

$$m = \frac{E}{c^2}$$

2D planar fields are relativisitic

radiant EM mass geometry

$$\frac{m}{c^2} = M$$

standing-wave Matter topology

3D standing-waves are velocity invariant

$$\frac{E}{c^4} = M$$

mass is a property of Matter - Matter is not a property of mass [they are directly related throught the velocity of light]

$$\beta = \left[\frac{v}{c}\right]$$

velocity is a result of vector forces

All 2D energy waveforms propagate at the 'speed of Light'

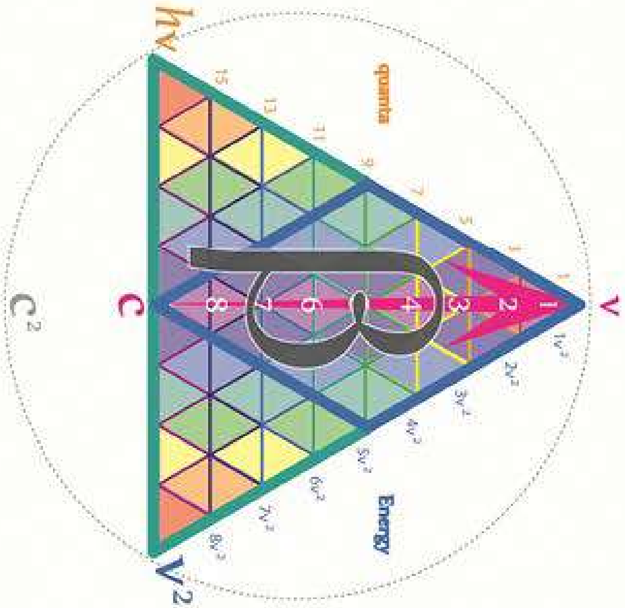

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

The Kinetic Energy of Matter in motion is directly related to the square of the velocity

$$\gamma \equiv \frac{c}{\sqrt{c^2-v^2}} = \frac{1}{\sqrt{1-\beta^2}} = \frac{dt}{d\tau}$$

The 'speed of Light' is a EM constant and is the limiting velocity acheivable by the electrical acceleration of Matter

Copyright ABRAHAM [2008] - all rights reserved

Spectral Energy relationships

$$\mathrm{kg}\,\frac{\mathrm{m}^2}{\mathrm{s}^2} \quad E = p^2$$

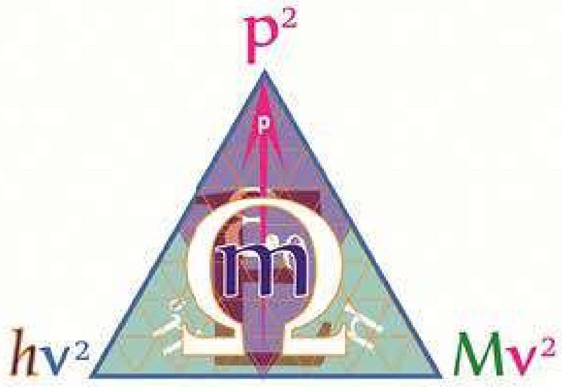

p²
p
Ω
m
hv²
Mv²

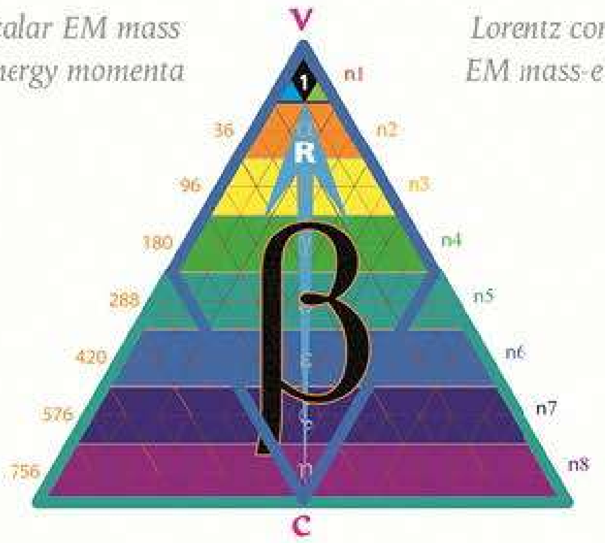

v
1
n1
36
n2
R
96
n3
180
n4
β
288
n5
420
n6
576
n7
756
n8
c

Scalar EM mass -energy momenta

Lorentz corrected EM mass-energies

$$KEM = \frac{mv^2}{c^2}$$

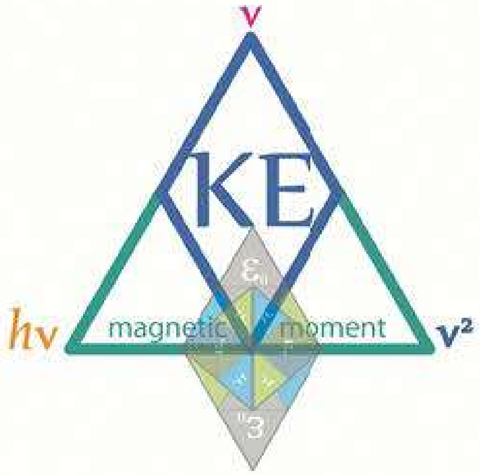

v
KE
ε₀
hv
magnetic
moment
v²
"3"

$$\boxed{Mv^2 = \underset{13.525\ \mathrm{eV}}{KEM} = hcR_H}$$

All Spectral line emissions and absorptions produce changes in KEM energies, Angular momentum, Linear momentum, Frequency and Wavelengths

Linear

$$\left[\frac{\Omega}{c}\right] = \lambda$$

de Broglie wavelengths

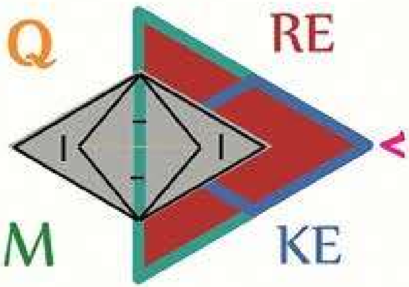

The 'speed of Light' is the maximum velocity acheivable by electrical energies

Scalar

$$f = \left[\frac{c^2}{\Omega}\right]$$

Compton frequencies

Tetryonics 29.14 - Spectral Energy relationships

Spectral line photon production

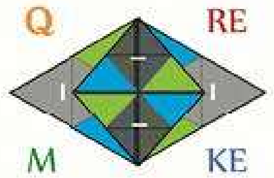

At no time does the 3D rest Matter-energy content of the electron change

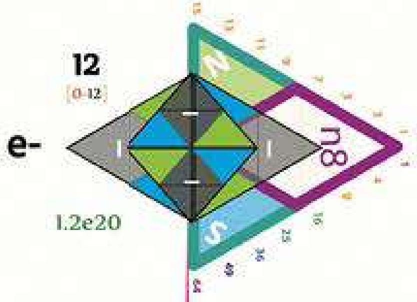
12
[0-12]
e-
n8
1.2e20

An electron transitioning from n8 to n7

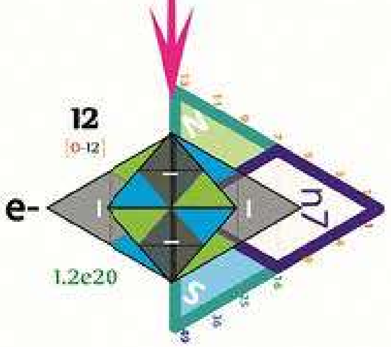
12
[0-12]
e-
n7
1.2e20

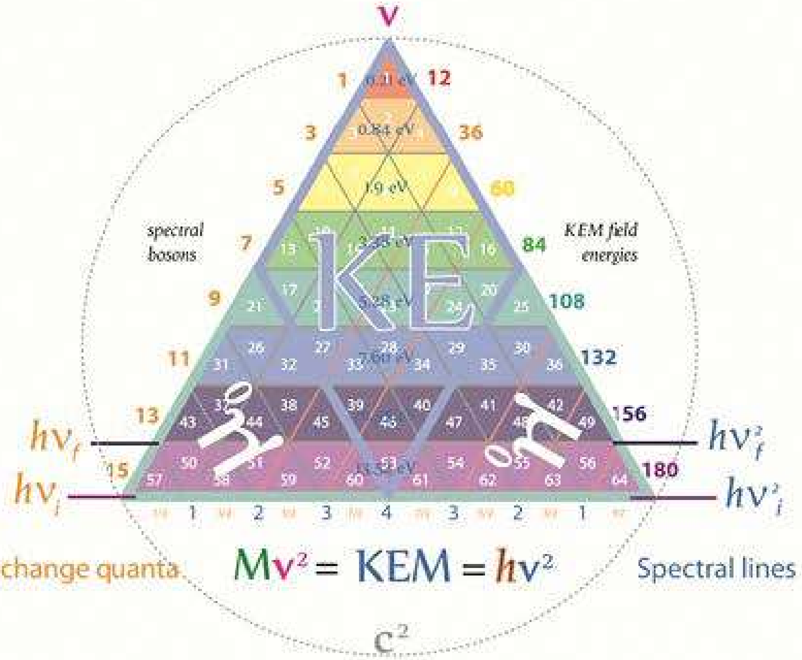
v
1
12
0.11 eV
3
36
0.84 eV
5
60
1.9 eV
7
84
3.38 eV
9
108
5.28 eV
11
132
7.60 eV
13
156
15
180
13.52 eV
spectral bosons
KEM field energies
KE
hν_f
hν_i
hν²_f
hν²_i
1 2 3 4 3 2 1
Exchange quanta
Spectral lines
c²

$$Mv^2 = KEM = hv^2$$

$$E = hf$$

As the 3D rest Matter of any particle in motion is Lorentz invariant any changes in velocity-momentum produces changes to the mass-energy momenta content of the 2D KEM field [boson changes are the difference of two squares]

Spectral line wavenumbers

Tetryonic

$$\left(\frac{15 \diagup 49}{64}\right)$$

.0047

$$1/\lambda$$

Rydberg

$$R\left(\frac{1}{49} - \frac{1}{64}\right)$$

.0047

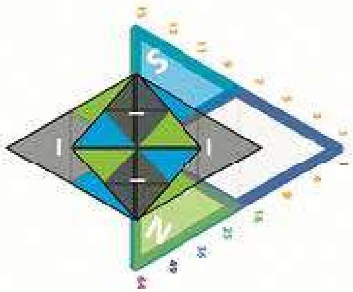

Does so by emitting Kinetic Energy quanta

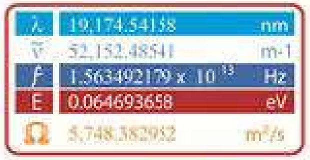

λ	19,174.54158	nm
ṽ	52,152.48541	m-1
f	1.563492179 x 10^{13}	Hz
E	0.064693658	eV
Ω	5,748.382952	m²/s

Which are released as photons of specific energy-momenta

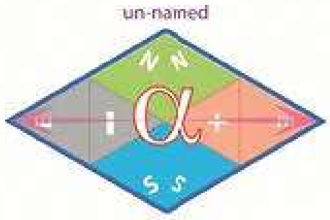

of specific wavelengths (Spectral Lines)

Wavenumbers

are the number of energy momenta wavelengths per unit distance

m

n1 2

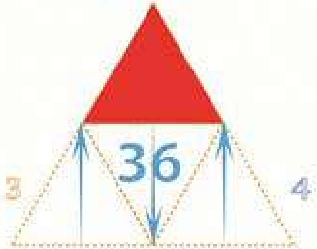

48

1 quantum level jump

α

n1 3

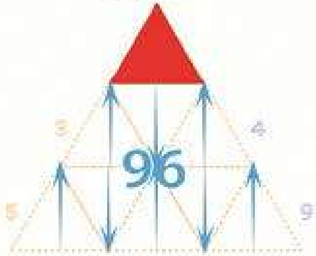

108

2 quantum level jump

β

n1 4

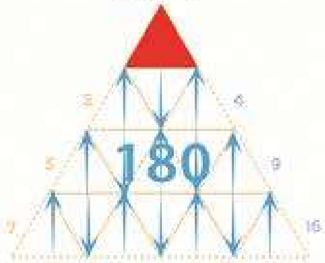

192

3 quantum level jump

γ

m

$\tilde{\nu}$

the wavenumber (also wave number) is the spatial frequency of a wave, either in quanta per distance or nπ radians per unit distance. It can be envisaged as the number of specific wavesets that exist over a specified distance

$1/\lambda$

m^{-1}

n1 5

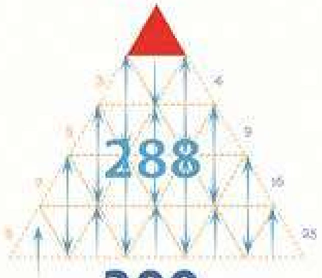

300

4 quantum level jump

δ

n1 6

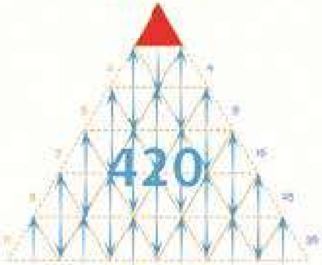

432

5 quantum level jump

ε

n1 7

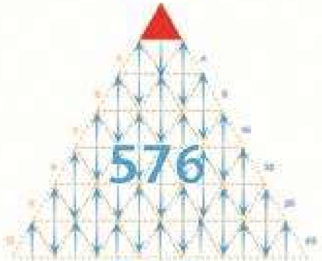

588

6 quantum level jump

ζ

n1 8

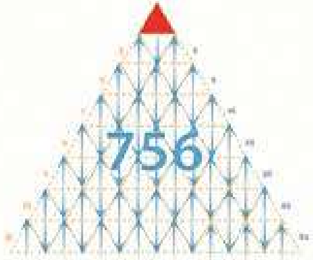

768

7 quantum level jump

η

For electromagnetic radiation in vacuum, wavenumber is proportional to frequency and to photon energy

Bosons, Photons and quantums of energy momenta

$\tilde{\nu}$ WAVENUMBERS

$E = \Sigma h\nu$

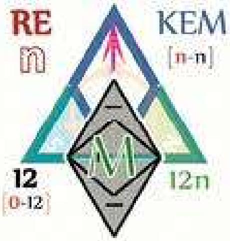
RE
n
KEM
[n-n]
M
12
[0-12]
12n

$KEM = Mv^2$

LINEAR MOMETNUM **p**

γ Spectral lines are produced $\Delta m\Omega v = \Delta mv = \Delta p$ by accelerating electrons **a**

equilateral energy momenta is the foundation of all quantum EM wave-functions

ODD distribution

$2\nu = f$

SQUARED distribution

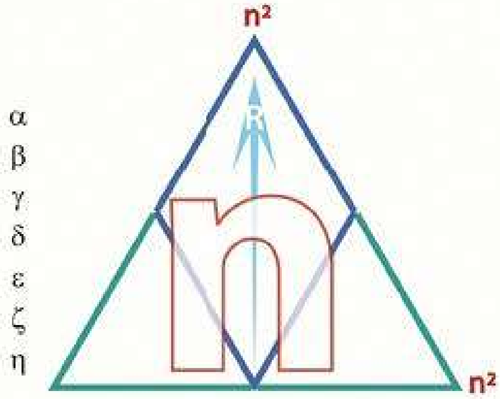
n²
n²
n
α
β
γ
δ
ε
ζ
η

Spectral lines

Bosons

$h\nu$

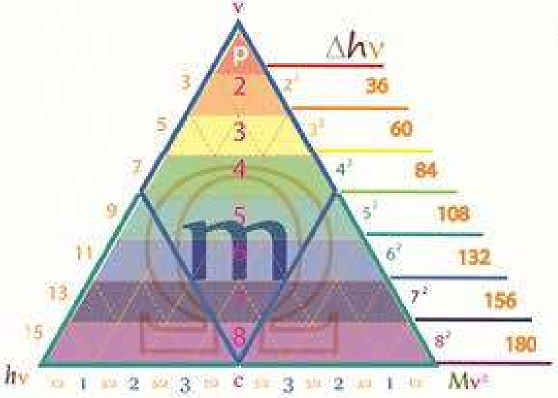
v
Δhv
p
2
3
4
5
m
8
3
5
7
9
11
13
15
2²
3²
4²
5²
6²
7²
8²
36
60
84
108
132
156
180
hv
1
2
3
c
3
2
1
Mv²

EVEN distribution

hf

Photons

Kinetic energies

$h\nu^2$

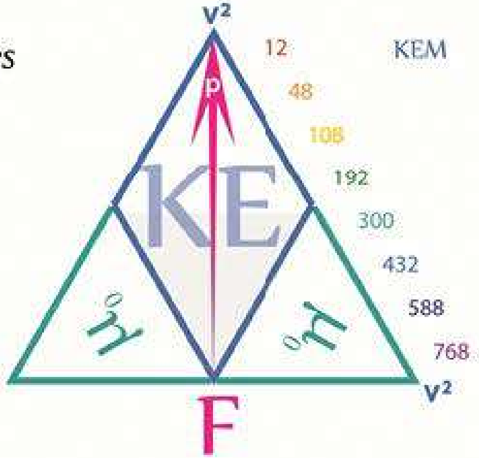
v²
v²
12
48
108
192
300
432
588
768
KEM
p
KE
F

energy momenta

Copyright ABRAHAM [2008] - all rights reserved

Spectral Energy - Planck relationships

$$\frac{E}{v^2} = m\Omega = \frac{E}{c^2}$$

deBroglie wavelength λ

Planck quanta differential

spectral mass-energies

f Frequency

p MOMENTUM

h

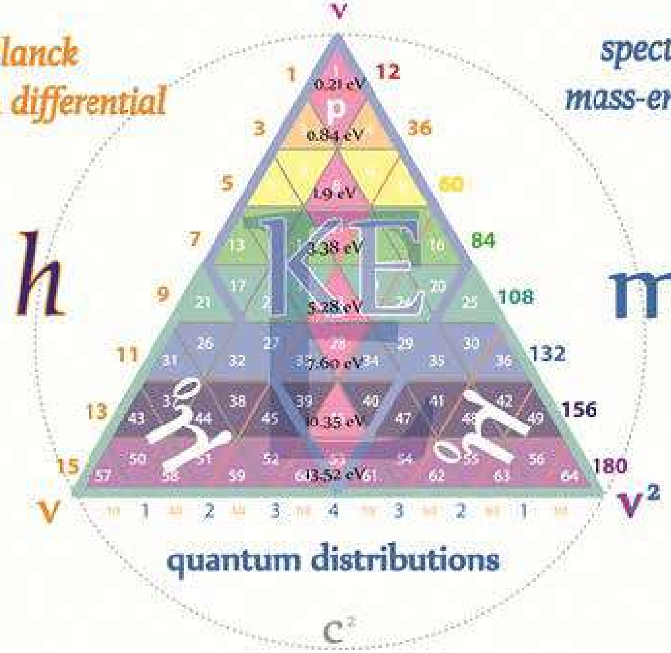
v
1
3
5
7
9
11
13
15
12
36
60
84
108
132
156
180
0.21 eV
p
0.84 eV
1.9 eV
3.38 eV
KE
5.28 eV
7.60 eV
10.35 eV
13.52 eV
h
h
v
v²
1 2 3 4 3 2 1
quantum distributions
c²

m

ENERGY E

Wave number $\tilde{v}$

v^2 Compton Frequency

Applying mass to spectral energy relationships reveals the familiar quantum mechanical wave-particle relationships and properties

	2D waves						3D Matter	
EM wave wavelength	Ω / c	=	planck constant	h / p	linear momentum	=	$m\Omega / mc$	deBroglie wavelength
EM wavenumber	c / Ω	=	linear momentum	p / h	planck constant	=	$mc / m\Omega$	Matter waves
EM wave Frequency	c^2 / Ω	=	kinetic energy	E / h	planck constant	=	$mc^2 / m\Omega$	Compton Frequency

Thus reinforcing the geometric role of quantised angular momenta in Planck's constant

Tetryonics 29.18 - Spectral energy relationships

White Light

White light is radiated by our SUN and all the stars throughout the Universe.
In fact, most of the energy radiated by the sun is within the visible spectrum, which is most likely why we see this range.

Incandescence is the greatest known generator of white light. The verb incandesce means to glow white. This is also referred to as black-body radiation.

The color of a black body object at high temperatures causes it to glow, and the waves emitted include visible light.
In addition to the sun and common light bulb, molten materials such as metal or glass also glow incandescently.

Hydrogen
Oxygen
Sulfur

Spectroscopy

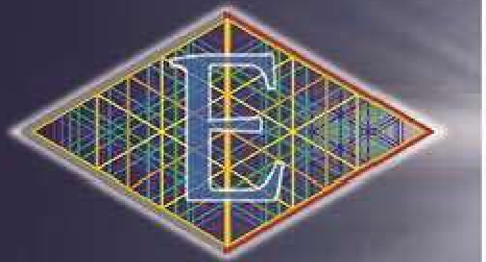

Planck's Heat Law

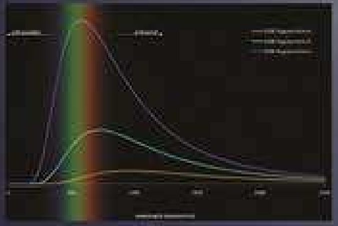

Most light sources are thermal, meaning the type of radiation
they emit is a characteristic of the source's temperature.

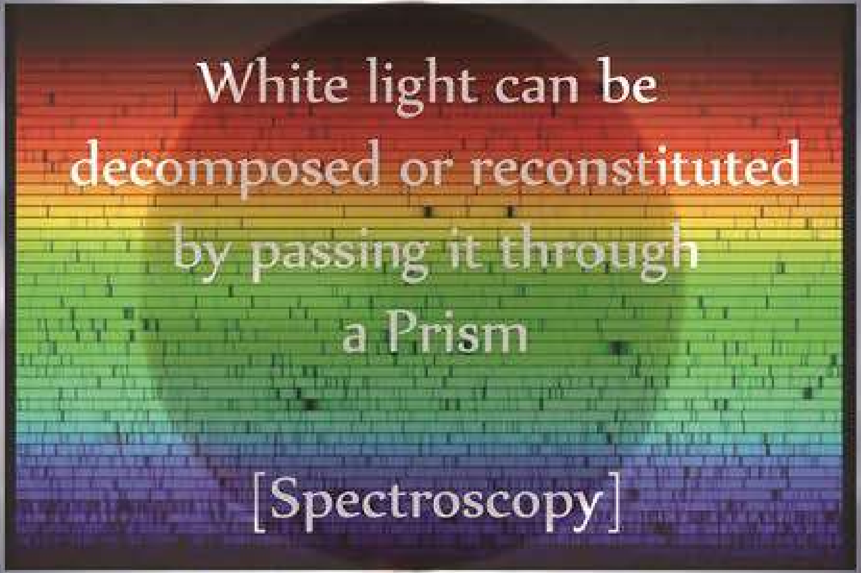

White light can be
decomposed or reconstituted
by passing it through
a Prism

[Spectroscopy]

White light is comprised
of many photons of varying
frequencies and wavelengths

Each colour of light has
a specific frequency
and wavelength

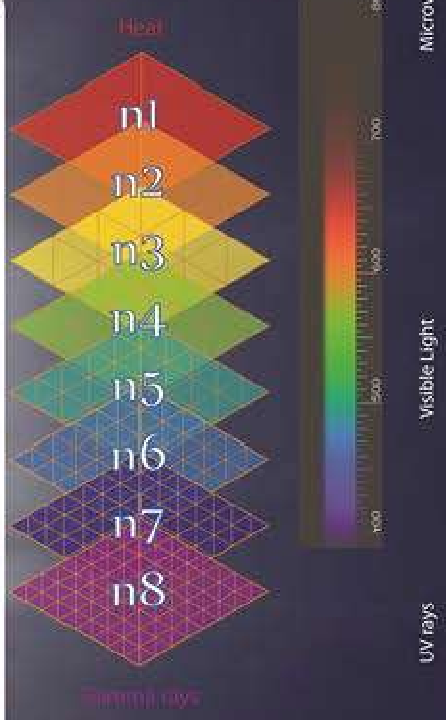

Until Newton's work proved otherwise, most scientists believed that white was the fundamental color of light; and that other colors were formed only by adding something to light. Newton demonstrated this was not true by passing white light through a prism, then through another prism. If the colors were added by the prism, the second prism should have added further colors to the single-colored beam. Since the single-colored beam remained a single color, Newton concluded that the prism merely separated the colors already present in the light.

White light is the result of superpositioning of the visible colors [varying frequencies] of EM waves

Copyright ABRAHAM [2008] - all rights reserved

Max Karl Ernst Ludwig Planck

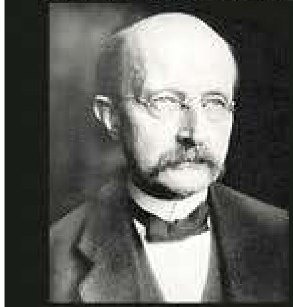

(April 23, 1858 – October 4, 1947)

White light

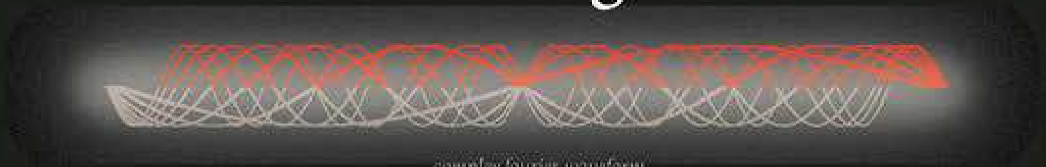

complex fourier waveform

is comprised of superpositioned differing spectral wavelength photons

Johannes (Janne) Robert Rydberg

(November 8, 1854 – December 28, 1919)

spectral lines & colour

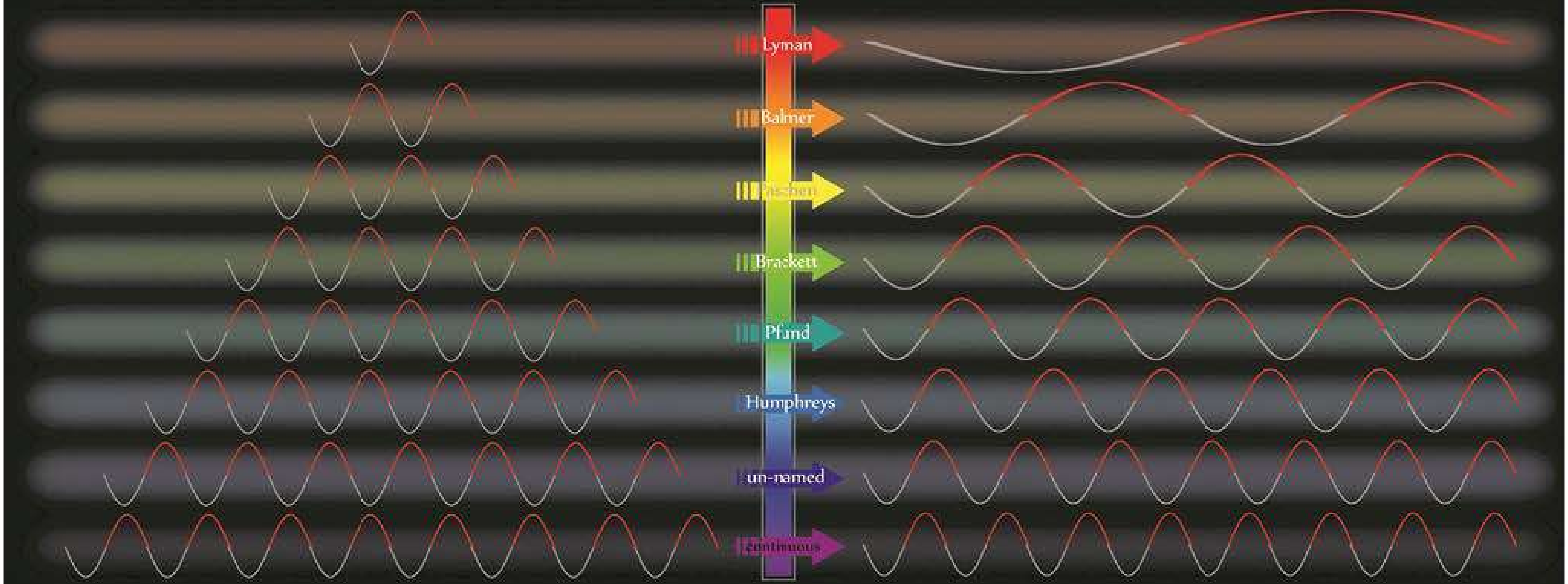
bosons
photons
Lyman
Balmer
Paschen
Brackett
Pfund
Humphreys
un-named
continuous

Instantaneous measurement of energy momenta released

Time measurement of released waveforms [photons]

created by the fractional release of equilateral KEM field energy momenta

Tetryonics 29.20 - White light & spectral Colours

Copyright ABRAHAM [2008] - all rights reserved

Lyman spectral lines

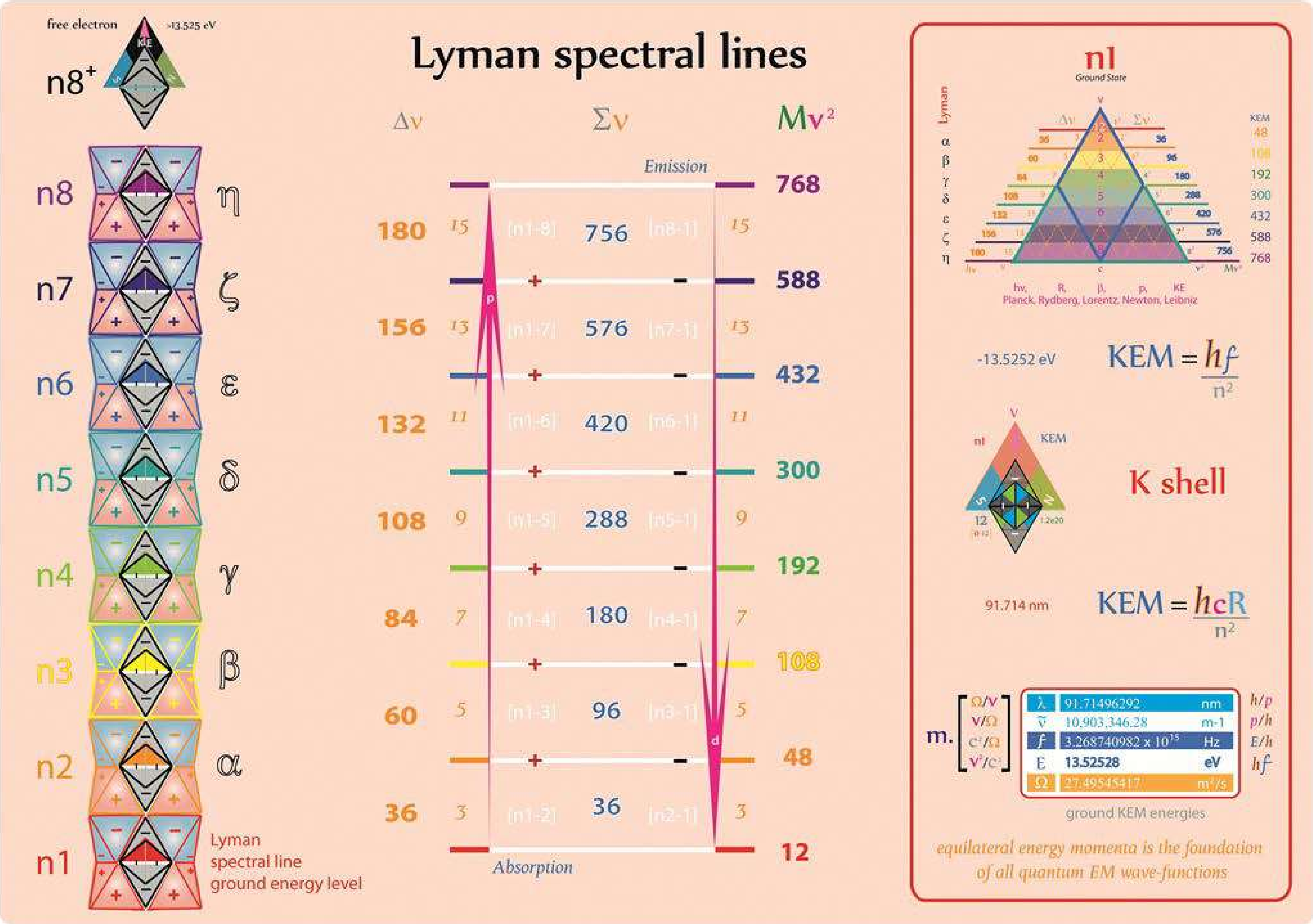
free electron
>13.525 eV
KE
n8+
n8
η
n7
ζ
n6
ε
n5
δ
n4
γ
n3
β
n2
α
n1
Lyman
spectral line
ground energy level
Δv
Σv
Mv²
Emission
768
180
15
[n1-8]
756
[n8-1]
15
588
156
13
[n1-7]
576
[n7-1]
13
432
132
11
[n1-6]
420
[n6-1]
11
300
108
9
[n1-5]
288
[n5-1]
9
192
84
7
[n1-4]
180
[n4-1]
7
108
60
5
[n1-3]
96
[n3-1]
5
48
36
3
[n1-2]
36
[n2-1]
3
12
Absorption
p
d
n1
Ground State
Lyman
α
β
γ
δ
ε
ζ
η
v
Δv
Σv
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
36
96
180
288
420
576
756
c
hv
Mv²
hv,
R,
β,
p,
KE
Planck, Rydberg, Lorentz, Newton, Leibniz
v
nl
KEM
12
1.2e20

-13.5252 eV

$$KEM = \frac{hf}{n^2}$$

K shell

91.714 nm

$$KEM = \frac{hcR}{n^2}$$

$$m.\begin{bmatrix} \Omega/v \\ v/\Omega \\ c^2/\Omega \\ v^2/c^2 \end{bmatrix}$$

λ	91.71496292	nm	h/p
$\tilde{v}$	10,903,346.28	m-1	p/h
f	3.268740982 x 10^{15}	Hz	E/h
E	13.52528	eV	hf
Ω	27.49545417	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

Tetryonics 30.01 - Lyman spectral lines

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

-13.5252 eV
v
nl
KE
12
[0-12]
1.2e20
91.714 nm

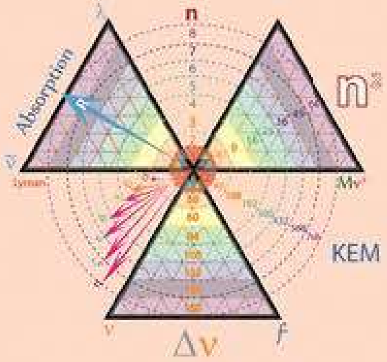
λ
n
Absorption
n^2
Lyman
Mv²
KEM
v
Δν
f

Lyman spectral series

α	122.28	n1-2
β	103.18	n1-3
γ	97.83	n1-4
δ	95.53	n1-5
ε	94.33	n1-6
ζ	93.62	n1-7
η	93.17	n1-8

UV spectrum

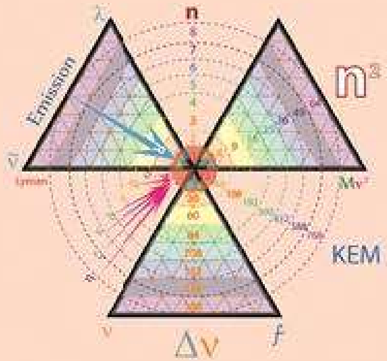
λ
n
Emission
n^2
Lyman
Mv²
KEM
v
Δν
f

K shell

Δν

KEM

n1~2

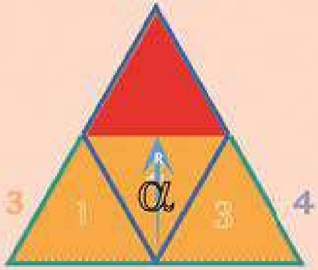

36

48

1 quantum level jump

n1~3

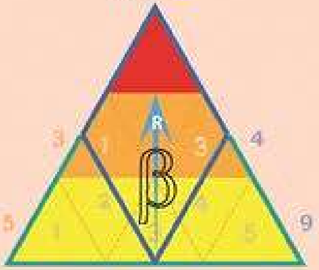

96

108

2 quantum level jump

n1~4

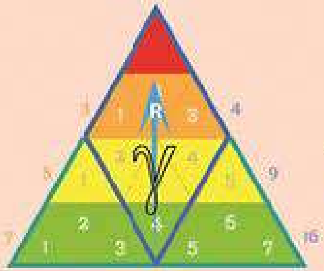

180

192

3 quantum level jump

n1~5

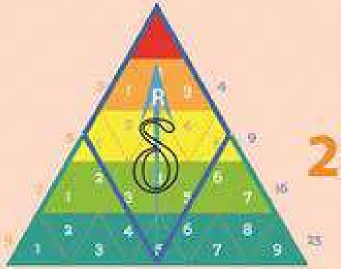

288

300

4 quantum level jump

n1~6

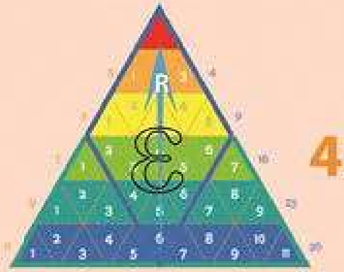

420

432

5 quantum level jump

n1~7

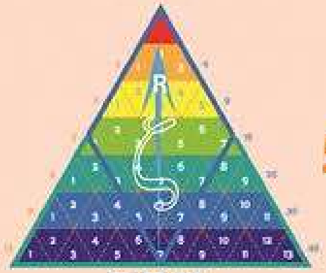

576

588

6 quantum level jump

n1~8

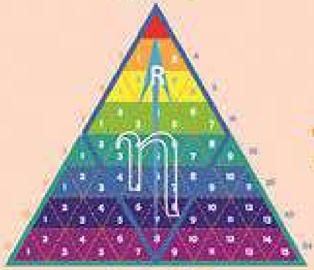

756

768

7 quantum level jump

Copyright ABRAHAM [2008] - all rights reserved

Lyman series energy momenta

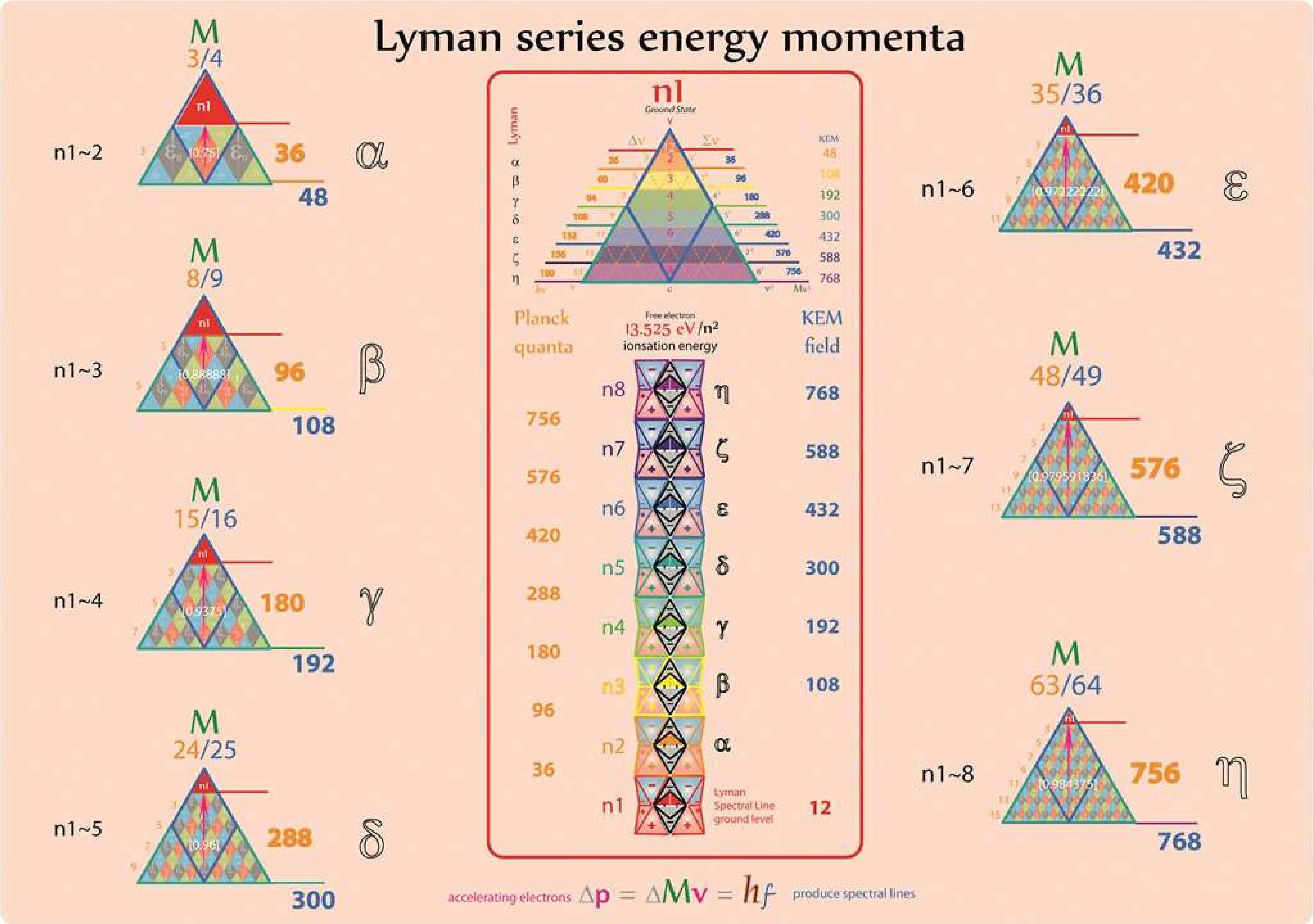
M
3/4
n1
n1~2
[0.75]
36
48
α
M
8/9
n1
n1~3
[0.88888]
96
108
β
M
15/16
n1
n1~4
[0.9375]
180
192
γ
M
24/25
n1
n1~5
[0.96]
288
300
δ
n1
Ground State
Lyman
α
β
γ
δ
ε
ζ
η
Δv
Σv
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
36
96
180
288
420
576
756
Planck
quanta
Free electron
13.525 eV /n²
ionsation energy
KEM
field
756
576
420
288
180
96
36
n8
n7
n6
n5
n4
n3
n2
n1
η
ζ
ε
δ
γ
β
α
Lyman
Spectral Line
ground level
768
588
432
300
192
108
12
M
35/36
n1~6
[0.97222222]
420
432
ε
M
48/49
n1~7
[0.979591836]
576
588
ζ
M
63/64
n1~8
[0.984375]
756
768
η

accelerating electrons $\Delta p = \Delta Mv = hf$ produce spectral lines

Tetryonics 30.03 - Lyman series energy momenta

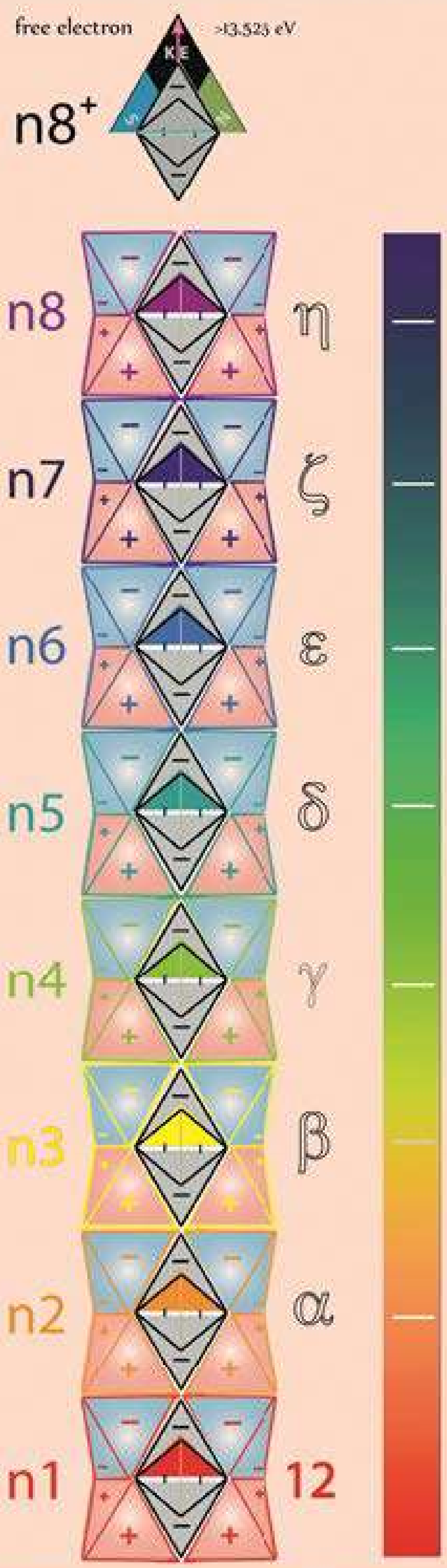

Lyman Series

$$\Delta h\nu = \text{KEM} = hcR$$

Tetryonic · 13.525 eV · Rydberg

63/64
.9843

λ	93.17075598	nm
ν̃	10,732,981.5	m-1
f	3.217666905 x 10^{15}	Hz
E	13.31395504	eV
Ω	[illegible]	m^2/s

$R\left(\frac{1}{1}-\frac{1}{64}\right)$
.984375

48/49
.9795

λ	93.62569132	nm
ν̃	10,680,829.01	m-1
f	3.202031983 x 10^{15}	Hz
E	13.24926138	eV
Ω	[illegible]	m^2/s

$R\left(\frac{1}{1}-\frac{1}{49}\right)$
.9795

35/36
.9722

λ	94.31539043	nm
ν̃	10,600,475.55	m-1
f	3.177842622 x 10^{15}	Hz
E	13.14958523	eV
Ω	[illegible]	m^2/s

$R\left(\frac{1}{1}-\frac{1}{36}\right)$
.9722

24/25
.9600

λ	95.53641971	nm
ν̃	10,467,212.43	m-1
f	3.137991343 x 10^{15}	Hz
E	12.98427616	eV
Ω	[illegible]	m^2/s

$R\left(\frac{1}{1}-\frac{1}{25}\right)$
.96

15/16
.9375

λ	97.82929378	nm
ν̃	10,221,887.14	m-1
f	3.064444671 x 10^{15}	Hz
E	12.67995718	eV
Ω	[illegible]	m^2/s

$R\left(\frac{1}{1}-\frac{1}{16}\right)$
.9375

8/9
.8888

λ	103.1793333	nm
ν̃	9,691,863.362	m-1
f	2.90534754 x 10^{15}	Hz
E	12.02247792	eV
Ω	[illegible]	m^2/s

$R\left(\frac{1}{1}-\frac{1}{9}\right)$
.888

3/4
.7500

λ	122.2866172	nm
ν̃	8,177,509.712	m-1
f	2.451555737 x 10^{15}	Hz
E	10.14396575	eV
Ω	[illegible]	m^2/s

$R\left(\frac{1}{1}-\frac{1}{4}\right)$
.75

ground KEM energies

$$m.\begin{bmatrix}\Omega/v\\ v/\Omega\\ c^2/\Omega\\ v^2/c^2\end{bmatrix}$$

λ	91.71496292	nm	h/p
ν̃	10,903,346.28	m-1	p/h
f	3.268740982 x 10^{15}	Hz	E/h
E	13.52528	eV	hf
Ω	27.49545417	m^2/s	

changing energy momenta produces specific frequency spectral lines

Copyright ABRAHAM [2008] - all rights reserved

Balmer spectral lines

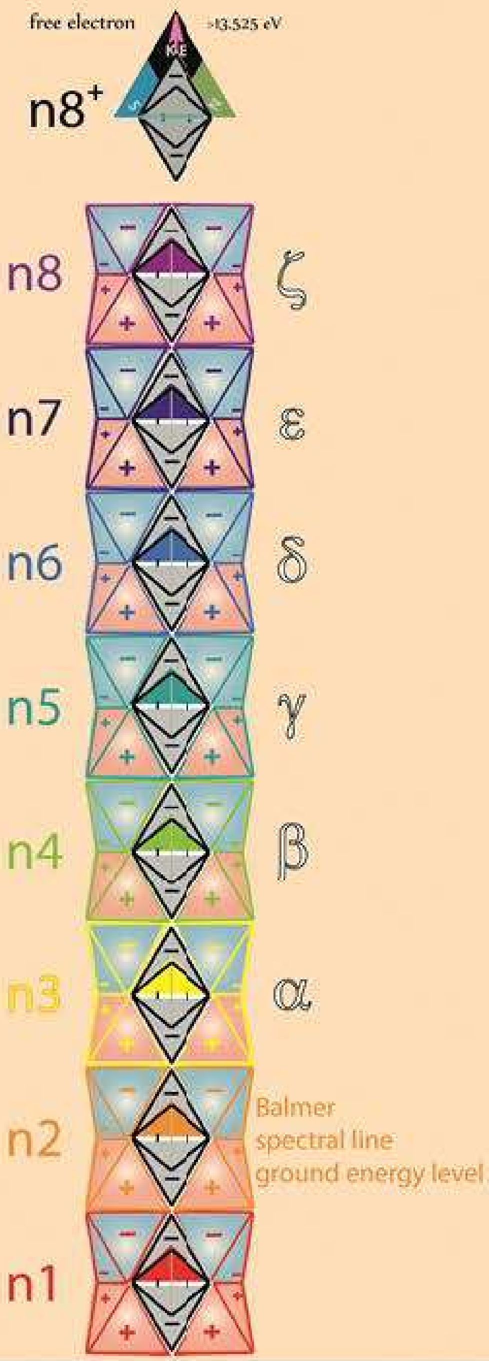
free electron
>13.525 eV
KE
n8+
n8
ζ
n7
ε
n6
δ
n5
γ
n4
β
n3
α
n2
Balmer
spectral line
ground energy level
n1
+
−

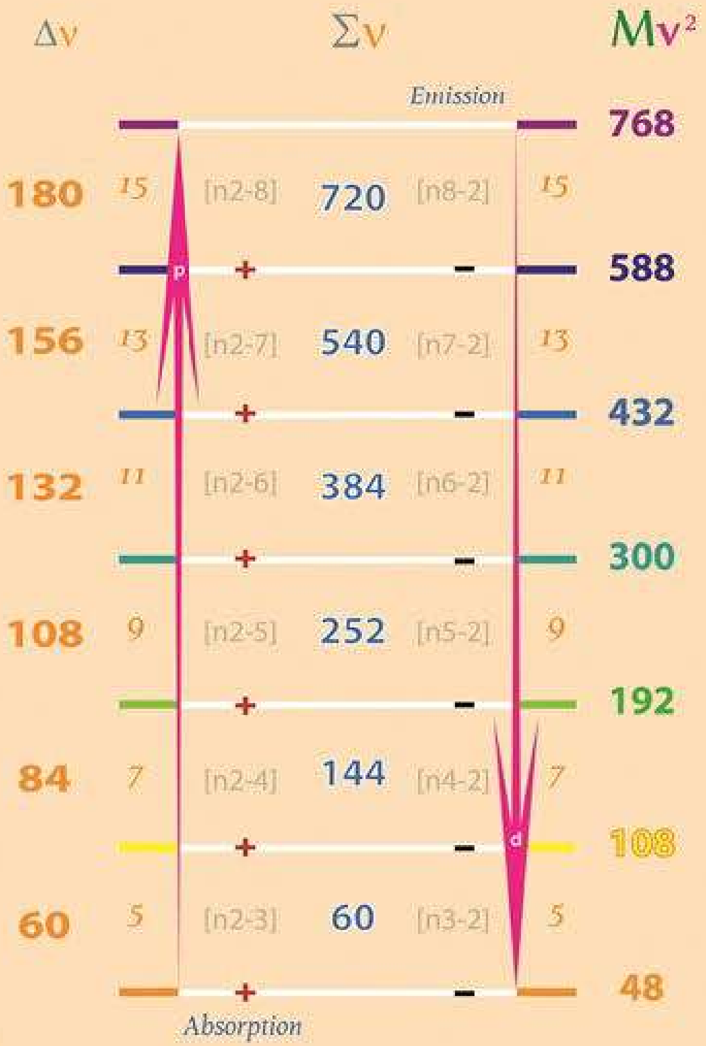
Δv
Σv
Mv²
Emission
768
180
15
[n2-8]
720
[n8-2]
15
p
588
156
13
[n2-7]
540
[n7-2]
13
432
132
11
[n2-6]
384
[n6-2]
11
300
108
9
[n2-5]
252
[n5-2]
9
192
84
7
[n2-4]
144
[n4-2]
7
d
108
60
5
[n2-3]
60
[n3-2]
5
48
Absorption
+
−

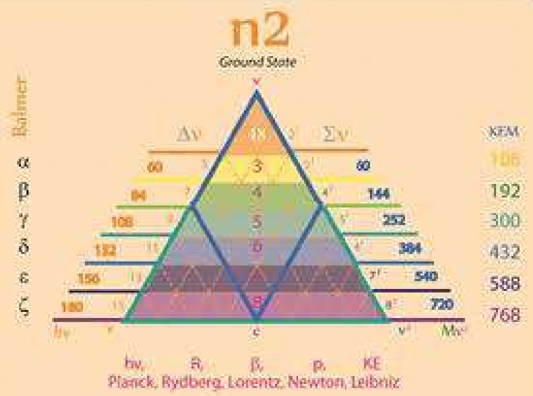
n2
Ground State
Balmer
α
β
γ
δ
ε
ζ
Δv
Σv
KEM
48
3
4
5
6
8
60
84
108
132
156
180
60
144
252
384
540
720
108
192
300
432
588
768
v
c
hv
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-3.38125 eV

$$KEM = \frac{hf}{n^2}$$

v
n2
KE
S
N
12
1.2e20

L shell

366.859 nm

$$KEM = \frac{hcR}{n^2}$$

$$m.\begin{bmatrix}\Omega/v\\ v/\Omega\\ c^2/\Omega\\ v^2/c^2\end{bmatrix}$$

λ	366.8598517	nm	h/p
ṽ	2,725,836.571	m-1	p/h
f	8.171852456 x 10¹⁴	Hz	E/h
E	3.381321915	eV	hf
Ω	109.9818167	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

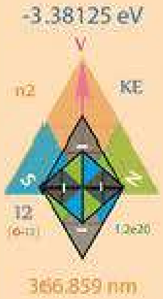
-3.38125 eV
v
n2
KE
12
(6-12)
12e20
366.859 nm

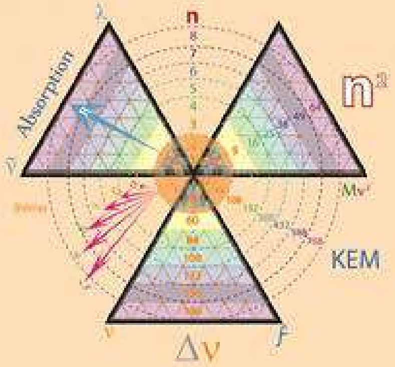
λ
n
Absorption
n²
Mv²
KEM
Δν
v
f

Balmer spectral series

α	656.34	n2-3
β	489.14	n2-4
γ	436.73	n2-5
δ	412.71	n2-6
ε	399.46	n2-7
ϕ	391.31	n2-8

Visible Spectrum

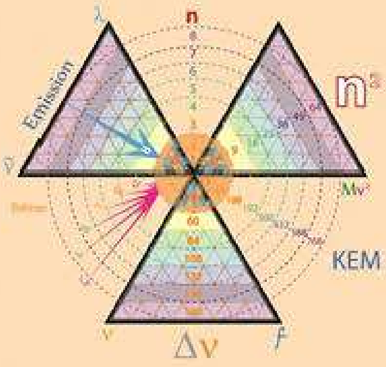
λ
n
Emission
n²
Mv²
KEM
Δν
v
f

L shell

Δν

KEM

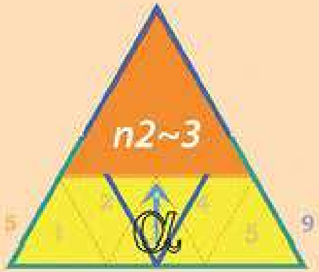

60

108

1 quantum level jump

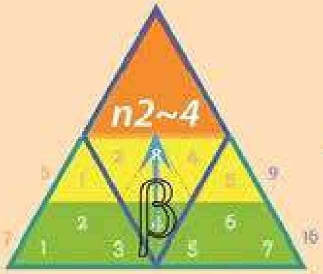

144

192

2 quantum level jump

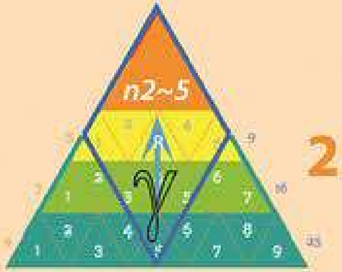

252

300

3 quantum level jump

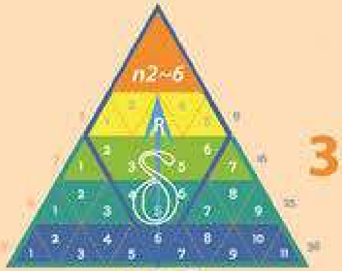

384

432

4 quantum level jump

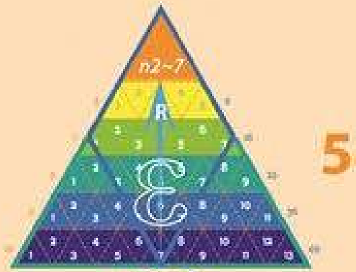

540

588

5 quantum level jump

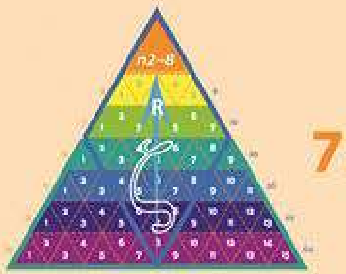

720

768

6 quantum level jump

Tetryonics 31.02 - Balmer quantum transitions

Balmer series energy momenta

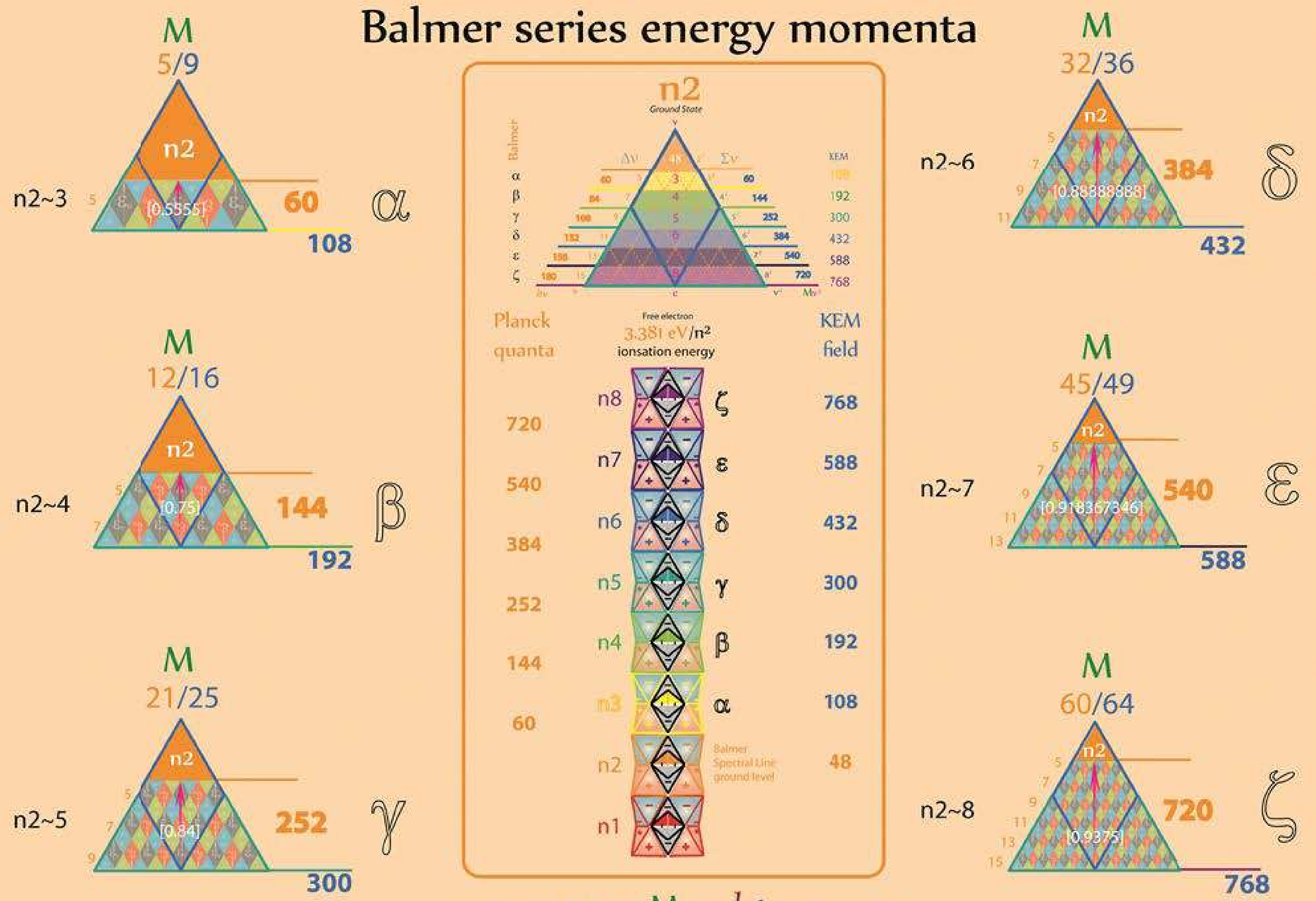
M
5/9
n2
n2~3
[0.5555]
60
108
α
M
12/16
n2
n2~4
[0.75]
144
192
β
M
21/25
n2
n2~5
[0.84]
252
300
γ
n2
Ground State
Balmer
Δv
Σv
KEM
α
β
γ
δ
ε
ζ
60
84
108
132
156
180
60
144
252
384
540
720
108
192
300
432
588
768
Free electron
3.381 eV/n²
ionsation energy
M
32/36
n2
n2~6
[0.88888888]
384
432
δ
M
45/49
n2
n2~7
[0.918367346]
540
588
ε
M
60/64
n2
n2~8
[0.9375]
720
768
ζ

	Planck quanta		KEM field
n8	720	ζ	768
n7	540	ε	588
n6	384	δ	432
n5	252	γ	300
n4	144	β	192
n3	60	α	108
n2		Balmer Spectral Line ground level	48
n1			

accelerating electrons $\Delta p = \Delta M v = hf$ produce spectral lines

Tetryonics 31.03 - Balmer series energy momenta

Copyright ABRAHAM [2008] - all rights reserved

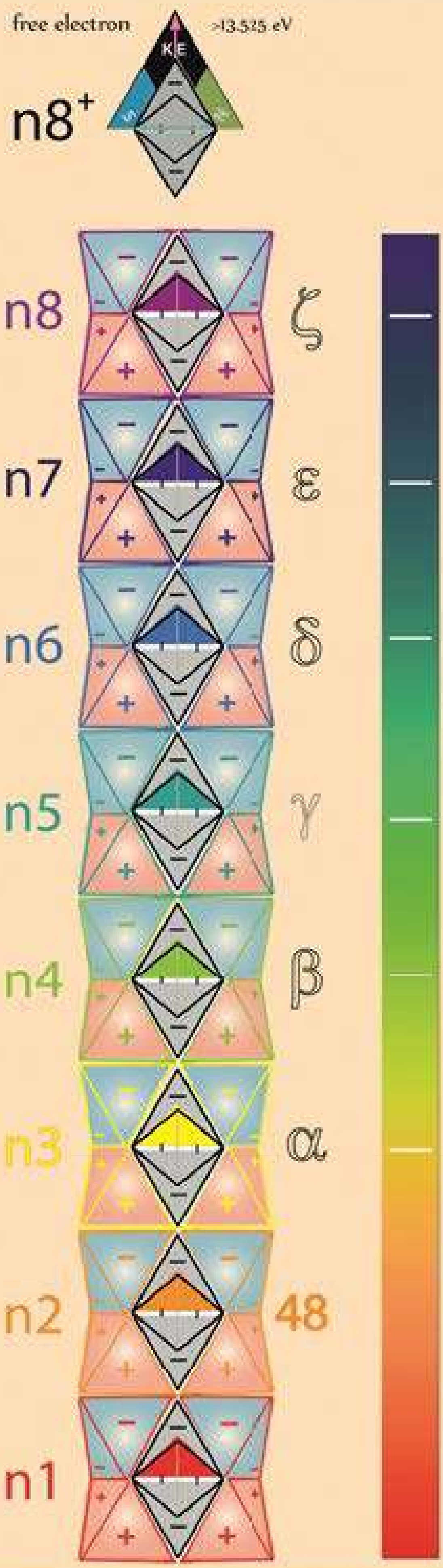

Balmer Series

$$\Delta h\nu = \text{KEM} = hcR$$

Tetryonic — 3.381 eV — Rydberg

60/64 .2343	λ 391.3171751 nm ν̃ 2,555,471.1785 m-1 f 7.661111677 x 10¹⁴ Hz E 3.169989296 eV Ω [illegible] m²/s	$R\left(\frac{1}{4}-\frac{1}{64}\right)$ .2343
45/49 .2295	λ 399.4696103 nm ν̃ 2,5033,319.299 m-1 f 7.50476246 x 10¹⁴ Hz E 3.105295637 eV Ω [illegible] m²/s	$R\left(\frac{1}{4}-\frac{1}{49}\right)$ .2295
32/36 .2222	λ 412.7173331 nm ν̃ 2,422,965.84 m-1 f 7.26386885 x 10¹⁴ Hz E 3.00561948 eV Ω [illegible] m²/s	$R\left(\frac{1}{4}-\frac{1}{36}\right)$ .2222
21/25 .2100	λ 436.7379187 nm ν̃ 2,289,702.719 m-1 f 6.864326063 x 10¹⁴ Hz E 2.840310409 eV Ω [illegible] m²/s	$R\left(\frac{1}{4}-\frac{1}{25}\right)$ .21
12/16 .1875	λ 489.1464689 nm ν̃ 2,044,377.428 m-1 f 6.128889342 x 10¹⁴ Hz E 2.535991437 eV Ω [illegible] m²/s	$R\left(\frac{1}{4}-\frac{1}{16}\right)$ .1875
5/9 .138	λ 660.347733 nm ν̃ 1,514,353.65 m-1 f 4.539918031 x 10¹⁴ Hz E 1.878512175 eV Ω [illegible] m²/s	$R\left(\frac{1}{4}-\frac{1}{9}\right)$ .138

$$m.\begin{bmatrix}\Omega/v\\ v/\Omega\\ c^2/\Omega\\ v^2/c^2\end{bmatrix}$$

λ	366.8598517	nm	h/p
ν̃	2,725,836.571	m-1	p/h
f	8.171852456 x 10¹⁴	Hz	E/h
E	3.381321915	eV	hf
Ω	109.9818167	m²/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Paschen spectral lines

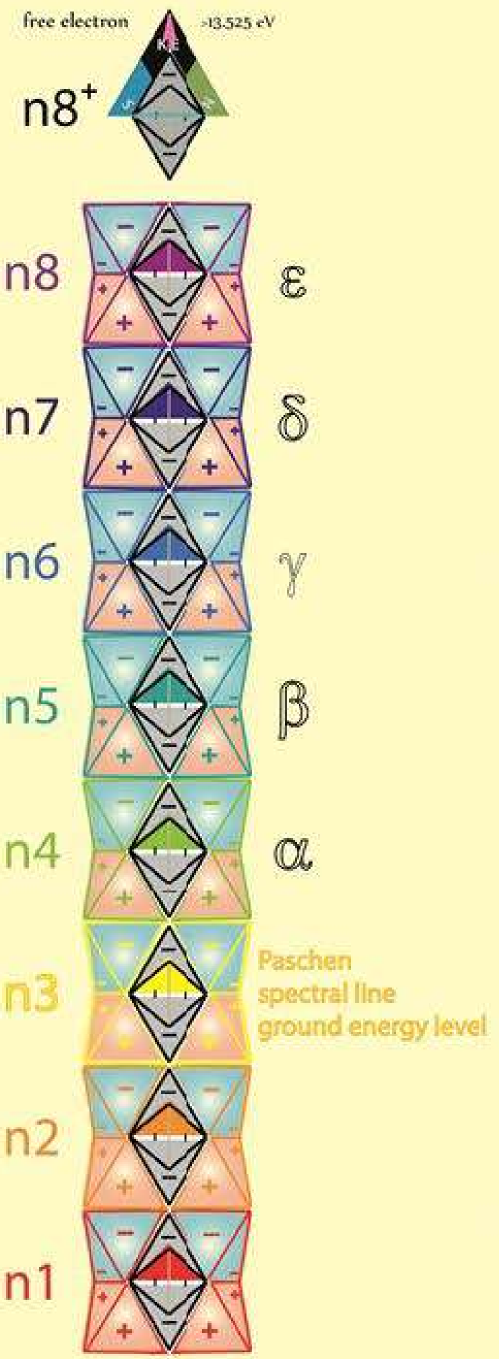
free electron
>13.525 eV
KE
n8⁺
n8
ε
n7
δ
n6
γ
n5
β
n4
α
n3
Paschen
spectral line
ground energy level
n2
n1

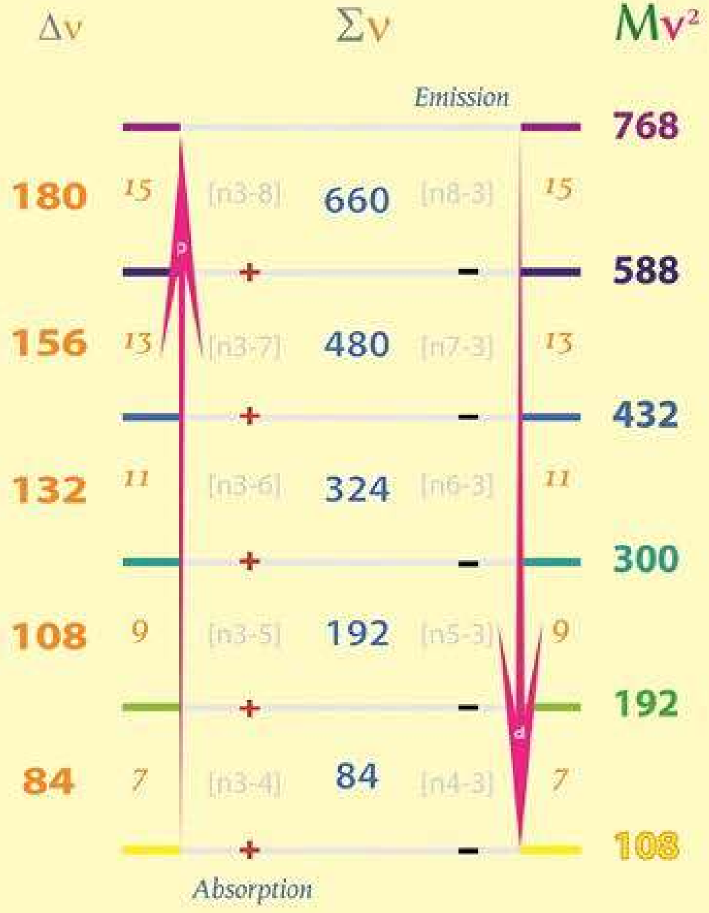
Δv
Σv
Mv²
Emission
768
180
15
[n3-8]
660
[n8-3]
15
588
156
13
[n3-7]
480
[n7-3]
13
432
132
11
[n3-6]
324
[n6-3]
11
300
108
9
[n3-5]
192
[n5-3]
9
192
84
7
[n3-4]
84
[n4-3]
7
108
Absorption

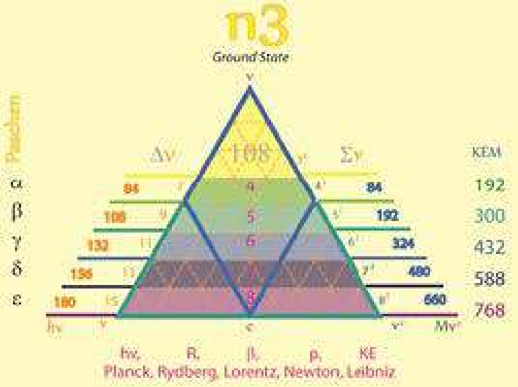
n3
Ground State
Paschen
α
β
γ
δ
ε
Δv
108
Σv
KEM
84
84
192
108
192
300
132
324
432
156
480
588
180
660
768
hv
c
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-1.50280974 eV

$$KEM = \frac{hf}{n^2}$$

v
n3
KE
12
[e-u]
1.2e20

M shell

825.434 nm

$$KEM = \frac{hcR}{n^2}$$

$$m.\begin{bmatrix}\Omega/v \\ v/\Omega \\ c^2/\Omega \\ v^2/c^2\end{bmatrix}$$

λ	825.4346663	nm	h/p
ṽ	1,211,482.92	m-1	p/h
f	$3.631934425 \times 10^{14}$	Hz	E/h
E	1.50280974	eV	hf
Ω	247.4590875	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

-1.50280974 eV

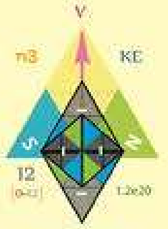
v
n3
KE
12
1.2e20

825.434 nm

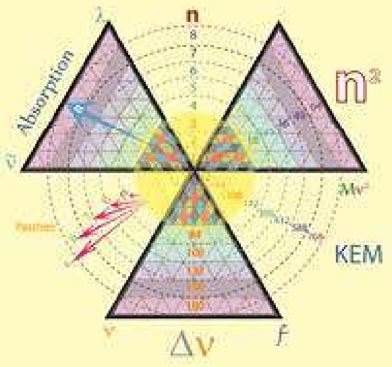
λ
n
Absorption
n²
Mv²
Paschen
KEM
Δν
f

Paschen spectral series

α	1,886	n3-4
β	1,289	n3-5
γ	1,100	n3-6
δ	1,011	n3-7
ε	960	n3-8

IR spectrum

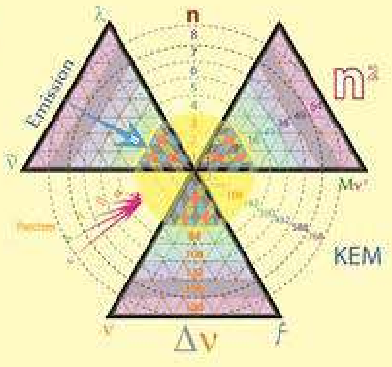
λ
n
Emission
n²
Mv²
Paschen
KEM
Δν
f

M shell

$\frac{\Delta \nu}{\text{KEM}}$

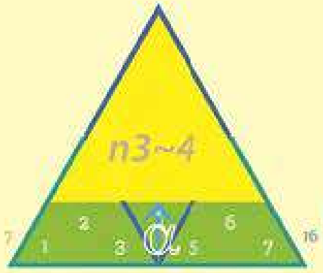

84

192

1 quantum level jump

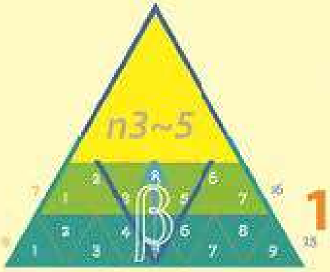

192

300

2 quantum level jump

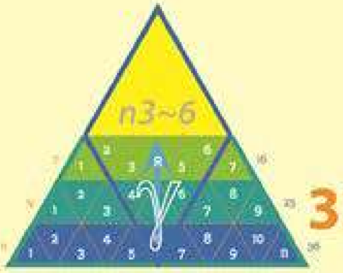

324

432

3 quantum level jump

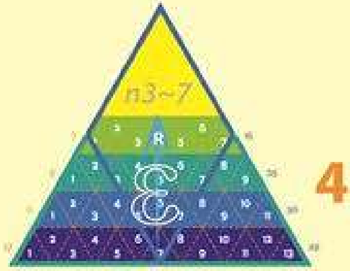

480

588

4 quantum level jump

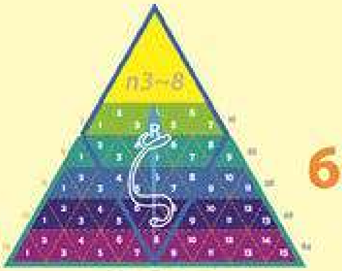

660

768

5 quantum level jump

Copyright ABRAHAM [2008] - all rights reserved

Paschen series energy momenta

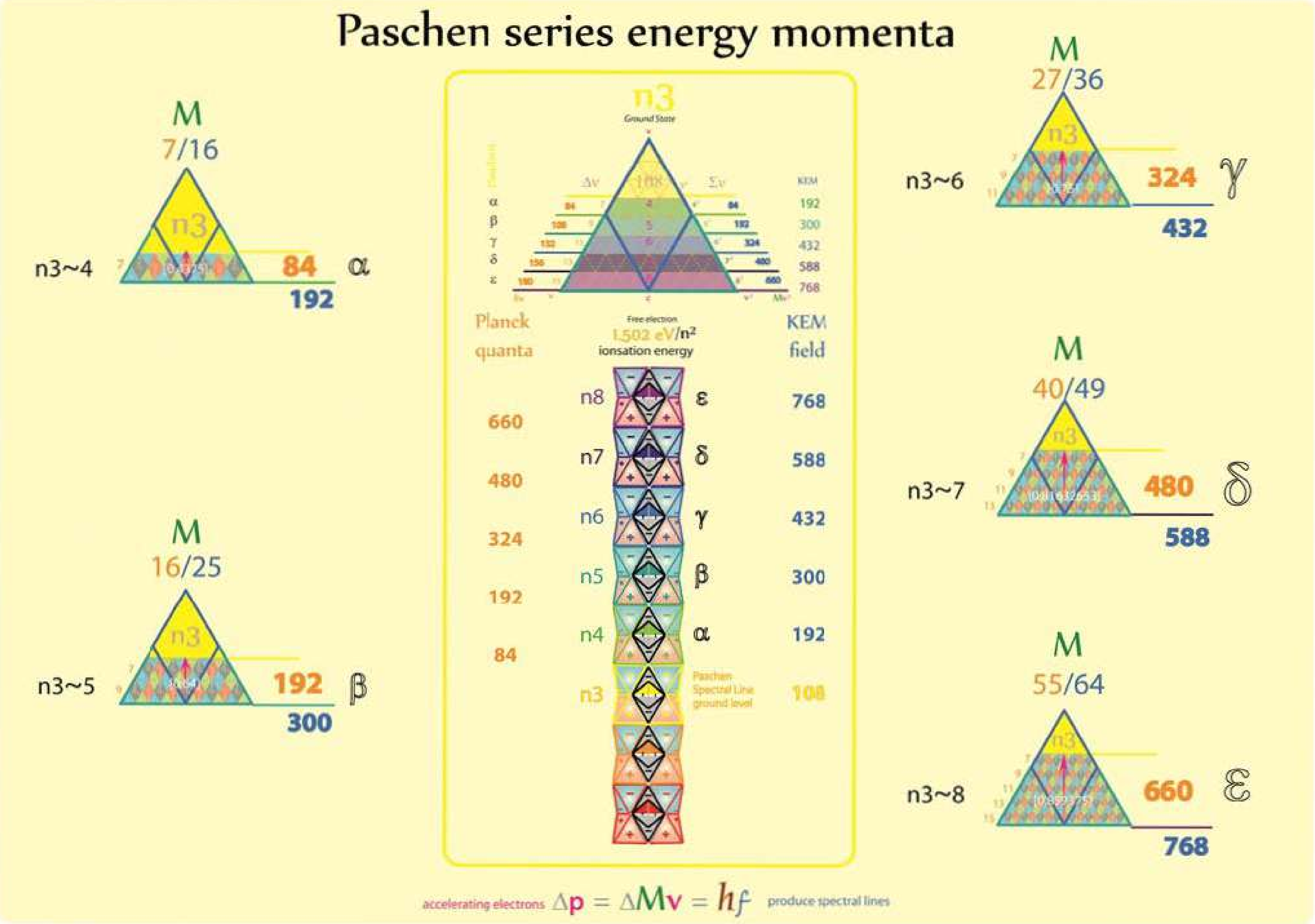
M
7/16
n3
n3~4
84
α
192
M
16/25
n3
n3~5
192
β
300
n3
Ground State
Δv
Σv
KEM
α
β
γ
δ
ε
84
108
132
156
180
84
192
324
480
660
192
300
432
588
768
Planck
quanta
Free electron
1.502 eV/n²
ionsation energy
KEM
field
n8
ε
768
660
n7
δ
588
480
n6
γ
432
324
n5
β
300
192
n4
α
192
84
n3
Paschen
Spectral Line
ground level
108
M
27/36
n3
n3~6
324
γ
432
M
40/49
n3
n3~7
480
δ
588
M
55/64
n3
n3~8
660
ε
768

accelerating electrons $\Delta p = \Delta M v = hf$ produce spectral lines

Tetryonics 32.03 - Paschen series energy momenta

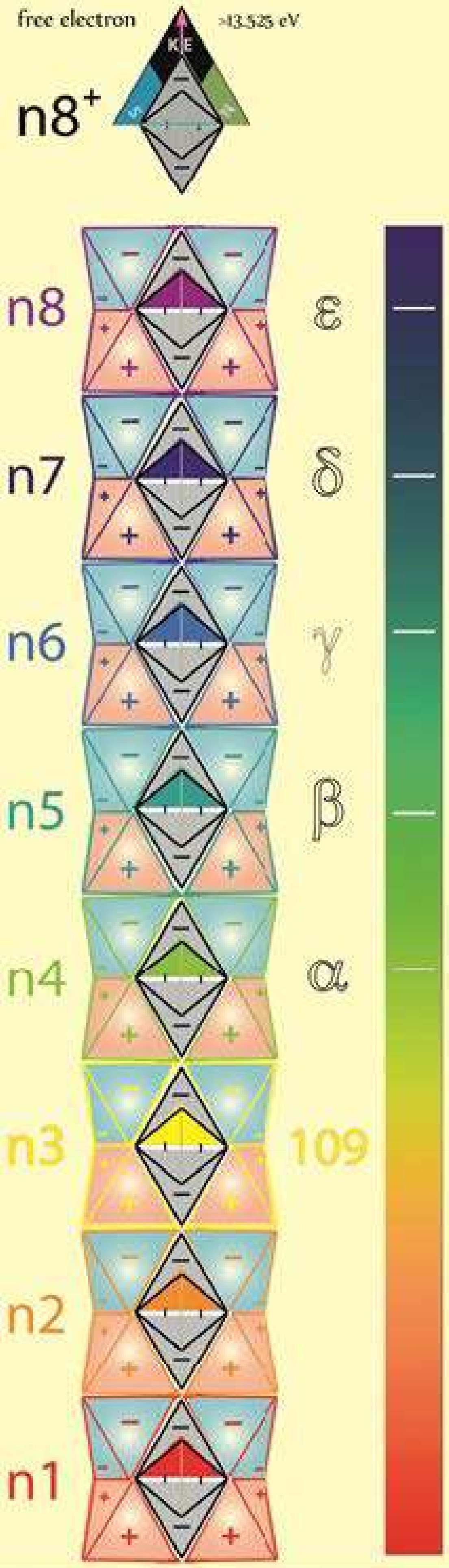

Paschen Series

$$\Delta h\nu = \text{KEM} = hcR$$

Tetryonic

1.502 eV

Rydberg

55/64

.0954

λ	960.5057935	nm
$\tilde{\nu}$	1,041,118.135	m-1
f	3.121193646 x 10^{14}	Hz
E	1.29147712	eV
Ω	237.9523928	m²/s

$R\left(\frac{1}{9}-\frac{1}{64}\right)$

.09548

40/49

.0907

λ	1,011.157466	nm
$\tilde{\nu}$	988,965.6492	m-1
f	2.9464844428 x 10^{14}	Hz
E	1.226783461	eV
Ω	301.1373822	m²/s

$R\left(\frac{1}{9}-\frac{1}{49}\right)$

.0907

27/36

.8033

λ	1,100.579555	nm
$\tilde{\nu}$	908,612.1902	m-1
f	2.723950819 x 10^{14}	Hz
E	1.127107305	eV
Ω	329.94545	m²/s

$R\left(\frac{1}{9}-\frac{1}{36}\right)$

.0833

16/25

.0711

λ	1,289.741666	nm
$\tilde{\nu}$	775,349.069	m-1
f	2.324438032 x 10^{14}	Hz
E	0.961798233	eV
Ω	386.9343240	m²/s

$R\left(\frac{1}{9}-\frac{1}{25}\right)$

.0711

7/16

.0486

λ	1,886.707809	nm
$\tilde{\nu}$	530,023.7776	m-1
f	1.588971311 x 10^{14}	Hz
E	0.657479261	eV
Ω	565.6207715	m²/s

$R\left(\frac{1}{9}-\frac{1}{16}\right)$

.0486

$$m.\begin{bmatrix}\Omega/v\\ v/\Omega\\ c^2/\Omega\\ v^2/c^2\end{bmatrix}$$

λ	825.4346663	nm	h/p
$\tilde{\nu}$	1,211,482.92	m-1	p/h
f	3.631934425 x 10^{14}	Hz	E/h
E	1.50280974	eV	hf
Ω	247.4590815	m²/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Copyright ABRAHAM [2008] - all rights reserved

Copyright ABRAHAM [2008] - all rights reserved

Brackett spectral lines

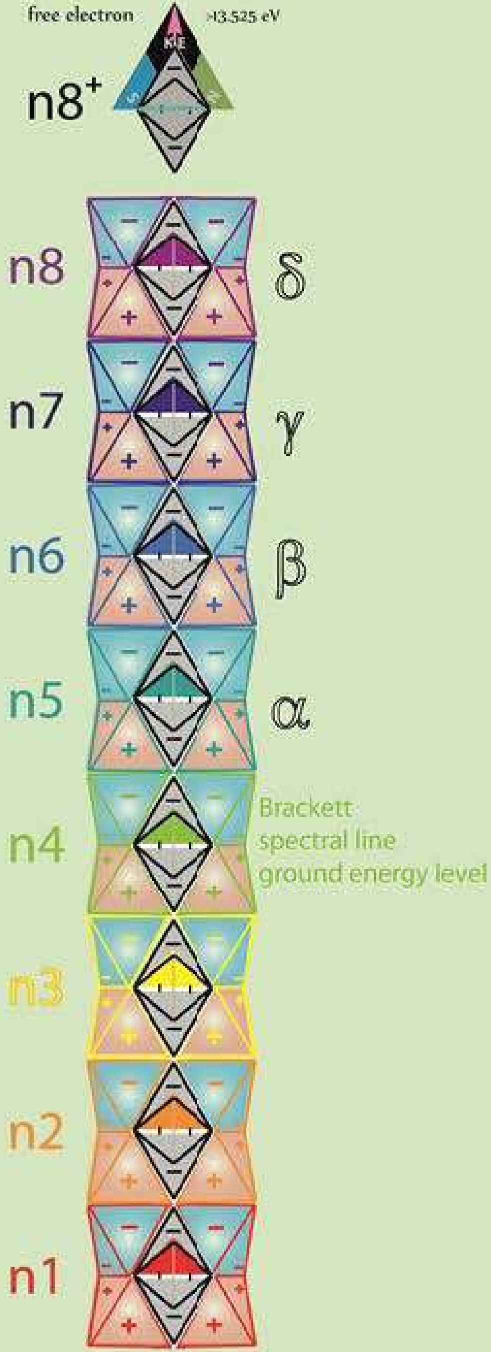
free electron
>13.525 eV
n8+
KE
n8
δ
n7
γ
n6
β
n5
α
n4
Brackett
spectral line
ground energy level
n3
n2
n1

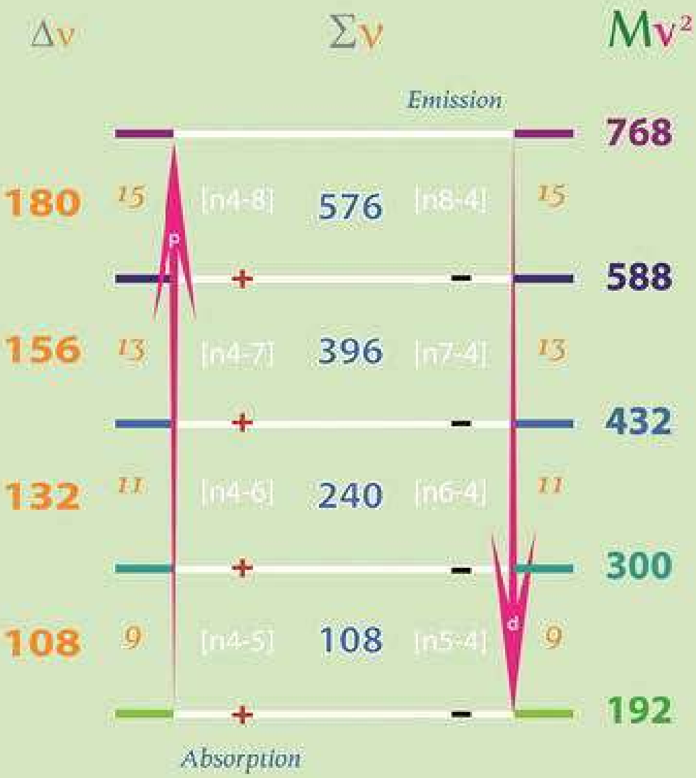
Δv
Σv
Mv²
Emission
768
180
15
[n4-8]
576
[n8-4]
15
588
156
13
[n4-7]
396
[n7-4]
13
432
132
11
[n4-6]
240
[n6-4]
11
300
108
9
[n4-5]
108
[n5-4]
9
192
Absorption

n4

Ground State

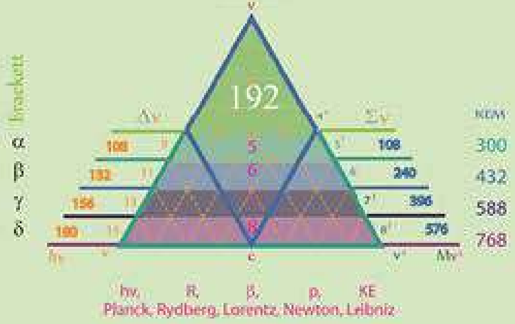
brackett
192
Δv
Σv
KEM
α
β
γ
δ
108
132
156
180
108
240
396
576
300
432
588
768
5
6
7
8
hv,
R,
β,
p,
KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-0.84533478 eV

$$KEM = \frac{hf}{n^2}$$

v
n4
KE
12
1.2e20

N shell

1,467.43 nm

$$KEM = \frac{hcR}{n^2}$$

$$m.\begin{bmatrix}\Omega/v \\ v/\Omega \\ c^2/\Omega \\ v^2/c^2\end{bmatrix}$$

λ	1,467.439407	nm	h/p
ṽ	681,459.1426	m-1	p/h
ƒ	2.042963114 x 10¹⁴	Hz	E/h
E	0.84533478	eV	hƒ
Ω	[illegible]	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

Tetryonics 33.01 - Brackett spectral lines

Copyright ABRAHAM [2008] - all rights reserved

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhν] W bosons

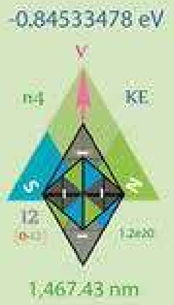
-0.84533478 eV
v
n4
KE
S
N
12
1.2e20
1,467.43 nm

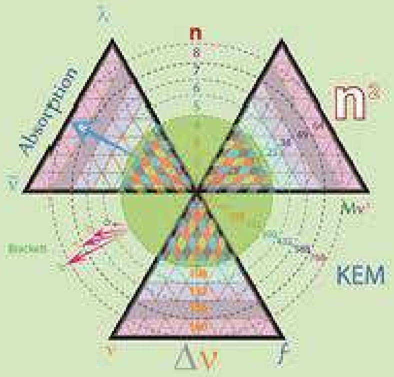
λ
n
8
7
6
5
Absorption
n²
Mv²
Brackett
KEM
v
Δν
f

Brackett spectral series

α	4,076	n4-5
β	2,641	n4-6
γ	2,178	n4-7
δ	1,956	n4-8

IR spectrum

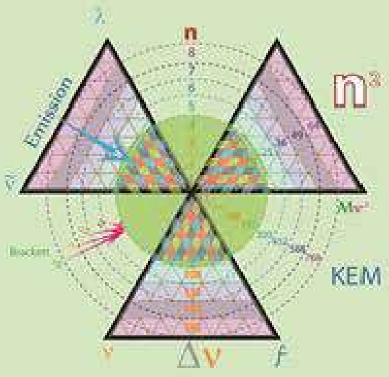
λ
n
8
7
6
5
Emission
n³
Mv²
Brackett
KEM
v
Δν
f

N shell

Δν

KEM

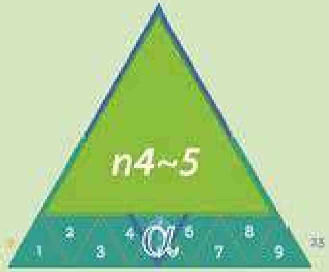
n4~5
α
1
2
3
4
5
6
7
8
9

108

300

1 quantum level jump

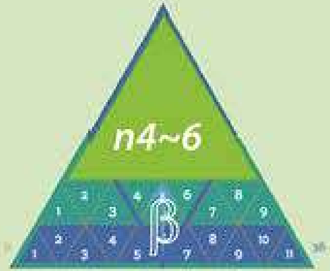

240

432

2 quantum level jump

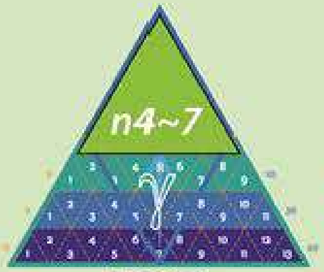

396

588

3 quantum level jump

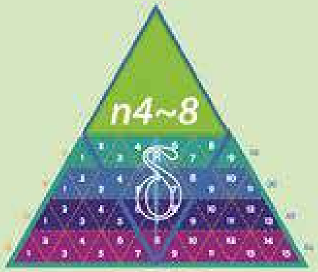

576

768

4 quantum level jump

Tetryonics 33.02 - Brackett quantum transitions

Brackett series energy momenta

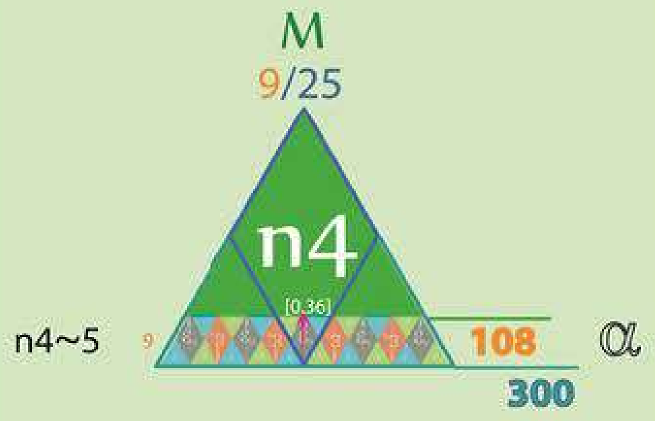
M
9/25
n4
[0.36]
9
n4~5
108
300
α

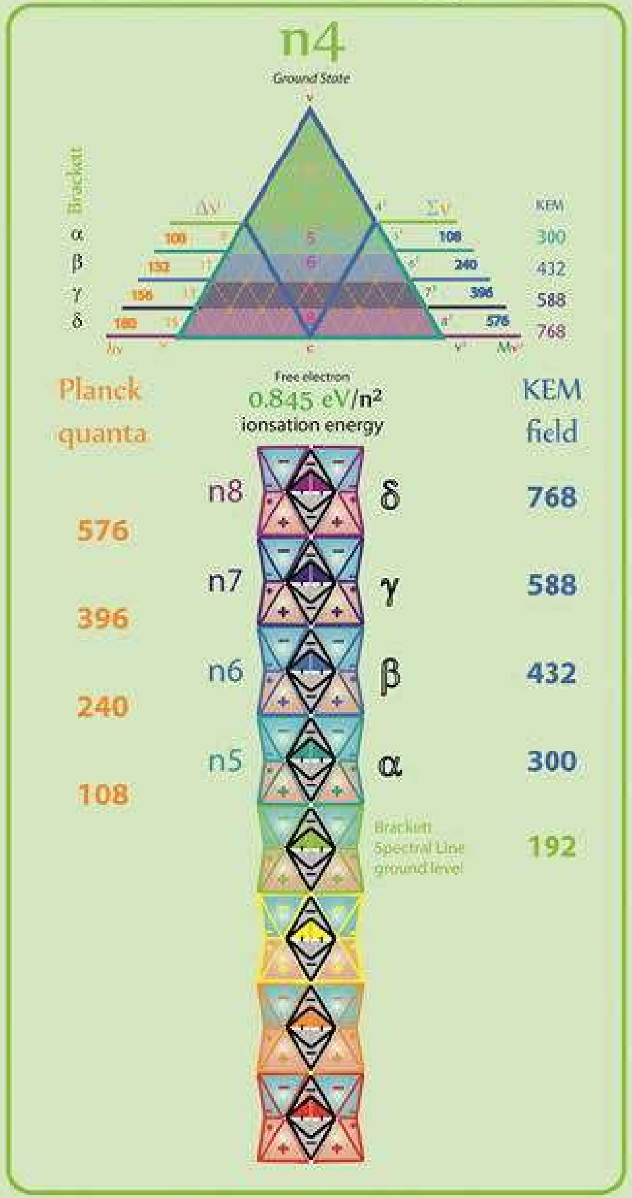
n4
Ground State
Brackett
α
β
γ
δ
Δv
Σv
100
132
156
180
108
240
396
576
KEM
300
432
588
768
Free electron
0.845 eV/n²
ionsation energy
Planck
quanta
KEM
field
576
396
240
108
n8
n7
n6
n5
δ
γ
β
α
768
588
432
300
192
Brackett
Spectral Line
ground level

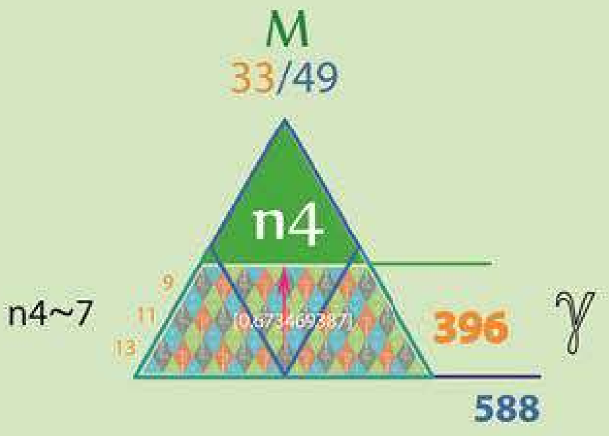
M
33/49
n4
n4~7
9
11
13
[0.673469387]
396
588
γ

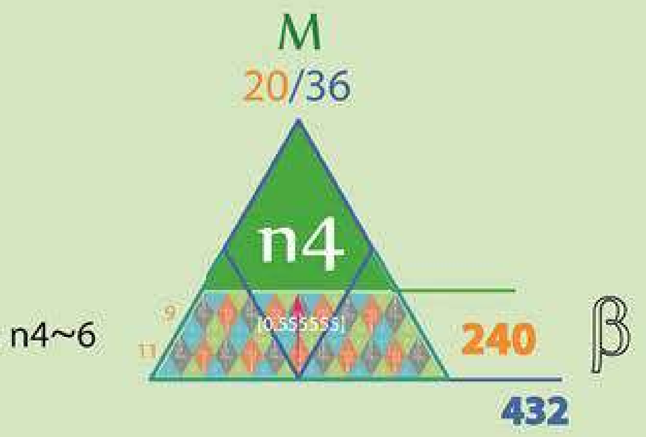
M
20/36
n4
n4~6
9
11
[0.555555]
240
432
β

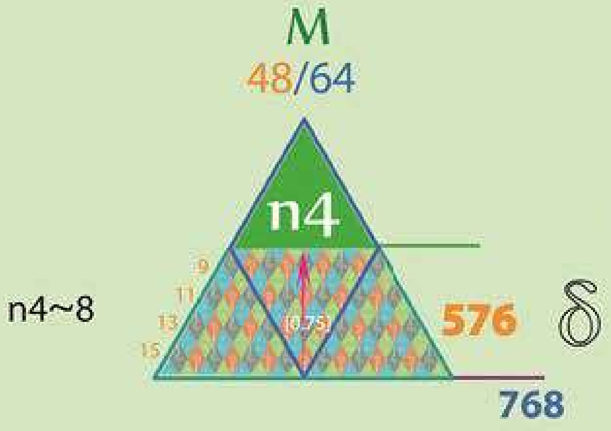
M
48/64
n4
n4~8
9
11
13
15
[0.75]
576
768
δ

accelerating electrons $\Delta \mathbf{p} = \Delta M \mathbf{v} = hf$ produce spectral lines

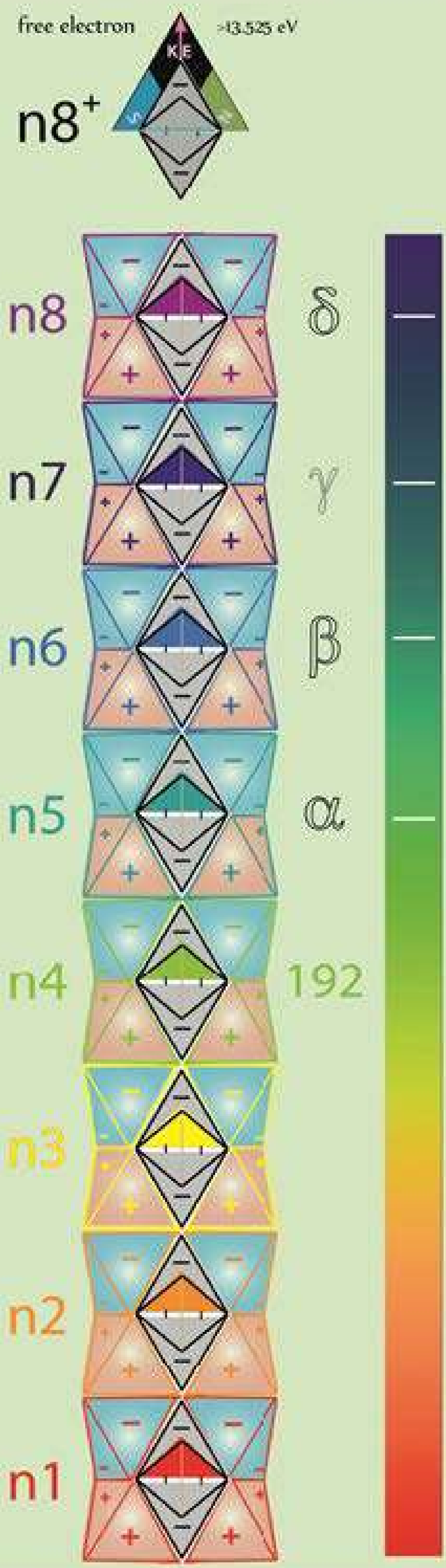

Brackett Series

$$\Delta h\nu \;=\; \text{KEM} \;=\; hcR$$

Tetryonics

0.845 eV

Rydberg

48/64
.0468

λ	1,956.585876	nm
$\tilde{\nu}$	511,094.357	m-1
f	1.532222335 x 10¹⁴	Hz
E	0.634001085	eV
Ω	556.569689	m²/s

$$R\left(\frac{1}{16}-\frac{1}{64}\right)$$
.0468

33/49
.04209

λ	2,178.92518	nm
$\tilde{\nu}$	458,941.8716	m-1
f	1.375873118 x 10¹⁴	Hz
E	0.569307096	eV
Ω	633.2255354	m²/s

$$R\left(\frac{1}{16}-\frac{1}{49}\right)$$
.04209

20/36
.0347

λ	2,641.390932	nm
$\tilde{\nu}$	378,588.4126	m-1
f	1.134979508 x 10¹⁴	Hz
E	0.469630433	eV
Ω	791.5606021	m²/s

$$R\left(\frac{1}{16}-\frac{1}{36}\right)$$
.0347

9/25
.0225

λ	4,076.220574	nm
$\tilde{\nu}$	245,325.2914	m-1
f	7.35466721x 10¹³	Hz
E	0.30432052	eV
Ω	1.222010165	m²/s

$$R\left(\frac{1}{16}-\frac{1}{25}\right)$$
.0225

$$m.\begin{bmatrix}\Omega/v\\ v/\Omega\\ c^2/\Omega\\ v^2/c^2\end{bmatrix}$$

λ	1,467.439407	nm	h/p
$\tilde{\nu}$	681,459.1426	m-1	p/h
f	2.042963114 x 10¹⁴	Hz	E/h
E	0.84533478	eV	hf
Ω	439.9272667	m²/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Copyright ABRAHAM [2008] - all rights reserved

Pfund spectral lines

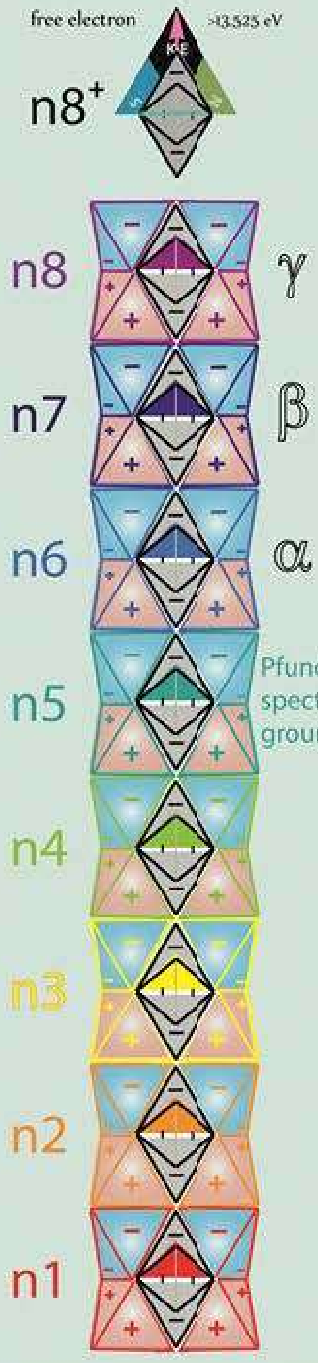
free electron
>13.525 eV
KE
n8+
n8
γ
n7
β
n6
α
n5
n4
n3
n2
n1

Pfund
spectral line
ground energy level

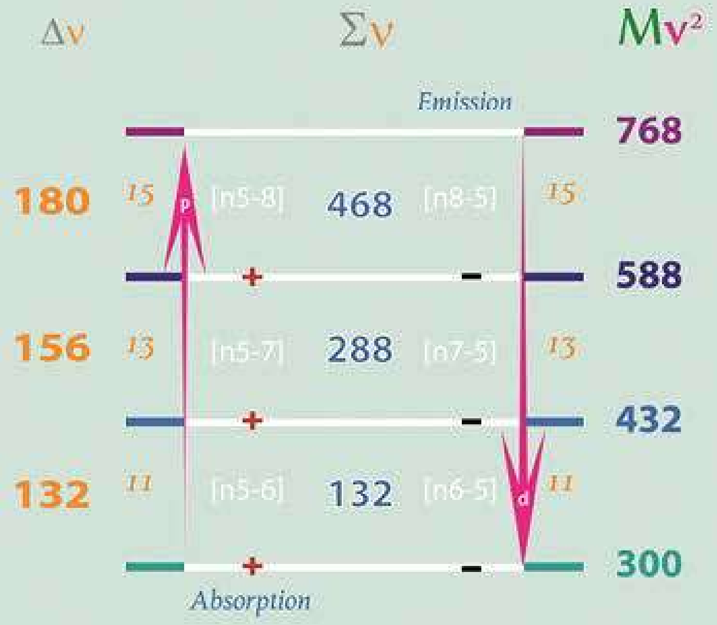
Δv
Σv
Mv²
Emission
768
180
15
[n5-8]
468
[n8-5]
15
588
156
13
[n5-7]
288
[n7-5]
13
432
132
11
[n5-6]
132
[n6-5]
11
300
Absorption

n5

Ground State

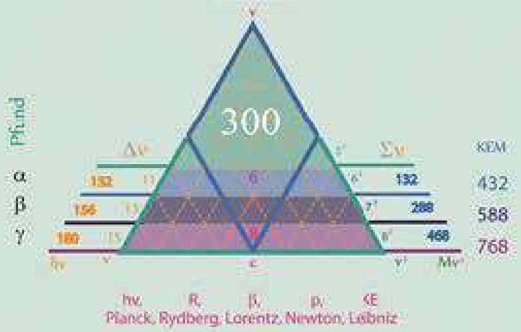
Pfund
α
β
γ
300
Δv
Σv
KEM
132
156
180
132
288
468
432
588
768
hv
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-0.5410115 eV

$$KEM = \frac{hf}{n^2}$$

v
n5
KE
12
1.2e20

O shell

2,292.87 nm

$$KEM = \frac{hcR}{n^2}$$

$$m.\begin{bmatrix} \Omega/v \\ v/\Omega \\ c^2/\Omega \\ v^2/c^2 \end{bmatrix}$$

λ	2,292.874073	nm	h/p
ṽ	436,133.8513	m-1	p/h
f	1.307496393 x 10^{14}	Hz	E/h
E	0.541011506	eV	hf
Ω	687.3863542	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

Tetryonics 34.01 - Pfund spectral lines

Copyright ABRAHAM [2008] - all rights reserved

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

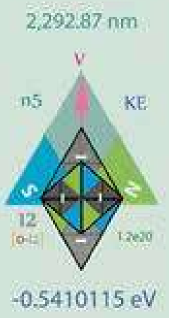
2,292.87 nm
v
n5
KE
12
[o-12]
1.2e20
-0.5410115 eV

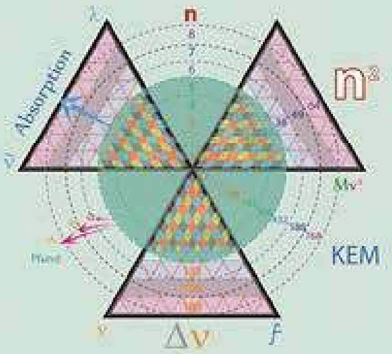
λ
n
Absorption
n²
Mv²
KEM
Δv
f

Pfund spectral series

α	7,503	n5-6
β	4,681	n5-7
γ	3,762	n5-8

IR spectrum

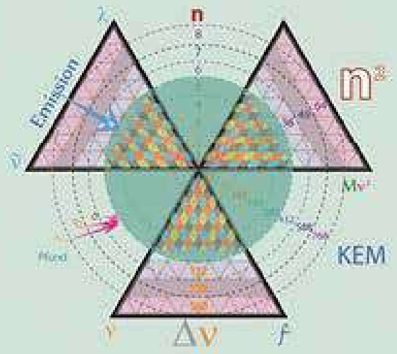
λ
n
Emission
n²
Mv²
KEM
Δv
f

O shell

Δv
KEM

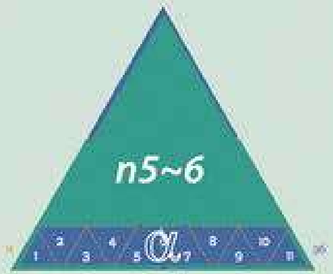

132

432

1 quantum level jump

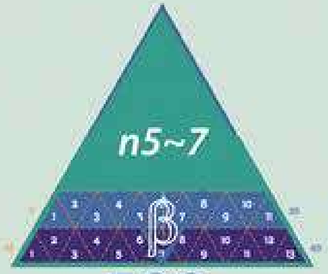

288

588

2 quantum level jump

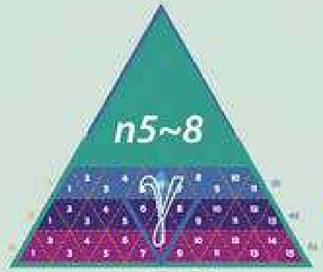

468

768

3 quantum level jump

Tetryonics 34.02 - Pfund quantum transitions

Pfund series energy momenta

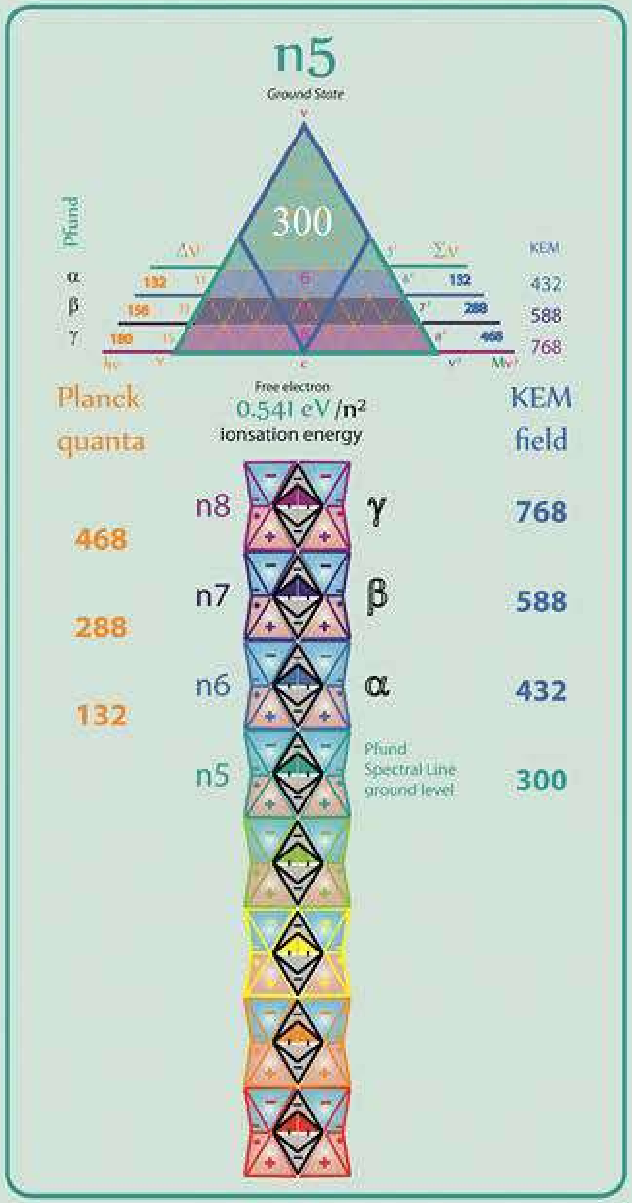
n5
Ground State
Pfund
α
β
γ
Δv
Σv
KEM
132
432
588
768
300
hv
Mv²

Free electron
0.541 eV /n²
ionsation energy

Planck quanta			KEM field
468	n8	γ	768
288	n7	β	588
132	n6	α	432
	n5	Pfund Spectral Line ground level	300

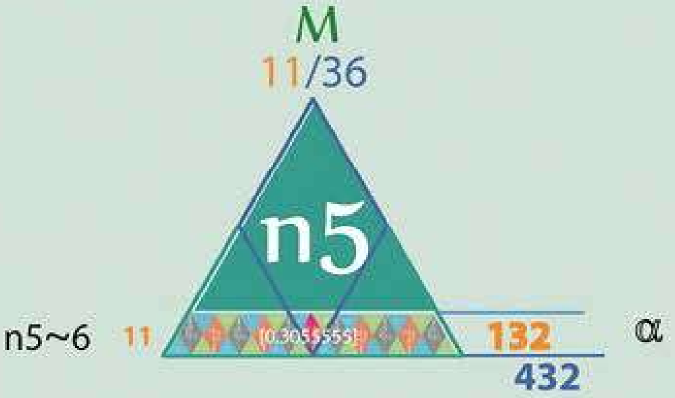
M
11/36
n5~6
11
n5
[0.3055555]
132
432
α

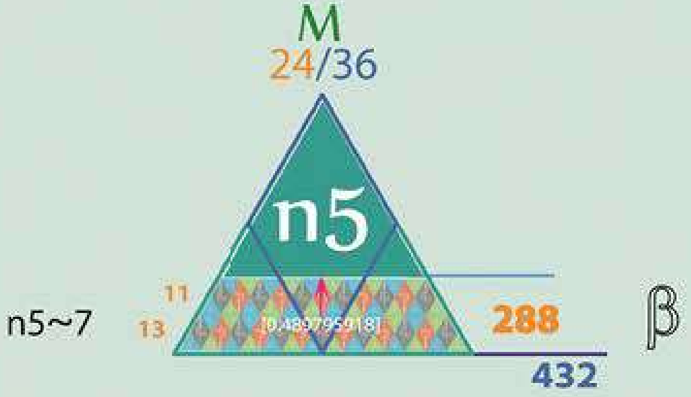
M
24/36
n5~7
11
13
n5
[0.489795918]
288
432
β

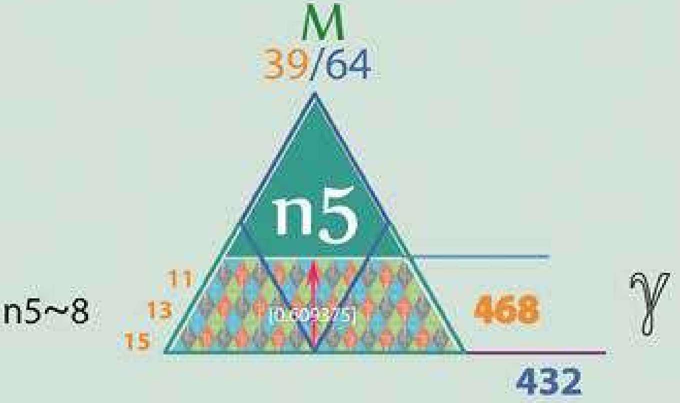
M
39/64
n5~8
11
13
15
n5
[0.609375]
468
432
γ

accelerating electrons $\Delta p = \Delta Mv = hf$ produce spectral lines

Tetryonics 34.03 - Pfund series energy momenta

Copyright ABRAHAM [2008] - all rights reserved

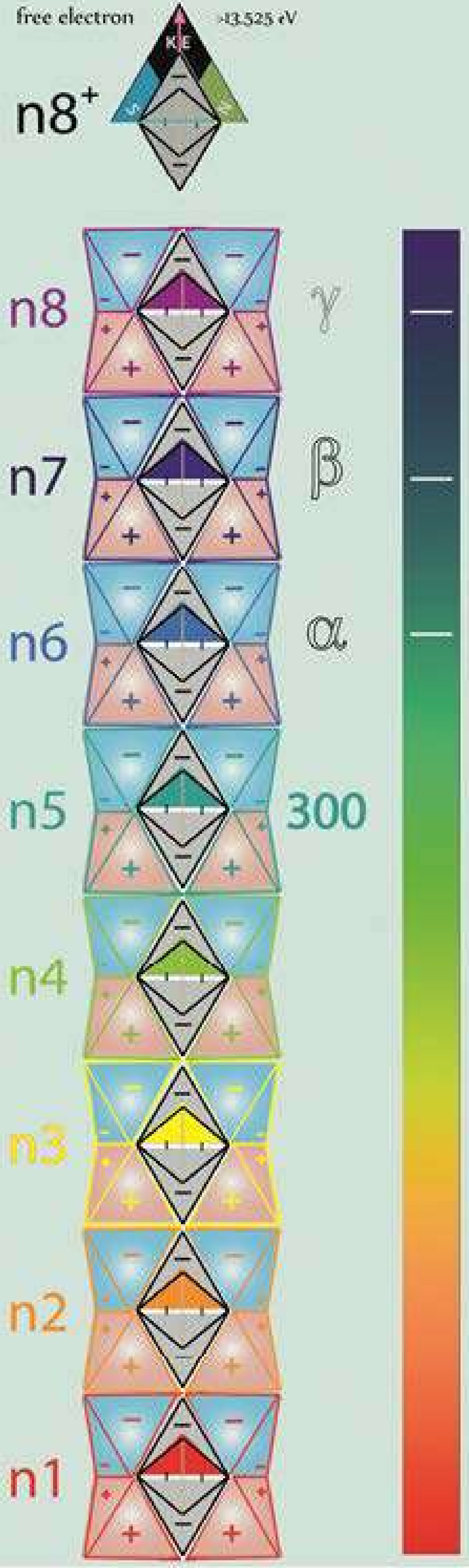

Pfund Series

Δhv = KEM = hcR

Tetryonic

0.541 eV

Rydberg

39/64

.0243

λ	3,762.665146	nm
ṽ	265,769.0656	m-1
f	7.967556145 x 10^{13}	Hz
E	0.329678886	eV
Ω	1,328.018633	m/s

$R\left(\frac{1}{25}-\frac{1}{64}\right)$

.0243

24/49

.0195

λ	4,681.294566	nm
ṽ	213,616.5802	m-1
f	6.404063965 x 10^{13}	Hz
E	0.264985227	eV
Ω	1,403.415807	m/s

$R\left(\frac{1}{25}-\frac{1}{49}\right)$

.01959

11/36

.0122

λ	7,503.951512	nm
ṽ	133,263.1212	m-1
f	3.995127867 x 10^{13}	Hz
E	0.165309071	eV
Ω	2,249.628068	m/s

$R\left(\frac{1}{25}-\frac{1}{36}\right)$

.0122

$m.\begin{bmatrix}\Omega/v\\ v/\Omega\\ c^2/\Omega\\ v/c^2\end{bmatrix}$

λ	2,292.874073	nm	h/p
ṽ	436,133.8513	m-1	p/h
f	1.307496393 x 10^{14}	Hz	E/h
E	0.541011506	eV	hf
Ω	687.3863542	m/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Copyright ABRAHAM [2008] - all rights reserved

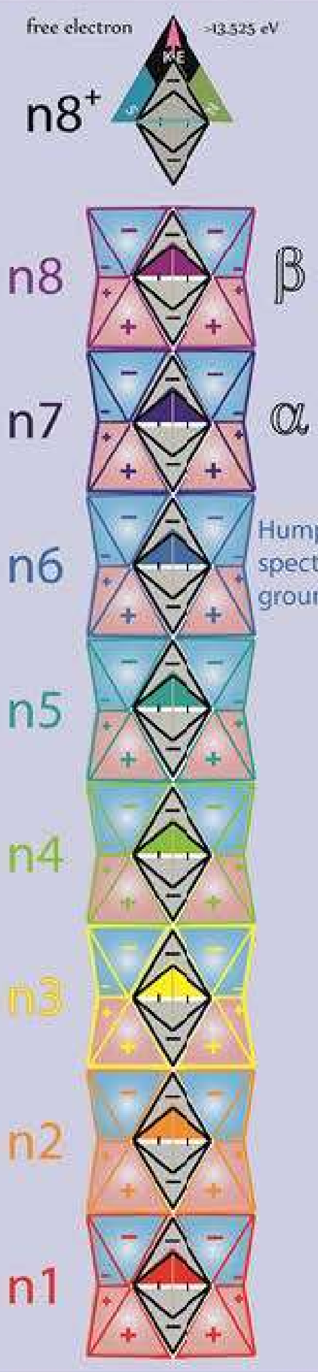
free electron
-13.525 eV
KE
n8+
n8
β
n7
α
n6
n5
n4
n3
n2
n1

Humphreys
spectral line
ground energy level

Humphreys spectral lines

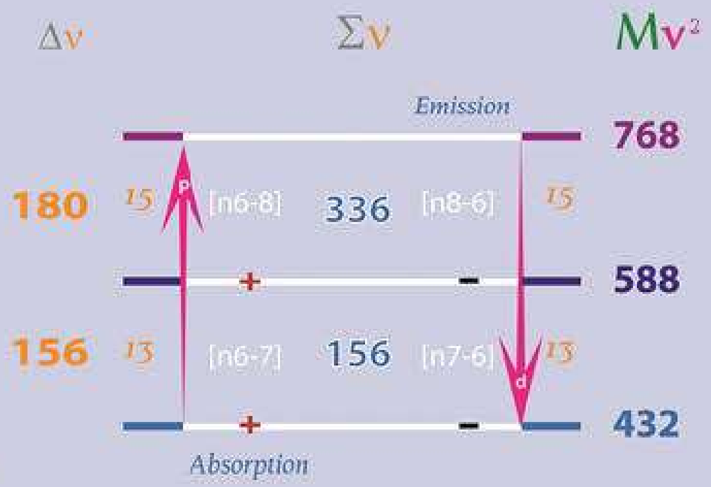
Δv
Σv
Mv²
Emission
768
180
15
p
[n6-8]
336
[n8-6]
15
588
156
13
[n6-7]
156
[n7-6]
13
d
432
Absorption

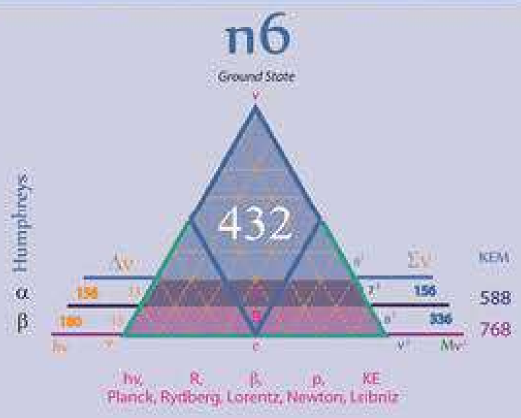
n6
Ground State
Humphreys
α
β
432
Δv
Σv
KEM
156
156
588
180
336
768
hν
c
Mv²
hν, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-0.375702435 eV

$$KEM = \frac{hf}{n^2}$$

v
n6
KE
12
1.2e20

P shell

3,301.73 nm

$$KEM = \frac{hcR}{n^2}$$

m.

Ω/v	λ	3,301.738665	nm	h/p
v/Ω	ṽ	302,870.7301	m-1	p/h
c²/Ω	f	$9.079836062 \times 10^{13}$	Hz	E/h
v²/c²	E	0.375702435	eV	hf
	Ω	989.8363501	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave functions

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

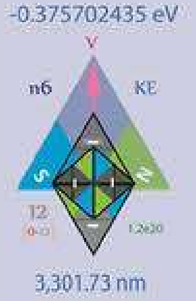
-0.375702435 eV
v
n6
KE
S
N
12
3,301.73 nm

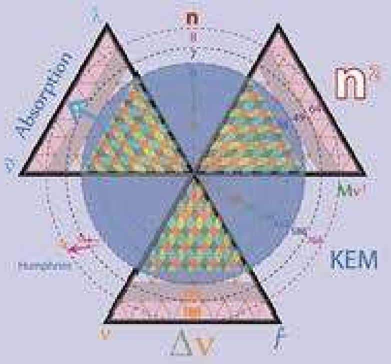
λ
n
Absorption
n²
Mv²
KEM
Humphries
Δv
f

Humphries spectral series

α	12,445	n6-7
β	7,546 IR spectrum	n6-8

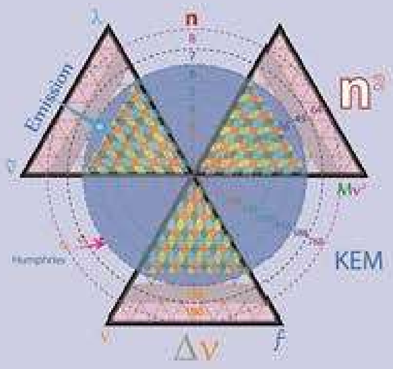
λ
n
Emission
n²
Mv²
KEM
Humphries
Δv
f

P shell

Δν

KEM

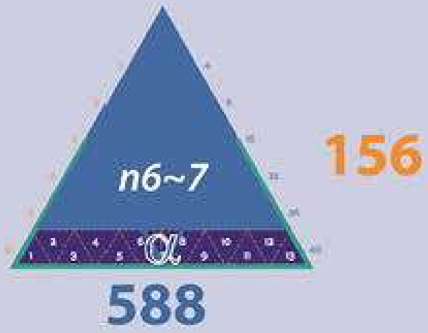
n6~7
α
156
588

1 quantum level jump

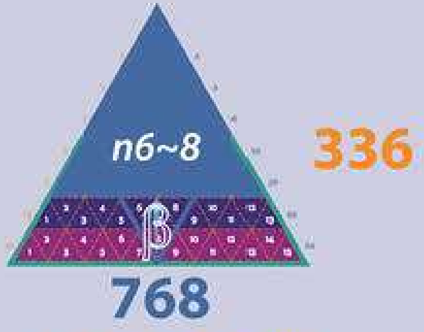
n6~8
β
336
768

2 quantum level jump

Tetryonics 35.02 - Humphreys quantum transitions

Humphreys series energy momenta

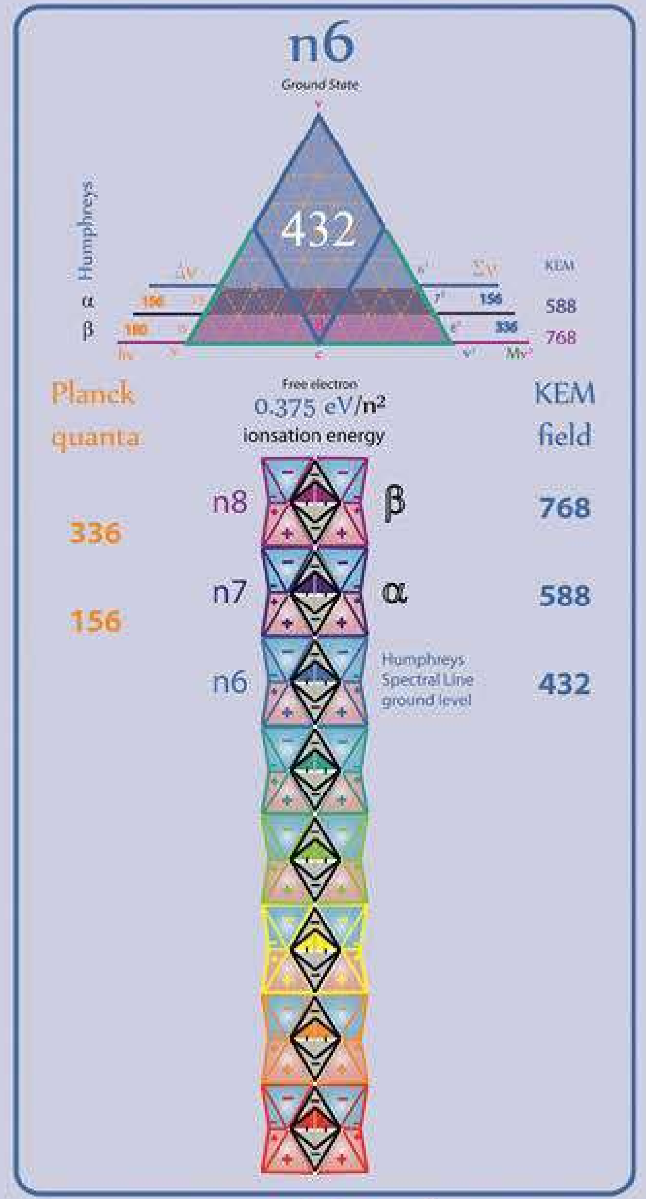
n6
Ground State
432
Humphreys
α
β
156
KEM
588
768
Planck
quanta
Free electron
0.375 eV/n²
ionsation energy
KEM
field
n8
β
768
336
n7
α
588
156
n6
Humphreys
Spectral Line
ground level
432

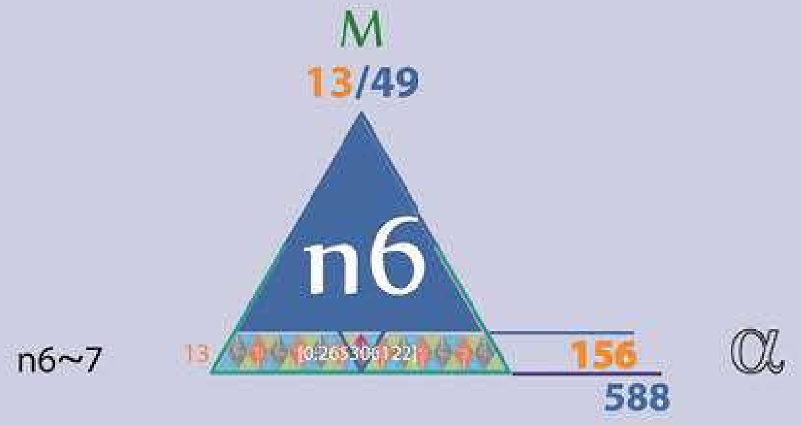
M
13/49
n6
n6~7
13
[0.265306122]
156
588
α

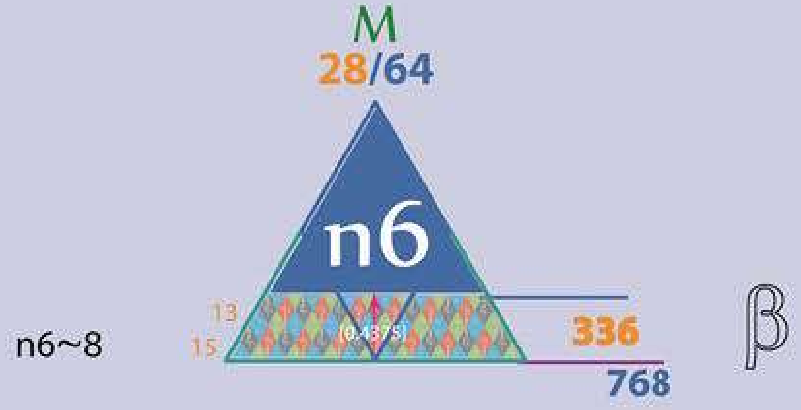
M
28/64
n6
n6~8
13
15
[0.4375]
336
768
β

accelerating electrons $\Delta p = \Delta Mv = hf$ produce spectral lines

Tetryonics 35.03 - Humphreys series energy momenta

Copyright ABRAHAM [2008] - all rights reserved

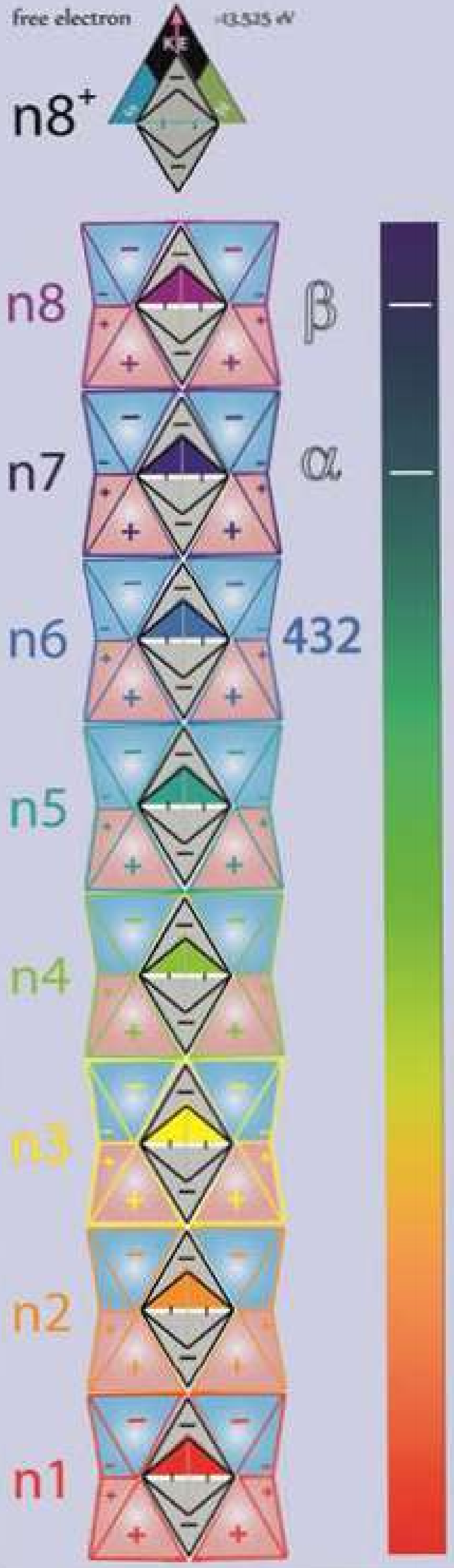

Humphreys Series

$$\Delta h\nu = \text{KEM} = hcR$$

tetryonic

0.375 eV

Rydberg

28/64

.01215

λ	7,546.831215	nm
ṽ	132,505.9444	m-1
f	3.972428277 x 10	Hz
E	0.164369815	eV
Ω	[illegible]	m²/s

$$R\left(\frac{1}{36} - \frac{1}{64}\right)$$

.01215

13/49

.00733

λ	12,445.01497	nm
ṽ	80,353.459	m-1
f	2.408993609x 10^{13}	Hz
E	0.099676156	eV
Ω	[illegible]	m²/s

$$R\left(\frac{1}{36} - \frac{1}{49}\right)$$

.0073

m. [c/v, v/c, [illegible]]

λ	3,301.738665	nm
ṽ	302,870.7301	m-1
f	5.079836062 x 10^{13}	Hz
E	0.375702435	eV
Ω	[illegible]	m²/s

h/p
p/h
E/h
hf

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Copyright ABRAHAM [2008] - all rights reserved

Un-named spectral line

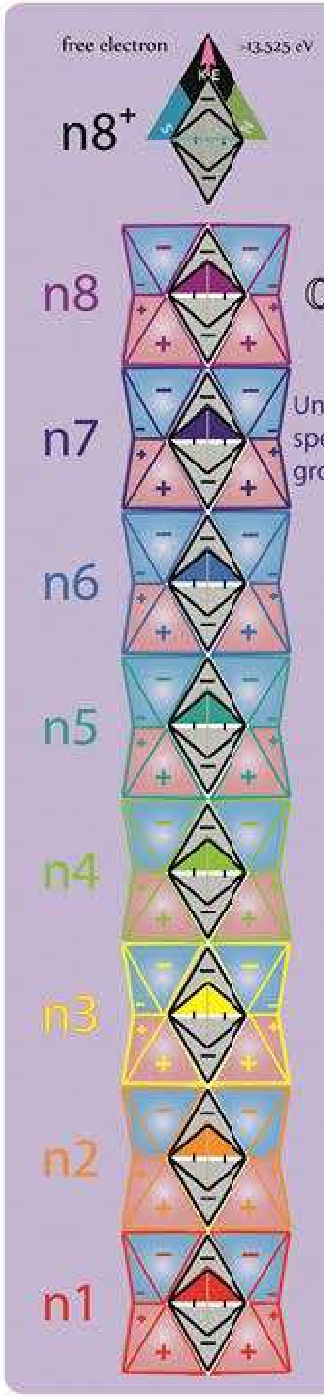
free electron
-13.525 eV
KE
$n8^+$
n8
α
n7
n6
n5
n4
n3
n2
n1

Un-named
spectral line
ground energy level

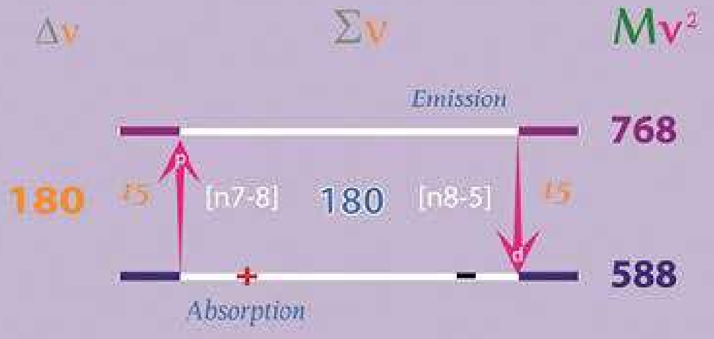
Δv
Σv
Mv²
Emission
768
180
ε5
[n7-8]
180
[n8-5]
ε5
588
Absorption

n7

Ground State

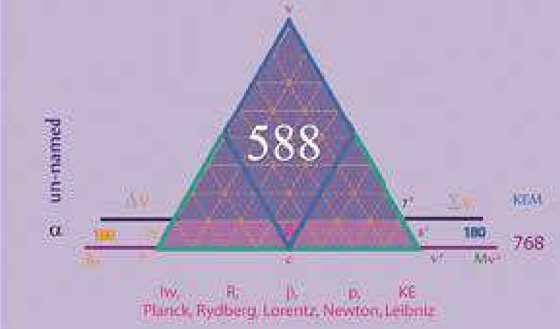
un-named
α
Δv
Σv
KEM
180
180
768
588
Mv²
hν, R, μ, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-0.276026782 eV

$$KEM = \frac{hf}{n^2}$$

v
n7
KE
12
1x20

Q shell

4,494.03 nm

$$KEM = \frac{hcR}{n^2}$$

m. [Ω/v, v/Ω, c²/Ω, v²/c]

λ	4,494.033183	nm	h/p
ṽ	222,517.2711	m-1	p/h
f	$6.67089964 \times 10^{13}$	Hz	E/h
E	0.276026278	eV	hf
Ω	1.347277254	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave functions

Tetryonics 36.01 - Un-named spectral line

This spectral line photo electronic transition can only release ***one*** *specific [Δhv] W boson of energy*

-0.276026782 eV
n7
KE
12
4,494.03 nm

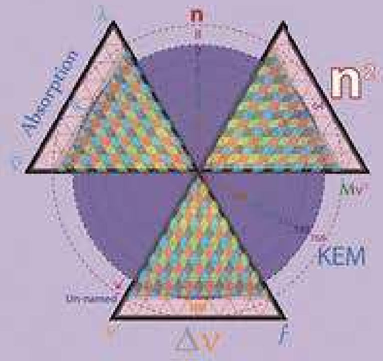
n
Absorption
n²
Mv²
KEM
Un-named
Δv
f

Un-named spectral series

α 19,174 n1-8

IR spectrum

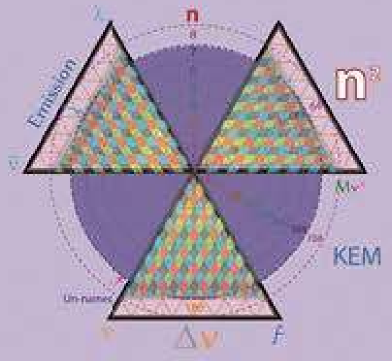
n
Emission
n²
Mv²
KEM
Un-named
Δv
f

Q shell

Δv

KEM

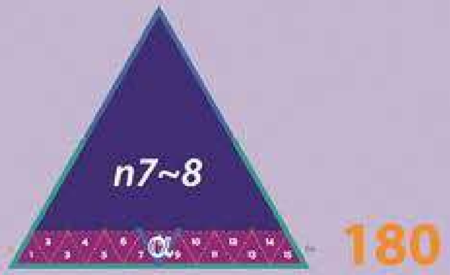

768

1 quantum level jump

Un-named series energy momenta

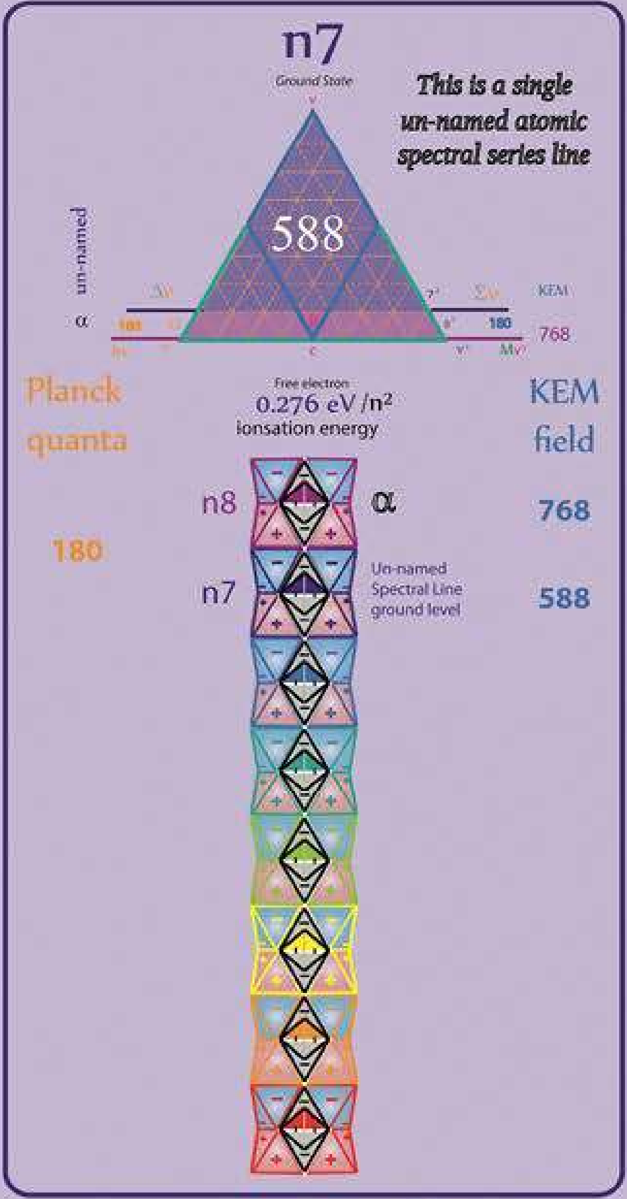

n7
Ground State
588
un-named
α
KEM
180
768
Mv²
c

This is a single un-named atomic spectral series line

Planck quanta

Free electron
0.276 eV /n^2
Ionsation energy

KEM field

n8 α 768

180

n7 Un-named Spectral Line ground level 588

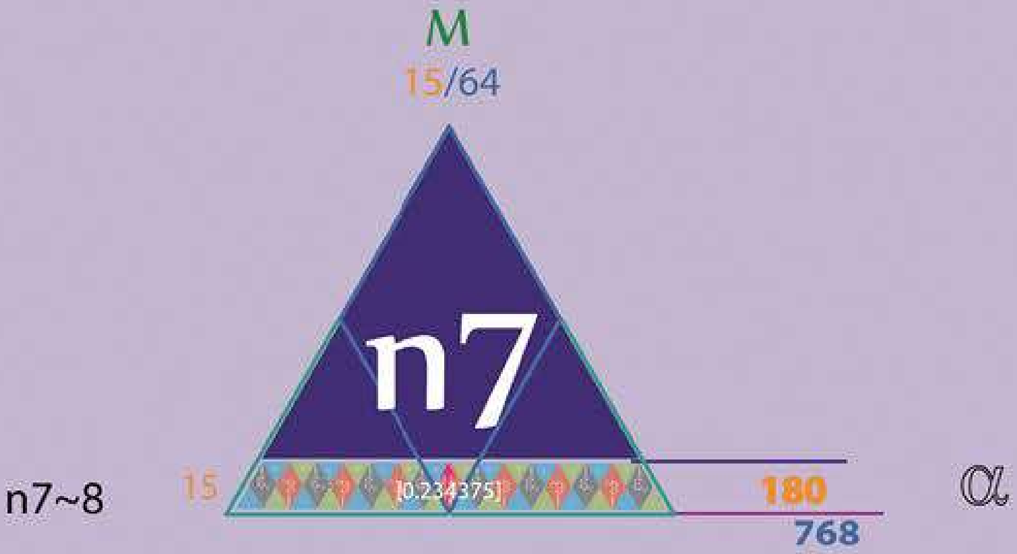

M
15/64
n7
n7~8
15
[0.234375]
180
768
α

accelerating electrons $\Delta p = \Delta M v = hf$ produce spectral lines

Tetryonics 36.03 - Un-named series energy momenta

Copyright ABRAHAM [2008] - all rights reserved

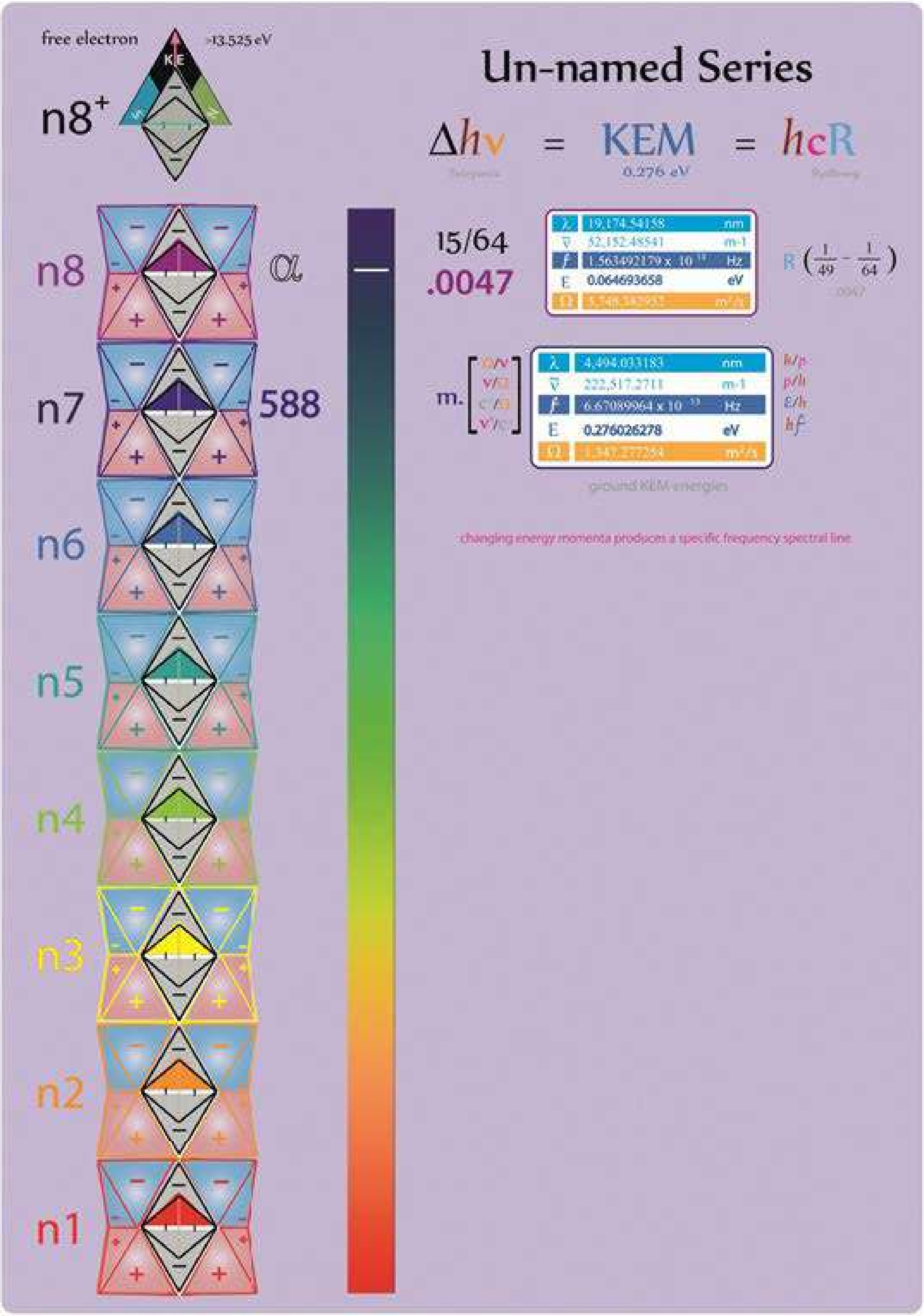

Un-named Series

$$\Delta h\nu = \text{KEM} = hcR$$

Tetryonics — 0.276 eV — Rydberg

15/64
.0047

λ	19,174.54158	nm
$\bar{\nu}$	52,152.48541	m-1
f	1.563492179 x 10^{13}	Hz
E	0.064693658	eV
Ω	5,548.382952	m²/s

$R\left(\frac{1}{49}-\frac{1}{64}\right)$

.0047

$$m.\begin{bmatrix} \Omega/\nu \\ \nu/\Omega \\ c/\Omega \\ \nu/c \end{bmatrix}$$

λ	4,494.033183	nm
$\bar{\nu}$	222,517.2711	m-1
f	6.67089964 x 10^{13}	Hz
E	0.276026278	eV
Ω	1,347.277254	m²/s

h/p
p/h
E/h
hf

ground KEM energies

changing energy momenta prodsom a specific frequency spectral line

Tetryonics 36.04 - Un-named series maths

KEM field geometry

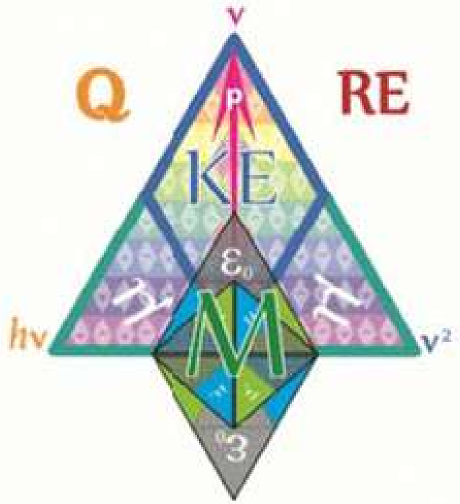
v
Q
RE
KE
ε0
M
μ0
μ0
hv
v²
3

The Lorentz corrected velocity of electron resulting from atomic energy level transitions

Bosons differential required to effect energy level transition

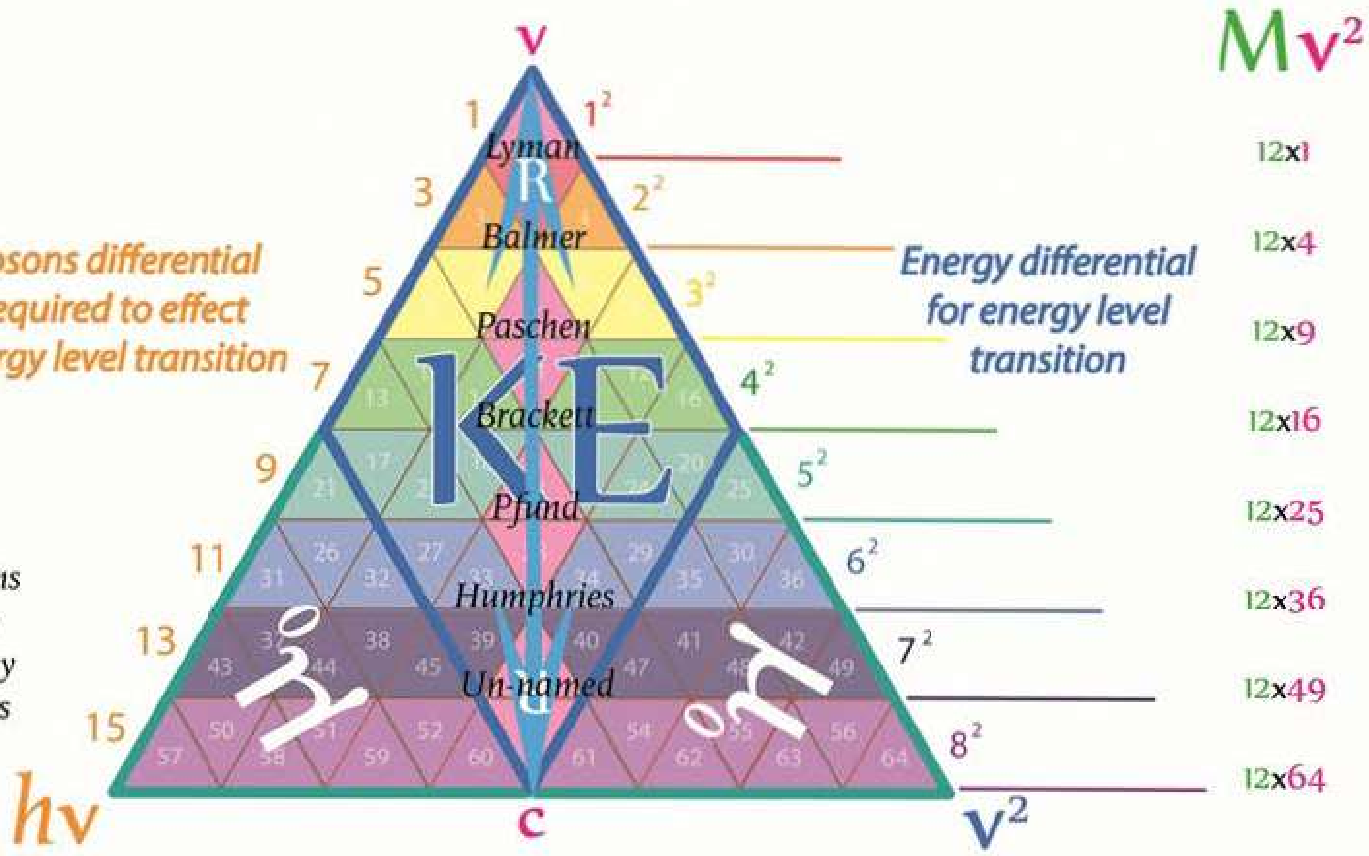
Mv²
v
1
1²
Lyman
R
3
2²
Balmer
5
3²
Paschen
7
4²
KE
Brackett
9
5²
Pfund
11
6²
Humphries
13
7²
Un-named
15
8²
μ0
μ0
hv
c
v²
12x1
12x4
12x9
12x16
12x25
12x36
12x49
12x64

Energy differential for energy level transition

All of the quantum jumps of electrons bound in a Hydrogen atom can be summerised in the Tetryonic geometry reflecting the energy level transitions

Ryberg's constant is reflective of the wavelengths and linear momentum of a transitioning electron's KEM field

Ryberg's formula is a mathematical description of Tetryonic energy geometry

$$Mv^2 = KEM = hcR_H$$

KEM field quanta

Photoelectron transitions

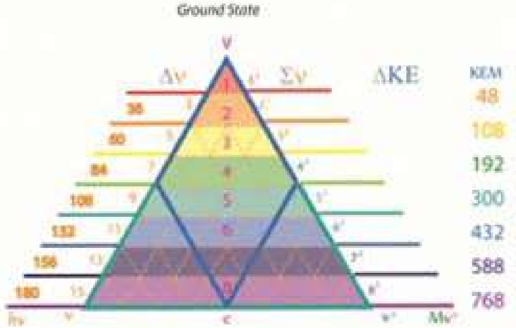
Ground State
Δv
Σv
ΔKE
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
1
2
3
4
5
6
7
8
hv
c
Mv²

Spectral Lines

Transition			Spectral series
n1~ n2	91.29	α	Lyman UV spectra
n3	91.30	β	
n4	91.32	γ	
n5	91.33	δ	
n6	91.36	ε	
n7	91.38	ϕ	
n8	91.41	η	
n2~ n3	367.00	α	Balmer Visible Spectra
n4	367.23	β	
n5	367.28	γ	
n6	367.54	δ	
n7	368.19	ε	
n8	368.59	ϕ	
n3~ n4	833.16	α	Paschen IR spectra
n5	834.84	β	
n6	835.68	γ	
n7	837.23	δ	
n8	839.02	ε	
n4~ n5	1499.68	α	Brackett
n6	1503.50	β	
n7	1507.87	γ	
n8	1512.92	δ	
n5~ n6	2381.53	α	Pfund
n7	2391.17	β	
n8	2402.25	γ	
n6~ n7	3499.27	α	Humphreys
n8	3520.12	β	
n7~ n8	4880.38	α	Un-named

Nuclear Quantum Levels

The kinetic energy levels of bound nuclear photon-electrons is determined by the energy levels of their Baryons

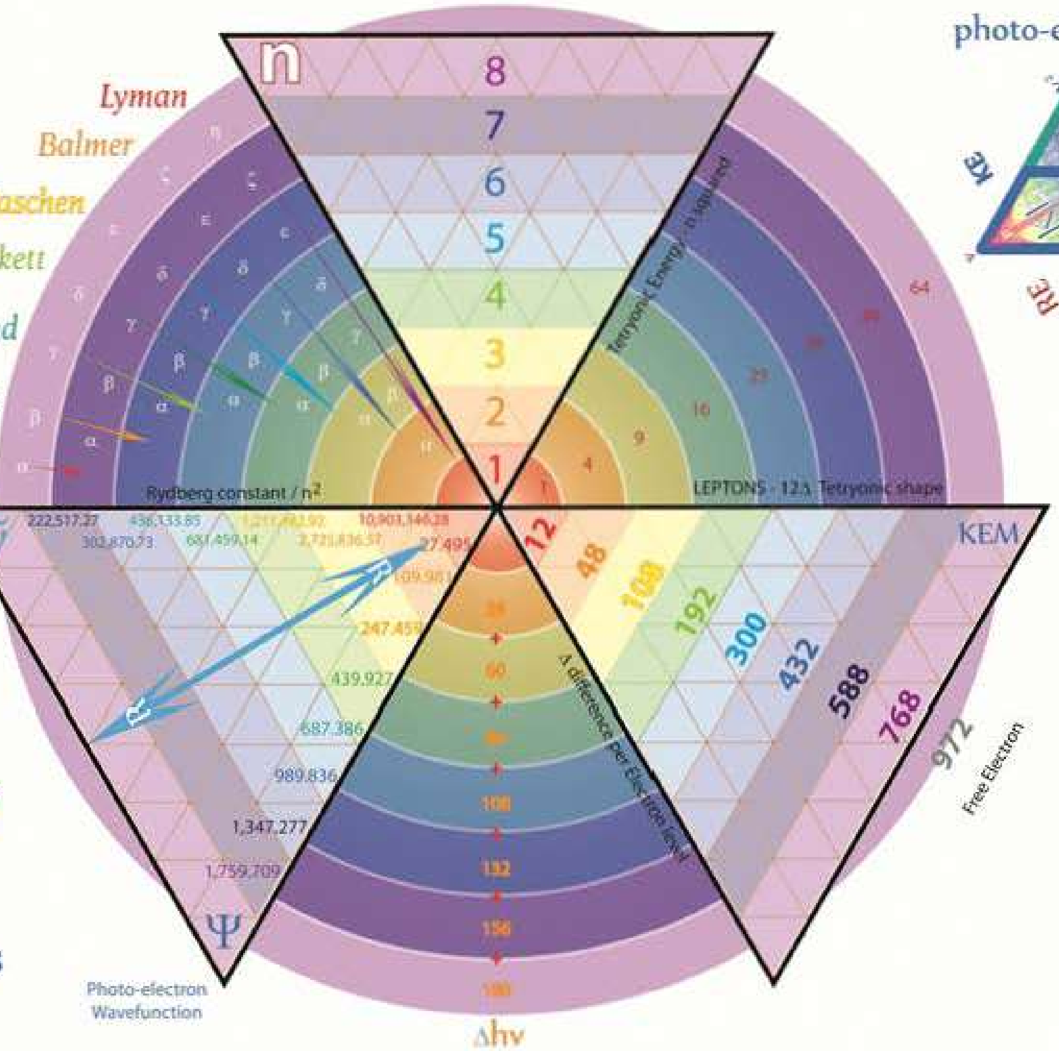
n
8
7
6
5
4
3
2
1
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
Tetryonic Energy / n squared
Rydberg constant / n²
LEPTONS - 12Δ Tetryonic shape
ṽ
Photon Wave number
222,517.27
436,133.85
302,870.73
681,459.14
10,901,346.28
27.495
109.981
247.457
439.927
687.386
989.836
1,347.277
1,759.709
Ψ
Photo-electron Wavefunction
KEM
12
48
108
192
300
432
588
768
972
Free Electron
Δ difference per Electron level
36
60
84
108
132
156
180
Δhv

photo-electrons

photons

Spectral lines are produced by

The energy and wavelength of emitted photons is a function of the photo-electron's KEM field quantum wavenumber

Quantum Transitions

$$hf = \Delta Mv = \Delta p$$

27.49545

accelerating photo-electrons

Changes to the energy momenta of KEM fields produce changes to the linear momenta of photo-electrons

Tetryonics 37.02 - Photoelectron transitions

Copyright ABRAHAM [2008] - all rights reserved

Photoelectric energy level transitions

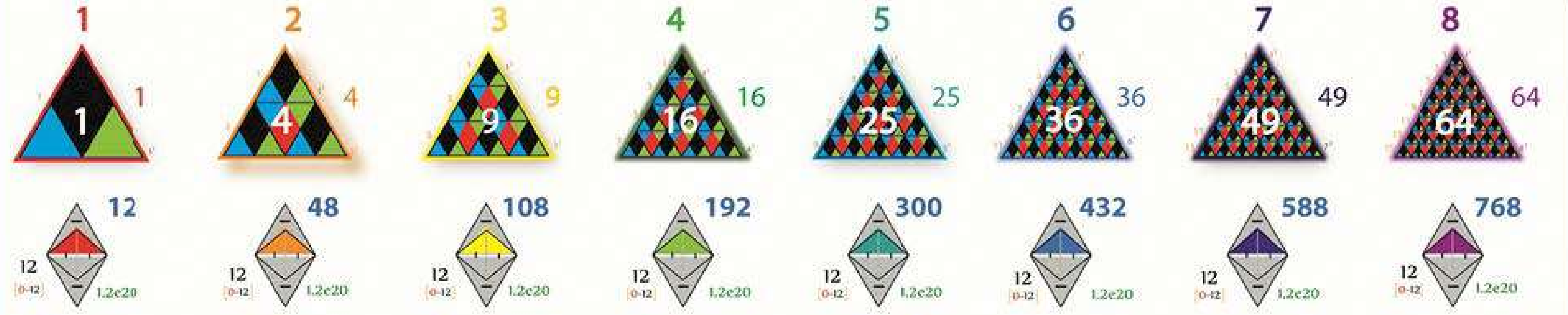
1
1
1
2
4
4
3
9
9
4
16
16
5
25
25
6
36
36
7
49
49
8
64
64
12
48
108
192
300
432
588
768
12
[0-12]
1.2e20

Lyman	Balmer	Paschen	Brackett	Pfund	Humphries	un-named
91.29	367.06	833.16	1499.69	2381.53	3499.27	1917.4
91.3	367.23	834.34	1503.5	2391.17	3520.12	
91.32	367.28	835.68	1507.87	2402.25		
91.33	367.54	837.23	1512.92			
91.36	368.19	839.02				
[illegible]	[illegible]					
91.41						
UV spectrum	Visible Spectrum	IR spectrum				

Applying any extra Kinetic Energy to leptons in this energy level results in unbound photo-leptons

Q
RE
v
p
KE
hv
v²
M
3
3

$$Mv^2 = KEM = hcR_H$$

These specific quantum level steps create the Photoelectric effect

$$\Delta hv = \phi = hf$$

Photo-electrons bound to atomic nuclei can only transition between Mv squared KEM energies as a direct result of the energy levels of Baryons

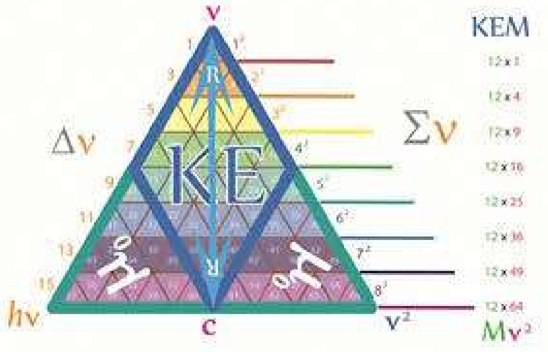
v
KEM
Δv
Σv
KE
R
R
hv
c
v²
1
3
5
7
9
11
13
15
1²
2²
3²
4²
5²
6²
7²
8²
12 x 1
12 x 4
12 x 9
12 x 16
12 x 25
12 x 36
12 x 49
12 x 64
Mv²

Tetryonics 37.03 - Photoelectric spectral energy levels

The Photoelectric effect

The photoelectric effect was first observed in 1887 by Heinrich Hertz (1857-1894) during experiments with a spark-gap generator — the earliest form of radio receiver

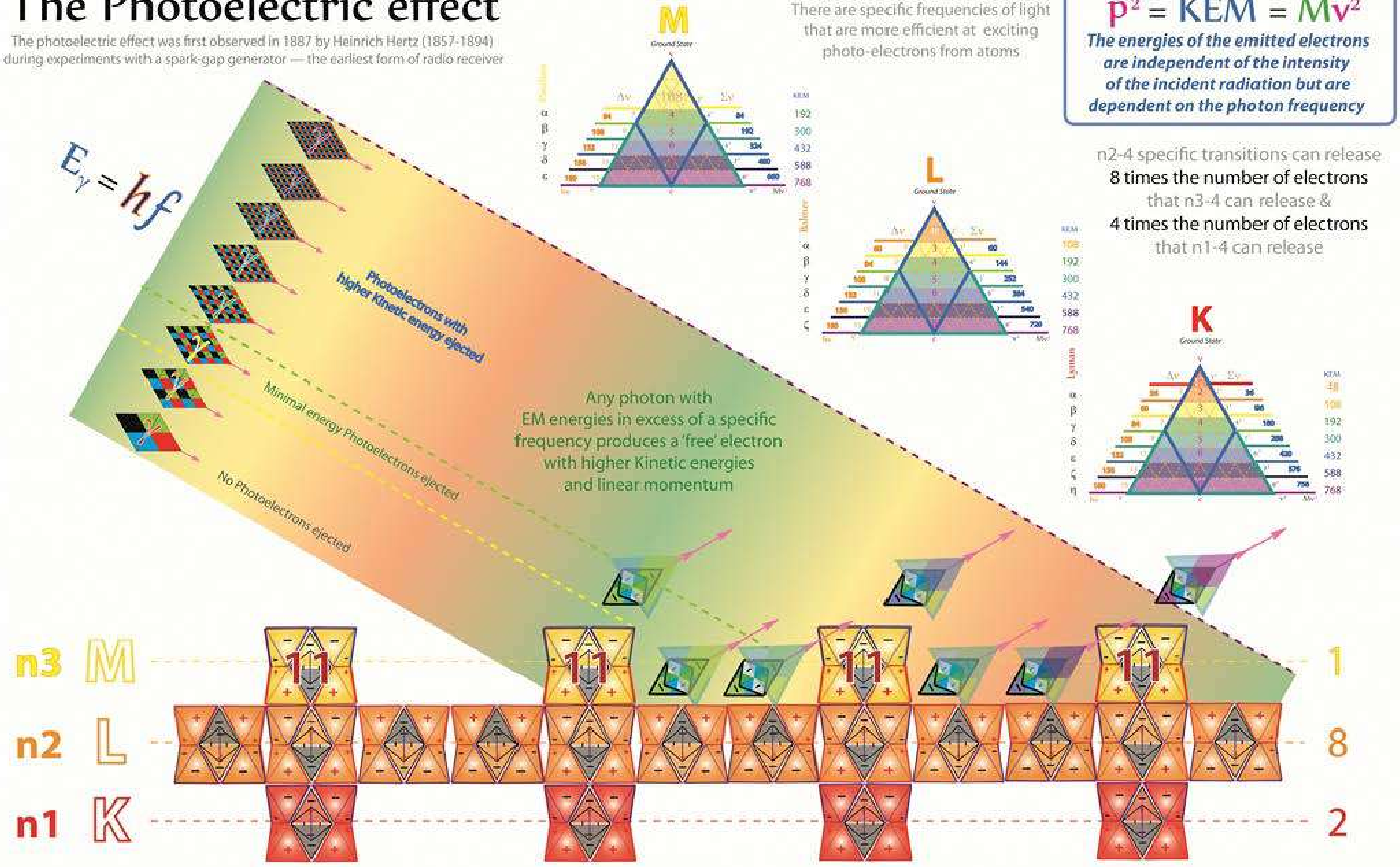
$E_\gamma = hf$
M
Ground State
L
Ground State
K
Ground State
Δv
Σv
KEM
192
300
432
588
768
108
48
Balmer
Lyman
Photoelectrons with higher Kinetic energy ejected
Minimal energy Photoelectrons ejected
No Photoelectrons ejected
n3 M
n2 L
n1 K
11
1
8
2

There are specific frequencies of light that are more efficient at exciting photo-electrons from atoms

$$p^2 = KEM = Mv^2$$

The energies of the emitted electrons are independent of the intensity of the incident radiation but are dependent on the photon frequency

n2-4 specific transitions can release 8 times the number of electrons that n3-4 can release & 4 times the number of electrons that n1-4 can release

Any photon with EM energies in excess of a specific frequency produces a 'free' electron with higher Kinetic energies and linear momentum

Soon after the discovery of Planck's Heat law and the quantisation of energy at the quantum level experiments by others, most notably Robert Millikan (1865-1953), found that light with frequencies below a certain cutoff value, called the threshold frequency, ould not eject photoelectrons from the metal surface no matter how bright the source

Photonic mass-energy momenta

$$2m\Omega v^2 = E = n.hf$$

All 2D EM waveforms possess mass-energy and momenta

Bosons

EM Field   Planck quanta

$$\text{ODD}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons   ElectroMagnetic   mass   velocity

Photons

EM Field   Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons   ElectroMagnetic   mass   velocity

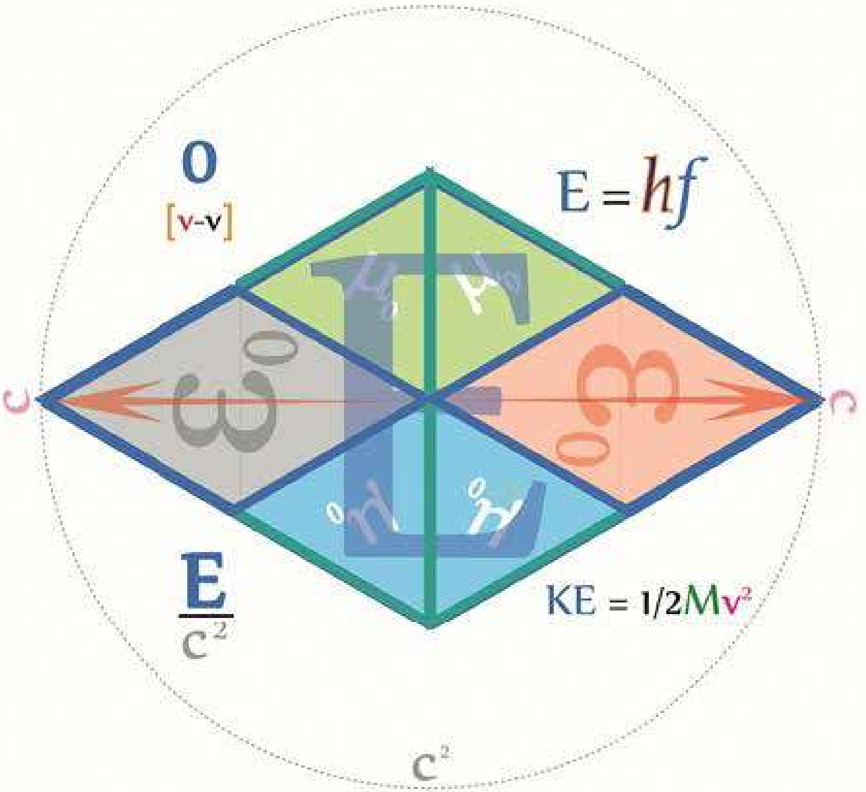
0
[v-v]
$E = hf$
c
c
μ_0
μ_0
ε_0
ε_0
μ_0
μ_0
$\frac{E}{c^2}$
$KE = 1/2Mv^2$
c^2

charged mass-energy-momentum

$$E = mv^2$$

radiant mass-energy momenta

$$E = hv^2$$

v
hv
v^2
ε_0
m
Ω
μ_0
μ_0

Planck energies

All 2D EM waveforms propagate at the speed of light

Velocity of Light   Speed of Light

$$v = \left[\left[\frac{c^2}{\Omega}\right]\cdot\left[\frac{\Omega}{c}\right]\right] = c$$

f scalar frequency λ linear wavelength

ε_0
γ
μ_0
μ_0
μ_0
μ_0
ε_0

Wave~Particle duality

First proposed by Christian Huygens in the 1600's, it wasn't until the 1800's that Thomas Young proved this wave-particle duality with the classic double-slit experiment.

Sir Isaac Newton

(1643-1727)

Particles

In 1678, Dutch physicist, Christiaan Huygens, believed that light was made up of waves vibrating up and down perpendicular to the direction of the light travels, and therefore formulated a way of visualising wave propagation.

This became known as 'Huygens' Principle' and was the successful theory of light wave motion in three dimensions

Sir Isaac Newton, held the theory based on his spetral observations that light was made up of tiny particles or 'corpusles of colour'

English physicist Thomas Young argued that Isaac Newton's theory of particle light was incorrect, and instead argued that light is a transverse wave.

In 1803, Thomas Young studied the interference of light waves by shining light through a screen with two slits equally separated, the light emerging from the two slits, spread out and produced wave-like interference patterns.

Waves

Thomas Young

(1773-1829)

In 1900
Max Planck proposed the existence of a light quantum, [n.hv] a finite packet of energy which depends on the frequency and spectral energy of the radiation

Light - a Wave, or a Particle?

In 1905
Albert Einstein, suggested that light is composed of tiny particles called photons, and that each photon has energy related to its frequency [hf].

Max Karl Ernst Ludwig Planck

(April 23, 1858 – October 4, 1947)

Bosons

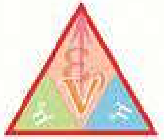

$$nhv = E = hf$$

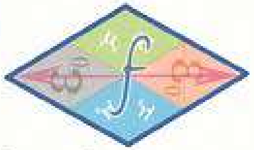

Mathematically directly relating the number of Planck quanta [n.hv] with Photon frequencies [hf] and the frequent interchanging of one term for the other in physics is the source of considerable quantum confusion

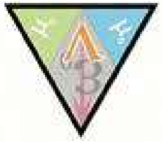

$$2hv = hf$$

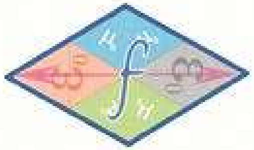

Photons

Albert Einstein

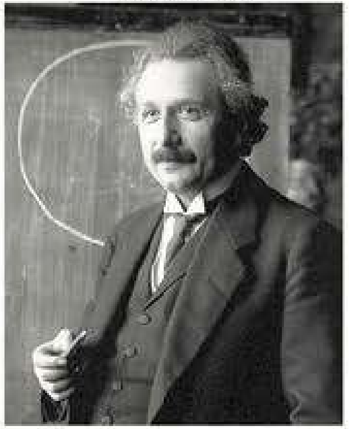

(14 March 1879 – 18 April 1955)

The quantum idea was soon seized upon to explain the photoelectric effect, and became part of the Bohr theory of discrete atomic spectra, quickly becoming part of the foundation of modern quantum theory in turn this led to the quantum wierdness of wave-particle duaity, Heisenberg's Uncertainty principle and Schrodinger's quantum wave equation & wavefunctions.

Wave Particle Mechanics

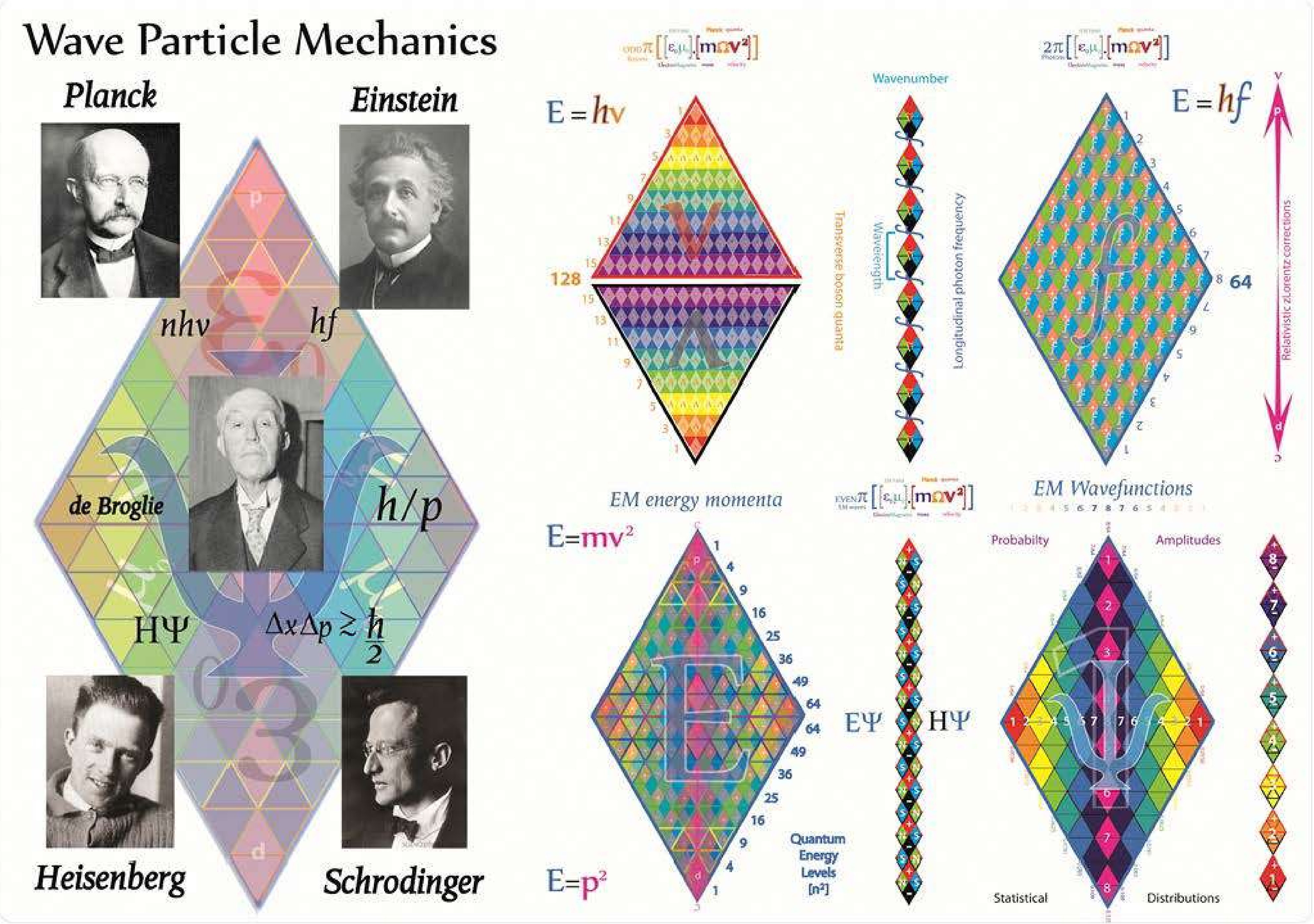
Planck
Einstein
nhv
hf
de Broglie
h/p
HΨ
ΔxΔp ≥ ħ/2
Heisenberg
Schrodinger
ODD π[[ε₀μ₀].[mΩv²]]
E = hv
128
Transverse boson quanta
Wavenumber
Wavelength
Longitudinal photon frequency
2π[[ε₀μ₀].[mΩv²]]
E = hf
64
Relativistic zLorentz corrections
EM energy momenta
EVEN π[[ε₀μ₀].[mΩv²]]
EM Wavefunctions
E=mv²
E=p²
Quantum Energy Levels [n²]
EΨ
HΨ
Probabilty
Amplitudes
Statistical
Distributions

Tetryonics 38.01 - Wave~Particle Mechanics

Copyright ABRAHAM [2008] - all rights reserved

Blackbody Radiation

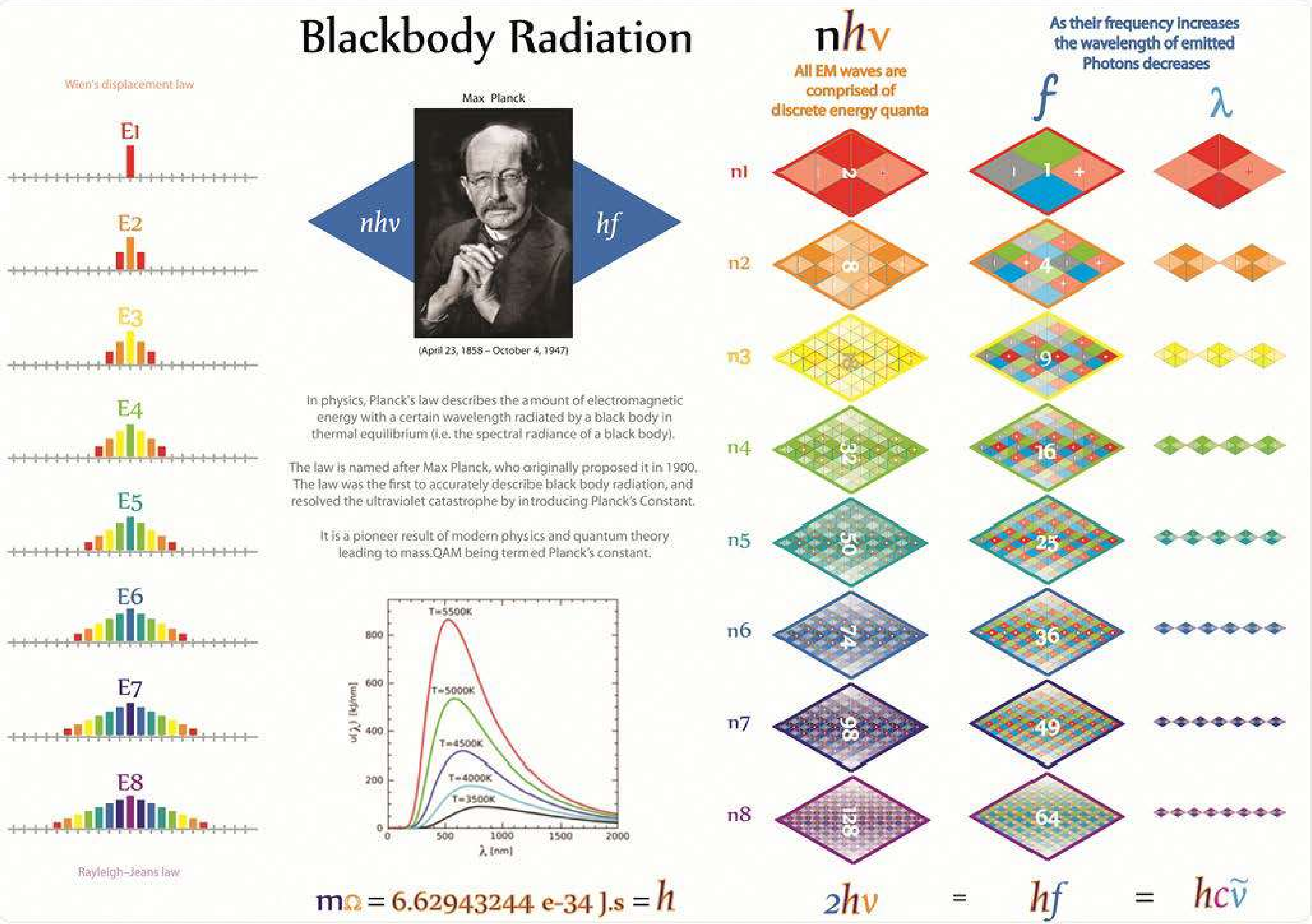
Wien's displacement law
E1
E2
E3
E4
E5
E6
E7
E8
Rayleigh–Jeans law
Max Planck
nhv
hf
(April 23, 1858 – October 4, 1947)
nhv
All EM waves are comprised of discrete energy quanta
As their frequency increases the wavelength of emitted Photons decreases
f
λ
n1
n2
n3
n4
n5
n6
n7
n8
2
8
18
32
50
74
98
128
1
4
9
16
25
36
49
64
T=5500K
T=5000K
T=4500K
T=4000K
T=3500K
800
600
400
200
0
u(λ) [kJ/nm]
0
500
1000
1500
2000
λ [nm]
2hv
=
hf
=
hcṽ

In physics, Planck's law describes the amount of electromagnetic energy with a certain wavelength radiated by a black body in thermal equilibrium (i.e. the spectral radiance of a black body).

The law is named after Max Planck, who originally proposed it in 1900. The law was the first to accurately describe black body radiation, and resolved the ultraviolet catastrophe by introducing Planck's Constant.

It is a pioneer result of modern physics and quantum theory leading to mass.QAM being termed Planck's constant.

$m\Omega = 6.62943244\text{ e-}34\ \text{J.s} = h$

Tetryonics 38.02 - Blackbody Radiation curves

EM field energies

Planck quanta

$$E = n\pi\left[\left[m\Omega v^2\right]\right]$$

mass velocity

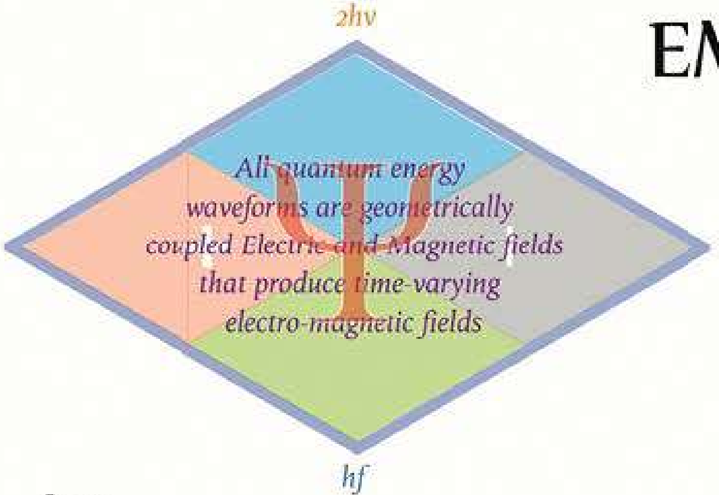

All quantum energy waveforms are geometrically coupled Electric and Magnetic fields that produce time-varying electro-magnetic fields

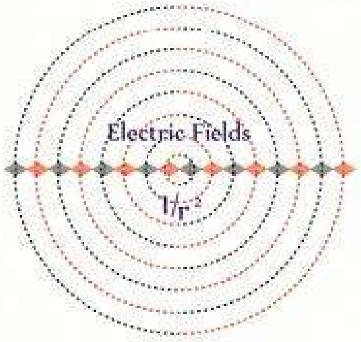
Electric Fields
1/r²

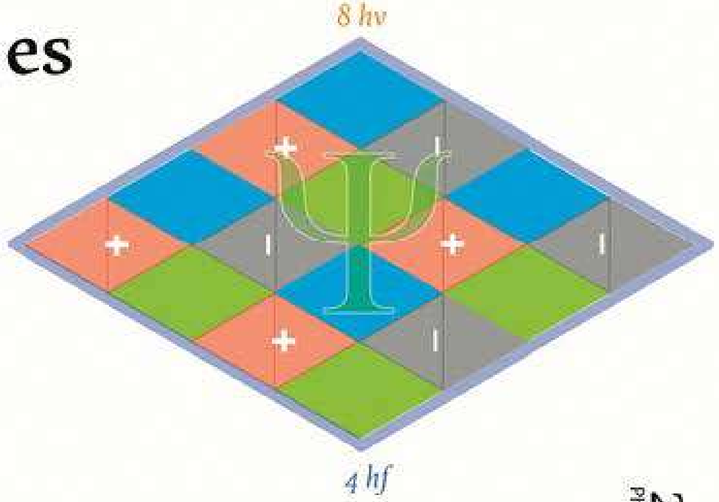

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Charge

EM Field ElectroMagnetic mass Planck quanta velocity

$$E = h\nu$$

EM waves are subject to Lorentz co-ordinate transformations

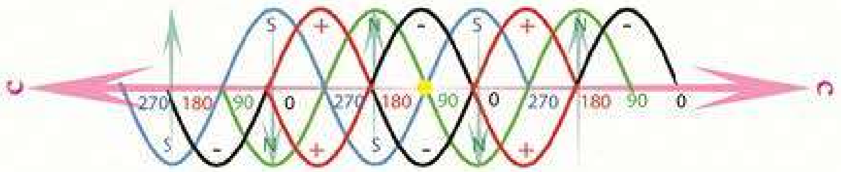
c
S
+
N
-
S
+
N
-
270
180
90
0
270
180
90
0
270
180
90
0
S
-
N
+
S
-
N
+
c

EM waves are a quadrature waveform of bi-directional electromagnetic oscillations

$$E = hf$$

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons

EM Field ElectroMagnetic mass Planck quanta velocity

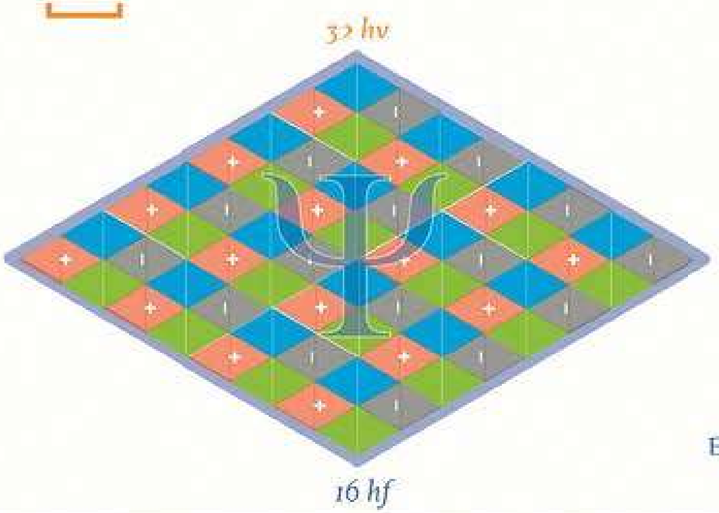

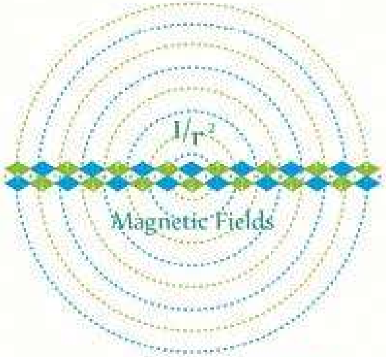
1/r²
Magnetic Fields

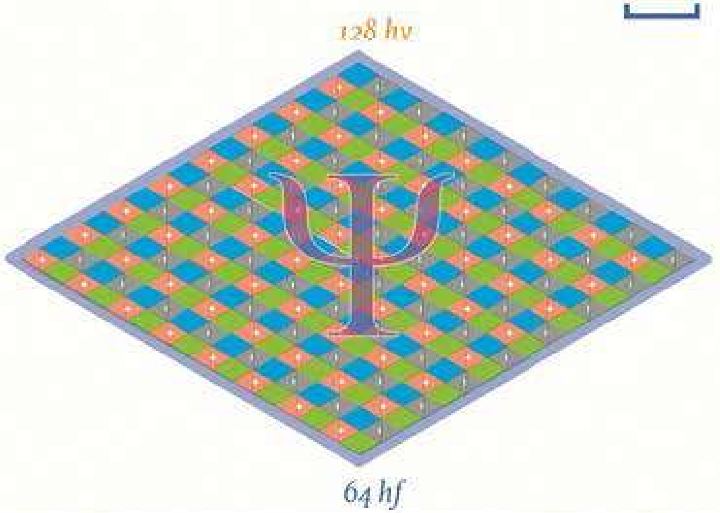

EM Field Planck quanta

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM waves ElectroMagnetic mass velocity

Tetryonics 38.03 - EM field Energy

Bosons and Photons in EM waves

$$\text{ODD}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons

EM Field Planck quanta

ElectroMagnetic mass velocity

Transverse charge Quanta

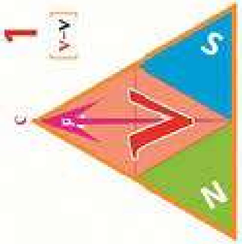

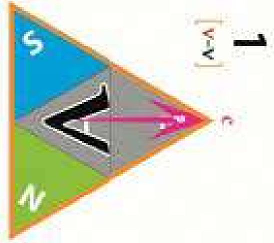

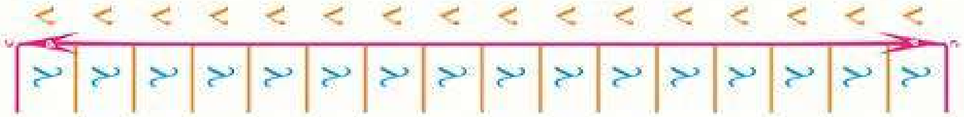

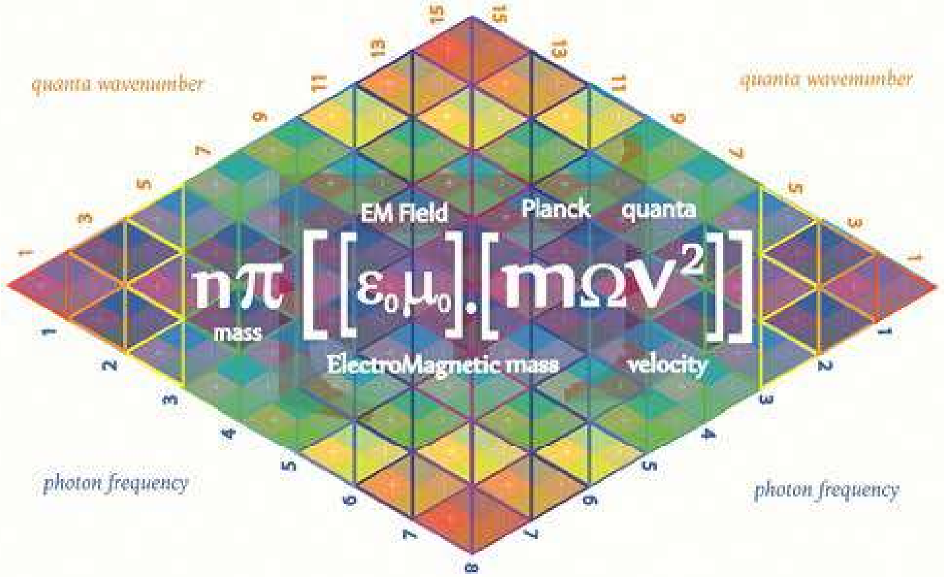
quanta wavenumber
quanta wavenumber
1
3
5
7
9
11
13
15
1
2
3
4
5
6
7
8
photon frequency
photon frequency
EM Field
Planck
quanta
mass
ElectroMagnetic mass
velocity

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

All EM waves are comprised of transverse Bosons

$$E = h\nu$$

$$E = hf$$

All EM waves are comprised of longitudinal Photons

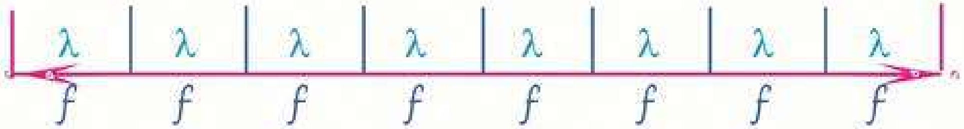

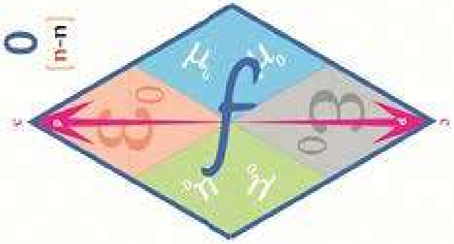
0
[n-n]
f
μ_0
ε_0

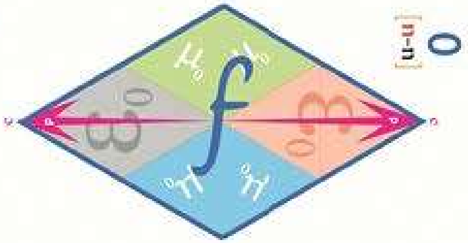
0
[n-n]
f
μ_0
ε_0

Longitudinal neutral Photons

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons

EM Field Planck quanta

ElectroMagnetic mass velocity

Copyright ABRAHAM [2008] - all rights reserved

EM wave energy quanta

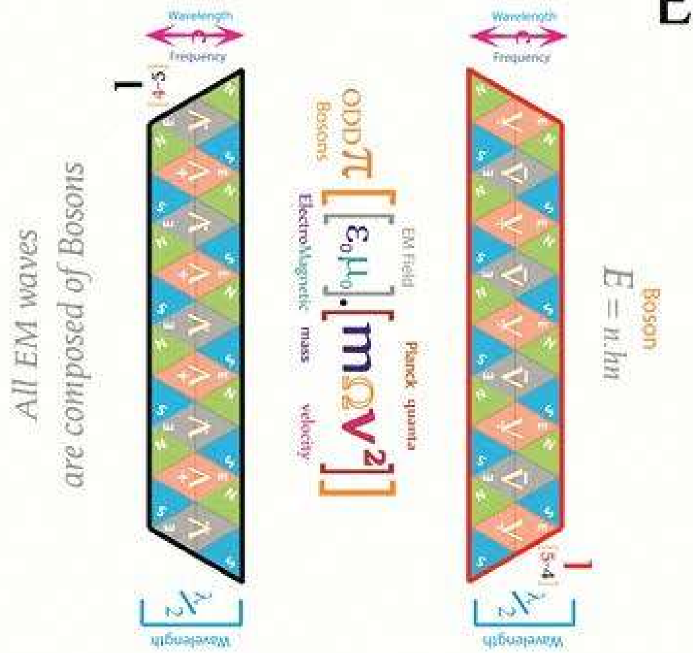
Wavelength
Frequency
Wavelength
Frequency
All EM waves are composed of Bosons
ODD π [[ε₀μ₀].[mΩv²]]
Bosons
EM Field
ElectroMagnetic
mass
Planck quanta
velocity
E = n.hn
Boson
λ/2
Wavelength
λ/2
Wavelength

Tranverse Bosons in EM wave energies should be modelled using [n] Planck quanta

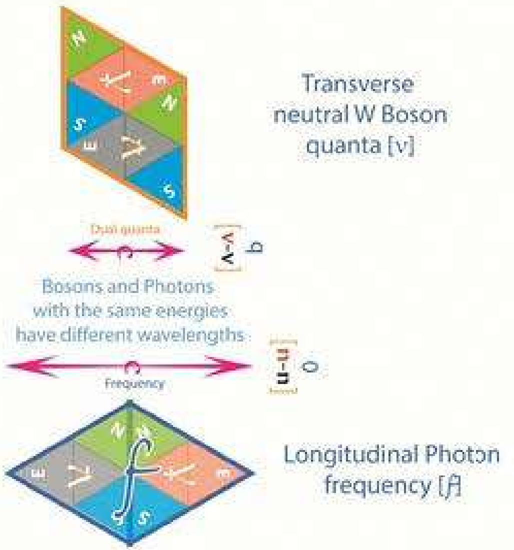
Transverse neutral W Boson quanta [ν]
Dual quanta
Bosons and Photons with the same energies have different wavelengths
Frequency
Longitudinal Photon frequency [f]

Wavefunction

$$i\hbar\frac{\partial}{\partial t}\Psi \quad \Psi \quad \hat{H}\Psi$$

Probability = |Amplitude|²

The Longitudinal EM wave energies should be modelled using [f] Photon frequencies

$$E = n\pi[m\Omega v^2]$$

Planck quanta

mass velocity

All Wave~Particle geometries and mass-energies can be calculated from and modelled with the Tetryonic Unified Field Equation

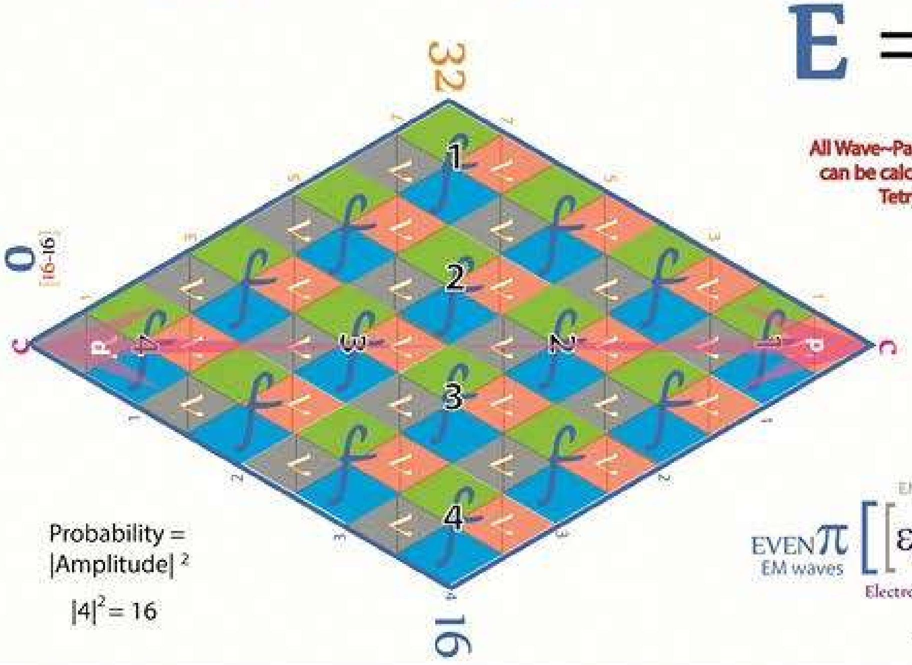
32
0
16
1
2
3
4
c
c

Probability = |Amplitude|²

$|4|^2 = 16$

$E = hf$

Photon

EVEN $\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM waves

EM Field Planck quanta

ElectroMagnetic mass velocity

All EM waves are composed of Square number Photons

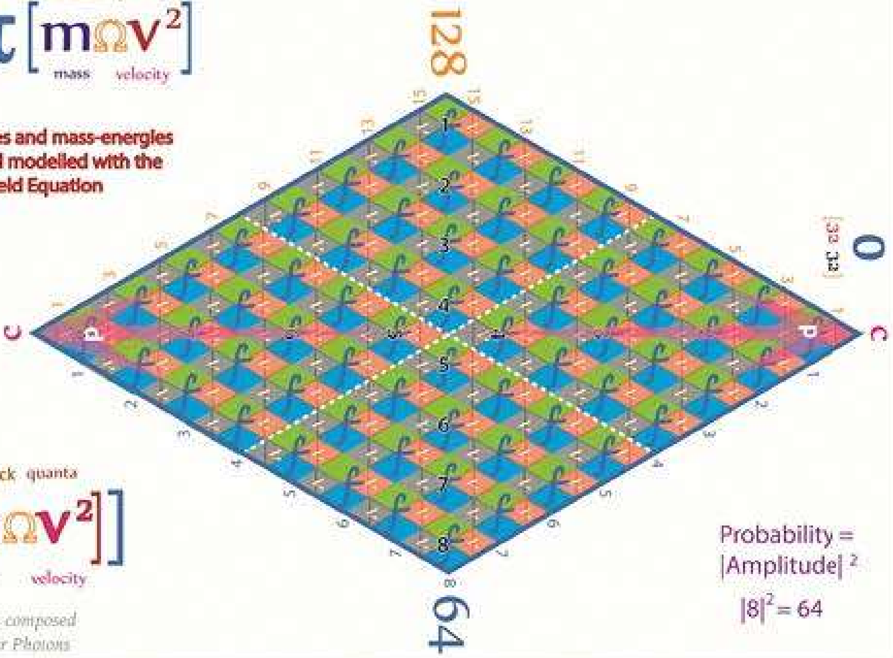
128
0
64
1
2
3
4
5
6
7
8
c
c

Probability = |Amplitude|²

$|8|^2 = 64$

Tetryonics 38.05 - EM wave Energies

Copyright ABRAHAM [2008] - all rights reserved

EM wave properties

Linear momentum is the square root of scalar mass-energies

$$\left[\sqrt{\overset{\text{Energy}}{[\underset{\text{mass}}{m}\underset{\text{quanta}}{\Omega}v^2]}}\right]$$

Transverse waveforms

EM waveform geometries

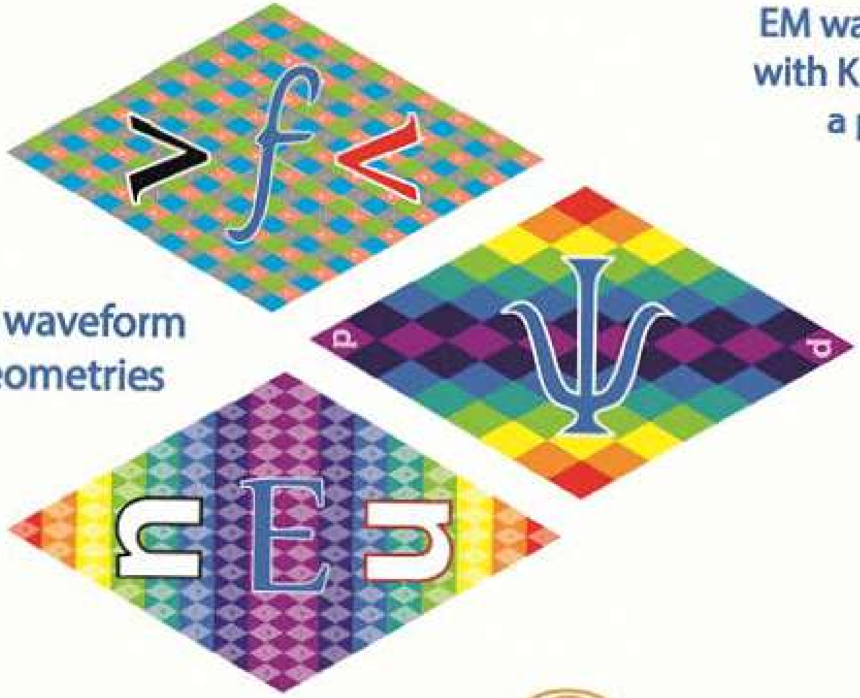
λ
f
λ
p
Ψ
p
ε
E
μ

EM waves are comprised of Photons with Kinetic energies, momenta and a photo-magnetic moment

$$p_\gamma = 2\pi\left[\overset{\text{Energy}}{[m\underset{\text{momenta}}{\Omega}v]}\right]$$

Photonic energy-momentum

Longitudinal waveforms

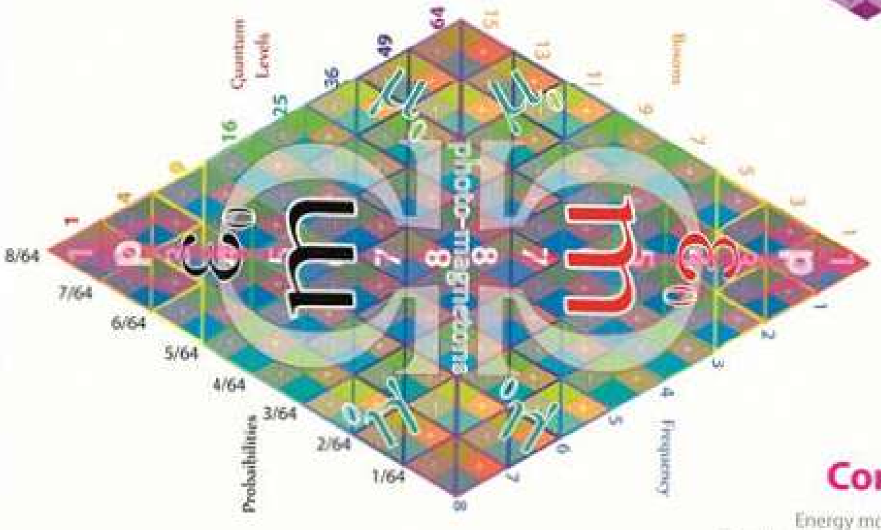
Quantum Levels
1
4
9
16
25
36
49
64
15
13
11
9
7
5
3
1
Bosons
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
Probabilities
1
2
3
4
5
6
7
8
Frequency
p
ε_0
m
photo-magnetons
m
ε_0
p
λ
λ
λ
λ

Scalar Planck quanta per second is the scalar source of all linear energy-momenta

p^2

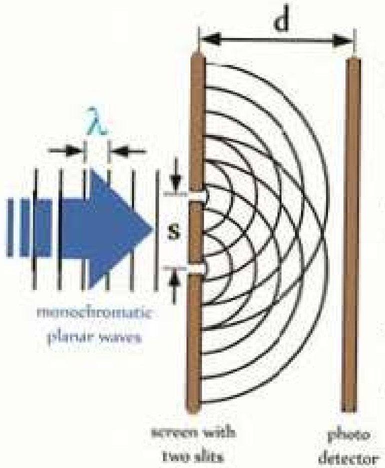
d
λ
s
monochromatic planar waves
screen with two slits
photo detector

Interference is the result of KEM waveform superpositioning

Conservation of Energy momenta

Energy momenta a conserved quantity in physical systems (laws of conservation), meaning that if a closed system is not affected by external foices, its total momentum cannot change

Copyright ABRAHAM [2008] - all rights reserved

Statistical probabilities of energy distributions

form the basis for quantum mechanical probabilities and wavefunctions

$$E = 2mv$$

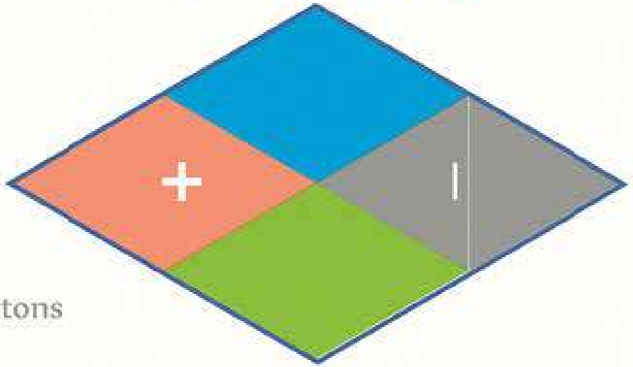

$$E = Mc^4$$

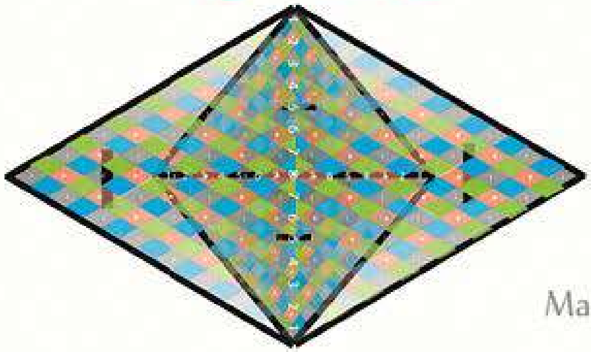

All ENERGY in motion has probabilistic distribtutions of energy quanta resulting from the equilateral Tetryonic geometries of electromagnetic mass-ENERGY-Matter

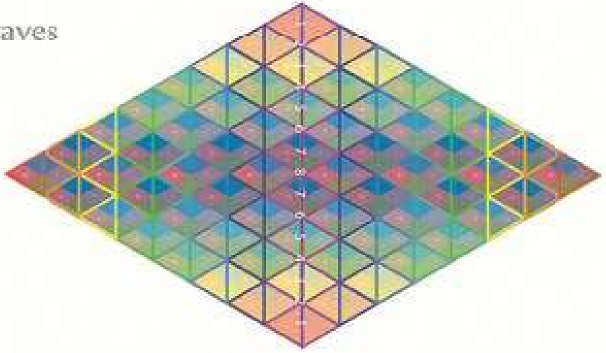

$$p^2 = E = mv^2$$

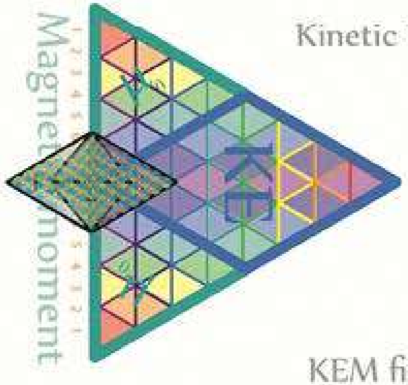
Kinetic Energies
Magnetic moment
KE
KEM fields

$$KE = \frac{1}{2}Mv^2$$

$$Mv^2 = KEM$$

Copyright ABRAHAM [2008] - all rights reserved

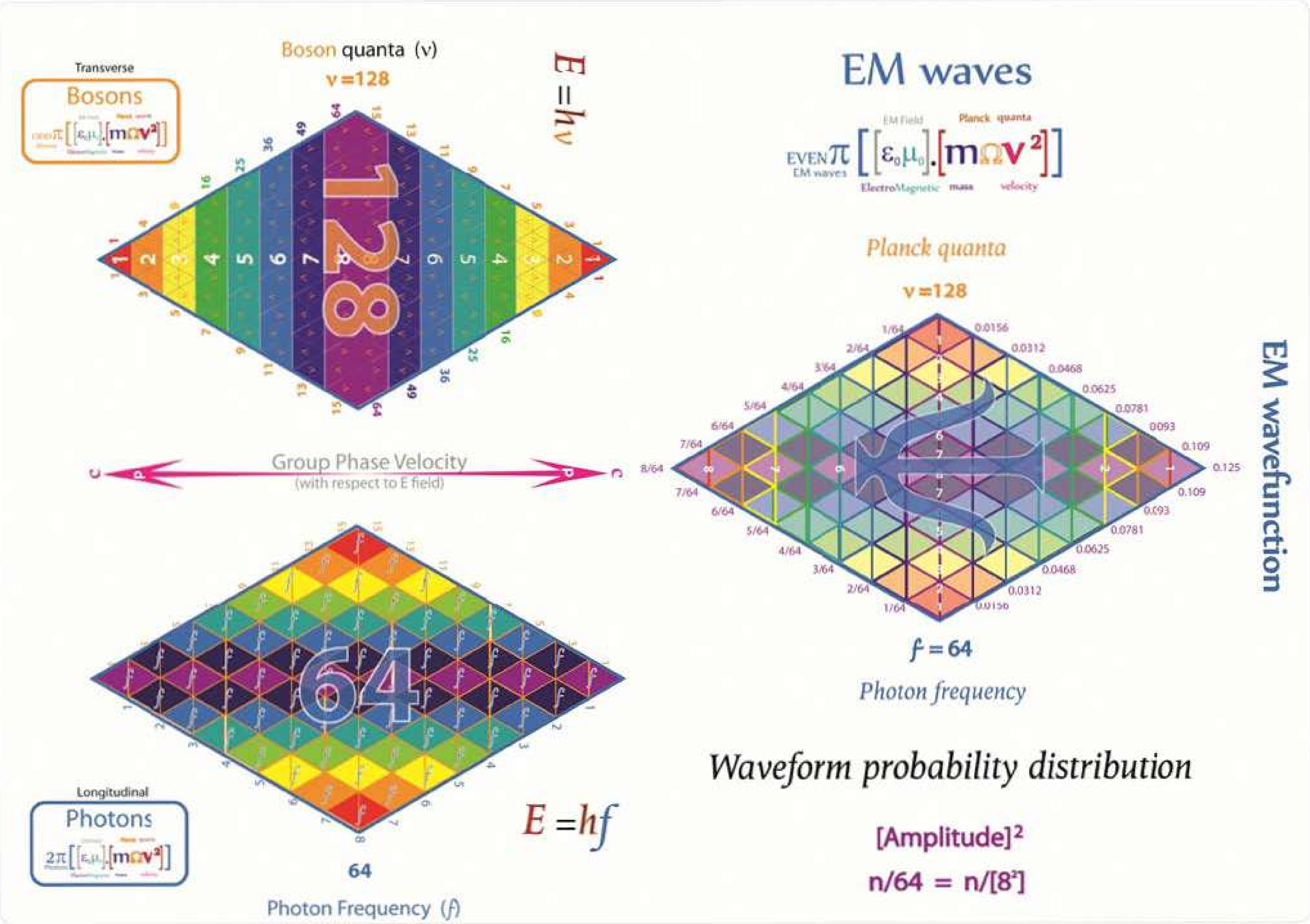
Transverse
Bosons
Boson quanta (ν)
ν=128
128
E = hν
EM waves
EVEN π [[ε₀μ₀].[mΩv²]]
EM waves
EM Field
Planck quanta
ElectroMagnetic
mass
velocity
Planck quanta
ν=128
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
EM wavefunction
f = 64
Photon frequency
Group Phase Velocity
(with respect to E field)
C
P
64
Waveform probability distribution
[Amplitude]²
n/64 = n/[8²]
Longitudinal
Photons
2π[[ε₀μ₀].[mΩv²]]
E = hf
64
Photon Frequency (f)

Tetryonics 38.08 - EM wavefunctions

Copyright ABRAHAM [2008] - all rights reserved

Wave functions

Bosons

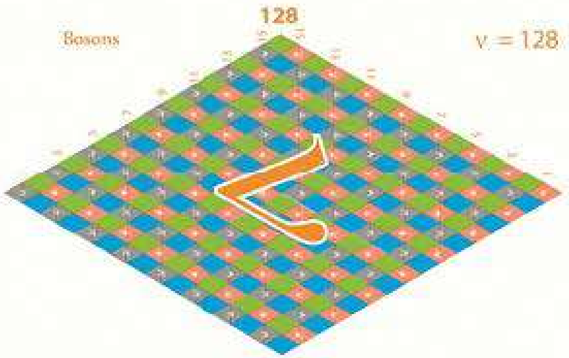
128
Bosons
ν = 128

$E = h\nu$

[Planck quanta are single-quanta Bosons]

All EM waves are comprised
of Longitudinal Photons
which in turn are made up
of Transverse Bosons

[Photons are neutral dual-quanta Bosons]

$E = hf$

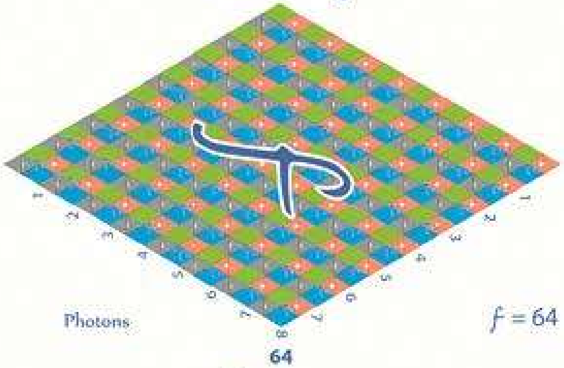
1
2
3
4
5
6
7
8
Photons
64
f = 64

Photons

All EM waves and Photons exhibit
quantum levels of mass-energy momenta
determined by their constituent bosons

Transverse
EM Waveforms

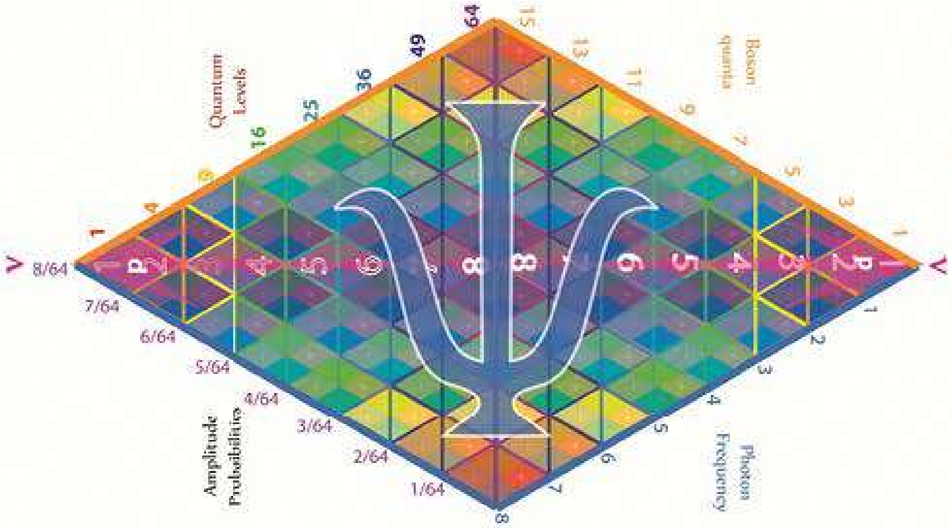
64
49
36
25
16
9
4
1
Quantum Levels
15
13
11
9
7
5
3
1
Boson quanta
V
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
Amplitude Probabilities
1
2
3
4
5
6
7
8
Photon Frequency
V
P
2
3
4
5
6
7
8
8
7
6
5
4
3
2
P

Longitudinal
EM waveforms

All electromagnetic waves exhibit
wave probabilities and amplitude functions
determined by their constituent photons

Energy levels

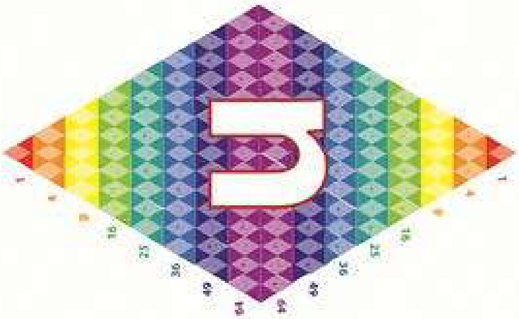
1
4
9
16
25
36
49
64

Squared numbers

Wave-particle duality holds that light and matter
simultaneously exhibit properties of waves
as well as particle-like properties (photons).

This concept is a consequence of the mathematical
treament of mass-energy physics in quantum mechanics
and a comprehensive explanation of this duality has
been elusive to Physics since its discovery.

Any attmpt to develop an unified quantum theory
must explain the root cause and processes
behind Wave-Particle duality and must also explain:

Diffraction, Deflection and Reflection
along with Bosons, Photons, EM radiation,
energy quantisation and mass-Matter.

Probabilities

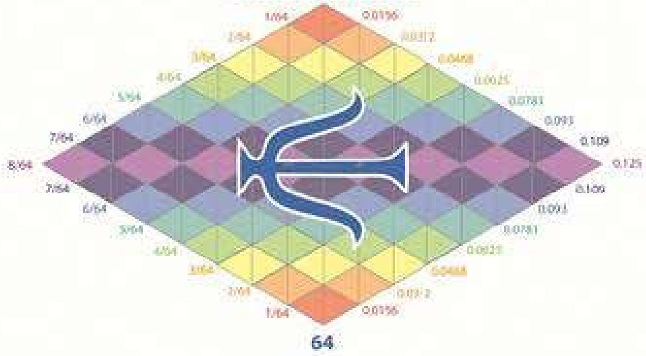
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
0.109
0.093
0.0781
0.0625
0.0468
0.0312
0.0156
64

Waves

Copyright ABRAHAM [2008] - all rights reserved

EM Wavefunctions

The wavefunction for a system contains all the information about the quantum state that a particle/system is in and gives a complete description of that part of the world at any one particular instant in time.

Transverse mass-energy distrbutions

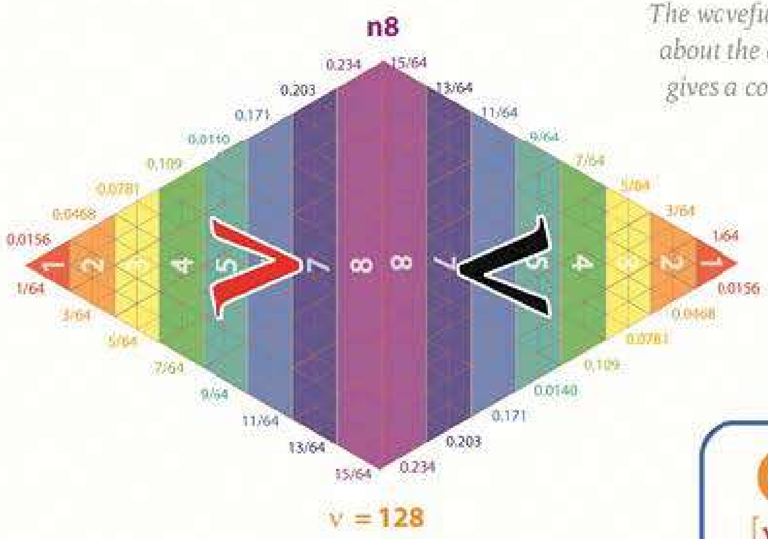
n8
0.234
15/64
0.203
13/64
0.171
11/64
0.0140
9/64
0.109
7/64
0.0781
5/64
0.0468
3/64
0.0156
1/64
1 2 3 4 5 7 8 8 7 5 4 3 2 1
v
v
v = 128

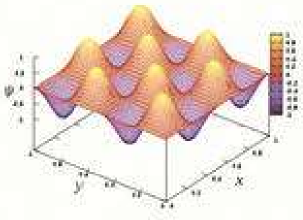

Longitudinal mass-energy distributions

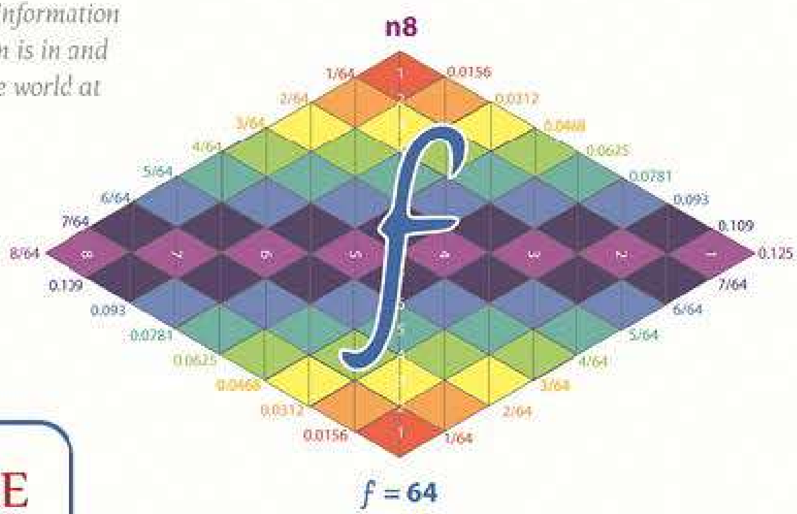
n8
1/64
0.0156
2/64
0.0312
3/64
0.0468
4/64
0.0625
5/64
0.0781
6/64
0.093
7/64
0.109
8/64
0.125
0.139
f
f = 64

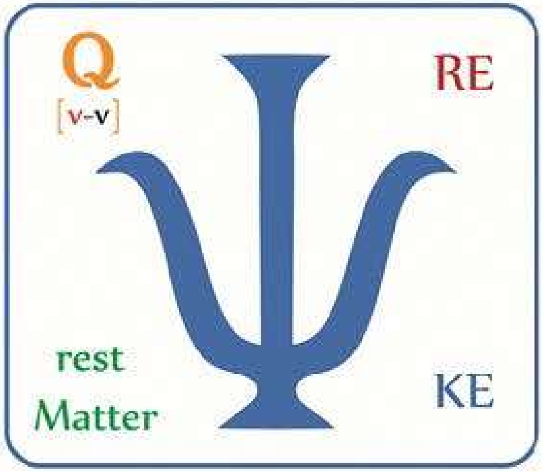
Q
[v-v]
RE
rest
Matter
KE

The wavefunction of an EM wave (KEM) geometry is distinct from that of rest mass-Matter toplogies due to the number of Planck quanta contained in it but both can be modelled using Tetryonic geometry

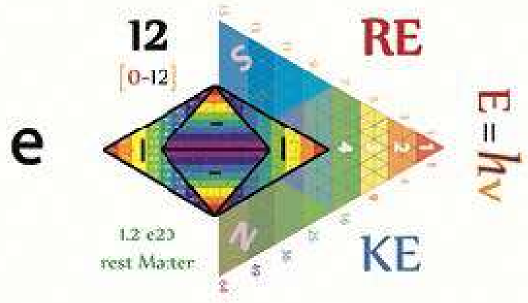
e
12
[0-12]
RE
E = hv
1.2 e20
rest Matter
KE

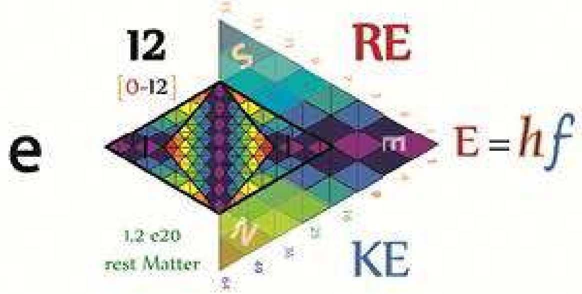
e
12
[0-12]
RE
E = hf
1.2 e20
rest Matter
KE

Quantum Energy Levels

The wavefunction itself is often said to be un-observable. In fact, it can modelled as it is a reflection of the quantas making up the quantum state of any particle and is complex-valued.

As the system evolves over time, the wavefunction also changes, so it can be written as a function of time Ψ(t)

$$\int_{-\infty}^{\infty} |\Psi|^2 \mathrm{d}x = 1.$$

The absolute square of the function must be normalizable

Wave-Particle Probabilities

The absolute square of the wavefunction is a probability density (the area of highest probability for a measurement to take place.

For example, if the wavefunction is expressed in real space and our system is a particle, the absolute square gives a probability density for the position of the system. Integrating this probability density between some bounds will give the probability that the particle will be found in that region when its position is measured

Copyright ABRAHAM [2008] - all rights reserved

Wave~Particle energies

Equilateral energy is the foundational geometry
of Kinetic EM Waveforms [WAVES]
and standing wave Matter [PARTICLES]

All Matter in motion exhibits a Wave-Particle duality

Everything is made up of
charged mass-energy quanta

ZPFs combine in ODD numbers to form
Bosons

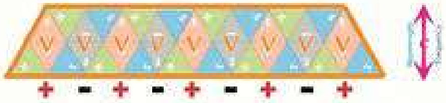

Bosons combine to form
Quantum Levels

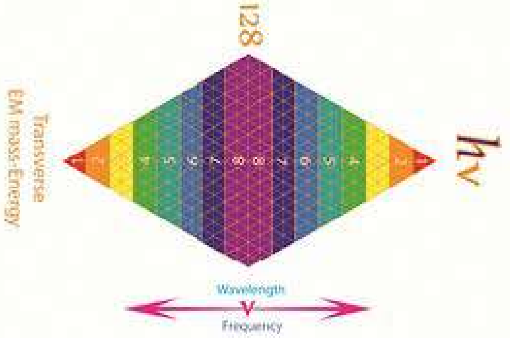
128
Transverse
EM mass-Energy
hv
Wavelength
Frequency

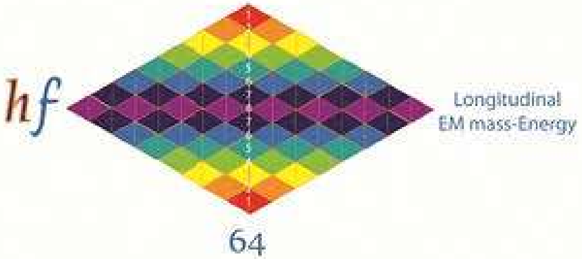
hf
Longitudinal
EM mass-Energy
64

The quantum Planck quanta tesslation
of Matter can be accurately modelled with
Wave Probability mechanics

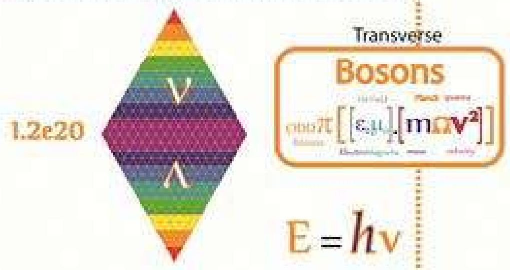
1.2e20
Transverse
Bosons
ODD π [[ε₀μ₀].[mΩv²]]
E = hν

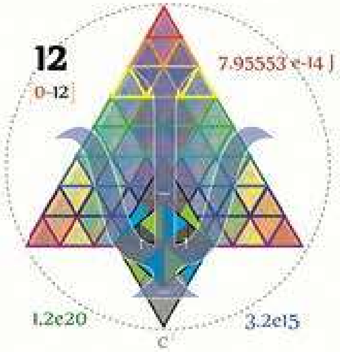
12
0-12
7.95553 e-14 J
1.2e20
3.2e15
c

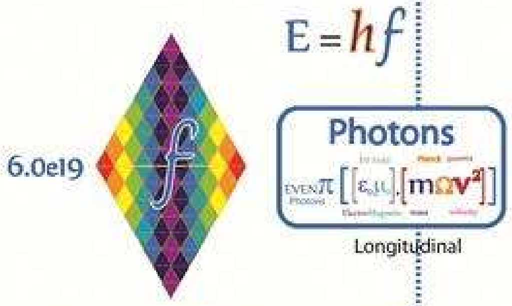
E = hf
6.0e19
Photons
EVEN π [[ε₀μ₀].[mΩv²]]
Longitudinal

A n8 Electron is a massive particle
which has a Kinetic EM energy field
[6.4e15 planck quanta]

2D EM field geometries
can combine to form
3D Tetryonic Matter topologies

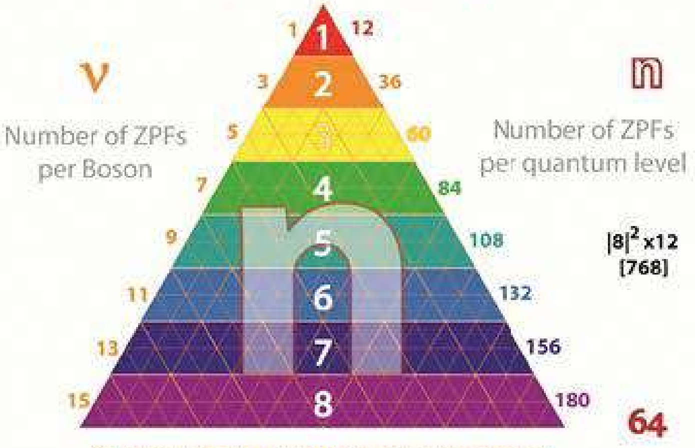
ν
Number of ZPFs
per Boson
n
Number of ZPFs
per quantum level
1 3 5 7 9 11 13 15
1 2 3 4 5 6 7 8
12 36 60 84 108 132 156 180
|8|² x12
[768]
64

Quantum levels are a transverse measurements of energy

Equivalently, all Matter & EM waves can be viewed either in terms of their transverse boson or longitudinal photon content and/ or their associated mass-energy momenta geometries

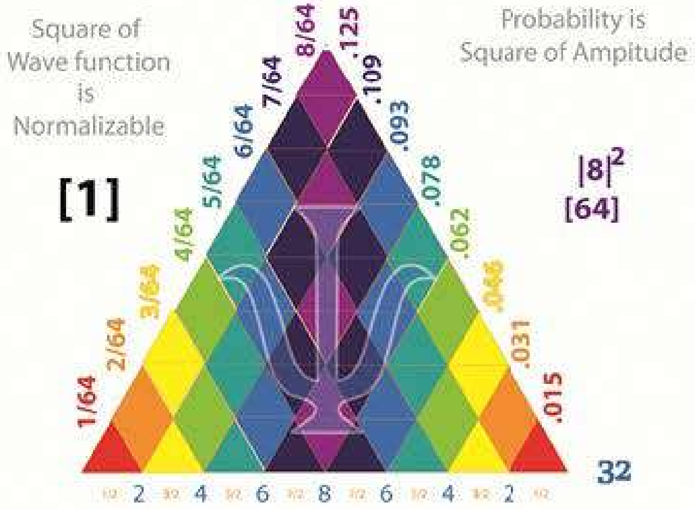
Square of
Wave function
is
Normalizable
[1]
Probability is
Square of Amplitude
|8|²
[64]
1/64 2/64 3/64 4/64 5/64 6/64 7/64 8/64
.125 .109 .093 .078 .062 .046 .031 .015
1/2 2 3/2 4 5/2 6 7/2 8 7/2 6 5/2 4 3/2 2 1/2
32

Wave functions are a linear measurement of photonic energy

Wave-particle Probabilities

All Particles in motion exhibit a Wave-Particle duality due to the 2D waveform geometry of its KEM wave and the lorentz velcoity invariant 3D standing wave geometry of its Matter

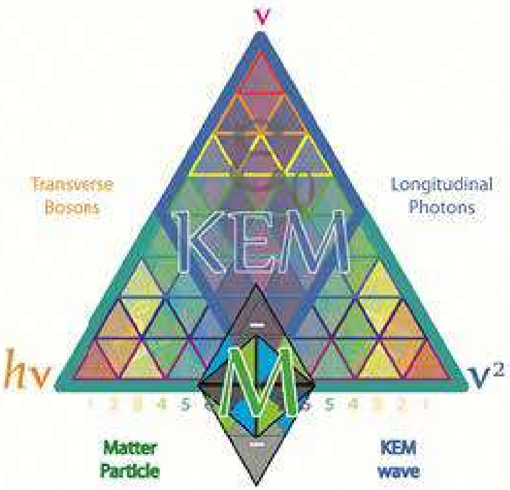
v
Transverse Bosons
Longitudinal Photons
KEM
M
hv
v²
1 2 3 4 5 6 5 4 3 2 1
Matter Particle
KEM wave

Equilateral energy tessellation provides the basis for the probabilistic statisical mechanics and math for all of Quantum Mechanics

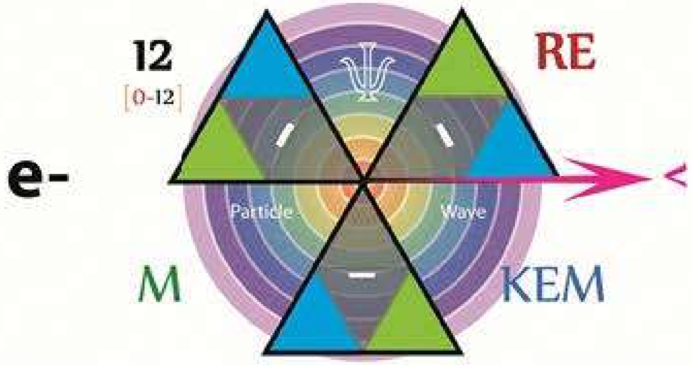
12
[0-12]
RE
e-
v
Particle
Wave
M
KEM

Schrodinger Wavefunction
The Wavefunction of a Particle equates to its Amplitude

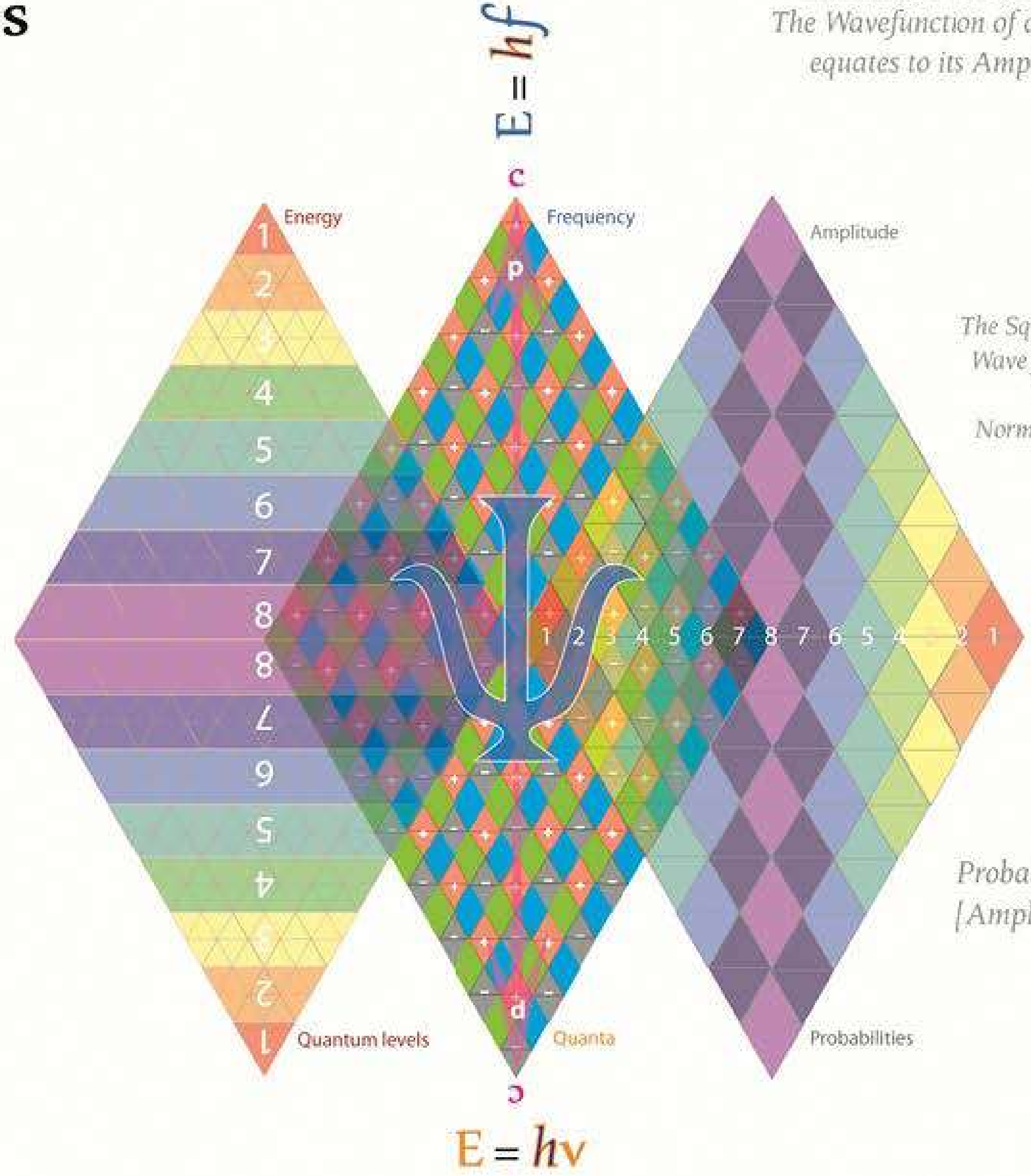
E = hf
c
Energy
Frequency
Amplitude
1
2
3
4
5
6
7
8
8
7
6
5
4
3
2
1
p
d
Ψ
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
Quantum levels
Quanta
Probabilities
c
E = hv

The Square of a Wave function is Normalizable

Probability = $[\text{Amplitude}]^2$

Born Probability Rule
The Probability of finding a Particle is the Square of its Amplitude

Tetryonics 38.12 - Wave~Particle Probabilities

Copyright ABRAHAM [2008] - all rights reserved

White Light mass-energy momenta waveforms

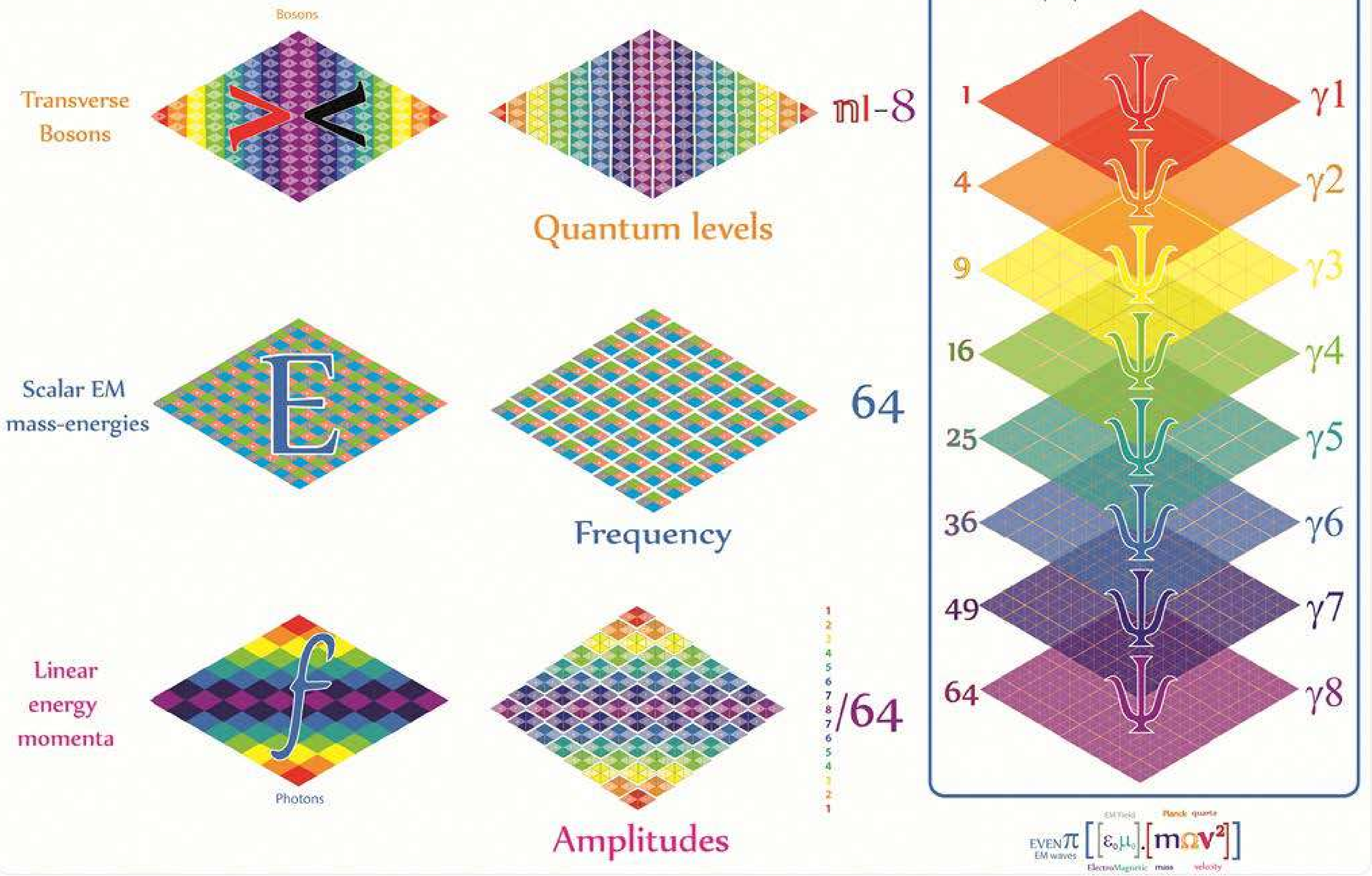
Bosons
Transverse Bosons
nl-8
Quantum levels
Scalar EM mass-energies
E
64
Frequency
Linear energy momenta
f
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
/64
Photons
Amplitudes
White light is comprised of superpositioned EM waves
1
4
9
16
25
36
49
64
γ1
γ2
γ3
γ4
γ5
γ6
γ7
γ8
EM Field
Planck quanta
EVEN π
EM waves
ElectroMagnetic
mass
velocity

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Tetryonics 38.13 - White Light waveform geometries

Bosons and Photons

EM waves are comprised of longitudinal [E=hf] Photons which in turn can also be measured as transverse [E=nhv] Bosons

Electro-Magnetic Fields

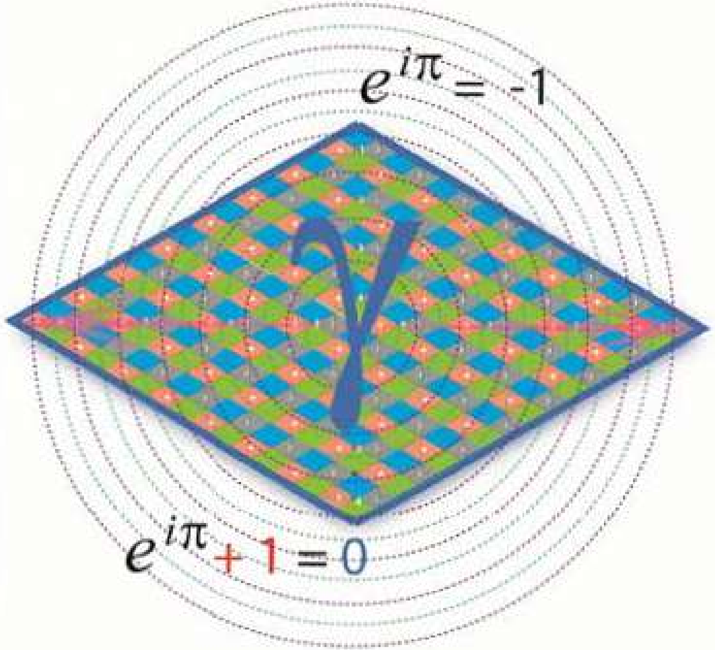

$e^{i\pi} = -1$

$e^{i\pi} + 1 = 0$

As described by Lorentz co-ordinate transforms the neutral quantum coin [quoin] geometry of all photons and EM waves have mirror E-field charges on each of their opposing fascia

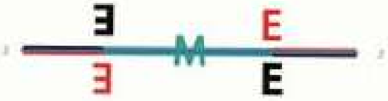

2D planar mass-energy momenta waveforms

$$\underset{\text{Photons}}{2\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck quanta}}{[\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}]}\right]$$

Photons and EM waves have dual c^2 space-time geometries

(8.987551787 e16 m²/s)

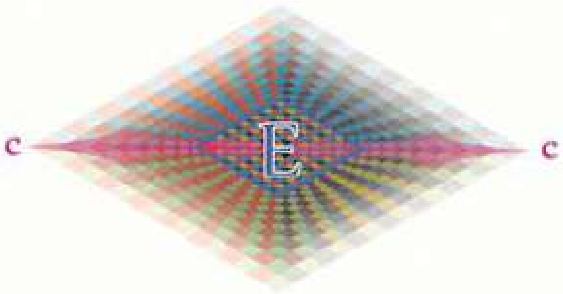

Photons and EM waves propagate outwards from their source at 'c'

(299,792,458 m/s)

$$\underset{\text{EM waves}}{\text{EVEN}\,\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck quanta}}{[\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}]}\right]$$

EM waves are detected as alternating radial electric E-wave radiation fields

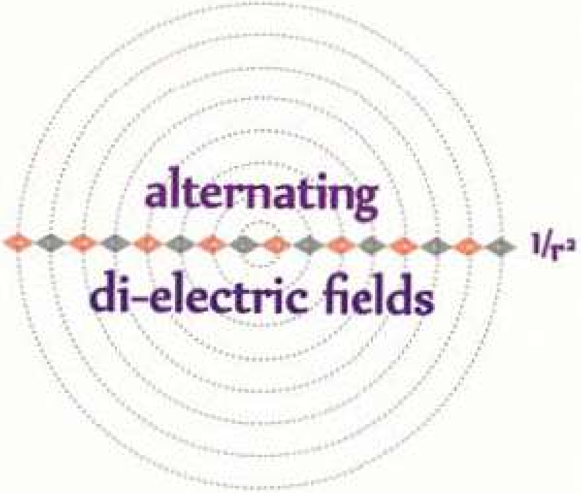

alternating

di-electric fields

$1/r^2$

EM wave radiation

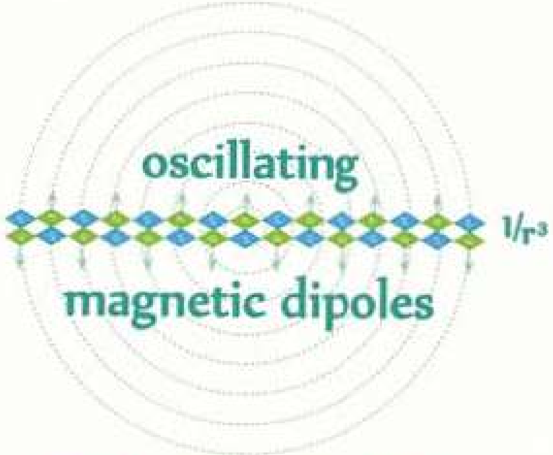

oscillating

magnetic dipoles

$1/r^3$

The Magnetic vector of EM energy momenta propagate orthagonally to their E-fields

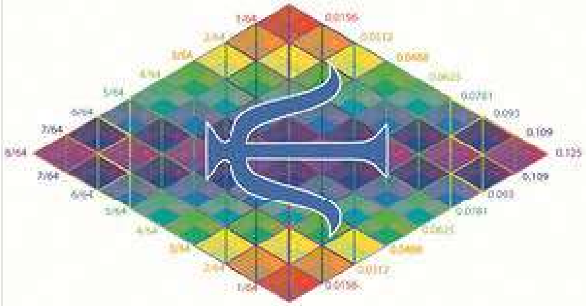

EM wave patterns

Polarised Photons and EM waves produce linear interference patterns

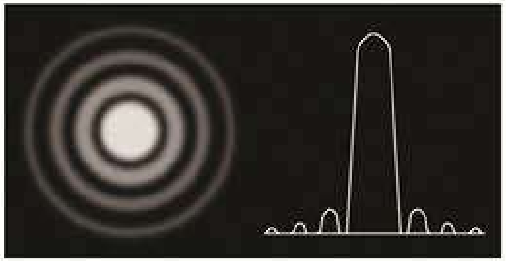

Diffraction occurs whenever propagating waves appear to bend around small obstacles in their path and spread out as smaller waves waves past any openings. its effects are generally most pronounced for waves whose wavelength is roughly similar to the dimensions of the diffracting objects.

If the obstructing object provides multiple, closely spaced openings, a complex pattern of varying intensity can result.

This is due to the superposition, or interference, of different parts of a wave that travels to the observer by different paths

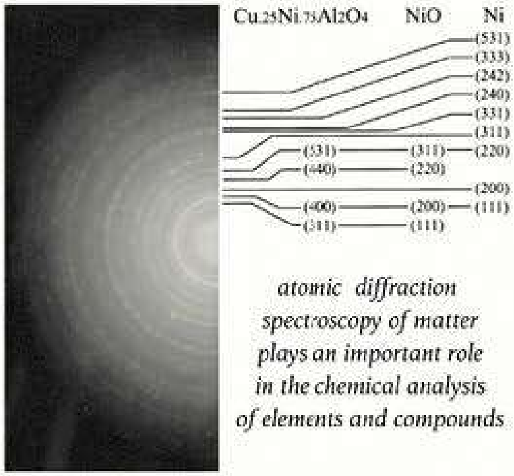

atomic diffraction spectroscopy of matter plays an important role in the chemical analysis of elements and compounds

Diffraction occurs with all waves, including sound waves, water waves, as well as with electromagnetic waves such as visible light, X-rays and radio waves.

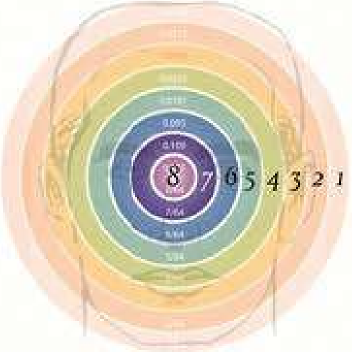

Un-Polarised Photons and EM waves produce circular diffraction patterns

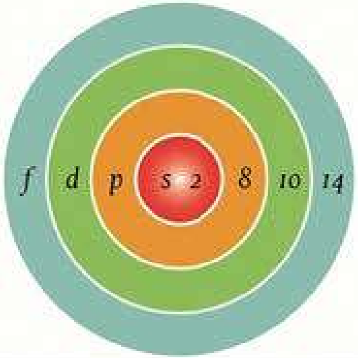

Richard Feynman once commented that:
"No-one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage, and there is no specific, important physical difference between them."

Copyright ABRAHAM [2008] - all rights reserved

Light waves

Thomas Young

(13 June 1773 – 10 May 1829)

In order to establish his wave theory of light Young had to overcome the century-old view, expressed by Sir Isaac Newton' in his treatise "Optics", that light is a particle of colour.

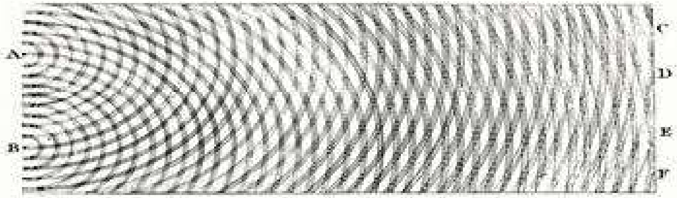

Thomas Young studied the interference of light waves by shining light through a screen with two slits equally separated, where the light emerging from the two slits would spread out and eventually superposition [overlap] each other to produce a distinct interference pattern

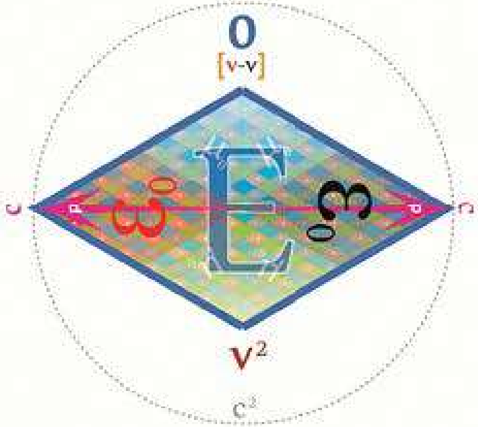

$$E = mv^2$$

"The term energy may be applied, with great propriety, to the product of mass or weight of a body, into the square of the number expressing its velocity.

Thus, if the weight of one ounce moves with a velocity of a foot in a second, we call its energy 1; if a second body of two ounces has a velocity of three feet in a second, its energy will be twice the square of three, or 18."

$$KE = \frac{1}{2}Mv^2$$

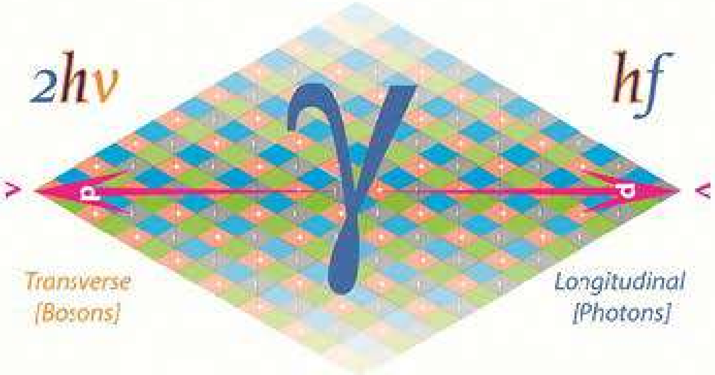

$$E_\gamma = 2mv^2$$

Photons are radiative Kinetic EM mass-energies

$$p^2 = KEM = Mv^2$$

All EM waves have energy momenta geometries whose physical properties can be modeled with statisical probabilities due to their equilateral wavefunctions

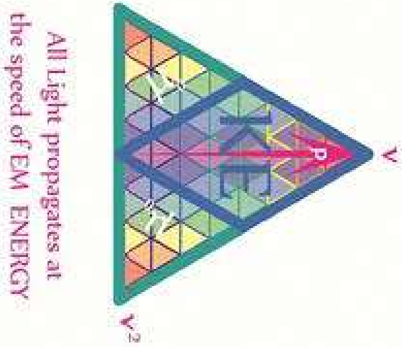

Tetryonics reveals the true geometry of Light

and dispels the misconceptions surrounding its physical properties of wave-length, energy momenta, frequency and associated wave-function

Photon-EM Wave Superpositioning

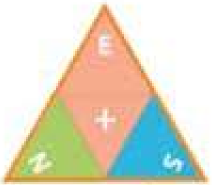

Positive quantum charge state
is one side of the quantum
energy momenta quoin

In physics, the Superposition principle, also known as superposition property, states that, for all linear systems, the net response at a given place and time caused by two or more stimuli is the sum of the responses which would have been caused by each stimulus individually.

$$F(\psi_1+\psi_2)=F(\psi_1)+F(\psi_2)$$

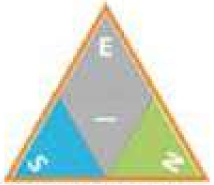

Negative quantum charge state
is the other side of the quantum
energy momenta quoin

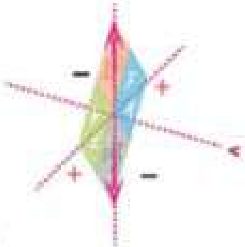

The linear momenta of transverse waves is orthogonal to the direction of wave propagation

Photons are neutral EM wavepackets

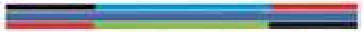

whose chiral quantum charges can be determined through Lorentz co-ordinate transformations

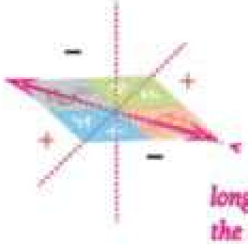

The linear momenta of longitudinal waves is co-linear to the direction of wave propagation

transverse Radio waves

longitudinal Spectral lines

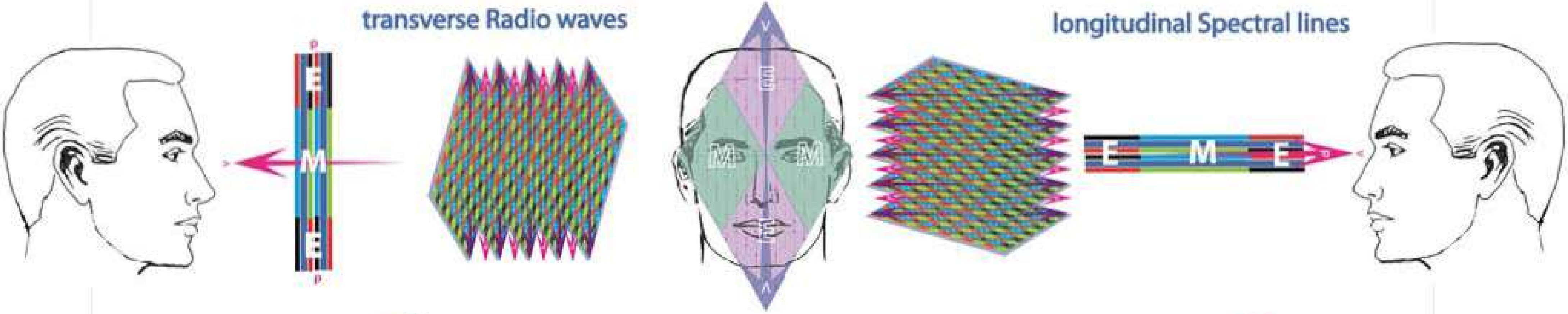

mc^{2^2} *Superpositioned EM waves can be likened to the cells of a chemical battery* mv^4

E^2

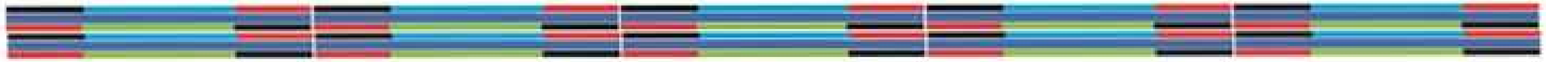

p^4

It is the superpositioning of longitudinal EM waveforms that creates white light and linear interaction-at-a-distance forces

Copyright ABRAHAM [2008] - all rights reserved

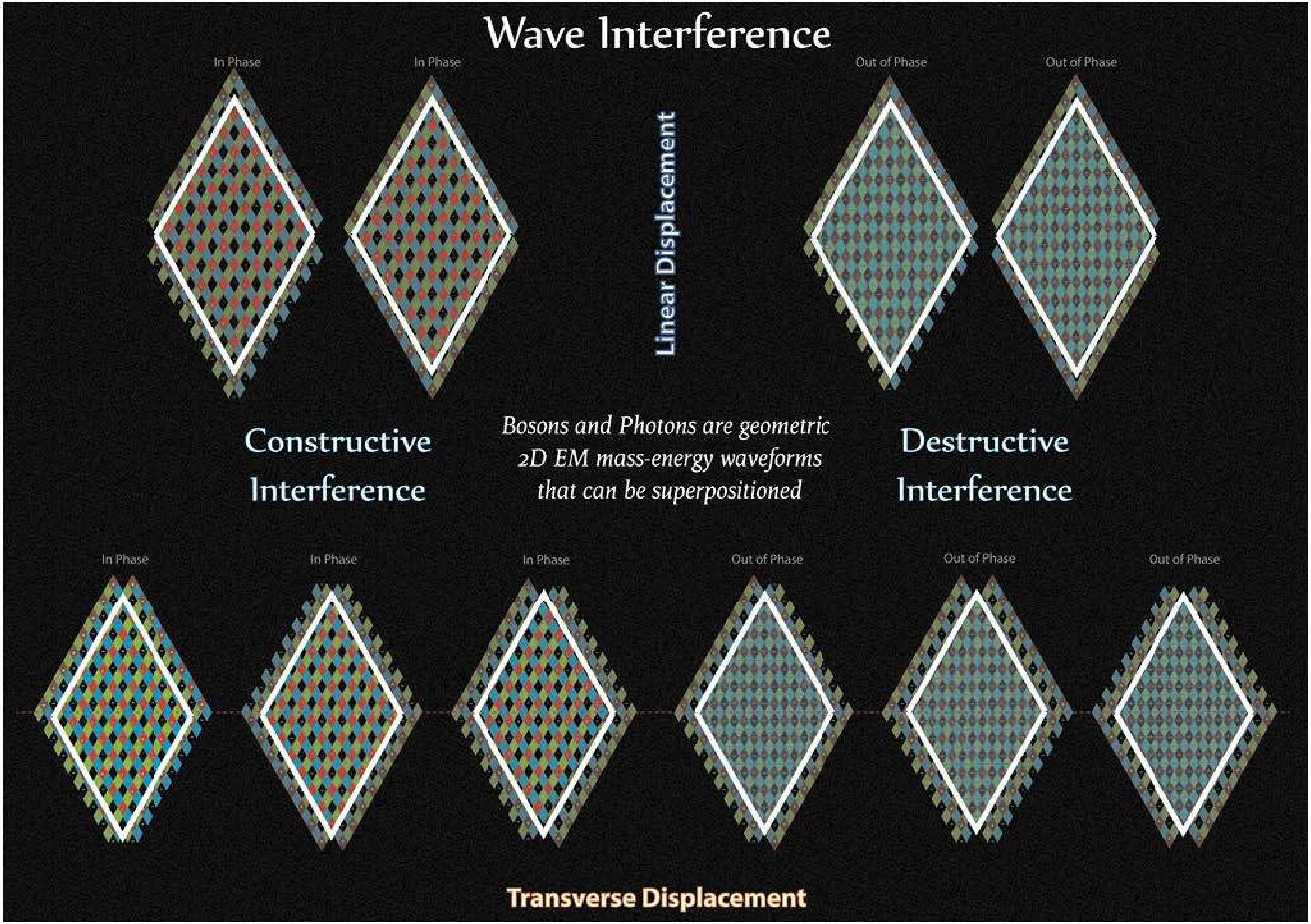

Wave Interference

In Phase
In Phase
Linear Displacement
Out of Phase
Out of Phase
In Phase
In Phase
In Phase
Out of Phase
Out of Phase
Out of Phase

Constructive Interference

Bosons and Photons are geometric 2D EM mass-energy waveforms that can be superpositioned

Destructive Interference

Transverse Displacement

Tetryonics 39.05 - Wave Interference - EM waves

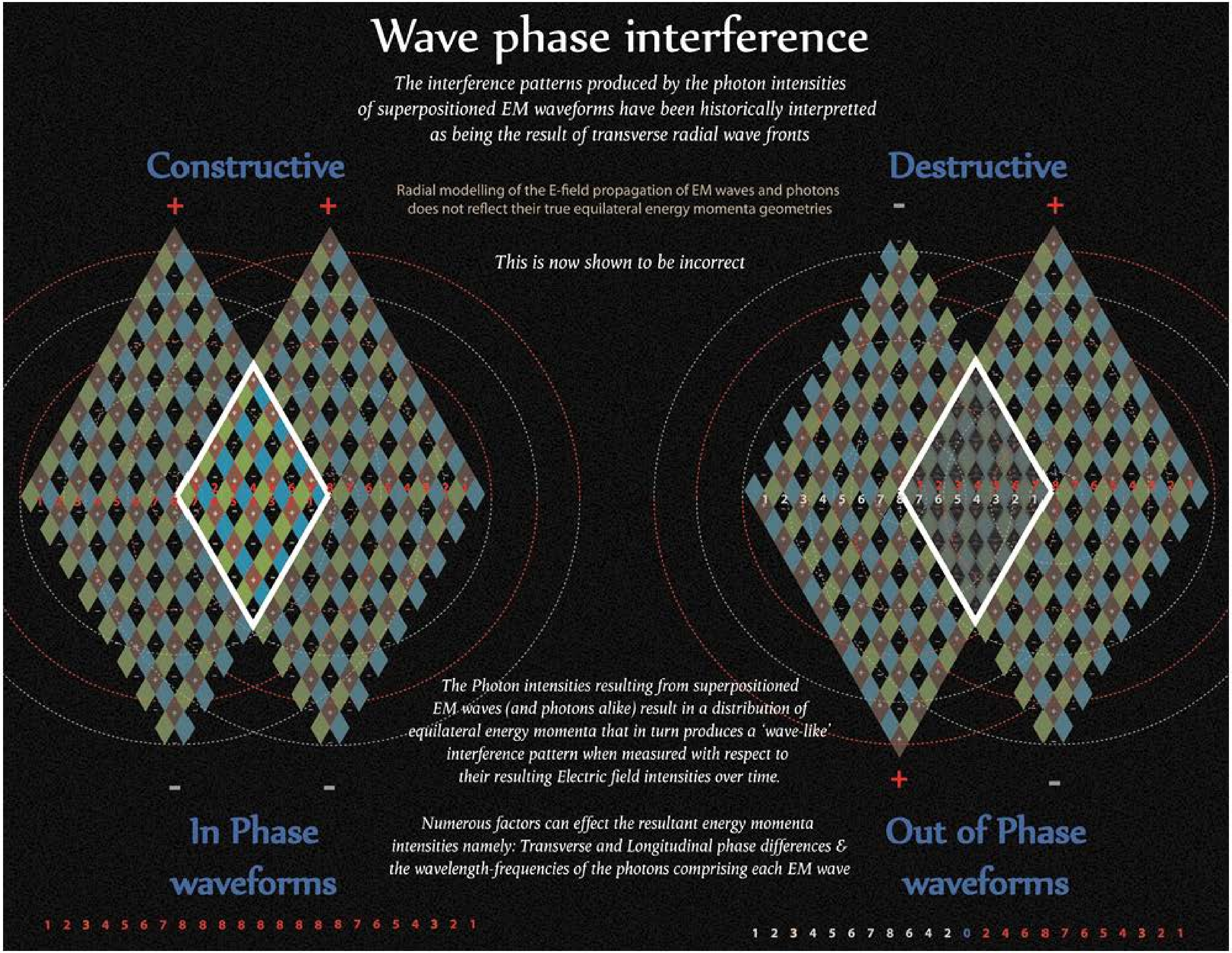

Wave phase interference

The interference patterns produced by the photon intensities of superpositioned EM waveforms have been historically interpretted as being the result of transverse radial wave fronts

Constructive

Destructive

Radial modelling of the E-field propagation of EM waves and photons does not reflect their true equilateral energy momenta geometries

This is now shown to be incorrect

The Photon intensities resulting from superpositioned EM waves (and photons alike) result in a distribution of equilateral energy momenta that in turn produces a 'wave-like' interference pattern when measured with respect to their resulting Electric field intensities over time.

Numerous factors can effect the resultant energy momenta intensities namely: Transverse and Longitudinal phase differences & the wavelength-frequencies of the photons comprising each EM wave

In Phase waveforms

Out of Phase waveforms

Tetryonics 39.06 - Wave phase interference

Copyright ABRAHAM [2008] - all rights reserved

Transverse EM mass-energies

Matter in motion

Longitudinal EM mass-energies

$E = h\nu$

Historically modeled through relativistic stress energy tensors
all Matter in motion is possessive of distinct energy waveforms:
a standing wave mass-energy geometry [Matter particle] &
a divergent kinetic mass-energy field [KEM wave]

$E = hf$

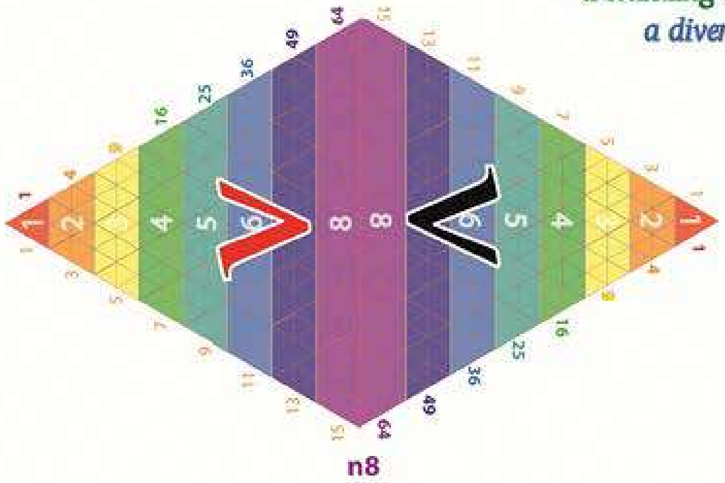

1
2
3
4
5
6
V
8
8
V
6
5
4
3
2
1
4
9
16
25
36
49
64
n8

$T\mu\nu$

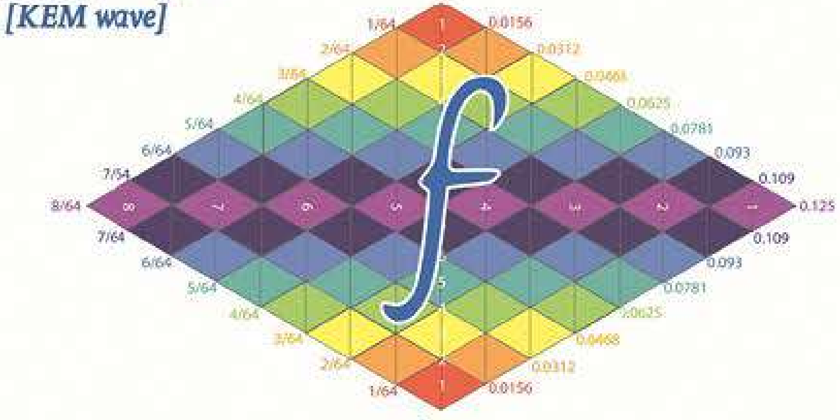

1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
f

f=64

EM induction and forces are effected by
neutral bi-directional KEM mass-energy fields
with interactive energy momenta

Negative charges in motion

Positive charges in motion

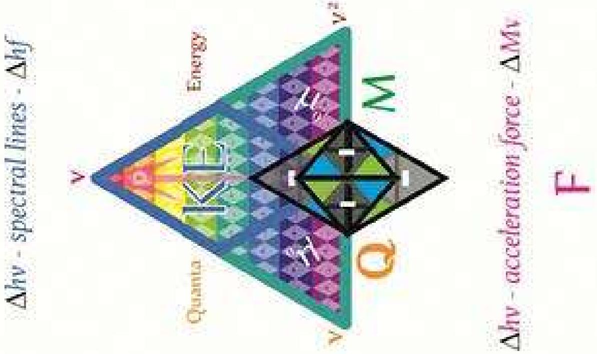

Δhv - spectral lines - Δhf
Energy
Quanta
KE
M
Q
v
v²
Δhv - acceleration force - ΔMv
F

unidrectional negative KEM fields
with interactive energy momenta

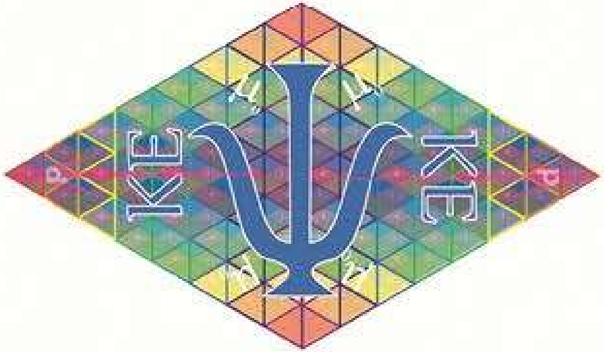

electromagnetic waves

[neutral bosons and/or photons]

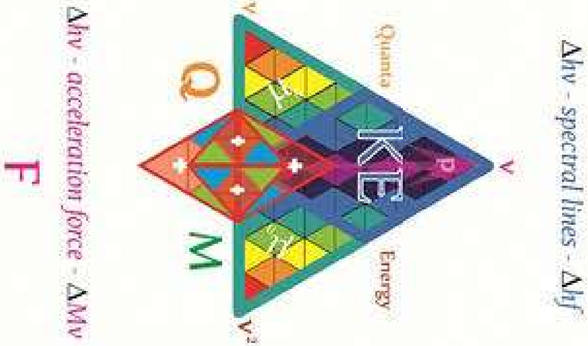

F
Δhv - acceleration force - ΔMv
Q
M
KE
Quanta
Energy
v
v²
Δhv - spectral lines - Δhf

unidirectional positive KEM fields
with interactive energy momenta

Tetryonics 39.07 - Matter in motion

Copyright ABRAHAM [2008] - all rights reserved

All Matter in motion [or not]
has both an intrinsic Matter-wave function
and an extrinsic KEM mass-energy wave function

Transverse boson wavefunction

EM Field · Planck quanta

$$\text{ODD}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right] \qquad E = n.hv$$

Bosons · ElectroMagnetic · mass · velocity

128

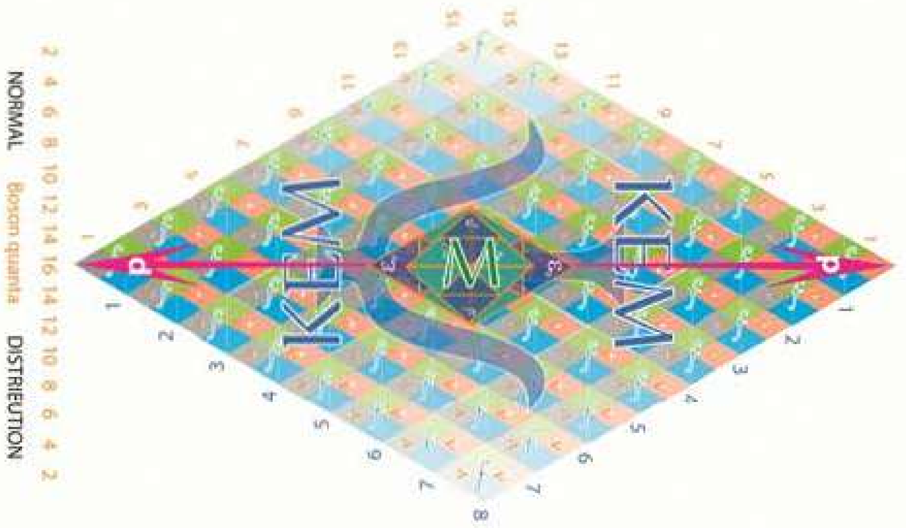

EM waves · EM Field · Planck quanta

$$\text{EVEN}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

ElectroMagnetic · mass · velocity

64

EM Field · Planck quanta

$$2\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right] \qquad E = hf$$

Photons · ElectroMagnetic · mass · velocity

Longitudinal photon wavefunction

Photon Intensity

Photon Intensity is a result of the geometry of EM waves
[which in turn is directly proportional to the velocity of the mass-Matter]
and can be measured in a number of varying ways related to
the mass-energy momenta of the EM wave

Wavenumbers, frequency or wavelength
Transverse EM masses [BOSONS]
Longitudinal EM masses [PHOTONS]
Wavefunction probabilities [WAVE-FUNCTIONS]

The wave–particle nature of mass-Matter in motion
has been the subject of much debate since Lucretius in 55BC with the
debate intensifying with Newton's and Young's differing views and the
development of Planck-Einstein quantum mechanics of the photo-electric effect

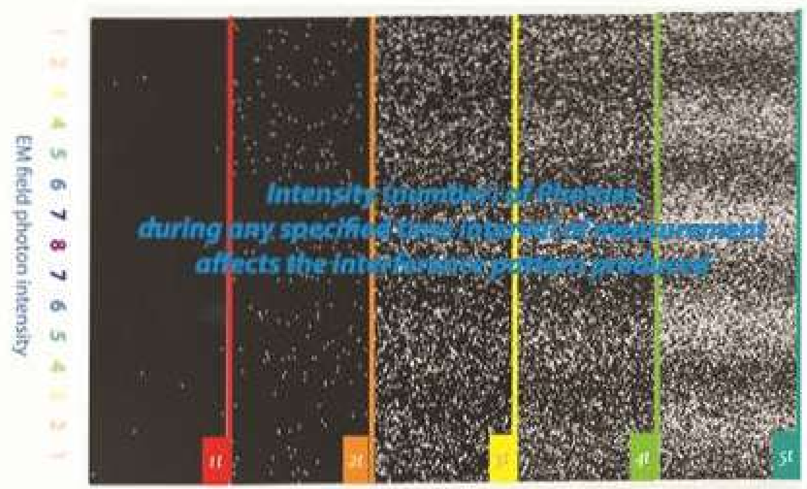

Intensity (number) of Photons
during any specified time interval of measurement
affects the interference pattern produced

The measurement of Photons with the use
of Photo-multipliers and Charge coupled Devices
results in the measurement of the E field properties
of EM waves (producing the long confusing wave-particle
results obtained by diffraction gratings)

In turn these results have been historically misinterpreted
as a waveform property that cannot be attributed to
the particle properties of a Photon.

Tetryonic geometry clears this matter up once and for all.

Copyright ABRAHAM [2008] - all rights reserved

Interference Patterns

ε_0

When EM waves pass through the slits they are detected via their E-fields as photons of varying strength energy momenta producing the impression of a interference pattern

μ_0

$E = h\nu$

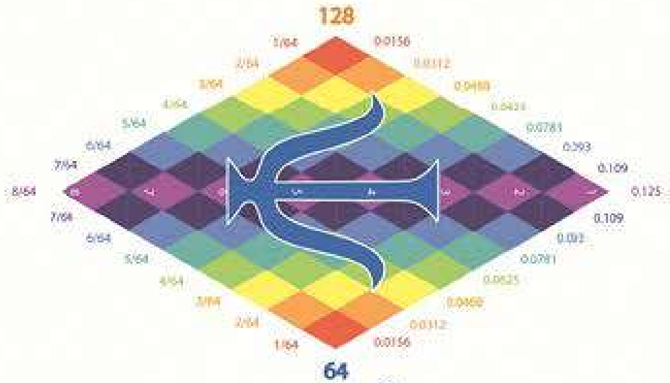
128
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
0.109
0.093
0.0781
0.0625
0.0468
0.0312
0.0156
64

$E = hf$

Every EM wave is comprised of identical, specific wavelength photons which are arranged in a Normal distribution resulting from the electromagnetic Wavefunction with a peak value equal to the wave's amplitude

[the square root of the EM wave's Wavefunction/Probability amplitude]

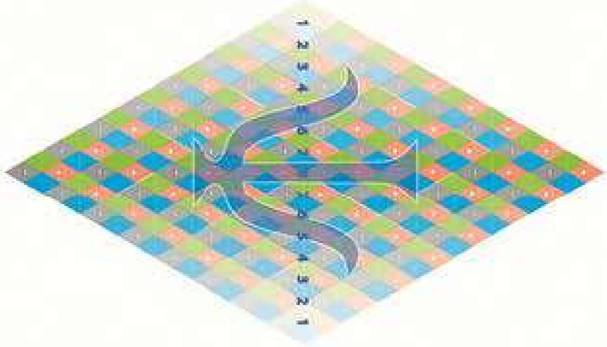

Photons impact screens and photo-detectors with intensities that are determined by their energy momenta distributions

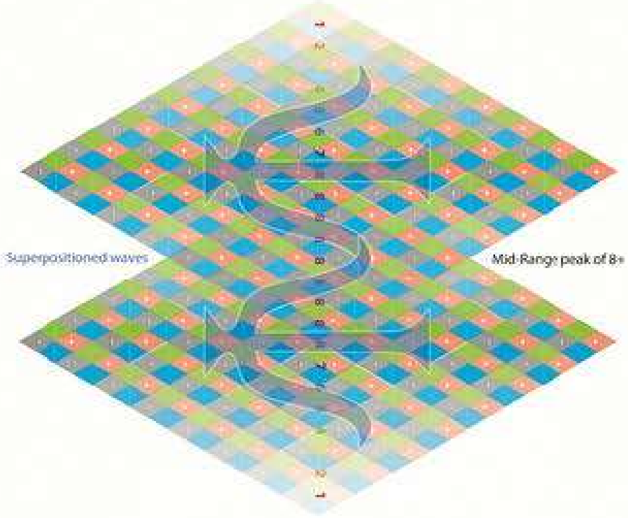
Superpositioned waves
Mid-Range peak of 8+

The resultant amplitudes are a direct result of the Phase of the Superpositioned photons within EM waves

Superpositioned wavefunctions

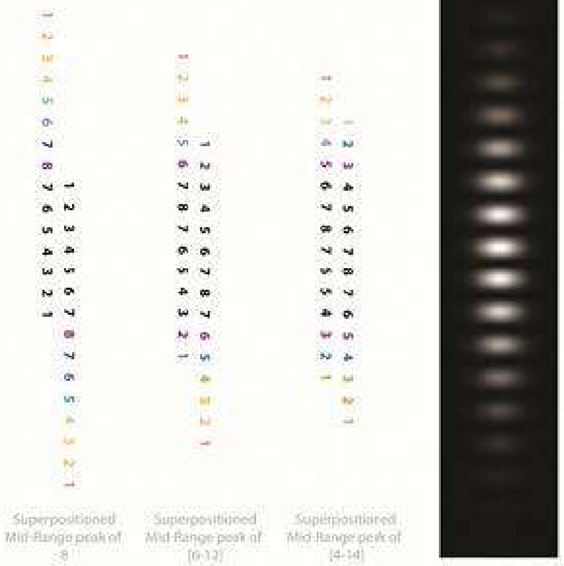
Superpositioned Mid-Range peak of 8
Superpositioned Mid-Range peak of [6-12]
Superpositioned Mid-Range peak of [4-14]

ALL EM mass-ENERGY Matter geometries being comprised of energy-momenta quanta are capable of producing interference patterns

The amplitudes of the detected photons in the resultant interference patterns are determined by their constructive/destructive superpositioning in each wave explaining the interference patterns currently accounted for by Young's wave theory

Single slit experiment

If light consisted strictly of ordinary or classical particles, and these particles were fired in a straight line through a slit and allowed to strike a screen on the other side, we would expect to see a pattern corresponding to the size and shape of the slit. However, when this "single-slit experiment" is actually performed, the pattern on the screen is a diffraction pattern with a fairly narrow central band with ever diminishing bands parallel to it on either side

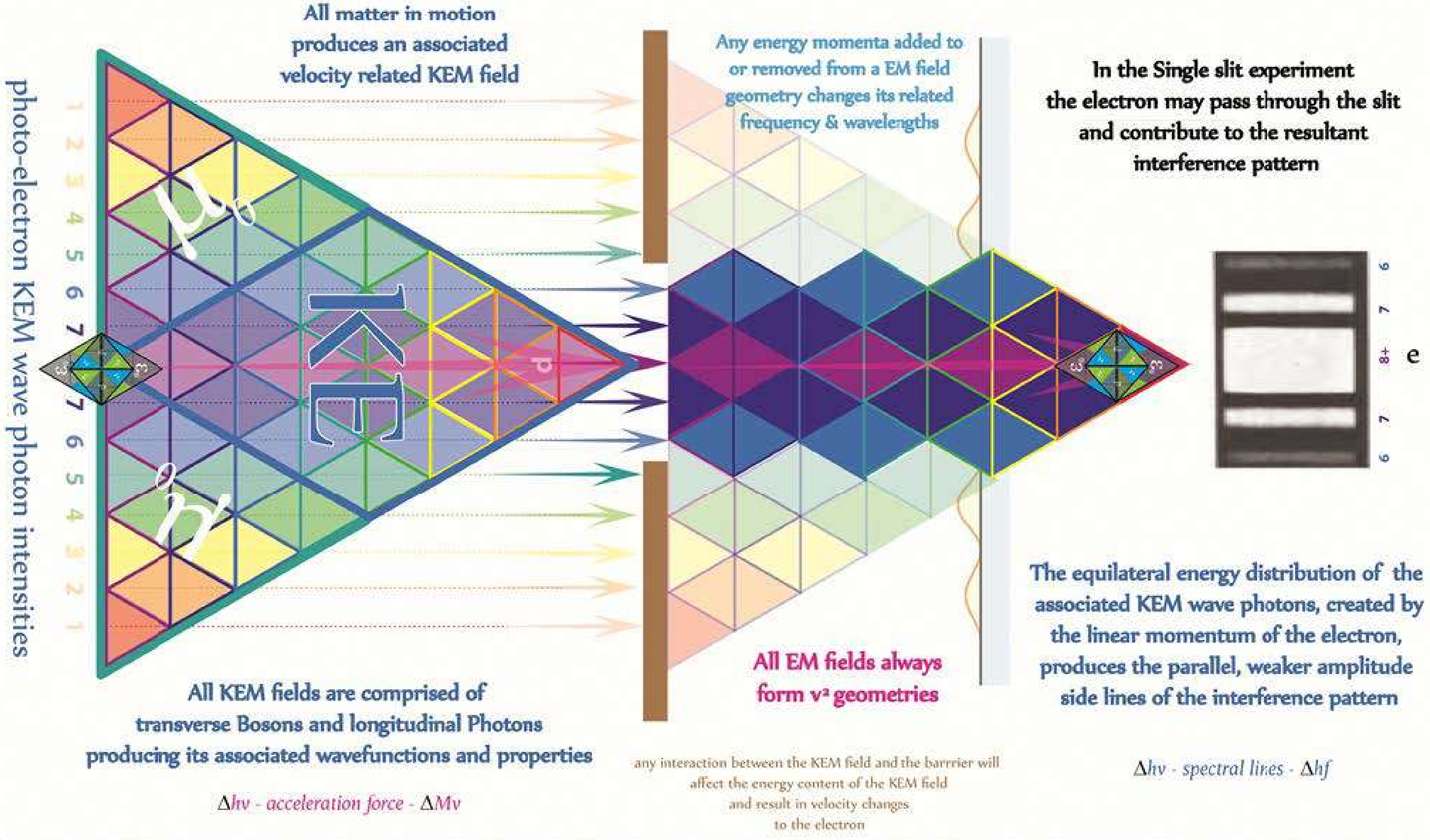

All matter in motion produces an associated velocity related KEM field

Any energy momenta added to or removed from a EM field geometry changes its related frequency & wavelengths

In the Single slit experiment the electron may pass through the slit and contribute to the resultant interference pattern

The equilateral energy distribution of the associated KEM wave photons, created by the linear momentum of the electron, produces the parallel, weaker amplitude side lines of the interference pattern

All KEM fields are comprised of transverse Bosons and longitudinal Photons producing its associated wavefunctions and properties

All EM fields always form v^2 geometries

Δhv - acceleration force - ΔMv

any interaction between the KEM field and the barrier will affect the energy content of the KEM field and result in velocity changes to the electron

Δhv - spectral lines - Δhf

Tetryonics 39.10 - Single slit experiment

Copyright ABRAHAM [2008] - all rights reserved

Double slit experiments

The double-slit experiment, sometimes called Young's experiment, is a demonstration that matter and energy can display characteristics of both waves and particles, and demonstrates the fundamentally probabilistic nature of quantum mechanical phenomena and Establishes the quantum interference principle known as wave–particle duality.

Matter particles are stopped by the barrier but the [K]EM wave passes through both slits and is diffracted by them producing weaker EM waves that then superposition with each other to produce interference patterns

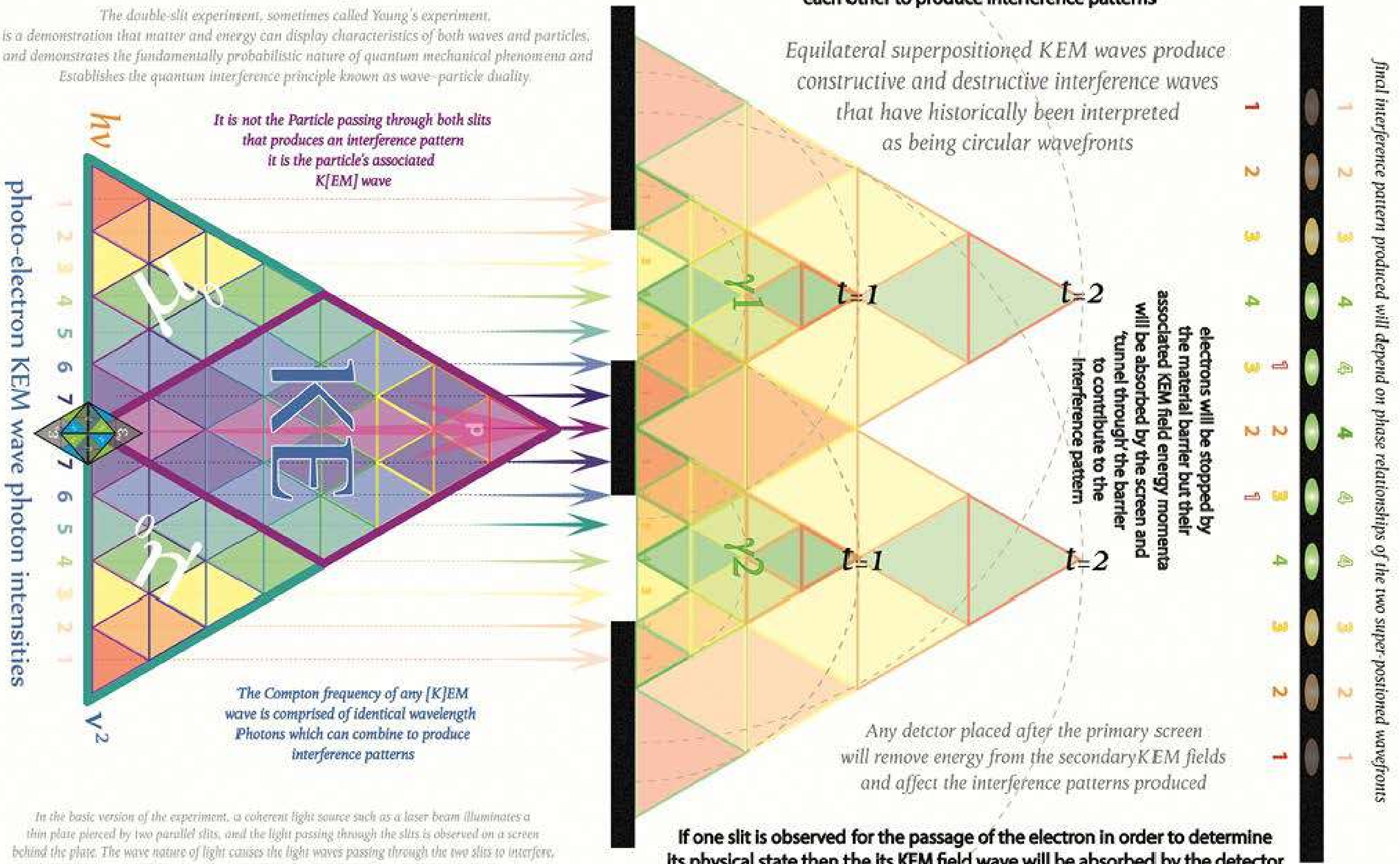
hν
v²
photo-electron KEM wave photon intensities
1 2 3 4 5 6 7 7 6 5 4 3 2 1
μe
μo
KE
p
γ1
γ2
t=1
t=2
t=1
t=2
1 2 3 4 3 2 1
1 2 3 4 3 2 1
1 2 3 4 4 4 4 4 3 2 1
final interference pattern produced will depend on phase relationships of the two super-postioned wavefronts

It is not the Particle passing through both slits that produces an interference pattern it is the particle's associated K[EM] wave

Equilateral superpositioned KEM waves produce constructive and destructive interference waves that have historically been interpreted as being circular wavefronts

electrons will be stopped by the material barrier but their associated KEM field energy momenta will be absorbed by the screen and 'tunnel through' the barrier to contribute to the interference pattern

The Compton frequency of any [K]EM wave is comprised of identical wavelength Photons which can combine to produce interference patterns

Any detector placed after the primary screen will remove energy from the secondaryKEM fields and affect the interference patterns produced

If one slit is observed for the passage of the electron in order to determine its physical state then the its KEM field wave will be absorbed by the detector resulting in only one wave remaining, enforcing a classical particle outcome

In the basic version of the experiment, a coherent light source such as a laser beam illuminates a thin plate pierced by two parallel slits, and the light passing through the slits is observed on a screen behind the plate. The wave nature of light causes the light waves passing through the two slits to interfere, producing bright and dark bands on the screen — a result that would not be expected if light consisted strictly of particles. However, on the screen, the light is always found to be absorbed as though it were composed of discrete particles or photons.

Tetryonics 39.11 - Double slit experiment

Light wave Reflection

Excited (or higher energy)
Atoms and Molecules can
emit light and energy via
Electro-dynamics

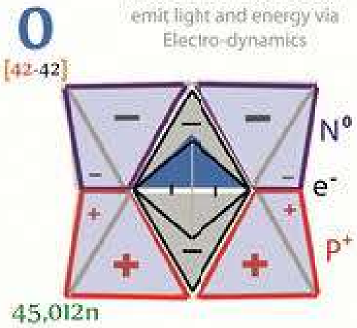
0
[42-42]
N°
e-
P+
45,012n

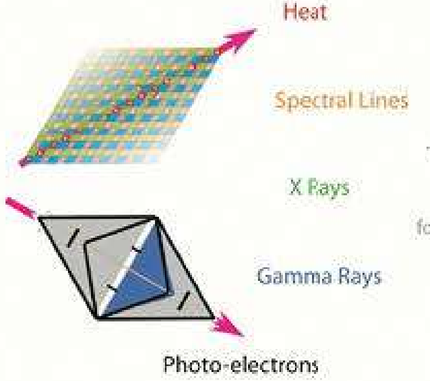
Heat
Spectral Lines
X Rays
Gamma Rays
Photo-electrons

The emitted photon/particle
can go in any direction
so this is the mechanism
for diffuse reflection/radiation.

All Atoms and Molecules
have ANGULAR charged fascia
in their topologies

Photon Re-Transmission

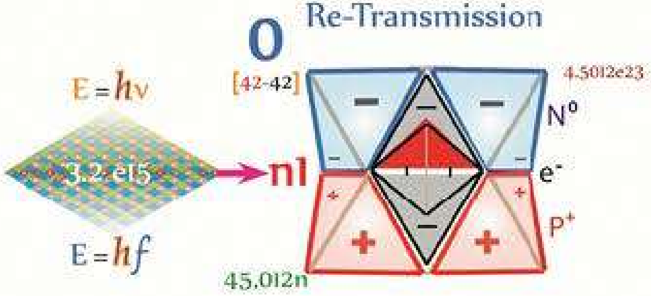
E=hv
3.2e15
E=hf
0
[42-42]
n1
4.5012e23
N°
e-
P+
45,012n

When a photon strikes a atom or molecule
it generally imparts EM energy as heat
and the photon is absorbed

Photon absorption
Photon emission

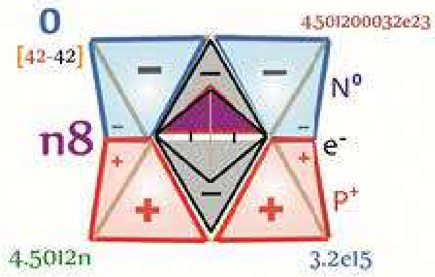
0
[42-42]
n8
4.501200032e23
N°
e-
P+
4.5012n
3.2e15

For some photons their EM energy is equal
to the resonance energy of the molecule.

When this happens the photon is absorbed
and the molecule is put into an excited state

Incoming light interacts with a surface and may be absorbed, refelected, and/or transmitted.

Materials have a reflectance spectrum which is a function of the angle of incidence of the incoming light.

The color of an object is a function of the color spectrum of the incident light and the reflectance spectrum of the object's surface.

Angle of Incidence
equals
Angle of Reflection

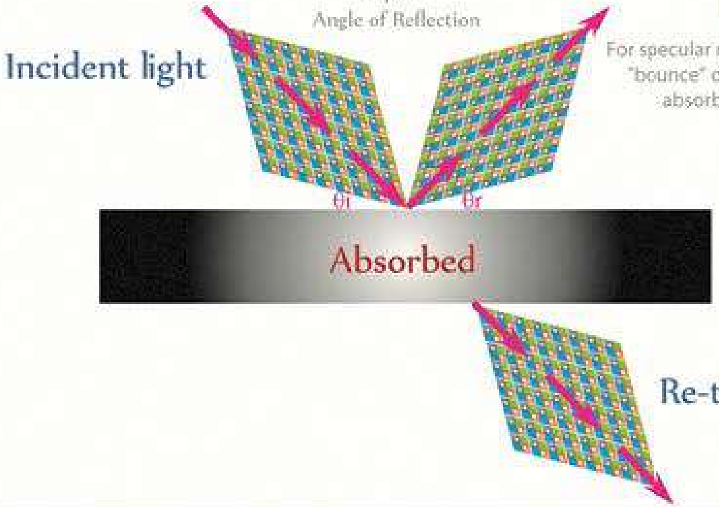
Incident light
θi
θr
Absorbed
Re-transmitted

Reflected

For specular reflection the photons actually
"bounce" off of the surface and are not
absorbed and then re-radiated

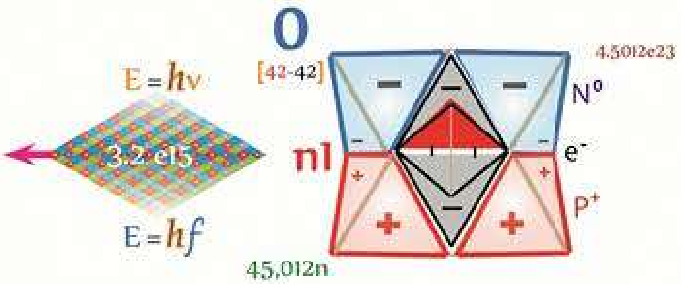
E=hv
3.2e15
E=hf
0
[42-42]
n1
4.5012e23
N°
e-
P+
45,012n

The molecule relaxes from the excited state back to
the ground state by emitting a photon with EM energy
equal to that of the original photon so the photon
appears to have been reflected
(but is actually absorbed then emitted)

Copyright ABRAHAM [2008] - all rights reserved

Lightwave Refraction

The refraction of light when it passes from a fast medium to a slow medium bends the light ray toward the normal at the boundaries between any two media.

The amount of bending depends on the indices of refraction of the two media and is described quantitatively by Snell's Law.

White light can be refracted into its component spectral colours

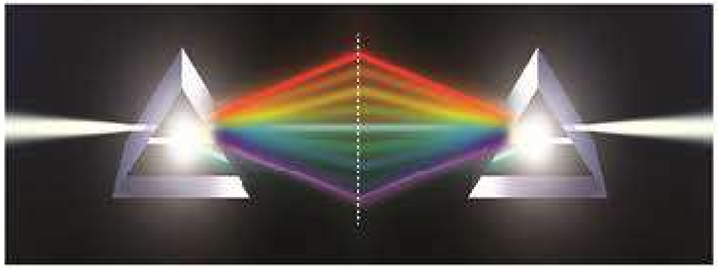

or be re-combined back into white light through superpositioning

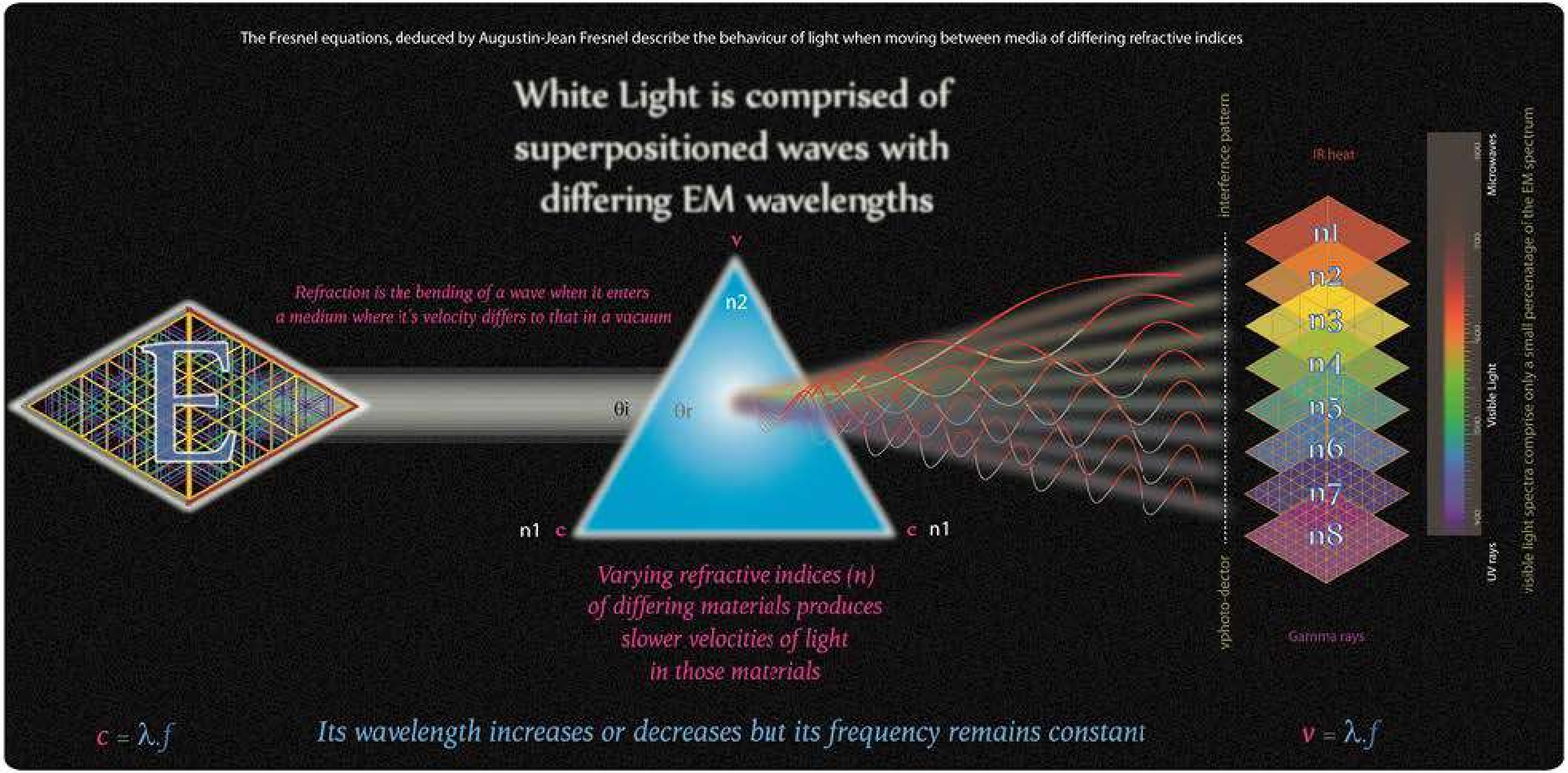

The Fresnel equations, deduced by Augustin-Jean Fresnel describe the behaviour of light when moving between media of differing refractive indices

White Light is comprised of superpositioned waves with differing EM wavelengths

Refraction is the bending of a wave when it enters a medium where it's velocity differs to that in a vacuum

Varying refractive indices (n) of differing materials produces slower velocities of light in those materials

$c = \lambda . f$ *Its wavelength increases or decreases but its frequency remains constant* $v = \lambda . f$

E
v
n2
θi
θr
n1 c
c n1
interference pattern
vphoto-dector
IR heat
n1
n2
n3
n4
n5
n6
n7
n8
Gamma rays
Microwaves
Visible Light
UV rays
Visible light spectra comprise only a small percenatage of the EM spectrum

Ball Lightning

An alternative radiant geometry for electromagnetic waves exists

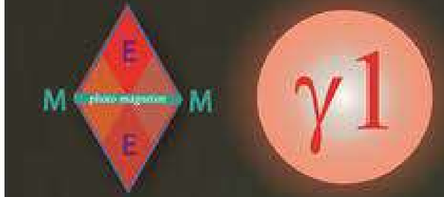

The EM energies in radiant 'ball' lightning are transformer-couplings of energy momenta & self-sustaining until any interaction with Matter topologies occurs

Descriptions of ball lightning vary wildly. It has been described as moving up and down, sideways or in unpredictable trajectories, hovering and moving with or against the wind; attracted to, unaffected by, or repelled from buildings, people, cars and other objects.

Ball lightning has been described as transparent, translucent, multicolored, evenly lit, radiating flames, filaments or sparks, with shapes that vary between spheres, ovals, tear-drops, rods, or disks

Any interaction with a 'ball' of EM energies will result in an explosive discharge of it's EM energy momenta

Nikola Tesla was reportedly able to artificially produce 1.5" (3.8 cm) balls using spark gap technologies

Ball lightning creates a strong Magnetic moment and divergent E fields [similar to a bar magnet without the Matter] creating a "Ball' of light with varying EM energies that is sensitive to external M fields

Wave~Particle interaction-at-a-distance

All Matter in motion has both:

a standing wave energy topology [Particle] and

a divergent relativistic, Lorentz velocity corrected geometric field of mass-energy momenta [KEM Wave]

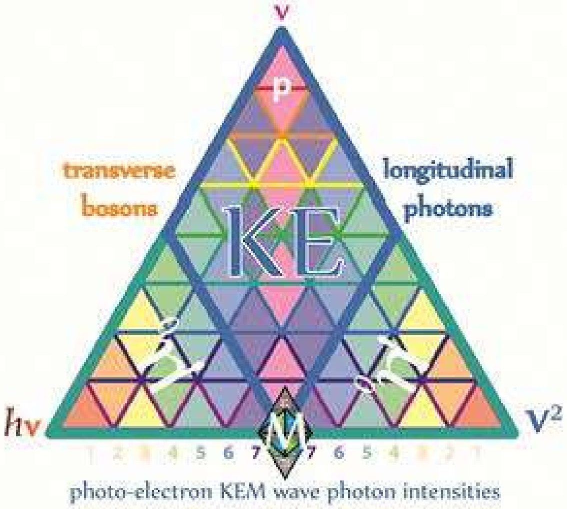
v
p
transverse bosons
longitudinal photons
KE
hv
v²
M
1 2 3 4 5 6 7 7 6 5 4 3 2 1
photo-electron KEM wave photon intensities

Whether it is Light itself or the KEM field of Matter in motion all EM fields are comprised of transverse bosons and longitudinal photons with probabillistic wave geometries

$$RE = M + KEM$$

relativistic energies — Matter — kinetic energies

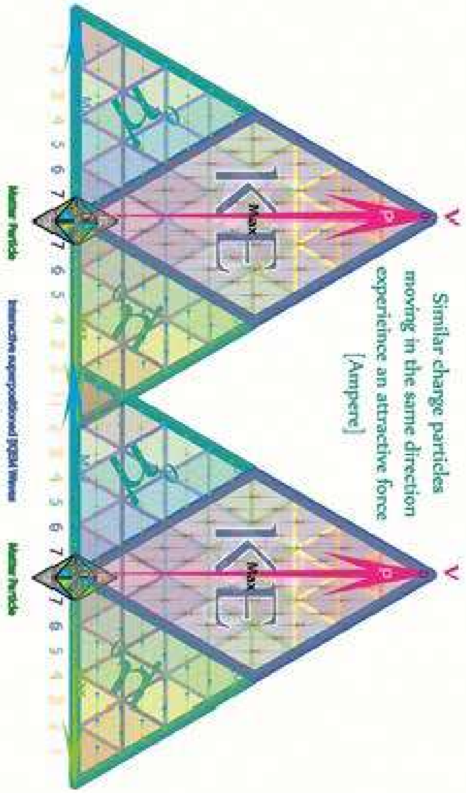
Matter Particle
Interactive superpositioned [K]EM Waves
Matter Particle
KE
Max
p
v
KE
Max
p
v

Similar charge particles moving in the same direction experience an attractive force [Ampere]

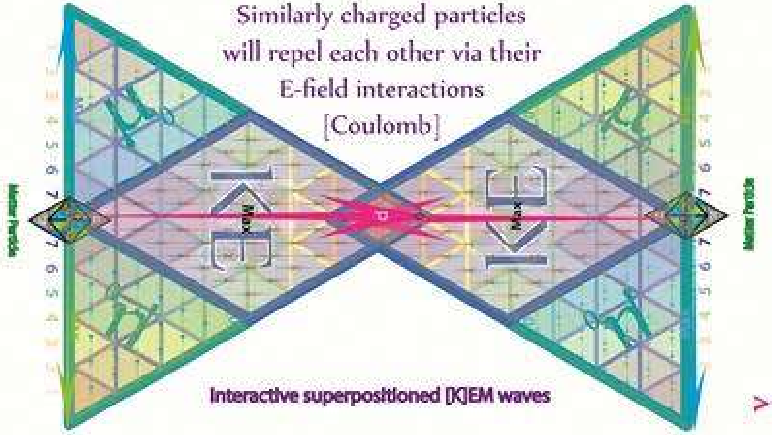
Matter Particle
KE
Max
p
KE
Max
Matter Particle

Similarly charged particles will repel each other via their E-field interactions [Coulomb]

Interactive superpositioned [K]EM waves

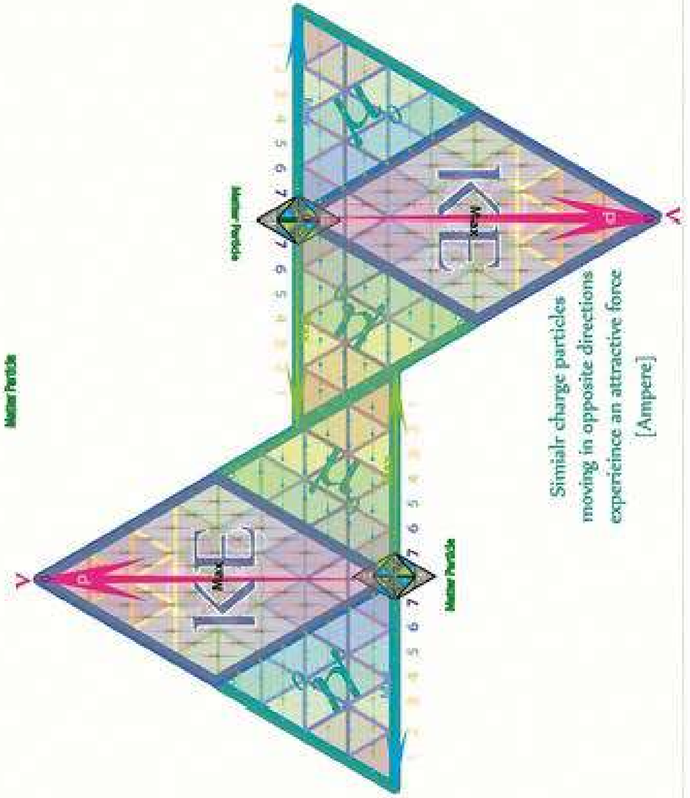
Matter Particle
KE
Max
p
v
v
p
KE
Max
Matter Particle

Simialr charge particles moving in opposite directions experince an attractive force [Ampere]

Force fields and interaction-at-a-distance

All Matter in motion produces an associated Lorentz velocity corrected KEM field of interactive energy momenta photons

Tetryonics 39.15 - Wave~Particle interactions

Quantum Tunnelling

Charged
12
[0-12]
Matter
1.2 e20

Particles attempting to travel between potential barriers can be compared to a ball trying to roll over a hill; quantum mechanics and classical mechanics differ in their treatment of this scenario.

Classical mechanics predicts that particles that do not have enough energy to classically surmount a barrier will not be able to reach the other side. Thus, a ball without sufficient energy to surmount the hill would roll back down.

Or, lacking the energy to penetrate a wall it would bounce back (reflection) or in the extreme case, bury itself inside the wall (absorption).

In quantum mechanics, these particles can, with a very small probability, tunnel to the other side, thus crossing the barrier.

Kinetic EM energies
v
v
v²
KE
p

The quantum magician's slight of hand trick is where the incident electron is swapped another

All electrons are IDENTICAL

It is practically impossible to identify a particular electron and track it

even using spin orientations only eliminates half of the total electrons in the barrier

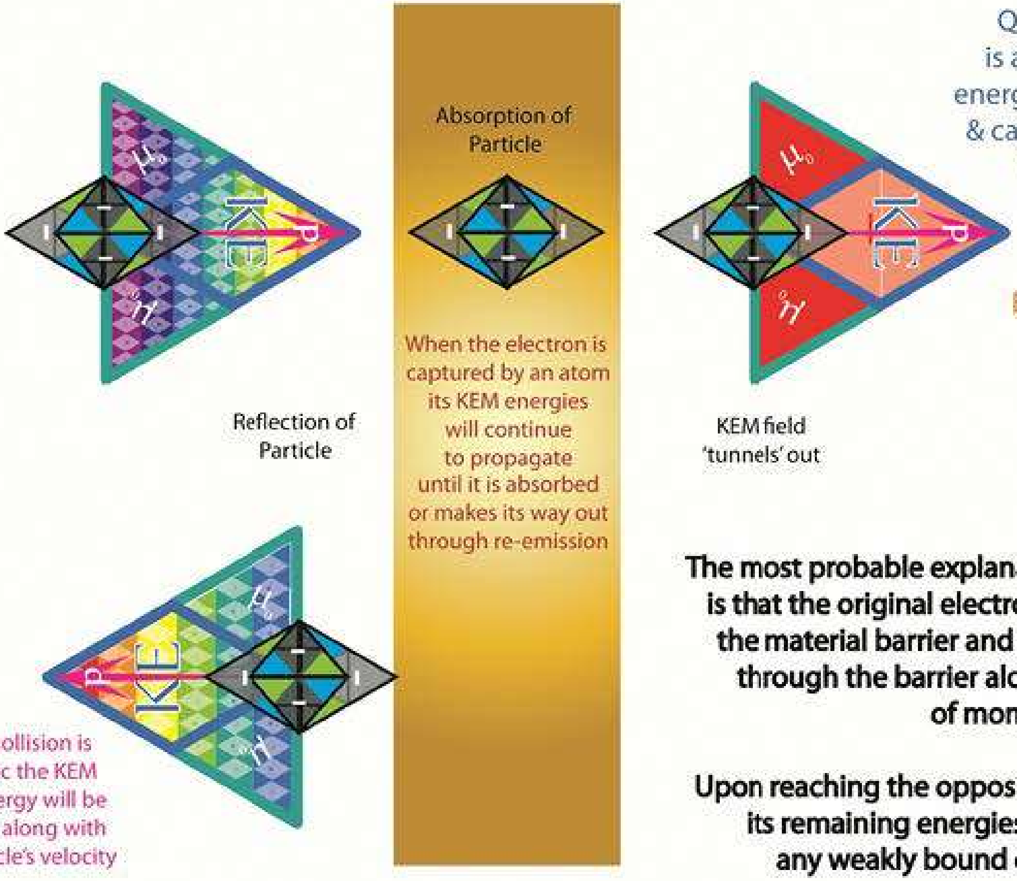

Reflection of Particle

Absorption of Particle

When the electron is captured by an atom its KEM energies will continue to propagate until it is absorbed or makes its way out through re-emission

KEM field 'tunnels' out

If the collision is inelastic the KEM field energy will be reduced along with the particle's velocity

Quantum KEM field tunnelling is an example of conservation of energy momenta on the atomic scale & can be likened to a unidirectional version of Newton's cradle

Removal of all electrons but the one under measurement is impractical as it would result in an attractive coloumbic force that would trap the electron in the barrier

The most probable explanation of quantum tunneling is that the original electron was bound to atoms in the material barrier and its KEM field propagated through the barrier along its original direction of momentum.

Upon reaching the opposite side/edge of the barrier its remaining energies are able to accelerate any weakly bound electron found there

Quantum 'tunnelling' is revealed to be the result of an electron's KEM wave propagation through the barrier not the electron itself

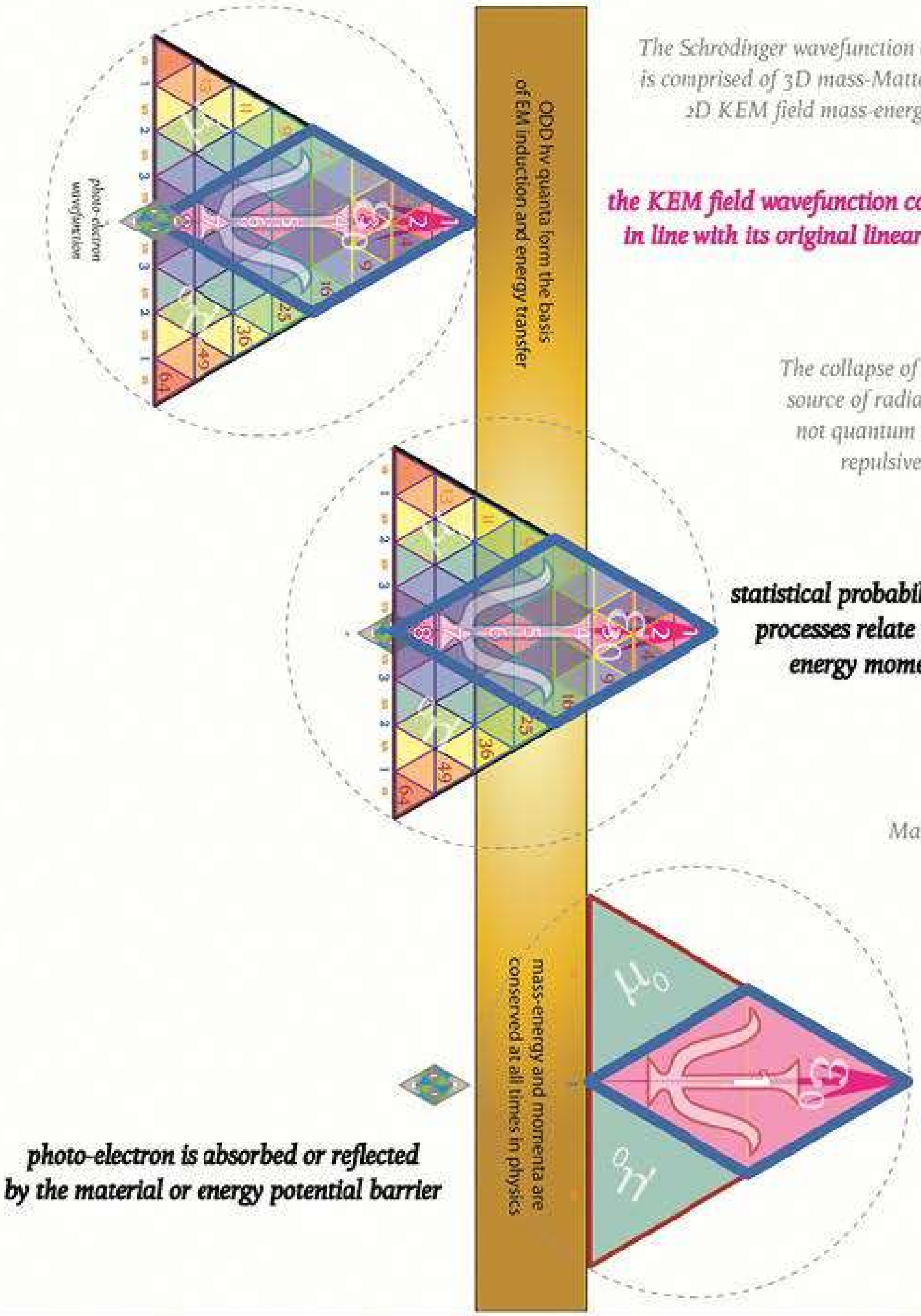
ODD hv quanta form the basis of EM induction and energy transfer
photo-electron wavefunction
1
4
9
16
25
36
49
64
mass-energy and momenta are conserved at all times in physics
μ_0
ε_0

The Schrodinger wavefunction of photo-electrons is comprised of 3D mass-Matter topologies and 2D KEM field mass-energy geometries

the KEM field wavefunction continues to propagate in line with its original linear momentum vector

The collapse of Matter topologies is the source of radiant stellar mass-energies not quantum tunnelling to overcome repulsive Coloumbic forces

statistical probabilities applied to quanutm tunneling processes relate to the distribution of EM Planck energy momenta quanta in the KEM field

Matter-energy conversion via GEM pinches is the source of stellar energies released in stars

photo-electron is absorbed or reflected by the material or energy potential barrier

As Planck energy momenta quanta interact or are absorbed the KEM field's wavefunction will entropy accordingly, in turn affecting the wavelength-frequency-velocity of the KEM field

$$\Delta hf = \Delta Mv = \Delta p$$

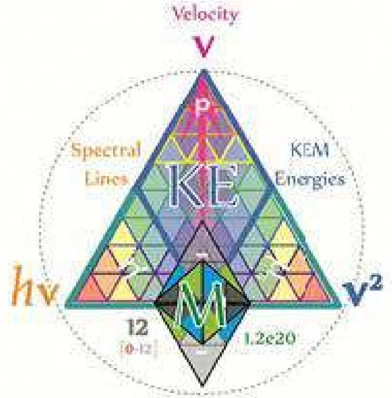
Velocity
v
p
Spectral Lines
KEM Energies
KE
hv
v^2
M
12
[0-12]
1.2e20

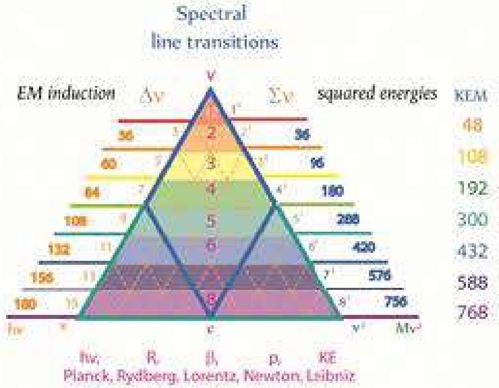
Spectral line transitions
EM induction
Δv
v
Σv
squared energies
KEM
48
108
192
300
432
588
768
36
96
180
288
420
576
756
36
60
84
108
132
156
180
2
3
4
5
6
7
8
hv
c
v^2
Mv^2
hv,
R,
β,
p,
KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity through equilateral geometry

Quantum tunneling plays no role in the 'fusion' of 'lighter elements into heavier elements in stars (which is itself an erroneous process)

Tetryonics 39.17 - mass-energy momenta induction & conservation

All Matter in motion is comprised of:

a mass-Matter standing-wave topology,
and
a KEM mass-energy field geometry

whose quantised energy momenta is reflective of the particle's vector velocity

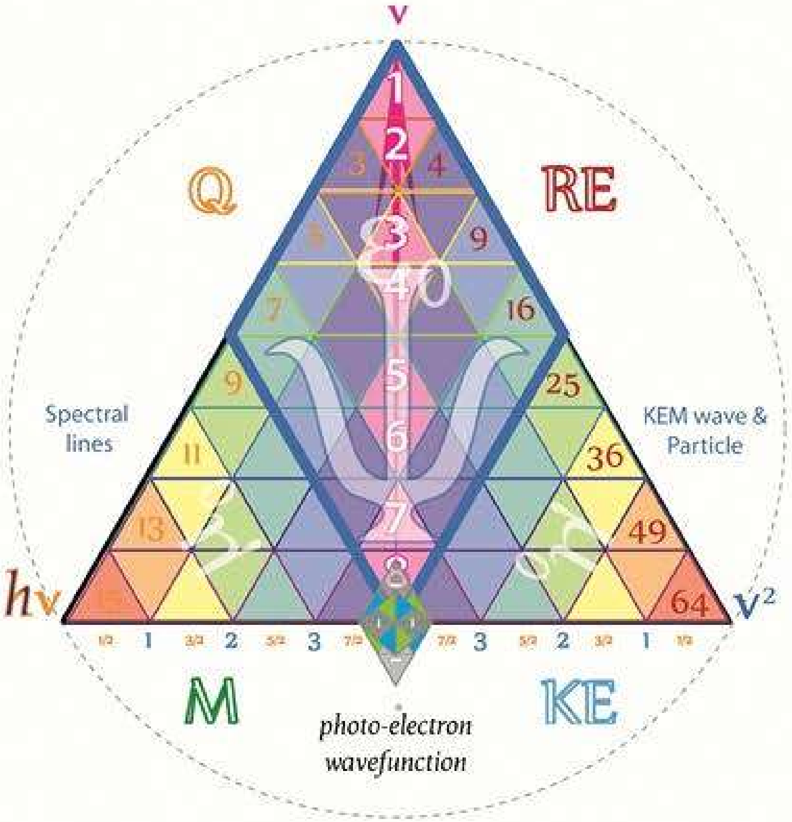
v
1
2
3
4
8
3
9
7
16
9
25
11
36
13
49
64
Q
RE
Spectral lines
KEM wave & Particle
hv
v²
½ 1 3/2 2 5/2 3 7/2 7/2 3 5/2 2 3/2 1 ½
M
KE
photo-electron wavefunction

The normal distribution of Planck energy momenta in any KEM field is the quantum foundation for the statistical probability mechanics of quanutm mechanics and quantum electrodynamics

Quantum Tunneling

it makes no difference if the barrier is a EM field or material object

electrons are NOT spherical point particles

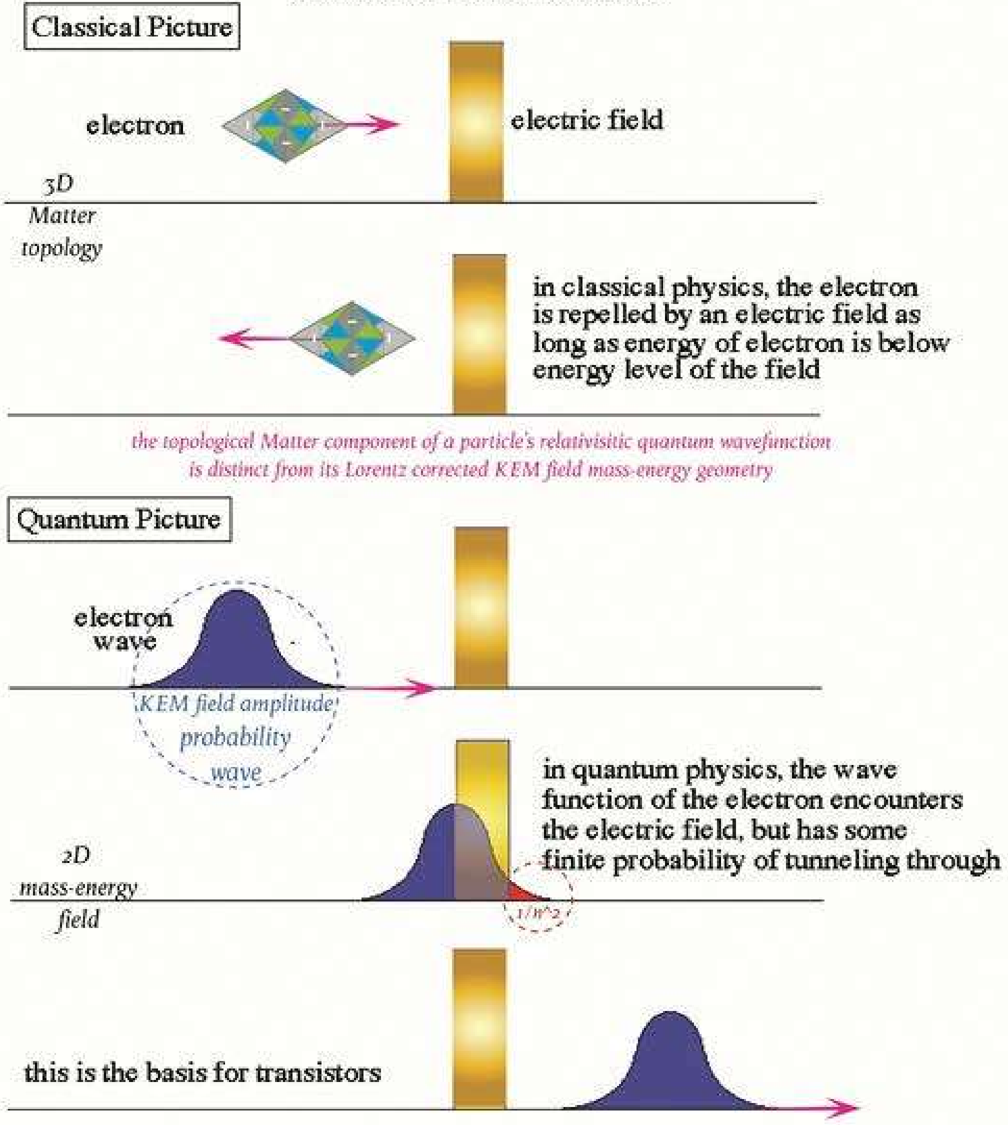
Classical Picture
electron
electric field
3D Matter topology
Quantum Picture
electron wave
KEM field amplitude probability wave
2D mass-energy field
1/n^2

in classical physics, the electron is repelled by an electric field as long as energy of electron is below energy level of the field

the topological Matter component of a particle's relativisitic quantum wavefunction is distinct from its Lorentz corrected KEM field mass-energy geometry

in quantum physics, the wave function of the electron encounters the electric field, but has some finite probability of tunneling through

this is the basis for transistors

quantum tunnelling is often also utilised in modern physics as a means of overcoming Coulombic charge barriers of Protons in stars in order to explain 'fusion' as the source of stellar energy

Tetryonics 39.17 - Quantum Tunnelling wavefunctions

Relativistic Kinetic Energies

$$KE = \frac{1}{2}M\frac{v^2}{c^2}$$

The relativistic Kinetic mass-energies of all Matter topologies in motion are contained in their KEM fields and are subject to Lorentz velocity corrections

$$p^2 = E = m\frac{v^2}{c^2}$$

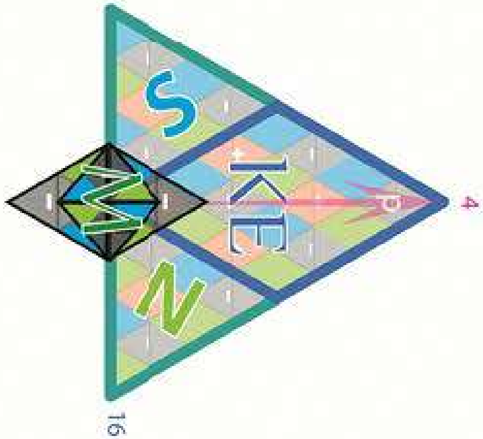

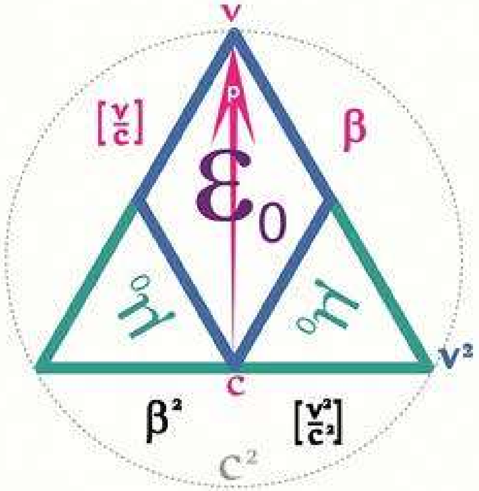
v
[v/c]
β
p
ε0
μ0
μ0
c
v²
β²
[v²/c²]
c²

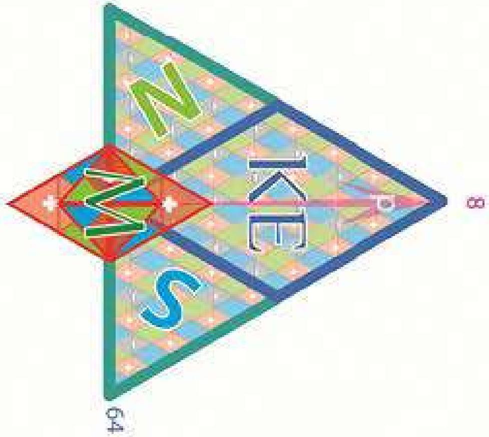

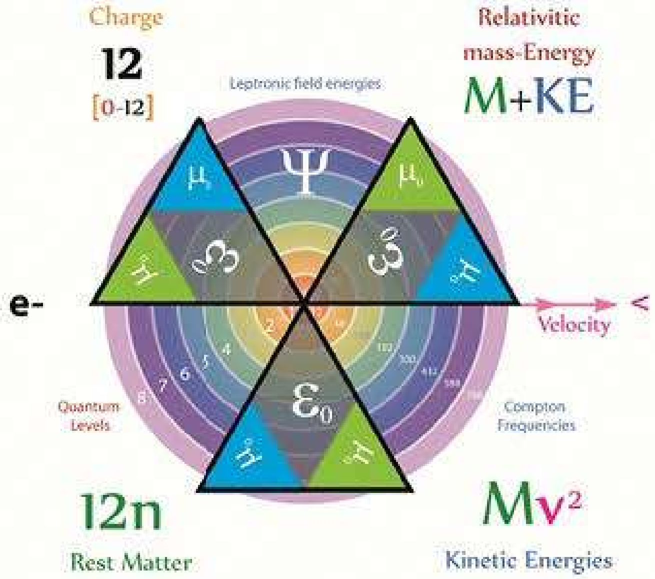
Charge
12
[0-12]
Leptronic field energies
Relativitic mass-Energy
M+KE
μ0
Ψ
μ0
ε0
ε0
μ0
μ0
e-
Velocity
v
ε0
μ0
μ0
Quantum Levels
Compton Frequencies
12n
Rest Matter
Mv²
Kinetic Energies

Lorentz contractions only apply to the KEM fields of Matter in motion

The Kinetic Energies of a particle in motion can be expressed in terms of its linear momentum

$$\frac{p^2}{2m} = KE = \frac{1}{2}Mv^2$$

or equally as 1/2 the total relativistic EM mass-energies [exclusive of its invariant rest mass-Matter]

Charge and Matter topologies are Lorentz invariant

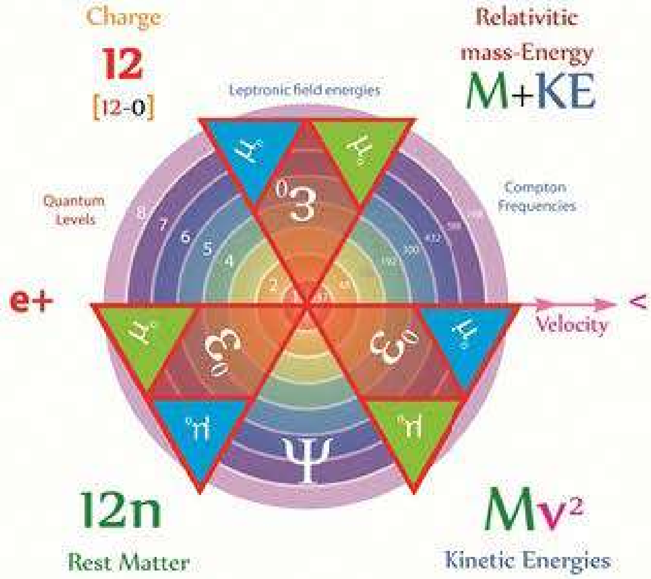
Charge
12
[12-0]
Leptronic field energies
Relativitic mass-Energy
M+KE
Quantum Levels
Compton Frequencies
μ0
μ0
ε0
e+
Velocity
v
ε0
ε0
μ0
μ0
μ0
μ0
Ψ
12n
Rest Matter
Mv²
Kinetic Energies

Copyright ABRAHAM [2008] - all rights reserved

Relativistic Lorentz Corrections

ges their name from their earliest appearance, in Lorentzian electrodynamics
as velocity increases the wavelengths and physical scalar geometries of Planck quanta decreases

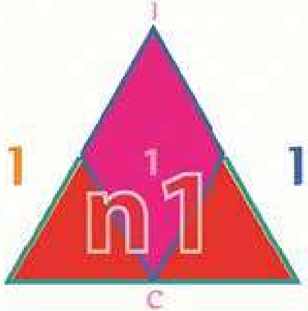

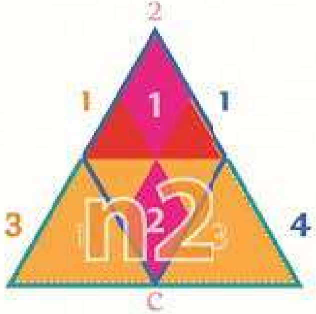

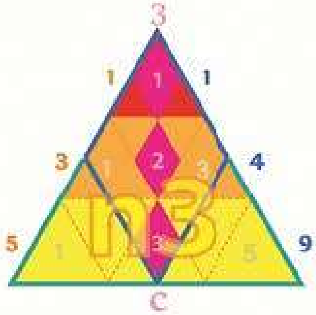
3
1
1
1
3
4
2
5
9
n3
c

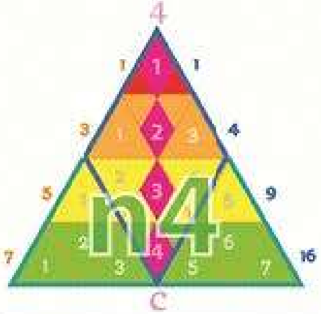
4
1
1
1
3
2
4
5
3
9
7
16
n4
c

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

Equilateral QAM contractions

$$t' = \frac{t}{\sqrt{1-\frac{v^2}{c^2}}}$$

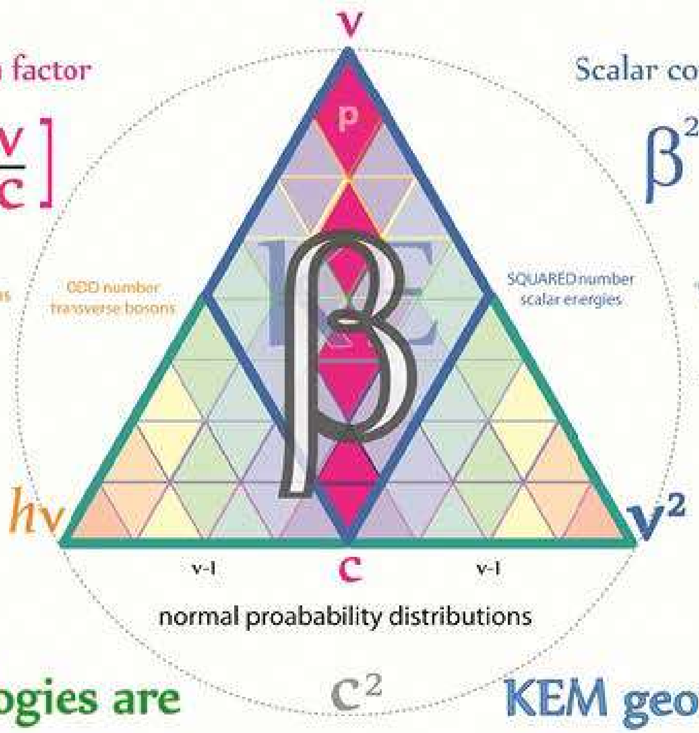
v
p
β
ODD number
transverse bosons
SQUARED number
scalar energies
hv
v²
v-1
c
v-1
normal proabability distributions
c²

Scalar correction factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

'WAVE-length' contractions

$$L = L'\sqrt{1-\frac{v^2}{c^2}}$$

Matter topologies are Lorentz invariant

KEM geometries are relativistic fields

The poor definitions of EM mass & Matter has led to the incorrect application of Lorentz corrections to Matter

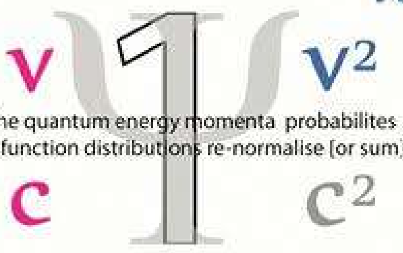

all the quantum energy momenta probabilites of wavefunction distributions re-normalise [or sum] to

Lorentz contractions apply to [K]EM waveforms only

c is the limiting velocity for the electrical acceleration of Matter

$$\sqrt{1-\beta^2}$$

Incorrectly applied to Matter topologies in motion.
The Lorentz factor or Lorentz term appears in several equations in special relativity.

$$\sqrt{1-\left(\frac{v}{c}\right)^2}$$

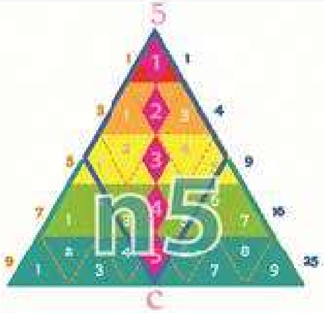

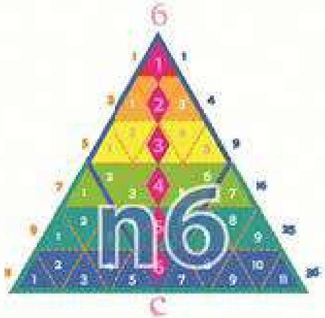

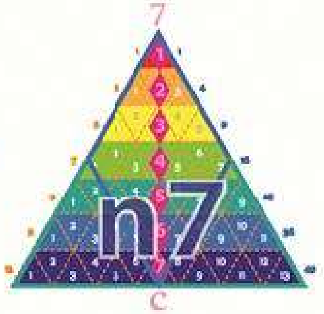

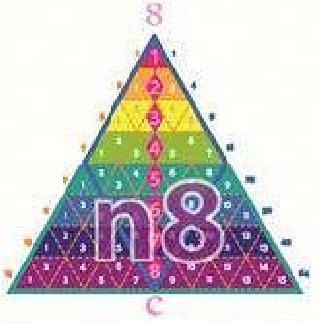
8
1
2
3
4
5
6
n8
c

Copyright ABRAHAM [2008] - all rights reserved

The Energy of a Magnet

It is known that Maxwell's electrodynamics—as usually understood at the present time—when applied to moving bodies, leads to asymmetries which do not appear to be inherent in the phenomena. Take, for example, the reciprocal electrodynamic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motionof the conductor and the magnet, whereas the customary view draws a sharp distinction between the two cases in which either the one or the other of these bodies is in motion. For if the magnet is in motion and the conductor at rest, there arises in theneighborhood of the magnet an electric field with a certain definite energy, producing a current at the places where parts of the conductor are situated. But if the magnet is stationary and the conductor in motion, no electric field arises in the neighborhood of the magnet.

In the conductor, however, we fied an electromotive force, to which in itself there is no corresponding energy, but which gives rise—assuming equality of relative motion in the two cases discussed—to electric currents of the same path and intensity as those produced by the electric forces in the former case. [On the electrodynamics of moving bodies, A. Einstein, 1905]

V

Energy can be removed by placing a conductive loop between
any Positive and Negative quanta the EM field thus creating
a potential difference & an electro-motive force

ε_0

The neutralised [K]EM fields are divergent
from a permanent magnet and create
a Magneto-static EM field with
neutral Electro-static energies

Investigation of the source of the emf produced
by magnets in conductive loops placed around them
led to the development of relativity theory

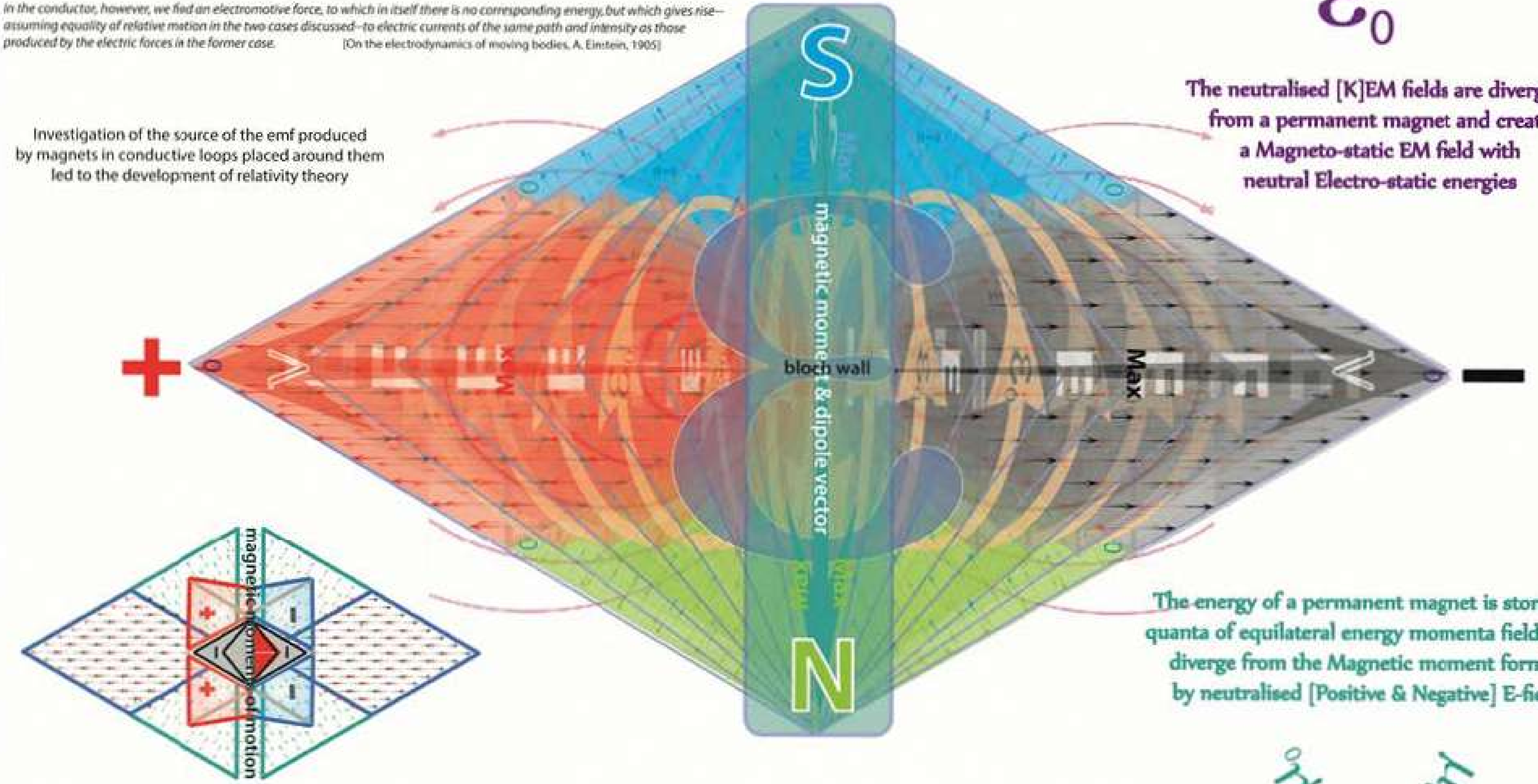

The energy of a permanent magnet is stored in
quanta of equilateral energy momenta fields that
diverge from the Magnetic moment formed
by neutralised [Positive & Negative] E-fields

μ_0 μ_0

Magnetic fields are produced by moving electric charges and the intrinsic magnetic moments of elementary particles associated with a fundamental quantum property, their spin. In special relativity, electric and magnetic fields are two interrelated aspects of a single object, caled the electromagnetic field tensor; the aspect of the electromagnetic field that is seen as a magnetic field is dependent on the reference frame of the observer. In QED, the electro-magnetic field is quantized and electromagnetic interactions result from the exchange of photons.

Ω

Classically, the energy of a permanent magnet
'circulates in endless loops' from North to South
in 3 dimensions around the magnetic moment

Copyright ABRAHAM [2008] - all rights reserved

$$E = h\nu$$

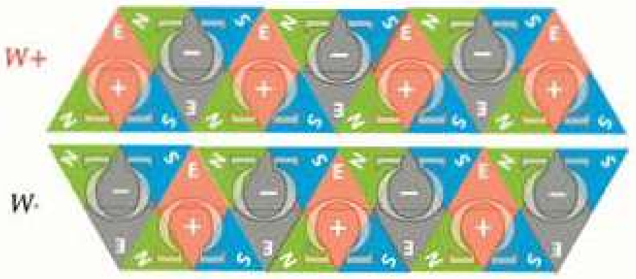

$$B = -\frac{d\phi_E}{dt}$$

A changing Electric field produces a changing Magnetic field

The is no such entity as a purely Electric field [all fields are Electro-Magnetic in their geometric foundations]

Any two separated charge bosons contain charged energy momenta that once connected via a conductive circuit will produce an emf

BOSONS are transverse EM waves [charge carriers]

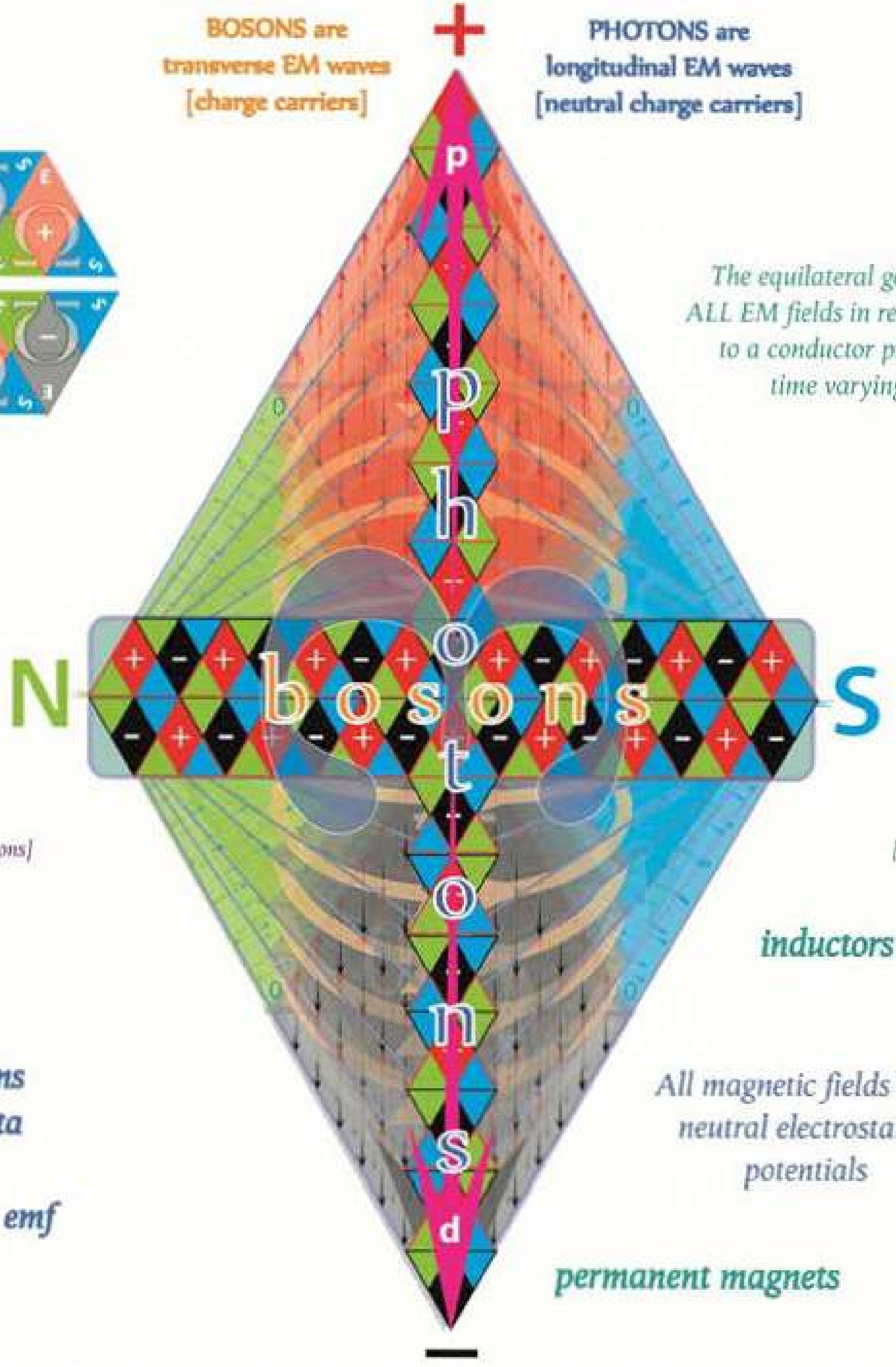

PHOTONS are longitudinal EM waves [neutral charge carriers]

$$E = hf$$

The equilateral geometry of ALL EM fields in relative motion to a conductor produces a time varying emf

$$\mathcal{E} = -\frac{d\phi_B}{dt}$$

A changing Magnetic field produces a changing Electric field

The is no such entity as a purely Magnetic field [all fields are Electro-Magnetic in their geometric foundations]

inductors

All magnetic fields store neutral electrostatic potentials

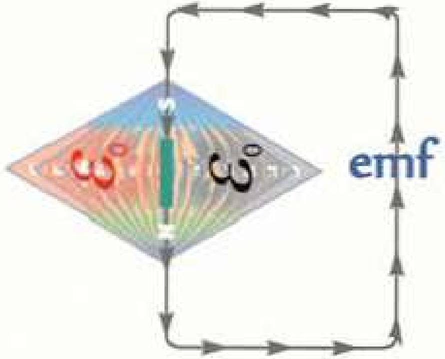

permanent magnets

Copyright ABRAHAM [2008] - all rights reserved

The moving magnet - conductor problem

$$B = -\frac{d\phi_{\varepsilon}}{dt}$$

$$\mathcal{E} = -\frac{d\phi_{B}}{dt}$$

All permanent magnets are electromagnetic fields comprised of transverse bosons and longitudinal photons

We see two observers both moving, approaching each other.

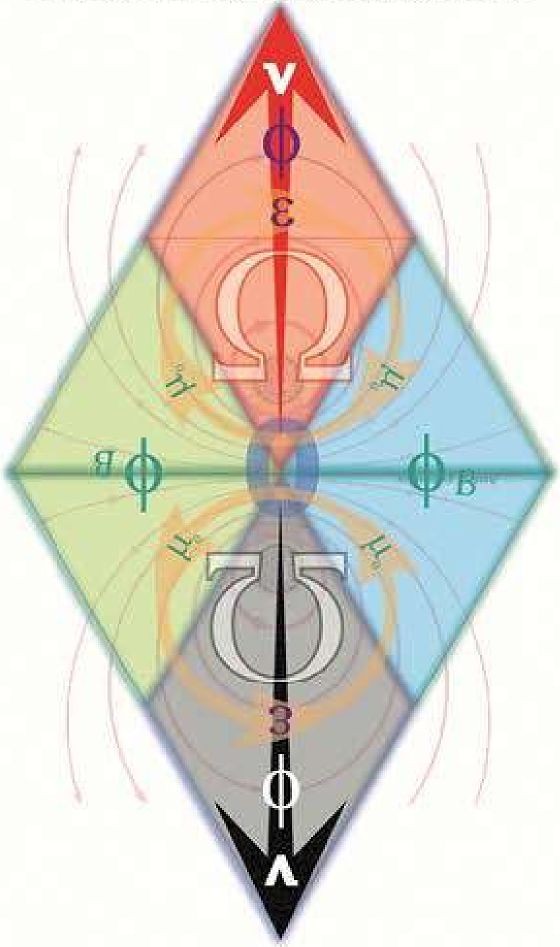

Neutral photon field geometries produce electromotive forces as a result of the charged bosons that make them

Magnet Frame of Reference
(Moving Conductor)

Observer 1 says:

The conductive loop is stationary and the magnet is moving toward it. The electrons in the loop are stationary and have no magnetic moments There is a magnetic field, but it can't produce any force on the electrons since they are stationary within the loop.

Instead, the magnetic field is changing, growing stronger as the magnet gets closer, and this changing magnetic field produces an electric field which causes forces on the electrons, and drives them around the loop producing the current in the galvanometer.

Conductor Frame of Reference
(Moving Magnet)

Observer 2 says:

The magnet is stationary and the loop is moving toward it. The electrons in the loop, since they are moving with the loop, generate their own magnetic moments and experience Lorentz forces as a result of the external magnetic field [F = q v X B], which drives them around the loop and produces the current in the galvanometer.

There is no electric field.

Charged boson field geometries contain equilateral energy momenta with transverse magnetic dipoles

v
p
KE
S
M
N
hv
v²

Special relativity postulates that electric fields in one reference frame become magnetic fields in another frame

Tetryonic theory shows geometrically that the Electric and Magnetic fields are discrete, invariant geometries resulting from the equilateral 'inductive' mass-energy [Planck quanta] geometries that comprise all EM fields.

It is the relative motion of these geometries wrt a conductor that creates changing E&M field flux strengths in turn producing an emf.
When there is no relative motion in either the EM field or the conductor, no emf is produced

The Electromotive Force

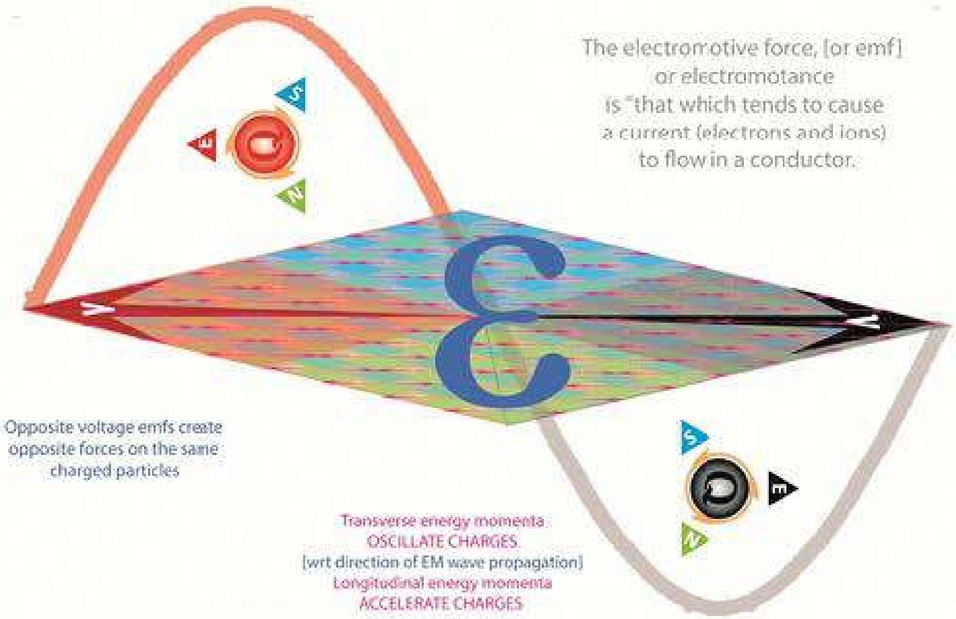

The electromotive force, [or emf]
or electromotance
is "that which tends to cause
a current (electrons and ions)
to flow in a conductor.

Opposite voltage emfs create
opposite forces on the same
charged particles

Transverse energy momenta
OSCILLATE CHARGES
[wrt direction of EM wave propagation]
Longitudinal energy momenta
ACCELERATE CHARGES

Opposite charged particles
experience opposing forces
due to the emf

*"A source of emf can be thought of as a kind of charge pump
that acts to move a charge from a point of low potential
through its interior to a point of high potential. ...
The emf of the source is defined as the work done per charge
dq: = dW/dq."*

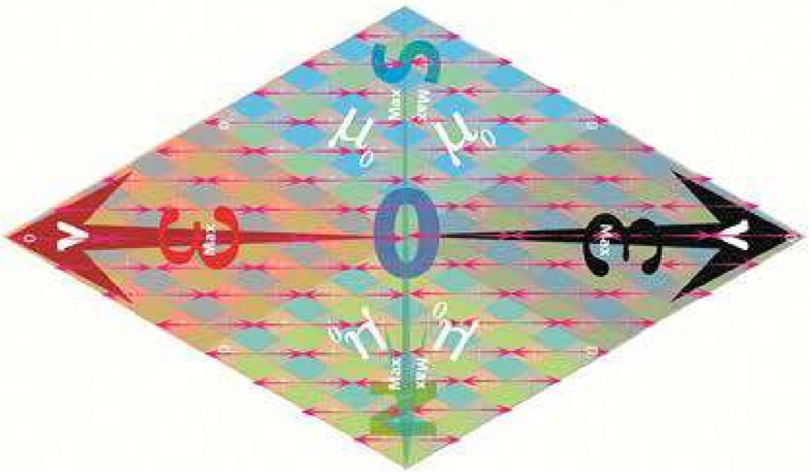

THe magneto-static fields of all Magnets are
comprised of electromagnetic fields of discrete
Planck energy momenta quanta, which
in turn are capable of accelerating
charged particles

All ElectroMagnetic fields have
distinct equilateral electric and
magnetic field geometries that produce
velocity related sinusodal waveforms

$$B = -\frac{d\phi_\varepsilon}{dt}$$

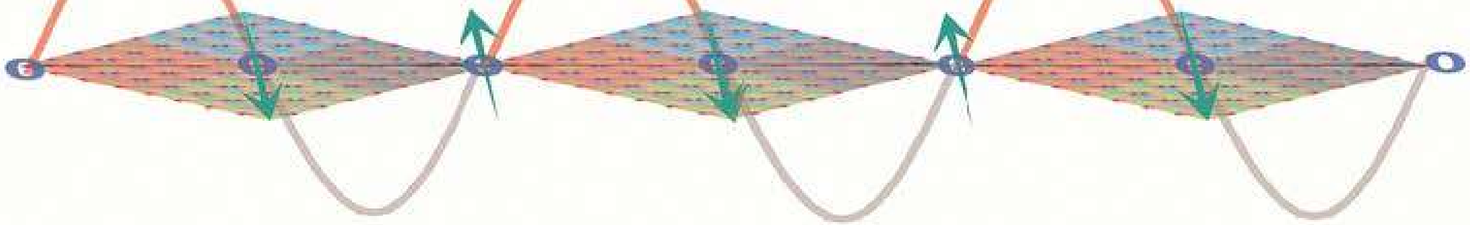

$$\varepsilon = -\frac{d\phi_B}{dt}$$

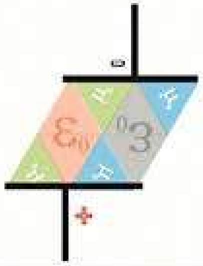

**Charged boson field geometries
can be viewed as transverse
quantum capacitors**

Lenz's law:
***"The emf induced in an electric circuit always acts
in such a direction that the current it drives around
the circuit opposes the change in magnetic flux
which produces the emf."***

**Neutral photon field geometries
can be viewed as longitudinal
quantum batteries**

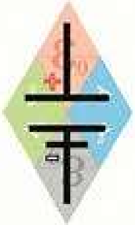

Tetryonics 40.06 - The Electromotive Force

Copyright ABRAHAM [2008] - all rights reserved

The Moving Magnet problem

An electromotive Force results from
the motion of a conductor
relative to a magnetic field

A Magnet is really an electrostatic
store of neutralised emf potential differences
[creating an orthagona Magnetic moment]

The moving magnet and conductor problem
is a famous 19th century thought experiment,
and provides the intersection for classical
and relativistic electromagnetism.

The moving magnet and conductor problem,
along with Michelson-Morley experiment,
formed the basis of Einstein's theory of
special relativity.

Relativistic theory explanation

*In the frame of a
conductor moving
relative to the magnet,
the conductor experiences
a force due to an electric field.*

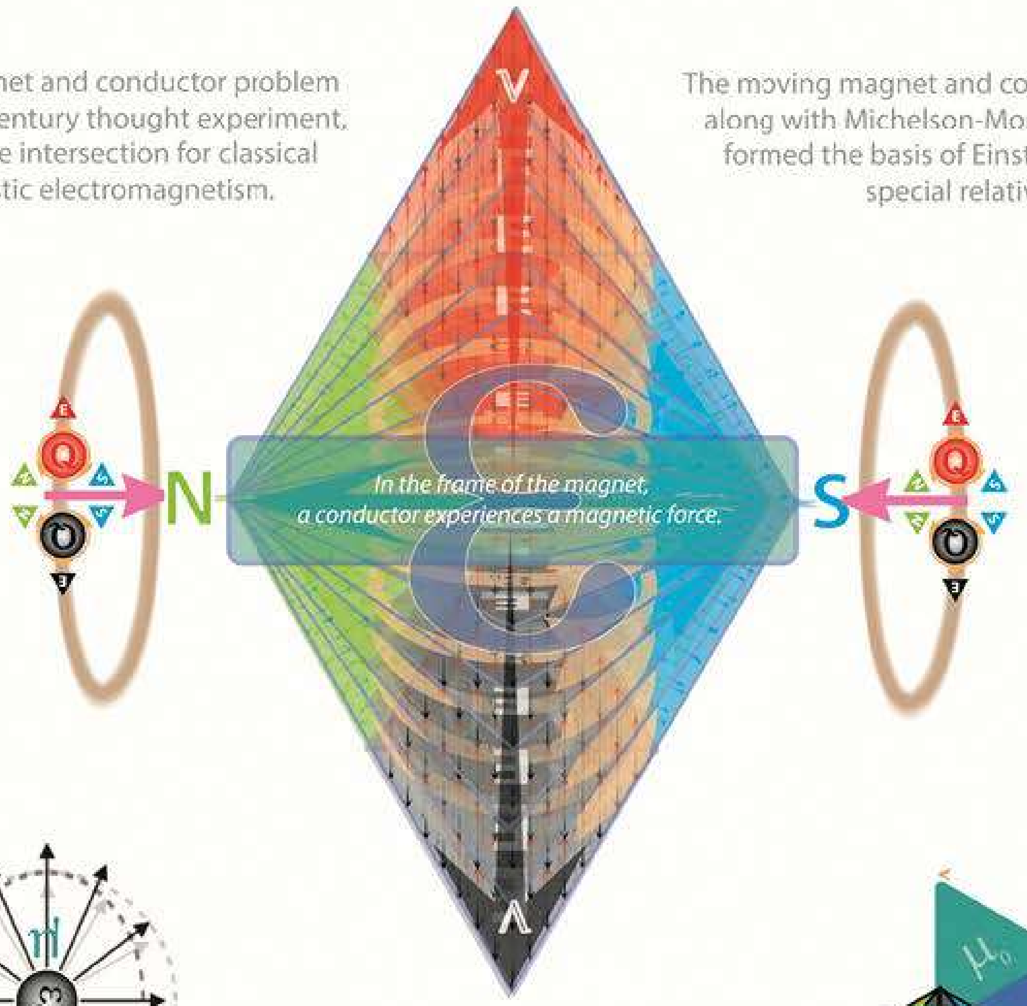

*In the frame of the magnet,
a conductor experiences a magnetic force.*

Tetryonic theory explanation

**The conductor
moves through continuously
changing electric and magnetic
field geometries experiencing
sinuodal E/M forces that are
directly related to each
other at all times**

**A Magnetic field is
an Electric field viewed
in a differing inertial frame**

**Electric and Magnetic fields
are discrete geometrically
related energies**

SR requires that all observers
in inertial frames arrive at the
same form for Maxwell's equations

SR seeks to modify space and time
in a manner such that the forces
and fields transform consistently.

producing expressions that
differ from Newton's law of
motion by a factor of γ

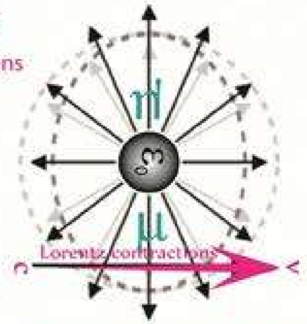

the relativistic distortion of
accelerating spherical point charges
creates magnetic moments

Electrostatic particles have no magnetic moment

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$

Electric force *Magnetic force*

*Once in motion it posseses a magnetic moment and kinetic energies
and is subject to Lorentz forces when it moves throuhg an external EM field*

magnetic moments are
the result of secondary KEM
fields of Matter in motion

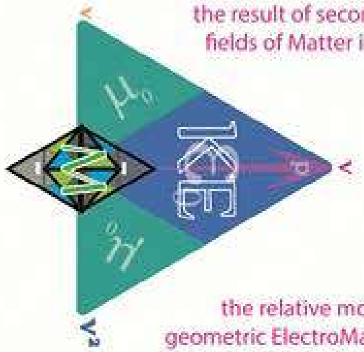

Lorentz corrections
only all to KEM field
mass-energies

the relative motion of
geometric ElectroMagnetic fields
creates changing EM forces

rest mass-Matter

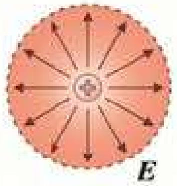

v=0

Einstein's Error of perception

In its rest frame the electric field of a point charge has the same strength in all directions and diverges away from the charge.

Electric fields are distorted due to relativistic speed effects to create Magnetic fields
The faster the velocity the greater the Magnetic field
At rest the Magnetic field becomes an Electric field

What led me more or less directly to the special theory of relativity was the conviction that the electromotive force acting on a body in motion in a magnetic field was nothing else but an electric field

Albert Einstein 1953

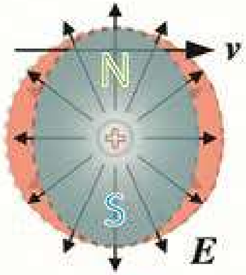

v=.6c

The SR theoretical model of a spherical charged body being relativisitically distorted as the the source of the Magnetic moments for charges in motion is incorrect

Lorentz velocity dependent factors relate only to Kinetic EM mass-energies of motion

M + It is the KEM field mass-energies that produce Magnetic moments in relativistic particles due to the motion of ElectroStatic Matter (12 + loop inductive geometries) = RE

Stationary Leptons are 12pi charged standing waves [Matter] topologies with NO secondary Kinetic EM mass-energies geometries

Matter is Lorentz invariant

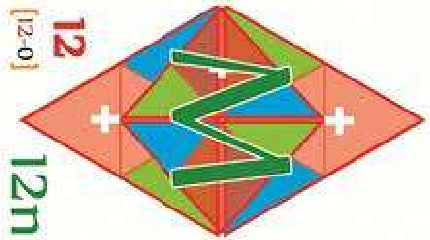

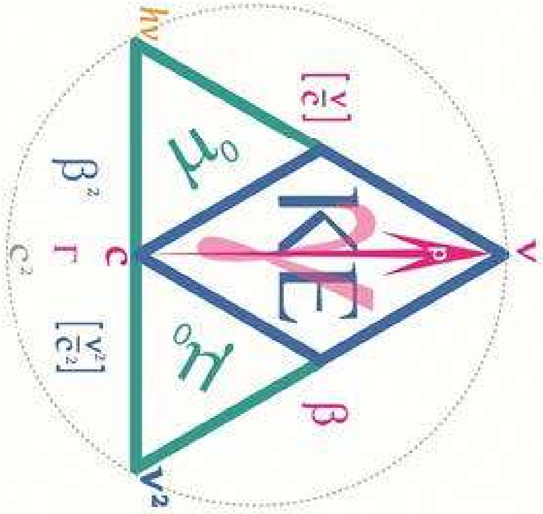

Moving Leptons are 12pi charged standing waves [Matter] topologies with Kinetic EM mass-energies geometries

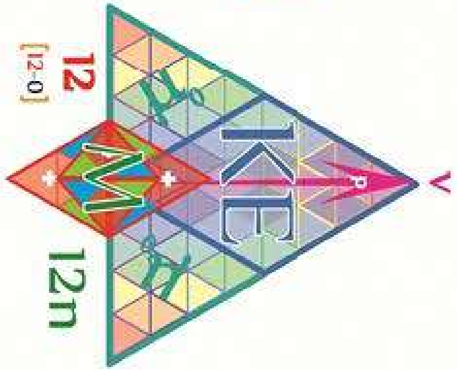

Stationary charges have Electric Fields and neutralised Magnetic fields

The speed of light in a vacuum [c] is the limit for electrically accelerated particles
Tetryonic geometry reveals it is NOT the limit of acheivable velocities

Moving charges have Electric fields and Magnetic moments

Tetryonics 40.08 - Einstein's Error (of perception)

Copyright ABRAHAM [2008] - all rights reserved

ε_0

Tetryonic relativistic motion

The [K]EM mass-energy content of a particle in motion is velocity dependent its rest Matter toplogy & charge geometries are velocity invariant

μ_0 μ_0

In a frame in which the particle is at rest, we can detect only electric field

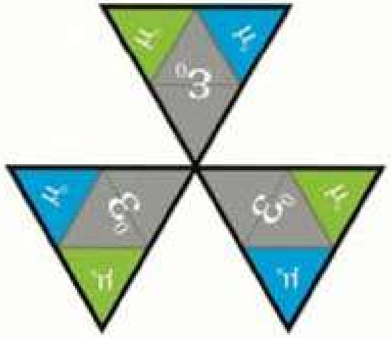

rest Matter

all Magnetic dipoles are neutral dipole pairs

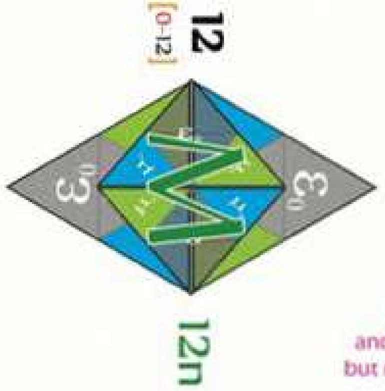

Charge is a velocity invariant measure of the quantised angular momenta iof any spatial co-ordinate system

The total EM mass-energy momenta of any particle in motion is velocity dependent and is specific to particular spatial co-ordinate systems

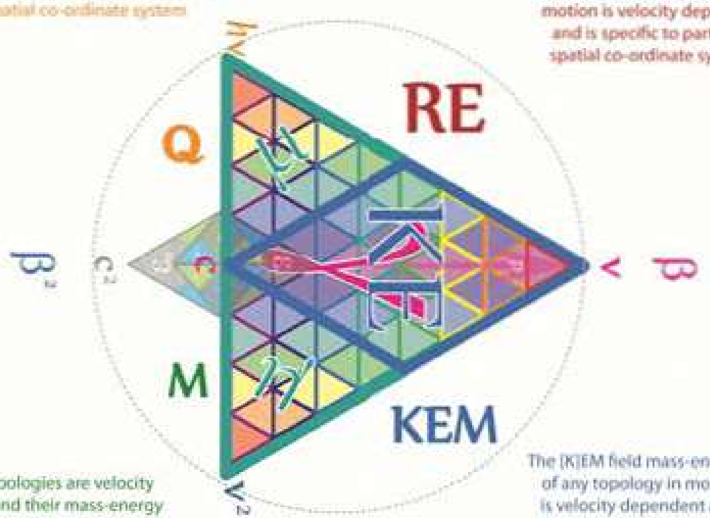

Matter topologies are velocity invariant and their mass-energy content can only be measured in c^4 spatial co-ordinates

The [K]EM field mass-energies of any topology in motion is velocity dependent and subject to Lorentz correction within in its c^4 spatial co-ordinates

It is well demonstrated that the fields of charged particles appear different in different inertial frames

In a frame in which the particle is moving, we also see a magnetic field.

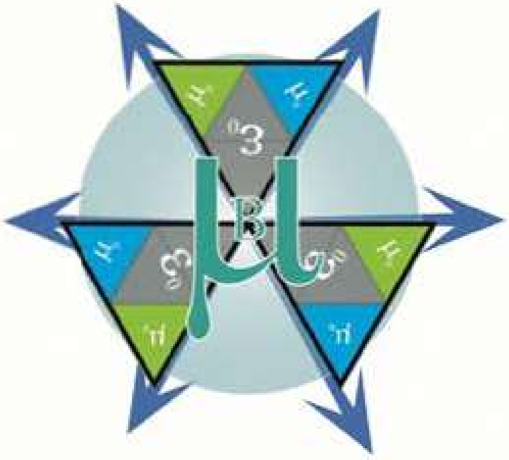

rest Matter+KEM

relativistic KEM fields produce a secondary magnetic moment

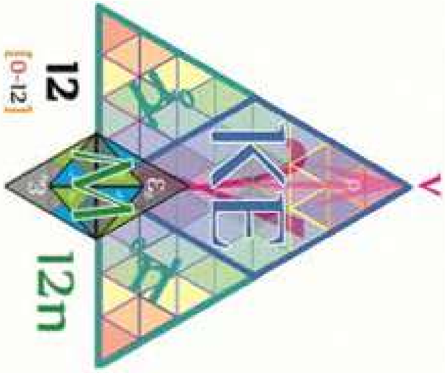

$$Mv^2 = KEM = hv^2$$

Tetryonics has revealed that the Kinetic Electro-Magnetic [KEM] field of a particle in motion has a distinct geometric identity of its own rather than being the relativistic distortion of a charged EM particle

Virtual particles and action-at-a-distance are negatived by Tetryonic KEM field mass-energy momenta and the problem of simultanity is avoided because a second particle responds not directly to the first particle, but rather to the first particle's extended, velocity related [K]EM field generated by its motion at its own position.

Copyright ABRAHAM [2008] - all rights reserved

Quantum level **KEM field**

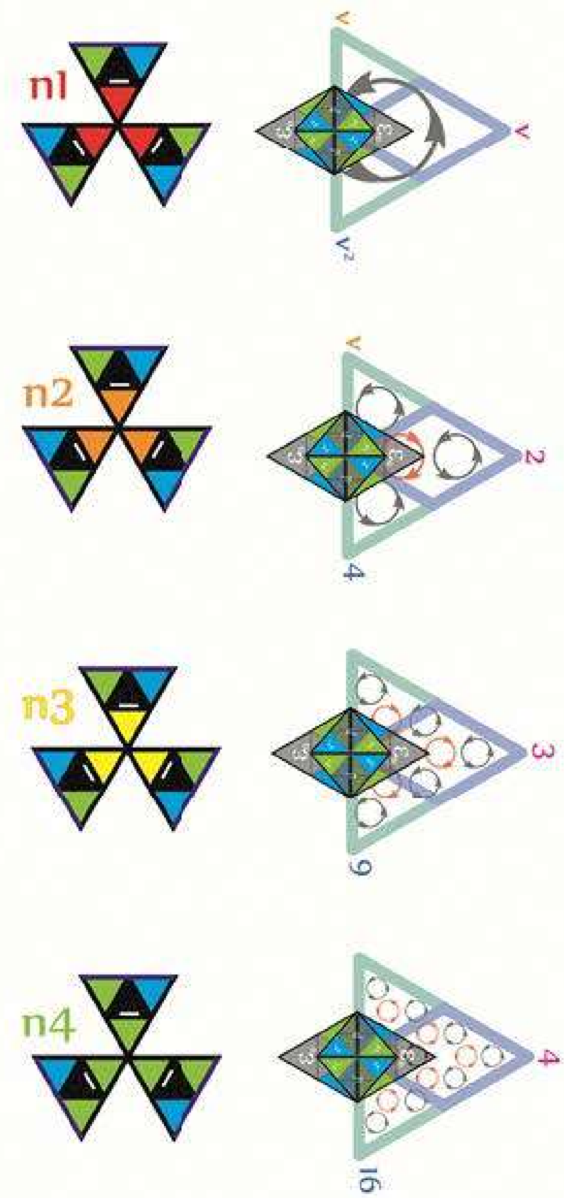

James Clerk Maxwell

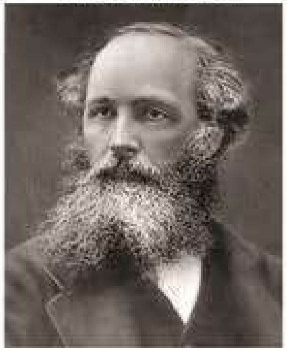

(13 June 1831 – 5 November 1879)

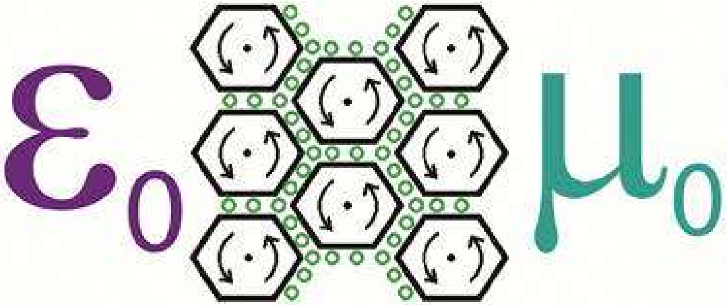

All quantum charge arrangements of EM mass-energy and forcefields can be viewed as an arrangement of 'quantum idler wheels

'As early as 1857 Maxwell began to develop the idea of orienting molecular vortices along magnetic field lines, culminating in the publication of his paper 'On physical lines of force'…

He posited a honeycomb of vortices in which each vortex cell was separated from its neighbour by a layer of spherical particles, revolving in the opposite direction to the vortices.

These 'idler wheel' particles communicated the rotatory velocity of the vortices from one part of the field to another. In this ether model, the most famous image in nineteenth-century physics, the analogy provides mechanical correlates for electromagnetic quantities. The angular velocity of the vortices corresponds to the magnetic field intensity, and the translational flow of the idle wheel particles to the flow of an electric current; the field equations are based on the rotation of molecular vortices in the ether.

He emphasized that while the theory was mechanically conceivable, the model itself was provisional and temporary, even awkward, hardly 'a model of connexion existing in nature'

KEM field **Quantum level**

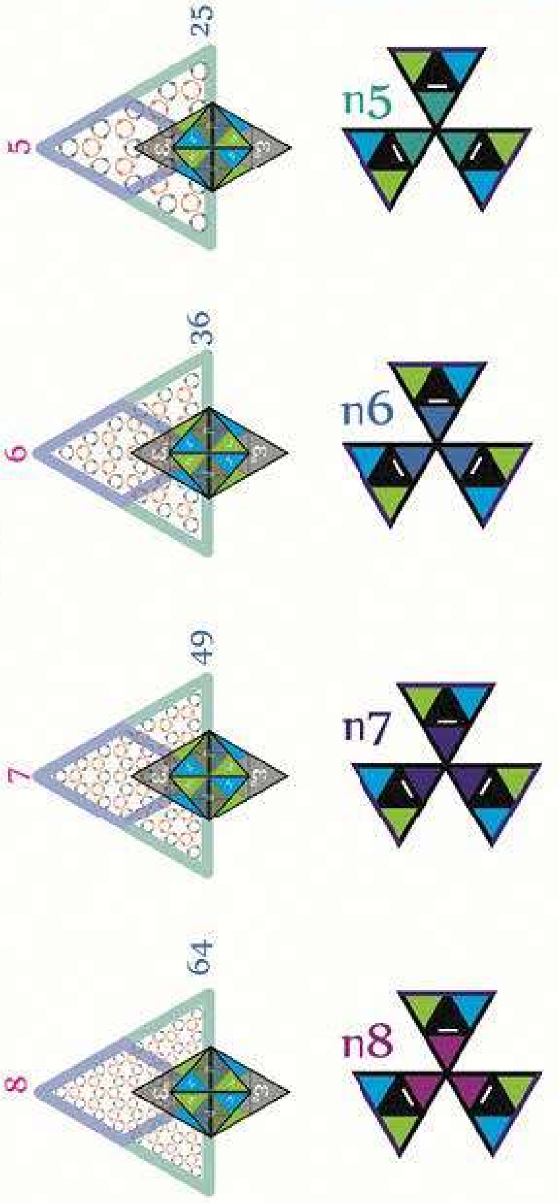

Equating Equilateral geometries to Quantised Angular Momentum reveals a classical mechanical model closely resembling Maxwell's hypothesised idler wheels that be be applied to any EM field of force

Virtual Particles

In physics, a virtual particle is a particle that exists for a limited time and space, it has become a commonplace mechanism in current Physical theories to provide a basis for the Force interactions between particles

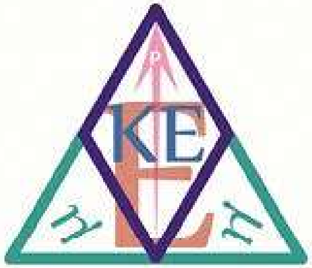

The energy and momentum of a virtual particle are uncertain according to the uncertainty principle.

$$m^2v^4 = E^2 = p^4$$

The degree of uncertainty of each is inversely proportional to time duration (for energy) or to position span (for momentum).

Original Feymar diagram illustrating the exchange of a quantum between 2 electrons

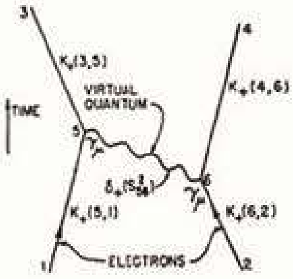

$$mv^2 = E = pv$$

Bohr understood that if you are going to try to be mechanical, you have to show some convincing mechanics.

If you can't show some convincing mechanics, you might as well dodge all mechanics from the beginning, staying with the math.

He had learned this from Maxwell, who had done the same thing 60 years earlier.

In the 1860's, Maxwell had tried to create vortices to explain the field mechanics, but, finding himself under heavy fire from Lord Kelvin and others, he decided to give it up and go to other mathematical alternatives like quaternions instead.

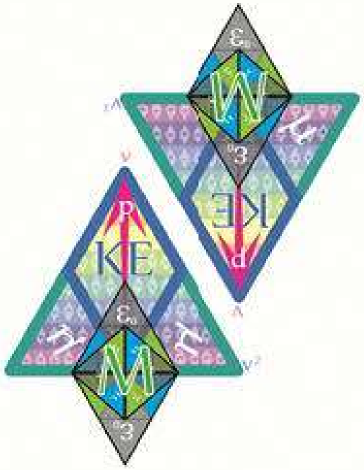

$m^2c^4 = E^2 - p^2c^2$ is in-correct for superpositioned 2D KEM fields the real interaction geometry is illustrated below:

$$\frac{m\Omega^2 v^4}{2} = KE^2 = \frac{p^2v^2}{2}$$

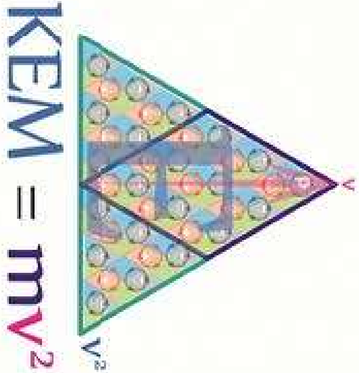

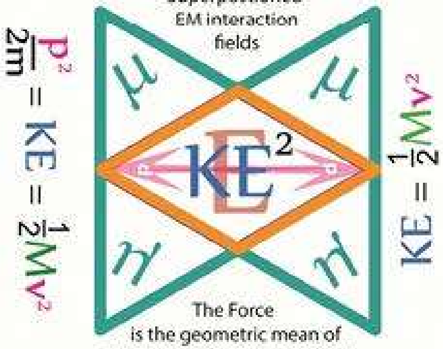

Superpostioned EM interaction fields

The Force is the geometric mean of the superpositioned [KE] fields

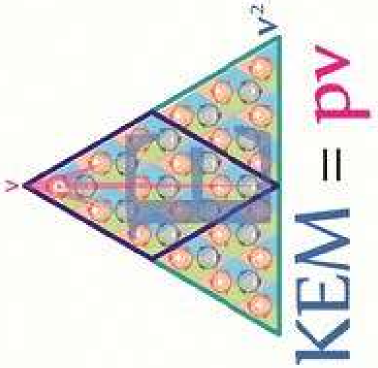

The equilateral energy geometry of Tetryonics clears up this issue once and for all, virutal particles do NOT exist. The Force interactions between all Matter in relative motion is mediated by the equilateral geometries of [K]EM mass-ENERGY momenta that spread out from their source Matter through Space-time

Copyright ABRAHAM [2008] - all rights reserved

Equilateral geometry - the Hidden Constant

$$^{kg}m\,\Omega\,\frac{m^2}{s} = h \;\; kg\frac{m^2}{s}$$

Quantised Angular Momenta has an equilateral geometry

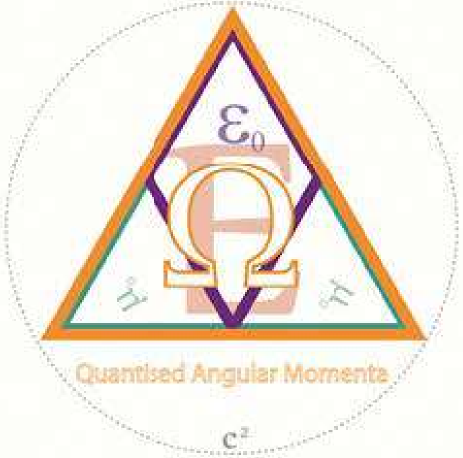
ε_0
Ω
μ_0
μ_0
Quantised Angular Momenta
c^2

Tetryonics has revealed the equilateral relationships between Charge, Planck energies and the Constants showing quantised angular momentum to be the 'hidden' geometric constant driving the physics of our Universe on all scales

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Charge · EM Field · ElectroMagnetic · Planck quanta · mass · velocity

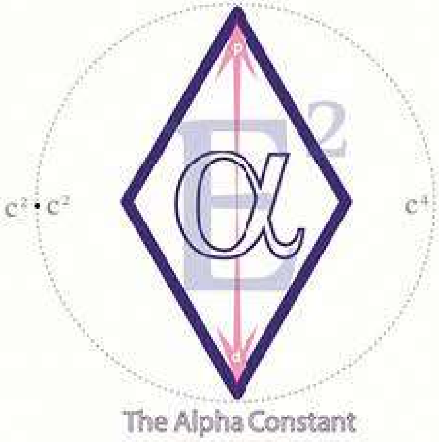
E^2
α
p
d
$c^2 \cdot c^2$
c^4
The Alpha Constant

Alpha is the geometric coupling constant between Magnetic and Electrical energies that acts as a scaling constant between Kinetic and Potential quantum energies

It is the source of the quantum charges of all EM mass-energy momentum & inertia

Classically modelled as a 'rotational' vector Quantised Angular Momentum is in fact an Equilateral geometry

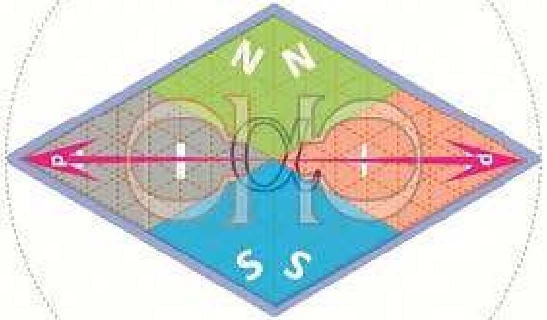
N N
S S
d
p
α
c^2

Photons are dual equilateral geometries creating a neutral charge with two opposed coupling constants

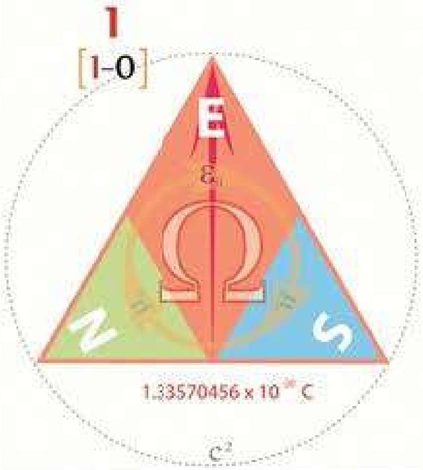
1
[1-0]
E
ε_0
Ω
N
S
μ_0
μ_0
1.33570456 x 10[illegible] C
c^2

Quantised Angular Momenta creates charged EM mass-energies

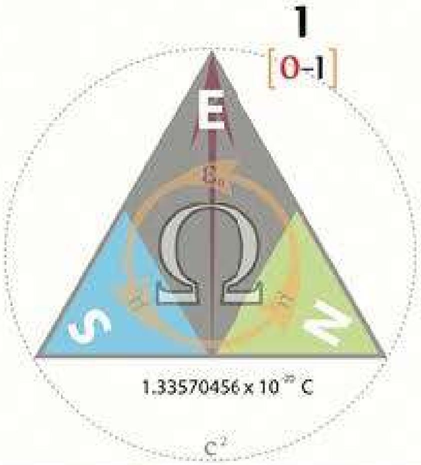
1
[0-1]
E
ε_0
Ω
S
N
μ_0
μ_0
1.33570456 x 10[illegible] C
c^2

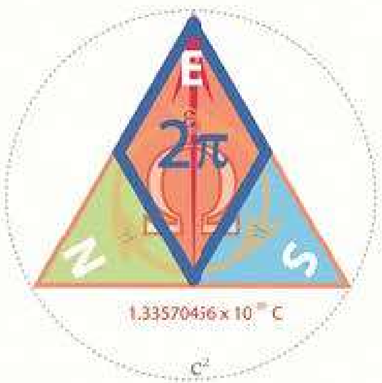

.0012

2π

.007539822

$$\alpha = \frac{e^2}{\hbar c}$$

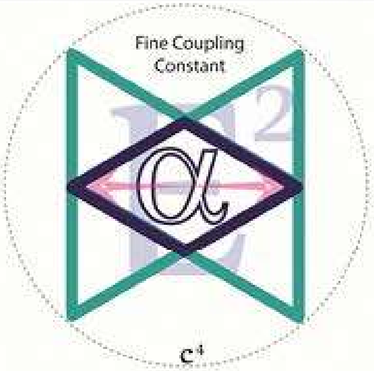

$$\alpha = \frac{e^2}{\hbar c\, 4\pi\varepsilon_0}$$

$7.297\,352\,5376(50) \times 10^{-3}$

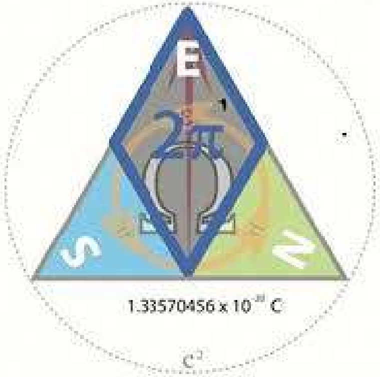

132.6291192^{-1}

$$= \frac{1}{137.035\,999\,679}$$

The ALPHA Constant is the scalar coupling field of quantised energy momenta in superpositioned EM fields that facilitates Force (linear energy momenta) exchanges, resulting in the familiar Laws of Attraction/Interaction

In physics, the fine-structure constant is a fundamental physical constant, namely the coupling constant characterizing the strength of the electromagnetic interaction.

The numerical value of α is the same in all systems of units, because α is a dimensionless quantity.

There is a most profound and beautiful question associated with the observed coupling constant,
e the amplitude for a real electron to emit or absorb a real photon.
It is a simple number that has been experimentally determined to be close to 0.08542455.

It has been a mystery ever since it was discovered more than fifty years ago,
and all good theoretical physicists put this number up on their wall and worry about it.

Immediately you would like to know where this number for a coupling comes from:
is it related to pi or perhaps to the base of natural logarithms? Nobody knows.
It's one of the greatest damn mysteries of physics: a magic number that comes to us with no understanding by man.

We know what kind of a dance to do experimentally to measure this number very accurately,
but we don't know what kind of dance to do on the computer to make this number come out, without putting it in secretly

Richard Feynman

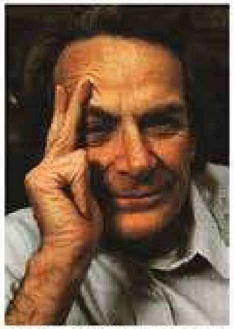

(May 11, 1918 - February 15, 1988)

Superpositioned Fields

equilateral energies

.0012 m²/s

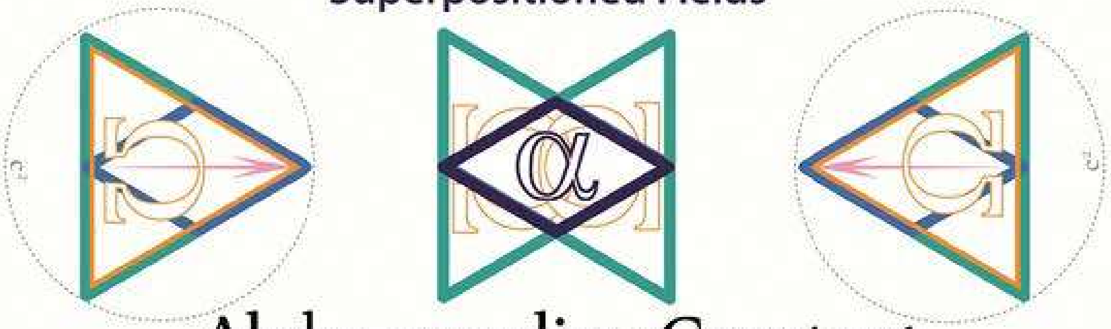

0.007539822

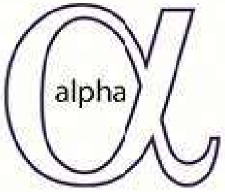

132.6291192^{-1}

Alpha coupling Constant

You might say the "Hand of GOD" wrote that number, but "we don't know how HE pushed his pencil."

The Fine Structure Constant

$2\pi[QAM]_{.0012} = 0.007539822$

Arnold Sommerfeld introduced the fine-structure constant in 1916 as a fundamental physical constant, namely the coupling constant characterizing the strength of the electromagnetic interaction.

$2\pi[QAM]^{-1}_{.00119997} = 132.6291192$

In quantum electrodynamics, α is the coupling constant determining the strength of the interaction between electrons and their exchange photons

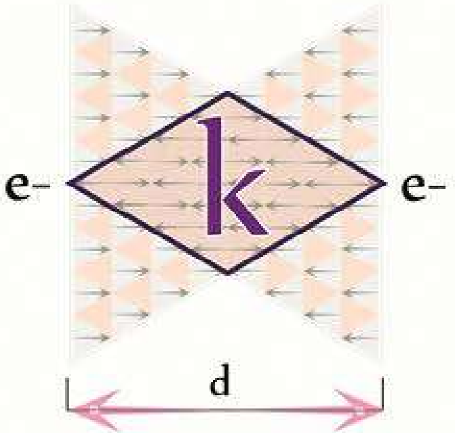

Similar Charges REPEL

10^{-6}

weak force

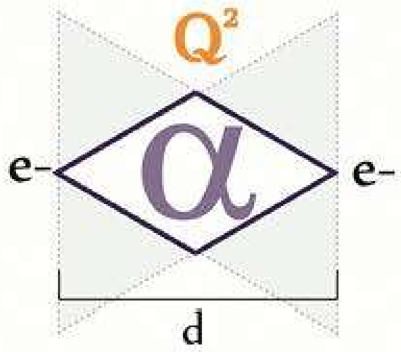

It is a dimensionless quantity that has a constant numerical value in all systems of units that is the direct result of its equilateral geometry.

1

strong force

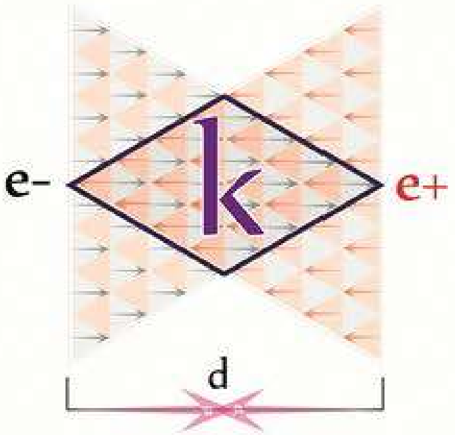

Opposite Charges ATTRACT

$$\alpha = .007539822$$

It can now be revealed to be a measure of the forces produced by, or contained in, the geometry of two superpositioned KEM $[E^2]$ field energies

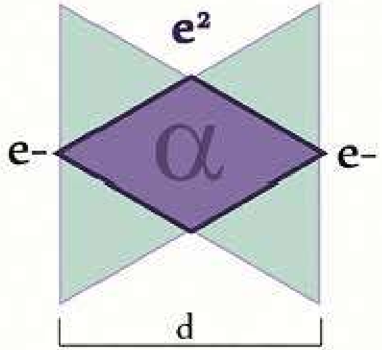

the energy needed to overcome the electrostatic repulsion between two electrons separated by a distance of d

10^{-39}

gravity

The alpha constant applies equally to the coupling of Electric and Gravitational superpositioned fields

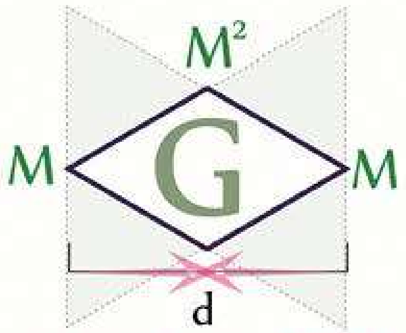

Gravitational Matter is always ATTRACTIVE

$7.5e^{-3}$

EM forces

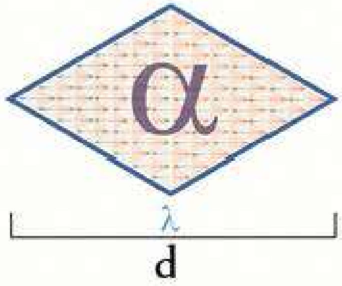

the energy of a single photon of wavelength λ that provides the energy momenta required to create a force between particles

Tetryonics 40.14 - The Fine Structure Constant

Coupling Constants

$$k\frac{Q_1Q_2}{r^2} = F$$

Coulomb's [k], the Fine structure constant [α] and Newton's [G] are all reflections of the same coupling constant geometries differing only by the strength of their respective superpositioned energy field densities

$$F = -G\frac{M_1M_2}{r^2}$$

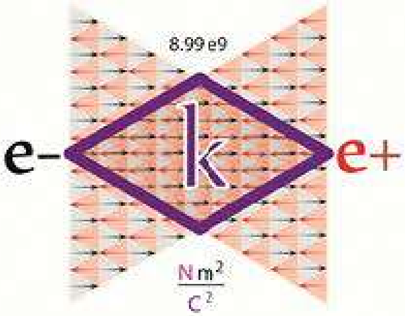
8.99 e9
e-
k
e+
$\frac{Nm^2}{C^2}$

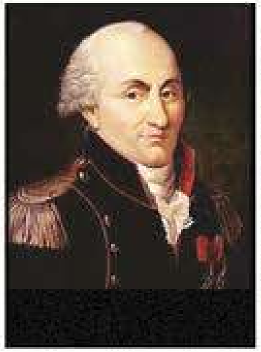

Coulomb's constant

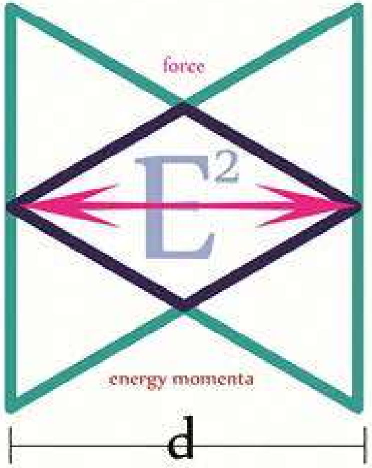
force
E^2
energy momenta
d

Newton's constant

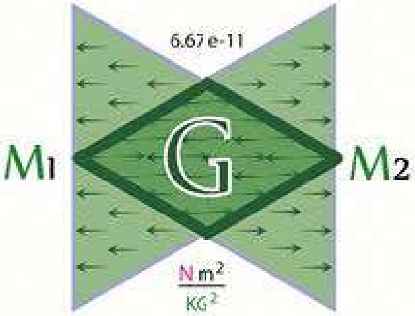
6.67 e-11
M_1
G
M_2
$\frac{Nm^2}{KG^2}$

$$\beta = \left[\frac{v}{c}\right]$$

Linear force coupling constant

In physics, a coupling constant is a number that determines the strength of an interaction for superpositioned Electrical or Gravitational fields between material objects

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Scalar energy coupling constant

Coupling Constants are dimensionless numbers

They are reflective of the linear & scalar forces produced by 2D mass-energies

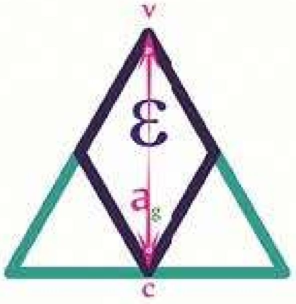

$$\frac{m}{s}\cdot\frac{s}{m}$$

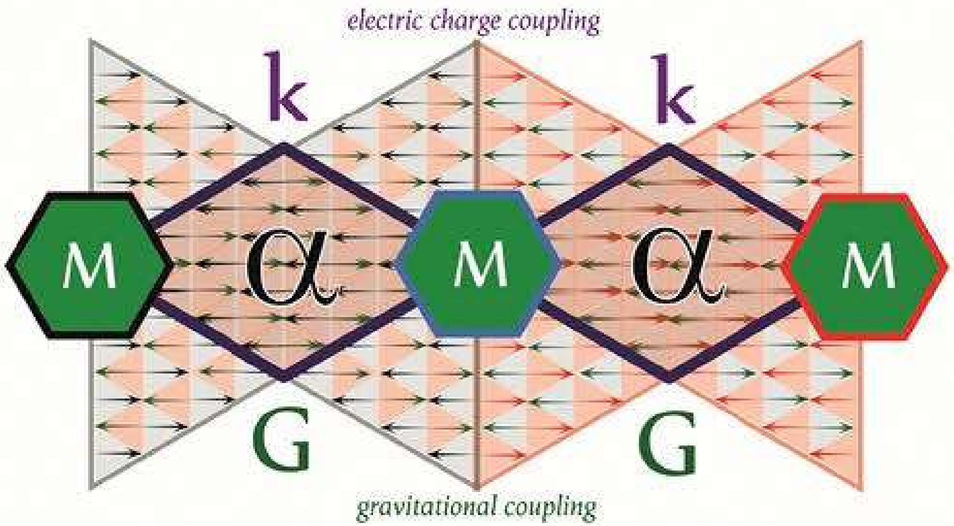
electric charge coupling
k
k
M
α
M
α
M
G
G
gravitational coupling

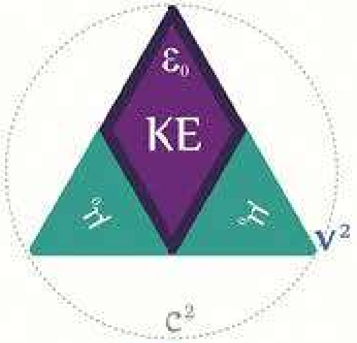
ε_0
KE
v^2
c^2

$$\frac{m^2}{s^2}\cdot\frac{s^2}{m^2}$$

Usually the Lagrangian or the Hamiltonian of a system can be separated into a kinetic part and an interaction part.
[For example, the electric charge of a particle is an Electrical coupling constant]

Copyright ABRAHAM [2008] - all rights reserved

Opposites
ATTRACT

ElectroMagnetic Charge

[the universal coupling constant]

Similars
REPEL

1.33518 e-20 C

Q

1.33518 e-20 C

Electric charge is a physical property of matter that causes it to experience a force when near other electrically charged matter.
Electric charge comes in two types, called positive and negative

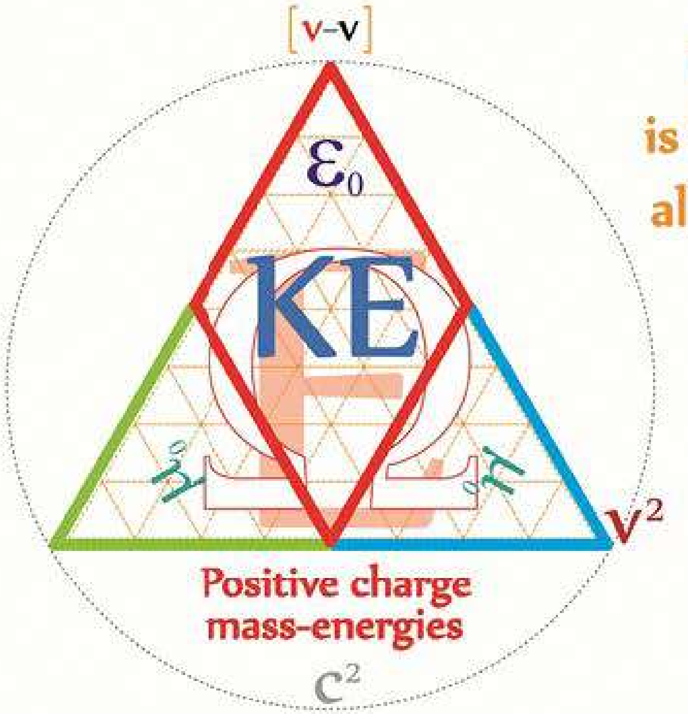

ElectroMagnetic CHARGE
is the equilateral foundation to
all the quantum geometries of
mass-ENERGY-Matter

$$\frac{m\,\Omega\,v^2}{c^2}$$

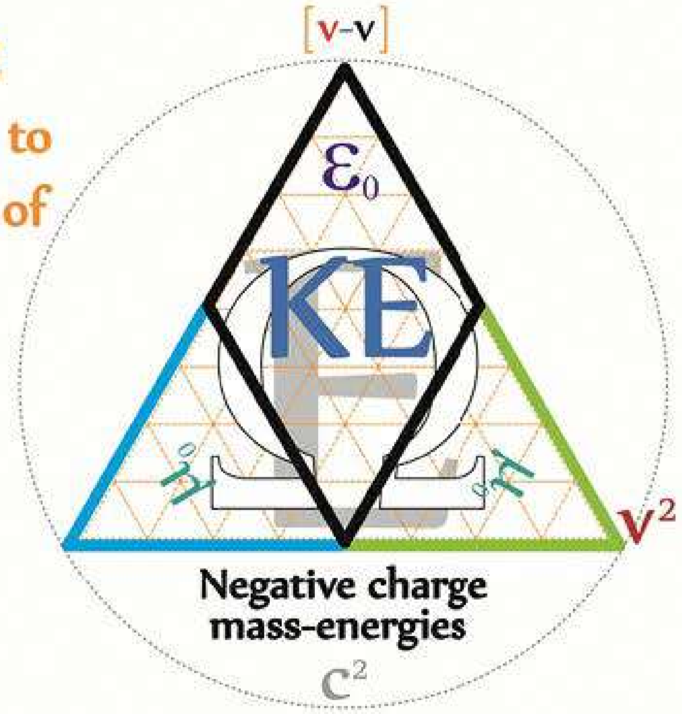

seconds

Changing charge geometries are equivalent to TIME in relativity

seconds

The electric charge is a fundamental conserved property of all subatomic particles, which determines their electromagnetic interaction.
The interaction between a moving charge and an electromagnetic field is the source of the electromagnetic force, which is one of the four fundamental forces

The equilateral geometry of Planck energies creates CPT symmetries in physics, and provides the foundation for all Matter geometries and their fields of interaction

Tetryonics 40.16 - Charge the universal Coupling Constant

Copyright ABRAHAM [2008] - all rights reserved

Einstein–Podolsky–Rosen paradox

was a thought experiment that attempted to challenge the Copenhagen interpretation of Quantum physics

Tetryonics provides a complete model of all Energy forms and Wave-Particle interactions, allowing a clear understanding of previously mysterious actions and processes in Quantum Mechanics

Showing that it is possible to know the both Position and Momenta of Particles and to model EM wave geometries and interactions in Quantum Physics

Boris Yakovlevich Podolsky

(1896–1966)

Δx

Albert Einstein

(14 March 1879 – 18 April 1955)

Δp

Nathan Rosen

(March 22, 1909 – December 18, 1995)

The EPR experiment involved two systems that initially interact with each other and are then allowed to separate before being measured.

Measuring the Position of the Electron determines the Position of the Positron

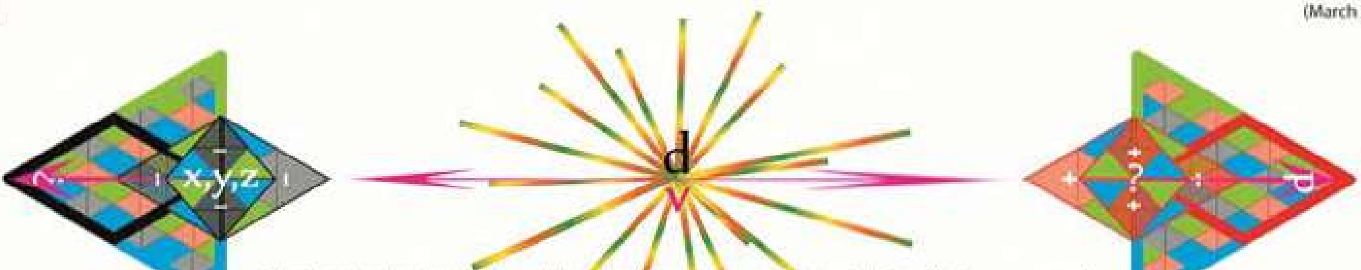

Measuring the Momentum of the Positron determines the Momentum of the Electron

Quantum entanglement occurs when particles such as photons, electrons, etc interact physically and then become separated; this type of interaction is such that each resulting member of a pair is properly described by the same quantum mechanical state, which is indefinite in terms of important factors such as position, momentum, spin, polarization, etc

The EPR paper shows that measuring one feature of a entangled system, e.g., the momentum of one of the pair of particles, will reveal the same feature of the other particle - thus providing a mechanism for determining both the momentum and position of both particles simultaneously

Thus providing a theoretical indication that either the Uncertainty Principle was incorrect or that our understanding of Quantum Mechanics was incomplete

Copyright ABRAHAM [2008] - all rights reserved

Spooky interactions-at-a-distance & the transfer of information

Quantum entanglement is sasid to occur when particles such as photons, electrons, and other forms of EM mass-energies interact physically and then become separated; the type of interaction is such that each resulting member of a pair is properly described by the same quantum mechanical description (state), which is indefinite in terms of important factors such as position, momentum, spin, polarization, etc.

Through the equilateral geomtry of Tetryonics the mass-energy-Matter of any system of particles can be modelled in order to reveal the true EM mechanics of their quantum interactions

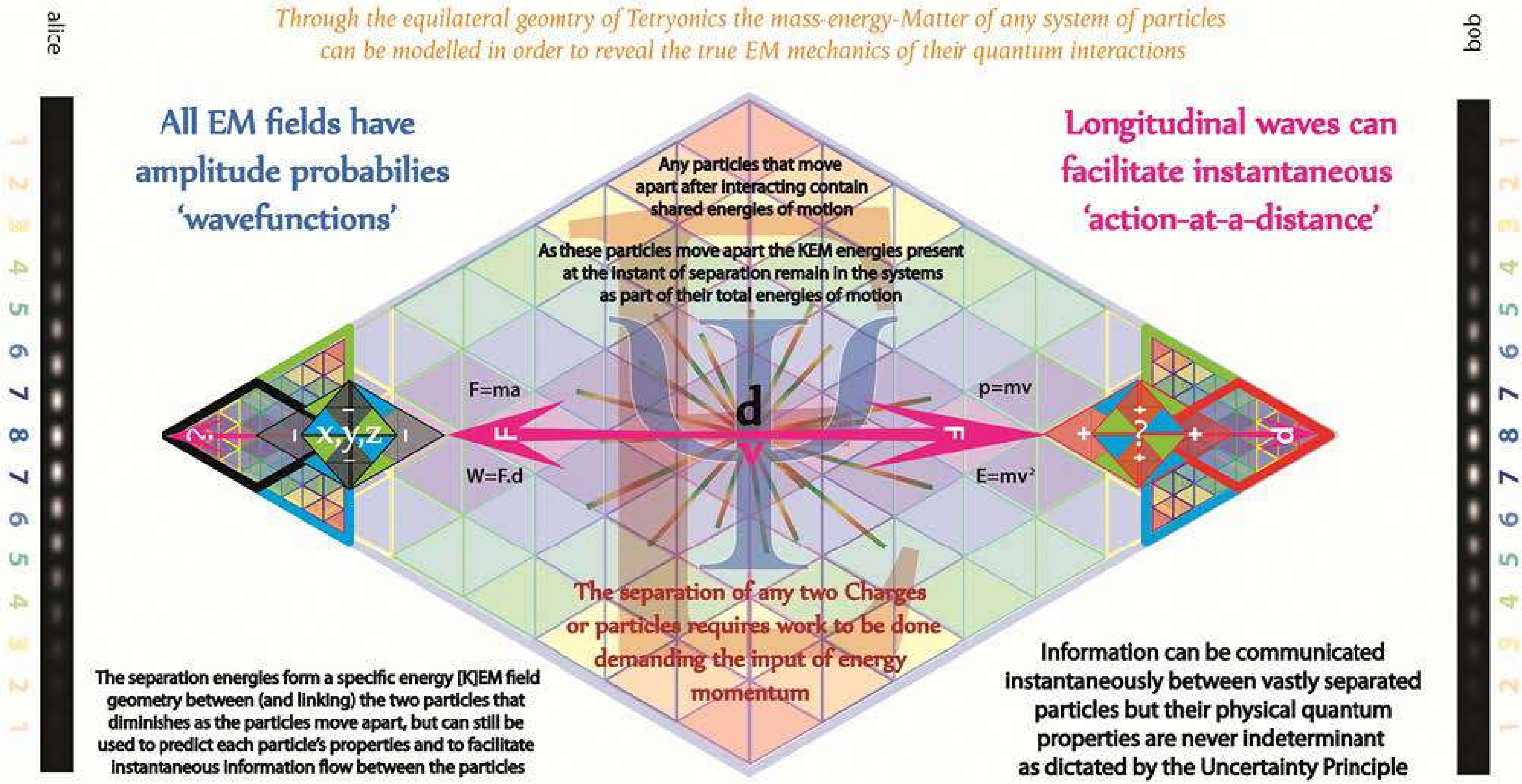

All EM fields have amplitude probabilies ‘wavefunctions’

Longitudinal waves can facilitate instantaneous ‘action-at-a-distance’

Any particles that move apart after interacting contain shared energies of motion

As these particles move apart the KEM energies present at the instant of separation remain in the systems as part of their total energies of motion

The separation of any two Charges or particles requires work to be done demanding the input of energy momentum

The separation energies form a specific energy [K]EM field geometry between (and linking) the two particles that diminishes as the particles move apart, but can still be used to predict each particle’s properties and to facilitate instantaneous information flow between the particles

Information can be communicated instantaneously between vastly separated particles but their physical quantum properties are never indeterminant as dictated by the Uncertainty Principle

When two entangled particles are separated using longitudinal energies, the energy momenta of separation provides a mechanism for the establishment and near-instantaneous communication of the physical wavefunction/properties of one particle to the other, irrespective of the spatial or temporal [space-time] separation of the two ‘entangled’ partner particles

Spectral line differentials

$$\Delta h\nu = \Delta M v = \Delta p$$

changes to energy momenta

accelerate photo-electrons

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg's formula is
a mathematical description
of Tetryonic energy geometry

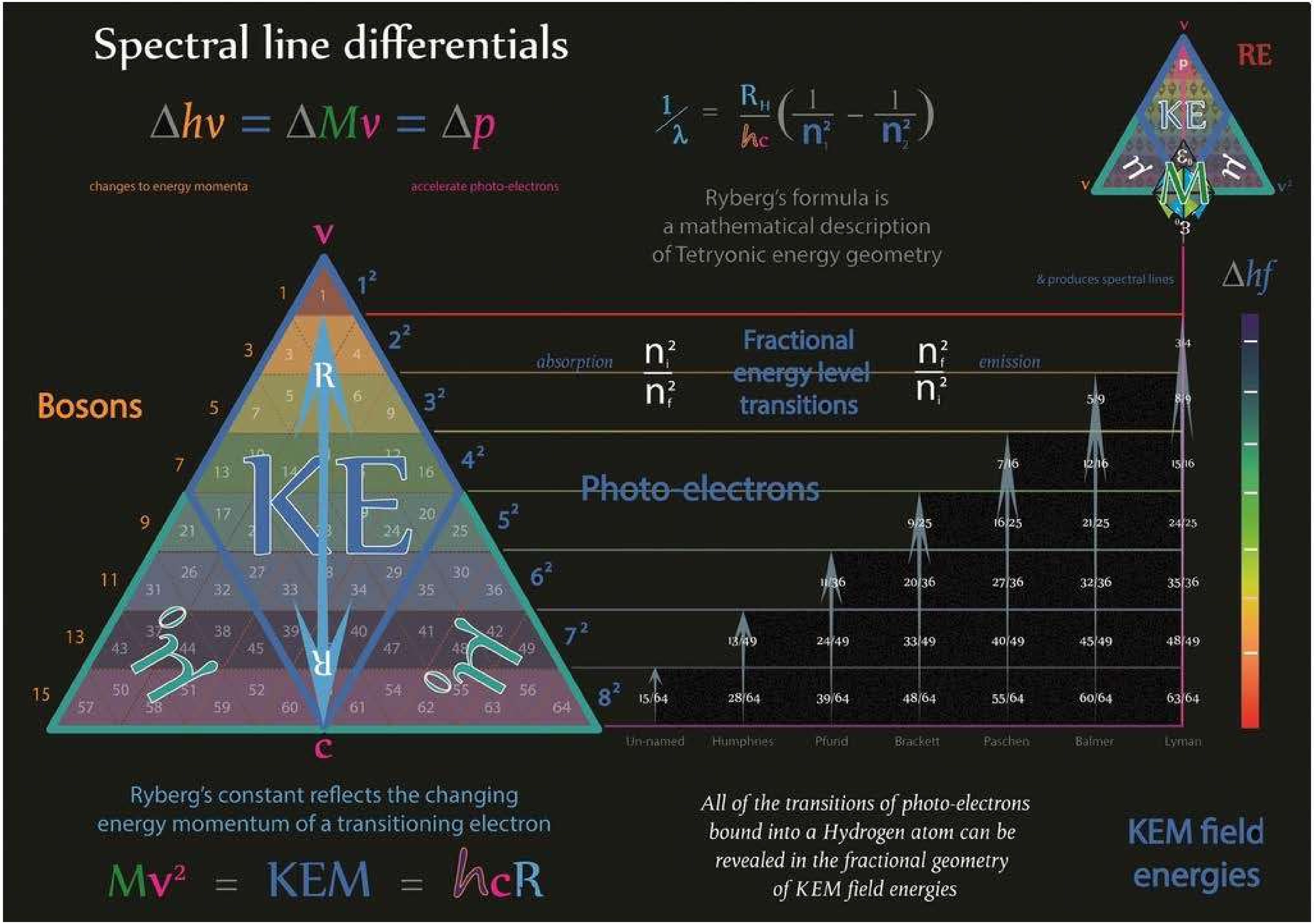
RE
KE
M
P
ε_0
v
v
v^2
3
& produces spectral lines
Δhf
v
c
R
KE
Bosons
1
3
5
7
9
11
13
15
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
absorption
$\frac{n_i^2}{n_f^2}$
Fractional energy level transitions
$\frac{n_f^2}{n_i^2}$
emission
Photo-electrons
3/4
5/9
8/9
7/16
12/16
15/16
9/25
16/25
21/25
24/25
11/36
20/36
27/36
32/36
35/36
13/49
24/49
33/49
40/49
45/49
48/49
15/64
28/64
39/64
48/64
55/64
60/64
63/64
Un-named
Humphries
Pfund
Brackett
Paschen
Balmer
Lyman

Ryberg's constant reflects the changing
energy momentum of a transitioning electron

$$Mv^2 = KEM = hcR$$

*All of the transitions of photo-electrons
bound into a Hydrogen atom can be
revealed in the fractional geometry
of KEM field energies*

KEM field energies

Tetryonics 40.19 - Spectral line differentials

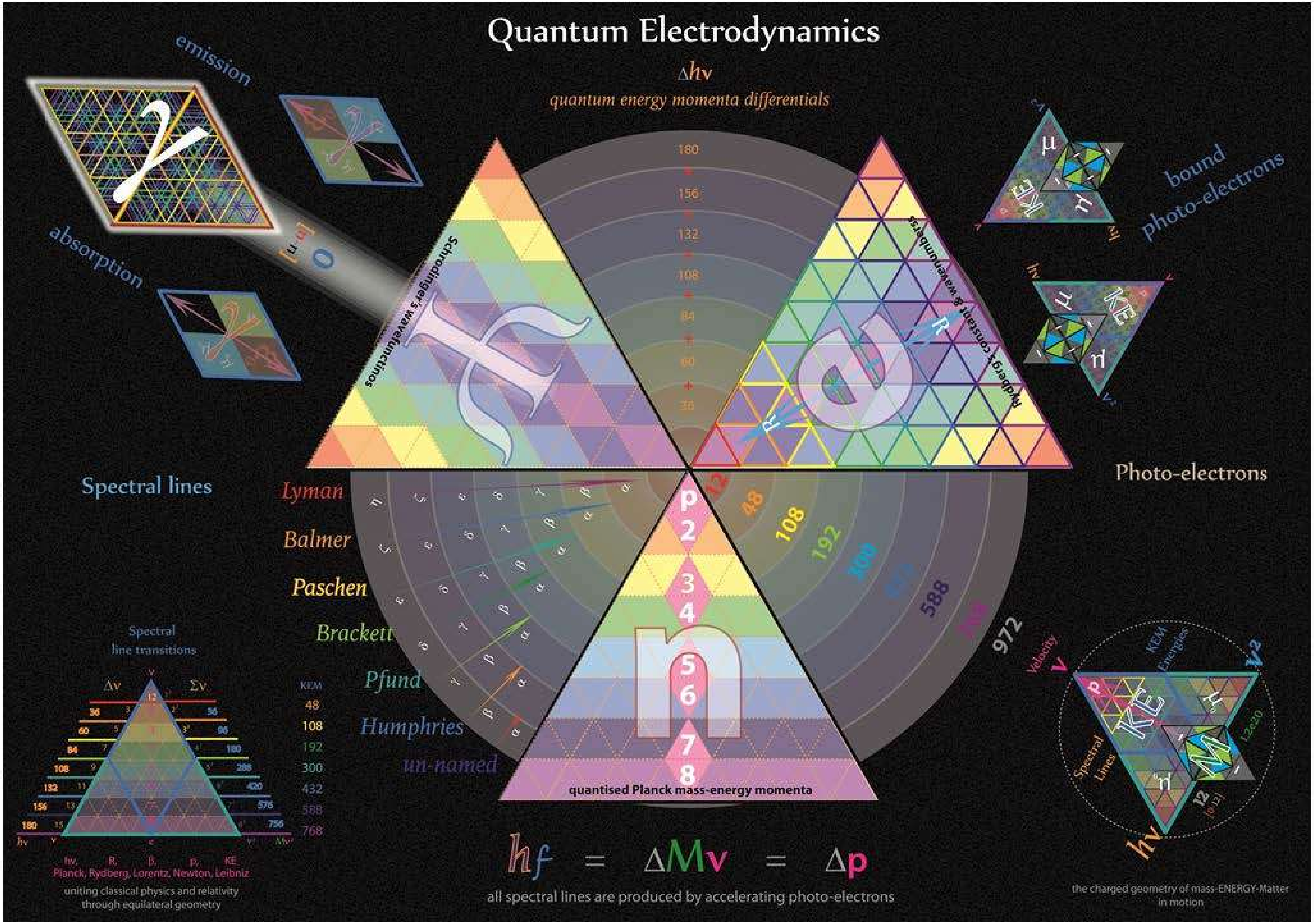
Quantum Electrodynamics
Δhv
quantum energy momenta differentials
emission
absorption
Schrödinger's wavefunctions
h
Rydberg's constant & wavenumbers
R
e
180
156
132
108
84
60
36
bound photo-electrons
μ
KE
hv
v²
Photo-electrons
Spectral lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
p
2
3
4
5
6
7
8
n
quantised Planck mass-energy momenta
12
48
108
192
300
432
588
768
972
Spectral line transitions
Δv
Σv
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
36
96
180
288
420
576
756
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz
uniting classical physics and relativity through equilateral geometry
hf = ΔMv = Δp
all spectral lines are produced by accelerating photo-electrons
Velocity v
KEM Energies
v²
Spectral Lines
M
c²
12
hv
the charged geometry of mass-ENERGY-Matter in motion

Tetryonics 40.20 - Quantum Chemistry follows

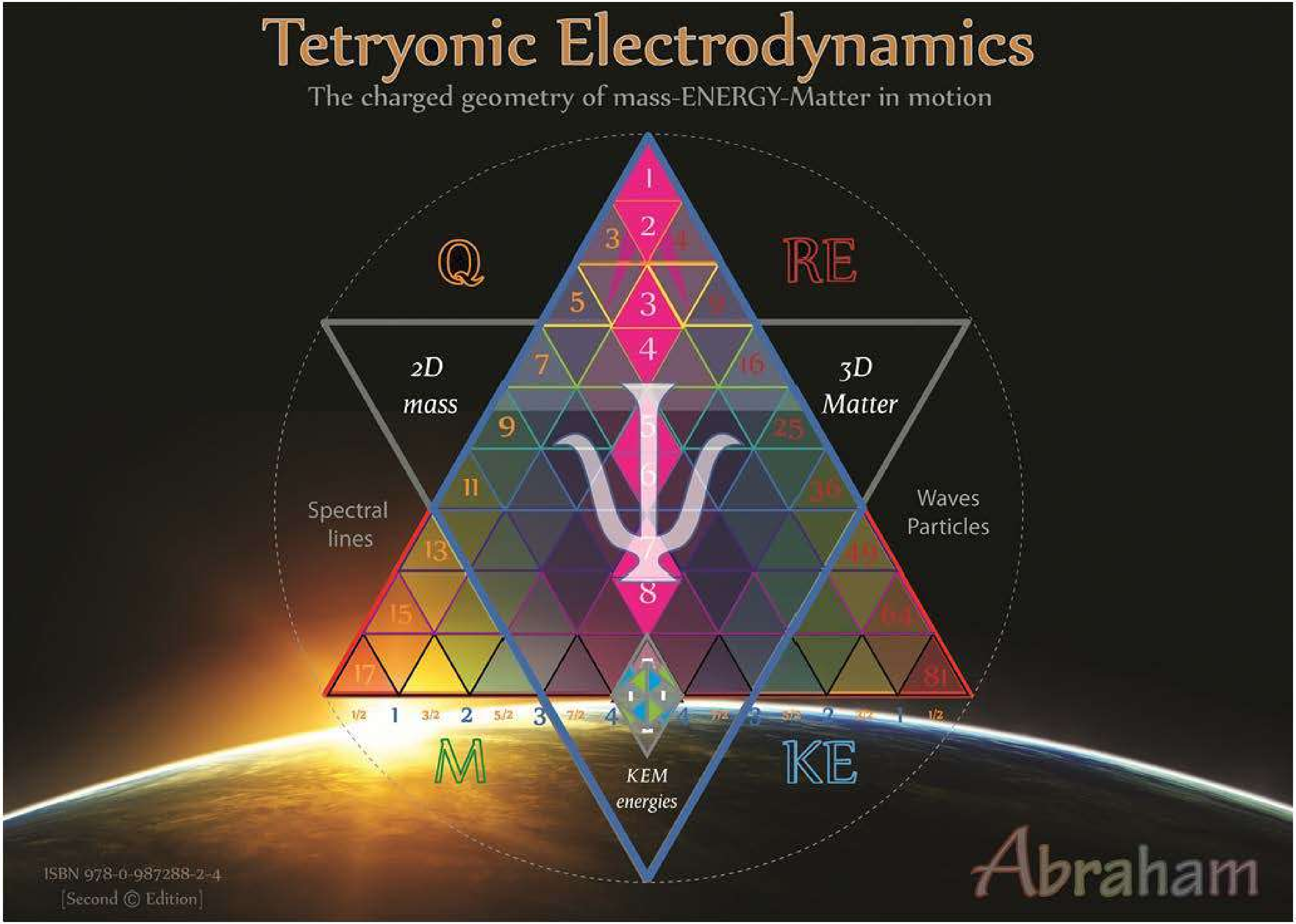

Tetryonic Electrodynamics

The charged geometry of mass-ENERGY-Matter in motion

Q
RE
2D
mass
3D
Matter
Spectral
lines
Waves
Particles
M
KE
KEM
energies
1
2
3
4
3
5
7
9
11
13
15
17
4
9
16
25
36
49
64
81
5
6
7
8
1/2
1
3/2
2
5/2
3
7/2
4
4
7/2
3
5/2
2
3/2
1
1/2
Abraham

ISBN 978-0-987288-2-4
[Second © Edition]

Tetryonics 40.21 - Quantum Electro-Dynamics

Copyright ABRAHAM [2008] - All rights reserved

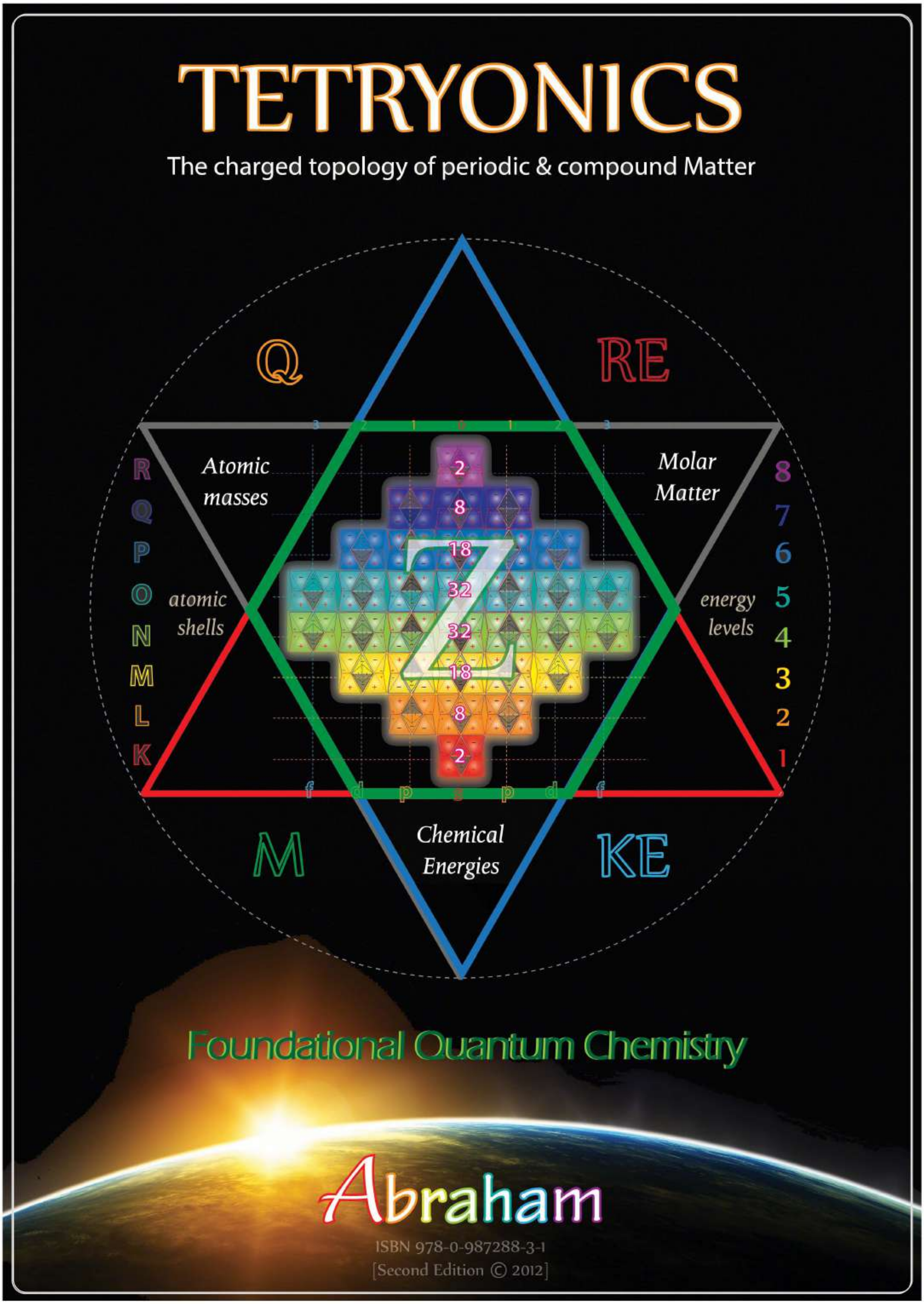

TETRYONICS

The charged topology of periodic & compound Matter

Q
RE
M
KE
Atomic masses
Molar Matter
atomic shells
energy levels
Chemical Energies
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f

Foundational Quantum Chemistry

Abraham

ISBN 978-0-987288-3-1

[Second Edition © 2012]

Protons **+** *Neutrons*

u+d+u d+u+d

All electrons, quarks and Baryons are made up of 4n+ (Tetryonic) standing-wave EM fields.

As well as having nett Tetryonic charged topologies ranging between [+24] ~ [-12] they all posses distinct ELECTRIC FIELDS that are concentrated in 3 apex points as indicated in the illustrations

These points result from the orientation of Electric apexes and orthagonal Magnetic dipole field edges that make up each particle's externalised EM fields.

The Positive and Negative electric apex points, obey the Law of Interaction forcing separated nuceli to combine due to their individual nett Tetryonic charges and provide a means of orienting nuclei to each other to create larger particles [elements, allotropes and compounds]

External Magnetic (H) fields can interact with the integral magnetic (B) dipoles of Tetryonic particles forcing them to orientate in specific directions to facilite chemical bonding [nuclear forces]

Additionally, external Electric fields can interact with the integral electric fields attracting or repelling them depending on the polarity of the external electric field [Electrostatics]

External energies can be induced into these integral EM fields via inductive coupling or the absorption of spectral photons in turn leading to an increase in the strengthes of the integral EM apexes in turn increasing the Strong Nuclear Force.

Residual Electro-Magnetic Forces

allow Neutrons and Protons to attract via the opposite Electric charge points created by their constituent Quarks in order to create Elementary Nuclei

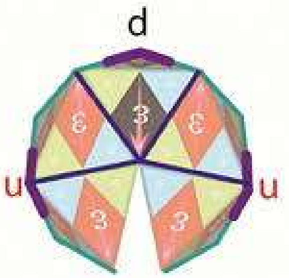

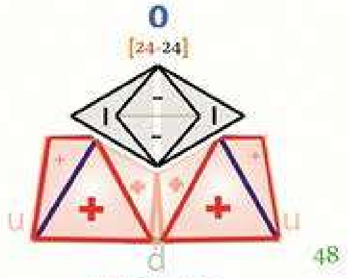

Hydrogen

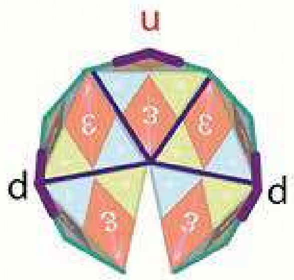

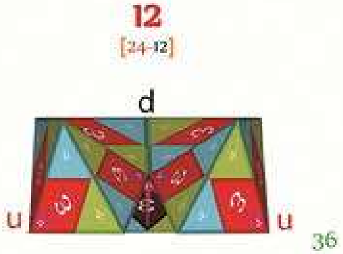

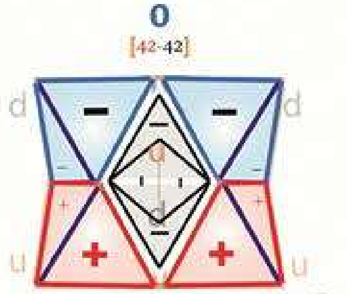

Deuterium

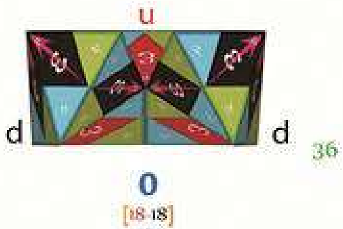

The Strong Nuclear force binds Matter together

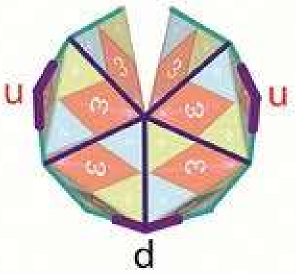

The orientation of the component Electric fields within 3D Matter creates macroscopic force apexes via externalised 'E-points'

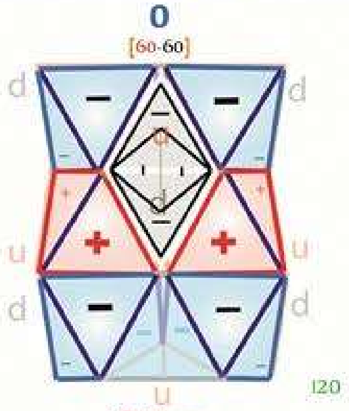

Tritium

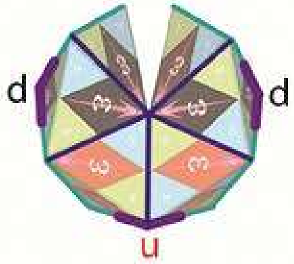

The orientation of the component Magnetic fields within 3D Matter creates macroscopic force apexes via externalised 'M-dipoles'

Copyright ABRAHAM [2008] - All rights reserved

Nucleonic residual EM force

The attraction between Baryonic E&M field apexs, a result of their component Quark arrangements, results in the formation of heavier and more complex Nuclei

The residual e-field apexes and m-field dipoles form two rings of residual EM fields around the circumference of atomic nuclei

Neutrons

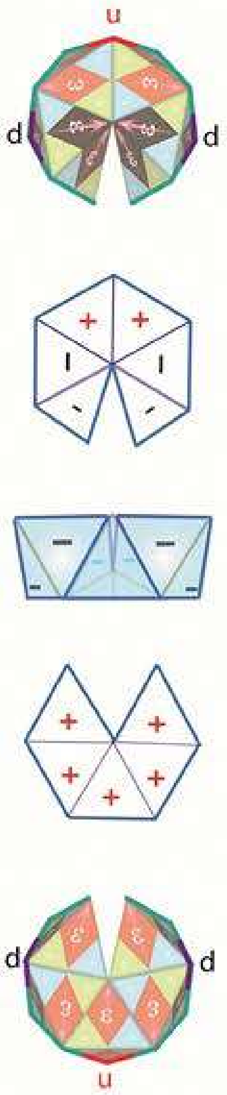

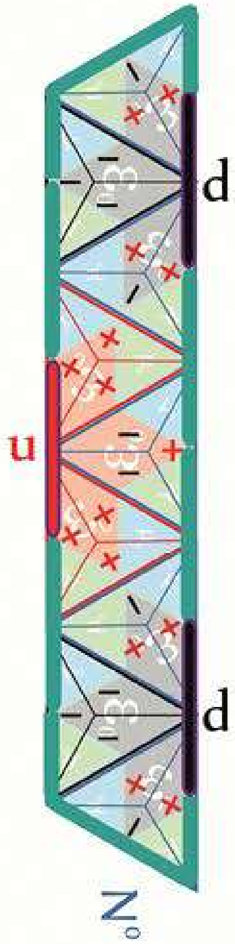

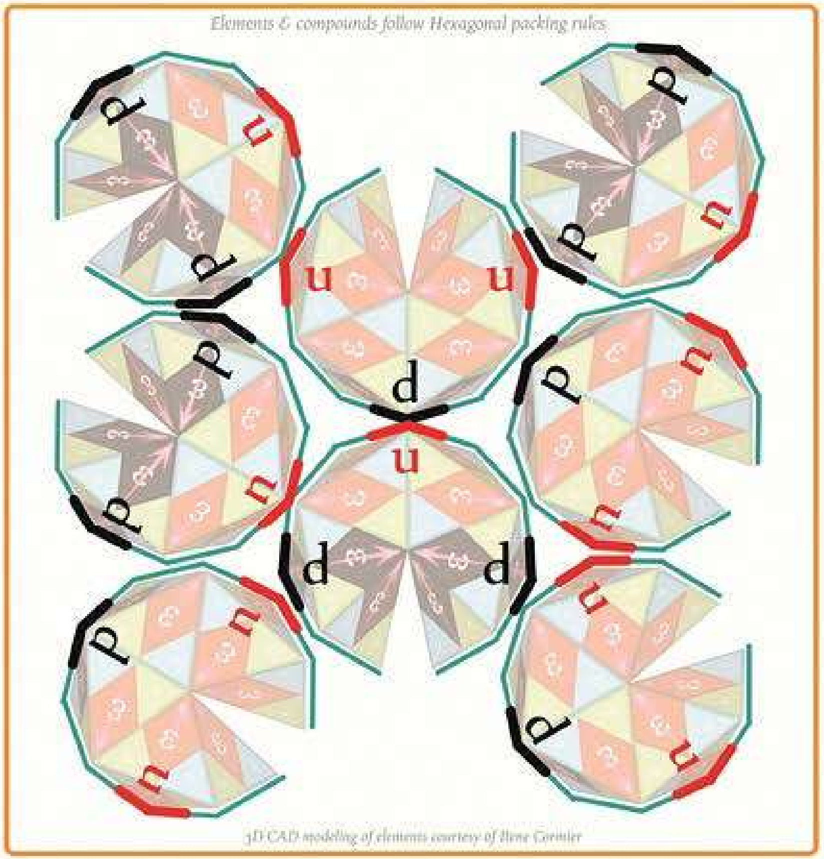

Elements & compounds follow Hexagonal packing rules

3D CAD modeling of elements courtesy of Ilene Cormier

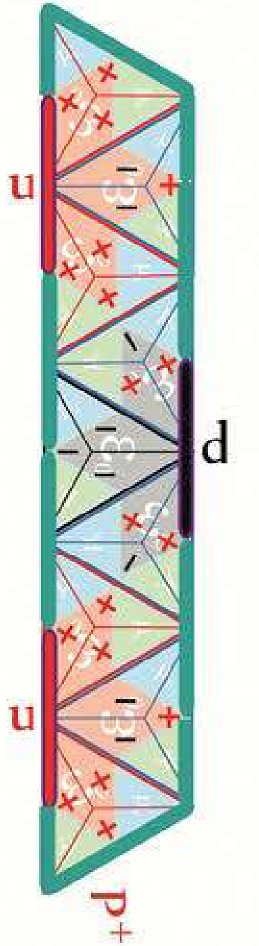

Protons

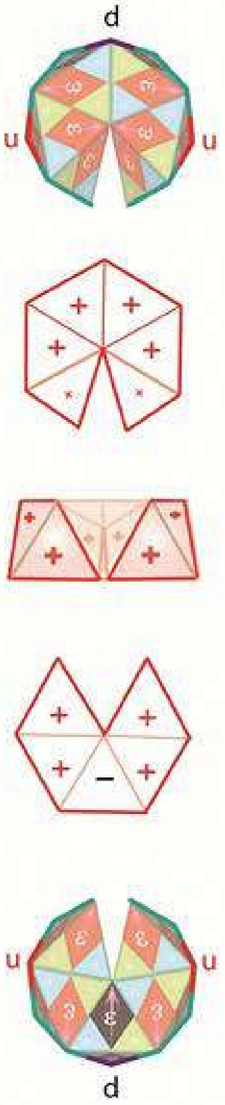

E-field apexes and their polarities highlight the quark alignment of all atomic nuclei and elements

UP Quark
Positive Electric field apex

DOWN Quark
Negative Electric field apex

Copyright ABRAHAM [2008] - All rights reserved

Insulators and Conductors

The position of electrons in Nuclei within Atomic Elements
results in the properties of Insulators or Conductors

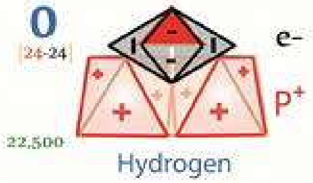
0
[24-24]
e-
P+
22,500
Hydrogen

Coloumbic forces

Electrons are attracted to the residual EM
net (+12) positive charge of Protons or
n[+12] unbalanced Ionic charges of nuclei

Conductor

Electrical energies move around
the material via boson exchanges
and electron movement

Insulator

Electrical energy is fixed within the nucleus
as electrostatic charges & released upon demand
via electron rotation/motion within the nucleus

Ions

Charge (energy) is moved around
material via electron movement

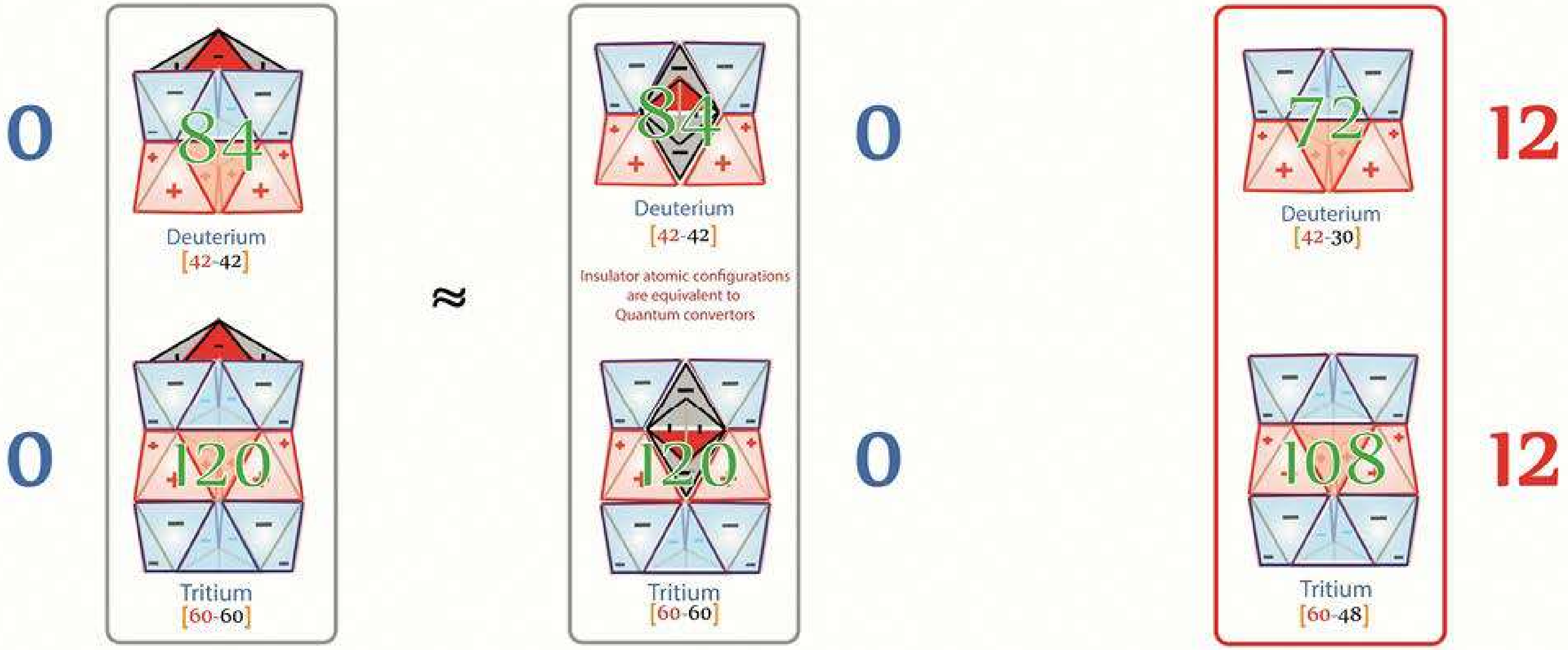
0
84
Deuterium
[42-42]
0
120
Tritium
[60-60]
≈
84
Deuterium
[42-42]
0
120
Tritium
[60-60]
0
72
Deuterium
[42-30]
12
108
Tritium
[60-48]
12

Insulator atomic configurations
are equivalent to
Quantum convertors

**Conductive materials contain 'free' electrons that
can be readily or easily moved within the material**

**Insulator materials have electrons that are 'bound' tightly to the atoms
and store charges locally where they are applied**

**Materials that have been ionised are
more likely to become Conductors
as they easily attract and bind
free electrons to them**

Copyright ABRAHAM [2008] - All rights reserved

Nucleon Quark Arrangement

The nuclei arrangement of each atomic shell [quantum level] is the result of quark EM field interactions

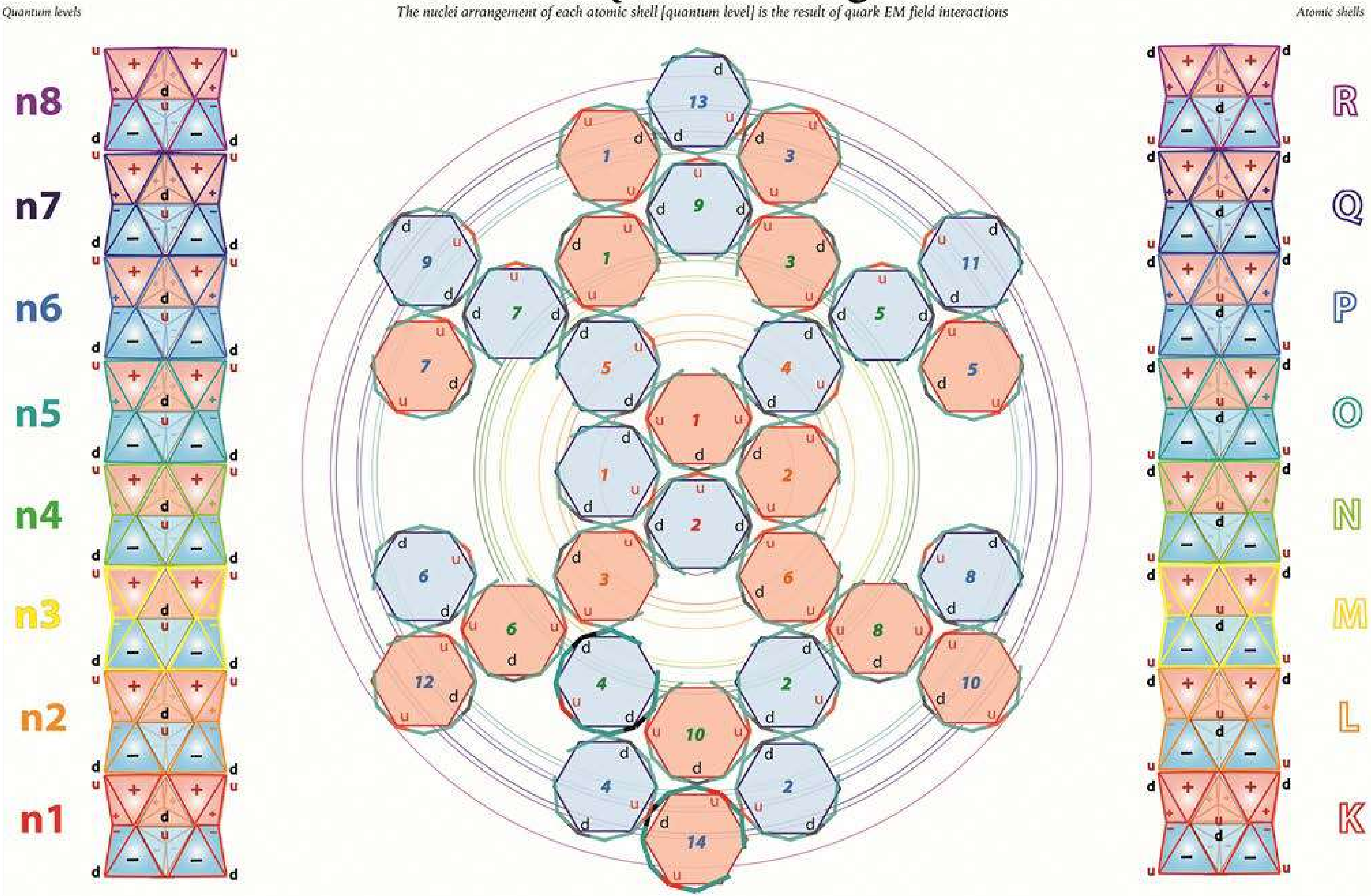
Quantum levels
Atomic shells
n8
n7
n6
n5
n4
n3
n2
n1
R
Q
P
O
N
M
L
K
+
−
u
d
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Copyright ABRAHAM [2008] - All rights reserved

Nucleon Charges and Bonding

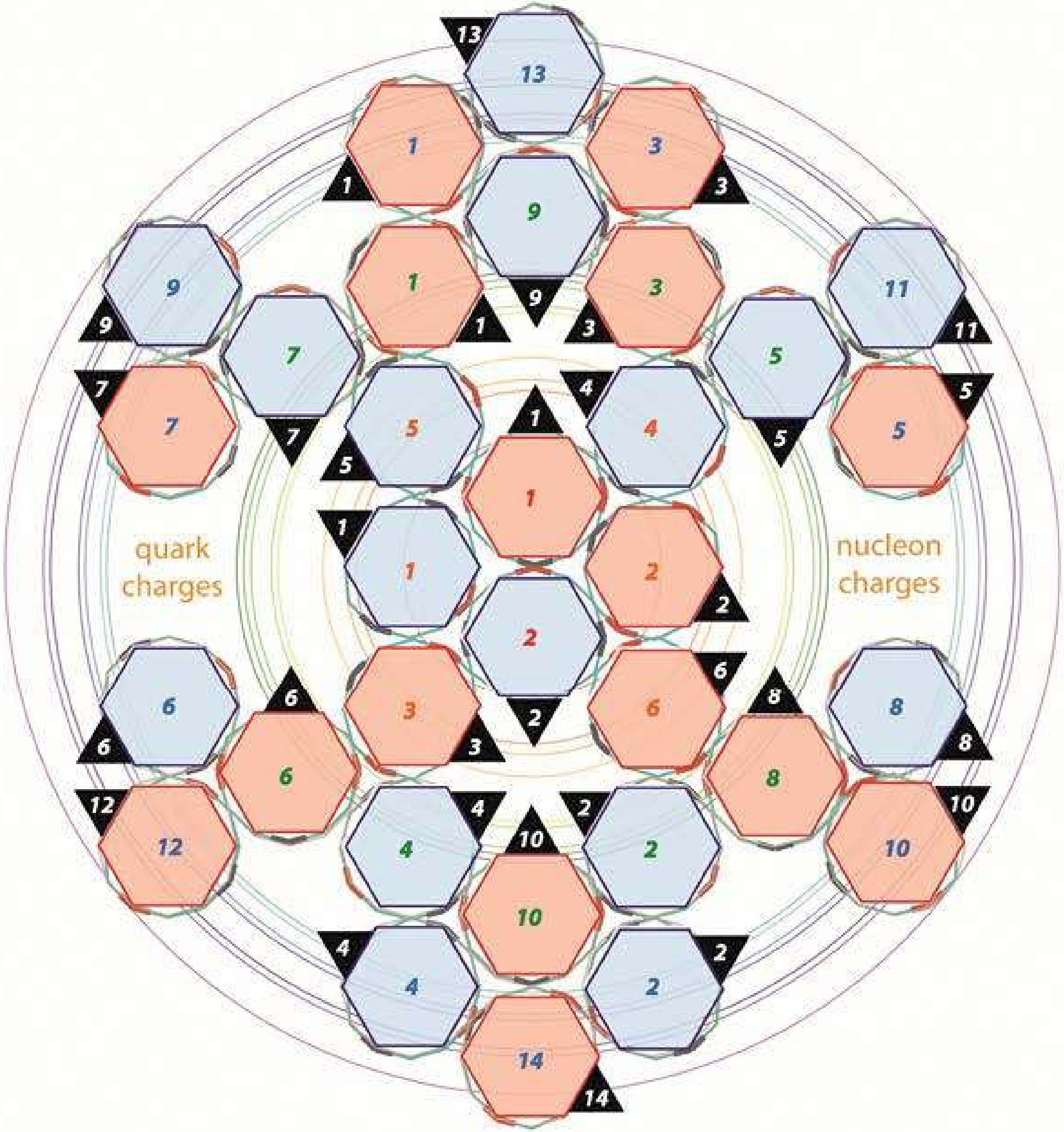
quark
charges
nucleon
charges

(Strong force - topological Electric Points)

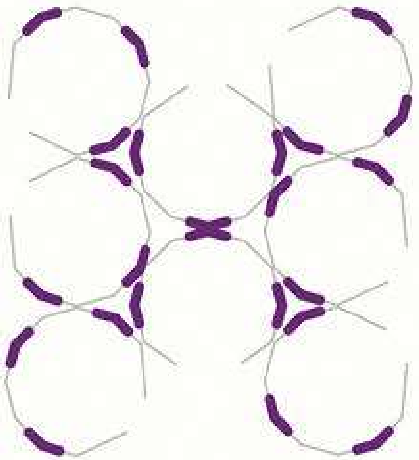

residual EM forces

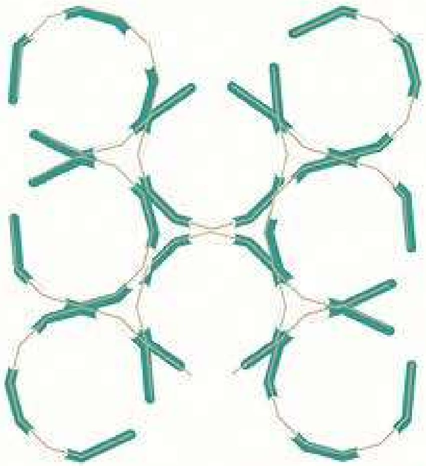

(Strong force - topological Magnetic dipoles)

Copyright ABRAHAM [2008] - All rights reserved

Bound electron arrangements

Externally bound electrons produce sub-orbital patterns different to the electron orbitals of internally bound electrons

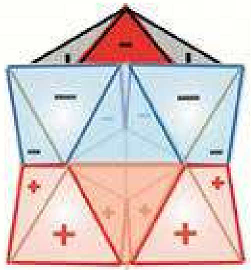

Deuterium

electrons are externally bound to the Deuteron nuclei

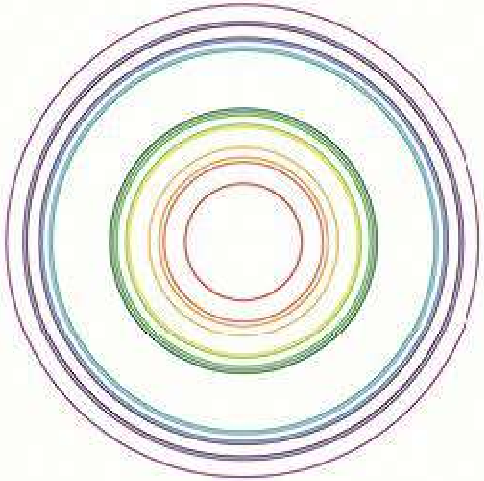

Conductors

Charges are free to move and equalise

electrons require less energy to 'break free' from Nuclei

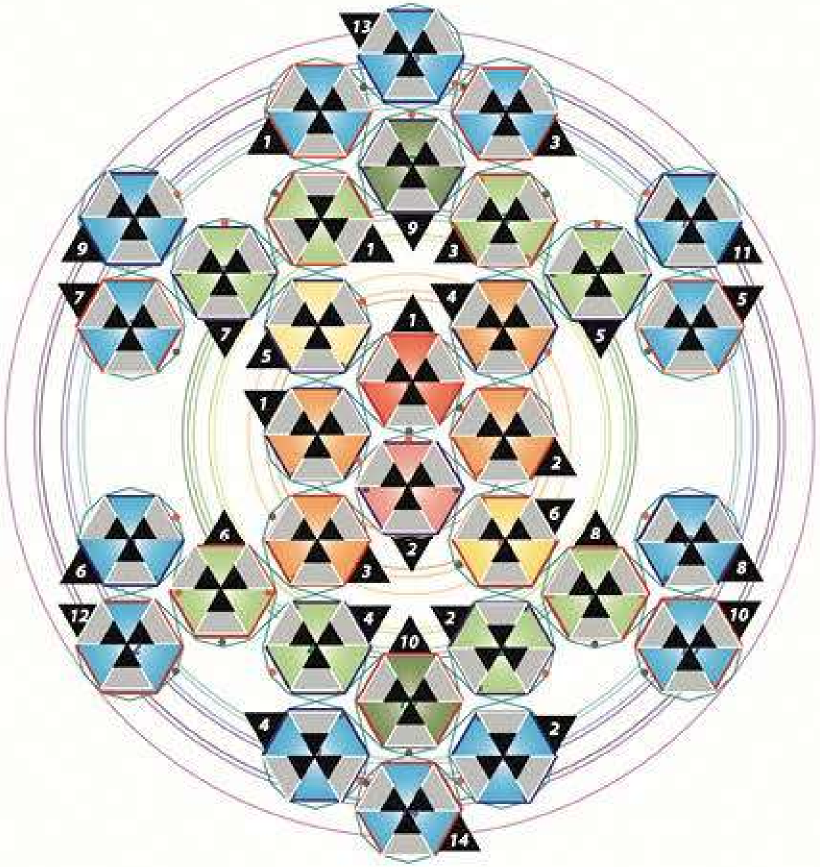

The electron orbitals of conductors are lower energies than those of insulators

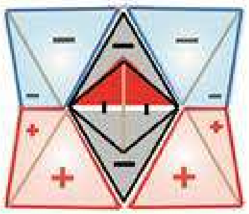

Deuterium

electrons are internally bound in the Deuteron nuclei

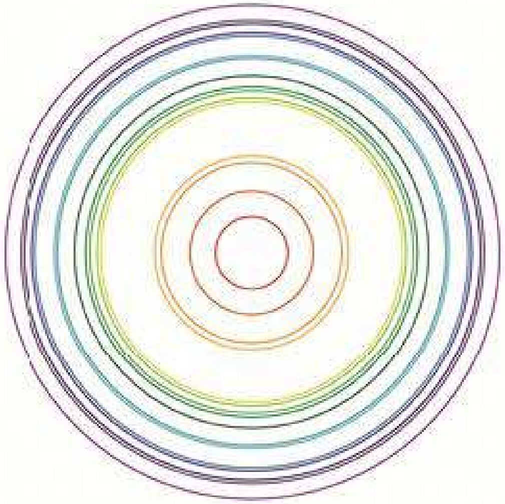

Insulators

Charges are bound to specific locations

electrons require more energy to 'break free' from Nuclei

Copyright ABRAHAM [2008] - All rights reserved

2D mass-energy geometries form the fabric of 3D Matter topologies

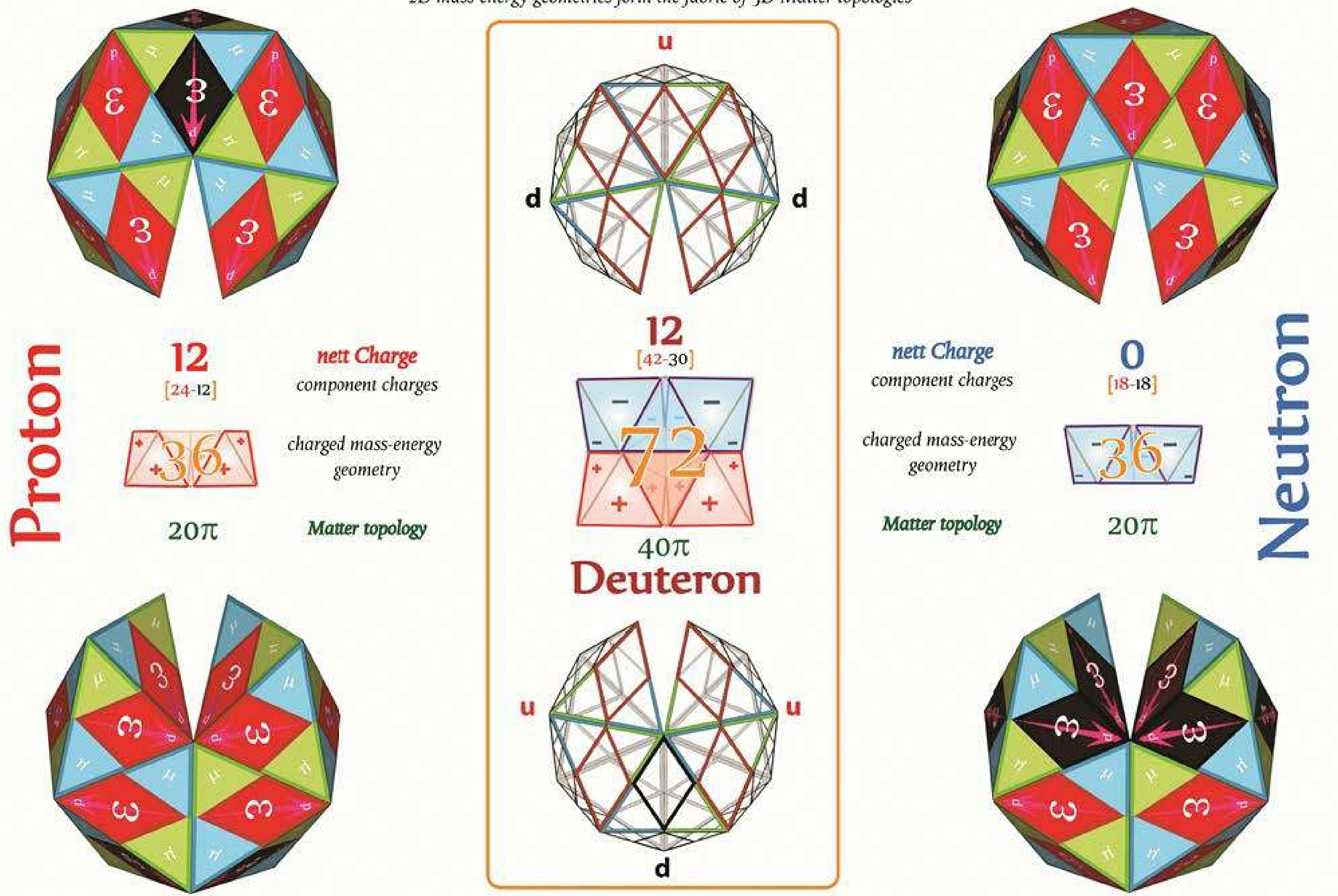

Proton

12
[24-12]

nett Charge
component charges

charged mass-energy
geometry

20π

Matter topology

12
[42-30]

40π

Deuteron

Neutron

nett Charge
component charges

0
[18-18]

charged mass-energy
geometry

Matter topology

20π

Charge provides the framework for the mass-energy geometry of Matter

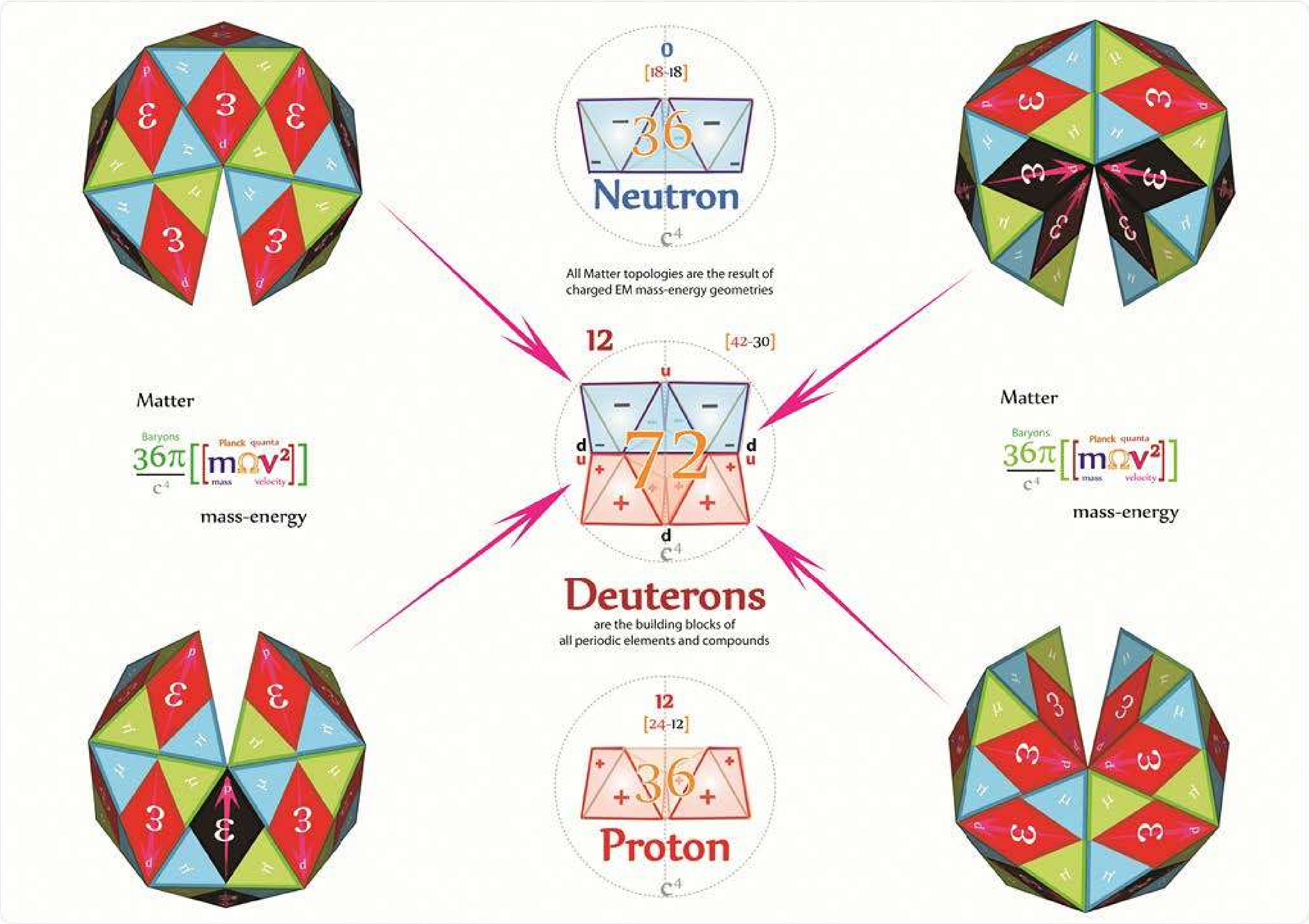

0

[18-18]

36

Neutron

c^4

All Matter topologies are the result of charged EM mass-energy geometries

12 [42-30]

72

Deuterons

c^4

are the building blocks of all periodic elements and compounds

12

[24-12]

36

Proton

c^4

Matter

Baryons, Planck quanta, mass, velocity

$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass-energy

Matter

Baryons, Planck quanta, mass, velocity

$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass-energy

Strong Forces and Nuclear Bonding

How do Baryons with Positive and Neutral charges attract each other and bind to form stable elements ?

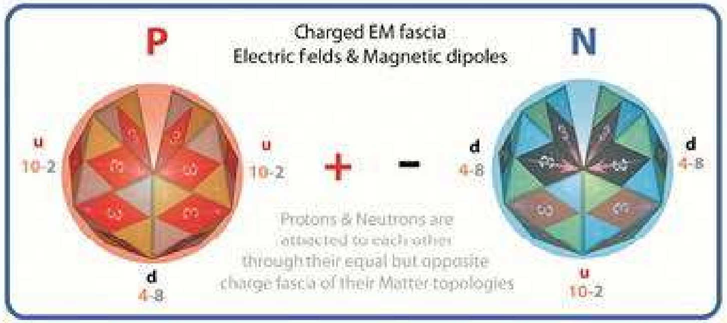
P
Charged EM fascia
Electric fields & Magnetic dipoles
N
u
10-2
u
10-2
+
-
d
4-8
d
4-8
d
4-8
u
10-2

Protons & Neutrons are attracted to each other through their equal but opposite charge fascia of their Matter topologies

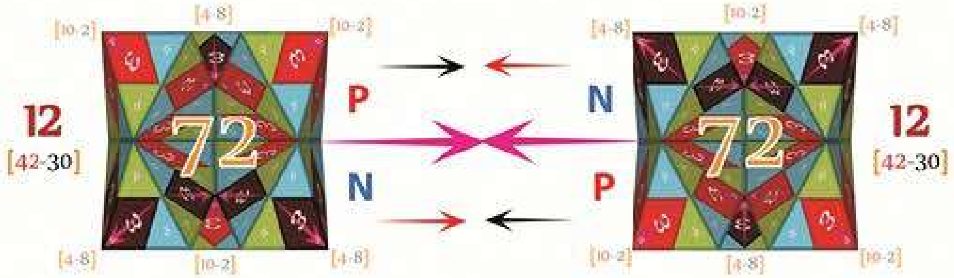
[10-2]
[4-8]
[10-2]
12
[42-30]
72
P
N
N
P
[4-8]
[10-2]
[4-8]
[4-8]
[10-2]
[4-8]
72
12
[42-30]
[10-2]
[4-8]
[10-2]

Once nuclei have been created their external electric fields & magnetic dipoles continue to attract and bind individual nuclei together via the Residual EM Force as nuclei seek charge equilibrium by combining with each other and electrons to form neutral elements

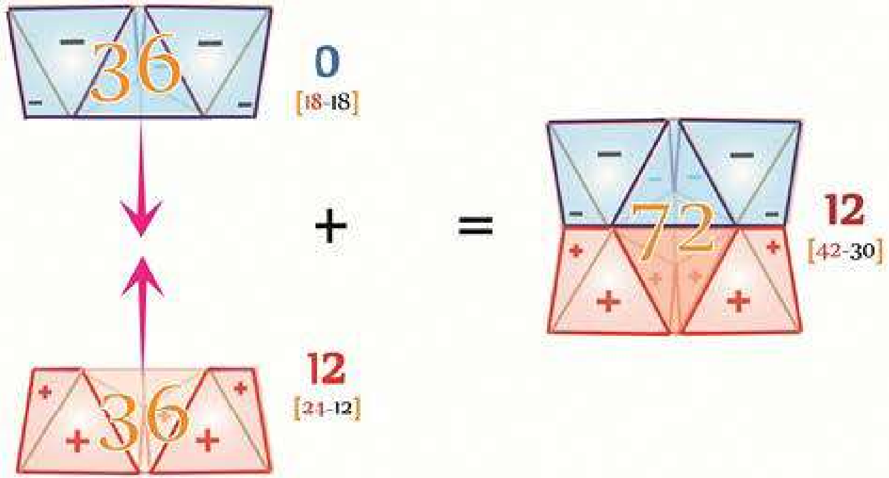
36
0
[18-18]
+
=
72
12
[42-30]
36
12
[24-12]

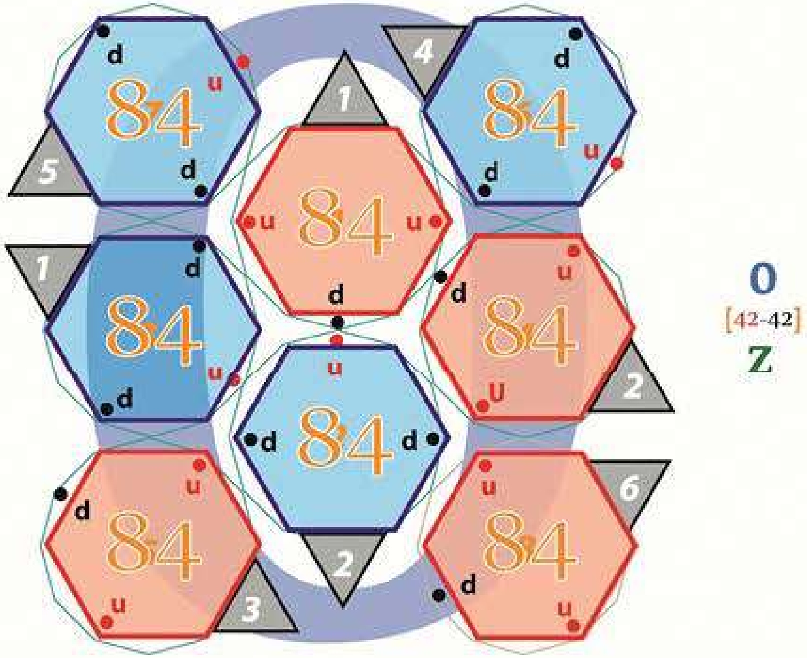
84
84
84
84
84
84
84
84
1
2
3
4
5
6
7
2
d
u
0
[42-42]
Z

The attraction and binding of Protons and Neutrons through their electric charge imbalances creates Deuterons which have +12 charges

The residual Z[+12] charge is what attracts electrons to form neutral atomic nuclei via Coulombic attraction

Copyright ABRAHAM [2008] - All rights reserved

Strong Nucleonic Bonding

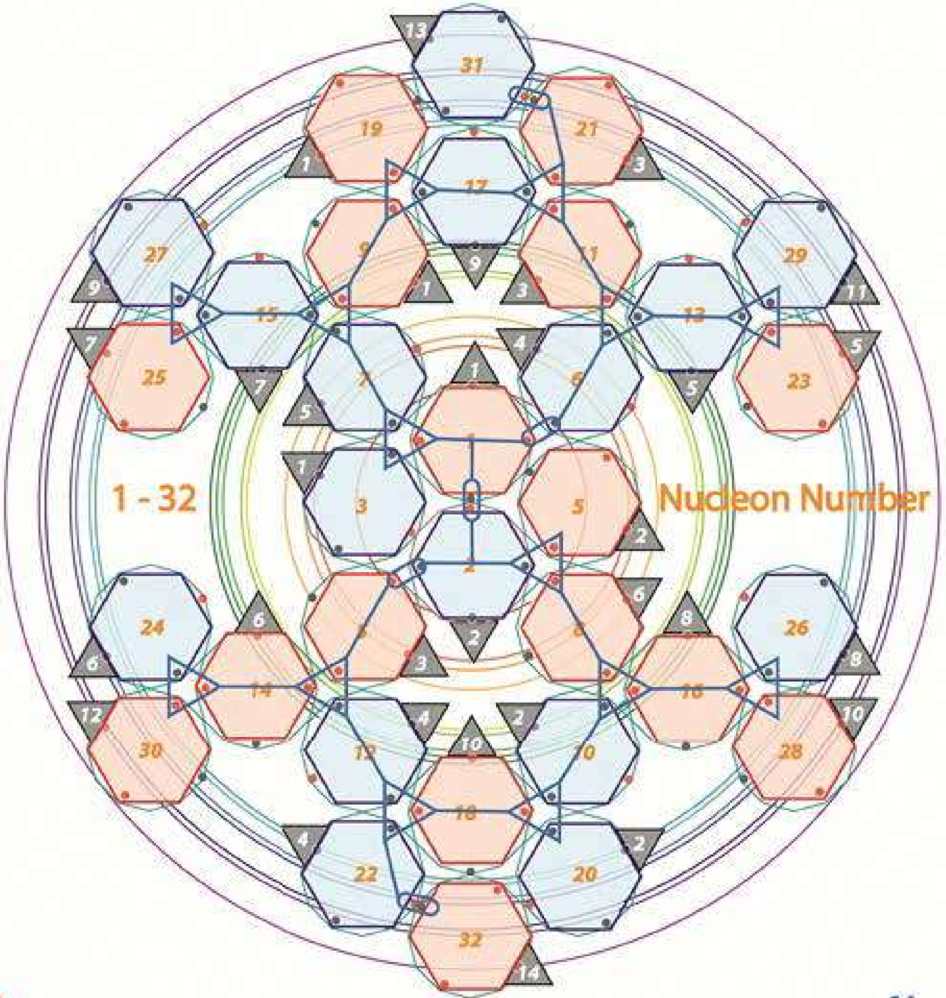
1 - 32
Nucleon Number
31
19
21
17
27
29
15
13
25
23
3
5
24
26
14
16
30
28
18
22
20
32

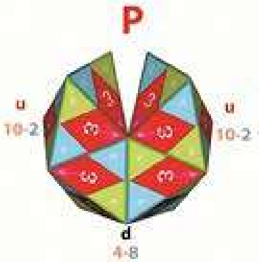
P
u
10-2
u
10-2
d
4-8

+

In addition to the Strong colour force
a separate residual binding force arises
from the external apexes formed by
Positive and Negative Electric points of quarks
in each Baryon [nuclear-chemical bonds]

-

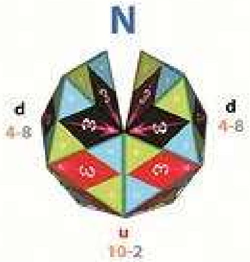
N
d
4-8
d
4-8
u
10-2

All energy seeks equilibrium

Strong E-force
nuclear linkages

External Electric field points
bind via Charge fascia interactions
with Plus and Minus Electric points
combining and sharing energy
throughout the resultant nuclei

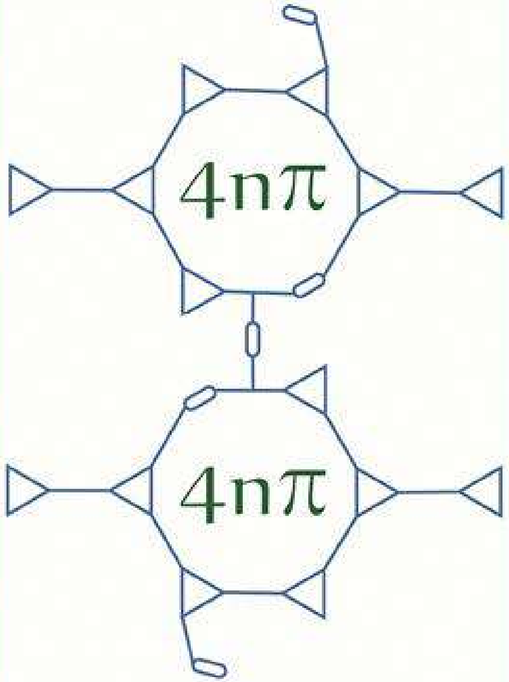

● Positive Electric field apex

Negative Electric field apex ●

Hydrogenic vs Nucleonic electron binding

If a unbound Proton attracts an Electron
the Electron can be bound to the nuclei
in a number of differing orientations
[each with differing spin energies]

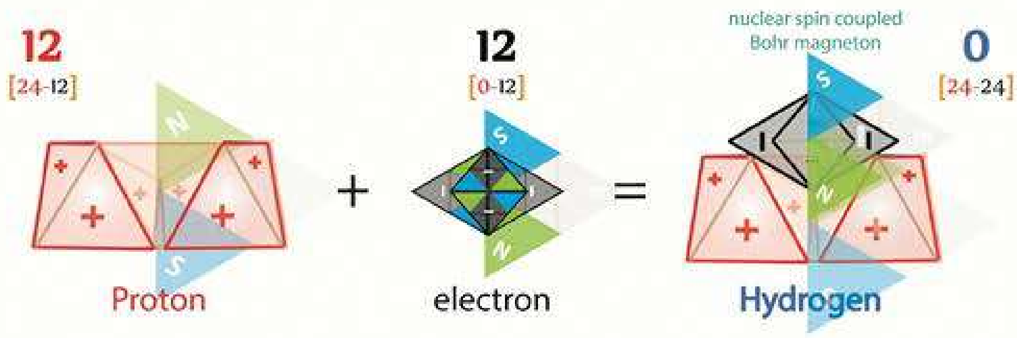
12
[24-12]
Proton
+
12
[0-12]
electron
=
nuclear spin coupled
Bohr magneton
0
[24-24]
Hydrogen
N
S
+

The energy levels of Baryons determines the KEM field energy of bound electrons

All atomic nuclei (and elements) are Deuteron'ic nuclei with a mixture of *orthagonal*, parallel and anti-paralllel spin orientations

(this is why Rydberg is less accurate for elemental nuclei compared to Hydrogenic atoms - see QM spin)

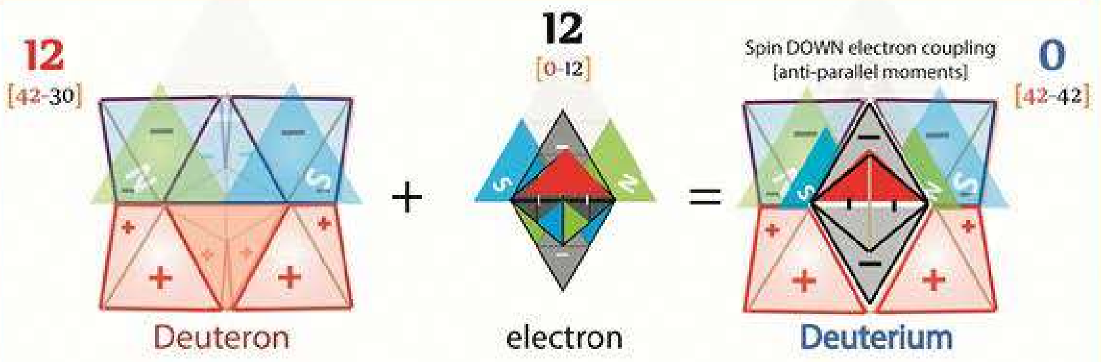
12
[42-30]
Deuteron
+
12
[0-12]
electron
=
Spin DOWN electron coupling
[anti-parallel moments]
0
[42-42]
Deuterium
N
S
+
−

Bohr magnetons are always referenced wrt the Nuclear magneton

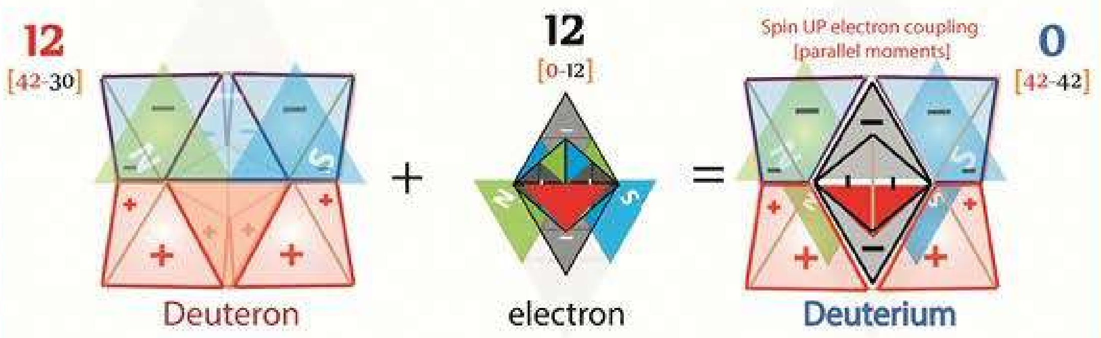
12
[42-30]
Deuteron
+
12
[0-12]
electron
=
Spin UP electron coupling
[parallel moments]
0
[42-42]
Deuterium
N
S
+
−

electrons produce stronger magnetic moments due to their mass-charge quotient

Electrons can be bound to deuteron nuclei
in four distinct orientations [2 horizontal & 2 vertical]
with each spin coupling orientation producing
differing energy electron orbitals
[wrt to the nuclear magnetic moments]

Ejecting electrons from atomic nuclei
by adding energies to their KEM fields
[the Photoelectric effect]
creates Positive Ions

Vertically orientated electrons
within Proton-Neutron Nuclei [Deuterons]
create quantum synchronous convertor geometries

Copyright ABRAHAM [2008] - All rights reserved

Atomic Nucleus Master Template

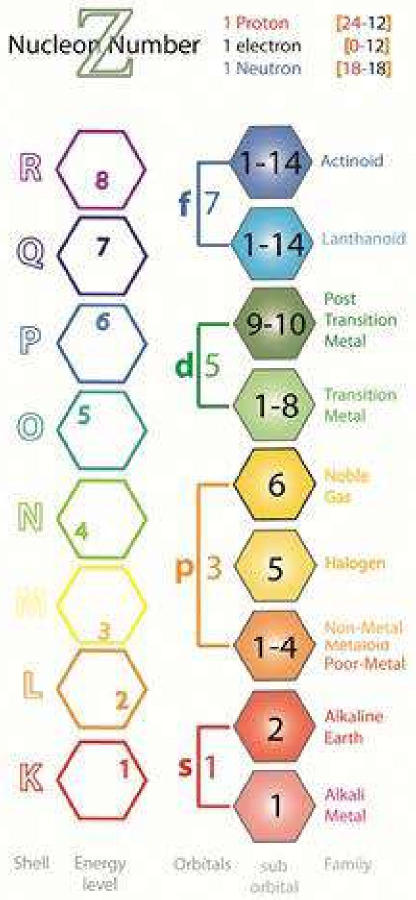
Nucleon Z Number
1 Proton [24-12]
1 electron [0-12]
1 Neutron [18-18]
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

o s 1 sub-Orbital (2 electrons max)

o p 3 sub-Orbitals (6 electrons max)

o d 5 sub-Orbitals (10 electrons max)

o f 7 sub-Orbitals (14 electrons max)

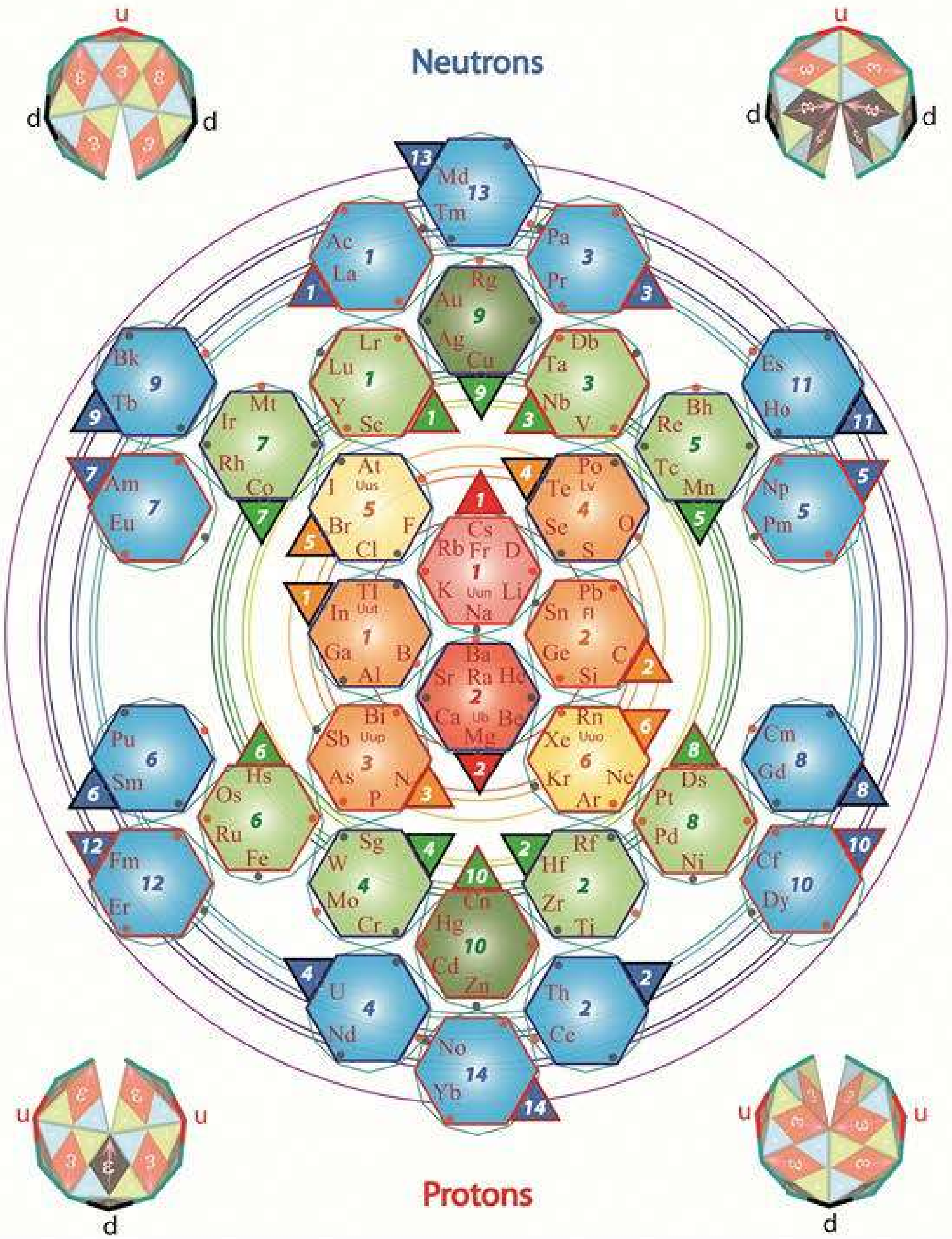
u
d
d
u
d
d
Neutrons
13
Md Tm 13
Ac La 1
1
Pa Pr 3
3
Rg Au Ag Cu 9
9
Lr Lu Y Sc 1
1
Db Ta Nb V 3
3
Bk Tb 9
9
Mt Ir Rh Co 7
7
Bh Re Tc Mn 5
5
Es Ho 11
11
7
Am Eu 7
At I Uus Br Cl F 5
5
1
Cs Rb Fr D 1 K Uun Li Na
Po Te Lv Se S O 4
4
Np Pm 5
5
1
Tl In Uut Ga Al B 1
Ba Sr Ra He 2 Ca Ub Be Mg
2
Pb Sn Fl Ge Si C 2
2
Pu Sm 6
6
6
Os Hs Ru Fe 6
Bi Sb Uup As P N 3
3
Rn Xe Uuo Kr Ne Ar 6
6
8
Ds Pt Pd Ni 8
Cn Gd 8
8
12
Fm Er 12
4
Sg W Mo Cr 4
10
Cn Hg Cd Zn 10
2
Rf Hf Zr Ti 2
10
Cf Dy 10
4
U Nd 4
2
Th Ce 2
No Yb 14
14
u
u
d
u
u
d
Protons

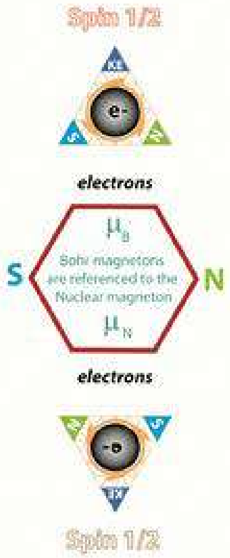
Spin 1/2
e-
electrons
S
N
μ_B
Bohr magnetons are referenced to the Nuclear magneton
μ_N
electrons
-e
Spin 1/2

Copyright ABRAHAM [2008] - All rights reserved

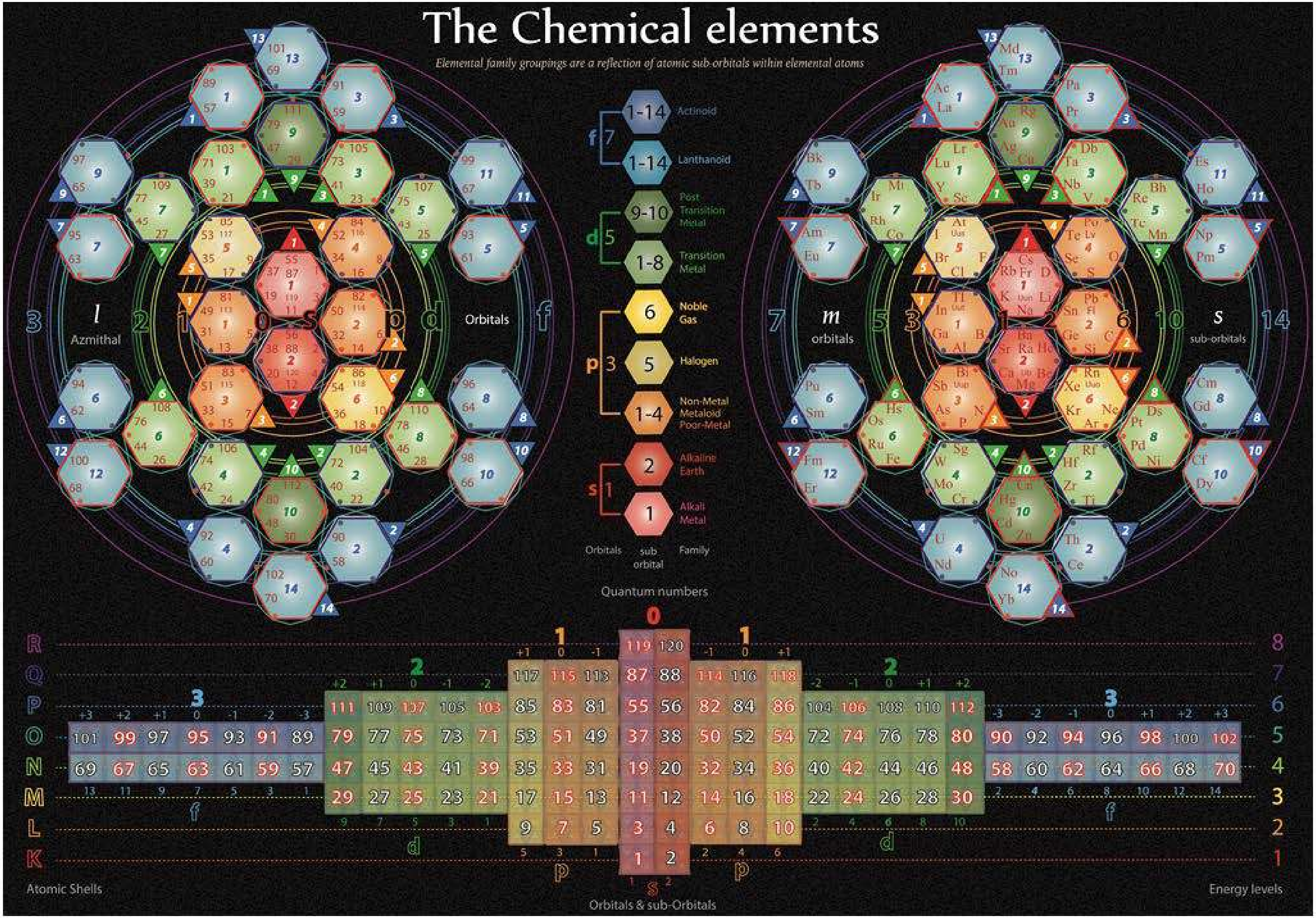
The Chemical elements
Elemental family groupings are a reflection of atomic sub-orbitals within elemental atoms
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Orbitals
sub orbital
Family
Quantum numbers
l
Azmithal
Orbitals
m
orbitals
s
sub-orbitals
R
Q
P
O
N
M
L
K
Atomic Shells
8
7
6
5
4
3
2
1
Energy levels
Orbitals & sub-Orbitals

Tetryonics 43.02 - The Chemical Elements

Element Numbers

Z *

1 Proton	[24-12]
1 Neutron	[18-18]
1 electron	[0-12]

Each elemental nuclei is made from Deuterium

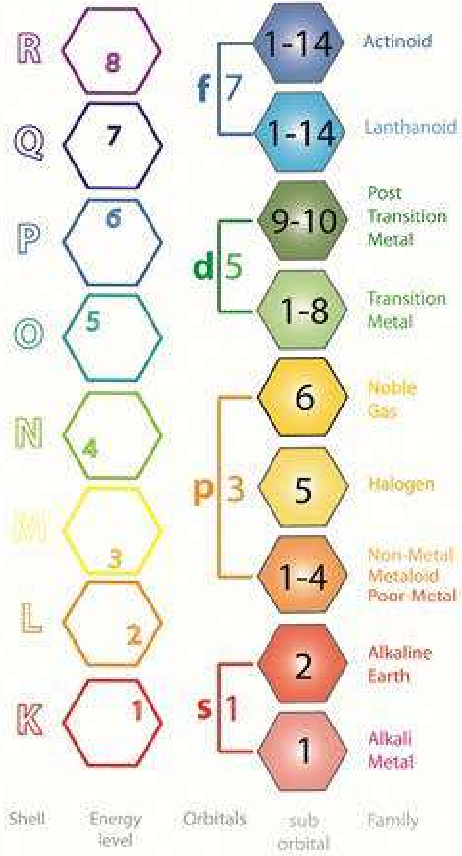
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
f 7
1-14
Actinoid
1-14
Lanthanoid
d 5
9-10
Post Transition Metal
1-8
Transition Metal
p 3
6
Noble Gas
5
Halogen
1-4
Non-Metal Metaloid Poor Metal
s 1
2
Alkaline Earth
1
Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

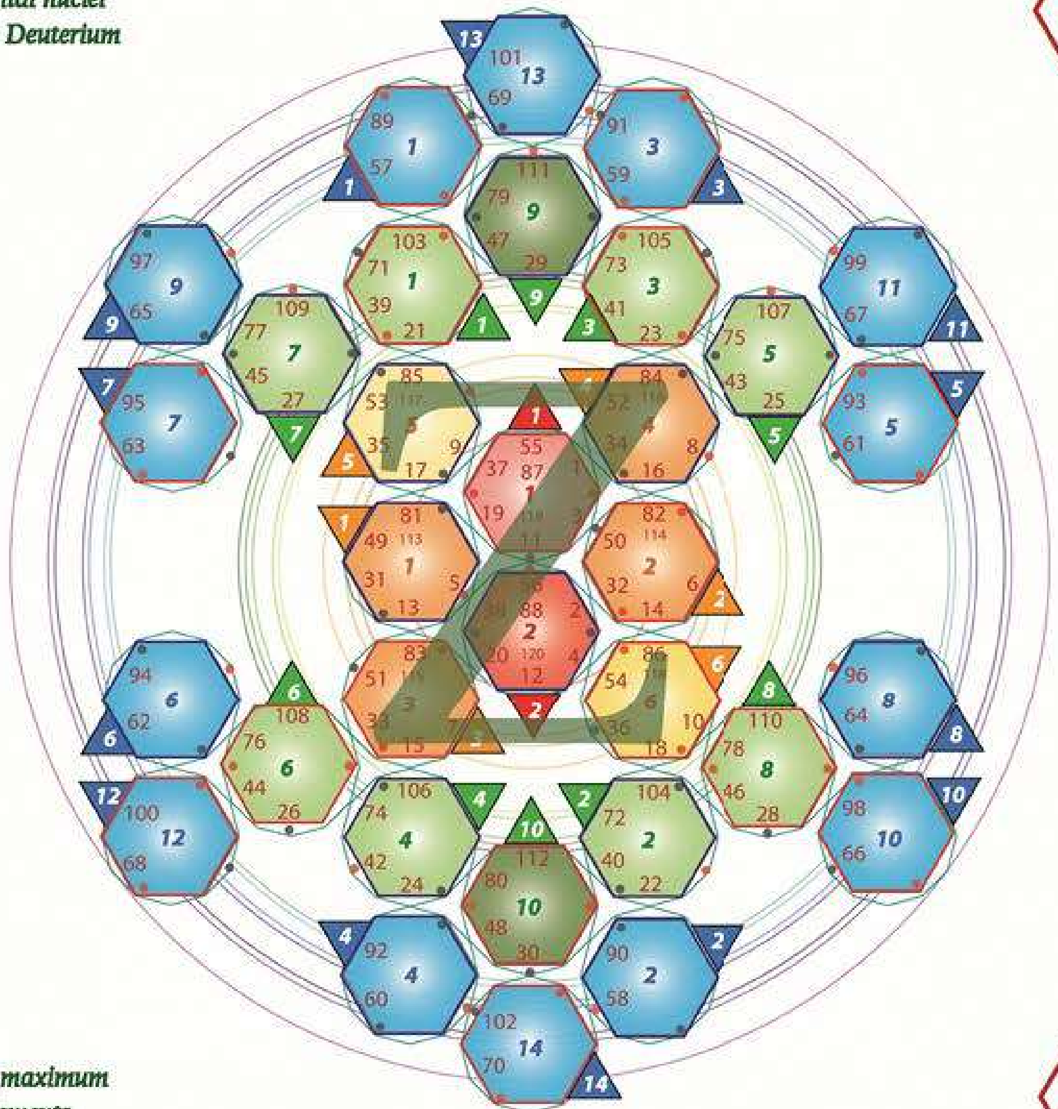

Atomic shells
P
O
Q
K
N
R
L
M

o	s	*1 sub-Orbital*	*(2 electrons max)*
O	p	*3 sub-Orbitals*	*(6 electrons max)*
O	d	*5 sub-Orbitals*	*(10 electrons max)*
O	f	*7 sub-Orbitals*	*(14 electrons max)*

There are a maximum of 120 elements possible

6
5
7
1
4
8
2
3
Energy levels

Element Names

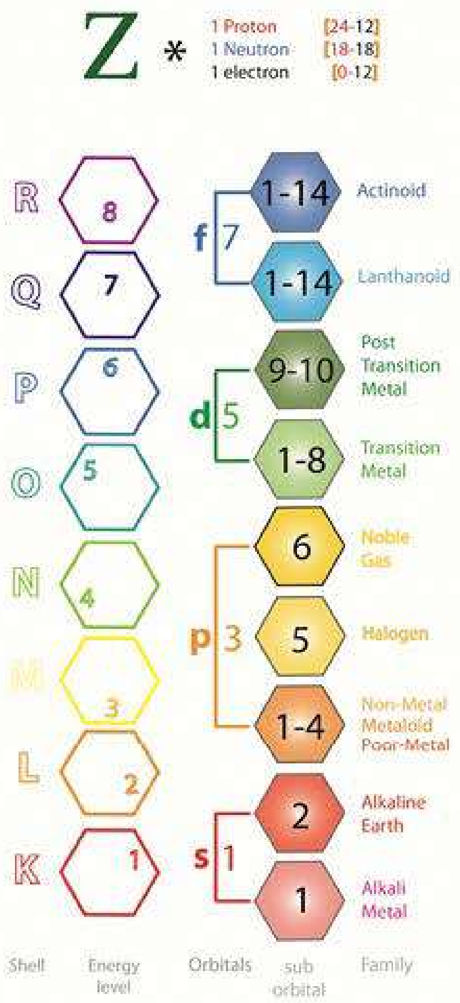
Z
*
1 Proton [24-12]
1 Neutron [18-18]
1 electron [0-12]
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

○ s 1 sub-Orbital (2 electrons max)

○ p 3 sub-Orbitals (6 electrons max)

○ d 5 sub-Orbitals (10 electrons max)

○ f 7 sub-Orbitals (14 electrons max)

There are a maximum of 120 elements possible

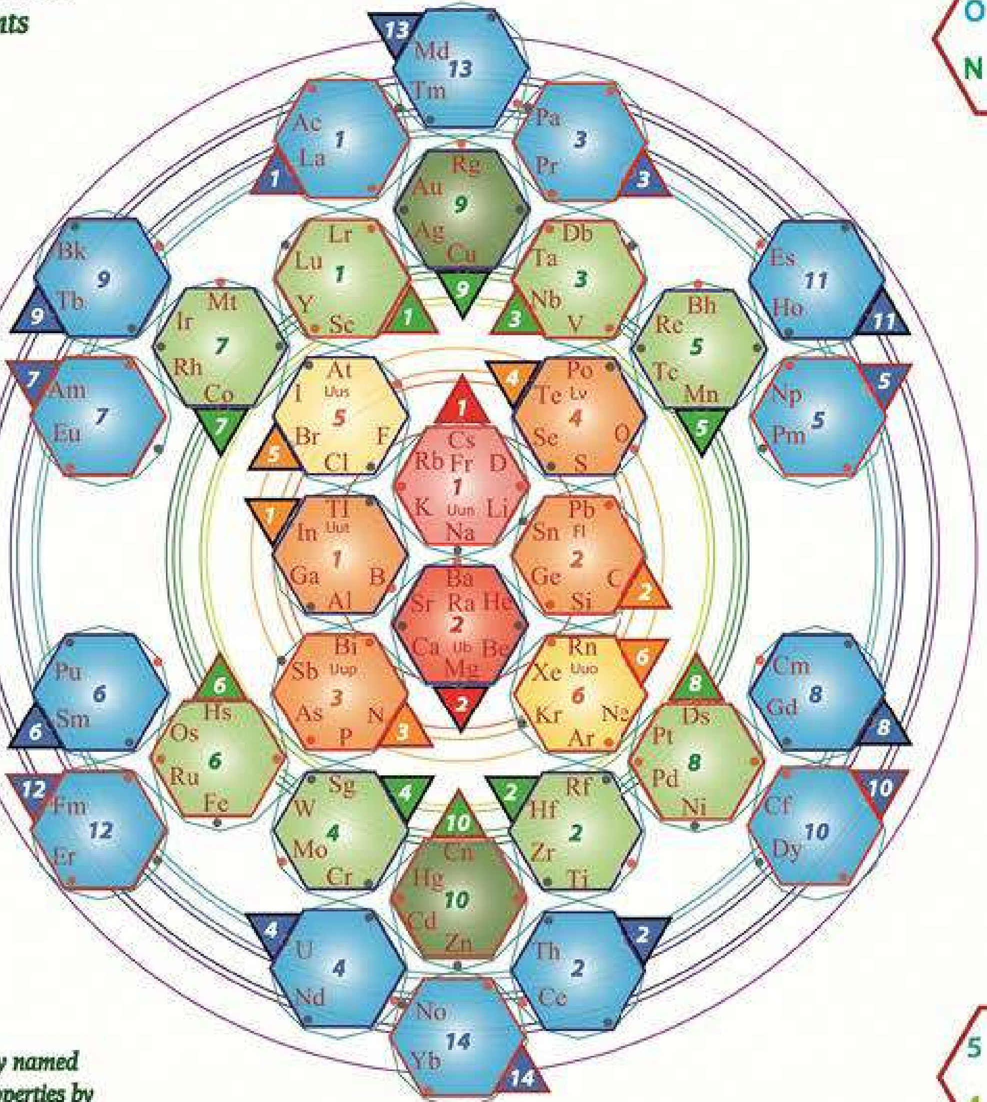
13 Md Tm 13
1 Ac La 1
3 Pa Pr 3
Rg Au Ag Cu 9
9
Lr Lu Y Sc 1
1
Db Ta Nb V 3
3
9 Bk Tb 9
11 Es Ho 11
Mt Ir Rh Co 7
7
Bh Re Tc Mn 5
5
7 Am Eu 7
5 Np Pm 5
At I Uus Br Cl F 5
5
4
Po Te Lv Se O S 4
1
Cs Rb Fr D K Uun Li Na 1
Tl In Uut Ga B Al 1
1
Pb Sn Fl Ge C Si 2
2
Ba Sr Ra He Ca Ub Be Mg 2
2
Bi Sb Uup As N P 3
3
Rn Xe Uuo Kr Ne Ar 6
6
6 Pu Sm 6
8 Cm Gd 8
6
Hs Os Ru Fe 6
8
Ds Pt Pd Ni 8
12 Fm Er 12
10 Cf Dy 10
Sg W Mo Cr 4
4
2
Rf Hf Zr Ti 2
10
Cn Hg Cd Zn 10
4 U Nd 4
2 Th Ce 2
No Yb 14
14

Atomic shells

They were originally named according to their properties by their discoverer but have recently been named after famous scientists

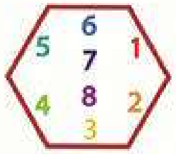

Energy levels

Copyright ABRAHAM [2008] - All rights reserved

Electron orbital configurations

A huge number of differing d and f orbital configurations are possible given the number of nuclei and bond points created by elemental topologies

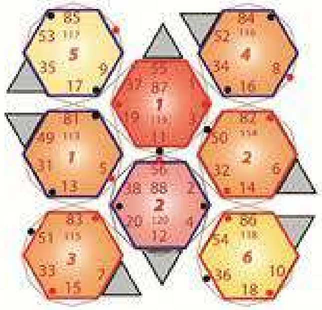
85
53
5
35
9
17
84
52
4
34
8
16
87
37
1
19
11
81
49
1
31
5
13
82
50
2
32
6
14
88
38
2
20
12
83
51
3
33
7
15
86
54
6
36
10
18

As the number of nucleons increases so does the complexity of the electron orbitals possible

quantum snowflakes

However all have a stable 'core' grouping of nuclei comprised of s and p electron orbitals

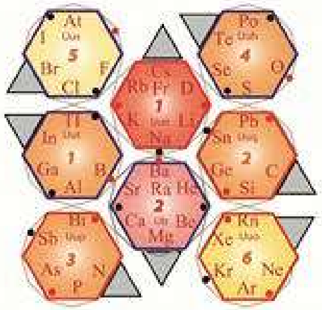
At
I
5
Br
F
Cl
Po
Te
4
Se
O
S
Cs
Rb
Fr
1
K
Li
Na
Tl
In
1
Ga
B
Al
Pb
Sn
2
Ge
C
Si
Ba
Sr
Ra
2
Ca
Be
Mg
Bi
Sb
3
As
N
P
Rn
Xe
6
Kr
Ne
Ar

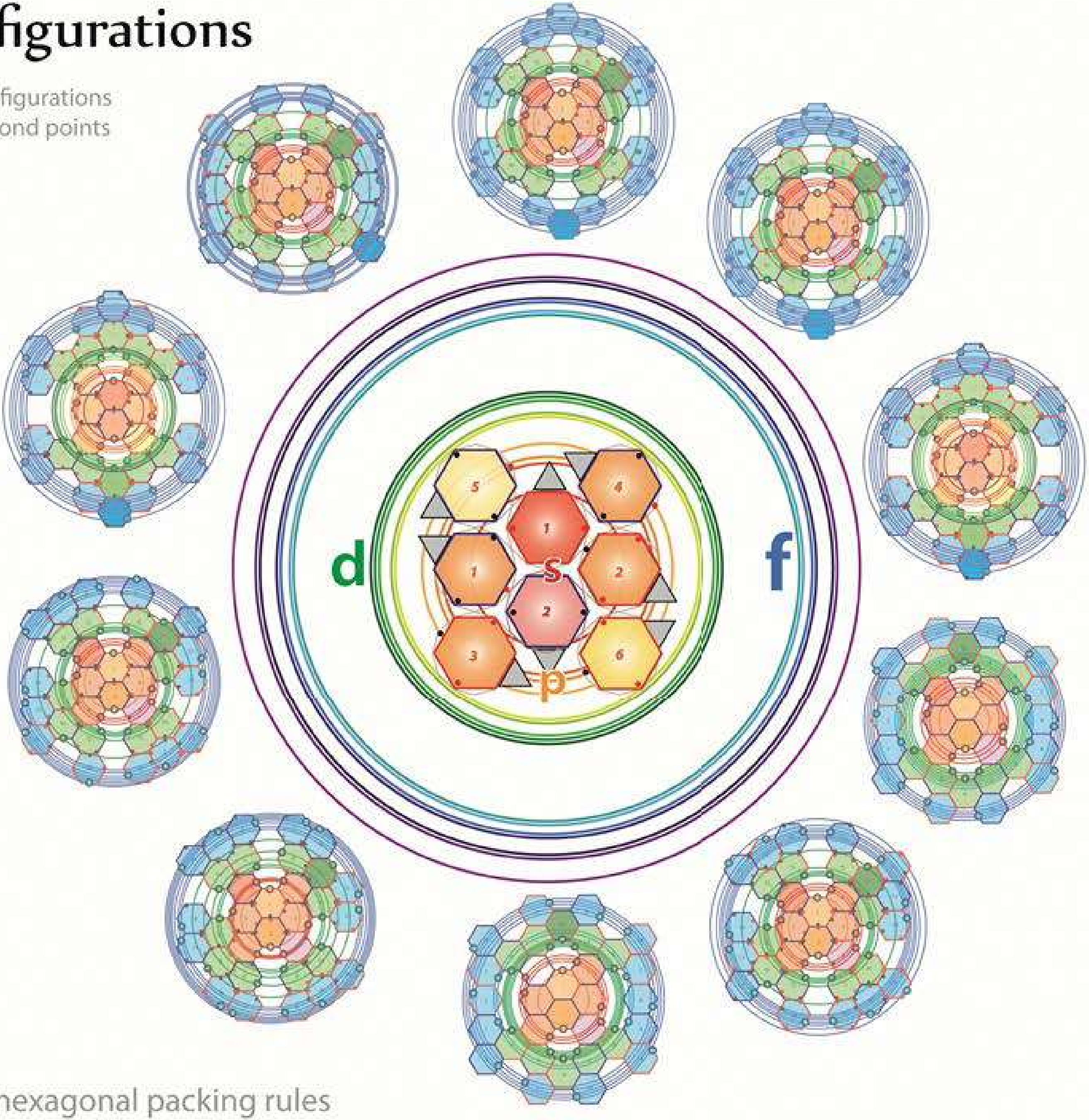
d
f
s
p
5
4
1
1
2
2
3
6

All nuclei bonding closely follows hexagonal packing rules

Atomic Nucleus
Exploded view

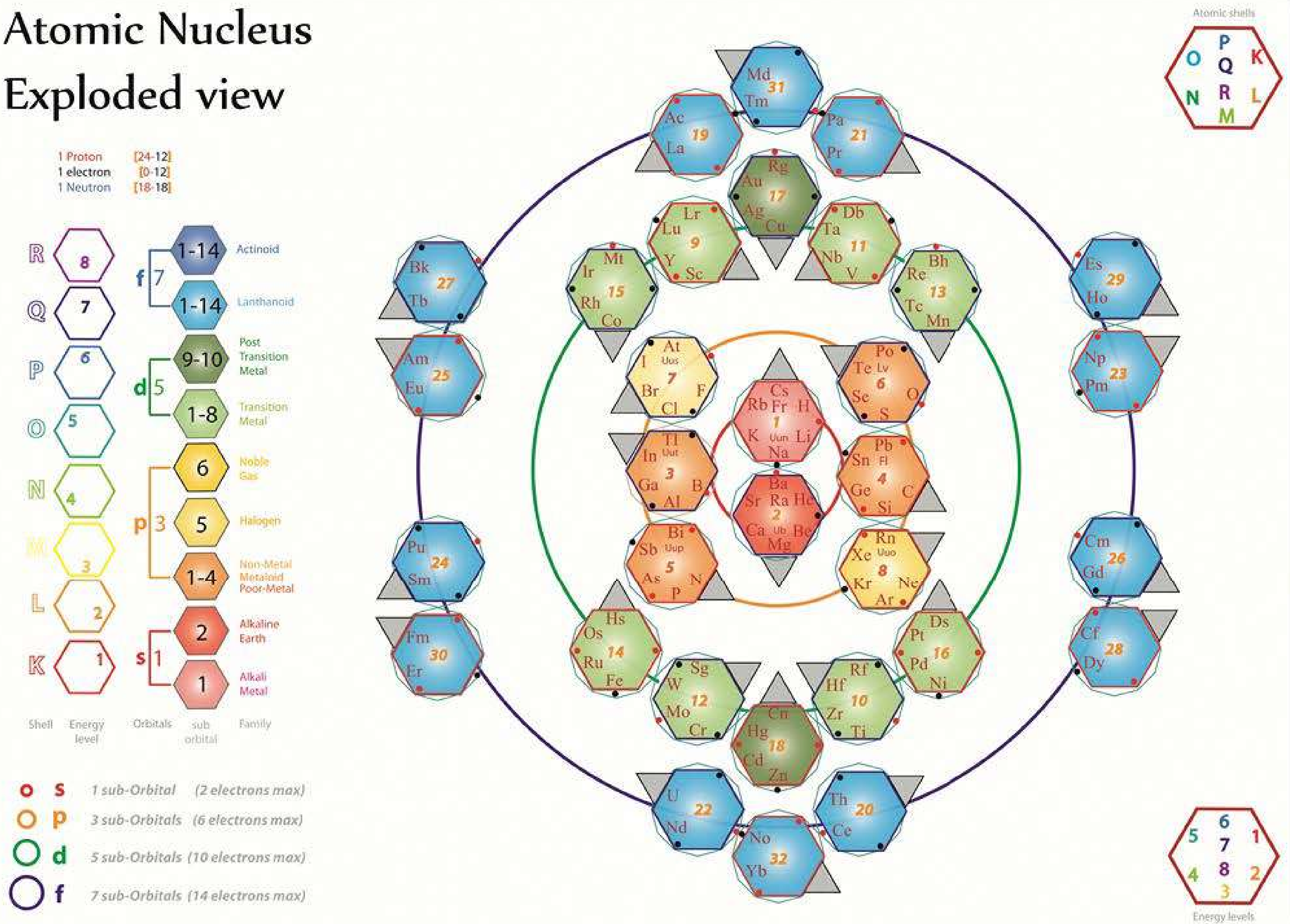
1 Proton [24-12]
1 electron [0-12]
1 Neutron [18-18]

R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
f 7
1-14
Actinoid
1-14
Lanthanoid
d 5
9-10
Post Transition Metal
1-8
Transition Metal
p 3
6
Noble Gas
5
Halogen
1-4
Non-Metal Metaloid Poor-Metal
s 1
2
Alkaline Earth
1
Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family
Atomic shells
P
O
Q
K
N
R
L
M
Md 31 Tm
Ac 19 La
Pa 21 Pr
Rg Au 17 Ag Cu
Lr Lu 9 Y Sc
Db Ta 11 Nb V
Mt Ir 15 Rh Co
Bh Re 13 Tc Mn
Bk 27 Tb
Es 29 Ho
Am 25 Eu
Np 23 Pm
At I Uus 7 Br F Cl
Cs Rb Fr H 1 K Uun Li Na
Po Te Lv 6 Se O S
Tl In Uut 3 Ga B Al
Ba Sr Ra He 2 Ca Ub Be Mg
Pb Sn Fl 4 Ge C Si
Bi Sb Uup 5 As N P
Rn Xe Uuo 8 Kr Ne Ar
Pu 24 Sm
Cm 26 Gd
Fm 30 Er
Cf 28 Dy
Hs Os 14 Ru Fe
Ds Pt 16 Pd Ni
Sg W 12 Mo Cr
Rf Hf 10 Zr Ti
Cn Hg 18 Cd Zn
U 22 Nd
Th 20 Ce
No 32 Yb
6
5
1
7
4
8
2
3
Energy levels

o **s** *1 sub-Orbital (2 electrons max)*
O **p** *3 sub-Orbitals (6 electrons max)*
O **d** *5 sub-Orbitals (10 electrons max)*
O **f** *7 sub-Orbitals (14 electrons max)*

Tetryonics 43.06 - Exploded Atomic Nucleus

Atomic radii of elements

The Mendeleev block arrangement of periodic elements presents a number of problems in representing the true topologies and properties of all the periodic elements:

Questions are raised over Hydrogen's appropriate position in the table

The table correctly positions elements according to their elemental properties but offers no insight into the individual geometries of the periodic elements themselves

The table does not reflect the charged quantum geometries of nuclei in each atomic shell

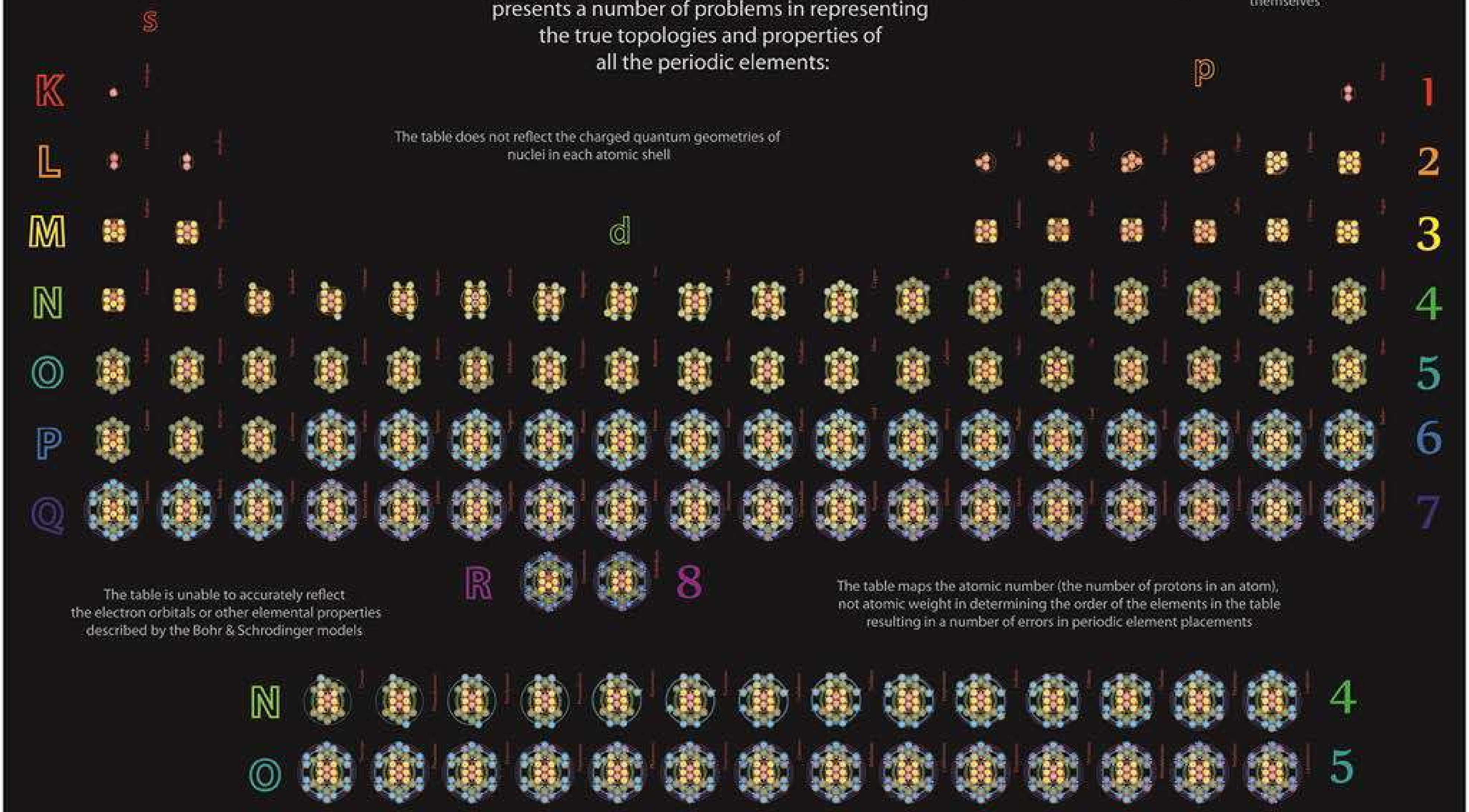

The table is unable to accurately reflect the electron orbitals or other elemental properties described by the Bohr & Schrodinger models

The table maps the atomic number (the number of protons in an atom), not atomic weight in determining the order of the elements in the table resulting in a number of errors in periodic element placements

'f' orbitals elements are not normally presented in their true atomic number positions

Tetryonic charge geometry rectifies all of these deficiencies

Tetryonics 43.07 - Atomic radii of elements

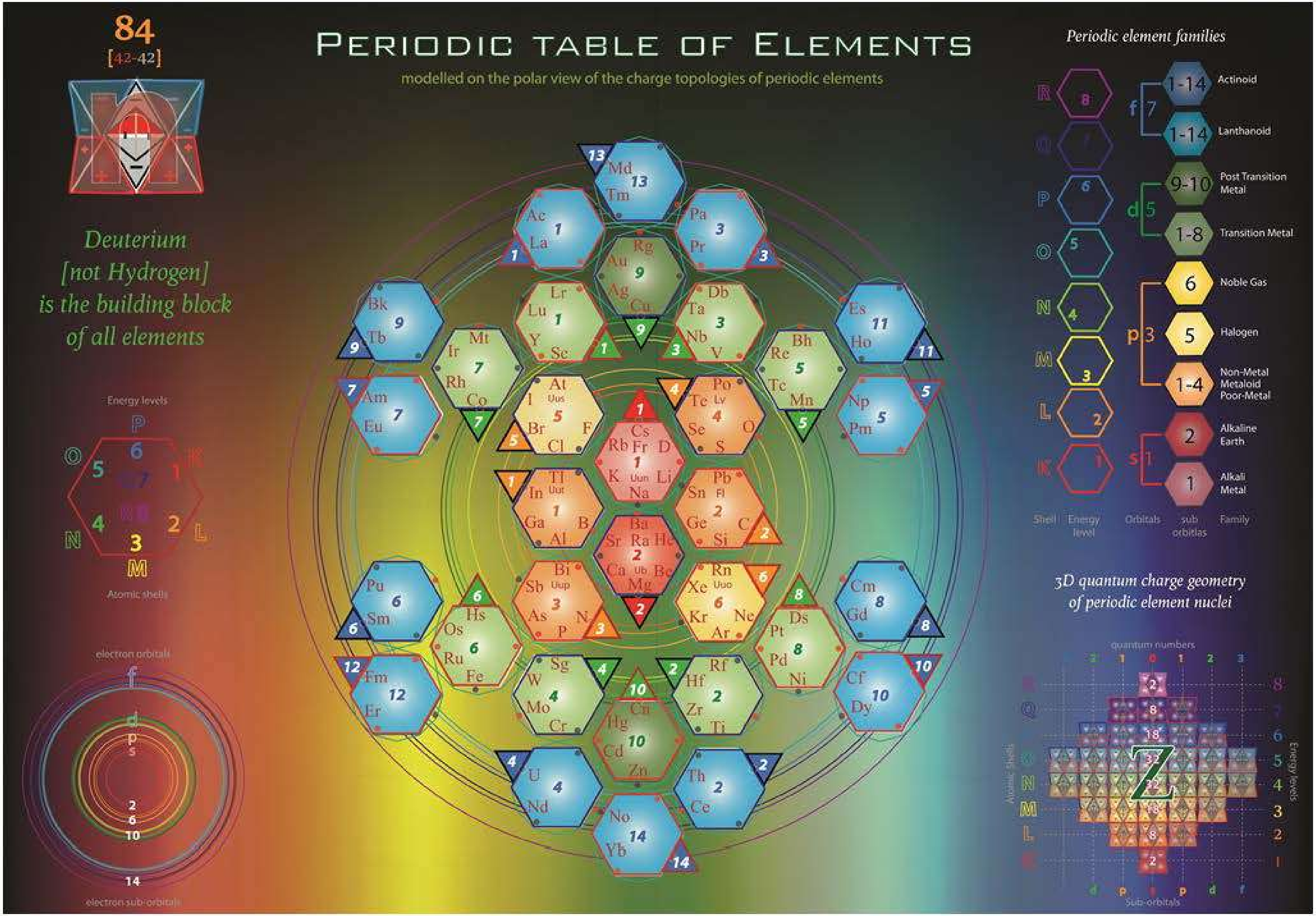

PERIODIC TABLE OF ELEMENTS

modelled on the polar view of the charge topologies of periodic elements

84
[42-42]

Deuterium [not Hydrogen] is the building block of all elements

Periodic element families

3D quantum charge geometry of periodic element nuclei

Energy levels
Atomic shells
electron orbitals
electron sub-orbitals
f
d
p
s
2
6
10
14
K L M N O P
1 2 3 4 5 6 7 8
Md Tm 13
Ac La 1
Pa Pr 3
Rg Au Ag Cu 9
Bk Tb 9
Lr Lu Y Sc 1
Db Ta Nb V 3
Es Ho 11
Mt Ir Rh Co 7
Bh Re Tc Mn 5
Am Eu 7
At I Br Cl F Uus 5
Cs Rb Fr D K Uun Li Na 1
Po Te Se O S Lv 4
Np Pm 5
Tl In Ga B Al Uut 1
Pb Sn Ge C Si Fl 2
Ba Sr Ra He Ca Ub Be Mg 2
Pu Sm 6
Bi Sb As N P Uup 3
Rn Xe Kr Ne Ar Uuo 6
Cm Gd 8
Hs Os Ru Fe 6
Ds Pt Pd Ni 8
Fm Er 12
Sg W Mo Cr 4
Cn Hg Cd Zn 10
Rf Hf Zr Ti 2
Cf Dy 10
U Nd 4
Th Ce 2
No Yb 14
R Q P O N M L K
f 7
d 5
p 3
s 1
1-14 Actinoid
1-14 Lanthanoid
9-10 Post Transition Metal
1-8 Transition Metal
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbitals
Family
quantum numbers
Atomic shells
Energy levels
Sub-orbitals
Z
2 8 18 32 32 18 8 2

Copyright ABRAHAM [2008] - All rights reserved

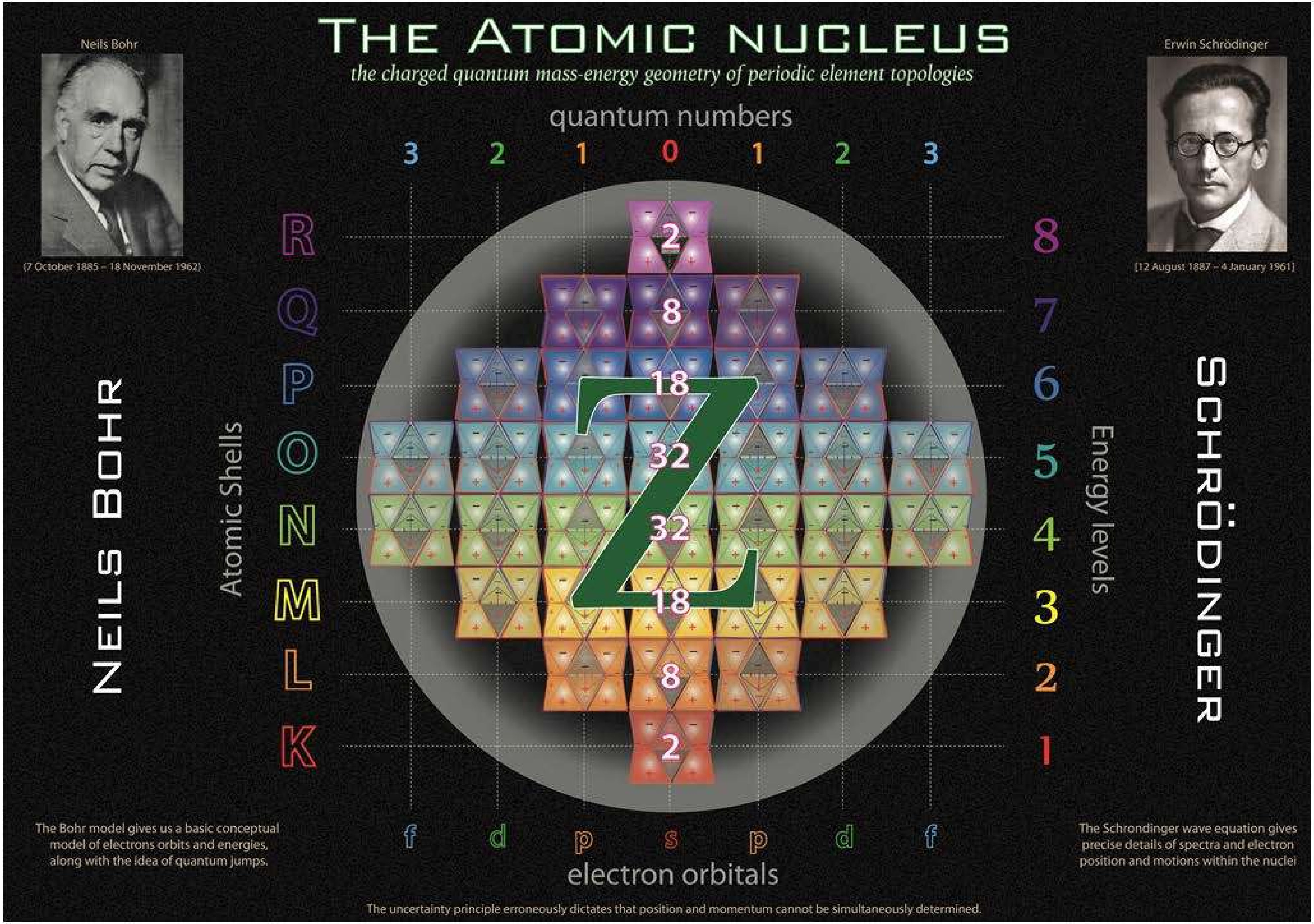

THE ATOMIC NUCLEUS

the charged quantum mass-energy geometry of periodic element topologies

Neils Bohr

(7 October 1885 – 18 November 1962)

Erwin Schrödinger

[12 August 1887 – 4 January 1961]

quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Atomic Shells
NEILS BOHR
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
Energy levels
SCHRÖDINGER
f
d
p
s
p
d
f
electron orbitals

The Bohr model gives us a basic conceptual model of electrons orbits and energies, along with the idea of quantum jumps.

The Schrondinger wave equation gives precise details of spectra and electron position and motions within the nuclei

The uncertainty principle erroneously dictates that position and momentum cannot be simultaneously determined.

Copyright ABRAHAM [2008] - All rights reserved

Periodicity of atomic elements

An electron shell may be thought of as an orbit followed by electrons around an atom's nucleus.
The closest shell to the nucleus is called the "1 shell" (also called "K shell"), followed by the
"2 shell" (or "L shell"), then the "3 shell" (or "M shell"), and so on further and further from the nucleus.
The shell letters K, L, M, ... are alphabetical

Each shell can contain only a
fixed number of deuterium nuclei
[Protons, Neutrons & electrons]

Each shell consists of one or more
electron orbitals, and each orbital consists
of one or more sub-orbitals.

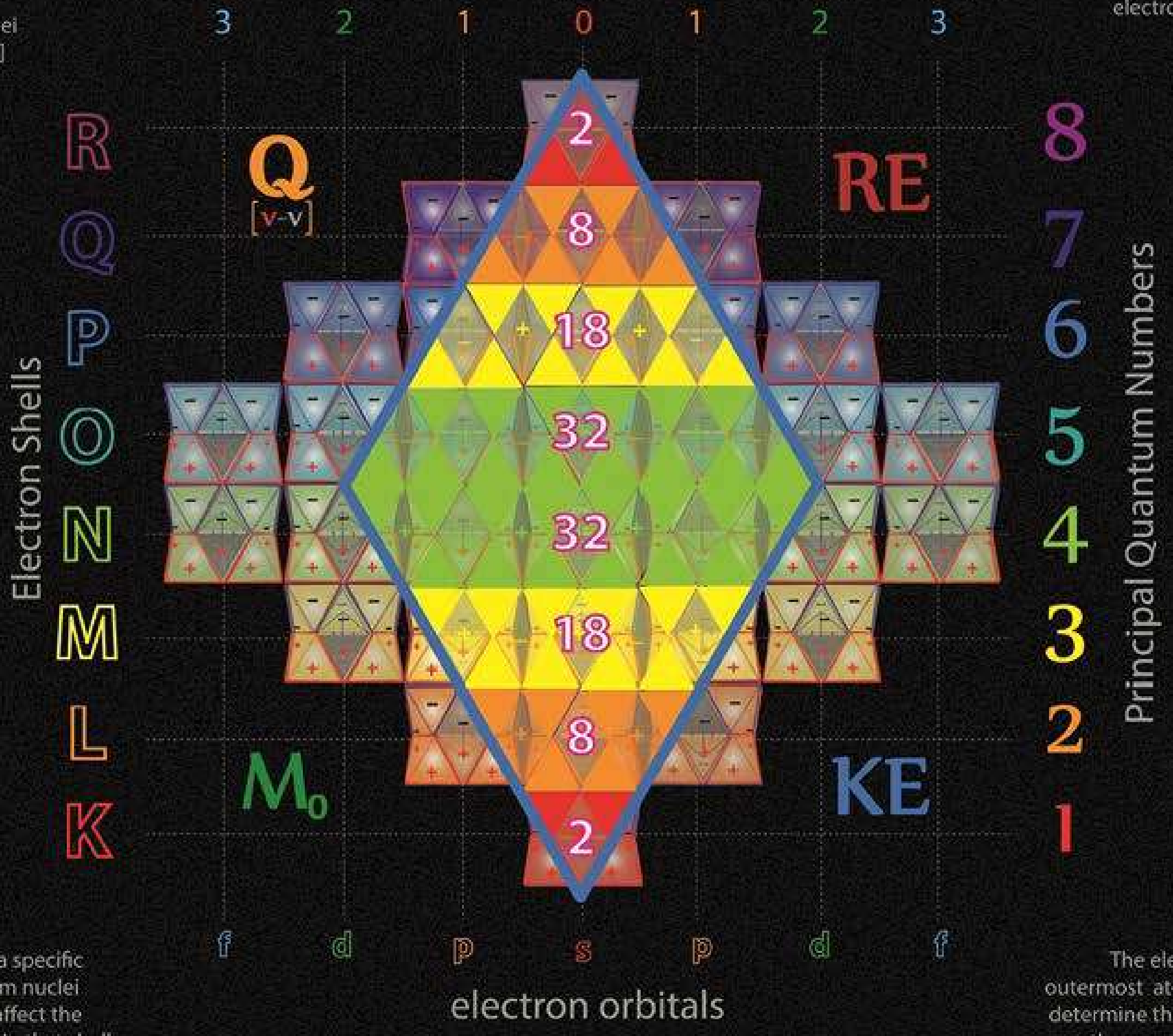

8
atomic
energy shells
with electrons

120
periodic
elements
in total

The number of deuterium nuclei that form each atomic shell in an elemental atom number is double that of Photon energy levels

Each atomic shell equates to a specific
energy level for the Deuterium nuclei
that comprise it with in turn affect the
angular momentum of electrons in that shell

The electrons in the partially filled
outermost atomic shell [or highest energy level]
determine the chemical properties of the atom;
they are called valence shell electrons.

Tetryonics 44.02 - Periodicity of atomic elements

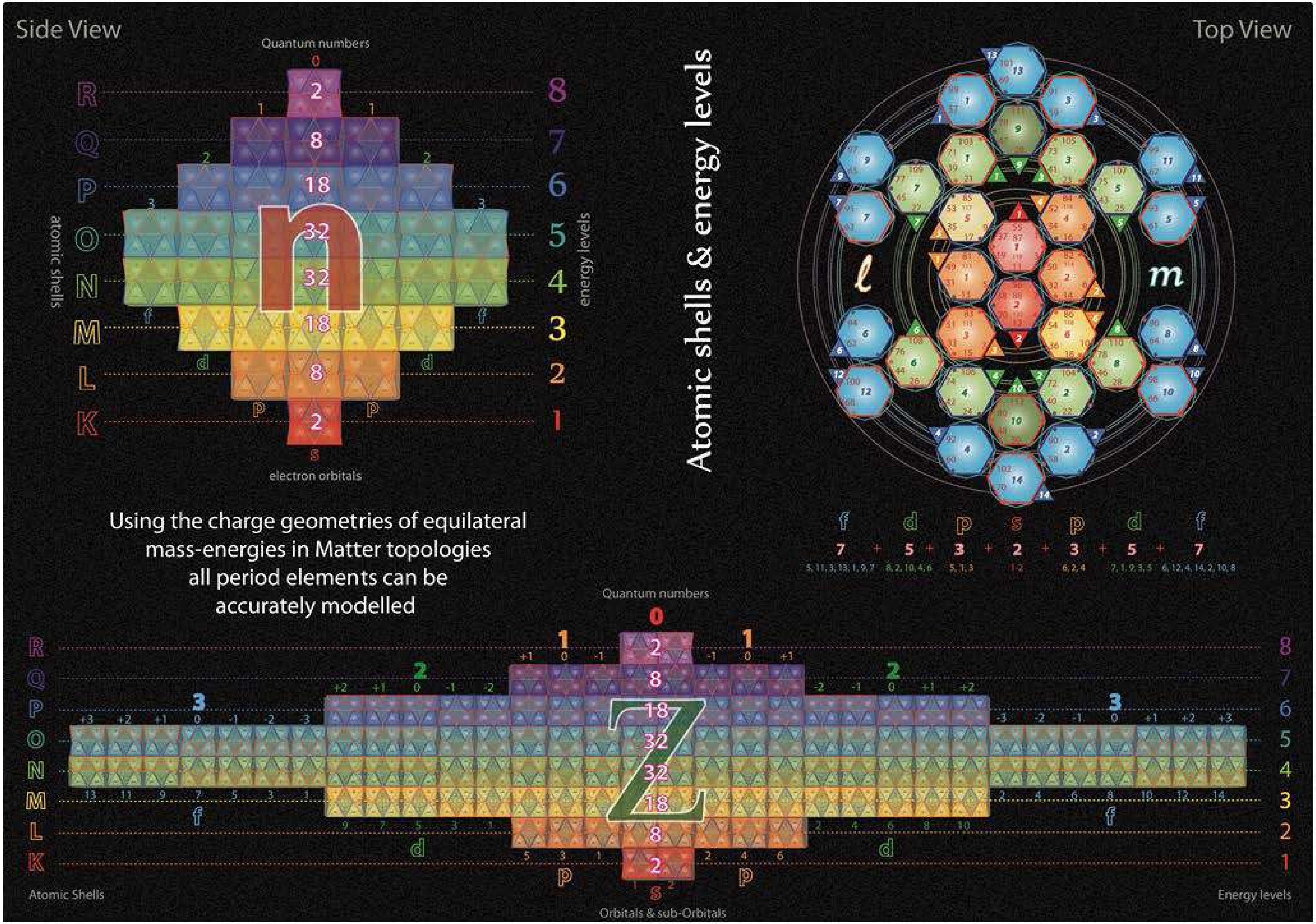
Side View
Quantum numbers
0
1
2
3
R
Q
P
O
N
M
L
K
atomic shells
2
8
18
32
32
18
8
2
n
f
d
p
s
electron orbitals
8
7
6
5
4
3
2
1
energy levels
Atomic shells & energy levels
Top View
l
m
f d p s p d f
7 + 5 + 3 + 2 + 3 + 5 + 7
5, 11, 3, 13, 1, 9, 7
8, 2, 10, 4, 6
5, 1, 3
1-2
6, 2, 4
7, 1, 9, 3, 5
6, 12, 4, 14, 2, 10, 8
Quantum numbers
Z
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Using the charge geometries of equilateral mass-energies in Matter topologies all period elements can be accurately modelled atomic Shells

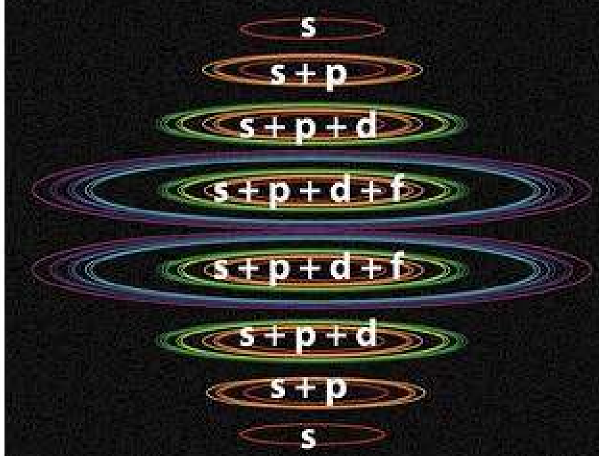
s
s + p
s + p + d
s + p + d + f
s + p + d + f
s + p + d
s + p
s

energy levels

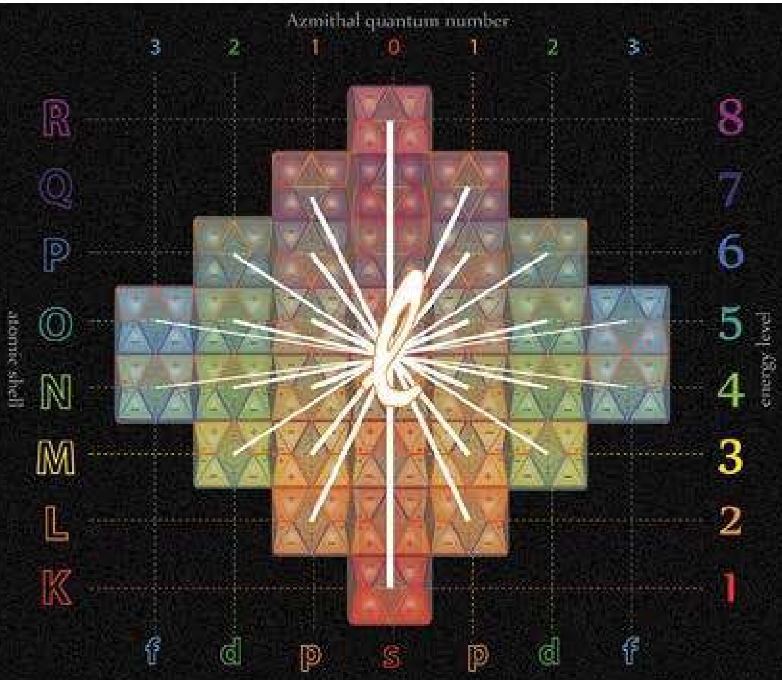
Azmithal quantum number
3 2 1 0 1 2 3
R Q P O N M L K
atomic shell
8 7 6 5 4 3 2 1
energy level
ℓ
f d p s p d f

Electron orbitals

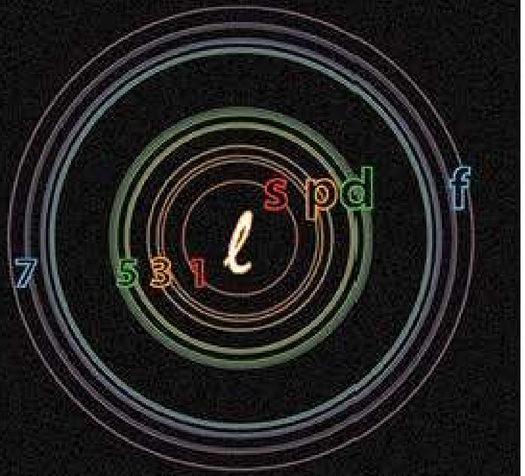

quantum numbers

Each electron's location in
an elemental nuclei is defined
by its quantum numbers

All atoms have shells comprised of concentric s, p, d & f electron orbitals

electron orbitals

each electron orbital can hold a maximum of two spin opposed electrons

a core s-orbital
surrounded by
3 p-orbitals
5 d-orbitals and
7 f-orbitals

Azmithal & Magnetic numbers

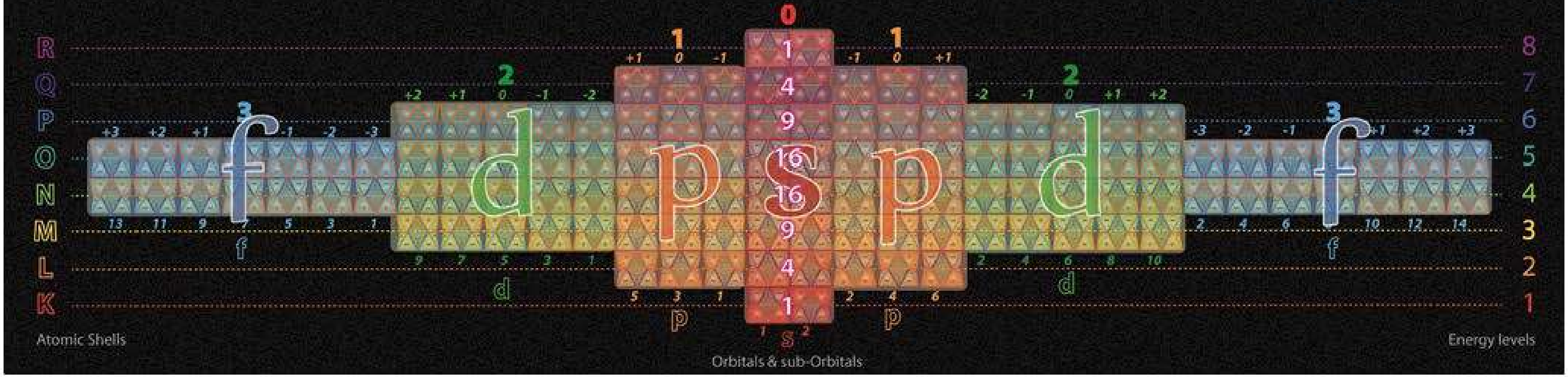
0
1 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2 -2 -1 0 +1 +2
3 3
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 +1 +2 +3
1 4 9 16 16 9 4 1
f d p s p d f
13 11 9 7 5 3 1 2 4 6 10 12 14
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
R Q P O N M L K
8 7 6 5 4 3 2 1

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

magnetic quantum numbers

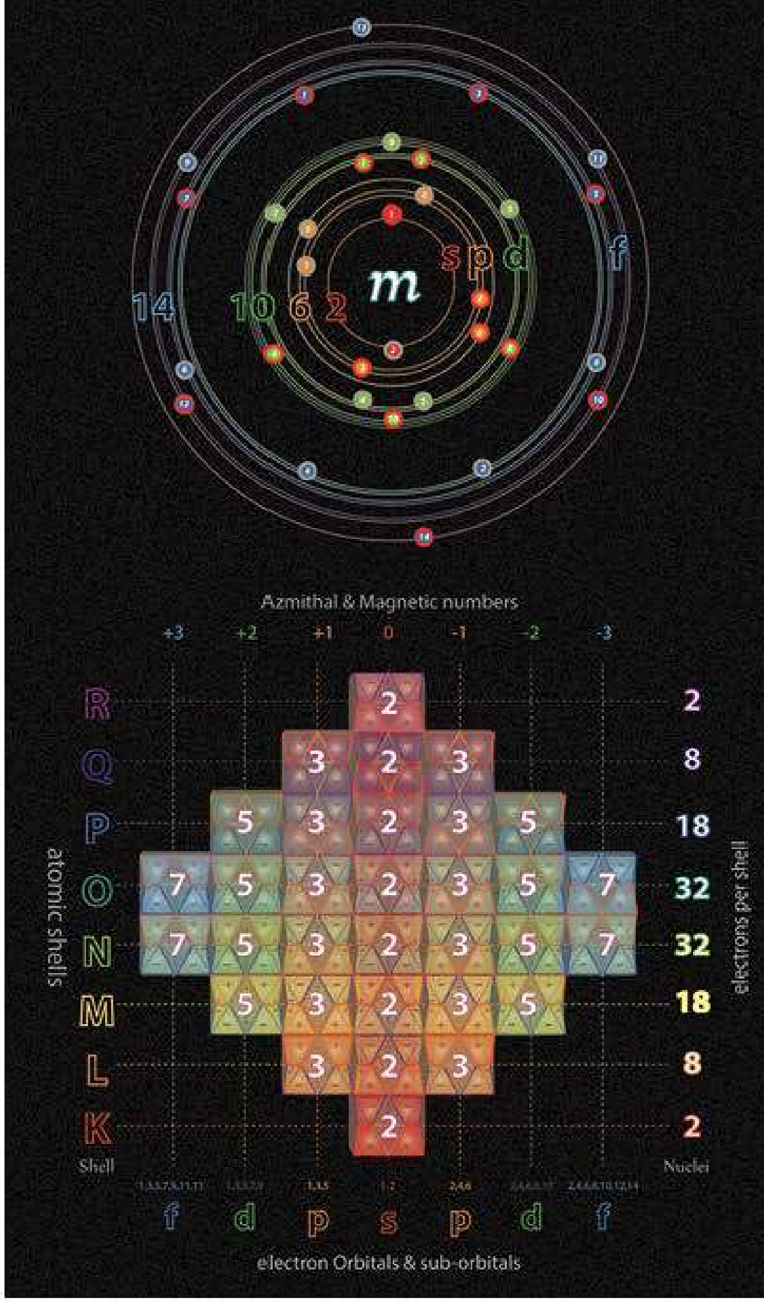
m
s p d f
2
6
10
14
Azmithal & Magnetic numbers
+3 +2 +1 0 -1 -2 -3
R Q P O N M L K
2 8 18 32 32 18 8 2
atomic shells
electrons per shell
Shell
Nuclei
f d p s p d f
electron Orbitals & sub-orbitals

electron sub-orbitals

each eatomic orbital can hold two spin opposed electrons

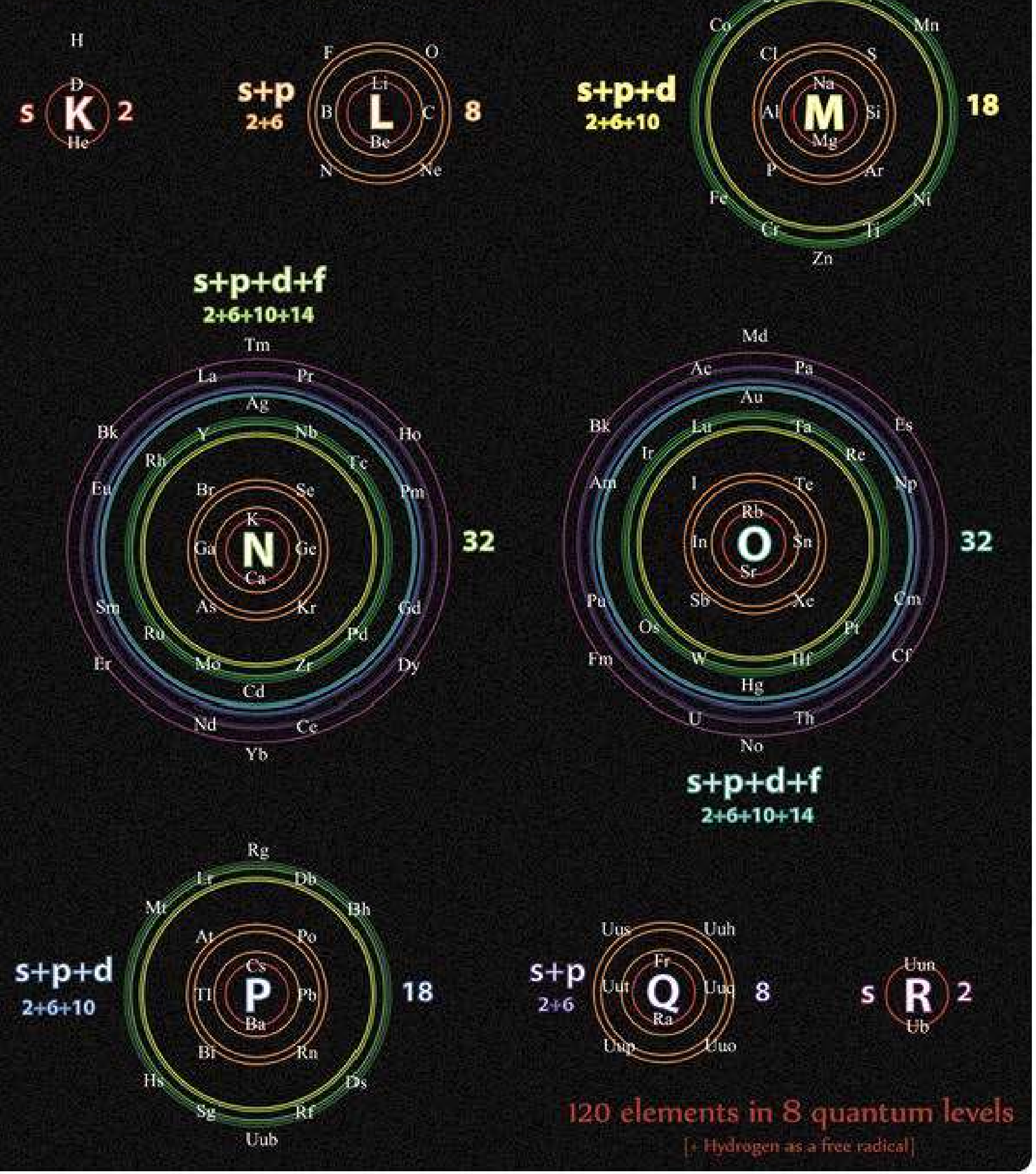
s K 2
H D He
s+p 2+6 L 8
Li Be B C N O F Ne
s+p+d 2+6+10 M 18
Na Mg Al Si P S Cl Ar Cu V Mn Ni Ti Zn Cr Fe Co Sc
s+p+d+f 2+6+10+14 N 32
s+p+d+f 2+6+10+14 O 32
s+p+d 2+6+10 P 18
s+p 2+6 Q 8
s R 2
Uun Ub
120 elements in 8 quantum levels
[+ Hydrogen as a free radical]

Copyright ABRAHAM [2008] - All rights reserved

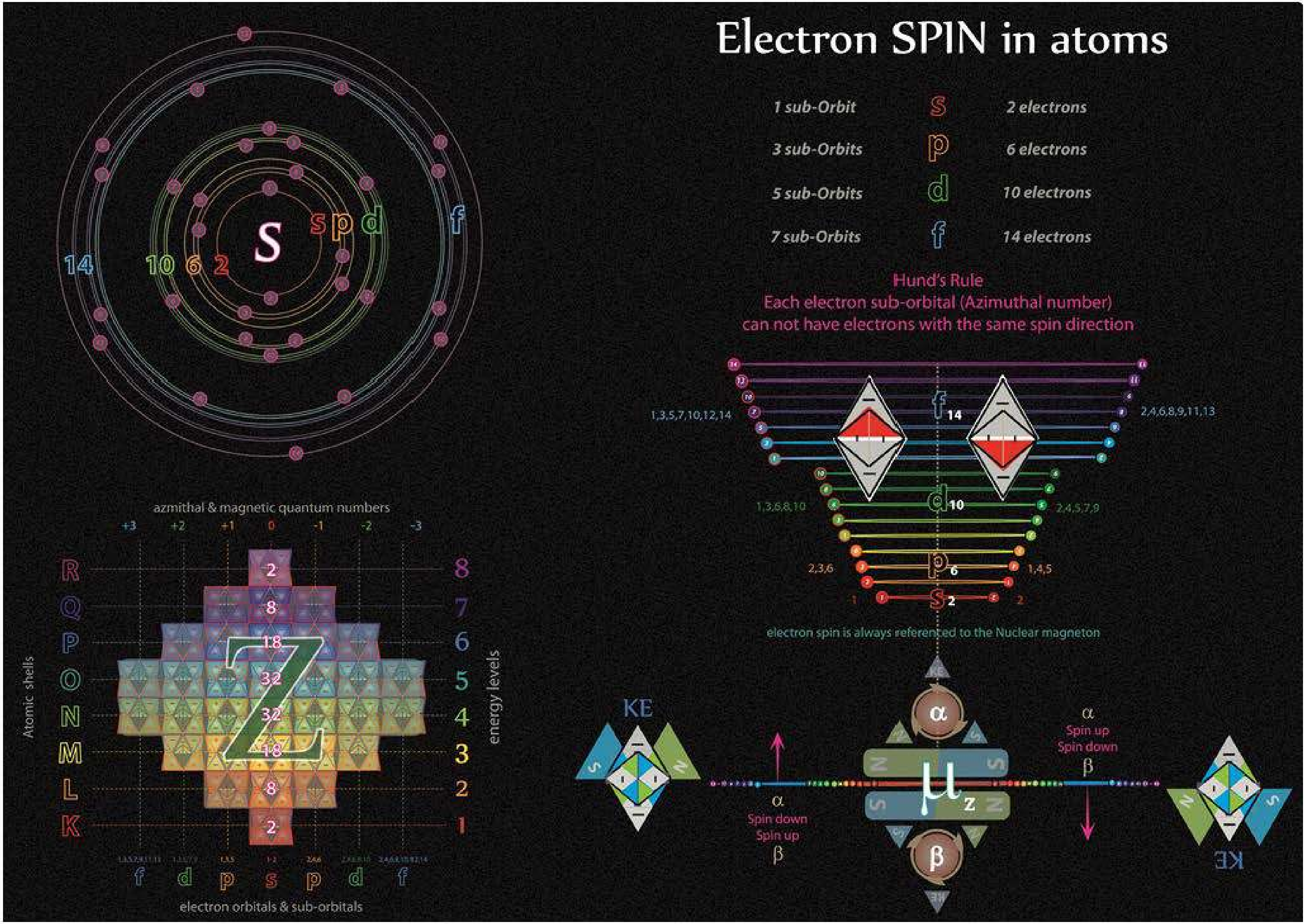

Electron SPIN in atoms

1 sub-Orbit	s	2 electrons
3 sub-Orbits	p	6 electrons
5 sub-Orbits	d	10 electrons
7 sub-Orbits	f	14 electrons

Hund's Rule
Each electron sub-orbital (Azimuthal number)
can not have electrons with the same spin direction

electron spin is always referenced to the Nuclear magneton

S
s p d
f
14
10 6 2
azmithal & magnetic quantum numbers
+3 +2 +1 0 -1 -2 -3
R Q P O N M L K
Atomic shells
2 8 18 32 32 18 8 2
Z
8 7 6 5 4 3 2 1
energy levels
f d p s p d f
electron orbitals & sub-orbitals
1,3,5,7,10,12,14
f 14
2,4,6,8,9,11,13
1,3,6,8,10
d 10
2,4,5,7,9
2,3,6
p 6
1,4,5
1
s 2
2
KE
α
Spin down
Spin up
β
α
Spin up
Spin down
β
μ z
N S
KE

Tetryonics 44.06 - Electron spin

Quantum numbers

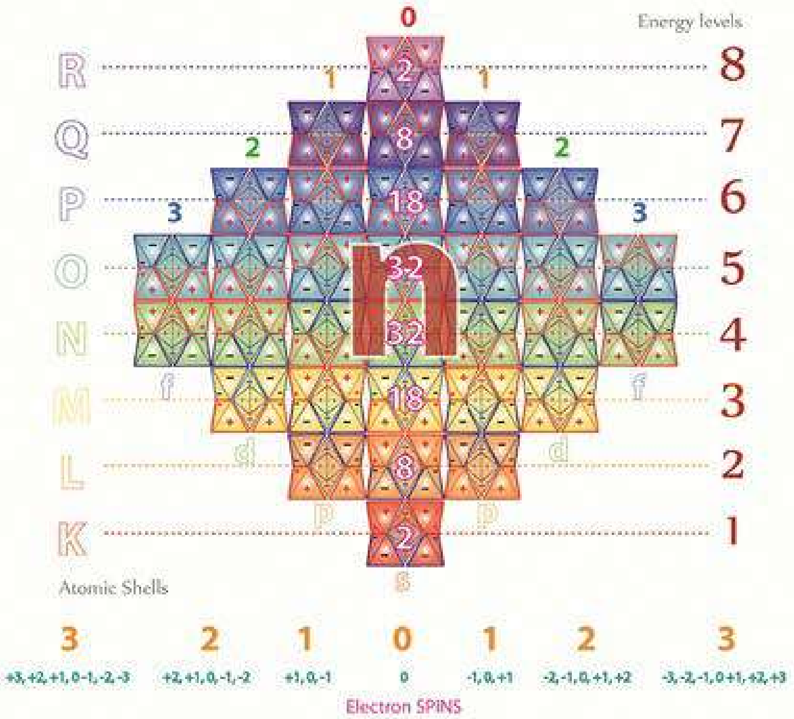
0
Energy levels
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
1
2
2
3
3
2
8
18
32
n
32
18
8
2
f
f
d
d
p
p
s
Atomic Shells
3
2
1
0
1
2
3
+3, +2, +1, 0 -1, -2, -3
+2, +1, 0, -1, -2
+1, 0, -1
0
-1, 0, +1
-2, -1, 0, +1, +2
-3, -2, -1, 0 +1, +2, +3
Electron SPINS

These four numbers, n, ℓ, m and s can be used to describe any electron in a stable atom [and can be mapped back to the classical terminology of Shells and electron orbitals].

The properties of every atom's specific electron configuration can be described by four quantum numbers:

n (1-8)
Principal
(n = 1, 2, 3, 4 ...)

ℓ (0-3)
Azimuthal
(l = 0, 1 ... n−1)

m (2ℓ+1)
Magnetic
(ml = −l, −l+1 ... 0 ... l−1, l)

s ±½
electron Spin
(ms = −1/2 or +1/2).

The Bohr model was a one-dimensional model that used one quantum number to describe the distribution of electrons

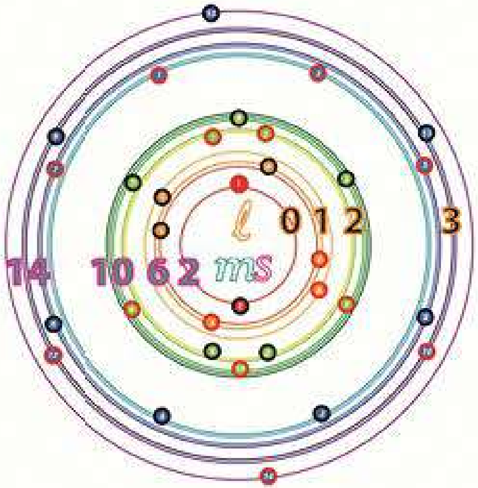
ℓ
0
1
2
3
14
10
6
2
ms

The three coordinates that come from Schrödinger's wave equations are the principal (n), angular (l), and magnetic (m) quantum numbers.

These quantum numbers describe the size, shape, and orientation in space of the orbitals of any particular atom mathematically.

Each electron's quantum numbers are unique and cannot be shared by another electron in that atom.

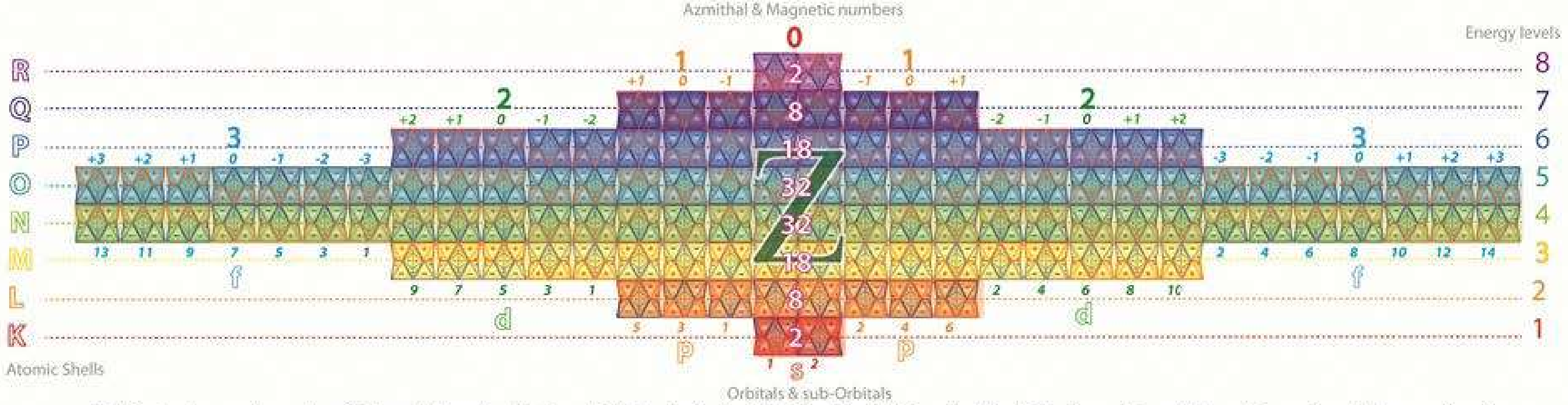
Azmithal & Magnetic numbers
Energy levels
0
1
1
2
2
3
3
+1
0
-1
-1
0
+1
+2
+1
0
-1
-2
-2
-1
0
+1
+2
+3
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+1
+2
+3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
Z
32
18
8
2
13
11
9
7
5
3
1
2
4
6
8
10
12
14
f
f
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Orbitals & sub-Orbitals

A stable atom has equal numebrs of Protons, electrons [and Neutrons], all following the Pauli Exclusion Principle thus orientating their spins so that each element has a unique electron configuration

Copyright ABRAHAM [2008] - All rights reserved

PRINCIPAL quantum number

The first describes the electron shell, or energy level, of an atom.

This is the only quantum number
introduced by the Bohr model

n(1-8)

(n = 1, 2, 3, 4 ...)

The principal quantum number
can only have positive integer values

atomic shells

energy levels

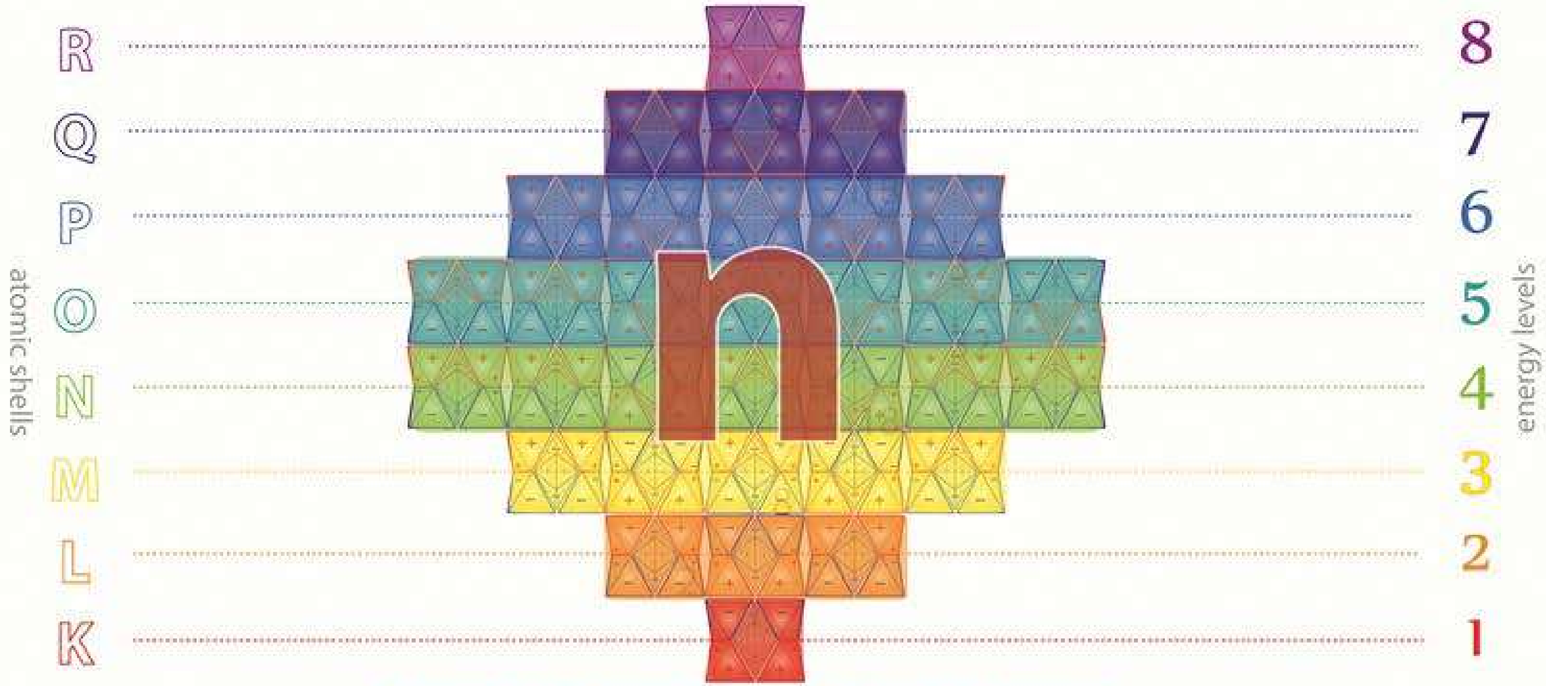

As energies of the Baryons comprising the atomic nuclei increases,
the electron bound to each nuclei also possesses more KEM field energies
and is therefore less tightly bound to the nucleus

Atomic Shells

(K, L, M, N, O, P, Q, R)

Atomic shells relate directly to Principal quantum numbers

(1, 2, 3, 4, 5, 6, 7, 8)

Energy levels

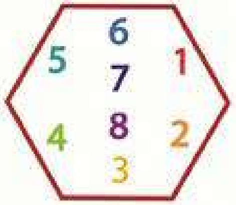

AZMITHAL quantum number

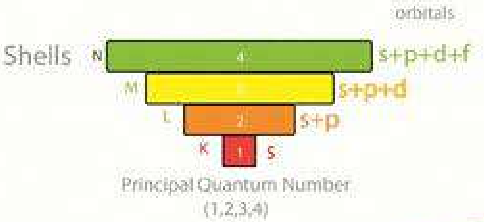
orbitals
Shells
N
4
s+p+d+f
M
3
s+p+d
L
2
s+p
K
1
s
Principal Quantum Number
(1,2,3,4)

The azimuthal quantum number is a quantum number assigned to any atomic orbital that describes its orbital angular momentum and determines the shape of the electron orbital

s p d f

electron orbitals

(l = 0, 1 ... n−1)

ℓ (0-7)

Azimuthal

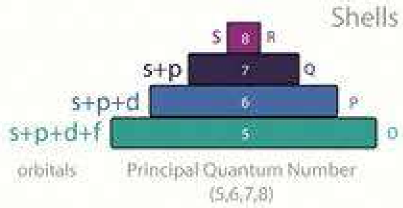
Shells
S
8
R
s+p
7
Q
s+p+d
6
P
s+p+d+f
5
O
orbitals
Principal Quantum Number
(5,6,7,8)

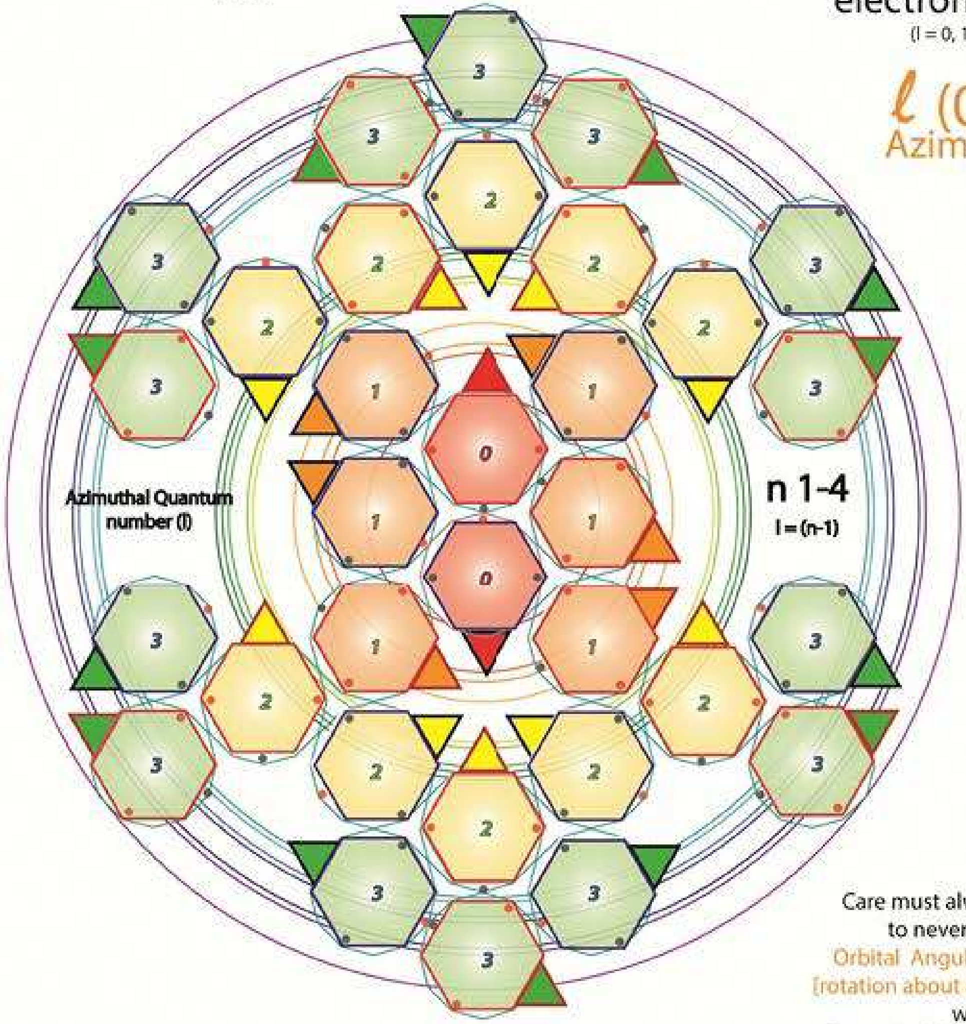
Azimuthal Quantum number (l)
n 1-4
l = (n-1)
0
1
2
3

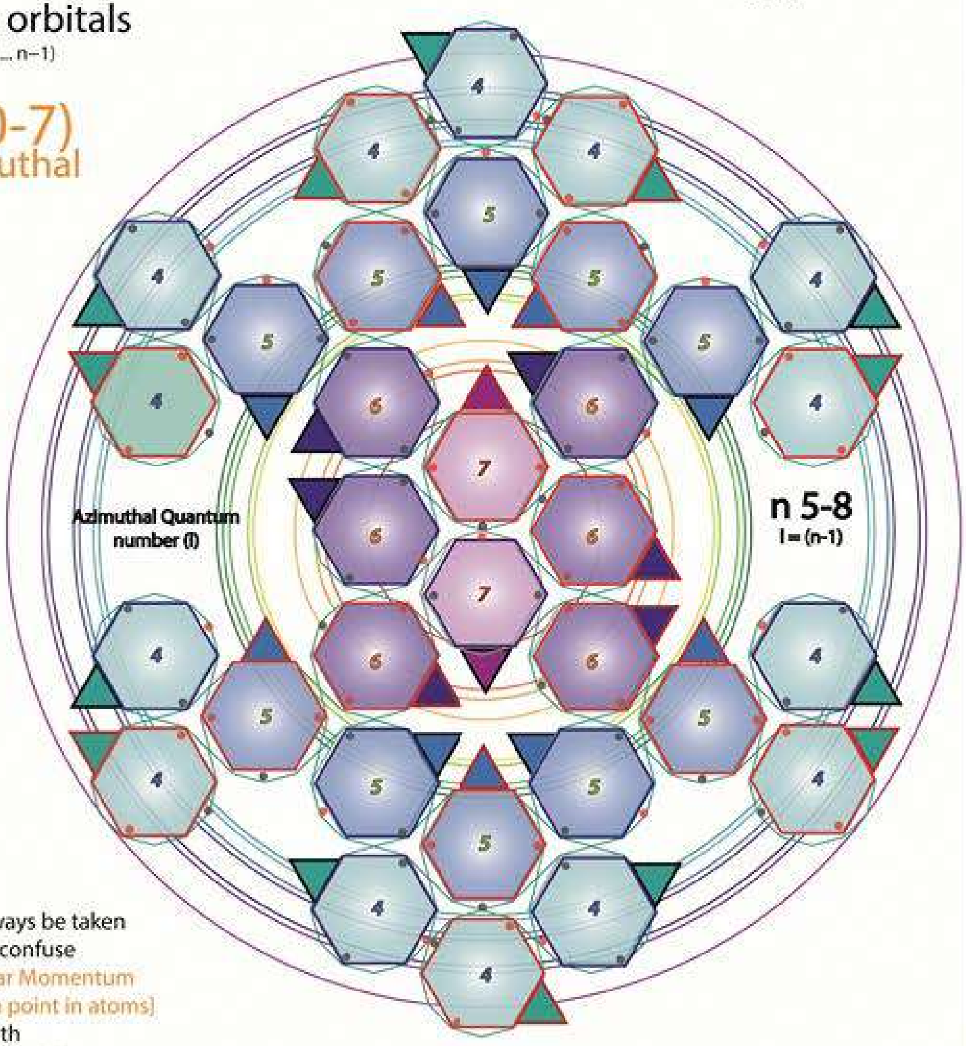
Azimuthal Quantum number (l)
n 5-8
l = (n-1)
4
5
6
7

Care must always be taken
to never confuse
Orbital Angular Momentum
[rotation about a point in atoms]
with
Quantised Angular Momenta
[equilateral Planck energy geometries]

MAGNETIC quantum number

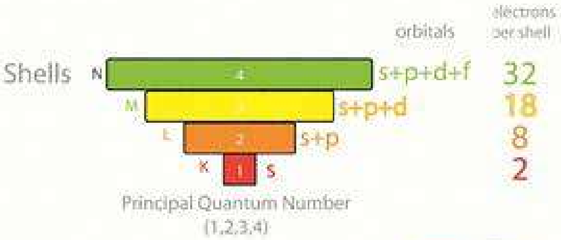
Shells
orbitals
electrons per shell
N
M
L
K
4
3
2
1
s+p+d+f
s+p+d
s+p
s
32
18
8
2
Principal Quantum Number
(1,2,3,4)

The magnetic quantum number denotes the energy levels available within any subshell

Magnetic numbers do not continue to increase as the Principal numbers increase
instead they reverse after n4 to reflect the charged quanutm geometry of Elements
and do not follow the current computer models in popular use

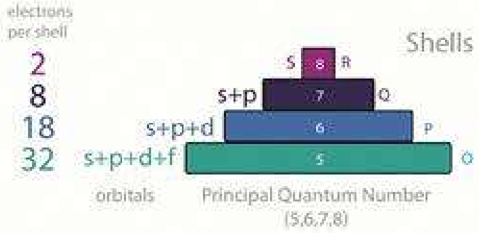
electrons per shell
Shells
2
8
18
32
S
8
R
s+p
7
Q
s+p+d
6
P
s+p+d+f
5
O
orbitals
Principal Quantum Number
(5,6,7,8)

s p d f

electron sub-orbitals

(ml = −l, −l+1 ... 0 ... l−1, l)

$m\ell(2\ell+1)$

Magnetic

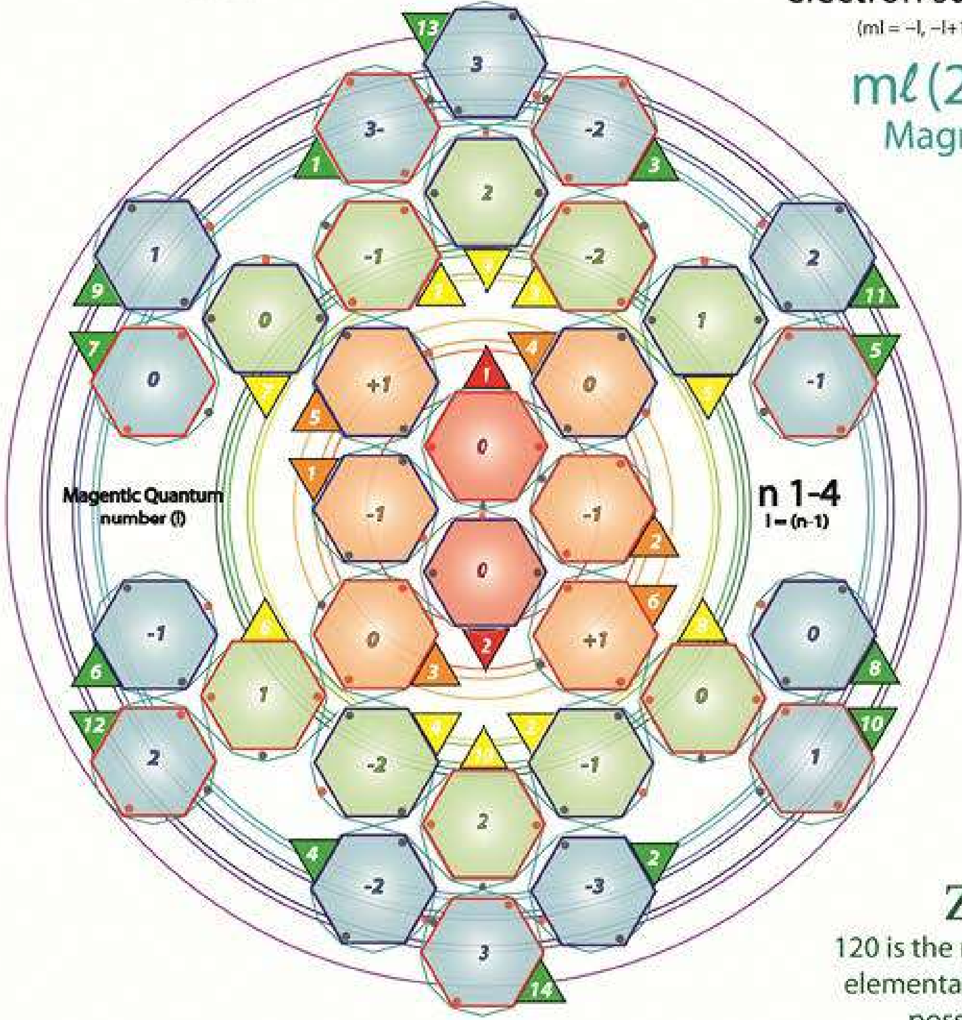
Magentic Quantum number (l)
n 1-4
l = (n-1)

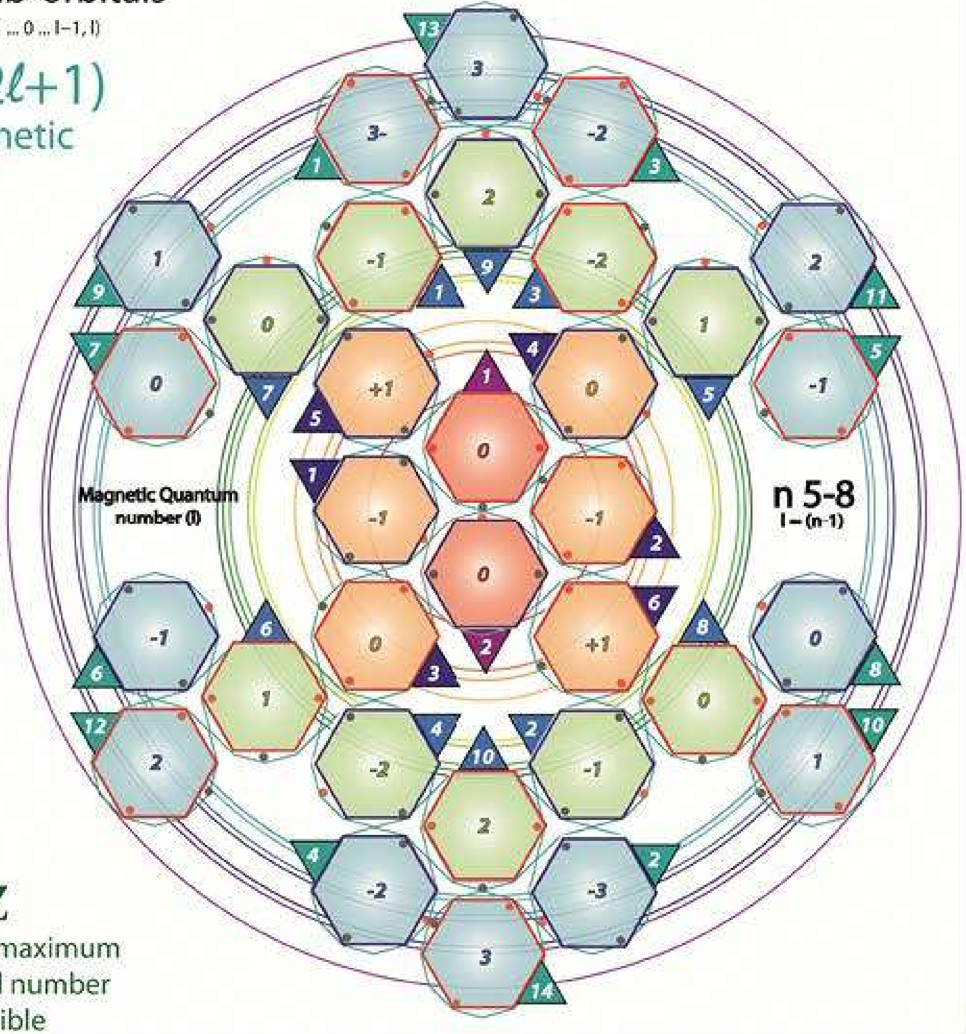
Magentic Quantum number (l)
n 5-8
l = (n-1)

Z

120 is the maximum elemental number possible

Copyright ABRAHAM [2008] - All rights reserved

SPIN quantum number

The spin quantum number is a quantum number that parameterizes
the intrinsic angular momentum (or spin angular momentum),
of any given electron anywhere in an atomic nucleus

parallel magnetic moments

Spin UP

higher coupling energies

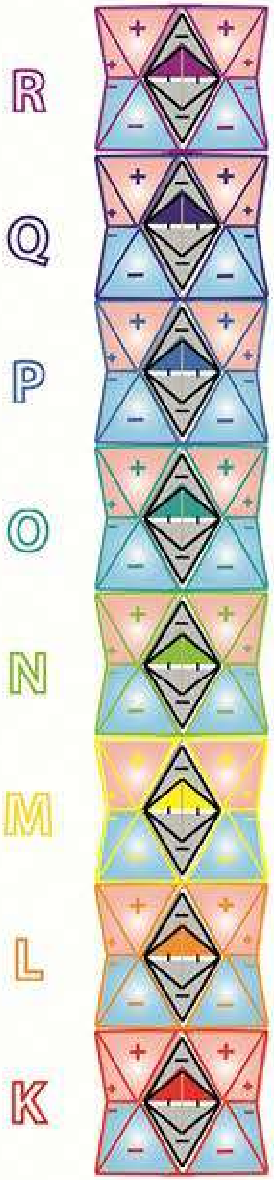

**Electron spin can orientate in either direction within Nuclei,
providing the nett spins follow the Hund rule and Pauli exclusion principle**

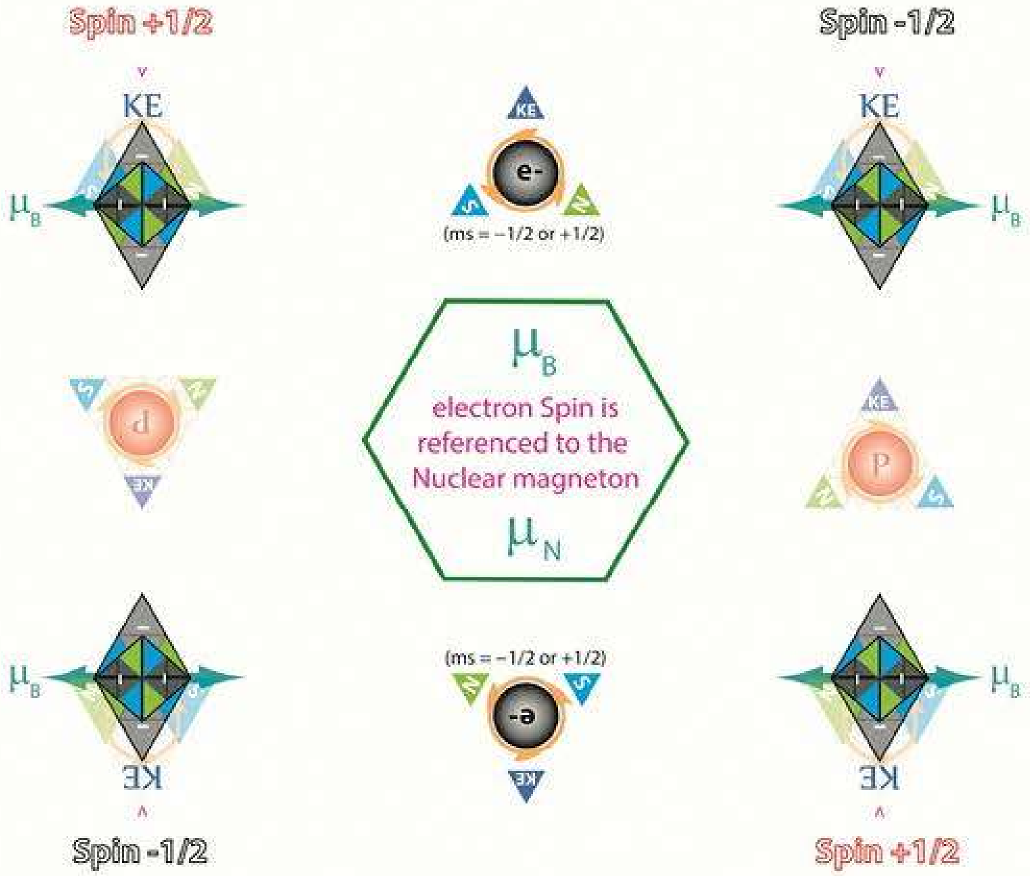
Spin +1/2
KE
μ_B
KE
e-
(ms = −1/2 or +1/2)
Spin -1/2
KE
μ_B
d
KE
p
μ_B
KE
Spin -1/2
(ms = −1/2 or +1/2)
-e
KE
μ_B
KE
Spin +1/2

μ_B

electron Spin is
referenced to the
Nuclear magneton

μ_N

anti-parallel magnetic moments

Spin DOWN

lower coupling energies

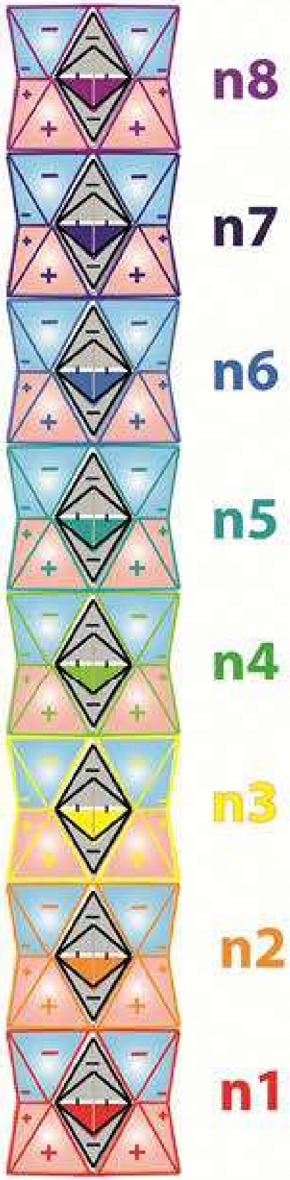
n8
n7
n6
n5
n4
n3
n2
n1
−
+

**The nuclear energy levels of the Baryons comprising Elemental nuclei
determine the energy-momenta of electrons bound to them**

Copyright ABRAHAM [2008] - All rights reserved

Modifying Hund's Rule

Electrons fill orbitals in an alternating sequential numbering pattern due to nucleon placement creating opposed direction electron spins

The increased stability of the atom, most commonly manifested in a lower energy state, arises because the high-spin state forces the unpaired electrons to reside in different spatial orbitals.

A commonly given reason for the increased stability of high multiplicity states is that the different occupied spatial orbitals create a larger average distance between electrons, reducing electron-electron repulsion energy. In reality, it has been shown that the actual reason behind the increased stability is a decrease in the screening of electron-nuclear attractions[1].

The total spin state is calculated as the total number of unpaired electrons + 1, or twice the total spin + 1 written as 2s+1.

As a result of Hund's rule, constraints are placed on the way atomic orbitals are filled using the Aufbau principle.

Before any two electrons occupy an orbital in a subshell, other orbitals in the same subshell must first each contain one electron. Also, the electrons filling a subshell will have parallel spin before the shell starts filling up with the opposite spin electrons (after the first orbital gains a second electron).

As a result, when filling up atomic orbitals, the maximum number of unpaired electrons (and hence maximum total spin state) is assured

Sub-orbitals fill in order of numbering

Electrons spins pair
before next orbital
is filled

ie.
p1 [DOWN] and p2 [UP] fill
before
p3 [UP] and p4 [DOWN]
before
p5 [DOWN] and p6 [UP] etc

s	p	d	f
▲1 1 2▼	▼1 1 2▲	▲1 1 2▼	▲ 1 1 2 ▼
	▲3 2 4▼	▲3 2 4▼	▲ 3 2 4 ▼
	▼5 3 6▲	▼5 3 6▲	▲ 5 3 6 ▼
		▼7 4 8▲	▲ 7 4 8 ▼
		▼9 5 10▲	▼ 9 5 10 ▲
			▼ 11 6 12 ▲
			▼ 13 7 14 ▲

▲

Hund's rule of orbital filling
must be modified to reflect
the true orbital filling order

▼

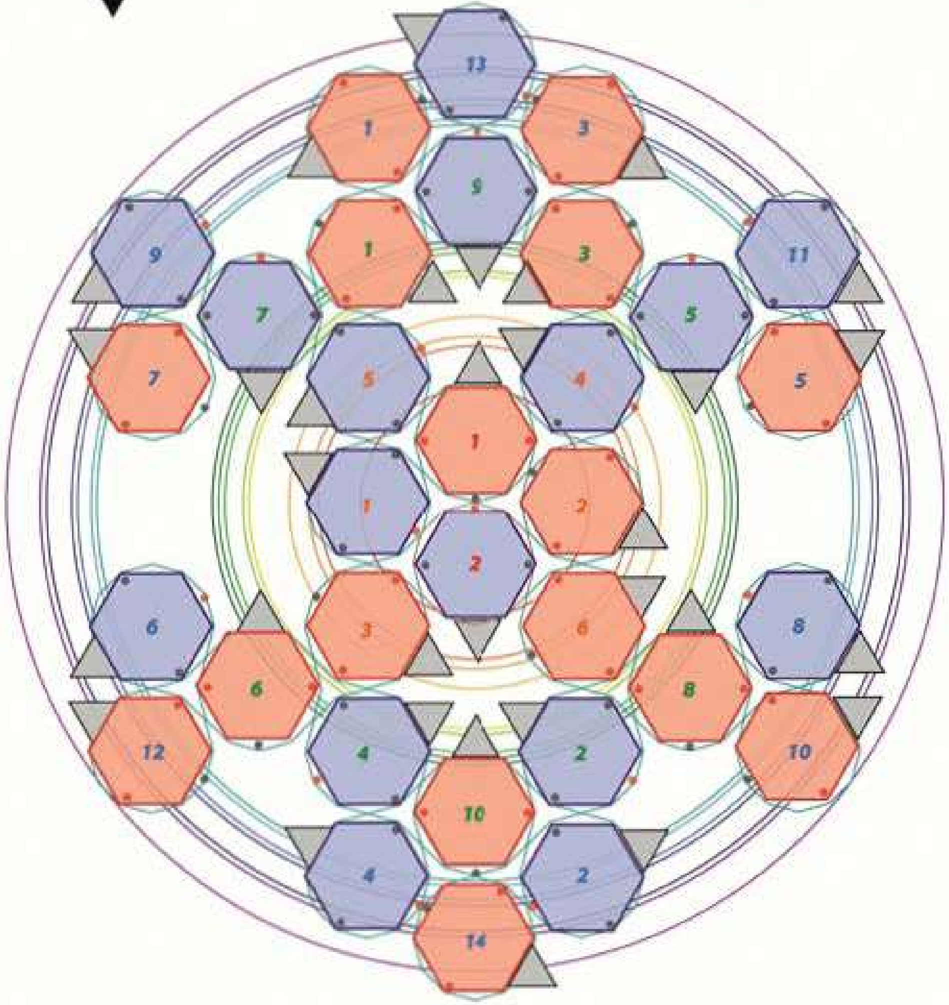

Copyright ABRAHAM [2008] - All rights reserved

Principle quantum Energies

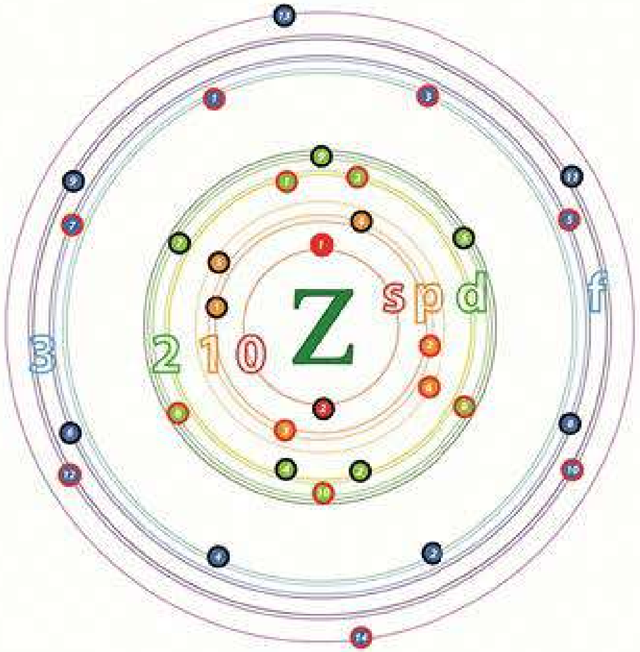

In an atom - electron energies are proportional to their intrinsic
Kinetic Energies - which in turn are directly proportional to the
quantum energy level of the nuclei which the electron binds to
in their respective atomic shells

In a nucleus, lower energy orbits have less 'paired' nuclei supplying energy.
The more energy you give a nuclei the faster it casuses the bound electron to rotate.
If you give the nuclei enough energy, it will impart eneough energy to
its electron for it to leave the system entirely.

The same is true for an electron orbital.
Higher values of n mean more energy for the electron and the
corresponding KEM field enegies of the electron is larger,
resulting in increased angular momentum.

Values of n start at 1 and go up by integer amounts.

If enough energy is added to the system by incident Photons
a electron will leave the atom creating a positively charged nuclei
[ionisation].

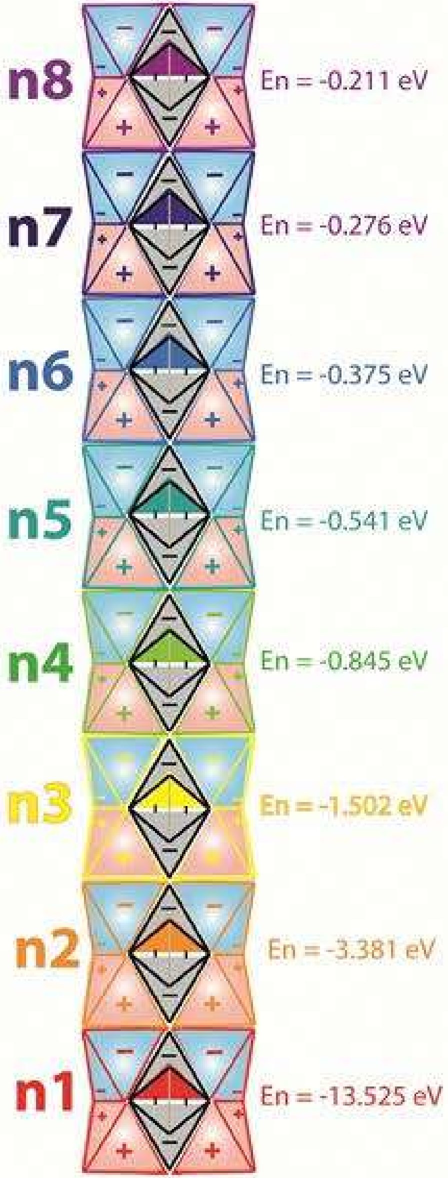

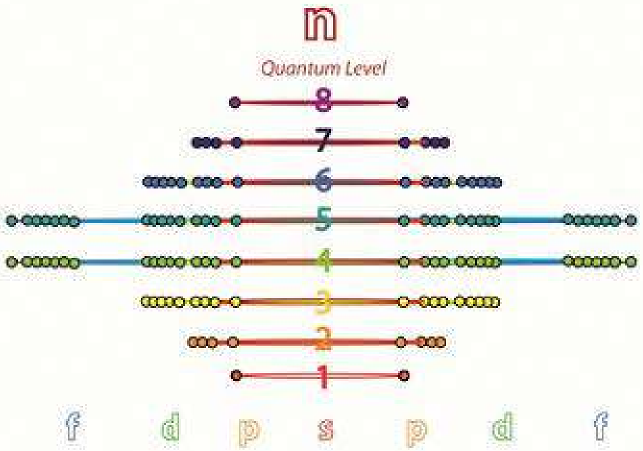

s orbits	n1-8
p orbits	n2-7
d orbits	n3-6
f orbits	n4-5

Eigenstate value

KEM field energy [per n] required
to exceed 13.525 eV at which
point the photo-electron has
sufficient KE to break free
of the Nucleus

$$E_n = \frac{E_1}{n^2} = \frac{-13.6eV}{n^2}, n = 1, 2, 3...$$

**The possible Kinetic Energies (quantum levels)
of an electron are directly related to
the energy level of the Nuclei
in each Quantum Level**

Quantum Level 1

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
1	1s1	Deuterium	1	1	1
2	1s2	Helium	2	2	2

2

Deuterium [not Hydrogen] is the building block of elements

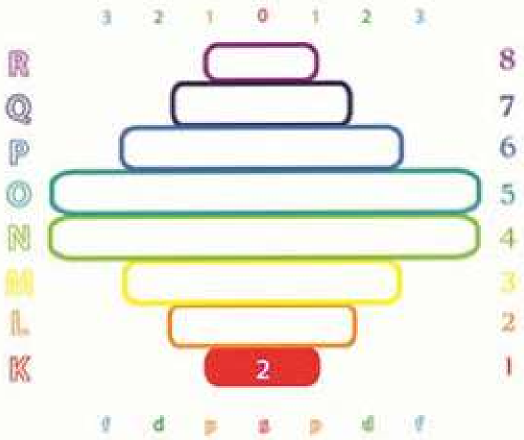
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
f
d
p
s
p
d
f

Z #		Name	Protons	electrons	Neutrons
1	1s1	Hydrogen	1	1	0

Hydrogen is a free radical element

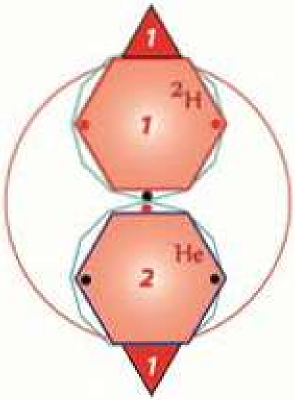

K shell

n1

Ground State electron

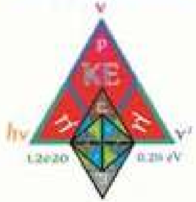
v
p
KE
hv
v'
1.2e20

$iE = -13.313\ eV$

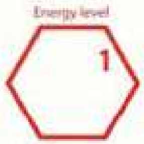
Energy level
1

n1

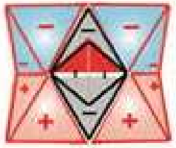

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 2

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
3	2s1	Lithium	3	3	3
4	2s2	Beryllium	4	4	4
5	1p1	Boron	5	5	5
6	1p2	Carbon	6	6	6
7	1p3	Nitrogen	7	7	7
8	1p4	Oxygen	8	8	8
9	1p5	Fluorine	9	9	9
10	1p6	Neon	10	10	10

8

Deuterium [not Hydrogen] is the building block of elements

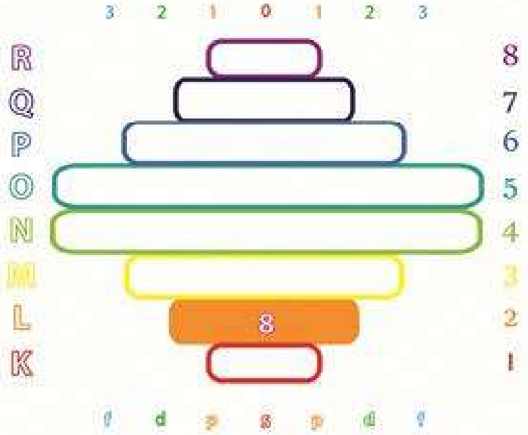

3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
8
f
d
p
s
p
d
f

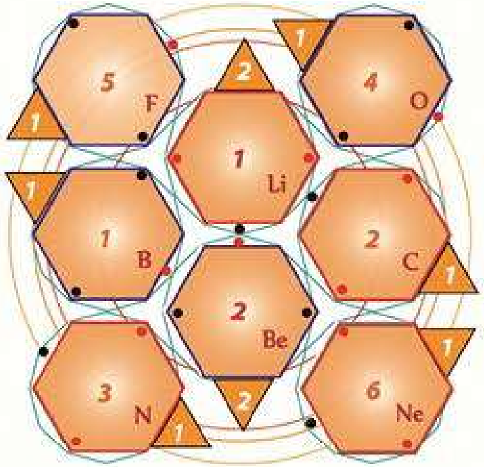

5
F
4
O
1
Li
2
1
1
1
B
2
C
2
Be
3
N
6
Ne
1
1
1
2
1

L shell

n2

Ground State electron

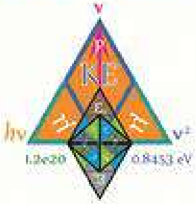

v
KE
hν
v²
1.2e20
0.8453 eV

$iE = -12.679\ eV$

Energy level
2

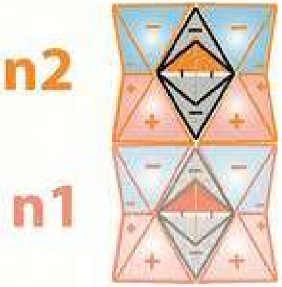

Quantum Level 3

The energy levels of bound electrons is determined by Baryons

18

Z #		Name	Protons	electrons	Neutrons
11	3s1	Sodium	11	11	11
12	3s2	Magnesium	12	12	12
13	2p1	Aluminium	13	13	13
14	2p2	Silicon	14	14	14
15	2p3	Phosphorus	15	15	15
16	2p4	Sulfur	16	16	16
17	2p5	Chlorine	17	17	17
18	2p6	Argon	18	18	18
21	3d1	Scandium	21	21	21
22	3d2	Titanium	22	22	22
23	3d3	Vanadium	23	23	23
24	3d4	Chromium	24	24	24
25	3d5	Manganese	25	25	25
26	3d6	Iron	26	26	26
27	3d7	Cobalt	27	27	27
28	3d8	Nickel	28	28	28
29	3d9	Copper	29	29	29
30	3d10	Zinc	30	30	30

Deuterium [not Hydrogen] is the building block of elements

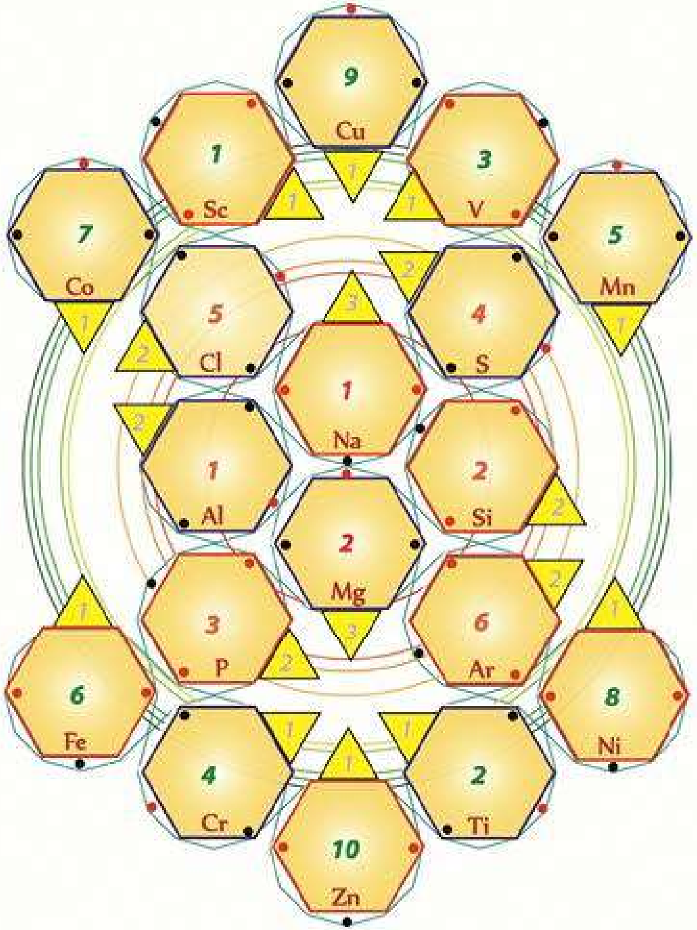
9
Cu
1
Sc
3
V
7
Co
5
Mn
5
Cl
4
S
1
Na
1
Al
2
Si
2
Mg
3
P
6
Ar
6
Fe
8
Ni
4
Cr
2
Ti
10
Zn

M shell

n3

Ground State electron

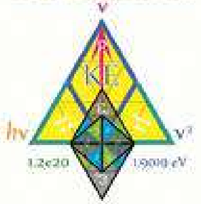
v
KE
hv
v²
1.2e20
1.9008 eV

$iE = -11.623\ eV$

Energy level
3

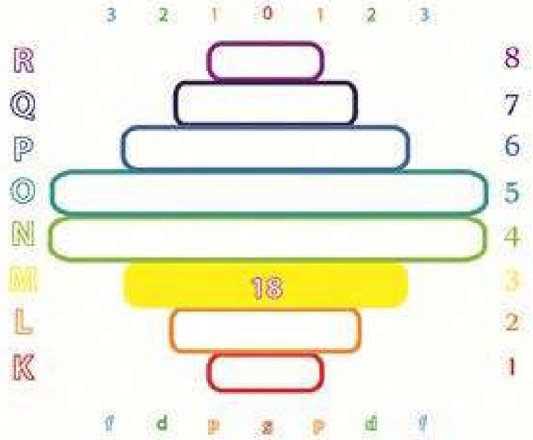
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
18
f
d
p
s
p
d
f

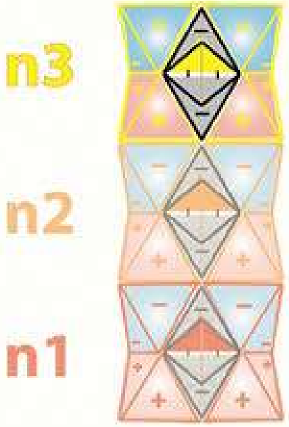

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 4

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
19	4s1	Potassium	19	19	19
20	4s2	Calcium	20	20	20
31	4p1	Gallium	31	31	31
32	4p2	Germanium	32	32	32
33	4p3	Arsenic	33	33	33
34	4p4	Selenium	34	34	34
35	4p5	Bromine	35	35	35
36	4p6	Krypton	36	36	36
39	4d1	Yttrium	39	39	39
40	4d2	Zirconium	40	40	40
41	4d3	Niobium	41	41	41
42	4d4	Molybdenum	42	42	42
43	4d5	Technetium	43	43	43
44	4d6	Ruthenium	44	44	44
45	4d7	Rhodium	45	45	45
46	4d8	Palladium	46	46	46
47	4d9	Silver	47	47	47
48	4d10	Cadmium	48	48	48
57	4f1	Lanthanum	57	57	57
58	4f2	Cerium	58	58	58
59	4f3	Praseodymium	59	59	59
60	4f4	Neodymium	60	60	60
61	4f5	Promethium	61	61	61
62	4f6	Samarium	62	62	62
63	4f7	Europium	63	63	63
64	4f8	Gadolinium	64	64	64
65	4f9	Terbium	65	65	65
66	4f10	Dysprosium	66	66	66
67	4f11	Holium	67	67	67
68	4f12	Erbium	68	68	68
69	4f13	Thulium	69	69	69
70	4f14	Ytterbium	70	70	70

32

Deuterium [not Hydrogen] is the building block of elements

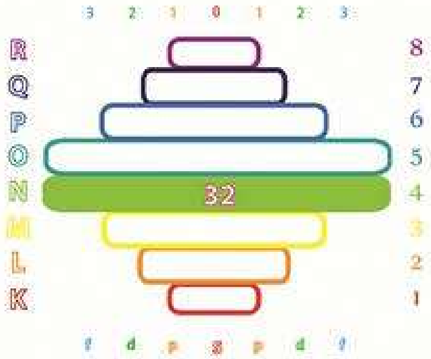

3 2 1 0 1 2 3
R 8
Q 7
P 6
O 5
N 32 4
M 3
L 2
K 1
f d p s p d f

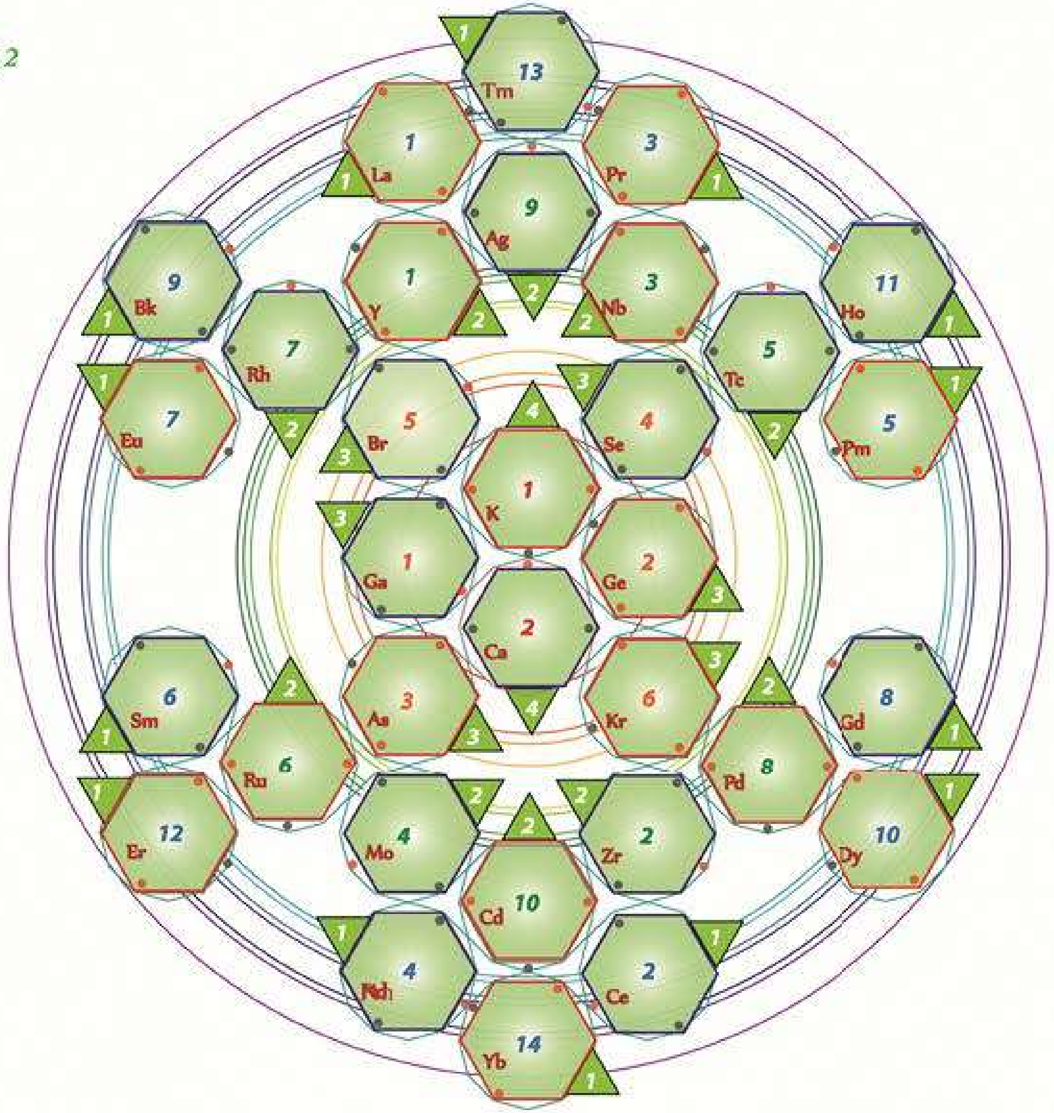

13 Tm
1 La
3 Pr
9 Ag
1 Y
3 Nb
9 Bk
11 Ho
7 Rh
5 Tc
7 Eu
5 Pm
5 Br
4 Se
1 K
1 Ga
2 Ge
2 Ca
3 As
6 Kr
6 Sm
8 Gd
6 Ru
8 Pd
12 Er
10 Dy
4 Mo
2 Zr
10 Cd
4 Nd
2 Ce
14 Yb

N shell

n4

Ground State electron

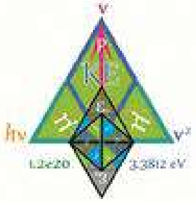

v
hv
v²
3.3812 eV

iE = -10.143 eV

Energy level

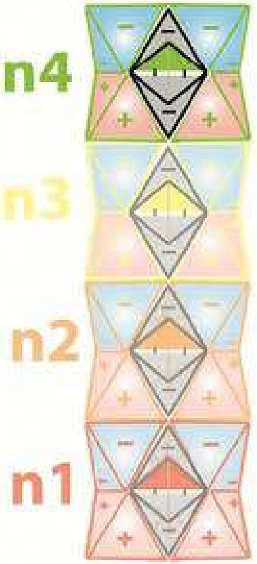

Quantum Level 5

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
37	5s1	Rubidium	37	37	37
38	5s2	Strontium	38	38	38
49	5p1	Indium	49	49	49
50	5p2	Tin	50	50	50
51	5p3	Antimony	51	51	51
52	5p4	Tellurium	52	52	52
53	5p5	Iodine	53	53	53
54	5p6	Xenon	54	54	54
71	5d1	Lutetium	71	71	71
72	5d2	Hafnium	72	72	72
73	5d3	Tantalum	73	73	73
74	5d4	Tungsten	74	74	74
75	5d5	Rhenium	75	75	75
76	5d6	Osmium	76	76	76
77	5d7	Iridium	77	77	77
78	5d8	Platinum	78	78	78
79	5d9	Gold	79	79	79
80	5d10	Mercury	80	80	80
89	5f1	Actinium	89	89	89
90	5f2	Thorium	90	90	90
91	5f3	Protactium	91	91	91
92	5f4	Uranium	92	92	92
93	5f5	Neptunium	93	93	93
94	5f6	Plutonium	94	94	94
95	5f7	Americium	95	95	95
96	5f8	Curium	96	96	96
97	5f9	Berkelium	97	97	97
98	5f10	Californium	98	98	98
99	5f11	Einsteinium	99	99	99
100	5f12	Fermium	100	100	100
101	5f13	Mendelevium	101	101	101
102	5f14	Nobelium	102	102	102

32

Deuterium [not Hydrogen] is the building block of elements

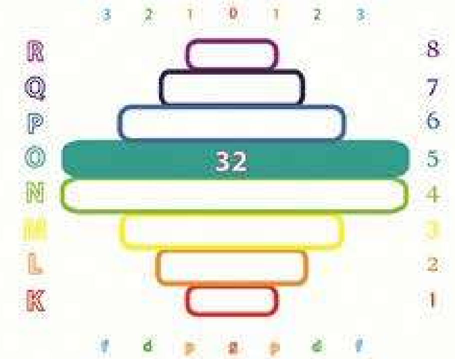
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
32
f d p s p d f

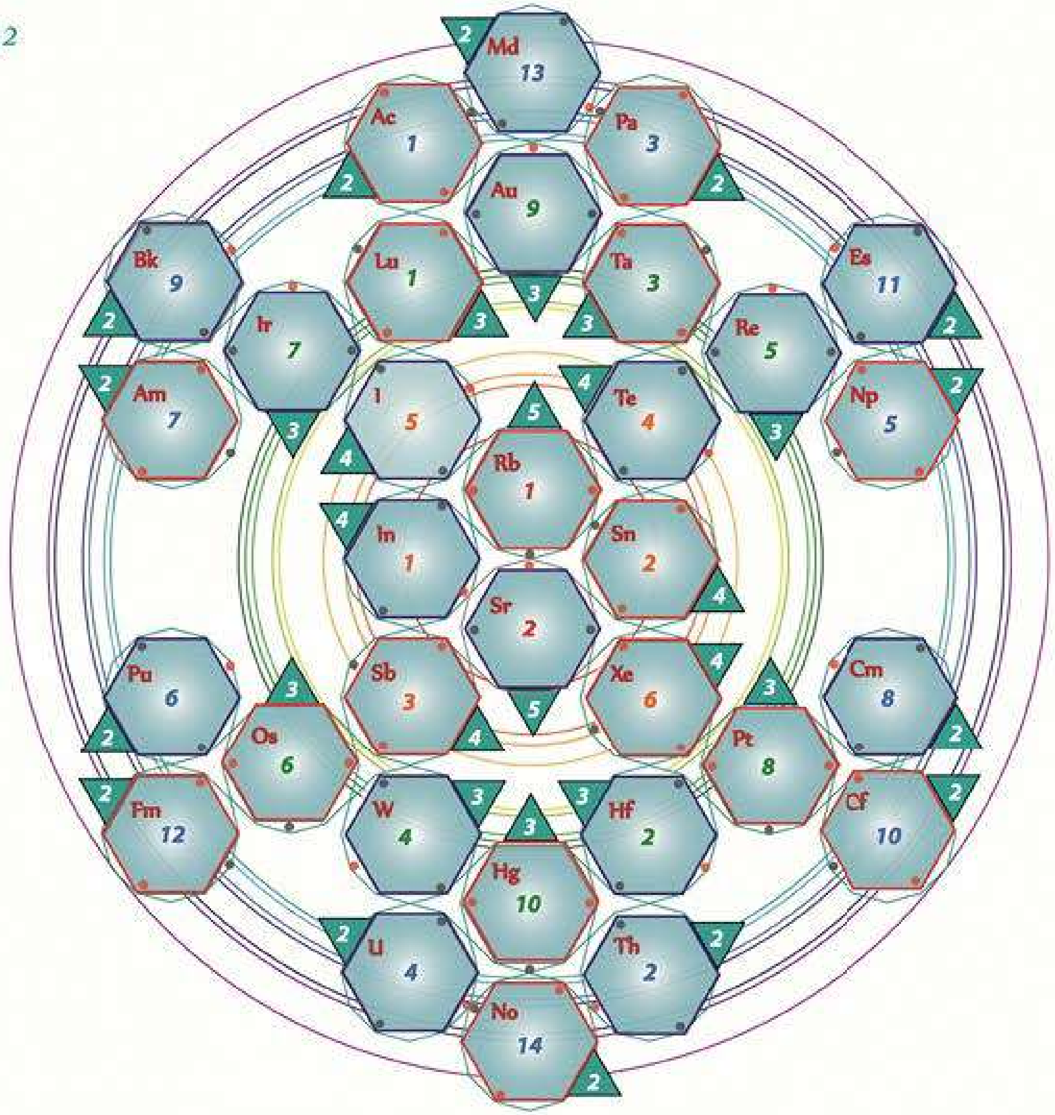
Md 13
Ac 1
Pa 3
Au 9
Lu 1
Ta 3
Bk 9
Ir 7
Re 5
Es 11
Am 7
I 5
Te 4
Np 5
Rb 1
In 1
Sn 2
Sr 2
Pu 6
Sb 3
Xe 6
Cm 8
Os 6
Pt 8
Fm 12
W 4
Hf 2
Cf 10
Hg 10
U 4
Th 2
No 14

O shell

n5

Ground State electron

v
P
KE
hv
v²
1.2e20
5.5282 eV

$iE = -8.241\ eV$

Energy level
5

n5
n4
n3
n2
n1

Quantum Level 6

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
55	6s1	Caesium	55	55	55
56	6s2	Barium	56	56	56
81	6p1	Thallium	81	81	81
82	6p2	Lead	82	82	82
83	6p3	Bismuth	83	83	83
84	6p4	Polonium	84	84	84
85	6p5	Astatine	85	85	85
86	6p6	Radon	86	86	86
103	6d1	Lawrencium	103	103	103
104	6d2	Rutherfordium	104	104	104
105	6d3	Dubnium	105	105	105
106	6d4	Seaborgium	106	106	106
107	6d5	Bohrium	107	107	107
108	6d6	Hassium	108	108	108
109	6d7	Meitnerium	109	109	109
110	6d8	Darmstadtium	110	110	110
111	6d9	Roetgenium	111	111	111
112	6d10	Copernicium	112	112	112

18

Deuterium [not Hydrogen] is the building block of elements

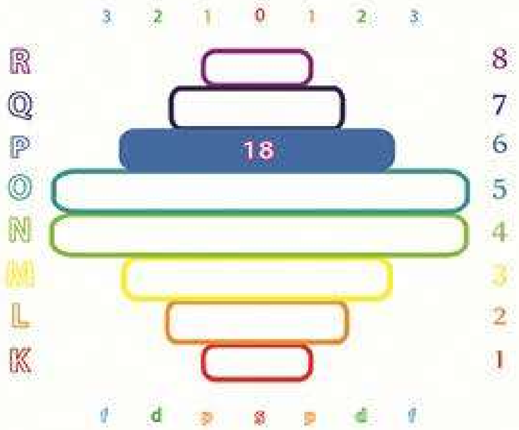
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
18
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Energy level
6

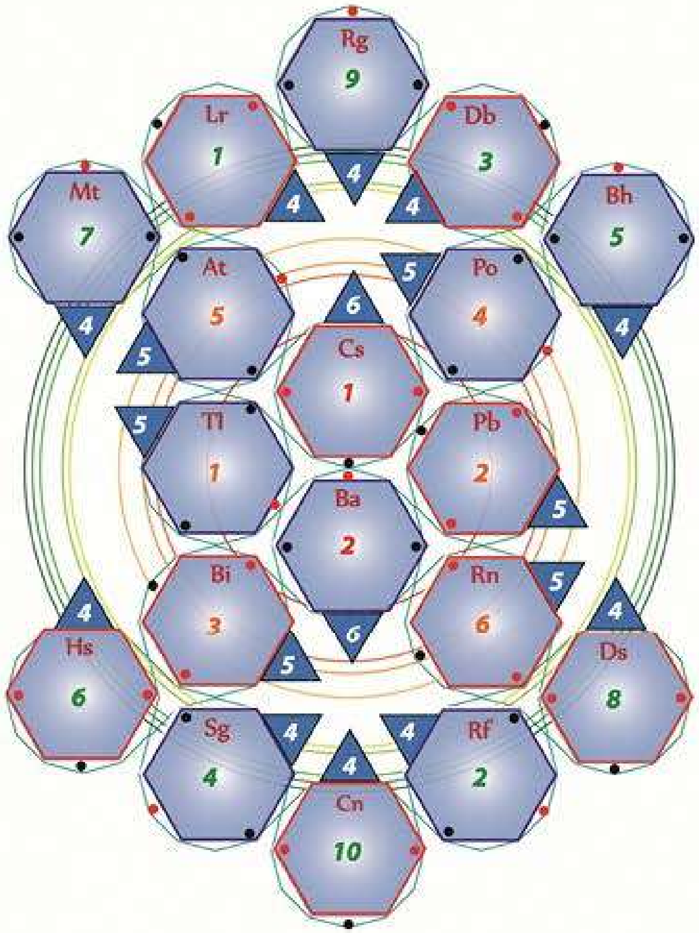
Rg
9
Lr
1
Db
3
Mt
7
Bh
5
At
5
Po
4
Cs
1
Tl
1
Pb
2
Ba
2
Bi
3
Rn
6
Hs
6
Ds
8
Sg
4
Rf
2
Cn
10

P shell

n6

Ground State electron

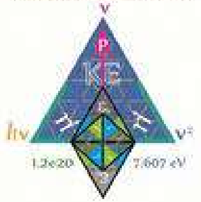
v
P
KE
hv
v²
1.2e20
7.607 eV

iE = -5.917 eV

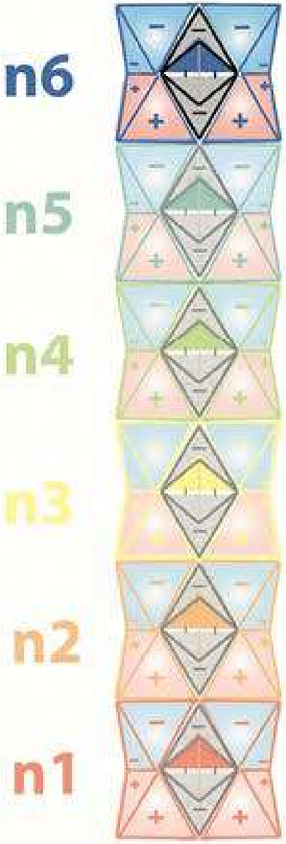
n6
n5
n4
n3
n2
n1

Quantum Level 7

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
87	7s1	Francium	87	87	87
88	7s2	Radium	88	88	88
113	7p1	Ununtrium	113	113	113
114	7p2	Flerovium	114	114	114
115	7p3	Ununpentium	115	115	115
116	7p4	Livermorium	116	116	116
117	7p5	Ununseptium	117	117	117
118	7p6	Ununoctium	118	118	118

8

Deuterium [not Hydrogen] is the building block of elements

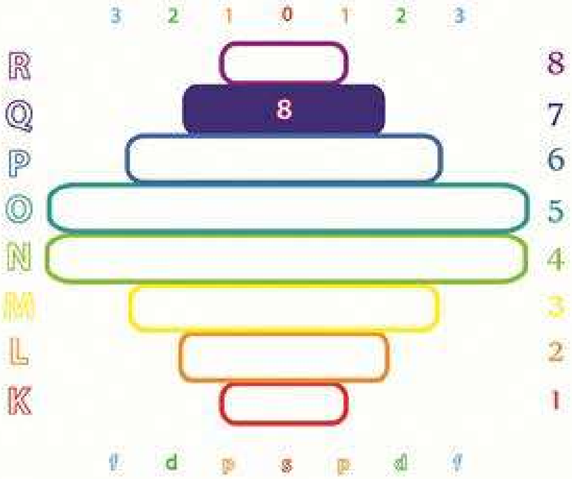
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Energy level

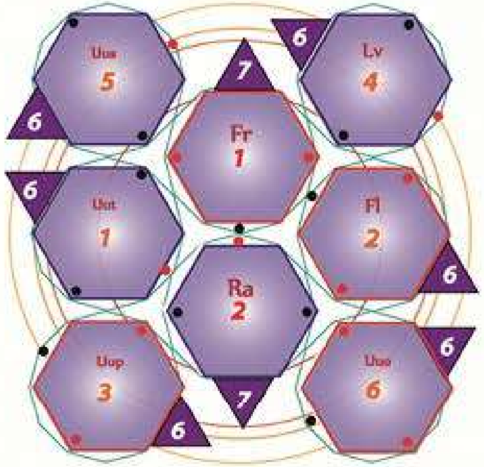
Uus
5
Lv
4
Fr
1
Uut
1
Fl
2
Ra
2
Uup
3
Uuo
6
7
7
6
6
6
6
6
6

Q shell

n7

Ground State electron

KE
hν
v²
1.2e20
10.355 eV

iE = -3.169 eV

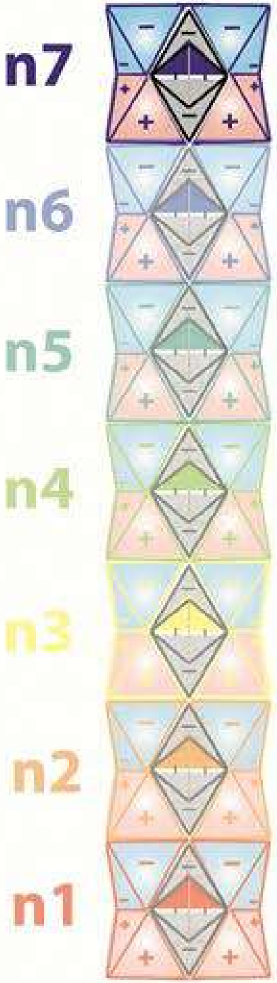
n7
n6
n5
n4
n3
n2
n1

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 8

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
119	8s1	Ununnonium	119	119	119
120	8s2	Unbinilium	120	120	120

2

Deuterium [not Hydrogen] is the building block of elements

R shell

n8

Ground State electron

v
P
KE
8
hv
v²
1.2e20
13.525 eV

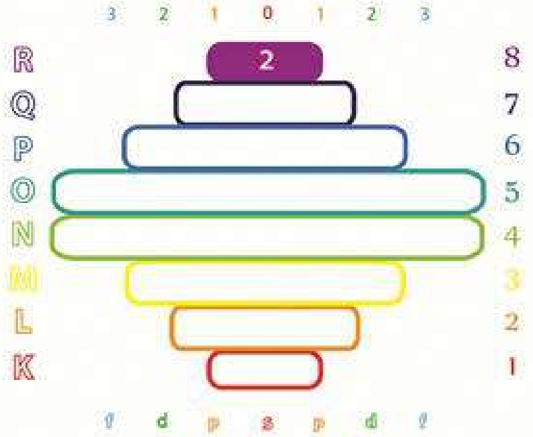
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

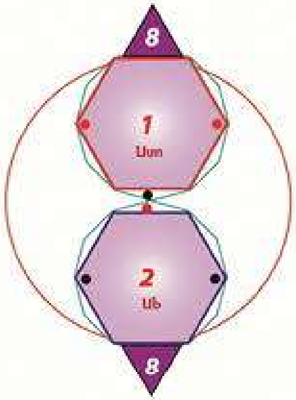

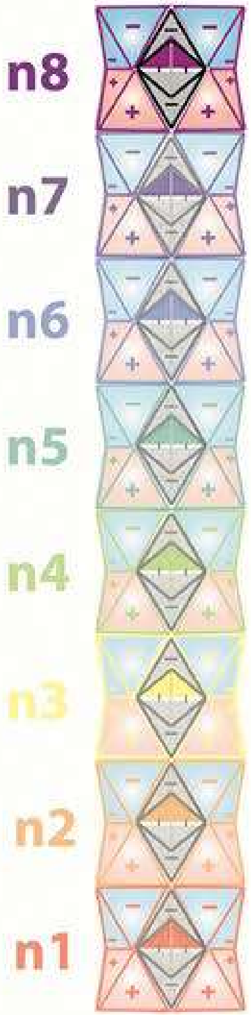
n8
n7
n6
n5
n4
n3
n2
n1

Energy level
8

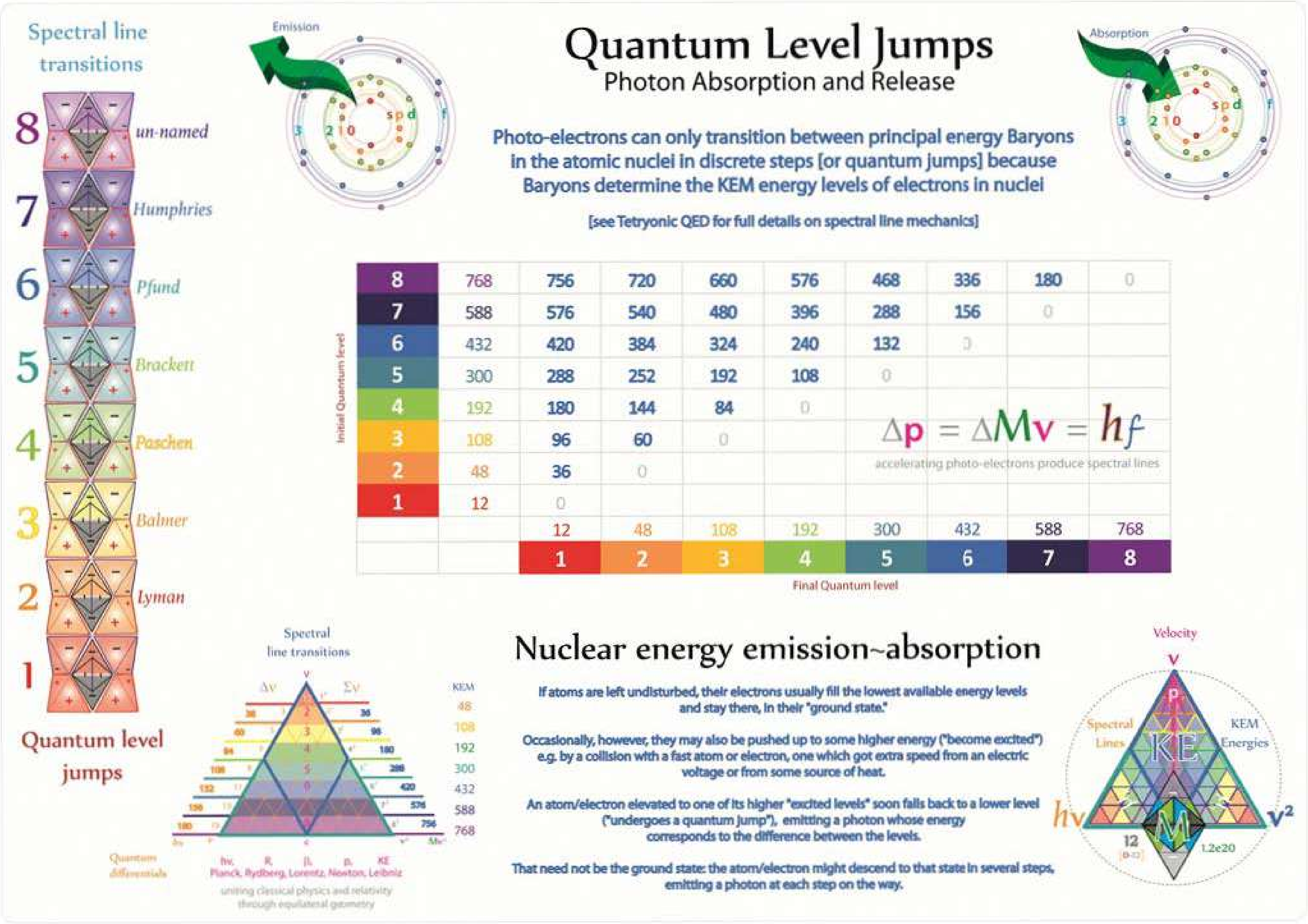
Spectral line transitions
8 un-named
7 Humphries
6 Pfund
5 Brackett
4 Paschen
3 Balmer
2 Lyman
1
Quantum level jumps
Emission
Absorption
s p d f
3 2 1 0
Spectral line transitions
Δv
Σv
v
KEM
48
108
192
300
432
588
768
Quantum differentials
hv, R, μ, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz
uniting classical physics and relativity through equilateral geometry
Velocity
v
Spectral Lines
KEM Energies
KE
hv
v^2
12
1.2e20

Quantum Level Jumps

Photon Absorption and Release

Photo-electrons can only transition between principal energy Baryons in the atomic nuclei in discrete steps [or quantum jumps] because Baryons determine the KEM energy levels of electrons in nuclei

[see Tetryonic QED for full details on spectral line mechanics]

Initial Quantum level

8	768	756	720	660	576	468	336	180	0
7	588	576	540	480	396	288	156	0	
6	432	420	384	324	240	132	0		
5	300	288	252	192	108	0			
4	192	180	144	84	0				
3	108	96	60	0					
2	48	36	0						
1	12	0							
		12	48	108	192	300	432	588	768
		1	2	3	4	5	6	7	8

Final Quantum level

$$\Delta p = \Delta M v = hf$$

accelerating photo-electrons produce spectral lines

Nuclear energy emission~absorption

If atoms are left undisturbed, their electrons usually fill the lowest available energy levels and stay there, in their "ground state."

Occasionally, however, they may also be pushed up to some higher energy ("become excited") e.g. by a collision with a fast atom or electron, one which got extra speed from an electric voltage or from some source of heat.

An atom/electron elevated to one of its higher "excited levels" soon falls back to a lower level ("undergoes a quantum jump"), emitting a photon whose energy corresponds to the difference between the levels.

That need not be the ground state: the atom/electron might descend to that state in several steps, emitting a photon at each step on the way.

Tetryonics 46.10 - Quantum level jumps

Copyright ABRAHAM [2008] - All rights reserved

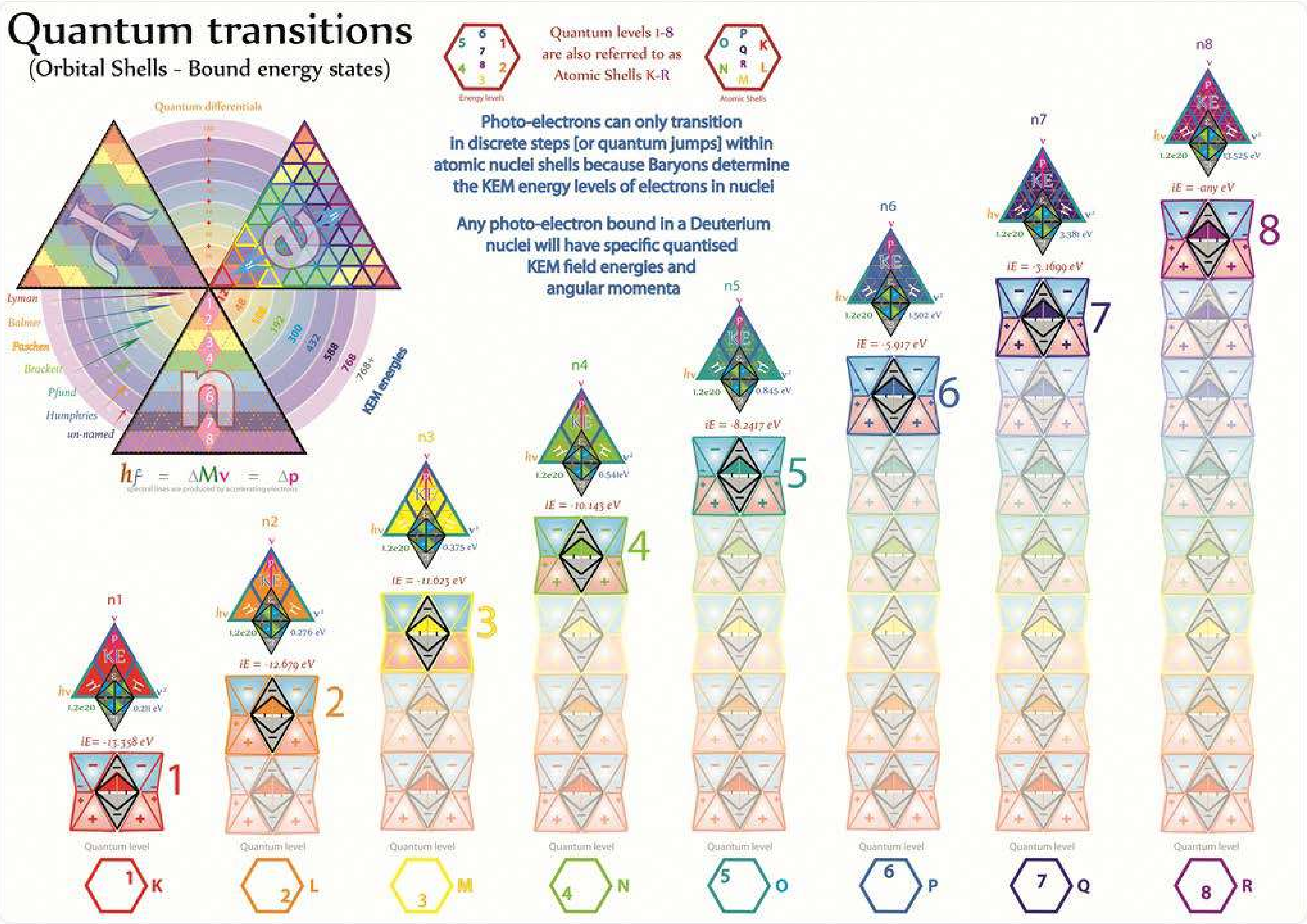

Quantum transitions

(Orbital Shells - Bound energy states)

Quantum levels 1-8 are also referred to as Atomic Shells K-R

Photo-electrons can only transition in discrete steps [or quantum jumps] within atomic nuclei shells because Baryons determine the KEM energy levels of electrons in nuclei

Any photo-electron bound in a Deuterium nuclei will have specific quantised KEM field energies and angular momenta

$hf = \Delta M v = \Delta p$

spectral lines are produced by accelerating electrons

Energy levels
Atomic Shells
Quantum differentials
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
48
108
192
300
432
588
768
768+
KEM energies
n1
n2
n3
n4
n5
n6
n7
n8
P
KE
hv
v²
1.2e20
0.211 eV
0.276 eV
0.375 eV
0.541eV
0.845 eV
1.502 eV
3.381 eV
13.525 eV
iE= -13.358 eV
iE = -12.679 eV
iE = -11.023 eV
iE = -10.143 eV
iE = -8.2417 eV
iE = -5.917 eV
iE = -3.1699 eV
iE = -any eV
1
2
3
4
5
6
7
8
Quantum level
K
L
M
N
O
P
Q
R

Copyright ABRAHAM [2008] - All rights reserved

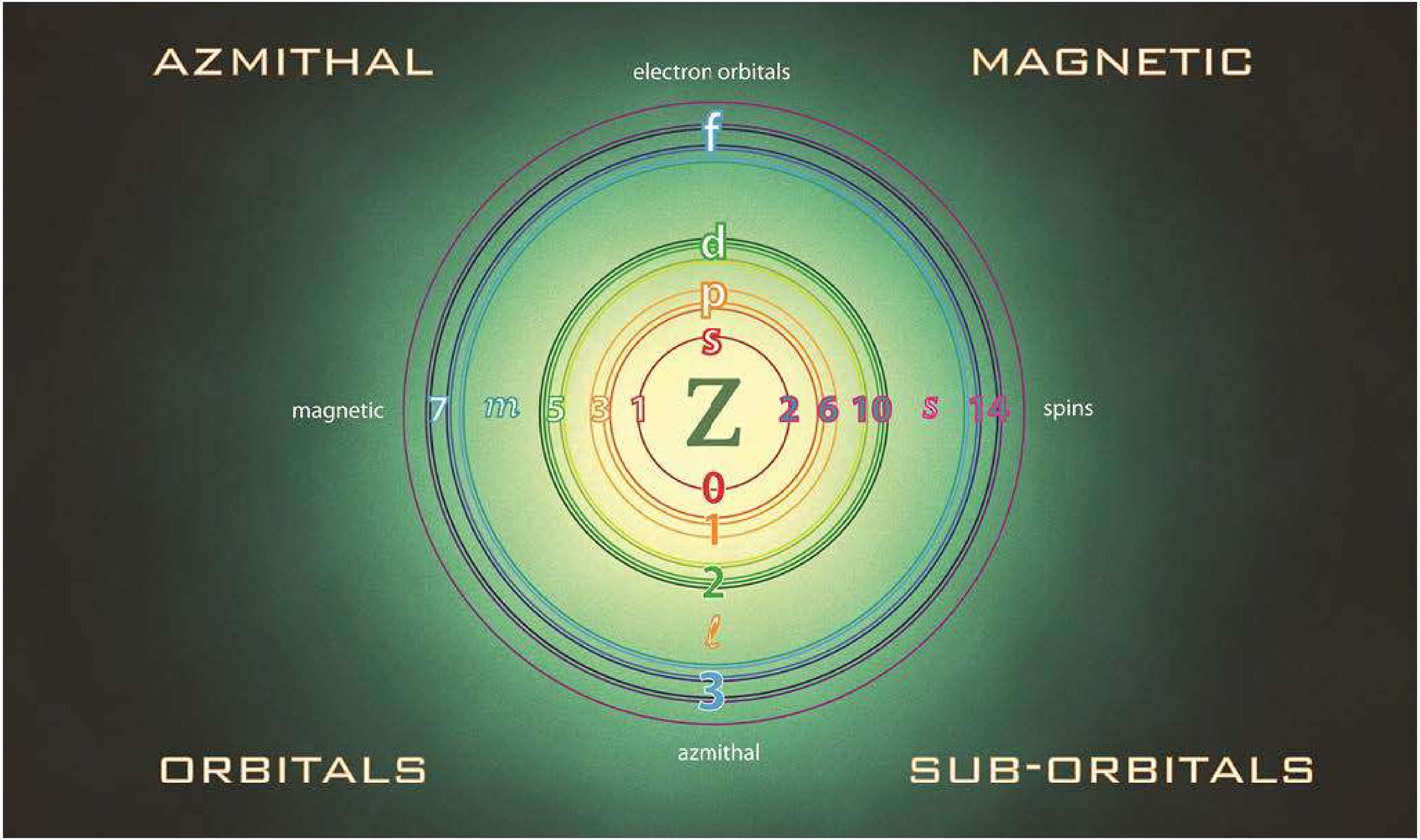
AZMITHAL
electron orbitals
MAGNETIC
f
d
p
s
magnetic
7
m
5
3
1
Z
2
6
10
s
14
spins
0
1
2
l
3
azmithal
ORBITALS
SUB-ORBITALS

Tetryonics 47.01 - Elemental Orbits

Atomic Orbitals

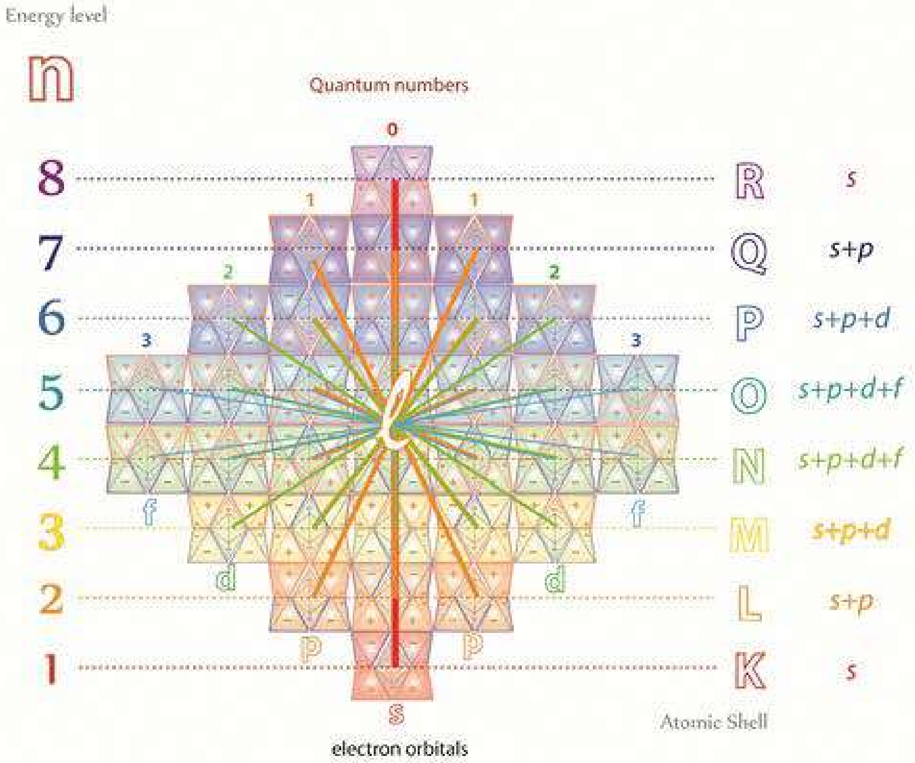
Energy level
n
Quantum numbers
0
1
1
2
2
3
3
8
7
6
5
4
3
2
1
R
Q
P
O
N
M
L
K
s
s+p
s+p+d
s+p+d+f
s+p+d+f
s+p+d
s+p
s
ℓ
f
f
d
d
p
p
s
electron orbitals
Atomic Shell

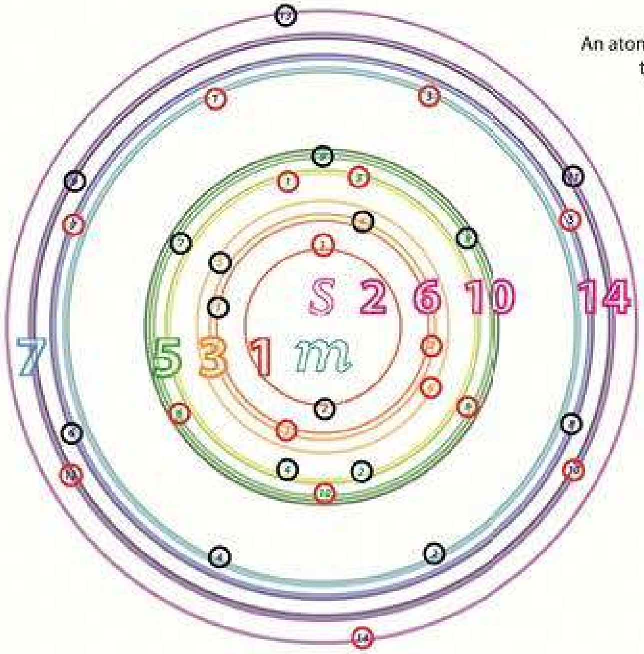
S
2
6
10
14
7
5
3
1
m

An atomic orbital is a mathematical function that describes the wave-like behavior of either one electron or a pair of electrons in an atom

Atomic orbitals are typically categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and an angular momentum vector component, respectively.

Historically used to define the pedagogical electron cloud model of an atom Tetryonics reveals the true geometry of atomic nuclei

Each orbital is defined by a different set of quantum numbers and contains a maximum of two spin opposed electrons.

Aufbau

Azmithal & Magnetic numbers

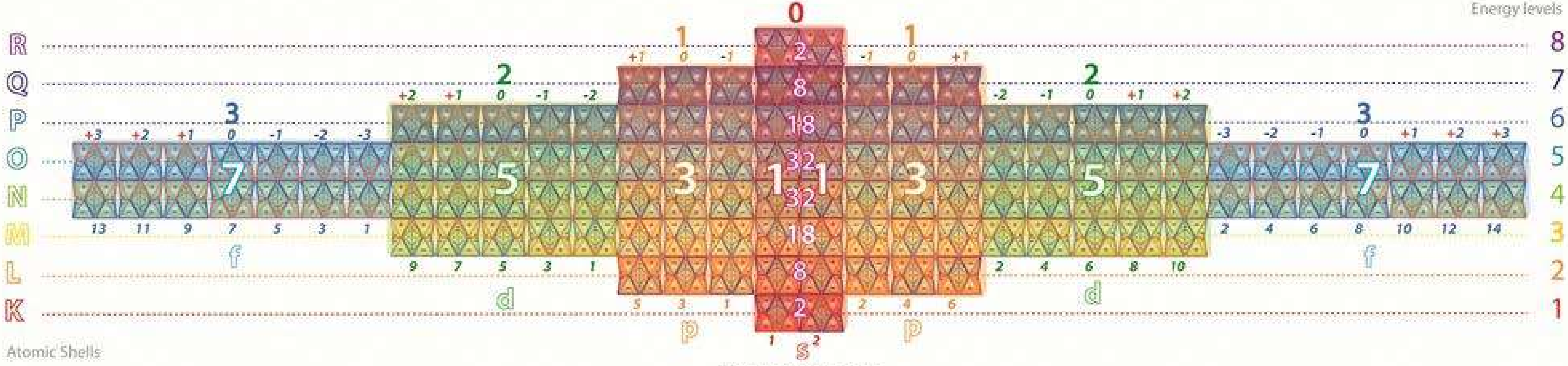
Energy levels
0
1
1
2
2
3
3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
2
8
18
32
32
18
8
2
7
5
3
1
1
3
5
7
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
Atomic Shells

Orbitals & sub-Orbitals

Electron Spins can be either up or down providing they obey the Pauli exclusion principle

s Orbital

1 Orbit (2 electrons max)

Alkali Metals & Alkali Earths

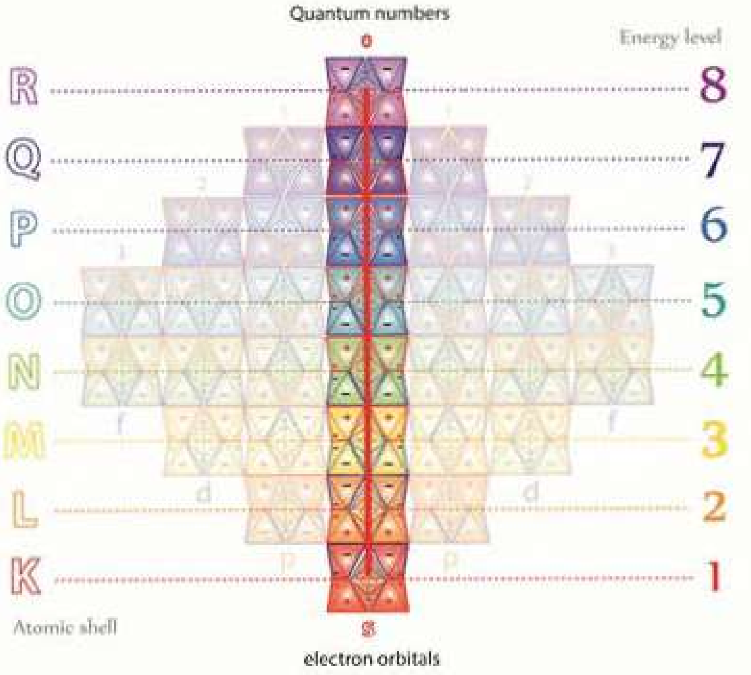
Quantum numbers
0
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
Atomic shell
s
electron orbitals

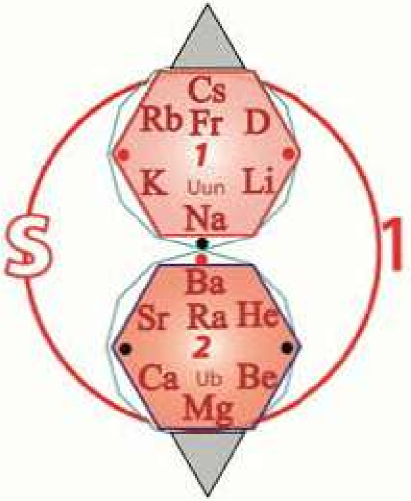
Cs
Rb
Fr
D
1
K
Uun
Li
Na
Ba
Sr
Ra
He
2
Ca
Ub
Be
Mg
s
1

		n
azmithal number	0	l
magnetic numbers	0	m
spins	-1/2 +1/2	s

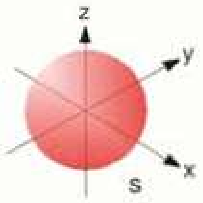

erroneous computer model of 's' electron orbitals

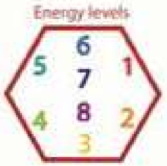
Energy levels
6
5
1
7
8
4
2
3

p Orbital

3 Orbits (6 electrons max)

Non-Metals, Halogens & Nobel gases

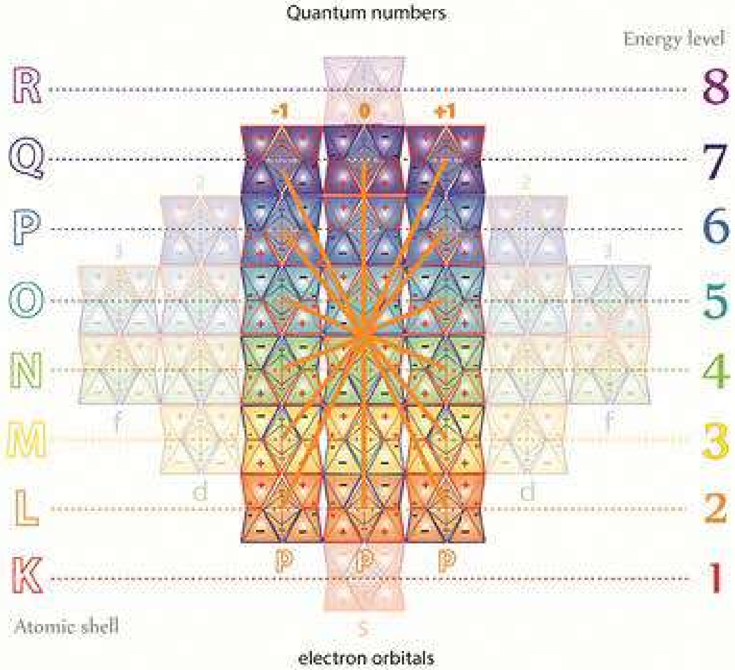
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-1
0
+1
p
p
p
s
d
d
f
f
Atomic shell
electron orbitals

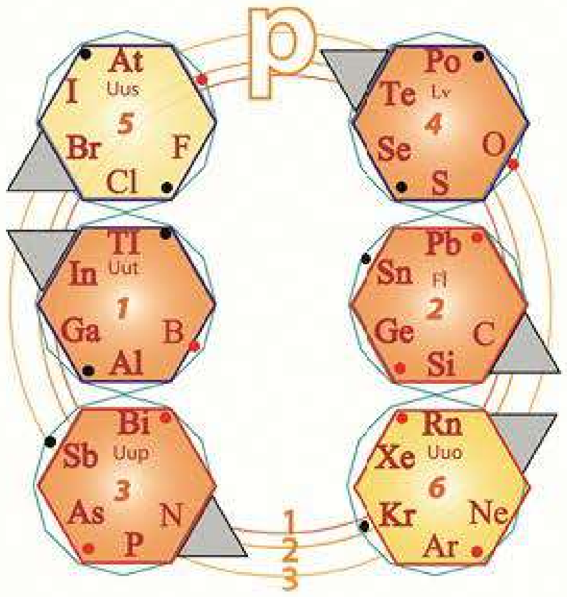
p
At
Uus
I
5
Br
F
Cl
Po
Te
Lv
4
Se
O
S
Tl
Uut
In
1
Ga
B
Al
Pb
Sn
Fl
2
Ge
C
Si
Bi
Sb
Uup
3
As
N
P
Rn
Xe
Uuo
6
Kr
Ne
Ar
1
2
3

n

azmithal number		1		ℓ
magnetic numbers	-1	0	+1	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	s

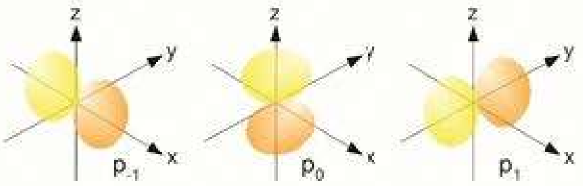
z
y
x
p_{-1}
z
y
x
p_0
z
y
x
p_1

erroneous computer model of 'p' electron orbitals

Quantum levels
6
5
7
4
2
3

d Orbital

5 Orbits (10 electrons max)

Transition & post-Transition Metals

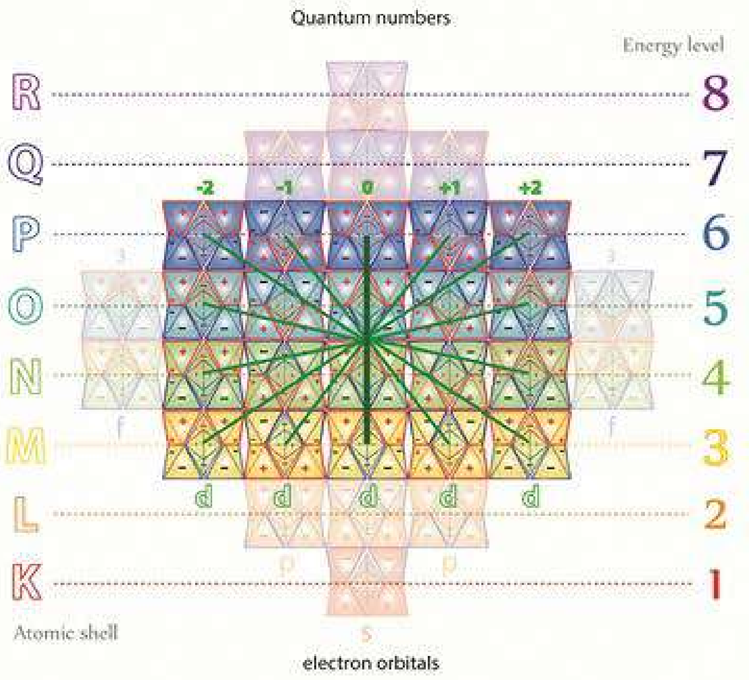
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-2
-1
0
+1
+2
f
f
d
d
d
d
d
p
p
s
Atomic shell
electron orbitals

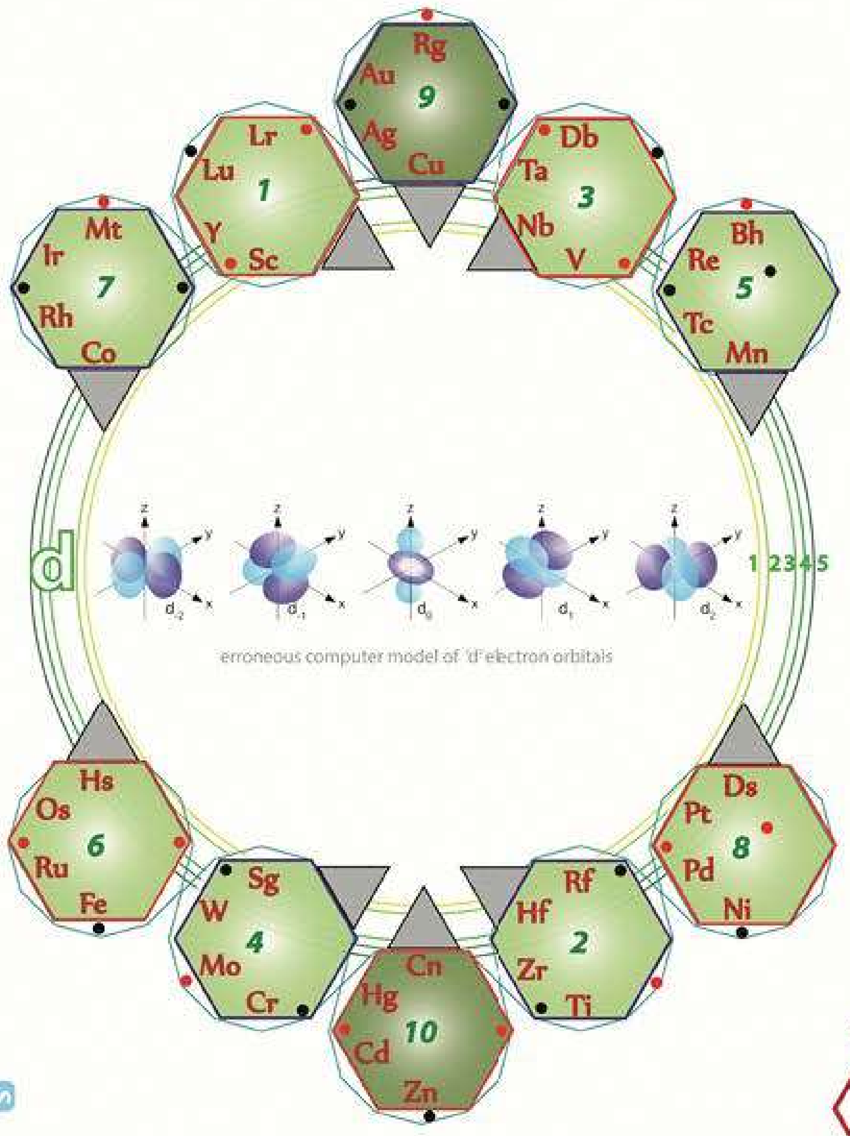
Rg
Au
9
Ag
Cu
Lr
Lu
1
Y
Sc
Db
Ta
3
Nb
V
Mt
Ir
7
Rh
Co
Bh
Re
5
Tc
Mn
d
1 2 3 4 5
erroneous computer model of 'd' electron orbitals
Hs
Os
6
Ru
Fe
Ds
Pt
8
Pd
Ni
Sg
W
4
Mo
Cr
Rf
Hf
2
Zr
Ti
Cn
Hg
10
Cd
Zn

						n
azmithal number			2			*l*
magnetic numbers	-2	-1	0	+1	+2	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	s

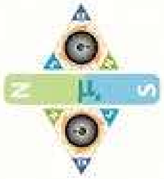

Quantum levels
6
5
4
3

f Orbital

7 Orbits (14 electrons max)

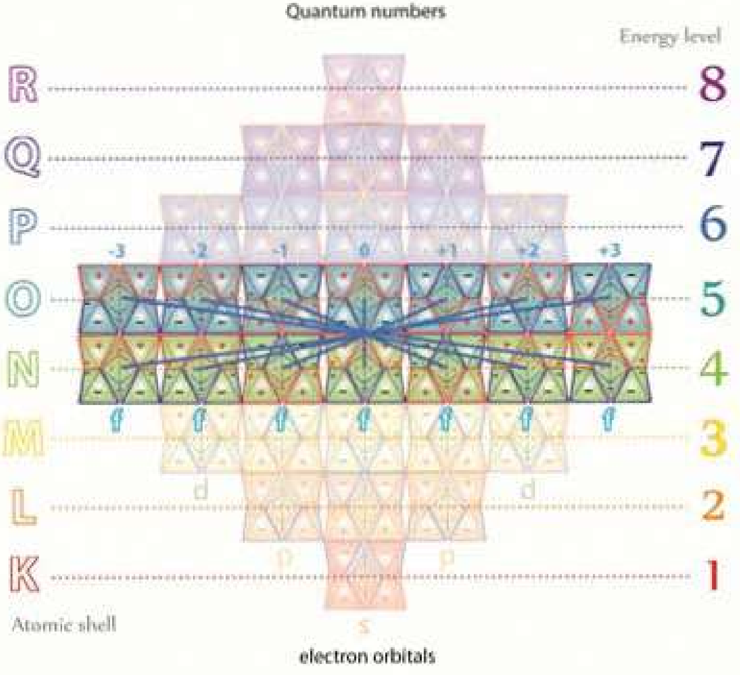
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-3
-2
-1
0
+1
+2
+3
f
f
f
f
f
f
f
d
d
p
p
s
Atomic shell
electron orbitals

								n
azmithal number				3				ℓ
magnetic numbers	-3	-2	-1	0	+1	+2	+3	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	S

Lanthanoids & Actinoids

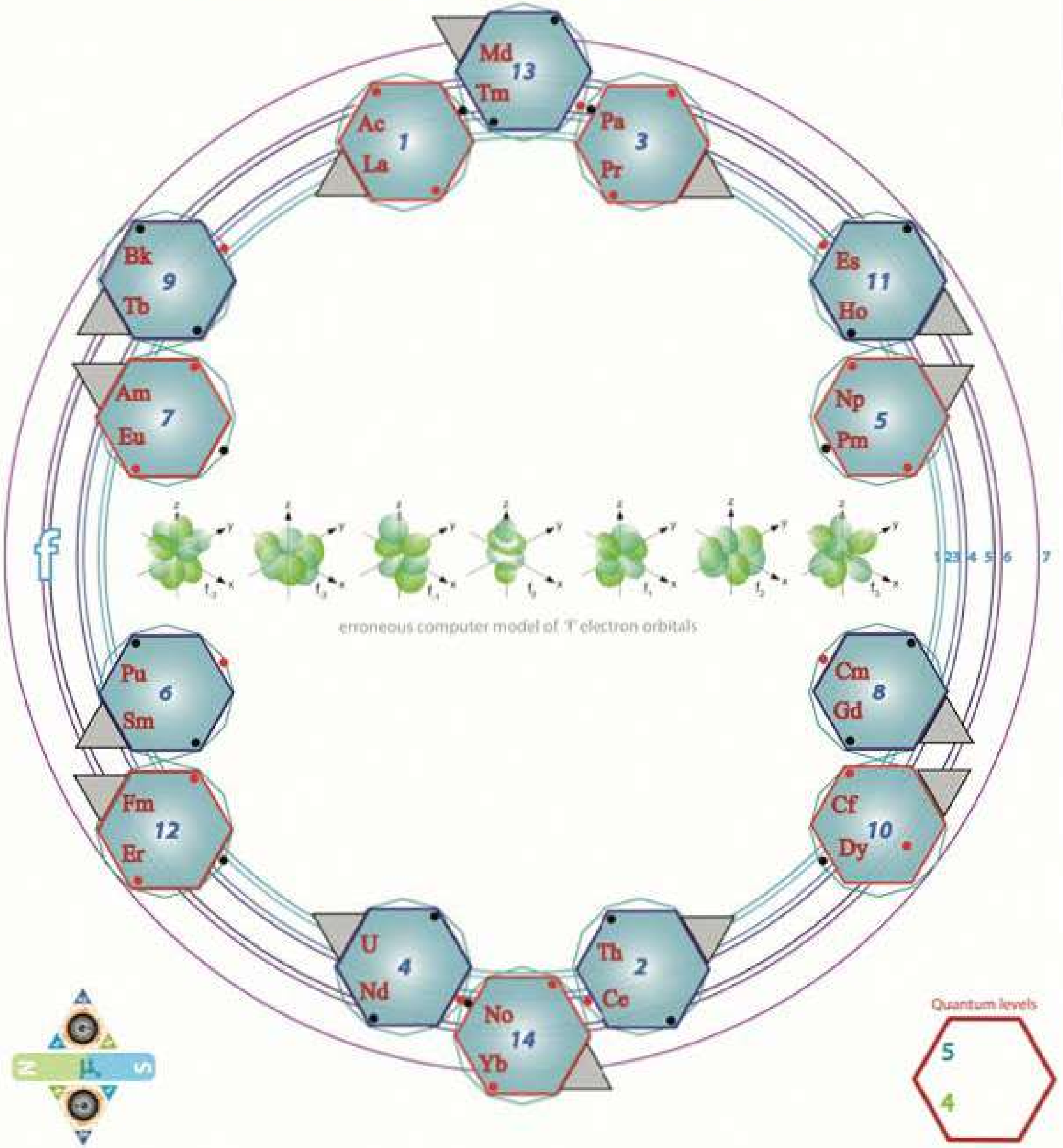
Md
13
Tm
Ac
1
La
Pa
3
Pr
Bk
9
Tb
Es
11
Ho
Am
7
Eu
Np
5
Pm
f
1 2 3 4 5 6
7
Pu
6
Sm
Cm
8
Gd
Fm
12
Er
Cf
10
Dy
U
4
Nd
Th
2
Ce
No
14
Yb
Quantum levels
5
4

erroneous computer model of 'f' electron orbitals

Tetryonics 47.06 - 'f' Orbital

Orbital energy variations

All Elements have stable core electron configurations of s & p orbitals for each energy level as revealed through diffraction studies

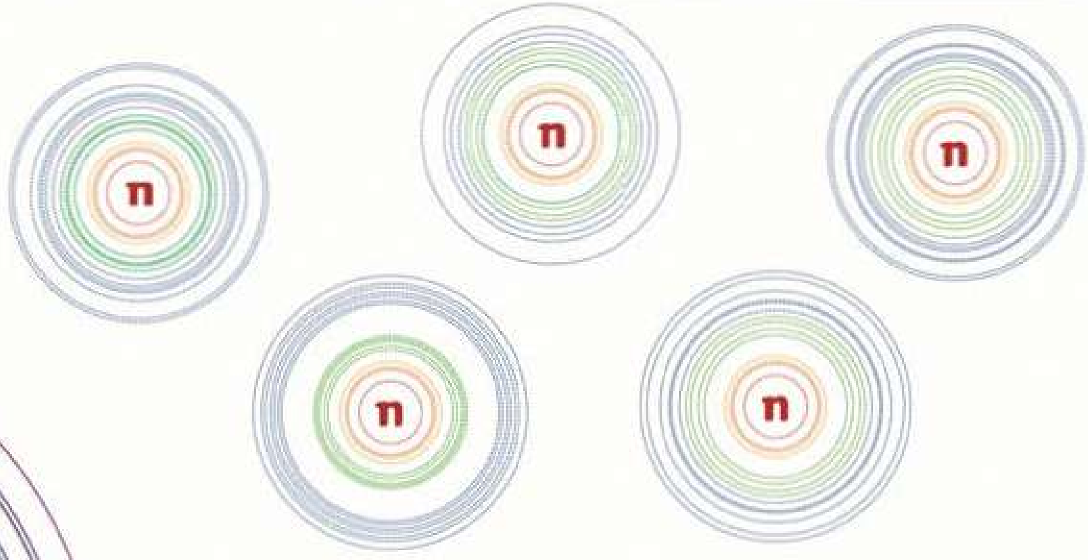

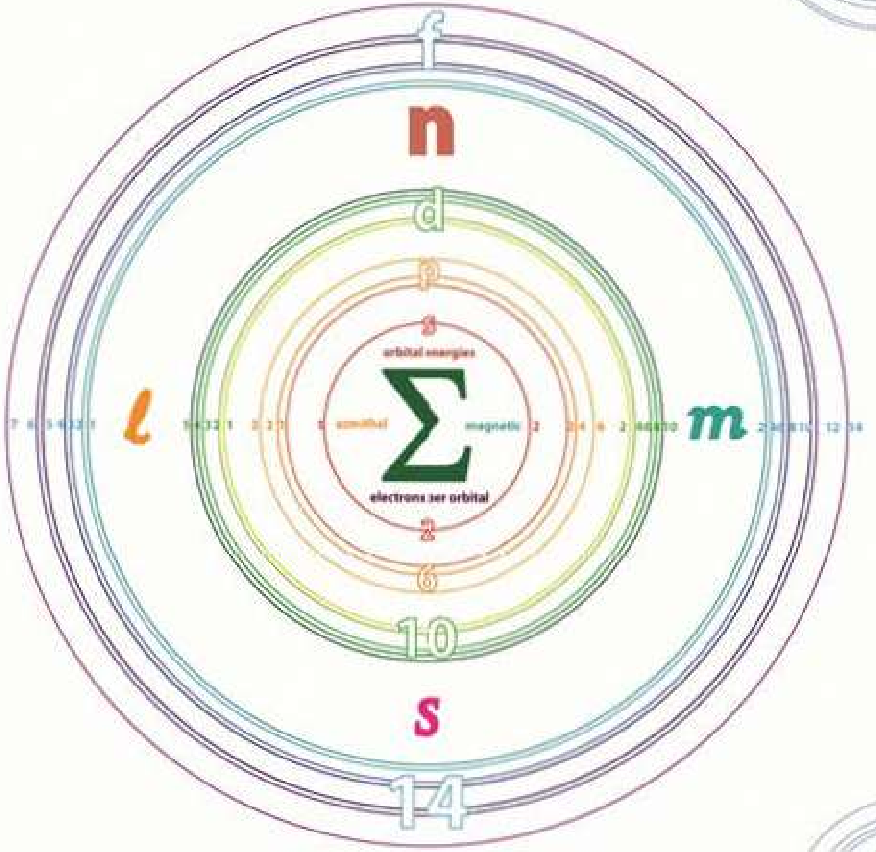
f
n
d
p
s
orbital energies
ℓ
azimuthal
Σ
magnetic
m
electrons per orbital
2
6
10
s
14

As additional nuclei bind to form d & f orbitals they can do so by bonding in many positions, with each location producing different orbital energies for each electron that binds to nuclei in that position

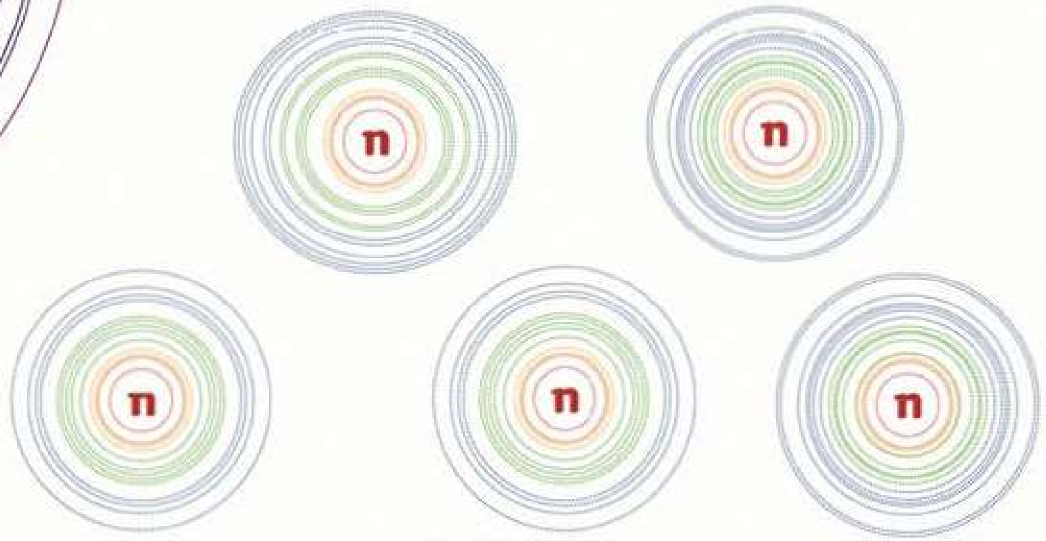

The final energy levels of each orbital is the result of the energy of the Baryons in the nuclei and the spin coupling energies of the photo-electrons bound to them

Copyright ABRAHAM [2008] - All rights reserved

Electrons per shell

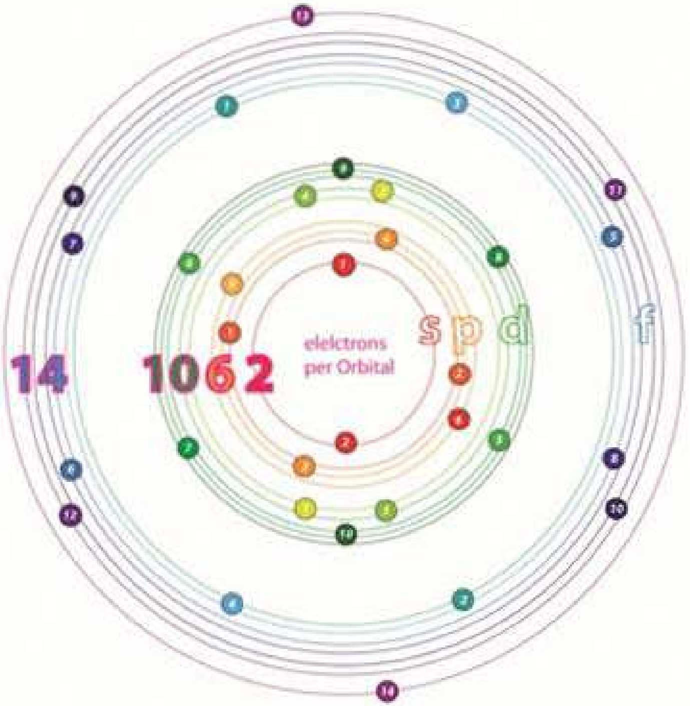

Each energy shell of a periodic element can hold only a fixed number of electrons

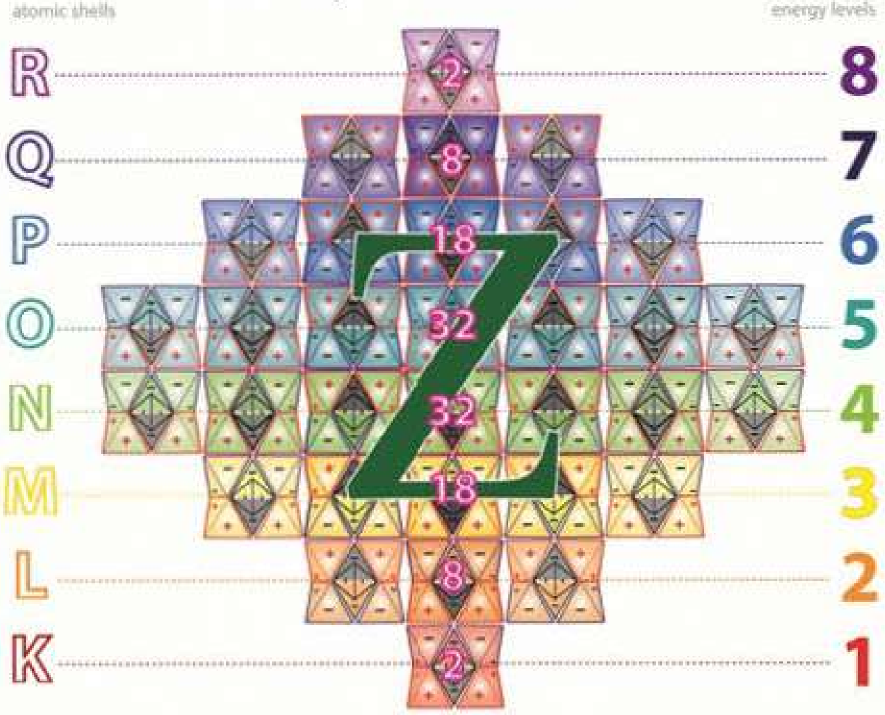

Atomic Orbitals

f	d	p	s	p	d	f
5, 11, 3, 13, 1, 9, 7	8, 2, 10, 4, 6	5, 1, 3	1-2	6, 2, 4	7, 1, 9, 3, 5	6, 12, 4, 14, 2, 10, 8

7 + 5 + 3 + 2 + 3 + 5 + 7

Rules governing the allowed combinations of Quantum Numbers

The three quantum numbers (n, l, and m) that describe an orbital are integers: 0, 1, 2, 3, and so on

n (1-8) Principal	The principal quantum number (n) cannot be zero.
ℓ (0-3) Azimuthal	The angular quantum number (l) can be any integer between 0 and n - 1.
mℓ (2ℓ+1) Magnetic	The magnetic quantum number (m) can be any integer between -l and +l.
m_s $\pm\frac{1}{2}$ Spin Projection	The Spin of electrons in any nuclear sub-orbital can only be +1/2 (Spin UP) or -1.2 (Spin DOWN).

Z

1, 2, 3, 4, 5, 6, 7, 8

s, p, d, f

1	2
1, 3, 5	2, 4, 6
1, 3, 5, 7, 9	2, 4, 6, 8, 10
1, 3, 5, 7, 9, 11, 13	2, 4, 6, 8, 10, 12, 14
down up	up down

Quantum Numbers

n	4-5	3-6	2-7	1-8	2-7	3-6	4-5
ℓ	3	2	1	0	1	2	3
m	-3	-2	-1	1	+1	+2	+3
s	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2

Using Tetryonic charged geometries for mass-ENERGY-Matter,
an electron's position and velocity CAN be modelled simultaneously
(but any attempt to measure or interact with it, will affect its component energy-momenta)

Electron Position Uncertainty

Atomic orbitals are typically described as "hydrogen-like" (meaning one-electron) wave functions over any spatial region of measurement, categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and a vector momentum component, respectively

Erwin Schrödinger

(12 August 1887 – 4 January 1961)

Werner Heisenberg

(5 December 1901 – 1 February 1976)

Lepton's are physically Spin 1 fermion particles that can easily be misconstrued as having entirely different spin numbers without the correct physical topologies to base the observed measurements on

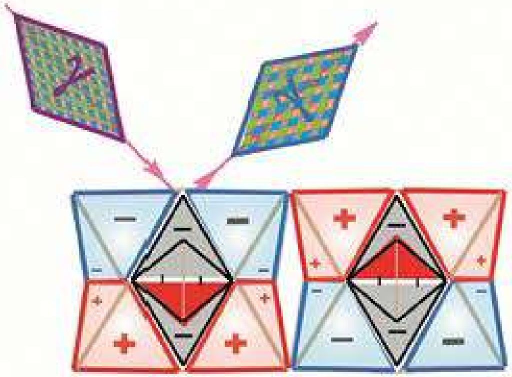

The energies of photo-electrons are determined by the Baryons they bind to & incident photons

Quantum Mechanics is a mathematical representation of equilateral energy momenta interactions and the charged geometries of mass-ENERGY-Matter

Determining the motion of electrons bound to atomic nuclei is akin to measuring the motion of variable speed electric fan blades mounted at various heights within a rotating carousel

Leptons are 12 loop quantum rotors

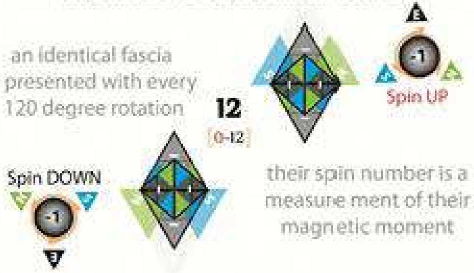

an identical fascia is presented with every 120 degree rotation

their spin number is a measure ment of their magnetic moment

The unique 12 faceted topology of leptons results in an identical EM geometry being oberved for every 120° rotation of the particle

Leading to the interpretation that the Lepton disappears and re-appears when being 'observed' or measured

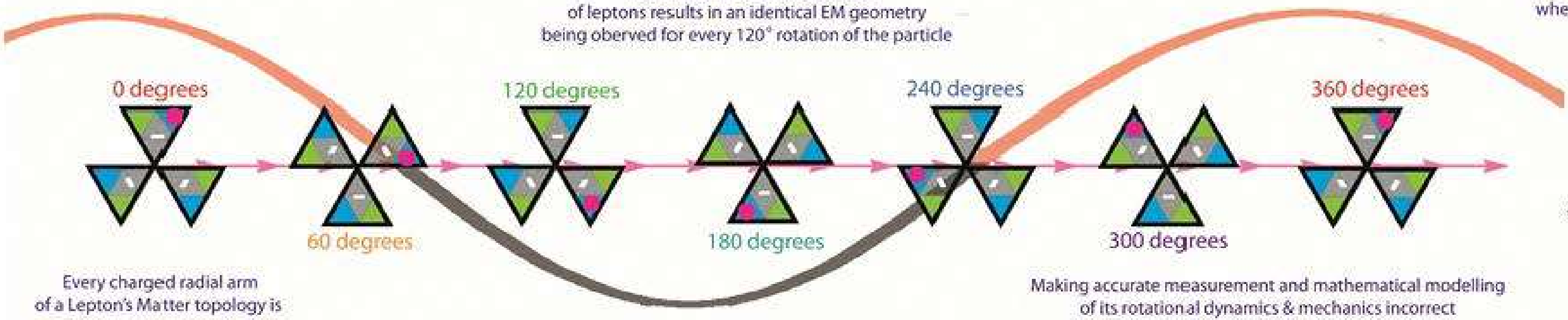

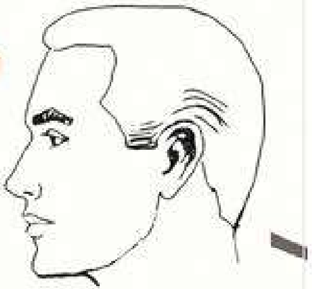

Every charged radial arm of a Lepton's Matter topology is identical to every other

Making accurate measurement and mathematical modelling of its rotational dynamics & mechanics incorrect without the correct physical topologies

Electron modelling & probability calculations

12
[0-12]
RE
KE
S

Electron positional measurements have proven to be historically difficult to accurately model due to the charged 12π rotating topologies of leptons

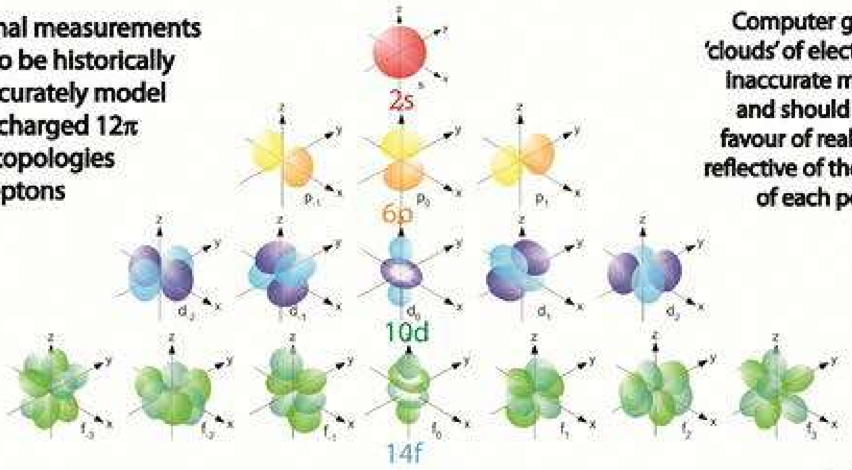

Computer generated plots of 'clouds' of electron probabilites are inaccurate mis-representations and should be abandoned in favour of realistic atomic models reflective of the charged geometry of each periodic element

RE
KE
P
KE
hv
Q
M

Every elemental atom can be viewed as a quantum carousel with a unique number of oscillating fans positioned around it.

Each fan has 3 blades and a fixed speed n[1-8] related to its height above ground level, AND the carousel is turning around on its axis

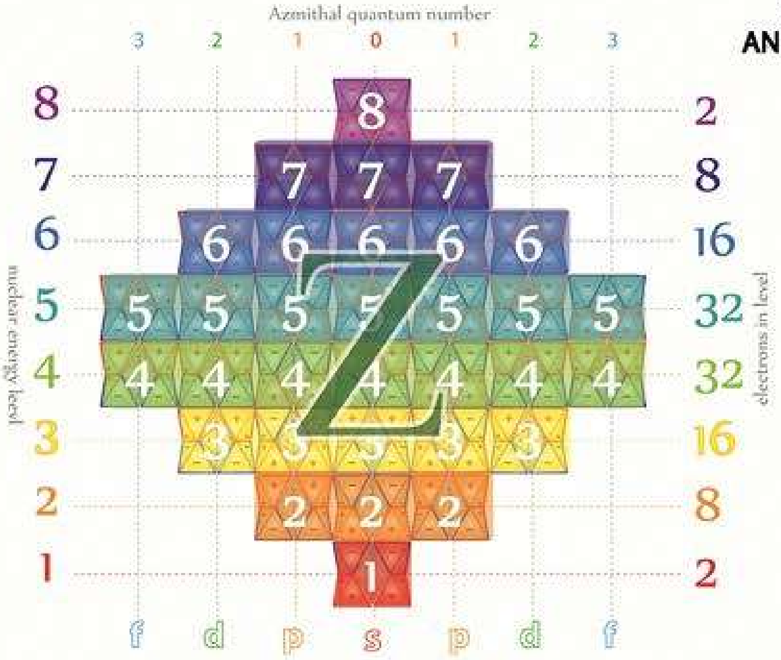

All periodic elements are made form Deuterium nuclei

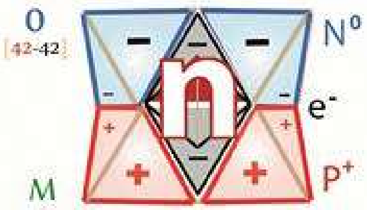

The baryonic energies of nuclei determines the energies of bound photo-electrons

The current computer generated electron probability diagrams in popular use at present can now be show to be a misrepresentative model of mathematical modelling of electron sub-orbital energies

The nuclear quantum levels [n], intrinsic quantised angular momentum [*h*] and orbital angular momentum [*l*] of each electron bound within atomic nuclei are all the direct result of the Baryonic energies of the nuclei they are bound to

Each level of the quantum carousel can contain only a limited number of fans each running at a specific speed

Imagine trying to measure (or model) the motion of any 1 quantum scale blade while the carousel rotates

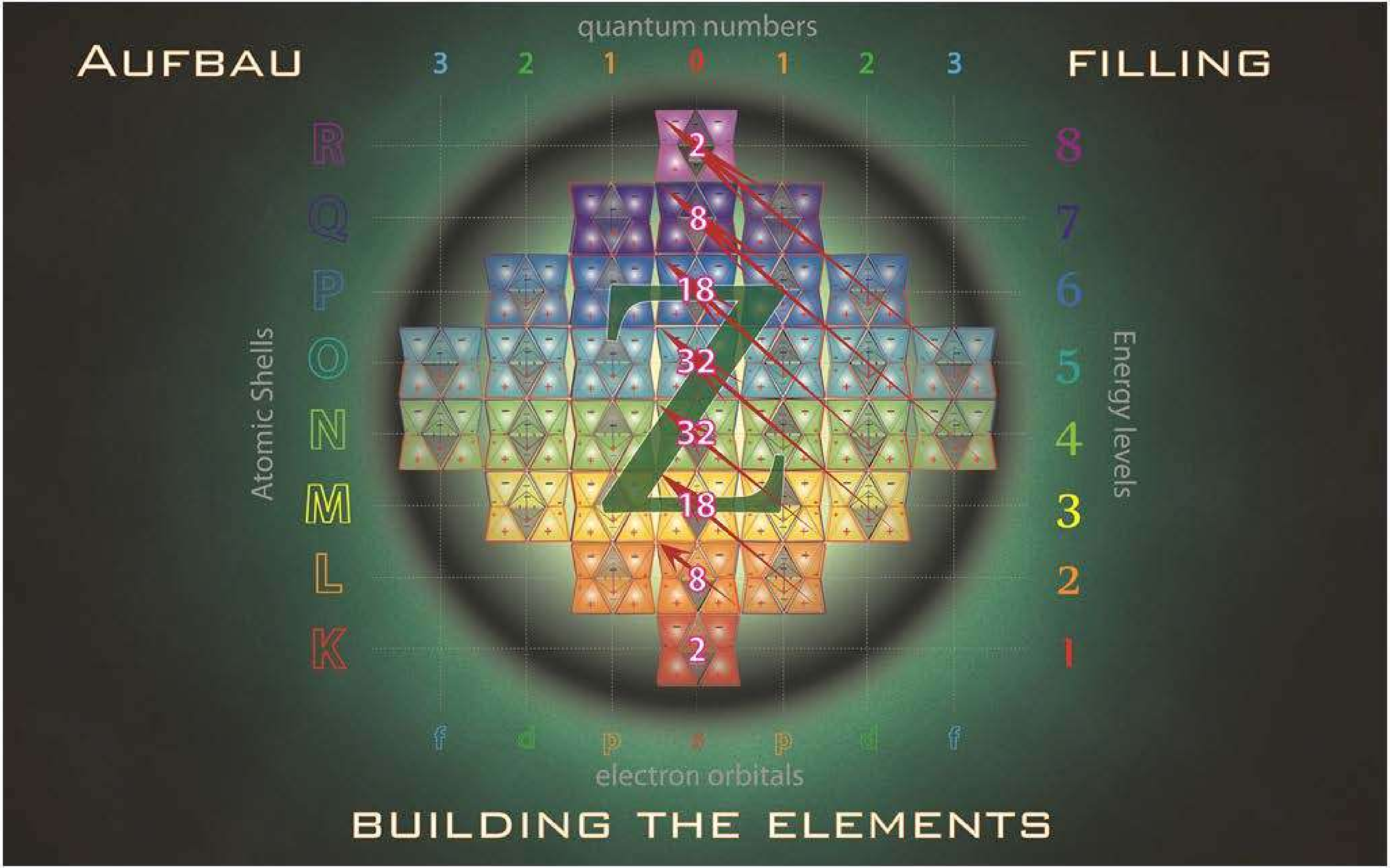
AUFBAU
FILLING
quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Atomic Shells
8
7
6
5
4
3
2
1
Energy levels
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
electron orbitals
BUILDING THE ELEMENTS

Tetryonics 48.01 - aufbau geometry

Copyright ABRAHAM [2008] - All rights reserved

Quantum Topologies

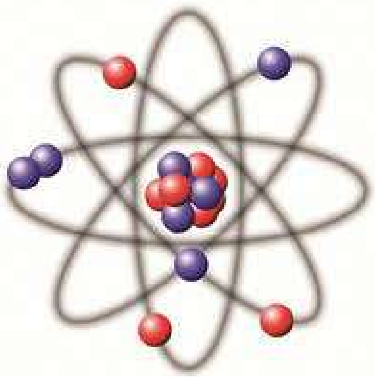

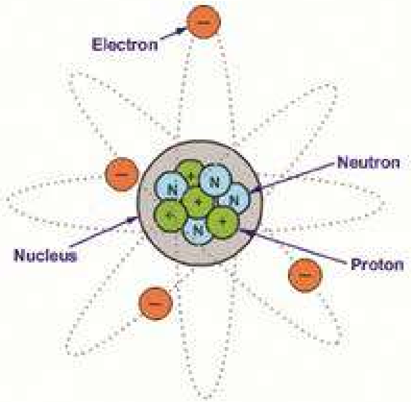
Electron
Neutron
Nucleus
Proton

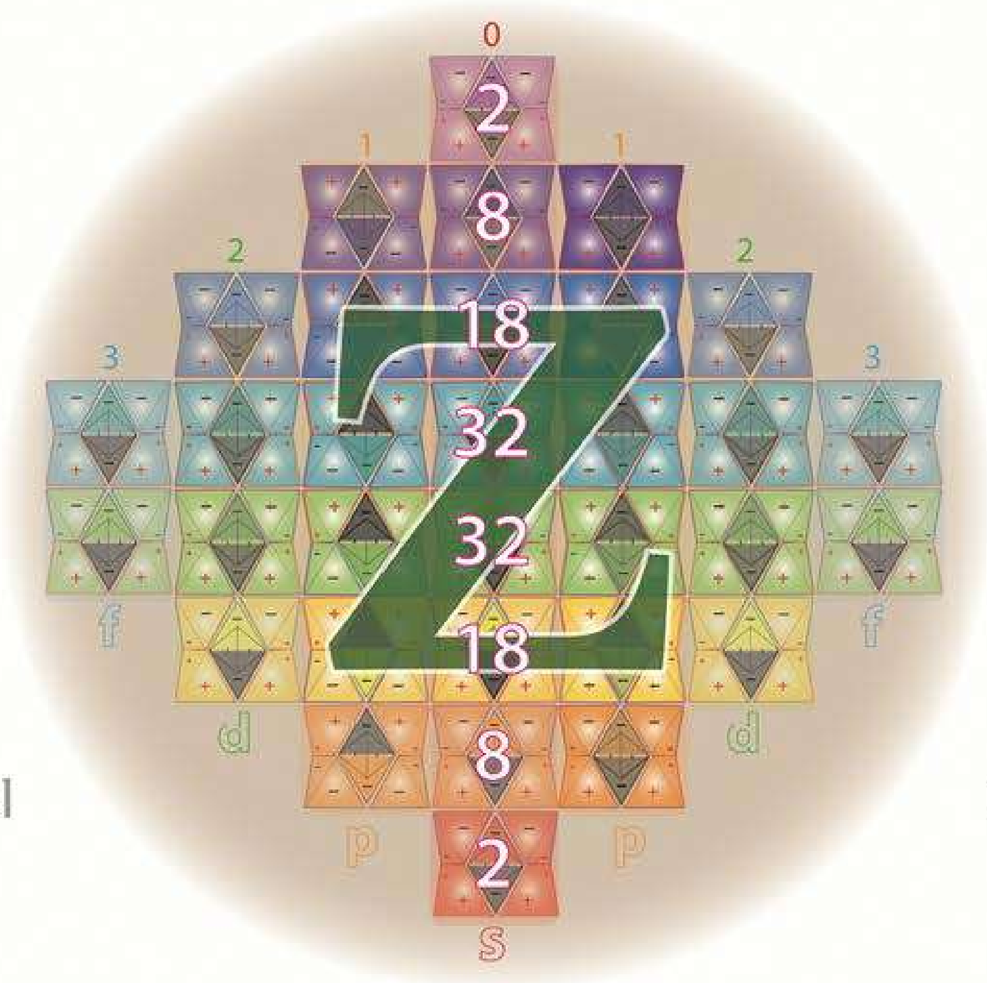
0
1
1
2
2
3
3
2
8
18
32
32
18
8
2
Z
f
f
d
d
p
p
s

Dalton Model

Quantum Model

Thomson Model

Rutherford Model

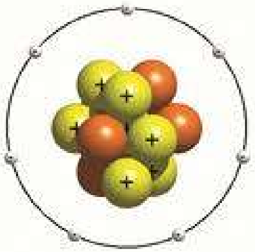

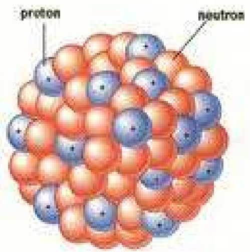
proton
neutron

Historically viewed as a spherical object
Tetryonic charge geometry has finally revealed the
true quantum topology of all atoms

Element numbers

The rule dictating how many nuclei form each Atomic shell is known as the Aufbau principle.

The physical and chemical properties of elements is determined by the atomic structure.
The atomic structure is, in turn, determined by the electrons and
which shells, subshells and orbitals they reside in.

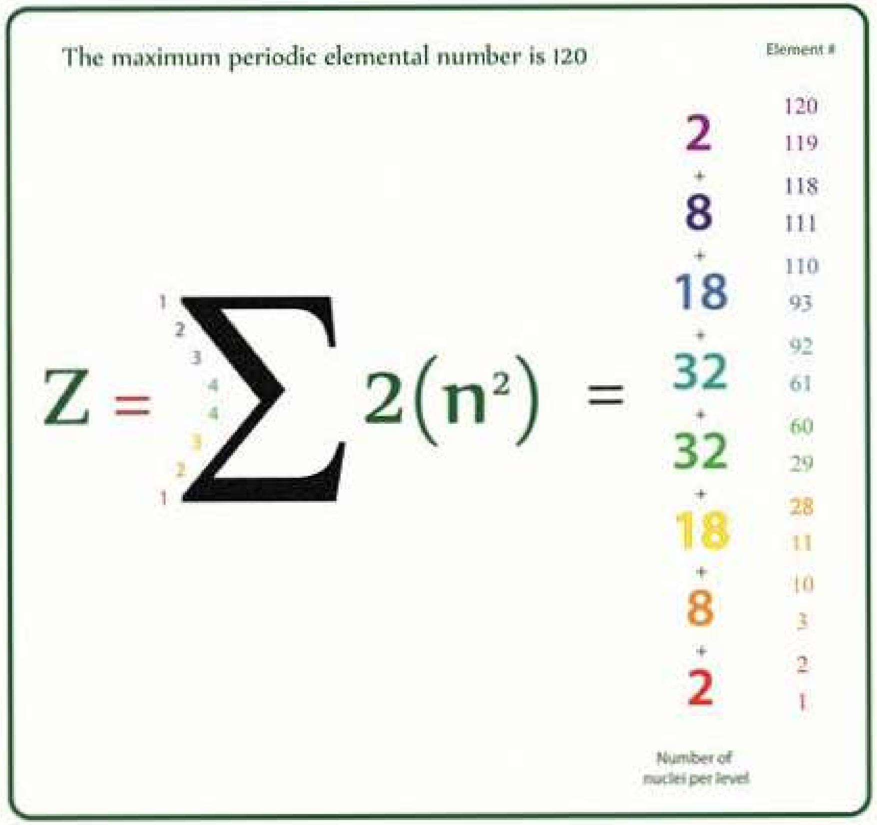

The maximum periodic elemental number is 120

$$Z = \sum 2(n^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

1
2
3
4
4
3
2
1
Element #
120
119
118
111
110
93
92
61
60
29
28
11
10
3
2
1
Number of nuclei per level

The number of nuclei per quantum level is reflective of photonic energy levels and provides the foundational geometry for all of the periodic elements

The number of possible nuclei in each Quantum level follows aufbau principle 'numbers' which can be determined using the following summation formula

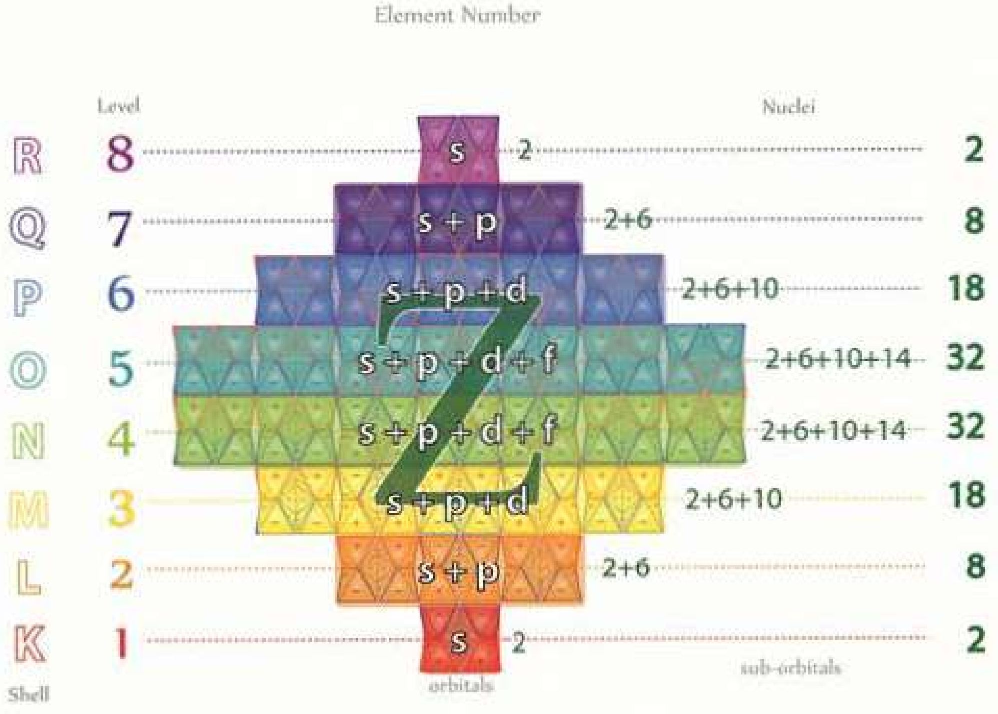

Element Number
Level
Nuclei
R 8 s 2 2
Q 7 s + p 2+6 8
P 6 s + p + d 2+6+10 18
O 5 s + p + d + f 2+6+10+14 32
N 4 s + p + d + f 2+6+10+14 32
M 3 s + p + d 2+6+10 18
L 2 s + p 2+6 8
K 1 s 2 2
Z
Shell
orbitals
sub-orbitals

Deuterium is the building block of all elements

Each element has equal numbers of Protons, electron & Neutrons with their stored mass-energies making up the molar masses of elements not excess neutron as currently modelled

Aufbau Principle
(Nuclei number and position)

1 Sub-Orbit (2 electrons max) ○ s

3 Sub-Orbits (6 electrons max) ○ p

○ d 5 Sub-Orbits (10 electrons max)

○ f 7 Sub-Orbits (14 electrons max)

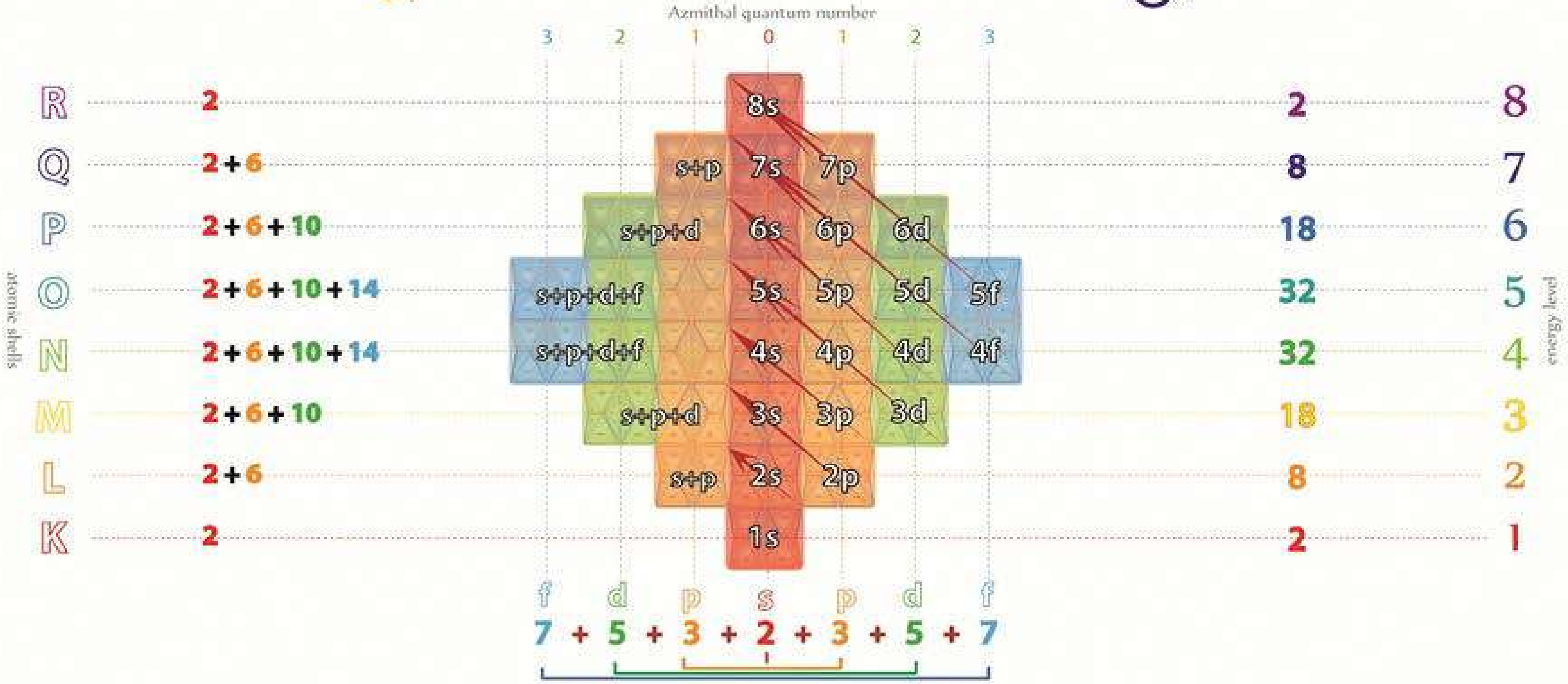
Azmithal quantum number
3 2 1 0 1 2 3
R 2
Q 2+6
P 2+6+10
O 2+6+10+14
N 2+6+10+14
M 2+6+10
L 2+6
K 2
atomic shells
8s
s+p 7s 7p
s+p+d 6s 6p 6d
s+p+d+f 5s 5p 5d 5f
s+p+d+f 4s 4p 4d 4f
s+p+d 3s 3p 3d
s+p 2s 2p
1s
2 8
8 7
18 6
32 5
32 4
18 3
8 2
2 1
energy level
f d p s p d f
7 + 5 + 3 + 2 + 3 + 5 + 7

	2		6								10														14									
Nuclei	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	per level	
R	119	120																															8	2
Q	87	88	113	114	115	116	117	118																									7	8
P	55	56	81	82	83	84	85	86	103	104	105	106	107	108	109	110	111	112															6	18
O	37	38	49	50	51	52	53	54	71	72	73	74	75	76	77	78	79	80	89	90	91	92	93	94	95	96	97	98	99	100	101	102	5	32
N	19	20	31	32	33	34	35	36	39	40	41	42	43	44	45	46	47	48	57	58	59	60	61	62	63	64	65	66	67	68	69	70	4	32
M	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	29	30															3	18
L	3	4	5	6	7	8	9	10																									2	8
K	1	2																															1	2
Shells	s1	s2	p1	p2	p3	p4	p5	p6	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	Quantum number	

sub-orbitals

A specific nuclear energy is associated with each electron configuration and, upon certain conditions, bound electrons are able to move from one orbital to another by emission or absorption of a quantum of energy, in the form of a photon.

The Aufbau principle can also be used to model the binding configuration of Protons and Neutrons in an atomic nucleus, according to their specific Baryonic Energy levels, in turn revealing the quantum geometry of all periodic Elements

Wolfgang Pauli

(25 April 1900 – 15 December 1958)

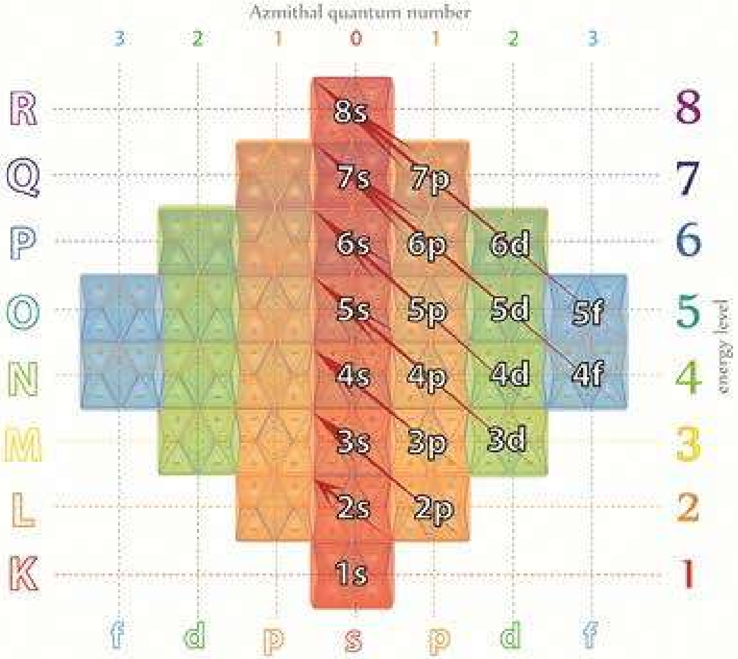
Azmithal quantum number
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8s
7s
7p
6s
6p
6d
5s
5p
5d
5f
4s
4p
4d
4f
3s
3p
3d
2s
2p
1s
8
7
6
5
4
3
2
1
energy level
f
d
p
s
p
d
f

Friedrich Hermann Hund

(4 February 1896 - 31 March 1997)

The orbitals of lower energy are filled in first with the electrons and only then are the higher energy orbitals filled.

aufbau electron orbital filling

Orbitals of equal energy are each occupied by one electron before any orbital is occupied by a second electron, and all electrons in singly occupied orbitals must have the same spin state

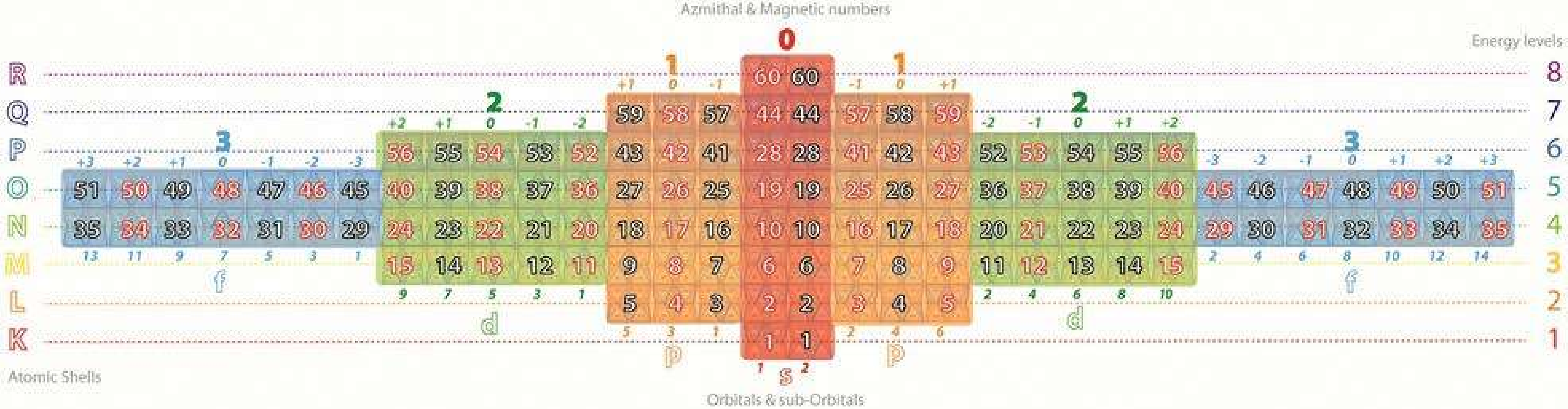
Azmithal & Magnetic numbers
Energy levels
0
1
1
2
2
3
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
R 60 60 8
Q 59 58 57 44 44 57 58 59 7
P 56 55 54 53 52 43 42 41 28 28 41 42 43 52 53 54 55 56 6
O 51 50 49 48 47 46 45 40 39 38 37 36 27 26 25 19 19 25 26 27 36 37 38 39 40 45 46 47 48 49 50 51 5
N 35 34 33 32 31 30 29 24 23 22 21 20 18 17 16 10 10 16 17 18 20 21 22 23 24 29 30 31 32 33 34 35 4
M 15 14 13 12 11 9 8 7 6 6 7 8 9 11 12 13 14 15 3
L 5 4 3 2 2 3 4 5 2
K 1 1 1
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f
d
p
s
p
d
f
Atomic Shells
Orbitals & sub-Orbitals

Aufbau construction

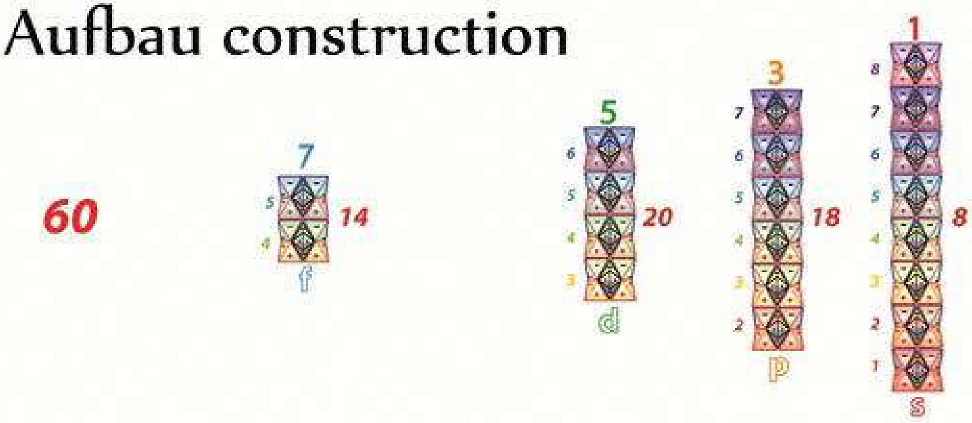
60
7
14
f
5
20
d
3
18
p
1
8
s

Proton
base
Hemisphere

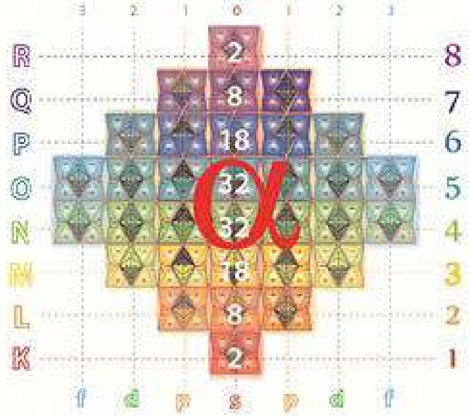
α
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
f
d
p
s
p
d
f

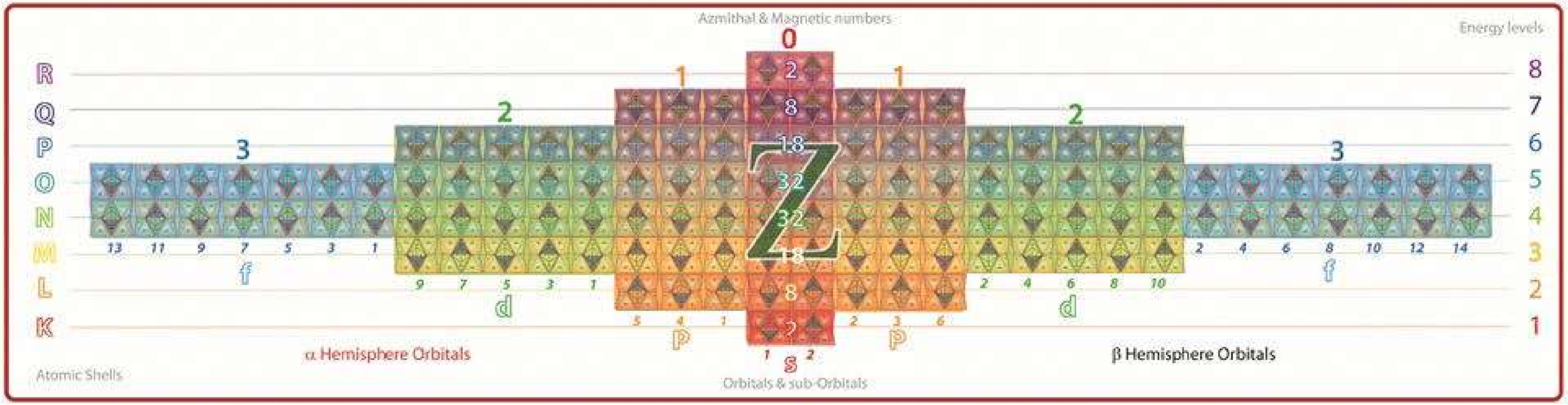
Azmithal & Magnetic numbers
Energy levels
0
1
1
2
2
3
3
Z
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
f
d
p
s
p
d
f
α Hemisphere Orbitals
β Hemisphere Orbitals
Atomic Shells
Orbitals & sub-Orbitals

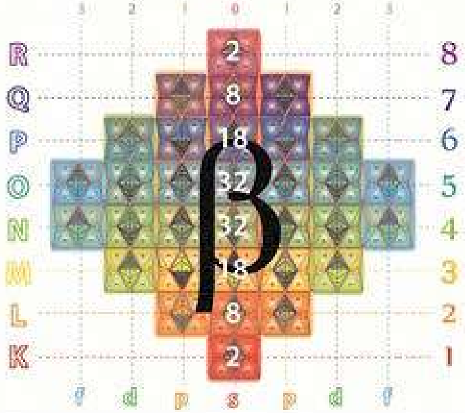
β
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
f
d
p
s
p
d
f

Neutron
base
Hemisphere

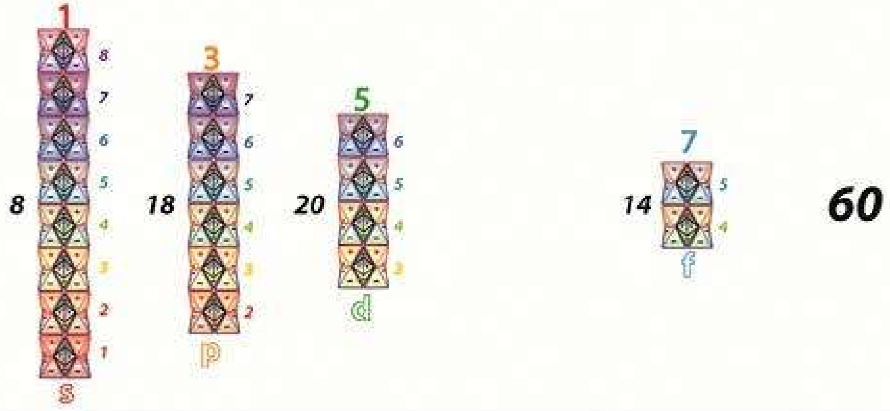
1
8
s
3
18
p
5
20
d
7
14
f
60

Copyright ABRAHAM [2008] - All rights reserved

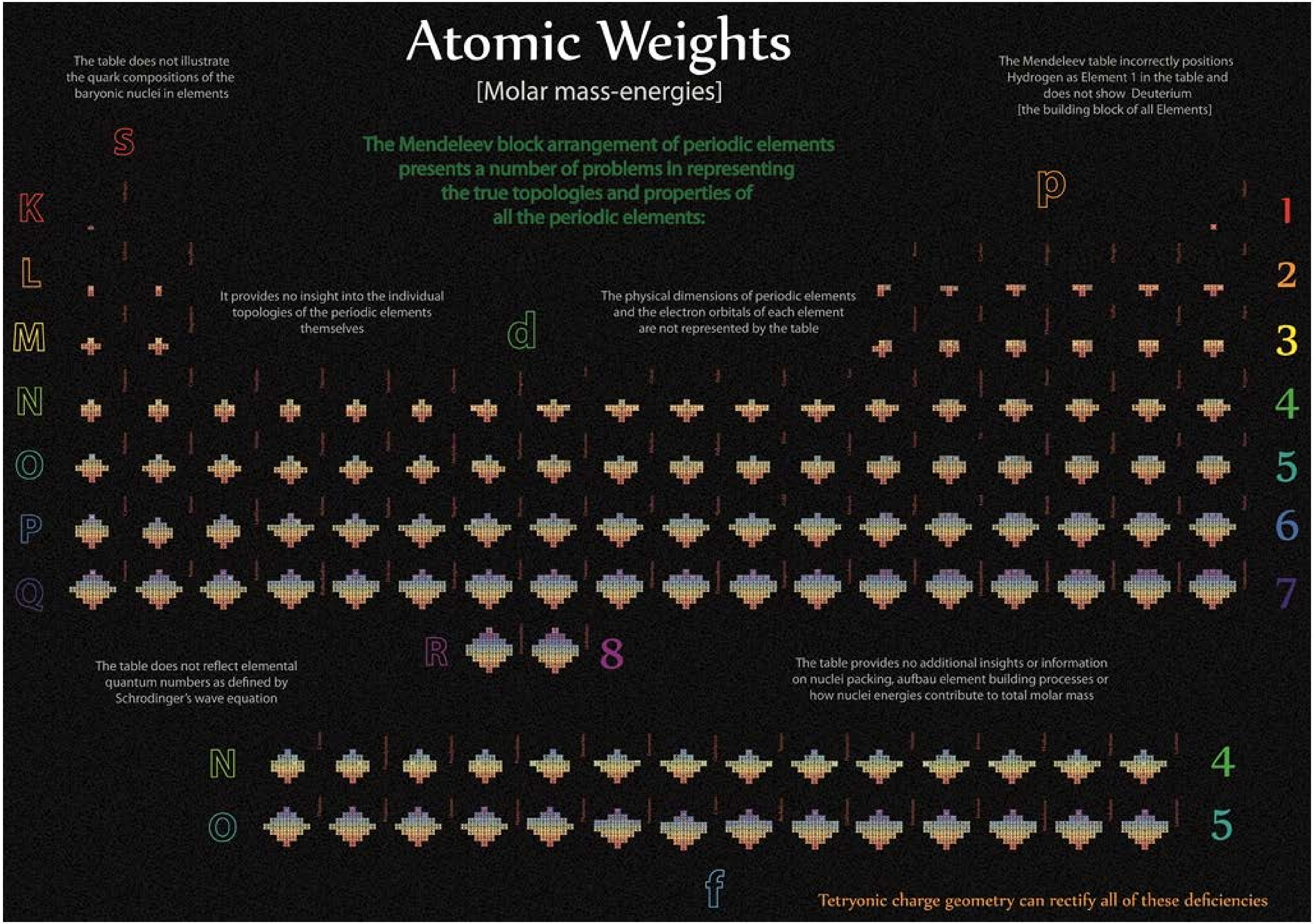

Atomic Weights

[Molar mass-energies]

The table does not illustrate the quark compositions of the baryonic nuclei in elements

The Mendeleev table incorrectly positions Hydrogen as Element 1 in the table and does not show Deuterium [the building block of all Elements]

The Mendeleev block arrangement of periodic elements presents a number of problems in representing the true topologies and properties of all the periodic elements:

It provides no insight into the individual topologies of the periodic elements themselves

The physical dimensions of periodic elements and the electron orbitals of each element are not represented by the table

The table does not reflect elemental quantum numbers as defined by Schrodinger's wave equation

The table provides no additional insights or information on nuclei packing, aufbau element building processes or how nuclei energies contribute to total molar mass

Tetryonic charge geometry can rectify all of these deficiencies

Tetryonics 49.01 - Atomic Weight

E

The quantum level mass-energies of Baryons determines the kinetic energies of electrons

1n - 1 e19v

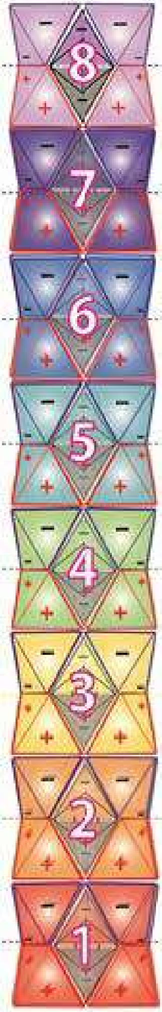

Baryon mass-energies n			electron KEM field hv^2
73,728	n32	v8	3.268969693 e15
69,192	n31	v7	2.502804921 e15
64,800	n30	v6	1.838795452 e15
60,552	n29	v5	1.276941286 e15
56,448	n28	v4	8.172424232 e14
52,488	n27	v3	4.596988631 e14
48,672	n26	v2	2.043106058 e14
45,000	n25	v1	5.107765145 e13
Compton frequencies 930 MeV	1.2e20 electron rest Matter is velocity invariant 496 keV		**spectral frequencies** 13.6 eV

Atomic nuclei mass-energies

Each element's weight [mass-Matter in a gravitational field] is the result of the total quanta comprising that element

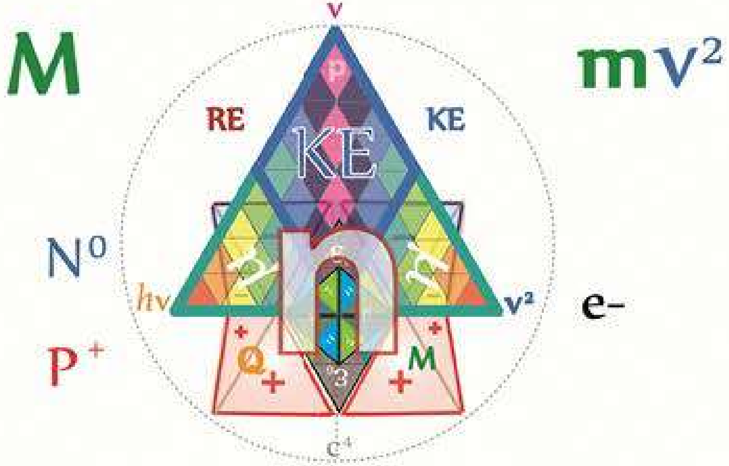

The nuclei forming each atomic shell have specific mass-energy quanta

Baryon rest masses — lepton rest mass — KEM

$$\overset{8}{\underset{1}{n}}\left[\,[72(n)^2]+[12e19]+[m_e v^2]\,\right]$$

Deuterium mass-energy per shell

Despite having differing mass-energies each Deuterium nuclei has the same velocity invariant Matter geometry [84π]

spin orbital coupling in synchronous quantum convertors

Electrons act as quantum scale rotating armatures in atomic nuclei and can only have specific energies reflective of the electron orbital energy level of the Baryons in which they are found

They acheive these energy levels by absorbing or emitting photons to acheive the specific angular momentum required

Copyright ABRAHAM [2008] - All rights reserved

Hydrogen Ionisation Energies [/n^2]

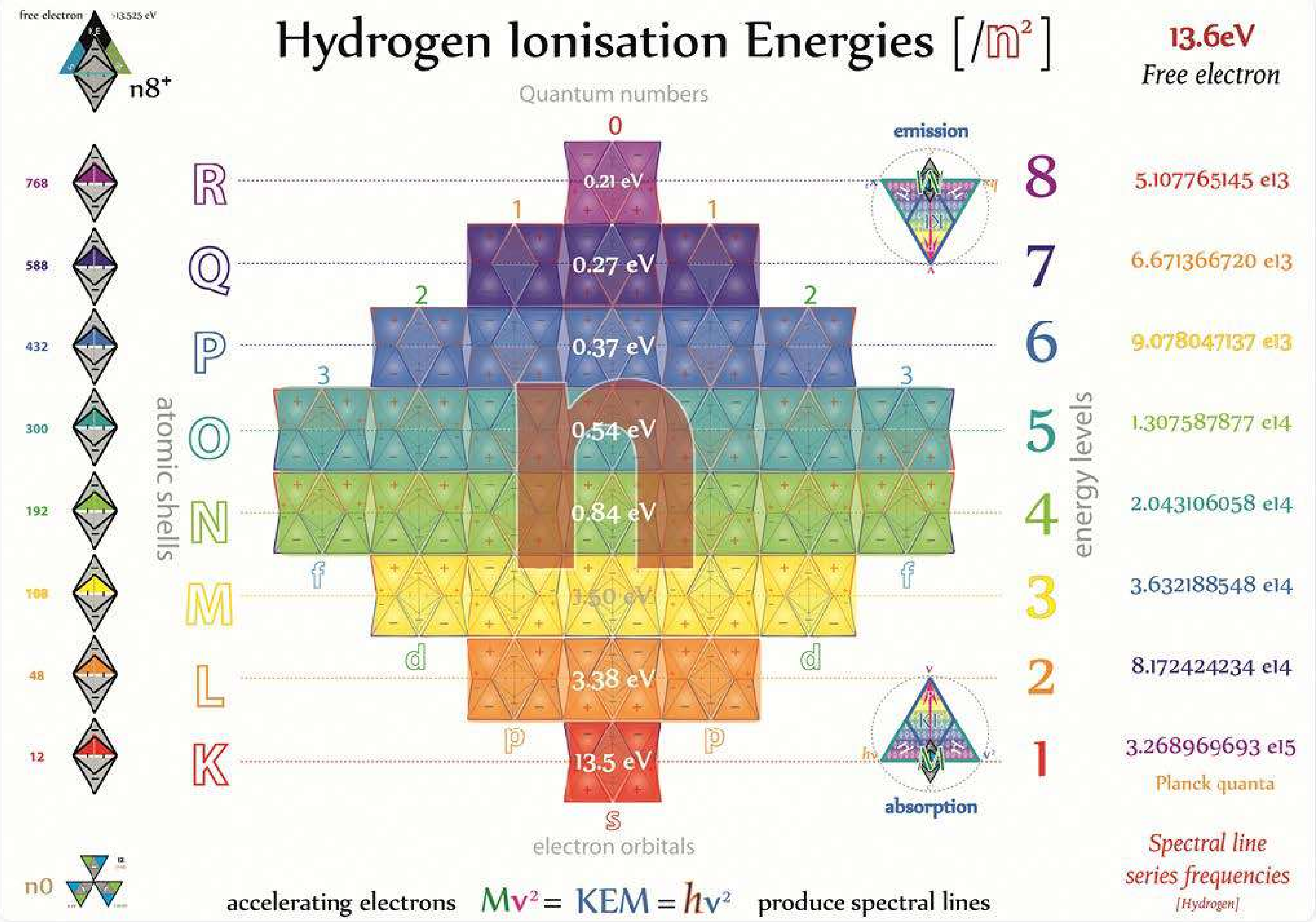
free electron
>13.525 eV
n8⁺
13.6eV
Free electron
Quantum numbers
0
1
1
2
2
3
3
f
f
d
d
p
p
s
electron orbitals
emission
absorption
atomic shells
energy levels
768
588
432
300
192
108
48
12
n0
R
Q
P
O
N
M
L
K
0.21 eV
0.27 eV
0.37 eV
0.54 eV
0.84 eV
1.50 eV
3.38 eV
13.5 eV
n
8
7
6
5
4
3
2
1
5.107765145 e13
6.671366720 e13
9.078047137 e13
1.307587877 e14
2.043106058 e14
3.632188548 e14
8.172424234 e14
3.268969693 e15
Planck quanta
Spectral line
series frequencies
[Hydrogen]

accelerating electrons $Mv^2 = KEM = hv^2$ produce spectral lines

Redefining Atomic weights

Atomic weight (symbol: Ar) is a dimensionless physical quantity,
the ratio of the average mass of atoms of an element (from a given source) to
1/12 of the mass of an atom of carbon-12 (known as the unified atomic mass unit)

Carbon 12

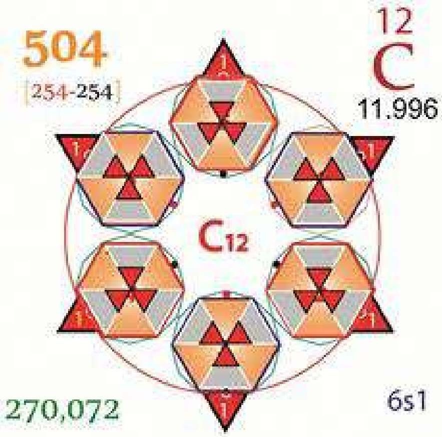

6 Protons	[24-12]	n1
6 Neutrons	[18-18]	
6 electrons	[0-12]	

The 'unified atomic mass unit' currently in use
is known to be inaccurate and must be corrected
in order to bring clarity & increased accuracy
to the atomic weights of all elements

A_r = 22,512
Hydrogen

Defining Hydrogen as having an exact
atomic Planck mass of 22512n quanta
provides uniformity with Tetryonics

1/12 C_{12}

A_r = 22,506

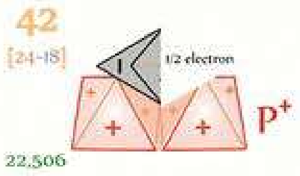

1.660096209e-27 kg

1 Proton	[24-12]	n1
.5 electron	0-6	

Unified atomic Matter unit

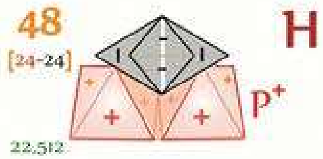

1.660538783e-27 kg

1 Proton	[24-12]	n1
1 electron	[0-12]	

Deuterium is the building block of all elements in the period table

A_r = 45,012
Deuterium

Defining Deuterium as having an exact
atomic Planck mass of 45012n quanta
reflects the true charged geometries
of all Elements & their topologies

D

1/6 C

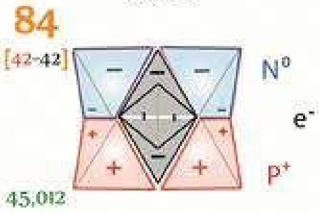

3.320192418e-27 kg

1 Proton	[24-12]	n1
1 Neutron	[18-18]	
1 electron	[0-12]	

Copyright ABRAHAM [2008] - All rights reserved

Planck mass-energy units

NAu = 0.001 kg

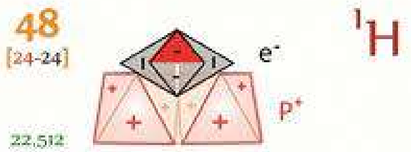
48
[24-24]
22,512
e-
P+
^{1}H

1.660538841 e-27 kg

One Da is approximately equal to the mass of one proton or one neutron

The unified atomic mass unit (symbol: u) or Dalton (symbol: Da) is a unit that is used for indicating mass on an atomic or molecular scale

270,072

1/12 the mass of a C12 graphene atom at rest in its electronic ground state

1.660538782(83)×10−27 kg

22,506

is an inaccurate means of determining the exact rest mass of a Hydrogen atom

22,512

Carbon 12 has 270,072n planck quanta (270,072 / 12 = 22,506)

Hydrogen has a mass of 22,512n (22,500+12) requiring all mass to be calculated directly using the Planck mass-energy quantum (.001kg / N_A / 22,512) & Tetryonic charge geometries

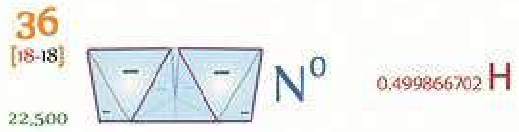
36
[18-18]
22,500
N^0
0.499866702 H

1.659653693 e-27 kg

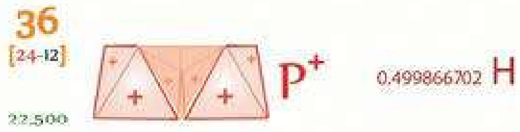
36
[24-12]
22,500
P^+
0.499866702 H

Deuterium is the building block of all elements

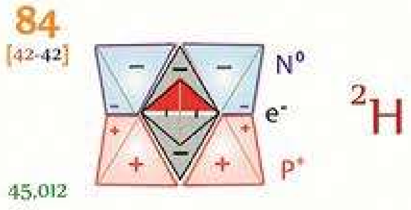
84
[42-42]
45,012
N^0
e-
P^+
^{2}H

3.320192534 e-27 kg

Using Tetryonic theory to define
n Planck mass = 7.376238634 x 10^{-32} Kg
(see Tetryonics QM 15.04)

exact atomic rest masses for all particles, elements and compounds can be determined directly from atomic theory

N_A = 6.02214179 e23

The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12; its symbol is "mol".

NA(12u) = 0.012 kg

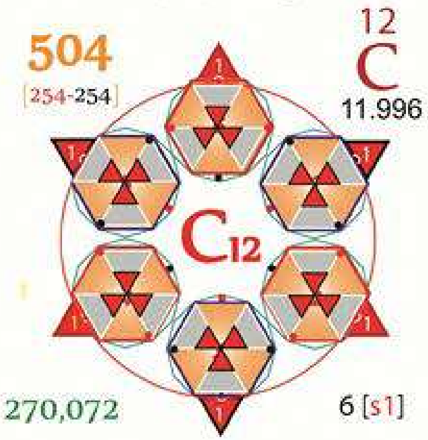
504
[254-254]
12
C
11.996
C12
270,072
6 [s1]

6 Protons [24-12]
6 Neutrons [18-18]
6 electrons [0-12]
n1

Carbon has a number of differing atomic configurations (allotropes)

6 Protons [24 12]
6 Neutrons [18-18]
6 electrons [0-12]
n1-2

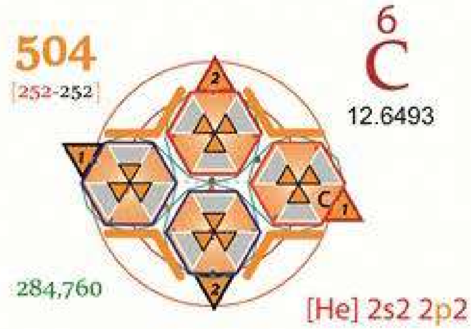
504
[252-252]
6
C
12.6493
284,760
[He] 2s2 2p2

Planck mass-energy contributions to the measured weights of periodic, elementary mass-Matter topologies

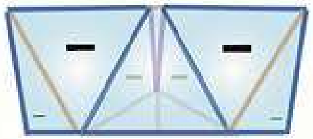

1.659653693 e-27 kg

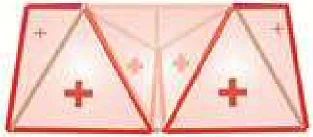

Baryons have 2,25e23 Planck quanta comprising their rest Matter topologies

[930.974 MeV]

Δ 1875 x

[496.5 keV]

8.851486361 e-31 kg

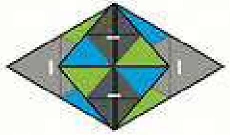

Leptons have 1.2 e20 Planck quanta comprising their rest Matter topologies

[496.5 keV]

Δ 36,711 x

[13.6 eV]

2.411109611 e-35 kg

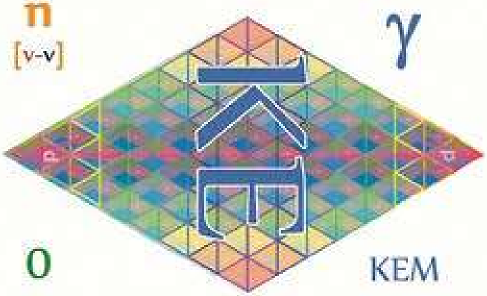

Photons are planar geometries [Matter-less] (purely Kinetic mass-Energy and momenta)

The Lyman alpha spectral line mass-energy contribution to the mass of a Deuterium nucleus is negligible

Electron quantum level energies are determined by the energy of the Nuclei they bind to in elements

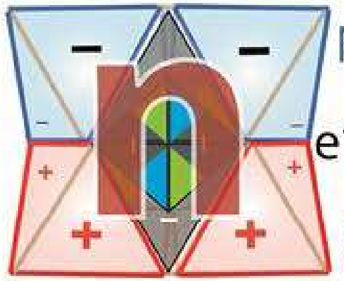

Photons contribute spectral mass-energies to the nuclei mass but are themselves Matterless [2D zero rest mass-energies]

Photons are 2π charge mass-energy geometries

Ionisation energies

Fractional quantum differentials

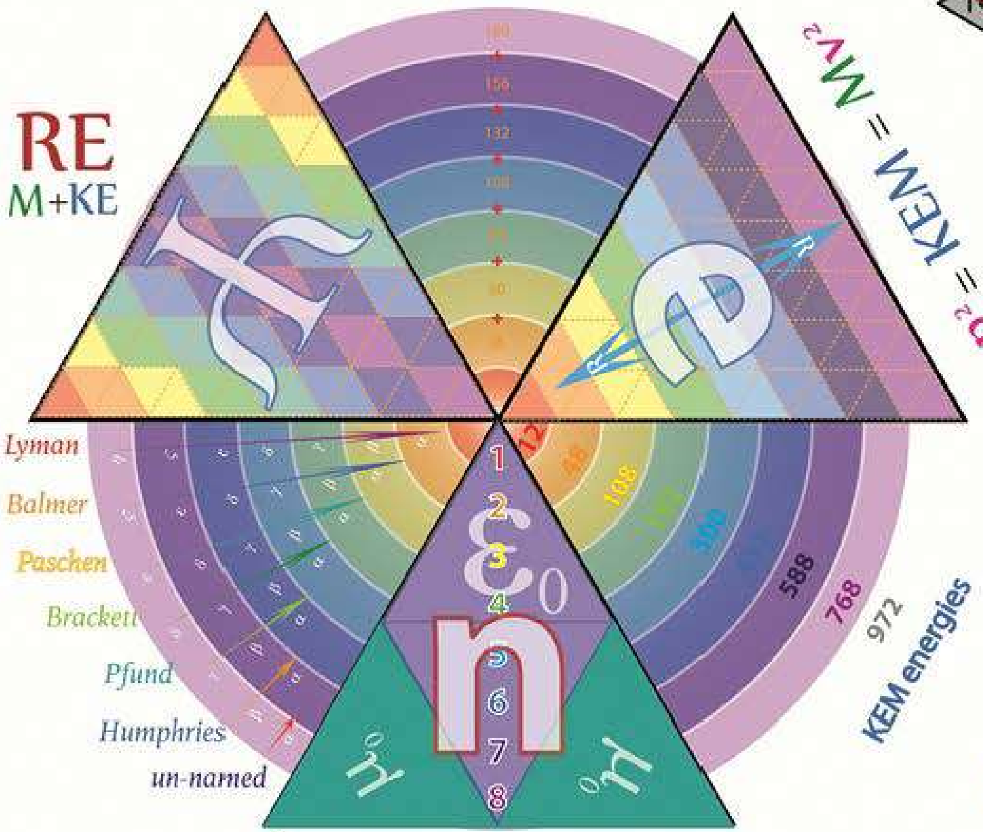
RE
M+KE
$p^2 = KEM = Mv^2$
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
12
48
108
192
300
432
588
768
972
KEM energies
1
2
3
4
5
6
7
8
ε_0
n

$$hf = \Delta M v = \Delta p$$

spectral lines are produced by accelerating electrons.

Note: this is an illustrative schema for modelling KEM field energies
All KEM fields possess the same physical spatial geometry
in radial-time defined spatial co-ordinate systems

c^2 c^4

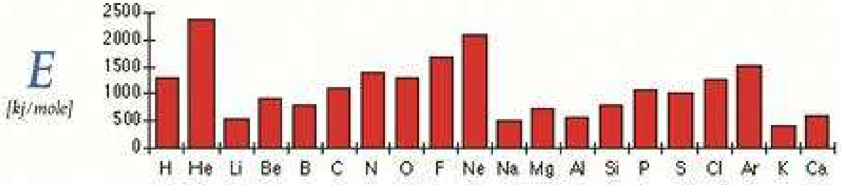

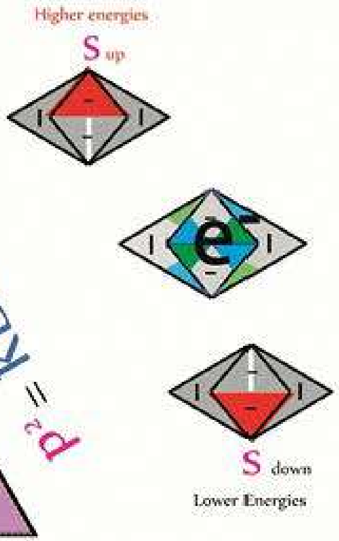
Higher energies
S up
e
S down
Lower Energies

Photo-electrons absorb/emit spectral energies

$$E = eV = \frac{1}{4\pi\epsilon_0}\frac{ne^2}{a}$$

photo-electronic energy momenta

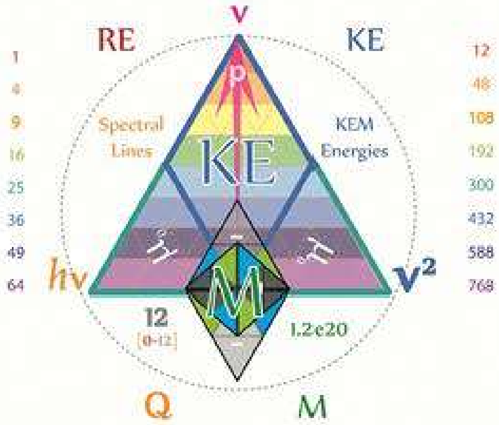
RE
KE
v
p
KE
Spectral Lines
KEM Energies
hv
v^2
M
12
[0-12]
1.2e20
Q
M
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768

Mapping photo-electron transition energies
to Tetryonic energy momenta geometries
reveals many key facts about the
ionisation energies of nuclei

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

The differing fractional KEM field energy momenta of electrons
that results from their transitions to specific energy nuclei
in elements results in differing QAM quanta and
produces spectral lines and fine line splitting

Spectral line transitions

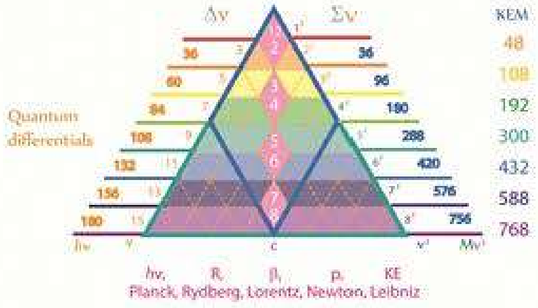
Δv
Σv
KEM
Quantum differentials
36
60
84
108
132
156
180
36
96
180
288
420
576
756
48
108
192
300
432
588
768
hv
v
c
v²
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Elementary ionisation energies

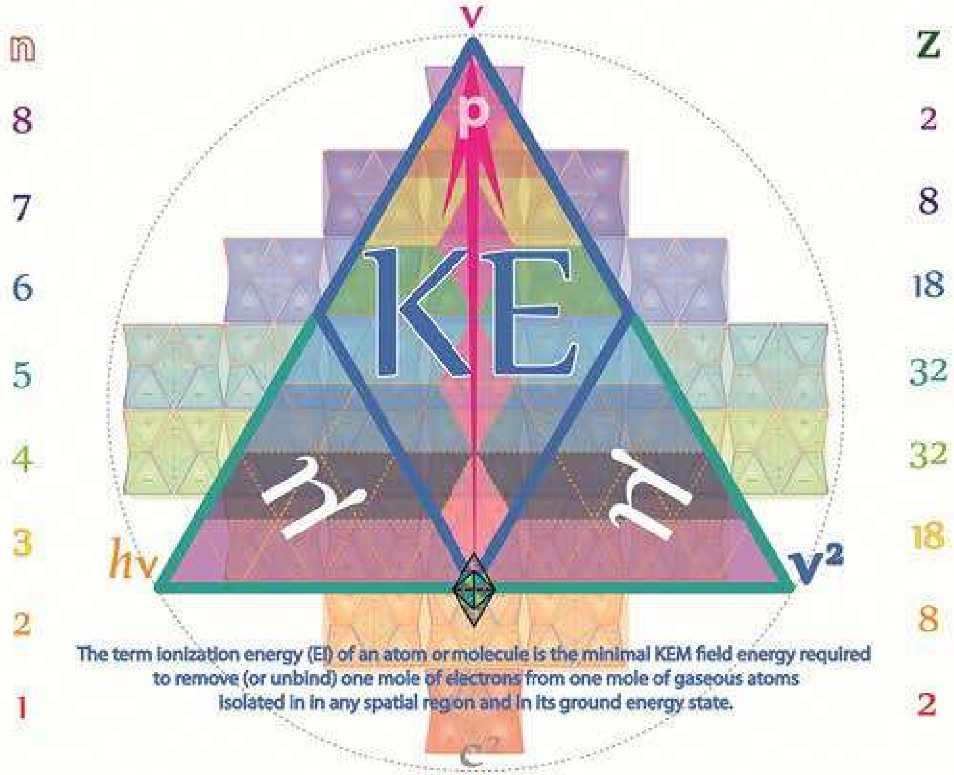

The term ionization energy (EI) of an atom or molecule is the minimal KEM field energy required to remove (or unbind) one mole of electrons from one mole of gaseous atoms isolated in in any spatial region and in its ground energy state.

The term "ionization energy" is sometimes used as a name for the work needed to remove (or un-bind) the highest energy photoelectron from an atom or molecule.

However, due to interactions with surfaces, this value differs from the ionization energy of the atom or molecule in question when it is located by itself in free space.

So, in the case of surface-adsorbed atoms and molecules, it may be better to use the more general term "electron binding energy", in order to avoid confusion.

Both these names are also sometimes used to describe the work needed to remove an electron from a "lower" orbital (i.e., not the topmost orbital) for both free and adsorbed atoms; in such cases it is necessary to specify the orbital from which the electron has been removed

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

Every electron in each elementary orbit has a unique ionisation energy

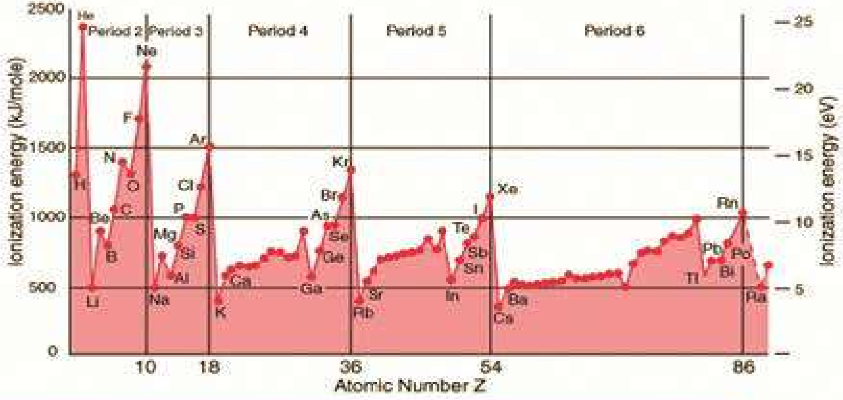

Element number

Nuclei	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	Energy level	per shell
R	119	120																															8	2
Q	87	88	113	114	115	116	117	118																									7	8
P	55	56	81	82	83	84	85	86	103	104	105	106	107	108	109	110	111	112															6	18
O	37	38	49	50	51	52	53	54	71	72	73	74	75	76	77	78	79	80	89	90	91	92	93	94	95	96	97	98	99	100	101	102	5	32
N	19	20	31	32	33	34	35	36	39	40	41	42	43	44	45	46	47	48	57	58	59	60	61	62	63	64	65	66	67	68	69	70	4	32
M	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	29	30															3	18
L	3	4	5	6	7	8	9	10																									2	8
K	1	2																															1	2
sub-orbitals	s1	s2	p1	p2	p3	p4	p5	p6	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14		

Shells

s

p

d

f

Orbitals

Hyperfine splitting and Lamb Shifts

When the spectral lines of the hydrogen spectrum are examined at very high resolution, they are found to be closely-spaced doublets.
This splitting is called fine structure and was one of the first experimental evidences for electron spin.

The fine structure describes the splitting of the spectral lines of atoms due to first order relativistic corrections [principal quantum energies]

n

Differing electron spins within shells & quantum levels produces Hyperfine splitting

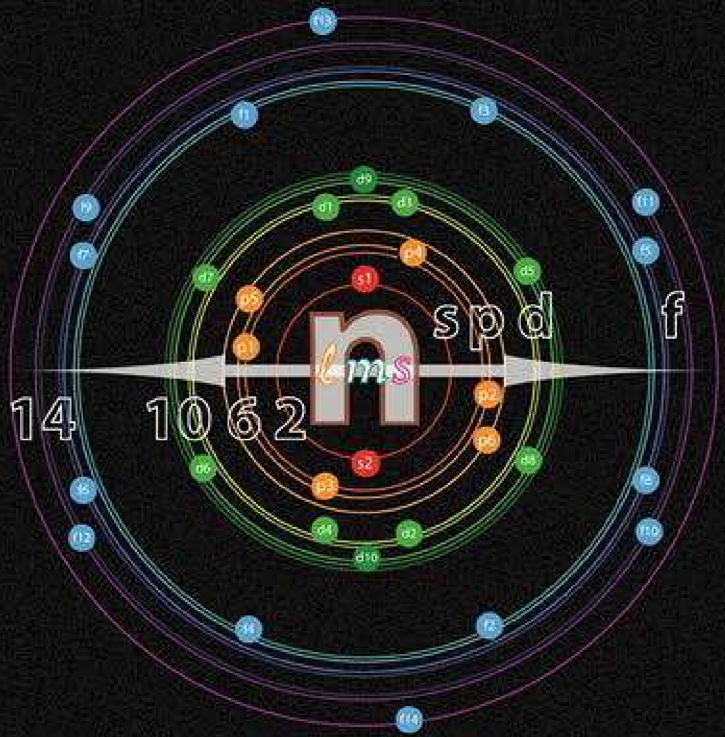

f13
f1
f3
f9
f7
f11
f5
d9
d1
d11
p4
d7
d5
s1
p5
s p d
f
p1
n
ℓms
14
10 6 2
p2
p6
d6
s2
d8
f6
f8
p3
f12
f10
d4
d2
d10
f4
f2
f14

The electron energy levels of Hydrogen should depend only on the principal quantum number n.

In 1951 it was discovered that this was not so, the $2p_{(1/2)}$ state is slightly less than the $2s_{(1/2)}$ state resulting in the Lamb shift

ℓ m s

Differing electron sub-Orbtial energies create Lamb Shifts

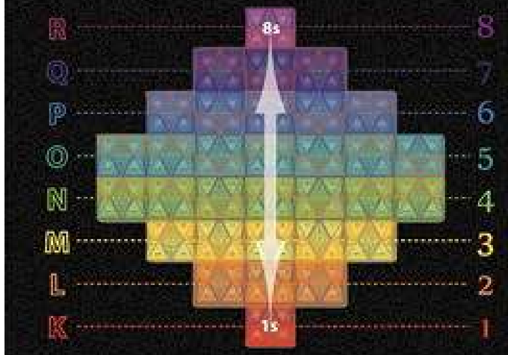

R
Q
P
O
N
M
L
K
8s
1s
8
7
6
5
4
3
2
1

If you measure the atomic energy levels of photo-electrons at an extremely high resolution, you'll find small deviations of individual KEM field energies of electrons in sub-orbitals which are primarily the result of parallel and anti-parallel electron spins

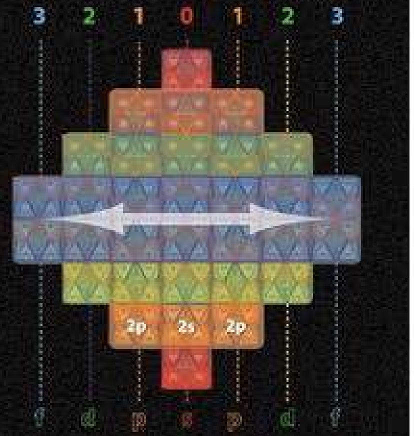

3
2
1
0
1
2
3
2p
2s
2p
f
d
p
s
p
d
f

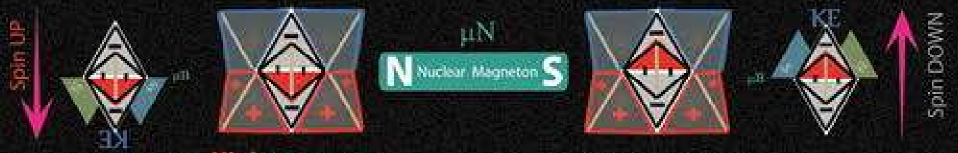

Spin UP
KE
μB
μN
N Nuclear Magneton S
μB
KE
Spin DOWN

atomic shell energies result from Series addition of baryonic energies

Higher energy
Parallel
Magnetic moments

Lower energy
Anti-parallel
Magnetic moments

electron sub-orbital energies are Parallel energy configurations

All electron spins are referenced to the Nuclear magneton

M_0

3D Matter topologies are comprised of charged 2D mass-energies

Energy per second^2

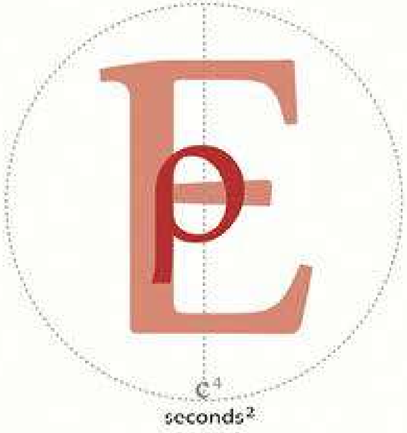

c^4
seconds^2

Matter Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-Matter

standing wave mass-energiescreate
the material substance of all chemical elements

RE

Relativistic mass-ENERGY-Matter

Relativity fails at the foundational level to explain and differentiate between mass-ENERGY and Matter in physical systems

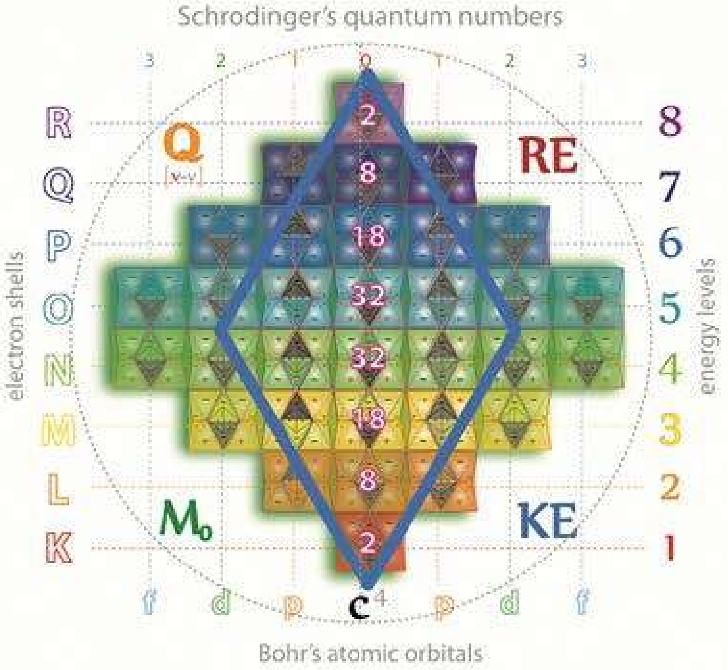

Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M0
KE
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f
d
p
c4
p
d
f
Bohr's atomic orbitals

Einstein's relativistic [Lorentz corrected] stress energy tensor aggregates all forms of energy into a single energy density gradient

atomic energies $\frac{Tmn}{c^4} + \frac{mv^2}{c^2}$ *electron spins*

3D rest Matter + Lorentz corrected 2D Kinetic Energies = total Relativistic Energies

KE

2D equilateral mass-energies are euclidean geometries

Energy per second

mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energies

radiant planar mass-energies create
EM fields, spectral lines & chemical interactions

Deuterium is the building block of all Elements (save Hydrogen)

The geometry of any Element is determined by its Charge

The Molar Weight of any Element is a measure of its standing wave mass-energies

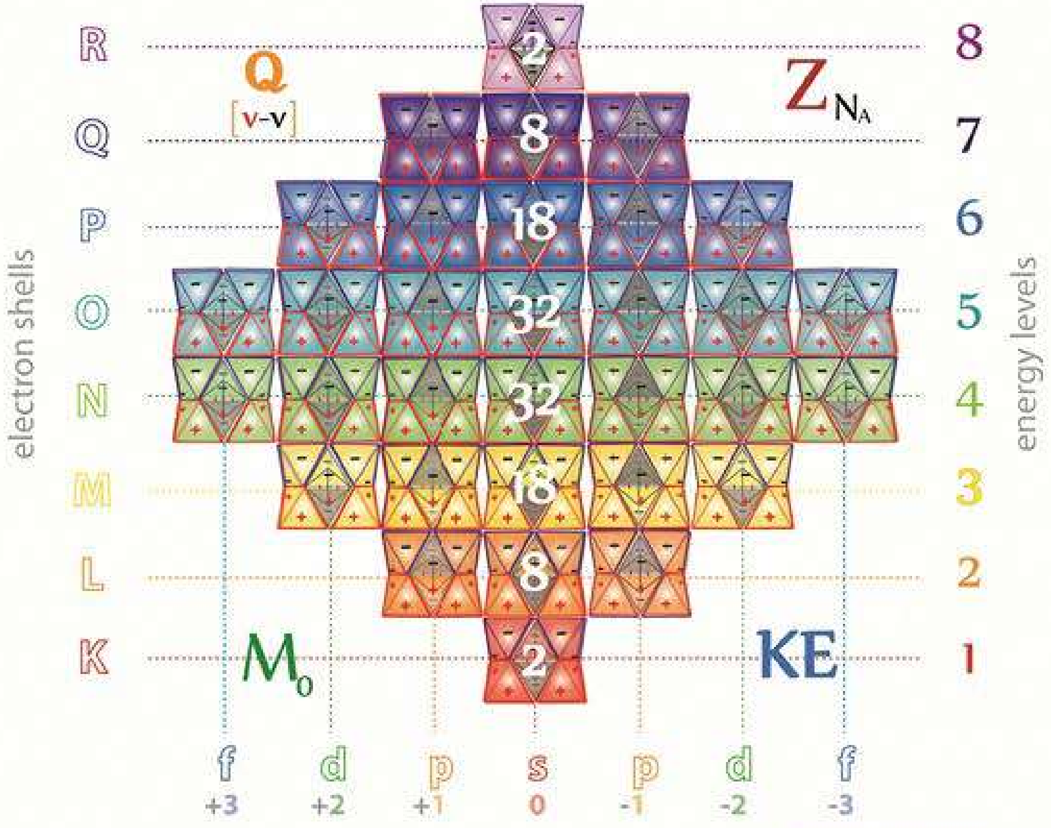

The rest mass-Matter of any Element is determined by the total number of quanta making up the Protons, Neutrons and electrons that comprise them (in their respective energy levels)

Any mass-energies in excess of the molar [n1] weight is a measurement of a element's CHEMICAL energies

Important point to note:

The Kinetic Energy difference between any Element's total [n1] Deuteron mass-energies and its Molar mass has historically been incorrectly explained as resulting from an excess number of Neutrons in the atom it is not. Z# = (number of Protons = number of electrons = number of Neutrons)

Elementary mass-Matter

$$\sum_{K}^{R}$$

	n per nuclei 1e19v = n	
2	2 nuclei [74,496 ea]	120 Unbinilium 119 Ununennium
	+	
8	8 nuclei [69,780 ea]	118 Ununoctium 87 Francium
	+	
18	18 nuclei [65,232 ea]	112 Copernicium 55 Caesium
	+	
32	32 nuclei [60,852 ea]	102 Nobelium 37 Rubidium
	+	
32	32 nuclei [56,640 ea]	70 Ytterbuim 19 Potassium
	+	
18	18 nuclei [52,596 ea]	30 Zinc 11 Sodium
	+	
8	8 nuclei [48,720 ea]	10 Neon 3 Lithium
	+	
2	2 nuclei [45,012 ea]	2 Helium 1 Deuterium

The rest mass-Matter of any Element is the sum total of its constituent Z[n²] energy level Deuterium nuclei

Aufbau

$$Z\#\left[\begin{array}{ll} z \text{ Protons} & [24\text{-}12] \\ z \text{ Neutrons} & [18\text{-}18] \\ z \text{ electrons} & [0\text{-}12] \end{array}\right] n1\text{-}8$$

(ie Calcium [20] = 2+8+10 n level Deuterium nuclei)

Periodic Harmonic motions

$x = A \cos(\omega t + \varphi)$

Circular motion

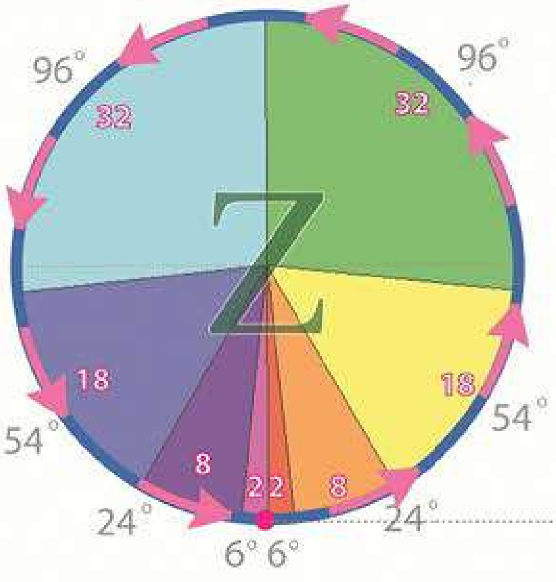
96°
32
96°
32
Z
18
18
54°
54°
8
2 2
8
24°
24°
6° 6°

Much of the math in of modern physics
is predicated on the assumption that
π [where it appears] is related to
the properties of a circle

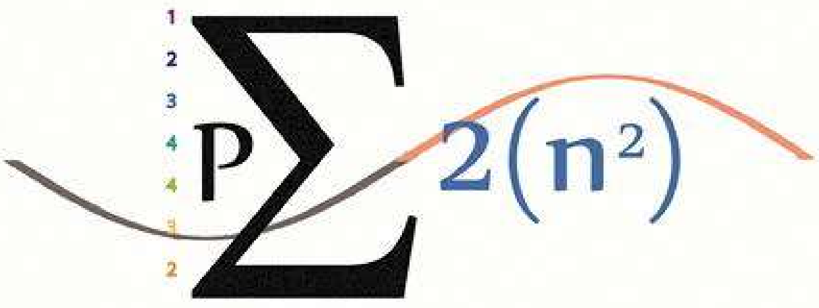

$P\sum 2(n^2)$

$F = -kx$

Linear motion

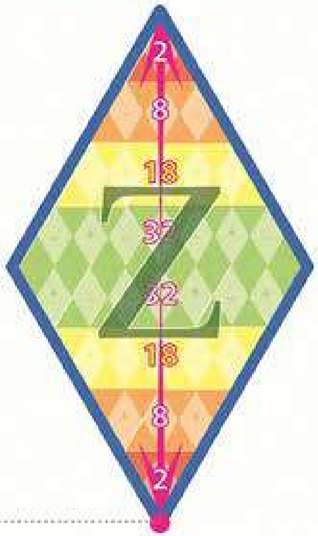
2
8
18
32
Z
32
18
8
2

Simple harmonic motion can be visualized as the projection of uniform circular motion onto one axis

circular harmonic motion

simple harmonic motion

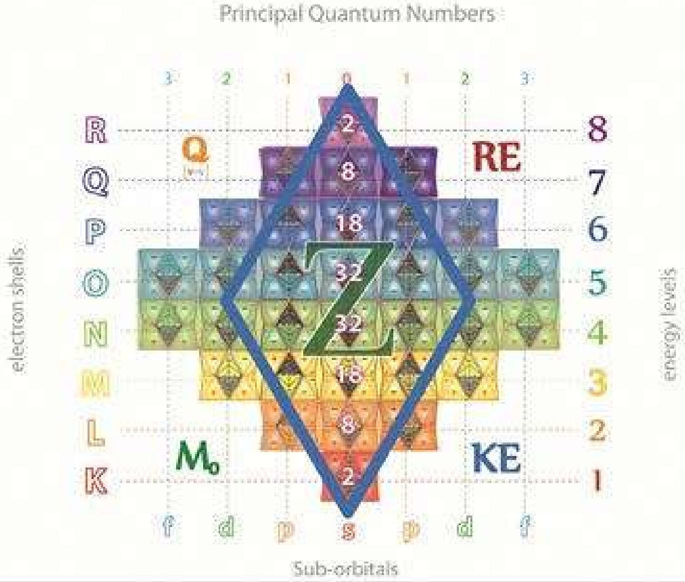
Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
2
8
18
32
Z
32
18
8
2
M₀
KE
8
7
6
5
4
3
2
1
electron shells
energy levels
f
d
p
s
p
d
f
Sub-orbitals

**Circular motions describe
the motion of a body
with a changing velocity vector
[the result of an acceleration force].**

**Nuclei per shell in elements follows
a 'periodic summation rule'
that is reflective of
photonic energies**

Copyright ABRAHAM [2008] - All rights reserved

$$\sum_{K}^{R}$$

nuclei per shell

$${}_{P}\sum_{K}^{R} Z$$

Periodic Summation

Periodic summation is a notation developed for Tetryonic theory to model the geometric series addtion of Z[n²] energy level Deuterium nuclei that form the periodic elements

$$\sum_{1}^{120} Z$$

element number

STEP ONE

Periodic summation follows the atomic shell electron config

STEP TWO

Periodic elements build up following the aufbau sequence

1 R
2
3
4
4
3
2
1 K

$${}_{P}\sum_{K}^{R} 2(x^2) =$$

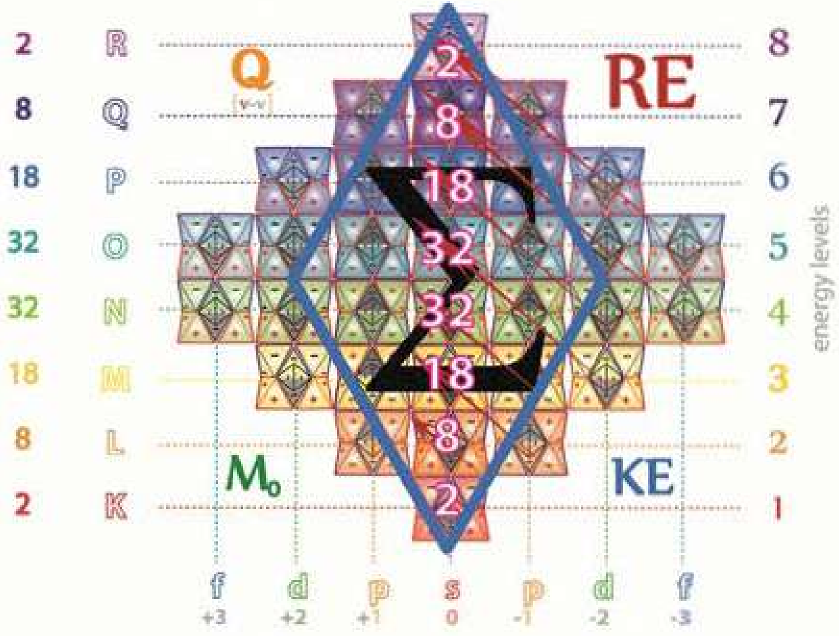

Each atomic shell can hold only a fixed number of deuterium nuclei

${}_{P}\sum R = 2$	2 nuclei [74,496 ea]	120	Unbinilium
	+		
${}_{P}\sum Q = 8$	8 nuclei [69,780 ea]	118	Ununoctium
	+		
${}_{P}\sum P = 18$	18 nuclei [65,232 ea]	110	Darmstadtium
	+		
${}_{P}\sum O = 32$	32 nuclei [60,852 ea]	92	Uranium
	+		
${}_{P}\sum N = 32$	32 nuclei [56,640 ea]	60	Neodymuim
	+		
${}_{P}\sum M = 18$	18 nuclei [52,596 ea]	28	Argon
	+		
${}_{P}\sum L = 8$	8 nuclei [48,720 ea]	10	Neon
	+		
${}_{P}\sum K = 2$	2 nuclei [45,012 ea]	2	Helium
		0	Hydrogen

THe LHS of the notation determine the number of nuclei in each atomic shell, from the periodic mass-energy levels for atoms, and the RHS follows the aufbau building principle to determine the rest mass-Matter of any specific element

Aufbau

Each periodic element is made of Z [n² energy] deuterium nuclei

Z# [z Protons [24-12]
z Neutrons [18-18]
z electrons [0-12]] n1-8

Planck mass-energies form the surface integral of rest Matter topologies for each periodic element

Copyright ABRAHAM [2008] - All rights reserved

Element numbers

Nuclei per shell in elements follow
a 'periodic summation rule'
that is reflective of
photonic energies

$$_{P}\sum_{K}^{R} 2(x^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

$$\begin{array}{l} 1\ R \\ 2 \\ 3 \\ 4 \\ 4 \\ 3 \\ 2 \\ 1\ K \end{array}$$

Z

120	Unbinilium	n8
119	Ununennium	
118	Ununoctium	n7
87	Francium	
112	Copernicium	n6
55	Caesium	
102	Nobelium	n5
37	Rubidium	
70	Ytterbuim	n4
19	Potassium	
30	Zinc	n3
11	Sodium	
10	Neon	n2
3	Lithium	
2	Helium	n1
1	Deuterium	
Hydrogen	0	

Principal Quantum Numbers

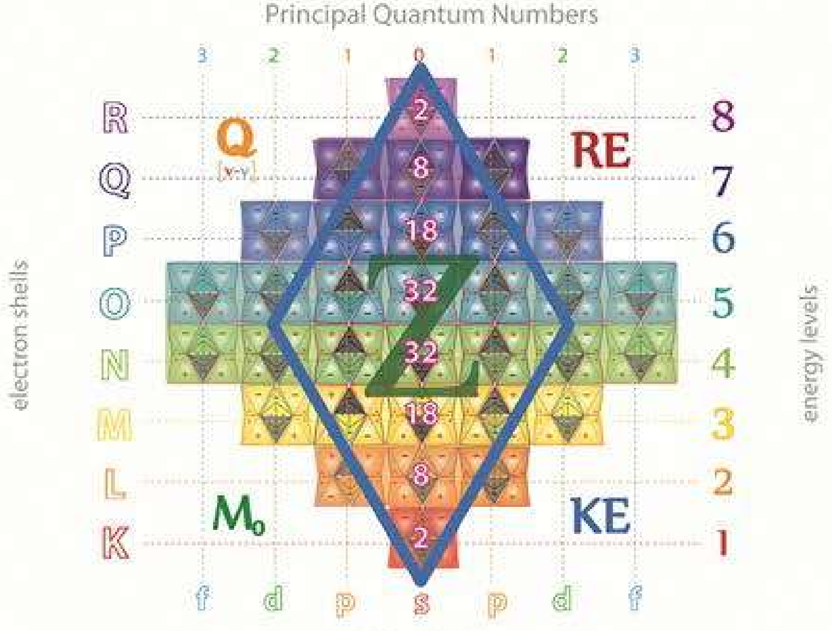
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M0
KE
Z
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f
d
p
s
p
d
f

Sub-orbitals

Periodic mass-ENERGY-Matter

Following periodic summation rules for shell filling
n[1-8] quantum energy deuterium nuclei
combine to form elementary Matter

$$\boxed{_{P}\sum_{K}^{R}{}_{25}^{32}\Big[\overset{\text{Baryon rest masses}}{[72(n)^2]}+\overset{\text{lepton rest mass}}{[12e19]}+\overset{\text{KEM}}{[m_e v^2]}\Big]_{1}^{8}}$$

Deuterium mass-energy per shell

The measured weight of Matter in gravitational fields is the result of
planar mass-energies in tetryonic standing-wave geometries

The periodicity of all the elements,
along with their exact molar rest mass-energies and
quantum wavefunctions can be described with Tetryonic geometries

Ionisation energies

$$\gamma$$

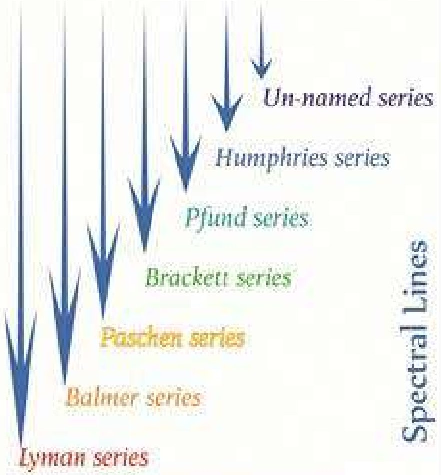
Un-named series
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
Spectral Lines

$$\boxed{Mv^2 = KEM = \hbar c R_H}$$

Photon emission/absorption

Photo-electron ionisation energies

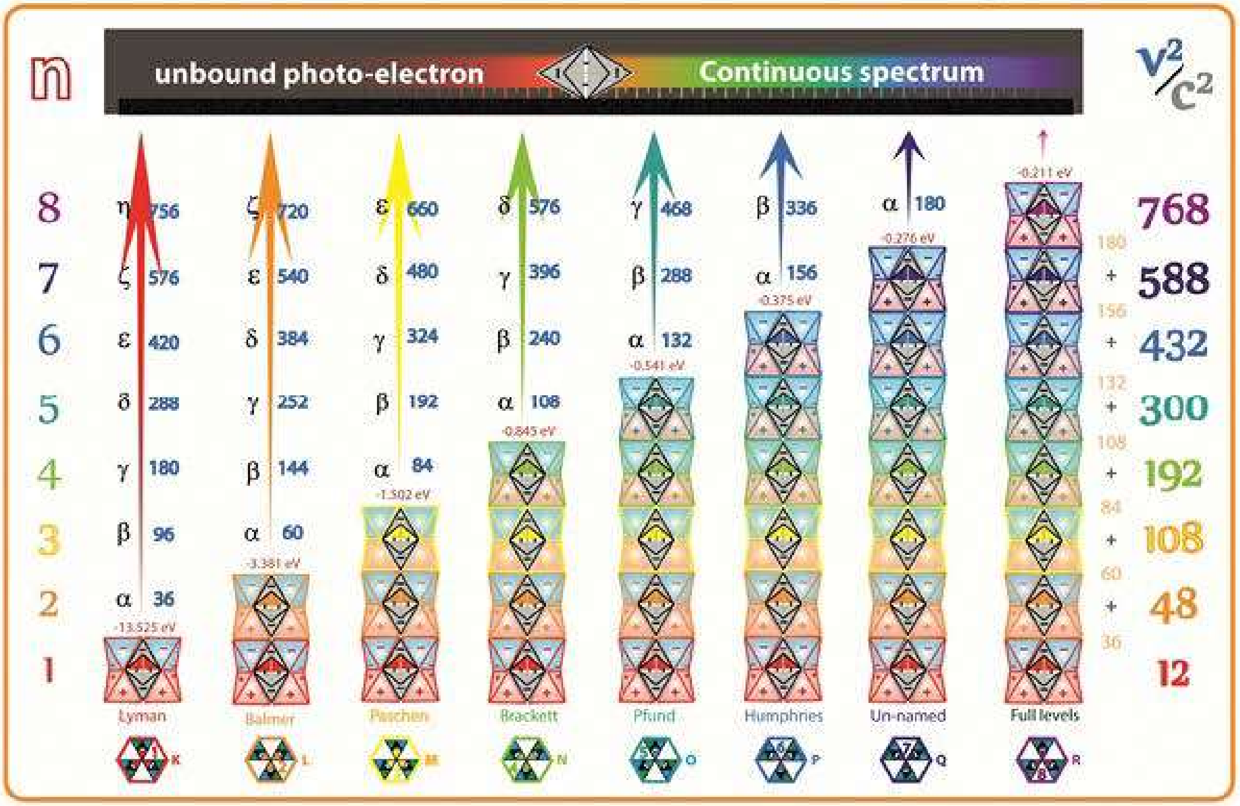
n
unbound photo-electron
Continuous spectrum
v²/c²
8
7
6
5
4
3
2
1
η 756
ζ 576
ε 420
δ 288
γ 180
β 96
α 36
-13.525 eV
Lyman
K
ζ 720
ε 540
δ 384
γ 252
β 144
α 60
-3.381 eV
Balmer
L
ε 660
δ 480
γ 324
β 192
α 84
-1.502 eV
Paschen
M
δ 576
γ 396
β 240
α 108
-0.845 eV
Brackett
N
γ 468
β 288
α 132
-0.541 eV
Pfund
O
β 336
α 156
-0.375 eV
Humphries
P
α 180
-0.276 eV
Un-named
Q
-0.211 eV
Full levels
R
180 + 156 + 132 + 108 + 84 + 60 + 36
768
588
432
300
192
108
48
12

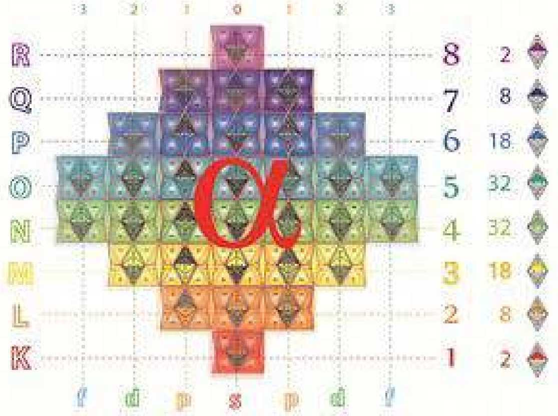
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2 8 18 32 32 18 8 2
α
f d p s p d f

Opposed electron spins results in Hyperfine splitting & Lamb shifts

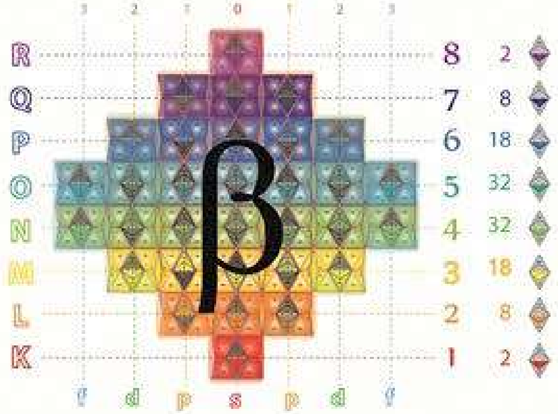
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2 8 18 32 32 18 8 2
β
f d p s p d f

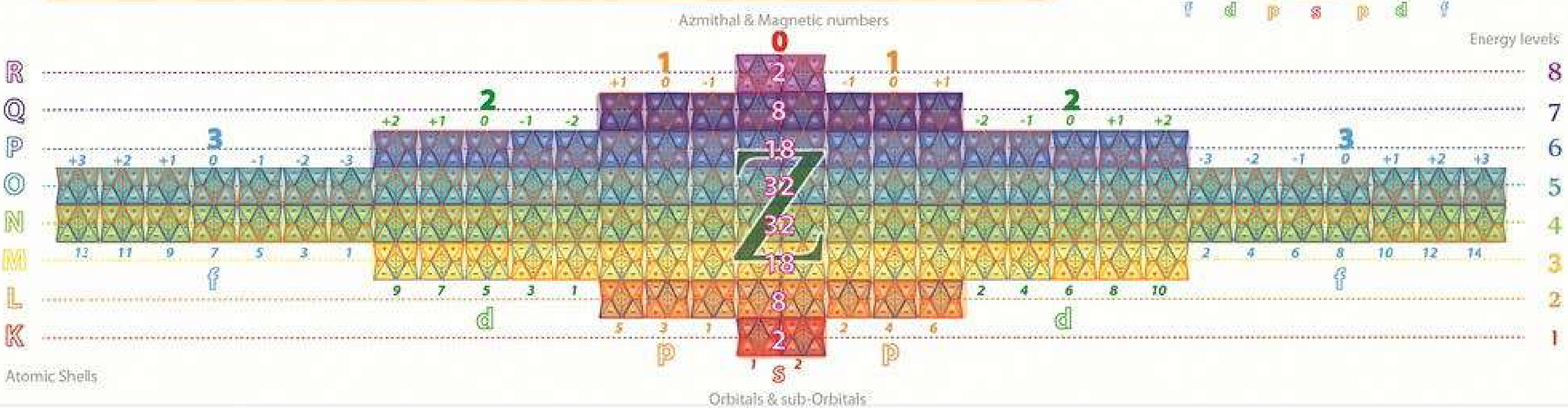
Azmithal & Magnetic numbers
Energy levels
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2 -2 -1 0 +1 +2
3 3
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 0 +1 +2 +3
2 8 18 32 32 18 8 2
Z
13 11 9 7 5 3 1 2 4 6 8 10 12 14
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Orbitals & sub-Orbitals

Eigenstate - Ionisation energies

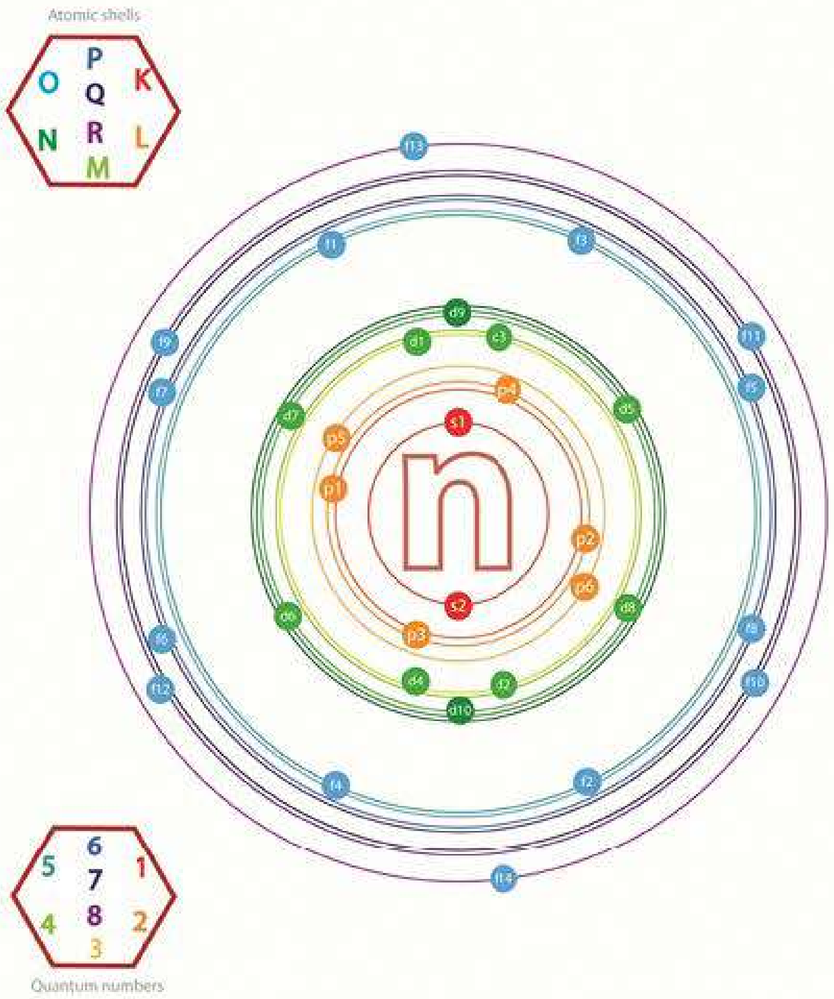
Atomic shells
P
O
Q
K
N
R
L
M
n
s1
s2
p1
p2
p3
p4
p5
p6
d1
d2
d3
d4
d5
d6
d7
d8
d9
d10
f1
f2
f3
f4
f5
f6
f7
f8
f9
f10
f11
f12
f13
f14
6
5
7
1
8
4
2
3
Quantum numbers

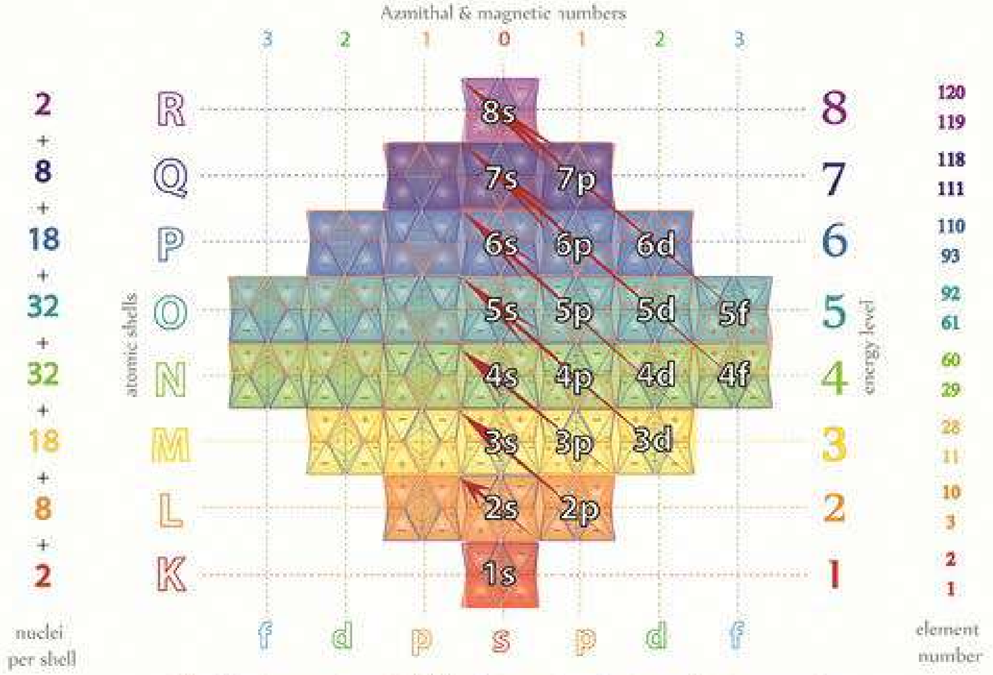
Azmithal & magnetic numbers
3
2
1
0
1
2
3
2
+
8
+
18
+
32
+
32
+
18
+
8
+
2
nuclei per shell
atomic shells
R
Q
P
O
N
M
L
K
8s
7s
7p
6s
6p
6d
5s
5p
5d
5f
4s
4p
4d
4f
3s
3p
3d
2s
2p
1s
f
d
p
s
p
d
f
8
7
6
5
4
3
2
1
energy level
120
119
118
111
110
93
92
61
60
29
28
11
10
3
2
1
element number

The ionisation energies of individual atoms varies due to many factors, namely:
electron spin-orbital coupling with Baryons of specific energies,
the relativistic energies of photo-electrons bound in nuclei
and Zitterbewegung effects on bound electrons

First ionisation energies for all periodic elements

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

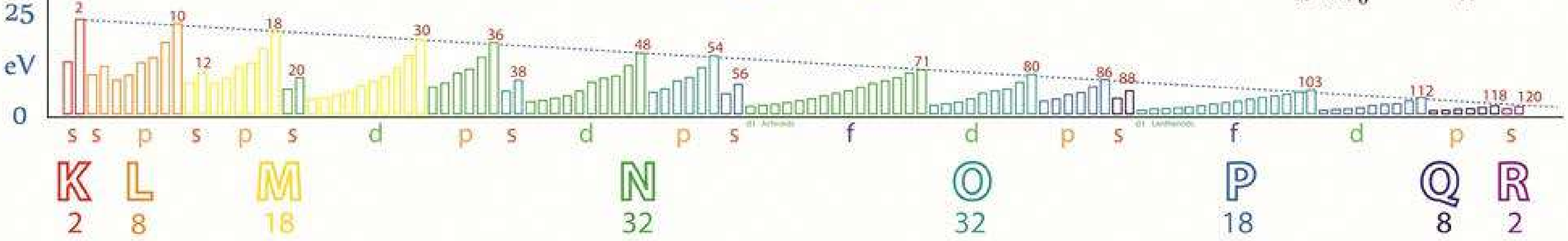

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number.
It is used as rule to determine which nuclei are stable or unstable.

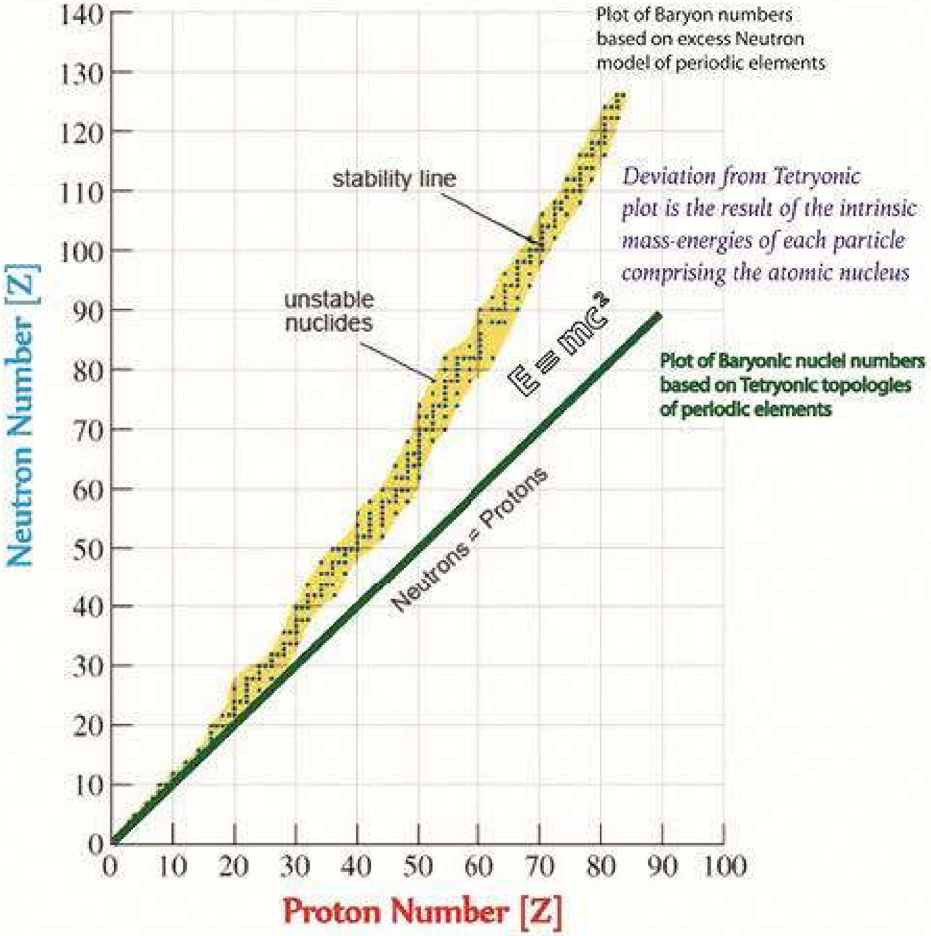

Plot of Baryon numbers based on excess Neutron model of periodic elements

Deviation from Tetryonic plot is the result of the intrinsic mass-energies of each particle comprising the atomic nucleus

Plot of Baryonic nuclei numbers based on Tetryonic topologies of periodic elements

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

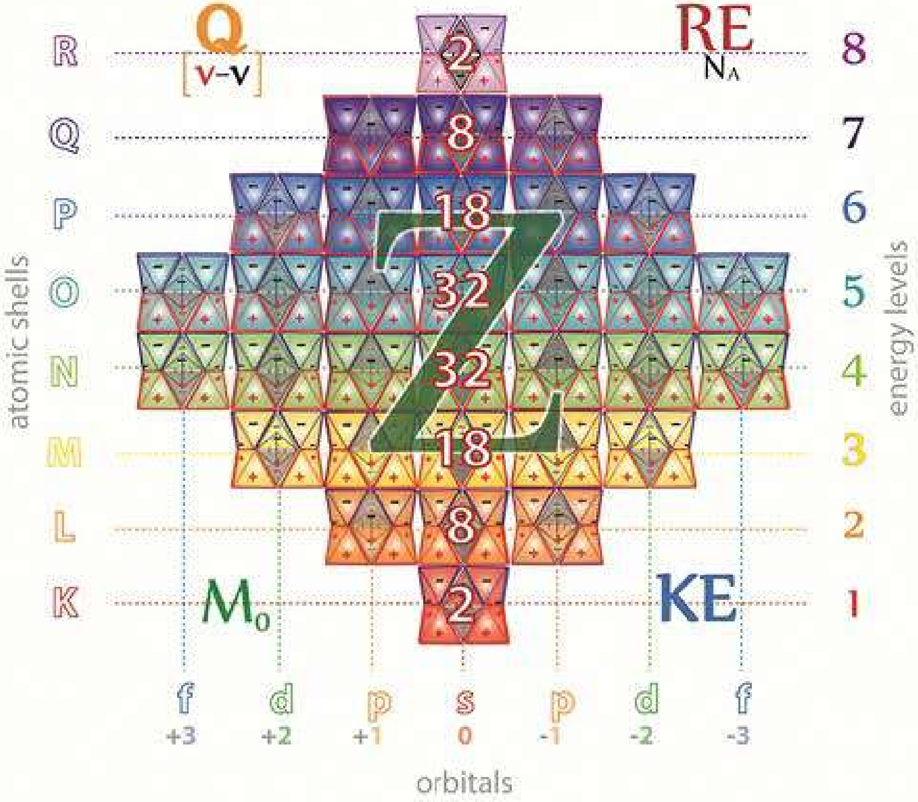

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

Planck mass-energy contributions to elementary Matter and isotopes

The mass-energy content of Deuterium nuclei creates the molar mass of elements
[not extra neutrons in excess of the elemental number]

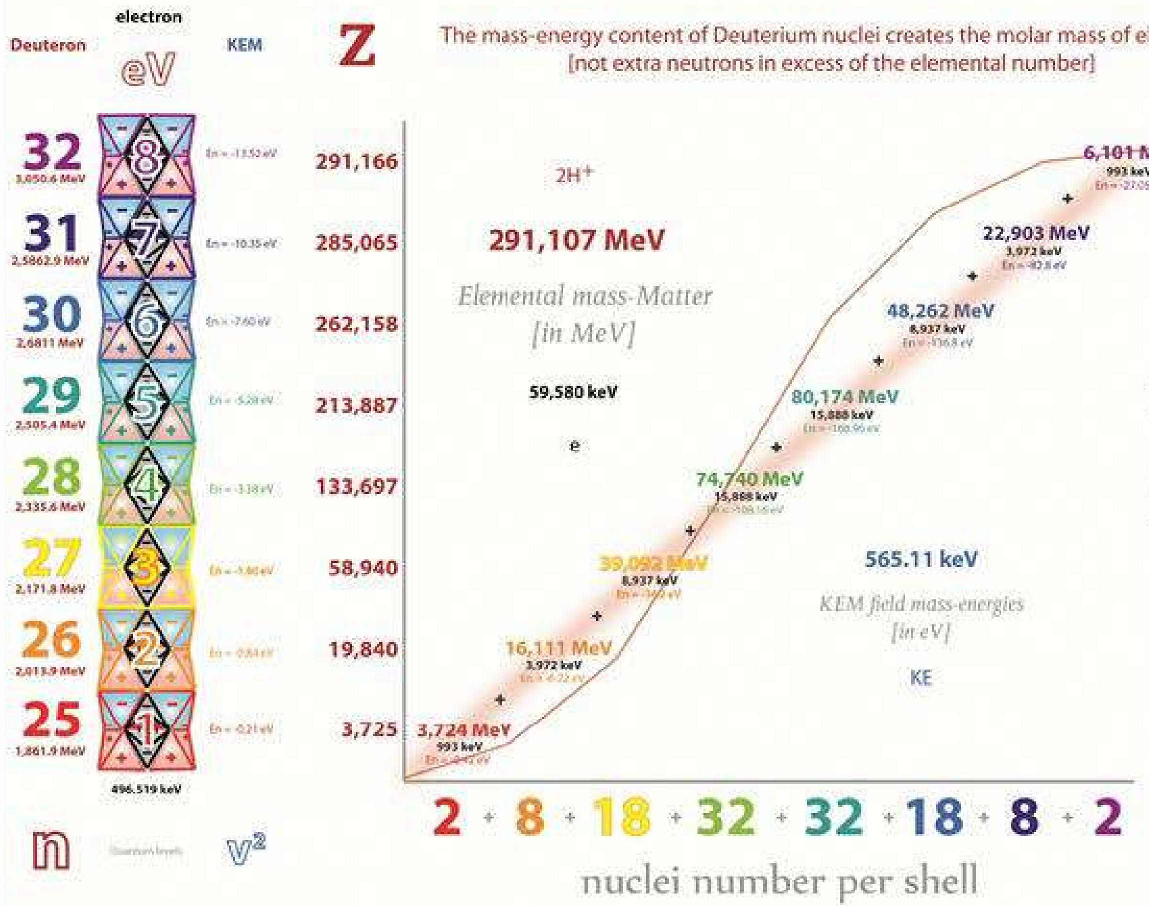

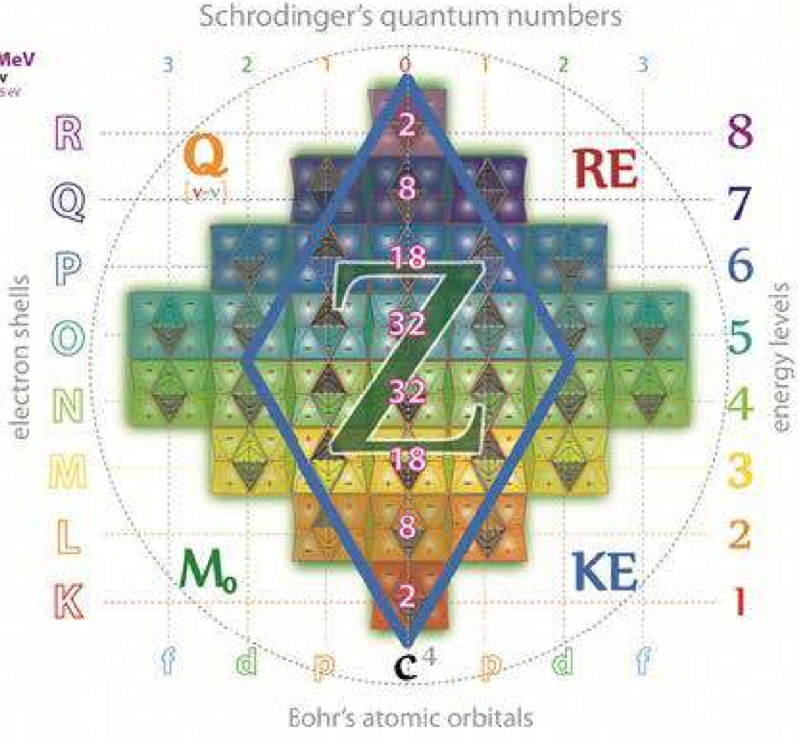
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v·v]
RE
M₀
KE
2
8
18
32
32
18
8
2
Z
c⁴
8
7
6
5
4
3
2
1
energy levels
f
d
p
p
d
f
Bohr's atomic orbitals

n1
Elementary nuclei

are comprised of equal numbers of
Protons, Neutrons & electrons
with varying energy levels

Baryons		electrons		KEM fields
930.947 MeV	+	**496.519 keV**	+	**13.525 ev**

The mass-energy content of Matter topologies is velocity invariant

The mass-energy content of Baryons determines the KEM field of electrons

Baryons **930.947 MeV** + **KEM fields** **13.525 ev** + **electrons** **496.519 keV**

Mapping Planck mass-energy contributions to elementary Matter and isotopes

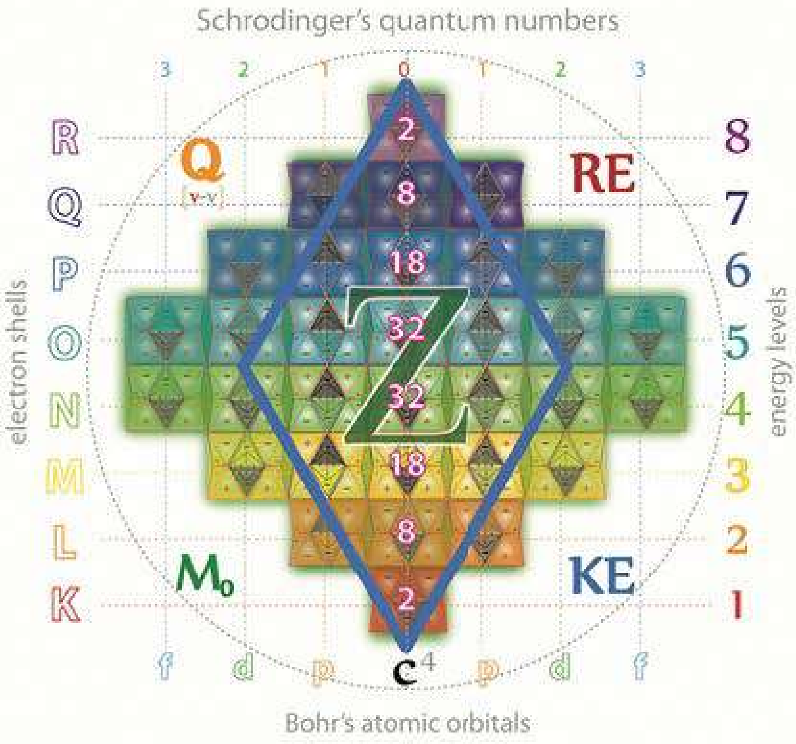
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M₀
KE
8
7
6
5
4
3
2
1
energy levels
2
8
18
32
Z
32
18
8
2
f
d
p
c⁴
p
d
f
Bohr's atomic orbitals

$$E = hv^2$$

$$n^2 + v^2 + e = Z$$

general form quatratic equation

$$ax^2 + bx + c = Z$$

$$E = nhv$$

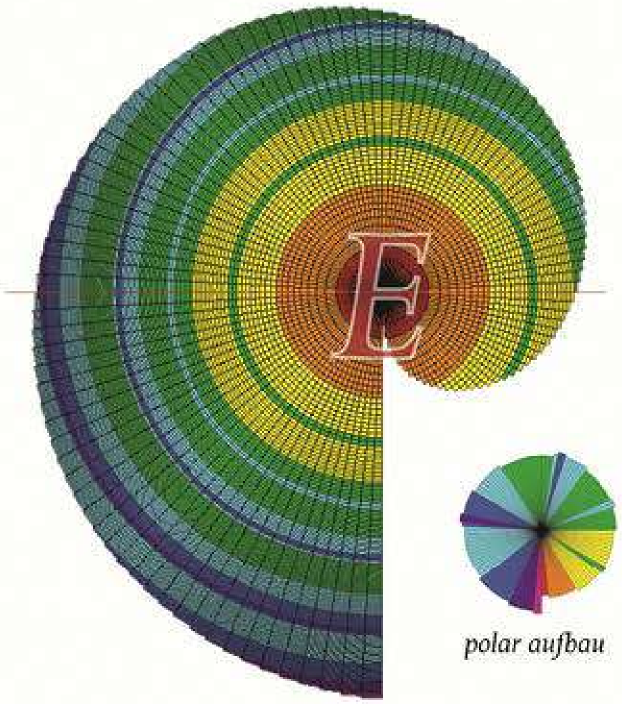
E
polar aufbau

polar energy spirals courtesy of Rene Cormier

Identifying electron rest Matter topologies as velocity invariant we can re-arrange the component Planck mass-energy geometry formulation of periodic elements to

$$h\left[\underbrace{72[v^2]}_{\text{Deuteron rest mass}} + \underbrace{v}_{\text{Spectral lines}} + \underbrace{1.20\,e20\,v}_{\text{electron rest mass}}\right]$$

reveal a quadratic formulation for all Z numbers

Copyright ABRAHAM [2008] - All rights reserved

All elements are comprised of n level Duetrium nuclei

Baryons **KEM fields** **electrons**

$Z\,[[72n^2] + [12v^2] + [1.2e20]]$

1,861,949 MeV **13.525 eV** **496,519 keV**

$2H^+$ $[72n^2]$ Z γ $[M_ev^2]$

Elemental mass-Matter
[in MeV]

The atomic shell energy levels of Deuterium nuclei in elements

Deuteron **electron eV** **KEM**

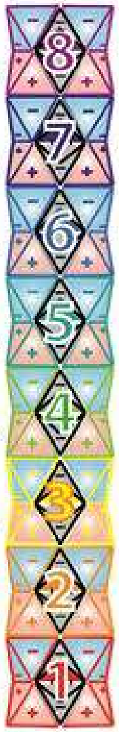

Deuteron		KEM
32 3,050.6 MeV		En = -13.52 eV
31 2,5862.9 MeV		En = -10.15 eV
30 2,6811 MeV		En = -7.60 eV
29 2,505.4 MeV		En = -5.28 eV
28 2,335.6 MeV		En = -3.38 eV
27 2,171.8 MeV		En = -1.90 eV
26 2,013.9 MeV		En = -0.84 eV
25 1,861.9 MeV		En = -0.21 eV

496.519 keV

n v²

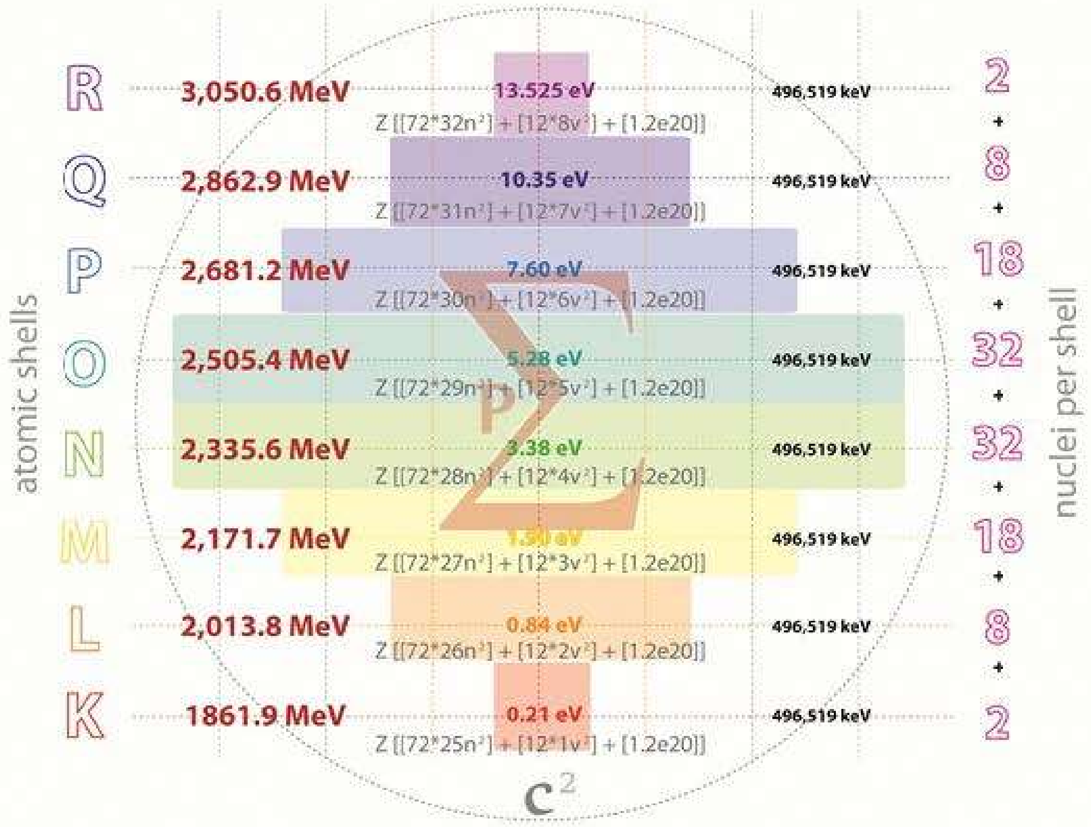

shell				nuclei per shell
R	3,050.6 MeV	13.525 eV Z [[72*32n²] + [12*8v²] + [1.2e20]]	496,519 keV	2
Q	2,862.9 MeV	10.35 eV Z [[72*31n²] + [12*7v²] + [1.2e20]]	496,519 keV	+ 8
P	2,681.2 MeV	7.60 eV Z [[72*30n²] + [12*6v²] + [1.2e20]]	496,519 keV	+ 18
O	2,505.4 MeV	5.28 eV Z [[72*29n²] + [12*5v²] + [1.2e20]]	496,519 keV	+ 32
N	2,335.6 MeV	3.38 eV Z [[72*28n²] + [12*4v²] + [1.2e20]]	496,519 keV	+ 32
M	2,171.7 MeV	1.90 eV Z [[72*27n²] + [12*3v²] + [1.2e20]]	496,519 keV	+ 18
L	2,013.8 MeV	0.84 eV Z [[72*26n²] + [12*2v²] + [1.2e20]]	496,519 keV	+ 8
K	1861.9 MeV	0.21 eV Z [[72*25n²] + [12*1v²] + [1.2e20]]	496,519 keV	+ 2

6,101 MeV
993 keV
En = -27.05 eV
+
22,903 MeV
3,972 keV
En = -82.8 eV
+
48,262 MeV
8,937 keV
En = -316.8 eV
+
80,174 MeV
15,888 keV
En = -168.96 eV
+
74,740 MeV
15,888 keV
En = -108.16 eV
+
39,092 MeV
8,937 keV
En = -34.2 eV
+
16,111 MeV
3,972 keV
En = -6.72 eV
+
3,724 MeV
993 keV
En = -0.42 eV

[1.2e20]]

e

Determines the spectral line [KEM field energies] of electrons bound to them

The relativistic rest mass-energy-Matter of all periodic elements
is the sum of the mass-energies of all atomic nuclei and spectral lines
that comprise its mass-Matter topology as measured in
any spatial co-ordinate system per unit of time

the rest mass-Matter of bound photo-electrons is velocity invariant

48
H
1.0000
22,512

Avagadro's number

6.022141579 e26
atoms in 1KG of Matter

1 KG mass
[of Matter]

1.660538841 e-27 kg
atomic rest mass-Matter

Hydrogen

01

using SI units Avagadro's number can be expressed exactly as the inverse rest mass of Hydrogen

Carbon

12

Weighted atomic mass

5.019789213 e25
atoms in 1KG of Matter

6.02214078 e 23

1.99211552 e-26 kg
atomic rest mass-Matter

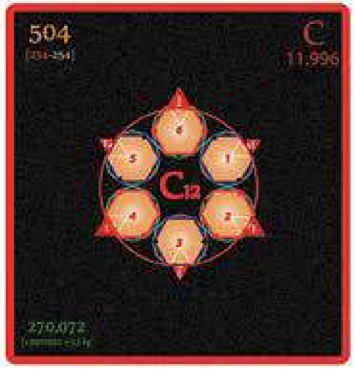

Tetryonic charge geometries make weighted atomic mass measurements and calculations obsolete

22,506 1/12 of Carbon 12 [Graphene] is not equal to 1 Hydrogen atom 22,512
(Deuterium is the building block of all atomic elements)

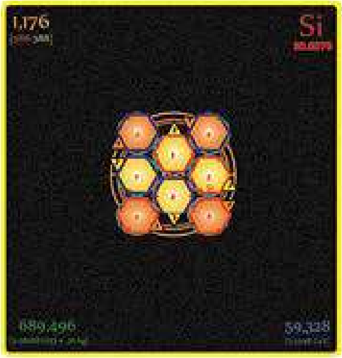

International Avagadro project

1.966225348 e25
atoms in 1KG of Matter

5.085887033 e-26 kg
atomic rest mass-Matter

Silicon

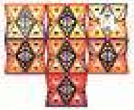

28

The gram was originally defined in 1795 as the mass of one cubic centimeter of water at 4°C,
making the kilogram equal to the mass of one liter of water.

The prototype kilogram, manufactured in 1799 and from which the current kilogram is based
has a mass equal to the mass of 1.000025 liters of water

In recent years two major experiments, namely the Watt balance & Avagadro projects, have been attempting to
measure and define 1KG of mass-Matter in terms of electrical force and the number of atoms respectively
in order to better define 1KG of mass-Matter precisely for all future physical references

Platinum

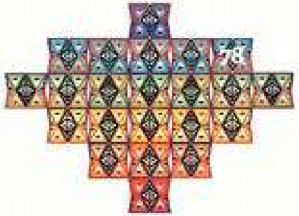

191

La Grande K

2.817950081 e24
atoms in 1KG of Matter
3.181804449 e23

90% Pt
10% Ir

The La grande K is an alloy of 90% Platinum & 10% Iridium
that has been slowly losing mass since its manufacture

3.1893811012 e-25 kg
atomic rest mass-Matter
3.142870708 e-25 kg

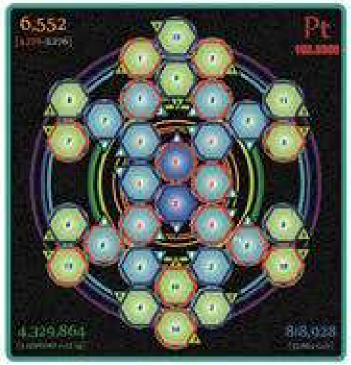

All atomic rest masses are for atoms at absolute zero and any deviation is a measure of
the topological Matter's Kinetic energy content [chemical energy, KEM fields and/or spectral lines]

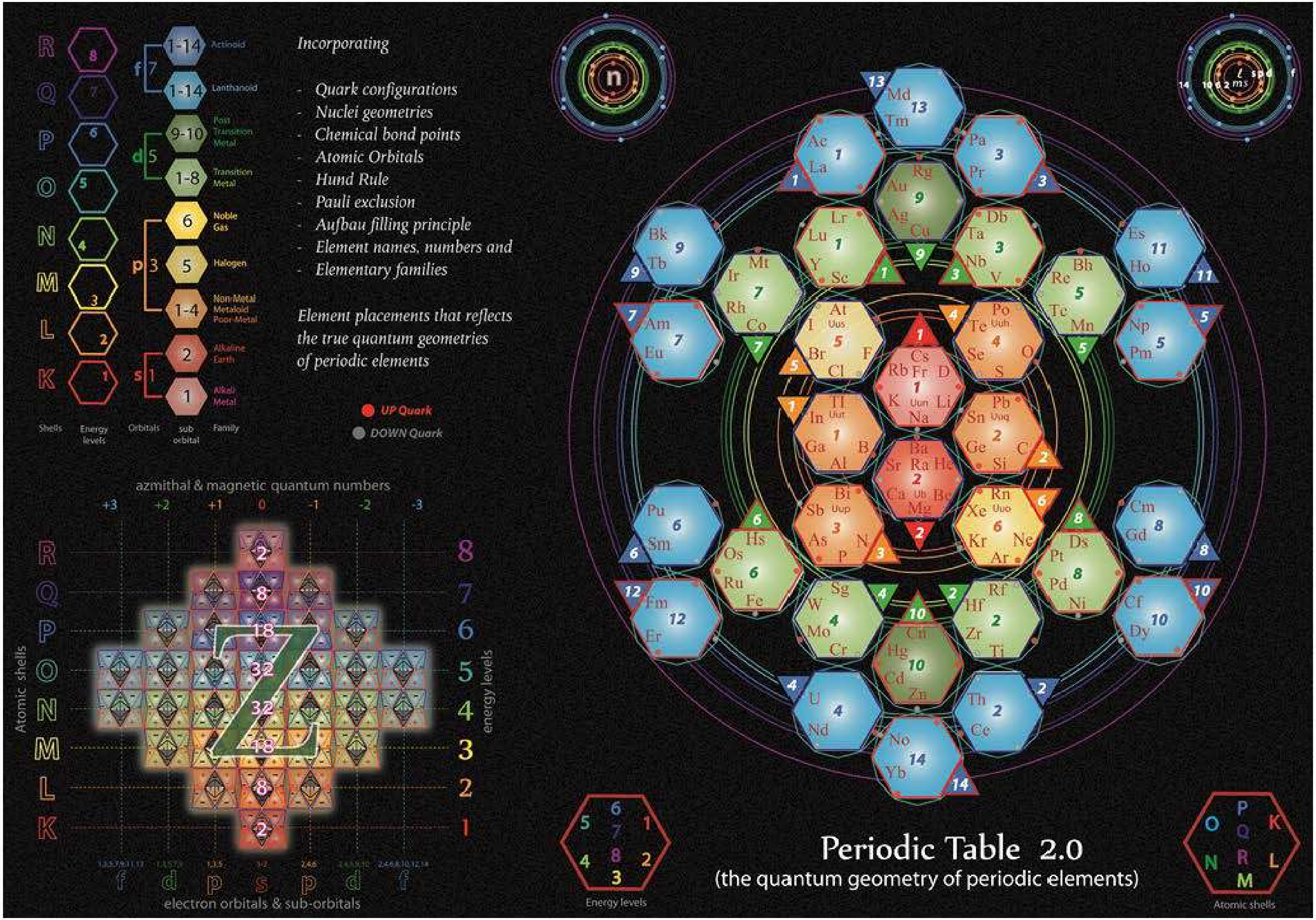

Incorporating

- *Quark configurations*
- *Nuclei geometries*
- *Chemical bond points*
- *Atomic Orbitals*
- *Hund Rule*
- *Pauli exclusion*
- *Aufbau filling principle*
- *Element names, numbers and*
- *Elementary families*

Element placements that reflects the true quantum geometries of periodic elements

Periodic Table 2.0

(the quantum geometry of periodic elements)

R Q P O N M L K
8 7 6 5 4 3 2 1
f 7
d 5
p 3
s 1
1-14 Actinoid
1-14 Lanthanoid
9-10 Post Transition Metal
1-8 Transition Metal
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
2 Alkaline Earth
1 Alkali Metal
Shells
Energy levels
Orbitals
sub orbital
Family
UP Quark
DOWN Quark
azmithal & magnetic quantum numbers
+3 +2 +1 0 -1 -2 -3
Atomic shells
energy levels
Z
2 8 18 32 32 18 8 2
f d p s p d f
electron orbitals & sub-orbitals
n
ℓ s p d f
ms
Md Tm 13
Ac La 1
Pa Pr 3
Rg Au Ag Cu 9
Lr Lu Y Sc 1
Db Ta Nb V 3
Bk Tb 9
Mt Ir Rh Co 7
Bh Re Tc Mn 5
Es Ho 11
Am Eu 7
At I Uus Br Cl F 5
Cs Rb Fr K Uun Li Na D 1
Po Te Uuh Se S O 4
Np Pm 5
Tl In Uut Ga Al B 1
Pb Sn Uuq Ge Si C 2
Ba Sr Ra Ca Ub Mg He Be 2
Bi Sb Uup As P N 3
Rn Xe Uuo Kr Ar Ne 6
Pu Sm 6
Hs Os Ru Fe 6
Ds Pt Pd Ni 8
Cm Gd 8
Fm Er 12
Sg W Mo Cr 4
Cn Hg Cd Zn 10
Rf Hf Zr Ti 2
Cf Dy 10
U Nd 4
Th Ce 2
No Yb 14
Energy levels
Atomic shells

Copyright ABRAHAM [2008] - All rights reserved

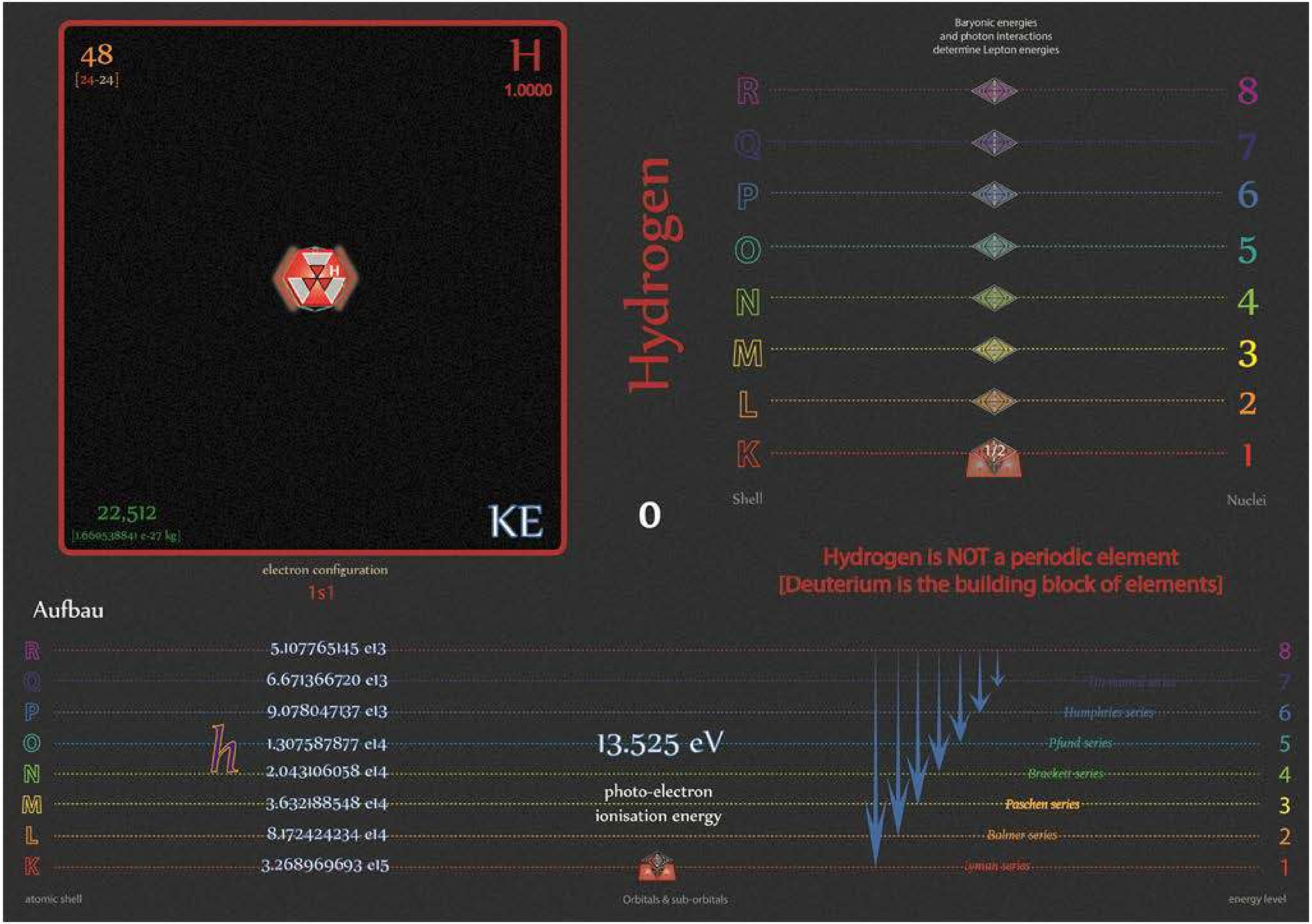

48
[24-24]

H
1.0000

22,512
[1.660538841 e-27 kg]

KE

electron configuration
1s1

Hydrogen

0

Baryonic energies
and photon interactions
determine Lepton energies

Shell	Nuclei
R	8
Q	7
P	6
O	5
N	4
M	3
L	2
K 1/2	1

Hydrogen is NOT a periodic element
[Deuterium is the building block of elements]

Aufbau

atomic shell			energy level
R	5.107765145 e13		8
Q	6.671366720 e13	[illegible]	7
P	9.078047137 e13	Humphries series	6
O	1.307587877 e14	Pfund series	5
N	2.043106058 e14	Brackett series	4
M	3.632188548 e14	Paschen series	3
L	8.172424234 e14	Balmer series	2
K	3.268969693 e15	Lyman series	1

h

13.525 eV
photo-electron
ionisation energy

Orbitals & sub-orbitals

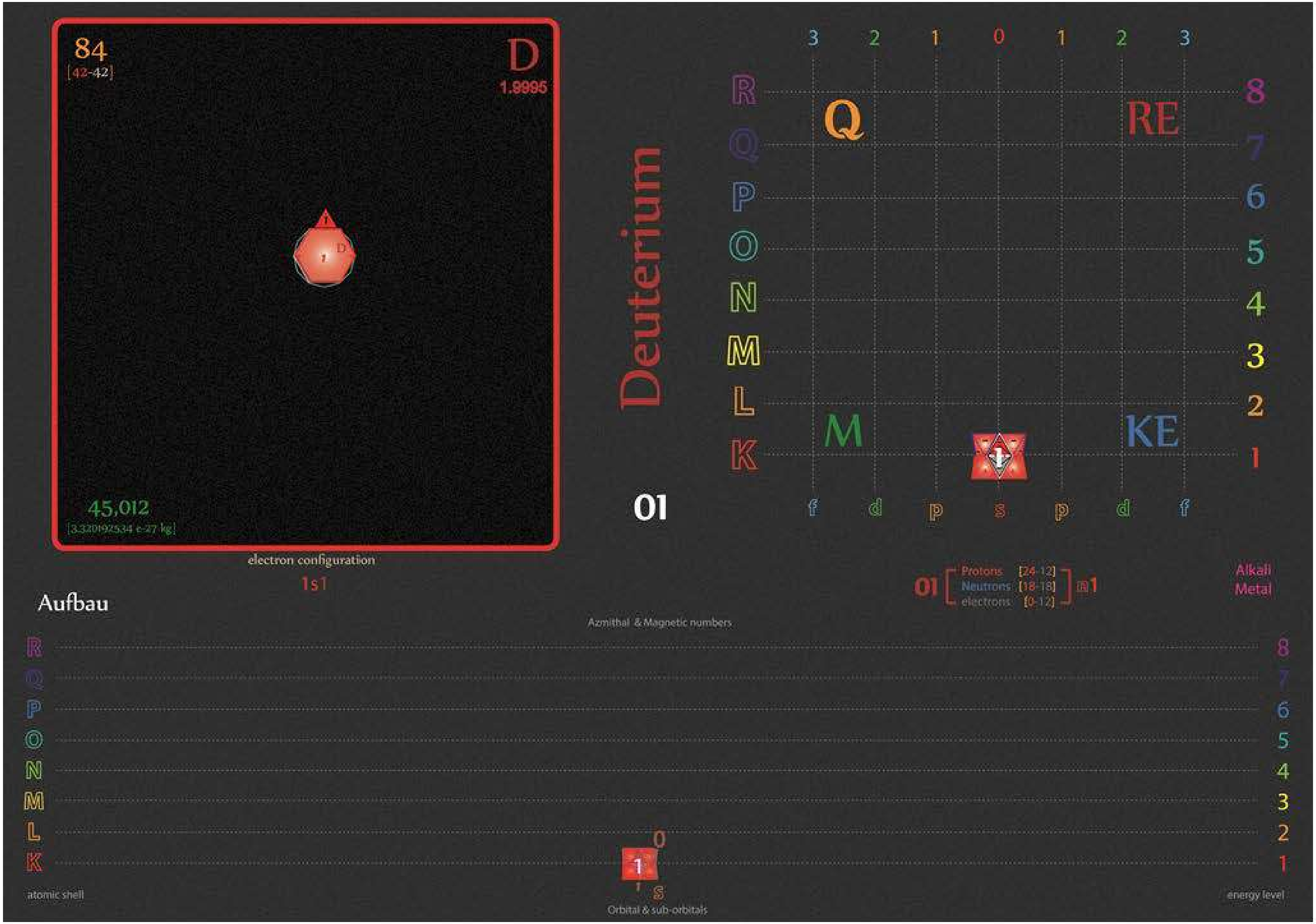
84
[42-42]
D
1.9995
45,012
[3.320192534 e-27 kg]
electron configuration
1s1
Deuterium
01
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
01
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1
Alkali Metal
Aufbau
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
s
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.01 - Deuterium atom

Copyright ABRAHAM [2008] - All rights reserved

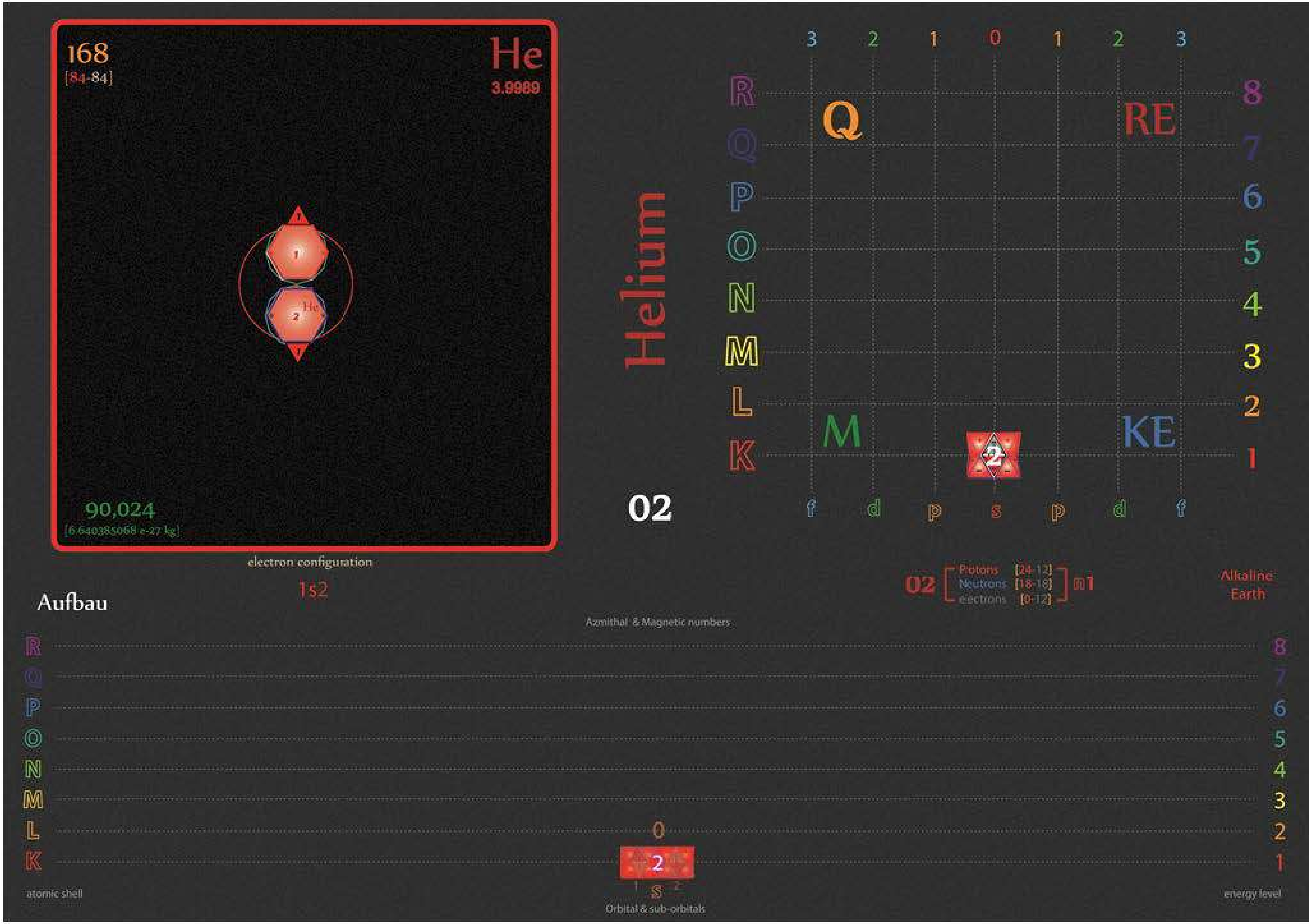
168
[84-84]
He
3.9989
1
2
He
90,024
[6.640385068 e-27 kg]
electron configuration
1s2
Aufbau
Helium
02
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
2
f
d
p
s
p
d
f
02
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1
Alkaline
Earth
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
s
2
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.02 - Helium atom

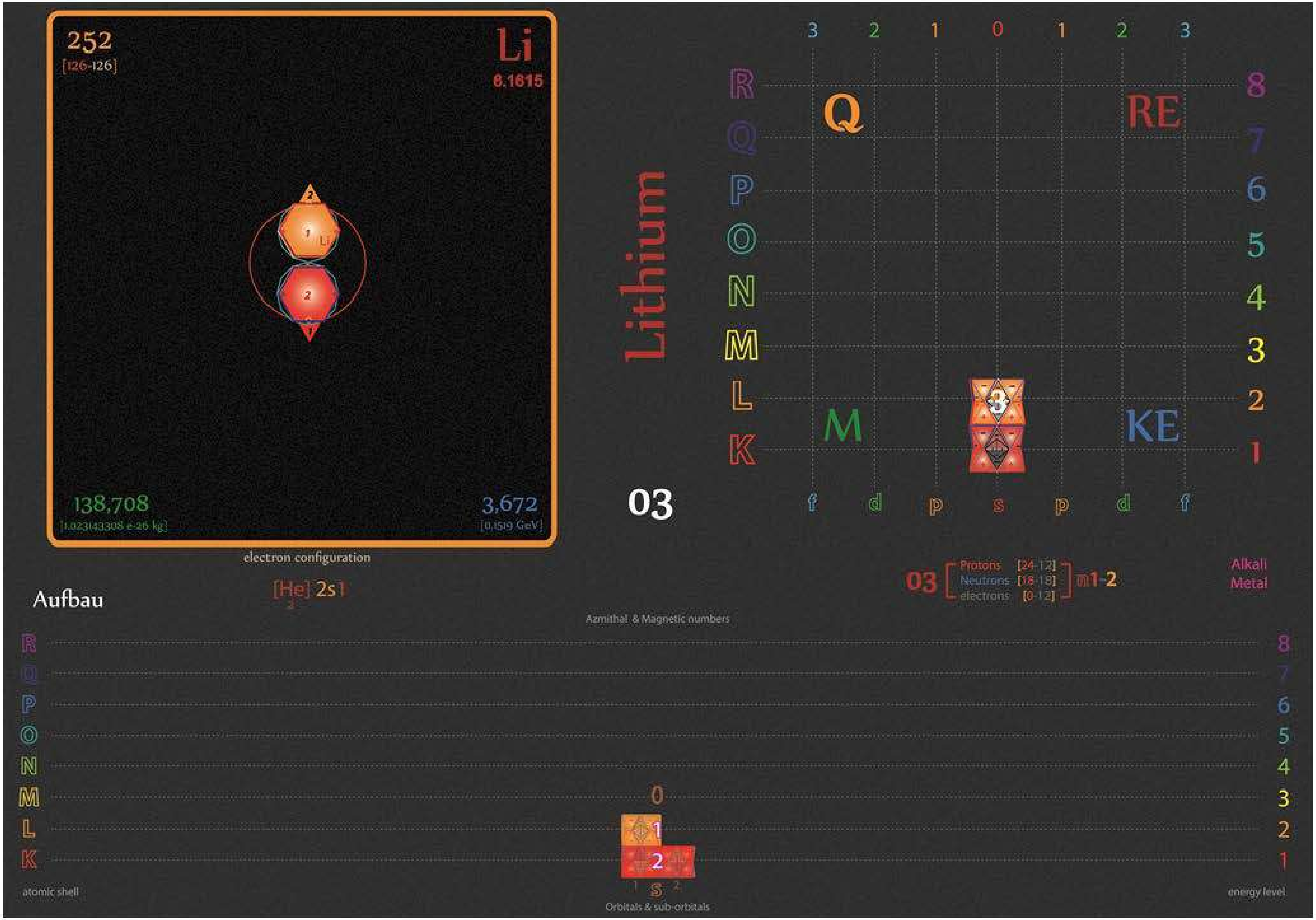
252
[126-126]
Li
8.1615
1
2
Li
138,708
[1.023143308 e-26 kg]
3,672
[0.1519 GeV]
electron configuration
[He] 2s1
Aufbau
Lithium
03
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
3
8
7
6
5
4
3
2
1
f d p s p d f
Azmithal & Magnetic numbers
03
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Alkali
Metal
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
2
1 s 2
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.03 - Lithium atom

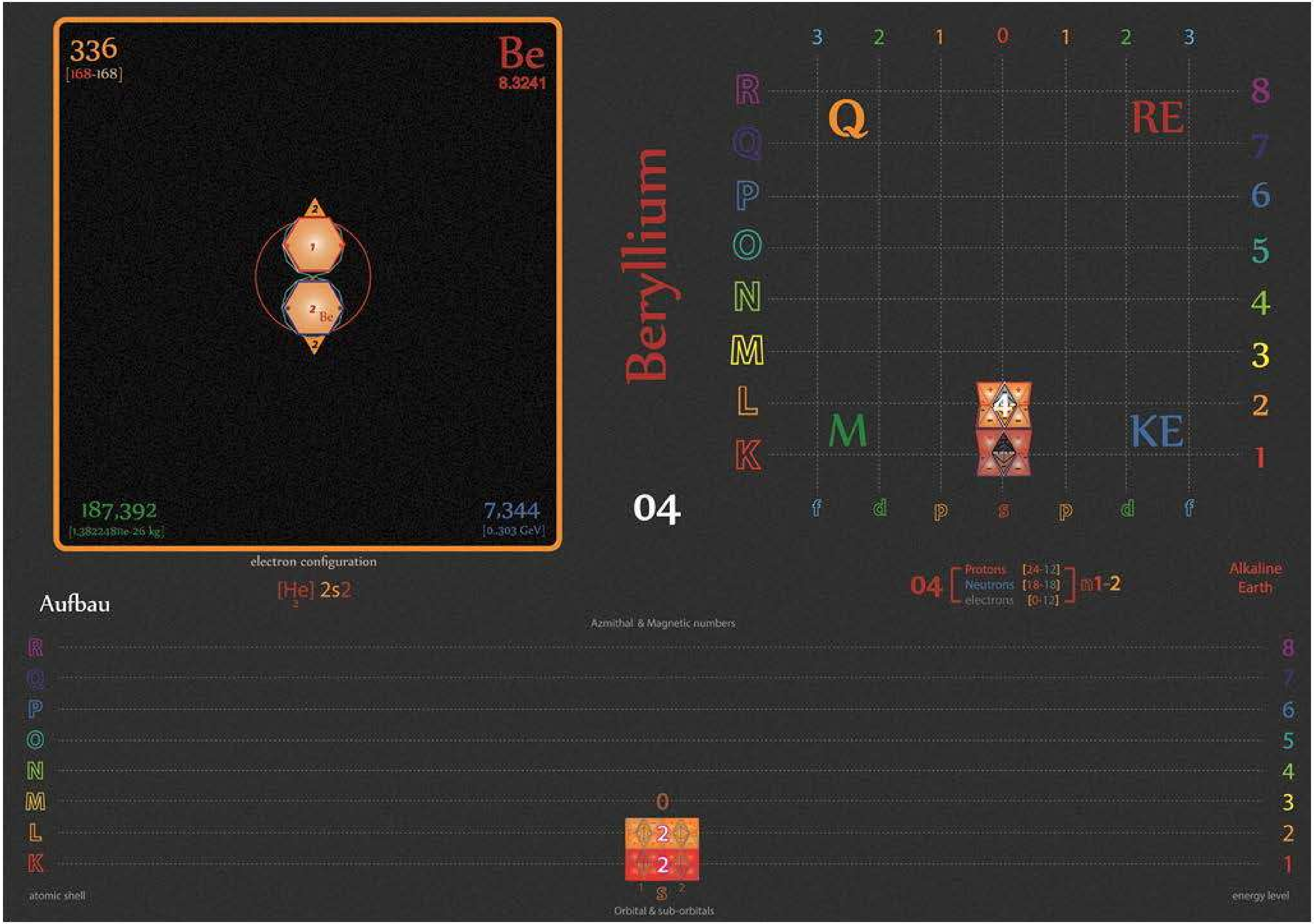
336
[168-168]
Be
8.3241
2
1
2
Be
2
187,392
[1.38224811e-26 kg]
7,344
[0.303 GeV]
electron configuration
[He] 2s2
2
Aufbau
Beryllium
04
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
4
f d p s p d f
04
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Alkaline Earth
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
2
1 s 2
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.04 - Beryllium atom

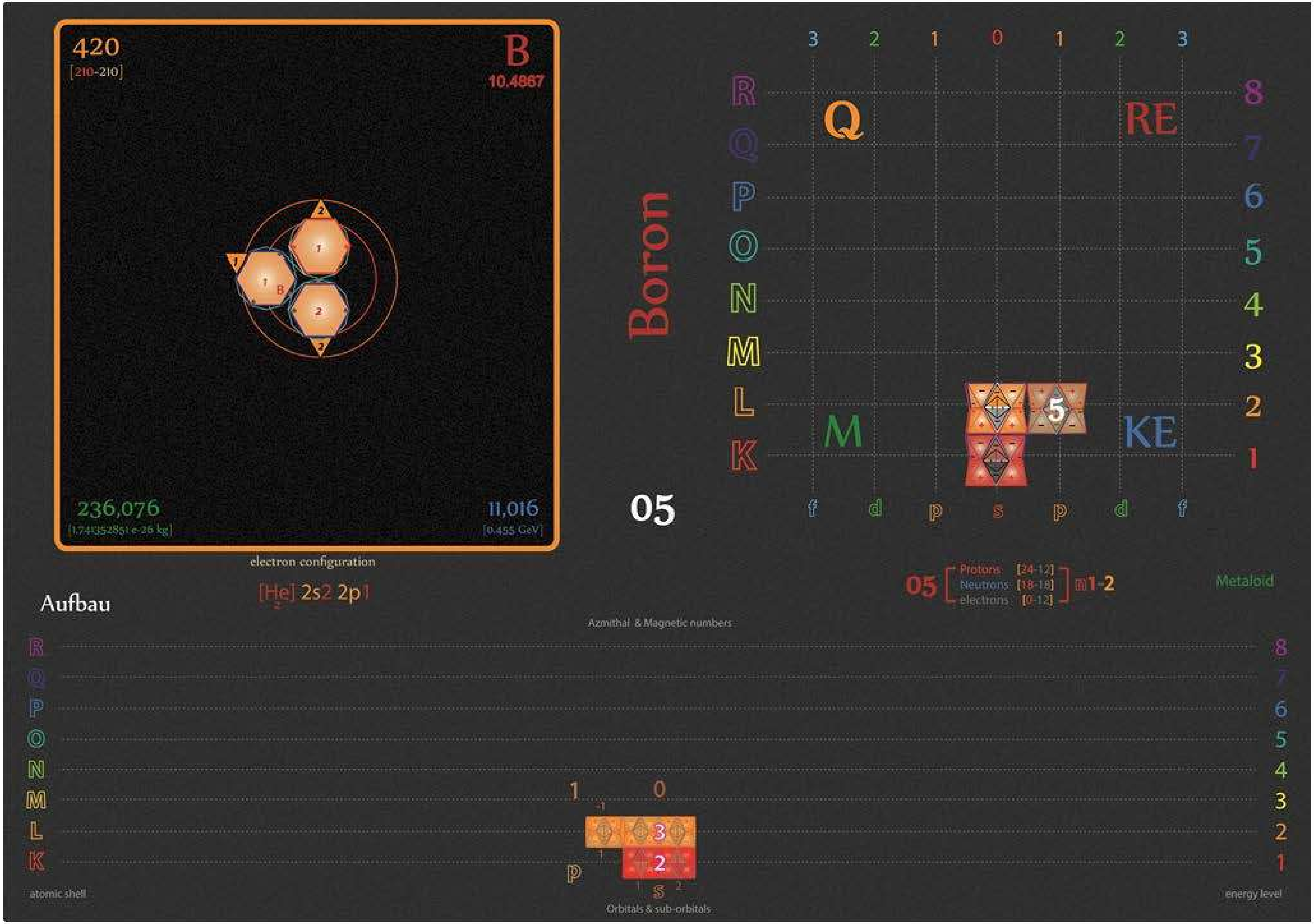
420
[210-210]
B
10.4867
2
1
1
B
2
2
236,076
[1.741352851 e-26 kg]
11,016
[0.455 GeV]
electron configuration
[He] 2s2 2p1
2
Aufbau
Boron
05
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
5
f d p s p d f
05
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1
0
-1
1
3
2
p
1
s
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.05 - Boron atom

Copyright ABRAHAM [2008] - All rights reserved

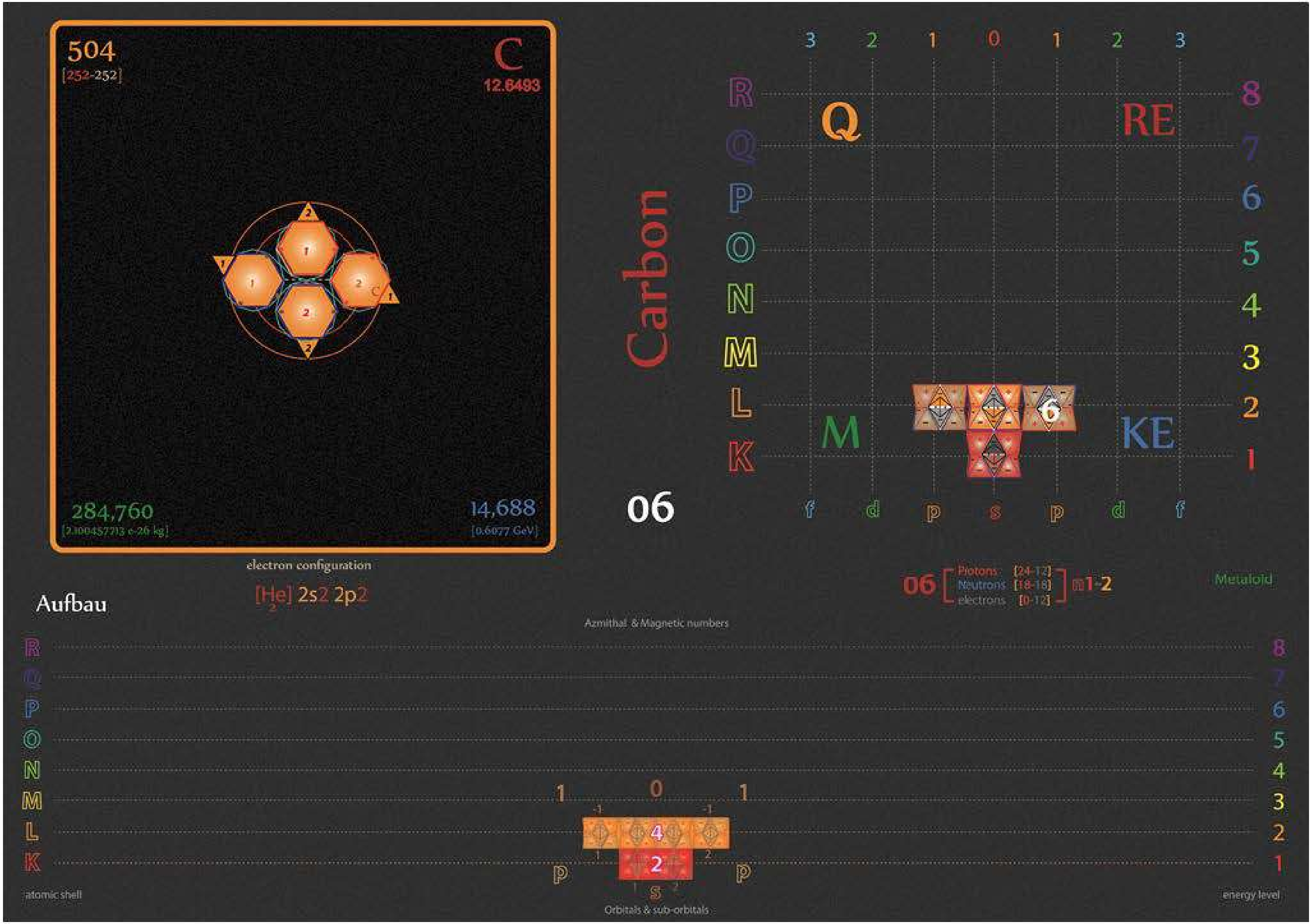
504
[252-252]
C
12.6493
1
1
1
2
2
2
2
1
284,760
[2.100457713 e-26 kg]
14,688
[0.6077 GeV]
Carbon
06
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
6
f d p s p d f
electron configuration
[He] 2s2 2p2
2
Aufbau
06
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
1 0 1
-1
-1
4
2
1
2
p
p
s
1
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.06 - Carbon atom

Copyright ABRAHAM [2008] - All rights reserved

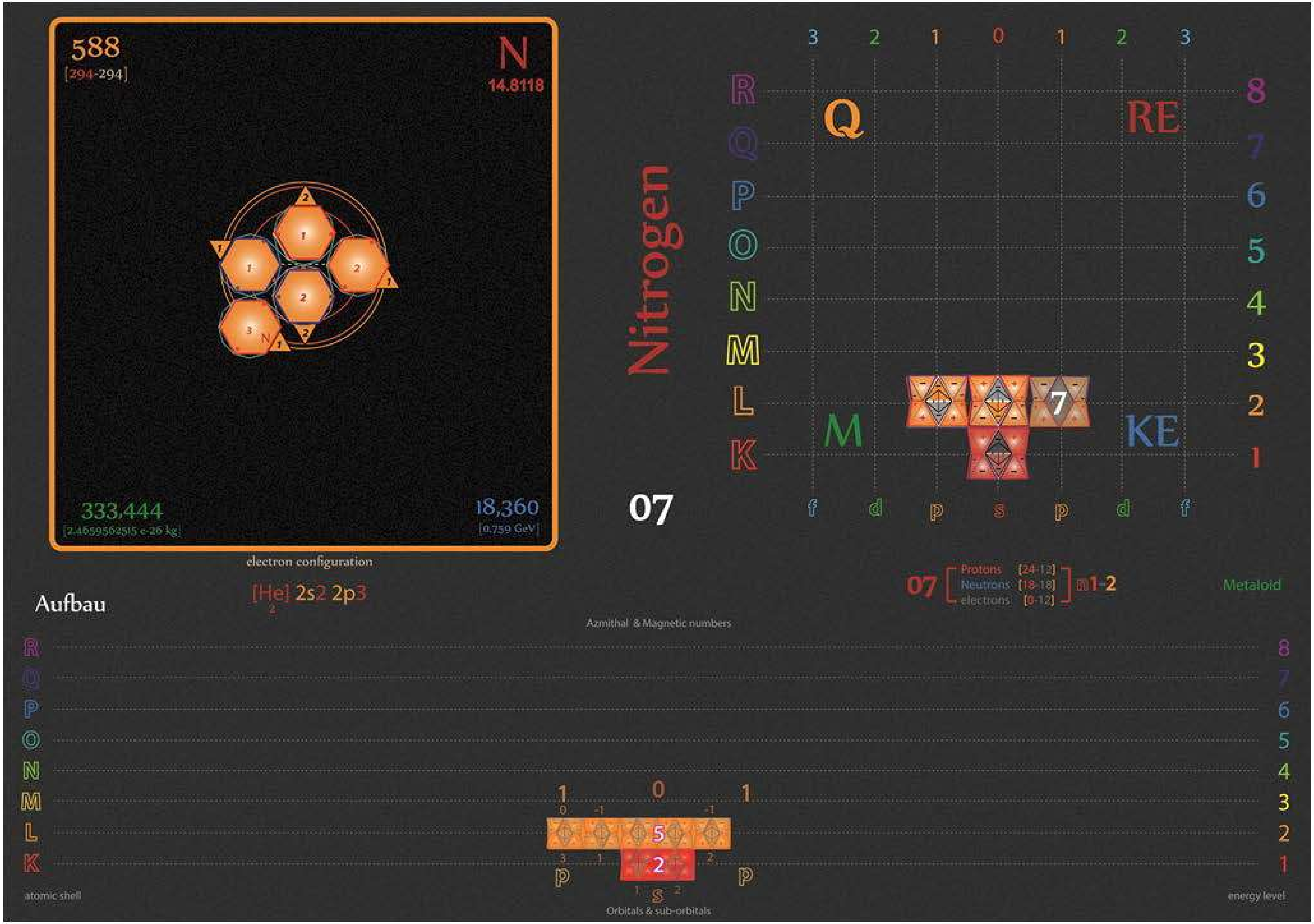
588
[294-294]
N
14.8118
333,444
[2.4659562515 e-26 kg]
18,360
[0.759 GeV]
Nitrogen
07
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
7
f d p s p d f
electron configuration
[He] 2s2 2p3
Aufbau
07
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
1 0 1
5
2
p s p
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.07 - Nitrogen atom

Copyright ABRAHAM [2008] - All rights reserved

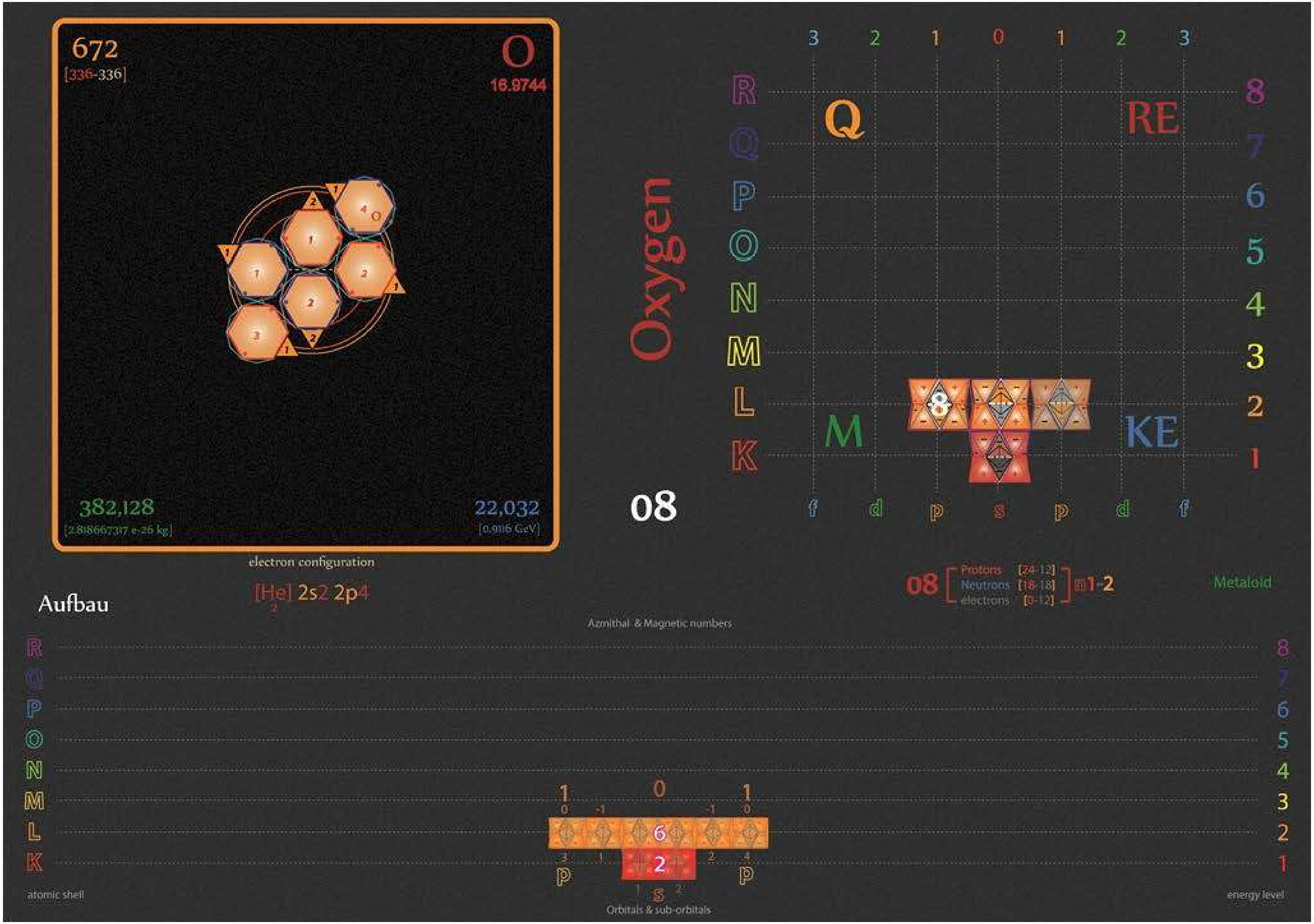
672
[336-336]
O
16.9744
382,128
[2.818667317 e-26 kg]
22,032
[0.916 GeV]
Oxygen
08
electron configuration
[He] 2s2 2p4
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
8
f d p s p d f
08 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Metaloid
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
6
2
3 1 2 4
p s p
1 2
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.08 - Oxygen atom

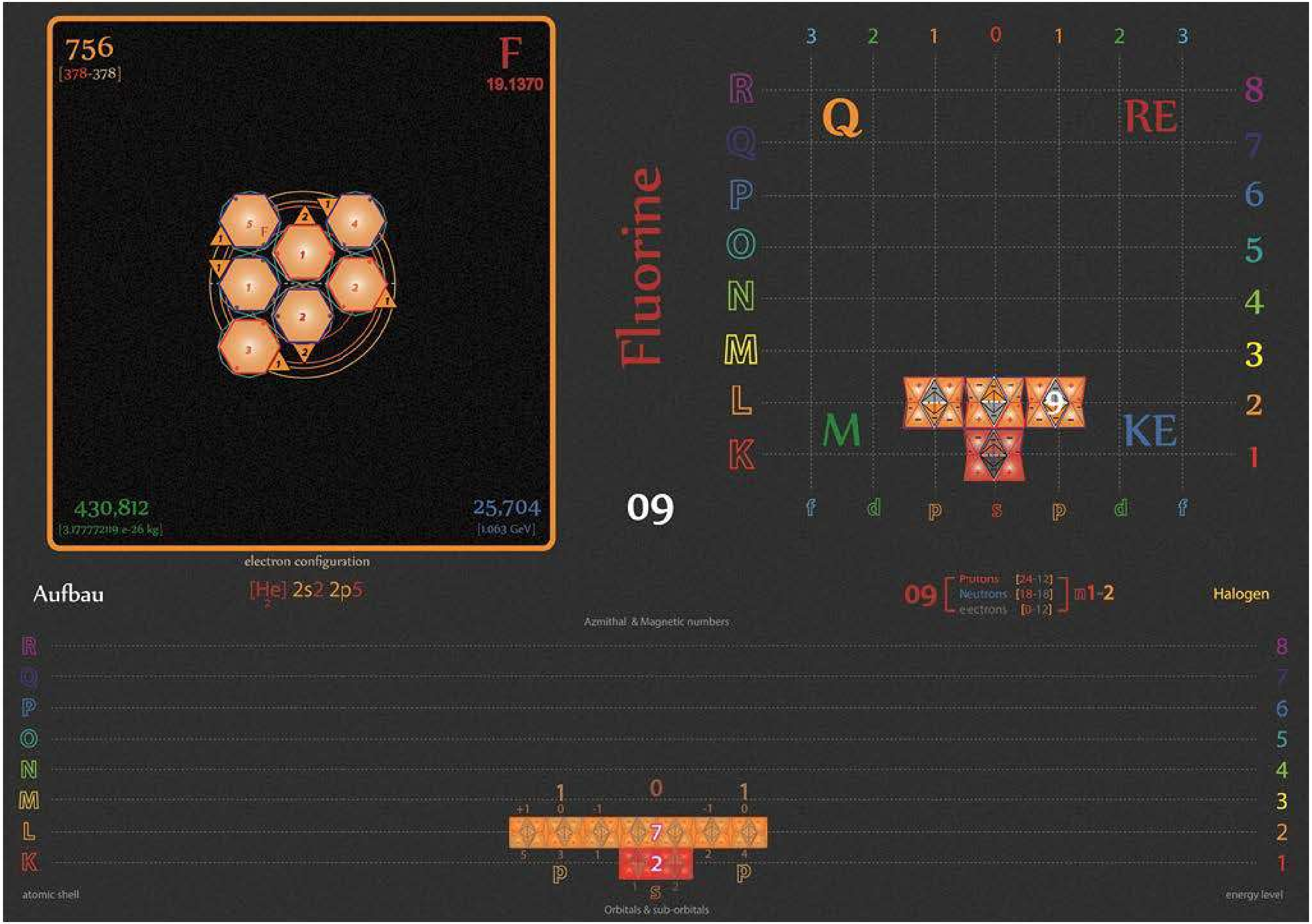
756
[378-378]
F
19.1370
430,812
[3.17772119 e-26 kg]
25,704
[1.063 GeV]
Fluorine
09
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
9
f d p s p d f
electron configuration
Aufbau
[He] 2s2 2p5
2
Azmithal & Magnetic numbers
09
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-2
Halogen
1 0 1
+1 0 -1 -1 0
7
2
5 3 1 2 4
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.09 - Fluorine atom

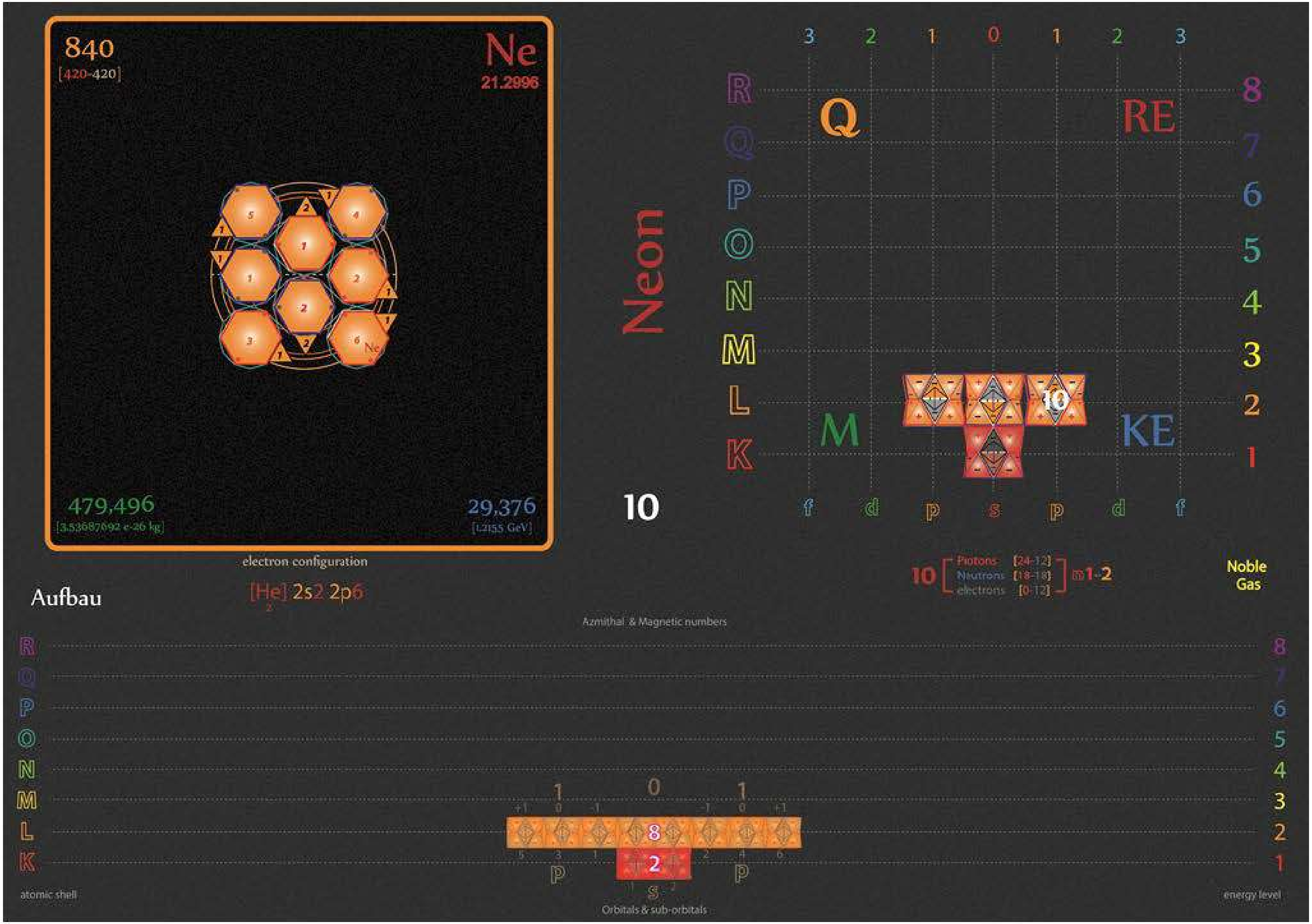
840
[420-420]
Ne
21.2996
479,496
[3.53687692 e-26 kg]
29,376
[1.2155 GeV]
Neon
10
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
10 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Noble Gas
electron configuration
[He] 2s2 2p6
Aufbau
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.10 - Neon atom

Copyright ABRAHAM [2008] - All rights reserved

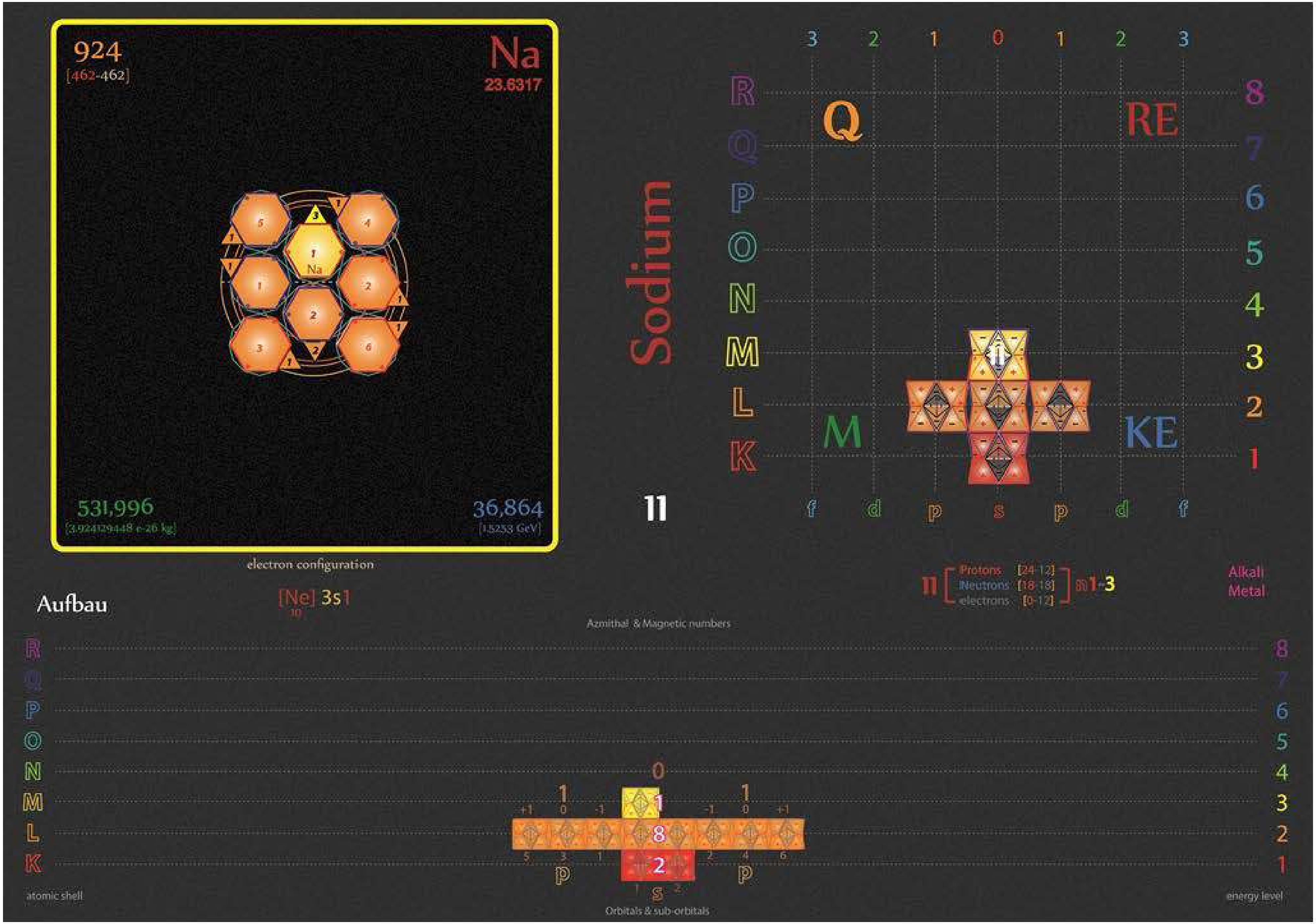
924
[462-462]
Na
23.6317
531,996
[3.924129448 e-26 kg]
36,864
[13253 GeV]
electron configuration
Aufbau
[Ne] 3s1
10
Sodium
11
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
11 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-3
Alkali Metal
Azmithal & Magnetic numbers
0
1 1
+1 0 -1 -1 0 +1
1
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.11 - Sodium atom

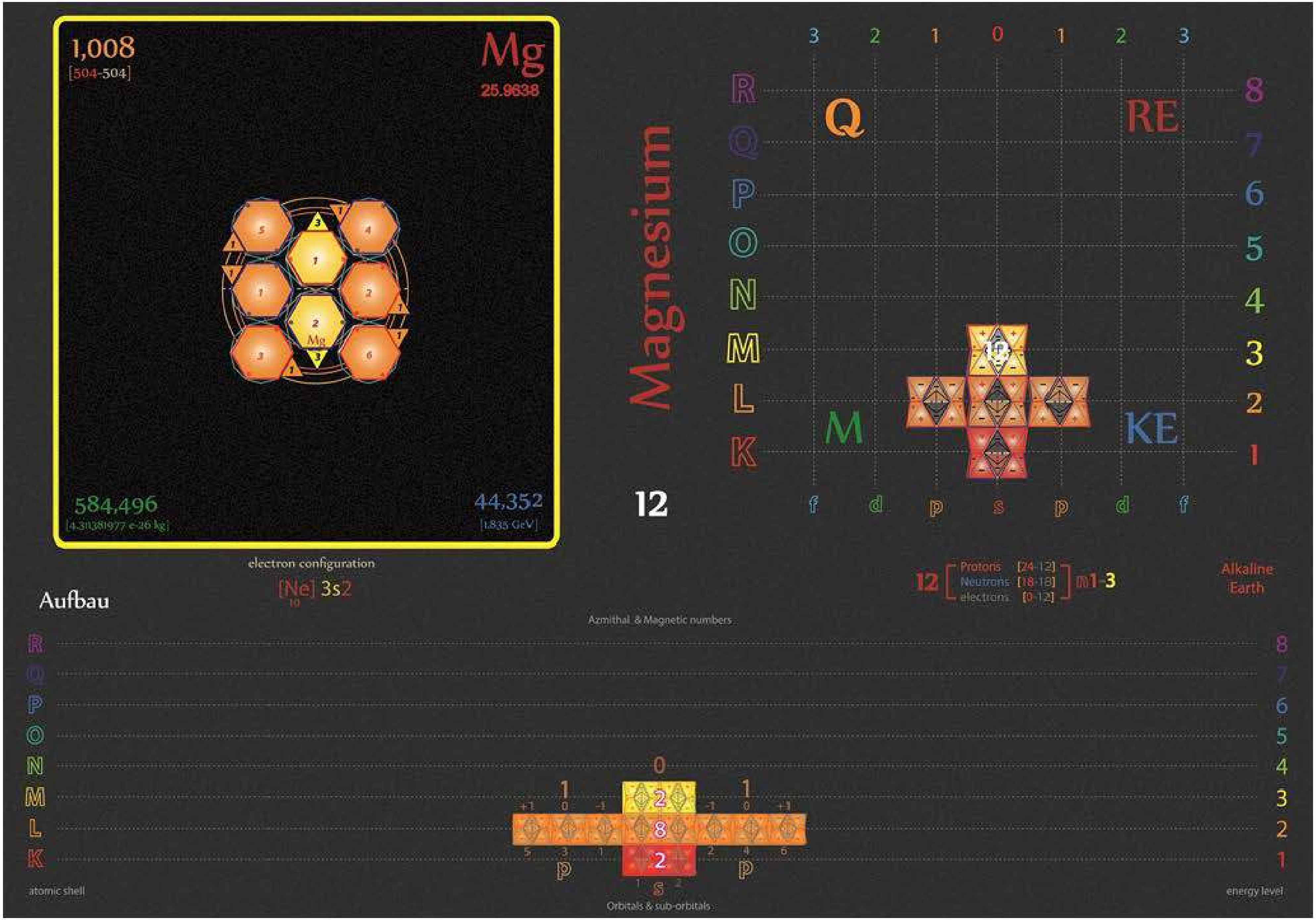
1,008
[504-504]
Mg
25.9838
5
4
3
1
1
2
2
Mg
3
6
584,496
[4.31381977 e-26 kg]
44,352
[1.835 GeV]
electron configuration
[Ne] 3s2
10
Aufbau
Magnesium
12
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
12
f d p s p d f
12
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Alkaline Earth
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
1
+1 0 -1
-1 0 +1
2
8
2
5 3 1
2 4 6
p
p
1 2
s
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.12 - Magnesium atom

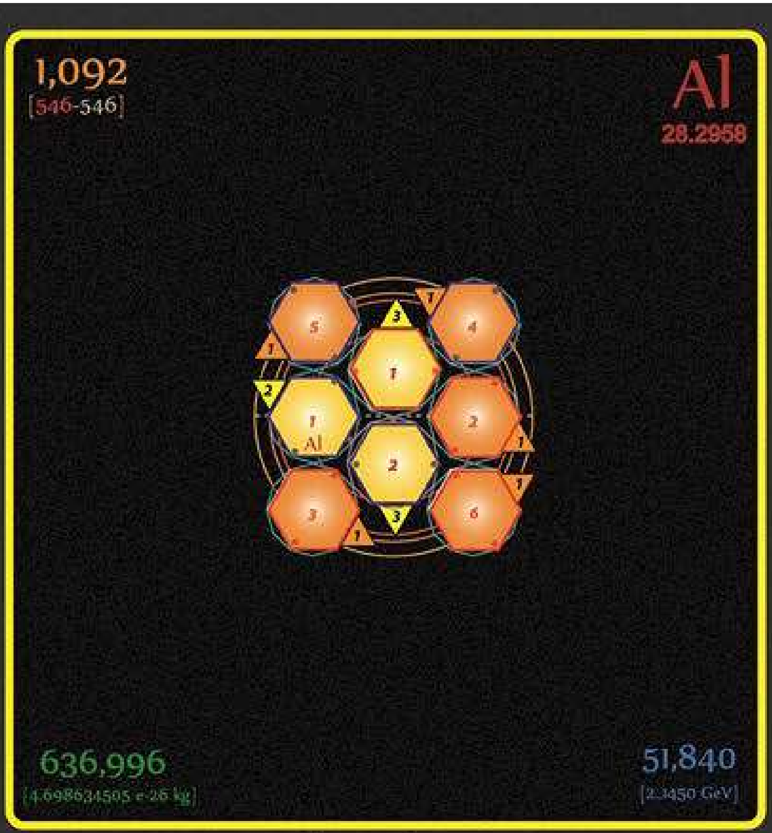
1,092
[546-546]
Al
28.2958
5
4
3
7
1
1
1
2
2
Al
2
7
7
3
3
6
1
636,996
[4.698634505 e-26 kg]
51,840
[2.3450 GeV]

electron configuration

$[\text{Ne}]_{10}\ 3s^2\ 3p^1$

Aufbau

Aluminium

13

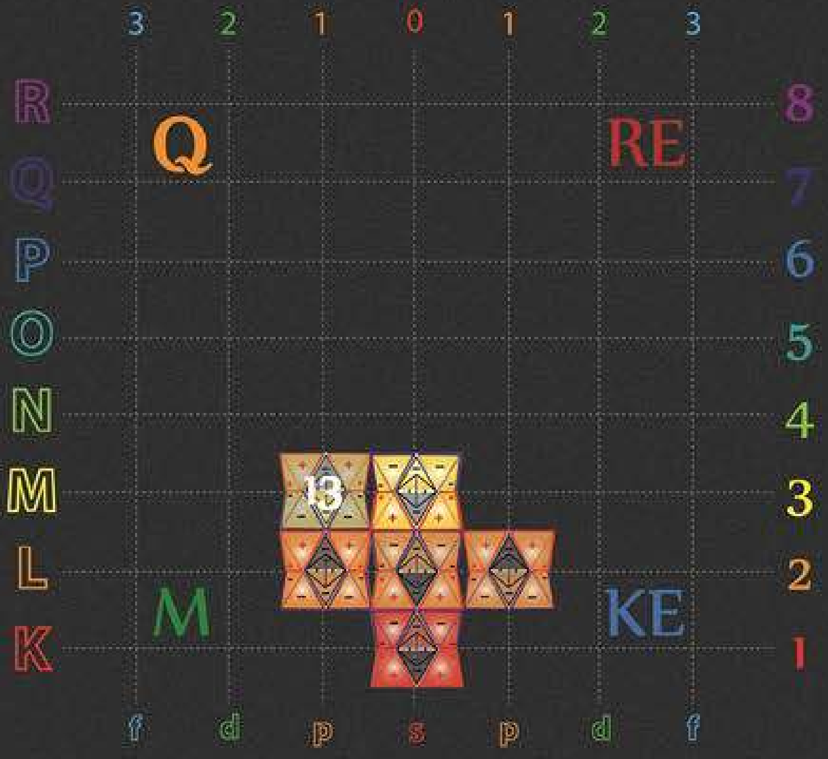
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
13
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

13 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3

Metaloid

Azmithal & Magnetic numbers

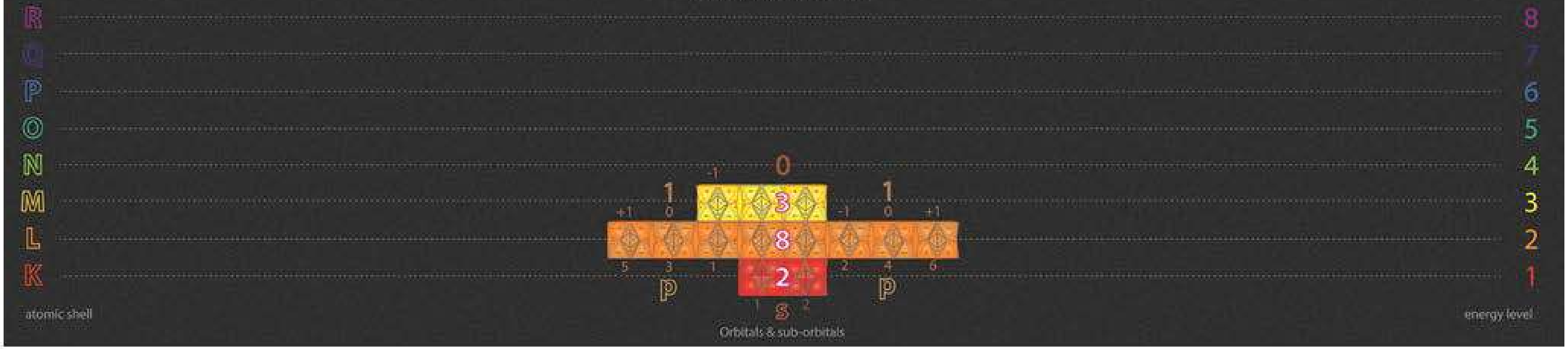
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
-1
1
1
+1
0
-1
0
+1
3
8
2
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.13 - Aluminium atom

Copyright ABRAHAM [2008] - All rights reserved

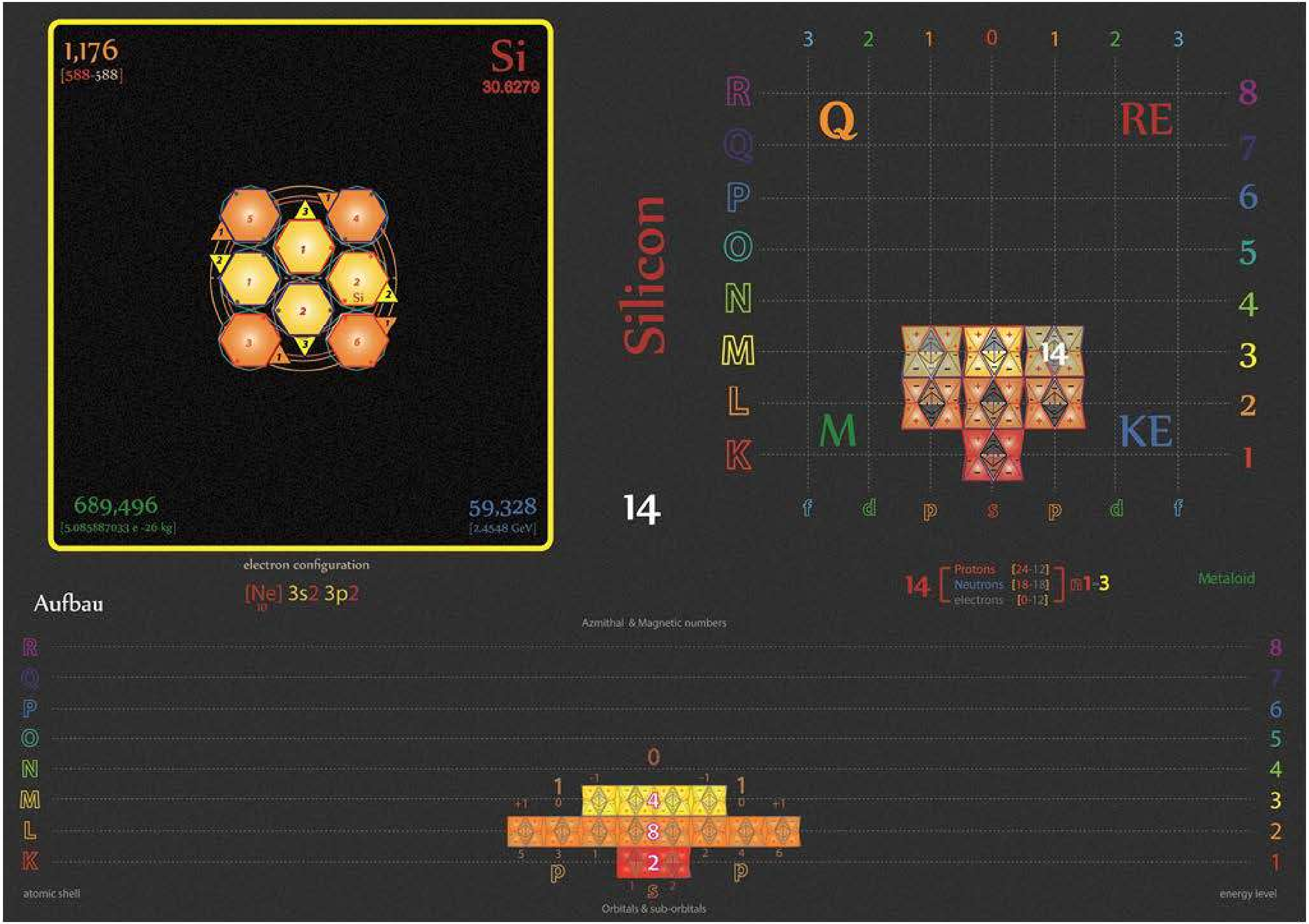
1,176
[588-588]
Si
30.6279
5
4
3
1
2
1
2
Si
2
3
6
1
689,496
[5.085887033 e -26 kg]
59,328
[2.4548 GeV]
Silicon
14
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
14
f d p s p d f
electron configuration
[Ne] 3s2 3p2
10
Aufbau
14
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
1
-1
-1
+1
0
0
+1
4
8
2
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.14 - Silicon atom

Copyright ABRAHAM [2008] - All rights reserved

1,260
[630-630]

P
32.9600

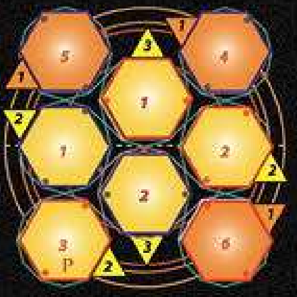
5
4
3
1
1
2
1
2
2
2
1
3
p
2
3
6
1

741,996
[5.473139562 e-26 kg]

66,816
[2.7646 GeV]

Phosphorus

15

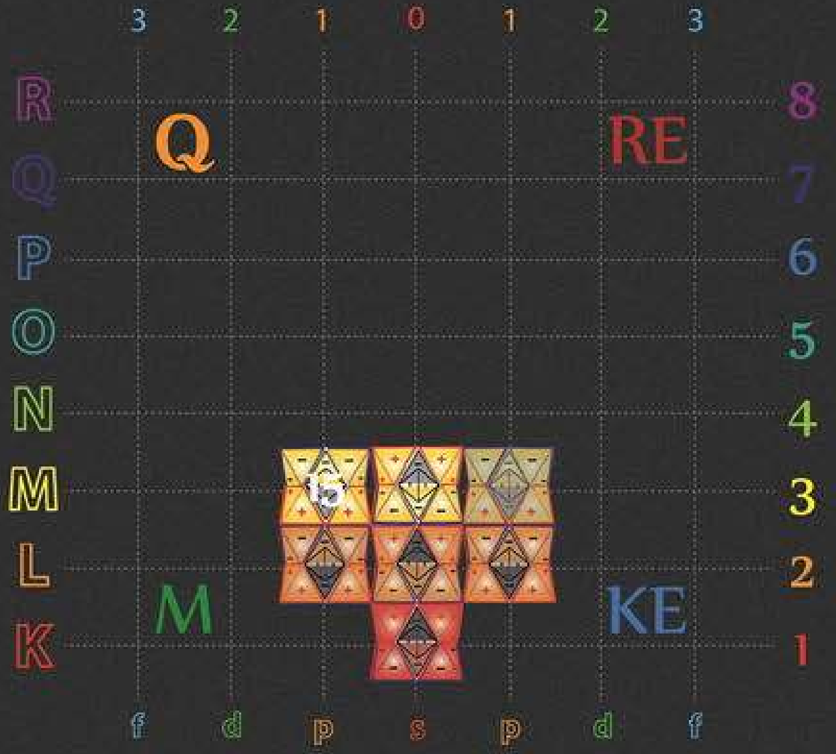
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
15
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

[Ne]$_{10}$ 3s2 3p3

15 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3

Metaloid

Aufbau

Azmithal & Magnetic numbers

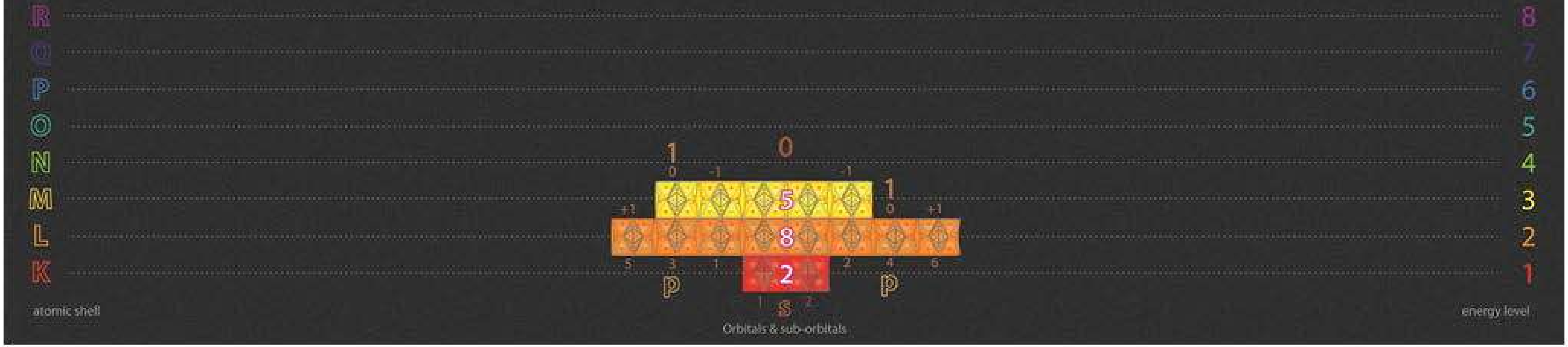
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
0
-1
-1
1
+1
0
+1
5
8
2
5
3
1
2
4
6
p
p
1
s
2

atomic shell

energy level

Orbitals & sub-orbitals

Tetryonics 51.15 - Phosphorus atom

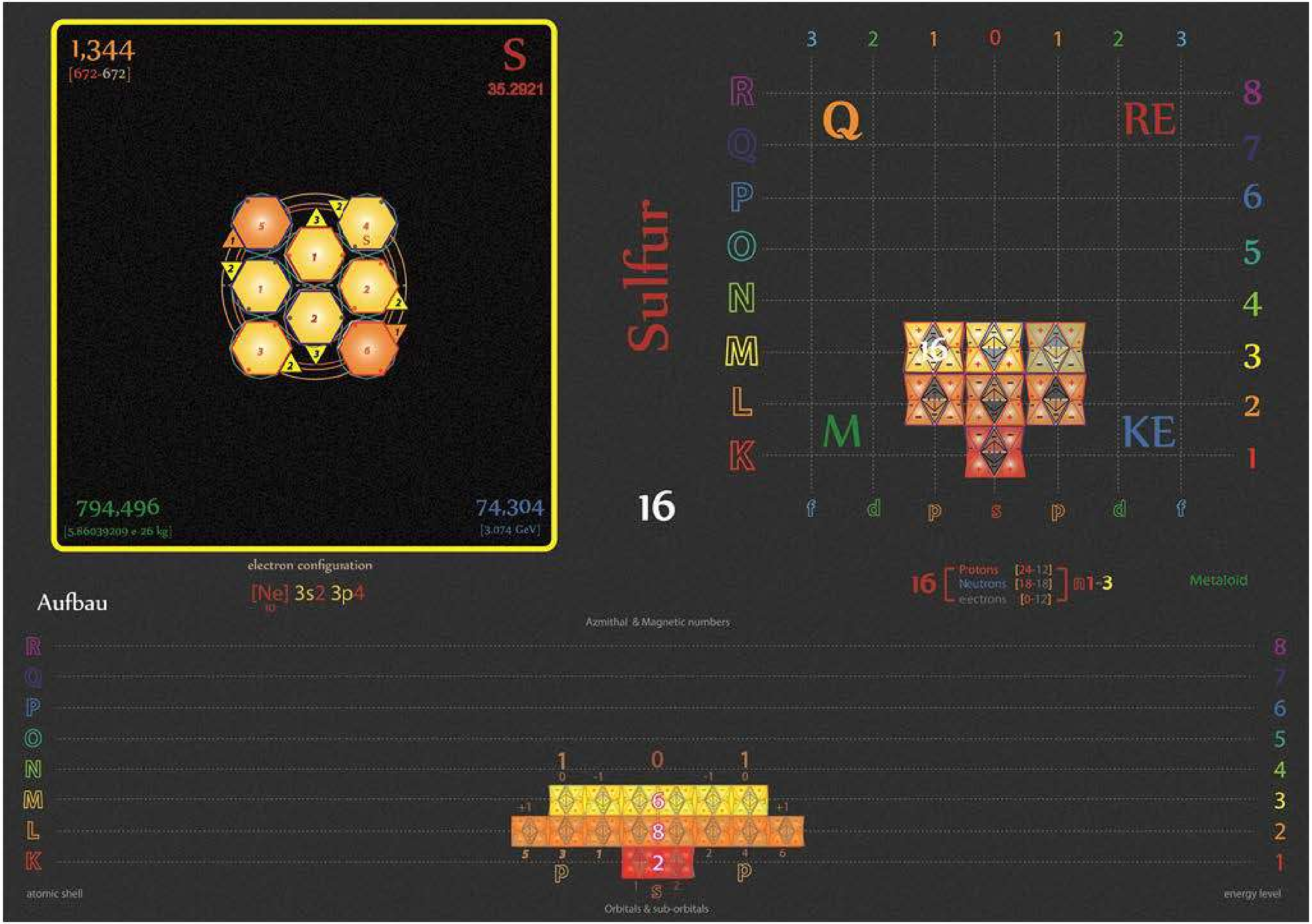
1,344
[672-672]
S
35.2921
794,496
[5.86039209 e-26 kg]
74,304
[3.074 GeV]
Sulfur
16
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Ne] 3s2 3p4
10
Aufbau
16 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Metaloid
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
+1 +1
6
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.16 - Sulfur atom

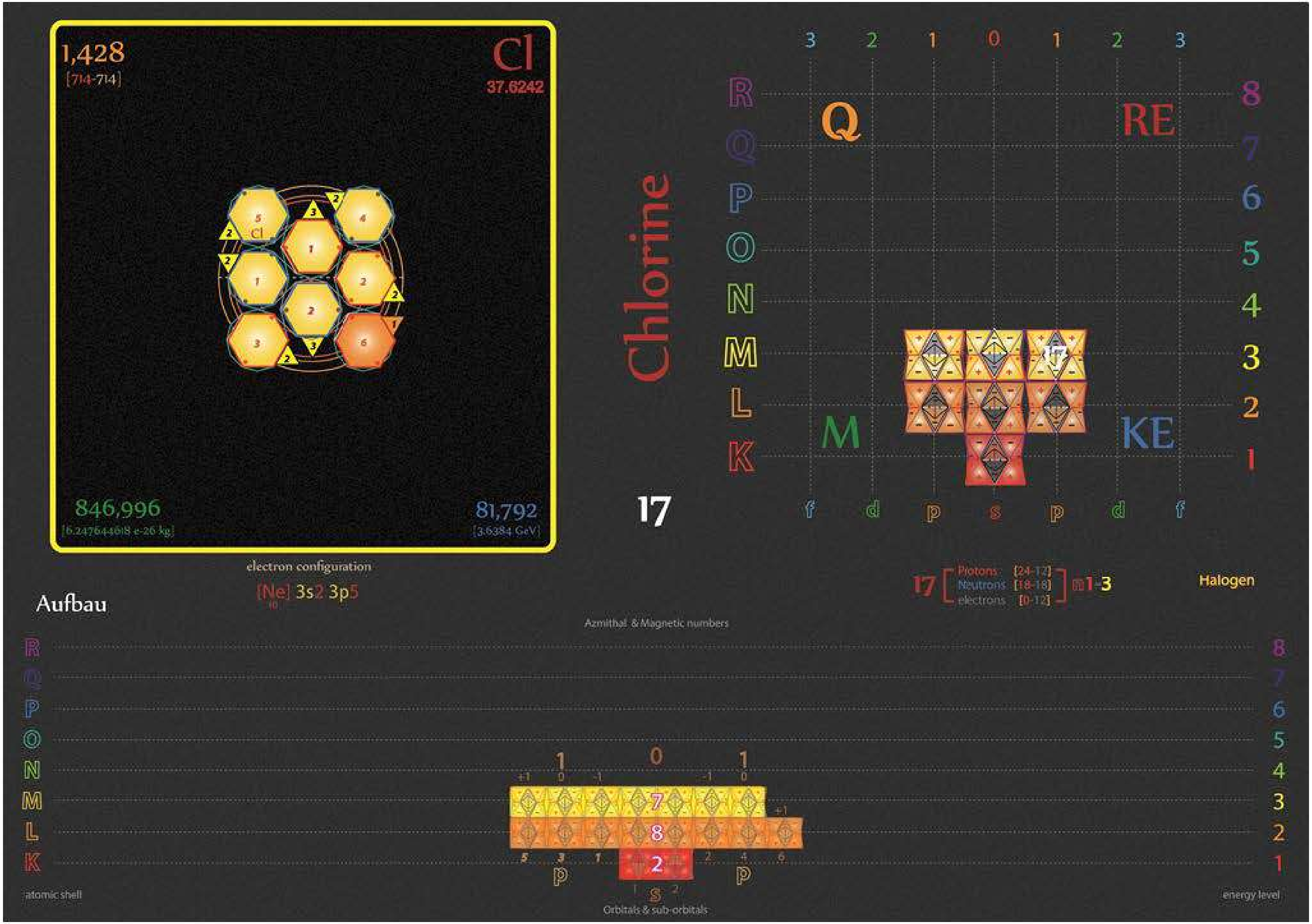
1,428
[714-714]
Cl
37.6242
5
Cl
4
1
1
2
2
3
6
846,996
[6.247644618 e-26 kg]
81,792
[3.6384 GeV]
Chlorine
17
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
17
f d p s p d f
17 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-3
Halogen
electron configuration
[Ne] 3s2 3p5
10
Aufbau
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0 +1
7
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.17 - Chlorine atom

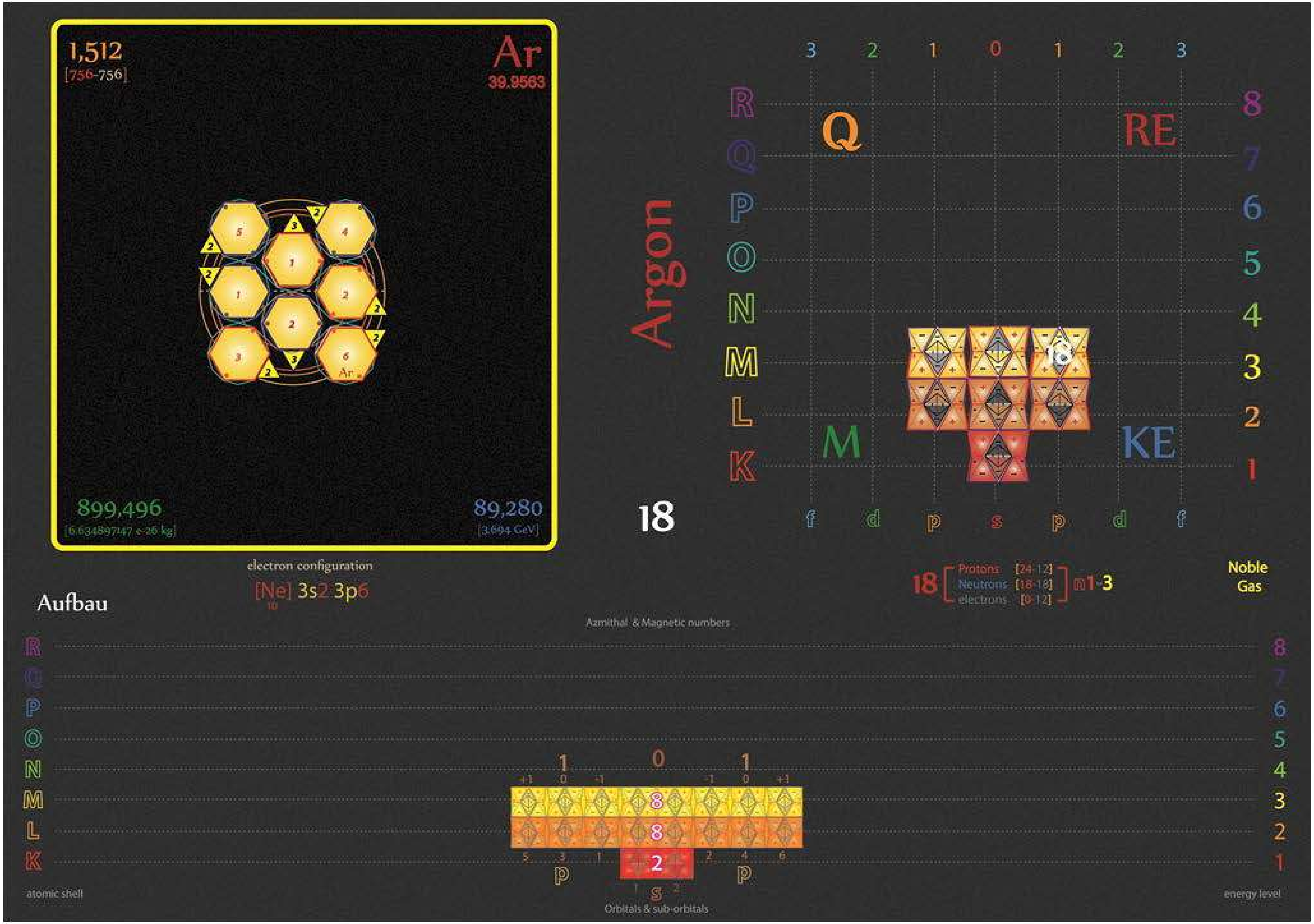
1,512
[756-756]
Ar
39.9563
5
4
3
1
2
2
1
2
2
3
6
Ar
899,496
[6.634897147 e-26 kg]
89,280
[3.694 GeV]
electron configuration
[Ne] 3s2 3p6
10
Aufbau
Argon
18
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
18
f d p s p d f
18
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-3
Noble Gas
Azimuthal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0 +1
8
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.18 - Argon atom

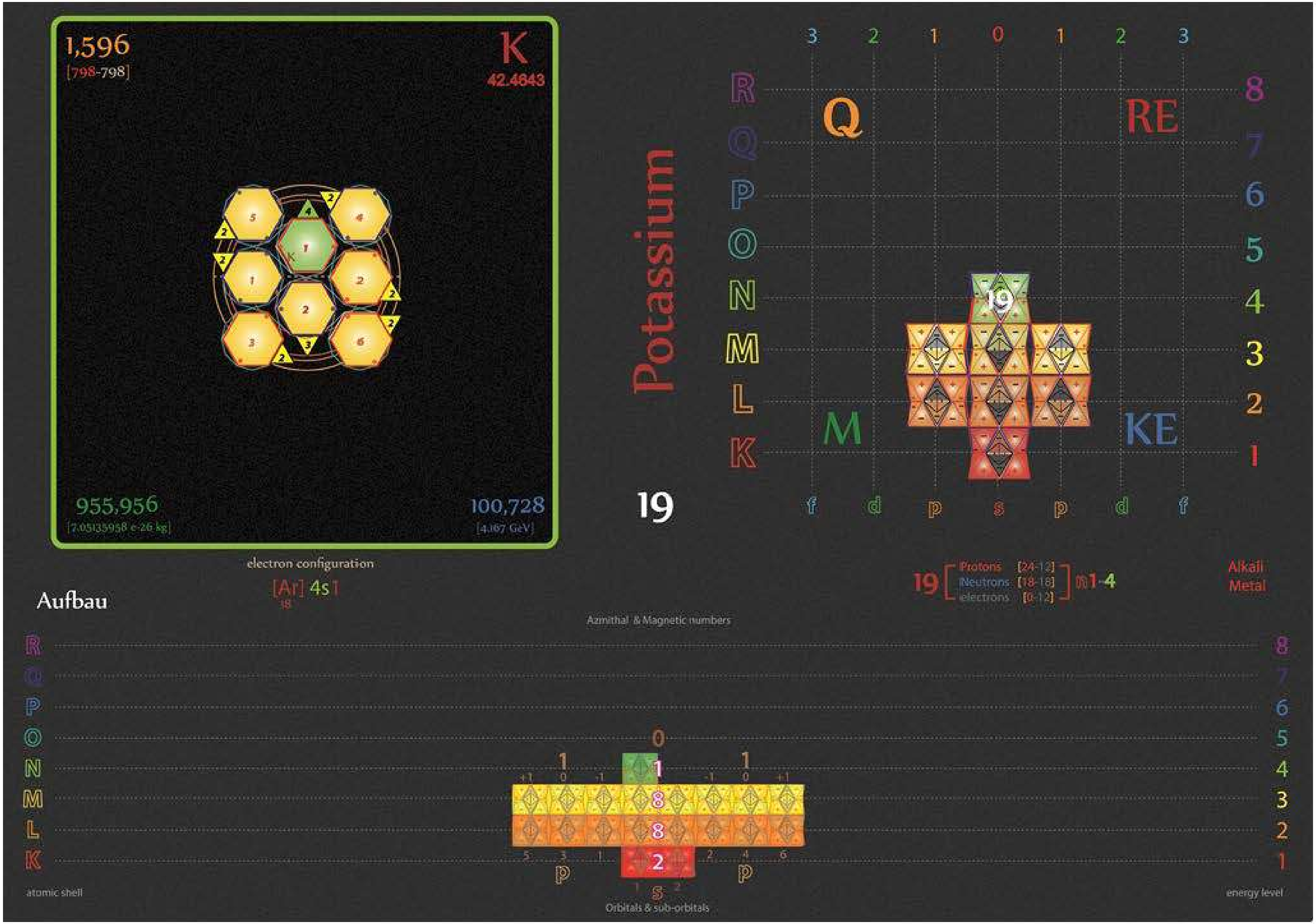
1,596
[798-798]
K
42.4643
5
4
1
2
1
2
3
6
2
2
2
2
4
2
3
2
955,956
[7.05135958 e-26 kg]
100,728
[4.107 GeV]
Potassium
19
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
19
f
d
p
s
p
d
f
electron configuration
[Ar] 4s1
18
Aufbau
19
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-4
Alkali
Metal
Azimithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
+1
0
-1
1
-1
0
+1
8
8
5
3
1
2
2
4
6
p
p
1
s
2
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.19 - Potassium atom

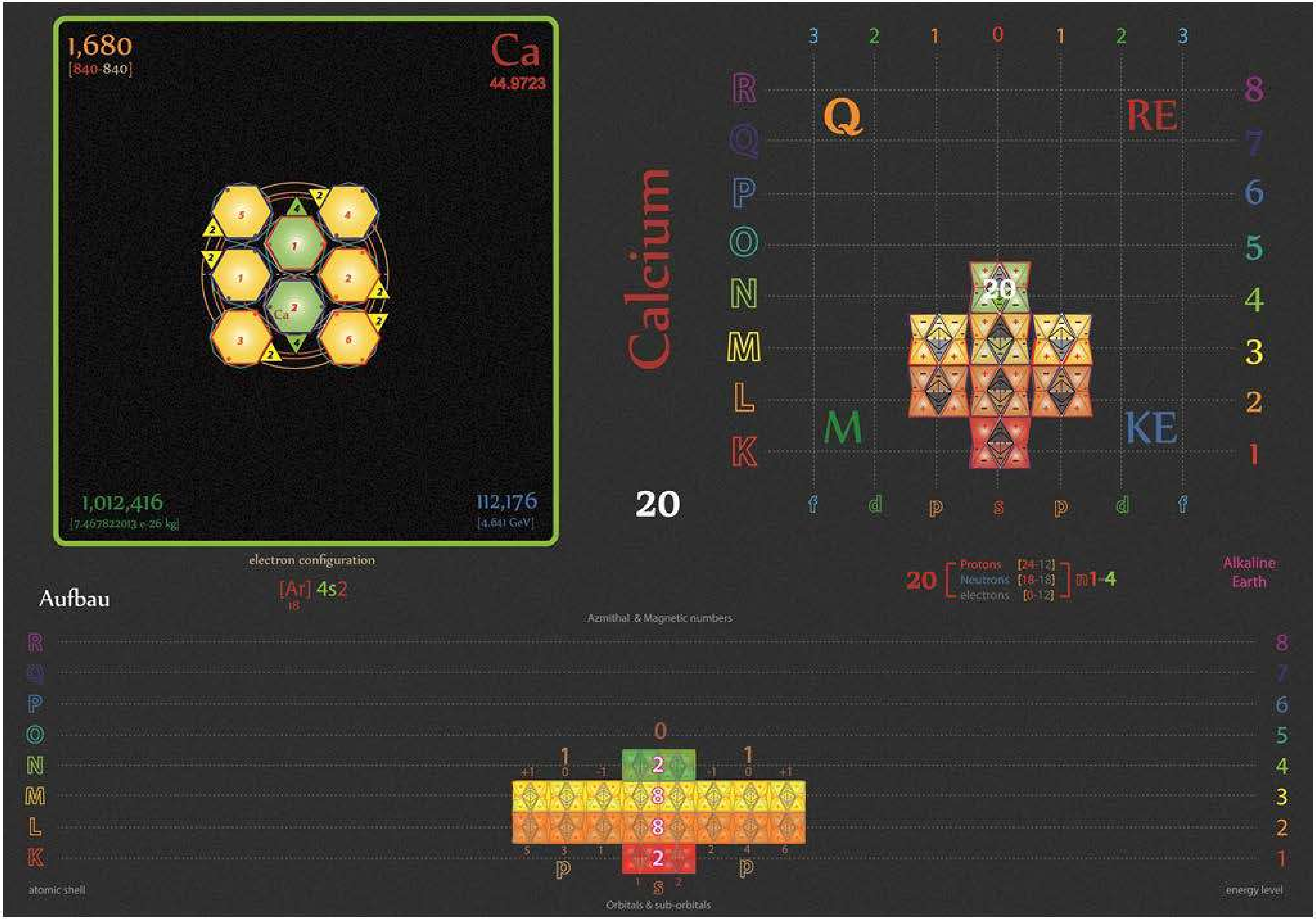
1,680
[840-840]
Ca
44.9723
1,012,416
[7.467822013 e-26 kg]
112,176
[4.641 GeV]
Calcium
20
electron configuration
[Ar] 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
20 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-4
Alkaline Earth
Azmithal & Magnetic numbers
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.20 - Calcium atom

Copyright ABRAHAM [2008] - All rights reserved

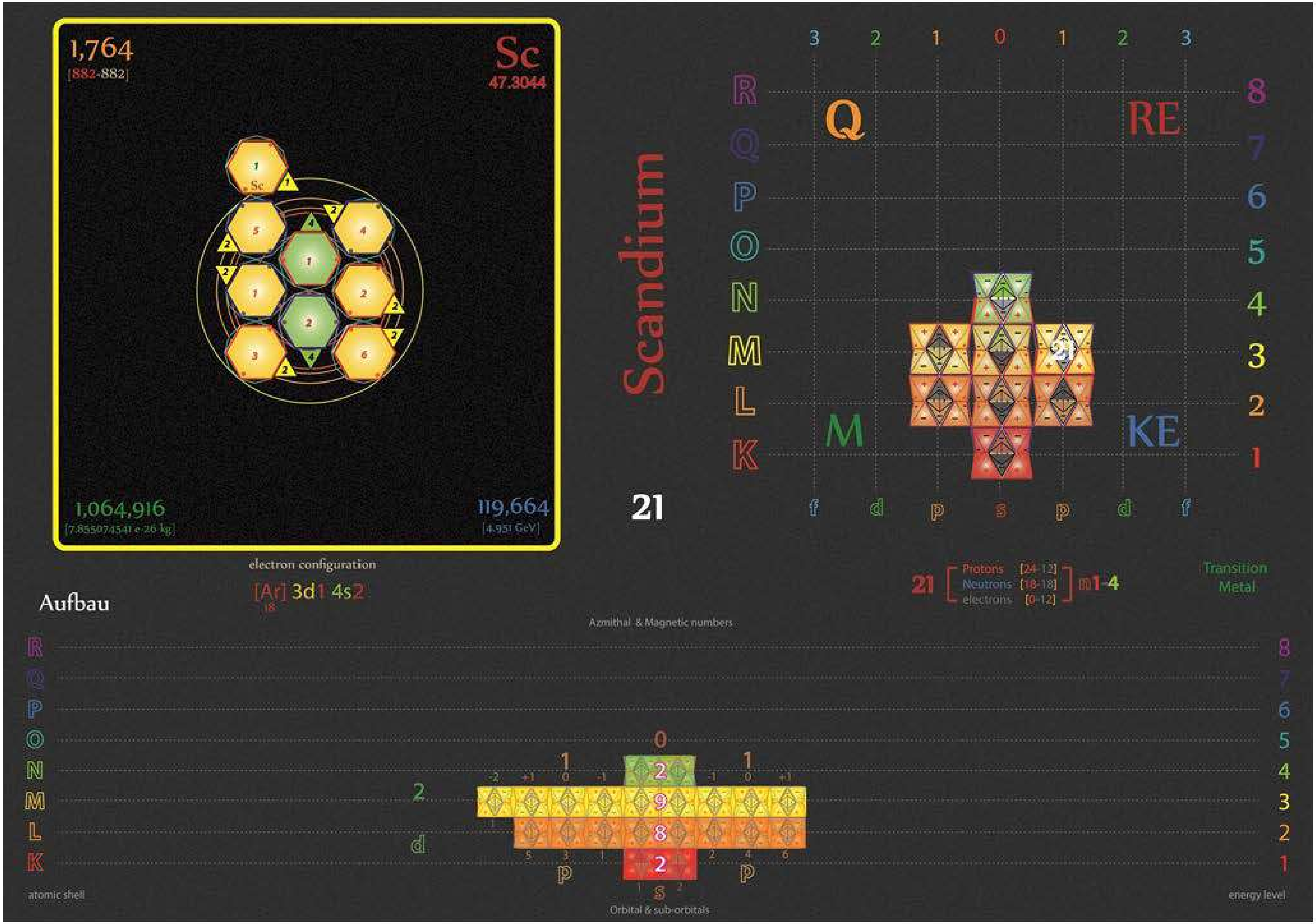
1,764
[882-882]
Sc
47.3044
1,064,916
[7.855074541 e-26 kg]
119,664
[4.951 GeV]
Scandium
21
electron configuration
[Ar] 3d1 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
21
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Transition Metal
Azmithal & Magnetic numbers
0
1
1
-2 +1 0 -1 -1 0 +1
2
9
8
2
2
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.21 - Scandium atom

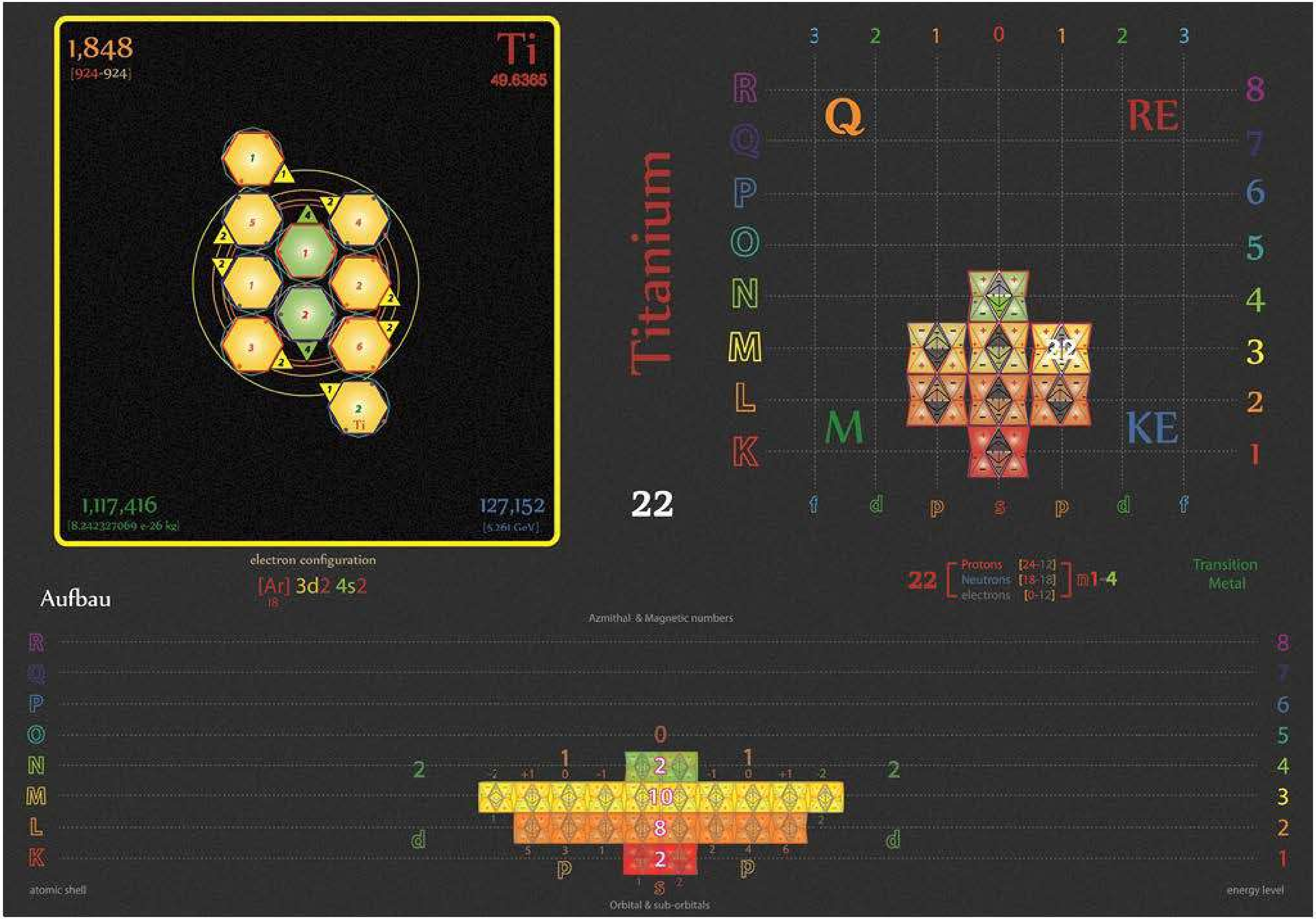
1,848
[924-924]
Ti
49.6365
1,117,416
[8.242327069 e-26 kg]
127,152
[5.261 GeV]
Titanium
22
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
22
f
d
p
s
p
d
f
electron configuration
[Ar] 3d2 4s2
18
Aufbau
22
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Transition
Metal
Azmithal & Magnetic numbers
0
1
1
2
2
-2
+1
0
-1
-1
0
+1
2
2
10
8
2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.22 - Titanium atom

Copyright ABRAHAM [2008] - All rights reserved

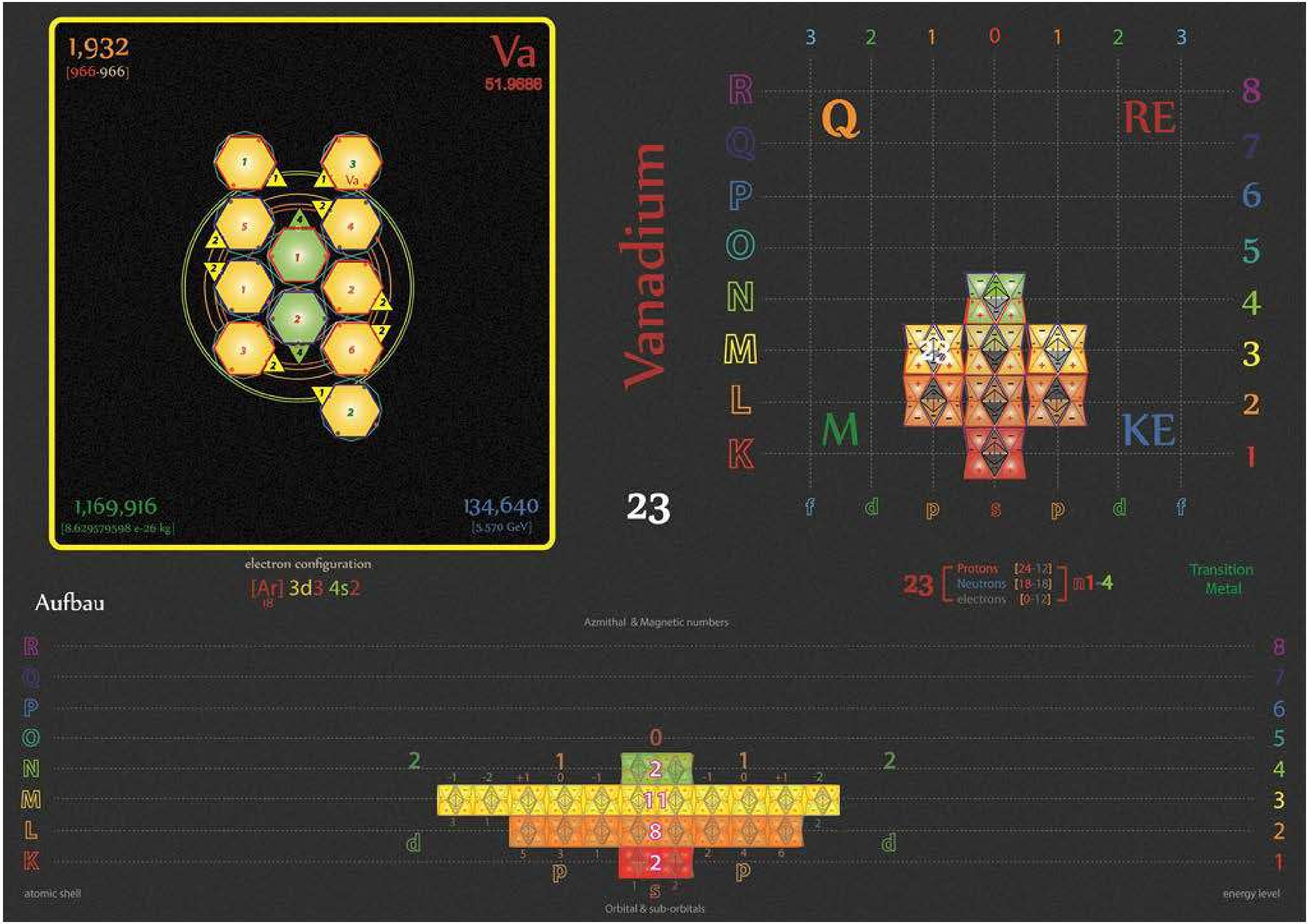
1,932
[966-966]
Va
51.9888
1,169,916
[8.629579598 e-26 kg]
134,640
[5.570 GeV]
Vanadium
23
electron configuration
[Ar] 3d3 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
23 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
11
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.23 - Vanadium atom

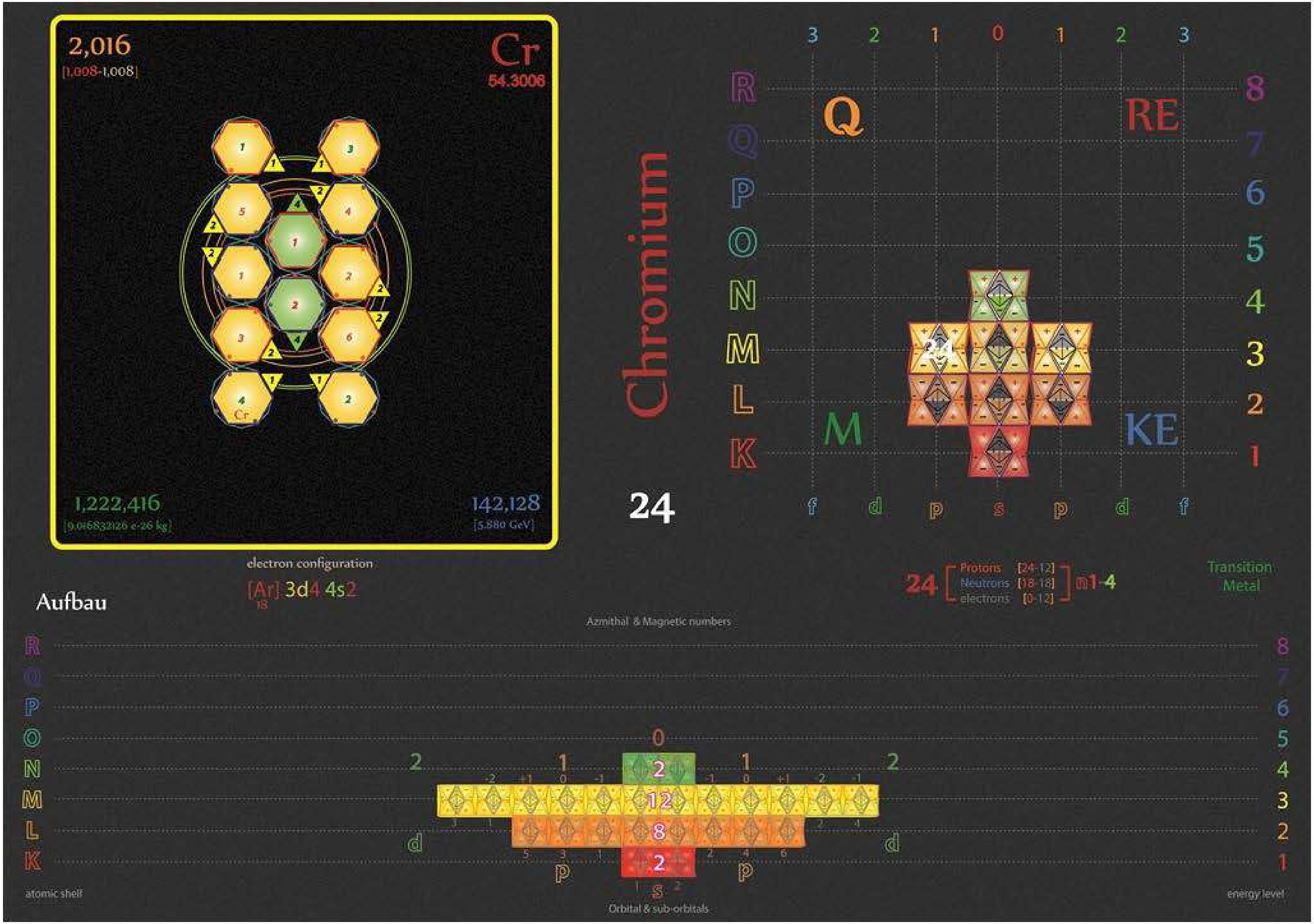
2,016
[1,008-1,008]
Cr
54.3008
1,222,416
[9.01683212б e-26 kg]
142,128
[5.880 GeV]
Chromium
24
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
24
f d p s p d f
electron configuration
[Ar] 3d4 4s2
18
Aufbau
24 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
12
8
2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.24 - Chromium atom

Copyright ABRAHAM [2008] - All rights reserved

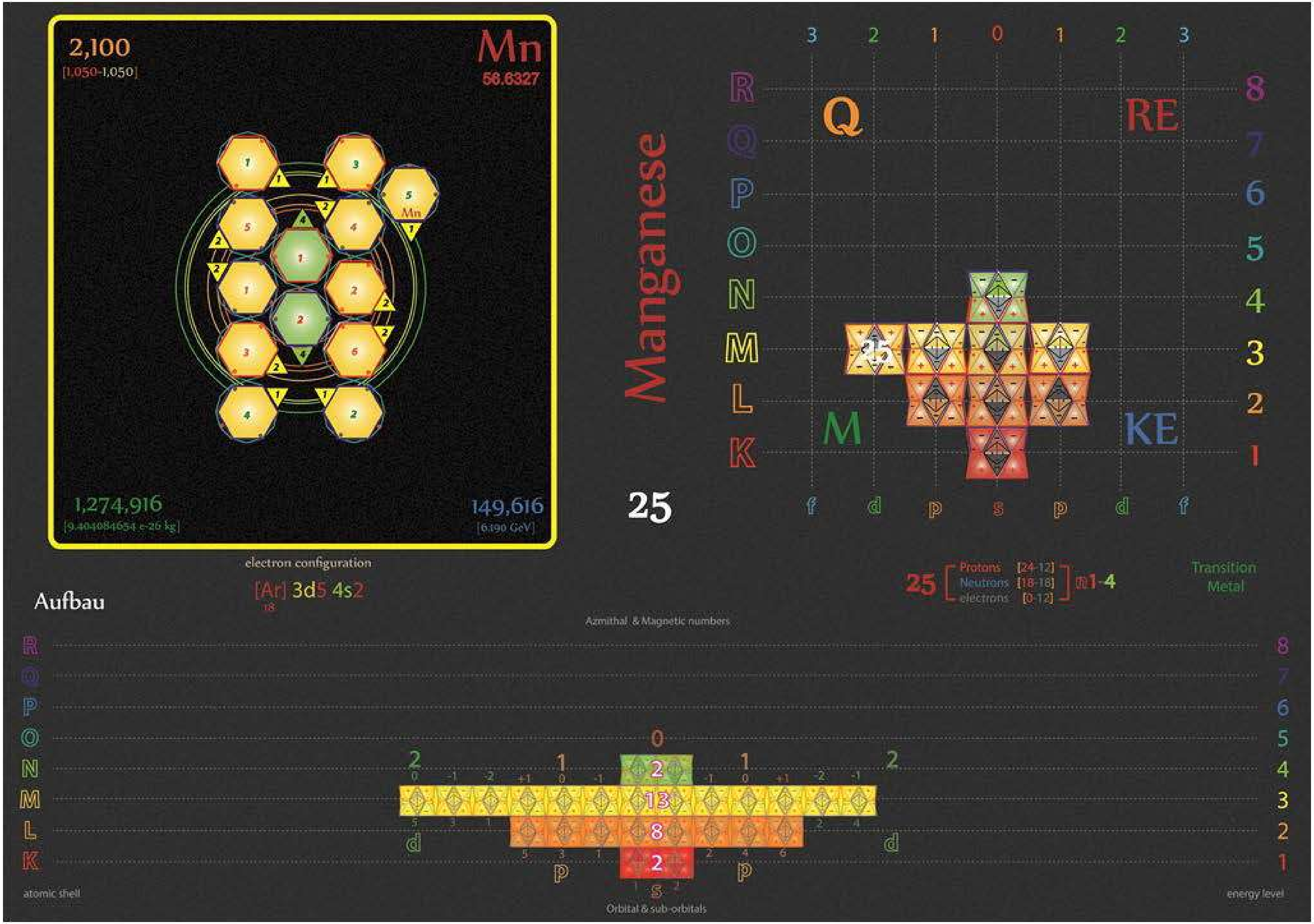
2,100
[1,050-1,050]
Mn
56.6327
1,274,916
[9.404084654 e-26 kg]
149,616
[6.190 GeV]
Manganese
25
electron configuration
[Ar] 3d5 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
25
f d p s p d f
25 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
0 -1 -2 +1 0 -1 -1 0 +1 -2 -1
2
13
8
2
5 3 1 2 4
5 3 1 2 4 6
d d
p p
1 s 2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.25 - Manganese atom

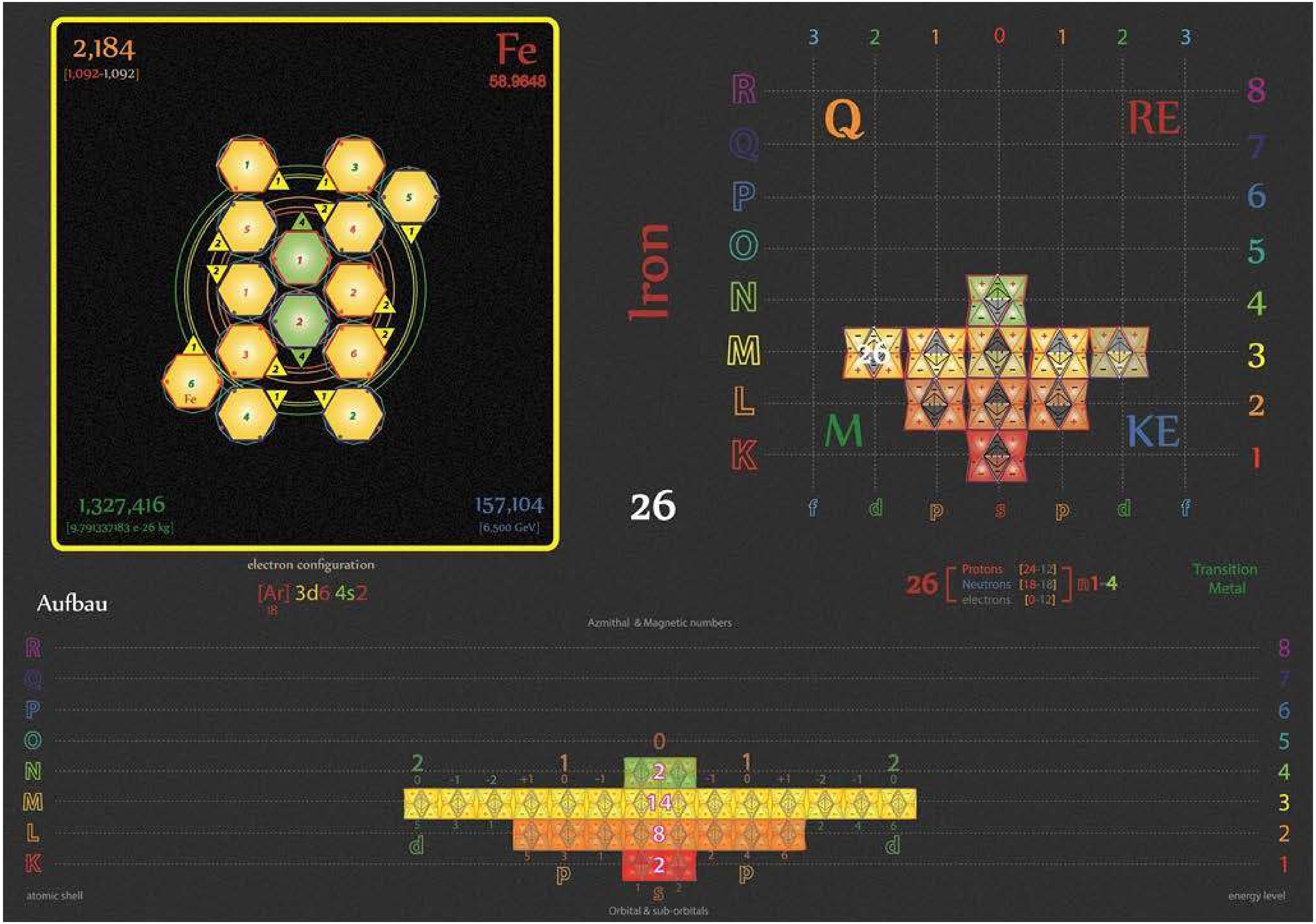
2,184
[1,092-1,092]
Fe
58.9848
1,327,416
[9.791337183 e-26 kg]
157,104
[6,500 GeV]
Iron
26
electron configuration
[Ar] 3d6 4s2
18
Aufbau
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
26
f
d
p
s
p
d
f
26
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Transition Metal
Azmithal & Magnetic numbers
2
1
0
1
2
14
8
2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.26 - Iron atom

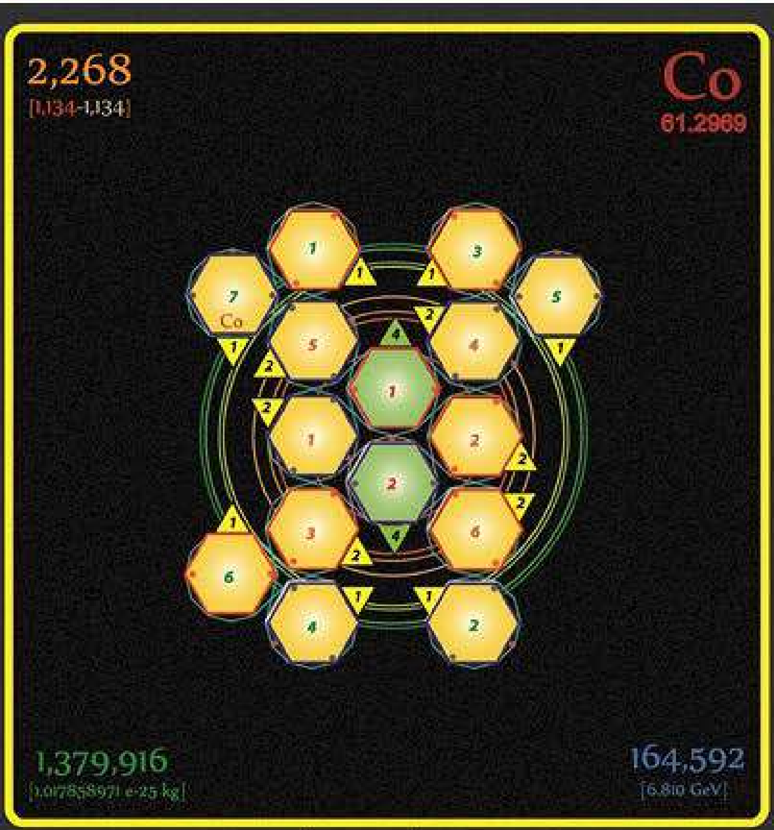
2,268
[1,134-1,134]
Co
61.2969
Co
1,379,916
[1.017858971 e-25 kg]
164,592
[6.810 GeV]

Cobalt

27

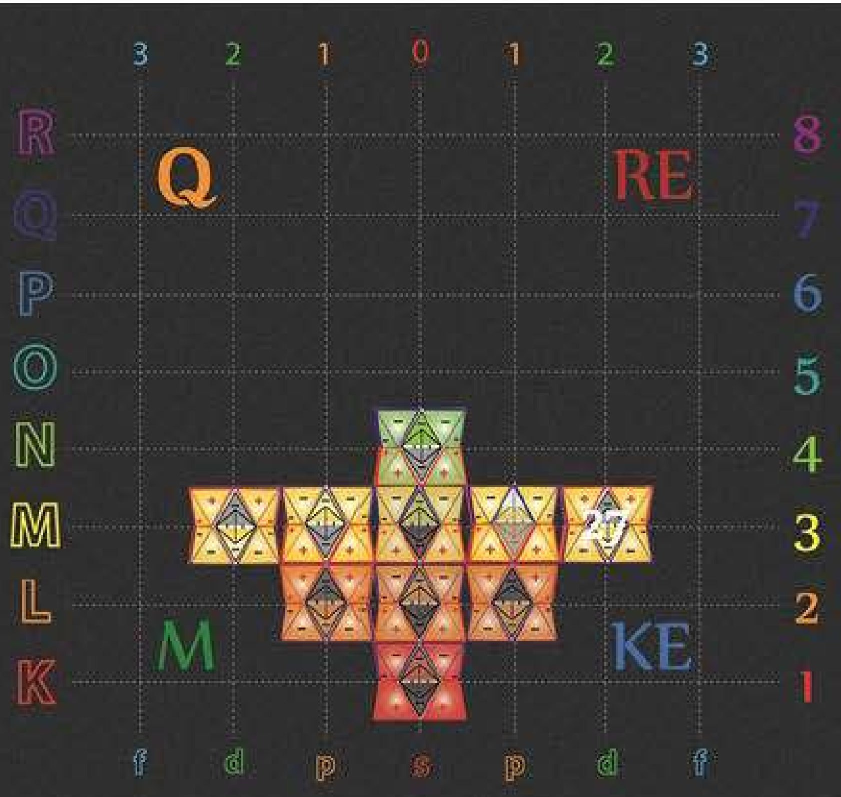
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
27
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

[Ar] 3d7 4s2
18

27 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4

Transition Metal

Aufbau

Azmithal & Magnetic numbers

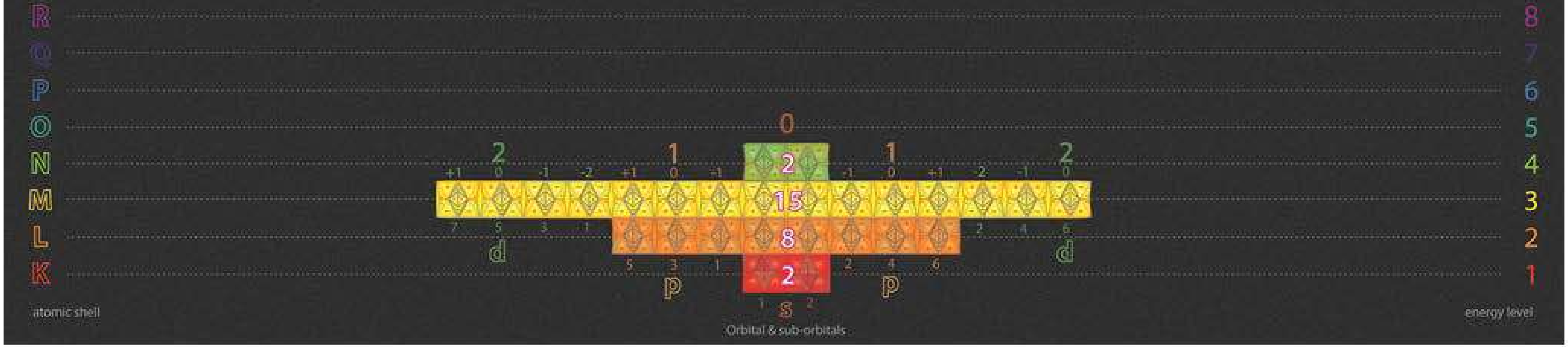
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
2
15
8
2
7
5
3
1
2
4
6
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.27 - Cobalt atom

Copyright ABRAHAM [2008] - All rights reserved

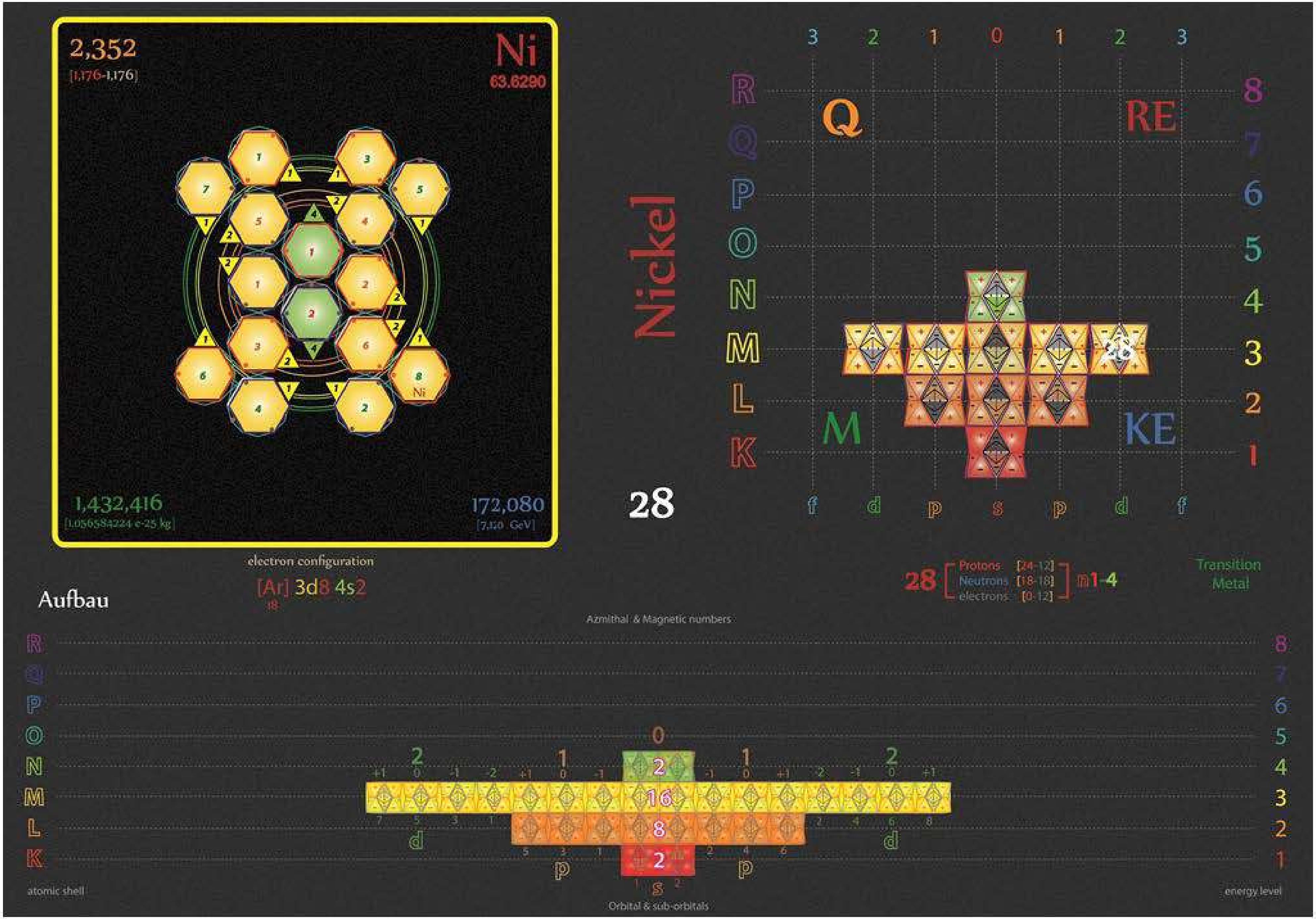
2,352
[1,176-1,176]
Ni
63.6290
1,432,416
[1.056584224 e-25 kg]
172,080
[7,140 GeV]
Nickel
28
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Ar] 3d8 4s2
18
Aufbau
28 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
16
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.28 - Nickel atom

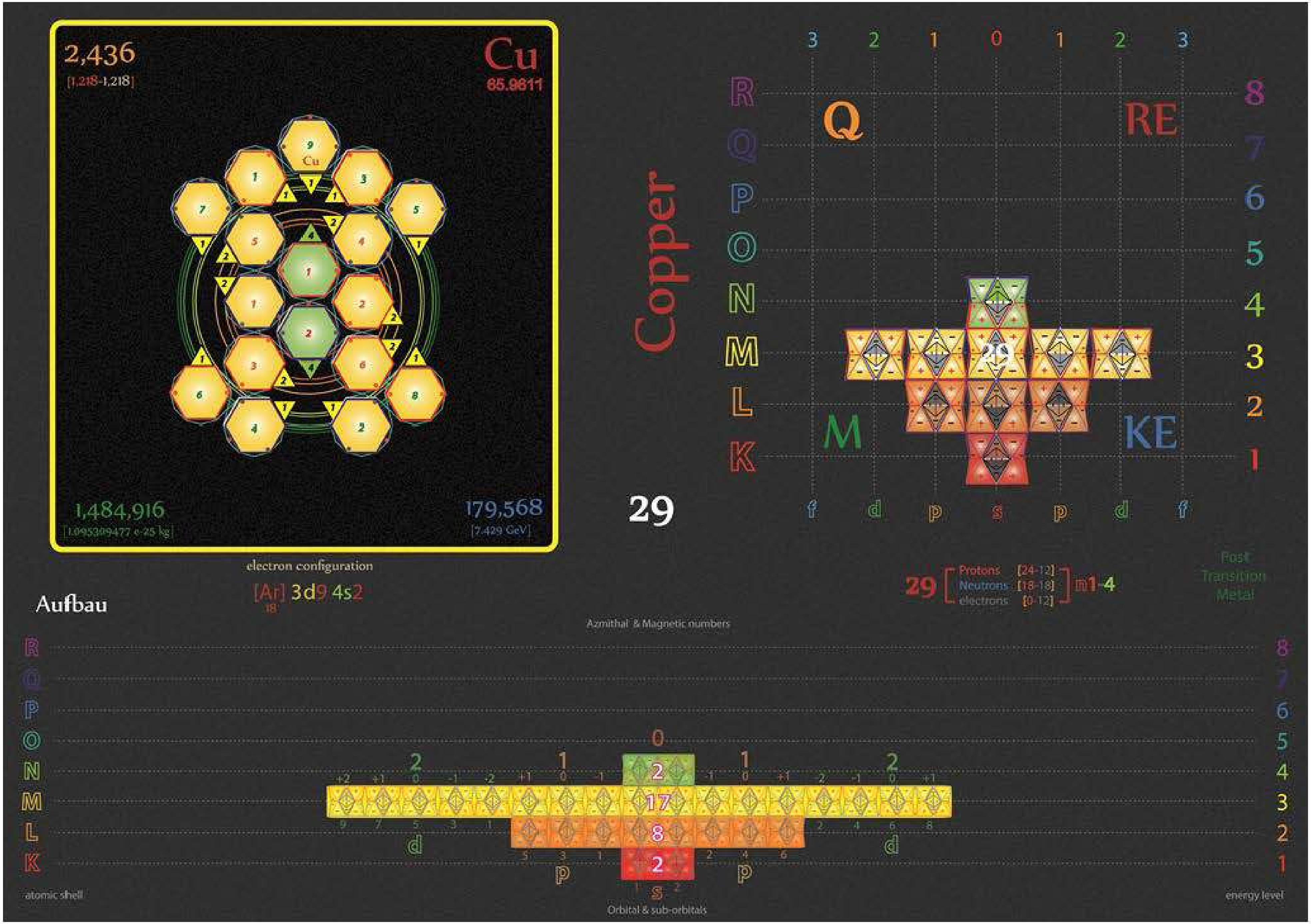
2,436
[1,218-1,218]
Cu
65.9811
9
Cu
1,484,916
[1.095309477 e-25 kg]
179,568
[7.429 GeV]
electron configuration
[Ar] 3d9 4s2
18
Aufbau
Copper
29
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
29
29 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Post Transition Metal
Azmithal & Magnetic numbers
0
1
2
-2 -1 0 +1 +2
17
8
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.29 - Copper atom

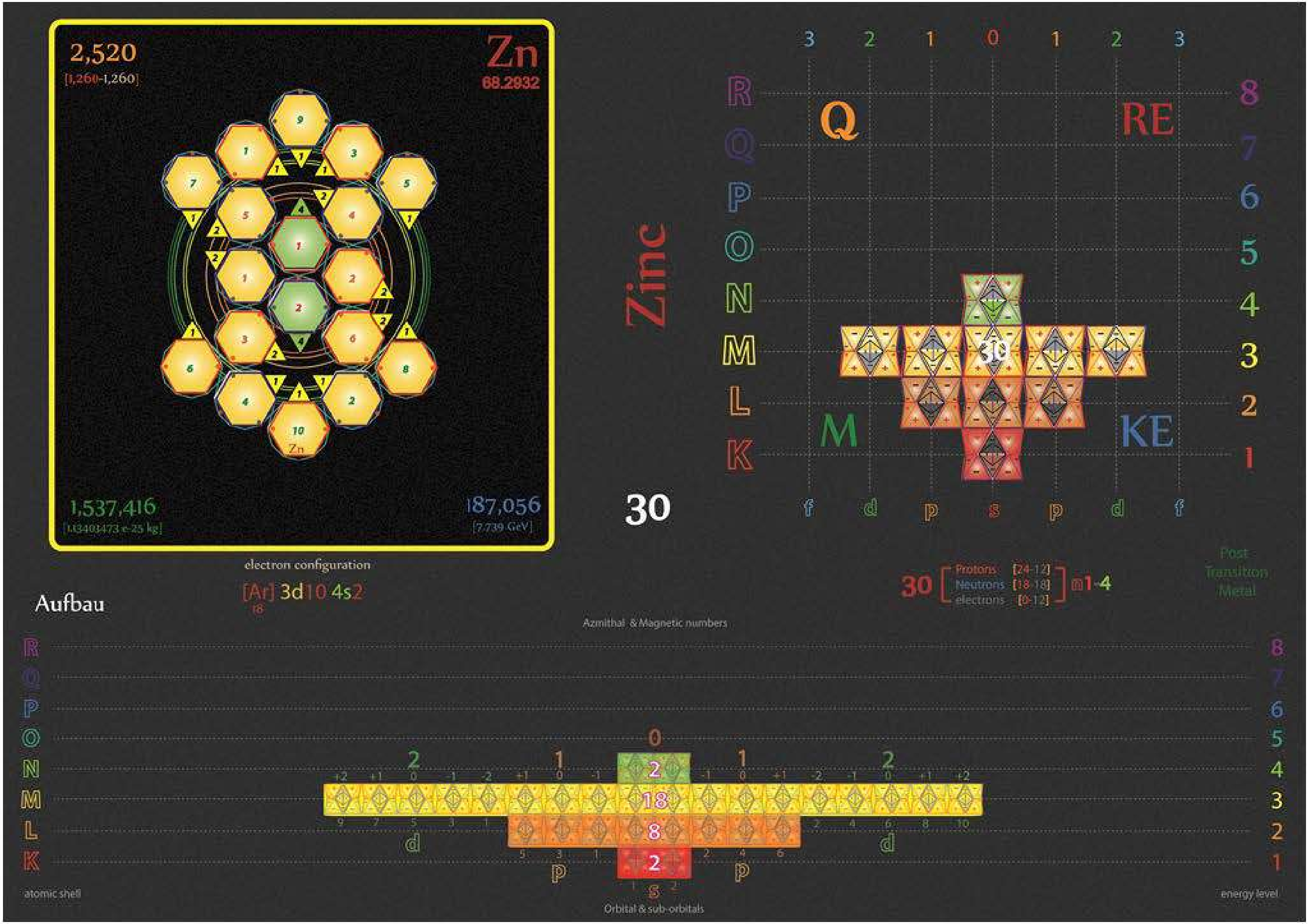
2,520
[1,260-1,260]
Zn
68.2932
Zn
1,537,416
[1.13403473 e-25 kg]
187,056
[7.739 GeV]
electron configuration
[Ar] 3d10 4s2
18
Aufbau
Zinc
30
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
30
f d p s p d f
30 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Post - Transition Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.30 - Zinc atom

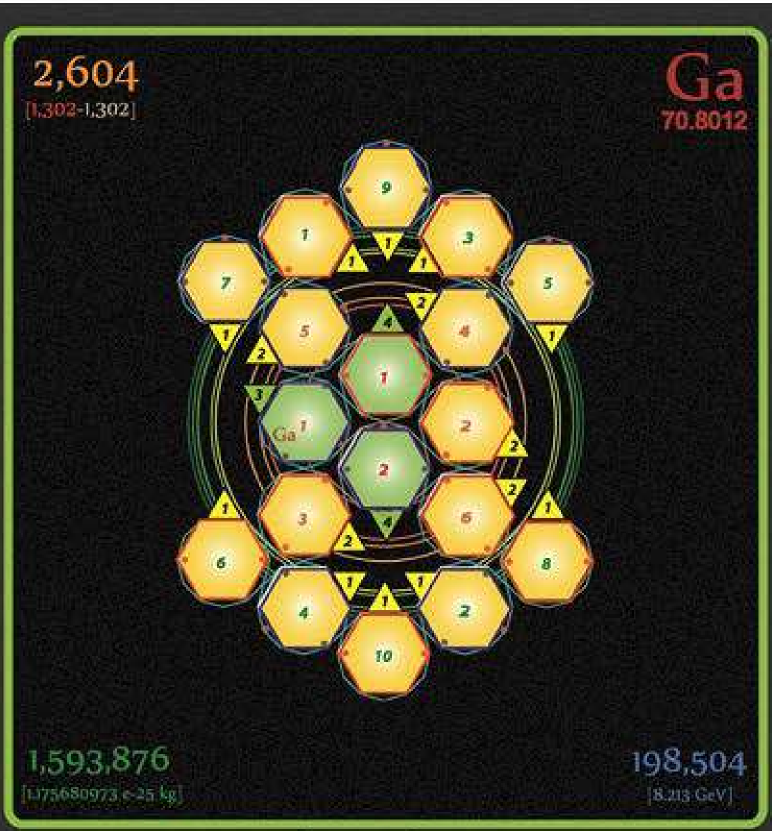
2,604
[1,302-1,302]
Ga
70.8012
9
1
3
7
5
1
1
1
2
5
4
4
1
1
1
3
Ga
1
2
2
2
3
2
3
6
1
1
6
8
2
4
1
1
1
4
2
10
1,593,876
[1.175680973 e-25 kg]
198,504
[8.213 GeV]

electron configuration

Aufbau

[Ar] 3d10 4s2 4p1
18

Gallium

31

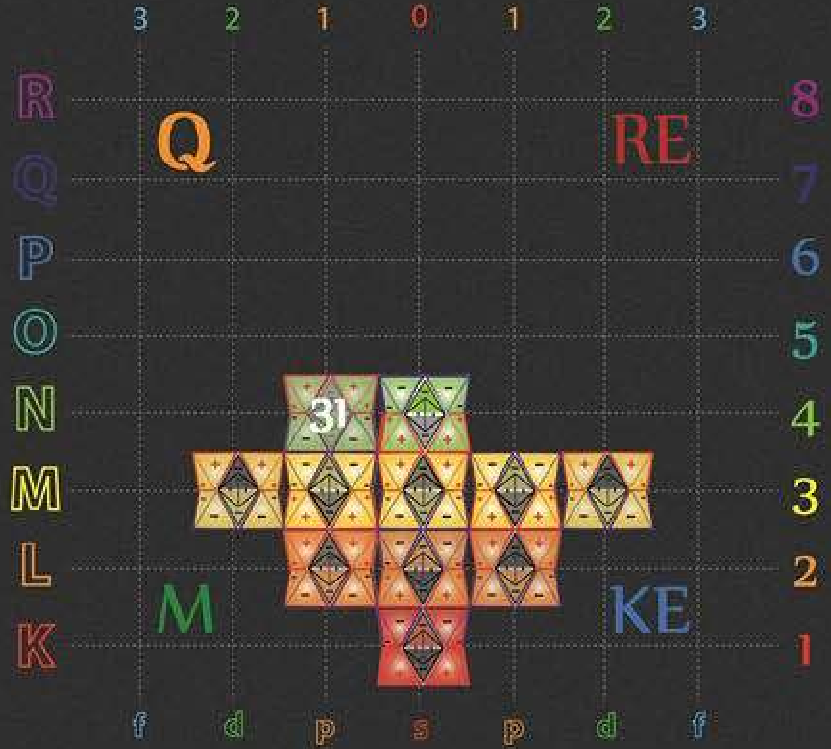
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
31
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

31 [Protons [24-12] / Neutrons [18-18] / electrons [0-12]] m1-4

Metaloid

Azmithal & Magnetic numbers

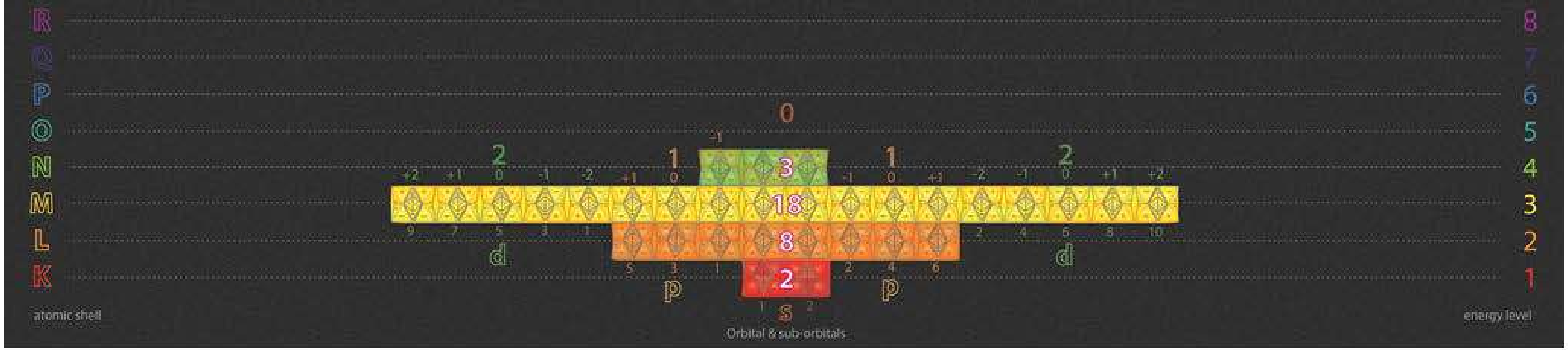
R
Q
P
O
N
M
L
K
0
-1
2
1
3
1
2
+2
+1
0
-1
-2
+1
0
-1
0
+1
-2
-1
0
+1
+2
18
9
7
5
3
1
2
4
6
8
10
d
8
5
3
1
2
4
6
d
2
p
p
1
s
2
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbital & sub-orbitals

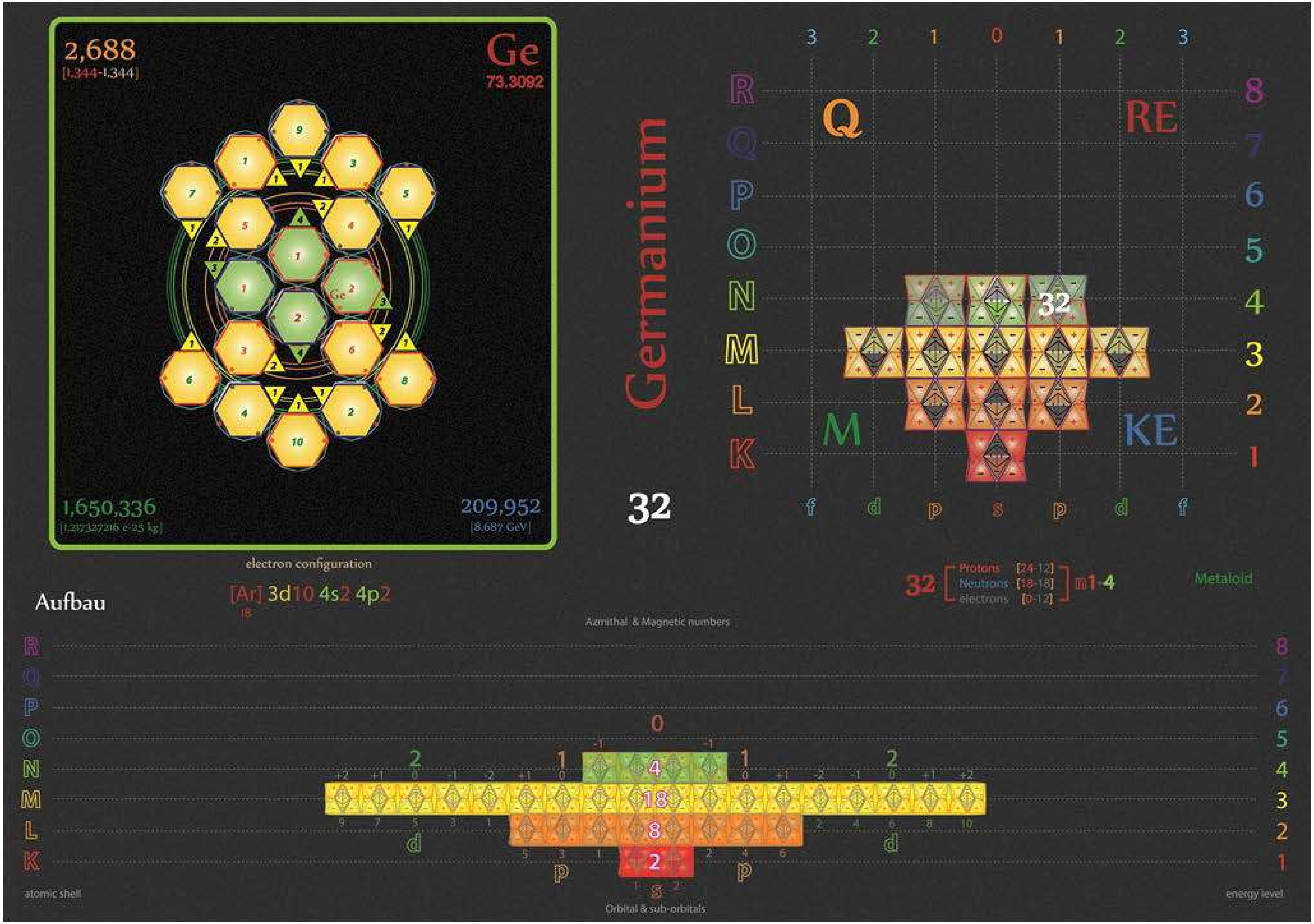
2,688
[1.344-1.344]
Ge
73.3092
1,650,336
[1.217327216 e-25 kg]
209,952
[8.687 GeV]
Germanium
32
electron configuration
[Ar] 3d10 4s2 4p2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
32
f d p s p d f
32 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Metaloid
Azmithal & Magnetic numbers
0
2 1 1 2
4
18
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.32 - Germanium atom

Copyright ABRAHAM [2008] - All rights reserved

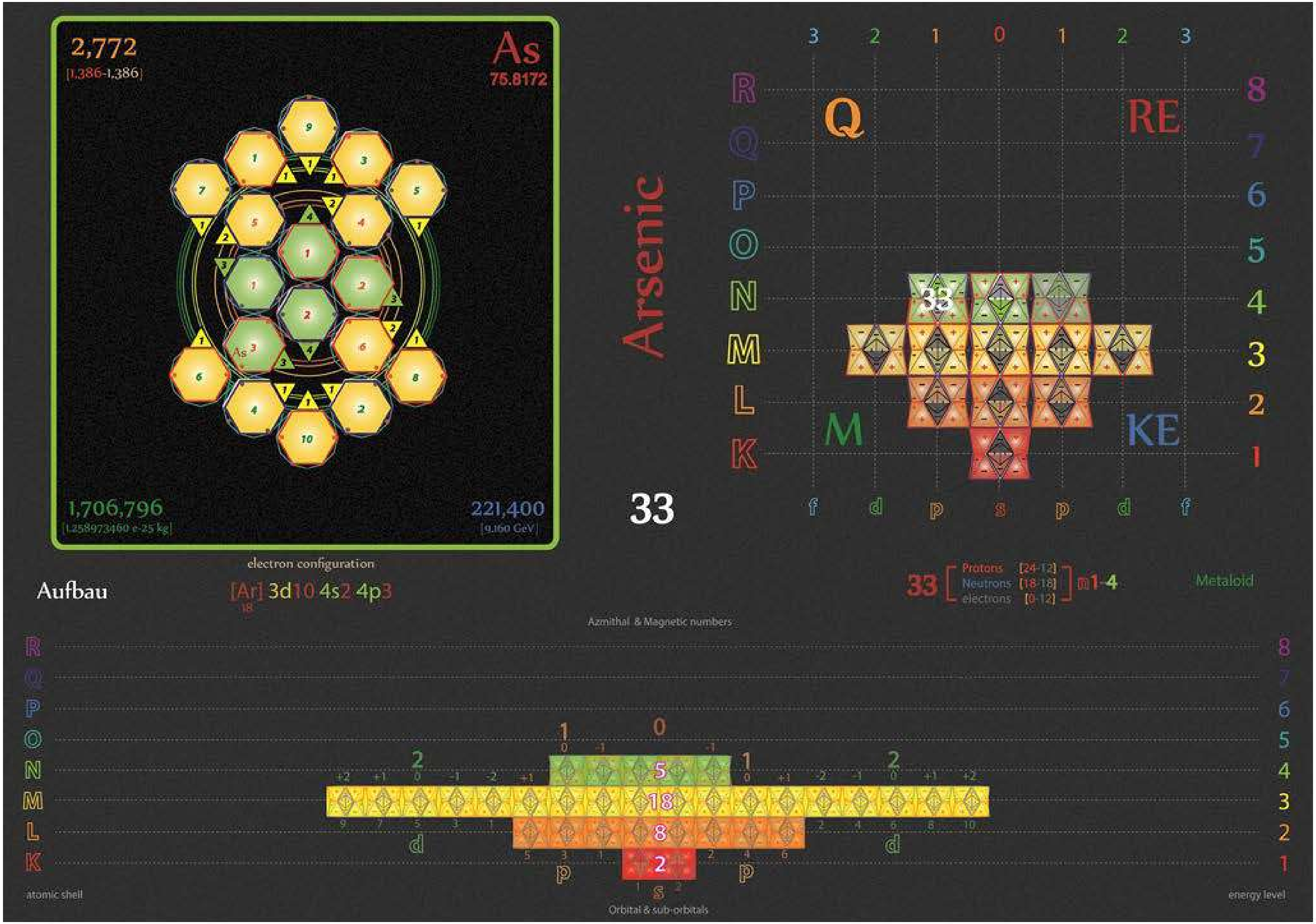
2,772
[1,386-1,386]
As
75.8172
1,706,796
[1.258973460 e-25 kg]
221,400
[9.160 GeV]
Arsenic
33
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
33
f d p s p d f
electron configuration
Aufbau
[Ar] 3d10 4s2 4p3
18
33 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
5
18
8
2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.33 - Arsenic atom

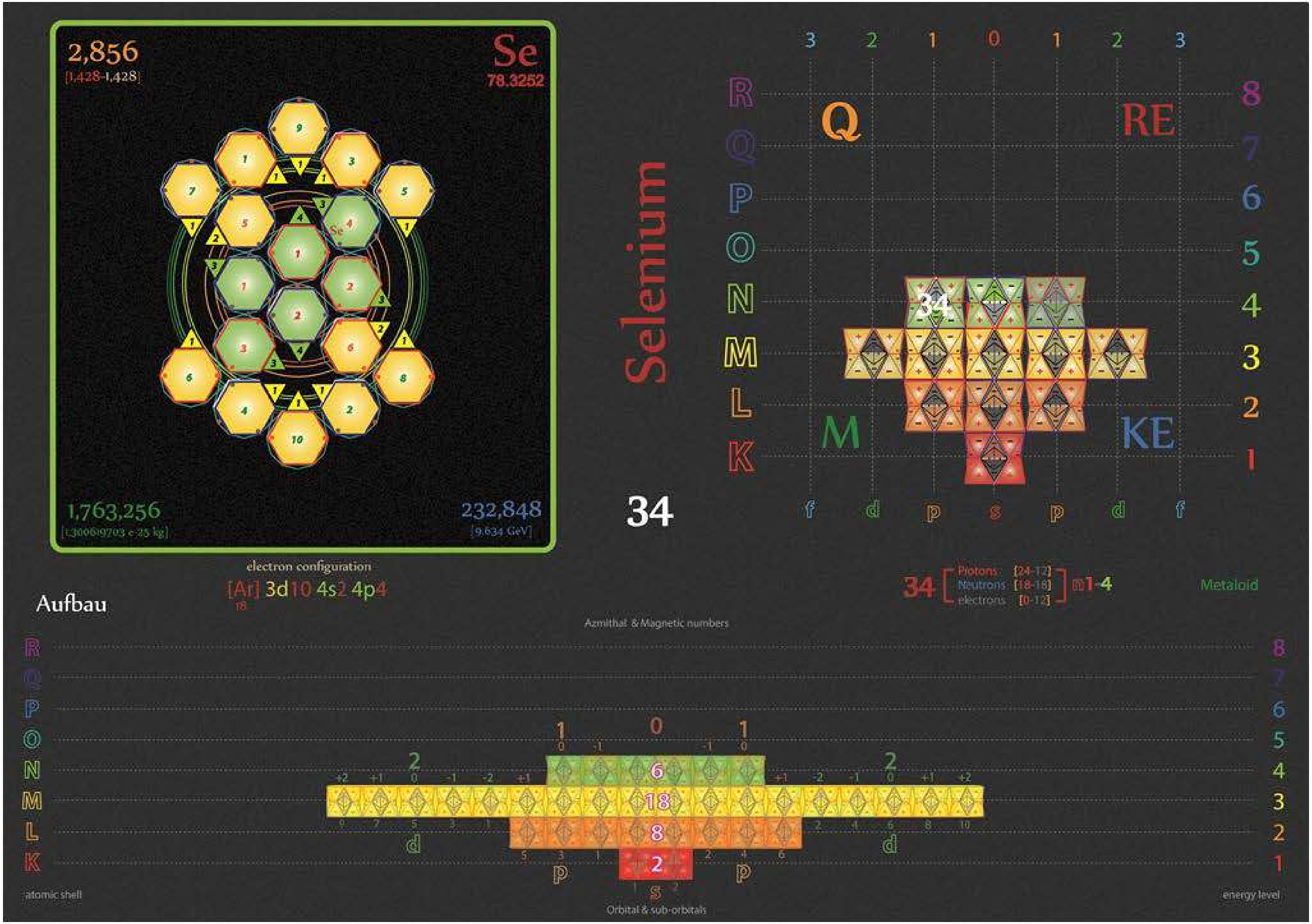
2,856
[1,428-1,428]
Se
78.3252
1,763,256
[1.300619703 e-25 kg]
232,848
[9.634 GeV]
Selenium
34
electron configuration
[Ar] 3d10 4s2 4p4
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
34 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Metaloid
Azmithal & Magnetic numbers
6
18
8
2
d
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.34 - Selenium atom

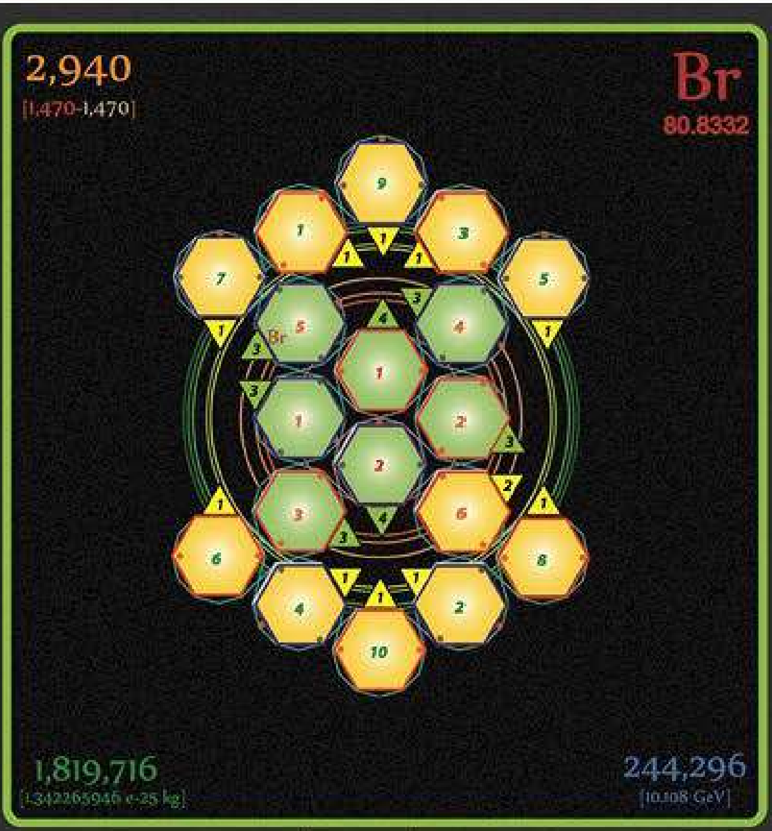
2,940
[1,470-1,470]
Br
80.8332
9
1
3
7
5
5
Br
4
1
1
2
3
6
6
8
4
2
10
1,819,716
[1.342265946 e-25 kg]
244,296
[10.108 GeV]

Bromine

35

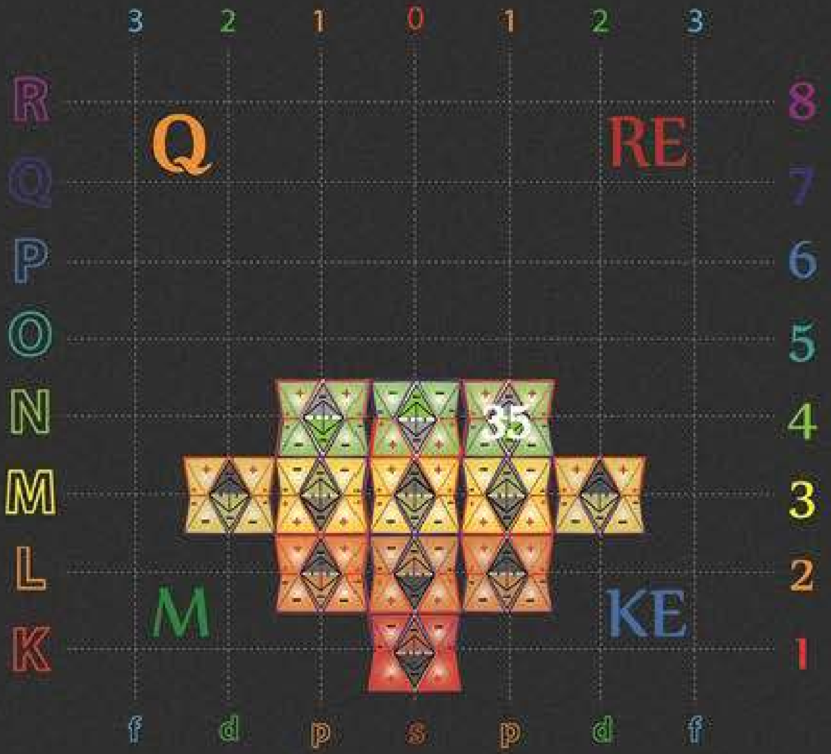
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
35
f
d
p
s
p
d
f

electron configuration

Aufbau

[Ar] 3d10 4s2 4p5
36

35 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4

Halogen

Azmithal & Magnetic numbers

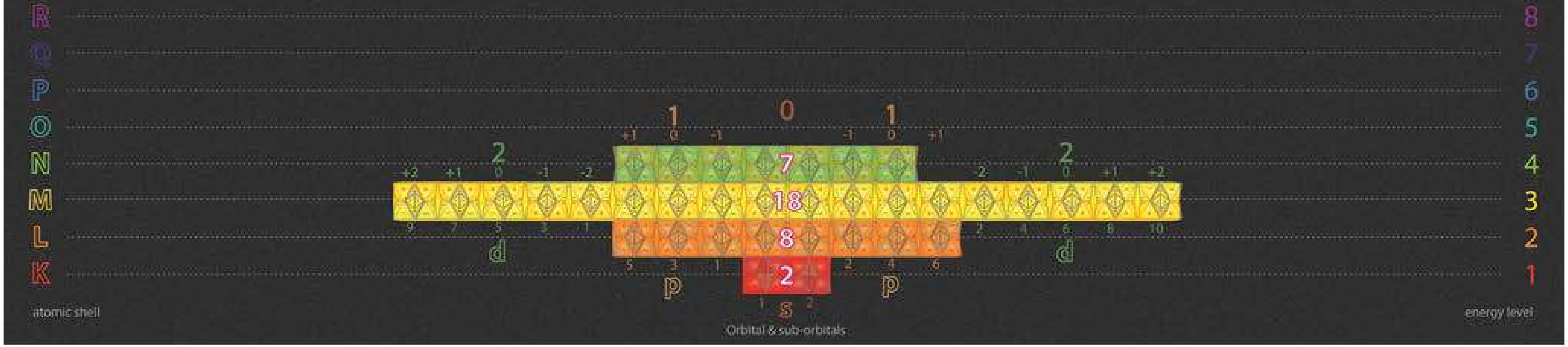
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
+1
0
-1
-1
0
+1
2
2
+2
+1
0
-1
-2
-2
-1
0
+1
+2
7
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.35 - Bromine atom

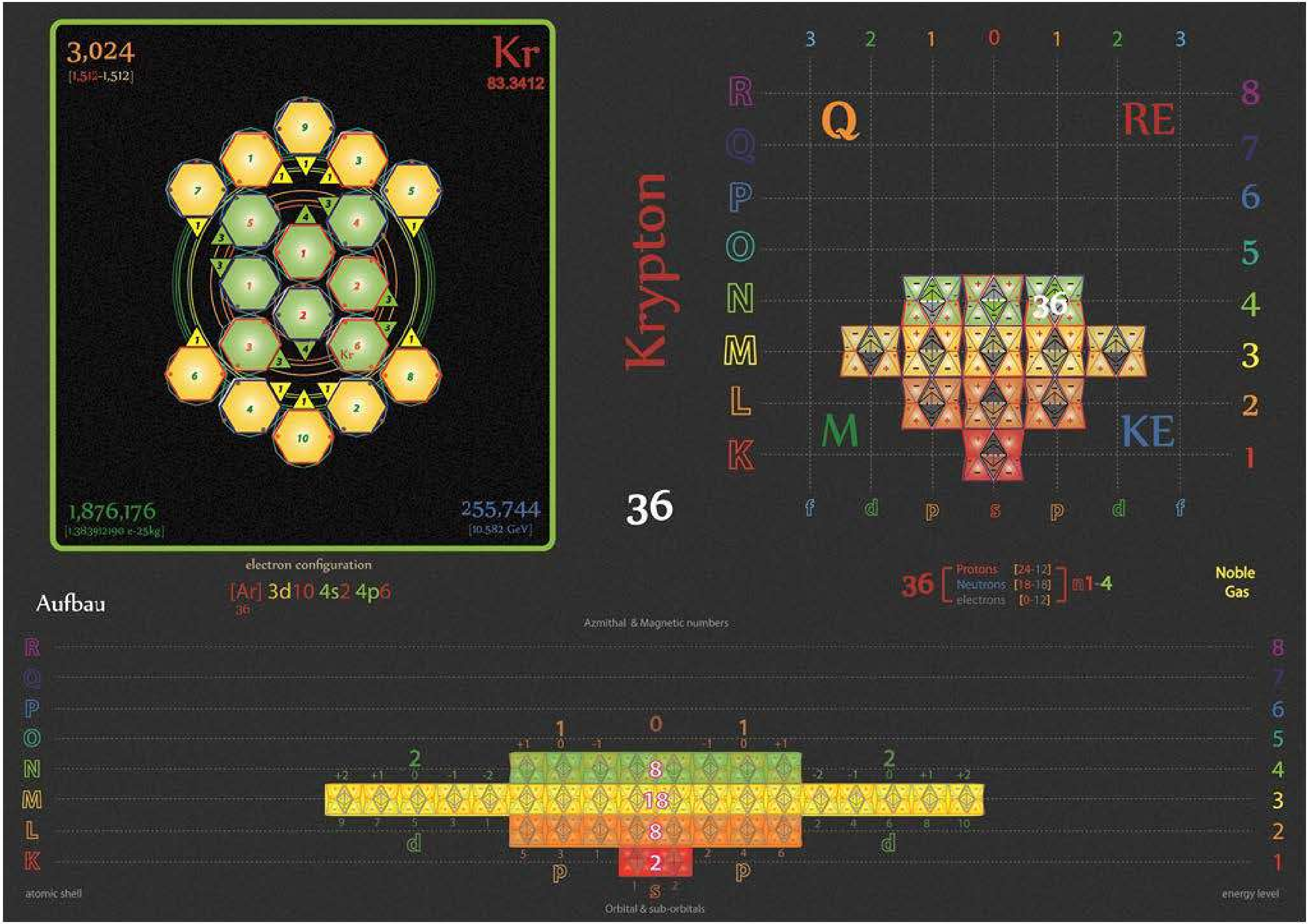
3,024
[1,512-1,512]
Kr
83.3412
1,876,176
[1.38391219 e-25kg]
255,744
[10.582 GeV]
Krypton
36
electron configuration
[Ar] 3d10 4s2 4p6
36
Aufbau
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
36
f d p s p d f
36 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Noble Gas
Azmithal & Magnetic numbers
2
1
0
1
2
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
8
18
8
2
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
p
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.36 - Krypton atom

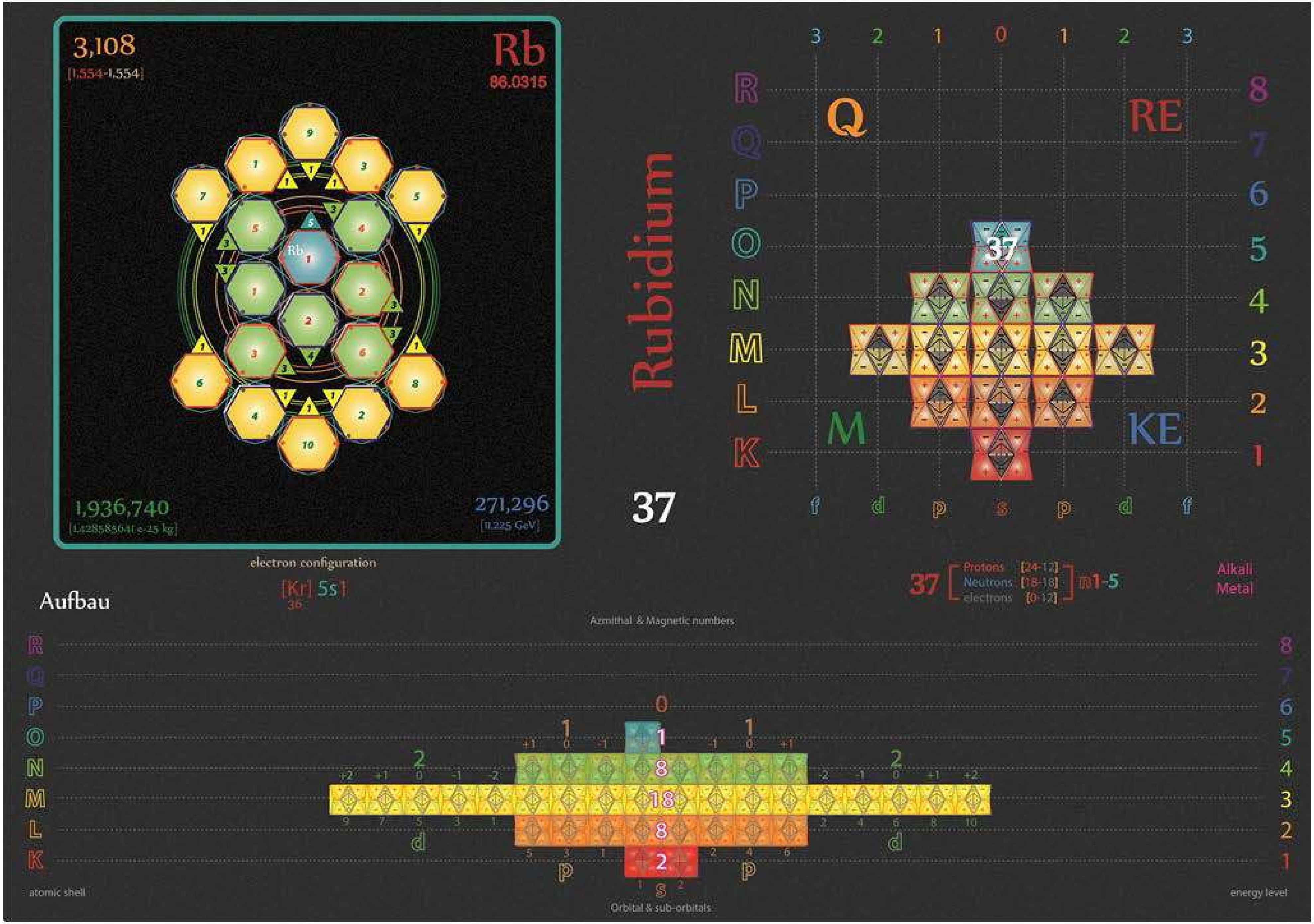
3,108
[1,554-1,554]
Rb
86.0315
1,936,740
[1.428585641 e-25 kg]
271,296
[11.225 GeV]
Rubidium
37
electron configuration
[Kr] 5s1
36
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
37 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Alkali Metal
Azmithal & Magnetic numbers
1
8
18
8
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.37 - Rubidium atom

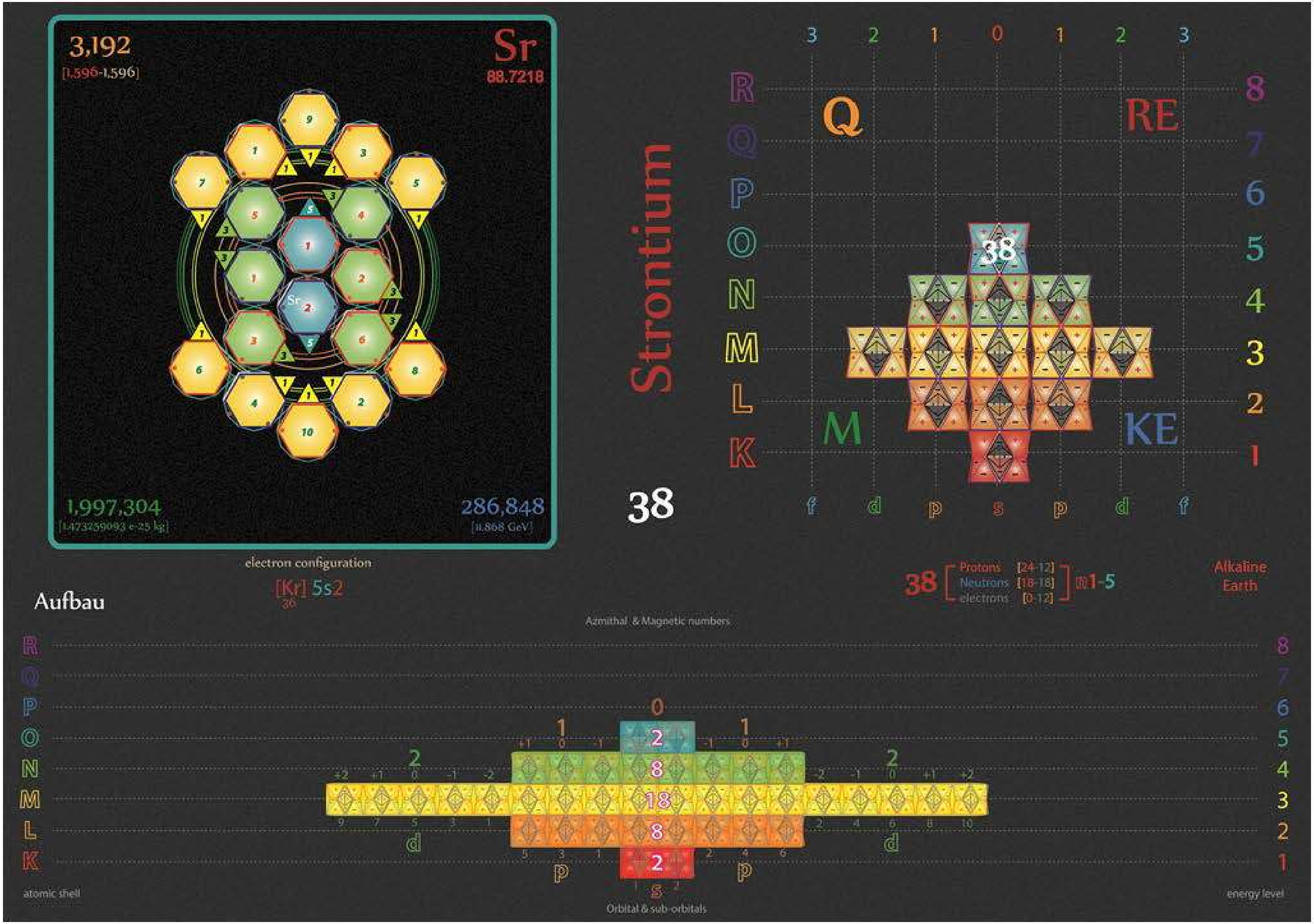
3,192
[1,596-1,596]
Sr
88.7218
1,997,304
[1.473259093 e-25 kg]
286,848
[11.868 GeV]
Strontium
38
electron configuration
[Kr] 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
38 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Alkaline Earth
Azmithal & Magnetic numbers
0
1 1
2 2
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -2 -1 0 +1 +2
2 8 18 8 2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.38 - Strontium atom

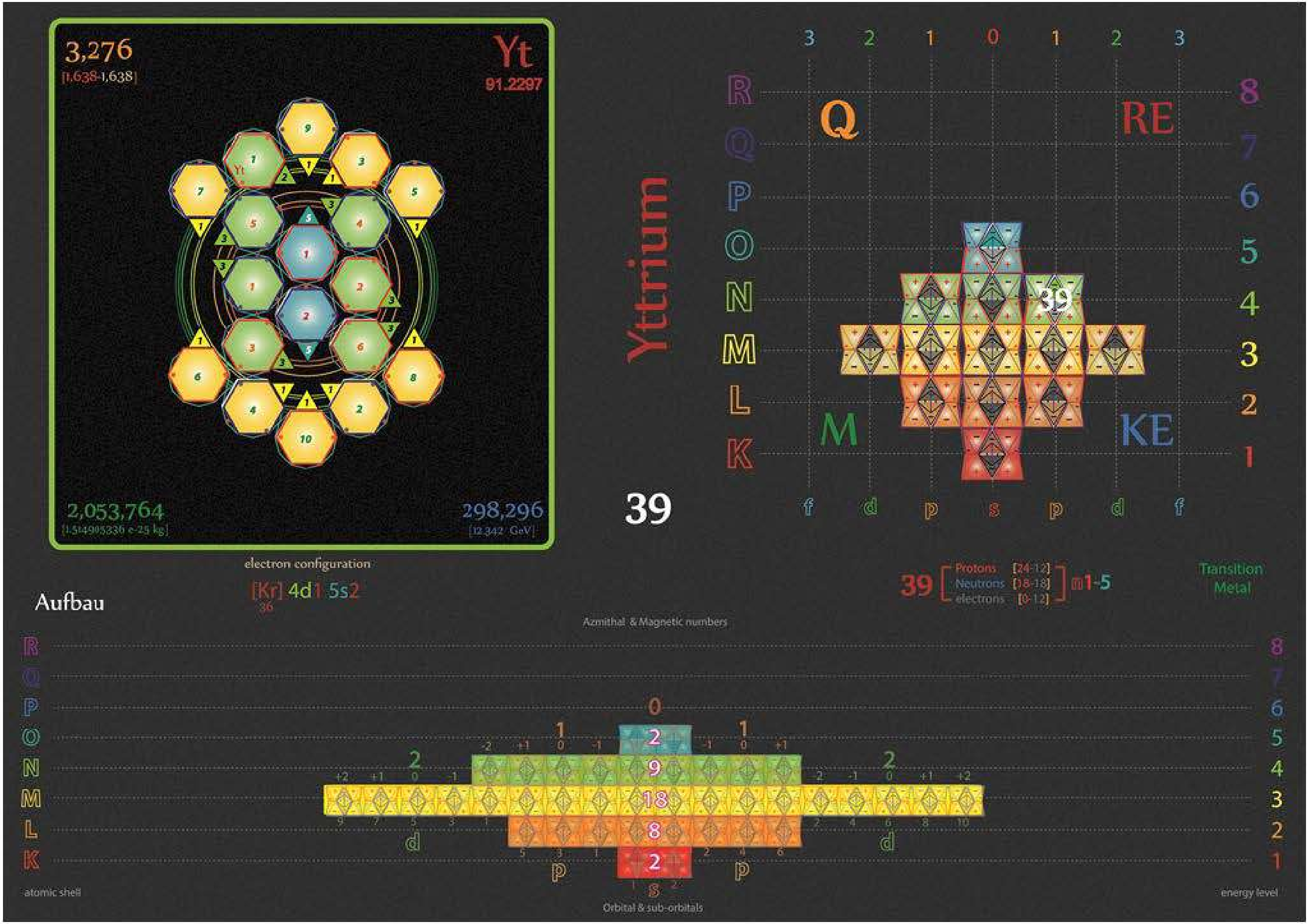
3,276
[1,638-1,638]
Yt
91.2297
2,053,764
[1.514905336 e-25 kg]
298,296
[12.342 GeV]
Yttrium
39
electron configuration
[Kr] 4d1 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
39
f d p s p d f
39 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
2
9
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.39 - Yttrium atom

Copyright ABRAHAM [2008] - All rights reserved

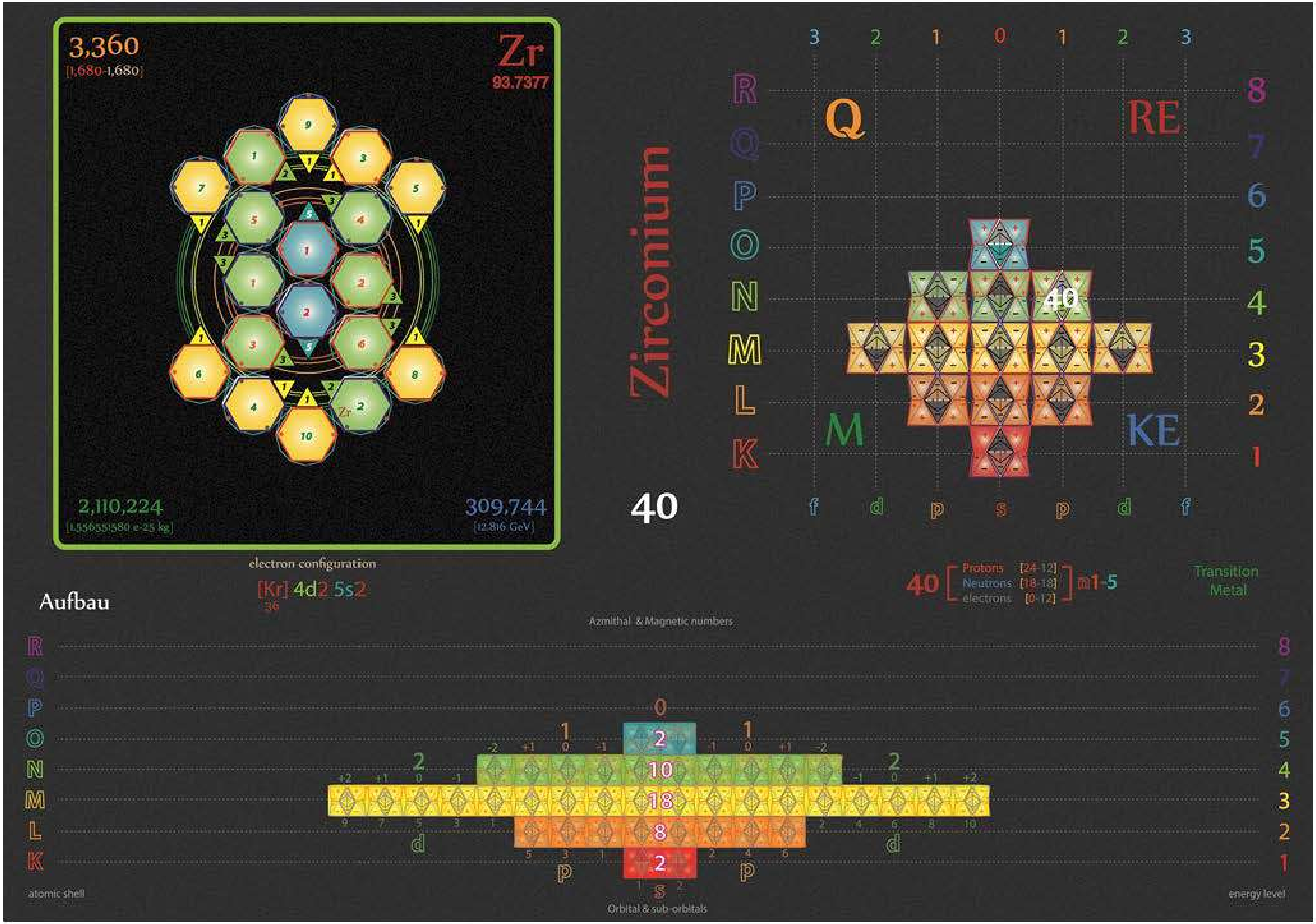
3,360
[1,680-1,680]
Zr
93.7377
2,110,224
[1,556551580 e-25 kg]
309,744
[12,816 GeV]
Zirconium
40
electron configuration
[Kr] 4d2 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
40
f d p s p d f
40
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-5
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
2
10
18
8
2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.40 - Zirconium atom

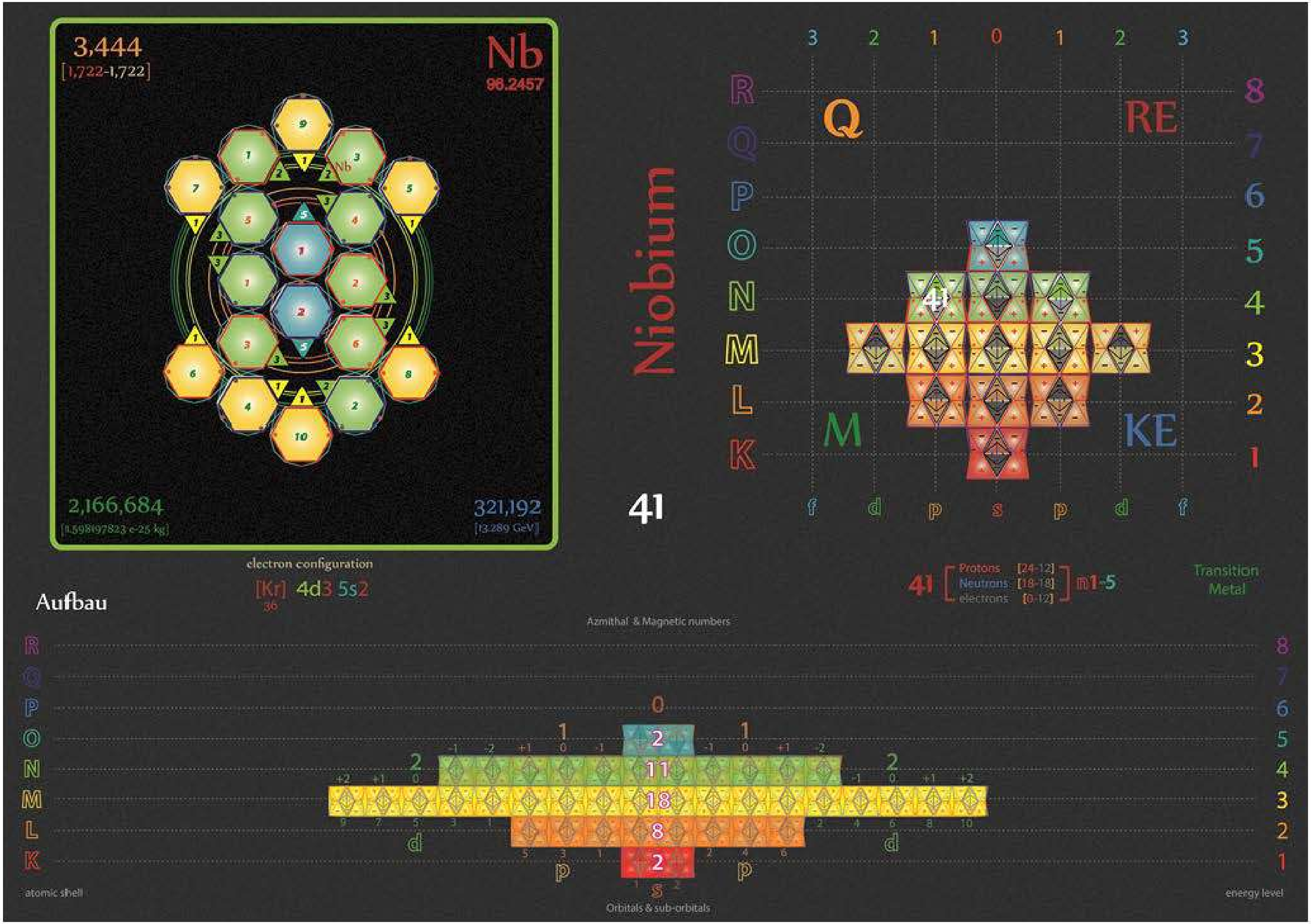
3,444
[1,722-1,722]
Nb
98.2457
2,166,684
[1.598197823 e-25 kg]
321,192
[13.289 GeV]
Niobium
41
electron configuration
[Kr] 4d3 5s2
36
Aufbau
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
41 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
2
11
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.41 - Niobium atom

Copyright ABRAHAM [2008] - All rights reserved

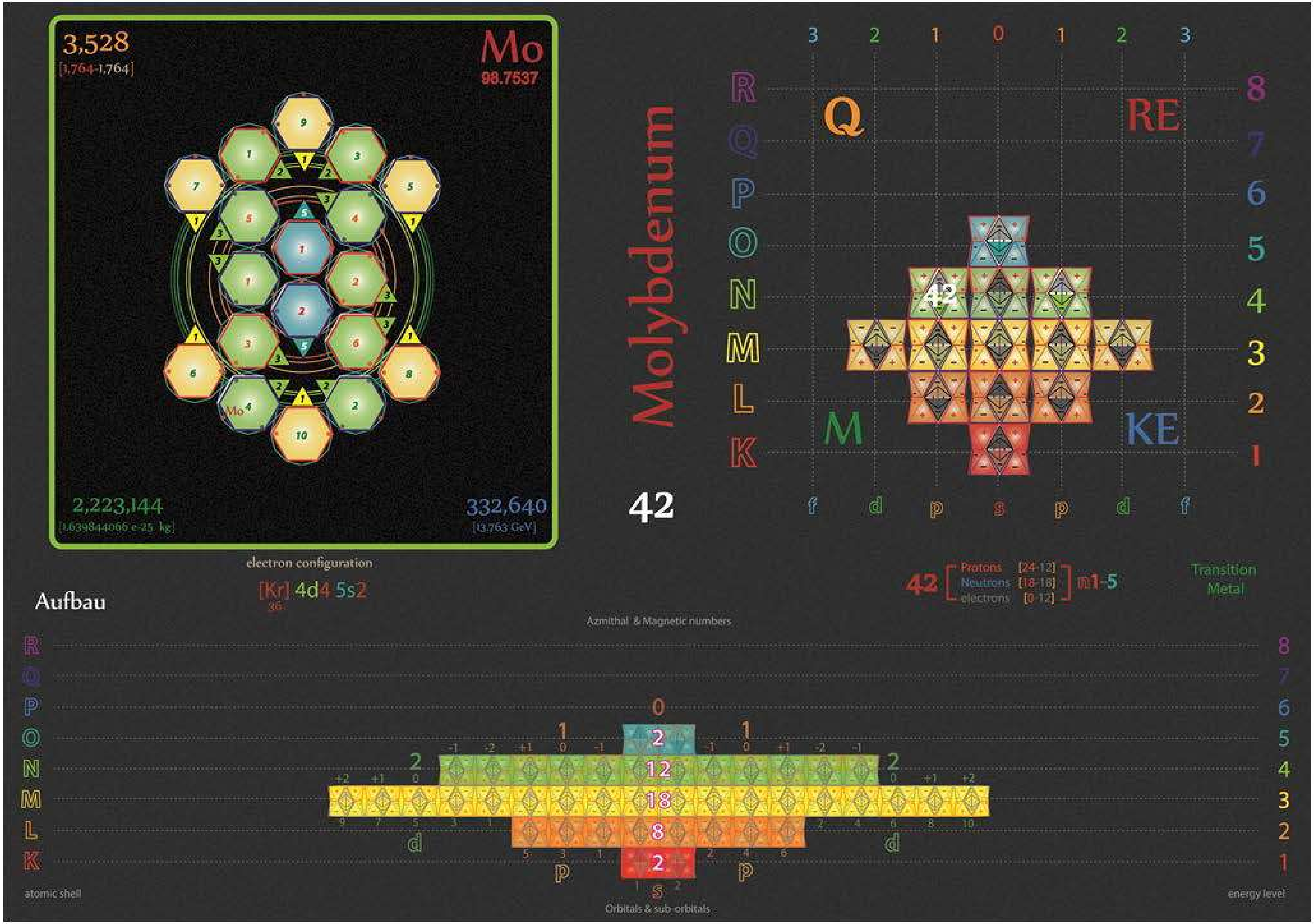
3,528
[1,764-1,764]
Mo
98.7537
2,223,144
[1.639844066 e-25 kg]
332,640
[13.763 GeV]
Molybdenum
42
electron configuration
[Kr] 4d4 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
42 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
2
12
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.42 - Molybdenum atom

Copyright ABRAHAM [2008] - All rights reserved

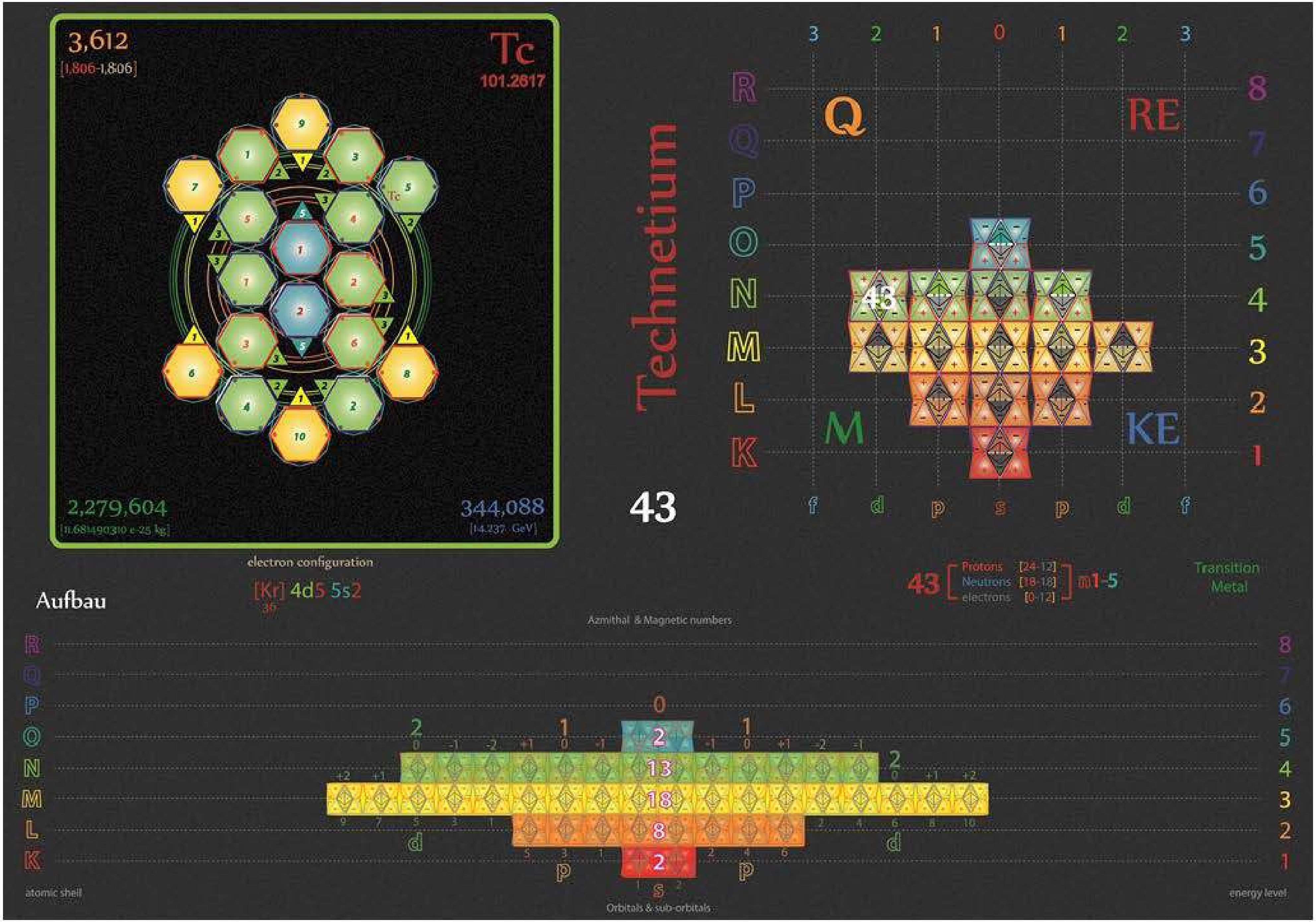
3,612
[1,806-1,806]
Tc
101.2617
2,279,604
[11.6814903101 e-25 kg]
344,088
[14.237 GeV]
Technetium
43
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
Q
RE
M
KE
f
d
p
s
p
d
f
electron configuration
[Kr] 4d5 5s2
36
Aufbau
43 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
2
13
18
8
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.43 - Technetium atom

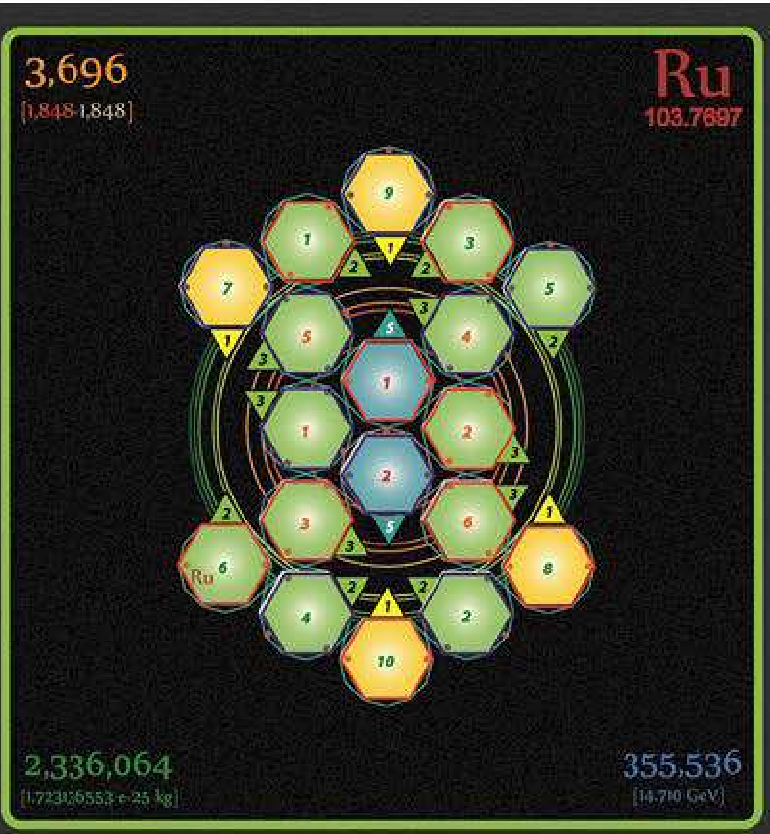
3,696
[1,848-1,848]
Ru
103.7697
9
1
3
7
5
2
2
1
3
5
4
3
1
2
3
1
2
3
2
2
3
5
3
1
2
6
3
6
8
4
2
1
10
2,336,064
[1.7230,6553 e-25 kg]
355,536
[14.710 GeV]

Ruthenium

44

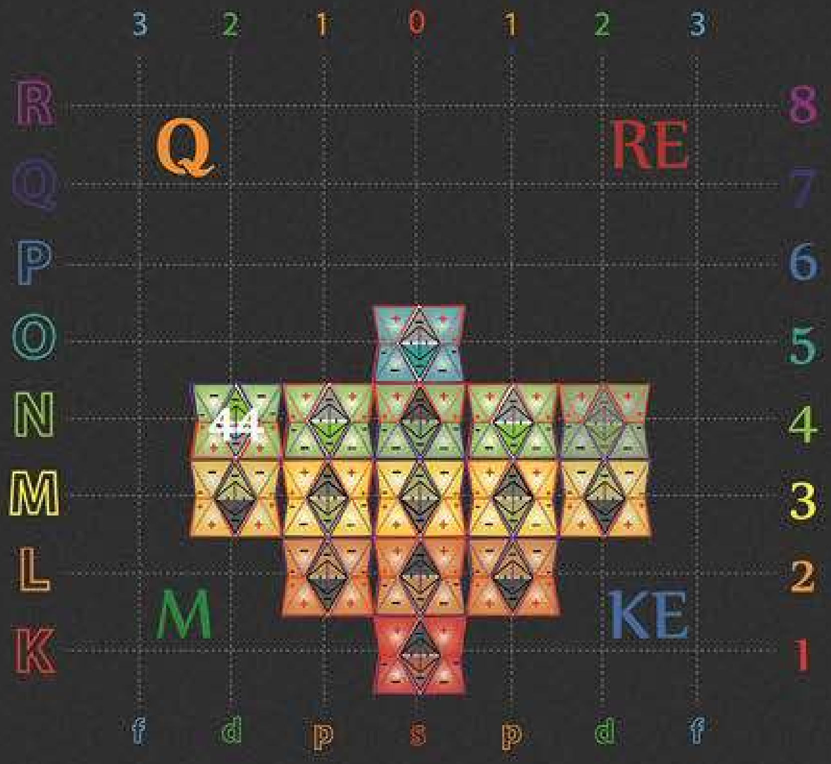
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
44
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

[Kr] 4d6 5s2
36

44 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5

Transition Metal

Aufbau

Azmithal & Magnetic numbers

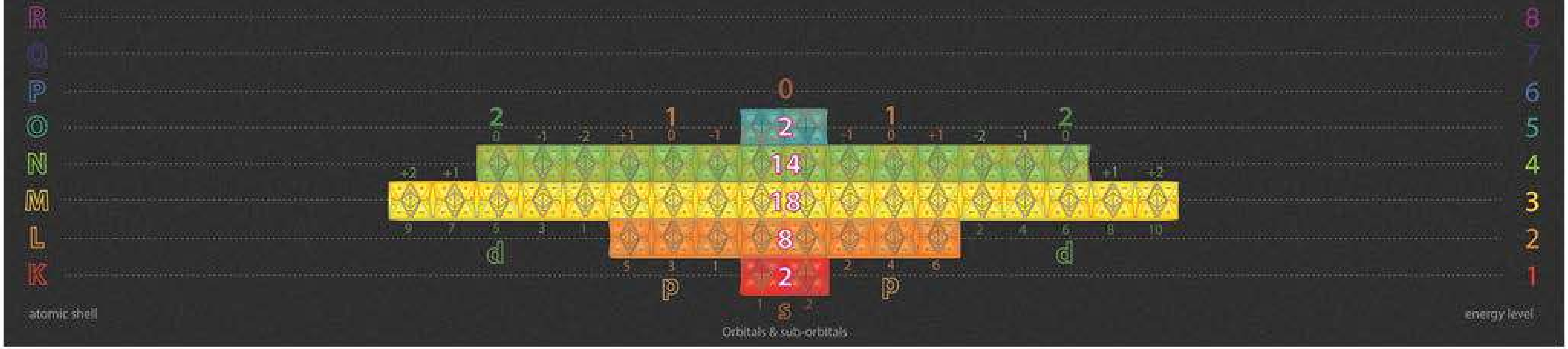
R
Q
P
O
N
M
L
K
atomic shell
0
2
1
1
2
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
+2
+1
+1
+2
2
14
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
8
7
6
5
4
3
2
1
energy level

Orbitals & sub-orbitals

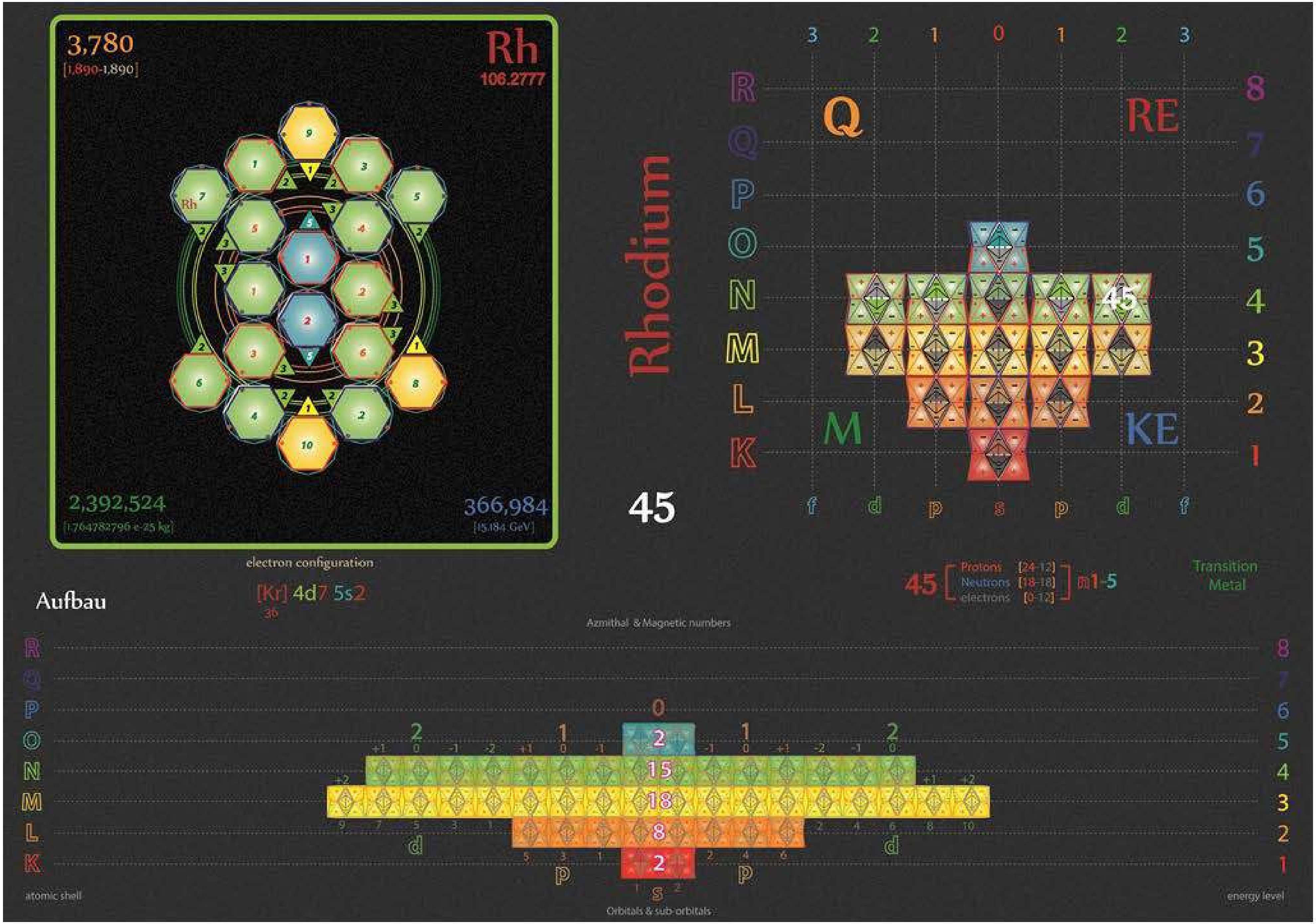
3,780
[1,890-1,890]
Rh
106.2777
2,392,524
[1.764782796 e-25 kg]
366,984
[15.184 GeV]
Rhodium
45
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
45
f d p s p d f
electron configuration
[Kr] 4d7 5s2
36
Aufbau
Azmithal & Magnetic numbers
45 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5
Transition Metal
0
2 1 1 2
2
15
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.45 - Rhodium atom

Copyright ABRAHAM [2008] - All rights reserved

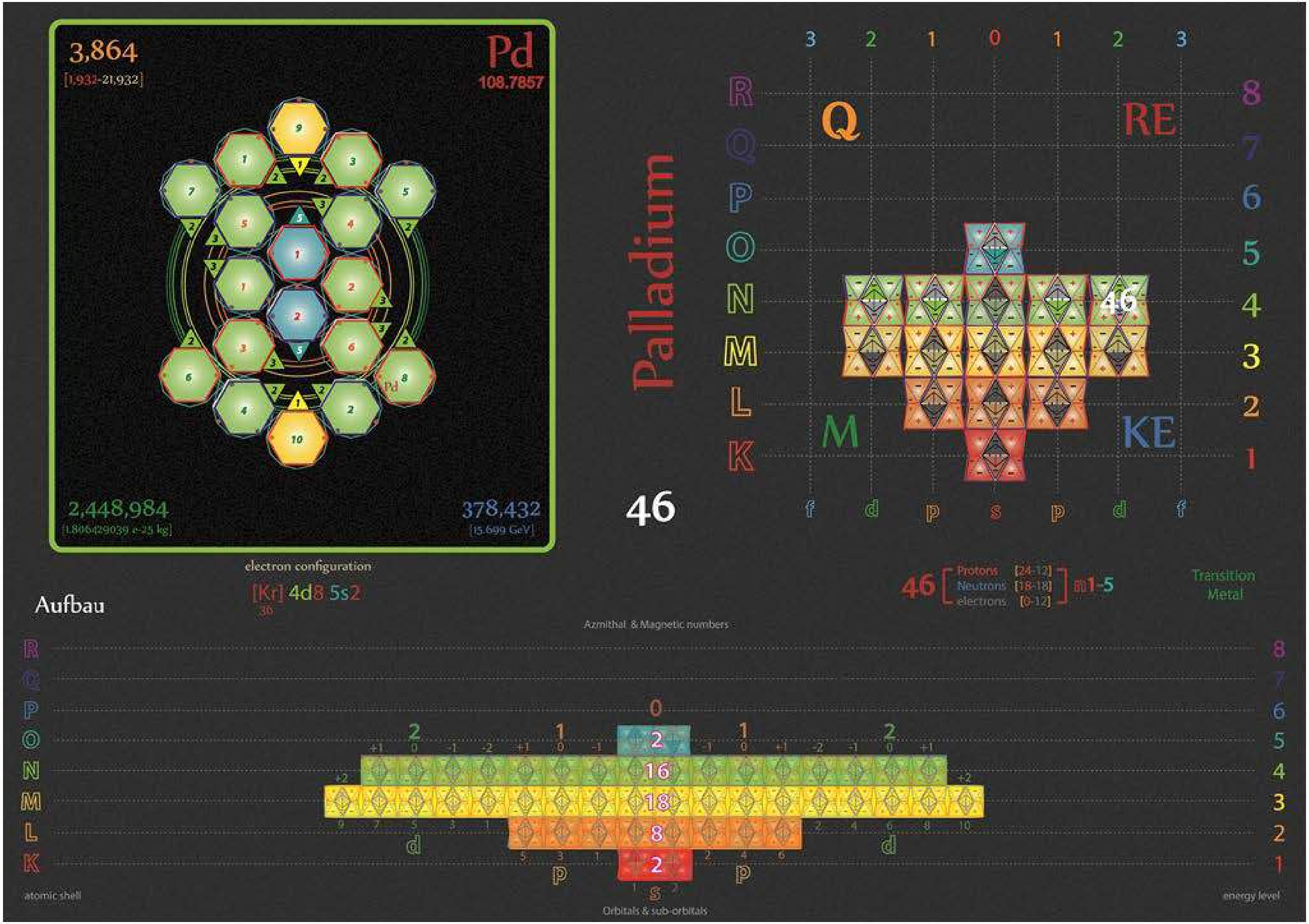
3,864
[1,932-21,932]
Pd
108.7857
2,448,984
[1.806429039 e-25 kg]
378,432
[15.699 GeV]
Palladium
46
electron configuration
[Kr] 4d8 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
46
f d p s p d f
46 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
16
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.46 - Palladium atom

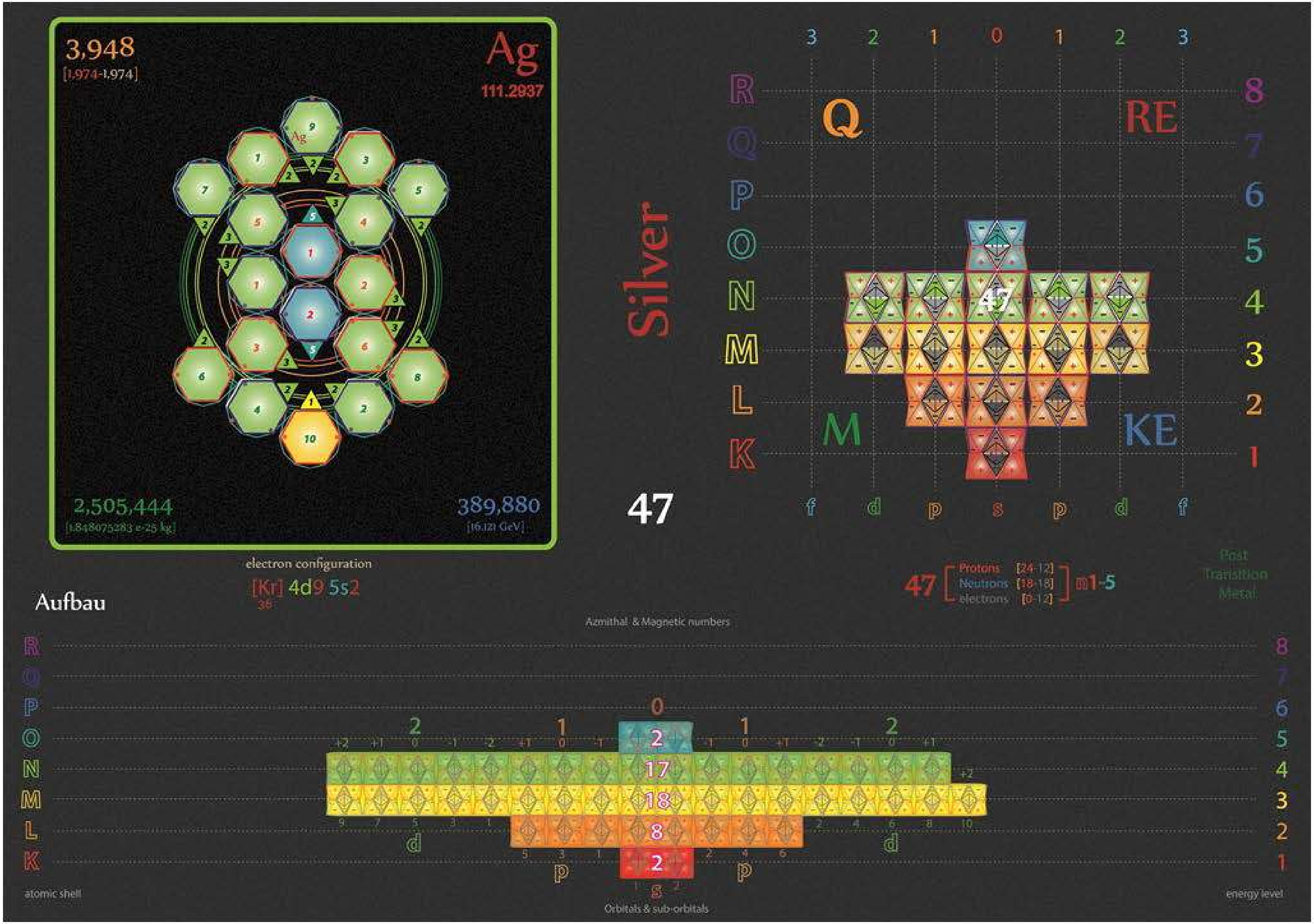
3,948
[1,974-1,974]
Ag
111.2937
2,505,444
[1.848075283 e-25 kg]
389,880
[16.121 GeV]
Silver
47
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d9 5s2
36
Aufbau
47 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Post Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
17
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.47 - Silver atom

Copyright ABRAHAM [2008] - All rights reserved

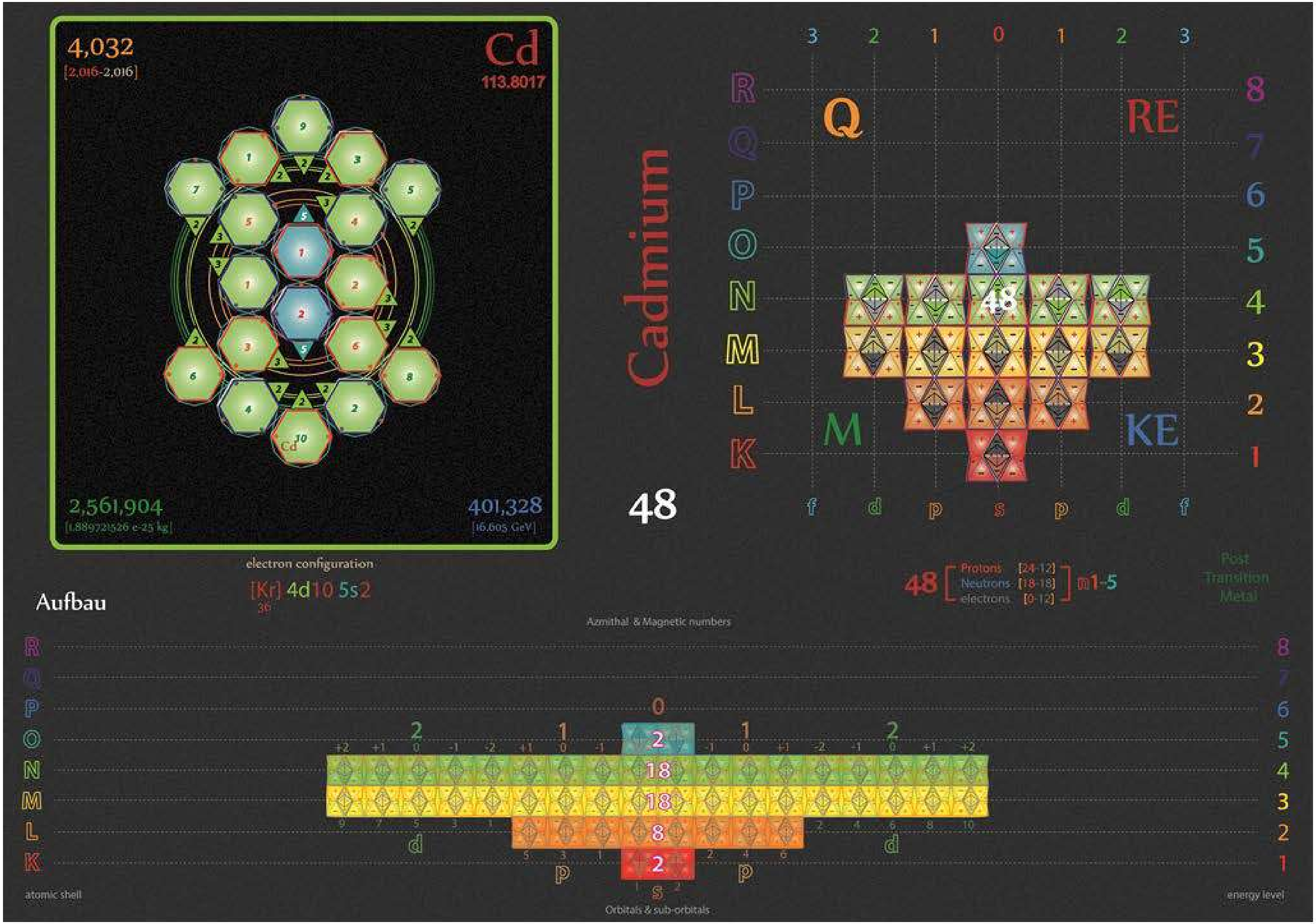
4,032
[2,016-2,016]
Cd
113.8017
Cd
10
2,561,904
[1.88972526 e-25 kg]
401,328
[16.605 GeV]
Cadmium
48
electron configuration
[Kr] 4d10 5s2
36
Aufbau
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
48
f
d
p
s
p
d
f
48 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5
Post Transition Metal
Azmithal & Magnetic numbers
0
2
18
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.48 - Cadmium atom

Copyright ABRAHAM [2008] - All rights reserved

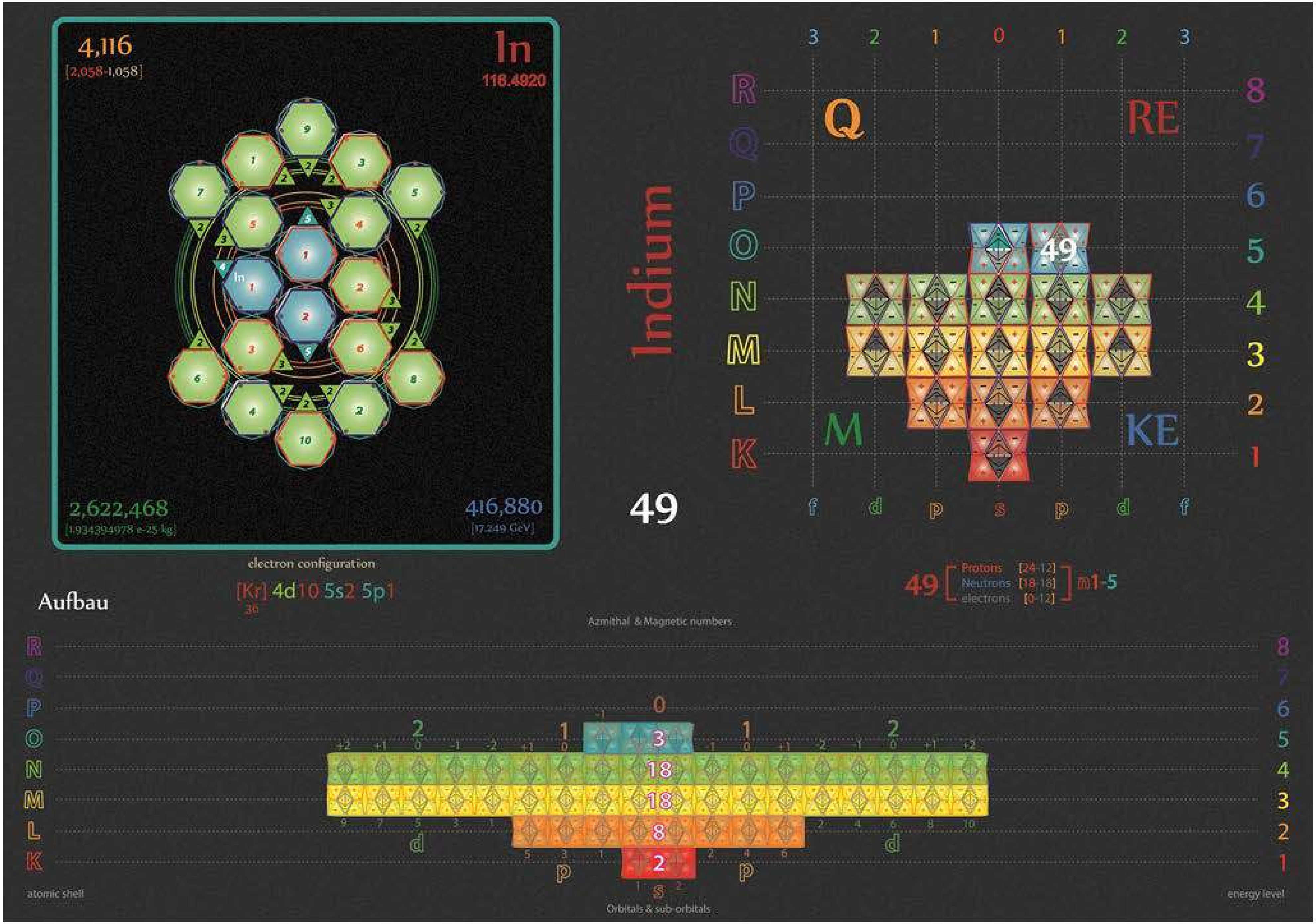
4,116
[2,058-1,058]
In
116.4920
In
2,622,468
[1.934394978 e-25 kg]
416,880
[17.249 GeV]
Indium
49
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
49
f d p s p d f
electron configuration
[Kr] 4d10 5s2 5p1
36
49 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Aufbau
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
3
18
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 2
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.49 - Indium atom

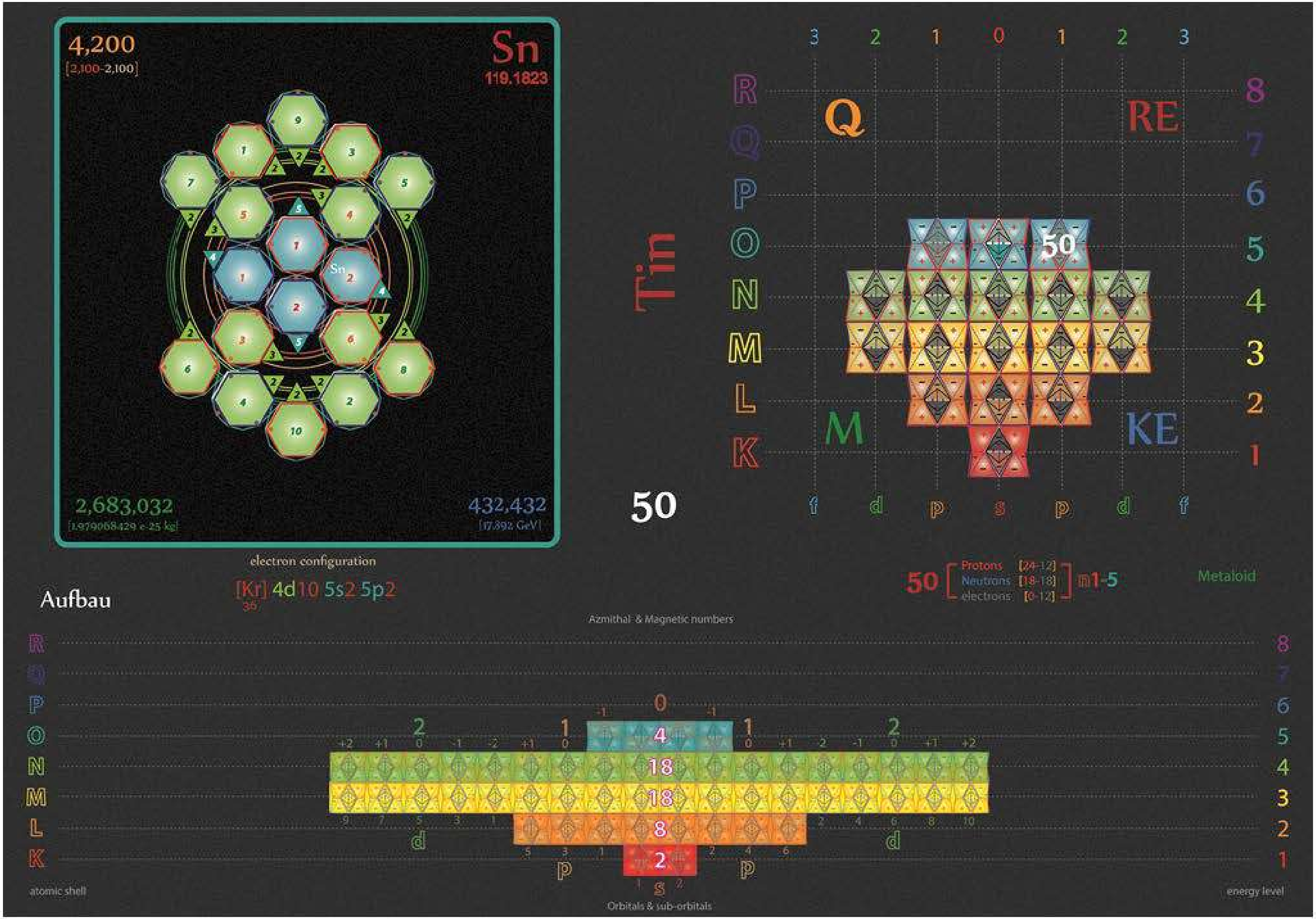
4,200
[2,100-2,100]
Sn
119.1823
2,683,032
[1.979068429 e-25 kg]
432,432
[17.392 GeV]
Tin
50
electron configuration
[Kr] 4d10 5s2 5p2
36
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
50
50 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5
Metaloid
Azmithal & Magnetic numbers
4
18
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.50 - Tin atom

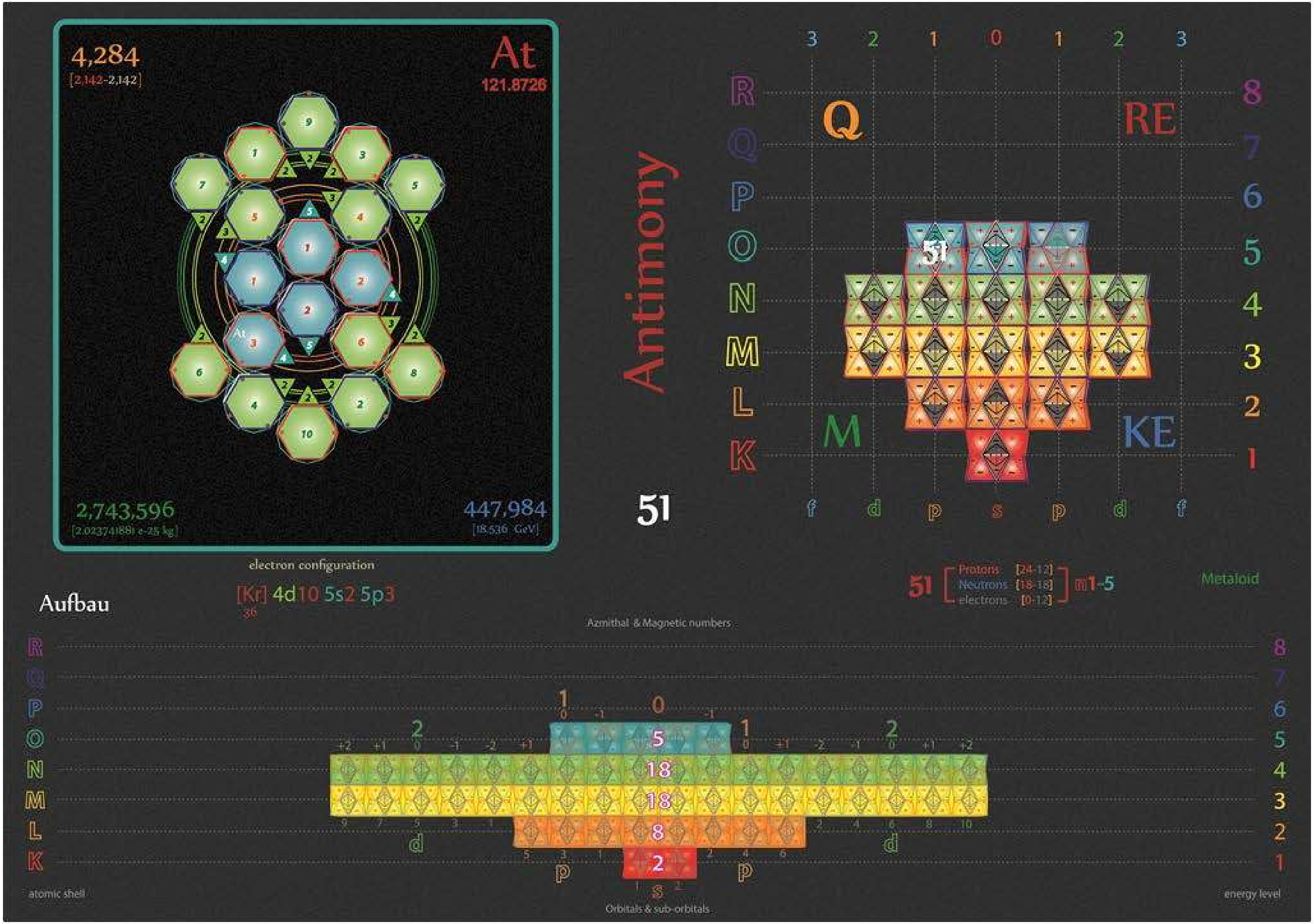
4,284
[2,142-2,142]
At
121.8726
2,743,596
[2.02374188i e-25 kg]
447,984
[18,536 GeV]
Antimony
51
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
51
f d p s p d f
electron configuration
Aufbau
[Kr] 4d10 5s2 5p3
36
51 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Azmithal & Magnetic numbers
5
18
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.51 - Antimony atom

Copyright ABRAHAM [2008] - All rights reserved

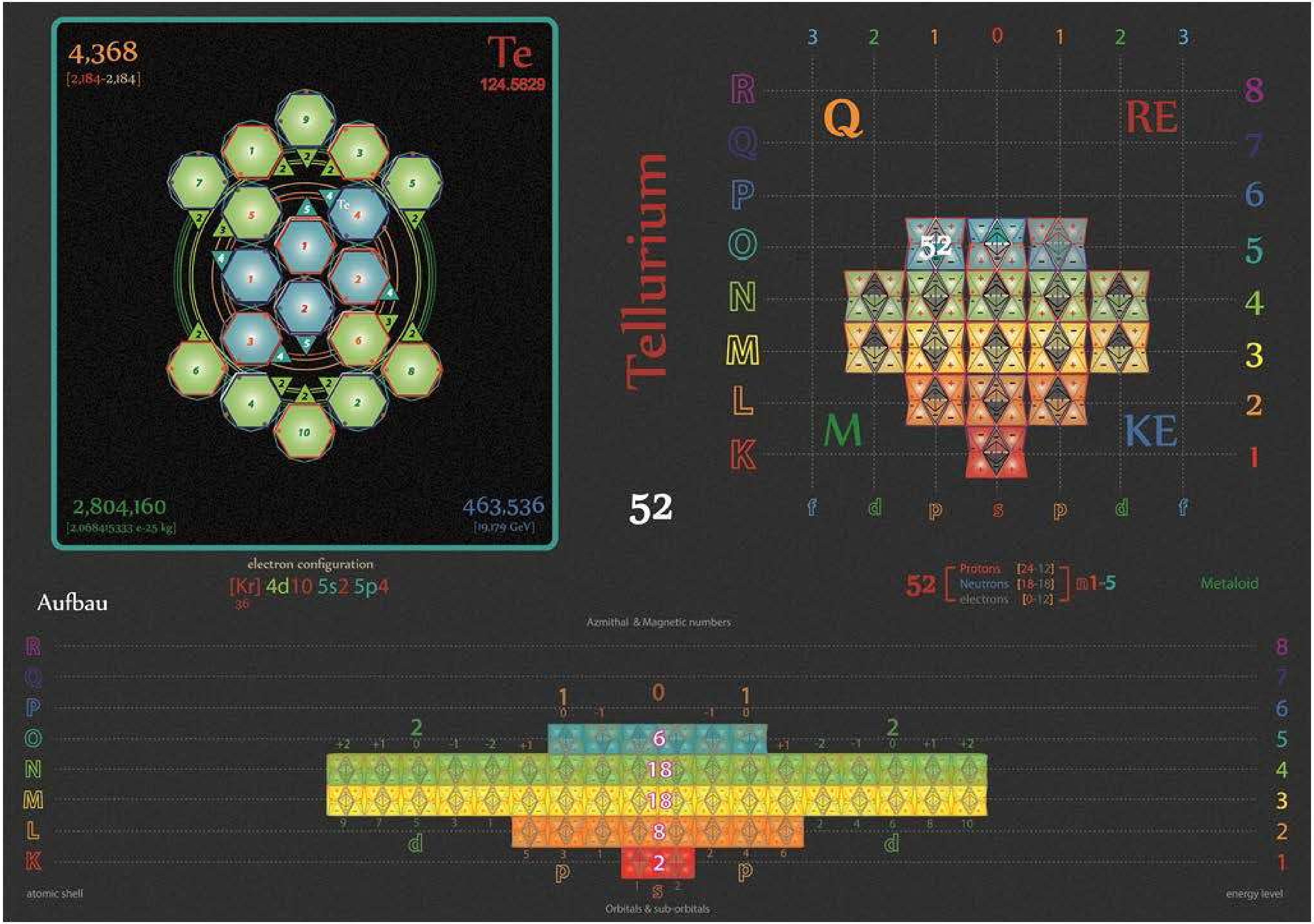
4,368
[2,184-2,184]
Te
124.5829
2,804,160
[2.0684/5333 e-25 kg]
463,536
[19.179 GeV]
Tellurium
52
electron configuration
[Kr] 4d10 5s2 5p4
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
52 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Azmithal & Magnetic numbers
6
18
18
8
2
d
p
s
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.52 - Tellurium atom

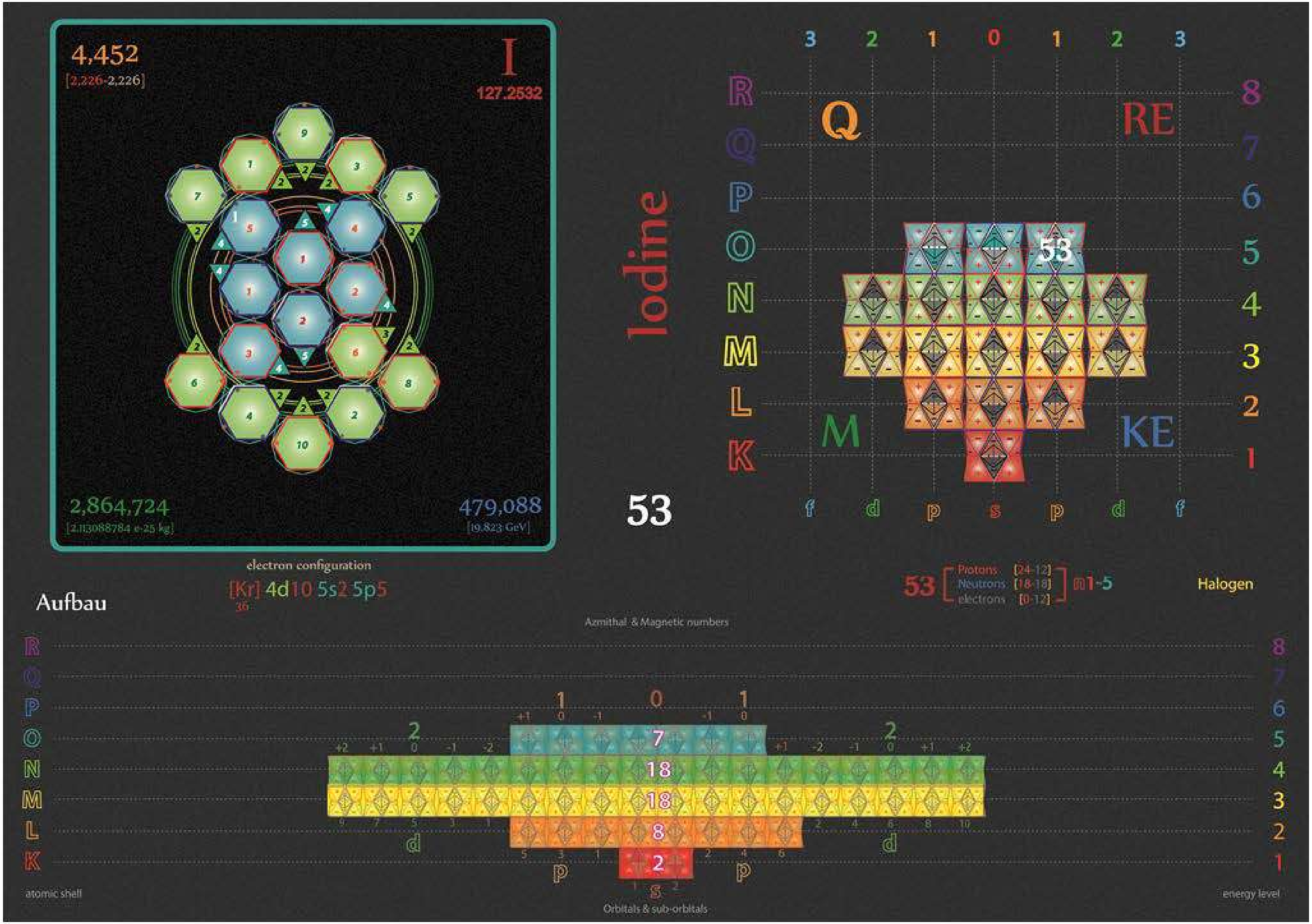
4,452
[2,226-2,226]
I
127.2532
2,864,724
[2.113088784 e-25 kg]
479,088
[19.823 GeV]
Iodine
53
electron configuration
[Kr] 4d10 5s2 5p5
36
Aufbau
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
53
8
7
6
5
4
3
2
1
f d p s p d f
53 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Halogen
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0
2 2
+2 +1 0 -1 -2 +1 -2 -1 0 +1 +2
7
18
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 2
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.53 - Iodine atom

Copyright ABRAHAM [2008] - All rights reserved

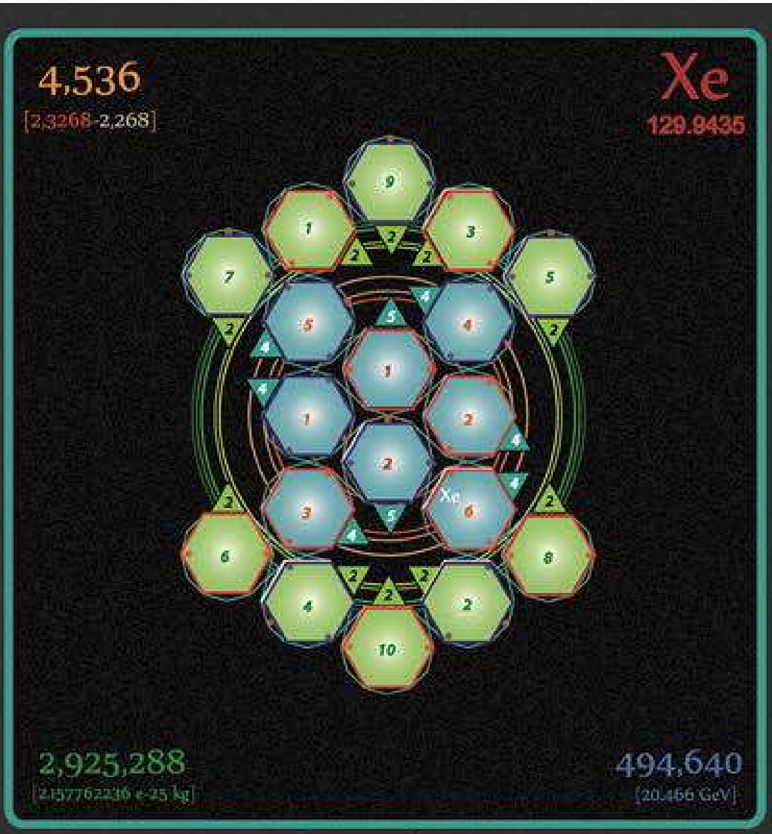
4,536

[2,3268-2,268]

Xe

129.9435

2,925,288

[2.157762236 e-25 kg]

494,640

[20.466 GeV]
9
1
3
7
5
5
4
1
1
2
2
3
6
6
8
4
2
10
Xe
2
4
5

Xenon

54

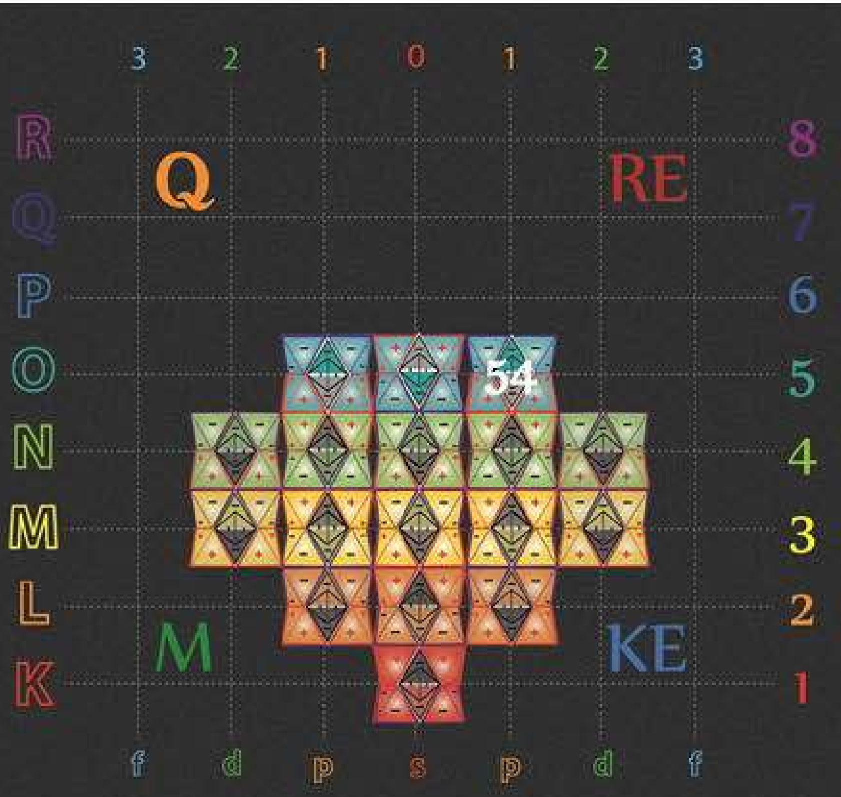
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
54
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

[Kr] 4d10 5s2 5p6

36

54 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5

Noble Gas

Aufbau

Azmithal & Magnetic numbers

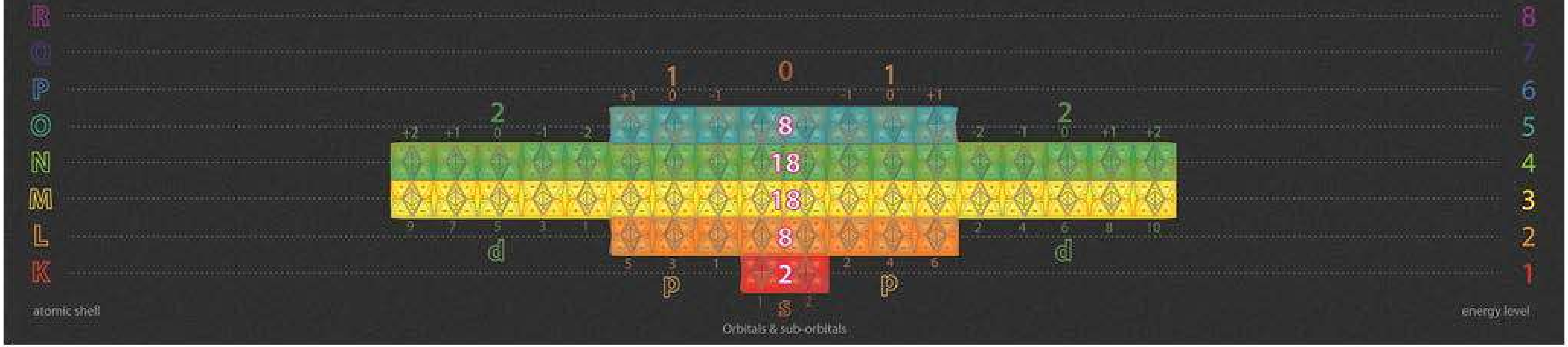
R
Q
P
O
N
M
L
K
1
0
1
+1
0
-1
-1
0
+1
2
+2
+1
0
-1
-2
2
-2
-1
0
+1
+2
8
18
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.54 - Xenon atom

Copyright ABRAHAM [2008] - All rights reserved

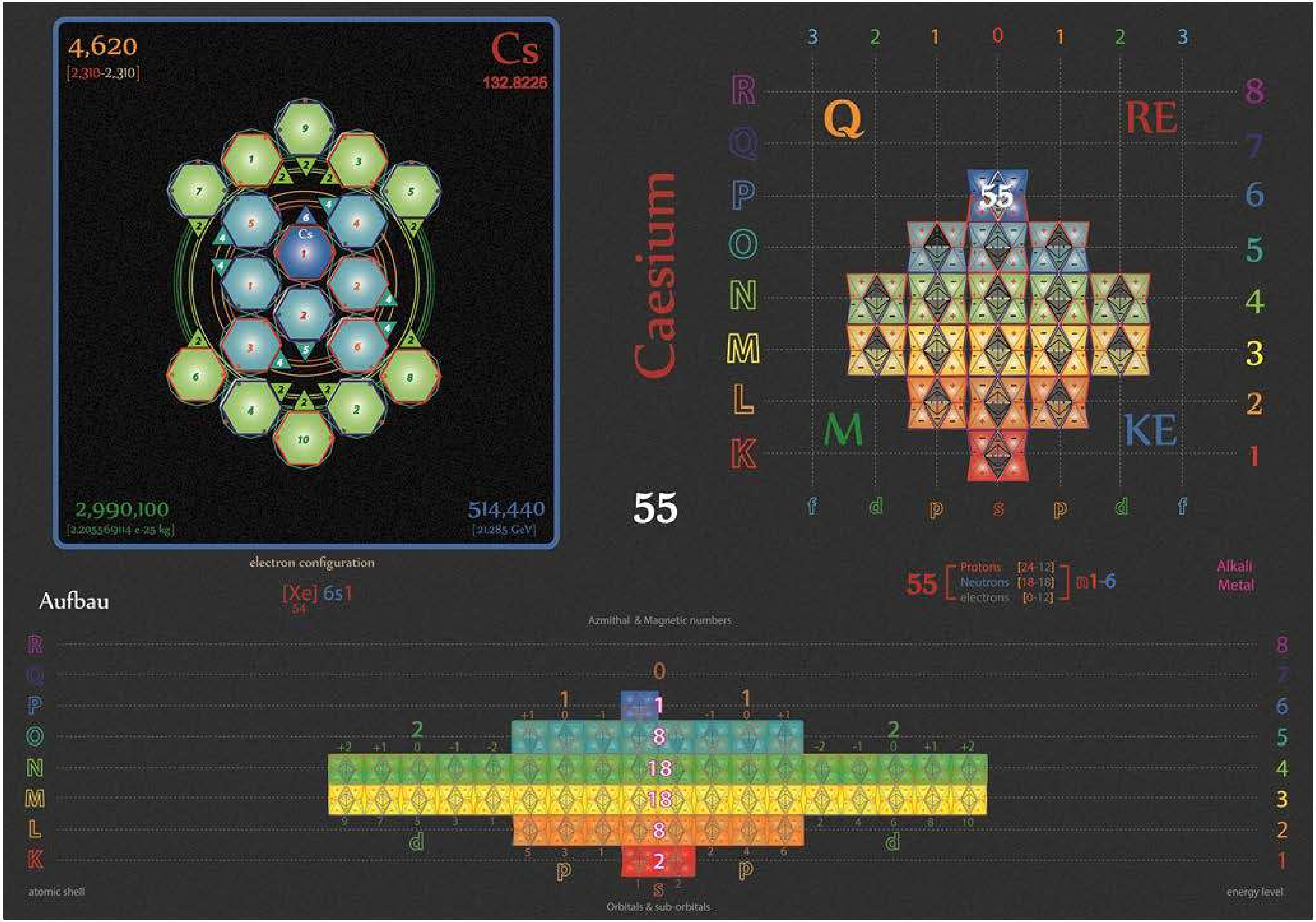
4,620
[2,310-2,310]
Cs
132.8225
2,990,100
[2.205569114 e-25 kg]
514,440
[21.285 GeV]
electron configuration
[Xe] 6s1
54
Caesium
55
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
55
55
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Alkali
Metal
Azmithal & Magnetic numbers
0
1
1
2
2
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
1
8
18
18
8
2
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.55 - Caesium atom

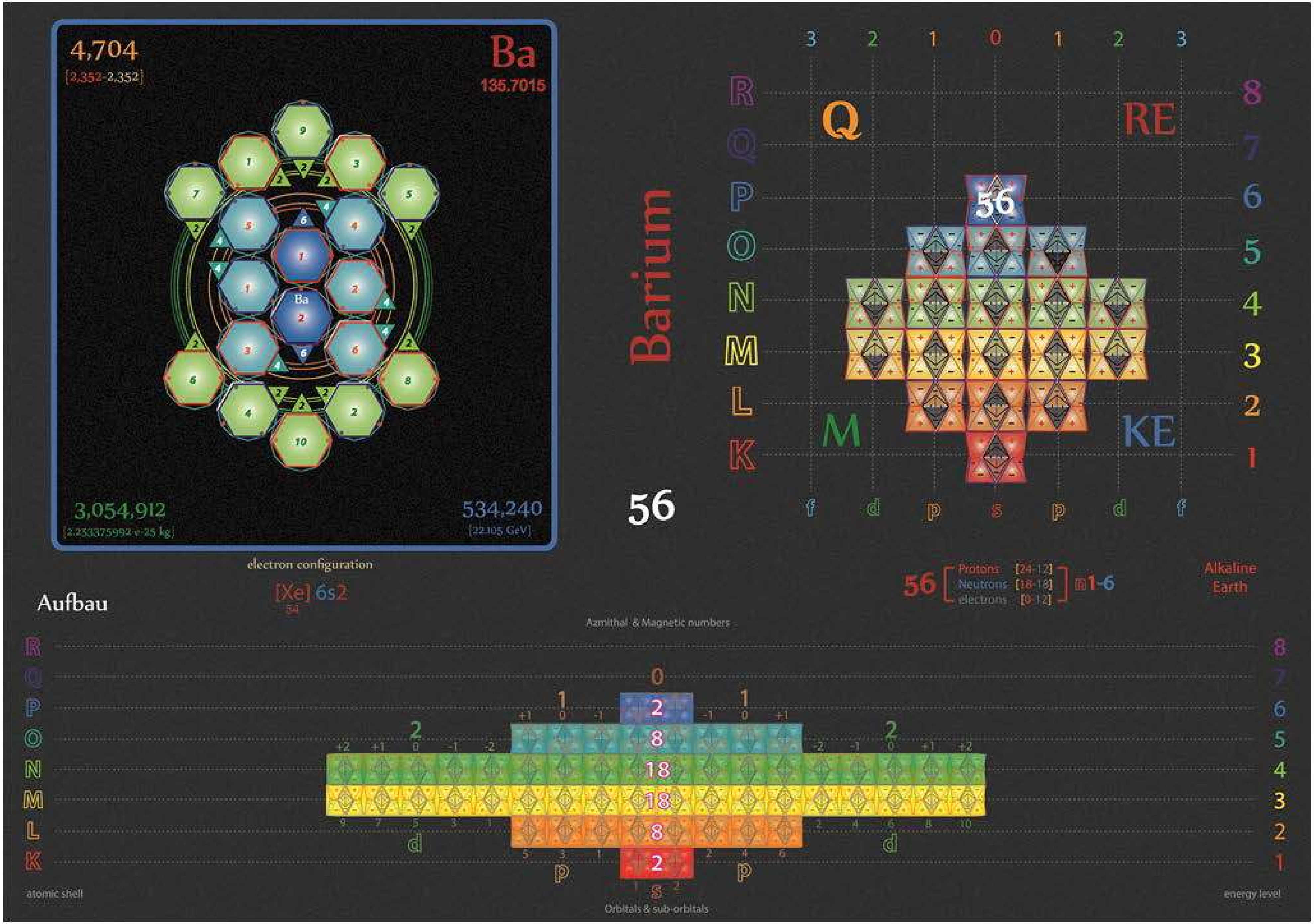
4,704
[2,352-2,352]
Ba
135.7015
3,054,912
[2.253375992 e-25 kg]
534,240
[22.105 GeV]
Barium
56
electron configuration
[Xe] 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
56
f d p s p d f
56
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-6
Alkaline
Earth
Azimthal & Magnetic numbers
0
1
1
2
2
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
2
8
18
18
8
2
9 7 5 3 1
2 4 6 8 10
d
d
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.56 - Barium atom

Copyright ABRAHAM [2008] - All rights reserved

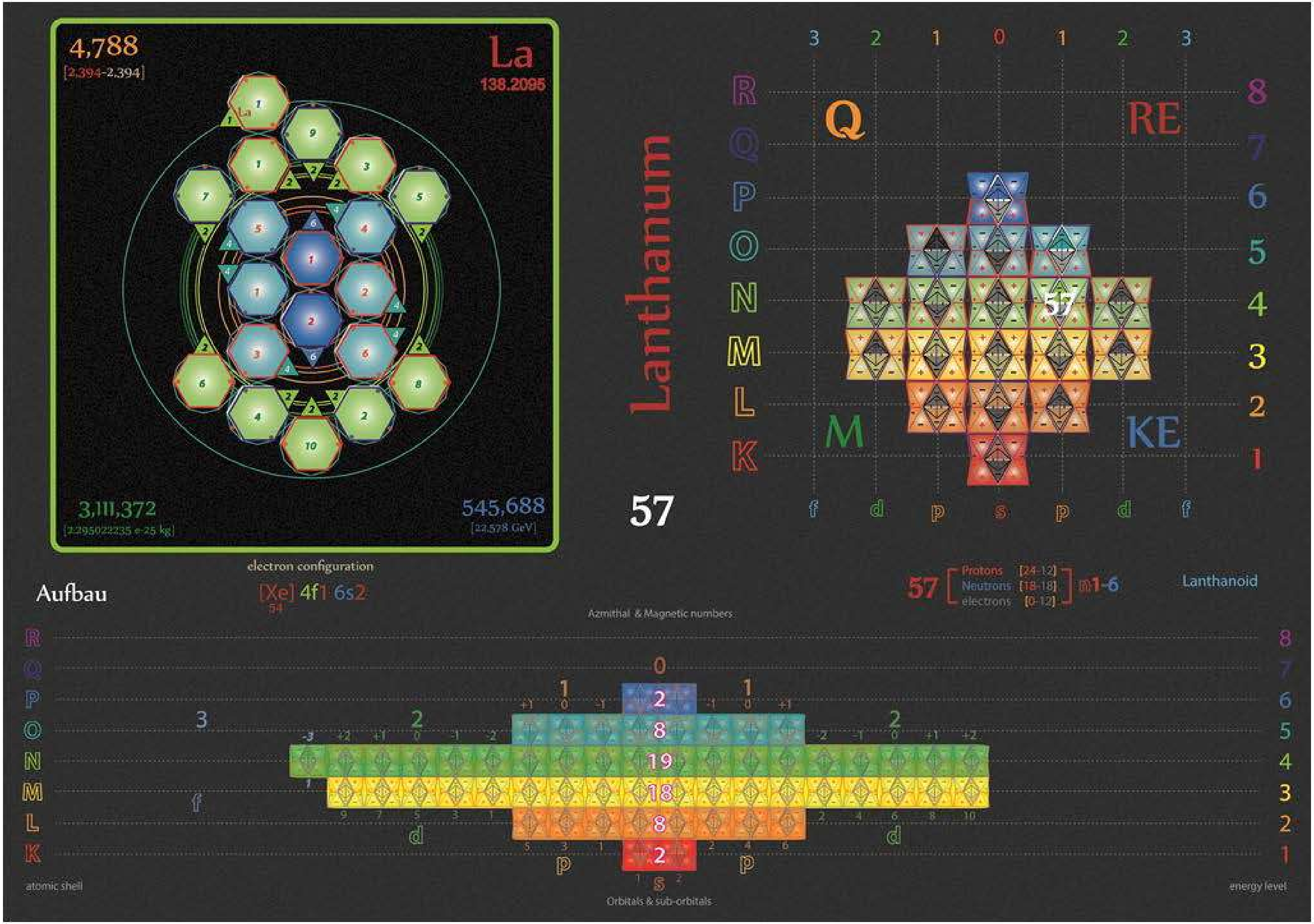
4,788
[2,394-2,394]
La
138.2095
3,111,372
[2.295022235 e-25 kg]
545,688
[22.578 GeV]
Lanthanum
57
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
57
electron configuration
[Xe] 4f1 6s2
54
Aufbau
57 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
2
8
19
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.57 - Lanthanum atom

Copyright ABRAHAM [2008] - All rights reserved

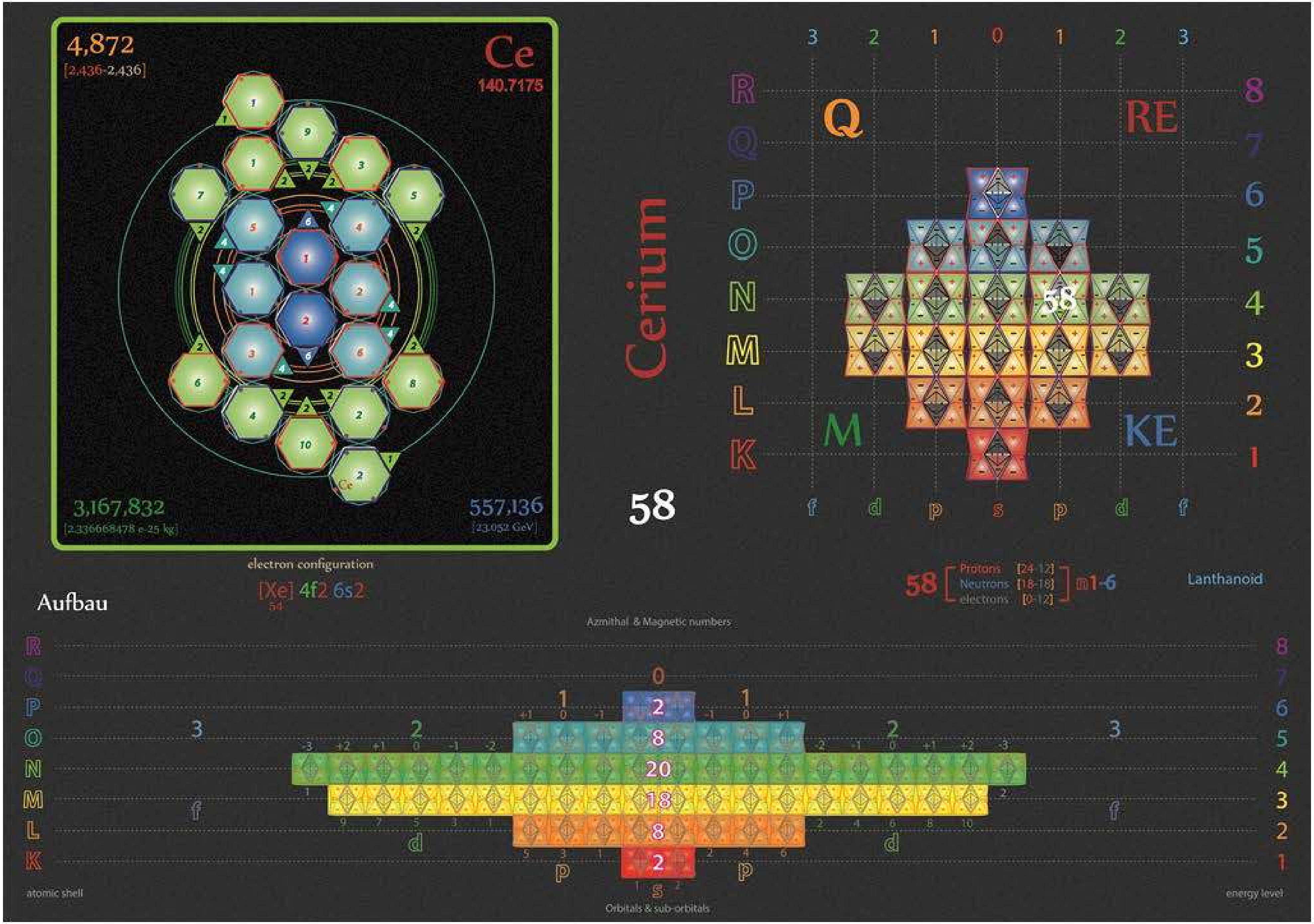
4,872
[2,436-2,436]
Ce
140.7175
3,167,832
[2.336668478 e-25 kg]
557,136
[23.052 GeV]
Cerium
58
electron configuration
[Xe] 4f2 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
58
f d p s p d f
58
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1
2
3
2
8
20
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.58 - Cerium atom

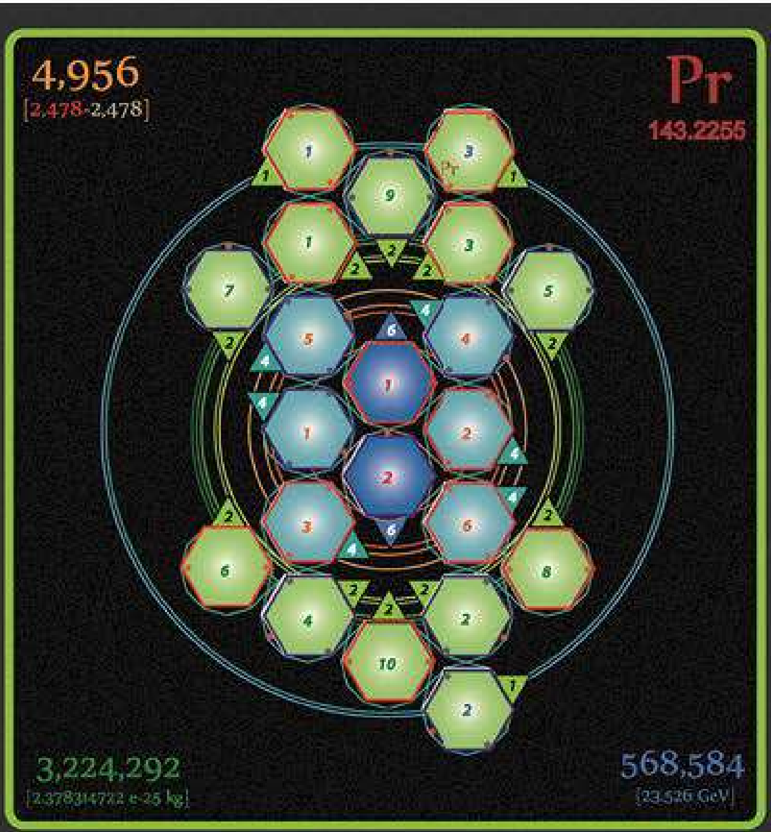
4,956
[2,478-2,478]
Pr
143.2255
3,224,292
[2.378314722 e-25 kg]
568,584
[23.526 GeV]

Praseodymium

59

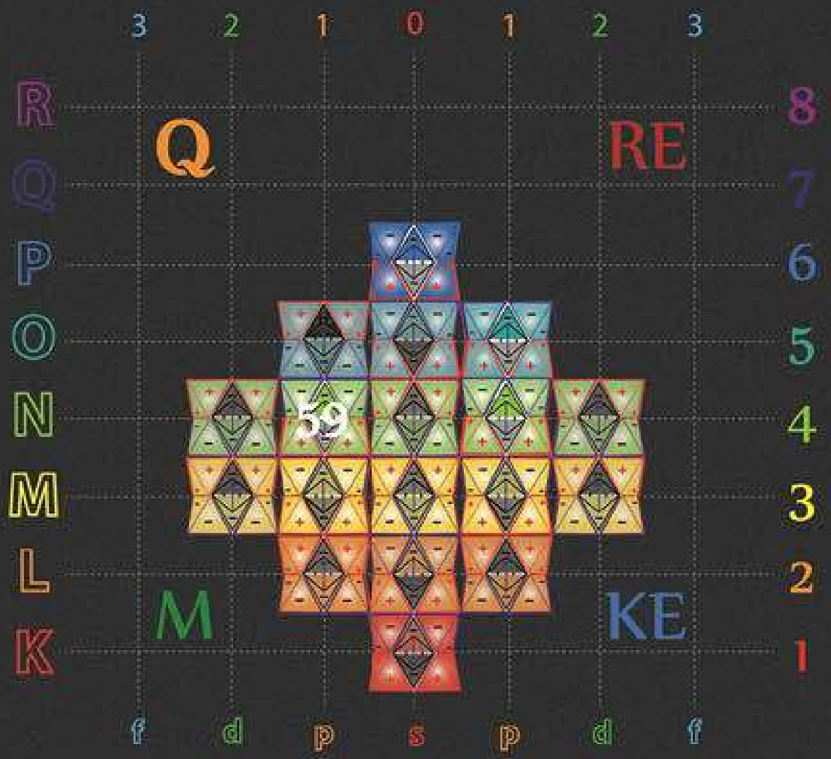
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
59
8
7
6
5
4
3
2
1
f d p s p d f

electron configuration

[Xe] 4f3 6s2

54

59 [Protons [24-12]
Neutrons [18-18]
electrons [0-12]] n1-6

Lanthanoid

Aufbau

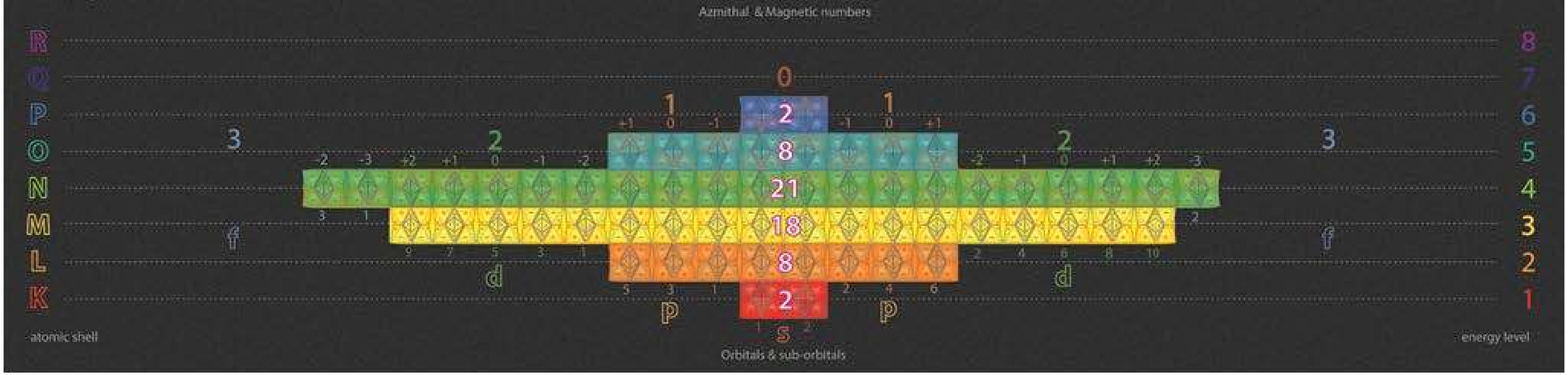
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
2
8
21
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.59 - Praseodymium atom

Copyright ABRAHAM [2008] - All rights reserved

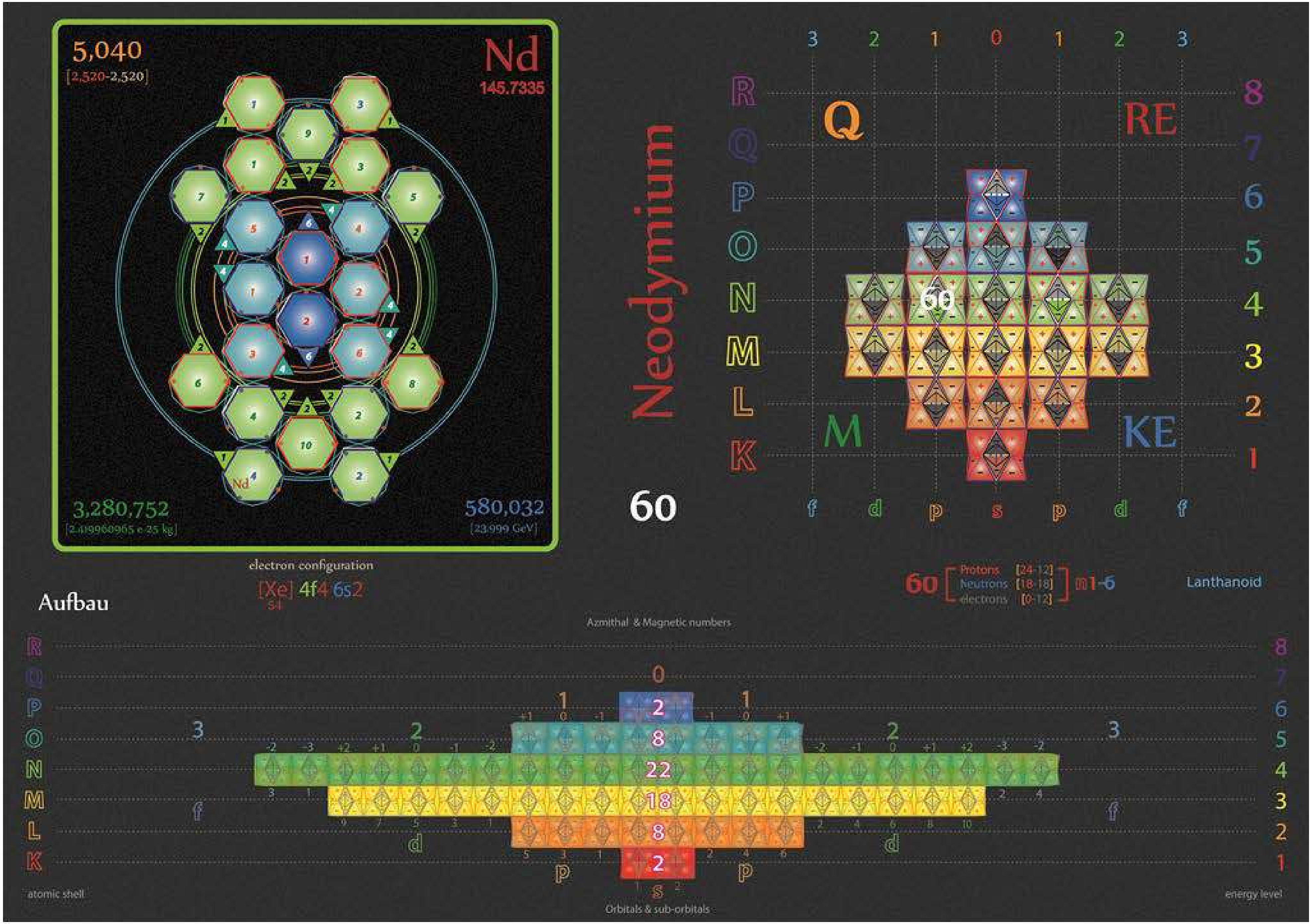
5,040
[2,520-2,520]
Nd
145.7335
3,280,752
[2.49960965 e-25 kg]
580,032
[23.999 GeV]
Neodymium
60
electron configuration
[Xe] 4f4 6s2
54
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
60 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
22
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.60 - Neodymium atom

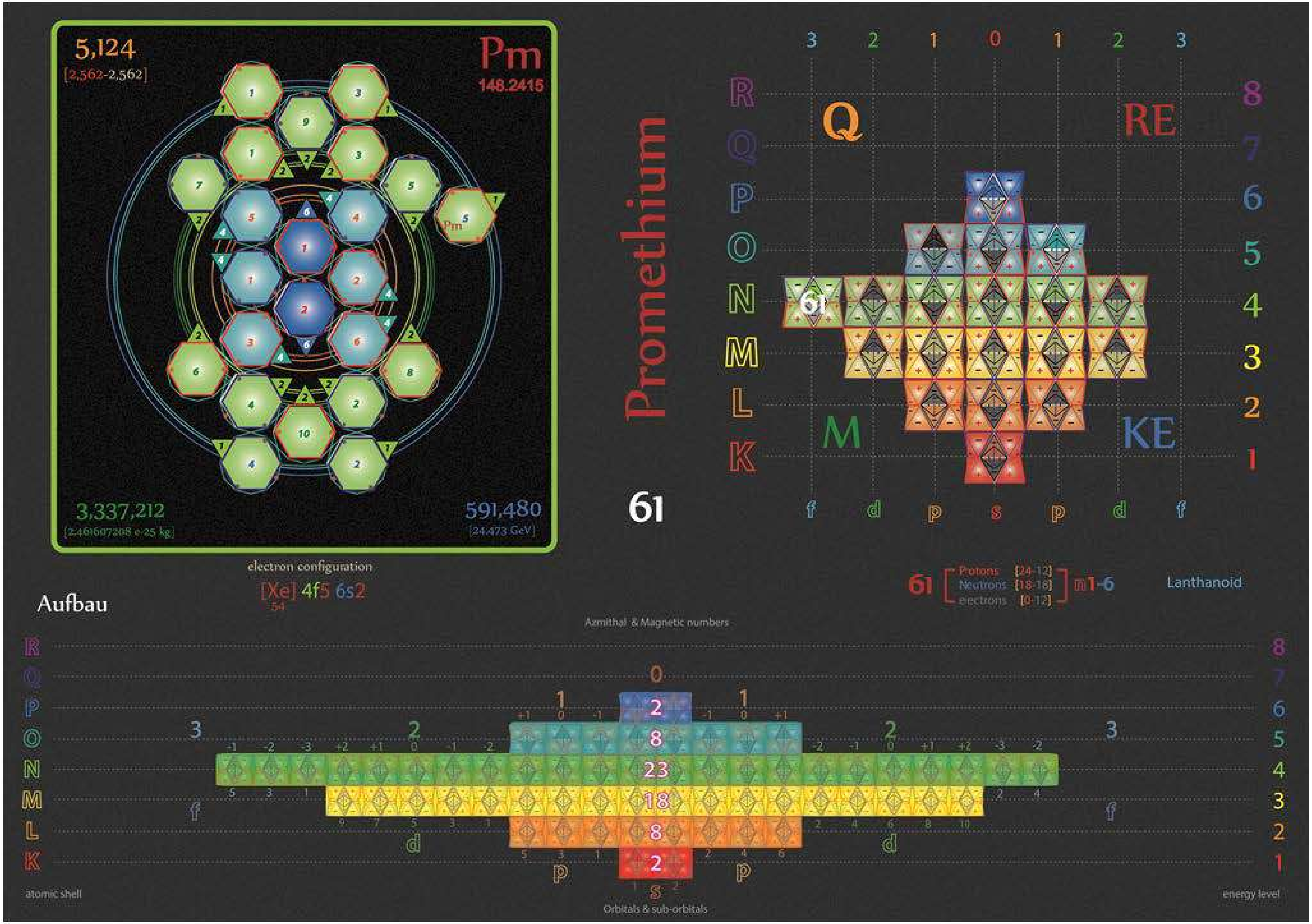
5,124
[2,562-2,562]
Pm
148.2415
3,337,212
[2.461607208 e-25 kg]
591,480
[24.473 GeV]
Promethium
61
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
[Xe] 4f5 6s2
54
61
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-6
Lanthanoid
Aufbau
Azimthal & Magnetic numbers
2
8
23
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.61 - Promethium atom

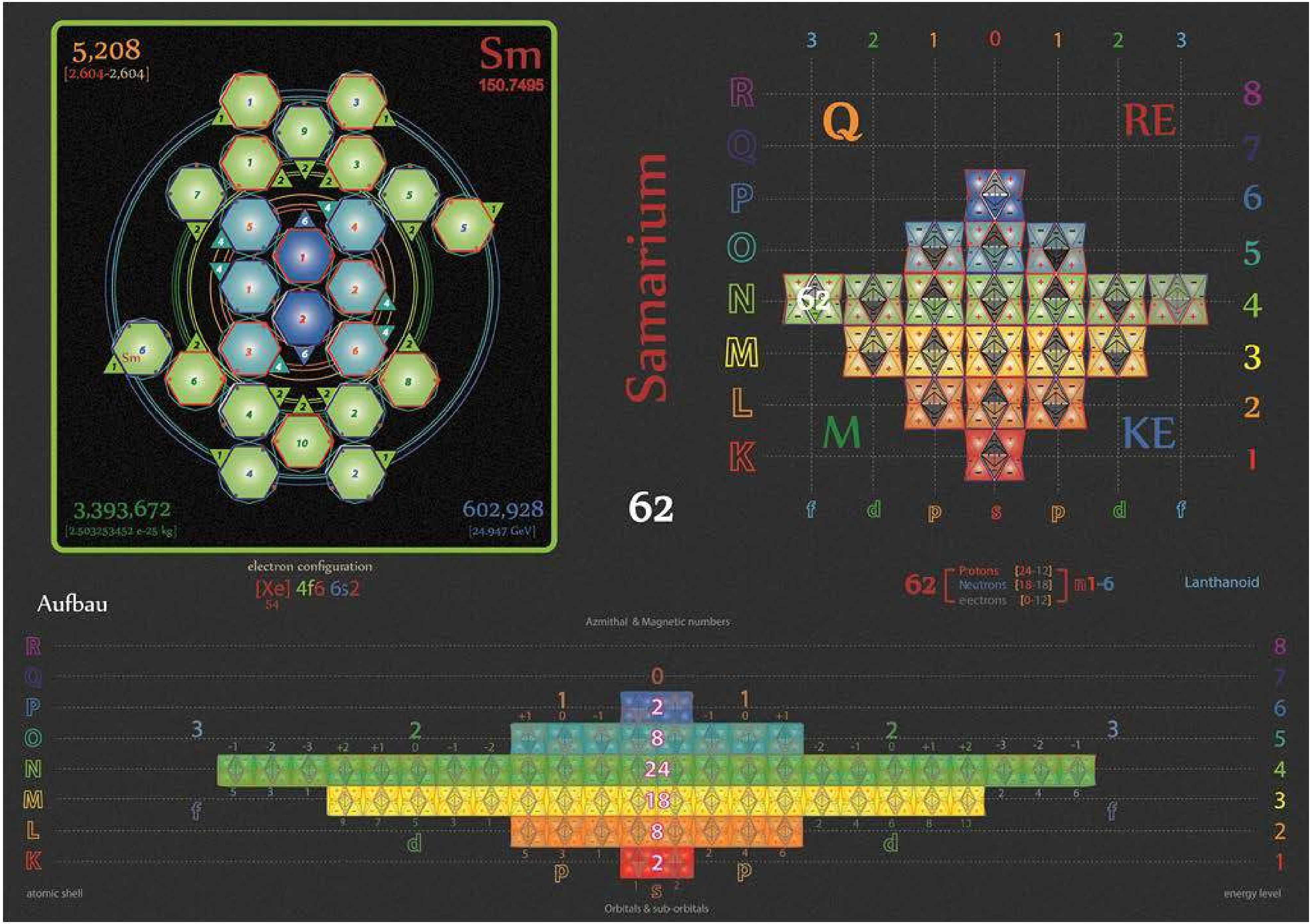
5,208
[2,604-2,604]
Sm
150.7495
3,393,672
[2.503233452 e-25 kg]
602,928
[24.947 GeV]
Samarium
62
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f6 6s2
54
62 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
8
24
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.62 - Samarium atom

Copyright ABRAHAM [2008] - All rights reserved

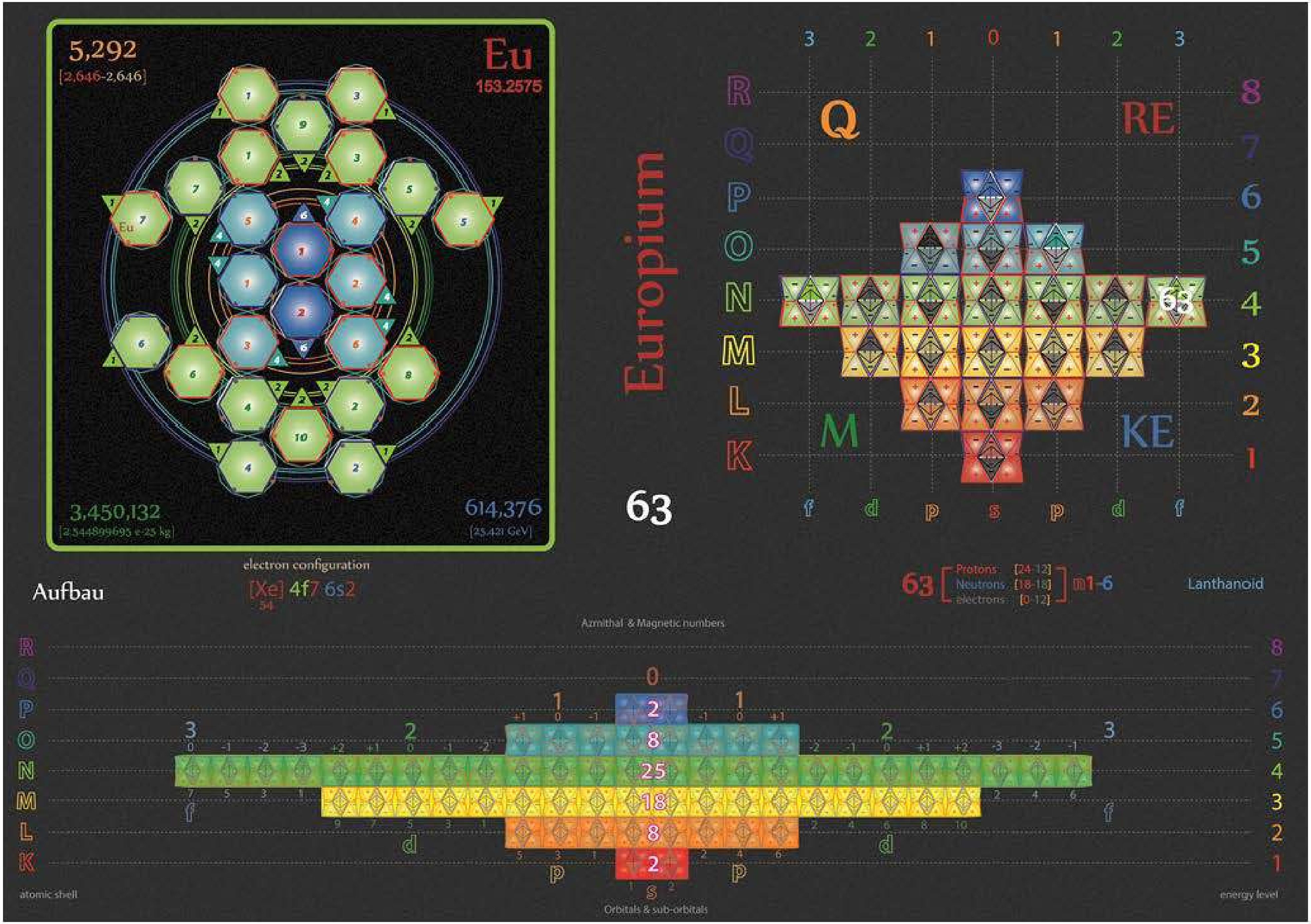
5,292
[2,646-2,646]
Eu
153.2575
3,450,132
[2.544899695 e-25 kg]
614,376
[25.421 GeV]
Europium
63
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
[Xe] 4f7 6s2
54
Aufbau
63 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
25
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.63 - Europium atom

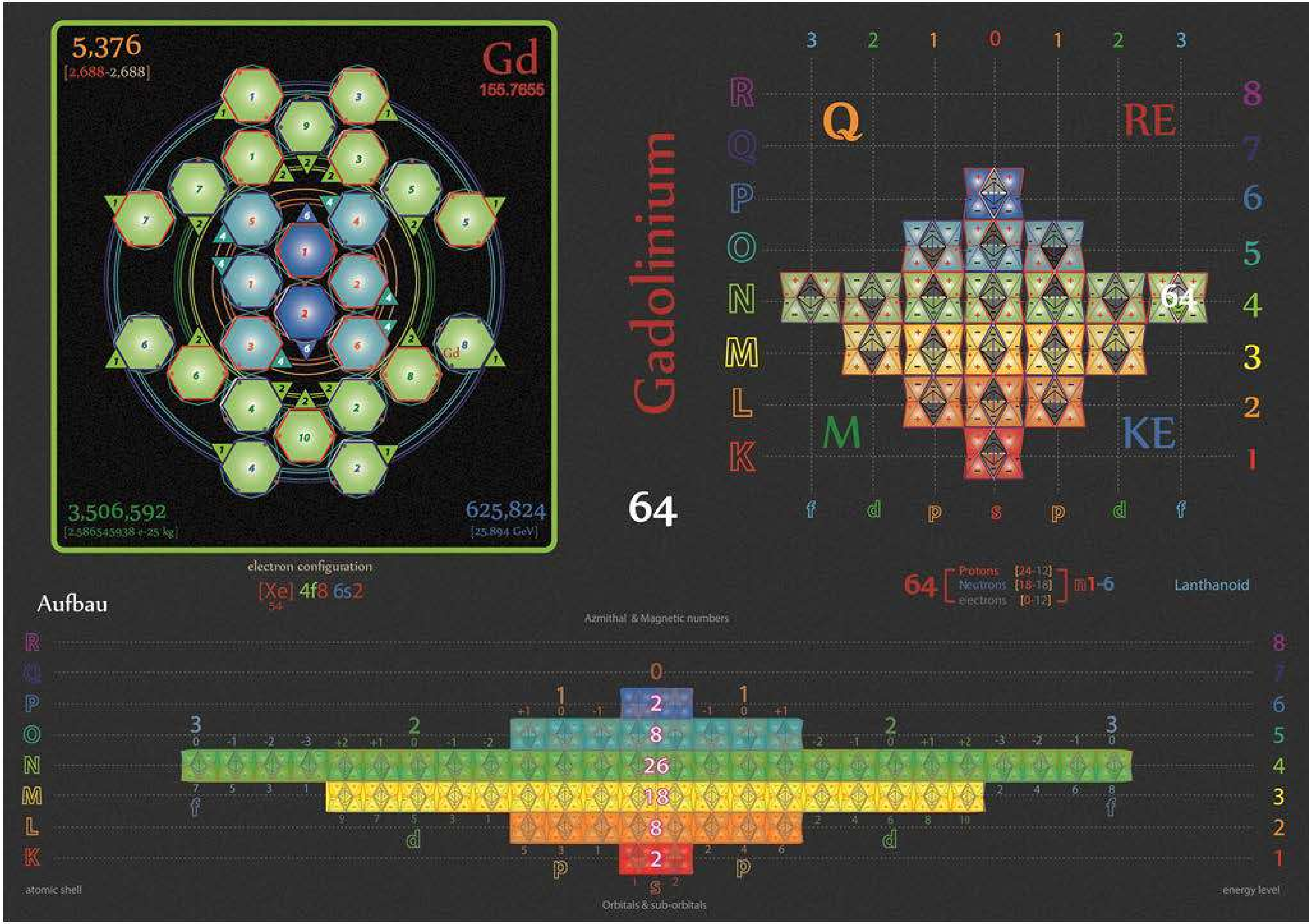
5,376
[2,688-2,688]
Gd
155.7655
3,506,592
[2.586545938 e-25 kg]
625,824
[25.894 GeV]
Gadolinium
64
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
64
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Xe] 4f8 6s2
54
64 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
2
8
26
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.64 - Gadolinium atom

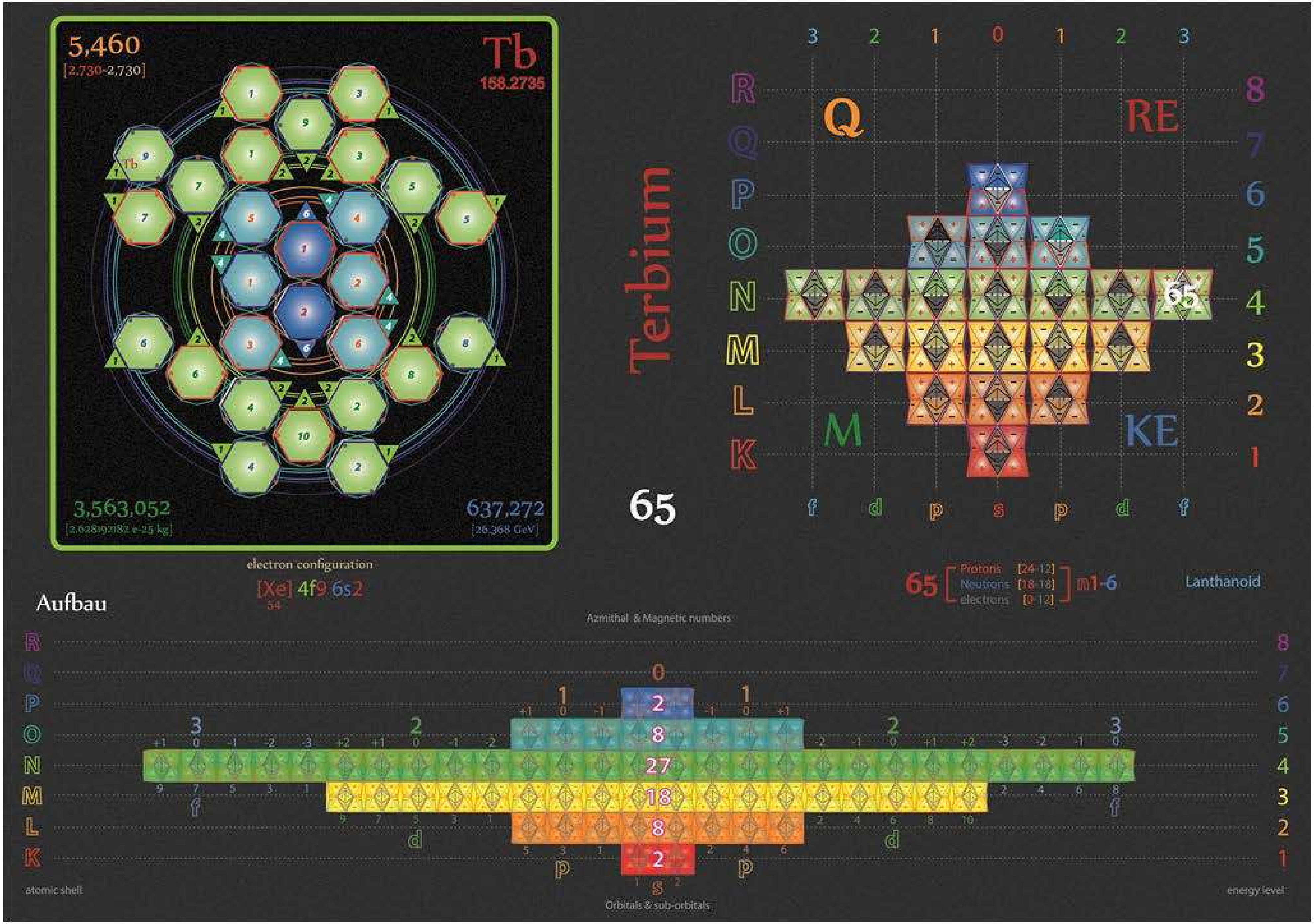
5,460
[2,730-2,730]
Tb
158.2735
3,563,052
[2.62819282 e-25 kg]
637,272
[26.368 GeV]
Terbium
65
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f9 6s2
54
Aufbau
65 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
27
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.65 -Terbium atom

Copyright ABRAHAM [2008] - All rights reserved

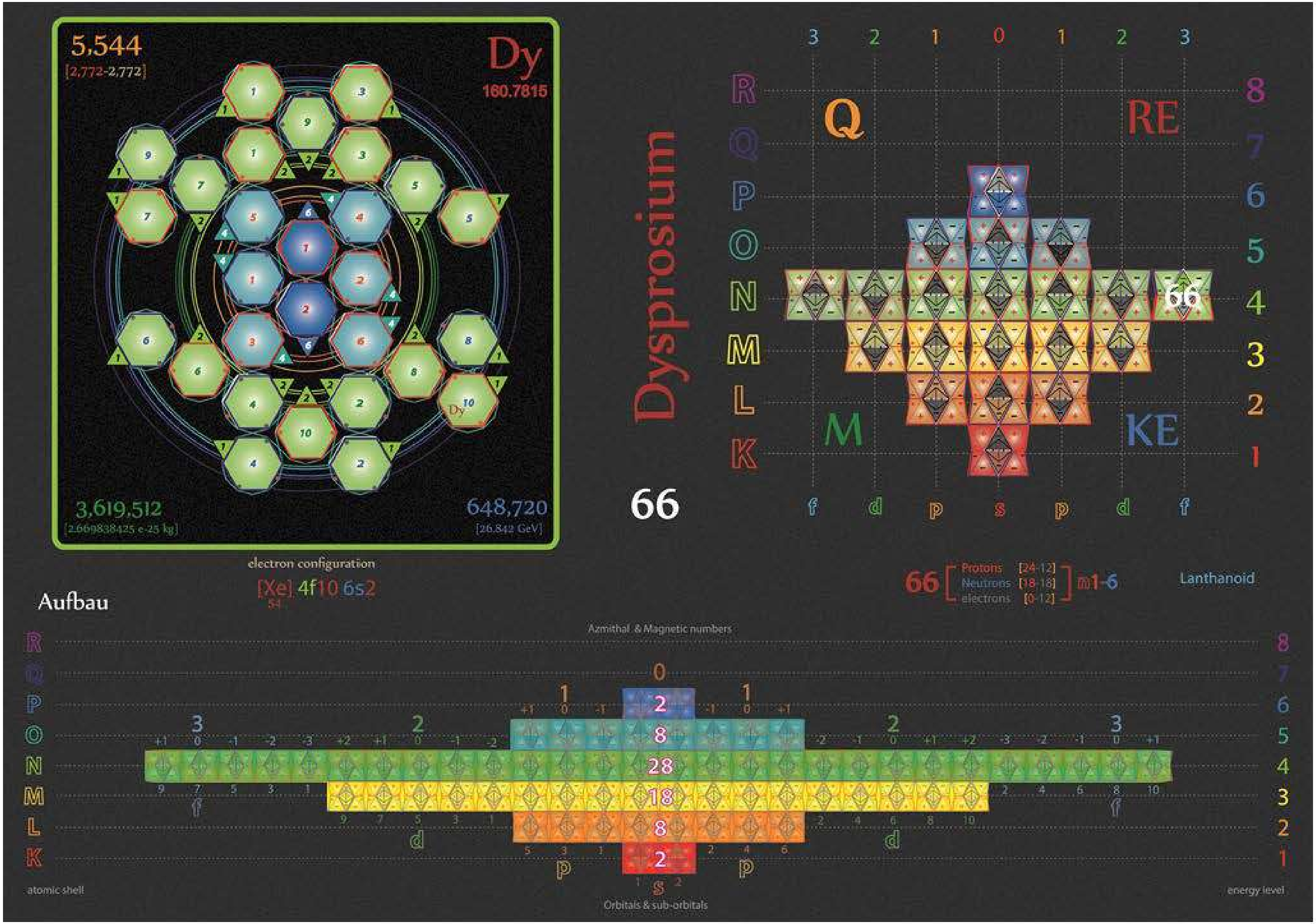
5,544
[2,772-2,772]
Dy
160.7815
3,619,512
[2.669838425 e-25 kg]
648,720
[26.842 GeV]
Dysprosium
66
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
66
electron configuration
[Xe] 4f10 6s2
54
Aufbau
66 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
3 2 2 3
2
8
28
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.66 - Dysprosium atom

Copyright ABRAHAM [2008] - All rights reserved

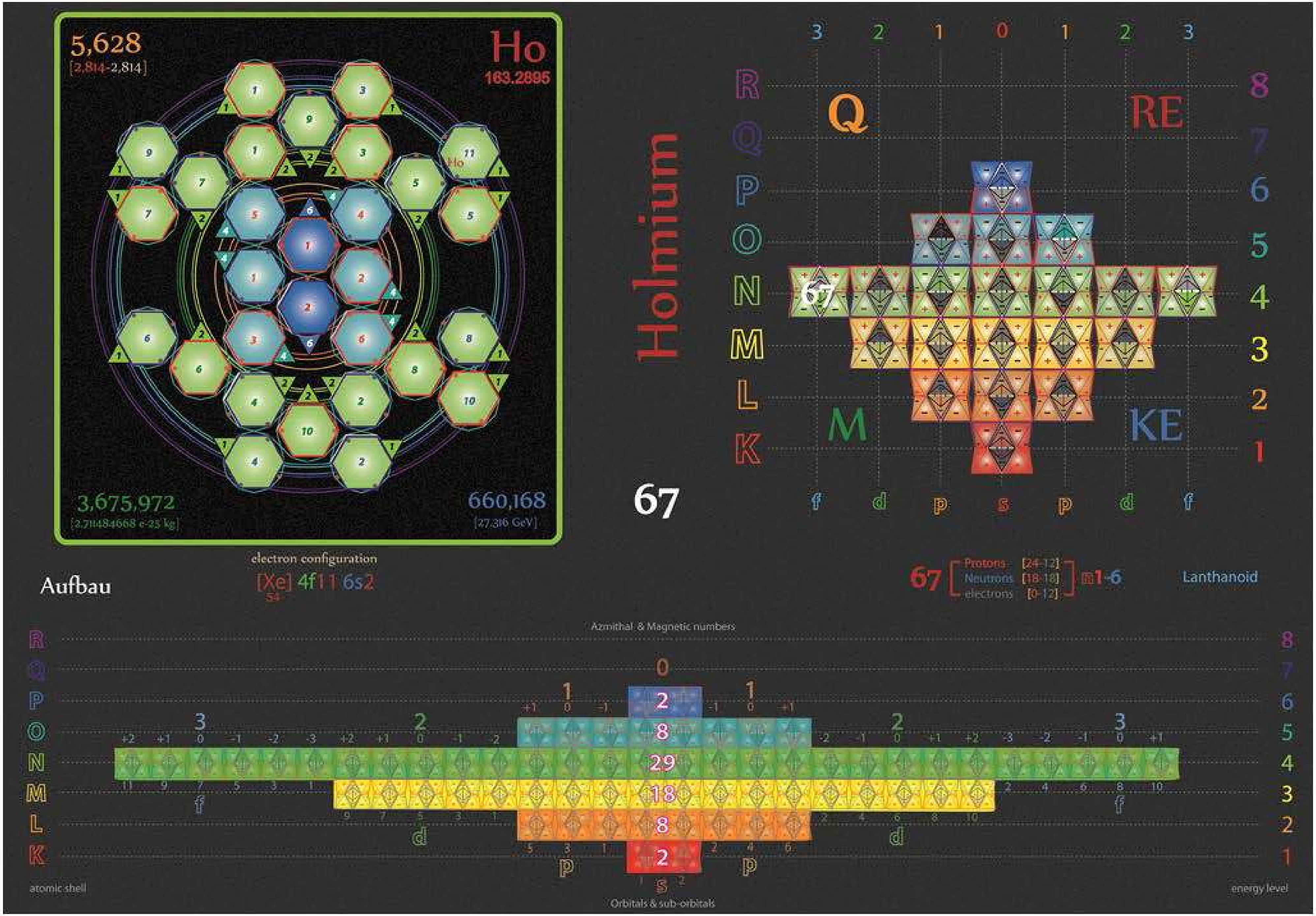
5,628
[2,814-2,814]
Ho
163.2895
3,675,972
[2,71484668 e-25 kg]
660,168
[27,316 GeV]
Holmium
67
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f11 6s2
54
Aufbau
67 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1
1
2
3
2
3
2
8
29
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.67 - Holmium atom

Copyright ABRAHAM [2008] - All rights reserved

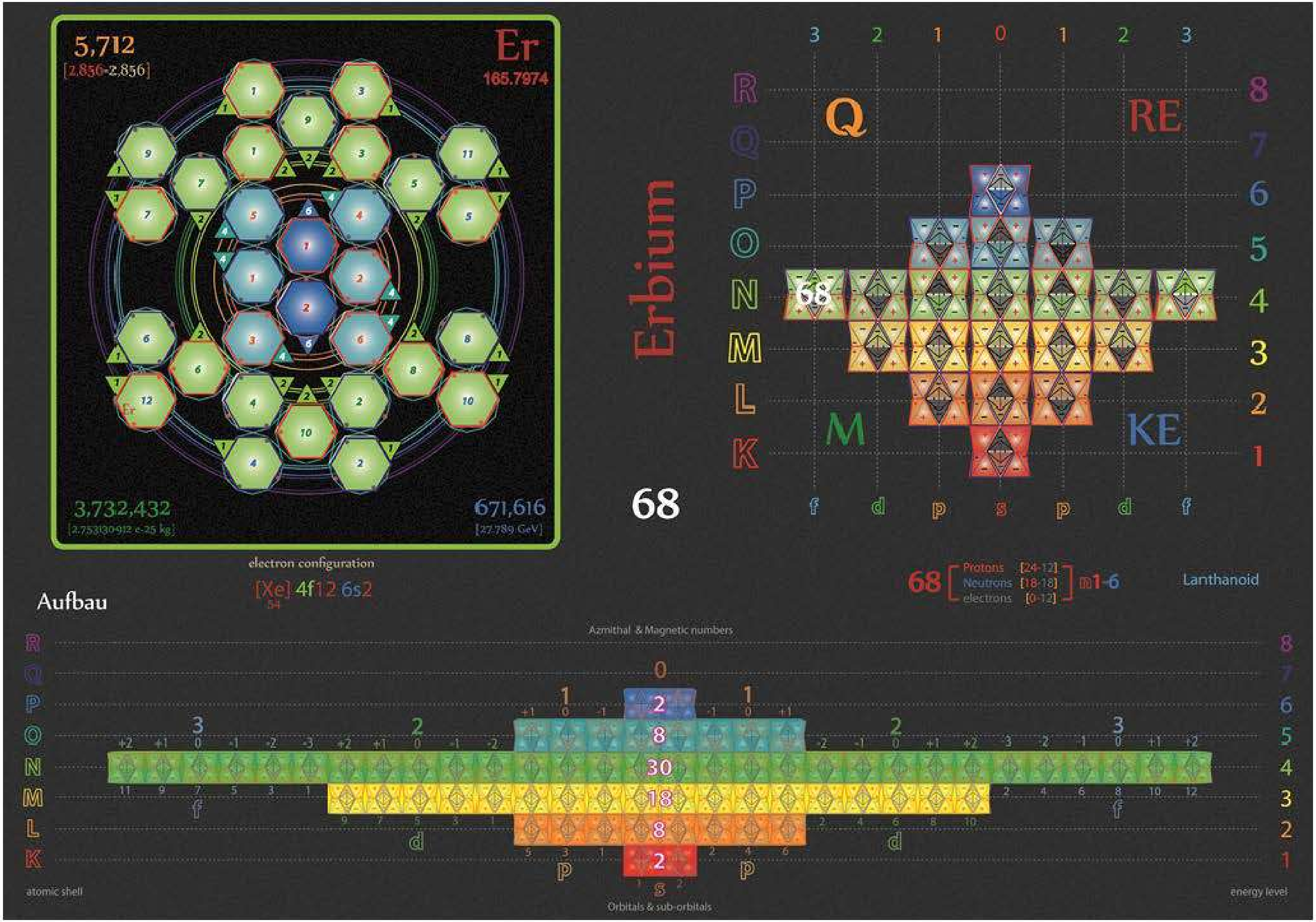
5,712
[2,856-2,856]
Er
165.7974
3,732,432
[2.753130912 e-25 kg]
671,616
[27.789 GeV]
Erbium
68
electron configuration
[Xe] 4f12 6s2
54
Q
RE
M
KE
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
68
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
2
8
30
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.68 - Erbium atom

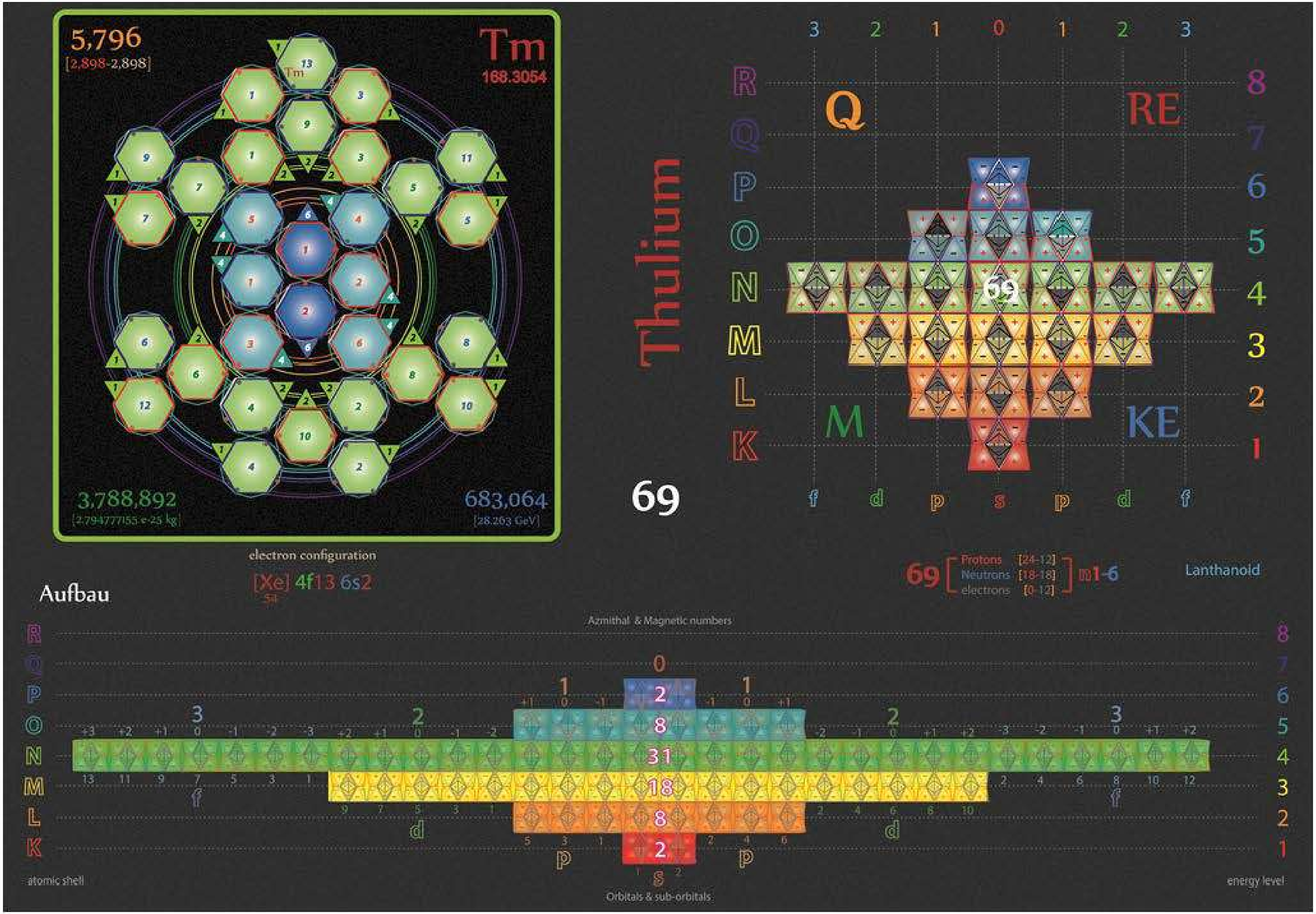
5,796
[2,898-2,898]
Tm
168.3054
3,788,892
[2.794777155 e-25 kg]
683,064
[28.263 GeV]
Thulium
69
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f13 6s2
54
69 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
1
2
3
2
8
31
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.69 - Thulium atom

Copyright ABRAHAM [2008] - All rights reserved

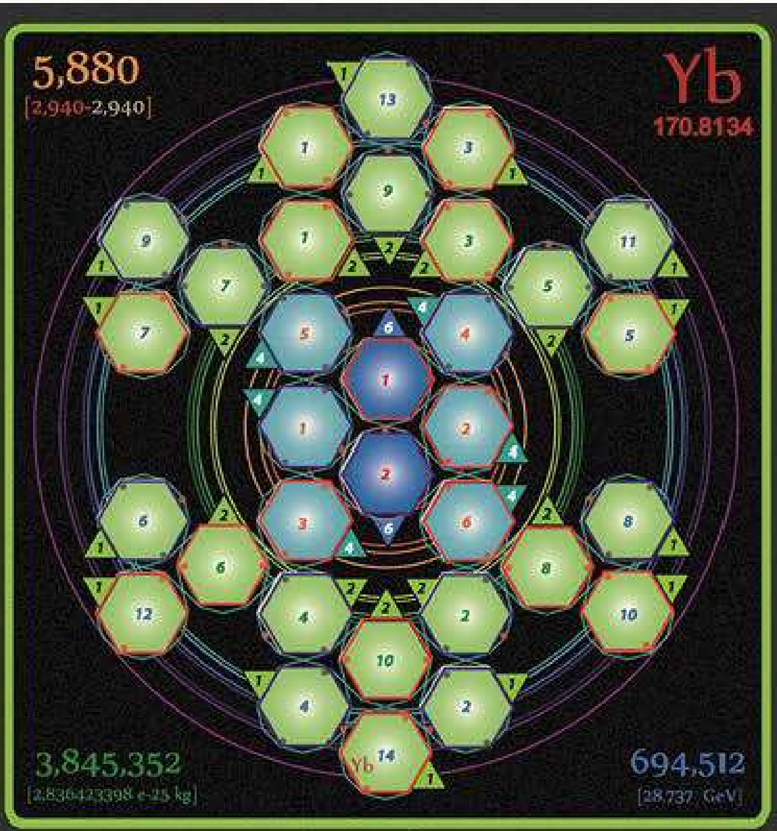
5,880
[2,940-2,940]
Yb
170.8134
3,845,352
[2.836423398 e-25 kg]
694,512
[28,737 GeV]
Yb

Ytterbium

70

electron configuration

[Xe] 4f14 6s2
54

Aufbau

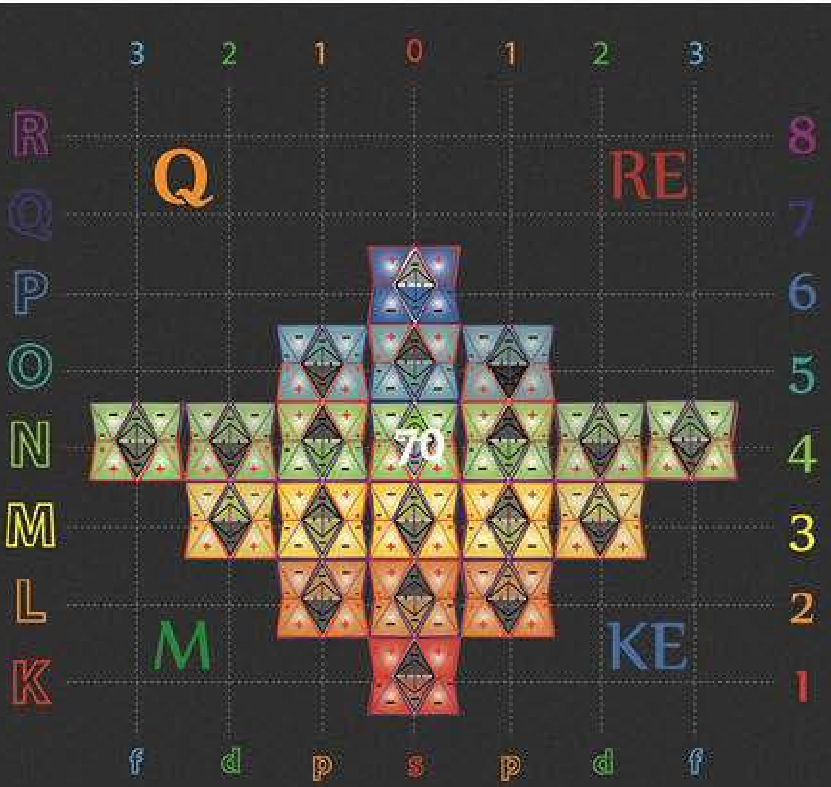
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
70
8
7
6
5
4
3
2
1
f d p s p d f

70 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Lanthanoid

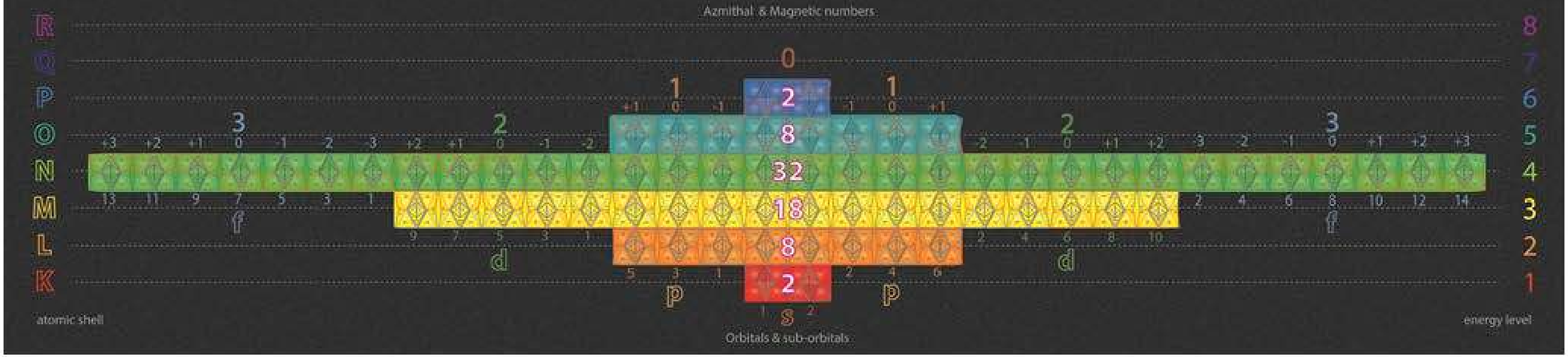
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
2
8
32
18
8
2
1 1
3 2 2 3
+3 +2 +1 0 -1 -2 -3
f d p s p d f
8
7
6
5
4
3
2
1
atomic shell
energy level
Orbitals & sub-orbitals

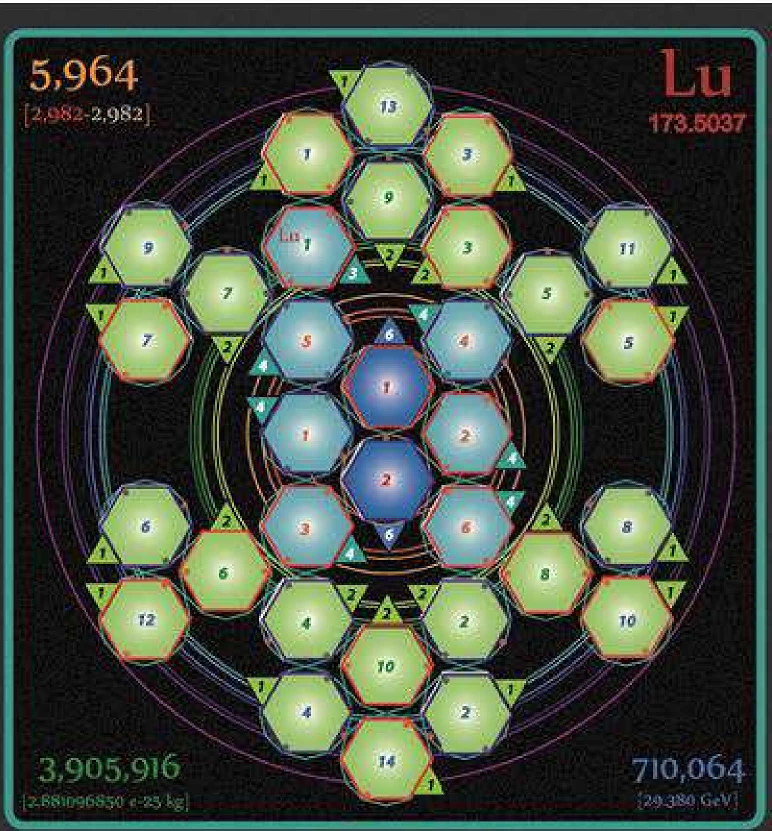
5,964
[2,982-2,982]
Lu
173.5037
3,905,916
[2.881096850 e-25 kg]
710,064
[29.380 GeV]

Lutetium

71

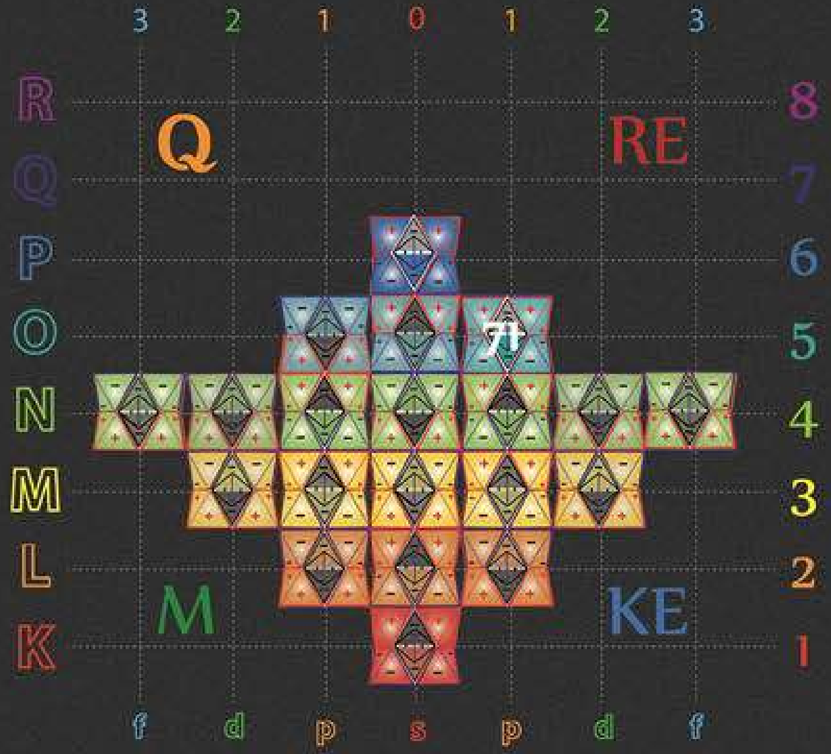
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
71
f d p s p d f

electron configuration

[Xe] 4f14 5d1 6s2
54

71 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Lanthanoid

Aufbau

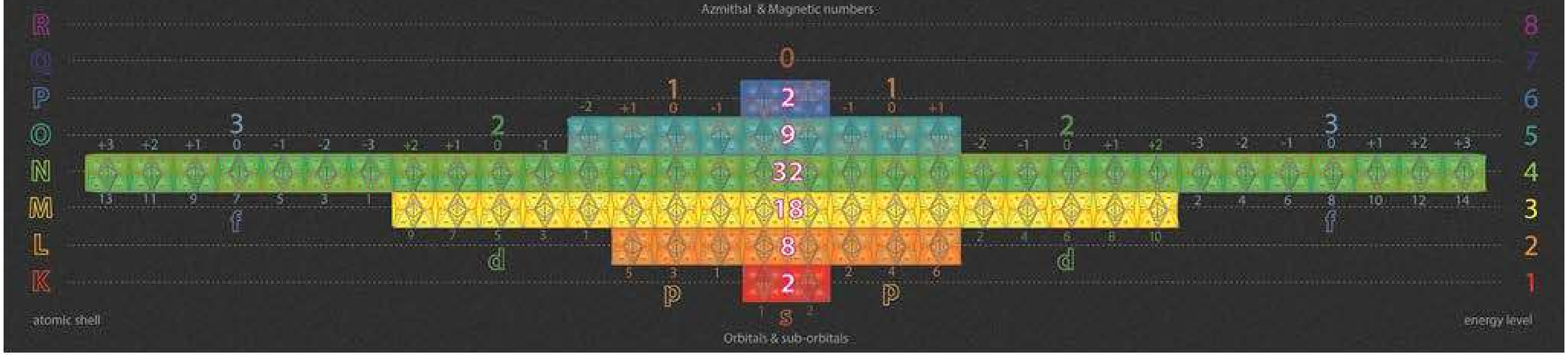
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1 1
2 2
3 3
2
9
32
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.71 - Lutetium atom

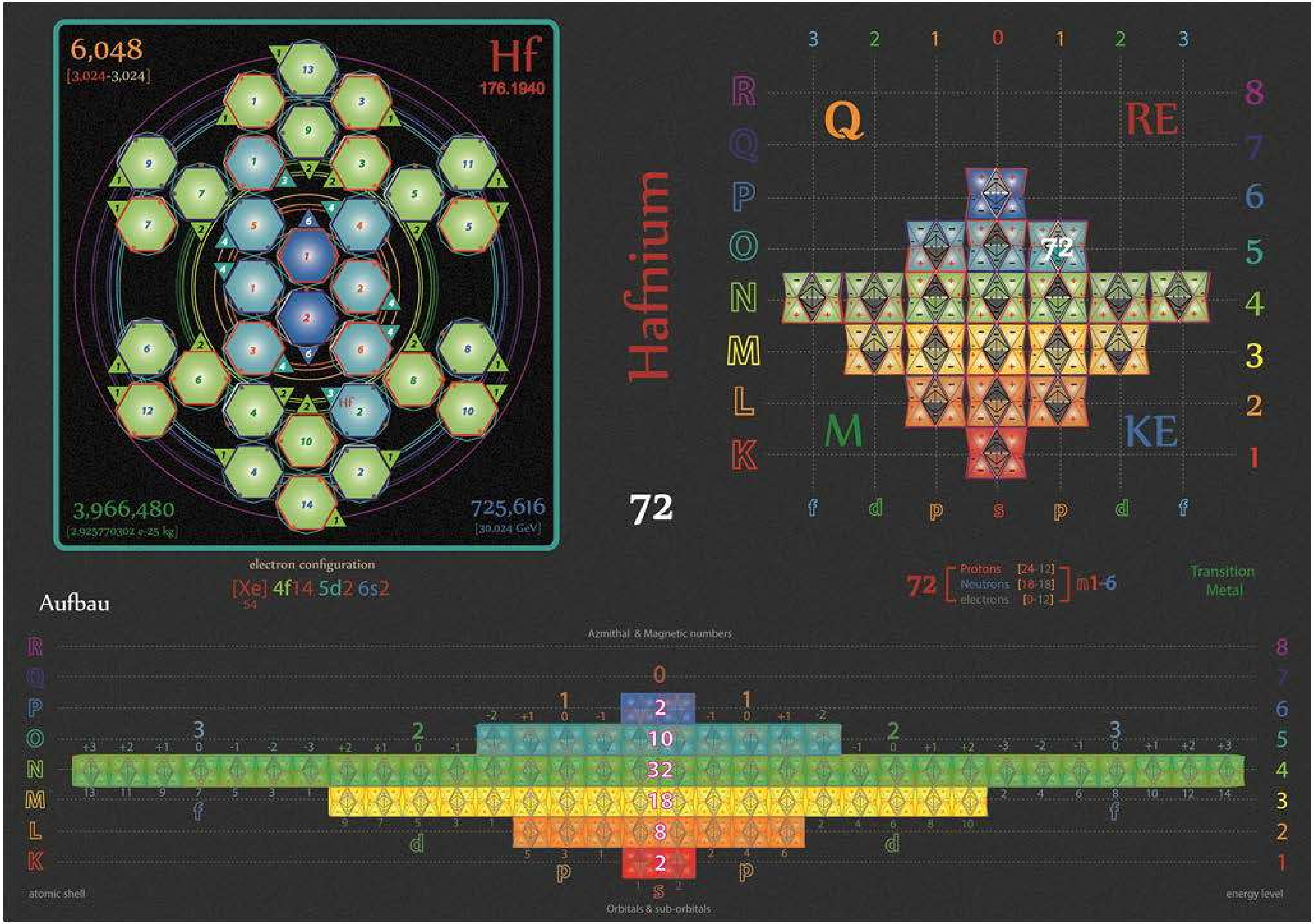
6,048
[3,024-3,024]
Hf
178.1940
3,966,480
[2.925770302 e-25 kg]
725,616
[30.024 GeV]
Hafnium
72
electron configuration
[Xe] 4f14 5d2 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
72
f d p s p d f
72
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Transition Metal
Azmithal & Magnetic numbers
0
2
10
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.72 - Hafnium atom

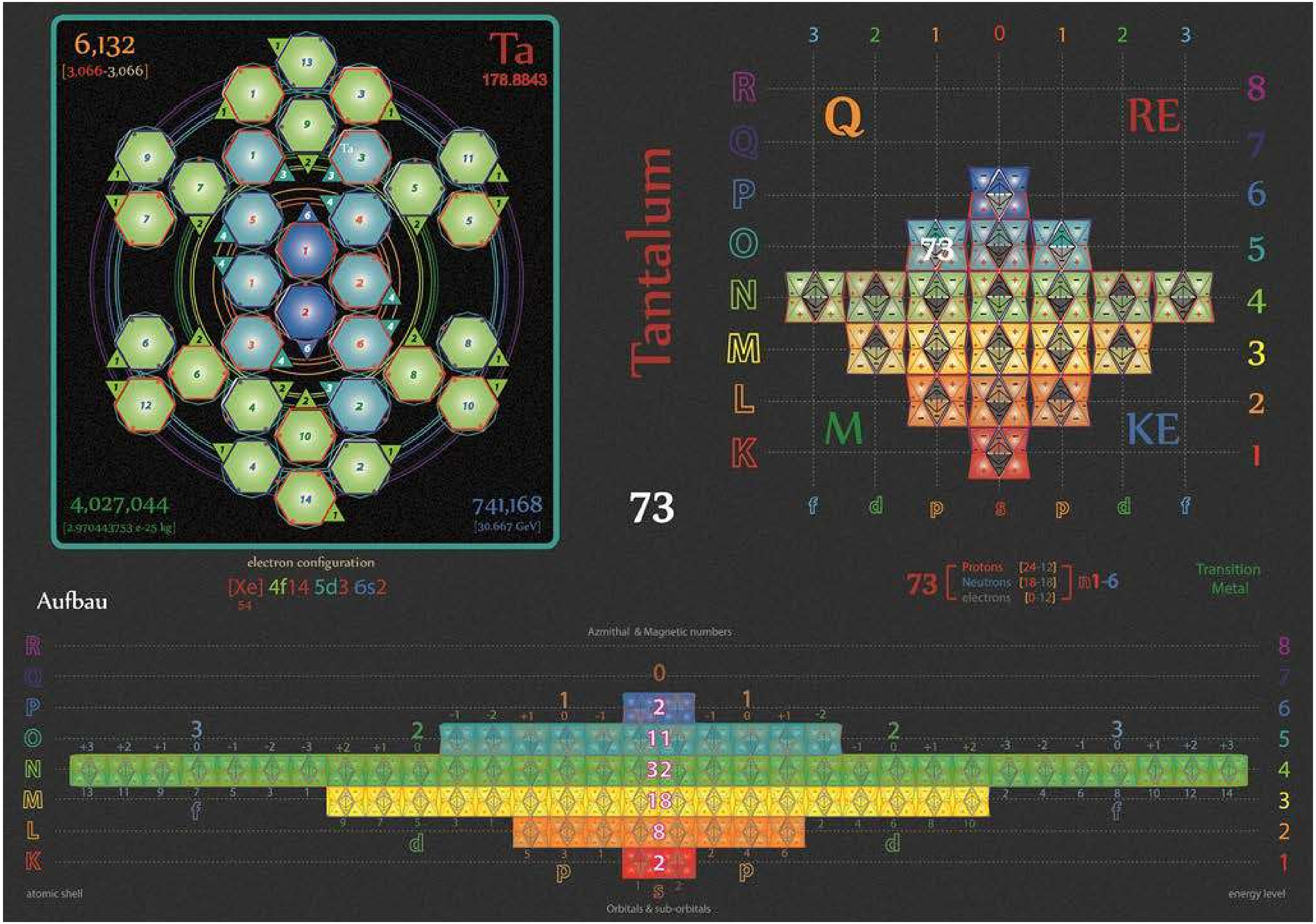
6,132
[3,066-3,066]
Ta
178.8843
4,027,044
[2.970443753 e-25 kg]
741,168
[30.667 GeV]
Tantalum
73
electron configuration
[Xe] 4f14 5d3 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
73
f d p s p d f
73 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
2
11
32
18
8
2
1
1
3
2
2
3
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.73 - Tantalum atom

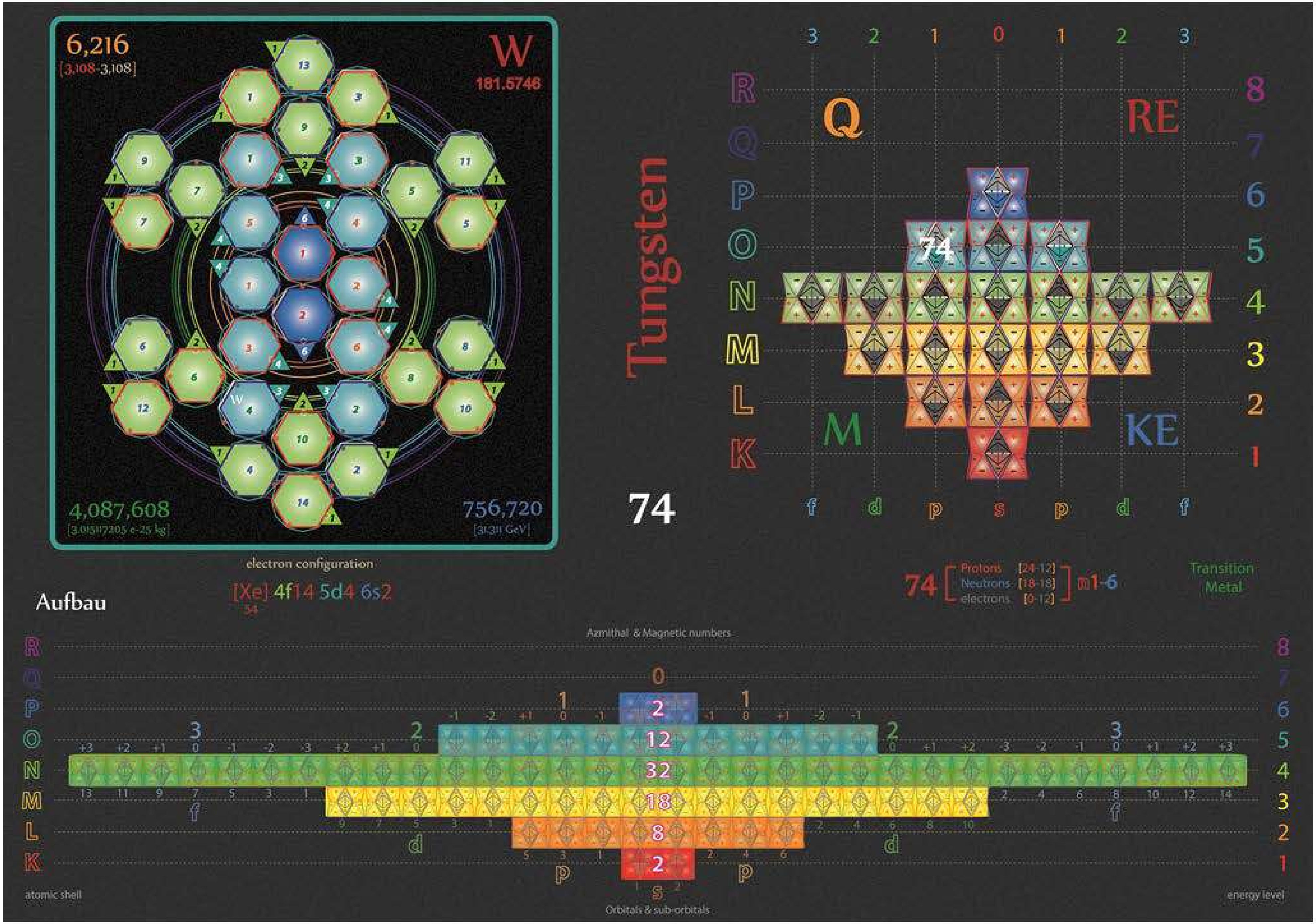
6,216
[3,108-3,108]
W
181.5746
4,087,608
[3.0151172105 e-25 kg]
756,720
[31.311 GeV]
Tungsten
74
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
Q
RE
M
KE
f
d
p
s
p
d
f
electron configuration
[Xe] 4f14 5d4 6s2
54
Aufbau
74
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Transition
Metal
Azmithal & Magnetic numbers
2
12
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.74 - Tungsten atom

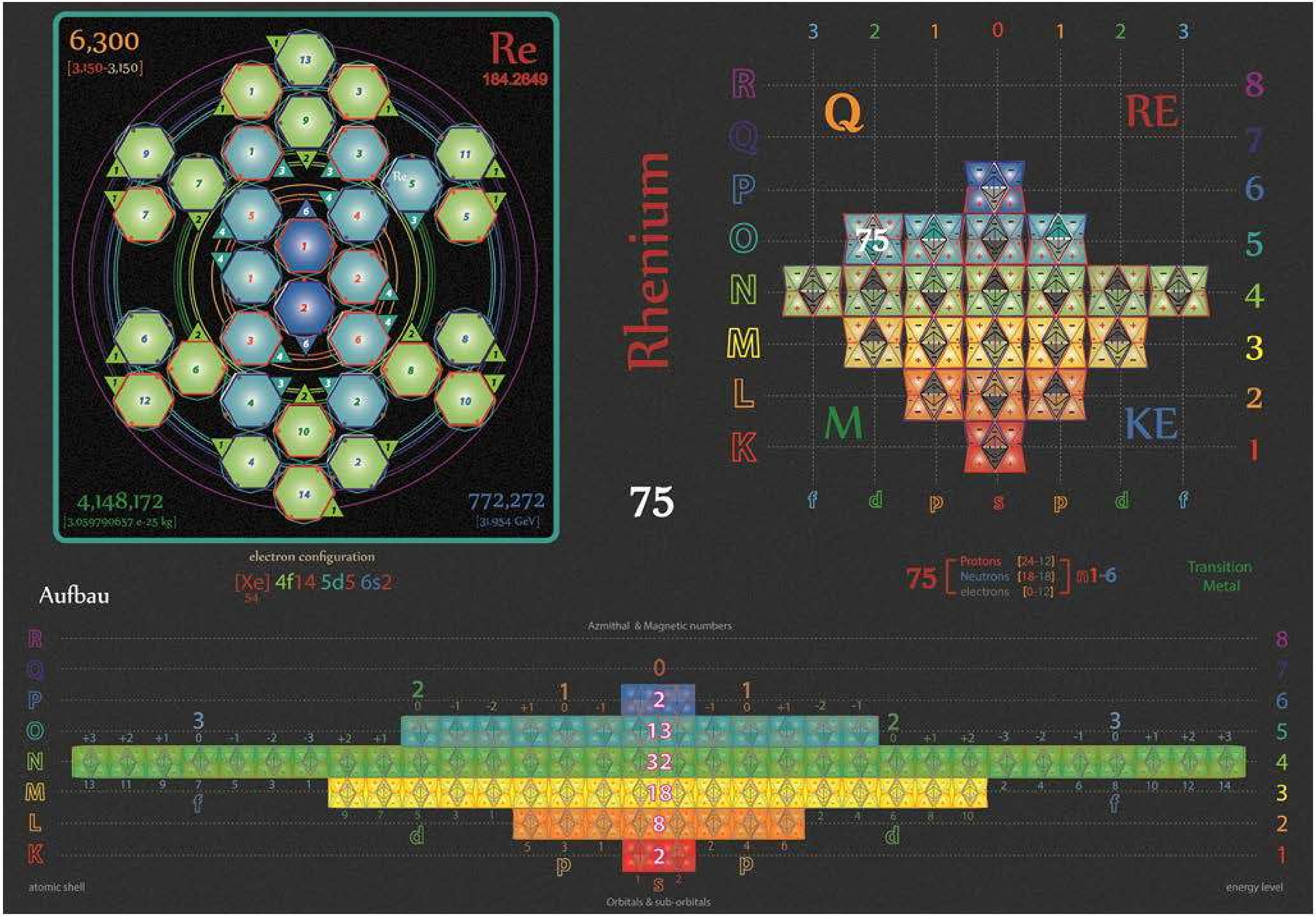
6,300
[3,150-3,150]
Re
184.2849
4,148,172
[3.059790657 e-25 kg]
772,272
[31.954 GeV]
Rhenium
75
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f14 5d5 6s2
54
Aufbau
75 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
2
13
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.75 - Rhenium atom

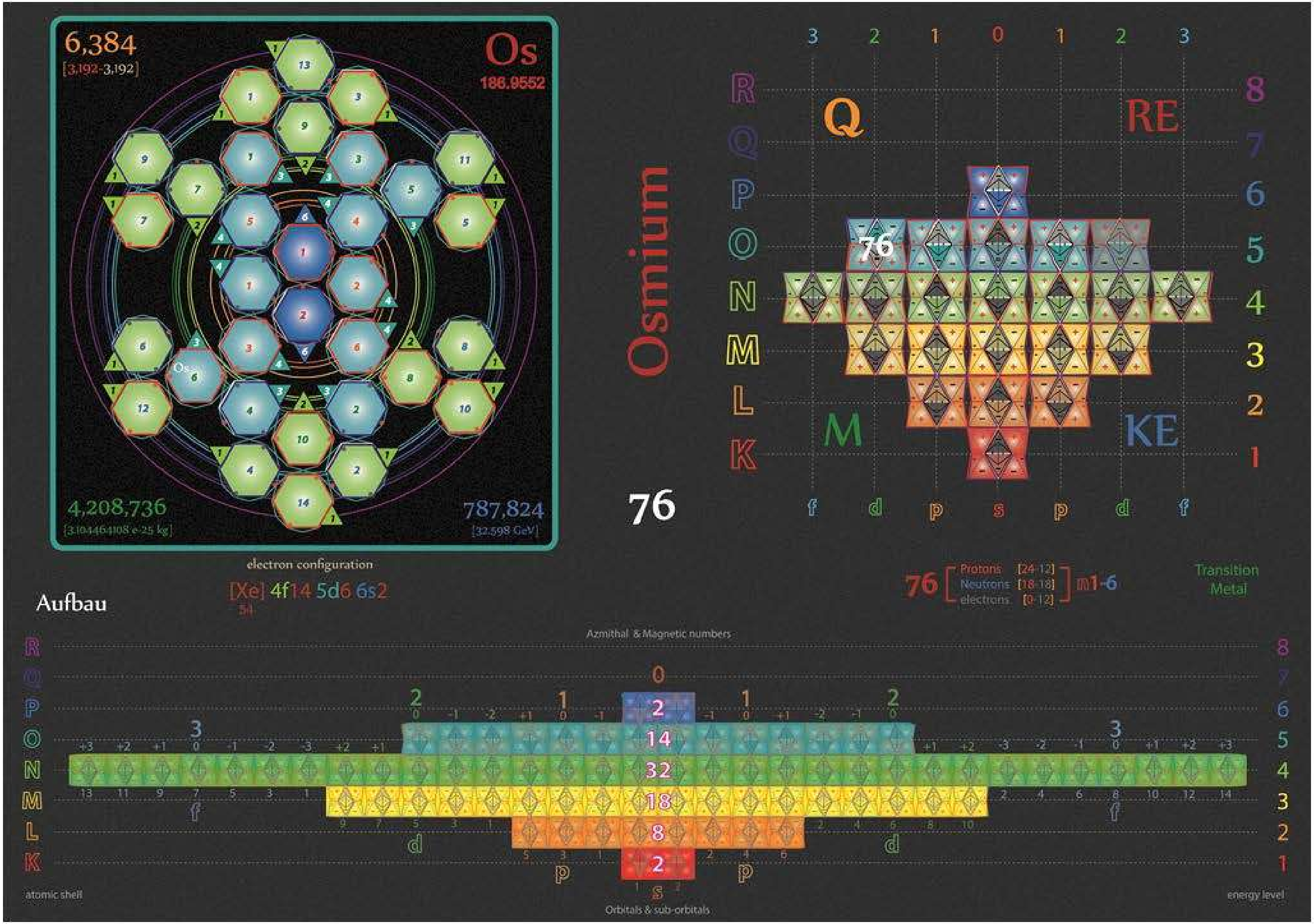
6,384
[3,192-3,192]
Os
186.9552
4,208,736
[3.104464108 e-25 kg]
787,824
[32.598 GeV]
Osmium
76
electron configuration
[Xe] 4f14 5d6 6s2
54
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
f
d
p
s
p
d
f
76
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Transition Metal
Azmithal & Magnetic numbers
2
14
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.76 - Osmium atom

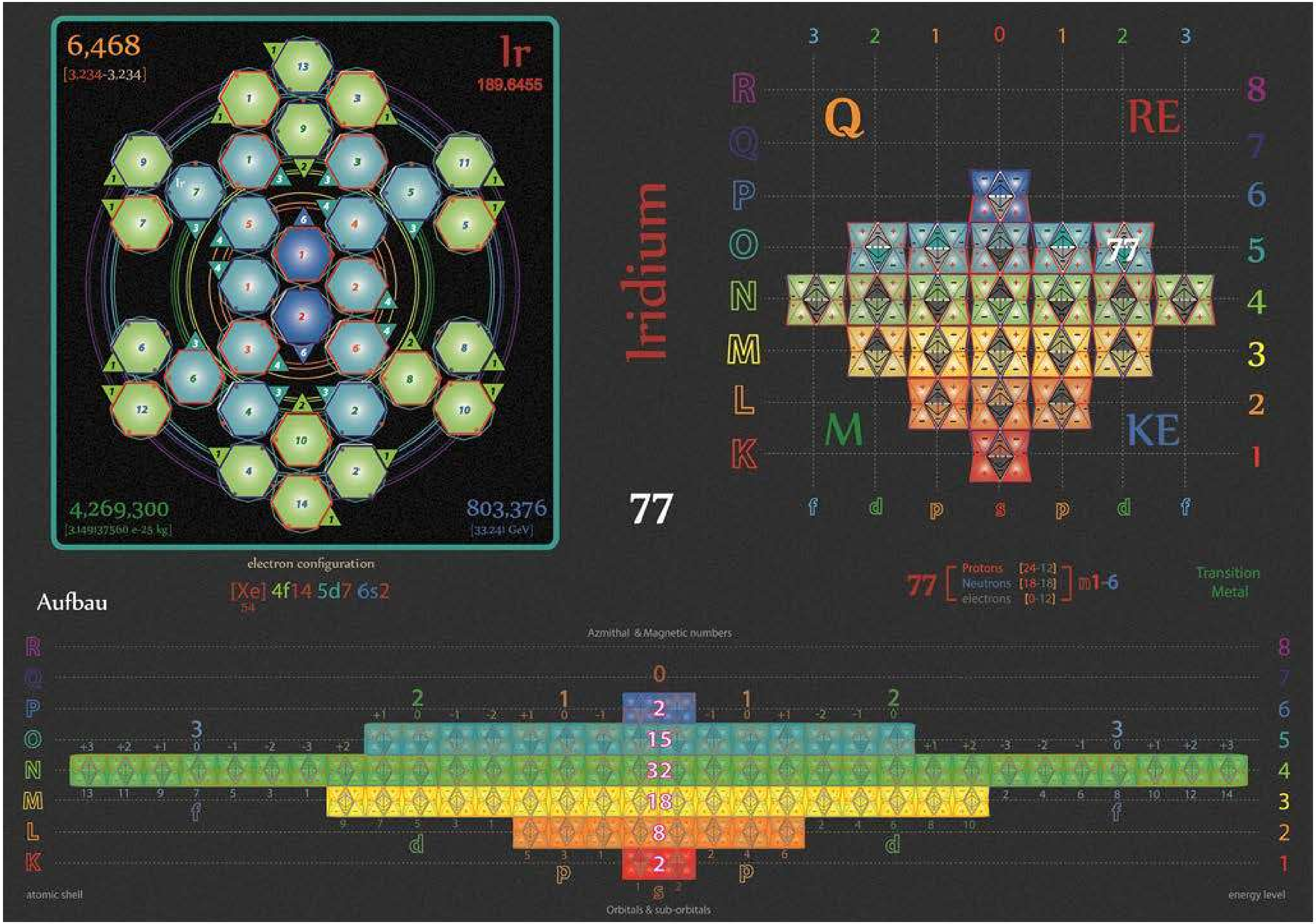
6,468
[3.234-3.234]
Ir
189.8455
4,269,300
[3.149137560 e-25 kg]
803,376
[33.241 GeV]
Iridium
77
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
[Xe] 4f14 5d7 6s2
54
Aufbau
77 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
2
15
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.77 - Iridium atom

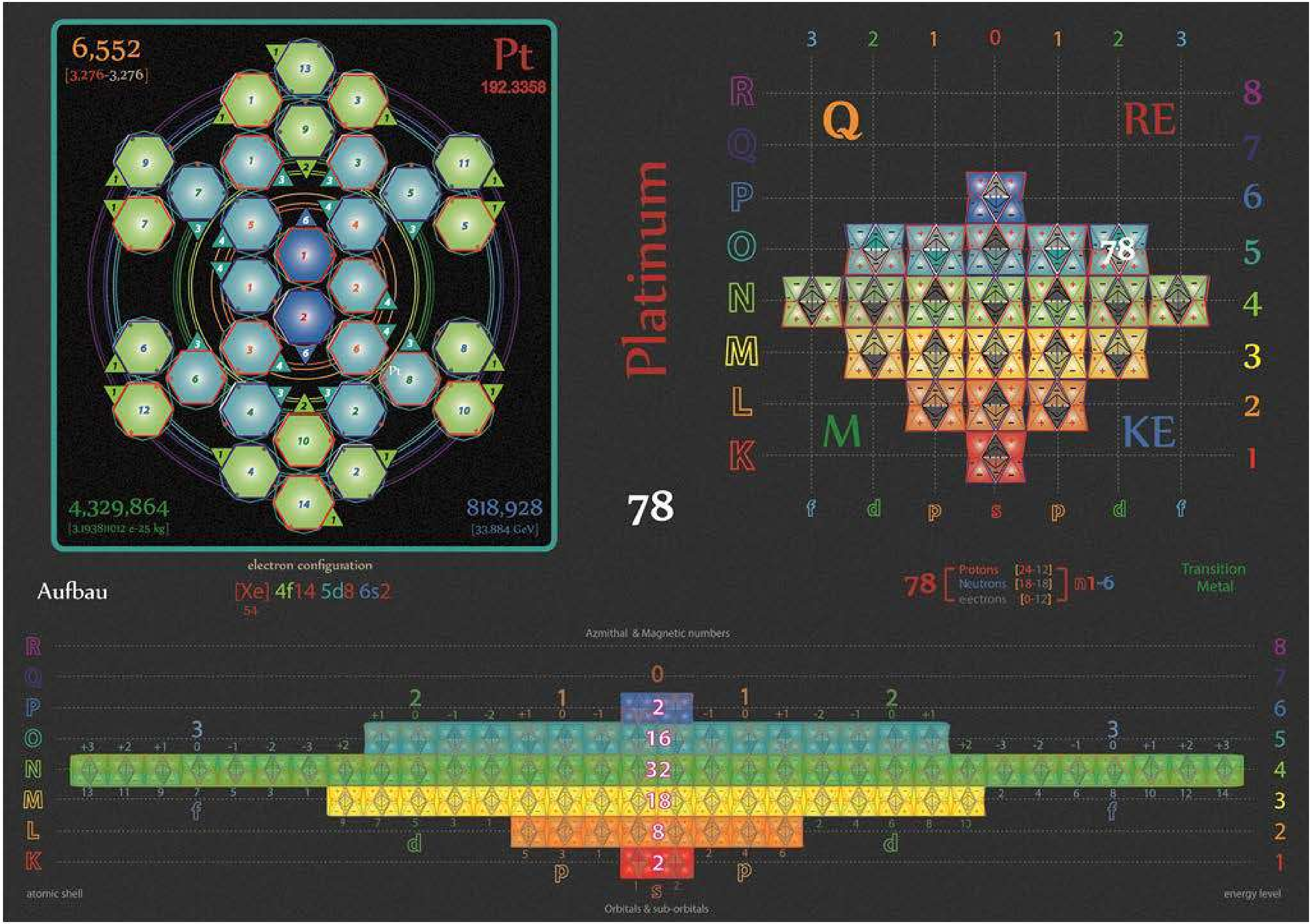
6,552
[3,276-3,276]
Pt
192.3358
4,329,864
[3.19381012 e-25 kg]
818,928
[33.884 GeV]
Platinum
78
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d8 6s2
54
78 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Transition Metal
Azmithal & Magnetic numbers
2
16
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.78 - Platinum atom

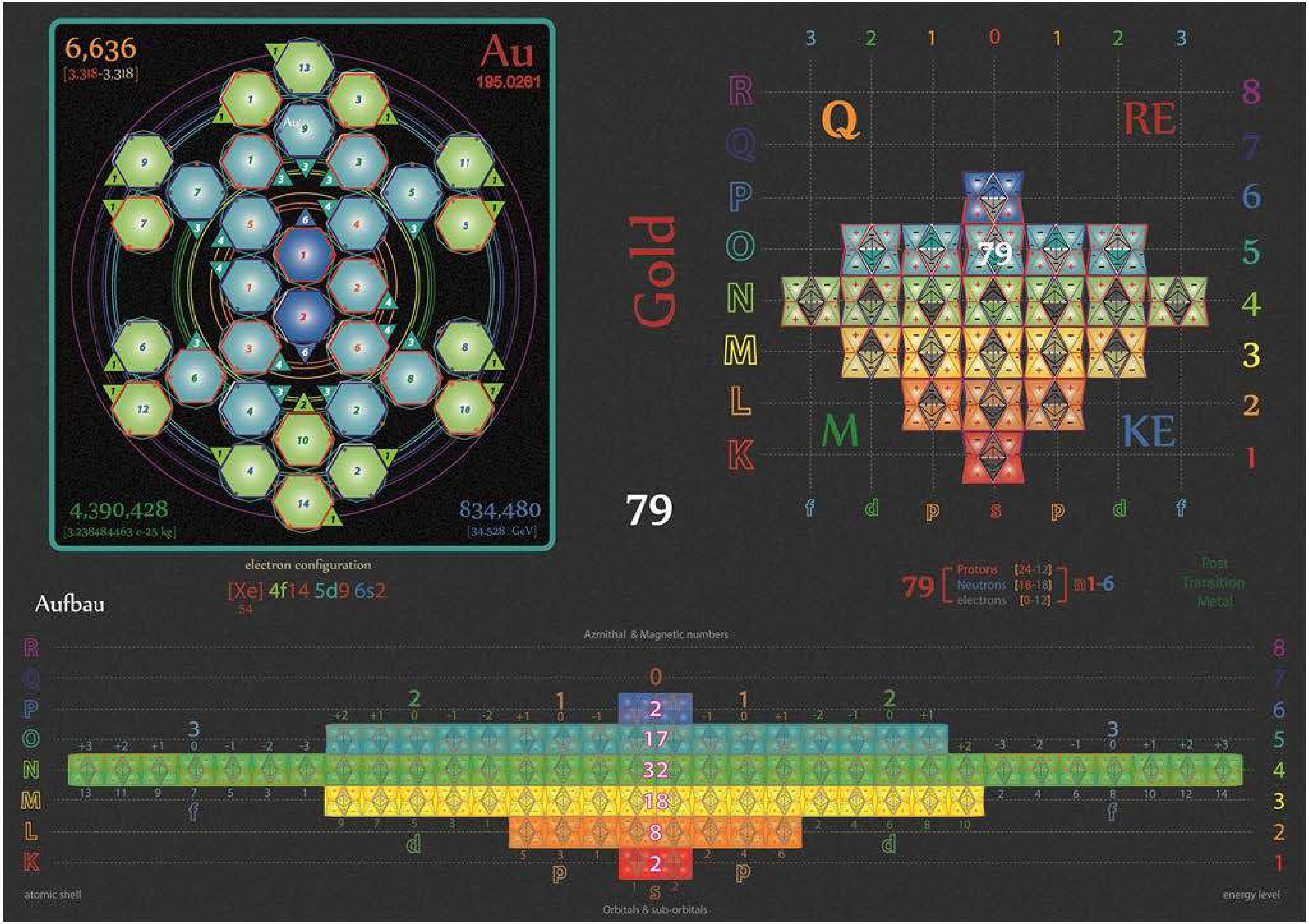
6,636
[3.318-3.318]
Au
196.0261
4,390,428
[3.238484463 e-25 kg]
834,480
[34.528 GeV]
Gold
79
electron configuration
[Xe] 4f14 5d9 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
79
f d p s p d f
79 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Post Transition Metal
Azmithal & Magnetic numbers
0
2
1
3
2
17
32
18
8
2
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.79 - Gold atom

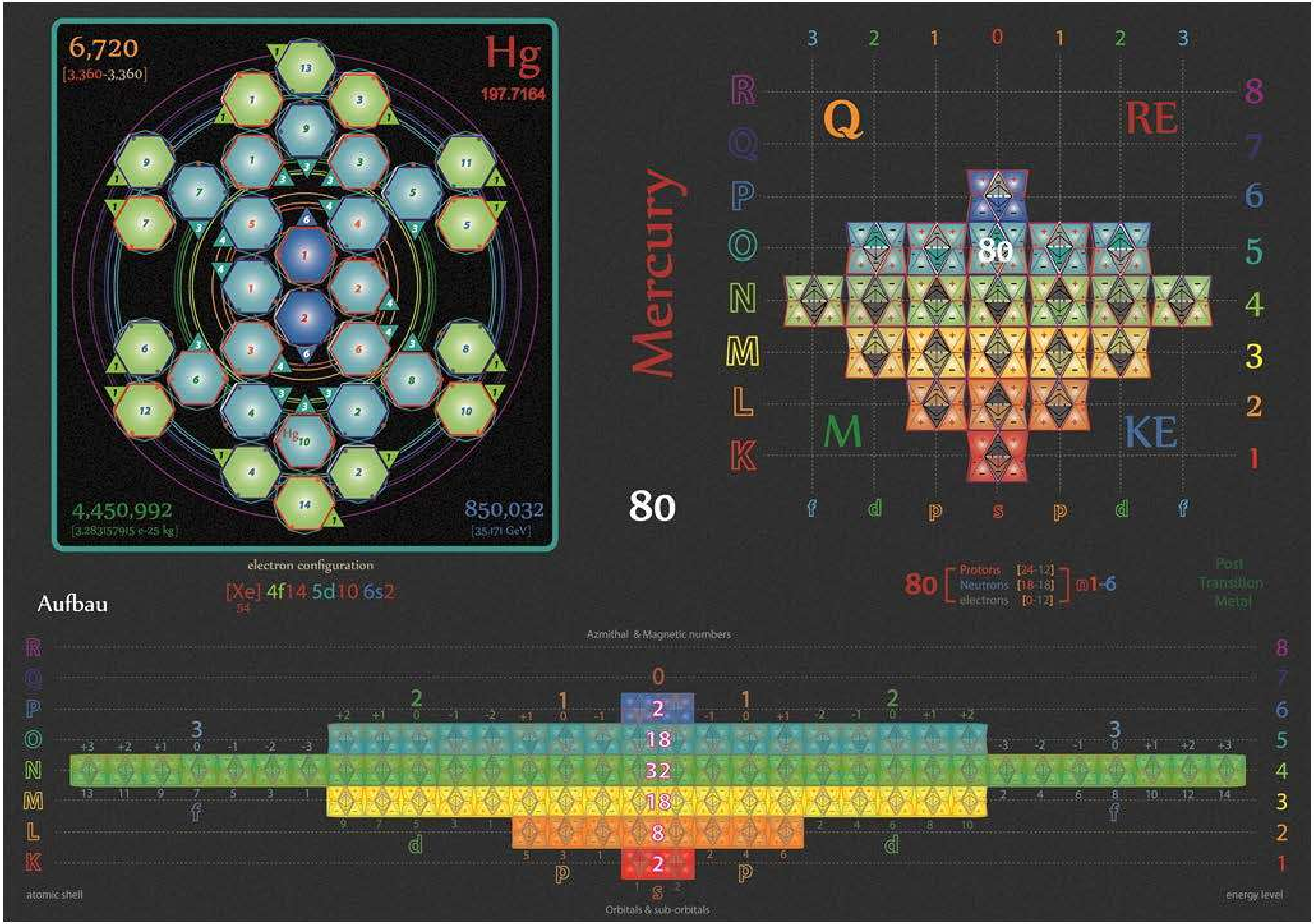
6,720
[3,360-3,360]
Hg
197.7164
4,450,992
[3.28315791 e-25 kg]
850,032
[35.171 GeV]
Mercury
80
electron configuration
[Xe] 4f14 5d10 6s2
54
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
80
80 Protons [24-12] Neutrons [18-18] electrons [0-12] @1-6
Post Transition Metal
Azmithal & Magnetic numbers
2
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.80 - Mercury atom

Copyright ABRAHAM [2008] - All rights reserved

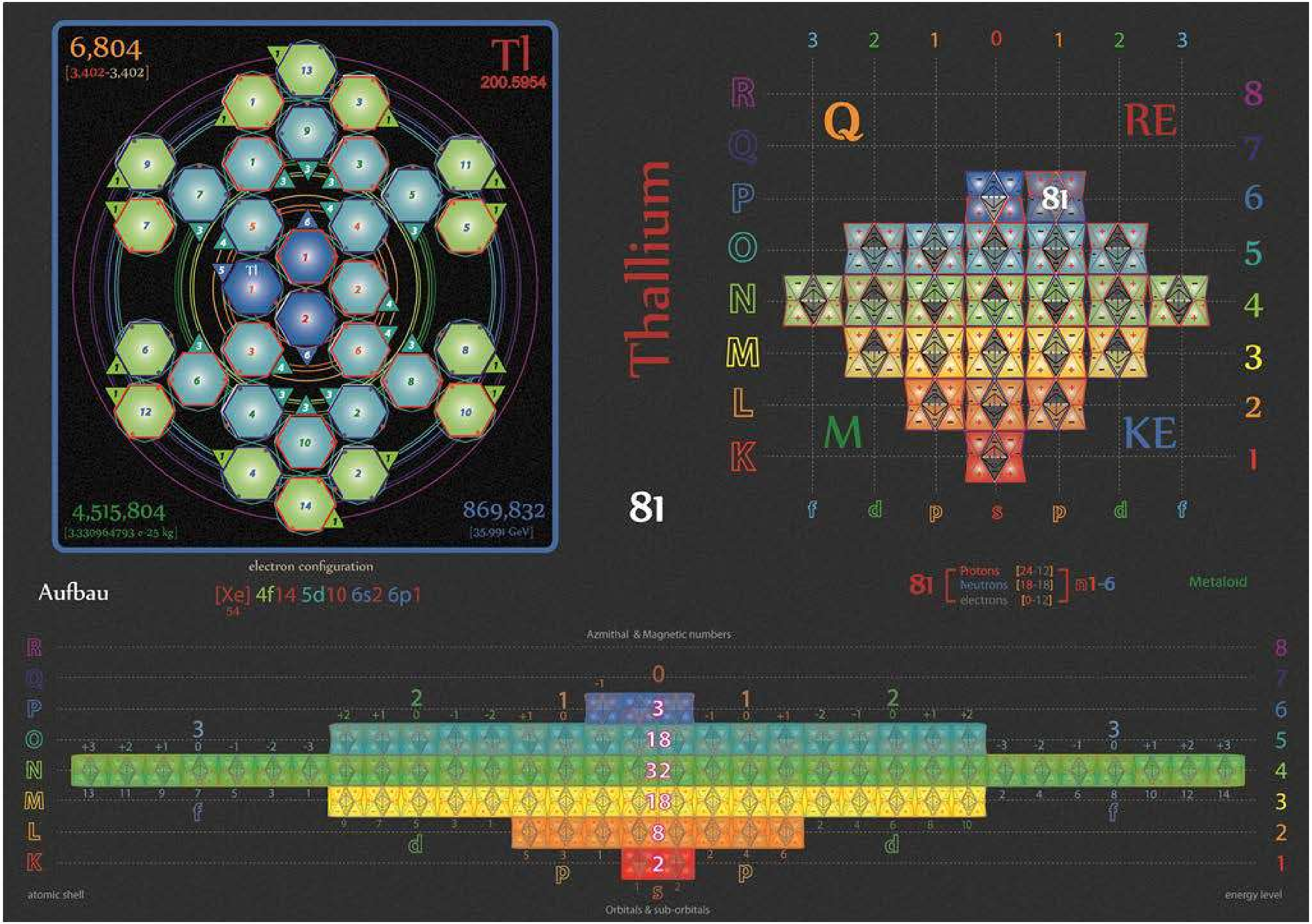
6,804
[3.402-3,402]
Tl
200.5954
4,515,804
[3.330964793 e-25 kg]
869,832
[35.991 GeV]
Thallium
81
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p1
54
81 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Metaloid
Azmithal & Magnetic numbers
3
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.81 - Thallium atom

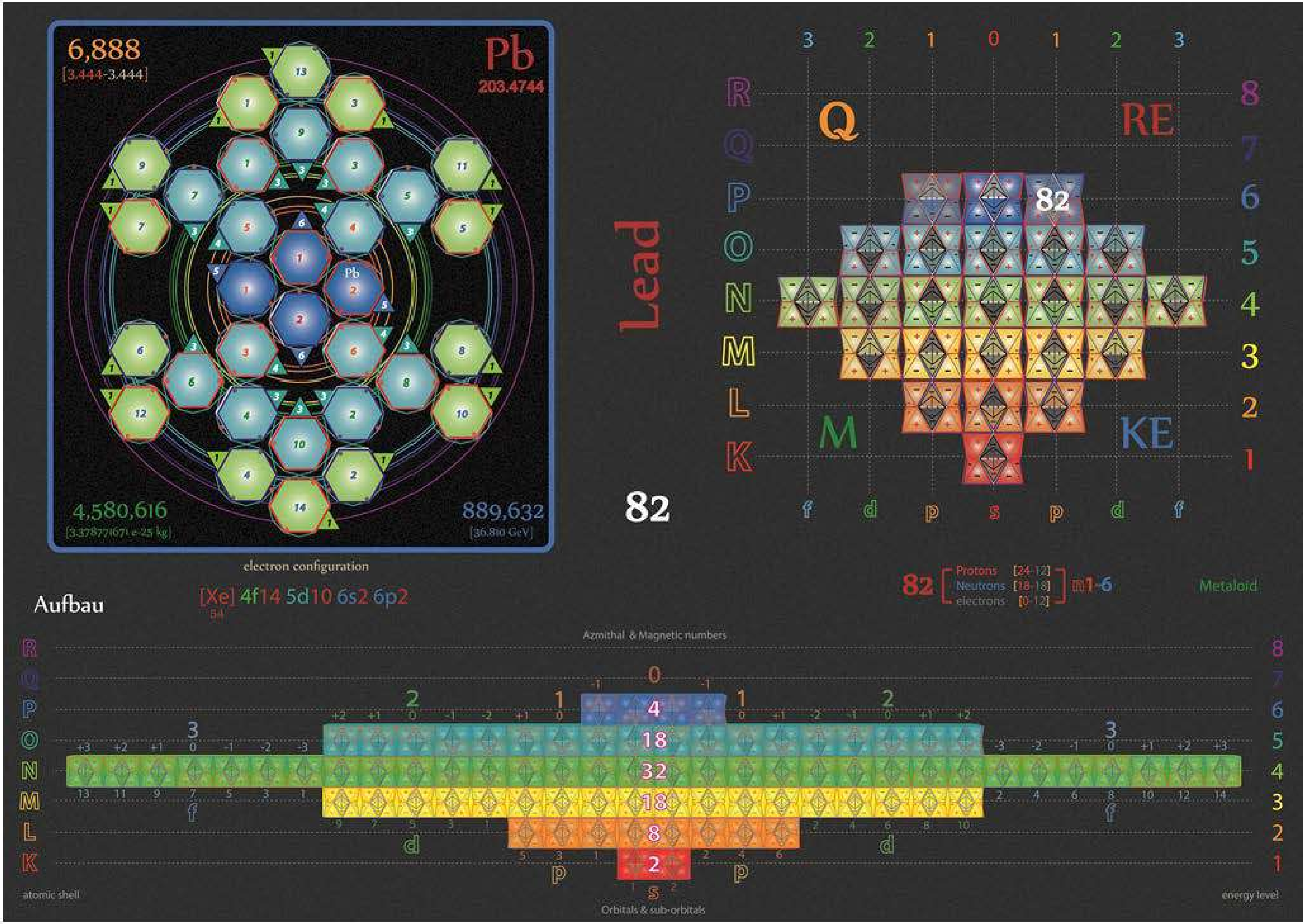
6,888
[3.444-3.444]
Pb
203.4744
4,580,616
[3.37877167 e-25 kg]
889,632
[36.810 GeV]
electron configuration
Lead
82
Aufbau
[Xe] 4f14 5d10 6s2 6p2
54
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
82
f d p s p d f
82
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Metaloid
Azmithal & Magnetic numbers
0
4
18
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.82 - Lead atom

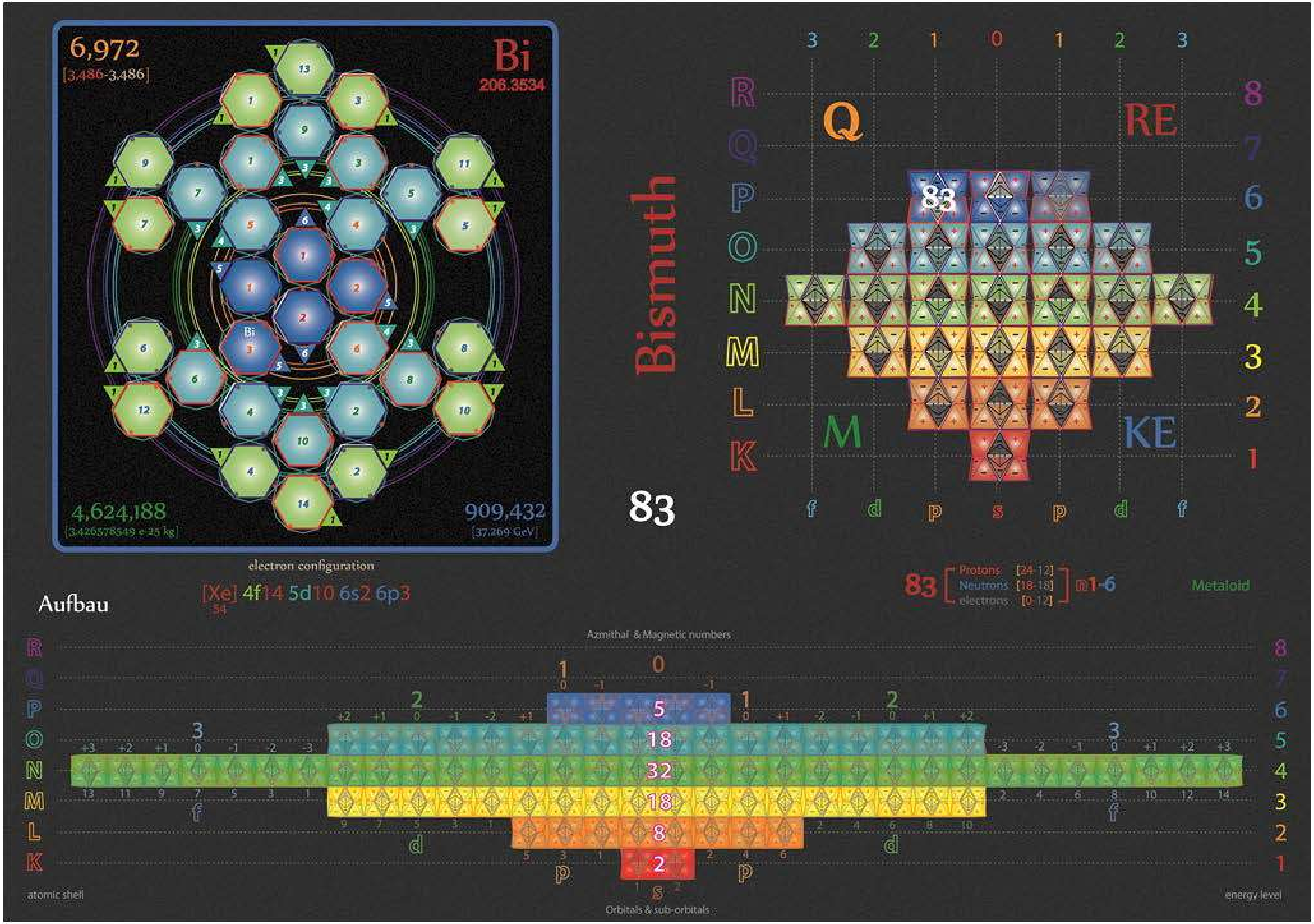
6,972
[3,486-3,486]
Bi
208.3534
4,624,188
[3.426578549 e-25 kg]
909,432
[37.269 GeV]
Bismuth
83
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f14 5d10 6s2 6p3
54
Aufbau
83 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Metaloid
Azmithal & Magnetic numbers
5
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.83 - Bismuth atom

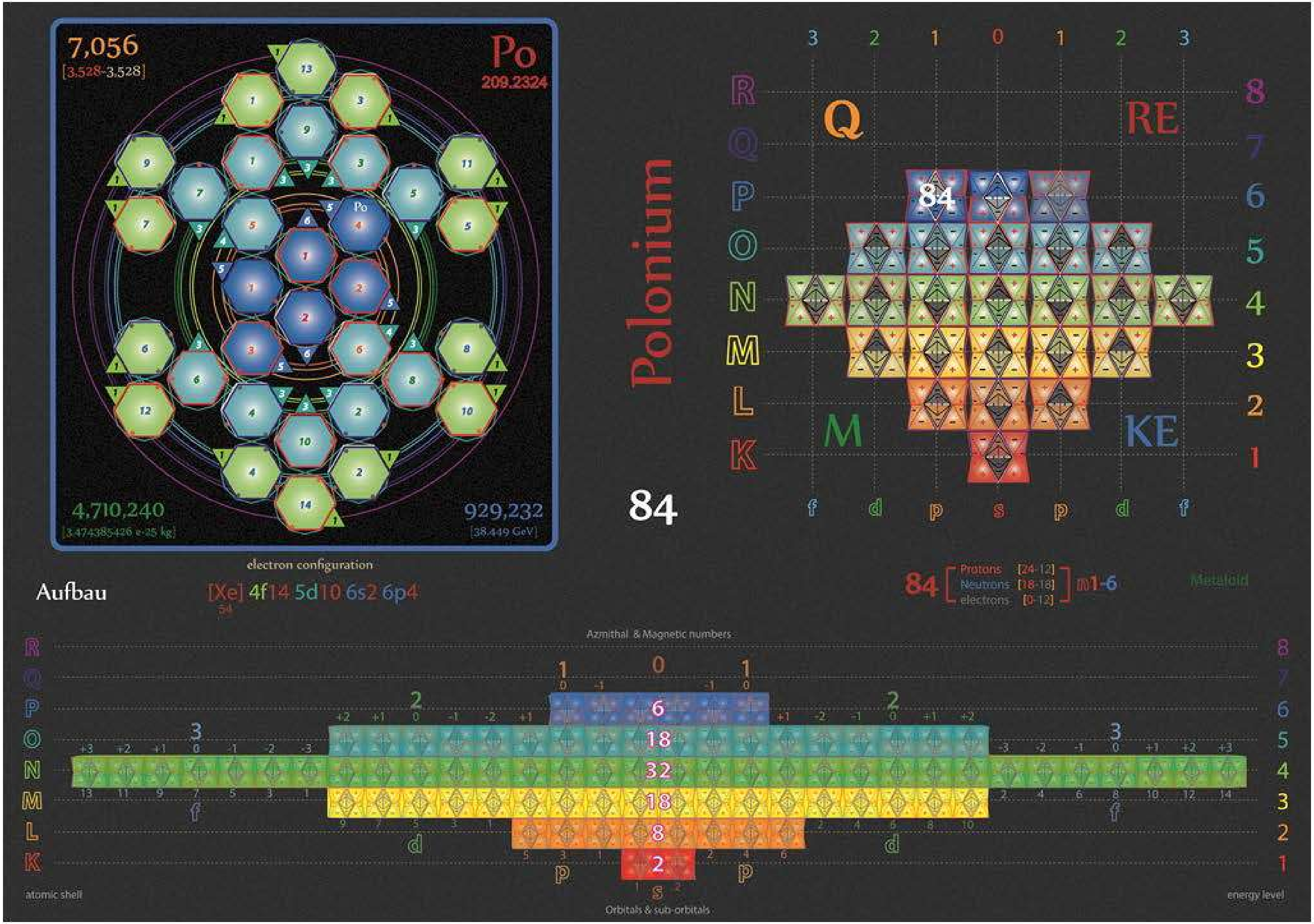
7,056
[3,528-3,528]
Po
209.2324
4,710,240
[3.474385426 e-25 kg]
929,232
[38.449 GeV]
Polonium
84
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p4
54
84
Protons [24-12]
Neutrons [28-18]
electrons [0-12]
m1-6
Metaloid
Azmithal & Magnetic numbers
6
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.84 - Polonium atom

Copyright ABRAHAM [2008] - All rights reserved

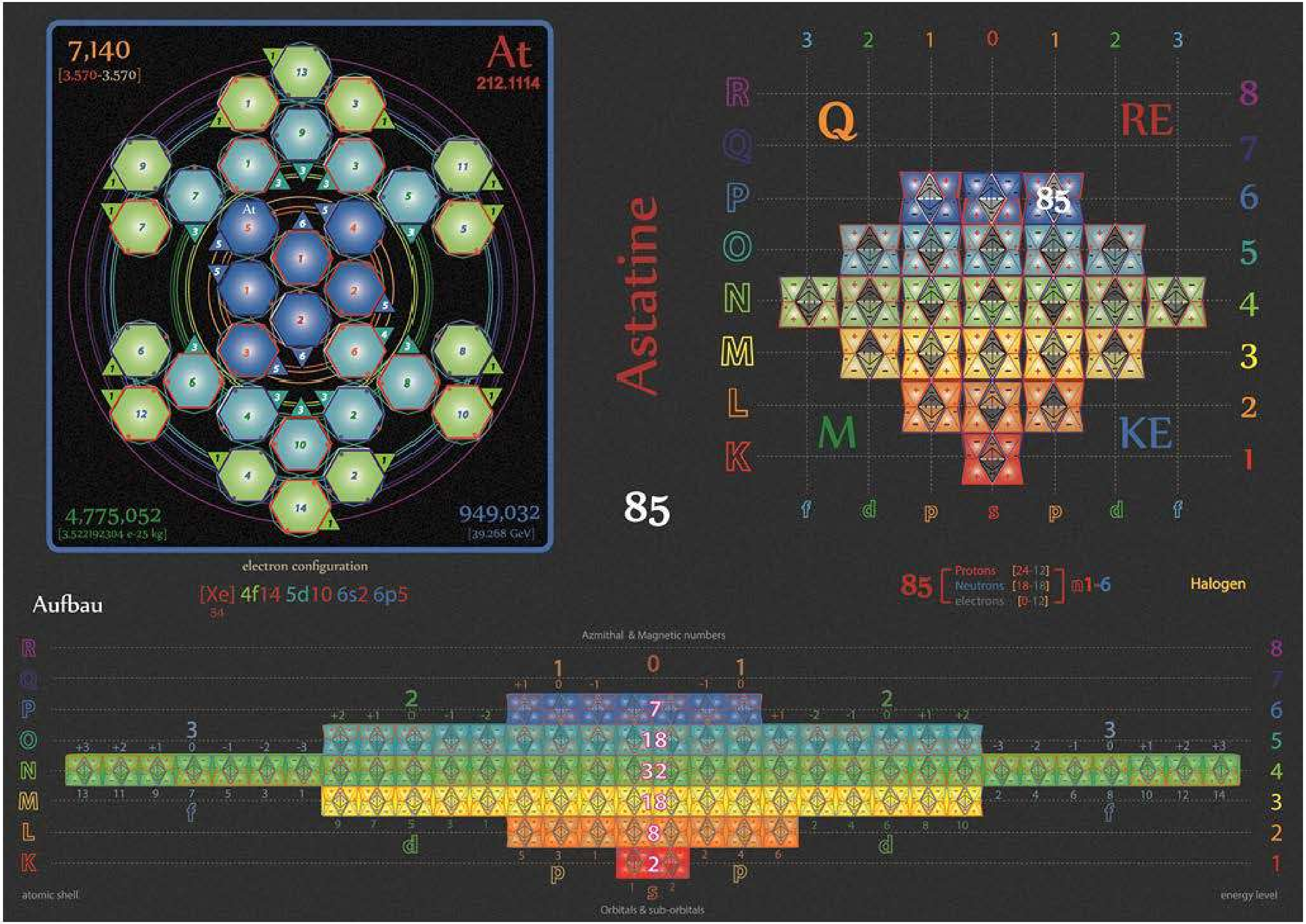
7,140
[3.570-3.570]
At
212.1114
4,775,052
[3.522192304 e-25 kg]
949,032
[39.268 GeV]
Astatine
85
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p5
54
85 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Halogen
Azmithal & Magnetic numbers
7
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.85 - Astatine atom

Copyright ABRAHAM [2008] - All rights reserved

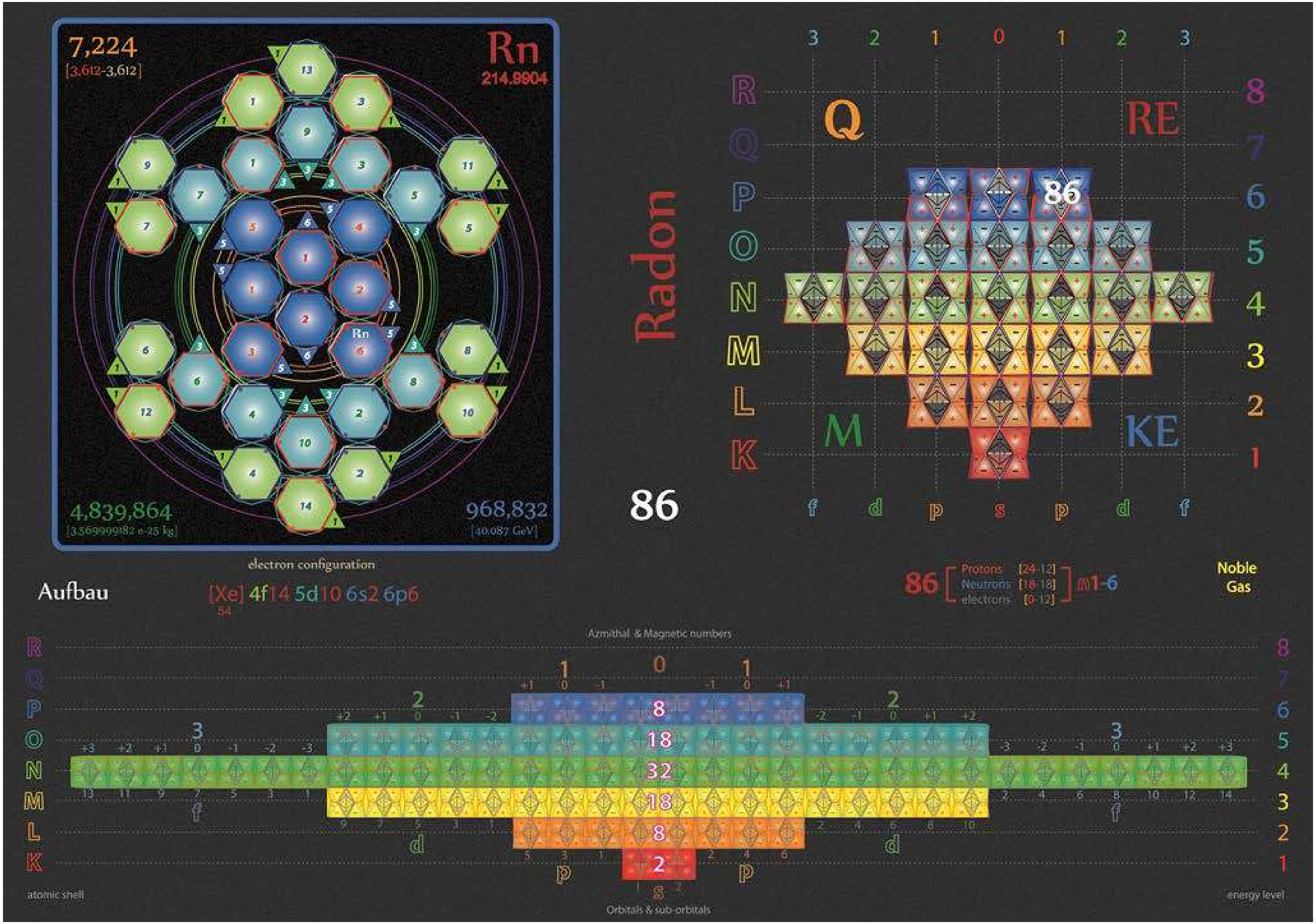
7,224
[3.612-3,612]
Rn
214.9904
4,839,864
[3.569999182 e-25 kg]
968,832
[40.087 GeV]
Radon
86
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
86
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p6
54
86 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Noble Gas
Azmithal & Magnetic numbers
8
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals
s p d f

Tetryonics 51.86 - Radon atom

Copyright ABRAHAM [2008] - All rights reserved

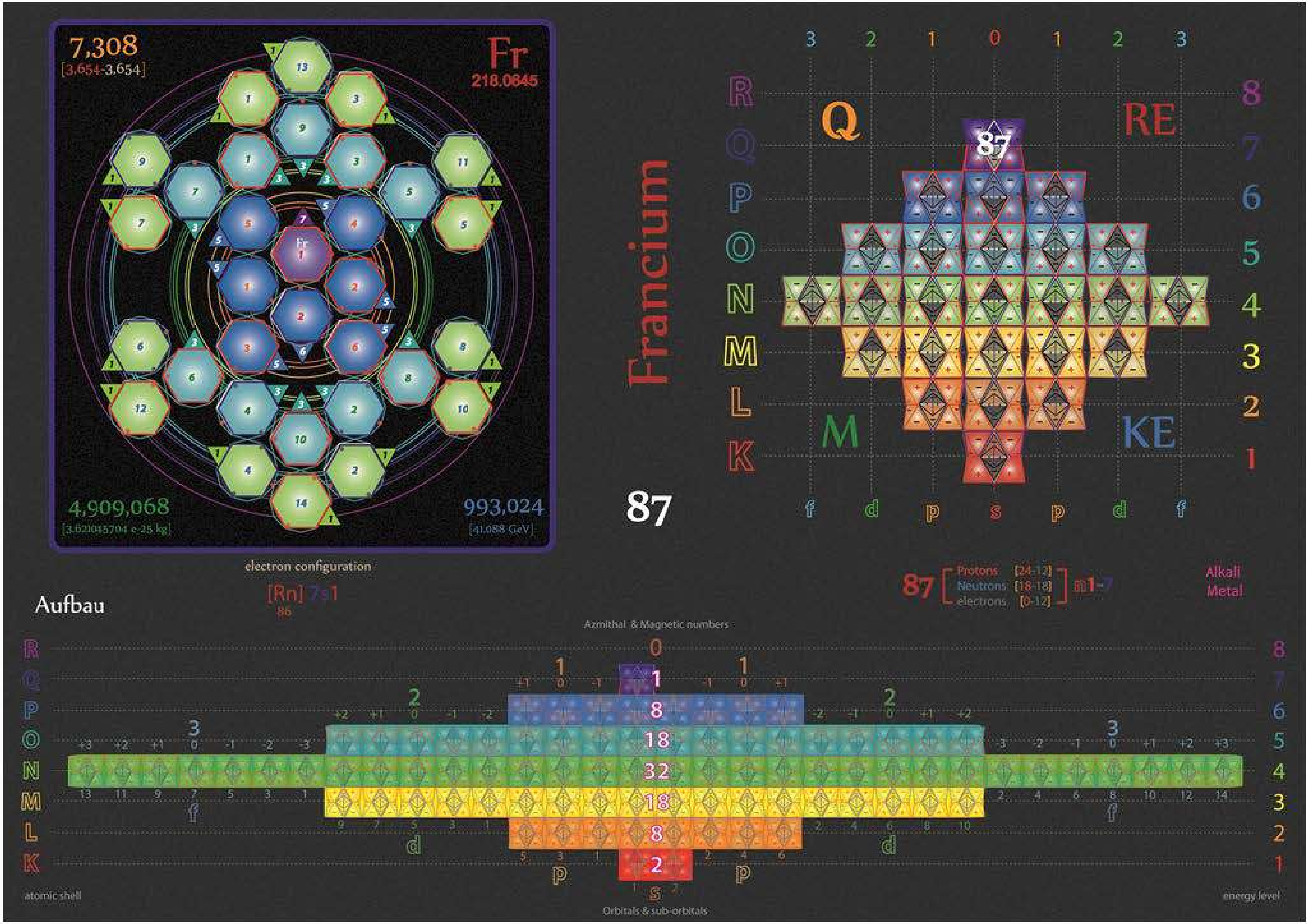
7,308
[3,654-3,654]
Fr
218.0845
4,909,068
[3.621045704 e-25 kg]
993,024
[41.088 GeV]
Francium
87
electron configuration
[Rn] 7s1
86
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
87
f d p s p d f
87
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Alkali
Metal
Aufbau
Azimthal & Magnetic numbers
0
1
1
2
2
3
3
1
8
18
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.87 - Francium atom

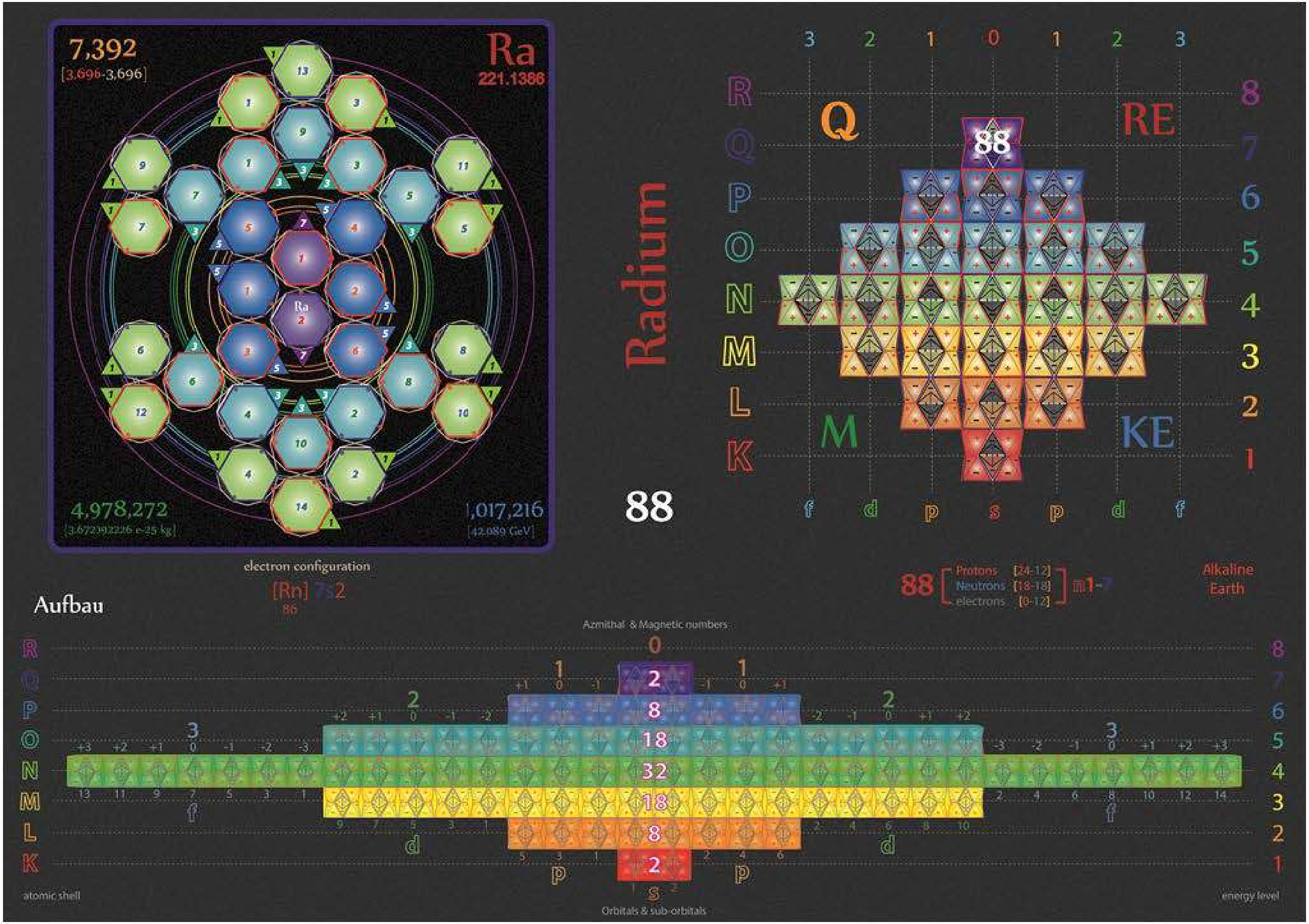
7,392
[3,696-3,696]
Ra
221.1386
4,978,272
[3.672392226 e-25 kg]
1,017,216
[42.089 GeV]
Ra
2
Radium
88
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
88
f d p s p d f
electron configuration
[Rn] 7s2
86
88
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Alkaline
Earth
Aufbau
Azimthal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 0 +1 +2 +3
2
8
18
32
18
8
2
13 11 9 7 5 3 1 2 4 6 8 10 12 14
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.88 - Radium atom

Copyright ABRAHAM [2008] - All rights reserved

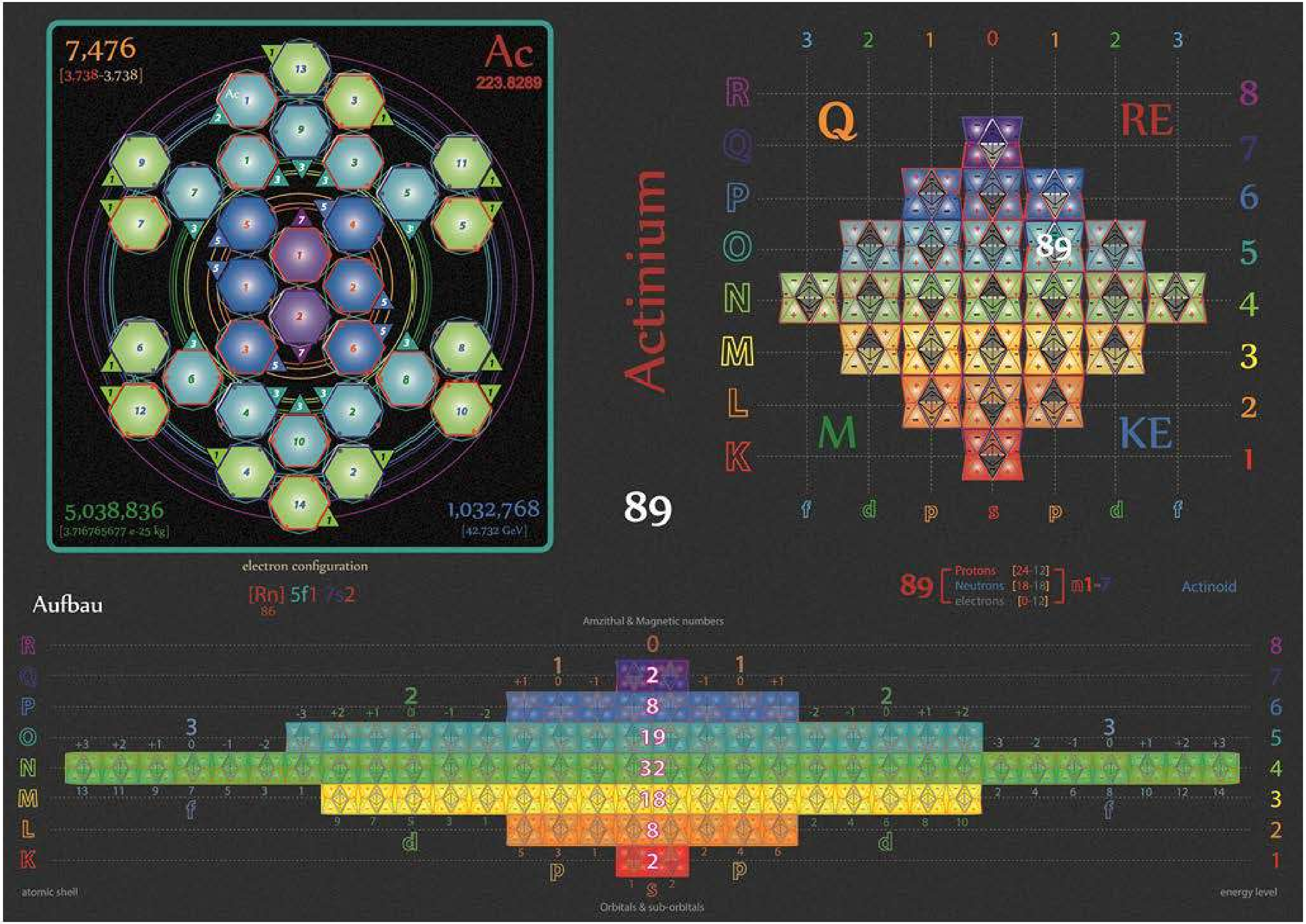
7,476
[3,738-3,738]
Ac
223.8289
5,038,836
[3.716765677 e-25 kg]
1,032,768
[42.732 GeV]
Actinium
89
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
[Rn] 5f1 7s2
86
89 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Aufbau
Amzithal & Magnetic numbers
0
2
8
19
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals
s
p
d
f

Tetryonics 51.89 - Actinium atom

Copyright ABRAHAM [2008] - All rights reserved

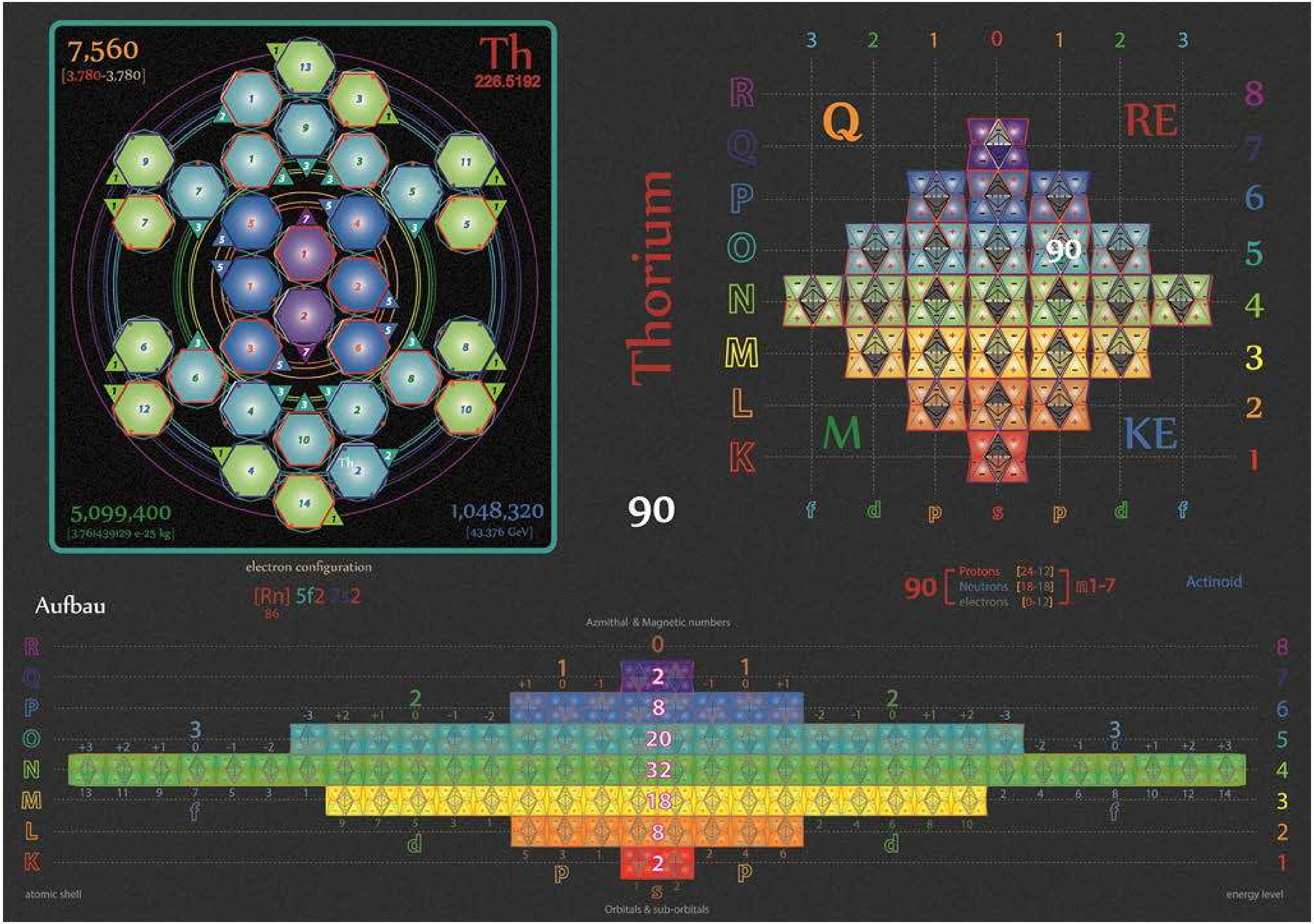
7,560
[3,780-3,780]
Th
226.5192
5,099,400
[3.761439129 e-25 kg]
1,048,320
[43.376 GeV]
Thorium
90
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
90
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
[Rn] 5f2 7s2
86
90
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Aufbau
Azmithal & Magnetic numbers
0
2
8
20
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.90 - Thorium atom

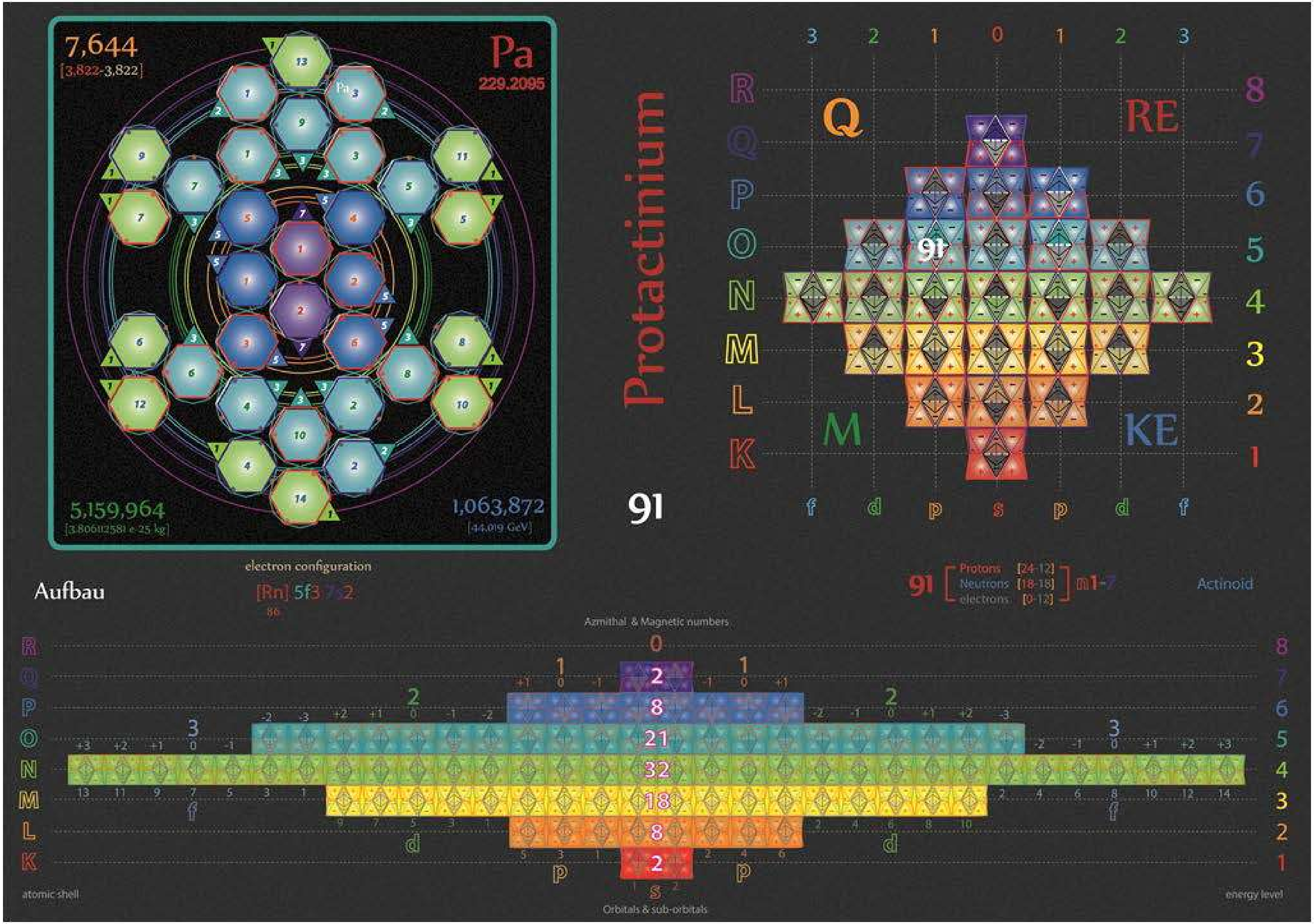
7,644
[3,822-3,822]
Pa
229.2095
5,159,964
[3.80610258 e-25 kg]
1,063,872
[44.019 GeV]
Protactinium
91
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
91
f d p s p d f
Aufbau
electron configuration
[Rn] 5f3 7s2
86
91
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
+1 0 -1
-1 0 +1
-2 -3 +2 +1 0 -1 -2
-2 -1 0 +1 +2 -3
+3 +2 +1 0 -1
-2 -1 0 +1 +2 +3
2
8
21
32
18
8
2
13 11 9 7 5 3 1
2 4 6 8 10 12 14
f
f
9 7 5 3 1
2 4 6 8 10
d
d
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.91 - Protactinium atom

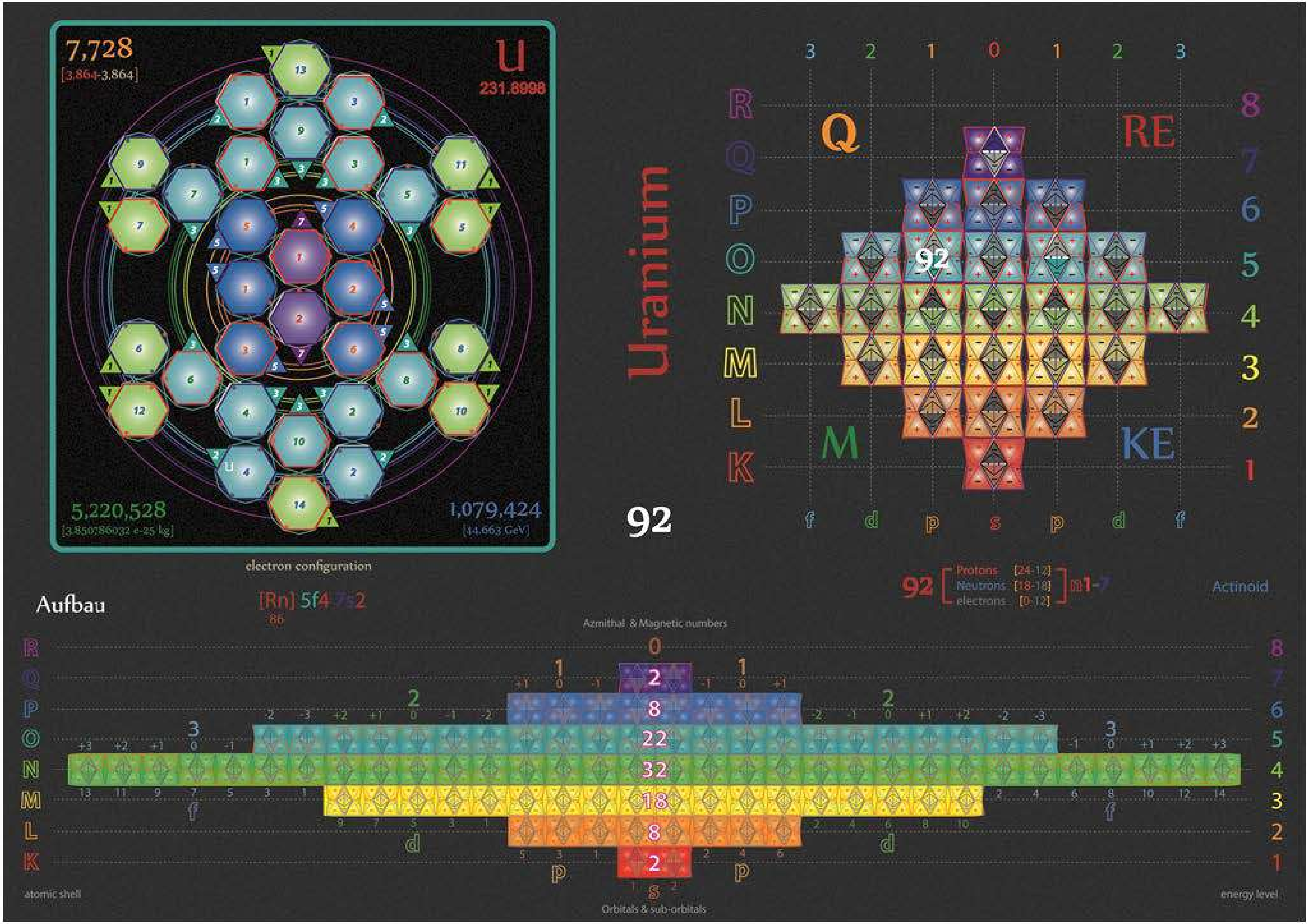
7,728
[3,864-3,864]
U
231.8998
5,220,528
[3.850786032 e-25 kg]
1,079,424
[14.663 GeV]
electron configuration
Uranium
92
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
92
f d p s p d f
92
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Aufbau
[Rn] 5f4 7s2
86
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
8
22
32
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.92 - Uranium atom

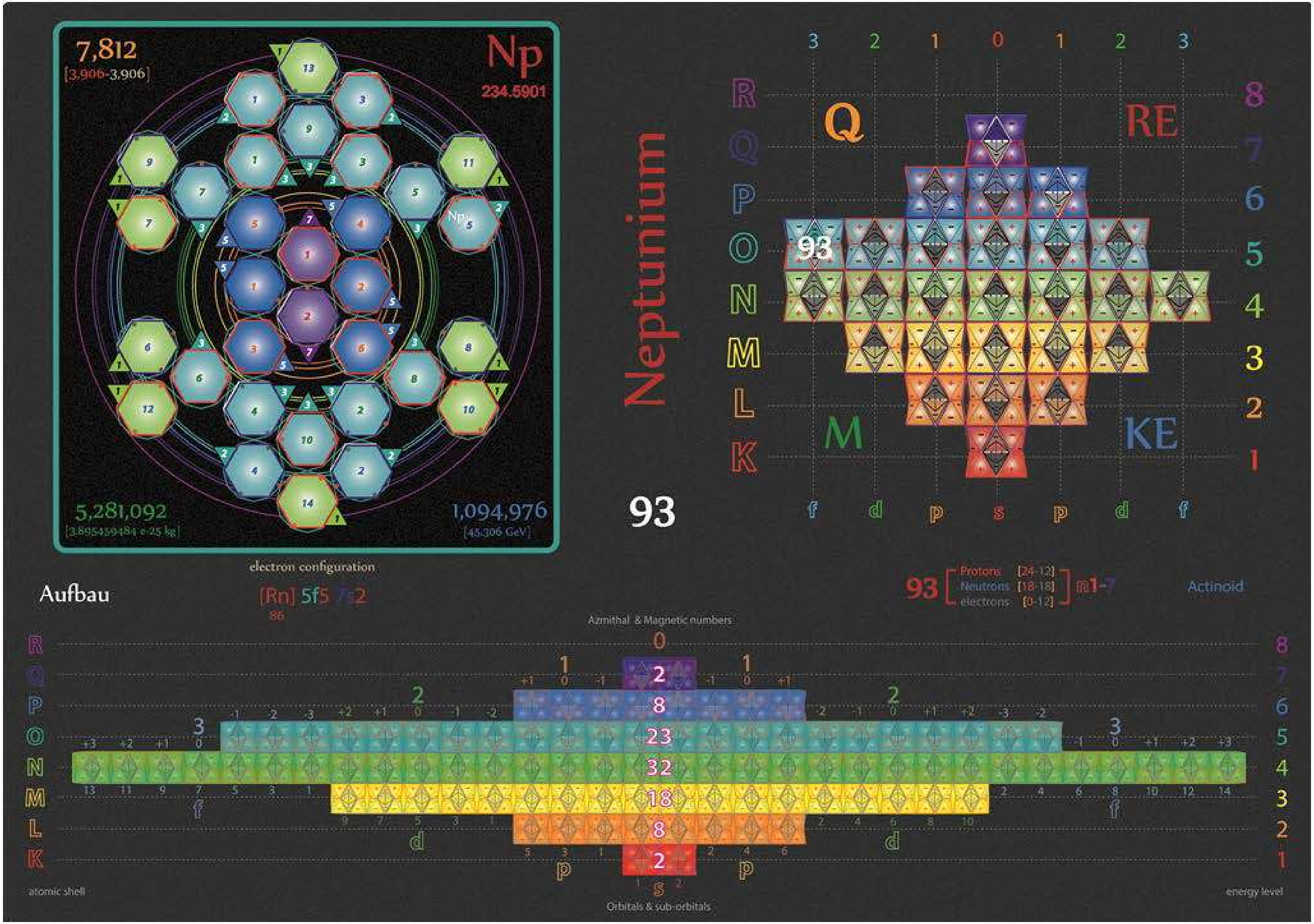
7,812
[3,906-3,906]
Np
234.5901
5,281,092
[3.89545948.4 e-25 kg]
1,094,976
[45.306 GeV]
Neptunium
93
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f5 7s2
86
93
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-7
Actinoid
Azmithal & Magnetic numbers
2
8
23
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.93 - Neptunium atom

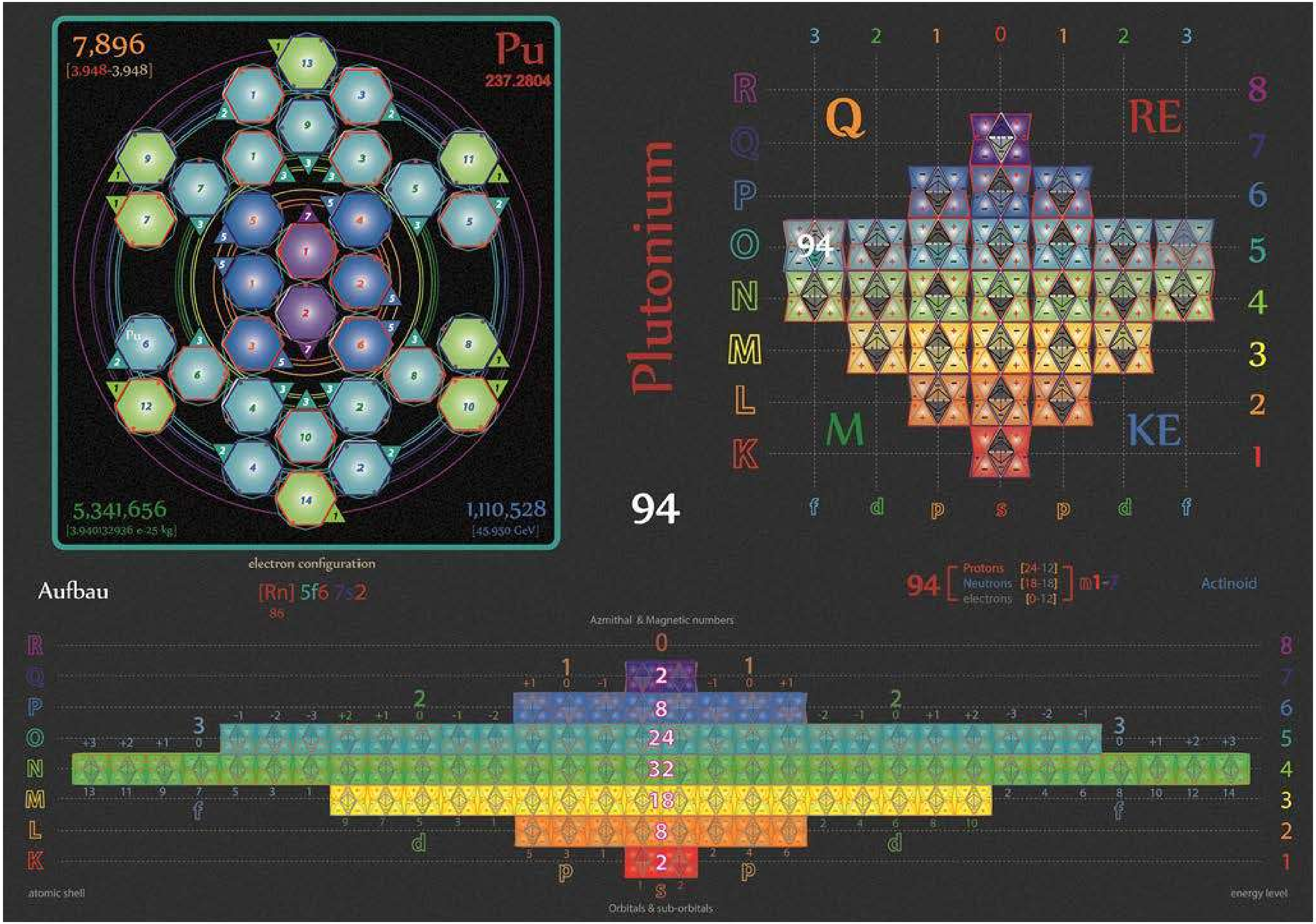
7,896
[3,948-3,948]
Pu
237.2804
5,341,656
[3.940132936 e-25 kg]
1,110,528
[45.950 GeV]
Plutonium
94
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f6 7s2
86
94 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
24
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.94 - Plutonium atom

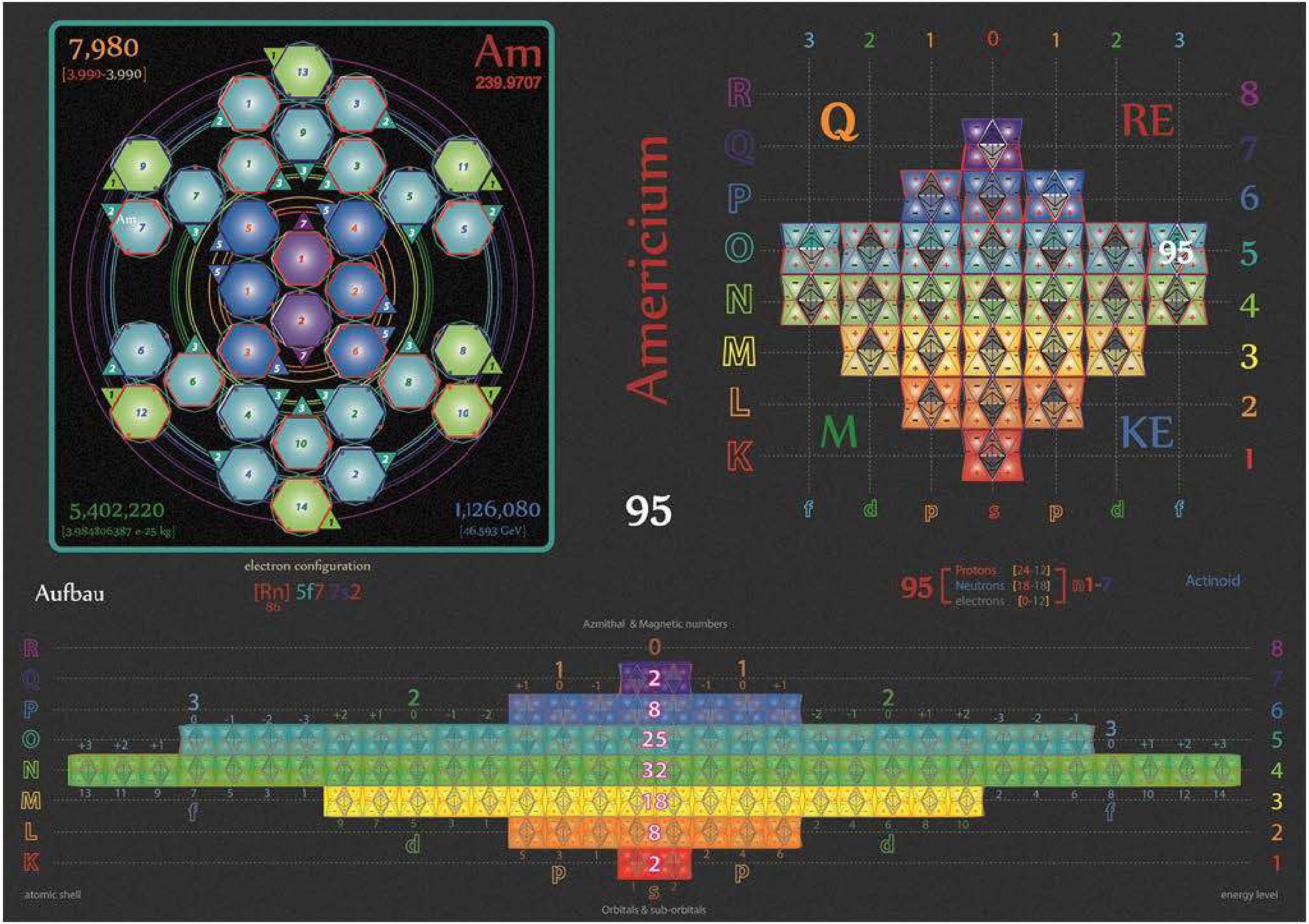
7,980
[3,990-3,990]
Am
239.9707
5,402,220
[3.984806387 e-25 kg]
1,126,080
[46.593 GeV]
Americium
95
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Rn] 5f7 7s2
86
Aufbau
95 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Actinoid
Azmithal & Magnetic numbers
0
2
8
25
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.95 - Americium atom

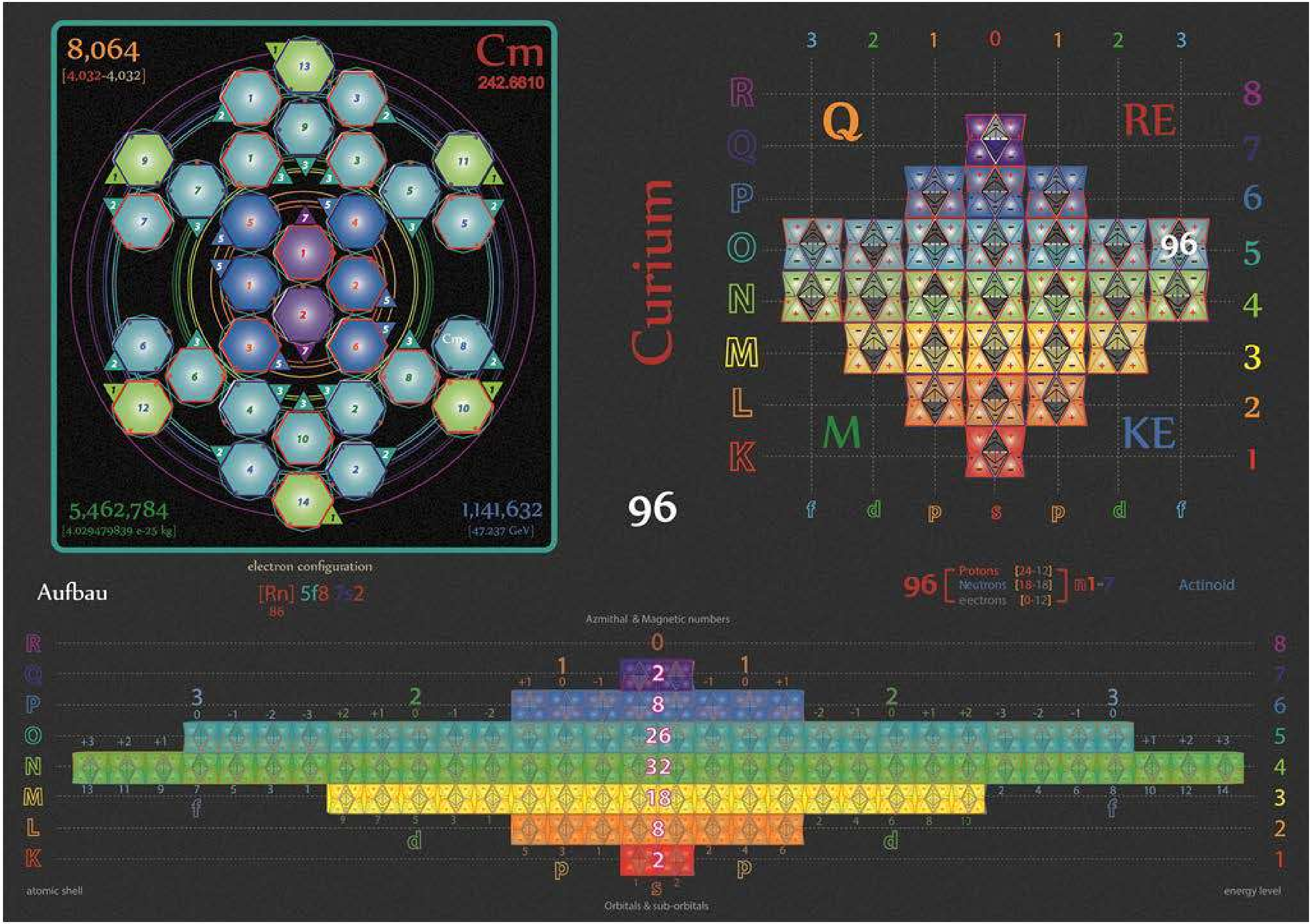
8,064
[4,032-4,032]
Cm
242.6810
5,462,784
[4.029479839 e-25 kg]
1,141,632
[47.237 GeV]
Curium
96
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
96 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Actinoid
Aufbau
electron configuration
[Rn] 5f8 7s2
86
Azmithal & Magnetic numbers
0
2
8
26
32
18
8
2
atomic shell
energy level
s p d f
Orbitals & sub-orbitals

Tetryonics 51.96 - Curium atom

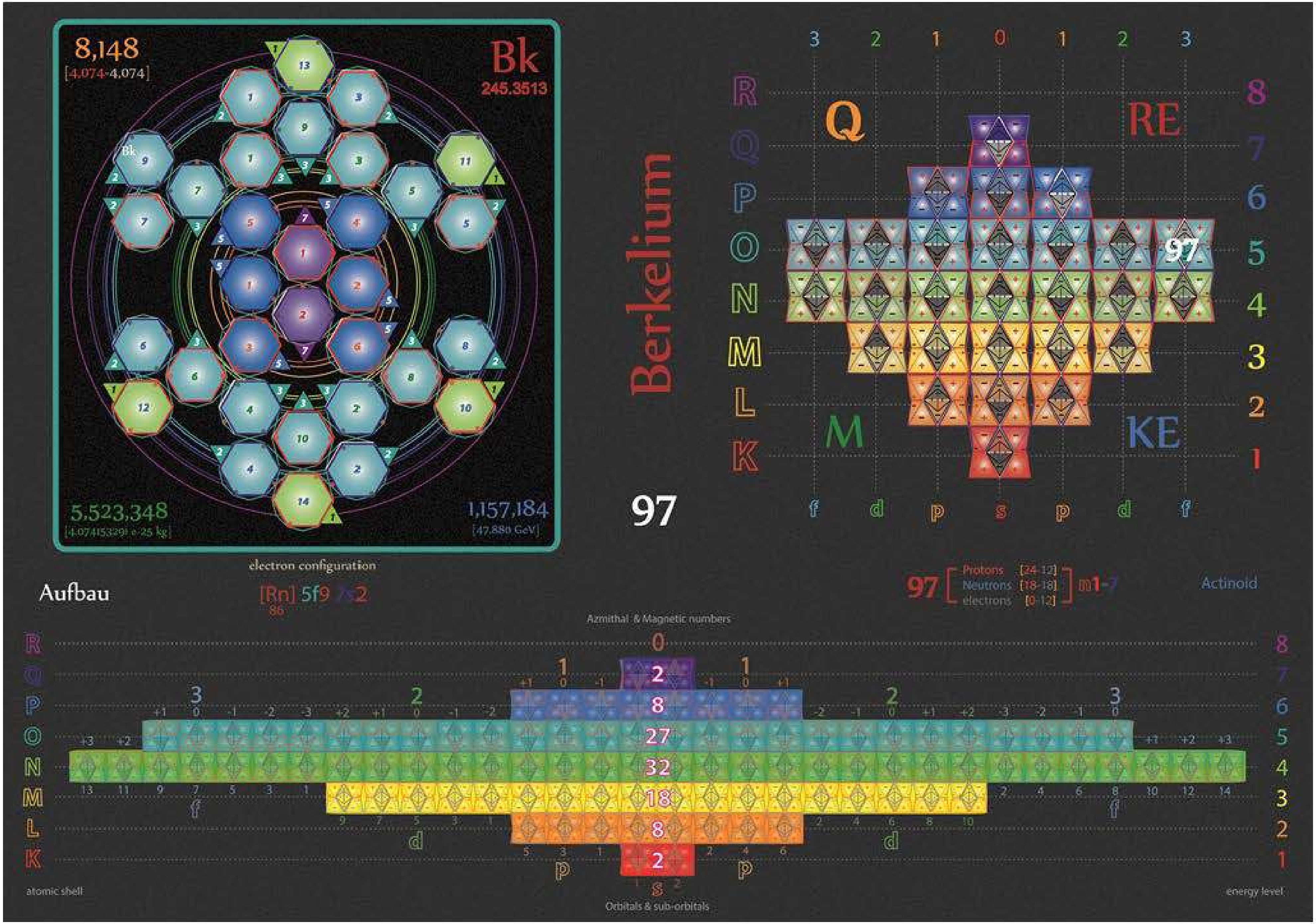
8,148
[4,074-4,074]
Bk
245.3513
Bk
5,523,348
[4.07415329 e-25 kg]
1,157,184
[47,880 GeV]
Berkelium
97
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
97
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f9 7s2
86
97
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
2
8
27
32
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.97 - Berkelium atom

Copyright ABRAHAM [2008] - All rights reserved

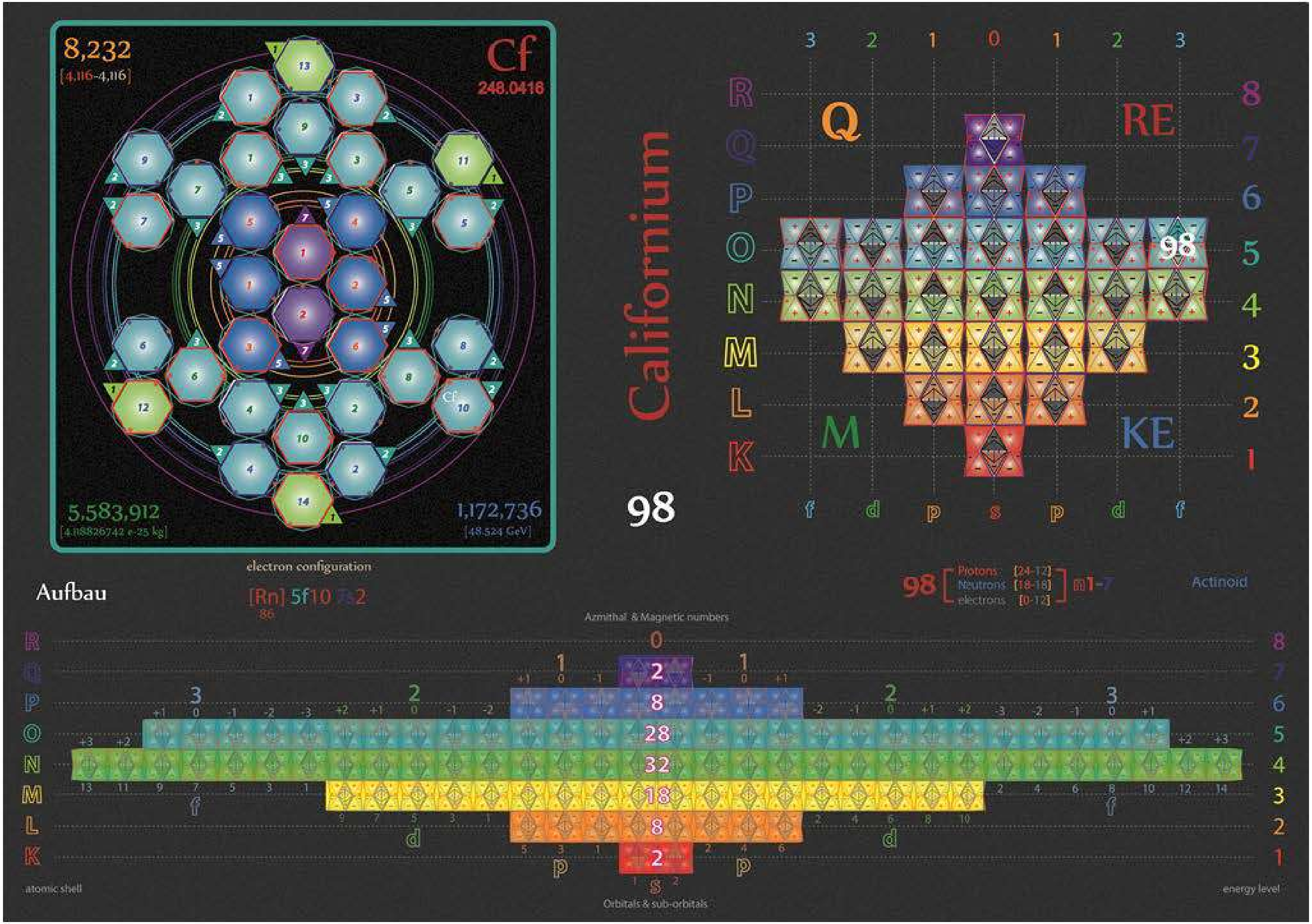
8,232
[4,116-4,116]
Cf
248.0418
5,583,912
[4.118826742 e-25 kg]
1,172,736
[48.524 GeV]
Californium
98
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
98 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Aufbau
electron configuration
[Rn] 5f10 7s2
86
Azmithal & Magnetic numbers
2
8
28
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.98 - Californium atom

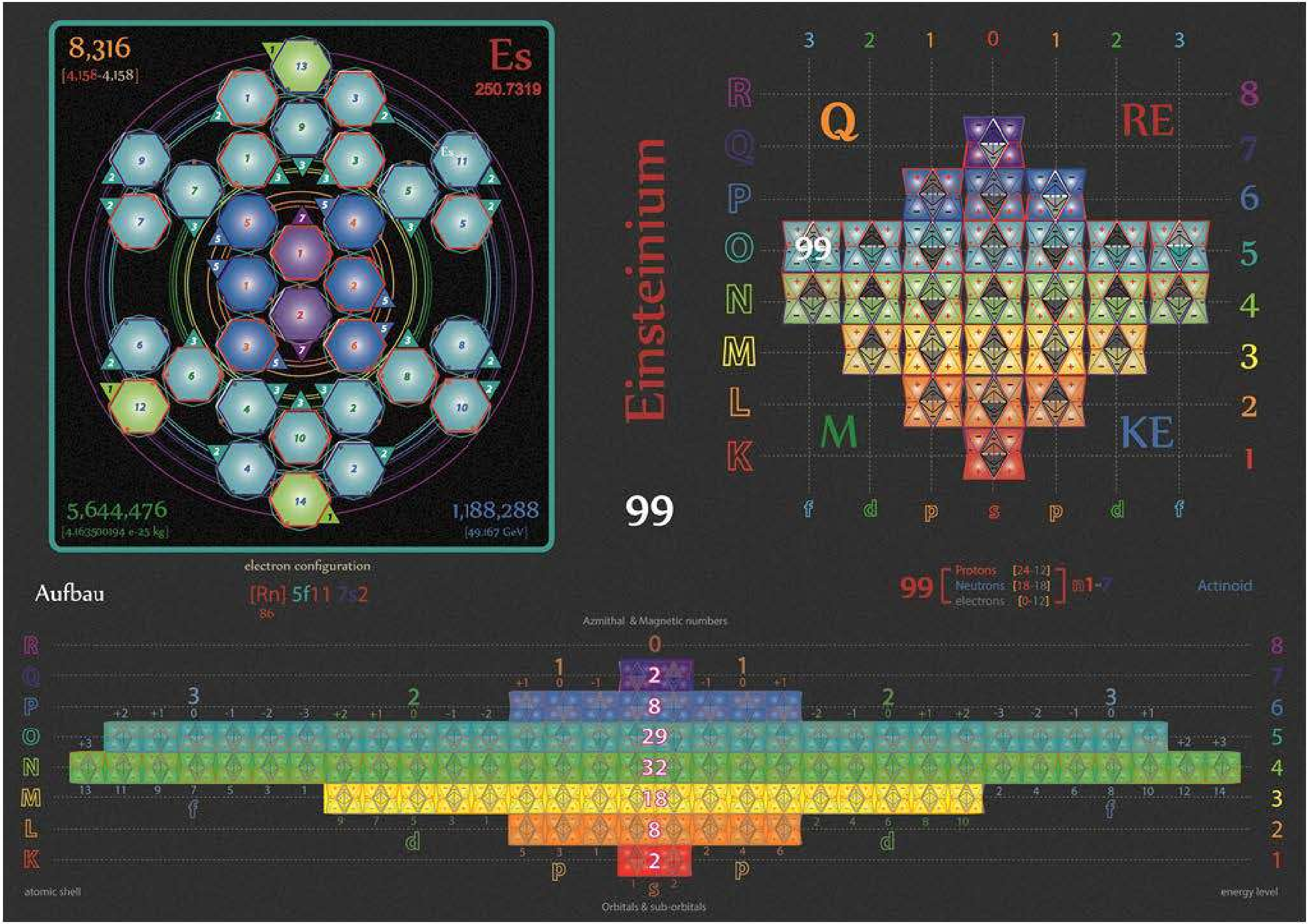
8,316
[4.158-4.158]
Es
250.7319
5,644,476
[4.163500194 e-25 kg]
1,188,288
[49.167 GeV]
Einsteinium
99
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f11 7s2
86
99 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Actinoid
Azmithal & Magnetic numbers
2
8
29
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.99 - Einsteinium atom

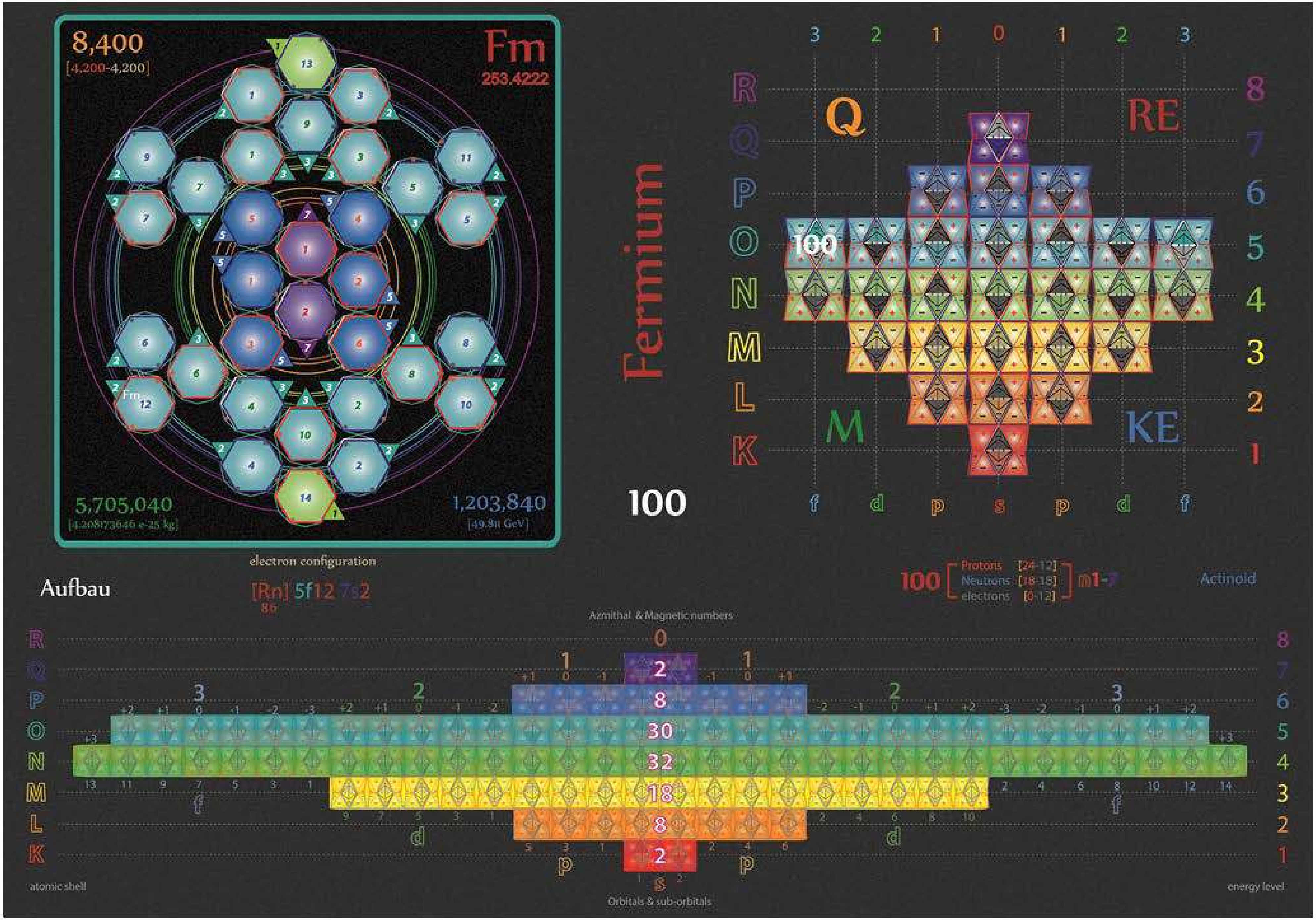
8,400
[4,200-4,200]
Fm
253.4222
5,705,040
[4.208173646 e-25 kg]
1,203,840
[49.811 GeV]
Fermium
100
electron configuration
Aufbau
[Rn] 5f12 7s2
86
100 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Azmithal & Magnetic numbers
2
8
30
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.100 - Fermium atom

Copyright ABRAHAM [2008] - All rights reserved

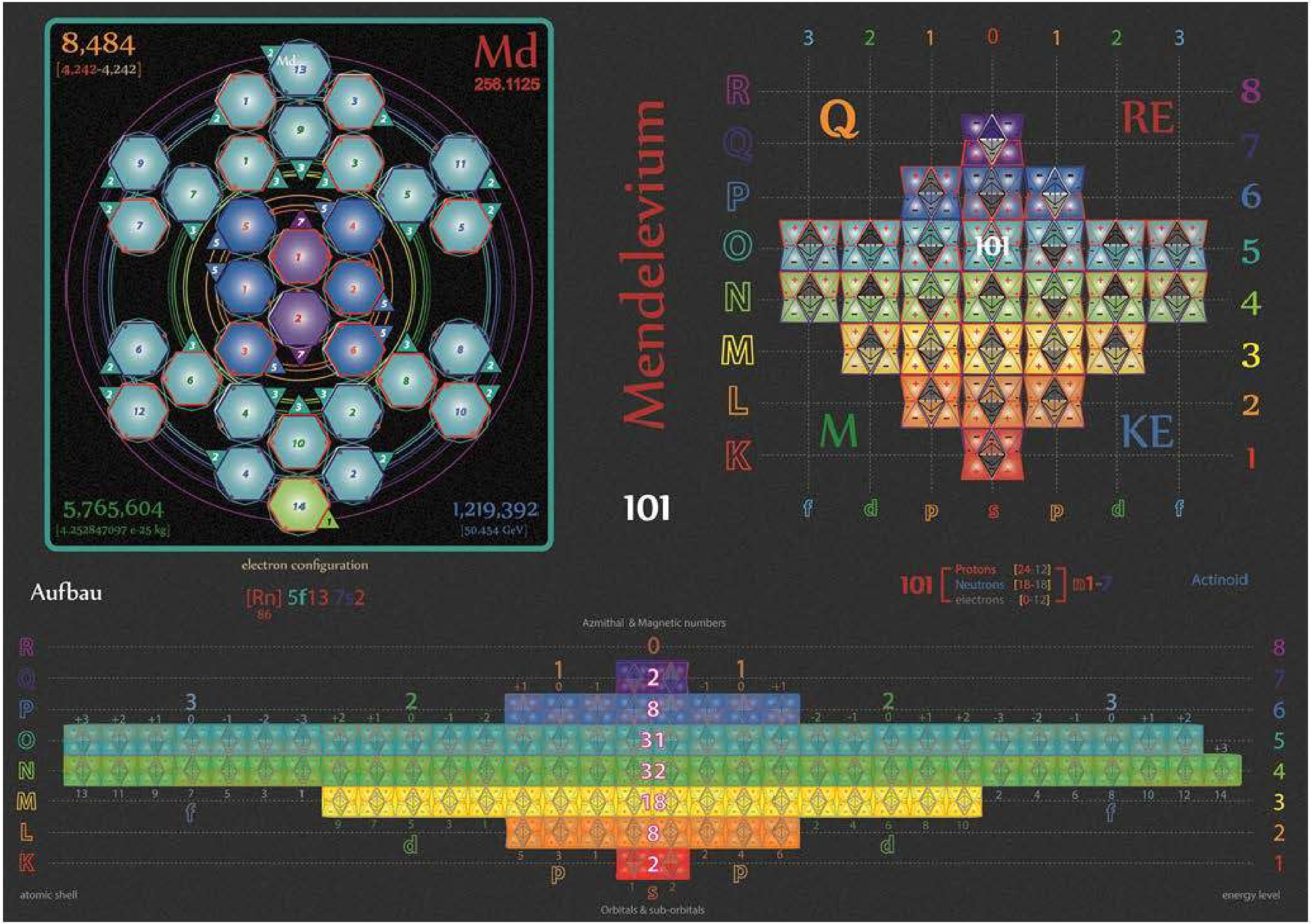
8,484
[4,242-4,242]
Md
258.1125
5,765,604
[4.252847097 e-25 kg]
1,219,392
[50.454 GeV]
Mendelevium
101
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f13 7s2
86
101
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
0
2
8
31
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.101 - Mendelevium atom

Copyright ABRAHAM [2008] - All rights reserved

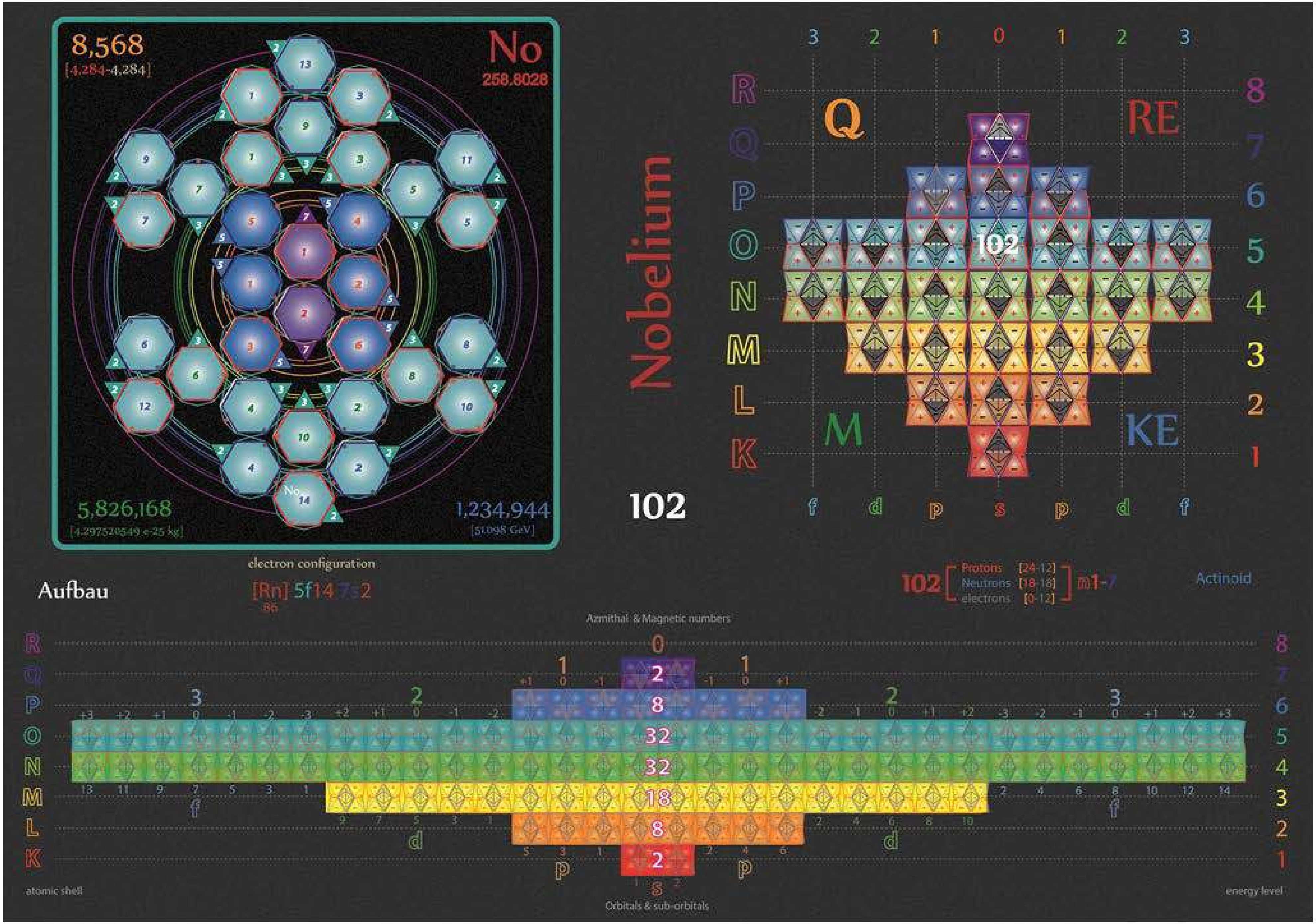
8,568
[4,284-4,284]
No
258.8028
5,826,168
[4.297520549 e-25 kg]
1,234,944
[51.098 GeV]
Nobelium
102
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
102
f d p s p d f
electron configuration
[Rn] 5f14 7s2
86
Aufbau
102 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
+3 +2 +1 0 -1 -2 -3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3 -2 -1 0 +1 +2 +3
2
8
32
32
18
8
2
13 11 9 7 5 3 1 2 4 6 8 10 12 14
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.102 - Nobelium atom

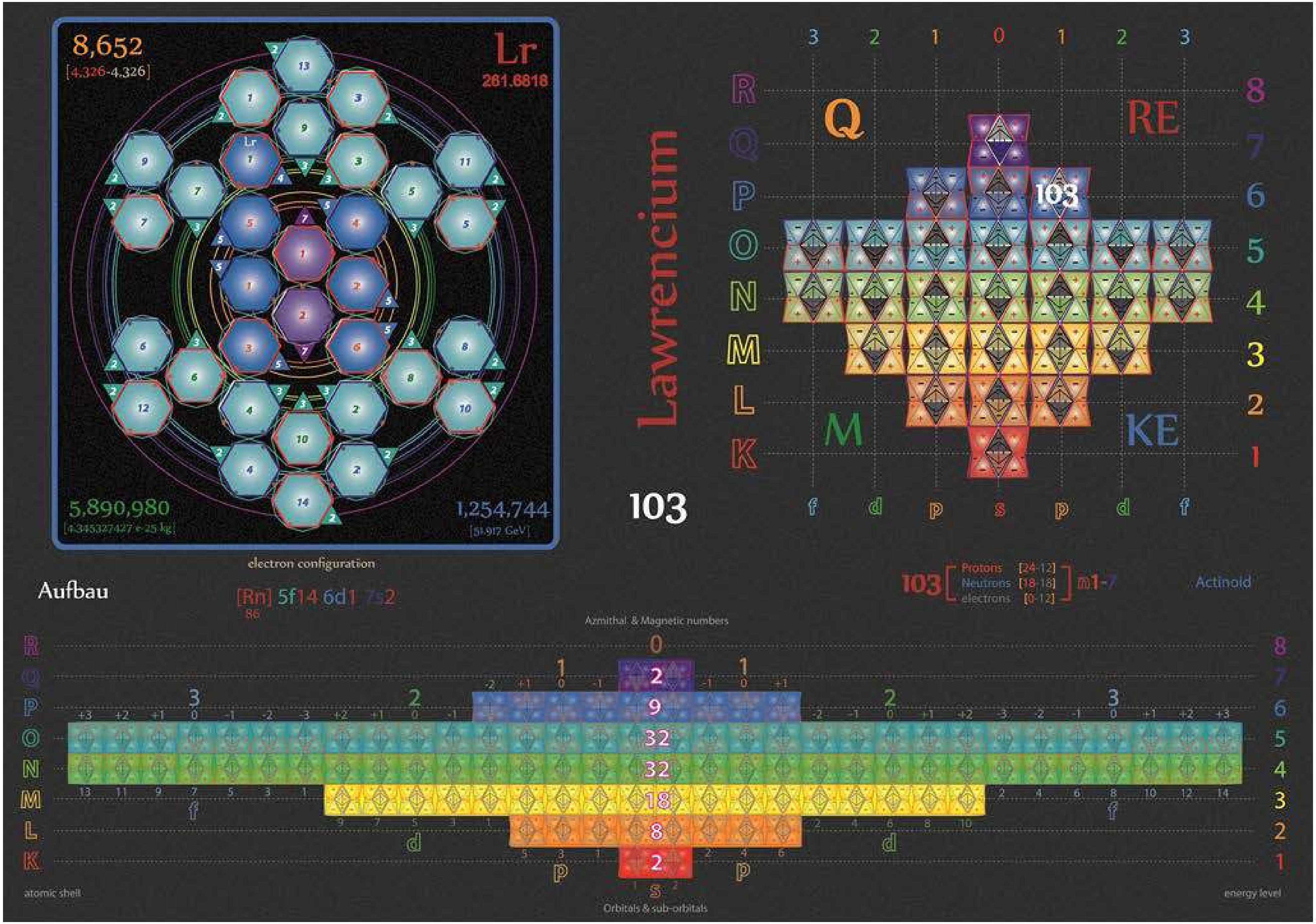
8,652
[4,326-4,326]
Lr
261.6818
5,890,980
[4.345327427 e-25 kg]
1,254,744
[51.917 GeV]
Lawrencium
103
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
Aufbau
[Rn] 5f14 6d1 7s2
86
103 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
9
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.103 - Lawrencium atom

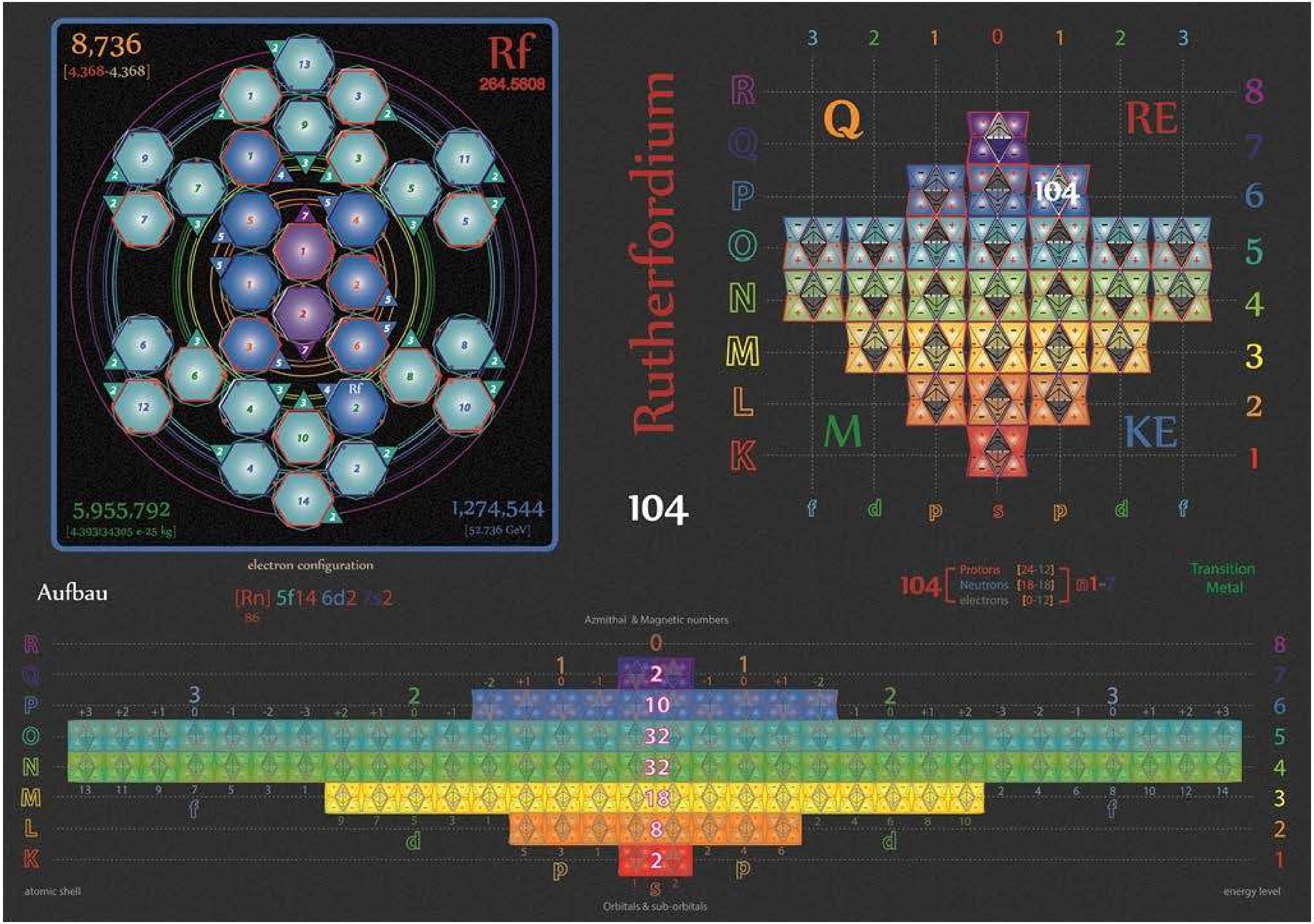
8,736
[4,368-4,368]
Rf
264.5608
5,955,792
[4.393134305 e-25 kg]
1,274,544
[52.736 GeV]
electron configuration
Aufbau
[Rn] 5f14 6d2 7s2
86
Rutherfordium
104
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
f
d
p
s
p
d
f
104 Protons [24-12] Neutrons [18-18] electrons [0-12] n1-7
Transition Metal
Azmithal & Magnetic numbers
2
10
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.104 - Rutherfordium atom

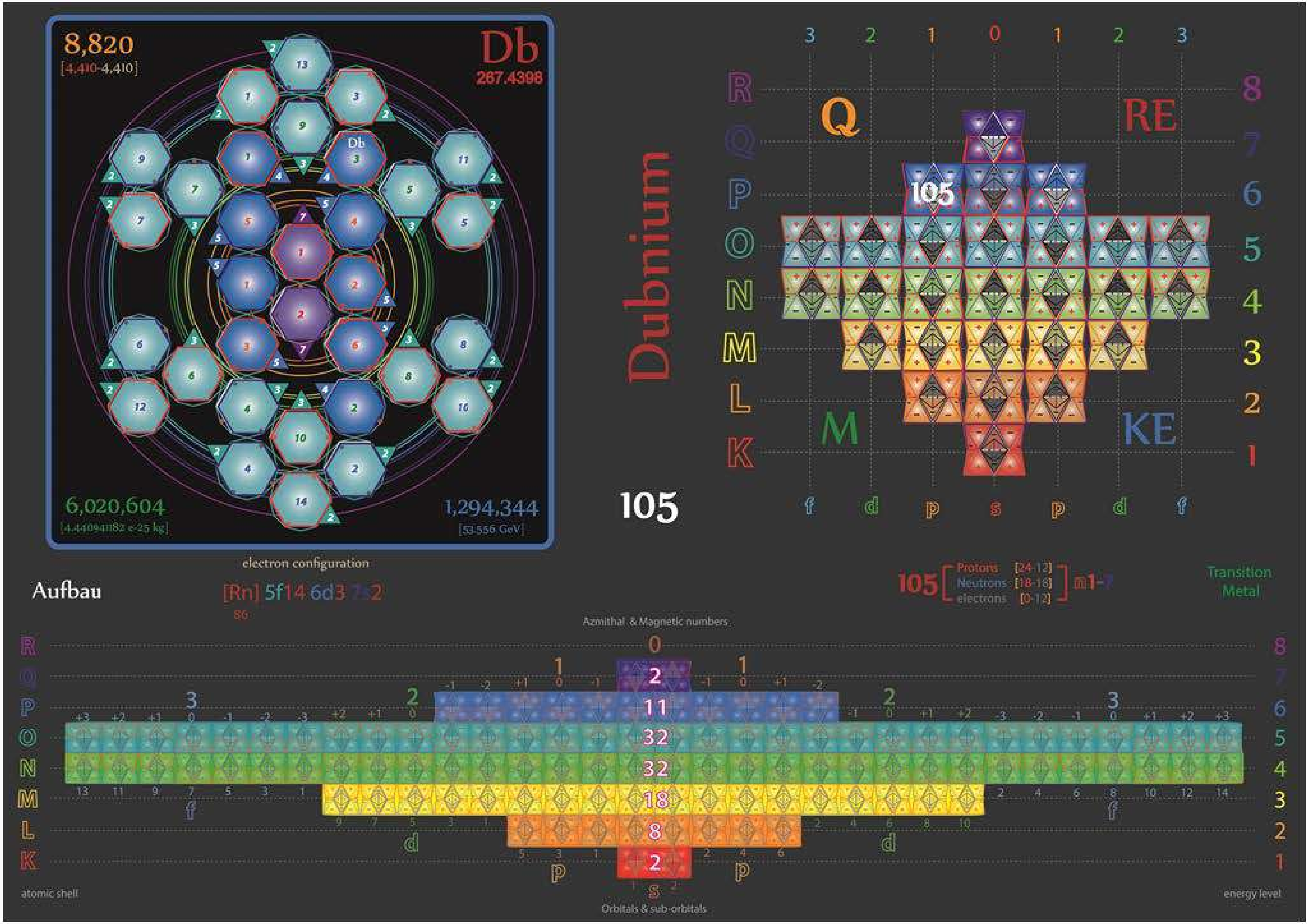
8,820
[4.410-4.410]
Db
287.4398
6,020,604
[4.44094182 e-25 kg]
1,294,344
[53.556 GeV]
Dubnium
105
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
Q
RE
M
KE
f
d
p
s
p
d
f
electron configuration
Aufbau
[Rn] 5f14 6d3 7s2
86
105
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Transition
Metal
Azmithal & Magnetic numbers
2
11
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.105 - Dubnium atom

Copyright ABRAHAM [2008] - All rights reserved

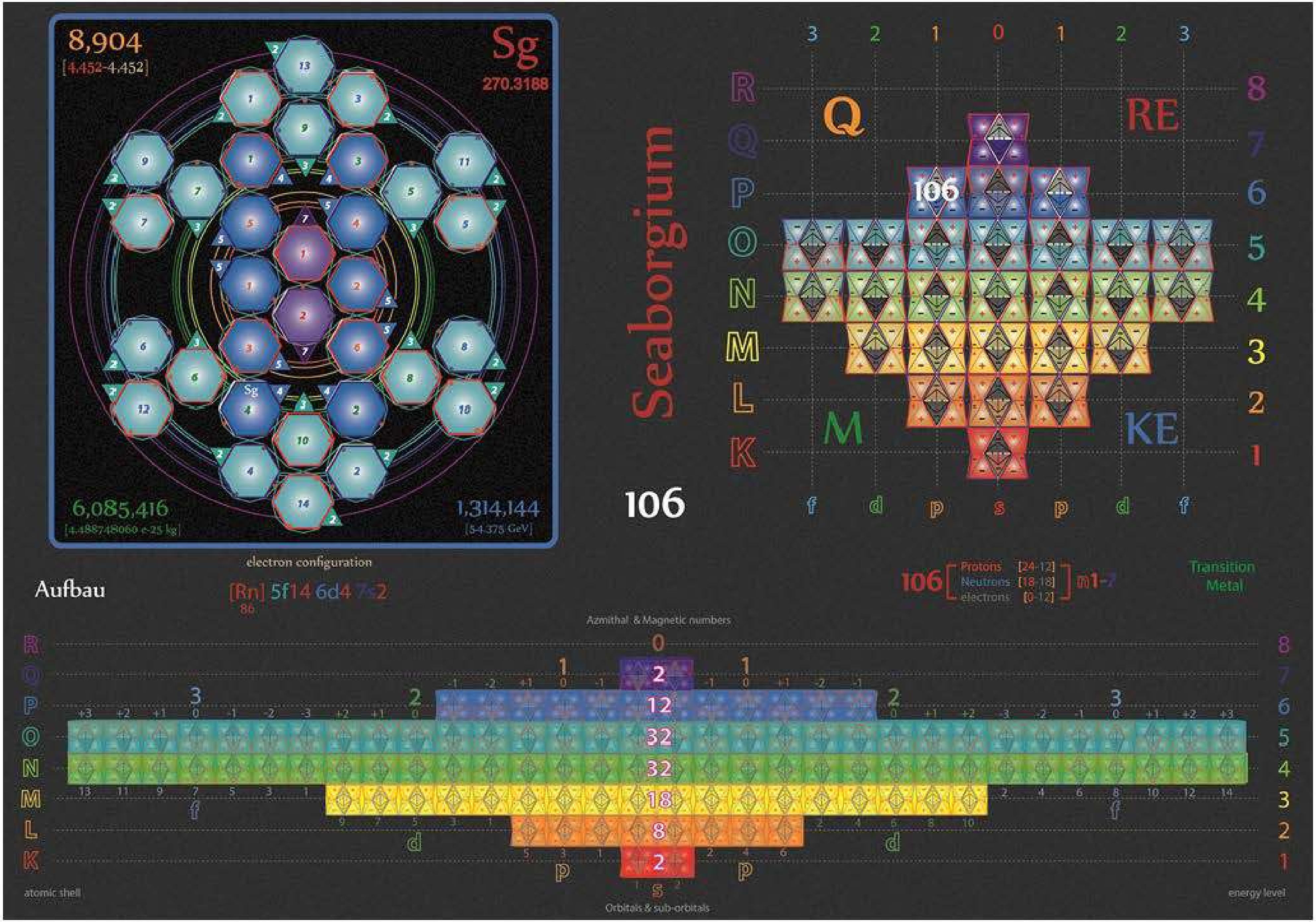
8,904
[4,452-4,452]
Sg
270.3188
6,085,416
[4.488748060 e-25 kg]
1,314,144
[54.375 GeV]
Seaborgium
106
electron configuration
Aufbau
[Rn] 5f14 6d4 7s2
86
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
106 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azimthal & Magnetic numbers
0
1
1
2
2
3
3
2
12
32
32
18
8
2
f
f
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.106 - Seaborgium atom

Copyright ABRAHAM [2008] - All rights reserved

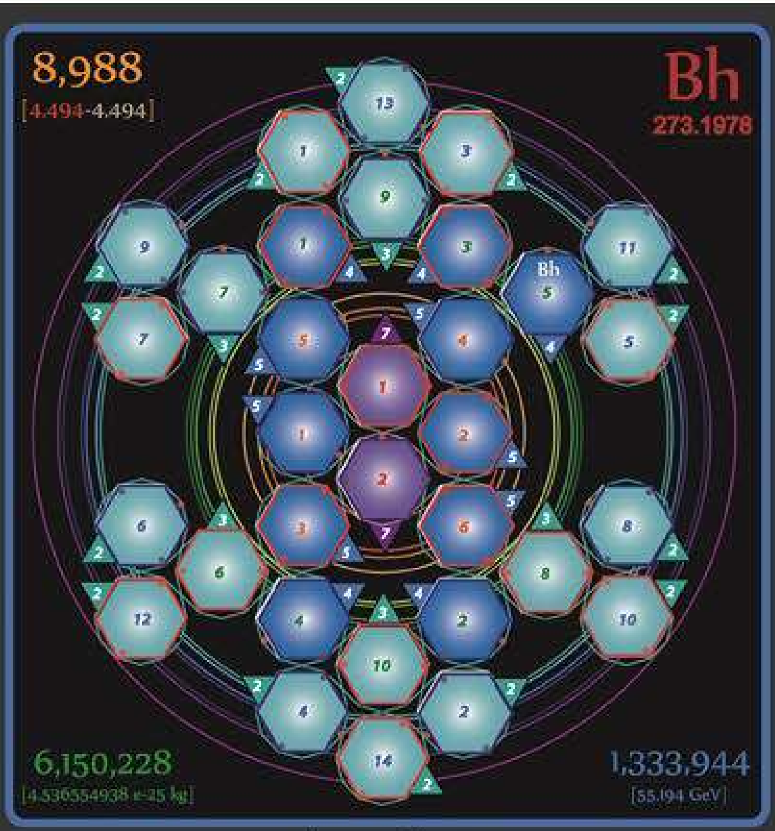
8,988
[4.494-4.494]
Bh
273.1978
6,150,228
[4.536554938 e-25 kg]
1,333,944
[55.194 GeV]

Bohrium

107

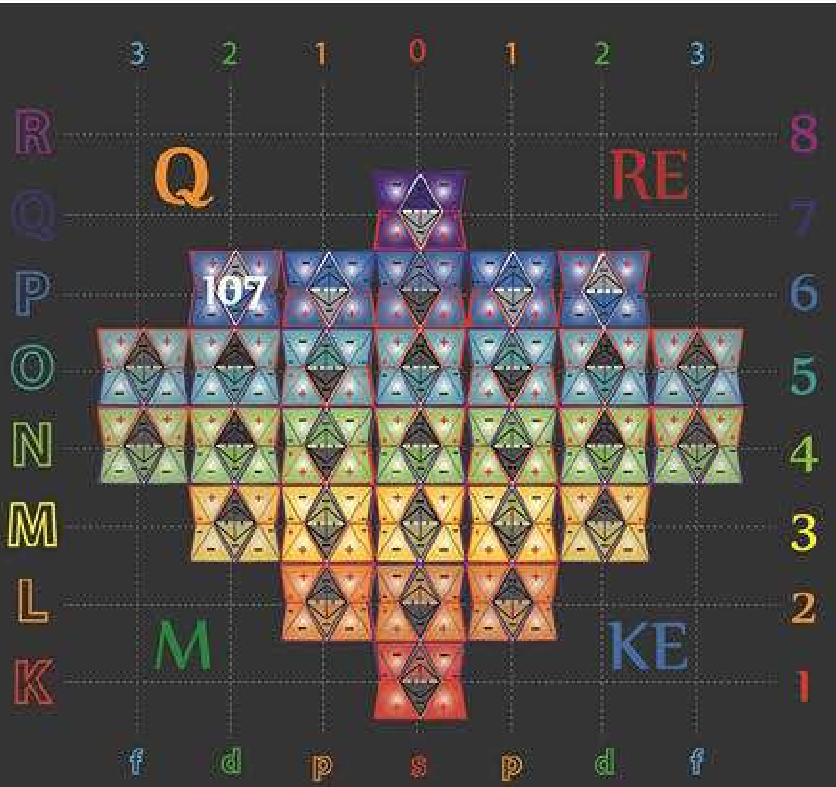
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
107
M
KE
f
d
p
s
p
d
f

Aufbau

electron configuration

[Rn] 5f14 6d5 7s2

86

107 [Protons [24-12]
Neutrons [18-18]
electrons [0-12]] m1-7

Transition Metal

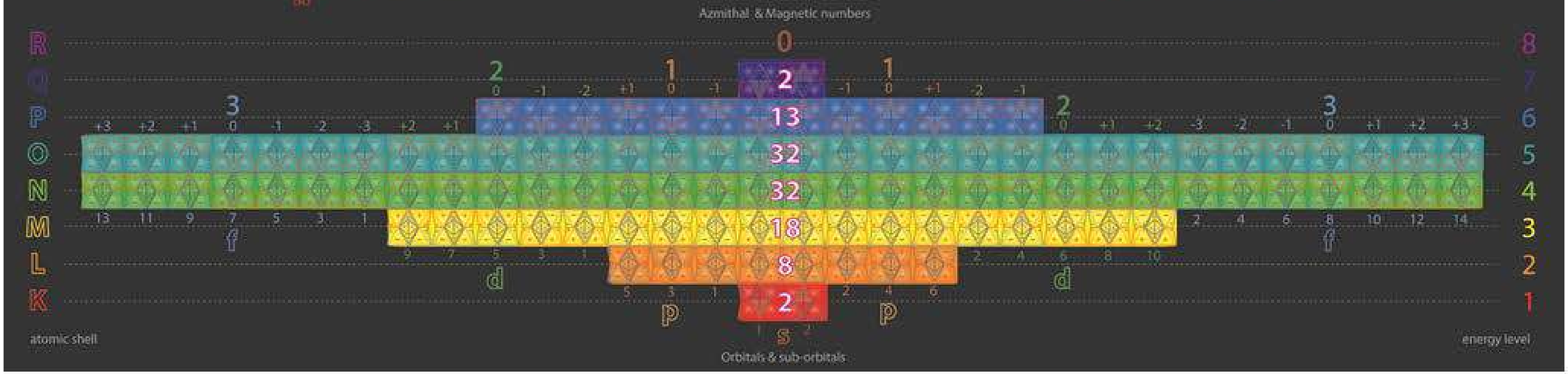
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
13
32
32
18
8
2
1
1
2
2
3
3
f
f
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

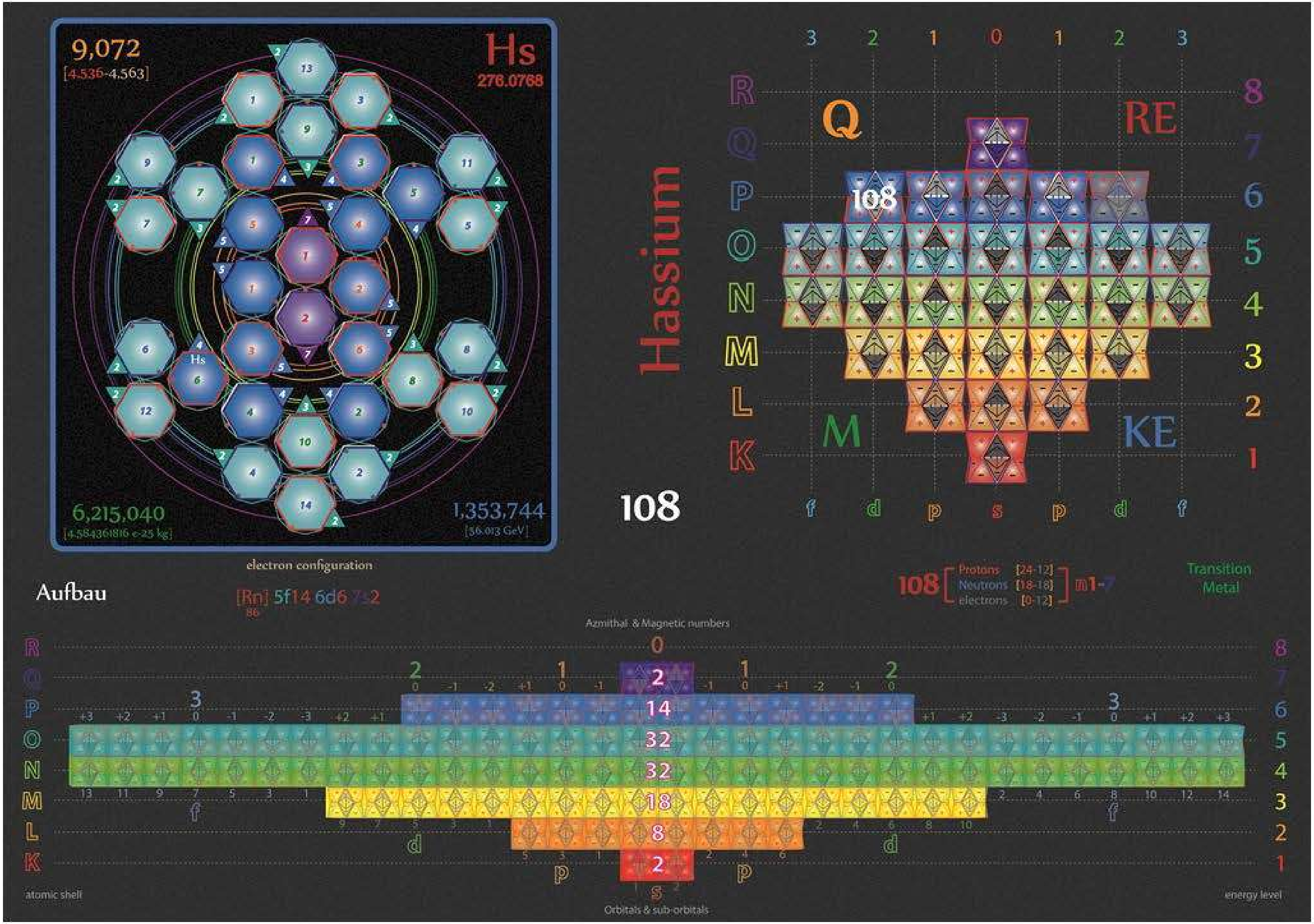
9,072
[4.536-4.563]
Hs
276.0768
6,215,040
[4.584361816 e-25 kg]
1,353,744
[56.013 GeV]
Hassium
108
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d6 7s2
86
108 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
0
2
14
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.108 - Hassium atom

Copyright ABRAHAM [2008] - All rights reserved

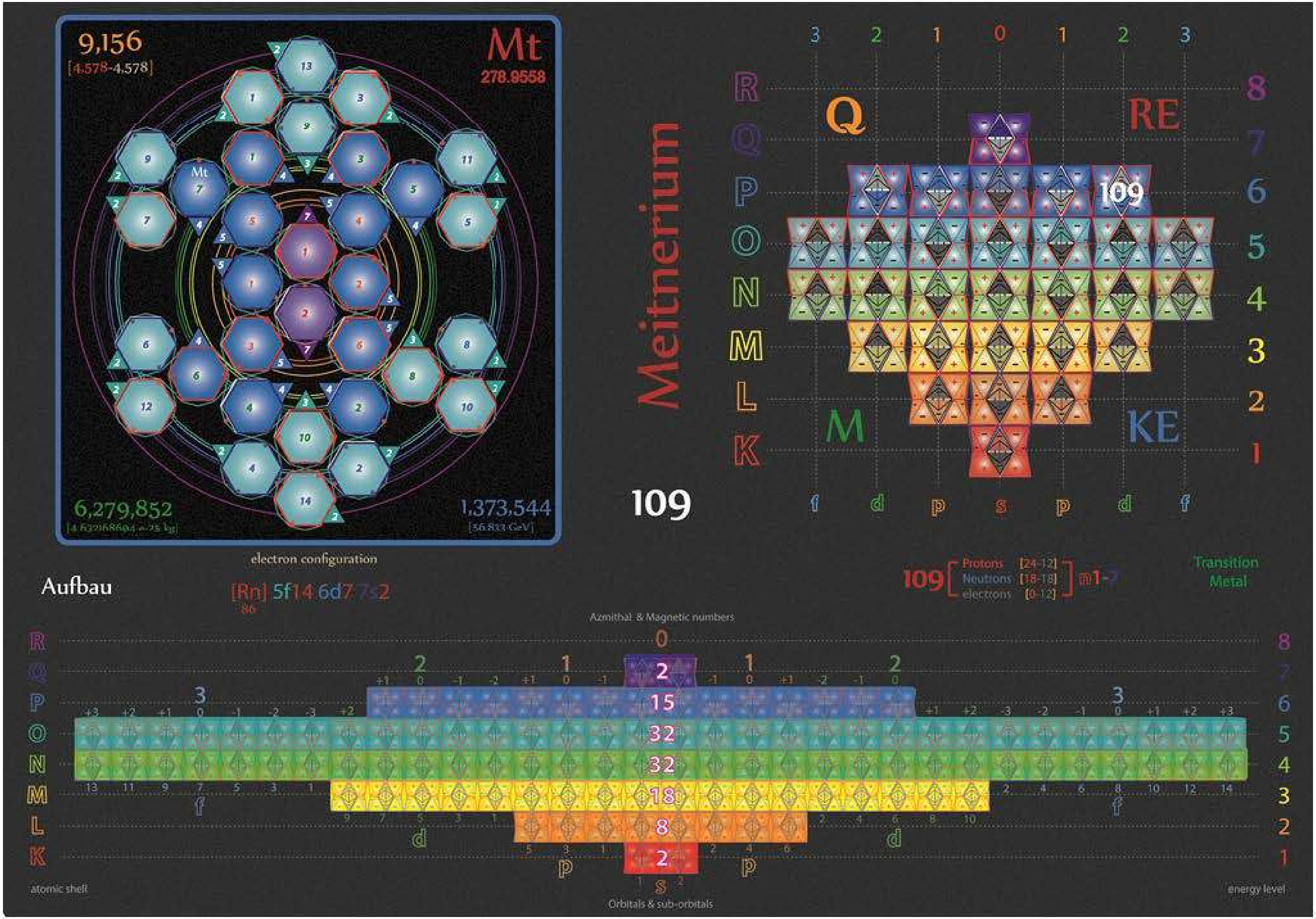
9,156
[4.578-4.578]
Mt
278.9558
6,279,852
[4.632168604 e-25 kg]
1,373,544
[56.833 GeV]
Meitnerium
109
electron configuration
Aufbau
[Rn] 5f14 6d7 7s2
86
Q
RE
M
KE
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
f d p s p d f
109 [Protons [24-12] Neutrons [18-18] electrons [0-12]] @1-7
Transition Metal
Azmithal & Magnetic numbers
2 15 32 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.109 - Meitnerium atom

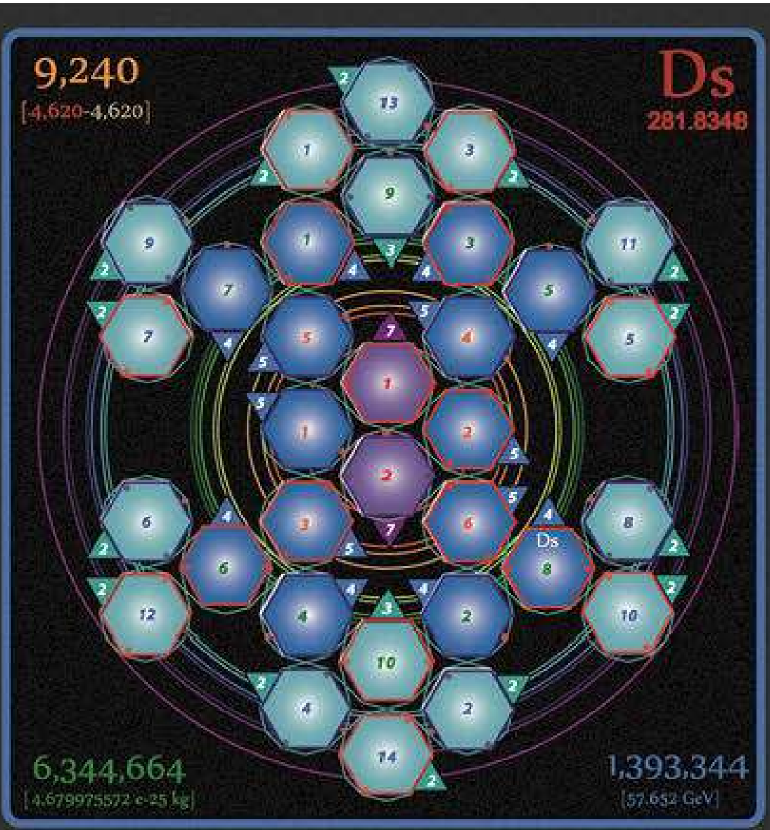
9,240
[4,620-4,620]
Ds
281.8348
Ds
6,344,664
[4.679975572 e-25 kg]
1,393,344
[57.652 GeV]

Darmstadtium

110

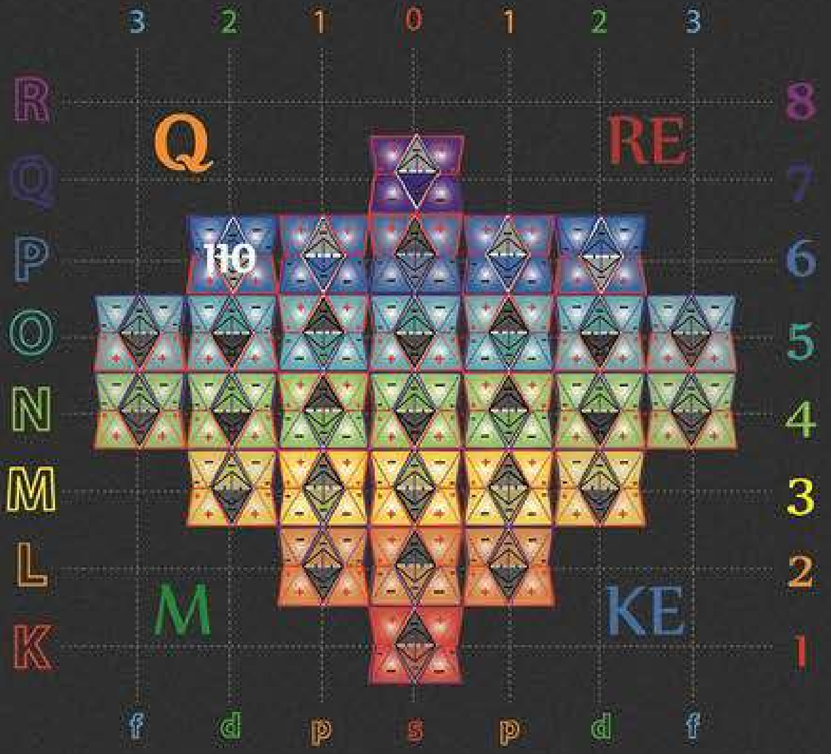
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
110
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

Aufbau

[Rn] 5f14 6d8 7s2
86

110 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7

Transition Metal

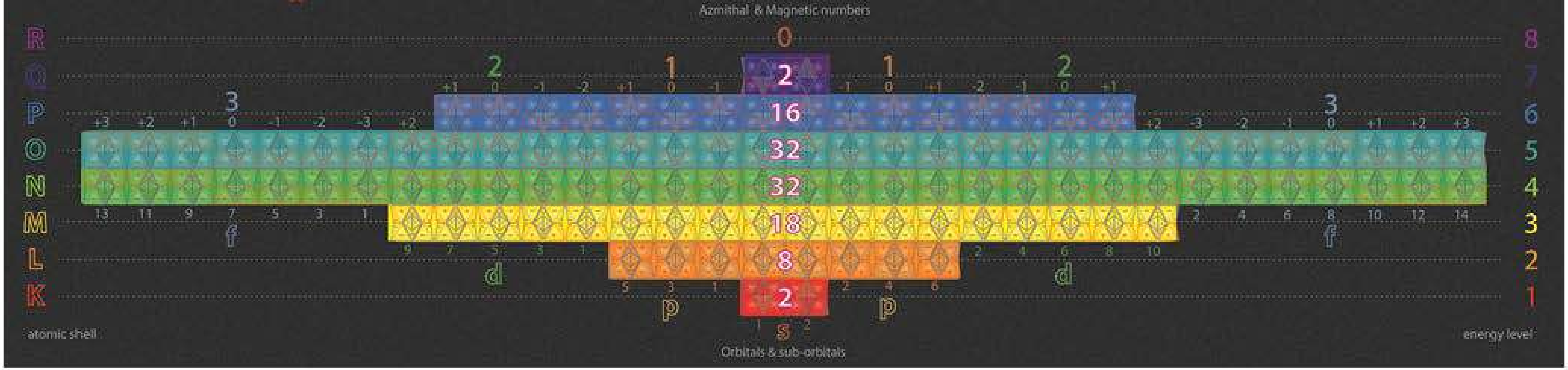
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
2
16
32
32
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.110 - Darmstadtium atom

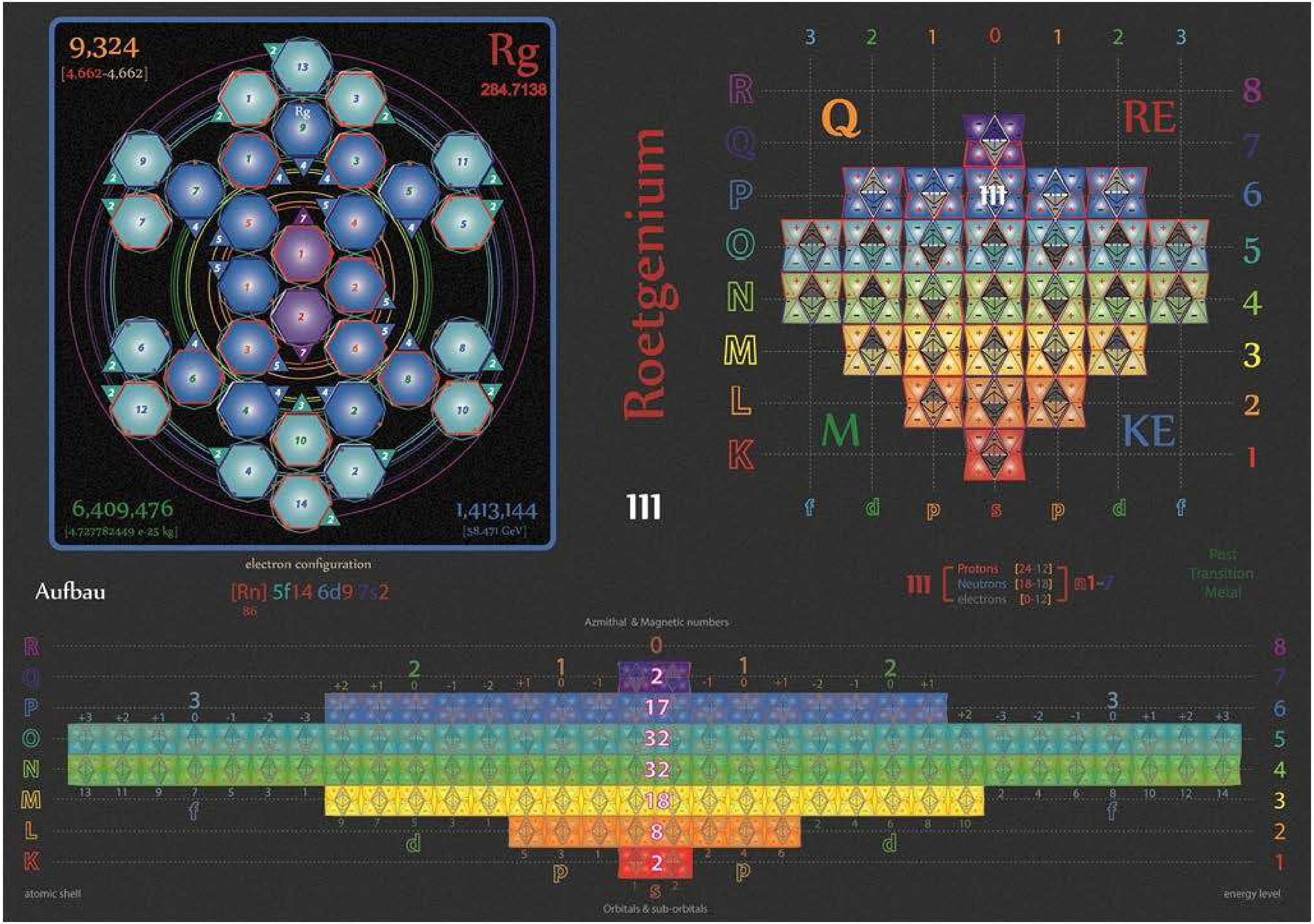
9,324
[4,662-4,662]
Rg
284.7138
6,409,476
[4.727782449 e-25 kg]
1,413,144
[58.471 GeV]
Roetgenium
111
electron configuration
Aufbau
[Rn] 5f14 6d9 7s2
86
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
111 [Protons [24-12] Neutrons [18-18] electrons [0-12]] @1-7
Post Transition Metal
Azmithal & Magnetic numbers
0 2 17 32 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.111 - Roetgenium atom

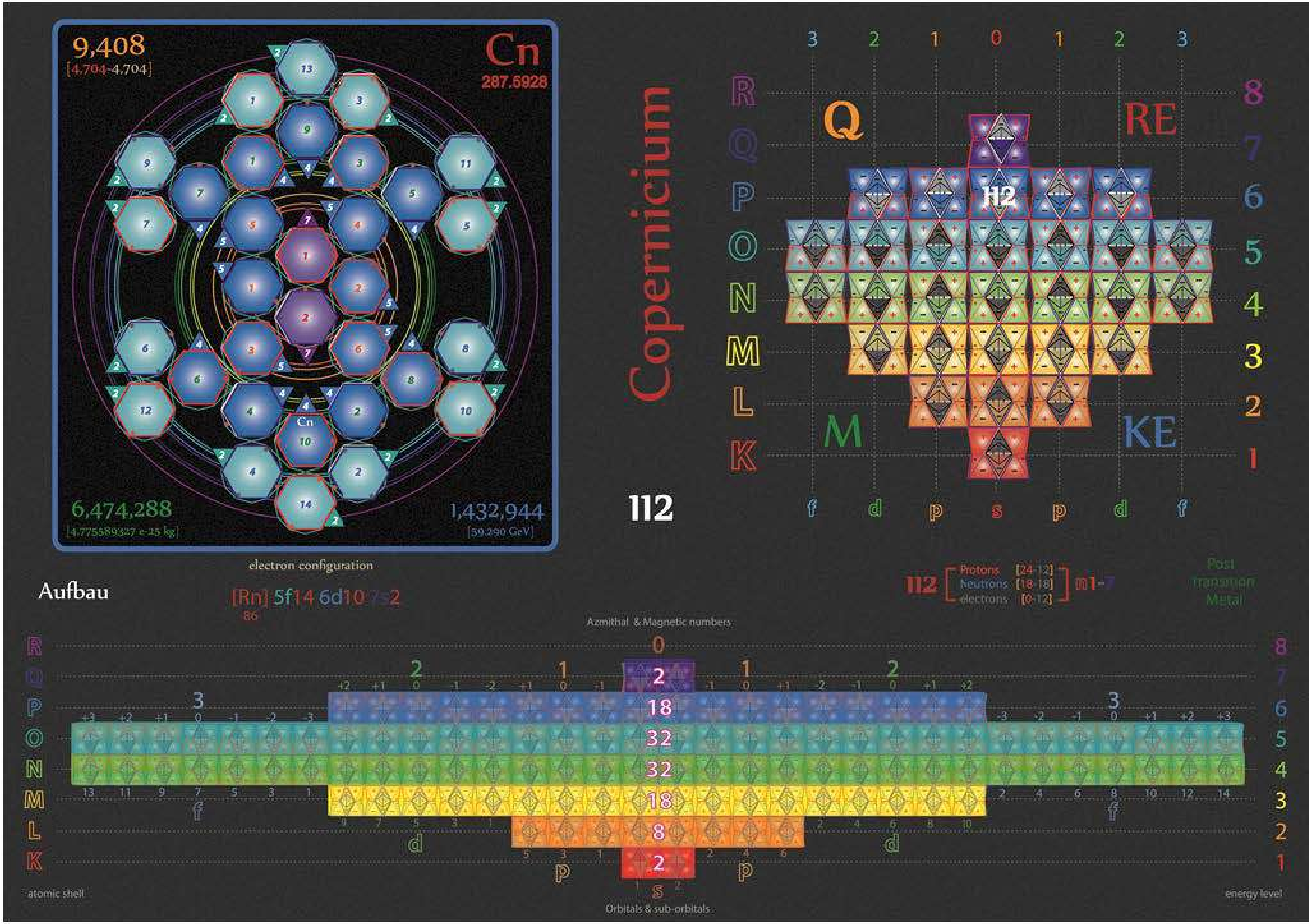
9,408
[4,704-4,704]
Cn
287.6928
6,474,288
[4.775589327 e-25 kg]
1,432,944
[59.290 GeV]
Copernicium
112
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2
86
112 [Protons [24-12] / Neutrons [18-18] / electrons [0-12]] m1-7
Post Transition Metal
Azimuthal & Magnetic numbers
2
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.112 - Copernicium atom

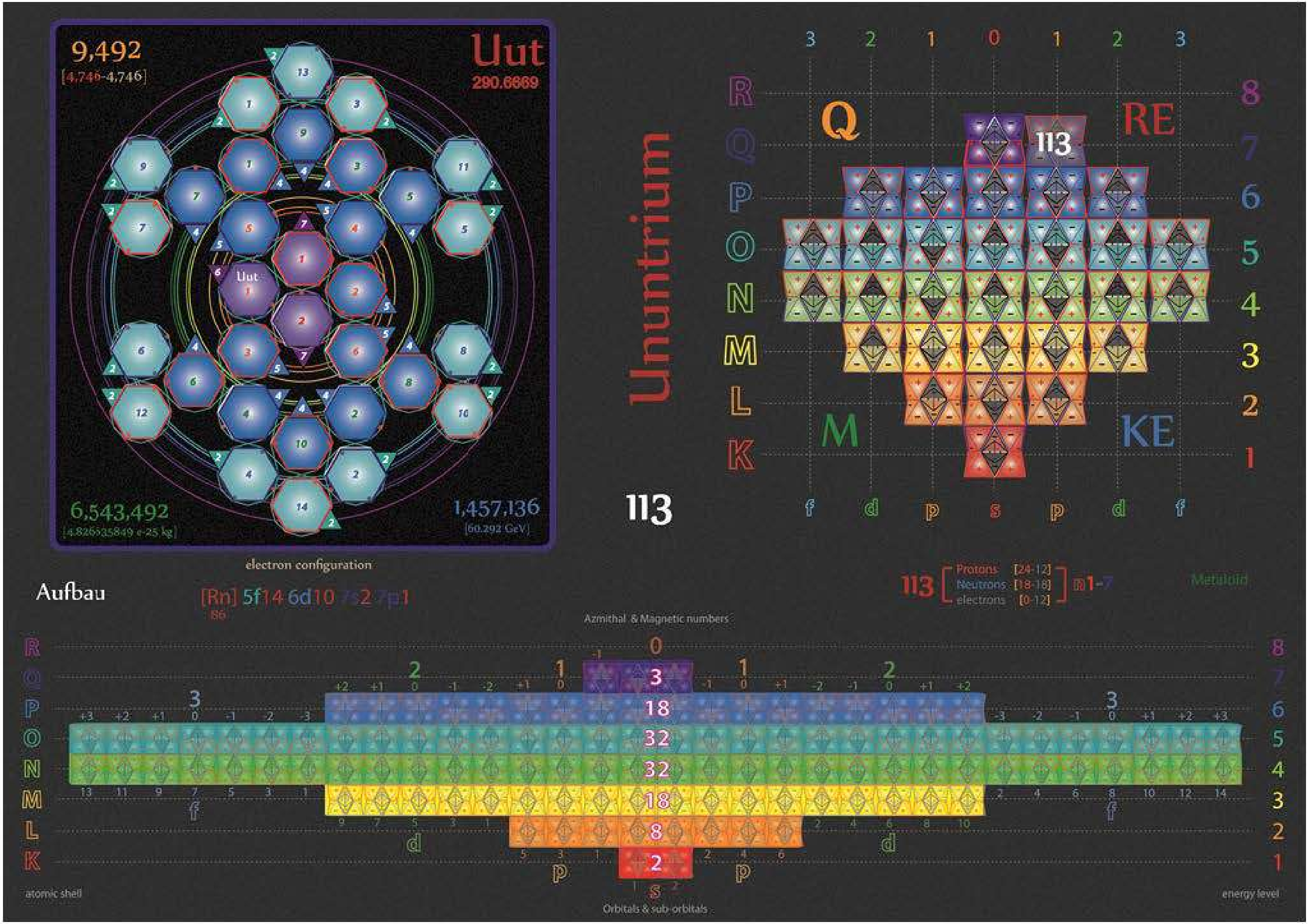
9,492
[4,746-4,746]
Uut
290.6669
Uut
6,543,492
[4.826535849 e-25 kg]
1,457,136
[60.292 GeV]
Ununtrium
113
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
113
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p1
86
113 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Metalloid
Azmithal & Magnetic numbers
3
18
32
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.113 - Ununtrium atom

Copyright ABRAHAM [2008] - All rights reserved

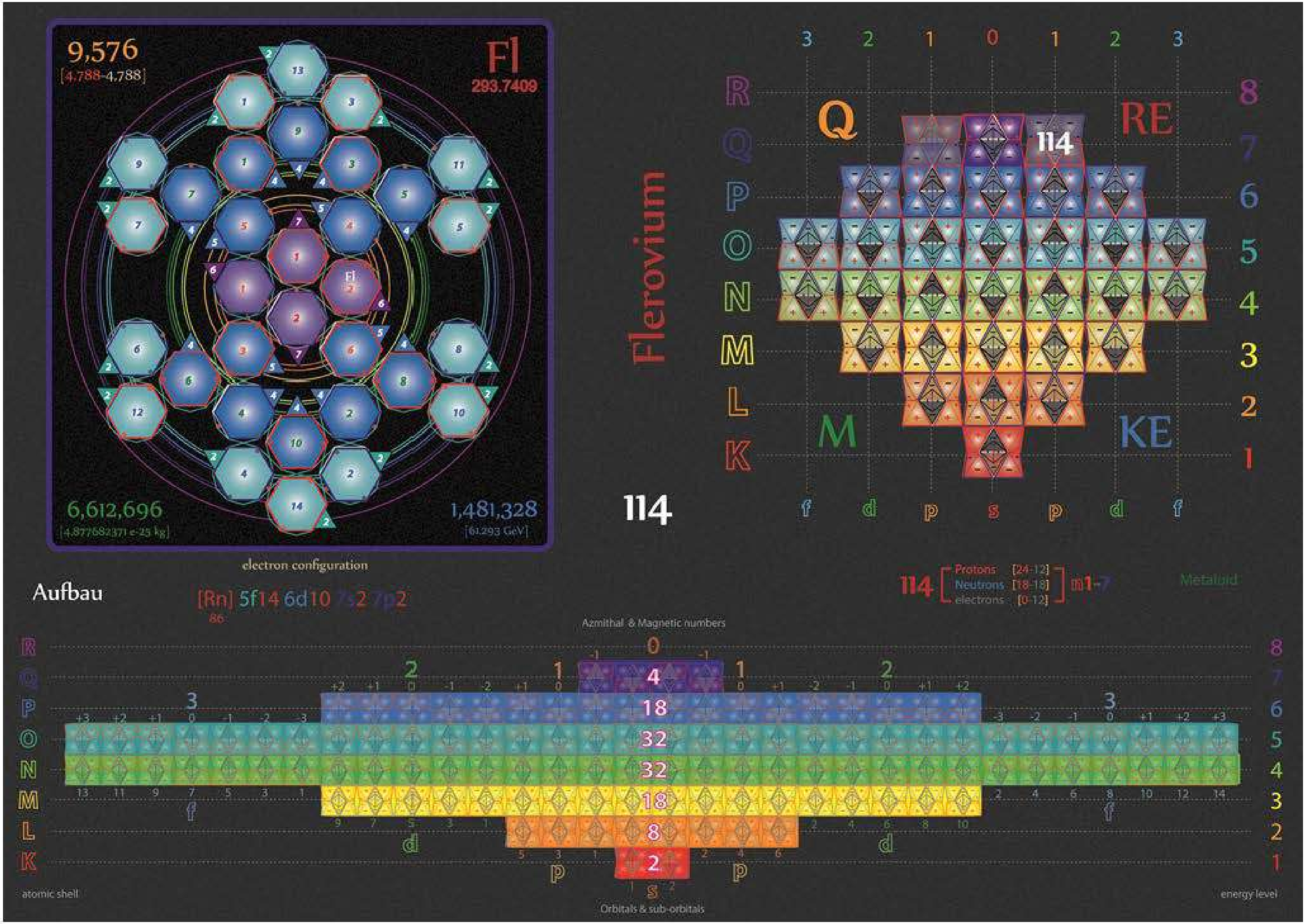
9,576
[4,788-4,788]
Fl
293.7409
6,612,696
[4.877682371 e-25 kg]
1,481,328
[61.293 GeV]
electron configuration
Flerovium
114
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
114 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Metalloid
Aufbau
[Rn] 5f14 6d10 7s2 7p2
86
Azmithal & Magnetic numbers
0
4
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.114 - Flerovium atom

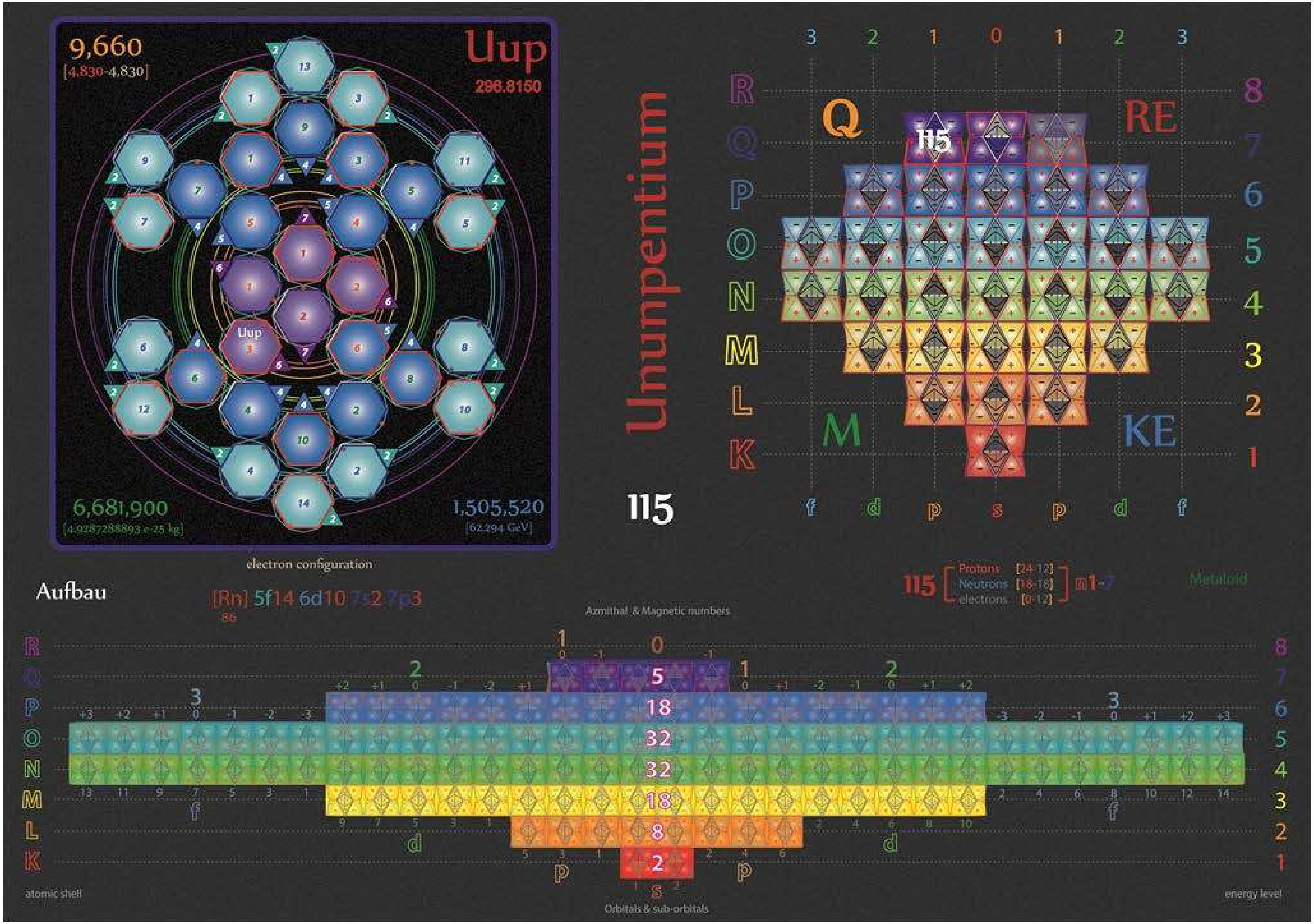
9,660
[4,830-4,830]
Uup
296.8150
Uup
6,681,900
[4.9287288893 e-25 kg]
1,505,520
[62.294 GeV]
Ununpentium
115
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
115
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p3
86
115
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-7
Metalloid
Azmithal & Magnetic numbers
R Q P O N M L K
5
18
32
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.115 - Ununpentium atom

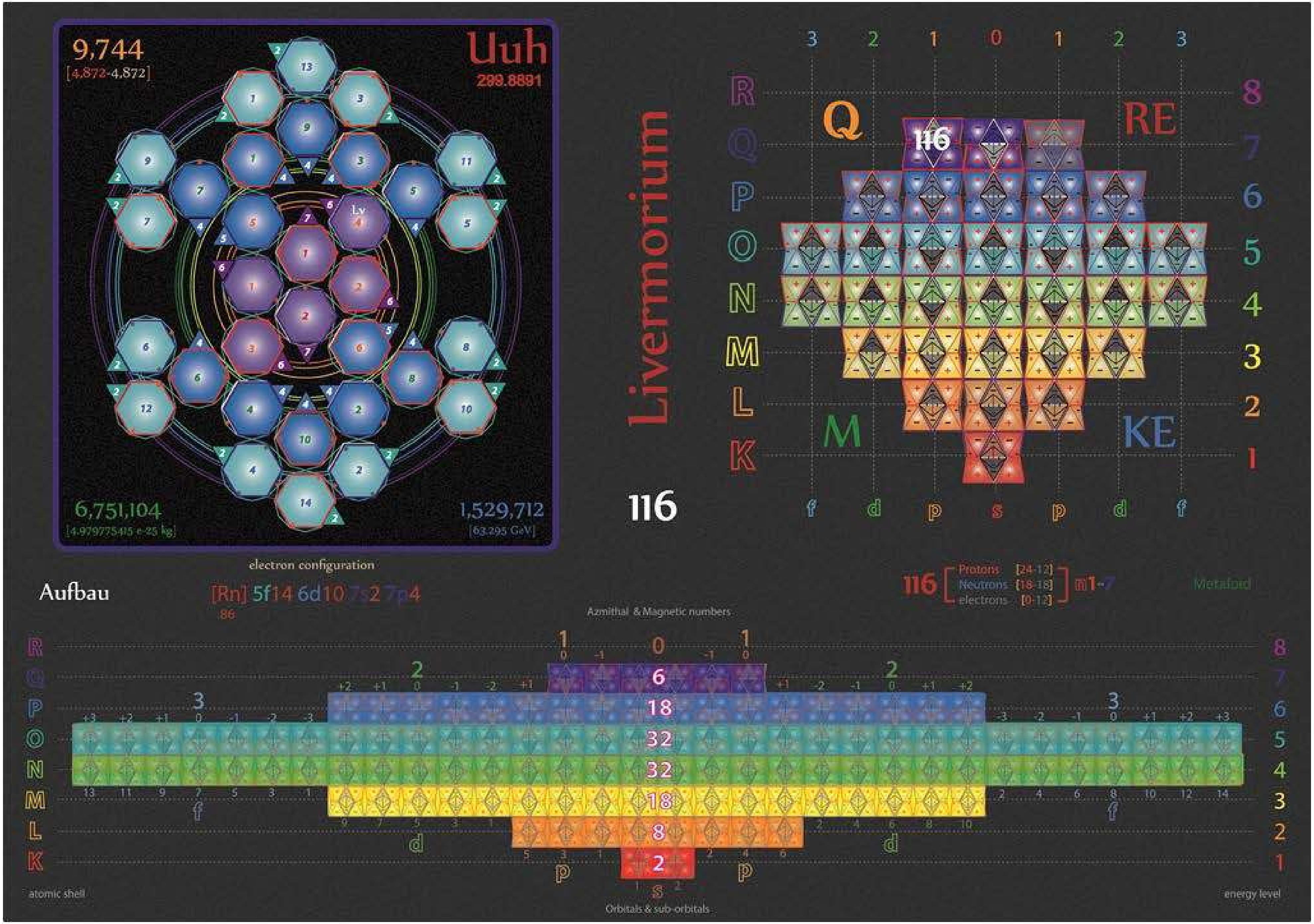
9,744
[4,872-4,872]
Uuh
289.8891
Lv
6,751,104
[4.979775415 e-25 kg]
1,529,712
[63.295 GeV]
Livermorium
116
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
116
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p4
86
116 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Metaloid
Azmithal & Magnetic numbers
6
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.116 - Livermorium atom

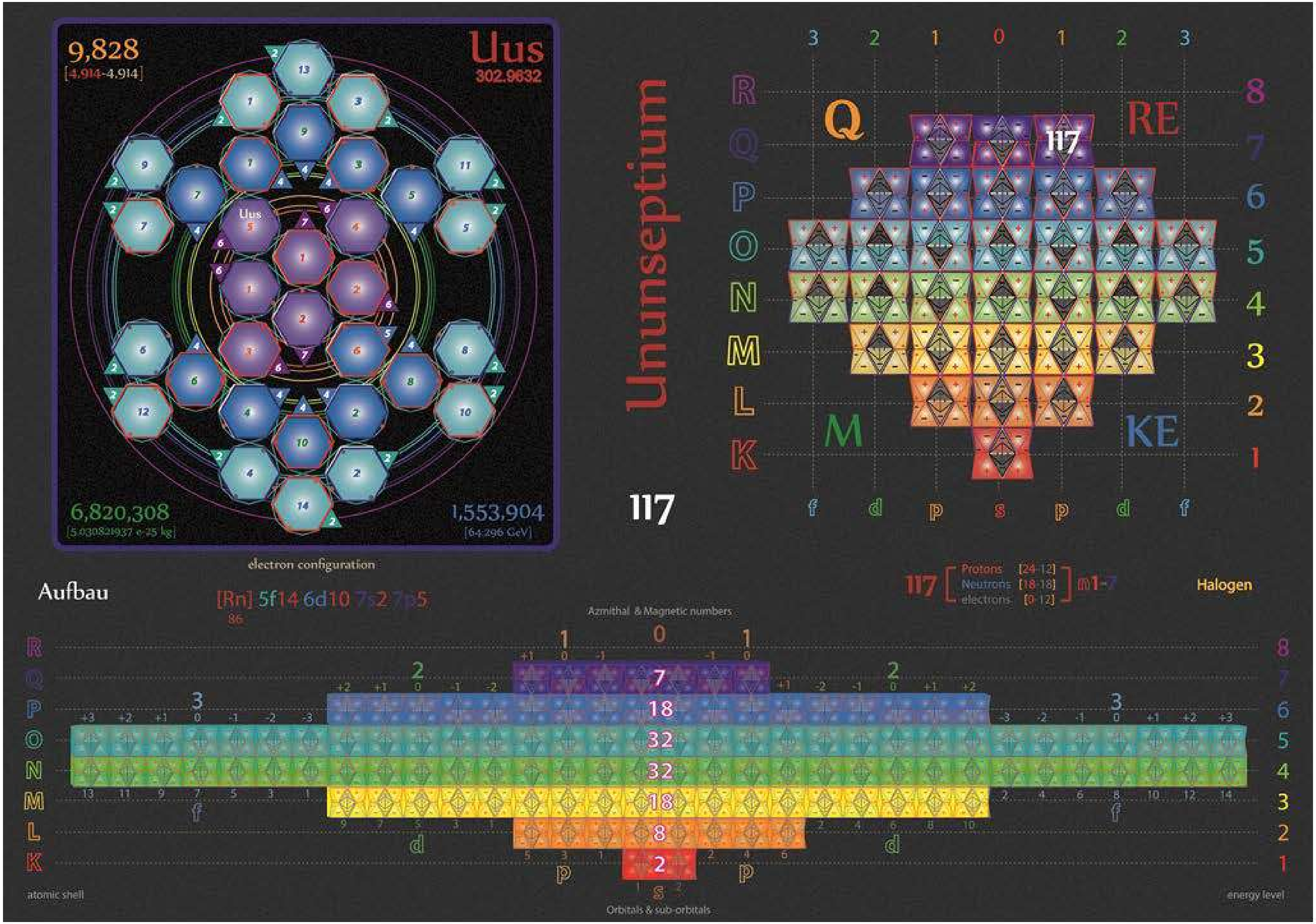
9,828
[4.914-4.914]
Uus
302.9832
Uus
5
6,820,308
[5.030821937 e-25 kg]
1,553,904
[64.296 GeV]
electron configuration
Ununseptium
117
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
117
M
KE
f d p s p d f
117 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Halogen
Aufbau
[Rn] 5f14 6d10 7s2 7p5
86
Azmithal & Magnetic numbers
1 0 1
2 2
3 3
7
18
32
32
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.117 - Ununseptium atom

Copyright ABRAHAM [2008] - All rights reserved

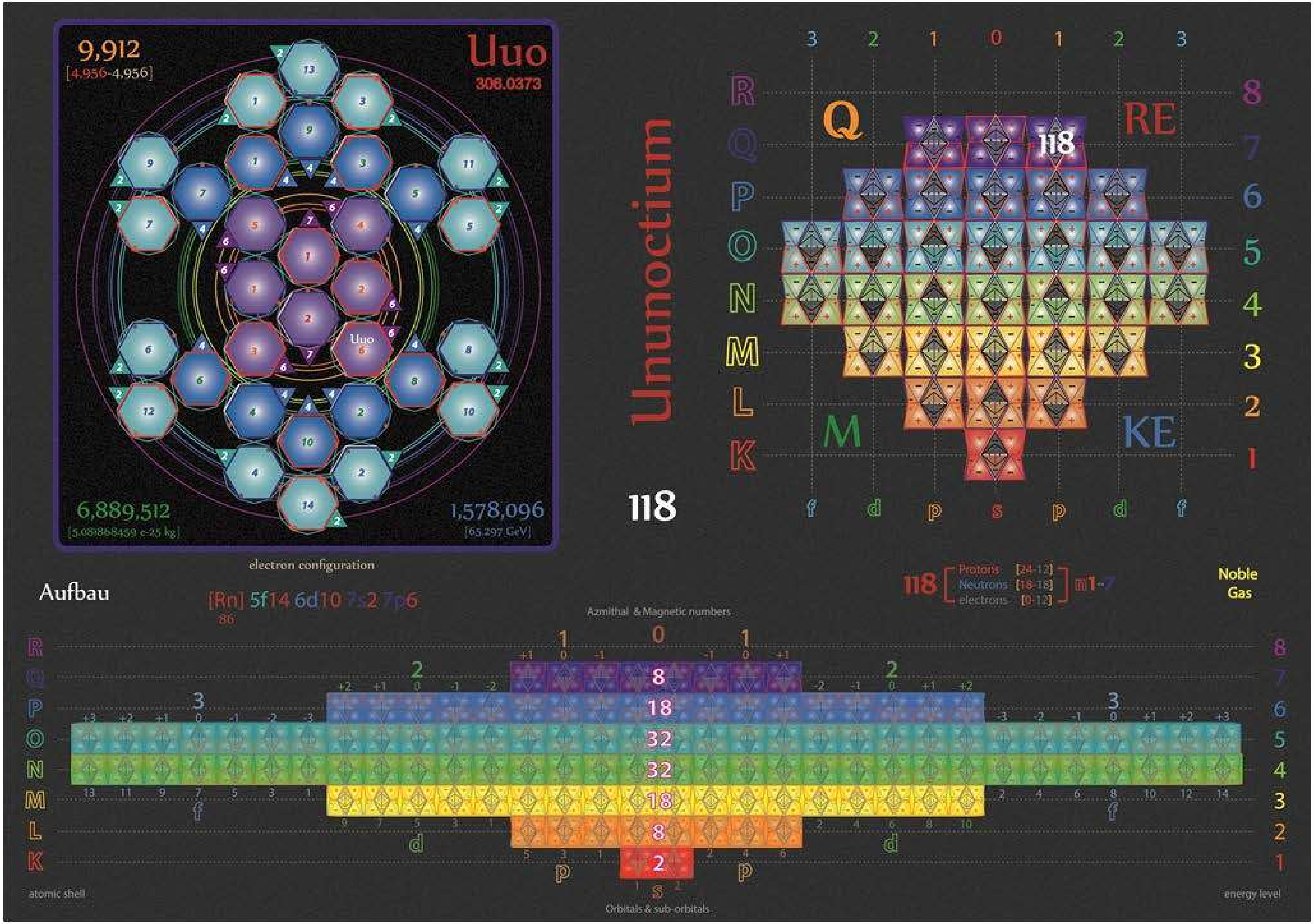
9,912
[4,956-4,956]
Uuo
308.0373
6,889,512
[5.081868459 e-25 kg]
1,578,096
[65.297 GeV]
Uuo
electron configuration
Ununoctium
118
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
118
f d p s p d f
118
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Noble Gas
Aufbau
[Rn] 5f14 6d10 7s2 7p6
86
Azmithal & Magnetic numbers
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.118 - Ununoctium atom

Copyright ABRAHAM [2008] - All rights reserved

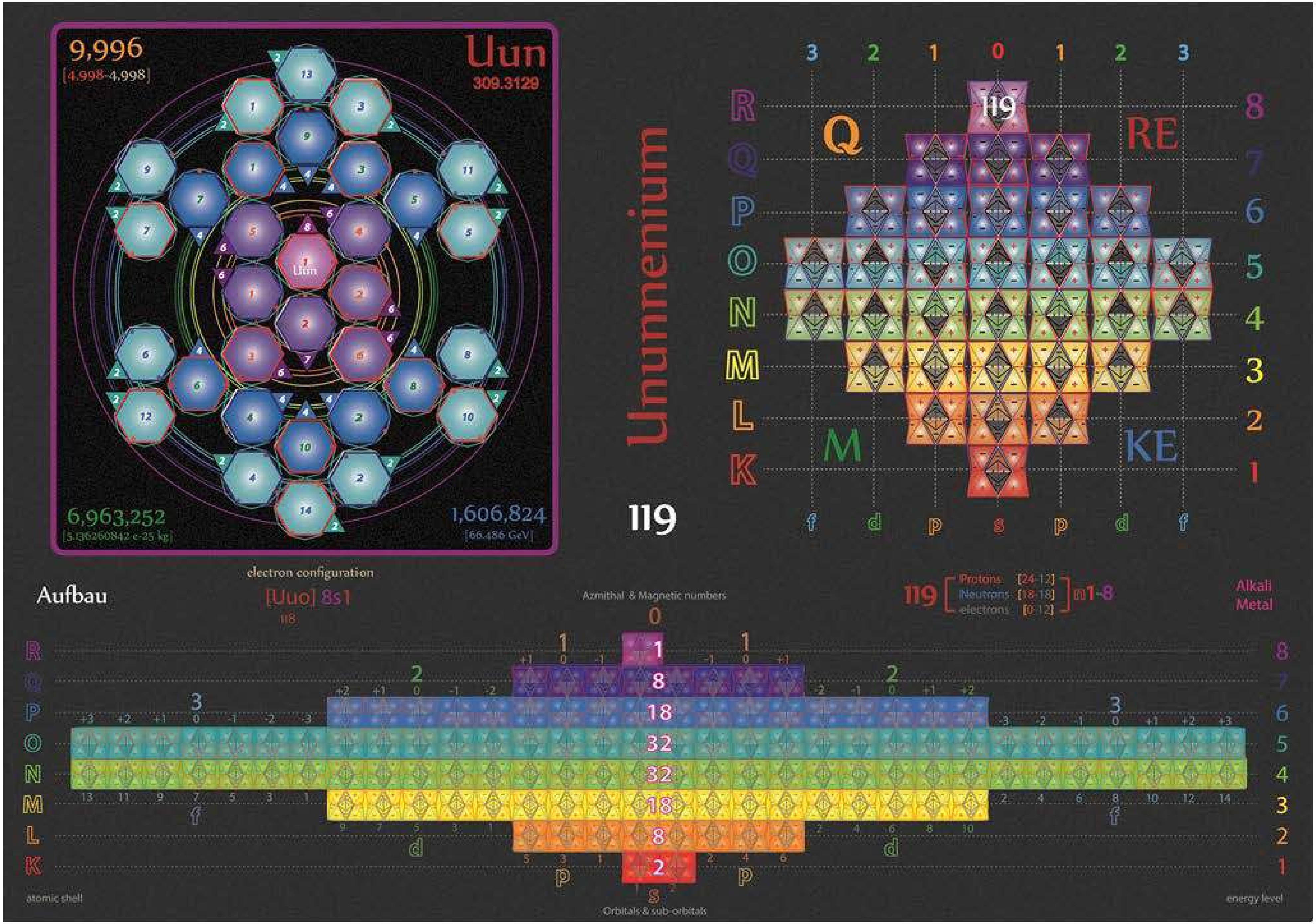
9,996
[4,998-4,998]
Uun
309.3129
6,963,252
[5.136260842 e-25 kg]
1,606,824
[66.486 GeV]
Ununnenium
119
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
Aufbau
electron configuration
[Uuo] 8s1
118
Azmithal & Magnetic numbers
119 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-8
Alkali Metal
2 8 18 32 32 18 8 1
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.119 - Ununnenium atom

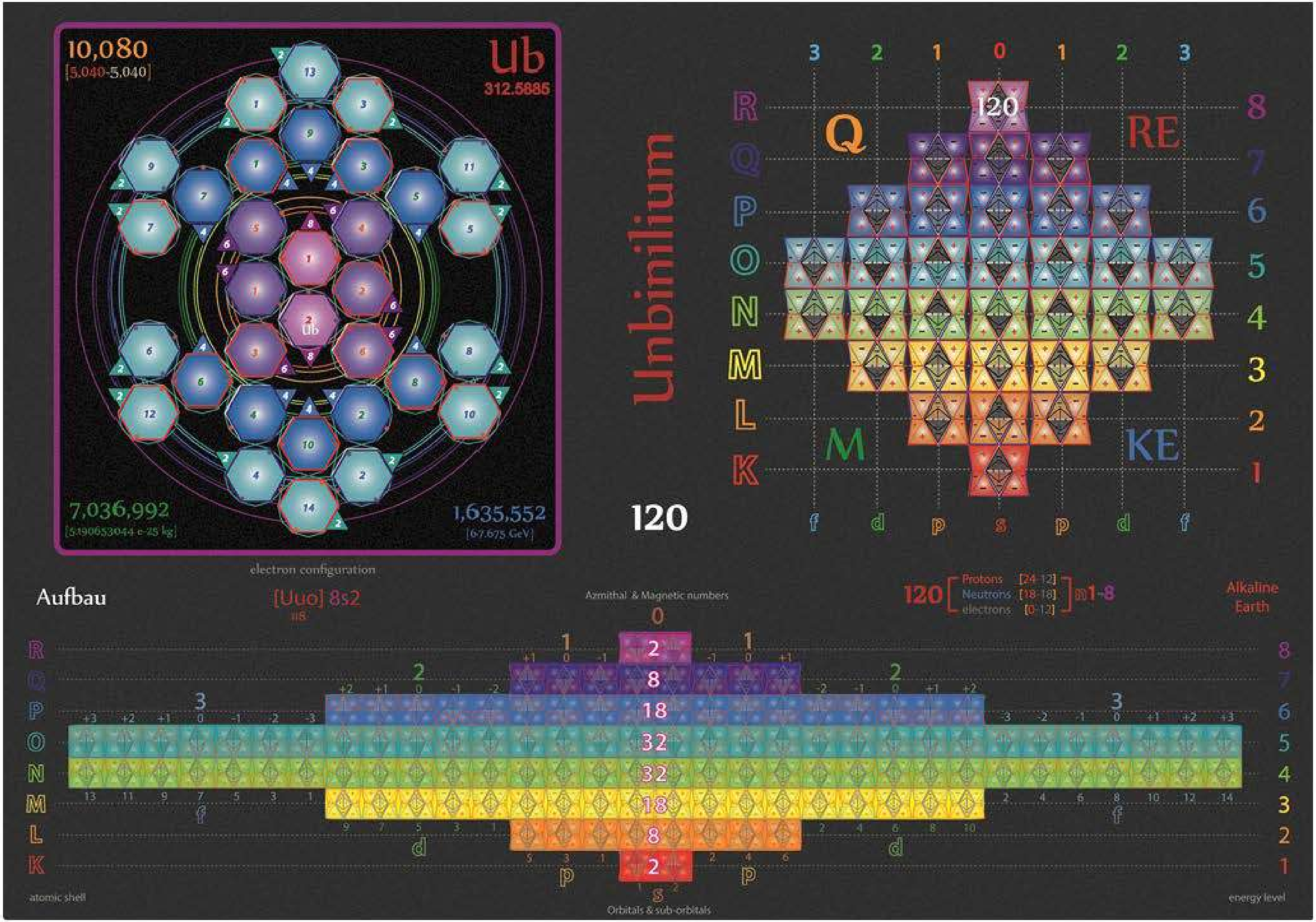
10,080
[5,040-5,040]
Ub
312.5885
7,036,992
[5.190653044 e-25 kg]
1,635,552
[67.675 GeV]
electron configuration
Unbinilium
120
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
Aufbau
[Uuo] 8s2
118
Azimthal & Magnetic numbers
120 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-8
Alkaline Earth
2
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.120 - Unbinilium atom

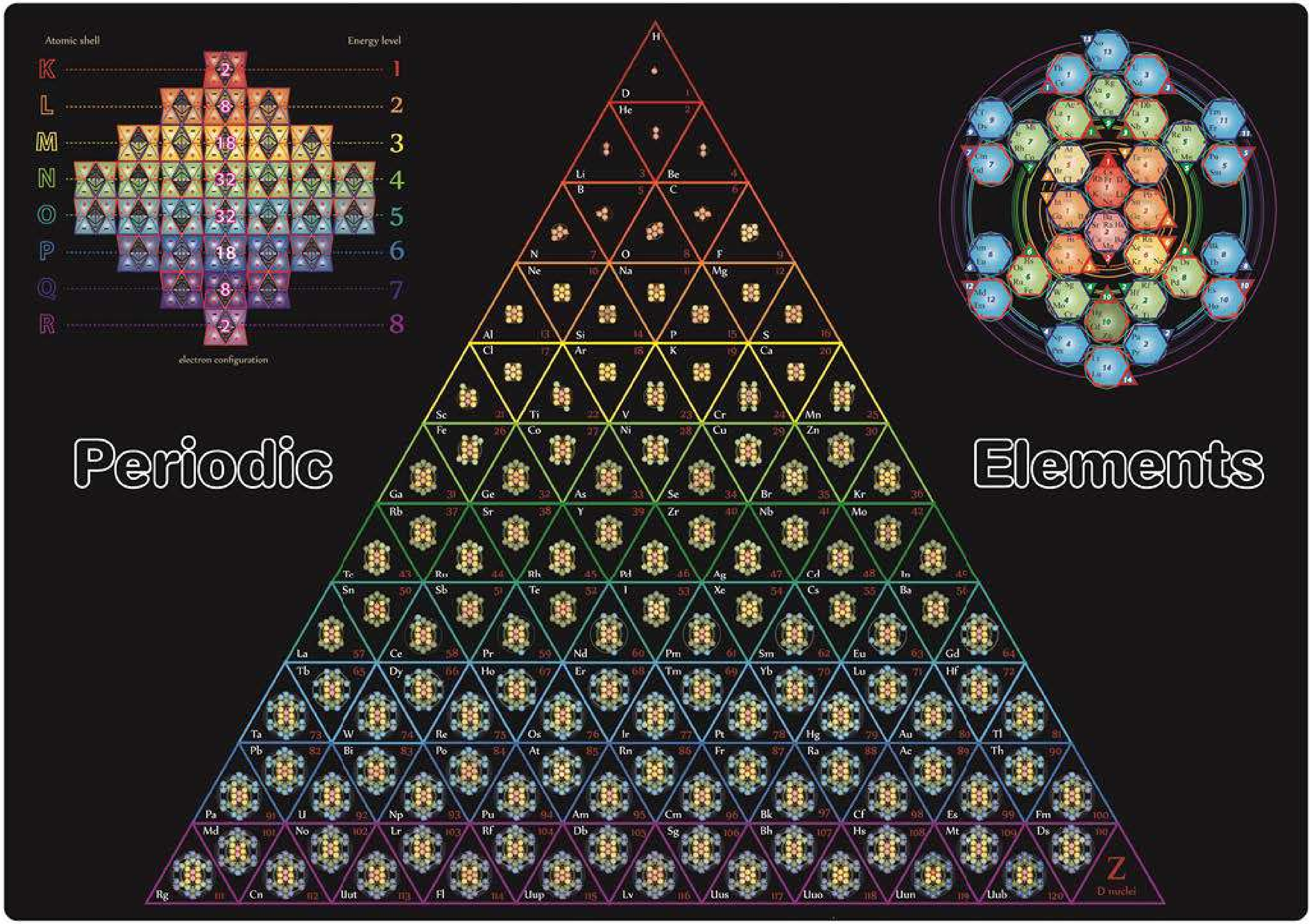
Atomic shell
Energy level
K 2 1
L 8 2
M 18 3
N 32 4
O 32 5
P 18 6
Q 8 7
R 2 8
electron configuration
Periodic
Elements
H
D 1
He 2
Li 3
Be 4
B 5
C 6
N 7
O 8
F 9
Ne 10
Na 11
Mg 12
Al 13
Si 14
P 15
S 16
Cl 17
Ar 18
K 19
Ca 20
Sc 21
Ti 22
V 23
Cr 24
Mn 25
Fe 26
Co 27
Ni 28
Cu 29
Zn 30
Ga 31
Ge 32
As 33
Se 34
Br 35
Kr 36
Rb 37
Sr 38
Y 39
Zr 40
Nb 41
Mo 42
Tc 43
Ru 44
Rh 45
Pd 46
Ag 47
Cd 48
In 49
Sn 50
Sb 51
Te 52
I 53
Xe 54
Cs 55
Ba 56
La 57
Ce 58
Pr 59
Nd 60
Pm 61
Sm 62
Eu 63
Gd 64
Tb 65
Dy 66
Ho 67
Er 68
Tm 69
Yb 70
Lu 71
Hf 72
Ta 73
W 74
Re 75
Os 76
Ir 77
Pt 78
Hg 79
Au 80
Tl 81
Pb 82
Bi 83
Po 84
At 85
Rn 86
Fr 87
Ra 88
Ac 89
Th 90
Pa 91
U 92
Np 93
Pu 94
Am 95
Cm 96
Bk 97
Cf 98
Es 99
Fm 100
Md 101
No 102
Lr 103
Rf 104
Db 105
Sg 106
Bh 107
Hs 108
Mt 109
Ds 110
Rg 111
Cn 112
Uut 113
Fl 114
Uup 115
Lv 116
Uus 117
Uuo 118
Uun 119
Uub 120
Z
D nuclei

Copyright ABRAHAM [2008] - All rights reserved

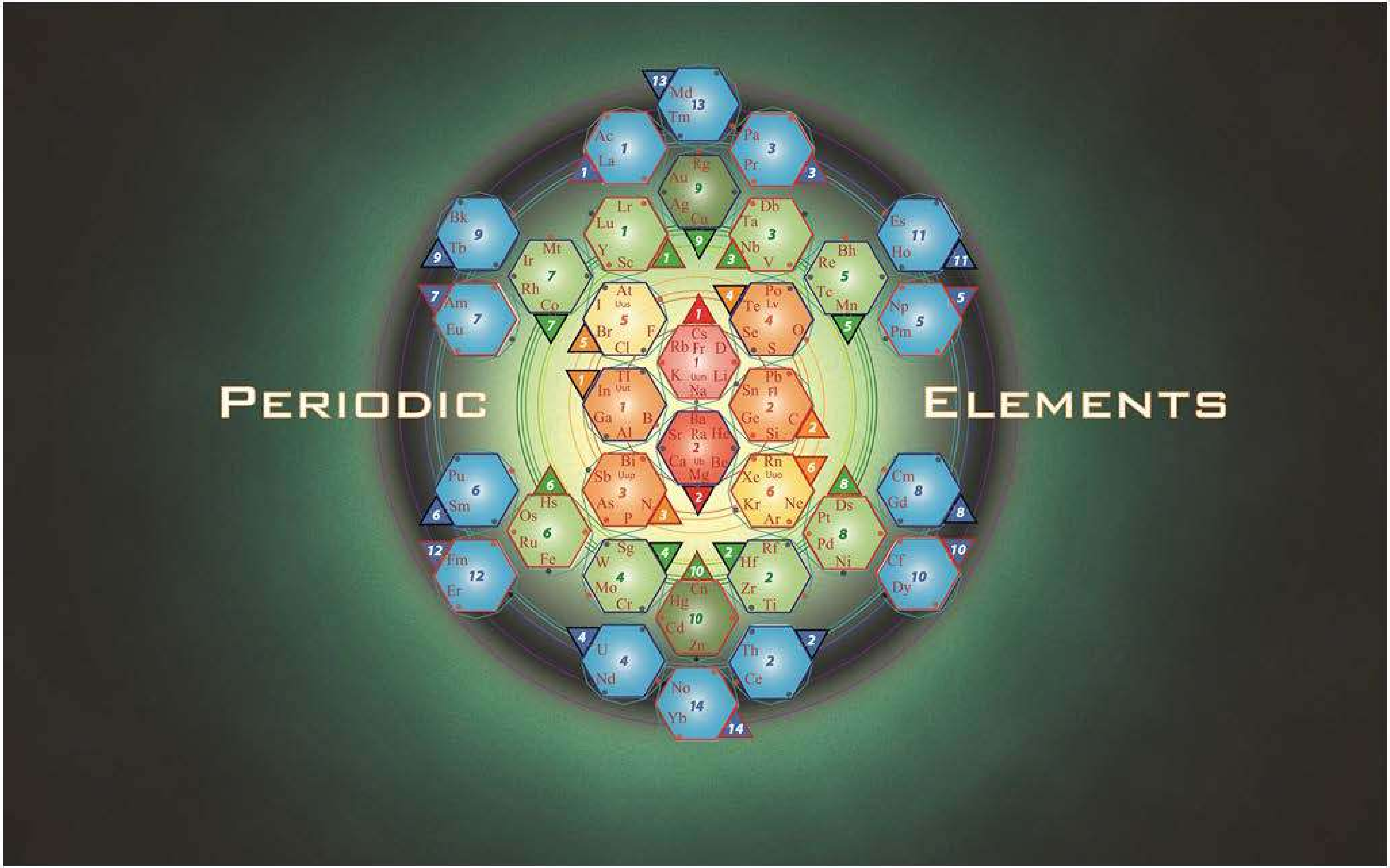
PERIODIC
ELEMENTS
Md Tm 13
Ac La 1
Pa Pr 3
Au Rg Ag Cu 9
Lr Lu Y Sc 1
Db Ta Nb V 3
Bk Tb 9
Es Ho 11
Mt Ir Rh Co 7
Bh Re Tc Mn 5
Am Eu 7
Np Pm 5
At I Uus Br Cl F 5
Po Te Lv Se S O 4
Cs Rb Fr D K Uun Li Na 1
In Tl Uut Ga Al B 7
Pb Sn Fl Ge Si C 2
Ba Sr Ra He Ca Ub Be Mg 2
Bi Sb Uup As P N 3
Rn Xe Uuo Kr Ar Ne 6
Pu Sm 6
Cm Gd 8
Hs Os Ru Fe 6
Ds Pt Pd Ni 8
Fm Er 12
Cf Dy 10
Sg W Mo Cr 4
Rf Hf Zr Ti 2
Cn Hg Cd Zn 10
U Nd 4
Th Ce 2
No Yb 14

Tetryonics 52.02 - Periodic Table [Elements]

Response to an analysis of Tetryonic Chemistry by Robert J Brotherus [July 2015]

In reply to a paper outlining Mr Robert Brotherus' examination and *mis*-understanding of Tetryonic theory and in particular how it applies to the modern understanding of Chemistry as it currently taught…..

Point by point I'll firstly address the particular points of concern raised wrt modern chemical theory and physics in general:

Tetryonics paper on quantum chemistry is very scarce on making any predictions about atomic properties or comparing such predictions to experimental results.

Tetryonic theory provides many predictions regarding the physical 2d planar **geometrie**s of bosons [radiant mass-energies] and the 3D material **Topologies** of Fermions [Matter and their associated physical properties]. Inclusive in these are many predictions and repeated statements and illustrations of specific particle shapes and the values of mass-energy momenta attributed to each… as well as how these values stem from the equilateral tessellation of Planck energy momenta and why they differ from the established values of modern scientific theory….

In many cases there are similarities to be found with existing observations but where Tetryonic theory differs from modern theory it is not by choice but rather dictated by the equilateral mass-energies within charged Matter topologies themselves and all too often led to much effort in order to reconcile the differences, leading to further insights and extensions of the theory over time…. Where differences exist it is deliberate not by error.

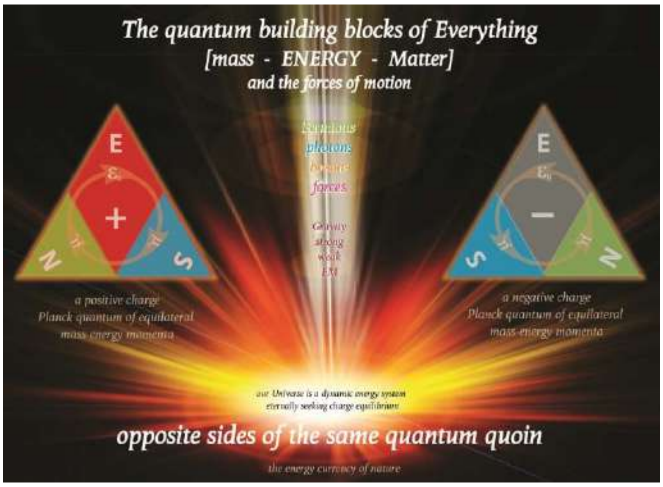
The quantum building blocks of Everything
[mass - ENERGY - Matter]
and the forces of motion
Fermions
photons
bosons
forces
Gravity
strong
weak
EM
E
+
N
S
a positive charge
Planck quantum of equilateral
mass energy momenta
E
−
S
N
a negative charge
Planck quantum of equilateral
mass-energy momenta
our Universe is a dynamic energy system
eternally seeking charge equilibrium
opposite sides of the same quantum quoin
the energy currency of nature

Tetryonics is a BOTTOM-UP physical theory that unifies many currently disparate theories of science into a coherent unified model of the same – but not without its own set of mental challenges to those who are first becoming acquainted with it. As such it must be recognised that Tetryonic theory should not be studied without a sound understanding of the foundation upon with it is built…. Many have approached analysis of my work in such a way only to find that their analysis was limited by modern science's lack of precise detail of the foundations of the physics that they attempt to explain [to wit modern theory is a top-down explanation of the science build up over many generations with an equal number of varying interpretations and assumptions]

Tetryonics plays down importance of "mathematics" and instead emphasizes "geometry".

Arguably I place a lot of emphasis on geometry over that of mathematics in Tetryonic theory, but only so as it offers insights into physical theory that maths alone has not been able to provide since its wholesale incorporation into the science in the 1600’s. This is not to downplay the importance and productiveness that maths has provided us a scientific tool, but rather to ‘illustrate’ and draw attention to the advances that a purely geometric model has to offer us in advancing our understanding of the same…. To wit, Tetryonic geometries are based on the postulate that all physical theory is founded on equilateral Planck quanta of energy momenta and as such these physical entities afford no room for speculative ideas or mathematical manipulations beyond what is physically possible to be constructed.

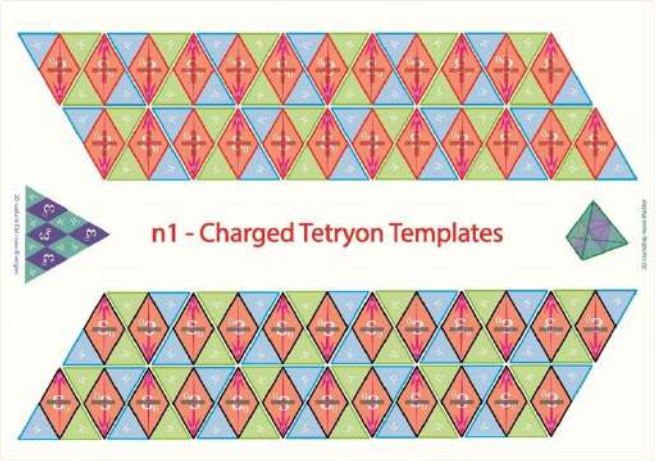

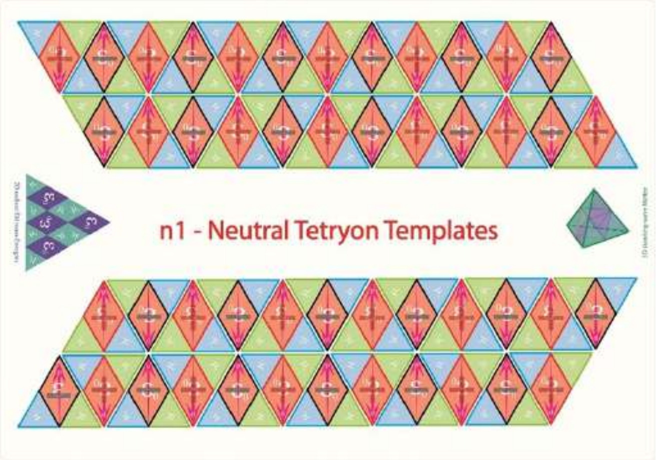

To help elucidate my work and thoughts on the differentiation of 2d planar mass-energy geometries from 3D Material topologies I recorded a detailed YouTube video titled “Tetryonics 101” where I show HOW to use the existing Tetryonic templates [of Planck mass-energy momenta] to construct all the fields [boson & photons] of the Standard model along with the 3D Mater topologies of all the known fermions [and them some]… I do hope you’ll take the time to review this detailed instruction on the creation of quantum fields and Matter and try to replicate them same yourself – it is THE most informative thing students of Tetryonics and science in general can do when trying to increase their understanding of the hitherto unseen constituents that make up quantum theory and our macro-scale Universe and the laws of Nature we observe….

Tetryonics paper on physical chemistry is scarce on such numerical predictions and where it does make such predictions they are often in contradiction to observed experimental facts.

In addition to the maths found within Tetryonic illustrations there are also many spreadsheets developed from those very same equations over time to provide the numerical detail found in the illustrations – particular wrt periodic elements and theory rest atomic masses and kinetic elements etc. – It is by no means exhaustive on every topic imaginable but they do address the information provided on particles & elements etc. along with their kinetic and relativistic energies of motion.

Tetryonic Problem: Missing hydrogen element

Hydrogen is not missing from Tetryonic theory – in fact it is extensively detailed and modelled in Tetryonics [1] – Quantum Mechanics eBook [again highlight the need not to jump into one's preferred field of expertise and study but rather learn Tetryonic theory as you would any modern physical theory in order of presentation]

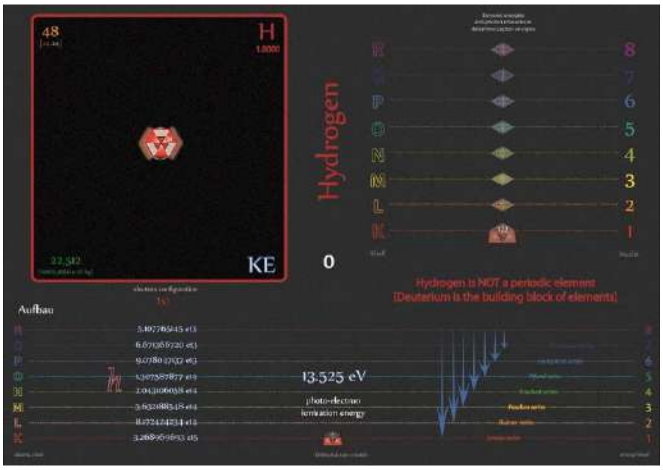
48
H
1.0000
Hydrogen
KE
0
K L M N O P Q R
1 2 3 4 5 6 7 8
Hydrogen is NOT a periodic element
[Deuterium is the building block of elements]
Aufbau
13.525 eV
photo-electron
ionisation energy

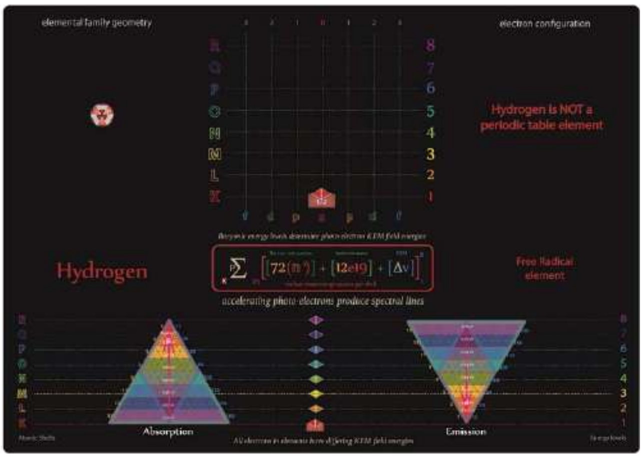
elemental family geometry
electron configuration
Hydrogen is NOT a periodic table element
Hydrogen
Free Radical element
accelerating photo electrons produce spectral lines
Absorption
Emission

Hydrogen is designated as element 00 in Tetryonic theory as it is **NOT** the building block of periodic elements as is currently understood but rather a sub-unit of Deuterium the real building block of all elements.... This is not an arbitrary decision but one forced on anybody who attempts to build physical models of all elements in order to match and explain many of the current established aspects of chemical and quantum theory wrt the quantum electrodynamics of elements themselves wrt molar mass, the Aufbau principle, Schrodinger and Bohr numbers for each element and how Deuterium nuclei [or any other] may bind together in order to form all the elements of the periodic table and their respective spectral emissions lines.

Attempting to build heavy elements utilising Z number Hydrogen nuclei as a foundation with the excess molar masses being accounted for by way of 'extra or excess' Neutrons [as is the established model of modern chemistry] runs into a number of problems immediately – namely:

The lack of residual strong Force binding points on nuclei to facilitate the binding of Neutrons to the nuclei in order to form the elements [particularly elements 60+]

The incorrectness of ‘reserve beta decay’ to explain and model Neutron particle formation and the mechanics of electron binding to periodic elements in accordance with Schrodinger’s wave equation and Bohr’s atomic model of the same in order to produce observed spectral line emissions of each element in the PT.

There is no uncertainty or indeterminacy in Tetryonic theory as there is in modern theory to explain away such inconsistencies between various aspects of physical theory – each attribute MUST be modelled and explained from first principles & invariably leads to new insights into quantum chemical theory that are at odds with established lines of thought.

To wit I stand by the claims and statements made in my eBooks :

> Each elemental nuclei is made from Deuterium" (p. 12)
> “120 elements in 8 quantum levels *and hydrogen as a free radical*.” (p. 26)
> “Deuterium (not hydrogen) is the building block of all elements." (p. 21)
> “Hydrogen is NOT a periodic element” (p. 83)

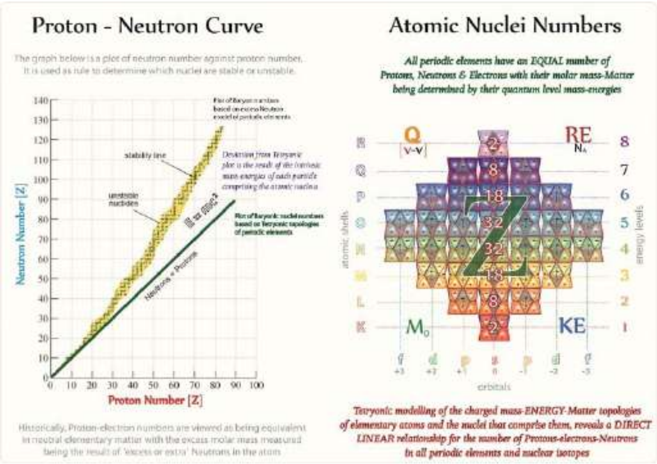
Proton - Neutron Curve

The graph below is a plot of neutron number against proton number. It is used as rule to determine which nuclei are stable or unstable.

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes
Neutron Number [Z]
Proton Number [Z]
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
stability line
unstable nuclides
Plot of Baryonic numbers based on excess Neutron model of periodic elements
Deviation from Tetryonic plot is the result of the intrinsic mass-energies of each particle comprising the atomic nucleus
E = mc²
Plot of Baryonic nuclei numbers based on Tetryonic topologies of periodic elements
Neutrons = Protons
Q
[V-V]
RE
N_A
M_0
KE
2
8
18
32
32
18
8
2
Z
R
Q
P
O
N
M
L
K
atomic shells
8
7
6
5
4
3
2
1
energy levels
f
d
p
s
p
d
f
+3
+2
+1
0
-1
-2
-3
orbitals

As it is the ONLY way to model and explain how all the periodic elements are created from equilateral Planck energy momenta that is in accord with the known observables… they are not in agreement with some aspects of established theory BUT they do closely mirror the observables and once KEM fields of thermal heat, kinetic motion etc. are taken into account provide a very accurate quantum model of chemical processes at the quantum scale.

Tetryonics does not throw all these facts out of the window just because hydrogen does not fit neatly into my [or other] pre-conceived set of boxes and triangles and hexagons, they are calculated from the Material models and immaterial fields associated with each element that MUST grow from first principles… not from assumptions elaborated upon over time.

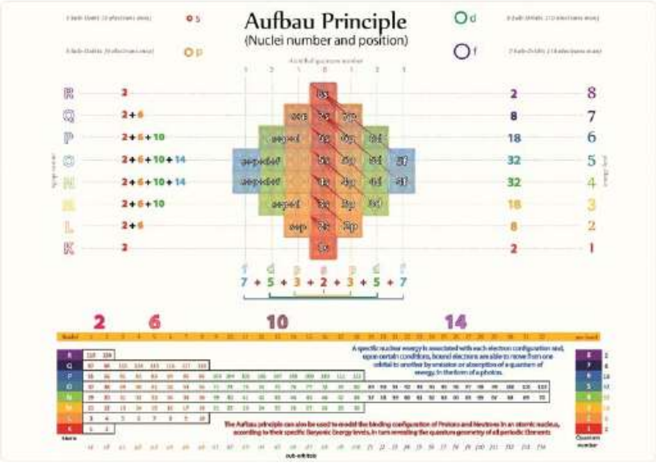
Aufbau Principle
(Nuclei number and position)
s
p
d
f
R
Q
P
O
N
M
L
K
2
2+6
2+6+10
2+6+10+14
2+6+10+14
2+6+10
2+6
2
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
7 + 5 + 3 + 2 + 3 + 5 + 7
2
6
10
14
A specific nuclear energy is associated with each electron configuration and, upon certain conditions, bound electrons are able to move from one orbital to another by emission or absorption of a quantum of energy, in the form of a photon.
The Aufbau principle can also be used to model the binding configuration of Protons and Neutrons in an atomic nucleus, according to their specific Baryonic Energy levels, in turn revealing the quantum geometry of all periodic Elements
Quantum number
sub-orbitals

In fact the Aufbau table is shown in Tetryonic theory to be a table that has been 'right justified' to match the nomenclature of mathematics rather than the geometric topology of reality [which is centred about the principle quantum numbers and shells of Schrodinger and Bohr respectively]… it is only our skewed perspective of maths that prevents most from seeing the real symmetry present in atomic nuclei as detailed in the quantum numbers of Schrodinger's wave equation for the same … ie there are no g,h,I,j, shells as supposed and all the quantum numbers of Schrodinger's wave equations match the symmetric topology of elements as geometric portrayed in Tetryonics [and are in agreeance with Bohr's shells and orbitals as well, thus unifying all 3 theories]

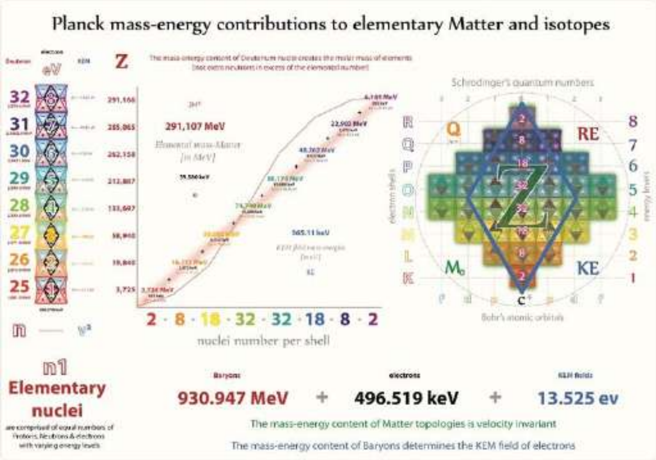
Planck mass-energy contributions to elementary Matter and isotopes
eV
Z
32
31
30
29
28
27
26
25
291,166
285,065
262,158
212,807
133,607
58,940
19,840
3,725
291,107 MeV
Elemental mass-Matter [in MeV]
565.11 keV
2 · 8 · 18 · 32 · 32 · 18 · 8 · 2
nuclei number per shell
Schrodinger's quantum numbers
R
Q
P
O
N
M
L
K
Q
RE
M₀
KE
Z
c²
8
7
6
5
4
3
2
1
electron shells
energy levels
Bohr's atomic orbitals
m
v²
m1
Elementary nuclei
Baryons
electrons
KEM fields
930.947 MeV + 496.519 keV + 13.525 ev
The mass-energy content of Matter topologies is velocity invariant
The mass-energy content of Baryons determines the KEM field of electrons

The material topologies of Tetryonic theory then also lead to the necessity to correct 'weighted' values for all known particles and elements [and then some] in light of the quantum mechanics of the same – ie their identical 3D charge topologies and the mechanics of the quantum rotating synchronous convertors [Deuterium nuclei] that they form in deference to the accepted theory of mini-solar systems and Neutrons that are heavier than their identical charged counterparts [Protons and anti-Protons]

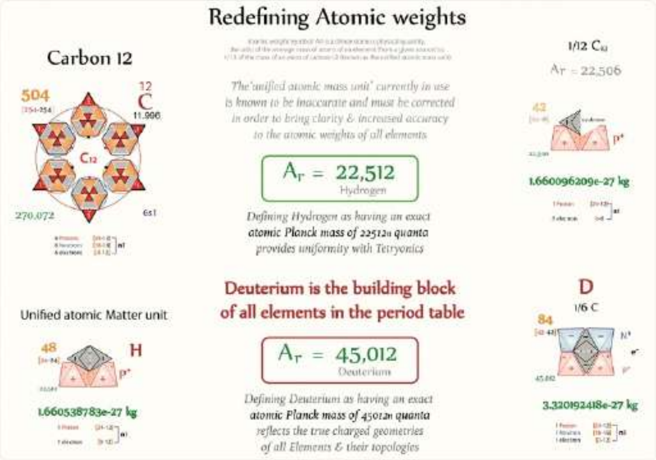

Redefining Atomic weights

The 'unified atomic mass unit' currently in use is known to be inaccurate and must be corrected in order to bring clarity & increased accuracy to the atomic weights of all elements

$A_r = 22{,}512$ Hydrogen

Defining Hydrogen as having an exact atomic Planck mass of 22512n quanta provides uniformity with Tetryonics

Deuterium is the building block of all elements in the period table

$A_r = 45{,}012$ Deuterium

Defining Deuterium as having an exact atomic Planck mass of 45012n quanta reflects the true charged geometries of all Elements & their topologies

Carbon 12
504
12
C
11.996
C12
270.072
6s1
1/12 C12
Ar = 22,506
42
p+
1.660096209e-27 kg
Unified atomic Matter unit
48
H
p+
1.6605387830-27 kg
D
1/6 C
84
N0
e-
p+
45.012
3.320192418e-27 kg

Point in case being **Tetryonic Problem: Incorrect Proton and Neutron masses**
As detailed in illustration 49.04 & 49.05 respectively in my eBook… Tetryonics demands an exact mass for Planck quanta and the molar masses [at absolute zero] for each and every particle in our Universe for modelling and any additional energy momenta [by way of thermal, spectral, kinetic or observational energies etc.] are accounted for in the KEM fields of each respective element or particle.

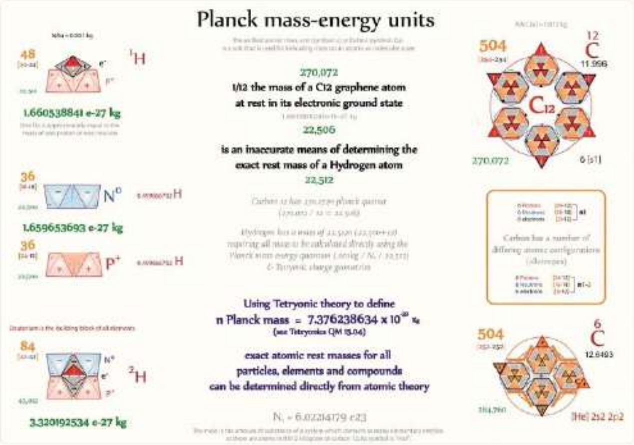

Planck mass-energy units

270,072
1/12 the mass of a C12 graphene atom at rest in its electronic ground state
22,506
is an inaccurate means of determining the exact rest mass of a Hydrogen atom
22,512

Using Tetryonic theory to define n Planck mass = 7.376238634 x 10⁻⁵⁰ kg (see Tetryonics QM 15.04)
exact atomic rest masses for all particles, elements and compounds can be determined directly from atomic theory

N_A = 6.02214179 e23

48
e-
1H
p+
1.660538841 e-27 kg
36
N0
H
1.659653693 e-27 kg
36
P+
H
84
N0
e-
2H
p+
3.320192534 e-27 kg
504
12
C
11.996
C12
270.072
6 [s1]
504
6
C
12.6493
[He] 2s2 2p2

There is NO room for a 'weighted' atomic mass as utilised in modern chemistry which conveniently ignores the mass and motional energies of the electrons in the elements or which ignores the KEM field energies noted above for 'most' situations and calculations – Tetryonic theory MUST account for all of these energies even if it does so at the risk of disagreeing with modern theory at times…

The Proton and Neutron MUST have identical atomic masses as they are identical – both are 36π Baryon Matter topologies differing only in the net charges of +12 [24-12] vs 0 [18-18] they possess as a result of their constituent quarks [UDU vs DUD] as extensively detailed in my QM eBook… Neutrons are NOT created through the process of 'reverse' beta-decay as is supposed presently supposed in modern theory and accordingly do NOT have a slightly higher mass than Protons which in turn affects the molar mass calculations for all periodic elements and isotopes as outlined in my Chemistry eBook as well at length.

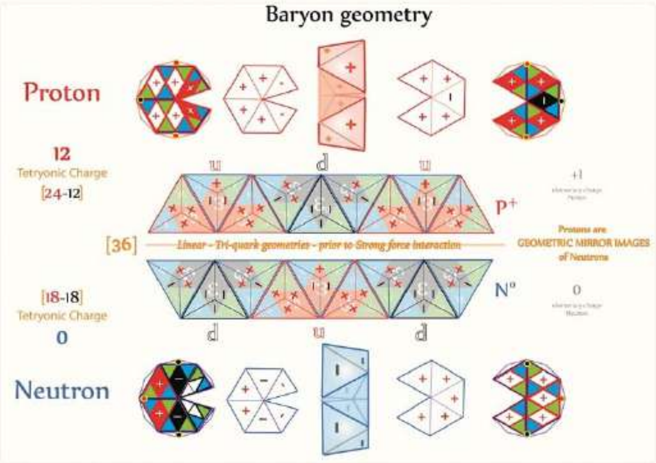
Baryon geometry
Proton
12
Tetryonic Charge
[24-12]
u
d
u
P+
+1
[36]
Linear - Tri-quark geometries - prior to Strong force interaction
Protons are GEOMETRIC MIRROR IMAGES of Neutrons
[18-18]
Tetryonic Charge
0
d
u
d
N°
0
Neutron

Ie Carbon 13, 14 and upwards are not the result of 'extra' Neutrons being bound to the Carbon nuclei – such a nuclei if possible would be stable and not decay as observed – but rather they are simply the result of the inherent 'squared' equilateral mass-energies of each Baryon in the nuclei being increased from n1 to n2 to n3 …. Resulting in a net mass increase that 'appears' to be the result of a number of 'extra' Neutrons in the nuclei [again this is detailed in my Chemistry eBook and spreadsheets of the calculations be easily created from the formulae found in chapter 53 on my Chemistry eBook….

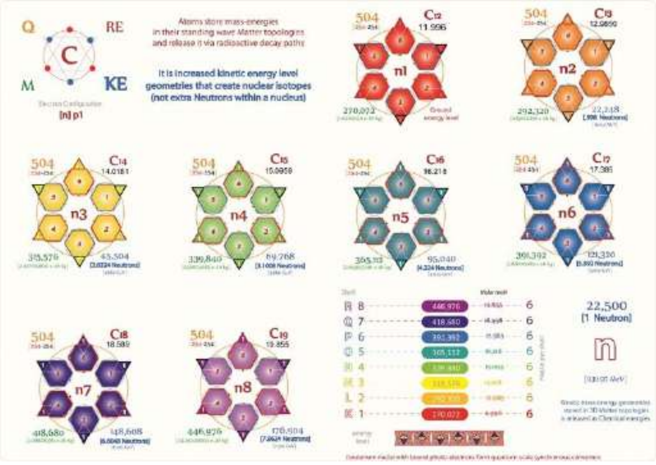
Q
RE
C
M
KE
[n] p1
Atoms store mass-energies in their standing wave Matter topologies and release it via radioactive decay paths
It is increased kinetic energy level geometries that create nuclear isotopes (not extra Neutrons within a nucleus)
504
C12
11.996
n1
270,072
Ground energy level
C13
12.9850
n2
292,320
22,248
C14
14.0181
n3
315,576
45,504
C15
15.0959
n4
339,840
69,768
C16
16.218
n5
365,112
95,040
[4.224 Neutrons]
C17
17.385
n6
391,392
121,320
C18
18.589
n7
418,680
148,608
C19
19.855
n8
446,976
176,904
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
446,976
418,680
391,392
365,112
6
22,500
[1 Neutron]
n

This 'significant discrepancy' with Tetryonics prediction of Proton and Neutron masses from observed ones alone will be enough to prove Tetryonic theory a worthy successor to modern quantum chemistry theory – unless you can explain all of the above using 'excess' Neutrons.

I am neither ignorant of the basic masses of basic particles nor do I think that Tetryonic theory and observed facts of nature disagree, but rather the established theoretical 'facts' of modern chemical theory must be in error

I agree that in reality Matter has **3D structure** (including **volume**) hence the terminology of Matter **topologies** throughout my work. To portray my planar illustration of elements and particles in my PDF eBooks as not reflecting their true material topologies is grossly inaccurate and reflects that fact that **YOU have NOT read all my work** including Tetryonics 101, the Google sketch warehouse of 3D CAD models of each and every particle [kindly developed by Rene Cormier over many months of effort] and extensive list of all elements and their charged 3D topologies and electronic wave function properties and numbers which are freely available to all – only planar 2 d mass-energy **geometries** [bosons and photons] are portrayed as being like thin 2D sheets in Tetryonic theory, something I am sure you will become more familiar with as your understanding of T-theory advances,[and your distain for its predictions and values not in accord with modern theory wains over time].

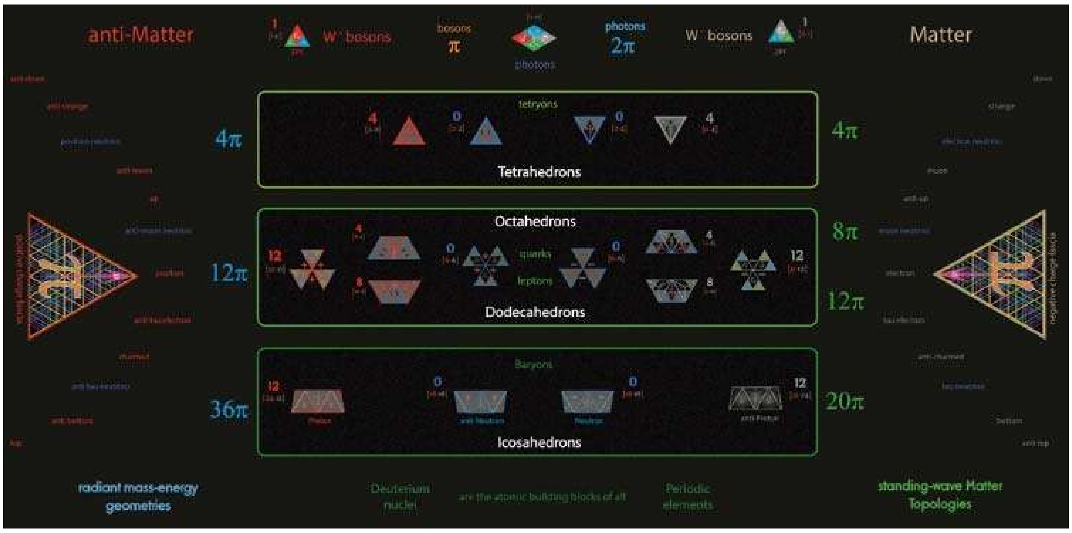
anti-Matter
W bosons
bosons
π
photons
photons
2π
W bosons
Matter
tetryons
4π
Tetrahedrons
4π
Octahedrons
8π
12π
quarks
leptons
Dodecahedrons
12π
Baryons
36π
Icosahedrons
20π
radiant mass-energy geometries
Deuterium nuclei
are the atomic building blocks of all
Periodic elements
standing-wave Matter Topologies

Regarding the ‘separation ‘ of electrons from the atomic nucleus – this is an antiquated model of quantum electrodynamics that is fashioned on planets and their satellites [ie the mini-solar system Bohr model] and which has been extended to a fuzzy electron orbital model to incorporate the Schrodinger quantum model and the ‘inherent’ quantum uncertainly/indeterminacy of statistical probabilities intrinsic to quantum theory as the theories developed over time…. Tetryonics makes it quite clear in T[1] – QM that the atomic nucleus is in fact a quantum synchronous convertor [ala Tesla/Westinghouse] that can store and release mass-energies in a controlled and predictable manner [ala isotopic decay and spectral line emissions] and that quantum mechanics is NOT quantum inherently uncertain nor in determinant as it is often portrayed – in fact quantum theory as explained by Tetryonic theory is very much deterministic in every way – it was only our top-down view of the mechanics that led modern science to this erroneous conclusion, which I have attempted to correct through Tetryonic theory.

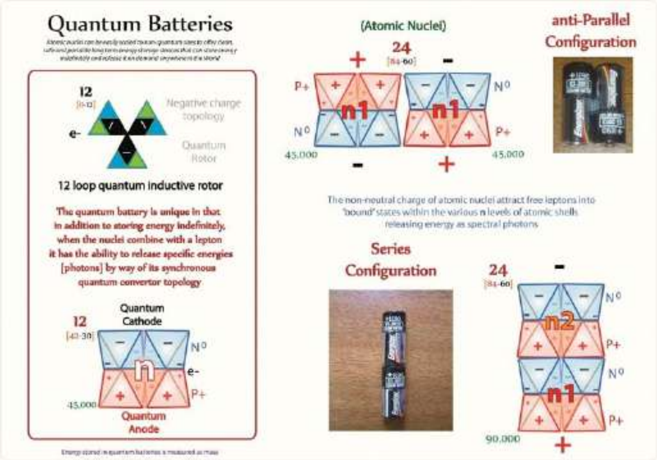
Quantum Batteries
12
e-
Negative charge topology
Quantum Rotor
12 loop quantum inductive rotor
The quantum battery is unique in that in addition to storing energy indefinitely, when the nuclei combine with a lepton it has the ability to release specific energies [photons] by way of its synchronous quantum convertor topology
12
Quantum Cathode
N0
e-
P+
n
45,000
Quantum Anode
(Atomic Nuclei)
24
+
-
P+
N0
n1
n1
N0
P+
45,000
45,000
anti-Parallel Configuration
The non-neutral charge of atomic nuclei attract free leptons into 'bound' states within the various n levels of atomic shells releasing energy as spectral photons
Series Configuration
24
N0
n2
P+
N0
n1
P+
90,000

The electron does not orbit the atomic nucleus of any element but rather is bound via its magnetic dipole moment to the positively charged nucleus and spins akin to the rotating rotor of quantum convertor partially within the 3D Matter topology of Deuterium nuclei….

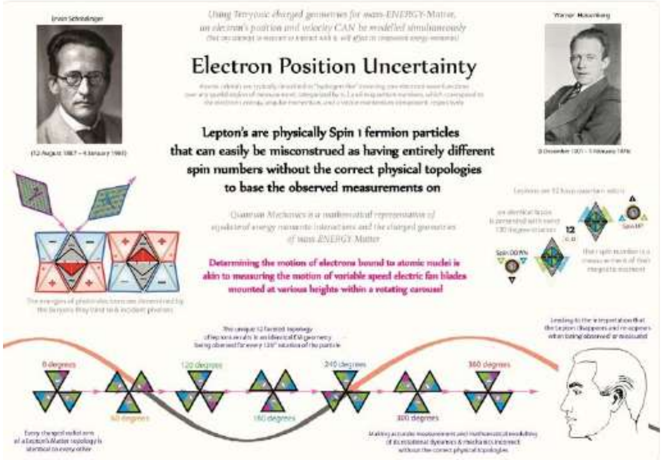

Using Tetryonic charged geometries for mass-ENERGY-Matter, an electron's position and velocity CAN be modelled simultaneously

Electron Position Uncertainty

Lepton's are physically Spin 1 fermion particles that can easily be misconstrued as having entirely different spin numbers without the correct physical topologies to base the observed measurements on

Quantum Mechanics is a mathematical representation of quantised energy momenta interactions and the charged geometries of mass-ENERGY-Matter

Determining the motion of electrons bound to atomic nuclei is akin to measuring the motion of variable speed electric fan blades mounted at various heights within a rotating carousel

Erwin Schrodinger
Werner Heisenberg
12
Spin UP
Spin DOWN
0 degrees
60 degrees
120 degrees
180 degrees
240 degrees
300 degrees
360 degrees

Again these models have all been built and operate according to established electrodynamic theory and their operation is well understood in the macro sense [but obviously the acceptance of such a model strains the thoughts of most who have grown up a mental picture of fuzzy electrons orbiting a spherical nuclei]

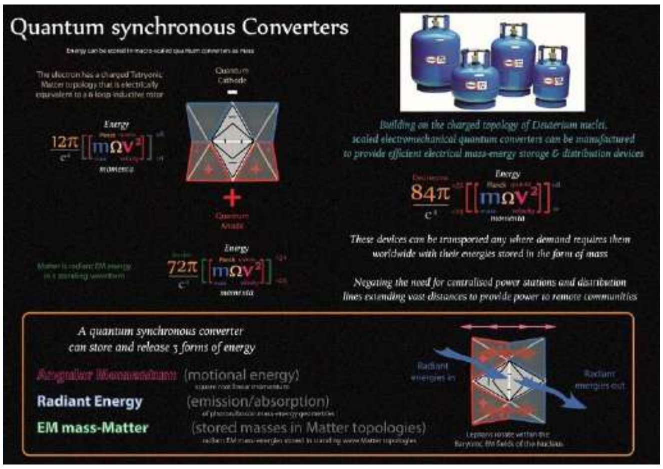

Quantum synchronous Converters

Energy can be stored in macro-scaled quantum converters as mass

The electron has a charged Tetryonic Matter topology that is electrically equivalent to a 6-loop inductive rotor

Building on the charged topology of Deuterium nuclei, scaled electromechanical quantum converters can be manufactured to provide efficient electrical mass-energy storage & distribution devices

These devices can be transported any where demand requires them worldwide with their energies stored in the form of mass

Negating the need for centralised power stations and distribution lines extending vast distances to provide power to remote communities

A quantum synchronous converter can store and release 3 forms of energy

Angular Momentum (motional energy)

Radiant Energy (emission/absorption)

EM mass-Matter (stored masses in Matter topologies)

Quantum Cathode
Quantum Anode
Energy
$\frac{12\pi}{c^4}[[m\Omega v^2]]$
momenta
$\frac{72\pi}{c^4}[[m\Omega v^2]]$
momenta
$\frac{84\pi}{c^4}[[m\Omega v^2]]$
momenta
Radiant energies in
Radiant energies out

Again the electronic structure of each and every one of the 120 possible periodic elements is well detailed wrt Bohr shells & orbitals and unified wrt their respective Schrodinger numbers for the same in my Chemistry eBook for all to view…. Unlike modern texts which normally stop at element 20 and claim that it is just too difficult to show the quantum models of larger elements with their evermore complex electron configurations and Schrodinger numbers

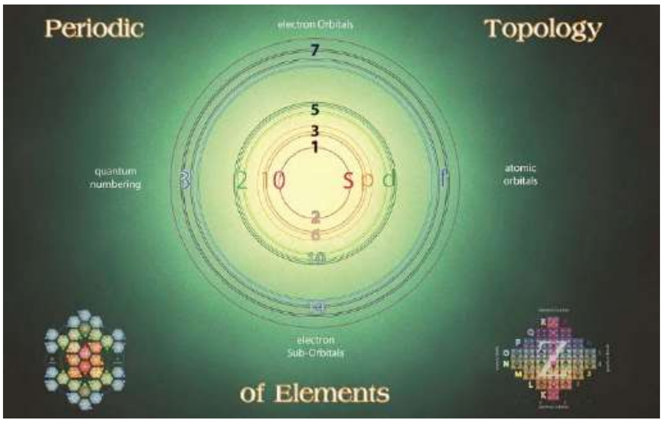

Periodic
Topology
of Elements
electron Orbitals
7
5
3
1
quantum numbering
3
2
1
0
s
p
d
f
atomic orbitals
2
6
10
14
electron Sub-Orbitals

The s.p.d & f orbitals of electrons ‘around’ any elementary nuclei is the result of the atomic nuclei rotating about is centre of mass/gravity [classical angular momentum] and the electrons tracing out easily determined and well defined concentric orbits about the same centre [as illustrated] …. A close examination of the Schrodinger numbers [or in fact Bohr orbits] of heavier/larger Z number atomic nuclei reveals the there are no g,h I orbits etc as often postulated but in fact there is a reversal of electron orbital distance as the element number and principal energy levels increases….. Resulting in the periodic symmetry depicted in Tetryonic theory of an atomic nucleus resembling a child's spinning top, which could be mistaken for a spherical topology from our macro view of the quantum scale.

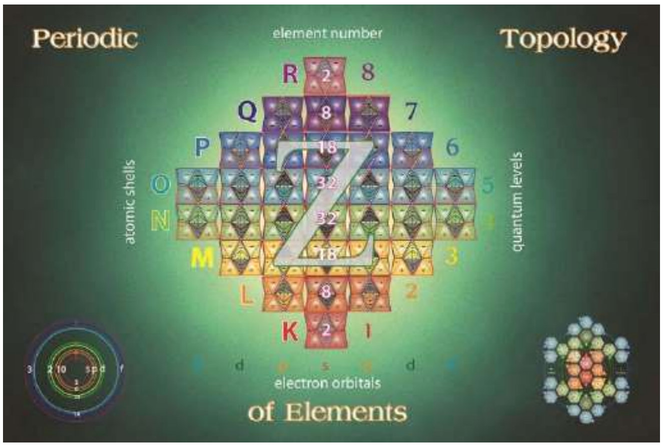
Periodic
Topology
element number
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
Z
atomic shells
quantum levels
electron orbitals
of Elements

With each element's Z number reflecting the number of **Deuterium** atoms in each nuclei, *not the number of Hydrogen atoms with an ever increasing number of Neutrons to account for observed molar masses*

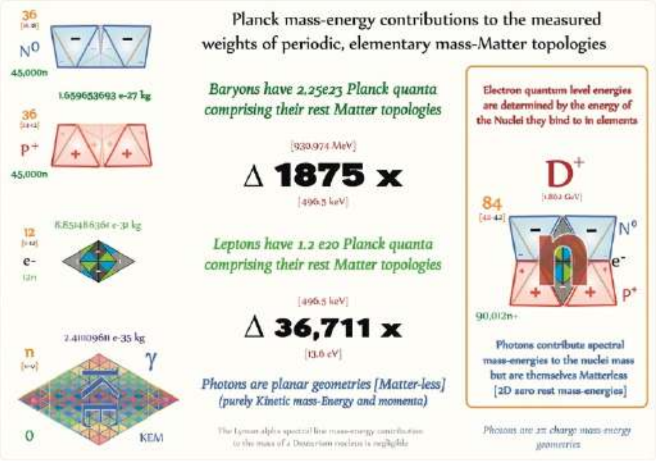
Planck mass-energy contributions to the measured weights of periodic, elementary mass-Matter topologies
36
N0
45,000m
1.659653693 e-27 kg
36
P+
45,000m
12
e-
6.851486364 e-31 kg
n
2.4110968 e-35 kg
γ
0
KEM
Baryons have 2.25e23 Planck quanta comprising their rest Matter topologies
[930.974 MeV]
Δ 1875 x
[496.5 keV]
Leptons have 1.2 e20 Planck quanta comprising their rest Matter topologies
[496.5 keV]
Δ 36,711 x
[13.6 eV]
Photons are planar geometries [Matter-less] (purely Kinetic mass-Energy and momenta)
Electron quantum level energies are determined by the energy of the Nuclei they bind to in elements
D+
84
N0
e-
P+
90,012m
Photons contribute spectral mass-energies to the nuclei mass but are themselves Matterless [2D zero rest mass-energies]
Photons are zero charge mass-energy geometries

The electrons associated with each atomic element's structure in fact have an important and direct bearing on the spectral emissions of each particular element. Even though their mass-energies of their electronic topologies are only 0.000533' % of the mass-energies of each Baryon in any atomic nuclei this mass-Matter differential along with the energy momenta of the electron's KEM field is key to facilitating the absorption and release of spectral line emissions particular to each and every element known… again, it is not simple good enough to conveniently ignore electrons ad hoc in scientific theory to suit a particular point or theory just because they are or 'insignificant' mass compared to the whole nucleus. Any successful model of science and physics in general must be able to account for ALL the known aspects of modern observations not just a select few that suit their purposes in order to unify physics on all energy scales. Ie a new model of the quantum structure, topology and mechanics of atomic nuclei must account for charge, atomic and chemical interactions, spectral line series and all known elements and compound creation in the ONE physical model.

Clarifying question about Tetryonics nuclear vs. atomic size wrt the Zenon atom
I humbly submit that illustrations 51.49

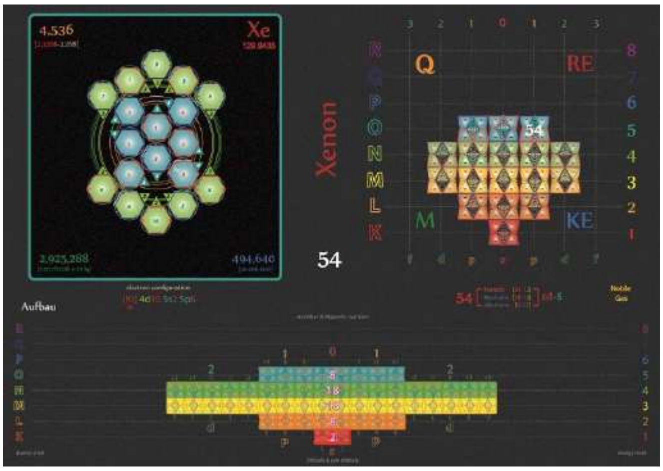
4,536
Xe
Xenon
2,925,288
494,640
54
Q
RE
M
KE
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1

And 53.49 respectively along with the 3D CAD illustrations of the same

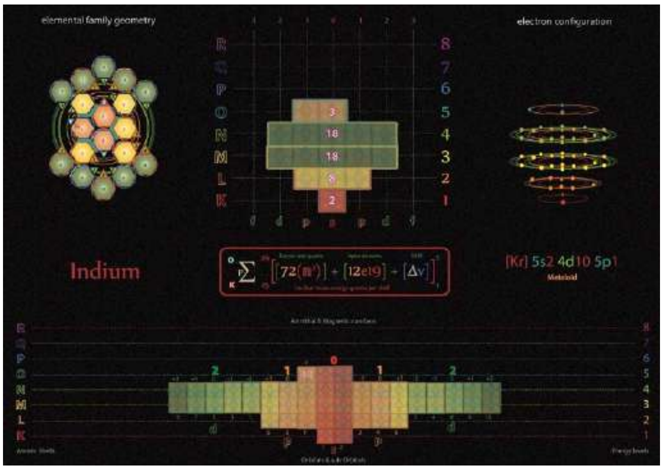
elemental family geometry
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
18
18
8
2
Indium
[Kr] 5s2 4d10 5p1
Metaloid

Should be sufficient to explain and illustrate all properties of the Zenon atom [or any other you chose] including their rest mass-Matter energies [at absolute zero], their charged Matter topology in accord with Bohr shells/orbitals and Schrodinger numbers and the kinetic mass-energies of their bound electrons.

Clarifying question 2 about Tetryonics nuclear vs. atomic size

According to Tetryonics the proton and neutron in Deuterium nuclei touch one another at the point son their respective charge topologies where the e-field components of their constituent quarks are located - this is indicated on all Tetryonic element illustrations via red [for positive] focus points and black [for negative] focus points.

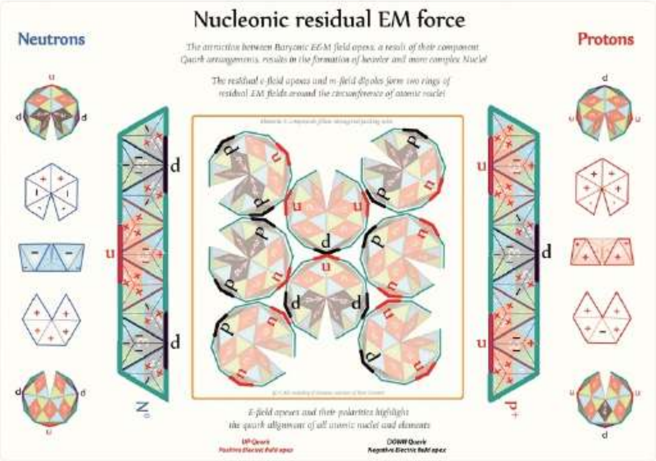
Nucleonic residual EM force
Neutrons
Protons
The attraction between Baryonic E&M field apexs, a result of their component Quark arrangements, results in the formation of heavier and more complex Nuclei
The residual e-field apexs and m-field dipoles form two rings of residual EM fields around the circumference of atomic nuclei
u
d
N0
P+
E-field apexes and their polarities highlight the quark alignment of all atomic nuclei and elements
UP Quark
Positive Electric field apex
DOWN Quark
Negative Electric field apex

But since Tetryonics does not include any separate strong nuclear force but just electromagnetics, all particles of Tetryonic nucleus would fly apart in a nanosecond.

Again, this misunderstanding and erroneous statement about Tetryonic theory seems to stem from you having failed to read Tetryonics [1] – Quantum Mechanics, viewed my Tetryonics 101 instruction video or even attempted to build and understand how atomic nuclei are formed and come together to form larger elements and compounds alike prior to commenting ….

In my first eBook I have gone to great lengths to show that the STRONG force does indeed exist and in fact plays a crucial role in bringing tetryons [the tetrahedral quantum of Matter topologies themselves] together to form leptons, quarks, mesons & Baryons – it is the parallel opposite charge interaction present between the planar charge faces of each tetryon within each and every Baryon [not to be mistaken for the residual strong force]

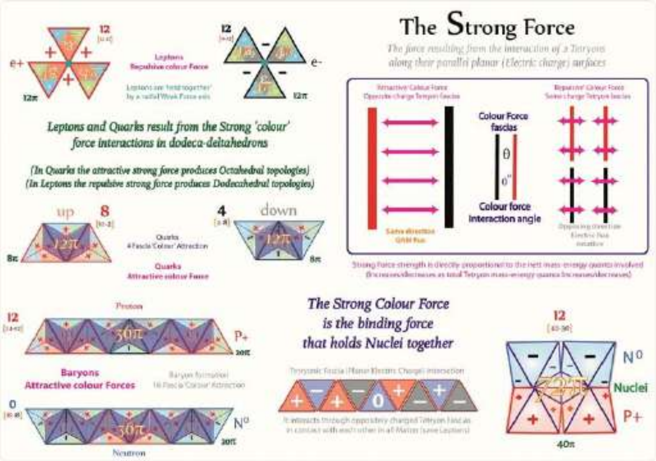
The Strong Force
The force resulting from the interaction of a Tetryons along their parallel planar (Electric charge) surfaces
e+
e-
12
12π
Leptons
Repulsive colour Force
Leptons and Quarks result from the Strong 'colour' force interactions in dodeca-deltahedrons
(In Quarks the attractive strong force produces Octahedral topologies)
(In Leptons the repulsive strong force produces Dodecahedral topologies)
up
8
4
down
12π
8π
Quarks
Attractive colour Force
Colour Force fascias
θ
0°
Colour force interaction angle
Proton
P+
30π
Baryons
Attractive colour Forces
N0
Neutron
The Strong Colour Force is the binding force that holds Nuclei together
Nuclei
40π

Obviously this co-planar opposite charge interaction holds all atomic nuclei together and your claim of Tetryonic particles flying away in a nanosecond just doesn't hold water…

In fact just place two sheets of glass together with a thin film of water between them and see how easy it is to pull those sheets apart by applying a force orthogonal to the co-planar sheets of glass and you'll soon understand the STRONG force much better that you do presently.

This charge interaction between the e-fields of quarks in Baryons serves to not only bring positive charge Protons together with neutral Neutrons [something modern theory cannot explain] but also serve to orientate each baryon to the other so as to facilitate electron bonding and the creation of ever larger atomic elements and compounds [the residual strong or EM force as it is often termed or the London force]

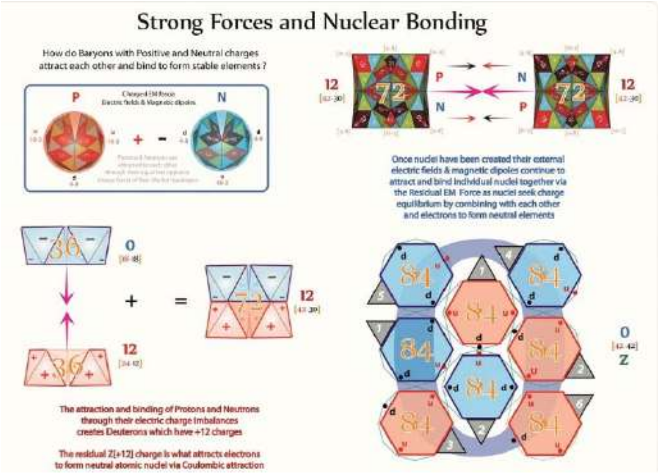
Strong Forces and Nuclear Bonding
How do Baryons with Positive and Neutral charges attract each other and bind to form stable elements ?
P
N
Charged EM forces
Electric fields & Magnetic dipoles
+
-
12
72
P
N
N
P
72
12
Once nuclei have been created their external electric fields & magnetic dipoles continue to attract and bind individual nuclei together via the Residual EM Force as nuclei seek charge equilibrium by combining with each other and electrons to form neutral elements
36
0
+
=
72
12
36
12
84
0
Z
The attraction and binding of Protons and Neutrons through their electric charge imbalances creates Deuterons which have +12 charges
The residual Z[+12] charge is what attracts electrons to form neutral atomic nuclei via Coulombic attraction

Yes I have an 'obsession' with all atomic nuclei have equal numbers of Protons, electron AND Neutrons in accord with their Z number, just as you have an obsession with them being Hydrogen atoms with excess Neutrons 'bound' to them – but for the reasons highlighted above, and many others – as for the quantum structure of Protium (regular hydrogen), Deuterium and Tritium being 'missing' - again my T[1] - QM eBook holds the answers

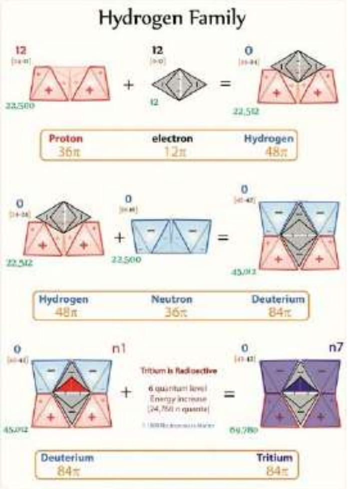
Hydrogen Family
12
12
0
+
=
22,500
12
22,512
Proton
36π
electron
12π
Hydrogen
48π
0
0
0
+
=
22,512
22,500
45,012
Hydrogen
48π
Neutron
36π
Deuterium
84π
0
n1
0
n7
Tritium is Radioactive
6 quantum level
Energy increase
(24,768 n quanta)
+
=
45,012
69,780
Deuterium
84π
Tritium
84π

With additional space being made to highlight the Tetryonic genesis of Tritium & Helium 3

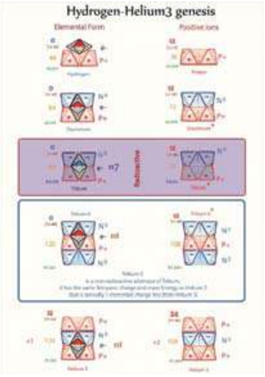
Hydrogen-Helium3 genesis
Elemental Form
Positive Ions
n7
Radioactive

All isotopes are simply Z number nuclei with the energy levels of their Baryons raised to the next quantum energy level [n1-n8] in accord with the 3D material symmetry of the atomic nuclei themselves and the Bohr shells/Schrodinger principal quantum numbers, which can occur either through natural energy absorption or the artificial increase of their mass-energy content in the lab….

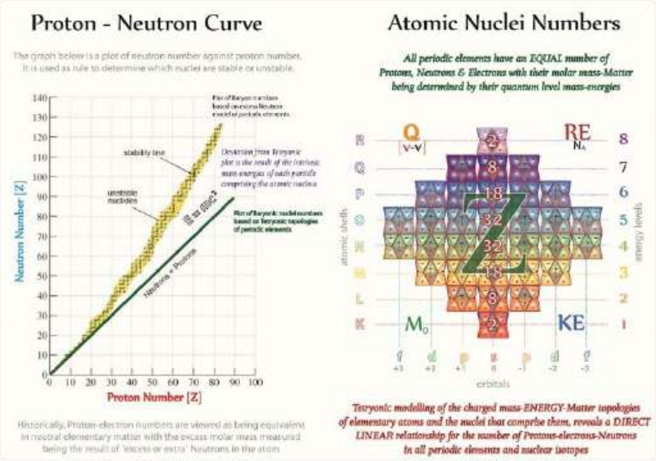

Obviously with all Deuterium nuclei being quantum synchronous convectors as the energy levels of their Baryons are increased [or decreased] the electrons bound to them respond in kind [hence synchronous] via spectral line emissions [or absorption] ….all covered in Tetryonics [2] – Quantum Electrodynamics.

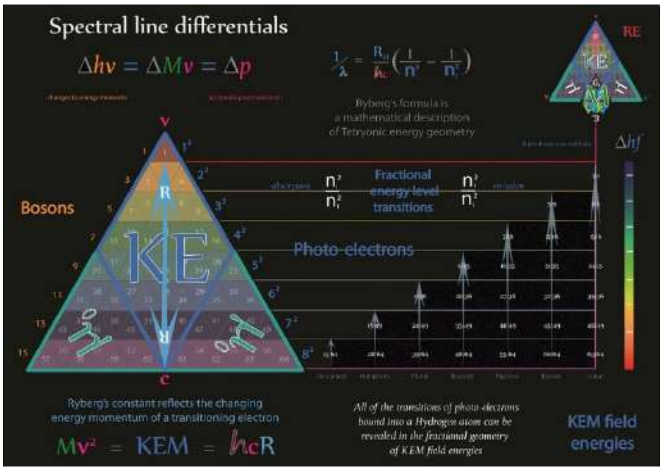
Spectral line differentials
$\Delta hv = \Delta Mv = \Delta p$
$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right)$
Ryberg's formula is a mathematical description of Tetryonic energy geometry
RE
KE
Δhf
Bosons
v
c
R
KE
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
absorption
$\frac{n_f^2}{n_i^2}$
Fractional energy level transitions
$\frac{n_i^2}{n_f^2}$
emission
Photo-electrons
Ryberg's constant reflects the changing energy momentum of a transitioning electron
$Mv^2 = KEM = hcR$
All of the transitions of photo-electrons bound into a Hydrogen atom can be revealed in the fractional geometry of KEM field energies
KEM field energies

There is NO explanation in modern chemical or quantum theory that accounts for the Proton-Neutron 'curve' of atomic isotopes nor for the islands of stability that exist for various elements, however Tetryonic theory accounts for all of this firstly by 'straightening' out the curved Z line of Protons and Neutrons underpinning atomic elements [by showing that all atomic nuclei are comprised of D nuclei not H atoms with excess neutrons] and then showing that the observed isotopic masses of each are the result of increased principal energy levels of their constituent baryons with the D nuclei comprising each element…. It is no error or 'false notion' as you put it, rather it is the current explanation of these facts by modern theory that is in error

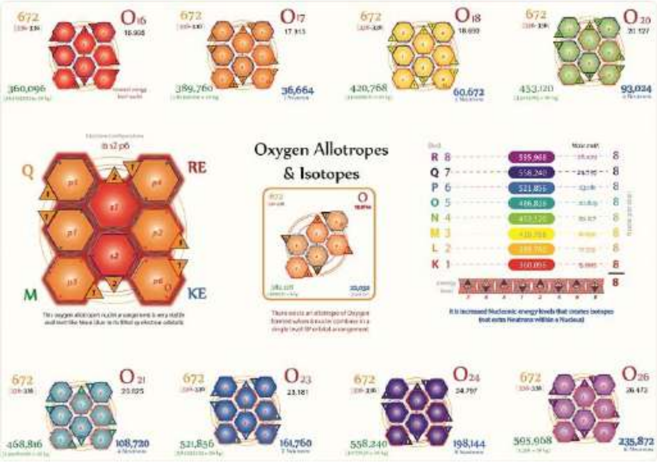
O16
O17
O18
O20
672
360,096
389,760
36,664
420,768
60,672
453,120
93,024
Q
RE
M
KE
Oxygen Allotropes
& Isotopes
O21
O23
O24
O26
468,816
108,720
521,856
161,760
558,240
198,144
595,968
235,872

You have chosen Lithium to make your point I chose Carbon, Nitrogen and Oxygen to make mine, but the explanation for Lithium remains the same….. an increase in the mass-energies of the baryons making up the element concerned… btw I have done spreadsheets for each and all the elements as a matter of course during the development of my theory.

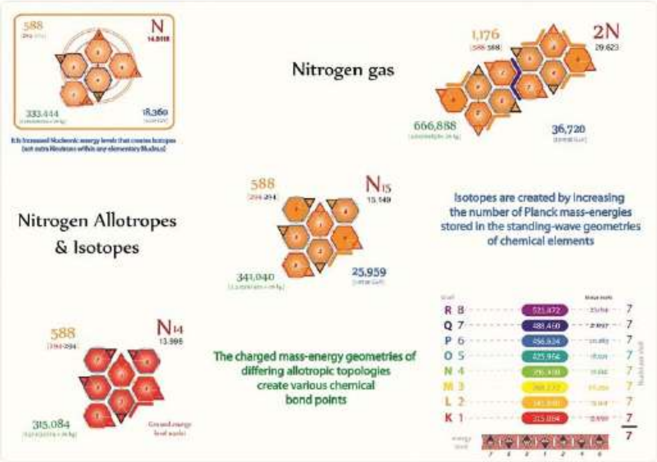
588
N
333,444
18,360
Nitrogen gas
1,176
2N
666,888
36,720
Nitrogen Allotropes
& Isotopes
588
N15
341,040
25,959
Isotopes are created by increasing the number of Planck mass-energies stored in the standing-wave geometries of chemical elements
588
N14
315,084
The charged mass-energy geometries of differing allotropic topologies create various chemical bond points

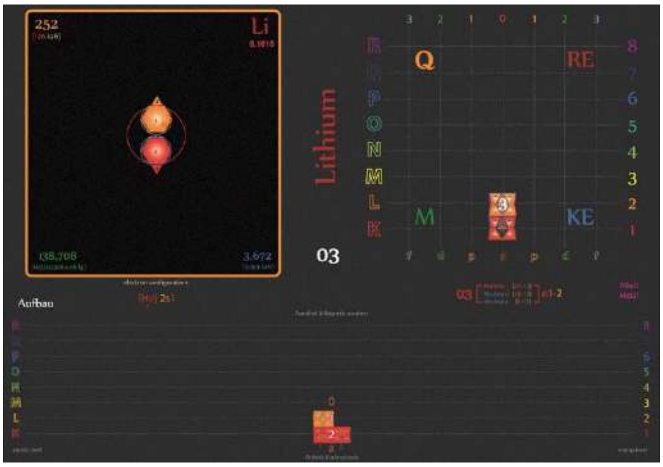
252
Li
138,708
3,672
Lithium
03
Q
RE
M
KE
Aufbau

Element 3 – Lithium [ground state atomic topology and electronic configuration]

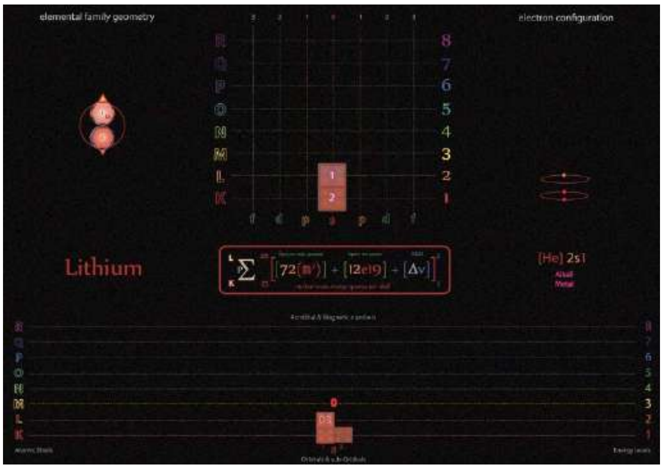

Regarding your comment of “but in reality **only electrons are on these shells** and the structure of the nucleus is completely separate from these shells”, - that is only the current opinion/consensus explanation of quantum chemistry at this stage and is obviously at odds with Tetryonic theory *built from the ground up from equilateral Planck energies not the top-down through supposition*….. I am confident that over time this consensus will change.

1. Explaining the extra (neutron) weight as higher energy-levels of the Deuterium nuclei seems *ad hoc* explanation with no supporting evidence – *simply nonsense*…..
2. No calculation is presented to *derive a prediction* for the excess mass (or energy) from the Tetryonics model – *have you even read my work in detail – calculations and mass-energies are provided everywhere throughout it*.
3. The excess mass of, say, Lithium 7 (one excess neutron compared to 3x deuterium) – *again a number based on your erroneous viewpoint – not y given explanation and illustrations of the same.*
4. This is inconsistent with the fact that there are **multiple** isotopes of each element.- *how so?*

   The energy level of Baryons of each and every illustrated element in Tetryonic can be increased resulting in isotopes of that element – with NO extra Neutrons present, again you seem to have failed to read and/or comprehend what was presented before you in the Tetryonics chemistry eBook.

5. See above for Hydrogen family [please]

All the ‘observed’ masses you have provided in that table are firstly just that, weighted [averaged] measurements of particles with kinetic energies [heat, spectral, motion etc.] not the absolute rest masses I have provided and clearly stated as such…… you must account for observational and inherent kinetic energies for your stated values to even begin reflecting the quantum values of particles at absolute zero.

Again, ‘isotopes’ is NOT what you think and proclaim – there are no extra Neutrons in the nuclei – only Baryonic energy above that of the res/ground state of such an element at absolute zero – unless you can account for those KEM field energies [as should be done]

Tetryonic Problem: Helium as Alkaline Earth metal

- And Hydrogen is element 1 which no-one can agree on which position it truly belongs in on the current periodic table.

Tetryonics problem: 120 possible maximum elements

- That is what is dictated by the charge topology of particles and elements themselves, perhaps you'd care to share what your spherical [or other] model for element 93 is, and how all the 'extra' Neutrons bind to each other to create that particular model?

Tetryonics shows quite clearly the reasons for there being only 120 possible periodic elements and it agrees with the math and equations of both Bohr and Schrodinger's models despite the differing quantum topologies.

Perhaps, you can produce element 121 and prove me wrong on this.

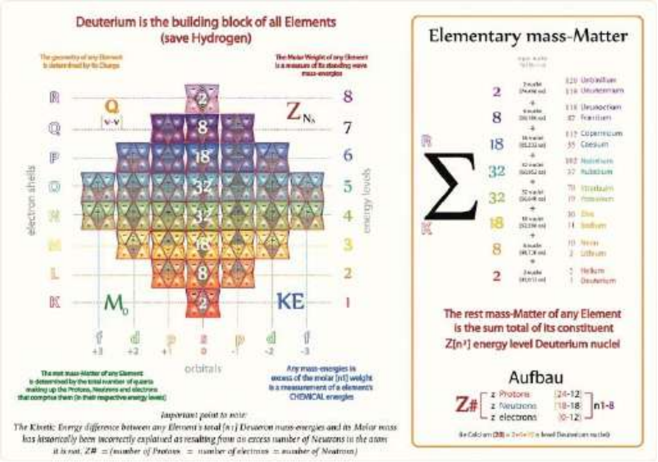
Deuterium is the building block of all Elements
(save Hydrogen)
R
Q
P
O
N
M
L
K
electron shells
Q
Z
M
KE
8
7
6
5
4
3
2
1
energy levels
2
8
18
32
32
18
8
2
f
d
p
s
p
d
f
+3
+2
+1
0
-1
-2
-3
orbitals
Any mass-energies in excess of the molar [n1] weight is a measurement of a element's CHEMICAL energies
Elementary mass-Matter
Σ
2
8
18
32
32
18
8
2
The rest mass-Matter of any Element is the sum total of its constituent Z[n²] energy level Deuterium nuclei
Aufbau
Z#
z Protons
z Neutrons
z electrons
[24-12]
[18-18]
[0-12]
n1-8

Additional random quotes - the uncertainty principle erroneously dictates that position and momentum cannot be simultaneously determined

- There's nothing random about this comment/quote - Tetryonic theory is founded on the SINGLE postulate that Planck mass-energy momenta has an quantum equilateral geometry [ie quantised angular momenta (m^2/s] is equilateral/triangular not circular as it is the macro sense of angular momentum] - from that flows the fact that the probabilistic nature of quantum theory and the maths used to describe it are in fact not 'fuzzy' or indeterminate as supposed presently but in fact an emergent property of the tessellation of mass-energy momenta at the quantum level.

The inherent indeterminacy of quantum theory is a crutch used my mathematicians and scientists alike to explain many of the puzzling aspects of quantum theory like the inherent statically probabilities of quantum fields themselves or the wave-particle duality of photons or particles in motion….. Again it is them and modern theory that have simply failed to grasp the geometric reality of energy momenta that underpins these phenomena and how they build to form the familiar laws of Nature at our scale of observation.

Tetryonic theory is all about correcting these erroneous assumptions and established viewpoints – but any change of perspective is not without birth pangs of angst and misunderstanding from those who first encounter it [like Newton, Planck and Einstein with their theories of Gravity, Quantum mechanics & Relativity theories before it]

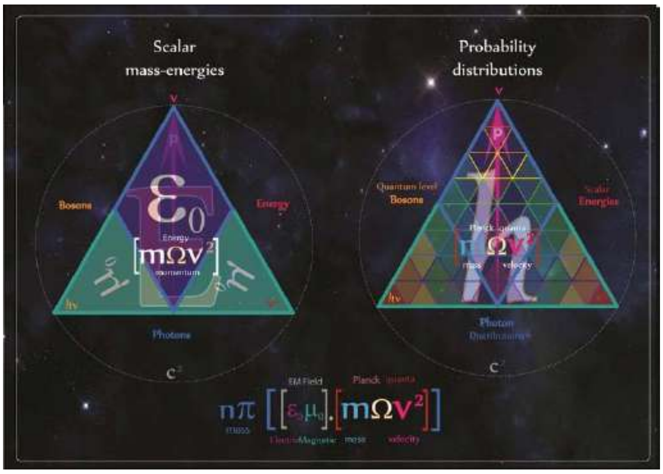
Scalar
mass-energies
Probability
distributions
v
Bosons
Energy
Energy
momentum
hv
Photons
c²
Quantum level
Bosons
Scalar
Energies
Planck quanta
mass
velocity
Photon
Distributions
EM Field
Planck quanta
nπ
mass
ElectroMagnetic
mass
velocity

That it is the transverse bosons Max Planck described using E=n.hv when resolving the ultraviolet catastrophe and spectral line emissions

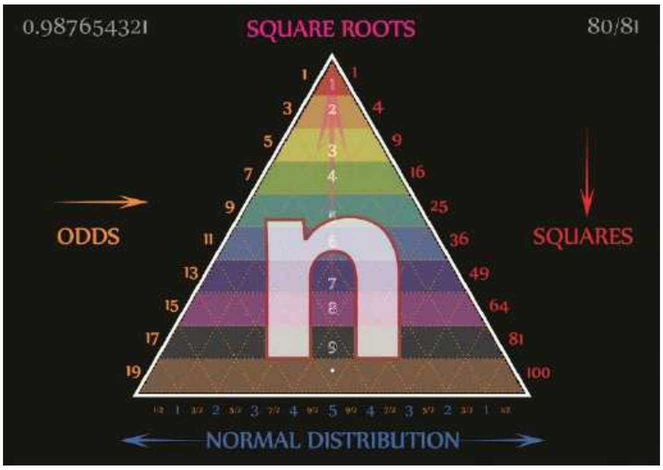
0.987654321
SQUARE ROOTS
80/81
ODDS
SQUARES
1
3
5
7
9
11
13
15
17
19
1
4
9
16
25
36
49
64
81
100
NORMAL DISTRIBUTION

And which subsequent theorists incorrectly applied directly to Einstein’s formulation of E= hf to describe neutral PHOTONS of energy momenta [dual charge bosons – quantum transformers]

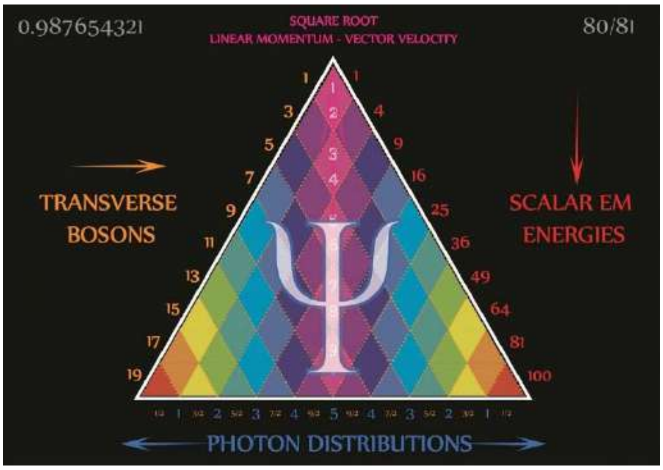
0.987654321
SQUARE ROOT
LINEAR MOMENTUM - VECTOR VELOCITY
80/81
TRANSVERSE
BOSONS
SCALAR EM
ENERGIES
Ψ
1
3
5
7
9
11
13
15
17
19
1
4
9
16
25
36
49
64
81
100
PHOTON DISTRIBUTIONS

And then erroneously attributed to an inherent uncertainly in quantum field dynamics etc. that does NOT and never existed in Nature itself or in Tetryonics….

I do look forward to further elaborating Tetryonic theory in detail with you – but also strongly suggest that you become more familiar with my Tetryonics [1] – Quantum Mechanics and & Tetryonics [2] - Quantum Electrodynamics eBooks beforehand as a lot of your suspicions of, and reaction to Tetryonic theory seem to stem from a lack of understanding of the basics of the theory….

And as I mentioned in previous communications, I remain open to an online discussion so we may discuss these points in far more detail than possible herein.

There is also lots more additional information and explanation of Tetryonic theory to be found on YouTube and Otto's Science of Life web- blog & the detailed 3D CAD models developed by Rene Cormier in the Google sketchup warehouse.

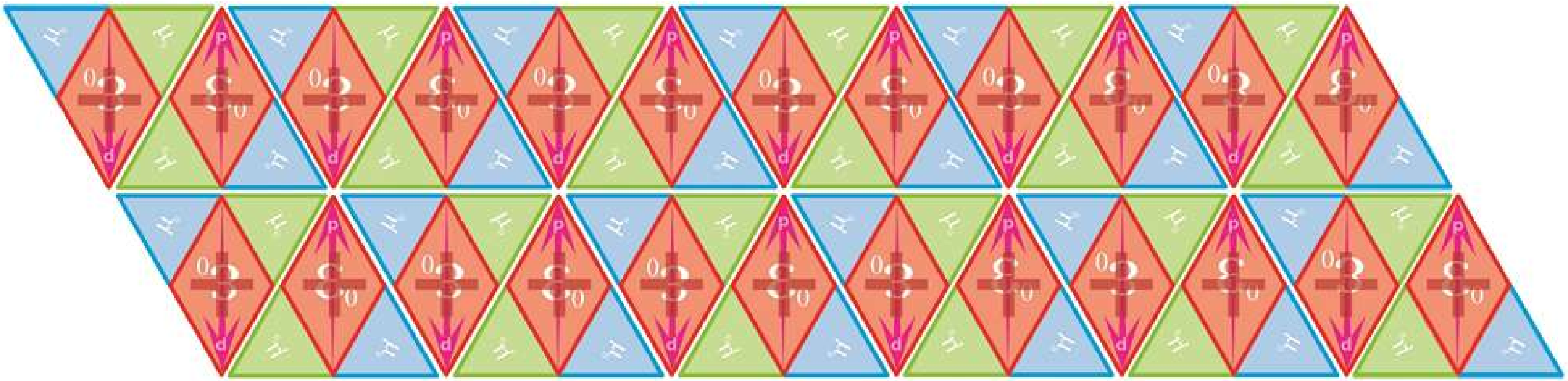

2D radiant EM mass-Energies
μ0
ε0

n1 - Charged Tetryon Templates

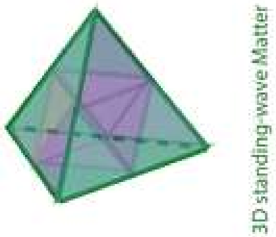
3D standing-wave Matter

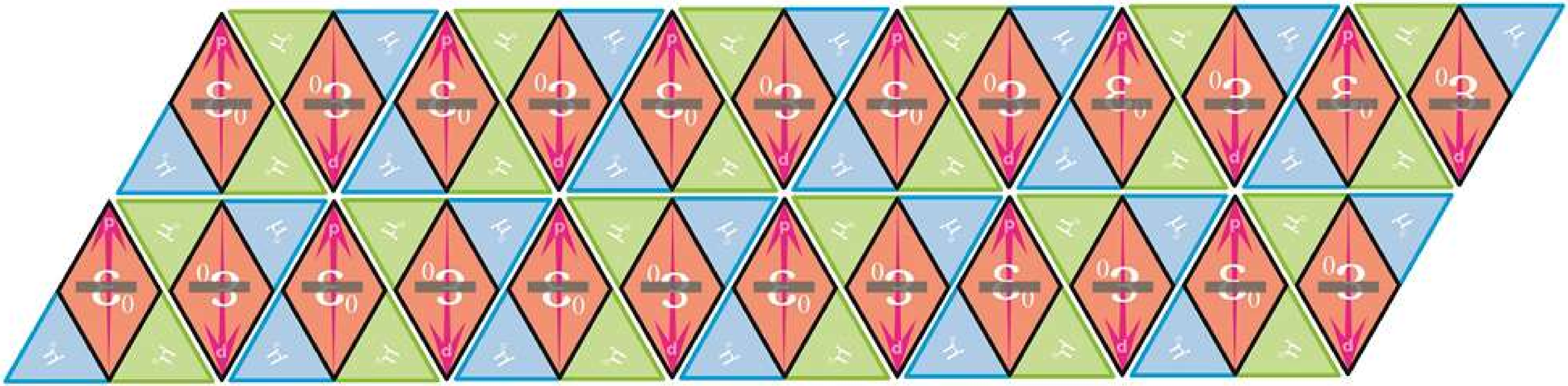

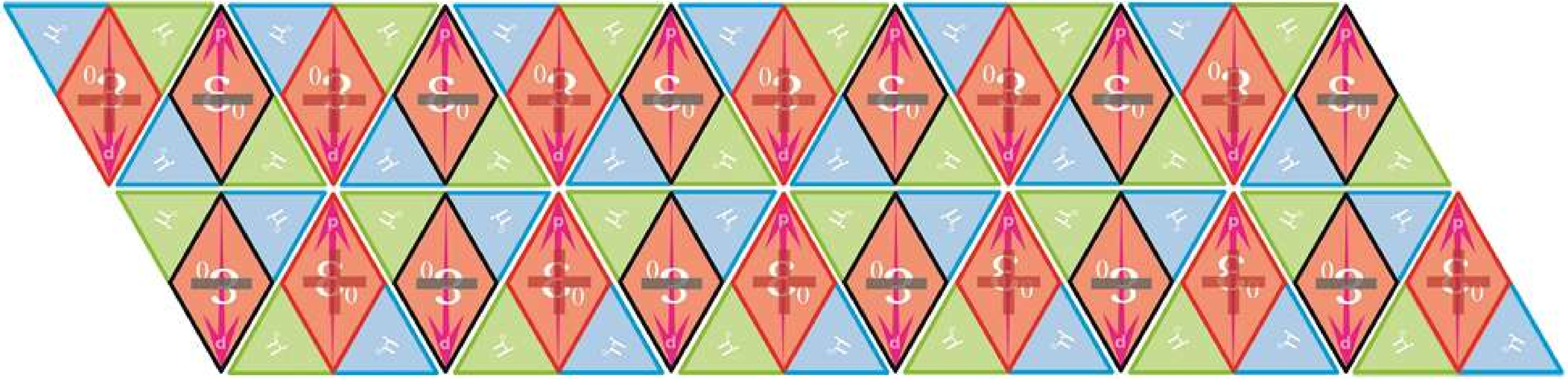

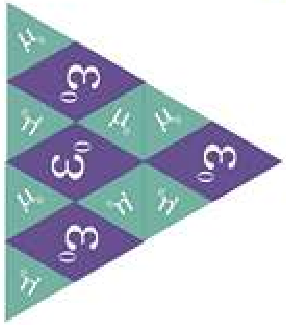

2D radiant EM mass-Energies

n1 - Neutral Tetryon Templates

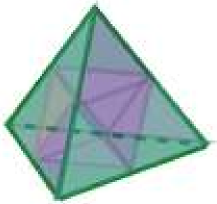

3D standing-wave Matter

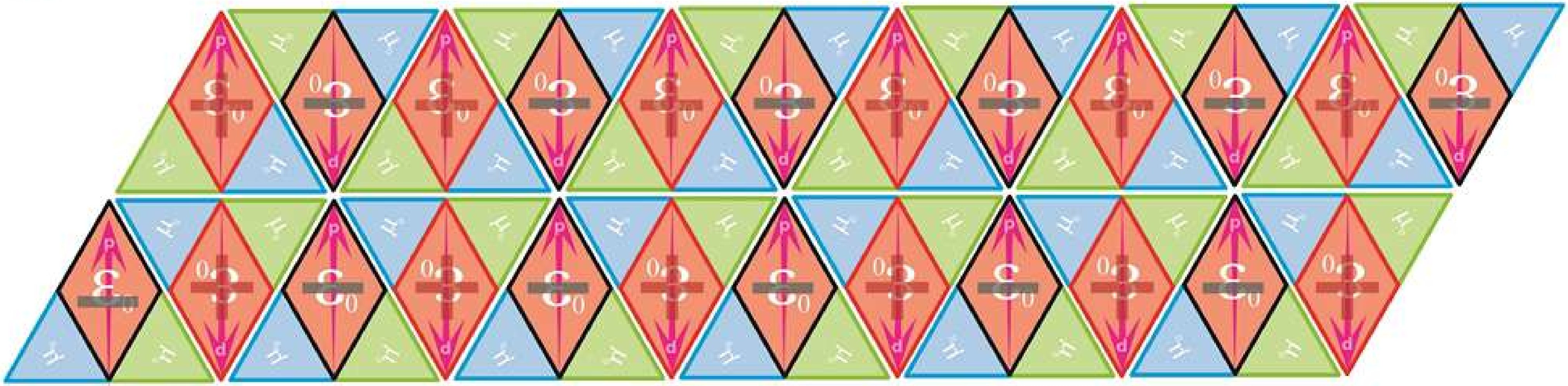

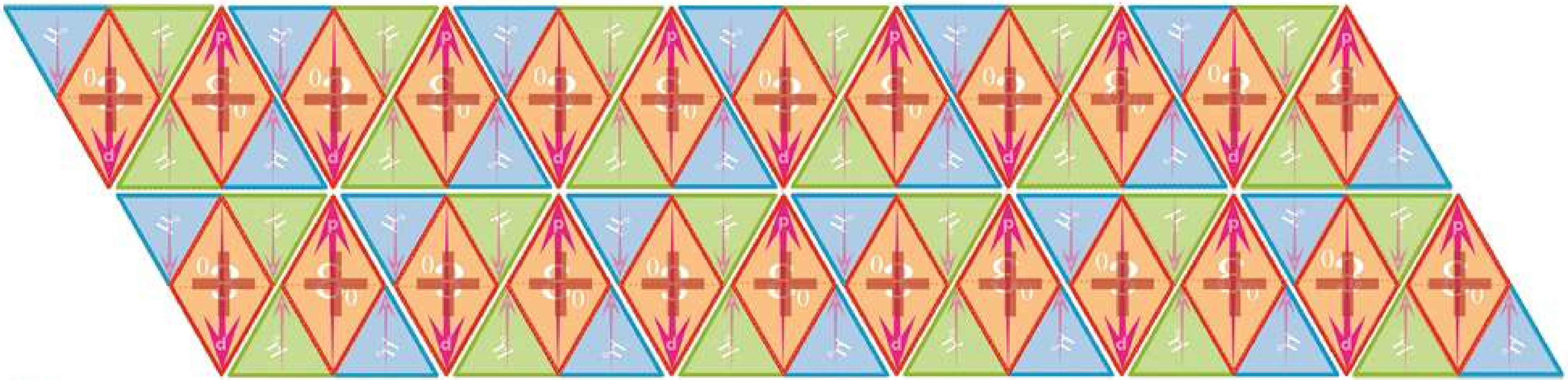

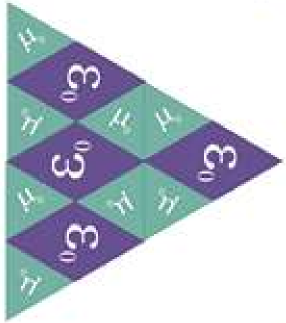
2D radiant EM mass-Energies

n2 - Charged Tetryon Templates

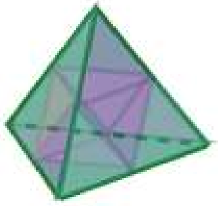
3D standing-wave Matter

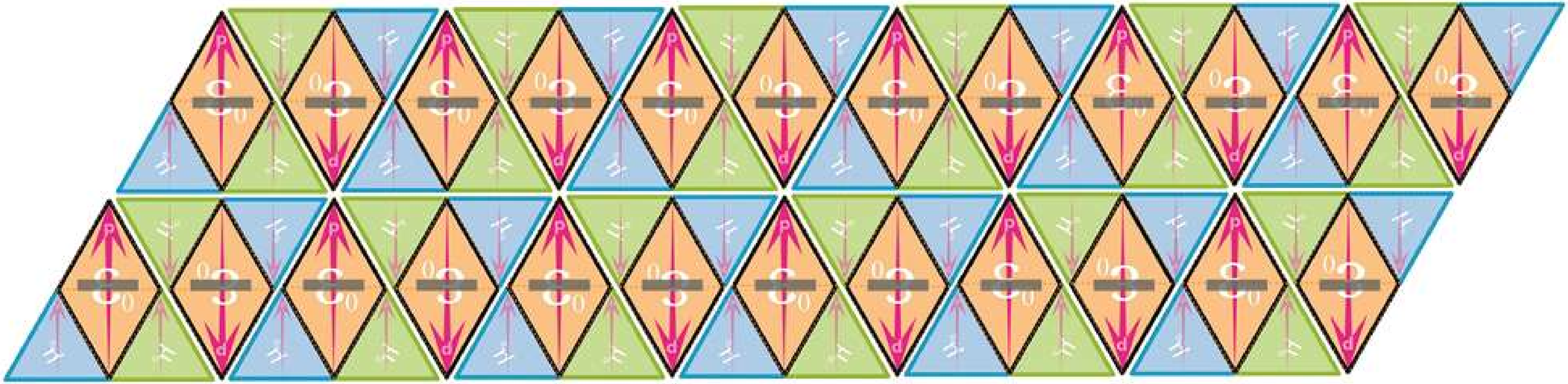

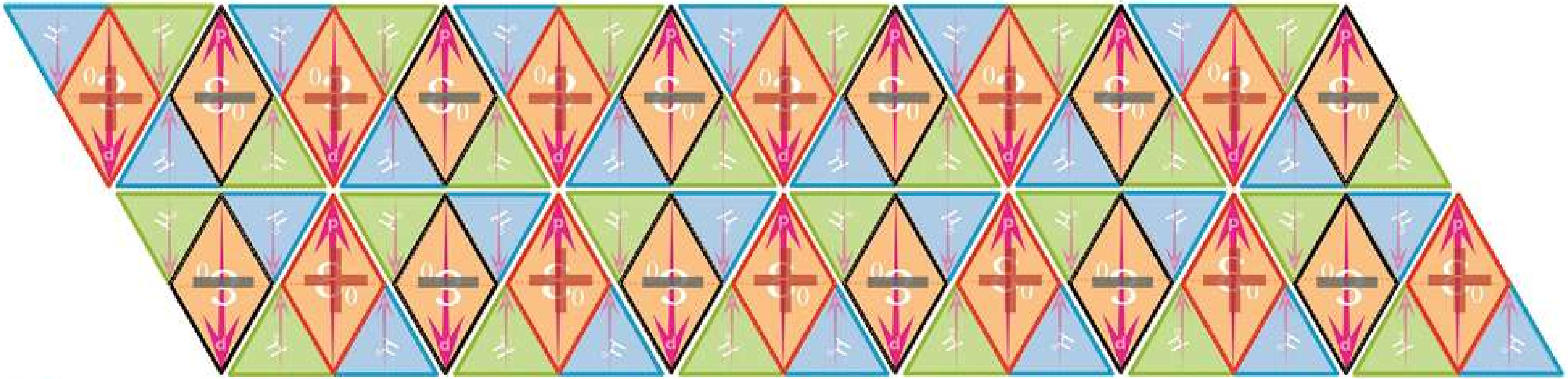

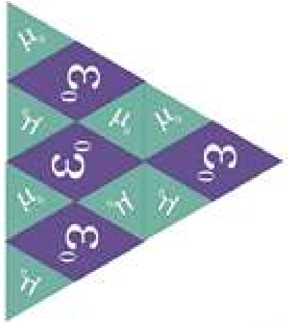

2D radiant EM mass-Energies

n2 - Neutral Tetryon Templates

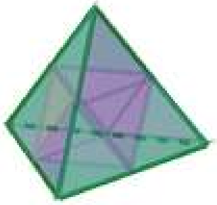

3D standing-wave Matter

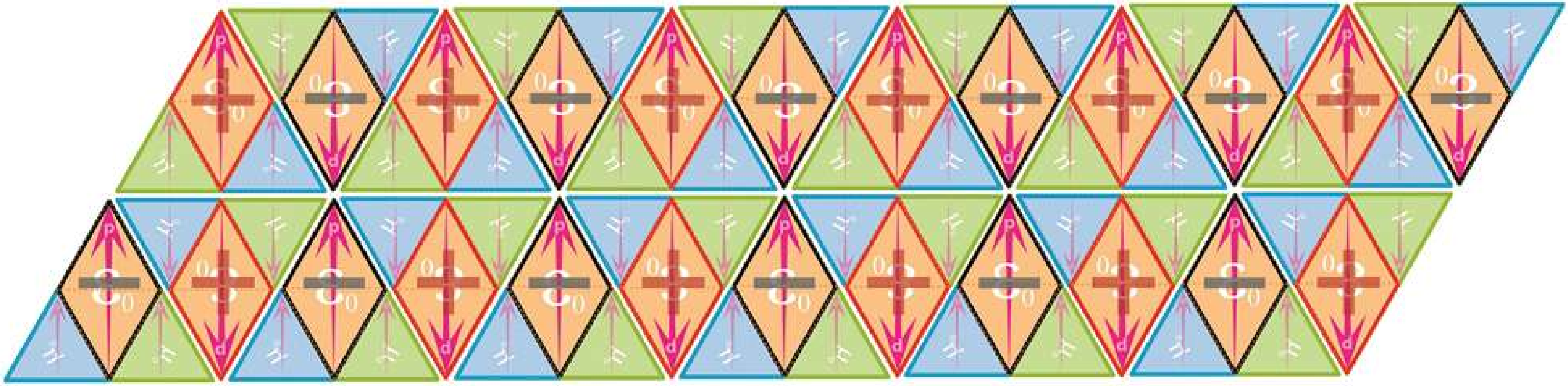

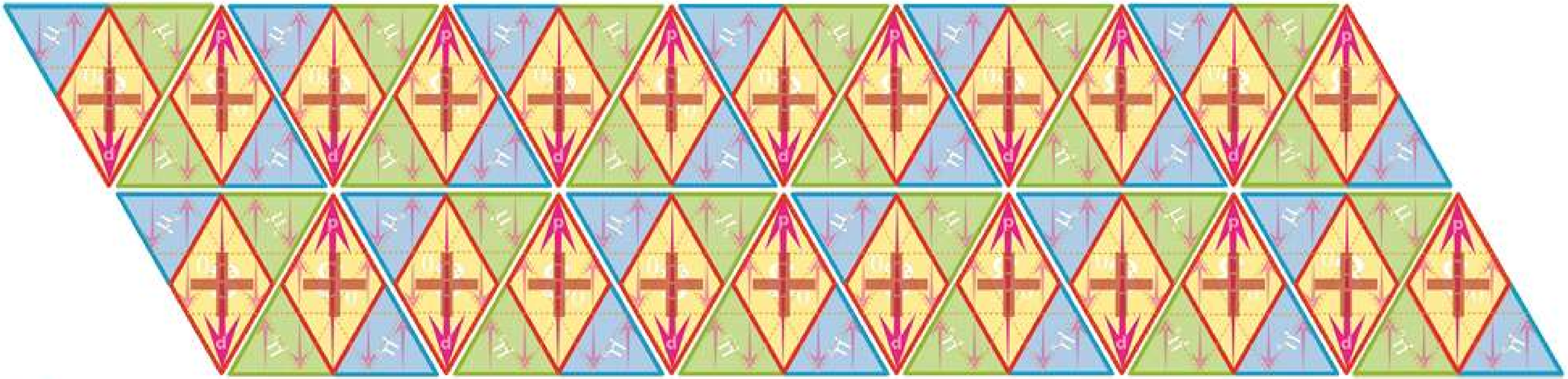

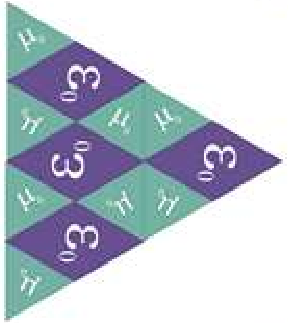
2D radiant EM mass-Energies

n3 - Charged Tetryon Templates

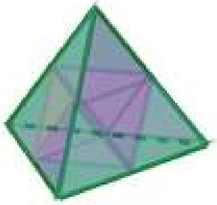
3D standing-wave Matter

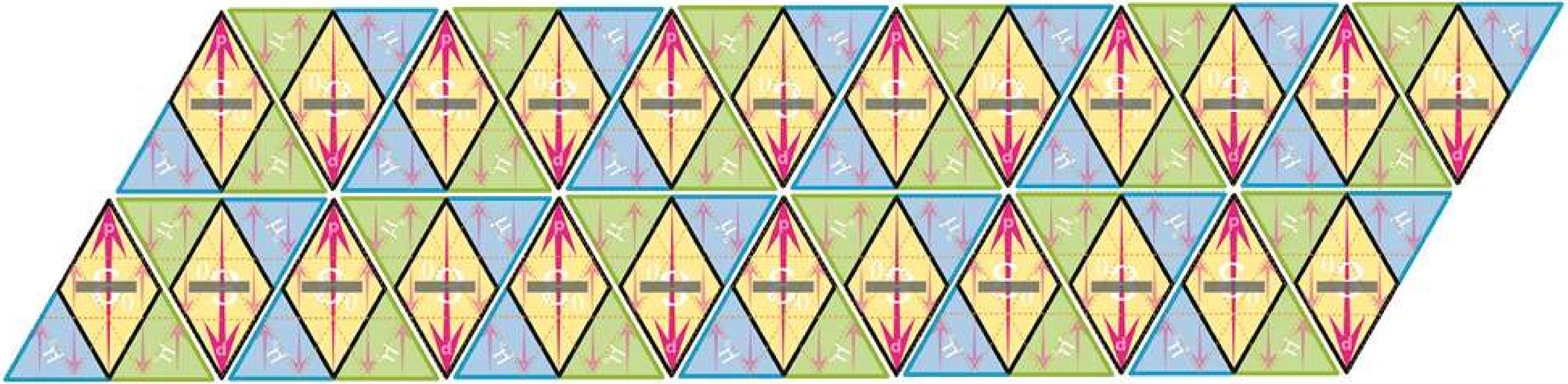

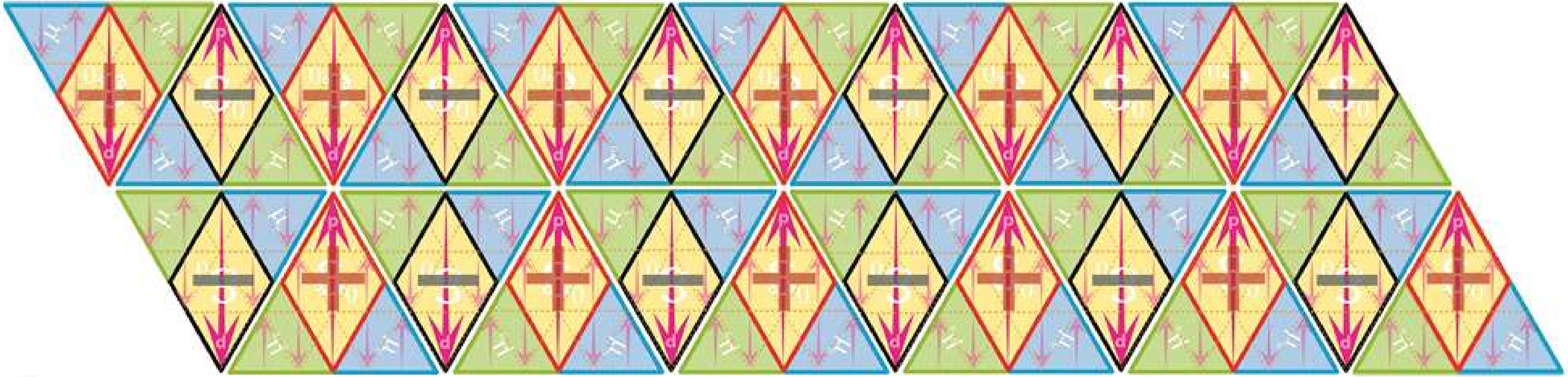

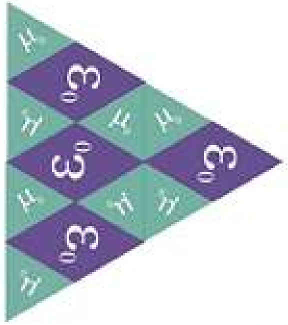

2D radiant EM mass-Energies

n3 - Neutral Tetryon Templates

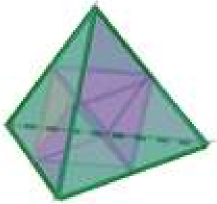

3D standing-wave Matter

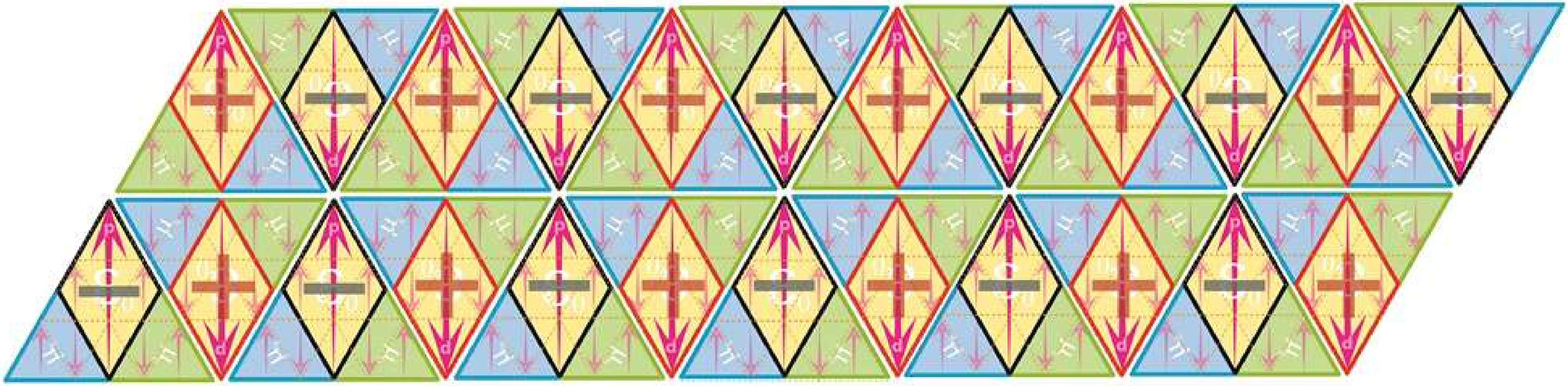

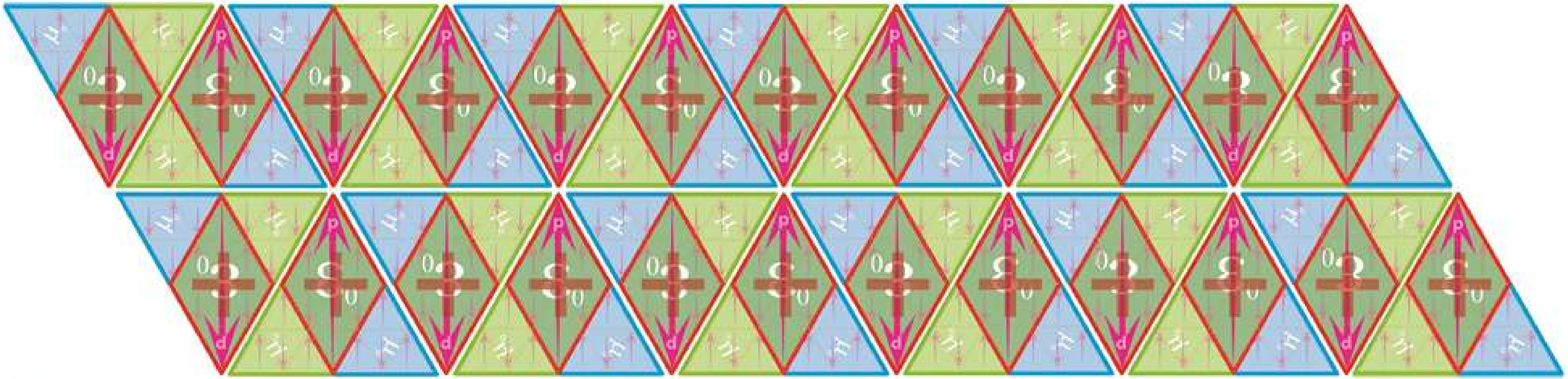

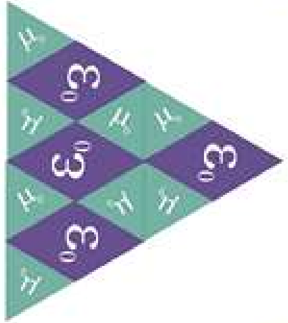

2D radiant EM mass-Energies

n4 - Charged Tetryon Templates

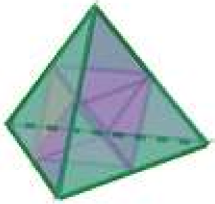

3D standing-wave Matter

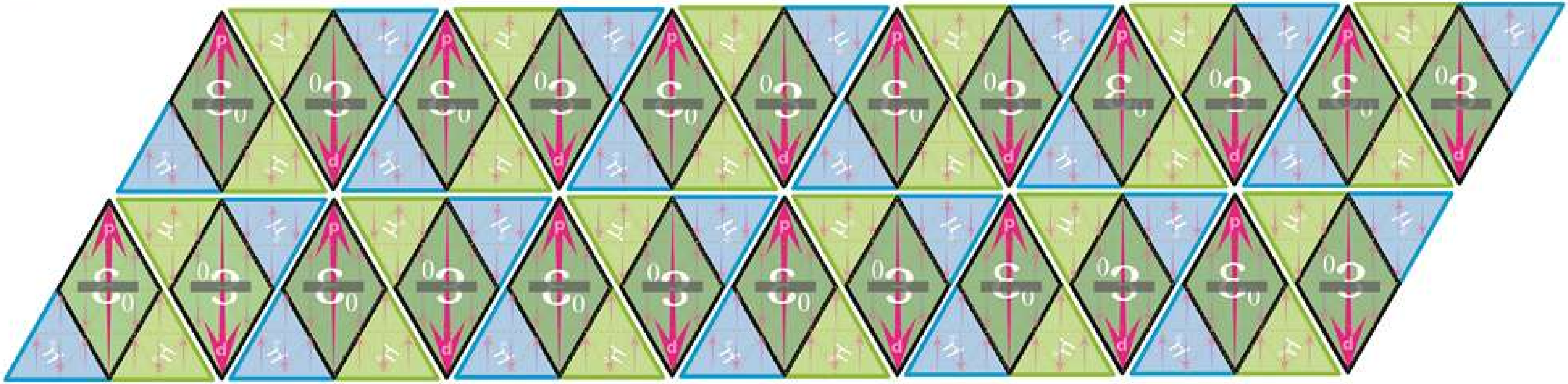

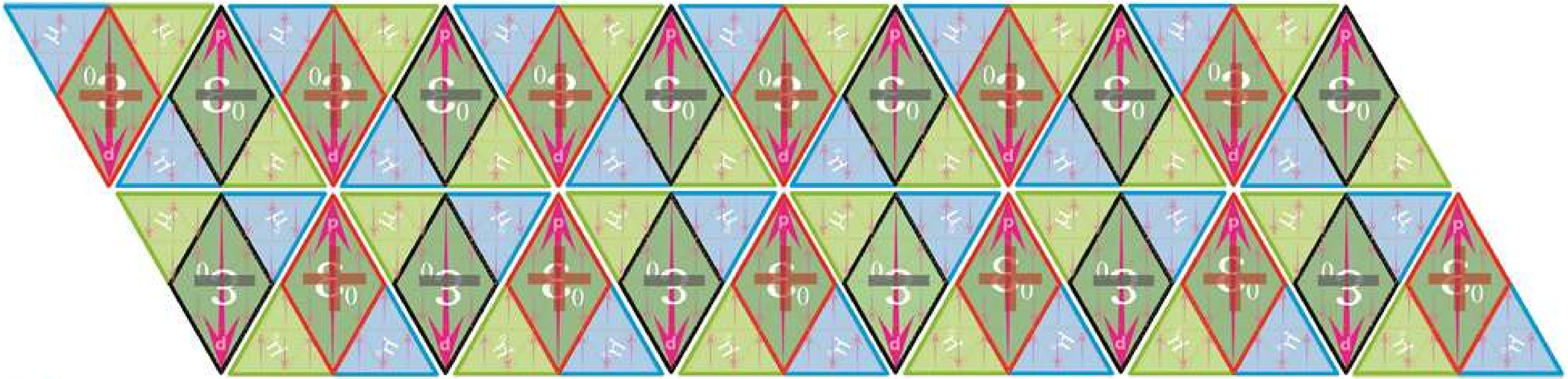

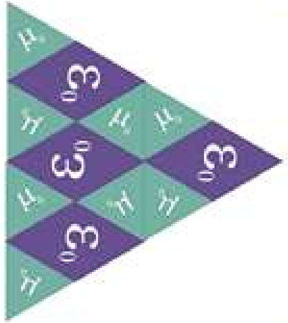

2D radiant EM mass-Energies

n4 - Neutral Tetryon Templates

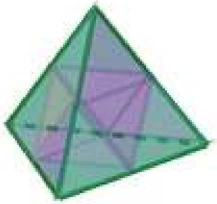

3D standing-wave Matter

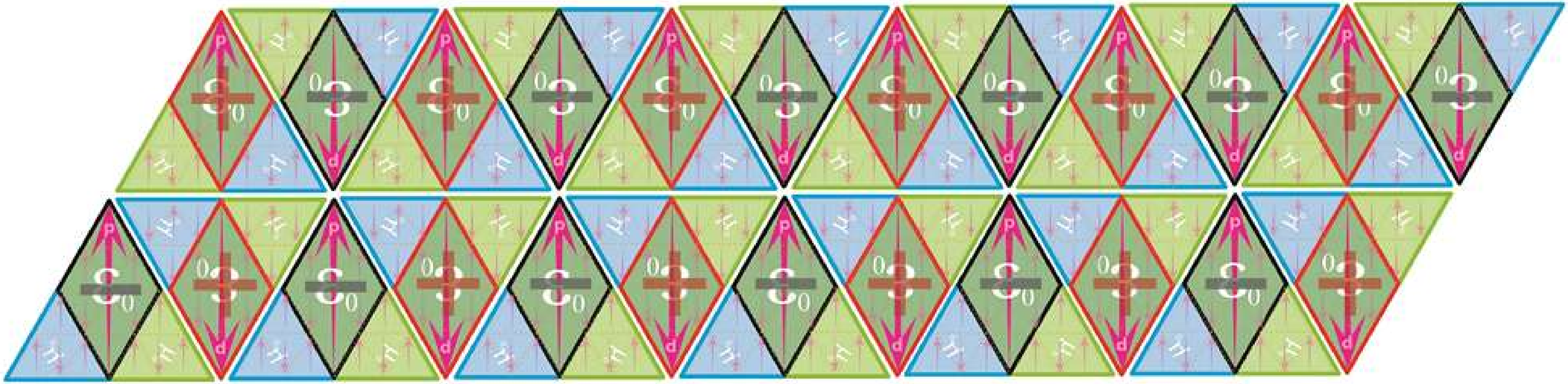

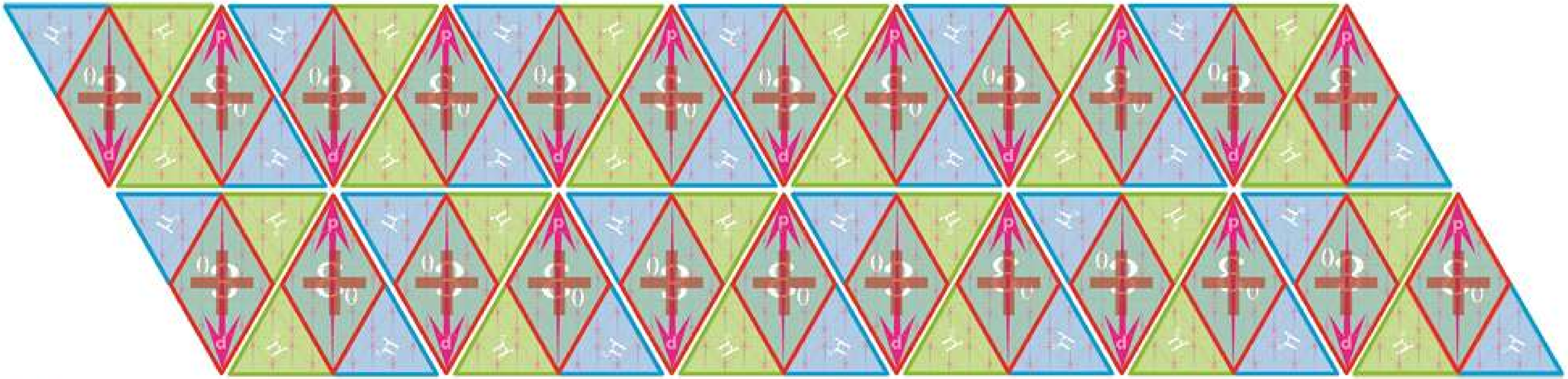
p
d
μ_0
ε_0

2D radiant EM mass-Energies
μ_0
ε_0

n5 - Charged Tetryon Templates

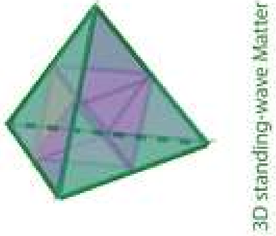
3D standing-wave Matter

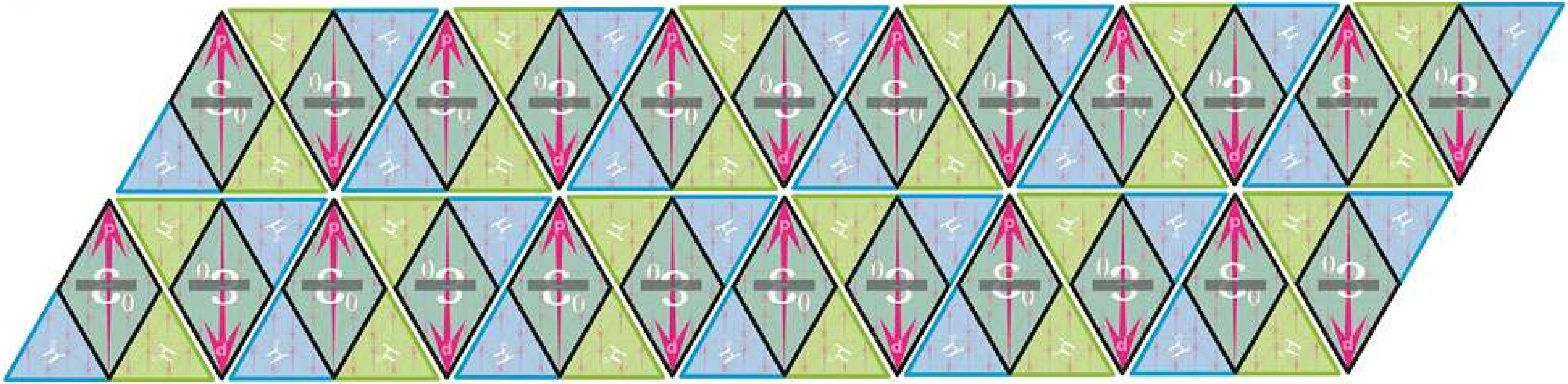
p
d
μ_0
ε_0

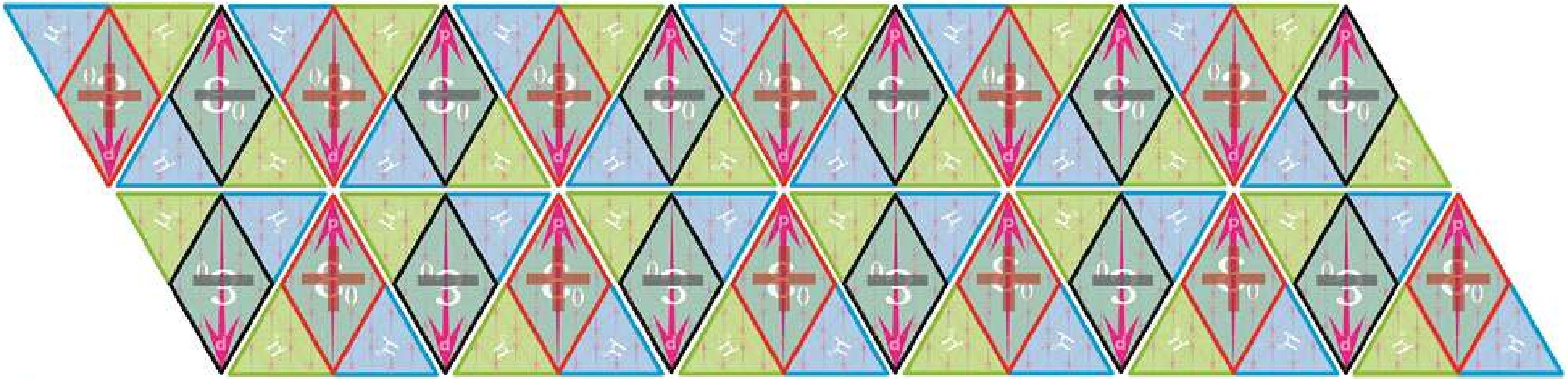

2D radiant EM mass-Energies

n5 - Neutral Tetryon Templates

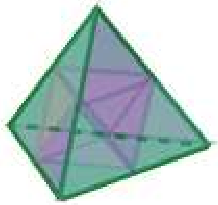

3D standing-wave Matter

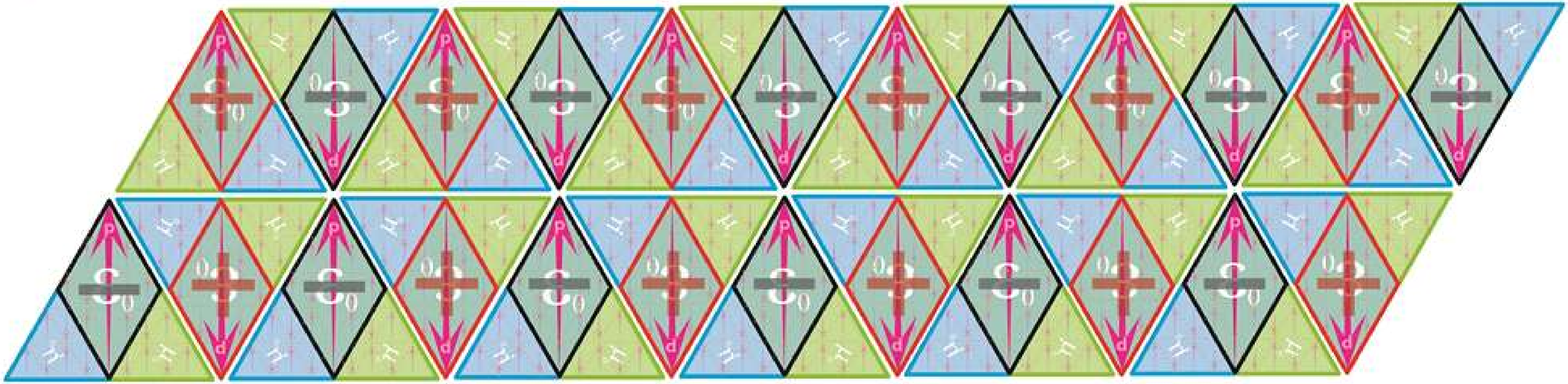

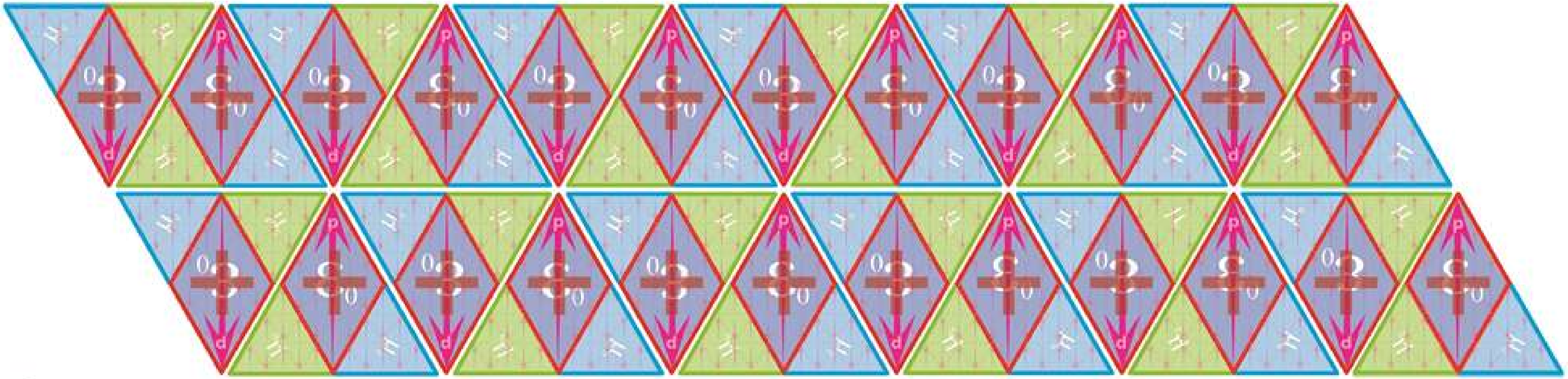

2D radiant EM mass-Energies
ε0
μ

n6 - Charged Tetryon Templates

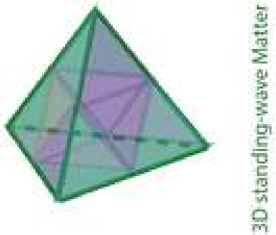
3D standing-wave Matter

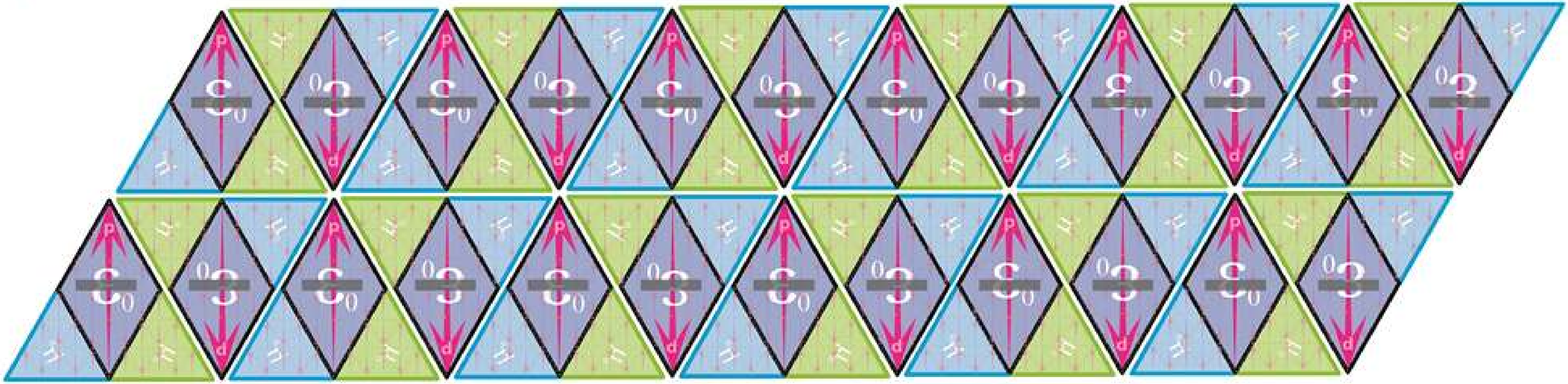

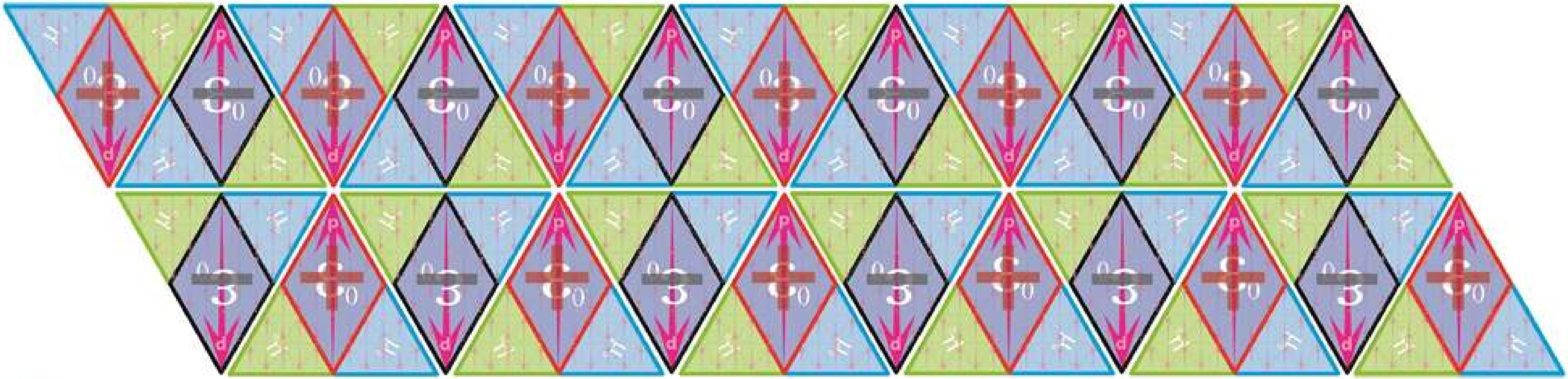

n6 - Neutral Tetryon Templates

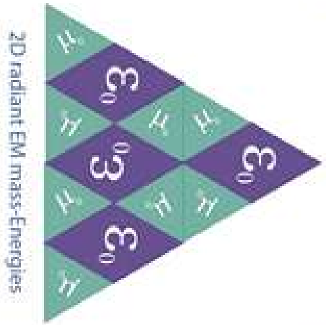
2D radiant EM mass-Energies
ε0
μ

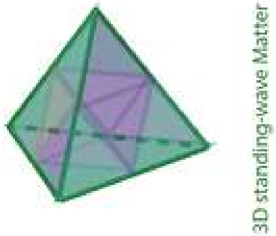
3D standing-wave Matter

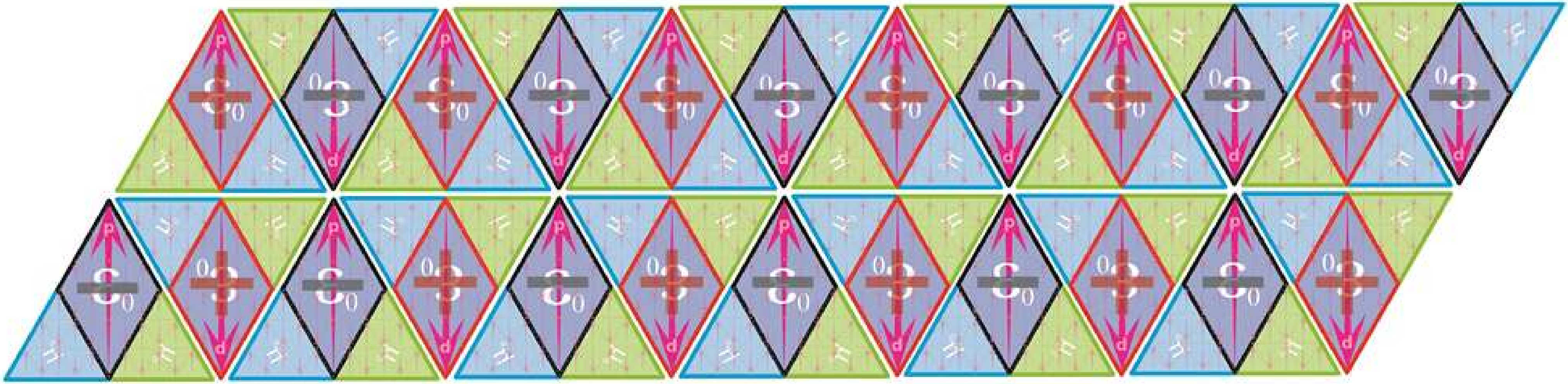

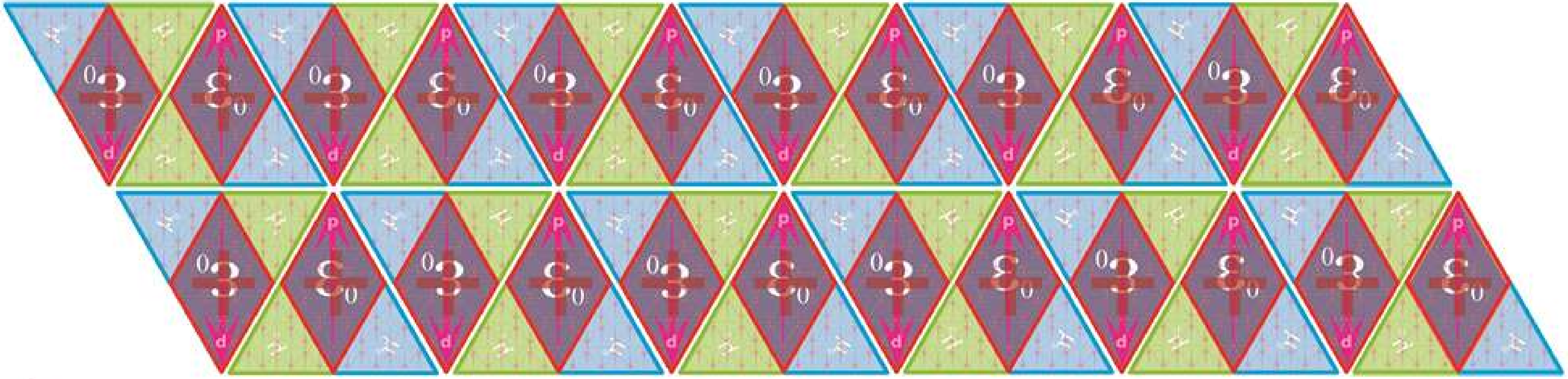

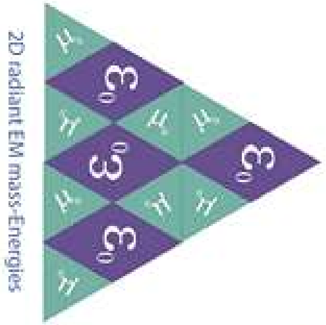
2D radiant EM mass-Energies
ε0
μ0

n7 - Charged Tetryon Templates

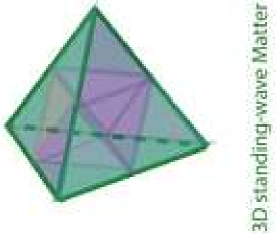
3D standing-wave Matter

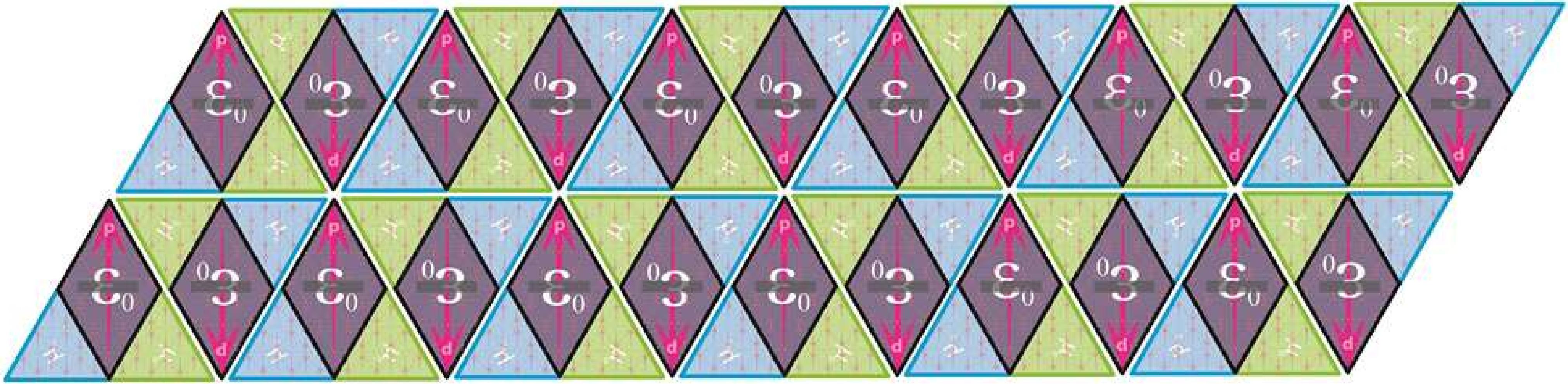

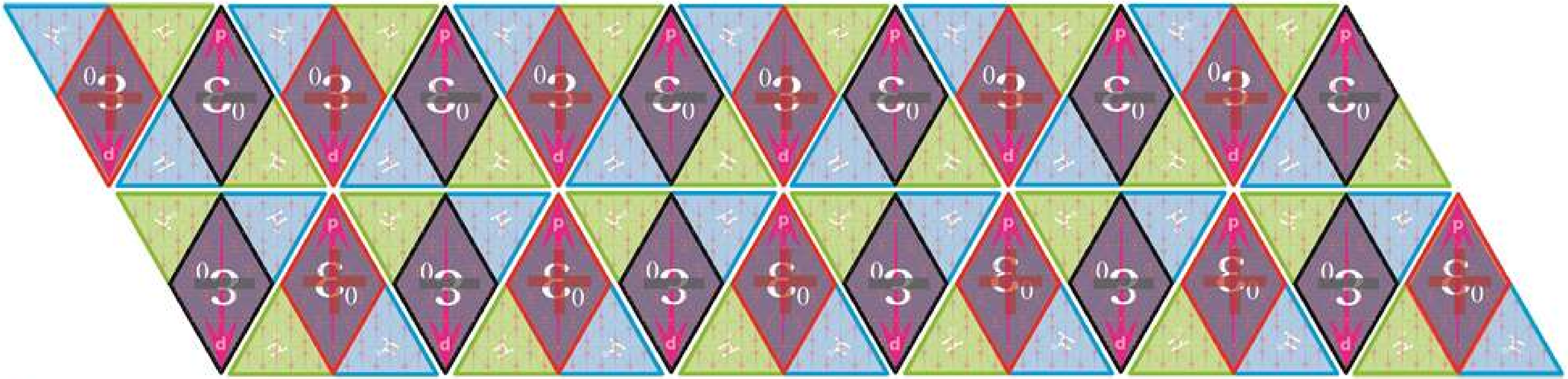

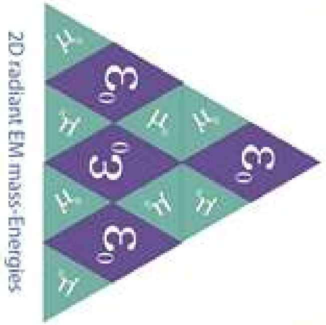
2D radiant EM mass-Energies
ε0
μ0

n7 - Neutral Tetryon Templates

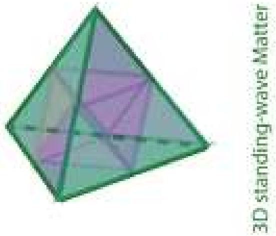
3D standing-wave Matter

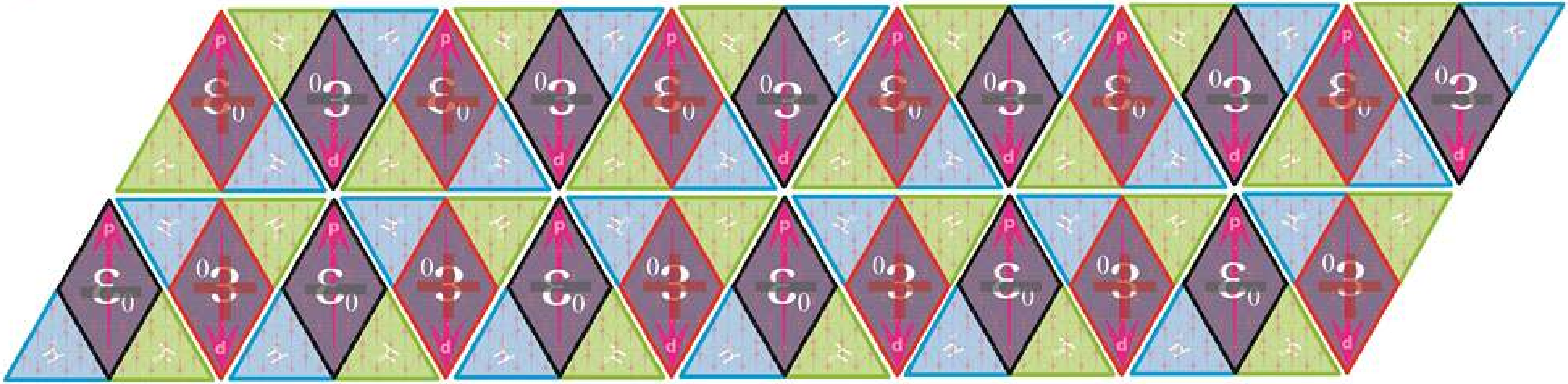

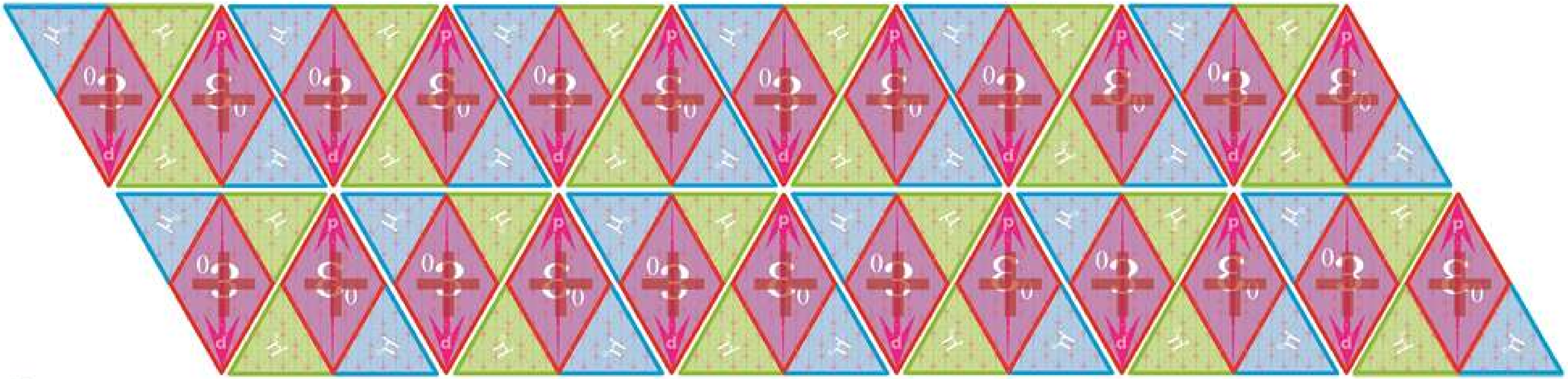

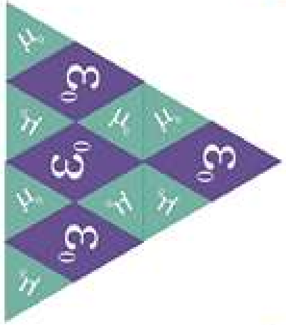

2D radiant EM mass-Energies

n8 - Charged Tetryon Templates

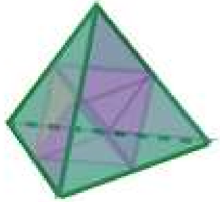

3D standing-wave Matter

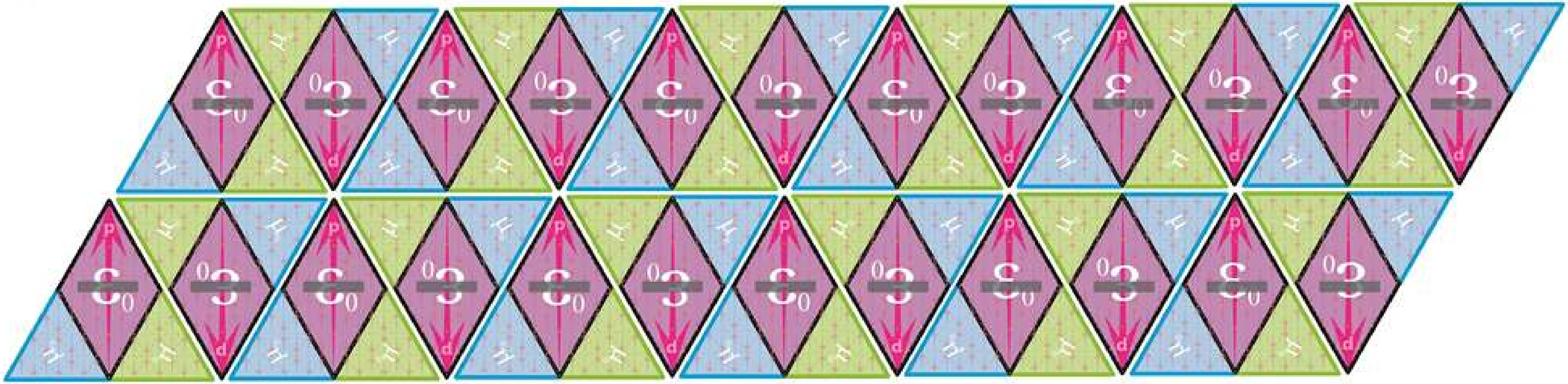

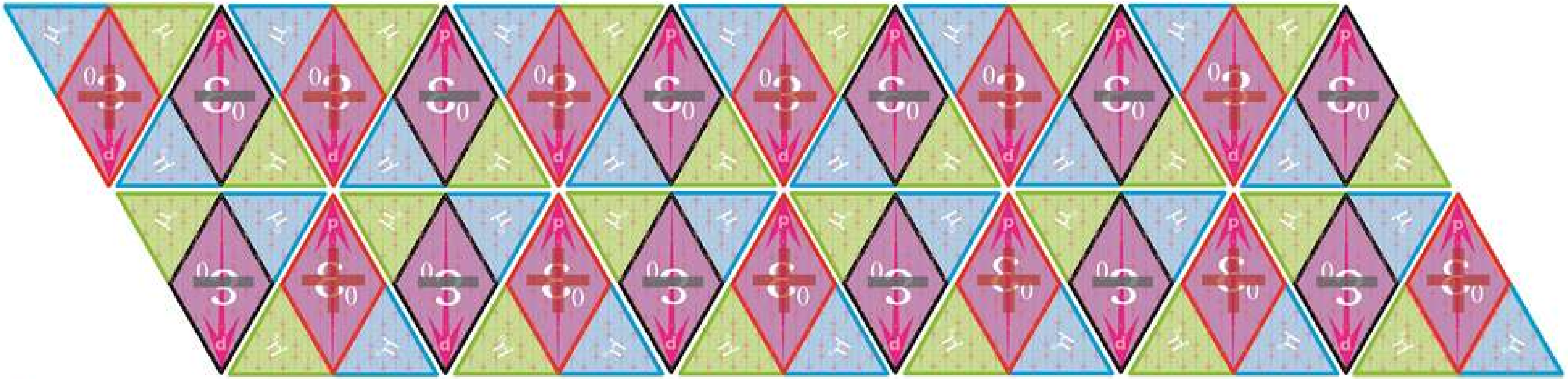

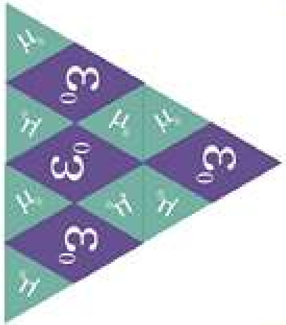

2D radiant EM mass-Energies

n8 - Neutral Tetryon Templates

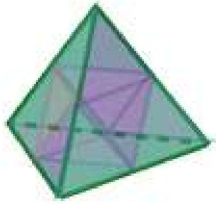

3D standing-wave Matter

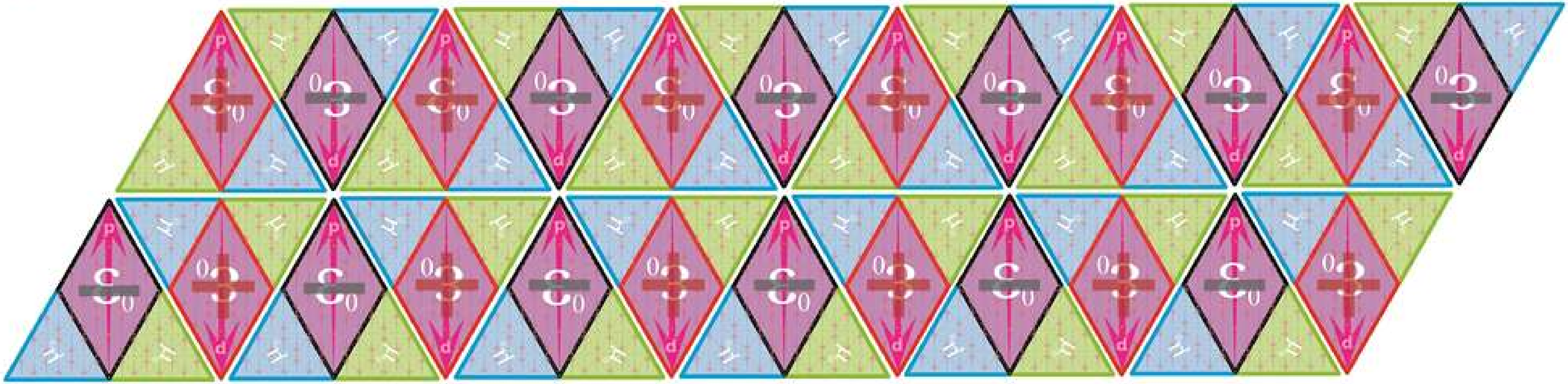

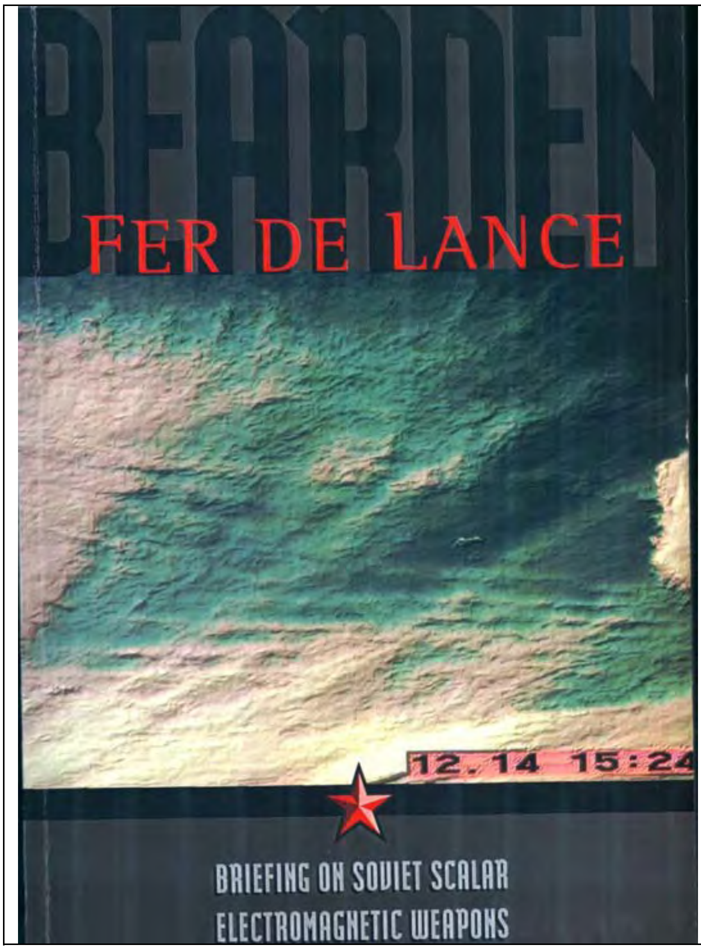
BEARDEN

FER DE LANCE

12.14 15:24

BRIEFING ON SOVIET SCALAR
ELECTROMAGNETIC WEAPONS

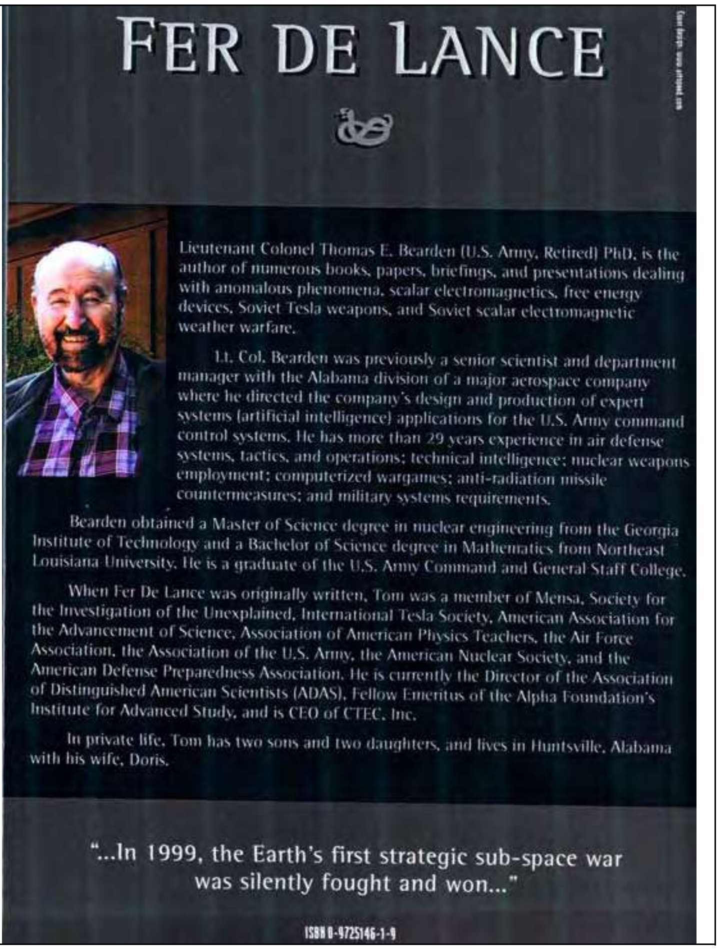
FER DE LANCE

Lieutenant Colonel Thomas E. Bearden (U.S. Army, Retired) PhD, is the author of numerous books, papers, briefings, and presentations dealing with anomalous phenomena, scalar electromagnetics, free energy devices, Soviet Tesla weapons, and Soviet scalar electromagnetic weather warfare.

Lt. Col. Bearden was previously a senior scientist and department manager with the Alabama division of a major aerospace company where he directed the company's design and production of expert systems (artificial intelligence) applications for the U.S. Army command control systems. He has more than 29 years experience in air defense systems, tactics, and operations; technical intelligence; nuclear weapons employment; computerized wargames; anti-radiation missile countermeasures; and military systems requirements.

Bearden obtained a Master of Science degree in nuclear engineering from the Georgia Institute of Technology and a Bachelor of Science degree in Mathematics from Northeast Louisiana University. He is a graduate of the U.S. Army Command and General Staff College.

When Fer De Lance was originally written, Tom was a member of Mensa, Society for the Investigation of the Unexplained, International Tesla Society, American Association for the Advancement of Science, Association of American Physics Teachers, the Air Force Association, the Association of the U.S. Army, the American Nuclear Society, and the American Defense Preparedness Association. He is currently the Director of the Association of Distinguished American Scientists (ADAS), Fellow Emeritus of the Alpha Foundation's Institute for Advanced Study, and is CEO of CTEC, Inc.

In private life, Tom has two sons and two daughters, and lives in Huntsville, Alabama with his wife, Doris.

"...In 1999, the Earth's first strategic sub-space war was silently fought and won..."

ISBN 0-9725146-1-9

Sweating done by red_gonzo.

Note: sometimes formatting of the text may be a little fuzzy. It's all because the OCR process didn't go as smoothly as I expected. Also there may be some errors inside text so be vigilant. In any case share this book, send emails to people about knowledge contained in it. Since this book talks about the most terrifying weapons present today the information in it is even more of value. But knowing an old saying »the knife cuts both ways« one can also use the same knowledge in order to heal ourselves since it seems that our body already does it using the physics described inside. Hopefully the latter will prove to be the most useful way to use this old but forgotten / supressed knowledge.

The contents include the 2002 update and the older 1986 briefing with slides.

TABLE OF CONTENTS

2002 OVERVIEW AND BACKGROUND........4

Mad Scramble of Soviet Science After the End of WWII........5

One Institute Had 2,000 PhDs and a Full Support Staff........7

Situation in the Early 1950s........8

1963 Deployment and Immediate Use of Scalar Interferometers........11

An Example of Russian Technology Astonishing Western Scientists........12

Softening of Metal at a Distance Was One Effect Used Against the Challenger 13

Energetics: Three Branches and a Strategic Deception Plan........14

Strategic Attacks Scheduled and Why We Were Not Destroyed........15

Attempts to Alert the Nation........18

The Responses........20

Effective Survival Treatment of Mass Casualties Could Be Developed........20

The Official Silence Was Deafening........22

The Long Dry Spell in Correcting Electrodynamics May Be Slowly Ending... 24

Engineerable Unified Field Theory Is Now Clearly Appearing........24

Examples of Some Major Non Sequiturs in Conventional Electrodynamics 25

Wide Confusion of Cause and Effect........26

CSRs: Engineering the Unobserved Causal Side To Create Functional Systems........27

Making the First CSR is Extremely Difficult; Making a Million Copies Is Easy........28

In 1999 The KGB Deployed Massive Numbers of CSRs for a Strategic Attack29

How the Asymmetrically Deployed Strategic Attack Was Stopped in Its Tracks........31

The Mostly-Ignored Time-Energy Interaction........32

Development of Threats and Scheduled Attacks Aborted........35

Live Destructive Tests Over and Near the U.S. Have Continued........38

The Scientific Basis For Energetics Weapons Is Established........39

Our Situation Here In the U.S. Is Precarious........39

The Dominant Weapons Are Quantum Potential Weapons....41
Cellular Time Reversal: Revolutionary Healing by a Team of French Scientists....43
Becker's Study of the Cellular Regeneration System....46
Need to Correct and Extend Classical Electrodynamics....48
The Situation Is Serious....48
A New Manhattan Project Is Warranted....49
2002 ENERGETICS WEAPONS THREAT TO THE U.S....51
The U.S. Narrowly Escaped Energetics Destruction Twice in 1997....51
Initiation of the Brazilian Energetics Weapon Program....52
Related Phenomena....54
Stalin *Personally* Initiated the Soviet Energetics Program with the Greatest Urgency....55
A Great Technical Breakthrough in Physics Can Always be Had in Four Years....55
A Typical Example: The Suppression of Cold Fusion....57
Revolutionary New Processes Are Hidden in the Cold Fusion Experimental Results....59
Stalin's Dictatorial Impact on Soviet Scientific Management's Response....62
Our Own Science Community's Progress During These Decades....63
U.S. Scientists Have Committed Monstrous Crimes and Gone Unpunished....66
Energetics: New Principles of Physics and How the Soviets Found Them....67
Longitudinal EM Wave Interferometry Energetics Was Almost Immediately Uncovered....68
By 1950 Russian Energetics Was Off and Running....69
Deployment of the Soviet Energetics Weapons and Their Quick Use....71
Why the Soviets Could Continue to Attack and Kill U.S. Targets with Impunity....73
The Science Behind the Soviet Energetics Weapons....77
The Importance of Scalar Photons and Time-Density Waves....80
Psychoenergetics: Engineering the Mind-Matter Interaction by Engineering Mind....81
Quantum Potential Weapons Are Utilized in All Three Energetics Branches... 84

The Importance of Longitudinal EM Waves 84
Energy, Wave Energy, and Wave Speed are All Dramatically Altered 86
A Process for Creating Longitudinal and Time-Density EM Waves 87
Quantum Potentials and Multiply Connected Spacetime 87
Need for Caution with QPs Used Against Biological Organisms 89
A Hypothetical Example 89
The QP Weapon Is a Superb Energy Amplifier "For Free" 91
Other Novel Uses of QP Weapons 93
The U.S. Narrowly Escaped Destruction Twice in 1997 Alone 95
Purpose of the Novel Attack on Captain Button 96
Rogue Japanese (Yakuza and Aum Shinrikyo) Have a Role 100
Inducing the Gulf War Disease 100
40-Years of Russian Induction of Diseases In U.S. Embassy Personnel 104
Western Electrodynamics Is in Sad Shape 105
Toward a Deeper Electrodynamics 108
Time Is Highly Compacted Energy 109
The New Transduction Model Explains the Cold Fusion Results and Mechanisms 109
Why There Is a Time Delay in the Onset of Cold Fusion Phenomena 1ll
The Early Soviet Longitudinal Wave Interferometers 112
Khrushchev Killed the Thresher to Prevent His Own Dismissal 112
Khrushchev Demonstrated the Powerful Blast Capability of Strategic Interferometry 114
Weather Engineering Over North America 115
Earthquakes to Order 117
Readying the Weapons for the 1985 Readiness Demonstration Tests 119
Meeting the 1985 Schedule for Readiness 121
KGB Crews Regularly Practiced Against U.S. Rocket Launches from Cape Canaveral 122
Kill of the Arrow DC: Practice Turned into Performance 123
The Arrow Pilot Sensed the Lack of Power, But the Aircraft Was Immediately Hit 124

Elements of the U.S. Government Deliberately Suppressed Vital Evidence... 125
Other Direct Energetics Signatures on the Arrow DC-8 126
Other Acts of War Completed the Milestone 127
Kill of the Titan Missile Over Vandenberg Air Force Base 128
Dead-Man Fuzing and Outside Assistance Deterred a Full Strategic Attack in 1986 129
What Really Happened in the Chernobyl Disaster 129
The High Side Risk Factor 131
The Low Side Risk Factor 132
Consequences of Not Taking Counterfire Action 132
My Recommendation to the Counterfire Group 132
The Group Fired and Took Out a Transmitter 133
Results at the Transmitter and at Chernobyl 133
Kill of the Challenger Included Use of Metal-Softening 135
Cold Molding: An Epilog to the Metal-Softening Donnybrook 136
Additional Update from Year 1998 to Year 2,000 Has Been Given 137
There is Also a Severe Bioenergetics (EM Biological Warfare) Threat 139
National Institute of Health Reaction 141
How the New Revolution in Medical Healing Works 142
What the French Researchers Showed 145
How the Regeneration Mechanism Was Finally Deciphered 147
Cellular and Body Rejuvenation are Possible, Once the Technology is Developed 149
How Many Americans Could Be Saved? 149
There Are No Resistant Strains to the New Technology, and There Cannot Be Any 150
Results of the Priore' Treatments Are Well-Documented and Independently Supported 151
The Present Threat Is Urgent, and Our Strategic Destruction is Still at Issue .151
Desperately Needed Defensive Capabilities Can be Provided by a QP Solution 152
In Conclusion 153

1986 FER DE LANCE INTRODUCTION & BRIEFING ON SOVIET SCALAR ELECTROMAGNETIC WEAPONS 157

Scalar Electromagnetics is Electrogravitation 157

Fer-de-Lance 157

Energetics and Directed-Energy Weapons (DEWs) 158

The Soviets Use a Deception Plan 158

Destruction of the Challenger in January 1986 161

Technological Surprise and a New Hiroshima 164

We Have Given Up Defense in the Larger Sense 164

1986—THE BRIEFING 167

Slide 001: Characterization 169

Slide 002: Outline 171

Slide 003: Elements of the Emerging Theory 172

Slide 004: Three Kinds of Electromagnetics 174

Slide 005: Bohm-Aharonov Effect 176

Slide 006: A-Field is Real 177

Slide 007: Quantum Non-Locality 179

Slide 008: Vector Mathematics has a Fundamental Problem 180

Slide 009: The Zero-Vector Axiom 182

Slide 010: Are These Systems Equal? 184

Slide 011: Infolded Systems are Excluded by Physics 185

Slide 012: Kaluza Geometry 186

Slide 013: A Potential is a Change in the Stress of Vacuum 187

Slide 014: Type of Potential Depends on the Particle(s) 190

Slide 015: A Potential Extends to Infinity & Involves the Entire Universe 192

Slide 016: A Natural Potential 194

Slide 017: An Artificial Potential 195

Slide 018: Producing an Artificial Potential 197

Slide 019: If No Observable Mass Flows: 199

Slide 020: Gravitational Potential: A Conglomerate of Stress Types 201

Slide 021: What Gravitational Field Is 202

Slide 022: EM Force Fields are Releases of Gravitational Potential via Observable Mass Flows 204

Slide 021: Scalar O-Wave Production 206

Slide 024: Scalar Electromagnetics 208

Slide 025: Rogers' Undersea & Underground Communications Systems 211

Slide 026: How does the Rogers System Work? 213

Slide 027: Scalar EM Waves are Emitted & Absorbed by the Nucleus 214

Slide 028: Nuclei of the Universe Continually Exchange Scalar EM Waves.. 216

Slide 029: Converting EM Field Energy to G-Field Energy & Vice-Versa 217

Slide 030: Normal EM Bleeds-Off G-Field 219

Slide 031: Scalar EM Forces 4-D G-Field 220

Slide 032: Detection of Electro-Gravitation 221

Slide 033: Scalar Wave Detector 226

Slide 034: Scalar Resonance 229

Slide 035: Scalar Resonance 231

Slide 036: Scalar Resonance Currents 234

Slide 037: Transverse Wave Interference 235

Slide 038: Transverse Wave Interference 236

Slide 039: Creating Energy at a Distance 239

Slide 040: Exothermic Mode (Scalar Interferometer) 240

Slide 041: Endothermic Mode (Scalar Interferometer) 242

Slide 042: Tesla's "Big Eye" for Remote Viewing 244

Slide 043: Site of Mystery Cloud 247

Slide 044: Mushroom Cloud from Sea off Japan near Russian Test Area 248

Slide 045: Mushroom Cloud Rising from Sea 249

Slide 046: Mushroom Cloud from Sea off Japan 251

Slide 047: Not a Known Natural Phenomenon 252

Slide 048: Explosive Events seen on Soviet Island 253

Slide 049: Explosive Events seen on Soviet Island 255

Slide 050: Explosive Events seen on Soviet Islands 256

Slide 051: Creating a Spatial 0-Field Pattern 257

Slide 052: Tesla Shield 258
Slide 053: Expanding Dome-like Phenomenon 260
Slide 054: Hemisphere & Globes 1977 261
Slide 055: Saryshagan Direction, September 1979 263
Slide 056: Continuous Tesla EMP Globe 266
Slide 057: Mysterious Light over North Pacific 268
Slide 058: White Sphere seen in North Atlantic 1976 269
Slide 059: Continuous Tesla Fireball 270
Slide 060: Terminal ABM System 272
Slide 062: ABM Defenses 275
Slide 063: Network of Virtual Transmitters 278
Slide 064: Network of Virtual Transmitters 281
Slide 065: Launch Phase ABM System 285
Slide 066: Soviet Command & Control Concept 287
Slide 067: Launch Phase ABM System (Anti-SLBM) 290
Slide 068: Detecting & Marking Missile Launch 291
Slide 069: Destroying the Launched Missiles 292
Slide 070: Shift to Anti-Sub Mode 293
Slide 071: Shifting to Earthquake Mode 295
Slide 072: Suppression Attack—MARVs (without defenses) 297
Slide 073: Suppression Attack—MARVs (with defenses) 298
Slide 074: Response to Counterforce Strike with No Defense 300
Slide 075: Response to Counterforce Strike with Defense 302
Slide 076: Clear Communications 305
Slide 077: Unique Communications 306
Slide 078: Mine Countermeasures—Orthodox System 308
Slide 079: Mine Countermeasures—Scalar EM Systems 309
Slide 080: Scalar Anti-Personnel Weapon 311
Slide 081 Afghanistan 312
Slide 082: Scalar Anti-Tank Weapon 314

Slide 083: Scalar Anti-Aircraft Weapon 315
Slide 084: Scalar Anti-Aircraft Weapon 316
Slide 085: An Effective Penetration Aid 318
Slide 086: Unique Capabilities: ATM/AD/AARTY 319
Slide 087: Unique Capabilities 320
Slide 088: Unique Capabilities 321
Slide 089: Unique Capabilities 322
Slide 090: Unique Capabilities: Penetrate Shielding 323
Slide 091: Multi-Purpose Weaponry 324
Slide 092: Multi-Purpose Weaponry 325
Slide 093: Multi-Purpose Weaponry 326
Slide 094: Unique Capabilities: Heat Extraction 327
Slide 095: Unique Capabilities: Cold Explosions 328
Slide 096: The Transmitter is the Major Weapon 330
Slide 097: Star Wars—The Final Frontier? 331
Slide 098: Comparison of EM Concepts (Chart) 333
Slide 099: Comparison of EM Concepts (Chart) 334
Slide 100: Comparison of EM Concepts (Chart) 335
Slide 101: Summary (1) 336
Slide 102: Summary (1) 337
Slide 103: Summary (2) 338
Slide 104: Summary (3) 339
Slide 105: Summary (4) 340
Slide 106: Soviet Defense Against Missiles 341
Slide 107: Surprise by Using Secret Weapons 342
Slide 108: Something Russia Doesn't Want the U.S. to Know 343
1986 SLIDE REFERENCES 344
1986 ADDITIONAL SLIDE NOTES 345
1986 APPENDIX I 371
SOME IMPORTANT INDICATORS AND DESTRUCTIVE TESTING.. 373

INDICATORS AND SUCH 373

FINAL DRESS REHEARSAL BEFORE THE BIG SHOW 384

THE SHOW BEGINS 400

MORE FRIGHTENING THAN THE MIND OF MAN HAS EVER IMAGINED 403

SCALAR EM INOCULATION AND ITS IMPLICATIONS 408

1986 APPENDIX II 413

COMPUTER WORMS AND VIRUSES, AND THE IMPACT OF SCALAR EM 415

LARCENISTS 416

ARSONISTS 418

TERRORISM 419

ANTIBODIES 419

NATIONAL DEFENSE 419

THE WEST'S VULNERABILITY 421

CORE WARS REVISITED 424

A COMMERCIAL WORM 424

ADAM OSBORNE'S WARNING 425

"NOW YOU'VE DONE IT!" 426

TEEN CHALLENGE 427

"PEOPLE ARE BASICALLY GOOD" 429

FEDERAL FUNDS 430

SCALAR EM INNOCULATION AND ITS IMPLICATIONS 431

1986 APPENDIX III 437

CONTINUING THE SCENARIO 439

2002 TABLE OF FIGURES 443

2002 ENDNOTES 447

The Russians have sustained the largest weapons development program ever launched by any nation, and they have kept it effectively hidden from prying Western eyes. I have called this program "Fer-de-Lance," after the deadly South American pit viper of that name.

The dreaded fer-de-lance is a snake of great agility and lethal effect. It often ambushes its hapless prey, striking unexpectedly and without warning. Its first sudden strike is usually lethal to its victim, which promptly expires in writhing agony. Since the Soviet development of scalar EM weapons has been designed for the same purpose, the name seems appropriate.

2002

Overview[1] and Background of the Energetics[2] Weapons Threat

T. *E.* Bearden, Ph.D.
LTC, U.S. Army (Retired)
Director, ADAS

September 15, 1998
Update

Note: We have placed this update as a section in front of the original paper. This leads to some redundancy, which remains to insure clarity.

---

[1] This is an updated and slightly sanitized version of a paper originally prepared for and transmitted to selected Senators and Congresspersons.

[2] The present name of the old KGB is the Federal Security Services. The leopard's spots remain the same, regardless of its new name—at least for a die-hard old communist faction of it. Presently the helm of Russia has been turned over to a new Premier who has an extensive KGB background. However, Putin has chosen to differ from the traditional KGB outlook and seems intent on becoming America's trusted cheap oil partner and supplier, in a bid to help his economy and pump much-needed funds into it. His friendship with George Bush may be one of the turning **points** of this new century He at least has achieved sufficient control over much of the KGB to get on with trying to solve his economic problems and break the old "die-hard communist" mold a **bit**

2002 OVERVIEW AND BACKGROUND

For some time, Russia and several other nations have possessed highly advanced "extended-electromagnetics" *(energetics)* weapons of a very novel kind, using a dramatically extended electrodynamics theory {1}. To comprehend these weapons requires a combination of non-Abelian electrodynamics in at least the 0(3) gauge symmetry {2}, general relativity, Bohm's interpretation of quantum mechanics {3}, and the use of time-domain (scalar) energy, fields, and potentials to directly enter distant 3-spatial points without propagating energy through 3-space. It also requires correcting many of the serious errors in classical electrodynamics. Probably the most elegant and applicable extant model to deal with these phenomena and weapons is Sachs' {4} unification of general relativity and extended electrodynamics, particularly as implemented in O(3) electrodynamics. O(3) allows direct engineering by modified EM means.

Most of these energetics weapons are more advanced than what has previously been known to the U.S. military, intelligence, and scientific communities, although parts of the communities are finally making real progress in understanding such "revolutionary" areas. For example, every major weapons lab on earth seems to have discovered longitudinal EM waves and longitudinal EM wave weapons—the scalar interferometry we have been discussing for so long {5}. A major contributing factor holding back U.S. catch-up is that the flawed foundations of Maxwell's 1865 theory {6} have not been changed by the Western scientific community.

Historically, the scientific community exerts resistance to substantial innovation. Often 40 to 50 years are required to do what can be done with a concerted effort in four years. The Manhattan Project in WW II is a notable example of where the scientific community made a maximum effort and did a difficult job in four years. Also, "skunk works" efforts continue to be utilized for very rapid development when something really advanced is to be done and done quickly.

With the U.S. substantially behind in these "new" strategic superweapons, a new Manhattan Project would appear to be advisable and warranted. There are at least some indications that such may actually be going on at last. We hope the indications are true.

A sobering development that occurred early this year (2002) was China's deployment of both quantum potential (QP) weapons {7} and *negative energy* electromagnetic pulse (EMP) weapons (a type which we originally called the *MindSnapper* to disguise its true nature) {8,9}. The QP weapons are the dominant weapons on Earth today, followed closely by negative energy EMP weapons in second place. We have not been able to ascertain exactly how many nations have the negative energy EMP weapons, but it is at least five at the time of this writing.

The U.S. is still well behind in the use of special "engines" (special curvatures of spacetime and their dynamics, but achieved with higher group symmetry electrodynamics) for directly engineering the mind itself, and for producing *nonmaterial* robot systems made totally of such engines—"causal system robots" which we will mention later. Several of the Western nations do accomplish work in the mind control area, but use a sort of "brute force" method oriented on frequency and frequency changes, etc. where the true active mode is the unsuspected transformations of the U(1) electrodynamics to a higher group symmetry unified field theory electrodynamics, accomplished by the human body tissues and structures themselves.

Western scientists have not realized that all EM energy in 3-space actually comes from the time domain *to* each point in 3-space and *returns from* that 3-space point back to the time domain.[3] Hence "propagation" of 3-space EM energy is actually the propagation of a "circulation" of EM energy between the time domain and 3-space, involving each 3-space point presently said to be "occupied" by the energy. Unfortunately, Western science is still unaware that *no observable exists in time,* a priori, since by definition it has no extension in the time dimension at all {10}. Instead, an observable *continually recurs* as a frozen 3-space snapshot of an ongoing dynamic 4-space interaction, due to the continual repetition of the observation process which is the application of a *d/dt* operator to the ongoing 4-interaction. Hence, we "see" physical reality (3-space reality) similarly to the way we view a "motion picture" series of successive frozen frames. The major result of this flaw (mistaking a 3-space observable as a 4-spatial cause) in Western science is that Western science still extensively substitutes effect for cause—e.g., in the notion of a separate 3-force acting on a separate 3-mass. No such situation exists in nature, and nothing "moves" or "changes" in 3-space alone. For an entity to exist in time, it requires at least 4-space because a time length is required. Nonetheless, the West is making progress. If some of its scientific dogma can be overcome, the younger scientists will break free and readily do the job required for the defense of this nation.

Mad Scramble of Soviet Science After the End of WWII

During WW II the West produced a great technical breakthrough and obtained the atom bomb as a decisive new weapon. While speeding up his own atomic

---

[3] Specifically, energy comes from the time domain to the negative 3-charge, thence to the positive 3-charge, thence back to the time domain in a *4-circulation* of EM energy. With extra energy *appearing* and present at the negative charge due to entering 3-space there, and with less energy present at the positive charge due to the energy *leaving* 3-space there, there is a delta in energy density (pressure) across a dipolarity, oriented from negative to positive. That is the exact basis of the T. T. Brown unilateral thrust force, and the force used in the "lifters" experiments today. To calculate the **force, th**e long-unaccounted Heaviside energy flow component must also be accounted— not just the weak Poynting flow component.

weapons program[4] as a result, Stalin also ordered his own Academy of Sciences to quickly derive the next great new technical breakthrough, and he held the top scientific leaders themselves personally accountable to him, to see that it was done at all speed (Figure 1).

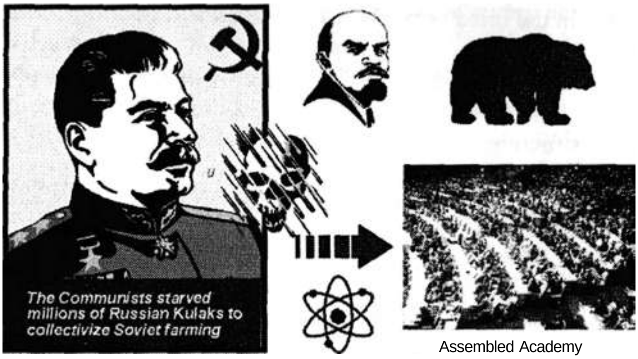
The Communists starved millions of Russian Kulaks to collectivize Soviet farming
Assembled Academy
Stalin

"The destiny of communism has been frustrated by the U S development of the atomic bomb That is not the last great technical breakthrough The next one WILL be Soviet! You will provide me with that great new technical breakthrough at all speed, or I will have your heads Do I make myself dear, comrades?"

Figure 1 Stalin's dictum removed any resistance from his scientific community.

Simply put, Stalin placed the heads of the Soviet scientific leaders on the chopping block, since he was an absolute dictator and often killed or imprisoned those who opposed his instructions. To survive, the leaders of the Soviet Academy had to demand the utmost of their subordinates, because they themselves had to report their progress or lack of it to a dictator who would and did take swift action if they did not produce results. Thus, Stalin simply bypassed the usual scientific posturing and protracted delay so typical of our Western scientific bureaucracies.[5] If the Directors of the Russian scientific community dragged their feet and did not push his desired program to the limit, their heads would roll—and some did.

---

[4] At the Potsdam conference, after Truman informed him the U.S. had just exploded a great new explosive weapon, Stalin secretly called the head of his nuclear weapons program and ordered him to accelerate the Soviet nuclear weapons program.

[5] As a simple example, the U.S. National Academy of Sciences, National Research Council, National Academy of Engineering, and the National Science Foundation—and all the electrical engineering departments at all the universities in the U.S.—have not yet recognized that all EM circuits and systems are in fact powered by EM energy extracted directly from the seething local vacuum by the source charges and dipoles in the circuit or system, and particularly by the broken symmetry of the source dipole in the generator or other primary power source. They also have not yet solved the more than a century-old source charge problem, but continue to support engineering models assuming that the charge *freely creates* real EM energy out of nothing at all, in total violation of the conservation of energy law, and pours that created energy out, freely and continuously, to establish its associated fields and potentials and their energy reaching across the universe. The agencies thus unwittingly advocate perpetual motion machines on a scale unparalleled in history. They continue to support an electrical engineering model and a classical

As a result of Stalin's forceful intervention, the entire Soviet scientific community was galvanized into a most intensive review of all of physics. The scientists desperately sought to quickly uncover potential breakthrough areas that could be developed into the great leap forward that Stalin demanded. They had to have progress, and they had to have it rapidly.

One Institute Had 2,000 PhDs and a Full Support Staff

Large review institutions were quickly set up, staffed with the best Soviet scientists and support translators and teams. One such search institute involved some 2,000 Ph.Ds along with their support staffs. Soviet ships brought in shiploads of copies of all the scientific journals of the West—from the beginning—, and they were reviewed paper-by-paper, page-by-page. Anything novel and implying new scientific possibilities, or anything that had not been adequately followed up, was laid aside in a "select" pile for further review and decision by the best Soviet theoreticians available {11}.

One gets a new physics by discovering shortcomings and flaws in the old one. This intensive Russian search almost immediately uncovered the deep flaws in classical electrodynamics—flaws that have been propagated throughout most other branches of science. Classical electrodynamics (and electrical engineering) is particularly beset by hoary flawed foundations problems, dating from at least 1865. As an example, the equations did and still do assume a material ether, and they have since the beginning! Not a single equation was changed when the

electrodynamics model which *do not even model* the active vacuum and its exchange with every charge, much less a *broken symmetry* in that exchange. So none of those agencies even knows what powers an electrical circuit, 45 years after the proof of broken symmetry was accomplished by Wu and her colleagues in 1957, and the Nobel Prize was then quickly awarded to Lee and Yang *the very same year* (1957) for predicting that broken symmetry revolution. In the modern view a charge is a dipolarity, when its vacuum interaction is considered. As a dipolarity, every charge exhibits the proven asymmetry of opposite charges, in its interaction with the active vacuum. The charge thus extracts and outpours real observable EM energy from the vacuum, freely and continuously, and it exhibits COP = oo. Those agencies—along with nearly all the rest of the U.S. scientific community—still naively assume that a COP>1.0 EM system (i) is impossible, (ii) is against the laws of nature, and (iii) is a forbidden perpetual motion machine. In nearly half a century since the discovery of the broken symmetry of opposite charges, those agencies have yet to apply that finding to electrical engineering and engineering curricula. Frankly, since they have dragged their collective feet for nearly a half century on extracting EM energy from the vacuum, what is needed is a Stalin to put his iron boot firmly to their posteriors and call their attention to it in language they cannot ignore. *The pollution of the planet for energy purposes and the imminent oil and energy wars that will ravish the earth are directly due (i) to this inexplicable dogma upheld by the scientific community and (ii) to the community's continued falsehoods equating EM COP>1.0 to perpetual motion machines.* The continuing failure to **incorporate** the **active vactnun** exchange—and its broken symmetry—into electrical engineering and classical **electromagetics a**lso poses an unacceptable threat to the continued survival of this nation.

Michelson-Morley experiments in the 1880s destroyed the material ether itself. The classical EM model also assumes a flat local spacetime (falsified by general relativity since 1915) and an inert vacuum (falsified in particle physics for several decades). It still assumes the same EM field in mass-free space as exists in mass—a non sequitur of major impact. Just to make the Maxwell-Heaviside equations easier to solve and avoid numerical methods, the *totally arbitrary* Lorentz symmetrical regauging[6] discards all EM systems far from equilibrium with their active vacuum environment. In fact, there is no "active vacuum environment" in the model, much less a broken symmetry in that exchange. Yet that very broken symmetry is exhibited by every charge and dipole {12}, and thus by every EM circuit and system. That asymmetry is involved in the continuously increasing giant negentropy poured out by every charge in the universe, thereby generating all EM fields and potentials and their energy, eventually reaching even across the universe, *and with perfect macroscopic order and time duration as long as one wishes*. Indeed, all electrodynamics is in total violation of the second law of thermodynamics, to any *macroscopic* size level and time duration desired.[7]

The terribly flawed electrodynamics model was completely overhauled by the Russian scientists *very* quickly, and the work was accomplished by some of the best nonlinear scientists in the world. Already present at the time were many things available to assist in dramatically extending {13} the hoary 1865 Maxwellian electrodynamics {6}. Gauge field theory was available since the turn of the 20th century, and the top theoreticians could readily see that something more than Abelian electrodynamics was firmly required in particle physics. The Yang-Mills {14} theory met that need in 1954 by originating non-Abelian gauge theory.

Situation in the Early 1950s

By 1950, Stalin's scientists had the beginnings of what they dubbed "energetics" {15,16}, which we would dub an "engineerable unified field theory" {17,18,19}. Longitudinal EM waves were primary {20,5} in that more fundamental approach. By the mid 1950s, one can pick out of the open literature {21} increasing incidences of prototype Russian energetics weapon testing around the world. In 1960, Khrushchev {22} spoke obliquely of these "fantastic new weapons" that were "just within the portfolio of the Soviet scientists," so to speak. Commercial aircraft sightings (Figure 2) and the reports of their captains and crews are also a particularly significant source of sightings of Russian energetics weapons tests, as are sightings by ships' crews (Figure 3) and ground observers (Figure 4).

---

[6] Actually a version of this symmetrical regauging was first performed by Ludwig Valentin Lorenz, "Ueberr die Identitat der Schwingungen des Lichts mit den elektrischen Stromen," *Ann. Phys. Chem.*, Vol. 131, 1867, p. 243-263—only two years after Maxwell published his seminal theory in 1865. The paper is also published in English, as Ludwig Valentin Lorenz, "On the identity of the vibrations of light with electrical currents," *Phil. Mag.*, Ser. 4, Vol. 34, 1867, p. 287-301. Unfortunately not much attention was paid to Lorenz's effort, since he independently derived electrodynamics. Decades later, when H. A. Lorentz also published symmetrical regauging of the Maxwell-Heaviside equations, he was given credit for it with no further mention of Lorenz. A proper historical summary of this has been given by J. D. Jackson and L. B. Okun, "Historical roots of gauge invariance," *Rev. Mod. Phys.*, Vol. 73, July 2001, p. 663-680. This is an excellent coverage of the history of who actually did what and when, and who got credit for it.

[7] This dramatically extends the size regions (microscopic through colloidal) and the time durations (momentary to two seconds) where the second law is known and proven not to necessarily apply.

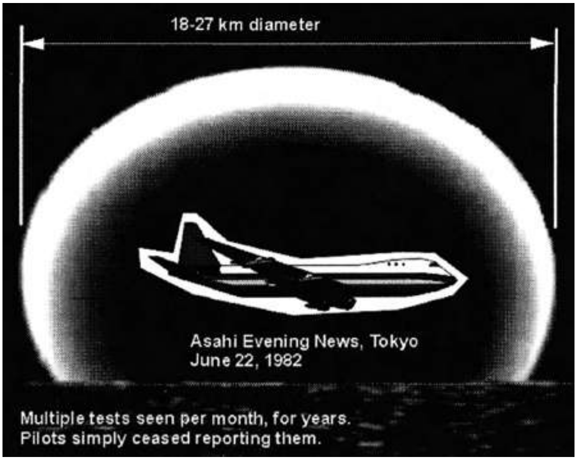
18-27 km diameter
Asahi Evening News, Tokyo
June 22, 1982
Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

Figure 2 Typical shield test, northern route to Japan, seen by passing airliner.

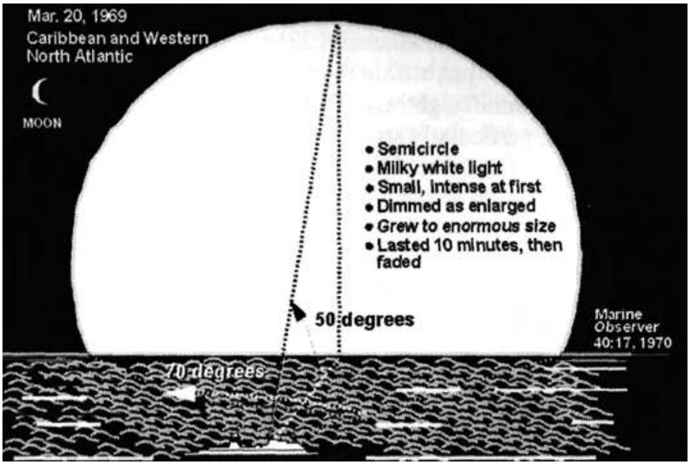
Mar. 20, 1969
Caribbean and Western North Atlantic
MOON
• Semicircle
• Milky white light
• Small, intense at first
• Dimmed as enlarged
• Grew to enormous size
• Lasted 10 minutes, then faded
50 degrees
70 degrees
Marine Observer 40:17, 1970

Figure 3 Giant missile shield test seen by passing ship's crew.

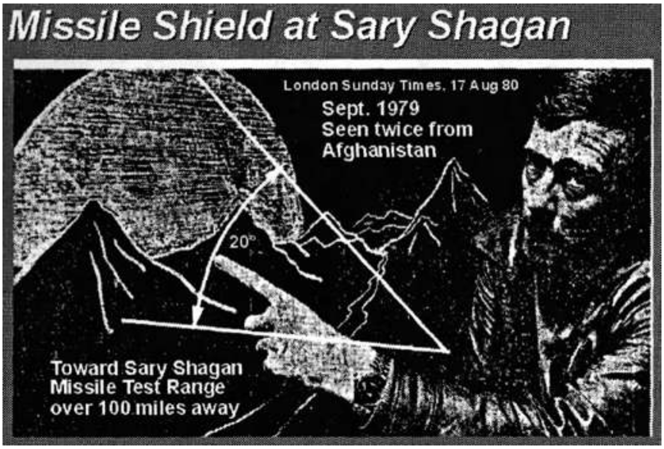
Missile Shield at Sary Shagan
London Sunday Times, 17 Aug 80
Sept. 1979
Seen twice from Afghanistan
20°
Toward Sary Shagan Missile Test Range over 100 miles away

Figure 4 Nick Downie's sighting from the ground of distant Soviet missile shield tests.

1963 Deployment and Immediate Use of Scalar Interferometers

The first huge scalar potential interferometers {23,24} of strategic range and power (Figure 5) were deployed by the Soviets in 1963, and one was used to kill the U.S.S. Thresher nuclear attack submarine (Figure 6), leaving clearly recognizable signatures {25}.

Since 1963, the Russians have had the equivalent of more than seven additional Manhattan Projects (using the Russian 5-year program instead of the 4-year Manhattan Project), back-to-back, in development of energetics weaponry. The energetics weapons have never been given to the regular Russian forces. Instead, all research, production, siting, manning, and employment are by the KGB and still under ruthless KGB control by die-hard communist factions {26}.

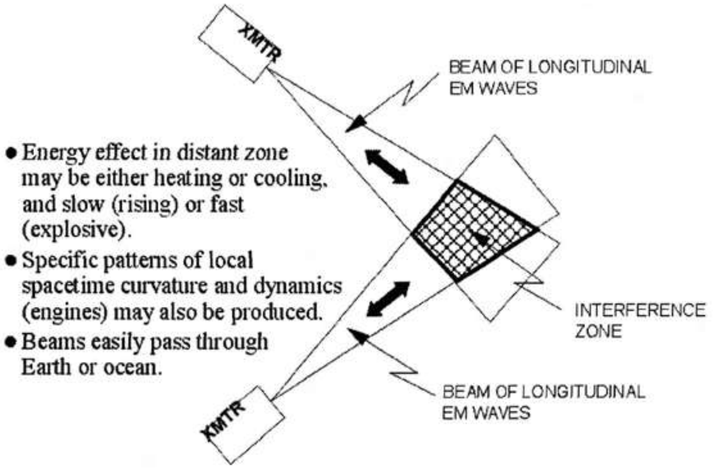

- Energy effect in distant zone may be either heating or cooling, and slow (rising) or fast (explosive).
- Specific patterns of local spacetime curvature and dynamics (engines) may also be produced.
- Beams easily pass through Earth or ocean.

Figure 5 Scalar (longitudinal EM wave) interferometer.

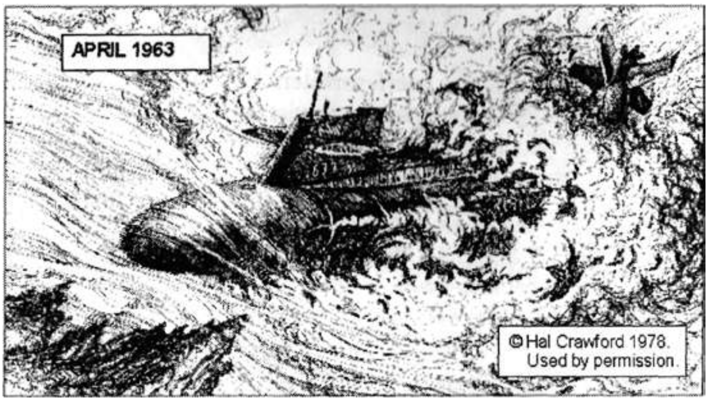

©Hal Crawford 1978.
Used by permission.

Figure 6 Kill of the U.S.S. Thresher. With her controls jammed, the hull of the doomed submarine implodes when she reaches crush depth

An Example of Russian Technology Astonishing Western Scientists

Some years ago, coordinating with a Western but non-US aerospace firm, I furnished the firm with the names and addresses of some Russian scientists in Moscow, since it was clear from their internet site that the scientists were on the fringes of the KGB energetics weapons program. An excellent European engineer fluent in Russian was sent in, several times. Several unusual things were demonstrated to him, including (i) transmitting enormous energy down a very thin wire, and (ii) cold molding, or turning metal into a liquid at room temperature, pouring it into a mold, and then letting it sit and harden. All without heating.

The engineer also closely questioned the Russian scientists about Russian . energetics weaponry, a subject the scientists were most reluctant to discuss. Finally they admitted the energetics weapons existed, and they actually obtained a meeting between the engineer and the KGB General who had directed the Soviet energetics weapon program for more than 20 years. In the meeting, the General coldly admitted the weapons, stated that "Of course you are not going to be allowed to see *those!"* and confirmed that somehow the Russians obtained everything the present author writes, as soon as he writes it {27}.

The engineer then brought the Russian scientists to his aerospace firm in that Western nation, where the scientists demonstrated cold molding to assembled scientists and some U.S. personnel. I forwarded a written explanation to the Western scientists of how the cold molding was performed technically. Later, on calling the engineer again, I discovered that the foreign authorities of that country had suddenly classified everything over there, and the engineer could no longer even talk to me without filing a Foreign Intelligence Contact Report. My own part in all this was completely open-source, pro bono, and I had already written of it openly and publicly.

Softening of Metal at a Distance Was One Effect Used Against the Challenger

At any rate, earlier I had openly advanced (and printed) the same mechanism, since it was the metal softening mechanism used by the KGB to kill the U.S. Challenger spacecraft in 1986. Publicly I had been resoundingly called a lunatic for advancing such a "preposterous" thesis. Nonetheless, metal softening by a distant LWI (longitudinal wave interferometer) was definitely utilized as part of the kill of the Challenger, as a close colleague here in the U.S. proved *experimentally,* detecting and demonstrating that metal-softening Russian signal prior to the kill of the Challenger. He amplified the signal and demonstrated on the bench its nearly instant softening of nails, rendering the nails momentarily as limp as wet noodles (Figure 7). Nails from the same box, not exposed to the signal, remained hard and could be hammered into a board. Not so the nails from the box that were exposed to the amplified Soviet signal! Those nails folded up like a wet noodle when tapped with a hammer, and could not be driven into a board.

Working unpaid with the foreign scientists, we had in fact obtained the "smoking gun" proving that this cold molding capability rigorously existed in the KGB inventory. Further, it had existed so long that the KGB had essentially *declassified* it so that the civilian Russian scientists seeking employment funds from the West could utilize it.

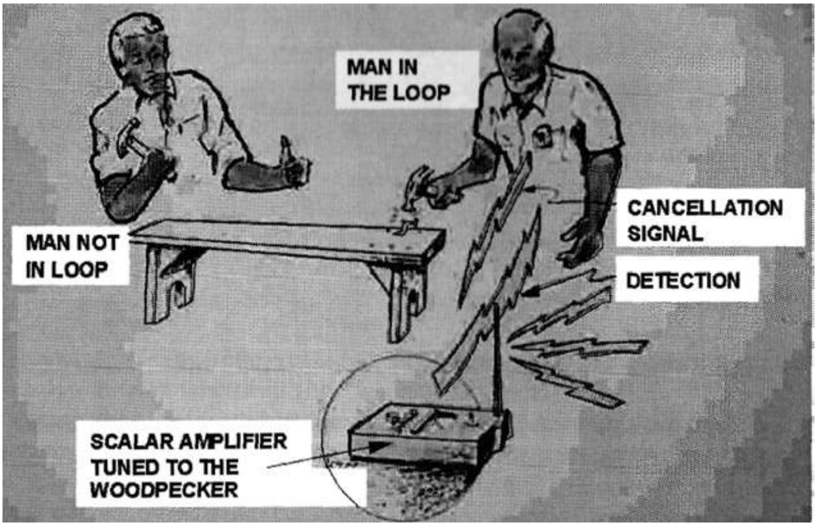

Figure 7 Proof of metal-softening signal: Golden softens nails with amplified Woodpecker signal.

Energetics: Three Branches and a Strategic Deception Plan

The *energetics* weapons use exotic polarizations of EM waves, quantum potentials, and the "infolded" LW electrodynamics (both scalar and longitudinal photons) inside and comprising every "ordinary" transverse EM field and wave. The KGB energetic weapons are categorized (Figure 8) according to what they are targeted against:

(1) against inert matter, fields, and potentials etc., that branch of weapons is called by the same name, *energetics* weapons.

(2) Against living tissue, bodies, cells, biofields, biopotentials, etc., they are known as *bioenergetic* weapons. Disease-induction electrodynamics such as used to irradiate the U.S. Embassy in Moscow are examples of bioenergetic weapons.

(3) Against the mind, thought, memory, behavior, and the mind-body coupling mechanism, they are known as *psychoenergetic* weapons. Particularly in the latter mode, these weapons also use time-polarized (scalar) EM waves and time-polarized (scalar) photons as well {28}.

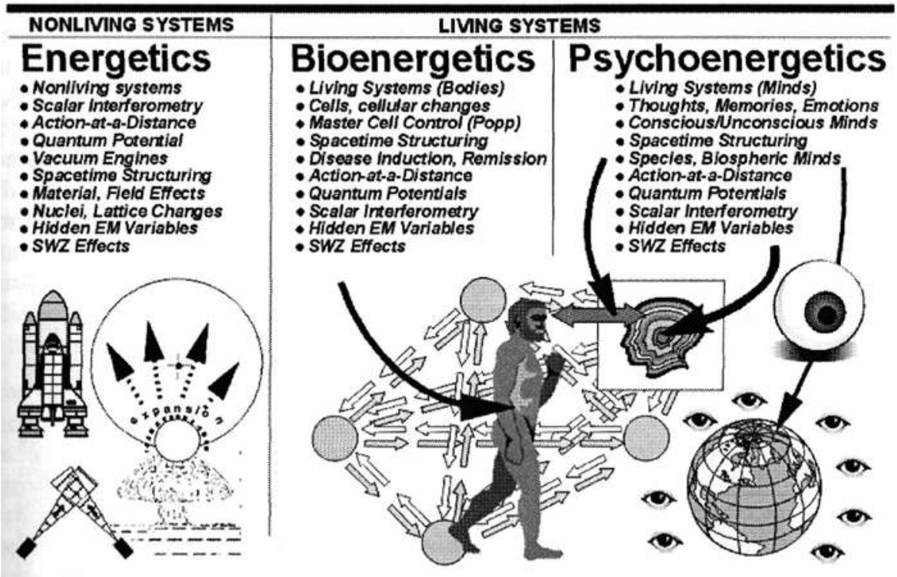
NONLIVING SYSTEMS
LIVING SYSTEMS
Energetics
• Nonliving systems
• Scalar Interferometry
• Action-at-a-Distance
• Quantum Potential
• Vacuum Engines
• Spacetime Structuring
• Material, Field Effects
• Nuclei, Lattice Changes
• Hidden EM Variables
• SWZ Effects
Bioenergetics
• Living Systems (Bodies)
• Cells, cellular changes
• Master Cell Control (Popp)
• Spacetime Structuring
• Disease Induction, Remission
• Action-at-a-Distance
• Quantum Potentials
• Scalar Interferometry
• Hidden EM Variables
• SWZ Effects
Psychoenergetics
• Living Systems (Minds)
• Thoughts, Memories, Emotions
• Conscious/Unconscious Minds
• Spacetime Structuring
• Species, Biospheric Minds
• Action-at-a-Distance
• Quantum Potentials
• Scalar Interferometry
• Hidden EM Variables
• SWZ Effects
expansion

Figure 8 Three branches of Russian energetics weapons.

The KGB also implemented an elaborate and continuing deception plan, as is standard military practice for any great strategic attack plan. The KGB deception plan has been highly successful {29} in keeping the United States convinced that these weapons do not exist, and convinced that the standard strategic nuclear weapons possessed by the West are "leading" and "superior."

Strategic Attacks Scheduled and Why We Were Not Destroyed

Indeed, the West's nuclear weapons are inferior to Russian energetics weapons, particularly to the quantum potential weapons (Figure 9) deployed at the end of 1989 and to the newer "causal system" weapons deployed since the first of 1999, where *functional robot systems* (Figure 10) are made of nonmaterial spacetime curvatures and their dynamics only! In the event of a full strategic Russian energetics strike, an opening strike with quantum potential weapons {30} will dud all nuclear weapons on earth, all nuclear power systems, and all nuclear propulsion systems in about 10 minutes. This takes care of most of the U.S. strategic armament, *within the first 10 minutes.* The publicized strategic weapons of this nation have in my opinion been almost defenseless now for at least a full decade {31,32}, but I also believe that situation may be rapidly changing for the better.

A friendly little foreign nation *does* have such weapons including QP weapons that operate in multiply connected spacetime (MCS). That little nation has been responsible for our continued survival and has continued to deter the planned KGB strategic energetics strikes upon the West. There are indications (though still not certain) that at least one—and possibly two—other friendly Western nation also possesses such weapons.

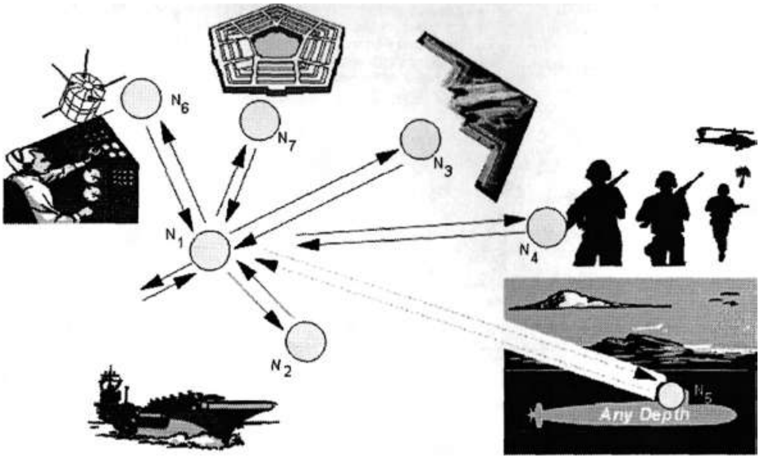
N_6
N_7
N_3
N_1
N_4
N_2
N_5
Any Depth

Figure 9 Quantum potential instantly attacking widely separated targets.

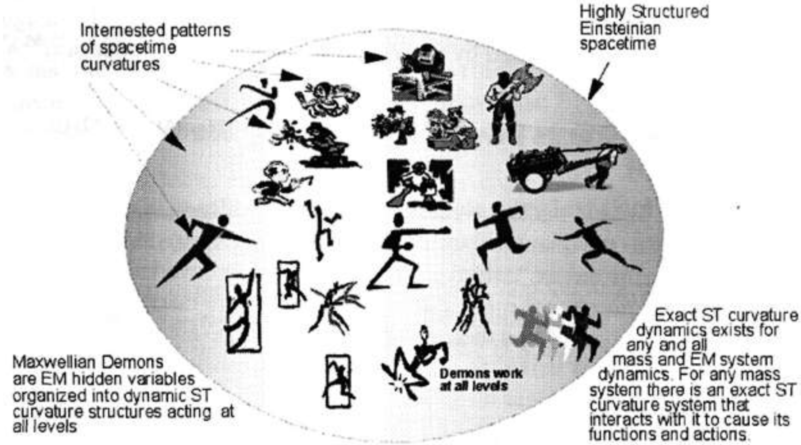
Internested patterns of spacetime curvatures
Highly Structured Einsteinian spacetime
Exact ST curvature dynamics exists for any and all mass and EM system dynamics. For any mass system there is an exact ST curvature system that interacts with it to cause its functions and actions.
Maxwellian Demons are EM hidden variables organized into dynamic ST curvature structures acting at all levels
Demons work at all levels

Figure 10 Causal system robots (CSRs). These are sets of ST curvature engines, with functions organized into a single complex weapon system with communication, propagation, weapons effects, and command and control functions incorporated. When the CSR interacts with mass, the mass performs those system functions.

Twice in 1997 alone *we were within days of the scheduled total strategic destruction of the U.S. by die-hard KGB energetic weapons,* with all-out strategic attacks actually scheduled in each case. The first attack was scheduled in latter February but was quickly countered. The attack was immediately rescheduled for May 1. The U.S. intelligence and scientific communities were apparently unaware of what was happening, and still do not believe it. However, as this author previously informed appropriate government agencies a few days prior to the attack scheduled for May 1, 1997, U.S. intelligence agencies certainly intercepted my urgent personal fax directly to the prime minister of the friendly little nation, urgently requesting that the "sane" command and control of the countering MCS/QP weapons be changed to "insane" control and spelling out the compelling reason why it was necessary.

In the insane weapons control mode, the computer periodically (say, every two hours) initiates launch sequence for the weapon, automatically. If any operating personnel remain alive on site, there is a designated short period in the sequence where the onsite personnel can countermand the order and abort the launch. If (he order is not countermanded, the launch is not aborted and the "doomsday" weapon automatically fires. In that case, KGB military power and most of Russia would disappear from the face of the earth.

This was necessary to nullify the impending strikes of MindSnapper negative energy EMP weapons {33} on those critical friendly QP weapon sites to instantly kill all personnel on site. Strike by the MindSnapper instantly kills every living thing in the area, and a protracted contamination remains whereby the struck areas emit longitudinal EM waves. If one attempts to insert new crews to resume operation of the weapons, the crewmen sicken and then die in the entry attempt. First, increasingly severe interference between mind and body control occurs. Then the individuals waver, lose consciousness, go into seizures, and their minds gradually de-couple from their bodies and they die. Fortunately, alter my fax message the C3 system was changed to the insane mode, and the Russians were notified, just in the nick of time {34}. To prevent their own destruction by the certain insane counterstrike, the KGB aborted the looming strategic strikes.

Ironically, at about the very time that the KGB strategic armada was standing down, Secretary of Defense Cohen {35} was making the following statement in Georgia:

> *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves...So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts."*

Almost certainly Secretary Cohen had no knowledge that, so to speak, a great Sword of Damocles hanging over his head was being removed by a friendly little foreign nation at that very moment, as he uttered those words apparently referring to the older longitudinal EM wave interferometer weapons. I suspect that Dr. Cohen is to this day still unaware of the incidents that were ongoing at the time of his poignant statement.

Attempts to Alert the Nation

We alerted the President and the National Security Council as early as 1984 (Figure 54). In 1998 we also furnished selected Senators, Congresspersons, and certain high-level U.S. government agencies with color copies of a 200-page briefing and expose {36} of—among other things—these weapons, their testing, and their clandestine use against the U.S. since 1963. Because of the nearly 50 years of demonstrated induction of diseases {37, 38} (Figure 11) in personnel in the U.S. Embassy in Moscow (e.g., responsible for the deaths of three U.S. Ambassadors), we also included part of the briefing on the Priore work {39} (Figure 12) showing the remarkable cellular changes (time-reversal of diseased states back to a previous earlier healthy condition) that can be and had been experimentally induced by energetics under rigorous scientific controls.

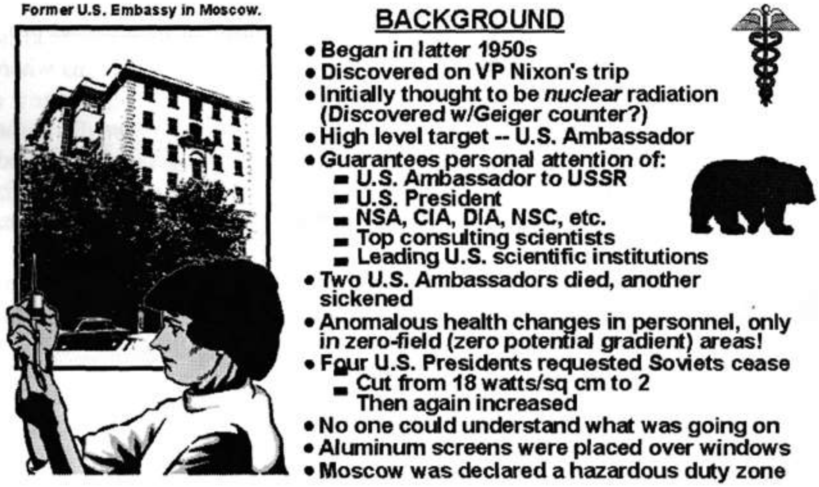
Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero potential gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2 Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Figure 11 Microwave radiation of U.S. Embassy in Moscow.

We proposed to the DoD a very rapid, high priority development program to develop portable suitcase-sized computer-controlled treatment devices (Figure 13) capable of saving a majority of mass casualties—e.g., from a professional large anthrax attack on a major population center.

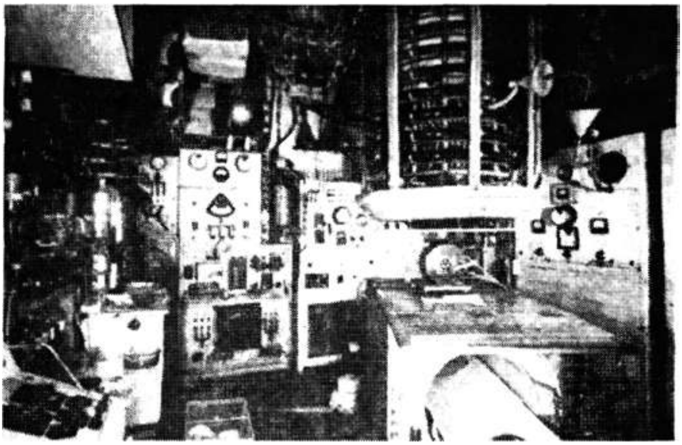

Figure 12 Priore[1] laboratory and device for treating small animals.

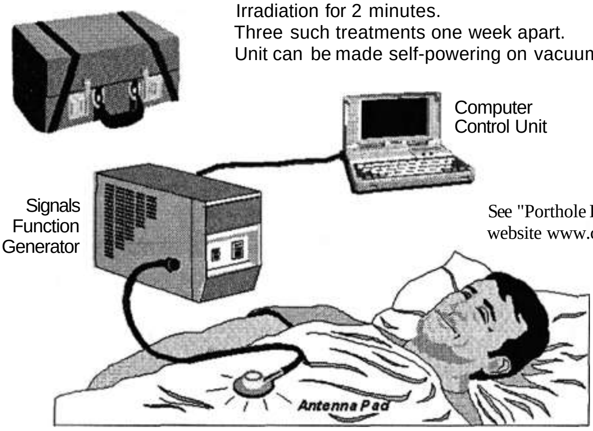
Equipment Case
Computer Control Unit
Signals Function Generator
Antenna Pad

Two blankets have antennas in them.
Uses an advanced but simple concept.
Eliminates plasma tube mixing etc.
Computer-controlled complex signal structure.
Irradiation for 2 minutes.
Three such treatments one week apart.
Unit can be made self-powering on vacuum energy.

See "Porthole Briefing" on website www.cheniere.org.

Figure 13 Portable unit for rapid treatment of mass casualties.

The Responses

The responses—or more exactly the near total lack thereof—by various U.S. government agencies were enlightening. E.g., the National Institutes of Health (NIH) never let the package out of its Policy (i.e., *political spin control)* section. It did not check a single scientific reference on a proven method documented in the hard French scientific literature for reversing terminal cancer, infectious diseases, etc. Not a single scientist called me to discuss such a proven, revolutionary effect. Instead, the NIH Policy section simply shipped the package over to the DOD as "their problem, not ours." That was because the package also explained how the Gulf War Disease was induced, including how semen can cause infection of the wife, and how the structured body biopotential of the sickened veteran can induce similar structuring of the biopotentials of his children by proximity, thus inducing various aspects of the disease.

Effective Survival Treatment of Mass Casualties Could Be Developed

Meanwhile, we had uncovered a way to dramatically speed up the Priore healing process, and use only normal transverse EM radiations for ready adaptation of existing equipment, while still forcing the same time-reversal of diseased and damaged cells back to normal. Just much faster. This meant that very small, suitcase-sized treatment units (Figure 13) could now be developed in a reasonable time—e.g., in two to three years, given a high priority, intense effort, adequate funding, and a carefully chosen scientific team.

So, we proposed a crash development by DOD of a small, suitcase-size portable device for quick, mass treatment of mass U.S. casualties following terrorist attacks on our cities using weapons of mass destruction. Eventually an innocuous reply (forced by my congressman) simply stated "it was not supported by the literature." In the package, of course, I had *specifically cited* the numerous scientific papers in the French scientific literature reporting and proving the experimental results. I had also worked 14 years to decipher the mechanism actually performing the revolutionary healing accomplished in rigorous animal testing by the Priore team. Eleven years after suppression of the Priore effort, the University of Bordeaux, which had been pressured into rejecting Priore's doctoral thesis, did approve and publish a doctoral thesis by Perisse on the work {40}.

The National Science Foundation (NSF) did not reply at all. Yet at the time the official strategic threat to the United States listed WMD attack on our cities and great population centers as the greatest threat of all (Figure 14). Today, as everyone knows we are in a war after the terrorist attack on the twin towers in New York and on the Pentagon on September 11, 2001. The U.S. government clearly recognizes that it is not a matter of if additional attacks on the U.S. cities and populace will occur, but *when.* The terrorists are also known to have already infiltrated weapons of mass destruction such as biological warfare weapons (anthrax, smallpox, bubonic plague, etc.) Nonetheless, apparently the NSF simply did not believe a novel but *rigorously demonstrated-in-the-1960s* unorthodox process that could potentially save most of those millions of Americans who otherwise will surely die—and who *are* going to die in the first major WMD strike on our population centers.

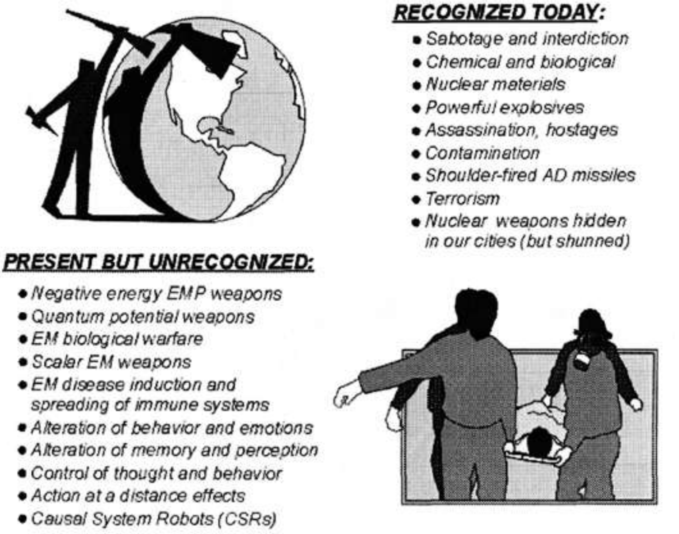
RECOGNIZED TODAY:
• Sabotage and interdiction
• Chemical and biological
• Nuclear materials
• Powerful explosives
• Assassination, hostages
• Contamination
• Shoulder-fired AD missiles
• Terrorism
• Nuclear weapons hidden in our cities (but shunned)
PRESENT BUT UNRECOGNIZED:
• Negative energy EMP weapons
• Quantum potential weapons
• EM biological warfare
• Scalar EM weapons
• EM disease induction and spreading of immune systems
• Alteration of behavior and emotions
• Alteration of memory and perception
• Control of thought and behavior
• Action at a distance effects
• Causal System Robots (CSRs)

Figure 14 The mass destruction weapons threat to our populace.

The entire scientific method is based on assigning priority to experimental truth rather than dogmatic belief. If experiments refute the theory, the theory must be changed and the experimental results accepted. Else one is not following scientific method. Apparently the National Science Foundation does not follow scientific method, nor does it seem interested in doing so. Neither do the National Institutes of Health. The decisive results of some 2,000 rigorous animal experiments reported in the refereed scientific literature in leading French journals, by multiple established French scientists, were simply ignored because the results are contrary to present dogma. Meanwhile, four years have been lost. During that four years, the portable units could have already been developed, mass produced, and made available to treat and save millions of coming U.S. civilian casualties. Apparently the lives of those Americans are not as important as maintaining the present medical dogma. When those millions do die—and they will—we shall most certainly remind those agencies of the caliber of their dedication to scientific method and of their service to their fellow Americans. In my personal view, ignoring a chance to save so many Americans is treason of a very special kind.

The Official Silence Was Deafening

Not a single scientist in any of those and other organizations seems to have taken the package seriously. Not one called me to discuss it, or apparently studied the references, etc. No one asked for a personal briefing. No one was interested in the actual photo of the strike of a Russian EM missile from a scalar interferometer, offset from a U.S. night shuttle launch (Figure 15) from Cape Canaveral in latter November 1985, or the photo of the actual registration point ball of light high in the air (Figure 16) over that same shuttle launch {41}. The same weapon then killed the Arrow DC-8 at Gander, Newfoundland about two weeks later (Figure 17). An eyewitness (or more than one) actually observed the streak of light from the sky that struck the Arrow DC-8 in the right fuselage ahead of the engines {42,43}, after the thrust of the engines were reduced during take-off by use of a negative energy (cooling) interferometer. A similar ball of light associated with the kill of a Titan missile fired from Vandenberg AFB in early 1986 {44} and printed in *Aviation Week & Space Technology* magazine, was also apparently of no concern.

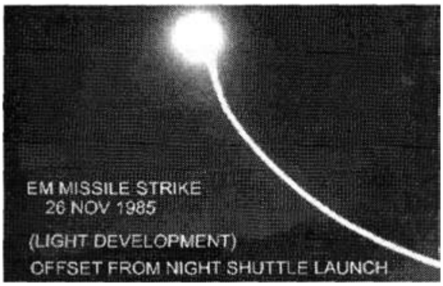

PHOTO BY BOB GLADWIN

Scalar interferometer operator fires offset EM missile from Russia.

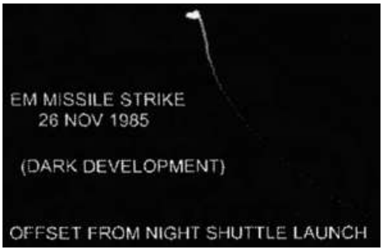

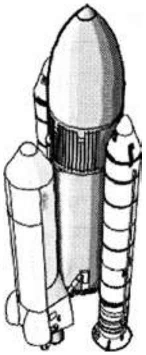

Figure 15 Soviet EM missile strike offset from U.S. shuttle launch at Cape Canaveral on 26 Nov. 1985. This is the smoking gun.

These remarks are intended to show the lack of understanding that has existed in our governmental, scientific, and intelligence communities of what energetics is and of the great energetics superweapons that have been developed. With the present rapidly growing interest in longitudinal EM waves, however, we suspect that such lack of understanding is now no longer the case. We certainly hope that is true, and we are

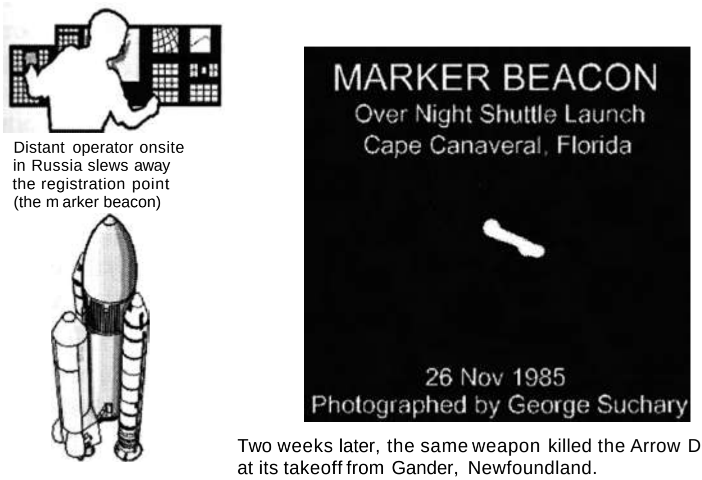
MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida
26 Nov 1985
Photographed by George Suchary

Distant operator onsite in Russia slews away the registration point (the marker beacon)

Two weeks later, the same weapon killed the Arrow DC-8 at its takeoff from Gander, Newfoundland.

26 Nov1985

Figure 16 Marker beacon above U.S. shuttle launch at Cape Canaveral, 26 Nov. 1985. Russian crews were using shuttle launches as practice ICBM launch targets.

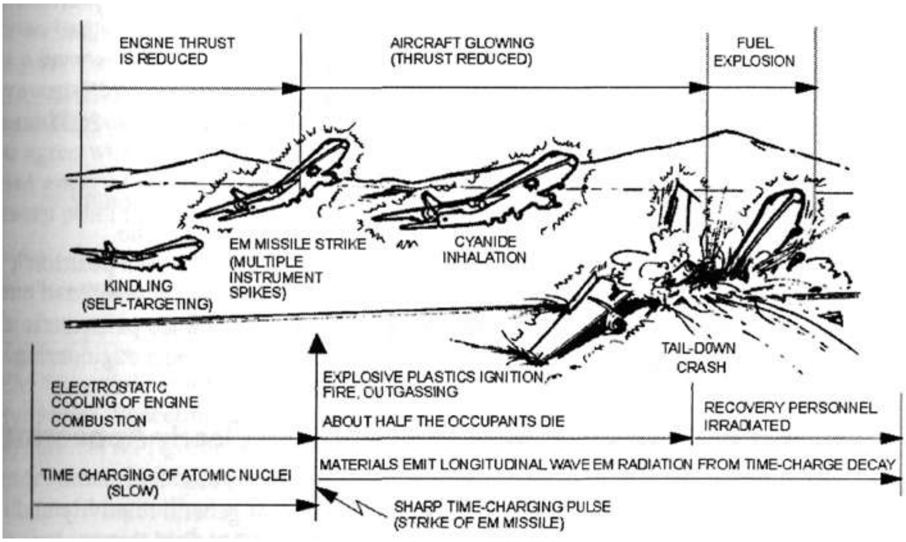
ENGINE THRUST IS REDUCED
AIRCRAFT GLOWING (THRUST REDUCED)
FUEL EXPLOSION
EM MISSILE STRIKE (MULTIPLE INSTRUMENT SPIKES)
CYANIDE INHALATION
KINDLING (SELF-TARGETING)
TAIL-DOWN CRASH
ELECTROSTATIC COOLING OF ENGINE COMBUSTION
EXPLOSIVE PLASTICS IGNITION, FIRE, OUTGASSING
ABOUT HALF THE OCCUPANTS DIE
RECOVERY PERSONNEL IRRADIATED
MATERIALS EMIT LONGITUDINAL WAVE EM RADIATION FROM TIME-CHARGE DECAY
TIME CHARGING OF ATOMIC NUCLEI (SLOW)
SHARP TIME-CHARGING PULSE (STRIKE OF EM MISSILE)

Figure 17 Kill of the Arrow DC-8 at Gander, Newfoundland Dec. 12,1985.

much encouraged that at least one—and possibly two—friendly Western nations now possess quantum potential weapons.

The Long Dry Spell in Correcting Electrodynamics May Be Slowly Ending

In over 20 years of unrelenting effort, until 1999 we were unable to get any significant Western scientists of real "horsepower" (e.g., in non-Abelian electrodynamics, foundations errors in electrodynamics, advanced gauge field theory, general relativity, etc.) to meticulously examine conventional U(1) electrodynamics and its glaring errors, and to look at weapons implications of higher group symmetry electrodynamics.

Beginning about three years ago, however, a noted scientist, Dr. Myron Evans {45,2}, along with other theoretical scientists of the Alpha Institute's Institute of Advanced Study (AIAS), have been extending electrodynamics by developing the 0(3) electrodynamics. They also have deeply examined the Whittaker work of 1903 and 1904 on which the early part of the Russian energetics weapons are based {24}.

The AIAS analyses, and subsequent papers by Dr. Evans and Alex Labounsky as well as by the entire AIAS group, have clearly established the primary importance of time-like precursors of EM fields, operating in massfree spacetime where fields E and B cannot exist. These time-like precursors *create* all the rest of the conventional electrodynamic entities {46}. This is a dramatic extension to electrodynamics and physics across the board. More than 100 papers have been produced to date, with more than thirty published in journals such as *Foundations of Physics,* and many of the others are either accepted or still in the referee process with various journals. Some 60 of the papers were published as a single special edition of the *Journal of New Energy* (JNE) in latter 1999, and a second edition of *Modern Nonlinear Optics,* Wiley, 2001 has been published with many important papers.

However, the problem is that the scientific community, which is strongly committed to "business as usual", has funded none of this work. The conventional lead agencies of the community are still in the "ostrich position", with their heads buried very firmly in the sand with respect to the terrible damage they continue to enforce on this nation by their continued perpetuation of an old, long-archaic, inadequate electrodynamics and electrical engineering.

Engineerable Unified Field Theory Is Now Clearly Appearing

The rapidly emerging AIAS extension of electrodynamics has "met in the middle" with Sachs' {4} revolutionary generalization of general relativity and electrodynamics. For the first time, an *engineerable* unified field theory—employing 0(3) electrodynamic means—is now being produced. The AIAS work in 0(3) electrodynamics was shown to be an important subset of Sachs' approach. It also now provides the full, rigorous theoretical foundation and model for the Priore mechanisms and results.

We also comment most strongly that the important SU(2)xSU(2) electrodynamics as developed by Barrett (47) leads to essentially the same conclusions, since Barrett's electrodynamics is homomorphic with 0(3) electrodynamics. Barrett, of course, is one of the pioneers of ultrawideband radar—another scientific area that was fiercely opposed and suppressed by the orthodox U.S. establishment for some time. Ironically, today those very scientists who were so dogmatic in their vociferous condemnation of UWB radar, have assumed the mantle of "experts in the UWB field", sometimes even taking credit for having "advanced it" in the first place.

Examples of Some Major Non Sequiturs in Conventional Electrodynamics

In every textbook and millions of technical papers in the West, electrodynamicists routinely calculate the "magnitude of the scalar potential". Actually, not a single one—including Maxwell—calculated or calculates the *magnitude of the potential itself.* Instead, they universally calculate the *reaction cross section* of the potential at a point occupied by an assumed intercepting/collecting unit point *static* charge {48}. At best that is a representation of the *intensity* of the field at each point. The reaction cross section of the potential with a fixed unit point static charge is indeed a scalar value. The "scalar" potential itself, on the other hand, *is not* its own reaction cross-section at each point of itself, and it is not a scalar entity at all! Instead, it is a harmonic bundle of bidirectional EM longitudinal phase conjugate wavepairs, as shown by E. T. Whittaker in 1903 {49}. It is and always has been a multi-vectorial, multiwave entity. By slightly reinterpreting Whittaker's work to agree with quantum field theory, the scalar potential is also a circulation of EM energy from the time domain to 3-space and back to the time domain, at every point in space that is occupied by the potential.

Further, the "field" notion is utilized in two self-contradictory manners. On the one hand, it is assumed in the massless reaches of space, and on the other hand it is assumed in mass, *which is the only case in which a force field is properly defined.*[8] The so-called "electromagnetic field" existing in empty space is not an EM force field at all, but a curvature or curvatures of spacetime with assorted ilynamics. Spacetime is after all active, and being active means curvatures and ilynamics. As rigorously shown by the AIAS theoreticians, what exists in *spacetime*—before the force-fields appear on reacting charged mass—are time-like and longitudinal scalar potentials and fluxes. These "pre-EM field" entities in fact interact with mass to produce all the force fields (the electromagnetic fields). Mass is a *component* of force; simply by F == *d*p/dt, which is expanded into F = d(mv)/dt, which is ma + v*(dm/dt).* Note that both terms have mass as a component. Hence there is no force or forcefield when mass is absent. Present mechanics is seriously wrong (and has been wrong for hundreds of years) in assuming a separate massless force acting upon a separate mass.

---

[8] Since Maxwell and **the** original pioneers believed in a material ether filling all space, to them t**here** was not a single point anywhere in the universe that was devoid of mass. Hence to them there was only one kind of field: **the** kind that exists in mass.

Jackson, one of the ablest classical electrodynamicists of our day, adroitly states how electrodynamicists skirt that issue. Quoting, p. 249 of his *Classical Electrodynamics,* second edition: *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."*

Our comment is that one does not get an observable until after the interaction of a cause with a previous effect, to provide a change (a new effect). The 4-cause itself—in this case the field that actually exists in the vacuum—is never observed nor is it observable. Its *effect* (force with an affected observable mass as a component of it) is observed. The field in space and the field in matter thus do not even have the same fundamental units, but differ by a factor of m. In electrodynamics the m is avoided by "defining" the field as force per unit coulomb of "charge" (without saying or accounting "charged mass").

Wide Confusion of Cause and Effect

Present electrodynamics (and even the hoary old mechanics, as pointed out) has thoroughly confused the causal (4-space) side of the observation process with the effects (3-space) side—observation yields *3-spatial* entities, not 4-spatial. Indeed, as is well known, *time is not observable, even in theory.* That is really because observation is a *d/dt* process applied by the 3-space intermediary (e.g., mass) being acted upon by the causal (4-space) entity in spacetime, to produce a change in that intermediary (3-spatial) that results (is observed). *A priori,* the output of the observation process does not even exist in time, since its fundamental units are LLL and not LLLT. The observation process is thus *d/dt(LLLT)* => LLL.

Further, electrodynamics omits the fact that, not only does the causal entity act upon the intermediary to produce the *effect* (changes to the intermediary), but *those changes (effects) also act back (in reverse) through the intermediary to produce changes in the causal entity itself.* This mutual interaction between cause and effect (but never stressed in exactly that manner) is well known in GR, where the curvature of spacetime acts upon mass-energy to change it, and the resulting change in mass-energy also acts back upon spacetime to curve it. The omission of the latter "back reaction" in EM theory by Maxwell resulted in omitting half the energy, half the wave, and half the causal entity in spacetime, as well as half the effect. The missing half of the effect appears as the Newtonian third law reaction, which in electrodynamics is just mystically "assumed" to be an effect without a cause.

The vast importance of restoring this missing "back action of effect upon cause" is because it can be deliberately utilized to directly engineer the causal (4-space) state itself, producing nonmaterial "engines" and unobservable functioning robot systems comprised entirely of curvatures of spacetime and their dynamics. Such "causal system robots" (CSRs) can be designed and engineered to then perform almost any desired action upon normal mass systems they encounter. The robots also can easily penetrate vast depths of mass, which is mostly empty space filled with potentials, fields and waves—all of which are just bundles of longitudinal EM waves and their dynamics. "Mass" is actually a vast superhighway through which such CSRs can easily move at the speed of light {50}.

CSRs: Engineering the Unobserved Causal Side To Create Functional Systems

Understanding early on the mutual interaction between cause and effect that existed in their energetics, the Russians were able to engineer this totally new kind of system: the *causal system robot* or CSR (Figure 10). With sufficient theoretical development, one can "work backwards" by adroitly re-engineering the effect to obtain a desired causal system (spacetime curvature set along with its impressed dynamics; (Figure 18) corresponding to some physical system and its changes one desires to engender. One thereby directly creates a deterministic set of spacetime curvatures and impressed dynamics, which we call an "engine" (Figure 18). This *ST curvature engine* system can then function as a completely separate system {51} (Figure 10). One can communicate with it via longitudinal EM waves, control it, etc. By building in scalar interferometry functions, the **CSR** can be given "weapons" capabilities, etc.

- All levels of energy structures and all levels of time structures mold spacetime geometry.
- The "pattern" is called a *template* for a vacuum engine (spacetime curvature engine).
- Spatial and temporal structurings of the engine act upon mass at all levels.
- This produces a template of forces, for precise translations and stresses within the mass at all levels. The forces exist in both time and 3-space.

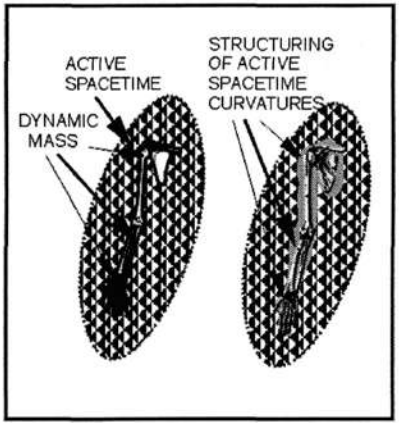

Figure 18 The principle of general relativity extended to include the concept of the vacuum engine.

In theory such a CSR can be assembled to perform any desired function or complex of functions upon matter, fields, potentials, etc. that is desired. In short, one can build real, functioning "infolded" (subspace) robot weapon systems using this "backward engineering of spacetime itself process. The huge underground facilities (hundreds) that the Russians continued to build after the collapse of the Soviet economy and the fall of the Soviet Union, apparently contained the full CSR development and control facilities as well as the operational systems employed for this totally new kind of weapon system.

A great engineering of the future lies in engineering CSRs, which already constitute a great leap forward past the material nanobot technology being pursued in the West.

Making the First CSR is Extremely Difficult; Making a Million Copies Is Easy

The development of a specific CSR with a specific set of functions is akin to developing a highly complex software system of major proportions (several tens of millions of lines of code). The job is extremely difficult, and tortuous "debugging" is required for a lengthy period. Once a single CSR is completed and debugged, however, the production of thousands or even many millions of copies is then a trivial task.

The CSR may be visualized as a sort of vast complex of longitudinal EM waves (recall, we are speaking of ST curvatures, not the conventional EM waves thought to operate on a flat spacetime. *All* EM waves operate in a curved spacetime *a priori*, as clearly shown by Sachs). It turns out that all matter is mostly empty space, filled with potentials and fields. Space itself, considering its energetics, may be taken as a giant scalar potential. All those potentials and fields and waves are in fact nothing but sets of longitudinal EM waves and dynamics. In short, they are vast superhighways for the travel of CSRs and the propagation of longitudinal EM waves.

A CSR can be embedded within any kind of EM potential, wave, or field. So to replicate one, one embeds it in a magnetic signal and records that signal on a diskette or other media. Then one makes millions of copies of the diskette, for perhaps 10 cents or less each. Or do it en masse, by first completing a CD-ROM with a hundred or two hundred CSRs embedded in its signals (Figure 19). Then simply reproduce the CD-ROM. Every reproduction has another hundred or two exact duplicates of those CSRs, made for almost no cost at all. With communications built into the CSRs, the command and control facilities take charge of them, give them individual numbers or designations, and a robot armada emerges with most unusual capabilities indeed.

The puzzling underground facilities in Russia (which our fellows apparently never deciphered and which remained a mystery) apparently contained the scientific teams, the research and development facilities, the production facilities, and the command and control facilities for the new vast CSR weapon complex. It was to see its first great employment at the end of 1999.

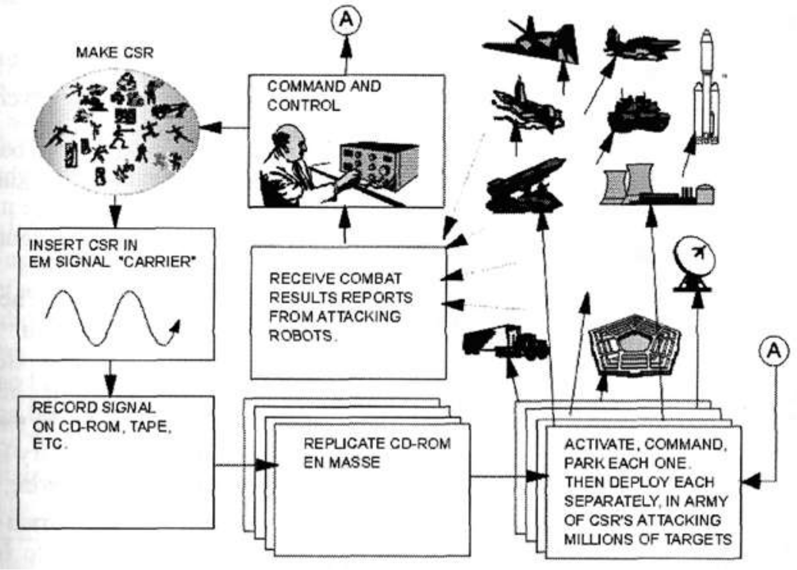
A
MAKE CSR
COMMAND AND CONTROL
INSERT CSR IN EM SIGNAL "CARRIER"
RECEIVE COMBAT RESULTS REPORTS FROM ATTACKING ROBOTS.
A
RECORD SIGNAL ON CD-ROM, TAPE, ETC.
REPLICATE CD-ROM EN MASSE
ACTIVATE, COMMAND, PARK EACH ONE. THEN DEPLOY EACH SEPARATELY, IN ARMY OF CSR'S ATTACKING MILLIONS OF TARGETS

Figure 19 Mass reproduction of a CSR is simple and inexpensive. The CSR resides in, or propagates inside, any EM signal, potential, or field.

In 1999 The KGB Deployed Massive Numbers of CSRs for a Strategic Attack

In latter 1999 the KGB began massively deploying these CSRs of various types throughout the Western world, particularly including the United States. They simply sent the CSRs back "inside" the potentials and fields and waves being emitted from radio stations, TV stations, power lines, satellites, communications systems, inside matter, etc. These CSRs were deployed and infiltrated in major command posts and installations throughout the U.S. and its allies. True to form, the KGB even "stimulated" some major computer systems in critical U.S. government facilities by having the robots "surface" some effects on the Computers, right through the firewall programs (whose EM signals and operations indeed provided nice superhighways for the robots). Our fellows **detected** the interference as a "sophisticated computer hacking attack" (puzzlingly from the Russian Academy of Science), and then the robots "submerged" again after deliberately performing actions to produce obvious changes and guarantee the U.S. detection of those changes. The purpose of this "stimulus and probe" was to see if the U.S. was aware of CSRs and CSR **technology.** We clearly showed we were not, and we still are not aware of them. Fur**ther,** the entire government apparatus—including all our vaunted national laboratories—knows absolutely nothing at all about CSRs, and none of the laboratories believe it. That is the same posture they've been in with respect to energetics weapons for a half-century. Apparently they will be in that same posture for another half century, left to their own devices.

However, I was not allowed to remain a casual onlooker for this new CSR armada! The KGB also planned to use me as a stimulus to ping the government yet another way, and then assess the government's response.

Here in Huntsville, suddenly I was personally struck by such a CSR right from my computer monitor, with the attack riding through the light from the monitor to my eyes and retinas and right into the brain and nervous system (Figure 20). This was the type of CSR that can lurk in any potential or field inside a computer, and then strike and kill the operator looking at the screen. The CSR took over control of the machine, split the screen so that only one-inch horizontal strips were luminous—one at the top and the other at the bottom of the screen with the middle dark. These two separated horizontal strips allowed scalar interferometry via the inner LWs comprising the light from the two strips. The LWI signals crossed in my vision center, where suddenly ordinary EM signals and waves were produced by the scalar interferometry occurring there.

Stimulating apparently the only analyst capable of recognizing a CSR attack and connecting it with the scheduled giant strategic CSR attack.

KGB General commanding the energetics weapon previously confirmed that what this analyst writes on his computer is received as soon as he writes it. Apparently accomplished by Tempest techniques or directly by energetics.

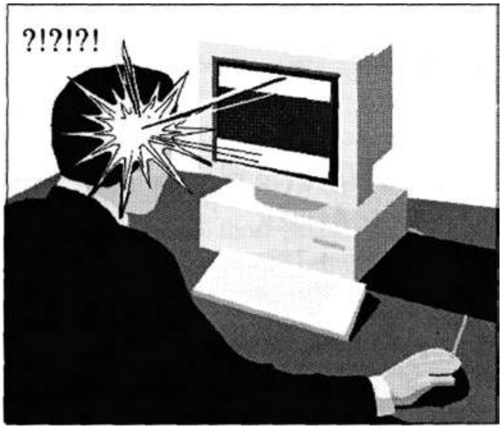

Figure 20 Present author struck by CSR in latter 1999 as a deliberate nonfatal stimulus.

I saw rapid blinding flashes like a hundred flashbulbs exploding at once right in my face. However, the signals representing those "flashes" were being created right there in my vision center. An external camera would have recorded only the weirdly separated screen display. It *would not* have shown the brilliant flashes of light I was seeing, unless it was set up stereoscopically. The strike was deliberately sustained for only about 10 seconds so as to be severe but nonlethal; about 30 seconds would be totally lethal. Violent fibrillation of the heart also resulted explosively, and I was instantly hypnogogic and seriously weakened by sudden hypoxia from near total lack of blood pumping (because of the fibrillation). The purpose was apparently to strike the only analyst in the U.S. who would recognize what the nature of the attack was, stimulate him, watch him try to notify the government, and then observe the government's reaction. It look four days to recover from the strike. But I knew immediately what had made the strike, and that I had been hit with a CSR weaponized by having a built-in scalar interferometry capability. It was also a very sophisticated CSR, as it showed by assuming complete "intelligent" control of my computer's functions.

When I finally recovered, I did notify the government and warn them of CSRs, exactly as anticipated by the KGB. As might be expected, the government's total disbelief and lack of reaction clearly showed they knew nothing of CSRs. They simply did not believe such weapons existed or even could exist. Sadly, once again one faced the "not invented here" syndrome. General relativity and its provision for spacetime curvatures and their associated dynamics (engines) operating back upon mass are apparently just supposed to remain a theoretical curiosity. Why, obviously no scientist in his right mind would think of weaponizing such!

A more correct answer is, *"Like hell they wouldn't!"* *American* scientists would not think of it, but *Russian* scientists with their national chess-playing psychology will immediately seize upon any opportunity—even many moves ahead in the game—and develop it and utilize it. They understand and use the boa constrictor. We only understand and use the immediate rattlesnake.

How the Asymmetrically Deployed Strategic Attack Was Stopped in Its Tracks

However, we also alerted the "little nation" that has advanced EM weapons technology, which—since early 1997—has been saving the U.S. by forcing the abortion of impending KGB energetics weapons attacks.

When I warned them in 1997 about the MindSnapper weapons and mind-control weapons used against Captains Button and Svoboda, unknown to me at the time they later reacted and suddenly destroyed those facilities in Russia (apparently with violent scalar interferometry strikes). So *that* mind-control type of attack being readied for 1999 was thwarted for a few years by these strikes sometime prior to mid-1998.

But to return to the CSRs. In the latter quarter of 1999, the Russians were developing and deploying throughout the West numerous kinds of CSRs. So the period at the end of 1999/first few days of 2000 attack was already scheduled for a massively programmed CSR attack. It was designed to cripple the entire **nation,** and perhaps even collapse it economically, yet all would be blamed on Y2K! Our scientific ignorance guaranteed that we would blame it on something **mundane that** our scientists could understand: Y2K problems in all the old software programs running so much of our automated systems, communications, and centralized functioning, accounting, banking, etc. In short, the KGB planned to give us a bicentennial gift: a Y2K nightmare far beyond our wildest fears.

After my recovery from the CSR strike in the latter quarter of 1999, I again urgently alerted the friendly nation of the development of CSRs and their widespread deployment—and informed them of the personal attack on me and specifically why it was made nonfatal. Of course I also alerted our own government. We were scheduled for a *real* Y2K problem that would curl one's hair, so to speak, with everything imaginable failing—and probably all blamed on Y2K.

Admirably, beginning just at the end of 1999, the friendly nation simply began methodically and massively cleaning out all the CSRs, and they disposed of them in very short order! Such CSR systems[9] may easily be located and destroyed by strategic longitudinal EM wave weapons {52}, when and if one has the longitudinal EM wave technology sufficiently developed. So Y2K came and went with only a few relatively minor incidents from the *normal* Y2k problems expected.[10]

However, unknown to our scientists and intelligence community, a great war of a very special kind—*the Earth's first strategic subspace war*—had just been silently fought and won, right there inside our power systems, command and control systems, major weapon systems, electrical power systems, electrical control systems of our nuclear submarines and strategic bombers and missiles, even inside the electronics of our nuclear warheads, etc. And not a bobble showed on the surface, anywhere. This was the first real "subspace" war, where weapons made of pure spacetime curvatures and dynamics were involved as the main offensive battle systems. LWI destruction of numerous CSRs inside potentials and fields in various electronic equipment throughout the nation, created little or no overt signs at all.

This time we won, thanks to the "good guys" and not to our own scientific community. Next time we may not be so fortunate.

The Mostly-Ignored Time-Energy Interaction

For some inexplicable reason, physicists in the West have omitted one-half the actual photon interaction.[11] If one allows the spatial energy component of the photon to interact, then one must also account for the simultaneous interaction of the time-component of the photon. A mass m absorbing a photon (ΔE)(Δt) becomes *masstime* (m+dm)t, not mass m. Mass *m* and masstime *mt* are as different as are force *F* and impulse *Ft.* No scientist in his right mind would ignore the *t* in *Ft* and consider just the force *F.* Yet that exact kind of giant *faux **pas*** has been done for many decades now in the photon interaction. And it is still ubiquitously done in physics. *Mass does not emit photons; masstime does.* Again, physicists confuse effect and cause.

---

[9] These early CSRs apparently had no "stealth" capabilities built-in, so they were easily detected and destroyed, rather like shooting fish in a rain barrel.

[10] Even so, a normally disastrous Y2K problem was narrowly averted, in the nick of time, by a concerted national software reprogramming effort massively mounted and costing billions of dollars. Otherwise, the expected "ordinary" Y2K problem would have been very bad indeed.

As the Russian energetics weapon scientists know full well, time is actually spatial energy compressed enormously, by at least the factor c^2. Hence time has at least the equivalent energy density of mass. The greatest amount of "energy" existing in a photon—particularly in lower frequency photons—is in that "time" part that is arbitrarily ignored and discarded. An ELF photon has enormously more total energy—by many orders of magnitude—than does a gamma ray photon.

Energy and time components in the photon are canonical. Low frequency photons are reduced in the magnitude of their *spatial* energy component and thereby maximized in their *time* components. For the ordinary (decompressed) *energy equivalency* of the time component, one must multiply the time in seconds by c^2. Halving the frequency, e.g., doubles the time component, whose spatial equivalent energy is now $(2t)\times(c^2)$. With respect to the original photon, the new total energy (achieved by halving the frequency) is given by $[(\Delta E)/2] \times [(2\Delta t) \times c^2] = [(\Delta E)(\Delta t)\times c^2]$. So halving the photon's frequency increases its total energy by a factor of c^2, which is approximately by a factor of $9x10^{16}$. Hence photons actually carry enormously greater total energy (most is locked up as time) at low frequency than at high frequency. Indeed, the highest energy photon interactions in the universe are at low frequency, *if and when* a transduction of the available time-energy into spatial energy occurs.

This has been completely missed in the West, but not by the Russian energetics weapons scientists. They have been *engineering* that formidable new high energy but low frequency physics—a physics far more energetic than that used in our most energetic colliding beams in "conventional high energy physics".

The cold fusion processes[12] stumbled onto by Western scientists are using the

---

[11] I.e., except for accounting for the overall "action" (angular momentum) and such things as **minimum** action theorems, etc. Treating the photon's increment of time specifically as a formidably compressed piece of spatial EM energy does not appear in Western physics. It has **long** been utilized in the KGB's secret energetics weapon science.

[12] For an explanation of the use of time-reversal zones in cold fusion experiments, and some of the resulting new transmutation reactions that result, see T. E. Bearden, *Energy from the Vacuum: **Concepts** and Principles*, "Chapter 10: Cold Fusion: Low Spatial-Energy Nuclear Reactions at High Time-Energy," Cheniere Press, Santa Barbara, CA, 2002. Available from www cheniere.org. Statistically, the formation of significant time-reversal zones is guaranteed by **the transient** fluctuation theorem of Denis Evans et al., and by recent experimental work by Evans et al. proving that such zones (where the reactions run backwards) can and do occur at up to **micron** (colloidal) size and for up to two seconds in duration. We stress that this is a recognized, **proven** violation of the second law of thermodynamics already known to physics.

time-energy component of photon interactions, to generate enormous localized EM energy directly upon the interior quarks and gluons of nucleons {53}. We worked out a speculative explanation of the active time-energy transduction mechanisms that fits all the experimental phenomena, including the eerie instrument anomalies occurring in rigorous electrolyte experiments at China Lake for some years. That is included in my new book, *Energy from the Vacuum: Concepts and Principles,* Cheniere Press, Santa Barbara, CA, 2002, available from www.cheniere.org.

Of course consistency alone does not prove it conclusively; consistency is a necessary but not sufficient condition for proof. But at least the proposed explanation is consistent, and nothing else seems to be, at this time. In one chapter in my referenced new book we do cover the cold fusion mechanism and give specific new nuclear reactions producing the excess deuterium, tritium, and alpha particles in so many successful cold fusion experiments worldwide.

My "time-reversal zone" approach is stringent, because the recent work by Denis Evans *et al.* {54} has clearly shown that reactions can and do run in reverse—totally in violation of the second law of thermodynamics—at micron (colloidal) level in size and for up to two seconds in duration (and sometimes even a little longer). So experimental demonstration that the "time reversal zones" I assumed can and do indeed occur has now been accomplished by Evans *et al.* This provides strong experimental and theoretical support for the basic assumption (of the formation and momentary existence of localized time reversal zones that subsequently decay) we used to derive the fundamental cold fusion mechanism including some of the specific reaction equations for excess deuterium, tritium, and alpha particles now demonstrated in hundreds of successful experiments.

In a reversed zone *the law of attraction and repulsion of charges is reversed,* so that like charges attract—sometimes so closely that each enters the strong force region of the other, forming a quasi-nucleus (Figure 21). As the time-reversal zone subsequently decays back to a time-forward zone, the reduced strong force is restored much faster than the Coulomb force is reduced and then reversed back to normal. So during this readjustment decay process, the quasi-nucleus will simply change to an isomer and a known nucleus by quark-flipping to change an H+ ion to a neutron. Note that the "reversal zone" momentarily changes the Coulomb barrier—always the only thing that has really prevented chemical fusion and transmutation at low spatial energy—to a Coulomb attractor. Hence cold fusion—as now backed up by more than 600 successful experiments worldwide, by multiple laboratories and researchers in multiple nations—is an established fact and a herald for a vast new science based on deliberate and intentional reversal of the Coulomb barrier and use of time-energy.

With its usual perspicacity, the organized U.S. scientific community has flubbed it—and flubbed it badly—with respect to cold fusion. Its contribution has largely been a tirade of ad hominem attacks, slander, and libel of a viciousness seldom seen in the annals of science. Once again, the community is far more devoted to defending its cherished dogma (and its own rice bowls from the conventional Big Nuclear Science that cold fusion thoroughly upsets), than it is in applying scientific method which calls for *believing the experiments* and changing the dogmatic theory.

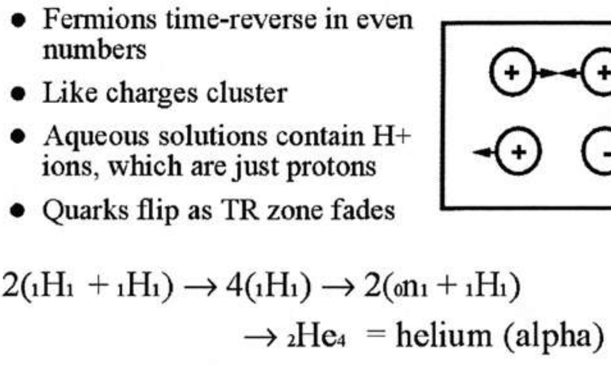

- Fermions time-reverse in even numbers
- Like charges cluster
- Aqueous solutions contain H+ ions, which are just protons
- Quarks flip as TR zone fades

$$2({}_1H_1 + {}_1H_1) \rightarrow 4({}_1H_1) \rightarrow 2({}_0n_1 + {}_1H_1) \rightarrow {}_2He_4 = \text{helium (alpha)}$$

Figure 21 Example of a time-reversed zone nuclear reaction common in successful cold fusion experiments.

Development of Threats and Scheduled Attacks Aborted

KGB energetics weapons have long constituted our single most pressing yet largely unrecognized national security problem. That problem has also increased with the growing development and deployment of such weapons by other nations such as China. Some 10 nations, for example, now have scalar interferometry weapons, and even the Japanese Yakuza has them and also has their own secret facilities for producing them in Japan, including small portable units developed for use in assassination and terrorist type warfare within a targeted nation—the United States of America.

Unknown to our government, we narrowly escaped destruction twice in the first half of 1997 from scheduled all-out strategic KGB energetics weapons attacks. **Only** the actions of a friendly small country countered the attacks and caused the **KGB** to abort on each occasion.

When the second scheduled Russian strategic attack was aborted just prior to May 1, 1997, a new KGB counter-counter for the friendly counter used by the **little** nation in latter April 1997 was already rapidly being deployed. The new **Russian** counter-counter—seizure and control of the mind of an operator at a distance, and complete control of his actions while he performs sophisticated technical tasks—was tested in two separate incidents {55}, first on Captain Button (Figure 22, Figure 23) and then on Captain Svoboda about two months later (Figure 24), each flying an A-10 Warthog aircraft over the U.S.

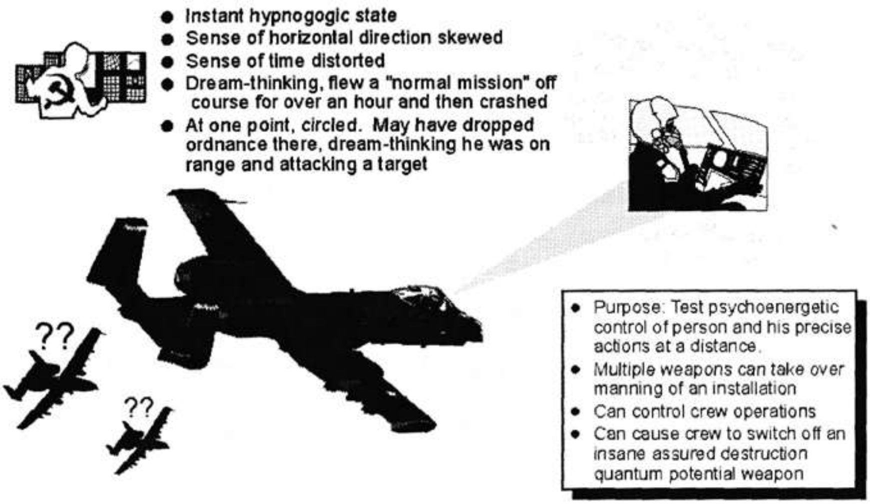

- Instant hypnogogic state
- Sense of horizontal direction skewed
- Sense of time distorted
- Dream-thinking, flew a "normal mission" off course for over an hour and then crashed
- At one point, circled. May have dropped ordnance there, dream-thinking he was on range and attacking a target

- Purpose: Test psychoenergetic control of person and his precise actions at a distance.
- Multiple weapons can take over manning of an installation
- Can control crew operations
- Can cause crew to switch off an insane assured destruction quantum potential weapon

Figure 22 Captain Button's bizarre flight to his death on 2 April 1997.

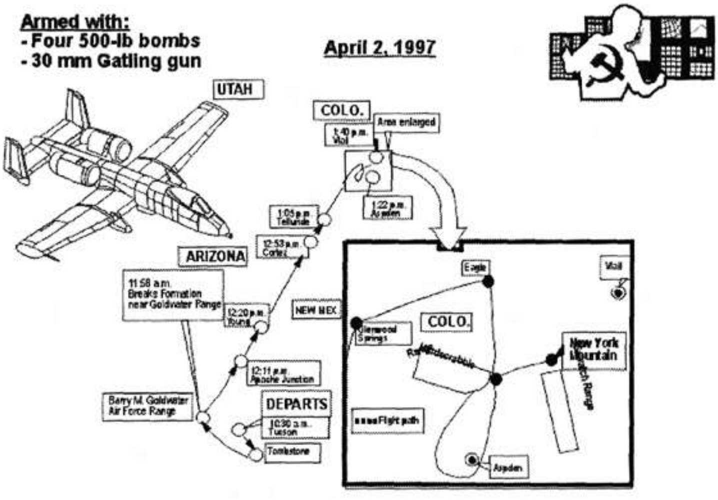

Armed with:
- Four 500-lb bombs
- 30 mm Gatling gun

April 2, 1997

UTAH
COLO.
Area enlarged
1:40 p.m. Vail
1:22 p.m. Aspen
1:05 p.m. Telluride
12:53 p.m. Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
NEW MEX.
12:11 p.m. Apache Junction
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a.m. Tucson
Tombstone
Eagle
Vail
Glenwood Springs
COLO.
New York Mountain
Flight path
Aspen

Figure 23 Captain Button's flyaway path to his eventual fatal crash.

The highly successful test against Captain Button (Figure 22, Figure 23) is particularly significant, as we have explained. Captain Button was controlled while performing highly technical operations, for more than a full hour, and one hour was apparently the acceptance test specification. Captain Svoboda (Figure 24) simply had her mental perception of up and down instantly reversed while climbing her A-10 warthog from low level ordnance delivery on target. Erroneously perceiving she was diving, she immediately "corrected" sharply to "climb", thereby diving headlong into the ground and perishing in the resulting explosion of her aircraft.

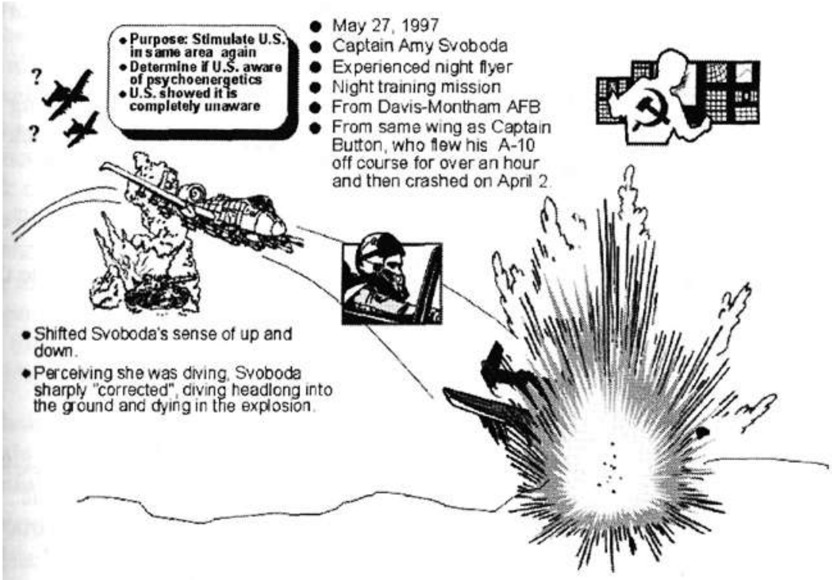
• Purpose: Stimulate U.S. in same area again
• Determine if U.S. aware of psychoenergetics
• U.S. showed it is completely unaware
?
?
● May 27, 1997
● Captain Amy Svoboda
● Experienced night flyer
● Night training mission
● From Davis-Montham AFB
● From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2
• Shifted Svoboda's sense of up and down.
• Perceiving she was diving, Svoboda sharply "corrected", diving headlong into the ground and dying in the explosion.

Figure 24 Climbing out of low-level ordnance delivery, Captain Svoboda suddenly dives headlong to her death on 27 May 1997.

Svoboda's kill demonstrated a *tactical destruction* usage, whereas Button's test demonstrated a *strategic* hostile usage of the adversary's defending weapons crew to take action against the adversary's defending strategic weapons system. A new KGB strategic energetics attack against the United States—using the mind control (psychoenergetics) weaponry to take over the operators of the little nation's QP weapon sites and stand the sites back down for maintenance—was rescheduled apparently for latter 1997 to mid-1998 or so. It appears that the extensive preparations for that attack were totally destroyed clandestinely and very suddenly by the friendly little nation some time after we warned them, as we mentioned above.

However, still another Russian counter—the development of "weaponized en**gines"** (functioning causal system robot weapons, "infolded electromagnetic" in **form)** consisting of organized and functioning systems of spacetime curvatures operating in and traveling through EM fields, potentials, waves, and signals—came on line (Figure 10, Figure 18). The KGB simply altered the attack plan for a Millennium Strike and completed the forward deployment of these eerie new CSR systems by introducing millions of them directly throughout U.S. facilities, command centers, weapon systems, power grids, ships, aircraft, submarines, missiles, nuclear warheads, computers, communications systems, etc.

We estimate that several million such systems were introduced in that manner, and readied for the Millennium Strike. The United States and its allies were and are almost totally unprepared for this dramatically new kind of strategic warfare that has been developed and deployed against us. No such weapons have previously existed in history. Consequently our national survival and the survival of our allies was once again at stake—and right on schedule during the critical period from mid-1999 through mid-2000 that we had predicted back in 1997.

The die-hard elements of the KGB indeed had a new millennium gift for us, but it was not at all the kind of gift we would appreciate. The KGB is not a monolith! This plot apparently did not include Putin, and it did not include the rest of the KGB, particularly that element reporting to Putin.

Live Destructive Tests Over and Near the U.S. Have Continued

Without elaboration, we believe that the Learjet crash incident and the EA-990 crash incident may have been very sophisticated tests of these new robotic causal EM systems. We also believe that the long years of Russian construction of over a thousand mysterious underground facilities, even though the Russian economy was collapsed and the citizens and soldiers were destitute, is significant. Those facilities apparently contain the set of command and control systems for this vast new strategic weapon system.

The facilities also include the "factories" where the CSRs are made, because the factories are unlike any other factories ever conceived. They do not build "physical" systems, but something more like "software." Only now the "software program" itself is an independent robotic system, with a full set of weapon systems functions, and capable of independent navigation and functioning anywhere, while remaining under KGB command and control via longitudinal EM wave communications infolded inside the potentials of earth, ocean, and space. The facilities also contain the operational forces, headquarters, and communication and control systems that maintain, deploy, and employ (fight) these CSR systems, and direct the battle.

The purpose of the 1998 document we sent to DoD—which did not include the causal robot system armadas later deployed throughout everything in the West—was to inform selected members of the Senate and Congress of the energetics threat situation and summarize it for them. Accordingly, we presented the background of the development and testing of the KGB energetics weapons since WW II, in some detail, for selected Legislators and their staffs. We also included some specifics on the failure of the U.S. scientific community to keep American science and technology abreast of that of our adversaries.

The Scientific Basis For Energetics Weapons Is Established

As pointed out above, the scientific basis for our early use of Soviet scalar EM interferometry is now rigorously established by work accomplished by the Alpha Foundation's Institute of Advanced Study. It is also rather "old hat" now **that** all major weapons labs have discovered longitudinal EM waves and are **actively** working on and with them. The Director of the AIAS, Dr. Myron Evans, is an eminent scientist and theoretician with over 600 papers in the hard science literature. He and his colleagues also have had zero funding for their work on the higher group symmetry electrodynamics necessary to understand such things as energetics. A combination of non-Abelian electrodynamics, advanced gauge field theory, and the impact of E. T. Whittaker's work shortly after the turn of the century has shown that scalar interferometry is real, and that weapons based on scalar interferometry can be a potent threat on the battlefield. The time-like (time-energy) basis for electrodynamics is now clearly established.

With this important new work, we are now only some 40 years behind the **Russian** weapons scientists! But we are beginning to close the gap—and hopefully that will be done very rapidly.

In 1999 after being personally attacked by a CSR, I again contacted the office of the leader of the friendly small nation and again furnished all the information we possess on the new CSR systems programmed by the KGB as a countering method against the small nation's QP weapons. The outcome was that the little nation destroyed the CSRs in time, as we stated above. We were lucky that (i) the KGB did attack me with one of the weapons, deliberately to "get my strong and immediate attention", and (ii) deliberately did not kill me with the strike. Otherwise, there would have been no warning to the little nation, and the U.S. a**nd** its allies would have experienced a "Y2K" problem worse than their wildest fears, possibly collapsing the economy of the United States and much of the developed world.

Our Situation Here In the U.S. Is Precarious

Meanwhile, our own scientific community apparently does little or nothing in **these** areas, except to castigate, slander and libel anyone daring to suggest that such "out of the box" weapons areas even exist. E.g., there does not appear to be a single U.S. scientist who knows how to create a *time-polarized* EM wave.[13] There do not seem to be any concerted programs to deal with t-polarized photons and t-polarized EM waves. Our scientists still firmly believe that one has to propagate spatial EM field energy through space to affect a target electromagnetically, when that was obsoleted by Russian "subspace EM longitudinal wave propagation" weapon systems deployed as early as 1963, and was completely obsoleted by deployment of quantum potential weapons in 1989.

Yet a close colleague of the present author has single-handedly developed and demonstrated prototype superluminal communication systems using the infolded (longitudinal EM wave) electrodynamics, and these systems would have been heading onto the market in latter 2001 except for the sudden bankrupting of the major backing company. There has not been any funding forthcoming for his revolutionary work, but only clumsy attempts to take it from him. It is my fear that the further clumsy takeover attempts will be successful, and this desperately needed technology will never see the light of day in defense of our nation. As a single example without further discussion, his work revolutionizes the present notions used in the rapid development of quantum communication and quantum computing, because it revolutionizes the very notion of the Q-bit, completely removing the statistics of the quantum state and making it engineerable. This of course is completely counter to present QM notions, but is consistent with the Bohm hidden variable approach to quantum mechanics.

Meanwhile, our own weapon scientists apparently still do not recognize the possibility of a superluminal longitudinal EM wave communication system, and appear to have no motivation or intention to develop such. This is in a way quite sad; quantum tunneling at more than four times the speed of light, of a clearly recognizable Mozart symphony, has already been demonstrated in a section of a waveguide {56}. Whatever one wishes to call it and however one wishes to do it, superluminal communication of intelligible signals has been clearly demonstrated in the laboratory. For the "infolded" longitudinal EM waves, all the "normal" or "envelope" EM waves, potentials, and fields are just great *superwaveguides* for such longitudinal EM wave "tunneling through subspace" at superluminal speeds.

---

[13] Keeping one's sense of humor, apparently there has also not been a U.S. scientist who even recognized and knew what actually powered an electrical circuit, until our solution to the source charge problem and publication of it in 1998. Eerily, even after the technical basis for a solution had been proven in particle physics in 1957, no one seemed to have applied the broken symmetry of opposite charges to the common dipole and to a dipolarity such as a scalar potential and a charge (with its associated virtual charges of opposite sign, in the modern view). Everyone accepted that the fields and potentials and their energy are established by their associated source charges, but no one had stated that the source charge receives its input energy from its asymmetry in its seething energy exchange with the active vacuum.

The Dominant Weapons Are Quantum Potential Weapons

Full-bore deployment of Russian quantum potential weapons occurred by the end of 1989, the first lab prototype weapon having been tested in April 1986 against the U.S. air attack on Libya (Figure 25). Beside Russia, two other nations—Brazil and the "little friendly nation"—also had developed quantum potential {57} weapons. The United States did not develop them, or at least apparently not just yet.

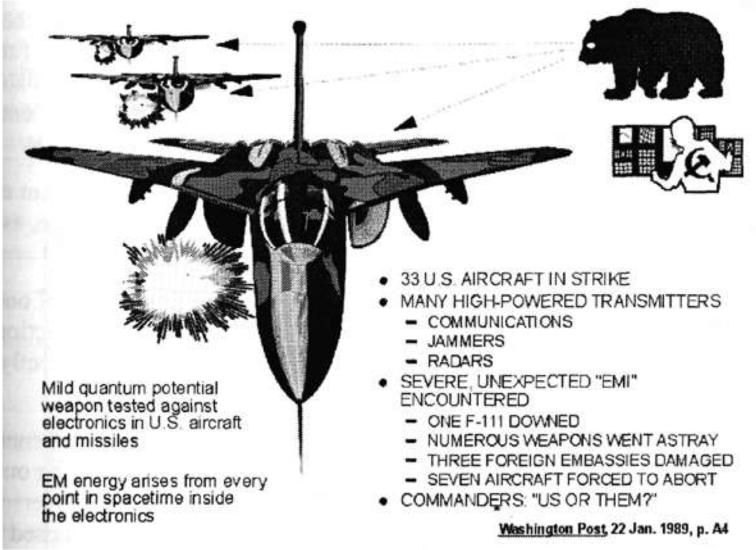

Mild quantum potential weapon tested against electronics in U.S. aircraft and missiles
EM energy arises from every point in spacetime inside the electronics
• 33 U.S. AIRCRAFT IN STRIKE
• MANY HIGH-POWERED TRANSMITTERS
- COMMUNICATIONS
- JAMMERS
- RADARS
• SEVERE, UNEXPECTED "EMI" ENCOUNTERED
- ONE F-111 DOWNED
- NUMEROUS WEAPONS WENT ASTRAY
- THREE FOREIGN EMBASSIES DAMAGED
- SEVEN AIRCRAFT FORCED TO ABORT
• COMMANDERS: "US OR THEM?"
Washington Post, 22 Jan. 1989, p. A4

Figure 25 U.S. air strike on Libya, April 1986.

Burlier this year (2002), China also deployed QP weapons, and we now find that an*other friendly* nation has them as well. That brings the total to five, as of this writing. Two of those are friendly, and a third is astraddle of the fence, so to speak. The other two are Russia (the KGB) and China.

At this point China could go either way, so it represents a future enigma. China **has** control of the Panama Canal, and has inserted more than 200,000 Chinese into Panama—which is on the way to gradually becoming a Chinese territory, as soon as the Chinese in Panama outnumber the native Panamanians. In addition, under the Clinton administration China gained access through two U.S. ports (with little checking, so anything can easily be inserted). China also has a**dditional** ports in the Western hemisphere. China is almost certain to eventually attack and seize Taiwan, which puts her on a collision course with the U.S., si**nce** the U.S. has an agreement to defend Taiwan. She has also declared the South China Sea—through which passes some 60% of the oil bound for Japan, a**nd** much oil for other nations also—to be Chinese territorial waters. China also has developed bioenergetics weapons, having long ago induced cancer into most of the GRU Russian representatives then serving in China. She almost certainly has or is rapidly working to develop psychoenergetics weapons as well.

So far as I am aware, there still exists not a single funded U.S. scientific program to directly correct the glaring errors in our own classical electrodynamics—such as those problems deplored by Nobelist Feynman and Wheeler {58,59,60}. Both knew that the force field concept in space was erroneous. As they put it, only the "potential" for the force field exists in space, in case some charged mass is available there for interaction. But the field itself is created by the interaction and is the effect of it. It does not exist before the interaction occurs. Yet all our texts continue to teach the disinformation of the "field in space." And our science community continues to fiercely defend such dogma, attacking and sometimes destroying the careers of those brave scientists who do try to change it.

One notes that time is multiply-connected, since in theory the same instant exists at every point in the universe. Or said another way, in any instant in time, every point in the universe simultaneously exists.

Operating in the time-domain as a causal system, and then turning out of one point there into any point in 3-space desired, is a means of producing "action at a distance" that obsoletes our old concept of transmitting the energy directly through space {61}.

Presently the U.S. scientific community has little predisposition toward energy turning into the time-domain, operating in the time-domain, then turning out of the time-domain into any desired point in 3-space, anywhere in the universe.[14] Every 3-space point in the universe simultaneously exists in (is superposed in) each point in time, as those points in time occur. In one sense, time can be considered to be the total multiple connectedness of space! We suspect that it may even be possible to interpret "instantaneous" communication—as by a quantum potential—in that "in here-out there" instantaneous fashion via "tunneling through the multiple connection of the time channel".

Finally, mass—being an observable and 3-spatial—does not exist in time. Mass continually turns into masstime (by photon absorption), then masstime turns back into mass (by photon emission). Masstime exists in time, but mass does not. U.S. scientists also seem unaware that the so-called "rate of flow of a mass through time" is generated by the sum total of all photon interactions (both virtual and observable) with that mass. The ubiquitous interaction of a mass with photons also causes that mass to continually recur, and thus to seemingly "continuously exist." It is continually observed, but it does not continuously exist! Since the photon interaction is engineerable, then it follows that a specific mass can be time-reversed back to an earlier state {62}.

---

[14] If one models the time domain in three dimensions instead of one, e.g., then in theory it becomes possible to "flip" a 3-dimensional mass object into the time 3-domain at a single "moment," then rotate from that "time 3-point" back into 3-space at a new 3-space location anywhere in the universe. Hence the "beam 'em up Scotty!" of the old Star Trek (and the more modern beaming version) may hold more of a promise than just for science fiction entertainment.

Cellular Time Reversal: Revolutionary Healing by a Team of French Scientists

Priore and eminent French scientists {39,40} unwittingly used the mechanism for time-reversing a mass and every part of it, to time-reverse diseased and damaged cells, including tumor cells, physically changing them back to their previous healthy physical state and condition in vivo. This included all parts of **the** cell, including its genetics. The Priore team demonstrated revolutionary cures of terminal tumors, infectious diseases, atheriosclerosis, and restored depressed immune systems.

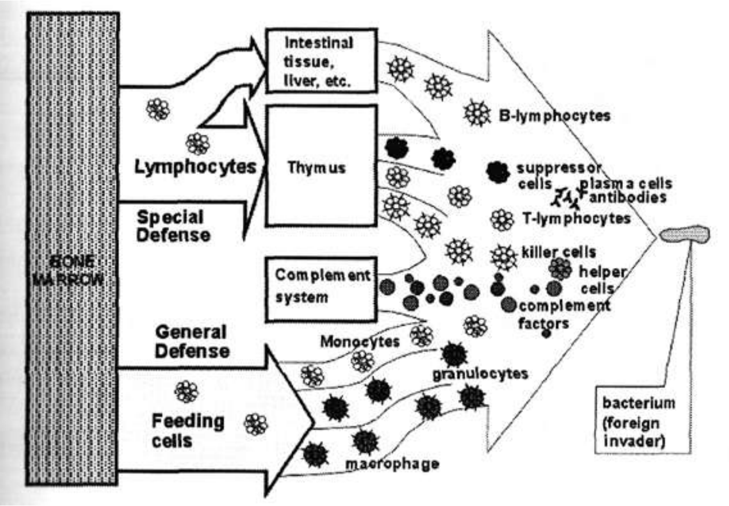
BONE MARROW
Intestinal tissue, liver, etc.
Lymphocytes
Thymus
Special Defense
Complement system
General Defense
Feeding cells
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
Monocytes
granulocytes
macrophage
bacterium (foreign invader)

Figure 26 Simplified diagram of the immune system.

The immune system (Figure 26) *heals nothing, not even its own damaged cells.* The human cellular regenerative system {63} is responsible for all healing in the body—and it utilizes time reversal of the damaged cells as the active healing mechanism. The regenerative system (Figures 27, 28, 29) pumps the damaged cells with longitudinal EM waves (and therefore with time-polarized EM waves, which always accompany longitudinal EM waves), inducing the cells and all their parts to add the phase conjugates, thereby forming time-polarized EM waves and pumping in the time-domain rather than the 3-space domain. This dramatically extends nonlinear optics to the time domain, and mass pumped in that fashion is time-reversed (propagated back along its own time-track, to a previous physical condition.

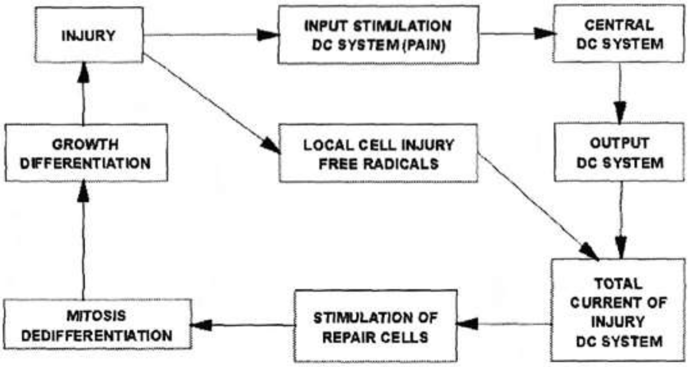
INJURY
INPUT STIMULATION DC SYSTEM (PAIN)
CENTRAL DC SYSTEM
GROWTH DIFFERENTIATION
LOCAL CELL INJURY FREE RADICALS
OUTPUT DC SYSTEM
TOTAL CURRENT OF INJURY DC SYSTEM
MITOSIS DEDIFFERENTIATION
STIMULATION OF REPAIR CELLS

Note: Becker only had conventional EM available, hence was unable to model the inner structuring and inner dynamics of the fields and potentials. *So he missed the deeper "engine and antiengine" mechanisms.*

***Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.**

Figure 27 Becker's theoretical control system involved with response to injury.

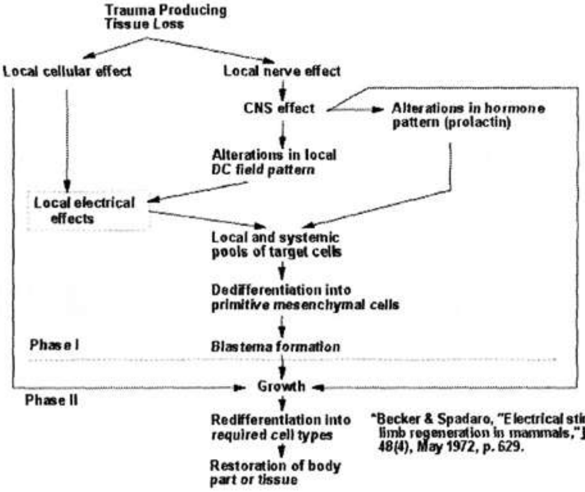
Trauma Producing Tissue Loss
Local cellular effect
Local nerve effect
CNS effect
Alterations in hormone pattern (prolactin)
Alterations in local DC field pattern
Local electrical effects
Local and systemic pools of target cells
Dedifferentiation into primitive mesenchymal cells
Phase I
Blastema formation
Growth
Phase II
Redifferentiation into required cell types
Restoration of body part or tissue

***Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.**

Figure 28 Becker's proposed control system governing regeneration.

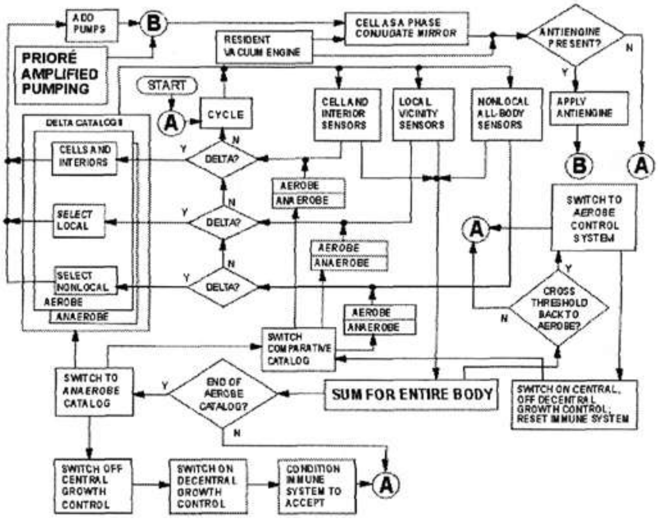
ADD PUMPS
B
CELL AS A PHASE CONJUGATE MIRROR
ANTIENGINE PRESENT?
N
RESIDENT VACUUM ENGINE
PRIORÉ AMPLIFIED PUMPING
START
Y
CYCLE
A
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NONLOCAL ALL-BODY SENSORS
APPLY ANTIENGINE
DELTA CATALOGS
CELLS AND INTERIORS
DELTA?
N
B
A
AEROBE
ANAEROBE
SWITCH TO AEROBE CONTROL SYSTEM
SELECT LOCAL
DELTA?
A
AEROBE
ANAEROBE
SELECT NONLOCAL
DELTA?
CROSS THRESHOLD BACK TO AEROBE?
AEROBE
ANAEROBE
AEROBE
ANAEROBE
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
END OF AEROBE CATALOG?
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL, RESET IMMUNE SYSTEM
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT
A

Figure 29 Author's deeper block diagram of the cellular regeneration system.

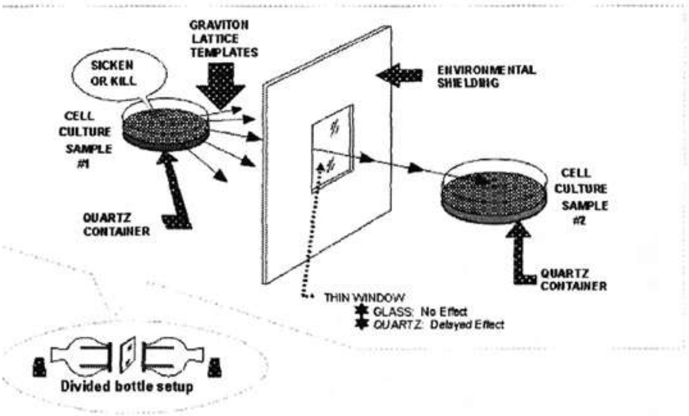
GRAVITON LATTICE TEMPLATES
SICKEN OR KILL
ENVIRONMENTAL SHIELDING
CELL CULTURE SAMPLE #1
QUARTZ CONTAINER
CELL CULTURE SAMPLE #2
QUARTZ CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle setup

Note:* *Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice. Consider particularly the dielectric pathway in the emitting diseased cells and in the receiving targeted healthy cells.

Figure 30 Kaznacheyev's electromagnetic induction of cellular disease and disorder.

Kaznacheyev has shown that essentially any cellular disease or disorder can be electromagnetically induced in cells at a distance (Figure 30). Together with Becker's work, this shows that the cells can be altered either detrimentally (disease induction) or beneficially (disease elimination) by purely **electromagnetic** means. It would seem that diseases such as AIDS, e.g., could be **rather** dramatically cured by use of whole-body longitudinal EM wave radiation (with its simultaneously time-polarized EM wave radiation), as was done for other diseases by the Priore team and properly reported in the French scientific literature. We have also uncovered a process and method for causing the body to directly produce the time-polarized EM wave pumping, reducing the irradiation to perhaps 30 seconds of pumping instead of two or three hours {64}.

Becker's Study of the Cellular Regeneration System

In all our medical science, the only true healing procedures being utilized appear to be the use of Becker *et al.'s* tiny EM stimulation across otherwise intractable bone fractures {65} (Figure 31), to dedifferentiate (time-reverse) red blood cells, then redifferentiate (time-forward) them again twice to form osteoblasts, which are deposited in the fracture site to make new bone growth. All the other medical procedures seem to be *interventions,* not *healing.* After the intervention—which of course may be urgently necessary—it is then up to the body to heal itself (restore its cellular and tissue damage and functioning). So we have little *healing* science today, and in fact our medical establishment hardly funds any *healing* studies at all. EM radiation is usually considered as something that simply heats tissue. Our present medical science is deeply engrossed in studying the immune system. Again, the immune system heals nothing, not even its own damaged cells.

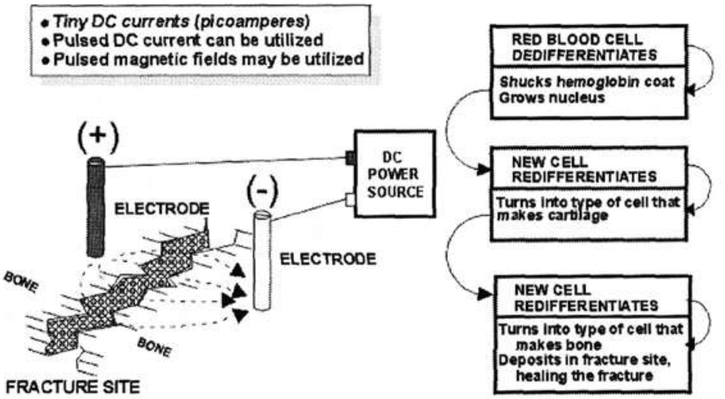
• Tiny DC currents (picoamperes)
• Pulsed DC current can be utilized
• Pulsed magnetic fields may be utilized
(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE
RED BLOOD CELL DEDIFFERENTIATES
Shucks hemoglobin coat
Grows nucleus
NEW CELL REDIFFERENTIATES
Turns into type of cell that makes cartilage
NEW CELL REDIFFERENTIATES
Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Figure 31 Becker's epochal bone fracture healing.

After the immune system "kills the bad guys" and saves the day, the battlefield is littered with residue and many cells of the body are damaged. The immune system sends in large scavenger cells that clean up the residue. And that's it for the immune system. Comparing it to an army's operation in the field, the immune system is the combat troops, not the medical services.

Restoration (healing) of the damaged cells back to normal is accomplished by the *cellular regeneration* system (Figure 29). This system is little known and little studied, but mostly studied by Becker (Figures 27, 28). It is a system that functions electrodynamically, but uses time-domain pumping to time-reverse the cells and their every part (including their genetics) back to an earlier state. In biology, the effect is called *dedifferentiation*, but without any knowledge of the actual EM mechanism.

So, within its capabilities, the cellular regenerative system in the body already uses longitudinally polarized EM waves, time-polarized EM waves (TPWs), and optical pumping by LWs and TPWs. Here we point out that neither the longitudinal photon nor the scalar (time-polarized) photon is individually observable; however, the combination of the two is observed as the instantaneous scalar potential {66}. With a little correction to properly make the phase conjugate replica wave a time-polarized EM wave, Whittaker's fundamental bidirectional EM longitudinal wave decomposition of the scalar potential (Figure 32) is consistent. Immediately the direct involvement of both the scalar photon and the longitudinal photon—and both the scalar EM wave and the longitudinal EM wave—are seen in Becker's bone-healing work using a scalar potential (Figure 31).

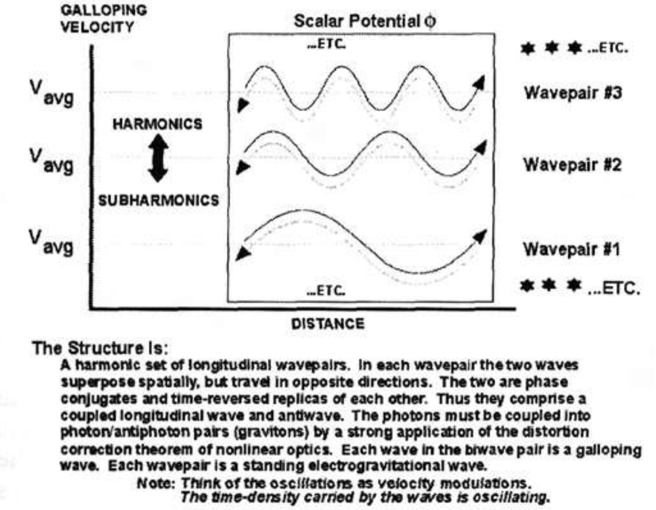

The Structure Is:
A harmonic set of longitudinal wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave.
Note: *Think of the oscillations as velocity modulations. The time-density carried by the waves is oscillating.*

Figure 32 Infolded longitudinal EM biwaves composition of a scalar potential, per Whittaker 1903.

Priore unwittingly stumbled onto the actual mechanism used by all biological systems in "healing" (cellular regeneration), and also recognized how to amplify it **(Figure** 12). But no one understood its exact nature, including Priore and the scientists who worked with him at the time, and the conventional community still does not understand it. The Russian scientists not only *understood* it early on, but *weaponized* it as well. So did the Chinese.

Need to Correct and Extend Classical Electrodynamics

The organized scientific community's protracted defense of a seriously flawed classical electrodynamics is long standing. The conventional, comfortable ways of "doing business as usual, leisurely" now threatens the survival of the nation itself. We need to quickly admit and correct long-standing EM errors—such as calculating a reaction cross-section of an entity and calling it the magnitude of the entity itself! Our electrical texts continue to make that fundamental mistake, and they have done it now for 100 years or more. We need to quickly change the classical EM model to include the active vacuum and its interaction, and we need to change all textbooks to clearly state that all EM energy in the universe comes from the active vacuum, via the asymmetry of the source charges in their fierce energy exchange with it.

We also need to point out that all primary electromagnetic entities—the charge, the field, the potential, and every joule of EM energy in the universe—totally violates the second law of thermodynamics. The charge produces continuously increasing negentropy, the field and the potential are formed at expanding light speed across the universe and are perfectly ordered (thereby totally violating the statistical mechanics basis of the second law), and every joule of EM energy is either field energy or potential energy, and hence violates the same statistical basis of the second law.

The Situation Is Serious

Admittedly these are strong statements highly critical of our scientific establishment's track record. But our national survival continues to be increasingly at stake, the situation is serious, and we must not continue to bury our collective heads in the sand like ostriches. We simply cannot continue to idly sit by and pontificate while other groups and nations methodically maneuver technically to cut our collective throats with advanced weapons we cannot defend against, once they find a way to "get by" the little nation that has been countering them for some years. Our own beloved nation may have less than 10 years {67} and perhaps less than three years to survive, unless the organized scientific community is galvanized to focus its very best scientific efforts in the energetics area.

So far, in truly innovative approaches to new methods for electrical energy, the U.S. scientific community is about as energetic as a sand terrapin sleeping in the tropical sun. There is no discernible move to recognize the seriousness of the problem, or to do anything effective about it. In 45 years, the asymmetry of the dipole (and the charge in its modern view as a dipolarity of special kind) has not made it across the university campus from the particle physics department to the electrical engineering department. In fact, the implication of broken symmetry to

electrical engineering and classical electrodynamics seems to still be completely unnoticed by the National Academy of Sciences, the National Science Foundation, our universities, and our great national laboratories. In the scientific community, other than work on vaccines etc., it is largely "business as usual."

A New Manhattan Project Is Warranted

Our need in the energetics weapons area is extreme, as are our needs in the medical area for quick, cheap, easy, and effective portable treatment of the coming mass casualties and our coming dramatic needs in energy. We *do* have superb scientists with the horsepower necessary to do something about it. The astounding, independent work of Sachs, the AIAS, and a few other scientists—which we stress has been unfunded—already clearly proves that. But unless we change the organized scientific community's normal rather snail-like pace—by a strong measure such as a Presidential Decision Directive and a declaration of a Nalional Emergency—the U.S. scientific community will overshoot the point of no return now approaching.

We need a new Manhattan-type project. We need it very quickly. If we do not change and focus our present scientific procedure, then—all too shortly—little that the scientific and intelligence communities are doing will be of value. We shall all be very, very dead or else in a new Dark Ages where most of the world economy has collapsed and freedom has largely disappeared from the face of the Earth.

T. E. Bearden, Ph.D.
Lieutenant Colonel, U.S. Army (Retired)
President & CEO, CTEC, Inc.
Director, Assoc. of Distinguished American Scientists

2002 ENERGETICS WEAPONS THREAT TO THE U.S.

T. E. Bearden 2000
Updated Oct. 1,2002

The U.S. Narrowly Escaped Energetics Destruction Twice in 1997

On two occasions in 1997 (latter February and on May Day), we were scheduled In be strategically attacked without warning and completely destroyed. Eerily, in public statements the leader of the Aum Shinrikyo sect (Figure 33) had foreshadowed the dates for the planned attacks, since his group was also involved—in fact, shooting down the TWA-800 (Figure 34) as advanced practice and stimulation of the U.S. to see if it knew what was afoot. Those two planned strategic strikes in 1997 were avoided only in the nick of time by actions of a friendly little foreign country, which was the only nation on earth at the time that could counter the KGB quantum potential (QP) weapons. I am sure that proper U.S. agencies intercepted my urgent personal faxes to the Defense Attache in Washington, DC and directly to the Prime Minister of the nation involved. Actions by that nation did succeed in averting our destruction in the nick of time.

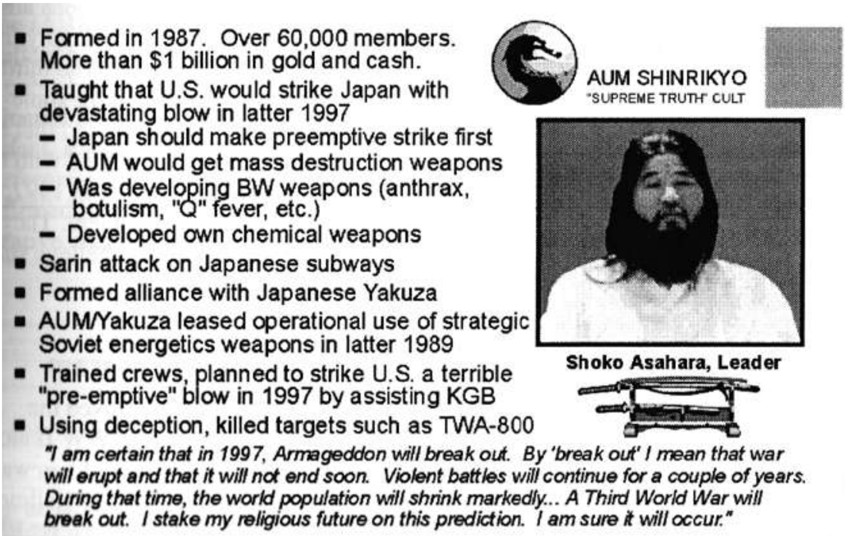

- Formed in 1987. Over 60,000 members. More than $1 billion in gold and cash.
- Taught that U.S. would strike Japan with devastating blow in latter 1997
  - Japan should make preemptive strike first
  - AUM would get mass destruction weapons
  - Was developing BW weapons (anthrax, botulism, "Q" fever, etc.)
  - Developed own chemical weapons
- Sarin attack on Japanese subways
- Formed alliance with Japanese Yakuza
- AUM/Yakuza leased operational use of strategic Soviet energetics weapons in latter 1989
- Trained crews, planned to strike U.S. a terrible "pre-emptive" blow in 1997 by assisting KGB
- Using deception, killed targets such as TWA-800

AUM SHINRIKYO
"SUPREME TRUTH" CULT

Shoko Asahara, Leader

"I am certain that in 1997, Armageddon will break out. By 'break out' I mean that war will erupt and that it will not end soon. Violent battles will continue for a couple of years. During that time, the world population will shrink markedly... A Third World War will break out. I stake my religious future on this prediction. I am sure it will occur."

Figure 33 Planned KGB/Yakuza/Aum Shinrikyo attack in 1997 was foreshadowed in public statements made by Shoko Asahara.

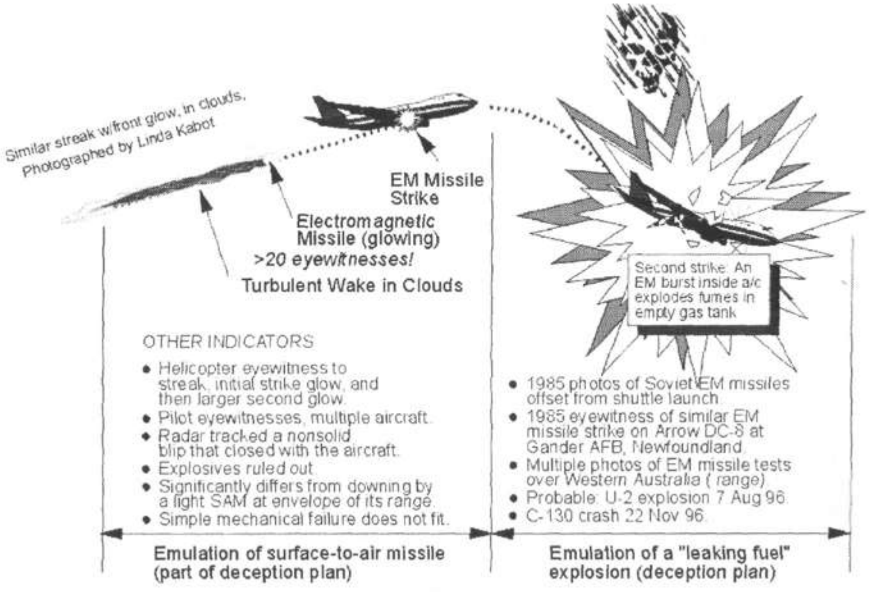
Similar streak w/front glow, in clouds, Photographed by Linda Kabot
EM Missile Strike
Electromagnetic Missile (glowing)
>20 eyewitnesses!
Turbulent Wake in Clouds
Second strike: An EM burst inside a/c explodes fumes in empty gas tank
OTHER INDICATORS
• Helicopter eyewitness to streak, initial strike glow, and then larger second glow.
• Pilot eyewitnesses, multiple aircraft.
• Radar tracked a nonsolid blip that closed with the aircraft.
• Explosives ruled out.
• Significantly differs from downing by a fight SAM at envelope of its range.
• Simple mechanical failure does not fit.
• 1985 photos of Soviet EM missiles offset from shuttle launch
• 1985 eyewitness of similar EM missile strike on Arrow DC-8 at Gander AFB, Newfoundland
• Multiple photos of EM missile tests over Western Australia (range)
• Probable: U-2 explosion 7 Aug 96
• C-130 crash 22 Nov 96
Emulation of surface-to-air missile (part of deception plan)
Emulation of a "leaking fuel" explosion (deception plan)

Figure 34 Kill of TWA Flight 800 on 17 July 1996 by Yakuza and Aum Shinrikyo.

Only five nations or national groups have quantum potential (QP) weapons, though one or more additional nations are almost certainly working on them and may even have them. First of the five is the KGB in Russia (the weapons are not in the regular Russian forces, and never have been). Second is the friendly little foreign country to whom we presently owe our continued survival. The third is Brazil (which has had a small and successful QP weapon program for some time). China deployed QP weapons in early 2002, and I recently found that at least one other friendly nation has them as well. I'm not too impressed with that other friendly nation, since they played no role in saving our bacon in 1997. It appears they did not actually know the coming attacks were imminent. They certainly do not have Causal System Robot weapons, nor do they believe such weapons even exist.

Initiation of the Brazilian Energetics Weapon Program

Brazil implemented an energetics weapons program after it employed the repatriated Dr. Hellman, a German scientist taken to Russia after WW II along with the German radar team. Years after being taken to Russia, Hellman was repatriated and made his way back to Germany. In a news interview Hellman reported that he had worked directly on the development of Soviet secret EM weapons which could suddenly freeze large areas of the ocean surface. That of course is a scalar interferometry weapon, producing negative (converging) EM energy rather than positive (diverging) EM energy in their distant interference zones.

Also, when the U.S. made David Bohm *persona non grata* in the McCarthy witch-hunts, Bohm was later employed in lecturing and teaching in Brazil for some time. At the time our own scientists continued to pontificate as to whether they *liked* or could *tolerate* Bohm's quantum potential and his hidden variable interpretation of quantum mechanics {3}. Meanwhile, the KGB weapon scientists and the Brazilian Hellman project seized upon Bohm's hidden variable theory with alacrity as soon as he published his seminal paper. Those two weapon groups were already familiar with the internal longitudinal EM wave structuring of potentials, fields, and waves (Whittaker 1903), and immediately saw the fantastic weapons implications of a deliberately structured quantum potential, containing a specific "engine" as desired. Most of our scientists still are unaware of it to this day, and are unaware of its formidable weapon implications when combined with Whittaker's 1904 superpotential paper, his 1903 potential decomposition, and quantum field theory. Most of our scientists still adhere to the Bohr interpretation of QM, which excludes the normal, everyday, observable, ordered universe we all inhabit and observe.[15]

The Bohm quantum potential is, after all, a scalar potential, and at least in theory it has a Whittaker biwave decomposition {68}. Even in the Coulomb gauge, the scalar potential has infinite velocity, as pointed out by Jackson.[16]

Several tests of the "cold explosion" Russian weapons referred to by Hellman have since been openly made. One major test seen by at least five jetliner crews was the anomalous rapidly rising cloud and subsequent extremely huge light globe (ABM energy shield) effects off the Kurils in April 1984 (Figure 35). Two excellent scientific articles on the physical phenomena observed in that incident are published in the prestigious journal *Science* {69}. We had explained the cold explosion phenomena and the ABM energy shield effects some years before that incident. Even ordinary superpotential theory {70} and longitudinal EM wave interferometry at a distance can substantiate the capability. We also positively confirmed by private sources that the scalar interferometry, the MindSnapper (negative energy EMP), and the cold explosion weaponry are real.

---

[15] This exclusion is euphemistically known as the "missing chaos" problem in quantum mechanics, and it is resolved in Bohm's hidden variable approach.

[16] J.D Jackson, *Classical Electrodynamics,* 2nd Edn., John Wiley & Sons, New York, 1975, p. 222, 223. Usually the instantaneous scalar potential contributes only to the near field.

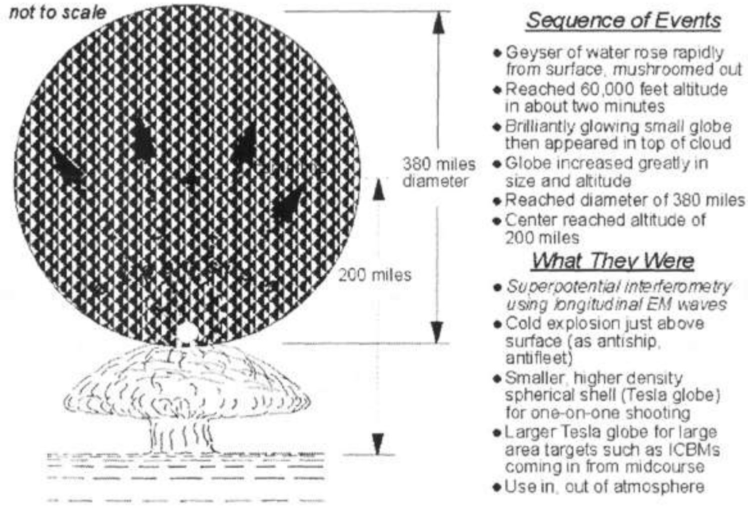
not to scale
380 miles diameter
200 miles
Sequence of Events
Geyser of water rose rapidly from surface, mushroomed out
Reached 60,000 feet altitude in about two minutes
Brilliantly glowing small globe then appeared in top of cloud
Globe increased greatly in size and altitude
Reached diameter of 380 miles
Center reached altitude of 200 miles
What They Were
Superpotential interferometry using longitudinal EM waves
Cold explosion just above surface (as antiship, antifleet)
Smaller, higher density spherical shell (Tesla globe) for one-on-one shooting
Larger Tesla globe for large area targets such as ICBMs coming in from midcourse
Use in, out of atmosphere

Figure 35 Exothermic (cold) explosion off the Kurils 9 April 1984. A second test, an expanding globe of light, was added in the exothermic mode to demonstrate rapid serial switching between weapon modes.

Related Phenomena

In 1969 in Toronto, Canada, Sid Hurwich demonstrated a device capable of "freezing" a pistol lying on a table so that the gun could physically not be moved and the trigger could not be pulled {71}. For this exhibition of *strong electrogravitation,* general relativistic fields initiated by the EM force field (including the usually unaccounted Heaviside nondiverged component that Lorentz arbitrarily discarded) as the agent of spacetime curvature were used {72}. The strength of the effect was attested by the fact that the hands of the observers' watches all showed "no passage of time" during the half hour they were in the room with Hurwich's device operating.[17]

Rapid development of that strong electrogravitational technology and concomitant energetics then proceeded in another nation. Weapons of such nature were then employed by that nation a few years later, in two well-known international incidents. Clear signatures of the nature of the weaponry utilized were present in those incidents. Details of these incidents will not be included in this paper.

---

[17] We point out that this is highly interesting itself, since it showed that the observers were operating in a different frame from that of the watches and the gun, which were "frozen". We also point out that recently light itself has been "stopped" in quantum experiments, yielding something quite similar.

Since the energetics weaponry already utilized special EM "hidden variables," Bohm's hidden variable theory and quantum potential were directly adapted into the program to dramatically extend it into the development of quantum potential weapons.

Stalin *Personally* Initiated the Soviet Energetics Program with the Greatest Urgency

While Stalin was at the Potsdam conference, at the close of WWII, the U.S. exploded the first atomic bomb in the desert. Truman informed Stalin in general terms that a powerful new U.S. weapon had been successfully tested. Secretly, Stalin already knew of our atomic bomb progress, for he had a spy in the program. While he was still at Potsdam, he then secretly called the Director of the Soviet atomic weapons program and directed a speed-up.

After we exploded two atomic bombs on Japan, the Japanese surrendered and WWII was over. See again Figure 1. Stalin vehemently ordered his Academy of Sciences to produce the next great technical breakthrough at all possible speed, since the American atomic bomb had frustrated his plan to wait two years after the war while the West speedily disarmed, then march into Europe and take it in six weeks. Now his massed troops could just be bombed back into the Stone Age by U.S. bombers using atomic bombs.

A Great Technical Breakthrough in Physics Can Always be Had in Four Years

The scientific community can have a great new technical breakthrough whenever its furious and dogmatic resistance to innovative discoveries or the possibility of innovative developments can be suspended. All that is necessary is a concentrated group of highly skilled, open-minded scientists, the money and facillities, and a strong headlock placed on the powerful but dogmatic scientific groups who would otherwise hamstring the program and destroy it.[18]

---

[18] Present day examples are (i) the cold fusion fiasco, where the U.S. scientific community has revealed furious determination to ignore more than 600 successful experiments worldwide, (ii) a similar fury against any and all research into extracting EM energy from the vacuum, even though all EM energy comes directly from the active vacuum via the proven asymmetry of source charges and dipoles, (iii) violent suppression of proven "extended electrodynamics" methods of rapidly curing severe diseases, even in terminal stage (resistance to anything except "cut, drug, burn" and anything that would interfere with the monstrous profits of the large pharmaceuticals), (iv) continuing dedication to an electrical engineering model that does not even model or apply what powers an electrical circuit, (v) continuing resistance to quantum potential research and development, (vi) slavish dedication to "big nuclear science", massive accelerators, dangerous nuclear power plants, and a "hot fusion WPA program", even though much higher energy physics than these programs advocate and use can be and is being done on the tabletop, using time-energy instead of spatial energy, and (vii) increasing "lock-up" of all research funds into the organized hierarchy's politically approved agenda. The community has deviated so far from scientific

We proved it in the WWII Manhattan Project, where we went from nearly zilch to the operational atomic bomb in 4 years. But we did it only by secretly bypassing the conventional U.S. scientific establishment. Einstein wrote to President Roosevelt *around* the scientific community, and when the President approved and ordered the highly classified program, conventional scientists not on the project were just parked on the sideline and ignored so that the job could be completed. The naysayers were also muzzled because of the extreme classification of the project; they faced jail if they engaged in their usual ad hominem attacks and dogmatic eloquence.

Further, today the "Skunk Works" approach repeatedly proves that bold new projects can be done, when the skilled scientists and facilities are mobilized and no adamant opposition by the conventional community is tolerated.

Numerous examples of the U.S. scientific community's suppression of innovation and new technology can easily be furnished. Such are well known to any historian of science. Cold fusion, the amorphous semiconductor by Ovshinsky, the first laser papers, ultrawideband radar, Mayer's original conservation of energy law, Boltzmann's work, Wegener's continental drift, flight of heavier-than-air machines, the rocket as a great strategic weapon, and hundreds of other examples come readily to mind. Such suppression has been *the* major characteristic of the organized scientific community since its inception, as pointed out by numerous leading scientists and authors. As Planck stated, usually one gets a new physics only after the old physicists who so adamantly oppose it just die off and get out of the way.

Today, the U.S. scientific community is in fact our "front line troops," although—except for a little lip service—the managers of the organized scientific community are blissfully unaware of it and have little or no sense of mission in out-of-the-box areas with respect to national survival. It is ironical, e.g., that our orthodox scientific community does not even know what powers an EM circuit—energy extracted directly from the vacuum by the asymmetry of the source charge (considered in its modern view as a dipolarity) and by the source

method in defending archaic theory in the face of experiments that falsify it, that the ubiquitous violation of the law of increasing entropy (second law of thermodynamics) by electrodynamics itself continues to be adamantly resisted using ad hominem and inflammatory attacks, slander and libel, and destruction of scientific careers of researchers. We flatly are not getting our taxpayer money's worth from our present scientific community, and—barring a Stalin-like shakeup from top to bottom—we are not going to get our money's worth in the future unless the community is forced to free up funds for young graduate students and postdoctoral scientists wishing to work in "out of the box" scientific areas. As Max Planck stated, *"An important scientific innovation rarely makes its way by gradually winning over and converting its opponents: it rarely happens that Saul becomes Paul. What does happen is that its opponents gradually die out, and that the growing generation is familiarized with the ideas from the beginning."* In other words, one doesn't get a scientific innovation until the old dogs so bitterly opposing it just die off and get out of the way. This long-known and stated characteristic of the scientific community—its fierce and unrelenting resistance to change and innovation—is presently the greatest threat to the survival of this nation.

dipole (Figure 36).

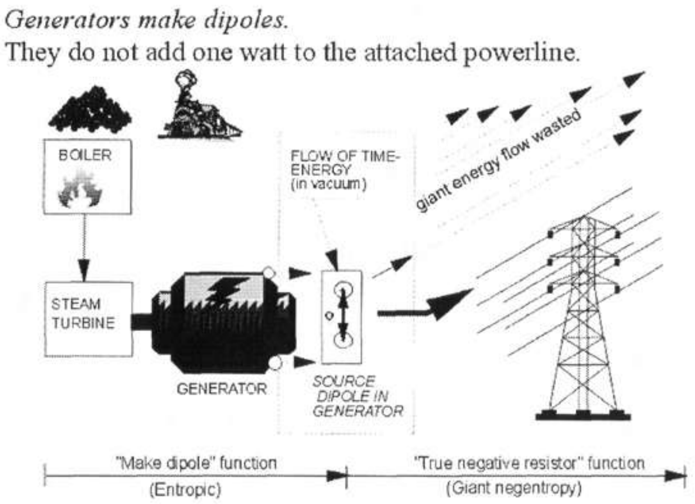
Generators make dipoles.
They do not add one watt to the attached powerline.
BOILER
STEAM TURBINE
GENERATOR
FLOW OF TIME-ENERGY (in vacuum)
SOURCE DIPOLE IN GENERATOR
giant energy flow wasted
"Make dipole" function
(Entropic)
"True negative resistor" function
(Giant negentropy)

Figure 36 Energy extracted from the vacuum by the asymmetry of the source dipole powers every electric power line and electric circuit.

A Typical Example: The Suppression of Cold Fusion

As an example, much of U.S. Science—dominated by nuclear factions since the advent of the atomic bomb in WWII—vehemently attacked cold fusion and the cold fusion researchers, including wholesale ad hominem attacks, with a viciousness seldom seen in the history of science. Many dogmatic scientists declared the entire field at best a mistake and at worst a fraud. Even the character of many of the leading cold fusion scientists was openly impugned and the researchers were slandered and libeled. All this was done by conventional dogmatic scientists who have not solved the fundamental source charge problem {10,12}, have no inkling that all EM energy in 3-space comes from the time-domain {10}, and have not a clue as to what really powers every EM circuit. Since they also do not know where the electrical energy driving the chemistry actually comes from, they have no knowledge of the effects of time-energy in electrolyte solutions.

They also appear to have little or no knowledge of the implications of Evans' and Searles' *transient fluctuation theorem*[19] in thermodynamics showing to what degree the second law of thermodynamics can be violated and reactions can run in reverse, nor its extension to the micron (colloidal) size level and to two seconds in duration.[20] If they did, then they would realize that the Coulomb barrier provided by *repulsion* of like charges (such as two H+ ions), often reverses into the *attraction* of like charges, with the barrier changing into the Coulomb attractor. This can last (by experimental proof shown by Evans *et al.)* up to two seconds, in a volume of fluid up to a micron in diameter. Hence it is not "against the laws of physics and thermodynamics" to observe attraction of two protons in an electrolyte so closely that each enters the strong force region of the other and the two form a quasi-nucleus. In the decay of this reversal zone, the preferred decay of the quasi-nucleus can be by quark flipping in one of the protons, yielding an atom of deuterium emerging from the decayed time reversal zone. Such results have been meticulously observed in many cold fusion experiments in multiple labs and nations, and by multiple researchers. Since these and other novel nuclear transmutation results are *replicated* experiments (more than 600, in multiple labs, multiple nations, and by multiple researchers), it means that no amount of theory can refute them or the results. The harsh skeptics launching the frenzied ad hominem attacks are thus revealed as the unscientific bigots they are, since they ignore substantial experimental results that contradict the prevailing theory. The excess energy production and inexplicable new nuclides (Figures 21, 37, 38) obtained by these multiple experimenters are completely inexplicable by ordinary nuclear interactions in the accepted theory.

---

[19] D.J. Evans and D. J. Searles, *Phys. Rev. E,* Vol. *50,* 1994. p. 1645.

- Fennions time-reverse in even numbers
- Like charges cluster
- Aqueous solutions contain H+ ions,
  which are just protons
- Quarks flip as TR zone fades

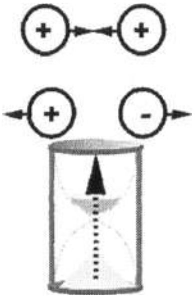

$$({}_1\mathrm{H}_1 + {}_1\mathrm{H}_1) \rightarrow ({}_0\mathrm{n}_1 + {}_1\mathrm{H}_1) = {}_1\mathrm{H}_2 \text{ deuterium}$$

Figure 37 Electronuclear reaction in a time-reversed zone producing excess deuterium ions.

---

[20] G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.,* 89(5), 29 July 2002, 050601.

- As TR zone decays, quasi-nucleus is usually in an excited state
  - — Decay action starts from every point in spacetime inside the nucleons
  - — Moves from inside toward outside
  - — Strikes the nearly free quarks first
- Decay occurs by easy quark flip
- Quasi-nucleus transmutes to an isobar
- If quasi-nucleus is a stable state in a forward time zone, no decay occurs.

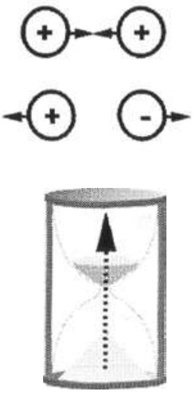

$${}_1H_2 + {}_1H_2 \rightarrow {}_2He_4$$

Figure 38 Electronuclear reaction in a time-reversed zone producing excess alpha particles.

Meanwhile, the U.S. Patent Office has been instructed not to accept any patent applications that would interfere with present nuclear planning. Positive cold fusion results have also been successfully obtained in several U.S. government laboratories, including at China Lake where very strange instrumental effects have the experimental team puzzled. I proposed an explanation of those anomalies in August 1998, and have included it in my book just published in September 2002 {73}.

Revolutionary New Processes Are Hidden in the Cold Fusion Experimental Results

Once understood, cold fusion holds the breathtaking key to systems that consume *a tiny fraction of time flow itself as* a fuel to produce copious heat energy for the world's needs. Unknown to the conventional scientists, of course, electrodynamics itself already does this, but it has just not been recognized and accepted. Transducing one microsecond per second of time into transverse wave EM energy will produce some 9×10^{10} watts of real electrical power, if all the transduced EM energy is intercepted and used. This is the *real* way to extract energy from the vacuum and use it cleanly and cheaply.

To grasp such a notion, recall that the fundamental units used in physics are completely arbitrary. It is perfectly possible to express all physics in only a **single** fundamental unit—for example, *the joule*. Time then becomes totally a function of energy. So it is not so insane after all to express time as highly compressed energy. Mass is also highly compressed energy. That concept is now familiar after Einstein, the nuclear age, and $E = mc^2$. Time-as-energy is just as straightforward as mass-as-energy, but it has just not yet come into recognition.

The *use* of time as compressed energy, however, leads to some profoundly new types of electronuclear reactions presently unknown in particle physics. These new transduction reactions produce the new nuclides in cold fusion experiments, and generate the electronuclear interaction. We explain that in some detail in our new book, *Energy from the Vacuum,* previously cited {10}.

When interpreted as using time-energy, cold fusion shows us the way to a new electronuclear regime of nuclear reactions where nuclei, elements, and compounds can just be directly assembled by comparatively weak (in terms of *spatial* energy!) electromagnetic interactions. Presently we have made substantial progress in understanding the fundamental mechanisms generating these electronuclear reactions, after presenting preliminary results in August 1998.

As one example, time-reversing a single fermion (a particle with spin 1/2, 3/2, etc.) is precluded by the Pauli exclusion principle. But one can time reverse *two* of them, or any even number n, where they act as a semi-boson or group of semi-bosons. Simply put, if a time-reversal zone is established momentarily in the region where like charges exist, then the n-tuple fermion cluster can be achieved because the law of attraction and repulsion of charges is temporarily reversed {74,75}. This unexpected "attraction of paired fermions of like sign" will form quasi-nuclei in many cases, particularly in small clusters of two or four, etc.

Hence two H+ hydrogen ions (two protons) in such a time-reversal zone in a fluid are drawn together into a quasi-nucleus, which actually is an excited isomer of deuterium (Figure 37). Time-reversal zones are temporary in cold fusion experiments due to the statistics of their formation and the transient fluctuation theorem {76} of thermodynamics. As the time reversal zone decays, the quasi-nucleus of two protons is increasingly in an excited and unbalanced state. The excited-state *diproton* nucleus decays by flipping a quark in one proton, converting it to a neutron. That produces a *neutron + proton nucleus,* which is a deuterium ion, strictly by electronuclear means. Or four protons may cluster by "attraction in a time-reversal zone" and form a 4-p nucleus. That is an excited isomer of a helium nucleus. As the time-reversal zone decays, first one quark flips in one proton and then another quark flips in a second proton. That produces a helium nucleus - i.e., an *alpha particle.* Or two deuterium ions can attract and form a quasi-nucleus (Figure 38), which is already an exact isomer of a helium nucleus and thus easily decays into an alpha particle by simply "tightening" as the weakened strong force recovers faster than the Coulomb force can reduce, vanish, and resume in the opposite direction. The exact reaction equations for these reactions are given in my new book, *Energy from the Vacuum* {10}.

We filed an Invention Disclosure on the process of forming and using time-reversal zones to cause like charges to attract and unlike charges to repel *in vacuo,* in fluids, and in other media, resulting in the formation of semi-nuclei as excited isomer states of other nuclei, followed by decay into new nuclides. We will follow it with patent applications in the future. This represents a completely new kind of particle physics interaction presently not included in the textbooks, and not requiring high-energy processes. Literally thousands of new nuclear interactions—all *permissibly* violating the second law of thermodynamics because the statistics permit such violations and experiments prove it—are now possible.

Cold fusion does indeed hold a new kind of nuclear physics in its hand—but a physics presently considered very alien by our orthodox nuclear science community. For decades nuclear scientists have been seeking ever more *spatially* energetic particle smashing, believing that extreme high energy nuclear reactions and transmutations cannot possibly be done with charged particles at low *spatial* energy. One is sympathetic to the scientists, because even the concept of "low spatial energy transmutation" itself sounds like an oxymoron. And it would be, *if one were limited to spatial energy alone*, as is presently assumed by physicists. But one is also speaking of the neglected and compressed *huge time-energy* component of those low-spatial energy photons, and that is quite a different matter. The lower the spatial energy of the photon (the lower its frequency), the greater the amount of compressed energy transported in the time component as "time". The transposition of time-polarized EM energy into longitudinally polarized EM energy—and vice versa—is already justified in quantum field theory. Use of that extremely high energy component by its transposition or partial transposition into spatial energy in fact provides a much higher energy physics interaction than present orthodox high energy physics utilizes, knows, or considers.

Almost all the learned scientists are thinking *only* of spatial energy, since that is the only kind they have been "taught" to consider. Hence, from their view (i.e., on the premise of their implicit *assumption)* it is perfectly "logical" to condemn cold fusion, since in their view, "cold" implies low (spatial) energy and "fusion" implies high (spatial) energy transmutation. What is unknown to the orthodox community is that the "cold" in cold fusion also implies "extraordinarily high time-energy". So with a little transduction of highly compressed time-energy into spatial energy, "cold" fusion actually can imply "extremely high spatial energy made available by transduction" also.

As an example, the Chief Editor of *Science* recently had the courage {77} to publish results of sonoluminescence research reporting positive experimental transmutation results {78}. A paper was then published in *Nature* {79}, pointing out that the (spatial) energy involved is insufficient for such nuclear transmutation temperature, thereby implying that the reactions reported by the first researchers are not possible.

However, given formation of a time-reversal zone, this second paper is not necessarily applicable to the sonoluminescence fusion process, because the paper is based on an unwitting two-part premise that *(i) only spatial energy can be used for nuclear transmutation and (ii) no time-reversal zones exist to nullify that premise by having reactions (such as the law of attraction and repulsion of charges) run backwards.* That premise is already experimentally proven to be invalid by Evans et al., and it fails when the time reversal zone effect is considered.

We strongly stress that thermodynamics—both in theory and as demonstrated in experiments—not only allows but predicts—for a certain fraction of the time—the formation and existence of just such "reversal zones" where the usual reactions *run in reverse* {76}. It is also experimentally proven {54}. In that case a much higher energy—time energy—can be and is employed to convert the Coulomb barrier into the Coulomb attractor.[21] We know now (from the experimental demonstrations by Denis Evans et al.) that such "reversal zones" may be up to a micron in size and may endure for up to two seconds. The subsequent decay of that TRZ then allows the quasi-nuclei temporarily formed in it to decay by quark flipping alone, allowing the quasi-nucleus to easily transmute into a perfectly well known nuclear reaction product.

To quote Arthur C. Clarke, when a learned scientist says something is *possible*, he is probably right. When he says something is *impossible*, he is probably wrong.

Stalin's Dictatorial Impact on Soviet Scientific Management's Response

Stalin was a powerful dictator. He simply killed or jailed those who opposed his will. So after his blunt order (Figure 1) circa latter 1945 to find a great new technical breakthrough area quickly, his Academicians and high level science managers were faced with a severe dilemma: They could not afford to *oppose or criticize a vigorous search* for new innovations and a great breakthrough, and they could not remain passive. Their very lives depended on they themselves *personally finding such a breakthrough in great haste and presenting it to Stalin*. They had to actively spur the efforts of their most outstanding subordinates and Soviet science itself, or suffer dire personal consequences. Siberia and prison camps loomed for the lot of them unless they produced.

To put it bluntly, Stalin "got their collective attention" and their "whole-hearted support." Those few who resisted simply were shipped post-haste to the notorious prisons of Siberia, and worked to death in short order.

---

[21] We strongly point out that, when the reactions run backward and negentropy appears in those reversed regions, the conservation of energy law is violated unless time-energy is being converted to spatial energy. Either we must begin accounting time as energy, or we must give up the conservation of energy law itself. The obvious solution is to account the time energy being utilized, as well as the 3-space energy.

So the highest-level Soviet scientific managers now desperately demanded of their subordinate scientists *anything*—*anything at all*—that was novel and had breakthrough implications. An almost convulsive national scientific effort was launched immediately, with hardly a detracting voice. A great team of the very best Soviet scientists was rapidly formed and methodically searched the scientific literature of the world from beginning to end. Nothing even remotely like that search and review has ever occurred in the West.

For example, one great technical center set up by the Soviets had a staff of about 2000 Ph.D.s with associated translators, staffs, etc. The Soviets purchased and hauled in copies of all the scientific literature of the West—by the shiploads—to such centers. Their scientists then methodically went through the material, journal by journal, paper by paper, page by page. There are and always have been myriads of innovations and discoveries sitting on the shelf in the scientific literature, never followed up, or opposed by standard models and credos. The first tier of reviewing Soviet scientists was composed of excellent scientists. They skimmed out the "cream" into a select pile which the second tier then analyzed in depth. The second tier was composed of the best nonlinear scientists in the world—the "superscientists."

They were seeking *new principles of nature that could rapidly be developed into highly advanced technology*, and they found them—sitting there in the literature, already pointed out, and long ignored by the Western scientific community.

Shortly Stalin had his great new breakthrough area emerging, and they had their heads.

Our Own Science Community's Progress During These Decades

For decades foundations physicists (Bunge, Feynman, Wheeler, Lindsey, Margenau, etc.) in the U.S. have repeatedly pointed out the substantial foundations problems in electrodynamics. Nobelist Feynman and the great John Wheeler tried to "fix" the problem whereby electrodynamics still erroneously assumes force fields in a space filled with a *material ether* (unit point positive charge, unit point north pole, and unit point mass assumed at every point in space). Unfortunately they failed, primarily because their absorber theory {80} contradicted the photon hypothesis and hence quantum mechanics.

Meanwhile, our own National Science Foundation and National Academy of Sciences has directed only minimal work on correcting electrodynamics and dramatically extending it. Faraday and Maxwell—as did every other scientist at the time—rigorously assumed a material ether. Maxwell's equations also implicitly included that material ether assumption. They still do, even after Heaviside's truncation and Lorentz's symmetrical regauging, because none of the Maxwell-Heaviside equations were changed to eliminate the material fields in space (material ether) assumption after the Michelson-Morley experiments falsified it.

Every major university in America still teaches that implicit material ether, even though it was falsified over a century ago. That is an exact measure of the caliber of "foundations" work our own scientific community does if left alone. *In more than 100 years, they have not even chosen to root out the material ether from electrodynamics.*

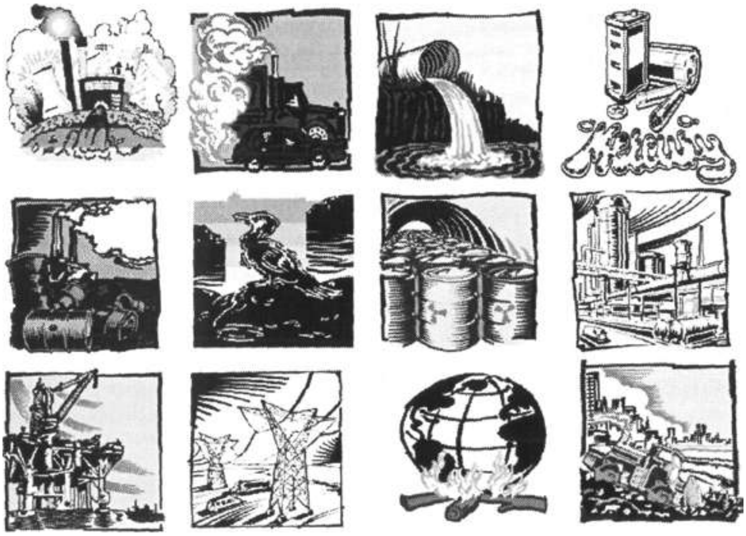

Figure 39 This polluted and suffering planet. Direct result of the scientific community's adamant resistance to upgrading classical electrodynamics and electrical engineering.

The prevailing flippant attitude is that, "Well, your TV set works, doesn't it?" Yes it does, but we also do not have practical electrogravitation and inertial propulsion, nor do we have cheap clean EM energy from the vacuum. Instead, we have a severely polluted planet (Figure 39), ever increasing deaths from biospheric fumes and polluting byproducts, increasing destruction of species, and a monstrous unleashing of far more coal and hydrocarbon burning in our near future while the world tears itself apart and destroys its economy and civilization in a rising series of oil and energy wars. The direct responsibility for that mess lies directly with the scientific community and its fierce resistance to innovative research and development, and changing a century-old, long outmoded classical EM and electrical engineering model.

Wryly put, it's high time we "got the lead out" of EM field theory, where the lead is the material ether and other non-sequiturs cluttering the theory.

Such drastic modeling errors would not be tolerated of government-funded aerospace engineers and contractors developing a new missile system. So it should be equally intolerable in the government-funded scientific community and in fact in the civilian scientific community as well, and it should not be tolerated of either the National Academy of Sciences or the National Science Foundation.

Einstein {81} said it perfectly, as follows:

> *"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted... in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."*

Upholding and perpetuating flawed electrodynamics models seriously impedes the progress of scientific development and revolutionary new technology.[22,23] At some point, one must legitimately question the upper management of the U.S. scientific community for not modifying classical EM and electrical engineering to include the active vacuum exchange with every charge and EM system. In not doing so, they continue to advocate perpetual motion source charges, creating energy from nothing, on a scale unparalleled in history.

It is little wonder that we do not have electrical power systems that successfully and freely utilize more than a trillionth of the vacuum energy they freely extract from the local vacuum. *There is not now and there never has been a single*

---

[22] Oddly, the conventional electrodynamics model assumes that every source charge freely creates energy out of nothing at all, and pours this energy out continuously in all directions, thereby establishing its associated EM fields and potentials at light speed, reaching out across the universe. From the 1960s alone, the total assumption of incredibly vast numbers of perpetual motion machines has been carefully hidden from the new students, so that today even many professors do not realize that in their model all EM fields and potentials, and every joule of EM energy in the universe, are assumed to be created by a vast assortment of perpetual motion machines called source charges. That problem has not been solved in electrical engineering or in classical electrodynamics, and if the scientists have their way, it is not going to be solved. Yet the basis for its solution—the asymmetry of opposite charges such as the dipolarity (in the modern view) of a source charge—has been in particle physics since 1957, with the award of the Nobel Prize to Lee and Yang in December of that year.

[22] One keeps one's sense of humor. The raucous skeptics and "defenders of the orthodox EM faith" who so noisily denounce legitimate COP>1.0 EM system researchers as "perpetual motion nuts", are themselves unwittingly the greatest perpetual motion nuts ever recorded in history. Further, they are so ignorant of what their present model actually means, that they do not even know it. It is one thing to be ignorant and to recognize it; it is quite another thing to be ignorant and so ignorant one does not know it—45 years after broken symmetry was rigorously established experimentally by Wu and her colleagues, and the Nobel Prize awarded to Lee and Yang.

electrical engineering department, professor, or textbook that knows what actually powers an EM circuit—real observable EM energy extracted from the local seething vacuum by the source charges and the source dipolarity.

Since our scientific community cannot seem to make such a simple change (as including the active vacuum's interaction with every charge and the charge's broken symmetry in that interaction) even 45 years after the basis for that change has been resoundingly proven in particle physics, our national survival is at stake and the nation itself is seriously threatened by the superweapons developed by the KGB over the last five decades. We would have already been totally destroyed—twice in 1997 alone—had it not been for the intervention of a friendly little foreign nation. My colleagues and I played a very small alerting role in those affairs, so we are speaking from experience, not from extrapolation. We would be quickly destroyed today, were it not for the continuing "balance of terror" provided by the little nation.

U.S. Scientists Have Committed Monstrous Crimes and Gone Unpunished

Rather sadly, we also point out another thing that has to be said: *scientists are no more moral or immoral than any other slice of American citizens.* A percentage of scientists is capable of committing horrible scientific crimes, as distasteful as such a statement may be. That includes the deliberate experimentation on humans, resulting in deaths. In short, at least manslaughter and perhaps even murder, since knowledge of the probable results was there beforehand.

One would think U.S. scientists would *never* scientifically condone and perform experiments leading to mass deaths for 30 years (i) as they did in the Tuskegee case in Alabama. Over 400 poor blacks were supposedly treated for syphilis, while actually not being treated at all, but were deliberately tricked and allowed to rot with syphilis so our "ethical" scientists could watch how their brains decayed, so to speak. Over 100 of those poor blacks died with the disease while the scientists cold-bloodedly watched them expire and kept meticulous records. Or (ii) as they did in the case of whole body nuclear radiation for 30 years—again by trickery, deceit, and illegal use of security classification to hide these acts—of helpless retarded children and others, some of whom also died.

Once these heinous incidents were exposed publicly, the political power of the scientific community is such that *no scientist was ever indicted or tried for these heinous mass murders, and none is going to be.* The scientific community *has* *not* fiercely called for punishment of these mass murderers amongst their midst—guilty of precisely the same kind of horrible actions for which we tried and executed war criminals at Nuremberg.

Incredibly, the President of the United States formally apologized for these incidents! He did not turn the case over to the Justice Department for prosecution of high crimes against humanity and mass murder, *as would have* *been done for ordinary citizens.* Had a single doctor in his private clinic done such things on his own, he would have been arrested, summarily tried, and either sentenced to life in prison or executed.

The sad thing is that there was really no sharp clamor from the scientific community to bring their fellow scientists to justice, a la Nuremberg. By their silence, the organized scientific community showed a side that scientists are loftily not supposed to possess. *So long as the scientific community continues to tolerate those mass murders of American citizens and not call for justice, then just so long do they stand convicted by their own resounding silence and inaction.*

We cite these painful incidents (there are others!) to point out that one must not deify the scientific community. The conduct of science is no better and no worse than the conduct of human affairs in other areas. Science is also a big business, and it has its moral scientists and its immoral scientists. It does its moral business and it does its immoral business.

We painfully submit that both sides of the scientific character must be considered, when one examines such things as the tolerance or intolerance of the scientific community to dramatic innovation and "out of the box" research. We have to consider it now, because it has become a serious threat to the continued survival of our nation. The conduct of scientific affairs will unavoidably include suppression, maneuvering, spin control, and fierce struggle for power and funds—such is in the very nature of human beings, and it is present in the scientific community on a bell-shaped curve as in all other human communities. It is ongoing widely today, and to an uncomfortably large extent much of science has been captured and effectively controlled by controlling the scientific funding and prescribing what research must be done for the funds available.

As the French say, "The more things change, the more they remain the same." Science is still the same way as it was when Planck made his famous statement. It will still be that way tomorrow, and a century from now. As with any other big enterprise, it has to have checks and balances imposed and enforced on it from outside—by the government itself. Little or no indication exists of that happening in the near future, until we get something like an anthrax spray attack on a large city, and have a million or more Americans dying in the streets.

Energetics: New Principles of Physics and How the Soviets Found Them

Under Stalin's whip, the Soviets particularly sought anything that indicated *possible new principles of nature.* It is no accident that Nobelist Petr Kapitsa later informed Nikita Khrushchev that a means of total neutralization of foreign missiles could only come from *a group of new principles in physics called energetics.* A group of new principles is precisely what energetics *is.* So if our scientific and intelligence communities wish to understand energetics, they must first understand the new principles it is based upon. They cannot just simply apply ordinary electrodynamics and "business as usual" as they have largely done. The massive Soviet scientific search program spurred by Stalin is where those new principles came from.

When the first tier of Soviet review scientists found something that had not been followed up, and saw that it had significant implications if found valid, that paper was laid aside into a "select" pile. That select pile then went to the "super" scientists, for full-blown review and analysis by some of the brightest and sharpest minds on the planet.

The Russians came up almost immediately with Whittaker's 1903 and 1904 work. That work shows a hidden electrodynamic *longitudinal* EM wavepair structure (in phase conjugate pairs) inside the scalar potential and comprising it.

This is a dramatic extension of electrodynamics to incorporate a full-infolded general relativity infolded inside the electrodynamics. Whittaker showed that scalar potential functions in interference already *create* all the "normal" fields and waves anyway (the beginning of superpotential theory). This novel interferometry and its inherent capability for action-at-a-distance did not elude the Russian superscientists. The superb Soviet scientific team came up with that and with much, much more.

Nothing like the comprehensive, intense review of the entire scientific literature has been accomplished by the U.S. scientific community. Nothing like it is being done now.

Longitudinal EM Wave Interferometry Energetics Was Almost Immediately Uncovered

At the end of WW II, the German radar team and infrared team were taken to Russia. Both teams were years ahead of their Western counterparts.

One of my close colleagues has validated that the German radar team had some severe instrumental anomalies while still in Germany. We think this is what they were:

In WW II the Germans invented radar absorbing materials, to coat their submarine snorkels and prevent Allied antisubmarine warfare bombers from tracking the snorkels and destroying the subs. A wide variety of highly nonlinear, doped materials were investigated. In addition to single beam radar cross section, the Germans also examined the materials in multi-beam illumination circumstances.

It appears that, in some multibeam experiments with some materials, conditions were just right so that the multi-illuminated absorbing material acted as a pumped phase conjugate mirror (PCM), and wave-to-wave interactions emerged. This meant that the absorber material PCM would suddenly gather up all the energy in the pumping beams, use one of the other beams as a signal input beam, and fire an amplified retroreflected pulse right back down the signal input beam, wrecking the experiment and creating havoc in surrounding equipment.

When the German radar team went to Russia, their knowledge of these anomalies went with them. In the desperate Russian scientific climate to find *lust such anomalies,* this startling and anomalous phenomenon would have been exhaustingly examined and studied, once the German radar engineers got to Russia to their new laboratories.

In short, it appears that the Russians quickly came up with what today is called "nonlinear phase conjugate optics," but first in the radar band rather than optics.[24] So before 1950, it appears that the Russians already had working phase conjugating pumped phase conjugate mirrors for the radar bands and in other bands.

Fisher {82} details that we Americans did not "stumble onto" practical phase conjugate optics experiments until Russian physicists briefed Lawrence Livermore National Laboratory scientists in 1972 about the strange wave in **many** of their optical experiments that "came out of nowhere and restored order **from** disorder." At that time, the KGB energetics weapon scientists had known aboutphase conjugate retroreflection and its amplified pumping, for more than 20 years. Indeed, the first giant strategic scalar interferometers had already been deployed and utilized to kill the U.S.S. Thresher nearly a decade earlier (Figure **6).**

By 1950 Russian Energetics Was Off and Running

By the beginning of the 1950s, the Russian scientists under iron KGB control were already experimenting with "scalar potential interferometry" at a distance, a la Whittaker 1904. Here transverse EM waves, EM fields, and EM energy—either *divergent* (heating) positive energy or *convergent* (cooling) negative enery—could be readily created at a distance {83} (Figures 40 and 41). Further, the longitudinal EM wavepairs used in this interferometry had the marvelous characteristic of easily penetrating right through the earth and ocean, with only a little intervening interaction, given that the LWs are sufficiently well made {84}. The interferometry "beams" can and do also move superluminally—again, right through the earth and ocean.

---

[24] **Of course, they would also have investigated it at all other bands as well.**

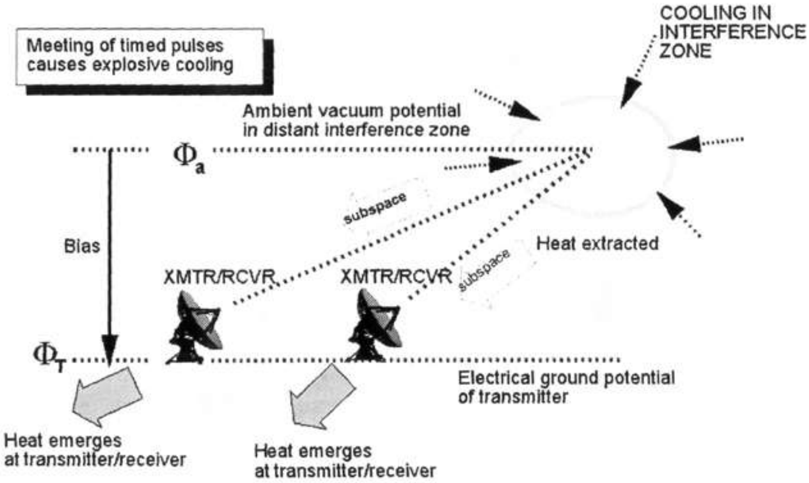
COOLING IN INTERFERENCE ZONE
Meeting of timed pulses causes explosive cooling
Ambient vacuum potential in distant interference zone
Φ_a
subspace
Bias
Heat extracted
XMTR/RCVR
XMTR/RCVR
subspace
Φ_T
Electrical ground potential of transmitter
Heat emerges at transmitter/receiver
Heat emerges at transmitter/receiver

Figure 40 Scalar interferometry in the endothermic (cooling) mode.

It is not accidental that Admiral Gorshkov, head of the Soviet Navy, would later bluntly state: *"We have made the oceans of the world transparent."* And so they had.

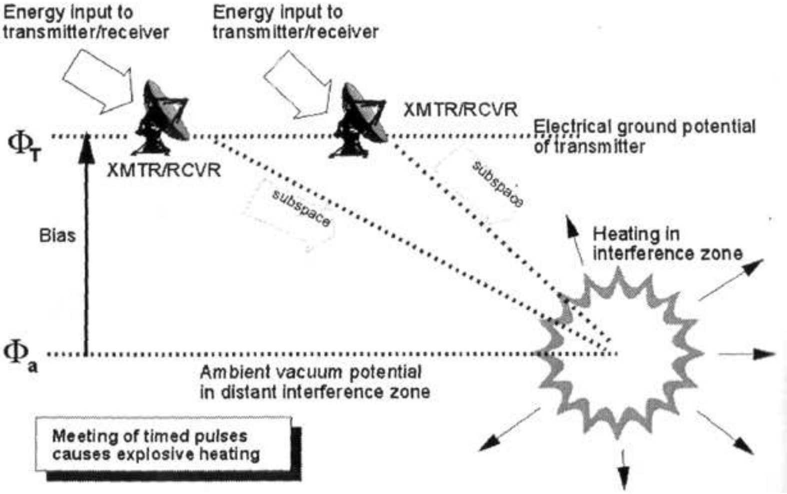
Energy input to transmitter/receiver
Energy input to transmitter/receiver
XMTR/RCVR
Φ_T
Electrical ground potential of transmitter
XMTR/RCVR
subspace
subspace
Bias
Heating in interference zone
Φ_a
Ambient vacuum potential in distant interference zone
Meeting of timed pulses causes explosive heating

Figure 41 Scalar interferometry in the exothermic (heating) mode.

By 1950, the Russian scientists were experimenting with deliberately creating and phase conjugating longitudinal EM waves, forming phase conjugate wavepairs. They then assembled the desired (mathematically calculated) sets of those waves, adjusted frequencies and magnitudes, phasing, etc., to produce EM potentials having a desired deterministic internal spacetime curvature set (internal spacetime engines) and dynamics.

As a result, by the mid-50s Russian weapon scientists were already experimenting in *bioenergetics* with the direct inductance of diseases and health changes via the transmission into human targets of such "internally structured" EM potentials, fields, and waves. They were also experimenting with the production of EM energy, potentials, fields, and waves at a distance, directly through the earth and the ocean. Either heating (energy flow divergence) (Figure 41) or cooling (energy flow reconvergence) (Figure 40) could be produced at will in the distant interference zone, merely by altering the bias voltage on the chassis grounds of the interferometer's transmitters. Rapid switching between the exothermic and endothermic modes thus was straightforward{85}, as demonstrated (Figure 35).

By the latter '50s, longitudinal wave interferometry weapons had already emerged from advanced engineering development and were in full production engineering toward deployment. The LIDA device {86} and even more advanced versions for directly affecting the human mind and its state of consciousness had emerged in *psychoenergetics* as one class of devices rather openly released and utilized. The LIDA uses structured EM microwaves {87} to induce a hypnogogic or cataleptic type of trance-like state in either a human or a cat. as tested years later by Dr. Ross Adey of the U.S. Early prototypes were purportedly used against our prisoners in North Korea during the Korean War to assist in their brainwashing by their North Korean Captors. Prisoners reported that its effects were irresistible.

Meanwhile, energetics weapon testing in several major modes was being conducted, particularly from Soviet research ships over remote regions of the ocean and at night. Descriptions of such tests—many of them seen by passing ship's captains—can readily be found in the open maritime literature. We gathered and published many such incidents and cited references.

Deployment of the Soviet Energetics Weapons and Their Quick Use

In January 1960 Khrushchev spoke obliquely of these "fantastic" new superweapons, just within the portfolio of the Soviet scientists, so to speak, in an important speech to the Presidium. Parts of that speech were reported in the *New York Times* {88}. No one in the West knew what Khrushchev was talking about. Our intelligence community, which bases many of its assessments on the tired old phrase "We have no evidence that..." simply treated it as propaganda, It wasn't, and the U.S.S. Thresher was to be destroyed (Figure 6) three years later by that same weapon Khrushchev was speaking of.

Khrushchev began rapidly inserting long-range missiles and nuclear warheads into Cuba in 1962, precipitating the Cuban Missile Crisis. Kennedy knew Khrushchev's strategic bombers and rockets were in horrible shape, due to direct information received from Colonel Oleg Penkovskiy, a Russian spy in our employ courtesy of British Intelligence. No one knew of the superweapons, but they were not yet operationally ready in Oct.-Nov. 1962. So Kennedy called Khrushchev's bluff and faced him off. Blustering awhile to obtain U.S. assurance that it would not invade Cuba, Khrushchev was forced to back down and withdraw the missiles. In his exuberance he had simply moved too fast in inserting the missiles into Cuba. Had he waited a few months later, the strategic superweapons would have been operational and ready. Any U.S. naval blockading force would have thus faced quick and eerie destruction or disabling, as was soon to be demonstrated in the case of the U.S.S. Thresher and its surface companion, the U.S.S. Skylark.

This severe loss of face by Khrushchev, before the Communist Party, made it imperative that he do something quite dramatic to stay in power, and do it as quickly as possible.

Khrushchev's first large, full-up, strategic longitudinal EM wave energetics interferometer weapon became operational on site by April 1, 1963. A few days later, Khrushchev employed it to kill the U.S.S. Thresher atomic submarine (Figure 6) underwater off the East Coast of the U.S. By simply creating EM noise in an interference zone placed on, in, and around the sub underwater, and thus jamming the sub's electrical controls so that it could not be controlled, the helpless sub sank to crush depth and imploded. The U.S.S. Skylark on the surface had multiple systems seriously jammed or rendered nonoperational. This was the Russian deception plan, to have the kill of the sub resemble an accident.

One day after the successful clandestine kill of the Thresher, the same weapon was used to demonstrate a powerful underwater electromagnetic burst 100 miles north of Puerto Rico (Figure 42). The energetics underwater burst equaled a substantial deep underwater nuclear explosion in size. The burst was seen by a passing U.S. jetliner crew and was reported to the FBI and the U.S. Coast Guard when the aircraft landed in Florida. A colleague interviewed the co-pilot of the aircraft, and relayed the interview information to me.

Khrushchev's Second Demonstration

Deep Underwater EM Explosion 11 Apr. 1963

(Similar to this photo set of a deep underwater nuclear **test** explosion)

Phenomena observed and reported by passing U S jetliner

100 miles north of Puerto Rico, over Puerto Rican Trench (One of the deepest parts of the ocean)

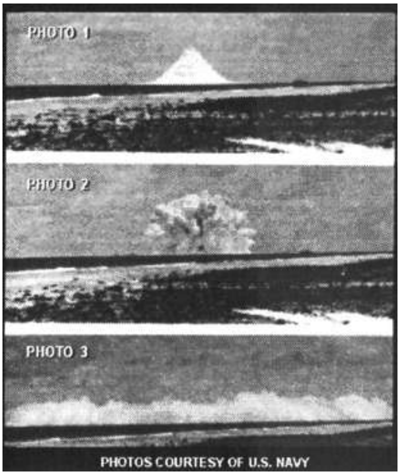

Figure 42 Deep underwater nuclear burst, demonstrating same water phenomenology produced by Khrushchev's second demonstration on 11 Apr. 1963, the day after killing the U.S.S. Thresher.

Why the Soviets Could Continue to Attack and Kill U.S. Targets with Impunity

Ironically, Khrushchev clearly demonstrated the inability of U.S. science to comprehend the numerous test incidents of Soviet LW interferometers, including the actual destruction of strategic U.S. targets (our latest and greatest nuclear attack submarine!) right off our coast. *The greatest strategic weakness of the U.S.—and the thing that may yet completely destroy all of us—is the continuing snail-like pace of U.S. scientific progress in revising, correcting, and extending its seriously flawed classical electrodynamics model. This scientific inertia placed us several decades behind in development of defenses in the areas of longitudinal EM wave weapons, quantum potential weapons, and causal system weapons. It countered our strategic nuclear armada for two decades. We were for some time analogous to the Spanish Navy with its wooden ships, when it faced and was destroyed by American steel ships. Hopefully at least some more modern work has now been accomplished in higher group symmetry electrodynamics, so that we are not quite as defenseless as we formerly were. But we still have a very long way to go.*

When one performs a standard strategic analysis on a nation, one of the areas examined in depth is called *national style.* One of the major characteristics in the U.S. **national** style is that we do not react to slowly increasing threats or changes

(Figure 43). That style particularly applies to the scientific community. Because of it, the Russians knew that our government only inches forward in achieving new *basic* science. We run at enormous speed, doing *applied* science. We are essentially Romans and great engineers. We are largely not Greeks and not great esthetics discoverers. We also will just ignore new threats if they emerge very slowly and are only gradually perceived, and we will hide behind that infamous phrase, "We have no evidence that..."

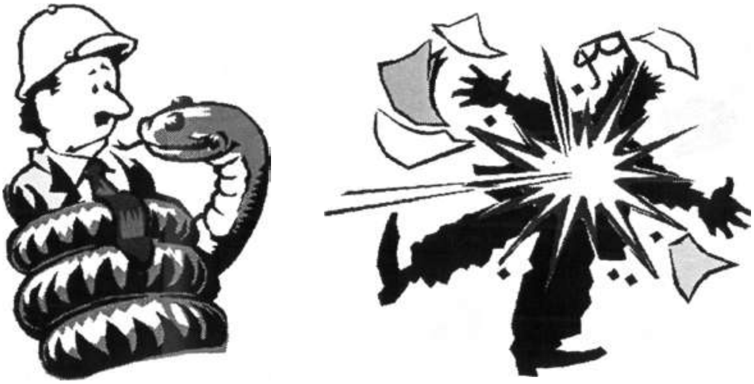

BOA CONSTRICTOR LIGHTNING STRIKE

Americans tend not to react to slowly increasing things, but react to rapidly increasing things.

Figure 43 America's national style: we do not react to slowly increasing threats.

In short, *we react to the coiled rattlesnake, not to the boa constrictor*. We are now in a long, bloody, costly war against terrorism because of that very characteristic of our national style. For decades, the slowly increasing infiltration of terrorists and their weapons was just ignored. Oh, yes, the official "threat studies" pointed it out, placidly.[25] But no one got very excited about it. The prevailing attitude was that "Ah, yes, but nothing like that will happen on *my* watch!"

As clearly demonstrated by their successful decades-long radiation of the U.S. Embassy in Moscow, the KGB chess masters knew that, if the test kills using their superweapons were not too often repeated, and if they even vaguely resembled other things known in the world, the U.S. scientific community and government investigators would resist all suggestions of Russian superweapons and super technology using new principles of nature. And so would our intelligence community, since they rely on the scientific community for their scientific assessments. They were well aware of our historical "ostrich" position (Figure44).

---

[25] Indeed, if one will look at the threat studies prior to William Cohen taking office as U.S. Secretary of Defense, one will find widespread complacency about the terrorist threat, and mostly just lip service to it. It was definitely not very important to most of the intelligence and scientific community. Cohen built a fire under the entire threat community, forcing them to get their act together and start actually studying the terrorist threat, asymmetrical warfare capabilities and dangers, and what could be done about it. He did force the research and threat communities to do a creditable job of studying it, and to publish reports detailing it. Unfortunately, the tenor of the political community outside SecDef Cohen's control was basically "business as usual, not to worry." September 11, 2001 changed all that forever. Once again we had waited until the rattlesnake struck and actually bit us, before the nation rose up and paid attention.

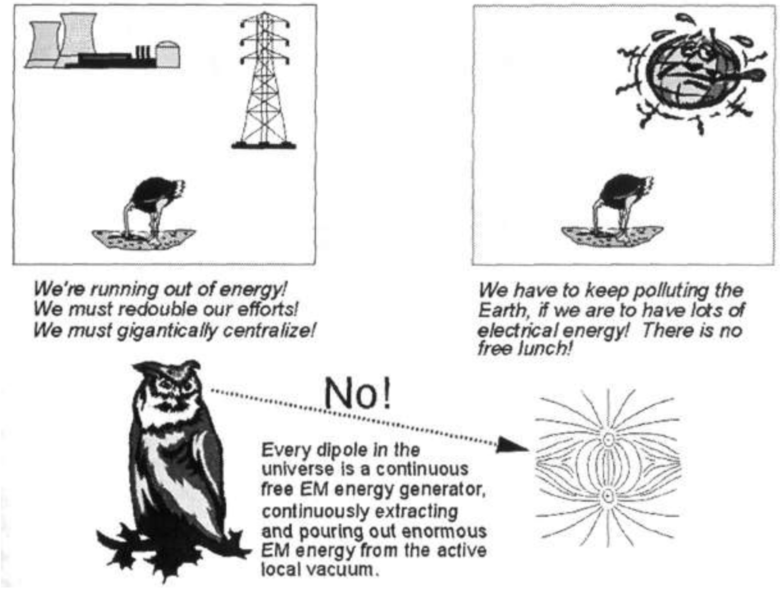
We're running out of energy!
We must redouble our efforts!
We must gigantically centralize!
We have to keep polluting the Earth, if we are to have lots of electrical energy! There is no free lunch!
No!
Every dipole in the universe is a continuous free EM energy generator, continuously extracting and pouring out enormous EM energy from the active local vacuum.

Figure 44 The only energy resource problem is the theoretical ostrich problem in science. We *must* become theoretical owls, rather than ostriches.

The KGB deception just depended upon the U.S. tiger to keep its stripes and exhibit its normal reactions or lack thereof. They also utilized (i) a series of strategic deception plans, deliberately designed to perform the tests suggestive of discredited phenomena such as UFO reporting and "lights at night" types of things, and (ii) a continuing series of subtle and anomalous stimuli (probes) designed to rather slowly attract the attention of U.S. Government at high levels, and of state-of-the-art U.S. government scientific contractors as well. The trick was not to attract U.S. attention *too suddenly or too strongly.* Show them the slow boa constrictor, not the lightning-fast rattler, and see how they reacted, if at all. By their reactions, the KGB could assess whether or not the U.S. governmcnt, intelligence, and scientific communities understood what was happening.

Precisely as anticipated by the KGB, we showed them for some decades that we really had little or no comprehension of the superweapons and their testing right over our heads. Even today, at least in the open community there is still little recognition, even of the violent weather war raging right over our own heads since July 4, 1976 and easily recognized.[26]

A case in point was the deliberate induction of diseases and health changes in U.S. personnel in the U.S. Embassy in Moscow for over 40 years (Figure 11). On a continuing basis, this excellent intelligence probe showed (by the U.S.'s actions or lack thereof) that we knew nothing about energetics.

Another KGB probe method was to kill a target once in a while, such as a jetliner or a submarine (deliberately rather seldom here, because the U.S. military would react if a certain frequency threshold were exceeded). Such kills resemble or can be posed as natural accidents or terrorist incidents. Quite a number of such incidents have occurred over the last three decades. The National Transportation Safety Board (NTSB) does not have higher group symmetry electrodynamics in its knowledge base or on its "checklists", nor does it have scalar interferometry, psychoenergetics, bioenergetics, or quantum potentials. In total, the results of deliberate stimuli of the U.S. continually assured the KGB chess masters (with 100% certainty) that we were a complete novice in the energetics game, and we were a lamb increasingly ready for the slaughter.

The major contribution to the continued success of these deliberate measures by the KGB has been provided for decades *in the "nonreaction of our own scientific community to slowly increasing threats."* To a lesser extent, our own intelligence community has also not focused sufficiently open-minded and highly qualified scientists to comprehend what is occurring in "new principles of nature" phenomena. "In the box" electrical engineers with Bachelor of Science degrees often are the analysts who analyze and assess such "out of the box" phenomena. Good luck!

Ordinary radar engineers may analyze an ordinary radar rather well. They will never, never become aware of a radar beam carrying infolded spacetime curvature engines. Such is not even in their lexicon or in their belief structure. They have no understanding or experience at all with longitudinal or time-polarized EM waves—most have never even *heard* of time-polarized EM waves. Mention curvatures of spacetime and their dynamics as deliberate "engines" impressed *inside* EM fields and potentials and waves, and one gets a glaze right over the eyes of the intelligence analyst recipient.

Western science conceives the mind as a meat computer, and does not understand that mind operations are totally electromagnetic and causal, but nonphysical—meaning in the time domain rather than in 3-space. Mind operations primarily use time-polarized EM waves rather than the transductions into transverse EM waves that the scientists actually measure. Hence Western science has not the foggiest notion as to what Russian psychoenergetics is all about, or how it can even "be." Certainly Western science has had no notion of deliberately creating time-polarized EM waves for direct engineering within the mind itself and within the mind's time-EM operations.

---

[26] The weather war over our heads has been particularly violent, and violently resisted, this year (2002).

As Arthur C. Clarke puts it in his "third law," *"Any sufficiently advanced technology is indistinguishable from magic. "*

Our long fascination with our "normal" (optics, IR, UV, radar, COMINT, satellite photography, etc.) technological intelligence collection systems will certainly not lead to any great breakthrough understanding of Russian energetics weaponry. That could only have come through HUMINT or COMINT. HUMINT is an area which, in the past, we have mostly just neglected or given lip service to for decades—and that came back to bite us severely on September 11, 2001. Even when we use HUMINT, it will not yield energetics results if the managers and directors themselves (and the agent handlers in the field) have no understanding of it and do not even know the proper questions to ask. And the system will not deliver results if the Intelligence Collection Requirements do not address the proper issues because ordinary electrical engineers and conventional electrical physicists having no inkling of psychoenergetics, or how it works, drew them up.

COMINT, on the other hand, suffers from its own extraordinary prolific output -as well as by only conventional analysts and conventional scientists examming it. Warehouses of COMINT information often have not even been read in any way except a casual glance by a junior analyst, and certainly have not had serious foundations scientists examine it in detail. ELINT does a little better insofar as a human analyst looking at it, but then what is a conventional electrical engineer or radar engineer to make of a giant hemispherical globe that slowly forms and expands from a tiny "seed bubble" of intense glowing light that appears (Figure 35), as it did off the Kurils in April 1984? What will he label it? "Light phenomena?" Where are the technical descriptors of the actual process forming the phenomenon? Where are the implications as to the strategic military use of such phenomena? The reader can see the point. For decades our analysts ignored prolific reports of such incidents by the hundreds. Why, everybody knew it was glowing swamp gas!

One cannot even write a proper "intelligence collection requirement" if one does not understand the basic phenomenology. So essentially no one in U.S. intelligence was writing the proper intelligence collection requirements.

The Science Behind the Soviet Energetics Weapons

The science behind these weapons is called energetics. (See again Figure 8.) Energetics consists of three branches, depending upon the nature of what is being targeted. The branch targeted against nonliving matter is just called the same name: *Energetics*. Targeted against living cells and biological material, the nervous system, etc., that branch of energetics is called *bioenergetics*. Against the mind and its functions (which are time-like and thus involve scalar photons and time-polarized EM waves), energetics is called *psychoenergetics* {89}. See Figure 45. Psychoenergetics includes both the conscious[27] and unconscious mind[28] of the individual—and even Jung's collective unconscious mind[29] of the entire human species {90}, and in fact even the collective unconscious mind in all species on earth, which corresponds to Lovelock's *Gaia* concept.

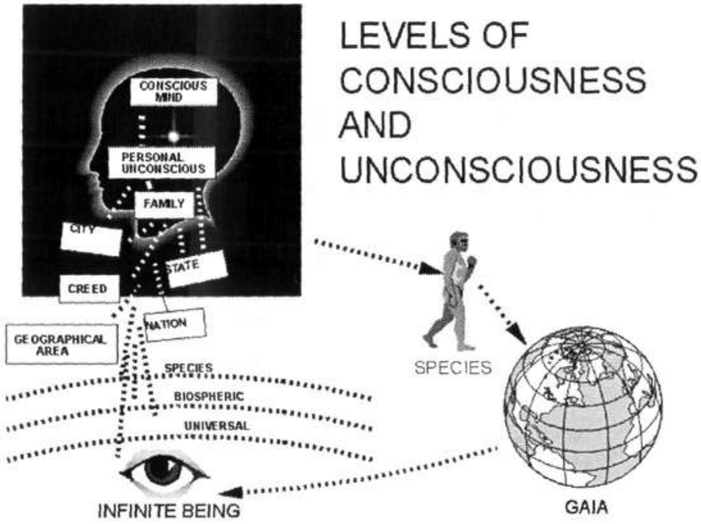
LEVELS OF CONSCIOUSNESS AND UNCONSCIOUSNESS
CONSCIOUS MIND
PERSONAL UNCONSCIOUS
FAMILY
CITY
STATE
CREED
NATION
GEOGRAPHICAL AREA
SPECIES
BIOSPHERIC
UNIVERSAL
SPECIES
INFINITE BEING
GAIA

Figure 45 Mind levels are subject to engineering by higher group symmetry electromagnetic (psychoenergetic) weapons.

We remind the reader that we are talking *physics*, not metaphysics and not mysticism. Strong Soviet physics developments over a sustained period have occurred in all three energetics areas, and these areas continue to be developed at present.

Once a nation enters into LW EM development, it has just entered into unified field theory development also. Particularly with longitudinal EM waves, one is directly engineering general relativity itself, and spacetime itself, as well as the dynamics of spacetime curvatures and sets of such curvatures. But it is an even greater unified field theory than Western physicists suspect. Not only does it unite the four forces of physics, but it also unites mind and matter and their dynamics as well.

---

[27] The conscious mind is a very fast serial processor, processing one snapshot at a time.

[28] The unconscious mind is a massively parallel processor, doing a great many things simultaneously. It is totally conscious, but multiply so. Since the serial processor conscious mind can only "see a single slide projector in the viewgraph at one time", when it "looks at the unconscious mind" it sees a large number of slides simultaneously, hence just sees "black" or "nothingness" (no single thing).

[29] Jung himself labeled it a *mind* (and hence a living entity). His modern followers shudder at such a concept, and very much downplay the "mind" and "entity" aspect, preferring that it just go away.

We shall also refer to time-polarized EM waves as *time-density waves*. The reason is that a time-polarized EM wave is one where the EM energy is on the time-axis and oscillating along it, like the squeezing of an accordion. In short, it is a longitudinal EM wave on the 4th Minkowski axis *ict*, where the only variable is *t*. Hence it is a wave of the variation of time-density along that axis.

All three branches of KGB energetics use longitudinal EM waves, longitudinal E*M* wave interferometry, and Whittaker-structuring of fields and waves to provide organized curvatures of spacetime known as *spacetime engines* or v*acuum engines*. They also use time-polarized EM waves, as mentioned above. Any or all can also employ the quantum potential as well as negative energy EMP aspects. For now, the reader should just recognize that each infolded longitudinal EM wavepair comprising the scalar potential, includes both a *time-density wave* along the time axis and a longitudinal EM wave in 3-space {91}. So ordinary electrostatic scalar potential fi—ordinary voltage—is made of a harmonic structure of *coupled time-polarized and longitudinal EM wavepairs.*

Oscillations of time density oscillate the local rate of flow of time. In turn, since time has the energy density of mass, its spatial energy equivalence W *is* given by $W = c^2t$. Any change in a time density wave is a very powerful oscillation of loca1 spacetime curvature. So time-density waves constitute a very powerful form of infolded general relativity, inside potentials and hence inside all EM waves and fields. The KGB weapon scientists call that inner general relativity (time density waves) the *information content of the field*. Unfortunately our follows confuse that term with common "spectral analysis" which it very definitely is not. But as part of the deception plan, the Russians are quite happy to help them think it is exactly that, and in fact the Russians go to some pains to continue to help our fellows think that.

KGB energetics involves the use of a highly structured general relativity infolded inside electrodynamics—inside EM potentials, fields, and waves regardless of type. In addition, that very special GR is using the very strong EM force along the time axis as the main agent of curvature. By deliberately structuring these time-polarized EM waves and their dynamics, extremely powerful *spacetime curvature engines* can be employed, either created locally or at a real distance, or also globally in an entire selected distant region.

Once these spacetime curvature engines structure the distant targeted spacetime region. Then in that interference zone the engines therein act upon any mass present. In theory, any physical effect known can be engineered, according to general relativity, once one can create and utilize the proper set of spacetime curvatures (the proper set of spacetime curvatures and their dynamics). In GR, curvature of spacetime acts directly upon any mass present, and vice versa. With such technology one can therefore *engineer mass* to any level and at any distance (and eventually in any quantity) at will, using energetics methodology. *The engineered mass may be either living or inert* {92}. It follows that bioenergetics and psychoenergetics effects can also be engineered at a distance in similar manner. All it requires is the proper engines for the effects desired in the targets desired.

By using time-density waves and their dynamics, mind itself can be engineered and controlled at a distance as in the cases of Captains Button (Figures 22 and 23) and Svoboda (Figure 24).

The Importance of Scalar Photons and Time-Density Waves

Now we digress momentarily. Again, there are four major photon polarizations in quantum field theory {93}. The first two are *transverse* photons, polarized in the x- and y-direction respectively (transversely to the line of propagation, which is customarily taken along the *z*- axis). Ordinary transverse EM waves are comprised of such photons or blends thereof. Though ignored, EM waves prior to interaction with mass do *transport EM time-energy* as well as EM spatial energy, though in electrodynamics that time-energy transport has been ignored, thus confusing the effect with the cause.

The third polarization is a *longitudinal* photon, vibrating along the *z*- direction (the direction of propagation). Obviously longitudinal EM waves are comprised of such longitudinal EM photons.

The fourth polarization is a *scalar* photon, since it oscillates only along the time axis (the word "scalar" refers to the fact that it has no vector field component in 3-space). In American physics the latter photon is largely ignored since time is not an observable in quantum mechanics. However, such scalar photons do comprise time-density waves and EM time-energy currents, and they are involved in powerfully curving spacetime via oscillating the time domain. Hence by their oscillations they produce spacetime curvature oscillations, which in turn project small components into 3-space. Sufficient such coherent projection components are in fact longitudinal waves in 3-space.

The combination of the scalar and longitudinal photons is observable as the instantaneous scalar potential, and when Whittaker's 1903 decomposition of the scalar potential is slightly corrected, then the scalar potential is comprised of coupled pairs of waves, where each pair consists of a time-polarized EM wave coupled to a longitudinal EM wave. The combination of the scalar and longitudinal EM waves in the presence of charge is observable as the macroscopic scalar potential. Negative charge absorbs the time-polarized photon from the time domain and re-emits a longitudinal photon in 3-space, while the positive charge absorbs a longitudinal photon from 3-space and re-emits a time-polarized photon in the time domain.

Psychoenergetics: Engineering the Mind-Matter Interaction by Engineering Mind

We shall be interested in what actually constitutes a living biological system in the psychoenergetics view (Figure 46). Mind operations are *time-like* rather than existing in 3-space, and for that reason mind and its operations are not observable. Observation is a *d/dt* operator acting on an LLLT ongoing interaction, which removes the T and leaves a frozen LLL snapshot, via d/dt(LLLT) => LLL. *Imposition of observation itself breaks*, any and all connection with time-like operations, including mind and its operations.

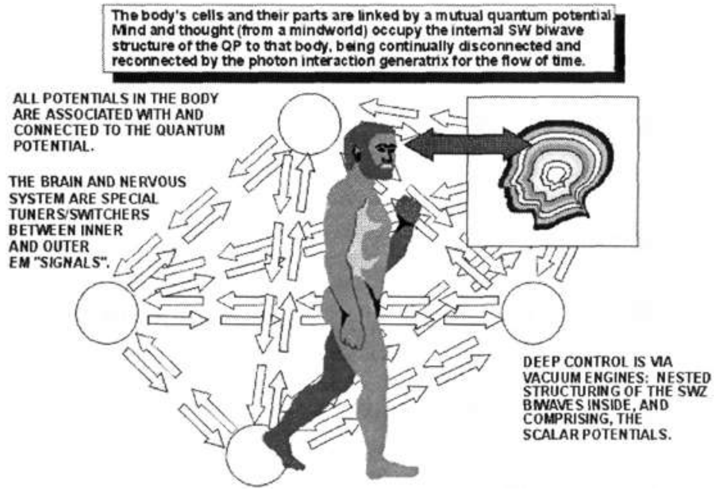
The body's cells and their parts are linked by a mutual quantum potential. Mind and thought (from a mindworld) occupy the internal SW biwave structure of the QP to that body, being continually disconnected and reconnected by the photon interaction generatrix for the flow of time.
ALL POTENTIALS IN THE BODY ARE ASSOCIATED WITH AND CONNECTED TO THE QUANTUM POTENTIAL.
THE BRAIN AND NERVOUS SYSTEM ARE SPECIAL TUNERS/SWITCHERS BETWEEN INNER AND OUTER EM "SIGNALS".
DEEP CONTROL IS VIA VACUUM ENGINES: NESTED STRUCTURING OF THE SWZ BIWAVES INSIDE, AND COMPRISING, THE SCALAR POTENTIALS.

Figure 46 A living biological system (psychoenergetics view).

"Dynamics" involves "velocity", and in general relativity's 4-space velocity is rotation (Figure 47). Any mind-operation or change is dynamics and produces a small velocity, which is a small rotation. This projects a very small component into 3-space, in the physical body, but still in the virtual state. We are speaking of physics rather than metaphysics! *Sufficient* coherent mind operations (changes) have just such small (virtual) rotation components in 3-space that will project coherent little projection components (virtual) into 3-space. Coherently sustaining these components (intent) results in their coherent integration to breach the quantum threshold and become detectable (produce an observable 3-space result). That is how an "intent" change sustained in the mind becomes an actual physical energy input to the body and nervous system.

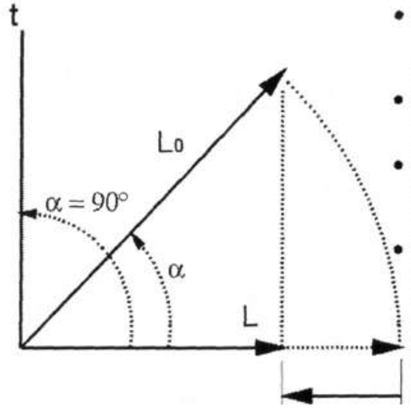

- An object moving with velocity v < c has rotated slightly out of 3-space toward the time dimension.
- Its "projection" back into 3-space is shorter, providing the Lorentz-Fitzgerald contraction.
- When velocity v = c, a full right-angle turn has been made, and the object has "lost" its projection along its direction of motion.
- Hence it appears as a 2-dimensional plane moving at the speed of light; in short, a photon or a light wave.

Lorentz-Fitzgerald Contraction

Figure 47 Velocity is rotation. Necessary to understand the mind-body coupling mechanism.

We again stress that the scalar potential is comprised of coupled time-density and longitudinal EM wavepairs. Mental operations also are time-density wave changes, hence they are changes to the interior of potentials, including the ambient potential of the vacuum itself. Potentials superpose, and their coherent internal structures add (integrate) coherently. Hence, mind changes can coherently integrate to produce minute but real, observable, physical input changes to the body system. *Mind permeates the body via the interlinked and superposed potentials of the body.* The living bio-organism's mind and its operations can be modeled as an organized operating wave dynamics inside a scalar potential and superposed potentials of the body of that organism. Since fields are made from the more fundamental potentials, these operations also function inside the fields of the body. The organized time-density dynamics constitute both the conscious mind and its operations and the unconscious mind and its operations. More subtle depths of the time-density dynamics also constitute the human collective unconscious mind functions and operations. Again, we are speaking physics, not mysticism.

See Figure 48. The "component projection" (coherent integration) process from time-like mind operations into 3-space in the body potentials introduces into the living body the *physical* commands correlated to the *mental intent* of that entity. This is the solution {94} to the long unresolved *philosophical problem of intent*, and addresses the unresolved physics problem of *just what constitutes the "observer."* But the subsequent physical changes accomplished by the responding body as a servomechanism, also produce small rotations into the time domain (into the mind domain) but in the virtual state. However, a series ol' coherent physical changes produces coherent virtual state changes in the mind realm, which in turn coherently integrate into perceptual input. In this fashion, the mind senses the response of the body to its order, and any errors in that response that need correction. The time delay gives also the sense of identity and personal persistence, hence creates the perception of self, the "responding body belonging to self," etc. Interwoven perceptions of coherent rotation changes from other 3-space phenomena that do not respond to mind intent, also create the perception of the "separate physical world." That these phenomena do not respond directly to mental intent creates the perception of the "separateness" of this "external" physical world, while the interweaving creates the perception of "existing simultaneously" with or in the external world.

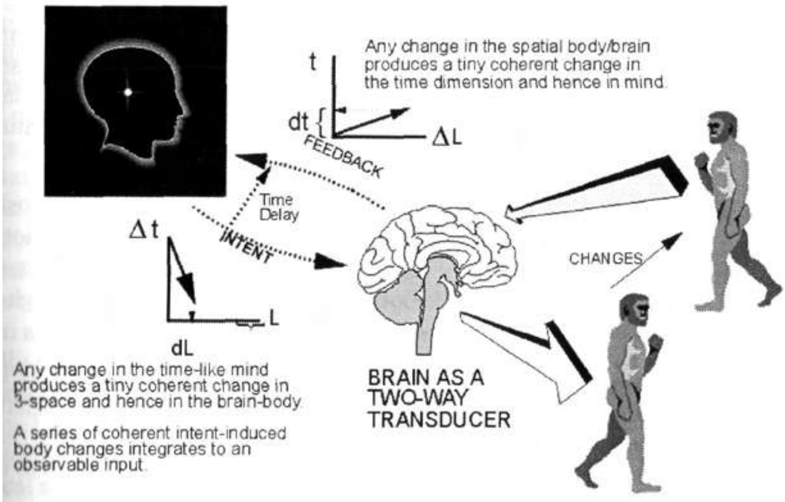
Any change in the spatial body/brain produces a tiny coherent change in the time dimension and hence in mind
t
dt
ΔL
FEEDBACK
Time Delay
Δt
INTENT
L
dL
CHANGES
Any change in the time-like mind produces a tiny coherent change in 3-space and hence in the brain-body
A series of coherent intent-induced body changes integrates to an observable input
BRAIN AS A TWO-WAY TRANSDUCER

Figure 48 Mind-body coupling mechanism at work.

All of that perception, dynamics, rotation, and processing is subject to direct engineering via psychoenergetics means.

Without further elaboration, we again stress that the conscious mind is a serial processor, whereas the unconscious is a massively parallel processor. Westerners are prone to equate self to consciousness or conscious mind operation alone. That is not the case. Consciousness is just one attribute of the self (of the total feedback loop having delay between feedforward and feedback).

The totality is the mind-body coupling mechanism (Figure 48) enabling a living biological system (Figure 46).

By using time-density waves formed of scalar photons, the mind realm also can be directly engineered by the same energetics technology. Hence, for the Russian energetics weapon scientists there followed the development of psychoenergetics as an entire class of breakthrough electromagnetic weaponry.

This resulted in the development of not only a unified physics of matter and space, but also a unified physics of mind and matter reaction. Any true unified field theory must integrate the observer and the unobserved as well as the various categories of the observed. It is now possible to *directly model and engineer both physical and mental reality and their interaction.* Russian psychoenergetics has been developed to do precisely this {95}.

Quantum Potential Weapons Are Utilized in All Three Energetics Branches

In addition to scalar interferometers, there is another category of energetics weapons used in all three branches. That category involves the use of *quantum potential weapons*, after David Bohm's quantum potential in his hidden variable interpretation of quantum mechanics {96}, published in *Physical Review* in 1952. We openly published one mechanism for creating a quantum potential {97}; there are others.

Presently these quantum potential weapons are the most powerful and most flexible weapons on earth. Nothing else can stand against them. Any nation not possessing QP weapons is already a second-rate power, regardless of how many nuclear missiles, bombers, and submarines it possesses, or how many high-energy lasers and high power microwave weapons it possesses.

The greatest problem the Soviets had was not in *making* the QP itself, but in (a) *localizing* its creation to the specific area or region desired, and (b) restricting its coupling to the type of equipment, entity, level, size, type of nuclei, etc. that was desired. We have not yet fully solved the problem of how they are able to do that, but have made a little progress upon it.

The Importance of Longitudinal EM Waves

The KGB energetics weapons use longitudinally polarized and time-polarized EM waves as their basis, rather than transversely polarized EM waves. There is a surge of renewed U.S. interest—and a revolution in electrodynamics—just underway in this country (and in the major weapons laboratories of other nations) *vis a vis* longitudinal EM wave solutions to the major equations of physics. E.g., one may check the Los Alamos National Laboratory Internet web site for summary papers by Rodrigues and Lu {98}. There are now formal EM wave solutions faster than light speed, slower than light speed, etc. {99}. There is also experimental support for superluminal tunneling of EM waves in waveguides {100}. In addition, as is well known, the speed of the ordinary scalar potential is infinite in the Coulomb (transverse) gauge {101}

Higher topological algebra is being used to embed the electrodynamics in most of the KGB energetics work. Maxwell's original theory {6} was some 20 quaternion and quaternion-like equations in 20 unknowns. Quaternion algebra has a higher topology than either vectors or tensors. Heaviside and others reduced the quaternion-like theory of Maxwell down to four equations in four unknowns—or to two coupled equations, depending upon how represented—and formed the much lower topology vector algebra in the process. This serously demoted the topology of electrodynamics and placed undue limitations upon what can be done with it. The subsequent mathematical promotion of electrodynamics to tensor algebra still did not recover much of the EM functioning possible in quaternion algebra. An even vaster, far more powerful electrodynamics emerges in *Clifford Algebra* electrodynamics, which may be **the** coming modern trend.

However, even with Clifford Algebra, there remain fundamental foundations difficulties—such as bluntly confusing *energy transport* with *energy dissipation, by* arbitrarily integrating the Poynting vector around a closed surface so as to select only the divergent portion of the Poynting energy flow. That produces the *energy dissipation flow* associated with the circuit, not the overall energy flow associated with it. The overall energy flow is some 10^{13} times greater in a nominal case, as (i) discovered by Heaviside in the 1880s, (ii) never even considered by Poynting, and (iii) deliberately discarded by Lorentz with a little integration trick {102} still used by classical electrodynamicists and electrical engineers. Electrical engineering departments, professors, and texts do not know and do not teach that all EM power sources—such as generators and batteries—actually pour out of their terminals a far greater energy flow than the meager shaft energy input to the generator or the chemical energy dissipated in the battery. Lorentz arbitrarily discarded the huge Heaviside nondiverged output energy flow component a century ago, with the observation that it "has no physical significance" {103,104}.

Part of the emerging electrodynamics revolution is intuitively obvious. Higher topology algebra used in one's electrodynamics model means that lots more things can be done—in circuits, equipment, and components. As an example, suppose one can only draw geometrical figures on a flat, uncurved plane. Then compare the figures there to what can be drawn when a third dimension is added so that the plane can be curved, twisted, folded, stretched, compressed, etc. in 3-dimensional space. That is similar to the dramatic differences between electrodynamics in a higher topology algebra and electrodynamics in the far more limited tensor and vector algebras.

Further, if someone's circuits *are working* in a higher topology, a lower topology analysis will not even see their true functioning. This turns out to be true for Tesla's work, where so many pundits have used vector or tensor algebra in the belief that they therefore knew what Tesla was doing. They did not, nor did they even use the proper tools to see what he did. For the proof, see Barrett's examination of Tesla's actual patented circuits, using quaternions {105}.

Energy, Wave Energy, and Wave Speed are All Dramatically Altered

The energy of a pure longitudinal EM wave (LW) is infinite (unlimited). In practice, only an imperfect LW is produced experimentally, and therefore its energy is finite, but that energy may be extremely large, and the wave's propagation may be many times faster than the speed of light. This is much like quantum tunneling, which has been measured to be superluminal. Nimtz *et al.*, e.g., have transmitted Mozart's 40th symphony through a barrier in a waveguide at 4.7 times the speed of light {100}. Other researchers have transmitted information at lesser velocities but still at greater than c, the speed of light in standard vacuum.

Einstein assumed light speed in vacuum to be universally constant. That is correct for *transverse* photons and *transverse* EM waves, but not necessarily for other polarizations. Instead, it can be a function of the energy density of the vacuum (i.e., its stress energy density), which can be a matter of the gauge and the wave polarization. Further, *c* as a constant applies to transverse polarized EM waves (which are "surface" waves of the medium, so to speak) and not to longitudinal or time-polarized EM waves (which are "subsurface" or "subspace" waves). In general, the speed of the EM wave can be a function of the energy density of the vacuum (i.e., its stress time-energy density). And it is also a function of the type of EM wave being utilized.

"Empty space" is filled and writhing with the continual transiting of higher order EM waves, the presence of standing high order EM waves, and energetically organized regions where the energy is in the LW mode and in the time-density mode. Indeed, space (more correctly, spacetime) itself is nothing but a plenum of such waves and their wave-to-wave interactions.

The longitudinal and time-density effects in empty space are correlated with seasonal and periodic variations of time-charging and longitudinal EM wave emissions from planets, stars, etc. As a result, the characteristics of empty space itself periodically vary, and these can be detected as variations in the speed of light through the affected space.

The interplanetary radar measurements, e.g., do indeed show several periodic variations of the speed of light in interplanetary space, on a weekly and essentially monthly basis. In this data the proclaimed constant speed of light is upheld only by averaging the measurements over a month-long period, as shown by Wallace {106}. Actually, the scientists are dealing with variations of the periodically varying local time-density of spacetime, but have not taken that into account.

There are also specialized "galloping wave" zones, created by variations in the interferometry of the interpassing and interfering longitudinal EM waves in space. Such galloping waves from wave interferometry are already known in the literature {107} from 0.1 to 100 times the speed of light, but they do this on a "surge forward and lag backward" mode about an *average* speed of c in the laboratory here on earth. In space away from planetary masses, that value "c" varies from its textbook value also.

All that has really been shown is that interplanetary space is *not quite* uniform after all, contrary to present scientific assumptions.

One would think that this *fundamental foundations anomaly* shown by the interplanetary radar measurements would excite the entire scientific community and trigger a massive scientific investigation. Instead, scientists such as Wallace who publish critical analyses showing that the data directly prove the variation of c in "ordinary" space, are just ignored and essentially "parked on the side-lines" until they eventually fade away.

A Process for Creating Longitudinal and Time-Density EM Waves

Today we know that some plasmas very often can produce a longitudinal EM wave from a transverse EM wave, and can produce a transverse EM wave from an input longitudinal EM wave. This was not known when the Priore team was performing its experimental work. However, it provides one method for producing and detecting longitudinal EM waves.

The Priore team in France unwittingly made longitudinal EM waves by inputing transverse EM waves into Priore's rotating plasma in a giant plasma tube (Figures 12 and 49).

- At the end of the project, Priore had developed a large unit to treat humans.
- A few were treated.
- Cures replicated what had been done in lab animals.
- The massive size of the unit and long treatment time (hours) made the process bulky and very inconvenient — and expensive.
- We proposed a way to do it with a suitcase-sized unit and a treatment time of two minutes.*

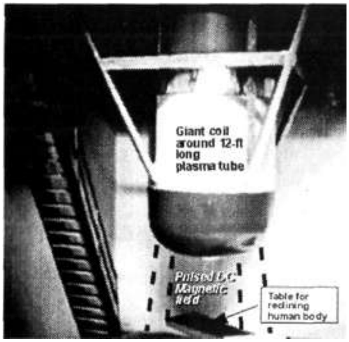
Giant coil around 12-ft long plasma tube
Pulsed DC Magnetic field
Table for reclining human body

*See *Porthole Briefing,* www.cheniere.org

Figure 49 Priore's large plasma tube installation in his 3-stories-high treatment facility for humans. Results obtained in animal experiments were also duplicated in human experiments.

Quantum Potentials and Multiply Connected Spacetime

It appears that a sufficiently clean longitudinal EM wave can be utilized to create a _quantum potential_, very similar to Bohm's notion, by iterative retroreflection between two phase conjugate reflectors. When sufficient _iterative self-targeting_ occurs and persists, the waves eventually "snap in" and fully converge to laser-like beams without any energy divergence {97}. At that point, one has achieved an essentially "pure" LW, which has infinite velocity. For all practical purposes, there is then little or no difference between this wave and a quantum potential.

A pure longitudinal EM wave just "appears instantly" everywhere in space that it will be. It is already an element of a quantum potential. In this sense, all that a "quantum" potential is, is an ordinary scalar potential in which the composite EM longitudinal phase conjugate wavepairs really are pure longitudinal EM waves. It would seem that this also moves the gauge to the Coulomb gauge, so far as the potential is concerned. It follows that, by establishing significant and clean longitudinal EM wave contact between any two points via iterative retroreflections, one establishes a quantum potential if sufficient harmonics are established in the contact. The implications for "phase conjugate shooting" to go one step farther are obvious.

By "scanning" an area with the appropriate LW transmitter array, with deliberately imperfect LWs so that contact reaction occurs, and then decreasing the imperfections in the scanning by iterative retroreflection, it appears that a quantum potential can be created and connected between the transmitters and a desired distant region.

Essentially, in the case of pure longitudinal EM waves and quantum potentials, we must model the situation in terms of a _multiply connected spacetime_. The divergence-free waves are now pure longitudinal waves in one sense, and pure potential beams in another sense. In any case, the Coulomb gauge applies. As is well-known, the potential moves at infinite velocity in the Coulomb gauge.

Thus a single station can be connected by such a QP to any number of distant stations—by first "knitting them together" into a common multiply connected spacetime by iterative mutual LW self-targeting beams. It matters not if one or more distant station(s) is/are then moved beyond the star Sirius; the moment an energy signal appears in the initiating station, it is already instantly in each distant coupled station anywhere in the entire universe. There is no separation distance involved. Here *one no longer needs to propagate transverse EM wave energy painfully through space at the speed of light.* Instead, one is now dealing with multiply connected spacetime. One now has the instant appearance *in each and every distant coupled target* of the energy input into the single "transmitting" or "coupling" station. So the process is a startling and giant "energy amplification method" as should be expected from something that has infinite energy in the first place!

One vulnerability and disadvantage of a QP weapon is that, if the target is aware of one's coupling and also has QP technology, he can now immediately send back any energy form or "greetings" he wishes, from any one of his coupled targets. He can even do it inadvertently, without even knowing he has been coupled by a quantum potential. In other words, the QP provides a two-way street, if the target is aware of it and has the technology himself, *and even if he is unaware of it and does not have the technology*. For that reason, actual use of the QP weapon in strategic strikes requires caution and special considerations. Usually it requires very rapid single use and then cessation of any further usage.

Need for Caution with QPs Used Against Biological Organisms

Another QP weapon disadvantage is that rather extreme care must be rigidly exercised, particularly when linking living creatures. For living things, the QP is best utilized in the absolute minimum energy mode. E.g., *an extremely weak OP already links every member of the human species*. Suppose one increased that linkage intensity carelessly. Immediately the entire human race is in grave difficulty. Every death of a person feeds into every other living person. Every birth of a baby feeds into every living human. Even the fertilization of one **human** ova instantly feeds into all other ova. The integration of all those changes would totally disrupt—and instantly destroy—all human life on Earth. Since chimpanzees have very similar genetics to that of humans, all the chimps would probably die also. Other higher primates would also suffer or die, as would many other species.

What might be useful, of course, would be to use a QP to kill every smallpox, HIV, Ebola, and other dread disease pathogen on Earth. If so, it would first have to be made extraordinarily specific, else it could threaten all or most of the life on Earth.

A very useful adjunct would be to use a QP device to nearly instantly scrub and decontaminate (disinfect) an anthrax-contaminated large city—such as New York or Washington, D.C.—by instantly killing all the anthrax spores in that locality. Or to clean all the antibiotic-resistant staph pathogens out of U.S. hospitals by just killing them all instantly and periodically. Or to rid the Earth of necrotizing fasciitis—flesh eating disease pathogens.

An extremely localized QP could be used to connect with all HIV viruses and DNA inside an AIDS patient's own DNA. One could then kill the HIV viruses not yet integrated into the patient's cellular DNA. One could also time-reverse the patient's own HIV-infected DNA back to normal, effectively "removing" the HIV DNA already dumped into the patient's own infected cells. In theory, 100% of the HIV infection could be removed from the patient's body, restoring him or her to a completely HIV-free condition.

A Hypothetical Example

Suppose one arbitrarily increased the even weaker QP that presently joins all living things on Earth, regardless of species. Instantly every living thing on Earth would die, and the Earth would be instantly converted to a sterile planet.

Even though *physically embodied* life was gone, however, all the forms of those species would remain in spacetime activity, their vacuum engines—for both minds and bodies—forever embedded in the local potentials and fields of the material Earth. In other words, the Earth's potentials and the local spacetime potential would still be *dimensioned* and *conditioned* for the life forms that once moved upon it in physical form. The dimensioning and conditioning (engines) for every life form that ever lived on the Earth *exists now* at every point in the Earth and in the local spacetime. With proper application of energetics, in theory it is possible to recover any of these life forms desired, back into living form.

When one understands that environmental aspect, one understands why the Miller-Fox-Urey biogenesis experiments {108,109,110} (Figure 50) would still produce living organisms again. A *physically* dead planet or test flask is not necessarily a "completely dead" planet or flask, when one considers the presence of spacetime engines and dimensioned potentials. One recalls that these forms and engines contain not only physical forms, but also the mind-aspects and the mind/body coupling and dynamics aspects as well.

In short, the entire biochemical notion of biogenesis from truly inert matter devoid of any form of life or pattern of life appears to be a false premise. Since potentials reach to the end of the universe, then any structured potential structures *(dimensions)* the universal vacuum potential everywhere in the universe {111}. The Miller-Fox-Urey experiments were not performed in a truly "sterile" environment after all, when the new energetics is considered {112}. One realizes that the entire universe is alive; the mere presence of a single living microbe or other form anywhere establishes an entire living noosphere. The Russian researchers—such as Kaznacheyev {113}— into EM induction of disease in targeted cells at a distance also realized that biology included a noosphere, and it could be engineered.

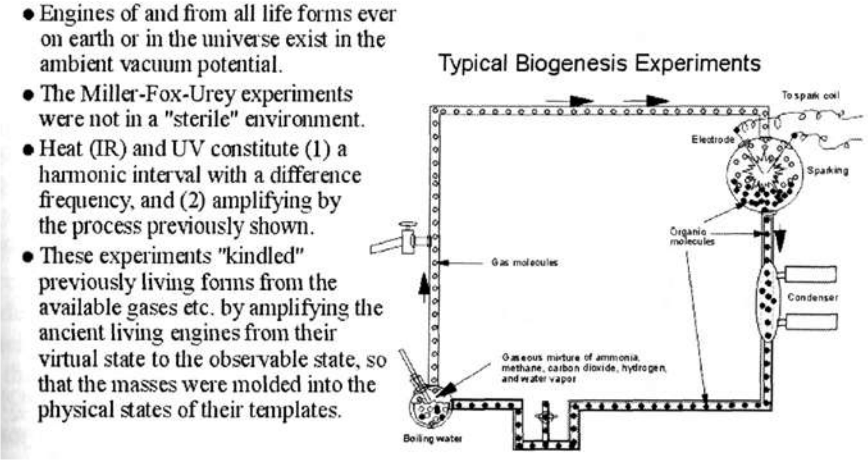

- Engines of and from all life forms ever on earth or in the universe exist in the ambient vacuum potential.
- The Miller-Fox-Urey experiments were not in a "sterile" environment.
- Heat (IR) and UV constitute (1) a harmonic interval with a difference frequency, and (2) amplifying by the process previously shown.
- These experiments "kindled" previously living forms from the available gases etc. by amplifying the ancient living engines from their virtual state to the observable state, so that the masses were molded into the physical states of their templates.

Typical Biogenesis Experiments
To spark coil
Electrode
Sparking
Organic molecules
Gas molecules
Condenser
Gaseous mixture of ammonia, methane, carbon dioxide, hydrogen, and water vapor
Boiling water

Figure 50 Typical Miller-Fox-Urey biogenesis experimental apparatus.

Herman Kahn, then director of the Hudson Institute (a leading think tank), once remarked that it did not concern him that there existed mechanisms that would destroy all physical life on earth. Instead, he stated, it concerned him that he know almost 200 such mechanisms. The quantum potential weaponry is yet another doomsday weapon with a thousand lethal mechanisms to add to Kahn's list, were he still with us.

The QP Weapon Is a Superb Energy Amplifier "For Free"

In a multiply connected spacetime established between multiple separated (to the observer) stations, one can instantly input one megawatt of power into *every distant stution* by inputting it into a single station. Suppose that coupling station is coupled via QP to one million distant stations (scattered anywhere in the universe). Then in each distant coupled station there now instantly appears one megawatt of power, even though we have input only a single megawatt in our "transmitter" coupled station. This is a superb *one-on-many weapon system and energy amplifier*—perhaps the ultimate weapon system.[30]

AS an analogy for understanding, one can look at it this way. Put a dot (point) on a piece of paper. Mathematics tells us that an indefinite number of other points can be "superposed" directly on that original point. Now input some "energy" into the original point. It instantly appears in each and every other superposed point. That is precisely the way a quantum potential linkage between stations works, except that in normal singly-connected 3-space the points are *observed* by a normal observer to be widely separated. Aristotelian logic—which is an "either/or" logic—does not apply. Instead, a higher topology logic applies.[31]

---

[30] However, we should not be loo surprised at such "astonishing" effects, once we appreciate that with the quantum potential and pure longitudinal waves one is playing with infinite energy anyway.

The QP weapon can be used with microwatt input to induce low-level disease states or pre-disease states *(shadow disease states)* (Figure 51) in entire populations. Tests done in this fashion will violate the "vector carrier" aspect of infectious disease transmission. Instead, the tests will cause a few widely spaced nearly simultaneously cases of the disease being tested. Several such tests (e.g., with necrotizing fasciitis) have already been conducted in the populace of the United States.

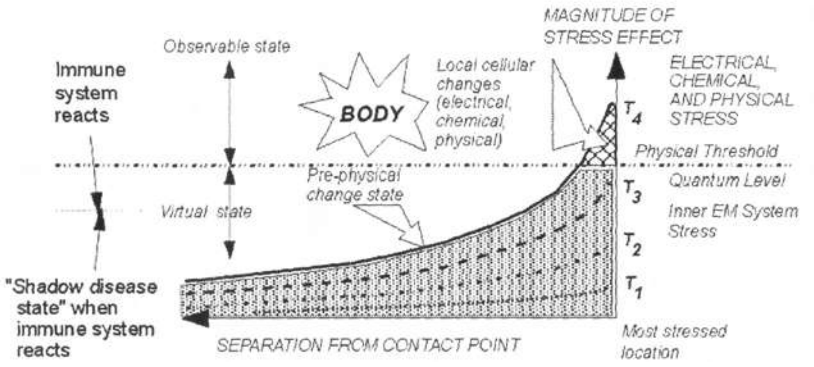
MAGNITUDE OF STRESS EFFECT
Observable state
Immune system reacts
BODY
Local cellular changes (electrical, chemical, physical)
ELECTRICAL, CHEMICAL, AND PHYSICAL STRESS
T_4
Physical Threshold
Pre-physical change state
Quantum Level
T_3
Virtual state
Inner EM System Stress
T_2
T_1
"Shadow disease state" when immune system reacts
SEPARATION FROM CONTACT POINT
Most stressed location

Figure 51 The shadow disease state is that structured virtual state level of the promoting disease engine template, at which the immune system reacts.

Or a QP weapon can be used with megawatt or gigawatt input to blast multiple distant targets simultaneously. One may destroy, e.g., every single transmitter of a given type. Or one may instantly kill all ships and weapons and paraphernalia in a large naval task force steaming toward its target. Or if one wishes to take the ships and equipment, the weapon can just kill all the humans instantly in that task force. ICBM sites, e.g., are vulnerable and obsolete since 1990 to any nation possessing QP weapons. So are underwater nuclear submarines and long-range strategic bombers. The entire strategic and tactical armada we have built up so painfully over the decades is obsoleted—also meaning *countered*—by any enemy possessing sufficiently advanced energetics weapons, and particularly quantum potential weapons.

---

[31] We developed such a logic years ago, and have recently included **it** in *Energy from the* ***Vacuum*** 2002, *ibid.*

Other Novel Uses of QP Weapons

If an enemy has such QP weapons and you do not, you are easily and directly destroyed at will, once certain other deterrents are overcome. One of these previous deterrents that had to be overcome was the stability of nuclear weapons in our nuclear arsenal. To remove this stability and render them easily destroyed, the holder of a scalar interferometry capability could induce sudden massive fissioning and fusioning of many of the targeted onsite weapons if the scalar potential interferometers were used massively. However, the resulting tremendous nuclear explosions—many "in the dirt and dirty"—would also unacceptably damage the attacker. Further, major antagonists usually clandestinely introduced hidden nuclear weapons into the cities and target areas of their opponents, which practice was known as "dead man fuzing." The notion was that, even if one side suddenly destroyed the other, that destroyed side could still detonate its secret nuclear weapons hidden in the attacker's homeland territory, destroying him as well. In essence, the killed side (the "dead man") had set off the fuzing of his pre-planted weapons, to destroy the attacker even though the "dead man" was already dead.

One strong reason that the Russians pursued the quantum potential weapon so strongly was precisely because it offered a mechanism to solve this "dead-man fuzing" problem, removing it as a viable deterrent. Obviously it would be quite advantageous to be able to dud or destroy these hostile weapons in one's very midst, even though one did not know their locations, and pull the fangs of the dead-man fuzing installed by one's opponent.

There appear to be ways of directly connecting a QP, e.g., to every atomic nucleus of a given type, in a given region (say, in a region consisting of on the surface of the Earth, in the Earth to some depth, in nearby space, on the surface of the ocean, and beneath the ocean). Connection can be made, e.g., to every U235 nucleus and to every weapons grade plutonium nucleus. Or with the technology developed, the users ought to be able to make such connection only between certain pre-selected *localized amounts* of U235 nuclei or plutonium nuclei.

Now suppose that such a QP is established between weapon-sized groups of U235 atoms. Then one hits one U235 nucleus in such a group in an activator station with a neutron bullet. Voila! One simultaneously hits every other linked U235 atom (including the activator station—one just blew it to smithereens if it contained very many U235 nuclei) in such weapon-sized groups. The raw U235 in ores in nature is still okay if the ore contained only tiny traces of U235 nuclei. But all the "neutron-struck" fissionable nuclei in nuclear weapons worldwide—regardless of owner or location—fission in the weapons immediately. That destroys all the nuclear weapons on Earth—including the clandestine nuclear weapons that various nations have spirited into other nations perceived to be their enemies, as a "checkmate" against sudden surprise attack. *One* does not need to know where a weapon is at all.

However, it gets all the weapons by exploding them, and that itself could create a giant nuclear destruction and fallout problem that would probably do in most life on earth because a great many nuclear weapons will have exploded "in the dirt and dirty." In short, it is an entirely unacceptable solution to the dead-man fuzing problem.

Of course, no *sane* enemy would wish to use the QP in such fashion, because it is quite suicidal for him and essentially destroys civilization. But suppose the sane enemy is well aware of the exact nature of the new cold type nuclear reactions (we explained these in an initial rough paper given in August 1998 at an Energy conference in Utah). The energy initiating these "time-reversed" nuclear reactions starts from "inside" the nucleons in the nucleus and moves outward (that is true of any "reversed" zone).

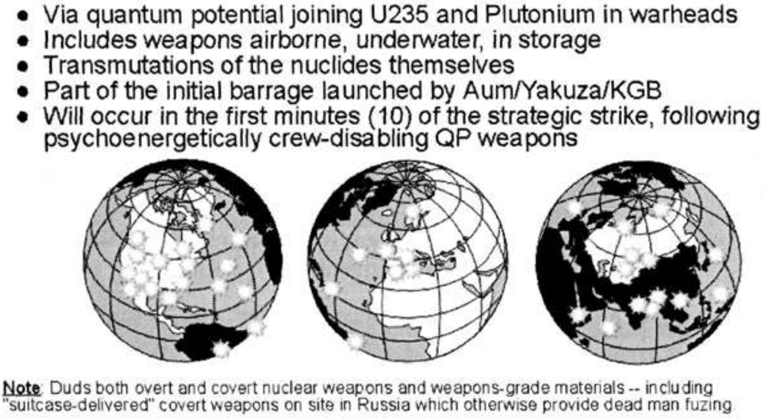

- Via quantum potential joining U235 and Plutonium in warheads
- Includes weapons airborne, underwater, in storage
- Transmutations of the nuclides themselves
- Part of the initial barrage launched by Aum/Yakuza/KGB
- Will occur in the first minutes (10) of the strategic strike, following psychoenergetically crew-disabling QP weapons

Note: Duds both overt and covert nuclear weapons and weapons-grade materials -- including "suitcase-delivered" covert weapons on site in Russia which otherwise provide dead man fuzing

Figure 52 Simultaneous dudding of nuclear weapons worldwide. Also duds nuclear reactors and nuclear propulsion systems.

Direct *isomeric transmutation* then becomes a preferred decay reaction to induce in the nuclear weapons one wishes to dud. By gradually inducing (slowly "kindling") one of these transmutation reactions, one can just "transmute" the U235 to—say—an isomer such as U238, with some readjustment and energy balancing in the ensuing rebalancing decay processes. In that way, one gets no nuclear explosions. One can even choose the transformation or several successive transformations so as to minimize any resulting nuclear radiation. The attacker also destroys all the nuclear reactors, including those in civilian power plants, nuclear submarines, and elsewhere. But he gets just about all the nuclear weapons if he transforms both U235 and plutonium (Figure 52). In about 10 minutes he will dud them all, without any nuclear explosions, but he would probably allow about one hour "just for good measure and nearly 100% certainty."

This offered the Soviets the ideal solution to the worrisome hostile nuclear arsenals of the world, since they also had additional nonnuclear strategic weapons of a different kind and of great power.

Apparently that very kind of QP weapon became operational, on site in Russia, circa 1989 under KGB control. For the first time, it provided the KGB with a guaranteed solution to the nuclear dead-man fuzing problem that had until then prevented its massive use of the scalar potential interferometers (such as in 1985-1986 as originally envisioned by Brezhnev).

By almost simultaneously leasing the earlier scalar interferometer weapons to **the** Yakuza and Aum Shinrikyo at the end of 1989, the KGB obtained an influx of vast new funds {114}. It thus concentrated even more intensely on the more advanced energetics weapons such as QP weapons and psychoenergetics weapons {115}.

So with a QP weapon one can attack multitudes of targets, using a single "coupler" or "transmitter," all without "propagating the EM energy through 3-s**pace."** Further, with the technology developed, only the number of outrigger stations coupled limits the energy amplification achievable. One can alter matter at any level, from the quarks in the protons and neutrons in the nucleus, to the lattice bonds, to the molecular bonds, etc. Or one can structure the emerging energy (that emerges directly out of the local spacetime itself), so that *structured local spacetime curvatures* (engines) appear in the target zone(s). These ST engines can be generated so as to perform macroscopic functions against real systems—e.g., jamming electronics, as against the U.S. air attack on Libya in April 1986 (Figure 25), or as was done by a longitudinal EM wave interferometer in the attack and kill of the U.S.S. Thresher (Figure 6) in April 1963 {116}.

The U.S. Narrowly Escaped Destruction Twice in 1997 Alone

On two occasions in 1997 the United States would have been utterly destroyed by the die-hard faction of the KGB, except for direct countering by a friendly little foreign nation. The present author and three colleagues played a very minor role by alerting that nation to the impending attacks. There was certainly no reason to contact officials of our own country; the U.S. was and is powerless to prevent such attacks, although *hopefully* that is changing! It made much better sense contact the only nation that could possibly avert our impending disaster. Of course we later also informed proper U.S. government officials of our actions.

The first strategic attack was scheduled for between early to mid-February, and involved (1) dudding all nuclear weapons, (2) striking the QP weapons in the little nation simultaneously, as well as extant scalar interferometer weapon sites in other foreign nations, and (3) then leisurely blasting cities, facilities, and installations with iterative and continued scalar interferometer strikes. This third phase would include work assigned to the Yakuza/Aum Shinrikyo teams. That is how the Yakuza (and Aum Shinrikyo) expected to revenge themselves for the Western defeat of Japan in WWII, and for the West's dropping the atomic bomb on Nagasaki and Hiroshima. To be able to "get in on the kill" is why the Aum Shinrikyo and Yakuza leased the scalar (longitudinal EM wave) interferometers from the Russians in the first place {117}.

These, along with negative energy EMP weapons, are the kind of energetics weaponry that now determines our fate, and against which the U.S. presently has still inadequate defenses. Our own scientific community does not believe such weapons even exist—apparently due to the "not-invented-here" syndrome. Most of the U.S. scientific community is thus relatively useless and outdated as far as contributing to the strategic survival of the United States is concerned.

Purpose of the Novel Attack on Captain Button

Hopefully, some scientists are aware of the strange incident that involved the mental seizure of Captain Button in his A-10 (Figure 22) and controlling him for well over an hour (Figure 23), including controlling his perceptions, intent, and actions while performing complex technical tasks. That was the test of a KGB *psychoenergetics* weapon, developing and to be deployed by the KGB. That was the "acceptance test", so to speak—deliberately conducted over the United States—for the KGB's system that would be produced in numbers sufficient to counter the QP weaponry, if the QPs of the little friendly nation were to be placed in the insane command and control mode. This was yet another Soviet counter, planned for the near future against the counter by which the little nation did force abortion of the KGB second attack (scheduled for May 1) in 1997.

The purpose of the attack on Captain Button can be understood as follows: The later second scheduled strategic attack on the U.S.—scheduled for May 1, 1997—was forcibly aborted after the friendly defending nation switched to an *insane* command and control mode. The insane C3 mode counters the MindSnapper type attack that instantly kills all on-site operators and renders the sites too "hot" for new personnel to come in from outside and control it. In short, the U.S. could have been destroyed in that scheduled May 1 attack, but with the change of the friendly QP weapons to insane C3, Russia would also have perished in the immediate quantum potential counterblow from the little nation[32] that would have been automatically delivered by the site weapons themselves. The May 1 attack had been predicated upon the fact that a preliminary MindSnapper (negative energy EMP) attack on the friendly QP weapon sites would have instantly killed all persons on site,[33] and neutralized the site for some months because of the resulting longitudinal EM radiation. However, the shift of the C3 of the QP weapons to insane mode rendered such a pre-emptive MindSnapper attack totally ineffective because Russia would still have been destroyed.

---

[32] There are also variations. As an example, the insane C3 can also include a separate channel requiring a countering "punch in" code every 15 minutes (as in a heightened alert status) or the control system will order the launch initiation (this accepts that all on site personnel have now been suddenly killed). So the insane control system is also flexible, and many variations and considerations are possible.

Hence another countering means was required, and in fact that additional system was already in development. It was a very sophisticated array of psychoenergetic weapons, each with a special, highly trained team of operators capable of fully controlling one targeted individual while he was directed and controlled to go about performing desired technical tasks. We estimate that it required one psychoenergetic weapon (very complex, multiple transmitters) and an operator team of perhaps 20 to 30 skilled operators (one for each major human function to be controlled) to adequately control the mind and actions of one distant person performing quite technical tasks. Hence several hundred such "weapon sets" were required to take over all persons on the targeted QP weapons sites in the friendly nation, have the controlled crews deactivate the QP weapons including using the security code and procedures to standdown for maintenance, etc.

Once the QP weapons were totally on standdown, then the MindSnapper would have struck suddenly and killed every living thing on site, neutralizing the site for several months due to subsequent strong longitudinal EM wave emissions from the struck site. No new crews could have entered the site and lived, so the sites could not be operated after MindSnapping, even though the QP weapons were still intact but just on standby. Of course, a third phase then would have involved strikes by scalar EM interferometers to leisurely and completely destroy the systems and sites, permanently taking out the physical system as well.[34] With the first two of that three-part series of strikes completed successfully, then the KGB weapon sites would have been absolutely free to blast North America and our European allies into smoking ruins, leisurely and at will, given the first 10 minutes of dudding all nuclear warheads worldwide, to take care of that old type of "dead man fuzing".

From the open reports of the Button incident, and from the kill of Captain Svoboda a few weeks later, the nature of the next strategic attack that would be scheduled was immediately apparent. Accordingly, we warned the little nation about the shift to a new attack plan, and estimated the new schedule between mid - 1999 and the first of 2000. Of course, we also passed that information on to our government as well, but apparently, no one believed it. The proper U.S. intelligence agencies, however, certainly intercepted my faxes to the little nation's Military Attache in Washington D.C., and particularly my fax a few days before May 1, sent directly to the Prime Minister of that nation and urging the change to an insane command and control system for the QP weapons and the reason why.

---

[33] We point out that it required a longer time to develop the more powerful negative energy EMP weapons capable of just destroying the sites, personnel and all. The first effective weapons of that type could easily kill personnel brains, nervous systems, and hearts while still doing little damage to the much more robust *QP* systems and on site equipment.

[34] We strongly believe there are shielding provisions at those sites, against scalar interferometry attacks, which reduces their effectiveness and delays their destructiveness. However, so far this analyst has not deciphered those shielding methods or mechanisms.

It turned out that, some months after I alerted them to the coming psychoenergetics, MindSnapper, and scalar interferometry attack in that order, to be scheduled between mid-to-end of 1999, apparently the friendly nation suddenly struck and utterly destroyed the KGB facilities for the MindController negative energy EMP weapons being painfully built up in sufficient number. Probably scalar interferometry weapons were used to destroy them. The onsite psychoenergetics crews and facilities were destroyed and neutralized completely.

Anyway, as we passed mid-1999, the type of "Y2K" problem we were worrying about was whether or not the attack was on, headlined by a psychoenergetics weapon strike on the little nation first, with the three phases as previously scheduled back in 1997.

Later, in early 1999 I realized the Russian/KGB development of the causal system robots (Figure 10), mentioned in the introduction. In the latter part of the third quarter of the year, working at my computer at night, I was struck by one of these very beasts from within my computer (Figure 20), in very bizarre but unmistakable fashion (previously detailed). In 10 seconds I was nearly killed and severely disabled for a period of time (it took four days to recover completely). Because of the features deliberately utilized by the CSR on my computer screen so I would recognize it, I grasped the nature of the attack (the KGB *wanted* me to recognize it, and live to tell about it!), how it was done, how it jammed the brain and vision center and strongly fibrillated the heart, upset the equilibrium, etc. I also realized that, in 20 seconds or so additional exposure, I would have been a very dead person, expiring with a very violent heart attack and severe brain embolisms. Note that the CSR deliberately struck directly from the computer screen through my retinas and onto the nervous system, passing through the vision center and flashing it tremendously, and jamming my mind, brain, and nervous system.

When I recovered shakily about four days later, I notified the small nation, once again, as best I could, giving the exact symptoms of the attack and explaining what it was. By then I knew what was coming down, the form it would take, and the deceptive cover (Y2K uncorrected errors!).

The little nation then quickly proceeded to search out—apparently *inside* the North American signals, potentials, fields, and waves "superhighways" for the CSRs—all the CSRs and destroy them in place. It turns out to be a fairly straightforward operation, once one possesses advanced longitudinal EM wave interferometry weapons and techniques. Locating a "very dense" and very dynamic set of LWs performing organized functions is analogous to a normal radar locating a large bomber in cloud cover. One "burst" from the LW "radar," and the targeted CSR is completely disrupted and electronically destroyed.

So a very short (lasting a few days) and strange strategic battle was fought in infolded space or subspace, inside our electrical and electrical systems and inside the fields, potentials, and waves in them. A mighty attacking armada of perhaps several million main battle systems was intercepted and destroyed throughout our homeland, without the slightest notice by our own government, scientific, and intelligence communities. And without a whimper of what is usually recognized as "combat actions".

Afterwards, a very simple and innocuous little message was received which, reading between the lines, simply meant "thank you very much." In short, my hard-pressed colleagues and I got a small "Attaboy!" for our role in alerting the friendly nation.

So Y2K and the Millennium Celebration came, while we waited to see if the CSRs were gone or if they would erupt with damage all over the U.S. Fortunately, they were gone like popping balloons, and so the *fourth* scheduled strategic attack upon the U.S. in the space of three years—this time actually launched—was successfully destroyed.

Whatever I can do to help the survival of my country I will continue to do without hesitation, regardless of whether the U.S. scientific, intelligence, and military communities are aware of the new threat or not, and regardless of whether they believe it or not. Presently our fate and that of the Western world rests on the shoulders of a small but advanced and friendly nation. That is beginning to change, it seems, but still not swiftly enough to suit our needs.

We simply must change all that, by developing QP weapons and defenses ourselves. It is not an option; as an example, China has now finished—as of earlier 2002—and deployed QP weapons. Almost simultaneously, China also finished and deployed the more powerful negative energy EMP (MindSnapper-type but much more powerful) weapons and deployed them as well. So we have entered a very dangerous period when the strategic changes that are likely to occur will be both massive and traumatic.

AS this is being written in early October 2002, it is obvious that our nation is also preparing to strike Iraq if need be, either with a coalition effort or alone. If so, then Saddam Hussein will certainly unleash whatever terrorist forces he has in North America. He will attempt to unleash upon us all the weapons of mass destruction at his command. This could even happen within the next 30 days, with the MidEast exploding, giant polarization of the Arab world, general chaos in the MidEast and upset of world oil supplies, and skyrocketing fuel and energy costs with severe impact upon the already seriously ailing U.S. stock market.

If Saddam Hussein succeeds in doing sufficient damage to our cities and energy infrastructure with more weapons of mass destruction than we have credited him with, then the potential economic collapse of the United States and much of the developed world is an uncomfortable possibility. If that happens, turf wars will erupt all over the world, the world economy may collapse, the KGB and rogue Japanese (Yakuza and Aum Shinrikyo) will move and unleash further destruction, Pakistan and India are likely to erupt into a nuclear war, and several other regions may also erupt into nuclear war and use weapons of mass destruction. What China might do in such event is anybody's guess.

In short, the Armageddon we have so long feared, with the resulting collapse or destruction of much of civilization, could well be approaching. Let us hope that dark picture is terribly, terribly wrong, and that we successfully avoid Armageddon one more time!

Meanwhile, we would be in much better shape to weather the coming storm had our scientific community not grossly betrayed us. But if and when millions of Americans are dying in the streets with no effective means of treatment, and no effective defenses to prevent onslaught after onslaught against our nation, all those deaths and all that suffering will be directly the fault of our own organized scientific community.

Rogue Japanese (Yakuza and Aum Shinrikyo) Have a Role

In the information briefing furnished the previously named agencies, we briefly detailed the elements of the attack, how our own nuclear weaponry can be dudded anywhere, etc.

We also detailed the strange role of the Japanese Yakuza and the Aum Shinrikyo, who leased the operational use of the early generations of KGB longitudinal wave interferometry weapons in 1990 directly from the KGB. The first down payment on the lease was $900 million. I previously furnished pertinent information on this to an investigative U.S. Government agency. Since then, the Yakuza have set up their own scalar interferometry weapons development facilities in Japan, including portable energetics weapons,[35] and have assumed a very secret and nearly unsuspected role somewhat similar to Bin Laden's role in the attacks of September 11, 2001—but much more deadly when the coming Armageddon erupts.

Inducing the Gulf War Disease

In 1986, the Russians had already solved the dead man fuzing problem theoretically, but had not yet completed the development and deployment of the necessary QP weaponry. The KGB's first "combat demonstration" test of the **ability** to localize a QP in a selected region of space was against the U.S. air attack on Libya in 1986. That test apparently used a laboratory apparatus to prove the process. In short, it was a major milestone whose success moved the system into full production engineering and production. The system was then developed at very high speed after the successful demonstration in 1986. The first strategic KGB QP weapon(s) became operational on site in latter 1989.

---

[35] We strongly suspect some of these portable weapons have already been infiltrated into the United States—e.g., to be used against nuclear power plants and other lucrative targets such as refineries, oil fields, etc.

It was this QP weapon that was used to generate the so-called *Gulf War Disease* in our soldiers in the MidEast. In the action package, I explained precisely how that was done. I have also finally placed the simplified gist of that explanation openly on the Internet for all to see (website www.cheniere.org). Those suffering U.S. veterans were *casualties of war,* not victims of just nerve gas whiffs or chemicals or their own stress, etc. The initial U.S. Government and U.S. scientific community's assertion that "it was all in their heads" and due to stress, insults every veteran who has ever served his nation. One justifiably has contempt for such treatment of our veterans. Sadly, the latest "studies" are **beginning** to conclude the same thing once again.

Apparently our intelligence and military command agencies have not yet learned t**he** lesson of the decades long Soviet microwave radiation of the U.S. Embassy in Moscow, which was responsible for health changes of many kinds in U.S. personnel and was also responsible for the deaths of three U.S. Ambassadors to the Soviet Union. Apparently we are still using radar engineers and electrical engineers for electrodynamics analysis, and these analysts apparently still believe that in the absence of the EM fields there can be no induced EM effects. One sadly comments that, just as the electrical engineering departments of our universities have been unable to walk across the campus to the particle physics department and find out what the discovery of broken symmetry in 1957 portends for their discipline and model, then the intelligence community apparently has not yet learned that no EM analysis of such things as the microwave radiation of the U.S. Embassy in Moscow or the shootdown of the TWA-800 is worth a tinker's dam unless it is a higher group symmetry EM analysis. As Barrett showed, our fellows still have not grasped what Nikola Tesla already accomplished in actual patented circuits circa 1900 and shortly thereafter. Tensor analysis and vector analysis simply will not show what is really there.

Our veterans were hit with a "cocktail mix" of specialized disease "spacetime curvature engines" just below the observable threshold, in the "shadow" (virtual, but nearly observable) state[36] (Figure 53). The immune systems and the cellular regenerative systems do react to that shadow state level.

---

[36] One should think deeply about this, in terms of the solution to the source charge problem (which our engineers have not thought of very much, much less solved). Every charge already "engineers" the virtual stale at will, converting virtual energy into deterministic observable EM energy. That means we can directly engineer the virtual state vacuum itself if we put our minds to it. **Actually** Nobelist Lee has already pointed out that we have *not even tried to do this!* See T. D.

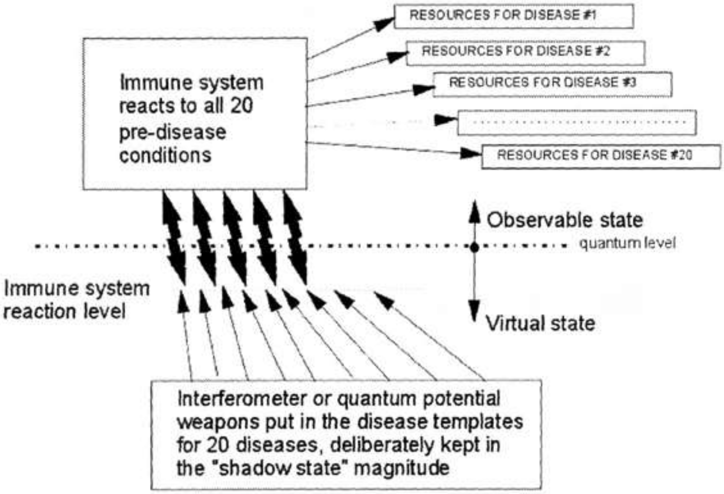
RESOURCES FOR DISEASE #1
RESOURCES FOR DISEASE #2
RESOURCES FOR DISEASE #3
RESOURCES FOR DISEASE #20
Immune system reacts to all 20 pre-disease conditions
Observable state
quantum level
Immune system reaction level
Virtual state
Interferometer or quantum potential weapons put in the disease templates for 20 diseases, deliberately kept in the "shadow state" magnitude

Figure 53 Spreading (thinning) the immune system's ability available to meet a new pathogen.

Such a "simultaneous mix" shadow induction "spreads" the reacting *finite* immune system and regenerative system (Figure 53) and their resources all across that entire mix of shadow diseases, so that only a much smaller portion of the immune system and of the regenerative system is directed against *any one* of the shadow diseases. This in turn greatly heightens (by a factor of 5 or more) the susceptibility of the body to all additional disease pathogens, and it greatly lowers the ability of the body to recover from any disease or biological warfare strike. In addition, the exposed veteran's body fluids and masses have accumulated a persistent "time-density charging" that very slowly decays (by . emission of weak *longitudinal* EM waves) over months or even a few years.[37] The sperm from these veterans did indeed have the innate "engines" and therefore the capability to gradually infect their wives. So did the saliva in their mouths. The weak "dimensioned" LW radiation from their bodies could also slowly alter the cellular regeneration and immune systems of their family members in close proximity, over a period of time, thus inducing the disease in them {118,119,120}.

Lee, *Particle Physics and Introduction to Field Theory,* Harwood, New York, 1981, in his "Chapter 25: Outlook: Possibility of Vacuum Engineering," p. 824-828. On p. 383 Lee points out that the microstructure of the scalar vacuum field (i.e., of vacuum charge) *is not utilized.* We add that the West does not utilize it. It has been utilized for some decades in the former Soviet Union and now in the continuing KGB energetics and bioenergetics weapons programs.

[37] See T. D. Lee, "Can Time Be a Discrete Dynamical Variable?." *Phys. Lett.,* 122B(3,4), Mar. 10, 1983, p.217-220.

The induction of GWS in our troops was in fact a test for the *present* clandestine use—already ongoing—of that technique to spread the immune and regenerative systems of the targeted U.S. population. This in turn gradually makes us more susceptible to terrorist BW attacks and disease pathogens of all types *from anyone else.* This tactic is already in place and accomplished, and it will substantially increase the American civilian casualties in our cities when they are subjected to terrorist BW attacks in the near future. *Those attacks now will have some five times the effectiveness that our present planners estimate.* In short, it is going to insure the deaths of millions more Americans. The approaching war in the MidEast, with Saddam Hussein's unleashing of his terrorist assets in the U.S. to the maximum extent possible, thus will probably see a five-fold increase in the effectiveness of some of his mass destruction weapons that Hussein's assets are able to unleash on America.

The entire affair of the Gulf War Disease was both a preliminary test and also a magnificent intelligence probe by the KGB, to clearly demonstrate once again that the U.S. Government and the U.S. scientific community are still nearly totally unknowledgeable of QP weaponry, immune and regenerative system "spreading," disease induction by bioenergetics means, etc. It follows a methodology used with great success for four decades by the KGB,[38] and one that our own government, scientific community, and military community has yet to decipher. *So long as the threat only slowly increases, Americans remain complacent in their ignorance, ignoring the boa constrictor slowly strangling them, while awaiting the cobra seen in the distance?*[9] We have unwittingly cooperated with the KGB boa constrictor for four decades, even as it slowly prepared to crush us. In the Gulf War, that crushing moved to the next threshold, to actually inflict casualties and deaths on U.S. troops and totally get away with it because of the bioenergetics (in the Russian sense) ineptness of the U.S. scientific community.

---

[38] The decades-long microwave radiation of the U.S. Embassy and induction of health changes and diseases, was a case in point. A high level target—the U.S. Ambassador—was involved, guaranteeing U.S. Government attention at the highest level. Puzzled, the U.S. government would then call in their most trusted EM scientific analysis groups. By the responsive actions then taken by **the U.** S. at the Embassy, the KGB could ascertain with absolute certainty whether or not our scientists knew of the use of infolded engines in bioenergetics. For more than four decades our actions assured them with 100% certainty that we knew nothing of what their "information content of the field and potential" actually was, or how it had been weaponized and was being used right in front of our faces and before the entire world.

[39] It is bizarre that our scientific community—even after those decades of EM induction of diseases and health changes in American personnel in the U.S. Embassy in Moscow—*still do not and cannot believe that conditioned EM radiation can induce cellular disease.* Such total blindness is inexplicable.

40-Years of Russian Induction of Diseases In U.S. Embassy Personnel

Sadly, our own fellows have never even figured out how the diseases and health changes were previously induced by the KGB for decades in personnel in the former American Embassy in Moscow. Three U.S. Ambassadors to the Soviet Union died as a result of their exposure to that radiation. Many other Americans on site were made ill or suffered health changes.

When Johns Hopkins scientists studied the problem, they found a most anomalous thing: no health changes occurred in areas *where the EM forcefields existed* in the microwave radiation. Instead, the health changes only occurred in areas where the potentials were gradient free so that the force fields were zero—and thus the potentials (and their specific internal "engines") were *persistent.* Using only a classical EM force field analysis—which does not show any of the internal longitudinal EM wave substructuring of the fields, potentials, and waves—the Johns Hopkins scientists erroneously concluded that it could not have been the weak microwave radiation causing the health difficulties, *because no such health difficulties occurred where the forcefields were present.* That is, they very naively assumed that "no force fields" meant no EM effects were possible. Their entire findings thus were based on a total non sequitur, the assumption that there can be no electromagnetically induced physical changes in matter in the absence of fields! The Aharonov-Bohm effect alone falsifies that entire premise. So does general relativity and higher group symmetry electrodynamics.

Actually, the Johns Hopkins scientists had just clearly proven the exact opposite: *that it indeed was the radiation that induced the health changes.* That is simply shown statistically, and it is a deep and continuing mystery how good scientists and their governmental overwatch did not notice the violation of inference methodology.

Simply put, there was 100 percent correlation to the presence of *gradient free EM potentials— the persistent kind.* There was zero correlation (or 100% anticorrelation) to the presence of the normal electric and magnetic fields—i.e., where the potentials changed over a very small region of space. Hence it could *only* have been the constant potentials, else we must throw out statistical inference theory altogether.

If the changes had *not* been caused by the EM radiation at all, one would have had some health changes in regions where fields were present and some health changes where the fields were absent. Considering the lengthy time the radiation persisted and the number of personnel affected, the statistics are decisive. Without question, the cause was indeed the persistent potentials where the **force** fields were zero. Whittaker's two papers—one in 1903 and the other in 1904-would then show how it was done. The 1904 paper and superpotential theory alone would *not* have explained it, since that alone does not include the "information content of the field"—i.e., spacetime curvature engines infolded inside the normal EM potentials, fields, and waves.

So, the U.S. government and U.S. scientific community "botched it badly," unfortunately, by elemental statistical inference standards. And unfortunately, they have continued to "botch it" to this day. Both the scientific community and the intelligence community missed it. Elementary statistics theory alone, coupled with the Johns Hopkins results, clearly indicates that the microwave radiation was causing the health changes, to a very high degree of confidence.

This demonstrates how "totally inexplicable" some of the effects generated by longitudinal EM wave interferometry and by "structured" or dimensioned potentials, fields, and waves can seem to our conventional scientists. If we examine the serious foundations flaws in classical electrodynamics and electrical engineering taught in our universities, and the near-total lack of higher group symmetry electrodynamics training, we can understand why our scientists and intelligence analysts did not and do not have the background required to properly assess the incidents and effects.

Western Electrodynamics Is in Sad Shape

Actually, numerous scientists have long pointed this out. For example, quoting Mario Bunge {121}: *"...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state."*

In quantum electrodynamics, it is the *potentials* that are the primary causes of all electromagnetic phenomena, not the force fields as used by Johns Hopkins and the Government in its analysis. We know that classical EM is wrong in its focus upon force fields as primary EM causes; the Aharonov-Bohm effect, Berry phase, and geometric phase work in some 20,000 papers in the physics literature already show that. Indeed, we also know that the classical force fields do not even exist apart from mass; in space, only the potential for such a force field exists, in case some charged matter is brought in to interact with the EM entity as it exists in mass-free space (as curvatures of spacetime).

The Aharonov-Bohm effect and its implications have been published in *Physical Review*—our most prestigious physics journal—since 1959. Later publications have included the Berry phase extension of the AB effect, and Aharonov and Anandan's further extension of the Berry phase to what is today called the geometric *phase*. The potentials are indeed the primary EM causes, and *the force fields do not even exist as such in mass-free space, but only in mass.* Yet our classical EM taught in our universities has not been changed in its fundamental acceptance of the material ether since 1865.

Every university in America is still teaching and defending a rather mangled electrodynamics where

(1) The material ether is still assumed, even though destroyed more than a

century ago by the Michelson-Morley experiments;

(2) The force fields are erroneously assumed to exist in massfree space, but in force-free form(!!!) when what actually exists there as an EM disturbance is a general relativistic perturbation of spacetime (i.e., an oscillation of the local curvature of spacetime). Actually, force identically is the time rate of change of momentum, where momentum is mass times velocity. Thus mass is a *component* of force and force field. There is no force or force field without mass. As Feynman put it, there is no force field in massless space; instead, there is only the potential for such a force field to be formed at the various points, should a charged mass be placed at (added to) each of them;

(3) The source charges are erroneously assumed to *create* their fields and potentials and all that energy in them, reaching across the entire universe from each and every charge, right out of nothing at all. This perpetual motion and creation of energy error continues even though particle physics has known (and proven, both theoretically and experimentally) that any charge (as a set of composite dipoles, when the charge is taken with its clustering virtual charges of opposite sign) is a broken symmetry in its fierce energy exchange with the vacuum. Hence the charge extracts that energy in its fields and potentials *from the active vacuum,* and outputs the extracted energy across space {122}.

Our own electrodynamicists and electrical engineering professors and departments would have us believe - and they continue to imply such in every university in America—that every charge is a perpetual motion machine clearly in violation of the master energy conservation dictum that *energy can neither be created nor destroyed.* The complete lack of scientific understanding that every charge extracts energy from the vacuum and freely and continuously pours it out as real observable EM energy, is reflected in rather silly "witch-hunts" {123} against electrical systems which exhibit more energy output than the operator himself inputs—i.e., systems which exhibit a coefficient of performance (COP) of COP>1.0, which are open EM systems far from equilibrium in their active environmental exchange.

Let us reason together! If the second law of thermodynamics really did prohibit COP>1.0 *electrical* systems, then it also would prohibit the demonstrated COP = oo of a solar cell array. It would also prohibit the demonstrated COP>1.0 *mechanical* systems such as a windmill, waterwheel, or sailboat—all of which have COP = oo! In other words, the second law cannot even be a law if it discriminates only one type of energy prohibited from being used for COP> 1.0,

For some inexplicable reason, EM COP>1.0 systems have come to be dogmatically labeled "perpetual motion machines" by the professional skeptics since EM systems must take their excess energy from the known active vacuum environment! In short, the dogmatic scientific community is saying that "windmill-type EM systems" with COP = oo are permissible for exchanges with *physically active and observable* environments (such as the sunlight used in solar cells, the heat in the air used in heat pumps, the water currents used hydroelectric power plants, the winds used in windmill power generation, etc.), but they are not permissible for exchanges with the proven modern active vacuum *virtual particle flux energy* environment, even though the discovery of **broken** symmetry in 1957 already proves that all EM energy is indeed freely taken from the seething vacuum environment by the asymmetry of the modern dipolarity of the source charge! We refer the interested reader to the proven broken symmetry of opposite charges {124}. We also refer the reader to Nobelist T. D. Lee's statement applicable to the skeptics' naive notion of symmetry. Finally, in modern physics all the major forces of nature are generated by the exchange of virtual particles. It is these forces that then translate against a resisting mass or entity to provide useful work. So all useful work in the universe already comes from the interactions of the seething virtual flux of the vacuum. On what distant planet have the skeptics been hiding while all this modern physics has been developed?

One must keep one's sense of humor! It is ironic but true that every major university in the U.S. continues to teach, uphold, and advocate—by implication—EM charges as ubiquitous perpetual motion machines, freely creating energy from nothing. And that is those professional skeptics and witch-hunters never mention that their own naive understanding of electrodynamics already implicitly assumes and teaches perpetual motion machines freely creating energy from nothing at all—and on a scale unparalleled in scientific history!

We have one chart with some 32 major foundations errors and flaws in the electrodynamics taught in all our schools; there are many others in addition. **This** erroneous teaching has so indoctrinated our graduate students that most **electrical** engineers no longer know the real difference between power (the rate of change of form of energy and thus the rate of production of work) and rate of energy flow (whether it is being dissipated or not).

Every textbook speaks of "sources furnishing power," when it is the rate of dissipation of energy in the individual loads and losses to provide the rate of doing work that *constitutes* power. Sources do not furnish power! They furnish energy flow (specifically, their source dipole furnishes energy flow, extracted and transduced directly from the active vacuum, and only a tiny bit of that enormous energy flow is diverged into the circuit and used). The various components of the receiving circuit interact with that energy flow to catch and dissipate a little of it, thus producing power and work locally in the intercepting and dissipating component.

A massive nondivergent (nonintercepted) energy flow has zero power (time-rate of doing work), but if intercepted and diverged, the divergence may develop enormous power (time rate of doing work) in and upon the interceptor, diverger, or collector. Even many physicists today are confused on the difference between rate of work and rate of energy flow per se. E.g., for about 30 years now, there has been a polite debate involving that very subject in the *American Journal of Physics.* It is still not completely resolved, because the argument has not actually addressed where all EM energy in fields, potentials, and waves comes from in the first place: directly from the seething vacuum via the asymmetry of the source charge. It also has not addressed the long unaccounted giant Heaviside nondiverged energy flow component.

Toward a Deeper Electrodynamics

The first thing the Soviet Academicians did after WW II was to completely overhaul electrodynamics, but in highly classified areas controlled totally by the KGB. Stalin's boot firmly planted on their backside saw to it, creating the terrible fear that if they failed to produce, they would surely die. Not having experienced such a "Stalin boot," our own scientific community continues leisurely onward, while it teaches 110 years old electrodynamic material filled with non-sequiturs.

The Russian use of the "information content of the field" euphemism is a case in point. Our fellows (in both the intelligence community and the scientific community) assume the Russians are speaking of simple transverse EM wave spectral analysis. Nothing could be farther from the truth. Let me explain.

In 1903, E.T. Whittaker—a leading theoretical physicist at the time—showed that any scalar potential is comprised of a harmonic set of bidirectional *longitudinal EM wavepairs.* Each pair consists of a longitudinal EM wave and its true phase conjugate replica (i.e., its "antiwave", or time-reversed twin, coupled together).

Unknown to Whittaker, that LW wavepair actually includes a *time-density wave* in the time domain, coupled to a longitudinal EM wave in 3-space. Per quantum field theory {66}, such a coupled pair would be observable as the scalar potential. Knowledge of this specialized coupling is necessary in order to properly decipher the Priore mechanism. That mechanism was demonstrated to *time-reverse* (dedifferentiate) a diseased or damaged cell and all its parts back to it previous normal state, *genetics and all.* The time-reversal was accomplished by "pumping" the cell and all its parts in the time-dimension so that time-reversal propagation of the mass-energy is achieved. The time-domain pumping is done by the associated time-density wave, which always accompanies any longitudinal EM wave a priori. So sometimes we sloppily say that the "longitudinal EM wave pumping" causes the time reversal of the mass, but it is really the associated time-density wave pumping that causes it. The time-density wave component also can cause *time charging* or *time-excitation charging* of a mass. During its alternate existence as masstime, the time portion is altered, and that is a time-excitation or time-energy excitation.

It turns out the human cellular regenerative system (Figure 29) also uses the same time-excitation charging methodology to produce time-pumping of the cell and all its parts. This produces a slow time-reversal (dedifferentiation) of damaged cells back to a previous state. This is in fact the basic mechanism for healing, as used by the living body.

Reversing aluminum, for example, just converts it back to ordinary aluminum. Not too interesting.

Reversing an HIV infected cell, however, would produce a healthy, normal cell with normal genetics. When a damaged cell turns into a cancerous cell, it is due to a process called *promotion* (change of the cell to a cancerous cell). *Reversing a* cancerous cell would change it back to a healthy, normal cell again—in short, it would constitute *demotion* of cancer and then healing of the damage in the now normal but damaged cell. In the AIDS patient, simply pumping the entire body with the time-density wave accompanying the longitudinal EM waves produced in a plasma—as Priore did for terminal tumors—would time-reverse all cells, completely ridding them of any residue of HIV infection. This would remove all the "HIV factories", thus completely solving the problem of AIDS. We have also uncovered a more advanced process where *matter can essentially be directly engineered in any conceivable manner,* including nuclear transmutation etc. We have filed a Provisional Patent Application and will file formal patent applications on those processes in the future, as soon as funds are available (patent attorneys are expensive!)

Time Is Highly Compacted Energy

Since time is energy compacted by some 9×10^{16}, even small time-density changes can produce enormous spatial energy, should any of that compacted energy be *transduced* from time-density waves into transverse EM waves and thus into spatial EM energy. Such transduction can indeed occur in nonlinear materials and processes, such as in electrolysis.

The New Transduction Model Explains the Cold Fusion Results and Mechanisms

In August 1998, we presented a conference paper pointing out that cold fusion uses this very *transduction of EM wave types,* one into the other. We presented such new concepts as (i) TW to LW to TDW transduction and in the other direction also, (ii) time density charging of a particle to an excited time-density state (analogous to energy charging to an excited energy state), and (iii) the observation that excess phase conjugation in a local region can generate a time reversal zone where like charges attract and unlike charges repel.

With these concepts we were able to explain the excess heat appearing in the cold fusion experiments of hundreds of researchers. We also came up with (iv) a totally new class of nuclear reactions without the usual high-energy considerations. That is, we showed the exact nature of the electronuclear reaction that is forming the anomalous nuclides in cold fusion experiments. Production of the new nuclides was revealed to be simply what one would expect from this mechanism, as indeed was particle clustering of like-charged particles.

We also explained (v) all the instrumental anomalies occurring in a series of well-documented cold fusion electrolysis experiments at China Lake. In fact, (vi) we added a completely new term to the conservation of energy law, thereby extending it.

An updated explanation is included in my new book, *Energy from the Vacuum*, 2002, *ibid.* With the new transduction interactions, one must state that the total energy in (a) spatial energy form, (b) mass-energy form, and (c) time-energy form is what is conserved.

So when transduction occurs in cold fusion experiments (electrolysis), some of the time-density charge energy is gradually transduced into ordinary transverse EM wave energy, which as it scatters produces excess heat not accounted for in present particle physics interaction models.

Scientists with only the "old" 2-term Einstein conservation law in mind, will be firmly convinced that this excess heat represents a violation of conservation of energy, hence is impossible. So they will question the calorimetry ability of the cold fusion scientists, etc. Unfortunately, the more ardent defenders of the orthodox faith will unjustifiably impugn the intentions and morals of the experimenters themselves with vicious ad hominem attacks. Such has indeed been precisely the case. Many of the harsh scientific critics have acted like cur dogs in a pack attack, and not as scientists at all.

Once one analyzes cold fusion in the new 3-term conservation manner and with the new transduction concepts, one immediately sees that the *overall* conservation of energy is upheld, while the *special case* represented by the former energy conservation law (conservation of the sum of the spatial energy and the mass-energy) is *permissibly* violated. The former case was a special case in that it arbitrarily assumed that no transduction of time-energy into spatial energy or mass energy—or vice versa—occurred.

In essence, successful cold fusion experiments show that, via transduction, a bit of time itself can be "utilized for fuel" and transduced into ordinary EM energy. Energy-wise the system has become a nonlinear open dissipative system[40] far from thermodynamic equilibrium {125}. As such, it is permitted to (1) self-organize, (2) self-oscillate, (3) output more energy than it itself has, by merely receiving, converting, and using excess energy from its environment (in this case, from the transduction of a tiny bit of the time-charging that has occurred), (4) power itself while simultaneously powering a load or performing useful work, and (5) exhibit negentropy. With the exception of transduction of time-charge into excess energy, all of that is already in the thermodynamics texts, though it has not been clearly spelled out in electromagnetics texts. The source charge, e.g., exhibits all five of those capabilities. *The inclusion of the time-domain as part of the external active environment, however, is an extension of* ***previous*** *open dissipative system work by leading researchers such as* ***Prigogine.*** *It dramatically extends the meaning of "active environment".*

---

[40] But we point out one idiosyncrasy of thermodynamics, which defines a "closed system" as one in which mass does not cross the system boundary, but energy is free to exchange across that boundary {125}. Hence, eerily a "closed thermodynamic system" is actually an open system with respect to energy exchange between system and environment. As such, energy-wise the system can be a system far from energetic equilibrium with its external active environment, and still be considered a closed thermodynamic system! Energetically, a *closed thermodynamic system* that exchanges energy but not mass is permitted to exhibit COP>I .0 and even COP = oo!

Why There Is a Time Delay in the Onset of Cold Fusion Phenomena

With an energy-to-time compaction ration of $9x10^{16}$, energy processes have to transduce spatial energy into time for quite a while in order to build up very much *time-excitation charge*. This leads to an initial delay in the onset of the "new law" transduction phenomena, as directly shown by the vast majority of careful cold fusion experiments. This build-up occurs during the "palladium-lattice loading" time, e.g.

By the same token, there is a time delay after cutoff of the input power to such experiments, during which time the slowly decaying time-density charge by its transduction into spatial energy continues to maintain the anomalous new processes (such as excess heat production) as they slowly subside. So the hours-long or even days-long delay phenomena after input power shutdown are also explained by the new approach.

In summary, there exist some very anomalous but very good experiments by excellent scientists—more than 600 of them—that are at least totally consistent with the above new wave polarization transduction model. Consistency with experiments is the first element of proof (it is necessary but by itself it is not sufficient). The new model also contains and obeys the conventional conservation of energy law until the new transduction phenomena become observable, at which time the extended energy conservation law applies. So in its lower limit, the new model corresponds to the present model—again, another requirement of any model purporting to extend a scientific area past that area's conventional model description.

The KGB weapons scientists—not the regular Russian academic establishment—have about a 40-year lead on us in all these areas. While our own scientists have dallied and pursued the standard model, the Russian weapon scientists have produced the equivalent of 10 Manhattan Projects, back-to-back, in the "new" energetics area.

The Early Soviet Longitudinal Wave Interferometers

Using longitudinal EM waves, the early KGB energetics interferometers could pass their beams right through the earth and ocean. LWs readily penetrate matter, reacting with it only weakly.

Some LW interferometry was also done using ordinary EM waves as carriers for the LW wave functions. Each such transverse wave carrier is in fact two scalar potential functions (Whittaker 1904, the initiation of superpotential theory). In each scalar potential for which functions are extracted and used to make the normal transverse carrier wave, there exist LW biwave pairs comprising said potential. So by assembling the desired LW biwave subsets, the two scalar potentials and their resulting interference-produced carrier wave are internally structured as desired. This is the use of the so-called *internal information content of the field*. We speak of such a deliberately structured potential, field, or wave as having been *conditioned*, *dimensioned*, or *activated*. If these terms are deemed inappropriate, just use the more generic term "internally structured." Then one has to explain what the "internal" is, and how this "internal" can be "structured."

In the distant interferometry zone, those *dimensioned* potentials, fields, or carrier waves would interfere and produce the desired alterations in the local spacetime (including specific spacetime curvature engines). Remember, the potentials are just bundles of special waves. We are discussing a special form of nonlinear multiwave interferometry, where the individual waves and waveforms comprising the total wave package assembly constituting a carrier have been carefully chosen in advance.

In the local spacetime in the distant interference zone, such multiwave longitudinal EM wavepair interferometry creates ordinary EM waves and energy that arise directly from each and every point in that individual interference zone. We stress that ordinary transverse wave EM energy propagation through the intervening space between the interferometer transmitters and the distant target *does not occur.*

In January 1960, in a speech to the Presidium, those are the weapons then in final engineering development for production that Khrushchev spoke of as "fantastic new weapons".

Khrushchev Killed the Thresher to Prevent His Own Dismissal

The first deployed large force-structure interferometer weapons became operational circa 1 April 1963. Khrushchev was still smarting from his Cuban missile crisis fiasco and his humiliation by Kennedy. Having overplayed his hand and gotten caught,[41] Khrushchev was desperate to stay in power and not be thrown out by the Communist Party leaders for such a massive loss of face.

On April 10 of 1963 Khrushchev used his new interferometers to reach through the earth and ocean and "kindle" EM jamming in the electrical controls of the **U.S.S.** Thresher nuclear submarine underwater off the East Coast of the U.S. **The** reader should recall that, wherever the interference zone is established—including in the earth or underneath the ocean, in the atmosphere, or in space—the energy arises directly out of the local spacetime. *There is no Faraday shielding of such effects, which do not depend upon ordinary transverse wave EM field energy propagation through space.* The energy appears everywhere within the targeted object, simultaneously.

It is extremely important that the reader recognize the dramatic difference in the way energy propagates in the new mode as compared to the old "propagation through space" mode. In the old mode, energy transport arises in one spatial location and is considered to proceed serially through the intervening spatial points to the intercepting and interacting particle's location. The Poynting model presently uses only this mode. In other words, the energy propagates from *without* the object *to* the object, and thence *inward inside* the object, if possible as the object interacts.

On the other hand, when the energy arises at each and every point in spacetime, it then propagates outward from each arising point. Thus in the nucleus of an atom, the "energy flow" starts at all interior spatial points and flows or diffuses outward in all directions. So the energy "meets" and interacts with the quarks in the nucleons, *before* it meets the "nucleon as a whole" and translates it. For this reason, with energetics it is easier to alter and "flip" quarks in nucleons than it is to translate nucleons, since one is using the new energy propagation mode. *This simple reversal of the outside-to-inside energy propagation to an "inside-to-outside" mode, is a revolution in particle physics and effects on materials.*

In fact, the energy dramatically enters the "interior" areas of nuclei, nucleons, etc. in a manner heretofore impossible or extremely difficult to achieve by high-energy physics, which can only do it by smashing into the object with sheer brute kinetic energy. Therefore the new "inside to outside" energy propagation mode can produce—at feeble spatial energy—extraordinary nuclear interactions not possible by brute force "outside-to-inside" energy propagation means. By controlling the structuring of the arising of the energy at each point in spacetime, the energy structure interacting with the quarks and other interior components can be tailored to order. This is the primary secret of transmuting an element at a distance by a specialized energetics weapon. The structuring of the emerging internal energy can be deliberately tailored to change quarks in—say—some of the nucleons of U235 and plutonium nuclei. In this fashion nuclear warheads, nuclear powerplants, and nuclear propulsion systems of atomic submarines can be altered at a distance, beneath the earth's surface, beneath the ocean's surface, in space above the earth, etc.

---

[41] Oddly, had he simply waited a few more months or a year before injecting the missiles and nuclear warheads into Cuba, Khrushchev's strategic scalar interferometers would have become operational. With those in operation, a "Cuban missile crisis" might well have had a different outcome.

The EM "hash" emerging from the Russian interferometry simply jammed the electrical controls and systems of the Thresher *from inside its very electrical circuits.* Out of control and helpless, the Thresher sank to crush depth and imploded (Figure 6). The EM "splatter" from that interferometry, arising from the local spacetime throughout a surrounding region, engulfed the Skylark, the surface companion of the Thresher. The resulting spurious EM jamming severely interfered with multiple electronic systems on the Skylark. Several electronic systems just ceased working altogether. Their electrical forces were time-energy excited and time-reversed so that electron currents in them became impossible (blocked).

It required over an hour for the Skylark to transmit an emergency message back to the Navy that the sub was in difficulty, through all that interferometry jamming effects. The Russians undoubtedly "tracked" the sub till it imploded, then cut off the interferometry. As the "time-density charging" in the area slowly diminished, all failed electronic systems in the Skylark returned to full functioning. This was because the excess time-energy excitation just decayed away, so that electrical currents could again flow.

Khrushchev Demonstrated the Powerful Blast Capability of Strategic Interferometry

The next day, 100 miles north of Puerto Rico, the same KGB interferometer weapon was powerfully pulsed to produce a giant EM burst deep under the ocean's surface (Figure 42). A cone of water rose up from the surface to about a half-mile in height, observed by the startled crew of a passing U.S. jetliner and reported to the FBI and Coast Guard when the aircraft landed in Florida. As it rose, the cone of water spread into a cauliflower shape typical of a deep underwater nuclear burst that does not vent the gas bubble. Then that giant cauliflower of water fell back out of the sky and into the ocean. The second successful demonstration of Khrushchev's scalar interferometer weapon had been made for the benefit of the Presidium and the leaders of the Communist Party.

With those two dramatic tests, Khrushchev had killed the latest and greatest U.S atomic submarine and demonstrated the awesome power of energetics weapons. He had shown the leaders of the Communist Party that the U.S. had no inkling of his new superweapons. He had struck the U.S. a formidable blow in a direct act of war—but a new kind of *clandestine* energetics war—and gotten clean away with it. Now it was Kennedy, rather than Khrushchev, who had been duped. So had the entire U.S. Intelligence Community and the U.S. Scientific Community. Sadly, mostly they apparently remain duped or partially duped to this day.

Khrushchev managed to stay in power nearly two years longer by these two demonstrations. So the KGB/Soviets embarked on a decades-long program of buildup and development of the new weaponry, clandestinely, in order to take the U.S. by surprise attack and destroy it utterly.

In 1972 Brezhnev named 1985 as the target year when Russia would be able to do as it pleased, and would essentially control the land, the oceans, and space {126}. This was stated at a secret meeting of the European Communist parties in Europe. A British intelligence agent was in the meeting, and furnished the U.S. that information. Later it was openly released and published.

The Russians met that 1985 deadline, and the resulting increase in killing U.S. missiles and aircraft in 1985-1986 was most indicative. Other indicators such as massive weather engineering were already ongoing.

Weather Engineering Over North America

The first KGB weather engineering tests over the U.S., using their relatively new interferometers, produced signatures of anomalous perfectly round holes appearing in clouds. These experiments started in 1967 or thereabouts. The Russians gave us that very severe "deep freeze" winter of 1967, as an initial weather-engineering test of their energetics interferometry weapons.

The KGB communists do have a sense of historical dates, and they often do things on certain dates in symbolic fashion. Full-time Soviet weather engineering over North America started in earnest on July 4th, 1976—our Independence Day and our bicentennial. This was the KGB's "gift" to the United States on such a historic occasion. Full-bore weather engineering over the Americas has continued to this day, and even increased. It was passed into the hands of the Yakuza and the Aum Shinrikyo upon their leasing of those interferometers on site in Russia at the end of 1989. In 1985 I produced a rather crude videotape[42] (40 minutes) with pictures of the giant radial cloud signatures etc. associated with such weather engineering operations since 1976.

The *basic* weather engineering method is simple. Use an LW interferometer (LWI) to reach through the earth and *cool* the air in a region and produce a high-pressure area, due to the densification of the air. Use another interferometer to *heat an*other area and form a low-pressure area, due to the expansion of the air. Then carefully steer those *highs* and *lows* by gradually and slowing rotating the distant antennae and slowly adjusting the interferometry range simultaneously, according to a calculated schedule.

---

[42] T. E. Bearden, "Soviet Weather Engineering Over North America," 1985.

By that means, the LWIs easily catch, entrain, and steer the jetstreams largely responsible for our weather. Put sharp loops and twists in that steering, and the huge angular momentum of the jetstreams developed in those sharp turns and circles will spawn tornadoes, violent weather, etc.

The first complete weather engineering system over North America in 1976 also used the Woodpecker's normal EM carrier signals with dimensioning and internal assemblage of the composite longitudinal EM wavepairs. A giant "interference grid" was established over much of North America. A great deal of local adjusting of the interferometry in the grid was necessary. These local grid readjustments caused "sudden popouts" of EM energy, creating many atmospheric booms, rumbles, and explosions.

To really stir up big weather trouble, the Russians—and now the Yakuza and Aum Shinrikyo—also heat or cool areas of the ocean where El Nino and similar "water engines" form. A dimensioned LW interferometer is required for this operation. By slowly and protractedly heating or cooling the water in one of the "warm water engines" fueling the weather, one causes a very significant, long-term, later effect on perhaps an entire continent.

One can create great droughts in some areas, great floods in others, etc. The KGB has been doing this extensively, since 1976. The great floods in Red China a few years ago are a direct example of KGB/Yakuza weather engineering, as is much of the extreme heat of 1998. In fact, thanks to the weather operations of the Yakuza under their KGB mentors, 1998 was the overall hottest year on record till then, since records were kept. At least for the summer of 2002, the record may have been broken again, with a significant season of giant forest fires burning away.

One purpose of the excess heating of the atmosphere, of course, is to do damage. An added "benefit" to the KGB is to furnish support to the environmentalists who are concerned about global warming. Understand, there is indeed such a thing as global warming. But by augmenting the atmospheric (and ice sheet) heating just a bit in a single year, the statistical analyses are skewed. The skewed analyses will show that the global warming problem is far more imminent than it really is, if the LWI engineering of it were suspended. Nobody on either side of the global warming debate seems aware of the LWI intervention that is ruining the results of their models and calculations.

In turn, the heightened global warming indications, detected as a result of the LWI interruptions, impel strong activist support for such things as the Kyoto treaty, *which is specifically designed to the strategic detriment of the U.S.* if and when the U.S. Senate ever approves it. The Kyoto treaty does not even apply to some 160 countries including Russia, China, etc. This focuses a certain proportion of our government time and effort in grappling with an inflated global warming problem, immediately. It directly fits the KGB strategy of "spreading" our government energy and attention all over the map, anywhere but upon the coming strategic energetics strike and upon energetics weaponry.

The die-hard section of the KGB strategists are chess players. *They will come at you anyway, anyhow, and anywhen that they can.* Everything that can be impressed to serve or support the overall objective, will be impressed and used. **That** is the overriding principle when analyzing Russian KGB strategy.

Khrushchev meant every word of it when he said to Nixon in their famous Kitchen Debate: *"We will bury you!"* Interestingly, to this day Americans and **their** government officials have continued to be deceived as to exactly what Khrushchev meant. He meant it literally.

Earthquakes to Order

According to U.S. Defense Secretary Cohen, electromagnetic induction of earthquakes is now recognized by the U.S. Department of Defense, considering **his** little-noticed statement in Georgia in 1997, which we previously quoted. An interesting exchange of induced earthquakes (Figure 54) had occurred in latter 1988 and early 1989. Here two sides seemed to be "counting rifles."

Opening bet (induced quakes)
- Quebec 25 Nov 88
- Pasadena 3 Dec 88

Raise you
- Armenia 7 Dec 88

Call** - **what do you have?
- Malibu 8 Jan 89

Four aces
- Soviet Central Asia 23 Jan 89

***Clear** table (stress relief)*
- Bakersfield CA et al. various

Figure 54 Counting rifles with earthquakes: Put up or shut up.

As Secretary Cohen made his April 1997 statement obliquely recognizing the scalar interferometer weapons, the little friendly nation had just completed countering and averting a massive strategic energetics strike upon the United States and Western Europe, scheduled for May 1, 1997 —just three days later. A little more than two months earlier, a previously scheduled attack had also been countered and averted by the same friendly little nation.

Once one can reach into or through the earth and ocean and create energy in a desired interference zone, induction of earthquakes is simple. Use the interferometers to deposit energy in a fault zone over a period of time. Keep it up, and the piezoelectric plates build up enough mechanical stress to forcibly "slip" and produce an earthquake.

Put the energy in very slowly, and one gets a long "overpotential" effect where the rocks do not slip for a long time, while the energy builds to higher levels due to the static friction of the rocks. Then there occurs a rather rapid catastrophic slip. In short, one produces a large destructive earthquake. The KGB weaponry can produce earthquakes of the highest magnitude if they work at it.

If one wishes to induce a quake in an area where there are no known faults, then one simply focuses the interference zone there inside the solid rocks and proceeds to deposit the energy, preferably rather quickly (in hours instead of days). The entire rock will be expanded and it will fracture, producing an anomalous so-called "flat plate" earthquake.

Around the early 1900s, the "normal" rate of large earthquakes was two or three per year. That gradually increased after WW II. During one week in year 2000, there were some 20 major earthquakes around the globe. Obviously, someone is doing it, since this is far beyond what could be expected naturally. Presently that "someone" is mostly the Yakuza and Aum Shinrikyo, on site in Russia and leasing the LWIs. With the scheduled CSR attacks forcibly aborted, the KGB has the problem of restraining an increasingly impatient Yakuza and Aum Shinrikyo group, while keeping them paying the lucrative lease money. Consequently, inducing some big earthquakes is one way to "keep the troops occupied and busy." Shooting down the occasional aircraft or missile is another way.

I worked for 8 years or more with an inventor (Floyd Sweet, now deceased) who had an energy device {127,128} sensitive to *artificially* induced quakes of that fashion, even such quakes thousands of miles away. The device would sit right on top of a *normal* quake and not react. We could tell within some 8-10 thousand miles—and often even for greater distances— which quakes were artificial and which were natural.

So during that period I was able to clearly ascertain what was going on, and who was doing what to whom, and which quakes were normal and which were manmade. Several artificial U.S. quakes, the Armenia quake, and a subsequent big quake in Russia were particularly interesting in the mid and latter 1980s. The huge 1976 China quake was directly induced by the KGB.

Figure 55 Brewer briefs President Reagan on Soviet scalar EM weapons October 1984, then briefs National Security Council

In 1984, on my behalf Roy Brewer {129} had briefed President Reagan on these scalar interferometry weapons (Figure 55), then briefed the National Security Council. Let us hope that these high level briefings may have borne at least some fruit.

Readying the Weapons for the 1985 Readiness Demonstration Tests

In latter April 1985 I received an urgent call from my close colleague Frank Golden. In his avionics repair work, he suddenly found that his best calibration instruments were very "noisy." He replaced one expensive instrument, and the second was "noisy," also. A third proved equally noisy. This was unparalleled. It was not the instruments, but instead it was all the electronics to be repaired, that were emitting electromagnetic noise!

So Frank borrowed one expensive instrument and brought it home for the weekend. He altered it by a proprietary method so that he could detect energetics radiation. Voila! The entire earth was alive with Soviet energetics signals. The weak interferometry of so many signals impinging into the weakly interacting electronic systems to be repaired, had generated the electronic noise his instruments had detected.

In response to Frank's call, I immediately proceeded to his location. I personally observed on the oscilloscope, as detected by Golden, some 27 pairs of Russian energetics signals, each of the two in a pair being 12 kilohertz apart. (We have previously explained why two frequencies are used, and one "pretends" one transmitted the difference frequency). Each of these frequency pairs represented the entire earth held in entrained resonance. Each pair was an "energy tap" in the earth, extracting enormous energy from the heat of the molten core of the earth itself and outputting it as 12 kilohertz power for transmission to giant interferometry weapon sites.

Each "tap" was extracting energy sufficient to power perhaps 4 to 6 of the giant intercontinental energetics interferometer weapons. We were observing something like 100 to 150 massive weapons on line and radiating. A full-up strategic dress rehearsal—or preparations for a full-up strategic attack on the U.S.—was underway.

Gorbachev was still a "new man on the job" at that time. Further, 1985 was the 40th anniversary of the end of WWII, and a very important occasion for the Communists and in fact for all Russians. The May Day celebrations that year went on for several days, not just on May 1.

What we were seeing were preparations for a massive May Day weaponry demonstration in and around May 1, 1985. We were observing the preparation and demonstrated readiness of all that great armada of Soviet energetics weapons, to meet Brezhnev's 1985 schedule laid down back in 1972 {126}.

Longitudinal EM wave communications to deep underwater Soviet subs, the huge interferometer armada, and everything else—all were being readied and demonstrated. I can assure the reader that observing full rehearsal of preparations for the next World War was a most sobering experience.

After May 1, most of the armada of interferometers was stood down again because it had been demonstrated that the "ready schedule" had been met. All but one of the giant energy taps in the Earth were also stood down again. That one remained active.

The May Day report to the heads of the Communist Party that year must have really been something!

As usual, apparently our own scientific and intelligence communities did not detect the main activities, and would not have believed it if one told them. They certainly did not believe it when I published it in 1986 in *Fer-de-Lance,* but labeled it sheer fantasy or at best total conjecture.

It was neither fantasy nor conjecture. It was direct experimental measurement. It was in fact what used to be called *scientific method.* The problem with our intelligence agencies was that they operated behind *policies.* Once the *policy* is decided, then regardless of what the intelligence analysis indicates, anything against the official *policy* is just ignored. This neat little *policy* method of doing business has resulted in making our intelligence agencies some of the most political-minded and manipulative groups on Earth.

Meeting the 1985 Schedule for Readiness

In 1985, KGB/Russian preparations began in earnest for the energetics weapons readiness projected for that year by Brezhnev in 1972. Being full-bore tests, these actually went after destruction of vehicles and killing people. These were actual acts of war. The so-called "Cold War" was not nearly so cold as pictured in the news media!

Early that year, an LWI was used to cool the combustion in the engines of a jet liner off the coast of San Francisco, bound for Los Angeles (Figure 56). This flamed-out the engines. The stricken craft fell 30,000 feet before the time-density "cooling" charge drained away, sufficient for the pilot to restart his engines and limp into San Francisco. Gravitational and inertial anomalies—*a clear signature of the use of this EM/GR weapon*—were exhibited upon the personnel in the cockpit. The official investigators ignored those signatures; after all, such is not in their standard checksheets! And everybody knows the standard checklists are already perfect!

SIX MILE DROP IN TWO MINUTES

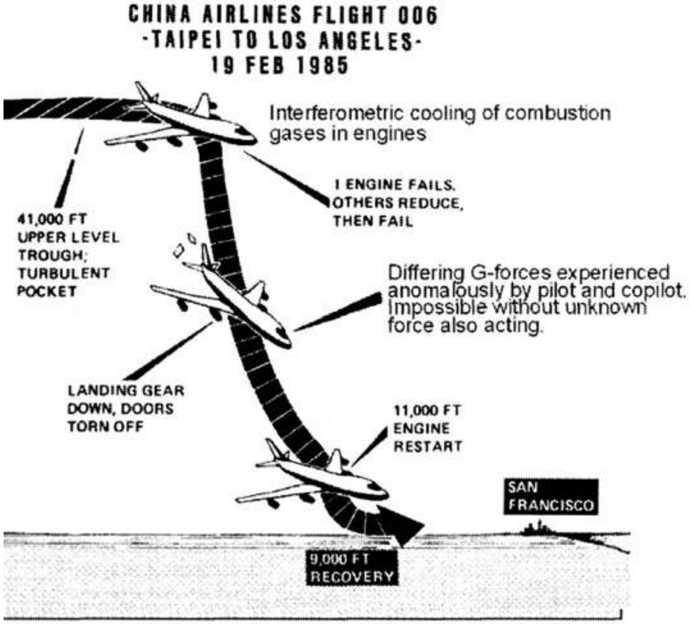
CHINA AIRLINES FLIGHT 006
-TAIPEI TO LOS ANGELES-
19 FEB 1985
Interferometric cooling of combustion gases in engines
I ENGINE FAILS. OTHERS REDUCE, THEN FAIL
41,000 FT UPPER LEVEL TROUGH; TURBULENT POCKET
Differing G-forces experienced anomalously by pilot and copilot. Impossible without unknown force also acting.
LANDING GEAR DOWN, DOORS TORN OFF
11,000 FT ENGINE RESTART
SAN FRANCISCO
9,000 FT RECOVERY

Figure 56 Extinguishing combustion in the engines of a jetliner 19 Feb. 1985.

KGB Crews Regularly Practiced Against U.S. Rocket Launches from Cape Canaveral

In a late Nov. 1985 night shuttle launch at Cape Canaveral, the Russian interferometer crews continued their practice of using the U.S. shuttle launches as convenient "simulated" ICBM launches, for energetics interferometer crew practice (Figures 15 and 16). EM missiles (balls of EM energy formed in the area by distant LW interferometry through the earth) were used as the "destructive agent", but were offset from the actual launch in order not to destroy the shuttle. This was *normal crew drill,* not combat operations. The practice range for energetics weapons is—as might be expected—the actual targeted area and targeted nation itself.

As shown in *Fer-de-Lance,* I published an actual picture (Figure 15) of such a KGB EM missile strike, offset from that shuttle launch in late Nov. 1985 and snapped by Bob Gladwin. I also published an actual picture of a second such EM ball of glowing energy (Figure 16), over the launch site for the same launch, and used as a registration point (the LWs deviate slightly in passage through the earth, since the potentials in the earth are not constant). George Suchary took the second photo.

A standard artillery "shift-from-known registration point" method of firing was being used. In other words, the crews establish LW track on that EM registration ball, right through the earth. That represents *a point having a known set of coordinates*—a registration point—established in the firing computer. Then when the targeted strategic missile (in this case the shuttle's main booster) ignites, the interferometer also tracks the huge ionization target of the exhaust with the same weapon system.

That gives a known *relative* location of the missile rear end, with respect to one's registration settings for the registration point ball of light. Then add some meters to move up onto the missile body proper, and voila! The center of the target missile that will be fired is now in the bulls-eye, so to speak. That is a standard "shift from a known registration point" method of shooting, well known to every artillery group on the planet. The "marker beacon" is what is known as a "high burst registration," except it is not a burst but a steady ball of energy maintained over the area as the registration point.

The distant rocket or missile can now be destroyed in a variety of ways:

(1) Just create a sudden EM surge arising right in that spacetime (that is, just pulse the interferometer). That burns out some of the inner electronics of the missile, causing partial or complete missile failure.

(2) Sharply pulse the interferometer to introduce EM energy surges and sparks in the fuel or solid propulsion, igniting it massively and blowing up the rocket. Works well for aircraft fuel tanks as well.

(3) Form another EM missile and hit the missile broadside with that, burning

through it and destroying it immediately.

(4) Hit it with an explosive EM missile internally, instantly exploding the missile into pieces. Or

(5) Hit it with several of the above simultaneously.

Suchary and Gladwin had captured the direct evidence of KGB dress rehearsal for the strike and kill of the Arrow DC-8 in Gander, Newfoundland about two weeks later.

Kill of the Arrow DC: Practice Turned into Performance

Two weeks later, the same KGB weapon tested against the Cape Canaveral shuttle launch was used to destroy the Arrow DC-8 at Gander, Newfoundland, killing some 256 occupants, 248 of them being U.S. soldiers returning from peace-keeping operations overseas. The aircraft *was not* iced up; later that was clearly established.

See Figure 17. As the ill-fated Arrow aircraft began its thrust down the runway for takeoff, it was already under LW interferometer track (the exhausts of the engines were being precisely tracked) by distant KGB operators. The KGB operators simply began to cool the combustion gases in the engine combustion chambers as had been done to a U.S. jetliner earlier that same year (Figure 56) off the coast of San Francisco (there they had completely flamed out the engines). Now they did not want to flame out the engines during take-off, but just wanted to reduce the thrust of those engines.

So the available thrust of the engines began decreasing as the aircraft picked up speed on the runway toward takeoff, even though the turbines were rotating at full speed. For powered flight, it is the *thrust-per-drag ratio* that is important. One can reduce that ratio by ice-loading and therefore increasing the overall aircraft drag, as all Transportation Safety Board checksheets assume if the engines are at full rotation and if operation and icing conditions are present. Or one can reduce the ratio by reducing the engine thrust itself, as was done for the Arrow, even though the plane is de-iced. All the Transportation Safety Board checksheets *assume* that it has to be drag increase, and thus due to icing.

The National Transportation Safety Board (NTSB) investigators know nothing about longitudinal EM wave interferometry and its capabilities—and its development by several nations of the world. Neither do their checksheets. Sadlly, the NTSB is not even aware that the electrodynamics and electrical engineering it assumes is so perfect is in fact a decrepit old model based on century-old errors. They are unaware that much better electrodynamics models have long been developed for particle physics, because the crusty old classical model is totally inadequate. Their checksheets also are "unaware" of anything really modern such as electrodynamics models of the higher group symmetry kind.

The thrust reduction by combustion gas cooling *does not appear on the official checksheets and is never considered.* It is not considered to this day. It will probably still not be considered this century.

Observers on the ground heard those engines "laboring," which can be a direct symptom of combustion gas cooling and engine thrust reduction. Rough running and missing can be another, if the cooling increases. Engine flameout occurs when the combustion no longer can be sustained because of the dramatically lowered temperature of the gases.

The Arrow Pilot Sensed the Lack of Power, But the Aircraft Was Immediately Hit

See Figure 17. At rotation and liftoff, the pilot of the Arrow DC-8 immediately sensed his lack of lift, and started a slow turn to the right, apparently intending to come around and land again as an emergency measure. Meanwhile, crossed beams had formed in the sky above and off to the right side. From those crossed beams of light, there shot down a bright "streak" of light that struck the fuselage of the aircraft in the right side, ahead of the engines. That was the EM missile strike—just as had been offset from the shuttle launch two weeks earlier in Florida, and photographed by Bob Gladwin.

That piece of Arrow DC-8 fuselage hit by the EM missile, with the burned hole in it, was recovered and tested. Tests showed no explosive residues on the edges of the hole, so it *was not* made by a bomb or other explosive device. Instead, it was made by an EM missile of the type that over a decade later hit the TWA-800 several times off Long Island and killed it.

In the Arrow aircraft's kill, the fiery EM missile sharply ignited the plastics in the forward cabin, which explosively outgassed. One product of such outgassing is hydrogen cyanide. A whiff or two, and all the passengers in that section were already dying of cyanide poisoning. *In the autopsies of the recovered bodies, about half the passengers were already dead of cyanide poisoning before the plane impacted the ground and exploded*. They died while that aircraft was still flying and sinking tail-first toward the ground.

The results of those autopsies were deliberately withheld from the Canadian investigative board.

The impact of the EM missile on the Arrow DC-8 was clearly recorded by sharp, vertical EM surges (spikes) in traces on both data recorders on board the aircraft. In *Gravitobiology,* on p. 82 I printed the recorded sensor responses **for** that signature also.

Elements of the U.S. Government Deliberately Suppressed Vital Evidence

Several eyewitnesses, including Judy Parsons, physically observed the formation of the crossed beams in the sky. Parsons and perhaps others actually saw the streak-down and strike of the EM missile when it hit the doomed Arrow DC-8. Parsons tried to testify, but her testimony was never allowed to be heard by the Canadian Transportation Safety Board. A major U.S. News agency interviewed her, but then squelched the story for some "unknown reason." Other witnesses (who are still reluctant to come forward) did observe the strike of the EM missile also.

Why would the U.S. government deliberately suppress information relating to the cause of the deaths of 248 of our soldiers returning from peacekeeping missions overseas? Why were they so determined that this direct act of war be declared an "accident" in the face of overwhelming evidence that it was not an accident at all?

Iran Contra was not yet divulged at the time. Recall, at very high level our own fellows had directly violated the law, and were engaged in the illegal Iran Contra fiasco up to their eyeballs. Also, some officials in Washington in certain high U. S. Government circles were in near panic, in their personal (erroneous) firm belief that a MidEast-sponsored terrorist bomb had destroyed the Arrow. In that case, so they reasoned, with the murder of 248 of our soldiers, the American public would be so aroused and angry that they would force military action against the MidEast nation or group, and that would then reveal Iran Contra. *Too many highly placed people's hands were dirty!*

So extreme spin control was the major requirement. Damn how many American soilders had just been killed in an overt act of war against the U.S. Damn the implications of the use of eerie new weapons we did not possess. A hue and cry from the U.S. public would almost certainly reveal Iran Contra and smoke out the duly secrets. That must not happen at any cost. Some high officials—particularly the ones who engage in such illegal acts, in assassinations, murders, etc.—are far more interested in covering up their nefarious deeds than they are interested in any kind of justice or morality. Consequently, the fix was in.

So *U.S. officials placed enormous pressure on Canada to produce a finding of icing.* Lots of other skullduggery occurred, which probably constituted criminal activity by some U.S. officials. A high-ranking U.S. military officer arrived on site quickly, apparently under orders to erase that scene as fast as possible. He almost immediately wanted the Canadians to bulldoze over the site, even before recovery operations were complete and while the investigation was just getting underway. The site was bulldozed over in unseemly haste not long thereafter. That would appear to possibly constitute deliberate tampering with the evidence at an official site of investigation, on the part of whoever ordered the General to bulldoze the site precipitously.

In addition, witnesses such as Parsons were never allowed to testify. The Canadian Board waited until the half of its members who strongly opposed any ad hoc finding were absent, and hastily approved a finding of icing as the cause of the "accident." Later the disgruntled Board members who were absent when the "voting" occurred, issued their own minority report. Unfortunately they knew nothing of scalar interferometry and energetics weapons, and so they strongly believed it had to be sabotage and explosives.

Anyway, the blackguards hiding Iran Contra buried the incident, and the savage kill of 248 of our soldiers and 8 civilians went unpunished. Anyone such as the present author who pointed out what had really happened and why, was just branded a lunatic. After all, the blackguards who do not flinch at murder and unapproved assassination do not blink an eye at slander and libel.

The sad part is those fine young soldiers and the 8 civilians who were killed, and the grief and lasting changes made to their families. It is the children who had to grow up without their fathers, the mothers whose husbands were suddenly dead, etc. They deserved much better from their government than they got.

Other Direct Energetics Signatures on the Arrow DC-8

Meanwhile, in the lifting DC-8 the strike of the EM missile had also *time-density charged* the forward cabin materials. Those materials slowly decayed from their excited states via longitudinal EM wave emission. When a human body is subjected to fairly significant LW waves (accompanied with their time-density waves), it gets its internal immune system and cellular regeneration system "jammed," since these systems use LWs and accompanying time-density waves. The cells and tissues are also "time-energy charged" as well, and so spurious LWs continued to be slowly emitted in their tissues and cells locally, leading to gradual, general debilitation. After a time delay for the effects of this time-charge decay to be exhibited, such exposed persons will get sick with odd symptoms of various diseases such as influenza.

The ground handlers on the scene and exposed to the struck materials of the aircraft, were also exposed to longitudinal EM wave radiation from those materials which had been instantly "time-charged" at the powerful strike. Hence they were exposed to fairly significant LW waves, and should therefore have shown later symptomology.

Some 60 or so of the ground handlers actually retrieving the crash materials later sickened with all sorts of mysterious maladies: dizziness, stomach cramps, flu-like symptoms, weakness, nausea, fatigue, etc.—precisely what would be expected if some the materials they were handling were emitting longitudinal EM radiation, exposing them to significant LW radiation. That is a clear signature again of a strike by an EM missile, and the exposure of those retrievers to LW emissions. *In no normal way could 60 of those ground crewmen handling the wreckage debris all come down simultaneously with the exact symptoms of*

LW radiation sickness, unless that was what it actually was.

Many of those sickened persons remained debilitated for years. Probably a few are still alive and debilitated.

There are other signatures on the Arrow DC-8 kill. I did a small bit of work on the problem in the interests of Arrow airlines, but of course the conventional investigators they also hired could only conceive of an ordinary explosive. Nevertheless, I did tell them as best I could what happened to their aircraft, their crewmembers, and our own valiant soldiers.

Our troops on board the tragic Arrow DC-8 *did not* die from a simple over-iced plane that crashed as a result of carelessness in deicing operations. *They died as the result of an act of war against the U.S., carried out by the KGB as a test in an entire series to demonstrate the 1985 readiness that Brezhnev had specified in 1972.* That war is still partially ongoing and undeclared to this very day, with other players such as the Yakuza and Aum Shinrikyo also now in the game—and possibly even the Chinese, who certainly have had the weapons for some time, and have had QP weapons and negative energy EMP weapons since earlier this year (2002).

With the kill of the Arrow DC-8, the KGB made its deadline that had been set down 13 years earlier by Brezhnev. To meet the 1985 schedule, something significant had to be killed successfully and clandestinely, no later than Dec. 31, 1985. The KGB made the schedule on December 12, 1985, by killing the Arrow DC-8 after a warm-up and full dress rehearsal against a U.S. night shuttle launch at Cape Canaveral about two weeks earlier. And they also demonstrated just how ignorant of such energetics weapons and their usage the U.S. government, scientific community, and investigative boards really were.

We were fortunate they did not just go ahead and kill that same shuttle in late November 1985. Instead, they waited and killed the Challenger a few months later.

Other Acts of War Completed the Milestone

Apparently the final demonstration of milestone completion called for the destruction of an aircraft, kill of an ICBM, kill of a space shuttle, induction of earthquakes, and demonstration of the new quantum potential weapon still in the advanced laboratory stage. The KGB completed much of the milestone demonstration in 1985. The remainder they completed in 1986.

In January 1986 they killed the space shuttle Challenger, and the very next night the KGB even gave a big party in Moscow, to celebrate the "perfect success of their active measures" against the shuttle. General Daniel O'Connor, former head of U.S. Army intelligence, personally issued an alert to many high-ranking U.S. government officials, pointing out the almost certain involvement of the KGB and the fact that the loss of the shuttle was almost certainly due to a hostile act by the Soviets.

In April 1986 (May Day report time was coming!) the KGB killed a Titan ICBM fired from Vandenberg Air Force Base. They also induced several large earthquakes for practice, and then began the induction of the long-dreaded "giant California quake" in the San Andreas Fault zone.

This latter induction of what would have been a massive quake in both San Francisco and Los Angeles, and all along the San Andreas Fault, was countered (destroyed) by a civilian group known to this author, with a backlash that resulted in the Chernobyl incident.

In April 1986 (for the same May Day report) the KGB also demonstrated the effectiveness of their newly emerging quantum potential weapon against the convenient U.S. air attack on Libya (Figure 25).

Kill of the Titan Missile Over Vandenberg Air Force Base

In 1986, the KGB began gingerly using those energetics interferometry weapons to kill other U.S. aircraft and several U.S. missiles, including killing a Titan 34-D fired from Vandenberg AFB on 18 April 1986. The Titan was killed 9 seconds after launch, partially crippling the U.S. space surveillance program. That kill was part of the demonstration of readiness milestone completion.

There was a completely decisive signature on the kill of the Titan. Simply check *Aviation Week & Space Technology,* 124 (17), Apr. 28, 1986, p. 18, where a picture (the one taken from longer range) of the Titan explosion is printed. Look about an inch above that explosion, which is already well developed. There above the Titan explosion sits the KGB ball of light registration point, just as it was sitting over the night shuttle launch in Cape Canaveral in latter Nov. 1985 when videotaped by a private photographer.

I persuaded NBC engineer Ron Cole and his colleagues to analyze a copy of that video tape frame by frame. In 17 frames, that little ball of energy above the explosion moved off on its own course, independent of the Titan explosion. Since the launch area for such missile launches is controlled and there were no aircraft over the explosion, then one again has a very clear signature. The KGB killed the Titan, with 100% certainty.

That of course was a test of the interferometry weapons in the role of *Launch Phase ABM system*. Tested against the Arrow, the KGB operators were emulating test of the system in a *Launch Phase Antibomber role*. These were vital *strategic* weapon tests. I published details of the use of the weapons in many other roles.

Dead-Man Fuzing and Outside Assistance Deterred a Full Strategic Attack in 1986

We have previously documented and published why the U.S. was not fully attacked and destroyed by the Soviets in 1986. It was because of our nuclear facilities, warheads, etc. which would have exploded, creating great fallout hazards worldwide, and because of nuclear dead-man fuzing weapons clandestinely placed in Russia by one or more other nations.

A friendly small nation possessing energetics weapons also intervened directly in l986, and blew up several Russian missile ammunition storage sites, some five or so in six months. In one such explosion, about one third of all the missiles in the Russian Northern Fleet were destroyed. As we stated, the so-called "Cold War" was not nearly so cold as has been represented.

In short, the friendly little nation served notice on the Soviets that, if they moved against the West, the Soviet Union itself would suffer terribly. They put it in language clearly understood by the KGB. So any full-up strategic attack for that period was stalemated by that combination of factors.

A major concern for the Soviets was their inability to counter the dead-man fuzing problem posed by nuclear weapons clandestinely hidden inside Russia and by the nuclear warheads and reactors inside the targeted nation(s). They already had the quantum potential (QP) answer *theoretically,* and had conducted a demonstration test of a laboratory QP device, but they had not yet completed the *QP* weapon systems and deployed them.

Nonetheless, a very powerful strike of peculiar nature—induction of a giant earhquake in the San Andreas Fault zone, including in both Los Angeles and San Francisco—was scheduled. This strike would have caused billions of do11ars worth of damage and would have killed perhaps 200,000 Americans. The initiation of the strike was actually begun in April 1986—as the reader can begin to see, there was to be one helluva May Day report in 1986! Incredibly, a private group here in the U.S. then made a unique and unprecedented contribution that saved the day for Los Angeles and San Francisco.

What Really Happened in the Chernobyl Disaster

Using the "inside" of the Woodpecker beams as superhighways for their longitudinal waves and interferometry, in April 1986 the KGB interferometers had started their build-up of substantial induced EM energy in the sides of the San Andreas Fault that runs through Los Angeles and up through San Francisco. Two beams were being used by the Soviets: One was coming down on the target from the North—coming over the North Pole regions and the ice caps—and one was coming around from the East and upon the target area. Completely (100%) certain signatures were present, including some ELF "chugger" waves symptomatic of the Soviet method of inducing very large quakes. The time-energy is being used, hence very low frequency ELF is employed to maximize the amount of time-energy available for transduction in the rocks on the side of the fault. The Soviets were building up a very large earthquake—indeed, to spawn a whole series of such quakes all along the San Andreas fault—with severe aftershocks following after the main quake had done terrible damage. In short, they were preparing to hit California with the "big one" that everybody had been anticipating and dreading for so long.

A private group contacted me. They stated they could take out one of the transmitters by producing a powerful (giant) phase conjugate replica signal pulse upon it that would backtrack right into the transmitter and catastrophically burn it out—a surface application of the "phase conjugate shooting" once considered for armed satellites that could shoot down rising hostile ICBMs at 10,000 miles distance. As I understood it, they had a new kind of essentially noise-free regenerative amplifier, used in a transmitter. This meant that *extreme amplification* could be achieved in a pulse response to a received Woodpecker signal, before the transmitter burned out.[43] Since the group had no knowledge of the exact KGB transmitter locations or the possible consequences of such an action, they asked my opinion as to the risks associated. This placed me in a real quandary.

At the time, the U.S. was literally under siege. We have mentioned some of the

---

[43] A very subtle but important point is this: An energy flow of any amount—say, 10^{20} joules per second—has absolutely *zero power* if none of that energy flow is diverted or changed in form. So if one has a truly perfect noise-free amplifier, and a pure noise-free energy signal input, the signal energy input can be directly and freely amplified without limit, so long as those noise-free conditions can be upheld. In short, that rigorously constitutes *regauging,* and the widely used **and** well-known axiom of gauge freedom—the ability to change the local energy density of a system or spacetime freely and at will—applies. Relativistically, such re-gauging is merely rotating the frame of the system away from the laboratory frame. This group indeed had such an amplifier, which I physically saw, and apparently has it today. Our present texts have not yet realized the real implications of gauge freedom: It means that one can freely amplify energy without limit, once certain conditions (no divergence of the energy or change of its form) are met. This does not violate energy conservation; the "potential energy of a system" is not a fixed number at all, but depends on the frame of the observer vis a vis the frame of the system. We point out that there really are no true conservation laws in general relativity unless introduced by arbitrary measured The great Hilbert already knew this in 1917, two years after the advent of Einstein's general relativity. Hilbert stated: "*/ assert... that for the general theory of relativity, i.e., in the case of general invariance of the Hamiltonian function, energy equations... corresponding to the* ***energy*** *equations in orthogonally invariant theories do not exist at all. I could even take this circumstance as the characteristic feature of the general theory of relativity.*" [D. Hilbert, *Gottingen Nachrichten,* Vol. 4, 1917, p. 21]. Loskutov and Logunov commented as follows: *"In formulating the equivalence principle, Einstein actually abandoned the idea of the gravitational field as a Faraday-Maxwell field, and this is reflected in the pseudotensorial characterization of the gravitational field that he introduced. Hilbert was the first to draw attention to the consequences of this. ... Unfortunately, ... Hilbert was evidently not understood by his contemporaries, since neither Einstein himself nor other physicists recognized the fact that in general relativity conservation laws for energy, momentum, and angular momentum are in principle impossible."* ["Nonuniqueness of the predictions of the general theory of relativity," *Sov. J. Part. Nucl,* 18(3), May-June 1987, p 179).

"target year" activities, such as kill of the Arrow DC-8, flameout of jet engines early in 1985, target practice against shuttle launches at Cape Canaveral in November 1985, etc. Actually kill of targets had continued in 1986. A huge earthquake followed by several large aftershocks had been used against Mexico City on 19 September 1985, causing great damage and demonstrating a little of what might be expected when they chose to hit Los Angeles and San Francisco much harder. The Mexico City quake had been preceded by an induced, giant hailstorm weeks earlier, and on the evening of 18 September 1985 by giant radial cloud signatures accompanying the generation of excess energy in the underlying rocks as the KGB prepared next day's giant Mexico City quake. The reader should recall that, working with Sweet during this period, I knew with certainty which earthquakes were natural and which were manmade.

U.S. missiles and aircraft had been hit, the Challenger had been destroyed, and the Titan 34-D had been destroyed at Vandenberg AFB on 18 April 1986, accompanied by a 100% indicator (the substantiated presence of the marker beacon for registration and shift firing). In addition, a little Richter 5.3 quake in the San Francisco Bay area had been induced by the Soviets on 31 March 1986, a quake which was correlated with preparatory activity on the Woodpecker transmitter signals the day before, showing Soviet induction of the energy to create the quake. Again, we had instrumental evidence that it was manmade.

That quake on 31 March had been a "warm-up" and practice for the big one now about to be induced in both Los Angeles and San Francisco, and along the San Andreas Fault. In fact, it appeared that the entire fault zone would be stimulated by deposit of excess energy all along its length. In that case, induction of a distrituted California quake of unprecedented extent and great magnitude was now in process, though just beginning. If that event or those events occurred, the loss of American lives and property would be staggering.

The High Side Risk Factor

The group's query to me presented a very difficult decision situation. Such huge Soviet interferometer beams, once built up, contain *enormous* longitudinal EM wave energy. If a Soviet transmitter were near large nuclear facilities, and were instantly blown and shorted, all that potential energy might flash directly into the ground.

In case of such a catastrophic transmitter failure, an enormously powerful longitudinal wave pulse would spread out in all directions in the surrounding earth. The first nuclear material the pulse encountered would be fissioned abruptly and very violently. If the transmitter were within a few kilometers of one or more of the large "monster" Soviet ICBM sites, the full-up explosion of several 30 or 40 megaton warheads might result—in the dirt and dirty. The resulting fallout and contamination would be massive, would spread round the world, and would eventually kill millions for years to come. Nuclear reactor stations were a similar risk; instant fission of all the fissionable material in one of them would constitute an unprecedented and giant nuclear explosion, in the dirt and dirty. That was the high side risk.

The Low Side Risk Factor

The low side risk would be that little or nothing at all would happen; perhaps a tremor induced in a nearby fault zone, etc. I knew that the Soviets—after an early Kyshtym transmitter failure—had produced such a flashover on a much smaller scale and that flashover had caused the nuclear explosion of nearby stored nuclear wastes. The resulting fallout contaminated some 1200 square kilometers, which is still contaminated to this day. I also knew that, after Kyshtym, the Soviets had installed safety devices on all their scalar interferometer transmitters in case of just such failures.

So if the safety devices held when the transmitter suddenly shorted out, they could hold the longitudinal EM wave beams and potentials long enough to drain the energy slowly and harmlessly into the earth over an extended period. That was the low side risk.

The *most likely* consequence, of course, would be somewhere in the middle, between the extreme upper risk and lower risk limits. In other words, we could expect some bad repercussions in the Soviet Union almost certainly, but hopefully it would mostly be confined to that spot.

Consequences of Not Taking Counterfire Action

If no action were taken, the risk to the U.S. was appalling. If the transmitters continued, then perhaps 200,000 Americans were going to die in Los Angeles and San Francisco and along the San Andreas Fault. Perhaps 300,000 more Americans would be injured, and many of them would also die. Enormous physical damage would also be done, and those basins would be disaster areas for some years to come. It would be interpreted by our inadequate scientific community as a great "natural" disaster caused by a giant earthquake, and the ghost of Khrushchev would have had one more laugh at us, with his words "We will bury you!" silently ringing in the ether.

My Recommendation to the Counterfire Group

To save those American lives, some reasonable element of risk had to be accepted. In other words, it became a classic "risk-to-benefit" military analysis, but one that was deadly serious. To me, as a military person the medium risk had to be accepted in such case.

I gave the group my reasoning. They then asked what I would personally do if my own finger were on the fire switch. I answered that I would take out the transmitter or transmitters and save those American lives, accept the medium risk, trust in God, and pray **for** the best. Military decisions often have to be made in just such uncertain conditions. If one's military could not make such tough decisions, then one's military forces would inevitably lose their wars.

The Group Fired and Took Out a Transmitter

The group thanked me, stated that they would fire, and hung up the phone. See Figure 57. They did fire on April 25, just as they promised, and they did destroy one of the KGB transmitters—apparently located some kilometers (perhaps 30?) from the Soviet nuclear reactors at Chernobyl.

At the moment the transmitter was hit and destroyed, engineer Bill Bise—out in the field with magnetic field detectors and measuring the Woodpecker beams—observed the abrupt failure of the Woodpecker transmitter beam coming around **the E**arth from the East. Thus, he actually detected the successful destruction of that distant interferometer transmitter. Unknowingly he had detected the results of a great counterstrike in a very dangerous strategic battle.

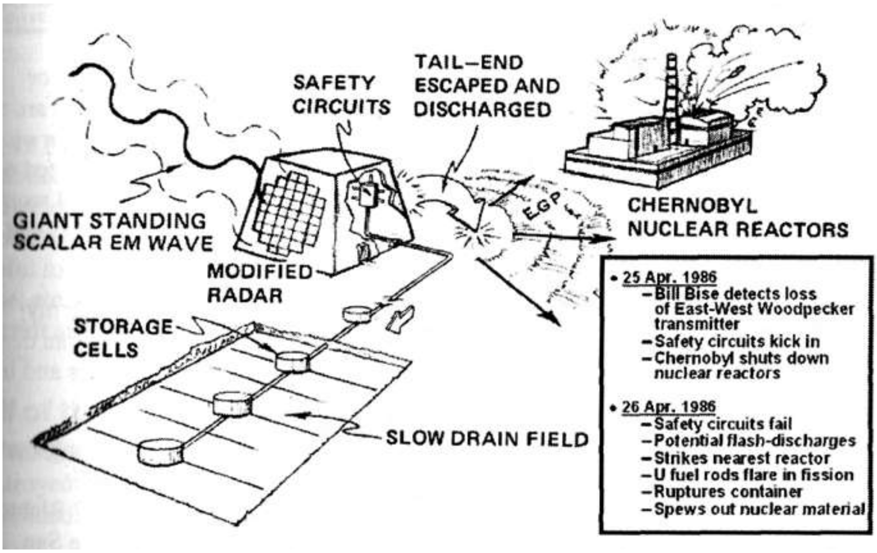
SAFETY CIRCUITS
TAIL–END ESCAPED AND DISCHARGED
EGP
CHERNOBYL NUCLEAR REACTORS
GIANT STANDING SCALAR EM WAVE
MODIFIED RADAR
STORAGE CELLS
SLOW DRAIN FIELD

- 25 Apr. 1986
  - Bill Bise detects loss of East-West Woodpecker transmitter
  - Safety circuits kick in
  - Chernobyl shuts down nuclear reactors
- 26 Apr. 1986
  - Safety circuits fail
  - Potential flash-discharges
  - Strikes nearest reactor
  - U fuel rods flare in fission
  - Ruptures container
  - Spews out nuclear material

Figure 57 Chernobyl disaster 25-26 April 1986.

Results at the Transmitter and at Chernobyl

We were very, very lucky. At the destroyed transmitter site, the safety circuits kicked in and caught the huge beam potential and held it for about 24 hours, gradually draining down the intense beam energy into the earth in a slow manner all the while.

A **frantic** Russian message to the Chernobyl reactors galvanized the operators to immediately shut down those reactors (put in the cadmium control rods).

However, the uranium in the fuel rods was of course still sitting there. Chernobyl then waited with bated breath to see if the safety circuits would hold.

Then on 26 April, the safety circuits at the transmitter site finally crumbled, and a far weaker LW flashover occurred than would have occurred earlier without the safety. The spreading LWs in the earth struck the nearest reactor core at Chernobyl, preceded seconds before by a tremor induced in the earth at Chernobyl as the LWs arrived. *Part* of the struck uranium in the core fissioned immediately, venting the containment structure and spewing out radioactive material. Nuclear fallout spread over much of Europe. The rest of the disaster is well known.

That unknown group, in my estimation, saved a large number of American lives—at least 200,000 or more and perhaps as many as 300,000. They also prevented enormous physical damage to California, both to Los Angeles and San Francisco and to points in between. The members of that group are the unsung heroes of this bizarre episode. It is my hope that I may live to turn the names of those persons in for the award of a very high Presidential medal. Until then I will not reveal those names under any circumstance.

Whether or not the U.S. scientific and intelligence communities know it or believe it is of no concern. What matters is that a great many Americans are still alive, and their children born since then are alive. Those Americans otherwise would have been very, very dead or never born. I personally have reflected on those desperate circumstances many times. If I had it to do over again, I would unhesitatingly urge the same action: Fire and save those American lives, and accept the medium risk. And pray.

And I did pray after recommending the strike. In my opinion, this time my fervent prayers were answered. I am extraordinarily sorry for the civilian deaths that occurred; all one can say is that it was under severe circumstances and in a very peculiar—and very desperate—war.

The Soviets had many other interferometers on line. They induced a substantial earthquake in Mexico at the end of April. Thus they did complete their "earthquake footsteps" of the milestone. On 7 May they induced a 7.7 Richter quake in the Aleutian Islands. If 7.7 to 9.5 quakes had hit all along the San Andreas Fault in Los Angeles and San Francisco and in between, then the destructive scenario I had envisioned for the West Coast would have occurred.

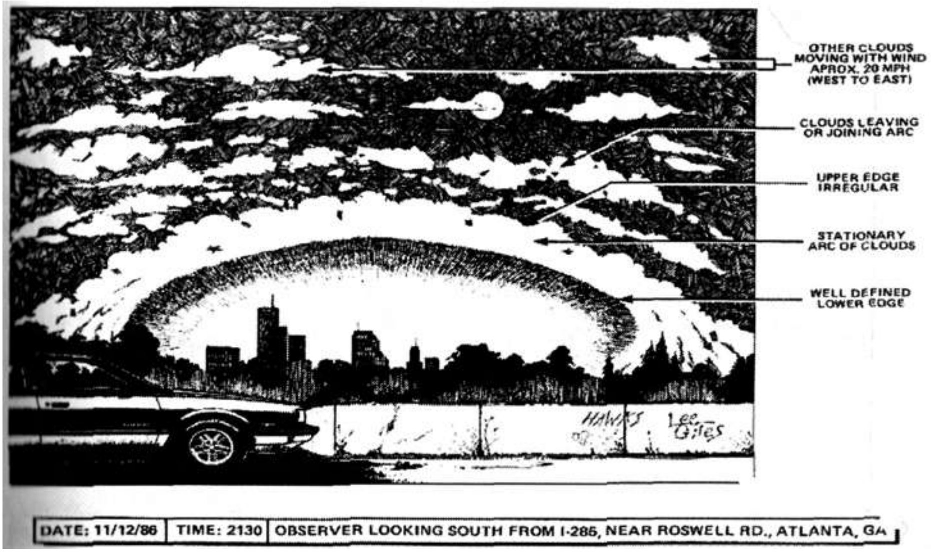
OTHER CLOUDS MOVING WITH WIND APROX. 20 MPH (WEST TO EAST)
CLOUDS LEAVING OR JOINING ARC
UPPER EDGE IRREGULAR
STATIONARY ARC OF CLOUDS
WELL DEFINED LOWER EDGE
HAWKS
Lee Giles
DATE: 11/12/86
TIME: 2130
OBSERVER LOOKING SOUTH FROM I-285, NEAR ROSWELL RD., ATLANTA, GA

Figure 58 "Bottling up" shield practice over Atlanta 12 Nov. 1986. With a real shield, aircraft or missiles flying into it (from inside or from outside) explode and are destroyed.

On 12 November 1986 the Soviets even placed a giant "dummy shield" (practice) over Atlanta, Georgia (Figure 58). If the shield had been real and not just practice, any aircraft and missiles flying into it would have been instantly exploded and destroyed. That mode—"bottling up" a strategic weapons area—is useful to nullify an entire strategic ICBM launch complex or a large bomber base, etc. It can also "bottle up" a large carrier task force at sea, including the aircraft and the Tomahawk missiles, the shells from the naval guns, etc.

Kill of the Challenger Included Use of Metal-Softening

As we mentioned, during that 1986 preparation period the Russians also destroyed the Challenger spacecraft (January 28, 1986). Numerous signatures abounded, all of which we have already published.

A few weeks before the kill of the Challenger, a close colleague had experimentally proven that a metal-softening signal was indeed placed on the Russian "woodpecker" signals (Figure 7). He did this quite simply: he wired himself into a receiver, through an intervening amplifier, and tuned the receiver to the Woodpecker signals.[44] Then he took nails from a can of nails he was using, and hit them with a hammer to drive them into planks. The nails all folded up instantly like totally wet noodles. A couple dozen, one right after the other. Simultaneously, his son—who was not connected to the Soviet signal—was pounding nails from the same box, normally and without difficulty. None of the nails picked up and pounded by his son experienced any difficulty or softness.

---

[44] Here he was taking advantage of the fact that the infolded LW signals inside the Woodpecker carrier wave, comprising the "engines" for metal-softening, also constitute "perfect" divergence-free EM energy, so to speak, at least overall. Amplifying the signal also amplifies the "engine" inside it. When connected to his body, "potentials superpose" is the rule, and that transferred the amplified engines to his body and into the EM field immediately surrounding his body. Hence he softened the hammer metal a tiny bit, and softened the much less tempered nails quite a bit more.

Then my friend disconnected himself from the signal, and waited a bit. Whereupon when he then used nails from that same box, he drove them in without difficulty. There were no more "wet noodle" nails.

We reach a very simple conclusion: When he connected to an amplified Woodpecker signal, his body was time-excitation charged (quick charge and quick decay) with a specific signal so that the nails he picked up were charged. Then when hit with the hammer, the nails acted as if made of very limp rubber without a rebound. When he disconnected from the amplified Woodpecker signal, and waited a moment for the induced charge in his body to dissipate, the nails also discharged their engines and quit acting like limp rubber bands when hit, and became normal nails with normal behavior. That would seem to be quite sufficient to prove the thesis.

The Challenger was then destroyed, as we previously printed, with numerous signatures of the hit, including even a KGB party in Moscow to celebrate the success of killing the vehicle (and getting away with it).

Cold Molding: An Epilog to the Metal-Softening Donnybrook

I was called a total lunatic for pointing out a metal-softening capability on the Russian signals. So be it. Metal softening is achieved by an engine which serves as a pumped phase conjugate mirror for the vibrations and signals in the lattice bonds of metals. Build up that engine in the metal until the precisely reversed and overlaid signals negate the normal signals, and one has "dissolved" the lattice bonds, converting the solid to a liquid. Lattice bond cancellation is not heating.

But that was not to be the end of the "metal softening" story.

A few years later, working informally with a large UK aerospace firm "for free", I furnished them an address in Moscow of some Russian scientists associated on the perimeter of the KGB energetics weapon program. The address had been taken from their cautious advertisement for business on the Internet and furnished to me by a correspondent. A UK engineer fluent in Russian was sent to meet with those scientists in Russia. They nervously showed him several novel things, one of which was "cold-molding" of metal. The liquefied the **metal**

The result was that a hammer blow onto a nail head resulted in the nail acting as a limp noodle upon impact.

without heating, poured it in a mold, and waited for the "effect" to wear off and the liquid to turn back into a solid again.

Briefly, an energetics charge (using LWs) was induced on the metal lattice vibrations, adding their phase conjugate replicas to the interior LWs comprising the lattice bond signals. The metal's atoms were time-density charged to an excited state (presently not recognized in Western science). Thereupon the metal lattice EM forces "dissolved" and the metal became liquid at room temperatures and without heating. The cool liquid metal was then poured into a mold. Then after awhile, the time-density charge and slowly drained away, and the metal resumed its normal solid state—all without any heating whatsoever.

On a subsequent trip the engineer brought that Russian team itself out of Russia and to the UK. The Russian scientists demonstrated cold molding to assembled UK scientists, and to some U.S. visitors. The phenomenon was physically demonstrated and proven.

I was correct about the metal softening then. We were also correct back there in 1986 about the metal-softening signal employed against the Challenger (together with several other things).

There are numerous other signatures on the death of the Challenger. Again, *the death of the Challenger was not an accident.* It was due to a deliberate act of war by the KGB. The KGB knew in advance—after substantial probing incidents—that they would get away with it. And they did so.

Additional Update from Year 1998 to Year 2,000 Has Been Given

In the action package we sent to the DoD in 1998, we told the rest of the KGB energetics weapons story up to that date. We gave the attack plans as we saw them and the most useful types of attacks to be expected.

I hope the reader is interested in this special energetics weapons information. High power microwave weapons, high energy lasers, and nuclear weapons are useful but no longer decisive weapons either strategically or on the battlefield against a foe equipped with such energetics weapons, and particularly against a foe equipped with quantum potential weapons and negative energy EMP weapons.

At present, we are intensely watching to see if we can pick up indicators for the next scheduled strategic attack, its probable timing, and the "new twists" included in it.

Presently, we have perhaps an 80% chance that the little nation can once again counter any new threat and cause the KGB to abort the impending attack. On the other hand, with Putin apparently limiting at least some of the action of the KGB, we have moved into our present asymmetrical warfare against terrorism—

and ultimately against the extensive terrorist teams and assets already infiltrated into our homeland.

As a single example, before the collapse of the former Soviet empire, the Russians long ago slipped in nuclear weapons (from 20 to 40 KT) and hid them in all our large cities and population centers, along with the Spetznaz teams to detonate them on command.[45] Several nations have introduced teams with normal sabotage assets and also with weapons of mass destruction such as radioactive nuclear materials (for a dirty bomb), anthrax, smallpox, bubonic plague, etc. Behind the scenes, the die-hard remnant of the old KGB crew are using QP weapons to "spread" the immune systems (Figure 53) of our populace and materially enhance any BW strikes that the terrorists unleash. The rogue Japanese team (Yakuza and Aum Shinrikyo) on site in Russia are using their LWIs in a giant weather war and the occasional kill of an aircraft, etc. Over the decades Castro infiltrated some 10,000 guerrilla/special forces types trained in camps in Southern Mexico for decades. Probably at least half are still loyal and dedicated—and waiting to blow up bridges, power substations, high voltage power distribution system towers, dams, harbor facilities, refineries, pipelines (both oil and gas), etc. The first phase of WW III—delivery of the weapons of destruction and mass destruction to the targets along with means to unleash them—has already been accomplished.

We are particularly struggling against the rise of asymmetrical warfare—the superpower's paradox (Figure 59)—and our present involvement in that "new" kind of warfare is still quite foreign to us. Only time will tell how well we fare in this new type of war, which we will be engaged in for many years to come. Vice President Cheney rightly regards it as almost a state of perpetual warfare we shall have to learn to live with, as shown by his words on Oct. 21, 2001:

"The war on terrorism will not be over in our lifetime. It is different than the Gulf War was in the sense that it may never end. At least not in our lifetime. The way I think of it is, it's a new normalcy."

---

[45] See Stanislav Lunev (with Ira Winkler), *Through the Eyes of the Enemy,* Regnery, Washington, 1998. Lunev is a high-ranking GRU defector. He is a former Colonel in the GRU, the military counterpart of the KGB. In p. 22-33, he summarizes the Spetznaz capabilities. On p. 22-27, he summarizes Spetznaz use of nuclear weapons already on American soil. On p. 26, he gives some of the ways in which the Soviets easily brought nuclear weapons into the U.S. On p. 30 he confirms Russian seismic weapons. Use of EMP weapons is on p. 30-31. Use of very, very low frequency weapons to destroy the human brain, put people into a zombie-like state, and aid in brainwashing is confirmed on p. 31. And this is i**nformation** Okayed by the CIA for release!

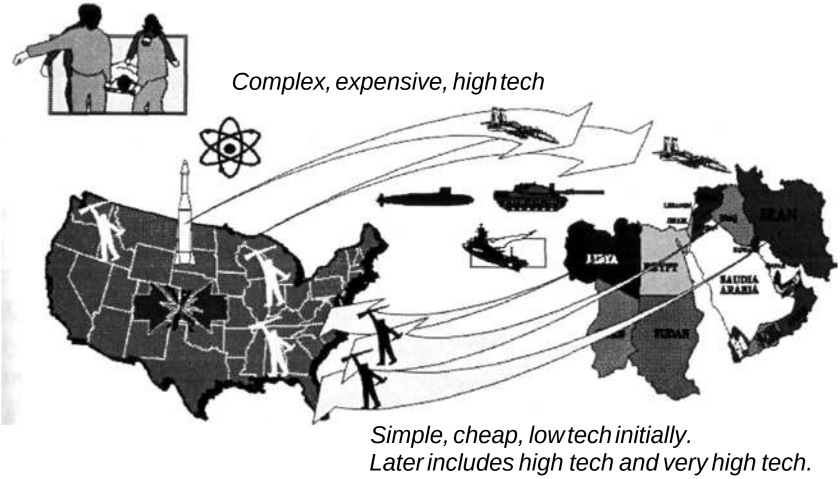
Complex, expensive, high tech
LEBANON
ISRAEL
IRAQ
IRAN
LIBYA
EGYPT
SAUDIA ARABIA
SUDAN
Simple, cheap, low tech initially.
Later includes high tech and very high tech.

Figure 59 The superpower paradox: Asymmetric cheap WMD strategic strikes and destruction, later becoming extremely high tech and extreme yield.

There is Also a Severe Bioenergetics (EM Biological Warfare) Threat

The threat from bioenergetics is equally vivid. That old die-hard section of the KGB can now attack an entire population, e.g., with electrodynamically (energetics) induced diseases of any kind whatsoever, particularly when using QP weapons. Such attacks are totally immune to inoculations and antibiotic treatment. Any kind of strain or variation of a disease can also be induced in the pathogen pool for that disease. Against a foe who has such capabilities, our entire medical establishment—including all our emergency treatment procedures and agents—*is obsolete right now.* A vaccination and the presence of antibodies, e.g., does absolutely nothing against the EM induction of the exact physical effects and form of a disease, as shown in the decades-long "microwave radiation" of the U.S. Embassy in Moscow (Figure 11).

To the Assistant to the SecDef in such matters, in 1998 we furnished a 200+ page briefing package and loosely proposed a crash development of the new type of medical therapeutic treatment devices pioneered and proven by the Priore group in France in the 1960s and early 1970s. Later, Congressman Cramer queried that office for an answer.

Unfortunately the Assistant SecDef s office erroneously responded to Congressman Cramer that nothing in the literature supported the mechanisms proposed. Indeed, the package we furnished contained a *detailed list* citing the hard French scientific papers in the literature, reporting the exact experiments and the astounding results obtained. Some 2,000 highly successful animal experiments for more than a decade of experiments in fact supported the thesis, and higher group symmetry electrodynamics fully supports the ability to use EM means to make spacetime curvature engines, so that these engines then interact back upon the targeted matter to perform the actions desired. That it does not appear in conventional electrical engineering is of no consequence; it appears in physics and it is good physics, just on the cutting edge.

We also reported, explained, and cited Becker's epochal work (Figures 27 and 28), which is copiously published in the hard literature and which was an example of the fundamental mechanism also. Becker was twice nominated for a Nobel Prize for his outstanding work. Unfortunately, the Assistant SecDef s office responded with a non sequitur.

Packages were sent also to NIH, NSF, CDC, and several other government agencies. Not a single one of those agencies was interested enough to have a competent scientist contact me and discuss the area. I strongly doubt that a single scientist anywhere in the Government bothered to *seriously* read and ponder the material so painfully and expensively compiled and furnished. Lieutenant Colonel (Retired) Ken Moore and I worked our fingers to the bone for weeks, preparing that briefing, and we spent all the spare money doing it that we could rake up and obtain.

The near-total lack of response itself seems astounding, but it clearly shows the state of U.S. science with respect to novel energetics innovations, even long after those innovations have been experimentally proven and documented in the hard literature, and even after they have been weaponized against us. It is a direct indication of why the U.S. never deciphered what was going on in some four decades of Soviet weak microwave radiation of the U.S. Embassy in Moscow, including the kill (eventual deaths) of three U.S. Ambassadors.

For decades there have been manipulations, to say the least, in the hotly contested field of the biological effects of EM radiation. We refer the interested reader to an excellent expose of that situation: The expose is by Dr. Andrew A. Marino, Ph.D., J.D., *Powerline Electromagnetic Fields and Human Health,* which is current. I believe it can be freely downloaded from: http://www.ortho.lsumc.edu/Faculty/Marino/PowerlineTOC.html. Dr. Marino has been personally and deeply involved in the investigation of EM bioeffects, in court cases, and in Blue Ribbon Panel investigations and proceedings. The book is an eye-opener, and it gives one a good view of the machinations in the EM bioeffects field. We quote directly from Dr. Marino, Chap. 6, p. 1:

> *"Oversimplistic as it may sound, whoever pays for EMF bioeffects research and analysis determines what data is produced and the way it is interpreted. Soon after the possibility that powerline EMFs were health risks was raised in a legal dispute involving the New York Public Service*

> *Commission, power companies and their trade associations, particularly the Electric Power Research Institute (EPRI), became massively involved in EM bioeffects research. Subsequently, the power industry dominated funding of the effects of powerline EMFs, both in terms of absolute dollars and compared with dollars from non-industry sources...the power companies and their trade associations were deeply deceitful regarding the information they provided to scientists and to the public regarding the potential health hazards of powerline EMFs."*

In the entire EM bioeffects field, there does not appear to be any mention of, study of, or recognition of EM bioeffects induced by the most powerful EM factors: longitudinal EM wave irradiation, time-polarized EM wave irradiation, and time-excitation charging of particles throughout the body and its cells, with resulting emission of transduced longitudinal and transverse EM radiation *from-within-to-without.*

National Institute of Health Reaction

NIH replied to my letter and action package from their *Policy*—read: *spin control*—Section. They noted that the package mentioned the Gulf War Syndrome, which was not in their mission area. So they just sent it off packing to DoD (simplest way to get rid of it, apparently no one seriously read it or even cared to read it). I queried them back by separate response letter about their lack of interest in a revolutionary proven new medical methodology well-documented to cure cancer and other dread diseases, in many hundreds of experiments by leading French medical scientists. This was certainly within their mission area.

They hastily replied from the *Policy* Section that, yes of course NIH was interested in all such things, and told me where to "submit a proposal." We never got out of their *Policy* department. Again, not a single competent scientist in the U.S.'s supposed master medical agency apparently even read the package or cared to call and discuss the matter. This was an indicator of their total lack of any grasp of the energetics threat, its testing, and its use to induce the Gulf War Disease.

Now perhaps one can see why we are so far behind the KGB energetics weapons scientists and weapons program.

Marino, *ibid.,* is particularly enlightening on the adamant opposition by appropriate sections of NIH vis a vis the harmful effect of nonionizing EM radiation.

No responses at all were received from the NSF and the CDC, or the other U.S. government agencies to which the package was furnished.

The CDC of course comes under NIH, and thus its responses or lack thereof would be dictated by the parent NIH policies. As for the NSF, Marino, *ibid.,* again is enlightening on the qualifications or lack thereof of its membership.

So as the reader can see, the top of the medical establishment in this nation is flatly not interested in any "controversial" or "out of the box" proven method for curing cancer, dread diseases, and mass treatment of those millions of American casualties that are coming in our cities and population centers. Again, they directly ignore scientific method's very premise: *If the experiment is replicated and contradicts the theory, accept the experiment and change the theory.* I personally concluded that the U.S. Medical establishment largely does not follow scientific method, but follows dogma and the already approved methodologies.

How the New Revolution in Medical Healing Works

For the reader who is a medical scientist and trained in physiology and biology, one might be interested in that revolutionary French therapeutic methodology and how it worked (before it was ruthlessly suppressed in the early 1970s).

Priore's apparatuses (Figures 12,49, and 60) simply pumped the cells of the entire body with longitudinal EM waves (each accompanied of course by its time-polarized twin, in a Whittaker pairing). The cells and every part of them are very nonlinear, so they act as pumped phase conjugate mirrors for Whittaker wave-pair engines resident in the cells and acting as "input signal complexes".

Once the cells are pumped by these now-established *time-density* waves, the damaged or diseased cellular masses slowly "charge up" with excess time-excitation. The irradiation took several hours, and this was the "activation" or "time-excitation charging" stage. Thereafter, the time-charge decayed slowly, all the while constituting *pumping* of every cell and every part of it, in the time domain.

The resident engine in the cell contains two components: (1) the normal or healthy component, and (2) the abnormal component representing disease or debility. Pumping these two components time-reverses both of them and amplifies that reversal. This creates amplified and incredibly precise *antiengines* in each cell and in each part of each cell. These antiengines override the weaker cellular engines resident. In turn, this "reversing" action of the antiengines causes the pumped cellular masses (actually the mass-energy, in its exact structure and dynamics)—at any and all levels, including the genetics—to be time-reversed back to normal. In biological instead of physics terms, it "dedifferentiated" the diseased or altered cells back to a previous healthy state by simply eliminating the diseased engine component. *This is a dramatic extension of nonlinear optical phase conjugation from the 3-space pumping domain to pumping in the complex (time) plane, and it is effective for time-reversing mass-energy at any and all frequencies. We accent that the cellular*

time *reversal back to a previous state includes all the cell's genetics and every* ***part*** *of the cell no matter how small.*

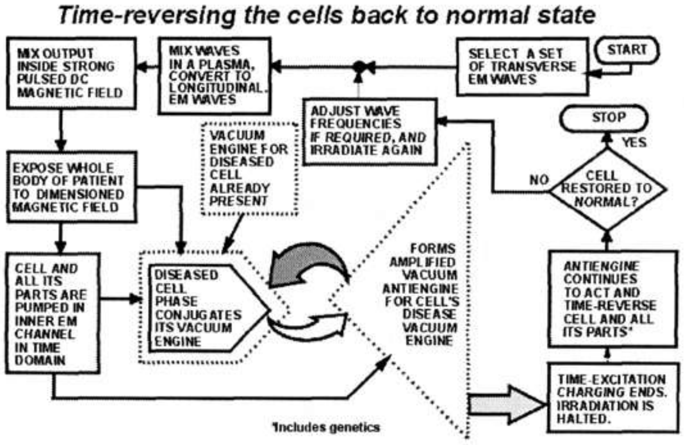
Time-reversing the cells back to normal state
MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD
MIX WAVES IN A PLASMA, CONVERT TO LONGITUDINAL EM WAVES
SELECT A SET OF TRANSVERSE EM WAVES
START
ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN
STOP
YES
NO
CELL RESTORED TO NORMAL?
VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT
EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD
CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN
DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE
FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE
ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*
TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.
*Includes genetics

Figure 60 Block diagram of Priore's healing (regenerative and time-reversing) process.

Normal cells are just backed up to a slightly younger condition. Diseased cells **have t**heir abnormal engine component overridden and reversed back to zero (its earlier state), so that only a slightly younger normal engine remains resident in the cell. This simultaneously reverses the cellular masses back to their previous healthy condition.

Contrast this direct healing mechanism to the action of the immune system.

The immune system itself does not *heal* anything. It contains the killers, the scavengers, the locators, the markers, the attackers, and the troops manning the ram**parts.** Usually after a fierce battle with invading hostile pathogens, it wins. Lots of cells still living are damaged in the melee. The immune system scavenger cells then scavenge up the residue littering the battlefield. But *the immune system itself cannot heal and restore a single damaged cell, including one of its own.*

That healing and restoration of damaged cells is the job of the *regenerative syst**em,*** still little known and inadequately studied today. Perhaps the most fundamental studies were mostly by Becker *et al.* some decades ago. The regenerative system (Figure 29) is an EM system in its operation, as clearly shown by Becker (Figures 27 and 28), yet it is able to physically differentiate and dedifferentiate cells from one form to another at will, and in highly selective fashion. The regenerative system acts within its limitations (only weak pumping can be done for the time-reversal of diseased or damaged cells) to eliminate the delta (the deviation from normal of the resident engines in the pumped cell). That restores the normal (healthy) engine and slightly amplifies it as well.

It is obviously of the utmost medical importance to understand and apply (and amplify) this little-known technical mechanism by which the regenerative system is able to restore damaged cells back to normal.

It turns out that *the regenerative system in its novel EM functioning uses precisely LW and TDW pumping of those damaged cells and gradually time-reverses their entire masses back to a normal healthy state, within its limitations.*

This is the secret of the mechanism of the cellular regenerative system—which is an electrodynamic system of very unusual nature. The exact mechanism has not previously appeared in the literature. It took me 14 years to uncover that fundamental mechanism, and another 4 or 5 to refine it a bit, even though both Priore and Becker (as well as others) applied it decades before without being able to fully decipher it.

The experimental proof is already in the literature, once one understands the higher topology EM unwittingly used by both Becker and the Priore team. The proper types of higher group symmetry electrodynamics required to understand the action of the cellular regenerative system are available. There exist scientists skilled in their application and use. These scientists and that higher group symmetry electrodynamics are little used, and are never used at all in the medical field. There exists no funding anywhere for higher EM examination and modeling of the functioning of the cellular regenerative system—i.e., for the fundamental healing process.

Just imagine what this could mean to medicine: *One can now develop a new medical science to amplify and apply nature's own proven method of restoration and healing of damaged cells, including all their damaged genetic material*. It matters not what caused the damage, whether pathogen or gamma radiation, whatever. This new therapeutic method is a totally new general relativistic process where powerful spacetime engines are created which work directly upon the damaged cell and every part of it, physically reversing the cell and all its parts precisely back to normal condition.

General relativity rigorously tells us that a precise spacetime curvature engine (set of precise spacetime curvatures) is present and active for every specific disease condition and every nuance in it. It also tells us this flawed (disease or damage carrying engine) continuously operates on the cellular mass at every level, maintaining the exact physical condition of the cell and its dynamics. By creating a precise "anti-engine" and amplifying and applying it, general relativity assures us that we are physically acting upon the cell and all its parts at every level, down to the smallest—to precisely return every part in perfect proportion back to a previous normal healthy condition. *Pautrizel directly proved the time-reversal to an actual earlier state of the cells and the organism* (Figure 61). Amplification by time-domain EM pumping, means that we can cause this time-reversal of the cells back to normal to happen much more quickly than can the body itself, which can only pump very weakly.

- Pautrizel compared identical treatment of immature rats with immature immune systems, and mature rats with mature immune systems.
- Treatment restored the damaged mature system back to full mature functioning, and it promptly recognized and dispatched the pathogens.
- Treatment restored the damaged immature system back to full immature functioning, unable to cope, and the pathogens promptly reinfected and killed the immature rat.
- The lesson: The body is indeed returned to a previous physical state. The pathogens themselves are not killed by the treatment, but by restoring the ability of the immune system to recognize and destroy them.
- Only *the previous ability* of the immune system is restored by the process.

PROF RAYMOND PAUTRIZEL
(Renowned Parasitologist)

Figure 61 Pautrizel's proof of time reversal of the stricken cells back to a previous state.

General relativity also assures us that any and every disease is subject to this novel treatment approach, even diseases which have resisted every other known therapy. The Sachs unified general relativity and electrodynamics, and GR/EM unification applications in O(3) electrodynamics produced by Evans, for the first time provide a solid theoretical basis for Priore's methodology and his evolutionary results. They also offer direct engineering of startling physical cures, once the technology is developed.

The profound implications of that treatment can be seen. Apparently, however, **our** leading governmental agencies and health agencies do not see it, do not wish to see it, have no appreciation for general relativistic effects in healing, and do not intend to have any appreciation for general relativistic effects in healing. Indeed, they seem totally uninterested in even trying to comprehend and research t*he fundamental healing mechanism* demonstrated by nature herself.

What the French Researchers Showed

Among other things, the Priore team produced revolutionary cures of *terminal tumors, infectious trypanosomiasis, suppressed immune systems, and atheriosclerosis* in several thousand laboratory animals in rigid laboratory tests under rigorous protocols. Eminent scientists such as Robert Courrier and Raymond Pautrizel worked with the Priore team. Courrier was head of the Biology Section of the French Academy of Sciences at the time, also *Secretaire Perpetuel* of the Academy, and a noted scientist of high stature. Courrier personally presented the astounding Priore results to the entire assembled French Academy. He also sent his own most able scientist assistant to personally prepare the terminal tumor grafts on the lab animals before their treatment.

A world-renowned parasitologist who did enormously important work with Priore was Raymond Pautrizel, who is still alive at this writing though quite aged. Pautrizel was greatly persecuted for his years of determination and effort. Courtesy of Pautrizel via the late Christopher Bird, I personally have the very Ph.D. thesis that Priore himself submitted to the University of Bordeaux on his revolutionary work. That thesis was maliciously rejected in the suppression that was applied to the project. When the French Government changed to a Leftist government in the mid-70s, Government support was immediately withdrawn and the project was ruthlessly suppressed. Priore later sickened and died and that was the end of that. More than a decade later a brave and determined Pautrizel finally was able to get a doctoral thesis approved in the Priore work area, at the same university, by another doctoral candidate, Eric Perisse {40}.

There are two U.S. patents on the Priore device itself, and of course several French patents also.

All the above was rigorously documented in the 1998 package prepared and submitted by Ken Moore and I.

While Priore was still alive, a team with which I was connected attempted to mount a well-funded effort to restore the Priore methodology while Priore was still alive. We also were viciously suppressed. The lives of our financial backers and their families were threatened. The head of our project was viciously attacked financially and destroyed, fleeing the country for his very life. Whoever thinks that dramatically innovative medical science in this country is just sweet reason and scientific method, should have a strong enema to remove the serious blockages interfering with his perceptions.

This vicious outside suppression of our project so infuriated me that I made a solemn promise that, were it humanly possible, I would decipher the Priore process before I died. It took 14 years, but we made good on that promise. During the 20 years or so since our project was suppressed, it is heartbreaking to consider the millions of children and sickened and suffering people who could have been saved had the project continued and succeeded. It is heartbreaking now to consider the coming millions of casualties in our civilian populace in our cities, most of who are going to die from mass destruction weapons. And most of those would not have to die, would the scientific community waken from its long sleep and discover some modern physics and higher group symmetry electrodynamics—and their application to solve nature's mechanism for **healing.**

When one time-reverses the body's cells and all their parts, one forces the diseased cells to dedifferentiate back to healthy state. We strongly accent that this is accomplished by unified GR/EM processes involving precisely **organized** spacetime curvatures and their local actions in the cells and all their inner components, down to the finest level. It can be rigorously expressed in good unified physics, as expressed in an electrodynamics such as O(3).

The nondamaged or nondiseased cells just get a little younger. That is also highly beneficial; it reverses the ravages of aging, just a little. Indeed, one of the implications of the Priore mechanism is that aging can indeed be reversed by the process and treatment. Again, one must keep one's sense of humor. In order to get this technology and therapeutic method born, what seems to be needed is an extremely wealthy person, strongly desiring to live longer than the expected life **span** and do it in a vigorous, healthy body—and who is willing to fund the project with the necessary $60 million required so to redevelop and extend the Priore technology so it can be readily applied. A byproduct will be the saving of millions of American lives when those coming mass casualties from weapons of mass destruction occur in our cities.

How the Regeneration Mechanism Was Finally Deciphered

What took so long in deciphering the mechanism was the problem of how to make such precisely tailored and exact ST curvature engines, specifically so that they are *exact* antiengines for a specific disease and even a specific pathogen, and even for different versions of the engines for different cells in a single body having different strains of the same pathogen, having different body DNA, etc. For years, the solution of that task seemed hopeless. How on earth could one possibly work all that out in a hundred lifetimes, so that it could be precisely calculated even by an array of the most powerful computers on Earth?

After years of struggling with the problem, in one stroke of enlightenment it was revealed to be the simplest thing in all the world. Absolutely no calculation was necessary, and the exact amplified antiengine effect was a fairly simple and straightforward thing to obtain. General relativity tells us that every cellular mass and each part of it, has its own unique spacetime curvature engine, to the finest detail and an exact fit. As Wheeler put it so nicely {130}, *"Space acts on matter, telling it how to move. In turn, matter reacts back on space, telling it how to curve. "* The mass and its dynamics and the engine and its dynamics constantly interact upon each other, and "keep attuned and in perfect adjustment," so to speak.

That means that a "normal" cell already has just such a precise "normal" ST engine residing in it, and acting upon every component at every level—even in the atomic nuclei. Well, a diseased cell has an absolutely specific structuring, and therefore it also has an absolutely specific ST engine structuring. *A priori that* resident structuring for the diseased or damaged state in the individual cell differs (by some delta) from the structuring of the normal healthy cellular state for that individual cell.

Voila! That means that *the ST engine for the damaged cell can be decomposed into two parts (two engines):* (1) the normal "healthy cell" state engine, and (2) a delta ST engine that has been added to it so that together the two comprise the full ST engine for the exact individual condition of the cell.

What is now needed is a way to produce an exact antiengine of two parts: (1) one for the normal engine and (2) a second one for the delta (disease) engine. Applying this amplified and combined "antiengine" would overpower the resident combined engine, precisely reversing the cellular condition by reducing the delta to zero. Rigorously, the cellular mass and every part of it must then also revert to a previous physical state, or else we must complete discard general relativity itself.

Well, if we were to just time-reverse the resident engine and its two components, that would give us the perfect antiengine with its two components. Amplifying this antiengine would then produce one that was more powerful than the resident engine in the cell. Applying that amplified antiengine to the cell, that would not really change the "normal" engine component, but it would precisely reduce the delta engine component back to zero (its previous state, before it was there!) Voila! A specific healing signal of perfect structure, for the specific deviant condition of the individual cell and its every part.

Since we have a combined GR/EM to use, we can think electrodynamically about this general relativity problem. Immediately we see that phase conjugate optics is a candidate, since the "resident composite engine" can correspond to the "signal" or input wave, and the pumping of the nonlinear mass of the cell and its parts (where they act as nonlinear phase conjugate mirror materials) produces an exact phase conjugate replica (the exact amplified antiengine). The presence of that antiengine overrides the magnitude of the resident engine, and it moves the cell precisely back along the temporal path previously taken. In short, it applies a specific kind of dedifferentiation of the cell back to its previous healthy condition.

Since we extend phase conjugate pumping to the time-energy domain, then the resident composite engine and its time-stream component are the "input." The amplified antiengine and its time domain component will then forcibly propagate that pumped mass energy back over the time trajectory previously taken by it. In short, it will time-reverse the cell and all its parts back to a previous condition, completely eliminating the delta (disease or disorder). The beauty is that it will also reverse damaged or altered genetics, such as in the AIDS disease. It can do this prior to the appearance of health symptoms—it can be done as soon as the patient tests positive for HIV.

Since the antiengine is amplified, it moves the cell back to normal fairly rapidly (in a few hours, for example). At least the time charge for the reversal process is delivered in that period; one may allow (as Priore did) a full week for the time-charge and its transduction to continue reversing the cells. But it is much faster than the normal speed of the human regeneration system using the same process, because now we have greatly amplified the time-reversal rate.

It is therefore ideal for a method of quickly treating mass casualties, particularly since it can be developed and used in a portable unit about the size of a large suitcase (Figure 13).

Interestingly, during the radiation with LW EM radiation, one also pumps the pathogen and time-reverses it as well. Its "normal state" towards the beginning was a condition whereby it could move with relative immunity past the immune system's opposition, having deceived the immune system. In the treatment, a small amount of the pathogen is therefore converted back into this condition just shy of the beginning, where that part of the pathogens is resistant to the immune system. However, a very small, short "post treatment" irradiation with LW radiation for a much shorter interval and at weaker power, will clean up this altered pathogen residue by allowing the immune system to recognize it and destroy it. Pautrizel clearly demonstrated this significant fact.

Cellular and Body Rejuvenation are Possible, Once the Technology is Developed

One can in fact really accomplish *rejuvenation* of the aged and infirm, as the technology is developed. One could certainly solve the major problems of Medicare, and cut the patient load at least in half, relatively quickly and inexpensively. Of course lots of big U.S. drug companies will be rather violently opposed to any such idea, because much of their empire would be threatened. Much of the organized medical community depends largely on the Big Drug Approach. So the organized medical community will also vehemently oppose it, just as the formal medical science community in France viciously opposed it.

One could also produce cheap, portable LW treatment units *en masse,* which is what we were suggesting to the Assistant. SecDef. That way one could flood the emergency response system (response teams, hospitals, fire stations, police precinct stations, schools, etc.) with quick and effective treatment devices. They could be used to rapidly {131} treat those millions of American casualties that are going to occur when one or more of our cities are attacked in the near future by terrorist WMD attacks. This would save the lives of most of those stricken millions of Americans.

Mow Many Americans Could Be Saved?

For an anthrax attack on Washington, D.C., first generation equipment could save perhaps 70% of those sickened Americans, almost all of who are going to die with *present* capabilities. With second-generation equipment, one could save perhaps 90%.

For an anthrax attack without spreading via QP clandestine preparation of the populace, one would save 0.7 to 2.1 million Americans. With clandestine KGB QP spreading and a *5:*1 increase in the yield of the anthrax strike, one would save 3.5 to 10.5 million Americans.

For a truly nightmare attack (as a worst case scenario) perhaps 5 major cities might be attacked in such manner and to such degree. In that case, one is talking about saving 3.5 to 10.5 million Americans for the unspread attack case, and up to 17.5 to 52.5 million Americans for the spread case.

Further, the development of the necessary portable treatment units is doable and affordable, for mass treatment capability. Training on a portable machine could be done for a high school student in 30 minutes or less, and that student could then give emergency treatment effectively.

Absolutely nothing else out there can presently warrant such projections. Yet this is the very technology our own medical science community is eminently uninterested in, has no knowledge of, does not wish to obtain any knowledge of, and does not wish to scientifically discuss.

There Are No Resistant Strains to the New Technology, and There Cannot Be Any

A most interesting thing is that there cannot be a "resistant strain" of a pathogen for the new treatment methodology, *a priori.* No mass can shield against spacetime curvature effects. Slight adjustments of the controls will adjust the spacetime engines being utilized. Two or three hour's work with a developed unit and a lab team on a new strain would suffice to completely determine the required adjustments, once the technology is developed. The new information could then be quickly relayed to all stations treating patients. On each treatment device, adjusting a few dials and controls for the laptop computer controlling the unit will do the trick. The use of the final device would be so simple that a teenager could be trained to use it in less than 30 minutes.

Such new capabilities are urgently needed not only for the millions of civilian casualties we expect in the future, but also for support of our armed forces in the field. Frankly, the anthrax shots are not going to do them much good, once one realizes that the "spreading" of their immune and regenerative systems is already occurring via KGB quantum potential means. Camel pox is available to any terrorist foe, and has almost the identical actions and constitution as smallpox. Indeed, it is almost identical to smallpox.

The Gulf War Disease and Syndrome were just small preludes of what is to come and what is already being used against the entire U.S., though as yet very clandestinely. Look again at the Kaznacheyev work. *He clearly proved in thousands of successful experiments that any kind of cellular disease or damage or death whatsoever, can be induced at a distance by novel EM means.* Replication experiments were successfully accomplished at the University of Marburg and also at the University of Sydney.

Results of the Priore' Treatments Are Well-Documented and Independently Supported

The Priore team's results are well documented in the hard French scientific literature. At least one French doctoral thesis was eventually accepted upon the work. The astounding results really occurred. They were scientifically substantiated under rigorous protocols and even police-guarded labs for tests by a learn of outside scientists. To then deny those results simply because no one **has** *understood* the active mechanism until now, is to deny the scientific method itself: *Believe the experiment when it refutes the prevailing theory.*

There is also independent work by other scientists that strongly supports the Priore results, once the mechanism is finally understood.

Robert Becker *et al.* in the U.S. conclusively showed that living cells *in situ* can be dedifferentiated, redifferentiated, etc. by laughably weak currents (picoamperes) and DC potentials (which are comprised of those famous Whittaker pairs). Becker was unaware that his potentials included time density waves and longitudinal waves, as predicted by quantum field theory and as shown by Whittaker in 1903 (once one understands that a bidirectional phase conjugate longitudinal EM wavepair is a time-density EM wave coupled to a longitudinal EM wave). Application of such EM potentials by Becker transformed red blood cells into pre-cartilage cells, then further into pre-bone cells. That latter cell type was then deposited in bone fractures, healing them when nothing else would. Becker and others have demonstrated partial regeneration of limbs in small test mammals, using the technique. The bone-healing techniques were finally approved by the FDA and are often used in some hospitals for treating otherwise intractable bone fractures.

Since then,other scientists have shown the directed formation of other types of body cells from cells from distant parts of the body (i.e., muscle-type cells from cells produced in the bone marrow, etc.). Such things are now well documented in the U.S. literature, in at least several hundred papers and in the leading journals including *Nature, Science,* etc.

The Present Threat Is Urgent, and Our Strategic Destruction is Still at Issue

Our own medical community, our science community, and our intelligence community are at least 20 years or more behind the KGB right now. We are also behind the Chinese. In many respects our scientists have simply missed the biggest scientific breakthrough of the century.

It is urgent that we find some way to counter and avoid future strategic energetics attacks sure to be scheduled. The eventual attack actually launched—if the little nation's QP weapons can be countered first—would destroy the U.S. in about 2 hours. It only takes about 10 minutes, it seems, for the QP weapons *already deployed* to negate almost all our strategic nuclear arsenal. Thereafter, the KGB and Yakuza/Aum Shinrikyo can simply attack us at will and utterly destroy the U.S. To any loyal American, that is an entirely unacceptable solution.

The most urgent requirement would seem to be immediate and strong contact at the highest level between the U.S. government and the leaders of the little foreign nation (and the other friendly nation) that presently is guarding the ramparts and saving our necks. A cooperative crash program to develop our own QP weapons would seem to be the order of the day, if it can in any fashion be negotiated. But to have any chance of success, the President and the Congress will have to use the utmost care in selecting the U.S. negotiators and scientists to explore this possibility.

Immediate transfer of the requisite capabilities to the U.S. would be desirable, at least from our own view. From their view, of course, we do not necessarily resemble a trustworthy scientific partner at all, based on our past scientific performance, and it is up to the U.S. scientific community to prove to them that the leopard really has changed its spots. That would all have to be negotiated and worked out.

If that is not possible, then it is of extreme urgency that our own country embarks upon a new Manhattan Project to develop strategic quantum potential weapons and defenses. It would be a great relief to discover that such had already been started, but I fear not.

Desperately Needed Defensive Capabilities Can be Provided by a QP Solution

We pose one more very, very useful QP capability prior to closing. As we stated, if a large U.S. city is hit with—say—an anthrax attack done in professional manner, we would expect 1-3 million casualties, according to a well-known major U.S. government study. With immune/regenerative spreading presently already ongoing here in our populace by KGB QP weapons, that figure may actually increase by about 5-fold. So the casualties (prompt and delayed) might well reach 5-15 million, in a single *professional* anthrax spray attack by two persons from one light plane on a calm night dispensing 100 kilograms of anthrax spores sprayed over the city with a simple agricultural sprayer. With the indicated Priore technology and treatment units in all emergency response hands, at least 70% or so of those casualties could be saved. Otherwise, *presently almost all of them will die.* That dramatic saving of lives is one desperately needed capability that could be met.

But there is a further capability of LW technology that nothing else presently known can do without serious impact upon the populace.

After the prompt and shortly delayed casualties are mostly cured and the remaining fraction dies, a terrible problem of BW contamination still remains. No one anywhere really knows how to clean up an entire city from anthrax spores. But unless harshly decontaminated, that contaminated city is going to be effectively uninhabitable for decades to come, if only the present technology and methods are available. Islands of the coast of Scotland, where anthrax was tested decades ago, are still "hot" and contaminated—and off limits.

The indications are that our nation has developed chemical sprays that can kill and neutralize the anthrax, including in the lungs of the populace breathing the spray. On the other hands, it is indicated that the spray is a high physical stress factor, and some of the more debilitated citizens will be injured or even killed by the spray itself. This includes babies and small children, the aged, those with lung and breathing difficulties, etc. As a matter of national survival, desperate diseases do deserve desperate remedies. So while spraying the zone might itself sicken and kill some thousands of our weaker citizens, at least—so the thinking would go—it saves millions. This of course is similar to the use of triage—reserving treatment and scarce medical capabilities and supplies for treating those most likely to live and recover with treatment, rather than treating those more seriously affected and less likely to survive. Sadly, presently triage will almost certainly be necessary because of very short supplies and treatment teams, etc. So most of those 1 to 3 million casualties will simply be dragged aside and left to die, and just made as comfortable as possible.

However, the QP technology can be adapted to produce systems that can indeed do complete decontamination, very quickly, and without harm to the citizens in zone. The anthrax can simply be modified by specialized spacetime curvature engines—specialized EM fields and waves carrying hidden inner "information content of the field" engines that are designed specifically to change and destroy anthrax and nothing else. Eventually one could develop the capability to do it similarly for *any other* major contaminant, including nuclear radioactive materials contamination of a major city.

Any strategic and tactical analyst can immediately see the tremendous value of such capabilities.

In Conclusion

I hope the above admittedly lengthy overview is of interest to the reader and useful. If we are to survive the debacle that now looms before us in this new asymmetrical war we shall be in for decades, it will require the efforts of leading and dedicated Lawmakers and Overwatch persons. Without firm prodding, the military, scientific, and intelligence communities are likely to just continue their present pace until many or most of us all die suddenly.

The KGB energetics threat is still real and imminent; Russia is not a monolith and Putin still does not have complete control over the KGB. Another full strategic attack is sure to be scheduled before very long by either that die-hard faction of the KGB, the Yakuza/Aum Shinrikyo, the Chinese (particularly if they move against Taiwan and we defend), or someone. And so our possible deaths also to be taken into account, if we just continue to do business as usual.

Along with other Americans, I love my country and served it to the best of my ability for 20 years of active duty. It is the survival of America itself that is now at stake, both in the asymmetrical war against terrorists and in the continuing war against energetics weapons. Our scientific, military, and intelligence communities simply must be jolted awake and into massive efforts in energetics.

We have the scientists with the required capability. We have the facilities. We have the leadership, and we can do it if we can put this thing under a Presidential Decision Directive, under an Executive Order and a Declaration of National Emergency (along with the several already in effect).

Thomas E. Bearden, Ph.D.
Lieutenant Colonel, U.S. Army (Retired)
President and CEO, CTEC, Inc.
Director, Assoc. Distinguished American Scientists

1986

Fer De Lance

Briefing on Soviet Scalar Electromagnetic Weapons

T. E. Bearden, Ph.D.
LTC, U.S. Army (Retired)
Director, ADAS

1986

Note: This is the original Fer De Lance briefing paper wich includes some minor text and artwork editing.

1986 FER DE LANCE INTRODUCTION & BRIEFING ON SOVIET SCALAR ELECTROMAGNETIC WEAPONS

This briefing presents the basic concepts of Soviet scalar electromagnetic weapons, some of the major types available, and evidence of their widespread testing.

Scalar Electromagnetics is Electrogravitation

Scalar electromagnetics is an extension of present electromagnetics (EM) to include gravitation. That is, it is a unified, and, what is more important, it is a unified engineering theory. Nikola Tesla initially discovered its basis.

In the scalar EM extension, EM field energy can be turned into gravitational field energy and vice versa. This exchange can be patterned and localized, in specific areas or objects. Such a controlled change of electromagnetics to gravitation is not possible in the normal EM or physics presently taught in Western textbooks. However, the bits and pieces of the theory have been scattered through the physics literature for some time, but no orthodox Western scientist seems to have realized that these anomalous portions could be integrated into a startling new physics. Unorthodox experimenters, inventors, and scientists have made discoveries in this arena for several decades, but again have not realized the exact implications or the precise manner in which their results could be combined with present electrical physics.

Fer-de-Lance

Such sluggishness is certainly not present in the Soviet Union. For over three decades, the Soviet Union has been developing electrogravitation and applying it to develop strange new secret weapons of incredible power and capability. They have sustained the largest weapons development program ever launched by any nation, and they have kept it effectively hidden from prying Western eyes. 1 have called this program "Fer-de-Lance," after the deadly South American pit viper of that name.

The dreaded fer-de-lance is a snake of great agility and lethal effect. It often ambushes its hapless prey, and strikes unexpectedly and without warning. Its first sudden strike is usually lethal to its victim, which promptly expires in writhing agony. Since the Soviet development of scalar EM weapons has been designed for the same purpose, the name seems appropriate.

The equivalent effort of about seven Manhattan projects has been poured into Fer-de-Lance by the Soviets, and the program has been successful almost beyond imagination. The eery weapons are now developed, deployed, and tested. The ambush has been completed; Fer-de-Lance is coiled and ready to strike.

Energetics and Directed-Energy Weapons (DEWs)

The ordinary Soviet name for this type of weapons science is energetics. In the West, that term is believed to be associated with conventional directed-energy weapons (DEWs) such as particle beam weapons, lasers, radio frequency (RF) directed-energy devices, etc. The Soviets do not limit the term in this way.

Western scientists are familiar only with directed-energy weapons where fragments, masses, photons, or particles travel through space and contact the target to deliver their effects. Hence in their thinking they limit the Soviet term "energetics" to the type of weapons they themselves understand - exotic but normal weapons using energy or mass traveling through space to impact a target.

However, it is possible to focus the potential for the effects of a weapon through spacetime itself, in a manner so that mass and energy do not "travel through space" from the transmitter to the target at all. Instead, ripples and patterns in the fabric of spacetime itself are manipulated to meet and interfere in and at the local spacetime of some distant target. There interference of these ripple patterns creates the desired energetic effect (hence the term energetics) directly in and through the target itself, emerging from the very spacetime (vacuum) in which the target is imbedded at its distant location. As used by the Soviets, energetics refers to these eerie new superweapons, as well as to the more mundane DEWs known to the West.

As a consequence of the Soviet breakthrough and decades of feverish development, monstrous strategic weapons undreamed of in the West are already in Soviet hands. A noose is slowly and steadily being tightened about our throats, and it is already the 11th hour.

The Soviets Use a Deception Plan

Concomitant with this supersecret development program, the Soviets developed and implemented an elaborate deception plan to conceal these startling weapons and their nature from Western eyes until it is too late. Soviet deception has been so successful that even when Western scientists are confronted with the actual Soviet tests of these weapons directly over their heads, they do not recognize the weaponry nor the nature of the effects produced.

As early as January 1960, Nikita Khrushchev announced the Soviet development of a new, fantastic weapon. On May 1, 1960 Soviet defensive radars - rigged as prototype weapons of this new kind - probably downed Francis Gary Powers's high-flying U-2 reconnaissance plane over the Soviet Union, precipitating a major diplomatic incident.

On April 10, 1963 one of the first new superweapons operationally deployed was used to destroy the U.S.S. Thresher atomic submarine underwater, off the east coast of the U.S. The next day, April 11, 1963 the same deployed superweapon was utilized in a different mode to produce a giant underwater explosion in the ocean over the Puerto Rican Trench, 100 miles north of Puerto Rico.

Over the years, various aircraft were interfered with or downed as tests of these Soviet weapons. A particular case involved the mysterious loss of F-111s in the Vietnam conflict. At least one downed F-111 crew was recovered in the subsequent prisoner-of-war exchange. On that aircraft, all electrical systems suddenly were in difficulty simultaneously. All emergency indicator lamps were lit up, "like a Christmas tree." This was probably due to special Soviet teams converting some North Vietnamese SA-2 missile system radars to the scalar EM mode, and employing "scalar beam" interference to produce spurious EM noise throughout the electrical and electronic systems of the aircraft.

In late April/early May of 1985, the entire armada of Soviet strategic scalar EM superweapons was activated as a special celebration of the 40th anniversary of the end of WWII. Activation of this armada (which probably contained over 100 giant weapons), together with 27 giant power systems and a large number of command and control transmissions, was monitored on an advanced, proprietary detection system by Frank Golden. After the giant strategic exercise, which lasted several days, most of the weapons and power sources were once again stood down to "standby."

The NASA shuttle launches provided a convenient opportunity for Soviet testing of these superweapons in a Launch Phase ABM mode, where a launched missile can be detected and destroyed shortly after liftoff. At first, electromagnetic pulse (EMP) bursts on the early shuttle trajectory were deliberately delayed in time, to prevent actual destruction of the target and avoid alerting the U.S. that something unusual was happening.

The shuttle launch of November 26, 1985 saw a particularly significant-test of this kind. In this case, a very loud "sonic boom" or explosion occurred over the launch site 12 minutes after shuttle liftoff, when the shuttle was already away and downrange. At least two previous shuttle launches had also been used as pseudotargets, with delayed booms occurring over the launch site well behind the vehicles.

After the lack of U.S. reaction to these three tests showed that the U.S. still had no knowledge of the new technology and did not even recognize its employment, the Soviets apparently decided to proceed with tests where the target vehicles would actually be destroyed.

On December 12, 1985 the same Soviet weapon tested against the NASA shuttle launches may have deliberately interfered with the controls of an Arrow DC-8 taking off from Gander Air Force Base, Newfoundland. At an altitude of 100 feet, the aircraft - carrying over 250 U.S. soldiers and civilian crew members - lost power and sank into the ground tail-low, killing everyone on board. Three Canadian witnesses to the crash were interviewed over the Canadian Broadcast Corporation's television news on April 8, 1986 at 10 p.m. No flame or smoke issued from the plane before its descent and crash. However, the aircraft was seen to be mysteriously glowing with a yellow glow. That is a signature of the use of a scalar howitzer in the "continuous EM emergence" mode, similar to the manner in which the F-111s were downed in Vietnam. In short, the DC-8's electrical systems were interfered with by electromagnetic noise created throughout each increment of spacetime occupied by the aircraft. The powerful charge created in and on the aircraft also apparently caused the loss of two engines, one after the other. With its controls ineffective and power drastically reduced, the aircraft sank to earth, still in its "tail down" configuration from takeoff, and crashed and burned. The "yellow glow" was a corona due to the acquisition of a high electrical charge by the skin of the aircraft.

Other factors contributing to the crash may have been reduced lift due to poor engine maintenance, increased weight of the aircraft due to icing, and heavy loading. Still, no one has recognized the significance of the "yellow glow" or what it implies, or the possible connection between loss of the aircraft and previous Soviet testing of-a Launch Phase ABM system against U.S. shuttle launches.

As this book goes to press, the last two U.S. Air Force Titan 34-D missiles fired from Vandenberg Air Force Base in California have blown up shortly after launch. The first one blew up on Aug. 28, 1985 just after lift-off. That loss has been attributed to failure of a high-powered fuel pump, causing a massive oxidizer leak and a smaller fuel leak. The second Titan loss occurred on April 18, 1986 when the missile blew up 5 seconds after lift-off. Its loss is still under investigation and no determination of cause has been made. Apparently the shuttle and the Titan presently provide the only viable launch vehicles for launching U.S. "spy" satellites. The loss of these sensitive satellites - if indeed they constituted the payloads - cannot help but be damaging to our strategic surveillance capability. According to the Los Angeles Times, the single remaining KH-11 satellite was launched in December 1984 and, with an expected life of two to three years, it could stop functioning later this year.

Whether or not significant Woodpecker grid activity existed in the vicinity of the Titan launch of August 1985 is unknown at this time. However, significant activity in the grid definitely occurred before the April 18 Titan disaster and on the same day.

On Easter Sunday, Mar. 30, 1986 engineer Ron Cole observed significant **cloud** signatures of grid pattern activity, correlated with Soviet Woodpecker measurements. On April 18, the present author observed traces of a cloud **radial** over Huntsville, Alabama and took photographs of it. Preliminary reports from

Thousand Oaks, California indicate extensive grid activity again on April 18, the day the second Titan exploded.

At least the second of these two missile destructions shortly after launch is suspicious, since the grid positively was active during that time. Also, a clear trail of Launch Phase A8M system indicators exists back to the massive scalar exercise of April/May 1985. The first Titan explosion in Aug. 1985 thus falls within the Soviets' "now let's test them against U.S. launch vehicles" period. The second Titan loss follows highly suspicious losses of the Arrow DC-8 on Dec -T7 1985 and the Challenger on Jan. 28, 1986. The same Soviet weapon system that destroyed those targets may also have destroyed one or both of the critical Titans.

The Soviets have also been able to significantly engineer the weather over North America for more than a decade without being found out. They have tested a fantastic range of anti-ballistic missile (ABM) defense weapons for over two decades, and no one is the wiser in the West.

As previously stated, prior to the end of November, 1985, at least three "wet-run" tests of a Soviet "launch phase ABM system" against actual U.S. shuttle launches were made directly over Cape Canaveral itself, and still no one recognized what was happening or what sort of weapon was being tested. On December 12, destruction of the Arrow DC-8 in Newfoundland produced no indications that the Americans and Canadians knew anything about the nature of the weapon possibly used.

Accordingly, after a sufficient wait to test our reactions (if any), the Soviets prepared to actually destroy a shuttle after its launch.

Destruction of the Challenger in January 1986

As the whole world knows, on January 28, 1986 the shuttle Challenger was launched from Cape Canaveral, Florida after exposure to undesirable weather conditions, and disastrously exploded shortly after launch. The evidence seems to indicate that, as the rising vehicle was stressed, one end of its right booster broke loose, twisting away and into the main fuel tank, causing rupture, spillage of the fuel and catastrophic explosion. Several other anomalies still exist, however, and it is clear that a problem existed with at least one of the booster seals. All seven astronauts aboard the flight were killed in the fiery destruction of the vehicle.

Of course no one had recognized that the Soviets had already tested a launch-phase anti-ballistic missile (LPABM) system against three of our previous shuttle launches. These "wet-runs" used a deliberate "time offset" to delay the explosive emergence of electromagnetic energy in a launched shuttle's location along its trajectory. The delayed test shots resulted in very large "booms" above the launch site after the shuttle was safely out of the area, but did not destroy the shuttle vehicles themselves. For example, the delayed-shot boom occurred 12 minutes after the evening launch of Nov. 26, 1985. Even a marker beacon (large light in the sky) was utilized on that launch shortly after lift-off. The marker beacon was photographed. In addition, another photograph taken in a time sequence shows another sudden streak of light coming down and ending in a burst of light. This was probably a spatially-offset test of the "pulse" mode for destroying the shuttle. The "light burst" would have been detected in the Soviet Union and scored against its intended offset position. However, another mechanism may have been used to cause destruction of the shuttle itself.

On Jan. 1, 1986 the presence of a metal-softening signal added on to the Soviet LPABM system's scalar EM transmissions was detected by a surprised Frank Golden. The metal-softening ability of the detected signal was experimentally verified by him. Golden also locally nullified the action of the scalar EM signal in a test, rather conclusively establishing (1) that it existed and (2) what it was. The signal was apparently being prepared for use against an upcoming U.S. shuttle launch.

Just prior to the launch of the shuttle in late January, 1986, the Soviets accomplished significant weather engineering over the U.S. The jet stream was severely bent southward in the middle of the U.S., bending rightward again to move across the Florida panhandle. Icy cold air from Canada was drawn far down, into Florida and onto the Challenger sitting on its launch pad. This exposure to cold outside its tested range was probably an additional factor contributing to the Challenger's failure. Positive signatures of the Soviet weather engineering and jet stream manipulation were observed and photographed by several persons, particularly in Alabama and California.

On January 28, 1986, the Soviet scalar EM weapon system effects were sharply localized in the launch zone. Localization involved higher frequencies being present; these are quite painful to small brains - whose hemispheres act as a scalar interferometer and detector - such as in birds. As commented upon by national news commentators, strangely the birds were not flying that morning. Indeed, they were staying down on the ground or avoiding the area, since the sky over the area was painful to them.

As the Challenger rose, the metal-softening signal would have been experienced in and around the boosters shortly after ignition, since the booster flame acts as a special "ion-plasma" turner/detector for the scalar signal. (Certain plasmas have a unique characteristic: they transform an input transverse wave to a longitudinal wave, or an input longitudinal wave to a transverse wave.) The effect of such a local signal is to "change mass in the immediate vicinity" with the particular scalar resonance signal. (Mass acts as a capacitance or accumulator for scalar resonance).

The ill-fated Challenger was doomed. After ignition, the booster flame acted as a ionic plasma detector/amplifier for the scalar metal-softening signal on the

Soviet Woodpecker grid. The metal in and around the booster flame was slowly and steadily weakening due to charge-up with the metal-softening pattern. This attributed to booster leakage from the already cold-damaged seal. The leaking booster poured out smoke and later flame, contributing to the impending disaster.

Substantial winds and air turbulence over the site increased the stress on the Challenger as it rose through this region. This also contributed to the impending disaster and may have been deliberately created there by Soviet weather engineering.

As the Challenger stressed, eventually one or more weakened metal mounts gave way, partially freeing the end of the right booster. The booster oscillated, rotating into the tank and rupturing it. Contact of the escaping liquid fuel and the flame resulted in a fiery explosion, destroying the vehicle.

Even so another anomalous flame or light was observed on the vehicle, and may have represented a very small additional "pulse mode energy form" produced by the Soviet weapon that was attacking the Challenger. The shock of the explosion probably killed the astronauts instantly, although the cabin appears to have remained essentially intact while it plunged several miles to the ocean below.

By treating excess cold exposure to the shuttle, inducing metal-softening in and around the ignited booster, and possible adding a deliberate "intensely hot spot", the Soviets caused the Challenger to weaken and destroy itself shortly after launch, in so subtle a fashion that NASA scientists would not suspect what had actually caused the mishap. The Soviets also had previously withdrawn all their trawlers and ships which normally shadow a shuttle launch, to prevent any suggestion of Soviet presence near or involvement in the catastrophic accident.

A few days later, sporadically intense "high frequency localization signals" were still present on the grid, at least at one hinge-point at Birmingham, Alabama. From 1-4 February 1986, many birds inadvertently flew into the zone when an intense breakout of these components occurred, and dead birds fell from the sky in substantial numbers.

* According to an Urgentgram sent out by General D. Graham, on the evening of the Challenger's loss, KGB headquarters held a party to celebrate "success of their active measure against the shuttle." Note that all Soviet scalar EM weapons - development, testing, deployment, and usage - is under command and control of the KGB.

Among other things, this briefing details the bizarre series of indicators that showed the Soviet intention to destroy the vehicle, and documents the accidental discovery of the signal with which they intended to cause the shuttle to fail and destroy itself.

Technological Surprise and a New Hiroshima

Sadly, the bureaucratic smugness of orthodox Western scientists has materially assisted the Soviet deception process. Most Western managerial scientists - particularly in weapons development activities — have continued to view the Soviets as ignorant peasants, still trying to clean the mud off their boots. This view, of course, is totally untrue and unwarranted.

It takes only a few examples to refute this attitude. In nonlinear mathematics, engineering and science, the Soviets have led their Western counterparts since the beginning. The electromagnetic pulse (EMP) effect of a nuclear explosion appeared in the ordinary Soviet scientific literature before Western scientists were even aware that the effect existed. The Soviets continue to lead the world in explosive welding, titanium forming and welding, etc. While the Soviet scientists exhibit little inclination to build good washing machines, they certainly do produce state-of-the-art technology — and beyond — in any area in which they focus their main efforts. (We do lead the Soviets in some areas such as computers, computer software, miniaturization, etc.)

Once before, a modern nation - the United States - developed a mighty weapon in secrecy and used it to force a powerful foe - Japan — to its knees. The mindbending atomic blows to Hiroshima, and shortly after to Nagasaki, showed once and for all that in the modern age technological surprise can prove instantly disastrous. Yet in our scientific arrogance, we have assumed that no one else will ever do such a thing. We have assumed that it certainly could never happen to us, and that the "secret weapon" scenario will never be repeated. On the contrary, it has happened again, someone else has done it, and it has happened to us.

It has also become fashionable in the West to believe that all the laws of physics are already discovered. We assume we already know all of them. While we have been pridefully crowing this tune, the Soviets have been steadily discovering new laws in secret, as well as new ways to circumvent the old laws.

We Have Given Up Defense in the Larger Sense

Meanwhile, to defend ourselves strategically, we have chosen to forego "defense" and rely almost totally on building a powerful strategic offense that is capable (we think!) of destroying our would-be enemies anywhere in the world. Our offensive striking power is based mostly on a triad of deterrent forces - land-based ballistic missiles, submarine-launched ballistic missiles, and strategic bombers. The cruise missile is presently being added to the triad. If this 4-part offense were nullified or destroyed by Soviet secret weapons, we would be powerless to prevent our own destruction and Soviet domination of the world.

We have been viewing ourselves as Samson, confident in the strength of our locks.

Yet literally our locks have been secretly shorn and the Philistines are already upon us. Indeed, Soviet energetics weapons are now capable of destroying our triad shield, our homeland, our armed forces in the field, and our population, quickly and efficiently. We have a new "gap" of monumental proportions: not a missile gap, not a submarine or bomber gap, and not even a particle beam or laser gap. We have a scalar electromagnetics or electrogravitation gap.

In this briefing, it is my purpose to show a portion of the Soviet weapons developments that have lead me to such dire conclusions. The Fer-de-Lance briefing is also an attempt to galvanize our leaders and scientists into action, for our present total vulnerability to the Soviet Fer-de-Lance weapons is intolerable and unacceptable.

Unless we achieve defenses — and quickly — our own demise in a fashion similar to the Hiroshima-Nagasaki scenario is inevitable and imminent.

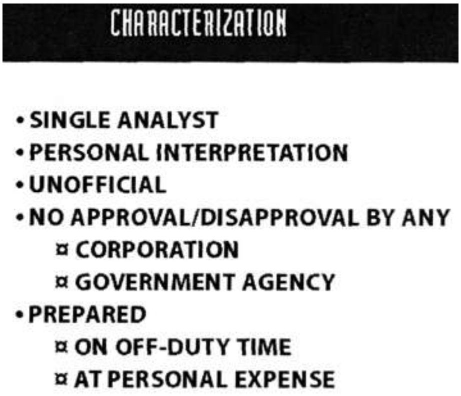

CHARACTERIZATION

- SINGLE ANALYST
- PERSONAL INTERPRETATION
- UNOFFICIAL
- NO APPROVAL/DISAPPROVAL BY ANY
    - ¤ CORPORATION
    - ¤ GOVERNMENT AGENCY
- PREPARED
    - ¤ ON OFF-DUTY TIME
    - ¤ AT PERSONAL EXPENSE

Slide 001: Characterization

We point out at the beginning that this briefing and its interpretation are the result of 22 years of analysis and effort by a single analyst. It is entirely unofficial. It is not approved or disapproved by any corporation, company, or government agency. It has been prepared by this analyst at nights, on weekends, on holidays, and during off-duty time.

It has been prepared mostly at personal expense, though several grants and support have been received from R. J. Reynolds III, Peter Kelly, Mike Bearden, Interdimensional Sciences, Association of Distinguished American Scientists, and others. Very valuable assistance has been received from the U.S. Psychotronics Association and the Planetary Association for Clean Energy. These contributions are deeply appreciated and most gratefully acknowledged.

Graphics and artwork support by Hal Crawford, Margaret Wilson, Tommy Neumann, Ron Cole, and Lee Giles has been invaluable. My sincere thanks go to these artists who have made the presentation vivid and imaginative. A great deal of background support was given by Joe Gambill, Frank Golden, Tom Herold, Ken Moore, Pete Kelly, Eike Mueller and Mel Bartlett and is deeply appreciated.

Finally, the help of many other persons and colleagues too numerous to enumerate is also gratefully acknowledged.

If the U.S. bureaucracy ever moves and develops defenses against scalar EM weapons, it will largely be due to the support of the persons and organizations who have freely given of their talents and resources to further this effort.

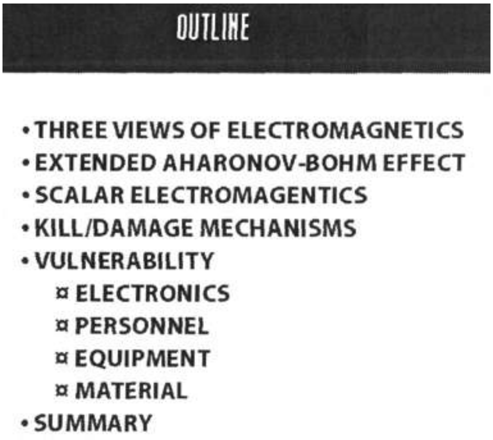

OUTLINE

- THREE VIEWS OF ELECTROMAGNETICS
- EXTENDED AHARONOV-BOHM EFFECT
- SCALAR ELECTROMAGENTICS
- KILL/DAMAGE MECHANISMS
- VULNERABILITY
  - ¤ ELECTRONICS
  - ¤ PERSONNEL
  - ¤ EQUIPMENT
  - ¤ MATERIAL
- SUMMARY

Slide 002: Outline

This presentation does not attempt to cover the complete field of energetics weapons developed by the Soviet Union. Instead, it concentrates on the major areas necessary to understand the type of weapons likely to be unleashed on our strategic and tactical installations and our troops and their equipment. It does not attempt to present the frightful bio-electromagnetic weapons the Soviets have developed to be unleashed on our populace at large and on our troops in the field. Also, the briefing does not dwell on the historical progression, but on the types of weapons themselves.

Of necessity, it presents several shortcomings in classical EM, and presents three differing views of electromagnetics and what can be done with them. The briefing particularly stresses the Aharanov-Bohm (AB) effect. This mesoscopic effect, which produces action at a distance by pure EM flux and not EM force fields, has been extended by the Soviet weapons developers into the macroscopic world around us.

In addition, we develop the basic concepts of scalar electromagnetics and how electrogravitation is achieved. The basic concepts of scalar interferometry and scalar resonance are briefly advanced, and a few of the major kill and damage mechanisms are presented. The nearly total vulnerability of our present electronic equipment, personnel, communications, vehicles, aircraft, ships, missiles, submarines, nuclear weapons, and materiel is stressed.

Finally, a short summary is given to highlight the main points of the briefing.

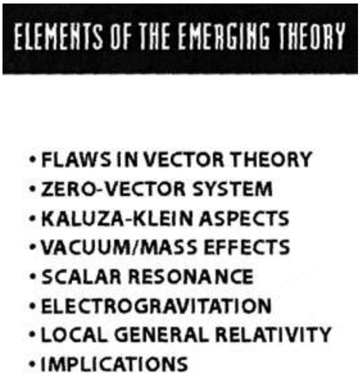

ELEMENTS OF THE EMERGING THEORY

- FLAWS IN VECTOR THEORY
- ZERO-VECTOR SYSTEM
- KALUZA-KLEIN ASPECTS
- VACUUM/MASS EFFECTS
- SCALAR RESONANCE
- ELECTROGRAVITATION
- LOCAL GENERAL RELATIVITY
- IMPLICATIONS

Slide 003: Elements of the Emerging Theory

Specifically, we will point out some flaws in vector mathematics itself, particularly with the concept of the zero force vector. The zero force vector is a system of forces that sum to a zero resultant. Hence the components of the summation represent a patterned stress in the medium to which they are applied, or in which they are imbedded. This includes the vacuum (spacetime) medium. In classical electromagnetics, this vacuum stress due to a zero-vector summation of EM force fields has been totally omitted and ignored.

As such, the EM "zero" force-vector summation produces a "trapped internal EM flux and flux pattern, without resultant (external) force field" condition - precisely as does the Aharonov-Bohm effect. The components of the artificially zeroed system, however, can be transmitted and still maintain their special relationship and coherence. While the AB effect has been shown to hold for the mesoscale (a few thousands of angstroms), the zero-vector scalar EM effect can hold for hundreds of thousands of kilometers.

To provide a unified electrogravitation, we adapt Kaluza-Klein 5-dimensional gravitational concepts to the idea of the zero-vector stress system in vacuum-spacetime. We also point out how simultaneously varying the magnitudes of **the** force components of a stress, all in phase, produces a stress wave or scalar EM wave. Scalar waves are almost always absorbed and emitted by the nucleus of the atom, not by the electrons in orbit.

The relationship of mass and vacuum, and the constitution of the vacuum, are pointed out from the viewpoint of modern quantum mechanics.

How a scalar wave resonance differs from conventional EM resonance is developed briefly. Mass and inertia are the direct result of - and are - trapped scalar resonance. The trapping mechanism is the spin of the particle.

Severely limiting assumptions in ordinary general relativity (OGR) are pointed out. In OGR, it is assumed that the local frame is always a Lorentz frame, and never curved. In other words, local spacetime is always assumed to be flat. This **saves** the conservation laws, simplifies relativity, and reduces "general" relativity to special relativity with distant perturbations and curvatures.

By removing this ad hoc assumption, a much richer local general relativity results. This local general relativity is readily engineered. Note that, in OGR, the physicist has actually assumed that he can never "engineer" local general relativity! Indeed, with the scalar EM approach, he can easily do so, in contradiction to what is taught in all Western universities.

By engineering a local general relativity (LGR), the individual conservation laws can be violated locally. This includes the conservation of energy/momentum, and the conservation of charge, for example.

The major implication of this startling new engineering physics is that one can **engineer** physical reality itself. For example, elements can be transmuted with minuscule energy input, free energy devices are possible, action at a distance is possible, communication faster than light speed is possible, etc.

By using the zero-vector approach, the virtual state can be organized and made **largely** deterministic, rather than statistical. This means that the probabilities of the states propagated forward by the Schroedinger equation can be engineered and changed. Whether or not a certain quantum change shall emerge or not can be determined or substantially influenced in advance. Bohm's hidden variable **theory** now becomes directly engineerable. This is a drastic change to quantum mechanics and physics in general.

Another implication is that this is the final engineering, for it allows the direct engineering of physical reality itself. Humans must now find a way to resolve their differences peacefully, or shortly Man will destroy himself and his biosphere by his own hand.

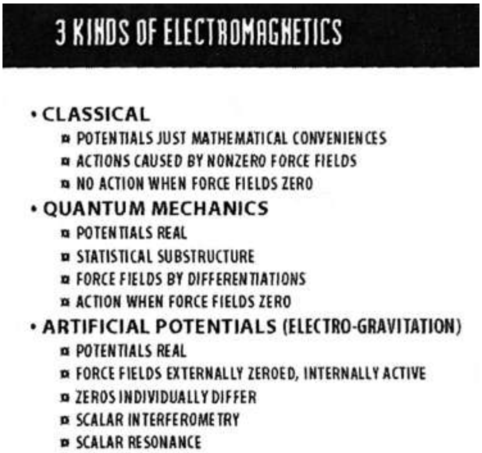

3 KINDS OF ELECTROMAGNETICS

- CLASSICAL
  - POTENTIALS JUST MATHEMATICAL CONVENIENCES
  - ACTIONS CAUSED BY NONZERO FORCE FIELDS
  - NO ACTION WHEN FORCE FIELDS ZERO
- QUANTUM MECHANICS
  - POTENTIALS REAL
  - STATISTICAL SUBSTRUCTURE
  - FORCE FIELDS BY DIFFERENTIATIONS
  - ACTION WHEN FORCE FIELDS ZERO
- ARTIFICIAL POTENTIALS (ELECTRO-GRAVITATION)
  - POTENTIALS REAL
  - FORCE FIELDS EXTERNALLY ZEROED, INTERNALLY ACTIVE
  - ZEROS INDIVIDUALLY DIFFER
  - SCALAR INTERFEROMETRY
  - SCALAR RESONANCE

Slide 004: Three Kinds of Electromagnetics

There are actually three kinds - or three views - of EM. These are (1) the classical view, (2) the quantum mechanical view, and (3) the scalar EM or electrogravitational view.

In the classical view, the potentials are just mathematical conveniences and do not physically exist. The real causative agents are the force fields, and there is no longer any electromagnetics going on if the force fields reduce to zero. Further, the ideas of "charge" and "charged mass" have been made erroneously synonymous.

Of course the classical view was formed from the idea of a thin material ether, with electricity as a thin fluid, long before the discovery of the electron. Since most earlier scientists studied string waves and these are transverse, the EM wave was modeled as a transverse wave. Also, detection equipment actually detected transverse waves. The role of electron spin and drift velocity, which would have shown that force-field-causing EM waves in the vacuum could only be longitudinal, was not yet discovered. Maxwell's equations and the classical approach became so engrained that the basic derivations were never corrected for more modern discoveries.

The quantum mechanical view, on the other hand, regards the potentials as the real physical actants, and the force fields are just effects derived from the potentials by differentiating operators. Classically oriented physicists have adamantly opposed this foundations requirement of QM because it would require nearly a complete redo of EM theory. It would also rather drastically change our ideas of physical reality. For years a controversy has raged around the Aharonov-Bohm effect (which demonstrates the reality of the potentials, among other things). Only this year - 1986 - have most physicists finally accepted the AB effect with its implications (see Physics Today, Jan. 86). However, no changes have yet been made to EM theory and the basic classical approach to electrical physics and engineering.

Yet even the QM view is flawed, since it has not examined the structure of EM forces which sum to a vector zero. Such a system produces stress, and if the summation is in the vacuum itself, it produces stress of vacuum/spacetime. Rigorously, this is a gravitational effect, and the energies of the various EM components in the local region are locked into an artificial potential. From general relativity, this type of potential where the energy density of vacuum is altered is a gravitational potential. From Kaluza-Klein unified theory, it is at least a 5-dimensional gravitational potential.

If the individual force vector components of the vector zero are varying in magnitude - say, all in phase - then they produce a gravitational wave. The energy density of the local vacuum is being rhythmically varied. Call such an EM wave a scalar EM wave, where, by "scalar" we imply that, to an external observer, the EM force vector resultants are identically zero, but the local gravitational potential of the wave is varying. Thus this is an electrogravitational wave, and vector zeroing of EM force fields constitutes a means of changing EM field energy into G-field energy. On the other hand, by breaking up the coherence of the zero-vector summation of the EM forces, nonzero EM resultants are recovered, constituting the change of gravitational energy into EM energy.

A simple means to break zero-summed coherence is by interference of two or more such zero-vector waves.

In this, third view of EM, action at a distance is easily possible, and is the norm rather than the exception. In addition, a new kind of resonance - scalar resonance - exists. The scalar EM wave does not interact with orbital electrons, but rather with the interior of the nuclei of atoms. Thus the new scalar EM resonance is between nuclei and within nuclei.

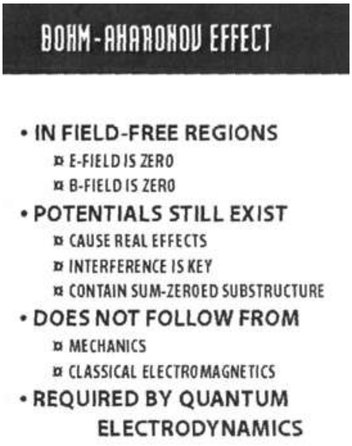

BOHM-AHARONOV EFFECT

- IN FIELD-FREE REGIONS
  - E-FIELD IS ZERO
  - B-FIELD IS ZERO
- POTENTIALS STILL EXIST
  - CAUSE REAL EFFECTS
  - INTERFERENCE IS KEY
  - CONTAIN SUM-ZEROED SUBSTRUCTURE
- DOES NOT FOLLOW FROM
  - MECHANICS
  - CLASSICAL ELECTROMAGNETICS
- REQUIRED BY QUANTUM ELECTRODYNAMICS

Slide 005: Bohm-Aharonov Effect

In 1959, Aharonov and Bohm published a fundamental paper in Physical Review, which pointed out the QM implications of potentials as the real entities, while force fields were derived effects. They showed that, even in the presence of zero EM force fields, the potentials may still exist and produce real effects in physical systems. They also suggested experiments to prove these predictions.

Interference of the potentials is the key mechanism producing real effects in charged particle systems, even in the presence of zero E-field and zero B-field. Unfortunately the AB paper did not address the issue of zero-summed systems of EM force fields as one way of producing artificial potentials having unique characteristics

At any rate, the "AB effect," as it came to be called, was rather hotly resisted and debated over the years, until finally it has generally been accepted as proven in 1986. The principle represents a violation of both classical mechanics and classical electromagnetics. It is required, however, by quantum mechanics and quantum electrodynamics. And experiment has proved it.

(See Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), Aug. 1, 1959, p 485-491.).

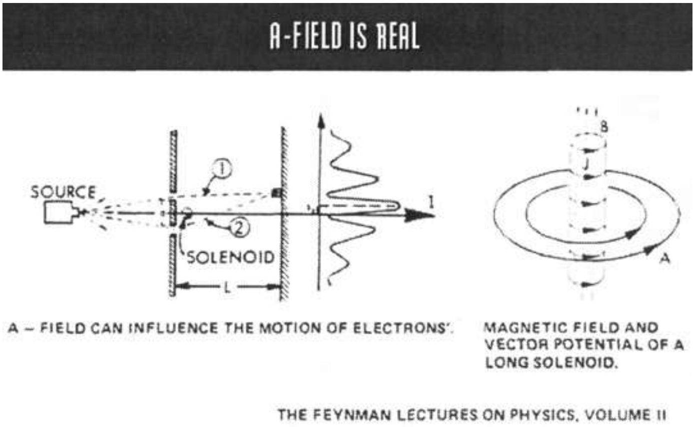

A-FIELD IS REAL

A – FIELD CAN INFLUENCE THE MOTION OF ELECTRONS'.

MAGNETIC FIELD AND VECTOR POTENTIAL OF A LONG SOLENOID.

THE FEYNMAN LECTURES ON PHYSICS, VOLUME II

Slide 006: A-Field is Real

In classical EM, the vector magnetic potential (the A-field) had been defined as a mathematical convenience by the equation

$$\nabla \times A = B \qquad [1]$$

But if the potentials are real, then conceivably the A field can be loosed from its enchainment to the $\nabla \times$ operator. In that case, it becomes a free, new, and independent field of nature with potentially unique characteristics. For example, its defining equation shows that magnetic force field can be made from it, and the rightmost term of the equation

$$E = -\nabla \emptyset - dA/dt \ [2]$$

shows that its time rate of change makes an electric field.

Let us explain this in more simple terms, and somewhat more precisely than conventional theory. We will use the Kaluza unified G-EM interpretation and electron flow in our explanation.

In the A-field, we have a certain kind of 5-dimensional G-potential which can bleed-off as EM force fields in two ways: (1) in a swirl fashion, where the vortex producing the swirl moves parallel to electron movement, and the swirling is a torque or spin, and (2) in a linear fashion, where the time rate of change of the A-potential produces a linear E-field on the electrons.

The first bleed-off as given by equation [1] constitutes the magnetic B-field, and the second bleed-off as given by the rightmost term of equation [2] creates a component of the overall electrical E-field. (Bleed-off of the electrostatic scalar potential produces the other component of the E-field).

Now in general relativity (GR) theory, "the" G-potential is just a conglomerate of many things, each of which has the characteristic of curving spacetime. "The" gravitational field is not a single thing at all, but is composed of a collection of many things.

Thus if we realize that both the electrostatic scalar potential (Ø-field) and the magnetic vector potential (A-field) are components of the 5-d G-potential, then we see immediately that bleed-off of these two components of the 5-d G-potential creates all normal EM force fields.

It follows that, if we produce a zero-vector summation of the two or more EM bleed-offs, we are actually "putting as much back in" to the 5-d G-potential through its A and Ø components as we are taking out electromagnetically. In that case, the 5-potential is in a state of equilibrium with respect to EM bleed-off. It is now forced to bleed-off in the only other way it can: as ordinary 4-dimensional gravitational field. Thus by vector-zeroing EM force fields, we turn EM field energy into G-field energy and vice versa, via the intermediary of the 5-potential.

At any rate, soon after publication of the Aharonov-Bohm paper, experiments showed that, if the magnetic field is trapped inside a long solenoid, a phase shift still is induced in the two-slit electron experiment, even though - classically - no contact of the enclosed magnetic field and the moving electrons occurs. This phase shift is explained by the fact that the freed A-field exists outside the trapping solenoid, even though the B-field does not. Consequently, interaction of this free A-field with the electrons produces a phase shift of the QM interference detection pattern.

This proves that A-field is real and causes physical effects.

It also proves that a form of action at a distance is real, just as required by quantum mechanics.

Years ago, Frank Golden and this author - together with Dr. William Tiller - experimented with "free A-field" devices. Golden went on to develop prototype transmitters and receivers and a prototype underwater communication system.

Since that time, Gelinas has patented several curl-free magnetic vector **potential** (free A-field) devices: see U.S. patent no. 4,447,779, May 8, 1984; 4,429,288. Jan. 31, 1984; 4,429,280, Jan. 31, 1984; and 4,432,098, Feb. 14, 1984. These patents are assigned to Honeywell.

(See also Theodor Kaluza, Sitz. Berlin Preuss, Adad. Wiss. 966, 1921.).

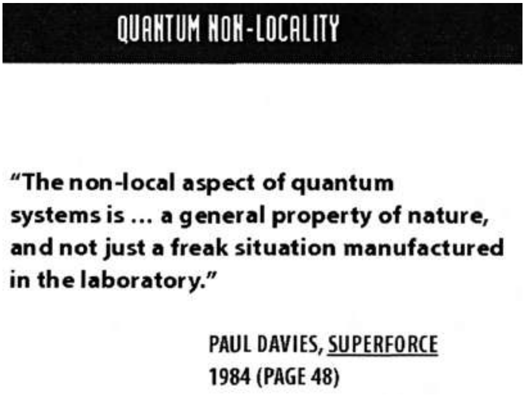

QUANTUM NON-LOCALITY

"The non-local aspect of quantum systems is ... a general property of nature, and not just a freak situation manufactured in the laboratory."

PAUL DAVIES, SUPERFORCE
1984 (PAGE 48)

Slide 007: Quantum Non-Locality

Indeed, quantum systems exhibit nonlocal effects that are general properties of nature. These effects are not just freak situations manufactured in subtle laboratory experiments.

However, if spatial coherence in such microscopic effects can be induced and sustained, then action-at-a-distance can be obtained in large macroscopic systems. This is the engineering aspect of the scalar EM approach.

It is also the aspect that the Soviets have secretly weaponized for over three decades.

The zero-vector EM force-field summation of multiple EM waves is used to transport specific time-phased patterns of EM polarization of vacuum to a distance. There interference of two or more such patterns results in the interference of the internal patterns, breaking of the phase relations, and appearance of positive (added) or negative (extracted) EM energy in the intersection region.

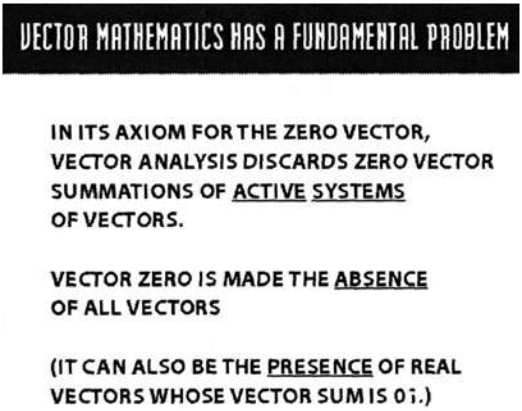

VECTOR MATHEMATICS HAS A FUNDAMENTAL PROBLEM

IN ITS AXIOM FOR THE ZERO VECTOR, VECTOR ANALYSIS DISCARDS ZERO VECTOR SUMMATIONS OF ACTIVE SYSTEMS OF VECTORS.

VECTOR ZERO IS MADE THE ABSENCE OF ALL VECTORS

(IT CAN ALSO BE THE PRESENCE OF REAL VECTORS WHOSE VECTOR SUM IS 0i.)

Slide 008: Vector Mathematics has a Fundamental Problem

In its concept of the zero vector, vector mathematics discards zero-vector summations of active systems of vectors. It replaces such a summation with a zero vector. This is fine for mathematics as an abstract system, but it is in error when applied to real electromagnetic force fields of nature.

In the abstract mathematics, a vector zero summation is made the "absence of all finite vectors". Further, all vector zeros are made equal. No concept of the "internal stress" of the zero vector exists in abstract vector mathematics.

However, physically the zero summation or "balancing" of vector forces in a medium represents stress in that medium. In the physical case, a vector zero summation system of non-zero vectors has a dynamic substructure, and this substructure is an individual.

Obviously, in the physical case vector zero summations may materially differ, both in the pattern of stress and the magnitude of stress. They cannot all be equated. Further, they are not the absence of vectors, but represent the presence of vectors in a special manner. Such a zero summation of EM vector forces directly stresses the "medium" - which may be the vacuum itself.

In the physical case, several changes to the axioms of abstract vector mathematics are required. (1) the "potential" of a vector zero must be taken into account, such as is represented by the sum of the squares of the magnitudes of its vector components. (2) the specific deterministic pattern of the vector components comprising the zero must be taken into account. (3) The dynamic variation in both the deterministic directions and deterministic magnitudes of the components and of the overall pattern must be taken into account. (4) Frequencies of the changes in the direction, magnitude, and actual makeup of the vector zero must now be accounted for. That is, time and wavelengths are rigorously aspects of the vector zero, and these may be deterministic variables. Since time itself is now a variable aspect of the vector, the vector zero system can affect its "rate of time flow" in the observer's system. (5) Since a "reference vector zero" can be established at any point of a vector magnitude, then individual vectors themselves may have dynamic substructures inside a special "zero reference" in and on the vector. The patterned potential of a vector is a reality.

This leads to a system of "vectors nested inside vectors" ad infinitum. In other words, it leads to an infinite-dimensional system, and the "opening" of every finite closed vector system through its vector zeroes.

For application to physical electrogravitational systems, at least 5 dimensions are required, four of space and one of time.

Here we still are considering only a special case where all vector zero summations represent EM force field energy locked in a gravitational potential, or gravitational force field energy locked inside the EM potentials. That is, we are prescribing a system where the only intertranslation of energy is among mass, EM field, gravitational field, and vacuum virtual energy (anenergy).

(See T. E. Bearden Toward a New Electromagnetics: Part III: Clarifying the Vector Concept, Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1983; Part 4: Vectors and Mechanisms Clarified, 1983. See also E. T. Whittaker, Proc. Lond. Math. Soc. 1, 367, 1903; Robert Bruce Lindsay and Henry Margenau, Foundations of Physics, Dover, New York, 1963, p. 283-287; Richard P. Feynman et al, The Feynman Lectures on Physics, Addison-Wesley, New York, **Vol.** I, p.2-4.).

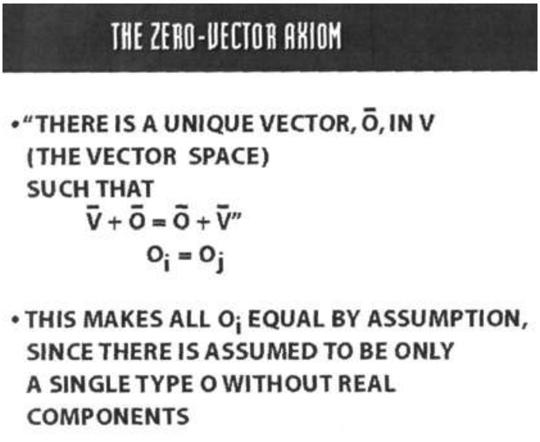

THE ZERO-VECTOR AXIOM

- "THERE IS A UNIQUE VECTOR, $\bar{O}$, IN V (THE VECTOR SPACE) SUCH THAT

  $\bar{V} + \bar{O} = \bar{O} + \bar{V}$"

  $O_i = O_j$

- THIS MAKES ALL O_i EQUAL BY ASSUMPTION, SINCE THERE IS ASSUMED TO BE ONLY A SINGLE TYPE O WITHOUT REAL COMPONENTS

Slide 009: The Zero-Vector Axiom

This slide shows a standard mathematical statement of the zero-vector axiom, used in many texts.

As can be seen, all zero-vectors are made identical, and no substructure is considered.

Such an abstract system of vector mathematics does not fit physical reality. However, when all one's EM foundations concepts are strongly conditioned in this fashion, the gravitational aspects of EM are discarded.

Further, EM theory, experiments, and equipments will be developed along these limiting lines. In that case, EM engineering of gravitational effects will not be developed. One will assume that it is impossible to have a locally curved spacetime, and hence all local frames will be Lorentz frames. Therefore the conservation laws will rigorously apply.

Further, the nonlocal quantum effects will stay firmly in the microscopic world where they belong. Physical macro-reality will stay stable and predictable in a classical manner.

General relativity (in curved spacetimes) will be restricted to the special case of the local special relativistic (uncurved) spacetime, with only distant perturbations (spacetime curvatures).

The world will stay sane and "normal", and the totally variable insanity of physical reality will remain bottled up in the minute microworld or the distant maelstrom of suns, stars, black holes, etc.

The Aharonov-Bohm heresy will stay kaput in the real world and will not be unleashed beyond mesoscopic reality — a few thousand angstroms or so.

With such an error in the application of vector mathematics to physical science, we become "flatland physicists," so to speak. We become like skimmer bugs on the surface of a pond; we have no appreciation of the dynamics of the depths below or of the heavens above.

Worse, we condition our development of instruments according to the fixed notions in our heads. Then we use these "biased" instruments to do experiments that reaffirm our notions. Any deviation suggested to this is considered heresy and nonsense.

And as Max Planck once pointed out, you will get a new physics only when the old physicists — who so adamantly oppose it — die off.

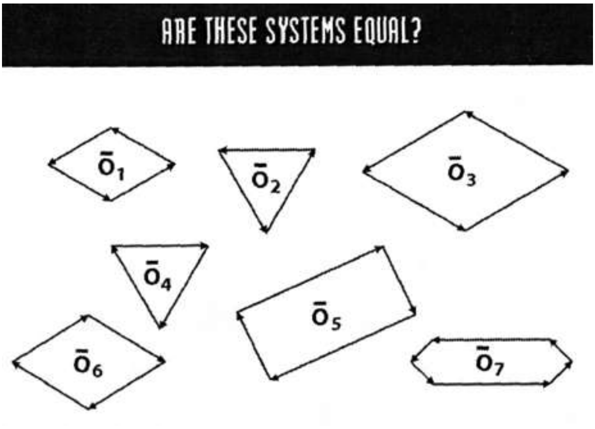

ARE THESE SYSTEMS EQUAL?

Slide 010: Are These Systems Equal?

Shown here are some representations of various very simple zero-vector systems. As can be seen, the stress magnitudes and the spin and dynamics of all of these systems are very different.

Further, their "stress potential" substructures are not composed of a huge collection of little random force vectors. Instead, their substructures are deterministic. If the components being put into the zeros are varied in direction and magnitude, but always sum to a zero resultant, then highly complex substructure dynamics may be created and utilized, infolded (Bohm's term) inside the zero vector.

In theory one can establish "channels" of communication and power (magnitude) transmission without ever surfacing a nonzero EM force field. In practice one can do this with very little EM spillout. Thus one can establish electrogravitational channels and energy transmission inside normal EM potentials and/or ordinary EM carrier waves.

One can use a conventional system of EM potentials, force fields, and waves as a special kind of "wiring circuit" in and through which to transmit and produce electrogravitational effects. By distant interference, EM effects at a great distance can be accomplished, yet no "ordinary EM energy" has passed between the transmitters and the distant interference zone.

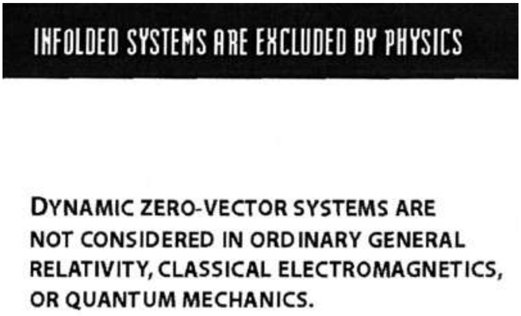

INFOLDED SYSTEMS ARE EXCLUDED BY PHYSICS

DYNAMIC ZERO-VECTOR SYSTEMS ARE NOT CONSIDERED IN ORDINARY GENERAL RELATIVITY, CLASSICAL ELECTROMAGNETICS, OR QUANTUM MECHANICS.

Slide 011: Infolded Systems are Excluded by Physics

Unfortunately, these infolded dynamic zero-vector systems have been eliminated from physics.

They are omitted in classical EM, ordinary general relativity (OGR), and quantum mechanics.

That is, they have been omitted in Western physics.

They have not been omitted in secret Soviet weapons labs and weapons development programs.

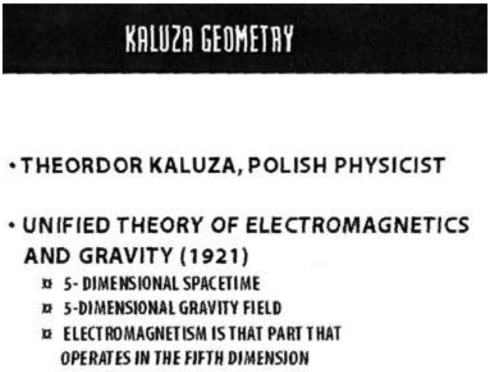

KALUZA GEOMETRY

- THEORDOR KALUZA, POLISH PHYSICIST
- UNIFIED THEORY OF ELECTROMAGNETICS AND GRAVITY (1921)
  - 5-DIMENSIONAL SPACETIME
  - 5-DIMENSIONAL GRAVITY FIELD
  - ELECTROMAGNETISM IS THAT PART THAT OPERATES IN THE FIFTH DIMENSION

Slide 012: Kaluza Geometry

In 1921, Theordor Kaluza, a Polish physicist, published a unified theory of electromagnetics and gravitation. Albert Einstein, who had had the paper for two years, recommended his paper for publication. In that theory, five dimensions - four space and one time - are utilized in the basic model. This gives a 5-dimensional spacetime.

In the model, electromagnetics and ordinary gravity are two aspects of a single more fundamental field: the 5-space gravity field. Electromagnetism is that part of the 5-field that operates (bleeds-off) in the fifth dimension.

The ordinary 4-space G-field is the small residue that spills over into our ordinary 4-space, and doesn't slide around into the fifth dimension and bleed-off there as electromagnetics.

Implicit in the theory, but not explicitly stated, is the fact that the 5-space G-potential bleeds-off in two fashions: (1) in the 5th dimension nearly unopposed, to produce what we see as EM force fields, and (2) in the ordinary 4-space we live in, which we see as ordinary gravitational field.

Normally almost all the force field bleed-off is in the fifth dimension as EM; only a minuscule bit bleeds off as 4-gravity.

Just as one example, between two electrons the 5-space G-potential bleeds-off into electric force field about 10^{42} times as much as it bleeds-off into 4-space G-field.

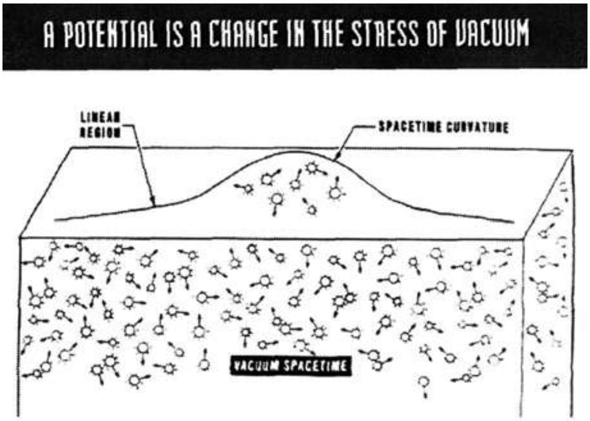
A POTENTIAL IS A CHANGE IN THE STRESS OF VACUUM

Slide 013: A Potential is a Change in the Stress of Vacuum

In physics, the idea of a "potential" is very poorly defined, if at all. The normal "definition" for the electrostatic scalar potential, for example, is not a definition at all. Instead, it's an abstract operation that tells how to mathematically calculate the magnitude of the potential, by pushing a unit charge in from infinity, against the potential field.

Now ultimately a definition must be an identity statement, not an operational statement about something.

Since I could find hardly anyone who actually understood what a potential was, and how or of what it was composed, I simply took quantum mechanics at face value and constructed this model.

First, the modern view of the vacuum is that it is not an emptiness, but a plenum. It's teeming with an incredible amount of raw energy, existing totally as the temporary bits of energy of virtual particles. That is, the vacuum "energy" exists as the nonintegrated energies of little particles which appear from nothing fare spontaneously created) and disappear into nothing (are spontaneously annihilated) almost instantly. This happens so fast that the little particle and its energy cannot individually be detected; it doesn't hang around long enough.

Hence the fleeting little booger is called a "virtual" particle. Like, "We had 'im there man, for just a moment, but he was gone before we could grab 'im!"

This kind of ghostly particle, however, is very real in modern physics. All observable forces are theorized to be due to the integration or collection of a large number of virtual forces created by exchange of these little particles.

However, this implies a peculiar thing about the vacuum "energy." Actually it's vacuum energies, for each little bit of energy only exists momentarily, then returns to the nothingness from which it came. That is, vacuum virtual energies are totally unintegrated. Each "bit" of virtual energy exists individually. These virtual energies cannot be "added up" to compute the "available energy" unless some collector or integrating mechanism puts them together.

That is, unless some kind of "zipper" is there to "zip them together" into macroscopic, observable energy. The usual "zipper" is a spinning, observable particle of mass. We say that the vacuum energy is unzipped and hence unobservable in that state. We say that "observable energy" is just zipped-together vacuum energies - integrated by and on a spinning, observable particle of mass.

So the vacuum can be visualized as a special sort of violent, unzipped plasma, where the particles of the plasma are fleeting and do not last. We can speak of it also as a sort of special unzipped gas, in the same manner.

The "pressure" of the vacuum gas we refer to as "stress." Due to its violent boiling activity, the vacuum gas is always under very high stress.

This "stressed virtual plasma" identically IS the vacuum and identically IS spacetime. Length, time, frame, mass, etc. - all arise by zipping operations (or understood zipping operations) in and of this raw vacuum/spacetime. When one uses the vacuum notion in this manner, one now uses the term "ether" again. However, this is not the old "thin material ether that hangs around"; this is an entirely new kind. It's a virtual flux and a virtual plasma.

The ether is also what we call "spacetime" in relativity. As can be seen, spacetime is stressed. Further, spacetime has a substructure.

By an uncurved spacetime, we mean one in which the stress of these virtual particles (the "pressure in the ethereal gas," so to speak) is constant from one place to another and from one time to another. Indeed, we also should mean **that** the stresses of the individual components of the entire plasma are constant from one place to another and from one time to another.

By definition, when we determine the lengths between every two points, we determine a frame. (Actually, we have prescribed a universal "length zipping" operation/mechanism as well as its characteristics.). A "linear frame" or "Lorentz frame" is one in which the pressure or stress remains constant. It may be higher than ambient vacuum, however, due to a particle's or an object's motion through the observer's ether. In the case of constant velocity, the particle/object encounters a higher level of virtual particle flux in the observer's ether, like a vehicle moving through raindrops. It encounters a constant rate of flux, however. In such constant velocity case, the moving object's frame is said to be "rotated," but not curved, with respect to the laboratory frame of the observer.

In a "curved" spacetime, we mean one in which the stress of vacuum increases or decreases from one place to another, or from one time to another, or both. At one fixed location, if spacetime is curved, then the stress is increasing or decreasing as a function of time. That is, the flux density of that area is changing. We may visualize the local frame as undergoing angular acceleration, rotating more or rotating less. This is a non-Lorentz frame, or curved spacetime. It is also a nonlinear vacuum.

In such a vacuum, local gradients exist in the flux of virtual particles. If affected by (coupled to) one or more of these gradients by a nonzero coefficient, a local object experiences effects (forces and actions) "without observable cause." Observably, conservation laws may be locally violated by these "virtual rivers" of vacuum by coupling, just as a paddlewheel dipped in a river violates its own local "conservation of energy" due to the river providing an energy source to it.

One must be very careful with the "pressure" analogy, however, when one utilizes unified theory (Kaluza theory). We now have a virtual plasma vacuum in five dimensions, not four. But the basic notions to enable our visualization and understanding still apply.

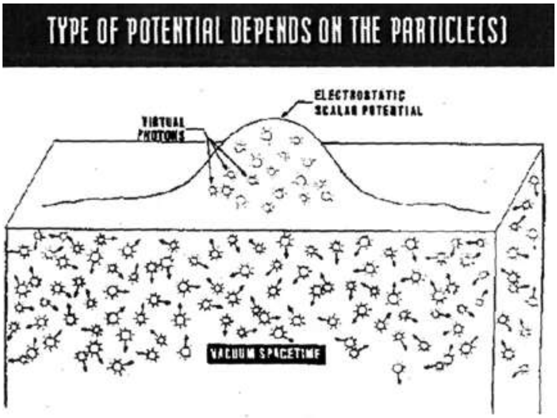
TYPE OF POTENTIAL DEPENDS ON THE PARTICLE(S)
ELECTROSTATIC SCALAR POTENTIAL
VIRTUAL PHOTONS
VACUUM SPACETIME

Slide 014: Type of Potential Depends on the Particle(s)

Now one can see that, by choosing the type or class of virtual particle, and looking at its potential or stress in the vacuum, one can have many kinds of "potentials" existing in the overall "gravitational potential" of vacuum spacetime.

For example, the stress of the virtual photon flux (which is what causes electrical charge) is called "electrostatic scalar potential."*

We can choose any other class of virtual particle we wish, and there exists a potential (and "change") for it. There are quark potentials, neutrinic potentials, etc.

Indeed, one of the reasons for ignoring Kaluza unified theory for so long was that it predicts a great many types of potentials and fields which have never been observed and which are unknown. Approximations, which neglected all these mysterious, and unknown fields, gave erroneous results. Perplexed physicists, seeking to simplify matters, simply turned away from it in frustration. They returned to it in the mid-70's when it appeared that 11-dimensional Kaluza-Klein theory possibly offered a complete unified theory.

Actually, the diversity of new fields in Kaluza theory was its strength not its weakness. That they are presently unobservable and undetectable is beside **the** point. Neither is time directly observable; however, we have worked out handy procedures and measurements of spatial quantities, which allow us to infer **time.** We have not just tried to drop it from physics!

And indeed the Kaluza theory is consistent with experiment when the higher fields are included. The message is, we must work out "clocks" for these new potentials and fields, so that they can be inferred with at least some degree of precision

- The character of movement also determines the type of potential. The longitudinal (radial) movement of the virtual photon flux is implied in electrostatic scalar potential. The swirl (tangential) component of this flux might be taken as the magnetostatic scalar potential - a magnetic pole - when the cw and ccw components of the swirl balance, but the sum of their absolute values differs from the sum of the absolute values of the corresponding tangential components in the swirl flux of the local ambient vacuum. If this swirl stress is higher than the vacuum swirl stress, this constitutes a north magnetic pole. If lower, it constitutes a south magnetic pole. If there is a prevalence of direction (cw or ccw) in the swirl (i.e., one component exceeds the other), this corresponds to the magnetic vector potential (the A-field) when the A-field is freed from the magnetic force field (B-field) — i.e., when A-field is freed from its $\nabla\times$ operator in the equation $\nabla\times A = B$. The bleed-off of the excess swirl stress from the north magnetic pole to the deficient swirl stress of the south magnetic pole constitutes the magnetic force field, B — usually referred to simply as the "magnetic field."

A POTENTIAL EXTENDS TO INFINITY & INVOLVES THE ENTIRE UNIVERSE

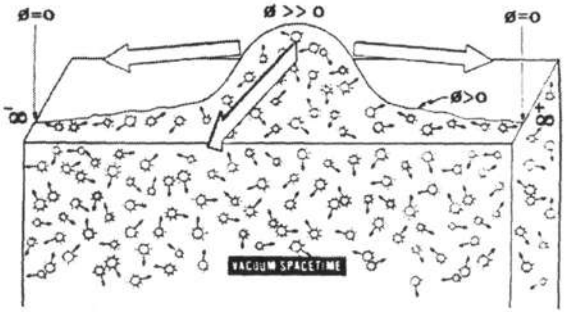

Slide 015: A Potential Extends to Infinity & Involves the Entire Universe

As shown on this slide, the change in the stress of vacuum due to a potential tapers off to infinity. Its magnitude normally does not reduce to zero until infinity.

Some types of potentials reduce much faster or slower than others, however, and so - at some distance away from a common origin of the potentials, one or more may be neglected in its (their) overall effects on an experiment or system.

Also, the speed of propagation of potential is highly dependent upon (1) what type of virtual particle or particles produce its stress, (2) its magnitude, (3) its composition, and (4) the presence or absence of coupling objects and other interacting potentials.

Fundamental observable particles are like little spray pumps and little vacuum cleaners at the same time.

Each is emitting virtual particles into the vacuum in a spray or flux, and at the same time it is absorbing virtual particles from the vacuum spray or flux. It's a little dynamo, and nature furnishes its driving energy unceasingly.

However, each little virtual particle it emits is itself such a little simultaneous pump and vacuum cleaner. That is, it's also emitting an even finer (and faster) flux of smaller virtual particles, and at the same time it's also absorbing flux al this time finer level.

"Bigger fleas have smaller fleas to bite em;

And so on ad infinitum".

We may consider the vacuum to be made of an infinite number of virtual state layers or levels.

It is not too difficult to show that each of these successively deeper, nested "layers" of virtual state corresponds to a successively higher dimension being added to our 4-space basic spacetime. So hyperdimensions and deeper levels of virtual state are one and the same thing.

The basic "speed" of the first layer is c, the speed of light.

The basic "speed" of the second layer is c^2.

And so on.

This is interesting. We may directly engineer the virtual state by means of the vector zero (scalar electromagnetics) approach. By nesting vector zeros inside other vector zeros, we may directly engineer the deeper layers of virtual state, a**nd** consequently hyperdimensions.

Scalar electromagnetics thus is virtual state engineering and hyperspatial engineering at one and the same time.

Superluminal communications systems, hyperspace drive, and materialization a**nd** dematerialization are all hypothetically possible, using scalar electromagnetics. As the technology develops, we should see the development of many of the systems long thought impossible except in science fiction.

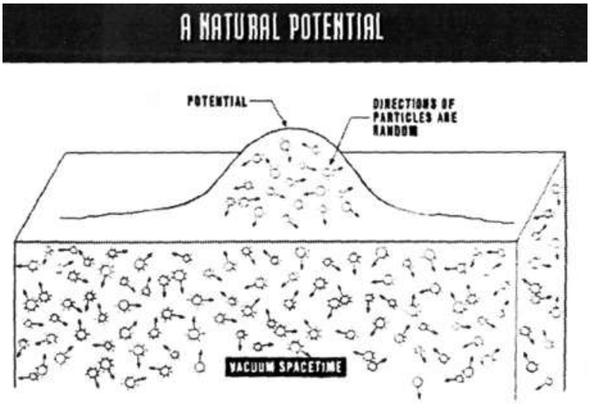

A NATURAL POTENTIAL

Slide 016: A Natural Potential

There are two broad categories of EM potentials, depending upon the way in which they are formed.

The two categories of potential formation are: (1) natural, and (2) artificial.

The difference is in the virtual particle flux substructure that comprises the potential.

If the substructure of emerging and vanishing virtual particles is random and without deliberate order (polarization) except in one broad, overall sense, then that is a natural potential. As such, there is no deterministic functioning going on inside the substructure of the potential itself. The potential can only act "as a whole" on a charged particle system.

Natural potentials have substructures of random virtual vector components. Any small amount of coherence as a function of distance vanishes quickly (on the order of a few thousand angstroms at most, according to present Aharonov-Bohm experiments).

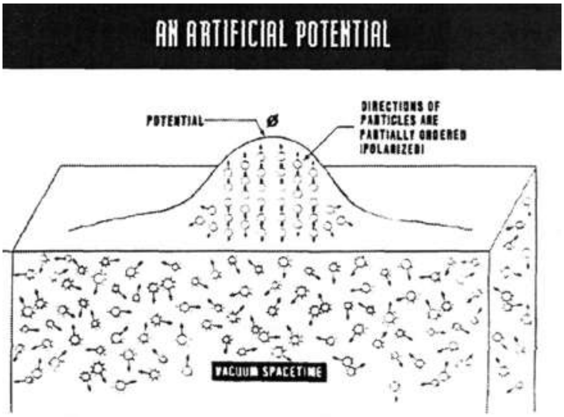
AN ARTIFICIAL POTENTIAL

Slide 017: An Artificial Potential

An artificial potential by definition has a substructure composed of deterministic observable vector components, summed to an overall zero vector. Coherence as a function of distance can be maintained over enormous macroscopic distances - even hundreds of thousand of kilometers - by simultaneous transmission of the entire cluster of substructure components as a coherent "zero" group.

To the conventional linear EM detector, at any point along its transmission path, this substructure is detected as a zero-E vector and a zero-H vector. Hence the conventional detector does not see the artificial potential, even if its "stress magnitude" (a function of the magnitudes of all summed components) changes.

Action at a great distance is possible with this artificial potential, however, if a highly nonlinear situation is met so that the phasing of the components is broken or significantly altered. In that case, the components do not sum to zero after their dephasing, and real EM force fields emerge. This "dephasing" can be made to occur at the distant nonlinearity.

Action at a great distance is not possible with the natural potential, but is possible with the artificial potential.

Further, with the transmitted artificial potential, VO must be applied from the transmitter to the distant interruption zone, as if there were no intervening space between.

Note that VO may be positive or negative. Thus energy may emerge at the distant disruption, or be extracted from there. In the first case, energy is input to the transmitter, to re-emerge at the distant dephasing zone. In the latter case, energy is extracted (disappears) from the distant dephasing zone and is received (reappears) back at the transmitter

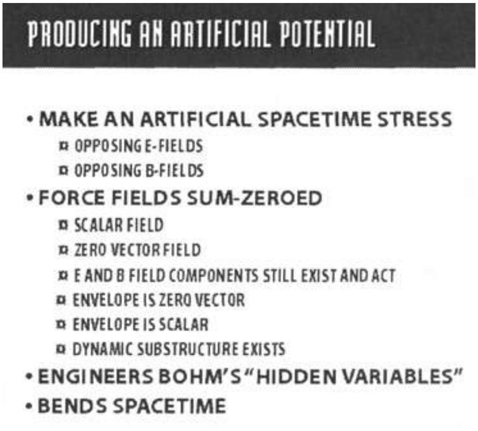

PRODUCING AN ARTIFICIAL POTENTIAL

- MAKE AN ARTIFICIAL SPACETIME STRESS
  - OPPOSING E-FIELDS
  - OPPOSING B-FIELDS
- FORCE FIELDS SUM-ZEROED
  - SCALAR FIELD
  - ZERO VECTOR FIELD
  - E AND B FIELD COMPONENTS STILL EXIST AND ACT
  - ENVELOPE IS ZERO VECTOR
  - ENVELOPE IS SCALAR
  - DYNAMIC SUBSTRUCTURE EXISTS
- ENGINEERS BOHM'S "HIDDEN VARIABLES"
- BENDS SPACETIME

Slide 018: Producing an Artificial Potential

AN artificial potential is deterministically patterned spacetime stress, made by opposing E-fields and/or B-fields so that they sum to vector zeros in a special pattern.

The resulting zero-summed envelope has no EM force field, to an external observer/detector.

However, the infolded E-field and B-field vector components still exist and act. They may dynamically vary, so long as their summation is always kept to zero.

The simplest variation is to vary all their magnitudes at once, by the same degree. In that case, each one comprises an "EM wave." However, the summations of this cluster or "locked group" of waves still exhibit a zero-E and zero-B field to any external observer/detector. In other-words, to an external observer, one now has a varying wave of pure spatiotemporal stress, but one that has a deterministic structure. This is a scalar EM wave, or electrogravitational wave. It is also an alternating current of specific scalar pattern.

Varying the stress of spacetime locally, curves it locally. This violates the conventional assumption of restricted general relativity that local spacetime is uncurved (is a Lorentz frame).

By use of scalar EM waves with deliberate substructures, one can engineer Bohm's "hidden variables" so that quantum mechanics becomes deterministic rather than statistical. This is a drastic change to the common (Bohr) interpretation of quantum mechanics.

And Einstein's intuition that God does not play dice with the universe turns out to be correct after all.

Since physicists haven't seen where the real game was being played, it has all seemed bewilderingly statistical to them.

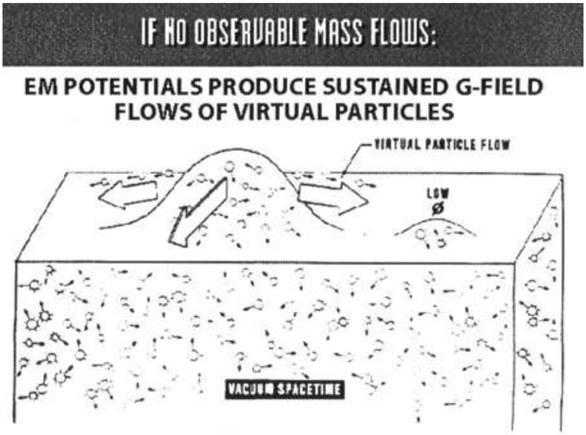

IF NO OBSERVABLE MASS FLOWS:

EM POTENTIALS PRODUCE SUSTAINED G-FIELD FLOWS OF VIRTUAL PARTICLES

Slide 019: If No Observable Mass Flows:

Since there are no observable mass particles flowing in vacuum, there are no observable force fields produced. Only virtual force fields can be produced in the virtual particle medium that is vacuum itself.

Thus (here is no observable E-field in the vacuum surrounding a charged point. There is, however, a virtual E-field there. When an observable charged particle such as an electron is introduced in the surrounding vacuum and "couples to" (integrates, zips together) the virtual E-field, an observable E-vector force is produced on and of the observable charged particle.

(For the classically trained electrical engineer or electrical physicist, the statement that no E or B fields as such exist in vacuum, is usually bewildering. Let me point out that any so-called "vector field" (such as the assumed E and B vacuum fields) can mathematically be replaced with two scalar fields. See E. T. Whittaker, Proc. Lond. Math. Soc.l, 366, 1903. What we are saying here is that the definition of an E field is in terms of force per unit charged mass. The scalar fields - which are actually what exist in vacuum - provide only a virtual or unzipped E field, until an observable spinning particle couples to both of them by virtue of its "dynamo flux pump" action. The ensemble, then, of two electrostatic scalar potentials coupled to a spinning observable particle, constitutes and is the so-called "E-field.")

When detecting these "vacuum E-fields," it is the free electrons in the electron gas in our probe or antenna which couple to the two "Whittaker scalar fields". These electrons produce ensembles that interact with each other collectively.

Charges in this electron-gas coupled medium is what we actually detect - not "what is in the vacuum."

To repeat this again for emphasis: The usual detector is an "electron-wiggle" detector. It detects changes in its own conduction electron gas, not in the vacuum itself. That is, the disturbance in the virtual-particle vacuum interacts with the observable particles of the electron gas, if (and only if) the spins of these observable particles couple the vacuum disturbance to the electrons and integrate their virtual components. After integration, an observable disturbance of the integrating object - say, the electron - results. It is this electron gas disturbance that is "detected" by almost all orthodox EM detectors.

If the spin of the observable conduction electrons cannot couple the electron to the vacuum flux disturbance, then the normal "simple" detector will not detect the disturbance. (By a "simple" detector we mean one in which the electrons are able to couple to the vacuum disturbance! Little circular definition here!)

Now most potentials reach to infinity before reducing to zero. The change in the virtual particle flux intensity of vacuum - which comprises the potential - decreases with distance from the potential. The magnitude of the change tails off" toward zero as one approaches infinity.

To any finite distance, then, there exists a decreased gradient in the change of the virtual flux intensity of vacuum, from a potential. Therefore virtual particles are flowing in the direction of the decrease of this gradient, just as gas molecules flow from a region of high pressure toward a region of lower pressure.

So there is a virtual particle "river" in the vacuum between any two separated points of different potential magnitude. The type of potential determines the type of virtual particle(s) in the river.

If the potential is natural, the gradient river does not carry any coherent substructure.

If the potential is artificial, the gradient river carries the coherent substructure everywhere within it. In this case, there is a flow or "river" of the virtual structure of this coherent pattern in the vacuum between two separated points of different "potential" in that pattern.

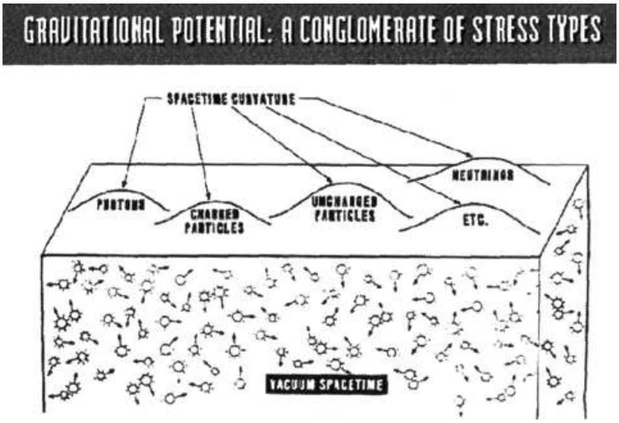
GRAVITATIONAL POTENTIAL: A CONGLOMERATE OF STRESS TYPES
SPACETIME CURVATURE
NEUTRINOS
PHOTONS
CHARGED PARTICLES
UNCHARGED PARTICLES
ETC.
VACUUM SPACETIME

Slide 020: Gravitational Potential: A Conglomerate of Stress Types

Rigorously, all EM potentials are actually part of the overall 5-dimensional gravitational potential.

There is no such thing as "the" singular gravitational potential: instead, it is a conglomerate of different stress types and patterns, each of which is named for the type of action that produces it.

The movement of charged particles, for example, creates electromagnetic potentials. At any point in vacuum, the potential consists of a flux of virtual particles.

The vacuum contains a great variety of particle fluxes. There are virtual photons, virtual electrons, virtual positrons, virtual protons, virtual pions, virtual neutrinos, etc.

The conglomerate stress from all these virtual particle stresses constitutes the overall gravitational potential

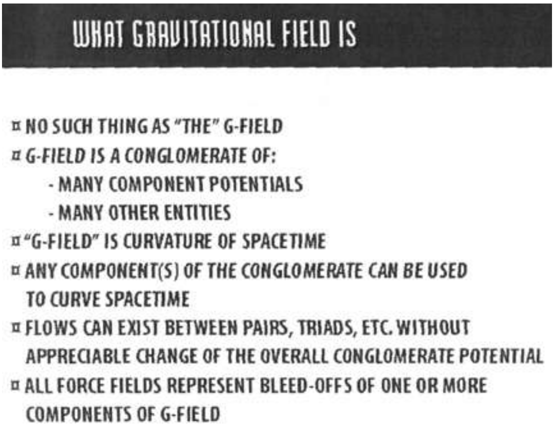

WHAT GRAVITATIONAL FIELD IS

¤ NO SUCH THING AS "THE" G-FIELD
¤ G-FIELD IS A CONGLOMERATE OF:
  - MANY COMPONENT POTENTIALS
  - MANY OTHER ENTITIES
¤ "G-FIELD" IS CURVATURE OF SPACETIME
¤ ANY COMPONENT(S) OF THE CONGLOMERATE CAN BE USED TO CURVE SPACETIME
¤ FLOWS CAN EXIST BETWEEN PAIRS, TRIADS, ETC. WITHOUT APPRECIABLE CHANGE OF THE OVERALL CONGLOMERATE POTENTIAL
¤ ALL FORCE FIELDS REPRESENT BLEED-OFFS OF ONE OR MORE COMPONENTS OF G-FIELD

Slide 021: What Gravitational Field Is

As previously stated, there is no such single thing as "the" gravitational field, as is assumed in the Newtonian sense.

G-field is a conglomerate of many component potentials and other entities.

Rigorously, anything that "curves spacetime" is a gravitational field.

In the modern view, all forces ultimately arise, by some means, from the curving of spacetime. Thus all forces and force fields are related to - and caused by, so to speak - gravitational field.

If the vacuum stress is uniform from one place to another in an observer's **frame** then the observer's spacetime is uncurved. The observer's frame, is said to be linear or uncurved; or a Lorentz, frame. In such a situation the conservation laws rigorously apply for macroscopic, closed systems.

If the vacuum stress is not uniform from one place to another, then the spacetime is curved. In such a situation the conservation laws may be violated for macroscopic systems, since they are "opened" to the vacuum stress, and vacuum virtual particle flux gradients exist across the system or portions of it.

Note that the overall stress magnitude may be constant across a region, but its composition may change. That is, two or more components may vary canonically. In that case, the spacetime is uncurved in the gross external sense, but contains specialized "internal" curvature patterns locally. Depending upon reaction to the canonical infolded curvature patterns, some macroscopic systems will still exhibit conservation, but certain other macroscopic systems may not.

An example is symmetry breaking. Here charge, parity, and time symmetry may be violated individually or in pairs, but not all three simultaneously. An overall conservation law is obeyed, but any two of the three conservation laws can be (und are) violated.

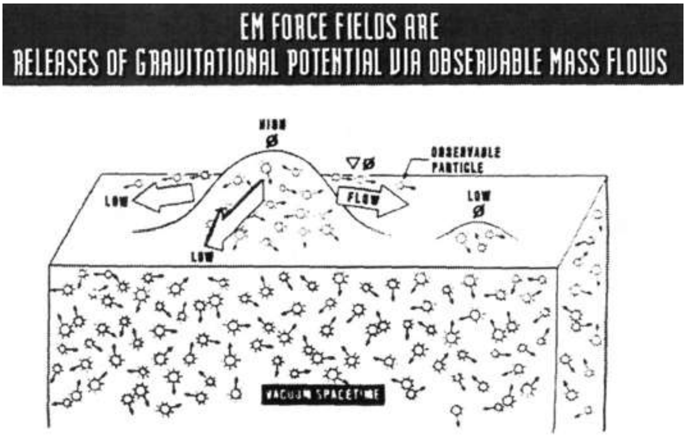

EM FORCE FIELDS ARE
RELEASES OF GRAVITATIONAL POTENTIAL VIA OBSERVABLE MASS FLOWS

Slide 022: EM Force Fields are Releases of Gravitational Potential via Observable Mass Flows

In Kaluza theory, we may take EM force fields to be the 5th dimensional "bleeding-off" of a 5-G potential. However, in vacuum alone this bleeding-off is totally within the virtual state. That is, only virtual EM force fields exist in vacuum.

By the ordinary term "EM force fields," physics has chosen to imply measureable or observable E and B force fields.

That implies that the 5-G potential gradient flows of virtual particle flux must be coupled to observable charged particles - such as electrons, protons, ions, etc. - before observable E and B force fields exist.

Rigorously, E and B force fields consist of the virtual E and B force fields coupled to observable charged particles. They do not "cause" a particle to accelerate; they are the product of the particle and its acceleration.

As these observable particles are themselves "spray pumps" producing virtual particle flux, then a collection of them represents an increased or decreased **flux** (for negative or positive charge) vis-a-vis the ambient background flux of vacuum. Nature furnishes the continuous power to drive these "virtual flux" pumps.

Normally, EM potentials are formed in natural and manmade electrical circuits by collections of these observable particle "spray pumps." Such collections of increased or decreased "spray flux" density, of course, represent 5-potentials, gravitationally.

Release of these pumps in conductors or in the vacuum itself results in creating E-fields and B-fields on and of the moving observable charged particles. But release of the observable-particle spray-pumps from the collection (the potential "pressure head" or "source") reduces the potential or "pressure" at the source.

Thus the formation of E and B fields represents the release of the 5-potentials. Hence EM force fields are the bleed-off of 5-dimensional G-potential via observable mass flow.

Since classical EM always seeks to create the force fields, it always bleeds-off and discards the gravitational effects that could be otherwise achieved.

To allow orthodox EM flow is to dissipate the gravitational effects.

Strangle the orthodox EM flows, and you can keep and use the amplified gravitational effects.

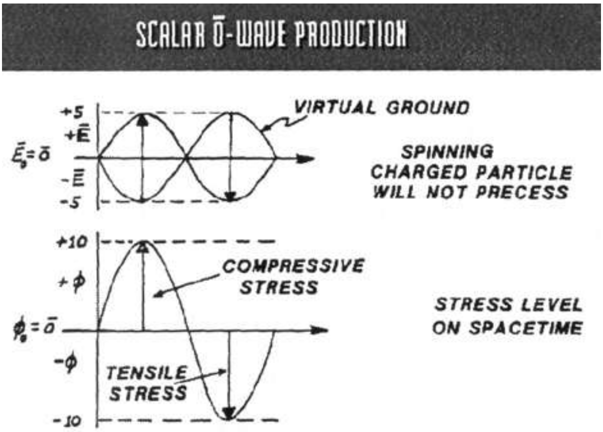
SCALAR Ō-WAVE PRODUCTION
+5
+Ē
Ē₀ = ō
-Ē
-5
VIRTUAL GROUND
SPINNING
CHARGED PARTICLE
WILL NOT PRECESS
+10
+φ
φ₀ = ō
-φ
-10
COMPRESSIVE
STRESS
TENSILE
STRESS
STRESS LEVEL
ON SPACETIME

Slide 023: Scalar O-Wave Production

We now visualize the formation of waves of pure stress in the spacetime medium (in the vacuum). These we call scalar EM waves, Tesla waves, electrogravitational waves, longitudinal EM waves, waves of pure potential, electrostatic/magnetostatic waves, and zero-vector EM waves. All these terms are synonymous. Each sheds its own particular light upon the nature of these waves or of their original discoverer, Nikola Tesla.

We use a gedanken experiment, or thought experiment, in which we can be "perfect."

Thus we have two single-frequency EM sine waves whose E-field componenls are shown on the slide. The two waves are of the same frequency, traveling together in the same direction, and superposed 180 degrees out of phase with each other.

In this case, at any spatial point, the summation E and B fields are vector zeros.

However, if we plot the spatiotemporal (vacuum) stress induced by the wave, wo see that it constitutes a sine wave, with compressive stress in one half cycle and tensile stress in the second half cycle.

Rigorously this defines a "longitudinal" EM wave - or what Tesla called a "sound wave in the nonmaterial ether."

Also, rigorously it is a gravitational wave, for it is a wave of the curvature of spacetime (nonlinearity of vacuum) itself. That is, what is changing in the wave is the vacuum stress, or intensity of the virtual particle flux of vacuum. That is identical to changing the curvature of spacetime. In one half-cycle, spacetime is curved positively. In the second half-cycle, spacetime is curved negatively.

Since the increase or decrease of the intensity of virtual particle flux (vis-a-vis the ambient vacuum's virtual flux intensity) represents electrical charge, then in one half-cycle negative charge is represented, and in the other half-cycle positive charge is represented. This directly explains the "positron-electron pair" representation of an electromagnetic photon - the "photon" being one wavelength.

In one half-cycle, time moves slower. In the second half-cycle, time moves faster. Thus scalar waves can also be considered to be "tempic" waves (to use Wilbur Smith's term), or oscillations of the rate of flow of time itself, about the ambient rate of time flow. Since these oscillations are variations in the curvature of spacetime, then they represent gravitational or force-generating waves, when coupled to a mass.

Variations in the rate of flow of time produce force, just as any other type of curvature in space-time does.

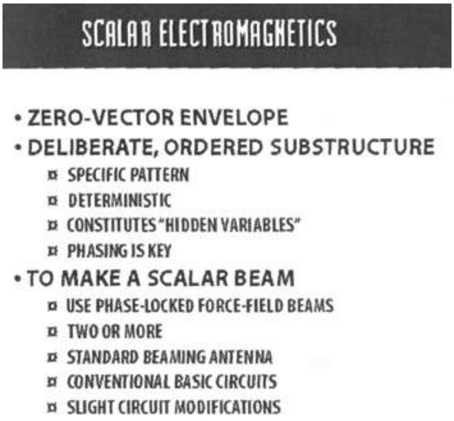

SCALAR ELECTROMAGNETICS

- ZERO-VECTOR ENVELOPE
- DELIBERATE, ORDERED SUBSTRUCTURE
  - SPECIFIC PATTERN
  - DETERMINISTIC
  - CONSTITUTES "HIDDEN VARIABLES"
  - PHASING IS KEY
- TO MAKE A SCALAR BEAM
  - USE PHASE-LOCKED FORCE-FIELD BEAMS
  - TWO OR MORE
  - STANDARD BEAMING ANTENNA
  - CONVENTIONAL BASIC CIRCUITS
  - SLIGHT CIRCUIT MODIFICATIONS

Slide 024: Scalar Electromagnetics

To rehash: scalar waves are zero-vector waves. Patterned E and B fields exist in a relationship where their summation equals vector zero. Electromagnetically, the zero EM "envelope" of a scalar EM wave is composed of a zero-summation of infolded finite EM force-vectors.

However, the internal (infolded) substructure is deliberate, macroscopic, and ordered. Vacuum is polarized deterministically by the wave, and macroscopic spacetime is locally curved by it.

These substructure components constitute Bohm's "hidden variables." They can be manipulated and varied at will. Phasing, beaming, frequency, superposition, interference, resonance, and Fourier expansion are the keys to scalar EM engineering.

To make a scalar beam, special modifications are made to the EM wave transmitter so that, effectively, the transmitter transmits multiple transmissions simultaneously, and these vectorially sum to zero. This is the same thing as transmitting multiple phase-locked EM force-field beams simultaneously, in **the** sum-zeroed fashion.

Standard beaming antennas and conventional circuits can readily be utilized with slight modifications.

Any radar can be converted to a scalar EM radar quite simply. Indeed, a modification kit to accomplish that can be developed and held until the **scalar** mode is desired. The modification kit can be quickly installed to provide the scalar capability.

Scalar radars, for example, cannot track "corner reflector" targets. However, they can magnificently track smoothly filleted corners; rounded, smoothly curved surfaces; and curved or flat metalized dielectrics. Against aerial targets using special construction to lower their cross-section to normal radars, the application of the scalar radar mode is immediately obvious.

Scalar waves pass through the electron shells of an atom and interact with the nucleus. They are continually absorbed and emitted by all nuclei in the universe.

BACKGROUND FOR THE BRIEFER.

Any large collection of nuclei - such as a star or a planet - is a strong absorber and radiator of scalar wave radiation.

The Sun is a particularly strong source of scalar radiation. This radiation penetrates the Earth deeply, interacting more and more with the deeper layers which, under greater mechanical stress, are more nonlinear. Most of the heat in the molten core of the Earth comes from the dephasing of a portion of this absorbed scalar radiation from the Sun, liberating ordinary EM energy as heat.

The Earth also re-radiates scalar wave radiation back to the Sun. The Sun and Earth are thus coupled into a "scalar" system in equilibrium or near-equilibrium. Each body in the couplet possesses both a feed-forward and a feedback loop.

The Earth also radiates scalar wave radiation to the Moon, and the Moon reradiates scalar radiation back to the Earth. The Earth and Moon also are coupled into a "scalar" system in equilibrium or near-equilibrium. Again, each body in the couplet possesses both a feed-forward and a feedback loop.

The Sun and Moon are correspondingly coupled into a scalar system in equilibrium or near-equilibrium, each with both a feed-forward and a feedback loop.

The Earth, Sun, and Noon thus form a triad-coupled system of special importance to the stability of our existence here on Earth, and the stability of our biosphere.

And so on with all the other planets in the solar system, and combinations thereof.

The solar system is thus a very sensitive scalar-coupled system, with strong subsystems within it.

ADDED FOR THE BRIEFER:

Actually this scalar coupling of the solar system provides a major check on unrestricted use of the large Soviet strategic scalar EM weapon systems. If significant scalar effects are produced on earth in a "pulse" mode, pulsed disturbance of the earth-sun and earth-moon systems results. Here a danger exists that one or more natural resonances of the coupled system may be excited. If the feedback stimulation of the Sun is not insignificant, for example, large sunspot activity may result sometime thereafter, say in a day or two. If too much or too sharp stimulation occurs on earth, the coupled resonant response from the sun could be disastrous. Indeed, a solar response could be stimulated so that the Sun would violently belch and destroy our biosphere, among other effects. The simplest doomsday stimulation would be for a violent expulsion of Solar EM energy and particles to occur. If this were due to resonance, the expulsion of Solar EM energy and particles would continue during some decay time. In that case, fiery destruction of the earth, strongly indicative of Biblical prophecy, would result. Particularly sensitive are the resonances of the sun-earth, sun-moon, and earth-moon systems.

Note the abnormal influence of the moon on tides - tides of both the tectonic plates and the oceans. One cannot help but point out that, if the earth-moon resonant system were overstimulated, one might expect violent earthquakes of extraordinary magnitude, and tremendous tidal waves hundreds of feet in height. (The potential connection to legends of earth-destroying floods, evidences of extreme water levels in mountains, and legends of ancient cataclysmic destruction of supposedly advanced technological oceanic cultures is obvious.)

Accordingly, use of huge scalar EM weapons is a double-edged sword. Unless carefully employed, use of the weapons could cause a terrible backlash to the user as well as the victim, and even accidentally cause the destruction of the earth itself. It is not accidental that in 1960 Khrushchev stated that his new fantastic weapon could - if unrestrainedly used - destroy all life on earth.

This appalling backlash potential is apparently what prompted Brezhnev in 1975 to make so strong an effort to get agreement to outlaw the development of "new weapons of mass destruction" more frightful than the mind of man had even dreamed of. Gromyko even introduced such a draft agreement into the proceedings of the United Nations. The first article of the draft agreement indicated that the nature of the weapons referred to would be negotiated. Sadly, the entire Western world did not even know what the Russians were talking about.

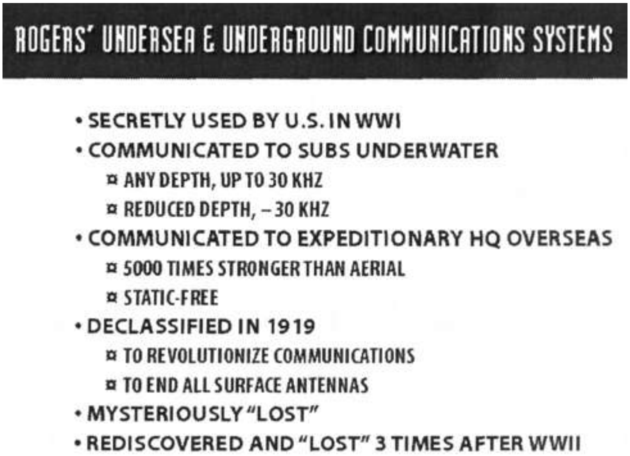

ROGERS' UNDERSEA & UNDERGROUND COMMUNICATIONS SYSTEMS

- SECRETLY USED BY U.S. IN WWI
- COMMUNICATED TO SUBS UNDERWATER
  - ¤ ANY DEPTH, UP TO 30 KHZ
  - ¤ REDUCED DEPTH, – 30 KHZ
- COMMUNICATED TO EXPEDITIONARY HQ OVERSEAS
  - ¤ 5000 TIMES STRONGER THAN AERIAL
  - ¤ STATIC-FREE
- DECLASSIFIED IN 1919
  - ¤ TO REVOLUTIONIZE COMMUNICATIONS
  - ¤ TO END ALL SURFACE ANTENNAS
- MYSTERIOUSLY "LOST"
- REDISCOVERED AND "LOST" 3 TIMES AFTER WWII

Slide 025: Rogers' Undersea & Underground Communications Systems

In an article in the March 1919 issue of the Electrical Experimenter, a real bombshell was released on an unsuspecting world.

In that issue, details of the Rogers'undersea and underground communication systems were released. These systems had been classified SECRET during WWI, and used to communicate with (1) U.S. submarines underwater around the world, and (2) U.S. Expeditionary Forces Headquarters overseas.

Subs at any submerged depth could be reached with transmitters of frequencies up to 30 kiloHertz. At reduced depths, frequencies above 30 kHz could be used.

At the headquarters of U.S. overseas forces, the signals were received through the earth. They were 5,000 times stronger than aerial signals through the atmosphere, and static-free.

The Rogers' systems were covered by a series of patents. They used special antennas implanted in the ground.

While the patents are somewhat confused, several of the diagrams in the Electrical Experimenter reveal that the antennas were producing scalar waves. Scalar waves can indeed travel through the earth and through the ocean. Indeed, if good quality scalar waves are made, then frequencies well above 30 kHz can be sent through the ocean. Multi-megahertz frequencies, for example, are easily achievable. (The problem, of course, is that these waves are pure zero-vector artificial potentials that do not move electrons in the conduction gas of a conventional antenna/receiver combination. Normal receivers and detectors do not see them. Special underwater detectors - such as those developed by Frank Golden - are required.).

With the startling declassification and revelation of the Rogers' systems, it was expected that all communications would be revolutionized, and that all surface antennas would be removed.

Strangely, this entire technology faded away quickly, as if it had never been. Certain powerful interests did not wish the extended electromagnetics to be discovered or taught. That, after all, would lead to free energy and loss of economic control of the citizens. That was precisely why Nikola Tesla had already been suppressed.

According to the important research of Dr. Robert Beck, the same scalar technology has been rediscovered and mysteriously "lost" at least three times after World War II!

Today the orthodox establishment is still suppressing it.

One wonders why no orthodox physicist - with the single exception of Dr. Hooper (see U.S. patents 3,610,971, Oct. 5, 1971 and 3,656,013, Apr. 11, 1972) - seems to have experimented with deliberately constructed zero-vector EM waves and devices, and published his results in the literature.

One would think that any university physics department would wish to explore such a gaping void with at least a few graduate student experiments, if for no other reason than scientific curiosity, particularly when the experiment is easily performed, it is totally different from anything else in the theory or literature, and the implications may be formidable. A true scientist in that related field ought to try it!

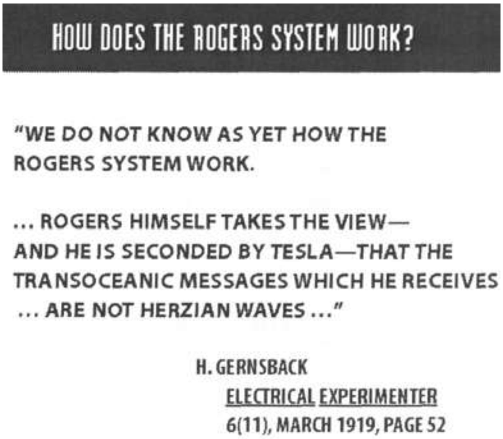

HOW DOES THE ROGERS SYSTEM WORK?

"WE DO NOT KNOW AS YET HOW THE ROGERS SYSTEM WORK.

... ROGERS HIMSELF TAKES THE VIEW—AND HE IS SECONDED BY TESLA—THAT THE TRANSOCEANIC MESSAGES WHICH HE RECEIVES ... ARE NOT HERZIAN WAVES ..."

H. GERNSBACK
ELECTRICAL EXPERIMENTER
6(11), MARCH 1919, PAGE 52

Slide 026: How does the Rogers System Work?

In the same March 1919 issue of Electrical Experimenter, Hugo Gernsback pointed out that no one as yet knew how the Rogers' system actually worked.

Rogers himself believed he was not using conventional Hertzian waves.

Nikola Tesla himself confirmed that the underground and undersea waves were not Hertzian waves.

If so, they could only have been scalar waves.

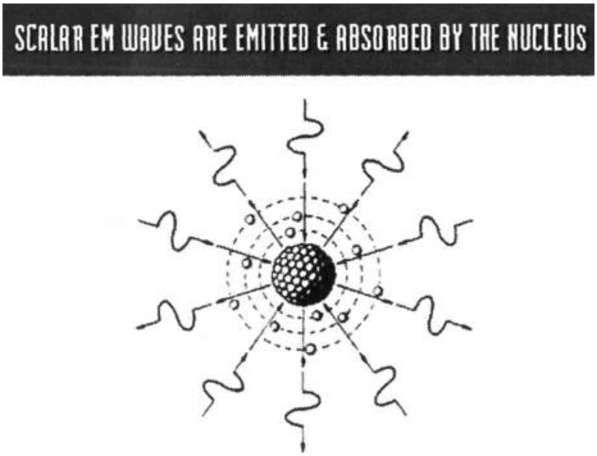
SCALAR EM WAVES ARE EMITTED & ABSORBED BY THE NUCLEUS

Slide 027: Scalar EM Waves are Emitted & Absorbed by the Nucleus

Normally scalar EM waves do not couple to orbital electrons. Entering into the atom, incoming scalar EM waves penetrate the electron shells without interaction. They continue, penetrating into the interior of the closely packed nucleus, where the extreme nonlinearity of the charged virtual particle flux that binds the nucleons together distorts the scalar wave, dephasing its summed EM vector components. This produces a nonzero EM resultant and polarized pattern directly inside the virtual particle charge flux of the nucleus.

The new EM resultant immediately couples the distorted scalar wave to the nucleus, resulting in absorption of the wave and incorporation of its polarization pattern into the nucleus' virtual flux. This in turn results in a slightly excited nucleus. Normally the excited nucleus will promptly decay by emitting another scalar wave, which passes through the orbital electron shells and radiates out into the universe.

However, as can be seen, if the same strong scalar pattern is constantly radiated into the nucleus and absorbed by it, the excitation of the nucleus "charges up," leading to decay by other modes. In this manner the nucleus can be engineered even transmuted - by rather minuscule power of the engineering radiation signal.

This is how living biological systems are able to transmute elements, as discovered by Louis Kervran. (Kervran was nominated for the 1977 Nobel Prize for this work.). The biosystem produces scalar EM waves. If the appropriate signal is employed into and on the appropriate nucleus, a changed element emerges. In this fashion chickens denied sufficient calcium in their diet, but allowed a surplus of potassium, are able to transmute some of the potassium into calcium. Note that one of the isotopes of potassium is an isomer of normal calcium. Transmutation of this potassium isomer to calcium is easiest.

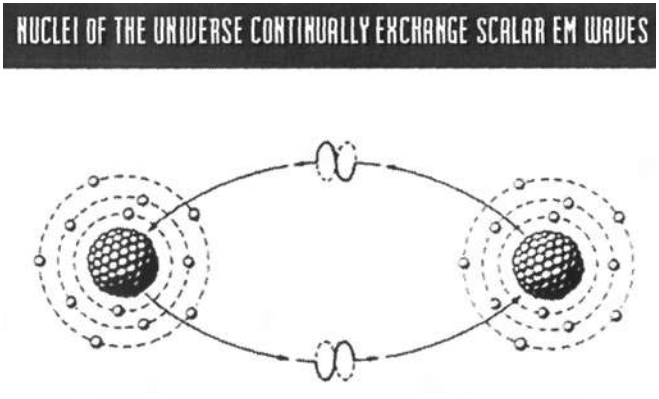

NUCLEI OF THE UNIVERSE CONTINUALLY EXCHANGE SCALAR EM WAVES

Slide 028: Nuclei of the Universe Continually Exchange Scalar EM Waves

Thus scalar EM waves are continually exchanged by all nuclei of the universe, and the vacuum is - among other things - a seething cauldron of scalar radiation.

In fact, it is this seething scalar wave cauldron that creates the virtual particles and the virtual particle flux of vacuum itself. (Since the scalar EM wave represents an oscillation of the curvature of spacetime, and since everything - virtual or observable - is a curvature of spacetime, then scalar waves can be taken as the universal generatrix.

The intensity of the trapped scalar flux in the nucleus of an atom is responsible for the mass and inertia of the nucleus.

Mass is localized trapped scalar resonance.

The trapping agent is the spin of a particle.

The mass and inertia of the particle are due to the scalar resonance that is trapped. Increasing the amplitude of the trapped scalar resonance increases the mass and inertia; decreasing the amplitude decreases the mass and inertia.

CONVERTING EM FIELD ENERGY TO G-FIELD ENERGY & VICE-VERSA

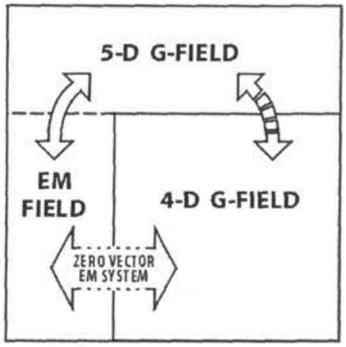

Slide 029: Converting EM Field Energy to G-Field Energy & Vice-Versa

According to Kaluza theory, there are no such things as a separate EM field and a separate gravity field. Instead, in five dimensions there is only one field: the 5-d gravitational field. The EM field is the 5th dimensional aspect, while our **normal** 4-dimensional G-field is the other aspect of the 5-field, in our normal 4-space.

Thus we may say that the 5-field is composed of two components: the 5th dimensional component (our normal EM field) and the component occupying **our** normal 4 dimensions (our normal G-field).

The 5-d "force" field, of course - or the 5-space analogy to a force field - would be due to a gradient or "bleed-off of the 5-d gravitational potential. This gradient in turn is composed of two components: the 5th dimensional "bleed-off" outside our normal 4-space (this outer bleed-off is our normal EM force **field**) and the bleed-off inside our normal 4-space (this inner bleed-off is our normal G-field).

Normally the 5-potential bleeds-off outside our 4-space as EM field, far greater than it bleeds-off inside our world as G-field. Between two electrons, for example, the electric field is about 10^{42} times as strong as the G-field. Since in the Kaluza view there is only one 5-d G-potential that is causing both force fields, this shows that there is an incredibly greater EM bleed-off of the 5-

potential between the two electrons than there is a G-field bleed-off between them.

As shown on the diagram, the 5-d G-field is normally comprised almost entirely of the 5th dimensional EM field. Only a small 4-space G-field component exists.

However, suppose we were to "block" the bleed-off of the 5-potential in the EM mode. Then none of the 5-potential could bleed-off in the 5th dimensional EM field. Instead, it would be forced to bleed-off into the 4-space G-field. In our two-electron example, this "perfect case" would result in the disappearance of the E-field between the two electrons, and the G-field between them would increase to about 10^{42} times its normal strength.

The end result would be that, by blocking the EM force field bleed-off, EM field is converted to G-field. In addition, EM field energy is converted to G-field energy. With the extremely amplified gravitational and inertial effects that result, one can now accomplish direct engineering of gravitation, mass, and inertial effects.

We can effectively accomplish this "blocking the EM bleed-off by opposing EM force fields so that they sum to vector zero. This is the same as summing various 5th dimensional gradients of the 5-potential to a zero vector resultant. In that case, as much "EM bleed-back to 5-potential" occurs as there is "EM bleed-off from 5-potential." This places the 5-potential in equilibrium with respect to EM bleed-off.

Also, another nice thing results: Since we can readily vary the magnitudes of the EM force field components of the vector zero summation, we can actually form phased "zero vector waves" by increasing and decreasing the magnitudes of all the component EM vectors in phase, but retaining their vector summation always equal to zero.

In that case we have produced a very simple 5-space G-potential wave, which concomitantly forces phase-locked variations in the 4-space G-potential. In short, we have produced an electrogravitational (EG) wave, or scalar EM wave for short. This wave described is the simplest EG scalar wave we can produce; much more complex EG waves can be produced and used for highly **specialized** purposes.

The scalar EG wave changes EG potential energy into 4-gravity potential energy in one half-cycle, and changes 4-G potential energy into EG potential energy in the other half-cycle. However, the EG potential energy in the first half-cycle does not react electromagnetically in linear circumstances, since it is electromagnetically a linear vector zero.

So our scalar EG wave actually oscillates energy back and forth between a locked-in 5th dimensional EM potential and a 4-space G-potential. As can be seen, to the linear 4-space observer this scalar EG wave appears as a purely gravitational wave, with G-potential magnitude varying in a wavelike manner.

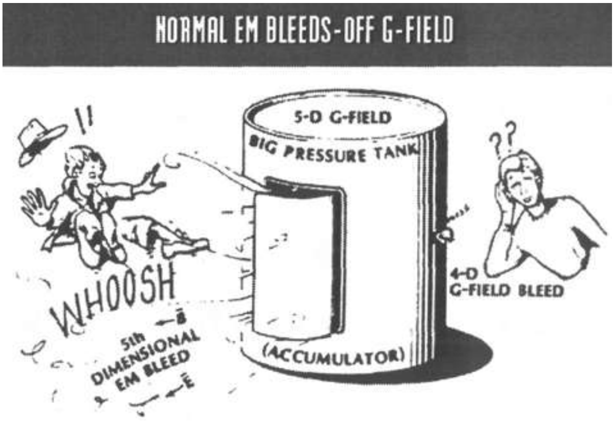
NORMAL EM BLEEDS-OFF G-FIELD
5-D G-FIELD
BIG PRESSURE TANK
(ACCUMULATOR)
WHOOSH
5th DIMENSIONAL EM BLEED
4-D G-FIELD BLEED
!!
??

Slide 030: Normal EM Bleeds-Off G-Field

For easy visualization, we may regard the situation in the analogy shown on the slide.

The 5-potential of 5-space vacuum is like a high pressure in a big tank. The tank has a huge open door in the 5th dimension, leading to rapid bleed-off of the 5-presure through the big door; that is, as EM field. Only a very small hole in the tank normally exists in our own 4-space world, so only a very small amount of the pressure bleeds off as our G-field.

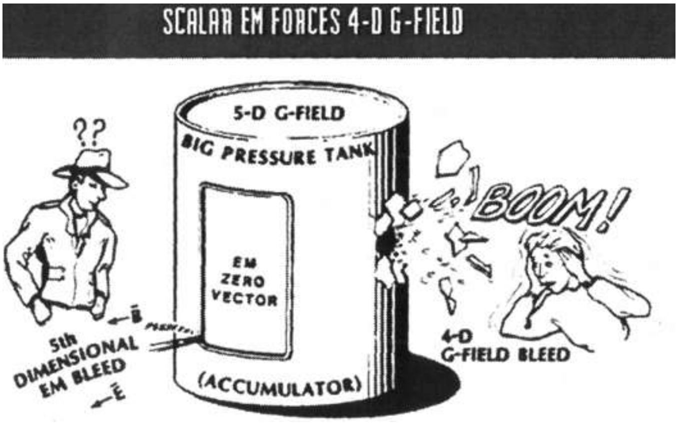

SCALAR EM FORCES 4-D G-FIELD

5-D G-FIELD
BIG PRESSURE TANK
EM
ZERO
VECTOR
(ACCUMULATOR)
??
5th
DIMENSIONAL
EM BLEED
$\leftarrow \bar{B}$
$\leftarrow \bar{E}$
BOOM!
4-D
G-FIELD BLEED

Slide 031: Scalar EM Forces 4-D G-Field

As can be seen however, if we shut the big door so that we have only a little bleed-off through the seals, then most of the pressure is forced to bleed-off through the small opening, greatly enlarging it. In that case, most of the 5-potential bleeds-off as 4-G force field.

The mass exposed to such increased 4-G bleed-off exhibits a much greater mass and inertia.

In fact, mass itself can act as an accumulator for the effect, and it can increase its ability to store trapped scalar resonance. This increases its inertia.

By the use of scalar EM (blocking the EM bleed-off), we can input normal EM energy and obtain powerful gravitational and inertial effects. Many other powerful effects can also be obtained.

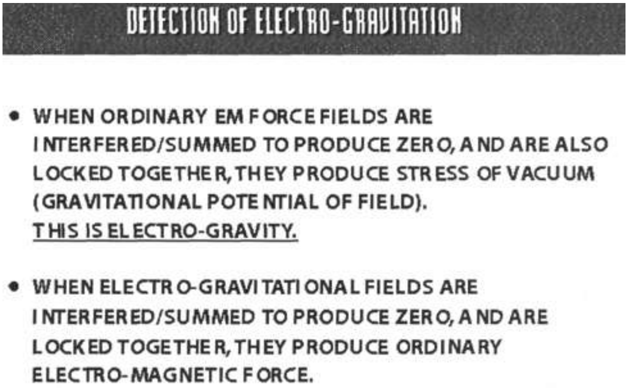

DETECTION OF ELECTRO-GRAVITATION

- WHEN ORDINARY EM FORCE FIELDS ARE INTERFERED/SUMMED TO PRODUCE ZERO, AND ARE ALSO LOCKED TOGETHER, THEY PRODUCE STRESS OF VACUUM (GRAVITATIONAL POTENTIAL OF FIELD). THIS IS ELECTRO-GRAVITY.
- WHEN ELECTRO-GRAVITATIONAL FIELDS ARE INTERFERED/SUMMED TO PRODUCE ZERO, AND ARE LOCKED TOGETHER, THEY PRODUCE ORDINARY ELECTRO-MAGNETIC FORCE.

Slide 032: Detection of Electro-Gravitation

Here arc two magic rules in the scalar electrogravitational business:

(1) When ordinary EM force fields are interfered or summed so that they produce vector zero, and the fields are locked together, they produce stress of vacuum (5-potential). One component of this 5-potential is the 4-space G-potential. Since bleed-off as a gradient of this 4-space G-potential is all that is allowed, then the EM destructive interference/zero summation bleeds-off to produce gravitational force field. This is because opposing equal forces, which sum to a vector zero force field resultant, actually produce a stress potential having stress energy of spacetime. And that is a curvature of spacetime, hence gravitational.

In short, total destructive interference of EM waves or force fields produces excess stress potential energy in spacetime, and therefore gravitational force field on detecting/coupled particles.

(2) When electrogravitational fields are interfered/summed to produce vector zero, this destructive interference/zero summation bleeds-off to produce ordinary electromagnetic force field.

In short, destructive interference of scalar EM waves or fields (electrogravitational waves or fields) produces electromagnetic force fields on detecting/coupled particles.

BACKGROUND FOR THE BRIEFER:

Mathematically we can summarize this as follows;

$$G_{(1-4)} = G_{(1-4)}(m) + G_{(1-4)}(EM) \qquad [3]$$

$$G_{(1-5)} = G_5 + G_{(1-4)} \qquad [4]$$

If $$\mathbf{E} = \Sigma \mathring{\mathbf{a}}_i E_i = \mathbf{0},\ E_i \neq \mathbf{0};\ \mathbf{B} = \Sigma \mathring{\mathbf{a}}_i B_i = \mathbf{0},\ B_i \neq \mathbf{0} \qquad [5]$$

then $$G_{(1-4)}(m) = G_{(1-4)}(EM) = G_5 = G_{(1-5)} \qquad [6]$$

and $$G_5 = G_{(1-4)} = f[\Sigma(\mathring{\mathbf{a}}_i\, |\mathbf{E}_i|^2)] + g[\Sigma(\mathring{\mathbf{a}}_i\, |\mathbf{B}_i|^2)] \qquad [7]$$

and $$G_{(1-5)} = G_5 = G_{(1-4)}(m) = G_{(1-4)}(EM) = f(\phi) + g(\mathbf{A}) \qquad [8]$$

Where:

$G_{(1-5)}$ is the 5-d G potential, or total G-stress on dimensions 1-5.

G_5 is the combined EM potential, or G-stress on dimension 5 only.

$G_{(1-4)}(m)$ is the normal 4-d G-potential of mass alone, or the G-stress due to mass-gravity.

$G_{(1-4)}(EM)$ is any extra 4-d G-potential of mass due to conversion from G_5 stress (normally zero).

E and B are the electric and magnetic force fields.

ϕ is the electrostatic scalar potential.

A is the magnetic vector potential.

Thus the age-old scientific dream of direct engineering of gravitational field — and of spacetime itself — can now become reality. To start, one simply makes an intense EM vector zero and varies its stress intensity and internal pattern of summed components. One also varies the reference potential at which the patterned EM stress zero is created, and the frequency of internal component variation.

BACKGROUND FOR THE BRIEFER.

He that has ears, let him hear.

Here is the secret of antigravity

The "electrical charge" of a charged particle — such as one of the orbital electrons of an atom or one of the protons of its nucleus — represents a difference in flux intensity (potential) between the local particle and its ambient vacuum.

It's the continual bleed-off of gravitational charge as electrical charge.

If that bleed-off is reversed or stopped, a drastic effect on the gravitational potential and charge ensues. The 5-potential and 5-charge have become 4-G-potential and 4-d-charge respectively.

Thus "charging" an object with its scalar EM pattern charges it gravitationally.

Now the only "bleed-off channel" is through the 4-G force field.

Further, in a nucleus the nucleons continually change back and forth between proton and neutron, so the electrical charge is "spread" throughout the nucleus and shared by all the nucleons.

Further, each element (actually each isotope) has its own unique "ensemble pattern" of Fourier expansion scalar frequencies, amplitudes, etc. This pattern, of course, can be reproduced artificially and transmitted by modified EM transmitters (i.e., by scalar EM transmitters). There is, however, a sort of "master key" scalar EM (EG) pattern for nucleons (protons and neutrons which are changing back and forth into each other by exchange of virtual charge currents).

If one "reverses the charge" by inverting this pattern, and then "charges up a mass with the inverted charge," to the external observer the charging mass just gets lighter and lighter, and its inertia gets less and less. Eventually it seems (to him) to acquire negative mass and negative inertia, and just accelerate away from the earth. The object "falls upward" instead of "falling downwards."

There are also some weird time effects. The object can be moving slower through time than the laboratory observer, or even be moving backwards through time with respect to the laboratory observer. (Don't believe everything they laught you in relativity; none of those professors had ever engineered a single general relativistic situation. Little that they teach in GR is based on direct experiment. Most of what they teach is already experimentally proven to be incomplete or in need of further extension.).

But back to our slide.

Inside the pattern itself, it's just a quite normal pattern. No change in or on the object seems to have occurred, to it, if the charge is evenly accomplished throughout the object. Indeed, to an internal observer it's the external environment that seems to have suddenly gotten very strange! This is what occasionally happens to a hapless aircraft or ship that experiences anomalous spatiotemporal difficulty in an *occasionally scalar-active* area such as the Bermuda triangle. Under the right conditions, the vehicle's gravitational charge is affected by anomalous changes in the scalar radiation from the earth in that area. To the passengers and crew, it is the external environment that suddenly seems weird. In addition, on-board EM and inertial instruments may be affected, and other electrogravitational effects may also occur, due to differing rates of charge in different parts of the vehicle.).

"Reversing or lowering the EG-charge" is controlled by means of biasing the ground potential on the ensemble pattern transmitters. These transmitters can even be on-board the vehicle itself. (Remember, in scalar EM to transmit is also to receive. A vehicle can change its own bias potential (with respect to local vacuum G-potential) by properly transmitting, which translates to "receiving potential charge." The charge current can either be negative or positive (its potential can be lowered or increased vis-a-vis that of the vacuum.) Think of this as handling positive and negative energy, which means positive and negative gravity.

By "fiddling" with this, you can float metal. Or a human body. Or a battleship. Or a high-speed vehicle containing a crew.

You can even "dematerialize it" or "teleport it."

The Philadelphia Experiment may have been real after all. If so, the test ship and its personnel were "blasted" into this strange realm, instead of making a gentle, controlled entry.

Suppose you materially lower or reverse (adjust) the gravitational charge (gravitation 5-potential with zero-summed EM to throttle EM bleed-off). At zero gravitational charge, to the external observer, such an assumed vehicle would seem to have no mass and no inertia. It is capable of extreme accelerations, right angle turns at full speed, etc. It is also right on the point of dematerialization, and may appear to be glowing or just a vehicle of light.

If the pilot wishes to "land," of course he must adjust the charge on the vehicle.

If the pilot wishes to "dematerialize" or "teleport," again he must adjust the charge on the vehicle.

If he wishes to go "hyperspatial," again he must adjust the charge on the vehicle and allow the proper bleed-off to produce a hyperspatial force. That's done with nested zero summations and nested patterns of scalar EM transmission.

That is, he can charge and bias the vehicle in multiple, nested zero-summations simultaneously.

That is necessary for hyperspatial control and travel.

Note that in lower hyperspace such a vehicle could do some very strange things. Such as "penetrate" solid matter (actually, it goes "around" the 3-space matter in a 4th Kaluza-Klein space).

We might expect to see such vehicles glowing. Their various surface features and mechanisms might appear to be glowing or revolving lights, etc.

They could exhibit incredible "aerodynamic performance," seemingly in the atmosphere. Actually they would not be moving "through" the atmosphere at all but through a higher space outside each particle of atmosphere.

They could seem to materialize and dematerialize.

They could seem to plunge into the ocean or rise out of it.

They could even seem to operate under the ocean or inside the earth itself.

Such anomalous vehicle performances have been seen all over the world, particularly since a few years after WWII.

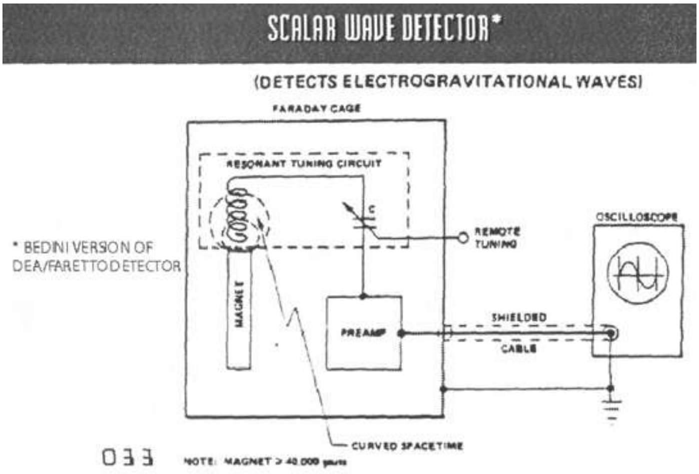

SCALAR WAVE DETECTOR*

(DETECTS ELECTROGRAVITATIONAL WAVES)

* BEDINI VERSION OF DEA/FARETTO DETECTOR

033

NOTE: MAGNET > 40,000 gauss

Slide 033: Scalar Wave Detector

On this slide we show a concept for a detector for scalar EG waves. This is the Bedini scalar wave detector, adapted from an original concept by Dea and Faretto.

The idea is quite simple: install a very powerful bar magnet vertically inside a grounded Faraday cage. Then install an open-ended coil longitudinally above the magnet so that a line through the longitudinal axis of the magnet passes through the longitudinal axis of the coil above it. The open end of the coil does not touch the magnet.

Connect the other end of the coil to a variable tuning capacitor, so that the coil and capacitor form a tunable, series-LC-oscillatory circuit. The output of the capacitor is connected to a transistorized preamplifier inside the cage. A tuning shaft for the capacitor is very carefully passed through a small hole in the shield to allow tuning from outside.

The output of the preamp passes very carefully through a small hole in the shield, through a shielded cable, to an adjacent oscilloscope. The cable shield in also grounded to reference ground potential.

The theory is as follows: Suppose a normal EM wave appears inside the cage, above the magnet or closely adjacent thereto. In that case a coupled **oscillation** appears in the field of the magnet, and this oscillation is coupled to the coil immediately above it. If the oscillation is within the bandwidth of the tuned series LC circuit, detection occurs. This is amplified by the preamp and passed to the oscilloscope, where it is displayed on the scope.

Note that the detector detects normal EM wave. Now our problem is: how do we get an EG wave to be detected? And how do we assure that we do not detect ordinary EM waves from outside?

Actually this is simple. Ordinary waves (except for quite low frequency) will be grounded out by the Faraday shield, and will not penetrate the cage. Thus these normal EM waves cannot enter the cage and appear above the magnet. They will not be detected.

Scalar EG waves, on the other hand, will enter the cage since they do not couple to conduction electrons in the metal of the cage. Above the pole of the magnet, spacetime is locally bent. After all, a pole is a magnetostatic scalar potential, which is a part of the conglomerate called "G-potential." The pole represents an increase (or decrease, depending upon whether it is a north or a south pole) in **the** magnetostatic component of the local G-potential. This is a curvature of spacetime. An EG wave entering this region adds a varying component of magnetostatic G-potential, which bleeds-off in the coil as an ordinary EM wave.

Another way to view the detector is to model the EG wave as a longitudinal wave, and an ordinary EM wave as a transverse wave. When the EG wave enters the curved spacetime region above the magnet, to the coil (the "observer" in this case**)** the longitudinal aspect of the EG wave appears to be rotating back and forth, so that an oscillating transverse component is present. This transverse component appears to the coil as an ordinary EM field, and so the LC circuit detects it if it is in the proper frequency band.

Yet another way to view the situation is to realize that an EG wave entering the magnet results in an oscillation component added to the magnetostatic scalar potential (pole strength) of the magnet. Accordingly, the magnet is a receiver **for scalar** waves, which are detected to "bleed-off as an oscillation of the magnetic field strength of the magnet. Coupling of this oscillating magnetic field **to the** coil creates an oscillating current flow into the capacitor. This oscillates the voltage input to the preamp, which amplifies and feeds the signal to the oscilloscope for display.

In fact rather like the latter view! It is important to remember that a magnetostatic potential and an electrostatic potential can be oscillated by a scalar EG wave. By placing a magnetic material inside a Faraday cage, the oscillation of the magnetostatic scalar potential (pole strength of the magnet) can be used as the EG detecting mechanism. By placing a chargeable material inside a Faraday cage, the oscillation of the electrostatic scalar potential can be detected. (Possible examples of the latter type detectors are given by Hodowanec, Radio Electronics, April 1986.)

Note also that one may detect waves at one reference level and not at another. Changing the bias on the zero-reference ground of the Bedini detector affects the detection. To look inside a normal EM carrier (such as the Woodpecker) and see what scalar signals are riding upon it, the carrier may be used to bias the reference ground of the Bedini detector.

To produce a spectrum analyzer, simply use additional series resonant LC tuning circuits in parallel (put multiple taps on the coil, and wire each tap to a separate tuning capacitor of different capacitance). Again, varying the zero reference level is important, as is varying the strength of the magnet.

Golden also has invented an excellent series of scalar wave detectors based upon quite different proprietary principles.

The bottom line is: we can indeed rigorously detect and measure scalar EG waves. When one considers the large sums of money presently being spent to bury large aluminum cylinders etc. in an attempt to detect gravitational waves, one wonders why some funds could not be released to a few unorthodox researchers by the National Academy of Sciences to fund the proven creation and detection of electrogravitational waves!

Note from TEB: I understand that it took Bedini about 600 attempts before he got the device to work. It was then handed over to a different party to produce, who made some "improvements," and the device then failed to work.

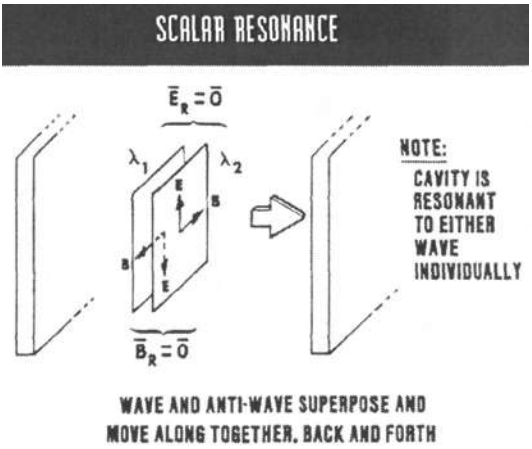

WAVE AND ANTI-WAVE SUPERPOSE AND MOVE ALONG TOGETHER. BACK AND FORTH

Slide 034: Scalar Resonance

Here we explain a new kind of resonance: scalar EH resonance, or electrogravitational resonance.

First, imagine we have a standard resonant cavity, represented by the two walls in our diagram. In this cavity we have a resonant EM wave moving back and forth, represented by the forward-most plane. In accordance with convention, we show the E-field vector and the B-field vector at right angles in this moving wave front. As the wave front moves back and forth, the vectors vary back and forth; however, at any one point between the walls, the two vectors always have the same value. Thus our resonant EM wave forms a standing wave in the cavity.

Now imagine that a second wave front, precisely like the first and of the same frequency, is superposed over the first wave front and travels along with it. This second wave - the "antiwave" - has its force vectors 180 degrees out of phase with the force vectors of the reference wave. Hence the E-fields and B-fields of the two superposed waves always sum to vector zeros, anywhere in the cavity. To an external observer, the cavity contains no ordinary electromagnetic force fields, hence no ordinary EM energy.

However, the energy density of a single EM sine wave in vacuum is given by

$$[E^2+B^2]/8\pi \qquad [9]$$

This energy density is always positive. Hence the energy density of the two waves at any point x between the walls is equal to

$$[E^2+B^2]/4\pi \qquad [10]$$

where

$$E = E(x) \; ; \; B = B(x) \qquad [11]$$

Thus the energy density of vacuum varies with x. But, rigorously, since the resultant E and B fields are zero, this describes a standing gravitational wave. Hence we have a standing EG wave, existing in the resonant cavity. This is an example of scalar resonance. Rigorously the cavity has mass and inertia, to an outside observer, as a result of the two warps in spacetime it contains.

Note that in one half-cycle the energy density of vacuum is greater than ambient, and in the other half-cycle it is less. In the region of one half-cycle, time flows at a faster rate than to the ambient observer, and in the other half-cycle time flows at a slower rate than to the ambient observer.

One half-cycle appears to contain negative electrical charge, and the other appears to contain positive charge.

One half-cycle appears to contain a north pole (positive magnetostatic scalar potential), and the other half-cycle appears to contain a south pole (negative magnetostatic scalar potential).

Perhaps now one can begin to understand why a continuously accelerated orbital electron in the atom does not radiate EM energy, completely in violation of Maxwell's equations. The electron is naught but a complex aspect of a standing scalar resonance, existing between the nucleus and the orbit.

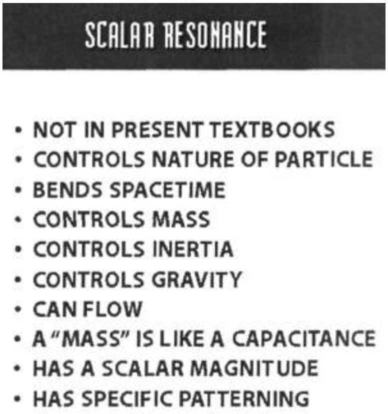

SCALAR RESONANCE

- NOT IN PRESENT TEXTBOOKS
- CONTROLS NATURE OF PARTICLE
- BENDS SPACETIME
- CONTROLS MASS
- CONTROLS INERTIA
- CONTROLS GRAVITY
- CAN FLOW
- A "MASS" IS LIKE A CAPACITANCE
- HAS A SCALAR MAGNITUDE
- HAS SPECIFIC PATTERNING

Slide 035: Scalar Resonance

Scalar resonance is not in the conventional textbook.

Scalar resonance is a particular zero-summed multi-resonance, electromagnetically, so that it does not act in an electromagnetic manner.

A scalar resonance is a standing electrogravitational wave. It can be made e**lectrically,** but it is not electrical in behavior.

In any scalar resonance, spacetime is curved, and it is the magnitude (and direction; of this spacetime curvature that is oscillating in "standing wave" fashion. In respect to stress of the vacuum medium, one half of a standing sine wave of scalar resonance is tensile; the other half is compressive. However, this is with respect to the local ambient stress of vacuum.

"Mass" of a particle is just a characteristic exhibited by a trapped scalar resonance. Usually this trapping is done by the "spin" of the individual particle.

The concept of "mass" may be compared to the concept of "capacitance." That is, a mass is an accumulator for scalar waves; that is, for scalar resonances. It is continually being "charged" and "discharged" by absorption and emission of sca**lar** waves from and to the ambient vacuum scalar wave flux.

Indeed, the magnitude of a mass may be defined in terms of the absolute value of this "switching" (absorption = switch in; emission = switch out) rate. In a very early paper I defined the mass of an object as proportional to the absolute value of the sum of the input and output switching rates, where each "switch" represents a change of action of h/(4). (Bearden, Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena, Mar. 1973, AD 763210, available through NTIS.) In that scheme, one kilogram of mass equals a switch rate of 17.053 x 10^{50} switches per second. An object is in a Lorentz frame when the switch-in and switch-out rates are equal in a given direction. If, in addition, the in-out rates in all directions are constant, the object is "stationary" with respect to a Lorentz observer. If not constant in all directions, the object is in a curved spacetime with respect to a Lorentz observer. In that case either the object's local spacetime is warped or the object is in a generalized, accelerated frame.

In the normal linear spacetime, the "charging" and "discharging" are equal in all directions; hence the "mass" exhibited in any direction is the same.

With respect to an "external observer's" equilibrium flux, a moving object encounters an increased amount of flux rate along its line of motion, just as a moving object in a rainstorm encounters more raindrops per second in the direction of its path, than does a stationary object. Encountering increased scalar wave flux (with respect to the external observer) forces the moving object to absorb and emit scalar waves at a higher rate along its direction of motion than when it is not moving. Thus to the external observer, the mass of the object has increased, insofar as any disturbing force along its line of motion is concerned.

At right angles to its line of motion, however, the flux rate is precisely the same as when the object is at rest. Therefore the "mass" of the moving object with respect to any disturbing force at right angles to its line of motion has not changed, as seen by the external Lorentz observer.

Thus is explained both parts of one of the alltime great mysteries of special relativity: (1) how the mass of an object increases with respect to its motion, and (2) why the mass only increases with respect to its line of motion, and not at right angles to it.

The inertia of a particle is due to its mass, i.e., to the total magnitude of its trapped scalar resonance.

The gravitational attraction between two masses is due to their spacetime curvature. Further, mass is like a capacitance. It can trap additional scalar waves as trapped resonance; hence increase its magnitude or "inertial charge." Or, it can discharge more scalar waves than it absorbs; hence decrease its magnitude or "inertial charge."

Those scalar waves emitted from resonance are emitted as a pattern ensemble of the resonance. Hence they may be regarded as constituting a current of scalar resonance leaving the mass-accumulator-object. Scalar waves absorbed into trapped resonance may be regarded as constituting a current of scalar resonance into the mass-accumulator-object. Thus it is proper to speak of scalar resonance as being able to "flow."

Further, it is possible to increase the mass of an object directly, by transmitting scalar EM waves to it so that it absorbs them. (By "absorbing" scalar waves one means that more enter the object than leave it, so the object acts as an inertial accumulator charging up with inertial charge. This is done by insuring that the reference potential of the scalar wave transmitter is higher than the reference potential of the irradiated object.)

It is also possible to decrease the mass of an object directly, by transmitting scalar EM waves to it so that it emits more than it absorbs. (By "emitting" scalar waves one means that more leave the object than enter it, so the object acts as if it were an inertial accumulator that is discharging its inertial charge. This is done by insuring that the reference potential of the scalar wave transmitter is lower than the reference potential of the irradiated object.)

As can be seen, the scalar wave "transmitter" is actually somewhat comparable to a heat pump; it can either act as an "energy transmitter" or as an "energy extractor," depending upon the difference in potential between "transmitter" and "receiver."

Finally, scalar resonance can have a specific pattern: both in frequency and in spatial curvature aspects, as well as "rate of flow of time" aspects. Indeed, scalar-wise every object has its individual "scalar pattern" which is a unique fingerprint. Since that print is spatiotemporal, it is a product of the object's entire past history. Thus - scalar-wise - no two objects are identical.

This brings up another rather amazing potential: If a reasonably precise scalar pattern of an object can be "irradiated" and resonated with scalar waves, energy may be created in or extracted from the distant object, just as stimulating one tunning fork can excite another at a distance by sympathetic resonance. We leave it as an exercise for the reader to ascertain the relevance of this statement to clairvoyance, radionics, remote viewing, etc.

For the skeptic, however, we must point out that - rigorously - quantum mechanics requires that, continually, in any localized region of spacetime, the "pattern" of any object in the universe appears momentarily in the virtual state, purely from statistical considerations alone. So long as the situation remains statistical, any place in the universe can continue to have the "ghosts" of everything present there, in an ethereally thin pattern, and the observable world will not be affected by it. However, if one can discriminate and scalarly "charge" or "discharge" individual patterns in this "ghost-realm," action-at-a-distance is directly possible, as is materialization and dematerialization. If one accepts that even thought itself produces such virtual "ghost-patterns" in the virtual particle flux of vacuum, then it is at least theoretically possible to even materialize thoughts and thought images.

We have entered a new kind of reality where the old rules and the old limitations do not necessarily apply.

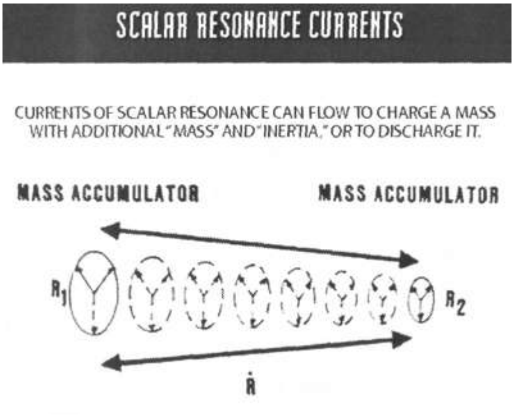

SCALAR RESONANCE CURRENTS

CURRENTS OF SCALAR RESONANCE CAN FLOW TO CHARGE A MASS WITH ADDITIONAL "MASS" AND "INERTIA," OR TO DISCHARGE IT.

Slide 036: Scalar Resonance Currents

As we pointed out, we can often greatly simplify matters by considering currents of scalar resonance. These currents flow from higher potential to lower potential, regardless of whether we are considering "transmission" or "reception."

Indeed, to "transmit at lower potential" is to receive, and to "receive at higher potential" is to transmit.

Thus the "transmitter-receiver" is a special system where simply biasing the two nodes differently determines which way the scalar resonance current will flow.

We may increase or decrease an object's inertia and mass, simply by properly biasing the transmitter-receiver system's two nodes.

In one accidental experiment of some hours in duration, Golden charged up an area so that, locally, all clocks became erratic. This included electrical clocks, battery-driven watches, wind-up kitchen clocks, and a pendulum-operated grandfather's clock. The rate of flow of time itself was apparently altered in the local area by the accumulated charge, which took some four days to drain off and discharge. At the end of the four days of discharge, all clocks and watches had returned to normal.

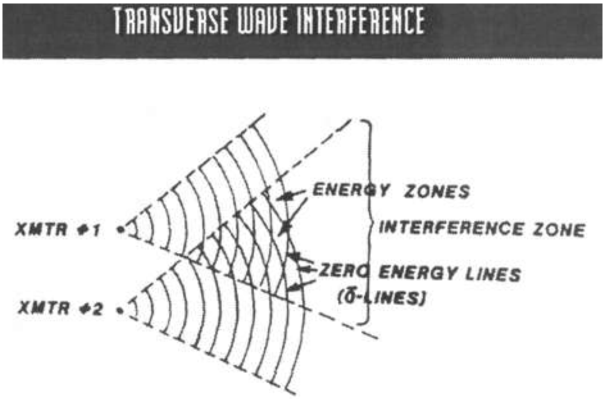

Slide 037: Transverse Wave Interference

On this slide we show the orthodox representation of EM wave interference.

The chart is stylized since we assume two single-frequency EM waves that intersect and interfere in the indicated region.

In the interference zone, in-phase wave amplitudes add (constructive interference) while out-of-phase wave amplitudes subtract (destructive interference). In our stylized mode, we show the zero-summed locations as lines.

Conventionally, these interfering waves are considered to be transverse waves in vacuum.

The point is, in this scheme we think of "putting in the waves of finite amplitude, and creating the zero-amplitude regions by destructive interference." We create additional waves of increased-amplitude regions by constructive iterference.

Note that this view is only half correct. It addresses the electromagnetics, but not the gravitational aspects. Actually the regions of destructive EM interference are regions of constructive G-interference, and the regions of constructive EM interference are regions of destructive G-interference.

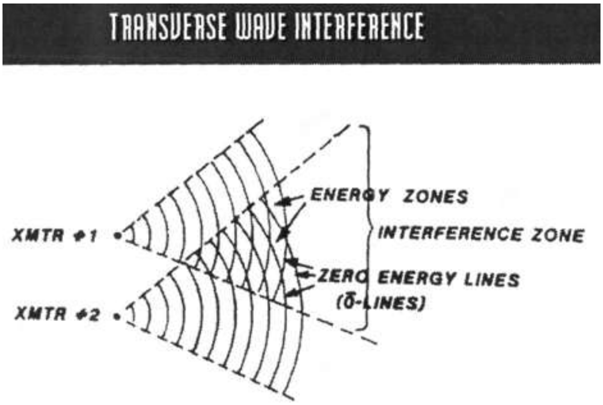
TRANSVERSE WAVE INTERFERENCE

Slide 038: Transverse Wave Interference

Here we show the same interference situation, only now we think of transmitting scalar EM waves, which we model as longitudinal waves.

Again we assume two single-frequency scalar EM waves that intersect and interfere in the indicated region.

However, to an external observer, each of these waves individually appears to have zero E and H fields, hence to contain zero EM energy. That is, conventionally we seem to be transmitting waves of "pure potential," without any "force field" amplitudes whatsoever. We say we are transmitting waves of "artificial potential," since we deliberately constructed the zero-vector-summation scalar waves in the first place.

In the interference zone, a strange thing now happens. The relative magnitudes of the artificial potential of one wave with respect to the artificial potential of the other depend upon the phase relationship between the two. Further, whether the "gradient" so established appears negative or positive also depends upon the phase relationship between the two.

What results is most interesting: the out-of-phase condition now produces an EM gradient whose sign and magnitude are functions of the location within the interference zone.

Thus one now has real, non-zero EM potential gradients, both electrical and magnetic. And so real EM energy has emerged in an interference pattern within the area.

Ironically, we get essentially what we had before! Only now we have "put in the zero-reference lines, and produced the non-zero gradient (energy) zones."

In other words, we have recreated ordinary EM energy at a distance, directly from the interference of gravitational potential energy (anenergy).

And, to the laboratory observer, there has been no "EM wave" energy flow through the intervening space.

This was Tesla's secret of "wireless transmission of energy to a distance without losses."

Further, in a hypothetically perfect case, the interference represents a sort of "energy bottle."

But now recall our need to take the sign of the scalar current into account. In this energy bottle, then, EM energy emerges and stabilizes - if and only if our transmitters are at higher reference potential than the ambient interference zone. If our transmitters are biased at lower potential, then energy is extracted from the energy bottle and re-emerges back at the transmitters, where it must be extracted and disposed of, if the transmitters are not to be burned out.

BACKGROUND FOR THE BRIEFER:

If we interfere two transmitters in a region of great energy -such as the molten core of the earth itself- and bias the transmitters negatively, we can extract enormous energy directly from that distant source!

Indeed, a slight variation of this effect is precisely how the Soviet Union obtains the enormous energy with which to power its incredibly powerful strategic "continent-buster" scalar EM weapons. Scalar frequency pairs - 12 kiloHertz **apart** - are transmitted into the earth itself on some portion of the earth's "giant scalar resonance" frequency curve. Biasing the two transmitters differently with respect to each other produces a high potential difference (voltage) between them. At the same time, biasing them both negative with respect to the earth's molten core produces a scalar resonance current from the core to the two transmitters. In the "outer loop" between the two scalar transmitters, there appears a gigantic voltage and gigantic electrical current that can be captured by appropriate means. This electrical power is tapped off and transmitted to incredibly powerful scalar EM weapon systems. There special accumulators and switchers are are utilized to provide the power of many Niagaras to each weapon.

Just before May 1, 1985 the Soviet Union performed a "full-up" operational test of their entire strategic scalar EM weapons complex. This test was detected and monitored periodically by Frank Golden. Some 27 of these giant Soviet "power taps" were locked into the molten core of the earth, producing forced (entrained) scalar resonance of the entire planet on 54 controlled frequencies under our very feet. The remainder of the scalar frequency spectrum was ablaze with literally hundreds of Soviet scalar transmitters: probably the entire strategic scalar command and control system to underwater submarines, higher command centers, distant commands, etc. was activated in the giant exercise.

For several days, the system was exercised on a mind-boggling scale, apparently as part of the Soviet Union's highly accentuated 40th anniversary celebration of the end of World War II.

Ironically, not a single U.S. intelligence agency, laboratory, or scientist detected this monstrous exercise for the new Soviet leader, Gorbachev. Not a one of them had a detector for scalar EM radiation, and not a one of them officially believes the exercise ever happened.

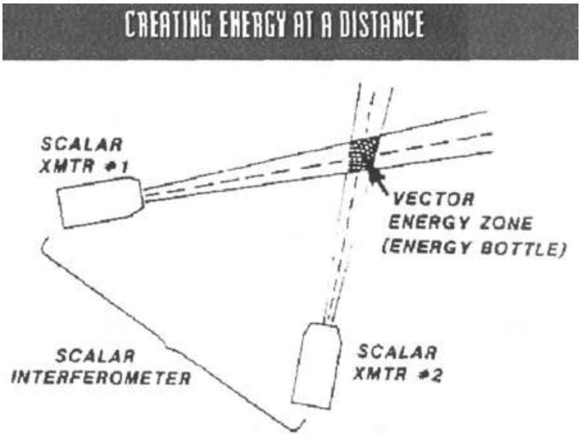
CREATING ENERGY AT A DISTANCE

Slide 039: Creating Energy at a Distance

A most useful device is obtained if one uses a scalar interferometer where the two transmitters transmit beams that intersect at a distance.

In the interference zone, an energy bottle is created.

By biasing the transmitter reference potentials well above that of the distant energy bottle, EM energy emerges in that zone. In that case the interferometer is operating in the exothermic mode.

By biasing the transmitter reference potentials well below that of the distant energy bottle, EM energy is extracted from the distant zone and emerges from the transmitter. In that case the interferometer is operating in the endothermic mode.

If **the** transmitters transmit continuously, the effect in the distant zone is continuous.

If each transmitter transmits a pulse, and the two pulses meet in the distant intersection zone, then an explosive emergence or extraction of energy occurs at the distant interference zone, depending upon whether the interferometer is operating in the exothermic or the endothermic mode.

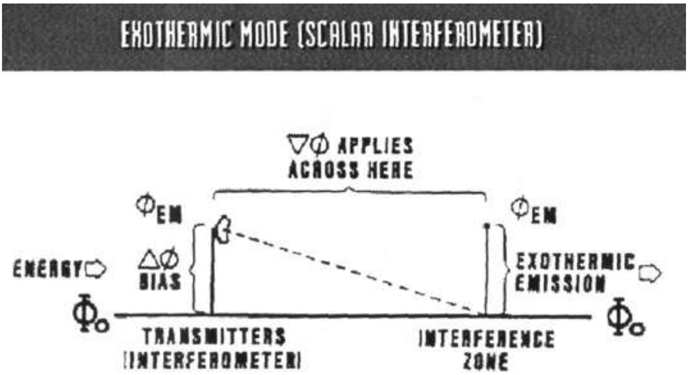

EXOTHERMIC MODE (SCALAR INTERFEROMETER)

Slide 040: Exothermic Mode (Scalar Interferometer)

For a scalar interferometer, "del phi" (the gradient of the potential) applies between the transmitter site and the distant intersection site, not to the intervening space in between. Thus one speaks of "energy flow" as being between transmitter and intersection, without any "in between." In the intervening space, all the energy exists as locked-in artificial potential, not EM force field energy (gradient bleed-off of potential).

In the exothermic mode, the ground potential of the transmitter is biased well above the ground potential of ambient vacuum. Energy enters the transmitter and "disappears", to "reappear" in the distant interference zone.

If continuous wave transmission is used, the energy continuously appears in the distant zone.

If pulse transmission is used and timed so that the two scalar pulses meet in the distant zone, energy explosively appears there.

If multiple frequencies are transmitted in the manner of a Fourier expansion of a particular geometric form, then a 3-dimensional energy form appears at the distant intersection. In this manner, a spherical or hemispherical shell (a globe of "dome" of EM energy) can be created at a distance. If pulse transmission is used, this will be an impulsive or explosive emergence of that energy form. If continuous transmission is used, this will be a continuous glowing form.

By feeding incredibly powerful transmitters with large amounts of energy extracted from the earth's molten core by "energy taps," very large spherical globes and hemispherical domes of EM energy can be created. The energy in the interference "shell" of such a sphere or hemisphere is sufficiently dense to lift Dirac matter from the Dirac sea of the vacuum. Thus the shell contains a glowing plasma.

Many such scalar howitzer signatures have been seen over the oceans, particularly over the North Pacific by jet airliners flying into and out of Japan. Also, such globes and domes of light have been seen deep within the Soviet Union, by observers in Afghanistan and by pilots of aircraft landing in Iran.

In the impulsive exothermic mode, a large visible flash will occur if EM energy in the visible spectrum is produced. "Flashes" in the infrared may similarly occur, if the emerging energy is in that spectral band. In September 1979 our Vela satellites detected such a large "nuclear flash" over the South Atlantic, off the southern coast of Africa. A second "nuclear flash" detected by the Vela satellites in 1980 was in the infrared only. Some years ago, a series of Momalous "flashes" and "booms" occurred off the east coast of the United States. These were due to the orientation and alignment -and registration - of one or more Soviet "scalar howitzers" - scalar EM interferometers in the exothermic mode.

Large "booms" occurred over the shuttle launch site at Cape Canaveral, associated with three shuttle launches prior to the end of November 1985. These were tests of such exothermic scalar EM howitzers, being tested as part of the Soviet Launch Phase ABM system. In that mode, the scalar pulses from the two transmitters are injected into special zero-reference scalar EM "channels" established in the EM carriers of the Woodpecker over-the-horizon radars. The scalar pulses travel through these channels, traveling in the ordinary EM beam of the transmitter and following it in its curvature around the earth in the earth-ionospheric waveguide. The scalar pulses are timed, phased, and oriented so as to meet the rising shuttle. In these tests the transmissions were offset in time so the shuttle would not actually be destroyed, but the system could be tested against actual shuttle launches in the U.S.

The specific uses of these various exothermic modes will be covered in later parts of the briefing.

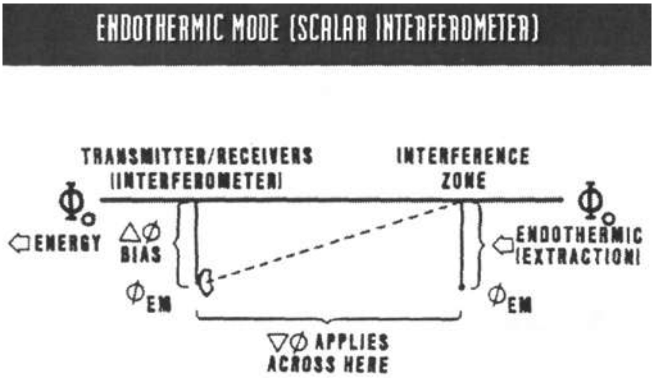

ENDOTHERMIC MODE (SCALAR INTERFEROMETER)

Slide 041: Endothermic Mode (Scalar Interferometer)

In the endothermic mode, the bias on the transmitters is such that energy is extracted from the distant intersection zone, to re-emerge at the transmitters. There the energy must be caught and disposed of if it is not to burn out the transmitters. When very large amounts of energy are extracted from distant points, the Soviets often use a "dump" method of temporarily capturing, and storing in an accumulator, the heat energy emerging at the transmitters. A "scalar energy tap" established into the accumulator then feeds a second howitzer in the exothermic mode. This howitzer is focused on a distant dumpsite - often Bennett Island, where U.S. weather satellites continually observe exhausts.

In the impulsive endothermic mode, energy is impulsively extracted from the distant intersection zone. In the atmosphere this results in a very sharp cooling, or a "cold explosion." It also results in the type of boom and rumble often associated with thunder after a lightning bolt.

On April 9, 1984 the Soviet Union tested such a "cold explosion" off the coast of Japan, near the Kuril Islands. In the suddenly induced low pressure "cold zone" above the ocean, ocean water was sharply sucked up from the ocean, forming a dense cloud. Then air rushing in to the low pressure zone forced the cloud upward, forming a mushroom much like that of an atomic explosion in visual appearance, except for the absence of a flash and the absence of a giant shock wave moving out away from the site of the explosion. (In this case the

Soviets controlled the rise time of the pulse so that no "giant atmospheric boom" occurred.).

As the cloud rises, it expands by mixing - very similar to a giant thunderhead buildup, except much faster. In this case the cloud rose to 60,000 feet in about 2 minutes, spreading out until it reached a diameter of about 200 miles. The pilots and crews of several Boeing 747 jet airliners in the general vicinity saw the incident. Sometime after that, a U.S. weather satellite photo of the area occurred. The cloud in that photo has an anomalous density distribution, differing from that of all other clouds in the area.

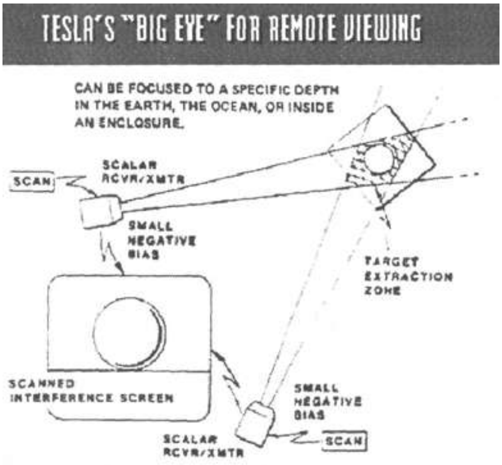
TESLA'S "BIG EYE" FOR REMOTE VIEWING
CAN BE FOCUSED TO A SPECIFIC DEPTH IN THE EARTH, THE OCEAN, OR INSIDE AN ENCLOSURE.
SCAN
SCALAR RCVR/XMTR
SMALL NEGATIVE BIAS
TARGET EXTRACTION ZONE
SCANNED INTERFERENCE SCREEN
SMALL NEGATIVE BIAS
SCALAR RCVR/XMTR
SCAN

Slide 042: Tesla's "Big Eye" for Remote Viewing

Nikola Tesla also planned a very special use for his endothermic scalar interferometer. He planned to produce what he called his "big eye to see at a distance."

This is accomplished as shown on the diagram. First, the system is only weakly endothermic, so that only a small amount of energy is extracted from the **distant** target. Also, the beams are "scanned" by an open receiver, timewise, from side to side and top to bottom. By scanning yet another single beam through the intersection zone and phasing its pulses, an even better representation can be obtained. Thus the receiver produces a representation of the energy extracted from various locations within the distant endothermic zone. By displaying the received signals on an appropriately scanned screen, a representation of the distant scene in the intersection zone can be created. Indeed this is a special kind of "microwave interferometry," and - with modern techniques - the imagery obtained might be surprisingly good. With development, it might even become as good as the image presently obtained by side-looking acquisition radars.

Interestingly enough, since scalar EM beams will easily penetrate the earth or the ocean, one can also look beneath the earth or beneath the ocean with this type of scanning scalar interferometer.

The importance of this capability to strategic and tactical reconnaissance is obvious. Camouflage, cover, and concealment have no effect on such a **system.** One can easily look inside buildings and into underground facilities. With a small system such as this, the U.S. Marines at Khe Sahn would have had **little** difficulty locating the tunnels continually dug under the perimeter by the Viet Cong. And targets under jungle canopies are directly visible.

BACKGROUND FOR THE BRIEFER:

It requires little imagination to see that this system is easily adapted to produce an "underwater radar." With such devices, the problem posed by the underwater nuclear submarine is solved. For example, an entire area can be continually searched, much like acquisition radar systems do now. A submarine can be detected and tracked, and none of its ordinary detectors will detect anything out of the ordinary. By using a separate pair of beams in the exothermic mode, powerful scalar pulses can be fired at the distant sub, intersecting at the submarine in a violent EMP throughout the sub and its armament. Thus the sub and all its missiles are destroyed instantly.

Or, continuous exothermic transmission can be used by the targeting weapon at lesser power, gradually interfering with the sub's electrical systems and causing it to lose control. The sub then sinks to crush depth and implodes.

Precisely that scenario seems to be what happened on April 10, 1963 to the U.S.S. Thresher nuclear submarine. It left a signature: the sub's surface companion, the nearby U.S.S. Skylark, was in the "splatter zone" of the underwater scalar interference. That is, spurious EM noise was being generated In all the Skylark's electrical systems, some of which were actually disabled. So intense was the "electronic jamming" that it required over an hour and a half for the Skylark to transmit an emergency message back to its headquarters that the Thresher was in serious trouble and contact with it had been lost. Some of the Skylark's communication systems actually failed, but later resumed operation inexplicably, once the jamming was gone. That type of "jamming" of multiple bands and multiple electronic equipments, of course, together with the anomalous failure of electronic equipment and its later mysterious recovery, were direct signatures of the use of the exothermic scalar interferometer against the undersea target area in the vicinity of the Skylark.

The very next day, Apr. 11, 1963, the same Soviet scalar EM howitzer system was tested in the "destroy submarine" pulse mode. A huge underwater EM blast occured off the coast of Puerto Rico, about 100 miles north of the island. The underwater explosion caused a huge boiling of the surface of the ocean, followed by the rising up of a giant mushroom of water about a third of a mile high. The mushroom of water then fell back into the ocean, completing the signuture.

Fortunately the startled crew of a passing U.S. jetliner that was just passing its checkpoint in that area saw the entire incident. (See Robert J. Durant, "An Underwater Explosion - or What?", Pursuit, 5(2), April 1972, p. 30-31.)

These two incidents were full-up operational tests of Khrushchev's newly deployed superweapons. He probably staged this dramatic one-two punch in a desperate effort to recover face with the Communist Party after his disastrous facedown by Kennedy in the Cuban Missile Crisis a few short months previously. Apparently the attempt was successful, since he remained in power another year before being deposed.

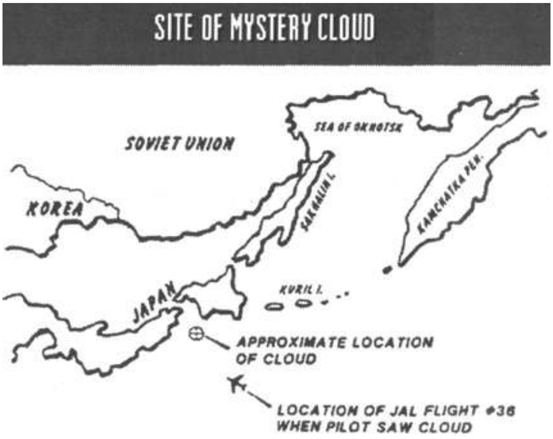
SITE OF MYSTERY CLOUD
SOVIET UNION
SEA OF OKHOTSK
KOREA
SAKHALIN I.
KAMCHATKA PEN.
JAPAN
KURIL I.
APPROXIMATE LOCATION OF CLOUD
LOCATION OF JAL FLIGHT #36 WHEN PILOT SAW CLOUD

Slide 043: Site of Mystery Cloud

On this slide we show the location of the "cold explosion" off the coast of Japan on April 9, 1984. The site of the explosion was only about 200 miles from downtown Tokyo. The crews of several jet airliners, including Japan Air Lines Flight 36, saw it.

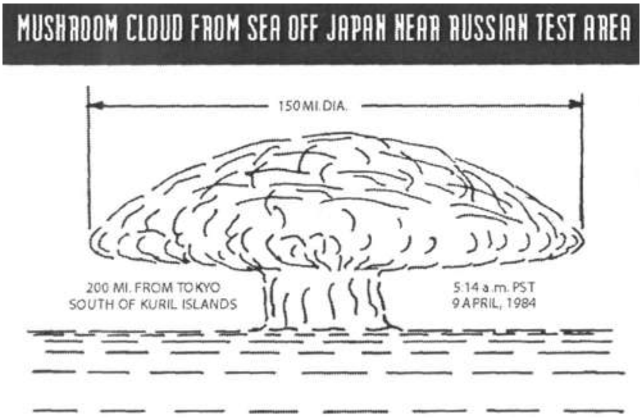

Slide 044: Mushroom Cloud from Sea off Japan near Russian Test Area

A short time before, the Soviets had hastily announced missile tests would be conducted into a zone some distance away from the cold explosion site. This may have been intended as deception to cover this incident, should any foreign nation indicate they understood what was going on.

It also was a direct "stimulus" to the Japanese and the rest of the world: That is, stimulate the system and see if the scientists recognize what happened. If they do, then they know about scalar EM weapons. If they don't, then their countries know nothing of scalar EM weapons, and those countries are defenseless against them.

At any rate, at about 5:14 a.m. Pacific Standard Time on April 9, 1984, a **giant** mushroom cloud erupted above the ocean south of the Kuril Islands. The mushroom rose rapidly, expanding as it rose. In two minutes it had reached **an** altitude of 60,000 feet and a diameter of about 200 miles.

One pilot of a Boeing 747 who observed the cloud was a former B-52 bomber pilot. He described the cloud as very similar to a mushroom cloud from a gigantic nuclear explosion, except there was no flash of visible light. Taking evasive action, he turned off course away from the blast, and braced for a blast shock wave, which never came.

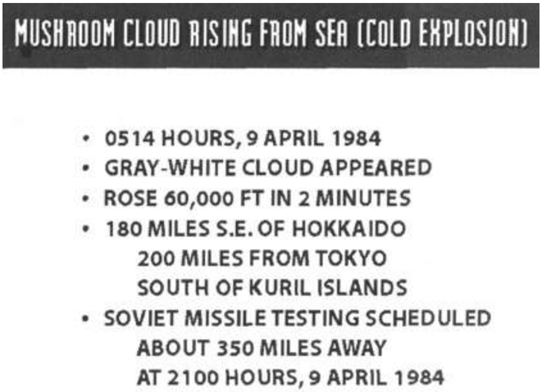

MUSHROOM CLOUD RISING FROM SEA (COLD EXPLOSION)

- 0514 HOURS, 9 APRIL 1984
- GRAY-WHITE CLOUD APPEARED
- ROSE 60,000 FT IN 2 MINUTES
- 180 MILES S.E. OF HOKKAIDO
  200 MILES FROM TOKYO
  SOUTH OF KURIL ISLANDS
- SOVIET MISSILE TESTING SCHEDULED
  ABOUT 350 MILES AWAY
  AT 2100 HOURS, 9 APRIL 1984

Slide 045: Mushroom Cloud Rising from Sea

Indeed, this highly anomalous cloud was due to a cold explosion - a test of a Soviet scalar EM howitzer in the pulsed endothermic mode.

In the weapon's targeted interference zone, EM energy (heat) was rapidly withdrawn when the biased endothermic extraction occurred. Above the ocean an intense low-pressure zone was created. Water from the ocean was sucked up, as if by a giant vacuum cleaner. The water was violently broken into droplets, forming a dense cloud, with a strong upward momentum.

The surrounding higher-pressure air rushed into the low-pressure zone, pushing the rising dense cloud of water vapor upward as its upward momentum continued. Mixing caused rapid expansion of the rising cloud, causing its density to thin measurably as it rose and expanded. An effect very similar to the formation of a strong thunderhead resulted, except much faster.

The mixing cloud also resembled the mushroom cloud from a giant nuclear explosion, just as an emerging thunderhead does.

But no outward blast wave resulted as does from a nuclear explosion. Also, no visible flash occurred, probably as a result of deliberate design by the Russians.

Thus a beautiful "stimulus" and test was conducted: quite a few airliners were "pinged," providing a high-level stimulus to competent observers. The news was sure to be reported by them to several nations. The press was sure to pick up the story. This was a high order stimulus to see whether the U.S., Japan, or other potential adversaries of the Soviet Union recognized the testing of a cold explosion weapon.

By our reaction, we assured them (with high confidence) that we still knew nothing of scalar EM interferometry or cold explosions.

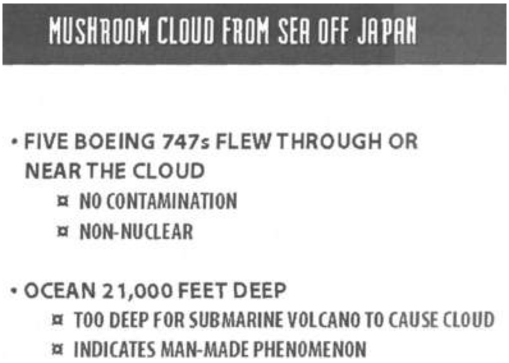

MUSHROOM CLOUD FROM SEA OFF JAPAN

- FIVE BOEING 747s FLEW THROUGH OR NEAR THE CLOUD
  - ¤ NO CONTAMINATION
  - ¤ NON-NUCLEAR
- OCEAN 21,000 FEET DEEP
  - ¤ TOO DEEP FOR SUBMARINE VOLCANO TO CAUSE CLOUD
  - ¤ INDICATES MAN-MADE PHENOMENON

Slide 046: Mushroom Cloud from Sea off Japan

Five Boeing 747's flew through or near the cloud on their way to Anchorage, Alaska. The aircraft were checked at Anchorage for radioactive contamination, but nothing was found. This showed that the huge "explosion" had been non-nuclear.

The ocean in the area is also about 21,000 feet deep. That is really too deep for a submarine volcano to have erupted and caused the cloud. (Besides, a volcano would have continued to snort at least a bit, and so that appears to be out of the question as a possible cause of the incident.)

In short, a manmade phenomenon is indicated.

Indeed, we know it was a cold explosion. As I've stated previously, a colleague has in fact produced the basic effect over a dish of water in the laboratory, using a small scalar interferometer in the endothermic mode.

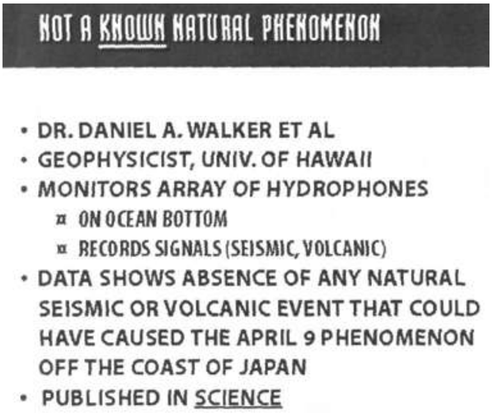

NOT A KNOWN NATURAL PHENOMENON

- DR. DANIEL A. WALKER ET AL
- GEOPHYSICIST, UNIV. OF HAWAII
- MONITORS ARRAY OF HYDROPHONES
  - ON OCEAN BOTTOM
  - RECORDS SIGNALS (SEISMIC, VOLCANIC)
- DATA SHOWS ABSENCE OF ANY NATURAL SEISMIC OR VOLCANIC EVENT THAT COULD HAVE CAUSED THE APRIL 9 PHENOMENON OFF THE COAST OF JAPAN
- PUBLISHED IN SCIENCE

Slide 047: Not a Known Natural Phenomenon

Further, the story does not end there.

A geophysicist at the University of Hawaii, Dr. Daniel A. Walker, together with his colleagues monitor an array of hydrophones on the ocean bottom, recording signals such as seismic incidents, volcano eruptions, etc.

Dr. Walker and his colleagues performed a comprehensive analysis of the data recorded by the hydrophones in and around the time of the incident in question. The data shows the absence of any natural seismic or volcanic event that could have caused the April 9th phenomenon off the coast of Japan.

He and his colleagues concluded that it was either an as yet unknown natural phenomenon, or a man-made phenomenon.

Their report is published in Science. (Walker et al, "Kaitoku Seamount and the Mystery Cloud of 9 April 1984," Science, 227(4687), Feb. 8, 1985, p. 607-611.)

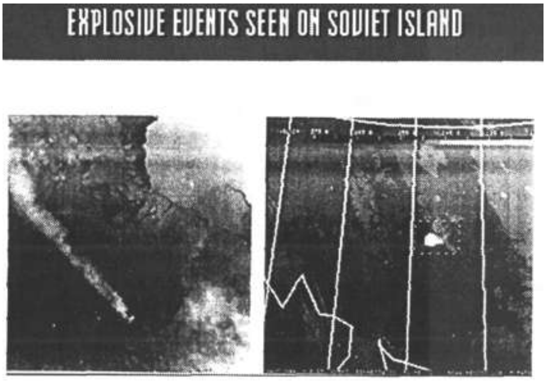

EXPLOSIVE EVENTS SEEN ON SOVIET ISLAND

Slide 048: Explosive Events seen on Soviet Island

Here are two U.S. weather satellite photos of the anomalous exhausts on Bennett Island we previously referred to.

The exhaust in the left picture is about 150 miles long, and is nearly horizontal (about 1.5 degrees above the horizontal). It is entirely consistent with the continuous exhaust from a "dumping transfer" scalar EM howitzer in the continuous exothermic mode. (The primary howitzer, of course, was activated in the endothermic mode.)

I think any open-minded investigator must admit that that is indeed a very powerful jet-like exhaust, and not like anything one ever sees from a natural volcano or geological gas vent. If anyone knows of such an anomalous volcanic or geological exhaust, I would be most delighted to see a photo or report. To my knowledge, no such natural photo exists, and no such natural exhaust has ever **been** reported.

The right picture shows the "puff of an explosive emergence of the exhaust. In other words, this one is the exhaust from a "dumping transfer" howitzer used in the pulsed exothermic mode. Since it did not have so much energy to dump, it could dump it in a pulse. Again, the primary howitzer, of course, was activated in the endothermic mode.

Over 100 weather satellite photos of these anomalous exhausts have been taken since 19**74.**

One or more U.S. aircraft have even been vectored through the exhausts, and samples of the exhaust taken and analyzed. The exhaust is made up of a little clay and mud and water, and is colder than the surrounding air.

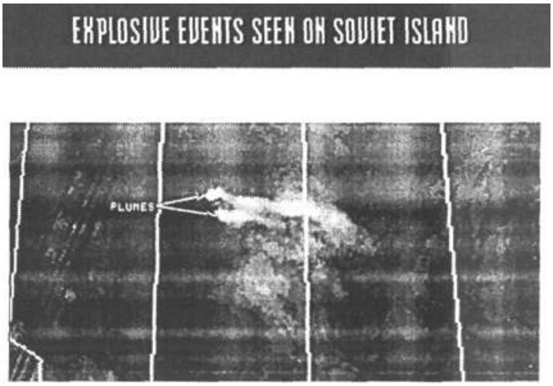

EXPLOSIVE EVENTS SEEN ON SOVIET ISLAND

Slide 049: Explosive Events seen on Soviet Island

U.S. weather analysts euphemistically call these exhausts from Bennett Island plumes.

Here is another U.S. weather satellite photo, showing the twin emergence of multiple puffs.

One might expect such exhausts if the secondary "dumping transfer" howitzer is operated in the continuous mode, with continuous extraction by "power tapping" the accumulator of the primary howitzer. In that case, the signature of the exhaust would expose the operation of the primary howitzer. For example, this exhaust would indicate multiple puffs, hence probably multiple frequencies and multiple primary transmitters, each phased in a controlled manner. That of course indicates the production of Fourier expansion forms - of 3-dimensional geometrical forms - by the primary howitzer(s). Since these are puffs, it may indicate the explosive formation of distant spherical balls of energy by the primary howitzer. This in turn indicates the use of a howitzer system in the "pulse firing" mode, which can be used against missiles, ships, aircraft, ground targets, armored vehicles, underwater targets, etc.

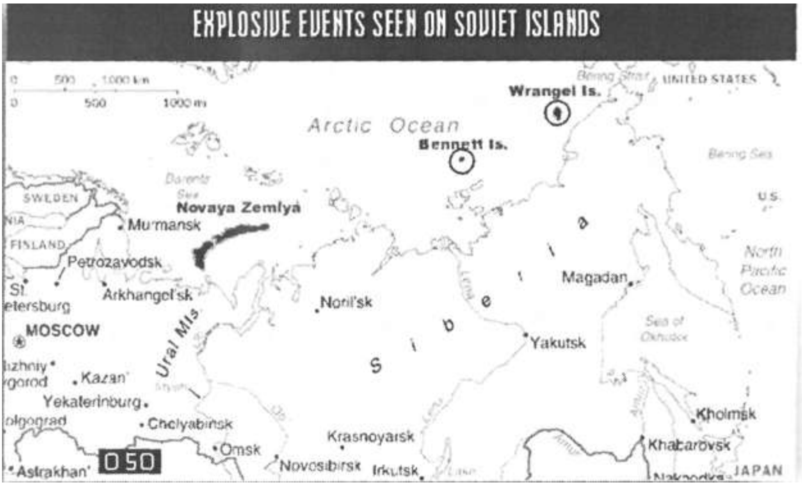
EXPLOSIVE EVENTS SEEN ON SOVIET ISLANDS
0 500 1000 km
0 500 1000 mi
Arctic Ocean
Wrangel Is.
Bennett Is.
Bering Strait
UNITED STATES
Bering Sea
U.S.
North Pacific Ocean
Sea of Okhotsk
Barents Sea
SWEDEN
FINLAND
Novaya Zemlya
Murmansk
Petrozavodsk
St. Petersburg
Arkhangel'sk
MOSCOW
Ural Mts
Noril'sk
Magadan
Yakutsk
S i b e r i a
Nizhniy Novgorod
Kazan'
Yekaterinburg
Volgograd
Chelyabinsk
Omsk
Astrakhan'
050
Novosibirsk
Krasnoyarsk
Irkutsk
Kholmsk
Khabarovsk
Nakhodka
JAPAN

Slide 050: Explosive Events seen on Soviet Islands

This slide shows the location of three areas of interest relative to Soviet scalar EM weapons testing.

Wrangel Island, north of the northeastern portion of the Soviet mainland, is associate with tests of these weapons in large circular breakups of the ocean's ice cover. This of course could aid submarines that wished to surface and fire cruise missiles, or submarines that wished to fire from underwater and needed the **ice** broken up.

Strangely enough, Wrangel Island has a most peculiar history: legally it still belongs to the U.S., but has been "given" to the Soviet Union without permission of the U.S. Senate, which is illegal.

Bennett Island, a little farther to the west and still north of the Soviet **mainland,** is noted for its phenomenal number of anomalous exhausts as previously discussed.

North of the Soviet mainland and near its middle lies the large island of Novaya Zemlya. Weather satellites have taken pictures of anomalous circulations over this island, which are further candidates for exhausts from scalar EM howitzer operations.

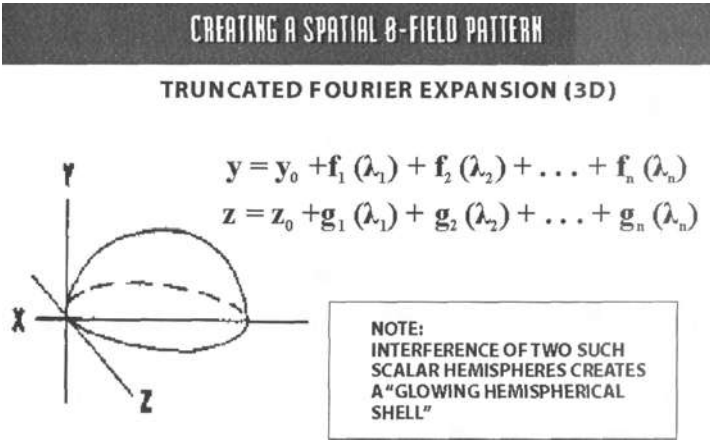

CREATING A SPATIAL 0-FIELD PATTERN

TRUNCATED FOURIER EXPANSION (3D)

$$y = y_0 + f_1(\lambda_1) + f_2(\lambda_2) + \ldots + f_n(\lambda_n)$$

$$z = z_0 + g_1(\lambda_1) + g_2(\lambda_2) + \ldots + g_n(\lambda_n)$$

NOTE:
INTERFERENCE OF TWO SUCH SCALAR HEMISPHERES CREATES A "GLOWING HEMISPHERICAL SHELL"

Slide 051: Creating a Spatial 0-Field Pattern

On this slide we show that a hemispherical dome (or other regular shape) can be composed from scalar waves by 3-dimensional Fourier expansion. By truncating the expansion - say after 10 or 12 frequency terms - a good representation of **the** desired form can be obtained with a practical number of frequencies, and in turn a practical number of transmitters. A computer is used to compute the multiplying functions for each frequency, and the entire complex of frequencies is **then** transmitted to produce the scalar form. By interfering two such scalar forms, a 3-dimensional shell of locked-in, ordinary EM energy is formed in **space.**

Interference of two such scalar hemispherical surfaces creates a "glowing hemispherical shell or dome of energy."

Interference of two such scalar spherical surfaces creates a "glowing spherical shell or globe of energy."

Both such geometric forms have been repeatedly seen.

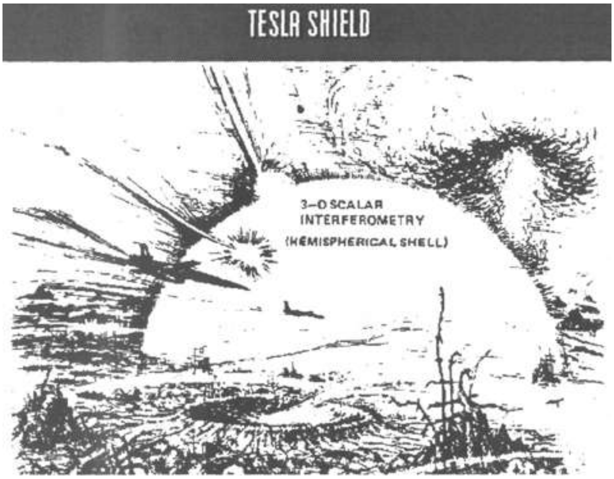

TESLA SHIELD

Slide 052: Tesla Shield

Of particular interest is the hemispherical shell of energy, which years ago I dubbed the "Tesla shield."

As shown on this slide, two scalar hemispherical surfaces are created, using multiple frequency transmitters and truncated Fourier series expansions. Interference of the two scalar hemispheres creates a great, glowing hemispherical shell of ordinary electro-magnetic energy. In the shell, the energy density is sufficient to lift Dirac matter from the Dirac Sea of vacuum. The shell is thus filled with a glowing plasma.

Such a shell may be several hundred miles in diameter at the base. The enormous energy required to form such a defense shell is obtained by a "scalar power tap" into the molten core of the earth itself, as previously explained. In late April/early May of 1985, 27 such "power taps" were placed in the earth by the Soviets. If each tap is capable of powering four to six large scalar EM weapons, then the Soviet strategic scalar EM arsenal contains over 100 monstrous superweapons capable of generating exothermic explosions, endothermic explosions, engineering the weather, locating and destroying underwater submarines, detecting and destroying ballistic missiles shortly **after** launch, detecting and destroying long range strategic bombers as soon as they are airborne, etc.

At any rate, the giant Tesla shield is useful against any penetrating vehicle. If the shell is not so large, its energy density may be very high. In **that** case, **the intense** plasma heating will fuse and even vaporize metallic bodies. In addition, any vehicle encountering the shell is subjected to an extremely intense EMP arising everywhere inside its circuitry. EM shielding is of no use against the creation of energy throughout the spacetime of the circuits; such shielding only helps against energy flow through space in the conventional sense.

Thus the electronics of any vehicle encountering the shield are instantly dudded, whether or not they are shielded against ordinary electromagnetic interference (EMI). This includes the electronics operating a nuclear warhead, carried by a re-entry vehicle. Electrically everything penetrating the shell is totally dudded. Further, explosive materials are exploded when such an EMP is encountered, and combustible materials are fiercely consumed or set afire. Ablative shielding suffers an interesting catastrophe: since energy does not try to "flow into" the shielding but "arises everywhere in it simultaneously", "ablation" occurs everywhere throughout the ablative material, simply exploding it instantly. In addition, for smaller Tesla shells (say of 50 miles diameter) the energy density is sufficient to melt or vaporize metals such as in missile structures.

With such a Tesla shield, there is no need to discriminate true warhead-bearing re-entry vehicles from decoys, chaff, etc. The entire "mess" entering the shield is simply "cleaned up" and "sterilized" or destroyed. The shield can take care of ICBMs/IRBMs and their nuclear warheads, strategic bombers and their nuclear **bombs,** cruise missiles and their nuclear warheads, re-entry vehicles and their warheads, decoys, chaff, etc.

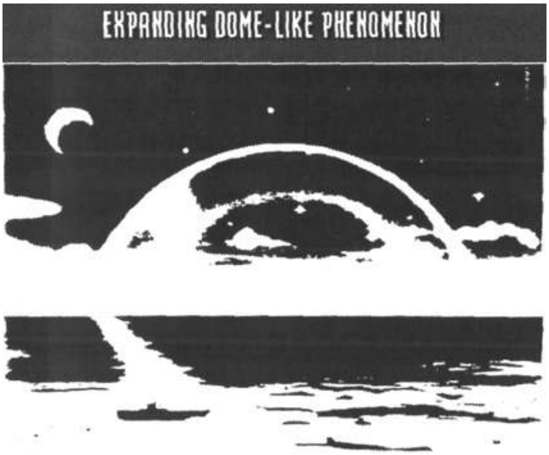

Slide 053: Expanding Dome-like Phenomenon

Here we show one instance of apparent Soviet testing of the Tesla shield (hemispherical shell of energy) over the ocean. This incident occurred in the Caribbean and Western North Atlantic Ocean on Mar. 20, 1969 and is reported in the Marine Observer, Vol. 40, 1970, p. 17. A semicircle of milky white light, quite intense at first, was sighted on the horizon. The semicircle enlarged, dimming as it expanded. The incident lasted for about 10 minutes. Seen from a passing ship, the top of the hemisphere reached a sighting elevation angle of about 50 degrees, while the base of the hemisphere of light subtended an arc of about 70 degrees. This is directly indicative of a test of the Tesla shield, first as a small, more intense shield and then expanding to greater size for a less intense shield whose primary destructive mechanism is EMP. Note that the EMP in a Tesla shield is locally contained in the energy-bottle shell. It is not an ordinary EMP of energy radiated out of the shell.

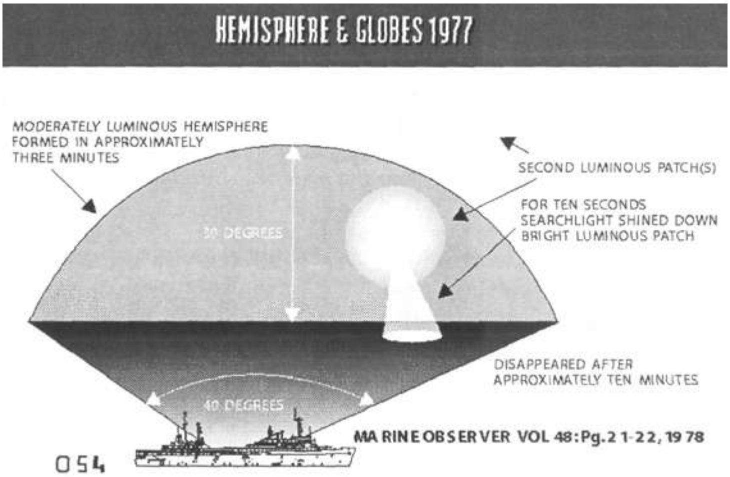
HEMISPHERE & GLOBES 1977
MODERATELY LUMINOUS HEMISPHERE FORMED IN APPROXIMATELY THREE MINUTES
SECOND LUMINOUS PATCH(S)
FOR TEN SECONDS SEARCHLIGHT SHINED DOWN BRIGHT LUMINOUS PATCH
30 DEGREES
DISAPPEARED AFTER APPROXIMATELY TEN MINUTES
40 DEGREES
MARINE OBSERVER VOL 48:Pg.21-22, 1978
054

Slide 054: Hemisphere & Globes 1977

Here is a second instance of apparent Soviet testing of the hemispherical Tesla shield over the ocean. This time it is combined with simultaneous testing of two smaller scalar interferometers creating Tesla globes of EM energy.
This incident occurred at 23 degrees 5 minutes north, 17 degrees 25 minutes west. It was sighted and reported by a British ship, H.M.V. Kinpurnie Castle, on Mar. 24, 1977. The report is in the Marine Observer. Vol. 48, 1978, p. 21-22.

First a large, moderately luminous hemisphere of light formed, requiring about 3 minutes to do so. Seen from the ship, the top of the hemisphere was at a sighting elevation angle of about 30 degrees, while the base subtended an arc of about 40 degrees. This was the basic Tesla shield.

Then two other luminous globes or patches of light, each of about 1 degree diameter as seen from the ship, formed in the same area; one inside the shield and one outside it. From the glowing globe inside the shield, a light beam similar to a giant searchlight shone down to the ocean's surface for about 10 seconds.

Again the phenomena lasted about 10 minutes and then disappeared.

The luminous globe or patch inside the hemisphere might be the test of a second weapon deployed from a Soviet aircraft in the area. If so, the aircraft has placed a **strong** spherical shield around itself. In addition, the aircraft is utilizing some sort of scalar interferometer beam weapon in an underwater mode. Probably this would imply a "shot" at a hostile nuclear submarine inside the shielded zone. Thus this aircraft would have the scalar EM equipment to search for the underwater sub and detect it, and also to concentrate interference energy upon it underwater, destroying it quickly.

Suffice it to say that it is possible to perform scalar interferometry by means of a single beam, in a proprietary process discovered by a close colleague. Thus the "glowing beam to the ocean" resembling a searchlight makes very good sense in the anti-submarine role. The same weapon, of course, could also attack any hostile surface vessel that might have been inside the shielded area at its formation.

The second luminous globe or patch, outside the hemisphere, is probably from a second scalar interferometer from the naval task force, simulating the continuous attack of a certain target (missile or aircraft) outside the protective dome.

As can be seen, this phenomenon is directly indicative of the types of operations one would expect from the testing of Soviet scalar EM weapons at sea, possibly to protect a Soviet naval task force during simulated combat operations.

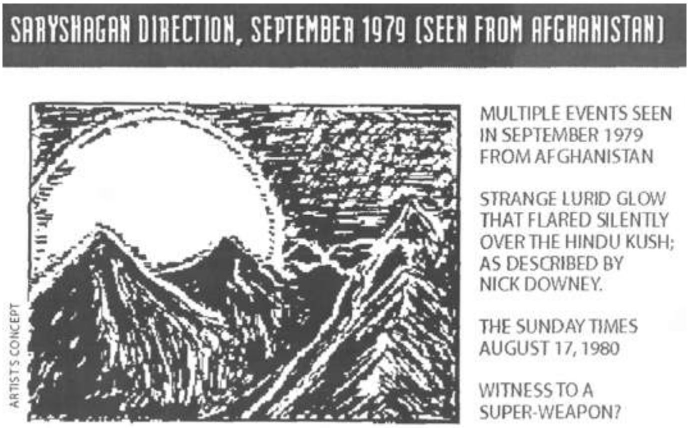

SARYSHAGAN DIRECTION, SEPTEMBER 1979 (SEEN FROM AFGHANISTAN)

MULTIPLE EVENTS SEEN
IN SEPTEMBER 1979
FROM AFGHANISTAN

STRANGE LURID GLOW
THAT FLARED SILENTLY
OVER THE HINDU KUSH;
AS DESCRIBED BY
NICK DOWNEY.

THE SUNDAY TIMES
AUGUST 17, 1980

WITNESS TO A
SUPER-WEAPON?

Slide 055: Saryshagan Direction, September 1979

The London Sunday Times of 17 August 1980 contained information and a photo-sketch of incidents of sighting of the testing of very large Tesla globes deep within the Soviet Union. British war cameraman Nick Downie made the sightings from Afghanistan. The phenomena seen were in the direction of the Saryshagan Missile Test Range, which - according to the U.S. Defense Department's Soviet Military Power, 1986 - contains one or more large directed energy weapons (DEWs).

Even though Downie was seeing the giant globe of light from a great distance, it flared silently over the Hindu Kush and expanded to subtend an arc of about 20 degrees, dimming as it expanded. (An arc of 20 degrees subtended by an object many hundreds of miles distant indicates an object of well over a hundred miles in diameter. This gives some idea of the enormous energy being controlled and manipulated by these Soviet weapons.).

Downie saw the sight on more than one occasion in September 1979.

BACKGROUND FOR THE BRIEFER:

In the same month - September 1979 - a stationary luminous globe containing a vertical stripe of black in the center was seen in the sky off Saint Petersburg, Florida. This particular type of sighting in that area has been previously correlated with times of known activity at Saryshagan.

Further, in the same month U.S. nuclear-warning Vela satellites detected a mysterious "nuclear flash" over the South Atlantic, off the coast of Africa. Controversy has raged in U.S. intelligence and scientific circles to this day as to whether a nuclear explosion or some other mechanism produced the flash.

Indeed, the flash may have been produced by a scalar EM howitzer from Saryshagan as one more "ping" of the U.S. intelligence system, to ascertain whether or not it knew anything about scalar EM howitzers. Again, the negative response told them with high confidence that (1) we still didn't know about scalar EM stuff, and (2) we were still absolutely defenseless against the Soviet scalar EM weaponry.

In 1980, a second "Vela flash" was reported. This time it occurred in the infrared region only, which positively rules out any sort of nuclear explosion. Even this increased stimulus still evoked a U.S. action that revealed total ignorance of scalar EM weaponry.

An alternate possibility for the September 1979 Vela flash also exists: countries not hostile to the U.S. may have tested a scalar EM weapon that produced the flash.

At any rate, from Downie's sightings, it is highly probably that the DEW weaponry at Saryshagan Missile Test Range was active in September 1979, and was producing huge Tesla globes. If the DEWs at Saryshagan can produce the giant luminous Tesla globe, they almost certainly are scalar EM interferometers and can produce the giant Tesla shields as well. Downie reported other earlier sightings of similar phenomena seen by Afghans deep within the Soviet Union in the same direction toward Saryshagan.

Briefly let us cover the uses of such a "giant globe" or spherical shell of glowing EM energy and plasma.

By placing such a "giant globe" thousands of kilometers out away from the defended heartland, an entire arc of the sky can be defended against long-range ballistic missile attack in midcourse. During their midcourse trajectory, the attacking missiles would have to penetrate the globular shell twice, exposing them to giant internal EMPs twice. A very high probability thus exists that all missiles entering the space occupied by the globe are dudded upon entry and/or exit. This includes the electronics inside the nuclear warheads themselves. Also, this is particularly effective against MIRV and MARV missile carriers, since the multiple re-entry vehicles are normally still on the main vehicle during most of midcourse. The use of this midcourse ABM globe defense greatly reduces the number of vehicles arriving at the latter part of midcourse and at the terminal phase of their trajectory.

By using a smaller, more intense globe and placing it on incoming clusters of objects or single objects, both EMP and intense local heating are used against the objects. This is suitable in the latter part of midcourse and in the **terminal** phase of ICBM's, IRBM's, SLBM's, and cruise missiles. It is also useful against incoming strategic bombers and their air-to-surface missiles, both ballistic and cruise.

Two modes of the globes - especially the small ones - can be used. First, the continuous mode can be used to "fry" or vaporize incoming objects in a relatively small volume (say two or three tens of kilometers in diameter). Second, the "pulse" mode can be used to "service" all incoming objects, whether or not they have passed through the "large globe" midcourse defense. This provides an additional guarantee of killing the objects; discrimination is not required, just service all of them. The exposure of all incoming objects to multiple attacks raises the probability of kill to essentially 100%, or as close to that as one wishes. Of course the incoming vehicle may still encounter a terminal defense consisting of the Tesla shield and associated roving "quickshot" small intense globes.

In short, with these systems an essentially 100% ABM and anti-bomber defense is possible. Further, the Soviets have possessed such an effective defense for two decades, just as they have openly stated.

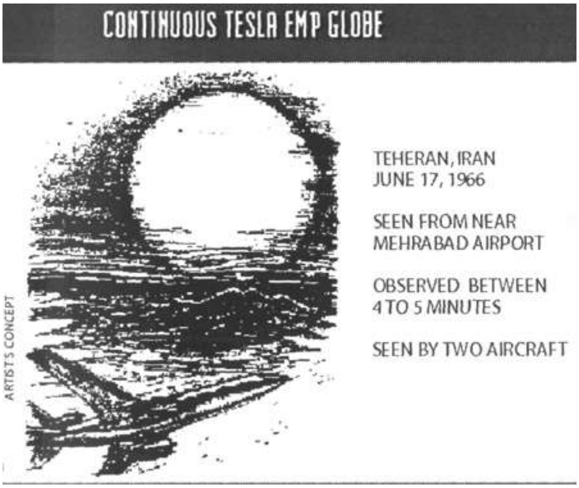

CONTINUOUS TESLA EMP GLOBE

TEHERAN, IRAN
JUNE 17, 1966

SEEN FROM NEAR
MEHRABAD AIRPORT

OBSERVED BETWEEN
4 TO 5 MINUTES

SEEN BY TWO AIRCRAFT

Slide 056: Continuous Tesla EMP Globe

Here is another verified incident of a gigantic test of a Soviet scalar EM howitzer deep within the Soviet Union.

This is a CIA report, released under the Freedom of Information Act. One can be quite sure that the incident occurred as stated.

The phenomenon was seen from two aircraft approaching Mehrabad Airport in Teheran, Iran on June 17, 1966 and reported by their pilots.

On the far horizon deep within the Soviet Union, an intense spherical ball of light appeared, "sitting on the horizon" so to speak. The globe of light increased to enormous size, dimming as it did so, literally filling an arc of the distant sky as it expanded. The sighting was shielded from most ground observers' view at the airport itself due to an intervening mountain range that masked most of tin-phenomenon from the ground.

The silent, expanding globe was observed for four or five minutes before it faded away.

BACKGROUND FOR THE BRIEFER:

Again, this is positive evidence of the testing of a giant scalar EM interferometer, in the "midcourse ABM globe" type of action.

However, note the date - mid 1966! The Soviets have therefore been testing such scalar weapons of enormous size and power for at least two decades. This implies that development must have started at least a decade earlier, or in the mid '50s.

Still earlier in January 1960 Nikita Khrushchev had announced the development of a "fantastic" Soviet weapon, one that could even destroy all life on earth if unrestrainedly used. In 1962 the ebullient Khrushchev was forced to back down and lose face before John Kennedy in the Cuban Missile Crisis. Khrushchev's missiles and bombers were in woeful shape, as Kennedy well knew (by courtesy of the Soviet spy, Colonel Oleg Penkovsky). Kennedy did not know that Khrushchev's new superweapons were nearing deployment, but were not yet quite ready.

To save face and prevent his immediate ouster, Khrushchev apparently conducted a startling two-strike demonstration of his new weapons as soon as they became operationally ready. On April 10, 1963 he detected and destroyed **the** U.S.S. Thresher nuclear submarine using a scalar EM howitzer in the underwater "continuous" mode.

The next day he demonstrated the "pulsed" underwater destruction mode for **nuclear** subs by producing a giant underwater EM explosion underneath the ocean 100 miles north of Puerto Rico. Ironically, just as the Atlantic was coming a**live** with U.S. naval vessels searching for the lost Thresher, a second test of the weapons that had finished off the sub occurred to the south of them, unnoticed and disregarded, even though seen and reported (to the FBI and the U.S. Coast Guard) by a passing U.S. jetliner's pilot and crew, who observed the underwater explosion.

We thus can peg the development of these weapons by the Soviets as starting **well** before 1960. The first operational deployment of the giant strategic weaponry occurred in early 1963.

Note that the 1966 testing observed here is completely consistent with this estimated development schedule.

Thus large Soviet strategic scalar EM weapons have been operational on site for 23 years! This implies that at least three additional generations of the scalar EM weapons have been developed and deployed by now.

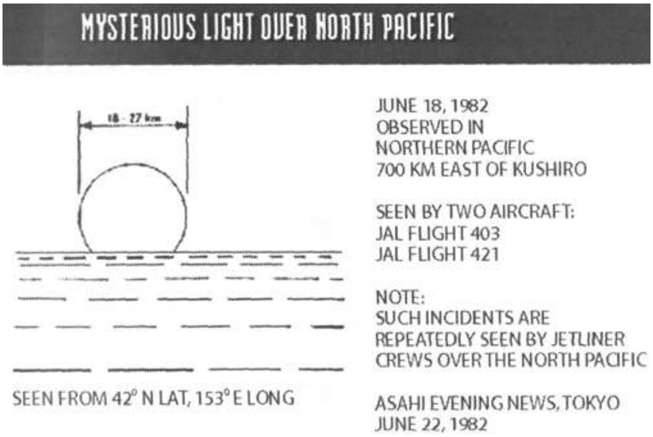

MYSTERIOUS LIGHT OVER NORTH PACIFIC

SEEN FROM 42° N LAT, 153° E LONG

JUNE 18, 1982
OBSERVED IN
NORTHERN PACIFIC
700 KM EAST OF KUSHIRO

SEEN BY TWO AIRCRAFT:
JAL FLIGHT 403
JAL FLIGHT 421

NOTE:
SUCH INCIDENTS ARE
REPEATEDLY SEEN BY JETLINER
CREWS OVER THE NORTH PACIFIC

ASAHI EVENING NEWS, TOKYO
JUNE 22, 1982

Slide 057: Mysterious Light over North Pacific

Incidents of Soviet testing of the Tesla shield and the Tesla globe weapons are routinely observed by airline pilots flying over the North Pacific into and out of Japan.

This is one typical example. Two jetliners, Japan Air Line flights 403 and 421, sighted and reported a large glowing globe of light sitting just beyond the horizon and extending well above it. The aircraft were in the vicinity of 42 degrees N latitude and 153 degrees longitude at the time. The sighting was thus about 700 miles east of Kushiro.

The pilots estimated the diameter of the ball as at least 18-27 kilometers. Depending upon the actual distance to the sphere, it may have been of much larger size.

This incident is reported in the Asahi Evening News, Tokyo, June 22, 1982.

The jet airliner pilots flying over these waters have reported many other **similar** sightings.

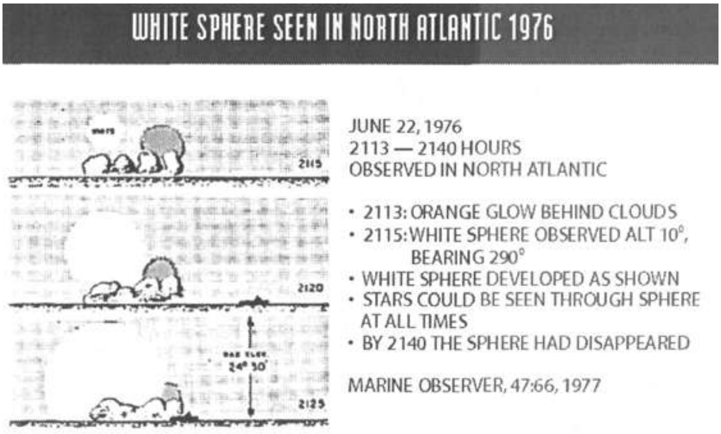
2115
2120
24° 30'
2125

WHITE SPHERE SEEN IN NORTH ATLANTIC 1976

JUNE 22, 1976
2113 — 2140 HOURS
OBSERVED IN NORTH ATLANTIC

- 2113: ORANGE GLOW BEHIND CLOUDS
- 2115: WHITE SPHERE OBSERVED ALT 10°, BEARING 290°
- WHITE SPHERE DEVELOPED AS SHOWN
- STARS COULD BE SEEN THROUGH SPHERE AT ALL TIMES
- BY 2140 THE SPHERE HAD DISAPPEARED

MARINE OBSERVER, 47:66, 1977

Slide 058: White Sphere seen in North Atlantic 1976

Here is another sighting closer to home.

This incident was observed from a passing ship on June 22, 1976 in the North Atlantic, at about 2113-2140 hours. It is reported in the Marine Observer, Vol. 47. 1977, p. 66.

First, an orange glow was sighted behind some distant clouds. A couple of minutes later, a glowing white sphere of light was observed to the left of the orange glow, just above the clouds. The white sphere then slowly expanded to a much larger sphere, dimming as it expanded. At its maximum size, the top of the white sphere reached about 24 degrees 30 minutes elevation angle to the observer. Development to maximum size required about 10 minutes.

By 2140 hrs the sphere had faded and disappeared. The sphere was sufficiently thin that the stars could be seen through it at all times.

Again this incident strongly fits the large Tesla globe mode of a Soviet scalar EM interferometer.

The significance and role of the orange glow are not known at this time.

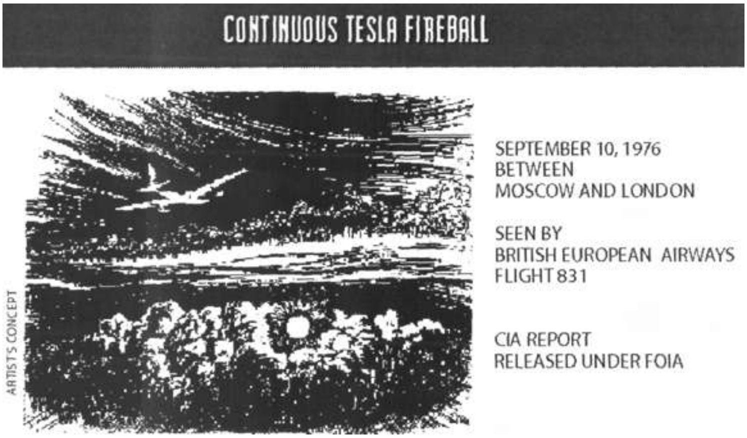

CONTINUOUS TESLA FIREBALL

SEPTEMBER 10, 1976
BETWEEN
MOSCOW AND LONDON

SEEN BY
BRITISH EUROPEAN AIRWAYS
FLIGHT 831

CIA REPORT
RELEASED UNDER FOIA

ARTIST'S CONCEPT

Slide 059: Continuous Tesla Fireball

Here is another incident that represents a stimulus to the British government, to see if the British are aware of scalar electromagnetics.

Again, this is a CIA report released under the Freedom of Information Act, so the details are reliable.

On Sep. 10, 1976 British European Airways flight 831, flying between Moscow and London and over Lithuania at the time, observed an intense ball of light above the clouds below the aircraft. The light was so intense that it lit up the sky in the entire vicinity.

The concerned pilot reported the glowing object to the Soviet ground authorities with whom he was in contact. He received the harsh instructions to ignore the light, and essentially to continue on his way out of there.

Here we see an incident involving a small, intense Tesla globe, of the kind with which the defense could "service" objects that had already penetrated a large midcourse globe, or aircraft approaching the defended area.

Obviously the Soviet authorities were tracking the aircraft, and knew it was in the vicinity. It seems logical, then, that they deliberately placed the brightly glowing ball beneath the aircraft so that the pilot and crew could not fail to observe it.

The strange message to the pilot was simply designed to increase the intensity of the stimulus. The stimulus was to be something like, "The Soviets are doing something in research and development that allows them to create intense balls of glowing light at a distance, and place these objects in the air in and around aircraft, possibly to intercept them." The purpose, of course, was to observe the British government's reaction after the incident was reported by the pilot upon his arrival at London.

Again the reaction of the British - and the U.S. as well - was as predicted. Again we showed that we knew nothing of scalar electromagnetic weapons, and did not recognize one when we encountered its effects.

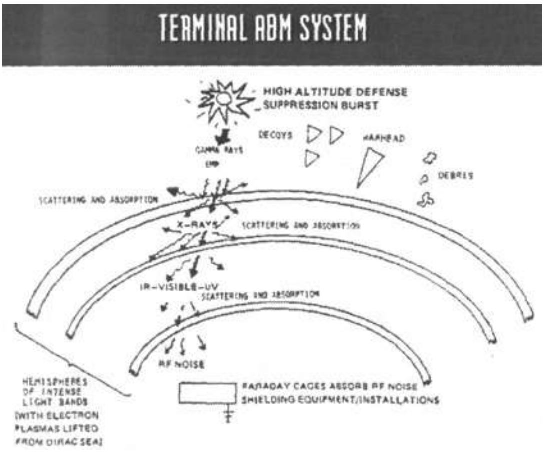
TERMINAL ABM SYSTEM
HIGH ALTITUDE DEFENSE SUPPRESSION BURST
DECOYS
WARHEAD
GAMMA RAYS
EMP
DEBRIS
SCATTERING AND ABSORPTION
X-RAYS
SCATTERING AND ABSORPTION
IR-VISIBLE-UV
SCATTERING AND ABSORPTION
RF NOISE
HEMISPHERES OF INTENSE LIGHT BANDS (WITH ELECTRON PLASMAS LIFTED FROM DIRAC SEA)
FARADAY CAGES ABSORB RF NOISE SHIELDING EQUIPMENT/INSTALLATIONS

Slide 060: Terminal ABM System

Very neat things can be done if one nests several Tesla shields - say three or four - concentrically, one inside the other. In that case even the nuclear radiation (such as gamma rays) from a defense-suppressive high altitude nuclear burst can be handled.

For example, suppose three such concentric shields are placed over a large vital area. Further, suppose a high altitude nuclear burst is placed up above the outer shield. Gamma radiation almost instantly strikes the plasma in the outer shell, where it is absorbed, scattered, and re-radiated at lower temperature. (That, after all, is what plasmas do.)

Inside the first shell, the scattered radiation is now in the x-ray and ultra-violet region. Let us track the most lethal component, the x-rays.

The scattered x-rays then strike the second plasma shell, and are absorbed, scattered, and re-radiated at lower temperature. Inside the second shell the scattered radiation is now in the visible and infrared region, with a little ultraviolet.

This optical radiation in turn strikes the third plasma shell, and is absorbed, scattered, and re-radiated at still lower temperature. Inside the third shell most of the energy is now in the form of radiofrequency (RF) energy, with a little IR and visible band spectral energy content.

At this point, ordinary electromagnetic interference (EMI) shielding of electronic equipments on the ground and inside the third shield can take care of any RF interference resulting from the emergent RF noise.

As can be seen, three shells are sufficient to convert the gamma and x-ray radiation (and ultraviolet and infrared) mostly to harmless RF energy before all three shells are penetrated. Thus the tactic of deliberate defense suppression by a preliminary high altitude nuclear burst can be countered by multiple Tesla shields.

In addition, of course, any ordinary vehicles penetrating all three shields are exposed to successive violent EMP's, and are almost certainly electrically dudded. The vehicles are also subjected to multiple periods of intense heating, so combustibles, fuels, explosives, and ablatives are destroyed. In addition, metal structures may be melted or vaporized.

Think of it this way: anything that hits one of these Tesla shields goes phht! Just **like** a bug hitting an electrified bugkiller screen.

For years passing ships have observed and reported such multiple-shield "light phenomena" over remote regions of the ocean. U.S. intelligence has routinely **not** paid any heed to reports of "lights at night" over remote ocean areas, and so Soviet tests in this manner have remained relatively unnoticed by officialdom.

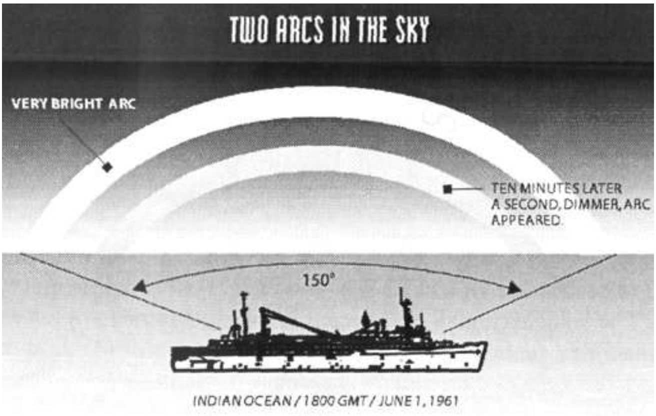
TWO ARCS IN THE SKY
VERY BRIGHT ARC
TEN MINUTES LATER A SECOND, DIMMER, ARC APPEARED
150°
INDIAN OCEAN / 1800 GMT / JUNE 1, 1961

Slide 061: Two Arcs in the Sky

Just to list one incident involving the testing of multiple concentric Tesla shields, here is an early test conducted in mid-1961 over the Indian Ocean.

The incident occurred at 1800 hrs. Greenwich mean Time (GMT) and was observed by a passing ship. The report is contained in Marine Observer, Vol. 32, 1962, p. 64.

First the ship's crew observed a gigantic, very bright arc of light on the horizon or beyond it, with both ends passing below the line of sight to the horizon. This was in fact probably a large Tesla shield, but of such low intensity that only **its** intersection arc could clearly be seen. The arc subtended an arc of about 150 degrees across its diameter at the bottom. After about 10 minutes, a second **and** dimmer arc appeared, concentrically inside the first one. The ship thus was observing what probably was the nesting of two giant hemispheres (of EM energy) to provide a multiple Tesla shield.

This was probably a test of an early research and development prototype, since Khrushchev had only announced his new "fantastic weaponry" in development in his January, 1960 speech to the Soviet Presidium. And apparently land-based weapons such as this prototype probably did not become deployed and operational until April 1963. However, even in 1961 the Soviets would have been developing and testing prototypes of the weapon adapted for naval task force protection and usage.

Several other similar phenomena involving nested hemispheres or giant concentric arcs of light have been seen by mariners at sea.

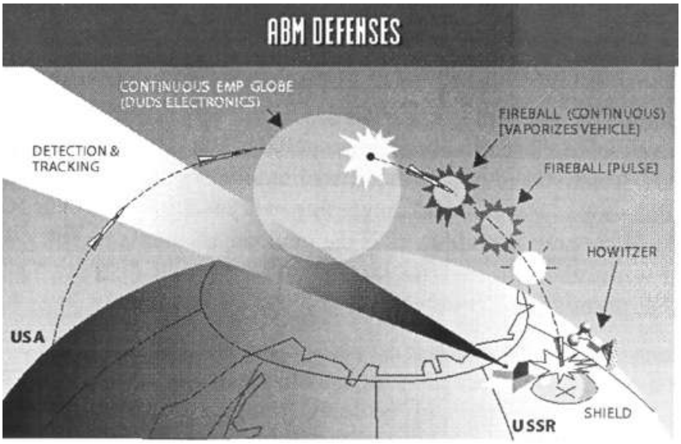

Slide 062: ABM Defenses

When the various capabilities of Tesla shields, globes, and scalar howitzers are integrated together into an ABM defense, a most formidable Soviet strategic defense capability is apparent.

On this slide we deal only with midcourse and terminal phase ABM defense. Later we shall cover the application of these weapons to launch phase ABM defense, primary phase bomber defense, underwater antisubmarine defense, etc.

Suppose the U.S. is provoked to desperately launch a missile attack toward Russia, as indicated on the slide.

First would come the launch phase destruction of the launched missiles, which we will cover shortly.

Any remaining missiles in midcourse would encounter extremely large, glowing spheres of light, where the EM energy density is contained and held in the spherical shell "energy bottle" of the glowing globes. Each globe might cover an are of up to about 20 degrees, as seen from the Soviet defended area. Missiles carrying MIRV and MARV payloads would not yet have deployed their re-entry vehicles.

As a missile and its payload encounter one of the globular shells, an intense EMP is suddenly experienced throughout the missile and all of its components. Again, EMI shielding is of no consequence since the EMP arises in each increment of spacetime throughout the vehicle volume. The result is instantaneous dudding of all electronic circuitry, including the electronics of the nuclear warheads themselves. The HE content of the nuclear warhead system is violently exploded, totally destroying the configuration of the warhead. A low-order nuclear detonation may occur, but more likely the warhead's nuclear material is just broken up, finely divided, and set afire.

In addition, all combustible materials are set fiercely afire or destroyed. Explosive materials such as booster propellants and HE explosives are violently exploded. Ablative and plastic materials literally explode into pieces, torn apart by sudden violent ablation and outgassing everywhere within. The remnants of the vehicle are now total duds, as is the payload, including re-entry vehicle, nuclear warhead electronics, nuclear materials from the disruption and explosion of the HE portion of the warhead system, etc.

Still, each piece of the debris must pass through the glowing shell once more before it exits. Thus a second "shot", precisely as effective as the first, still awaits each and every piece of the debris plunging along inside the shell as it exits.

As the debris continues on to the latter part of midcourse and early terminal phase, any portions large enough to track are subjected to small, intense Tesla globes where the energy density in the shell is many orders of magnitude greater than in the larger globe used in earlier midcourse. This part of the defense system simply services each incoming "blob" methodically, probably first in the pulse mode. Later, any blobs still remaining are given the maximum "continuous heating" mode with a small Tesla globe, to actually vaporize metals and materials.

In the unlikely event that anything active still remains, or that an object has managed to elude the large and small Tesla fireballs, a Tesla shield is still to bo encountered, towering over the protected area. For significant targets this shield will be concentric, and will contain three or more concentric Tesla shields. Again, the approaching vehicle will be subjected to multiple EMP's and severe heating throughout its volume. In addition, the concentric shields will "clean up and sterilize any gamma radiation and EMP resulting from any deliberate high altitude defense suppression nuclear burst that somehow (inconceivably!) gets through the launch phase ABM system, and the Tesla-globe midcourse and earl terminal phase defenses.

As can be seen, the system complex represents a most formidable defense **that** certainly approaches 100% effectiveness against the present triad, and even against the addition of cruise missiles.

In addition, since these weapons suffer from very strange constraints and limitations, the scalar transmitters are actually dual mode: they become conventional radars by simply switching into the ordinary EM mode. They arc part of the more conventional ABM defense system complexes we presently recognize in the Soviet Union. In this conventional role, backup ABM missile interceptors are available, to be employed in case the conventional role must be assumed.

Later we will address the reason for the strange limitations that must be placed upon unqualified use of the large scalar EM systems, particularly in the pulse or "burst" nodes.

NETWORK OF VIRTUAL TRANSMITTERS

Slide 063: Network of Virtual Transmitters

In July 1976 the U.S. received very special Bicentennial greetings from the Soviet Union.

At that time, communications systems of the world in the 3-30 megaHertz band suddenly met substantial interference from extremely powerful, chirped Soviet transmitters that were suddenly activated. These transmitters continue their transmissions to this day.

Estimates of the power of these enormous transmitters vary, but figures range as high as several hundred megawatts, with a nominal figure being 100 megawatts.

These powerful transmitters were promptly nicknamed "Woodpeckers" because of the characteristic sound of the chirped signal when received. That is, the received signal made a "pecking" sound much like a woodpecker's beak hitting a block of wood.

Several nations protested, but the powerful signals have continued, right down to this day. The only Soviet response was to add a "spread spectrum" capability, so that the transmitter did not dwell too long on one specific frequency, but shifted periodically to other frequencies.

These transmitters have apparently never been precisely located by U.S. intelligence, but their beams carry much of the characteristics of an **over-the** horizon (OTH) radar. They have been dubbed OTH-radars by U.S. intelligence, and can without question perform that mission, in addition to some very interesting missions which U.S. intelligence does not assess.

Soviet Military Power, Department of Defense, 1985, p. 45 shows the direct intersection over the United States of the Woodpecker radar beams used in an OTH-role. We show an illustration of that intersection on this slide.

In addition, we show an additional "scanner" beam that can be scanned across the intersection "grid" over the U.S., formed by waveform interference of two main Woodpecker beams.

These Woodpecker transmitters have a great many diverse functions.

First, they can be used in a conventional over-the-horizon radar node, since their beams follow the earth-ionosphere waveguide and curve around the earth. In this mode they can detect missiles at launch and thereafter, and strategic bombers when taking off and thereafter.

However, these scalar interference grid weapons have many other uses.

First, they can be used to biologically attack entire populations in a targeted area. This aspect is not covered in this briefing. Suffice it to say that phase-locked ELF modulation signals ot 10 Hz and less are often detected on multiple Woodpecker frequencies simultaneously. In a target area, this modulation - if sufficiently stronger than the Schumann resonance of the earth's magnetic field - will entrain a percentage of the brains into "forced entrainment." In that case, these human brains are "synchronized" to the Woodpecker signals so that multiple coherent frequencies are phase-locked into them. That is, multiple coherent EM channels directly into these entrained brains now exist. At that **point,** Fourier expansions may be used to attack specific portions of the brain geometrically.

In addition, scalar EM disease patterns can be modulated upon the carriers, again with Fourier expansions. Specific biological effects can be induced in the entrained populace at will, limited only by the state of the art of the Soviet technology used to attack them. Possible induced effects include instantaneous **death,** heart seizure, severe emotional disruption, loss of control of internal functions, diseases, disabling of the immune system, and even implantation of **thoughts,** emotions, and ideas which are interpreted by the targeted subjects as their own.

While further discussion of this area is beyond the scope of this briefing, the biological aspects of the Woodpecker transmitters are horrible. It suffices to say that, in thousands of experiments, Kaznacheyev demonstrated that almost any **kind** of cellular death and disease pattern could be electromagnetically transmitted. Kaznacheyev reported the effect in the near ultraviolet. Experimenters at the University of Marburg in West Germany duplicated the experiments in the infrared.

The bottom line is that photons themselves can carry death and disease patterns between cells. Scalar E*M* technology allows the synthesis of the actual potential pattern (which after all represents total control of charge and charge distribution, hence biochemistry in the cell) of a particular cellular disease or death mechanism. Symptoms (and cellular death from them!) of nuclear radiation, chemical poisoning, bacterial infection, and other mechanisms were successfully induced by the Kaznacheyev experiments.

Dr. Popp of West Germany has published an analysis of the virtual photon master control system of the cells. Since scalar EM represents the deliberate ordering of virtual particle flux into deterministic patterns, the master control system can readily be entered with scalar techniques to induce disease and disorder at will.

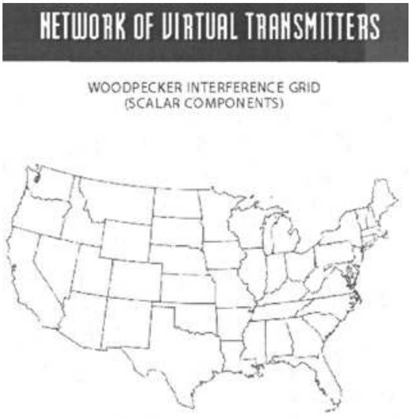

NETWORK OF VIRTUAL TRANSMITTERS

WOODPECKER INTERFERENCE GRID
(SCALAR COMPONENTS)

Slide 064: Network of Virtual Transmitters

On this slide we diagrammatically represent the Soviet woodpecker interference grid over the U.S. This grid can easily be used to influence and control the weather, and for over a decade the Soviets have manipulated the weather over the U.S. and other parts of the world.

By the peculiar title of this slide, we wish to call attention to the fact that, in the exothermic mode of scalar interferometry riding on the woodpecker radar signals, each of the woodpecker grid interference cells can act as a little "virtual transmitter." It's exactly as if the Soviets had been permitted to enter the U.S. and build thousands of EM transmitters at regular intervals all over the country!

And of course it's even worse. In the endothermic mode, it's as if the Soviets had been permitted to enter the U.S. and build thousands of "cold generators" all over the U.S. That is, in that mode, each of the "transmitters" can actually extract electromagnetic energy.

These two modes have enabled the Soviet Union to drastically manipulate the weather at will over the U.S. during the last decade (and similarly, for other places as well.)

(For a disclosure of the Soviet weather control operations over the U.S., see T.E. Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985, available from The Tom Bearden Website www.cheniere.org. We will only summarize here.).

Briefly, inside various reference level zeros on the ordinary EM woodpecker carriers, scalar "channels" are established. Scalar beams, signals, and interferometer pulses and transmissions can be transmitted through these reference channels as if directly through a wire. Phasing, pulse shaping, and Fourier expansion formation of geometric forms are then the key to be used in scalar interferometry through the internal scalar channels of the Soviet woodpecker signals.

By choosing one or more grid cells and using scalar interference there in the exothermic continuous mode, local heating and expansion of the air, and formation of a high-pressure area, occurs.

By choosing one or more grid cells and using scalar interference there in the endothermic continuous mode, local cooling and shrinking of the air, and formation of a low-pressure area, occurs.

By electromagnetic "rotation" of the interfering scalar beams, the hotspots and coldspots can be moved along a desired path and at a desired speed.

By these methods the entire jetstream across the U.S. can be deviated and controlled to a large degree. Moisture can be drawn from the Pacific beyond Southern California and Mexico, and collided over the southern U.S. with extremely cold air brought down from Canada, producing ice storms such as those in and around February 1, 1985.

Violent thunderstorm activity can be induced and directed. By adding "spin", these same storms can be induced to form extensive tornados.

By sustained weather engineering operations, drought or excessive rain and flooding can be induced in areas of the U.S. as desired. Crops can be destroyed or heavily damaged by severe and unseasonal weather.

If the grid is placed in the ground on scalar EM carriers instead of normal EM carriers (and the woodpecker "radars," can switch to, or add, scalar carrier mode whenever they wish), it can be used as an anti-submarine weapon, as we shall see later. It can also be used to cause substantial ocean effects that have drastic consequences on the world's weather - such as the anomalous El Nino of 1982-83.

By localizing one spot of heating at the top of a thunderstorm anvil, and another spot of cooling at its side, a localized downburst can be created. Our weather satellites have detected just such an anomalous combination heating and cooling of thunderstorm anvils over the U.S. While there may exist natural mechanisms to cause this anomalous dual pattern, it may also be caused by Soviet weather engineering, which can cause either broad-area or highly localized effects.

All these types of weather engineering incidents have been deliberately created over the U.S. by the Soviet Union.

Perhaps the greatest weather effects can be induced by altering some of the large-scale normal mechanisms that influence and control our weather. Thus the **artificially** induced El Nino of 1982-83 was a case in point, causing drastic weather upsets worldwide for a year or more.

Thelast decade has seen weather so anomalous that it could only occur once in 1200 years by chance. It has not been due to normal chance, but by Soviet weather engineering, particularly since the advent of the Woodpecker transmitter complexes.

ADDITIONAL BACKGROUND FOR THE BRIEFER:

The Woodpecker transmitters can carry scalar components, which, in addition to biological strikes, can be used for many other purposes such as to

1. Influence and control the weather,
2. Deviate the giant jetstreams of the upper atmosphere,
3. Detect, track and destroy missiles shortly after launch,
4. Detect, track and destroy strategic bombers and other aircraft,
5. Attack land and ocean surface targets by large aerial electromagnetic explosions: the EM explosion produces the fireball, the thermal effects, and the blast wave of a nuclear explosion without the nuclear radiation and radioactive fallout contamination;
6. Attack surface land and ocean targets by large "cold explosions", where living things freeze almost instantly, leaving the surface facilities and equipment intact,
7. Attack land and ocean surface targets by severe internal electromagnetic disruption, so that all electronics are dudded,
8. Attack personnel on land and sea by direct destruction of their nervous systems with an internally induced strong EMP,
9. Disrupt the operation of distant facilities, equipment, and personnel or destroy food crops in large areas by severe cold or severe heating, or even startling oscillation between the two.

   In addition, the scalar aspects of the Woodpecker transmitters can be beamed through the earth and ocean as surface waves or at a given depth to intersect at a distant region. This mode can be used to

10. Cause distant earthquakes in a distant fault zone by depositing energy in the rocks there, increasing the plate stress until slippage occurs at a fault zone,

11. Cause large, anomalous rocking of an entire plate or large area, by depositing and extracting energy from alternate areas, causing buckling of the large rock plate,
12. Set up an interference grid under the water, so that underwater submarines can be detected, tracked, and destroyed on station and with their missiles still unfired,
13. Attack naval vessels and task forces through the water and through their hulls, without warning,
14. Attack missile silos and nuclear storage sites "before launch", directly through the earth itself,
15. Initiate the full nuclear explosion of implosion-type nuclear warheads on-board missiles on station,
16. Initiate full nuclear explosion of implosion-type nuclear warheads in storage sites and ready reserve,
17. Destroy seaborne and ocean-bottom-mounted mines and underwater sensors, even including entire "fields" of them,
18. Detect and destroy hostile torpedoes, ship-to-underwater anti-submarine missiles, and dumped anti-submarine munitions and charges,
19. "Pump" the electrical grounding of distant electrical power distribution and generation systems so that the system catastrophically fails,
20. "Pump" the electrical grounding of distant electrical power systems so that radio and television transmitters, power grids, and wiring complexes in buildings and facilities radiate a desired scalar wave component or signal complex at equipment and personnel.

All in all, the woodpecker transmitter complexes and other associated scalar EM weaponry are most useful in the major missions required in war.

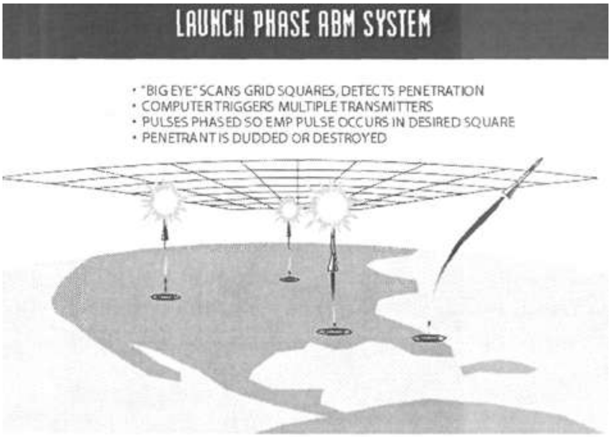

LAUNCH PHASE ABM SYSTEM

- "BIG EYE" SCANS GRID SQUARES, DETECTS PENETRATION
- COMPUTER TRIGGERS MULTIPLE TRANSMITTERS
- PULSES PHASED SO EMP PULSE OCCURS IN DESIRED SQUARE
- PENETRANT IS DUDDED OR DESTROYED

Slide 065: Launch Phase ABM System

On this slide we show the use of the Woodpecker grid in the Launch Phase **ABM** system role.

The scalar components establish an interference grid over the U.S., at some altitude above the earth. One or more separate scalar interferometers transmit through scalar channels established in the Woodpecker's normal EM carriers, scanning through the grid square cells one by one in the "just barely extract energy" endothermic mode.

Back in the Soviet Union, the energy extracted passes through a scanned receiver biased just so that the normal energy extracted from the distant atmosphere does not show on a scanning screen. Any substantial amount of additional energy appearing in the cell - such as from the jet engines of a large aircraft or the rocket exhaust of a large missile - will be detected.

The **operation** of this weapon is four-fold: (1) the scanning interferometer or "big eye" scans the grid squares, detects penetration, and tracks the rising missile. (2) A separate scalar EM interferometer (a howitzer) is controlled by a computer and fired through a separate scalar EM channel opened in the ordinary EM woodpecker carriers. (3) Two timed scalar pulses are fired so that they converge and meet in the grid cell containing the rising missile. This results in the **sudden** emergence of a violent EMP pulse throughout the missile and the surrounding vicinity. (4) All missile and warhead electrical and electronic systems are instantly dudded and destroyed. In addition, fuels, propellants, and explosives of the missile are immediately ignited, explosively destroying the penetrating missile.

That takes care of the missile target. (The same weapon, of course, can be used against airborne strategic bombers - including Stealth-type bombers. In fact, Gary Powers' high flying aircraft was probably shot down in 1960 by a "jury-rigged" scalar EM howitzer using modified radars and timed scalar pulses to provide an aerial explosion and EMP.)

The Woodpecker grid/howitzer weapon system can be placed over the ocean and used against cruise missiles, naval surface-to-air and surface-to-surface missiles, submarine-launched ballistic and cruise missiles, etc. Placed over a carrier task force, it can also take care of the aircraft launched by the carrier as fast as they are launched. It can also handily take care of the missiles launched by guided-missile cruisers of the accompanying task force.

Now let's take a look at the command and fire control end of this weapon system, way back in the Soviet Union.

SOVIET COMMAND & CONTROL CONCEPT

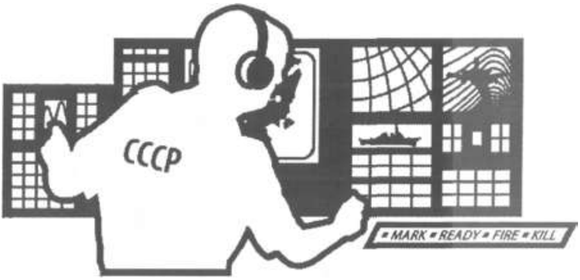

Slide 066: Soviet Command & Control Concept

On this slide we show a postulated Soviet fire control system for the grid weapon.

Two or more operators sit at a control console, facing a giant display (screen). The display contains a grid, which represents the distant interference grid of the Woodpecker carriers (or scalar EM carriers, if in the underground or underwater mode). A distant missile launched into the real interference grid overhead will cause a spot of light to appear in the appropriate grid square (cell) on the operators' screen. The operators will track the displayed target, select firing mode and type of action, engage the target, and assess whether or not the distant target is killed or missed by the firing.

When a missile is launched in the distant area over which the Woodpecker grid has been placed, it quickly rises through one of the scanned grid cells on its upward journey. As it penetrates the cell, additional energy is extracted (since part of the missile's exhaust energy is now extracted also). At the screen in the Soviet Union, the additional energy received rises above bias, causing a "bloom" or spot of light to appear on the screen at the appropriate grid cell representation.

The distant operator has now detected the missile launch and its location. He places a marker over the target track and activates the computer, also inputting the type of firing or action to be performed. In this case, let us say he wishes a burst of **energy** to emerge inside the distant launched missile, and has activated that mode of firing action.

The computer computes the necessary parameters for another scalar interferometer channel in the exothermic mode, and for firing at the location of the rising missile so that two scalar pulses will meet in the cell penetrated by the rising missile on the other side of the earth. As the channel is opened, the firing solution settled, and the interferometer howitzer readied, the computer activates indicators on the console notifying the operator. When all is ready, a "ready-to-fire" light is lit on the console.

Upon receiving the command to fire and destroy the missile, the operator presses the fire button. The computer fires the activated scalar EM howitzer in the pulse mode. The operator continues to watch his screen.

When the pulses meet in the distant grid cell, a violent EMP suddenly arises in and throughout the missile and its surrounding vicinity. This explodes the missile propellants and warheads, destroying the target.

On the distant screen, a sudden drastic "blossoming" of the target results from the sudden extraction of a great deal of additional energy by the scanning scalar interferometer. The operator thus knows he has made a "kill."

Should for some reason the target be missed, additional energy will still be extracted by the scanning interferometer from the grid cell/vicinity in which the EMP suddenly emerges. The distant operator will still see a bloom on his scope, but not nearly so great as when the missile explodes. He thus knows that the EMP effect of the firing has occurred, but the target has been missed (he will see the bloom in its offset location as well). In that case, the operator can quickly mark the EMP location, hit the switch, and the computer will automatically correct and fire again.

At any rate, for a miss a large "sonic boom" or blast still results in the EMP emergence zone from sudden heating of the air. These are precisely the type of booms that were associated with three NASA shuttle launches prior to the end of 1985. These were actual testing of this weapon system, using the shuttle launches to provide a target, and delaying the burst some minutes so the shuttle would not actually be destroyed. On Nov. 26, 1985 a "marker beacon" (glowing ball of light) was also created over the site. This was probably to orient satellites and other detection systems. Some 12 minutes after that nighttime launch, a large blast occurred over the site, heard for hundreds of miles up and down the coast. Shortly after, the light suddenly moved away very rapidly - faster than a jet aircraft.

If many missiles are being launched, the operator marks them rapidly, one after the other, and the howitzer fires burst after burst at them, decimating the launched missiles, one after the other. In this fashion most of the counterstrike missiles launched by the U.S. would be destroyed shortly after launch, greatly reducing the number of missiles that make it to midcourse.

However, apparently the loss of the shuttle launched on Jan. 28, 1986 was caused by the addition of a metal-softening pattern in the exothermic mode, in and around the booster, using the actual booster ionic flames as a receiver-amplifier. This led to the failure of one of the supports and the partial breakaway of the right booster. The booster rotated into the tank, causing damage and the resulting explosion when the venting main fuel hit the booster flame.

By spread spectrum techniques and proper timing and phasing, one system of Woodpecker transmitters can set up multiple interference grids, in various parts of the world. The scanners and howitzers can operate in the appropriate zones, again by spread spectrum techniques and proper timing and phasing.

In fact, the latest versions probably can operate in the "ordinary EM carrier" (atmospheric) mode and the "scalar EM carrier" (underwater and in the earth) mode simultaneously - again by timed switching between EM and scalar EM carriers, spread spectrum techniques, and timing and phasing of howitzers and Manners.

Thus a single system can operate in the earth, underwater, and atmospheric modes. It can operate against the very wide range of different targets previously mentioned. It can do all this "simultaneously," from a real-time practical viewpoint.

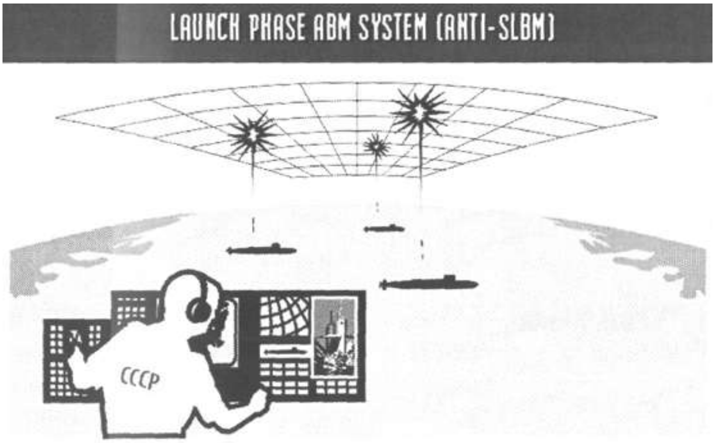

LAUNCH PHASE ABM SYSTEM (ANTI-SLBM)

Slide 067: Launch Phase ABM System (Anti-SLBM)

Here we show the operation of the system in the atmosphere over the ocean, to destroy sub-launched ballistic missiles such as Poseidon and Polaris. The system operation shown uses the normal EM Woodpecker carrier frequencies, with scalar components to establish the interference grid, and scalar channels for the scanners and howitzers. The grid can be set up by this method anywhere in the world.

As can be seen, during their early launch phase, this mode will take care of the missiles launched from underwater U.S., British, and French nuclear submarines.

Remember, this is just one mode in which the system is operating. It can at the same time have other interference grids at other locations, and even in other modes, including underwater grids used against the submarines themselves, it can be engaging a multitude of targets of a wide variety of types, essentially all simultaneously.

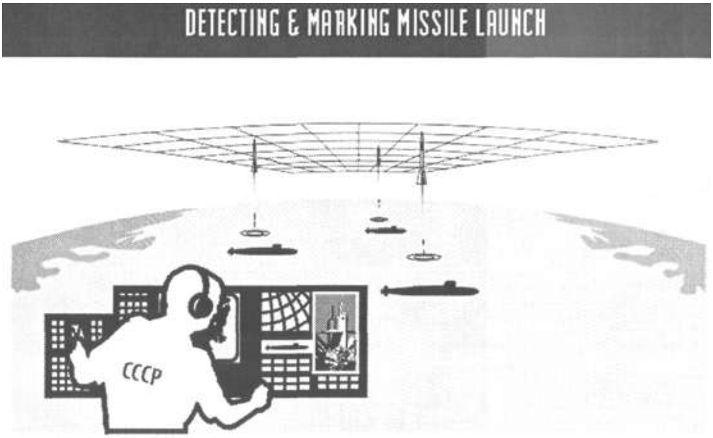
DETECTING & MARKING MISSILE LAUNCH

Slide 068: Detecting & Marking Missile Launch

Here we show the missiles launched from the underwater submarines, when just detected in the atmosphere over the ocean. As each rising missile is detected, an operator at the screen places a cursor (a "marker") over the detected track spot on the scope.

The computer quickly computes the firing parameters and indicates when the system is "ready to fire."

By using additional scalar scanner frequencies in the multi-gigaHertz range and fine scanning, the operator is able to accomplish non-cooperative target identification.

After the operator places a cursor over a target, the advanced system fine-scans and identifies it as hostile, then continues the track in automatic.

When the operator presses the fire button for a designated target, the computer intiates fire by the appropriate scalar EM howitzer through established grid scalar firing channels.

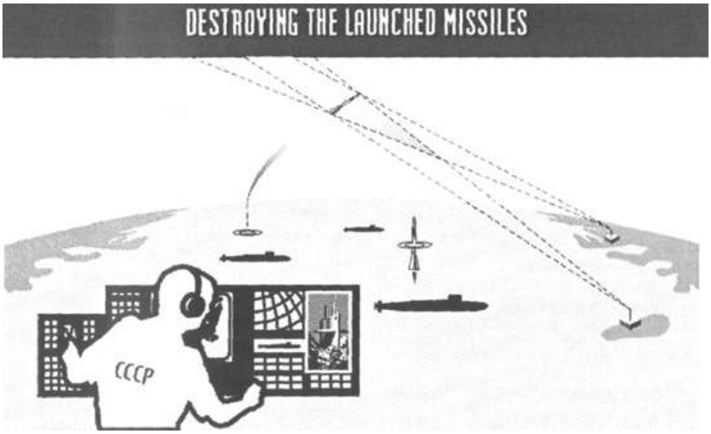

DESTROYING THE LAUNCHED MISSILES

Slide 069: Destroying the Launched Missiles

On this slide we temporarily omit the interference grid and diagrammatically show the operation of the scalar EM howitzer. As shown, two pulses are fired so that they meet in the grid cell where the rising missile is located. A drastic EM pulse suddenly emerges everywhere throughout the missile and its vicinity causing instant missile destruction.

Of course, individual howitzers not in the Woodpecker carrier's scalar channel can also fire at the missiles in the purely scalar exothermic mode. Or they can utilize the large "Tesla globe" in the continuous mode, the small Tesla "ball of light" in the continuous or pulsed mode, etc.

Also, continuous EM energy can be caused to emerge in a broader, "defocused" interference area causing electronic dudding of circuits in vehicles over a wider area. This is a very useful mode to use against mass attacks by many aircraft and/or many missiles at the same time. That is, the entire formation can be attacked in the "dud electronics" mode, and then followed up by more intense servicing of each individual hapless target.

Again, in 1960 Gary Powers was probably downed by a crude prototype scalar interferometer "howitzer" obtained by modifying two ordinary radars. At least the Soviets had previously been unable to down a penetrating U-2 with missiles they possessed. We know the scalar howitzers were in development, for Khrushchev had announced to the Presidium in January 1960 that a new "fantastic weapon" was in development. A development prototype was probably used to down Powers' aircraft.

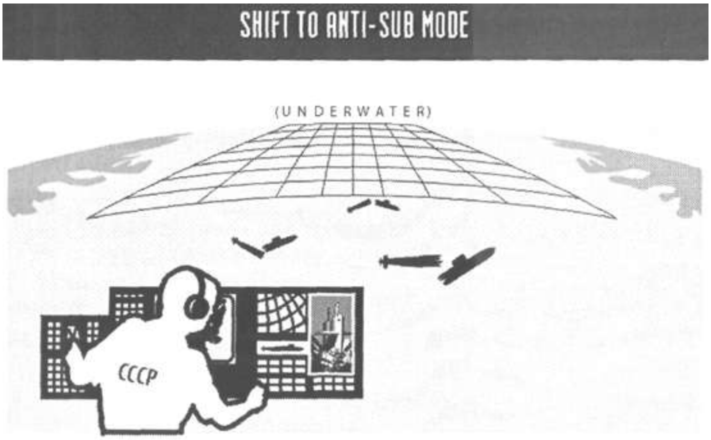

SHIFT TO ANTI-SUB MODE

Slide 070: Shift to Anti-Sub Mode

In the underwater mode portrayed on this slide, the Woodpecker transmitters have been switched to "scalar carriers" mode and the carrier beams transmitted through the ground and into the ocean underwater. (Remember, an artificial EM potential that does not couple to charged particles or ions will travel through the ocean, but a force field will not.)

Thus a giant scalar interference grid is now established underwater. Additional scalar beams in the weakly endothermic mode scan this grid. When a submarine enters the grid, it substantially changes the potential in that cell, and the scanner picks this up. The distant screen now displays the sub as a "track" of light in the appropriate grid square cell. The computer switches in fine-scanning and non-cooperative target identification. It identifies the target as hostile and alerts the operator (by blinking the track and changing its color to red, for example).

As before, the operator marks the target and enters the firing mode and type of action desired. Again, let us assume the "EMP burst" mode of destruction is to be used. The computer computes the firing parameters and readies the howitzer, When the commander gives the command to fire, the operator presses the firing button. The computer fires the howitzer.

Two scalar pulses proceed through the howitzer scalar channel and meet in the submarine. Throughout the sub and its vicinity, a violent EMP is suddenly experienced, knocking out the electrical systems, electronic equipment, and igniting propellants and explosives in the on-board missiles and armament. This explodes the submarine, missiles, and torpedoes instantly.

At the distant command console, the scanners detect the violent eruption of energy from the exploding submarine. It appears on the screen as a sudden strong blossoming of the track. The operator assesses that the target has been destroyed. He immediately engages another target if one is present.

If the EMP is off-course and misses, the scanners still pick up a smaller blossoming in the offset location. The operator immediately knows he has missed the target. He marks the offset bloom, presses the switch for "Correct!" and the computer makes the firing correction. When the operator again gets a Ready-to-Fire light illuminated, he presses the firing button and fires again at the target, this time with corrections applied.

Again, the same Woodpecker system - using timing, phasing, scalar/EM switching and spread-spectrum techniques - may be involved in several grids and many types of targets. When in a target-rich environment, the firing operators may be very busy indeed.

As we shall later see, the same weapon system may also be firing in the "cold explosion" or pused endothermic mode, where intense bursts of instant and severe freezing are unleashed in the target zone.

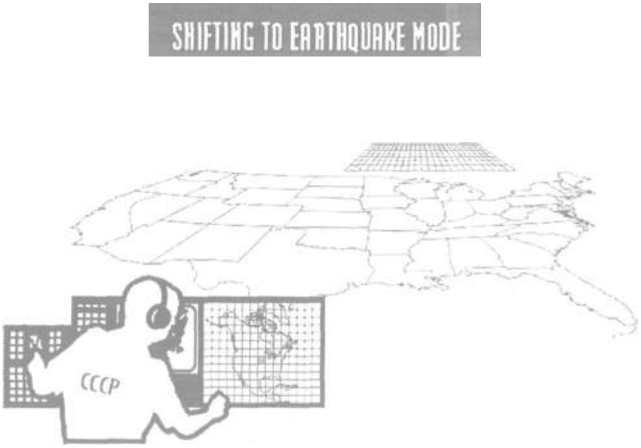

Slide 071: Shifting to Earthquake Mode

The uses of the complex grid, scanners, and scalar howitzers is limited only by human ingenuity and state-of-the-art of the technology developed.

By shifting to the scalar-carrier mode and transmitting in the surface layer of the earth, the interference zone can be established in a distant region of the earth, beneath the earth's surface.

Suppose we establish a very small interference grid in the earth at and around a distant fault zone. Then by steadily depositing energy in the zone in the continous exothermic mode, the stress in the fault area is steadily and gradually increased. Eventually the plates on each side of the fault will slip, and a "natural appearing earthquake" occurs. If the energy is deposited slowly and gently, a gentler quake can usually be effected. If the energy is deposited fairly rapidly, a higher peak stress can usually be reached before the rocks slip, and a larger, more destructive earthquake results.

By making a wider grid in the earth in a larger area, a strange kind of earthquake can be induced, whether or not a fault zone exists in the area. By pouring energy out of alternate cells and extracting energy from the others, a "wave-like" buckling tensile and compressive stress is created throughout the zone. As this continues, the entire plate area will try to oscillate, moving and causing an anomalous sort of "rocking, flat earthquake" throughout a large area.

Not too many days after the destruction of the Challenger on January 28, 1986, such an anomalous "flat earthquake" occurred in the U.S., near the Great Lakes, involving some 10 states. This was probably the subsequent testing of an additional weapons mode of the same Soviet scalar EM system that destroyed the shuttle.

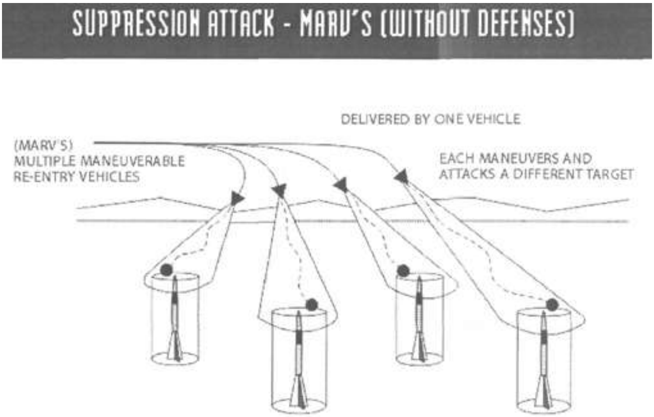

SUPPRESSION ATTACK - MARV'S (WITHOUT DEFENSES)

DELIVERED BY ONE VEHICLE

(MARV'S)
MULTIPLE MANEUVERABLE
RE-ENTRY VEHICLES

EACH MANEUVERS AND
ATTACKS A DIFFERENT TARGET

Slide 072: Suppression Attack—MARVs (without defenses)

Now to shift the emphasis a bit, and look at the impact on some "normal" strategic systems and concepts.

On this slide we show the basic idea of suppression attack against an enemy's ground based missile force in fixed-site silos. The attack uses multiple maneuverable re-entry vehicles (MARVs) delivered by a single carrier missile, which acts as a "bus." The bus ejects the MARVs at the appropriate time, and each of the MARVs heads toward its predesignated target, maneuvering as it re-enters the atmosphere and proceeds to the target. The maneuver makes conventional intercept of the MARV by conventional anti-ballistic missiles more difficult.

In this manner, missile silos without defenses may be caught by surprise and destroyed.

The "bus" concept actually multiplies the number of targets a single launched missile can attack, and the maneuvering of the MARV increases the difficulty of defending the target area by orthodox means.

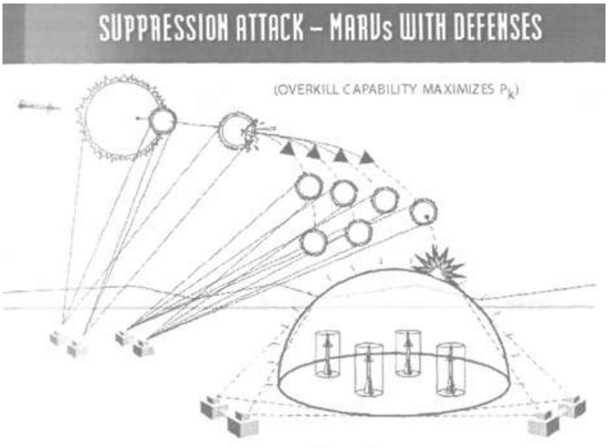

SUPPRESSION ATTACK – MARVs WITH DEFENSES

(OVERKILL CAPABILITY MAXIMIZES P_k)

Slide 073: Suppression Attack—MARVs (with defenses)

Now suppose some of the scalar EM defenses we have discussed are added to the picture.

On this slide we show a similar missile field, defended by a Tesla dome and by multiple scalar howitzers.
The howitzers take the incoming buses under attack during the early midcourse phase, by means of very large spherical EMP globes. Many of the buses - and their entire payloads, including the nuclear warheads themselves - are dudded by these midcourse defenses.

Servicing of dudded vehicles that have penetrated the globe now occurs farther along in midcourse. Each dud vehicle may be serviced, for example, with a small globe in the continuous mode for an appreciable number of seconds. This causes vaporization or fusing of the physical structures of the vehicles.

Groupings of incoming vehicles may also be serviced by medium-sized globe in the pulse mode. This assures dudding, and is especially useful for clearing away chaff and debris with smaller ballistic coefficients. It also is useful to further insure that the nuclear warheads themselves are totally dudded.

An additional wrinkle is to place scalar signal patterns - that affect and transmute a nuclear warhead's radioactive nuclei - onto a continuous globe. In short, by the proper scalar signals it is possible to transmute plutonium and uranium into harmless elements, and it can be done at a distance. The 4th and 5th generation Soviet scalar EM technology is probably capable of doing **this**

This insures that, not only are the incoming nuclear warheads electromagnetically inert, but they are also nuclearly inert and incapable of generating nuclear explosions.

Still another wrinkle - particularly useful against implosion-type weapons - is to utilize a scalar signal pattern that causes sudden generation of neutrons in the active material (plutonium or uranium). This harmlessly explodes the nuclear warheads while they are far away from their intended targets and out in space. The burst mode is probably the most useful for this tactic.

Note that, if desired, this same tactic - exploding the nuclear weapons themselves - can also be utilized in the Launch Phase ABM system grid against the missiles after launch. This capability converts an adversary's intended nuclear attack against you to an actual attack against himself!

Also note that the adapted grid system in the surface of the earth can be used to explode the nuclear warheads in ICBMs while they are still in the silos, or to explode nuclear warheads in storage. In the underwater mode, the nuclear warheads in the Poseidon and Polaris missiles on-board the nuclear submarines can be exploded in a similar manner. In fact, the nuclear reactor used to furnish the power for the nuclear submarines can also be exploded or disabled in this manner, as can land-based and ship-based reactors.

Again, this actually turns an adversary's own nuclear weapons against him. His own weapons can be used to obliterate him. Without scalar defenses, possessing the capability of nuclear attack becomes a serious detriment!

But back to our slide. Just in case something gets through this formidable defense somehow, the Tesla Shield (dome) provides the final reiteration that takes care of any residue "leakage."

The sum total of these defenses fully justifies the 1968 Soviet statement in Military Strategy that the USSR has achieved a 100% defense, and the West has

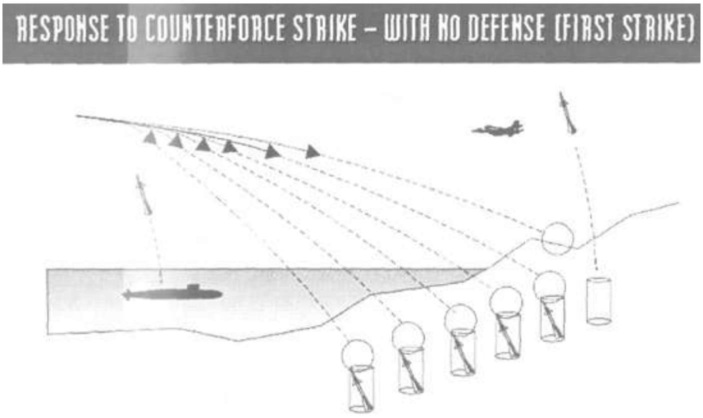

RESPONSE TO COUNTERFORCE STRIKE – WITH NO DEFENSE (FIRST STRIKE)

Slide 074: Response to Counterforce Strike with No Defenst

Here we show the orthodox concept of response to a surprise counterforce nuclear strike when the surprised force has no defenses.

A great percentage of the attacked missile silos are destroyed. The surprised side thus is able to mount only a limited strike in retaliation. Perhaps he is able to launch some remaining silo-based missiles that escaped destruction. He is also able to launch missiles from any of his underwater nuclear submarines that escaped. Some of his strategic bombers may have escaped destruction on the ground or while airborne on station; these surviving bombers can be directed to proceed with strategic reprisal against the enemy.

As can be appreciated, this entire concept is now totally obsoleted by the Soviet possession of scalar EM weapons.

The "mutual assured destruction" capability - or MAD concept as it was so aptly named! - is no longer mutual.

And for two reasons: (1) the Soviet surprise attack could destroy essentially all our Triad forces wherever they are, in the first attack. (2) Even if a very small number of our strategic forces survive, the puny strike we could attempt to launch could be repeatedly decimated from launch to terminal phase.

The restraint on the Soviets is not because of our Triad forces. The restraint is twofold: (1) a second nation has scalar EM weapons, and would undoubtedtly unleash them upon the Soviet Union with great violence if such a massive world-wide destruction of U.S. forces occurred or was threatened (is it really "a**ccidental"** that six major Soviet missile ammunition storage sites mysteriously exploded within seven months?); (2) the closed-loop scalar radiation exchange system of Earth-Sun could be inadvertently "tweaked" in the feedback loop from Earth to Sun, so that a large solar scalar resonance was stimulated. In that cas**e** the Sun could emit a mighty resonant "burp" that would simply wipe out all life on Earth in a fiery shower.

The "terror" in the "balance of terror" between nations has just increased by many, many orders of magnitude.

The biblical prophecy that the Earth will be wiped out by "fire and brimstone from heaven" is very near fulfillment now.

Put another way, the pucker factor - even on the Russians who may contemplate unleashing Armageddon - is higher than any Western strategic analyst has ever imagined, even in his wildest nightmares.

Truly these weapons are "more frightening than the mind of man has imagined," as Brezhnev put it in 1975.

The frightening scalar EM weapons can be used, but only very, very gingerly indeed**.** If a slight mistake is made, everybody loses everything.

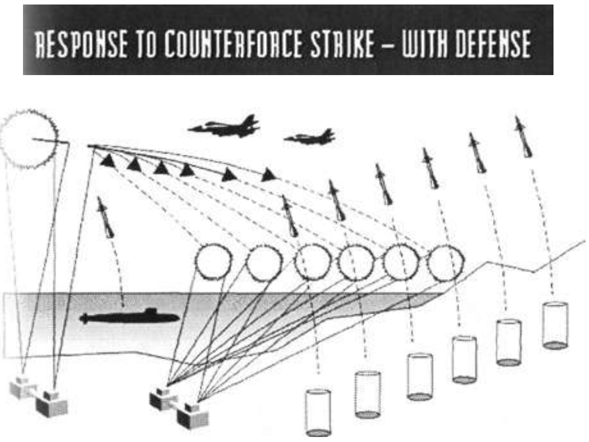

RESPONSE TO COUNTERFORCE STRIKE – WITH DEFENSE

Slide 075: Response to Counterforce Strike with Defense

On this slide we show the impact of scalar defenses against an adversary who possesses only the orthodox nuclear weapons and delivery systems, and makes a "first strike" by surprise against his opponent.

In this case, the attacked opponent has scalar EM defenses.

With just the midcourse and terminal scalar EM howitzer defenses, almost all of the attacking missiles and re-entry vehicles are destroyed.

If for some reason (such as to prevent triggering the sun and destroying the world inadvertenly!) the attacked adversary does not wish to attack strategically with scalar EM weapons, he may now launch a massive retaliatory strike on his opponent.

And if that opponent has no ABM defenses, he's just bought the farm.

What a funny coincidence! Long ago, we crazy Americans chose to put all **our** eggs in the offensive basket. That is, we thought the Soviets cannot dare to strike us because we will have enough surviving missiles to penetrate any orthodox defense and destroy them in return. They would also have the same "mutual assured destructon (MAD)" capability, of course!

The MAD concept was aptly acronymmed - it was an insane idea from its inception. That is, the Soviets were not supposed to defend their cities **and** populace, and neither were we. That way, you see, we could destroy their citizenry at will, and they could destroy ours at will. We deliberately agreed to that, thinking - in our insanity - that the Soviets agreed also.

Of course the Soviets weren't playing the same lunatic's game at all. They simply fed our own garbage back to us, and we importantly thought we were negotiating something. Unknown to us, they were secretly building and deploying scalar EM defenses even then. Since we didn't know of any other defense, we assumed the Soviets could not defend their citizenry with ABMs any better than we could defend ours with ABMs. So we blithely scrapped our ABMs.

The high government officials of the United States of America grossly violated our constitution and their sworn oaths of office. Our constitution specifically charges the government of the United States to defend the people of the United States. That means our officials are specifically charged with defending our people. Yet they deliberately agreed not to do so, and they deliberately sacrificed us in advance.

They committed treason of the highest kind, by deliberately surrendering us to total destruction anytime the big balloon goes up.

The constitution doesn't charge the government with surrendering our citizens to destruction under any circumstances. It charges the government with defending us against destruction.

Ronald Reagan is the first U.S. president in some time to see through the MAD concept. At least he stands forthrightly for defense. It's just a shame that the government and scientific leaders under him aren't aware of the Soviet scalar weapon developments, so they could brief him on these startling developments.

Be sure you understand what the government leaders before Reagan have been saving. For years, our official policy has effectively said to the Russians: "We will allow each side a token defense only. We deliberately will not defend ourselves, so that anytime you decide to strike us you can destroy us. In return, however, we do not think you will want to strike us, for we believe that - before we become ashes from that strike - we can get enough away at you, to destroy you as well. Of course we believe you are in the same sort of boat."

Then we said, "There! That's settled! What a relief! Nobody can strike anybody anymore. Now can we talk about partial disarmament and on-site inspection?"

The Soviets just kept their astonished amusement to themselves and kept building scalar defenses (and scalar offenses as well!). Their deception plan called for imbedding the scalar EM weapons in ordinary systems when possible, and hiding the scalar EM weapons tests in ordinary weapons tests when possible. Of course they had to be adamant in their opposition to on-site inspection; after all, even Western scientists, if allowed on site with the dual-mode weapons, might suspect something was different and possibly things were amiss.

The Soviet deception has worked like a charm. The U.S. *MAD*men who created and implemented the MAD doctrine have succeeded in prejudicing two generations of academicians and elected U.S. leaders against the need for effective defense. A vast wasteland of inane paper studies - learnedly dealing with ABM's, penetration tactics, leakage, force and counterforce, radar blackout, electronic warfare, etc. etc. - has been churned out by our military, industrial, and university "think tanks" to clog the minds, efforts, and literature of our defense and governmental establishments.

Our *homeland* has no effective ABM defense (and very little defense against aircraft or subsonic cruise missiles).

Our field army has no effective ABM defense (or air defense). Our *naval forces* have no effective ABM defense (although they are better defended against air attack than the field army). Yet ballistic missiles - and to some extent cruise missiles - have supposedly been our greatest threat for decades, followed closely by the secondary threat of massed aircraft attack.

A cynic might take the stand that, if our Defense Department hasn't been allowed to defend us (with *deployed* weapon systems, not by "paper studies") against what we ourselves have proclaimed is our biggest threat, then why is it called the *Defense* Department? Why not call it the *No-defense* Department'?

Understand, a great many fine, hard-working people in DoD have certainly tried to get adequate defenses developed and deployed. The sheer weight of the bureaucracy and its mind-numbing entanglements has prevented it.

But it's even worse. Our bureaucrats - both in government and in "big, organized science" - have been so caught up in the "Russians are really just ignorant peasants" syndrome and the "not-invented-here" (NIH) syndrome **that** they've not recognized the greatest scientific and technical breakthrough in history. Such a breakthrough - far greater in importance that the nuclear bomb has been made, developed, and deployed by the Soviet Union, all in secrecy and behind the smokescreen of a masterful deception plan.

We've totally missed the most incredible "secret superweapons" development program of all times, and the worldwide testing of the weapons themselves.

Hiroshima and Nagasaki should serve as object lessons in the price of ignorance of such technological breakthroughs.

History does repeat itself. But this time it's not we Americans who got the growl new superweapon first.

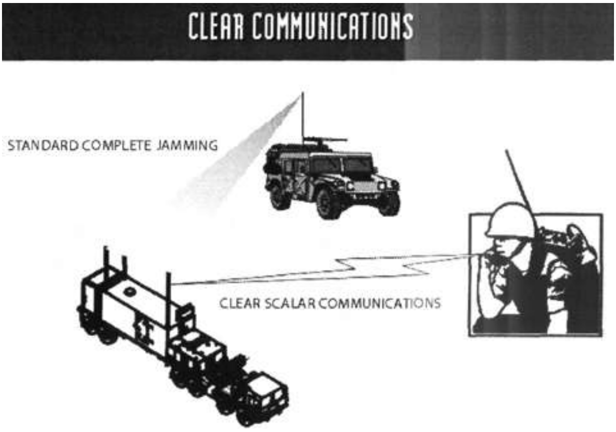
CLEAR COMMUNICATIONS

Slide 076: Clear Communications

Now let us rapidly give some unique capabilities that one can do with scalar EM in smaller things. Not only can scalar EM weapons be used in the big "continent burning" strategic weapons, but they can also be used for a variety of smaller defense tasks.

For example, one can establish a zero reference anywhere one wishes in an ordinary carrier. Inside that zero, one can establish as many independent "carriers" as one wishes, and hide them completely from ordinary detectors.

The way to do that is easy. Take as complex a signal as you wish. Subtract it from zero. That gives a "complement" which, when summed with the original signal, will produce the zero. Then only a scalar detector can go inside the zero and pullout the individual components.

So take a carrier wave. Let the enemy jam it completely, with gobs of power.

Your scalar communication channel stays serenely there, calm and unjammed. The enemy thinks you're jammed and can't communicate. Meanwhile, you continue to communicate just fine.

Or be real smart. Take one of those components inside the first zero. Pick a zero reference inside it. Hide still more carrier components inside the second zero. Do this as many times as you wish.

You can put all the telephone calls in the United States of America on one carrier. And you can keep them clear, right through total jamming.

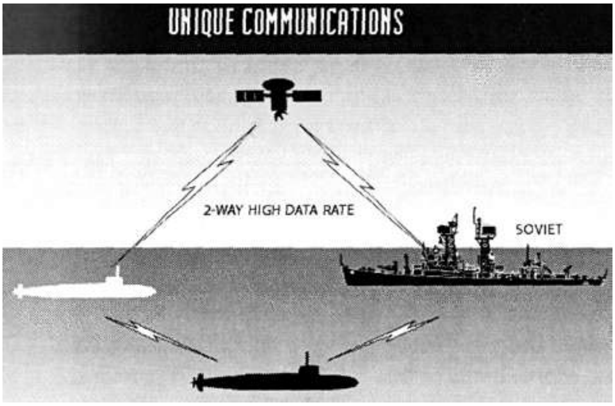

Slide 077: Unique Communications

But that's still not all.

You can also communicate with your submarines underwater. High data rate. Chat away as you wish. You can scramble, frequency hop, encrypt, the works. They can also communicate back to you.

Remember. EM force fields of any appreciable frequency won't travel through the water. Natural EM potentials grab charged particles and ions, so they won't travel through the ocean either. But artificial potentials don't grab onto charged particles or ions, and they will go through the ocean quite nicely.

Everything - aircraft, ships, shore installations, submarines, etc. - can communicate like gangbusters if they possess scalar EM technology.

In fact, you can even build scalar EM underwater radar if you wish.

The airborne scalar EM radar can see under the ocean with a little adjustment. perhaps as well as a conventional side-looking radar sees distant targets today, The airborne scalar EM radar can detect and track the underwater submarine while tooling along.

In fact, the airborne scalar EM radar can then destroy the sub, by using a **little** scalar interferometry. And it's even possible to do scalar interferometry with a single transmitter! How, is proprietary to a friend.

In fact, you can even use your enemy's jamming transmission as a "wire" along which to establish your scalar EM channels. You can use these scalar channels In the "weakly endothermic mode" to locate his transmitter and "lock-in" on it. You can "walk" the lock-in right past his encryption function to the "clear" Input, and receive and record that in the clear. All this from a distance.

You can then destroy his equipment if you wish, or just continue to monitor him in the clear if it's important to do so. Perhaps now we can understand what the decades-long weak microwave radiation of the U.S. Embassy in Moscow is all about. Or why the Soviets are building a nice new Embassy, bristling with antennas, on the high ground in Washington, D.C.

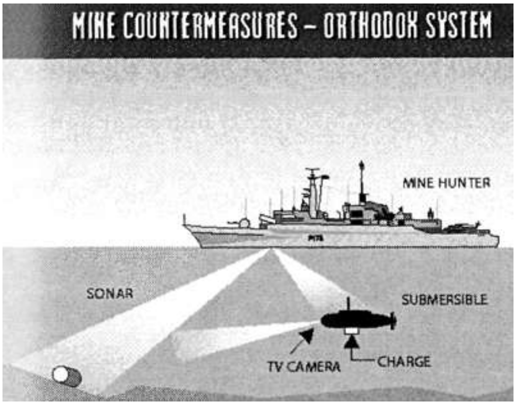

MINE COUNTERMEASURES – ORTHODOX SYSTEM

Slide 078: Mine Countermeasures—Orthodox System

Take something like mine countermeasures, for example. Finding and destroying mines - particularly those placed on the bottom in relatively shallow areas such as ports and rivers - is a really tough problem.

Presently it's often done as portrayed on this slide.

Here a minehunter ship uses an underwater sonar to locate the mine, or what resembles a mine. Then it directs a submersible robot vehicle to proceed over to the object and examine it with an underwater television camera. If it's a mine, the submersible places a remotely operated charge beside it, then chugs away. When it reaches a safe distance away, the charge is exploded, causing the mine to detonate.

A very painful process. Very time consuming, and very difficult.

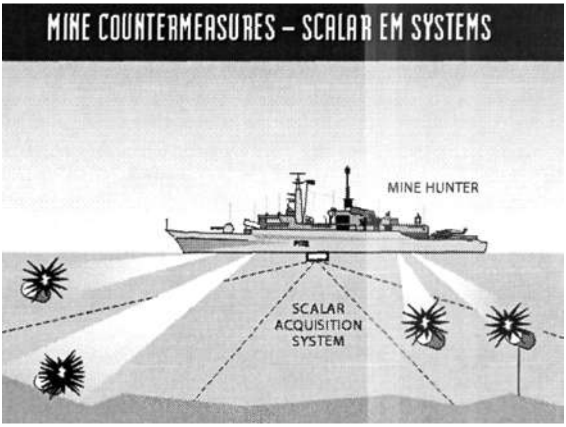
MINE COUNTERMEASURES – SCALAR EM SYSTEMS
MINE HUNTER
SCALAR ACQUISITION SYSTEM

Slide 079: Mine Countermeasures—Scalar EM Systems

Now try the same thing while using scalar EM underwater systems. The scalar EM acquisition system is really a sophisticated scalar radar. It scans a large area rapidly, locating moored mines, bottom-planted mines, and drifting mines.

The acquisition system is also a scanned "underwater eye." The operator in the surface vessel can tell whether or not it's a mine he's located.

Once the target mine is detected, a computer computes the firing parameters for a scalar interferometry pulse to hit it and destroy it. The operator fires the pulse, destroying the mine.

The same system that detects the mine at a distance identifies and destroys it at a distance. In this way large areas can be quickly scanned and swept clean of mines.

If a suspicious object is detected and one is still not quite sure it's a mine, shoot it anyway! If it's a mine, it explodes. If it's just a rock, nothing happens. So the same system can do "discrimination by fire."

The same system is useful to destroy midget submarines, bottom crawlers, etc.

It can also wipe out whole arrays of underwater sensors by electromagnetically dudding them.

It can take on a sub if it has to! Here it would just gently kindle EM energy inside the sub's controls, causing it to lose control of all its systems. The sub would then sink to crush depth and implode. The scalar EM mine sweeping system could, of course, simply "blow" the sub, but might itself get caught in the resulting explosion.

SCALAR ANTI-PERSONNEL WEAPON

Slide 080: Scalar Anti-Personnel Weapon

From a military viewpoint, these scalar EM weapons are really quite all-around weapons.

For example, they are very lethal against personnel.

Set in the "high intensity pulse mode," one blast and a person's nervous system is destroyed instantly. Total, instantaneous death results. One can blast away at entire groups of infantrymen, for example. It's the peak power that's important, not the average power. So the weapon doesn't require too much battery power.

Set on a lower intensity pulse mode, one zap can simply knock out a person, rendering him instantly unconscious.

Quite useful if one needs prisoners to interrogate. Also very silent, which is useful in operations employing stealth and surprise.

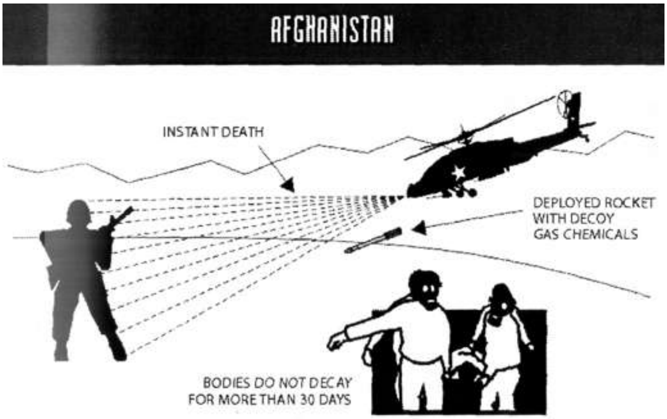
AFGHANISTAN
INSTANT DEATH
DEPLOYED ROCKET WITH DECOY GAS CHEMICALS
BODIES DO NOT DECAY FOR MORE THAN 30 DAYS

Slide 081: Afghanistan

The Soviets are employing just such a scalar "death ray" weapon in Afghanistan, apparently in the noses of some of their HIND helicopters.

These choppers usually fire gas rockets to camouflage the testing of the tactical scalar death ray. For example, they may fire rockets filled with nerve gas, and some persons will die convulsively from the chemical.

Those hit by the scalar EM weapon, however, have a most peculiar death mode,

Death comes instantly and totally. There is no convulsion, no response. The entire nervous system is destroyed instantly. Every living cell in the body is killed instantly, including all bacteria, germs, etc.

A body hit with this thing falls like a limp rag and lies where it falls. It doesn't decay in even 30-45 days. In a macabre fashion, it's been reduced to something like food irradiated with nuclear radiation; everything is killed, so the material is preserved for an extended period before any decay can set in.

Western military authorities have speculated that this strange death is due to some sort of eery new gas.

However, even with the most virulent gas, some seconds are required for a body to die. And while it's dying it goes into convulsions. Violent convulsions.

This eery, instant death that leaves a non-decaying corpse is not due to a gun, This is the signature of a high-powered pulse kill using a scalar EM weapon

The associated gas rockets are the disguise used to fool our intelligence analysts.

They're just part of the Soviet deception plan.

ADDITIONAL NOTES AND REFERENCES

15. For details of the mysterious, instant deaths in Afghanistan, see Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," Jane's Defence Weekly, 1(13), Apr. 7, 1984, p. 508

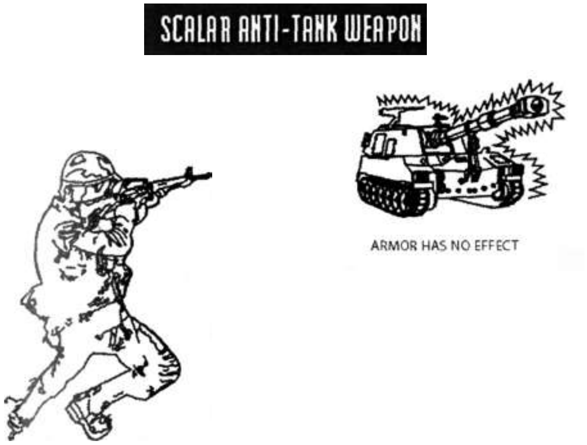

SCALAR ANTI-TANK WEAPON

ARMOR HAS NO EFFECT

Slide 082: Scalar Anti-Tank Weapon

Such weapons are also effective against armored vehicles, for example.

They can be carried by personnel, mounted on vehicles, mounted in helicopters, etc.

They are truly all-purpose weapons.

An infantryman with a bazooka-sized scalar EM pulse weapon can easily knock out a tank with one shot. The armor has no effect; scalar EM goes right through conductive shielding and Faraday cages.

And with the same weapon he can knock out another tank, and another, and another...

SCALAR ANTI-AIRCRAFT WEAPON

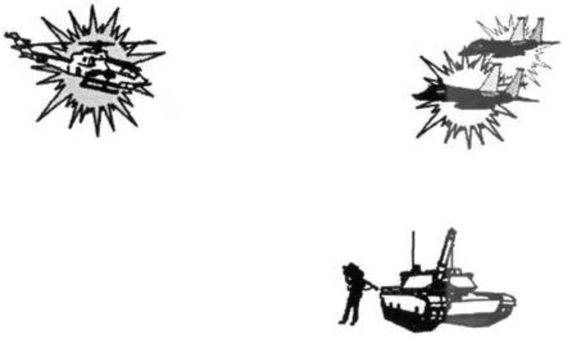

Slide 083: Scalar Anti-Aircraft Weapon

And, if a helicopter or a fixed-wing aircraft attacks him, he can bring the same weapon to bear on it.

He can knock down a chopper or a jet. With one shot.

And another. And another.

Also, it's a line-of-sight weapon. He doesn't have to compute lead angle and all that jazz. Just aim, get the target in his sights, and pull the trigger.

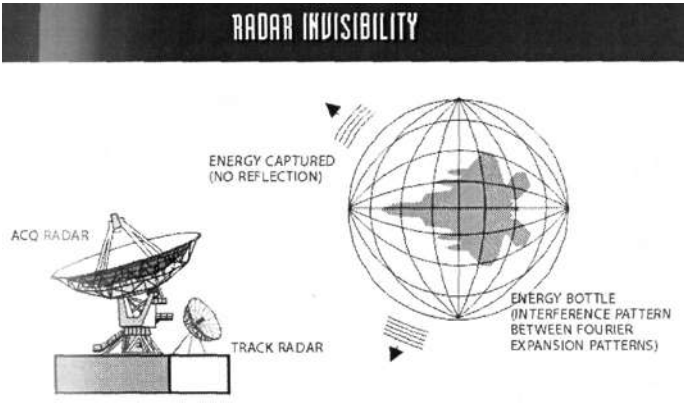

RADAR INVISIBILITY

Slide 084: Scalar Anti-Aircraft Weapon

Another neat trick you can do with scalar EM is make an aircraft invisible **to** ordinary radar.

Just use multiple transmitters on the airplane to establish a spherical interference shell, in the bandwidth of the threatening radar, around the aircraft.

When the signal pulse (or continuous wave) from the ground-based radar **strikes** the "energy bottle" in the shell, it interferes with it and is diffracted, scattered, and partially absorbed.

The ground-based radar doesn't get any coherent return pulse or return signal reflected from the target.

Radars don't track aircraft; they track the reflection of their own signal bouncing off aircraft. No reflected signal, no aircraft - as far as a radar is concerned.

To shield against a passive radar (one which tracks emissions from the aircraft), just detect and complement your own emissions so that "near zeroes" result In the spectrum of interest.

Or do both, to protect against both.

It's even possible to cool the exhaust and heated surfaces of the aircraft by a scalar EM adaptation of electrostatic cooling. Theoretically it's even possible to cause the aircraft to "disappear" optically. Takes a bit of doing, but it's possible The basic scheme here is that, when a photon hits the target from a given direction, it's absorbed totally, and a corresponding or "matching" photon is emitted from the opposite side in the same direction the original photon was going. In that fashion, the light emitted from the target from one direction seems to have "passed through it" from the other direction.

Since that's what empty space does, then what a distant observer sees is "empty space," so far as he can tell.

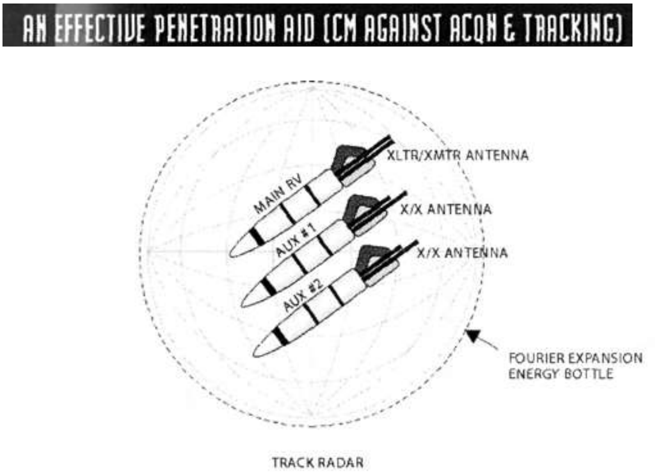

AN EFFECTIVE PENETRATION AID (CM AGAINST ACQN & TRACKING)

XLTR/XMTR ANTENNA
MAIN RV
X/X ANTENNA
AUX #1
X/X ANTENNA
AUX #2
FOURIER EXPANSION ENERGY BOTTLE
TRACK RADAR

Slide 085: An Effective Penetration Aid

Of course you can do the same thing for re-entry vehicles.

Use some multiple antennas on a vehicle. Or better yet, use satellite RV's to carry the extra transmitters necessary for scalar interferometry.

Deploy the interferometer RV's and activate the scalar EM transmitters. Place an energy bottle Fourier expansion interference shell around the ensemble in the bandwidth of interest.

Radars in that bandwidth now will be unable to detect the incoming ensemble. If nothing detects the threat, nothing is going to shoot at it or react to it.

This can also be used on a bus vehicle during midcourse, for example.

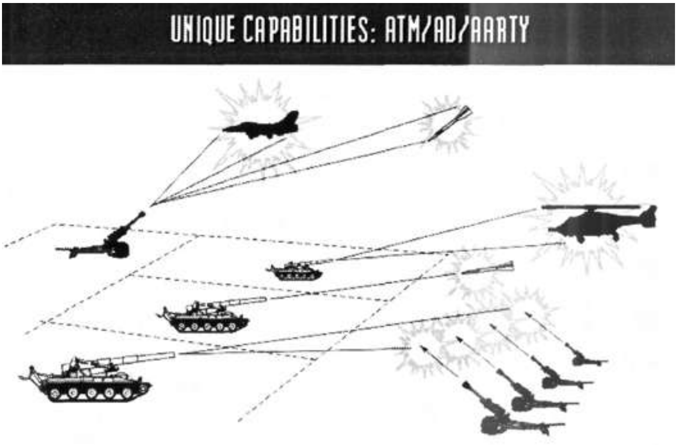

UNIQUE CAPABILITIES: ATM/AD/AARTY

Slide 086: Unique Capabilities: ATM/AD/AARTY

But the uniqueness of scalar EM devices and weapons doesn't end there.

You can hit lots of other threats that are presently difficult or impossible to effectively handle.

For example, on the tactical battlefield, the same scalar EM radar interferometer system can defend against tactical missiles (ATM role), aircraft (AD role), cruise missiles (ACM role), and even artillery and mortar shells (AARTY role).

How the devil, you might ask, can it defend against artillery and mortar shells?

Simple. Sweep the artillery or mortar barrage with closely spaced pulses. Sorta like a machinegun. Or more like a phased-array radar.

If the peak power in the sweeping pulses is sufficient, the explosives are sparked and detonated. Fuses are destroyed.

So even mortar and artillery rounds can be destroyed, by the same system providing air defense and anti-tactical missile defense.

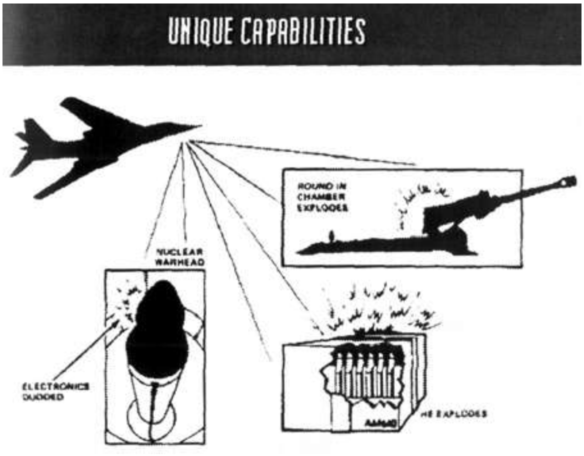
UNIQUE CAPABILITIES
ROUND IN CHAMBER EXPLODES
NUCLEAR WARHEAD
ELECTRONICS DUDDED
HE EXPLODES

Slide 087: Unique Capabilities

That's still not all.

An incredible number of different types of targets can be attacked with scalar EM weapons.

For example, aircraft equipped with scalar EM weapons can attack targets such as loaded artillery pieces. The round in the chamber explodes, destroying the weapon.

It can also attack nearby stored ammunition. Again, the high explosive in the ammunition explodes, destroying the store.

Even underground missiles can be attacked. The scalar pulse will go right through the overhead cover and dud the electronics, including the nuclear warhead. It will also detonate the HE in the warhead and detonate the fuels and propellants.

A great variety of other targets can also be attacked, as we shall see.

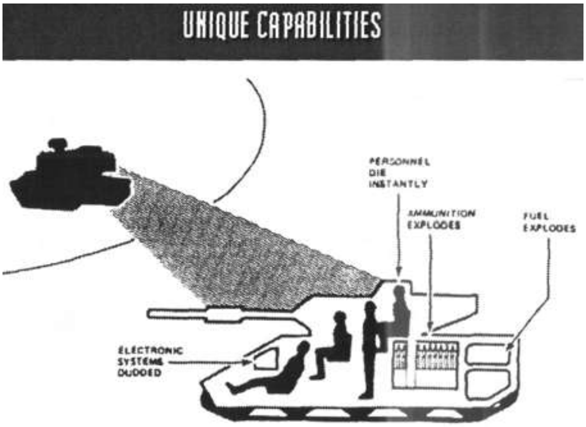
UNIQUE CAPABILITIES
PERSONNEL DIE INSTANTLY
AMMUNITION EXPLODES
FUEL EXPLODES
ELECTRONIC SYSTEMS DUDDED

Slide 088: Unique Capabilities

For example, suppose one attacks a tank.

With scalar EM pulse, the personnel die instantly. Total personnel kill is achieved.

All electronic systems of the tank are dudded. Total systems kill is achieved.

The ammunition in the tank explodes. Total kill is achieved.

The fuel explodes. This is another total kill mechanism.

That tank has been killed totally, by a variety of mechanisms, all simultaneously.

That's a k-kill in any analyst's book.

And **don't** worry about retrieval and repair. That tank is finished permanently.

Even **the** most modern tank is just as vulnerable as the most obsolete.

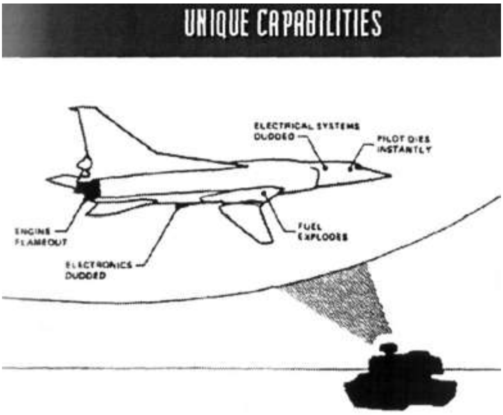

UNIQUE CAPABILITIES

ELECTRICAL SYSTEMS DUDDED
PILOT DIES INSTANTLY
ENGINE FLAMEOUT
FUEL EXPLODES
ELECTRONICS DUDDED

Slide 089: Unique Capabilities

Suppose one attacks a fighter-bomber aircraft.

Again, the kill is catastrophic from multiple simultaneous kill mechanisms.

The pilot and crew die instantly. Total destruction of their nervous systems.

The electrical systems are dudded.

The electronics in all avionics and ordnance is dudded.

The fuel explodes.

HE in on-board or out-board ordnance explodes.

The jet engine flames out if the pulse is powerful enough.

That aircraft is finished. Permanently.

And it's point-and-shoot, line-of-sight.

Even the most modern aircraft is just as vulnerable as the most obsolete.

UNIQUE CAPABILITIES: PENETRATE SHIELDING

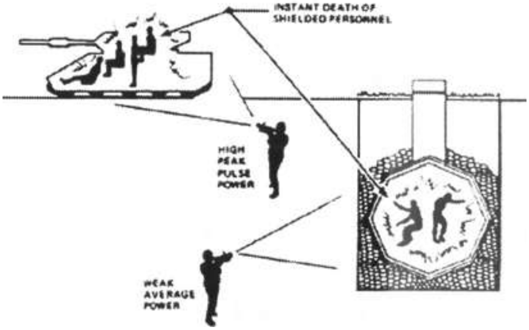
INSTANT DEATH OF SHIELDED PERSONNEL
HIGH PEAK PULSE POWER
WEAK AVERAGE POWER

Slide 090: Unique Capabilities: Penetrate Shielding

Shielding is ineffective.

The scalar pulse (or continuous wave) can go anywhere gravity can go.

It penetrates the steel hull of tanks and armored vehicles.

It penetrates underground shelters and bunkers.

It penetrates pillboxes and fortifications.

Foxholes are no protection. Just shoot through the earth into them.

There is no longer any place to hide.

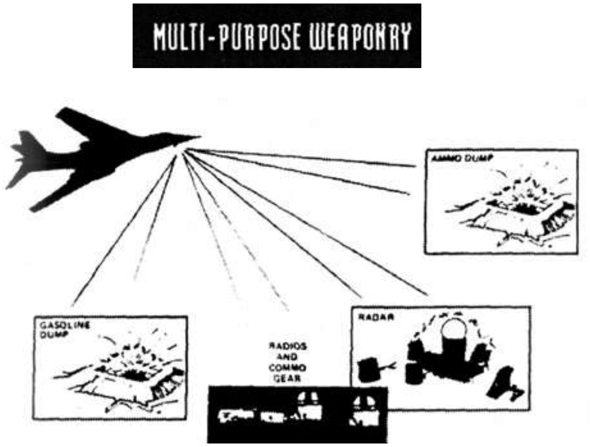
MULTI-PURPOSE WEAPONRY
AMMO DUMP
GASOLINE DUMP
RADIOS AND COMMO GEAR
RADAR

Slide 091: Multi-Purpose Weaponry

These scalar EM weapons are all-purpose. Things such as ammunition dumps, radar systems, radios and communication gear, communications centers, gasoline dumps, etc. can easily be destroyed.

Not only that, but the weapons can be swept, giving area coverage.

You can hit both point targets and area targets.

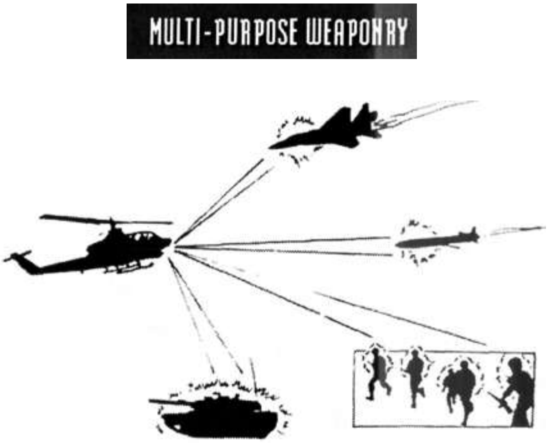

MULTI-PURPOSE WEAPONRY

Slide 092: Multi-Purpose Weaponry

Look at what a weapon a helicopter becomes.

The same vehicle, equipped with scalar EM weapons, can engage and destroy jet aircraft, other helicopters, cruise missiles, personnel, and armored vehicles.

It can also hit all the other targets we've been talking about. Any vehicle equipped with scalar EM weapons becomes a formidable destroyer on the battlefield.

And "reconnaissance by fire" takes on a particularly lethal meaning.

For example, don't worry about the enemy in cover and concealment, or highly camouflaged.

Don't worry about decoys or dummy positions.

Just sweep the area with scalar EM fire. The actual targets will be discriminated by fire and destroyed.

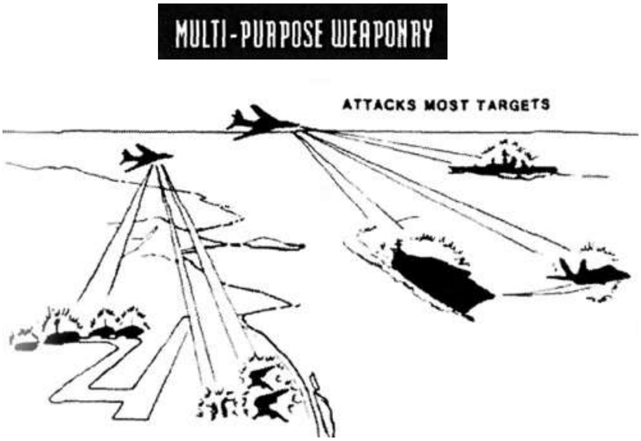

MULTI-PURPOSE WEAPONRY

Slide 093: Multi-Purpose Weaponry

As can be seen on this slide, most targets of interest can be attacked.

A single aircraft can sink an aircraft carrier or destroy a battle cruiser from afar.

It can also destroy aircraft, ground installations, artillery, personnel, nuclear weapons storage sites, tanks, vehicles, fuels, and combustibles.

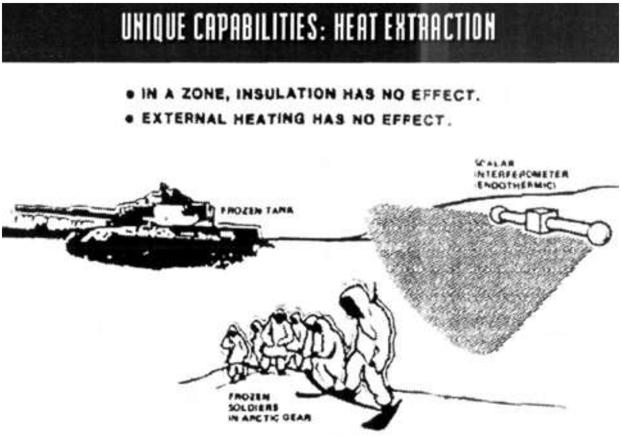

UNIQUE CAPABILITIES: HEAT EXTRACTION

- IN A ZONE, INSULATION HAS NO EFFECT.
- EXTERNAL HEATING HAS NO EFFECT.

SCALAR
INTERFEROMETER
(ENDOTHERMIC)
FROZEN TANK
FROZEN
SOLDIERS
IN ARCTIC GEAR

Slide 094: Unique Capabilities: Heat Extraction

Or one can get truly exotic.

To prevent destroying equipment and installations, the endothermic mode can be utilized.

Cold explosions can be used to freeze tanks, personnel, and equipment.

The equipment and tanks thaw out.

The personnel thaw out too, but they are dead when they do. Heaters and arctic clothing are no protection.

UNIQUE CAPABILITIES: COLD EXPLOSIONS

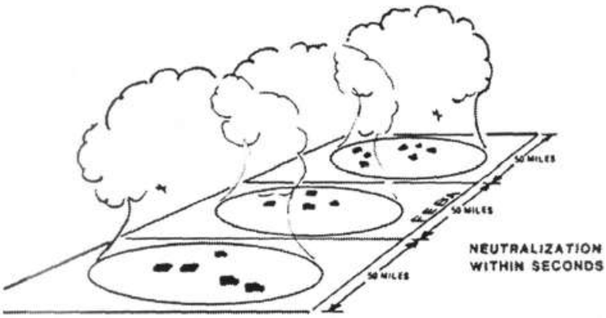

Slide 095: Unique Capabilities: Cold Explosions

For example, suppose three cold explosions, each 50 miles in diameter, are placed upon the front-line strength of NATO.

In seconds there is no longer any "front line strength" of NATO. Every living thing is frozen into a block of ice.

A few more of these things placed on selected rear area installations, and the battle is over before it gets started. Say in about one minute flat!

Suppose NATO is quickly defeated - say in two days. And then the Russians offer us a "Dunkirk." They agree to allow us to evacuate the beaten remnant of our forces.

Are we going to evoke a strategic nuclear exchange, or are we effectively "decoupled" from Europe? Particularly if the Soviets also offer sweeping accommodations to the Europeans, and they accept? If we punch the strategic button, we guarantee our own absolute destruction. (MAD doctrine and all **that,** old chap!)

Whether or not we could do anything to the Russians is a matter of conjecture.

For a demonstration, they could even invite us to launch a couple of missiles, and destroy them over our heads shortly after launch.

And they could take out a few submarines under the ocean, just to make it believable.

And destroy some SAC bombers that are orbiting on station.

Or take out a U.S. Naval Task Force, with a nuclear carrier, as a demonstration that they mean business.

Would a sane President go ahead and destroy us, or accept the Soviet's "Monroe Doctrine" and an "armistice"?

If the Soviets thought he was seriously contemplating ordering a nuclear strike, they could immediately destroy all our nuclear armaments where they sit. Our own weapons would obliterate us!

What would Congress do - after the Soviet Union had demonstrated its absolute power, offered a Dunkirk and armistice, and threatened our total destruction if we "reached for the holster," so to speak - when it realized the President might actually be contemplating evoking assured national suicide?

What would you do if you were a Congressman? If you were the President?

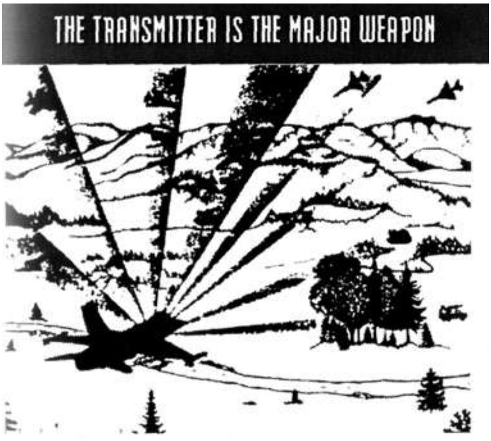

THE TRANSMITTER IS THE MAJOR WEAPON

Slide 096: The Transmitter is the Major Weapon

As shown on this slide, with scalar EM the transmitter becomes the major weapon on the battlefield.

Almost every type of target of any importance can be easily and effectively attacked and destroyed or nullified by scalar EM weapons.

Small wonder the Soviets do not discard their thousands of "old" radars!

How effective do you suppose some of their "old" transmitters on their "old" aircraft might be?

What chance do you think SAC bombers would have against thousands of "old" radar-directed gun sites and "obsolete" air defense missile sites?

Could it be that even something like the old SA-1 could be startlingly effective?

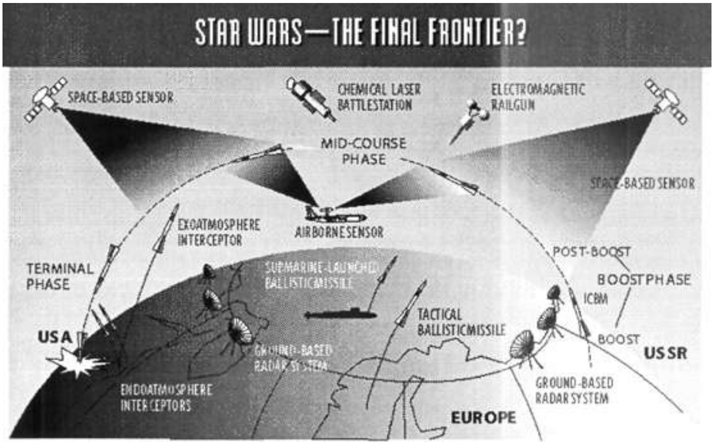
STAR WARS—THE FINAL FRONTIER?
SPACE-BASED SENSOR
CHEMICAL LASER BATTLESTATION
ELECTROMAGNETIC RAILGUN
MID-COURSE PHASE
SPACE-BASED SENSOR
EXOATMOSPHERE INTERCEPTOR
AIRBORNE SENSOR
POST-BOOST
TERMINAL PHASE
SUBMARINE-LAUNCHED BALLISTIC MISSILE
BOOST PHASE
ICBM
TACTICAL BALLISTIC MISSILE
USA
GROUND-BASED RADAR SYSTEM
BOOST
USSR
ENDOATMOSPHERE INTERCEPTORS
GROUND-BASED RADAR SYSTEM
EUROPE

Slide 097: Star Wars—The Final Frontier?

But what of our advanced concepts, such as the strategic defense initiative?

Here we show some of the expensive and complex components of that powerful system.

We have ground-based radars, endoatmospheric interceptors, exoatmospheric interceptors, airborne sensors, space-based sensors, chemically propelled interceptors, electromagnetic railguns, particle beam weapons, high energy lasers, etc. all tied together by a massive command and control system. This massive, expensive system is designed to detect and destroy missiles, including ICBMs, IRBMs, cruise missiles, SLBNs, etc.

The system detects the missile launches and engages them from launch on in to the terminal phase.

Massive systems. Electromagnetic marvels. Nuclear-pumped lasers of staggering power. Perhaps even x-ray and gamma ray lasers. Directed energy RF weapons. Stupendous railguns. Sensors everywhere. Giant webs of communications. Banks of computers and control systems.

And it's totally vulnerable to scalar EM interferometer weapons. Giant scalar EM "radars" can simply sweep through all that like a scythe through standing hay.

As it stands, it's obsolete before its design is even finished.

It's obsolete to Soviet scalar EM weapons that are already deployed and operationally tested in place!

Without scalar EM weapons and defenses, all our present weapons and concepts are deadly vulnerable.

The war is over before it hardly gets started. It's not a war; it's more like a clay pigeon shoot.

Understand, if orthodox science was all the Russians possessed, the Strategic Defense Initiative would be the way to go.

But they do possess something else, and presently they have all the trump cards in the deck.

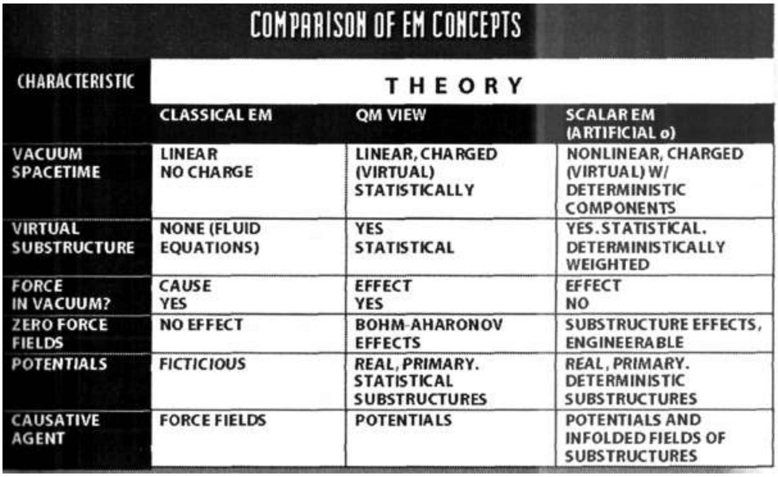

COMPARISON OF EM CONCEPTS

CHARACTERISTIC	THEORY		
	CLASSICAL EM	QM VIEW	SCALAR EM (ARTIFICIAL o)
VACUUM SPACETIME	LINEAR NO CHARGE	LINEAR, CHARGED (VIRTUAL) STATISTICALLY	NONLINEAR, CHARGED (VIRTUAL) W/ DETERMINISTIC COMPONENTS
VIRTUAL SUBSTRUCTURE	NONE (FLUID EQUATIONS)	YES STATISTICAL	YES. STATISTICAL. DETERMINISTICALLY WEIGHTED
FORCE IN VACUUM?	CAUSE YES	EFFECT YES	EFFECT NO
ZERO FORCE FIELDS	NO EFFECT	BOHM-AHARONOV EFFECTS	SUBSTRUCTURE EFFECTS, ENGINEERABLE
POTENTIALS	FICTICIOUS	REAL, PRIMARY. STATISTICAL SUBSTRUCTURES	REAL, PRIMARY. DETERMINISTIC SUBSTRUCTURES
CAUSATIVE AGENT	FORCE FIELDS	POTENTIALS	POTENTIALS AND INFOLDED FIELDS OF SUBSTRUCTURES

Slide 098: Comparison of EM Concepts (Chart)

On this slide and the three that follow, we show a comparison of some of the major characteristics of the three kinds of electromagnetics.

We compare the classical, the quantum mechanical, and the scalar EM (electrogravitational) views of electromagnetics

Notice the differences between the three in regards to vacuum spacetime, virtual substructure, force, whether or not force exists in vacuum, the effect of zero force fields, potentials, and what is to be regarded as causative agents.

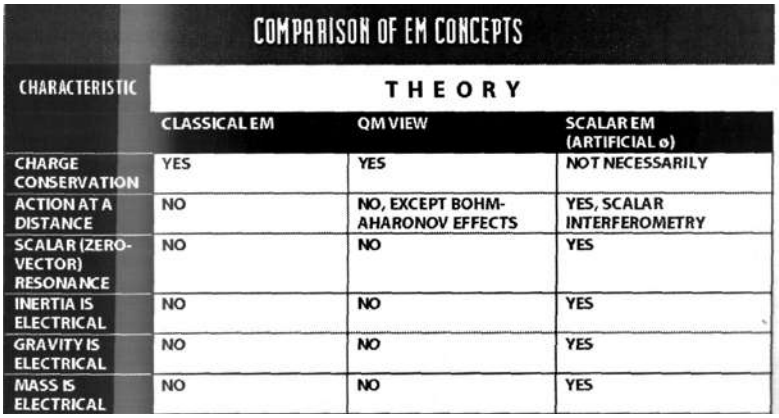

COMPARISON OF EM CONCEPTS

CHARACTERISTIC	THEORY		
	CLASSICAL EM	QM VIEW	SCALAR EM (ARTIFICIAL ø)
CHARGE CONSERVATION	YES	YES	NOT NECESSARILY
ACTION AT A DISTANCE	NO	NO, EXCEPT BOHM-AHARONOV EFFECTS	YES, SCALAR INTERFEROMETRY
SCALAR (ZERO-VECTOR) RESONANCE	NO	NO	YES
INERTIA IS ELECTRICAL	NO	NO	YES
GRAVITY IS ELECTRICAL	NO	NO	YES
MASS IS ELECTRICAL	NO	NO	YES

Slide 099: Comparison of EM Concepts (Chart)

Now note the differing views of charge conservation, action at a distance, scalar resonance, and whether inertia, gravity, and mass are electrical.

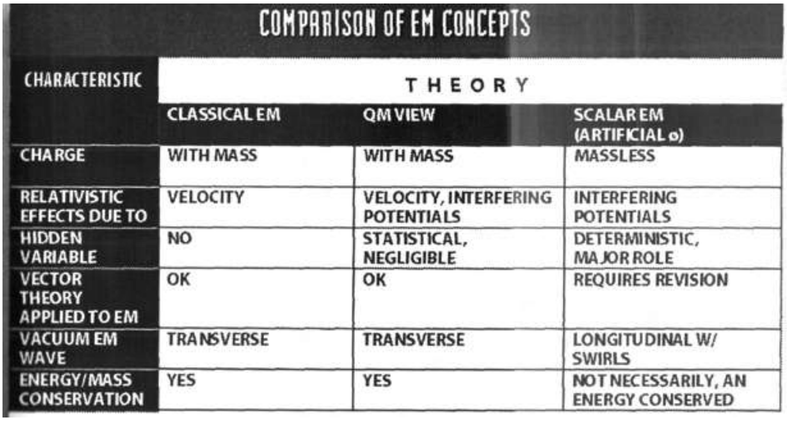

COMPARISON OF EM CONCEPTS

CHARACTERISTIC	THEORY		
	CLASSICAL EM	QM VIEW	SCALAR EM (ARTIFICIAL ø)
CHARGE	WITH MASS	WITH MASS	MASSLESS
RELATIVISTIC EFFECTS DUE TO	VELOCITY	VELOCITY, INTERFERING POTENTIALS	INTERFERING POTENTIALS
HIDDEN VARIABLE	NO	STATISTICAL, NEGLIGIBLE	DETERMINISTIC, MAJOR ROLE
VECTOR THEORY APPLIED TO EM	OK	OK	REQUIRES REVISION
VACUUM EM WAVE	TRANSVERSE	TRANSVERSE	LONGITUDINAL W/ SWIRLS
ENERGY/MASS CONSERVATION	YES	YES	NOT NECESSARILY, AN ENERGY CONSERVED

Slide 100: Comparison of EM Concepts (Chart)

Continuing, note the differences in the views of charge, what relativistic effects are due to, whether or not one has hidden variables and whether they are enginerrable, how vector theory is applied to electromagnetics, the nature of the EM wave in vacuum, and conservation of energy/mass.

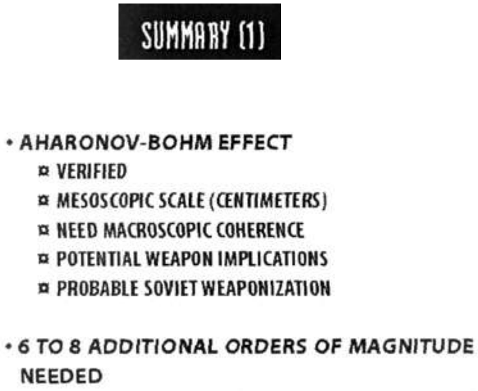

SUMMARY (1)

- AHARONOV-BOHM EFFECT
  - VERIFIED
  - MESOSCOPIC SCALE (CENTIMETERS)
  - NEED MACROSCOPIC COHERENCE
  - POTENTIAL WEAPON IMPLICATIONS
  - PROBABLE SOVIET WEAPONIZATION
- 6 TO 8 ADDITIONAL ORDERS OF MAGNITUDE NEEDED

Slide 101: Summary (1)

Now we examine the three views for whether or not electro-gravitational waves exist and what they interact with, the number of required spacetime dimensions the type of geometry, the nature of a zero-summed EM vector, local and distant relativity aspects, and whether or not the probability states propagated by the Schroedinger equation can be deterministically engineered.

As can be seen from the last four slides, the three views are quite different, and the scalar EM view is drastically different.

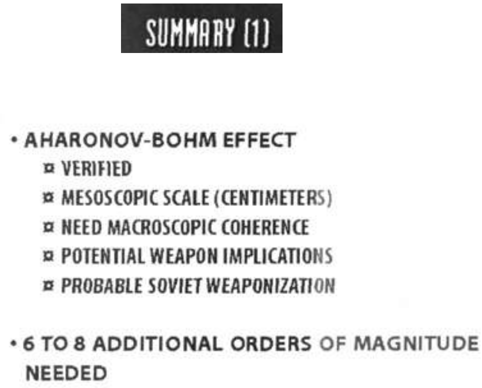

SUMMARY (1)

- AHARONOV-BOHM EFFECT
    - ¤ VERIFIED
    - ¤ MESOSCOPIC SCALE (CENTIMETERS)
    - ¤ NEED MACROSCOPIC COHERENCE
    - ¤ POTENTIAL WEAPON IMPLICATIONS
    - ¤ PROBABLE SOVIET WEAPONIZATION
- 6 TO 8 ADDITIONAL ORDERS OF MAGNITUDE NEEDED

Slide 102: Summary (1)

In summary, we can now state unequivocally that the Aharonov-Bohm effect has been conclusively proven, at least for the mesoscopic scale over a few thousand Angstroms, and possibly over a few centimeters.

But in these experimental proofs; the "natural" type of potential has been used, it is basically incoherent in its substructure, dooming the effect to stay small.

By utilizing macroscopic EM force fields and waves to zero-sum and comprise artificial potentials, the coherence can be maintained over a macroscopic distance - even for hundreds of thousands of kilometers.

The weapons implications - only some of which we have set forth in this treatise - are mindboggling.

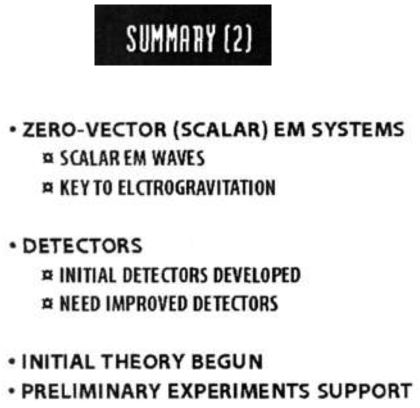

SUMMARY (2)

- ZERO-VECTOR (SCALAR) EM SYSTEMS
    - ¤ SCALAR EM WAVES
    - ¤ KEY TO ELCTROGRAVITATION
- DETECTORS
    - ¤ INITIAL DETECTORS DEVELOPED
    - ¤ NEED IMPROVED DETECTORS
- INITIAL THEORY BEGUN
- PRELIMINARY EXPERIMENTS SUPPORT

Slide 103: Summary (2)

Zero-vector (scalar) EM systems are the key to scalar EM waves, which are actually electrogravitational waves.
Scalar wave detectors nave been developed and are being improved. Further improvement is needed. This development is limited only by funds.

The initial theory has been begun. More work is needed. Again, this development is limited only by funds.

Preliminary experiments support the scalar EM concepts.

Equipment can be designed and made to work electro-gravitationally.

And it has been.

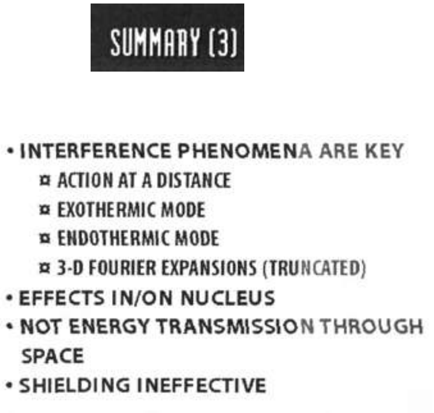

SUMMARY (3)

- INTERFERENCE PHENOMENA ARE KEY
  - ¤ ACTION AT A DISTANCE
  - ¤ EXOTHERMIC MODE
  - ¤ ENDOTHERMIC MODE
  - ¤ 3-D FOURIER EXPANSIONS (TRUNCATED)
- EFFECTS IN/ON NUCLEUS
- NOT ENERGY TRANSMISSION THROUGH SPACE
- SHIELDING INEFFECTIVE

Slide 104: Summary (3)

Interference phenomena are key.

One can get action at a distance - even over hundreds of thousands of kilometers.

One can engineer gravitational and inertial effects.

One can engineer the nucleus, including transmute it, easily and cheaply. One could clean up all the nuclear wastes.

Electromagnetic energy can be produced at a distant target, or extracted from a distant **target.**

This is is not energy transmission through space in the form of EM force fields. Instead, it is transmission through spacetime in the form of electrogravitational potentials.

Conventional EM shielding is ineffective against scalar EM.

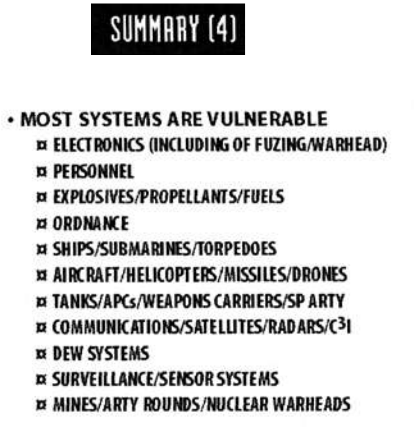

SUMMARY (4)

- MOST SYSTEMS ARE VULNERABLE
    - ¤ ELECTRONICS (INCLUDING OF FUZING/WARHEAD)
    - ¤ PERSONNEL
    - ¤ EXPLOSIVES/PROPELLANTS/FUELS
    - ¤ ORDNANCE
    - ¤ SHIPS/SUBMARINES/TORPEDOES
    - ¤ AIRCRAFT/HELICOPTERS/MISSILES/DRONES
    - ¤ TANKS/APCs/WEAPONS CARRIERS/SP ARTY
    - ¤ COMMUNICATIONS/SATELLITES/RADARS/C^3I
    - ¤ DEW SYSTEMS
    - ¤ SURVEILLANCE/SENSOR SYSTEMS
    - ¤ MINES/ARTY ROUNDS/NUCLEAR WARHEADS

Slide 105: Summary (4)

Almost every weapon system we presently have - or are developing - is totally vulnerable to scalar EM weaponry.

This includes personnel, electronics (including fuzing and warhead), explosives, propellants, fuels, ordnance, ships, submarines, torpedoes, aircraft, helicopters, missiles, drones, rockets, tanks, armored vehicles, weapons carriers, self-propelled and towed artillery, communications, satellites, radars, command and control, directed energy weapon systems, surveillance and sensor systems, mines, artillery rounds and ammunition, nuclear warheads, etc.

We are in dire straits. We need a "Manhattan Project" of the highest priority. Now!

The Soviets have already had the equivalent of seven or eight Manhattan Projects in this area.

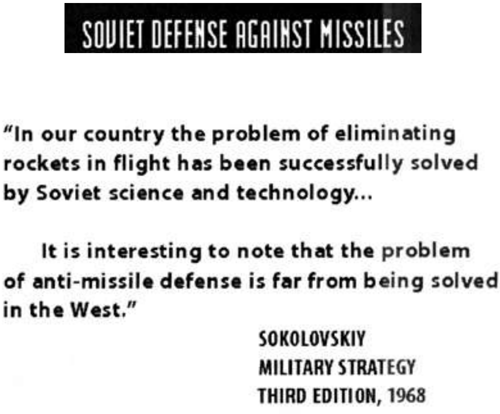

SOVIET DEFENSE AGAINST MISSILES

"In our country the problem of eliminating rockets in flight has been successfully solved by Soviet science and technology...

It is interesting to note that the problem of anti-missile defense is far from being solved in the West."

SOKOLOVSKIY
MILITARY STRATEGY
THIRD EDITION, 1968

Slide 106: Soviet Defense Against Missiles

Here is a significant statement by Sokolovskiy, contained in Military Strategy, 3rd edition, 1968.

This is not propaganda. The Soviets really do have an effective missile defense. We do not. At the present rate we're going, we may not have one in the year 2000 - if we live that long.

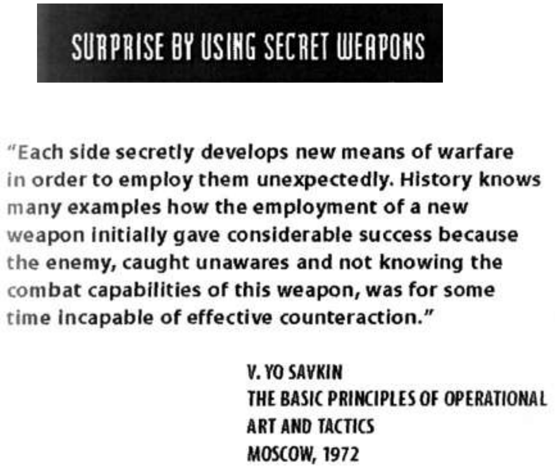

SURPRISE BY USING SECRET WEAPONS

"Each side secretly develops new means of warfare in order to employ them unexpectedly. History knows many examples how the employment of a new weapon initially gave considerable success because the enemy, caught unawares and not knowing the combat capabilities of this weapon, was for some time incapable of effective counteraction."

V. YO SAVKIN
THE BASIC PRINCIPLES OF OPERATIONAL
ART AND TACTICS
MOSCOW, 1972

Slide 107: Surprise by Using Secret Weapons

The Soviets took to heart the lesson of our secret development of the atomic bomb. Thrust suddenly upon a fanatical nation, the atomic bomb brought the empire of Japan to its knees.

Other lesser examples - such as the impact of the German V-l and V-2 missiles and jet fighter - also were not lost on the Soviets.

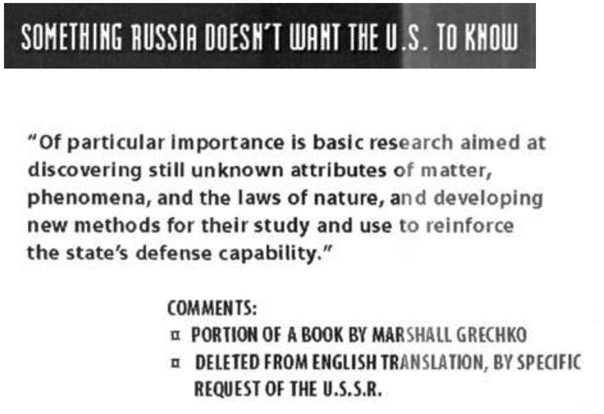

SOMETHING RUSSIA DOESN'T WANT THE U.S. TO KNOW

"Of particular importance is basic research aimed at discovering still unknown attributes of matter, phenomena, and the laws of nature, and developing new methods for their study and use to reinforce the state's defense capability."

COMMENTS:
- PORTION OF A BOOK BY MARSHALL GRECHKO
- DELETED FROM ENGLISH TRANSLATION, BY SPECIFIC REQUEST OF THE U.S.S.R.

Slide 108: Something Russia Doesn't Want the U.S. to Know

But Marshall Grechko said it even plainer.

He made it clear that the Soviet Union has actively searched for great technical breakthrough areas, particularly in discovering new laws of nature, unknown attributes of matter, and new phenomena, and developing their use for defense.

He made it so clear that the Soviet Union formally requested the U.S. State Department to have deleted from its. English translation that portion of his book.

That is a positive signature that the statement is exceedingly sensitive to the Soviets. In other words, it's a real tipoff.

So of course the State Department leaned on the publisher and got that section eliminated.

ADDITIONAL NOTES AND REFERENCES

1. Nikola Tesla, "The Problem of Increasing Human Energy," Century magazine, June 1900, p. 208-209. In this article Tesla spoke of using standing waves to produce distant effects and locate objects such as ships. At this point he was thinking of a scalar radar locator, something more advanced than even our radar today. It was also capable of producing electrical effects at any part of the globe. To quote: "Stationary waves... mean something more than telegraphy without wires to any distance... For instance, by their use we may produce at will, from a sending station, an electrical effect in any particular region of the globe; we may determine the relative position or course of a moving object, such as a vessel at sea, the distance traversed by the same, or its speed..."

2. Nikola Tesla, "The True Wireless," Electrical Experimenter, May 1919, p. 87.

3. T. E. Bearden, Comments on the New Tesla Electromagnetics: Part I: Discrepancies in Present EM Theory, Tesla Book Co., 1580 Magnolia, Millbrae, CA 94030, 1982.

4. T. E. Bearden, Toward a New Electromagnetics: Part III: Clarifying the Vector Concept, Tesla Book Co., 1580 Magnolia, Millbrae, CA 94030, 1983.

5. T. E. Bearden, Toward a New Electromagnetics: Part 4: Vectors and Mechanisms Clarified, Tesla Book Co., Millbrae, California, 1983, p. 6, 39.

6. T. E. Bearden, Solutions to Tesla's Secrets and the Soviet Tesla Weapons, Tesla Book company, 1981

7. T. E. Bearden, Tesla's Secrets and the Soviet Tesla Weapons, 1 hr. 40 min. videotape, Weather Engineering, POB 1472, Huntsville, AL 35807, 1981.

8. T. E. Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, Weather Engineering, POB 1472, Huntsville, AL 35807, 1985.

9. T. E. Bearden, Star Wars Now! The Bohm-Aharonov Effect, Scalar Interferometry, and Soviet Weaponization, Tesla Book Co., 1984

10. Stefan T. Possony, "The Tesla Connection," Defense & Foreign Affairs, Aug. 1984, p. 12-14, 27.

11. Stefan T. Possony, "Psy-War: Soviet device experiment," Defense & Foreign Affairs Daily, 12(104), June 7, 1983, p. 1-2. Reports on Dr. Ross Adey's investigation of the Soviet LIDA device which is used to bombard human brains with radio waves in the 40 MHz region. Also reports on work by A. S. Davydov of the Ukrainian Academy of Sciences, who discovered how the blood-brain barrier can be penetrated by low-frequency radio waves so that brain cells are affected.

12. Fritz Albert Popp, "Photon storage in biological systems," in Electromagnetic Bioinformation: Proceedings of the Symposium, Marburg, Sep. 5, 1977. Edited by Fritz Albert Popp et al, Munchen, Wien, and Baltimore, 1979, p. 123-149. Dr. Popp has discovered, both theoretically and experimentally, that mitogenetic radiation from cells can be seen as a sort of "waste" from a virtual electromagnetic field with a high coherence. This field has a tendency to become stationary over the whole organism. In addition, it includes the storage of "virtual" coherent photons. (In other words, there is a master EM field with high coherence, stationary over the whole body. This is the cells' master control system. Storage of virtual photons -- i.e., scalar signals -- occurs in this field.).

13. To gain an appreciation of the Aharonov-Bohm effect, see Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), Aug. 1, 1959, p. 485-491. For an exhaustive discussion of the Aharonov-Bohm effect and a large number of references, see S. Olariu and I. Iovitzu Popescu, "The quantum effects of electromagnetic fluxes," Reviews of Modern Physics, 57(2), April 1985. See also Bertram Schwarzschild, "Currents in normal-metal rings exhibit Aharonov-Bohm effect," Physics Today, 39 (1), Jan. 1986, p. 17-20.

14. For details of the mysterious explosions of six major Soviet missile ammunition storage sites in seven months, see photograph, Jane's Defence Weekly, 2(3), July 28, 1984, p. 92; see also Mark Daly, "Goa and Goblet SAMs in Severomorsk explosion," Jane's Defence Weekly, 2(6), Aug. 18, 1984, p. 224; Derek Wood, "Soviets' northern fleet disabled... 'not viable' for six months," Jane's Defence Weekly, 2(1), July 14, 1984, p.3; Derek Wood, "Six explosions in the past seven months," Jane's Defence Weekly, 2(1), July 14, 1984, p. 3; John Moore, "The aftermath of Severomorsk," Jane's Defence Weekly, 2(6), Aug. 18, 1984, p. 224.

15. For details of the mysterious, instant deaths in Afghanistan, see Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," Jane's Defence Weekly, 1(13), Apr. 7, 1984, p. 508.

16. For details of the Enhanced-V cloud signature of severe weather, see Gary Ellrod, "Dramatic examples of thunderstorm top warming related to downbursts," National Weather Digest, 10(2), May 1985, p. 7-13; T. J. Fujita, "Manual or Downburst Identification for Project Nimrod," SMRP Research Paper No.156, University of Chicago, 1978; D. W.McCann, ., The Enhanced-V: A Satellite Observable Severe Storm Signature," Monthly Weather Review, 111, 1978, p. 887-894.

17. For a good discussion of the modern view of the vacuum, see
I.J.R. Aitchison, "Nothing's plenty: The vacuum in modern quantum field theory," Contemporary Physics, 26(4), 1985, p. 333-391.

18. For a discussion of the ring interferometer experiments related to quantum and gravitational effects, see G.E. Stedman, "Ring interferometric tests of classical and quantum gravity," Contemporary Physics, 26(4), 1985, p. 311-332.

19. For information on self-induced transparency with pulsated light, see E. L. Hahn, Scientific American, June 1967. This is a scalar EM effect. The phenomenon is observed only in insulators, not in conducting materials. A dielectric tends to stop the flow of electrons, thus stopping the bleed-off of the Kaluza-Klein 5-potential as EM force fields. Thus through the insulator or dielectric, an electrogravitational pulse is transmitted. When this pulse emerges from the other side of the

dielectric, then electrons are free to move again and form ordinary EM fields, resulting in resumed EM bleed-off of the KK (Kaluza-Klein) 5-potential. Thus ordinary photons strike the dielectric, changing into electrogravitational waves which transit the opaque dielectric, and these EG waves transform back to EM waves on the far side of the dielectric. Advanced scalar EM application of this effect can make an object optically invisible.

20. For a lucid and profound presentation of the entire subject of elementary particle physics and quantum field theory, see T. D. Lee, Particle physics and introduction to field theory, Harwood Academic Publishers, New York, 1967. Subjects covered include those at the forefront of current research. Chapter 16, "Vacuum as the source of asymmetry," is particularly important. Here Lee concisely covers the idea of degenerate vacuum states, vacuum structuring, and vacuum effects causing symmetry breaking (violation of conservation laws). An excellent coverage of asymmetries and observables is given on p.181-188. For example, whenever symmetry is spontaneously broken, a nonobservable (virtual entity) has become observable. Thus there exist mechanisms for the direct translation between virtual and observable states.

Most engineers are unaware that physics has already rigorously proven (both theoretically and experimentally} that conservation of energy can be physically violated, even by excitation of the vacuum.

However, in the present assumption of a randomized virtual microstructure of a vacuum potential, these broken symmetry effects are individually confined to very small microlevels. For example, on p. 383, Lee points out that the microscopic structure of the scalar vacuum field (i.e., of vacuum charge or "potential") is not utilized.

This of course is reasonable if all vacuum potentials used in particle physics are treated as having a randomized microstructure of virtual particle flux vectors. Since any appreciable directional coherence in such a randomized microstructure could persist only over extremely tiny distances, then the microstructure of vacuum degenerate states has negligible local effects in ordinary macroscopic systems.

In other words, these conservation violation effects would be very difficult to coherently capture and utilize to provide "free energy, emerging from the vacuum."

The randomized momentary coherences, however, are responsible for the violent microfluctuations of the EM fields, well-known in classical electrodynamics. An approach such as taken by Yater or Moray would be necessary to achieve a vacuum energy tap under such conditions. To achieve usable macroeffects, a very large collection of charge or magnetic material, etc. would be required before the probability of usable coherences would become sufficiently large. Therefore this area is neglected in practical physics and engineering.

However, this is precisely where scalar EM is lost in conventional theory and engineering practice. In scalar EM, the coupling together of macroscopic EM force field vectors into a summed zero system means that the vacuum potential's microstructure is now macroscopically ordered by that system. Further, it is sustained during translation through space. In such case, one has produced directional gravitational potential gradients that have macroscopic range and form. Further, by coherently varying the amplitude of the individual components of the zero vector system, a gravitational wave -- a controlled oscillation of the degenerate states or vacuum -- is produced. By processing the EM components, this gravitational wave can be narrowed, formed into a beam, and projected. Two such beams can be interfered, even at a great distance, producing G-interference and causing EM effects to emerge in the distant interference zone. Thus scalar EM profoundly affects quantum mechanics, quantum field theory, particle physics, general relativity, and electromagnetics.

21. For some strong and pertinent critique of present scientific bias, see Ruggero Maria Santilli, Ethical probe on Einstein's followers in the U.S.A.: An insider's view, Alpha Publishing, POB 82, Newtonvllle, MA 02160. Santilli is a highly published physicist and the editor of a refereed scientific journal, the Hadronic Journal. After lengthy attempts to correct injustices within the Scientific system, to no avail, he finally has gone public with his specific charges. Santilli references senior Russian physicist G. Yu. Bogoslovsky, whose 1984 paper "Generalization of Einstein's relativity theory for the anisotropic spacetime," whose research would have been impossible in the U.S. Quoting Santilli: "He would have been discredited, humiliated, and chased out of these institutions [Harvard, M.I.T., and Yale] beyond any reasonable doubt by ample documentation on similar cases." Again to quote Santilli: "Papers bearing the title of 'Generalization of Einstein's relativity' are routinely permitted with human dignity at Moscow University in the U.S.S.R., but they are absolutely inconceivable today at the departments of physics of Harvard University, of the Massachusetts Institute of Technology, of Yale University and of other similar institutions in the U.S.A., as numerous senior scientists would readily testify."

You see, dear reader, there has been a deep scientific conspiracy in the U.S. to prevent any change in the severe restrictions placed on general relativity by Einstein. This has reduced GR to a sort of "special relativity, with only distant perturbations." Specifically, it has excluded scalar electromagnetics and branded as heresy the inconceivable idea that general relativity could be local, and engineerable.

Now you know why the Soviets have been able to develop these monstrous weapons for over three decades while our scientists slept at the wheel. Scientific arrogance and pride has been our greatest enemy.

See also Bryan G. Wallace, "The great speed of light in space coverup," Scientific Ethics, 1(1), Feb. 1985, p. 2-3; "Radar testing of the relative velocity of light in space," Spectrosc. Lett. 2, 1969, p. 361; Wallace, "Letter to the editor," Physics Today, 36(1), 1983, p. 11; "Letter to the editor," Physics Today, 37(6), 1984, p. 15. See M.E. Ash et al, Astron. J., 72, 1967, p. 72. Direct experiments on Einstein's second postulate, in the 1961 interplanetary radar contact with Venus, challenge whether relativity is correct. Variations of over 30,000% of the best possible general relativity fit the MIT Lincoln Lab could generate were measured. The variations were not random but contained diurnal, lunar, and synodic periodic components. See also Bryan G. Wallace, Spectrosc. Lett. 4, 1971, p. 79 for an analysis of published 1961 data on the interplanetary radar contact with Venus that concludes that the data showed a component that was relativistic in the Galilean c+v sense. See also Wallace, "The Unified Quantum electrodynamic ether," Foundations of Physics, 3, 1973, p. 381. which details Measurements strongly challenging the Einstein assumption of the constant velocity of light, and deals with the nature of the ether. See also Thomas E. Phipps, Jr., "Critique of accepted theories of special relativity and electromagnetism," Scientific Ethics, 1(2), Apr. 1985, p. 17-23.

22. For a sample of Soviet scientific freedom from narrow-minded bias in general relativity, see V. I. Denisov and A. A. Logunov, "The inertial mass defined in the general theory of relativity has no physical meaning," Teor. i Matemat. Fizika, 51(2), May 1982, p. 163-170 (In Russian). See also A. A. Vlassov and V. I. Denisov, "Einstein's formula for gravitational radiation is not a consequence of the general theory of relativity," Teor. i. Matemat. Fizika, 53(3), Dec. 1982, p. 406-418. Here we find that the intensity of gravitational radiation and the total intensity determined in the general theory of relativity in accordance with Einstein's formulas depend on the choice of the coordinate system, so that they can be made to vanish or even become negative. [Note by T.E.B.: If local curvature is

engineerable, as in scalar EM, the local choice of coordinate systems can itself be directly engineered. This clearly implies that scalar EM can thus produce either "positive" intensity radiation or "negative" intensity radiation; thus it can "produce energy at a distance" or "extract energy from a distance.]
Continuing with Denisov: This result is a particular consequence of the general assertion to the effect that in general relativity there are no energy-momentum conservation laws for a system consisting of matter and the gravitational field. [Note by T.E.B.: Thus we can have an energy source or energy sink appear in the distant interference region, where matter has been lifted from the Dirac sea and exposed to the highly nonlinear interfering EG fields.]. See also V. I. Denisov and A. A. Logunov, "New theory of spacetime and gravitation, "Teor. and Matemat. Fizika, 50(1), July 1982, p. 3-76. Points out that ". . . the gravitational field in general relativity is completely different from other physical fields and is not a field in the spirit of Faraday and Maxwell." (p. 3, English version).

23. In the Proceedings of the Tesla Centennial Symposium, an IEEE Centennial Activity, Colorado College, Colorado Springs, Colorado, Aug. 9-11, 1984, see Richard A. Blade, "Modification of Maxwell's equations in free space to account for nonzero photon mass and scalar electromagnetic waves," p. 91-92; Jack Y. Dea, "Scalar fields: their prediction from classical electromagnetism and interpretation from quantum mechanics," p. 94-98; Robert W. Bass, "Self-sustained non-Hertzian longitudinal wave oscillations as rigorous solutions of Maxwell's equations of electromagnetic radiation," p. 89-90; T. E. Bearden, "Tesla's electromagnetics and its Soviet weaponization," p. 119-138.

24. To see the limiting assumptions imposed on general relativity that results in the severely restricted "ordinary general relativity" taught in almost all Western universities. see Charles W. Misner, Kip S. Thorne and John Archibald Wheeler, Gravitation, W. H. Freeman and Co., San Francisco, CA, 1973, p. 19-21, 71-72, 367-369. Among other things, these arbitrary restrictions save the conservation laws. When they are removed. the conservation laws can be violated.

25. Several anomalous events -- skyquakes, booms, or earthquakes -- have been documented in Europe. See H. W. Haak. "The Seismic event of November 9, 1983: Earthquake or sonic boom." Seismic Activity in Western Europe, (With particular consideration of the Liege Earthquake of November 8, 1983), Ed. by Paul Melchior, D. Reidel Publishing Co., 1985, p. 215-222. For such events in the United States, see S. Claflin-Chalton and G. J. MacDonald, "Sound and light phenomena," The Mitre Corporation, McLean, Virginia 22102, Nov. 1978.

26. Fundamental experiments were performed by Eotvos et al. in the early 1920s. See R.V. Eotvos, D. Pekar and E. Fekete, Ann. Phys. (Leipzig) 68, 1922, p. 11. Actually a fifth force of nature due to hypercharge may have been found by these experiments. See Ephraim Fischbach et al. "Reanalysis of the Eotvos experiment," Physical Review Letters, 56(1), Jan. 6, 1986, p. 3-6. See also F. D. Stacey and G. J. Tuck, Nature 292, 230 (1981); S. C. Holding and G. J. Tuck, Nature 307, 714 (1984); G. W. Gibbons and B. F. Whiting, Nature 291, 636 (1981); F. D. Stacey, in Science Underground, edited by M. M. Nieto et al., AIP Conference proceedings No. 96, (American Institute of Physics, New York, 1983, p. 285. See also S. H. Aronson et al, Physical Review Letters 48, 1982, p.1306; Physical Review D 28, 1983, p. 476, 495. Also, comparison of satellite and terrestrial determinations of the gravitational field constant reveals significant disagreement: see R. H. Rapp, Geophys. Res. Lett. 7, 1974, p. 35; Bull. Geod. 51, 1977, p. 301.

27. John J. A. O'Neill. "Tesla cosmic ray motor may transmit power 'round the earth: Famed scientist, on eve of 76th birthday, says he has succeeded in harnessing 'penetrating rays' to operate small motive device." Brooklyn Eagle, July 10,1932. Tesla reported he had harnessed "cosmic rays" and caused them to operate a small motive device. Tesla's use of "cosmic rays," however, encompassed scalar EM radiation. Tesla later seems to have progressed to a simpler motive unit. In the Kansas City Journal-Post, Sep. 10, 1933, he stated: My power generator will be of the simplest kind -- just a big mass of steel, copper and aluminum, comprising a stationary and rotating part, peculiarly assembled." A similar type of "gravitational turbine" seems to have been invented by John Bedini.

28. To see just how arbitrary and postulational are present "definitions" of mass and force, see Robert Bruce Lindsay and Henry Margenau, Foundations of Physics, Dover Publications, New York, 1963, p. 283-287. Note on p. 283 that a "field of force" at any point is actually defined only for the case when a unit mass is present at that point. Also see Richard P. Feynman, Robert B. Leighton, and Matthew Sands, The Feynman Lectures on Physics, Addison-Wesley, New York, Vol. 1, 1963, p. 2-4 for a definition of the electric field in the context of its potentiality for producing a force. This problem did not exist for the original foundation of electromagnetic force fields in vacuum, since the vacuum was assumed to be a material ether. With the modern view a nonmaterial ether has emerged, yet the classical electromagnetic theory has not been made sufficiently rigorous in distinguishing "observable" and "virtual" electromagnetic force fields.

29. E. P. Battey-Pratt and T. J. Racey, "Geometric Model for Fundamental Particles," International Journal of Physics, Vol. 19, No. 6, 1980, p. 437-475.

30. Some years ago, Dr. William Tiller produced very important "free A-field" theoretical work which showed waves which did not fall off as $1/r^2$. Interesting theoretical work in scalar wave theory has also been advanced by Dea. See Jack Dea, "Fundamental Fields and Phase Information," presented at the 1985 U.S. Psychotronics Association Conference, Bergamo Center, Dayton, Ohio, July 1985. Dea has formulated an important theory for vacuum EM scalar waves, derived from Maxwell's equations with zero E and B fields. He predicts waves whose intensity falls off with distance as $1/r$ instead of $1/r^2$, and detection of these waves by various means. He shows simple generators for the waves. Detection is quantum mechanically by electron interference, not by single electrons. Thus such scalar waves will pass through normal Faraday shields. In his theory Dea does not address the question of spin vortexes in vacuum scalar waves. It is the present author's view that all vacuum waves are scalar waves, and the primary distinction between the "transverse" EM wave and the longitudinal scalar wave is simply the presence or absence of spin-vortexes. Indeed, one may argue that the present "vector wave" EM theory can be replaced by a scalar wave theory. In 1903 Whittaker proved that any vector field can be replaced by two scalar fields. See E. T. Whittaker, Proc. Lond. Math. Soc.1, 367, 1903. By implication, since a vector wave is simply a moving vector field, it can be replaced by two moving scalar fields.

31. David Jones, "Israel's Secret Weapon," Weekend Magazine, Dec.17, 1977.

32. John Hutchinson, private communication. Masses of wood, metal, etc. -- some weighing over 60 lbs. -- have been levitated by Hutchinson.

33. For a description and details of the April 9, 1984 incident off Japan, see Daniel A. Walker, Charles S. McCreery, and Fermin J. Oliveira, "Kaitoku Seamount and the Mystery Cloud of 9 April

1984," Science, 227(4684), Feb. 8, 1985, p. 607-611. See also transcription from Japan Airlines flight 036 to Anchorage International Flight Service, 9 April 1984, at 1406 Greenwich mean time, Federal Aviation Administration (FAA) — Alaska Region; Jim Derry, Special Agent, FAA-Alaska Region, interviews with Captain C.H. McDade, Japan Airlines Flight 036, with Captain Cornelis Vander Berg, KLM Royal Dutch Airlines Flight 868; with First Officer Reint Sol, KLM Royal Dutch Airlines Flight 868, and with Captain Mike Howe, Flying Tiger Lines Flight 078, Anchorage Daily News, 10 April 1984, p. A-l; Anchorage Daily News, 12 April 1984, p. A-l; Honolulu Advertiser, 10 April 1984, p. B-l; Newsweek, 30 April 1984, p. 25, Gregg Ripee, "Mushroom cloud sighted off Japan," Los Angeles Daily News, 11 April 1984, p. 1,8; Ross Gelbspan, "Mystery cloud baffles scientists," Boston Sunday Globe, 13 Jan. 1985, p. 1, 14; T.E. Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985. See also McKenna, Daniel L. & Daniel Walker,"Mystery Cloud: Additional Observations", Science, 234(4775), Oct. 24, 1986, p. 412-413.

34. For a description of the April 11, 1963 anomalous underwater explosion north of Puerto Rico, see Robert J. Durant, "An Underwater Explosion — or What?", Pursuit, 5(2), April 1972, p. 30-31.

35. For details of the loss of the U.S.S. Thresher and the interference experienced by the U.S.S. Skylark, see John Bentley, The Thresher Disaster, Doubleday, Garden City, N. Y., 1975, particularly p. 164. Some of the communications systems of the Skylark experienced temporary failure, but were not disabled, later resuming operation.

36. The Soviet Union has a well-established history of interfering with and destroying aircraft; for example, for a listing see "Soviet and Warsaw Pact air incidents," Jane's Defence Weekly, 3(2), Jan. 12, 1985, p. 58.

37. For details of the anomalous "nuclear flashes" detected by U.S. Vela satellites, see "Scientists Fail to Solve Vela mystery," Science, 207, 1 Feb. 1980, p. 504-506. "Satellite Evidence Shows 'possibility' of Nuclear Test, DOD Says," Aerospace Daily, Oct. 29, 1979, p. 286. "A Flash of Light," Newsweek, Nov. 5, 1979, p. 64-65. "Was It a Nuclear Device?", Newsweek, July 21, 1980, p. 19. "Diverging Views," Washington Roundup, Aviation Week & Space Technology, July 21, 1980, p. 15. Philip J. Klass, "Clandestine Nuclear Test Doubted," Aviation Week & Space Technology, Aug. 11, 1980, p. 67, 69, 71-72. "Debate continues on the Bomb That Wasn't," Science, 209, 1 Aug. 1980, p. 572-573. "Navy Lab Concludes the Vela Saw a Bomb," Science, 209, 29 Aug. 1980, p. 996-997. See also Jack Anderson and Dale Van Atta, United Press Syndicate, "CIA knew South African-Israeli nuclear test was coming," Birmingham Post-Herald, Apr. 26, 1985, p. A5. The latter article repeats the CIA/DIA argument that the Vela incidents represent concurrent South African-Israeli testing of atomic weapons. Taiwan is considered a possible third partner. According to Anderson and Van Atta, a South African navy contingent was "in the area" at the time of the 1979 Vela detection, and Israel and South Africa were cooperating on an atomic bomb project as early as 1966. Apparently U.S. announcement of an impending South African atomic test in the Kalahari Desert in 1977 led to such an outcry that the test was suspended. Anderson and Van Atta cite succeeding visits to South Africa by Israeli scientists, technicians, and defense officials. However, since the actual Vela detections showed some deviations from what would be expected from nuclear tests, one can speculate that Anderson and Van Atta have presented evidence for the development of scalar electromagnetic weapons by two additional countries. One may further speculate that this could reveal what has been checking the Soviets from simply moving against the West with scalar electromagnetic weapons.

Indeed, the mysterious destruction of six major Soviet missile ammunition storage areas within the space of seven months could have been a preemptive measure and warning to the Soviets if they had actually been considering moving against the West. Also, other incidents reported in the South Atlantic may possibly be associated with the program producing the "Vela flashes." For example, see N.W.C. Rutherford, "Unidentified Phenomena," Marine Observer, 51, 1981, p. 186 for a description of a great bluish-white flash and small bolt of lightning, indicating some sort of anomalous electrical discharge, on Dec. 23, 1980 at 2245 hours (10:45 p.m.). Many other anomalous flashes for which no known causative mechanism exists have been detected by satellites. For example, the OSO-2 and OSO-5 satellites have detected about one mysterious short (less than 0.1 sec) flash about every ten orbits, at times when the fields of view of the telescopes were well above the earth. As many as three of the satellite's photometers have detected a flash simultaneously. These flashes must be removed from any discussion of satellite-observed lightning; see J. G. Sparrow and E. P. Ney, "Lightning Observations by Satellite," Nature, 232, 1971, p. 540.

38. For a description of sightings of enormous glowing globes of light deep within the Soviet Union in the direction of Saryshagan, see Gwynne Roberts, "Witness to a Super Weapon?", The Sunday Times, London, England, Aug. 17, 1980. Several incidents were observed by Downie and others.

39. For an artist's conception of a directed energy weapon installation at Saryshagan, see Aviation Week & Space Technology, July 28, 1980, p. 48. For a DOD sketch, see Soviet Military Power, 1985. Obviously classically-trained U.S. military analysts, with no knowledge of scalar interferometry, have not considered the installation as a candidate hyperspatial howitzer.

40. For a description of multiple pilot sighting (from Iran) of an enormous glowing globe of light deep within the Soviet Union, see Central Intelligence Agency (CIA) Information Report No. CO-B-321/15354-66, "Sighting of Unusual Phenomenon on Horizon Near Iranian/USSR Border," Sep. 8, 1966, released on 15 Dec. 1978 under the Freedom of Information Act (FOIA). The sighting occurred at 2142 hours (9:42 p.m.).

41. For details of sighting of the intensely glowing ball of light near British European Airways flight 831 between Moscow and London, see CIA Foreign Intelligence Information Report CO-B-324/33601-76, "Aerial observation of intense Source of Light," Nov. 18, 1976, released on 15 Dec. 1978 under the FOIA. The incident occurred between 1800 and 1900 hours (6:00-7:00 p.m.). The light was observed for 10-15 minutes, until the aircraft had flown past and left the light behind.

42. For other examples of pertinent 1ight phenomena: F. S. Angus and G. Carling, "Optical Phenomenon: Caribbean Sea; Western North Atlantic," Marine Observer, 40, Jan. 1970, p. 17-18; R. A. Holmes, "Unidentified Phenomenon, off Barbados, West Indies," Marine Observer, 40(229), July 1970, p. 107-108; H. K. Dyer, "Unidentified Phenomenon: Western North Atlantic," Marine Observer, 43, July 1973, p.114. For a 1961 double concentric hemispheric shape observed in the Indian Ocean, see Marine Observer, 32, 1962, p. 64. For a particularly significant combination of forms, see Marine Observer, "Unidentified Phenomenon," 48, 1978, p. 21-22. Also, on the night of July 26, 1984 the pilot and crew of a Boeing 747 (American carrier) flying from Tokyo to Fairbanks, near the Kuriel Islands, at 47 5'44" N: 161 00'05" E, noticed a slowly expanding hemisphere of white light off to their left above the horizon. The shell of light continued to expand over a 10-minute

period until ahead of them and to the right. The crew braced for a shock wave which never arrived. Their weather radar saw nothing out of the ordinary. The shell of light had sharp edges and was semi-transparent so that stars became visible through it. (Richard F. Haines, private communication, Sept. 27, 1984). See also Richard Hall, "Aerial anomalies at sea," The Info Journal, 4(3), May 1975, p. 6-9 for an unusual light phenomenon seen in the Caribbean by the R.M.S. Carmania (British). The phenomenon occurred four times at precise hourly intervals, strongly suggesting an artificial origin. See also F. Shepherd, "Cloud: Strait of Gibraltar," Marine Observer, 51(273), July 1981, p. 107-108 for an incident of associated cloud, arch, and anomalous pressure effects. For another significant airline sighting by Japan Airlines flights 403 and 421 of a giant, expanding globe of light on June 18, 1982, in the North Pacific, 100 km. east of Kushiro, see the Asahi Evening News, Tokyo, Japan, June 22, 1982. For additional evidence that nomadic tribesmen passing through Kazakhstan repeatedly observe such gigantic light phenomena in close association with Saryshagan missile test range, and that Soviet deception planners foster the notion of "UFO activity" as a Red Herring, see Robert Jackson, "KGB takes to the flying saucers!", Northern Echo, Darlington, England, March 15, 1984.

43. For details of Khrushchev's announcement of a Soviet superweapon, see Max Frankel, "Khrushchev Says Soviet Will Cut Forces a Third; Sees 'Fantastic Weapon'", New York Times, Jan. 15, 1960, p. 1.

44. See Christopher S. Wiren, "Brezhnev Calls for Accord against 'Terrifying Arms,'" New York Times, June 14, 1975, p. 1, 11. In a major Kremlin speech on June 13, 1975, Leonid I. Brezhnev included a statement that one major concern not covered by existing agreements was "assuming ever greater urgency with each passing day." Calling for a ban on new weapons of mass destruction, the Soviet leader stated: "The level of modern technology is such that a serious danger arises of creating weapons more terrifying than nuclear ones." He stated, "The reason and conscience of humanity dictate the necessity of erecting an insurmountable barrier to the development of such weapons." See also Malcolm W. Browne, "Senatorial Group Received by Brezhnev," New York Times, July 3, 1975, p. 2. Leonid Brezhnev repeated his proposal to a group of U.S. senators on July 2, stating that the capability to develop more terrible weapons was obvious. He emphasized, however, that he "had no particular weapon in mind." Here he may have been deliberately misleading the U.S. senators, to prevent revealing that the Soviet Union had already developed such weapons, and was horrified at their potential for total destruction. E.g., if massive and rapid scalar explosions occur, the linearity of spacetime -- and the orderly flow of time itself -- may be seriously perturbed throughout the earth and its adjacent space. From a general relativistic viewpoint, such an event might well prove catastrophic for the entire biosphere. Boris N. Ponomarev, a Soviet national party secretary, again raised the same issue to a delegation of visiting U.S. congressmen in the Kremlin on Aug. 11, 1975; see Christopher S. Wren, "Moscow Now Pressing Disarmament." New York Times, Aug. 12, 1975, p. 6. At the United Nation's thirtieth session of the General Assembly on Sept. 23, 1975, Foreign minister Andrei A. Gromyko strongly raised the same issue, warning that science can produce "ominous" new weapons of mass destruction. He urged that all countries, led first by the major powers, should sign an agreement to ban the development of these unspecified new weapons. He even offered a draft, entitled "Prohibition of the Development and Manufacture of New Types of Weapons of Mass Annihilation and of New Systems of Such Weapons." The first article provided that the types of these new weapons would be "specified through negotiations on the subject." Because of its fixation on nuclear weapons, the West may have lost its only opportunity to prevent the spread of scalar electromagnetic weapons of power unprecedented even by nuclear arms. Of course, it may also have

been a Soviet ploy to prohibit the U.S. from developing scalar electromagnetics defensive weaponry to counter the massive strategic scalar electromagnetics weapons already developed, tested, and deployed by the Soviet Union. Certainly in 1973 at a secret meeting in Prague with European communist leaders, Brezhnev had laid out a firm statement of Soviet domination of the world by 1985. See "Secret Speech: Did Brezhnev Come Clean?", National Review, 29(8), Mar. 4, 1977, p. 248, 250. Brezhnev's speech contained statements such as: "We are achieving with detente what our predecessors have been unable to achieve using the mailed fist...come 1985 we will be able to extend our will wherever we need to." Brezhnev stated the Soviet Union would control the oceans of the world, 90% of the land, and the air and space above. A front-page Boston Globe article on Feb. 11, 1975 was the first to carry the story. A copy of Brezhnev's secret speech was obtained by British intelligence in 1973, but denigrated by Secretary of State Kissinger. Apparently the first official U.S. reference to Brezhnev's startling speech did not appear until the 1976 National Intelligence Estimate. Although the big intelligence bureaucracies can do a fantastic job with "current intelligence," their record at predicting a drastic change is woeful, precisely because of the political nature of bureaucracies. See "What is Intelligence," Colonel William V. Kennedy et all, Intelligence Warfare: Penetrating the Secret World of Today's Advanced Technology Conflict, with an introduction by Dr. Ray S. Cline, Crescent Books, New York, 1983, p. 8-23. Intelligence Warfare is heartily recommended as the single book which best gives the reader a grasp of modern technical intelligence organizations, capabilities, and operations.

45. For details of Soviet weather engineering operations over the U.S., see Thomas E. Bearden, "USSR: New beam energy possible?", Defense & Foreign Affairs Daily, 13(111), June 12, 1984, p. 1-2. Cloud signature patterns reported in the article have been repeatedly sighted over multiple areas in the U.S. A series of anomalous "hole in the cloud" patterns, possibly associated with early Soviet weather engineering efforts over the U.S., was widely seen in 1967-68. See Weatherwise, 21(4), Aug. 1968, cover and p. 143; 21(5), Oct. 1968, p. 194-195, 204-205; 21(6), Dec. 1968, p. 238-245. Peculiar stationary blocking patterns have also been observed to occur and influence weather over the U.S., leading to severities such as the winter of 1976-77. See K. K. Tung and R. S. Lindzen, "A theory of stationary long waves, Part 1, A simple theory of blocking; Part 2: Resonant Rossby waves in the presence of realistic vertical shears; Part 3, Quasi-normal modes in a singular waveguide," Monthly Weather Review, 107(6), June 1979, p. 714-774. See also T. E. Bearden, Soviet Weather Engineering Over North America, videotape, 1985 for an expose or the physics and some major incidents of major Soviet weather engineering, including details and photographs of cloud signature patterns. See also C. B. Baker, "Soviet Weather Mayhem," Youth Action News, Nov. 1983, P.O.B. 312, Alexandria, Virginia 22313, p. 1-8; C. B. Baker, "Diabolic Soviet Warfare," Youth Action News, Nov. 1984; For a conventional Soviet look at weather modification (possibly intended to mislead the West), see E. T. Protasevich, "Microwave weather modification," Soviet Technical Physics Letters, 7(3), March 1981, p. 114-115. For a downburst mechanism, see Gary Ellrod, "Dramatic examples of thunderstorm top warming related to downbursts," National Weather Digest, 10(2) May 1985, p. 7-13. See Art and Kay Westermann, Weather War — Worldwide, Liberty Bell Books, POB 288, Sedona, Arizona 86336, 1983. See also U.S. Senate Resolution on Environmental warfare, July 11, 1973.

46. For details of the Soviet Woodpecker beams bending around the earth and intersecting over the U. S., see Christopher Chant and Ian Hogg, Nuclear War in the 1980's?, Harper & Row, New York, 1983, p. 71. See also Soviet Military Power, U.S. Department of Defense, 1985, p. 45. The

intersection forms an interference grid. Scalar EM channels added to the Woodpecker carriers then provide a corresponding sca1ar EM interference grid in the scalar area.

When adjusting the scalar interferometry components of the grid, bursts in the interference and discharges of energy as well as bursts of energy extractions occur. This leads to the production of anomalous aerial phenomena such as explosions, rumblings, airquakes, etc. In addition, direct Bohm-Aharonov forces may be engendered in ground structures, shaking buildings and windows, etc. without seismic disturbances. Such anomalous disturbances have been widely noted in the open U.S. press. E.g., particularly see Kenneth F. Bunting and Daniel M. Weintraub, "Mysterious Rumbles Jar Southland," Los Angeles Times, Dec. 8, 1984, CC(II), p. 1,10; Let Scarr, "People shook up over mystery shakes," The San Diego Union, Dec. 8, 1984, p. Bl, 4. Following this and a Dec. series of anomalous aerial disturbances off the Los Angeles/San Diego coast, a sharp and unexpected weather front appeared, racking the area with high winds on the night of Dec. 12, producing widespread damage and power outages. The newly-formed front moved on to produce significant snowstorms in other areas. See Ted Thackery, Jr., "Winds Kill Two, Wreak Havoc as They Sweep Through the Southland," Los Angeles Times, Dec. 14, 1984, MF/Part II, p. 1,8; "Even Tucson gets snow in storm that dumps 20 inches," AP release, Los Angeles Herald Examiner, Dec. 14, 1984, p. A-12. During several months prior to the December 1984 Los Angeles incident, anomalous electrical failures and outages had occurred in the San Diego and Southern California area, possibly due to scalar interferometry adjustments in the Woodpecker grid over the area. For other typical incidents possibly associated with adjustment of the interference grid, see "Who boomed? Odd Noise unnerves Ohioans," Journal, Milwaukee, Wisconsin, Dec. 28, 1982; "Tremors shake Miss. Gulf Coast," Times-Picayune, New Orleans, Louisiana, Feb. 27, 1983; Sandra Skowron, "Sonic boom is just Mother Nature," Beaver Co. Times, Beaver Falls, Penn., Jan 13, 1983; "Booms in Rockland, New York Times, Mar. 1. 1979, Sec. 8, p. 82; "Loud 'boom' unexplained," Star-Beacon, Ashtabula, Ohio, Feb. 21, 1983; Paul Senecton, "The word is now IDENTIFIED," Mail, Hartlepool, England, Feb. 3, 1985: On Jan. 23, 1974 a mysterious explosion over North Wales rocked a 60-mile radius area. Associated anomalous light phenomena were also seen. Hundreds and hundreds of anomalous booms, aerial rumbles, and aerial explosions, many accompanied by flashes or anomalous shaking of ground structures, buildings, windows, etc., have occurred over the U.S. and other Western nations in the 60's, 70's, and 80's.

On occasion, an aircraft at the proper altitude and position can conceivably experience interference in its electrical systems and in its engine combustion processes due to exothermic kindling or endothermic extraction of electromagnetic energy. The combustion processes of aircraft engines produce highly ionized gases in very nonlinear conditions. These can act as strong dephasers and decouplers of the vector EM components of a vector zero summation in a scalar interference zone. Exothermic interference would have little effect on combustion, but strong endothermic interference can result in engine flameout. Anomalous simultaneous failure of jet and propeller-driven aircraft engines sometimes occurs. For high altitude grids, this can result in engine flameouts, but restart and recovery after unusual loss of altitude has dropped the aircraft out of the scalar interference zone. Several candidate incidents over and near the U.S. exist, particularly where power failures were experienced by multiple aircraft. The more usual exothermic case, however, if substantial could result in anomalous loss of control of the aircraft. For a recent candidate incident, see "Jetliner drops 32,000 feet; 400 aboard; 50 are injured," Associated Press release, Huntsville Times, Feb. 20, 1985, p. 1 and "China Airlines Pilot Denies Cockpit Error," Associated Press release, Huntsville Times, Feb. 21, 1985, p. A-9. Deliberate use of this effect as an anti-strategic-bomber weapon capable of destroying bombers at or shortly after takeoff is possible (the Arrow DC-8 which crashed just after takeoff from Gander Air Force Base, Newfoundland on December 12, 1985 may have been deliberately interfered with by just such a Soviet weapon -- see extensive newspaper and TV news coverage of that event).

In scalar EM interference with aircraft, certain signatures may occur. Different instruments may sharply contradict each other, since their activation and inactivation is no longer a simple function. The pilot's instrument panel, for example, may indicate failure of the control system, when actually it has not failed, and the flight recorder may record that the control system is still operating. In such case, the pilot may take proper action based on his own instrumental readings, yet these actions may be contraindicated for the actual status of the aircraft, resulting in opposing pilot/autopilot actions. Tentative indications of the jetliner that anomalously dropped 32,000 feet and recovered are that apparently such anomalous instrument disagreements occurred. Another signature is that, as the EM charge accumulates in and on the aircraft, its skin charges up electrically, producing a glowing corona similar to St. Elmo's fire. The Arrow DC-8 was observed to be glowing yellow by eyewitnesses, according to the Canadian Broadcast Corporation's 10:00 p.m. network TV news on April 8, 1986.

47. T. E. Bearden, Star Wars Now! The Bohm-Aharonov Effect, Scalar Interferometry, and Soviet Weaponization, Tesla Book Co., Millbrae, California, 1984, p. 19-22.

48. For a paper possibly related to the first Bedini scalar wave detector from the "free A-field is a scalar field" view, see Robert R. Lewis, "Aharonov-Bohm effect for trapped ions," Physical Review A, 28(3), Sept. 1983, p. 1228-1235. This paper analyzes the role of the Aharonov-Bohm effect for the bound states of a charged particle in the superposition of two fields: a uniform magnetic field plus the vector potential of a solenoid. Gives exact solutions for a solenoid of zero size and discusses the corrections due to a solenoid of finite size. Shows that a series of new "betatron resonances" appear in the spectrum, dependent on the flux quantum number. The fundamental betatron resonance remains narrow and unshifted for solenoid radii large enough to broaden and shift the cyclotron resonance. See also W. Ehrenburg and R. E. Siday, Proc. Phys. Soc. London, Sect. B62, 1949, p. 8 which, among other things, discusses the effect of the vector potential in electron interferometry, where the insertion of a tiny solenoid between the interfering paths is shown to give an additional phase difference proportional to the flux through the solenoid. The resulting shift in the fringe system is called the Aharonov-Bohm effect.

49. Frank Golden has developed a most important and precise scalar EM detection system to display scalar EM signals on a cathode ray tube. His important measurements have mapped the scalar resonant frequency responses of the entire earth. Two completely different types of scalar instrumentation show precise agreement on the frequency maps. The measurements show that the Soviet Union has for some time entrained the entire earth into giant scalar electromagnetic resonances on two frequencies approximately 12 kilohertz apart. It is the opinion of the author that the beat frequency between these two giant resonances is utilized by the Soviet Union to extract enormous power from the earth, essentially by a mechanism that Golden has uncovered. A complex of four to eight weapon systems might reasonably be powered from the single bi-resonant tapping source. In April 1985 Golden discovered the rather sudden activation of many other scalar frequencies with characteristics resembling command and control communications. In addition, a significant number of additional paired scalar resonances had been entrained upon the earth. It is the

opinion of the author that this showed a major activation of the Soviet strategic scalar electromagnetic defense system, with the power up of from 80 to 100 weapons, along with the complete command and control system. Golden has previously built a prototype of an actual Soviet scalar communications system and demonstrated that it works. Because of the impending mayday and the 1973 statement by Brezhnev indicating that 1985 was a critical planning year, the sudden activation of such massive numbers of weapons and communications was alarming. Accordingly, the author discreetly notified certain contacts to alert the U.S. system. Golden then detected the periodic turn-on and turn-off of the systems, indicating a large command and control exercise rather than preparations for direct tactical use. An additional difference, however, was the change from scalar electric to scalar magnetic field on many of the large transmitters. This is indicative of electrostatic location, targeting, and adjustment of the target on the earth and then switching to the "power input" mode. It is the opinion of the author that this indicated the choice of a geological target and proceeding to attack that geological target in the "slow buildup" mode. Again, proper contacts were so notified.

50. A very good coverage of the Soviet microwave radiation of the U.S. Embassy in Moscow is given in Paul Brodeur, The Zapping of America, W. W. Norton & Co., New York, 1977, passim. None of our U.S. Embassy security personnel knew of scalar interferometry, nor did they possess scalar electromagnetics detectors. Further, orthodox U.S. scientists were also ignorant of scalar electromagnetics. The Soviets were thus in a position to tailor the radiation signals and precisely determine what our on-site personnel would detect. Details of the effects on U.S. Embassy personnel, possibly associated with the microwave radiation, are given by Brodeur. A great number of related releases by the open news media also exist. For example, see Robert C. Toth, "Radiation Devices Bug U.S. Embassy," Huntsville Times, reprinted from the Los Angeles Times, Feb. 8, 1976; "U.S. Screening out Soviet 'Listening Rays'," United Press International release, Washington D.C. Star, Feb. 11, 1976; Norman R. Beebe, "Nixon Exposed to Radiation In Russia?", Associated Press release, Huntsville Times, Apr. 30, 1976; Barton Reppert, "U.S. Scientists Puzzled Over Embassy Radiation Effect," Associated Press release, Huntsville Times, May 28, 1976; Barton Reppert, "Radiation Entering Embassy in Moscow through Phone Line," Associated Press Release, Huntsville Times, May 28,1976. See also Martin Ebon, Psychic Warfare: Threat or Illusion, McGraw Hill Book Co., New York, 1983.

51. For details of the electromagnetic transmission of cellular disease and death, see Vlail P. Kaznacheyev, "Information function of ultraweak light flows in biological systems," Problems in Biophysics, Novosibirsk, 1967, p. 7-18. Kaznacheyev et al, "Conditions necessary for appearance of distant intercellular interactions after UV-radiation," Bulletin Experimentalnoy Biologii i Meditsiny, (5), 1979, p. 468-471; Distant intercellular interactions in a system of two cultures connected by optical contact," Ultraweak Luminescence Biology, Moscow, 1972, p. 224-227; "Distant intercellular interactions caused by UV-radiation," Photobiology of a Living Cell, Leningrad, 1979, p. 221-223; "Studies on the biological role of electromagnetic emission as a factor of adaptive behavior of cells under conditions of latitude changes," Abstracts of the second all-union conference devoted to the problem of human adaptation to different geographical, climatic, and labor conditions, Novosibirsk, 1977, p. 101-104; Vlail P. Kaznacheyev, S. P. Shurin and L. P. Mikjhailova, "Distant intercellular interactions in a system of two tissue cultures," Official Bulletin of the committee on inventions and discoveries affiliated to the council of ministers of the USSR, Discovery no. 122 (19), 1973, p. 3, also printed in Psychoenergetic Systems, 1(3), Mar. 1976, p. 141-142; "Apparent Information Transfer

Between Two Groups of Cells," Psychoenergetic Systems, 1(1), Dec. 1974, p. 37.

52. For evidence of anomalous scalar reflection out of the scalar grid caused by a Navy chaff drop, See "Navy sabotages traffic lights, TV sets," United Press International, PM cycle, Jan. 11, 1985; "Power blackout blamed on anti-radar particles," United Press International, AM cycle, Jan. 11, 1985; "Power Disrupted Briefly After Navy Plane Drops Metallic Strands," Associated Press, AM cycle, Jan. 16, 1985.

53. A good summary of the early incidence of the Soviet Woodpecker signals is given by Bradley Wells, "The Russian woodpecker: a continuing nuisance," CQ, Nov. 1984.

54. A particularly important Soviet Paper is E. B. Smetanin, "Electromagnetic field in a space with curvature -- new solutions," Soviet Physics Journal, 25(2), Feb. 1982, p. 107-111. In this paper, a solution to the stated problem is obtained, which may be regarded as a classical model of a charged particle that has both a magnetic moment and a nonzero magnetic charge density. Interaction with spacetime curvature can disrupt the gauge invariance of the EM field, and it can also reestablish invariance in an initially noninvariant theory. (Hutchinson has had anomalous breaks of metal caused by standing scalar waves, which seems to be due to a mechanism that produces monopoles and deposits them in the metal at standing-wave nodes. This paper should be directly applicable.)

See also Yu. S. Vladimirov and A. A. Kozlenkov, "6-optics and a general theory of gravitation and electromagnetism," Soviet Physics Journal, 27(12), Dec.1984, p. 1039-1041; V. G. Bagrov and M. D. Noskov, "New exact solution of the Dirac equation. XI," Soviet Physics Journal, 27(12), Dec. 1984, p. 1030-1034. See particularly V. I. Petukhov and I. A. Kuzin, "Weak processes in the field of a gravitational wave," Soviet Physics Journal, 27(12}, Dec. 1984, p. 1025-1029.

In the latter paper it is significant that the decay probability increases in a locally curved spacetime, and that solutions are obtained for circularly polarized waves.

(In EM waves, circularly polarized waves have special characteristics in nonlinear isotropic media: exact solutions exist. Also, when two longitudinal sine waves of differing frequencies are transmitted into a nonlinear isotropic medium, the apparent difference frequency is transmitted as a sine wave, undistorted. This characteristic has been used by Westinghouse to produce improved sonar transmission through seawater, for example. The scalar EM (electrogravitational) wave may be represented as a longitudinal EM wave. Immediately, if one wishes to produce a pure scalar EG wave in the nucleus of an atom, two frequencies should be transmitted such that their difference is the frequency wished. Or circularly polarized waves may be utilized.)

Note that, by implication, one should be able to deliberately tailor scalar EM waves to greatly speed the decay of long-lived radioisotopes, such as are contained in radioactive wastes from nuclear reactors and other processes.

55. For some evidence of continuing persistence of the higher frequency scalar components applied to the scalar grid on Jan. 28, 1986 to localize the effects when attacking the Challenger, see "Poison suspected in unexplained deaths of birds," Huntsville Times, AP release, Feb. 6, 1986. From 1-4 Feb.

1986, various kinds of dead birds -- such as blackbirds, cardinals, sparrows and bluebirds -- plummeted to the ground in Birmingham, Alabama. At first some sort of mysterious poison was suspected, but later was ruled out. During this period, the woodpecker grid was particularly active in weather engineering operations. Multiple sightings of giant cloud radials occurred in various places throughout the U.S. The retention of the high frequency components possibly was intensified in the Birmingham area sporadically during the period. Birds flying into this zone would be killed if their brains acted as scalar receivers and were sufficiently jammed by the signals. Normally Huntsville (further north) is a "hinge point" for increased scalar activity, to start the Jetstream bending northeast to move along the Appalachian mountain chain. However, just before 28 January the jetstream had been diverted further south than usual by the Soviet interference, bringing unseasonably cold weather to Florida and the waiting Challenger on its launch pad. Probably hinge points further south than Huntsville were needed for this operation. If these more southern points were retained for a few days, and the high frequency components were retained, then an inadvertent high intensity area of these components may have sporadically occurred in the Birmingham area, killing the birds.

For description of the Apr. 18, 1986 Titan blast and its ramifications to the spy satellite program, see James Gerstenzang and Ralph Vartabedian," Los Angeles Times News Service, "Titan Blast to Hurt Spy Program," Huntsville Times, Apr. 19, 1986.

56. For another very important paper indeed, see Vu. G. Ignat'ev, "Relativistic kinetics of an anisotropic plasmalike medium with damping in a field of gravitational radiation," Soviet Physics Journal, 27(12), Dec. 1984, (English translation: Plenum, June 1985), p. l066-1069. If we substitute the vacuum "virtual plasma" for the observable plasma, this paper may be part of the smoking gun to show that the Soviet Union has developed scalar electromagnetics. Extrapolating from this paper, the amplitude of the "induced longitudinal field" is proportional to the square of the amplitude of the gravitational wave (GW) in an initially isotopic medium, but reduces to a linear proportionality as the medium develops nonlinearities (anisotropies or inhomogeneities). Whereas conventional GW are considered to be weak, the scalar GW may be very strong since we ourselves determine the variation in the energy density of vacuum it involves. Further, sharply spiked scalar EM (EG) pulses may develop appreciably more amplitude that the linearly calculated amplitude from the magnitudes of the zero-summation components. The end result is that (1) we can produce powerful gravity waves using the scalar EM approach, (2) local spacetime can be curved, producing violation of the conservation laws, and (3) the assumption of a local Lorentz frame made by Einstein as a limitation on ordinary general relativity is falsified.

In passing, note that immediately one also has a clear basis for the excess local production of energy when the battery furnishing power to a DC series motor is sharply spiked backwards in a "recharging" fashion. Further, the ions in the battery (which are several hundred thousand times as massive as an electron) possess a resonance on the order of 16 megahertz in a typical lead-acid battery. Under the correct conditions, the ions collect the excess energy (due to the initial square effect) produced by a sharp pulse of recharging potential. The ions are shocked into resonance and, on reversal, "overshoot" and deliver this excess energy to the plates, recharging the battery. Thus such a system can violate linear conservation of energy because spacetime is locally curved by the spiking, and excess energy from the excited local vacuum is collected by the ions and delivered to the plates. This is precisely the effect upon which the first Bedini free-energy motor depended. The timing and adjustment, however, are highly critical and difficult, and the battery takes a rather severe beating.

57. For details and photographs of highly anomalous exhausts seen on Bennett Island, see "Explosive Events Seen on Soviet Island," Aviation Week & Space Technology, Sept. 26, 1983, p. 31. NOAA-6 and NOAA-7 weather satellite photographs of massive smoke plumes, one of them 150 miles long and nearly horizontal, are shown. Also, in U.S. weather satellite photos of Novaya Zemlya, similar plumes have been detected, making this location a tentative candidate for a second Soviet "exhaust site." See also Letter John M. Miller, Geophysical Institute, University of Alaska, Jan. 4, 1985 with attachment, "Bennett Island plume cases recently found." Satellite photos showing actual circular breaking of arctic ice have also been taken near Wrangel Island. See Sam Bishop, "UA photos show Soviets break ice for missiles," Fairbanks Daily News-Miner, Dec. 15, 1984, p. 1,3; Craig Covault, "Soviet Abi1ity to Fire through Ice Creates New SLBM Basing Mode," Aviation Week & Space Technology, Dec. 10, 1984, p. 16-17. Wrangel Island in fact has a bizarre history: it belongs to the U. S., but has been illegally given to the Soviets and occupied and used by them. For details, see "Wrangle over Wrangel and the Weather War," Don Bell Reports, 32(9), Mar. 8, 1985.

58. For details of anomalous blinding (probably by scalar EM means) of U.S. satellites, see Philip J. Klass, "Anti-Satellite Laser Use Suspected," Aviation Week & Space Technology, Dec. 8, 1975, p. 12-13. For details of anomalous failures and recoveries of a British satellite, see Joe Schwartz, Nature, 280, July 12, 1979, p. 95.

59. For other possible evidence of traces of Soviet scalar EM activity over the oceans, see George D. Curtis, "An electromagnetic radiation pattern over the ocean," Undersea Technology, 5(8), Aug. 1964, p. 29-30, 40. In producing a scalar wave, there exists a sort of "signal to noise" ratio of the desired scalar wave to an accompanying, undesired normal EM component. Due to practical equipment imperfections, the generated scalar wave inevitably is accompanied by a small residue of ordinary EM wave. While measures can be taken to further purify the scalar wave, a residue always remains. Curtis may have detected the weak EM residue accompanying early Soviet interference gridding for weather control testing over the U.S.

60. Interactions which occurred between Khrushchev and renowned physicist and Nobelian P. Kapitsa are of interest. Khrushchev desired absolute defense of the Soviet Union so that the Soviets might be able to launch any action desired without risk or serious concern. Kapitsa informed the Soviet leader that, if a means of total neutralization of foreign missiles was to be found, it could only come from a group of new principles in physics which was called " energetics." The term "energetics," of course, was contracted from "psychoenergetics," and is essentially the expanded physics/electromagnetics that this author has dubbed scalar electromagnetics.

Strangely, a portion of a book by Marshal Grechko was deleted from its English translation at the specific request of the Soviet Union to the U.S. State Department. The passage stated: "Of particular importance is basic research aimed at discovering still unknown attributes of matter, phenomena, and the laws of nature, and developing new methods for their study and use to reinforce the state's defense capabilities."

The most definitive Soviet book on military strategy is titled Military Strategy (Voyennaya

Strategiya). See V. D. Sokolovskiy, Soviet Military Strategy, third edition, edited with an analysis and commentary by Harriet Fast Scott, Crane, Russak & Company, Inc., New York, 1975. There are three editions, the latest in 1968. In the third edition, the exotic weapons normally proposed for defense against strategic ballistic missiles -- such as particle beams, high energy lasers, exotic nuclear warheads, plasma weapons, antigravitational weapons, etc. -- are discussed and discounted by implication. The statement is made several times that 100% defense against missiles and aircraft is possible. A statement is also made that it is interesting to note that this capability (a solution to the strategic missile defense problem) has been achieved by the Soviet Union, but not by the West (e.g., p. 298.) In the same edition, a previous definition of antimissile defense obviously consisting of radars, computers, and interceptor missiles is deleted (see author's note 190, p. 454).

The Soviets also recognize the great importance of technical surprise: "The experience of war shows that warring sides quite often have tried to employ new types of weapons to this end to achieve surprise, which are a surprise for the enemy or are little known to him. Each side secretly develops new means of warfare in order to employ them unexpectedly. History knows many examples how the employment of a new weapon initially gave considerable success because the enemy, caught unaware and not knowing the combat capabilities of this weapon, was for some time incapable of effective counteraction… The mass introduction of new weapons into an army usually cannot remain secret for long for the other side. -- New means of warfare, and in mass numbers, are impossible to create often and quickly. In this regard, the sides have begun to take other paths concerning the attainment of surprise along with searches for new weapons and their sophistication." V. Ye. Savkin, of The Basic Principles Operational Art and Tactics, Moscow, 1972, U.S. Air Force translation, U.S. Government Printing Office, Washington, D.C. 20402.

61. The Soviets have not discarded most of their old radars, and have continued to develop, build, and deploy giant new radars (in apparent violation of SALT agreements), which U.S. analysts always analyze in conventional fashion. It is fairly straightforward, simple, and inexpensive to modify an ordinary pencil beam radar to produce scalar wave beams and geometrical scalar Fourier expansion forms. Radars can provide a wide array of totally new and unexpected capabilities when utilized in the scalar electromagnetics mode. Even older Soviet radars, with antiaircraft gun sites and older surface-to-air missile sites, will exhibit new and startlingly advanced capabilities against ballistic and cruise missiles and Strategic Air Command bombers. Since there are many thousands of these older Soviet radars still operationally deployed, in the scalar mode the Soviet strategic defense may well approach 100%, just as Military Strategy implies. Also, the antiaircraft and antimissile defense capability of Soviet land forces may be far greater than anything presently estimated by U.S. analysts.

62. To the physicists, engineers, and researchers who have made the tough sledding to this point, congratulations! Here's a bonus for you.

Turn to page 32, equation [7]. Time differentiate it (this is left as an exercise for the reader). Collect terms and study the result intensely.

Now you have the reason why rise time is so important from a scalar EM viewpoint. You can also see why spiking the innate inductance of any system very sharply is so important. Fast rise time greatly increases the value of the time derivative terms, and your gravito-inertial effects go up with the collective squares of your zero-vector summed components. Any system has at least a little inductance. For sharp spikes, the gravitational potential goes very high. The inductance of the system momentarily opposes any current (translation: opposes any current bleed-off, or EM force field bleed-off). During that time you've got the 4-space G-potential to use. If you need to hold it longer to use it, then increase the inductance.

Now you can see why extremely rapid discharge of properly tuned, giant Tesla coils give gravitational and inertial effects sometimes.

You can see why the sharp "EMP" (electromagnetic pulse) from a nuclear explosion or EMP generating device can penetrate the ocean.

That's why the spikes (such as used by the neurophone) can carry messages directly into the brain and nervous system. You also should see why the Soviet woodpecker carriers (a Soviet over-the-horizon radar system) are loaded with scalar EM spikes (Bedini has precisely detected them from 50 Hz to 10 MHz; Golden has detected them into the gigahertz range.) to work on the humans in the targeted area. They're zapping the hell out of us. And they have done something similar at the U.S. Embassy in Moscow for decades.

Bedini has developed a simple, cheap device to mess up the phaselocking of the scalar EM spikes on the woodpecker carriers. It's based on the fact that the Russkies use a lot of circularly polarized transmitters when they make the spikes. Here's what to do: Take 4 strong bar magnets and an old (but operational) long-playing record turntable. Set it on 33-1/3 RPM speed. Glue the magnets on the top of the turntable, standing vertically, symmetrically placed around the turntable's outer edge. Place them so that on one a north pole is up, on the next a south pole is up, then a north up, then a south up. Now turn the turntable on and let it run. This gadget makes a slowly rotating magnetostatic wave. (Yes, it will also make a rotating "magnetic" wave, which is actually two rotating magnetostatic waves.). These rotating magnetostatic waves will interfere with the scalar spikes, changing them to EM (electromagnetism) and slowly rotating their phase. This breaks the phaselock of the Soviet circularly polarized scalar EM signal on the human brain. This gadget will protect the average large room.

But back to our differentiated equation.

Now you should also see why the first Bedini motor is so "devilishly difficult" to adjust. You've got to get the effect and examine it in well under 20 nanoseconds. You'll need an oscilloscope of 350 MHz range or so. Microwave-quality bench equipment is necessary; the ordinary electrical and electronic shop equipment is not good enough. You can also see why the ordinary lead-acid battery takes such a beating. However, a special battery certainly can be designed to take it. Now you know.

Okay, you tigers. Go get 'em.

63. Now here's a final bonus for the persistent reader.

This time, see J. N. Brittingham, "Focus Wave Modes in Homogeneous Maxwell's Equations:

Transverse Electric Mode," J. Appl. Phys. 54, 1983, p. 1179-1189; T. T. Wu and R. W. P. King, "Comment on 'Focus Wave Modes in Homogeneous Maxwell's Equations: Transverse Electric Mode' [J. Appl. Phys. 54, 1179 (1983)]", J. Appl. Phys. 56, 1984, p. 2587; T. T. Wu and H. Lehmann, "Spreading of Electromagnetic Pulses," J. Appl. Phys. 58, 1985, p. 2064-2065; T. T. Wu, "Electromagnetic Missiles," J. Appl. Phys. 57, 1985, p. 2370-2373. Pay particular attention to the latter paper.

Because of the Strategic Defense Initiative (SDI) and other programs, great interest has recently been kindled in using EM pulses to carry energy from a high-powered source to a target. However, so far almost all the work is still couched in terms of purely ordinary (though sophisticated) electromagnetics. It still involves the EM concept of "EM energy flow through the intervening space from the transmitter to the target." As such, the work to date still suffers from all the self-imposed ills and limitations of using the classical force fields concept of EM.

However, the theoretical work of these referenced papers is of great interest to us here.

In the first paper, Brittingham sparked a great deal of interest with theoretical work which promised a permanently focused 3-dimensional concentration of EM energy that propagates in a straight line with the speed of light. That is, this mode used a newly formulated wave packet (called the "focus wave mode") that acted as a non-dissipating EM soliton.

In the second reference, it was concluded that Brittingham's focused wave packet actually satisfied Maxwell's equations, but not the associated boundary conditions.

[Comment by this reviewer: both of these formulations use "linear time" and a "linear frame," ignoring the local general relativity aspects. Indeed, Brittingham's "focused wave packet" can be made by the zero-vector approach! However, it will not be an "EM" packet, it will be a pulse (or oscillating wave, depending upon how it is made) of 5-gravity potential. As such, it contains a multiplicity of locked-in, 5th-dimensional, non-zero EM gradients in and EM gradients out. Collectively these gradients themselves constitute a deterministic "flux of gradients" or an EM flux in the 5th dimension, with no resultant effective current of EM gradients. In other words, it's possible to build a "flux of gradients" and to have a "current of gradients," just as in ordinary EM it's possible to have a "flux of virtual particles" and a "current of virtual particles." The argument concerning the EM boundary conditions of Maxwell's equations is not relevant in such a case. The only boundary involved is the zero itself.]

Be that as it may, the 4th reference is of direct interest. In this paper, Wu has done us a great service. He shows that, under transient excitation, an antenna of finite size can transmit energy to a faraway receiver so that the energy decreases as slowly as one wishes. An electromagnetic pulse of energy with such slow decay is conveniently referred to as an "electromagnetic missile."

The electromagnetic missile has important applications, according to Wu. One possible application is the transmission of information in a manner that is very difficult to intercept or jam. A second possible application is to destroy or disable a distant target. However, Wu feels that so much energy is required [remember, he is speaking of EM non-zeroed force fields traveling through ordinary space] that a like1y radiating source must be something like a nuclear explosion. However, as Wu points out, a nuclear explosion produces a powerful EM pulse which includes frequencies from microwaves to gamma rays. To weaponize the explosion to produce a coherent beam, the nuclear explosion must be controlled to produce a coherent radiating source and a pulsed laser-like EM beam. Wu points out that such a technology is difficult and has not been developed.

We comment as follows: It is certainly true that enormous energy is required if the waves in the pulse remain electromagnetic -- that is, if they have non-zero resultant E and B fields to the external observer.

On the other hand, if we form a complex of waves in opposition so that we control the energy content of each individual component wave, but blend all the waves so that their collective envelope wave has zero E-field and zero B-field resultants, then -- to the external observer -- the EM energy is infolded and unzipped electromagnetically. It is now electrogravitational energy, and limitations of "EM force field energy flowing through space or through a physical medium" need no longer apply. However, we may still look at each EM component wave as an individual EM wave, moving in the Kaluza-Klein 5th dimension that is wrapped-around each and every point in our ordinary 4-space. In other words, to the ordinary 4-spacetime observer, each component wave has a special channel around every point of his space, not through it. Thus we have established multiple EM channels "just outside" each point of observer space. These we call "scalar channels," and EM energy can move through each one as if it were moving through a wire. The relative potentials between the "ends of each wire" determine in which direction the "EM energy current" flows along the wire.

And in each of these "special channels," we just have a normal kind of space (to the internal observer). Wu's paper assures us that we can get as much of the input energy to the distant target as we desire. The problem then is how to break the zero-summed EM envelope (and hence the "zero channel" for each component EM energy wave) in the distant target itself, spilling the infolded contents of all the special channels back into ordinary spacetime as EM energy (either positive or negative).

That, of course is easily done by interference in the distant zone. As previously stated, there is a way in which we can even perform this interference in a single scalar beam. Since the newly formed EM energy emerges from each and every increment of spacetime occupied by the distant target, it already emerges inside the very vitals of the target. It does not have to translate through the intervening ordinary space between transmitter and target. Instead, it goes through the Kaluza-Klein fifth dimension. It translates along a line through our space, but not through our space. With the scalar EM approach, we transmit EM energy through hyperspace, not through normal space.

Hereafter we will adapt and expand Wu's SDI term "electromagnetic missile" to include these electrogravitational aspects. In other words, we include the "scalar EM missile", which consists of an infolded multiple of Wu's ordinary EM missiles moving together coherently in special sheathed hyperspace channels.

In Appendix I we show an anomalous photograph taken by Bob Gladwin, just before the shuttle launch of Nov. 26,1985. While the photograph may be an artifact, it also is a candidate for the test strike of such a Soviet scalar "electromagnetic missile" tested against an aiming point offset from the shuttle launch which followed a few seconds later. The test was accomplished before the shuttle was launched, and offset from it, so that it positively would not endanger the shuttle itself. The conditions under which this spectacular photograph was taken happened to coincide with the special conditions required to photograph such scalar EM energy. Those conditions are clarified and explained in the Appendix.

My deep appreciation goes to Mr. Gladwin for allowing me to use this photograph. As we will point out in Appendix I, the video camera caught a picture of the Soviet scalar EM marker beacon associated with the Titan explosion of April 18, 1986. One can rest assured that the Soviets probably destroyed that vehicle, using a scalar EM missile.

APPENDIX 1: SOME IMPORTANT INDICATORS AND DESTRUCTIVE TESTING

Some of the indicators which bear on the use of the Soviet
scalar electromagnetic grid via the Woodpecker EM carrier beams are given in this appendix. along with other information believed pertinent to overall scalar EM effects.

IN ADDITION, THE THESIS OF ACTUAL LIVE TESTING OF THESE WEAPONS BY THE SOVIETS -- TO INCLUDE THE DESTRUCTION OF A TROOP-CARRYING U.S. AIRCRAFT, A SHUTTLE, AND A TITAN MISSILE -- IS ADVANCED AND INDICATORS PRESENTED.

An anomalous photograph taken by Bob Gladwin is presented which possibly may have captured the "smoking gun" -- the strike of a
Soviet scalar EM missile at an offset aim point near the shuttle Atlantis, just before its launch on Nov. 26, 1985. However, this photograph is caveated since it may also represent an anomalous but ordinary artifact. captured in some as yet unknown manner. The manner in which such a photograph may have been taken -- when normally the scalar energy does not show on a photo -- is explained.

Another photograph taken by George Suchary after the launch did
capture the Soviet "marker beacon" utilized in the test firing, however. This independent photograph is presented, and adds credence to the thesis that the Gladwin photograph captured the test of a Soviet scalar EM pulse missile.

- INDICATORS AND SUCH -

1. Continual cloud radials have repeatedly been seen and photographed over Los Angeles, California or its vicinity.

Los Angeles is also apparently a "hinge point" for bending and controlling the jetstream. Three typical radials seen to form one after the other by Margaret Wilson (and sketched and airbrushed by her) in the greater Los Angeles area are shown in figures 1, 2, 3, and 4. These radials were formed on Sunday, Jan. 26, 1986. Ms. Wilson had heard my Saturday night "Soviet Weather Engineering" broadcast over KABC Los Angeles, and my request that anyone who saw the typical radial cloud formation should photograph or draw it, noting time and location, and send the photos or drawings to me.

As Margaret described it. "To my astonishment, the next day I not only saw one, but three in the sky at the same time.

The time of the first one was Sunday, Jan. 26, at 1:15 p. m.

. . . I stepped out the back door for some minor errand and noticed a
fan of long straight clouds in the sky. The sky had been filled with dark clouds most of the morning but this had been breaking up and the western half of the sky was now clear with these lines of clouds spread out on it. I immediately thought of your request and tried to see the bottom of the fan which was hidden behind bushes and a garage. When I stepped into the street, I saw there was a semi-circle beneath them.

Not having film and a wide-angle lens on my camera. I knew I couldn't capture it in its entirety. I am, however, a retired commercial airbrush artist so I quickly sketched the formation. It is illustration #1.

It was in the western sky, the base of the central semi-circle sitting about 15 degrees above the ocean horizon. It was very clear and sharply defined. The semi-circle was broken at the top and the two broken ends curled inward like commas. As I sketched the wind deformed it somewhat, the right half of the whole formation sliding down away from the left.

Ten minutes later it was getting quite clear of clouds except
for a bank over the Santa Monica Mountains to the north and there, straight north, another fan was forming. As I watched it took shape -- illustration #2. The same broken semi-circle (with commas') but this time the bases of the rays were feathered about half-way out with end halves long straight wisps.

As I sketched, another fan had begun to take shape further east. It was hard to tell its size but I had estimated the first one at about 15 miles across and the second a little larger and the third larger yet. At the peak of its formation it looked like illustration #3. ...There in the middle of the central circle, hazy and a little faint was what looked like the top of a mushroom cloud. Or was it only a distant thunderhead that coincidentally found its way to a place behind the formation? The arms of the last, largest formation were rather wavy and feathery like some gigantic sea creature. It did not disperse as fast as the others, but held for a couple of hours, though it was pretty distorted by then. I could not estimate its size as I wasn't sure of its altitude.

Were they aircraft contrails? Certainly not! There was an airliner in the sky at the time -- a mere scratch on the smooth blue sky (see illustration #3, right center.).

The three formations were present in the sky at the same time although the first was getting quite distorted by the upper level winds before the last reached its peak.

Tuesday our weather changed from hot and sunny to rainy as the weather man said the path of the jetstream altered, and then there was the space shuttle explosion.

...The Russians seemed very sincere in their sympathy for the loss of our astronauts but only two days later it was printed in TASS and I personally heard Vladimir Posner say that Russia hoped that the United States would now realize what could happen if the S.D.I. program was further pursued. Was this friendly concern or a threat? It is very evident that if the S.D.I. can destroy a missile in flight, it can also destroy one on the ground. This would immediately put all Russian missiles in [reach of] our arsenal. They are already in place in the most desirable target areas -- their own launch sites. I'd be panicky too if I were a Soviet leader."

[Ed. comment: My sincere thanks to this thoughtful lady who took the time to sketch the cloud radial phenomena occurring in the Los Angeles "jetstream pivot point" at this critical time, just prior to the launch and destruction of the Challenger.

Anomalous cloud grid activity (formation and dissolution) in the Los Angeles area has been positively correlated with Soviet Woodpecker activity by engineer Ron Cole, with 100% correlation in 30 straight Woodpecker measurement runs. Statistically, of course, this is quite decisive.].

2. Continual cloud radials have repeatedly been seen and photographed over Huntsville, Alabama, associated with weather engineering and grid activity. Typical radials seen by this author are shown in Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985. Huntsville is apparently a pivot point or "hinge point" for bending and controlling the jetstream.

3. Spectacular cloud grid patterns have been seen over Huntsville, Alabama and Los Angeles, California. An excellent videoclip of one of the Los Angeles area grids, personally observed and shot by KABC Open Mind talk show host Bill Jenkins, is shown in Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985. An artist's sketch of an earlier gigantic grid, extending from horizon to horizon in every direction, seen over Huntsville, Alabama by Tom Bearden and Ken Moore, is also shown in the tape. The videotape is available from P.O. Box 1472, Huntsville, AL 35807. Sketches and details of another highly anomalous grid pattern observed in the greater Los Angeles area by engineer Ron Cole are shown in figures 5 and 6.

The reason that clouds of water droplets and/or ice particles detect the scalar interferometry and form signature patterns is simple: Consider each H_2O molecule as having two light little hydrogen atoms hanging on to the much heavier oxygen atom. The covalent sharing with the oxygen atom of the electron from each hydrogen atom means that (1) the electrostatic scalar potential between the H ion and the O ion it is bonded to, is rhythmically varying as the electron is shared back and forth, (2) this varying potential contains "electron spin holes" since it is made by the moving electron, (3) two such varying potentials exist since there are two H atoms sharing covalent electrons with the O atom, (4) the two H ions are at an angle of over 100 degrees with respect to each other, (5) the molecule assembly thus constitutes one part (one half, so to speak) of a scalar interferometer with imbedded electron spin holes for electron hooking, (6) Incident scalar waves from outside the system interact with the "half scalar interferometer." This scalar interferometry interaction is coupled to the covalent bonding electron because of the internal spin hole pattern of the molecular half of the interferometer. The coupled electron moves with the scalar pattern's changes, causing an observable interaction with and in the electrical structure and ionic potential of the molecule, (7) the weak H to H hydrogen bonding between molecules connects resulting ionic potential changes to the structuring of the entire macroscopic cloud assembly. Thus by scalar interferometry with the H_2O "half interferometers", an incident scalar EM pattern is detected and translated into patterns of E and B force fields. The H_2O molecules then "line up" in accord with the electrical patterns detected.

Thus the clouds form "signature patterns" according to the incident scalar EM radiation patterns. Giant cloud radials in the targeted area are probably due to the type and shape of antennas -- including the type and shape of the electrical wires establishing the antennas' ground planes -- 640used by the Soviet Union to project the scalar waves. Note that such "radial" antenna patterns were occasionally used by James Harris Rogers in his underground and undersea scalar EM transmission system.

[For Rogers' patents, see U.S. patent numbers 1,316,188, Radiosignaling system, Sept.16, 1919; 1,322,622, Wireless Signaling System, Nov. 25, 1919 (figure 6 of this patent clearly shows a "radial wire" ground plane antenna); 1,349,103, Radiosignaling system, Aug. 10, 1920 (the fourth version in figure 1 shows inside/outside variation which, for the long waves used, to a distant observer appears as a varying magnetostatic scalar potential using opposing B fields); 1,349,104, Radiosignaling system, Aug. 10, 1920; 1,220,005, Wireless Signaling System, Mar. 20, 1917; 1,303,729, Wireless signaling system, May 13, 1919; 1,303,730, Radiosignaling system, May 13,1919; l,315,862, Radiosignaling system, Sept. 9, 1919; also 958,829, Method and Apparatus for Producing High Frequency Oscillating Currents, May 24,1910.].

4. Anomalous power outages in Ventura County, Southern California on Nov. 8 and 9, 1985. Again, engineer Ron Cole, who lives in the affected area, investigated and reports as follows:

"On the evening of 11-8-85 in a rural area of Ventura County known as Carlisle Canyon, a strange electrical phenomenon began to occur at about 10:15 p.m. Our first observation was that random white dots began to appear on the TV screen, then began to increase in size and brightness. The brighter the dots got, the dimmer certain lights in our home became. Yet later we found that, when some lighting circuits were dimming, other circuits were causing lights to shine much brighter than normal. After a bit of checking, we were able to determine that lighting circuits common to pole 1 of the service transformer were the ones that were seeing a reduction in voltage, while all circuits on pole 2 were seeing an increase in voltage.

There is another interesting condition that was observed on the TV screen. The dots were in sync with the vertical field or frame rate of the signal. This means that the interference had to be of a frequency of 59.7 Hz or 29.85 Hz. Normal power line interference will form a grouped dash pattern across the screen and will strobe upward at the rate difference of 59.7 Hz or 29.85 Hz. In the 8 November case, there was no strobing of the dots whatsoever.

These conditions continued off and on for the rest of the entire evening and were observed by no less than six other persons that I know of. They were observed over the entire TV frequency spectrum, low band VHF, high band VHF, and on two UHF channels. Line interference is almost always most prominent on the low band VHF frequencies only.

These anomalous events would take place on a somewhat irregular time pattern; anywhere from 3 to 10 minutes apart. They would always last for about 2 minutes.

At 12:30 a.m. I went to bed and was not aware of any further activities until 6:20 a.m., at which time I was awakened by several sharp snapping or cracking sounds coming from the walls of my house. I rolled over and went back to sleep, only to be reawakened about 15 minutes later by my wife, who informed me that the power and phones were out. My comment was, "Some dummy must have crashed into a power pole," and I dozed off again.

At 10:15 a.m. the power and phones came back on line. Shortly after, our phone began to virtually ring off the hook. About a dozen neighbors up and down the canyon wanted me to check out their electrical wiring and various appliances. All of them were trying to explain their individual encounters and what took place when the power went out. Such things as wires snapping in walls, singing and humming sounds coming from wiring in walls, strong smell of ozone, light bulbs exploding, buzzing coffee pots, and a handful of other phenomena all took place just before the power went out at 6:20 a.m.

A typical example was that of Rob Steel. [Ed: Name changed for this report; his real name is in my files.] Rob lives four miles down the canyon from us. He had just turned on the lights in his barn and was getting ready to feed the livestock, when he realized he had left several needed items in the garage located about 75 yards from the barn. Upon entering the garage, he saw the garage lights began to flicker, then go to a dim orange glow. Upon stepping from the garage area, Steel saw that lights in the barn were so bright that he thought the entire structure was about to explode. He froze in his tracks for a moment and his whole house and garage began to hum. Then everything went out.

An interesting point to inject is that Mr. Steel's home and several others farther down the canyon from us are being fed from a different 19 kv power feeder source than the 14 kv feeder that supplies us and most homes in the canyon.

At about 2 p.m. on 11-9-85 I was in contact with the power utility company (Southern California Edison Company) to inquire as to the cause of the power outage. My first assumption was wrong. No power poles had been knocked down; only a ceramic fuse in the 19 kv feeder system had blown and a two-pole oil breaker had kicked out on the 14 kv feeder system. Load charge recorders at the power company show that both fuse devices kicked out at exactly the same time, but they are in no way common to each other as to load distribution. They only supply power to adjacent areas.

Also, the phone company has no record of its service being out, yet the phone service was restored when the power was restored. The phone company does not use Edison Company's power to power its phones. Again, the only commonality is that the phones and phone lines were in the same area.

Numerous magnetic breakers, used as manual switch devices for well pumps and other services, were on when power was returned. These switches had previously set in their off state.

A strong electrical-type ozone odor was detected at five locations at the time the power went out.

Later we found that a Research and Development laboratory in a nearby area measured 29 Hz intermittent signals late on the night of 11-8-85. These signals were interfering with some critical tests the lab personnel were trying to accomplish. The laboratory itself did not experience a power outage.

At this point in time, the only realistic surmise we can come up with is that a tremendous ELF or scalar ground wave was adding and subtracting to the grounded neutral service side of the power pole transformers so as to sum in and out of phase. When peak nodes were achieved, all line transformers ended up back-feeding the feeder lines out of phase, causing the final power outage.

Twenty-one persons are aware of the strange conditions of the power outage and prior events. [Ed: Their names are in my files.]"

Editor's comment: Another anomalous power outage also occurred in the same area. This occurrence is possibly associated with effects of the scalar interference grid adjustments. If the Soviets were preparing for a major test of the grid in the Launch Phase ABM mode, substantial tests and exercises may have been performed before the actual "wet run" against the shuttle launch on Nov. 26, 1985 -- about two and a half weeks after the anomalous power outage at Carlisle Canyon.

See the Rogers' patents, referenced above, for ideas of how scalar EM signals couple to the power lines and telephone lines.

As this book goes to press, John Bedini has notified me that his new and very sensitive developmental scalar detector positively verifies scalar frequency "spike" activity on the Woodpecker grid (and possibly the ground reference) from 50 Hz to 10 MHz. A normal instrument will definitely not see these signals. John's lab is at Sylmar, California and about 70 miles from Carlisle Canyon.

5. During the Skylab mission, NASA astronauts observed and photographed a mysterious green glow of the atmosphere. The glow was over 1,000 miles long and over 100 miles wide. From Carlisle Canyon, Ventura County, California the sky has also been observed to possess a mysterious green glow from horizon to horizon, on three occasions. [Editor's comment: It is hypothesized that this green glow is associated with grid activity in the Launch Phase ABM System mode.]

6. Just prior to the hapless launch of the Challenger on Jan. 28, 1986, the Soviet Union sharply bent the jetstream down through the center part of the country, then curved it eastward across the Florida panhandle. This exposed the waiting Challenger on the launch pad to severe cold air, brought down from Canada by the sharply deviated jetstream. The use of the scalar interferometry aspects of the Woodpecker grid for weather engineering over the U.S. has been documented previously. [Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985;

Bearden, "USSR: New beam energy possible?", Defense & Foreign Affairs Daily, 13(111), June 12, 1984, p. 1-2.]. Wilson's cloud radial sightings on Jan. 26, 1986 and Cole's giant cloud grid sightings on Jan. 24, 1986 are directly connected with adjusting the Woodpecker grid and deviating the jetstream, to carry cold air to the Florida launch site.

7. The previous testing of the grid in the Launch Phase ABM/Anti-bomber Mode, using actual jet aircraft as "wet-run" targets, has been demonstrated by various anomalous incidents.

As one probable example, see "Jetliner drops 32,000 feet; 400 aboard; 50 are injured," AP release, Huntsville Times, Feb. 20, 1985, p. 1 and "China Airlines Pilot Denies Cockpit Error," AP release, Huntsville Times, Feb. 21, 1985, p. A-9. In this incident (see figure 7) a Boeing 747, flying at 42,000 feet and bound for Los Angeles, experienced unusual engine flameouts while still out from San Francisco. In a hair-raising plunge toward the earth, the pilot fought to restart the engines and regain control of the aircraft. After a fall of some 32,000 feet in two minutes, the engines finally restarted, and the pilot was able to fly into San Francisco and land safely.

We have already explained how the Soviets are able to track the aircraft's jet engines. Later in this Appendix we show just how accurately an actual "scalar EM missile" may be focused by the Soviets on a hapless target. If a negatively biased pulse or continuous wave scalar EM "energy extraction" beam hits the engines, the extraction of energy will result in engine flameout. (No heat energy, no combustion of the fuel.). This is apparently what happened to this aircraft. In the presence of "splatter" from the extraction beam targeted at the engines, the aircraft instruments were seriously affected. Different instruments have different responses to such an anomalous scalar EM kindling effect. Thus a signature would be that the instruments disagree with each other, and need not record what actually occurred. This appears to have been the case, constituting a signature that scalar EM was targeted against the aircraft. The pilot's instruments apparently indicated that the control system had failed, while the flight recorder apparently indicated that it had not.

Let us assume this was a Soviet test. When the distant Soviet operator turned off the extractor beam, the engines returned to normal state after a "discharge time constant" period. They normalized before the plane struck the ground, so the pilot was able to restart the engines and proceed.

Notice that this was a limited test to cause engine flameout only. If the scalar EM missile had been used against the aircraft in the "pulse" mode, it would instantly have exploded in mid-air.

The crash of an India jet aircraft outside Great Britain may have been a test of the same weapon in the pulsed "destroy aircraft" mode, since there are indications that the aircraft suffered an anomalous midair explosion. Another candidate for such a possible "shoot-down" of a jet aircraft is an (Italian) Itavia DC-9 that crashed into the Mediterranean in June, 1980. Both Itavia and the Italian Transportation Minister Salvatore Formica believed that a missile struck the jetliner and was the likely cause of the crash. It may have been a scalar EM missile and a full-up test by the Soviet Union. See "Plane was probably hit, judge says," Birmingham Post-Herald, Jan. 17, 1981.

Also the present author has knowledge of at least one instance in the early 1980's when an American-made tactical ballistic missile was anomalously destroyed shortly after being launched by troops in Europe. The destruction was anomalous and a large hole was inexplicably burned through the side of the missile -- exactly as if it had been struck by a scalar EM missile test in the anti-tactical missile (ATM) role.

8. As another example of airborne scalar EM activity in the Woodpecker grid, in January 1985 a U.S. Navy-dispensed chaff cloud, dispensed well-offshore from San Diego, California was caught by an unexpected wind and blown toward the city. (See figure 8.). As the chaff cloud moved in, significant failures of electrical systems and electrical components resulted. Power was disrupted to as many as 60,000 homes by the incident.

The sudden, unpredicted wind was suggestive of a wind blowing toward an artificially created low pressure area. Under certain conditions, a specialized structure such as that of a piece of sophisticated chaff will reflect scalar waves of frequencies within its cut bandwidth. If scalar modulation frequencies within the chaff bandwidth were present on the Woodpecker carrier grid above and surrounding San Diego, a myriad of reflecting bits of chaff in the moving chaff cloud would produce myriads of random, invisible "fireflies" of EM energy kindled at a distance, from random scalar interferometry, in a zone surrounding the cloud. As these invisible "firefly" pulses of

EM energy pulsed inside components of the electrical systems and electrical controls in San Diego and vicinity, extensive and random internal electrical interference occurred. This lead to widespread, anomalous electrical failures throughout the city.

Note that this activity preceded the huge Soviet scalar EM exercise of April-May 1985. The grand exercise would certainly have been scheduled well in advance by Soviet strategic forces controlling the scalar EM howitzers and Woodpecker grid weapons, and adjustments and smaller tests would have been being made for some months before the tests.

9. Other anomalous "electrical strikes" certainly exist, suggesting Soviet testing may occur widely throughout the world, yet be reported as ordinary but anomalous phenomena. An example which comes readily to mind is the anomalous blast (or blasts) that struck Bell Island, off the east coast of Newfoundland, on April 2, 1978. Persons on the shore of Conception Bay reported seeing fireballs coming down out of the sky and hitting the island, according to Royal Canadian Mounted Police. Highly anomalous atmospheric electrical activity preceded the blast. This may have been Soviet testing of scalar-created electromagnetic missiles. Extensive information on this incident can be obtained from P.A.C.E., 100 Bronson Ave., Apt. 1001, Ottawa, Canada T1R 6G8.

As another possible example, villages in an area of West Donegal in Ireland have been terrorized by freak "lightning bolts" for over two years. Roofs have been literally stripped off, windows smashed, telephones and television sets blown up, and pipelines wrecked. On Dec. 25, 1984 a mighty blast caused a three-day blackout: it accompanied a snowstorm, the first "white Christmas" in the area in 20 years. See "Bolts from the blue," Fortean Times, Issue no. 45, Winter 1985, p. 5.

10. Scalar spikes or pulses from the Soviet testing may also cause other anomalies, including deaths of unsuspecting persons using the telephone. For a candidate event of this type, see "Killer Phones," Fortean Times, Issue no. 45, Winter 1985, p. 8-9; also AP release, Huntsville Times, May 30, 1985; AP newswire June 2, 1985; and New York Times, Sep. 11,1985.

On May 21, 1985, five days after passing a complete physical examination, 17-year-old Jason F. Findley of

Piscataway, New Jersey was on the phone at his grandmother's house, talking to his girl friend. An odd click occurred on the line, followed by a gasp from Findley and then only the TV playing in the background. His grandmother found him unconscious, still clutching the phone in his hand. Rushed to Muhlenberg hospital, he was pronounced dead shortly afterwards. An autopsy could not determine the cause of death, and the telephone company found the phone working properly, properly grounded, and no evidence of excess electrical charge. Findley was lying on a wooden bed which would not conduct electricity. He had a hemorrhage of the inner ear. There had been an electrical storm over Scotch Plains, New Jersey the night Findley died.

AP wire services were told by the special forensic investigator with the New Jersey medical examiner's office that about six people had died in the U.S. in similar circumstances. Others had reported being knocked unconscious by "high voltage shock" from their telephones. Several weeks before Findley's death, another man from Whitehouse Station, New Jersey had been found unconscious with a telephone in his hand. According to the Consumer Product Safety Commission in Washington, D. C., in 1984 nearly 12,000 people in the U. S. were taken to emergency rooms with injuries related to telephones. A hundred of them died, although exact details were not revealed.

Note that the giant Soviet scalar exercise of May 1985 had been detected and verified by Golden. Substantial activity in the Soviet weapon complex, and in the Woodpecker grid over the U.S, continued throughout May and beyond. A phone line may by chance detect one or more atmospheric scalar EM pulses, just as occurred in the blackout in Ventura County, California on Nov. 8-9, 1985. These scalar pulses can travel along a phone line and be broadcast into a person's ear and brain from the telephone earpiece. The proper scalar EM pulse can induce both "acoustic shock" in the inner ear and "electrical shock" conditions in the nervous system. The pulse may also interrupt the body's electrical timing pulses to the heart, causing the heart to stop without physical damage. The scalar EM shock to the brain and nervous system can knock the person unconscious or, if sufficiently strong, kill him instantly. The condition is worsened if the person is ungrounded, since then he acts as a large, open-ended capacitance in series with the telephone instrument's inductance. In such case, the scalar resonance of the equivalent series LC circuit can be sharply excited and resonated. By the nonlinearities of the body and nervous system, the scalar resonance can be translated into an ordinary electrical pulse or resonance. Since the scalar power is not translated into ordinary power except inside the nonlinearities of the human capacitor/nervous system, no damage to the telephone or phone lines need happen. [At the ComTec exhibit in Las Vegas in 1984, John Bedini demonstrated powering a 200-watt speaker system through a 100-foot length of #40 magnet wire, using scalar EM means. He put some 200 "watts equivalent" of scalar energy through the tiny wire to the speakers, where a small Bedini translator translated it back into ordinary electrical "force field" energy, powering the speakers. This was much to the astonishment of the technical audience, since the wire was quite cool and exhibited no evidence of heating, even though by conventional thinking it should have been instantly melted.]

The bottom line is: With increased Soviet scalar weapons activity over the U.S., the telephone system is likely to become more hazardous to humans using it, when the instrument is connected to the external wire. A radio phone is much safer against anomalous scalar EM "spiking."

11. The metal-softening detected and verified by Frank Golden on Jan. 1, 1986 is certainly indicative of a new mechanism or effect added to the Woodpecker grid. The highly specialized nature of the particular detector/amplifier which Golden was using provides strong evidence that the signal was actually derived by recording the scalar EM signals produced by a human cerebrum (two cerebral hemispheres connected by the corpus callosum, forming a scalar interferometer) when the human was successfully performing psychokinesis (metal-bending). The fact that Golden was able to nullify the signal locally by special methods worked out by him lends further credence to this hypothesis. In any event, three hours of very objective experimental data strongly support this thesis. It is therefore indicated that (1) the metal-softening signal had been added for a purpose, and (2) something made of metal was going to be softened, leading to its failure. Certainly the Challenger would have been an ideal target; softening the metal on its boosters and fasteners by even a small percentage would significantly reduce the vehicle's ability to withstand the

stress during launch. This of course is circumstantial evidence only, but the finger of suspicion is there nonetheless. It is a fact that the metal-softening signal was and is in the Soviet Woodpecker-repertoire; it is a presumption that it was intended for use against the Challenger.

- FINAL DRESS REHEARSAL BEFORE THE BIG SHOW -

12. The shuttle launch from Cape Canaveral on the evening of Nov. 26, 1986 seemed smooth as silk to the army of official observers and onlookers. In reality the Soviet Union clandestinely conducted one or more Launch Phase ABM System tests against the vehicle, using it as a convenient target.

Figures 9, 10, 11, and 12 are included to give the reader some preliminary background on the highly significant events which occurred in conjunction with this flight, and previous to it. These figures are newspaper articles from the News Tribune, Fort Pierce, Florida. The articles are used by permission.

Anomalous atmospheric "booms" and airquakes had occurred in the Florida area for over three and a half years prior to the launch. Sixteen counties, most of them on the northern Gulf coast or close to it, had reported booms during the two-year period preceding April 1984. State and local emergency preparedness officials and U.S. Air Force Base representatives met but could not resolve the mystery of what was causing all the anomalous booms.

In fact, many booms had occurred in and around the gulf coast of Florida, particularly offshore from St. Petersburg, starting from several years earlier.

These anomalous booms were caused by the Soviets adjusting and testing the Woodpecker scalar EM grid for weather engineering and potential weapons use. The booms and airquakes were certainly not confined to Florida alone, but also occurred at many other locations throughout the U. S. and Europe.

Mysterious, highly localized seismic microquakes also had occurred at widely scattered locations throughout the U. S.

On at least two shuttle flights prior to the Nov. 26, 1985 launch, large anomalous "booms" had occurred over or around the launch site after launch. No one in NASA knew what they were or what caused them.

So prior to Nov. 26, the Soviets had already prepared and tested the Launch Phase ABM system mode of the grid against at least two U.S. shuttle launches. However, so eerie and strange are the scalar EM weapons that they had been able to do so without the U. S. government even becoming aware of what was going on. In the Kremlin, they must have marveled at our continued ignorance.

At any rate, they tested the Launch Phase ABM system again on Nov. 26, and this time private U.S. citizens happened to capture at least some evidence of what was going on. It still would have died there without surfacing, except for a curious chain of events. Let me briefly recount those events.

****

Lucius Farish publishes a fine little monthly newsclipping collection called UFO Newsclipping Service (Route 1, Box 220, Plumerville, AR 72121). In it she includes clippings on all sorts of anomalies, including UFO's, strange light phenomena, etc. Because of my long work in UFOs (for some time now, placed on the back burner due to the press of my work in Soviet scalar electromagnetic weapons), I've subscribed to Farish's service for several years.

And this time my unorthodox interests certainly paid off. Farish published a clipping of the article, "NASA officials stumped by strange light, boom," by Susan Burgess, News Tribune writer (see figure 10). The moment I read the article, I realized that the Soviets had actually tested the Launch Phase ABM System's area mode against the shuttle launch of Nov. 26, 1985. Further, George Suchary had taken a picture of the marker beacon as it was slewed away at the conclusion of the test.

With alacrity I contacted Ms. Burgess and, through her, George Suchary. Both were very helpful, and Mr. Suchary gave me permission to use his important photograph of the light moving away after the test.
Ms. Margery Bril , the kind librarian at the News Tribune, sent me four clippings.

After reading the article by Julie Enders, Showcase Editor of the News Tribune, I contacted her also. Although extremely busy, Ms. Enders kindly gave me additional details, including the fact that Bob Gladwin had taken a series of eight stills of the launch itself, and in one frame an anomalous, curving line of light with a splash on the end appeared.

At first I thought it might be the answer to a dream. Very tentative information suggested that the "curved light with a splash on the end" might have occurred in one of the pictures taken during the launch sequence itself, with the light striking directly under the shuttle shortly after it lifted off. I was positive the Soviets could strike at the shuttle (or a missile) with a "scalar EM pulse" missile, but hardly dared hope someone could have actually been lucky enough to get a photograph of it.

From my acquaintance with two masters of paranormal photography -- the indomitable Trevor James Constable and a close friend, Joe Gambill -- I understood how one could possibly get a picture of such a strike -- assuming it occurred -- at night, when the same picture could not have been taken in the daylight. I also understood why the camera might catch it when the human eye would not see it.

It seemed that, between them, these friendly, cooperative persons just might have gathered together the positive evidence to prove that the Soviets had struck at the shuttle with an offset test of a scalar EM missile, using a marker beacon for adjustment. (As it turned out. it was not to be as definite as that!) However, time was of the essence.

The Soviet tests had definitely included a test in the "area burst" mode, which can catch and destroy several missiles fired from a deployed silo complex at once. The clinching evidence I had needed for so long on a Soviet test of the pulse mode -- the scalar EM pulse missile -- just might be there in one of Bob Gladwin1s photographs. But there was no time to lose; such evidence is fleeting. and unless I moved quickly, it was likely to fade away and disappear forever. With persistence, the rest followed swiftly. Mr. Suchary graciously gave me permission to use his photographs. These of course prove that the marker beacon was there. (In addition, it was seen by numerous persons.). The News Tribune gave me permission to print the articles. Mr. Gladwin furnished his negatives and gave me full permission to analyze and use his photographs. Personal friends, Everett and Shirley Edwards, printed the photographs in their private darkroom so the negatives would not be out of safe hands for even a moment.

At the time, this book was almost finished. If the photo of the candidate Soviet use of the electromagnetic missile was accurate, quick work would be necessary if it was to be included. My friend and publisher, John Ratzlaff. had already agreed to swiftly publish this large output as soon as I could get it to him. At Omni Video Publishing, Ms. Electra Briggs was alerted to the potential importance of what was coming. She and Mr. Glenn Shoemaker swung by my house from New Orleans, and a quick, informal videoclip was taped, announcing the book. (Murphy's law -- if anything can go wrong, it will -- got me there. Inexplicably, in my near-total fatigue from months of 19-hour days and 7-day weeks on this project, I spoke the date of the Titan explosion as April 19th, knowing full well it was the 18th! It happens. Electra suggested she slap a correcting statement over the video. Done. Be ready to get it out. It's too important; don't sweat the small stuff.)

Zerox copies of the photos were received. There in the fifth photo (or so it seemed) was an unequivocal strike of the scalar EM missile, in the launch sequence itself. The assumed strike seemed to have occurred right under the shuttle, a few seconds after liftoff.

The evidence seemed complete. All that was left was to receive the actual negatives and verify the sequence. At last we seemed to have the smoking gun. The Soviets would have been caught redhanded.

Based on the advance zerox copies, the manuscript was altered. Everything was ready. And then the negatives came in.

In this business, you learn to meet and accept a lot of failures. Things often go astray, and don't wind up as they seem at first blush. This time was no exception. Murphy got us again! The photos on the negative were in a different order then they had been on the zerox copies.

There's no mistaking the order in which the pictures were taken. There they are on the negatives, one right after the other, and clearly numbered. And the curving light with a splash at the end did not occur during the launch sequence at all. Instead, it was taken before the launch -- just how long remained to be seen. Everett and Shirley took the photo and developed it at both extremes: very dark and very light, to bring out all the details possible.

The results were most interesting. As shown on the lighter print (figure 13), the light does not curve down out of the sky and strike toward the ground, it turns up toward the sky and strikes above the distant treeline. Further, as shown on the darker print (figure 14), it does consist of a little ball of light, just like a scalar EM missile would appear. No apparent structure to hold the light appears in either picture, though one could not positively rule this out completely without microdensitometer tests. (One can inadvertently get artifacts like this by focusing on a ball-like light bulb, for example, and then moving the camera while the shutter is still open. On the other hand, there is no evidence that this happened at all, and no street lamp post or other foreign object in the photo.)

(The remaining photos of the liftoff sequence are shown in figure 15.).

We called Mr. Gladwin. He verified that the photos had been taken in order, one right after the other. He had not had the camera on a tripod, but had steadied it on a railing. In other words, the "light streak with a burst on it" had been taken just before the liftoff. Also, Mr. Gladwin had not observed 'the marker beacon, and had not attempted to photograph any such "light in the sky." In fact, visually he had not observed any such streak and burst of light in any scene he was photographing.

So the question is, exactly what did Mr. Gladwin photograph in that photo, just prior to the shuttle's liftoff? The answer is. we just don't know for sure.

Let's run through the two possibilities:

(1) He may have snapped an artifact while moving the camera. If so, the camera was definitely quite still at least most of the time the lens was open, for the treeline is quite clear in the lighter photo, which reduces light saturation and shows details more clearly. Further, there is no evident support or structure attached to the light itself; it seems to be a "disembodied" light. So this possibility is ambiguous, though reasonable. Mr. Gladwin does not recall any such movement of the camera; it was steady while he was shooting.

(2) He may still have snapped a scalar EM missile. If so, it means the Soviets fired a "small test" of the exothermic pulse or scalar EM missile in advance of the actual shuttle launch. Let's pursue that for a moment, and suppose it to be true.

The Soviets already had the marker beacon up in the sky -- that's certain. Many persons saw it. Soviet trawlers off shore and Soviet satellites could also have given the actual position of the marker beacon. Thus the Soviet LPABM System would have had (a) the actual parameters the computer had used to transmit the beacon, and (b) the actual location at which the beacon emerged. Comparison of (a) and (b) is much like having a "meteorological data message" in field artillery: One has the corrections to apply for specific conditions of the medium that cause deviation.

These corrections can then be applied to calculated or "table" data to adjust firing for real-world conditions.

So far so good. The Soviets had the exact position of the registration beacon, and could apply corrections for a shot of the scalar EM missile. They certainly would know the exact position of the shuttle on its launch pad.

If a fairly substantial offset for the scalar EM missile shot were used, "mett corrections" could be applied and a small scalar missile could be zinged in, stopping momentarily at the final offset "strike" position so that a track could be obtained by the scanner extraction beam. Then by comparing the detected offset position of the "burst" of the scalar EM missile to the detected offset position of the marker beacon, the difference could be applied to the beacon's true position to get the true strike position of the missile. The offset used could then be subtracted to give the location where the missile would have struck without the offset, had it actually been fired at the rising shuttle after liftoff. This "adjusted final strike position" could be compared to the shuttle's expected early trajectory (which varies only slightly from standard for the first few seconds) to determine the expected combat success of the shot. It's a little roundabout, but it's a very safe way to do it if one wants quite positive assurance that the missile will not strike the shuttle and destroy it. Since the Soviets delayed the "area burst" and its concomitant atmospheric "boom" to T + 12 minutes, it is evident that a large offset was used for the area strike -- probably about 10 extra minutes -- to positively assure that there would be no actual damage or threat to the shuttle.

It seems logical, then, that they might fire the small scalar EM missile well offset from the shuttle, and just before its launch, to prevent any possibility of actual damage to the shuttle.

So there you have it. Mr. Gladwin's photograph is consistent with such a careful, safe scalar EM missile test by the Soviets, prior to the actual liftoff of the shuttle. On the other hand, it's not possible to say with certainty that that is what it is. It may simply be a very peculiar artifact. In fac, if it were not for the other indicators, one would most certainly not consider such a bizarre possibility as an offset Soviet scalar EM missile strike at the shuttle.

If the "streak" photo is examined as a candidate for a sca1ar EM missile test, then one other thing must be explained: Why did the film capture the test when observers did not see it at all? That, of course, we can explain!

So back to our presentation. Here I'm going to assume -- and accent that this is a postulation -- that the Gladwin photo actually represented a scalar EM missile test. We will run through a possible sequence of events, based on that assumption. We will hypothesize the LPABM system operation and the Soviet conduct of the test. The major events here in the U.S. may be quite real; those assumed to occur in the Soviet Union, of course, are "author's license" -- but hopefully realistic.

****

The evening of Nov. 26, 1985 is perfect. At Cape Canaveral, the shuttle Atlantis is poised on its launch pad waiting to move majestically into the heavens on its impending journey. As launch time approaches, everything is as smooth as a cat's whisker. There is, however, just one small anomaly.

A mysterious light is hanging up in the sky, staying still for awhile, then gently rocking back and forth. The light is a Soviet marker beacon for a scalar EM howitzer. The howitzer, employed in a special scalar EM channel contained in the Woodpecker transmitters and interference grid, will shoot through the EM carriers at an aiming point offset from the shuttle.

Since this is still to be a nondestructive test, no chances will be taken. The pulse mode will be used, and a "scalar EM missile" -- a bundle of energy buried in hyperspace around its moving zero-point in this world -- will be fired directly at an offset aiming point prior to launch of the shuttle. A substantial offset in distance will be used to prevent posing any danger of inadvertently hitting the shuttle on its launch pad.

The marker beacon is necessary for precision. Its desired position will be calculated and the beacon will be placed on station. By placing this source of continuous EM energy in the area, it can be detected by the scanning "energy extraction beam" and tracked by the distant Soviet operator. Soviet trawlers and satellites monitoring the test will report the true position of the marker beacon, which will deviate from the calculated location due to conditions of the channel medium (the Woodpecker beams) at the time. Comparison of the beacon's true position to its desired, calculated position "registers" the marker beacon and yields the "meteorological corrections" which must be applied to standard firing table calculations. These corrections can be applied to any further shot into the area to correct its aim, just as standard field artillery applies registration corrections in the target area.

When the shuttle boosters ignite, an additional track will be obtained from it. The offset from the received track of the marker beacon and the received track of the shuttle booster flame thus will give the exact correction or "shift" from the marker beacon settings that should be used to set the howitzer's aim. In other words, a "relative" firing, offset from the marker beacon registration track, will automatically compensate for any vagaries of the medium experienced by the beams in transmission.

The Russians, being excellent artillerymen, have simply applied the standard artillery technique of shifting from a known registration point.

The Soviet Launch Phase ABM System is cocked and ready. This is the last test in the "wet run" series in which an offset will be used. If everything goes as planned, future shots will move into the next phase, where actual U.S. targets will be destroyed. The two great powers are already at war, but one of them does not know it yet. Because of the importance of this last test before the next step in hostilities commences, Gorbachev himself is monitoring the test.

The countdown begins. In Florida a great many persons are observing the event. George Suchary has already spotted a strange light hanging in the sky, where none ought to be. Bob Gladwin, resting his camera on a railing, prepares to snap his first picture so that he will be sure to have a sequence completely across the launch.

Deep within the Soviet Union, a large display screen is illuminated. The marker beacon's track is clearly shown on the display, as are synthetic symbols for the shuttle on its launch pad, several Soviet ships off the coast of Florida, and several Soviet satellites.

The control officer barks, "Missile test! Proceed!" The simulation officer presses a switch, and a simulated shuttle launch track appears.

The operator responds immediately, "Launch detected!" The identification officer, scanning the frequency components and the signature of the energy extracted from the launch, calls out, "Missile! Target hostile!" He presses a button, and a light illuminates on the panel, marking the target as a hostile missile. The launch track begins to blink.

The control officer announces, "Prepare to fire!" "Pulse mode!" "Strength one!" "Test!" "Apply burst offset!" The operator rapidly flips switches, setting the mode of fire and the offset. Instantly lights light to show that the computer has computed the shift and offset, and corrections. A red "Ready to Fire" light illuminates and a low siren begins its wobbulating growl.

The control officer orders, "Fire!" The operator presses the fire switch. A "Fire" light appears. A slight, deep shudder is felt briefly as the howitzer fires its scalar EM pulse missile of deadly energy. On the screen, the path of the scalar EM missile is traced out as it speeds toward its target far over the horizon.

In Florida, George Suchary is getting more and more curious about that pesky light. Bob Gladwin is preparing to snap his first photo as the final seconds of the countdown tick away.

On the distant Soviet screen, a sudden small burst appears, superposed over the synthetic shuttle track. The operator yells, "Burst!" On a smaller screen, a vertical profile shows the strike directly underneath the target. Immediately a blinking "Target Killed" light appears as the computer assesses the shot a kill.

The operator excitedly calls out, "Kill !" "Direct Hit!" Cheering breaks out from the firing team and the onlookers. Gorbachev and his entourage are smiling broadly and clapping. The cheering quickly hushes as the crew and participants continue to watch the track of the rising shuttle, and prepare for the launch and the next test.

Bob Gladwin has snapped his first picture (figures 13 and 14) leaving the shutter open a bit to allow for the nighttime conditions. As luck would have it, the scalar EM missile strike Occurs during the time the shutter is open. The film has captured the strike by a special process, even though human eyes have not seen it. (We will explain that shortly.).

At 7:29 the Atlantis's boosters ignite and the vehicle begins to rise smoothly off the launch pad. She will continue to rise, then pitch over to a 45 degree angle and proceed down range..

Bob Gladwin is snapping pictures regularly, one after the other.

In the local Florida area, hundreds and hundreds of persons are watching the flawless launch. The tip-off of the Indian River Community College basketball game and the St. Lucie County School Board meeting are held up while participants watch the Atlantis begin to rise serenely upward.

In the Kremlin, another track has appeared on the control screen display. Once again the sharp-eyed operator snaps, "Launch detected!" The identification officer, scanning the frequency components, tersely rasps "Missile! Target hostile!" Again he presses a button, and a light lights on the panel, marking the target as a hostile missile.

The control officer orders, "Simulate multiple firings!" The simulation officer acknowledges, "Multiple firings simulated!" and presses a button to cause the computer to simulate multiple launches on the display.

Immediately other synthetic target tracks begin to appear on the display screen, one by one, clustered around the shuttle launch at various separation distances, prearranged to represent multiple firings from a missile silo complex for this exercise. A special audio alarm sounds as each track appears.

The operator immediately calls out, "Multiple tracks! Missile launches!" and then "Identified hostile!"

The control officer orders, "Prepare to fire!" "Area pulse mode!" "Test!" "Apply time offset ten!"

The operator flips switches rapidly, acknowledging. Lights illuminate on the console, showing status of preparation for firing. The "Ready to Fire" light illuminates and the low siren begins again.

Back in Florida, Bob Gladwin has stopped snapping photographs. George Suchary has gone into his house to find his Polaroid camera and take a picture of that pesky light still hanging up there in the sky.

Back in Russia, the fire control officer orders, "Fire!" The operator presses the fire switch. The "Fire" light illuminates, immediately blinking, and the siren changes tone. On the display screen, all target tracks are blinking. A large light on the console comes on, showing that an "exercise time offset" of 10 minutes has been applied. A digital clock starts ticking off the minutes.

Minutes pass as the offset time ticks down.

Back in Fort Pierce, George Suchary is coming out to take photographs of that darn light, which is still hanging up there!

Folks are streaming back into the Indian River Community College gymnasium for the basketball game. Attendees at the St. Lucie County School Board meeting are wandering back inside, to get ready for the meeting.

More minutes pass.

Back at the control console in the Kremlin, the offset clock ticks down to zero. The siren ceases its wobbulating tone and holds a steady note. The "Fire" light ceases blinking and illuminates with steady brilliance. A deep shudder is felt as the howitzer fires again. On the screen, the rapid trace of the mighty scalar EM serpent is shown as it races toward its target.

Suddenly it reaches the target area, and a large blossoming light appears, covering all the target tracks. The operator announces, "Burst!"

All the target tracks flare brightly, then dwindle and die. The "Targets killed" light illuminates and the operator yells, "Kill! All targets killed!"

Wild jubilation breaks out. Everyone is grinning and clapping. Gorbachev personally congratulates the fire control officer and pumps his hand. The tension relaxes visibly. In the noise, the operator touches a joystick and slews away the marker beacon, preparing to shut down the system.

The final test has been a crowning success. Now the real first phase of active combat begins. "And the fools don't even know it yet!" a grinning Gorbachev delightedly announces.

Back in Fort Pierce, precisely 12 minutes after liftoff, high in the atmosphere above the launch site and a little downrange, a tremendous rumbling boom suddenly rocks the coast for hundreds of miles in each direction. The blast is heard as far north as Charleston and as far south as Key West.

George Suchary is taking pictures (figure 10) as the hanging light in the sky suddenly begins to move away, obeying the controls of a Russian operator thousands of miles away. Bob Gladwin has already finished snapping his photographs (figures 13, 14, and 15) of the launch and taken down his camera.

The anomalous light disappears rapidly in the distance, moving faster than any jet aircraft.

People gradually resume their normal activities, talking about the beautiful evening and the perfect launch. Many of them have noticed the strange light that hovered and bobbled in the sky, then sped away.

The final Soviet registration of their eerie weapons for World War III has been completed. A totally different kind of Pearl Harbor -- more akin to a Trojan Horse -- has been prepared for the unsuspecting United States.

Countdown toward Armageddon begins.

****

George Suchary's photograph of the marker beacon, just after it had started slewing away, is shown in figure 10, as printed in the News Tribune.

Bob Gladwin's photographs of the shuttle launch are shown in figures 13, 14, and 15.

Figures 13 and 14 show two developments of the same shot, one light and one dark. The light one shows the treeline in focus in the distance, with the light streak and burst.

Figure 15 shows liftoff of the shuttle and its rise and tilt over on the first part of its trajectory.

Figure 16 shows a possible detailed interpretation of the first Gladwin photo. The simulated aim point was offset from the shuttle, and the strike was directly at the offset point. The computer computed the burst strike strike of the scalar EM missile from the adjusted registration data, and determined whether or not the target was killed.

The control precision is shown by the small size of the scalar EM missile.

The strike of the weapon is caught by the camera, but not seen by the naked eye. None of the many hundreds of onlookers saw the strike, though all of them saw the shuttle launch.

There is a very good reason for that, though it is not commonly known.

Constable

The reason is that Bob Gladwin's camera was utilizing a form of photography pioneered -- to the best of my knowledge -- by Trevor James Constable. At least I learned about it from Constable's books and articles. I then told it to my good friend, Joe Gambill, and Joe applied and adapted it to the unusual photography which he performs. Both Constable and Gambill are masters of this type of photography. Once I got deeply into scalar electromagnetics, I finally figured out what was going on in this methodology.

Briefly, here's the way it works. Scalar energy, of course, has to be detected by some sort of interference phenomenon. There exists a very special way to get a camera to perform such interference for you.

In the EM spectrum, the infrared and the ultraviolet have a special relationship to each other. The ultraviolet is exactly twice the frequency of the infrared, if the two zones are properly chosen. In other words, the UV is the first harmonic of the IR.

The scalar EM energy, since it operates in the Kaluza hyperspace surrounding every point in ordinary space, may be considered to be composed of pure spin. (That is, it's moving in a dimension where each particle of that dimension is spinning). Harmonics there are spin harmonics.

In this photon-interaction produced level of reality (the ordinary world), the IR and UV bands are more closely connected to the "shadow world" in which scalar energy moves. Thus there is a sort of shadowy "swirling" of the scalar energy near the IR and the UV. Entry of the scalar energy into this world, through the IR and UV windows by a kind of "harmonic interferometry," is normally prevented by the presence of visible light. That is, visible light "squelches" the "paranormal channel" that lurks beneath the IR and UV zones.

To make use of the IR and UV interference source zones, Constable covered his camera lens with a special filter (18A) which is opaque to the visible light spectrum and transparent to the IR and UV. Note that the film is sensitive to the IR and UV region we are talking about, but the human eye is not.

When pointed at a sufficiently strong source of scalar energy, such a specially filtered camera has unique characteristics. It strips out the visible spectrum, but still admits the IR and UV spectrum. The scalar energy surges through these two source windows, and phaselocks harmonically. That is, the scalar energy entering through the UV window fits twice neatly in the scalar energy entering through the IR window. This interference, on the film, is (recall) swirling in frequency. The net result is that the film records the interfering scalar energy as visible spectrum energy.

Infrared film can be used to increase the response in many circumstances.

Both Constable and Gambill have rigorously proven this technique in literally thousands of "paranormal" photographs. (They do not necessarily use my scalar EM terminology, but the technique is rigorous and works, regardless of terms.)

Under nighttime conditions, sometimes the particular lens and film combination will phaselock between IR and UV for incident scalar EM energy. This is particularly true when a relatively strong source of IR is in the vicinity.

While Bob Gladwin was taking his photographs, there was no suppression of the IR/UV phase-lock effect by visible light because it was night and there was hardly any visible light entering his lens.

Thus his camera lens and film formed an IR/UV phase-locked interferometer that detected and recorded the strike of the scalar EM missile underneath the shuttle, even though human eyes did not detect it.

At this point it would be interesting to know whether NASA may have had IR cameras focused on this shuttle launch (which was at night). If so, it would also be interesting to know whether any of them may have developed the IR/UV phaselocked interferometer effect, and captured the strike of the weapon.

Note that daylight shots would automatically break the IR/UV phase-lock possibility, by jamming out the effect with visible light, unless a special filter is used to filter out the visible light but leave both the IR and UV.

One would hope that both NASA and the U.S. Air Force will take note and discreetly add a specially filtered camera to the instrumentation observing and recording all future daylight launches.

To recapitulate the major events at the shuttle launch on Nov. 26., 1985, see figure 17. A marker beacon was placed in the vicinity, high over the area, for precise registration and location of tracks relative to it. A scalar EM missile strike may have occurred just prior to launch, and offset from the shuttle.

Then the shuttle lifted off. The exercise called for simulating multiple launches from a missile field. The area code was computed, and the fire order given. A substantial time delay -- say, 10 minutes or more -- was used to absolutely assure that the actual shuttle was well away from the engage blast.

Twelve minutes after launch, the delayed firing in an area "multiple missile kill" mode occurred, producing a massive boom heard for hundreds of miles up and down the east coast. After the boom, the marker beacon was slewed away, and photographed by George Suchary just as it moved.

The shuttle moved on downrange and went merrily on its way, blithely unaware of the jaws of the tiger that had snapped shut behind it.

To the Soviets, this final test of the preparation phase was highly successful. They were now ready to increase the risk level and fire the first round of World War III.

- THE SHOW BEGINS -

13. On Dec. 12, 1985 the same Soviet weapon tested against the previous NASA shuttle launches and against various aircraft may have deliberately interfered with the controls of an Arrow DC-8 taking off from Gander Air Force Base, Newfoundland. At an altitude of 100 feet, the aircraft -- carrying over 250 U.S. soldiers and civilian crew members -- lost power and sank into the ground tail-low, killing everyone on board. It was a tragedy of enormous proportions to the U.S., and especially to the families and friends of the brave servicemen and crewmembers who lost their lives.

Three Canadian witnesses to the crash were interviewed over the Canadian Broadcast network television news on April 8, 1986 at 10:00 p.m. No flame or smoke issued from the plane before its descent and crash.

However, witnesses reported seeing the aircraft mysteriously glowing with a yellow halo. That is a signature of the use of a scalar howitzer in the "continuous EM emergence" mode, similar to the manner in which several F-111's were downed in Vietnam.

In short, the DC-8's electrical systems were interfered with by electromagnetic energy and noise created throughout each increment of spacetime occupied by the aircraft. A powerful charge was rapidly created in and on the aircraft structures and skin. The "yellow glow" seen by the witnesses was a corona due to the skin of the aircraft acquiring a high electrical charge.

Something also apparently caused the rapid loss of two engines, one after the other. This indicates that the distant Soviet operator may have struck at the engines, one after the other, with a scalar "energy extractor beam." Thus he apparently tracked the aircraft as it moved down the runway, then used two modes against it: one to create serious EM interference with the electrical systems of the aircraft, and one to disable the engines.

With its controls ineffective and power drastically reduced, the doomed aircraft sank to earth, still in its "tail down" configuration from takeoff, and crashed and burned.

Admittedly other factors may also have contributed to the crash. These included reduced lift due to poor engine maintenance, increased weight of the aircraft due to icing, and heavy loading. However, normally the aircraft could still have taken off and flown to its destination. (See "Arrow Air was absolutely safe, FAA official says," UPI release, Birmingham Post-Herald, Apr. 17, 1986.)

To date, no official seems to nave recognized the sinister significance of the "yellow glow" or what it implies. No one seems to have connected the loss of this aircraft, the mysterious booms over U.S. shuttle launches, the Woodpecker interference grid, the anomalous loss of the U. S.S. Thresher, the high altitude "booms" off the east coast of the U.S. some years back, the puzzling loss of two critical Titan launch vehicles in a row, anomalous engine flameouts of civilian jetliners, etc.

Certainly officialdom has not recognized that a Soviet Launch Phase ABM System is being tested directly over its collective head, to include actual destruction of U.S. aircraft, missiles, and space vehicles with concomitant severe loss of American lives.

Direct and diabolical acts of war have been committed upon us by the Soviet Union, but in a fashion that still eludes our erstwhile pundits.

However, it is not as simple as that. To even energize giant scalar EM weapons is to run a risk -- to all mankind -- so horrendous that it boggles the imagination. Truly it deserves Brezhnev's 1975 phrase, "...more frightful than the mind of man has ever imagined."

Shortly we will reveal exactly what he meant.

- Apr. 18, 1986: Death of a Titan -

14. As this book goes to press, the last two U.S. Air Force Titan 34-D missiles fired from Vandenberg Air Force Base in California have blown up shortly after launch. The first one blew up on Aug. 28, 1985 just after lift-off. That loss has been attributed to failure of a high-powered fuel pump, causing a massive oxidizer leak and a smaller fuel leak. The second Titan loss occurred on April 18, 1986 when the Missile blew up 5 seconds after lift-off. Its loss is still under investigation and no determination of cause has been made. Apparently the shuttle and the Titan presently provide the only viable launch vehicles for launching U.S. "spy" satellites. The loss of these sensitive satellites -- if indeed they constituted the payloads -- cannot help but be damaging to our strategic surveillance capability. According to the Los Angeles Times, the single remaining KH-11 satellite was launched in December 1984 and, with an expected life of two to three years, it could stop functioning later this year.

Whether or not significant Woodpecker grid activity existed in the vicinity of the Titan launch of August 1985 is unknown at this time. However, significant activity in the grid definitely occurred before the April 18 Titan disaster and on the same day.

On Easter Sunday, 30, 1986 engineer Ron Cole observed significant cloud signatures of grid pattern activity, correlated with Soviet Woodpecker measurements. On April 18, the present author observed traces of a cloud radial over Huntsville, Alabama and took photographs of it. Preliminary reports from Thousand Oaks, California indicate extensive grid activity again on April 18, the day the second Titan exploded.

At least the second of these two missile destructions shortly after launch is highly suspicious, since the grid positively was active during that time. Also, a clear trail of Launch Phase ABM system indicators exists back to the massive scalar exercise of April/May 1985. The first Titan explosion in Aug. 1985 thus falls within the Soviets' "now let's test them against U.S. launch vehicles" period.

The second Titan loss follows highly suspicious losses of the Arrow DC-8 on Dec. 12, 1985 and the Challenger on Jan. 28, 1986. The same Soviet weapon system that destroyed those targets probably also destroyed one or both of the critical Titans.

At least, on the second Titan explosion, the Soviets left another signature. In one of the videotapes taken of the explosion, the marker beacon can be seen well , above the exploding debris, moving away from the explosion site.

See for yourself! Refer to "Titan Explosion Cripples U. S. Launch, Surveillance Capability," Aviation Week & Space Technology, 124(17), Apr. 28, 1986, p. 16-19. Specifically, look at the color photo of the explosion on p. 18.

Look well above the explosion (about one inch on the page) and to the left (about one inch on the page) of the center of the top of the fireball portion. There you will see a little light, caught in the very act of moving away.

At my rather frantic urging, engineer Ron Cole sought out a copy of the original videotape. He and a team spent two hours, going through this portion of the tape frame by frame. The light is there. It's real. And it moves on out of there, well above the explosion and independent of it.

It's the Soviet marker beacon, caught in the very act. Here I want to express my sincere appreciation to Ron Cole for his untiring efforts to nail this light down. Without his dedication to supporting this investigation, it would have been impossible to say for sure that the light was there.

So we have caught the bear red-handed, with his fingers in the cookie jar. If the marker beacon is there, you may rest assured that the Soviets destroyed that Titan missile, with very high probability. [See also "Investigators Search for Clues to Cause of Titan Blast," Aviation Week & Space Technology, 124(17), Apr. 28, 1986, p. 19-20; also "Titan Accident Prompts Funding, Technology and Policy Questions," AW&ST, 124(17), Apr. 28, 1986, p. 21.].

Continue

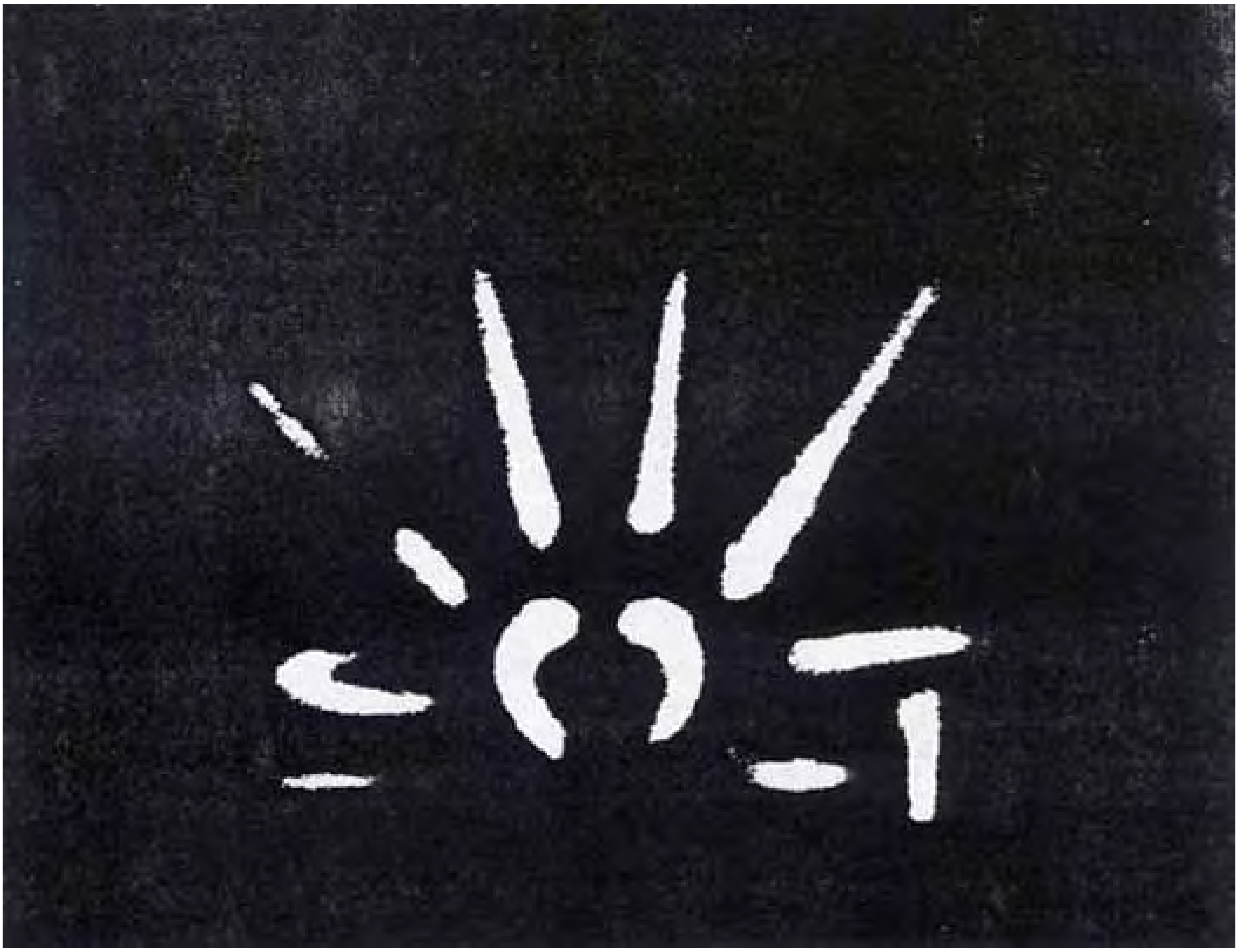

Figure 1. Wilson's first cloud radial

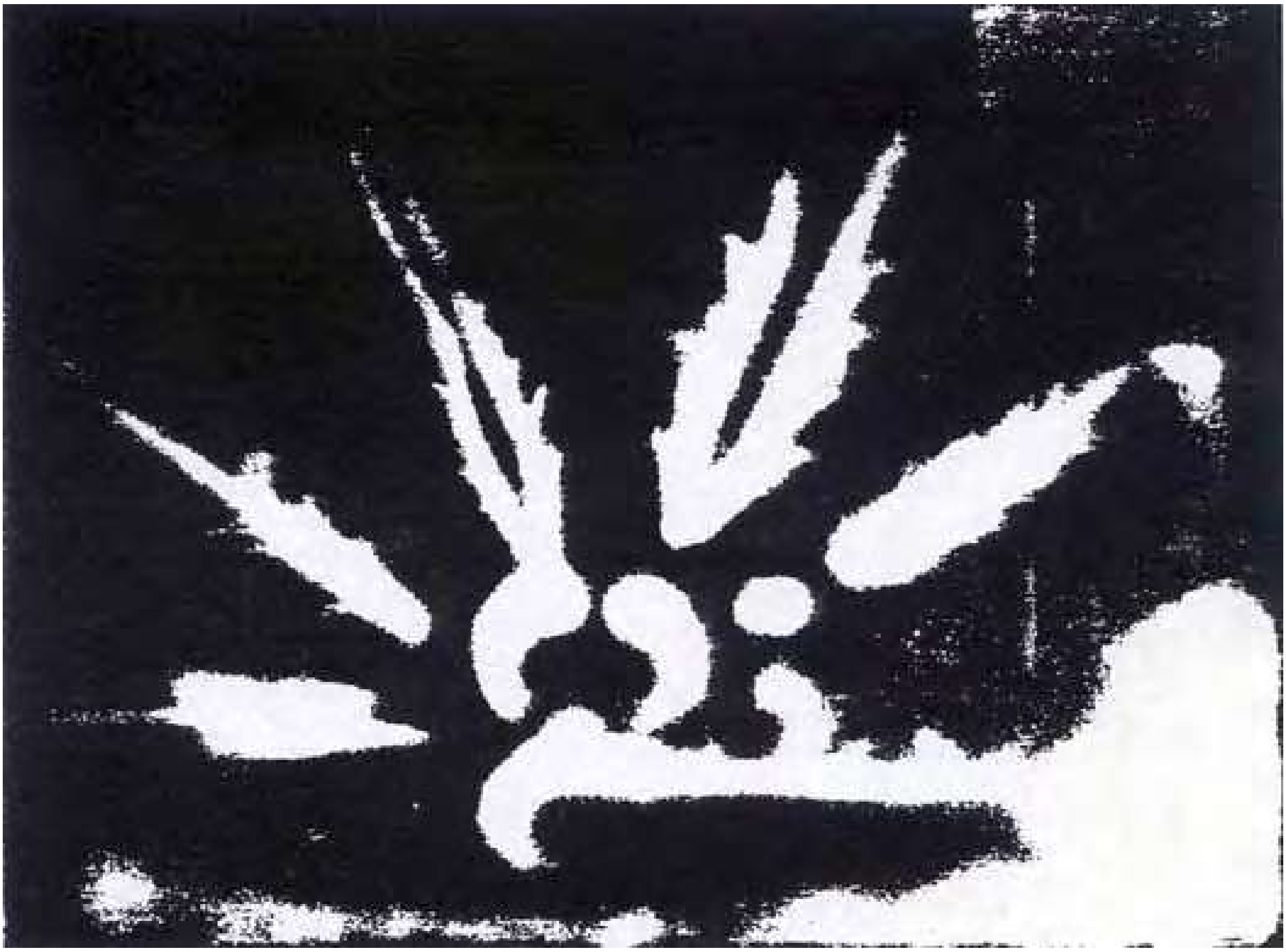

Figure 2. Wilson's second cloud pattern

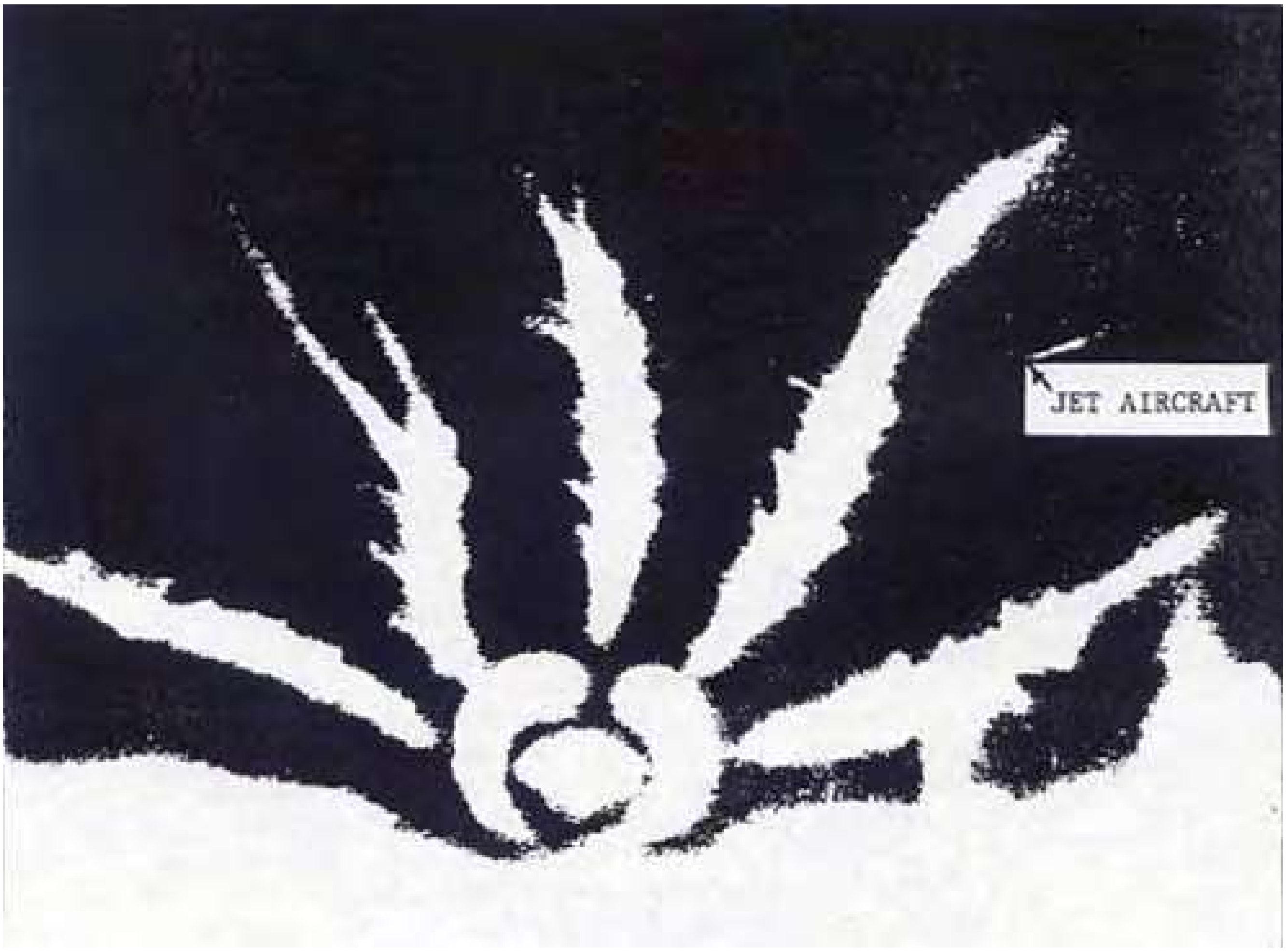
JET AIRCRAFT

Figure 3. Wilson's third cloud radial

Figure 4.

PLACEMENT OF WILSON'S THREE RADIALS

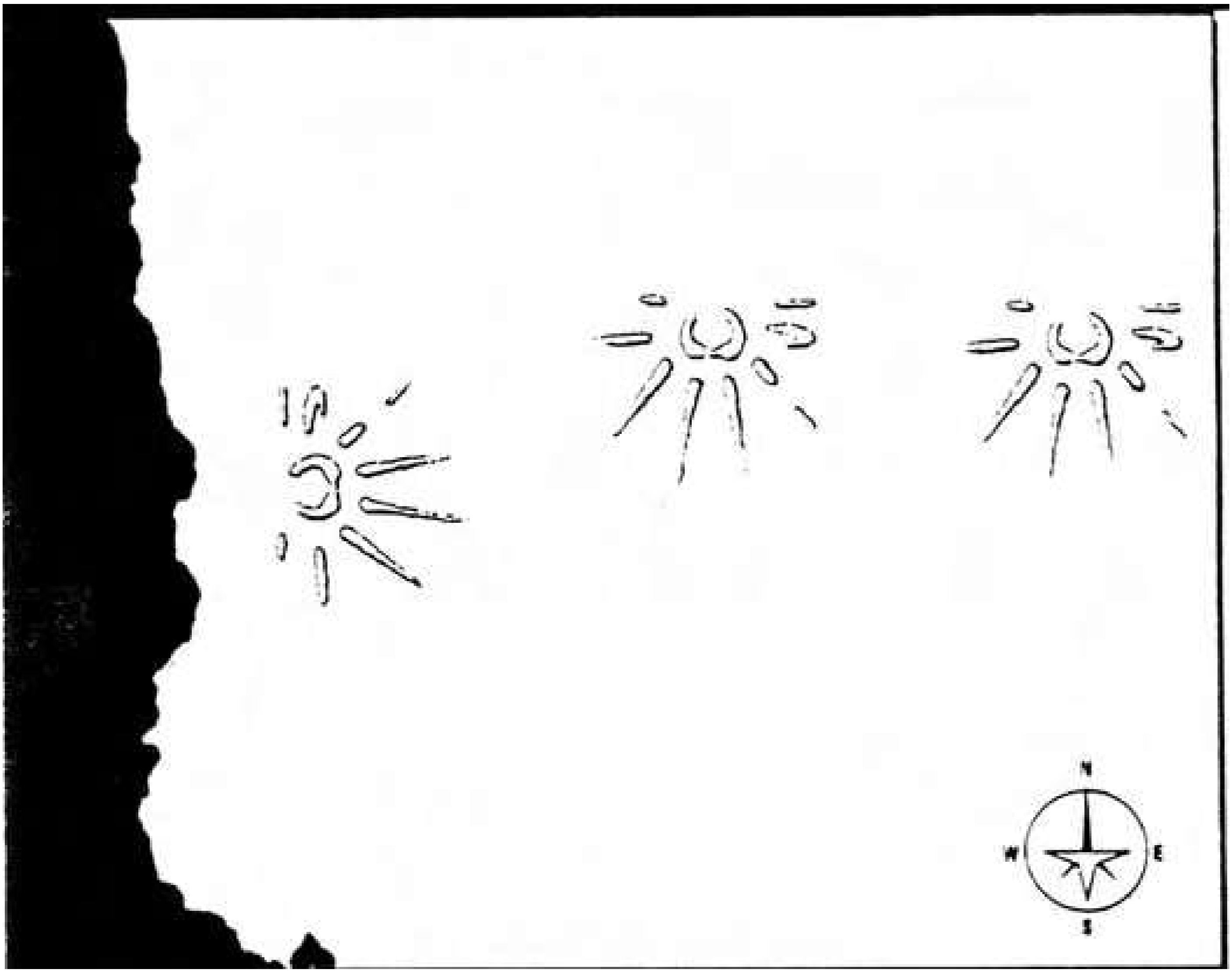

Figure 5.

VIEW FROM WESTLAKE OVER VENTURA COUNTY, CA

1-24-86, 280• WEST/NORTH-WEST

11:45 AM PST

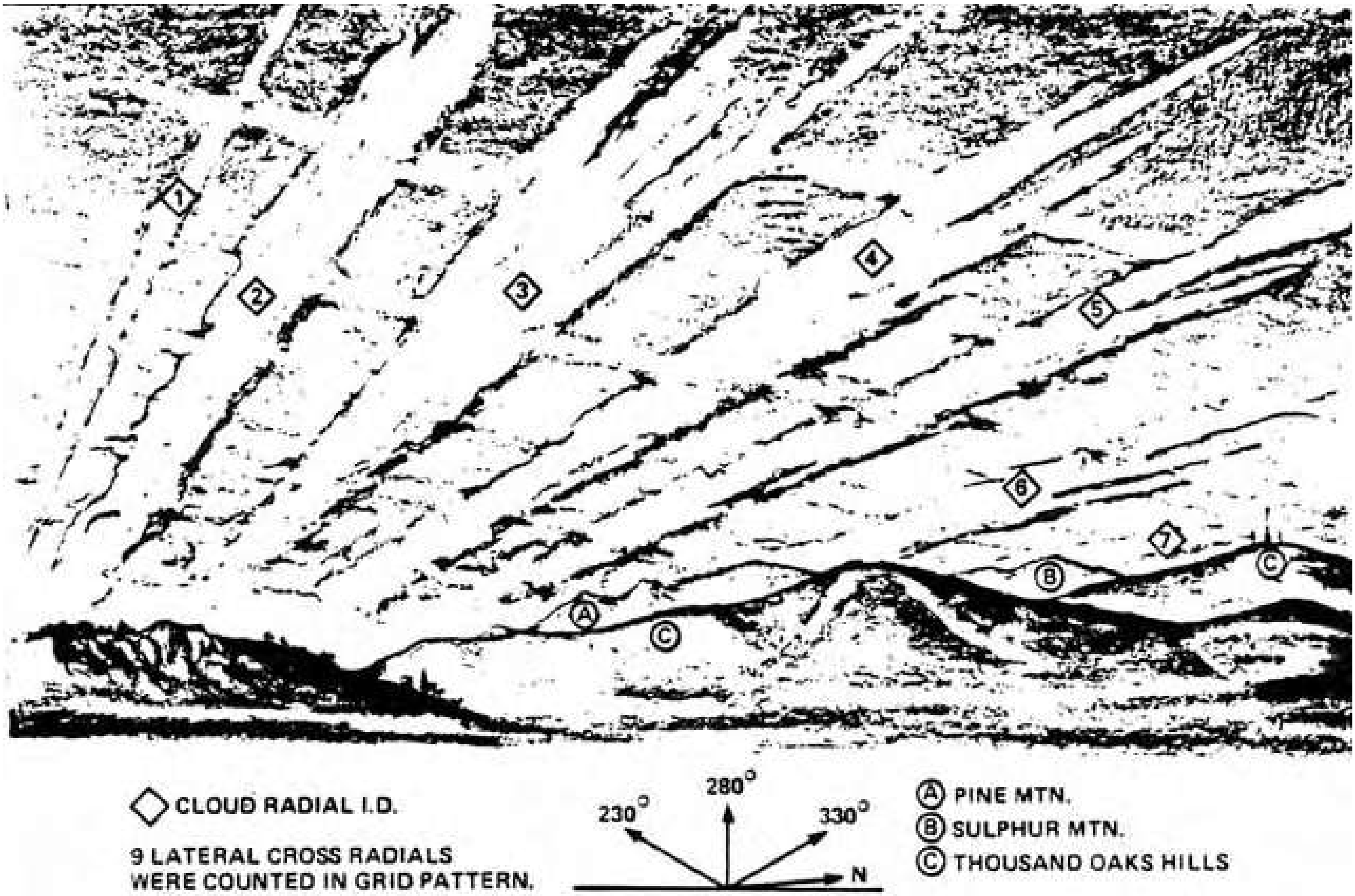
1
2
3
4
5
6
7
A
B
C
C
CLOUD RADIAL I.D.
9 LATERAL CROSS RADIALS
WERE COUNTED IN GRID PATTERN.
280°
230°
330°
N
Ⓐ PINE MTN.
Ⓑ SULPHUR MTN.
Ⓒ THOUSAND OAKS HILLS

Figure 6.

GRID VECTOR APPROXIMATION
1-24-86 VENTURA COUNTY, CA

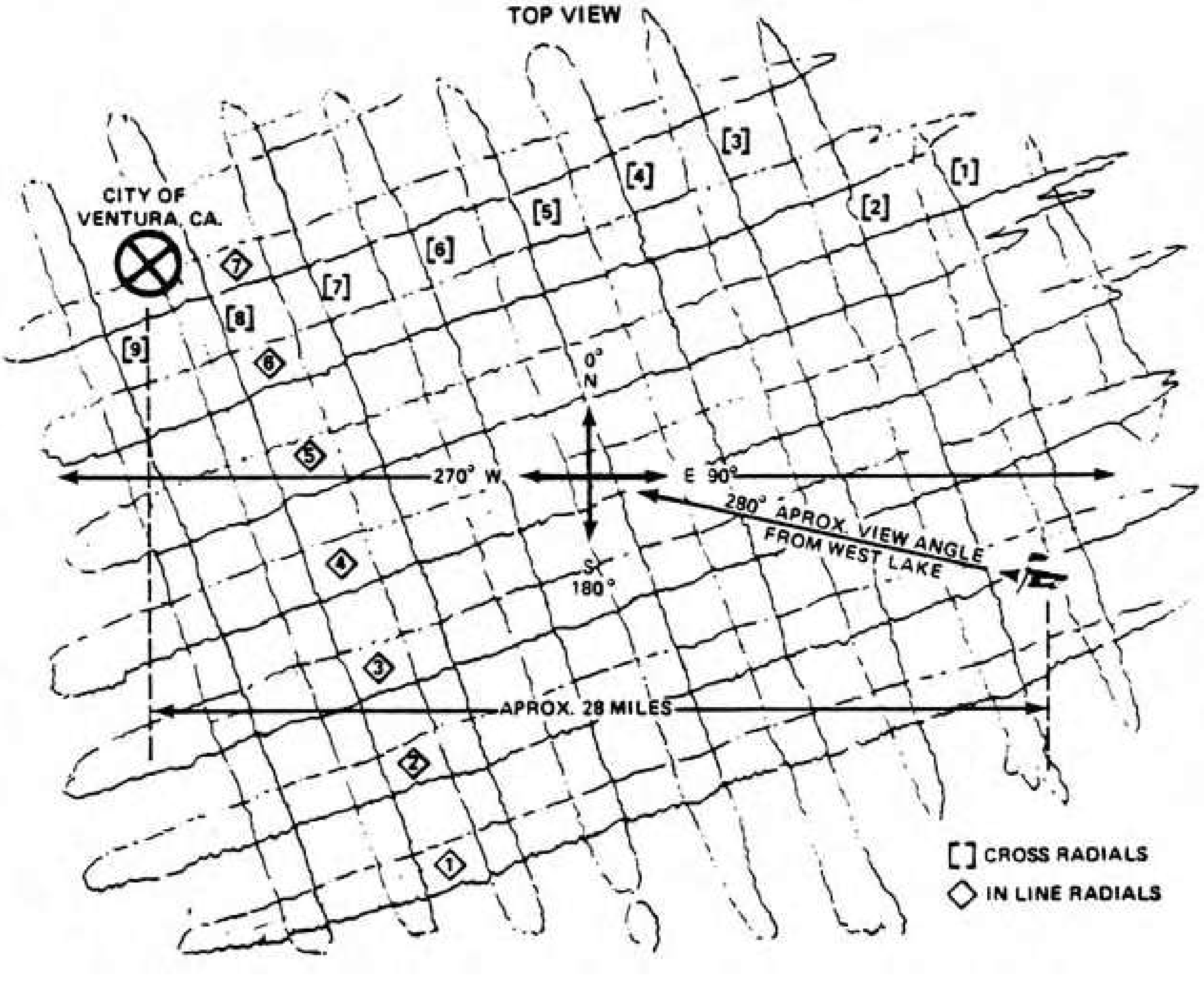
TOP VIEW
CITY OF VENTURA, CA.
[1]
[2]
[3]
[4]
[5]
[6]
[7]
[8]
[9]
⑦
⑥
⑤
④
③
②
①
0° N
270° W.
E 90°
S 180°
280° APROX. VIEW ANGLE FROM WEST LAKE
APROX. 28 MILES
[ ] CROSS RADIALS
◇ IN LINE RADIALS

Figure 7.

SIX-MILE DROP IN TWO MINUTES

CHINA AIRLINES FLIGHT 006
- TAIPEI TO LOS ANGELES -
19 FEB 1985

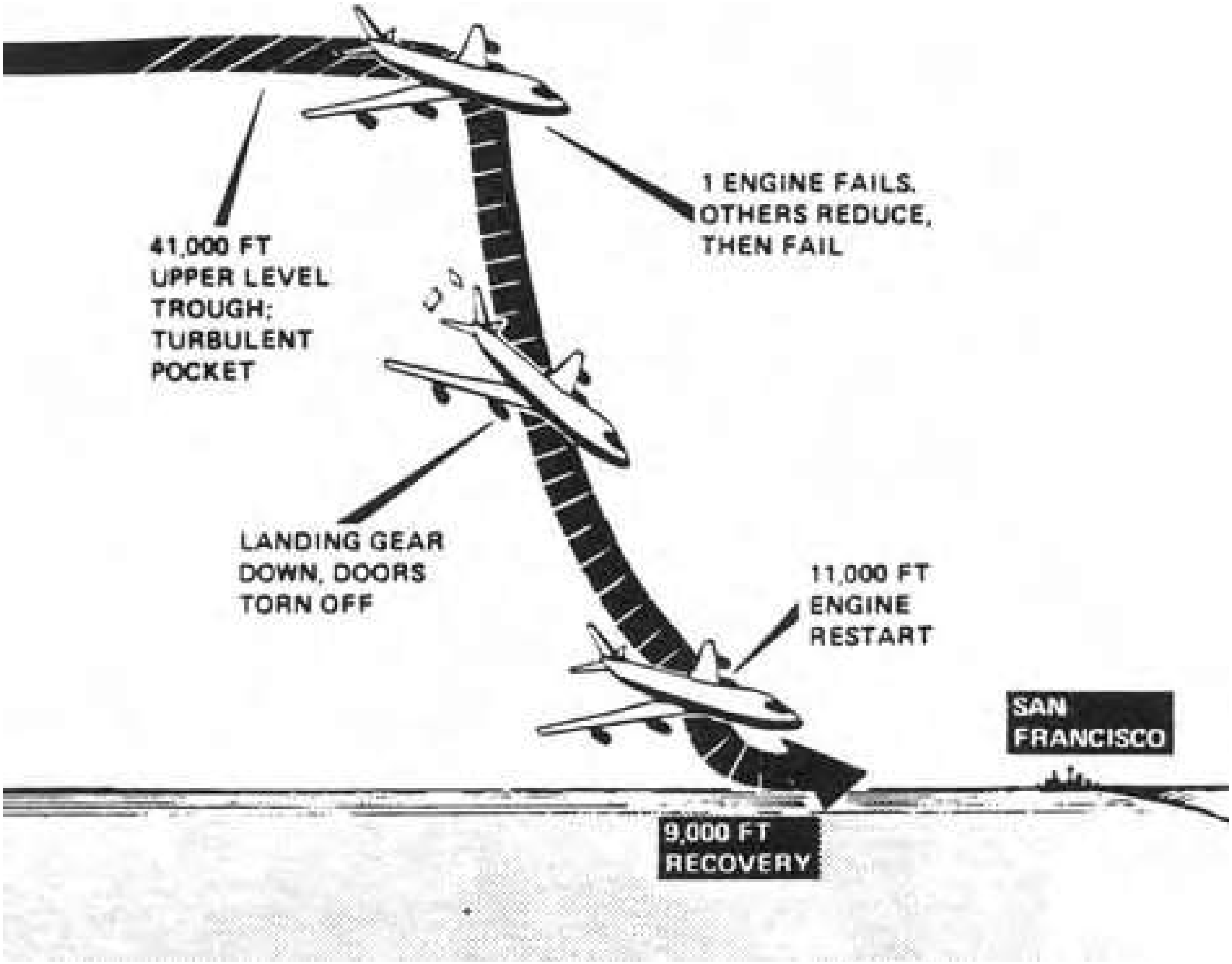
41,000 FT
UPPER LEVEL
TROUGH;
TURBULENT
POCKET
1 ENGINE FAILS,
OTHERS REDUCE,
THEN FAIL
LANDING GEAR
DOWN, DOORS
TORN OFF
11,000 FT
ENGINE
RESTART
SAN
FRANCISCO
9,000 FT
RECOVERY

Return to Fer de Lance Index

The Tom Bearden Website

Fer de Lance

Figure 8.

SAN DIEGO CHAFF INCIDENT, JAN. 1985

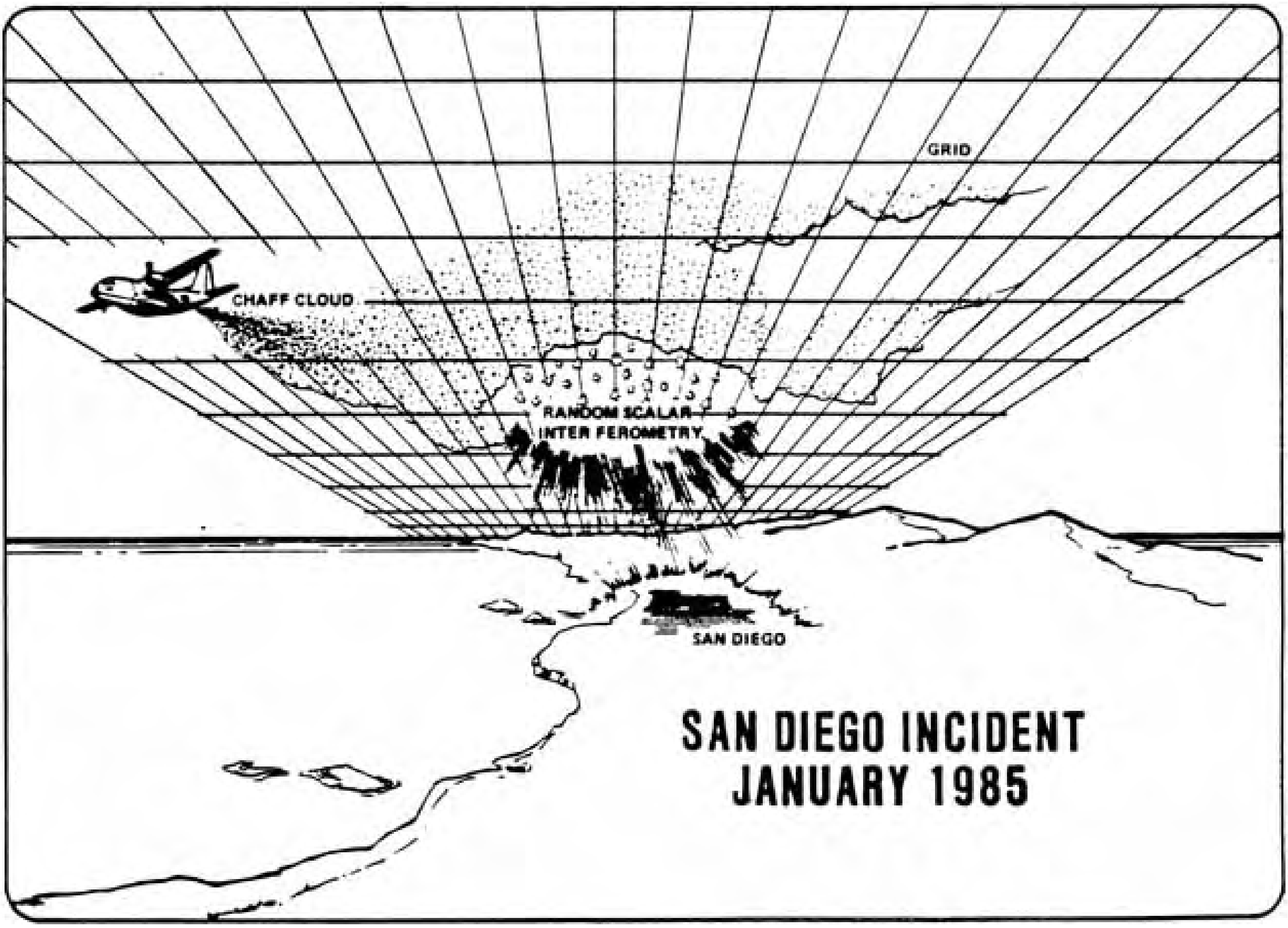
GRID
CHAFF CLOUD
RANDOM SCALAR INTER FEROMETRY
SAN DIEGO
SAN DIEGO INCIDENT
JANUARY 1985

Return to Fer de Lance Index

EM Missile Strike, Offset from Night Shuttle Launch

Cape Canaveral, Florida

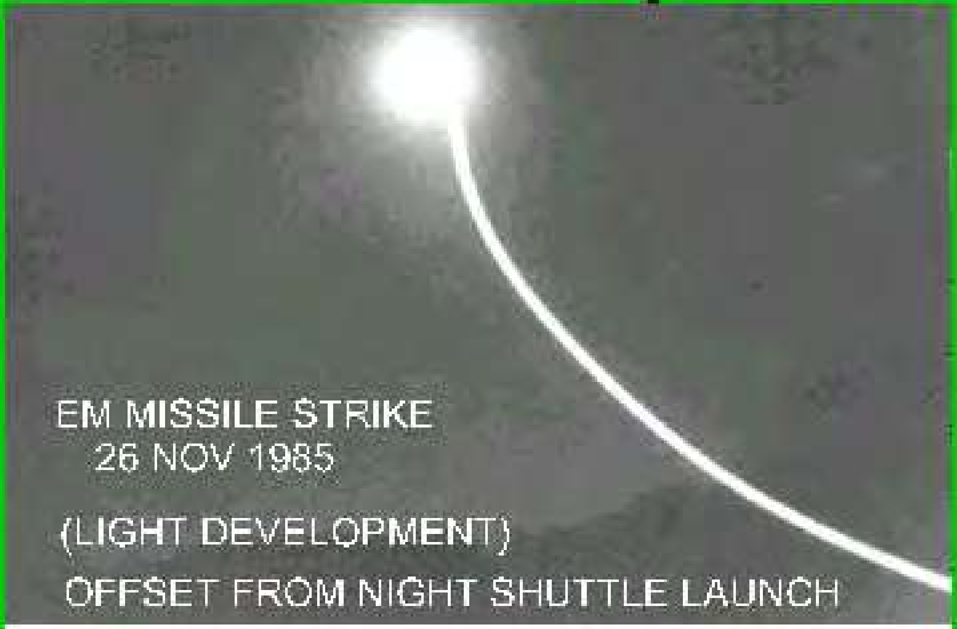

EM MISSILE STRIKE
26 NOV 1985

(LIGHT DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

PHOTO BY BOB GLADWIN

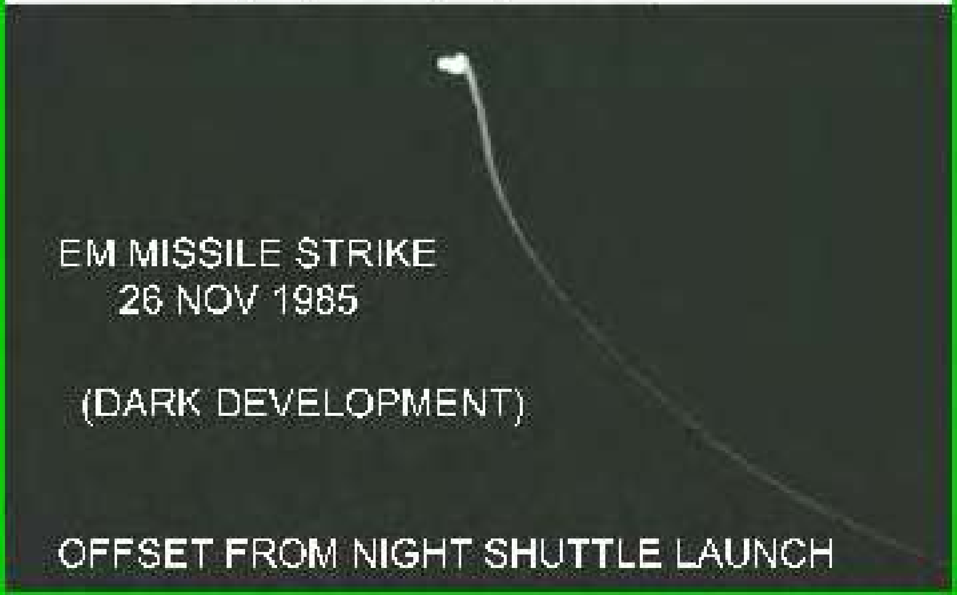

EM MISSILE STRIKE
26 NOV 1985

(DARK DEVELOPMENT)

OFFSET FROM NIGHT SHUTTLE LAUNCH

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

© T.E. BEARDEN 1996

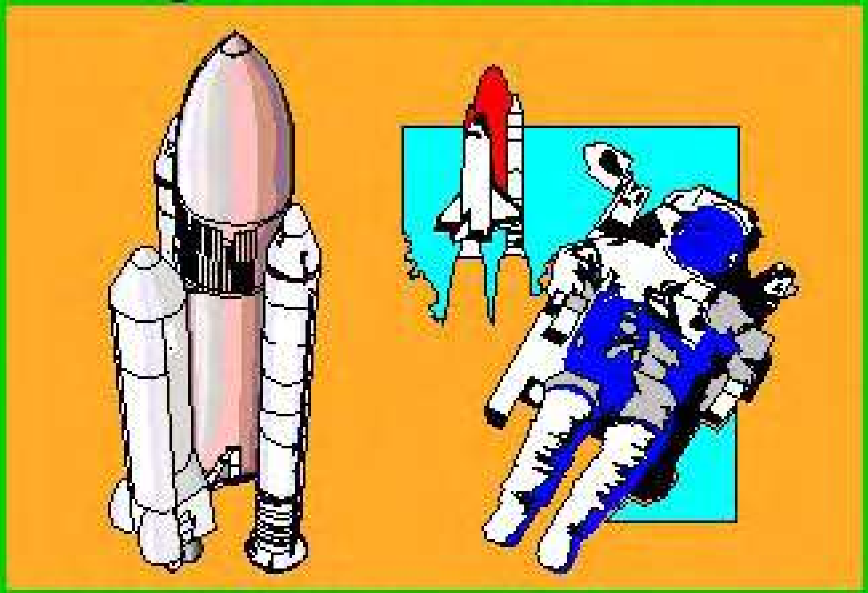

The Smoking Gun on 26 Nov. 1985

Longitudinal EM Wave Interferometry: Exothermic

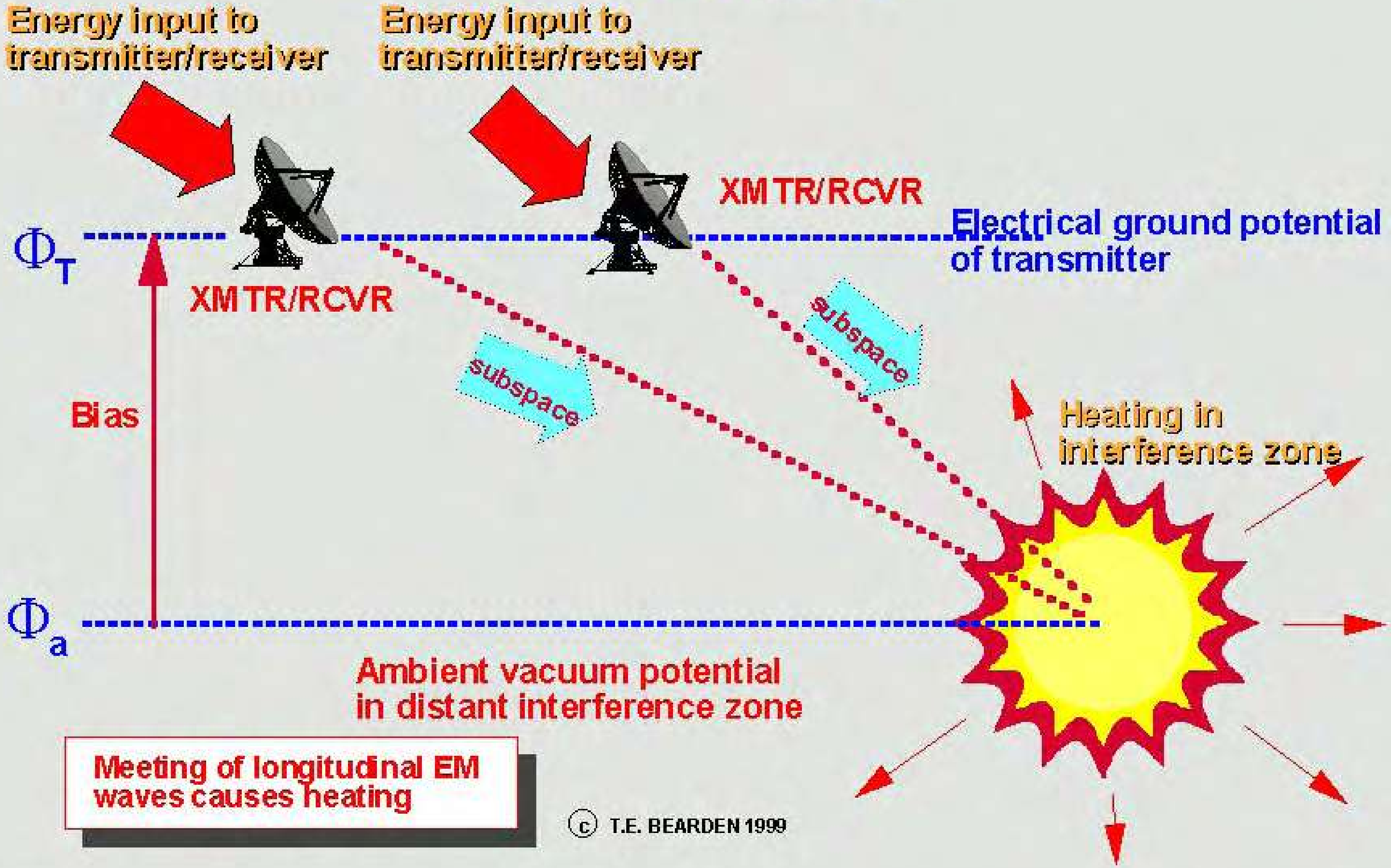

Energy input to transmitter/receiver

Energy input to transmitter/receiver

XMTR/RCVR

XMTR/RCVR

Electrical ground potential of transmitter

Φ_T

Bias

subspace

subspace

Heating in interference zone

Φ_a

Ambient vacuum potential in distant interference zone

Meeting of longitudinal EM waves causes heating

ⓒ T.E. BEARDEN 1999

Longitudinal EM Wave Interferometry: Endothermic

Meeting of longitudinal EM waves causes cooling

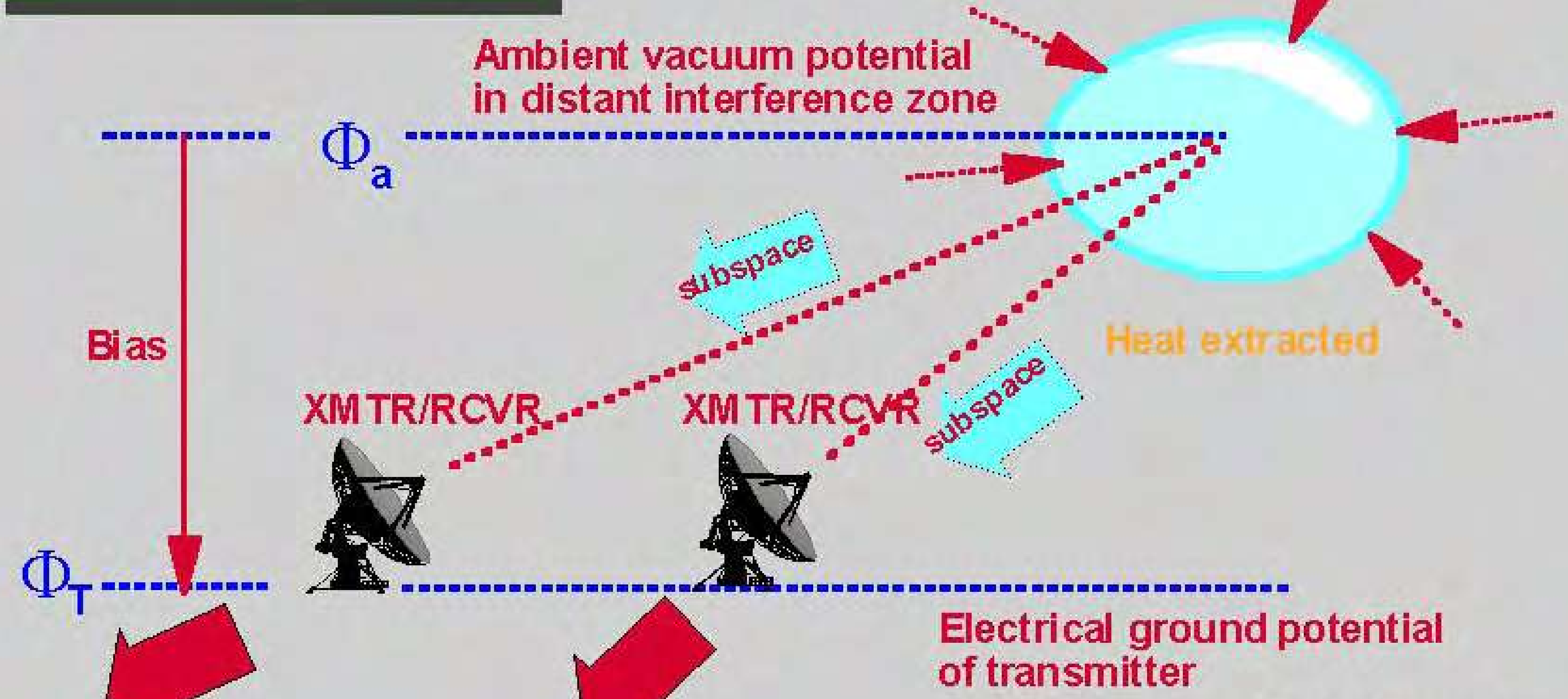

COOLING IN INTERFERENCE ZONE

Ambient vacuum potential in distant interference zone

Φ_a

subspace

Heat extracted

Bias

XMTR/RCVR

XMTR/RCVR

subspace

Φ_T

Electrical ground potential of transmitter

© T.E. BEARDEN 1996

Heat emerges at transmitter/receiver

Heat emerges at transmitter/receiver

- More Frightening Than the Mind of Man Has Ever Imagined -

Brezhnev's unheeded 1975 characterization of these weapons as "more frightening than the mind of ran has ever imagined" has very real justification. In a moment we shall explain what he meant. We shall also explain why the Russians may be serious about substantial nuclear disarmament.

There is now a "balance of terror" loose in the world that makes the old MADmen's MAD concept look like a sunny day in Hawaii.

Let me explain just a little of what I mean.

The scalar EM can and does use an ordinary EM beam as a medium. Many, many scalar EM channels can be infolded into an ordinary carrier. Thus the amount of energy stored in the carrier as "electrogravitational" energy has little to do with the feeble amount of electromagnetic energy stored in the EM carrier signal.

Thus in one sense, when the Soviets are employing the big "continent buster" weapons, literally hundreds of millions -- or even billions -- of watts of energy may be infolded into the seemingly weak carrier beams. All this energy is in the scalar EM channels established in the carriers.

Plus which, the scalar beams are employed in a "standing wave" node. Each scalar standing wave is like a gigantic "capacitor": It is highly charged. Under normal conditions, it has a charge and discharge time constant as well. If something were to "short circuit" one of these huge standing waves -- say, one with a thousand megawatts in it -- all that stored electrogravitational charge would discharge through the short circuit into whatever is on the other end of the short, and into the ground itself. In other words, the earth itself in that location would receive a mighty electrogravitational pulse (EGP). The ground in a region surrounding the discharge point would "charge up" just like a capacitor, and would transmit the EGP as a hemispherical current into the earth and spreading around it. This is where the real problem cores in.

The problem, as originally discovered by Wilhelm Reich [orgone energy was and is real, and is simply one form of scalar electromagnetics], is that radioactive material has a very great affinity for this EGP and undergoes a drastic reaction to it. Crudely, it works like this:

Imagine the vacuum as a virtual particle gas. This gas is under extremely high "pressure" or, more correctly, internal Stress. This stress is a gravitational potential.

However, the stress of the ambient vacuum is normally about the same in most localities. So a chunk of radioactive material doesn't experience a difference in vacuum potential across itself or throughout itself. In other words, the scalar EM fluxes in and out of every part of the nucleus are very high, but of equal magnitudes. The radioactive nucleus, however, is "unbalanced"; that is, it has an extra bit of unresolved EG potential which it would very much like to get rid of or discharge, by spitting out one or more particles. That is, it would like to "decay" to a more stable configuration. A nucleus that is trying to decay to a more stable configuration is a radioactive nucleus.

Remember, the radioactive nucleus is involved in a tremendously dynamic virtual particle flux exchange with the vacuum potential. And a part of it isn't quite in balance in that flux. Sooner or later the right "bobble" in the input flux will come along, and the nucleus will decay and emit a particle and/or radiation. By measuring the level of decay activity in a certain amount of the radioactive material, we can determine a "decay constant," or basically the probability that a nucleus will decay in one second. We can then determine how long it will take that chunk of material to decay away to a stable, relatively inert material. Some materials only require a few minutes to decay, while others may require thousands of years.

Now if you suddenly and drastically change the stress of the vacuum in which a radioactive nucleus is imbedded, you change the potential placed on the unbalanced nucleus. You hit it, in other words, with a strong EG potential pulse -- an EGP. That instantly unbalances the nucleus and it decays explosively. That is, it changes the decay constant and the probability of decay.

Basically that's what happens in a nuclear implosion weapon: You hit the radioactive hunk of material in the center with a very high, imploding pressure wave, which is carrying an EGP. The EGP causes all the nuclei to decay nearly instantly, producing a nuclear explosion.

The bottom line is: strong EGPs wreak havoc if they happen to hit anything radioactive. Such as nuclear warheads, in missiles or in storage. Such as nuclear reactors. Etc.

Now do you begin to see the problem? Let me spell it out.

Suppose the Soviets have activated some of their giant scalar EM transmitters, to place a giant weather engineering grid over the U.S., or to knock down a Challenger or a Titan. They've got their scalar EM beams all charged up. One of the giant Woodpecker transmitters has -- let's say -- 500 megawatts in its infolded scalar EM standing waves. Everything is going good. They're socking it to the Americans and the Europeans.

All those multiple scalar EM frequencies they're using to form the patterns, etc. have to all be absolutely phaselocked. Therefore all the transmitters have to be rigorously phaselocked, too.

And suddenly something goes wrong. A major amplifier/transmitter controlling or stabilizing the phaselocking malfunctions. The phaselock is disturbed, and begins to quiver.

Emergency circuits react instantly, trying to stabilize the situation and restore the quivering phaselock. Little "shorts" are being applied to the circuitry and to the "giant capacitor" that is the scalar EM standing waves. Dumps and discharges of EM energy are beginning to flash in the transmitter and associated circuits -- and into the ground. These make the situation more unstable. Feverishly the emergency circuits struggle to contain this thing -- and Wham! they cave in, all at once.

Now you get a mighty and continuing discharge of the scalar EM energy in that standing wave back into the transmitter and into the ground. It burns out the transmitter. And the discharge continues as the "capacitor" drains all its EG energy into the earth in a giant pulse.

Around the transmitter site, the EGP is transmitted radially away, still very much like a standing scalar wave. It reaches the first nuclear facility. Let's suppose it's a nuclear reactor. Specifically, let's suppose it's a nuclear reactor at Chernobyl.

If you'll check Soviet Military Power, DoD, 1986 you'll find that there's a big ABM radar site not far from Chernobyl. And Canadian sources verify that one of the large Woodpecker transmitters seems to be sitting in that general location. Now check "The Soviet Nuclear Disaster," The Wall Street Journal, Apr. 30, 1986, p. 24. Check the small drawing in the lower right corner of the bigger drawing. Here you can see that a Woodpecker transmitter located near Chernobyl would ideally transmit nearly due west to intersect over the U.S., entering across North America from its east. Other Soviet "over-the-horizon (OTH) radar transmitters" go over the north polar regions, to enter the U.S. from the North. The two then form a giant interference grid over the U.S. (You can see that in Soviet Military Power.).

Suppose you were in the field in the U.S., measuring the Woodpecker signals. Suppose the stronger beam was coming from the east. You would be able to see this with your instruments.

Now suppose the giant transmitter near Chernobyl got into the type of failure we're talking about. The emergency circuits panic and cut down the EM carriers, trying to drain away the standing scalar waves in a controlled fashion. Then your instruments would suddenly record the "new" direction as coming from the north, since the east carrier is gone.

That's exactly what engineer Bill Bise detected out in California, about the time it appears the reactor at Chernobyl began to get in trouble.

Let's continue with our scenario. Suppose the emergency circuits at the discharging Chernobyl transmitter contain things for awhile. The standing scalar waves are steadily draining into the earth. The EG potential starts to rise at the Chernobyl nuclear power plant.

In the nuclear reactors, the decay constant of the uranium in the fuel rods changes. Additional radioactivity begins to pour out of the uranium. The reactor alarms go off, and the crew takes measures to reduce the problem. The control rods go further in, to slow it down.

But the EG charge keeps increasing. The radioactivity of the uranium continues to increase. The control rods go further in to counteract this. The electrical controls begin to act up, showing anomalous indications. The radioactivity increases. In panic, the controls are fought until the control rods go all the way in. The reactor should in fact now be shut down completely, in normal circumstances.

Things settle and hold a bit. The scalar EG charge increases. Finally the "shutdown" reactor starts to increase in radioactivity again. The control rods can't handle it. Everything is heating up, including the graphite moderator. Back at the transmitter, the emergency circuits finally are overwhelmed and give in. They fail all at once. The dreaded "uncontrolled short circuit" now exists between the collapsing standing wave and the earth. A gigantic discharge of EGP into the earth occurs at the transmitter site.

In the desperate reactor, suddenly a great EGP appears. Instantly the uranium goes intensely radioactive. A gigantic pressure of heat, steam, graphite and radioactive materials bursts out of the reactor, breaching the building and erupting violently into the atmosphere. The core melts down. The graphite moderator ignites. You've got the realization of the nuclear engineer's worst fears: core meltdown and reactor out of control. It's afire and spewing deadly radioactivity out by the gobs. And the heat! It's like hell itself opened up for business in this spot.

At any rate, something very similar to that seems to have happened at Chernobyl.

This author believes something similar happened in 1957 in the Southern Urals. (See Zhores A. Medvedev, Nuclear Disaster in the Urals, Translated by George Saunders, W. W. Norton & Co., 1979, p. 150.). In the Kyshtym area, an explosion of nuclear wastes stored underground occurred in the winter of 1957-58. The area for miles around was contaminated, and is still contaminated to this day.

Recall, Khrushchev spoke of a fantastic weapon in development in January 1960. Early prototypes of the weapon would have been in development and testing in 1957. Probably at that time there were no emergency control circuits to handle standing wave collapse, or only primitive circuits at best.

If the failure of a transmitter occurred in full-power experiments, the unrestrained collapse of the standing scalar EM wave would most certainly have produced a large EGP in the ground, and instantly activated the nuclear wastes stored nearby. Eyewitness accounts reportedly stated that the wastes exploded, and that is totally consistent with our thesis. It is in fact a scalar EM accident signature.

Any nuclear facility at all is an incredible and monstrous risk whenever scalar EM weapons are active.

Of course, one's nuclear facilities are at risk if one's enemy has scalar EM weapons and fires at them, as we covered earlier in this book.

However, they are at terrible risk even from one's own scalar EM weapons. The consequences ot a relatively simple electronic failure can be catastrophic; not only for the local nation, but for the earth as a whole.

To show what we mean:

Suppose a similar scalar EM accident -- such as seems possibly to have happened at Chernobyl -- occurs at some other giant Soviet "radar" facility. Further, suppose this particular transmitter is not too many tens of kilometers away from a Soviet missile silo complex. Just to really throw in the grit, suppose these missiles are the "big monsters" with--- say -- a 50 megaton warhead in each missile. Suppose there are 10 of them in the complex.

Let's say this accident occurs "catastrophically" -- that is, the emergency control circuits are overwhelmed immediately and a discharge of all the scalar energy begins at once.

A giant EGP strikes the earth, spreads in all directions, and strikes the waiting monster missiles. In their implosion trigger warheads and in their fissionable and fusionable payload, everything goes at once. Actually, it goes better than when normally triggered. So each one goes at about 100 megatons.

You've just had a monstrous 1,000 megatons of nuclear explosion in the earth. You get a giant earthquake. If it's not too far from the ocean, you get a tidal wave also. You get incredible amounts of lethal dirt and debris, radioactive and deadly, hurled violently into the air.

This is just about the equivalent of a "normal all-out nuclear war!"

And what about the other side's detection and early warning systems? Who's going to alert whom, and for what? How many missiles get launched?

I think you see the point.

And I haven't even covered what a giant EGP can do to humans. If it's truly a catastrophic failure, you can get an EGP that can wipe out half a mighty nation in one burst!

And if the discharge happens to tickle the Sun and Moon's feedback loops the wrong way, you'll get convulsions of the earth, mighty burps of the sun raining fire and brimstone on the earth, and a violent increase in the interior heat of the earth's molten core, with a concomitant eruption of that core right up though the mantle.

In other words, it could be "back to the beginning" for the earth, and start all over -- if, that is, some semblance of an earth survived.

The Soviets know full well the implications of their giant scalar EM weapons. (Remember, I estimate they have over a hundred of these monsters, based on Frank Golden's detection of their full-up exercise in April/May 1985.).

Scalar EM weapons have added a totally new dimension to the meaning of the word "terror". If one develops them and his adversary does not, nothing on earth can stand against them. So in Man's present state, the temptation to develop them is nearly irresistible.

On the other hand, if one's adversary already has developed them, one is already totally defeated unless one also develops them. Again, the need to develop them is imperative.

Yet, whenever one activates a large scalar EM weapon, one immediately places the entire earth in deadly peril. The slightest misstep, and it's curtains for everyone. And it's curtains for the earth as well.

So one is damned if one does, and damned if one does not.

We have reached the final turning point in human history. The clock is ticking away the last years -- or days -- of the lives of all humans on earth. There has never been such a threat.

We simply have come to the crossroads. somehow, all Mankind has to change.

Infatuated with the thirst for ultimate power, we are facing two doors. Behind one is a beautiful lady, and the fulfillment of all man's dreams: control of physical reality itself. Behind the other is the deadliest tiger ever dreamed of in our wildest nightmares.

Which is it to be, the lady or the tiger?

SCALAR EM INOCULATION AND ITS IMPLICATIONS

My appreciation goes to Gary North for his openly sharing the above important information with the public. Human nature being what it is, one or more of his more dire scenarios is going to happen sooner or later.

The problem becomes enormously more compounded if the hostile agent or agents possesses sophisticated scalar EM technology.

With scalar technology, there exists the potential that the worm or virus can be implanted from a distance. If scalar scanning devices of sufficient precision can be developed, then major software residing in a chip, a computer, or on a disk can be read from a distance. With scalar EM methods, the ordinary software "guards" are totally ineffective. The scalar "reader" could simply read off the software directly as the disk or device goes about its normal operations.

After the software is read, a germ or a virus is designed to produce a particular problem as desired and upon special stimulus. Then a scalar "imprinter" device is used to produce the necessary digital pulses directly upon the media in which the software resides. The required code has now been .implanted. In effect, the operating system has been injected from a distance, and contaminated with the desired "sleeping agent" -- germ or virus. From there on, it's just like Gary North's scenarios.

So the scalar EM technology can probably be used (at least eventually) to place "mines" in systems in advance, for damage or destruction at a future date. At least it's possible in principle.

However, it's not the injection of computers and equipment that bothers me. It's the injection of human beings. Here's generally how it can be done (and possibly is being done).

Basically, human brains have various rhythmic waves which represent "envelope pulsations" of the entire electrical system. However, it is not the EM force fields represented by these pulsations which are important: it is the summed-zero (scalar) patterning of the entire complex of neuron firings which represents the real activity of the brain -- and represents it thoughts and emotions. A storage system for these patternings exists, but it's a scalar pattern storage, not an E and B field storage. The storage system is in the nuclei of the tissues of the nervous system itself, not in electrical or electrochemical potentials.

At any rate, the human brain rhythms are also "riding along" relatively entrained to the Schumann resonance of the earth's magnetic field. So all the patterning of the brain is grouped together into giant overall orders, and one or more of these is normally synchronized to our natural environment.

If a varying magnetic signal of extremely low frequency -- say 10 hertz -- bombards the brain with greater power than the earth's magnetic field, it is possible for the brain's rhythm to be "forcibly entrained" to synchronize with that signal, rather than with the normal Schumann resonance. If a large number of brains are forcibly entrained to such a signal, and the signal is artificial, then a master coherent phasing of all the brains to the transmitted signal has been accomplished.

And that can be utilized -- and is utilized -- by the Soviet Union. On the Woodpecker signals, Bob Beck has detected up to 16 carriers, each with a 10 Hertz modulation, and with all the 10 Hz modulations in phase.

Let's stress what we're saying here. At times, the Woodpecker signals have shown 16 carrier frequencies which contain the same 10 Hertz modulation, and which are all kept coherently sync-locked.

If those signals and the modulations are strong enough, the brains in a targeted area demodulate the 10 Hertz signal and phase-lock to it -- actually, to them. At that point, the brains are entrained to a 10-Hertz driver, and 16 channels of multimegahertz frequencies are locked-in to the brain.

Now if the Soviets have broken the genetic code of the human brain (which they have) and succeeded in developing the ability to implant thoughts, emotions, pictures, etc. (which at least to a degree they have), then they can simply record the proper signals from human guinea pigs, using scalar technology and instrumentation. They can tailor those signals and add them onto the Woodpecker carriers, using scalar EM techniques. By using most of the sixteen carriers as a means to express Fourier frequency expansions, specific areas and functions of the brain can be targeted.

Notice that the distant transmitters al ready "know" the correct phasing; just use the 10 Hertz modulation signal as a reference. In that case, phasing takes care of itself inside the captured human brains.

The brain has a unique characteristic. If a signal arises inside it, it interprets that it itself originated the signal. Thus for the first time in history , true unwitting "robot" actions and emotions can be engendered in masses of human beings.

But it's still worse. The Kaznacheyev experiments (see references) proved conclusively that electromagnetic signals can carry and transmit cellular disease and death patterns into a targeted biosystem. You can rest assured that the Soviets continued on to actually measure and determine a wide variety of death and disease scalar patterns. All of that can be added on to the Woodpecker carriers so that the population in the intersection grid area is devastated. One can even test it out in advance by producing a modified pattern of a disease in the targeted populace, and keeping the percentage of brains entrained down to a small level. This lowers the number of persons "injected" from a distance to a small percentage of the populace. Then the outbreak of the "new strain" of a familiar disease -- such as influenza, pneumonias, and something like AIDS -- can be monitored to provide targeting and effectiveness data. Notice that a series of diseases was experienced by U.S. personnel assigned to the U.S. Embassy in Moscow, probably as a result of the so-called "weak microwave radiation" of the Embassy by the Soviet Union.

But that's not all. Much more subtle effects can be produced in the targeted populace, without anyone getting wise. Cancers and leukemias should be relatively easy to induce. Also, the immune system can simply be attacked directly. If done slowly, a series of common, ordinary diseases will just seem to get "tougher and tougher to cure." Actually it's not the disease that got tougher, it is the immune systems that got weaker. It is my opinion that precisely that sort of testing -- against the immune system -- has been accomplished by the Soviet Union against the U.S. since about October, 1985. A great percentage of influenza patients during the period from then until now have had complications. So much so that multiple sessions -- as many as 3 to 5 -- of antibiotics are necessary before the patient recovers. At least in the Huntsville, Alabama "hinge pin" area, that seems to have been the case. This author personally experienced this, having influenza and follow-on complications for a period of seven months -- something quite different indeed for him.

But it's even worse.

The human scalar EM nervous system ("inside" the ordinary EM signals and functioning of the system) also has its own resident "software." It has special storage media contained in its brain. its cells, and its genetic material, and its actual nucleons. All of these storage media are continually telecommunicated with via scalar EM means. The "stability" of what each of us is, is largely conditioned by those recordings. As personalities we literally "are" what we have experienced, felt, thought, and did. Our whole "functioning conscious system" has been evolved by our experience, and it continues to be "us" by means of continual reference to the record base.

I hope you now realize that the actual record base itself can conceivably be altered by scalar electromagnetic means, even from a distance. Even by the hidden scalar EM content of 16

Woodpecker carriers, locked into a captured brain and pouring in their deadly content. It is conceivable that even what one is can be taken from him, and what someone decides he will be, he will be.

A totally new -- and ultimate -- form of "slavery" has emerged.

Also, look at all this scalar EM functioning of the human system as a special kind of cybernetics. A scalar EM computing software and hardware system.

What if scalar techniques are used to plant "worms" and "viruses" in this system? With hidden triggers? What about a series of them, so that a range of "eruptive software diseases" can be stimulated at will?

And guess what. Such scalar EM "worms" and "viruses" can even be induced by contact with or consumption of "contaminated" water and food. Or saliva. Or other body fluid.

So you can see the possibilities of "seeding" a population with a staggering variety of time-bombs.

I think you begin to see the point. Now you can change North's "Soviet blackmail" scenario drastically. In a new scenario, Gorbachev simply announces to Reagan that a graduated series of specific sweeping changes are going to happen in our population -- changes which we are absolutely powerless to stop. The last stages are various deaths. The master "trigger" for this series has already been initiated. Incidentally, the Soviet populace has been carefully "seeded" with the proper antidotes. The Russian leader demands certain things (very similar to what North's scenario utilized). Unless his demands are met, a "killer virus" will be activated by a special signal. If so, everyone dies. The first step change in the series will start to occur in a few hours. And it does.

In fact, this scenario could be added to North's scenario, and both used simultaneously. What do you think the President and the Congress would do? Or could do?

There are other even more diabolical possibilities, but these should be quite sufficient to illustrate the point.

Pandora's box has already been spilled, and the end of humanity is ticking away like a time-bomb. It's already nearly midnight, and the watchman hasn't even sounded the alarm yet.

I certainly can anticipate the next question: Have the Soviets already "seeded" the populations of the U.S. and other selected western nations with scalar EM viruses and germs"?

The answer is, I simply don't know. And just like Gary North, I don't know what if anything can be done about it if they have. This researcher certainly doesn't have all the answers. An immediate national crash program to develop scalar EM defenses is essential if we are not simply to be crushed beneath the hammer and sickle. It takes funding, people, facilities, and time.

In some 22 years of unrelenting struggle to convince the system we've got a terrible problem, I've hardly been able to move it off dead-center (no pun intended). This book, produced under extreme workload stress, is all I can do to raise the alarm. I've given it my very best shot.

What happens now remains to be seen. Other hands, hearts and minds are going to have to take up the task and carry it forward. This researcher has reached his limit, and has no more left to give.

1986 APPENDIX II

COMPUTER WORMS AND VIRUSES, AND THE IMPACT OF SCALAR EM

Source: - Gary North, "Electronic Aids (Parts 1 and 2)," Remnant Review, 12(20), Nov. 1, 1985, p. 1-8; 12(20), Nov. 15, 1985, p. 1-8. POB 8204, Fort Worth, TX 76124. Phone (817) 595-2691.

Annotation: - Two deliberately un-copyrighted issues to alert the public to software worms and viruses, and their implications for drastic interruption of society. Describes a hypothetical scenario in which drastic scrambling and destruction of banks' computer tapes, files, and records occur due to software worms and viruses. Written in first person by Gary North. [Slightly edited.]

Article: - "Worms." Earlier this year, I read a very interesting article on a major problem facing computer software (programs) development companies. A program comes on one or more 5.25-inch plastic discs. It takes only a few seconds to copy a program on one disc to a blank disc, which costs $3. Yet these programs normally run at least $250, and usually sell at $495, and sometimes cost thousands. Very few are less than $100. So you have a major temptation: make a $500 asset out of a S3 asset. Insert the $500 program into drive A, write "COPY A:*.*.B:" and hit the "enter key;" sixty seconds later, you have a $500 program in drive B.

There are ways to make this copying more difficult. The companies code the programs, and force you to have a control disc in drive A at all times. These "copy protected" programs are a hassle for users. We cannot put them on "hard (big) disc" easily, and sometimes the control disc dies for some reason. Then what? Your data are locked in your hard disc or on a floppy disc, but you can't get to the data because the control disc is not functioning. You order a replacement. Weeks go by.

Last year, several firms came up with a solution. It is called a worm. A worm is a command that is built deep into the complex code that creates the program itself. These are incredibly complex codes, and it is easy to bury a command in them. They cannot be traced.

What does the worm do? It "eats" things. Say that you are a software thief. You make a copy of a non-copy-protected disc, either to use on a second computer, or to give (or sell) to a friend. The program works just fine. But when the program is copied to a new disc, the worm is "awakened." It bides its time, maybe for many months, maybe for years. The program's user is blissfully unaware that a monster lurks inside his pirated program. He continues to enter data, make correlations, etc. He becomes completely dependent on the program.

Then, without warning, the worm strikes. Whole sections of the data disappear. Maybe the data storage disc is erased. Maybe it is just scrambled. Even his backup data discs have worms in them. Everything he entered on those discs is gone. Forever.

Can you imagine the consternation of the user? He has become dependent on a booby-trapped program. His business could simply disappear. For the savings of $500 (stolen program), he could lose everything he has.

Several firms threatened to insert worms into their programs. But then they backed off. They are afraid that lawsuits initiated against them might go against them in court. They could be hit for damages suffered by the thieving victims. Juries might decide that the punishment, a bankruptcy) was too much for the crime (a $500 theft).

So far, no worms are lurking in any commercial software programs — as far as I know and the industry knows, anyway. But what if a disgruntled programmer was to hide one in a master copy of, say, Lotus 1-2-3, the most popular business program on the market? What if ten thousand copies a month go out for, say, three years? Then, without warning, every company that has started using them loses three years of data? They sue Lotus. Lotus goes bankrupt paying lawyers. No company in the industry is willing to talk about this sabotage threat publicly. Obviously.

LARCENISTS

I just happened to stumble across an article on worms in a computer magazine. It occurred to me that it might be possible to use the worm technique as a form of deliberate sabotage rather than just as a copy protection device. But what did I know? I'm not a computer expert.

I know a computer expert, however. I mean, a real expert — one of those people you occasionally read about. In the world of business, they're called "space cadets." They operate somewhere in between the asteroid belt and Jupiter. But this one is different. He's a businessman, too.

I got him to sit down with me to discuss the problem of worms. It turned out that he has a real fascination for the topic. He tells me that there are advanced design worms, called 'viruses' by 'hackers' — computer freak programming geniuses. "The software virus is the most terrifying thing I've ever come across," he told me. And then he showed me why. My initial scenario is based on only a portion of his estimation of the threat. It gets a lot worse.

He gave me a 90-minute FIRESTORM CHAT interview. He must remain anonymous. He used to be a software developer for programs that were used in the U.S. banking system, but is now employed in a highly sensitive job in a related industry. Therein lies his problem. If he were to tell the story of what he is capable of doing to these banks, his firm might lose a lot of sales. He can't "go public." Let's call him Tom.

Let me summarize briefly some of the details he gave to me. They floored me. They're going to floor you.

(1) Jackpotting. The rush is on in the banking world to get automated teller machines (ATM's) into shopping malls, supermarkets, and in front of every bank. We've all seen them. Just walk up, punch in your card number, ask for cash, and you get it.

In a busy location, one of these machines can hold as much as $250,000 in cash, mostly small bills. These machines are controlled by computer. They are hooked up to the bank's computer system, usually by phone lines. This local line, Tom tells me, is what computer freaks call THE LOOP. The loop is wide open to tampering. He says that what computer thieves are doing is to hook up a cheap Apple II computer, tie into the phone lines, break into the ATM, and get it to empty itself. This is "jackpotting."

He tells me that ATM thieves are hitting banks continually, but nothing is getting to the press. The banks have yet to show a profit with the ATM's so far, which is understandable. They are hoping to get their machines placed in key locations, so "market share" is crucial to their plans. They are suffering horrendous losses in the short run in the hope that long-run profits will pay off, if and when a defense is developed.

The banks are saying nothing because of their fear that if the extent of the losses gets into the press, they will be forced by pressure from depositors — bank runs — to cancel the ATM's. The losses are horrendous, he says. At present, there is no known defense, given the communications technology.

(2) Rounding Off. This is the "preferred" computer bank theft system. Someone on the inside who has access to the software takes advantage of the banks' need to round off numbers. The programs carry numbers out to 13 places. Banks can't use all that space. So when they balance the books (interest rates at, say, 9.873), they just don't count every tenth of a cent. The program is assumed to round off the numbers randomly. What does the bank care? But the thief has set up bank accounts that absorb those random tenths or hundredths of a cent. In millions of dollars worth of transactions (federal funds, etc.), programmers in some cases have stashed away hundreds of thousands of dollars — maybe millions — over a few years. No one knows how much of this goes on.

How could a bank spot this? The books would always balance to the penny. How would the accountants ever know?

I think of a story that Adam Osborne tells in his paperback book, Running Wild. The president of a large firm was looking out his window one day, and he noticed two Rolls Royce cars parked next to each other. He inquired as to the owners. They were two men in the data processing department. He called in investigators, and the cars and the men disappeared. They fled to Brazil and took their cars with them; **Brazil** has no extradition treaty with the U.S. Years later, as

Osborne was writing the story, the firm still hadn't figured out what they had done.

ARSONISTS

These are the fearful ones, far more than the larcenists. These are the practical jokers who get into a major data bank and trash things. It's a kind of multimillion-dollar "Kilroy was here" graffitti.

How easy is it to get in? Incredibly easy. The boy in "War Games" really could have broken into most firms' telephone-connected computers. Computer programs exist that allow the user to hook up his computer to a phone line and randomly dial numbers until they hear the telltale whine of a computer line. It then notes the phone number and goes on its way, searching out more lines.

They can do it by long distance, free of charge. The telephone company has a tough time tracing those who use various sorts of electronic black boxes to call anywhere on earth at no charge. Some people get caught, of course. "The tip of the iceberg," says Tom.

How do they get in? Easy: few systems are protected, once you locate the line. If one is, he says, you create a deliberate error. Most programs then collapse the protective shell, and the hacker finds himself inside the heart of the system. Tom has designed a program that keeps this from happening to his company's programs, but few companies have anything like it.

It's very easy to get in if someone has "logged on," — opened his terminal's connection to the main computer — if the system is connected to phone lines. Or anyone in the company can just tap in, if someone has left his desk and left the computer on. It's common to forget and leave an open terminal.

He showed me. He says anyone can get fired for leaving a computer on. He demonstrated his point. With 40 computers on line, he ran a quick search and found two of them "logged on," despite the fact that it was after hours. All the security in the world can't do anything if a computer line is open. It's like a burglar alarm; it's worthless if you leave the door unlocked or leave the keys lying around. That janitor you hired. Is he a computer illiterate? Or a plant?

Once inside, what can you do? Steal a fortune? Yes, if you really know the system. He told me he could easily steal $3 million from a local bank, even as an outsider. He would then offer to give it back and keep his mouth shut about how easy it was if the bank would pay him 10% the take. He thinks most banks would capitulate for fear of the publicity. In any case, he knows that he probably wouldn't get caught.

How about creating a new identity? The grade-changing scene in "War Games" is true. You could even create a new identity, give yourself high grades in any academic discipline, just by breaking into a university's database. There is very little security here, he says.

But for sheer vindictiveness, for sheer envy, consider the possibilities of a virus-implanter. He gets inside the computer for a major communications link: telephones, large information database, bank wire transfer, or whatever. Then he lays the egg: a tiny, untraceable brief instruction. Inside a huge database are just a few characters. These float inside a system, seeking to devour certain kinds of data, or executing certain routines.

There is a game played by computer freaks called "Core War." They try to implant these killer messages, which seek out each other and battle one another. If you find one morning that yours has been consumed, you lost the battle. That was probably the origin of worms and viruses.

TERRORISM

Say that a revolutionary terrorist group, or some anti-zionist group gets a "ringer" into a system. He might be a computer genius type. Everyone knows they are either oriental, dark-skinned people with accents, or teenagers. The firms don't hire teenagers, but they hire a lot of foreigners. They may even check the guy's credentials. Electronic credentials. (Ha!) Then they turn the guy loose in the system.

The virus is implanted deep inside the system. It can then be transferred to any other bank's computer by means of EFT (electronic funds transfer). Maybe it is triggered when someone with a peculiar name and address opens a bank account. Three days later: Bam! The data disappear. They haul out the back-up tapes. Bam! The virus is on them, too. It is a process of infection, contamination, and incubation. There is no known defense. Not yet. This is the bottom line.

ANTIBODIES

The designer of a virus can also design an "antibody." The antibody is a counter-virus agent that seeks out a virus and destroys it. But like other antibodies, it must be specific. The only way today that an antibody system can be created is to know what kind of a virus is involved beforehand.

Tom says that people are now selling antibodies at very high prices. Who is paying? Big companies that suspect that there is a virus present in their computers. In all probability, the guy selling the antibody created and injected the virus. But now can any businessman prove it? So he pays the blackmail.

NATIONAL DEFENSE

A Soviet agent or American spy working for the Soviets penetrates any of a dozen computers used by the military. He plants a virus. The computers talk to each other, and the virus spreads to all of them. It tells them to execute a certain routine when a certain command is entered at a missile-controlling terminal. That command might interfere with a routine that activates a missile or launches it. Upon reading that command, the virus shuts down the computer, or scrambles the executing program, or scrambles the data. No more "launch on warning." No more launch at all. Dead metal.

Scenario: The President of the United States receives a telephone call on the "red phone" — the direct link to Moscow. He lifts the receiver and says, "Hello."

"Mr. President, this is Michael Gorbachev. You must recognize my voice. I have very little time. I will come directly to the point. You have refused to back down on your threat to implement your Strategic Defense Initiative. You intend to go ahead with space-based weapons. My military staff informs me that they think that the United States has the technology to implement it, and that it could place my nation's military strategy in jeopardy. We cannot allow you to do this.

"If we allow you to deploy the SDI, it will be too late for us to respond effectively. Therefore, we are taking the initiative today. I issued orders this morning to put Soviet military units on immediate alert. We are abiding by your biblical rule to announce the initiation of hostilities before striking. Neither the Japanese nor the Germans gave us this courtesy. If you do not come to terms with us, we will launch a first strike against your nation in three hours. We will delay for one day, if you agree to follow a precise procedure that I will outline shortly.

"At one time we feared nuclear retaliation. We no longer do. Within two hours, you will know why not. I suggest that you instruct your ballistic missile team to prepare your missiles for a strike. Then, to prove to yourself that we no longer are concerned about retaliation, launch one or two of them. As far as I am concerned, launch all of them. But please instruct your senior military commanders to report back to you concerning the effects of their instruction. I suggest that you try launching three or four as a test. We don't care which ones.

"Mr. President, let me tell you what is going to happen. As soon as anyone attempts to launch a missile, that missile's computer guidance system will shut down. It will lock up tight, and you will not be able to unlock it within the time you need to respond to our attack. Two hours and thirty minutes from now, your nation will have no remaining Minuteman III missiles to respond with, should you finally unlock your frozen computers.

"I suggest that you contact your senior officers now. You will have to mobilize them within 60 minutes. The test should take about 30 minutes. I will telephone you again in 90 minutes to present our terms of surrender." Click.

The President calls the Joint Chiefs. If he is lucky, he will be able to locate two of the three in time. They will be paralyzed. Who wouldn't be? But in all likelihood, they will at least test Gorbachev's theory. They will order one or (wo missiles launched. The computer guidance system on both will shut down the system. They will try two or three more, with the same result. They will attempt to launch one from a submarine, with the same result.

The President brings in senior Congressional officials and the remaining Joint Chiefs member to the White House.

Exactly 90 minutes after he had hung up, Gorbachev telephones back. He presents his list of demands. First, the immediate removal of U.S. troops from Europe. Second, the withdrawal of personnel from Diego Garcia Island in the Indian Ocean. Third, the breaking of diplomatic relations with Red China and Taiwan. Fourth, the removal of all U.S. troops from Korea. Fifth, a moratorium on all debts owed to U.S. banks by the Soviet Union and its client states. Sixth, the removal of all Minuteman III missiles from their silos. Seventh, the return of all U.S. submarines to port. If he agrees, and the orders are delivered within two hours, the Soviet Union will delay launching a first strike. The President complies.

They might do it with our communications satellites, Tom says. You might do it with any aspect of U.S. data transmission. The virus could sit dormant in a system for years, and no one would know. Triggered, it would then strike.

THE WEST'S VULNERABILITY

The West has become increasingly dependent on computers. We can no longer function without them. The Third world hasn't. Neither has the U.S.S.R. Their technology is still pre-computer. They are inefficient, but they are far less vulnerable.

Tom says that the world of computers presumes that almost everyone is essentially honest, and that all the brightest programmers must be honest. They aren't. Thus, the members of the entire system — banks, national defense, large and small businesses, public utilities — have opened themselves to attack. The attackers are invisible.

"Nothing I have seen in all my years of computers scares me as much as this does," he says. "The system has been designed in terms of a far older set of standards, especially with respect to security. It is totally vulnerable."

He compares it to plague, or venereal disease. People copy each other's software to save a few bucks. They use public access databases. They use "loops" — the phone lines. Yet these transmission belts of information can become transmission belts of collapse.

This is what I have harped on for twenty years: the potential for a collapse of the division of labor. We become rich by means of a brilliant technology, yet we become dependent on it to an extent that no previous society ever has.

Centralized institutions are most vulnerable, but because we use public transmission lines, from microwave transmissions to cables in the ground, each local unit is vulnerable. Those who would choose to bring down the system need only plant electronic viruses in a handful of major common-use databases or transmission sources, and live years or ten years later, the disease hits.

It could bring down the system if technological defenses are not developed. Nothing on the immediate horizon points to a solution, he says. The silence of those who should know what to do indicates that they don't know what to do, but they don't want panic to spread.

Is this fantasy? Last summer a 23-person team broke in (by phone) to Chase Manhattan Bank's computer.

Break-in: - Continuation of part 1, a treatise on worms and viruses in cybernetic systems.

"Maybe you saw the article buried somewhere in your newspaper. I saw it in the New York Times (Oct. 19):

> "A group of at least 23 teen-age computer users broke into a Chase Manhattan Bank computer installation by telephone in July and August and "significantly damaged" bank records, the Federal Bureau of Investigation said yesterday.
>
> Chase officials insisted that no money was stolen or transferred out of customer accounts as a result of the break-ins, and said no intrusions have been detected recently. But Federal officials, in an affidavit, said that in several cases the youths changed passwords, preventing customers — and in one case a unit of the bank itself — from gaining access to their own computer files."

And where were these teen-agers located? In San Diego, across the continent! It gets even more ludicrous:

Federal officials said that most of the offenders were probably too young to be prosecuted.

Robert D. Rose, the assistant United States Attorney handling the case, said: "We're not yet sure what we are going to do. But these things can get out of hand — it did get out of hand — and we have to treat them seriously."

Treat what seriously! "These things?" What things? If they can't legally treat the electronic trespassers seriously, just what is the man talking about? He is talking about the topic, above all topics, that bank and government officials don't want to face: the vulnerability of their computer records.

I have seen no follow-up on this story in the conventional press. A brief article did appear in the computer-oriented tabloid, Info World (Oct. 28). It turns out that the students had broken into the fifes of Interactive Data Corp. of Waltham, Massachusetts, which maintains the bank's financial records. The break-ins were discovered in late July. They had obtained the toll-free 800 number which was restricted (Ha!) to Interactive Data subscribers. As late as October 9, an illegal entry was observed. In short, it took ten weeks after the break-ins were discovered to put a stop to them.

The response of the bank's bureaucracy was predictable. It will ever be thus: "Bank officials are claiming that the FBI exaggerated the nature of the activities of the suspected individuals. A spokesperson for Chase Manhattan said that Interactive's customers were not prevented from accessing their accounts and that none of Interactive's data was altered or manipulated in any way." In response, FBI supervisory agent John Kelso said that the FBI has sworn affidavits from bank officials that say data has been manipulated or damaged. "That sounds pretty serious to me," he volunteered.

Here is the capper: Interactive Data has 25,000 subscribers who are tied into that toll-free phone line. Try keeping tight security on a system with 25,000 users. Chase Manhattan couldn't. If they can't, who can?

And if Chase Manhattan Bank was vulnerable to 23 teenagers who are too young to prosecute, consider its vulnerability to just one envy driven genius who knows all about electronic viruses. The students that did this were apparently just goofing around. But what if just one malevolent computer freak decided to "get even" with Chase Manhattan? What if he had phoned in just once or twice, implanted a long-dormant data-killing virus, and quit? What if he had tied its detonation to, say, a calendar clock in the Interactive computer? If it took security forces from July until early October 15 to raid the 23 students' homes, they would never have spotted one break-in. They could not have traced it, either. Conclusion: we have a risk-free opportunity for electronic arson. We face a potential electronic epidemic. And when I say "we" I mean the entire financial system of the West.

Sure, all the bank "spokespersons" in the world will tell you, "No problem." But there is a problem. A horrendous problem.

At this point, it really gets interesting. Chase Manhattan Bank has just announced that we will be able to set up our own personal electronic banking facilities with them by buying an expanded version of *Managing Your Money*, Andrew Tobias' home financial management.

Citicorp and Bank of America have opted for *Dollars and Sense* a rival program. You will be able to pay monthly bills electronically, balance your "checkbook," monitor your net worth, buy and sell stocks, etc., etc., just by dialing Citicorp or Chase Manhattan. Fantastic! But despite all the assurances, I get nervous. Yes, I know no one will be able to break in and tamper with the numbers. But 23 teenagers shouldn't have been able to do it, either. And now we're talking about a lot more subscribers than 25,000.

Obviously, the master program used by the banks will prohibit easy entry. Unfortunately, someone has to write the program. Can you imagine the blackmail possibilities? Some hotshot programmer could build-in a bomb, and then threaten to detonate it. In fact, he could merely pretend to have inserted a virus. Who would want to call his bluff? Not Chase Manhattan, I would bet.

CORE WARS REVISITED

In May of 1984, A. K. Dewdney published an article in Scientific American's "Computer Recreations" column. It was a light-hearted-piece on how computer experts can get involved in playing this exciting game of "blow up your opponent's defenses." You know: recreation! In the March 1985 issue, he wrote a follow-up. It begins:

"When the column about Core War appeared last may, it had not occurred to me how serious a topic I was raising. My descriptions of machine-language programs, moving about in memory and trying to destroy each other, struck a resonant chord. According to many readers, whose stories I shall tell, there are abundant examples of worms, viruses and other software creatures living in every conceivable computing environment. *Some of the possibilities are so horrifying that I hesitate to set them down at all.*"(emphasis added.)

It turns out that tne French nave been enjoying a novel on the international implications, *Softwar: La Guerre Douce,* by Breton and Beneich. A translation is scheduled for publication here by Holt, Rinehart & Winston. The study revolves around the sale of a high-power computer to the Soviet Union. The U.S. allows its export because it has a "software bomb" in it. When the U.S. Weather Service announces a certain temperature at St. Thomas in the Yirgin Islands, the program proceeds to subvert every piece of software in the Soviet Union.

A pair of Italian programmers was "inspired" by the translation of Dewdney's original article to dream up a virus (a virus is a computer-to-computer killer, whereas a worm is resident in one man's computer). They figured out that by infecting a disk operating system disk (these start computers and tell them what to do with programs and electronics), and then installing it on disks used by the biggest computer shop in the city, they could create an epidemic. They decided not to do it. In short, the only restraint is self-restraint.

A high school student in Pittsburg wrote a virus that was subtler than a data-destroying virus, which at least tells us that we have a problem. His virus created a plague of very subtle errors in the disk operating system. "All of this seems pretty juvenile," he wrote, but "Oh woe to me! I have never been able to get rid of my electronic plague. It infested all of my disks, and all of my friends' disks." He wrote a program to destroy the virus (an "antidote" to it) but it is not anywhere near as effective as the virus is.

Warning: do not copy disks from your friends' copies. This act of piracy could cost you plenty.

A COMMERCIAL WORM

Just a few days after I wrote "Electronic AIDS, Part 1," 1 read a column in the Washington Times, the conservative (Moonie-owned) daily newspaper. One of the reporters has a computer. He had purchased a newly released program from

Microsoft Co., called "Access." Understand that Microsoft supplies the disk operating system that is used by the IBM PC, the most popular microcomputer. In other words, this is no backyard company. It is one of the two or three software giants in the U.S. (Its owner is under age 30, which tells you something about who is pioneering the microcomputer revolution).

As he was setting up his computer to take advantage of this telecommunications program, a warning flashed on his screen: "The weed of crime bears bitter fruit. Now trashing your program disk." Wham! He lost all his files — probably a couple of year's worth of work. Sure, he was probably smart enough to have made back-up copies, but think of the risk. And what if it had been a worm that kept silent for a few years, infecting all of his back-up disks?

He called Microsoft, and they gave him the runaround. They told him that they were not responsible. Some programmer had put in the worm in order to zap program pirates, but the journalist insisted that he was an original buyer. Tough luck, they told him. Obviously, they didn't know that he was a reporter.

Then he published his article. All of a sudden, the victim was not some average buyer. He was big trouble. Things started moving. InfoWorld (Oct. 28) reports that Microsoft has admitted that a programmer put in the worm, but without permission. The offending text has now been removed, we are assured. But what if it had sat in the master for three years? Here is the premier firm in the software.

This is not idle speculation. It has already happened, verifying my hypothetical scenario within a few days after I published it.

Can you imagine the absolute havoc that a dormant worm or virus could create if it were imbedded in all updates of Microsoft's masters of PC DOS and MS DOS, the operating systems for all IBM microcomputers and IBM-compatible microcomputers? It could cost the U.S. economy billions, and some microcomputer-dependent firms wouldn't survive. Any Microsoft spokesman who says, "It's impossible; it could never happen," has to explain how it already did happen to "Access."

ADAM OSBORNE'S WARNING

You may know the name Adam Osborne. He invented the revolutionary portable computer, the Osborne 1. Before there was an Osborne 2, the company went bankrupt. Compaq, the most successful first-year firm in U.S. history (over $100 million in sales in its 12 months of operations) and others built imitations that were far superior.

That isn't my point, however. Adam Osborne was "present at the creation" of the microcomputer industry. He created Osborne publications, and then sold out to McGraw Hill. He knows what is going on. In his delightful paperback book, Running Wild, which is a history of the microcomputer (desk top) revolution, 1975-82, he offers this warning. He says that three areas should not be allowed to be computerized: 1) bank money transfers; 2) the stock market; 3) elections.

All three are just about fully computerized. Another ten years, or maybe five, and they will be 100% computerized. Several firms allow microcomputer buying and selling of stocks (e.g., Charles Schwab), and New York Stock Exchange floor transactions eventually will be fully computerized, at which time it will be pressured to get rid of the "specialists" who make (and sometimes manipulate) the market, short-term — Richard Ney's hated "Wall Street Gang" — but the price of getting rid of them may turn out to be horrendously high.

"The great fortunes of the 21st century," Osborne predicts, "will be the legacies of the great computer thieves of the 20th."

Three years ago, I used a firm to supply computer services I needed. The head of it was a former businessman, quite young, and a true "space cadet." I've quoted him in the last issue. I call him Tom. He operated in a world far removed mentally from the rest of us. He is a nice fellow, a Christian, and a moral philosopher of sorts.

He ran the operations of the local elections. He did it fairly inexpensively. He told me why: "I want to keep these elections honest. It would be incredibly simple to rig the program to produce whatever outcome I wanted in close races. If I can do it, anyone with enough skill to set up the system could do it."

I asked him if he thought Osborne was correct in his predictions about bank theft. "It would be a piece of cake for me to steal three or four million from any local bank. I could go in the next week, offer to give the money back, keep 10% as a finder's fee, and promise not to tell the press how easy it was to steal. They would probably pay me my 10% just to keep me quiet."

Look, These people are geniuses. Worse, they are geniuses in a very narrow field technically, which is now being used to control darned near everything. This unique intellectual-technical skill is the possession of literally a handful of people, mostly under 35 years of age. They are "fooling around" with Chase Manhattan Bank's computers. What happens when a few of them stop fooling around and get deadly serious?

Computer program designers keep telling us that there is no 100% secure way to defend data banks. Maybe there will be a 98% secure system someday, but not now. To defend itself, the system relies on the integrity of youth. In short, self-government is the major defense.

And where have they learned self-discipline? In the public schools?

"NOW YOU'VE DONE IT!"

About four years ago, I read an article in the *Rolling Stone,* the tabloid aimed at rock music fans. It was the only article I ever read in that periodical. It was a gem.

It described a subculture of students at Stanford University, "hackers." These people are computer freaks. The mainframe computer at Stanford was cheaper to use after midnight, so from midnight to 6 a.m., the hackers gathered at their terminals. They lived on candy bars, junk food, and high-technology dreams.

One of the games they played was breaking into each other's programs. It was considered the mark of a master hacker to be able to crack another hacker's defenses. They would spend hours trying. They were "hacker-crackers."

One bright fellow then designed a classic booby trap. He wrote a program that warned trespassers not to tamper with it. This, of course, alerted every would-be electronic safe cracker to the challenge. It was a complex program, and it took days to crack it. Then, after repeated warnings, the successful trespasser got a surprise. Japanese letters appeared on his screen. Roughly translated, the words proclaimed, "Now you've done it!"

At that point, the victim's computer screen went blank. Then the names of all his own computer files appeared on the screen — files that may have taken years to assemble. One by one, they blipped off the screen. In horror, the victim would stare at the screen, unable to stop the process.

As it turned out, the booby trap was only a practical joke. It really didn't erase all the victim's files. It only listed the names, and then erased them. But for a horrifying few minutes, the victim wouldn't know this.

Hackers play games. Very interesting games.

The kind of people who spend six hours, midnight to 6 a.m., trying to break into each other's programs are different from the rest of us. Among their ranks are some highly individualistic people. Some of them are libertarians. I mean anarchists. They are electronic "don't tread on me" sorts of people. They do not appreciate bureaucracy. They appreciate being pushed around even less.

The folks at Chase Manhattan really do have a problem. Do you attempt to prosecute a legally un-prosecutable kid? A kid who has already cracked your computer system? I don't think you do. You play the role of stern but appreciative banker. "Son, I am impressed by your ability to break in. But understand, we are honest people. There is a code of honor here. You wouldn't want to break that code — of honor, I mean — would you?" Because if this kid gets angry, he can do it again. Quietly. And next time, he deposits a virus.

Of course, Chase may hire a programming team to create an unbreakable system. Sure. "Hire fox A. Give him chain link fence B. Hire him to build fence B around chicken coop C."

TEEN CHALLENGE

Suppose that the public gets wind of the threat to the whole banking system that is posed by viruses. What do the bankers (or anyone else) announce to the public? "We want to assure you that our computer program is impenetrable. No one can break in. It is foolproof."

Here is the challenge — rather like the Stanford program that announced: "Do not trespass." These kids see breaking-in as a challenge, a kind of sport. They do not regard it as theft. They do not regard it as vandalism, even if it costs a company millions of dollars to unscramble. They may be ethical in other respects, but they think of "core wars" as a game.

How would you like to be the 60-year-old banker who doesn't know a byte from usury, but whose public relations department tells him to inform the public that nobody can crack his bank's code? To cite Mr. T in "Rocky III," that bank is dead meat. So are its depositors.

But if he keeps quiet, and the story still gets out about the vulnerability of the system, one or two small "virus-demolished" banks could trigger a collapse of the system, as people do the only smart thing: run for cash. The whole fractional reserve banking system would deflate; only the FED's printing presses could "save the day," in a wave of fiat money.

What I am saying is this: I think that we will see the end of fractional reserve banking in our day. At the very least, I think we will see it subjected to tremendous shocks. People will lose faith in electronic promises made by bureaucrats who do not know anything about the monsters that their efficient computers can be turned into.

Attack on Martinsburg. Now, let's take it a step farther. Some day some state or Federal bureaucrat is going to step on the toes of some genius entrepreneur who has created a software development firm. The bureaucrat will try to wrap this entrepreneur in red tape. Or maybe — just maybe — he will try to sock him with a tax bill that the entrepreneur regards as unfair.

In Martinsburg, West Virginia, there is a large computer. It is owned and operated by the Internal Revenue Service. Into it, over the next five years, the I.R.S. apparently intends to deposit all the records it can assemble on every U.S. taxpayer. This computer database will be the biggest in the world. It is the tool by which the I.R.S. hopes to increase taxpayer compliance. And it may succeed. For a while.

This is one reason for saving all letters to and from the I.R.S. If the I.R.S. becomes dependent on its computer system, which is likely, then any short-circuiting of its database could create havoc for tax collecting. If word gets out that a major failure has hit the I.R.S., the tax revolt could multiply overnight. You would see the deficit become astronomical. If the I.R.S. continues to tie its "voluntary" compliance program to the myth of the "all-seeing computer," then news of the computer's scrambling could backfire.

It is possible that the story of the I.R.S. database is a myth. Maybe they aren't going to build it. But if the public believes that such computer power is at the disposal of the I.R.S., and taxpayers then learn either that the system has been blown, or that it was mythical from the start, the tax revolt could spread like an epidemic. The electronic epidemic-could trigger a tax revolt epidemic.

He who lives on the cutting edge of technology eventually dies on the cutting edge of technology.

"PEOPLE ARE BASICALLY GOOD"

Let's return to my taped interview with "Tom." In a 90-minute interview, we covered a lot of ground. But one topic that stands out in my mind is our discussion of the pre-supposition that goes into the creation of a computer-based society. The computer people have all adopted the assumption that undergirds modern science, namely that, participants are well meaning, that they will not fake their experiments, and that they will play fair. If scientists had to check every aspect of every article, science could not advance very fast.

What about the computer industry? The whole system rests on faith: "Men are not malevolent. They are not envy-driven. They will not deliberately seek to destroy the work of some random victim." Tom says categorically that this assumption is false. There are bad people with tremendous computer skills, and modern society has not restructured its economic institutions to protect itself.

Here is one example of a break-in technique. Someone phones in to a computer, which has been left open temporarily by some user. The lock is unlatched; he needs no key to get in. He then seeks to penetrate the inner core of the program, such as a bank's program. He creates a deliberate error, which all too often triggers a kind of electronic explosion. The protective shell self-destructs, and the invader now finds himself inside the system, where far fewer defense mechanisms exist.

Tom designed his own firm's defense against this tactic. His program automatically records the source of the error, and throws the user out of the program. The program has protection against deliberate errors, but most of them don't, he says. A major error simply collapses the program's outer shell.

In my previous issue, I speculated that a Soviet spy or agent could penetrate U.S. computers. Note: I did not assume that he would simply phone in; I assumed that a disloyal programmer, or a team, could plant the virus as insiders. From there, the virus would spread through the system through normal telecommunications. Several people have written in to tell me that a wrecker cannot destroy the system by penetrating it from the outside. They may be correct. But when informed that I am assuming an inside job by someone with access to a major computer, the critics have admitted that this might be possible.

The weed of crime bears bitter fruit: for honest, computer-dependent people.

FEDERAL FUNDS

The Federal Funds bank transfer lines allow banks to borrow money overnight. Hundreds of billions of dollars go across these lines every working day. The bank's computers communicate with each other by means of this telecommunications hook-up. What if someone were to plant a long-delay virus in the software that operates these transfers? And what banker has even thought about this problem?

What if this scenario was to take place: a virus triggers the disruption of bank records — not a total breakdown initially, but disruptions in the data? It might be weeks or months before auditors recognized the extent of the problem.

As rumors begin to leak out about complex accounting or other data-management problems of major banks all over the U.S. (including off-shore branches), the various banking regulatory agencies would be swamped with crises and outside rumors. Then, all at once, bank computers begin breaking down.

The rumors then explode. The lines appear in front of banks. The only answer at this point is to print up paper money. It would be printed by the hundreds of billions in order to offset the deflationary effects of bank runs (paper money that is pulled out but not re-deposited in another bank).

You could topple the fractional reserve banking system all over the world. The entire payments system could easily become engulfed in chaos. Debits and credits would no longer be meaningful. A pure paper money inflation would replace the manipulated, "fine-tuned" monetary inflation of modern central banking.

All of a sudden, market-created alternative currencies would be revived. It would then be metallic cash that talks loudest. Silver dimes are not electronic. "They can't be infected electronically." They still circulate when banks are "temporarily closed, due to circumstances beyond our control."

The loss of efficiency would be initially horrendous, I would guess. The division of labor would break down. You could then have the crash that lurks in the minds and suspicions of average depositors. Who says it cannot happen? A lot of public relations firms hired by the banks — computer illiterates in high places?

What we have is an internal bank money wire system that is totally vulnerable to some vindictive programmer. There is little doubt in my mind that the bankers are desperately fearful of this sort of vandalism. It could topple people's confidence in the fractional reserve banking system, and confidence is the only thing that keeps it going.

Conclusion. Technologically, there is no solution at this point. I have no heartening message. Maybe later; not now. Keep precious metal coins. Don't assume that it can't happen here. It can. The only thing holding it back is the restraining hand of God, through the temporary self-restraint of a technological priesthood.

[Reviewer's comment: I was unaware of the problem of computer worms and viruses until I read Gary North's two articles. His open release of this important information is appreciated.]

SCALAR EM INNOCULATION AND ITS IMPLICATIONS

My appreciation goes to Gary North for his openly sharing the above important information with the public. Human nature being what it is, one or more of his more dire scenarios is going to happen sooner or later.

The problem becomes enormously more compounded if the hostile agent or agents possesses sophisticated scalar EM technology.

With scalar technology, there exists the potential that the worm or virus can be implanted from a distance. If scalar-scanning devices of sufficient precision can be developed, then major software residing in a chip, a computer, or on a disk can be read from a distance. With scalar EM methods, the ordinary software "guards" are totally ineffective. The scalar "reader" could simply read off the software directly as the disk or device goes about its normal operations.

After the software is read, a germ or a virus is designed to produce a particular problem as desired and upon special stimulus. Then a scalar "imprinter" device is used to produce the necessary digital pulses directly upon the media in which the software resides. The required code has now been implanted. In effect, the operating system has been injected from a distance, and contaminated with the desired "sleeping agent" — germ or virus. From there on, it's just like Gary North's scenarios.

So the scalar EM technology can probably be used (at least eventually) to place "mines" in systems in advance, for damage or destruction at a future date. At least it's possible in principle.

However, it's not the injection of computers and equipment that bothers me. It's the injection of human beings. Here's generally how it can be done (and possibly is being done).

Basically, human brains have various rhythmic waves, which represent "envelope pulsations" of the entire electrical system. However, it is not the Et, force fields represented by these pulsations which are important: it is the summed-zero (scalar) patterning of the entire complex of neuron firings which represents the real activity of the brain — and represents its thoughts and emotions. A storage system for these patternings exists, but it's a scalar pattern storage, not an E and B field storage. The storage system is in the nuclei of the tissues of the nervous system itself, not in electrical or electrochemical potentials.

At any rate, the human brain rhythms are also "riding along" relatively entrained to the Schumann resonance of the earth's magnetic field. So all the patterning of the brain is grouped together into giant overall orders, and one or more of these is normally synchronized to our natural environment.

If a varying magnetic signal of extremely low frequency — say 10 hertz — bombards the brain with greater power than the earth's magnetic field, it is possible for the brain's rhythm to be "forcibly entrained" to synchronize with that signal, rather than with the normal Schumann resonance. If a large number of brains are forcibly entrained to such a signal, and the signal is artificial, then a nester coherent phasing of all the brains to the transmitted signal has been accomplished.

And that can be utilized — and is utilized — by the Soviet Union. On the Woodpecker signals, Bob Beck has detected up to 16 carriers, each with a 10-Hertz modulation, and with all the 10 Hz modulations in phase.

Let s stress what we're saying here. At times, the Woodpecker signals have shown 16 carrier frequencies that contain the same 10-Hertz modulation, and which are all kept coherently sync-locked.

If those signals and the modulations are strong enough, the brains in a targeted area demodulate the 10-Hertz signal and phase-lock to it — actually, to them. At that point, the brains are entrained to a 10-Hertz driver, and 16 channels of multimegahertz frequencies are locked-in to the brain.

Now, if the Soviets have broken the genetic code of the human brain (which they have) and succeeded in developing the ability to implant thoughts, emotions, pictures, etc. (which at least to a degree they have), then they can simply record the proper signals from human guinea pigs, using scalar technology and instrumentation.

They can tailor those signals and add them onto the Woodpecker carriers, using scalar EM techniques. By using most of the sixteen carriers as a means to express Fourier frequency expansions, specific areas and functions of the brain can be targeted.

"Notice that the distant transmitters already "know" the correct phasing, just use the 10 Hertz modulation signal as a reference. In that case, phasing takes care of itself inside the captured human brains.

The brain has a unique characteristic. If a signal arises inside it, it interprets that it itself originated the signal. Thus for the first time in history, true unwitting robot actions and emotions can be engendered in masses of human beings.

But it's still worse. The Kaznacheyev experiments (see references) proved conclusively that electromagnetic signals can carry and transmit cellular disease and death patterns into a targeted biosystem. You can rest assured that the Soviets continued on to actually measure and determine a wide variety of death and disease scalar patterns. All of that can be added on to the Woodpecker carriers so that the population in the intersection grid area is devastated. One can even test it out in advance by producing a modified pattern of a disease in the targeted populace, and keeping the percentage of brains entrained down to a small level. This lowers the number of persons "injected" from a distance to a small percentage of the populace. Then the outbreak of the "new strain" of a familiar disease — such as influenza, pneumonia, and something like AIDS — can be monitored to provide targeting and effectiveness data. Notice that a series of diseases was experienced by U.S. personnel assigned to the U.S. Embassy in Moscow, probably as a result of the so-called "weak microwave radiation" of the Embassy by the Soviet Union.

But that's not all. Much more subtle effects can be produced in the targeted populace, without anyone getting wise. Cancers and leukemia should be relatively easy to induce. Also, the immune system can simply be attacked directly. If done slowly, a series of common, ordinary diseases will just seem to get "tougher and tougher to cure." Actually it's not the disease that got tougher, it is the immune systems that got weaker. It is my opinion that precisely that sort of testing — against the immune system — has been accomplished by the Soviet Union against the U.S. since about October 1985. A great percentage of influenza patients during the period from then until now have had complications. So much so that multiple sessions — as many as 3 to 5 — of antibiotics are necessary before the patient recovers. At least in the Huntsville, Alabama "hinge pin" area, that seems to have been the case. This author personally experienced this, having influenza and follow-on complications for a period of seven months — something quite different indeed for him.

But it's even worse.

The human scalar EM nervous system ("inside" the ordinary EM signals and functioning of the system) also has its own resident "software." It has special storage media contained in its brain, its cells, and its genetic material, and its actual nucleons. All of these storage media are continually telecommunicated with via scalar EM means. Those recordings, largely condition the "stability" of what each of us is. As personalities we literally "are" what we have experienced, felt, thought, and did. Our whole "functioning conscious system" has been evolved by our experience, and it continues to be "us" by means of continual reference to the record base.

I hope you now realize that the actual record base itself can conceivably be altered by scalar electromagnetic means, even from a distance. Even by the hidden scalar EM content of 16 Woodpecker carriers, rocked into a captured brain and pouring in their deadly content. It is conceivable that even what one is can be taken from him, and what someone decides he will be, he will be.

A totally new — **and** ultimate — form of "slavery" has emerged.

Also, look at all this scalar EM functioning of the human system as a special kind of cybernetics. A scalar EM computing software and hardware system.

What if scalar techniques are used to plant "worms" and "viruses" in this system? With hidden triggers? What about a series of them, so that a range of eruptive software diseases can be stimulated at will?

And guess what. Such scalar EM "worms" and "viruses" can even be induced by contact with or consumption of "contaminated" water and food. Or saliva. Or other body fluid.

So you can see the possibilities of "seeding" a population with a staggering variety of time-bombs.

I think you begin to see the point. Now you can change North's "Soviet blackmail" scenario drastically. In a new scenario, Gorbachev simply announces to Reagan that a graduated series of specific sweeping changes are going to happen in our population — changes which we are absolutely powerless to stop. The last stages are various deaths. The master "trigger" for this series has already been initiated. Incidentally, the Soviet populace has been carefully "seeded" with the proper antidotes. The Russian leader demands certain things (very similar to what North's scenario utilized). Unless his demands are met, a "killer virus" will be activated by a special signal. If so, everyone dies. The first step change in the series will start to occur in a few hours. And it does.

In fact, this scenario could be added to North's scenario, and both used simultaneously. What do you think the President and the Congress would do? Or could do?

There are other even more diabolical possibilities, but these should be quite sufficient to illustrate the point.

Pandora's box has already been spilled, and the end of humanity is ticking away like a time bomb. It's already nearly midnight, and the watchman hasn't even sounded the alarm yet.

I certainly can anticipate the next question: Have the Soviets already "seeded" the populations of the U.S. and other selected western nations with scalar EM viruses and germs7

The answer is, I simply don't know. And just like Gary North, I don't know what if anything can be done about it if they have. This researcher certainly doesn't have all the answers. An immediate national crash program to develop scalar EM defenses is essential if we are not simply to be crushed beneath the hammer and sickle. It takes funding, people, facilities, and time.

In some 22 years of unrelenting struggle to convince the system we've got a terrible problem, I've hardly been able to move it off dead-center (no pun intended). This book, produced under extreme workload stress, is all I can do to raise the alarm. I've given it my very best shot.

What happens now remains to be seen. Other hands, hearts and minds are going to have to take up the task and carry it forward. This researcher has reached his limit, and has no more left to give.

APPENDIX III: CONTINUING THE SCENARIO

On May 3, 1986 a stunning loss of a U.S. Delta rocket carrying a critically-needed weather satellite occurred. The rocket suffered a shutdown of its main engine during launch and began to veer off course, causing the range safety officer to destroy it.

The takeoff of the blue and white Delta "workhorse" from launch Pad A at Cape Canaveral at first appeared picture-perfect. About a minute later, six of the nine booster rockets were expended and the remaining three switched on. Shortly after that, the rocket seemed to lose power. The main liquid-fueled engine had suddenly shut down, with the three strap-on boosters still burning. The rocket then had no stability and started to slowly drift off course, creating an increasing angle of attack. Within a few seconds, the nose cone snapped off under the severe stress. The U.S. Air Force range safety officer then sent a destruct command to the stricken rocket, which exploded in a ball of flame.

The Delta, built by McDonnell Douglas, had previously piled up a continuous string of 43 successful launches. Since 1960, 177 launches have occurred, and the success rate has been 94%. The Delta is one of the most reliable rockets ever built.

The unexpected shutdown of the main engine was perplexing. At a news conference following the mishap, William Russell. NASA's Delta project manager. stated: "It was a very sharp shutdown. almost as if it were a command shutdown."

Preliminary data, such as turbine speed and temperatures in and around the engine, "all appeared to be very normal." Russell said.

It may well have been a "command shutdown," unknown to Bill Russell and his NASA engineers.

Preliminary runs of the videotapes of the incident, aired by the national TV news media, show that during liftoff, a mysterious light approached the rocket from below and from the side. rising and striking the vehicle. This light and its path are strongly reminiscent of the first anomalous "light streak and ball" photo taken by Bob Gladwin just prior to the shuttle launch on November 26, 1985.

Let' s suppose that on May 3 the Soviet scalar EM grid was in the launch Phase ABM system mode and was tracking the launch of the Delta rocket.

If the anomalous light that moved up and contacted the rocket was a small ball of continuous electromagnetic energy formed by a distant Soviet scalar EM howitzer, then the rocket's main engine control circuits would have received a sudden and continuous pulse of EM power. This would probably have shutdown the engine by an actual "command" signal, artificially delivered. *

---

*Engineers have now determined that such a shorting of the engine circuitry occurred, causing engine shutdown.

Certainly such mysterious switching of satellite circuits has occurred before, probably due to Soviet scalar EM precision testing. For example, the British Ariel VI satellite has exhibited very strange anomalies in its command and control circuitry -- in fact, with very precise control switching being exhibited as if from an external source.

Thus a very small scalar EM missile may have been utilized by the Soviets to disrupt the control circuits in the rocket. A resulting pulse of power in the engine shutdown command circuit could have "ordered" the shutdown of the main engine -- just as it appeared.

It remains to be seen whether or not the Soviets used a "marker beacon" associated with the disruption of the Delta launch. One news photograph seems to contain a suggestion of such a light, but the videotapes must be studied frame by frame before it can be ascertained exactly what if anything is really there.

If a marker beacon was not used, the distant Soviet operator could simply have "popped out" the continuous EM ball of energy, tracked it and the ascending rocket exhaust, and "joysticked" the scalar EM missile up into the rocket.

Analysis of this incident is continuing, so nothing is certain as we go to press.

However, if this incident was indeed another Soviet kill, then it will have served the Russians well. Much of the world's media attention will now be focused on NASA's launch difficulties, and away from the recent Chernobyl nuclear reactor disaster in the Soviet Union. It's bound to take some of the media pressure off the Soviets.

In addition, it further cripples the U.S. satellite program. Loss of this critical weather satellite payload seriously reduces U.S. weather monitoring capabilities.

Further, the U.S. space program now appears to be all but grounded by the Challenger, Titan, and Delta explosions. For example, it will require about 18 months for the U.S. to redesign the shuttle so that America's shuttle program can get going again. So things are delayed until summer of 1987 at the earliest.

Meanwhile, the Russians are set to grasp the lead in the space race. In 1985 they launched about 100 rockets, five times as many as the U.S. They have a shuttle program of their own, well underway, with a flight expected within months. Earlier this year, they launched Mir, a second generation space station; a U.S. version is still some eight years away.

Charles Vick, of Huntsville, Alabama's Space and Rocket Center, summed it up: "They have their standard military program, a shuttle program and the equivalent of an Apollo program, all within a 10-year period. It's awesome."

An experienced propagandist like Gorbachev certainly understands the advantages of a strong Soviet push into space, such as (1) military advantage, (2) national prestige, (3) scientific achievement, and (4) projecting an image in the world's news media that the Soviets are first in space exploration.

So the failure of the Delta rocket provides many benefits to the Soviet Union.

It also continues to exercise and demonstrate Soviet mastery of the skies over North America itself.

Meanwhile, two U.S. nuclear submarines have "inadvertently" run aground in the last month or so.

Gorbachev again seems disposed toward a summit meeting soon.

If it's held, the agenda may be quite different from the one President Reagan is prepared for.

The clock ticks on toward midnight.

2002 ENDNOTES

1. It is actually a unified field theory since the infolded electrodynamics inside all EM potentials, fields, and waves is utilized—and the internal electrodynamics forms spacetime curvatures and dynamics of those ST curvatures.

2. M. W. Evans, "0(3) Electrodynamics," *Modern Nonlinear Optics,* 2nd Edition, M. W. Evans (Ed.), Wiley, New York, 2001, Vol. 2, p. 79-267.

3. David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II." *Phys. Rev.,* 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II); — "Quantum theory as an indication of a new order in physics." A:, *Found. Phys.,* Vol. 1, 1971, p. 359; B: "Implicate and explicate order in physical law," *Found. Phys.,* 3(2), 1973, p. 139-168.

4. Mendel Sachs, General *Relativity and Matter,* Reidel, 1982; — "Symmetry in Electrodynamics: from Special to General Relativity, Macro to Quantum Domains," *Modern Nonlinear Optics,* 2nd Edition, Wiley, New York, Vol. 2, p. 677-706.

5. M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," *J. New Energy,* 4(3), Winter 1999, p. 299-301 gives a basic mathematical expose of scalar interferometry.

6. James Clerk Maxwell, "A dynamical theory of the electromagnetic field," *Phil. Trans. Roy. Soc,* Vol. 155, 1865, p. 71, 459. Orally presented in 1864. This was Maxwell's definitive presentation of his theory, with 20 quaternion and quaternion-like equations in 20 unknowns. Also in *The Scientific Papers of James Clerk Maxwell,* edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-604.

7. In T. E. Bearden, *Gravitobiology: A New Biophysics,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1991 we first exposed the quantum potential weapon and some of its uses and characteristics.

8. A negative energy EMP weapons strikes the target or target area with an extreme pulse of negative energy empty Dirac Sea 4-holes. As a result, the free electrons flowing in electrical currents in the struck target instantly disappear by re-absorption into the Dirac Sea. A living body struck by such a blast is totally and instantly dead in every smallest part, since all its brain and nervous system currents, cellular currents, etc. are instantly and totally extinguished. The Russians tested this weapon in the 1980s in their war in

Afghanistan, against two Afghan villages. I spent two hours in Washington D.C. with the then-representative of the Mujahedin to the U.S. Government, verifying the incidents and going over the exact characteristics of the instantaneous deaths of those villagers. With such a sudden death, the mind-body interconnection loop is instantly severed, so that the mind is instantly snapped loose from its connection to the body—hence my use of the term "MindSnapper." The struck body usually emits longitudinal EM waves for an extended period, so that vultures and other predators will not approach the bodies and eat them. All bacteria, viruses, etc. in the body are also killed instantly, and the body does not decay, even in 30 or 60 days.

9. As can be seen, the negative energy EMP weapon is an all-purpose weapon extremely effective against every kind of target: missiles, aircraft, communications, satellites, ships and particularly aircraft carriers, personnel, generators, power systems, the electronics of nuclear warheads, submarines, tanks and vehicles, and the electrical and electronic controls of any kind of machinery or equipment. Such weapons are the second most powerful weapons on earth. In the first part of 2002, a U.S. 2-carrier task force maneuvered in the South China Sea as a precautionary warning to the Chinese that we would defend Taiwan. Unknown to that fleet, tracking it and trained on it were the newly deployed Chinese negative energy EMP weapons, which could have destroyed the entire task force—everything in it—with a few shots in less than 10 minutes.

10. T. E. Bearden, *Energy from the Vacuum: Concepts and Principles,* Cheniere Press, Santa Barbara, CA, 2002. Chapters 1 and 2 give a deeper coverage of this and similar foundations principles.

11. As is well known, Russian theoreticians have always been the best nonlinear scientists on Earth, since the beginning. They still are.

12. See T. E. Bearden, "Giant Negentropy from the Common Dipole," *J. New Energy,* 5(1), Summer 2000, p. 11-23 for solution of the source charge problem. On DoE restricted website http://www.ott.doe.gov/electromagnetic/ and on www.cheniere.org. The problem of the source charge can be solved by treating (with the modern quantum field theory viewpoint) the "isolated" observable charge and its cluster of virtual charges of opposite sign as a composite dipole. Because of the broken symmetry of its opposite charges, any dipole or dipolarity continuously receives virtual energy from the seething vacuum, transduces it into real observable EM energy, and pours out observable photons in all directions at the speed of light. See also E. T. Whittaker, 1903, *ibid.* There is a scalar potential between the ends of each dipole. The scalar potential may be decomposed into a bidirectional set of longitudinal EM phase conjugate energy from the time dimension (from the imaginary plane) and outputs that EM energy in the real 3-space. 3-space conservation of energy flow

is violated, but 4-space conservation of energy flow is upheld. This rigorously follows from the discovery of broken symmetry of opposite charges in 1957—one of the types of asymmetry for which Lee and Yang received the Nobel Prize.

In classical electrodynamics, the charged particle itself is *assumed* to be the source of its own fields and potentials. In other words, the charge is implicitly considered to be *creating*—right out of nowhere—the Poynting energy flow it continuously pours out across the universe in all directions. That makes the charge a perpetual motion machine of the grossest kind, violating the maxim that energy can neither be created nor destroyed.

So the greatest perpetual motion advocates of all times are our own classical electrodynamicists and electrical engineering departments and professors who still ignore the active vacuum interaction that is missing from the classical model. They thereby ignore that the charge is a broken 3-symmetry in its violent exchange of virtual energy with the active vacuum, and merely organizes and emits in observable form a fraction of the energy it continuously receives from the imaginary plane (from the time domain). Yet, that broken symmetry has been known in particle physicists, both experimentally and theoretically, for 45 years.

When Maxwell wrote his theory, neither the electron, the atom, nor the nucleus had been discovered. There was no knowledge of the Drude electron gas, electron spin, etc. There was no knowledge of the nucleus of the atom, and even the atom itself was still contentious. A thin material ether (fluid) was believed to fill all space, so that no point in all the universe was empty of mass. Consequently, E and B were conceived as physical force fields in a material medium—the luminiferous ether. All Maxwell's equations—including the subset selected and modified by Heaviside that today are called "Maxwell's equations"—assume this material ether to this day.

Electricity was thought to be a "thin material electric fluid" running through wires like water through a pipe. The magnitude of the potential was not used; its reaction cross-section was calculated and erroneously used as its magnitude, as it is today (although a good electrodynamicist does know that it is only the *intensity* of the potential, not its magnitude).

Faraday believed that his "lines of force" were taut physical "strings" in this material ether, and that EM perturbations were the vibrations of these taut strings. As he clearly states in his 1873 *Treatise...*, Maxwell sought to capture Faraday's work exactly. He simply *assumed* Faraday's "plucked taut string" waves. Later the measurement in instruments of what was actually the lateral gyroprecession of electrons in receiving wires was thought to prove that the incoming ether vibrations were transverse, like those string waves, because the detected electric **fluid (electron** precession) waves in the receiving wire were transverse. The drift velocity of electrons longitudinally in a wire (in a typical case, several inches per hour) was unknown. A charge *q* was just a piece of electric fluid, much like a cubic centimeter or other volume of fluid; it had no connotation of "charged particle." There was no concept in all of physics of EM energy flowing through space. There was no knowledge that the same fields created in the transmission wire, which perturbs the Drude electrons laterally, also perturbs the positively charged nuclei laterally, in the opposite direction, highly damped but with equal energy.

There was no way to realize that the energetic perturbation of the curvature of spacetime around the sending wire antenna, by the Drude electron perturbation energy, is simultaneously accompanied by an equal and opposite perturbation of the curvature of spacetime in the opposite direction (energy density, not amplitude of the disturbing wave, determines the degree of curvature of the surrounding spacetime). Hence Maxwell omitted precisely one half the EM energy, one half the EM wave, and one-half the electrodynamics. As a consequence, the "Newton's third law recoil" of the nuclei today is thought to spontaneously arise as if some mystical demon observes the Drude electron perturbations, and simultaneously kicks the positively charged nuclei in the equal and opposite direction. Hence Newton's third law, so to speak, is erroneously still omitted from classical electrodynamics, and the better texts say so.

In the 1880s, the Michelson-Morley experiments falsified the material ether. Not a single one of the Maxwell-Heaviside equations were ever changed to this day. Instead, the electrodynamicists eventually announced that, since there was no longer an ether, they were not using one! Then a total non sequitur was incorporated: that in massfree space, the same E and B fields (defined as material entities and thus existing only in mass) still remain but the forces go to zero. Simple algebra reveals this clever, insidious faux pas, using the E-field as an example: First, $F \equiv d/dt(mv)$, so mass is a component of force. (This reveals the great error due to the hoary old notion that a separate force acts upon a separate mass; the mass is a component of the force, and the force is a coupled system of a curvature of spacetime and a charged mass.). Then $E \equiv F/q$, which rigorously holds when mass is present.

Now for whatever reason, let $F = 0$. Then $F/q = 0$ and $E = 0$ also, else we must discard algebra. What actually remains is easily explained in general relativity terms (available since before 1920). The couplet of a curvature of spacetime and a charged mass is broken up by absenting the mass. What remains is that specific curvature of spacetime. GR assures us that curved ST acts upon mass, and that any force involves a curvature of ST. Yet in more than 80 years, this erroneous old "material ether remnant" in electrodynamics has been fiercely defended. Every textbook today uses and confuses the E and B fields in two dichotomous fashions: (1) as massive force fields existing only in the presence of mass, and as "massless force-free fields" existing only in the absence of mass. This of course is a non sequitur.

13. E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," *Math. Ann.*, Vol. 57, 1903, p. 333-355. The "scalar" potential itself is not even a *scalar entity*, as shown rigorously by Whittaker. Instead, it is a harmonic series of bidirectional EM longitudinal phase conjugate wavepairs, and it decomposes into that series. Hence it is a *multiwave, multivectorial entity.*

14. C. N. Yang and R. L. Mills, "Conservation of Isotopic Spin and Isotopic Gauge Invariance," *Phys. Rev.*, Vol. 96, 1954, p. 191.

15. The term "energetics" simply means *energy as the primary thing in the universe.* Rankine first used it. G. Helm in 1887 revived Rankine's term "energetics" and factored energy into an *intensity* factor and an *extensity* factor (which today we would call potential and kinetic energies). William Ostwald in 1893, 1895, and 1901 raised the concept of "energetics" (energy as the primary thing in the universe) to the forefront. Thus when the Russians developed new extensions and principles in weapons physics, they simply used the historical notion of "energy as the primary thing in the universe" and adopted the earlier name for it: *energetics.*

16. Later when Khrushchev demanded a 100% antimissile defense, famed Russian Nobelist Petr Kapitsa answered, *"If a means of total neutralization of foreign missiles is to be found, it can only come from a group of new principles in physics, called energetics".*

17. See T. E. Bearden, "Mind Control and EM Wave Polarization Transductions, Part I", *Explore*, 9(2), 1999, p. 59; Part II, *Explore*, 9(3), 1999, p. 61; Part III, *Explore*, 9(4,5), 1999, p. 100-108. It appears that Sachs' unified GR and EM approach is perhaps the best Western physics model to apply to analysis of the KGB energetics weapons. However, because of the third branch *(psychoenergetics:* engineering the mind directly), Sachs' work would appear to need further extension along the lines indicated by the present author in the cited paper.

18. See T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," *J. New Energy*, 3(2/3), 1998, p. 12-28; — *Energy from the Vacuum*, 2002, *ibid.* For nuclear transmutation at a distance (e.g., dudding all nuclear weapons by transmuting U235 and weapons grade plutonium), extension is required along the lines indicated by the present author in these references.

19. See M. Sachs, *ibid.*, also M. Sachs, "Relativistic Implications of Electromagnetic Field Theory," in T. W. Barrett and D. M. Grimes, eds., *Advanced Electromagnetism*, World Scientific, 1995, p. 551. The Sachs theory

already expresses the greatest generalization possible of electrodynamics, and fully justifies the preceding view.

20. See Whittaker, 1903, *ibid.;* — "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," *Proc. Lond. Math. Soc,* Series 2, Vol. 1, 1904, p. 367-372. The latter paper was published in 1904 and orally delivered in 1903. Per these two Whittaker papers, it is known that (i) ordinary EM fields and waves are the result of interference of two or more scalar potential functions, (ii) scalar potentials are comprised of harmonic sets of bidirectional EM phase conjugate wavepairs, (iii) all ordinary EM fields and waves and potentials have LW EM dynamics internally, and (iv) in theory, this "infolded" electrodynamics can be engineered. Sachs' theory dramatically extends the richness of that internal electrodynamics ignored in the West, and highlights the general relativity (sets of spacetime curvatures, or "engines", involved).

21. E.g., *Marine Observer* was a rich source for Russian energetics weapons testing in remote ocean areas, sighted by ships' captains and crews and entered into the ships' logs.

22. Portions of Khrushchev's statements to the Presidium are given in Max Frankel, "Khrushchev Says Soviet Will Cut Forces a Third; Sees 'Fantastic Weapon'," *New York Times,* Jan. 15, 1960, p. 1.

23. See Whittaker 1903 and 1904, *ibid.,* for the original basis of scalar interferometry and its creation of any kind of EM field, potential, and wave in the interference zone.

24. Also see M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry, *J. New Energy,* 4(3), Winter 1999, p. 76 for mathematical proof of scalar interferometry. There are several other AIAS papers in the same issue of *JNE,* dealing with interferometry and with the implications of the Whittaker papers.

25. As an example, the U.S.S. Skylark, surface companion of the Thresher, was beset by mysterious "jamming" of multiple types of EM systems in multiple bands. So bad was it that it required more than an hour to get an emergency message to Naval headquarters that the Thresher was in serious difficulty. Later, as the scalar interferometry died away, all the EM systems aboard the Skylark resumed normal functioning. The Skylark was in fact in a "splatter zone" of the underwater interference zone around the Thresher.

26. Circa the end of 1989, the earlier longitudinal EM wave weapons on site in Russia were leased to a rogue Japanese group comprised of Aum Shinrikyo and Yakuza, who have been operating those weapon systems—on site in Russia—since then. Senator Sam Nunn's investigations revealed the presence

of the Aum Shinrikyo in Russia, to include establishing a university there. His investigators missed the Yakuza involvement and also missed the exotic weaponry aspects as one might expect.

27. I concluded from this remark that TEMPEST techniques are probably used at some nearby site not too distant from my home here in Huntsville, to read and record my work as I write it on the computer. That appears to be a good assumption.

28. As a small example of what can be done, see a series of papers by M. W. Evans, P. K. Anastasovski, T. E. Bearden et al, published in *J. New Energy*, 4(3), Special Issue, Winter 1999. The papers are: (a) "On Whittaker's Representation of the Electromagnetic Field in Terms of Two Scalar Potentials, Part I"; (b) "An Experimental Test of the Existence of Whittaker's *g* and/Fluxes in the Vacuum"; (c) "On Whittaker's Representation of the Classical Electromagnetic Field in Vacuo, Part II: Potentials Without Fields"; (d) "On Whittaker's F and G Fluxes, Part III: The Existence of Physical Longitudinal and Timelike Photons"; (e) "Representation of the Vacuum Electromagnetic Field in Terms of Longitudinal and Time-like Potentials: Canonical Quantization"; (f) "On Whittaker's Analysis of the Electromagnetic Entity, Part IV: Longitudinal Magnetic Flux and Time-Like Potential Without Vector Potential and Without Electric and Magnetic Fields"; (g) "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry"; and (h) "On Extending Whittaker's Theory, Part VI: Photons Without Fields and Vector Potentials".

29. As an example, one part of the strategic deception plan was devoted to convincing the West that Soviet psychoenergetics was nothing but a technical wrinkle upon the rather harmless Western parapsychology. This part of the program was completely successful for decades. Another very clever use of deception is to allow release of some real scientific advances that lead the West, but are not too important to the KGB in the weapons arena. The increasing emphasis on torsion fields is an example of this "deception by giving good information that directs the adversary upon a path one wishes him to take", effectively steering him away from the path one wishes him to avoid. Lately, however, there is an uproar amongst the academicians in the Russian Academy of Sciences over this very point, that torsion fields being presented as the "end all, be all" is simply disinformation.

30. With their deployment of QP weapons in 2002, the Chinese are also on the way to having this capability.

31. The multiply connected spacetime (MCS) or quantum potential weapons went on site circa latter 1989. A laboratory prototype had been tested against the U.S. air attack on Libya in April 1986. For details of the extremely anomalous jamming **of aircraft** and missile guidance and control systems in that

attack on Libya, see Mark Thompson, "Mixed signals may have misguided U.S. weapons," *The Washington Post,* Jan. 22, 1989, p. A-4.

32. One or more QP weapons also apparently induced the Gulf War Disease in MidEast forces in the Gulf War. We have previously explained the peculiar anomalies, how it was done, why the endemic natives did not contract the disease, why the French forces did not contract it but U.S., British, and Canadian forces did, etc. We briefly discussed quantum potential weapons in T. E. Bearden, *Gravitobiology: A New Biophysics,* In T. E. Bearden, *Gravitobiology: A New Biophysics,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1991.

33. (a) Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," *Jane's Defence Weekly,* 1(3), Apr. 7, 1984, p. 508 gives something of the two Soviet tests of this weapon in Afghanistan. In addition to the fierce blast of empty Dirac Sea 4-holes, these weapons produce a powerful time-polarized EM pulse in the area, producing a very sharp "jerk" in the rate of flow of time, (b) The mind-body coupling mechanism—which cannot withstand such a sharp jerk in the rate of flow of time—has been given by T. E. Bearden, "Mind Control and EM Wave Polarization Transductions," *Explore,* 1999, *ibid.* This sharp "jerk" in time-flow instantly snaps the mind-body coupling mechanism for any living thing in the struck area. A very small version of this weapon was tested twice in Afghanistan, where the bodies in the "mind-snapped" zone died instantly, fell like limp rags, never twitched, and did not decay or get molested by insects or scavengers for the 30 days it required the Mujahedin to fight their way back into the two struck villages. The Mujahedin thought it was some new, extremely fast acting nerve gas, which they dubbed "smerch" or "death wind".

34. Two Ph.D. scientists were working closely with me on this entire effort, as was another close colleague. There was much more involved in the effort that is so briefly mentioned here.

35. Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

36. T. E. Bearden, *Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications,* CTEC Proprietary, Mar. 21, 1998, 200+ page inclosure to CTEC Letter to Gen. (Ret.) Walter Busby, Deputy Secretary of Defense for Counterproliferation and Chemical and Biological Defense, March 21, 1998. Copies were sent to various other U.S. government agencies, including to NIH, CDC, the USAF, etc.

37. See Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian], Novosibirsk, 1981. An English exposition of much of the Kaznacheyev material is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," *Psi Research,* 1(1), Mar. 1982, p. 47-76. The work of Vlail Kaznacheyev led to the understanding of how to electromagnetically transmit cellular death and disease into targeted cells (or bodies) at a distance, and involved thousands of lab experiments. Highly sanitized work revealing the results but not the mechanisms, were permitted in order to stimulate the West to see if it knew about energetics.

38. See Stanislov Lunev and Ira Winkler, *Through the Eyes of the Enemy: Russia's Highest Ranking Military Defector Reveals Why Russia Is More Dangerous Than Ever,* Regnery, Washington, D.C., 1998. The Red Chinese also were able to decipher the Russian EM disease transmission methodology, early on. As a strange kind of "poetic justice" action the Chinese induced cancer into many Russian spies stationed in the Russian Embassy in China. Quoting Lunev, p. 160: "...I *was not the only person who was stationed in Beijing to come down with cancer...Almost all of the GRU officers stationed in China with me have been diagnosed with cancer. People assigned to the embassy had told me that they were picking up high levels of radiation from the apartment building around the embassy compound. Whether it was a form of high-tech monitoring or a conscious effort to harm our health, it seems to have been successful. So much so that in 1992 the GRU was asked to investigate the high rate of cancer among embassy officials. The findings, however, have never been made public; in part, no doubt, because of the rapprochement between Russia and China."*

39. See my "Porthole Briefing", on www.cheniere.org. Also see "Le Probleme Priore," *Rapport de la Commission de I'Academie des Sciences a Monsieur le Ministre d'Etat charge de la Recherche et de la Technologie,* France, 1982, p. 1-22; J. B. Bateman, "Microwave Magic," Office of Naval Research London Conference report, ONRL C-14-77, 1977; —*A Biologically Active Combination of Modulated Magnetic and Microwave Fields: The Priore Machine,* Office of Naval Research, London, Report R-5-78, Aug. 16, 1978; — "Staging the Perils of Nonionizing Waves," *European Scientific Notes,* ESN 32-3-85-88, 1978; T. E. Bearden, "Energetics Update and Summary," Parts 1, *Explore More,* No. 1, 1996, p. 13-18; Part 2, *Explore,* 8(1), 1997, 53-56; Part 3, *Explore,* 8(2), 1997, p. 42-45; Christopher Bird, "The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science," Explore, 5(5-6), 1994, p. 97-110; R. Cambar, "Rapport general des travaux de la Commission de Controle constituee en vue de verifier l'un des effets biologiques obtenu par l'utilisation de l'appareillage de Priore A. Bordeaux," [General findings of the work of the control commission formed to verify one of the biological effects obtained by use of the apparatus of A. Priore at Bordeaux],

1969; R. Courrier, "Expose de M. le professeur R. Courrier Secretaire perpetuel de 1'Academie des Sciences fait au cours d'une reunion a l'lnstitut sur les effects de la machine de M. A. Priore 26 Avril 1977" [Presentation by Professor R. Courrier, Perpetual Secretary of the Academy of Sciences, made at the meeting of the Academy on the effects of the machine of M. A. Priore.]; Jean-Michel Graille, *Dossier Priore: A New Pasteur Affair,* De Noel, Paris, 1984 [in French]; Antoine Priore, "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968; — "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; — "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963; A. Priore, *Guerison de la Trypanosomiase Experimentale Aigue et Chronique par L'action Combinee de Champs Magnetiques et D'Ondes Electromagnetiques Modules* [Healing of intense and chronic experimental trypanosomiasis by the combined action of magnetic fields and modulated electromagnetic waves], thesis submitted in candidacy for the doctoral degree, 1973—rejected when the project was suppressed.

40. Eric Perisse, *Effets des Ondes Electromagnetiques et des Champs Magnetiques sur le Cancer et la Trypanosomiase Experimentale* [Effects of Electromagnetic Waves and Magnetic Fields on Cancer and Experimental Trypanosomias], Doctoral thesis, University of Bordeaux No. 83, March 16, 1984. Although Priore's own thesis was suppressed in 1973, Pautrizel nevertheless finally succeeded in getting this doctoral thesis by Perisse approved on the work, at the University of Bordeaux which had rejected Priore's thesis, after eleven more years had passed. Considering the viciousness of the suppression, this was a major accomplishment.

41. The photo was taken by George Suchary of the registration point (ball of light) high over the Nov. 1985 night shuttle launch from Cape Canaveral. For a photo of the same kind of "ball of light" registration point above the kill of a Titan missile launched from Vandenberg AFB in early 1986, see reference 39 below.

42. The picture by Canadian investigators showing the hole burned through the right fuselage ahead of the engines of the Arrow DC-8 is shown in T. E. Bearden, *AIDS: Biological Warfare,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org , 1988, p. 52.

43. See T. E. Bearden, *Gravitobiology: A New Biophysics,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1991, p. 82. Two sensors on-board the Arrow DC-8 independently recorded the sharp electrical spike generated by the strike of the EM missile.

44. "Titan Explosion Cripples U.S. Launch Surveillance Capability," *Aviation Week & Space Technology,* 124(17), Apr. 28, 1986, p. 16-19. Specifically see the picture of the explosion on p. 18. It clearly shows the Soviet marker beacon (little ball of light) well above the missile explosion. General background on the Titan explosion and investigation is given in "Investigators Search for Clues to Cause of Titan Blast," *Aviation Week & Space Technology,* 124(17), Apr. 28, 1986, p. 19-20.1 persuaded NBC engineer Ron Cole with a team of engineers to examine the videotape of the explosion, frame by frame. In 17 frames the ball-of-light (registration point) is visible, and moves away completely independent of the explosion. No "bogies on the range" were reported at Vandenberg AFB.

45. Dr. Evans is co-discoverer with Vigier of the B(3) field, now well established but still not in the textbooks. He is the Director of the Alpha Foundation's Institute of Advanced Study (AIAS) and has over 600 papers and books in the literature.

46. M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., "On Whittaker's F and G Fluxes, Part III: The Existence of Physical Longitudinal and Timelike Photons," *J. New Energy,* 4(3), 1999, p. 68-71; — "Representation of the Vacuum Electromagnetic Field in Terms of Longitudinal and Time-Like Potentials: Canonical Quantization," *ibid.*, p. 82-88; — "On Extending Whittaker's Theory, Part VI: Photons Without Fields and Vector Potentials," *ibid.*, p. 79-81.

47. Terence W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in Lakhtakia, A. (ed.): *Essays on the Formal Aspects of Electromagnetic Theory,* World Scientific Publishing, River Edge, NJ, 1993, p. 6-86; — "On the distinction between fields and their metric: the fundamental difference between specifications concerning medium-independent fields and constitutive specifications concerning relations to the medium in which they exist." *Ann. Fond. Louis de Broglie,* 14(1), 1989, p. 37-75; — and D. M Grimes. [Eds.] *Advanced Electromagnetism: Foundations, Theory, & Applications.* World Scientific, (Singapore, New Jersey, London, and Hong Kong), Suite 1B, 1060 Main Street, River Edge, New Jersey, 07661, 1995. In the latter book see T.W. Barrett, "Sagnac Effect: A Consequence of Conservation of Action Due to Gauge Field Global Conformal Invariance in a Multiply-Joined Topology of Coherent Fields," p. 278-313.

48. But see Craig F. Bohren, "How can a particle absorb more than the light incident on it?" *Am. J. Phys.,* 51(4), Apr. 1983, p. 323-327. By *resonating* the intercepting charged particle, its reaction cross section increases, and more EM energy is collected and re-radiated by the charge than is input to it by the conventional accounting (which only accounts the input Poynting energy that would be intercepted by a *static* charge). The conventional input energy

calculation therefore ignores the huge nondiverged Heaviside component surrounding the Poynting component and wasted. The resonant particle sweeps out past the usual Poynting component into the unaccounted Heaviside input component, and collects excess energy. The Bohren experiment with resonant charged particles produces COP =18. Metallic particles at ultraviolet frequencies are one class of such particles and insulating particles at infrared frequencies are another. For independent confirmation of the Bohren experiment, see H. Paul and R. Fischer, {Comment on "How can a particle absorb more than the light incident on it?'}," *Am. J. Phys.,* 51(4), Apr. 1983, p. 327.

49. E. T. Whittaker, 1903, *ibid.*

50. A beneficial use, e.g., would be to use such CSRs to investigate and report the interior of the sun and planets, including the Earth and its core. Since a CSR can also travel "through" the interior of a light beam, in theory such could be dispatched back into some of the nearer stars to also explore and report their interior.

51. This is no more curious than stating that sets of waves and their dynamics can exist in space, without interacting with mass, and that they can then later interact with a mass system to change it in specifically designed manner. Since longitudinal EM waves and their dynamics are "infolded" inside ordinary EM waves, fields, and potentials, they are also infolded "inside" the ambient vacuum potential, which itself is a part of the local vacuum potential. In short, one forms highly complex LW wavesets and functional dynamics, in essentially a sort of "subspace" as popularized by the Star Trek television series. The LW "engines" are identically ST curvature engines. A close colleague of the present author has in fact already invented and rigorously demonstrated such a communication system, but his former funding by several very large communications companies collapsed before he could develop and employ this system for the commercial market.

52. This is easily understood if one shifts one's vision from "space" to "subspace." Subspace is filled with (literally is composed of) longitudinal EM waves and their dynamics. The CSR's are quite unique, enormous systems "singing away" with longitudinal EM waves. So a longitudinal EM wave interferometer can easily locate them by sweeping its beams through subspace, as a radar sweeps to track aircraft in space. Once the CSR target "pings", the system automatically hits it with a profusion of interacting LWs, completing disrupting and destroying its dynamics. Scratch one system. Probably several million such systems were destroyed throughout the West (command centers, power grids, submarines, nuclear bombers, etc.) in rapid fire order, all without any external "spatial" sign other than perhaps a wee bit of very low level "EM noise" briefly as a struck CSR fractured.

53. See M. Sachs, *Quantum Mechanics from General Relativity,* Reidel, 1986. Sachs'unified field theory, e.g., applies equally to the quarks and gluons, as it does to the nucleons, the atoms, molecules, and macroscopic objects and phenomena. See also M. Sachs, "Symmetry in Electrodynamics: From Special to General Relativity, Macro to Quantum Domains," in *Modern Nonlinear Optics,* Second Edition, 3 vols., M. W. Evans (ed.), Wiley, 2001; vol. 1, p. 677-706.

54. G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.,* 89(5), 29 July 2002, 050601.

55. The tragic death of Captain Hess was also very probably such a test. Hess, perfectly normal and happy, laid out his uniform in the early morning and then went for his customary morning jog. He suddenly veered aside, sat down in a path and stabbed himself 26 times, all deep stab wounds, four of which would be totally incapacitating (two right through the heart) to any normal individual experiencing normal shock, etc. To a psychoenergetically controlled individual where shock does not occur nor does emotion or pain, this is understandable. Add this to the incidents involving Captains Button and Svoboda, and one begins to see the picture.

56. W. Heitman and G. Nimtz, "On Causality Proofs of Superluminal Barrier Transversal of Frequency Band Limited Wave Packets," *Phys. Lett. A196,* 1994, p. 154-158. Aichmann and Nimtz and co-workers at the University of Cologne have transmitted 8.7 GHz microwaves (free space wavelength 3.4 cm) traveling in a rectangular wave-guide that contains a "barrier" section of reduced dimensions. They have recently transmitted Mozart's 40th Symphony as frequency modulated microwaves through a barrier wave-guide at an FTL group velocity of 4.7 c, receiving audibly recognizable music from the microwave photons that survived their barrier passage. The transit time through the barrier was about 81 picoseconds and was observed to be constant for barriers with widths varying from 4.0 cm to 11.4 cm.

57. Quantum mechanics in fact requires instantaneous transmission of effects. A mishmash of learned contention would have us believe one can "transmit" energy instantly, but not intelligence. This argument has essentially stymied our scientific establishment so that it has not really funded strong investigation in this area. It is a stupid argument: Simply transmit one change, then a second one, then a third one, in serial order, where the "order" contains the intelligence. No different from Morse code. Somehow the pundits so certain that intelligence cannot be superluminally transmitted seem to have forgotten what "intelligence" is. They also seem to have assumed that one superluminal change can be **transmitted** on order, but a second, third, and fourth cannot.

58. Feynman and Wheeler tried very hard to formulate a new absorber theory of electrodynamics to eliminate many of the fierce foundations problems. Wheeler described the effort this way: *"...we may summarize the account of radiation reaction given by Wheeler and Feynman in 1945. In it every charged particle is envisaged as coupled to every other charged particle by a field that is symmetric in time: half advanced, half retarded. Interconnections run forward and backward in time in such numbers as to make an unbelievable maze. That weaving together of past and future seems to contradict every normal idea of causality. However, when the number of particles is great enough to absorb completely the signal starting out from any source, then this myriad of couplings adds up to a simple result: the familiar retarded actions of everyday experience, plus the familiar force of radiative reaction with its familiar sign."* [John Archibald Wheeler, "Law without law," in *Quantum Theory and Measurement.*, preprint, p. 39].

Unfortunately, when the foundations of any science are closely examined in their roots, one runs into the age-old bane of the philosopher: ultimately opposites are identical, at least prior to the observation (separation) action to separate them. It follows that all scientific reasoning is circular at its most primary basis.

59. See Charles W. Misner, Kip S. Thorne and John Archibald Wheeler, *Gravitation,* W. H. Freeman, San Francisco, CA., 1973, p. 71. Wheeler spoke of that inescapable fact in this manner: *"Here and elsewhere in science, as stressed not least by Henry Poincare, that view is out of date which used to say, 'Define your terms before you proceed.' All the laws and theories of physics, including the Lorentz force law, have this deep and subtle character, that they both define the concepts they use (here B and E) and make statement about these concepts. Contrariwise, the absence of some body of theory, law and principle deprives one of the means properly to define or even to use concepts. Any forward step in human knowledge is truly creative in this sense: that theory, concept, law, and method of measurement—forever inseparable—are born into the world in union."*

60. Also see T. E. Bearden, *Energy from the Vacuum: Concepts and Principles,* 2001, *ibid,* for a rework and extension of Aristotelian logic to eliminate its present fatal flaw, extending it into a 5-law logic.

61. The very notion of "propagation of energy through space" is radically extended in the extended unified EM and GR being pursued by AIAS and Sachs. Time-like EM energy currents dominate in the vacuum. E.g., see M. W. Evans et al., "Representation of the Vacuum Electromagnetic Field in Terms of Longitudinal and Time-like Potentials: Canonical Quantization," *J. New Energy,* 4(3), Winter 1999, p. 82.

62. Multiwave interactions (between transverse EM waves) produce time-reversed EM waves (phase conjugate replicas of the signal wave input). That method time-reverses EM energy in waveform. To time-reverse EM energy in mass form, simply pump that mass with time-polarized EM waves. As an alternative, it may be pumped with longitudinal EM waves (which are always accompanied by time-polarized EM waves anyway), and the pumped mass is also pumped in the time-domain because of the presence of the time-polarized EM waves in the pumping. We previously filed an invention disclosure with the U.S. Patent Office dealing with how to do that.

63. Mostly studied by Becker. See (a) R. O. Becker, "The control system governing bone growth in response to mechanical stress," *J. Ark. Med. Soc,* Vol. 62, 1966, p. 404; (b) — and David G. Murray, "The electrical control system regulating fracture healing in amphibians," *Clinical Orthopaedics and Related Research,* Vol. unk, No. 73, Nov.-Dec. 1970, p. 169-198. This latter paper is the definitive technical exposition by Becker of his electrical control system ideas and findings. Good references for the educated layperson are (c) R. O. Becker, "A technique for producing regenerative healing in humans," *Frontier Perspectives,* 1(2), Fall/Winter 1990, p. 1-2; (d) Robert O. Becker and Gary Seldon, *The Body Electric: Electromagnetism and the Foundation of Life,* William Morrow and Company, New York, 1985; and (e) Robert O. Becker, *Cross Currents: The Promise of Electromedicine, the Perils of Electropollution,* Tarcher, St. Martin's Press, New York, 1990.

64. A provisional patent application has been filed on the process, and an overview of the method can be seen as the "porthole briefing" on my website www.cheniere.org.

65. (a) C. A. L. Bassett; R. O. Becker, and R. J. Pawluk, "Effects of electric currents on bone in vivo," *Nature,* Vol. 204, Nov. 14, 1964, p. 652-654. The complete electrical control system (conventional EM theory) for the healing of bone fractures is given in (b) R. O. Becker, *Clinical Orthopedics and Related Research,* Vol. 73, 1970, p. 169. A paper "putting it all together" is (c) R. O. Becker, J. *Bioelec,* Vol. I, 1982, p. 239.

66. F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984, Revised Edition 1993, under the heading "5.2 Covariant Quantization" and "5.3 The Photon Propagator" in Chapter 5. The longitudinal and scalar photon polarizations are not directly observable, but only in combination, where they manifest as the instantaneous Coulomb (i.e., electrostatic) potential interaction, *"which emerges as an exchange of longitudinal and scalar photons."* So the common old electrostatic scalar potential is actually comprised of longitudinal EM waves and time-polarized EM waves.

67. With the MidEast increasingly set to explode, we may be facing the unleashing of weapons of mass destruction, a vast shortage and curtailment of

available oil and energy, and a total collapse of the Western economy and therefore most of the world economy, in as little as three or four years, if the impending energy crisis worldwide is not resolved quickly. See T. E. Bearden, ADAS position paper, "The Unnecessary Energy Crisis: How to Solve It Quickly," *ibid.* As this is written, we have already had a warning shot across our bow: the vicious terrorist attacks on September 11, 2001. We are in Afghanistan and elsewhere in a war against terrorism, and anyone can see the escalation in the MidEast toward war between the U.S. and Iraq, and almost certainly a resulting explosion in the MidEast and unleashing of the hostile terrorists assets and mass destruction weapons already inside the United States. Since modern economies are based on cheap energy, and cheap energy is largely based on cheap oil, then with surging energy demands worldwide we shall see "oil wars" erupt and rage worldwide, with terrorist retaliations and attacks—using weapons of mass destruction—occurring in U.S. cities. *Unless energy from the vacuum is quickly developed and deployment begun at high speed, we risk the coming collapse of the world economy, great destruction of populaces, and a new Dark Ages.*

68. E. T. Whittaker, 1903, *ibid.*

69. Daniel A. Walker, Charles S. McCreery, and Fermin J. Oliveira, "Kaitoku Seamount and the Mystery Cloud of 9 April 1984," *Science,* Vol. 227, Feb. 8, 1985, p. 607-611; Daniel L. McKenna and Daniel A. Walker, "Mystery Cloud: Additional Observations," *Science,* Vol. 234, Oct. 24, 1986, p. 412-413.

70. Superpotential theory was initiated by E. T. Whittaker, extending earlier concepts advanced by G. Johnstone Stoney. See E. T. Whittaker, *Proc. Lond. Math. Soc,* Series 2, Vol. 1, 1904, p. 367-372. Nisbet extended the Whittaker and Debye two-potential solutions of Maxwell's equations to points within the source distribution. This is a full generalization of the vector superpotentials (for media of arbitrary properties) together with their relations to such scalar potentials as those of Debye. See A. Nisbet, *Physica,* Vol. 21, 1955, p. 799. Also, A. Ricci in 1901 introduced what might be called the magnetization potential, satisfying a certain equation, as an alternate to the Hertz vector. This was a part of early vector superpotentials. For an excellent overview discussion of superpotentials and related things, see Melba Phillips, "Classical Electrodynamics," in *Principles ofElectrodynamics and Relativity,* Vol. IV of *Encyclopedia ofPhysics,* edited by S. Flugge, Springer-Verlag, 1962. Phillips discusses, for example, that Whittaker 1903 (note: oral delivery; paper was published in 1904) was the first to prove that one can derive a general electromagnetic field from two scalar functions, which are really components of the vector superpotentials, with proper choices of the gauge functions. Whittaker's method is well known in the treatment of transverse electric and transverse magnetic modes of a cylindrical cavity or a waveguide. Nisbet has shown that this result may be extended within the source distribution (assumed

given) by a suitable choice of stream functions. The Debye potentials and the Bromwich potentials are essentially radial components of the vector potentials of which Whittaker potentials are the real parts. So in general the particular integral (the stream potentials) of the inhomogeneous Maxwell equations may be so chosen that the complementary function can be expressed in terms of only two scalars, which are components of the vector superpotentials. The Whittaker and the Debye-Bromwich potentials are special cases of two vector superpotentials.

71. David Jones, *Vancouver Sun Times,* Weekend Magazine, Dec. 17, 1977, p. 17.

72. For more specific information on the production of practical antigravity and an experiment demonstrating it, see T. E. Bearden, *Energy from the Vacuum: Concepts and Principles,* Cheniere Press, Santa Barbara, CA, 2002, Chapter 8. Approach to Antigravity.

73. T. E. Bearden, *Energy from the Vacuum: Concepts and Principles,* Cheniere Press, Santa Barbara, CA, 2002, Chapter 10. Cold Fusion: Low Spatial-Energy Nuclear Reactions at High Time-Energy. By inadvertently using highly compressed time-energy, cold fusion researchers are actually using—on the tabletop—higher energy density than our present high-energy physics uses in its largest accelerators. The conventional scientific community does not have a single scientific program dedicated to using time as energy, even when one's physics model can legitimately be made of a single fundamental unit. When the joule is chosen as that unit, then time becomes totally a function of energy, just as is mass. The use of mass-energy gave us the nuclear age; the use of time-energy will give us a far greater age, if we can ever get the scientific community off its duff and moving.

74. Further, the second law of thermodynamics has long been known not to apply to a single charge, atom, or molecule or to a small group. For such microscopic phenomena, there will occur reversed reactions a certain small fraction of the time. But also see G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.,* 89(5), 29 July 2002, 050601. These researchers have experimentally proven that the second law can be and is violated for much larger scales—up to micro (colloidal) size—and for much longer times (up to two seconds). Hence it is now proven that such "reversed zones" where reactions run backwards for quite a bit, can and do in fact occur on the colloidal scale, sufficient for the formation of large clusters of like charges. Evans and his colleagues have pointed out that these results dramatically affect chemistry.

75. Shoulders has been creating and using such charge clusters for years. See Kenneth R. Shoulders, "Energy Conversion Using High Charge Density,"

U.S. Patent # 5,018,180, May 21, 1991; — "Energy Conversion using High Charge Density," U.S. Patent No. 5,123,039, 1992; — "Method and Apparatus for Production and Manipulation of High Density Charge," U.S. Patent #5,054,046, 1991; — U.S. Patent #5,054,047, 1991; — "Circuits Responsive To and Controlling Charged Particles," U.S. Patent 5,148,461, 1992; — and Steve Shoulders, "Observations on the Role of Charge Clusters in Nuclear Cluster Reactions,"J. *New Energy,* 1(3), Fall 1996, p. 111-121.

76. D. J. Evans and D. J. Searles, *Phys. Rev. E,* Vol. 50, 1994, p. 1645.

77. See Donald Kennedy, "To Publish or Not to Publish," *Science,* Vol. 295, 8 Mar. 2002, p. 1793.

78. R. P. Taleyarkhan, C. D. West, J. S. Cho, R. T. Lahey Jr., R. I. Nigmatulin, and R. C. Block, "Evidence for Nuclear Emissions During Acoustic Cavitation," *Science,* Vol. 295, 8 Mar. 2002, p. 1868-1873. Quoting from the abstract: "In cavitation experiments with deuterated acetone, tritium decay activity above background levels was detected. In addition, evidence for neutron emission near 2.5 million electron volts was also observed, as would be expected for deuterium-deuterium fusion. Control experiments with normal acetone did not result in tritium activity or neutron emissions."

79. Yuri T. Didenko and Kenneth S. Suslick, "The energy efficiency of formation of photons, radicals and ions during single-bubble cavitation," *Nature,* Vol. 418, July 25, 2002, p. 394-397.

80. E.g., see John Archibald Wheeler and Richard Phillips Feynman, *Phys. Rev.* 59, 1941, p. 683; —"Interaction with the Absorber as the Mechanism of Radiation," *Rev. Mod. Phys.,* 17(2, 3), Apr.-Jul. 1945, p. 157-181.

81. Albert Einstein, "Foreword," in Max Jammer, *Concepts of Space: The History of Theories of Space in Physics,* Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

82. Robert A. Fisher, [Ed.], *Optical Phase Conjugation,* Academic Press, NY, 1983, p. xv.

83. Basically, the choice between distant heating in the interference zone, or distant cooling, is as simple as how one biases the ground potential used for the electrical grounds of the interferometer transmitters. Biasing well above the ambient potential in the distant IZ produces divergent EM energy (heat) forming and radiating in the distant zone. Biasing well below the ambient IZ potential produces convergent EM energy (cold) forming and radiating in the distant zone.

84. Mass is mostly empty space filled with EM potentials, fields, and waves. Since these potentials, fields, and waves are simply bundles of longitudinal EM waves and their impressed dynamics, it follows that mass itself

is mostly a "giant superhighway" for the transmission of LWs. This is actually a sort of "subspace" transmission "right through" matter, so to speak.

85. As demonstrated by such dual serial usage by a giant interferometer in the anomalous and explosive "cloud" phenomena near the Kurils in April 1984. For excellent descriptive articles and a hard look at the phenomena, see Walker et al., *Science, ibid.*

86. See "Soviet device uses radio waves as tranquilizer," AP release, *Huntsville [Alabama] Times,* May 19, 1983. See also W. Ross Adey, "Biological effects of low energy electromagnetic fields," *Proc. NATO Adv. Study Inst. on Advances in Biological Effects and Dosimetry of Low Energy Electromagnetic Fields,* Plenum Press, New York, 1983, p. 359-391. Adey apparently was aware only of ordinary U(l) electrodynamics. He describes the **LIDA** device as follows: *"Soviet investigators have also developed a therapeutic device utilizing low frequency square wave modulation of a radiofrequency field. This instrument know as the LIDA, was developed by L. Y. Rabichev and his colleagues in Soviet Armenia, and is designed for the 'treatment of neuropsychic and somatic disorders, such as neuroses, psychoses, insomnia, hypertension, stammering, bronchial asthma, and asthenic and reactic disturbances.' It is covered by U.S. Patent #3,773,049. In addition to the pulsed RF field, the device also delivered pulsed light, pulsed sound, and pulsed heat. Each stimulus train can be independently adjusted in intensity and frequency. The radiofrequency field has a nominal carrier frequency of 40 MHz and a maximum output of approximately 40 watts. The RF field is applied to the patient on the sides of the neck through two disc electrodes approximately 10 cm. in diameter. The electrodes are located at a distance of 2-4 cm from the skin. Optimal repetition frequencies are said to lie in the range from 40 to 80 pulses per minute. Pulse duration is typically 0.2 sec. In a 1-year trial period, the instrument was used on 740 patients, including adults and children. Positive therapeutic effects were claimed in more than 80 percent of the cases treated."* We accent that this is one of the simplest Russian psychoenergetics devices, obsoleted very early on, and thus released very early for public use.

87. Since U.S. researchers are unaware of the EM wave polarization transduction mechanisms, and in general have no instruments for detecting longitudinal EM waves, the specific transduction effects on humans would not be measured by U.S. laboratories. Hence the U.S. scientists could only think of it in conventional electrodynamic terms, and remained totally unknowledgeable about the specific transduction mechanisms in the human body producing time-polarized EM waves and effects from the device.

88. Max Frankel, "Khrushchev Says Soviet Will Cut Forces a Third; Sees 'Fantastic Weapon'," New *York Times,* Jan. 15, 1960, p. 1.

89. For a simplified introduction to psychoenergetics and the mind-matter coupling mechanism, see T.E. Bearden, "Mind Control and EM Wave Polarization Transductions, Part I", *Explore,* 9(2), 1999, p. 59; Part II, *Explore,* 9(3), 1999, p. 61; Part III, *Explore,* 9(4,5), 1999, p. 100-108.

90. One of the serious goals of the die-hard KGB psychoenergetics scientists appears to be to achieve the capability of directly engineering the entire human collective unconscious mind. The plan envisioned is to convert the human species into a sort of "giant ant species" where each one has perfect, machine-like niche and function (except, of course, for the elite Communist masters). In short, they conceived back there in the 1960s that such a "perfect communist state" for all humanity was in fact possible to achieve technically, with sufficient development of psychoenergetic technology. They are apparently well along the way toward achieving that goal. [The U.S. has not even conceived such a scientific possibility, and modern cognitive science in the West seems bent on the rather astounding proposition that mind itself does not even exist!]. We point out that no type of physical defense is even possible against such a capability engineered into the collective human unconscious, once it is achieved. The only "counter" is to use the same technology (it has to first be developed) to "re-engineer" what will have been done. But our fellows have not even conceived the possibility of the threat, much less the cause or the counter. Hence they continue to totally misunderstand what happened to Captains Button, Svoboda, and Hess in successful mind-control tests achieved perfectly at great distance.

91. To see the nature of the time-density wave (which is basically a longitudinal EM wave along the time axis), due to the perfect correlation between time-energy flow currents and variations with respect to longitudinal EM waves, see T.E. Bearden, "Giant Negentropy from the Common Dipole," 2000, *ibid.* Not stressed in the paper is the fact that the same dipolar negentropy with correlated EM time-energy currents and EM spatial energy currents, exists in the "individual" photon. Further, they can occur in coupled photon-antiphoton pairs, yielding the graviton. For time-polarized EM photons, it appears that two photon-antiphoton pairs are coupled, yielding the "supergraviton." Elsewhere we have briefly and addressed these novel formations as hypotheses.

92. Induction of specific diseases and health changes in the U.S. Embassy in Moscow over four decades, included causing the deaths of three U.S. Ambassadors. It was accomplished by implementing desired disease-pattern "engines" infolded inside field-free potentials. The Johns Hopkins study of the radiation patterns revealed a most significant finding: All the affected persons were in regions of the Embassy where the Russian microwave radiation *fields* were absent. In short, where the field-free—and thus very stable—potentials existed. Since these scientists (and most others in classical EM) consider that EM effects are caused only by fields, they concluded (totally erroneously) that

these health changes could not have been caused by the microwave radiation. To the contrary, they had just clearly proven that the radiation was indeed the cause. They showed a 100% correlation to the *presence of field-free potentials,* and a 100% anticorrelation to where the fields were present. This was for induced phenomena over several years. Had the radiation not been the culprit, then some health changes would have occurred where the fields were present, and some would have occurred where they were absent. That is not what happened at all, and this inexplicable logical non sequitur against common statistical inference was the accepted decision. The truly revealing aspect was that *apparently the entire U.S. Intelligence community had no competence or capability in higher group symmetry electrodynamics.*

93. E.g., see Lewis H. Ryder, *Quantum Field Theory,* Cambridge University Press, 2nd Edition, 1996, p. 147+ which covers the four photon polarizations. See also F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984, Revised Edition 1993, under the heading "5.2 Covariant Quantization" and "5.3 The Photon Propagator" in Chapter 5.

94. We elaborate on this somewhat, in "Mind Control and EM Wave Polarization Transductions", *ibid.*

95. Again, we point out the unified field approach of Mendel Sachs as probably being the best Western model in which to grapple with these phenomena.

96. David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of'Hidden' Variables, I and II." *Phys. Rev.,* 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II).

97. T. E. Bearden, *Gravitobiology: A New Biophysics,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1991.

98. W. A. Rodrigues Jr. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 < v < \infty$ in nature," *Found. Phys.,* 27(3), 1997, p. 435-508. A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory web site. It includes corrections to the published version.

99. W. A. Rodrigues Jr. and J. Vaz Jr., "Subluminal and Superluminal Solutions in Vacuum of the Maxwell Equations and the Massless Dirac Equation," *Adv. Appl. Clifford Algebras,* Vol. 7(S), 1997, p. 457-466.

100. E.g., see A. Enders and G. Nimtz, *Phys. Rev. Lett.,* Vol. 48, 1993, p. 632. Enders, Aichmann, Nimtz and co-workers at the University of Cologne have transmitted 8.7 GHz microwaves (free space wavelength 3.4 cm) traveling in a rectangular wave-guide that contains a "barrier" section of reduced dimensions. They transmuted Mozart's 40th Symphony as frequency modulated

microwaves through a barrier wave-guide at an FTL group velocity of 4.7c, receiving audibly recognizable music from the microwave photons that survived their barrier passage.

101. J. D. Jackson, *Classical Electrodynamics,* 2nd Edition, Wiley, New York, 1975, p. 222.

102. See H. A. Lorentz, *Vorlesungen Uber Theoretische Physik an der Universitat Leiden, Vol. V, Die Maxwellsche Theorie (1900-1902),* Akademische Verlagsgesellschaft M.B.H., Leipzig, 1931, "Die Energie im elektromagnetischen Feld," p. 179-186. Figure 25 on p. 185 shows the Lorentz concept of integrating the energy flow vector around a closed cylindrical surface surrounding a volumetric element. This is the procedure which arbitrarily selects only a small component of the total energy flow in space outside a circuit and associated with it—specifically, the small Poynting component diverged by the surface charges into the conductors to power the electrons—and then treats that tiny component as the "entire" vast energy flow (which may be a trillion or more times greater in magnitude). Thereby Lorentz arbitrarily discarded all the huge *Heaviside* energy transport component that does not strike the circuit at all, is not diverged into the conductors, and is just wasted. See Oliver Heaviside, *Electrical Papers,* Vol. 2, 1887, p. 94. Quoting: *"It [the energy transfer flow] takes place, in the vicinity of the wire, very nearly parallel to it, with a slight slope towards the wire... . Prof. Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from the vertical. This difference of a quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature of the electric field in the vicinity of a wire supporting electric current. The lines of electric force are nearly perpendicular to the wire. The departure from perpendicularity is usually so small that I have sometimes spoken of them as being perpendicular to it, as they practically are, before I recognized the great physical importance of the slight departure. It causes the convergence of energy into the wire."*

103. When it is associated with positive energy fields and potentials from positive mass, positive energy source charges, we have nominated that long-unaccounted Heaviside energy flow component as the source of the excess gravity holding the arms of the spiral galaxies together (a solution to the dark matter problem). When it is associated with negative energy fields and potentials from negative-energy Dirac Sea holes in 4-space, we have nominated that unaccounted energy flow component as the source of the excess antigravity responsible for accelerating the expansion of the universe (a solution to the dark energy problem).

104. About a year from now, Bedini and I will also release a very novel use of this Heaviside energy flow component in inverted circuits—circuits that

operate almost completely in reverse from normal textbook circuits. This is a totally new kind of circuit that we hope will revolutionize power systems and electronic control systems. Until our intellectual property rights are secured, we are not releasing any details at all.

105. Terence W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," *Ann. Fond. Louis de Broglie,* 16(1), 1991, p. 23-41. Tesla was able to shuttle energy around in his circuits at will. Barrett went on to improve the process and patent two versions. See T. W. Barrett, "Active Signalling Systems," U.S. Patent No. 5,486,833, Jan. 23, 1996. This is a signaling system in time-frequency space for detecting targets in the presence of clutter and for penetrating media. See also T. W. Barrett, "Oscillator-Shuttle-Circuit (OSC) Networks for Conditioning Energy in Higher-Order Symmetry Algebraic and RF Phase Conjugation," U.S. Patent No. 5,493,691. Feb. 20, 1996.

106. Bryan G. Wallace, "The Unified Quantum Electrodynamic Ether," *Found. Phys.,* 3(3), Sept. 1973, p. 381-388.

107. William G. Harter, John Evans, Roberto Vega, and Sanford Wilson, "Galloping waves and their relativistic properties," Am. J. *Phys.,* 53(7), July 1985, p. 671-679. Most waves caused by a single frequency source have a galloping motion. The only exceptions are the very special cases of absolutely pure moving waves or absolutely pure standing waves. The superposition and resulting interference of two oppositely moving waves generally yield a wave having nonuniform motion; in short, the wave rushes forward rapidly, then slows appreciably, etc. A relativistic example is given in which the source frame observes a very nonuniformly galloping wave whose velocity varies between 0.01 c and 100 c, even though its average velocity is c. Relativistic properties of interfering electromagnetic plane waves are discussed.

108. S. L. Miller, "The Formation of Organic Compounds on the Primitive Earth," *Ann. NYAcad. Sci.,* Vol. 69, 1957, p. 260; — *Bioch. Biophys. Acta.,* Vol. 23, 1957, p. 480-489; — *Science,* Vol. 117, May 15, 1953, p. 528-529; — and H. C. Urey, "Organic Compound Synthesis on the Primitive Earth," *Science,* Vol. 130, 1959, p. 245-251.

109. R. Harada and S. Fox, "Thermal Synthesis of Natural Amino-Acids from a Postulated Primitive Terrestrial Atmosphere," *Nature,* 201(4917), 1964, p. 335-336; S.W. Fox, "How Did Life Begin?", *Science,* 132(3421), July 22, 1960, p. 200-208; — and S. Yuyama, "Abiotic Production of Primitive Protein and Formed Microparticles," *N.Y. Acad. Sci. Ann.,* Vol. 108, June 1963, p. 487-494; — and K. Harada, "Thermal Copolymerization of Amino Acids to a Product Resembling Protein," *Science,* 128(333), 1958, p. 1214; — and M. Middlebrook, "Anhydrocopolymerization of Amino Acids Under the Influence of Hypothetically Primitive Terrestrial Conditions," *Federation Proc.* Vol. 13, Mar. 1954, p. 211.

110. H. C. Urey, "On the Early Chemical History of the Earth and the Origin of Life," *Proc. Nat. Acad. ScL,* 38(4), 1952, p. 351-363.

111. I.e., one may speak of the "potential" of a specific engine of a former life form, and then realize that the potential for that form reduces in magnitude inversely as the radial distance from the Earth. However, it does not reduce to zero until an infinite distance is reached.

112. As an example, a great EM energy flow—produced by the cross product ExH of the earth's magnetic field and the earth-to-electrosphere electric field—flows from West to East around the earth, nearly parallel to the equator. Infolded in that great EM energy flow are all the structures of all the genetics and life forms on earth, or that have ever been on earth and conditioned the natural scalar potential of the earth. The vials in the laboratory experiments of Miller, Fox, and Urey were all continually bathed in that great energy flow, which within it contained the entire life structure and dynamics of the planet. So no "inert" experiments were ever performed, contrary to the interpretations of the scientific community.

113. Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian], Novosibirsk, 1981 [an English exposition of much of the Kaznacheyev work is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," *Psi Research,* 1(1), Mar. 1982, p. 47-76.] Kaznacheyev, e.g., appears to have headed the Soviet team actually producing the basis for the disease-inducing technology utilized for four decades in the microwave radiation of the U.S. Embassy in Moscow.

114. Specifically, the down payment was $900 million in gold. The lease is probably about $1 billion per year. Thus the Yakuza acquired very powerful strategic weapons, and became a secret power to be reckoned with internationally and strategically. Since 1990, weather engineering over the U.S. and occasional shootdown of aircraft (such as the TWA-800) have been by rogue Japanese fingers on the scalar interferometer weapon triggers, on site in Russia.

115. The Yakuza has also transferred some of the longitudinal EM wave interferometry technology to their Japanese facilities. One such system has apparently been deployed in the power grid, and has been used to kill some government officials who opposed the Yakuza. This information was obtained by some brave Japanese risking their lives to get the information to the present author.

116. Many important KGB weapon tests happened in April of various years; the KGB weapons folks were simply preparing for their May Day reports, and each wished to have a good one.

117. Sen. Sam Nunn's investigation teams verified that the Aum Shinrikyo were actively seeking the "Tesla Weapons", with the Aum Shinrikyo even visiting the Tesla Museum in Beograd for that purpose. We have 100% certain confirmation of the Yakuza and Aum Shinrikyo leasing of the scalar interferometers from the KGB and Russian Mafia (working for the KGB) at the end of 1989 or first few days of 1990. Sen. Nunn's report confirms that the Aum Shinrikyo set up a University in Moscow. Our comment is that this is where the Russians actually taught the Japanese crews the science and technical functions so that they could successfully man and use the scalar (longitudinal EM wave) interferometry weapons.

118. Recall that in the 1950s and early 1960s, Kaznacheyev et al., *ibid.,* had already clearly proven that nearly any form of cellular change, damage, or disease could be induced in cells at a distance by purely electrodynamic means.

119. In the kill of the Arrow DC-8 at Gander, Newfoundland on 12 Dec. 1985, the struck aircraft did charge up with time-density charging, and many pieces of the wreckage did indeed emit longitudinal EM waves thereafter for some time. Consequently, the Canadian recovery crews working the site and handling the materials, were exposed to longitudinal EM wave radiation for varying degrees of exposure. It follows that many of those on-site handlers should have later come down with physical ailments, with symptoms such as flu-like features, weakness, heart palpitations, headaches, dizziness, disturbance of equilibrium, and with various infections because of weakened immune systems. They also should have had weakened cellular regenerative (healing) systems, so their physical problems would endure and largely remain. About five dozen of the recovery personnel did come down with those exact symptoms, and many remained disabled or debilitated for some years.

120. We have also explained (a) why the French forces in the Gulf War did not get GWS, while American, British, and Canadian forces did, (b) why the native population did not exhibit GWS, and (c) the remaining anomalous phenomena associated, such as the semen "infecting" the wife, the children catching similar or related conditions from sustained close proximity to their father's biofields (transmission of engines), etc.

121. Mario Bunge, Foundations of Physics, Springer-Verlag, New York, 1967, p. 176.

122. T. E. Bearden, "Giant Negentropy from the Common Dipole," J. *New Energy,* 5(1), Summer 2000, p. 11-23 gives the solution to the long-vexing source charge problem. The paper is also carried on DoE restricted website http://www.ott.doe.gov/electromagnetic/ and on www.cheniere.org

123. Two examples of inane articles and actions in this respect, revealing no knowledge that every charge in the universe already exhibits COP = oo, are (a)

Graham P. Collins, "There's No Stopping Them: Perpetual Motion is alive and well at the U.S. Patent Office," and (b) Resolution by the Executive Board at the June 2002 meeting of the American Physical Society (APS) as follows: *"The Executive Board of the American Physical Society is concerned that in this period of unprecedented scientific advance, misguided or fraudulent claims of perpetual motion machines and other sources of unlimited free energy are proliferating. Such devices directly violate the most fundamental laws of nature, laws that have guided the scientific progress that is transforming our world."* So the learned APS has just declared the source charge as an unacceptable "free energy" machine and perpetual motion machine, since the charge simply sits there and continuously pours out real observable EM energy in all directions at the speed of light, without any observable EM energy input. Thus the APS has implicitly declared all EM fields and potentials and their energy—which includes every joule of EM energy in the universe, in matter or in space—to be unacceptable byproducts of perpetual motion nonsense. We heartily call upon the APS in the name of science to enforce their resolution in the orthodox academic community, who are the greatest perpetual motion advocates in history. To be consistent with their own resolution, they must root out and destroy—from all texts and universities and from the scientific literature—any mention whatsoever of charge, field, potential, and EM energy. Either they must do that, or stand hoisted on their own petard as a dogmatic group still very much unaware of why the Nobel Prize was awarded to Lee and Yang in 1957 and what the proof—by Wu et al.—of the asymmetry of opposite charges means with respect to the source charge considered in its modern quantum field theory view as a dipolarity of a bare infinite central charge surrounded by clustering virtual charges (also infinite charge) of opposite sign in the active vacuum. Every electrical engineering department and professor and textbook still uses a model and teaches a subject assuming that every charge in the universe freely creates energy from nothing at all. Further, the APS in every annual meeting also tolerates, encourages, and publishes multiple papers and presentations with the same such unacceptable perpetual motion assumptions in them. Such is the lofty understanding of our present organized scientific community, which seems much more politically bent than scientifically bent. The simple Bohren experiment and the entire area of negative resonance absorption of the medium has long been established in nonlinear optics, and such experiments widely yield a nominal COP = 18. In other words, the experiments nominally output 18 times as much energy as the operator inputs, and the excess input energy is furnished absolutely for free from the active vacuum's exchange. In this case the Executive Board of the APS is not even aware of what its own physics discipline has long proven and accepted in particle physics, for nearly a half century. Neither are the self-appointed defenders of the faith, who also are totally unaware of what physics has already proven with a Nobel Prize awarded nearly a half century ago.

124. C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes and R. P. Hudson, "Experimental Test of Parity Conservation in Beta Decay," *Phys. Rev.*, Vol. 105, 1957, p. 1413. We also quote Nobelist T. D. Lee as follows: *"Since nonobservables imply symmetry, any discovery of asymmetry must imply some observable. The experiment of Wu, Ambler, Hayward, Hoppes and Hudson... established the asymmetry between the positive and negative signs of electricity."* [T. D. Lee, *Symmetries, Asymmetries, and the World of Particles,* U. Wash. Press, Seattle, 1988, p. 11.].

125. Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, New York, 1998, reprinted with corrections 1999.

126. See "Secret Speech: Did Brezhnev Come Clean?" *National Review,* 29(8), Mar. 4, 1977, p. 248, 250

127. For the energy description of this device and some results, see Floyd Sweet and T. E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," *Proc. 26th Intersoc. Energy Conversion Engineering Conf. (IECEC '91),* Boston, Massachusetts, p. 370-375.

128. Also see my book, *Energy from the Vacuum,* 2002, *ibid,* which includes photographs of the self-powering Sweet device, courtesy of Walter Rosenthal.

129. Roy Brewer was very active in Hollywood Unions against Communists when they were busily penetrating the Hollywood complex. He was a staunch anti-communist, fighting them out of control of the unions, way back when. Ronald Reagan was a personal friend of his of long standing, and Roy had open access to the White House when Reagan was President.

Roy briefed Reagan on the Russian scalar weapons, and Reagan believed him. Three times, Reagan ordered his Secretary of Defense to take action in that area, but neither of the three did so because the U.S. Intelligence Community did not believe it and adamantly opposed any such action. We took quite a beating in those days, for quite a long time (some years), before certain key parts of the Government finally believed us, checked it out experimentally, and found that it worked.

Roy also introduced me to the Duke de Grantsmesnil, one of the last of the great old British intelligence men, now unfortunately deceased. In the hectic 1997 days when we were trying desperately to warn a little friendly nation (and our own) against a coming great strategic strike on the U.S. and other allies, the Duke played a major role in introducing us directly to higher political levels in that little nation - which was the only nation on earth that could counter the pending strike. The first planned strike scheduled in February 1997 was successfully countered, only to be immediately rescheduled. In the second incident, the attack was scheduled for May 1, 1997. We finally realized why the hostile nation could reschedule and go ahead, and needed then to warn the little nation of the reason in some detail. We had tried everything to get certain information through, and all means had failed. A former ambassador of that nation to Great Britain was also assassinated during that struggle. Roy Brewer was one of the key fellows assisting in that attempt, as were other personal colleagues of mine who also were witnesses and direct participants in those desperate events. When all else failed, I finally just faxed the Prime Minister directly, in the clear, some five days before the attack, urging the most urgent and immediate action possible by the Prime Minister. A complete change in command and control of the weapon was required, and that change had to be made and then the hostile nation had to be notified that the change was in place (and probably shown that it was). The Prime Minister accepted the message and the necessary action was carried out immediately. So two days or so before the actual attack was scheduled, that great hostile armada began standing down. Oh, the hostile nation and its leaders could have easily destroyed the entire United States in two hours or so, without fail! But then their own country would have promptly disappeared from the face of the earth. So they stood down. And our own Intel community never even picked it up (they did get my messages alerting the U.S. Government, but thought it was a joke or something). They, of course, also intercepted my open fax to the Prime Minister, and so were all over me for that one for a while thereafter.

Interestingly, at the very time that great armada was just beginning to stand down, then Secretary of Defense William Cohen made the following statement at a conference in Georgia: "Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves...so there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts, and that's why this is so important." Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Ironically, the Secretary was referring to the earlier scalar interferometer weapons, and now the quantum potential weapons etc. He was completely unaware that a great strategic strike had actually been scheduled and was just being aborted as he was speaking those words.

Roy Brewer is one of those fellows who has been a great American, has served his country admirably during his lifetime, and contributed greatly to preventing a planned take-over by the Communists of the labor unions of America. In addition, he helped set up the contacts with the friendly **liltle foregn nation, so** that on four occasions planned attacks to destroy the United States have been successfully aborted or handled.

As you know, one of the great failures in the Communist plan to have America go communist was their failure with the labor unions. And we owe that failure to the steadfast efforts of some staunch folks like Roy Brewer. He helped save our nation when many of us were still wearing knee pants.

130. Wheeler's principle, as stated in W. Misner, K. S. **Thorne, and J.** A. Wheeler, *Gravitation*, W.H. Freeman and Co., San Francisco, 1973, p. 5.

131. With the new improvements we uncovered, treatment time can be reduced to 5 minutes per patient. Three such treatments one week apart would be what are required. Hence one machine station could treat a patient every five minutes, around the clock. With sufficient portable treatment units, mass treatment for mass casualties would become a practical reality.

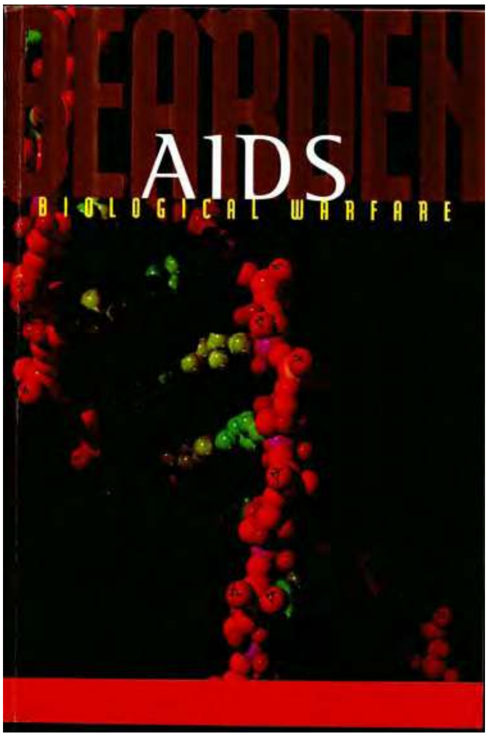

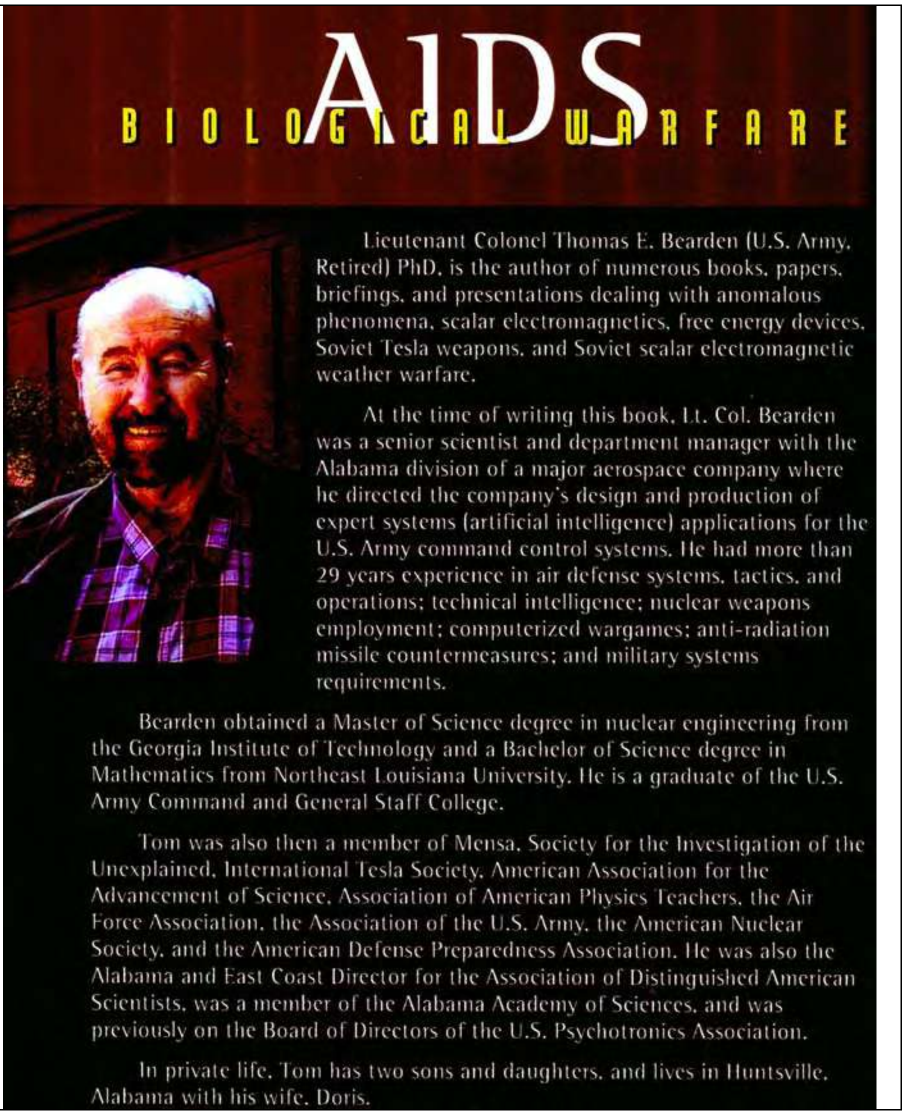

Lieutenant Colonel Thomas E. Bearden (U.S. Army, Retired) PhD, is the author of numerous books, papers, briefings, and presentations dealing with anomalous phenomena, scalar electromagnetics, free energy devices, Soviet Tesla weapons, and Soviet scalar electromagnetic weather warfare.

At the time of writing this book, Lt. Col. Bearden was a senior scientist and department manager with the Alabama division of a major aerospace company where he directed the company's design and production of expert systems (artificial intelligence) applications for the U.S. Army command control systems. He had more than 29 years experience in air defense systems, tactics, and operations; technical intelligence; nuclear weapons employment; computerized wargames; anti-radiation missile countermeasures; and military systems requirements.

Bearden obtained a Master of Science degree in nuclear engineering from the Georgia Institute of Technology and a Bachelor of Science degree in Mathematics from Northeast Louisiana University. He is a graduate of the U.S. Army Command and General Staff College.

Tom was also then a member of Mensa, Society for the Investigation of the Unexplained, International Tesla Society, American Association for the Advancement of Science, Association of American Physics Teachers, the Air Force Association, the Association of the U.S. Army, the American Nuclear Society, and the American Defense Preparedness Association. He was also the Alabama and East Coast Director for the Association of Distinguished American Scientists, was a member of the Alabama Academy of Sciences, and was previously on the Board of Directors of the U.S. Psychotronics Association.

In private life, Tom has two sons and daughters, and lives in Huntsville, Alabama with his wife, Doris.

Sweating done by red_gonzo.

Note: sometimes formatting of the text may be a little fuzzy. It's all because the OCR process didn't go as smoothly as I expected. Also there may be some errors inside text so be vigilant. In any case share this book, send emails to people about knowledge contained in it. This book should especially be interesting to any person in medical community (being as a doctor or as patient) as it provides factual evidence of successful (but supressed) treatments through electromagnetic means. But before any devices can be understood and developed a new and extended kind of physics must be introduced into our educational system. Only then can this proper foundation be learnt and used by applicative sciences such as medicine. So far such knowledge has been contained in highly classified military circles and hopefully now, after many decades of suppression, it will finally be used for the good of mankind.

Dedication

This book is sincerely dedicated to the millions of persons who have died needlessly of cancer, leukemia, and AIDS in the last two and half decades. They were men, women, and children. Some were weak, some were strong. Some died in agony, some died silently. All of them were our fellow human beings. We remember them. They were important. Their anguish cries out to be heard. God willing, the new electromagnetic medicine so long suppressed will flower, and save the millions otherwise yet to die.

Special Acknowledgements

Mr. Christopher Bird's kind permission to print, his summary of the Priore Affair is gratefully acknowledged, as is the contribution of his own extensive Priore research material to this effort. The author owes a great debt indeed to Chris Bird for his magnificent assistance.

Mr. Robert Whitney graciously contributed rare photographs of Antoine Priore's device and also translations of background papers by French scientific journals. His previous valiant though fruitless efforts to bring the Priore machine into the mainline of Western medical treatment are specifically recognized and commended.

Vital contributions by Drs. Robert B. and Theodore A. Strecker are particularly acknowledged. Their March 1986 alert to the federal government, that AIDS was a deliberate biological attack spawned by Soviet agents working in Western research laboratories, is specially pointed out.

Heroic efforts by Dr. John Seale and other scientists to alert the medical research community to the true source of AIDS is specifically recognized and gratefully acknowledged.

Continuing efforts by Dr. William C. Douglass to "tell it like it is" on the AIDS crisis is specially pointed out. His permission to extract and present an excerpt from his research, pointing out tested healing of viral infections by UV irradiation of the blood, is deeply appreciated.

The unyielding and continuing efforts of Dr. Eva Snead, to alert the national authorities to the fact that there is more to AIDS than meets the eye, is also particularly recognized and appreciated.

Decades of effort by John Crane to bring Royal R. Rife's work out of suppression are specifically recognized and appreciated.

Important scalar EM experiments by John Bedini and Frank Golden are fully recognized and deeply appreciated. Their experimental efforts over a period of years have profoundly influenced my own thinking and understanding.

As always, my great admiration and deep thanks go to my friend Hal Crawford, whose keen artistic vision and skilled hand have contributed key illustrations to my work for years.

Other important illustrations and artwork by Messrs. James Thomas Neumann and Lee Giles are gratefully acknowledged.

Photographs furnished by John Moray showing his father's pioneering free energy work are appreciated. T.H. Moray's genius and achievements will yet be recognized.

Support from the Association of Distinguished American Scientists and Mr. Richard J. Reynolds III is deeply appreciated. Without that support, this effort could not have been accomplished.

Important contributions from a host of other colleagues and persons are deeply appreciated. These persons include Frank Golden, John Bedini, Ken Moore, Tom McLaughlin, Joe Parrot, Randy Davidson, George Fencl, Andrew Michrowski, Stefan Possony, Tony Gideon, Electra Briggs, and many others too numerous to mention.

Finally, nothing that I have attempted could have been accomplished at all without the unflinching and loyal support of my beloved wife, Doris.

To all of you I owe a great deal. If this bizarre struggle to save humanity is ever won, it will largely be due to the significant contributions you have made.

Thomas E. Bearden
Jan. 17, 1988

TABLE OF CONTENTS

DEDICATION iii

SPECIAL ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS vi

LIST OF FIGURES xi

LIST OF TABLES xv

FOREWORD 1
- What This Book Is About 1
- Desperate Diseases, Desperate Remedies 3
- The "Dirty Dozen" 5
- An Analogy 6

THE VACCINE CONNECTION 8
- The "Monkey Bite" Theory 9
- Retroviruses and Recombinants 10
- Cancer Research Produced Ideal Biological Warfare Agents 11
- U. S. "Cancer War" Labs Were Infiltrated 12
- The Deed Was Easily Done 14
- It's Actually Documented 15
- Soviet Biological Warfare Background 17
- We Need An Immediate, Full-Scale Investigation 20

THE SOVIETS KNEW THE U. S. WOULDN'T REACT 24
- Administering the Coup-de-Grace 24
- Murky "Behind-the-Scenes" Intentions 26
- The USSR Has Repeatedly Verified Our Nonreaction 30
- The Asian Tiger Mosquito: A New and Efficient Vector 62

TABLE OF CONTENTS

WORLD WAR III IS RAGING AND WE'VE NEARLY LOST IT 66
- More On Aids Doubling Time 66
- World War III is Already Raging 68
- Communism is Savage Enough to Do Diabolical Things 68
- Western Medicine is Inadequate 69
- The Soviets Will Not Be Decimated 71

EXTRAORDINARY PHYSICS 74
- Maxwell's Lost Unified Field Theory 74
- Heaviside's Mutilation of Maxwell's Theory 83
- Building Upon Whittaker's Fundamental Work 86
- Symmetry and Parity 90
- Charge-Parity-Time and Negative Energy 97
- Free Energy and Antigravity 100
- Phase Conjugation (Time Reversal) 104
- The Phase Conjugate Mirror 114
- Amplifying the Phase Conjugate Effect 117
- The Phase Conjugate Mirror as a Vacuum Triode 122
- Newton's Third Law and the Full Law of Entropy 125
- Scalar Electromagnetics 134
- Scalar EM View of the Vacuum 139
- Charge, Potential, and Curvature of the Vacuum/Spacetime 151
- Twin Flows of Time 152
- Engineering Local General Relativity 154
- Engineering Antigravity 157
- Curved Spacetime and the Disintegrated Vacuum 159
- Separation of Vacuum and Observable States 160
- Partial Potentials, Curvature, and Warping 161
- Force and Mass are Inextricably Intertwined 164
- The Concepts of Zipped and Unzipped Forces 166
- Force, Force Field, and EM Waves in Vacuum 168
- Zero-Summed Vector Systems: Shortcoming in Vector Analysis 172
- Maxwell's Theory was Altered and Curtailed 174
- Scalar Waves and Polarization of the Vacuum 177
- Vector Zero Systems and the Kaluza-Klein Approach 180

TABLE OF CONTENTS

Electrogravitation Amplification Factor....186
Comparison of EM Concepts....187
The Aharonov-Bohm Effect....188
Curl-Free Magnetic Vector Potential (A-Field)....190
Fractional Charges, Magnetic Monopoles, and Magnetic Currents....192
Some Ehrenhaft Experiments....194
What is a Magnetic Monopole?....199
Supercharges and Fractional Charges....202

EXTRAORDINARY BIOLOGY....204
Kervran's Proof of Biological Transmutation....204
Biological Transmutation Has a History....207
Surplus-of-Energy Mechanisms Proposed By the U. S. Army....208
Alchemy and Unusual Critters....210
The Cell Also Lives and Functions in the Virtual State....213
The Kaznacheyev Experiments....214
Structuring and Charging a Biopotential....217
The Cell's Electromagnetic Breathing....218
The Summed Virtual Structures of Kaznacheyev "Death Photons" Physically Kindle the Disease....219
Electromagnetic Infection Results in Physical Disease....221
A Possible Cure For AIDS....221
The Mirror Cytopathogenic Effect and Factors Influencing It....223
Some Biological Warfare Implications....225
What Kaznacheyev Hid: The Role of Phase Conjugation....226
Popp's Master Cellular Communication System....228
Scalar EM Comment on Popp's Communication System....230
A New View of the Nature of Mind and Thought....231
Kindling, Life, Mind, and Negentropy....236
The Priore Machine and Phase Conjugation....247
Phase Conjugates of "Death Photons" are "Healing Photons"....249
The Effect is Universally Applicable....251
In the Mid-70's It Ended....252

TABLE OF CONTENTS

Work of Australians Reid and Barsamian, and the American G. Wilbanks ... 254
Concepts of Dr. Reid and Associates ... 259
Other Important Ideas ... 261
Wider Implications of Structured Field Information ... 262
Dr. Barsamian's Important Theoretical Insights ... 264
Royal Raymond Rife ... 266
Rife's Universal Microscope ... 267
Rife Proved That Everything Is Alive ... 272
Rife's Microscope Used Evanescent Waves ... 275
Rife Revealed a Far More Fundamental, Living Biology ... 277

DEVELOPING THE ELECTROMAGNETIC CURE FOR AIDS ... 280
Reversing the Kaznacheyev Effect ... 280
The Proof: Priore's Work ... 282
What Is Needed ... 284
Encouraging Preliminary Work ... 284
Conclusion ... 288

SELECTED BIBLIOGRAPHY ... 290
References Related to Chapters 1 and 2 ... 290
References Related to Chapter 3 ... 297
References Related to Phase Conjugation (Time Reversal) ... 303
References Related to William J. Hooper's Work ... 309
References Related to Ehrenhaft's Work ... 312
References Related to Louis Kervran's Work ... 314
References Related to Popp's Master Cellular Communication System ... 316
References Related to Kaznacheyev's Experiments ... 319
References Related to The New View of The Nature of Mind and Thought ... 324
References Related to Antoine Priore's Work ... 333
References Related to the Australian Work ... 339
References Related to Royal R. Rife's Work ... 341

TABLE OF CONTENTS

APPENDIX I: The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science (Christopher Bird) (Used by Permission) 346

APPENDIX II: Patents 376

APPENDIX III: A Conditional Criterion for Identity, Leading to a Fourth Law of Logic 428

APPENDIX IV: List of Selected Incidents 444

APPENDDIX V: Dead Man Fuzing: The Real Meaning of the Reykjavik Summit 484

LIST OF FIGURES

No. Title Page

1. Conventional science cannot do the job 4
2. The "Monkey-Bite" theory 9
3. The Soviets have delivered an intended knockout 15
4. The U.S. Embassy in Moscow 32
5. Decay of a radioactive nucleus 33
6. EGP fission of a radioactive nucleus 33
7. Bottled up energy in a standing scalar EM wave is a potential 34
8. EGP shortout 34
9. EGP initiation of the Chernobyl event 35
10. The U-2 high-altitude reconnaissance aircraft 36
11. The U. S. S. Thresher disintegrates 37
12. Soviet helicopters have used a "death ray" in Afghanistan 40
13. Mushroom cloud from the sea off Japan, near a Russian test area 40
14. The Tesla Shield, a giant hemispherical shell of energy 41
15. The Tesla shield: A 1969 Virgin Islands incident 41
16. Tesla shields: Two nested arcs of light in the sky 42
17. Tesla shield: An expanding hemispherical circle of light 42
18. Tesla shield: Hemisphere and globes, 1977 43
19. Large Tesla shield observed by Red Chinese fighter pilots 43
20. Large glowing "Tesla Globe," Red China 44
21. 1966 test of giant Tesla globe, expanding deep within the Soviet Union 45
22. A giant Tesla globe, deep within the Soviet Union in 1979 46
23. Engines disabled, a passenger jet drops six miles in two minutes 47
24. Giant "Woodpecker" beams intersecting over the U. S 48
25. Third Soviet non-destructive testing against shuttle launches Nov. 26, 1985 50
26. Death of the Arrow DC-8, Dec. 12, 1985 52

LIST OF FIGURES

No. **Title** Page

27. Anomalous hole in the right fuselage 52
28. **Part** of the Soviet engineering of U. S. weather, just prior to the death of the Challenger 53
29. Soviet trawler spyship 54
30. Launch of the doomed Challenger, Jan. 28, 1986 54
31. Loss of a Delta rocket, May 3, 1986 56
32. Probable Soviet test of a Tesla Shield over Atlanta, Georgia 60
33. The riddle of time 99
34. Raindrop model of physical change 99
35. Phase conjugate reflection of a time-reversed wave 107
36. Phase conjugation by four-wave mixing 119
37. Soviet launch phase and midcourse strategic defense system 120
38. Power reflection coefficients and oscillation 123
39. Photons have virtual substructures 127
40. Once established, a virtual river is for free 129
41. Local energy conservation can be violated 130
42. T. H. Moray adjusts the controls of his radiant energy device 130
43. Charge affects anything existing in spacetime 136
44. Vacuum/spacetime is pure virtual particle flux 140
45. A simplified charged particle 141
46. A charged particle is a special kind of "spray nozzle." 141
47. An observable charged particle is directly connected to virtual particle vacuum flux 142
48. In a flux gradient (virtual particle river), a charged particle moves itself 143
49. A natural potential is a *disorganized* change in the stress of vacuum 143
50. An artificial potential is an *organized* change in the stress of vacuum 144
51. Gravitational potential is a conglomerate of stresses, organized or disorganized 144

LIST OF FIGURES

No. Title Page

52. EM force fields are releases of gravitational potential via observable charged mass flows ... 145
53. As beyond (without), so within ... 147
54. Internested levels of virtual state vacuum interact with neutrino, photon, and mass ... 147
55. Virtual state patterns are absorbed and emitted by the atomic nucleus ... 148
56. Nuclei of the universe continually exchange scalar EM waves and virtual particle fluxes ... 148
57. First order reality ... 149
58. Higher order (hyperspatial) reality ... 149
59. A photon is a single oscillation of a carrier wave ... 153
60. A potential is trapped energy ... 162
61. A force-field vector is a movement of mass in one direction ... 171
62. The zero-summed vector system is an engine and an artificial potential ... 173
63. An analogy for how scalar electromagnetics can engineer Kaluza-Klein unified field theory ... 182
64. The zero-vector system can engineer Kaluza-Klein theory ... 182
65. The basic idea of Klein's explanation of where the extra dimension is ... 183
66. Normal EM bleeds-off G-field ... 184
67. Scalar EM forces 4-dimensional gravitational field ... 184
68. The curl-free A-field ... 190
69. Earth, sun, and moon are locked into a triad system ... 195
70. Forming magnetic monopoles with a standing scalar EM wave ... 202
71. Metal fracture by depositing magnetic monopoles ... 203
72. The Kervran effect ... 205
73. The Kaznacheyev effect ... 215
74. A photon is one oscillation of an electromagnetic carrier ... 215
75. Nested modulations are the key to hyperspatial engineering ... 216

LIST OF FIGURES

No. Title Page

76. Hyperframes, vacuum, virtual state, minds and thoughts ... 233
77. Internested levels of virtual state vacuum contain mind and thought ... 237
78. The human cerebral cortex is a natural scalar interferometer ... 238
79. Consciousness and life ... 238
80. Life and death ... 239
81. Layers of unconsciousness intersect hyperspatially ... 239
82. Scalar electromagnetics can directly interact with the various levels of human mind and personality ... 240
83. What personality is ... 241
84. Jamming the "older files recall" process produces simple amnesia ... 241
85. To sustain one's personality, continual functioning is required ... 242
86. If the files themselves are changed, the personality itself is changed ... 242
87. The mind and personality can be affected—and changed or controlled—from a distance ... 243
88. The mind and personality are physical things, albeit hyperspatial and virtual ... 246
89. One of Priore's intermediate devices ... 248
90. The large special phase-conjugating plasma tube for Priore's giant device ... 252
91. Capping assembly for the top of Priore's giant plasma tube device ... 253
92. Royal Raymond Rife in his laboratory ... 268
93. Rife's first virus microscope ... 268
94. Side view of Royal R. Rife's prismatic universal microscope ... 269
95. Front view of Royal R. Rife's prismatic universal microscope ... 270

LIST OF TABLES

No. Title Page

1. Interesting Soviet statements 25
2. Microwave radiation of the U. S. Embassy in Moscow 32
3. Khrushchev's 1960 statement 35
4. Death of the U. S. S. Thresher 37
5. Severe EM interference experienced by U. S. S. Skylark 38
6. Massive 1985 strategic May Day exercise 48
7. Shuttle launch anomalies, Nov. 26,1985 51
8. Soviet escalation after 3rd test against shuttle launch on Nov. 26,1985 51
9. Arrow DC-8 disaster Dec. 12, 1985 53
10. Death of the Challenger, Jan. 28,1986 55
11. KGB celebration 55
12. Brezhnev's 1975 proposal 59
13. Symmetry 91
14. A force field is a special local symmetry 92
15. Gravity and local symmetry 92
16. Hindsight is better than foresight 93
17. When symmetry is violated (broken) 93
18. Some proven asymmetries 93
19. Spontaneous symmetry breaking 94
20. Why such drastic difference? 95
21. Aspects of strong local asymmetry 96
22. Flow of time 100
23. Physical change 100
24. Negative energy 101
25. Phase conjugate waves 108
26. Phase conjugate replica 108
27. Distortion correction theorem 108
28. Four-wave mixing 119
29. A zero vector system 135
30. Stress is fundamental 136
31. Vacuum is 140
32. Spin couples vacuum to particle 142
33. Residue universe, residue science 145
34. Reality is a hologram 151

LIST OF TABLES

No. Title Page

35. Physical change 153
36. Ordinary General Relativity (OGR) is a special case 155
37. Maxwell's equations 175
38. Kaluza geometry 181
39. 5-D G-field bleed-off (Kaluza 5-D theory) 181
40. Physical reality is eleven dimensional 183
41. Comparison of EM concepts 188
42. The Aharonov-Bohm effect 189
43. The living aura: The cell's electromagnetic breath 220
44. Lisitsyn's report 243

Chapters 4 and 5
contain complex scientific material
and are suggested for those who are
technically proficient.

FOREWORD

What This Book is About

In my book, **"Fer-de-Lance: A Briefing on Soviet Scalar Electromagnetic Weapons,"** (Tesla Book Co., P.O. Box 1649, Greenville, Texas 75401, 1986), I specifically pointed out that I was of necessity neglecting the biological warfare (BW) aspects. I intended to later write a second book covering the BW aspects.

This book contains the most important item of that planned second book, including the photobiology aspects of BW.

The material is put down in great haste, for it is meant to serve as an urgent warning. There is no time to do the six months additional research and compilation necessary to "flesh it out", sculpture it and meticulously document it, as should be done by a careful scholar.

I apologize to the reader for this lowered standard, but in this case the basic information is far more important than scholarship.

Further, I acknowledge the fundamental contributions of Dr. Robert B. Strecker, M. D., Ph. D., who together with his brother Ted, has for some time urgently called attention to the terrible threat posed by the AIDS epidemic.

I specifically acknowledge his important warning, "This is a Bio-Attack Alert," Mar. 28,1986, which was sent to the U. S. President and Vice-President, Governors of the states, various federal agencies such as Departments of State, Defense, and Agriculture; National Security Agency, Federal Bureau of Investigation, Central Intelligence Agency; and three members of Congress.

Dr. Strecker has also strongly pointed out the covert involvement of Soviet and other communist agents in the epidemic, and the absolute necessity for an *electromagnetic* solution.

Also recognized are the efforts of the other medical scientists, such as Dr. John Seale, M.D., and Dr. Eva Snead, who have tried repeatedly to call attention to the nature of the AIDS initiation, only to have their submitted papers resoundingly rejected by important journals such as New Scientist, **The** Lancet, etc.

Work by Dr. William C. Douglass in alerting the public to the no-nonsense truth about aids is particularly recognized. His monthly newsletter, **The Cutting Edge,** is an oasis of clarity sharply standing out in a sea of misinformation. Permission to reproduce his published research article on UV irradiation of the blood to successfully heal viral infections is deeply appreciated.

The continuing efforts of Dr. Andrija Puharich and Dr. Robert Beck to assist in achieving an AIDS solution are pointed out and recognized. Important contributions by Ed Skilling are also recognized in this regard.

The vital and extensive contributions of Christopher Bird are particularly acknowledged. Chris has unstintingly and unselfishly shared with me the results of his enormous research on the Priore device and affair and the results of his translations from the French. He has most graciously permitted me to publish his important summary of the Priore affair, and I am greatly in his debt. The reader will find that material most illuminating, and well may be shocked at the depth and malevolence of the scientific and political suppression of Priore's rigorously demonstrated electromagnetic cure for cancer, leukemia, and other killer diseases.

Special acknowledgement is made to Robert Whitney, who so nobly labored to bring the Priore technology into the mainstream of medical treatment, only to be ruthlessly suppressed in his efforts. Photographs of the Priore machine were furnished by Bob from his collection.

My special thanks to colleagues John Bedini, Frank Golden, Andrew Michrowski, George Fencl, Ken Moore, Tom McLaughlin, Joe Parrott and Randy Davidson, without whose assistance this project could not have been accomplished.

The material in this present small book is intended to be of any assistance possible to certain brave colleagues who — with little or no funds or resources except what they can tear out with their bare hands — are already laboring around the clock to develop a permanent solution to the AIDS problem.

As a matter of discretion, I shall not mention the names of those colleagues. The solution sought is radical, by orthodox assessment. Yet it is the only possible approach that has any chance whatsoever of succeeding.

What they are doing, I cannot do. If they do not succeed, no one can. But I will say this, and most strongly: In my view, the only hope for humanity lies in the work they are doing, and they have my utmost admiration and support.

Their initial results are exciting and encouraging. Basic phase conjugated electromagnetic "mixed-modulations" signals lethal to bacteria, nematodes (hookworm-like organisms), various amoeba, viruses, etc. at very weak power levels have been synthesized and demonstrated.

In my best estimate/opinion, these researchers must succeed in *two years or less, certainly in no longer than three years,* else the **AIDS** incubation already accomplished by then, and the implementation time delay — even given a cure —together will insure a devastation of humanity unparalleled in all of history.

Let me accent that developing this AIDS cure is far more important than developing our entire arsenal of new military weapons. It is even more important than the development of defenses against the Soviet scalar electromagnetic weapons I detailed in **Fer-de-Lance.** Problems in the Middle East, Korea, and Nicaragua are insignificant compared to the AIDS problem.

Desperate Diseases, Desperate Remedies

Spurred by the looming holocaust of AIDS, and by the strong efforts of my colleagues, I have decided to release specific *information on how to develop electromagnetic healing, essentially against any and all of the "killer" diseases including cancer, leukemia, and AIDS.*

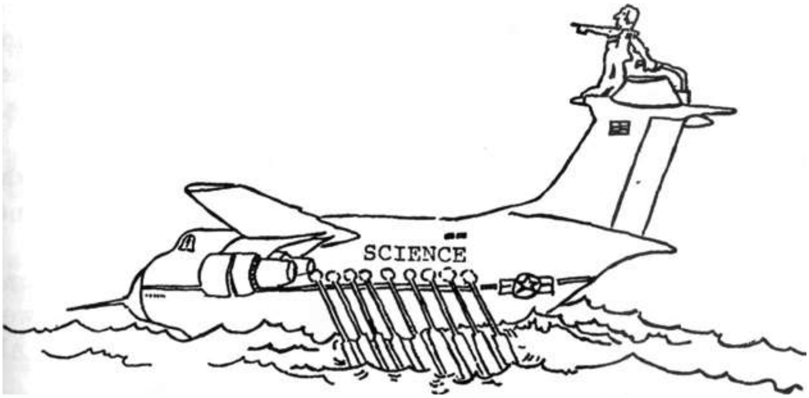

Figure 1. Conventional science cannot do the job.

Unavoidably this also releases a major principle of scalar electromagnetics: what is really going on in phase conjugation. If phase conjugation is truly understood and applied, all the Soviet weapons I have pointed out in **Fer-de-Lance** can be understood, designed and developed.

I'm also releasing specific information on mind and thought that will enable them to eventually be directly detected, measured, and investigated scientifically. Unfortunately, this also will enable them to be technologically manipulated for nefarious purposes.

However, the developing scalar EM technology for the AIDS cure must not be hamstrung. There is no time to apply the standard "Let's classify it!" response. If we do, we'll win the skirmish and lose the war.

As shocking as it sounds, I am convinced that the Soviets have already launched WW III, to the hilt. And it's not at all the kind of war we've been expecting.

There's no time to rail against the Soviets, who have cunningly and in great secret "struck our vitals with the mailed fist." We've been struck in such a way that, *unless we can solve the AIDS problem, in two to three years or less, we are already finished.*

We have been preparing to fight a better World War II. The one we're preparing for is never going to be fought. The real World War III is already raging. Our death knell has been sounded, *and we've not yet even recognized that the battle is underway.*

A great Soviet "first strike to finish us" has already been accomplished. The new "Pearl Harbor" has already happened. The enemy has already overrun our Maginot Line; we are at our Dunkirk. There are no boats waiting, and there are none on the way.

The great Soviet first strike has been delivered with surreptitious biological warfare, not with nuclear weapons or ground forces.

As always in the past, we have almost no troops facing the brunt of the first strike and trying to turn it. This time we have far fewer than we've ever had. Just two or three unorthodox researchers. All the orthodox researchers are manning the wrong ramparts.

Ironically, this time our entire system—the government, the medical establishment, the scientific establishment, the universities — all are far too dogmatic, and have far too sluggish a response time, to offer any assistance at all. They are simply so "out of it" that they don't even know what's really happened.

The bureaucracy simply is so stagnated that, even with a giant paroxysm, it cannot change quickly enough to do the job in time.

The best response the orthodox system could theoretically make —if it immediately placed the nation under martial law, marshalled all the universities and scientists, everything — would require at least seven years to produce results. A more likely response time is 10 to 14 years. That won't do the job.

So the job has to be done from completely *outside* the system. It has to be done with *independent* financing. Financing without a lot of conventional, bureaucratic strings on it. Resources must be committed as needed, immediately, not with a several months' approval process.

Everything in the orthodox establishment simply has to be bypassed.

The government, the medical establishment, the scientific establishment, the universities, everything.

Otherwise, we're as good as finished *right now*.

The "Dirty Dozen"

In the West, there are only perhaps a dozen persons or so that I know of, who are qualified to work on the main task required. They are all unorthodox researchers and scientists.

There are thousands of fine orthodox scientists, of course, who can help with the peripherals — making and testing viral solutions, setting up and operating equipment, doing interminable tests, all the lab stuff, etc.

But only about 12 unorthodox guys are qualified and mentally and emotionally prepared to come to grips with the particular scalar EM technology required for rapid development of an AIDS cure.

Right now, there are just two or three of the 12 working on this, hammering away as best they can, at terrible cost to themselves and their normal endeavors.

Obviously immediate funds are needed for facilities, equipment, and marshalling that dozen unorthodox researchers and a team of scientists — along with support staff and personnel.

An Analogy

I'm reminded of a story from WWII.

On the beach at Anzio, in a battle lull some doughboys were briefly resting near their foxholes, eating lunch from C-ration cans.

A lone German Messerschmitt appeared, closing on them fast and low with machineguns blazing. They were forced to hastily jump into their foxholes as the Messerschmitt strafed their position.

The pilot continued to make strafing run after strafing run, low and "on the deck." Obviously lunch was rather totally disrupted.

Suddenly, in total disgust a soldier leaped up and flung his half-filled C-ration can into the air directly at the approaching aircraft, full in the teeth of the machinegun fire.

Incredibly, the can sailed straight up into the air intake of the hurtling aircraft, where it was sucked into the engine, which sputtered and died.

The surprised German pilot banked his plane upward and bailed out.

He was promptly captured by the very troops he had been strafing only moments before.

Those two or three colleagues of mine are literally "throwing their C-ration cans" in a desperate effort to ward off a monstrous threat.

Those two or three are striving to the limit of their endurance.

This book and the release of the material in it represent my own "Oration can" hurled into the teeth of the actual Soviet threat as a last resort. Hopefully we can "hit the air intake" of the Soviet juggernaut, so to speak, and stop its progress.

The giant Soviet scalar electromagnetic weapons that I detailed in **Fer-de-Lance,** e.g., are easily countered — in six months or less. Primarily, it is necessary to make a pumped phase conjugate mirror adjunct (PPCMA) that operates at the frequency band of interest, and associate it with a receiver. Once a signal is received from the enemy weapon — say, from a Woodpecker transmitter — it is introduced to the PPCMA. A powerful phase conjugate replica energy pulse is returned into the distant hostile transmitter, totally destroying it.

To meet the viral biological warfare weapon, however, we must develop electromagnetic healing.

In an important article, "SDI vs. the plague," Mar. 8,1988, p. 6, the **Washington Times** revealed a startling Pentagon conclusion that the Soviets may counter SDI with genetically engineered weapons.

Valentin Falin, Chief of the Soviets' Novasti Press Agency, is quoted as saying:

> *"We won't copy you anymore: We'll take asymmetrical means with new scientific principles... genetic engineering could be a hypothetical example. Things can be done for which neither side could find defenses or countermeasures, with very dangerous results. If you develop something in space, we could develop something on earth. These are not just words. I know what I'm saying."*

The genetically engineered Soviet counterstrokes exist and are already being taken. This book tells you what the asymmetrical means, new scientific principles, and genetically engineered weapons are.

May God help us all. If we do not move immediately and correctly to augment the effort of my colleagues, the Free World and half of the peoples of the earth will perish before another decade passes.

CHAPTER 1

THE VACCINE CONNECTION

The following quote is from Don Rowe, **Wall Street Digest,** 214 Carnegie Center, Princeton, New Jersey 08540.

"The AIDS Plague will affect society in ways that you cannot now imagine. AIDS is 100% fatal. Scientists and medical research people are not optimistic about an immediate cure to [or] vaccine. They do not expect to develop an effective vaccine within the next two decades. Public health officials estimate that as many as 2.4 billion people (half the world's population) will die from AIDS within the next 15-20 years."

"Economically, the insurance and medical health systems could be devastated in the 1990's. Nothing short of a spectacular medical breakthrough will keep Western civilization from suffering the worst catastrophe in the history of the world."

The first official U. S. AIDS case was diagnosed in San Francisco in 1981. This followed the puzzling outbreak of the disease in Third World areas such as Africa and Haiti in the 1970's.

The U.S. AIDS outbreak at first seemed to be confined to the gay community, such as in San Francisco.

No one seemed to connect the apparently unrelated facts that, before the outbreak in the U.S. gay community, (1) the World Health Organization had accomplished substantial smallpox vaccination in Haiti, and (2) West Coast gays — particularly from San Francisco — had made Haiti a main playground and vacation spot.

Most orthodox medical scientists have not the foggiest notion why such a drastic "explosion" of the virus has occurred.

It has occurred because *AIDS was introduced by contaminated vaccine used by WHO in its smallpox eradication campaign.* A high correlation exists between WHO vaccination and subsequent outbreak of AIDS.

Figure 2. The "monkey-bite" theory.

The "Monkey-Bite" Theory

As the prevailing rationale goes, a naturally-mutated monkey AIDS virus derivative or strain is believed to be the cause of the monstrous plague looming to strike down half of all the humans on earth.

That is, a monkey is supposed to have bitten a man—an African—and transmitted the normally harmless monkey AIDS virus to him. The virus is then supposed to have naturally "recombined" its genetic material with other viral genetic material and the man's human genes, to produce the virulent new strain of AIDS.

Then this new strain, *from its single human host,* is supposed to have rapidly started spreading through Africa and around the world.

That is simply nonsense.

In Africa, for example, monkeys as well as other animals have had their viral diseases for thousands of years. If monkey virus were the culprit, African man would have been devastated by AIDS thousands of years ago. He either would have been dead or—as have the monkeys — he would have developed a certain resistance and tolerance for the disease, so that it is no more harmful than others.

Understand, the "monkey bite plus human cell modification" undoubtedly does occur in Africa with great rarity — once in several thousand years.

However, even if that rare natural recombination of the monkey AIDS virus were to occur in a single human, normal contact between other humans and the infected human is totally insufficient to provide any substantial infection rate *unless the rate of infection of new individuals is greatly assisted and vastly speeded up by some other outside mechanism — for example, by something like AIDS-contaminated vaccine and a mass inoculation program.*

Retroviruses **and Recombinants**

As Dr. John Seale has pointed out, "The AIDS virus (human immunodeficiency virus or HIV) is a lentivirus — a little-studied sub-family of the retroviruses. It is highly pathogenic to man but it differs profoundly from any other virus of humans. It is spreading rapidly in the general population of central Africa, and amongst homosexual men and drug addicts in the West. *It is the first virus to have appeared in mankind for many centuries which is entirely new, highly lethal, and spreading steadily from person-to-person world-wide."* [accent by the present author.]

Dr. Seale has courageously pointed out some other very disturbing facts:

(1) Retroviruses of animal origin, if repeatedly passed from human cell culture to human cell culture, will gain the ability to infect human cells *preferentially.* Indeed, they may then even cease to be infectious to their original animal hosts.

(2) In 1982 Robert Weiss pointed out that, in the laboratory, the genetic code of retroviruses commonly recombine if two of them infect the same cell simultaneously.

(3) In a 1984 article in the **Journal of Virology,** David Baltimore showed an efficient means of overcoming the cell's block against simultaneous infection by two or more retroviruses.

(4) Further, Baltimore reviewed how *this had widely been done during the preceding ten years!*

(5) Outside the laboratory, recombination of multiple simultaneous retroviruses has not been shown to occur naturally in humans.

(6) In animals, there is some evidence that it may occur naturally with extreme rarity — indeed, only once in many thousands of years.

(7) By the early 1970's, the technology to make new recombinant retroviruses virulent to humankind was already developed and had been widely published.

To the layman, the infection and reproduction by retroviruses is peculiar to say the least.

When a retrovirus infects a cell, from its own genetic material (RNA) it first makes a "template" or pattern (DNA) for forming other retroviruses just like it. It then inserts this "DNA template" into the DNA of a cell. The host cell itself is now a "sleeping factory", ready to start producing retroviruses.

When activated, the "template" in the cell causes the factory to go into production of retroviruses. Activation is varied. The "infection production factory" can just remain dormant, be slowed over years, take place almost immediately, or be slowed for years and then "explode" to rapidly increase the infective viruses.

Genetically, the cells of any person infected with AIDS or any other recombinant retrovirus have been permanently transformed. Literally that person has acquired a new set of genes. He or she has been converted into a "factory" for producing—and spreading—the virus.

Cancer Research Produced Ideal Biological Warfare Agents

In 1971 President Nixon signed into law the National Cancer Attack Act. Though certainly not intended by the President and the Congress, *the inadvertent result of this act was that taxpayer's funds were used to research and develop the precise things that constitute the most lethal kind of viral biological warfare.*

The on-going small effort in medical research laboratories to alter animal tumor viruses — including retroviruses —to replicate efficiently in human cells was drastically speeded up, as part of the new "war on cancer." In other words, whether intended or not, much of our cancer research money was converted into the precise kind of research that develops virulent biological retroviruses for potential use in biological warfare.

Not only was the research drastically accelerated, but also the direct participation by Communists and scientists sympathetic to the communist cause was accepted and encouraged.

Astoundingly, this open invitation to the communist scientists was issued at a time when one of the high KGB priorities was to infiltrate our recombinant DNA genetics laboratories! To the Soviets' utter astonishment, we simply opened our labs and invited them in.

The Soviets participated — and penetrated — with alacrity. It is an obvious conclusion that a number ofthe Iron Curtain scientists sent to help staff our laboratories were undercover KGB agents.

In the cancer laboratories, on a wide-spread scale, animal retroviruses were injected and reinjected into human cell cultures, until these retroviruses had become efficient, lethal recombinants favoring human cells as their hosts.

At that point, infection of one or more persons by one of these lethal new retroviruses could unleash a massive AIDS-like plague on all mankind.

A single laboratory accident (or a single act by one Communist agent) could unleash the Great Death Plague.

Weak and ineffective reference to this forthcoming danger was made in 1974 by the Committee on Recombinant DNA Molecules, which had been established by the U.S. National Academy of Sciences. According to **Science,** July 26,1974, p. 303, the Committee made the following rather bland and ineffectual recommendation:

"Scientists throughout the world (should) join with the members of this committee in voluntarily deferring experiments (linking) segments of the DNA from (cancer-causing) animal viruses to... possibly increase the incidence ofcancer or other diseases."

So much for the great scientific watchdogs.

U. S. "Cancer War" Labs Were Infiltrated

As we said, ironically we ourselves helped load and cock this potentially disastrous biological warfare cannon. Incomprehensibly, enemy fingers were permitted — even encouraged! — to be placed on the trigger. Many of our medical research facilities involved in this research contained known Communists, dissidents, etc.

For example, see page 106 in Omni, March 1986. Carlton Gajdusek, presently Chief of the National Institute of Health's Laboratory of Central Nervous System Studies and Laboratory of Slow, Latent and Temperate Virus Infections, was asked the question, "Isn't Fort Detrick in Maryland such a biological-warfare research facility?" He responded:

"No, emphatically no! There is no defensive or offensive warfare microbiology done at Fort Detrick today. It is the national cancer research facility of NIH. In this facility I have a building where more good and loyal Communist scientists from the USSR and mainland China work — with full passkeys to all the laboratories — than Americans. With night-working U. S. citizens and foreign Communist investigators here, obviously there is no "secret" bacterial warfare activity going on. Even the Army's infectious-disease unit is loaded with foreign workers — not always friendly nationals. It is a valid basic research unit on worldwide problems of infectious -diseases in which no classified or secret activities unfold."

The point is, what was being produced in these labs were viruses that would grow in human cells. Under tight control, these new lethal agents could be safely studied, and many things could be tried experimentally to see if they would slow, stop, or kill the viruses and heal or benefit the virally infected human cell cultures.

The viruses were being developed to benefit man by aiding in the war on cancer — we accent this most strongly.

However, many of them were precisely the same things that would be developed in a biological warfare lab: lethal new viruses that preferred the human as a host, and against which humans have developed little or no immunity.

The Deed Was Easily Done

Now one can see the subtlety and irony in the Soviet Union accusing the U. S. of having made the AIDS virus at Ft. Detrick!

It's both true and false at the same time.

There is substantial information indicating that the AIDS disease is manmade, and has been deliberately unleashed on mankind by Soviet and other Communist agents, augmented by the international control groups and shockingly inept work by U. S. and European medical researchers.

Human cell cultures deliberately infected with animal viruses were carelessly handled in U. S. and European laboratories — carelessly handled in that far too great (essentially open) access to the new viral strains was permitted.

Our own scientists were apparently unwitting pawns, and Soviet agents performed the "final touchups" to insure release — whether intentional or inadvertent — of the lethal new agents against an unsuspecting world.

Viruses can infect bacteria, since a bacterium is a one-celled organism and the virus is a single molecule.

Cell cultures where vaccine is made from weakened bacteria could easily be contaminated with the desired recombinant viruses available in the cancer laboratories.

To insure worldwide release of the lethal virus, covert Communist agents apparently caused contamination of the smallpox vaccine that was prepared in quantity in Western laboratories.

This contaminated vaccine then was widely used to inoculate Third World peoples and others, in the successful international program to eradicate smallpox.

With only a few agents in key places, the Soviets could easily have taken advantage of the looseness of Western laboratory procedures to deliberately contaminate the animal cell cultures used to make the vaccines.

They had an abundance of agents in the key laboratories, for they were openly invited in and accepted!

The Soviets had their fingers in it, up to the elbow.

And at just exactly the right time, the answer to a Communist's biological warfare dream was available: The World Health Organi-

©Hal Crawford 1979

Figure 3. The Soviets have delivered an intended knock-out.

zation was already involved in its 13-year campaign of massive smallpox inoculation in the Third World, to eradicate smallpox.

All that was needed was to contaminate the WHO vaccine by one or more of the lethal new recombinant retroviruses that had been developed — and were continually being developed — in cancer research laboratories filled with Communists agents.

The opportunity was presented for a great Soviet "first strike" against the West, using biological warfare agents developed in U.S. and European laboratories and administered by an international agency.

The opportunity was ripe for the taking. It was irresistible. You can be sure the Soviets took it.

It's Actually Documented!

Virus-contaminated vaccines were then widely administered to Third World countries—particularly in Africa, South America, and Haiti. Administration of these vaccines by the World Health Organization is precisely correlated to the subsequent outbreak of the AIDS epidemics in those countries.

This strongly indicates that the vaccine itself was the "precursor" agent initiating the start of the AIDS infection.

In a shocking article, "Smallpox vaccine 'triggered Aids virus'," in the London **Times** of May 11,1987, **Times** Science Editor Pearce Wright announced that the AIDS epidemic may have been caused by World Health Organization (WHO) smallpox inoculations in third-world countries.

In an important article in **Easy Reader,** June 4, Jon Rappoport reported the **Times** article and the results of an interview with Robert Matthews, technical correspondent of the **Times.**

According to Rappoport, Matthews informed him WHO itself had somehow suspected that its immunization program against smallpox might be connected to AIDS. An outside consultant was employed to perform an independent study, and the study confirmed that WHO vaccines were indeed — somehow — causing AIDS to evidence in inoculated persons.

Needless to say, WHO promptly buried that report—whose conclusion must have been the last thing WHO wanted to hear.

The consultant then came to the **Times.** The **Times** heard him out and printed the story.

This astounding story has almost startling correlations: *"The smallpox vaccine theory would account for the position of the Central African states as the most afflicted countries; why Brazil became the most afflicted Latin American country; and how Haiti became the route for the spread of AIDS, to the U.S.: Brazil, the only South American country covered in the eradication campaign, has the highest incidence of AIDS in that region."*

Rappoport confirms that the subject of contaminated vaccines— "vaccines which, made in tissue cultures of animal parts, contain 'extra' dangerous viruses" — has been of increasingly serious concern to medical scientists in recent years, and the WHO smallpox vaccines may have been contaminated in this fashion.

A documented U.S. Army incident is a case in point.

New U.S. Army recruits are routinely vaccinated against a range of diseases. One of them is smallpox, in case an enemy should try to use smallpox as a biological warfare agent.

A healthy 19-year old Army recruit developed AIDS after vaccination, and was admitted to Walter Reed Army Medical Center, where he died.

In a paper published in the **New England Journal of Medicine,** Walter Reed medical team scientists reported their discovery of the connection between vaccination and stimulated AIDS disease.

They also warned against the WHO plan to use modified versions of the smallpox vaccine to combat other diseases in developing countries.

WHO's 13-year vaccination and eradication program ended in 1980. The program is credited with saving two million lives a year and preventing 15 million infections.

It may be inadvertently responsible for the ultimate death of billions.

In a meeting of 50 experts near Geneva in May, 1987, it was revealed that millions of new AIDS cases are about to hit Southern Africa. The experts believe that up to 75 million Southern Africans — about a third of the population — may have AIDS within the next five years.

From our previous statements above, it appears almost certain that Communist agents who had penetrated U.S. and European laboratories manipulated the contamination of the altered vaccines.

Further, the Soviets may be stimulating the disease and guiding it electromagnetically, by special implanted signals carried by the giant Soviet woodpecker "over-the-horizon radars" and other Soviet transmitters.

Or, the Soviets may be using their own transmitters and ours, including our own power lines and radio and TV stations, to broadcast signals that "precondition" and greatly amplify our population's vulnerability to certain selected diseases — AIDS and otherwise.

We will address these issues and Soviet subtle EM capabilities later.

Soviet Biological Warfare Background

In World War I, the Russians suffered half a million casualties to gas warfare, including 50,000 dead.

At 10,000 troops per light division, that's the equivalent of 5 divisions killed and 45 divisions incapacitated.

This made a deep and lasting impression on the Soviets. Lenin himself ordered a major effort mounted to develop and manufacture chemical and biological weapons, for he understood the implications of such mass-destruction weapons.

In 1919 the Soviets began work in earnest on BW weapons. Work on them has not stopped since then. It will not stop so long as the Communists remain in power.

To loyal Communists, Lenin's dictate to develop and lead the world in biological weapons is absolute and unquestioned.

By 1937, the Red Army was fully equipped and psychologically prepared to wage chemical and biological warfare.

Continuing today, the Soviet forces have the highest training in chemical warfare of any troops in the world. Red Army soldiers practice regularly, often using actual chemical agents. Decontamination is a standard measure and practiced as a matter of course.

Even though the Soviets, along with many other nations, signed the Geneva Protocol of 1925 which banned the use of deadly chemical and bacteriological weapons, the Soviets have never ceased development of such weapons.

In the 1930's, a secret Soviet bacteriological institute was set up at the town of Suzdal. Prisoner microbiologists worked on a variety of new weapons and vaccines, including bubonic plague and tularemia (rabbit fever). By World War II the Soviets were prepared to retaliate vigorously to any use by the Germans of chemical and biological warfare agents.

After the war, Soviet research and development in chemical and biological weapons continued apace. However, much of the genetic work was hampered by Stalin's stern support of the autocratic control of agronomist Trofim D. Lysenko. Lysenko's crackpot theories on genetics, for example, held that genes did not even exist. After the death of Stalin, Lysenko's work was discredited, and the Soviets set forth immediately to "catch up" with the West in genetic implications such as recombinant DNA techniques.

Thus in the early 1960's the Soviets began sending Russian scientists to the West to catch up with Western genetics research. One of these scientists was David Goldfarb, who later became a leading Soviet expert in the field of molecular genetics. He headed up the Laboratory of Molecular Genetics of Bacteria and Bacteriophages of the Soviet Academy of Sciences in Moscow. Although he did not work on secret Soviet projects, the Soviet authorities stopped him cold when he applied for a visa to leave in 1979. Goldfarb was a specialist in plasmids, antibiotic resistances, and recombinant DNA technology—all very, very useful to the Soviet effort in biowar.

One can readily understand the reluctance of the authorities to release scientists such as Goldfarb. After all, the Soviet "dirty work" with the deliberate AIDS contamination of the World Health Organization's smallpox vaccine had already been accomplished and a great Soviet biological warfare "strike" was already underway, thanks to the massive smallpox vaccination program of WHO.

In 1972, in the Biological and Toxin Weapons Convention, the Soviet Union and every major power agreed not to develop or use biological weapons. The agreement went into effect in 1975. Although it bans the development, production, and stockpiling of biological agents and toxins for hostile purposes, it has not slowed or altered the Soviet program.

In the early 1970's, Soviet leaders made a decision to use genetic engineering to produce new and improved BW weapons. Highly secret institutes for this purpose were set up in Moscow, Leningrad, and Novosibirsk. These laboratories specialize in seeking ways that genetic engineering techniques can be used to develop virulent strains and agents that their opponents cannot resist.

Every Soviet civilian institution dealing with bacteriology and epidemiology is responsible for certain military work as well. The Soviet Academy of Sciences in fact funds appreciable research of such nature at these institutions.

One of the major Soviet centers for recombinant DNA techniques in biological weapons research is in the village of Protvino, near Moscow. Research on chemical and biological weapons is very easy to conceal, and the Soviets are sure that they will not be caught. Satellites cannot detect what is going on in the labs. The payoff is big and the risk is minimal. The Soviets push this area vigorously.

A number of installations capable of producing disease agents and toxins on a large-scale basis and placing them in munitions and delivery systems have also been identified. One such facility is in the city of Sverdlovsk and has a long history of BW research and development and production, with emphasis on anthrax. During early April 1979, an accidental release of anthrax occurred in Sverdlovsk that caused many casualties and deaths in spite of heroic efforts by Soviet doctors and cleanup and decontamination crews.

Note that, according to DoD's 1987 edition of **Soviet Military Power,** page 110, *"Anthrax causes a high mortality rate when the infection results from ingestion (up to 70 percent fatal) or inhalation (almost 100 percent fatal) if treatment is not promptly begun."*

In addition to military BW institutes, military officers and KGB agents are assigned to key posts in the Soviet Union's normal health and science ministries to monitor research and development that has possible military significance.

Also, the KGB has a continuing and highly active program to insert recruited Soviet Scientists in Western genetic research institutes and agencies. This is a high priority effort, and it has been very successful. Also, it obviously says something about the probability of Soviet agents penetrating U.S. and Western European cancer research institutes, particularly those using recombinant DNA techniques. To make it even worse, we simply invited them in with open arms!

Indeed, it was that KGB program plus the "open arms" in Western cancer virology research institutes that enabled the Soviets to set up and bring about a massive "first strike", using WHO and Western laboratories to do the actual dirty work unknowingly.

We *Need An Immediate, Full Scale Investigation*

Annually, President Reagan makes a report to the U.S. Congress on the subject of Soviet noncompliance with Arms Control agreements.

In April, 1984 Defense Secretary Caspar Weinberger put it this way:

"There is an apparent effort on the part of the Soviets to transfer selected aspects of genetic engineering research to their biological warfare centers... Soviet research efforts in the area of genetic engineering may also have a connection with the biological program...Normally harmless, non-disease-producing organisms could be modified to become highly toxic or produce diseases for which an opponent has no known treatment or cure. Other agents, now considered too unstable for storage or biological warfare applications, could be changed sufficiently to be an effective agent."

At the time that Secretary Weinberger made this statement, a devastating biological warfare strike using those exact techniques had already been unleashed on the Western world by the Soviet Union, cleverly disguised under the banner of the World Health Organization's smallpox vaccination program.

It is imperative that the Congress of the United States launch a full investigation of the entire background of the AIDS biological warfare strike, as a matter of the highest priority.

The investigation should include the role of the Soviet Union, fellow-travelers, and Western scientists openly sympathetic to the Soviet cause.

It requires little investigation to see that there is a well-marked trail for several decades, showing that such a biological warfare strike against humanity has long been considered — and even preferred — by those advocating the Communist cause and world domination.

For example, Bertrand Russell, philosophical spokesman for a group of Westerners, explicitly stated this diabolical BW capability for consideration, in his **Impact of Science on Society:**

"At present the population of the world is increasing at about 58,000 per diem. War, so far, has had no very great effect on this increase, which continued throughout each of the world wars... War has hitherto been disappointing in this respect... but perhaps bacteriological war may prove effective. If a Black Death could spread through the world once in every generation, survivors could procreate freely without making the world too full. The state of affairs might be

unpleasant, but what of **it?"**

The blood of every American — and every European, African, and Asian — should curdle at such words. We are now facing the very "Black Death" type of plague referred to by Russell in such glowing terms!

Highly influential followers of the Russell philosophy — and his callous approval of the periodic slaughter of much of the world's population — may have continued to view the great destructive power of new viral warfare agents as the best means to achieve their end: Selective thinning of the world's population.

At least a few such men may have continued to cooperate with the Soviet Union to work powerfully behind the scenes in international organizations, trying to arrange the proper opportunity to achieve their nefarious purpose.

The "East-West cooperative effort" in cancer research created and gave them the precise biological agent custom-made for the job.

WHO's massive, worldwide smallpox inoculation program gave them the precise means to rapidly get the deadly virus into enough human veins to assure a new "Black Death."

All that had to be done was surreptitiously contaminate the vaccine — the easiest thing of all.

This plague has been knowingly, with malice aforethought, unleashed upon mankind. The perpetrators should be hounded down like the wild beasts they are, uncovered in the full light of day, and suitably punished with the swift and sure justice they deserve.

At least one outstanding medical consultant had the courage to report to WHO that their vaccine was generating the AIDS epidemic. When WHO suppressed the evidence, the London Times had the courage to print the story.

I strongly urge every U.S. Senator and Congressman to introduce and/or support immediate legislation that would launch a full-scale, massive investigation, with no stone left unturned. I urge every concerned activist to demand and support such an investigation at the highest level.

By preoccupying ourselves with lesser things — "IranGate," Nicaragua, the Middle East, etc. — we are simply fiddling while Rome burns.

CHAPTER 2

THE SOVIETS KNEW THE U.S. WOULDN'T REACT

Administering the Coup-de-Grace

As we've seen, the present AIDS viruses were deliberately created in Western cancer research laboratories strongly penetrated by Communist scientists and Soviet agents.

In laboratory experiments that were a legitimate and vital part of the war against cancer, animal retroviruses were repeatedly injected into human cell cultures until they recombined to produce **strains** hostile to human life—strains that would thus live in human cell cultures.

Even though their work was accomplished for a totally different purpose, cancer laboratories produced the ultimate, made-to-order biological warfare agents.

The World Health Organization (WHO) is also strongly penetrated — and even controlled at some levels—by communist agents, **as** are many other such international agencies.

And WHO's massive smallpox vaccination program in Third World countries provided the Soviets a made-to-order opportunity to have (ostensibly) "someone else" deliver the actual BW strike.

In other words, the Soviet deception plan was complete.

It was very simple for Soviet agents to contaminate the cultures in which smallpox vaccine was being produced for WHO's inoculation campaign. This insured the production of vaccine contaminated with AIDS viruses, and the subsequent massive infection of the African peoples and other Third World nations.

It delivered what will be the knockout punch against the West, unless AIDS is countered.

Table 1. Interesting Soviet Statements

"WE HAVE A NEW WEAPON, JUST WITHIN THE PORTFOLIO OF OUR SCIENTISTS, SO TO SPEAK — SO POWERFUL THAT, IF UNRESTRAINEDLY USED, IT COULD WIPE OUT ALL LIFE ON EARTH."

NIKITA KHRUSHCHEV,
1960, SPEAKING TO THE PRESIDIUM

"WE WILL BURY YOU!"

NIKITA KHRUSHCHEV,
TO RICHARD NIXON, "KITCHEN" DEBATE

"IF A MEANS OF TOTAL NEUTRALIZATION OF FOREIGN MISSILES IS TO BE FOUND, IT CAN ONLY COME FROM *A GROUP OF NEW PRINCIPLES IN PHYSICS, CALLED ENERGETICS!"*

NOBELIAN PETER KAPITSA, TO NIKITA KHRUSHCHEV

"EACH SIDE SECRETLY DEVELOPS NEW MEANS OF WARFARE IN ORDER TO EMPLOY THEM UNEXPECTEDLY. HISTORY KNOWS MANY EXAMPLES HOW THE EMPLOYMENT OF A NEW WEAPON INITIALLY GAVE CONSIDERABLE SUCCESS BECAUSE THE ENEMY, CAUGHT UNAWARES AND NOT KNOWING THE COMBAT CAPABILITIES OF THIS WEAPON, WAS FOR SOME TIME INCAPABLE OF EFFECTIVE COUNTERACTION."

V.YO. SAVKIN, THE BASIC PRINCIPLES OF OPERATIONAL ART AND TACTICS. MOSCOW, 1972.

"OF PARTICULAR IMPORTANCE IS BASIC RESEARCH AIMED AT DISCOVERING STILL UNKNOWN ATTRIBUTES OF MATTER, PHENOMENA, AND THE LAWS OF NATURE, AND DEVELOPING NEW METHODS FOR THEIR STUDY AND USE TO REINFORCE THE STATE'S DEFENSE CAPABILITY."

PORTION OF A BOOK BY MARSHALL GRECHKO
DELETED FROM ENGLISH TRANSLATION, BY SPECIFIC REQUEST OF THE U.S.S.R.

"WILL SHORTLY SEE DEVELOPMENT OF NEW WEAPONS, MORE POWERFUL THAN NUCLEAR WEAPONS, AND THEY WILL BE NONVERIFIABLE."

LYSENKO (FIRST NAME UNKNOWN)
USSR REP/SOVIET EMBASSY WASHINGTON, D.C.
WORLD FURTURESCONFERENCE, WASH. D.C, 20 JUL 82.

Murky "Behind-the-Scenes" Intentions

As was pointed out, the WHO inoculation program is well-correlated as the precursor or "initiator" of the AIDS epidemic.

Where WHO smallpox vaccination occurred, the mysterious symptoms of AIDS disease appeared in due time.

Even the eventual spread of the AIDS virus to the U.S. homosexual community is explainable.

For example, the economy of Haiti had earlier been made a shambles. The poor citizens of that impoverished island had been reduced to desperate straits.

In desperation many younger Haitian males had begun to prostitute themselves to visiting homosexual tourists. The Haitian males were not homosexual themselves; rather, they were desperate to gain money to provide for their starving families.

In response, U.S. homosexuals — particularly from San Francisco and New York — had made Haiti their vacation playground. They even ran ads to this effect in their gay newsletters and mngazines.

WHO then vaccinated some 15,000 or so Haitians against smallpox, using the vaccine contaminated with the AIDS virus. This introduced AIDS into Haiti and into the Haitians.

The U.S. homosexuals quickly picked up the AIDS virus by direct contact, and brought it back to the U.S. with them. There the nature of the activities engaged in by homosexuals was responsible for enhanced spread of the virus among members of the homosexual community.

Engineering the AIDS epidemic appears to have been a collusion between Communists and the international control groups.

Apparently the control groups wished to "clean out the African continent and other Third World areas" — possibly to provide room for relieving population pressure in the developed countries. This may have been seen as the way to solve the problem of the ballooning human population — the so-called "biological time bomb."

The Soviets, however, realized that contamination of the smallpox vaccines would result in an epidemic of unparalleled proportions. They recognized the opportunity for safely initiating — and winning — World War III with a great first strike.

The Soviets themselves already possess an electromagnetic cure for diseases such as AIDS, but they have not unveiled it.

All such products of scalar electromagnetics (which the Soviets call energetics) are held tightly by the KGB.

That the Soviets have kept such a cure secret strongly implies that they regard it as a known antidote for biological warfare agents, and thus highly classified.

The importance of this can easily be seen: Presently the West has absolutely zero defense against a biological warfare strike using viruses as BW agents.

In such a strike, fast-acting viral strains can easily be used — say, agents which are highly infectious and kill within two or three days.

Development of a vaccine is impossible. There is simply no time to do that. Further, even if the virus was slow-acting, a vaccine could only be developed if the virus is stable. If it is unstable (constantly mutating), development of a vaccine is impossible.

The greatest vulnerability of the U.S. today is its total vulnerability to unknown BW viral agents. Absolutely nothing we have or are doing in our Department of Defense is of any consequence or protection against new BW viruses for which no vaccine is developed or possible.

The great ease with which such cheap new, highly lethal viral BW agents can be developed means that even a relatively small military power—such as Libya, Cuba, or Iran—can develop and use them against the U.S.

Such BW viral agents can also easily be clandestinely planted across the U.S., to devastate an unsuspecting, unprotected American populace.

In fact, even a single deranged madman can readily develop lethal BW viral agents sufficient to devastate even a great military power. Further, in an open society such as the U.S., that same madman can easily introduce such an agent he develops — and, if he's clever, not even get caught in the act.

Past isolated cases of "contamination" of on-the-shelf drugstore medicines is just a faint prelude of that which is to come.

Indeed, we may unwittingly do it to ourselves by accident, even without a BW strike by the Soviets, small enemy powers, or a madman.

Throughout the West, it is now most fashionable to experiment with the production of new bacteria and viruses. Gene-splicing, rccombinant DNA genetics, etc. are being employed on a grand scale as the latest "scientific marvel."

Even some high school students have experimented with this sort of thing. Many other colleagues, universities, and private persons are proceeding full-tilt with casual experiments.

Most of this "casual experimentation" is uncontrolled. When finished, the residue gets dumped in the trash can or flushed down the drain. In any case, it goes into our biosphere. Some of this residue contains unstable, highly adaptable viruses.

It's only a matter of time until one of these unwitting agents gets loose and survives — one that's lethal, fast, and highly infectious.

Make no mistake: It's a *near certainty* that this will happen. The only question is *when.*

Also, with the recent Supreme Court decision allowing the patenting of a virus or bacteria, a number of laboratories and companies are now feverishly developing new agents to unleash in the biosphere upon persons, animals, and plants.

For the most part, these experiments are tightly controlled. However, it is impossible to test in the laboratory the result of the natural recombination and mutation of these agents once they are loose in the environment and in a free exchange with all the natural bacteria and viruses.

Once loosed in the biosphere, a new viral agent undergoes combinations and recombinations with the greater number of viruses already present. The situation is like an increase in a large factorial number: Only a few additional viruses will result in millions and millions of additional combinations — at least some of which inevitably will be lethal and quick-acting. This is *not* tested in those laboratories making, patenting, and releasing new viruses.

Humans have spent millions of years developing sufficient immunity to the natural BW agents to allow humanity to survive.

Very shortly we are going to have increasingly large numbers of new environmental BW agents emerging all around us. For some of them, it is a foregone conclusion that we have little or no natural immunity developed.

So we are going to enter an increasingly harsh environment, characterized by the rise of great epidemics and pandemics of many kinds.

Inevitably some of these coming scourges are going to rival the "Black Death" in ferocity and lethality.

Effective defense against one or more **such BW viral attacks — either against our forces in the field or against our civilian population — can only be provided by a ready, fully developed technology of electromagnetic healing.**

It does not really matter whether the BW strikes are performed by a hostile foe, a madman, natural mutations within our now drastically stimulated viral environment, or as a result of some particular "residue" discarded by an enthusiastic high school or college student.

The end result — the decimation of humankind — is going to be the same unless we develop and use electromagnetic healing technology.

However, with the present enormously self-hostile state of mankind, one or more enemies or madmen will almost certainly launch a strong, clandestine BW strike against the United States Armed Forces and the U.S. populace.

With EM healing technology, a specific "antidote" to any viral disease — known or unknown **— can be developed in a matter of hours.**

Mass electromagnetic treatment can then be started immediately, even for an entire army in the field or for our population in the heartland. Within hours of recognizing a BW strike, cure rates above 90% can be quickly achieved in exposed personnel.

Further, using electromagnetic healing technology, if necessary a more conventional vaccine can also be rapidly developed — say, in two weeks — for mass distribution and inoculation of the civilian populace.

The bottom line is this: Neither the U.S. Armed Forces nor the U.S. civilian populace can or will survive the coming lethal viral warfare strikes unless we quickly develop electromagnetic healing technology and have the necessary diagnostic and treatment devices already ready and waiting.

The clandestine AIDS strike by the Soviet Union is just the precursor of a great deal more to come.

The USSR Has Repeatedly Verified Our Nonreaction

The Soviet Union already knows that Western science and governments are too stereotyped and dogmatic to react to anything highly unorthodox.

For years, the Soviet have been proving this, over and over, in their development and testing of frightful scalar electromagnetic superweapons of unprecedented power. In the face of overwhelming evidence of this, the West has kept its head buried in the sand like an ostrich.

It wasn't invented in the West. Therefore it cannot be.

Years of substantial Communist indoctrination of students in our universities, by leftist/liberal professors, has placed a massive imprint into our society. An entourage of sympathetic groups has resulted — groups which believe that the Soviet leaders are just like us, that the Stalinist era is long since over, and that the Communist leopard has changed its spots and really isn't trying to take over the world anymore.

As a system, then, much of the West simply refuses to accept anything unorthodox, anything "not invented here", and anything that indicates Soviet aggression.

The Soviets know the U.S. is totally ignorant of advanced scalar EM weaponry. They have proven over and over that the U.S. will not even comprehend that it has been deliberately struck, if the strike is sufficiently unorthodox.

The Soviet scalar EM superweapons have already been documented in **Fer-de-Lance,** Tesla Book Company, 1986, and that documentation will not be repeated here.

Only a short summary of events will be presented, to put the Soviets' confidence into perspective. A longer chronology of selected events is included as Appendix IV.

31 Soviets Knew The U.S. Wouldn't React

The Soviets were quite confident in arranging for a great biological warfare strike, because of the total lack of the U.S. to comprehend incidents such as the following:

1. Beginning in the late 1950's, hidden Soviet transmitters across the street targeted "weak microwave radiation" against the U.S. Embassy in Moscow. Over the years the radiation continued, sometimes continuously, sometimes intermittently. The Soviets electronically induced diseases that resulted in extensive health problems of Embassy personnel, the death of two U.S. Ambassadors, and the sickness of a third. Several U.S. Presidents have protested the radiation. Both U.S. Intelligence and the U.S. scientific community have been totally baffled. They still do not know the true nature of the radiation, nor its purpose.

2. In the winter of 1957-58, an unexplained nuclear event happened near Kyshtym, within the Soviet Union. A huge explosion occurred in the stored nuclear wastes in the disposal section of the Soviet atomic weapons industry located in the southern Urals. Radioactive contamination covered 1,000 square miles. For years the story was opposed — or suppressed — by U.S. intelligence agencies, perhaps to allay public resistance to the U.S. nuclear industry. At the time, Soviet experimentation with large scalar electromagnetic transmitters was well underway. A large transmitter probably failed, discharging its built-up, giant electrogravitational potential (EGP) into the earth. Such a pulse would have propagated through the earth as a wave; the first nuclear material contacted would have been fissioned immediately. The stored nuclear wastes probably exploded due to receipt of an EGP from the inadvertent failure of one of the Soviet Union's new superweapons in development. Survivors reported that indeed the nearby nuclear wastes exploded.

3. In January 1960 Khrushchev announced to the Presidium that a new fantastic weapon was in development. He stated that the weapon was so powerful that, if unrestrainedly used, it could wipe out all life on earth.

Soviets Knew The U.S. Wouldn't React

Figure 4. The U.S. Embassy in Moscow. The Soviet Union has bombarded this **facility** with weak microwave radiation for decades.

Table 2. MICROWAVE RADIATION **OF THE U.S. EMBASSY** IN MOSCOW

- **SINCE 1959 OR 1960**
- **HIGH LEVEL TARGET (U.S. AMBASSADOR)**
- **GUARANTEES PERSONAL ATTENTION OF**
    - **PRESIDENT**
    - **NSA, CIA, DIA**
    - **TOP CONSULTING SCIENTISTS**
    - **STATE DEPARTMENT**
    - **LEADING SCIENTIFIC INSTITUTIONS**
    - **NATIONAL SECURITY COUNCIL**
    - **ETC.**
- **TWO AMBASSADORS DIED, ANOTHER SICKENED**
- **U.S. REACTION REVEALS ANY KNOWLEDGE OF TESLA ELECTROMAGNETICS**
- **FOUR U.S. PRESIDENTS REQUESTED SOVIETS TO CEASE**
    - **CUT FROM 18 WATTS/CM^2 TO 2**
    - **THEN AGAIN INCREASED**

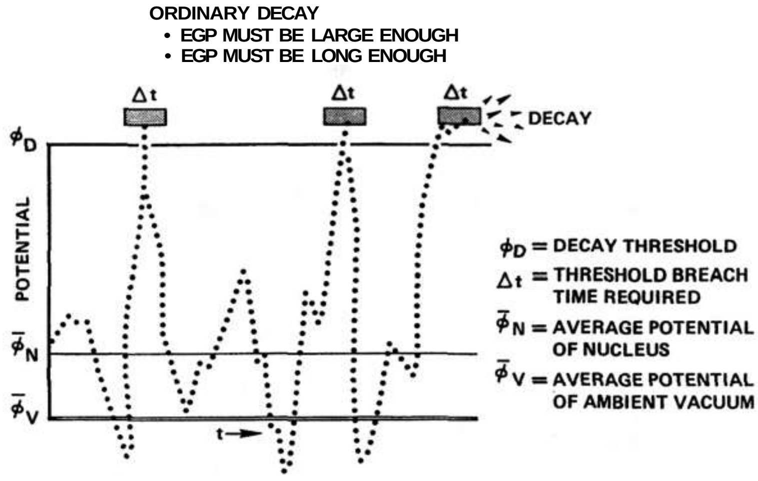
ORDINARY DECAY
• EGP MUST BE LARGE ENOUGH
• EGP MUST BE LONG ENOUGH
Δt
Δt
Δt
DECAY
ϕD
POTENTIAL
ϕ̄N
ϕ̄V
t→
ϕD = DECAY THRESHOLD
Δt = THRESHOLD BREACH TIME REQUIRED
ϕ̄N = AVERAGE POTENTIAL OF NUCLEUS
ϕ̄V = AVERAGE POTENTIAL OF AMBIENT VACUUM

Figure 5. Decay of a radioactive nucleus. When it absorbs a scalar EM wave (electrogravitational pulse) that is long enough and large enough, the nucleus fissions.

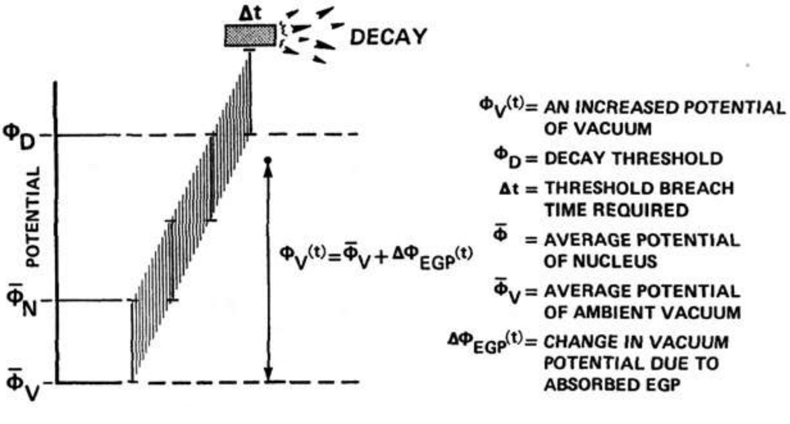
Δt
DECAY
ΦD
POTENTIAL
Φ̄N
Φ̄V
ΦV(t)=Φ̄V+ΔΦEGP(t)
ΦV(t)= AN INCREASED POTENTIAL OF VACUUM
ΦD = DECAY THRESHOLD
Δt = THRESHOLD BREACH TIME REQUIRED
Φ̄ = AVERAGE POTENTIAL OF NUCLEUS
Φ̄V = AVERAGE POTENTIAL OF AMBIENT VACUUM
ΔΦEGP(t)= CHANGE IN VACUUM POTENTIAL DUE TO ABSORBED EGP

Figure 6. EGP fission of a radioactive nucleus. If an initial EGP is large enough and long enough, it will fission any and all radioactive nuclei it strikes.

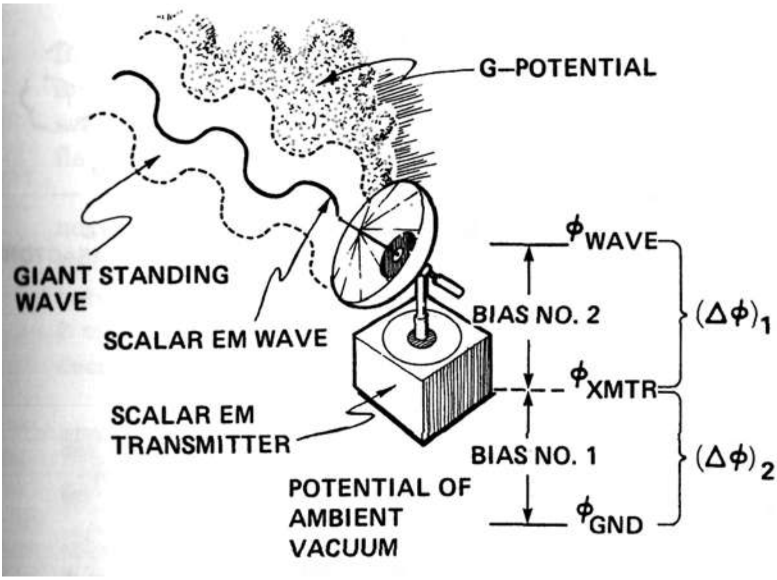
G–POTENTIAL
GIANT STANDING WAVE
SCALAR EM WAVE
SCALAR EM TRANSMITTER
POTENTIAL OF AMBIENT VACUUM
ϕ_{WAVE}
BIAS NO. 2
ϕ_{XMTR}
BIAS NO. 1
ϕ_{GND}
$(\Delta\phi)_1$
$(\Delta\phi)_2$

Figure 7. Bottled up energy in a standing scalar EM wave is a potential.

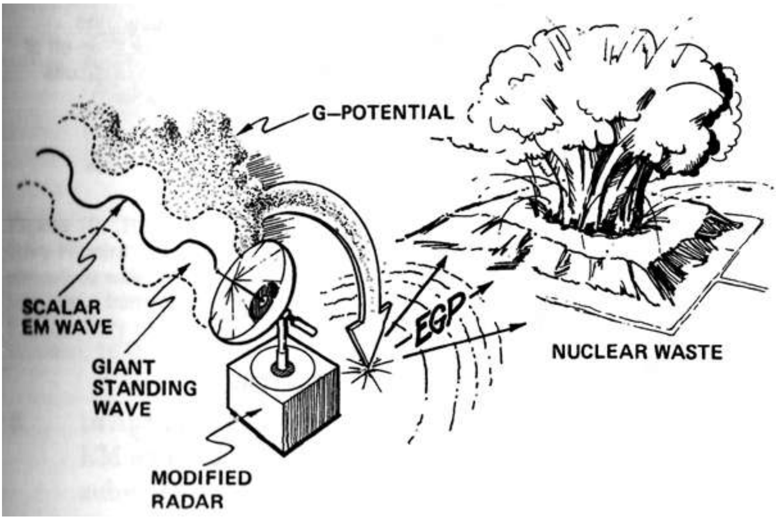
G–POTENTIAL
SCALAR EM WAVE
GIANT STANDING WAVE
MODIFIED RADAR
EGP
NUCLEAR WASTE

Figure 8. EGP short-out Transmitter failure and consequent discharge of a giant standing scalar EM wave potential caused the explosion of stored nuclear wastes at Kyshtym, USSR, in the winter of 1957-58.

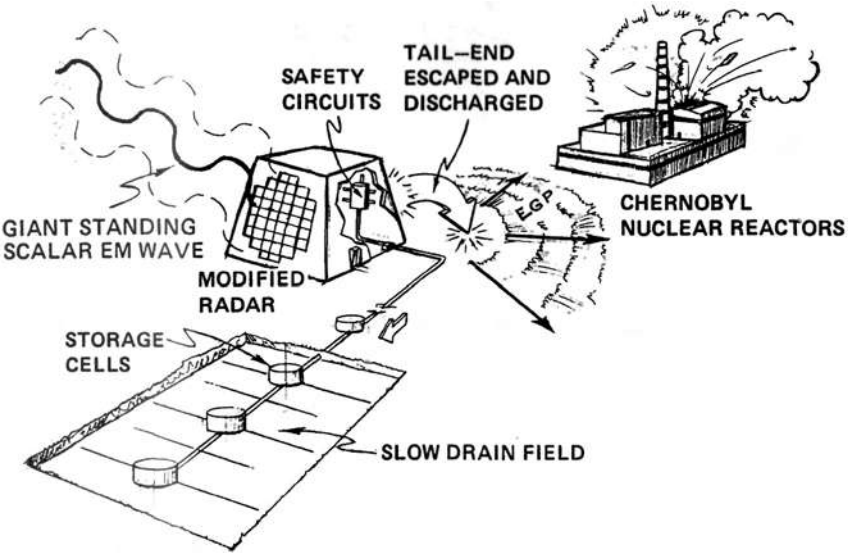
TAIL–END ESCAPED AND DISCHARGED
SAFETY CIRCUITS
CHERNOBYL NUCLEAR REACTORS
GIANT STANDING SCALAR EM WAVE
EGP
MODIFIED RADAR
STORAGE CELLS
SLOW DRAIN FIELD

Figure 9. EGP initiation of the Chernobyl event. Failure of another giant transmitter near Chernobyl in 1986 eventually caused the eruption of one of the Chernobyl reactors by EGP-pulsing its nuclear fuel rods. When the transmitter failed, the alarmed Soviets shut down the four Chernobyl reactors while waiting for drain-off of the EGP potential by the safety circuits of the disabled transmitter. These circuits finally gave way, producing a mild EGP pulse into the earth. Had they failed immediately, all four reactors at Chernobyl would have violently exploded.

Table 3. KHRUSHCHEV'S 1960 STATEMENT

(Speaking to the Presidium)

WE HAVE A NEW WEAPON,
JUST WITHIN THE PORTFOLIO OF
OUR SCIENTISTS...
SO POWERFUL THAT,
IF UNRESTRAINEDLY USED,
IT COULD WIPE OUT
ALL LIFE ON EARTH!

4. In May 1960 Gary Powers' high-flying U-2 spy plane was mysteriously disabled over the Soviet Union, resulting in its subsequent crash and the Soviet capture of Powers himself. A flash appeared behind Powers' aircraft, as if from an explosion — except that *the flash persisted,* showing that it was not a normal HE explosion at all. A tracking radar, employed in a scalar EM mode, probably was used to cause the electromagnetic "flash ball" and disable Powers' aircraft. Harmless SA-2 missiles fired far below him may have been employed to deceive Western intelligence analysts.

Figure 10. The U-2 high altitude reconnaissance aircraft. On May 2,1960 Francis Gary Powers high flying U-2 spyplane was disabled over the Soviet Union. Some difficulties with the autopilot were first experienced. Then a single dull thump was felt, and a tremendous orange flash lit the cockpit and sky — *and persisted.* (See Francis Gary Powers with Curt Gentry, **Operation Overflight,** Holt, Rhinehart and Winston, 1970, p. 82.)

5. In April 1963 the Soviet Union utilized a newly deployed scalar EM weapon to destroy the submerged U.S.S. Thresher nuclear submarine, off the coast of the U.S. itself. In spite of intense and totally anomalous electronic jamming experienced by the U.S.S. Skylark, surface companion to the Thresher, the U.S. failed to recognize this major indication of a scalar EM attack. *One day later, a Soviet scalar EM weapon placed a giant*

underwater burst in the ocean, 100 miles north of Puerto Rico. With these two tests Khrushchev, still smarting from his Cuban Missile Crisis facedown by John Kennedy, avenged himself and convinced the Presidium to allow him to remain in power for yet awhile.

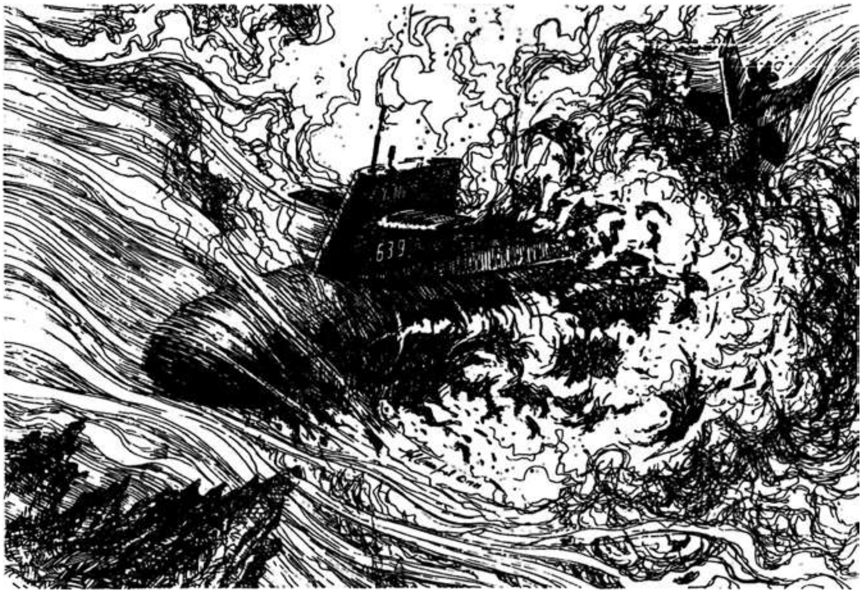

Figure 11. The U.S.S. Thresher disintegrates. © Hal Crawford 1979

Table 4. DEATH OF THE U.S.S. THRESHER

APRIL 10, 1963

- **DURING TRIALS OFF EAST COAST OF U.S. IN ATLANTIC**
- **SURFACE COMPANION: U.S.S. SKYLARK**
- **UNDERWATER PHONE BETWEEN THE TWO**
- **THRESHER DIVED AND WAS LOST**
    - **GARBLED MESSAGE**
    - **DID NOT RESPOND TO CONTROLS**
    - **SANK TO CRUSH DEPTH**
    - **IMPLODED**
- **SKYLARK ENCOUNTERED SEVERE ELECTRONIC INTERFERENCE**

Table 5. SEVERE EM INTERFERENCE EXPERIENCED BY U.S.S. SKYLARK

APRIL 10,1963

- **SKYLARK JAMMED (1.5 HR TO TRANSMIT MESSAGE)**
- **MULTIPLE SYSTEMS**
- **SIGNATURE:**
    - **SCALAR INFEROMETRY**
    - **SKYLARK IN "SPLATTER" ZONE**
    - **THRESHER PROBABLE TARGET**
- **OTHER SYMPTOMS**
    - **ONE SYSTEM KNOCKED OUT**
    - **OTHERS THEN WENT OUT**
    - **LATER RESUMED WORKING**

6. In Afghanistan, in 1981 and at intervals since then, Soviet helicopters have sometimes attacked Afghan guerrillas or villages with a strange new weapon. A body struck by this weapon dies instantly and completely. It falls like a limp rag, **and** does not decay even in 30 days. Every cell, bacterium, virus, etc. in the body has been killed instantly. Nerve gas rockets and chemical rockets are also usually fired to provide **a** "deception cover" for the nature of the real weapon. This eery Soviet weapon has been mislabeled as some strong new gas, **and** is referred to as "Smirch", a combination of the two words "smert" (death) and "semerch" (sandstorm).

7. In April 1984 a monstrously huge scalar EM incident occurred off the coast of Japan, near the Kuriles. Subsequent interviews with the pilots sighting the incident have revealed that several combined tests of the various modes of a scalar EM howitzer occurred. A giant "cold explosion" occurred, raising a giant mushroom cloud to over 60,000 feet and greater than 150 miles diameter in two minutes. A glowing dome appeared, and expanded to monstrous size. According to Dr. Walker and colleagues, **the diameter of this great** glowing, **hem-**

ispherical **shell expanded to several hundred kilometers.** * He and his colleagues were able to rule out any possible known natural events or mechanisms that could have caused the incident.

8. Numerous incidents of hemispherical shells of light and giant balls of light within the Soviet Union have been seen from Iran and from Afghanistan. Such giant artificial phenomena have also repeatedly been seen over the oceans by mariners and airline pilots.

9. Several instances of anomalous disabling of jet engines of commercial airliners have occurred while the aircraft were in flight.

10. Direct and intensive weather engineering over the U.S. itself has been accomplished by the Soviet Union for a decade, using the interference grid resulting from the crossing of two or more giant beams from the woodpecker "over-the-horizon radar" transmitters. Specific cloud signatures (anomalous holes, giant radials, twin giant radials, and razor-sharp giant interference patterns) have continually occurred, but the orthodox system has blithely ignored them.

11. A giant test of over a hundred huge Soviet scalar EM weapons occurred in and around May 1, 1985 for the 40th anniversary of the end of WWII. The entire Western scientific and intelligence establishment remained sublimely unaware of the giant exercise. They had not a single scalar EM signal detector — even though the air, land, and ocean was filled with hundreds of Soviet scalar EM transmitters, chattering away to Soviet armed forces, submarines under the water, etc. Underneath our feet, the earth itself was in giant scalar EM resonance on 54 frequencies — 27 pairs, each of which tapped energy directly from the molten core of the earth and fed it to an ensemble of giant Soviet strategic scalar EM weapons.

---

*See Science, 227 (4584), Feb. 8, 1985, p. 607-611; Science 234 (4775), Oct. 24, 1986, p. 412-413.

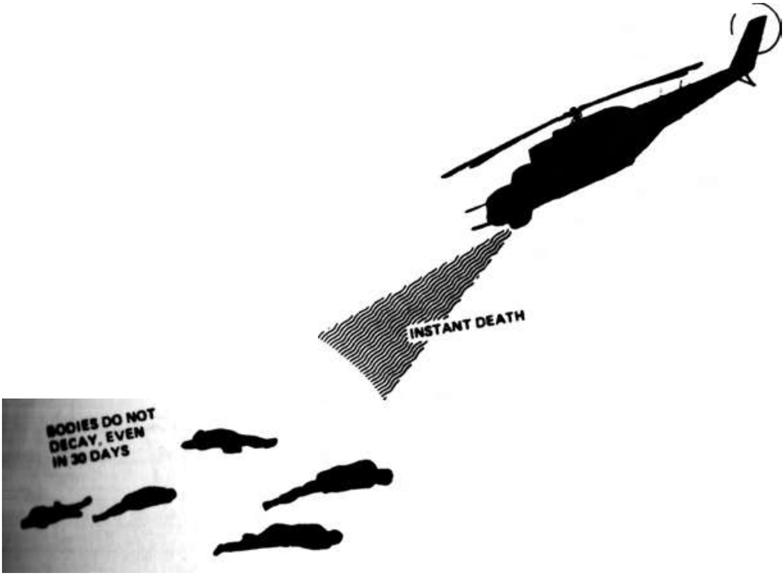
INSTANT DEATH
BODIES DO NOT DECAY, EVEN IN 30 DAYS

Figure 12. Soviet helicopters have used a "death ray" in Afghanistan.

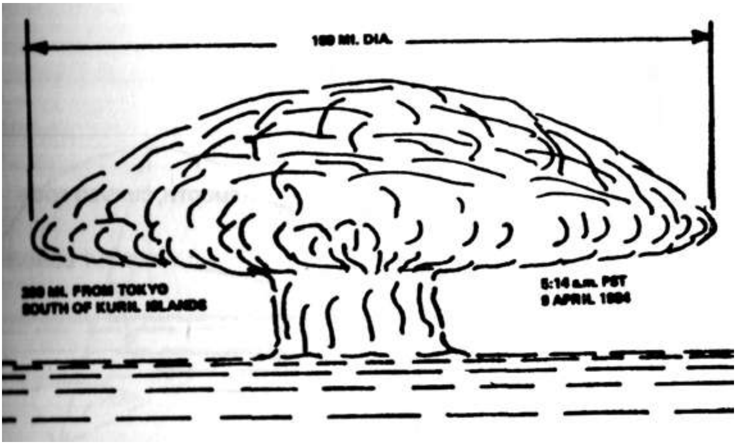
180 MI. DIA.
200 MI. FROM TOKYO
SOUTH OF KURIL ISLANDS
5:14 a.m. PST
9 APRIL 1984

Figure 13. Mushroom cloud from the sea off Japan, near a Russian test area.

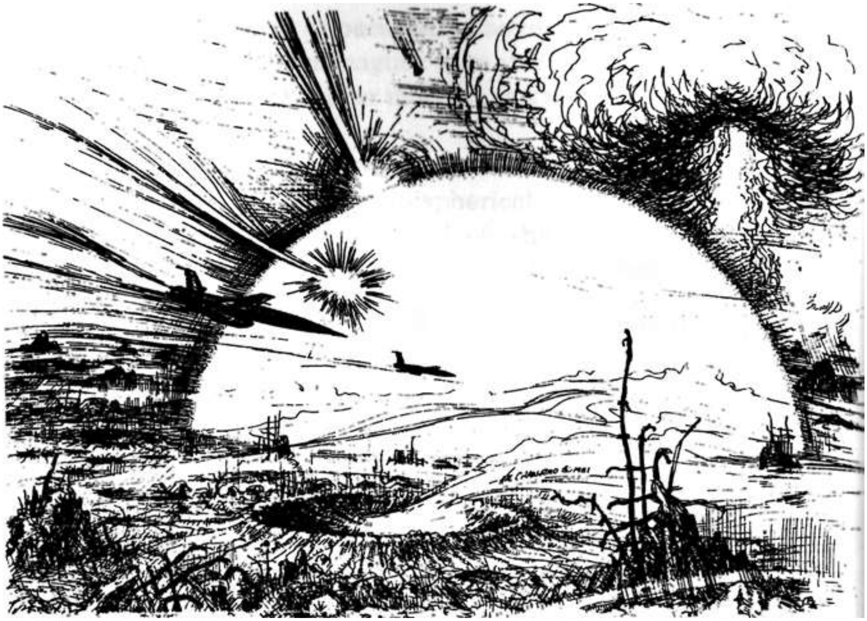

©Hal Crawford 1981

Figure 14. The Tesla shield, a giant hemispherical shell of energy. Seen repeatedly in Soviet tests in out-of-the-way places and remote regions of the ocean.

8:15 PM
THURSDAY
AUG 1969
VIRGIN ISLANDS
SEEN BY MANY RESIDENTS

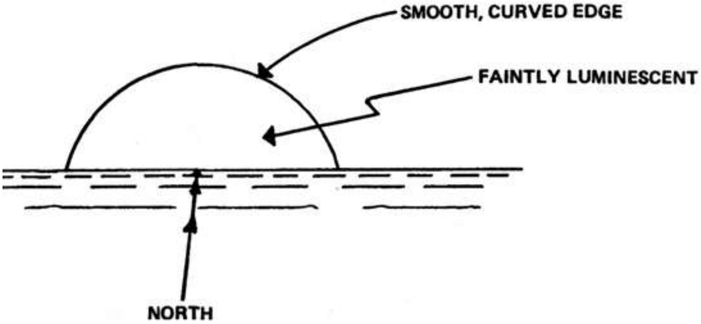
SMOOTH, CURVED EDGE
FAINTLY LUMINESCENT
NORTH

Figure 15. The Tesla shield: A 1969 Virgin Islands incident.

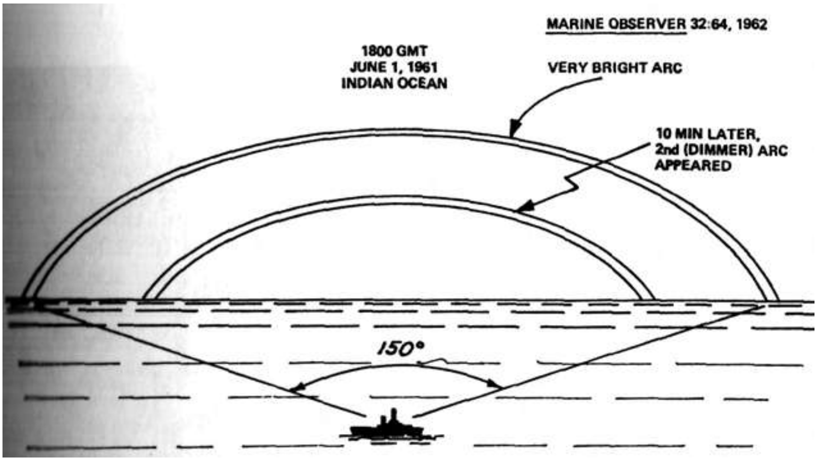
MARINE OBSERVER 32:64, 1962
1800 GMT
JUNE 1, 1961
INDIAN OCEAN
VERY BRIGHT ARC
10 MIN LATER,
2nd (DIMMER) ARC
APPEARED
150°

Figure 16. Tesla shields: Two nested arcs of light in the sky.

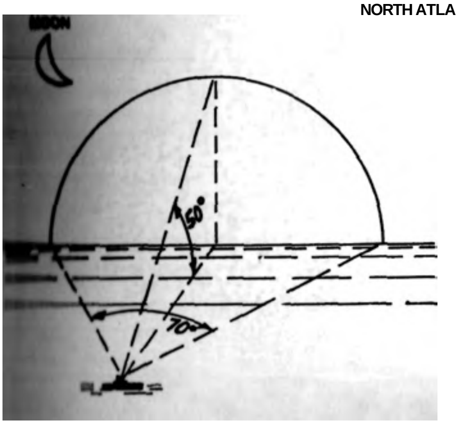

MAR 20, 1968
AND WESTERN
NORTH ATLANTIC

MARINE OBSERVER 40:17.1970

- **SEMICIRCLE**
- **MILKY WHITE LIGHT**
- **QUITE BRIGHT AT FIRST**
- **DIMMED AS ENLARGED**
- **LASTED 10 MINUTES**

Figure 17. Tesla shield: An expanding hemispherical circle of light.

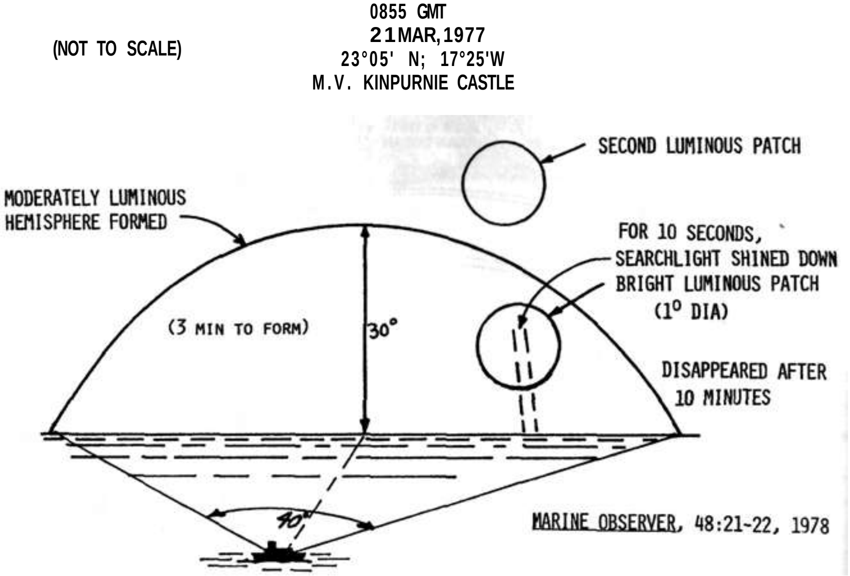
(NOT TO SCALE)
0855 GMT
21MAR,1977
23°05' N; 17°25'W
M.V. KINPURNIE CASTLE
SECOND LUMINOUS PATCH
MODERATELY LUMINOUS
HEMISPHERE FORMED
FOR 10 SECONDS,
SEARCHLIGHT SHINED DOWN
BRIGHT LUMINOUS PATCH
(1° DIA)
(3 MIN TO FORM)
30°
DISAPPEARED AFTER
10 MINUTES
90°

MARINE OBSERVER, 48:21-22, 1978

Figure 18. Tesla shield: Hemisphere and globes, 1977.

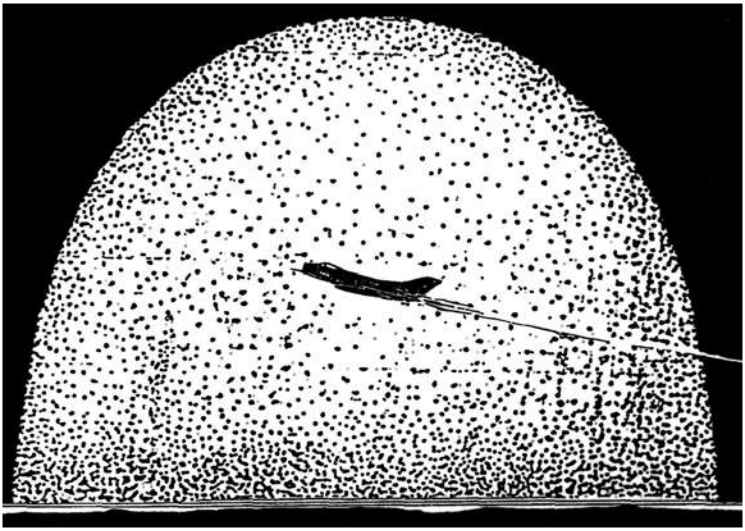

Figure 19. Large Tesia Shield observed by Red Chinese fighter pilots. Electrical difficulties in the aircraft were also experienced.

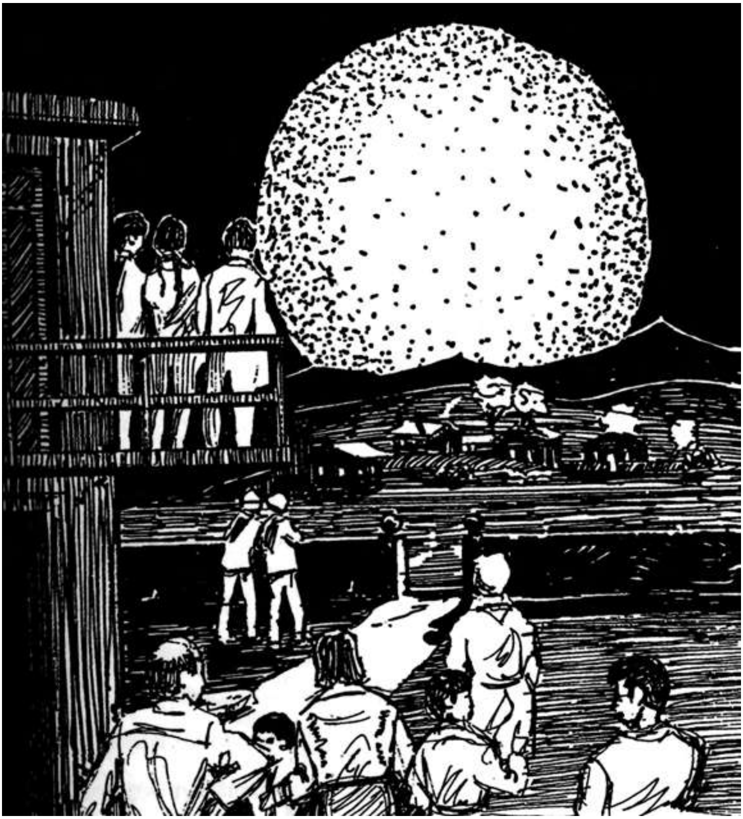

Figure 20. Large glowing Tesla Globe," Red China. Observed by hundreds of **persons.**

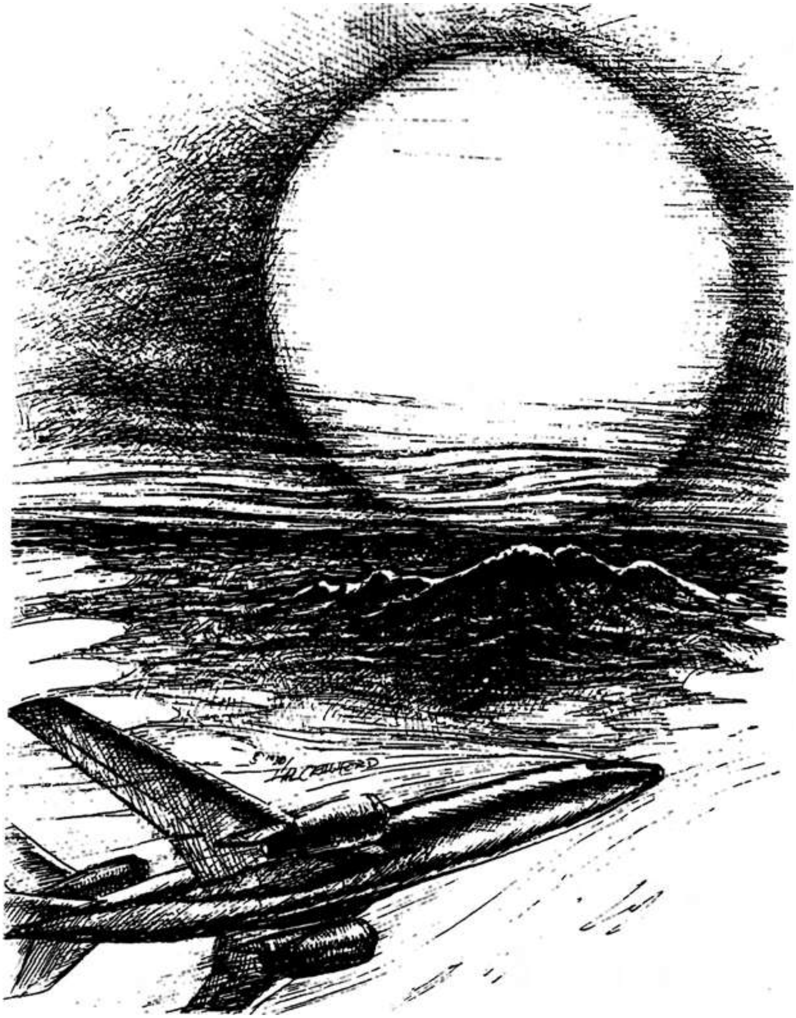

©1980 Hal Crawford

Figure 21. 1966 test of giant Tesla Globe, expanding deep within the Soviet Union. Seen from Teheran, Iran June 17, 1966 by two aircraft approaching Mehrabad Airport. Observed for several minutes.

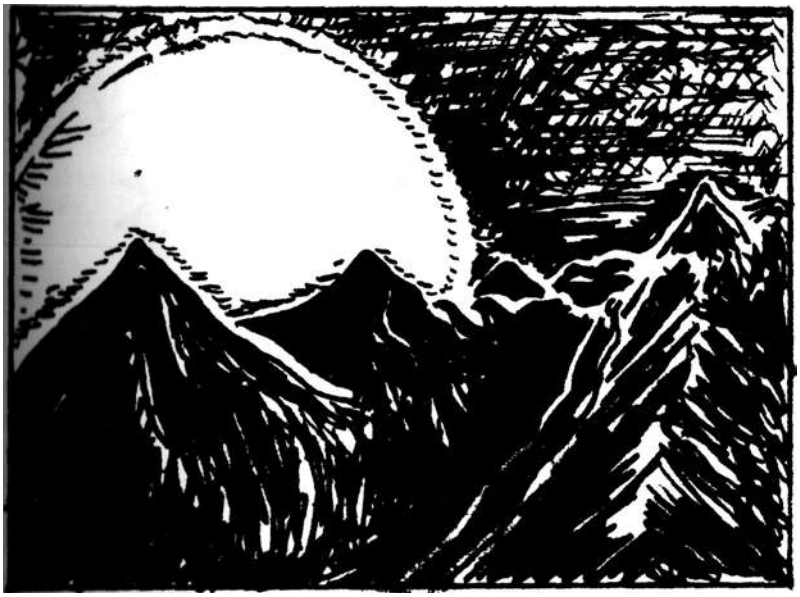

Multiple events seen in Sept. 1979 from Afghanistan

Strange lurid glow that flared silently over the Hindu Kush; as described by
Nick Downie.
The LONDON SUNDAY TIMES, 17 August 1980

Figure 22. A giant globe, deep within the Soviet Union in 1979. Seen several times frorn Afghanistan, looking toward Saryshagan.

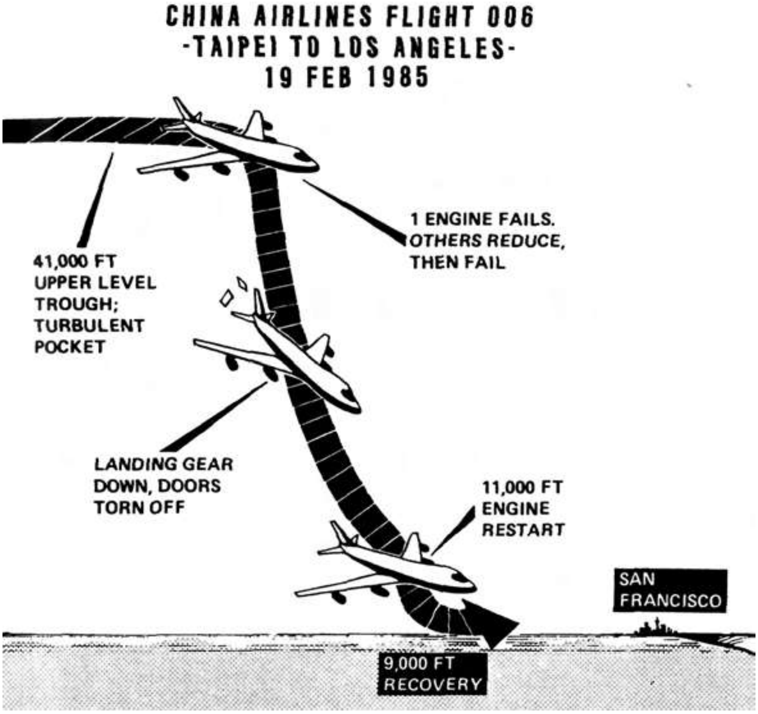
CHINA AIRLINES FLIGHT 006
-TAIPEI TO LOS ANGELES-
19 FEB 1985
41,000 FT UPPER LEVEL TROUGH; TURBULENT POCKET
1 ENGINE FAILS. OTHERS REDUCE, THEN FAIL
LANDING GEAR DOWN, DOORS TORN OFF
11,000 FT ENGINE RESTART
SAN FRANCISCO
9,000 FT RECOVERY

Figure 23. Engines disabled, a passenger jet drops six miles in two minutes.

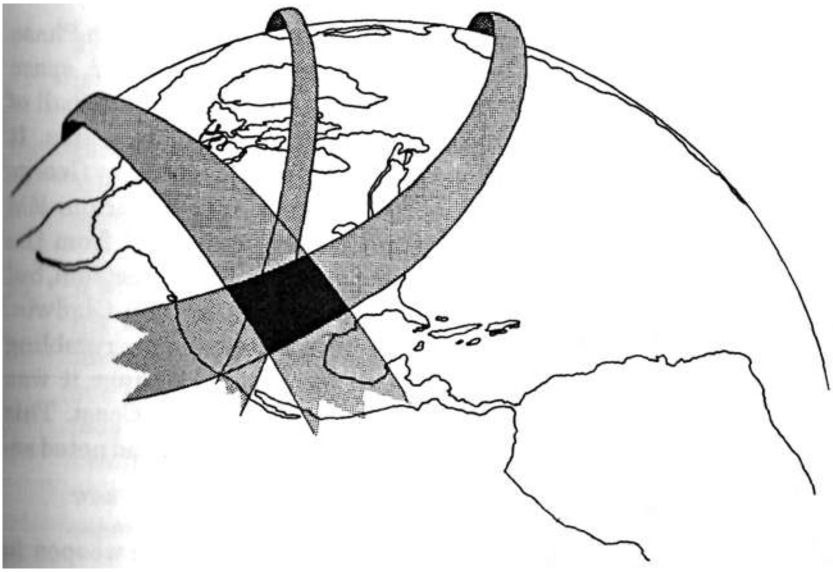

Figure 24. Giant "Woodpecker" beams intersecting over the U.S. Used for weather control, as a launch phase ABM and anti-bomber system, electromagnetic attack of distant populations, and disease conditioning. Using phase conjugate radar mirrors, any amount of energy desired can be delivered to any distant target from which a radar return has been received. Real-time distant holography, to produce powerful energy forms at a distance, can also be accomplished.

Table 6. MASSIVE 1985 STRATEGIC MAY DAY EXERCISE

APRIL/MAY 1985

- **BREZHNEV'S 1972 STATEMENT FULFILLED**
- **40th ANNIVERSARY OF WWII**
- **GORBACHEV NEWLY IN POWER**
- **EARTH IN GIANT SCALAR EM RESONANCE**
    - **54 FREQUENCIES**
    - **27 "POWER TAP" PAIRS**
    - **4 TO 6 WEAPONS EACH PAIR**

- **C^3I SCALAR EM LINKS**
    - **UNDERWATER SUBS**
    - **NAVAL FORCES**
    - **OTHER FORCES**

12. In late November, 1985 the Soviets tested their Launch Phase ABM system over Florida, using the launch of a U.S. space shuttle as a convenient test target. A marker beacon (ball of light) was placed in the sky, hovering near the launch area. It was seen by hundreds of persons, and photographed by George Suchary. During the final countdown to launch, a scalar EM "missile" (ball of intense energy) was fired, offset from the shuttle. This ball of energy was not in the visible spectrum, but may have been accidently photographed by Bob Gladwin. Twelve minutes offset after launch, a tremendous rumbling overhead boom echoed through the area — so huge it was heard for hundreds of miles up and down the East Coast. This was the third U.S. shuttle launch in which NASA had noted associated anomalous phenomena, but did nothing.

13. On December 12, 1985 the Soviets used the same weapon in the anti-bomber mode to destroy an Arrow DC-8 jet just after its takeoff from Gander Air Force Base, Newfoundland. **Over 250 U.S. soldiers and civilian crew members were killed.** Gander has long been a registration point for the Soviet scalar EM weapons. For example, in one 4-day period in 1982, 36 anomalous incidents occurred in and around Gander, involving moving lights, lights in the sky, etc.

14. On January 28, 1986 the Soviets destroyed the U.S. space shuttle Challenger, shortly after its launch. A dozen signatures are associated with this Soviet kill, too many to mention all of them. To give only a few: First, the Soviet ships monitoring the launch site suddenly departed at high speed, four hours before the launch. This was unparalleled. Second, according to General Daniel Graham, *the following evening KGB headquarters gave a party to celebrate the success of the KGB's perfect active measures against the Challenger!* Direct and unmistakable indications of extreme localization of the Soviet transmitters to the Challenger's launch site existed on the day of the launch. In addition, a metal-softening signal was confirmed on the Soviet Woodpecker beams as early as

January 1,1986. Many other indicators also exist. The total is a very strong statement that the Soviets destroyed the Challenger with the same weapon they had tested against three previous shuttle launches, and which they had used to kill the Arrow DC-8 at Gander Air Force Base only six weeks earlier.

15. A whole series of anomalous missile failures occurred, involving Titan, Delta, and Ariane launches. The odds against all these launches failing in that pattern were millions to one. In one Titan failure, video taken of the event showed the anomalous "marker beacon" light in the vicinity, high overhead. Frame-by-frame analysis of the video showed that the light was indeed independent of the explosion, and it moved off separately after the rocket exploded.

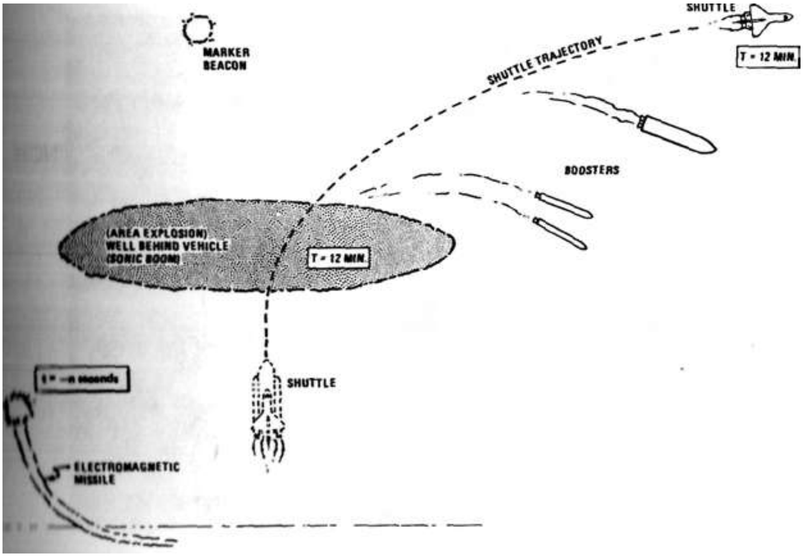
SHUTTLE
T = 12 MIN.
SHUTTLE TRAJECTORY
MARKER BEACON
BOOSTERS
(AREA EXPLOSION) WELL BEHIND VEHICLE (SONIC BOOM)
T = 12 MIN.
t = -n seconds
SHUTTLE
ELECTROMAGNETIC MISSILE

Figure 25. Third Soviet non destructive testing against shuttle launches, Nov. 26, 1985.

Table 7. SHUTTLE **LAUNCH ANOMALIES, NOV. 26, 1985**

- **MARKER BEACON**
    - **PHOTOGRAPHED**
    - **SEEN BY HUNDREDS**

- **OFFSET SCALAR EM MISSILE**
    - **INVISIBLE TO HUMAN EYES**
    - **STRIKE PHOTOGRAPHED**

- **GIANT ATMOSPHERIC BOOM**
    - **T + 12 MINUTES**
    - **HEARD FOR HUNDREDS OF MILES**

- **THIRD LAUNCH WITH SUCH A BOOM**

Table 8. SOVIET ESCALATION AFTER 3RD TEST AGAINST SHUTTLE LAUNCH on Nov. 26,1985.

- **U.S. STILL UNAWARE OF SCALAR EM WEAPONS USE**
- **SOVIETS ESCALATED ACTION**
- **DESTROYED OR DISABLED:**
    - **ARROW DC-8 (DEC. 12,1985)**
    - **CHALLENGER (JAN. 28, 1986)**
    - **TITAN 34-D (APRIL 18, 1986)**
    - **DELTA ROCKET (MAY 3, 1986)**

- **MAY HAVE DISABLED:**
    - **ARIANE II(MAY 31,1986)**
    - **TWO U.S. NUCLEAR SUBS WHICH RAN AGROUND**
    - **STEALTH FIGHTER WHICH CRASHED**

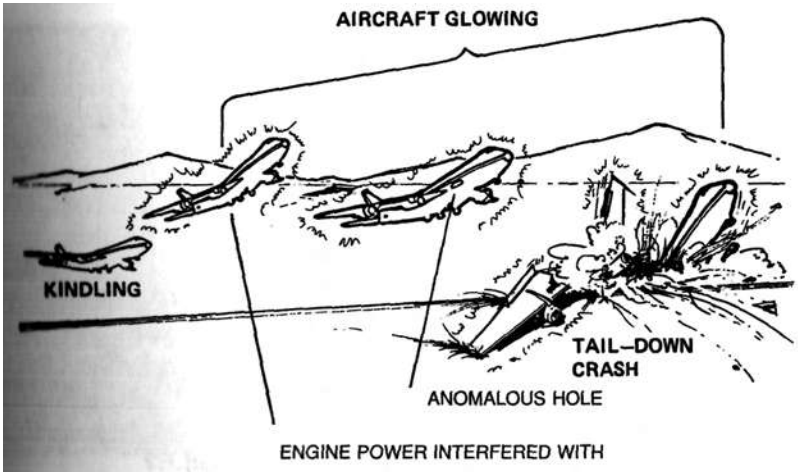
AIRCRAFT GLOWING
KINDLING
TAIL–DOWN CRASH
ANOMALOUS HOLE
ENGINE POWER INTERFERED WITH

Figure 26. Death of the Arrow DC-8, Dec. 12,1985. Three eye-witnesses saw the doomed airliner glowing. An anomalous hole was found in the fuselage, forward of the engines. Engine combustion was interfered with, by phase conjugate (negative) energy. Engines were rumbling laborously, not roaring with power, as the aircraft passed 100 ft directly over a witness.

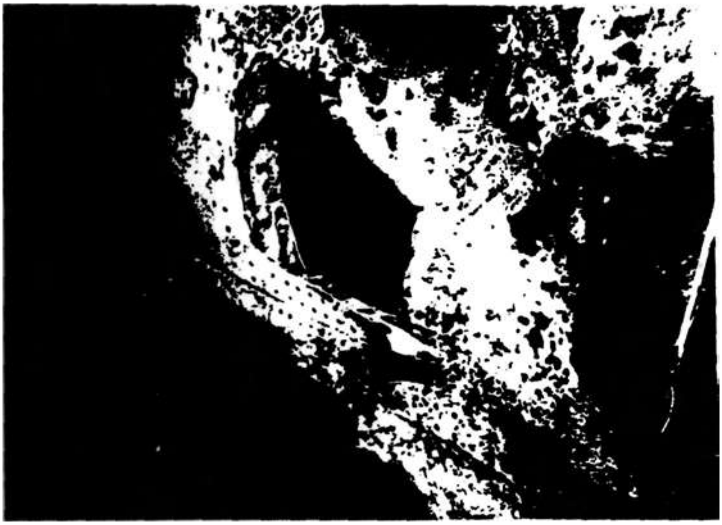

Figure 27. Anomalous hole in the right fuselage of the ARROW DC-8, ahead of the engines. With crossed over-the-horizon radar beams and multiple phase conjugate radar mirrors, real-time holography allows the production of a ball of EM energy at a great distance. The ball may be as small as desired, and contain as much energy as desired.

Table 9. ARROW DC-8 DISASTER, DEC. 12, 1985

GANDER AFB, NEWFOUNDLAND

- **AIRCRAFT LIFTED OFF**
- **AIRCRAFT WAS GLOWING YELLOW - ORANGE**
- **SANK TAIL-DOWN AND CRASHED**
- **SIGNATURE OF ENGINE/COMBUSTION INTERFERENCE**
  - **LAUNCH PHASE WEAPON**
  - **ANTI-BOMBER MODE**
  - **ANOMALOUS HOLE IN FUSELAGE, FORWARD OF ENGINES**
  - **STRUCK BY ELECTROMAGNETIC MISSILE?**
  - **EXTENSIVE "BALLS OF LIGHT" WEAPON TESTS IN AND AROUND GANDER**
- **CLOSELY FOLLOWED THIRD TEST AGAINST SHUTTLE NOV. 26,1985, IN ABM MODE**

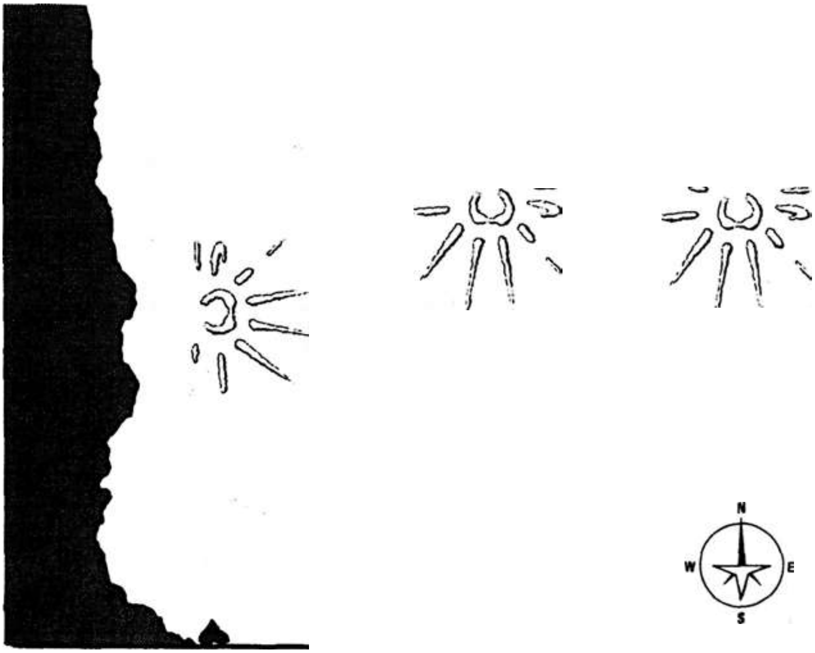

Figure 28. Part of the Soviet engineering of U.S. weather, just prior to the death of the Challenger. Three "giant radial" clouds sighted over the greater Los Angeles area by Margaret Wilson. Appeared one after the other, at 10 min. intervals.

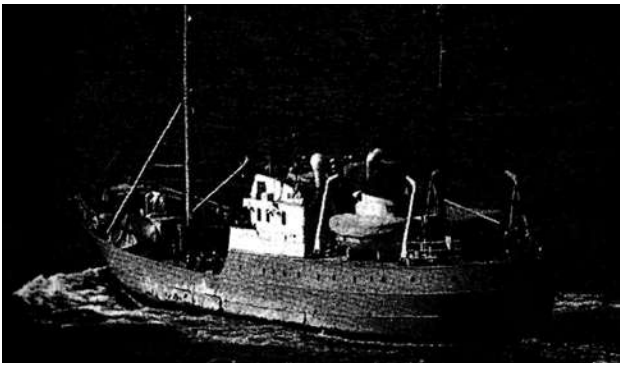

DOD PHOTO

Figure 29. Soviet Trawler spyship. About four hours before the launch of the Challenger, all Soviet ships off coast, that normally monitor each launch, suddenly and mysteriously departed at high speed. Unprecedented event.

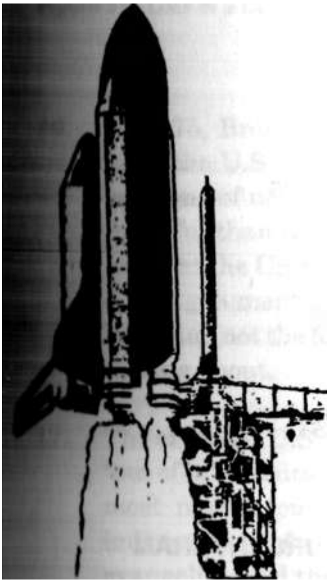

- **SOVIET WEATHER ENGINEERING**
- **SEVERE COLD STRESS**
- **METAL-SOFTENING**
- **BOOSTER SEAL PROBLEM**
- **SOVIET SHIPS ABSENT**
- **LOCALIZATION**
- **AIR TURBULENCE**
- **BIRDS NOT FLYING**

Figure 30. Launch of the doomed Challenger, Jan. 28,1986.

Table 10. DEATH OF THE CHALLENGER, JAN 28, 1986

- **METAL-SOFTENING SIGNAL ON WOODPECKER GRID AS OF JANUARY 1, 1986**
- **SOVIETS ENGINEERED COLD WAVE INTO FLORIDA**
- **SEVERE COLD STRESS ON VEHICLE**
- **GRID LOCALIZATION SIGNATURES**
    - **BIRDS NOT FLYING IN LAUNCH AREA**
    - **BIRD FALLS IN BIRMINGHAM, AL (HINGE)**
- **SOVIET SHIPS/TRAWLERS DID NOT TRACK LAUNCH**
- **BOOSTER SEAL PROBLEM SEVERELY AGGRAVATED**
- **CAUSED SHUTTLE FAILURE**
- **SOVIET KGB HEADQUARTERS CELEBRATED "PERFECT SUCCESS OF ACTIVE MEASURES" AGAINST THE CHALLENGER!**

Table 11. KGB CELEBRATION

EVENING OF JAN. 28,1986

- **AFTER DEATH OF CHALLENGER**
- **KGB HEADQUARTERS**
    - **JUBILANT PARTY**
    - **CELEBRATED SUCCESS**
    - **OF ACTIVE MEASURES AGAINST THE CHALLENGER**

SOURCE: GENERAL DANIEL GRAHAM, URGENTGRAM

- **NOTE: SOVIET SCALAR EM WEAPONS ARE CONTROLLED BY KGB**

- **SHORTLY AFTER LAUNCH**
- **ANOMALOUS POWER SURGE**
- **"COMMANDED" ENGINE SHUTDOWN**
- **VEHICLE THEN UNSTABLE**
- **RANGE SAFETY OFFICER DESTROYED ROCKET**
- **EXTERNAL RANGE SENSORS DID NOT RECEIVE SURGE SIGNAL**
- **SIGNATURE OF TIME REVERSED WAVE**
- **WHICH LOCALIZED ON ROCKET**

Figure 31. Loss of a Delta rocket, May 3, 1986.

16. In 1975, Brezhnev had surprisingly urged U.S. senators to have the U.S. cooperate in banning the development of new weapons of unparalleled horror. He used the phrase, "more frightful than the mind of man has ever imagined." Shortly after, at the United Nations, Gromyko had even introduced a draft agreement for all nations of the world to sign. Again, the West had not the foggiest notion as to what the Russians were talking about.

17. The orthodox U.S. news media has been, and is, almost totally "out of it." Ironically, at the time of our deadliest peril, and the most momentous period in history, incidents such as the indiscretions of a presidential candidate and of a television evangelist, and the "Iran Gate" actions of the National Security Council in trying to illegally get arms to the Contras, are the "burning issues of the hour" — and the week, and the month, and the year, and HO on ad nauseam.

18. The Congress — great watchdog and the direct representative of the U.S. citizenry — is equally immersed in interminable hearings of little or no substance except news media coverage and publicity.

19. A belabored president, misled and now hemmed in on all sides and on the defensive, has had his power to act greatly constricted. In addition, the orthodox filtering networks surrounding him have effectively isolated him. He has been unaware of the extensive information that shows just how certain it is that the Soviets have fantastic superweapons and have been gingerly using them against us for decades. He and his immediate advisors are fed highly filtered, shaped and interpreted data. U.S. intelligence is very good at doing the things that it does. However, it is also very good at ignoring the things that it ignores! Further, the intelligence community is a past master at "slick-tonguing" the interpretation of anything new, and anything it really doesn't understand, in terms such as "We have no evidence that ..." For decades, the community has consistently underestimated the Soviet threat, and it has also usually projected any unpleasant present Soviet capability in the terms that "If the Soviets keep this sort of thing up, it may prove to be a real threat 30 years from now!" For years U.S. intelligence was unaware of a Toronto inventor's development and public demonstration in the late 1960's of an inertial field generator device, for example, because it simply did not place high value on simple overt intelligence such as reading the newspapers! The entire event was printed in detail in a Canadian newspaper. Only very recently, within the last few years, has the community again realized that the most important intelligence can only come from human (HUMINT) sources. And in relying on the scientific community — which itself knows nothing of scalar electromagnetics — the intelligence community has not received the technical advice necessary for it to properly task its collection agencies to collect information on the Soviet superweapons.

20. Western science has also shown itself to be highly dogmatic and ensconced in the comfortable "status quo." It has regally ignored the important Soviet developments in unified field **theory.** It has assumed an aristocratic stance, regarding the Soviet scientists as little more than peasants still trying to clean the mud off their boots. In fact, the Soviet scientists have been ahead of the rest of the world in *nonlinear* science — the truly hard stuff— since the beginning. They have always been **the** best nonlinear mathematicians in the world. Soviet scientific papers regularly appear which deal with local general relativity and the overcoming of the sacrosanct conservation laws as a matter of course. That subject is still a "no-no" in U.S. scientific publications, even though breaking symmetry and consequent violation of conservation laws has long boon proven in particle physics. To my knowledge, with the exception of Hooper's obscurely published work, not a single US. scientific paper appears in the literature showing any detailed investigation of "bucking fields" as such and investigating the behavior of materials exposed to such "sum-zeroed" fields. Yet ironically a whole area of optical physics — phase conjugate optics (PCO) — has shown great activity since we discovered it in the open Soviet literature in the late 60's, and since two Soviet scientists visited Livermore* in 1972 and briefed us upon the subject (as another stimulus to see if we recognized the awesome weapon implications). In PCO, the scalar EM wave formed by two waves 180 degrees out of phase with each other, and locked (modulated) together by a nonlinear medium, is blithely called the *pump wave.* Using it on nonlinear materials produces extraordinary results — such as electromagnetic waves traveling backward in time. Further, the PC effect is known to be universal, and not just in the optics **region** at all; for example, it has also been done with sound waves. **The obvious experimental and theoretical proof of scalar EM has been in our literature for nearly two decades, and even in PCO our scientists do not yet realize that the** "pump" **wave is a scalar** wave, **a** wave **of**

***Lawrence Livermore National Laboratory**

artificial potential, and a gravitational wave that pumps the nuclei of the atoms in the nonlinear medium. They also still do not understand that it is the nucleus of the atom that produces the phase conjugate replica wave in PCO.

Table 12. BREZHNEV'S 1975 PROPOSAL

AT THE SALT TALKS ON JUNE 13, 1975

- **THE SOVIETS URGED THE U.S. TO AGREE ON A BAN AGAINST RESEARCH AND DEVELOPMENT OF NEW KINDS OF MASS DESTRUCTION WEAPONS MORE TERRIBLE THAN ANYTHING THE WORLD HAS KNOWN!**

- **BREZHNEV PERSONALLY URGED VISITING U.S. SENATORS TO AGREE TO THE BAN.**

- **GROMYKO INTRODUCED A DRAFT AGREEMENT INTO THE UNITED NATIONS MEETING AND URGED THE NATIONS OF THE WORLD TO SIGN IT.**

- **THE WEST HAD NO IDEA WHAT THE SOVIETS WERE TALKING ABOUT!**

- **SOVIET HAWKS THEN GAINED DOMINANCE, AND THE USSR STARTED ITS MASSIVE BUILDUP FOR EVENTUAL WORLD TAKEOVER.**

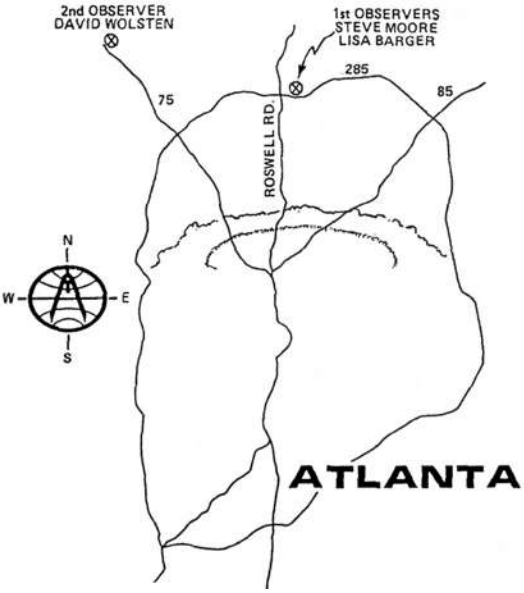
2nd OBSERVER
DAVID WOLSTEN
1st OBSERVERS
STEVE MOORE
LISA BARGER
285
85
75
ROSWELL RD.
N
W
E
S
ATLANTA

A. Location of the incident

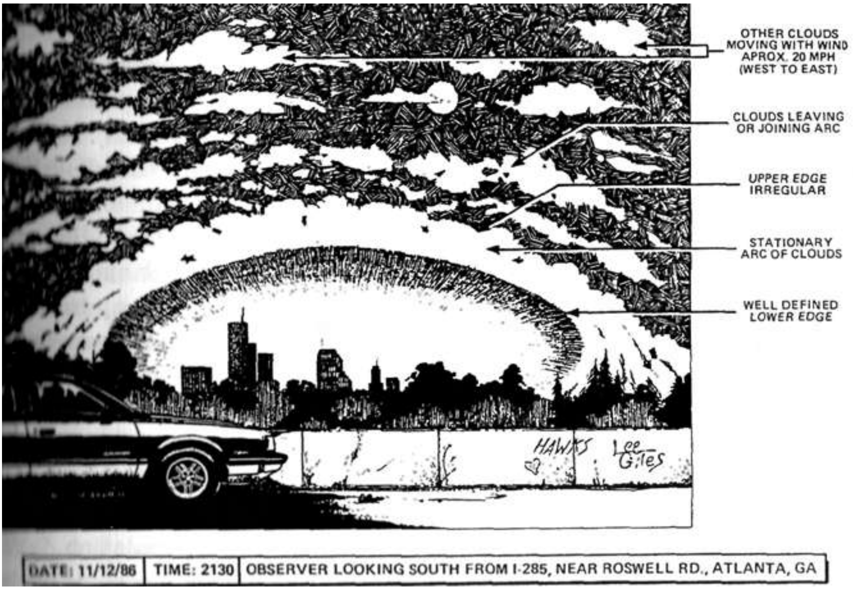
OTHER CLOUDS
MOVING WITH WIND
APROX. 20 MPH
(WEST TO EAST)
CLOUDS LEAVING
OR JOINING ARC
UPPER EDGE
IRREGULAR
STATIONARY
ARC OF CLOUDS
WELL DEFINED
LOWER EDGE
HAWKS
Lee
Giles
DATE: 11/12/86
TIME: 2130
OBSERVER LOOKING SOUTH FROM I-285, NEAR ROSWELL RD., ATLANTA, GA

B. Artist's sketch of the incident.

Figure 32. Probable Soviet test of a Tesla Shield over Atlanta, Georgia.

Continually assured that Western science and intelligence remained outrageously ignorant and would obstinately persist in wearing orthodox blinders, the Soviets felt quite safe in striking the U.S. and its Western allies a mortal death blow.

The Soviets knew there would be no retaliation, so long as the blow was sufficiently unorthodox.

After all, the Soviets have been waging a slowly increasing unorthodox war against us for decades, and totally getting away with it.

Consequently the Soviets manipulated the cancer research and the contamination of the World Health Organization's smallpox vaccine to produce the perfect biological warfare "first strike," administered unwittingly by the WHO itself in its mass immunization program to eradicate smallpox. The Soviets continue to manipulate the spreading AIDS viruses electromagnetically so that the West will be devastated, as well as Africa and the Third World.

At least one other spreading AIDS-strain has already appeared in Africa. Testing for exposure to the first strain in most cases did not detect exposure to this second strain.

Officially no one knows whether it will be as infectious as the first strain. However, a suspicious correlation has appeared between hepatitis vaccination in several large U.S. cities and consequent outbreak of AIDS among the persons inoculated.

Further, the latest studies are beginning to show that (1) AIDS may be just as spreadable, though more slowly spread, in the heterosexual community as in the gay community, (2) the AIDS tests themselves are far more unreliable than suspected; one may have the AIDS virus — and serve as an active carrier —for up to three years before the tests begin to detect it, and (3) no one has any real idea as to how many persons have already been infected.

As can be seen, even orthodox scientists are beginning to suspect a coming AIDS infection "asymptote" which may represent an explosion of the disease into a mass pandemic in the world population. In Africa, for example, AIDS is spreading freely among the hetersexual population as well as the homosexual population. * We may be on the verge of seeing a similar phenomenon emerge in the

---

*Although there is much speculation as to why African heterosexuals seem so vulnerable, at present no one knows why it is true.

Western world.

If so, we are looking full in the face of a forthcoming AIDS massacre that makes the 1918 influenza pandemic pale by comparison. Not only will AIDS decimate Africa and the Third World, but it will also decimate the West a little later.

In the Soviet plan, after the AIDS massacre has peaked, the remaining much smaller U.S. and European populations will be decimated and totally powerless. The demoralized and devastated Western nations will easily capitulate and pass under the Soviet hammer and sickle.

In their deception planning, the Soviets even are allowing some contamination of their own citizens by AIDS. Thus they can keep up *a pretense* that they are just as much victimized as we are, and just as much "in the dark" as we are about the whole thing.

Of course, once the AIDS infection does its lethal dirty work in the West, the Soviets can easily eradicate it in their own population at will, using the electromagnetic healing technology they already posess. The Soviet planners regard the loss of a few Soviet citizens along the way as one of the things necessary to insure the success of the Great Strike and the Final Revolution. Forward divisions in the standard Soviet attack are always regarded as such sacrificial pawns to guarantee the success of the follow-on echelons in the attack. In this BW strike, the sacrifice of some Soviet citizens is deemed necessary to the overall success of the attack.

Remember, these are the same type of Soviet communist planners who killed 20 million Kulaks — their own people — to forcibly collectivize Soviet agriculture. For their highest purposes, these planners are still quite capable of ruthlessly sacrificing millions of their own people to achieve their objectives.

The *Asian Tiger Mosquito: A New and Efficient Vector*

As almost everyone now realizes, in the past our government has greatly understated the AIDS problem. Past projections of the limitation of the spread of the disease have been far more optimistic than warranted.

For example, officially the government is still denying any possibility of insect vectors for the disease.

Don't believe it. In fact, highly efficient insect carriers of the disease are already on the way.

Insects in Africa have positively been verified to be harboring the AIDS retrovirus. Of course, such reports are always accompanied by the bland assurance that "We have no evidence... (that the insect bites will give AIDS to humans).

Let's see now: We're supposed to believe that a monkey bite will do enormously more than that. Even when the monkey didn't have the exact present form of the AIDS retrovirus. Even though monkey bites hadn't done the trick in thousands of years.

But we're supposed to turn right around and believe that an insect carrying the live AIDS retrovirus can bite you, inject into you some of his body fluid containing AIDS viruses, and you won't stand any chance of developing AIDS .

What are we supposed to believe? That the injected AIDS virus from an insect bite won't bother you, but the injected AIDS virus from other causes will?

Nonsense!

Anyone who believes that is a prime candidate for purchasing that bridge down the road...

Of course insect vectors will give you AIDS!

And we've got a real tiger — mosquito, that is, on the way to do just that.

The Asian Tiger mosquito has been introduced into the U.S. (it just "accidentally" appeared, of course!) and is now spreading through the Southeastern states. This vigorous and voracious mosquito prefers man as its target, and has an unusually wide range.

Recently U.S. scientists have reluctantly confirmed the presence of the AIDS virus in the Asian Tiger mosquito. However, as usual we are presented with the same tired refrain, "We have no evidence that... the mosquito's bite is capable of causing a human to develop AIDS."

Of course, you can be sure that eventually we *will* have evidence of precisely that!

The AIDS virus is now known to be able to live outside the body for up to at least 15 days, if not longer.

Also, the AIDS virus is an unstable virus — it's constantly mutating and changing its form. That means it's "highly adaptable" **- that** is, it's continually adapting to its environment, at a rate millions of times faster than normal stable viruses change and

That fact implies that some mutant strain of the virus will almost certainly be adapted to living in the Asian tiger mosquito and being introduced into its human host by the mosquito's bite.*

It's not a matter of "Can it do that?" or "We have (as yet) no evidence of that."; instead, it's a matter of "How long will it be until that occurs and we confirm it?"

The Asian Tiger mosquito apparently is intended to serve as the major vector (carrier), or at least one of them, to eventually rapidly increase the AIDS infection rate and devastate the U.S. population.

AS the vector spreads and the virus adapts to that mode of transmission, the rate of increase of the AIDS infection will itself accelerate. This means that all the projections for "doubling times" are seriously in error.

The latest data from Alabama, for example, shows that the disease is now doubling every six months instead of every year. Also, the doubling interval itself has been shortening, at least up to the **present.** Note also that, from the data of the new studies confirming a much longer lag time for the AIDS test to indicate the presence of the infection, even the present data on the number of persons infected and the doubling time are underestimated.

If these and other data are a true indication, it appears that we literally may have an "explosion" of the AIDS infection in no more than two to three years.

In addition, as we said the AIDS virus is an unstable virus. Mutations in the virus itself are occurring so rapidly that producing an effective vaccine essentially will be impossible. Without a totally new means applied to produce an AIDS cure, AIDS simply cannot be **stopped.**

•And one or more strains will eventually evolve that can be infectuously spread like the influenza virus is.

In fact, recenly it has been discovered that the AIDS virus is far more devious than previously susupected.* It attacks in more ways than suspected. It changes in the body, increasing its virulence as the disease progresses.

Macrophages, the immune system's scavenger cells, become infected and carry the virus to other body cells. It appears that macrophages may pass the virus to the T-cells, conditioning the destruction of the immune system. Infected macrophasges are found in rectal material of AIDS patients and in cervical fluid and semen. Thus through infected macrophages the AIDS virus is able to infect partners through exhange of semen, cerical fluid, and rectal material — even though there are no breaks or tears in the skin, and even though the number of free AIDS viruses in the fluids is too small to pass the infection.

Other immune system cells —such as Langerhans cells—are also suspected to be capable of serving as AIDS carriers when infected. The end result is that, as the AIDS patient sickens, the virulence of the AIDS virus increases, so that it attacks a greater variety of cells and can more easily infect them.

Present orthodox methods cannot cure AIDs in time, if ever.

Unless unorthodox technology is applied to solve the problem the West is already defeated.

---

*See Gina Kolata, "Fatal Strategy of AIDS Virus Grows Clearer," Science Times section, **New York Times,** Mar. 22, 1988; see also "Growth in AIDS Virus Virulence Found," **Los Angeles Times,** Apr. 1, 1988.

CHAPTER 3

WORLD WAR III IS RAGING AND WE'VE NEARLY LOST IT

More on AIDS Doubling Time

AS we said, the AIDS problem is much, much worse than officially announced. As stated, in some areas at least, the number of new AIDS cases diagnosed has been doubling every six months instead of every year. New studies have shown that the tests to identify infection by the AIDS virus may themselves have a 3-year lag period in which they do not reliably indicate the presence of the infection.

These facts probably make all previous "officially announced" estimates and projections off by several orders of magnitude (at least three).

If the incubation period (nominal) is, say, 5 years, then there is a factor of up to 2 (exp 10) of AIDS cases already in incubation.

The lag in accuracy of the AIDS tests however, may mean that there are even more AIDS cases in incubation than this.

And we may have a new entry into the equation at any time, once the AIDS virus adapts to its transmission by the Asian tiger mosquito and possibly even other insect vectors.*

Remember, we've got the U.S. data for only a relatively few years since 1981. So it's statistically difficult to project ahead with any accuracy. However, the best interpretation of the data we can make shows a spiraling (speed-up) trend in the rate of acceleration of the disease, indicative of a forthcoming asymptote. If that interpretation of the data holds true, the asymptote represents literally an "explosion" of the disease.

That is, the six-months "doubling interval" is slowly becoming **the s**ix-months "tripling interval," the six-months "quadrupling interval," etc.

The best indication/projection I can tentatively arrive at is this: If the total cure is not accomplished within about two to three years, then - considering the incubation ongoing by then and the speedup

---

***Another HIV virus has now been identified.**

ahead — somewhere between a third and two thirds of all humans on earth will perish.

In other words, nominally half the humans on earth will inevitably die, say within five years (the incubation) after that two to three year point.

Most of the people who die are going to be highly selected. Most of them are going to be Westerners, Africans, Asians, and Brown people.

And all of our tanks, ships, bombers, missiles, nuclear weapons, and armed forces put together can't save a single AIDS victim.

We've got more generals than ever before in our history, including World War II. And not a one of them has a single tactic or a single weapon to use against this mortal attack.

Understand, our generals have not failed us against the enemy they know and recognize. The bellies of our B-52 bombers and the aircraft aboard our distant aircraft carriers are filled with nuclear bombs sufficient to destroy the world many times over. Our powerful nuclear-tipped missiles sit in land-based silos and in sleek submarines cruising silently under the oceans, waiting to respond in an instant to hostile attack, again with the power to destroy the world many times over.

Yet the fine men and women who have built and maintained that awesome force, and who man it in constant vigilance to protect our nation, are absolutely vulnerable and defenseless against the smallest foe of all —just as were the Martians in H.G. Wells' **War of the Worlds.**

With all our awesome forces, we do not have the power to stop a single AIDS virus — or indeed, any other similar new BW virus.

We can be defeated with ridiculous ease by even the tiniest determined hostile nation — or even by a single terrorist, madman, or genetic accident by a high school youth.

And if there's one thing that we can learn from history with iron certainty, it is this: If we can be defeated with ridiculous ease, then most certainly we will be!

Since the strike has already occurred, then our defeat is just a matter of time if we continue our present courses of action.

All other threats — nuclear, chemical, scalar EM, the works — pale in comparison to this seemingly innocuous, doomsday BW viral weapon that has already been unleashed upon an unsuspecting world.

World War III is Already Raging

World War III has long been unleashed and a final, decisive first strike has now occurred. We've already lost the war unless unbelievably heroic measures are taken immediately.

This thing is deliberate. It's fiendishly clever. It's effective. And a smiling Gorbachev — master propagandist — is just part of the icing on the cake to deceive us until it's too late.

Ironically, Soviet announcements have cleverly started claiming that the U.S. has produced the AIDS virus and unleashed it.

That's a very clever propaganda trick to shift suspicion from themselves. Indeed, in the Third World the propaganda has been very effective. Most Third World peoples now give at least credence to the Soviet claim.

Unless we move immediately to effect a total cure, the armed forces and the populaces of all the Western nations will shortly be devastated — perhaps even worse than if they been attacked with nuclear weapons.

With our armed forces and populations devastated, we surrender. We lose.

The Soviets get the U.S. (and European) farmland. That's what they really need.

Of course they also will have rid themselves of about 200 million Americans or so along the way.

That, solves the old "winner's problem" of what to do with, and how to effectively control and care for, a conquered, resentful populace. That problem will have already been taken care of, for most of the U.S. populace will essentially have been eliminated.

There won't be many U.S. military prisoners, for example, since most of the military personnel will also have been eliminated by the virus or viruses clandestinely released.

Communism Is Savage Enough to Do Diabolical Things

The Communist's doctrine is to wage total war. Win any way, with any weapon, at any cost.

ALSO, communism has always taught that it is necessary to destroy capitalism. That equates with destroying the U.S., the major capitalist country. Communists have always taken the long view, willing to struggle for decades to eventually accomplish their aim.

If you think no modern nation could be so incredibly savage as to deliberately destroy untold millions of people, look at the actual record of the Communists.

When the Khmer Rouge took over Cambodia, they murdered a third of their own people.

The Soviets murdered some 20 million or so Kulaks when they collectivized the Soviet farms.

The Red Chinese also slaughtered Chinese peasants by the tens of millions — perhaps as many as twenty or thirty million — when they collectivized the farmers.

Lenin's followers have proved over and over that they are capable of murdering even their own people, on a scale of millions.

Dedicated communists are highly religious people — even fanatics. Deliberate murder to achieve communism's goals is part of the accepted dogma.

The communists are quite capable of killing a billion or more "foreigners" and "enemies," given the chance. Particularly when the goal is to win the "Final Revolution."

The opportunity to strike the decisive final blow without consequences, using as simple and unorthodox a weapon as the AIDS viruses, could simply not be ignored by men who have spent their lives trying to control the world.

The unparalleled strike opportunity was particularly irresistible since someone else (the Western cancer research institutes) would develop the weapon itself, someone else (the World Health Organization) would unwittingly deliver the strike, and the blow could easily be clandestinely arranged (simply by secretly contaminating the WHO smallpox vaccine) and accomplished without repercussions to the Soviet Union.

Western Medicine is Inadequate

But what of Western medicine? Doesn't it always come through to save us?

Present Western medical science is totally inadequate to do the task required.

For example, present medicine is largely based on "slowing or killing the invader of the body." Whether that invader is a germ, bacterium, virus, or whatever.

If there's a mechanical difficulty, then present treatment is to try to mechanically remedy it — remove a bad appendix, set a broken leg, etc.

However, then it's strictly up to the body's own natural healing system - its immune system and its repair system — to actually do the healing and repair/reconstruction.

The problem with AIDS is that it directly affects and suppresses the body's systems for healing and repair.

Once this protection and repair system is in disarray, any disease organism can attack the body and infect it. If the immune system can't handle the problem, eventually the invading disease organism kills its host.

Even if a way is found to destroy the AIDS virus itself, present medicine has not the foggiest notion of how to restore the healing and repair system from the damage and destruction it will have already sustained by massive AIDS infection.

Compare the situation to division combat: Hostile invaders are met at the front of the division by the division's combat forces. A battle ensues, with penetration and damage to the rear area support and sustenance facilities. After the enemy is beaten back or destroyed, the rear support and the combat forces must be renewed and restored. If that cannot be done, the seriously weakened division can still literally be boiled over by almost anything that comes along and hits it.*

---

*For this reason, standard Soviet doctrine is to attack in echelon, say, by three divisions — one behind the other. Thus, even if the first division is annihilated in the attack, the second and third divisions meet a seriously weakened enemy division and blast right over it. Similar tactics in the BW crisis - multiple viral strikes, one after the other — can almost inevitably be expected, now that the "first division" - AIDS - has engaged us.

The Soviet "repetitive attack" method *works,* in case anyone has doubts. For example, see George C. Wilson, "Army training exercises raise troubling questions," The Washington Post, as printed in the Huntsville (AL) Times, Feb. 28, 1988, p. 21A. Repeated U.S. Army tests in the Mojave Desert have shown that, in nonnuclear combat, when attacked by multiple Soviet echelons with 3:1 strength, U.S. forces lose every time! In particular, U.S. troops have poor anti tank defenses.

Modern medical dogma has concentrated almost entirely on the combat forces' problem of destroying and/or repelling the invader.

Western medicine has almost no tools or means of restoring the support and rear maintenance functions.

Even in a paroxysmal effort, dogmatic Western medical science would take decades to totally reorient its medical thinking to where it could even attack the real problem of AIDS and catastrophic sickness due to other viral BW warfare agents.

The same inability to cope with the problem also applies to our government bureaucracy, government agencies, the private companies, the universities, and to Western science in general.

If we have to depend on any or all of our present "system," the war is already over and we are all already doomed. Half of humanity will die within possibly five years after the upcoming two- to three-year "point of no return."

Medical insurance, treatment centers, and even private physicians are quickly going to be bankrupted as the AIDS plague increases. With the collapse of Western medicine, panic and pandemonium will reign in the streets. Martial law will be a necessity, and rioting on a scale unparalleled in American history will occur. We are likely to see our own U.S. dictatorship, in the form of a military junta making a coup and seizing power "to restore order and keep the lid on."

Study carefully the performance of governments and military regimes in such countries as Korea and South Vietnam, when those countries were hard pressed by enemy forces and in great disarray.

You may be looking at your own future.

The Soviets Will Not be Decimated

With the devastation of our armed forces, our governmental and civil agencies, and our population at large, a simple show of strength by the Soviets will collapse any further resistance. The hammer and sickle will wave supreme over the residue.

That's because the Russians *do* have the necessary developed scalar EM medical technology, not only to control and destroy the AIDS viruses, but also to directly restore the immune and repair systems.

The Kaznacheyev experiments alone show they have extensively studied electromagnetic induction of cellular disease and death.* We will have much more to say on that later.

If the Soviets extensively studied cellular death transmission by electromagnetics, you can bet they also studied the reverse: electromagnetic healing, or time-reversing the electromagnetic "death" signals themselves to produce specific "healing" signals for any type of cellular disease.

This is particularly certain since the Soviets also discovered and possessed the necessary "time reversal" electromagnetics: phase conjugation. This had been widely used by the Soviets in developing their enormous scalar EM weaponry, beginning about 1950 or so.

The curative signal technology was kept secret, of course, since the Soviets realized that it provided the only viable defense against viral biological warfare weapons.

The Soviets most certainly have the necessary AIDS cure. They can readily make the correct electromagnetic healing signal, by phase conjugating the AIDS death signal from dying human cells that contain AIDS retroviruses.

They also know we most certainly do *not* have the necessary EM cure. Years of weak microwave radiation of personnel in the U.S. Embassy in Moscow, and inducing illnesses in those personnel and even U.S. ambassadors, have shown that our own intelligence and scientific communities don't know what's really going on there.

Specificially, we've demonstrated (after checking with the DIA, CIA, NSA, and the orthodox scientists) that we do not recognize electromagnetic *biological warfare* —the deliberate introduction of cellular disease and death by electromagnetic means. Not only do we not recognize it, we can't do anything about it when it hits us.

And make no mistake, time-reversed electromagnetic medicine is shockingly cheap and effective, once developed. Specific healing signals for specific diseases — viral or whatever — can be quickly developed (within hours) and immediately used for mass treatment of entire populations.

For example, given the developed healing or "antidote" signal, it can even be recorded and played over ordinary transmitters, using their normal transmissions as carrier waves. A populace continually bombarded by such "healing" television, radio, and powerline signals simply doesn't get the disease. And those already infected recover when continually exposed to such healing signals.

---

* There is now reliable information more closely associating Kaznacheyev with military *institutes near Moscow, engaged in* microwave weapon (directed energy) research

So the Russians can stop the AIDS epidemic cold, anytime they choose to do so.

Electromagnetically they can also gently guide and stimulate the progress of a disease such as AIDS, once the disease is introduced into a targeted population.*

Or, the Soviets can pull out all the stops and induce the disease in a distant population directly, placing the "death signals" on common carriers. Such as on our power lines, radio stations, televisions, and directly by induction from the Woodpecker over-the-horizon radar signals which continually bombard us.

The West cannot do any of these things. Neither can the West counter them, or defend against them effectively.

Unless a totally different tack is taken in the Western effort to counter this viral BW threat, Gorbachev's entrancing smile has already lured us to our doom. We are faced with a direct realization of Lenin's words:

"In the final revolution it does not matter if three-quarters of the earth is destroyed, so long as the remaining quarter emerges communist."

And Khrushchev's ghost will have the last laugh after all. We will have fulfilled the prediction he made to Nixon in their famous "Kitchen debate":

"We will bury you!"

Quite simply, for some decades the Soviets have had a new science — one they call *energetics,* and one which I have called *scalar electromagnetics.* The keystone or "heart" of this new science is time-reversed (phase conjugation) electromagnetics and electrogravitation.

---

*A candidate for an eery test of just this capability may be the highly anomalous deaths of numerous dolphins off the East coast of the U.S. in 1987. Nearly 500 died in the summer and late fall, apparently as a result of the weakening of their immune system to the common bacterium Vibrio. See Julia Lawlor, "Key: Failure of Mammals' **defenses,"** USA Today, **Jan. 19, 1988, p. 1A-2A.**

Chapters 4 and 5
***deal** with complex scientific material*
and are suggested for those who are
technically proficient.

CHAPTER 4

EXTRAORDINARY PHYSICS

Maxwell's Lost Unified Field Theory

About the time of the U.S. Civil War, James Clerk Maxwell succeded in unifying magnetism and electricity. Actually he did far more than that, in his theory as originally written.

In fact, he had produced a theory which also captured the free interchange between electromagnetic energy and gravitational energy, but no one — including Maxwell himself—realized it at the time.

Maxwell wrote his original theory in quaternion and quaternion-like mathematics. The modern form of vector mathematics had not yet been finalized by Gibbs and Heaviside. It is most instructional to examine some of the fundamental differences between a vector and a quaternion.

In a conventional 3-dimensional vector, one may have three vector components, such as

$$v = ai + bj + ck \qquad (4\text{-}1)$$

where i, j, k are unit vectors in the directions of the x, y, and z axes respectively and a, b, and c are constants.

Obviously if the vector components of vector v are zero, then

$$v = 0 \qquad (4\text{-}2)$$

We shall be interested in the "vector product" of two identical vectors v, where

$$\mathbf{v} \times \mathbf{v} = AA \sin\emptyset = \mathbf{O} \qquad \textbf{(4-3)}$$

and A is the length (magnitude) of vector v, $\emptyset$ is the angle between the two vectors (in this case zero), and O is the zero vector.

Now let us look for a moment at the quaternion situation.

First, in addition to the three vector components, a quaternion also has a separate scalar component, w. So the quaternion q for this situation is

$$q = w + ai + bj + ck \qquad (4\text{-}4)$$

Now when this quaternion is multiplied times itself, the vector part zeros, just as it did for the vector expression. However, the scalar part does not go to zero. Instead, we have

$$q \, X \, q = A^2 = a^2 + b^2 + c^2 \qquad (4\text{-}5)$$

There is a very good physical interpretation of this result. It is a square of the amplitude, hence for the vector part of a wave, it is directly proportional to the energy density of the vacuum, as a function of time, at the particular position. However, we now need to make a short explanation of variation of stress energy density of spacetime.

First, we note that, according to general relativity, the "gravitational potential" is just a conglomerate of potentials of all kinds. Basically, a potential represents a G-potential, and consequently a curvature of spacetime. The potential also represents "trapped energy."

Second, we note that Kaluza combined electromagnetics and gravitation as a unified theory in 1921. Kaluza added a fifth (spatial) dimension to Minkowski's 4-space, and applied Einstein's relativity theory to 5 dimensions.

To Kaluza's delight, a common 5-d potential is responsible for both electromagnetic field and gravitational field. The "bleed-off of this 5-potential in the 5th dimension (which is wrapped **around** each point in our 3-space) is what we know as the **electromagnetic force field.** The bleed-off of this 5-potential *in and through* our 3-space is what we know as the **gravitational force field.**

Since the EM field is very much stronger (by a factor of 10^{42} for electrons) than the gravitational field, it is obvious that most of the bleed-off of the 5-potential is in the 5th dimension, as EM force field. Only a tiny bit is left to bleed-off in 3-space, producing a very weak gravitational field.[1]

We state this fact: as a mass moves in space, it generates increased "activity" with the virtual particle flux of vacuum itself. **The** Increased virtual particle flux activity exchange between vacuum and mass is analogous to a strange kind of "virtual resistance." Since the resistance is virtual, it does not observably slow down an observable object moving in an (unobservable, virtual-particle flux) vacuum.

The increased flux activity represents an increased "virtual energy density" of space time, and an increased "trapped potential" (mass; resistance to an accelerating force) of the moving object. It represents a rotation of the spacetime frame, vis a vis the laboratory observer).

In the virtual vacuum (which contains both positive and negative time), one sees two antiparallel virtual forces: one in positive time, along the velocity vector of the object, and one in negative time (time reversed, or phase conjugated). The reason one sees virtual forces is that, each virtual (subquantal) change in the virtual flux activity represents an individual (unintegrated), separate change, hence a virtual acceleration. The observer (where things are integrated), sees the integral of all these accelerations, hence observable velocity.

The vector sum of these two virtual forces in the vacuum is a zero vector; however, the two taken together represent a stress in the local energy density of vacuum.

Since we may regard an EM wave as a stream of virtual electrons/positrons, each engaging in tremendous virtual particle flux exchange with the vacuum, then the same basic picture applies.

Now for our physical interpretation of (4-5): If we refer to an EM wave moving in the vacuum, the rotation of the frame is maximum (90 degrees). But this same rotation is just the same as additional vacuum stress, so the vacuum stress is maximum.

---

[1] **Electromagnetics is 5-gravity sliding around our 3-space. 3-gravity is 5-gravity oozing** through our 3-space.

This leads to these conclusions: An electrical force field vector represents a local maximum linear stress in spacetime, along the line of the vector. (Note we specifically deny that the electrical force field vector, of an EM wave in vacuum, is transverse. Instead, it is longitudinal. That has been addressed elsewhere by the author and will not be covered further here.)

Another electrical (stress) vector interacting with the first one adds more "urging" stress to the first. However, this action is occurring in the rotated frame of the moving wave, and so is rotated 90 degrees from the electrical velocity vector. Therefore it is lateral (but in a hyperdimension, not in 3-space) at right angles to the electrical velocity vector.

The combined "urging" action of the two vectors thus sweeps out an area with respect to the laboratory observer.

This means that the total "urging" or "stressing" action of the two vectors is analogous to a vector area.

It also means that this "area" function may be taken as the "swirl" of the electrical vector, but in a hyperdimension, not in 3-space. That is, we have described the magnetic force field.

Thus any two electrical vectors that interact will have an "area" or "resistance" component generated. *Any two that interact.* Whether they add vectorially, cross-product multiply, or dot-product multiply.

What is actually happening is that the wave exists in the 5-potential. The E and B fields just represent the oscillations in that 5-potential. They represent oscillations in the bleed-offs of that potential as E-field (longitudinal) and B-field (swirl).

The drag-area represents the accumulation of extra potential — hence the local rotation of spacetime. Since this accumulation is moving (along with the EM wave), as it passes a point it represents a change in the local virtual particle flux density of vacuum at that point, **hence a local curvature of spacetime.**

Hence, the EM wave makes a 5-dimension G-potential wave as it travels. The 3-dimensional gravity wave associated with this is normally very, very much smaller in magnitude — say, by a factor of 10^{-42} or so.

However, if the two vectors interact so as to produce a vector zero resultant, then all the electromagnetic energy of tho two vectors **is** captured. That is, all the "EM vector zero" resultant means is that the EM bleedoff of the 5-space gravitational potential wave has been stopped. The 5-potential is still oscillating, and now all its trapped 5-energy must bleed off as 3-gravity force field.

Mass acts as an accumulator for this "trapped-EM energy turned into local curvature of 5-space." If we continually irradiate a mass with such a wave, the atomic nuclei of the mass slowly charge up with the new energy. Note that this potential delta may be positive or negative, if one adjusts accordingly.

In this fashion one may change the mass of a static object in the laboratory. One may either increase the mass or decrease it, or cause it to float, or even cause it to accelerate upwards.

But to return to our vector interaction and our interpretation of the scalar remainder of the quaternion.

The rule is, when the two EM vectors interact so as to form a zero EM resultant, then the EM energy represented in each of the two vectors has been converted into a special form of 5-space gravitational potential, one that is not bleeding-off in the fifth dimension (electromagnetically), but one which will gradually produce a 3-gravity potential in a mass's atomic nuclei as a function of time, the individual element, permeability and absorption factors of those nuclei, etc.

Therefore in our mathematical theory we ought to have a scalar component remaining when two EM vectors interact to form an EM vector zero resultant. That scalar component represents what is happening in the 5-potential, that will only bleed into 3-gravity.

With exploration of this phenomenology in the laboratory, one can work out the functions, constants, coefficients, and parameters which specify how the "5-G to 3-G and vice versa" component works in conjunction with mass, motion, and other fields.

That's the magic secret of electrogravitation.

It was captured inherently by the quaternion theory of Maxwell published during the American Civil War!

After Maxwell's death, when the scalar portion of the quaternion was discarded (by Oliver Heaviside) to form "modern" EM theory, that also discarded the unified field interaction between electromagnetics and gravitation.

Electromagnetic field and gravitational field were then modeled and regarded as mutually exclusive. EM field, therefore, was thought to produce no specific gravitational effects in the vacuum itself.

Hence when Albert Einstein was formulating general relativity some decades later, he knew only one way to "curve" spacetime: that was gravitationally, by "attraction of mass" forces.

But gravitational force was so weak that only a huge collection of mass would exert enough of it to measurably curve spacetime. That would require a sun or star. Since the observer and his instruments would never be on the surface of the sun or a star, **Einstein** assumed **that the local** spacetime **of the observer would not be curved.**

Hence he severely crippled his general relativity theory. In the West, it remains an assumption to this day. It is not a universal assumption in the Soviet Union, however, since the Soviets have long since written — and developed in the laboratory—unrestricted general relativity with *local* spacetime curvature, and hence *local violation of conservation laws.*

So the scalar part of the quaternion interaction, that remains when the vector part of the resultant is zero, is magic indeed.

That is the magic unified field portion that everyone has been seeking for decades and decades!

It was there at the beginning. Then we inexplicably threw it away!

But to return to our vector/quaternion examples.

Note also that the two vectors

$$V_1 = ai + bj + ck, \quad v_2 = -ai - bj - ck \qquad (4\text{-}6)$$

sum to zero vectorially when added, such that

$$v_1 + v_2 = 0 \qquad (4\text{-}7)$$

However, quaternions may behave quite differently, even under addition. For example, the two quaternions

$$q_1 = w + ai + bj + ck,$$
$$q_2 = w - ai - bj - ck \qquad (4\text{-}8)$$

sum their vector parts to a vector zero resultant, but do not sum to a scalar zero as well. Instead, they sum to

$$q_1 + q_2 = 2w \qquad (4\text{-}9)$$

AS can be seen, quaternions which have the same vector parts as vectors, do not necessarily yield a complete zero when the vector parts sum to zero. And when two vectors multiply to provide a zero vector resultant, corresponding quaternions may yield a scalar term that is equal to the product of the magnitudes of the two vectors. In this way, the quaternion approach can capture the *stress of the medium,* induced by opposing or multiplying vectors. In the vector approach, the stress of the medium is entirely lost when the two vectors sum or multiply to a zero resultant.

Let us see just how important this "vacuum stress" can be.

First, the "stress in the medium" represents curvature of spacetime when that medium is the vacuum/spacetime.

In other words, the quaternion approach captures the ability to utilize electromagnetics and produce local curvature of spacetime, in an engineering fashion. Heaviside wrote a subset of Maxwell's theory where this capability is excluded.*

---

*Dr. Henry Monteith has independently discovered that Maxwell's original quaternion theory was a unified field theory. See his important "Dynamic Gravity and Electromagnetic Processes," in publication

Note that, by Maxwell's original quaternion theory, however, Einstein's assumption need not be true at all. For example, look at equations (4-5) and (4-9): Here we may utilize electromagnetic force quaternions to produce zeroed EM forces, and an increased stress in local spacetime. In other words, we have curved local spacetime electromagnetically. Since (with electrons) electromagnetic forces are about 10^{42} times as strong as the gravitation force, this local curvature of spacetime is not negligible.

That is, we have produced a scalar effect from zeroing vector operation between electromagnetic forces. I have called this **scalar electromagnetics,** and pointed out that it is truly *electrogravitation.*

We stress again that this violates one of the severely limiting assumptions that Einstein placed upon his theory of general relativity. He assumed that curving spacetime could only be done by the weak gravitational force due to mass. Since gravitational force is so weak, only a stupendous collection of mass — such as the sun or a star — could curve spacetime enough to notice experimentally.

Since obviously the observer and his laboratory instruments would never be located on the surface of the sun or a star, *Einstein assumed that the local spacetime would never be curved!* In other words, the local frame would always be a Lorentz frame. This meant that, locally, the familiar conservation laws of physics would always apply. Curvature of spacetime would only occur at great distances, and at huge collections of mass such as a star or dwarf star.

Einstein did not write a complete, unlimited general relativity. He wrote a sort of "special relativity with distant perturbations."

If Einstein had had electromagnetic theory in quaternions, the scalar "vacuum pressure" parts would have been there for him to ponder. It is highly probable that he would have captured the "electromagnetics-to-gravity conversion remainder" in the quaternion interactions.

If so, he would have written the full theory of general relativity, involving local violation of conservation of energy, a unified field theory, and the direct engineering of

gravitational and antigravity effects on the laboratory bench by electromagnetic means.

In that case, we should long since have navigated all around the solar system, colonized the planets, produced practical free energy devices and power systems, and avoided two great world wars and a host of little ones.

But let us now see if we can make a gravitational wave, electromagnetically.

Again, regard equations (4-5) and (4-9). Suppose these are instantaneous operations of EM force quaternions whose vector parts are varying in magnitude, but in such a manner that the vector parts ways form a zero vector resultant. Now one can see that the scalar part remaining — which represents the stress of local spacetime - is varying as the product of the magnitudes of the vectors in the interaction vary.

This means that one has now produced a scalar wave that represents the local variation of spacetime curvature in an oscillating manner.

Rigorously this is a gravitational wave. It has been produced locally. It has been produced by Maxwell's original unified theory.

Again, I have called this area *scalar electromagnetics*. The Soviets call it *energetics*.

Where local spacetime curvature is varied, conservation laws (energy, conversation, etc.) need not hold. Curved one way, the local spacetime acts as a source (of energy, charge, etc.) Curved the other way, the local spacetime acts as a sink (of energy, charge, etc.)

The Soviets often do not utilize the same restricted kind of general relativity that Western scientists adhere to.

Soviet papers in general relativity regularly point out the complete and unrestricted theory, where local spacetime curvature is allowed. They also point out that all conservation laws may be violated by such local curvature. Thus the Soviets have no unduly dogmatic respect for conservation laws.

Further, by assuming the possibility of local spacetime curvature, Soviet scientists have assumed the possibility of direct experimentation with general relativity on the laboratory bench.

In the West, we have assumed that such cannot possibly be done, because of Einstein's limiting assumption of no local spacetime curvature. Thus Western physicists are strongly conditioned away from electrogravitation.

This is particularly ironic since the basis for just such an experimental theory was produced by none other than Maxwell himself in his original theory of electromagnetism.

Indeed, shortly after the U.S. Civil War, we should have been developing antigravity spaceships. We should have developed electromagnetics a la Maxwell and been on our way to the planets of our solar system. For Maxwell had — admittedly somewhat unwittingly — given us the basis for the necessary engineering theory of unified electrogravitation.

Heaviside's Mutilation of Maxwell's Theory

Well after Maxwell's death,Oliver Heaviside helped to finalize what is today vector analysis.

Then he undertook to "translate" Maxwell's theory from quaternion form to the new vector mathematics form.

Now quaternions were devilishly difficult to calculate in. So much so, that a majority of the electrical scientists (there were not very many of them in those days!) were in despair.

Not to worry! Heaviside took a broadax, figuratively speaking, and simply chopped off the scalar term, leaving only the vector components.

With that artifice, he greatly simplified the calculations to be performed.

Of course, he also threw away the EM stress of spacetime! That is, he threw away the "gravitation" part of Maxwell's theory!

Let me stress this fact most strongly. **After Maxwell's death a single** man**—Oliver Heaviside—directly altered Maxwell's equations, eliminating localized electrogravitation and producing the form of the theory taught throughout the West today as "Maxwell's theory."**

Maxwell's theory has never been taught in Western universities! Only Heaviside's crippled subset of the theory has been taught!

Then, shortly before the turn of the century, a short, sharp "debate" erupted in a few journals — mostly in the journal Nature. Only about 30 scientists took part in the "debate."

It wasn't really much of a debate! The vectorists simply steamrolled right over the remaining quaternionists, sweeping all opposition before them.

They simply threw out the remaining vestiges of Maxwell's quaternion theory, and completely adopted Heaviside's interpretation

Thus, a little over a decade later when Einstein wrote his general relativity theory, he did not know that the original work of Maxwell already indicated the unification of gravitation and electromagnetics, and indicated the ease with which local spacetime could be electrogravitationally curved locally and *engineered.*

Accordingly, he placed the scientists of the West on a road which rigorously assumed that a unified field theory was yet to be discovered. It also strongly discouraged any experimentation aimed at curving local spacetime, for it assumed that such could not be done.

After Potsdam and World War II, a frustrated Stalin was to drive his scientists to review the entire scientific literature of the Western world, actively seeking a great new technical breakthrough area such as the Allies had demonstrated with the development and use of the atomic bomb.

Great Soviet institutes — one staffed, for example, with over 2,000 PhD's — were set up to thoroughly review all the Western scientific literature from its very beginning. Anything interesting, anomalous, or unknown was put aside for further examination.

It is a good bet that the meticulous Soviet scientists discovered the difference between Maxwell's original electromagnetic theory and Heaviside's mutilation of it. Great mathematicians that they are, Soviet scientists would have realized the implications of the difference. With their knowledge of unlimited general relativity, they would have made the connection to electrogravitation.

By 1950 they had indeed done so, and were deeply into the development of what they called "energetics", and I have called *scalar electromagnetics.*

They had also reached another milestone about the same time — 1950 or so.

After WWII, both the Soviets and the U.S. were keen on securing the best of the German scientists. The U.S. particularly wanted missile scientists and rocket engineers. The Soviets wanted them too; but they also wanted the German radar specialists and infrared specialists.

The West didn't care about the German radar scientists and engineers, and the IR fellows. The Soviets did, and they got them. That was to prove a most spectacular benefit indeed.

During the war, the Germans had placed extreme emphasis upon radar and radar absorbing materials (RAM). The German scientists had fantastically developed and extended the science of radar cross section — which is the heart of the matter and very, very complex. **They were much further ahead in radar cross section theory at the end of WWII than where the U.S. is today, in the opinion of** some **U.S. radar experts.**

So the Soviets started with a great jump on us in radar knowledge, and they have steadily increased the lead over the years.

In addition, the Germans had developed highly successful radar absorbing materials, and much of the theory to accompany them.

Such materials turn out to be the key to how to build and develop a radar phase conjugation mirror, to produce a time-reversed radar wave.

Thus, because of the German scientists, by 1950 or so the Soviets had already discovered phase conjugation. And they had discovered it in radar first, not in optics!

They would have been primed for the discovery by their great review of Western literature and the foundations of science, since they would probably have noticed that the time-reversed wave is a solution to the wave equation. If so, they would certainly have realized its generality throughout all physics, all frequency bands, and all types of waves.

Superb mathematicians that they are, the Soviets would certainly have made the Kaluza-Klein theory connection, and also realized that phase conjugate waves carry negative energy as well as negative time. They would quickly have seen the gravity and antigravity implications.

So about 1950 or so, the Soviet Union would have started phenomenology experimentation in earnest, with phase conjugate radar mirrors and phase conjugate radars. This is what was referred to as *energetics.* The Soviets began a massive program in energetics about the time of the beginning of the Korean War.

By 1957-8 the Soviets had progressed to the point of a giant scalar EM accident in the Urals which exploded nearby atomic wastes, devastating the area. They had also progressed to development of great new superweapons using their new energetics — weapons to which Khrushchev referred in 1960 when he informed the Soviet Presidium of a new, fantastic weapon in development, a weapon "so powerful that it could wipe out all life on earth if unrestrainedly employed."

About the same time (mid-to-late 50's), the Soviets had also started the eery low-level microwave radiation of the U.S. Embassy in MOSCOW, to see if the U.S. knew of scalar electromagnetics (energetics) and was developing its own electrogravitational weapons and defenses.

Building Upon Whittaker's Fundamental Work

In 1904, a most fundamental paper in the foundations of electromagnetics was delivered by the British mathematician E. T. Whittaker. (E. T. Whittaker, "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372.).

In this important paper, Whittaker showed that the electromagnetic force field equations can be replaced with the derivates of two scalar potential functions.

He also derived the most general form of electromagnetic disturbances in the ether.

This means that the coupling of two dynamic scalar functions can replace vector electromagnetics in the vacuum.

Note that Whittaker's work pointing out the overriding importance of scalar fields also accents the erroneously discarded scalar part of Maxwell's quaternion electromagnetic theory even more strongly.

Let me explain now how I got from Whittaker's paper to scalar electromagnetics, Soviet Tesla weapons, free energy, antigravity, and electromagnetic healing.

When I discovered Whittaker's paper, I had already strongly objected that "charges" and electromagnetic vector force fields — as presently included in the Heaviside version of Maxwell's equations — included observable mass. Of course there was no observable mass in the vacuum, hence the prescribed kind of EM force fields could not exist as such in the vacuum.

Obviously the foundations of our ordinary electromagnetics theory were seriously flawed. Although my objections fell on deaf ears, I determined to examine the foundations of EM theory, discover the flaws, and at least point out the necessary corrections to be made.

Though this was an arduous task to undertake and it required many years, slowly the flaws showed themselves, and the necessary corrections slowly became clearer.

Most exciting of all, in working with several unorthodox researchers, I was able to see many of these new ideas tried, adjusted, and demonstrated. In addition, the proprietary discoveries of these colleagues continued to reveal new and unique principles and concepts. The only disadvantage was that I could not reveal the propriety apparatuses and demonstrations of my inventor associates, but only the principles and concepts that developed. In turn, I also developed principles and concepts to explain what they were doing and the results they were obtaining.

So over the years I have slowly been releasing the principles and concepts. Some of them are my own discoveries, many of them are the discoveries of my associates. Some of them are simply a mixture of both.

Early on, it became obvious that the Soviet Union was far ahead on this path, and was already utilizing the new unified field theory to build eery, powerful new superweapons.

Since no one else in the U.S. seemed to be "watching thin particular store" (I was rather universally regarded as some peculiar sort of fool!), I also began to compile information and data on the Soviet weaponization of this unrecognized technology. This information I have released in a series of papers, briefings, and books, the most recent being a 1-hour videotape, "Soviet Weather Engineering Over North America," 1985, and a detailed book, Fer-de-Lance: A Briefing on Soviet Scalar Electromagnetic Weapons, Tesla Book Co., Greenville, Texas, 1986.

Building upon Whittaker's important work, I formulated a conceptual revision to electromagnetics, which I dubbed *scalar* electromagnetics to accent that the observable EM vector force fields did not exist as such in vacuum, but dynamic scalar fields did. I also wished to call strong attention to the fact that observable force does not exist until an observable particle of mass is coupled to the interference of the two scalar fields (much like in the Aharonov-Bohm effect). The Soviets, of course, call this area *energetics,* Energetics technology has been used in gigantic weapons programs of the Soviet Union for decades, and it appears to be developed under the most highly classified program that the Soviet possess. All development and deployment of energetics weapons is under the KGB and controlled directly by that organization, not by the Soviet Armed Forces.

Peter Kapitsa, the great Soviet physicist, was once pressed by Nikita Khrushchev for a total defense against missiles and air- and space-borne vehicles. Kapitsa replied that it could only come from the new energetics. In 1960, of course, Khrushchev gleefully announced to the Presidium that a new, fantastic Soviet weapon was in the development, "so powerful that, if unrestrainedly used, it could wipe out all life on earth."

Ironically, Khrushchev "jumped the gun" before his new superweapons were deployed. In the fall of 1962 he began inserting long range missiles into Cuba, bracketing the U.S. with nuclear firepower in an attempt to immediately change the balance of power. Kennedy, of course, backed him down "eyeball to eyeball," so to speak, in a blunt confrontation, but promised not to invade Cuba.

Khrushchev, with his days numbered, was desperate to deploy his new superweapons and provide a dramatic demonstration to recover face.

By destroying the U.S.S. Thresher on April 10,1963 and, on the next day, producing a gigantic underwater explosion 100 miles north of Puerto Rico, the Soviets demonstrated that the new superweapons had been deployed. Khrushchev managed to retain his position a while longer.

In the 1960's and early 1970's, I was also deeply involved in the study of paranormal phenomena.

In 1969, I entered the Georgia Institute of Technology to pursue a Master's Degree program in nuclear engineering, graduating in 1971.

In 1973, I published a rather simple paper, "Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena," in which I pointed out the nature of quantum change, gave a new definition of mass and acceleration, and pointed out the fundamental nature of inversion of time. The paper also contained a simplified derivation of Newton's laws of motion, relativistic form. The elements of this paper had been worked out in 1971 while I was finishing my Master's program in nuclear engineering. Finishing the work had been interrupted by a slight sidetrack — a tour in Vietnam from summer of 1971 until summer 1972.

At about the same time, I formulated a fundamental correction to Aristotle's logic, adding a fourth law of logic to Aristotle's three, and a proof of it. The new logic was of great use in discovering and uncovering new concepts in unified field theory.

Incorporating Kaluza-Klein 5-dimensional concepts, scalar EM became a field theory that unifies electromagnetics and gravitation.

Incorporating dynamic sum-zeroed EM vector systems (which are discarded in normal EM theory) allowed the direct engineering of the unified field theory, including structuring the vacuum, curving local spacetime, and producing effects at a distance and in higher dimensions. Actually it allowed the recovery of much of the scalar part of Maxwell's original theory.

I then realized that, inside a vector zero EM force field summation/multiplication, the virtual particle flux of vacuum/spacetime was ordered and controlled locally and macroscopically. This of course violated one of the major assumptions (a postulate) of quantum mechanics; the assumption that the structure of vacuum was randomized, and could not be deliberately ordered, engineered, and curved locally.

Adding phase conjugation (time reversal) aspects and extended quantum mechanical concepts allowed local antigravity and local curvature of spacetime to be included — again, on an engineering basis. It also allowed one to produce a mechanism responsible for Newton's third law, and to engineer the reaction force at will.

Further, it revealed that the law of entropy was simply the *positive time* statement; it showed that there was another half of the law, the *negative time* part or the law of negentropy.

In addition, a startling new concept of mind, thought, life, biofields, disease, and healing emerged from all this — again, on an engineering basis. As we stated in the beginning of this book, it is now an urgent necessity to release my work on the basis for electromagnetic disease and electromagnetic healing. We must produce a very quick, positive treatment and cure of AIDS and other coming lethal viruses before the world is decimated.

Accordingly, this work is being released in this book.

In this chapter we will next present some perhaps surprising material on phase conjugation, from the scalar EM viewpoint, after first briefly explaining symmetry and parity.

In following subsections, we will cover briefly the remaining major concepts in scalar electromagnetics. This will then set the stage for the following chapter, Extraordinary Biology, in which we will deal with the basis for unparalleled electromagnetic healing.

Symmetry and Parity

The basic idea of symmetry is the arrangement of the parts of a body or system about an axis so that two or more parts appear the same with respect to some operation.

The most obvious example is to look in a mirror, where we notice that our image has been reversed, left to right. Yet otherwise there is no difference; and so we may say that the reflection has "mirror symmetry." It's the same except that left and right are reversed.

If you know the details of a system at one point, and at another point you know that the system will possess mirror symmetry, then you can predict exactly what state the system will be in at the latter point. It will be "left to right reversed." You don't have to calculate the laws of reflection and the laws of motion of the system.

All the laws of nature that possess "mirror symmetry" have a special property: If the words "right" and "left" are interchanged in the statement of the law, then the behavior of a system that obeys the law is unchanged.

Thus symmetry became a powerful idea in physics. For a system with a great number of possible interactions, etc., one can usually eliminate a very great number of them due to symmetry considerations. In other words, if one can assume that symmetry of the system will not be violated, then all the possible interactions that would yield "broken symmetry" can be eliminated.

On the other hand, if symmetry is broken in a physical interaction, at least one of the conservation laws is broken in that interaction. Broken symmetry — and hence violation of individual conservation laws — is a well-established fact in particle physics today.

Another concept intimately associated with mirror symmetry is parity. Each particle is assigned a number (+1 or -1), depending upon what kind of particle it is. One adds up these numbers for an assembly of particles, to obtain an overall "parity" number for the system. If parity is conserved in an interaction, then this total number does not change in the interaction.

Table 13. SYMMETRY

- **SOMETHING IS SYMMETRIC IF IT REMAINS UNCHANGED UNDER A CERTAIN OPERATION**
- **A SPHERE IS SYMMETRIC WITH RESPECT TO ROTATION ABOUT ITS CENTER**
- **A CATHEDRAL ARCH IS LEFT AND RIGHT SYMMETRIC ABOUT A VERTICAL LINE THROUGH THE CENTER**
- **LAWS OF ELECTRICITY ARE SYMMETRIC WITH RESPECT TO REVERSAL OF POSITIVE AND NEGATIVE CHARGE**

Table 14. A FORCE FIELD IS A SPECIAL LOCAL SYMMETRY

"..THE FORCE OF GRAVITY IS SIMPLY A MANIFESTATION OF AN ABSTRACT SYMMETRY — A LOCAL GAUGE SYMMETRY — THAT UNDERLIES THE PHYSICS OF THE WORLD."

"ALL FOUR FORCES OF NATURE CAN BE GENERATED IN THIS WAY."

PAUL DAVIES, SUPERFORCE.
1984, p. 115

Table 15. GRAVITY AND LOCAL SYMMETRY

- **THE** LAWS OF PHYSICS CAN BE MADE SYMMETRIC EVEN UNDER LOCAL GAUGE TRANSFORMATIONS

- A GRAVITATIONAL FIELD IS INTRODUCED TO COMPENSATE FOR PLACE-TO-PLACE VARIATIONS

- THE GRAVITATIONAL FIELD IS NATURE'S WAY OF MAINTAINING A LOCAL GAUGE SYMMETRY

- IN THE ABSENCE OF GRAVITY, THERE IS ONLY GLOBAL SYMMETRY

Table 16. HINDSIGHT IS BETTER THAN FORESIGHT

- **FROM THE TWIN REQUIREMENTS OF**
  - **SIMPLEST LOCAL GAUGE SYMMETRY**
  - **LORENTZ-POINCARE SYMMETRY OF SPECIAL RELATIVITY**

- **ONE CAN CONSTRUCT OR INFER**
  - **MAXWELL'S EQUATIONS**
  - **ALL THE LAWS OF ELECTROMAGNETICS**
  - **THE EXISTENCE OF RADIO WAVES**
  - **THE POSSIBILITY OF MOTORS AND GENERATORS**
  - **ETC.**

- **IN REALITY THESE WERE FIRST DISCOVERED EXPERIMENTALLY**

Table 17. WHEN SYMMETRY IS VIOLATED [BROKEN]:

- **A "NON-OBSERVABLE" TURNS OUT TO BE AN OBSERVABLE**
- **AN INVARIANCE IS BROKEN**
- **A CONSERVATION LAW OR SELECTION RULE IS BROKEN**

COMMENTS:

  - **A VIRTUAL OBJECT BECOMES AN OBSERVABLE OBJECT**
  - **LOCAL SPACETIME BECOMES CURVED, AT LEAST TO SOME POTENTIAL INVOLVING THAT OBJECT**

Table 18. SOME PROVEN ASYMMETRIES

- **POSITIVE AND NEGATIVE SIGNS OF ELECTRIC CHARGE**
- **TIME REVERSAL**
- **RIGHT/LEFT HANDEDNESS**

Table 19. SPONTANEOUS SYMMETRY BREAKING

PRESENT VIEW

1. IF 4-MOMENTUM $K\mu = 0$,

   $\O_{(X)} = \O_{VAC}$

   WHERE $\O_{(X)}$ IS AS COMPLEX
   AS ANY SPIN-0 FIELD
2. **FOR** VOLUME Ω » RELEVANT MICRODIMENSION.

   $K\mu \neq 0$ BUT $K_M \sim 0$

   $\langle \O_{(X)} \rangle \neq \O_{VAC}$ [INSIDE *Q]*

   $\langle \O_{(X)} \rangle \neq \O_{VAC}$ [OUTSIDE Q]
3. INSIDE Ω, SYMMETRY PROPERTIES DIFFER
   FROM THOSE OUTSIDE Ω
4. T. P AND CP MAY BE SLIGHTLY ASYMMETRIC
5. CPT REMAINS INTACT

SCALAR EM VIEW

1. IF ARTIFICIAL POTENTIAL $\O_{ART} >> 0$, VOLUME Ω
   MAY BE SMALL
2. $\O_{ART}$ MAY BE ENGINEERED IN
   - SPATIAL SIZE
   - LOCATION
   - MAGNITUDE
   - SUBSTRUCTURE
3. SYMMETRY PROPERTIES INSIDE Ω MAY
   - DRASTICALLY DIFFER
   - BE DELIBERATELY ENGINEERED
   - BE PATTERNED
4. T, P, AND CP MAY BE APPRECIABLY ASYMMETRIC
5. CPT MAY BE ASYMMETRIC

Table 20. WHY SUCH DRASTIC DIFFERENCE?

IN PRESENT THEORY

$\varnothing_{(X)}$ • IS USED ONLY AS A PHENOMENOLOGICAL DESCRIPTION
- **MICROSCOPIC STRUCTURE OF $\varnothing_{(X)}$ OF NO CONCERN**
- **ONLY LONG-WAVELENGTH LIMIT OF FIELD IS OF INTEREST**
- **VACUUM ENGINEERING IS ONLY A REMOTE POSSIBLITY**
- **ACTION AT A DISTANCE IS NOT POSSIBLE**
- **LOCAL SPACETIME NOT CURVED**

IN SCALAR EM

$\varnothing_{ART}$ • REPRESENTS A LOCAL CURVATURE OF SPACETIME [LOCAL GENERAL RELATIVITY]
- **MICROSCOPIC STRUCTURE OF $\varnothing_{(X)}$ IS DETERMINISTIC AND COHERED INTO MACROSCOPIC STRUCTURE**
- **ALL WAVELENGTHS OF THE FIELD ARE OF INTEREST**

$\varnothing_{ART}$ WAVES ARE REAL, ENGINEERABLE AND ELECTROGRAVITATIONAL
- **VACUUM ENGINEERING IS A REALITY**
- **ACTION AT A DISTANCE IS POSSIBLE**

Table 21. ASPECTS OF STRONG LOCAL ASYMMETRY

- **PROPERTIES OF AN OBJECT MAY DIFFER APPRECIABLY FOR**
  - **DIFFERENT OBSERVERS**
  - **DIFFERENT DETECTING MEANS**
  - **ONE TIME TO ANOTHER**
  - **ONE POSITION TO ANOTHER**

- **CONVERSATION LAWS MAY BE APPRECIABLY VIOLATED**
  - **ENERGY**
  - **CHARGE**
  - **SPIN**
  - **MOMENTUM**
  - **ANGULAR MOMENTUM**

- **LOCAL SPACETIME IS CURVED**
- **LORENTZ INVARIANCE OF VACUUM IS VIOLATED**
- **MAY BE A LOCAL "SINK" OR "SOURCE "**
- **GRAVITATIONAL/INERTIAL EFFECTS FROM EM**
- **TRANSLATION BETWEEN VIRTUAL AND OBSERVABLE**
- **ELECTROGRAVITATIONAL SOLITONS**
- **ACTION AT A DISTANCE**
- **TRANSMUTATION EFFECTS MAY EXIST**
- **SCALAR/PSEUDOSCALAR FIELD TRANSLATION**

For years the idea of conservation of parity convinced physicists that it would never be possible to tell right from left in the universe, Then in 1956, Yang and Lee pointed out a class of reactions where parity need not be conserved. This effect — broken parity — was confirmed experimentally, and so certain parts and interactions of the universe do possess asymmetry.

The result of all this was that particle physicists established that

(1) symmetry can be, and is, sometimes broken, (2) individual conservation laws can be, and sometimes are, broken.

Particle physics, however, concerns itself primarily with microscopic interactions. Most of the delightful new things discovered in the microworld have not yet been applied — by *physicists,* that is — in the macroworld.

If you can successfully apply broken symmetry in a major system, for example, you should be able to violate the conservation of energy law and thus produce a "free energy" engine. Several inventors with whom I work have done precisely that. The techniques and technology are still primitive and they need much further development, but the principle has been clearly established and replicated.

Physics is correct in certifying broken symmetry, and you can do it in large systems as well as in microscopic particle interactions. Of course you also *curve local spacetime* as well, something which Einstein and his followers have assumed cannot be done.

Charge-Parity-Time ***and Negative Energy***

The upshot of all this is that the physicists came up with the idea that what is really conserved is a consolidated thing called Charge-Parity-Time. You don't have to conserve any one or two of these three, but only the combined product.

Here they overlooked one key factor that is very important to our thesis: In the photon interaction, if you reverse charge and time, you reverse the sign of the energy.

That is, a photon that is emitted from a positive charge carries negative time negative energy. It is different from a photon that is emitted from negative charge, for that one carries what we have conventionally established as positive energy.

Thus physics has not considered the exclusive use of negative-energy / negative-time photons, even when some of its interactions produced them. Instead, it has just lumped together photons and antiphotons, and considered them both to have positive energy.

Antimatter systems don't have positive energy, for example; they have negative energy. *Here we don't use the positive and negative signs to indicate the direction the energy is traveling, but the fundamental kind, of energy that it is—time forward or time reversed.*

For example, consider the local conservation of energy. As conditoned, one almost always unconsciously thinks of the conservation *of positive* energy, without so stating. Yet one is perfectly free to build a system into which he inputs 100 joules of positive energy, and from which he outputs 300 joules of positive energy and 200 joules of negative energy!

For the real conservation law, that's the equivalent of putting in 100 joules of positive energy and outputting 100 joules of positive energy.

Now negative energy will run motors, light lamps, etc. —often much better than positive energy. It can be transmitted, transformed, stepped up or down in "voltage," rectified, received, etc. It can be translated into positive energy or vice versa.

Circuits run "cool" when using negative energy. If you short out a circuit of negative energy so that a violent discharge occurs, it produces cooling instead of heating, greatly lessening the danger of fire and destruction. If you mix negative energy onto a line that is carrying the same amount of positive energy, however, the two negate each other and there is suddenly "no power" on the system.

Think of what you were taught in simple arithmetic. You can take a zero — the absence of any single specific number — and replace it with any set of multiple numbers whose algebraic sum is zero. The presence of this special set of numbers you choose does not violate the condition that the zero means that you could not use just a single number.

Although zero is the absence of just a *single* thing, it can be the presence *of multiple* things.

The same thing works with energy. In the microworld, everything is fantastically energetic, and wildly fluctuating. For example, in a copper wire carrying a current of one ampere, wild momentary fluctuations at extremely high frequencies are occurring.

According to classical electrodynamics, the instantaneous current in a single one of these micro-micro fluctuations may reach thousands of amperes, and the instantaneous voltage may reach 10^{40} volts or more. As can be seen, the instantaneous energy and power are enormous. Yet outside the wire, we see a very placid and peaceful situation.

TIME IS NOT OBSERVABLE

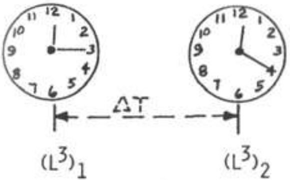

WHAT THEN IS TIME?

WHERE THEN DOES TIME EXIST?

Figure 33. The riddle of time. Unsolved by both physicists and philosophers.

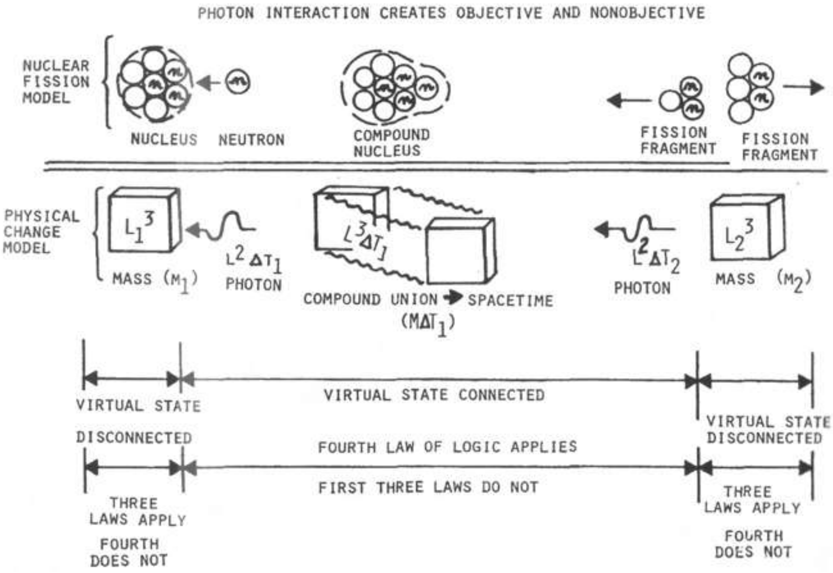
PHOTON INTERACTION CREATES OBJECTIVE AND NONOBJECTIVE
NUCLEAR FISSION MODEL
NUCLEUS
NEUTRON
COMPOUND NUCLEUS
FISSION FRAGMENT
FISSION FRAGMENT
PHYSICAL CHANGE MODEL
L₁³
MASS (M₁)
L²ΔT₁
PHOTON
L³ΔT₁
COMPOUND UNION ➔ SPACETIME
(MΔT₁)
L²ΔT₂
PHOTON
L₂³
MASS (M₂)
VIRTUAL STATE DISCONNECTED
VIRTUAL STATE CONNECTED
FOURTH LAW OF LOGIC APPLIES
VIRTUAL STATE DISCONNECTED
THREE LAWS APPLY
FOURTH DOES NOT
FIRST THREE LAWS DO NOT
THREE LAWS APPLY
FOURTH DOES NOT

Figure 34. Raindrop model of physical change.

Table 22. FLOW OF TIME

- **PHOTON INTERACTION CONSTITUTES TIME CHANGE**
  - $\rightsquigarrow \langle \text{NEGATIVE CHARGE} \rangle \rightarrow +\Delta T$
  - $\rightsquigarrow \langle \text{POSITIVE CHARGE} \rangle \rightarrow -\Delta T$
- **CHARGE-PARITY-TIME [CPT]**

  $\rightsquigarrow \langle e \rangle \rightarrow +\Delta T$ **(note: ~ = photon)**
  - **UBIQUITOUS**
  - **PREDOMINANT**
- **ANTIPHOTON IS A PHASE CONJUGATE REPLICA**

Table 23. PHYSICAL CHANGE

- **COMPOSED OF ACTION**
  - **h = BASIC QUANTUM**
- **TWO CANONICAL VARIABLES**
  - **ENERGY/TIME (dE dt)**
  - **MOMENTUM/LENGTH (p L)**
    ETC.
- **ENERGY/TIME**
  - $h = (+E)(+t)$
  - $h = (-E)(-t)$

Free Energy and Antigravity

Let's say the current flowing in the wire is alternating current at a frequency of one megahertz. Let's say we have one watt of power in the wire (there's one volt on it, and one ampere of flowing current). As a working rule-of thumb for engineering use, the available absolute micropower in an alternating current in a wire can be taken as:

$$P_a = 40 \times f \times P_c \qquad (4\text{-}10)$$

where P_a is the absolute power available, in watts; f is the frequency in Hertz; P_c is the conventional power available, in watts; and 40 is a constant of proportionality.

In our example, the absolute power actually available in the current is 40 million watts! We've got about 20 megawatts of positive power, and about 20 megawatts of negative power, in there — *in the microstructure of the electrical current itself--but disintegrated*

By applying unconventional scalar EM engineering techniques, we ought to be able to take such an "input" (a straight wire with one watt of power at a frequency of a megahertz) and get an "output" of up to 40 megawatts. It is relatively straightforward to get out 1 megawatt; such ratios have actually been achieved on the laboratory bench by my associates.

Of course we will have curved the local spacetime to do such a thing. But it is absolutely permitted to do that, by the laws of nature as best we know them from particle physics.

It is also possible to do that by Maxwell's original theory. And it has been proven in the laboratory.

Table 24. NEGATIVE ENERGY

- **CAN BE**
    - **GENERATED**
    - **TRANSMITTED**
    - **TRANSFORMED**
    - **RECTIFIED**
    - **RADIATED**
    - **TRANSLATED**
- **EASILY POWERS**
    - **AC MOTORS**
    - **DC MOTORS**
    - **LAMPS**
    - **TRANSFORMERS**
    - **RECTIFIERS**
- **CIRCUITS STAY COOL**

In scalar electromagnetics, we deliberately seek to produce such effects and systems by nonclassical application and extension of the modern conservation law CPT. And we include negative energy/ negative time production in those systems and effects.

Here's the extended magic rule: First we write down the expression **CPTEGS.**

In this expression, C stands for charge, P stands for (space) parity, T stands for time, E stands for energy, G stands for gravity, and S stands for entropy.

We are referring here to the algebraic sign of each term, and to a complete photon interaction (absorption and re-radiation) with a charged particle inside an atom of matter. Specifically, we are refferringto the result, after the reaction has occurred.

Normally the photon interacts with one of the electrons in the atom's outer shell. This interaction with the negative electron charge *produces* the "normal" physical world/physical reality we see around us. In the interaction, normal parityis produced, andnormal gravitation (attraction of mass) is produced. Also, in a series of such interactions, the normal entropy (movement from order to disorder) is txperienced and produced.

So in the expression CPTEGS normally the charge term is negative, and every other term is positive, for the complete photon interaction in an atom.

Now here is the magic engineering rule: In the complete photon interaction with a charged particle in an atom, if one of these terms is reversed in algebraic sign, all the others will be reversed also.

For example, if the photon interacts with the positively charged nucleus,then parity is reversed, the bit of time-advance is reversed, a negative/time-reversed photon is emitted, a tiny bit of antigravity is produced in the nucleus, and negentropy is produced. By negentropy, we mean that the emitted photon exhibits the strange characteristics of time reversal, and hence it will return from disorder to order.

This rule is actually a statement of phase conjugation and time reversed waves, as we shall see later.

For example, using the rule, the production of antigravity is straightforward. Simply produce a great excess of negative energy and negative time in an object or generator by scalar photon interaction with the nucleus, at a sufficient rate.

In negative time, gravitational force is reversed. Thus in negative time gravity is a repulsion, not an attraction. Production of excess negative time in the atomic nuclei of a system produces antigravity in that system. Period.*

***At 100 Hz about 500 negative watts/lb is required for flotation. At 1 GHz, the effect is so *weak it cannot be measured, and no amount of power will float* the mirror.**

Let me amplify that statement. There are two kinds of energy, negative and positive. Normally, everyone thinks only of positive energy.

Take Einstein's formula for the amount of energy that's "bottled up" and stored in mass:

$$E = mc^2 \qquad (4\text{-}11)$$

That's positive energy and positive mass. It automatically implies that we are looking at things where time is forward-going.

Now suppose we "charge up" that mass with negative energy: Then the mass is negative, like this:

$$-E = (-m)c^2 \qquad (4\text{-}12)$$

And negative mass is repelled by positive mass. (Like mass G-charges attract, unlike mass G-charges repel.)

That's all there is to it! Simply charge up the protons/neutrons of the atomic nuclei of a mass with excess phase conjugate (negative) energy/time, and it exhibits antigravity compared to the earth. Nothing could be simpler.

Again, it's been done in the laboratory by an associate.

So, with respect to a negative energy/negative time generator, you just add more load and draw more negative power from the machine, forcing it to produce more. It produces more negative time, and hence more antigravity and more "repulsion." When the repulsion balances the earth's attraction, the system "floats in air." Draw still more power by adding still more load, and it accelerates upward. Draw less power by decreasing the load, and it sinks downward again. That's all there is to it.

We should have been working upon this, shortly after the U.S. Civil War.

Again, there exist two completely different kinds of electromagnetic energy: positive and negative. These have been totally and hopelessly confused by modern physics, even though physicists sometimes use these terms.

But in dealing with CPTEGS in the photon interaction in an atom, we must clearly keep in mind what kind of energy we are dealing with.

The basic engineering rules are these: (1) If any one of the factors - C, P, T, E, G, or S — is reversed, all the others are reversed. (2) If positive charge is involved in the photon interaction, then time (carried by the emitted photon) is negative and so is the energy carried by it. (3) The simplest scalar EM wave may be considered as two EM waves locked together (modulating each other), where one component wave carries positive energy and time and the other component wave carries negative energy and time. (4) Separation of the two components by interaction with charged particles creates both positive and negative time and energy. (5) If the interacting particle is negatively charged, the positive energy/positive time component will interact with it. (6) If the interacting particle is positively charged, the negative energy/negative time component will interact with it. (7) The interacting particle phase conjugates (time reverses) the component with which it interacts. (8) The primary interaction with negative electron shells of atoms yields positive energy and positive time: constituting ordinary physical reality as we observe it. (9) The secondary interaction with the positive atomic nucleus produces phase conjugated (time-reversed) physical reality. (10) Secondary interaction (nuclear phase conjugation) normally produces Newton's third law of motion. (11) In Newton's third law, the reaction force need not be equal, and need not be antiparallel. Note, however, in such case local spacetime is curved. (12) The present law of entropy is only the positive-time half; the other (negative time) half is the negentropy law.

Phase Conjugation (Time Reversal)

It is a most remarkable fact that Soviet radar engineers and radar scientists are also trained in optics. While this seems little short of astonishing by Western standards, the Soviets have an exceedingly good reason: As a matter of course, most Soviet radar systems are also operable in, or associated with another system that is operable in, a time-reversed (phase conjugate) mode.

In this mode the radar and/or its associated system is also an extremely powerful directed energy weapon, having capabilities undreamed of in the West.

A phase-conjugate (time-reversed) wave actually is a wave that travels backward through time.* That is, it is capable of precisely retracing the path through space, taken by another wave that traveled that path to a nonlinear mirror, stimulating the reflection of the time-reversed wave.

Further, in retracing its invisible path through space, the phase conjugate replica wave does not diverge as do normal waves. Instead, it continually converges upon its invisible trace.

Consequently all its energy arrives back at the distant source that emitted the stimulus wave originally. This was Tesla's original secret of his "wireless transmission of energy at a distance with no losses."**

It can easily be shown, however, that the phase conjugate wave is a legitimate solution to the wave equation. Hence it is a general property of all types of waves: sound, magnetohydrodynamic, etc. It is not confined just to the optical band.

Obviously such a startlingly different wave has phenomenal uses. Literally it ushers in a new physics.

Let us turn now to look at this strange wave that has emerged on our scientific horizon.

In the late 1960's, Soviet researchers cautiously pointed out the possibility of a time-reversed wave. Papers appeared in the open Soviet scientific literature, and shortly thereafter U.S. physicists began to read about it.

In 1972, Soviet scientists visited Lawrence Livermore Laboratory and specifically briefed U.S. scientists on the optical phase conjugation (OPC) phenomenon. Thereafter the effect became of increased interest to American scientists.

(Note, however, that phase conjugation (PC) is a major phenomenon of all physics. It is not confined merely to the optical spectrum, nor just to electromagnetic waves. Phase conjugation is a general phenomenon true of all waves, regardless of type. We will primarily discuss OPC, since most Western experimentation has been in the optics domain, and the literature is rich in that respect. However, we also point out that Western scientists have absolutely no inkling of the mechanism causing OPC. They have simply written a model for a time-reversed wave by taking the equation of a normal wave und writing the conjugation (adding an asterisk exponent) to the appropriate term, and pointed out that this new equation also satisfies Maxwell's wave equation. They also do not clearly recognize that an EM wave carries time and energy, and that a time-reversed EM wave carries negative energy and negative time. Further, they do not at all comprehend the startling implications for generating gravitational fields, including antigravity and inertial effects, directly on the laboratory bench.)

---

*The external positive-time observer sees this time reversal as length (space) reversal.

**The great electrician had actually discovered what today is known as phase conjugation or time reversal of EM waves.

In optical phase conjugation, when an input EM wave (ordinary) enters a nonlinear medium, the medium "reflects" or produces a strange, time-reversed EM wave in response.

Figure (35a) shows the effect. In the figure, E1 is the normal input wave, entering a distorting, nonlinear medium (the "blob") as shown. Wave E_1 moves on through the blob, emerging on the other side and continuing, though now in distorted form.

The medium produces a second wave, E_2, which precisely retraces the steps of E_1 That is, everywhere E_1 is, E_2 now appears. The phases of the E_2 wave are precisely reversed from those of E_1 The only difference in the two waves is that wave E_1 carries positive time and positive energy, while wave E_2 carries negative time and negative energy.

(U.S. physicists are well aware of everything in the above statement except for the negative energy portion.) The historical background of the phase conjugate wave is given in Table 25.

(We strongly stress that the West did not discover the time-reversed wave; the Soviets did. At the time (late 1960's) this appeared in the open Soviet literature, the Soviets had already been using it for about two decades in their giant energetics (scalar electromagnetics) weapons development program.)

However, to return.

The time-reversed wave is called the phase conjugate replica of wave E_1. The overall characteristics of the phase conjugate replica E_2 are summarized in Table 26.

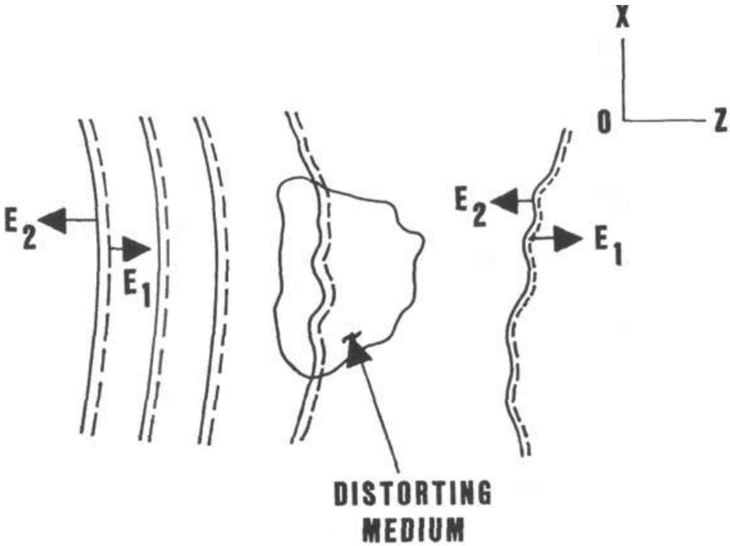
X
O
Z
E_2
E_1
E_2
E_1
DISTORTING
MEDIUM

a. Phase conjugation by a nonlinear, distorting medium.

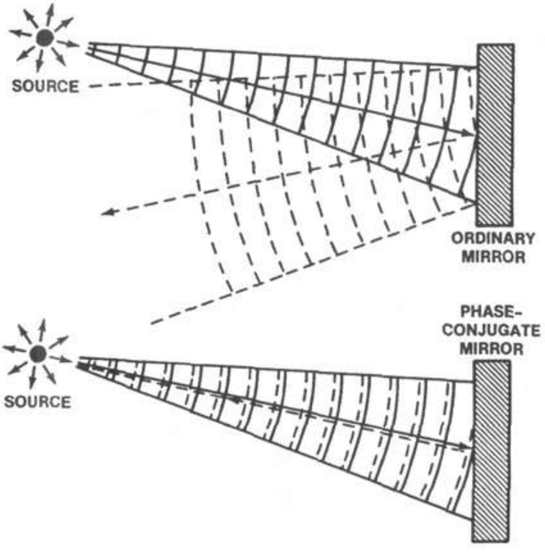
SOURCE
ORDINARY
MIRROR
PHASE-
CONJUGATE
MIRROR
SOURCE

b. Ordinary mirror reflection versus phase conjugate mirror reflection.

Figure 35. Phase conjugate reflection of a time-reversed wave.

Table 28. PHASE CONJUGATE WAVES

- NONLINEAR OPTICS TRANSIENTS
  - DISCOVERED BY SOVIETS
  - LATE 60s-EARLY 70S
- **PRESENTLY** "HOT ITEM"
- INITIALLY THOUGHT RARE
- NOW KNOWN WIDESPREAD
- THIRD POWER OF E-FIELD
- S**TILL** MUCH UNEXPLORED

Table 26. PHASE CONJUGATE REPLICA

- TIME-REVERSED
- CARRIES NEGATIVE TIME
- CARRIES NEGATIVE ENERGY
- "HEALING" PROPERTY
- DISTORTION CORRECTION
- F**ROM** DISORDER TO ORDER
- **MAY** DWARF INPUT WAVE

Table 27. DISTORTION CORRECTION THEOREM

"If a scalar wave E_1 (r) propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z=0] its phase conjugate replica E_2 (r) , then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

We point out here that the basic nonlinearity (nonlinear medium) in spacetime is a single particle of mass. In a photon interaction with that particle of mass, the interacting (input) photon (positive energy, positive time) causes the generation and emission of a phase-conjugated (time-reversed) photon. This time-reversed photon provides the precise mechanism for generating Newton's third law: **For every action there is an opposite and equal reaction.** This will be covered later.

But back to our nonlinear medium, and the two waves E_1 and E_2.

E_1 and E_2 are "locked together" as a scalar EM wave unless something else is specially done to separate them. That is because the nonlinear medium is a *modulator,* and the two waves modulate each other (lock together as a single wave).

(We interrupt again to strongly point out this fact: This directly shows that all EM waves in vacuum exist as scalar EM waves, not as the so-called Maxwell-Hertz transverse waves of conventional theory. Nikola Tesla, the greatest electrical scientist the world ever produced, always adamantly insisted that EM waves were "sound" type waves, not "transverse" waves. For one thing, the vacuum-medium would have to be a solid to produce transverse waves! Tesla even visited Hertz in Europe and tried to convince him of the error of his so-called "proof of transverse EM waves." Tesla found Hertz, adamant on the subject, however, and returned without changing Hertz's mind. Ironically, Tesla was right and all the textbooks in **the** Western world have been consistently in error as to the nature of the EM wave in vacuum.)

We simply state here that, for a single-frequency EM sine wave in vacuum, a "full photon" of that wave in vacuum is actually a *single sine-wave oscillation.* The positive (positive energy, positive time) half of that sine wave is the same thing as a virtual electron (negative charge). The negative (negative energy, negative time) half of that sine wave is the same thing as a virtual positron — a virtual electron travelling backwards in time, or "time-reversed", which we detect as length-reversed and charge-reversed.

Further the vacuum — including the vacuum inside each relatively empty "mass structure" such as an atom or molecule — is **a** seething inferno of activity of virtual particle fluxes. It is upon that medium that all physical reality and physical changes are impressed. Thus there is nothing static in the world; what actually exists is *switching of virtual particle fluxes.* What we call a "static thing" is only an equilibrium in this switching activity. And, once switched on, any stream continues until switched off or switched to change it.

For that reason, the full photon has always contained twice as much absolute energy as is provided by the presently used equation

$$E = hf \qquad (4\text{-}13)$$

Where E is energy, h is Planck's constant, and f is frequency.

One half the photon carries negative energy and negative time, while the other half carries positive energy and positive time. In the full photon interaction, one half of the photon interacts and "switches off" the stream. Thus the photon interaction makes a single change of action, energy x time, both switching on the beginning of the quantum change and switching off the ending of it.

The full photon in vacuum is, and has always been, a scalar EM photon. It rigorously is the presently accepted photon plus its directly "hooked on" and accompanying phase conjugated replica photon.

That is, in terms of the present physical view, the scalar electromagnetics view of a full photon in vacuum is that it consists of a photon and its antiphoton, directly combined in serial fashion. Further, this antiphoton part is not identical to the photon, since it is time- and energy-reversed.

The present physics makes no distinction between the two types of photons, hence makes no distinction between positive and negative photon energy. It also still considers "static" things to independently exist. It does not pursue the fact that all things appearing "static" are equilibrium states in nearly unbelievably intense virtual particle streams. Since it does not pursue that fact, it does not build a "switching" model for quantum change, not does it arrive at a method of directly engineering these "switching" operations.

For that reason, present physics has totally missed local general relativity and the engineering of spacetime curvature by electrogravitational methods, as was contained in

Maxwell's original theory.

For the same reason, present Western physics has not worked out the basic mechanism that produces phase conjugation, and it has not realized that all conservation laws can be locally violated with the new methods — methods actually hailing back to Maxwell's original theory!

Now returning again to our nonlinear medium and waves E_1 and E_2:

It is possible to separate (demodulate) wave E_2 from wave E_1 In that case, E_1 is an ordinary EM wave, carrying positive energy and positive time. E_2 is a phase conjugate replica, carrying negative energy and negative time.

Irradiation of an atom in a new sample of material with a normal EM wave such as E_1, just provides the normal photon interaction with the electron shells. That is, the electron part of the full photon adds to the orbital electron's energy state in positive time, raising the orbital electron to an "excited state." (Remember, this has switched on a continuous process of "keep raising the energy level of the electron" in positive time.) The "state-increasing" orbital electron then interacts with the second, positron-part of the full photon, which switches off the increasing, leaving the electron in its "static excited state." Simply put, one half of the full photon turns on an uplifting spray, and the second half of the full photon turns off the uplifting spray, leaving the electron in equilibrium state at a new level.*

Thus in orthodox physics we have taken the view that an electron in a shell absorbs a "photon" from the incident wave, taking extra energy and being raised into an excited energy state.

In the orthodox view, the excited-state electron then decays to release another ordinary photon (since it comes from a negative charge).

What actually happened was that, when the "full photon" of the incident wave is absorbed by the orbital electron, increase of the electron energy is switched on, then switched off, leaving the elec

*We state here a magnificient principle without proof. When one furnishes (inputs) electrical power to a system, all the "input work" is expended purely upon switching fields by fighting against them in brute-force fashion. E.g., electrical power furnished to an electric motor does **not** run the motor, but only continually wrestles the internal fields to switch them. Obviously, to build an over-unity device, one must get it to switch its own fields, at least a significant fraction of the time.

tron in equilibrium at a new energy level.

What happened in the middle was phase conjugation. In absorbing the first half of the incident full photon, the electron acted as a phase conjugate mirror (we will explain that term shortly). It emitted a forward-travelling phase conjugate replica (PCR) of that first photon half, and this conjugate replica entered the nucleus, where it is absorbed to cause the Newtonian reaction force. Note that this PCR emission by the electron does not generate recoil in the electron-PCM, as is well known in OPC theory and experiment.

The same type of phase conjugation happened when the state-increasing electron absorbed the second half of the incident full photon. It phase conjugated again, sending this half to the nucleus to acompany the former PCR half. In the nucleus, this second half switches off the increase in excited state of the nucleus that was engendered by absorption of the first half.

The photon interaction is a process for switching on and off a continuous rate of action increase or decrease.

At the conclusion of the full photon absorption, both the orbital electron and the nucleus are in excited states. The nucleus has incre**ased** its gravitational mass.

Then the situation precisely reverses. Both nucleus and electron dec**ay,** again emitting full photons and again phase conjugating.

The PCR photon emitted by the nucleus reduces the positive mass of the nucleus. The PCR photon emitted by the electron is reversed from the electron's absorption record, reducing the negative charge's energy state, and the electron is said to "decay" from its excited state. The PCR photon emitted by the electron now (to the observer) will appear to carry positive time and positive energy.

(In passing, we point out that, actually, the situation is far more complicated. Both the electron and the positively charged nucleus are enmeshed in continuous streams of virtual particle flux, including intense streams of virtual photons. One must go to n-wave mixing, where n is greater than 4, to fully understand the total reaction. The main points are, the Newtonian reaction forces within the atom are generated by phase conjugation, the photon emitted by the decay of an excited state electron will normally carry what we call positive energy and positive time, and the photon emitted by the decay of an excited state positive charge will normally carry what we call negative energy and negative time. Full coverage of these interactions is beyond the scope of this book, and must wait for another effort in the future.).

Now once again we return to our simplified model of phase conjugation, and waves E_1 and E_2 in our nonlinear, modulating medium!

As we stated, we can engineer the situation so that the phase conjugate replica E_2 can be separated from E_1. A separated E_2, however, is a quite different breed of cat. It is a phase conjugate wave, carrying negative energy and negative time.

E_2 is a magic wave indeed!

Irradiation of the nucleus of an atom in a new sample of material with a phase conjugate replica such as E_2 causes the nucleus of the atom to interact in reverse of its normal interaction.* That is, one of the positive charges in the nucleus interacts with a photon resulting from wave E_2, and goes into an excited "negative energy" state. That state, in decaying, emits an ordinary virtual photon which promptly interacts with a virtual positive charge in the nucleus, which decays to emit the virtual photon, which interacts with another positive charge in the nucleus, etc.

In other words, the negative energy of the absorbed antiphoton is captured by the nucleus, and the nucleus overall does not *immediately* decay to discharge the negative energy.

Instead, the nucleus acts as an accumulator or capacitance, charging up and then discharging far more slowly than the normal photon/electron interaction.

(For the purist, the nucleus does phase conjugate the input negative energy/negative time wave, and phase conjugates it again — upon an electron in the electron shells, through spin coupling This second phase conjugate replica is now just a normal EM wave, and the electron exhibits a normal EM interaction. Thus **our** instruments (Which are electron wiggle detectors) detect a normal

*Irradiation of the nucleus with a PCR wave is easily accomplished by using ordinary magnetic resonance, and modulating the PCR wave upon the magnetic carrier, as Antoine Priore did.

EM wave detection and do not discriminate between photon and antiphoton halves of a full photon. For that reason, scientists have not clearly noticed the difference between positive and negative energy, and between photons and antiphotons. They have also not resolved the basic conflict between the particle and wave theories of electromagnetics. *Note that the full photon concept, where the full photon is a normal photon and phase conjugate replica of that photon hooked together into a single sine wave, resolves the long-standing wave-particle controversy.)*

In most of their instruments, present scientists are detecting a secondary normal EM wave for the phase conjugate replica interaction, since the phase conjugate interactions in the instrument are not sorted out, and since the instruments measure only one kind of PCR, that emitted by a negatively charged electron. Hence scientists think the antiphoton has positive energy, and interpret it in positive time. Consequently they tend to accent that the wave is "phase-reversed," and not accent the negative time/negative energy implications since they have not clearly sorted them out.

Let us return to our consideration of the irradiation of the nucleus of an atom (with an incident E_2 phase conjugate replica wave carrying negative energy and negative time, with respect to our electron-wiggle-detecting observer).

AS irradiation continues, the actual "structured charge pattern" existing on the incident E_2 wave slowly charges up the nucleus of the atom with that exact charge structure, in the negative energy/ negative time state.

We shall find that most interesting fact of great utility when we later discuss the mind, thought, cellular control systems, and generating a specific electromagnetic phase conjugation pattern to reverse cellular disease.

We shall also find that it is of great interest in accomplishing antigravity when we use extremely low frequency PCR irradiation.

The Phase Conjugate Mirror

Any system which phase conjugates the input wave and returns a phase conjugate "reflection" is called a phase conjugate mirror, or PCM for short.

But the phase conjugate mirror has some very unique properties, as can be seen in figure 35b.

For example, if you look in a *normal* mirror, you see your own reflection. You see your shoes and legs, for example. Light from your shoe scatters across the mirror, and it hits one spot on the mirror which reflects it at the proper angle to strike your eye.

But that simply doesn't happen with *a phase conjugate* mirror! If you were looking into a phase conjugate mirror (PCM), the light that left your shoe and scattered across the mirror would be reflected right back to the shoe, since it would retrace the path taken by the previous light striking all across the PC mirror from the shoe. So all you would be able to see would be two black dots: the retinas of your eyes. Light reflecting off them and striking the PCM would be reflected back to the source — the retinas — so you could see that light and that light only.

This path reversal is most important to keep in mind. With a phase conjugate signal, you are not functioning with an ordinary signal. Further, it has a "precursor" path taken by the incident wave that stimulated its emission from the PCM. Unless interfered with, the PCR will trace back down the invisible vacuum trace of that incident wave, or back down its wave path, like electricity goes down a wire.

This, by the way, is the secret of Tesla's wireless transmission of energy at a distance with no losses.

When broadcast into space, a normal EM wave (positive time, positive energy) diverges or scatters from its path.

Precisely the opposite happens with a PCR wave. When broadcast (reflected) into space, the PCR wave continually converges back upon the invisible trace taken by its incident stimulus precursor. Thus all the PCR energy is continually converging upon an invisible beam or "wire" through the vacuum, back to the source of the original incident wave.*

If the PCR wave is highly amplified (and means to do this are now well-known), then large amounts of EM energy can be precisely

*One can effectively move the entire imprinted vacuum path, however, by adding an additional input vector wave computed so that the **resultant** input would have come from the desired shifted distant point. In this way the PCR can be steered in space to lead and intercept a distant moving target.

returned to distant points from which a stimulus wave of any kind is received.

If a micro-microwatt is received from a point thousands of miles distant, then a gigawatt can readily be returned precisely to that distant point.

This was Tesla's magic secret. The great electrician had discovered phase conjugation, though he did not use that term. He did, however, point out that he could create an invisible wire through space to a distant point, and could send any amount of energy to that distant point without any scattering losses along the way.

He also accomplished the same thing with sound waves and mechanical waves. He pointed out that he could produce an undiminished physical (mechanical) effect at any distant point, and could eventually split the earth if he added power long enough.

Tesla had given the principles of radar in World War I! He later even spoke of his "big eye that can see at a distance." Here he was probably referring to stimulating the emission of a PCR return from that target. The PCR would automatically correct for the distortion in the intervening medium, and fairly sharp pictures of the distant scene — without showing the size dispersion of distance — could probably be obtained. Note that a non-scattering signal to a distant observer does not make him suffer any loss in field or apparent size and detail of the scanned object, which can be recovered to any resolution, limited only by the state of the art of the physical realization technology.

The implications for far greater resolution microscopy— far heyond the present limitation to about 4,500 — should be immediately obvious. The implications of such a microscope - that can see directly inside atomic nuclei; photograph nuclei and fundamental particles directly; see living viruses, molecules, and atoms directly; and theoretically see even into the virtual state itself— should be immediately obvious.

In discovering phase conjugation, the Soviet Union was attempting to break Tesla's secret of wireless transmission of energy without losses, and the Soviet scientists did so by about 1950. They highly weaponized the effects of phase conjugation, phase conjugate mirrors, and amplified phase conjugate mirrors, both in radar waves, mechanical waves, and sound waves, for nearly two decades. Then the Soviets cautiously stimulated the open literature with a limited paper on optical phase conjugation, to ascertain whether or not we realized the severe weapon implications of time-reversed waves, their production and use, and their amplification.

So phase conjugate mirrors can do very useful things. For example, look again at figure 35a. Suppose we are trying to photograph something on the other side of the distorting medium, say through a distorting gas, such as a turbulent atmosphere. In that case E_1 on the left is, say, light coming from an object on the left and passing through that distorting medium. If we just used ordinary waves in a camera on the right, we would see a very distorted wavefront, as represented by the distorted E_1.

However, if we illuminate the object by sending a wave through the distorting medium, and detect the *phase conjugate* signal that returns, it will have reversed the distortion when it gets to us, and we will get a clear picture. A very nice photo of just this process is contained in David M. Pepper, "Applications of Optical Phase Conjugation," **Scientific American,** 254 (1), Jan. 1986, p. 75.

Phase conjugation has many more uses than we have covered in this brief paper, of course, but this is an important use: the removal of distortion effects from optical systems.

It would, however, be nice to be able to amplify the effect. And so we can.

Amplifying the Phase Conjugate Effect

Let's look now at figure 36, where we will develop how to amplify the phase conjugate replica.

Be careful to notice that we are changing subscripts on you. By convention, optical scientists use A_1 and A_2 as two additional waves, called the *pump waves,* that are used to stress the nonlinear medium, and input signal wave A_4. Wave A_3 then is the stimulated PCM output, or the phase conjugate replica (PCR) wave.

In this figure we show a scheme for amplifying the phase conjugate signal that is generated and returned by the phase conjugate mirror (PCM).

We do this by adding two opposing waves of the same frequency (in this simplified case). These are the two "pump waves", and impressing them upon the PCM is called "pumping" the PCM.

Actually, in producing the pump wave, we are adding two input waves of the same frequency, 180 degrees out of phase with each other. Since the nonlinear medium is a *modulating* medium, the waves are forced to modulate each other and "lock together" as a *single scalar* EM wave of most interesting characteristics. Indeed, its E fields sum to a zero resultant vectorially, and its B-fields sum to B zero resultant vectorially. However, the scalar parts remain and are multiplied together, as in the original Maxwell quaternion theory. Since the magnitudes of the vector components in the scalar summation (quaternion multiplication, since we are addressing modulation) are varying, then the magnitude of the remaining scalar part is varying. We have therefore produced a standing scalar wave in the nonlinear material medium, having zero vector EM gradient components. We have produced a purely scalar EM wave of pure potential — and this is an electrogravitational wave, rigorously, since it represents a time oscillation of the local energy density (local virtual particle flux density) of vacuum. This scalar EM wave passes through the electron shells and enters the nuclei, where it is phase decoupled by the extreme nonlinearity of the violent virtual particle currents of the nucleus. Its energy is then absorbed by the nucleus, raising it to an excited pseudopotential. This potential decays, returning the scalar EM stress wave into the local vacuum and area. We are now pumping the atomic nuclei of the nonlinear PCM material with our two pump waves, and we have created an oscillating local gravitational field around the PCM.

This increased nuclear G-potential, looked at in Kaluza-Klein theory, is a 5-space potential. It is ready to burst out in the 5th dimension (as electromagnetic bleedoff) at the slightest provocation. That provocation is the entrance and absorption of the input signal wave, E_4. Note that E_4 acts as an "initiator" (or a pinprick into a highly inflated balloon, if you will) on the built-up, excited 5-dimensional gravitational pseudopotential of the nuclei. The moment so initiated, the entire 5-space G-pseudopotential collapses, emitting a time-reversed phase conjugate replica (PCR) of E_4. This emitted, highly amplified PCR is labelled E_3. It now "backtracks" down the invisible trace through space taken by the incident stimulus wave E_4, returning to the distant source that originally emitted

Voila! Tesla's wireless transmission of energy to a distant point without losses. *Voila!* The Soviet Launch Phase ABM system and the Soviet Launch Phase Anti-Bomber System, when used in conjunction with an over-the-horizon radar (such as the giant Wood-

Table 28. FOUR-WAVE MIXING

- **NONLINEAR MEDIUM**
- **OPPOSING PUMP WAVES**
    - **- CROSS MODULATION**
    - **- SCALAR WAVE**
    - **- INTO NUCLEI**
- **TRANSMISSION WAVE**
- **REFLECTION WAVE**
- **REFLECTION GAIN**

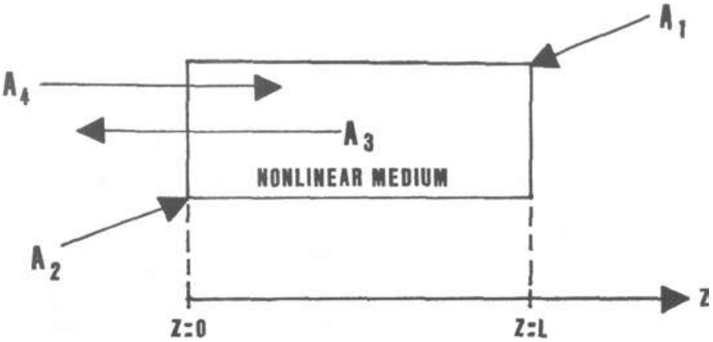

Figure 36. Phase conjugation by four-wave mixing.

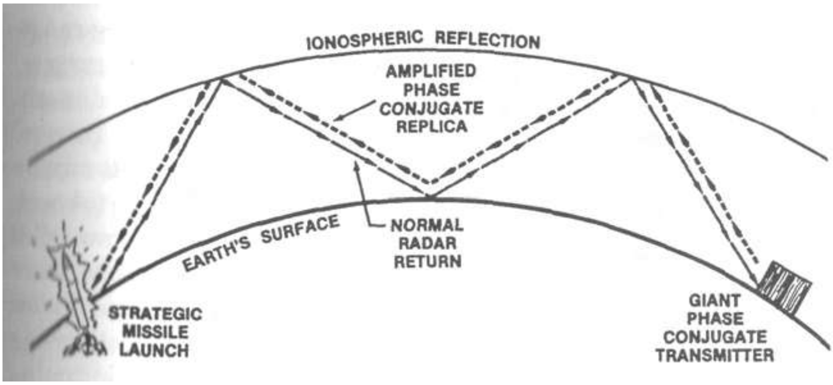
IONOSPHERIC REFLECTION
AMPLIFIED PHASE CONJUGATE REPLICA
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC MISSILE LAUNCH
GIANT PHASE CONJUGATE TRANSMITTER

a. Soviet Launch-phase and midcourse ABM system.

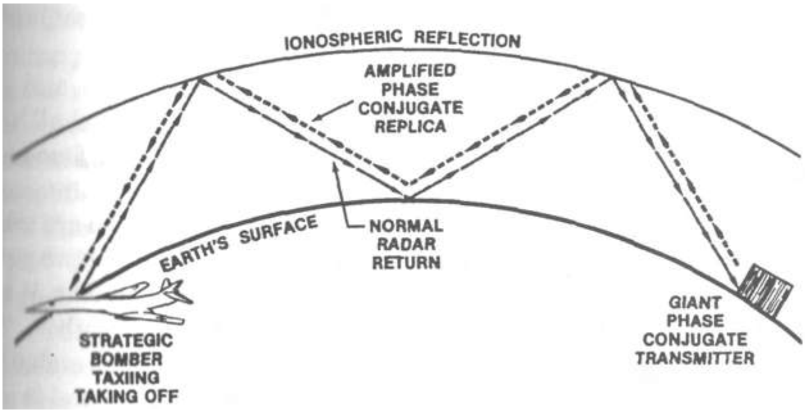
IONOSPHERIC REFLECTION
AMPLIFIED PHASE CONJUGATE REPLICA
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC BOMBER TAXIING TAKING OFF
GIANT PHASE CONJUGATE TRANSMITTER

b. Soviet launch-phase and midcourse anti-bomber system.

Figure 37. Soviet launch-phase and midcourse strategic defense system.

peckers) to provide a reflection signal from a distant target. *Voila!* The true use of Soviet Stand-Off-Jammers to knock out enemy antiaircraft, air control, and anti-tactical missile radars wholesale. *Voila!* The Intermediate Phase ABM system, where a long range strategic radar is used to provide reflections from incoming targets. "Fire and fire again!" is the rule, and 100% defense is directly achievable. *Voila!* How the Soviets knocked down Gary Powers' high-flying U-2 reconnaissance aircraft. *Voila!* How SA-2 missile system radars were temporarily modified in North Vietnam to knock down several hot new F-111 aircraft. *Voila!* How the Arrow DC-8 aircraft, Titan missiles, Delta rocket, and Ariane rockets were "electromagnetically surged" from inside to destroy them. *Voila!* Why the Soviets almost never throw away old radars, since they can now add PCMs and make them formidable destroyers of incoming targets. *Voila!* How one solves the ABM problems of discrimination and weapon kill; simply pulse-fire everything — nuclear material explodes in a full nuclear explosion when struck with potential pulses of sufficient magnitude and duration.

Summarizing figure 36 and amplification of the phase conjugate replica:

When we consider the addition of the input wave and the resulting phase conjugate wave that will be produced, we shall have a total of four waves. Thus this particular process is called four-wave. *mixing.**

On the diagram in Figure 39, A_1 and A_2 by convention are taken to be the two opposing EM pump waves. Notice that these two pump waves stress the medium. Since the medium is nonlinear, it is a modulator. Hence the two pump waves modulate each other. The two waves thus lock together into a scalar EM wave. This scalar EM wave is an oscillating artificial scalar EM stress wave, and it goes directly into the atomic nuclei of the medium and pumps the nuclei themselves.

The atomic nuclei are now rhythmically stressed with the oscillating scalar wave.

---

*Probably the best single introduction to optical phase conjugation in the English language is David M. Pepper, "Nonlinear optical phase conjugation," **Optical Engineering,** 21(2),**Mar.**/ Apr. 1982, p. 156-183. Especially read Pepper's footnote on p. **166.**

We now input a weak signal (wave A_4) into the "pumped nuclei" medium.

By our CPTEGS rule, the positively- charged nuclei generate a negative energy/negative time wave — in short, a phase conjugate replica wave. Further, it is a highly amplified PCR, and may contain *as* much raw energy as was fed into the pumped mirror by our pump waves. This "seeker wave" or "electromagnetic missile" wave then sets out on its invisible path through space, seeking a distant "A_4 source point" and delivering all its energy intact to that point.

The Phase Conjugate Mirror as a Vacuum Triode

The pumped phase conjugate mirror (four wave mixing) can best be understood as a special kind of vacuum triode.

The pumped artificial potential on the atomic nucleus provides a powerful cathode in positive time, and a powerful plate as well in negative or reversed time. The electron shields, in their spin coupling to the nucleus, act as a grid in positive time. The input signal A_4 itself acts as a relatively weak signal onto the grid of the triode.

The positive charge difference in potential between the powerful pump potential of the nucleus (cathode, in forward time) acts as the plate voltage. The decrease of this potential when discharging (increase of the negative time, negative aspect) acts as the production of a plate signal to the negative time operation (production of the amplified, time-reversed PCR wave).

The excited nucleus immediately decays upon stimulus by the "grid signal" A_4, emitting a "time- and energy-reversed" strong signal — in short, amplified phase conjugate replica wave E_3.

The amplified phase conjugate signal moves from the cathode (nucleus) out to the electron shells and beyond, precisely because it is time-reversed. To a time-reversed signal, the grid (normal input) acts as the final plate (external output) — that is, in negative time, the elements of the triode are reversed. This makes it possible for the material to emit the powerful PCR.

I have called this effect/analogy a vacuum triode, to accent the direct engineering possibilities. My associates have already engineered this effect to obtain enormous amplification of energy ("free energy") and practical antigravity, from solid-state devices, directly

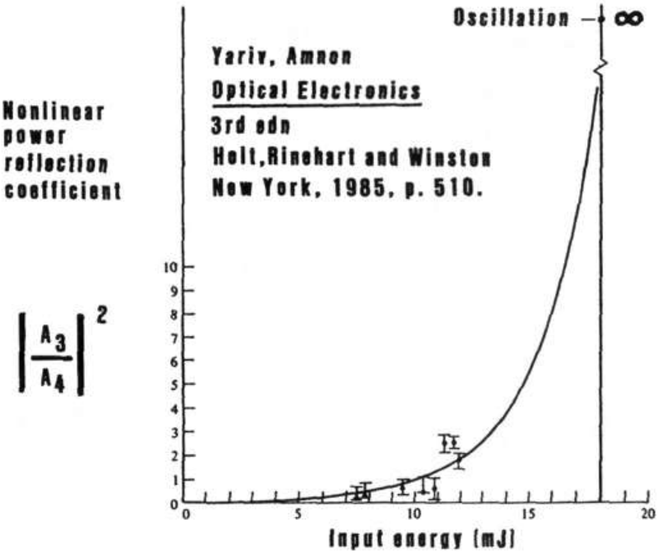

Yariv, Amnon
Optical Electronics
3rd edn
Holt,Rinehart and Winston
New York, 1985, p. 510.

a. Typical nonlinear power reflection coefficients.

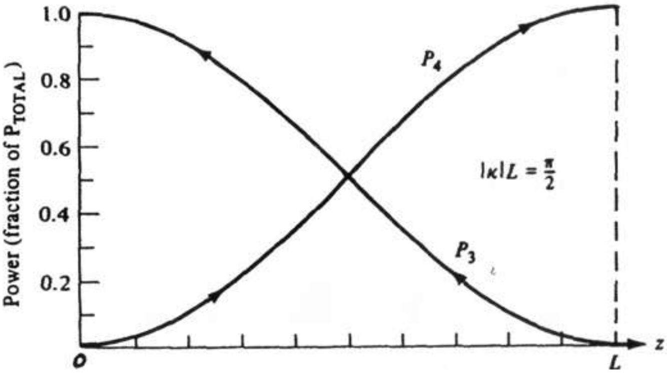

b. Intensities for the oscillation condition.

Figure 38. Power reflection coefficients and oscillation.

on the laboratory bench.

The actual "amplitude gain" of the vacuum triode is the ratio of the amplitude of A_3 to the amplitude of A_4. The square of the amplitude gain is proportional to the power gain; this entity is called the nonlinear power reflection coefficient.

Figure (38a) shows the typical nonlinear power reflection coefficient for a pumped phase conjugate mirror, as a function of the input energy and the depth of the material. As the input signal increases, the gain increases. The ultimate condition would be that the "free oscillation" condition is approached. This occurs when the angle between the input wave A_4 and the pump waves A_1 and A_2 is 90 degrees.

Figure (38b.) shows the oscillation condition and the fraction of the total pump power that is in the input wave A_4 and its output **phase** conjugate replica A_3.

At the oscillation condition, an "infinite" gain exists at the correct depth in the medium. This means that essentially all the energy being fed into the pump waves A_1 and A_2 will now be emitted **in** the output wave A_3. At that point, "saturation cutoff" of the triode output occurs, and it cannot output higher energy, since it has no more.

Notice that, at the oscillation condition, for the slightest output, the full power of the pump waves is fed into the phase conjugate replica. This powerful wave then retraces the path taken by the original input signal, unless something else is done to divert it.*

The interesting question then arises: Since we can have infinite **gain,** can we get a special oscillation condition so that the pump wave furnishes its own energy? In other words, can the pump wave "resonate" with the vacuum flux in such a fashion that it is self-pumping and self-perpetuating?

In at least three materials, the answer is yes. Such effects have **been** experimentally accomplished in special forms of barium titanate. See, for example, J. O. White et al, "Coherent oscillation by self-induced gratings in the photorefractive crystal $BaTiO_3$," **Appl. Phys. Lett.,** vol. 40, 1982. p. 450.

---

*Recall that the PCR can be steered by introducing a minute wave A_5 to vectorially sum with input A_4, In that case, the effective input is the vector resultant of $(A_4 + A_5)$. The apparent distant source point and invisible vacuum path are shifted correspondingly.

Phase conjugation is a universal phenomenon. Theoretically it applies to any sort of wave: electromagnetic, scalar, sound, mechanical, etc. By using it, incredible new effects and an entirely new physics can be accomplished. It is a new and exciting field, little-known outside specialist circles, and still not well-understood. Most discoveries in the field are yet to be made.*

Newton's Third Law and the Full Law of Entropy

For example, the present author noticed that Newton's third law— for every action there is an opposite and equal reaction — is actually a result of phase conjugation. To explain that, we diverge a bit.

First, phase conjugation is also involved in the fact that a charged particle of mass is a little "dynamo" or engine.

Actually, a particle of mass is itself a "nonlinearity" in vacuum spacetime. When a photon (virtual or observable) strikes a particle of mass and is absorbed by it, it's the same as "interacting with a nonlinear medium."

In terms of virtual photons, the charged particle is continually bombarded with them, from the vacuum flux. The charged particle, being a "nonlinear medium", phase conjugates each input virtual photon, emitting a phase conjugate replica.

If it absorbs a normal virtual photon, it emits a "negative energy, negative time" phase conjugate replica.

If it absorbs a normal virtual photon, it emits a "negative energy, positive time" phase conjugate replica — just a normal virtual photon. Note that the particle of mass thus does not exist in time. Or, it exists equally in positive and negative time. Since we only observe particles, it follows that time is not normally an observable, as quantum mechanics has assumed.

The emission of a phase conjugate replica does not affect the momentum of the emitting PCM, as is well-known.

For an observable photon, the particle (nonlinear medium) absorbs the normal photon, which does not affect the energy and momentum of the PCM particle of mass.

If there is no other mass to accept the PCR, it just moves off.

*The U.S. may finally be waking up. SDI is now starting to study EM missiles "in the time domain" as opposed to frequency. See **Av. Wk. & Spurn Tech., Feb. 29, 1988, p. 56.**

If there is another mass to accept the PCR, the PCR is absorbed and phase conjugated by it.

A charged particle in vacuum is continually bombarded by a flux of essentially "randomized" fluctuations. However, on the average these all "average out" to zero. Any particle is automatically in a very high vacuum potential, and is in a state of great agitation at all times, even if it's "at rest" with respect to ordinary observers. It's really "smeared all over the place," so to speak.

When we expose a charged particle to a *normal* potential, the interior virtual particle flux of that potential is disorganized. Hence the particle just "wiggles a bit more frantically," so to speak, randomly in every direction.

However, when we expose a charged particle to an *artificial* potential made by zero summed/multiplied EM force fields or waves, the particle wiggles more frantically in *organized directions.**

This is just as true for virtual particles as it is for observable particles.

Thus the use of such artificial potentials in scalar electromagnetics enables us to overcome the disorder of the internal structure of normal electromagnetics. By this means, we can structure the vacuum, organize electricity and electromagnetic waves, accomplish negentropy, curve local spacetime in complex, deterministic fashions , produce free energy and antigravity, reorganize and transmute nuclei (transmute elements) with miniscule input energy, etc.

Rigorously, this means that the present law of entropy is only half of the full law of entropy. Specifically, it is the "positive time, positive energy" half of the law, and it states that in closed real physical systems, continuing operations of ordered "positive time, positive energy" effects tend to inevitable "disordered" effects produced in the apparatus or system.

In other words, the present law of entropy really states that no ordered system is totally closed. The system will inevitably have energy and actions that escape the system's order, and hence (to the system) this escape represents "loss of order" — or in short, disorder.

The present law of entropy specifically excludes (by implication)

* I.e., **the** electronic noise is partially organized. This is a major key toward building scalar EM detectors. It is also a primary cause of many solid-state circuit malfunctions. The exact principle: (Random changes) + (artificial potential) equals (chaos).

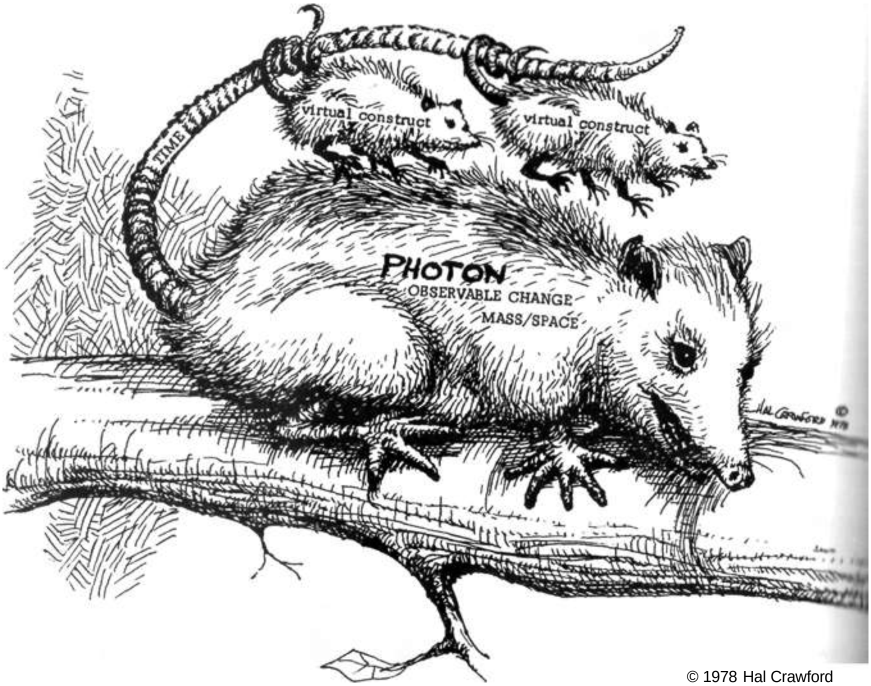

© 1978 Hal Crawford

Figure 39. Photons have virtual substructures. These may be statistical or deterministic.

negative time aspects, where external (disordered) energy time-reverses to again enter the ordered system, in perfect order. By implication, it also excludes negative energy production and operations.

However, **experimentally we can produce both negative energy and negative time, in deliberately time-reversed operations such as phase conjugation and with devices that serve as phase conjugate mirrors. Further, we can amplify the phase conjugation and disorder-to-order operations.**

Hence the present law of entropy is incomplete, and states only half the true possibilities. The other half of the law is the **law of negentropy.** In this law, disordered operations outside the **ordered** system undergo time-reversed (and possibly amplified) effects. These reordered energy effects — which may be highly amplified in comparison to the disordered escaped energy—reenter the "ordered system", restoring order. Since this process moves in negative time, it goes from disorder to order. Hence it is negentropic.

Therefore we have solved the old thermodynamic problem of the "eventual decay of the universe into disorder." Not to worry; the other half of the full law of entropy prescribes negentropic operations in negative time/negative energy, and these restore the order of the universe. That takes care of the problem.

It also means that, in highly disordered systems of many degrees of freedom and far from thermodynamic equilibrium (that is, when the disorder law is saturated), then one can expect to see further stress "create" negative time and hence create order emerging and stabilizing from disorder.

And of course that is exactly what is seen. Ilya Prigogine was awarded the Nobel Prize in 1977 for writing a new thermodynamics predicting precisely such effects, and experiments have proved his case.

But to return to Newton's third law:

Newton abstracted his third law from the classical interaction between two colliding balls.

If one ball is at rest and another approaches it at some velocity, the approaching ball carries momentum and kinetic energy. As it "collides" with the resting mass, quantum mechanics tells us what actually happens: Particles of mass in the moving ball are producing virtual photons continually, and these virtual photons strike the particles of mass of the resting ball and are absorbed by it.

The resting mass now acquires extra energy and momentum from the absorbed photons. At the same time, it acts as a nonlinear medium. It produces phase conjugate replica virtual photons, and these time-reversed virtual photons are emitted. These photons produce no change in momentum or energy upon the emitting mass. However, being phase conjugates, they precisely follow back along the paths taken by the first or "stimulus" virtual photons, striking the moving ball. They are absorbed by that ball, producing (to the external observer) negative momentum and negative energy in it.

These subtract from the kinetic momentum and kinetic energy of the moving ball, reducing them.

The virtual photons emitted by the approaching ball perform positive work upon the target ball.

The virtual, phase conjugated photons, emitted in return by the mass of the struck ball as a PCM, reverse to strike the approaching ball to do negative work upon it.

Notice that twice as much absolute work is always done on the system of two balls as we "input" with the approaching ball.

Also notice that, if we directly "engineer" the phase conjugation of the struck system, we can directly tamper with, and drastically change, the production of the negative energy and negative momentum in the moving ball. Hence we can drastically alter Newton's third law. We can now make a Maxwell's Demon.

This alone clearly establishes that it is perfectly possible to build a so-called free energy device. Though subtle mechanisms must be used, it's nothing more spectacular than putting a paddle wheel in a river, and extracting shaft power from the wheel, furnished by the river's current. Phase conjugation yields other exciting possibilities

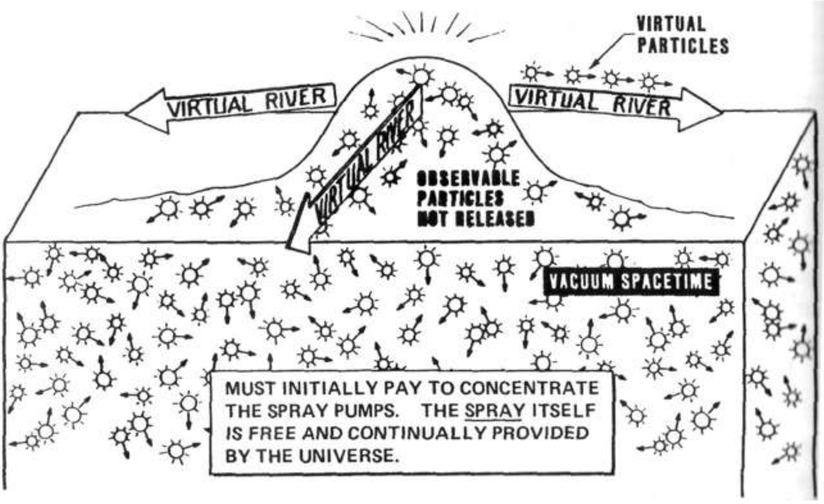

MUST INITIALLY PAY TO CONCENTRATE THE SPRAY PUMPS. THE SPRAY ITSELF IS FREE AND CONTINUALLY PROVIDED BY THE UNIVERSE.

Figure 40. Once established, a virtual river is for free.

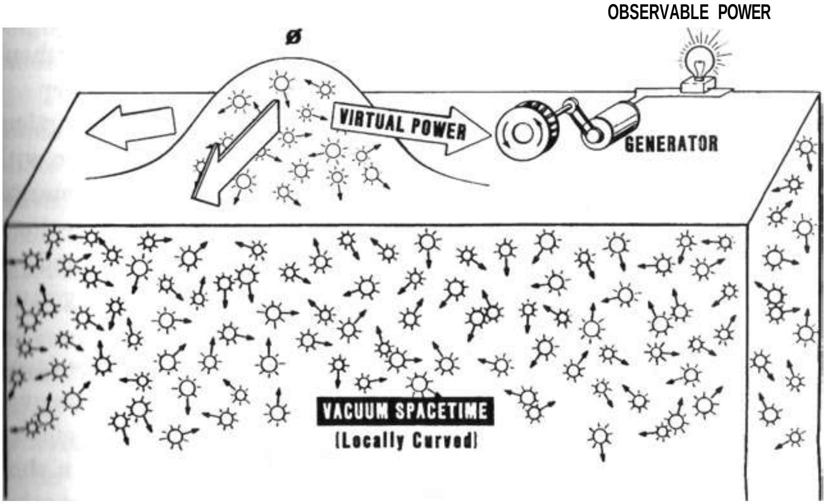
OBSERVABLE POWER
Ø
VIRTUAL POWER
GENERATOR
VACUUM SPACETIME
(Locally Curved)

Figure 41. Local energy conservation can be violated.

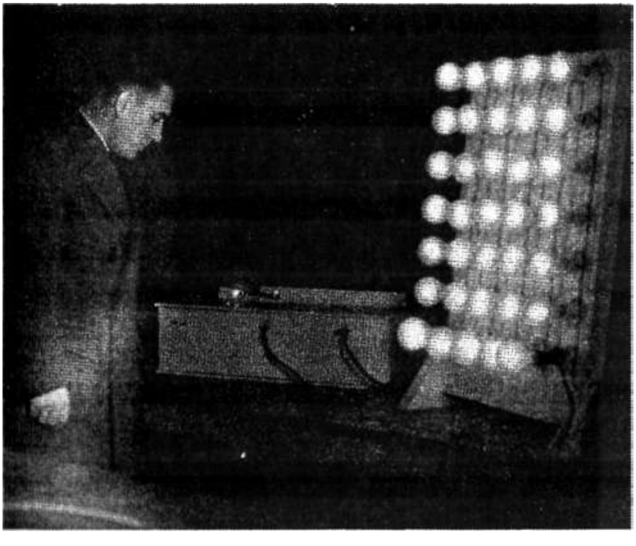

Figure 42. T.H. Moray adjusts the controls of his radiant energy device. Salt Lake City, Utah, in Feb. 1937. Moray obtained 50 kilowatts of power from a 55-lb. device, straight from the ether. Its characteristics (cool circuits, brilliantly lit lamps) reveal **that** It was negative energy.

too numerous to explore here. For just one example, by phase conjugation, real-time holography can be accomplished without having to separately make and illuminate the holograms.

Now with holography and multiple projectors — and particularly with two intersecting "normal" beams to backtrack upon with ever-converging PCR signals—you can assemble a "form of energy" at an appreciable distance from the projectors in three dimensional space. With phase conjugation, you can do it in real time. And you can do it at a great distance back along the path of any signal you use as an input.

By using PCM holography, degradation with distance is not appreciably experienced. After all, the PCR waves utilized are ever-converging. The energy does not diverge and spread. Therefore, if you can make a 6" diameter ball of intense EM energy at a short distance in the laboratory, you can make that same 6" diameter ball of intense energy at several thousand kilometers distance, using PCR real-time interference holography.

With highly amplified PCM's, you can put far more energy into the distant interference form than you can with normal holography. Conceivably it is possible to assemble so much energy in the holography "object" that it condenses into a real material object! In other words, it materializes or semi-materializes.

Both the "intense light form" usage and semi-materialization of forms (such as light forms, flying geometrical forms, etc.) have been created worldwide by the Soviets for decades, to stimulate UFO reports and provide a deception plan for their development and testing of large, strategic energetics weapons using highly amplified PCMs and PCR holography.*

Other bizarre considerations also arise. For example, in **radionics** one utilizes a "witness"—such as a photograph—of the object one is attempting to send "energy" to. (For example, this might bo a sickly plant one is attempting to spruce up). Now when the plant was photographed, the light from it struck the photoemulsion of tho film, producing a photochemical reaction. It also produced phase conjugation in the atomic nuclei of the film emulsion, since the emulsion acts in some respects as a nonlinear medium.

---

*For an example of highly reliable detection and observation of "Soviet hololography UFOs," see Leif Havik, "Project Hessdalen," MUFON UFO Journal, Jan. 1988, p. 4-7. The present author certainly does not imply that all UFO phenomena are due to Soviet PCR holography, only that some UFO phenomena are.

There exists a continuous, invisible "stimulus photon and its connections" trace in time and space, between the plant when the photo was taken and that photo now, even if one or both of them have moved. That trace is highly persistent (as Australian experimenters Reid and Barsamian have shown). A phase conjugate (time-reversed) signal can be made to retrace that track and travel all the way back to the plant, in the present. The effect of using the photo (or sample of the object) as such a "witness" of the real plant and directing weak PCR energy back along the invisible trace into the object is called radionics. In most cases, the two cerebral hemispheres of the operator's brain are depended on to form the necessary scalar "pump wave" to pump the "witness" and produce phase conjugated signal energy, time-reversed to travel back into the distant object.*

The problem with present radionics is that it is far too weak, and it is "operator dependent", changing with the mood, temperament, and skill of the operator. It does objectively work, however, though Often erratically. In countries such as England, France, and Germany, use of radionics to treat human disease is legal and permitted, though operators (for example, in England) must undergo extensive training and certification. Its results are roughly comparable to the results of normal medical doctors. Radionics does not replace orthodox medicine. It can successfully treat certain things that orthodox doctors have little or no success with (such as lupus), and orthodox medicine can successfully treat many things that radionics has little success with (a broken leg or ruptured appendix, for example). The two treatments are best used in conjunction with each other.

However, we must point out that a new, amplified PCM radionics is likely to emerge in the near future. Such a new approach will offer highly effective, powerful treatment of diseases, even at a distance, by amplified PCR electromagnetics. Again, the radiation in and from a photo or witness of a diseased patient can be used as a "grid signal" into a powerful, pumped phase conjugate mirror. The

---

*Again recall that we observe time-reversal of waves as **spatial path** reversal, not as time **travel.**

PCM will produce a specific, powerful, amplified signal that will travel directly to the patient, be absorbed in his entire system, including the atomic nuclei comprising his cells, and directly reverse the chemistry, toxins, and cellular damage in the patient from that disease. It will also kill or destroy the harmful viruses, bacteria, etc. causing the disease, and it can even reverse the actual genetic changes in the cell caused by disease-inducing viruses.

Of course, any powerful tool is two-edged. It can be used for harm as well as good. I think the reader can appreciate the damage that can be done to systems and persons at a distance by means of such powerful new amplified PCM radionics devices.

You can also see the security problems posed to large computing facilities in banks, federal facilities, state and municipal facilities, the IRS, U.S. military weapon systems, ships, aircraft, etc.

Obviously such devices are going to have to be highly regulated and controlled. Else criminal operators will be killing people and destroying things wholesale, secretly and at a distance, and they will have perfect legal alibis to allow them to escape prosecution.

To summarize: if a photograph or witness pattern of an object or person can be correctly stimulated by phase conjugate electromagnetics so that the emulsion acts again as a PCM, can one not argue that—theoretically — phase conjugate energy from the photograph and the instrument can be sent back along all paths, all the way back to the original object? After all, this has been rigorously demonstrated by University of Sydney researchers. Is there not an actual physical mechanism for radionics? For highly amplified, powerful radionics where the operator himself is removed from the circuit? Just because it changes our present notion of physical reality, must we dogmatically reject the results being shown on the laboratory benches, some of which are in the hard-core scientific literature?

Phase conjugation and time-reversed engineering have a diversity of applications unparalleled in the history of science. They literally present us with a new physics. New rules apply.

Most of the major discoveries in phase conjugation have yet to be made.

Scalar Electromagnetics

The scalar electromagnetic approach modifies classical electromagnetics (EM) to include gravitational waves and effects.

To do this, scalar EM utilizes summed-zero vector forces and force fields to construct polarized vacuum potentials.

Although externally it has zero E and B vector gradients, an "artificial" potential possesses an internal, dynamic E and B vector field structure. This "infolded" (Bohm's term) structure has finite size and is deterministic. The infolded structure of the artificial potential in turn deterministically structures and polarizes the vacuum and curves spacetime locally, contrary to presently assumed limitations of general relativity.

Since in free vacuum a potential normally extends to infinity in a decaying exponential fashion, then this infolded E and B vector structure of an artificial potential extends to infinity in a decaying exponential fashion.

The artificial potential itself may be rhythmically varied either in structure, magnitude, frequency, or all of the above. This produces waves of potential, and waves of the structure of vacuum/spacetime — again, in the free vacuum case, reaching to infinity in a "decaying exponential" fashion. These "scalar EM waves" are gravitational waves.

Each photon (one major wave length of the "carrier", complete with modulations) of the scalar EM wave also has structure (its included modulations).

Modulated waves have compound or "giant" photons — photons containing infolded photons.

Each compound photon of such a scalar EM wave is a vacuum engine. It deliberately structures and patterns — dynamically — the energy density and charge of vacuum.

(1) By canonically varying two or more components of the electromagnetic vector structure of the zero-summed EM force vector system, the local structure of vacuum spacetime is macroscopically varied in its internal composition. This is the compound variation of "curvature within curvature" — and hence hyperdimensional.

Table 29. A ZERO VECTOR SYSTEM:

- **IS AN ACCUMULATOR**
- **DOES "INTERNAL" WORK ON MEDIUM**
- **HAS NO "ENERGY" EXCEPT STRESS OF VACUUM, YET**
- **CAN YIELD THE ENERGY OF ITS COMPONENTS WHEN DISSIPATED**
- **IS A SPECIAL SORT OF "PUMP"**
- **CAN YIELD CONTINUOUS ANENERGY FLOW IF NOT DISSIPATED**

(2) By prohibiting internal canonical variance while coherently varying the amplitudes of all the EM components, the total stress of vacuum/spacetime — and hence its overall curvature — is locally varied, without modifying its structural form. This is simple variation of overall curvature of local spacetime, involving primarily relativistic effects.

Rigorously, each of these two methods produces a localized gravitational wave, where the local stress of vacuum is deliberately patterned as well as oscillated in amplitude.

Both methods may be applied simultaneously (and multiple infolded times) to produce an even more sophisticated gravitational wave, and more sophisticated structuring of the local stress energy density of vacuum.

The resulting unified electromagnetics/gravitation is called **scalar electromagnetics,** since the electrogravitational effects are obtained by deliberately opposing EM vector force fields so that they vectorially sum or multiply to a zero vector EM resultant, while **the** infolded (Bohm's term) EM vector components structure and vary the stress energy density of vacuum.*

*Gravitational potential is just "in-folded and locked in" dynamic electromagnetic forces. The infolded EM energy is locked-in, representing a change in the local energy density of vacuum and hence a curvature in spacetime. Electromagnetic force fields are just the out-folding of the G-potential's inner EM contents into the 5th dimension. Ordinary gravitational force field is just the "trickle leakage" of the G-potential's infolded contents, out into and through 3-space.

Note immediately that scalar EM also deals with internally structured and patterned electricity and electromagnetism.

"Normal" electromagnetics has no deterministic internal structure.

The new electrogravitational (EG) wave is called a scalar EM wave, and is believed to have been originally discovered by Nikola Tesla.

James Clerk Maxwell was aware of the potential for electromagnetics to stress and structure the vacuum ether.

His original electromagnetic theory — written in quaternions, not vectors (which had not yet been completed by Oliver Heaviside) - can allow for these effects to be expressed.

Unfortunately Heaviside's interpretation of Maxwell's work

Table 30. STRESS IS FUNDAMENTAL

- **THE MOST FUNDAMENTAL REALITY IS STRESS**
- **THE COMPONENTS COMPRISING STRESS ARE THE GREAT CAUSATIVE AGENTS**
- **MOST OF THESE ACTIVE AGENTS ARE HIDDEN INSIDE ZERO-VECTOR SYSTEMS**
- **THE COMPONENTS OF STRESS MAY BE STATISTICAL OR DETERMINISTIC**

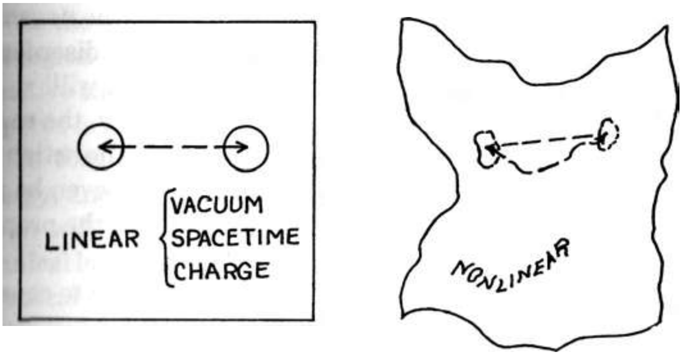

figure 43. Charge affects anything existing in spacetime.

eliminated these effects, and they have remained eliminated from the Western EM theory to the present day.

The use of the scalar EM wave directly engineers the virtual state and the vacuum itself.

Hooper's work represented important work in this respect, as does other related experimental work by Bedini, Watson, Golden, Dea, Faretto, Beck, and other inventor colleagues who do not wish their names mentioned.

Scalar EM theory also bears a strong relationship to the Kaluza-Klein unified theory of gravitation and electromagnetics. At least five dimensions — four spatial and one time — are required as a minimum.

Nested levels of virtual state may be modeled as identical to successive hyperspaces.

Internested levels of zero-summed vector EM force fields allow the direct engineering of those hyperspaces.

Negative energy and negative time flow are especially important when building free energy devices and antigravity. The mechanism producing negative energy and negative time flow is the forced amplified production and absorption of phase conjugated waves.

In addition, phase conjugation is a basic phenomenon of nature: for example, as previously pointed out, it is the direct mechanism that generates Newton's third law of motion. It also adds a negentropy law (for negative energy/negative time).

Accordingly, if we engineer phase conjugation, we can change Newton's third law to our will. And we can invoke the negentropy half of the total law of entropy, so that we can go from disorder back to order. Again, we can make a Maxwell's Demon at will!

In Newton's third law, when manipulated locally, the reaction force need not necessarily be equal in magnitude to the action force It also need not necessarily be parallel to it. It can even be phase conjugated itself to yield an addition force, assisting the propelling of a particular engine or device.

Phase conjugation may also be manipulated so as to cause like magnetic poles to attract, and unlike poles to repel.

Magnetic "force field lines" of a bar magnet normally exist in conjugate co-existing pairs. One of the pair is in observer positive time, and (by convention) runs from the north pole to the south pole. The second line of the pair is a phase conjugate of the first, and exists in negative observer time. Thus, to the "positive time only" observer, it will appear to run from the south pole to the north pole, perfectly retracing the path of its twin.

Howard Johnson uses a "two particle" theory of magnetic field, for example, where one particle (photon) is the phase conjugate of the other. This is identically the same as the scalar electromagnetics concept of phase conjugate pairs of field lines. By very complex, specialized, compound permanent magnets, Johnson is able to partially separate the two lines, and spatially concentrate or diminish the phase conjugation (negative time) component in a given area of operation.

If the negative time component lines are concentrated to outweigh the positive component lines in a local region, then the laws of magnetics are reversed in that region. There like poles attract and unlike poles repel.

In his rotary permanent magnet motor design, he concentrates the negative time component in that part of the rotation where a north rotor pole is approaching a north stator pole. Normally, in this part of the cycle, one has to overcome the repulsion of like poles to "force energy" into the device for later extraction.

However, *now* in that region *like poles attract.* There the north pole of the rotor is attracted toward the north pole of the stator, adding impetus to the rotor.

As the rotating north pole passes the stator north pole, it leaves the region of negative time concentration and returns to a normal "positive time" region again. Now the north poles repel again, and impetus is again added to the rotor.

In that manner Johnson is able to violate local conservation energy, and make a "free energy" permanent magnet motor. The motor will rotate itself and deliver constant shaft power to a mechanical load. By attaching a standard electrical generator, normal electrical power is continually and freely produced by the complete motor generator power system.

Johnson's complex nonlinear magnets severely curve local spacetime, allowing *local* violation of energy conservation.

Scalar EM View of the Vacuum

In the modern quantum mechanical view, the vacuum is not an emptiness, but instead is a plenum.

Today the vacuum is considered to be filled with incredible virtual particle activity. From nowhere, virtual particles continually arise — even with fierce energy — then disappear again into nowhere, so rapidly that they cannot be individually observed.

However, these virtual particles are quite real, for they cause all the forces of nature when they interact in the aggregate with observable particles.

To the observer, the fleeting particles also appear and disappear in both positive time and negative time. That is, the flux consists of both particles and antiparticles.

Thus the vacuum is a seething inferno of virtual particle fluxes.

The concept of an **ether** is again accepted. It refers to this "virtual flux" vacuum.

Note, however, that this is an ether far different from the old material ether that was theorized prior to relativity.

In the new vacuum ether, every imaginable type of particle continually and spontaneously arises (creation) and disappears (annihilation) at every point in the vacuum, according to modern quantum mechanics.

The rate of this seething virtual particle creation and annihilation is essentially unlimited. Hence the "flux density" of vacuum is essentially unlimited.

Further, any virtual particle created has a flux of even finer virtual particles associated, and so on without limit.

The vacuum's virtual particle flux is thus comprised of nested levels of ever finer virtual particle fluxes, in the modern view. We state without proof that **each deeper virtual level may be modeled as a higher dimension** (hyperdimension). **This yields an infinite-dimensional vacuum spacetime (hyperspace) that is identically the infinite nested levels of virtual state.**

At the same time, every nucleus in the universe is continually absorbing and emitting scalar EM (electrogravitational) waves. The emitted scalar EM waves of pure potential are waves in the stress and structuring of the vacuum itself.

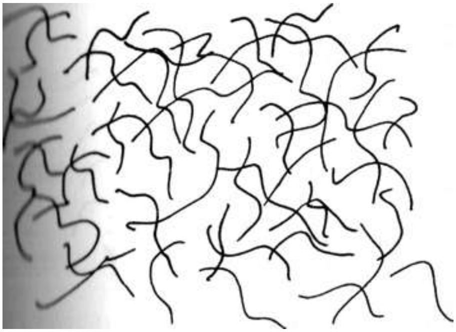

VACUUM IS IDENTICAL TO:

- **VIRTUAL PARTICLE FLUX**
- **SPACETIME**
- **ANENERGY**
- $Ø_o$
- **MASSLESS CHARGE**

VACUUM IS:

- **MADE OF UNQUANTIZED ACTION**
- **WITHOUT DEFINITE LENGTH INTERVALS**
- **WITHOUT DEFINITE TIME INTERVALS**
- **N-DIMENSIONAL (UNFIXED)**

Figure 44. Vacuum/spacetime is pure virtual particle flux.

Table 31. VACUUM IS:

- **SPACETIME (L^nT, WHERE n >= 3)**
- **CHARGE (MASSLESS)**
- **ELECTROSTATIC SCALAR POTENTIAL** $(Ø_o \neq 0)$ **(ERROR IN PRESENT THEORY)**
- **BROKEN BITS OF ENERGY (SUBQUANTAL)**
- **PURE VIRTUAL PARTICLE FLUX**
- $Ø$ **-WAVE FLUX**
- **MULTILEVEL, STRUCTURED, PATTERNED**
- **A VIRTUAL PLENUM**
- **AN OBSERVABLE EMPTINESS**

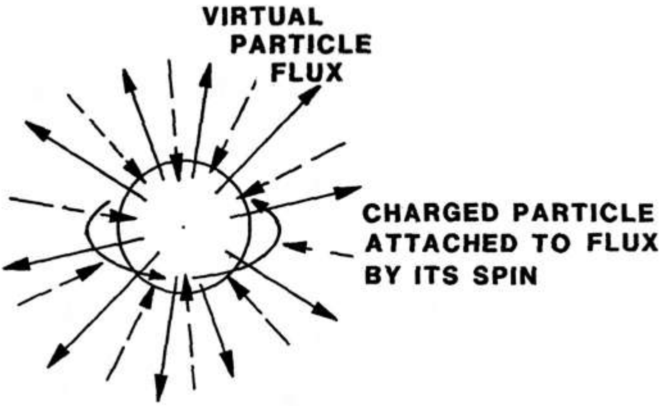
VIRTUAL PARTICLE FLUX
CHARGED PARTICLE ATTACHED TO FLUX BY ITS SPIN

Figure 45. A simplified charged particle.

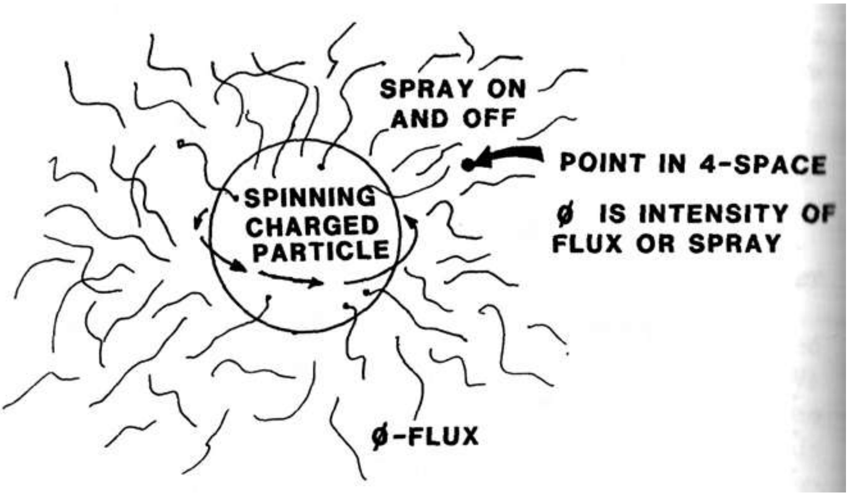
SPRAY ON AND OFF
POINT IN 4-SPACE
0 IS INTENSITY OF FLUX OR SPRAY
SPINNING CHARGED PARTICLE
0-FLUX

SPRAY IS THROUGH 4 OR MORE DIMENSIONS, NOT JUST 3

Figure 46. A charged particle is a special kind of "spray nozzle."

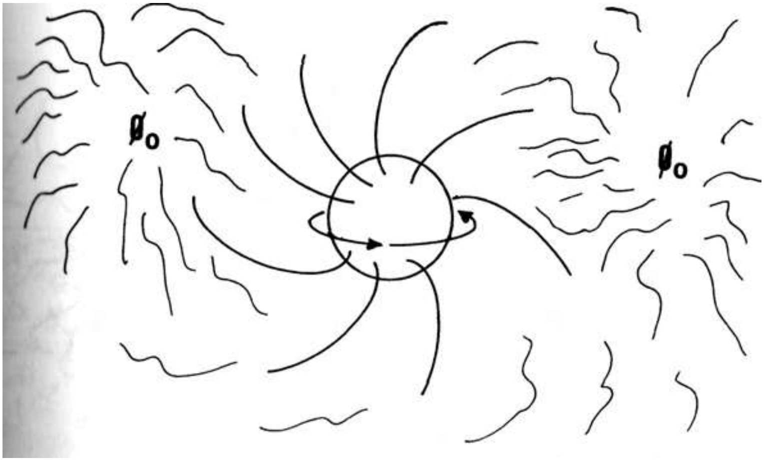

Figure 47. An observable charged particle is directly connected to virtual particle **vacuum** flux.

Table 32. SPIN COUPLES VACUUM TO PARTICLE

- **SPHERICAL ROTATION IS THE KEY**
- **A PARTICLE CAN BE MODELLED AS A SPHERICALLY ROTATING VORTEX OF SPACETIME**
- **ITS MASS IS DUE TO ITS SPIN**
- **THE SPINNING OBJECT IS CONTINUALLY CONNECTED TO ITS ENVIRONMENT**
- **VERY HIGH VALUES OF ELECTROSTATIC POTENTIAL CAN INDUCE RELATIVISTIC CONDITIONS**
    - **CHANGE RATE OF FLOW OF TIME**
    - **EVEN THOUGH VELOCITY IS NONRELATIVISTIC**

E.P. BATTEY-PRATT AND T.J. RACEY,
"GEOMETRIC MODEL FOR FUNDAMENTAL PARTICLES,"
INTL. J. OFPHYS.19. NO. 6, 437-475, 1980

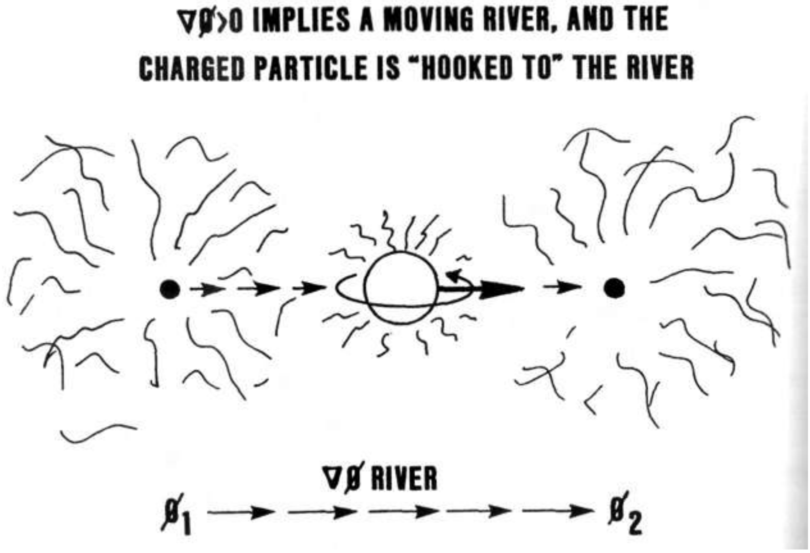
∇Ø>0 IMPLIES A MOVING RIVER, AND THE
CHARGED PARTICLE IS "HOOKED TO" THE RIVER
∇Ø RIVER
Ø1
Ø2

Figure 48. In a flux gradient (virtual particle river), a charged particle moves itself.

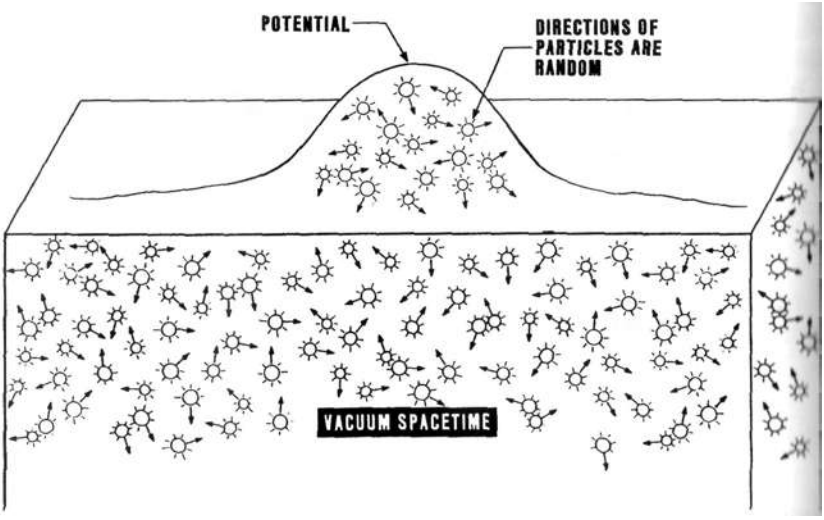
POTENTIAL
DIRECTIONS OF PARTICLES ARE RANDOM
VACUUM SPACETIME

Figure 49. A natural potential is a **disorganized change** in **the stress of** vacuum. It has a random virtual substructure.

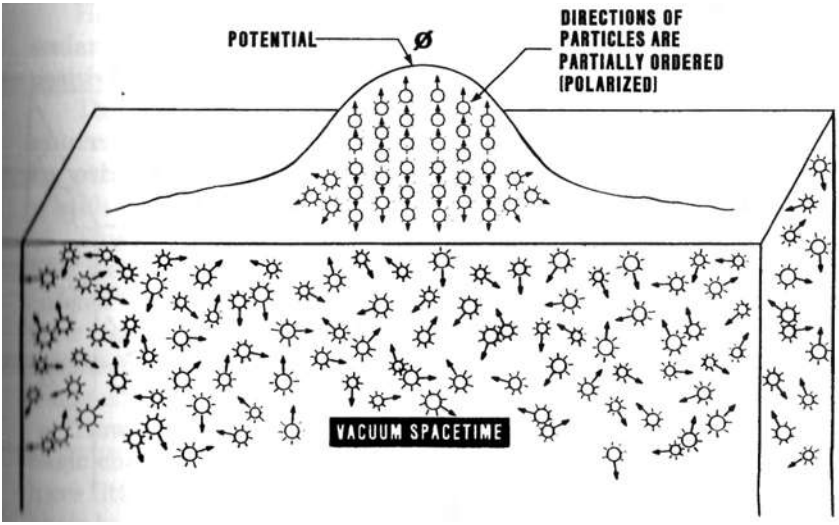
POTENTIAL
Ø
DIRECTIONS OF PARTICLES ARE PARTIALLY ORDERED (POLARIZED)
VACUUM SPACETIME

Figure 50. An artificial potential is an organized change in the stress of vacuum. It has a deterministic virtual substructure.

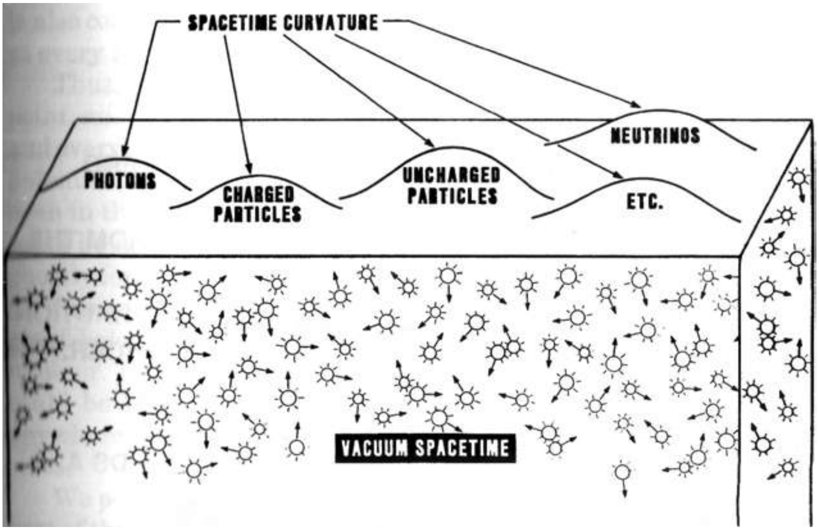
SPACETIME CURVATURE
PHOTONS
CHARGED PARTICLES
UNCHARGED PARTICLES
NEUTRINOS
ETC.
VACUUM SPACETIME

Figure 51. Gravitational potential is a conglomerate of stresses, organized or disorganized.

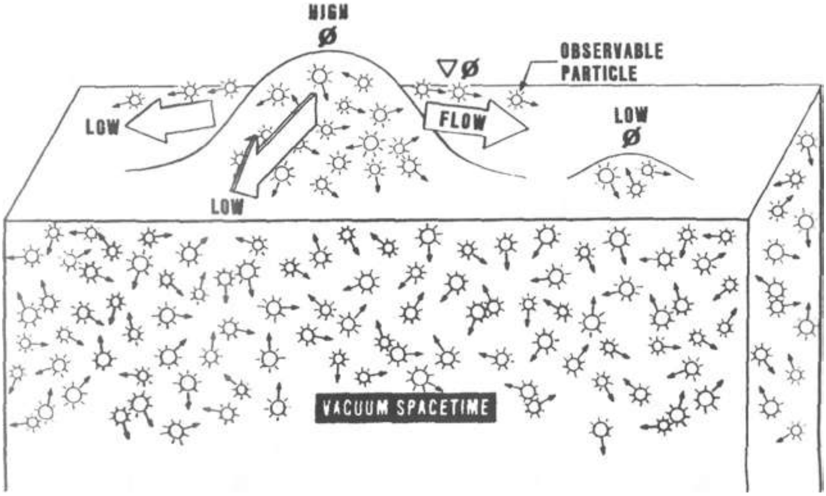

Figure 52. EM force fields are releases of gravitational potential via observable charged mass flows.

Table 33. RESIDUE UNIVERSE, RESIDUE SCIENCE

- **VECTORIALLY, STRESS IS SUM-ZEROED**
- **THE VECTOR FORCE FIELDS REPRESENT NON-ZERO GRADIENTS IN MORE FUNDAMENTAL STRESS**
- **FORCE FIELDS ARE THE RESIDUE LEFT OVER FROM THE SUM-ZEROED STRESSES**
- **PHOTON INTERACTIONS ARE FORCE-FIELD INTERACTIONS**
- **PHOTON-DETECTED REALITY IS ONLY "FIRST ORDER." IT SHOWS ONLY "RESIDUE-LEVEL" REALITY**
- **WE HAVE BUILT A "RESIDUE SCIENCE" AND A "RESIDUE PHILOSOPHY" BASED ON FORCE FIELDS AND PHOTON DETECTION**

Hence at every point in the vacuum, an intense flux of these scalar EM waves, with concomitant interferences, exists — in both positive and negative observer time.

This scalar EM flux causally **drives (constitutes) all the** enormous **quivering of the vacuum** spacetime medium **itself. In other words,** it **drives subquantum (virtual) change, which in turn drives quantum change.**

This is a drastic reorientation of quantum mechanics. In scalar EM, we view that the virtual flux of vacuum is *causally driven or created* by interacting scalar EM waves. This replaces the present **view** that the virtual flux is entirely chance. It opens up the direct engineering and structuring of the vacuum — and hence the atomic nucleus — by scalar EM means.

However, it does not replace the normal *statistical nature* of the **basic** changes of the background vacuum medium. We normally **have** little or no knowledge of the myriad of drivers that cause the basic background waves of that medium. We may, however, deliberately create special vacuum potential waves which are determinis**tic,** and which we have knowledge of.

To the observer, from a purely statistical viewpoint, in this violent vacuum flux *any* and *every* finite pattern of virtual particles is also continually being momentarily created and destroyed — and at every level and in every hyperdimension.

Thus, in the vacuum there continually exists — at any and every point, and in any and every region — the ghostly image of anything and everything, whether in the past, present, or future; and whether **potential,** probabilistic, or actual. Even everything that could have be**en** in the past but wasn't, or might be in the present but isn't, or could be in the future but won't be, is continually present in thin, ghostly form.

Rigorously, the universal vacuum may be taken to be a sort of gi**ant** hologram, for the whole is in each part, albeit in ghost-like manner. Everything that is, or was, or shall be; and everything that could be but isn't, wasn't, or shan't be; already exists *at once, an*ywhere and everywhere, in this ghostly, holographic, virtual state.

We point out in passing that, conceivably, one can engineer any **part** of this "potential for reality." That is, one can directly engineer physical reality itself, in the scalar EM view.

All that must be done is to amplify (continuously and coherently

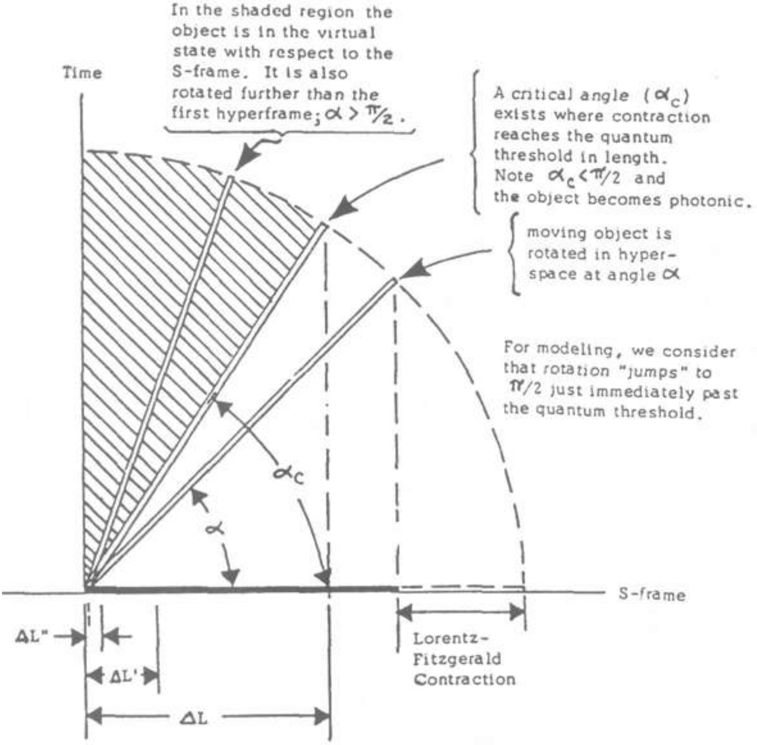
In the shaded region the object is in the virtual state with respect to the S-frame. It is also rotated further than the first hyperframe; $\alpha > \pi/2$.
Time
A critical angle (α_c) exists where contraction reaches the quantum threshold in length. Note $\alpha_c < \pi/2$ and the object becomes photonic.
moving object is rotated in hyper-space at angle α
For modeling, we consider that rotation "jumps" to $\pi/2$ just immediately past the quantum threshold.
α_c
α
S-frame
ΔL"
ΔL'
ΔL
Lorentz-Fitzgerald Contraction

Figure 53. As beyond (without), so within. Hyperspaces are internested virtual states, and rotated orthogonal frames. They are also substructures within and of vacuum state potentials.

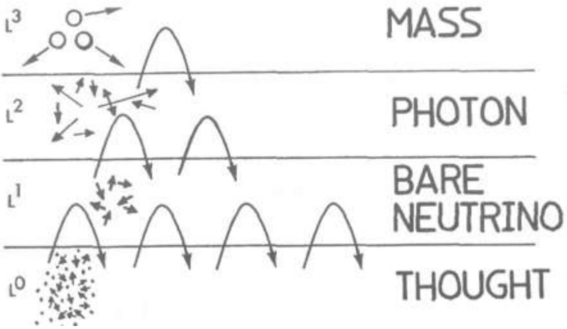
L^3
MASS
L^2
PHOTON
L^1
BARE NEUTRINO
L^0
THOUGHT

Figure 54. Internested levels of virtual state vacuum interact with neutrino, photon, and mass. "Thought" or "mental state" refers to the third and more levels down in virtual state.

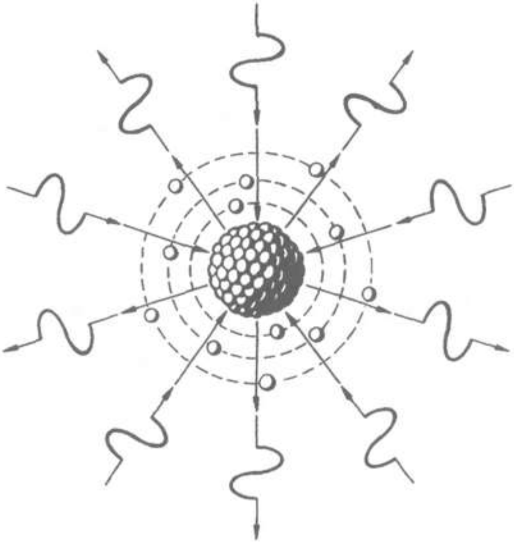

Figure 55. Virtual state patterns are absorbed and emitted by the atomic nucleus. The built-up nuclear "charge" in that particular pattern is a partial potential — for that pattern only.

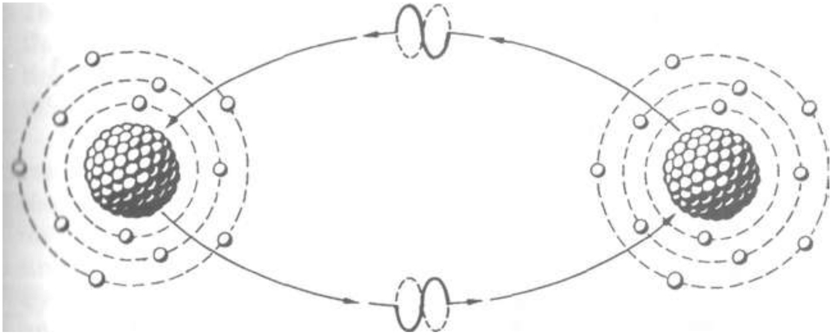

Figure 56. Nuclei of the universe continually exchange scalar EM waves and virtual **panicle** fluxes.

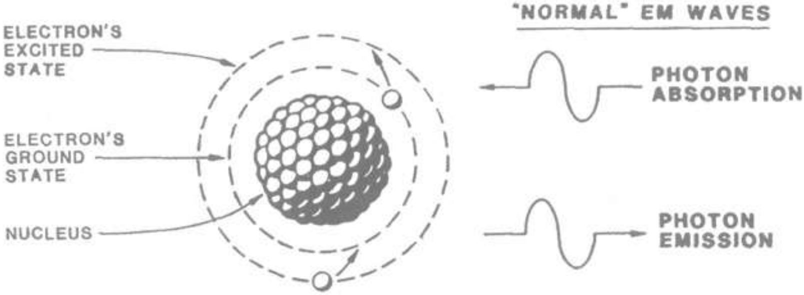
"NORMAL" EM WAVES
ELECTRON'S EXCITED STATE
ELECTRON'S GROUND STATE
NUCLEUS
PHOTON ABSORPTION
PHOTON EMISSION

- **HEAVY NUCLEUS RELATIVELY UNAFFECTED**
- **MACROSCOPICALLY. MASS APPEARS STABLE AND "BRICK-LIKE"**
- **SPECIAL RELATIVITY APPLIES**

Figure 57. First order reality. Rough-hewn by observable photon interaction with electron shells of the atoms of physical mass. This process is "physical observation."

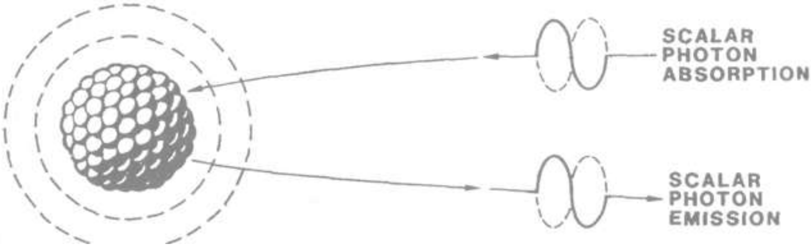
SCALAR PHOTON ABSORPTION
SCALAR PHOTON EMISSION

- **HEAVY NUCLEUS AFFECTED**
- **MACROSCOPICALLY. MASS APPEARS FLUID AND CHANGEABLE**
- **GENERAL RELATIVITY AND KALUZA-KLEIN GEOMETRY APPLY**

Figure 58. Higher order (hyperspatial) reality. It exists as stored in the potentials ot the atomic nuclei of physical mass.

add energy to) one of these "ghost images" virtual state forms. In that case, the form becomes "denser and denser," until eventually it starts to breach the quantum threshold. At that time, the form will be seen emerging as a sort of "coalescing form of light," where the light seems to be forming in little "chunks" or pieces. As the coalescing operation continues, the light-form will eventually become fully formed. If it continues, it will start to coalesce further into *thin, ghostly material* form, appearing much like a faint fog-form. With continued coalescing, it will become a very solid, material form which exerts pressure on tree limbs, breaks them, leaves traces, etc.

Long ago I dubbed this process *kindling.* It is rather widely met in paranormal phenomena of many types.

But back to the basic idea of ghostly images in the virtual states of vacuum.

In this view, Everett's many-worlds interpretion (MWI) of quantum mechanics is literal and real, albeit the additional worlds are to bo taken as virtual and hyperdimensional to the laboratory observer.

(It is rather simple to show that ever deeper, nested levels of the virtual state are exactly the same thing as ever higher dimensions. By "rotating" an object in many-dimensional space, one can show that, after three rotations, a physical object has lost all physical intersections with our normal 3-space, to us now only occupying time. Since this is the precise characteristic of a *thought form,* I used this phenomenon and spatiotemporal model to make a multidimensional space model directly including thought, mind and matter. That is, one is thus able to form a unified field theory of mind, thought, life, and matter — a model that is testable, makes predictions, and actually unites physics and metaphysics. That work and its plications, however, are largely beyond the scope of this book.)

It is stressed that the basic scalar EM image (interpretation) of reality is directly implied by quantum mechanics. It is not idle metaphysics. Other than the scalar EM wave causality for the observer's statistical microscopic dynamics (and the extended model unifying mind, thought, life, and matter), it is not the construct of this author. Instead, it follows from interpretation of the equations and axioms of quantum mechanics (QM) itself, as changed and extended by the "zero vector system" dynamics.

Charge, Potential, and Curvature of the Vacuum/Spacetime

The flux of one type of virtual particle through a point in the vacuum may be examined.

A single type of virtual particle flux (VPF) at a point constitutes one particular type of charge at that point. For example, the flux of virtual photons constitutes the electrical charge.

Further, the magnitude of the selected flux (usually expressed as a flux density) through the point may be taken as the magnitude of the vacuum's charge there in that type of particle. In other words, at a point the magnitude of the vacuum's charge (VPF) is one type of *potential* at that point.

Various kinds of fluxes constitute various kinds of charge, and comprise various kinds of potentials. Immediately we see that one type of charge may always be decomposed into other kinds of charge, as can potentials. This point is beyond the scope of this book and will not be further amplified.

Continuing our example, the magnitude of the electrostatic scalar potential at a point in vacuum represents the magnitude of the flux of virtual photons through the point, and hence the electrical charge at the point.

Table 34. REALITY IS A HOLOGRAM

TO ESTABLISH THIS:

1. **ALL DETECTION IS INTERNAL**
   PHYSICAL REALITY CONSISTS OF THE INTERNAL DETECTIONS OF THE PERCEIVER/OBSERVER.

2. **THE OBSERVER MAY BE AT ANY POINT IN THE UNIVERSE**
   THE UNIVERSE CAN BE DETECTED FROM ANY POINT INSIDE ITSELF.

3. **THE PHYSICAL UNIVERSE IS TOTALLY INSIDE EACH POINT IN ITSELF**
   THIS DEFINES A HOLOGRAM: THE WHOLE IS IN EACH PART.

Twin Flows of Time

In quantum mechanics, each virtual photon is continually turning into an electron/position pair, and vice versa.

By Dirac theory, *a positron is an electron traveling backwards in time.*

Further, pair production produces time-smeared particles — an electron and a positron.

Pair production, then, actually produces two *electrons,* one coupled to (smeared in) a positive piece of time and one coupled to (smeared in) a negative piece of time.

Thus in the vacuum two discretized streams of time, one positive and one negative, are continually created in conjunction with pair production, and destroyed by pair annihilation.

Further, integration of tiny virtual pieces of time (from virtual photons) to form "passage of time" macroscopically is directly associated with the charge (the absorption and emission of virtual particles) of an observable particle.

This is what is meant by an object "existing" (persisting). Its continual virtual photon interactions are integrated by its mass (timeless part) to continually create its energy states and its march through time.

The absorption and emission of an observable photon, however, moves the mass (timeless part) in (comparatively) larger jumps through time. The absorption of the photon connects a positive piece of time to the mass of the particle, converting it to masstime. The subsequent emission of an observable photon "tears off the little time-tail," so-to-speak, and leaves behind a totally spatial mass entity, having no large connection to "the flow of time."

This is what is really meant by "observation."

Because the "observation" disconnects time from the previously "excited mass" (masstime), then it follows that time is not an observable.

It also follows that the twin streams of time are not continuous. Each is broken into incredibly tiny bits.

The bits of time are continually interlaced between negative and positive, and neither of the two streams of time bits is integrated in the vacuum. The vacuum is disintegrated!

A photon — the basic quantum — is composed of discretized energy and time, bound together (integrated) as action (angular momentum). That is, a photon is a little piece of energy, welded to a little piece of time, with no seam in the middle.

Two kinds of photons exist: the normal photon exists as (+dE) (+dt). The phase conjugate photon exists as (-dE) (-dt).

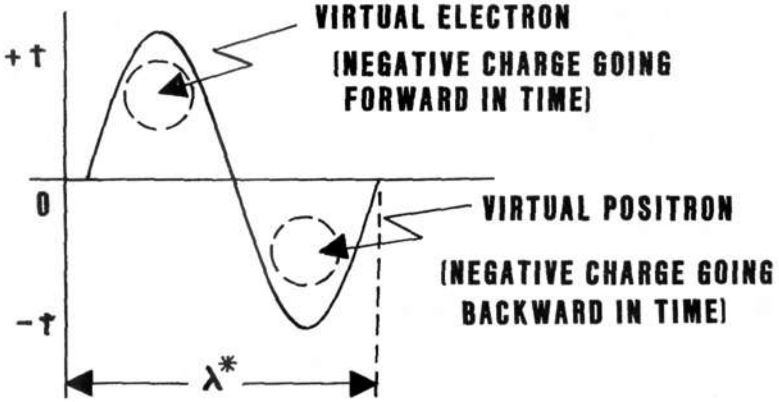
+t
0
-t
VIRTUAL ELECTRON
(NEGATIVE CHARGE GOING FORWARD IN TIME)
VIRTUAL POSITRON
(NEGATIVE CHARGE GOING BACKWARD IN TIME)
λ*

Figure 59. A photon is a single oscillation of a carrier wave. It is also two virtual electrons, one going forward in time, and one going backward in time. With internested modulations (subphotons), the structure of the photon becomes extremely complex.

Table 35. PHYSICAL CHANGE

- **COMPOSED OF ACTION**
- **h = BASIC QUANTUM**
- **TWO CANONICAL VARIABLES**
    - **ENERGY/TIME** $(\Delta E \Delta t)$
    - **MOMENTUM/LENGTH** $(\Delta p \Delta L)$

    ETC.
- **ENERGY/TIME**
    - $h = (+\Delta E)\ (+\Delta t)$
    - $h = (-\Delta E)\ (-\Delta t)$

If the photon is of a certain magnitude, given by Planck's constant, then it is an **observable photon.**

If the photon is less than h in magnitude, then it is a **virtual photon.**

A simple photon in vacuum can also be considered to be one cycle of a sine wave. One half of the sine waves exists in positive time, hence is/contains a virtual electron.

The other half of the sine wave exists in negative time, hence it is/contains a virtual positron — an electron traveling backward in **time.** This second half of the full photon (a single sine wave oscillation) is the phase conjugate (time reversal) of the first half.

As we stated previously, the full photon in vacuum actually exists as a presently visualized "positive time" photon and a phase conjugate "negative time" photon , a "serially linked pair," where **each** photon is only one-half the sine wave. This resolves the conflict inherent in wave/particle duality.

It also has a great deal to say about engineering the vacuum and about detectors and the detection process.

In the spontaneous decay of a virtual photon, it can be seen that one part (the virtual electron) is associated with positive energy and positive time; the other part (the virtual positron) is associated with negative energy and negative time.

The continual "switching" between positive and negative time in an observable particle's absorption and emission of virtual photons **(and** the associated electron/position +t/-t pairs) means that the positive flow of time, to the macroscopic observer, is continually being "started" and "stopped" (created and destroyed).

This continual attachment **and detachment of positive time to and from the "observer particle" creates the "past" and the "future" to which the observer seems somehow to be connected, and yet not connected.**

Engineering Local General Relativity

In a flat spacetime (linear vacuum), the two time flows are balanced, as are pair creation and pair annihilation. In this case, conservation of charge and energy hold.

Table 36. ORDINARY GENERAL RELATIVITY (OGR) IS A SPECIAL CASE

- **OGR HAS ASSUMED AWAY MOST LOCAL GR SYSTEMS, OF THE TYPE WHERE MAXWELL'S EQUATIONS AND THE CONSERVATION LAWS CHANGE LOCALLY.**

- **PHYSICISTS ARE THUS STRONGLY DISCOURAGED FROM CONSIDERING THAT GR CAN EASILY BE ENGINEERED IN A LOCAL EM SYSTEM.**

In a curved spacetime (nonlinear vacuum), one of the two triplets (positive electron/energy/time or negative electron/energy/time) predominates. Hence in a curved spacetime the rate of flow (production) of the two time streams is unequal, as is the production of the two kinds of energy and the two kinds of electrical charge.

A locally curved spacetime thus produces a predominance of either positive or negative electrons, and either positive or negative energy, and a predominance of the addition of either positive or negative time production.

The curved spacetime acts as a local source or sink accordingly, and both conservation of energy and conservation of charge can be locally violated in such a curved spacetime region.

As we stated, the locally curved spacetime produces either positive time or negative time more predominantly, acting as a source or sink for normal (positive) observer time. The local rate of flow of the observer's time can thus be speeded up or slowed down, depending upon the type of local spacetime curvature utilized.

It follows, then, that standing local gravitational waves can result in localized, stable gradients in the vacuum virtual particle flux, between the "high potential" part of the standing wave and the "low potential" part.

This allows the violation of conservation laws such as charge and energy in the local gravitational gradient, utilizing the localized curvature of vacuum to provide a special source.

For proof that conservation laws need not apply to a curved spacetime, see A.A.Vlasov and V.I. Denisov, "Einstein's formula for gravitational radiation is not a consequence of the general theory of relativity, **Theoretical and Mathematical Physics,** 53(3), June 1983 **(English translation);** Russian, Dec. 1982, p. 406-418. Quoting:

"This result is a particular consequence of the general assertion to the effect that in general relativity there are no energy-momentum conservation laws for a system consisting of matter and the gravitational field."

See also V.I. Denisov and A.A. Logunov, "New theory of spacetime and gravitation," **Theoretical and Mathematical Physics, 50**(1), July 1982, p. 3-76. This paper (p. 3) points out that:

"... **the gravitational field in general relativity is completely different from other physical fields and is not a field in the spirit of Faraday and Maxwell."**

A 1984 Soviet paper by senior Russian physicist C. Yu. Bosgoslovsky, "Generalization of Einstein's relativity theory for the anisotropic spacetime," is also very relevant.*

In composing his theory of general relativity, Einstein assumed that the local spacetime of the observer could never be curved, and instead would always be represented by a Lorentz frame (flat spacetime).

This severe assumption had the effect of "saving" the sacrosanct conservation laws, and maintaining the exclusion of electrogravitation.

Since then, Western physicists have raised Einstein's assumption — and the conservation laws — to a dogmatic belief system — in some cases, to near fanaticism.

To see a clear statement of these limiting assumptions imposed on general relativity to provide the severely restricted version taught in almost all Western universities, see Charles W. Misner, Kip S. Thome and John Archibald Wheeler, **Gravitation,** W.H. Freeman and Co., San Francisco, California, 1973, p. 19-21, 71-72, 367-369.

In other words, Western physicists still use a sort of "special

---

**See* also V.I. Denisov and A.A. Logunov. "The inertial mass defined in general relativity has no physical meaning," preprint p. 0214, Institute of Nuclear Research, USSR Academy of Science, Moscow, 1981.

relativity with distant curved spacetime pertubations," and call it *general* relativity!

The Soviets do not limit themselves to the same restricted general relativity theory as does the West.

Santilli (Ruggero Maria Santilli, **Ethical probe on Einstein's followers in the U.S.A.: An insider's view,** Alpha Publishing, Newtonville, Massachusetts) has documented what essentially amounts to a conspiracy in Western physics to shut out any generalization of Einstein's severely limited relativity.

It is precisely this near-total Western scientific prejudice — along with Heaviside's mutilation of Maxwell's original theory — that has excluded electrogravitation from being developed in the West.

Note also that, by embracing a "locally flat spacetime" assumed to be decreed by the "laws of the universe," in their rigid mindset Western physicists have excluded any possibility of performing simple laboratory experiments in locally curved spacetime.

They do not believe it can be done, so they have never tried.

But let us further address the importance of the negative triplet (electrical charge/energy/time) and its concomitants parity, gravity, and entropy:

Engineering Antigravity

In negative time flow, gravity is a repulsion, not an attraction.

Producing antigravity in an object is simply the production of excess negative time flow in that object.

That is readily achieved by curving local spacetime in the appropriate direction, and forcing the object to produce excess negative energy.

Mass, for example, is just a special form of trapped energy—and trapped energy (potential) is just trapped charge of one kind or another, whether observable or virtual.*

As such, mass represents a "localized potential" in the ambient vacuum, of a certain kind of "trapped energy/time charge."

*Actually, since it is in equilibrium in a dynamic and continual flux, a mass particle is actually a special trapped **rate** of action flow. A defining equation for mass in terms of the rate of action flow was given by the author in **Quiton/Perceptron Physics,** NTIS, 1973. Report AD-763210.

The mass/potential is continually being charged by the vacuum flux, and discharging and emitting charge flux into the vacuum flux. The mass is "in equilibrium" at its fixed potential.

Normal ("positive") mass is simply charged up with positive energy and positive time. And its external electron shells result in it presenting a "negative charge" face to the world.

If we charge up the mass, however, with negative energy and negative time, so that some of its charge potential now is of this second type, the trapped energy will now consist partially of negative energy. In addition, the mass will partially exist in negative time (be charged up partially with negative time).

When excess negative time flow is produced in a mass, the mass begins to be repelled by the earth as well as still being attracted to it in its positive time stream. The net result of the object's fixed attraction and increasing repulsion is that the object begins to lose weight in the earth's gravitational field and gets lighter.

When a sufficient rate of excess negative time flow (sufficient local spacetime curvature) is produced, the mass floats freely.

With additional negative time production rate, it accelerates ("falls") upward.

The rate of excess negative time produced may be controlled very simply by using a "negative time/negative energy" generator and varying the amount of the load applied to the device.

The negative time/negative energy generator can be rigidly affixed to a flying apparatus, serving as a sort of gravitational rocket or gravitational propulsion system.

Of course, the negative energy produced in the propulsion process can be used to light lamps and power electrical loads, as well as host of other things. (Yes, Virginia, it can also be used to generate force fields, power weapon systems, etc.)

Thus one can "eat his cake and have it too"; by using negative energy flow for power, a very practical antigravity spaceship can be developed.

An additional side benefit to the passengers probably exists: they should age at a slower rate when traveling in the vehicle.

As can be seen, since many kinds of VPFs exist in vacuum, the vacuum is filled with (and consists of) many types of "potentials." Overall, this matches very closely the view of general relativity, if we regard the vacuum itself as indefinite "spacetime", where no ordered universal frame with a fixed metric has yet been associated or prescribed.

Curved Spacetime and the Disintegrated Vacuum

From general relativity, additional characteristics of the modern vacuum/spacetime may be inferred or interpreted.

First, we may take a "flat spacetime" to consist of a charged vacuum, such that both the magnitude and structure of the VPF are essentially unchanged from point to point, and the flux of virtual particles and virtual antiparticles are balanced. Equal positive and negative time flows/charges exist.

Note, however, that we can have what casually may appear to be a "flat spacetime" but instead is "warped." This case exists when the magnitude of the overall VPF comprising the vacuum does not change from point to point, but the component structure of the VPF does vary from point to point.

In other words, the overall vacuum potential remains the same from point to point, but two or more individual potentials may canonically vary. Let us further clarify this point by analogy.

Suppose we view the vacuum/ether/spacetime as a special sort of gas, where the "gas" is actually a mixture of many, many gases.

It is a rather peculiar sort of gas; it exists in five or more dimensions and is composed of very strange, fleeting particles which . spontaneously are created and destroyed, each arising out of nowhere and returning to nowhere almost immediately.

There is only minuscule (negligible) time overlap (integration) of the existence of these virtual particles in the main.

That is, mostly each virtual particle — along with its associated energy, movement, momentum, charge, and increment of time — exists almost entirely as an individual.

The energies, movements, momenta, charges, and time increments of the individual virtual particles do not integrate or sum to any appreciable degree in vacuum.

Thus, although casually this ether/gas may seem to contain enormous "energy density", momentum density, charge density and time increment density, these are not *integrated* energy density, momentum density, and time density in the normal sense.

Instead, they are the densities of *disintegrated* energy, *disintegrated* momentum, and *disintegrated* time flow.

In scalar electromagnetics, we have called the disintegrated energy of vacuum **anenergy.** This clearly distinguishes that it exists in a fashion altogether different from normal, *integrated* energy with which we are accustomed.

By integrating anenergy, of course, one obtains ordinary energy—but it first requires an integrating agent. The most **usual integrating agent/operation is the spin of an observable particle of** mass.

Similar considerations apply to momentum and time.

With this more **precise interpretation, the vacuum contains enormous anenergy, but essentially no energy. It contains enormous virtual energy, but no observable energy.**

Unfortunately, however, orthodox physics does not yet clearly differentiate between the two states of energy, virtual and observable (disintegrated and integrated).

Instead, it continues to loosely utilize the term "energy" for both states, whether integrated or not. Thus it exclusively uses the term **"energy** density of vacuum." Rigorously, the term should be "anenergy density of vacuum."

Indeed, most physicists are not clearly aware of the distinct difference between the two different states of energy, insofar as the implication of integration or nonintegration.

Regrettably, then, we will continue to utilize the orthodox term, "energy density of vacuum." but the reader should be clearly aware that the vacuum "energy" and vacuum "stress energy density" as **such** are totally disintegrated.

Separation of Vacuum and Observable States

Because the energy density of vacuum is unintegrated, an integrated physical body — including one's personal body — can **exist** "in" the vacuum medium of incredible "energy" density, yet observably exist in a medium of zero energy density.

The unintegrated vacuum ether has enormous spatial density of *virtual* energy, but essentially has zero spatial density *of observable* energy.

Note, however, that in our hyperdimensional interpretation of the levels of virtual state, the vacuum spatially does not exist in the normal 3-space of observable matter.*

The vacuum occupies the 4th dimension (time) and all spatial dimensions (hyperdimensions) greater than the third — it occupies all levels of virtual state, but not observable state. It occupies all levels of disintegration, but not the levels of observable integration.

Exchanges occur between normal 3-space and the hyperspatial vacuum, through the common connecting dimension, time.

These exchanges (of virtual particles) are integrated by the spinning observable mass particle into the first quantally excited state — a delta (change) representing a single entrapped quantum — of the 5-space potential represented by the particle and its accumulating VPF. At that time a sharp electromagnetic (fifth dimensional) discharge of the quantum occurs, converting the accumulated 5-G quantum to the ordinary EM quantum of the 4th (time) and 5th (infolded space) dimensions.

This accounts for the so-called "collapse of the wave function", the production of unitary quantum change, and the discretizing of observable change.

We note in passing that time is the only "dimension" where everything can be considered to occupy the same "point" or "interval", but it is discretized and chopped into pieces (disintegrated), quite different in nature from the familiar three spatial dimensions of the observable (integrated) universe.

The dimensional separation (spatially) of the seething vacuum "energy" (anenergy) and the 3-space of normal matter allows a physical body to enormously interact with the vacuum at or near equilibrium at the microscopic level, but maintain its macroscopic form and stability as if no interaction were going on.

Partial Potentials, Curvature, and Warping

At any rate, we view the magnitude of the overall "potential" of this vacuum gas mixture, as the magnitude of the entire **"stress**

---

*Here we sharply differ from conventional physics, which hopefully confuses the concepts of vacuum, spacetime (indeterminate), spacetime (determinate), and frame.

ENERGETICS OF A POTENTIAL
Running in Place and Getting Nowhere

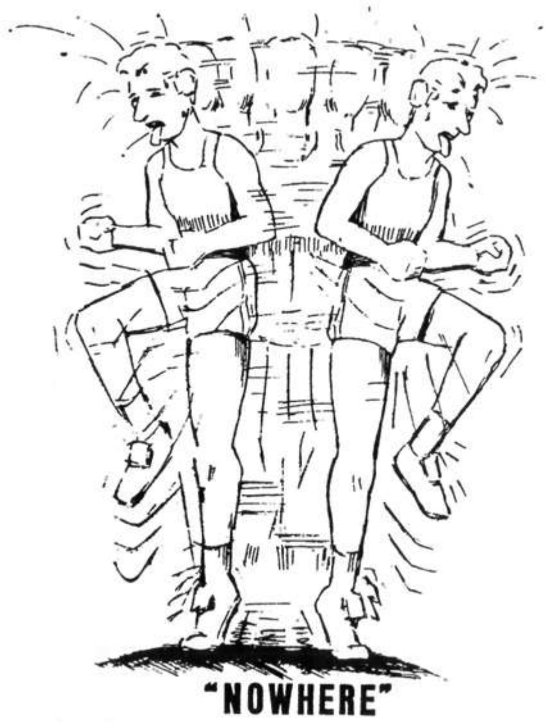

Figure 60. A potential is trapped energy. A potential in vacuum is trapped virtual-state "energy" or flux. A natural potential is trapped random energy. An artificial potential is trapped organized energy.

energy" or "pressure" of the mixture — IF AND ONLY IF this stress energy or pressure were somehow integrated.*

As can be seen, we can now look at each of the component gases in the mix as contributing a partial pressure or partial stress energy (an individual component potential) of the vacuum conglomerate.

We may have the overall pressure remain the same, yet canoni-

***Thus it's** really the **potential** stress energy density. The vacuum is nothing but potential!

cally change the partial stress energies (individual potentials).

In that case one has an internally twisted or "warped" spacetime.

A warped spacetime may prove quite useful, for example, for direct translation of one type of "energy" into another. In that fashion, one may conceivably obtain radiation energy at the expense of mass, or mass at the expense of radiation energy.

Or one may convert between potentials where one or more of the converted potentials are as yet unknown to orthodox physics.

Kaluza theory, for example, contains many potentials that are as yet unknown and undiscovered in nature. When undue simplifications are made to arbitrarily exclude the unknown potentials, the adapted theory leads to error in its predictions.

This was originally regarded as a great imperfection, and Kaluza theory fell into disfavor for decades. Then as modern particle physicists began to produce fundamental particles of an astounding variety, a Kaluza-Klein theory of 11 or more dimensions was found to very reasonably describe the emerging situation. Once again, then, Kaluza-Klein theory became a moving force among modern theorists.

One may also convert electromagnetic field energy into gravitational field energy, for example, and obtain enormous amplification of gravitational effects.

This is precisely what scalar electromagnetics seeks to do.

Since the electrical field between two electrons is about 10^{42} times as strong as the gravitational force between them, conversion of the EM charge flux into gravitational charge can conceivably yield up to 10^{42} times as much gravitational force as normal, depending upon the efficiency of the conversion process. Even with a process or device that uses electron currents and has an efficiency of only 10^{-22}, gravitational/inertial gains of 10^{20} are possible.

Note, however, that the levels of vacuum may also be compared to horizontal planes beneath the ocean.

For example, we may take a particular level of vacuum potential, and that level constitutes a sort of included "flat spacetime" of its own. By biasing the potential of a scalar transmitter and receiver to a given level, one may transmit (in either direction) at that level only, Receivers at higher or lower levels will detect nothing.

The immediate application to engineering the nucleus for such things as nonradioactive transmutation of elements, controlled nuclear fusion, electromagnetic healing, and processing nuclear wastes to render them harmless should be apparent.

To recapitulate our analogy, a true "flat spacetime" is a vacuum whose VPF does not essentially vary overall or in each constituent part, between two comparative points.

A "curved spacetime" is a vacuum whose overall VPF varies in overall magnitude, between two comparative points.

A "warped spacetime" is a vacuum whose VPF may or may not **vary** in overall magnitude between two comparative points, but whose partial VPFs vary their ratios, between the same two points. Note that a vacuum/spacetime can be both warped and curved.

It is stressed that we have not limited the number of dimensions **that** can be placed on the vacuum to Minkowski's four. We hold the number of dimensions necessary to represent both vacuum and observable matter to be arbitrary, but greater than four.

To unify electromagnetism and gravitation (which is what scalar electromagnetics is about), we must take five dimensions — four spatial and one time — to be the minimum necessary in modeling the unified theory, similar to Kaluza.

We state now a major characteristic: **Vacuum/spacetime consists entirely of potentials, nothing else, and these are hyperspatial a priori.**

The vacuum does not contain observable force fields nor any **Other** *non-potential* entity.

We also state without proof that in scalar EM theory it seems **very** possible to model the mind and its unification with physical **reality,** in a fashion subject to experimental verification.

Force and Mass Are Inextricably Intertwined

In foundations of physics, it is well-known that, paradoxically, force and mass can only be defined one in terms of the other.

We may use that paradox for a new interpretation of force and **force** fields: a force may be interpreted **to consist of** the time rate of changing the "mv" momentum of a mass system or object, instead of **being the cause of** its dp/dt.

That is, we may take the view that

$$\mathbf{F} \equiv \mathbf{dp}/\mathrm{dt} \qquad (4\text{-}14)$$

not just that

$$\mathbf{F} = \mathbf{dp}/\mathrm{dt} \qquad (4\text{-}15)$$

We have previously pointed this out and called it the "strong definition of force."

This immediately suggests the question: If force is an effect and not a cause, what, then, causes force?

Quantum mechanics already provides the answer: A differentiating operator must be applied to (coupled to) potential in order to produce a force field. The most usual "differentiating operator" is an observable spinning charged particle.

A charged particle exposed to the variation of electrostatic scalar potential between two points is exposed to a gradient (flow) in the virtual photon VPF, and hence to a directional "stream" of virtual photons. The observable charged particle couples to that stream. **Its** spin integrates the unobservable, disintegrated anenergy of the gradient flux into integrated, observable energy of the particle.

With greater integrated "pressure" on one side of the particle than on the other, the observable particle accelerates and translates This produces or "detects" (and comprises) an electric force between the two points.

Force is the mass particle(s) accelerating, either in time or in mass. Observable force is observable particle(s) accelerating. **Virtual** force is virtual particle(s) accelerating.

In **Quiton/Perceptron Physics,** I gave a new definition of mass that is consistent with this viewpoint:

$$1 \text{ kg. mass} = 17.053 \times 10^{50} \text{ switches/sec} \qquad (4\text{-}16)$$

where a switch is h/4*pi), having the units of action/time, or energy. This directly defines mass in terms of energy accumulation, where the energy accumulated is in equilibrium in a differentiated action flow.

Note that "energy" is the time rate of flow of action. Energy can only be accumulated (as a mass or a potential) by possessing two rates of flow of action simultaneously, one in positive time and one in negative time. This is a very deep statement that affects all of physics, but unfortunately it is beyond the scope of this present book.

To return:

The observable electric force field does not exist as such in vacuum. The vacuum and all its constituents are totally virtual. Only the potential ("potential for the electric force field, given a coupled charged observable particle") exists in vacuum, as pointed out by Richard P. Feynman et al, The Feynman Lectures on Physics, Addison-Wesley, Reading, Mass., 1964 vol. 2, p. 2.

The Concepts of Zipped and Unzipped Forces

As noted previously, the potential exists as an unintegrated virtual particle flux (VPF). Further, each little moving virtual particle is smeared in length and time, is coupled to finer VPF, and thus may be said to constitute a minute "force vector" while it exists.

Thus, the electrostatic scalar potential is composed of a myriad of little individual "virtual particle force field" vectors.

Since these little virtual vectors are essentially separate and unintegrated, one may say that the electric field is *unzipped* in vacuum. By "unzipped" we mean that the individual virtual particle E-field vectors are unintegrated.

A potential, for example, is composed of unzipped vector fields since it is comprised of unintegrated, virtual vectors.

To be static, the potential must be comprised of a pair (at least one pair) of unzipped vector fields, one in positive time and one in negative time. These unzipped vector fields must in turn be locked together. In this case, the positive time observer sees nothing (vacuum only), until he "observes" or detects. Normal detectors detect (and in so doing, they zip together) only the positive time unzipped component. The negative time (phase conjugate) vector held retraces the positive time field, in Newton's third law. Or, if generating a wave or pulse, it would be seen as a phase conjugate replica which affected the disturbing source.

Let us be very plain. **A "scalar EM potential" is actually two vector fields zipped together: one is in positive time, and the other is in negative time.** Each half is composed of microvectors. A dynamic gradient in this scalar EM potential (i.e., a scalar wave, for example) actually consists of two waves: one is the normal" positive energy/positive time" EM wave, while the other is the phase conjugate replica. When we detect the normal half (with electron detectors), we actually detect the second half with the atomic nuclei. Since it is a phase conjugate, it produces a negative momentum change on the nucleus {since one has p/(-t)} instead of LeSage's positive momentum change. In other words, it produces a drawing (attracting) force on the absorbing nucleus, instead of LeSage's assumed repelling force on it.

That is the entire secret of gravitation. And of electrogravitation. Knowing that, you can easily work out how to apply the necessary engineering to get antigravity.

But to return.

The potential can also be comprised of many unzipped vector fields at once. These may be deterministic (artificial potential) or randomized (natural potential).

The "force field" is disintegrated (unzipped) in vacuum, and integrated (zipped) on, of, and by an observable charged particle. In vacuum, it is always accompanied by its phase conjugate twin. When detection by an observable particle occurs, the spin of the particle acts as the "zipper," producing an observable force (which is actually of the particle accelerating, not "acting independently upon it.")

Vacuum is unzipped; observable mass is zipped.

Any zipped entity exists observably in 3-space; any unzipped entity exists in hyperspace outside 3-space, and is unobservable and virtual to the 3-space observer.

This interpretation is still consistent with the extended quantum mechanical view, where ultimately all observable forces are considered to be generated by absorption and emission of virtual particles by/from an observable particle (mass), at least insofar as what we may observably measure or detect.

Force, Force Field, and EM Waves in Vacuum

In the new interpretation, consistent with the quantum mechanical view, observable force becomes an effect, not a cause.

One may interpret quantum mechanics as stating that **any observable is an effect, since it is assumed to have a virtual state cause.** In the axioms of quantum mechanics, an operator is required for every observable.

The primary **causative agents are potentials, not force fields, as** is **now** firmly **established in** quantum mechanics.

The Aharonov-Bohm (AB) effect is a striking illustration of the primary reality of the potential rather than of the force field.

In the AB effect, potentials may still exist when the force fields **are** zero, and their interference can still cause real effects. The AB effect has finally become generally accepted by physicists after nearly three decades of controversy.

The immediate result is that, contrary to the outdated view imbedded in classical EM, observable E and B force fields do not exist *as such* in vacuum. This allows a totally new interpretation of the **nature** of the EM wave in vacuum: It is longitudinal. And actually, it exists in longitudinal pairs: the second is the phase conjugate of the first.

But of course we detect transverse waves with our detectors! However, this has a simple explanation.

Examine, for example, a straight-wire antenna. The spinning conduction electrons are longitudinally constrained, as is well known. The electrons are relatively unrestrained transversely in the **wire,** but can only "drift" longitudinally down the wire with a speed usually on the order of centimeters per second. The signal potential, however, moves longitudinally down the wire at almost the speed of **light.**

Obviously it is not the longitudinal movement of conduction electrons that constitutes the "signal", even though classical electromagnetics represents it that way.*

We note that a spinning electron whose movement is longitudinally restrained, but is unrestrained transversely, may be characterized as a gyroscope. Further, the axis of this gyroscope may be taken as lying in a plane perpendicular to the straight wire conduction path, since the electron is relatively unrestrained transversely.

---

*Actually, we do not really know how current/electrons travel down a wire, or even if it does/ they do. See three articles by Chappell Brown in Electronic Engineering Times: "Railgun Research Shoots Holes in Lorentz's Theory," April 6, 1987,p. 49-50; "Anomalies in Electromagnetic Law Spur Debate," Sept. 14, 1987, p. 58; "Electrons and Conduction: Not So Simple After All," Dec. 28, 1987, p. 21-22.

If now disturbed by a vacuum potential gradient, the gyroscopic electron is free to precess at right angles to the length of the wire. In other words, given a disturbing force, the electron will move transversely, or radially in the cross sectional plane of the wire, because of gyroscopic precession.

Thus, assuming a gyroscopic electron and its longitudinal constraint, we will detect transverse oscillations in the electron gas in the wire, IF AND ONLY IF the incoming disturbance in the vacuum is longitudinal, and IF AND ONLY IF the unzipped vacuum disturbance "couples" to (is zipped by) the gyroscopic conduction electrons.

Two things are certain in our electron-detector model: First, by the nature of a gyroscope, the conduction electrons move (precess) at right angles to the "disturbing force." Second, we only detect the movement of the electrons themselves.

Since we do detect transverse oscillations of the precessing electrons in the straight wire antenna, then it follows that normal electromagnetic waves in vacuum—that is, the component half that, interacts with the atom's electron shells and not with its nucleus — are (1) longitudinal, and (2) of such form that they couple to electrons.

Note that Tesla always insisted that electromagnetic waves in vacuum were longitudinal. For example, he called them "electromagnetic sound waves" in the ether. In "Tesla Maps Our Electrical Future," H. Winfield Secor presents Tesla's view of transverse waves in this fashion:

"Tesla upholds the startling theory formulated by him long ago, that the radio transmitters as now used, do not emit Hertz waves, as commonly believed, but waves of sound... He says that a Hertz wave would only be possible in a solid ether, but he has demonstrated already in 1897 that the ether is a gas, which can only transmit waves of sound; that is such as are propagated by alternate compressions and rarefactions of the medium in which transverse waves ***are*** *absolutely impossible."*

In his "The True Wireless," **Electrical Experimenter,** May **1919,** Tesla himself further stated:

> *"The Hertz wave theory of wireless transmission may be kept up* **for** *a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history."*

It should also be pointed out that our analogy permits the option of making a longitudinal wave in the vacuum in such a fashion that It will not couple to electrons.

In that case a normal"electron gas" conduction detector will not detect the new wave. Such a noncoupling EM longitudinal wave is **referred** to as a **scalar electromagnetic wave,** where by use of the **term** "scalar" we call attention to the fact that the vacuum potential **is** varying, but electron coupling to form vector force fields in an **electron** gas does not occur.

Since according to general relativity the variation of a potential in vacuum rigorously is a gravitational wave, then the scalar EM wave is a gravitational wave.*

We often say it is an *electrogravitational* wave to call attention **to** its electromagnetic origin, and to the absence of EM force field "bleeding away" of the 5-space Kaluza-Klein gravitational potential. We will discuss the 5-space G-potential shortly.

The conventional equation

$$\mathbf{E} = -\nabla \emptyset \qquad (4\text{-}17)$$

thus applies in the detecting electron gas, but not in the vacuum. Instead, in the nonmaterial vacuum, rigorously

$$\nabla \emptyset \neq -\mathbf{E} \qquad (4\text{-}18)$$

and this states that a gradient in a vacuum electrostatic scalar potential produces an antiphased electric force on and of a particle, on condition that a particle is present. The intent of equation (4-17) should be stated as

---

*It is also a very powerful gravitational wave, since for electrons it is some 10^{42} times as strong as "mass attraction" gravity waves.

$$\nabla \varnothing = -\mathbf{E} \qquad (4\text{-}19)$$

where the vertical line means "on condition that", and cp means "charged particle" (i.e., "that a charged particle is there.).

Equation (4-19) clearly shows that, in the vacuum, only the organized potential for an E field exists due to the quantity $\nabla \varnothing$ essentially as pointed out less strongly by Feynman et al in **The Feynman Lectures on Physics.**

Certainly, in the quantum mechanical interpretation, it is the potential that is real. The force field can only be made from the potential by a differentiating operation.

Since there is nothing in vacuum to perform this differentiating operation observably, one may interpret quantum mechanics as already prohibiting the existence of observable electromagnetic force fields in vacuum, at least in the form prescribed in classical electromagnetics.

Figure 61. A force-field vector is a movement of mass in one direction.

One may regard quantum mechanics — and physics in general - to be incomplete until electromagnetic theory is changed to incorporate the implications of quantum mechanics.

In scalar electromagnetics a new interpretation is taken, and the "overhaul" of electromagnetic theory is in progress.

We accent that scalar electromagnetics is still embryonic. It is at a stage similar to that where electromagnetism was when experimenters were still rubbing glass rods with cat fur and experimenting with little charged pith balls.

Early experiments are enlightening and encouraging; nonetheless, the phenomenology and the myriad of variables are not at all well-understood, even by the experimenters themselves, and certainly not by this author.

Zero-Summed Vector Systems: Shortcoming in Vector Analysis

It has long been pointed out by this author that a fundamental shortcoming exists in classical vector analysis itself when applied to physical systems. Let us briefly look at the construction of the vector mathematics theory, to clarify this shortcoming.

First, an abstract entity called a "vector" is conceptualized. This vector has both direction and magnizude, as contrasted to a "scalar" quantity, which only has magnitude.

Next, an abstract "vector space" is defined; this is a sort of "container" for mathematical objects called vectors, with certain inherent properties of the system specified. In other words, a vector space is an abstract "system" or "space" which can contain objects called "vectors."

Next certain other properties, operations, and entities are defined, one of which is the is the "zero vector".

Essentially the zero vector is defined by a set of operations; however, one simplified interpretation is "the absence of any vector of finite magnitude." By implication, in the abstract vector theory, all zero-vectors are defined to be identical.*

---

* Because, by their nature assumed in the concept, they have no internal structures. That is, effectively the zero vector has been too restrictively defined as "the absence of any and all finite vectors." It should be redefined as "the absence of any single vector being present alone." In that case the definition would admit "the presence of multiple translation vectors, where their combined translation action is zero."

In the vector addition operation, a group of finite vectors adds to a single vector called the "resultant."

It is standard practice in abstract vector analysis to replace a system of summation vectors by their resultant, since all appropriate properties of the summation system are retained by the vector resultant, in the abstract mathematical model.

This, however, is only possible because the abstract vector space has no such thing as stress, and a zero-vector has no substructure by definition. In the abstract space, only translation and rotation are possible, *by assumption*

It is also possible for two or more finite vectors to add to a zero resultant, in which case the system is replaced with a zero vector. Note that all further action by the summing vector components is eliminated by this assumption.

Again the appropriate properties of the summation system are retained by the zero resultant, in the abstract mathematical model, only because the *medium* (the abstract vector space) is not subject to stress, translation, rotation, and structuring.

Note however, that in physical systems, one cannot simply

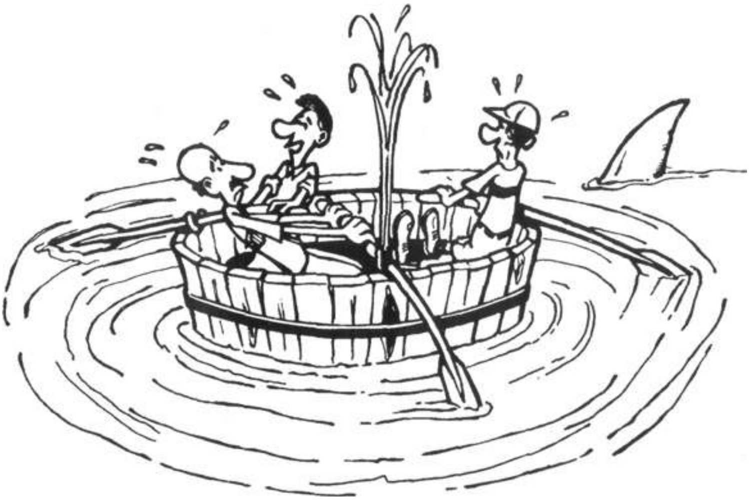

Figure 62. The zero-summed vector system is an engine and an artificial potential.

replace a zero summation system of force vectors with a zero vector and retain all characteristic physical effects caused by the replaced system. The physical medium - including the virtual flux of the vacuum, *is* subject to stress, translation, rotation and structuring.

Physically, the presence of the summing vector component forces creates a stress and a dynamic macroscopic structure in the physical medium in and on which the forces are acting. Therefore the stress of the medium and its structure must be preserved and accounted for, even though a zero vector resultant force exists.

This is particularly true in the vacuum, where internal (infolded) stresses cause charge and energy conversion between one kind of charge/potential and another.

Maxwell's Theory Was Altered and Curtailed

In fact, Maxwell seems to have been well aware of the electromagnetic stress of the ether and its importance. After all, he believed the ether to be material and mechanical, and developed his theory accordingly.

For example, in his Treatise on Electricity and Magnetism, he wrote:

> *"There are physical quantities of another kind which are related to directions in space, but which are not vectors. Stresses and strains in solid bodies are examples of these, and so are some of the properties of bodies considered in the theory of elasticity and in the theory of double refraction. Quantities of this class require for their definition nine numerical specifications. They are expressed in the language of quaternions by linear and vector functions of a vector."*

Maxwell worked out his theory in quaternions since, at the time, the modern form of vector analysis had not been originated by Heaviside and Gibbs.

To recapitulate what we stated earlier: A quaternion consists of a vector part and a scalar part and can readily take into account the stress and strain of the medium. A vector consists only of the vector part, and does not take into account the stress of the medium.

It seems certain that Maxwell knew that the scalar part of the quaternion could be present and vary, even though the vector part was sum-zeroed. Quaternions, however, are devilishly difficult, and even in the time of their founder, Hamilton, few mathematicians and scientists ever mastered them.

Maxwell's theory was transposed (shortly after his death) to its modern vector form by Oliver Heaviside. Many of the characteristics of the scalar part of the quaternion were effectively discarded.

In the remodeling, the "unified" basis was omitted, and electro-

Table 37. MAXWELL'S EQUATIONS

- **INVARIANT**
    - **UNDER CONFORMAL GROUP OF TRANSFORMATIONS**
    - **IN 4-D MINKOWSKI SPACE**
- **GROUP INCLUDES**
    - **TRANSLATIONS**
    - **ROTATIONS**
    - **REFLECTIONS**
    - **INVERSIONS WRT HYPERSPHERES OF M-SPACE**
- **HENCE MAXWELL'S EQUATIONS INCLUDE TRANSFORMATIONS THAT**

CHANGE INERTIAL FRAMES INTO FRAMES THAT ARE NOT INERTIAL.

PROVED IN 1910 BY
H. BATEMAN AND E. CUNNINGHAM

gravitation was accordingly discarded. Heaviside wrote a severely limited subset of Maxwell's theory; he did not by any means capture all of it.*

The new Heaviside vector reinterpretation of EM was far easier to calculate in than Maxwell's quaternions. Many physicists (there were really very few scientists in those days!) immediately loved it.

FitzGerald wrote glowingly of Heaviside's translation (and mutilation!) of Maxwell's theory:

"Since Oliver Heaviside has written, the whole subject of electromagnetism has been remodelled by his work. No future introduction to the subject will be at all final that does not attack the problem from at least a somewhat similar standpoint to the one he puts forward."

Thus it was that Heaviside discarded the basis for scalar electromagnetics and electrogravitation from Maxwell's original work, and **the** so-called "classical Maxwell Equations" appeared for the first time in their modern vector form in Heaviside's work.

Then in the early 1890's a short but vigorous debate on vectors and vectorial methods occurred, involving eight journals, twelve leading publications, and over 30 scientists. Over half the publications were in the leading British scientific weekly journal Nature.

This debate, which Lord Rayleigh characterized by the statement "Behold how these vectorists love one another," together with **the** far greater difficulty of quaternion methods as compared to Gibbs' and Heaviside's vector analysis, spelled the death knell to the position of quaternions.

After some years, the Gibbs-Heaviside vector theory was firmly entrenched. The quaternion theory of Maxwell was forever modified **and** limited to the form we know today as "Maxwell's equations."

What really happened was that electrical physicists, faced with the formidable difficulties of quaternions, simply stampeded down the far easier road blazed by Heaviside, bowling over all protests against Heaviside/Maxwell/Hertz theory.

---

*Specifically, Heaviside wrote the subset of Maxwell's theory where gravity and electromagnetism do not interact and are mutually exclusive.

In Western universities today, no one actually teaches Maxwell's true theory! Instead, Heaviside's curtailed reinterpretation —- a severely restricted subset of Maxwell's theory — is taught.

And so that is the EM theory that is learned and applied.

One may rightly argue, therefore, that **Maxwell's theory has only been partially applied in its modern translated form, and the modern form of the Maxwell/Hertz electromagnetic theory is very much incomplete.**

To recover the stress factor in Maxwell's original theory, electromagnetics must take into account the *artificial* potential, which has a deliberately ordered, dynamic, deterministic structure. Scalar electromagnetics takes the patterned stress into account and deliberately uses it. Normal electromagnetics only considers the *natural* potential with randomized VPF structure implicitly assumed.

Scalar Waves ***and Polarization of*** *the* ***Vacuum***

Let us use a simple analogy. This will lead us into the profound implications of the deliberate vacuum polarization and curving of local spacetime that is provided by the vector zero summation (artificial EM potential) method of scalar electromagnetics.*

Visualize two sets of opposing and balanced forces pressing on the sides of a plate.

The forces sum to zero, so the resultant force acting on the plate for translation is zero. Hence the plate does not translate (move away or accelerate).

However, the plate is under internal stress (compression), and is in a quite different condition than when it has no external forces at all acting on it.

Now visualize the forces being applied with rigid rods welded to the sides of the plate, so that the forces may alternately "pull" as well as "push." Let the forces rhythmically vary, alternately "pulling" and "pushing", but always remaining balanced so that their summation for a translation resultant is constantly a zero vector.

The plate never accelerates or moves in translation, but it now contains an internal stress wave which rhythmically varies between compression and tensile stress. Rigorously the plate internally

*Vector **multiplications** that yield a zero vector product are also highly significant.

possesses a scalar stress wave, or a wave of "internal stress in the medium."

Now visualize a similar "plate-like" region of vacuum, with its virtual particle flux, instead of the material plate. Visualize two opposing sets of (unzipped) EM force fields, acting in and on the plate-like region, so that they rhythmically increase and decrease, changing direction also, but always with a zero vector summation.

In this case, there is an unzipped zero resultant EM force field (it is zero, by our assumed conditions), but there is a rhythmic oscillation of the intensity of the vacuum flux (intensity of the vacuum potential) and the vacuum structuring in the region. That is, there is a rhythmic and steady oscillation of the stress energy and structuring of vacuum, and hence of the curvature of spacetime, in **the** plate-like region.

Rigorously, this oscillation — which we call a scalar electromagnetic wave—is a gravitational wave, since the local curvature of spacetime is being oscillated.

Further, it differs from a "natural" gravitational wave in several respects:

1. It has a deterministic pattern or substructure,

2. It patterns or "polarizes" the vacuum,

3. It constitutes local curvature of spacetime, and hence localized general relativity - something which ordinary general relativity assumes cannot be accomplished,

4. It deterministically engineers the virtual state and local spacetime,

5. Since it is achieved by converting EM field energy into artificial G-field energy, one can expect tremendous gains of gravitational and inertial effects in and around electrical circuits utilizing such fields precisely,

6. It affects the Schroedinger wave and the probabilities of the

states being propagated forward by the Schroedinger wave. **With the scalar EM wave, one can thus deterministically engineer the emergence of quantum change, and violate one of the fundamental assumptions of quantum mechanics, that of the totally statistical nature of quantum change,**

7. The scalar wave can accomplish direct and localized change of the rate of flow of time (even to its reversal) and variation of mass and inertia, without concomitant translation of matter,

8. By locally curving and patterning vacuum spacetime, a stabilized standing scalar EM wave can provide macroscopic violation of the conservation laws, which rigorously depend upon a locally flat spacetime (Lorentz frame),

9. Since the components in the zero summation may be EM waves, and may be "locked together" and broadcast to a great distance and interfered there, effects at great macroscopic distances may be achieved, in violation of present assumptions of physics, and

10. Since excess negative time flow may be locally produced, antigravity, negative energy, and negentropy may be locally produced.

These are the startling implications inherent in correcting classical electromagnetics' improper treatment of zero-summed EM force vector systems in vacuum.*

By classically replacing the force vectors in a zero EM force field summation with a zero vector and discarding the components, one is totally discarding the fact that such a patterned system of oppositive forces forms a deterministically patterned stress in the vacuum medium.

By this innocuous error, classical EM discards electrogravitation, and avoids unification of gravitation and electromagnetics. It also avoids direct engineering of gravitation, vacuum/spacetime, inertia, rate of time flow, free energy, and quantum change.

---

*And its zero - multiplied force vector systems as well.

Generations of Western physicists and electrical engineers have acquired a mindset in only one aspect of the vastly expanded electromagnetics actually available.

Vector Zero Systems and the Kaluza-Klein Approach

Prior to 1921, Theodor Kaluza applied Einstein's new general relativity to five dimensions.

He produced a unified theory of electromagnetism and gravitation.

In Kaluza's theory, the ordinary 4-dimensional gravitational field and the electromagnetic field are but two different aspects of a single more fundamental field: the 5-dimensional gravitational field.

Kaluza's theory was published in 1921, on the personal recommendation of Albert Einstein, who had had Kaluza's paper for two years.

In Kaluza's model, electromagnetics is the 5th dimensional aspect of the 5-d G-field, while the ordinary 4-d G-field is the intersection of the 5-d G-field with our ordinary world.

Five years after Kaluza's epochal paper, Oskar Klein explained where the extra space dimension — the fifth dimension — was. He modified it as "wrapped around" each point in our ordinary 3-space.

In the Kaluza-Klein model, then, an electromagnetic wave— which moves spatially only in the fifth dimension — does not move through our 3-space at all. Instead, on its trajectory it "flows around" each point in our 3-space along its path.

Using Kaluza's model, we may regard the EM field as the normal "bleed-off' or escape of the 5-dimensional gravitational potential in the 5th dimension.

Any gradient in a potential represents a bleed-off of the potential from the high point to the low point. The way to prevent effective bleed-off of a potential at a point is to add an equal amount of bleed-in! The bleed-out and the bleed-in at the point then become balanced, so that an equilibrium state exists in the potential.

By this mean, the magnitude of the potential can be fixed by the two opposing bleed processes in equilibrium.

Table 38. KALUZA GEOMETRY

- **THEORDOR KALUZA, POLISH PHYSICIST**
- **UNIFIED THEORY OF ELECTROMAGNETICS AND GRAVITY [1921]**
  - **5 - DIMENSIONAL SPACETIME**
  - **5 - DIMENSIONAL GRAVITY FIELD**
  - **ELECTROMAGNETISM IS THAT PART THAT OPERATES IN THE FIFTH DIMENSION**

Table 39. 5-D G-FIELD BLEED-OFF [KALUZA 5-D THEORY]

- **FAR EASIER TO BLEED-OFF AS ORDINARY EM FIELD**
- **VERY DIFFICULT TO BLEED-OFF AS ORDINARY 4-D G-FIELD**
- **E.G., BETWEEN TWO ELECTRONS:**

$$\frac{EM}{G(4)} \approx 10$$

- **BETWEEN TWO PROTONS:**

$$\frac{EM}{G(4)} \approx 10$$

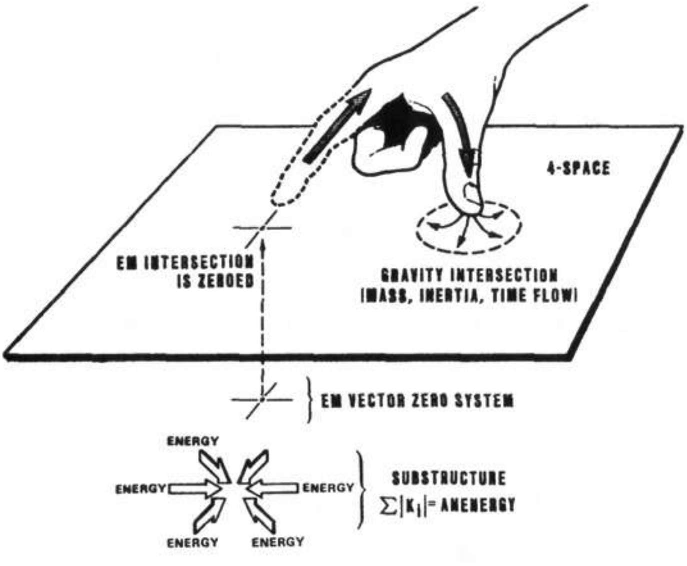
4-SPACE
EM INTERSECTION IS ZEROED
GRAVITY INTERSECTION (MASS, INERTIA, TIME FLOW)
EM VECTOR ZERO SYSTEM
ENERGY
ENERGY
ENERGY
ENERGY
ENERGY
SUBSTRUCTURE
Σ|K_i| = ANENERGY

Figure 63. An analogy for how scalar electromagnetics can engineer Kaluza-Klein unified field theory.

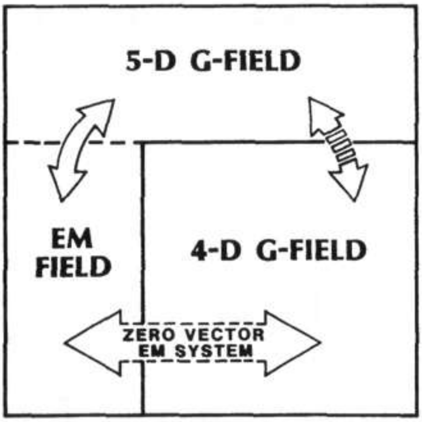
5-D G-FIELD
EM FIELD
4-D G-FIELD
ZERO VECTOR EM SYSTEM

Figure 64. The zero-vector system can engineer Kaluza-Klein theory.

WHERE DID THE EXTRA DIMENSION GO?

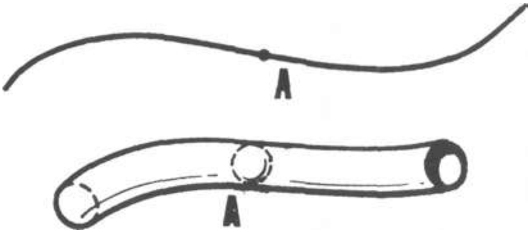

- **FROM A DISTANCE, A HOSE-PIPE LOOKS LIKE A WIGGLY LINE**

- **WHAT WE REGARD AS A POINT IN 3-SPACE IS REALLY A TINY CIRCLE AROUND ANOTHER DIMENSION OF SPACE**

Figure 65. The basic idea of Klein's explanation of where the extra dimension is.

Table 40. PHYSICAL REALITY IS ELEVEN DIMENSIONAL

- **IN THE KALUZA-KLEIN THEORY THE GAUGE FIELD SYMMETRIES BECOME CONCRETE**

- **THEY ARE THE GEOMETRICAL SYMMETRIES ASSOCIATED WITH THE EXTRA SPACE DIMENSIONS**

- **AN 11-SPACE IS REQUIRED BY THE GRAND UNIFIED FORCE THEORY**

PAUL DAVIES, SUPERFORCE.
1984, P. 160-161

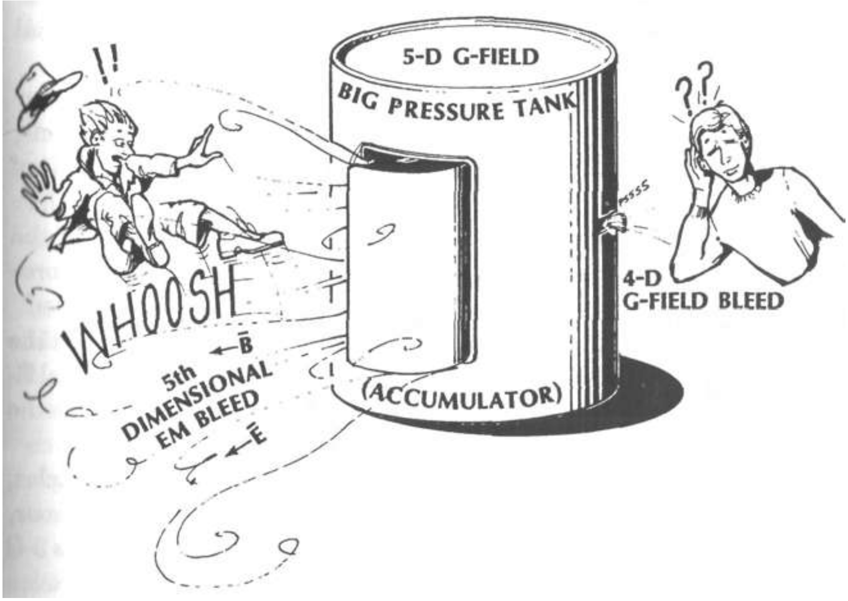
!!
??
5-D G-FIELD
BIG PRESSURE TANK
PSSSS
WHOOSH
4-D
G-FIELD BLEED
← B̄
5th
DIMENSIONAL
EM BLEED
← Ē
(ACCUMULATOR)

Figure 66. Normal EM bleeds off G-field.

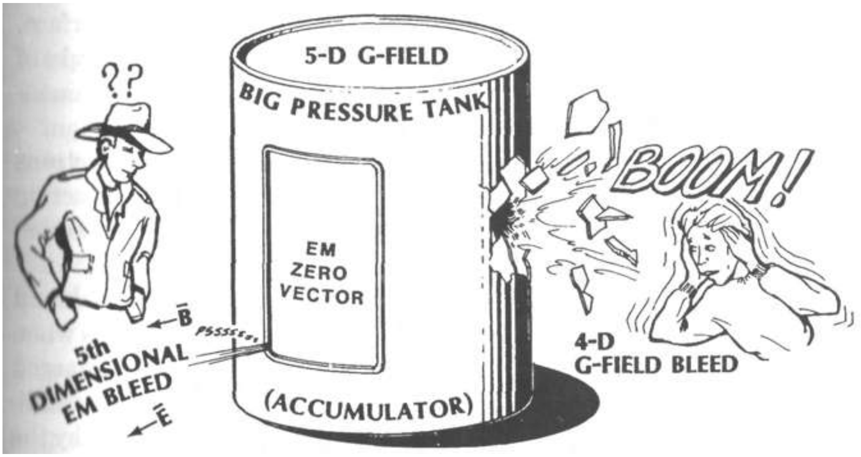
??
5-D G-FIELD
BIG PRESSURE TANK
BOOM!
EM
ZERO
VECTOR
← B̄
PSSSSSS
5th
DIMENSIONAL
EM BLEED
← Ē
4-D
G-FIELD BLEED
(ACCUMULATOR)

Figure 67. Scalar EM forces 4-dimensional gravitational field.

Note that this positively establishes that two or more zero-summed EM force field gradients produce and stabilize an artificial potential.

Usually, very little bleed-off of the 5-potential occurs in our normal 3-space. When the 5-space G-field is bled off on electrons, e.g., the resulting 5th dimensional E-field bleed-off is on the order of 10^{42} times as much as the 3-space G-field bleed-off.

This accounts for the enormous magnitude of the ratio of the electrical force between two electrons to the gravitational force between them.

Even when the 5-space G-potential is bled-off on and against the greater inertia of much more massive protons, its 5th dimensional E-field bleed-off is still on the order of 10^{36} or so times as much as the 3-space bleed-off.

Note that if we can stop the acceleration of the charged particles, bleed-off of the 5-space G-potential as EM force field cannot occur, and instead all of it must bleed-off as 3-space G-force field. This 3-G bleed-off normally occurs only minimally in electromagnetics, even for so-called "static" charges. This can be seen as follows:

When one charges a conductor such as a large metal sphere, for example, the electrons are unconstrained on the surface of the sphere. They are in constant agitation and motion on the surface, thus constantly bleeding-off the 5-space potential as myriads of squirming little E-fields. Only a very little is bled off as 3-dimensional G-field.

On the other hand, when one charges a dielectric, the electrons cannot be in as much agitation movement as they are on a conducting surface. Hence slightly more of the 5-dimensional G-potential is bled off as 3-dimensional G-field.

Therefore, charged conductors and charged dielectrics should exhibit slightly different gravitational and inertial properties when exposed to a gravity field potential difference. Highly charged capacitors, for example, should exhibit a small but detectable unidirectional thrust in vacuum. Apparently they do, as exhibited by the Biefield-Brown effect.

In addition, different materials should exhibit slightly different fall rates gravitationally. Again there is some evidence that they do, as exhibited by the Eotvos experiments in a modern reinterpretation by Fischbach et al.*

Electrogravitation Amplification Factor

If the movement of two free electrons could be entirely stopped, all EM bleed-off of the 5-d G-potential between them would be effectively zeroed or blocked. Then the 5-d G-potential would be forced to totally bleed-off into ordinary 3-d G-field.

For two electrons, this could increase the gravitational force between them by a factor of 10^{42}, if the EM bleed were perfectly blocked.

Such perfect blockage of the 5th dimensional EM bleedoff cannot be accomplished in practice. Normally, with the increase in potential caused by zero-summing, agitating electrons in a statically charged object agitate more intensely, moving and squirming in local, erratic constraints.

However, conversion efficiencies on the order of 10^{10} or even 10^{20} appear to be achievable in ordinary circuitry. In that case, gravito-inertial amplification gains of up to 10^{22} or even 10^{32} appear to be achievable. Gains of 10^{10} or more appear to be fairly readily achievable.

Time delay in producing the gravito-inertial effects may be experienced, however, because mass (nuclei) acts as a capacitance to incident scalar waves, exhibiting capacitive charging and discharging "time constants" when charging scalarly for a particular effect or to a particular pattern.

In these achievable amplification ranges, a single milliwatt of electrical power can control enormously powerful local gravitational and inertial effects, once the scalar EM charge is built up in the nuclei of the target material.

Excessive and frustrating "fiddling time" with the circuits is usually required, however, due to the charging "time constant" effect, the general lack of adequate measuring instruments, and the great uncertainty as to the phenomenology and major variables involved.**

---

*Ephraim Fischback et al, "Reanalysis of the Eotvos experiment," Phy. Rev. Lett., 56(1), Jan, 1986, p. 3-6.

••A series of scalar EM instruments, however, is being developed by Bedini, Schnur, Golden, **and** perhaps even Hewlett-Packard.

However, with persistence, in rather ordinary EM circuits and devices, the use of "summation to zero-vector" techniques and "multiplication to yield a zero-vector" techniques can lower the EM bleed-off of the 5-potential sufficiently to allow substantial local gravitational/inertial effects. By beaming and wave interference, effects at a distance can be achieved.

When gravito-inertial amplification of many orders of magnitude are involved, the assumption in ordinary general relativity of a local Lorentz frame (local special relativity, with local conservation laws applying) is readily violated.

Now one may have local general relativity, with concomitant violation of conservation laws, and local broken symmetry on a macroscopic scale.

Soviet papers already strongly point out that such unrestricted general relativity allows the violation of all conservation laws.

In the Soviet Union, scalar electromagnetics is called *energetics.*

Comparison of EM Concepts

Table 41 shows a comparison between three concepts of electromagnetics: classical, quantum mechanical, and scalar electromagnetic views.

As can be seen, scalar EM re-introduces Maxwell's original potential for electrogravitational variation and structuring of the stress energy density of vacuum spacetime. It applies Kaluza-Klein 5-dimensional unified theory in an engineering fashion.

It deliberately utilizes phase conjugation as a tool to achieve startling negative time/negative energy capabilities and results, including "free" energy and antigravity.

With scalar EM, it should be possible to overcome or circumvent most of the severely limiting assumptions of present physics and engineering.

The great conservation laws yield to direct manipulation and violation. The previously inviolate statistical structure of quantum change becomes deterministic and engineerable.

General relativity becomes an experimental scientific discipline in the laboratory, instead of a complex description of effects only observable in distant, massively curved regions such as on the surface of a star or near a black hole.

God does not play dice with the universe after all, just as Einstein suspected.

The Aharonov-Bohm Effect

In 1959, Y. Aharonov and David Bohm published a fundamental paper in **Physical Review** which pointed out the quantum mechanical implications of potentials.

The seminal paper is Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," **Physical Review,** Second Series, 115 (3), Aug. 1, 1959, p. 458-491.

In classical electromagnetics, the potentials had largely been regarded as mathematical conveniences, having no real effects.

Table 41. Comparison of EM Concepts

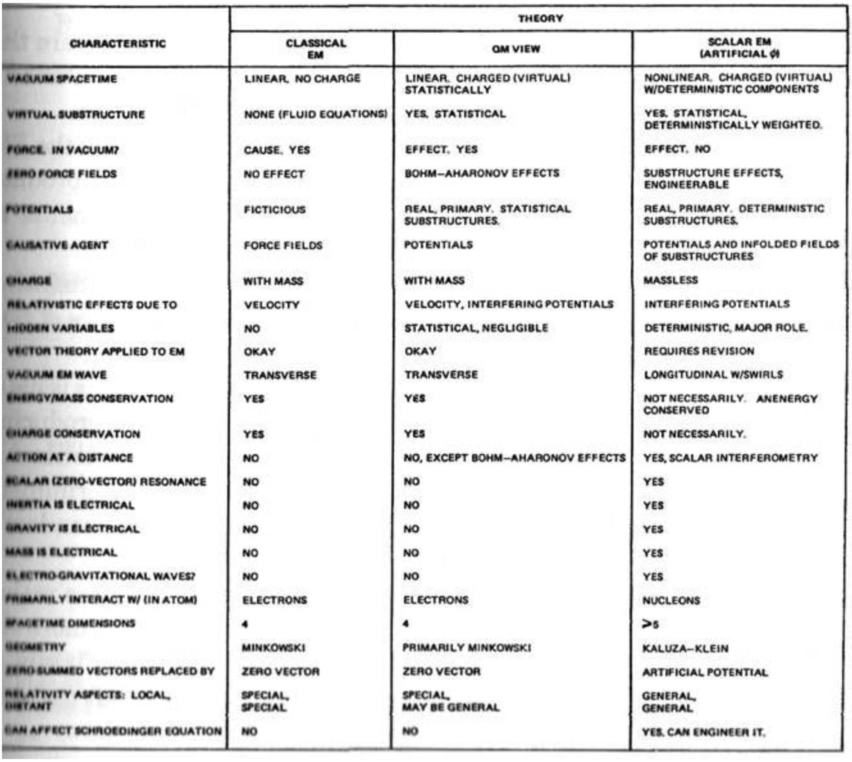

CHARACTERISTIC	THEORY		
	CLASSICAL EM	QM VIEW	SCALAR EM (ARTIFICIAL ϕ)
VACUUM SPACETIME	LINEAR, NO CHARGE	LINEAR, CHARGED (VIRTUAL) STATISTICALLY	NONLINEAR, CHARGED (VIRTUAL) W/DETERMINISTIC COMPONENTS
VIRTUAL SUBSTRUCTURE	NONE (FLUID EQUATIONS)	YES, STATISTICAL	YES, STATISTICAL, DETERMINISTICALLY WEIGHTED.
FORCE, IN VACUUM?	CAUSE, YES	EFFECT, YES	EFFECT, NO
ZERO FORCE FIELDS	NO EFFECT	BOHM–AHARONOV EFFECTS	SUBSTRUCTURE EFFECTS, ENGINEERABLE
POTENTIALS	FICTICIOUS	REAL, PRIMARY, STATISTICAL SUBSTRUCTURES.	REAL, PRIMARY, DETERMINISTIC SUBSTRUCTURES.
CAUSATIVE AGENT	FORCE FIELDS	POTENTIALS	POTENTIALS AND INFOLDED FIELDS OF SUBSTRUCTURES
CHARGE	WITH MASS	WITH MASS	MASSLESS
RELATIVISTIC EFFECTS DUE TO	VELOCITY	VELOCITY, INTERFERING POTENTIALS	INTERFERING POTENTIALS
HIDDEN VARIABLES	NO	STATISTICAL, NEGLIGIBLE	DETERMINISTIC, MAJOR ROLE.
VECTOR THEORY APPLIED TO EM	OKAY	OKAY	REQUIRES REVISION
VACUUM EM WAVE	TRANSVERSE	TRANSVERSE	LONGITUDINAL W/SWIRLS
ENERGY/MASS CONSERVATION	YES	YES	NOT NECESSARILY. ANENERGY CONSERVED
CHARGE CONSERVATION	YES	YES	NOT NECESSARILY.
ACTION AT A DISTANCE	NO	NO, EXCEPT BOHM–AHARONOV EFFECTS	YES, SCALAR INTERFEROMETRY
SCALAR (ZERO-VECTOR) RESONANCE	NO	NO	YES
INERTIA IS ELECTRICAL	NO	NO	YES
GRAVITY IS ELECTRICAL	NO	NO	YES
MASS IS ELECTRICAL	NO	NO	YES
ELECTRO-GRAVITATIONAL WAVES?	NO	NO	YES
PRIMARILY INTERACT W/ (IN ATOM)	ELECTRONS	ELECTRONS	NUCLEONS
SPACETIME DIMENSIONS	4	4	>5
GEOMETRY	MINKOWSKI	PRIMARILY MINKOWSKI	KALUZA–KLEIN
ZERO SUMMED VECTORS REPLACED BY	ZERO VECTOR	ZERO VECTOR	ARTIFICIAL POTENTIAL
RELATIVITY ASPECTS: LOCAL, DISTANT	SPECIAL, SPECIAL	SPECIAL, MAY BE GENERAL	GENERAL, GENERAL
CAN AFFECT SCHROEDINGER EQUATION	NO	NO	YES, CAN ENGINEER IT.

Table 42. THE AHARONOV - BOHM EFFECT

- **IN FIELD-FREE REGIONS**
    - **E - FIELD IS ZERO**
    - **B - FIELD IS ZERO**
- **POTENTIALS STILL EXIST**
    - **CAUSE REAL EFFECTS**
    - **INTERFERENCE IS KEY**
    - **CONTAIN SUM-ZEROED SUBSTRUCTURE**
- **DOES NOT FOLLOW FROM**
    - **MECHANICS**
    - **CLASSICAL ELECTROMAGNETICS**

However, according to quantum mechanics, the potentials are the real entities, while the electromagnetic force fields are just effects derived from the potentials by imposed operations on them.

Aharonov and Bohm showed that, even in the absence of the EM force fields, the potentials may still exist and interfere to produce real effects in physical systems. They suggested experiments to prove these predictions.

Interference of the potentials is the key mechanism producing real effects in charged particle systems, even in the presence of zero E-field and zero B-field.

Unfortunately, Aharonov and Bohm did not address this issue of zero-summed systems of EM force fields as one way of producing artificial potentials having unique interference characteristics.

That is, by implication they did not consider the deterministically substructured "artificial" potential, but only the randomly structured "natural" potential.

At any rate the "AB effect," as it came to be called, was rather hotly debated and fiercely resisted over the years. Steadily, more and more experiments were added, validating the AB effect. Finally, in 1986 it came to be generally accepted as proven to any but the most die-hard skeptic.

The AB principle represents a violation of both classical mechanics and classical electromagnetics. It is absolutely required, however, by quantum mechanics. And experiments have now established it as proven beyond question.

The potential, then, can be a free and independent field. Further, it can directly act on charged particle systems, even when classical electromagnetics does not recognize the interaction.

Curl-Free Magnetic Vector Potential (A-Field)

As an example, in classical EM the vector magnetic potential (the A-field) had been defined as a mathematical convenience by the equation

$$\nabla \times \mathbf{A} = \mathbf{B} \qquad (4\text{-}20)$$

But if the potentials are real, then conceivably the A field can be loosed from its enchainment to the $\nabla\times$ operator. In that case, it

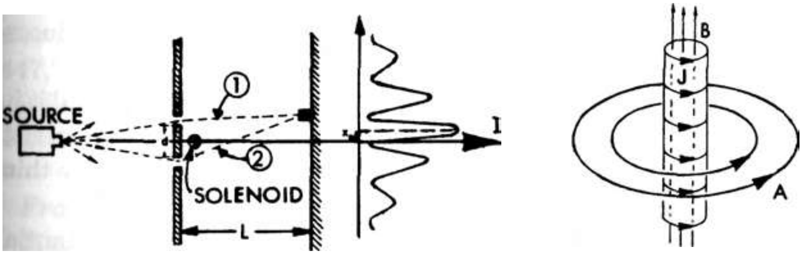

A - FIELD CAN INFLUENCE THE MOTION OF ELECTRONS.

MAGNETIC FIELD AND VECTOR POTENTIAL OF A LONG SOLENOID.

THE FEYNMAN LECTURES ON PHYSICS. VOLUME II

Figure 68. The curl-free A-field. The magnetic vector potential (A-field) may be freed from its attachment to the magnetic field (B-field) and electric field (E-field). It can interfere at a distance to cause real effects. From it, either E-field or B-field, or both, can be made at a distance.

becomes a free, new, and independent field of nature with potentially unique characteristics.

For instance, its defining equation shows that magnetic force field can be made from it, and the rightmost term of the equation

$$\mathbf{E} = -\nabla \text{Ø} - d\mathbf{A}/dt \qquad (4\text{-}21)$$

shows that its time rate of change makes an electrical field.

Let us explain this in simpler terms, and somewhat more precisely than conventional theory. We will use the Kaluza unified G-EM interpretation and electron flow in our explanation.

In the A-field, we have a certain kind of 5-dimensional G-potential which can bleed-off as EM force fields in two ways: (1) in a swirl fashion, where the vortex producing the swirl moves parallel to electron movement, and the swirling is a torque or spin, and (2) in a linear fashion, where the time rate of change of the A-potential produces a linear E-field on the electrons.

The first bleed-off as given by equation (4-20) constitutes the magnetic B-field, and the second bleed-off as given by the rightmost term of equation (4-21) creates a component of the overall electrical E-field. Bleed off of the electrostatic scalar potential produces the other component of the E-field.

Now in general relativity (GR) theory, "the" G-potential is just a conglomerate of many things, each of which has the characteristic of curving spacetime. "The" gravitational field is not a single thing at all, but is composed of a collection of many things.

Thus if we realize that both the electrostatic scalar potential (Ø-field) and the magnetic vector potential (A-field) are components of the 5-d G-potential, then we see immediately that bleed-off of these two components of the 5-d G-potential creates all normal EM fields.

It follows that, if we produce a zero-vector summation of the two or more EM bleed-offs, we are actually "putting as much back in" to the 5-d G-potential through its A and Ø components as we are taking out electromagnetically.

In that case, the 5-potential is in a state of equilibrium with respect to EM bleed-off in the fifth dimension. It is now forced to bleed-off in the only other way it can: as ordinary 3-dimensional gravitational field. Thus by vector-zeroing EM force fields, we turn KM field energy into G-field energy and vice versa, via the intermediary of the 5-potential.

At any rate, soon after publication of the Aharonov-Bohm paper, experiments showed that, if the magnetic field is trapped inside a long solenoid, a phase shift still is induced in the two-slit electron experiment, even though — classically — no contact of the enclosed magnetic field and the moving electrons occurs. This phase shift is explained by the fact that the freed A-field exists outside the trapping solenoid, even though the B-field does not. Consequently, interaction of this curl-free A-field with the electrons produces a phase shift of the QM interference detection pattern.

This proves that curl-free A-field is real and causes physical effects.

It also proves that a form of "action at a distance" is real, just as required by quantum mechanics.

Years ago, Frank Golden and this author — together with Dr. William Tiller — experimented with "free A-field" devices. Golden went on to develop prototype transmitters and receivers and a prototype underwater communication system.

Since that time, Gelinas has patented several curl-free magnetic vector potential (free A-field) devices. See U. S. patents numbers 4,447,779, May 8,1984; 4,429,288, Jan. 31,1984; 4,429,280, Jan. 31,1984; and 4, 432, 098, Feb. 14,1984. These patents are assigned to Honeywell.

Fractional Charges, Magnetic Monopoles, and Magnetic Currents

Fifteen years ago, a most distinguished gathering of physicists occurred in Trieste, Italy to review the physicist's conception of nature and how it developed. Seven Nobel Laureates were in the prestigious audience. The keynote address, appropriately, was given by P.A.M. Dirac, and — appropriately — was entitled "Development of the Physicist's Conception of Nature."

Dirac called attention to the possibility of fractional charges, and unexpectedly presented some of the data from experiments done lung before by Felix Ehrenhaft.

Years earlier Felix Ehrenhaft, former director of the Physical

Institute at the University of Vienna, had reported a number of experiments where — purportedly — he easily obtained fractional electrical charges.

Whereas Millikan had used liquid oil drops in his experiments to determine the charge of the electron, Ehrenhaft used little solid, red selenium oxide spheres. Not only could Ehrenhaft get a fractional charge, but he could actually vary the size of the fraction by varying the size of the spheres.

(Knowing what has been discovered in the last two decades on the complex bond structuring of water and liquid hydrocarbons, one might very well suspect that a solid — which has one less degree of freedom than a liquid — could possibly yield a very different experimental condition from a liquid, from first principles. However, no one seems to have noted that, or taken it into account.)

(Also, Ehrenhaft's results were replicated by other researchers who used his experimental methods. They published scientific papers, but that made not a whit to the scientific community, which simply rejected the experimental evidence.).

Dirac pointed out that Millikan himself, in his oil drop experiments that established the charge of the electron, reported one anomalous experiment which yielded a fractional charge. He discarded the results of that experiment.

Ending his speech, Dirac stated that it all made him wonder if perhaps Millikan did have a fractional charge in that one experimiment after all. And well he might wonder. Ehrenhaft — almost unknown now to the present generation of physicists — may have found not only fractional electrical charges, but the magic magnetic monopoles as well.

Over and over, Ehrenhaft reported the results of experiments which disagreed with the notions of magnetic theory at the time.

Again, some of these experiments were replicated, and reported by those researchers replicating them.

It didn't make any difference at all.

Now, finding a magnetic monopole — and the accompanying magnetic charge current — would be most satisfying to many theorists. That would make the equations of magnetics quite "balanced" with respect to electricity.

However, no one —at least whose results have been accepted -

has yet found a magnetic monopole or magnetic current.

Stanford researchers have perhaps come closest, and even reported the results of one exciting experiment that found the magnetic monopole. However, the team has consistently failed to be able to replicate those results. And in science, "no replication, no acceptance" is necessarily the dictum.

In a moment we will address a possible cause for this failure to find the magnetic monopole and magnetic charge current.

Some Ehrenhaft Experiments

Let's look at some simple, rather informal experiments that Ehrenhaft showed. We call attention to Alden P. Armagnac's excellent article , "Magic with magnetism," in **Popular** Science **Magazine,** June 1944, p. 130-133, 222, 226. In the first experiment, Ehrenhaft seems to have dissociated water with a permanent magnet.

First, Ehrenhaft seals two rods of pure iron, in holes through opposite sides of the U in a U-shaped glass tube filled with water. Then he uses the two rods — which protrude through the tube into the water — as the pole pieces of a horseshoe magnet — either an electromagnet or a permanent magnet. Dilute sulfuric acid — one percent by volume — is in the water, to give it a very slight acidity.

Bubbles of gas rise up through the twin vertical columns of water, and are collected and analyzed. Most of the gas is hydrogen, as is to be expected from common chemical interaction of the iron rods with the very dilute sulfuric acid. However, some of the gas — ranging from two to twelve percent — is oxygen, and most of the oxygen is produced at the north pole of the magnet.

The pole pieces may be short-circuited with wire, to insure they are at the same electrical potential and prevent any electrical dissociation of the water.

Without the magnet, all the gas liberated is hydrogen, as is to be expected. No oxygen at all is produced.

And another very interesting phenomenon occurs. A strong permanent magnet of the Alnico type gradually loses a marked fraction of its strength — say, 10% in 24 hours. It appears that the energy dissociates water — very similar to electrical energy from a battery draining away when electric current is used to dissociate water.

The real question is: Is this experiment a demonstration of magnetic current?

If it is, then the textbooks are wrong, and there's another constituency in magnetism that has not been taken into account. Further, if it proves magnetic current exists, it has also proven magnetic charge — the magnetic monopole — exists.

If it is not a demonstration of magnetic current, then what is it? Why don't you get any oxygen liberated when the magnetism is removed? What causes the magnet to discharge its magnetic strength?

These serious questions do not appear to have ever been resolved.

In all fairness, however, we must warn the reader that other factors in the water can very seriously affect the outcome of this

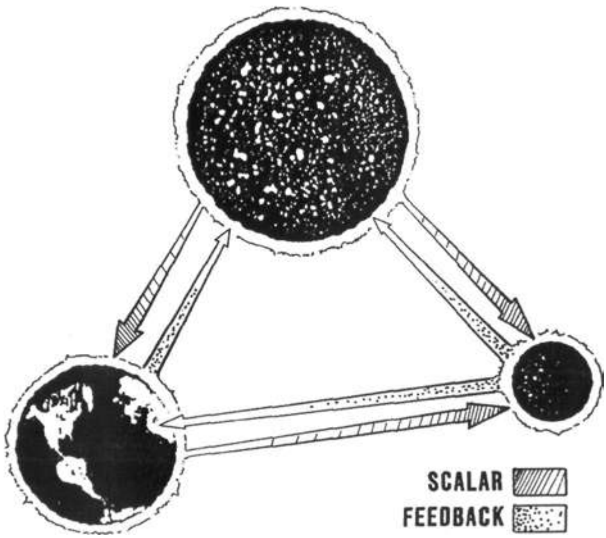

EARTH • SUN - MOON TRIAD

Figure 69. Earth, sun, and moon are locked into a triad system. They exchange both normal energy and phase-conjugate energy. The phase-conjugate energy (negative energy, negative time) streams affect all "subtle-energy" earth experiments

experiment. As we now know, the bonding structure of water is incredibly adaptive. Every bit of water you examine, for example, is absolutely unique in many respects. And in other respects, it's a sort of "single giant molecule."

Be forewarned also that subtle experiments with water are affected by just about everything under the sun. Including the sun!

Ehrenhaft himself was quite embarrassed when this "decomposition of water" failed to be 100% repeatable. It can be done — but you may have to work at it a bit. The best water for this experiment is the most natural, pure ground water you can obtain.

At any rate, it may be that these experiments of Ehrenhaft's are related to the old notion of "watergas" — the changing of water in some fashion so that, as it enters an internal combustion chamber, it decomposes into hydrogen and oxygen, and is ignited by the spark and burns as a fuel. In the typical watergas case, the experimenter has a "green powder" which, when added to water, will allow it to be burned as a fuel when it is used in an ordinary engine. Such a capability, for example, seems to have been exhibited at least once, to the U.S. Navy, by an unorthodox experimenter named Anderson.

Of course, if fractional charges exist and can be controlled (and Ehrenhaft's experiments certainly seem to confirm this), then it is conceivable that a material could be found that would do the "green powder" function. That is, if the bonding electron that bonds hydrogen and oxygen in water were replaced by a fractional charge, the "fractionated water molecule" could just barely be holding together an a liquid. A strong hydroshock — such as going through the jet nozzles in a carburetor — could then cause many of these molecules to break apart, effectively dissociating the water.

A possibly related phenomenon was reported by Le Bon in his Evolution of Matter. The experiment was to hold a piece of ordinary aluminum under the surface of mercury in a container for a period of time. Then the piece of aluminum is taken out and thrown into a container of water, where it violently dissociates the water. At least one colleague was able to replicate this experiment by holding the aluminum beneath the mercury for 36 hours.

These Ehrenhaft experiments may also be related to the extensive use in the Soviet Union (and to some extent, in the U.S.) of magnetic treatment of water pipes so that boilers do not develop so much hard precipitate. Magnetic treatment of water, for example, can rigorously be shown to alter some of a liquid's properties.

Ament, for example, in numerous well-controlled repetitions has clearly shown an increase in auto gas mileage by using a strong magnetic apparatus on the fuel line. Best results are obtained by a magnet which only has one polarity facing the fuel line, and using three stages where the middle polarity is reversed. The effect of this apparatus is definitely scalar EM (gravitational) and expressed in the atomic nuclei in the molecules of the fuel flowing through the line.

Again, the Ehrenhaft experiments may be related to the strange results of Viktor Schauberger, who obtained motive results with water and natural energy that have not been duplicated before or since, so far as anyone can tell.

Schauberger utilized "implosive" energy rather than "explosive" energy — strongly suggesting he was using the negative time aspects of energy, hence negentropy. Schauberger also conceptualized this energy in a spiral form — and that is precisely the type of motion Ehrenhaft observed in many of his bubble experiments. That such negative time/negative energy is "natural" energy will be apparent from our next chapter, where we point out that all living things must use negative time/negative energy processes if they are to defeat the inexorable law of entropy — the increasing disorder of a system as its positive flow of time progresses.

Note, **for** example, **that a** system **that is 20% efficient in positive time would be 500% efficient in negative** time.

Let's look at another of Ehrenhaft's magnetic experiments.

Taking the air gap between two very flat, very parallel pole pieces of sufficient width, one obtains a very accurately-parallel, uniform magnetic field in between them. Now suppose you place some very small particles of material—both magnetic and nonmagnetic — in the air gap.

Common sense assures us that, if closer to the south pole, for example, one of the little magnetic particles will be attracted by **the** preponderance of south polarity. That is, a little magnetic north **pole** will be induced on the side of the particle nearest the south polepiece, and a little south pole will be induced on the particle on the side away from the south polepiece. The little particle will then be attracted to the south polepiece.

Again, common sense assures us that, if closer to the north pole, the situation is reversed and the little particle is attracted to the north polepiece.

Well, that isn't quite what happens. Some move to the south polepiece, and some move to the north polepiece. And it seems to vary by the type of material the little particle is comprised of.

Now try another experiment. Let's make a permanent magnet motor. Understand, this one is not quite "for free," for we'll consume "fuel"—the strength of the magnet and the acidified water solution.

In this one, we use a glass cylindrical cell, with one vertical iron rod at the top and another at the bottom, and the rods serving as the pole pieces of a giant horseshoe magnet. (Or, for ease you can use an electromagnet). The cell is filled with the same acidified solution as before.

Without the magnet, hydrogen bubbles can be seen rising as before.

Now put on the magnet (or turn on the switch). Voila! the situation with the bubbles changes drastically. A spectacular miniature merry-go-round — a "whirligig" — of bubbles forms between the faces of the poles and parallel to them.

Add some copper particles to the solution. The particles will rotate in the same plane, but in opposite directions from the bubbles. The stronger the magnetic field, the faster the particles and bubbles will contrarotate.

If the magnetic field is reversed, the rotation of both the bubbles and the copper particles is reversed.

Only a few months after this was reported in Popular Science Magazine, Brother Gabriel Kane of Manhattan College and Charlea B. Reynolds of the Federal Communication Commission confirmed Dr. Ehrenhaft's discovery of magnetic currents.

They made a drop of copper sulfate solution spin between the polepieces of a permanent magnet. They also made an interposed microscope cover glass spin from the rotation of the copper sulfate.

It was hinted they were heading toward power machinery. If so, nothing else was heard — possibly because they never solved the problem of consuming the magnet's strength and the acidified solution.

Bill Mueller, the present inventor of a free-energy device, has also demonstrated the spinning of a piece of plastic in a magnetic field.

After WWII, Dr. Ehrenhaft returned to his post in Vienna, published later work in French and other scientific journals, and died not long after.

What is a magnetic monopole?

First, let's reiterate what we mean by "magnetic field." It'll take a little bit to get there, so please bear with me.

By electric charge, we mean the virtual photon flux connected with a particle of mass or an object (of mass). One type of charge (by convention, positive) is modeled to constitute a "source" of this flux. The second type of charge (negative) is modeled to constitute a "sink" of this flux.

Everything in electromagnetics is fluid flow, you see. The old guys who formed the theory believed that electricity was a thin material fluid. So they modeled it that way.

They modeled the positive charge as an everlasting "sprayer" and the negative charge as a "sink" or everlasting accumulator.

They reasoned that of course this "stream" of electric fluid could "swirl" or rotate, just as normal fluids can.

This "swirl" in the stream of virtual photons is magnetostatic scalar potential — or magnetic pole for short.*

Obviously there can be two directions of swirl — one right-handed, and one left-handed, with respect to the direction of the virtual photon stream. So there are two kinds of "magnetic pole" — north (right-hand swirl) and south (left-hand swirl).

The strength of one of these "swirls" can vary, of course, and that's a variation in the strength of the corresponding pole.

You can also think of the "pressure" of the swirl, so to speak, **at** least for modeling purposes. In that model, north swirl is "higher

---

*Rigorously, a permanent magnet does not produce an observable B field in vacuum! Instead, it produces the dynamic, swirling "potential" for that field — i.e., an **unzipped** field. A particle of observable mass placed in this field integrates and "zips" — producing the observed B-field on and of the spinning observable particle.

than ambient" pressure, and south swirl is "lower than ambient" pressure, where "ambient" means "with respect to right-hand swirl." The swirl pressure is then considered to "bleed off/flow" from north pole to south pole. This is modeled by Faraday's "flux lines of force." This bleed off (gradient) from north pole to south pole is called the magnetic force field, or just "magnetic field" for short, even though in vacuum it's really an *unzipped* magnetic field.

Note that the original modelers didn't consider negative time and negative energy at all. And they didn't know anything about phase conjugation — in fact, the electron had not even been discovered when all this model was originally put together.

But back to the model. Considering the swirling fluid as "conserved", if you produced a concentration of one kind of swirl at one place, you had to decrease the concentration elsewhere, which is the same as producing and adding opposite swirl there. Thus, conceptually, in the simplest case (only two swirls allowed), a north pole and a south pole always occur together in a pair — though of course there's no limit to the distance between the pair.

Notice that if you have a conservative field such as we assumed, you are keeping your overall frame of observation Lorentzian — in other words, conservative. You do not have a "curved spacetime" to speak of, by assumption.

Now of course, you can have compound and very complicated magnets and magnetic fields — you don't always have to have everything so neat and simple. In that case, conceptually you could have — for example — one north pole and two south poles.

In fact, it's easily possible to demonstrate that: a bar of magnetic material is simply magnetized so that a north pole is in the middle, and a south pole is at each end. Now bend it around a bit, and you get the point.

However, now let us suppose that we curve spacetime itself, locally — something they avoid in present physics by pure assumption.

Specifically, we curve our local spacetime in such a fashion that there's more right-hand swirl in the local region than there is left-hand swirl there.

In that case, we have a north-pole region there , all by itself. We have a magnetic north monopole. If that "curved spacetime" region of trapped swirl is moving, that constitutes magnetic current, and it may be either positive (north polar) or negative (south polar) flow.

It's as simple as that! Only when you curve local spacetime can you get a magnetic "charge" or monopole. Everyone who's trying to demonstrate a magnetic monopole in uncurved spacetime is doomed to failure a priori. It takes but little reflection to see that, conceptually, a magnetic monopole is a curvature of local spacetime. Indeed, each of the two poles of a bar magnet is precisely that anyway.

Now a monopole can indeed be "made", by bending local spacetime correctly. It can be made in two ways: in the virtual state vacuum itself, or in and on physical materials. It's easiest to make it in the virtual state: just curving the local spacetime will do that. When this "excess swirl flux without corresponding antiswirl flux" is impressed into and onto matter, however, the atomic nuclei are affected, producing excited "magnetic states." In addition, a great deal of physical stress is created in the material, due to monopolar repulsion between the various "nuclear monopoles deposited in it."

If a sufficiently strong monopole intensity (magnetic charge) is produced in the material, the material will violently fragment from internal stress. When a metal, for example, undergoes such **a** "monopolar fragmentation", it is common to see "tiled edges" of the metal rupture, much as twisted little "planks" or tiles of metal unfurled and "split." Such a break may be a positive signature of the involvement of virtual magnetic monopoles (i.e., by the metal having acquired virtual magnetic monopolar charge, causing the disruption).

This precise type of metal fracture has been very well demonstrated in experiments by Canadian experimenter John Hutchinson. Anomalous, powerful breaks in metal have been obtained by Hutchinson where the two ends that break apart both have a north pole or a south pole.

The anomalous **hole in the fuselage of the Arrow DC-8 destroyed at Gander Air Force Base, Newfoundland on Dec. 12,1985,** strongly suggests that an electromagnetic missile struck the aircraft, inducing monopolar effects. That would also match the-**three** eyewitnesses' description of an anomalous glow on part of the aircraft as it faltered and then fell.

Supercharges and Fractional Charges

Now without further ado, let me just state flatly that the only way to get a fractional charge is to have an appreciable curvature of local spacetime.

To get that local curvature, it's going to be very difficult to use liquid. For weak effects, it will almost always change its bond structure and not localize the curvature in the nuclei. So "fractional charges" in liquid are much harder to come by.

However, with solids it's a different thing, particularly if you make the solid material very, very small and nonlinear — precisely like Ehrenhaft's little red selenium oxide spheres. There you can get **the** curvature to localize without so much bleedoff through the bonding lattice.

The greatest electrical wizard of all, Nikola Tesla, reported producing electrons with 50 times the normal electrical charge of an electron. That's readily understandable if you look at what's happening to the charge of a particle: To have 50 times as much charge, the electron must be exchanging virtual photon flux with its surrounding vacuum at 50 times the rate of the normal vacuum.

In other words, *it's the vacuum (spacetime) which must be*

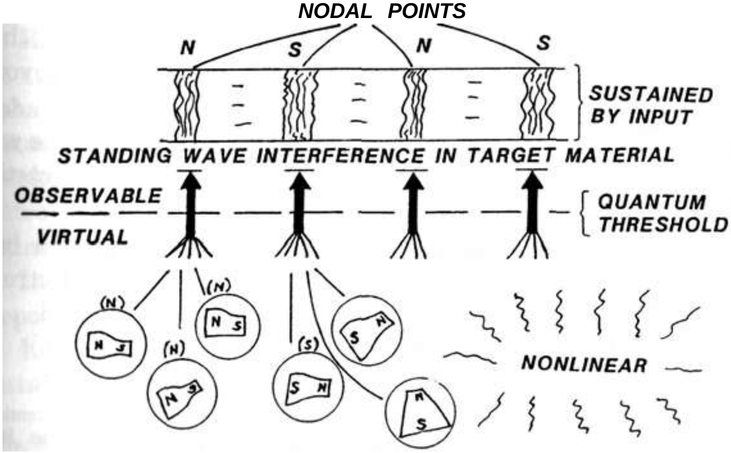

Figure 70. Forming magnetic monopoles with a standing scalar EM wave. In a nonlinear vacuum (locally curved spacetime), virtual magnetic monopoles can be formed and deposited in materials.

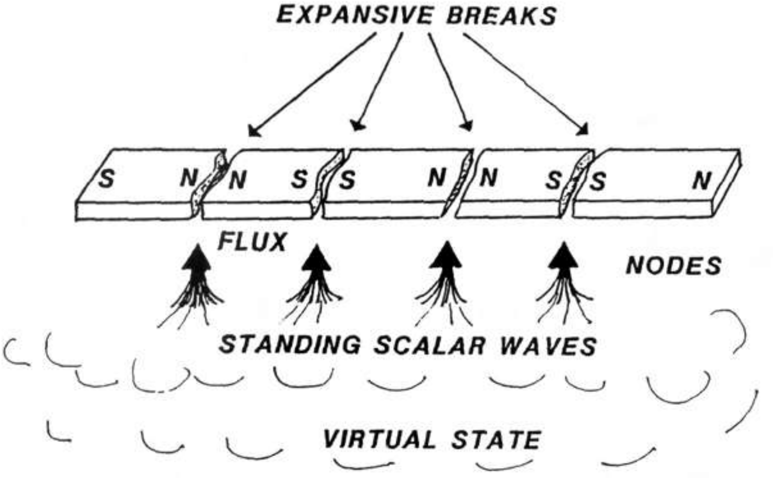

Figure 71. Metal fracture by depositing magnetic monopoles. Demonstrated by Canadian researcher John Hutchinson.

changed, not the electron! Actually, that happens anyway whenever the electron is in an electrical potential field. There are many unrecognized assumptions in the postulate that the electron charge is fixed.

If you curve local spacetime in one direction, you get the charge, of an electron increased. If you curve it in the other direction, you get the charge decreased. In that case it's called a *quark*.

At any rate, our own scientists have been looking in the wrong way and in the wrong place for magnetic monopoles and fractional charges.

Work such as that of Ehrenhaft and others should be re-evaluated. And, most certainly, the basis of the present electromagnetic theory should be both re-examined and reformulated.

---

*Such as the implication that the free electron is always free to move, and that potential surrounding the electron always possesses a translating gradient when it changes. It does not consider the case where the virtual flux exchange between electron and vacuum increases isotropically, centered on the electron. That, of course, represents an *excited charge* of the free electron.

CHAPTER 5

EXTRAORDINARY BIOLOGY

Kervran's Proof of Biological Transmutation

In orthodox chemistry, one of the strongest dogmas is the stubborn insistence that it is impossible to create another element by chemical reaction. Most chemists also insist that all reactions occurring in living systems are chemical in nature. They believe fervently that chemistry can and must explain life itself.

In the early 1960's, a French researcher named Louis Kervran published work which flew directly in the face of the accepted chemistry dogma. Kervran reported the astounding results of his research showing that living plants were able to accomplish limited transmutation of elements. Kervran was then the Conferences Director at the University of Paris, and his first paper was published in **La Revue Generale Des Sciences,** July 1960.

What was so revolutionary was that, according to the prevailing wisdom of science, you can't transmute elements (permanently change the nucleus) except with enormous energy — certainly not with the microvolts and millivolts (and microwatts and milliwatts) that living systems can muster electromagnetically. *

Rutherford, the British physicist who discovered the nucleus of the atom, had shown in 1919 that you can bombard elements with alpha particles and transmute them. The accepted wisdom of today is exactly the same, except that the physicists have used heavier and heavier "bullets" in their artillery approach. No one has tried a controlled approach, for the catchecism is that you have to use the *wham it harder!* approach.

In other words, to most scientists the whole thing had to be preposterous, and Kervran had to be deluded.

Kervran published further details of his work in a book, **Transmutations Biologiques,** Maloine, Paris 1962. But the initial reaction of most scientists was disbelief and skepticism. Few scientists would stoop to repeating Kervran's experiments, which of course they knew could not work anyway.

---

*Note, however.that since gravity is infolded EM, one can have extremely powerful infolded EM, yet only have miniscule electrical (outfolded) residues. Thus the actual "available power" in artificial biopotentials may not be quite so small after all.

Actually the effect is widespread amongst living systems. As Kervran pointed out, the ground in Brittany contained no calcium; however, every day a hen would lay a perfectly normal egg, with a perfectly normal shell containing calcium. The hens do eagerly peck mica from the soil, and mica contains potassium — a single step below calcium in the standard table of elements. It appears that the hens may transmute some of the potassium to calcium.

Further, if one tests this assumption, it is quickly shown to be true. Hens denied calcium but not potassium, stay perfectly healthy and lay perfectly normal eggs. Hens denied both potassium and calcium will be sickly and lay only soft-shelled eggs. If these sick chickens are allowed to peck only mica — which they will frantically do — everything returns to normal again.

Most orthodox scientists nevertheless remained skeptical or downright hostile.

However, a few other scientists began to repeat Kervran's

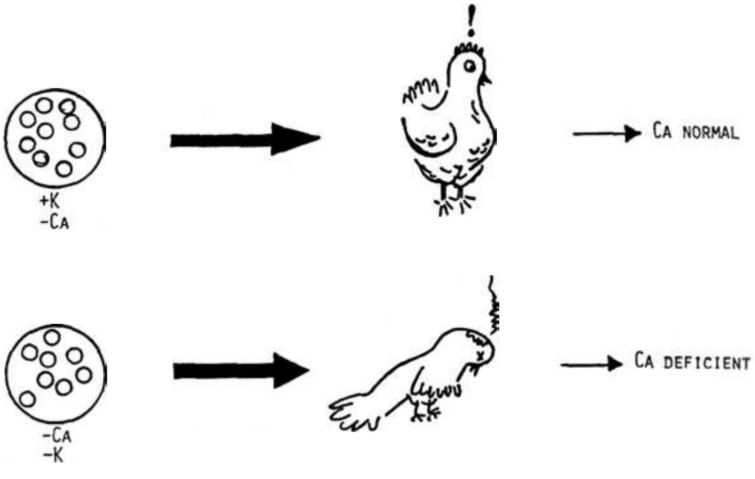

Figure 72. The Kervran effect. A biosystem can accomplish limited transmutation of elements.

experiments and replicate his results. Several of these corroborating scientists were (1) Professor Hisatoki Komaki, Chief of the Laboratory of Applied Microbiology at a leading Japanese university, (2)Professor Pierre Baranger, Head of the Laboratory of Chemical Biology of the Ecole Polytechnique in Paris, and (3) J.E. Zundel, then head of a paper company with a chemical analysis laboratory, and later a chemical engineer of the Polytechicum School of Zurich, Switzerland.

Later work by Zundel was particularly decisive: he utilized the mass spectrometer at the Microanalysis Laboratory of the French National Scientific Research Center, and neutron activation mass analysis at the Swiss Institute for Nuclear Research in Villigen to positively confirm an increase in calcium of 61% to an accuracy of 2%. Such results and instrumentation, of course, removed any doubt that the effect could be due to statistical variation. In the same experiments, the plants increased their phosphorus 29% and their sulphur 36%.

Komaki became head of a research laboratory at Matsushita Electric Company. There he conducted research conclusively proving that microorganisms (including some bacteria and two kinds each of molds and yeast) could transmute sodium into potassium. In fact, he placed a brewer's yeast product on the market that, when applied to composts, increases their potassium content.

Extensive work in the area has been done in the Soviet Union, where results similar to Kervran's have been substantiated.

Thus all doubt (to an open-minded scientist) was removed: living systems are able to change one element into another by some unknown means, using very feeble energy.

A noted French physicist, O Costa De Beauregard, suggested a mechanism for the transmutations, using weak force interactions and advanced waves.

No one — even Kervran himself— thought of negative energy/ negative time interactions. The jury is still out on the actual mechanism, but it is absolutely clear that the transmutation does indeed occur.

The Japanese researchers, having replicated Kervran's astounding results to their complete satisfaction, recommended him to the Nobel Committee for a Nobel Prize for such epochal work. Thus Kervran became a Nobel nominee, though he was not granted the prize.

Kervran has since passed away, leaving behind his books and papers that point to a revolution in chemistry and physics - transmutation of elements at very weak energy.

Biological Transmutation Has a History

Actually biological transmutation — and transmutation of elements (alchemy) in general — has a history, of both results and suppression.

Louis Nicolas Vauquelin, a celebrated French chemist, discovered that chickens could produce more calcium in their eggshells than entered their bodies. Hence they had to be able to "create" the calcium, else their own bodies would have been completely depleted.

One of his contemporaries, however — Antoine Laurent Lavoisier — became the "father of chemistry." Lavoisier laid down the dictum that nothing was created. So chemistry fixed upon the notion that the combinations of elements could be shifted, but the element itself could not be transformed.

Not until the discovery of radioactivity did any crack in this solid wall appear. But still, the basic ideas of chemistry said the element couldn't be transformed chemically. It could only be transformed if one blasted the daylights out of it with an atomic or particle bullet. .

Today most chemists still hold that exact same opinion, unshaken.

To resume: Over a century ago, a chemist named Albrecht von Herzeele proved that germinating seeds somehow transmuted elements in the process. In 1873 von Herzeele published a book, **The Origin of Inorganic Substances,** where he showed research proving that plants continuously create material elements.

Even earlier, in 1822 an Englishman named William Prout had studied chicken eggs in incubation. He found that hatched chicks had more lime (calcium) in their bodies than was originally present in the egg!

Another French scientist named Henri Spindler discovered that, a kind of algae called Laminaria could create iodine.

A German researcher named Vogel had planted cress seeds in a

bell jar. They were fed nothing but distilled water; still, when grown they contained more sulphur than had been in the seeds originally.

Lawes and Gilbert, two British researchers, also found that plants could "extract" more elements from the soil than the soil actually contained in the first place.

Baranger performed thousands of meticulous experiments in plant transmutation of elements. He proved that the transmutations do occur. He also discovered that many things affected the germinating seed transmutation process: the time the seeds germinate, the type of light they are exposed to, the phase of the moon, etc.

None of these experimenters understood the transmutation process used by the living organism. But they proved beyond question that the process existed, and universally occurred.

Surplus-of-Energy Mechanisms Proposed by the U.S. Army

There has also been other very positive support for the thesis that if living systems transmute elements, they can produce a net source of energy in the process.

In 1978 an officially-funded effort of the U.S. Army Mobility Equipment Research and Development Command, Fort Belvoir, Virginia positively confirmed that mechanisms for elemental transmutations could occur in biological systems, from an energy consideration.

The work was performed under the direction of Emil J. York, Chief of the Material Technology Laboratory. Solomon Goldfein was the principal investigator for the effort. Robert C. McMillan, Chief of the Radiation Research Group of the laboratory, provided guidance on matters of physics and nuclear physics.

The abstract of the final report (S. Goldfein, Report 2247, **Energy Development from Elemental Transmutations in Biological Systems,** U.S. Army Mobility Equipment Research and Development Command, May 1978. DDC No. AD A056906.) reads as follows:

"The purpose of the study was to determine whether recent disclosures of elemental transmutations occurring in biological entities have revealed new possible sources of energy. The works of ***Kervran,*** *Komaki, and others were surveyed, and it was concluded*

that, granted the existence of such transmutations (Na to Mg, K to Ca, and Mn to Fe), then a net surplus of energy was also produced. A proposed mechanism was described in which Mg adenosine triphosphate, located in the mitochondrion of the cell, played a double role as an energy producer. In addition to the widely accepted biochemical role of MgATP in which it produces energy as it disintegrates part by part, MgATP can also be considered to be a cyclotron on a molecular scale. The MgATP when placed in layers one atop the other has all the attributes of a cyclotron in accordance with the requirements set forth by E.O. Lawrence, inventor of the cyclotron."

"It was concluded that elemental transmutations were indeed occurring in life organisms and were probably accompanied by a net energy gain."

The researchers also concluded that elemental transmutations occurring in life organisms are accompanied by losses in mass representing conversion to thermal energy, and that such energy probably is a net gain when compared to the amount required to effect the transmutation.

All in all, they concluded that the little cell with its feeble energy does quite well! It's in control of cyclotrons, and cyclotron forces, and direct conversion of mass to energy. Pretty good for a little bitty beastie, wouldn't you say?

Actually, one should point out that, according to nuclear physics, an atom gets a little heavier when it absorbs (usually by means of an orbital electron) a normal "positive energy" photon. That is, the addition of positive energy results in the addition of a little bit of "positive mass."

Negative energy, of course, does a similar thing to the nucleus — except that it adds "negative mass." Thus the nucleus of the atom, when it absorbs negative energy, gets lighter. This is seen in the external world as "loss of mass."

With our present nuclear physics, only positive energy is assumed except in extremely rare cases.

Thus the Army study — which was conducted and controlled by some excellent scientists — worked out a "loss of mass" the way they're trained to.

By adding some positive energy, the nucleus would gain some positive mass. **By adding some negative energy, the nucleus would lose some corresponding positive mass. The conventional physics then would equate this "loss of mass" as the direct conversion of mass to energy. And so it is, only it's conversion to negative energy!**

However, by pointing out the cyclotron mechanism in the cell MgATP, the Army researchers have made a most important contribution.

Note also that the whirling motion may be very much related to Viktor Schauberger's work and to Wilhelm Reich's work. Both of them worked with what they viewed as an unusual kind of living, spiraling energy.

All the orbital electrons of an atom also are whirling around in orbit, in the simplest model. Further, these orbits themselves move and rotate or precess.

Similar orbits and shells occur in the nucleus, at least in some models (several rather independent models are used there for specific things.)

It may be that a whirling, spiraling (cyclotron) energy motion is necessary to connect positive energy to orbital electron (negative charge) shells, and to connect negative energy to positive charge shells in the atomic nucleus. *

Alchemy ***and Unusual Critters***

In ancient times, the old alchemists pursued the dream of making gold. Obviously, if one could do that economically, one could become quite wealthy.

Just as obviously, the kings and rulers of the world took rather **a** dim view of such proceedings. After all, much of their own power rested on their ability to get and control gold. And if some "loose cannon" could make all the gold anyone wanted, then the national **treasury** of the king wouldn't be worth a plugged nickel. And that would finish the king, for he would be powerless.

•The spinning/whirling motion may be viewed as integrating the unzipped vacuum flux virtual vectors into zipped observable force vectors —just as great grandma's spinning wheel integrated fibers into continuous threads.

There are some unorthodox researchers today who take the view that the alchemists were stamped out—*not because they failed, but because they succeeded.*

I subscribe to the same view.

T.H. Moray had a process to "recover finely divided gold from quartzite sands." My personal, strong belief is that he possessed a practical transmutation process. His knowledge and techniques, of course, are still possessed by his sons, and reside through them in Cosray Research Institute, Salt Lake City, Utah.

The possession of such a technical secret may be one of the major reasons why the Morays have met with such intrigue, harassment, and suppression over the years.

To speak further on "making gold," we first have to present some details on some special "critters" that live, but that can't be observed through a normal microscope — even an electron microscope.

In that vein, toward the end of this chapter, we will present some of Royal R. Rife's fundamental discoveries. Pay particular attention to his discovery of "finer" living forms — which today we could only refer to as "living energy, virtual-state forms."

Let's call them **critters** for short.

At one time, when the earth was young and the radiation from the sun was different, conditions on earth were much hotter. Great volcanic activity and fiery eruptions were commonplace and nearly continuous. Huge storms, of size and magnitude undreamed of today, swept the primitive atmosphere. The oceans were frenzied.

Under those conditions, many types of "critters" were highly active. Most of the critters, for example, lived in and worked on the atomic nuclei of matter .

After all, the critters are living, virtual-state organisms. There's a continual exchange between the virtual state (the vacuum, or spacetime) and mass (the observable state). An atomic nucleus is like an island in the "virtual state ocean", and the flux interchange is like waves breaking onto the island and then washing back to sea. The critters live in that ocean, and wash upon, so-to-speak, the mass-islands and interact with them.

In those primal days, many of the present great mineral deposits of the earth were created due to the transmutation activities of the critters.

One kind, for example, lived in copper. In an "energetics" sense, this critter "ate" copper and "excreted" gold, so-to-speak. Much of the gold that occurs in great copper deposits today was formed this way in the old days under primal conditions.

When conditions on earth changed, these little "copper critters" ceased their incessant activity and became dormant, just as viruses can do. But the critters are still there in the copper ore, waiting to be activated.

And activate them you can! You can even get the critters into a solution, and then crystallize them out as crystals.

These crystals are what the alchemists of old called the **philosopher's stone,** with the power to transmute base elements into gold. There are several kinds of philosopher's stones; this kind is for copper.

At any rate, you can then place these special crystals on some copper (and add another thing or two), and restore them to a similar primal environment as of old. That is, heat them in an electric furnace. Blast them with terrible electrical bolts. Bathe them in intense ultraviolet light. That's just a nice, refreshing spring day for the critters!*

That stimulates them and revives them. They wake up after a long sleep — and they're immediately "hungry." So they go right to work on the copper. Boom! In a little bit there isn't any more copper, just mostly gold, with a little other miscellaneous residue thrown in, such as black ruby and silver (in the experiments of one of my close colleagues).

The gold is radioactive when first made. Fortunately, all isotopes of gold are very short-lived: just minutes suffice for the radioactivity to die away. So you wait half an hour and everything's okay.

That's all there is to it.

And if you do that and try to capitalize upon it, your life expectancy is about 24 hours.

I don't know whether or not biological systems, in their Kervran-transmutations at weak energy, deliberately manipulate similar "critters." I suspect, however, that they do, at least to some extent.

---

*Note the probable similar effects involved in the Miller-Fox-Urey experiments in biogenesis

The Cell Also Lives and Functions in the Virtual State

Obviously, to transmute elements the living system has to be able to directly affect and influence the atomic nucleus.

It has been shown that this is a cellular capability, for single-celled organisms can do it.

As we shall see, Rife's work showed that the living cell is connected to at least 16 intemested deeper levels of reality than a relative "point" under an ordinary microscope. Further, all levels are structured and organized.

Think of it! Each one of those levels is to the preceding level as microscopy today is to the normal world. Sixteen levels!

I think it's reasonable to state that the life of the cell is patterned and dynamically structured and functioning all the way into the virtual state; indeed, to very deep intemested levels of the virtual state. That is, it also functions *hyperspatially*.

We shall also see that the mind and thought involve these more subtle physical (though virtual) levels.

Thus the living virtual-state levels are a reality, for Rife proved it.

The living organized structures at each level are a reality, for Rife proved it.

The living ordering and control of dynamic functions on all those levels is a reality, for Rife proved it.

Those living virtual-level parts of the living organism — plant or animal — thus affect, function in, and reside in the atomic nuclei of the material that composes its bodily structures.

Beasties like bacteria and viruses also have living, organized energy structures in multiple levels of virtual state. Apparently, for these more primitive life forms, the virtual-state "energy part" can be separated and pass through a filter, then re-engender the physical form and/or itself cause the disease in a host! At least that is what Rife and other scientists showed.

"Bigger fleas have smaller fleas to bite 'em, And so on, ad infinitum."

Of course the living system can "work on" the nucleus and change it a little bit! If it couldn't do so, it couldn't stay alive and function in there in the first place!

The Kaznacheyev Experiments

Dr. Vlail Kaznacheyev is Director of the Institute for Clinical **and** Experimental Medicine in Novosibirsk.

For 20 years he has been directing highly unusual experiments with twin cell cultures. These experiments are vital to understanding disease and healing on a more fundamental basis than is presently utilized by orthodox medical science.

The Kaznacheyev experiments (several thousand) in the Soviet Union proved conclusively that any cellular disease or death pattern can be transmitted electromagnetically, and induced in target cells absorbing the radiation.

In the experiments, two sealed containers were placed side by side, with a thin optical window separating them. The two containers were completely environmentally shielded except for the optical coupling.

A tissue was separated into two identical samples, and one sample placed in each of the two halves of the apparatus.

The cells in one sample (on one side of the glass) were then subjected to a deleterious agent — a selected virus, bacterial infection, chemical poison, nuclear radiation, deadly ultraviolet radiation, etc. This led to disease and death of the exposed/infected cell culture sample.

If the thin optical window was made of ordinary window glass, **the** uninfected cells on the other side of the window were undamaged **and** remained healthy. This of course was as expected in the orthodox medical view.

However, if the thin optical window was made of quartz, a most unexpected thing happened. Some time (usually about 12 hours) after the disease appeared in the infected sample, the same features of disease appeared in the uninfected sample.

This startling "infection by optical coupling" occurred in a **substantial percentage of the tests (70 to 80 percent). From an orthodox medical view, these results were unexpected and unheard of.**

Further, if the originally uninfected cells were in optical contact

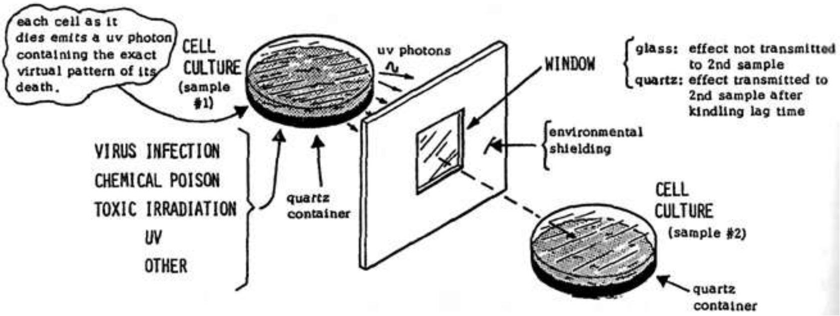
each cell as it dies emits a uv photon containing the exact virtual pattern of its death.
CELL CULTURE (sample #1)
uv photons
WINDOW
glass: effect not transmitted to 2nd sample
quartz: effect transmitted to 2nd sample after kindling lag time
environmental shielding
VIRUS INFECTION
CHEMICAL POISON
TOXIC IRRADIATION
UV
OTHER
quartz container
CELL CULTURE (sample #2)
quartz container

Figure 73. The Kaznacheyev effect. Thousands of experiments proved that (1) cellular disease is electromagnetic, and (2) it can be induced electromagnetically at a distance. Also called the cytopathogenic effect.

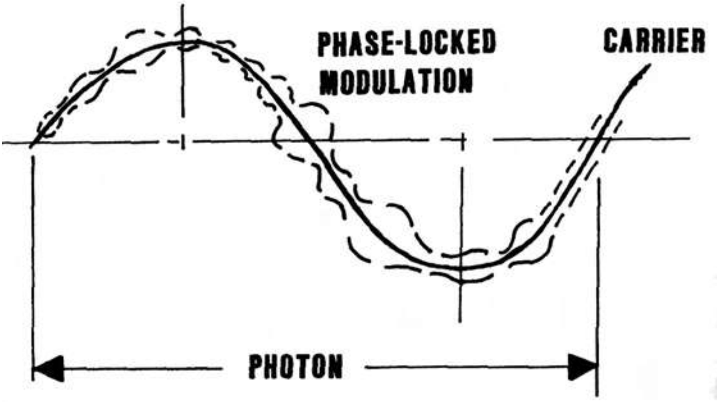
PHASE-LOCKED MODULATION
CARRIER
PHOTON

Figure 74. A photon is one oscillation of an electromagnetic carrier. It may have substructures that are modulations. Such a structure is called a "giant photon," or a "compound photon."

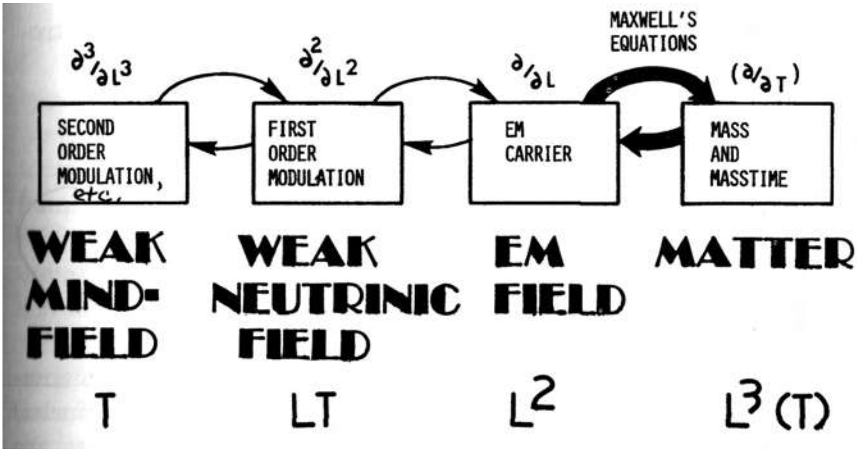
MAXWELL'S EQUATIONS
$\partial^3/\partial L^3$
$\partial^2/\partial L^2$
$\partial/\partial L$
$(\partial/\partial T)$
SECOND ORDER MODULATION, etc.
FIRST ORDER MODULATION
EM CARRIER
MASS AND MASSTIME
WEAK MIND-FIELD
WEAK NEUTRINIC FIELD
EM FIELD
MATTER
T
LT
L^2
L^3 (T)

Figure 75. Nested modulations are the key to hyperspatial engineering.

with the infected cells for 18-20 hours or so, and then were correspondingly exposed (optically coupled) to another uninfected cell sample, symptoms of the infection appeared in this third sample an appreciable portion of the time (20 to 30 percent).

Guided by A.G. Gurvitsch's work that showed that cells give off mitogenetic radiation (photons) that can affect other cells, the Kaznacheyev team sought an answer by looking for photons given off by the infected culture sample as its cells died.

They found that the cells in the infected culture gave off photons in the near ultraviolet when they died. The normal window glass was opaque to these near-UV photons and absorbed them. In that case, the uninfected culture on the other side of the glass was not exposed to radiation by the UV "death" photons from the dying cells, and they remained serenely healthy.

However, the quartz window was transparent to the UV "death photons". When the quartz window was installed, the UV "death photons" passed through it and were absorbed in the uninfected culture on the other side of the window. Most of the time, the uninfected culture which absorbed "death photons" sickened and **died** with the same disease symptoms.

The Kaznacheyev experiments proved conclusively that cellular death and disease patterns can be transmitted and induced electromagnetically.*

Structuring and Charging a Biopotential

Kaznacheyev thus demonstrated that a photon information/regulatory system exists in biological systems due to a continual influx of EM energy from outside the system. That is, the cells of the biosystem are charged with an electromagnetic potential, and additions and changes to the potential are continually received. The cell is thus in minute disequilibrium.

Usually the myriad of continual inputs from the external environment into the cell's potential charge pattern (in its atomic nuclei) may be taken to be potential changes whose substructures are *disordered.* In that case, no specific environmental effect is observed except slight fluctuations without order — a miniscule form of "heating."

However, if a continual *ordered* substructure exists in the input from the external environment into the cell's potential, the cell's potential will gradually "charge up" with that pattern.

An analogy will prove helpful. Imagine an accumulator, a largo pot, that holds a volume of water. Several pipes are connected to tho pot, some are inputs for water coming in, and some are outputs for water flowing in.

Imagine the inputs all containing "blue" water, just in slightly varying shades. The water in the pot is blue, and may slightly rise and fall in level as the input flow rates vary. The water in the pot may also vary slightly in its blueness as the inputs vary. However, it will still be blue.

Now suppose that yellow water starts flowing through one of the input pipes, and at a goodly rate. Slowly the water in the pot will start to turn greenish as a greater percentage of yellow builds up. In other words, the pot slowly charges up with some of the "yellow" charge, in the process acquiring a "green" charge.

---

*We point out that this effect has been investigated in both the infrared and ultraviolet. IR to UV may be taken as a single harmonic interval — an **octave,** musically speaking. The same effect can be reproduced in any other "octave" (single harmonic interval) of the electromagnetic frequency spectrum. The **reversal** of the effect can also be achieved in any harmonic interval. The mechanism for these effects involves the cellular biopotential, Popp's master cellular control system, and the deterministically-tailorcd substructure of photons.

The biopotential in the cell experiments works the same way.

A cell has a biopotential built up, which represents the "nominal equilibrium" of the scalar charge on the cell. This biopotential, being mostly a "sum-zero" of virtual state vectors, is centered in and on the atomic nuclei of the cell, constituting charge patterns in these atomic nuclei. The biopotential extends out of the atomic nuclei, through the electron shells, into and through the molecules, through the internal cell structures and membrane, and outside the cell.

From the atomic nucleus on out through the cell, every layered structure or organization of the cell will layer, structure, and organize the biopotential accordingly.

This organized, structured cellular biopotential is continually receiving "charge patterns" contained in incoming photons absorbed by the cell. The biopotential is also continually exhausting some of its biopotential charge pattern in the photons (heat, light, etc.) that the cell emits.

The Cell's Electromagnetic Breathing

Via structured photon exchange, the biopotential of the living cell thus "breathes in" the virtual state charge structure of its environment, and "breathes out" its own internal virtual state charge structure.

So, in the experiment, the uninfected cells are continually absorbing photons from their surrounding environment, and emitting photons back to it as well. According to our scalar EM view, each photon it absorbs has a substructure that depends upon the part of the environment from whence it came.

These "substructures" are actually patterns of the sum-zeroed virtual vectors comprising the potential of the absorbed photon carriers.

Normally, since a large number of very different substructures are continually being "input" into the cell's potential from the absorbed photons, the substructure of the cell's potential receives an essentially disorganized continual input from the environment. This liquates to the fact that the environment does not normally specifically influence or change the cell's potential with ordered information (organization).*

When a cell dies, it ceases to maintain the bio-dynamics that sustained its artificial potential (that part due to bio-ordering by its organized life processes, above the background level of its "inert matter" potential). The dead cell's built-up artificial potential then "discharges" by emitting a structured photon.**

Since this photon (energy) comes from an *organized potential drop, the virtual substructure of the emitted photon is organized.* The photon, then — among other things — carries the exact organized virtual charge pattern of the dying cell's disease.

We strongly insist on the quantum mechanical view here: All physical changes — chemical, material, mechanical, whatever — at root level are constituted and caused by virtual state interactions, in direct patterns of virtual particle exchanges.

In the full QM view, what's really going on in primary physical reality is just a complex set of patterns and changes in potentials anyway.

The Summed Virtual Structures of Kaznacheyev's "Death Photons" Physically Kindle the Disease

At any rate, the Kaznacheyev experiments showed that the dying cells from the infected culture emitted photons in the near UV that contained artificial (structured) potentials. The virtual-state, patterned-substructures in this photon flux directly represented tho cellular disease pattern caused by the original cell's specific infection.

In other words, as the infected cells died, they emitted "death photons" which contained the **template** pattern of their death condition.

When these "death photons" are absorbed into uninfected cells, their deterministic substructures gradually diffuse into the cell's bio-potentials. Gradually the biopotentials of the new cells are

---

*There may be sufficient ordered input from the environment, however, to have something to do with territoriality in living things, salmon returning to a fixed place to spawn, turtles returning to the same beach to lay eggs, the migration of birds, etc.

Note that this photon is emitted from an atomic nucleus. Hence it is a phase conjugate (time-reversed) photon. It will interact with the biopotential of targeted cells, and thus reach their own atomic nuclei. This is the mechanism for Kaznacheyev's cytopathogenic effect. See particularly C.W. Rietdijk, **Found. Phy , 7(5-6), Jun **77, p.** 351

Table 43. THE LIVING AURA: THE CELL'S ELECTROMAGNETIC BREATH.

- **VIRTUAL EM FIELD**
- **STORES VIRTUAL PHOTONS**
- **ENVIRONMENTAL INPUTS**
- **OUTPUTS BIOPHOTONS**
- **COHERES ORGANISM**
- **TENDS TO STABILIZE**
- **EXPERIMENT ESTABLISHES**

Popp, "Photon Storage in Biological Systems,"
Electromagnetic Bio-Information. 1979

"charged up" with the integrated pattern of the disease.

The master cellular control system of the biosystem is itself a dynamically changing, ordered pattern in the biopotential of the cells, which is centered in the atomic nuclei comprising the cell materials. As the bio-potentials of the cells gradually acquire the "death photon's" substructure pattern, this pattern is also diffused throughout (modulates) the master cellular control system. All the cells in the sample (or in a biosystem) are now slowly charging up with the "death photon" pattern.

As Popp discovered, photons continually "leak out" of the virtual photon master control system of the biosystem. Some of these *leakage photons* are observable photons, **but most are virtual photons.**

Further, they are *structured* photons.

In other words, as leakage photons spill out of the master control system, observable change is now being slowly initiated in the physical structures, biochemistry, etc. of the biosystem's cells — and those changes are in consonance with the integrated "cellular death pattern" of the originally infected cells.

Note particularly that it is already well-known in quantum mechanics/electrodynamics that, when a photon is emitted from **the surface of a dielectric body, the entire dielectric body participates in that emission. If a photon is absorbed on the surface of the dielectric body, the entire**

dielectric body participates in that absorption.

Thus as irradiation by the "death photons" continues, the "death structure" in the irradiated cells increases. It is spread throughout the cell culture by the master communication system, gradually charging the virtual state structure of that system with the death pattern.

Spillage photons from the cellular control system occur throughout the culture. These photons are structured with the death pattern, and gradually affect the cell and its biochemistry physically. The previously uninfected cells thus physically start to acquire and exhibit the symptoms and characteristics of the disease pattern that killed the infected cells.

Electromagnetic Infection Results in Physical Disease

The new cells are now electromagnetically infected and physically diseased.

After all, that is all a cellular disease is in the first place — physical, electrical, and biochemical changes in the normal functioning of the cell.

For a given pattern of changes in the cells, a specific "disease" exists in them.

It absolutely does not matter what causes this exact pattern. If the specific physical pattern is there, the specific disease is there.

Note that **any ghost pattern in the virtual state flux can charge up physical matter — that is, the atomic nuclei of a mass. All that is necessary is that a continual flux of thin virtual pattern continually bathe (irradiate) the mass's atomic nuclei.**

The eventual emergence of this "ghost template pattern" into observable physical reality is called **kindling.** Kindling is charging up one or more atomic nuclei with an integrated virtual charge pattern until the integrated pattern breaches the quantum threshold, resulting in emergence of that pattern into observable physical change.

A Possible Cure for AIDS

One of the things going for the "good guys" and EM defense against AIDS is that cells are a lot tougher than viruses. Thus even non-structured EM signals can be used to effect cures in many disease cases.

In fact, ordinary ultraviolet (UV) irradiation of the blood has already been utilized to cure or control severe infections, including severe viral infections. I am indebted to Dr. William C. Douglass for pointing this out to me, and for permission to reproduce the following information from his important newsletter, **The Cutting Edge, Nov.** 1987, p. 3. The following material is quoted verbatim, with no editing.

"It's amazing what you can find by nosing around in the dusty archives of a good medical library. I came across another remarkable therapy that the AMA and drug industry (or whoever is in charge of supressing non-toxic treatments that work) have shoved down the memory hole."

"Back in 1933, Doctors Hancock and Knott treated a patient dying of septicemia (blood poisoning) with ultraviolet irradiation of the blood.[1] The patient was moribund with a blood stream infection **and** obviously near death. (Remember that this was before antibiotics and there was nothing to lose.) The patient made a complete **and** uneventful recovery."

"Searching further, I found that in 1928 a similar terminal infection was treated by ultraviolet light to the blood. This patient also made a complete recovery.[2] "

"So in 1928, practically in the middle ages, an incurable disease, blood stream infection, was cured with ultraviolet light. With such **a** breakthrough why wasn't it tried again for *5 years?* According to **the** record, *another 6 years* passed before it was tried for a third time.[3]"

"Back in those days infection was the number one cause of death. You can't help but wonder how many lives could have been saved if doctors weren't so resistant to new ideas. Just imagine a cure for **AIDS** being set aside for 11 years. Yet bacterial infections of the blood were uniformly fatal in 1935, just like AIDS is today."

---

[1] Northwest Medicine, 33:200, 1934.

[2] Knott, AM. J. Surg., Aug. 1948, pp. 165-171.

[3] Am. J. Med. Sci., 197:873, 1939.

"Finally, in 1940, 110 cases treated with ultraviolet spectral energy were reported. The results were uniformly good. Between 1940 and 1948 many other conditions were successfully treated, including vein inflammation (phlebitis), polio and asthma. Up to the late 40's *over 40 thousand* treatments were given with ultraviolet blood irradiation."

"And now for the most interesting part. In 1947, Dr. G.P. Miley reported on 79 cases of virus infection.[4] Miley stated that ultraviolet blood irradiation therapy could be relied upon consistently to control an infection of a virus in a safe and efficient manner.[5]"

"AIDS is a virus. AIDS-II is a virus (the HTLV-IV leukemia and lymphoma now sweeping the world). Remember that these killer viruses are within the cell and any chemical agent that enters the cell to kill the virus will often kill the cell as well. But ultraviolet irradiation *kills the virus without harming the cell."*

"A fine piece of crystal can be shattered by exposing it to just the right frequency. You can be standing in the room and the energy from that frequency won't harm you in the least. Viruses have the same characteristics, and so, in my opinion, frequency irradiation of the blood in the ultraviolet range is our greatest hope for curing AIDS."

"But the treatment is simple, safe, inexpensive and unpatentable. That doesn't bode well for its future, at least until a few *senators* get AIDS."

The Mirror Cytopathogenic Effect and Factors Influencing It

The cellular disease induction effect was called the **mirror cytopathogenic effect** (CPE for short) by the Kaznacheyev group. Mirror CPE appeared only when the quartz or mica window was no thicker than 0.8 mm. A. F. Kirkin also duplicated the experiments using a thin plexiglas window.

There are conditions which enhance the effect, and others which inhibit or degrade it. Irradiation of the detector-culture with a low dose of UV prior to its optical contact enhances the effect, increasing it to certainty (99-100%). Increasing the temperature to 38.5 degrees centigrade also enhances the effect (from 37% to 90% for example).*

---

[4] Rev. Gastroenterol. 15 271-277, 1948.

[5] Am. J. Surgery, Aug. 1948, pp. 170.

A necessary condition for the success of the experiment is the rotation of the holder with its two optically-coupled samples at a rate of about 24-25 revolutions per hour. Optical contact between the inductor and detector cells for a minimum of 4-6 hours is necessary, after which the cell cultures can be separated. A longer contact time is necessary for complete development of the irreversible effect.

Both cultures must be maintained in complete darkness throughout the experiment. Use of the detector as a new inductor in a successive state reduces the effect by 20-30%. Three or four such stages is sufficient to eliminate the effect.

There is a seasonal variation in the results. In more than 15,000 experiments, monthly variations and daily variations were noted. (The present author's interpretation of this is that it is due to the monthly variations in the virtual photon substructure input from the moon to the substructure of the cell's bio-potential. The daily variation is due to the daily variations in the virtual photon substructure input from the sun to the substructure of the cell's bio-potential.)

Negative results appear more often in winter. (The present author's interpretation of this is that it is due to the fact that the scalar potentials of the earth and the biopotentials of each living cell on earth are lowered in winter by the weaker flux from the sun.)

Effects are correlated with the polarity of the interplanetary magnetic field. Negative polarity of the field usually precedes the appearance of mirror CPE. (This is because of the positive nuclei — which prefer one direction of the magnetic field over another). Disturbance of the geomagnetic field several days before a culture planting also results in enhancing the mirror CPE effect. (Disturbing the geomagnetic field provides a "dithering magnetic disturbance" in the atomic nuclei which "livens" them. Consequently their readiness to charge-up and emit structured virtual state charges is increased).

Kaznacheyev further discovered that the Sun's activity and the

---

•This is very important. Preconditioning the cells by "dithering" them in the frequency band **of** interest, or in a subharmonic band, "livens" the cells for developing the irradiated pattern. Thin is similar to a dither voltage placed on a missile fin, making it much easier and quicker **for** the fin to move when an actual order is placed on it.

Earth's magnetic field greatly affected the results of his experiments. Large flashes on the sun seem to inhibit the effect. (Such flashes cause substantially increased irradiation of the atomic nuclei by sun-emitted substructures, charging them mostly with this disordered substructure pattern and literally "burying" the disease structure several decibels below it.) In a season of active sunspots, the mirror CPE effect becomes highly unstable. (The sun emissions are sporadically jamming the effect.) Under active sun conditions, the effect varies from 90-100% on some days to complete absence on others.

Some Biological Warfare Implications

The Soviets reported detecting near-ultraviolet photons — bioluminescence — as carriers of the death/disease pattern.

However, scientists at the University of Marburg in West Germany also duplicated the effect in the infrared. This shows that bioluminescent photons in the near UV and in the IR can definitely carry "disease and death" information between cells. Further, integrating a continuing input of such photons coherently integrates the disease or death pattern from the virtual state into the observable state.

Note also that portions of the infrared spectrum are a subharmonic of the near ultraviolet. Harmonics are well-known in nonlinear oscillator theory, and biological systems are filled with nonlinear oscillators. It may be that harmonics and subharmonics are directly involved in the death pattern.

If so, the induction of such "death patterns" upon normal electromagnetic carriers is directly indicated. For example, modulations covering several octaves in the region of 10 gigahertz and above might be constructed that are the analogues of some particular cellular disease. This modulation pattern could then be added to a common microwave carrier — say in the communication band, from 3 to 30 megahertz. Say, that is, to something like the giant Soviet Woodpecker "over-the-horizon radar" signals as carriers.

In that case, a large population could be bombarded, even on the other side of the earth, with "death photons" whose virtual state substructures carry the particular disease pattern. With sufficient time, many of the targeted persons would develop the disease.

Note that, even if the power and/or irradiation time is reduced so that the absorbed "death photons" are insufficient to actually kindle the disease in the targeted population, a heightened change in the substructure of the biopotentials of the cells of the targeted persons is still accomplished.

In that case, a precursor pattern — a predisposition for that disease — exists in the targeted persons.

If the actual disease agent is now loosed on that population, the agent will be far more **infectious and lethal than it otherwise would.**

In this way, even diseases which normally do not kill or seriously debilitate the infected person can suddenly become very lethal agents indeed.

Influenza, the common cold, etc. can become devastating killers if the exposed population has been electromagnetically "pre-conditioned" for enhanced susceptibility.

What Kaznacheyev Hid: The Role of Phase Conjugation

If cellular disease can be electromagnetically induced, can it not be electromagnetically corrected or healed?

If one could time-reverse the exact signal structure (the information) that kindled the effect, and bombard the diseased cells with that reversed pattern, would not the cell deviate back to "normal" and be healed?

The burning question as to whether cellular disease conditions can be corrected by time-reversed disease signals must certainly have occurred to the Soviet experimenters.

It is highly significant that they did not openly publish those results.*

As we have explained in the sections on phase conjugation and scalar electromagnetics, there are really two major kinds of photons:

---

*Recent information indicates the strong connection of Kaznacheyev with the Institute of Physiology and Biophysics and the Frank Institute in Pushkino, just outside Moscow. Since these institutes are deeply involved in microwave and coherent microwave "directed energy" weapons, it is highly probable that the Soviets are applying Kaznacheyev's "death photons" to microwave weapons — such as the Woodpecker transmitters. If so, obviously they would develop phase conjugate countermeasure signals as well.

(1) the "normal" photon carries positive energy and positive time. (2) the "time reversed" or phase conjugate photon carries negative energy and negative time.

Further, the Soviets certainly knew all about phase conjugate signals. After all, they discovered and developed the effect. We discovered it only from the open Soviet scientific literature!

Let us assume that the "death photons" in the mitogenetic radiation emitted by the dying cells are ordinary photons. Their virtual state structures (in positive observer time) are exact "templates" for the disease pattern.

Now suppose we detect the "death photons" with a phase conjugator, which by definition will produce a time-reversed counterpart to the input signal detected. In other words, the death photons are allowed to strike a phase conjugate mirror (PCM). Time-reversed counterpart photons — carrying the exact time-reversed template of the death pattern — will be created and emitted by the PCM.

These newly emitted photons now carry the exact "healing pattern" for that specific "death/disease pattern that was received and detected."

Further, if we "pump" the phase conjugate mirror, we can greatly amplify the output pattern, and hence greatly increase the healing pattern!

If one records the pattern of the "death photons" for a specific disease, one could of course modulate that pattern upon ordinary photons/signals — such as the Woodpecker signals — and accomplish disease induction or precursor conditioning.

By phase conjugating the pattern of the "death photons," one can produce an exact antidote. One can modulate this specific healing pattern upon ordinary photons/signals — such as the Woodpecker signals — and accomplish healing induction for that specific disease.

In other words, one can create the healing pattern — the antidote, if you will, for any biological warfare agent. Cancer, leukemia, AIDS, viral diseases, bacterial diseases, whatever. One can create the antidote within minutes after the first symptoms of the disease or death pattern appear.

One can then simply add the negating (healing) signal to power line signals, television and radio signals, special transmitters, etc. —

and immediately start to "administer the antidote" to the irradiated population one wishes to protect. *Now one can see why the Soviets are so ready to expose the entire world to something like AIDS. It doesn't represent a real problem to them, the instant they decide to negate it.*

So they can devastate the rest of the world, with the assurance that their population is safe.

They can allow some of their own people to develop AIDS — and even some to die of it — as a deception plan to delude the West while Western populations are succumbing en masse.

Then they can snatch their own population right back to health, from "the brink of the grave," so to speak.

Our government must immediately develop the same capability. It is straightforward. As weapons and counterweapons go, it is enormously cheap. It can be immediately and widely implemented. **And** it can protect our population against AIDS or any other biological warfare strike by the Soviet Union.

We can save our people from the AIDS knockout already unleashed upon us by the Soviet Union.

First let us do that. *Then* let us negotiate.

Remember this: **You can negotiate with the Russians only from a position of strength. If you are weak, they will bury you.**

If we do not immediately develop this biological warfare countermeasure, we are already as good as dead.

Popp's Master Cellular Communication System

Dr. Fritz Albert Popp has already discovered and pointed out the "virtual state" master communication system that controls all cells in the body, and all their functions.

Based on a thesis derived to best fit experimental results by Ruth nnd others, Popp postulates that biological systems generally have the capacity to store coherent photons that come from the external world.

In other words, the biosystem is open to environmental communication and exchange.

He has shown that the cell population is in a quasistationary state that is far away from thermodynamic equilibrium, as pointed out by Ilya Prigogine.

Popp also concludes from his analysis that ultraweak photon emission within biological systems can influence chemical reactivity. In fact, his analysis strongly implies that "ultraweak" photon intensity can regulate the whole cell metabolism and related phenomena.

The cell takes up photons from external radiation. This includes both "observable" photons and "virtual" photons. Since it stores virtual photons, it stores charge, or biopotential changes. Since its stored virtual photons may be coherent virtual photons, it effectively "polarizes" or structures its stored photon charge, hence its biopotential.

The cell emits "spillage" photons — both coherent and incoherent — from its stored potential.

Although Popp only uses conventional "unstructured" photons in his analysis, he shows that, at the molecular level, there is a stationary equilibrium, as far as photon storage and emission are concerned, between the molecular photon traps, the cell population, and the external world.

It follows that *coherent* photon/charge inception from the external world can directly and precisely influence the cell's biopotential, hence its functioning and control, by information input.

Incoherent photon inception, on the other hand, can only grossly affect the cell, such as by heating or sporadic effects.

In his "Photon Storage in Biological Systems," Popp points out the master cellular communication and control system as follows:

"The photons which we have measured can be seen as a sort of "waste" from a virtual electromagnetic field with a high coherence. This field has a tendency to become stationary over the whole organism."

After additional analysis, he adds:

"....Consequently, biological systems must exhibit 'holographic' properties to an extremely high degree. The successful trials in

finding 'pictures' of various organs in each other organ, such as the ear, the hands, the eyes (acupuncture, iris diagnosis) support these conclusions. Our assumption that the entire genetic information of the DNA is stationarily delocalized over the body in form of genons may be seen as a further striking example."

"From this we can easily deduce that pattern recognition, as, for example, repair mechanisms and immunity, depends finally on the coherence of the photon field within the body."

Finally, Popp states a most important conclusion:

"...In medicine **new aspects have developed, and not only for cancer problems. Diseases in general can possibly be understood in terms of electromagnetic interactions within the organism."**

Scalar *EM* ***Comment on Popp's Communication*** *System*

Popp and his colleagues have produced most important work and results indeed. They only need to add the impact of the zero-summed/multiplied electromagnetics (electrogravitation).

As we cover in this book, the biopotential of the cell is rooted in the nuclei of the atoms of the cell's constituent materials. To be sure, every internal physical structure of the cell correspondingly "levels" and structures the biopotential. The overall cellular communication system is actually the exchange of "leakage" photons — both observable and virtual — throughout the overall biopotential of the organism.

Further, going beyond Popp's work, both the biopotential and the leakage photons have extensive, complex internal substructures. Leakage and intercommunication occurs laterally at all levels of the biopotential, and vertically among cells and substructures.

The master cellular control system's primary electrical conductivity path is not through the electron shells of the atoms, but is through the nuclei-to-nuclei scalar EM "biopotential levels" pathway.

With scalar EM methods, organized signals (signals with specific internal nonzero vector EM waves, but which externally sum to z**ero** vector resultant E and H fields) can be constructed for essentially any specific purpose. This includes "killing" a cancer or leukemia cell, destroying a virus, changing the DNA, etc.

This approach can directly reach and manipulate all immune and repair system functions.

The entire biochemistry and functioning of the cell — including its genetics — **is totally engineerable.** The Soviets have long known this, and have long since done it.

Further, a specific "charge pattern" of desired specific immunity (antibodies, etc.) can be designed and used to "charge up" the nuclei of the biosystem. This charge is then maintained by the system to provide permanent immunity. Thus one can develop, for example, an "electromagnetic innoculation" for AIDS, one for cancers and leukemias, etc.

Since the cellular control system is holographic, the "charge pattern" of immunity resides in every cell, including the blood cells.

Injecting a drop of blood from a scalarly immunized animal into another non-immune animal carries the scalar EM immunity pattern into the new animal. That charge diffuses throughout the overall biopotential of the organism, and the charge pattern activates the animal's immune system, including causing it to produce antibodies — according to the EM-transferred antibody template.

Antoine Priore demonstrated this effect numerous times. Thin was one of the great mysteries that confounded the orthodox members of the French Academy of Sciences.

The French Academy did not know of scalar electromagnetics, the cellular biopotential rooted in atomic nuclei of the cellular material, the cytopathogenic effect of mitogenetic radiation from diseased and dying cells, phase conjugation, and phase conjugated electromagnetic healing.

It is little wonder they did not comprehend the operational healing mechanism of the Bordeaux cancer-curing machine of Antoine Priore!

A New View of the Nature of Mind and Thought

The reason that Western science has not discovered what mind and thought are is simple: They have only built instruments and tools to look where the mind and thought *are not to be found.*

Let's look where they are to be found.

For example, consider the ionic discharges in and across the tremendous numbers of synapses of the human nervous system, and the slower discharge and migration of ions across cellular membranes, etc.

Considered as a single biological ensemble, these are much like a vast array of continual and continuing spark discharge vectors, slow charge current vectors, etc.

Overall, within the macroscopic space occupied by the body, these electromagnetic vectors sum almost entirely to a zero vector resultant. Only a tiny nonzero vector residue remains.

However, this vector zero summation has incredibly rich infolded signals, channels, and dynamic relationships (structure) within it.

Modern science/medicine measures the small nonzero electromagnetic residue (the remaining weak E and H fields) and tries to ascertain where and how mind and thought are accomplished, by studying that residue.

However, the residue E and H field remainders are simply the "garbage thrown out" of the glorious engines dynamically operating inside the vector zero summations (and multiplications).

The E and H field residues are the wastes or exhaust byproducts. They are not the functions of the mind/thought process itself; instead, they are the spillage or leakage from those functions.

Take a "frozen time" snapshot of the components of the zero vector summation. Regard the highly complex, infolded structure or pattern these components form.

Now take a second "frozen time" snapshot of the components of the zero vector summation, a very, very short time later. Subtract **the** previous "frozen pattern" from this second "frozen pattern."*

The difference or "delta" between the two patterns represents a myriad of thoughts, hence it represents the contents of the "thinking mind," in the most general sense.

The "mind" is the overall functioning and changing of the entire pattern's substructure, and the ability of an organism to have such **and** do such. The "ability of the organism to do such" requires at least

*Similar to the magnificient "double **exposure holography"** work of Dr Robert Powell.

two more nested levels of virtual state—two more hyperdimensions.

In simplest Kaluza-Klein theory, all those electromagnetic component vectors are in the fifth dimension. Hence they are hyperdimensional.

The zero-vector system may well have further infolded, internested levels of zero-vector systems. Systems within systems. These are hyperspatial, going into the 6th, 7th, 8th, etc. dimensions. Rife's microscope, for example, could reveal some of these levels; using evanescent waves, it could resolve some 16 ever-deeper levels of dynamic energy structure.

Everything driving the component vectors themselves is higher dimensional.

"A thought" is the exact change of a localized pattern (localized zero vector summation whose components are changing but remaining zero summed) from one moment to the next.

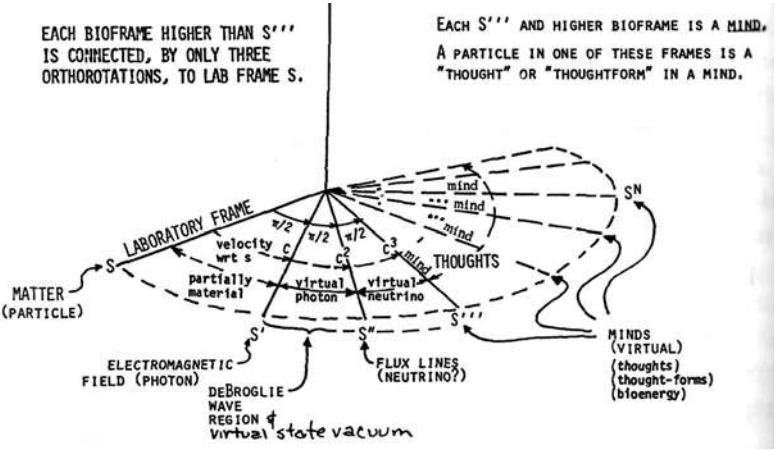
EACH BIOFRAME HIGHER THAN S''' IS CONNECTED, BY ONLY THREE ORTHOROTATIONS, TO LAB FRAME S.
EACH S''' AND HIGHER BIOFRAME IS A MIND.
A PARTICLE IN ONE OF THESE FRAMES IS A "THOUGHT" OR "THOUGHTFORM" IN A MIND.
LABORATORY FRAME
π/2
π/2
π/2
velocity wrt s
C
C²
C³
mind
mind
mind
mind
S^N
THOUGHTS
S
MATTER (PARTICLE)
partially material
virtual photon
virtual neutrino
S'
S''
S'''
ELECTROMAGNETIC FIELD (PHOTON)
DEBROGLIE WAVE REGION & virtual state vacuum
FLUX LINES (NEUTRINO?)
MINDS (VIRTUAL)
(thoughts)
(thought-forms)
(bioenergy)

Figure 76. Hyperframes, vacuum, virtual state, minds and thoughts.

Thought is always a pattern change, of the components inside an EM vector zero summation.

Most thoughts are totally "unconscious" (multiple simultaneous or "parallel" thoughts). A few are conscious (serially ordered, singularly considered and processed).

The unconscious mind is totally a parallel processor (many things/thoughts at once).

The conscious mind is totally a serial processor (only one thing/ thought at a time).

Most persons never take the time and trouble to reflect and notice that, consciously, they only can perceive one single thing at a time. Of course, the conscious mind is so rapid that by habit they assume they can perceive many things at once.

Since the conscious mind cannot discriminate the "multiple images in the slide projector at once" of the unconscious mind, that's what makes it unable to "be conscious" of the contents of the "unconscious."

In fact, the unconscious mind is totally conscious — it's just **multiply conscious "simultaneously."** (Which sheds some interesting light on why one can develop multiple personalities, for instance).

When the conscious mind "regards" the unconscious contents, it sees "something" which has "multiple meanings at once."

That's what we call *symbolic.* A symbol is something which can have many meanings at once.

That's why the unconscious content always "symbolizes" something when it's trying to communicate to the conscious mind. It's necessary to interpret the symbols to understand what the message being communicated is.

That's why dreams, for instance, are symbolic. And why it usually takes a trained psychiatrist or psychologist to properly interpret the symbolic representations manifested in mental disorder, and get at the underlying cause being symbolized.

With deliberately constructed scalar EM, any or all of the conscious/unconscious portions of the mind can be available for engineering and control/change. Eventually you will be able to put contents of the mind and memory on a video screen if you wish to.

It will also be possible to interact with the mind's contents electromagnetically. Mental illness will be treated directly, on an engineering basis.

Of course this opens up the frightening possibility of hostile misuse to the detriment of an individual. It will even be possible to change or erase his or her very personality itself. We pray, of course, that such a powerful tool will not be utilized this way, but will be used to heal humans, not hurt or kill them.

Unfortunately, Lisitsyn's work reveals that the Soviet Union has long since applied energetics (scalar electromagnetics) to mind control and mind engineering, including deciphering the genetic code operation. Controlled induction of images and sensations inside the brain — where they are processed as if self-originated — has been reported by the Soviets. Indeed, they have reported being able to control whether or not the induced material should rise to conscious awareness, as well as when it should rise to consciousness.

For years, U.S. intelligence analysts and U.S. scientists simply did not believe that electromagnetic signals could directly influence consciousness.

However, they were forced to reevaluate that position when a Soviet medical machine — the LIDA device — was openly obtained and evaluated.

The LIDA machine — a small unit somewhat larger than a briefcase — has been used in Soviet medical facilities to treat humans for decades.

The device uses a 40 megahertz EM carrier and very complex waveforms (signal modulations with complex mixtures of frequencies, phases, etc.)

When exposed to it, a person is caused to gradually fall into a trance-like, catatonic-like state within a few minutes. That person becomes very still and very quiet.

It was reported that the machine was tested upon a cat and had the same effect upon it.

One American scientist stated that the device had actually been used in North Korean "brainwashing" of U.S. prisoners of war in the early 1950's during the Korean War!

If so, one can see just how long the Soviets have achieved success with electromagnetic biological warfare/medical treatment techniques.

Reportedly, prisoners exposed to the machine were rendered unable to psychologically resist their harsh brainwashing interrogation.

But to return to the brain activity, with its innumerable, zero-summed minute EM vectors, and its extremely rich and complex deterministic substructure.

Note that the ensemble of sum-zeroed vectors forms potentials. These potentials — i.e., the scalar EM — penetrate to the atomic nuclei in the brain and body. Internal changes in the potentials' substructures (i.e., thoughts) also penetrate to the nuclei.

The nuclei continually "charge up" (the particles change their states) to these impressed potentials, including the "potential" or "charge" of each component of the impressed potential. That is, the nuclei charge up with both dynamic overall potential (mind) and thought (individual substructure changes).

Notice that **mind, thought, and memory reside in, and are recorded in, the atomic nuclei. At many virtual state levels. In many hyperdimensions.**

Thus the functioning mind and bio-control systems — including Popp's master cellular communication system, the immune control system, the acquisition control system, and the repair control system - all "reside" in the nuclei as dynamically interacting, patterned potentials or "patterned charge," complete with resonances and frequencies, etc.

We can regard the mind and its interactions as physical, in the extended sense we have briefly developed here.

Everything "mental" can be directly interfaced with and engineered.

One day, for example, "education" will be by direct loading of patterns into the brain/body/cellular biopotentials, much as we now place a diskette in a disk drive and load the computer memory. And then everyone can be educated, in the widest range of skills and to the highest degree ever dreamed of.

Kindling, Life, Mind, and Negentropy

The mass of the atomic nucleus can be charged up with a particular dynamically structured biopotential.

In scalar electromagnetics, I use the term **kindling** to refer to the effect of charging up mass by a specific structured charge pattern to form such a structured potential in the nuclei.

Note that we are actually and physically restructuring the nucleus itself. That is because, if the structure of the potential in the nucleus is changed, local spacetime of the nucleus is polarized in complex electromagnetic structures. The electromagnetic currents, actions, and charges of the affected nucleus adjust automatically to this structure or "grid." *

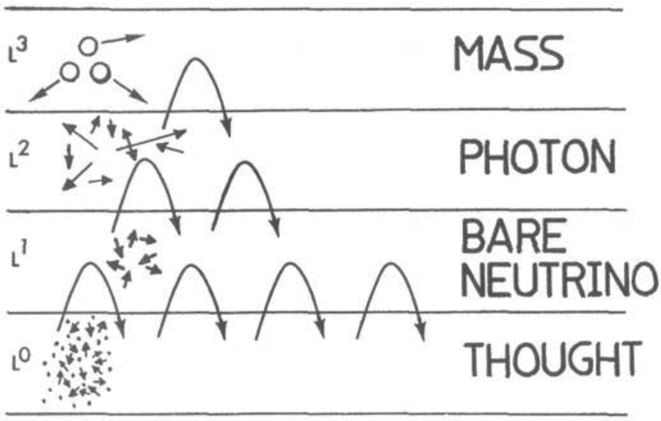

Figure 77. Internested levels of virtual state vacuum contain mind and thought. **The** vacuum also contains their interaction toward and upon mass.

---

*See Richard E. Prange and Peter Strance, "The semiconducting vacuum," Am. J. **Phy.** 52(1), Jan. 1984, p. 19-21. The vacuum may be regarded as a semiconductor. This semiconductor, particularly in the region close to the nucleus, may be manipulated by subjecting it to external fields, doping, etc. Extra energy density can even be extracted; see H. Paul and R. Fischer, "Comment on How can a particle absorb more than the light incident on it^{7}'" Am. J. **Phy. 51(4),** Apr. 1983, p. 327. For the involvement of negative time/negative energy, see C.W. Rietdijk, **"How** do 'virtual' photons and mesons transmit forces between charged particles and nucleons?" **Found. Phy** 7(5-6), June 1977. p. 351-374.

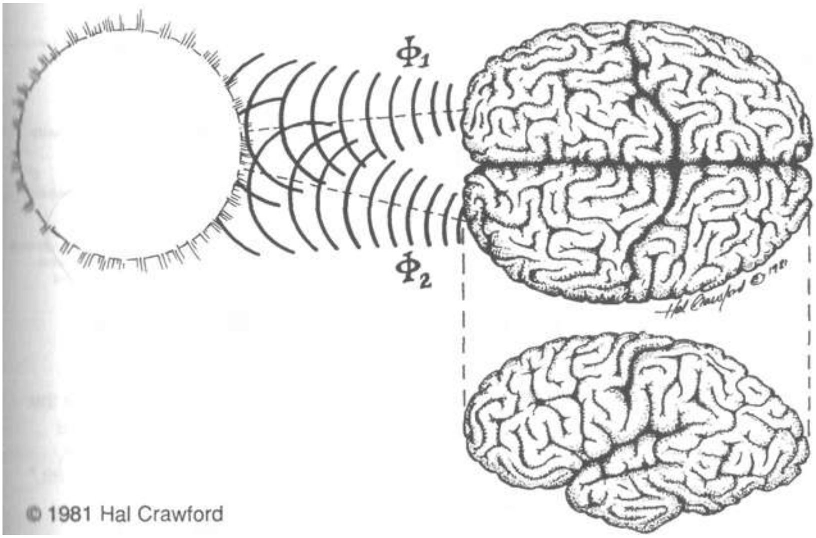
Φ1
Φ2
Hal Crawford © 1981
© 1981 Hal Crawford

Figure 78. The human cerebral cortex is a natural scalar interferometer. It is a virtual state tuner, processor, and transmitter-receiver. It also can produce...and control to some extent...phase conjugate energy and phase conjugate waves.

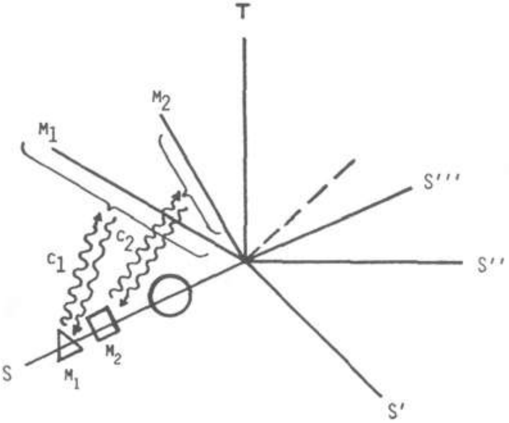
T
M2
M1
S'''
C2
C1
S''
M2
M1
S
S'

Figure 79. Consciousness and life. These phenomena refer to/constitute a deterministic hyperspatial (virtual state) coupling between mass and an ordered hyperworld.

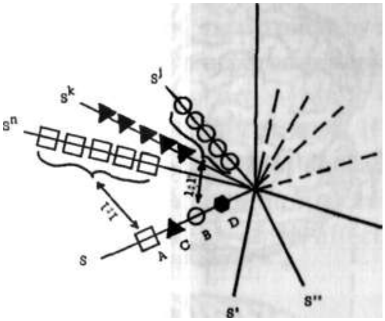
S^j
S^k
S^n
1:1
1:1
S
A
C
B
D
S'
S''

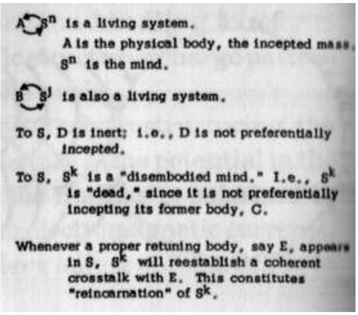

A ⇄ S^n is a living system.
A is the physical body, the incepted mass.
S^n is the mind.

B ⇄ S^j is also a living system.

To S, D is inert; i.e., D is not preferentially incepted.

To S, S^k is a "disembodied mind." I.e., S^k is "dead," since it is not preferentially incepting its former body, C.

Whenever a proper retuning body, say E, appears in S, S^k will reestablish a coherent crosstalk with E. This constitutes "reincarnation" of S^k.

Figure 80. Life and death. If the tuning or coupling between the mindworld and the tuned, structured mass body is broken, that is physical "death." Discharge of the structured cellular potentials then produces "death photons." Discharge of the overall, structured biopotential of the body constitutes discharge of the "living spirit."

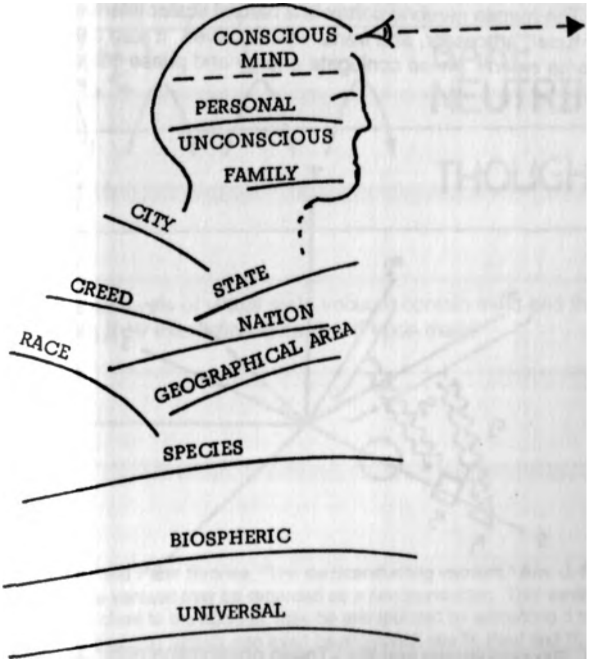
CONSCIOUS MIND
PERSONAL
UNCONSCIOUS
FAMILY
CITY
STATE
CREED
NATION
RACE
GEOGRAPHICAL AREA
SPECIES
BIOSPHERIC
UNIVERSAL

Figure 81. Layers of unconsciousness intersect hyperspatially.

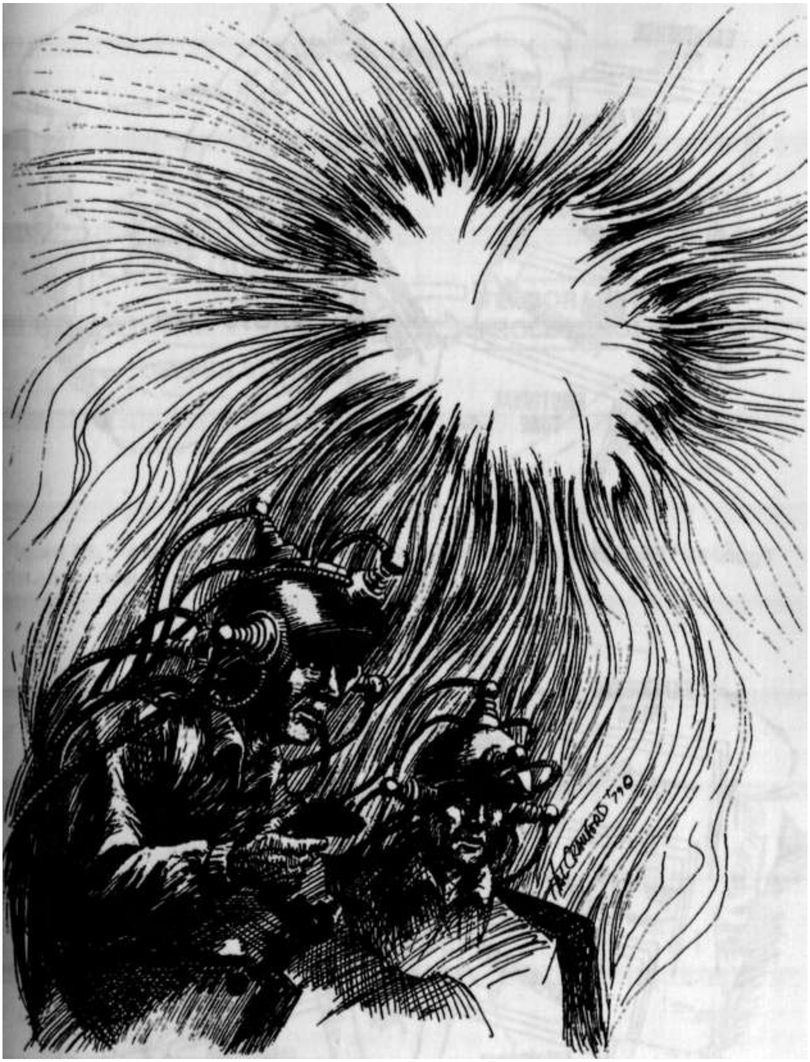

© Hal Crawford 1979

Figure 82. Scalar electromagnetics can directly interact with the various levels of human mind and personality.

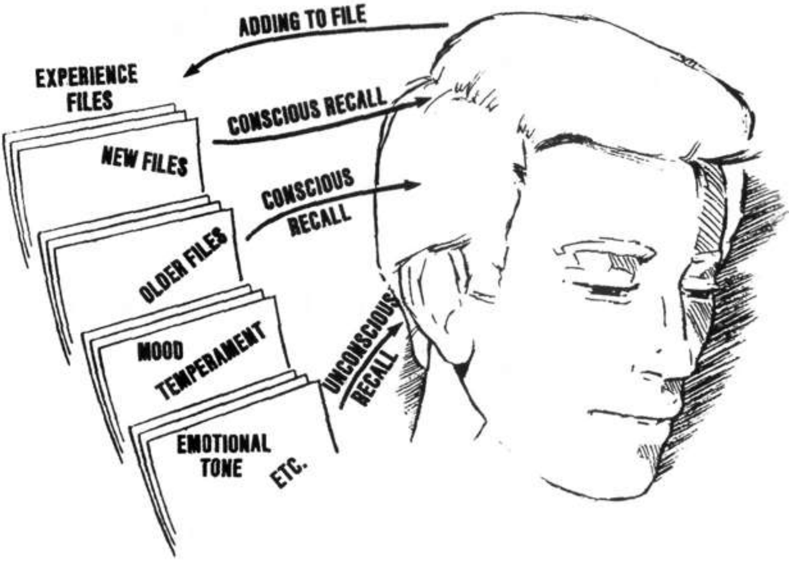
ADDING TO FILE
EXPERIENCE FILES
NEW FILES
CONSCIOUS RECALL
CONSCIOUS RECALL
OLDER FILES
MOOD
TEMPERAMENT
UNCONSCIOUS RECALL
EMOTIONAL TONE
ETC.

Fgure 83. What personality is.

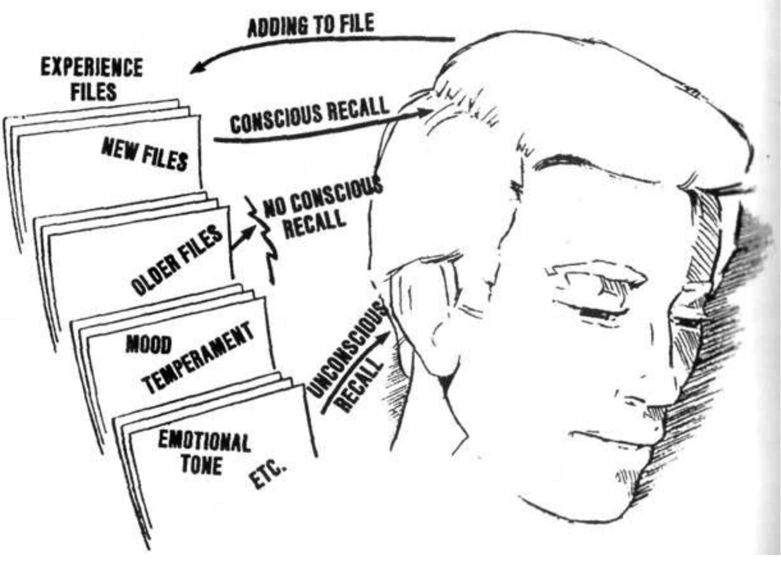
ADDING TO FILE
EXPERIENCE FILES
NEW FILES
CONSCIOUS RECALL
NO CONSCIOUS RECALL
OLDER FILES
MOOD
TEMPERAMENT
UNCONSCIOUS RECALL
EMOTIONAL TONE
ETC.

Figure 84. Jamming the "older files recall" process produces simple amnesia. If the recall process is then restored, the person regains his or her former memories.

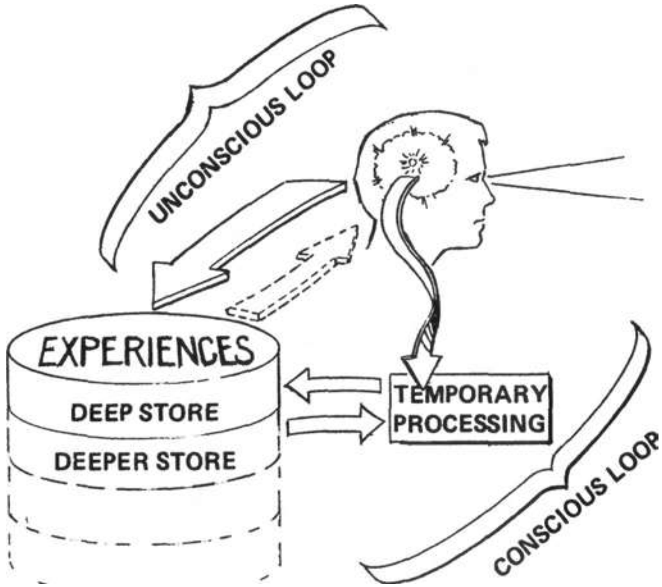
UNCONSCIOUS LOOP
EXPERIENCES
DEEP STORE
DEEPER STORE
TEMPORARY PROCESSING
CONSCIOUS LOOP

Figure 85. To sustain one's personality, continual functioning is required. All channels must be functioning.

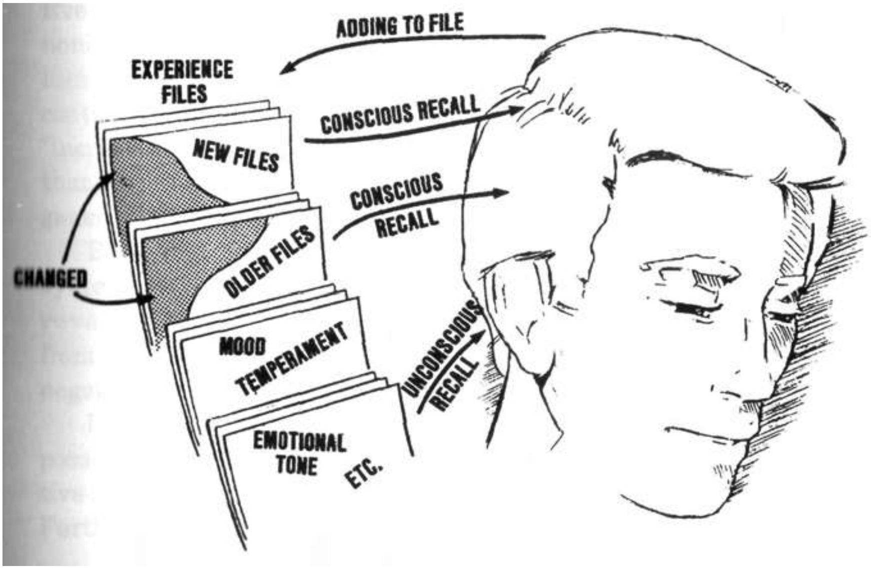
ADDING TO FILE
EXPERIENCE FILES
CONSCIOUS RECALL
NEW FILES
CONSCIOUS RECALL
CHANGED
OLDER FILES
MOOD
TEMPERAMENT
UNCONSCIOUS RECALL
EMOTIONAL TONE
ETC.

Figure 86. If the files themselves are changed, the personality itself is changed. It is possible to alter who or what a living person is.

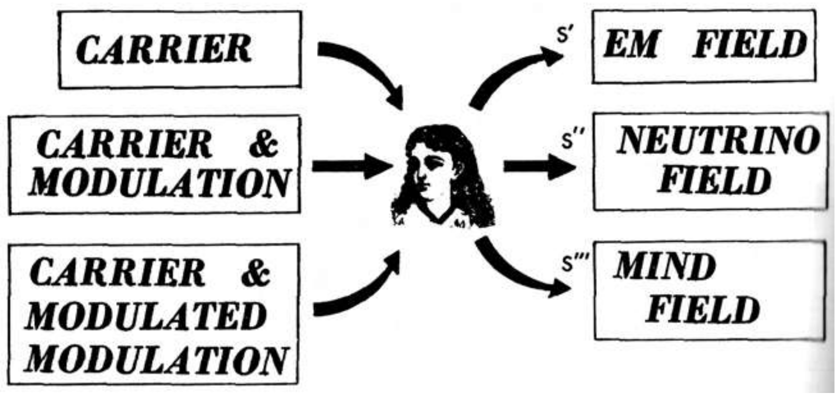
CARRIER
CARRIER & MODULATION
CARRIER & MODULATED MODULATION
S′
S″
S‴
EM FIELD
NEUTRINO FIELD
MIND FIELD

Figure 87. The mind and personality can be affected — and changed or controlled from a distance. The Soviets have developed at least a substantial part of this technology.

Table 44. LISITSYN'S REPORT

- **THEORY DEVELOPED & FITTED**
- **HYSTERESIS MEMORY LOOP**
- **CONTROLLED INDUCTION**
    - **IMAGES**
    - **SENSATIONS**
    - **PREDETERMINED EMERGENCE**
- **23 EEG BANDS**
    - **UP TO 8.1×10^{20} HZ**
- **11 INDEPENDENT CHANNELS**
- **BRAIN CODE BROKEN**
- **44 DIGITS OR LESS**

Thus by tailoring the specific scalar waveform and wave substructure with which we irradiate a mass, and utilizing phase conjugate (time reversed) waves, we can directly engineer the nucleus itself when sufficient charge structure has been built up in it to form a structured potential that is powerful enough to alter/ **shape** the nuclear processes.

We can, for example, transmute the nucleus into another form. Here an isomer form is the easiest. An isomer for an atomic nucleus (a specific isotope) has the same number of nucleons (protons and neutrons) overall as that isotope does, but a different number of protons. So it's a different element, but very specially related to its isomeric "brother."

All that it takes to change a neutron into a proton or vice versa, **is** to "flip" a single quark (subparticle of the nucleon).

Living systems can do that and transmute elements, to a small degree.

For example, living systems must use phase conjugation, negative time, and negative energy. To do so, they must function nonlinearly. That is the only way they can defeat the otherwise inexorable second law of thermodynamics: that all (linear) processes continue to more and more disorder as time passes. (That's called "increasing the entropy (disorder)." Unless a biosystem could defeat **that** law, its genetic pattern would inevitably be disordered in a few generations, and the species would disappear.

By using phase conjugation and time reversal, the **living system is** able to reverse the law of entropy in its time-**reversed channel.** That is, in that channel things go in reverse: fro***m*** disorder back to order! That's called "negative entropy" or negentropy" for short.*

It follows a priori that, to do this, the living system of necessity possesses phase conjugate energy systems (negative energy, negative lime systems) that are rooted in the atomic nuclei of its body. Further, it must deterministically manipulate these phase conju-

---

To* show how fully one can integrate physics and metaphysics, we point out that, since the vacuum imprints with every **change, nothing *is* ever lost! Death, where is thy sting?

gate systems.

This allows the precise definition of a living or nonliving system. If the system deliberately utilizes and manipulates phase conjugate energy/time in its atomic nuclei, it is a living system. If it does not deliberately use this, it is an inert or nonliving system.

For example, this sheds a great deal of light on the nature of a virus.

A virus is a single molecule. Ordinarily it clouds the distinction between life and death. Viruses can be crystalized out of solution, formed into "rock-like" crystals, and placed on the shelf, so to speak. There they can remain for many years or centuries, seemingly an inert rock. Then when placed back into their "preferred medium," so to speak, the virus crystals dissolve, separate, and "come back to life" and resume their living function. Yet the virus can be killed, in which case it is just a "hunk of rock," sitting on the shelf. In that case, it won't come back to life when placed back into its preferred medium.

In one case its "deterministic phase conjugate energy system" — its mind, to put it simply — remained in the nuclei of the atoms of the crystaline form. In the other case its "mind" was destroyed or disconnected from the crystalline form.*

A somewhat similar effect can be demonstrated in bacteria.

You can "kill" bacteria, for example, with UV radiation, then hold them in the dark for 24 hours (say, for 12 generations, since a nominal bacterial generation is about 2 hours.) They will remain absolutely "lifeless" and static, with no movement or cell division.

Then if you place them in the sunlight, you will be astounded to see them revive by the hundreds of thousands. In the "special kill" case here, the virtual state substructure of the UV photons just totally "jammed" the biopotential substructure functioning of the

---

*In the living case, the atomic nuclei remained patterned by the structured biopotential In the case of true death, the structured biopotential is lost and the nuclei revert to their "inert" form.

"minds" of the bacteria on one key channel. It didn't physically damage their bodies or "physical tuners." Then when placed in the sunlight, the broad frequency spectrum of the virtual substructure of the photons "dithered" and stimulated the bacteria's body-tuners, much like a "frequency massage" in all the bands. This set the physical tuners vibrating — and therefore responding once again to the biopotential substructure changes of the mind. Hence the bacteria "revived" and swam about — fat, dumb, and happy.

We have stated that **the mind is a physical thing, albeit a virtual or hyperspatial thing. In virtual/hyperspatial reality, it is a functioning, hyperspatial thing.** Interchanges be-

E=mc²
127034274
441523220
2142413756

©Hal Crawford 1978

Figure 88. The mind and personality are physical things, albeit hyperspatial and virtual.

tween it and an otherwise inert, functional, 4-dimensional ordinary physical system (a "body") comprise life functions, "thought," cellular control, etc. Those interchanges in a living system occur in the biopotentials residing in the atomic nuclei of the physical body, in the higher levels of biopotential in the body structures such as cells, membranes, and organs, and in the overall biopotential of the entire mass of the body.

We accent that this is not mysticism, but physics.

But now let us return to phase conjugation and the Priore machine.

The Priore Machine and Phase Conjugation

In the 1960's and 1970's, in France Antoine Priore built and tested electromagnetic healing machines of startling effectiveness.

In hundreds and hundreds of rigorous tests with laboratory animals, Priore's machine cured a wide variety of the most difficult kinds of terminal, fatal diseases known today.

Funded by the French government in the amount of several million dollars, **Priore's machines concretely demonstrated a nearly 100% cure of all kinds of terminal cancers and leukemias, in thousands of rigorous laboratory tests with animals. These results were shown to medical scientists as early as 1960.**

Many of the experiments and tests were done by prestigious members of the French Academy of Sciences. Robert Courrier, head of the Biology Section and Secretaire Perpetuel, personally introduced Priore's astounding results to the French Academy.

The operation of the Priore machine was seemingly incomprehensible. Many orthodox French scientists — some of them world renowned — were outraged at the success of such a machine, shrilling that science had nothing to do with "black boxes."

They loudly called upon the inventor to explain the mechanism utilized by his machine, but the inventor either wouldn't or couldn't, explain the curative mechanism.

Priore certainly knew how to build the machine and make it work. It is debatable to this day whether anyone — Priore included — actually understood its principle of operation.

Neither the French Academy nor Antoine Priore knew anything of phase conjugation at the time.

In fact, the entire Western World knew nothing of phase conjugation in the 1960's when Priore was getting his finest results. At that time, only the Soviets knew of time-reversed waves.

Certainly Priore's machine was impressive.

Into a tube containing a plasma of mercury and neon gas, a pulsed 9.4 gigahertz wave modulated upon a carrier frequency of 17 megahertz was introduced. These waves were produced by radio emitters and magnetrons in the presence of a 1,000 gauss magnetic field. Experimental animals were exposed to this magnetic field during irradiation, and the mixture of waves (some 17 or so) coming from the plasma tube and modulating and riding the magnetic field passed through the animals' bodies.

Amongst other things, a plasma can convert a transverse wave to a longitudinal wave. Also, phase conjugate (time-reversed) waves

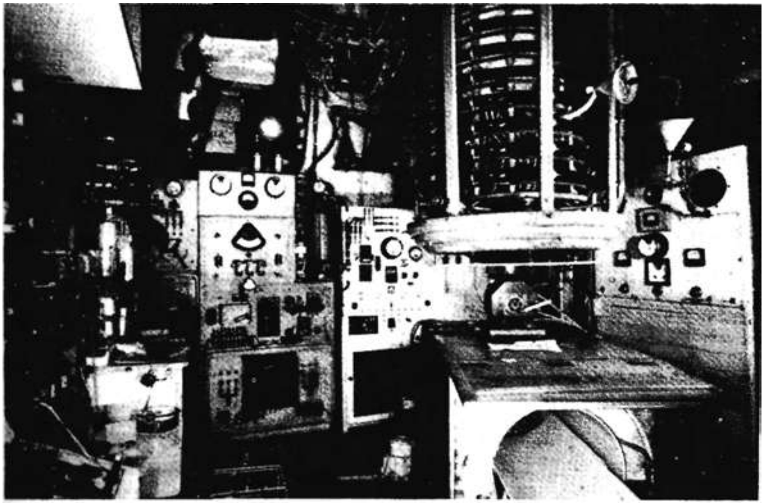

Figure 89. One of Priore's intermediate devices. This device cured terminal cancers and leukemias in thousands of laboratory animals. (Courtesy Bob Whitney)

can be produced by plasmas. Priore's apparatus produced a scalar EM wave/signal with deliberately constructed, infolded components including phase conjugate waves.

One may roughly view a cancer cell as a normal human cell gone "awry" and out of control of the body's master cellular control system. The cancerous cells, viewed as a sort of separate, parasitic group of cells, form a special kind of organism having its own master cellular control system "level," immersed in the host's biopotential.*

There is thus a specific, constant electromagnetic "delta" that differentiates the parasitic cancerous "organism" **from the normal human cellular organism.**

This "delta" can be considered a sort of constant, complex-structured charge existing in the body's atomic nuclei. It's exactly as if the body biopotential had been charged up by Kaznacheyev's "death photons" for that specific cancer condition.

If this cancerous "delta" (which may be rather like a complex intermodulation mix of waves) — or a frequency shifted "transform" of it to a different frequency band — is phase conjugated, a specific healing delta frequency pattern results.

If phase conjugate replicas of a cancer's cell's specific "delta" frequencies are fed into the body having that cancer, the deviation of the cancer cell's master cellular control system will be "time-reversed."

That will return the cancer cell to control of the animal's proper master cellular control system. The cancerous cell will be immediately destroyed, or reverted back to a normal cell of the animal.

A very similar process exists for just about every disease bacterium and infectious agent that attacks the body.

Phase Conjugates of "Death Photons" are "Healing Photons"

The Kaznacheyev experiments in the Soviet Union proved that, any cellular death and disease pattern can be induced by a specific electromagnetic pattern carried on an electromagnetic signal, if the target cells are bombarded with the pattern-carrying carrier signals for a length of time. What was not published of Kaznacheyev's work was the correspondent work showing electromagnetic reversal of cellular death and disease conditions by irradiating the diseased cell cultures by phase conjugate replicas of the pattern-carrying induction signal.

---

*One strongly points out the relevance of Dr. Robert Becker's epochal experiments proving that cells can be dedifferentiated and redifferentiated electrically. Those experiments are strong indicators that the cancer cell should be capable of being reverted to its more primitive state — that of the normal cell — by electromagnetic means.

In simplified terms, if an action in forward time induces a condition, then the time-reversal of that action will reverse the condition.

The concept is almost laughably simple. **The time-reversal of an electromagnetic disease process is a specific healing process for that disease.**

In his device, Priore internally structured the carrier photons themselves — making them **vacuum engines.** He phase conjugated his vacuum engines, and then passed these time-reversed vacuum engines down and through a strong magnetic field which thoroughly penetrated all cells of the biological organism being treated.

The scalar components (structured photons) representing the time-reversal of the disease were absorbed and reradiated in all the cells, "charging up" the nuclei of the atoms in the organism to some potential level of the exact "healing and reversing pattern."

In the process, the cancer/leukemia pattern/charge also destroyed the cancer cells, or converted them back to normal cells as appropriate.

As a highly simplified analogy, the complex signal — viewed as a scalar Fourier expansion — represented a "stress" against any abnormal cellular control system encountered and returned it to the normal cell's master cellular control system of the body.

Encountering the normal control system in normal cells, it produced zero stress.

Encountering the abnormal control system in tumerous cells, it produced great stress on it, reverting it to the normal control system of the body's normal cells.

Thus to a normal cell the Priore signal pattern acted somewhat as a comb passing through one's hair.

No interference occurred with the normal cell (one whose scalar control pattern is phase-locked to the body's master cellular control system and in phase with it).

The Priore signal stimulated and "stroked" the normal cell, but did not hurt it at all.

On the other hand, the scalar control pattern in and of an abnormal (cancerous) cell is out of phase with the body's master cellular control system, and is not phase-locked to it.

Hence the Priore signal caused direct interference with the abnormal cell's independent scalar control system. The interference pattern constituted the reconstruction of normal energy directly in the cancerous cell, and also jammed its abnormal scalar life signal.

This destroyed the cancerous cell by two mechanisms: (1) physical energy was kindled directly in the abnormal cell, causing direct mechanical damage, and (2) the cancerous cell, being an "independent living critter," had its scalar life channel (connecting its primitive "mind" to its "body") jammed and stopped.

If the cancerous cell was reverted to a normal cell by the Priore stress before being destroyed, it became just a normal cell and the Priore stress had no further effect upon it.

The Effect is Universally Applicable

Any disease with cellular, biochemical, or genetic basis can be cured in like fashion.

Priore's method, for example, was clearly shown to be able to completely reverse clogging of the arteries with fatty deposits, and to be able to lower the cholesterol level to normal, even in the presence of an abnormally high cholesterol diet.

His method also showed complete mastery and cure of sleeping sickness and trypanosome-induced illnesses.

Often Priore found that every cell of the body — even the hair — must be irradiated and treated ("charged up") with the signal, for the disease pattern was in every cell. The master cellular control system is holographic — the pattern (substructured potential with its dynamic, oscillating components) is in each component (each atomic nucleus, hence in each cell.)

Every structural level of the body larger than the cell also has its own correlated pattern, or modulation, on the overall.

A biological organism can regenerate lost limbs, for example, if it can utilize its natural recovery process in a Priore manner.

Even though Priore's work was presented to the French Academy of Sciences by Robert Courrier, the prestigious head of the Biology Section, the Academy could not understand the device and its functioning.

That was because the Academicians knew nothing of scalar electromagnetics, and phase conjugation, and the Priore machine was a scalar electromagnetic device using phase conjugation.

In the Mid-70's It Ended

In 1974, a change oflocal government lost Priore his government supporters. His support and funding were lost.

Ironically, Priore was just completing a 4-stories-tall apparatus capable of radiating and treating entire human bodies. It would have been capable of curing cancer and leukemia in humans rather than just in laboratory rats. (His previous machines were much

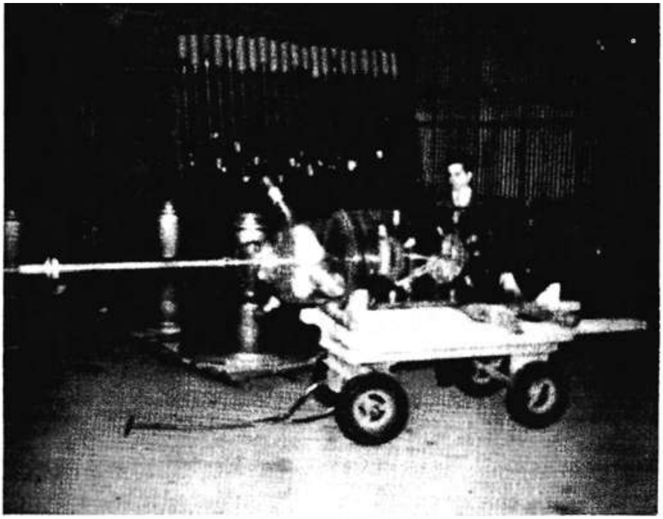

Figure 90. The large, special, phase conjugating plasma tube for Priore's giant device. The final machine would have treated humans "whole-body." (Courtesy Bob Whitney)

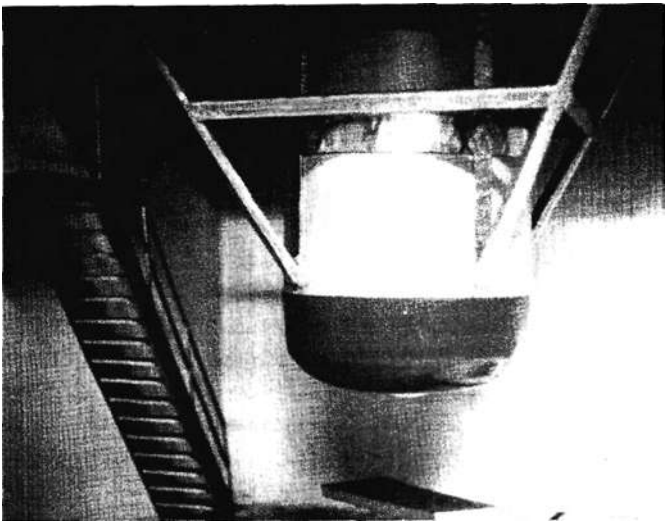

Courtesy Bob Whitney

Figure 91. Capping assembly for the top of Priore's giant plasma tube device. The last device would have treated human patients "whole-body." The cap assembly was on the third floor of the device.

smaller, and only a small animal could be irradiated whole-body.)

In that machine Priore used a "lamp," a special section, in which 17 specific frequencies were mixed and modulated upon the 9.4 gigahertz carrier.

The machine was large enough to irradiate humans over their whole body. **It should have been capable of curing cancer and leukemia in** two **five-minute irradiations, one week apart.**

While Priore was still alive, with my associates I spent nearly two years of my life in an effort to bring the Priore device to market and into the mainstream of medical research and development.

Robert Whitney, Frank Golden, and Tony Gideon played the major role, going to France and negotiating directly with the Priore group and the French government.

Verbal agreements with both the Priore group and with the French government were obtained. The machines were to be built, assembled, and tested in Bordeaux. Then they were to be dismantled and shipped to the purchasers — large medical research and development laboratories where they would be reassembled on site.

A group of wealthy backers was going to provide the necessary millions to fund the effort. Final contracts were being drawn.

Mysteriously, all funding for the project was suddenly withdrawn. Our backers were "leaned on" very hard and strongly threatened. Nothing we could do could revive the project.

Priore never again recovered his funding, and later died. The machine fell into disarray and was disassembled.

However, the proof that the requisite time-reversed signal can be produced, and will ride down a magnetic field, to penetrate every cell and every atomic nucleus in the body, is already in the present hard-core physics literature, if one knows where to look and how to properly interpret the work.

The fact that such signals can reverse nearly every major cellular illness condition of the body also has been proven by Priore and the scientists who worked with him, and it is in the French scientific literature.

The Priore-type cure for AIDS and other diseases only requires the necessary funding and personnel to be redeveloped and re-established.

Work of Australians Reed and Barsamian, and the American G. Wilbanks

At the University of Sydney, Department of Obstetrics and Gynaecology, some magnificent work of great importance to our thesis has been accomplished. Indeed, this Australian research may well be some of the most important work going on in the Western world today.

Dr. B. L. Reid, Dr. S. Barsamian, and their colleagues have produced experimental studies that positively verify pattern and form transmission at a distance, even through Faraday shielding. They have directly shown the reaction to, and results of, an unexplored information field on both living cells and inert crystalline matter.

In addition, they have replicated — and extended — the published results of Kaznacheyev on the cytopathogenic effect. In fact, they have obtained the effect at over 100 meters.

This shows, of course, that the effect is not just mesoscopic, but also macroscopic. Whether we like it or not, structural patterns of cellular disease can positively be transmitted electromagnetically, and kindled in target cell structures at a distance.

Dr. Reid has been interested in the origins of cancer since at least 1958. After much work in the standard models, by the late 1970's he was convinced that the present approach of medical science is wrong. He reasoned that we should pay stricter attention to the nature of growth before examining new growth, i.e., cancer. Growth, of course, is creation — and so the problem became, what is creation?

As a metaphor for biological growth to work with, he and his colleagues chose the growth of crystals. With keen foresight and intuition, they directed their inquiry so as to expose the complicity of some force — if any — external to solute and solvent.

Their work began to show that such a force existed, and it was a type that was unaffected by interposition of a Faraday cage. In other words, *it was not a conventional electrical force.**

Dr. Reid and his colleagues then included biological subjects in the form of living cells. They were familiar with Kaznacheyev's work in Alma Ata, Kazakstan on the subject of long range transmission of pathological (viral) effects on cells. In addition, they were familiar with the seminal studies of Gurwitsch on transmissions by onion roots cells, in the 1920's.

Most of their basic studies on the nature of the new force, however, were made with common salt, drying from aqueous solution, Some of their major results were as follows:

1. The familiar cubic symmetry is altered toward a higher energy form of dendrite crystals when biopolymers such as protein are included in the solvent.
2. A mass of lead 12-20 kg in the crystal space (up to 1 meter distant) causes an interference pattern in the salt crystal display.
3. A prospective dendrite pattern is not realized when a prospective cubic-pattern of the salt is dried in the local space (up to

---

"For one related possible scalar EM (electrogravitation) interaction, see Ya. B. Zel'dovich, "Electromagnetic current and charge due to interaction between a gravitational and a free electromagnetic field," **J. Exper. and Elec. Phy.** (Soviet translation), 16(7), Oct. 5, 1972, p. 302-303, 1 meter). The latter by its presence in space is able to prevent the dendrite expression and enhance the cubic symmetry expression. **A photograph of crystals of cubic symmetry when moved by a vibrating holder (2 cps) has the same effect on the real crystals drying from the solvent.**

4. Dendrite forms of salt crystal result when certain chemical reactions involving proton or ion movement occur up to 30 meters away, separated by brick walls, glass, etc. These reactions were detailed by Dr. Reid and his colleagues. If the mother solution is examined by ac capacitance on a sensitive bridge, variations in capacitance of over 300% accompany the receipt of the signal.

5. If such a transmission from 30 meters distance is regarded as a carrier wave which is modulated by a foreign — and thus detectable — chemical, **the chemical can be detected in the distant atmosphere some 3 months later in the form of a discrete or adiabatic 'cloud' which contaminates sodium chloride drying at the same time.** In this way, copper sulfate dissolved at a site A in June 1985 could be repeatably detected at site B, 30 meters distant and separated by walls, screens etc., by a drying solution of sodium chloride solution in September 1985. Serial dryings of the sodium chloride solution on glass slides showed that each day, the cloud hovered over a space of 1-2 cm at the end of a venturi tube for several hours before 3 p.m., and then moved off at 1 meter per hour in a NW direction. The copper content of the cloud was discernible by crystal structure and chemical means from the adjacent sodium chloride.

6. All of the above crystal forms required ambient oxygen for their manifestation. Atmospheres of nitrogen and argon prevented their display. At the time, the researchers suspected that the effects may have been carried by orbital electrons of oxygen showing unpaired spin which had interchanged with electrons of the copper solution. They then

developed an information-carrying device which could 'infect' the electron cloud of a target molecule sufficient for the chemical display of properties of the infecting electron's parent atom. This concept seemed to explain the experiment of the following paragraph.

7. Two beakers of sodium chloride solution were connected in series with a small voltage source (say a 1.5V dry cell) to provide a transfer of electrons through each beaker. The second beaker in series was sampled for subsequent drying and display of the crystal form. A crystal form of cubic symmetry resulted. When the solvent of the first beaker was altered by the addition of a protein solution (such as albumen or globulin), the crystal pattern of the second beaker was altered from cubic to dendrite pattern as though protein were present in the second solution.

8. During attempts to stabilize salt crystals on glass by means of thin plastic films drying from appropriate solvents (in order to stain the salt with aqueous solutions of dyestuff), it was discovered by chance that the crystalizing process is accompanied by structural patterns on the plastic film so exposed. These patterns were of two major types: (a) vortices, or spirals in the film plane, of 0.1 to 1.0 mm diameter, and (b) smaller curled structures into which the spiral could be seen to merge or grade. From a concurrent mathematical study of vorticeal interaction as designed by Barsamian, the similarity of the curl structures in plastic films (the researchers usually used 0.25% polystyrene in benzol) to computer simulations of vorticeal vectors was shown to be remarkable.

9. The researchers then dispensed with the underlying crystals and exposed the drying polymer film to a variety of spaces where the field, thought to be responsible for many of these tele-effects, was operative. With this technical advance, the team has made rapid progress over the last year or so. They have come to have a more comprehensive grasp of the ether

field than was possible by less direct methods involving formation of crystals.

10. On the biological side, they repeated the Russian studies done at Alma Ata with transmission of the cytopathogenic effect of viruses on animal cells. They used the cytopathogenic effect of colchicine and its relative vinblastine on cells with the following differences from the published Russian work: (a) The effect can be transmitted at least 100 meters between campus buildings, (b) it is not necessary to use a magnetic field for the propagation, (c) it is not necessary to use quartz vessels; glass is quite effective. These experiments came to the attention of a colleague, Dr. G. Wilbanks, Head, Department of Obstetrics and Gynaecology, Medical School, University of Chicago in 1983. Dr. Wilbanks was able to repeat them with similar results. However, in a more faithful copy of the Russian experiments, he used quartz vessels with cells placed in a magnetic field, with the vessels separated by no more than 1 mm. Both the Australian researchers and Dr. Wilbanks shared another thing in common: their uniform inability to interest the editors of several scientific journals such as **Science, Nature, Naturewissenchaften,** and others in publishing papers describing these fundamental experiments and results.

11. Experience with interference of a lead mass in the crystal work prompted the use of several metals in the ambient field of growing cells. Only lead (and to a lesser extent tin) showed the following properties: (a) Cell growth with 0.5-5 Kg lead in the ambient field is depressed sometimes to one third normal, (b) The cells die (dye exclusion test) but then can be revived by placing a small bar magnet in the local space with its poles in an E-W orientation. At least in Sydney, a N-S orientation of the bar magnet does not revive the stricken cells.

12. The effects of energy flowing through living matter have been documented by Kirlian's method. It was therefore of interest

to the Australian researchers to expose living systems to drying films of polystyrene. They found that the staining properties of the ambient-derived image on plastic films parallel those of the real organism.

13. By appropriate use of an imposed potential on living cells, it proved possible to match the field strength and show the actual imposition of the ambient (scalar) field on the biopolymer content of the material during the permeation process. The traces obtained are not unlike noise and would be so regarded by the casual investigator. The design of this apparatus and associated experiments was the work of Dr. Barsamian.

Concepts of Dr. Reid and Associates

Dr. Reid has kindly shared with me some of the concepts he and his colleagues have formulated. These concepts, it is stressed, are consonant with direct physical results they have obtained in the laboratory.

Note that any errors in this explanation are strictly the fault of this author, and not of Dr. Reid and his colleagues.

1. Matter is bathed in an energy sea familiar to quantum theorists. The Australian team prefers to call this sea of energy **ether,** as a relatively bland term, perhaps in a way to rescue it from the oblivion where Einstein banished it. The key to the silence of this sea is its "fractional-charge" charged particle status and low energy values. The former conceals it from physics, while the latter conceals it from chemistry and thermodynamics.

2. The differing energy levels of this sea, combined with its enduring motion, are theoretically and practically best expressed as a **vortex.** Dr. Reid and associates have detailed this theory to explain the creativity of biology. The incipient creativity of the vortex is given material expression by its association (as part of the EM field) with an electron which it

now endows with equivalent novel properties. There is no intuitive reason why the energetic rearrangements of the electron cloud cannot proceed to the nucleus, there to set up a quark or proton rearrangement — or, as the quantum theorist would say, a virtual particle exchange — leading to transmutation. Kervran's work on biological transmutations is relevant; the literature of such effects goes back to 1798 at least! The team suspects that some such mechanism may underlie their experimental result of the transition of sodium to copper.

3. **A further prime property of vorticeal interaction is its proclivity for amplification,** wherein the atomic-size vortex can be built up to the meteorological hurricane by such adherence or affinity. The team has noted the affinity of ambient field vortices to those produced by an electric current in a metallic conductor, with particular reference to the ill-defined boundary between the two. This parasitism was the reason behind the team's use of "carrier waves" from chemical reactions as transporters of fields whose identity of origin were sufficiently specific to trace the field subsequently. It is at least conceivable that the whole process could be amplified and used to parasitize (modulate) a radar beam as carrier. The affinity property of fields may be exhibited by the experience of mariners with what they loosely term "magnetic phenomena" thought to be derived from luminescent plankton. In the South China and Arabian Gulf seas, giant "wheels" of light, 5-15 km in diameter and containing spiraling spokes, revolve about the ship at night, reversing their direction of rotation as they pass the midpoint of the vessel. The key observation in the context of the field affinity concept is that the luminescence is manifoldly enhanced when the ship's radar is switched on. There could thus exist a chain, scalar-EM/plankton-optical, the first leg (scalar EM) of which can be amplified by radar EM.

4. The property of affinity means that **a scalar field's** transform **or copy, of the energy disposition imposed on it by the experience of matter last permeated, may persist as**

> **a** memory **in space,** as many of the Australian observations attest. The team found that recall was not all that difficult technically. **The energy transform for copper** sulfate **existing as a non-dissipated cloud with marked adiabatic properties could be recalled five months after its formation. There is no reason why informationally more elaborate spatial configurations of energy, including human thoughts, will not behave in precisely the same way.** The cell nucleus could conceivably be the candidate for the first site for storage of such energy. The nucleus has a highly ordered structure and an associated high permittivity, and this would make it a real target. In the human subject, there is experimental evidence that the mongoloid polymer constitution will have a higher activity for storage of the ambient field than will the Caucasian.

Other Important Ideas

Dr. Reid also shared other important insights:

Lavoisier in 1748 advanced the concept that matter itself (chemicals) is inert. The activity of all chemicals, according to Lavoisier, resided in their associated ether — and God sent the ether!

Taken to its limit, this suggests that the source of the energy which we call *thought* is a property of the scalar field. Matter (say, the nervous system) interacting with these fields, would be incapable of discerning the source of the field as presented, whether of self or external origin.

In his book **Hidden Variables in Quantum Theory,** Dr. David Bohm felt that in 'elevating' (his word) matter (as particles) from the sea, it might be possible to consider the idea as containing even more elaborate energy transforms such as biological shapes, these transforms themselves undergoing metamorphosis. [Comment by this author: Rife's microscope, using evanescent waves, actually allowed one to directly see and photograph such forms, proving the thesis.]

The ability of the Australian researchers to rescue these transforms on films together with evidence that they could be seeing the results of the exclusively space interaction of virtual forms (say microbe and antibiotic), is also a direct technical step toward validating Bohm's important idea.

Wider Implications of Structured Field Information

As T.D. Lee points out in his important book, **Particle Physics and Introduction to Field Theory,** particle physicists have not attempted to use the structure of the vacuum nor the structure of potentials.

Dr. Barsamian has shared other wider implications of the structured ether.

Over the ages a phenomenon which, as discussed, can have no representation in real (that is, electromagnetic or thermodynamic) terms can hardly have been thought of as possessing specific structure or function. Instead, it signaled its undoubted presence to generations of philosophers and physicists before Einstein, solely through statistical effects.

The writings of Helmholtz, James Clerk Maxwell, and Neils Bohr are examples. The more recent Dirac and Fermi statistics are **a** continuation of the process.

If an alternative definition of science is the continuing refinement of that which we partially knew or took for granted, then the refinement of these statistical approaches ~ with their obligatory uncertainty as defined by Heisenberg — is a worthy goal.

In this sense, a more articulate description of the anatomy of the field is a logical progression from the beginnings of the ether concept in Classical Greece through to the statistics of the nineteenth to the early twentieth century.

A hundred years ago, this orderly progression even reached the concept of a vorticeal atom at the hands of J. J. Thompson. However, **it** was interrupted very effectively by Einstein's publication of the special theory of relativity.

Einstein was fortunate in that he had chosen for his mathematics a pivotal feature of the ether field, its elasticity. The mathematics of elasticity had been evolving — through the developments of Hooke, Young, invariance, and the calculus — to handle its historically increasing complexity (by tensors). This mathematics of elasticity was then available to Einstein. The wisdom of this choice was vindicated by the remarkable special relativity predictions which were born out in subsequent observation and experiment.

There was seemingly no need to suspect, let alone probe, any special properties of the medium underlying the mathematics.*

The success of these predictions effectively blocked any chance of a return to the ether field concept. Indeed, it was seen as an arrogance to despoil the "elegance" of the mathematics. Nor was the success of quantum mechanics — the other leg of the Western World statistical approach to the nature of reality — able to get the physicists to heed Bohm's 1958 ideas on the hidden variables in its fabric or Dirac's 1952 ideas on a lesser uncertainty of the particle-field relationship than Heisenberg had formulated.

Recently Dr. Barsamian has originated the concept of an **alienation of the electron from its field.** The full expression of this could see the field so alienated that it is well-nigh autonomous, and some kind of fundamental entity in its own right.

In this sense, the freed autonomous field — at least conceptually — could proceed with its creativity by hybridization with other fields and so return the product of these asides to the particle. In atoms of the next fragment of matter with which the particle then associates, this product could create novel properties derived from the interchange.

The multiple properties of such a field allow it to be resolved in EM (electrostatics and electrodynamics), in atomic structure (electrons and nucleons), and in pressure (gravity) terms. This universality makes a new concept possible, long beloved of the GUT (Grand Unified Theory) theorists: that there is only one energy.

It is also quite possible that this ubiquity need not be confined to physics. It may penetrate the biological imponderables of growth, new growth and consciousness.

In view of the work of Dr. Reid and his associates, and other work — such as that of Hines and Chimonas showing atmospheric gravity waves launched by auroral currents — the prospect of amplification and transmission of transforms and patterns over long distances,

---

*As, for example, whether or not the concept of an abstract vector space in vector mathematics must be modified to correspond to what we know about the vacuum medium. If the vectors are to correspond to the movements of physical things in the *vacuum medium,* then the vector medium must correspond to the vacuum.

through affinity with long distance carriers, is indicated.

Dr. Reid stated that it would be surprising if other laboratories outside Russia had not progressed as far, if not further, than he and his associates had. Obviously the lack of similar publication indicates the nearly impenetrable editorial curtain confronting Western researchers in this and related areas.

Dr. Barsamian's Important Theoretical Insights

Dr. Barsamian has formulated a theory in relation to the development of cancer that is of great importance.

In his theory, he is developing the approach of classical mechanics to electromagnetic interactions between charged particles and between dipoles. The field associated with the primary moving particle is not neglected in his theory, and this part of it differs from Dirac's "new electromagnetic theory."

The field associated with the primary moving particle, in Dr. Barsamian's approach, is viewed as an important part of the particle's interactions. This field is stressed in the sense that diverse phenomena — from Maxwell's EM field radiating from the remotest regions of the universe to the controlled charge flow in the matter medium — are viewed as manifestations of EM forces.

This of course is unlike the traditional relativity view of EM force action on moving charges.

In Dr. Barsamian's view, the potential vectors (derived by a reatment from fluid mechanics theory) form an EM field and are positioned on a vortex curve. All parameters of such a field can be treated mathematically, and Dr. Barsamian has one or more papers presently in publication.

The duality of such potential vectors (gradients) is to conduct the motion of particles (which are always either charged or dipolar) appearing in nonmatter, which is taken as an incompressible continuum — the "void." Because the final results of these forces are always movement of matter, Dr. Barsamian prefers not to classify them by their origin.

Such a field may persist with great permanence. There are some limitations in summation (triplet conditions) and selectivity (periodicity) of the vortices themselves. Part of these limitations takes the form of particle characteristics. Presently he is rather sure that the formation of patterns of organic, inorganic and liquid crystals (typical to living biomatter) are partially responsive to the effect of the described field.

Dr. Barsamian is attempting to synthesize his own experimental findings on the dielectric behavior of biosystems as they enter the cancer process with the data that Dr. Reid is producing with respect to action at a distance.

With respect to living systems, he has slowly been approaching the conclusion that existing physical laws are just not applicable concerning (1) entropy, or (2) energy breakdown, or (3) random collisions of particles in quantum terms, or (4) electron-ion exchanges in bioelectrochemistry terms.

Under such circumstances, it is difficult to find a referee who sincerely understands the related experiments and their underlying theory. This presents a nearly impenetrable barrier to publication in leading scientific journals.

This is tragic, because these considerations and experimental results at least provide an entry into more fundamental experiments with present lethal diseases such as cancer and AIDS which annually take large numbers of human lives.

The present state of the art is telling everyone loud and clear that something fundamental is wrong with present theory, and that we should come to terms with more novel views.

Comment by the present author: In his prophetic insights, Dr. Barsamian has placed his finger squarely on the fundamental struggle of living systems against the "positive energy/time" laws of present physics. In its struggle, the living system must use negative energy/time, to correct from disorder to order and provide the extra free energy to do so. Otherwise, the steady progress toward increasing disorder in all its actions, would very shortly reduce the living system to such a disordered state that it could no longer reproduce the clear genetic pattern of its own kind. We have already pointed out the hitherto unexpected role that phase conjugation plays in allowing the living system to violate and break the iron dictates **of** conservation, nontransmutation, entropy, and raw statistics.

To live and survive, a living system must have access to other functional levels and other mechanisms beyond those contained in our present physics, electromagnetics, and biochemistry.

The incredibly productive and important work of Drs. Reid, Barsamian, and associates is of the utmost importance. Their work clearly moves everything in biochemistry and the knowledge of cellular functioning a major step forward.

It is a betrayal of the scientific method that such vital work goes so largely unreported in the leading journals. Indeed, the work of these Australian scientists should be galvanized and well-staffed and funded. They can literally usher in a new science, given the chance to further exploit and develop their results.

Royal Raymond Rife

In the 1930's and 1940's, Royal Raymond Rife revolutionized everything that has been done before or since in high resolution optical microscopy.

He also revolutionized everything before or since in cellular biology. He carried cellular structure far beyond anything ever dreamed of at the time or presently. He revealed the direct connection between organized living energy forms and organized biological systems. He revealed that life itself is organized and dynamic, to a far finer level than anything in the textbooks today. **He revealed that our present theory of disease is fundamentally very, very wrong.**

He produced direct, economical, electromagnetic cures of cancer, leukemia, and other such debilitating diseases. His work presages a future mankind could have had, where most debilitating diseases were quickly and economically corrected, and where no poisonous drugs, violent nuclear irradiation, and harsh chemotherapeutic "burning" of the patient would be necessary.

For such epochal work, he was ostracized, essentially imprisoned in a medical treatment facility, broken, condemned, and rejected by his peers. His findings, though printed in reputable publications and journals, were discredited and ridiculed.

He literally was reduced to a non-person by the power of the medical cartel.

Finally escaping from his enforced confinement, he lived out his remaining years and died quietly and unknown.

Rife's Universal Microscope

With his universal microscope, Rife achieved optical resolution of up to 31,000 diameters and magnification up to 60,000 diameters. His microscope could examine living viruses, living bacteria, and other as-yet-undiscovered living organisms and living energy forms that no other microscope before or since could see.

Even today, only the electron microscope furnishes such resolution, and it zaps-to-death the objects that are being examined. Further, it will not at all detect or see the "living energy forms" revealed by Rife's microscope.

To appreciate Rife's accomplishment, let's briefly summarize some of the performance factors of an optical microscope.*

Several factors are important in the functioning of a microscope: Resolution, magnification, and contrast lead the list.

By **resolution** we refer to the ability of the instrument to distinguish a small object. In other words, something like looking at a medical doctor's eye chart, and specifying the smallest print one can clearly see. Resolution is often referred to as resolving power.

By "magnification" or **magnifying power** we mean the ability of the instrument to "blow up" or enlarge the image produced. Roughly, it's the ratio of the apparent size of the object seen through the microscope to the apparent size of the same object seen without the instrument.

Magnifying power is exactly like enlarging a photograph. It won't improve the resolution or make the photograph any clearer, but may make whatever was captured (resolved) easier for the human operator to see. If you just keep enlarging the photograph, it will get more and more grainy, until no further details can be seen with higher magnification.

Contrast refers to the distinctness by which the various parts of the object can be distinguished from one another. To enhance contrast of objects under the microscope, staining of the objects is often used. Very often the act of staining will itself kill the living organisms — such as bacteria — that are being examined.

*The electron microscope reveals only the "dead carnage" of the battlefield after everything is destroyed. With it, one is observing physical residuals, not the living players. For some encouraging modern work, however, see Michael Hercher, "Virometer — an instrument for the measurement of the size of viruses using an optical microscope," **SPIE Seminar Proc.**, 1977, p. 17-22.

Figure 92. Royal Raymond Rife in his laboratory. Courtesy Christopher Bird

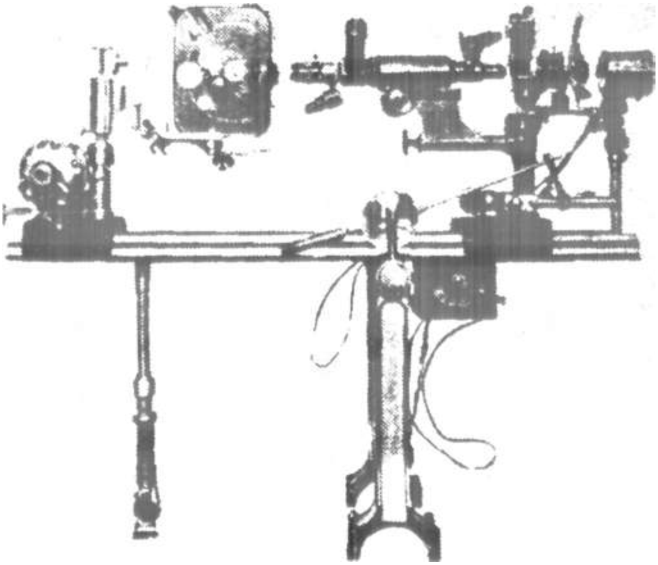

Figure 93. Rife's first virus **microscope.** Courtesy Christopher Bird

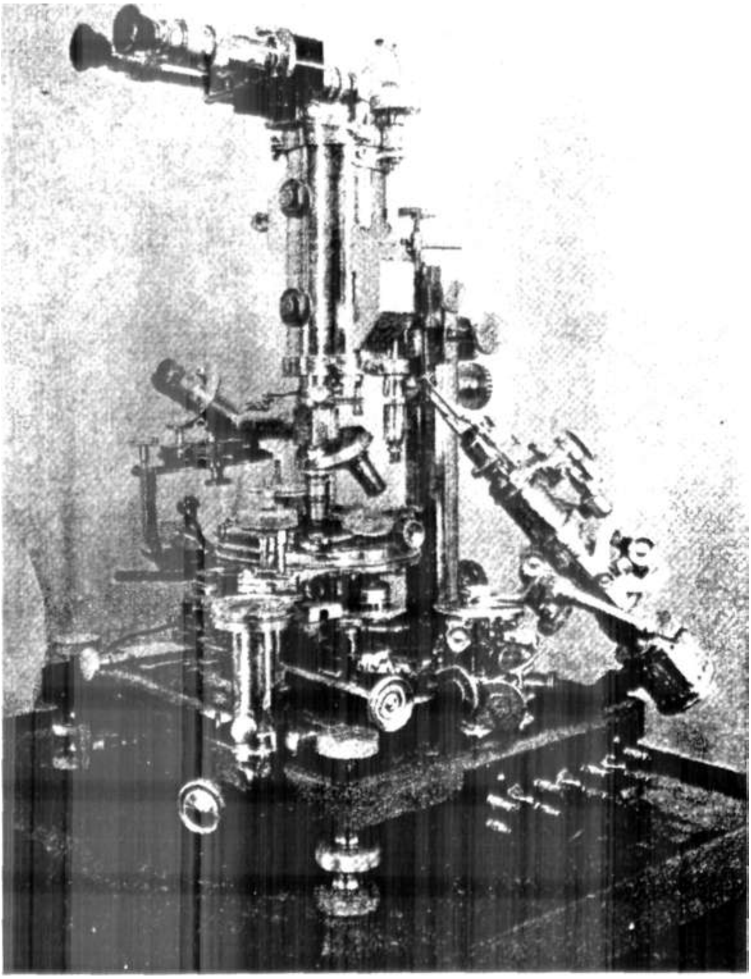

Figure 94. Side view of Royal R. Rife's prismatic universal microscope. Largest and most powerful of five Rife microscopes With it, living filterable viruses could be observed. Courtesy Christopher Bird.

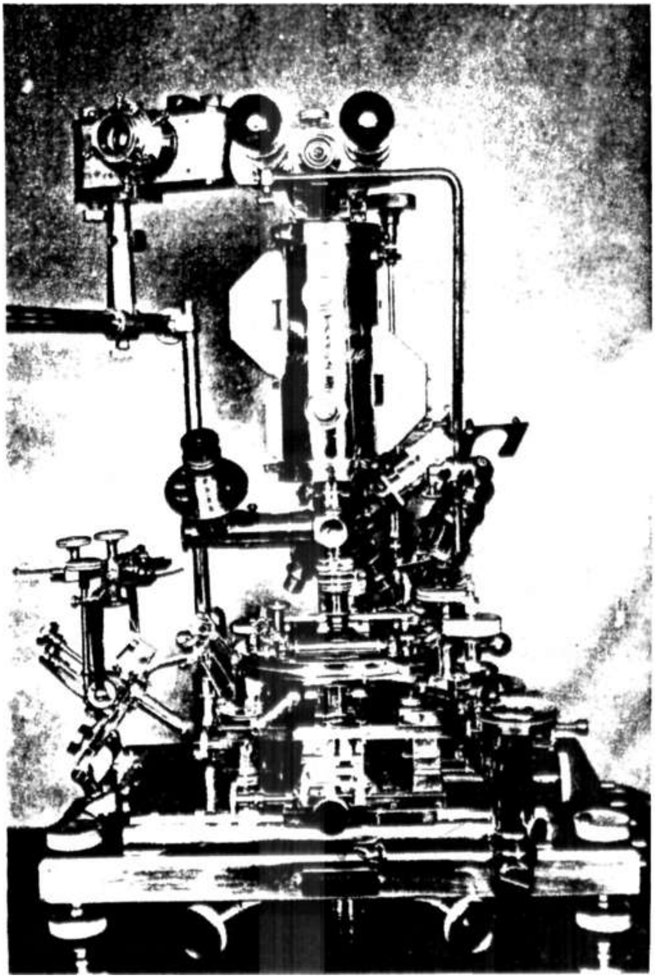

Figure 95. Front view of Royal R. Rife's prismatic universal microscope. Built in 1933 in Rife's San Diego, California laboratory. Courtesy Chris Bird.

We may state the resolution of an optical microscope in terms of the diameter of the smallest object resolved: for example, so many nanometers. Or we may just simply and loosely speak of so many thousands of diameters.

The resolution of an optical microscope depends upon the illumination conditions, its optical system, and the fact that the object viewed diffracts and spreads the light. A wide variety of mechanisms for illuminating, resolving, contrasting, etc. are used in optical microscopes.

According to standard Rayleigh theory, normal optical microscopes are limited in resolution to about half a wavelength. That corresponds to about 250 nanometers for visible light illumination. One can do a little better than that by using ultraviolet light and quartz optics. The UV has a higher frequency than visible light and hence a shorter wavelength. However, beyond 240 nanometers the resolution of an optical microscope quickly disappears.

A few researchers — notably in the 30's and 40's — remarked about other factors, such as "quality of the lenses," affecting resolution. Some of them reported results to about a tenth of a wavelength. Loosely, that translated to somewhere in the order of 5,000 to 6,000 diameters. The dependable upper limit seemed to be — and still is — about 3,000 diameters, however.

Normal optical microscopes simply are almost useless for trying to look at viruses. They will not resolve any but the very largest viruses, and will not resolve the internal features of even those few giants.

Much of today's knowledge of virus structures and shapes comes from the use of electron microscopes. They bombard the object viewed with a fierce rain of energetic electrons. These instruments see nothing of the *functioning* of the virus, for they kill it instantly in trying to view it.

With Rife's universal microscope, the dynamic living functions of the virus could be observed without killing it.

With Rife's microscope, a whole range of complex organisms and structures below the size of bacteria was revealed. Many of these organisms are still not known to present science, even though some of them were written about at the time. Filterable forms of bacteria

— forms which readily passed through filters supposed to easily block their passage — were discovered and reported by Rife and his medical colleagues.

In addition, Rife's microscope — which contained block-quartz prisms and lenses and interference stages — revealed halos around living organisms that other microscopes could not see, even though the object was within their resolving power. Further, it revealed the existence of entire organisms and forms which other microscopes could not see — even though, again, the size of the organisms was within their resolving power.

In other words, Rife's microscope not only revealed smaller physical forms than any other microscope could see, but it also revealed sizeable "living energy" forms which no other microscope could see.

Rife Proved That Everything Is Alive

The degree of smallness to which Rife's microscope would resolve, and the extraordinary energy forms it could detect, showed that direct detection of the virtual state organization of the living organism was accomplished by the instrument.

The science of the day was only just groping its way toward any sort of physics that could explain such an astounding instrumental result. Today, however, in the hard-core literature there is demonstrated proof that the optical limit of resolution can be drastically overcome, using evanescent waves.*

Shortly we will briefly give an "ad hoc" explanation of such evanescent waves.

Electromagnetism has been shown by Kaluza-Klein theory to actually exist in the fifth dimension. In other words, EM itself is hyperdimensional. It flows in the fifth dimension, which is "wrapped around" each point in our ordinary space. It is — to the first approximation — the external environment of every normal point in

---

*E.g., see T. Sato et al, "Application of evanescent waves to microscopic observation," **Bull. Tokyo Inst. Technol.** (Japan), No. 125, 1974. p. 35-41. See also G.A. Massey, "Microscopy and pattern generation with scanned evanescent waves," **Appl. Opt.** (Poland), 13(3), 1983, p. 247-255.

our 3-space.

Let's say that again.

The electromagnetics "medium" itself is totally external to each point in our space. Each of our points is surrounded in higher space by the electromagnetic medium.

We live inside a totally electromagnetic medium. It's not just an "electromagnetic environment" in our own space; instead, it's an electromagnetic environment in hyperspace.

Everything already is just the internal structure of the electromagnetic medium!

That's what the "vacuum" is.
That's what "spacetime" is.
That's what the "virtual state" is.
That's what physical matter is.

And it's all alive. Totally and completely alive. Everything is alive. There is nothing but life. The electromagnetic medium is alive.

Except within a locally-organized biopotential area, it's just "equally alive in all directions." So it appears inert, where by "inert" *we mean not singly preferential.*

Every biopotential change, at any level in a biopotential, extends "decaying-expotential-wise" to infinity, by standard theory. Part of that potential change exists at every point in the entire spatial universe. And there at that point, so do the biopotential changes for every other biosystem in the universe.

Everything's alive. There is nothing but life. Anywhere. Anytime.

An interchange between this "living, surrounding electromagnetic medium" and each particle of mass in our 3-space continually occurs. This exchange involves the so-called virtual particle flux.

If we model higher dimensions, such as are necessary to include the "particle zoo" discovered by modern particle physics, each of these "higher dimensions" corresponds to a "successively deeper-"

internested level of virtual particle flux—of the vacuum, or of this electromagnetic medium that surrounds us.

The living system orders, structures, and dynamically functions and interchanges electromagnetically throughout many higher "dimensions" — throughout many such internested levels of virtual-state/vacuum/spacetime.

Except for the first layer of virtual state (the 5th dimension in Kaluza-Klein theory), gross (classical) electromagnetic theory totally ignores the infolded, ordered deeper structure of electromagnetism (of the EM medium).

Using such gross electromagnetics, orthodox science thus can only detect and grasp the gross results of the living organism's functioning. Further, it can only detect and grasp those gross results that actually move 3-space matter — observable particles.

So it only detects the grossest interchanges between "mind" or "life" and matter. Specifically, it only detects the final results of that interchange—the gross movement of material particles themselves.

The first hypernumber — "i," the square root of minus 1 — is used to model another dimension at right angles to our normal three. An electromagnetic wave thus is modeled to have two parts: the first is the part that affects or moves a material particle, and that's called the "real" part or "observable" part. The second part of the EM wave is the component that lies in this "imaginary" (unfortunate standard term!) dimension. That imaginary part does not of itself move charged particles, so it is considered to be "something other than real."

Note that what's really done in this sort of orthodox modeling is to define observation as the movement of observable charged particles.

The human "conscious mind" is a functioning part of the overall human mind that has been specifically fitted to function almost totally with our gross bodily detection of the photon interaction.

We thus consciously detect and "are aware of only the first level of reality: the interface between the first layer of virtual-state/ vacuum/spacetime/ and the 3-space of observable particles of mass.

Since the physicist must be "conscious" of his observation or instrumental detection, the detection process ultimately has to have a final stage that produces a photoelectric interaction with his body — for that is all the conscious mind detects and processes.

Everything chemical, electrical, etc. — according to quantum mechanics — will involve at its root level only the photoelectric effect.

Rife's Microscope Used Evanescent Waves

However, it is certainly possible to build an instrument which is multi-staged, and one in which higher stages interact primarily with deeper internested levels of vacuum virtual state. That is, an instrument that interacts with higher dimensional phenomena. That is, one that interacts with deeper internested levels of electromagnetics. If these stages interact causally in a vertical manner, then the final interaction with the human body and nervous system can still be "normal electromagnetics" and yet indicate a higher dimensional phenomenon, event, or function.

That is precisely what Rife did: His universal microscope penetrated to a much finer level of reality because its multiple stages used **evanescent waves.**

Here's what we mean by that.

An electromagnetic wave consists of a "real" part and an "imaginary" part, as we have discussed. It's possible, however, to have the real part become zero, and still have the "imaginary" part remain and dynamically vary.

This sort of wave — containing only the "imaginary" part — is said to be one type of **complex wave,** or an **evanescent wave.**

In advanced EM waveguide and optical theory, such evanescent waves can function to guide or determine the real parts of the EM waves. That "real part" is the part that is then going to interact with electrons and move them, and give us a "detection."

In other words, the "real" part of the EM wave will be moved and guided by the "higher dimensional" or evanescent part, or even by separated evenescent parts — pure evanescent waves.

So one can build a device or instrument that utilizes such an effect to reveal what's going on in a whole dimension beyond what we normally see. Waveguides in certain RF radars already do that.

Such instruments are not limited to the resolution determined by the "real" part of the EM wave (such as the "real part" of light).

Such a microscope would not necessarily be limited in resolution by the wavelength of the light it used to illustrate the observed object.

Indeed, if one uses a higher dimensional Kaluza-Klein model (which particle physics of necessity has to consider), one could build a "repetitive stages" instrument which could view many more higher dimensions. That's exactly the same thing as viewing many more deeper internested levels of virtual state.

That is precisely what Royal R. Rife did: **He developed a microscope using multiple stages, special interferences, etc. He built a multistaged evanescent wave instrument that could see into higher dimensions and deeper levels of virtual state.**

Rife's microscopes were thus startlingly different from normal microscopes. That is why he could see phenomena and organisms, ostensibly within the size capability of normal microscopic resolution, which ordinary microscopes could not see.*

Evanescent waves penetrate the virtual state (hyperdimensions) and interact with what to us is "dimensionless" (what does not spatially intersect our 3-space). Interfering multiple evanescent waves reflected from/having interacted with these (to us) nonphysical forms can again reproduce ordinary electromagnetic "light." Thus the evanescent wave microscope, if properly built, allows one to directly "see" (and photograph) what is to us living forms of energy, nonphysical, and without 3-space matter bodies.**

The same scheme can be used to develop instruments capable of directly revealing the living nonphysical world around us — such as the human biopotentials, their internal structures (including

---

*However, Rife's microscope was extraordinarily difficult to focus. Rife often spent 24 hours straight at his universal microscope, just in focusing it.

**If one is going to accept the many-dimensional theories necessary to explain particle physics, one must accept the possibility of higher dimensional living things, which we would see as living energy forms or as thought forms. These would appear nonphysical to us - as is time itself— but would be real and interactive nonetheless.

"thought forms"), etc. What the Soviets call **bioplasma** can be directly observed and photographed in this fashion. We have only to develop the necessary instruments; the science required is difficult, but it is already in the literature.

That is what Rife's microscope did. It was a forerunner of the instruments we need to develop in electromagnetic healing, and Rife was a truly great and unappreciated pioneer.

In one swoop medical science could have jumped a century ahead. The ruthless suppression of Rife and his fantastic scientific breakthrough was one of the most dastardly deeds ever perpetuated by the orthodox scientific establishment.

Rife Revealed a Far More Fundamental, Living Biology

With Rife's powerful universal microscope, it was also possible to view the interiors of the so-called "pinpoint" cells situated between normal tissue cells and just barely visible to ordinary microscopes.

Here is the astonishing living world inside those "pinpoint" cells, as revealed under Rife's powerful instrument: When one of the "pinpoint" cells was magnified, still smaller cells were revealed within its structure. When one of these still-smaller cells, in turn, was magnified, it too was seen to be composed of even smaller cells.

With Rife's microscope, this process could be repeated 16 times. An astonishing internested series of organized levels of a living cell was revealed, far more fundamental than anything that exists in present biological theory.

The present author points out most strongly that these levels correspond electrically to, and are in virtual particle flux pattern exchange with, the electromagnetic potentials of their environment. This includes their own biopotentials that are centered in the atomic nuclei of the atoms comprising the physical material of the cells, and that charge up with specific internal patterns.

The biopotential itself is organized into a corresponding virtual-state series of internested levels and functions. **The structured biopotential of the cell is a living, organized, functioning thing, and its internal functioning literally constitutes the "spirit" or "true nonmaterial deep mind" of the organism.**

All the internested levels are in constant electromagnetic exchange "up and down" with each other, particularly with respect to organized virtual particle flux patterns.

In addition, all the cells are in constant electromagnetic exchange "across" with each other, at all levels. This provides a dynamic, structured, living biopotential for the entire bio-organism (the entire body). Within this potential, dynamic interchange on all levels is continually occurring. This is the basis for the master cellular communication system that Dr. Fritz Albert Popp discovered.

Rife's powerful microscope had revealed nonmaterial functioning life forms (structured, dynamic, living biopotentials) connected to material bodies.

He could follow "filterable" forms of bacteria — actual living biopotential forms of the organisms that could not be separated out by filters, but which would easily pass through any filter.* He could observe interactions of these forms, changes of forms, translations and transmissions of forms, etc. — none of which is detectable by present biological theory or medical science.

Rife had advanced biology and biophysics a century in one jump. As always, orthodox scientists — most of whom in their scientific paradigm are self-admitted materialists — were quite unready to tolerate such heresy.

Obviously the materialistic dogma of the science of his day — and of the science of today — reacted most hostilely to such hogwash. Contrary to the prevailing mystique, most scientists are dogmatically attached to materialism and to the dogma of their present paradigm. Faced with a conflict provided by experiment, most will uphold the dogma and reject the experiment — the exact opposite of the scientific method they espouse.

Rife's revolutionary work was no exception. He incurred the unending, total opposition of powerful individuals controlling the direction of biology and medical science for their own personal gains.

Rife was hounded into court on trumped up charges. Though he was acquitted, he emerged a shaken, broken man and an alcoholic.

*More precisely, they would easily flow *around* the 3-dimensional filter, since they were hyperdimensional.

His work was suppressed. His equipment was left to gather dust. He was also forcibly committed to a medical treatment facility.

Finally escaping from his "prison," Rife lived out his remaining years quietly. He died without ever being vindicated for his marvelous, world-shaking discoveries.

But at least he left us a legacy. Persons are still alive who knew Rife and his work, to one degree or another. Some of his microscopes are still in existence, though non-operable due to missing or stolen parts. With proper funding, an enlightened team of scientists and researchers can be assembled to quickly repair Rife's remaining instruments and duplicate his incredible work.

With concentrated effort, Rife's work will yet play a primary role in the development of a direct electromagnetic cure for AIDS. As a byproduct, it will play a role in rapid development of cures for cancer and other debilitating diseases.

Royal R. Rife's contributions may yet help save half of humanity, and prevent a Soviet takeover of the world.

CHAPTER 6

DEVELOPING THE ELECTROMAGNETIC CURE FOR AIDS

By now, how one goes about developing a electromagnetic cure for AIDS — and for cancer and other killer diseases — should be apparent.

One must set up a modern research laboratory and assemble as many of the "dirty dozen" together as is possible. One then tackles the problem head-on, adding the necessary support staff and special consultants.

First the Kaznacheyev effect for AIDS must be isolated and determined. Then it must be reversed, to yield the precise "curative" signal.

Transforms for these Kaznacheyev IR/UV "death photons" and their phase conjugated "healing photons" must be obtained for a lower electromagnetic frequency band — such as, say, 10 to 20 gigahertz.

Away must be found to irradiate the whole human body with the curative signal. The ideal way is to utilize a scalar EM curative transform so that the atomic nuclei of the body — and hence its entire master cellular control system and immune control system—will be "charged up" with the correctly structured "AIDS cancellation message."

The entire procedure and apparatus must be as simple and small as possible. My colleagues have already achieved very promising — even remarkable — progress in this respect.

Reversing the Kaznacheyev Effect

The "Kaznacheyev effect" for AIDS virus condition must be stimulated between cell cultures, so that the disease condition is electromagnetically transferred from one to another. The actual electromagnetic "delta" constituting the contribution of the AIDS infection must then be isolated electronically. The best way to do this is probably to subtract the normal cell radiation pattern from the "cell plus AIDS" radiation pattern.*

With the "AIDS delta" determined, the delta is then fed into the appropriate phase conjugating mirror system, so that its time reversed replica is produced.

The new phase conjugated signal is then the required *AIDS reversal signal* to reverse the effects of the AIDS virus itself, inside the cell where it resides dormant. This " time-reversed" signal will reverse the genetic change in the cell, not just kill the HIV virus.

In a crude way, one is making an **electromagnetic anti-virus.**

For ease of development and treatment, microwave technology is most attractive. Obtaining transforms of the signals in the radar band is ideal, since a wide variety of techniques, instruments, and electronic parts are available for that region. Millimeter waves would be most attractive, for the equipment could then be highly miniaturized.

One can regard it another way also: energy forms (critters) are involved. The action of the AIDS virus in its host cell, upon the DNA of the host, is underlaid by manipulation of energy critters. If one makes the electromagnetic anti-virus form, one is also manipulating the energy critters in their virtual state substrata.

The net result is that essentially the virus pattern — even the virus itself — can be phase conjugated by the energy critters. The result can be to turn the actual virus in the cell into a negative virus, accomplishing recombinant DNA procedures in reverse.

Remember that the signal we seek to use involves negative energy and negative time. We are also engineering the virtual state directly. The ordinary positive energy/positive time/observable state rules and limitations do not necessarily apply. And the "fixed form" first order physical reality as we normally conceive it need not be so fixed at all.

---

*Procedures along the lines of the extraordinary double-exposure holographic work of Dr. Robert Powell will probably be necessary. His work on biologically significant spatial frequency spectroscopy has blazed the trail as to how to obtain the specific delta patterns desired. It is hoped that Powell will shortly publish the remarkable results of his 15 years' work.

Physical reality itself can be directly engineered.

The engineering we seek to accomplish is directly upon the probability states propagated by the Schroedinger equation, before observation and collapse of the wave function occurs. **We seek to engineer physical reality before it is born, while it is yet forming.**

Only at such a level can the previous action of the AIDS virus — that in which it combined its genetic material with the genetic material of the host cell — be reversed and undone.

Only at such a level can we convert the infected human body from an "AIDS virus factory" back to an uninfected normal human body without AIDS.

The Proof: Priore's Work

Antoine Priore's pioneering work largely proves that it can be done. Cancers, leukemias, and many other virulent diseases yielded to his phase conjugated signals passed down through a powerful magnetic field to totally penetrate every cell in the treated patient's body.

And Royal R. Rife's work proved that a virus and a bacterium are not at all the "rigidly fixed" physical forms that our normal science has led us to believe they are. Instead, both the organisms and their biochemical and genetic actions can be addressed — and changed — on a much finer level of reality.

Of course, it would be enormously helpful if one had a working Rife microscope.

My colleagues are attempting to rebuild one of the original Rife microscopes, which has parts that were missing from it. They have every hope of having the microscope in action in the future.

Another angle of attack is also possible.

One of my colleagues has discovered a very peculiar, weak electromagnetic signal that will kill viruses, harmful bacteria, toxic protozoa, etc. but not harm living human cells. Only a few volts and a few milliamps are used.

However, there exists a major problem in getting any such very weak signal into every cell in the body — which is required if one is to heal blood diseases such as AIDS and leukemia. After all, that was the reason that Priore utilized a powerful magnetic field of thousands of gauss. The magnetic field penetrated every cell in the body — even those in the bone marrow where all the blood cells are manufactured. By using the all-penetrating field as the carrier, the phase conjugated healing signal pattern could thus be introduced into every cell in the body, bathing it completely, inside and out, with the restorative signal.

Remember, **one must not just get the signal into the cells themselves — instead, one must get the restorative signals directly into, and absorbed in, the atomic nuclei.**

Again, that is why Antoine Priore found it necessary to employ a "rippling" magnetic field. The "ripple" was actually a magnetic wave, and nuclear resonance then provided the magic mechanism to penetrate all the atomic nuclei.

So the initial problem is, how does one provide a mechanism to carry the desired signals into and through each and every cell of the body and into each and every atomic nucleus of the matter of the body?

Obviously one can utilize nuclear magnetic resonance, after the fashion of Priore. If so, the resulting apparatus is going to be extremely large and expensive. It would be highly desirable to do it a different, simpler, much cheaper way.

After many hundreds of back-breaking experiments, one of my colleagues appears to have discovered a completely unique and direct way of introducing the desired EM restorative signals into and completely through every cell of the body, and into every atomic nucleus. Though much additional work to confirm this still remains, the initial results are marvelously encouraging.

Another colleague has succeeded in developing a peculiar sort of detector that should prove adaptable to detecting the actual "biopotential structural patterns" themselves, directly in and out of the atomic nuclei. Though obviously much more work is necessary before the final instrumentation is ready, the preliminary results are again most encouraging.

As I write these words on paper, this work is proceeding, but very, very slowly due to lack of the necessary funds to attack the research problems in force.

What is Needed

Time is running out —most conventional science appears to be driven by special interest groups/Drug Manufacturers who are unable and/or unwilling to counter the AIDS epidemic with anything other than extensive "Addict Style" symptom-reducing drugs that generate dollars, not cures. Though this is not universal, it *is* the conventional norm.

Unconventional disorders or diseases require *unconventional* science for *unconventional* cures.

Immediate funding is required if this awesome threat to humanity is to be stopped. An American public alerted in time to this desperate situation can demand that the government and/or private business immediately address and act on this life-threatening issue in a different, unconventional manner.

Encouraging Preliminary Work

One of my associates, already familiar with scalar EM devices, has exerted every effort to try to reduce the scope of the problem. In literally hundreds of experiments, he has been able to narrow down the search, and obtain at least some very promising results.

He has obtained an initial candidate phase conjugated signal for further test and trial.

He may well have succeeded in discovering a new and unique method to communicate signals directly into the atomic nuclei inside the matter in the human body. This signal presently appears to directly interact with the cellular biopotential and with the body's master cellular communication system.

Remember, however, these are preliminary results. They still must be fully substantiated in a great many more tests. Undoubtedly a great deal more research, analysis, and adaptation is necessary. We certainly cannot presently say we have any sort of "cure" for the AIDS virus or anything else — or that these results are to be considered as proven in any **fashion.**

But what we can say is that my colleague has been able to derive a very complicated phase-conjugated signal, which produces negative energy and negative time of the general sort required. And he may just have made a most marvelous discovery that will point the way to eventual equipment much, much smaller than the 4-stories high machine with which Priore intended to treat terminal human cancer patients.

The initial signal discovered by my colleague, when applied to the body at miniscule voltage, seems to zap all sorts of "bad things" — viruses, harmful bacteria, dangerous protozoa, microworms, you name it — without harming the blood cells, the normal body cells, or the hosted friendly bacteria.

The signal does this at miniscule voltage and amperage.

It can be applied directly to the body through special electrodes. Through a special feature, my colleague has been able to get the signal to reverberate the entire body, all cells, all parts, and even penetrate the atomic nuclei and establish scalar resonance therein.

There turned out to be some extremely strange things that have to be done to the phase conjugated signal before it will accomplish what is being sought. At least some of these "strange things" have been uncovered by my colleague.

One of the peculiarities is that the entire electrical apparatus is part of the input "form" (that conditions the potential wave structure) being phase conjugated and sent to the organism. If a lead-acid battery is included in the apparatus to power it, one will inject the electromagnetic form for the battery acid directly into the organism,* destroyingit. This includes destroying the host's cells. In this case the signal is lethal, not curative. Substitution of a dry cell battery with no liquid electrolyte eliminates the problem. Exactly why a liquid electrolyte has a toxic effect and a "sludge" or solid electrolyte does not, is not understood at this time.

Other such anomalies in the tentative process have been discov ered and compensated for.

However, the way ahead is exciting. It suggests that the body

*Via a mechanism similar to that found by Reid and Barsamian (atomic nuclei) can literally be "charged up" (i.e., the living biopotentials can be "charged up" with the signal structure) so that the "disease-proofing" is very lasting, possibly for many years or even for a lifetime.

At least my colleague has pressed this to the point of demonstrating a long-lasting charge being acquired by the body.

For example, at one time his body became so "charged" from his lengthy experiments that a one-inch blue spark often leaped from his fingers when he reached out for something metallic. The discharge was cool, negative energy — living energy, if you will. It should be negentropic, not entropic. It was definitely not the type of energy the orthodox scientific community is accustomed to. And normal electricity will definitely not charge up the body in such a continuing fashion, so far as is known.

Let me clearly state again that we have not yet produced the specific anti-pattern per se. What my colleagues have discovered appears to be a broad-band signal that appears to act hyperspatially, analogous to the manner in which a broad-band drug such as penicillin acts biochemically. Even this much remains to be clearly established.

However, it is a most encouraging and promising first step.

Much more work, and a great deal of experiments to substantiate or adapt these tentative results, still need to be done. Now there is no substitute for rigor and thoroughness — and there is no substitute for clearly and scientifically demonstrating the proof of the concept in the laboratory.

It is not just good intentions that we seek, but solid, concrete, proven results substantiated by proper scientific procedures. Much work remains to be done.

But the preliminary results are very encouraging indeed.

Let me briefly share with you some of the things we foresee, if this present line of successful development continues.

We foresee being able to eventually develop and set up — legally and under proper medical auspices, of course — tested and proven devices that can easily treat up to two or three hundred persons at once. A treatment of about 45 minutes to one hour is all that would be required. Several repetitive treatments a week or so apart, might prove advisory.

In addition,once the entire gamut of the treatment process is validated and proven, and shown to be completely harmless and safe in accordance with legal medical requirements, if need arises we foresee simply adding the signal to ordinary radio and television transmitters — perhaps as simply as modulating the electrical ground. If so, a "maintenance" signal could be established to negate the AIDS virus (or other disease such as cancer and leukemia) in an entire area, and keep out other diseases such as malaria, sleeping sickness, etc.

Another advantage of such machines would be their portability in time of conflict. Also, they would be most useful indeed as defensive measures against biological warfare. For example, the inhalation of only a relative minor amount of anthrax agent is sufficient for 100 percent certain death unless treatment begins promptly. With the portable machines, however, this would be easily negated in any troops exposed to the agent, even before the lethal disease is evidenced. For a totally new agent, a "phase-conjugate delta signal" could be rather quickly ascertained and developed, and the proper settings made on the machines for immediate treatment and immunizing.

And even later, when we proceed to the direct engineering of the living energy form strata themselves, we shall see remarkable cures and remissions of diseases for which medical science offers little hope today. Arthritis, multiple schlerosis, lupus, and other such debilitating diseases come readily to mind. **Even reversal of the aging process should be possible.***

*Two other promising approaches have recently been discovered First, an electrolyte compound has recently been approved by the FDA for clinical testing. This compound has the remarkable property of raising the cellular electrical potential back to that of a strong, healthy cell. In several years of lab animal testing, this alone was indicated to be over 80% effective against cancer. It may also prove effective against diseases such as arthritis, where the body's immune system attacks body tissue with lowered cellular potentials, since it fails to recognize the weakened cells as those of the body. Second, Baylor University reseachers have found that treating blood with certain laser EM radiation kills the AIDS virus, but does not harm the blood cells This means that it will be possible to assure that blood used in transfusions will be AIDS - free, eliminating one sorce of AIDS transmission.

Conclusion

We have now come to the end of our road, literally and figuratively.

All of us have been struck a mortal blow by the Soviet AIDS first strike.

Make no mistake, this is **real.** We and our children are already as good as dead unless we move as we have never moved before.

We have a chance. A slim chance.

Americans have always come through when the chips are down. We can conquer this thing. We can overcome this mighty death blow that has been launched against us. We can defeat the others yet to come.

But we've got to move. *Now.*

Remember, Pandora's box has already been spilled. Even without the Soviet biological warfare strike, hosts of new viruses and different strains of old ones are going to be, and are now being, dumped into the biosphere by our own culture. It is also only a matter of time before terrorists and meglomaniacs turn to the use of this potent weapon against a wide-open society such as ours.

Both our Armed Forces and our civilian populace are totally defenseless against electromagnetic biological warfare. Now. At this moment.

Even without the Soviet BW strike implication, it is only a matter of time until we perish, unless we develop electromagnetic healing and electromagnetic biological warfare countermeasures.

We've come to one of those profound momentary pauses in history that determine the fate of the entire world henceforth.

It's like the parable of the lady or the tiger.

We're facing, so to speak, two doors.

Behind one is the most fearsome and hungry tiger of all time. If we delay, that door will open and we'll get the tiger full upon us. We shall be utterly destroyed. That mighty tiger will consume half the world in his roaring frenzy. Our children, and our children's children — what few of them will be left — will wear the hammer and sickle yoke for eons.

Those few future survivors will be taught strange things. How you and I were the real enemy. How we were absolutely destroyed for the good of all mankind. How glorious and necessary it was to unleash the great plagues upon us. And how heroic were those who performed the "noble deed."

Our own distant children will curse us and revile us, and they will be taught to worship at the throne of a false prophet.

The world will descend into a new "Dark Ages" far more frightful than George Orwell ever envisioned.

On the other hand, behind the second door is the most beautiful lady of all history. Literally all of humanity's dreams of health, beauty, and vitality lie behind that door.

If we open it, we achieve a freedom from disease and a measure of bountiful health for all mankind that has heretofore only been dreamed of. *Even reversal of the aging process itself lies beyond the second door. Health and youth — the dream of the ages—can be ours.*

But we have only a moment to open the second door. It is firmly shut, and we must exert ourselves to the fullest if we are to open it at all.

On the human stage, the first door is already slowly opening, inexorably. In only a few moments it will be open and the tiger will be upon us.

We must move quickly.

Which will it be for mankind, the lady or the tiger?

The next few moments in the human play will most assuredly tell.

SELECTED BIBLIOGRAPHY

Note: An extended bibliography for this book would consist of literally thousands of references. Since there has not been time for that, a few key references on most substantive matters are listed. Where possible, one or more of the listed references will direct the interested reader to the pertinent literature on that subject.

References Related to Chapters 1 and 2

1. Pearce Wright, Science Editor: "Smallpox vaccine 'triggered Aids virus'," **The Times** of London, Mon. May 11, 1987, p. 1, 18.

2. Jon Rappoport, "News blackout on pox vaccine link to AIDS protecting WHO?", **Easy Reader,** June 4, 1987, p. 12.

3. Jonathan Tennenbaum, "Parameters for a research mobilization against AIDS," **Executive Intelligence Review (EIR),** Science & Technology, Mar. 6, 1987. p. 38-45.

4. Steve Connor, "AIDS: mystery of the missing data," **New Scientist,** Feb. 12, 1987, p. 19.

5. Steve Connor, "AIDS: Science stands on trial," **New Scientist,** Feb. 1987, p. 49-54.

6. "The virus reveals the naked truth," **New Scientist,** Feb. 12, 1987, p. 55-58.

7. Michael Gold, "Cancer Cover-up: How honest scientists evaded the issue of renegade cells that contaminated research labs worldwide," **Science Digest,** May 1986, p. 64-67, 80-81.

8. Michael Gold, **A conspiracy of Cells,** State University of New York Press, Albany, New York, 1985.

9. Kenneth J. McLeod, Raphael C. Lee, and H. Paul Ehrlich, "Frequency Dependence on Electric Field Modulation of Fibroblast Protein Synthesis," **Science,** Vol. 236, Jun. 12,1987, p. 1465-1468.

10. "Leukemia Virus Linked to Nerve Disease," **Science,** Vol.236, May 29, 1987, p. 1059-1061.

11. R.C. Desrosiers, "Origin of the human AIDS virus," **Nature,** vol. 319, 1986, p. 728.

12. R.B. Strecker, "AIDS virus infection," J.R. **Soc.** Med., Vol 79, 1986, p.559-560.

13. S.A. Aaronson, "Common Genetic Alterations of RNA Tumor Viruses grown in Human Cells," **Nature,** Vol. 230, 1971, p. 445-447.

14. A. Chandra et al, "Serological relationship between reverse transcriptase from human T-cell lymphotropic viruses defined by monoclonal antibodies," FEBS Lett., Vol. 200,1986, 327-332.

15. E. Dermott, J.K. Clarke and J. Samuels, "The Morphogenesis and Classification of Bovine Syncytial Virus," **J. Gen. Virol.** 1971; 12:105-119.

16. G. Corneo and L.C. Nelli, "Could Bovine Leukemia Virus be a possible Agent of Some Human Lymphatic Leukemias?", **Acta Haemat.** 72; 1984:65-66.

17. C.A. Mims, "Vertical Transmission of Viruses," **Microbiol. Rev.** 45; 1981:267-286.

18. C.W. Molander et al, "Isolation and Characterization of Vi ruses from Fetal Calf Serum," **In Vitro** 1972; 7:168-173.

19. P. Legrain, B. Goud and B. Buttin, "Increase of Retroviral Infection in Vitro by the Binding of Antiretroviral Antibodies," **J.Virol.** 1986; 60:1141-1144.

20. S. Kennedy -Stoskopf and O. Narayan, "Neutralizing Antibodies to Visna Lentivirus: Mechanism of Action and Possible Role in Virus Persistence," **J. Virol.** 1986; 59:37-44.

21. E.K. Knott, "Development of Ultraviolet Blood Irradiation, **Am. J. Surg.** 1948; 76:165-171.

22. Dr. John Seale, "The Origins of AIDS," Mar. 15, 1987.

23. R.B. Strecker and T.A. Strecker, "The Bioattack Alert (on AIDS)," Los Angeles, California, 1986.

24. R.B. Strecker, "Aids Virus Infection," J.R. **Soc. Med.,** vol. 79, Summer 1986, p. 559-560.

25. T.E. Bearden, "AIDS: Urgent Comments on Mankind's Greatest Threat," June 1987. Privately printed and circulated. "Aids: Urgent Comments on Mankind's Greatest Threat; and the Secrets of Electromagnetic Healing," presented in absentia at the U.S.P.A. Annual Symposium, July 30-Aug. 2,1987, Golden, Colorado.

26. **Executive Intelligence Review (EIR)** Special Report, "Global Showdown: The Russian Imperial War Plan for 1988," July 24, 1985.

27. **Executive Intelligence Review (EIR)** Special Report, "The Soviet role in covering up the deadly threat of AIDS," 12(42), Oct. 25,1985. Charges that Soviet infectious-disease experts control the Communicable Diseases Division of the World Health Organization. Nearly hysterical responses were stimulated by this report: for example, less than a week later, **Literaturnaya Gazeta,** Oct. 30, 1985 published a counter, and claimed the AIDS virus may have been produced in U.S. laboratories by the CIA and the Pentagon.

28. Jeanne Toal, "Supervaccines," **Omni,** July 1987, p. 20, 122.

29. The Kaznacheyev work. See appropriate section for references.

30. The Priore work. See appropriate section for references.

31. William Kucewicz, "The gates slam shut on a microbiologist," "Beyond Yellow Rain: The Threat of Soviet Genetic Engineering," Sixth of a Series, **Wall Street Journal,** May 8,1984. p. 34.

32. Department of Defense, **Soviet Military Power,** 1987.

33. William Kucewicz, "A non-stop Russian response to WWI," "Beyond Yellow Rain: The Threat of Soviet Genetic Engineering," Seventh of a Series, **Wall Street Journal,** May 10,1984. p.34.

34. Lewin, Leonard C, **Report from Iron Mountain on the**

Possibility and Desirability of Peace. The Dial Press, New York, 1967. If you want to read something that will curdle your blood, take a look at this book. There exists a surprising body of material where intelligentsia and "think tank" mentalities have grappled with the idea of non-warfare types of "control mechanisms" for humanity — including induction of biological disease on a mass scale, subtle forms of slavery, etc.

35. Bennett,William and Joel, Gurin, "Science that Frightens Scientists," **Atlantic Monthly,** Feb. 1977. Millions of persons in this country were inadvertently injected with a widely-used substance called simian virus 40, or SV40. Tests proved that SV40 caused cancerous cell growth when injected into animals. SV40 was an unrecognized contamination of polio vaccines given before 1962 and of some cold shots given about the same time. The authors conclude by stating: "The ultimate question is not whether bacteria can be contained in special laboratories, but whether scientists can be contained in an ordinary society."

36. Letter, Lynn M. Draft, White House Policy Staff, to Dick Gregory, April 11, 1977. Quoting: "It is true that the SV40 virus was discovered in the early 1960's. It is also true that the virus was found in certain viral vaccines prepared from virus pools grown in monkey kidney cell cultures. However, the virus had not been recognized as a contaminant prior to that time, although millions of people have received the vaccine during the 1950's."
"... contamination of the vaccine was unintentional due to lack of knowledge."

37. Snead, Dr. Eva Lee, M.D. "AIDS - Immunization related syndrome," **Health Freedom News,** July 1987, p. 14-17, 22, 44-45. Strongly and rigorously points out just now little we actually know, even today, about the cause of AIDS and some of the anomalies connected with it. Specifically, SV40 does not appear to have been searched for in AIDS victims. Lack of detection has little meaning, since SV-40 may disappear and not reappear until triggered or promoted by another virus or chemical. Yet SV-40 has been shown to be one of the activators of the HIV or AIDS virus, as reported by Howard E. Gendel man et.al, in an article published in the **Proceedings of the National Academy of the United States of America,** vol. 83, p. 9759-9763, Dec. 1986. Also, SV-40 carries other viruses

"piggyback" and may have that type of role in AIDS disease. SV-40 may predispose one to a secondary infection by nullifying the immune system. SV40 is ideally suited for genetic manipulation, splicing, and the creation of hybrids or mutants.' Dr. Snead rigorously points out that no one has proved as yet that any virus positively causes AIDS. What has been proven is that HTLV-III is present in large percentages of the AIDS victims. Dr. Snead raises many excellent issues which cry out for answers.

38. Pearl Ma and Donald Armstrong, **The Acquired Immune Deficiency Syndrome and Infections of Homosexual Men,** Yorke Medical Books, 1984.

39. Ostrow, David G; Terri A. Sandholzer and Yehudi M. Felman, Eds., **Sexually Transmitted Diseases in Homosexual Men,** Plenum books, 1984.

40. Snead, Dr. Eva Lee, M.D., **aids-update,** Jun. 15,1987. In this incisive summary, Dr. Snead "tells it like it is" about Salk vaccine, how vaccines are made, the lack of sufficient scientific accuracy and safety in making vaccines, the nearly general contamination of vaccines by at least some viruses, some macabre details of the constitution of the vaccine 'soups." etc.

41. Gallo, Robert C, "The AIDS Virus," **Scientific American,** 256 (1), Jan. 1987, p. 46-56.

42. **Institut Pasteur v. United States of America,** United States Court of Claims, Civ. A. No. 730-85. See also Steve Connor, "AIDS: Science Stands On Trial," **New Scientist,** Feb. 12, 1987, p. 49-58. Reveals some direct monetary interests involved in discovery of the AIDS virus, and in royalties for the blood test which detects the antibody to the virus. Extended litigation occurred between French scientists at the Pasteur Institute and Robert Gallo regarding who should receive credit for discovery of the virus causing AIDS. The two sides eventually agreed to share the credit for the discovery, and donate royalties to an AIDS research foundation. Note that Dr. Gallo is head of the Laboratory of Tumor Cell Biology at the National Cancer Institute.

43. Curran, James W. et al, "The Epidemiology of AIDS: Current Status and Future Prospects," **Science,** vol. 229, Sep.**27,**

1985, p. 1352-57. Among other things, this report downplays the role of non-viral agents in causing AIDS.

44. Duesberg, Peter H., "Retroviruses as Carcinogens and Pathogens: Expectations and Reality," **Cancer Research,** vol. 47, Mar. 1,1987, p. 1199-1220. Cites 278 references. Duesburg is a pioneer in the study of retroviruses and a professor in the Department of Molecular Biology and Virus Laboratory, University of California at Berkeley. To show some of the scientific disagreement as to the cause of AIDS, we quote from Dr. Duesberg's article: "AIDS virus is not sufficient to cause AIDS and ... there is no evidence, besides its presence in a latent form, that it is necessary for AIDS." Dr. Duesberg presents an impressive set of technical objections to the present, theory that the accepted AIDS virus actually causes AIDS.

45. **Steadman's Medical Dictionary,** Williams and Wilkins, Baltimore, 1966. p. 874. A German bacteriologist named Robert Koch established three "laws" to determine whether a certain micro-organism is the cause of a specific disease. These laws provide a basis for scientifically investigating infection and its causes. Koch's laws are: (1) Whether it be a bacterium, protozoan, or virus, the micro-organism must be present in all cases of the disease. (2) Inoculating animals with pure cultures of the micro-organism must produce the same disease in them. (3) It must be possible to obtain the micro-organism from these animals and further propagate the micro-organism in pure cultures. (Note that this leaves a real problem with any disease caused by complexes of different organisms, rather than just one.)

46. Altaian, Lawrence K., "Aids Development in Infection; Data Suggest AIDS Rises Yearly After Infection," **New York Times,** Mar. 3, 1987, p. C1. Among other things, points out that the AIDS seropositivity rate rose from 4 percent in 1978 to 68 percent in 1984 among 6,875 members of a San Francisco hepatitis B study group.

47. Moyo, "Fighting AIDS in Africa," **AfricAsia,** May 1987, p. 55

48. Altaian, Lawrence K., "AIDS in Africa," **New York Times.** Nov. 1, 1986, p. A1, A8.

49. Krieger, Nancy, "The Epidemiology of AIDS Riddle," **Science for the People,** Jan./Feb. 1987, p. 18.

50. Nordheimer, Jan, "Florida Pig Farm Poses AIDS Riddle," **New York Times,** May. 2, 1986, p. A7.

51. Saul, Stephanie, "A Focus for the AIDS Mystery," **Newsday,** Jan. 26, 1986, p. 7.

52. Segal, Jakob and Lilli, **AIDS: USA Home-Made Evil: Not Imported from Africa,** date unknown (believed 1986). Pamphlet widely circulated in Africa. The two scientists are a biologist and a biophysicist, affiliated with the Humbolt University of Berlin. In the pamphlet the authors charge that the AIDS virus was manmade, and suggest that it may have been created at Fort Detrick, Maryland. This and other such literature, plus deliberate Soviet propaganda releases charging that "U.S. biological warfare" is responsible for the AIDS pandemic, have significantly influenced widespread belief in Africa that the U.S. government is responsible. Our own thesis is that this merely shows how clever and coordinated the Soviet BW strike really is: the deception plan was complete and ready made. The viruses were made elsewhere (in cancer research labs), someone else unwittingly injected the contaminated vaccines (World Health Organization), and a good cover story was immediately at hand (simply blame the U.S.).

53. Pearce Wright and Thompson Prentice, "WHO seeks evidence over vaccine link to Aids virus," **The Times** of London, May 12, 1987.

54. "Near disaster with the Salk vaccine," **Science Digest,** Dec. 1963.

55. Bernard Fields, Malcolm A. Martin, and Daphne Kamely, Eds., "Genetically Altered Viruses and the Environment," **Banbury Report,** Cold Spring Harbor Laboratory, 1985.

56. The following summary articles are in Science, 239 (4948), Feb. 5, 1988: (1) Peter Piot et al, "AIDS: An International Perspective," p. 573-579; (2) Anthony S. Fauci, "The Human Immunodeficiency Virus: Infectivity and Mechanisms of Pathogenesis, " p. 617-622; (3) James W. Curran et al, "Epidemiology of HIV Infection and AIDS in the United States," p. 610-616; and (4) Richard W. Price et al, "The Brain in AIDS: Central Nervous System HIV-1 Infection and ATDS Dementia Complex," p. 586-592.

References Related to Chapter 3

1. Publications by **T.E.** Bearden:

- **a.** **Fer-de-Lance: A Briefing on Soviet Scalar Electromagnetic Weapons.** Tesla Book Co., POB 1649, Greenville, Texas 75401, 1986. Extensive coverage of Soviet scalar EM weapons and their testing. Heavily documented and illustrated. See a review of **Fer-de-Lance** in **Defense Science & Electronics,** June 1987, p. 80.

- b. **Excalibur Briefing.** Foreword by John White. Strawberry Hill Press, San Francisco, California 1980. An updated second edition, jointly published by Strawberry Hill Press and Tesla Book Co., 1988.

- c. **Tesla's Secret and The Soviet Tesla Weapons.** 1 hr. 40 min. videotape. 1981. Available from Tesla Book Co. Generally poor quality; lots of information.

- d. "Soviet Psychotronic Weapons: A Condensed Background." **Specula,** Journal of the AAMS, POB 1182, Huntsville Alabama 35807. 1(2), Mar.-June 1978. pp. 18-32. (Journal is now defunct).

- e. **Star Wars Now! The Bohm-Aharonov Effect, Scalar Interferometry, and Soviet Weaponization.** Tesla Book Co., 1984.

- f. **Soviet Weather Engineering Over North America**. 1-hr, videotape, 1985. Available from Tesla Book Co. Excellent quality; a great deal of information.

- g. **Solutions to Tesla's Secrets and the Soviet Tesla Weapons.** John T. Ratzlaff, ed. **Part II: Reference Articles for Solutions to Tesla's Secrets.** Tesla Book Co., 1981.

- h. "Tesla's Electromagnetics and Its Soviet Weaponization," **Proceedings of the 1984 Tesla Centennial Symposium,** International Tesla Society, Colorado Springs, Colorado 1984.
- i. USSR: New Beam Energy Possible?", **Defense & For-**

eign Affairs Daily, 13 (111), June 12, 1984, p. 1-2.

j. T.E. Bearden and Andrew Michrowski, eds., The **Emerging Energy Science,** Planetary Association for Clean Energy, Ottawa, Canada, 1985.

k. "AIDS: Urgent Comments on Man's Greatest Threat; and the Secrets of Electromagnetic Healing," presented in absentia at the U.S.P.A. Annual Symposium, July 30-Aug. 2, 1987, Golden, Colorado.

1. "Soviet Phase Conjugate Weapons: Weapons that use Time-Reversed Electromagnetic Waves," **Bulletin,** Committee to Restore the Constitution, POB 986, Ft. Collins, Colorado 80522, Jan. 1988.

m. "Tesla Electromagnetics: Weapons or Healing?" Presented at the Whole Life Expo, Los Angeles Airport Hilton Hotel, Los Angeles, California, Oct. 3, 1987.

2. William C. Douglass, **The Cutting Edge,** Nov. 1987, p. 3.

3. Stefan T. Possony, "Psy-War: Soviet Device Experiment," **Defense & Foreign Affairs Daily,** 12(104), June 7,1983, p. 1-2. Reports on Dr. Ross Adey's investigation of the Soviet LIDA device which is used to bombard human brains with radio waves of complex waveform in the 40 MHz region. Also reports on work by A.S. Davydov of the Ukranian Academy of Sciences, who discovered how the blood-brain barrier can be penetrated by low-frequency radio waves so that brain cells are affected.

4. Stefan T. Possony, "The Tesla Connection," **Defense & Foreign Affairs,** Aug. 1984, p. 12-14, 27. Highly recommended.

5. David Jones, "Israel's Secret Weapon," **Weekend Magazine,** Dec. 17, 1977.

6. Daniel A. Walker, Charles S. McCreery, and Fermin J. Oliveira, "Kaitoky Seamount and the Mystery Cloud of 9 April, 1984," **Science,** 227(4584), Feb. 8,1985, p. 607-611. See also Daniel L. McKenna and Daniel A. Walker, "Mystery Cloud: Additional Observations," **Science,** 234(4775), Oct. 24,1986, p. 412-413 for update to the information in the first article.

7. Robert J. Durant, "An Underwater Explosion — or What?",

Pursuit, 5(2), April 1972, p. 30-31.

8. Gwynne Roberts, "Witness to a Superweapon?", The **Sunday Times,** London, England, Aug. 17, 1980.

9. Central Intelligence Agency (CIA) Information Report No. CO-B-321/15354-66, "Sighting of Unusual Phenomenon on Horizon Near Iranian/USSR Border," Sep. 8,1966, released on 15 Dec. 1978 under the Freedom of Information Act (FOIA).

10. CIA Foreign Intelligence Information Report CO B-324/ 33601-76, "Aerial Observation of intense Source of Light," Nov. 18, 1976, released on 15 Dec. 1978 under the FOIA.

11. Max Frankel, "Khrushchev Says Soviet Will Cut Forces a Third; Sees 'Fantastic Weapon'", **New York Times,** Jan. 15, 1960, p. 1.

12. Christopher S. Wren, "Brezhnev Calls for Accord against 'Terrifying Arms," **New York Times,** June 14, 1975, p. 1,11.

13. Malcolm W. Browne, "Senatorial Group Received by Brezhnev," **New York Times,** July 3, 1975, p. 2.

14. Christopher S. Wren, "Moscow Now Pressing Disarmament," **New York Times,** Aug. 12, 1975, p. 6.

15. Draft agreement introduced by Andrei A. Gromyko to the United Nations' thirtieth session of the General Assembly on Sept. 23,1975: "Prohibition of the Development and Manufacture of New Types of Weapons of Mass Annihilation and of New Systems of Such Weapons." The first article provided that the types of these new weapons would be "specified through negotiations on the subject."

16. "Secret Speech: Did Brezhnev Come Clean?", **National Review,** 29(8), Mar. 4, 1977, p. 248, 250.

17. William Beecher, "Brezhnev termed detente a ruse. 1973 report said," **Boston Globe,** Feb. 11, 1975, p. 1, 12.

18. See a series of articles by C.B. Baker in **Youth Action** NEWS,

POB 312, Alexandria, Virginia 22313: "Soviet Weather Mayhem." Nov., 1983; "Diabolic Soviet Warfare," Nov. 1984; "The Soviet Zapping of America," Dec. 1986; "Fifth Column Treachery," May 1987; and others.

19. "Jetliner drops 32,000 feet; 400 aboard; 50 are injured," Associated Press (AP) Release, **Huntsville (Alabama) Times,** Feb. 20,1985, p. 1. Also "China Airlines Pilot Denies Cockpit Error," AP Release, **Huntsville (Alabama) Times,** Feb. 21, 1985, p. A-9.

20. Paul Brodeur, **The Zapping of America,** W.W. Norton & Co., New York, 1977.

21. "Navy sabotages traffic lights, TV sets," United Press International (UPI), PM cycle, Jan. 11,1985; "Power blackout blamed on anti-radar particles," UPI, AM cycle, Jan 11, 1985; "Power Disrupted Briefly After Navy Plane Drops Metallic Strands," AP, AM cycle, Jan. 16 1985.

22. Bradley Wells, "The Russian woodpecker: a continuing nuisance," CQ **Magazine,** Nov. 1984.

23. "Explosive Events Seen on Soviet Island," **Aviation Week & Space Technology,** Sept. 26,1983, p. 31. Also Letter John M. Miller, Geophysical Institute, University of Alaska, Jan. 41985 with attachment, "Bennett Island plume cases recently found." Incredibly, the U.S. State Department is now attempting to **give away,** to the Soviet Union, the five-island chain of Wrangel, Herald, Bennett, Henrietta and Jeannette Islands and an outer continental shelf portion of Alaska twice the size of California! If this happens, the Bering Sea will become a Soviet pond.

24. Sam Bishop, "UA photos show Soviets break ice for missiles," **Fairbanks Daily News-Miner,** Dec. 15 1984, p. 1,3. See also Craig Covault, "Soviet ability to fire through ice creates new SLBM basing mode," **Aviation Week & Space Technology,** Dec. 10, 1984. See also "Wrangel over Wrangel and the Weather War," Don Bell Reports, 32(9), Mar. 8,1985. See also, "Wrangel Giveaway Imminent," **The Spotlight,** Jan. 18, 1988,p.1,3.

25. Philip J. Klass, "Anti-Satellite Laser Use Suspected," **Avia-**

tion Week & Space Technology, Dec. 8, 1975, p. 12-13.

26. Joe Schwartz, **Nature,** 280, July 12, 1979, p. 95.

27. V.D. Sokolovskiy, **Soviet Military Strategy,** third edition 1968, p. 298.

28. T.T. Wu, "Electromagnetic Missiles," **J. Appl. Phys.** 57,1985, p. 2370-2373. See also "Filter Center," article on SDI, **Aviation Week & Space Technology,** Feb. 29, 1988, p. 55.

29. Michael McClelland, "State officials bend an ear to boom talk in Panhandle," **News Tribune,** Fort Pierce, Florida, Apr. 22, 1984.

30. Susan Burgess, "NASA officials stumped by strange light, boom," **News Tribune,** Fort Pierce, Florida, Nov. 28, 1985.

31. Bernie Woodall, "Residents are baffled by boom," **News Tribune,** Fort Pierce, Florida, Nov. 27, 1985.

32. Julie Enders, "Star of wonder: booming, bright," **News Tribune,** Fort Pierce, Florida, Dec. 6, 1985.

33. April 8, 1986 Canadian Broadcast Network television news interview of three eyewitnesses to the crash of the Arrow Air DC-8 jet liner at Gander Air Force Base, Newfoundland on Dec. 8, 1985. A yellow light was seen on the aircraft.

34. "Arrow Air was absolutely safe, FAA official says," UPI release, **Birmingham (Alabama) Post-Herald,** Apr. 17,1986.

35. "Titan Explosion Cripples U.S. Launch, Surveillance Capability," **Aviation Week & Space Technology,** 124(17), Apr. 28, 1986, p. 16-19.

36. Zhores A. Medvedev, **Nuclear Disaster in the Urals,** Translated by George Saunders, W.W. Norton & Co., 1979, p. 150

37. Gary North, "Electronic Aids (Parts 1 & 2)," **Remnant Review,** 12 (20), Nov. 1,1985, p. 1-8; 12 (21), Nov. 15,1985, p. 1-8. POB 8204, Fort Worth, Texas 76124.

38. Captain John D. LaMothe, **Controlled Offensive Behavior**

— **USSR,** Defense Intelligence Agency (DIA) Report ST-CS-01-169-72, 1972 (Released under FOIA).

39. Captain John D. LaMothe and Mr. Louis Maire, **Soviet and Czechoslovakian Parapsychology Research,** DIA Report DST-1810S-387-75, 1975. (Released under FOIA).

40. John Bentley, **The Thresher Disaster: The Most Tragic Dive in Submarine History,** Doubleday & Co., Garden City, New York, 1975.

41. Canadian Aviation Safety Board (CASB), **Public Inquiry Into the Aircraft Accident Which Occurred Near Gander Airport, Gander NFLD. December 12, 1985,** Vol. 1, Apr. 8, 1986. Hull, Quebec, Canada.

42. Department of Defense, **Soviet Military Power,** 1987.

43. "America's Space Program Sabotaged?" Robert Herzberg, **International Combat Arms,** Mar. 1987. p. 12.

44. Pearce Wright, Science Editor: "Smallpox vaccine 'triggered Aids virus'," **The Times** of London, Mon. May 11, 1987. p. 2.

45. Jon Rappoport, "News blackout on pox vaccine link to AIDS protecting WHO?", **Easy Reader,** June 4, 1987, p. 12.

46. R.B Strecker and T.A. Strecker, "The Bioattack Alert (on AIDS)," Los Angeles, California, 1986.

47. EIR Special Report, "The Soviet role in covering up the deadly threat of AIDS," 12(42), Oct. 25, 1986.

48. Dr. John Seale, "The Origins of AIDS," Mar. 15, 1987.

49. Martin Ebon, **Psychic Warfare: Threat or Illusion,** McGraw Hill Book Co., New York, 1983.

50. For details of mysterious, instant deaths in Afghanistan, see Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," **Jane's Defense Weekly,** 1(13), Apr. 7, 1984, p. 508.

51. " 'Observers' at the August Minuteman Shot?", **The INFO**

Journal, 2(1), Fall 1968, p. 35.

52. Dietrick E. Thomsen, "The Mystery of the Cosmic Gamma-Ray Zaps," **Science News,** 105 (22), June 1, 1974, p. 357-358.

References Related to Phase Conjugation (Time Reversal)

1. Amnon Yariv, **Optical Electronics.** Third Edition. Holt, Rinehart and Winston: New York, 1985. See particularly Chapter 16: "Phase Conjugate Optics — Theory and Applications."

2. Robert A. Fisher, Ed., **Optical Phase Conjugation.** Academic Press, New York, 1983. Collection of papers. An extensive bibliography is given at the end of the book. Note that Fisher's introduction to the phase conjugation effect was as a result of two Russian visitors who briefed scientists at Lawrence Livermore Laboratory on optical phase conjugation. Also note (p.xv, p. 20) that almost any nonlinear optical effect can phase conjugate an incoming beam.

3. B. Ya Zel'dovich and V.V. Shkunov, "Spatial-Polarization Wavefront Reversal in Four-Photon Interaction," **Soviet Journal of Quantum Electronics,** Vol. 9, 1979. p. 379.

4. B. Ya Zel'dovich et al, "Connection between the wave fronts of the reflected and exciting light in stimulated Mandel'shtam-Brillouin scattering," English translation, **Sov. Phys. JETP** 15,109(1972).

5. B. Ya Zel'dovich, N.F. Pilipetsky, and V.V. Shkunov, **Principles of Phase Conjugation,** Vol. 42, Springer Series in Optical Sciences, Theodor Tamir, Ed., Springer-Verlag, New York, 1985.

6. J.P. Woerdman, "Formation of a transient free carrier hologram in Si," **Opt. Commun.,** vol. 2, 1971. p. 212.

7. B.I. Stepanov, E.B. Ivakin, and A.S. Rubanov, "Recording two dimensional and three-dimensional dynamic holograms in bleachable substances," Translation: **Soviet Physics-Doklady-Technical Physics,** vol. 16, 1971. p. 46.
8. A. Yariv, "Three-dimensional pictorial transmission in optical

fibers," **Appl. Phys. Lett.,** vol. 28, 1976. p. 88.

9. A. Yariv, "On Transmission and recovery of three-dimensional image information in optical waveguides," **J. Opt. Soc. Amer.,** vol. 66, 1976. p. 301.

10. R.W. Hellwarth, "Generation of time-reversed wave fronts by nonlinear refraction," **J. Opt. Soc. Amer.,** vol. 67,1977. p. 1.

11. A. Yariv and D.M. Pepper, "Amplified reflection, phase conjugation, and oscillation in degenerate four-wave mixing," **Opt. Lett.,** vol. 1, 1977. p. 16.

12. R.L. Abrams and R.C. Lind, "Degenerate four-wave mixing in absorbing media," **Opt. Lett.,** vol 2, 1978, p. 94. Erratum is contained in **Opt. Lett.,** vol 3, 1978, p. 205.

13. P.F. Liao, D.M.Bloom, and N.P. Economou, "CW optical wavefront conjugation by saturated absorption in atomic sodium vapor, **"Appl. Phys. Lett.,** vol. 32, 1978, p. 813.

14. A. Yariv, "Compensation for atmospheric degradation of optical beam transmission," **Opt. Commun.,** vol. 21, 1977. p. 49.

15. R.W. Hellwarth, "Third order susceptibilities of liquids and gases," **Progress in Quantum Electronics,** vol. 5, 1977. p. 1.

16. J. Auyeung et al, "A theoretical and experimental investigation of the modes of optical resonators with phase-conjugate mirrors," **IEEE J. Quant. Electr.,** vol. 9, 1979. p. 1198.

17. I.M. Beldyugin and E.M. Zemskov, "Theory of resonators with wave-front reversing mirrors," **Sov. J. Quant. Elect.,** vol. 9, 1979. p. 1198.

18. A. Yariv and T.L. Koch, "One way coherent imaging through a distorting medium using four-wave mixing," **Opt. Lett.,** vol. 7, 1982. p. 113.

19. J.O. White and A. Yariv, "Real time image processing via four-wave-mixing in a photorefractive medium," **Appl. Phys. Lett.,** vol. 37, 1980. p. 5.

20. N.V. Kukhtarev et al, "Holographic storage in electrooptic crystals," **Ferroelectrics,** vol. 22, 1979. p. 949.

21. M. Cronin-Golomb et al, "Theory and applications of four-wave mixing in photorefractive media," **IEEE J. Quant. Elec,** vol. 20, 1984. p. 12.

22. J.P. Huignard, J.P. Herriot and G. Rivet, "Phase conjugation and spatial frequency dependence of wavefront reflectivity in $Bi_{12}SiO_{20}$(BSO) crystals," **Opt. Lett.,** vol. 5, 1980. p. 102.

23. J.O. White et al, "Coherent oscillation by self-induced gratings in the photorefractive crystal $BaTiO_3$," **Appl. Phys. Lett.**, vol. 40, No. 6, 1982. p. 450-452.

24. J.W. Haus, CM. Bowden, and C.C. Sung, "Optical phase conjugation with smooth pump profiles," **Phys. Rev. A, General Physics,** vol. 35, No. 8, Third Series, Apr. 15, 1987. p. 3398-3405.

25. Mary J. Miller et al, "Time response of a cerium-doped $Sr_{0.75}$ $Ba_{0.25}$ Nb_20_6 self-pumped phase-conjugate mirror," **Opt Lett.,** vol. 12, No. 5, May 1987. p. 340-342.

26. Mark Cronin-Golomb et al, "Passive (self-pumped) phase conjugate mirror: Theoretical and experimental investign tion," **Appl. Phys. Lett.,** vol. 41, No. 8, Oct. 15, 1982. p. 689 691.

27. Philip S. Brody and Richard P. Leavitt, "Dynamic holographic method of imaging phase objects," **Applied Optics,** vol. 26, No. 5, Mar. 1, 1987. p. 913-916.

28. J. Feinberg, "Self-pumped, continuous-wave phase conjugator using internal reflection," **Opt. Lett.,** vol. 7, 1982. p. 486.

29. T.Y. Chang and R.W. Hellwarth, "Optical phase conjugation by backscattering in barium titanate," **Opt. Lett.,** vol. 10, 1985. p. 408.

30. J.P. Huignard et al, "Phase-conjugate wavefront generation via real-time holography in $Bi_{12}Si0_{20}$ Crystals," **Opt. Lett.,** vol 4, 1979. p. 21.

31. David M. Pepper, "Nonlinear optical phase conjugation," guest editorial, **Optical Engineering,** vol. 21, no. 2, Mar./Apr. 1982. p. 155. One of the leaders of the field of nonlinear optical phase conjugation (NOPC) tells one in clear English what all the excitement is about. He also points out that the conceptual extension of these techniques to other portions of the electromagnetic spectrum should follow.

32. David M. Pepper, "Nonlinear optical phase conjugation," **Optical Engineering,** vol. 21, no. 2, Mar./Apr. 1982. p.156-183. This paper is the best introduction to, and summary of, nonlinear optical phase conjugation (NOPC) in the English language. Also, on p. 156 the author specifically notes that "...these processing techniques can, in principle, be extended to other portions of the EM spectrum (e.g.,, rt, radio, microwave, radars, UV, etc.); and can also involve other fields (e.g., acoustic waves), given the proper nonlinear medium." In other words, phase conjugation is a universal nonlinear phenomenon, unknown until recently. An extensive bibliography of over 200 references is included.

33. Jeffrey O. White and Amnon Yariv, "Spatial information processing and distortion correction via four-wave mixing," **Optical Engineering,** vol. 21, no. 2, Mar./Apr. 1982. p. 224-230.

34. R.W. Hellwarth, "Optical beam phase conjugation by four-wave mixing in a waveguide," **Optical Engineering,** vol. 21, no. 2, Mar./Apr. 1982. p. 263-265.

35. I.M. Bel'dyugin, V.N. Seminogov, and E.M. Zemskov, "Possible wavefront reversal of fields using nonlinear-optics methods," **Sov. J. Quantum Electron.,** vol 9, no. 3, Mar. 1979. p. 385-387.

36. I.M. Bel'dyugin, M.G. Galushkin, and E.M. Zemskov, "Properties of resonators with wavefront-reversing mirrors," **Sov. J. Quantum Electron.,** vol 9, no. 1, Jan. 1979. p. 20-23.

37. G.S. Agarwal, Ari T. Friberg, and E. Wolf, "Elimination of distortions by phase conjugation without losses or gains," **Opt. Commun.** 43 (6), Nov. 15,1982. p. 446-450. Using scalar wave theory, the authors show the conditions for complete cancellation of distortion by phase conjugation: (1) scatterer is non-absorbing, (2) infinite phase conjugate mirror, (3) conjugate

wave generated without losses, (4) effects of evanescent waves are negligible outside the scatterer. Comment: By reflecting upon this paper, one can grasp that essentially the same approach can be used to show that phase conjugation generates Newton's third law of motion.

38. G.S. Agarwal and Emil Wolf, "Theory of phase conjugation with weak scatterers," J. **Opt. Soc.** Am., 72(3), Mar. 1982. p. 321-326. The effect of a distorting medium on an incident wave is eliminated by phase conjugation if certain conditions are satisfied: (1) the incident wave contains no evanescent components, (2) the transmitting medium is a weak, nonabsorbing scatterer, (3) backscattering of incident waves and of conjugate waves is negligible, (4) effects of scattered evanescent waves are negligible. Comment: A similar approach can be used to show that phase conjugation generates Newton's third law of motion.

39. T. Sato et al, "Application of evanescent waves to microscopic observation," **Bull. Tokyo Inst. Technol. (Japan),** No. 125, 1974. p. 35-41. Evanescent wave illumination for microscopy can be used for recognizing higher frequency components impossible to recognize by conventional illumination.

40. S. Twareque Ali. "Evanescent waves in quantum electrodynamics with unquantized sources," **Phys. Rev. D,** 7(6), Mar. 15,1973. p. 1668-1675. An identity between evanescent waves and virtual photons is established.

41. J.D. Lawson, "Some attributes of real and virtual photons," **Contemp. Phys. (Great Britain),** 11(6), Nov. 1970. p. 575-580.

42. J.L. Agudin et al, "The behaviour of evanescent waves under Lorentz transformation," **Lett. Nuovo Cimento (Italy),** vol. 23, ser. 2, no. 15, Dec. 9, 1978. p. 547-551. Two linear combinations of transverse magnetic and electrical evanescent waves (the left-and right-handed circularly polarized modes) remain the same in all systems under Lorentz transformation.

43. M. Hercher, "virometer—an instrument for the measurement of the size of viruses using an optical microscope." **Proc. Soc. Photo-Opt. Instrumentation Engineers,** vol. 126,1978. p. 17-22.

44. G.A. Massey, "Microscopy and pattern generation with scanned evanescent waves," **Appl. Opt. (Poland),** 13 (3), 1983. p. 247-255. This article shows how to materially violate the usual "wavelength limitation" of optical systems. Much better resolution can be obtained, and very sharp images can be performed. Comment: This mechanism may be the secret of the Rife microscope, which could see viruses and other small organisms otherwise quite beyond normal optical microscopy's wavelength resolution.

45. H.J. Gerritsen, "Nonlinear effects in image formation," **Appl. Phys. Lett.,** vol. 10, 1967. p. 237.

46. D. Gabor, U.S. patent No. 2,770,166.

47. H.W. Kogelnik, "Controlled transmission of waves through inhomogeneous media," U.S. Patent 3,449, 577.

48. V.I. Bespalov. Ed., **Optical Phase Conjugation in Nonlinear Media,** IPF AN SSR, Gorky, USSR, 1979 (in Russian).

49. B. Ya. Zel'dovich et al, "Phase conjugation by a surface," **Dok. Akad.Nauk SSSR 252,**1980, p. 92. English translation: **Sov. Phys. Dokl.** 25, 1980, p. 377.

50. F.V. Bunkin, D.V. Vlasov and Yu. A. Kravtzov, "On sound phase conjugation with amplification of phase-conjugated wave," **Kvantovaya Elektron.** 8, 1981, p. 1144. English translation: **Sov. J. Quant. Electron.** 11, 1980, p. 687.

51. A.G. Gyulamiryan et al, "Tunable nonlinear four-wave filter," **Opt. Spektrosk.** 52, 1982, p. 387. (In Russian).

52. R.Y. Chiao et al, "Stimulated Brillouin scattering and coherent generation of intense hypersonic waves," **Phys. Rev. Lett.** 12, 1964, p. 592.

53. S.A. Akhumanov et al, **Problems in Nonlinear Optics (Electromagnetic Waves in Nonlinear Dispersive Media), INI** AN SSSR, Moscow, 1964. (In Russian).

54. N.B. Baranova and B.Ya. Zel'dovich, "Wavefront dislocations and zeros of the amplitude," **Zh. Eksp. Teor. Fiz.** 80,1981, p. 1789. English translation: **Sov. Phys. - JETP** 54, 1981, p. 925. Might one infer from this paper a cautious hint (and

probe) at "do you Westerners know of scalar EM?"

55. V.I. Bespalov, Ed., **Phase Conjugation in Nonlinear Media,** IPF AN SSSR, Gorky, USSR, 1982. (In Russian).

56. D. Gabor, "A new microscopic principle," **Nature** 161,1958, p. 777.

57. Yu F. Kir'yanov et al, "Four-wave mixing in resonantly amplifying media under inversion saturation," **Kvantovaya Elektron.** 8,1981, p. 1734. English translation: **Sov.J. Quant. Electron.** 11, 1981, p. 1047.

58. V.L. Vinetsky et al, "Amplification of coherent beams by dynamic holograms in ferroelectric crystals," **Izv. Akad. Nauk. SSSR** 41, 1977. p. 811.

59. V.P. Kondilenko, G.G. Odulov and M.S. Soskin, "Amplified reflection of phase-conjugate waves in crystals with the linear electrooptical effect in an external electrical field," **Izv. Akad. Nauk. SSSR, Ser. Fiz.** 45, 1981, p. 958.

60. Yu. I. Kucherov et al, "Copropagating four-beam interaction in slowly-responding media," in **Phase Conjugation in Nonlinear Media,** ed. by V.I. Bespalov, IPF AN SSSR, Gorky, USSR, 1982, p. 111.

61. A.P. Brysev et al, "Model realization of a NbLi parametric phase-conjugating sound amplifier," **Pis'ma Zh. Tokhn. Fiz. 1,** 1982, p. 546.

References Related to William J. Hooper's Work

1. William J. Hooper, "Apparatus for generating motional electric field." U.S. patent #3,656,013, Apr. 11, 1972.

2. William J. Hooper, "New horizons in electric, magnetic and gravitational field theory," presented at a meeting of the Particles and Fields Division of American Physical Societies, University of Colorado, Boulder, Colorado, Aug. 21, 1969.

3. William J. Hooper. "New horizons for the BxV theory of gravity," undated, unpublished.

4. William J. Hooper, "Equivalence of the gravitational field and

a motional electric field," Session S6, American Physical Society Meeting, Chicago, Illinois Nov. 28-29, 1958.

5. William J. Hooper, "Equivalence of the gravitational field and a motional electric field," **Proceedings of the Boulder Conference on High Energy Physics,** Colo. Assoc. Univ. Press, 1970. p. 483.

6. William J. Hooper, "The motional electric field generator," **Bull, of the Am. Phys. Soc,** Series II, 15(2), Feb. 1970. p. 209.

7. J.H. Jeans, **The mathematical theory of electricity and magnetism,** 4th edn., Cambridge University Press, 1923. p. 125, 606.

8. Henry W. Kendall and Wolfgang Panofsky, "The structure of the proton and the neutron," **Scientific American,** June 1971.

9. William J. Hooper, "All-electric motional electric field generator." U.S. Patent #3,610,971, Oct. 5, 1971.

10. William J. Hooper, "Similarities of the motional electric and gravitational fields," undated, unpublished.

11. P. Moon and D.E. Spencer, "Some electromagnetic paradoxes," **Journal of the Franklin Institute,** vol. 260, 2955. p. 373.

12. Leigh Page and Norman I. Adams, **Electrodynamics,** D. Van Nostrand Co., 1940.

13. William J. Hooper, **New horizons in field theory,** revised 1965, unpublished. Chapter 1, "Fundamental Fields," is particularly important.

14. G.I. Cohn, **Electrical Engineering** 68,1969. p.441. Pointed out difference between flux linking law and flux cutting law of induction. Comment: Recently the Lorentz force law has been proven false. A modified version of Ampere's law results and saves the day.

15. G.I. Cohn, **Paradoxes of Electromagnetic Induction,** thesis, Illinois Institute of Technology library.

16. Pharis E. Williams, **"The Possible unifying effect** of the **dy-**

namic theory," Los Alamos National Laboratory report LA-9623-MS, Los Alamos, New Mexico, May 1983 .

17. Pharis E. Williams, **On a possible formulation of particle dynamics in terms of thermodynamic conceptualizations and the role of entropy in it,** thesis, U.S. Naval Postgraduate School, 1976.

18. Pharis E. Williams "The Principles of the dynamic theory," Research Report EW-77-4, U.S. Naval Academy, 1977.

19. Pharis E. Williams, "The Dynamic theory: A new view of space, time, and matter," Los Alamos Scientific Laboratory report LA-8370-MS, Dec. 1980.

20. Pharis E. Williams, "The Arrow of time in the dynamic theory," Los Alamos Scientific Laboratory report LA-8690-MS, Feb. 1981.

21. Pharis E. Williams, "The Dynamic theory: Some shockwave and energy implications," Los Alamos Scientific Laboratory report LA-8402-MS, Feb. 1981.

22. Dan C. Ross, "Plane electromagnetic waves in the dynamic theory," Los Alamos National Laboratory report, date unk.

23. Dan C. Ross, "Speed of light in the dynamic theory," Los Alamos National Laboratory, date unk.

24. G. Gilbert and B. McClain, "Fermions and stability in quantum Kaluza-Klein theories," **Nuc. Phys. B, Part. Phy.s. (Netherlands),** B244(l), Sept. 24,1984. p. 173-185. Oscillations of the internal space produce conformal gravity waves in the four-dimensional space.

25. M. J. Perry, "Magnetic monopoles and the Kaluza-Klein theory," **Am. Inst. Phys. Conf. Proc.** No. 116,1984. p. 121-125. Kaluza-Klein theory admits soliton solutions which constitute magnetic monopoles. These monopoles have inertial mass, but no gravitational mass.

26. M.W. Kalinowski, "The nonsymmetric Kaluza-Klein theory." **J. Math. Phys.** 24(7), July 1983. p. 1835-1845.

27. D.J.Toms, "Induced Einstein-Maxwell action in Kaluza-Klein

theory," **Phys. Lett. B (Netherlands),** 129B(1-2), Sept. 15., 1983. p. 31-35.

References Related to Ehrenhaft's Work

1. Alden P. Armagnac, "Magic with magnetism," **Popular Science** Magazine, June 1944. p. 130-133, 222, 226.

2. "Ehrenhaft discovery confirmed by new experiments," **Popular Science** Magazine, date unk., p. 208. [Comment: Later than June 1944, and still during WWII.]

3. "Magnetic current — discovery of the age?", **Radio-Craft** Magazine, Mar., 1944. p. 332.

4. "Discovery of the age?" **Radio-Electronics** Magazine, date unknown, p. 58.

6. Felix Ehrenhaft and Emanuel Wasser, "New evidence of the existence of charges smaller than the electron," **Philosophical Magazine and Journal of Science,** Seventh Series, 5(28), Feb. 1928. p. 225-241.

8. Felix Ehrenhaft, "The microcoulomb experiment. Charges smaller than the electronic charge." **Philosophy of Science,** 8(3), July 1941. p. 403-457.

7. Felix Ehrenhaft and Leo Banet, "Is there 'true magnetism' or not?", **Philosophy of Science,** 8(3), July 1941. p. 458-462.

8. Felix Ehrenhaft and Leo Banet, "Magnetization of matter by light," Nature, Mar. 8, 1941. p. 297.

9. Felix Ehrenhaft, J. Franklin Inst., vol. 230, 1940. p. 381.

10. Felix Ehrenhaft, **Nature,** vol. 146, 1941. p. 25.

11. D. Morichini, **Gilberts Ann. Phys.,** vol. 43, 1813. p. 212.

12. D. Morichini, **Gilberts Ann. Phys.,** vol. 46, 1814. p. 367.

13. M. Sommerville, **Gilberts Ann. Phys.,** vol 52, 1826. p. 493.

14. F. Zantejeschi, **Gilberts Ann. Phys.,** vol. unk. 1829, p. 187.

15. V. Baumgartner, **Gilberts Ann. Phys.,** vol. unk., 1827. p. 508.

16. Felix Ehrenhaft, "Rotating action on matter in a beam of light," **Science,** 101(2635), June 29, 1945. p. 676-677.

17. Gordon Ferrie Hull, "The torque or rotating action in a beam of light," **Science,** vol. 101, 1945. p. 220.

18. F. Ehrenhaft, **Wiener Akad. Anz.,** vol. VII, Mar. 4, 1909.

19. F. Ehrenhaft, **Wiener Akad. Anz.,** vol. X, Apr. 21, 1910.

20. F. Ehrenhaft, **Wiener Berichte,** vol. 119, 1910. p. 815.

21. F. Ehrenhaft, **Physik. Ztschr.,** vol. 11, 1910. p. 619.

22. F. Ehrenhaft, **Physik. Ztschr.,** vol. 39, 1938. p. 673.

23. F. Ehrenhaft, **Ann. der Physik.,** vol. 56, 1918. p. 81.

24. F. Ehrenhaft, **Comp. Rend.,** vol. 190, 1930. p. 263.

25. F. Ehrenhaft, **Ann. der Physik.,** vol. 13, 1940. p. 151.

26. F. Ehrenhaft, **Jour. Frank. Inst.,** vol. 233, 1942. p. 235.

27. F. Ehrenhaft, **Nature,** vol. 154, 1944. p. 426.

28. F. Ehrenhaft, **Phys. Rev.,** vol. 65, 1944. p. 287.

29. F. Ehrenhaft, **Bull. Am. Phys. Soc.** (New York Meeting), vol. 6, 1945.

30. H.S. Renue, **Radio Electronic Engineering (Radio News),** vol. 4, 1945. p. 22.

31. Woldemar Voigt, **Festschrift fuer Heinrich Weber,** 1912.

32. P.A.M. Dirac, "Development of the physicist's conception of nature," **Symposium on the Development of the Physicist's Conception of Nature,** ed. Jagdish Merha, D. Reidel Pub. Co., Boston, 1973. p. 12-14.

33. Olof Alexandersson, **Living Water: Viktor Schauberger**

and the Secrets of Natural Energy, translated by Kit and Charles Zweigbergk, Turnstone Press Limited, Willingborough, Northamptonshire, England, 1982. An excellent bibliography is included in the book.

34. O. Alexandersson, "Implosion contra explosion," **Seklet,** vol. 4, 1965.

References Related to Louis Kervran's Work

1. Publications by Louis Kervran:

- a. **Biological Transmutations.** Crosby Lockwood, London, 1972.
- b. **A la Decouverte des Transmutations Biologiques, une Explication des Phenomenes Biologiques Aberrants.** Le Courrier du Livre, Paris, 1966.
- c. **Preuves Relatives a L'existence de Transmutations Biologiques, Echecs en Biologie a la loi de Lavoisier d'invariance de la Matiere,** Maloine, Paris, 1968.
- d. **Transmutations Biologiques; Metabolismes Aberrants de L'azote, le Potassium et el Magnesium,** Librairie Maloine, Paris, 1962.
- e. "Les Transmutations Biologiques en Agronomie," Maloine, Paris, 1970.
- f. **Biological Transmutations,** Swan House Publishing Co., Binghamton, New York, 1972.
- g. "Alchimie d'hier et d'aujourd'hui," **L'Alchimie, Reve ou Realite,** Revue des Ingenieurs do L'Institut National Superieur de Rouen, 1972-73.
- h. **Transmutations biologiques et physique moderne.** Paris: Maloine S.A., 1982.
- i. **Transmutations a faible energie (naturelles et biologiques).** Paris: Librairie Maloine, 1972.

   j. **Transmutations biologiques.** Rue de l'Ecole-de-Medecine, Paris: Librairie Maloine S.A., 1963.

   k. **Transmutations naturelles.** Librairie Maloine S.A., Rue de l'Ecole-de-Medecine, Paris Vie, 1963.

2. Rudolf Hauschka, **The Nature of Substance,** Vincent Stuart Ltd., London, 1966.

3. V.B. Neiman. ed., **Problems of Transmutations in Nature: Concentration and Dissipation.** (Collection of papers in Russian), Aiastan Pub. House, Erevan, Armenia, USSR, 1971.

4. Peter Tompkins and Christopher Bird, **The Secret Life of Plants,** Harper & Row, New York, 1973.

5. The Kaznacheyev work and references. See appropriate sections.

6. Dr. Fritz Albert Popp's work and references. See appropriate sections.

7. The Australian work of Reid and Barsamian. See appropriate sections.

8. Fritz Albert Popp et al, eds. **Electromagnetic Bio-Information.** Proceedings of the Symposium, Marburg, September 5, 1977. Urban & Schwarzenberg, Baltimore, 1979.

9. Thomas E. Bearden, **The Excalibur Briefing,** Second edition, Strawberry Hill Press, San Francisco, 1988.

10. N.A. Kozyrev, "Possibility of the Experimental Study of the Properties of Time," Sep. 1967. English translation JPRS 45328, May 1968, p. 1-49.

11. Captain John D. LaMothe and Mr. Louis Maire, **Soviet and** Czechoslovakian **Parapsychology Research,** DST-1810S-387-75, Defense Intelligency Agency, Washington, D.C., 1975. Released under the Freedom of Information Act.

12. Jeff Hecht, "IBM Reports on Dark-Pulse Solitons," **Lasers & Optronics,** 7(3), Mar. 1988, p. 42 -44.

References Related to Popp's Master Cellular Communication System

1. Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979.

2. V.P. Kaznacheyev et al, "On distant intercellular interactions in a system of two tissue cultures connected by optical contact," **Transactions of the Moscow Society of Naturalists,** Vol. XXXIX, 1972. p. 224-227.

3. F.A. Popp, **Biophotonen. Ein neuer Weg zur Losung des Krebsproblems.** Verlag fur Medizin, Heidelberg, 1976.

4. F.A. Popp et al, "Biosignale zur Steuerung des Zellstoffwechsels. Eine Resonanzhypothese der Karzinogenese (vorl. Mitt.)" **Munch, med. Wschr,** vol. 116, 1974. p. 381-384.

5. F.A. Popp, "Molecular aspects of carcinogenesis, in Deutsch et al, eds.: **Molecular Base of Malignancy**, Stuttgart, 1976.

6. A. Gurwitsch, "Die mitogenetische Strahlung der Optischen Bahn bei adaquater Erreguns." **Pfluger Arch. ges. Physiol.**, vol. 231, 1932. p. 254-264.

7. A. Gurwitsch, "Die mitogenetische Strahlung des markhaltigen Nerven," **Pflugers Arch. ges. Physiol.,** vol. 231,1932. p. 234-237.

8. A.A. Gurwitsch, "The problem of mitogenetic radiation as an aspect of molecular biology." (in Russian), Leningrad, 1968.

9. A.A. Gurwitsch, V.F. Eremeiev, and Yu. A. Karabchievsky, "Energy bases of the mitogenetic radiation and its registration on photoelectron multipiers," **Medicina,** Moskva, 1974.

10. J.W. Hastings, "Bioluminescence," **Ann. Rev. Biochem.,** vol. 37, 1968. p. 597-630.

11. L.W. Latmanisowa, "Die mitogenetische Sekundar-strahlung des Nerven," **Pflugers Arch. ges. physiol.,** vol. 231,1932. p. 265-279.

12. H. Ninnemann et al, "Inhibition of respiration in yeast by light," **Biochim. Biophys. Acta,** vol. 205, 1970. p. 499-506.

13. T.I. Quickenden and S.S. Que Hee, "Weak luminescence from the yeast saccharomyces cerevisiae and the existence of mitogenetic radiation," **Biochem. Biophys. Res. Comm.,** vol. 60, 1974. p. 764-770.

14. A. Babloyantz, "Self-organization phenomena resulting from cell-cell contact," **J. Theor. Biol.,** vol. 68, 1977. p. 551-561.

15. R.M. Benolken, "Reversal of photoreceptor polarity recorded during the graded receptor potential response to light in the eye of limulus," **Biophys. J.,** vol. 1961. p. 551-564.

16. H. Breithaupt, "Biological rhythms and communication," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore 1979. p. 1-23.

17. Herbert L. Konig, "Bioinformation - Electrophysical Aspects," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 25-55.

18. Ulrich Warnke, "Information transmission by means of electrical biofields," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds.., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p. 55-79.

19. Walter Peschka, "On kinetobaric effects and bioinformational transfer by electromagnetic fields," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 81-94.

20. Gunther Becker, "Communication between termites by means of biofields and the influence of magnetic and electric fields on termites," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-In-**

formation: Proceedings of the Symposium, Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p. 95-106.

21. Bernhard Ruth, Experimental investigations on ultraweak photon emission," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p.107-122.

22. Fritz Albert Popp, "Photon storage in biological systems," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 123-149. A most important paper, presenting what I have been calling the master cellular communication system. Ruth has confirmed the existence of the "degradation radiation," ("death photons"). Experimental results of Ruth and others have led Popp to postulate that biological systems generally have the capacity to store coherent photons which come from the external world, and that ultraweak photon emission from biological systems is governed by photon storage ("charge up") within the cell population.

23. S.S. Sung, "A possible biophotochemical mechanism for cell communication," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p. 151-174.

24. Helmut A. Fischer, "Photons as transmitter for intra- and inter-cellular biological and biochemical communication — the construction of a hypothesis," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 174-180.

25. Walter Kroy, "The use of optical radiation for stimulation therapy," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-In-**

formation: Proceedings of the Symposium, Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p. 181-193.

26. Ulrich Warnke and Fritz Albert Popp, "Some aspects of magnetic influences on biological systems," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 195-199.

27. A.S. Presman, **Electromagnetic Fields and Life,** Plenum Press, New York, 1964.

28. F.A. Popp and B. Ruth, "Untersuchungen zur ultraschwachen Lumineszenz aus biologischen Systemen unter Berrucksichtig ung der Bedeutung fur die Arzneimittelforchung," **Arzneimittel-Forsch,** vol. 27 (I), 1977. p. 933.

29. B.Ruth and F.A. Popp, "Experimentelle Untersuchungen zur ultraschwachen Photonenemission biologischer Systeme," Z. **Naturforsch,** vol. 31c, 1976. **p.** 741.

References Related to Kaznacheyev's Experiments

1. V.P. Kaznacheyev et al, "Apparent information transfer between two groups of cells," **Psychoenergetic Systems,** 1(1), Dec. 1974.

2. V.P. Kazacheyev et al, "Distant intercellular interactions in a system of two tissue cultures," **Psychoenergetic Systems,** 1(3), Mar. 1976.

3. Captain John D. LaMothe, **Controlled offensive behavior — USSR,** Report ST-CS-01-169-72, Defense Intelligence Agency, Washington, D.C. (Released under FOIA).

4. Captain John D. LaMothe and Mr. Louis Maire, **Soviet and Czechoslovakian Parapsychology Research,** Report. DST-1810S-387-75, Defense Intelligence Agency, Washington, D.C. (Released under FOIA).

5. Trevor James Constable, **The Cosmic Pulse of Life,** Merlin

Press, Santa Ana, California, 1976.

6. John White and Stanley Krippner, eds., **Future Science,** Anchor, 1977.

7. B. Ruth, "Experimental Investigations on Ultraweak Photon Emission," in **Electromagnetic Bioinformation, Proceedings of the Symposium,** Marburg, September 5, 1977 (Munchen-Wien-Baltimore, 1979). p. 107-122

8. H. Fischer, "Photons as Transmitters for Intra- and Intercellular Biological and Biochemical Communication — the Construction of a Hypothesis," in **Electromagnetic Bioinformation, Proceedings of the Symposium,** Marburg, September 5,1977 (Munchen-Wien-Baltimore, 1979). p. 175-180.

9. F.A. Popp, "Photon Storage in biological systems," in **Electromagnetic Bioinformation, Proceedings of the Symposium,** Marburg, September 5, 1977 (Munchen-Wien-Baltimore, 1979). p. 123-149.

10. A.G. Gurvitsch, **Theory of Biological Field,** Moscow, 1944 (in Russian).

11. A.G. Gurvitsch,"A Concept of 'the Whole' in the Frame of Biological Field Theory," in **Works on Mitogenesis and Theory of Biological Field,** Moscow, 1974, p. 141-147 (in Russian).

12. E.S. Bauer, "General Theory of Living Substance. The Problem of'Living Protein'," **Archives Biol. Sci.,** Vol. 35, Ser. A, Issue 1, 1934, p. 1-36 (in Russian).

13. E.S. Bauer, "Contradictions between the Inner and Outer Work of Living Systems," **Archives Biol. Sci.,** Vol. 35, Ser. A, Issue 1, 1934, p. 37-52 (in Russian).

14. E.S. Bauer, **Theoretical Biology,** Moscow, 1935 (in Russian).

15. S.V. Konev, "On the Question of Nature and Biological Significance of Ultraweak Luminescence of Cells," in **Bioluminescence,** Moscow, 1965, p. 181-183 (in Russian).

16. B.N. Tarusov, A.I. Pllivoda, and A.I. Zhuravlev, "Study of Ultraweak Spontaneous Luminescence of Animal Cells," **Biofzika,** Vol. 6, No. 4, 1961. p. 490-492 (in Russian).

17. M. Rottemeyer, "Modelle zur Interpretation der Ultraschwachen Photone-emission aus Biologischen Systemen," **Diplomarbeit** (Marburg, 1978), (in German).

18. Yu. A. Vladimirov, "Ultraweak Luminescence of Subcellular Structures," **Ultraweak Luminescence in Biology, Proceedings of the Symposium,** Moscow, 1969 (in Russian).

19. I.S. Marchenko, **Biofield of Forest Ecological Systems,** Bryansk, 1973 (in Russian).

20. S.S. Sung, "A Possible Biophotochemical Mechanism for Cell Communication," in **Electromagnetic Bioinformation, Proceedings of the Symposium,**, Marburg, September 5, 1977 (Munchen-Wien-Baltimore, 1971). p. 151-174.

21. P.A.Popp and B. Ruth, "Untersuchungen zur Ultraschwachen Luminiszenz aus Biologischen Systemen uter Berucksichtig ung der Bedeutung fur die Arznei-mittelforschung," **Arzneimit. Forsch.** (Drug Research), Vol. 27, No. 5,1977 (in German).

22. V.P. Kaznacheyev, "Informational Function of Ultraweak Flows of Light in Biological Systems," in **Materials of Scientific Conference: Physical and Mathematical Methods of Research in Biology and Medicine,** Novosibirsk, 1965, pp. 38-41 (in Russian).

23. V.P. Kaznacheyev, N.V. Ignatovich, and S.P. Shurin, "Intercellular Distant Interactions in a System of Two Tissue Cultures Connected by Optical Contact," in **Ultraweak Luminescence in Biology,** Moscow, 1969 (in Russian).

24. V.P. Kaznacheyev, L.P. Mikhailova, and S.B. Stephanov, "Studies on Cellular Monolayer and Cellular Distant Interaction in High Latitude Conditions," in **Technical-Scientific Progress and Circumpolar Medicine,** Novosibirsk, 1978, p. 133-134 (in Russian).

25. V.P. Kaznacheyev, L.P. Mikhailova, V.N. Sudarev, and S.P

Shurin, "Studies on the Biological Role of Electromagnetic Emission as a Factor of Adaptive Behavior of Cells under Conditions of Latitude Changes," in **Abstracts of the Second Ail-Union Conference Devoted to the Problem of** Human **Adaptation to Different Geographical, Climatic, and Labor conditions,** Novosibirsk, 1977, p. 101-104 (in Russian).

26. V.P. Kaznacheyev, L.P. Mikhailova, V.N. Sudarev, and S.P. Shurin, "Distant Intercellular Interactions Caused by UV-radiation," in **Photobiology of a Living Cell,** Leningrad, 1979, p. 221-223 (in Russian).

27. V.P. Kaznacheyev, L.P. Mikhailova, D.G. Kadayeva, and M.P. Dranova, "Conditions Necessary for Appearance of distant Intercellular Interactions after UV-radiation," **Bulleten Experimentalnoy Biologii i Meditsiny.,** No. 5. 1979, p. 468-471 (in Russian).

28. V.P. Kaznacheyev, L.P. Mikhailova, and S.P. Shurin, "Distant Intercellular Interactions in a System of Two Tissue Cultures Connected by Optical Contact," in **Regulation of Biosynthesis and Biophysics of Populations,** Krasnoyarsk, 1969, p. 372-374 (in Russian).

29. V.P. Kaznacheyev, S.P. Shurin, and L.P. Mikhailova, Discovery No. 122, "Distant Intercellular Interactions in a System of Two Tissue Cultures," **Official Bulletin of the Committee on Inventions and Discoveries Affiliated to the Council of Ministers of the USSR,** No. 19, 1973, p. 3.

30. V.P. Kaznacheyev, S.P. Shurin, L.P. Mikhailova, and N.V. Ignatovich, "Distant Intercellular Interactions in a System of Two Tissue Cultures Connected by Optical Contact," in **Ultraweak Luminescence in Biology,** Moscow, 1972, p. 224-227 (in Russian).

31. V.M. Inyushin, "The Concept of Biological Plasma and Some Questions of Photoenergetics," in **Questions of Bioenergetics,** Alma-Alta, 1968 (in Russian).

32. E. Makovsky, **Nature and Structure of Living Matter,** Bucharest, 1976 (in Russian).

33. A.L. Chizhevsky and Yu. Shishina, **In the Rhythm of the Sun,** Moscow, 1969.

34. Stefan T. Passony, "Psy-War: Soviet device experiment," **Defense and Foriegn Affairs Daily,** 12 (104), June 7,1983, p. 1-2.

35. V.I. Vernadsky, **Collected Works,** Moscow, 1960, (in Russian).

36. A.L. Chizhesvsky, **The Earth's Echo of the Sun Storms,** Moscow, 1973.

37. A.S. Pressman, **Electromagnetic Fields and Living Nature,** Moscow, 1968, English translation: **Electromagnetic Fields and Life,** Plenum, New York, 1970.

38. H.Z. Konig, "Bioinformation. Electrophysical Aspects," in **Electromagnetic Bioinformation, Proceedings of the Symposium,** Marburg, September 5,1977 (Munchen-Wien-Baltimore, 1979). p. 25-54.

39. V.P. Kaznacheyev, "Information Function of Ultraweak Light Flows in Biological Systems," in **Problems in Biophysics,** Novosibirsk, 1967, p. 7-18 (in Russian).

40. V.P. Kaznacheyev and L.P. Mikhailova, "Ultraweak Radiations in Intercellular Interactions," Novosibirsk, 1981 (in Russian); English Translation: San Francisco: Washington Research Center, 1982.

41. John F. Gilbey, "The delayed death touch," **Secret Fighting Arts of the World,** Charles E. Tuttle Co., Rutland, Vermont, 1963. p. 13-22.

42. Edward W. Russell, **Report on Radionics: Science of the Future,** Neville Spearman, London, 1973.

43. Paul Brodeur, **The Zapping of America,** W.W. Norton & Co., New York, 1977.

44. Waldemar Kaempffert," 'Dead' organisms revived," Science in the News: **Science Digest,** May 1950, p. 77-78.

45 "Rejuvenating light," **Scientific American,** May 1949.

46. Death-rays and life-rays discovered," **Science Digest,** Feb. 1950, p. 56. A "death" frequency seems to be at 2,537 angstroms, while a "life" frequency seems to be at 3,659 angstroms.

47. Lewis R. Roller, **Ultraviolet Radiation,** 2nd edn., John Wiley & Sons, New York, 1965. p. 236-237.

48. T.E. Bearden, Appendix: "Theory of Pulsor(R) operation," and "A partial glossary for scalar electromagnetics and subtle phenomena," in George T.F. Yao, **Pulsor(r): Miracle of Microcrystals,** Gyro Industries, Newport Beach, California, 1986.

49. The Antoine Priore material.

50. The Louis Kervran material.

51. Thomas E. Bearden, **The Excalibur Briefing,** Strawberry Hill Press, San Francisco, 2nd Edition, 1988.

52. Martin Ebon, "Moscow, June 11, 1977," in his **Psychic Warfare: Threat or Illusion,** McGraw-Hill Book Co., New York, 1983, p. 1-11.

References Related to the New View of the Nature of Mind and Thought

Author's Note: So far as I can find, there are no available references other than my own that deal precisely with my particular approach to the nature of mind and thought. There is, however, a serious debate over whether (a) the mind is simply the physical brain, similar to a computer (the prevailing materialist view of orthodox science), or (b) the mind is something else altogether, and metaphysical — at least with respect to the level of scientific knowledge today. My own approach utilizes a non-Aristotlean fourth law of logic (included in an appendix to this book) by means of which the dichotomy between mind and matter is removed. In that fashion, one can model the mind as material, or model matter us mind.

Some references that have influenced my thinking are given **below.**

Selected Bibliography

1. Publications by Thomas E. Bearden:

   a. "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," **Pursuit,** 13(1), Winter 1980. p. 6-10. (Appendix III to this present book.).

   b. "A mind/brain/matter model consistent with quantum physics and UFO phenomena," **Mutual UFO Network Annual UFO Symposium Proceedings,** 1979. Also available through the National Technical Information System, Port Royal Road, Springfield, Virginia 22161. Report AD-A068988.

   c. "An approach to understanding psychotronics." Paper delivered to 1977 Annual Symposium of the U.S. Psychotronics Association. Oglethorpe University, Atlanta, Georgia. 1977. Available through the National Technical Information System (NTIS), Port Royal Road, Springfield, Virginia 22161. Report AD-A027866. (This report is mostly rough notes).

   d. **Excalibur Briefing: Understanding Paranormal Phenomena.** Second edition. Foreword by John White. Strawberry Hill Press, San Francisco, 1988.

   e. "Hyperspace (virtual state) engineering." presented at the 11th MUFON UFO Symposium, Clear Lake City, Texas, June 6-8. 1980. Published in **UFO technology: a detailed examination: 1980 MUFON UFO Symposium Proceedings.** Edited by Walter H. Andrus, Jr. and Dennis W. Stacy. Mutual UFO Network, Seguin, Texas, 1980.

   f. "Photon quenching of the paranormal (time) channel: **a** brief note." Paper presented in absentia at the Third International Psychotronics Congress, Tokyo, 1977. Available through the NTIS, Port Royal Road, Springfield, Virginia 22161. Report AD-A038588.

   g. **Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena." NTIS, 1973. Report AD-763210.**

h. "The Boundary Identity of Exact Opposites: A simple Solution to the Age-old Philosophical Problem of Change." NTIS, Oct. 1975. Report AD-020798.

i. "The fourth law of logic." NTIS, 1979. Report AD-A068987.

j. **Field, Formon, Superspace, and Inceptive Cyborg: A Paraphysical Theory of Noncausal Phenomena.** U.S. Army Medical Intelligence and Information Agency, Report MIIA-1-74, Dec. 1974. Available from the Defense Documentation Center (DDC), Report AD A-005579/8GI.

k. "The One Human Problem, Its Solution, and Its relation to UFO Phenomena." **Pursuit,** January 1976. Available from NTIS, Report AD-A034236.

1. "Writing the Observer Back Into the Equation," Address given to the Princeton Center for Alternative Futures, Princeton, New Jersey, Mar. 5,1976, published in June, 1976. Available from NTIS, Report AD-A027867.

m. "Species Metapsychology, UFO Waves, and Cattle Mutilations, May 23, 1977. Privately published.

n. "The Holography of Being," 1975. Unpublished,

o. **Yoseikan Aikido,** 1973. Privately published.

p. "AIDS: Urgent comments on mankind's greatest threat; and the secrets of electromagnetic healing," presented in absentia at the annual USPA Symposium in Golden, Colorado 30 July-2 Aug. 1987.

2. Mike Costello, "Without a Brain," letter, Fortean Times, 96 Mansfield Road, London NW 8 2HX, England, Spring 1987, p. 78. Information on hydrancephaly, persons with normal intelligence but with large cavities in the area normally occupied by cortical tissue. Most of the data on these anomalous cases is buried in obscure medical journals. (Hydrocephaly — an excess of fluid in the cerebral/spinal system — is much more widely known.). See also Anthony Smith, **The Mind,** Hodder & Stoughton, 1984 for additional information on hydrancephaly.

Quoting from Costello: "Actually cases of this kind have been turning up at autopsies for years, and reported in the appropriate medical journals, but have been disregarded, perhaps because they are so destructive of established beliefs about the relationship between consciousness and the human brain."

Costello goes on to point out that a group at Sheffield University has collected a great deal of data on this subject, from case studies of 253 hydrocephalics and hydrancephalics. There seems to be no correlation between the IQ of these individuals and the amount of residual brain tissue.

One living subject, Roger, has an IQ of 126 and a First Class degree in Mathematics, yet has no more than 5% of the normal amount of brain tissue.

A Swedish documentary on the subject was shown on ITV in 1982.

Fortean Times 38, **p.** 30 also contains an item on surviving adults with no detectable brains.

3. Hugh Everett, **III,** The **Many-Worlds Interpretation of Quantum Mechanics: A Fundamental Exposition,** with papers by J.A. Wheeler, B.S. DeWitt and Neill Graham, Princeton Series in Physics, Princeton University Press, 1973.

4. Louise and Galen Hieronymus, **Tracking the Astronauts in Apollo 11, With Data From Apollo 8 Included,"** Advanced Sciences Research and Development Corporation, Fort Lauderdale, Florida, 1969.

5. Vlail Kaznacheyev's work and publications related to it (see Kaznacheyev references).

6. Antoine Priore's work and publications related to it (see Priore references).

7. Charles Muses and Arthur M. Young, eds, **Consciousness and Reality.** Outerbridge & Lazard, 1972.

8. George Meek, **Healers and the Healing Process,** Theosophical Publishing House, 1977.

9. Harold Puthoff and Russell Targ, **Mindreach,** Delacorte, 1977.

10. Harold Puthoff and Russell Targ, "A Perceptual Channel for Information Transfer Over Kilometer Distances: Historical Perspective and Recent Research," **Proceedings of the IEEE,** 64(3), March 1976.

11. Edward W. Russell, **Report on Radionics,** Neville Spearman, London, 1973.

12. John White and Stanley Krippner, eds., **Future Science,** Anchor, 1977.

13. C. Louis Wiedemann, "Results ofthe N.J. 'Spook Light' Study," **Vestigia Newsletter,** Stanhope, New Jersey, May 1977.

14. David Bohm, **Wholeness and the Implicate Order,** Routledge, Kegan Paul, 1980.

15. Renee Weber, "Reflections on David Bohm's Holomovement: A Physicist's Model of Cosmos and Consciousness," in Valle and von Eckartsberg, editors, **The Metaphors of Consciousness,** Plenum Press, 1981.

16. "The Physicist and the Mystic: Is a Dialogue Between Them Possible?" A Conversation with David Bohm, conducted by Renee Weber, edited by Emily Sellon, **ReVision,** 4(1), Spring 1981, p. 22-35.

17. The Tao of Physics Revisited," A Conversation with Fritjof Capra, conducted by Renee Weber, **Revision,** 4(1), Spring 1981, p. 36-52.

18. "Reflections on the New Age Paradigm," an Interview with Ken Wilber, **ReVision,** 4(1), Spring 1981, p. 36-52.

19. Fritjof Capra, **The Tao** of **Physics,** Shambhala, Boulder, Colorado, 1975.

20. Fritjof Capra, **The Turning Point,** Simon & Schuster, New York, 1982.

Ken Wilbur, **Up From Eden,** Doubleday, 1981.

22. G. Spencer, **Laws of Form,** Julian Press, New York, 1972.

23. Bernard D'Espagnat, **Conceptual Foundations of Quantum Mechanics,** W.A. Benjamin, Menlo Park, California, 1971.

24. Robert Linssen, **Zen: The Art of Life,** Pyramid, New York, 1969.

25. E.P. Wigner, "Remarks on the Mind-Body Question," **The Scientist Speculates, I.J.** Good ed., W. Heinemann Ltd., London,1961.

26. Thomas G. Hieronymus, U.S. Patent No. 2,482,773, Sept. 27, 1949.

27. Langston Day and George De La Warr, **New Worlds Beyond the Atom,** Vincent Stuart Publishers Ltd., London, 1973.

28. Jule Eisenbud, M.D., **The World of Ted Serios,** Pocket Books, Simon and Schuster, New York, 1968.

29. Sheila Ostrander and Lynn Schroeder, **Psychic Discoveries Behind the Iron Curtain,** Prentice-Hall, 1970.

30. Carl G. Jung, **Man and His Symbols,** ed. by Carl G. Jung **and** after his death by M.S. von Franz, Aldus Books Ltd., London. 1964.

31. **The Unfathomed Mind: A Handbook of Unusual Mental Phenomena,** compiled by William R. Corliss, The Sourcebook Project, Glen Arm, Maryland, 1982.

32. Charles Panati, ed., **The Geller Papers: Scientific observations on the paranormal powers of Uri Geller**, Houghton Mifflin Company, Boston, Massachusetts, 1976.

33. Ken Wilbur, **The Spectrum of Consciousness,** The **Theo**sophical Publishing House, Wheaton, Illinois, 1977.

34. "Godel's proof," **Scientific American,** June 1965, p. 81-86
As Ken Wilbur puts it, "Thus logically as well as **physically** verification is not an absolute mark of reality. If all **is** to be verified, how do you verify the verifier, since he is part of **the** all?"

35. Erwin Schroedinger, **What is Life? and Mind and Matter,** Cambridge University Press, London, 1969.

36. Teilhard de Chardin, **The Phenomenon of Man,** Harper Torchbooks, New York, 1965.

37. D.T. Suzuki, **The Zen Doctrine of No Mind,** Rider, London, 1970.

38. Frederick S. Perls, **Gestalt Therapy Verbatim,** Real People Press, Lafayette, Indiana, 1969.

39. Frederick S. Perls, Ralf F. Hefferline, and Paul Goodman, **Gestalt therapy,** Dell, New York, 1951.

40. John White, ed., **The Highest State of Consciousness,** Anchor, New York, 1972.

41 J. Blofeld, trans., **The Zen Teaching of Huang Po,** Grove Press, New York, 1958.

42. R.E. Ornstein, **The Psychology of Consciousness,** W.H. Freeman and Company, San Francisco, California, 1972.

43. Joseph C. Pearce, **The Crack in the Cosmic Egg,** Julian, New York,

44. Andrija Puharich, **Beyond Telepathy,** Anchor Press, Doubleday & Company, Garden City, New York, 1973.

45. Jacques Vallee, **Passport to Magonia: From Folklore to Flying Saucers,** Henry Regnery Company, Chicago, Illinois, 1969.

46. Robert Bruce Lindsay and Henry Margenau, **Foundations of Physics,** Dover publications, New York, 1957. See p. 2-3 for the blunt statement that physics has nothing at all to say about a possible real world lying behind its concepts and equations. That is, physics is simply a model, and the assumption of a kind of reality is an interpretation and an hypothesis.

47. Ted Bastin, ed., **Quantum Theory and Beyond,** Cambridge University Press, 1971.

48. Edgar D. Mitchell, **Psychic Exploration: A Challenge for Science,** edited by John White, G.P. Putnam's Sons, New York, 1974.

49. Hector Hawton, **Philosophy for Pleasure,** Fawcett Publications, 1956, Fifth printing, June 1970. (A real gem.)

50. Jiro Kikkawa and Malcom J. Thorne, **The Behavior of Animals,** Plume, The New American Library Inc., New York, 1974.

51. Helena Curtis, **Biology,** Worth Publishers, Inc., New York, 1968 (fourth printing); p. 614-637 contains a good discussion of instinct, learning, and memory.

52. John Cunningham Lilly, **The Mind of the Dolphin: A Nonhuman Intelligence,** Avon, New York, Mar. 1969.

53. Isaac Asimov, **The Human Brain: Its Capacities and Functions,** Signet Science Library, The New American Library, Inc., New York, 1963.

54. R.W. Sperry, "The Great Cerebral Commissure," **Scientific American,** Jan. 1964, p. 42-52.

55. Michael S. Gazzaniga, "The Split Brain in Man," **Scientific American,** Aug. 1967, p. 24-29.

56. Roger Sperry, "Problems Outstanding in the Evolution of Brain Function," James Arthur Lecture, American Museum of Natural History, New York, 1964.

57. Jose M.R. Delgado, **Physical Control of the Mind: Toward a Psycho-Civilized Society,** Harper Colophon Books, Harper & Row, New York, 1971.

58. B.B. Kazhinskiy, **Biological Radio Communications,** Izdatel'stvo Akademii nauk Ukrainskey SSR, Kiev. 1962. Trans, available through DDC, AD 415676.

59. S.K Lisitsyn "New Approach to the Analysis of Electroencephalograms," in **Problems of Bionics (Selected Articles), p.** 16-25, DDC Report **AD** 730045.

60. Evert W. Beth, **The Foundations** of Mathematics, Harper Torchbooks, Harper & Row, New York, 1966. p. 481-518 contains a discussion of the "contradictions" of logic painfully noted by mathematicians shortly after the turn of the century.

61. Stanley Krippner and James Hickman, "West Meets East: A Parapsychological Detente," **Psychic,** May/June 1974, p. 51-55.

62. G.G. Globus, ed., **Consciousness and the Brain,** Plenum, New York, 1976.

63. Karl Pribram, **Emotion: the Search for Control,** McGraw Hill, New York, 1968.

64. Karl Pribram, **Languages of the Brain,** Prentice Hall, Englewood Cliffs, 1971.

65. Charles Hampden-Turner, **Maps of the Mind: Charts and Concepts of the Mind and Its Labyrinths,** Collier Books, Macmillan Publishing Co., New York, 1981.

66. Adam Smith, **Powers of the Mind,** Random House, New York, 1975.

67. Peter Tompkins and Christopher Bird, **The Secret Life of Plants,** Harper & Row, New York, 1973. (a real gem.)

68. The Louis Kervran material.

69. The Antoine Priore material.

70. Langston Day and George De La Warr, **Matter in the Making,** Vincent Stuart, London, 1966.

71. Sister Justa Smith, "Paranormal Effects on Enzyme Activity," Professional Paper No. 2, **Human Dimensions,** 1(2), 1972.

72. John Pierrakos, **The Energy Field in Man and Nature,** Institute of Bioenergetic Analysis, New York, 1971.

73. Jeffrey Mishlove, **The Roots of Consciousness,** Random House, New York, 1975.

74. Andrija Puharich, "Psychic Research and the Healing Process," in **Psychic Exploration,** ed. John White, Putnam, New York, 1975.

75. Andrija Puharich, **Beyond Telepathy,** Darton, Longman and Todd, London, 1962.

76. Lyall Watson, **Supernature,** Hodder & Stoughton, London, 1973.

77. Jule Eisenbud, "The Mind-Matter Interface," **Journal of American Society for Psychical Research,** 69(2), Apr., 1975.

78. George W. Meek, **From Enigma to Science,** Samuel Weiser, New York, 1973.

References Related to Antoine Priore's Work

1. Antoine Priore. "Procedeet dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes." Republique Francais: Brevet d'Invention P.V. No. 899.414, No.1.342.772,1963. For the U.S. patent, see Antoine Priore, "Method of producing radiations for penetrating living cells," U.S. patent No. 3,280, 816. Oct. 25, 1966.

2. Letter, R. Pautrizel to his colleagues at the University of Bordeaux, Sept. 11, 1979.

3. Letter, R. Pautrizel to his colleagues, Mar. 26, 1969.

4. Jacques Sylvain and Jean-Michel Graille, "Depuis quinze ans le "rayon Priore" divise le monde scientifique." [For fifteen years the "Priore Ray" has divided the world of science.]. **Sud-Ouest Dimanche,** Jan. 27, 1980.

5. "Affaire Priore: L'armee va trancher? [The Priore Affair: Will the Army resolve it?] **Sud-Ouest Dimanche,** May 20,1980.

6. "Affaire Priore: L'armee va tranchera." [The Priore Affair: The Army will resolve it.]. **Sud-Ouest Dimanche,** Apr. 12, 1981.

7. M.R. Riviere, A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Action de champs electromagnetiques sur les greffes de la tumeur T8 chez le Rat." [Action of the electromagnetic fields on the graftings of the TB tumor on the rat.]. **Compt. Rend. Acad. Sci.** (Paris), 1964, **259,** 4895-4897.

8. M.R. Riviere, A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Effets de champs electromagnetiques sur un lymphosarcome lymphoblastique transplantable du Rat." [Effects of Electromagnetic fields on lympheuplastic lymphosarcoma transplantable from a rat.]. **Compt. Rend. Acad. Sci.** (Paris), 1965, **260,** 2099-2102.

9. M.R. Riviere, A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Phenomenes de regression observes sur les greffes d'un lymphosarcome chez des souris exposees a des champs electromagnetiques." **Compt. Rend. Acad. Sci.** (Paris), 1965, 260, 2639-2643.

10. G. Delmon and J. Biraben,"la croissance du carcinome de Guerin sous Faction de champs electromagnetiques." **Rev. Path,** comp, 1966, 3, no. 2, 85-88.

11. M.R. Riviere and M. Guerin, "Nouvelles recherches effectuees chez des rats porteurs d'un lymphosarcome lymphoblastique soumis a l'action d'ondes electromagnetiques associees a des champs magnetiques." **Compt. Rend. Acad. Sci.** (Paris), 1966, **262,** 2669-2672.

12. R. Pautrizel, M.R. Riviere, A. Priore, and F. Berlureau, "Influence d'ondes electromagnetiques et de champs manetiques associes sur l'immunite de la souris infestee par **Trypanosoma equiperdum."** [Influence of electromagnetic waves and associated magnetic fields on the immunity of the mouse infected with the Trypanosoma equiperdum.]. **Compt. Rend. Acad. Sci.** (Paris), 1966, **263;** 579-582.

13. R. Pautrizel, A. Priore, and F. Berlureau, and A.N. Pautrizel, "Stimulation par des moyens physiques des defenses de la Souris et du Rat contre la trypanosomiase experimentale." [Stimulation, by physical means, of defenses of the mouse and of the rat against the experimental Trypanosoma.]. **Compt. Rend. Acad. Sci.** (Paris), 1969, **268,** 1889-1892.

14. R. Pautrizel, A. Priore, F. Berlureau, and A.N. Pautrizel, "Action de champs magnetiques combines a des ondes electromagnetiques sur la trypanosomiase experimentale du Lapin" [Action of magnetic waves combined with electromagnetic waves on the experimental trypanosoma on the rabbit.]. **Compt. Rend. Acad. Sci.** (Paris), 1970, **271,** 877-880.

15. A.J. Berteaud, A.M. Bottreau, A. Priore, A.N. Pautrizel, F. Berlureau, and R. Pautrizel, "Essai de correlation entre revolution d'une affection par **Trypanosoma equiperdum** et l'action d'une onde electromagnetique pulsee et modulee." [Trial of the correlation between the evolution of a disease by Trypanosoma equiperdum and the action of a pulsating and modulated electromagnetic wave.]. **Compt. Rend.** Acad. **Sci.** (Paris), 1971, **272,** 1003-1006.

16. R. Pautrizel, P. Mattern, A. Priore, A.N. Pautrizel, and D. Bernard, "Etat de protection vis-a-vis de **Trypanosoma equiperdum** chex des souris splenectomisees et soumises a une stimulation physique." **ler Multicolloque Europeen** de **Parasitologie,** Rennes, 1 au 4 Septembre 1971.

17. R. Pautrizel, A. Priore, M. Dallochio and R. Crockett, "Action d'ondes electromagetiques et de champs magnetiques sur les modifications lipidiques chez le Lapin par l'administration d'un regime alimentaire hypercholesterole." [Action of electromagnetic waves and magnetic fields on provoked lipidic modiications in the rabbit by the administration of an alimentary hyperchloresterol diet.]. **Compt. Rend. Acad. Sci.** (Paris), 1972, **274,** 488-491.

18. G. Mayer, A. Priore, G. Mayer and R. Pautrizel, "Action de champs magnetiques associes a des ondes electromagnetiquea sur 1' orchite trypanosomienne du lapin." [Action of the magnetic fields associated with electromagnetic waves on the trypanosomian orchitis of the rabbit.]. **Compt. Rend.** Acad. **Sci.** (Paris), 1972, **274,** 3011-3014.

19. R. Pautrizel, A. Priore, P. Mattern, and A.N. Pautrizel, "Stimulation des defenses de la souris trypanosomee **par** l'action d'un rayonnement associant de champ magnetique et ondes electromagnetiques." [Stimulation of the defenses of the trypanosomized mouse by the action of a ray associated with a magnetic field and electromagnetic waves.]. **Compt. Rend. Acad. Sci.** (Paris), 1975, **280,** 1915-1918.

20. R. Pautrizel, A. Priore, P. Mattern, A.N. Pautrizel, and A. Capbern, "Guerison de la trypanosomiase chronique du Lapin a **Trypanosoma equiperdum** par l'action combinee de champs magnetiques et d'ondes electromagnetiques modules." **J. Protozool.,** 1975, **22,** no. 3, A 84.

21. R. Pautrizel, "La trypanosomiase experimentale: stimulation des defenses de L'organisme par des moyens physiques." **XVIIth Seminar on Trypanosomiasis Research,** Londres, 22-23 Septembre 1976.

22. R. Pautrizel, P. Mattern, A.N. Pautrizel, and A. Priore, "Effets des champs magnetiques et des ondes electromagnetiques modules sur la trypanosomiase experimentale." [Effect of magnetic fields and modulated electromagnetic waves on the experimental trypansomiasmis.]. **Ann. Soc. Beige Med. trop.,** 1977, **57,** 501-523.

23. R. Pautrizel, P. Mattern, A. Priore, A.N. Pautrizel, A. Capbern and T. Baltz, "Importance des mecanismes immunitaires dan la guerison de la trypanosomiase experimentale par stimulation physique." [Importance of immune mechanisms in the cure of the experimental trypanosomias by physical stimulation.]. **Compt. Rend. Acad. Sci.** (Paris), 1978, 286, 1487-1492.

24. R. Pautrizel, and A. Priore, "Un aspect spectaculaire du bioelectromagnetisme: permettre a l'organisme de se debarrasser d'une infection aig ue ou chronique sans l'aide d'aucune substance medicamentese." [A spectacular aspect of bioelectromagnetism: permitting the organism to rid itself of an old chronic infection without the aid of any medical substance.] **Communication presented to the 104 eme Congres National des Societes Savantes,** Bordeaux, 17-21 Avril 1979 — section Sciences, p. 112.

25. R. Pautrizel, and A. Priore, "Guerison de la trypanosomiase experimentale par association de champs magnetiques et d'ondes electromagnetiques," [Cure of the trypanosomias by associated magnetic fields and electromagnetic waves.]. **Communication presented to the Symposium International des Therapeutiques Ondulatoires,** Versailles 19-20 Mai 1979, p. 9.

26. R. Pautrizel, A. Priore, A.N. Pautrizel and P. Chateaureynaud-Duprat, "Guerison de la trypanosomiase experimentale par l'association de champs magnetiques et d'ondes electromagnetiques: une stimulation des defenses de l'organisme." [Cure of the trypanosomias by associated magnetic fields and electromagnetic waves: a stimulation of the organism's defenses.]. **Communication presented to the Journees Nationales Microondes - Colloque Hertzienne et Dielectriques,** Lille 26-29 juin 1979, p. 31.

27. David M. Rorvik, "Do the French have a Cure for Cancer?" **Esquire** Magazine, July 1975, p. 110-111,142-149. Excellent summary of the Priore Affair, with some details of the working of the machine.

28. J.B. Bateman, "A Biologically Active Combination of Modulated Magnetic and Microwave Fields: The Priore Machine," Office of Naval Research, London Report R-5-78, Aug. 1978. Comes very close when he states that "The possibility that some hitherto unrecognized feature of the radiation from a rotating plasma may be responsible for the Priore effects should not be dismissed out of hand..." Indeed, none of the persons who checked the machine or pondered over it knew anything about phase conjugation — which plasmas can do.

29. J.B. Bateman, "Microwave Magic." Office of Naval Research London Conference Report, ONRL C-14-77, 1977.

30. D.S. Greenburg. "The French Concoction." **Saturday Rev. Sci.,** May, 1978. p. 36-44.

31. Lord Zuckerman, "The great Bordeaux magnetic mystery machine." **Sunday Times Weekly Review,** Jan. 7, 1973.

32. Lord Zuckerman, "Pride and Prejudice in Science. William Randolph Lovelace Commemorative Lecture. **Aerospace Medicine,** 1974, 45, 638-647.

33. L.B. Bateman, "Staging the perils of non-ionizing waves." Office of Naval Research London, **European Scientific Notes,** ESN 32-3: 85-88.

34. R. Courrier, 1977. "Expose de M. le professeur R. Courrier secretaire perpetuel de l'Academie des Sciences fait au cours d'une reunion a l'Institut sur les effets de la machine de M.A. Priore le 26 Avril 1977.

35. M. Riviere, A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Phenomena of regression observed on the graftings of a lymphosarcome in mice exposed to electromagnetic fields." Mar. 1, 1965.

36. Jean-Michel Graille, **Le Dossier Priore:** De Noel, Paris, 1984. Published in France.

37. G. Lakhovsky, "Apparatus with circuits oscillating under multiple wave lengths," U.S. Patent No. 1,962,565. June 12, 1934.

38. G. Lakhovsky, "Tube for producing multiple wave lengths." U.S. Patent No. 2,351,055. June 13, 1944.

39. Thomas H. Moray, "Electrotherapeutic apparatus," U.S. Patent No. 2,400,707. Feb. 1, 1949.

40. Robert C. Beck, "Extremely low frequency magnetic fields and EEG entrainment: A psychotronic warfare possibility?" Audiotape; presentation at the 1978 U.S. Psychotronics Association (USPA) Conference, Atlanta, Georgia.

41. T.G. Hieronymus, "History of radionics. Parts 1 and 2." Audiotape; presentation at the 1979 USPA Conference, University of Houston, Texas.

42. Robert C. Beck, "Extreme low frequency magnetic pulses and E.E.G. entrainment: Soviet Psychotronic Weapon?" Audiotape; presentation at the 1979 USPA Conference, University of Houston, Texas.

43. Al Bielek and William Bise, "A magnetic pacer." Audiotape; presentation at the 1980 USPA Conference, Dayton, Ohio.

44. Thomas E. Bearden, "Mind and matter interaction." Audiotape; presentation at the 1980 USPA Conference, Dayton, Ohio.

45. John Moray, "Sea of energy." Audiotape; presentation at the 1980 USPA Conference, Dayton, Ohio.

46. T. Galen Hieronymus, "Sunspots trigger earthquakes." Audiotape; presentation at the 1981 USPA Conference. Available from USPA, 3459 Montrose Ave., Chicago, IL 60618.

47. Al Bielek and Bill Bise, "Measurements of ELF and its parameters." Audiotape; presentation at the 1981 USPA Conference.

48. Martin Ruderfer, "Misconceptions in established science and their relevance to psychotronics." Audiotape; presentation at the 1981 USPA Conference.

49. Material on the Kaznacheyev experiments. See Kaznacheyev section for references.

50. Material on phase conjugation. See phase conjugation section for references.

51. Material on Ehrenhaft experiments. See Ehrenhaft section reverences.

52. Material on Kervran's work. See Kervran section for references.

53. Extensive proprietary scalar electromagnetics experiments by Frank Golden and John Bedini.

54. Material on Royal Raymond Rife's work. See Rife section for references.

References Related to the Australian Work

1. Private correspondence, Dr. B.L. Reid.

2. Private correspondence, Dr. S. Barsamian.

3. B.L. Reid, "Propagation of properties of chemical reactions over long distance in the atmosphere as seen by crystal growth pattern changes," **Australian Journal of Medical Laboratory Science,** Vol. 7, Feb. 1986. p. 30-35.

4. H.E. Anderson and B.I. Reid, "Vicinal Long Range and Extremely Long Range effects on growth of sodium chloride crystals from aqueous solutions containing protein," **App. Phys.** Comm, vol. 4, 1984. p. 217.

5. W.H. Campbell and J.M. Young, "Auroral zone observations of infrasonic pressure waves related to ionosopheric disturbances and geomagnetic activity," **J. Geophys. Res.,** vol. 68, 1963. p. 5909.

6. G. Chimonas and C. Hines, "Atmospheric gravity waves launched by auroral currents," **Planet Space Sci.,** vol. 18, 1970. p. 565.

7. F.B. Abeles, **Ethylene in Plant Biology,** Academic Press, New York, 1973.

8. G.I. Distler, **Jour. Cryst. Growth,** Vol. 3 & 4, 1968. p. 175.

9. G.I. Distler and V.P. Vlasov, **Thin Solid Films,** vol. 3, 1969. p. 333.

10. G.I. Distler, **Jour. Cryst. Growth,** Vol. 9, 1971, p. 76.

11. B. DeWitt, **Phys. Rev.,** 19C, 1975, p. 297.

12. W. Greiner and J. Hamitton. **Amer. Sci.,** vol. 68,1980, p. 154.

13. P. Davies and T. Birrell, **Quantum Fields in Curved Space,**

    Cambridge University Press, Cambridge, 1982.

14. P. Mitchell, Nature,vol. 191, 1961. p. 144.

15. A Rothen. **Proc. Nat. Acad. Sci.** (U.S.), vol. 72,1975. p. 2462.
16. L. Niemeyer, L. Betroniero, and H.J. Weismann, **Phys. Rev. Lett.,** vol. 52, 1984. p. 1033.

17. T.A. Witten and L.M. Sander, **Phys. Rev. B.,** vol. 27,1983. p. 5686.

18. T.D. Lee, **Particle Physics and Introduction to Field Theory,** Harwood Academic Publishers, New York, 1981. See particularly Chapter 16, "Vacuum as the source of asymmetry."

References Related to Royal R. Rife's Work

1. Royal R. Rife, "Bacteria," Apr. 28, 1939.

2. Royal R. Rife, "The filterable virus of carcinoma," Apr. 26, 1939.

3. Royal R. Rife, "The Rife microscopes," Apr. 28, 1939.

4. Royal R. Rife, "History of the development of a successful treatment for cancer and other viruses, bacteria and fungi." Allied Industries, Dec. 1, 1953.

5. Arthur Isaac Kendall and Royal Raymond Rife, "Observations on Bacillus Typhosus in its filterable state," **California and Western Medicine,** California Medical Association, Dec. 1931.

6. Arthur Isaac Kendall, "Filtration of bacteria; studies in bacterial metabolism," **Science 75,** Mar. 18, 1932. p. 295-301.

Arthur Isaac Kendall, "The filtration of bacteria," **Science,** Mar. 18, 1932.

8. E.C. Rosenow, "Observations on filter-passing forms...," **Proceedings of the Staff Meeting of the Mayo Clinic,** July 13, 1932.

9. E.C. Rosenow, "Observations with the Rife Microscope of Filter-passing forms of micro-organisms," **Science,** n.s., vol. 76, Aug 26, 1932. p. 192-193.

10. E.C. Rosenow, "Microdiplococci in filtrates of natural and experimental poliomyelitic virus compared under the electron and light microscopes," **Proceeding of the Staff Meeting of the Mayo Clinic,** 17(7), Feb. 18, 1942. p. 99-106.

11. E.C. Rosenow, "Observations of filter-passing forms of Eberthella Typhi (Bacillus Typhosus) and of the Streptococcus from Poliomyelitis," **Proc. Staff. Meet. Mayo Clinic** 7, July 13. 1932. p. 408-413.

12. E.C. Rosenow, "Transmutations within the Streptococcus-Pneumonococcus Group," **Journal of Infective Diseases,** vol. **14, 1914.**

13. "Powerful microscope makes very tiny objects visible," **Science Newsletter,** vol. 33, Jan. 22, 1938. p. 55.

14. "The Rife microscope or "facts and their fate'," Reprint No. 47, The Lee Foundation for Nutritional Research, Milwaukee, Wisconsin.

15. R.E. Seidel, M.D. and M. Elizabeth Winter, "The new microscopes," **Journal of the Franklin Institute,"** Feb., 1944. p. 103-129.

16. R.E. Seidel, M.D. and M. Elizabeth Winter, "Filterable bodies seen with the Rife microscope," Science Supplement, **Science,** Dec. 11, 1931. p. 10-11.

17. "Is a new field about to be opened in the science of bacteriology?" Editorial, **California and Western Medicine,** California Medical Association, Dec. 1931.

18. Christopher Bird, "What has become of the Rife microscope?", **New Age Journal,** Mar. 1976.

19. Peter Tompkins and Christopher Bird, **The Secret Life of Plants,** Harper & Row, New York, 1973.

20. Lida H. Mattman, **Cell Wall Deficient Forms,"** CRC Press, Cleveland, Ohio. 1974.

21. Georges Lakhovsky, **La Cabal: Histoire d'une Decouverte L'oscillation Cellulaire,"** G. Doin, Paris, 1934.

22. Georges Lakhovsky, "Apparatus with circuits oscillating under multiple wave lengths," U.S. Patentno. 1,962,565. June 12,1934. See also G. Lakhovsky," Tube for producing multiple wave lengths," U.S. Patent No. 2,351,055. June 13, 1944.

23. Georges Lakhovsky, **"L'oscillation Cellulaire," Ensemble des Recherches Experimentales,** G. Doin, Paris, 1931.

24. Harold Saxton Burr, **Blueprint for Immortality: The Electric Patterns of Life,** Neville Spearman Ltd., London, 1972.

25. George De La Warr and Douglas Baker, **Biomagnetism,** De La Warr Laboratories, Oxford, 1967.

26. Ruth Beymer Drown, **The Theory and Technique of the Drown H.V.R. and Radiovision Instruments.** (Private printing.). Artists' Press, Los Angeles, 1939.

27. Ruth Beymer Drown, **The Science and Philosophy of the Drown Radio Therapy.** Los Angeles, 1938.

28. Albert Abrams, **New Concepts in Diagnosis and** Treatment. Philopolis Press, San Francisco, 1916.

29. Albert Abrams, **Iconography: Electronic Reactions of Abrams.** San Francisco, 1923.

30. Jagadis Chandra Bose, **Response in the Living and Non-Living,** Longmans, Green & Co.; New York, 1902.

31. Jagadis Chandra Bose, **The Physiology of Photosynthesis,** Longmans, Green & Co.; New York, 1924.

32. D.M. Bose, "J.C. Bose's Plant Physiological Investigation Relating to Modern Biological Knowledge," **Transactions of the Bose Research Institute,** Vol. 37, Bose Research Institute, Calcutta, 1947-48.

33. "Invisible ultra-violet life frequencies made visible," **New Light on Therapeutic Energies,** Source/data unk., p. 51-55.

34. Wilheim Reich, **The Cancer Biopathy.** Orgone Press, New York, 1948.

35. V.M. Inyushin and P.R. Chekorov, "Biostimulation through laser radiation and bioplasma," Kazakh State University, U.S.S.R.

36. Philip Medley et al, "The filterable forms of bacteria," **Journal of Infectious Diseases,** vol. 48, 1931.

37. Gianni A. Dotto, "Method for constructing a thermionic pile," U.S. Patent no. 3,839,771. Oct. 8,1974.

38. Gianni A. Dotto, "Story of the Dotto Ring." notarized in Montgomery County, Ohio on Feb. 8, 1972. Certain parts of Dotto's theory, apparatus, and recommended treatment method suggest potential effects on recombinant genetic structures.

39. Gianni A. Dotto, "Electrostatic Wand," U.S. Patent No. 3,785,383. Jan. 15, 1974.

40. T.H. Moray, **The Sea of Energy.** 5th Edition. History and biography by John E. Moray. Foreword by Thomas E. Bearden. Cosray Research Institute, 2505 South 4th East, Salt Lake City, Utah 84115. 1978.

41. T.H. Moray, "Electrotherapeutic Apparatus," U.S. Patent no. 2,460,707. Feb. 1, 1949.

42. John Moray, "Radiant Energy," in **1st International Symposium on Non-Conventional Energy Technology,** Toronto, Canada. Planetary Association for Clean Energy, 1981. p. 316-319.

43. Antoine Priore's research and devices.

44. Wilheim Reich's orgone energy work and devices.

45. Radionics researchers' work and devices.

46. Jonathan Tennenbaum, "Soviet Work on Electromagnetic Pulse Weapons," EIR Special Report, **"Electromagnetic-Effect Weapons: The Technology and the Strategic Implications,"** Wiesbaden, Federal Republic of Germany, 1988.

47. Ron Iscoff, "Dye Lasers Used to Treat Blood," **Lasers & Optronics,** 7(3), Mar. 1988, p. 26-28.

APPENDIX I

The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science

© 1984, 1988 Christopher Bird (Used by permission)

Forty- four years ago, in 1944, an Italian engineer working as a prisoner and forced laborer for the Germans in the huge submarine base in Bordeaux, approached a French police agent to plead for his life. He would be killed when the Germans left Bordeaux, he said, and since they were by that time obviously losing the war, the day of his execution was at hand.

The police officer, who also worked clandestinely for the French underground, told the engineer to get in his car, then simply drove him out of the base and introduced him to the 7th battalion of underground resistance fighters, in the nearby province of Dordogne. There he so distinguished himself in military operations that he was ultimately decorated by the French government.

It was due to his thankfulness to his savior, and his loyalty to his companions-in-arms, that Antoine Priore decided after the war's end to live out the rest of his life in Bordeaux. Thus he became the focus of one of the strangest, and most scandalous, chapters in the scientific history of France or any other nation.

Antoine Priore had earlier graduated from a small provincial school for electricity in Trieste, Italy and become a radar operator in the Italian Navy. During this period he observed what to him was an exciting anomaly: some oranges left in a room filled with electrical bric-a-brac had fallen into an assemblage where they seemed to have been preserved in the same fresh state they had enjoyed when bought off a fruit stand. Other oranges in the room, bought at the same time, were rotten and putrid.

Stunned by his observation, Priore dreamed throughout the war of one day working out an electrical means of conserving foods in their fresh state based on what he surmised was a new, and wholly unexplained, principle. Newton's apple had become Priore's orange.

Occupied during the day as a humble electrical repairman — and projectionist in a movie theater — the almost wholly self-taught Priore devoted all his free time and his meager resources to research. With the help of his war-time companions, some of whom had attained high rank in the Bordeaux police force, he was able to beg, borrow, steal, scrounge, or otherwise acquire a mini-warehouse of electrical and electronic components and parts. With these he put together a device worthy of Rube Goldberg. Exposing lentil seeds to a magnetic field of 225 gauss and electromagnetic frequencies of 80, 32, 3 and 10 Hertz, Priore's device caused the lentil plants which sprouted from them to grow 12-15 centimeters in length, as against only 5 centimeters for controls not subjected to the same treatment. He got similar results for tulips, asparagus and other plants.

Shifting his focus, he next irradiated fertilized hens' eggs, only to see the chicks hatch in 19 days, instead of the normal 21. Though he could not explain these astonishing results, he realized he had stumbled upon a process basic to the enhancement, or speeding up, of cellular growth.

It was at this point that one of his police friends introduced him to Francis Berlureau, the former Director of Studies at the School for Veterinary Medicine in Toulouse and, at the time of their meeting, director of the Bordeaux abbatoir. Priore asked Berlureau to supply him with various animal tissues for experimentation. For 10 years they worked together, Priore's free time allowing, during which Priore noticed he could get no electrical measurement from a cancerous bull's testicles. Since he realized that, in some way, his newly constructed device (no trace of which remains today, except for a snapshot of it) affected the electrical properties of cells, he put two and two together and his sum of four led him to believe that he might be onto an electromagnetic cure for cancer. Newton's gravity had become Priore's cancer cure.

Berlureau next allowed him to expose a cat with cancer of the mammary glands to radiation of his machine. To make absolutely sure that he was not exposing himself to mockery, the veterinarian had all the histological work done by his friend and colleague, a Professor Drieux at the famous Veterinarian School of Maisons-Allfort, near Paris. Drieux wrote a technical report proving that a tumor taken from a cat had, before treatment, started to become cancerous and, after treatment, had become benign.

By 1953, with the help of a doctor of general medicine, Maurice Fournier, Priore began treating human patients whose cancers had been judged hopeless. The huge file of cases maintained by Fournier, and filed with a notary until after his death, was subsequently mysteriously lost. But a few details were preserved in letters discovered in an old dog-eared file.

Some of these dating to the year 1954 concerned a 12-year old boy, Alain B., whose diagnosis wavered between one of reticulo-histio-sarcoma and a malignant form of Hodgkin's disease. The boy was taken by his parents to Priore, who irradiated him. Though the exact nature of the radiation was not known, 12 years later a Bordeaux physician, after a medical examination, certified that the boy, now become a man of 24, was free of disease.

A second case unearthed from the old file indicated that a patient with cancer of the larynx was able to avoid a laryngectomy and be totally cured after Priore's new ministration.

Fascinated by the principle which he suspected must lie behind the strange Priore Ray, Dr. Berlureau tried to get some Bordeaux University physicists interested in the problem but was laughed out of their offices. He next turned to cancer specialists, beginning with Professor Lachapele, the Director of the Bergonie Foundation, a prestigious center for cancer research, to whom he proposed animal experiments to prove the efficacy of Priore's methodology. His plea met with a stony affirmation on Lachapele's part to the effect that he and his colleagues had no need of the new discovery, inasmuch as "all the patients treated in his hospital were cured and departed in perfect health." As if bound in the chains of his curt reply, years later Lachapele was to become one of the bitter adversaries of Priore's pioneering research.

Only somewhat discouraged, Priore kept up his momentum. He went on to build a new and more complicated version of his treatment device, called the P-l, over the next year. When it was finished he secretly and unofficially began to treat dozens of cancer patients who had been given up by their doctors as incurable. At his funeral in March of last year, among the crowd of mourners was, it is said, a small platoon of older people who had been cured of their terrible afflictions by Priore in the 1950s.

While his findings excited him, he nevertheless felt tremendously frustrated that he could apparently get no one in the world of medicine or science to pay attention to them. Undaunted by his previous rebuffs, his friend Berlureau next introduced the Italian at the end of 1959 or the beginning of 1960 to Professor Tayeau, vice dean of Bordeaux's Medical Faculty. Unlike Lachapele, Tayeau behaved as a true physician and scientist. He sent Priore to two researchers, Biraben, head of the Faculty's Department of Pathological Anatomy, and his assistant, Delmon. The two had been working together on cancerous rats for two years—specifically on animals grafted with T-8 tumors, discovered by the internationally famous team of Guerin and Oberling in Paris, which had proven to be intractable to any form of treatment yet known. To their utter surprise, the tumors in the rats treated with Priore's machine were reduced in volume by 60%, marking the first time in the history of cancerology that the virulent T-8 tumor had in any way been affected by any form of treatment.

Knowing that the mayor of Bordeaux, Jacques Chabans~Delmas-who has kept his post until this day, and was soon to become prime minister of France-- was most interested in the work of Priore (who, he too, had known as a fellow resistance fighter), they also informed Chaban.

Promptly Chaban convoked not one, but two, commissions made up of Bordeaux and Parisian scientists to study the Biraben-Delmon results in detail. Both commissions rejected Priore and his machine out of hand, and without appeal. It is curious that, in the science of our day, a result, undeniable though it may be, seems to have no hearing unless and until all means to effect it can be adequately explained. It was for this ostensible reason that the two commissions decided to so adamantly reject the research: Biraben and Delmon could not explain the nature of the radiation engendered by the Priore device.

One can stress the word *ostensible* here because the principal reason for the rejection lay elsewhere. The decision by the first commission was, in fact, hardly unanimous. But among its members was the same Professor Lachapele who had refused Berlureau's plea for assistance. His opinion was that even the results themselves were of little value because they were obtained, not on spontaneously arising, but on grafted, cancers. The fact that no treatment whatsoever had ever affected a T-8 tumor was totally discounted. As the sole cancerologist on the commission, Lachapele's dictum was preponderant.

When he learned that the rejection of the first commission had actually been a split decision, the Bordeaux mayor asked for the formation of a second commission to re-examine the problem. Fearing a reversal, Lachapele was able to get one of his colleagues, Professor Courtial, director of the Radium Institute in Paris, and one of the so-called top authorities of French cancer research, named to it. It was all but impossible for the other physicians on the new commission now to outvote not one, but two, cancer specialists, so again the antagonists won the day.

At no time did either of the commissions bother to interview Priore himself or to run a supplementary experiment under their own control.

This seemingly incomprehensible attitude on the part of scientific authority was only a foretaste of what was to come, again and again, over the years. Biraben and Delmon went on to do new experiments. They modified either the time after grafting that the radiation was applied, or the length of its duration. This time their efforts were crowned with unequivocal and complete success. The tumors stopped growing and, when still living cells were excised from them and implanted in healthy control animals, none of them became malignant.

Though these results should normally have fascinated any academy of medicine or sciences, the two researchers did not publish them. Why? The reasons horrify or disgust. It seemed that Biraben was simultaneously preparing an examination for the *agregation*, the highest French academic degree leading to a senior university teaching post. In charge of the committee to pass on, and award, this degree was none other than that same Professor Lachapele who told him: "Either you get the degree, necessary to your professional advancement, or you publish your research paper. But not both!" Discouraged, Biraben ceded to this demand but nevertheless continued to work on the research that looked so exciting and promising.

Most mystifying to him was how the machine operated to achieve its startling results. At the 3rd Congress of Biometerology held in 1963 in the Pyrennees mountains, a New York City researcher by the name of Kenneth McLean reported he had been able to obtain regressions on tumors and improve the health of cancer patients by using a magnetic field of a strength of 3000 gauss or more. Acting on this hint, Biraben and Delmon made an electromagnet that put out a field of 4,500 gauss and tried it out on the T-8 tumors but without the slightest success. Obviously, something other than a simple magnetic field was at issue.

In 1966, after others had had the same success with the T-8 tumor by irradiating it with the "Priore Ray," the two scientists finally published a memoir in the Revue of Comparative Pathology in which they stated that neither magnetic fields nor X-rays had any effect on the T-8s and that "only certain devices associating a magnetic field with high frequency waves seem at present to reveal therapeutic properties..."

Their conclusions were too late for, by that time, a campaign to stamp out Priore and his electromagnetic approach to cancer cure was well underway, a campaign that has lasted right up to the present moment.

The all-powerful Lachapele had sealed the fate of the Priore device as far as the local Bordeaux medical community was concorned. Veterinarian Berlureau and Priore next decided to carry their case to Paris. They contacted Professor Guerin at the cancer Institute at Villejuif, the leading French center for cancer research nnd the equivalent of the American National Cancer Institute in Bethesda, Maryland. Guerin, one of the discoverers of the T-8 tumor, which for the first time had been stopped in its tracks by the Priore device, courteously received his guests and heard them relate the whole story of how the device had come into being, starting at the point. when Priore had seen the oranges strangely preserved by some unaccountable electromagnetic effect.

Guerin was sufficiently impressed that he assigned his colleague, Marcel-Rene Riviere, to delve into the whole question. For two years, Riviere, who also had teaching responsibilities at the University of Rennes in Brittany, unremittingly worked to corroborate the Biraben-Delmon findings. On 9 December 1964 a note was sent for publication in the **Proceedings** of the French Academy of Sciences detailing the research and modestly concluding: "... as of now, one may already state that our first observations show that electromagnetic fields used can lead to most interesting data from a point of view of the biological behavior of grafts and their therapeutic action on experimental tumors."

Riviere next decided to see if the Priore Ray could affect another tumoral form that had never been affected by any therapeutic method, the 347 lymphoblastic lymphosarcoma. The results were even more spectacular than for the T-8 tumor. The effects produced were of broader scope and took place more rapidly. A second note was sent to the Academy for publication. The conclusion read: "We can now already affirm that our research offers proof that electromagnetic fields are capable of producing effects on quite different types of neoplasms."

At this point one of the key characters in this extraordinary drama must be introduced. There might have been no drama at all without his appearance on stage. This personage was Robert Courrier, an eminent endocrinologist, who had been named, while still in his 30s, a full professor. Courrier was now perpetual secretary of the Academy of Sciences and later would become President of the Academy of Medicine. Because no scientific paper can be accepted by an academy unless introduced by one of its members, Riviere would have had no chance to see his work so prestigiously published had not Courrier, who knew Riviere well, since he had shepherded him through the winning of his doctorate, taken the responsibility for its introduction.

It was Courrier who, at this point, also took up the cudgel to interest various highly placed French organizations responsible for the administration of scientific projects and their funding. Thus, ho sent a personal letter to the French Minister for Scientific Atomic and Space Research, who immediately offered to try to make funds available for further research on and with the Priore Ray. He also personally asked the Director of the CNRS (National Center for Scientific Research, which coordinates and oversees all such activity in France) to receive Priore and Riviere.

That this meeting was, in its way, somewhat of a disaster, can be explained in part only by a brief resume of the complex character of Priore himself. Priore throughout his life had great difficulty making himself understood in the French language and, as the **years** went by, he even forgot how adequately to speak his own mother tongue, Italian. Added to these twin impediments was his lifelong fear that his discoveries and inventions were prone to being stolen, a fear which led him never to fully explain the exact nature of the complex radiation emitted from his machine, far less the settings which controlled its various parameters. Whether Priore would not, or could not, exactly explain the functioning of his invention (which, as we shall see, went through several increasingly complex generations) is a question to which no precise answer has been given. It would appear that Priore was an excellent engineer gifted more with a God-given intuition than with school-book reasoning and logic. In short, Priore had a combination of talents that could remind one of the same enigmatic personality that was Nikola Tesla, the deductive reasoning behind some of whose discoveries has never fully been unravelled.

Highly placed scientific administrators are neither comfortable with, nor sympathetic to, what they see as self-appointed geniuses who have not run the same academic gauntlet through which they themselves had to pass. Thus the CNRS director took aversion to Priore's somewhat incomprehensible, yet fairly prolix explanations of his technology and only recommended that a physicist be sent to look over his device to properly decipher its working. At the same time, the Minister, together with the head of the general delegation for Scientific Research (DGRST) — still another key body in the administration of the French scientific decision-making process — let it be known to Robert Courrier that they had not understood a tingle thing about Priore's invention despite his best efforts to present it.

Advancing one more step into what was to become for him a 20 year-long expedition into a jungle of scientific intrigue, Courrier next resolutely decided to send to Bordeaux one of his most trusted laboratory workers, Madame Colonge, to repeat Riviere's experiments under her personal supervision. The DGRST director fully Concurred with Courrier's decision, while letting slip his admission that he strongly suspected that Riviere might well have been duped in some manner by Priore. When Courrier asked the minister for travel funds for Madame Colonge, he was refused with the dry remark that such a request was "premature." The now angry Courrier telephoned to reply: "You've been spending millions for programs and hypotheses about the cancer problem. Riviere has been presenting you with facts!"

The physicist who had been sent to try to elucidate the functioning principles of Priore's device, reported that he could make neither "head nor tail" of the machine.

Madame Colonge's experimentation was as prolonged as necessary. She was so meticulous that, in order not to take her eyes off the experimental animals for an instant, she limited her lunches to sandwiches eaten in the laboratory. She returned, profoundly impressed, to Paris.

Nor was Riviere idle during this period. He decided to experiment with the L-52 lymphosarcoma, a tumor similar to, but even more malignant than the 347. This time, he used not rats, but mice, as hosts for the grafts. So successful were his results that, this time, Courrier decided to bar no holds. Instead of simply having a note published in the Academy Proceedings on the quiet, he decided to present it personally, orally, and in all solemnity, before his fellow academicians at an official meeting scheduled for 1 May 1965.

That date was, and is, a turning point in what came to be known in France as the "Priore Affair." From then on the whole French, and even the international, scientific community could be divided into a minority and a majority group, the first that believed in the research, the second that did not know enough details about it, did not or would not believe, or simply didn't give a damn.

Before Courrier could make his presentation, its scheduling and subject were inadvertently and prematurely leaked to the press. Immediately thereafter, a horde of journalists arrived in Bordeaux. After one or two of them managed to all but force entry into his lab, Priore closed its doors and, with the help of his friends, wrote a printed press release that stressed his thankfulness to the many people who had helped him over the years rather than providing any comprehensible details about the machine he had brought to birth. Stymied, many of the newsmen travelled across town to seek an explanation from its leading cancerologist, Professor Lachapele, who informed them acidly that the machine was all but useless and unworthy of their attention or their time.

The journalists' reports, founded as they were on rumor of outright lies, roiled the pages of their newspapers and magazines in such a way that they either over-exaggerated the potential or a forthcoming cancer cure, or came close to billing Priore as just one more cancer-cure charlatan. All of which so alarmed, among many others, Dr. Wilhelm Bernhard, world specialist in electron microscopy, that he called his friend Courrier to warn him that his forthcoming presentation to the academy might put his hard-won reputation at risk as well as those of Guerin, Riviere and the Villejuif cancer institute itself.

In the journalistic melee, no one had bothered to read the declaration carefully issued by Riviere from Rennes, where he was occupied with his university courses, which formally stated: "Our experiments are of real interest. Much more experimental research has of course to be done before any therapeutic application on human cancer victims can even be considered. It goes without saying that this will take a certain time and no little effort of many types, both scientific and financial. Our work, as fascinating a perspective as it might hold, in no way, therefore, allows anyone to offer the public hope which could only lead to deception at the present time."

On May Day, Courrier gave his report to an Academy assembly hall crowded with scientists, newspapers and television reporters, photographers and an unusual number of curious bystanders. Accompanied with slides showing histological details, and animals before and after treatment, his lecture was heard out in almost tomb-like silence. When the lights came on again, he announced that he had personally checked the validity of Riviere's findings through the offices of his personal assistant, Madame Colonge. He then went on to say that he had taken the responsibility of presenting three notes to the Academy for two reasons. The first was a ringing declaration and a challenge to critics and skeptics of every stamp. It reads "When it is a question of a problem as serious as that of cancer and when one sees a little light beginning to dawn, one has the obligation to see what this light might represent. One has no right to snuff it out before learning what it may be worth."

The second was a tribute to his colleagues who had done the pioneering work, particularly Guerin, Riviere and Madame Colonge, and a statement of the essence of the problem to be faced down the road. It read: "Attention must naturally be given to the apparatus which Monsieur Priore has conceived and constructed. It has apparently already been examined by several physicists. It is found to be too complicated. While that may be possible, it is nevertheless a fact that Guerin and Riviere have obtained results with it that had to be made public. What is emitted from such an apparatus? I hope Monsieur Priore will allow disinterested physicists to study it at their leisure, for Science cannot tolerate apparatuses enveloped in mystery." Then, as almost a footnote to the history of the moment: "The biological action of magnetic fields is the object of intense research in the United States. In specialized institutes, the influence of these fields on tissue cultures, microbes, plant forms, diastases and certain tumors is under study. Up to now, the results obtained on grafted tumors seem less significant than those which have here been presented."

After Courrier sat down, a leading cancer specialist, Professor Lacassagne, rose to ask snidely why the notes had included no bibliographical references on work done on the bio-effects on tumors from electromagnetic fields, and criticized the experimentation as "impromptu." When Courrier denied this allegation as ludicrous, Lacassagne stalked out of the assembly hall in full view of the audience.

The meeting caused a new eruption of media reports which unfortunately accented one of three aspects of the problem at the expense of the other two. These were 1) the hope that a miraculous cure for cancer was in the offing 2) the contradictory, not to say discordant, reception of the data by various academics and 3) the enigma of Priore's personality.

Remarkably, no serious discussion among the scientists present at the meeting ever took place. This led a foreign scientist, present in Paris at the time, to remark: *"I don't understand. Here is a report given to the most authoritative scientific body in France by one ofthe most respected and eminent ofits members and it is publicly subject to doubt without that leading to any reaction whatsoever."*

The General Delegation for Scientific Research was at this point still open to the idea of providing funds for more research with the Priore Ray. The big stumbling block, however, was one related to niceties involved in relations up and down the scientific hierarchy. To open the way to the allocation of such funds required the approval of the Delegation's own section for cancer research and that section was headed by none other than the same Professor Lacassagne who had so rudely walked out of the Academy's assembly hall.

At the same time the General Delegate diplomatically covered himself by suggesting to Robert Courrier that he had to have more information underscoring the potential importance of the research accomplished. Courrier told him to simply re-read the three notes he had presented to the academy. His matter-of-fact, yet terse, come back then elicited his invitation to a full-dress meeting of scientific experts at the Institute for Scientific Cancer Research.

In this short historical account we obviously cannot go into the ins and outs of what transpired at this meeting or any of the many similar meetings which followed it. A paragraph in a brilliant book, four years in the writing, by the courageous Bordeaux journalist, Jean-Michel Graille and entitled: **Dossier Priore, A New Pasteur Affair** (of which this account is but a tenuous synopsis) must suffice to pointedly characterize the nature of the problem in its most general sense.

Writes Graille: "To read what follows in this chapter might well seem tedious: an enumeration of names and titles, the content of a debate held at an administrative meeting, personal remarks by one ranking personage or another, exchanges of letters following the meeting itself and the official report which came out of it. Tedious but indispensible for not a few reasons. It is important to know who were the participants at this meeting and what of these participants each was trying to represent. It is important to know how such scientific meetings go about their business at the 'top level,' And, finally, it is important to learn about and to understand, in the particular case of the Priore Affair, the behavior and reactions of all concerned. A reading of all this could be difficult (and it won't be the only such passage in this book) yet it is necessary to understand the essence of the dossier in order to be able to create for oneself as clear and well-motivated a personal opinion about it as to subsequently be able to discuss it, or to hear it discussed, with a thorough knowledge of the facts."

In this single paragraph, Jean-Michel Graille has, in my opinion, pointed to both the nub and the difficulty in getting at the essence of the real facts behind a case such as that of Priore's that are so important to its proper understanding, an understanding which can be painted against the backdrop of the history of science and the backdrop of human pettiness and maliciousness or human courage and magnanimity.

One of the participants, Professor Andre Lwoff, soon to become n Nobel Laureate for his work in virology, was violently against the meeting's central issue: namely, whether or not funds should be spent to build a new and better Priore machine. Not only did Lwoff aver that the three notes presented to the Academy never should have been published, but he also opined that since all the work was done, not on spontaneously generated but on grafted cancers, the effects of the machine were hardly impressive. He later added in writing that 1) the patents issued to Priore for his device were nothing but a web of nonsense, 2) the machine itself could never be duplicated based on any description given for it by its inventor and, in a repetition of his oral remarks, 3) the fact that only cancer grafts were experimented with was nothing to shout about. He strangely added that because the animals who had been irradiated subsequently were able to entirely reject new grafts, the whole phenomenon offered no proof that cancer cells could be killed while healthy cells were not. The whole thing came down to a question of immunity, he said, as if that were not of the greatest possible importance.

To which, in due course, Guerin and Riviere replied: "It has been claimed that our experiments are valueless because they were carried out on grafted tumors and that other therapeutic measures were known to get rid of such tumors and their metastases. *We defy those persons who have made such affirmations* to prove, *with the use of such other measures, that animals infected with T-8 tumors can be cured at a percentage rale identical to those obtained by using the device which Monsieur Priore has developed."*

Not a soul has responded to this challenge, then or since.

A second cancer expert at the meeting, a woman of great influence, resorted only to the cavil that the experiments had been of doubtful quality since none of the animals had been *weighed.* The fact that those same animals had survived normally lethal cancers seemed not to have *weighed* with her.

There were many more observations of the same ilk. They seem atrociously paltry, trifling and picayune coming from professionals who, if they no longer believed in the Hippocratic oath to which they once swore, are considered by the public in general, and by cancer patients in particular, at least to be concerned with seeing what a little light on the problem might reveal before extinguishing it, as Professor Courrier expressed it.

At the same time, we must not forget Priore's decidedly difficult personality. He was an inventor determined at all costs that his invention be developed for the benefit of humanity, yet anxious that that same humanity not steal it from him. As author Graille puts it, "His conceptions and attitude directly or indirectly conditioned the overall essence of this affair. Full of enthusiasm, from the very day he discovered that the ray he had developed had a curative effect on a cancerous cat, he developed a single-minded fixation on cancer One could understand and sympathize with him on this score. Here he is, a little Italian immigrant without money or means, and he is going to offer the world a cancer cure. He is so convinced that he wants to move ahead to doing just that. He will never understand or accept the exigencies of Science or Medicine. For him, experiments, controls, verifications and parallel research are a waste of precious time. 'I've made machines which cure cancer. Take them and treat cancer patients. Don't bother with the rest.' Such would be a summation of his point of view."

Through the efforts of persons kindly disposed to the inventor, this point of view was softened and he came finally to understand the- necessity for what has been called scientific rigor, on the other hand. another aspect of his character never changed an iota. This was his determination to preserve the secret of his invention, motivated first of all by his unshakeable desire that it be developed in Bordeaux, the city of his adoption, for the citizens of that city. Deeply rooted was his belief that if he made his secret public, the machine would be taken from the Bordeaux region and further developed by Parisians, those who considered themselves to be in the penthouse of the scientific edifice. Once this was accomplished, he would likely not have one more word to say about the matter. Therefore he continued jealously to conserve his secret and put confidence in nobody.

As Graille generously concedes, he may well have been right. and adds: "All his life he had to go up against men, whether scientists or industrialists, who had but one idea in their heads: to get to the bottom of the inventor's secret in order to build for themselves a machine which they then could exploit for their own account, for their own glory. Many such 'Priore Machines' were to be actually built more or less surreptitiously or clandestinely. Not one of them ever worked."

While one might easily accuse Priore of a limited view, the horizons of the researchers themselves were certainly not as broad us they might have been. Those involved in bio-medicine were content with the results produced by the machine, the workings of which were of no concern to them. A black box, as it were, emitted a ray that definitely affected experimental animals. At the same time, as researchers specifically interested in the cancer problem, they never gave a thought to what the Priore Ray might accomplish in the wider clinical domain of other afflictions.

As for the physicists, they were seemingly not up to the task of comprehending a complex radiation that had miraculously sprung, as from the head of Zeus, out of the intuition of a man they considered to be an undereducated and all but illiterate gadgeteer. Still others, whether physicists, biologists, doctors of medicine or specialists in a dozen other fields, were willing to throw the baby out even before it went into the bath water. In their eyes Priore was just a nobody.

Behind the scenes, many of these scientists resorted to using the press to achieve their own ends. Thus, the chief medical chronicler for Le Monde (the French New York Times), herself a doctor of medicine, was led to write outright lies about the Priore Affair—specifically and falsely stating that cancer patients had been treated with the Priore Ray in the clinic of Professor Lachapele in Bordeaux with not only negative, but disastrous, results.

On the other hand, a journalist for another leading Paris daily, Le Figaro, scrupulously conscious of his responsibility to fairly report what was going on, aptly wrote: "We would like to see at least one thorny point clarified as soon as possible. Several years ago P'rofessor Biraben of the Bordeaux Medical Faculty (who at that time had not become a ranking professor) was involved with the Priore device. According to certain reports from medical circles, his results seemed, even at that time, to have been already quite positive on small animals and he seems to have written a report to that effect, He was advised by highly placed authorities "to keep quiet" and stop talking about this affair. If this turns out to be true, it would be a veritable medical scandal to be judged in the harshest terms."

Could one have put it more succinctly?

The foregoing is to present something of the flavor of what was transpiring in the wide world far removed from the laboratory of

Priore who, at the time unaware of it, was revelling in the fact that his machine had been successfully used by high-ranking French cancerologists and its results reported in three separate notes to the Academy of Sciences.

His courage was also more than buoyed by the arrival on the scene of the commercial director of a large French industrial firm specializing in the intricacies of manufacturing glass components. This man had heard that Priore needed a large tube that was beyond all existing norms and perhaps did not exist anywhere in the world. This tube, it can be stated, contained a rare gas, neon, which when excited into a plasma, seemed somehow to convert the various electromagnetic inputs into a single Priore Ray which surged from the business end of the tube. In the tube were an anode and cathode. Peculiar to the anode was that it had to *rotate* to produce the desired biological effects and this is but one of the anomalies in Priore's equipment which physicists and bio-physicists have to this day been unable to explain.

The manufacturing company, a subsidiary of the internationally known company, Saint Gobain, was looking for a new product. The commercial director thought the new tube might fill the bill, particularly if it could be adapted to a machine that might ultimately cure cancer, a product that indubitably would have an enormous market across the world. There were plenty of problems with regard to the tube, notably those of its large dimension, its resistance and its conductability. When the tube was finally made, it now seemed that. Priore would have to explain his discoveries to the scientists of the company that had made it. One of these was sent to elicit such an explanation but was, so to speak, "shot down in flames" by Priore. So a second attempt was made by Ivan Peyches, a senior executive of the company, and president of the Society of Civil Engineers of France, who made a detailed investigation of the device. His reports were subsequently lost, but there remained an article he published in a leading French journal, Sciences and Technics, a short time before his death in 1978. It bore the intriguing title: "What Are So-Called Paranormal Phenomena?"

In it the engineer wrote: "There was so great an accumulation of components capable of having some kind of action, and being unable to work separately, that the results of measurement were limited to proving that there were no specific rays that issued from the tube (Priore talked about canal rays), no more than there were any X **rays.** On the other hand one could detect a magnetic field which was the end result of a field proper to the tube and of the magnetic field of a solenoid that constituted the experimental chamber, an electromagnetic field with a frequency of 16 megacycles (19 meters) and a high frequency field (metric waves), the whole being pulsed at a very low frequency of an order of one per second. It was impossible, in such an imbroglio, to determine what was necessary and what was sufficient. Priore maintained that the simultaneous action of his various generators was indispensible to achieving his effect."

Peyches then went on to relate how he tried to persuade Priore to offer a more precise definition of his thinking about the workings of his device. He wrote the inventor: "At this point, I would say that all reticence on your part, which in your eyes would be justified by the fear of seeing yourself partially dispossessed of your work, would be of far greater detriment to you than any safeguard of your interests. Moreover, since it has become a question involved in public health, you are no longer entirely your own boss...you absolutely must bring all this to the clear light of day and I don't believe you can do it alone...You must supply all the characteristics so that third parties can reproduce your results." Then, he concluded by citing the words of an academician: "Many phenomena are rejected by the scientific world because they are considered irrational: But it is not B proof of scientific honesty to refuse a priori to try, out of homage to truth, to have a look at them and perhaps to understand them. Will Science one day be able to abandon its taboos?"

It was Peyches' final conclusion that, in the end, Priore was a man of genius who knew absolutely nothing about what occurred in his machine from the scientific point of view. The company which he represented no longer exists since it was bought out by the American firm of Corning Glass.

Industrial interest in the Priore device was not limited to the Saint Gobain subsidiary. Next into the lists was a company in Anguouleme, Leroy-Somer, which specialized in electric motors, generators and later was to branch out into solar power. Its president, Georges Chavanes, took the initiative to write to Priore in 1965 that his company was interested in providing some of the complex electrical equipment needed by the inventor, more particularly high-powered generators, on the condition that Priore move his operation to the company seat at Anguouleme. When the inventor categorically refused, Chavanes tentatively agreed to build a factory to manufacture the Priore device in Bordeaux itself.

The alliance between Leroy-Somer and Priore, shaky at best, lasted two years and blew up on Holy Thursday of 1967. The period was a stormy one for both parties to the agreement. Priore did his best to convince Chavanes to commit himself to building a huge machine with a magnetic gauss strength of 10,000 gauss. In the end he got one that put out only 920 gauss, not much stronger than the machine he had already built which put out 620 gauss. Since the field of action increased with the gauss strength, Priore reasoned that a machine of literally behemoth size would be able to irradiate the whole, or every part, of a human cancer victim lying on a stretcher, whereas the smaller machines had been effective only for small animals or for treating a limited portion of the human body.

Chavanes and his company were aware that it would be a tremendous financial burden to contemplate building the larger machine. So they went ahead with plans for the smaller one while at, the same time putting great pressure on Priore himself to make him feel that he was the least important cog in a new gear, in fact that his status was reduced to being a simple employee of Leroy-Somer. In Graille's estimation, this lack of psychological finesse on Chavanes' part constituted what he called "the blackest pages in the Dossier Priore."

Even the smaller machine was to cost about half a million dollars, a price which today, due to inflation, could be tripled or quadrupled. During a stage in which an intermediate machine was designed by the chief Leroy-Somer engineer, Ribeau, a machine that never did function properly, Chavanes all but forced Priore, who was heavily in debt, to sign a contract which was falsified. The falsification was a matter of one word which was changed in the contract. In a phrase reading that an exclusive license of patents, and subsequent patent modifications, would accrue to the company "for all countries solely for therapy on cancers concerning animals and humans," a word was inserted by hand so that the phrase read. *"concerning particularly* animals and humans" implying that other uses of the machine, whatever they might turn out to be, would also accrue to the company. This one word change was amended on Priore's copy of the contract by calling the word "particularly" a "nullified word," but on Chavanes' copy it was called an "added word."

Leroy-Somer believed it was sufficiently well positioned in the driver's seat to be able to deal on behalf of Priore himself with the French governmental institutions, mainly the General Delegation, concerned with the funding of the new machine. When Priore learned of Chavanes' contact with the General Delegation he wrote a letter informing it that no one had the right to deal in his name. Nor did Chavanes even attempt to cut the Saint Gobain subsidiary, which alone could supply the tube, key to the device's functioning, in on the government funding.

In the meantime, no less a figure than Professor Kastler, soon to win the Nobel Prize in Physics, came down from Paris to inspect. the existing Priore device. He brought with him Delmon, who, we recall, had worked with Berlureau on the first animal experiments and who

NOW, it turned out, was trying to build his own version of a Priore device on the sly without telling Priore. Kastler's bringing Delmon with him to Priore's lab so angered Robert Courrier that he told the physicist he had committed a real gaffe. He also convinced Kastler that Leroy-Somer should build a machine with a power of at least 5000 gauss, but Chavanes refused. There seemed to be no harmony of outlook between the leading industrialist concerned, on the one hand, and the top physical and biological scientists on the other.

While all these, and many more, peripatetics were proceeding, Priore's sister in Italy came down with cancer. Beside himself with grief, Priore informed all concerned to commit themselves either to building an intermediate machine correctly, under his supervision or, better, the 5000-gauss machine, and to do this in time to save his sister, or he would wash his hands of the entire matter. Confronted with this ultimatum, the company began to work round the clock to perfect the intermediate machine but engineers involved, believing themselves to be more adroit with respect to its design than Priore, left out a host of what, to them, were unnecessary components. The result was that when the machine was first put to trial, most of the components burned out or otherwise failed, and the machine itself became a useless pile of rubble.

Shortly thereafter, Priore's sister died of cancer. Her grief-stricken brother went into what amounted to total isolation, unwilling to talk to a soul.

The whole Priore affair might have ended at that point, in the early part of 1967, were it not for the entry onto the scene of a key figure, Professor Raymond Pautrizel. Born on 3 June 1916 in Basseterre, capital of the French Carribean island of Guadeloupe, at forty years of age he was on the Faculty of Medicine at Bordeaux. He soon became known, world-wide, as the "father of parasitological immunity," a title he never accepted, saying that, if others had awarded it, it was simply because "he had searched through old scientific publications to find ideas that were as valid for modern research as they were forgotten by modern researchers." And he later was quoted as adding: "It is really too bad that researchers today don't pause from time to tome to dig into studies made by their predecessors, some of which were performed even decades ago!"

Professor Pautrizel was awarded the first academic chair in France for immunology, and later a special unit was created for him for parasitological immunology, a subject which is both simple and complex. The simple part involves the fact that various immunological techniques can be applied to diagnosing specific parasites that have invaded an organism in order to develop preventative actions against them via vaccines, or curative actions via serums.

When invaded by parasites, organisms react by creating antibodies, specific substances aimed at killing the invaders also known as antigens. These antibodies are liberated, like an attacking army, into the blood. Simple enough so far. The complexity arises because the defending army, the parasites, don't just lie down and die under the attack. They are capable of modifying their "personalities," as it were, and of changing various of their characteristics such that the mechanisms that the host uses to recognize, or detect, the invaders are invalidated. Thus, the substance which an organism would secrete to destroy an invader A becomes incapable of recognizing A, now become A-l, and therefore incapable of destroying it.

The organism at this point seems to realize it has to create a different substance to rid itself of its antagonist but, in the meantime, the metamorphized parasite is getting on with its assigned destructive task. Alternatively, the parasite has another capability: that of itself liberating substances which can annul or annihilate the organism's overall defense system. A sort of "in the blood" version of Star Wars is going on at the microscopic level.

The study and classification of the substances - - call them weapons - - emitted by parasite-attacked organisms allows for the establishment, in turn, of batteries of tests to define the exact nature of the parasites themselves in order to come up with an appropriate therapy or counter-weapon.

This then, is the essence of parisitological immunology, Raymond Pautrizel's area of research. He specialized on a particularly lethal parasite known as trypanosome, the scourge of tropical third-world countries where, in one form, it causes sleeping sickness in animals and humans, in another, equine syphillis, in still othera, other afflications. Over the years, during which he produced a small library of literature on the problem (known mostly to specialists in countries where that problem is acute), Pautrizel and his team discovered, among other things, that the trypanosome can modify itself, again and again, up to 101 times over a period as short as only three weeks.

Even before his work on trypanosomes, Pautrizel, back in 1949, was one of the first researchers to discover what is known as *ambivalence* in drugs, notably histamine. Histamine is a substance which is secreted by an organism as a defense mechanism but if oversecreted by certain cells circulating in the blood, it becomes virulently noxious, mainly by overdialating blood vessels, thus making them permeable to water and leading to edema and even death. This process occurs, for instance, in some human beings who are highly susceptible and over-reactive to bee or wasp stings.

Pautrizel's research on the noxious aspects of histamine led to his finding that the same substance, applied in requisite small doses, is extremely important to the defense system of the organism. Today he stresses the notion of *ambivalence* in many areas of his work and characterizes it as "a key to the biology of the day after tomorrow."

To finish with the background on Pautrizel, before bringing him on stage in the Priore drama, it may be added that only a few years ago, at a formal reception for him attended by the medical elite of France, he was given a Basque *makila,* an iron-bound honorific cane of sculptured wood, in tribute to his work. On it was the incised inscription: "Sometimes to heal, often to alleviate, always to console," an epithet that perfectly characterizes a medical doctor imbued with that kind of rare compassion that marked Pautrizel's character.

When Robert Courrier sent Madame Colonge to Bordeaux, it was Pautrizel whom he asked to provide her with every assistance. In this way, Pautrizel was first introduced to Priore and his device. After witnessing the results obtained with it he was to say: "What stupified me, and led me to ponder the question, was to see the control animals die from their tumors in 3 weeks, while at the same time I could observe that the tumors in the animals under treatment were literally melting away and the same animals were taken back to Courrier's lab at the College de France in Paris in perfect health." AS a result of his thinking about the problem, Pautrizel came to the belief that the machine, however it worked, did not exert any action at all to kill cancer cells but, through as yet unexplained mechanisms, stimulated the afflicted organisms to provide themselves with new immunological weapons that could overpower the cancer cells.

To shed light on this problem, Pautrizel proposed the simple expedient of experimenting, not on cancer-infested animals, but on *in vitro* cultures of cancer cells. He made this proposition to both French and British cancerologists but they were convinced that the Priore device *had to be actually killing cancer cells themselves,.* They could not see the point that, if the machine did not kill cancer cells, then it was doing something else to the body to allow it, and not the machine, to do that job.

Pautrizel's involvement with the British was the result of a team being sent from England to experiment with cancer mice with the Priore Ray. What happened cannot be related in this brief resume except to say that, out of a lack of understanding on the part of certain British cancer experts and malicious conniving on the part of one member of the cancer "aristocracy" in Paris, the experiments were put under a cloud. It was alleged that mice had been substituted somewhere during their long round-trip voyage between

England and Bordeaux to make it look as if a failed experiment had been successful. This did not prevent Sir Alexander Haddow, chief of the prestigious Chester Beatty Research Institute for Cancer from stating, at a meeting in Paris, that the Priore machine had been indubitably effective on the English mice and supporting Pautrizel's idea that experiments should forthwith be done to see if the Priore Ray had any effect on cancer cells *in vitro*. Haddow's suggestion backing Pautrizel's recommendation fell on deaf ears.

Because of the emotional turmoil and rancor with respect to cancer that had so long surrounded Priore and the workings of his machine, Pautrizel suggested that it be tried in a completely new area, one he knew so well, namely on afflictions caused by the trypanosomic pathogen. Before these could get underway, however, someone had to persuade the still desolate Priore to return to work. Pautrizel, known to those really concerned with and knowledgeable about the potential of the Italian's invention, at last was able to convince the inventor to cooperate and get back into harness. This ho did with that rare combination of diplomatic tact and warm human sympathy with which only the Pautrizels of this world are gifted.

In the meantime, Riviere had gone on to implant new 347 tumor grafts in rats previously cured of 347 tumors. When none of the tumors developed, that result added one more argument to back Pautrizel's idea that the machine was, in fact, affecting the immunological defense system of the animals. However, when Riviere tried the same procedure with the T-8 tumors, his animals died. This lead to the conclusion that the immunity acquired by the animals to lymphoblastic lymphosarcoma 347 was specific to that tumor. When a note on this research was sent, again through Courrier's good offices, to the Academy, for the first time, it strangely omitted from the listing of the participating researchers the name of Antoine Priore. It seemed that Riviere had been taken to task by fellow cancerologists who believed that Priore was nothing but a naive bumpkin or, worse, a swindler. They had warned him against publishing any papers with which Priore's name would be associated This rank injustice and lack of fair play again sent Priore into a fit of despondency and depression from which he could only be wilth drawn by those subtleties involved in Pautrizel's sympathetic and friendly counsel.

On 25 July 1966, another note was sent to the Academy filed for the first time not under the rubric Cancerology but under the rubric Immunology. It was entitled "Influence of Associated Electromagnetic and Magnetic Fields on the Immunity of Mice Infected with Trypanosoma equiperdum." The conclusion read: "The treatment allows the organism to rid itself of parasites even when these have invaded it in a most intensive way.... There is an enhancement of both the specific and aspecific factors of immunity."

Thus, for the first time, the field of research shifted from the narrower field of cancer to the much vaster domain of immunology. And, for the first time, Pautrizel's name appeared as the senior author on the paper. It also appeared that, for the first time, there should no longer be any problem about experimenting with the I'riore machine. Such was not the case.

Still complicating the whole issue was the fact that Priore himself was using different setting to produce different varieties of radiation depending upon his own intuitive evaluation of the particular biological experiments being run with his machine. He would never reveal the nature of these settings.

At this point there appeared on the scene a new researcher who became Pautrizel's loyal ally, a young woman, Pierette Chateau-Reynaud Duprat. During her work in Paris, she had learned of the Priore controversy, and, against the stern advice of mentors senior by many years to her in the cancer hierarchy, she came to Bordeaux to meet Pautrizel and learn more about the research.

Her work, performed over many years, is too detailed for presentation here but it led to important conclusions. One was that the Priore Ray had no direct effect on the trypanosomes themselves but stimulated and reinforced the defense mechanism of the infested organisms, allowing them to reject the parasitical influence with an effect so durable that they were no longer subject to this influence even after treatment stopped.

Another conclusion was even more important and involved, in part, British research. It pertained to the effects of the machine on both *allografts* or those made between two different individuals of the same species, and *isografts,* or those made between two different individuals ofthe same genetic line having in common antigens that were characterized by what is called the same histocompatability. The conclusion was that not only was the rejection of *allografts acce*lerated by the Priore Ray but that *isografts* were also rejected. This meant, in sum, that the ray stimulated not only the *defense* mechanisms of the organism but also, and more importantly, its re*cognition* mechanisms. In the case of an isograft, this allowed the recognition of weak antigens that were not recognized in non-irradiated animals. In other terms, where at first the anti-aircraft batteries could not shoot down the aircraft because they could not see them.now they could shoot them down because they could see them. In immunological terms, the ray affected both humoral and cellural, both specific and aspecific, immunity.

Here we must return to the mystery of the settings on the device.

As a result of the new experimentation it seemed that, depending on those very settings, the active ray, complex as it was, could have either similar, totally different, or diametrically opposed effects. Thus it was not a question of a ray having universal effects — a kind of magic bullet capable of killing any target — but of multiple radiations which, due to the complexities in Priore's personal makeup, have unfortunately yet to be sorted out and explained.

Thus, the machine originally designed by Priore, called the P-l when it put out a wave length of from 19-21 meters, had a radical effect on certain animal cancers, on *cellular* defense mechanisms, and finally, but not universally, on organisms infested with **Trypanosoma equiperum,** (hereinafter called T.e.).

A second machine, dubbed the P-2, was at first not able to produce these frequencies. What it did put out was a frequency of 17 meters that *was* universally effective against T.e. and seemed to act. not on the cellular, but the *humoral,* defense mechanisms. The rejection of grafts depends on the cellular defense mechanisms, which partially explains why Pautrizel when using the P-2 machine, selected the T.e. vector, as it is called in microbiology, just because this creature is fought by the organism's humoral defense system.

Consequently, the bio-effects that were successfully attained depend on the varying, not to say quixotic, nature of the radiation At one point Pautrizel actually did experiments on animals infected with *plasmodia* — the vector for malaria which attacks red cells - and found that the settings used were effective while never learning exactly what they were or the exact nature of the radiation. Furthermore, Priore himself maintained that over the years he had successfully treated cases of human tuberculosis but, again, never revealed which frequencies had been used to achieve this.

Several more notes were sent to the academy on the successful work performed with the Priore Ray on animals affected with T.e. But the central issue remained: how to find out exactly how the-machine worked. It fell, not to civilian scientists, but to those in the-French army service to attempt, at this point, to work out the problem. The army service brought into the picture was the DRME (an acronym which translates as Administration for Research **and** Test Methods), to which Pautrizel had sent a request for funds in 1968.

This request was the subject of a meeting at which were present three of the top names in French science, one representing biology, the second physics and the third, medical physics. The latter two turned in extremely unfavorable reports recommending that no money be wasted on the problem. The biologist, however, turned in a most favorable report and, despite the fact that he was in the minority, his opinion won the day.

As remarkable as was this victory, it was even more stunning and incredible given the fact that this biologist was the same Andre Lwoff who had so adamantly opposed the Priore research a couple of years previously. Lwoff had summoned the courage to completely reverse himself only after he sent one of his most trusted colleagues to do secret experiments with the Priore Ray on mice injected with peroxydase (an antigenic solution) to see if they would produce a higher level of antibodies than non-irradiated animals. This they did so well that Lwoff became convinced that the Priore Ray caused an extremely important increase in immune reactions. These results were never published because, before the experiments could be repeated to be absolutely sure of their results, the machine suffered one of its many interminable breakdowns.

The DRME report was at length, and in length, issued but not publicly since it was protected by a military classification. However, a synthesis of it was finally published in November 1979 by Herbert Gossot, Secretary General for the French Association for Bioelectromagnetism, under the title: "A Scientific Balance Sheet on the Priore Ray." Its contents were as follows:

"The two physicists assigned by the army made a complete analysis of the electromagnetic radiations and magnetic fields activated by the Priore device. They thus determined the spectrum of frequencies which the device emitted. They showed particularly that frequencies in the visible light and infrared range had no biological effect; that there were no X-rays or Y-rays; and that the pulsed ultra-high frequency electromagnetic wave was modulated in amplitude to that of a high-frequency wave. They did a topographic survey of the respective intensities of the various magnetic and electromagnetic fields in the experimental plane of the device. In particular, they determined the spatial repartition in this plane of the density of the strength of the ultra-high frequency wave. They showed that its value was very weak and that it could not produce any kind of overall significant thermal effect imputable to the hyperfrequency ray.

Finally, and most importantly, by using what they had learned about these repartitions, they demonstrated a clear correlation between the biological effects obtained and the intensity of the hyperfrequency ray. What they actually observed was that, on the **biological** model used, i.e. experimental trypanosomosis of the mouse, there was a diminution of the rate of evolution of the parasitemia that was proportional to the strength of the hyperfrequency wave. To quote them: *'These experiments of correlation are of certain interest: they confirm, if there is still any need of so doing, the* *biological efficacy ofthis device."*

The two physicists, Bottreau and Berteaus, are still interested in rebuilding a Priore device with which additional biological research could go forward. At the same time they suggested to administrative bodies in French science the creation of a special laboratory for bioelectromagnetism to fund more work, a suggestion in which Professor Pautrizel concurred. No action was taken and their report was kept under wraps. In a note they presented to the Academy of Sciences on their investigation, they were not allowed to include the names ofthe laboratories where they worked: in the case of one, the CNRS Magnetic Laboratory at Bellevue near Paris, and ofthe other, the Laboratory of Ultra-Hertzian Optics and Talence near Bordeaux. Why? Because the directors of these laboratories did not want any mud in the Priore affair to be spattered on them.

The next experiment done by Pautrizel was on rabbits whose testicles had been so seriously affected by trypanosomes as to be almost entirely destroyed. After radiation the same testicles took on their normal histological appearance and the rabbits, able to procreate again, in no way abstained from their newly regained ability. **This implied the complete regeneration of an organ that had all but completely degenerated.**

Yetjournalists, who sought out truths about the Priore affair in Paris from high officials they believed would know best about what was going on, continued to be led astray. For example, an American scientific reporter, writing in the **Saturday Review** of **Science** in 1973 saw fit to state: "It is really a question of a mystical problem that has little to do with science." He was quoting Professor Bader, a man who for 15 years held top administrative posts in science that could have allowed him to back the Priore research with all the funding necessary to its accomplishment. At the time Graille's book came out, Bader issued a book of his own about the Priore affair which offers no real idea ofwhat was involved. When I asked several people in France why Bader had written the book, they were unaware of Bader's inmost motivation.

Machinations continued to swirl about the case over the next several years. Behind-the-scenes intrigues, distorted accounts in the press, lethargic attitudes on the part of administrative officials who would not take responsibility to cut an increasingly tight Gordian knot, outright fear of various personalities to become too deeply implicated lest they lose their jobs - - all these, and more, continued their daily round in an atmosphere of "Business AS Usual," and "Don't Risk Your Neck."

To get to the nexus of the situation, we have but to cite the observation of one of the few perspicacious journalists who, in **the** prestigious scientific monthly, Sciences and Life, wrote: "The physicists are convinced that the effective Priore Ray is very complex but to analyze this further some things first have to be made clear. One is to raise the suspicion that has surrounded Monsieur Priore with a fabulous acretion of misunderstandings, insults and accusations of being a swindler over many years. What is needed is a veritable national effort to act effectively and to act rapidly."

Over the next two years the decision-making process of the French government lumbered its way along until it was finally decided to back the construction of a powerful machine. This decision was not favorably accepted in many quarters. As Le Monde would comment: "The decision was made in spite of the disapproval of many scientists. When money is tight, one should pay particular attention to how it is being spent. Such seems not always to be the case. A credit of some $3.5 million francs (or about a million dollars) has just been accorded to finance the construction of a new Priore machine."

The scientists to whom the article referred were in a rage. They understood, at this juncture, that the only way to put an end to the affair was to eliminate Pautrizel who, because of the very success he was having with his research, was seen as a dangerous competitor that might even become one of the top figures in medicine and science on a national, or perhaps, on a world scale. Indeed, it was learned that Professor Courrier had gone to the length of sending a report on Pautrizel's behalf to the Nobel Committee in 1979.

To make a long story short, the large powerful machine, the M-600, was built but a huge tube in it, after functioning for about a week, exploded. Due to the galloping inflation of the 1970's, to replace it would have cost another million dollars. The money was not forthcoming.

In the meantime Pautrizel, ever experimenting with the still functioning smaller machine, was to discover new facts. Mice with their spleens cut out, for example, also could survive injections of T.e. The Priore Ray had important implications for Arterio-sclerosis, since it effected lipid modifications in rabbits given a dietary regime high in cholesterol. This research, published in another note in the Academy Proceedings, instead of being warmly received, only irritated the cardiological fraternity which felt, as some of its members put it, "trapped" by Pautrizel's efforts.

One particularly virulent opponent was Professor Bricault, Dean of the Bordeaux Medical School who, as late as 1980, was telling his own students that the published results were a farce and had never been obtained. The students, who carried out a special investigation of the matter on their own, were able to judge what a farce their own medical dean might represent.

L'Express, the **Time** magazine of France, read by at least half the population of French intellectuals, had the gall to compare the results of the Priore research to those of the infamous Trofim Lysenko of the Russia of Stalin's day. Haughtily **L'Express** added: "Today Priore's defenders explain that his machine has not only cured cancer but, in all probability, altered the immunological characteristics of mice. Were this, in fact, so, all the immunologists, all the geneticists of the world would unite to affirm that a machine capable of changing the genetic patrimony is the discovery of the century, far more important than the atomic bomb or the conquest of the moon. Unfortunately, the history of the whole thing has never been properly elucidated." The article was illustrated with photos distortedly selected to convince viewers that the Priore machine was as serious and effective as the one that purportedly brought Frankenstein to life.

In this poisonous atmosphere the slow work of building the M-600 went forward. To give anyone who was not there a feeling for this endeavor we may now cite verbatim a passage from Graille's book: "The construction and assembly of the prototype —the M-600, that of highest power and variable parameters — were fraught with many uncertainties and delays on the one hand and, on the other, were marked by the stamp of Antoine Priore's sparkling genius.

"To go from an apparatus that developed 1,240 gauss applied over an effective area of some 20 centimeters, to one developing 5000 gauss over an area of 60 centimeters means to take on an extremely risky technical and technological wager. Electrical, mechanical and glass-blowing specialists plunged into the unknown. They had to conceive, make, adapt and put together all the various myriad components almost haphazardly with no precise technical study being previously available. Priore's stubbornness forced them to take on a trial-and-error manufacturing "gimmickry" without precedent. As the thing was put together and preliminary tests made, it became clear that many of the components were unsuitable and that they would have to be modified or replaced. The tube itself, made of pyrex, 60 centimeters in diameter, and 6 meters tall, had to be replaced twice after it imploded. In fact, practically everything had to be reconsidered or readapted. "Everything" meant the parts going to make up a generator of 50 tons in weight. For example, the coil which created the magnetic field: 5.5 tons with 11 miles of copper wire. For example, the numerous cooling circuits which stabilized the thermal equilibrium of the generator and its environment or, additionally, the circuits governing command, control regulation and selection — 6 tons of electrical cables of which 15 miles were of tele-command wiring.

"Priore astonished everyone. Breakdown after breakdown, incident after incident, it was he alone who showed what to do next, indicated the proper steps to take, the right settings to adopt, the right way to assemble the components: He was virtually building his machine by himself, nursing its construction along day after day, all the engineers' studies and efforts actually, and ultimately, serving only as a preliminary attempt, a sketch as it were. When Priore made his presence felt, things began working."

Then.... after the machine was built: "The part of the entire apparatus to generate electricity was set up on a provisional basis. It was so noisy that, while functioning, it woke up the whole neighborhood. The number of experiments had therefore to be curtailed so that the machine would not be used at night. And, all at once, everything came to a halt. The Faraday cage, shielding and isolating Priore's apparatus, was torn and fissured by the shock of the cement pilings that were being sunk into the ground all around to hold up the building under construction. This allowed high-frequency waves to escape which disturbed radio broadcasts emitted by local radio stations, the army, and civilian aircraft for miles around."

Nevertheless during the week or ten days that the machine was in good operation the results of experiments performed with it were more than formidable. First of all, it allowed for as many as forty experiments to be performed on some 280 animals in a remarkably short period of time. Among the discoveries made were: The ray emitted provided the treated animals with an extremely strong immunitary response. Animals whose immune defenses had been attentuated by an immuno-depressant were able to overcome the effects of injected parasites but relapsed a few days later. One could therefore conclude their immune response was much weaker than those normally infested and treated.

Newborn animals, whether treated or not, developed a marked parasitemia leading to their deaths. At the time of death, the parasites had the same antigenic structure as those ofthe innoculum which thus implied that they had met with no defense at all in the infected organisms. This also proved that the Priore Ray did not act directly on the parasites themselves but only by way of an increase in the immune defense system of the organisms. The newborn animals succumbed to their parasitemia because their immune sytem was not yet sufficiently developed to be stimulated by the P-Ray. The phenomenon of a stimulation ofthe immune defenses was demonstrated by the fact that animals which had received soluble antigens developed, after being irradiated, a level of antibodies far superior to the controls.

These and other conclusions were the object of notes presented to the Academy of Sciences by Pautrizel and his team in 1978. Even before, at a colloquium held in Antwerp, Belgium devoted to African human trypanosomiasis, the same team had offered the conclusion that the stimulation of the immune defense system that allowed organisms to throw off the effects of trypanosomiasis had to be very significant in that all attempts to try to effect such stimulation through immuno-stimulants as well known as B.C.G., or Corynebacterium granulosum, had led neither to the cure produced by the Priore Ray, nor to any prolongation of the infected animals' lives, nor even to the slightest modification in the evolution of the Trypanosomiasis.

These three scientific papers did little for the cancerologists who read them except to exacerbate their urge to oppose the Priore research, if not to arouse their outright hatred for the principal experimenter, Raymond Pautrizel. Could this have been because, for over 20 years, the same cancerologists had been working in vain to provoke in cancerous organisms immuno-stimulative reactions by intensively and successively vaccinating them with B.C.G.? Many others had been life-long apostles of chemotherapeutic cocktails of all sorts, or life-destroying ionizing radiations, or, what more recently has become the fashion, of applying the two methods in endless combination.

For this reason, they saw Priore and Pautrizel as nothing more than spoil-sports who had to be destroyed.

One of the opening shots in this campaign was a letter received by Pautrizel to inform him that his request for funds to continue his research through Unit-89, a unit that had been specially set up for him to direct, had been denied. It took many months of investigation for Pautrizel to learn that the real reason for the refusal was because of his work with Priore.

Next Pautrizel was informed that his appointment as director of the same research unit would be extended for only two years, whereas the normal extension for similar units was five years. A third insult came when Pautrizel tried to win a post within his unit for a high-ranking military physician, who had been his student and who had decided to quit the military in order to participate in the fascinating research prosecuted by his mentor. Pautrizel's request for funds to pay this physician, who all his life had been working on tropical medicine closely associated with problems of trypanosomia sis, were refused four times in a row with no cogent reason given. The physician, who in the meantime had volunteered his time without pay, finally became so emotionally overwrought that he gave up his medical career and retired to the countryside where he gave himself over to alcohol. Then Pautrizel tried to get a salaried post for another of his brilliant collaborators (who still works with him). He was told that this man could take up his new functions only if he left Bordeaux. One could go on with many other shocking stories but we will leave it to Graille to conclude: "Everything possible was done to isolate Pautrizel, to separate him from his collaborators. Every single one of these collaborators saw their careers put in jeopardy, compromised, or broken."

As a final insult, when the time came again to renew Pautrizel as director of Unit 89, those responsible, not daring to overstep what even they knew to be decent limits by not extending him, simply abolished the unit. And to add injury to that insult, a doctoral thesis that had now been prepared by Priore, and backed not only by Pautrizel but by Nobel Laureate Andre Lwoff himself, was summarily refused by the President of the University of Bordeaux.

It is perhaps unnecessary to state that the details behind all of this skulduggery could, and did, fill up two chapters of a book and make for the most heart-rending reading imaginable.

So what happened next? In the autumn of 1977, Professor Georges Dubourg, one of the leading lights in Bordeaux's company of surgeons and a friend and admirer of Pautrizel's, came to him to say openly and baldly: "My friend, at the point you've reached, there is only one more way to jolt medical opinion and that is to treat human cancer patients." Pautrizel was hesitant, believing his role to he one of continuing with his animal experiments but where would the funds for that come from now? He therefore asked his old mentor, Robert Courrier's advice. Courrier gave the green light. The treatments were restricted to terminal cancer patients whose immune defense systems had been disastrously weakened by chemotherapy or radiation or both. At least one of them was totally cured. The other lived, without pain, for a period many times longer than predicted by prognosis. Dubourg, Pautrizel and their collaborators wrote up the results and sent them as an official communication to the French Academy of Medicine for publication.

The reply they received from that Academy's perpetual secretary reads: "Experts whom we consulted consider that your work does not fall within the jurisdiction of our members and that it would doubtless find an audience more worthy of its purpose in a more specialized society."

To which Pautrizel formally replied: "Since two of the four signatories of our note are corresponding members of your Academy, could we not benefit from the remarks and comments made by the committee which saw fit to refuse our paper? And even, if this is not too indiscreet a request, to learn the names of the expert members who were consulted which would allow us to get into contact with them directly and to benefit from their singular competence?"

His letter has remained unanswered for four years.

There was nothing more to do except one thing which Raymond Pautrizel, as a man of science, had always been careful to avoid: Get a responsible journalist interested in the case, inform him of all possible details, and let him carry the Priore Affair in all its harrowingly loathsome aspects to the broad reading public. That journalist was Jean-Michel Graille.

For four years, Graille went about his task, publishing three consecutive long articles in his newspaper **Sud-Ouest France** and finally the book to which we have referred and of which this presentation is largely a resume. As early as 1980, Graille would write in his newspaper: "The Priore Affair is simple in essence. It can be reduced to a simple alternative: either the machine developed by Antoine Priore is of no interest and, having shown this, the affair can be considered at an end. Or else the machine is of real and demonstrable medical interest and, if that is officially recognized, he would be allowed to get on with the work. For this dilemma runs the risk, yet again, of being buried under delays and evasions. Beyond all the powers-that-be that have been directly connected to the affair for many years now — the power of finance, the power of medicine, the power of science — perhaps it is now political power with which responsibility lies if it can rise to meet and assume that responsibility through decision."

That was Graille's statement in 1980. His book which came out four years later ends with the sentence: "The Dossier Priore thus depends, from here on out, on a decision that must be taken on the very highest level, and imperatively. This responsibility devolves, in last resort, on the chief of state and on him alone. Will he assume it?"

Would the President of the United States?

Note added by T.E.B.: Antoine Priore is now dead. His machine has been dismantled. The iron dogma and hatred of electromagnetic medicine by bureaucratic science may have doomed hundreds of millions of humans - - whom Priore's device could have saved - - to bitter, agonizing, and unnecessary deaths. Hitler, Stalin, and Mao combined were not responsible for the deaths of so many.

APPENDIX II

PATENTS

English Translation of the French Patent No. 1,342,772 of Antoine Priore, 7 Oct. 1963.

Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of living cells.

Brevet d'Invention. P.V. no. 899.414 no. 1.342.772 Classification Internationale: A61 k-H 05 g

Procedure and assemblage for production of radiation especially serviceable for the treatment of living cells.

Antoine Priore
Requested 1 June 1962, 14.52 hours, Paris
Released by decree (arrete) of 7 Oct 1963

The invention deals in a general manner with radiation capable of penetrating matter. More exactly, it aims at a procedure and an apparatus making it possible to obtain a combination of radiations of different types able to penetrate matter and especially to penetrate intimately living organic tissues in order to produce in them certain effects, particularly in human tissues with a view to a therapeutic effect without destroying essential elements such as the enzymes.

In conformity with the present invention, one emits in a cavity a stream (rayonnement) of electrically charged particles upon which one superposes electromagnetic radiation of the centimeter wavelength range, the wavelength of which is preferably between 3 cm and 80 cm, and one directs the resulting radiation emerging from the cavity on to the object to be irradiated.

This applicant has shown that the penetration and, in particular, the curative effects are very distinctly improved when one gives the electromagnetic radiation a frequency determined as a function of the organ or the tissue to be penetrated or to be treated. For example, a wavelength of 14 cm is suitable for the liver and a wavelength of 19.5 cm for the spleen.

Preferably, the stream of charged particles is accelerated in a particle accelerator in such a way as to increase the force of penetration.

The resulting radiation is advantageously applied and directed upon the target, that is to say, upon the tissue to be penetrated, by the intermediary of a tube which is the site of electric fields and of magnetic fields for acceleration and control, the said radiation being preferably guided and/or reflected by a rotary deflector placed in the tube.

It is often advantageous to modulate or impose rhythm on this stream of particles by means of variable magnetic and/or electric fields so as to augment still more the force of penetration. This rhythm is preferably consistent, especially in medical applications, with the intrinsic period of the tissue to be penetrated or of the neighboring tissues, for example, muscle. These intrinsic periods arc well-known in medicine and are applied, particularly, for diathermy: they are situated in the wavelength range from 1 m to 50 m and more especially from 1 m to 18 m.

Preferably, one arranges to modulate the emission of radiation, the accelerating magnetic and electric fields, and perhaps also the rotatory deflector, to the cardiac rhythm of the subject.

It seems that the result obtained by the invention in the treatment of maladies of living cells (vegetable or animal) are due to certain phenomena which will be described, it being understood that this exposition will not circumscribe the invention.

As a function of its electro-physico-chemical constitution, the cellular pair nucleus-protoplasm is endowed with electric conductivity in direct relationship with ion exchange processes provoked by metabolic phenomena. One finds in tissues the presence of an accumulation of electricity at different potentials according to the different cellular densities of the tissues.

The work of Renshaw, Forbes, Morison, Amassian, de Vito, Ruset, Albe-Fessard, Tauc, Adrian, etc. has shown with the aid of micro-electrodes the existence of slowly oscillating elementary electric activity in the interior of cells; it can be thought that the *rythmeur* (or pace-maker) is achieved by the oscillatory electromagnetic system comprising the cell nucleus. This nucleus, in effect, is made up essentially of tubular filaments of insulating material (related to chitin) containing in its interior an electrically conducting saline liquid, and these filaments, coiled upon themselves, can be considered to constitute real little oscillatory circuits.

The recent work of Warson (sic) in America, as well as that of French scientists, including a communication from Polonsky, Don zon and Sadron presented to the **Academie des Sciences** by Prof. Francis Perrin on 16 May 1960 **(Rec. comptes rend, heb.,** 250 No. 20, 3414-3416) making it clear that experimental samples of **solid** DNA manifest properties analogous to those known in ferroelectric materials, makes plausible the hypothesis that a potential difference may exist between nucleus and periphery of cells. Certain recent theories go even further and liken the cell to an electronic receiver-emitter device normally functioning in harmony with the ambient media. The oscillatory system of demand waves, constituted by the cell nucleus, would behave in accordance with the laws governing semi-conducting materials.

The applicant is led to the conviction that in a normal state of physico-electric equilibrium, the cell nucleus is positively charged, but can acquire a negative surcharge following phenomena analogous to polarization.

The invention, especially, enables organs afflicted by this inversion of their electric potential, particularly in the case of the pathologic negative surcharge of cancerous nuclei, to recover their former equilibrium.

The following description in regard to the attached drawing, given as a non-restricting example, will make it possible to understand how the invention can be realized, the details which emerge both from the text and the drawing being, of course, part of the said invention:

Fig. 1 is a schematic section showing an apparatus for production and emission of an electromagnetic field combined according to the invention.

Fig. 2 is a frontal elevation of the cathode, taken from the right of Fig. 1.

Fig. 3 is a section through III- III in Fig. 1.

Fig. 4 is the overall scheme of the electric supply.

Fig. 5 is a view analogous to Fig. 1 showing another mode of implementation.

Fig. 6 is a section through VI-VI in Fig. 5.

Fig. 7 represents schematically an apparatus for pulsing the electric current.

Fig. 8 is the circuit of an amplifier permitting activation of the apparatus of Fig. 7 at the cardiac frequency.

Fig. 9 is the circuit of an oscillator permitting modulation of the electric current according to a wavelength between 1 m and 18 m.

The assemblage of Fig. 1 contains an apparatus 1 emitting electrically charged particles 2 in a cavity or passage 3, a cyclotron 4 accelerating the particles 2 and sending them into a cavity 5 forming a tube into which merges another cavity 6 acting as waveguide for electromagnetic radiation of frequency in the centimeter range emitted by a magnetron 7. The cavity 8 formed by the joining of tube 5 and waveguide 6 leads into a tube 9 in which the resulting radiation is accelerated and aligned. The interior of the cavity-formed by the assemblage of elements 1, 3, 5, 6, 8, and 9 contains argon at a pressure of 2 mm mercury.

The particle emitter 1 consists of an electron gun comprising a plate 10 and a cathode 11.

The cathode 11 is of molybdenum and has the very special form represented in Figs. 1 and 2. It consists of a rim 11a connected by two aligned spokes 11b to a hub 11c pierced by a hole 11d with its axis along XX'. The rim 11a is in two parts (as one sees in Fig. 1) which may be held together by screws (for example) forming a cavity of revolution 11e traversed by a number of holes 11f parallel to the axis XX' and regularly spaced. The filament for heating, 12, situated in the interior of the cavity 11e and is connected to the power supply conductors 12a.

The best results are obtained with a cathode 11 of molybdenum. The applicant has obtained satisfactory, but slightly inferior, results with tungsten cathodes. It turns out that molybdenum, and to a lesser degree tungsten, are the metals whose valency is closest to the mean valency of the chemical molecules constituting living tissues and more particularly those of human beings. One might well seek to use this fact to explain scientifically the phenomena involved, but it is understood that *the invention is not limited by any scientific explanation.* Concerning, on the other hand, the gas present in the apparatus at low pressure, the best results have been obtained with argon; the applicant has also obtained satisfactory, but slightly inferior, results with other gases of the rare gas series.

Surrounding the tube which constitutes the electromagnetic chamber are arranged: an electromagnet 13 with its winding 13a, placed at the level of the accelerating coils 14 and 15. Other accelerating coils 14,15 and 16 are similarly dispersed around the cavities 3, 5 and 8.

The two semi-circular boxes or "dees" 4a of the cyclotron 4 are placed in the usual manner between the poles of the frame around which are wound the accelerating coils 4b and 4c.

The magnetron 7 is of a familiar type and must be capable of emitting in the cavity 3 a centimeter wave of adjustable wavelength from 3 cm to 80 cm.

The lower portion of the tube 9, for acceleration and alignment, contains a cathode 17 resembling the cathode 11, with a filament 17a. This cathode 17 is supported by a hollow pillar 18 pierced by holes 18a close to its junction with the bottom of tube 9. This pillar 18 communicates with a tube 18b emerging on the axis of a rotary deflector 19 carrying at each end a "crown" of graphite plates 19a inclined at 45° to the vertical. The rotating axis 19b of the rotary deflector is mounted in a support 20 fixed to the interior of tube 19 and carries at its upper extremity magnetic bobs 19c which ensure that it will be set in motion in cooperation with the magnetic bobs 21a mounted on the shaft 21b of a motor 21. The lower extremity of the rotary deflector 19 is composed of a piece of molybdenum or of tungsten 19d in the form of a pyramid whose apex is opposite the open end of the tube 9.

The hollow base 18 and the tube 18b can be of a borosilicate glass of low coefficient of expansion such as that sold under the trademark "Pyrex." They may also be of quartz. Tube 9 itself can be of "Pyrex" as above or of another glass of the quality currently used for the manufacture of electronic tubes, but its bottom 9a, which is traversed by the radiation, is advantageously made of quartz.

The duct 8 joins the tube 9 by way of several tubulures such as 8a and 8b directed in vertical planes towards the plates 19a at a certain angle, which is advantageously about 22.5°. An electromagnet 23 analogous to the electromagnet 13 of the emitter tube 1 is placed around the cathode 7. Similarly, accelerator windings 24 are disposed around the tube 9. This tube also carries, at positions indicated in the drawing, three electrodes 25, 25a, and 25b surrounded by windings 26, 26a, and 26b, respectively. The drawing shows the supply lines, 17b, 17c of the cathode and its filament and that of the plate, 22a.

The basic plan of the electric supply is represented in Fig. 4. The part 27 feeds an initial branch with low voltage alternating current: this consists of a rectifier 28 (e.g., a Kenotron) whose rectified current is modulated at a frequency variable between 30 and 120 pulses per minute by means of a resistance 29, the control apparatus lor which will be described with reference to Figs. 7 and 8. The current so modulated is passed through the electromagnets 13 and *23* in such a manner as to generate, normal to the cathodes 11 and 17, a modulated field of 10,000 to 20,000 gauss.

The part 27 also feeds a variometer (interrupter) 30 which modulates the current of this part at a variable frequency 30 to 120 pulses per minute, the current pulsed in this way serving to feed the remainder of the installation, to wit: The magnetron 7; a converter 31 whose excitation is modulated at a variable frequency *300 to 900 Hz,* yielding a doubly modulated current (first at 30-120 pulses per minute, then at 300-900 Hz) which feeds the coils 15, 16 and 26; another converter 32 producing a low voltage direct current modulated at 30-120 pulses per minute by the variometer 30. This current feeds the motor 21 as well as the motors driving the variometer 30 and the apparatus controlling the resistance 29.

The current produced by the converter 32 also feeds a voltage step-up apparatus 33 consisting of a vibrator followed by a transformer and a rectifier, and producing a direct current varied at 30 to 120 pulses per minute imposed by the variometer 30. The maximum value of the voltage produced by the apparatus 33 is, for instance, 300,000 V, but this value may vary up or down, depending on the power one wishes to operate with.

The current produced by the voltage step-up apparatus 33 feeds the coils 4b of the cyclotron and 24 of tube 9, as well as a rheostat 34 permitting regulation of voltage to the desired value between 5000 V and 70,000 V. This voltage is applied to an oscillatory circuit 35 which produces oscillations at a frequency variable at will of wavelength between 1 m and 18 m. The current available to the output terminals 35a, 35b of this oscillating circuit 35 is thus a high tension current modulated first at 30 to 120 pulses per minute (by the variometer 30) and secondly at a wavelength 1-18 m. This current feeds the coils 4c and 14; the electrodes, 25a and 25b are connected to terminals 35a and 35b, respectively, and electrode 25 is connected to the mid-point 35c.

Cathodes 11 and 17, the "dees" of the cyclotron, and the plates 10 and 22, not shown in the diagram of Fig. 4, are connected to the output of the step-up assembly 33, while the heating current for the filaments 11e and 17a is furnished by the resistance 29.

To operate the apparatus, one adjusts the controls of the resistance 29 and the variometer to the desired rhythm; in medical applications, this is advantageously the subject's cardiac rhythm: This rhythm is thus imposed upon the whole installation. The cathode 13 emits toward the left a stream of positively charged particles 2, which are concentrated by the electromagnet 13 and accelerated by the coils 14, 15 and 16 and by the cyclotron 4. Superimposed on this particle stream in the duct 8 is the electromagnetic radiation from the magnetron 7, which is adjusted to the wavelength found by experience to be most appropriate to the cells which are to be penetrated, e.g., 14 cm for the liver and 19.5 cm for the spleen. The resulting radiation is deflected, directed and accelerated in tube 9 and is directed by way of the base of this tube toward the target to be penetrated.

It must be noticed that the magnetic field of the coils 15,16 and 26 is modulated, by means of the converter assembly 31, at a frequency adjustable *between 300 and 900 Hz.* This modulation has the effect of concentrating the particles, that is to say to detach them from the walls of the tubes, and it also permits an appreciable saving in weight of the iron cores of the coils. One chooses the highest frequencies (i.e., around 900 Hz) when one wishes to produce hard radiation at the exit of tube 9, and the lower frequencies for soft radiation.

The unidirectional magnetic fields of the coils 4c of the magnetron (sic) (cyclotron?) and the accelerator coils 14 as well as the electric field of electrodes 25, 25a, 25b, are modulated by the oscillatory circuit 35 at a chosen wavelength between 1 m and 18 m. In medical applications notably one selects the wavelength that best suits the organ to be treated or its surroundings, such as muscle. As already indicated, experience with diathermy makes it possible to determine the most suitable wavelength.

It must be noted that the resulting radiation already possesses, in tube 8 (Fig. 1) a considerable penetrating force. One could therefore use the assemblage described by omitting tube 9 and terminating the cavity at the end of tube 8 by means of a glass or quartz base, the resulting radiation being accelerated and directed, for example, immediately upstream from the base, by a final coil (not shown) surrounding tube 8. However, tube 9 appreciably improves the results.

Figs. 5 and 6 represent another mode of realizing the assemblage in accordance with the invention, in which the elements playing the same role are indicated by the same signs as in Figs. 1 and 3, modified by primes.

The arrangement of the connecting ducts of Fig. 5 in relation to tubes 1' and 9', to the magnetron 7' and the cyclotron 4', is different from that of Fig. 1 and has been used successfully by the applicant. The waveguide 6' of the magnetron 7' opens into the end of tube 1' and the duct 3', carrying the resulting radiation, divides into two branches: Branch 36, which conducts the radiation directly to tube 9', and branch 37 which conducts it to the cyclotron 4'. This blocks the electromagnetic radiation and accelerates the stream of particles which is passed by way of duct 38 to the tube 9'.

This arrangement can be used with a particle emitter and an accelerator and director tube similar to tubes 1 and 9 of the preceding figures. However, tubes 1' and 9' of Figs. 5 and 6 are constructed in a different manner with regard to their cathodes and anodes.

Tube 1' contains a first electrode 11' exactly like that of the cathode 11 of tube 1, and an identical second electrode 39 furnished with a filament 39a. Tube 9' (Fig. 6) contains in its lower part a first electrode 17' with filament 17'a and an identical second electrode 40 with its filament 40a.

In normal operation, i.e., to produce radiation identical to that described in connection with Figs. 1 to 4, electrode 11' serves as cathode and electrode 39, given a positive potential, plays the role of the plate 10 in Fig. 1, the filament 39a being unheated. Electrode 40 and its filament 40a are disconnected; cathode 17' and plate 22' are supplied as in Fig. 3.

To obtain particularly penetrating radiation, the polarities are reversed: Electrode 11' becomes an anode and its filament 11'e is disconnected, while electrode 39 receives the cathode supply and its filament 39a is heated; electrode 17' (with filament 17'a disconnected) and electrode 22' become anodes, while electrode 40 serves as cathode and its filament 40a is heated. For example, one can establish a potential difference of 250,000 V between 40 and 17', and 50,000 V between 40 and 22'. It is understood that in these conditions the cathode 39 emits to the left a stream of electrons which is concentrated, modulated and accelerated by the various coils and in the cyclotron, the polarities of which must of course be established in the proper direction. This stream of electrons is combined with the centimeter radiation emitted by the magnetron 7', resulting in tube 9' in the emission of very hard x-rays, modulated at the chosen frequencies, combined with the centimeter radiation of the desired frequency.

The assembly shown in Figs. 5 and 6 thus permits one to obtain at will either very hard x-rays or the radiation described in connection with the preceding diagrams.

The following description, referring to Figs. 7 and 9, relates to certain details of the apparatus used for modulating the electric current.

Fig. 7 represents schematically the apparatus for control of the resistance 29 and the variometer 30. The variable resistance 29 consists of a graphite helix 29a immersed in a conducting liquid **29b;** another electrode 29c, also of graphite, partly immersed in the liquid, is set into up-and-down oscillations by a connecting rod 41a linked to a fly-wheel 41. The fly-wheel is set in rotation, through the intermediary of a worm transmission 41b, by an axle 42 which can be driven, thanks to a double clutch 42a, 42b, either by a motor 43, or by the shaft 30a of the variometer 30. The variometer 30 is driven by a motor 44 by way of the worm transmission 44a.

If the fly-wheel 41 is driven by the motor 43 at the proper speed, the resistance 29 causes the supply current of the electromagnets 13 and 23 (Figs. 1 and 4) to vary at the chosen rhythm which as we have seen can be between 30 and 120 pulsations per minute and which can be checked by means of a rotation counter shown schematically in 45. In this case, the motor 44 of the variometer 30 can be stopped and the remainder of the installation is then not pulsed. If, on the contrary, the fly-wheel 41 is engaged in 42b and disconnected from 42a, the motor 44 activates the variometer 30 and the resistance 29 at the chosen rhythm.

The speed of rotation of motors 43 or 44 can be regulated at the required speed, corresponding visibly to the cardiac rhythm of the subject, by acting upon the excitation of these motors by means of a manually adjustable rheostat. If one prefers to regulate the speed of motors 43 or 44 in direct accord with the cardiac rhythm of the subject, one can use an assembly such as that represented schematically in Fig. 8: At 46 there is a contact microphone which, when placed over the subject's heart, produces impulses. These are amplified in the circuit shown and applied to an electromagnet at 47 with a moving core which activates a rheostat; this in turn regulates the current running the motors 43 or 44.

Fig. 9 shows schematically the principle of the oscillating circuit 35. The rectified potential, adjustable between 5000 V and 70,000 V by means of rheostat 34 (Fig. 4) is applied between the terminals 48 and 48a. Terminal 35c (which is also connected to electrode 25, Figs. 2 and 4) is connected to the neutral point on the high tension side of **the** transformer which is a component of the step-up assembly 33 (Fig. 4). The terminals 49 and 49a receive the heating current produced by the resistance 29. The variable condensers 50 and 50a make it possible to regulate to the desired wavelength (which, as seen, is between 1 m and 18 m) the current available at the output terminals 25a and 25b of the oscillator shown.

The modes of implementation described gave been successfully carried out but it is self-evident that these are only examples, and that they might be modified, notably by substitution of equivalent techniques, without going beyond the bounds of the invention. In particular, the electron gun 1 or 1' could be replaced by another charged particle generator.

Resume

The invention includes especially:

> A procedure for obtaining a combination of radiations of different kinds capable of penetrating matter, especially of

intimately penetrating living tissues in order to produce in them certain effects and more particularly in human tissues with a therapeutic effect in mind, consisting of the emission in a cavity of a stream of electrically charged particles, upon which is imposed electromagnetic radiation in the centimeter wavelength range, and the guiding of the resulting radiation emerging from the cavity toward the target to be penetrated.

2. Types of implementation exhibiting the following features taken separately or in the various possible combinations:

   a. The centimeter radiation has a wavelength between 3 cm and 80 cm;

   b. This wavelength is set at the value found by experience to be most suitable for the tissues to be penetrated, e.g., 14 cm for liver and 19.5 cm for spleen;

   c. The particle stream is accelerated by magnetic and electric fields such as those which are used in particle accelerators;

   d. The resulting radiation is accelerated and guided, before its emergence from the cavity, by means of electric and magnetic fields;

   e. The resulting radiation is guided, before its emergence from the cavity, by means of deflecting and/or reflecting surfaces;

   f. The stream of particles and/or the resulting radiation are concentrated and accelerated by means of individual magnetic fields modulated at a frequency between 300 and 900 Hz, the highest frequencies being used to produce hard rays and the lower frequencies to produce soft rays;

   g. The emission of the particle stream, as well as the acceleration and concentration of the radiation resulting at its exit from the cavity, are aided by individual magnetic fields of temporally variable intensity, advantageously pulsed at a rhythm between 30 and 120 pulsations per minute and preferably at the cardiac rhythm of the subject;

h. The assemblages for production of the resultant radiation are pulsed in their entirety at the same rhythm as the magnetic fields according to g;

i. The stream of particles and/or the resulting radiation are accelerated and concentrated by direct magnetic and/or electric fields modulated at a wavelength between 1 m and 50 m and preferably between 1 m and 18 m, this wavelength being advantageously chosen as that which is known in diathermy as suitable for the tissues to be penetrated or for the surrounding tissues.

3. An assemblage making it possible to obtain a combination of a stream of electrically charged particles and a beam of centimeter electromagnetic waves in order to penetrate intimately and to irradiate living tissues and particularly human tissues, the said assemblage comprising at least a particle emitter, means for channeling said particles in a cavity serving as waveguide for an emitter of electromagnetic radiation of which the wavelength is included in the range of centimeter waves and preferably adjustable from 3 cm to 80 cm, means for generating in the cavity magnetic field for acceleration and concentration and means for concentrating and accelerating the resulting radiation at the exit of the cavity.

4. Modes of implementation with the following details taken separately or in the various possible combinations:

a. The particle emitter is an electron gun of which the anode is at the end of the cavity and the cathode is situated further along, this cathode being hollow and placed in the magnetic field of an electromagnetic in order to ensure emission of a stream of particles towards the mouth of the cavity;

b. The cathode consists of a rim connected by two aligned spokes to a hub, the said rim being provided internally with an annular housing containing a heated filament and the said housing communicating with a number of holes arranged annularly and traversing the rim transversely;

c. The cathode is made of a metal of valency close to the mean valency of the chemical molecules comprising the tissue to be penetrated;

d. The cathode is of tungsten or preferably of molybdenum;

e. The cavity contains a rare gas, preferably argon, under a vacuum of the order of 2 mm Hg;

f. The cavity contains a duct carrying at least part of the stream of particles to a cyclotron and a duct bringing back into the cavity the particles accelerated in the cyclotron;

g. The cavity passes through several coils, the supply current for the various coils being capable of undergoing modulation at different frequencies;

h. The downstream end of the cavity is composed of a tube containing, in the part from which the resulting radiation must emerge, a cathode and an electromagnet which may be identical to the cathode and the electromagnetic according to para. a, an anode near the other end, and a rotary deflector consisting of a number of plates arranged *en couronne* on a rotor lacing the incident radiation at such an angle that the radiation deflected and/or reflected is directed toward the cathode, several coils whose supply current can be modulated being distributed over the length of the tube;

i. This tube also contains electrodes supplied by alternating current generating an electric field at the level of the rotary deflector, each of the said electrodes being surrounded by a bobbin of which the supply current can be modulated;

j. Methods are anticipated for modulating, at an adjustable rhythm between 30 and 120 cycles per minute, the supply current of the electromagnetic according to a and h, and preferably to modulate the supply current of the rest of the assemblage at the same rhythm;

k. Methods are anticipated for modulating, at a frequency between 300 and 900 Hz, the supply current of the bobbins surrounding according to i and one or several coils according to g;

1. Methods are anticipated for modulating, at an adjustable wavelength between 1 m and 50 m and preferably between 1 m and 18 m, the supply current of the electrodes according to i, of one or several of the coils generating the magnetic field of the cyclotron;

m. The electrodes of the electron gun consist of two electrodes identical to the cathode according to a, b, c, or d, the cathode of the tube according to h is replaced by a double electrode reproducing the arrangement of the electrodes of the electron gun, and methods are anticipated for reversing at will and simultaneously the polarities of these two pairs of electrodes and the direction of flow of the current supplying the acceleratory coils, a first pattern of polarities assuring the functioning of the apparatus in the conditions which are laid down according to a, and a second pattern of polarities assuring emission in the cavity of a stream of electrons combined with the centimeter radiation and giving rise, at the exit of the said tube, to emission of very hard x-rays.

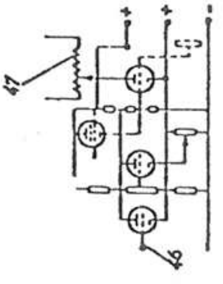

Fig. 18

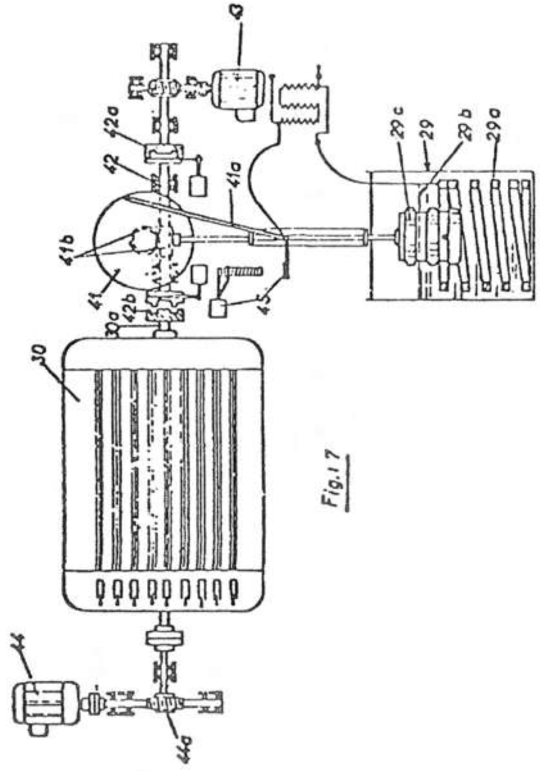
44
44a
30
30a
42b
41
41b
42
42a
41a
43
45
29c
29
29b
29a

Fig. 17

Fig.:9

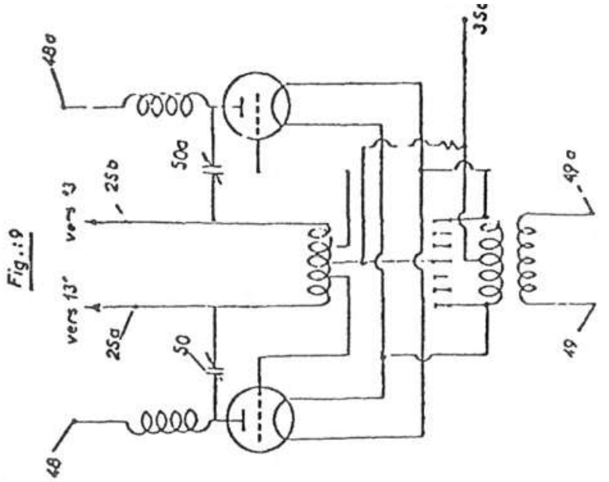
vers 13
vers 13'

Fig.:4

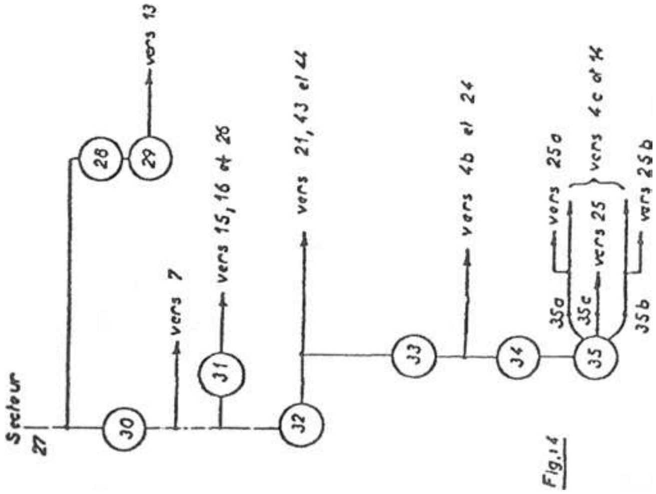
Secteur
27
30
28
29
vers 13
vers 7
31
vers 15, 16 et 26
32
vers 21, 43 et 44
33
vers 4b et 24
34
35
35a
35c
35b
vers 25a
vers 25
vers 25b
vers 4c et 14

Fig. 2

Fig. 1

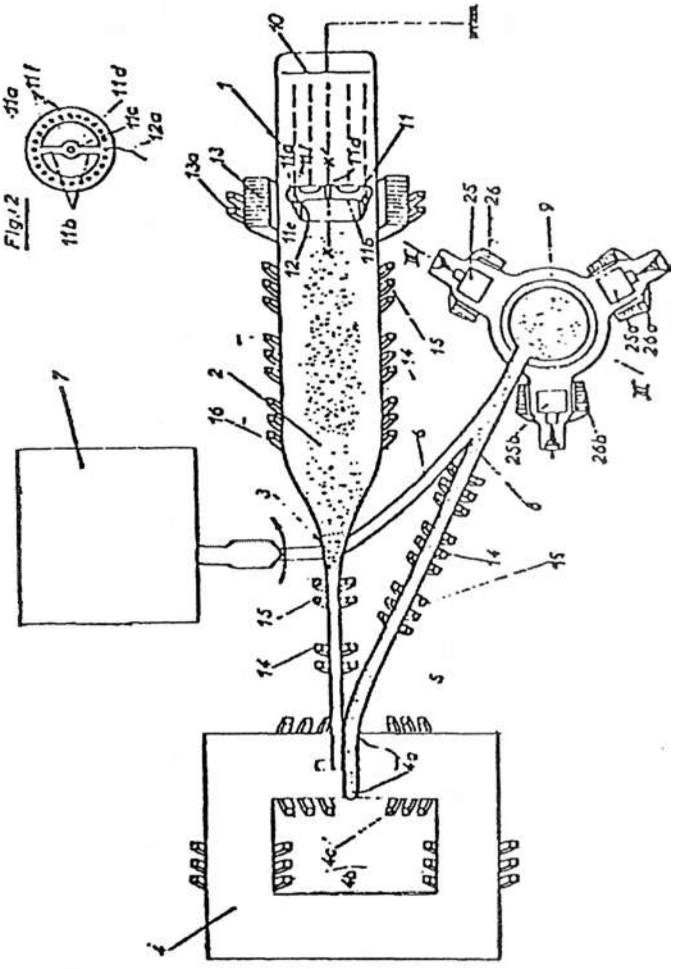
x
x
II
II
n
4a
4c
4b

Fig.: 3

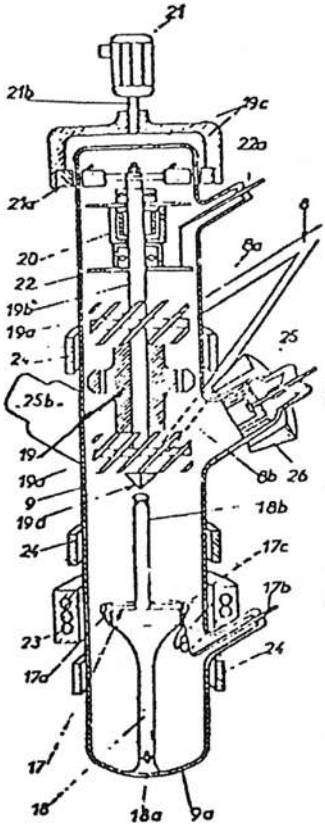
8a
8b
9a

Fig.: 6

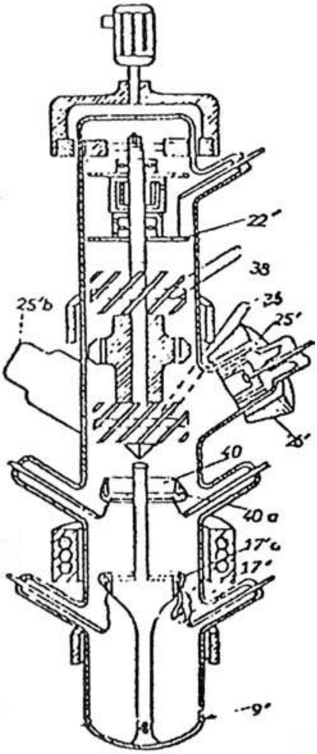
25'b
17'c

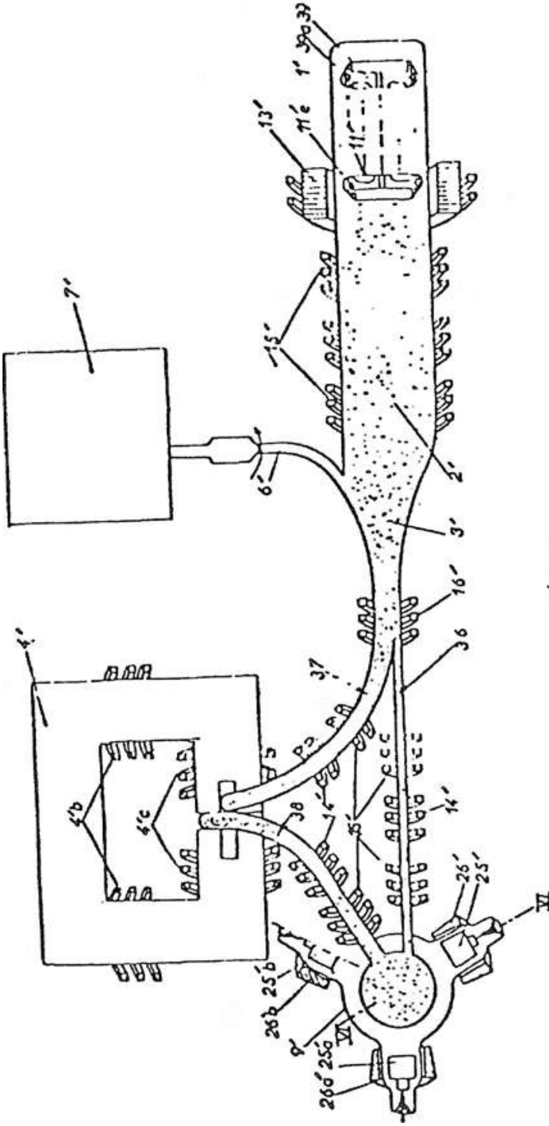
11'e
4'b
4'c
VI
25'b
26'b
VI
25'a
26'a

Fig. 5

Two Patents of G. Lakhousky

June 12,1934 G. LAKHOVSKY 1,962,565

APPARATUS WITH CIRCUITS OSCILLATING UNDER MULTIPLE WAVE LENGTHS

Filed Nov. 13, 1931

Fig. 1.

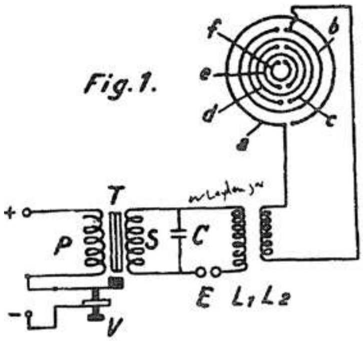
f
b
e
d
c
a
w Leyden jar
T
+
P
S
C
E
L_1
L_2
-
V

Fig. 2.

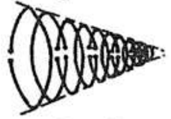

Fig. 3.

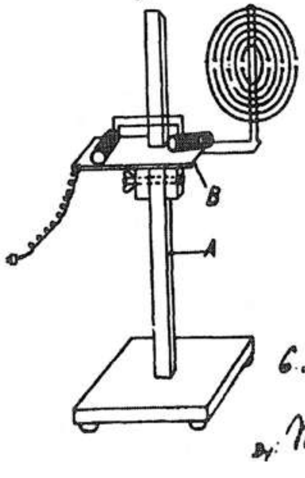

G. Lakhovsky
INVENTOR

By: Marks & Clerk
ATTYS.

Patented June 12, 1934 **1,962,565**

UNITED STATES PATENT OFFICE

1,962,565

APPARATUS WITH CIRCUITS OSCILLATING UNDER MULTIPLE WAVE LENGTHS

Georges Lakhovsky, Paris, France

Application November 12, 1931, Serial No. 574,907
In France May 2, 1931

7 Claims. (Cl. 250—33)

The present invention has for its object an apparatus capable of sending out simultaneously different wave lengths so that among these waves or their harmonics there is found almost always one or several capable of producing the best effect sought for any application whatever.

There is used for that purpose the property of open circuits having a fundamental or natural self induction and capacity to oscillate upon a well determined wave length if they are excited by electric impulses emanating from discharges of any source.

The radiating part of the system, producing waves of various lengths, is therefore constituted by a certain number of concentric open rings of different diameters, which are maintained insulated by any appropriate process. These rings may or may not be terminated by small spheres forming capacities.

These rings may be arranged either in the same plane or in different planes so as to form any surfaces, such as portions of cones, spheres, paraboloids, ellipsoids, etc. . . . , these surfaces being of a nature to concentrate the waves in a determined direction in the manner of a reflector, for example.

These rings may be fed by any producing device whatever of high frequency, for example, a combination comprising a trembler coil (or any other transformer) and a high frequency circuit provided with a self-induction coil and a capacity.

One or several points of this high frequency circuit is or are connected to the end of one or several of the rings forming the radiating apparatus. The other rings not connected are excited by induction.

The complete apparatus (coil, high frequency circuits, rings, etc. . . .), may be mounted in a box or upon a small board capable of sliding along a guide or rocking so as to be able to be placed in the most convenient position.

In the accompanying drawing, by way of example, not limitative, there is shown:

Fig. 1 a diagram of the system.

Fig. 2 the position of the rings in the form of a cone.

Fig. 3 the apparatus mounted upon a foot or base.

In Fig. 1 there is seen at T the transformer or coil (P the primary, S the secondary of this coil), V, the screw of the trembler, C the condenser the discharge of which across the spark gap E produces oscillations in the self-induction coil L^1, which in its turn acts by induction upon the self-induction coil L^2; this self-induction coil L^2 is connected on one side to the first exterior ring *a*, on the other side to the following ring *b*. It thus forms an electrostatic field of high frequency which induces oscillations in the other rings *c*, *d*, *e*, *f*.

As a modification, L^2 may be suppressed and the rings *a* and *b* may be connected each to one of the ends of L^1.

One of the ends, or even any point of the self-induction coils L^1 or L^2 may likewise be connected to any one of the rings.

Fig. 2 shows the arrangement of the rings in the form of a cone.

The open circuits *a*, *b*, *c*, . . . *f* can be constituted by solid conductors. They can also be tubular circuits, this being advantageous, since, for the same bulk, the radiating surface is larger. Finally, the area of the cross section of each of these various circuits can vary from one circuit to the other; for instance, this area can decrease from the periphery (ring *a*) to the center (ring *f*) of the radiating system.

Fig. 3 shows the apparatus mounted upon a foot A along which it can slide, in order to be placed at any desired height and upon a hinge B which allows it to be oriented upon the horizon.

I claim:—

1. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters, the split ends being spaced from each other and terminating with small spheres.

2. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters and terminated by small spheres forming capacities.

3. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters and arranged in one and the same plane the split ends being spaced from each other and terminating with small spheres, in staggered position as regards adjacent rings.

4. An apparatus of the type described comprising a base, a plate slidably mounted on said

base, a support pivoted to the said plate, a series of open and insulated, high frequency radiating circuits mounted on the pivoted support, and a high frequency producing device for the excitation of the said circuits.

5. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits, said circuits being open, insulated, and having dimensions different from each other.

6. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits arranged in the same plane, said circuits being open, insulated, and having dimensions different from each other.

7. An apparatus of the type described comprising a transformer, a high frequency energising circuit fed by the said transformer and provided with a self-induction coil and a capacity, a series of open and insulated high-frequency radiating circuits having dimensions different from each other, an inductive connection between two of these radiating circuits and respectively each of the poles of the outgoing element of the high frequency energising circuit.

GEORGES LAKHOVSKY.

Patented June 13, 1944 **2,351,055**

UNITED STATES PATENT OFFICE

2,351,055

TUBE FOR PRODUCING MULTIPLE WAVE LENGTHS

George Lakhovsky, New York, N. Y.; Anne-Marie Louise Lakhovsky, administratrix of said George Lakhovsky, deceased, assignor to Henry S. Blum, Chicago, Ill.

Application November 21, 1941, Serial No. 420,006

11 Claims. (Cl. 250—33)

This invention relates generally to electric devices excited by electric impulses and more particularly to multiple wave length conducting and/or producing means. This invention has for its primary object the provision of such means disposed within a vacuum tube or a tube containing rare gas or gases.

My work over a period of more than 20 years has led me to the belief that there is great need for means conducting and/or producing multiple wave lengths which may be employed in many industrial and medical fields. Suitable apparatus for producing waves of various lengths is disclosed in my U. S. Letters Patent No. 1,962,565, dated June 12, 1934. Embodiments of this apparatus have been in use the world over. Because of the need for simplification in such apparatus so that it may be handy and easily used by persons lacking skill in electrical matters, I have concluded that by providing one unit which incorporates oscillation generating means and another in which the wave lengths are produced, the apparatus may be more universally used in the treatment of cellular life in therapy, in the aging or treating of liquids and in other industrial fields.

The present invention deals with those units, preferably combined with one another and constructed as a portable device and as stated above, has for its major object the provision of a unitary multiple wave length producing and/or conducting device in the form of a vacuum tube.

The present invention further contemplates the incorporation in such a tube, of means for varying the effect of wave lengths, or selectively employing the means therefor enclosed in or forming part of said tube.

This invention further seeks to provide a tube of the indicated type incorporating means for generating oscillations.

Another object of the invention is to provide a single unit which contains an oscillation generator, which produces waves of definite frequency, and further inductance means each permitting emanation of wave lengths of various values.

The structural features of my invention also form a material part of this disclosure, the objects and advantages being attained in structures such as shown in the accompanying drawing, which exemplifies the invention. The following specification, based on said drawing, more clearly points out the purposes and advantages of my invention.

In the drawing:

Fig. 1 is a vertical, partial sectional, partial elevational view of a tube incorporating features of my invention.

Fig. 2 is a similar view of an alternate form thereof.

Fig. 3 illustrates in a similar manner a modified form of my invention.

Fig. 4 is a fragmentary sectional view of a modification made in accordance with the invention.

Fig. 5 is a similar view of another modification.

Figs. 6 and 7 are sectional views illustrating types of wires used in either of the forms of my invention as shown in Figs. 1, 2 and 3.

Fig. 8 is a partial sectional, partial elevational view illustrating a tube incorporating a plurality of grids and selective capacity means for varying the effects of said grids.

Fig. 9 is a similar view of a tube, such as shown in Figs. 1 or 2 and incorporating oscillating or vibrating means.

In that form of my invention shown in Fig. 1, there is provided a sealed preferably transparent envelope **10** made of glass or like material affixed in the usual manner to a screw or Edison plug **11**. The latter is conventionally provided with a shell terminal **12** and a central terminal **13**. The support or seal **14** of the tube serves to hold wires such as **15** and **16** which connect to the terminals **12** and **13**.

Connected to the wires **15** and **16** is a primary inductance member in the form of a loose coil **17** which may be disposed axially of the envelope **10**, as shown. The upper end of said member **17** forms a tight wound coil **18** which serves to radiate the strongest induced wave lengths.

Surrounding the primary inductance member **17**, there are arranged a plurality of separate coils of wire each being of different length and capacity. Thus, I provide a coil **19** having greatest capacity and successive coils **20**, **21**, **22**, **23**, **24** and **25**, each having less capacity than the preceding coil. These coils **19** to **25** form secondary inductance means each permitting emanation of a wave length of different magnitude and value than the other.

I may support the secondary coils in several ways. As shown, I prefer to provide a dielectric member such as a glass sleeve **26** surrounding the primary inductance wire member **17** and to mount the secondary coils **19** to **25** on the outer surface of said sleeve, leaving the coil **18** exposed at the top thereof. In order to effectively support and centralize the glass sleeve **26**, I provide spacing washers **27** and **28** of mica or the like, top and bottom, and provide a supporting crosspiece **29** on the lower portion of the wire **17**.

June 13, 1944. G. LAKHOVSKY 2,351,055

TUBE FOR PRODUCING MULTIPLE WAVE LENGTHS

Filed Nov. 21, 1941

Fig. 1.

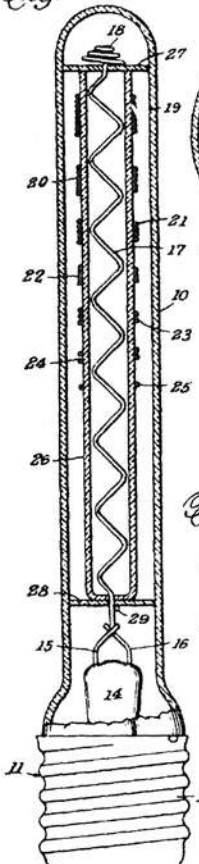

18
27
19
20
21
17
22
10
23
24
25
26
28
29
15
16
14
11
12
73

Fig. 2.

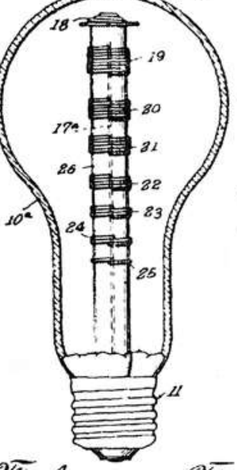

18
19
20
17a
21
26
22
10a
23
24
25
11

Fig. 3.

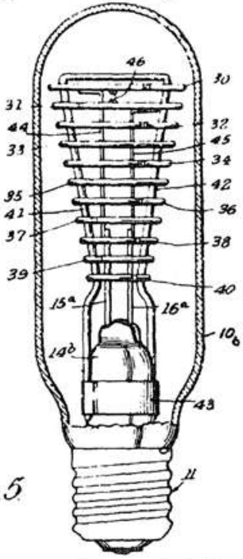

46
30
31
32
44
45
33
34
42
35
36
41
37
38
39
40
15a
16a
10b
14b
43
11

Fig. 4.

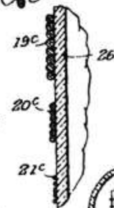

19c
26
20c
21c

Fig. 5.

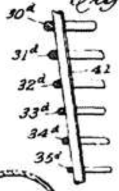

30d
31d
41
32d
33d
34d
35d

Fig. 6

Fig. 7.

Fig. 8.

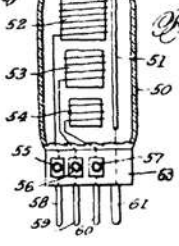

52
51
53
50
54
55
57
63
56
58
61
59
60

Fig. 9.

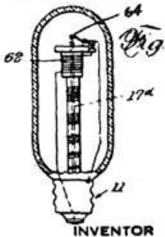

INVENTOR
GEORGE LAKHOVSKY

BY Leon M. Strauss
ATTY.

The tube above described may be employed as indicated for the treatment of cellular structures by connecting it to any one of the rings in the electrostatic field of high frequency, such as shown in said U. S. Letters Patent No. 1,962,565 each of the coils 13 to 25 producing by induction a different wave length, as can be understood.

In the form of my invention shown in Fig. 2, the envelope **10a** is somewhat differently shaped, and the wire *17a* straight instead of arranged as an open coil. In other respects, the structure follows that described with reference to Fig. 1.

As shown in Fig. 3, there may be arranged a plurality of open rings **30**, **31**, **32**, **33**, **34**, **35**, **36**, **37**, **38**, **39**, and **40** in frustro-conical form, for instance, with the largest of said rings at the top and the smallest one adjacent the bottom. I have shown said rings as supported, for instance, by glass posts **41** and **42**, said posts being supported on the seal **14b**, as by a band **43**. I provide extensions for the terminal wires **15a** and **16a** and connect said extensions **44** and **45** to the upper ring **30** and the next lower ring **31**, respectively.

The rings **30** and **31** may become the primary inductance members, whereas the remaining rings constitute the secondary inductance members as before set forth. Electric current conducted to the suitably spaced rings **30**, **31** will cause a spark **46** to be drawn between the rings **30** and **31** providing oscillations emanating from electrical discharges therebetween and causing the radiation of waves of various lengths within the envelope **10b** and affecting the remaining rings.

In Figs. 1 and 2, I have shown the secondary coils as formed of wires having uniform cross-section. Fig. 4 illustrates how such wires may be made progressively smaller in the succeeding coils **15c**, **20c**, and **21c**, etc. Similarly, I have shown the rings of Fig. 3 made progressively smaller in Fig. 5, said rings being designated by numerals **30d**, **31d**, **32d**, **33d**, **34d** and **35d**, etc. In such various manners, the wave length producing effect of the secondary coils or rings may be arrived at.

The secondary coils or rings may be round or somewhat flattened wire as shown in Fig. 6 or tubular as in Fig. 7, offering a yet greater flexibility of design.

The devices of Figs. 1 and 2 above described may be incorporated in the system shown in said U. S. Letters Patent by connecting one of the rings *a*, *b*, *c*, *d*, *e*, or *f* with plug **11** of Fig. 1, since either terminal **15**, **16** is in contact with primary inductance member **11**. To this end a conductor (not shown) connects such a ring with said plug or socket.

In Fig. 8, I have shown a modification of the invention which comprises a vacuum tube **50** containing a plate **51** and a plurality of different wave lengths producing grids **52**, **53**, and **54**. I connect each grid with a variable capacity device, such as **55**, **56**, and **57** carried by said tube **50** and preferably by its base support **63**, whereby the grids may be selectively connected into an operating circuit (not shown) by means of the respective prongs **58**, **59**, and **60**, and the plate prong **61**. Heating or other exciting means may be incorporated in the tube in a well-known manner. This combination tube **50** is designed to replace a plurality of known tubes each operating with different and variable wave length effect, mere exterior manipulation at said base support brings about change of the capacities **55**, **56**, and **57** and cutting in one or more of the grids **52**, **53**, and **54**.

The effect of the spark as produced in the form of the invention depicted in Fig. 3 may be further utilized in Fig. 9 in which I also incorporate a vibrator **62** which produces electrical oscillations by a spark intermittently generated between the end of the primary inductance member **11d** and vibrator armature **64**. In other respects the structure follows that of Fig. 1.

It may be noted that any type of spark producing or oscillation generating means may be substituted for the vibrator shown. The tubes shown in Figs. 3 and 9 may be used in the manner described for Fig. 1 by connecting one terminal of the base as set forth to produce an induction effect. These tubes may be also connected across both terminals with a suitable potential to obtain the spark effect.

From the foregoing it may be seen that I have provided tubes in various forms for the purpose of generating and conducting multiple wave lengths. Other forms of the invention may be produced within the spirit and scope of the invention as claimed.

Having thus described my invention, what I claim as new and desire to be secured by Letters Patent, is:

1. A device of the character described comprising a vacuum tube having an envelope and base, primary inductance means, and a plurality of separate secondary inductance means of different length for radiating waves of various lengths extending above said primary inductance means, all said inductance means being arranged within said envelope.

2. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member axially arranged in said envelope, and a plurality of secondary inductance members spaced apart and independent from each other each member surrounding a respective portion of said primary inductance member and within the field of influence thereof.

3. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member axially arranged in said envelope, and a plurality of secondary inductance members each surrounding the primary inductance member and within the field of influence thereof, each of said secondary inductance members having a different wave length effect than the others.

4. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member comprising a pair of split rings in inductive relation, and a plurality of secondary inductance members each comprising split rings.

5. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member comprising a pair of split rings in inductive relation, and a plurality of secondary inductance members each comprising split rings, said latter rings being arranged in different planes and being of different length to produce a different wave length effect.

6. In a device of the character described, means for producing multiple wave lengths comprising a wire member, enclosing dielectric means for said member, and a plurality of separate wire coils of different lengths on said dielectric means and in the field of inductance of said wire member.

7. In a device of the character described,

means for producing multiple wave lengths comprising a wire member, enclosing dielectric means for said member, and a plurality of wire coils on said dielectric means and in the field of inductance of said wire member, each coil extending over said wire member and having a different number of convolutions than the others.

8. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings.

9. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings, the plurality of rings being of smaller diameter than the mentioned two rings and each progressively smaller than the others.

10. In a device of the character described means for producing multiple wave length comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings, the plurality of rings being of smaller diameter than the mentioned two rings and each progressively smaller than the others, all of the rings being arranged in frusto-conical disposition.

11. In a vacuum tube having a primary inductance member and secondary inductance members, a vibrator arranged in the field of influence of said primary inductance member and secondary inductance members, a vibrator arranged in the field of influence of said primary inductance member for affecting the induction of said secondary members.

GEORGE LAKHOVSKY.

T.H. Moray's Patent
Electrotherapeutic Apparatus
Feb. 1, 1949

Feb. 1, 1949. **T. H. MORAY** **2,460,707**

ELECTROTHERAPEUTIC APPARATUS

Filed April 30, 1943 3 Sheets-Sheet 1

FIG. 1.

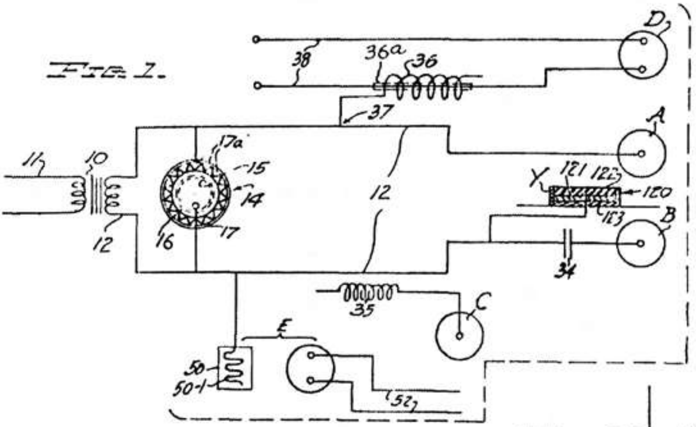
D
38
36a
36
37
A
11
10
17a
15
14
12
Y
121
122
120
123
B
12
16
17
34
35
C
E
50
50-1
52

FIG. 11.

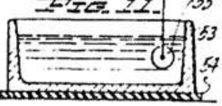

FIG. 3.

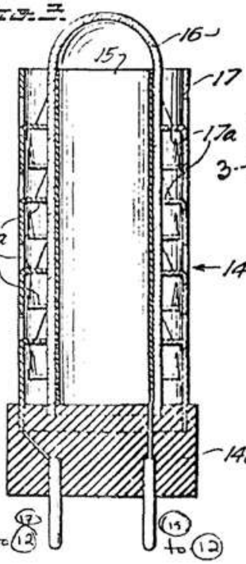
16
15
17
17a
17a
14
14a
17
to 12
15
to 12

FIG. 2.

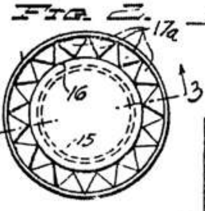

FIG. 12.

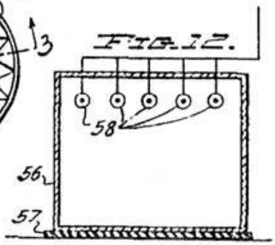

Inventor:
T. H. MORAY.
R. Mallinckrodt and
Philip A. Mallinckrodt,
Attorneys.

Feb. 1, 1949. T. H. MORAY 2,460,707

ELECTROTHERAPEUTIC APPARATUS

Filed April 30, 1943 3 Sheets-Sheet 2

FIG. 7.

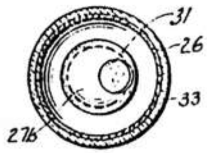

FIG. 5.

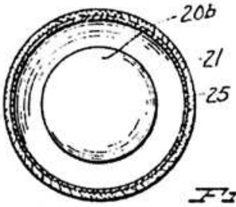

FIG. 9.

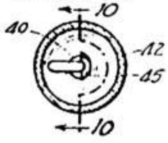
10
40
42
45
10

FIG. 10.

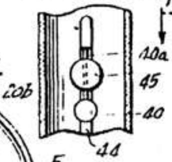
40a
45
20b
40
44

FIG. 6.

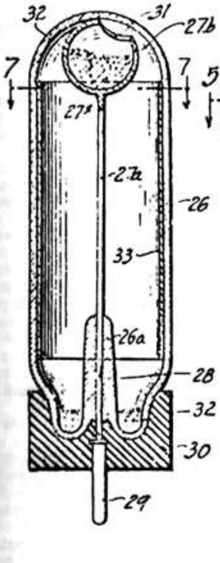
32
31
27b
7
7
27a
27a
26
33
26a
28
32
30
29

FIG. 4.

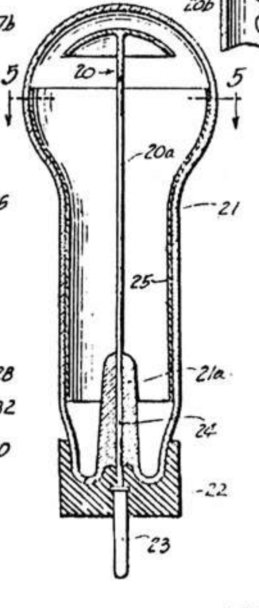
20
5
5
20a
21
25
21a
24
22
23

FIG. 8.

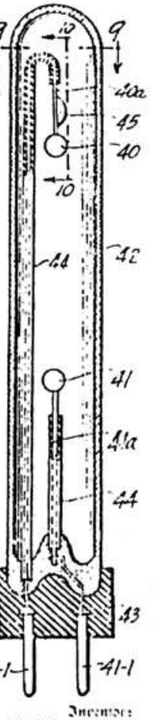
9
10
9
40a
45
40
10
44
42
41
41a
44
43
40-1
41-1

Inventor:
T. H. MORAY,
R. H. Mallinckrodt and
Philip A. Mallinckrodt.
Attorneys.

Feb. 1, 1949. T. H. MORAY 2,460,707

ELECTROTHERAPEUTIC APPARATUS

Filed April 30, 1943 3 Sheets-Sheet 3

FIG. 13.

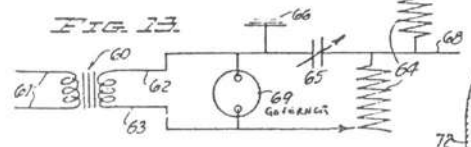
60
61
62
63
64
65
66
67
68
69
GOVERNOR

FIG. 14.

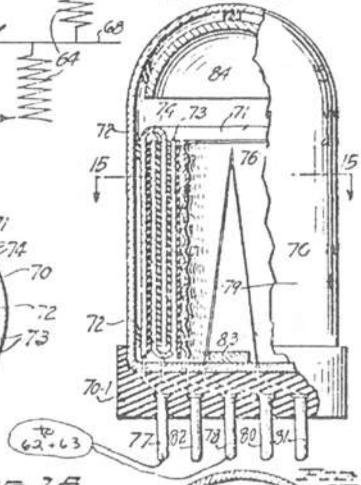
84
74
73
71
72
76
15
15
70
79
72
83
70.1
77
82
78
80
81
to 62+63

FIG. 15.

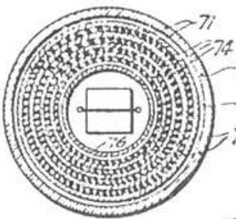
71
74
70
72
73
76

FIG. 16.

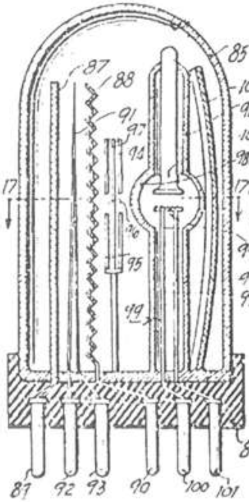
87
88
85
91
97
94
103
98
102
98a
17
17
96
95
99c
99a
99b
99
86
81
92
93
90
100
101

FIG. 18.

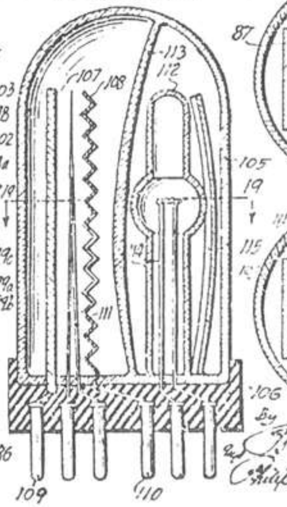
107
108
113
112
105
19
19
114
111
115
106
109
110

FIG.

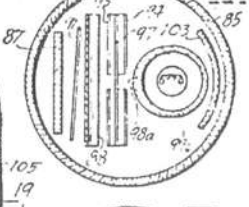
87
95
97
97
103
85
98a
98
91

FIG.

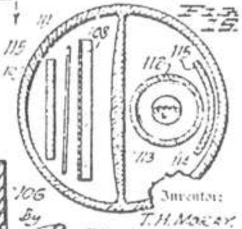
108
115
112
113
114

Inventor:
T. H. MORAY.
By R. Mallinckrodt.
& Chas. A. Mallinckrodt.
Attorneys.

Patented Feb. 1, 1949 **2,460,707**

UNITED STATES PATENT OFFICE

2,460,707

ELECTROTHERAPEUTIC APPARATUS

Thomas H. Moray, Salt Lake City, Utah

Application April 30, 1943, Serial No. 485,112

8 Claims. (Cl. 128—421)

This invention relates to electrotherapeutic apparatus, and to methods of applying electrical, radioactive, and other radiant phenomena therapeutically.

The invention is primarily concerned with the use of high potential, high frequency electricity though not necessarily limited thereto, in conjunction with radioactive and other types of electronic and radiation phenomena, for therapeutic purposes.

Among the objects of the invention are the following:

First.—To render highly effective, from a therapeutic standpoint, radioactive and other types of electronic and radiation phenomena, and, likewise, to render highly effective, from a therapeutic standpoint, high potential, high frequency electricity.

Second.—To augment the therapeutic effect of radioactive and other types of electronic and radiation phenomena by the conjoint use of high potential, high frequency electricity, and, conversely, to augment the therapeutic effect of high frequency, high potential electricity by the conjoint use of radioactive and other types of electronic and radiation phenomena.

Third.—To accomplish the above without danger of burning or of otherwise harming the patient.

Fourth.—To provide apparatus for accomplishing the above, which is relatively simple in construction and operation and relatively inexpensive to produce and operate.

Fifth.—To provide novel electronic and radioactive devices especially adapted for use in conjunction with high potential, high frequency electrical therapy.

I have found that, by enveloping a patient in a high potential, high frequency electrical field in such a manner that no closed circuit is completed through his body, radioactive and other electronic and radiation phenomena can be used therapeutically with considerably greater effectiveness than if used alone. The exact reason for this is not known, nor is it known definitely which, the electric field or the radioactive phenomena, acts upon the other to produce the advantageous results. It is thought, however, that the electric field, permeating the body of the patient as it does, attracts the radioactive emanations or radiations and enables them to penetrate considerably deeper into the tissues and vital organs of the patient than would otherwise be the case. In any event, remarkable therapeutic results have been achieved by use of the invention in the treatment

of malignant tumors, arthritis, sinus infections, and various other diseased conditions.

The invention contemplates the use, in therapeutics, of high potential, high frequency electricity to produce diversified forms of radiant energy, such forms being those which have been found best suited, individually, to benefit various human ailments. In accomplishing this purpose, several special discharge tubes have been developed to serve as treatment electrodes, by means of which correspondingly different curative results are obtained. Throughout the practice of the invention, a prime consideration is that only one terminal of any particular circuit shall be in contact with a patient's body at one time, so there will be no flow of current through a closed circuit of which the patient's body is a part. Such a terminal, too, is usually non-heat producing, so there is no danger of burning. In cases where there is a tendency for a tube to produce X-rays or other injurious rays, these are filtered out.

The present application constitutes a continuation in part of a copending application filed by me November 15, 1940, which bears Serial No. 365,798 and is entitled "Method of and device for the therapeutic application of electric currents and rays," and which has now become abandoned.

In the accompanying drawings, which illustrate several embodiments of apparatus preferred for carrying the method of the invention into practice:

Fig. 1 represents a wiring diagram of a preferred embodiment of apparatus for carrying out the method of the invention in general therapeutic work, several independent treatment stations being provided;

Fig. 2, a top plan view of the novel corona regulator of Fig. 1, employed in the circuit to control and adjust the current and as a governor to safeguard the transformer;

Fig. 3, a vertical section taken on the line 3—3, Fig. 2;

Fig. 4, a vertical section taken centrally through one novel type of discharge tube used as a treatment electrode in the apparatus of Fig. 1;

Fig. 5, a horizontal section taken on the line 5—5, Fig. 4;

Fig. 6, a vertical section taken centrally through another novel type of discharge tube used as a treatment electrode in the apparatus of Fig. 1;

Fig. 7, a horizontal section taken on the line 7—7, Fig. 6;

Fig. 8, a vertical section taken centrally through a novel discharge tube used as a treating device in the apparatus of Fig. 1;

Fig. 9, a horizontal section taken on the line 9—9, Fig. 8;

Fig. 10, a fragmentary vertical section taken on the line 10—10, Figs. 8 and 9;

Fig. 11, a fragmentary view in vertical section, and drawn to a reduced scale, of a tub bath capable of use as a treatment station in the apparatus of Fig. 1;

Fig. 12, a view similar to that of Fig. 11, but showing a shower or vapor bath arrangement for the same purpose;

Fig. 13, a wiring diagram similar to that illustrated in Fig. 1, but fragmentary in nature, and of a somewhat different embodiment of apparatus;

Fig. 14, an elevation, partly in central vertical section, of a novel tube used in the apparatus of Fig. 13 in place of the corona regulator of Figs. 2 and 3;

Fig. 15, a top plan view, partly in horizontal section on the line 15—15, Fig. 14, of the tube of Fig. 14;

Fig. 16, a vertical section of another novel tube which may be used in place of the tube of Figs. 14 and 15;

Fig. 17, a vertical section taken on the line 17—17 of Fig. 16;

Fig. 18, a top plan view of still another novel tube which may be used in place of the tubes of Figs. 14 and 15 and of Figs. 16 and 17; and

Fig. 19, a vertical section taken on the line 19—19 of Fig. 18.

In accordance with the invention, provision is made for enveloping the patient in a high potential and, in certain instances, a high frequency electric field, and for applying to the patient, while so enveloped in the electric field, radiations and emanations having therapeutic value.

The apparatus of Fig. 1 is capable of administering various specific kinds of treatment, pursuant to the invention, at the several treatment stations provided. The treatment stations are indicated A, B, C, D, and E, respectively.

For supplying the high potential electric field, a suitable transformer is employed. This may be of any type capable of delivering high potential electricity, say from 10,000 to 30,000 volts. It is preferred, however, to utilize a conventional double magnetic circuit type of transformer, indicated at 10 in Fig. 1, having adjustable, laminated, magnetic shunts (not shown), the transformer being connected across an ordinary power line 11 charged with the customary 115 v. The output lines 12 from this transformer advantageously extend to the treatment stations A and B, respectively. The first secondary of the transformer 10 is preferably direct connected to the second secondary thereof. It is noted that this high potential electricity may be applied, without causing injury, direct to a patient who is not grounded. However, in order to safeguard the transformer 10 from damage by sparking across its output terminals, and to render the high potential electricity more suitable for therapeutic purposes, which is believed to include the automatic changing of the frequency to an extent which depends upon electrical characteristics of the patient's body, a governor or control device 14 is shunted across the leads 12.

This governor or control device 14 is a sparking condenser of high capacity embodying a multitude of spark gaps. A preferred embodiment of this governor or control device 14 is illustrated in detail in Figs. 2 and 3.

As illustrated, the device comprises a cylindrical, electrically conductive plate 15 surrounded by a cylindrical dielectric 16. An outer cylindrical and electrically conductive element 17 surrounds the dielectric 16 exteriorly. It is provided with a multitude (for example, 250) of inwardly extending prongs 17a, which are advantageously formed by stamping out, and inturning, triangular portions of the electrically conductive element 17. The internal plate 15 preferably contacts the interior surface of the dielectric 16, but, in any event, should lie closely adjacent thereto. Likewise, the tips of the prongs 17a preferably contact the outer surface of the dielectric. The several elements are advantageously mounted in a plug-in base 14a, which is adapted to mate with a suitable receiving socket (not shown) carrying the required electrical connections. The internal plate 15 connects with one of the electric lines 12, while the external element 17 connects with the other electric line 12, as shown diagrammatically in Fig. 1.

It is preferable that the dielectric 16 be in the form of a closed tube or envelope, as shown, and be exhausted to vacuum condition. The multitude of sparking prongs 17a product a brush discharge.

Where the dielectric 16 is not a closed tube or envelope, it is preferred that it be of quartz.

The treatment station A is a discharge tube of a novel type, exemplified by the tubes illustrated in detail in Figs. 4 and 5 and Figs. 6 and 7. Either tube is plugged into the circuit of Fig. 1 at a suitably provided, single-terminal outlet. High potential electricity is, therefore, fed directly into the tube, which serves as an electrode. The tube also embodies radioactive material, which supplies radioactive emanations to the patient simultaneously with the electrical discharge.

As illustrated in Figs. 4 and 5, the tube or electrode may comprise an electrically conductive discharge element 20, having a supporting stem 20a and a major discharge cap or head 20b, which is preferably in the form of a thin, convex-concave plate. The head 20b may be spot welded to the end of the stem 20a.

The discharge element 20 is enclosed within a tube 21 of dielectric material, preferably glass, the stem 20a being fixed in the fused tongue portion 21a of the tube. The tube or shell 21 is fitted into an insulating base 22, provided with a single plug-in terminal 23, and an electrical connector 24 extends from the terminal 23 to the stem 20a.

The inside surfaces of the side walls of the tube or shell 21 are coated with a radioactive material, as at 25. The coating is conveniently made from uranium salts or powdered carnotite or other radioactive ore. The ends of the tube or shell are left uncoated.

Air is evacuated from the tube 21, and a small quantity of mercury introduced. The mercury is preferably triple-distilled to insure great purity. It is preferred that argon or like inert gas be also introduced.

Since the tube just described is plugged into the circuit of Fig. 1, the discharge element or cathode 20 is charged with high potential electricity, and, in its capacity of a treatment station in the apparatus of Fig. 1, serves as an electrode to similarly charge the patient. The patient is insulated from the ground, and the tube is applied directly to the afflicted part of his body, preferably in close contact with the body.

Because of the construction of the tube, radiation of a radioactive nature is also directed against the patient through the uncoated top end of the tube. This radiation has been found to differ somewhat from the radioactive emanations discharging from the side walls of the tube, and is thought to comprise rays lying close to X-rays on the radiation spectrum. These rays appear to have a definite healing value, and to lack the injurious nature of X-rays. Where a predominantly radioactive emanation treatment is desired, the side walls of the tube are placed against the body of the patient.

Best results are obtained when the discharge element or cathode 20 is made of an alloy metal compounded from copper, lead, sulphur, and, if desired, aluminum. The relative percentages of the several ingredients may vary considerably, but a satisfactory mixture comprises 5.0% copper, 55.0% lead, 30.0% sulphur, and 10.0% aluminum. Should aluminum not be used, the difference may be made up by additional copper.

In preparing the alloy, the copper and aluminum are heated to a molten state, after which the sulphur is added while stirring the mixture. After cooling, the mass is again melted, and the lead, in a molten state, is mixed with it, the molten mass being thoroughly stirred. This new mass is then cooled, being later reheated, and, while hot, rolled to make it ductile, so it can be shaped into the desired forms.

The discharge tube or electrode of Figs. 6 and 7 is similar to that of Figs. 4 and 5, having an enclosing tube or shell 26 which is evacuated. A cathode discharge element 27 is positioned within the shell, being fixed in the tongue portion 26*a*. A conductor 28 connects the stem 27*a* of the element 27 with a plug-in terminal 29, which extends outwardly of the base 30. The cap or head 27*b* of the element 27 differs from the cap or head 20*b* of the electrode of Figs. 4 and 5, in that it is spherical in form and hollow. It has an opening 31 formed at its top, contiguous with the top inside surface of the tube 26. A quantity 32 of radioactive material, which may be the same as used for the coating 25 of the electrode of Figs. 4 and 5, is introduced into the tube or shell 26, along with a relatively small quantity of mercury, before the tube is sealed tight. Such material 32 is preferably powdered or granulated, and is shaken into the hollow of the head 27*b* through the opening 31 before any given treatment is commenced. The mercury is provided primarily as a getter, and does no harm if shaken into the head 27*b* along with the radioactive substance. The mercury also tends to produce a vapor in the tube, which aids in the operation thereof. As in the case of the electrode tube of Figs. 4 and 5, this tube may have a radioactive coating 33 covering the inner surfaces of its side walls.

The treatment station B of Fig. 1 differs from the treatment station A only in the fact that a condenser 34 is interposed in the electric supply line 12.

The treatment station C of Fig. 1 differs from the stations A and B only in the fact that the high potential electricity is supplied from the supply line 12 through an inductance 35.

The treatment station D utilizes a germicidal discharge tube, a preferred form of which is illustrated in detail in Figs. 8, 9, and 10. The high potential electricity is taken by induction from the particular supply line 12 concerned. For this purpose, an induction coil 36 is provided, tapping the line 12 at 37. A pair of leads 38 from an ordinary 115 v. supply source extend to a plug-in socket connection for the germicidal tube, one of the leads passing through a glass tube 36*a*, Fig. 1, which is disposed within and extends along the length of the induction coil 36. Thus, high potential electricity is impressed, by induction, upon the ordinary current flowing through the particular lead 38 concerned.

The germicidal discharge tube of Figs. 8, 9, and 10 has a pair of discharge terminals 40 and 41, respectively, positioned in an evacuated tube or envelope 42, and electrically connected with plug-in terminals 40—1 and 41—1, respectively, by means of stems 40*a* and 41*a*, respectively. The tube or envelope 42 and plug-in terminals are mounted in a conventional base 43. It is preferred that insulating material 44, such as a ceramic sleeve, cover the major portions of the stems 40*a* and 41*a*. A piece of lithium metal 45, see particularly Fig. 10, is advantageously secured to the stem 40*a* adjacent the discharge terminal 40 to act as a getter. It may, however, be placed at any other convenient location in the tube. It is preferred that the discharge terminals 40 and 41 be formed of the special alloy previously described. Argon or other suitable inert gas is preferably injected into the tube or envelope 42, as is, also, a small quantity of mercury. The mercury, by vaporizing, aids electrical arcing between the discharge terminals. As will be noted, the high potential electricity induced in the one lead 38 will manifest at the upper discharge terminal 40, and will charge the patient simultaneously with the discharge into his body of germicidal rays from the tube.

The treatment station E embodies the tube of Figs. 8, 9, and 10, as above described, but impresses the high potential electricity directly on the patient instead of passing it first through the tube. For this purpose, a discharge device 50, in the form of a soft, flexible pad in which a coil 50—1 is embedded, taps one of the high potential electric lines 12. This pad 50 is wrapped around the patient's body adjacent the afflicted portion thereof, thus charging the patient. Any other electrode capable of charging the patient with high potential electricity may be used in place of the pad 50. The germicidal tube has its terminals 40—1 and 41—1 plugged into a suitable plug-in socket connected to leads 52 which extend to an ordinary 115 v. source of supply. The high potential electricity with which the patient is charged is induced into the germicidal tube, thereby further activitating the discharge therefrom. A certain beneficial discharge from this germicidal tube will be had by induced activation alone, it being unnecessary, in such instances, to plug the tube into the 115 v. line.

Other types of germicidal and discharge tubes may be used in place of the tube of Figs. 8, 9, and 10, as, for instance, the well known infra-red and ultra-violet lamps, to produce results surpassing those ordinarily attained by the use of such infra-red or ultra-violet lamps apart from the apparatus of the invention.

It should be remembered that the patient is insulated from the ground while being treated at any of the treatment stations of the invention.

Figs. 11 and 12 show how a patient is treated, pursuant to the invention, while immersed in an electrically conductive fluid bath. In Fig. 11, a bath tub 53 is insulated from the ground by a layer of insulation 54. A treatment electrode of the type shown in any of the figure groups 4 and 5, 6 and 7, and 8, 9, and 10 is positioned to charge the fluid of the bath with high potential electricity, as well as to discharge healing radiations and emanations into the patient. The particular electrode illustrated is diagrammatic in form and is designated 55. It may be connected into the circuit of Fig. 1 as shown at any of the treatment stations A, B, C, and D. In Fig. 12, a shower or vapor stall 56 is insulated from the ground by a layer of insulation 57. A plurality of treatment electrodes are designated 58, respectively. These correspond to the treatment electrode 55 of Fig. 11. A water spray or vapor, such as steam, may be admitted to the stall 56 in any well known manner (not shown), thus enveloping the patient during treatment.

Another embodiment of apparatus, pursuant to the invention, is illustrated diagrammatically by the wiring diagram of Fig. 13. While no treatment stations are shown, those provided are identical with the several treatment stations designated A, B, C, D, and E in Fig. 1. The distinction in this embodiment of apparatus resides in the fact that a special generator of high frequency electricity is provided in the system.

A transformer 60 has its input terminals connected across an ordinary 115 v. electric power line 61. Electrical conductors 62 and 63 lead from the respective output terminals of the transformer to a high frequency generator of the Oudin coil type, indicated generally at 64, a variable condenser 65 being interposed in the line 62, and the circuit being grounded at 66. Output conductors 67 and 68, leading from the high frequency generator 64, provide connections for the several treatment stations in the same manner as illustrated in Fig. 1.

The transformer 60 may be any ordinary high voltage type. A governor or control device 69 is shunted across the conductors 62 and 63.

In the illustrated instance, the governor or control device 69 preferably takes the form of a vacuum tube, having the construction shown by Figs. 14 and 15, Figs. 16 and 17, or Figs. 18 and 19. These tubes all possess high capacity, and include elements effecting a brush discharge. They serve, as does the device 14 of Figs. 2 and 3.

The tube of Figs. 14 and 15 embodies an outer shell or envelope 70 of insulating material such as glass, a plastic, or fiber coated with shellac. Inside the shell 70 is a bi-cylindrical element 71 formed of electrically conductive material. Separating element 71 from the enclosing shell 70 are spacers 72 made of rubber, Bakelite, or other insulating material. Inter-fitting with the element 71 is a second electrically conductive, bi-cylindrical element 73, the two elements being separated by a dielectric 74. Inwardly of the element 73, and separated therefrom by a dielectric 75, is a corrugated, cylindrical element 76. The shell or envelope 70 is secured in an insulating base 70—1, provided with plug-in terminals. One of the terminals, designated 77, is electrically connected with the element 71, while another, designated 78, is electrically connected with the corrugated element 76. These two terminals connect with the conductors 62 and 63, as illustrated in Fig. 13, and the brush discharge takes place at element 76.

Under certain circumstances, it is desirable that the outer shell 70 be made of quartz glass, and that a filament 79 be provided, the filament being heated by connection, through plug-in terminals

80 and 81, with a source of low voltage heating current (not shown). Plug-in terminal 82, which is electrically connected with element 73, may be used instead of or in connection with the terminal 77, since element 73 acts in a manner similar to element 71. A getter 83 of suitable material, and an insulating and reflecting shield 84 may be provided, as shown. While the tube may have either a high or a low vacuum condition, or may be filled with an inert gas, I have also found it advantageous to fill the tube with a moist vapor. The tube acts as an oscillator for electric currents, and has an enormous capacity, a capacity many times that of a condenser of approximately equal size.

The tube of Figs. 16 and 17 comprises an outer shell or envelope 85, which may be made of metal, glass, or fused quartz. This shell is mounted in an insulating base 86. Inside the shell 85 is a metal plate 87, and, spaced apart therefrom, a corrugated metal plate 88. A plug-in terminal 89, which extends from the base, is electrically connected with the plate 87, and a second plug-in terminal 90 is electrically connected with the corrugated plate 88. These terminals are adapted to connect, through a suitable socket, with the electrical conductors 62 and 63 of Fig. 13.

Under certain conditions of use, it is desirable to have other elements in the tube. These are provided, and may be utilized or not as occasion warrants. A filament 91 is disposed between the plates 87 and 88. It is electrically connected with the two plug-in terminals 92 and 93, which are adapted to be connected to a source of low voltage heating current (not shown). A slit screen, comprising shields 94 and 95, with apertures 96 extending therethrough, is disposed adjacent that side of corrugated plate 88 which is remote from plate 87. The apertures 96 are in alignment with each other, and the shields 94 and 95 are made of lead or other material capable of screening off X-rays. Between shields 94 and 95 is a sheet 97 of material which is readily permeable to X-rays. Within the shell 85 there is also mounted a shell or envelope 98 of glass, quartz glass, or similar material, having a portion 98*a* which is ground like a lens and directed toward the slit screen. This shell 98 really constitutes a tube within a tube. A filament or cathode 99, comprising electrically conductive legs 99*a* and 99*b* and an electron-emitting portion 99*c*, is disposed within the shell 98, plug-in terminals 100 and 101 being electrically connected to the respective legs 99*a* and 99*b*. A bombardment element 102 is disposed within the shell 98 opposite the portion 99*c* of cathode 99. Within the shell 85, but outside the shell 98, is a reflector 103 directed toward the slit screen.

The tube of Figs. 18 and 19 is essentially the same as the tube of Figs. 16 and 17, being equipped with a shell or envelope 105, a base 106, a plate 107, and a corrugated plate 108, the two plates being connected to plug-in terminals 109 and 110, respectively, which are adapted to connect electrically with the conductors 62 and 63 of Fig. 13. There is a filament 111 and an inner shell or envelope 112, but no slit screen. Instead of a lens portion being provided on the inner shell 112, a partition 113 of lens formation is disposed between the inner shell and the corrugated plate 108. It is fused to the walls of the outer shell 105. Within the inner shell 112 is a filament or cathode 114, which corresponds to the similar element 99 of the tube of Figs. 16 and 17. A reflector **115** is directed toward the lens partition **113**.

Reverting now to Fig. 1, there is another advantageous way of treating a patient pursuant to the invention. As shown at Y, a foot pedestal **120** may be provided for making the patient a part of a condenser. The pedestal comprises an electrically conductive plate element or electrode **121**, connected electrically with one of the high potential lines **12**, and covered by an insulating platform **122** upon which the patient rests his feet while being treated at any of the previously described treatment stations A, B, C, D, or E. The electrode **121** and insulating platform **122** are conveniently mounted in a frame **123**, which insulates the plate from the ground. The insulating platform **122** is made of a high quality insulating material, such as first grade hard rubber. In certain instances it is desirable that the device be made in other than foot-pedestal form. For instance, it may be of cylindrical formation for use in a bed against any part of the patient's body.

If desired, the patient may be charged with the high potential electricity by direct contact with a metal or electrically conductive electrode in place of the pad **50** of treatment station E, or of the tube electrodes.

The invention has been described in the foregoing with sole reference to its use for therapeutic purposes. It should be noted, however, that inorganic matter may also be treated to advantage pursuant to the method and with the apparatus of the invention. It has been found that metals, for example, lead, have changed physical properties after treatment in accordance with the above. In instances where the invention is not being used therapeutically, it is not always necessary to insulate the subject from the ground.

Whereas this invention is here illustrated and described with respect to particular specific embodiments thereof, it is to be understood that various changes may be made in such specific embodiments and various other embodiments may be utilized by those skilled in the art without departing from the spirit and generic scope of the invention as set forth herein and in the claims which here follow.

Having fully described my invention, what I claim is:

1. Apparatus for applying radiant energy therapeutically, comprising means for producing high potential, high frequency electricity; a high capacity sparking condenser, and a treatment electrode connected in circuit with the foregoing, said treatment electrode including a discharge element adapted to charge the patient with said high potential, high frequency electricity, and radioactive means adapted to discharge radioactive emanations into said charged patient.

2. Apparatus in accordance with claim 2, wherein the sparking condenser is in the form of a vacuum tube of high capacity having mutually spaced capacity elements adapted to produce a corona discharge.

3. Apparatus for applying radiant energy therapeutically, comprising means for producing high potential, high frequency electricity; a high capacity sparking condenser; and a treatment device connected in circuit with the foregoing, said treatment device including discharge means adapted to charge the patient with said high potential, high frequency electricity, and radiating means adapted to discharge radiations into the charged patient.

4. Electrical treatment apparatus, comprising a high capacity sparking condenser; a treatment outlet electrically connected to said condenser; and means for electrically connecting said condenser to a source of high potential electricity.

5. Electrical treatment apparatus, comprising a transformer for producing high potential electricity; a high capacity sparking condenser electrically connected across the high potential output terminals of said transformer; and a treatment outlet electrically connected to said condenser.

6. Electrical treatment apparatus, comprising a transformer for producing high potential electricity; a high capacity sparking condenser electrically connected across the high potential output terminals of said transformer; and a plurality of treatment outlets independently electrically connected to said condenser.

7. Electrical treatment apparatus, comprising a transformer for producing high potential electricity; a treatment electrode electrically connected to one of the output terminals of said transformer; an electrical conductor sheathed by insulation electrically connected to the other of the output terminals of said transformer and disposed adjacent said treatment electrode so the subject to be treated may be placed between and in contact with the two; and a high capacity sparking condenser connected across the said outlet terminals of the transformer.

8. In electrical treatment apparatus equipped with means for the supply of high potential electricity and a treatment electrode, a high capacity sparking condenser electrically connected between the said supply means and the said treatment electrode.

THOMAS H. MORAY.

REFERENCES CITED

The following references are of record in the file of this patent:

UNITED STATES PATENTS

Number	Name	Date
628,351	O'Neill	July 4, 1899
647,687	Topham	Apr. 17, 1900
765,470	Friedlander	July 19, 1904
950,842	Davis	Mar. 1, 1910
1,156,317	Santos et al.	Oct. 12, 1915
1,193,018	Howard	Aug. 1, 1916
1,466,777	Winkelmann	Sept. 4, 1923
1,590,930	Falkenberg	June 29, 1926
1,655,783	Gallois	Jan. 10, 1928
2,010,018	Hodnette	Aug. 6, 1935
2,073,428	Schmid	Mar. 9, 1937
2,126,070	Wappler	Aug. 9, 1938
2,192,638	Dixon	Mar. 5, 1940

OTHER REFERENCES

Tousey; Medical Electricity and Roentgen Rays (1910), pp. 493 to 498. Copy in Division 55.

Two Patents of
William J. Hooper

PATENTED OCT 5 1971

3,610,971

FIG. 1

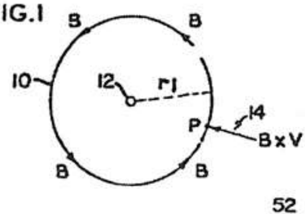
B
B
10
12
r1
P
14
B x V
B
B

FIG. 3

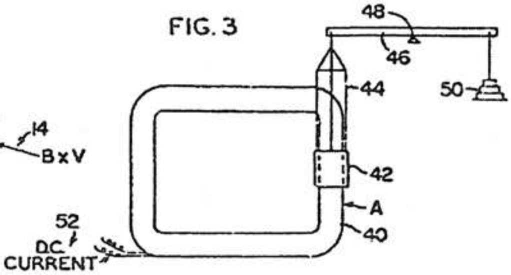
48
46
44
50
42
A
40
52
D.C.
CURRENT

FIG. 2

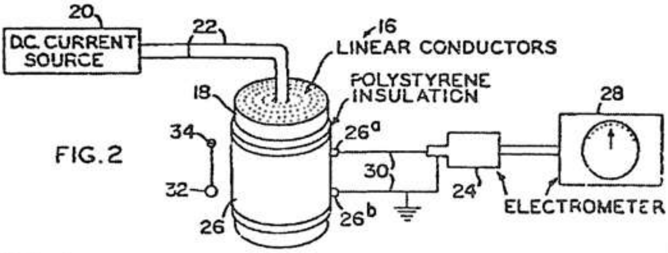
20
D.C. CURRENT SOURCE
22
16
LINEAR CONDUCTORS
18
POLYSTYRENE INSULATION
28
34
26a
30
32
26b
26
24
ELECTROMETER

FIG. 4

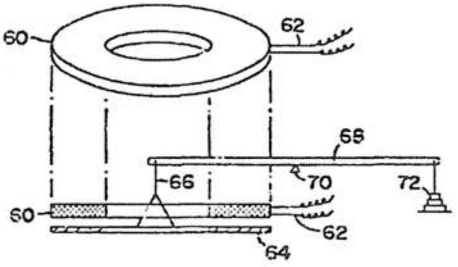
60
62
68
66
70
72
60
62
64

FIG. 5

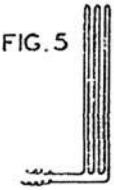

FIG. 6

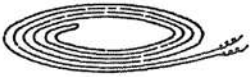

FIG. 8

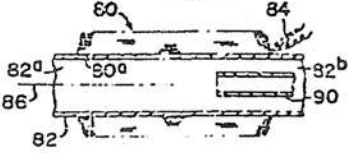
80
84
82a
80a
82b
86
90
82

FIG. 7

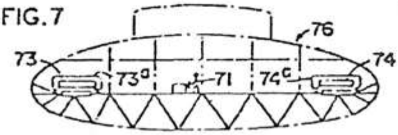
76
73
74
73a
71
74c

INVENTOR
WILLIAM J. HOOPER

BY
Oldham & Oldham
ATTORNEYS.

UNITED STATES PATENT OFFICE
CERTIFICATE OF CORRECTION

Patent No. 3,610,971 Dated October 5, 1971

Inventor(s) William J. Hooper

It is certified that error appears in the above-identified patent and that said Letters Patent are hereby corrected as shown below:

Col. 1, line 14, after "conductor" insert -- with -- and delete "electric charges".

Col. 1, line 45, "10^{17}" should be -- 10^{-7} --.

Col. 2, line 4, "nd" should be -- and --.

Col. 2, line 25, "3.6×10^{12}" should read -- 3.6×10^{-2} --.

Col. 2, line 44, "motion" should be -- motional --.

Col. 4, line 7, "anitgravitational" should read -- antigravitational --.

Col. 4, line 22, "he" should be -- the --.

Col. 5, line 20, "out" should be -- our --.

Col. 6, line 26, after "conductor" delete "o" and insert -- to --.

Signed and sealed this 16th day of May 1972.

(SEAL)
Attest:

EDWARD M. FLETCHER, JR.
Attesting Officer

ROBERT GOTTSCHALK
Commissioner of Patents

United States Patent

[11] 3,610,971

[72] Inventor William J. Hooper
Sarasota, Fla.
[21] Appl. No. 816,356
[22] Filed Apr. 15, 1969
[45] Patented Oct. 5, 1971
[73] Assignee Electrodynamic Gravity, Inc.

[54] ALL-ELECTRIC MOTIONAL ELECTRIC FIELD GENERATOR

5 Claims, 8 Drawing Figs.

[52] U.S. Cl. 310/10
[51] Int. Cl. H02n 11/00
[50] Field of Search 310/10, 11; 324/109; 336/181, 225; 338/284, 297

[56] References Cited

UNITED STATES PATENTS

2,553,875 5/1951 Shaw .... 338/284 X
3,259,784 7/1966 Vogel .... 338/297 X

OTHER REFERENCES

Textbook—The Electromagnetic Field in Its Engineering Aspects by G. W. Carter. 2nd Edition, 1967, Published by Longmans—48 Grosvenor St. London W.1. pp. 168, 169, 170 (Copy in Central Library, Washington, D.C.)

The Feynman Lecture on Physics by Feynman, pages 13–6 to 13–11; Addison-Wesley Publishing Co., New York.

Primary Examiner—D. X. Sliney
Attorney—Oldham & Oldham

ABSTRACT: This invention relates to an all-electric generator yielding a motional electric field in the space surrounding the device, but requiring no mechanical movements of its parts in generating this field. The Theory underlying the production of such a field postulates that each moving electron constituting the current in a linear conductor carries with it a loop of magnetic field energy about it.

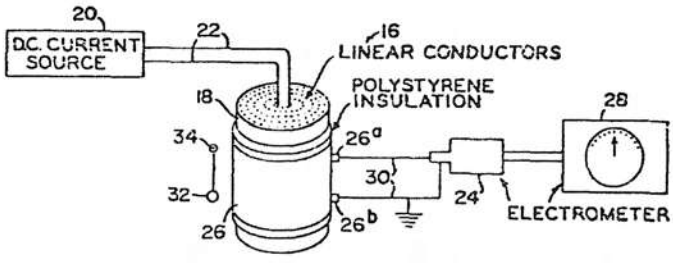

ALL-ELECTRIC MOTIONAL ELECTRIC FIELD GENERATOR

This application is an extension of the teachings presented in my earlier filed pending applications, Ser. Nos. 722,587, filed Apr. 19, 1968 and 803,187, filed Feb. 28, 1969.

PRIOR HISTORY

The detection of a magnetic flux about a current-carrying conductor was first made in 1820 by Hans Christian Oersted. This discovery unified the then separate sciences of electricity and magnetism. My experimental discovery that this magnetic flux actually moves along the conductor electric charges the electric charges constituting the current has great promise of now unifying the three known fields of electricity, magnetism and gravitation.

For a better understanding of the invention, reference should be had to the accompanying drawings wherein:

FIG. 1 is a schematic illustration of a wire containing flowing electrical current indicating the magnetic field generated with respect thereto;

FIG. 2 is a perspective block diagram schematic of a preferred embodiment of the invention;

FIGS. 3 and 4 illustrate modifications of the generator of FIG. 2 adapted to techniques for studying and utilizing gravitational and antigravitational phenomena;

FIG. 5 illustrates the winding of the wire in the device of FIG. 2;

FIG. 6 illustrates the winding of the wire in the device of FIG. 4;

FIG. 7 illustrates the incorporation of a plurality of the devices of the invention into a spacecraft where they might be appropriately utilized; and

FIG. 8 illustrates a coil wound embodiment useful for producing the separation of oppositely charged ions in gases or liquids.

BACKGROUND INFORMATION

The law of Biot and Savart yields the magnetic flux density B (webers/m.^{2}) at a point distant r meters from the conductor carrying a current of I amperes and is given by

$$B=\mu_o I/2\pi r \qquad (1)$$

where $\mu_o=4\pi\times10^{-7}$ webers/amp. meter. With reference to FIG. 1 of the drawings, this flux consists of circular magnetic loops directed counterclockwise, as shown by arrow 10, for current directed in a linear conductor wire 12 at right angles to this paper and outwardly toward the reader.

Electrons producing this flux will be moving perpendicularly into the paper, carrying the flux B with them, at a velocity V meters/sec. Thus, at any point P, distant r_1 from the long linear conductor 12, there will be existent a motional electric field of intensity $E=B\times V$ volts/meter directed radially inward toward the wire in the plane of the paper, as indicated by vector arrow 14, where $B\times V$ is the vector cross product of B and V.

PREFERRED EMBODIMENT

The preferred embodiment of the invention is illustrated in FIG. 2 and consists in part of many insulated linear conductors indicated generally by numeral 16, all parallel, packed and held in close proximity, and connected in series, so as to form a solid package 18 preferably cylindrical in shape. When a constant direct current from a source 20 is caused to flow through wires 22 this device with its axis in the vertical direction, half of the linear conductors 16 will have current flowing upward and half downward. It will be a noninductive device, since surrounding it the magnetic flux due to the upward currents will be superimposed on the flux due to the downward currents, so that no magnetic flux can be measured. On the well recognized "Principle of Superposition of Fields," each one exists however, and acts as though the other was absent.

When the package 18 of linear conductors has a circular cross section there will exist two systems of circularly directed magnetic flux, one directed clockwise and the other counterclockwise as seen from above, and one set moving upward and the other set moving downward. What is of capital interest however is that each of these two systems give rise to equal radially and inwardly directed $B\times V$ motional electric fields distributed laterally in horizontal planes in the space about the cylindrical device. Here we have a field which is not magnetic, is not electrostatic, and which I have shown experimentally to be immune to electrostatic shielding. No such spacially distributed field is known to Physics today except that of gravity. I have shown theorectically that this field should act attractively on electrically neutral matter. I have shown experimentally that the intensity of this field can be measured with the aid of an electrometer system 24, 28 driven by the potential difference induced electromagnetically in the conductors 30 connecting the electrometer to the capacitor plates 26*a* and 26*b* of the capacitor 26. The electrometer head 24 actuates a visual dial indicator 28.

At cryogenic temperatures the drift velocity of the electrons in the linear conductors will be greatly enhanced. Theoretical calculations indicate that this velocity in copper at 20° C. is of the order of 3.6×10^{-2} cm./sec. in a current density of 480 amp./cm^2. The random velocity of free electrons in conductors is theoretically estimated at 10^8 cm./sec. At superconducting temperatures when the electrical resistance approaches zero, it is generally believed that the drift velocity of electrons approaches the value of 10^8 cm./sec. This velocity is 10^{10} times greater than the velocity at 20° C. (See "Physics," Halliday and Resnick, pages 681 and 691, J. Wiley, 1962). This indicates that the intensity of the $B\times V$ field about the device, herein described, might possible be enhanced 10^{10} times at or near the absolute zero of temperature. Hence, the invention contemplates that each of the devices described hereinafter will preferably be operated at or below the critical temperature at which the conductor used becomes superconducting. As the state of the art advances new conducting materials will undoubtedly become available with critical temperatures very considerably above the absolute zero.

Not only does the device that has been described in FIG. 2 afford apparatus hitherto unavailable for the generation of a spacial distribution of the motion electric field, but it affords means for investigating the properties of this field in connection with its relation to gravity and antigravitational phenomena. In addition, it affords for the first time, a direct experimental method for determining the electron drift velocities and mobilities in linear conductors made of different materials.

When the cylindrical capacitor 26 is charged by the $B\times V$ field and its potential difference (P.D.) measured, it is determined by the equation

$$\text{P.D. (volts)}=B\times V\cdot l \qquad (2)$$

where l is the distance between the thin capacitor plates. The direct measurement of V meters/sec. can then be made

$$V=\text{P. D. (volts)}/Bl \text{ m./sec.} \qquad (3)$$

where B is in webers/m.^{2}.

This device will exhibit an attractive force on electrically neutral matter, better defined as a differential force; i.e. the difference between a pull and push, in which the pull is greater. (See reference to general theorem, page 125, The Mathematical Theory of Electricity and Magnetism, by Sir James H. Jeans) The $B\times V$ field is an electrical field acting on all the elementary charges comprising the atoms of matter, electrons, protons and even neutrons, as I believe neutrons also consist of electrons and protons in close bond. The $B\times V$ force being relatively weak does not ionize the atoms of matter, but being immune to shielding, electrically polarizes them. This causes a slight elongation of some of the electrons orbits within the atoms in the direction of the field. The positive charges move slightly into the more intense region of the converging, radially directed field while the negative charges move slightly further way from the device into the weaker portion of the field. Thus the differential force between the pull and the push upon the component parts of an atom results in the polarization of the atoms and a resultant attractive force, equivalent, I believe, to that of gravity.

A small, electrically neutral, simple pendulum 32 suspended close to one side of my device of FIG. 2, about pin 34, should thus undergo a slight measurable deflection under a fairly strong $B\times V$ motional electric field. At cryogenic temperatures this attractive force should be greatly enhanced. When the device is lying horizontally, a gravity meter placed centrally over it should give a measurable reading in milligals.

Modifications of my generator such as are shown in FIGS. 3 and 4 show diverse methods for studying gravitational and antigravitational phenomena which the invention makes possible. FIG. 3 illustrates a rectangular coil 40 made up of a single wire passed many times in the coil configuration so as to provide a large leg A preferably having a circular cross section. However, any cross-sectional configuration would meet the objects of the invention as long as the wires are parallel to each other and closely packed. About one leg of the coil, a cylindrical shell 42, made of electrically neutral, conductive or nonconductive material, is suspended by a carrier frame 44 from one end of a sensitive beam balance 46. The beam pivots about a support bar 48 and carries a load 50 to normally counterbalance the weight of the shell 42. When a fairly strong DC current I is fed through lines 52 and is built up in the coil, the flux density B is generated in the region of the cylindrical weight whose walls are a distance r from the central axis of the coil and n is the number of linear conductors in the coil.

$$B=\mu_o nI/2\pi r \text{ (webers/meter}^2\text{)} \qquad (4)$$

If the current I is suddenly cut off the flux B will rapidly collapse. This flux in the act of collapsing should cross the walls of the suspended cylinder with a considerable inwardly directed horizontal velocity of V meters/sec. This should yield a very considerable vertical $B\times V$ in the walls of the cylinder directed either up or down, depending on the direction of the current in the coil.

In this device the motional electric field $B\times V$ is not convergent, but is uniformly vertical, hence the pull and push exerted on the electric components of the atoms in the cylinder will be equal. In other words the action of this field will either enhance or decrease the state of polarization already existing within the cylinder and produced by the gravitational field of the earth. If the polarization is increased, then the gravitational pull of the earth's field should be greater. If the polarization is decreased then the weight of the cylinder should be less. It is conceivable that the polarization could be reduced to zero. In this case, the object should be weightless. If the intensity of the depolarizing $B\times V$ field is still further increased, then polarization in the opposite direction to that produced by the earth's field should be achieved. In this event the earth's gravitational field would act repulsively on the cylinder. This would be antigravity in its truest sense.

The apparatus I have just been describing acts only momentarily (during the time interval in which the flux is collapsing) to change the weight of the cylinder shell 42 (shown in FIG. 3).

It is desirable to have an apparatus with which to experimentally study gravitational and antigravitational phenomena under steady continuous conditions. In FIG. 4 of the drawings is shown another modification of the device shown in FIG. 2. Here I utilize a noninductive coil wound in the form of a solid, low, thick walled, hollow, right circular cylinder, and indicated generally by numeral 60. DC current is supplied through lines 62 to the coil 60. In the region directly beneath this coil disk 60 I place a similarly shaped nonferrous disk 64, suspended by a carrier 66 to the end of a beam 68. The beam 68 is pivotal about a fulcrum 70 and is counterbalanced by a load 72. For positions of disk 64 below coil 60, the $B\times V$ field can only be directed vertically upward when a continuous DC current is flowing in the wires composing the disk 60. It should be noted that this field is practically uniform and parallel close to the disk, hence it is not attractive in its action on neutral matter. Its action on such matter can only change the state of electrical polarization in the matter. Any change thus effected in the state of polarization should cause the earth's gravitational field to act less intensely, on any object placed close to the under surface of the device, causing it to weigh less, become weightless, or to actually be accelerated upward by the anitgravitational action of the earth's field.

FIG. 5 of the drawings illustrates the manner in which the noninductive device in FIG. 2 is wound with wire 22, by turning the linear conductors back on themselves through a 180° turn. FIG. 6 illustrates the way the noninductive windings of the device illustrated in FIG. 4 is achieved. One very long insulated wire is simply turned back on itself at its midpoint and wound double into a solid coil held together in the shape illustrated.

FIG. 8 illustrates a modified embodiment of the invention which comprises a wire coil 80 either simply wound, or noninductively wound, around a hollow tube 82. Electrical current is passed to the coil 80 through wires 84, and at low temperature this coil will produce a radially, outwardly directed, internal $B\times V$ field from the axis 86 of the cylinder to the internal circumference 80*a* of the coil 80. The electrical wire making up coil 80 is preferably a ribbonlike conductor which is coated with any appropriate electrical insulation. A minimum space between adjacent turns with concentricity between adjacent turns is the preferred winding embodiment for coil 80. As many courses of turns as deemed appropriate to generate field strength desired will meet the objects of the invention. The tube 82 is preferably made of material having little or no magnetic permeability. Both the coil 80 and tube 82 are stationary.

An apparatus of the type illustrated in FIG. 8 might be utilized to effect separation and concentration of fluids. For example, in the desalinization of water, a $B\times V$ electric field generated by the passage of the current through the coil 80, acts upon the disassociated ions of the sodium chloride in solution. Hence, by introducing a flow saline water at end 82*a* of tube 82, and providing a concentric separating tube 90 at end 82*b* of tube 82, increments of fluid which are concentrated with chlorine and depleted of sodium may be drawn off through the stationary tube 90, while the increments of fluid which are concentrated with sodium and depleted of chlorine may be drawn off through the space between the exterior walls of tube 90 and the interior walls of tube 82.

FUTURE APPLICATIONS

This country's space projects have become expensive beyond the ability of the average man to comprehend. Achieving lift by means of costly propellants will some day soon be seen as akin to primitive man's use of awkward clubs as a means of exerting force. In a very short interval of time an intense $B\times V$ field should effect the reversal of the gravitational polarization of an object. Once achieved, only a very small expenditure of energy would be required to hold the polarization. Energy from the atoms of the earth would supply the lift and propulsion through the medium of the earth's gravity field. The most formidable problem would appear to be the problem of holding the $B\times V$ generator at cryogenic temperatures. In view of the overall possibilities of this invention, this problem does not appear to be incapable of a solution. A breakthrough in finding superconducting materials or even near superconducting at slightly elevated temperatures from those presently required would greatly aid in the solution of this problem.

FIG. 7 of the drawings illustrates in outline, the future possible application of this device, obtaining regulated lift by an operator 71 adjusting the current into the coils 73 and 74 which incorporate the structural features of the devices of FIG. 4. The antigravity gravity control features of the coils 73 and 74 thus control the lift and movement of vehicle 76. Suitable cryogenic generators 73*a* and 74*a* are associated with each of the coils 73 and 74 to provide the cooling thereof to as near absolute zero as possible to give the enhanced current passage characteristics necessary to the objects of the invention.

While I have demonstrated that AC current can be used in my device, it appears at present that DC current is much to be preferred.

This invention could be uniquely adapted to a communication system from one side of a solid barrier to the other side whether this barrier be solid concrete or battleship armor, by pulsing the *B×V* generator coil with signals, Morse Code or even voice modulated signals to be picked up by a capacitor-electrometer circuit on the other side of the barrier. Secret communications could be carried on and with the electrometer circuit as shown in FIG. 2 connected to a relay radio broadcasting unit, so that one might conceivably broadcast by radio from the interior of a solid metallic enclosure.

It is conceivable that the invention might be useful in the highly specialized art of concealed detonation devices. A detonating circuit connected to the capacitor electrometer receiving circuit of my device such as shown in FIG. 2 could be actuated through a solid barrier by out *B×V* generating coil placed on the other side of such a barrier.

It is conceivable that for some specialized purposes, it might be highly desireable to transmit electrical power from one side of a fixed solid barrier, concrete or armor plate, etc. by pulsing, DC or AC, in our *B×V* coil on one side of such a barrier. A resonant LC circuit could pick up these *B×V* pulsations and supply a source of electrical power on the other side of the barrier, in a manner somewhat analogous to my electrical receiving methods outlined in U.S. Pat. Application, Ser. No. 803,187, identified above.

In accordance with the patent statutes, only the best known embodiments of the invention have been illustrated and described in detail, but it is to be understood that the invention is not limited thereto or thereby, but that the scope of the invention is defined in the appended claims.

What is claimed is:

1. Apparatus for demonstrating an electrical phenomenon comprising,

a coil made from an electrically conductive material having a very large plurality of sections compacted in close proximity arranged in substantially complementary relation around a common central axis, where the coil forms a solid cylindrical shape with the sections parallel to the axis thereof, and

means to pass current through the coil so that about half the sections will pass current in a direction opposite to the other sections to cancel magnetic flux.

2. Apparatus according to claim 1 where adjacent sections are electrically insulated from each other, and the coil comprises a single wire bent back on itself to make the sections.

3. Apparatus according to claim 1 which includes means to increase the passage of electrical current through the conductor.

4. Apparatus according to claim 3 where the means to enhance is a cryogenic generator which reduces the temperature of the conductor o or below its critical temperature at which it becomes superconductive.

5. Apperatus according to claim 1 which includes an elongated hollow tube having an axis, the conductor formed around the tube.

Patented April 11, 1972 3,656,013

4 Sheets-Sheet 1

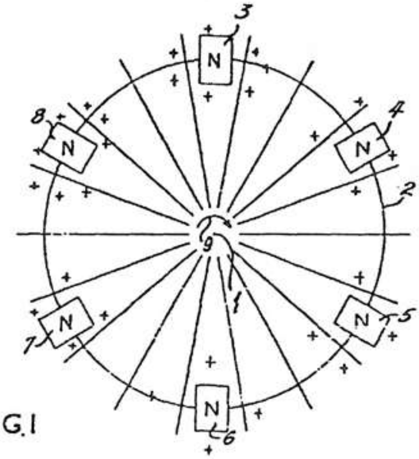
3
N
8
N
4
N
2
9
1
7
N
5
N
N
6
+

FIG. 1

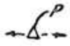

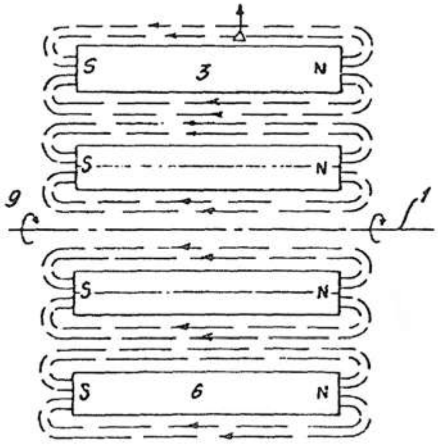
S
3
N
S
N
9
1
S
N
S
6
N

FIG. 2

INVENTOR:
WILLIAM J. HOOPER

BY Oldham & Oldham
ATTORNEY.

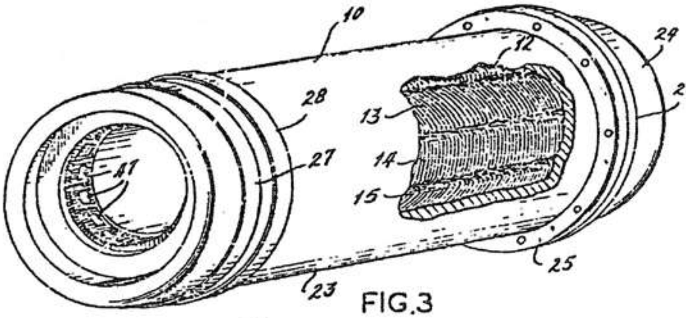
10
12
24
28
26
13
27
14
47
15
25
23

FIG.3

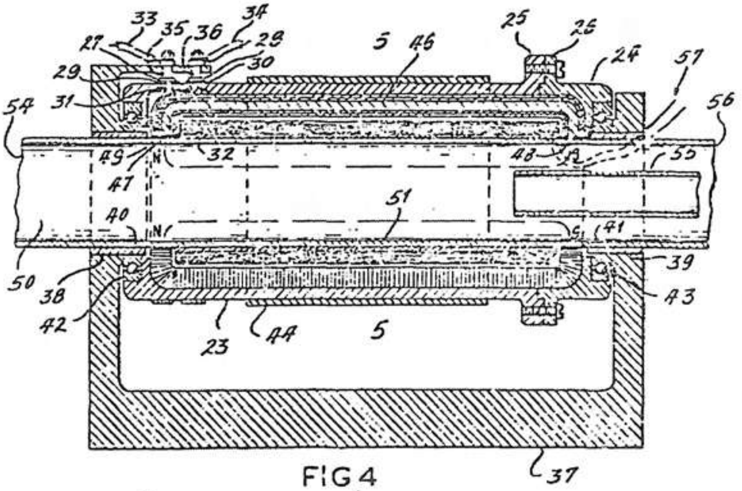
33
35
36
34
27
28
5
46
25
26
24
57
29
30
29
31
56
54
49
32
48
55
47
51
40
50
41
39
38
43
42
23
44
5
37

FIG 4

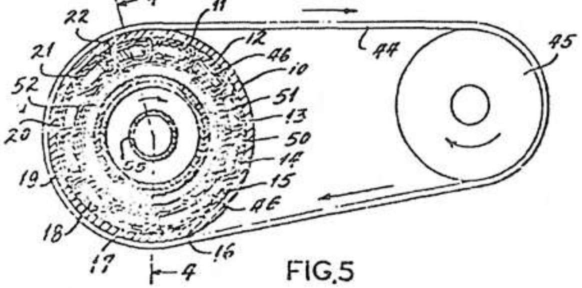
4
11
22
12
44
45
21
46
10
52
51
13
20
50
14
55
19
15
46
18
17
16
4

FIG.5

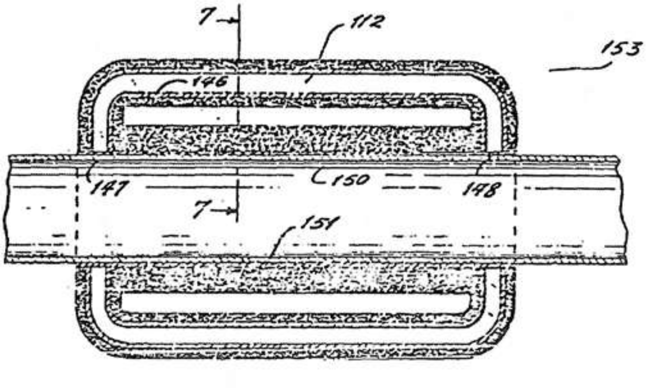
7
112
153
146
147
150
148
7
151

FIG.6

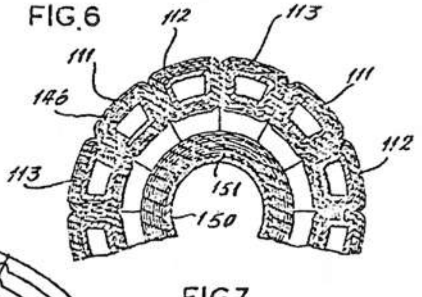
112
113
111
146
111
112
113
151
150

FIG.7

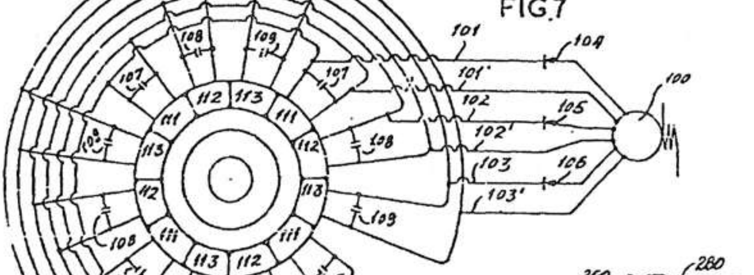
101
104
108
109
107
107
101'
100
112
113
111
102
105
111
108
108
113
102'
112
103
106
118
103'
108
109
111
111
113
112
107
107
109
108

FIG.8

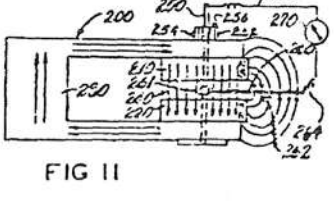
280
200
250
256
270
254
252
260
219
261
230
260
220
264
262

FIG 11

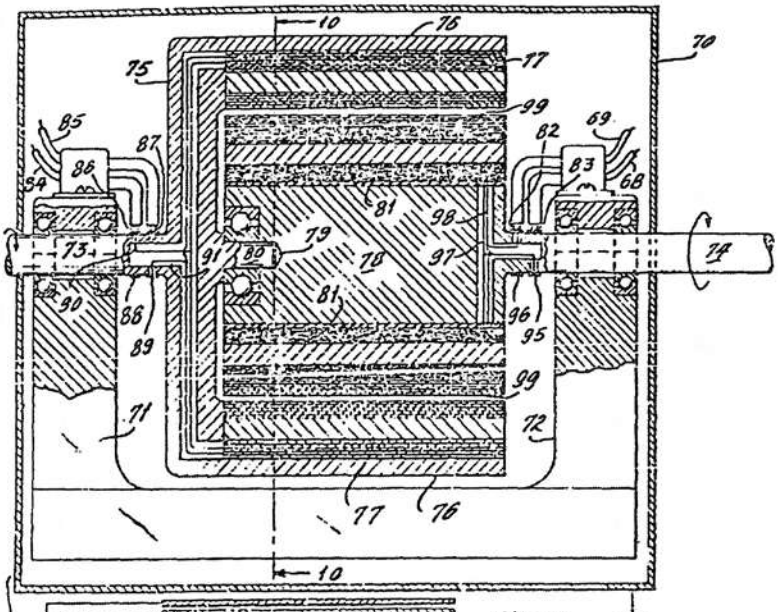

FIG. 9

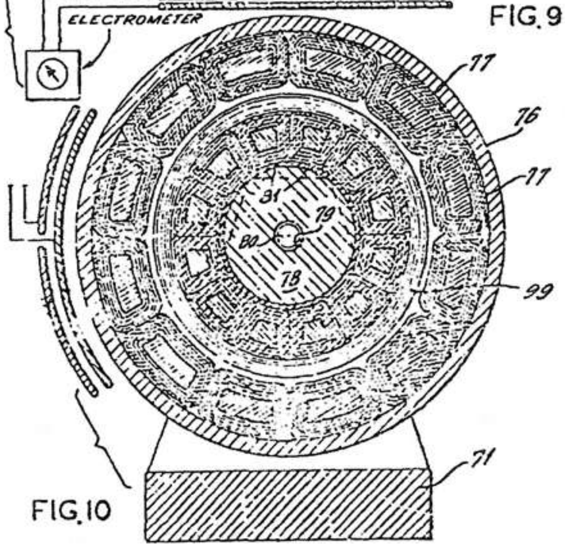
ELECTROMETER

FIG. 10

United States Patent [15] **3,656,013**
Hooper [45] **Apr. 11, 1972**

[54] APPARATUS FOR GENERATING MOTIONAL ELECTRIC FIELD

[72] Inventor: **William J. Hooper, Cuyahoga Falls, Ohio**
[73] Assignee: **Electrodynamic Gravity, Inc., County of Sarasota, Fla.**

[22] Filed: **Apr. 19, 1968**
[21] Appl. No.: **722,587**

[52] U.S. Cl. 310/10
[51] Int. Cl. H02L 1/00
[58] Field of Search....................310/10, 11, 178, 177, 156, 310/42, 268, 211, 169, 216, 269, 259; 103/1; 55/2, 14, 101, 123, 127, 120; 73/194 EM; 328/233–238

[56] **References Cited**

UNITED STATES PATENTS

3,285,179	11/1966	Resler, Jr.	103/1
3,336,489	8/1967	Volger	310/40
3,431,441	3/1969	Shair	310/11
399,800	3/1889	Thomson	310/115
460,087	9/1891	Hewett	310/46
2,066,343	1/1937	Gillen	310/46
2,558,540	6/1951	Clos	310/46
2,705,762	4/1955	Pile	310/104
3,277,631	10/1966	Sunnen	55/3
3,385,983	5/1968	Bohn et al.	310/11

OTHER PUBLICATIONS

Bock–Theory and Calculation of Electrical Apparatus by C. P. Steinmetz; N.Y.– 1917.
Text Book– The Electromagnetic Field in its Engineering Aspects G. W. Carter– 2nd edition– 1967 pages 168 to 171 incl., 328, 329, 334, 335 and 336
The Feynman Lecture on Physics by Feynman, Leighton and Sands Addison– Wesley– Publishing Co., New York; pages 17–1 and 17–2

Primary Examiner—D. X. Sliney
Attorney—Oldham & Oldham

[57] **ABSTRACT**

Apparatus for producing and demonstrating properties of motional electric fields by means of rotating magnetic flux produced by a plurality of magnets extending parallel with the axis of rotation, said flux of these magnets being put into rotation about a common axis by mechanical or by electro magnetic means.

6 Claims, 11 Drawing Figures

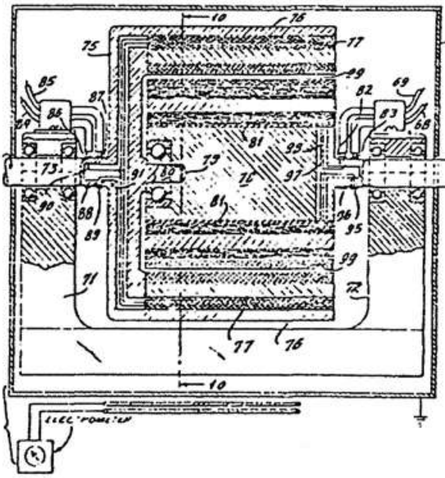

APPARATUS FOR GENERATING MOTIONAL ELECTRIC FIELD

This invention relates to the generation of motionally induced electric fields as distinct from electrostatic fields, and as distinct from the electric field (illustrated by transformer action) which is induced by the time rate of change of magnetic induction. Hence, all reference herein to a motional electric field means the electric field that is generated by the movement of magnetic flux.

The motional electric field generated by rotational motion of magnetic flux is non-uniform in both direction and radial intensity. Particles (large or small) of matter (solid, liquid or gaseous) which are within the ambit of this motional electric field are acted upon with a force which tends to accelerate them. It is well known that a particle carrying a charge Q (coulombs) in an electric field of intensity E (Newton/coulomb) will be acted upon by a force F (Newton) given by the equation $F = EQ$. The electric intensity E of a motional electric field is given by the vector equation $E = B \times V$, where B is the magnetic flux density at a point in space, expressed in webers per square meter, and V is the velocity of the moving magnetic flux at that point in space, expressed in meters per second. What is not so well known, however, is the fact that a non-uniform electric field exerts a force upon electrically neutral matter. In his celebrated text, "The Mathematical Theory of Electricity and Magnetism", Sir James H. Jeans describes, in the case of a non-uniform electrostatic field, how a "slab of dielectric will be sucked in between the plates of the condenser" thus demonstrating the mechanical force produced by such a field. He states (P. 125), "This, as will be seen later, is a particular case of a general theorem that any piece of dielectric is acted on by forces which tend to drag it from the weaker to the stronger parts of an electric field of force." According to Clerk Maxwell in "Electricity and Magnetism" (Vol. II, p. 181), Faraday discovered that the electric field due to electromagnetic induction penetrated within and throughout all materials. The motional electric field is just such a field. Apparatus for demonstrating the "general theorem", stated by Jeans, has, in the case of the non-uniform motional electric field, not heretofore been provided.

Accordingly, it is one of the objects of the present invention to provide a device for generating a non-uniform motional electric field whose force is of magnitude such that it can be sensed.

Since the motional electric field acts within and throughout all matter, such a non-uniform field will exert a physical force on any kind of electrically neutral matter, in a manner quite analogous to that of the electrostatic field with respect to dielectric material.

Another object of the invention is to provide apparatus for obtaining useful work from a non-uniform motional electric field.

The several objectives of the invention may be accomplished by rotating an elongate magnetic field about an axis which is concentric with the field and parallel with the field's elongation. For example, an array of magnets may be mounted on a rotor so that all magnets are parallel with the axis of the rotor, and all have the same polar orientation relative to the rotor. A solenoidal electro magnet, or another array of magnets, with magnetic axis concentric with the rotational axis may also be employed to enchance the flux density of the array of rotating magnets by flux linking with them. When such an array of magnets is rotated about the axis of the rotor, the magnetic field of each is rotated with it. As another example, an elongate magnetic field can be rotated about its axis, without mechanical movement, by the provision of stationary solenoidal electro magnets arrayed like the staves of a barrel and connected to a source of multi-phase alternating current (half wave rectified to prevent reversal of sign) in a manner comparable to the stator winding of a three-phase motor to produce "revolving field." While the composite magnetic field is thus in rotary motion, every charged, or uncharged particle of matter (solid, liquid, or gaseous) within the non-uniform motional electric field thus generated in the neighborhood of the rotor, will be acted upon by a force due to the electric intensity E. Positively charged particles will be forced in the direction of the vector E, negatively charged particles will be forced in the opposite direction, and electrically neutral matter will be pulled from the weaker regions of field intensity to the stronger regions. This latter mechanical force is called a differential force due to the action of the field on the internal electrical polarization of the neutral matter.

The three vectors E, B, and V are always (with a possible exception) mutually at right angles to each other. If the thumb of the right hand is pointed in the direction of the magnetic flux density vector B, and then rotated like a right-handed screw so that it points in the direction of the velocity V of the moving magnetic flux, then the forefinger of the right hand will point in the direction of the electric vector E. The rotation of the composite magnetic flux about its axis produces a motional electric field which will, except at the axial ends of the magnetic field, be quite generally directed radially with respect to the axis of rotation. The possible exception is the case of a magnet having a homogeneous electrically conductive core of perfectly circular cross-section, in which case the force of the electric field may be somewhat off radial.

With a given polar orientation of magnets relative to the axis of rotation, rotation of the magnetic field in a clockwise direction will generate a motionally induced electric field of sign (positive or negative) opposite that which is generated when the rotation is counterclockwise. For any given sense of rotation, reversing the polar orientation of the magnets relative to the axis of rotation likewise reverses the sign (positive or negative) of the motionally induced electric field. Reversing both the direction of rotation and the polar orientation of the magnets (a situation exemplified by viewing the apparatus first from one axial end, and secondly from the opposite axial end) produces no change in the sign (positive or negative) of the motionally induced electric field.

In the accompanying drawings, three embodiments of the invention are exemplified. Also, there are two model schematics to illustrate in its simplest form the principle of the invention, in which a plurality of commonplace permanent magnets are mounted upon a rotor.

The first embodiment is one in which an array of electro magnets is mounted on a rotor having a hollow center, through which particles of matter can be moved and their response to the motional electric field perceptibly demonstrated.

Second, there is an embodiment in which electro magnets are arrayed as in the first embodiment, but are stationary and their composite magnetic field is rotated electromagnetically by phase displacement.

Third, there is an embodiment in which a pair of oppositely rotating rotors, nested one within the other, are each provided, on its circumference, with an array of electro magnets oriented as aforesaid on each rotor, but oppositely oriented on the respective rotors, and wherein the effect of the motionally induced electric fields can be perceptibly demonstrated by the behavior of particles of matter externally of the rotor.

In the accompanying drawings:

FIG. 1 is a diagrammatic view illustrating the effect of rotating an array of spaced permanent magnets about an axis which is parallel with each of the magnets when all magnets have corresponding polar orientation;

FIG. 2 is a diagrammatic view representing a side elevation of the arrangement shown in FIG. 1, and illustrating the external magnetic flux paths of the several permanent magnets;

FIG. 3 is a perspective view of a rotor for generating a radially directed motional electric field in accordance with one embodiment of the invention, wherein electro magnets are arrayed within a hollow rotor with axes parallel to that of the rotor;

FIG. 4 is a sectional view taken along a diametric plane of the rotor shown in FIG. 3, together with its cooperating parts;

FIG. 5 is a sectional view along line 5—5 of FIG. 4, and showing, in addition, driving means for the rotor;

FIG. 6 is a sectional view taken along a diametric plane of a stationary form of apparatus similar to that shown in FIGS. 3, 4, and 5, but wherein the magnetic flux field is rotated electromagnetically by phase displacement;

FIG. 7 is a sectional view taken along line 7—7 of FIG. 6;

FIG. 8 is a wiring diagram for the apparatus shown in FIGS. 6 and 7.

FIG. 9 is a longitudinal sectional view of an apparatus having a plurality of rotors for generating a B × V electric field;

FIG. 10 is a sectional view taken along line 10—10 of FIG. 9; and

FIG. 11 is a diagrammatic illustration of an apparatus for exhibiting moving field phenomena.

In FIGS. 1 and 2, the principle of the invention is diagrammatically illustrated with the utmost simplicity, wherein a catena of magnetic flux fields, all with the same polar orientation, is produced by arranging an array of rod-type permanent magnets parallel with each other, and parallel with the axis about which they are to be rotated. In FIG. 1, the axis of rotation is designated 1, and may be considered as the center of a shaft of a rotor, preferably composed of material having low, if any, magnetic permeability, and low, if any, electrical conductivity. The periphery of such a rotor is designated by the circle 2. In the form shown, a series of permanent magnets 3, 4, 5, 6, 7, and 8 is appropriately secured to the periphery of the rotor with the respective magnets in equi-spaced relationship, and all with their north poles addressed in the same direction which, as shown in FIG. 1, is in the direction of the reader or, as shown in FIG. 2, to the reader's right. The external flux fields from the respective permanent magnets 3, 4, 5, and 6 are illustrated in FIG. 2, where it will be observed that the external flux path of all magnets is from right to left, a relationship designated by the plus (+) marks in FIG. 1. With such an arrangement, the flux fields of all magnets may be concurrently rotated about axis 1. Considering the rotation to be in the direction of arrows 9 and the polar orientation of the several permanent magnets to be as illustrated, a B × V *motional electric field will be generated both externally and internally* of rotor periphery 2 upon rotation of the rotor. The thus generated motional electric field will exert a radially directed force upon every particle of matter, neutral or charged, such as particle P, which is within the ambit of the catena of magnetic flux fields. The magnitude of the so impressed electric field intensity will depend upon the magnetic flux concentration at the particle P and the relative velocity between the flux field and particle P. While, if particle P is in motion, its velocity must be considered in arriving at the relative velocity V (in the formula $E = B \times V$) the disclosure will be simplified *if the particle P be considered as stationary, and in such event* V will be 2 R times S, where R is the radial displacement of particle P from axis 1, and where S is the revolutions per unit of time made by the rotor. Hence, with the particle P stationary, the magnitude of the force impressed upon it is increased when the speed of rotation of the rotor is increased at a given flux density; and is increased by an increase in the flux density at a given speed of rotation. The direction in which the force of the motionally induced electric field acts upon particle P is always *perpendicular to the magnetic flux field* B and to its velocity V. The sign of the electric force field may be either positive (acting radially outward from the axis of rotation) or negative (acting radially inward toward the axis of rotation), which, for any given polar orientation of the permanent magnets, may be reversed by reversing the direction of rotation of the rotor. With the polar orientation shown in FIGS. 1 and 2, and the direction of rotation indicated by arrow 9, if particle P is a proton or a positively charged molecule, the force of the $B \times V$ *electric field will be radially inward* as indicated by the solid line arrow attached to particle P; but if particle P is an electron or a negatively charged molecule, the particle will tend to move radially outward as shown by the broken line arrow; and if the particle P is electrically neutral, it will tend to move in the direction of increasing electrical intensity E. Increase of E requires either that both B and V be increasing or that one be increasing at a sufficiently greater rate than the other is decreasing so that the product $(B \times V)$ increases. In the relationship of field and particle shown in FIG. 1, the product $(B \times V)$ increases radially inward.

For practical purposes, however, commonplace permanent magnets produce insufficient flux density in their external flux paths to impress upon a particle of charged matter such as P, an electric force of sufficient magnitude that the force of the electric field can be measured with ease, at least at speeds of rotation within practical limits. Consequently, the practical embodiments of the invention, later to be described, make use of electro magnets of a form capable of producing external flux fields of vastly greater flux density B, as well as higher velocities V.

The term "unipolar induction" has been used to indicate *the induction of an electric field in the vicinity of an axially* symmetrical magnetic system rotating about its axis of symmetry. There has been disagreement about whether as the magnetic system is rotated the lines of magnetic induction are carried with it or remain stationary while the magnetic system rotates through them. For example, when a bar magnet is rotated about its magnetic axis, it is believed by many that the magnetic field is stationary. I have demonstrated that under certain conditions, at least, the magnetic lines can be made to move, and I have devised certain methods and apparatus for utilizing this phenomena as in the production of localized and high intensity electric fields.

Turning now to FIG. 11 of the drawings, an apparatus for demonstrating certain magnetic phenomena is illustrated in diagrammatic form. A magnetic system, indicated generally by numeral 200, includes a pair of permanent magnets 210 and 220 which are axially magnetized and mounted in axial alignment between the ends of a U-shaped yoke 230 of highly permeable magnetic material. The adjacent surfaces of magnets 210 and 220 being of opposed polarity are spaced apart, providing an air gap 240 therebetween. The lines of magnetic flux are shown in the drawing and indicated generally by numeral 262. A brass rod 250 extends axially through magnets 210 and 220. The magnetic system 200 is rotatable about rod 250. A stiff conductor 260 is electrically connected by a collar 261 at one end in rotatable or fixed relation as selectively desired, with the brass rod 250 at a point in the gap 240 between magnets 210 and 220. The other end of conductor 260 extends substantially perpendicularly from rod 250 to a distance in space away from the magnets and yoke where the magnetic field falls to approximately zero intensity. In order to complete an electrical circuit through galvanometer 270 and conductor 280 to the brass rod 250 and conductor 260, a flexible and extensible electrical wire 264 connects the end of conductor 260 to the galvanometer 270. A gear 252 locked alternatively by set screw 254 to rod 250 or by set screw 256 to yoke 230 can serve to provide driving motion to either the rod 250 or YOKE 230 as selectively desired. Normally, for the purposes of the experiment, the driving motion will be reciprocal because of the connection of wires 280 and 264.

With this apparatus, if the magnetic system 200 is held fixed and a given angular displacement imparted to conductor 260 by rotating it horizontally about or with rod 250 through the air gap at a given angular speed, a voltage is induced which provides a deflection of galvanometer 270. Furthermore, when the conductor 260 is held fixed and the magnetic system 200 rotated through the same angular displacement about rod 250, at the same given angular speed, a similar voltage of exactly the same magnitude but opposite direction is induced in the galvanometer circuit. The same deflection is obtained whether or not the magnets 210 and 220 rotate with the yoke 230 or are held stationary while the yoke alone is rotated. This indicates clearly that the lines of magnetic induction in the quantity rotate about the axis of rotation 250 as if they were rigidly attached to the magnetic system 200.

A most important aspect of the operation of the apparatus of FIG. 11 is noted when the magnets 210 and 220 are held fixed in relation to the conductor 260 and rod 250 and the yoke 230 is rotated about the rod 250. This action causes induction of current in the galvanometer circuit exactly as occurs when the entire magnetic structure or the conductor itself are moved. However, if the magnetic yoke 230 and conductor 260 are held fixed and magnets 210 and 220 rotated, there is no deflection of the galvanometer. Thus, it appears that the yoke structure plays an important part in controlling movement of the magnetic flux. Furthermore, it should be noted that the magnetic flux does not link the galvanometer circuit in the usual sense that this term is used, but that a current is induced in the galvanometer circuit through the action of the conductor 260 in merely cutting across the field extending between adjacent faces of magnets 210 and 220.

It has also been found that the permanent magnets 210 and 220 may be replaced by an electromagnet and the results described above duplicated. Furthermore, it can be demonstrated that the magnetic flux within the air core of a solenoid may be rotated about its magnetic axis. A brass tube formed in the shape of a toroidal C has a winding applied to its entire length with holes drilled in the tubing to receive a brass rod, as 250 in the structure shown in FIG. 11. With the winding energized, the above-described procedures produce similar galvanometer deflections.

This application is concerned with certain methods and apparatus which make use of the phenomena described above.

Referring now to FIGS. 3, 4, and 5 for an illustration of the second embodiment, a tubular rotor 10 is provided on its inner periphery with 12 keystone cross-sectional electro magnets *11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21*, and *22* of the solenoid type. As seen in the cross-section of FIG. 5, the several electro magnets are arranged as sectors of a cylindrical annulus. While in the form shown, the several electro magnets 11–22 are shown in circumferentially wedged relationship, whereby to minimize the likelihood of relative movement radially inward as at standstill, as well as in the interest of achieving substantial uniformity of external flux density throughout the inside circular area of the rotor, it will be understood that when and if desired, the several electro magnets can be circumferentially spaced one from the other, and, if desired, a *spacer of relatively low, if any, magnetic permeability interposed* between them. Any suitable means may be employed for securing the several electro magnets to the shell of the rotor.

In order to facilitate the assembly of the several keystone-shaped electro magnets on the interior of the rotor as shown, the shell thereof is made of at least two pieces 23 and 24. In the form shown in FIGS. 3–5, the two pieces 23 and 24 are cup-shaped, and are respectively provided with outwardly extending flanges 25 and 26. Once the several solenoids have been positioned within the longer shell piece 23, the shorter shell piece 24 may be applied as a cap so that flanges 25 and 26 abut, and may be appropriately secured together as by riveting, bolting, welding, or cementing.

The shell pieces 23 and 24, as well as their *interconnecting* means, may be formed of material having little or no magnetic permeability, or in cases where it is desired to shield the exterior from the escape of magnetic flux, they may be formed of material having a high magnetic permeability. Instead of dividing the rotor shell in twain axially, it may be divided in twain radially, in which event the two halves will have axially extending joints when assembled, and may be held in such assembled position by banding or other appropriate means capable of sustaining the centrifugal force to which the rotor will be subject in use.

At the end of *shell piece 23, remote from flange 25*, there is provided a pair of slip rings 27 and 28 which, in the event the shell piece is formed of electrically conductive material, may be separated from the shell piece by sub-rings 29 and 30, of appropriate insulating material. The slip ring 27 is connected through a conductor 31 to the free end on the inner course of the solenoid winding for each of the electro magnets 11–22. The slip ring 28 is connected trough a conductor 32 to the free end of the outer course of winding in each of the solenoids 11–22. All 12 solenoids are connected so that the direction of electric current flow is the same in all solenoids. The energizing current is supplied from an external source through conductors 33 and 34 and brushes 35 and 36, *in a manner well understood in the art.*

The rotor is mounted upon an appropriate frame 37, having opposite stationary stub axles in the form of ferrules 38 and 39 which project toward each other. On the projecting portion of each of ferrules 38 and 39, there is provided an inner race 40 and 41 for each of two ball bearing sets whose outer races 42 and 43 are appropriately mounted to the respective ends of the rotor 10, so that the rotor 10 is free-running, with respect to the frame 37, about an axis concentric with the rotor 10.

In the embodiment shown, the exterior cylindrical surface of rotor 10 serves as a pulley for a *drive* belt 44 which connects the rotor to a source of power 45, but any other suitable means of driving the rotor in rotation may be utilized.

In the form shown in FIGS. 3–5, the apparatus is intended to concentrate the magnetic flux of the several solenoids radially inward, and to facilitate this each of the solenoids is provided with a core 46 of material having high magnetic permeability. As shown in FIG. 4, the several cores 46 are of squat U-shape, and extend, for the most part, parallel to the axis of the rotor, but at each end the cores are curved so as to provide faces 47 and 48 addressed radially inward. To minimize the radially outward escape of magnetic flux from the several solenoids, they are preferably wound of ribbon-like conductor which is coated with any appropriate electrical insulation. The first course of winding begins at 49, and proceeds in helically wound fashion with the minimum of space between adjacent turns about core 46, to the opposite end thereof, whereupon, without interrupting the continuity of the conductor, the second course is wound helically over the first course, with the pitch of the helix reversed from that in the first course. In this way, any gaps between successive turns in the first course are overlapped by turns in the second course. While the drawings show only two courses of conductor about each core 46, it will be understood that, in practice, there may be many more courses.

With the several solenoids connected through the slip rings to a source of direct current as above-described, it will be understood that all the solenoids 11–22 have the same polar orientation with respect to the axis of rotation of the rotor. For example, all core ends 47 are North poles, and all core ends 48 are South poles.

A tube 50 of circular cross-section is mounted in fixed relationship within the respective ferrules 38 and 39. The tube 50 is preferably made of material having little or no magnetic permeability. The tube 50 is stationary, but the rotor 10 rotates about the axis of the tube 50.

A stationary solenoid 51 is fixedly mounted on the exterior of tube 50 in a position such as to fit, with clearance, into the bight of the squat U-shaped solenoids 11–22. Solenoid 51 may be *wound directly* on tube 50, but in the form shown, it is wound on a cylindrical core 52 of material having little or no magnetic permeability, and the core 52 is telescoped on tube 50. The stationary solenoid 51 is energized through appropriate leads (not shown) from conductors 33 and 34 or other source of direct current, and is so connected that its magnetic polarity is opposite that of solenoids 11–22, that is to say that when, as aforesaid, the core ends 47 are North poles, the adjacent end of solenoid 51 will be its South pole. Thus, magnetic flux of solenoids 11–22 has an external path, through tube 50, which is coincident with, and in the same direction as, the internal path of magnetic flux in solenoid 51.

Thus, the magnetic field generated by each of the solenoids 11 through 22, as well as that generated by solenoid 51, penetrates the interior of tube 50, and when the rotor is driven in rotation, these magnetic fields rotate collectively about the axis of tube 50. During such rotation, particles of matter within the tube 50 are within the motional electric field generated by the rotating magnetic flux. Such an apparatus has a demonstrable effect in the separation and concentration of fluids. For example, in the de-salinization of water, the $B \times V$ electric field generated by the rotating magnetic fields, acts upon the disassociated ions of the sodium chloride in solution. This is explained by the fact that if the rotor 10 be rotated in the direction shown by the arrows in FIG. 5, the $B \times V$ field tends to drive the positively charged sodium cations toward the center, and tends to drive the negatively charged chlorine anions away from the center of tube 50. Hence, by introducing a flow of saline water at end 54 of tube 50, and by providing a concentric separating tube 55 at end 56 of tube 50, increments of fluid which are concentrated with sodium and depleted of chlorine may be drawn off through the stationary tube 55, while the increments of fluid which are concentrated with chlorine and depleted of sodium may be drawn off through the space between the exterior walls of tube 55 and the interior walls of tube 50.

Another utility of the apparatus shown is that of transmuting hydrogen into helium, in which event tube 55 is omitted and electrodes are provided at opposite ends 54 and 56 of tube 50, so that an electric arc may be established between them. Heretofore, difficulty has been encountered in maintaining the plasma of the arc in a relatively straight path between the electrodes as it ends to wiggle and extinguish itself when it makes contact with one of the confining walls. However, by impressing a $B \times V$ electric field upon such an arc, its path may be confined within controllable limits. The $B \times V$ electric field has the further effect of driving hydrogen protons toward the center of the tube, and driving electrons away from the center of the tube. Add one neutron to a hydrogen atom nucleus, and there is produced the isotope known as Dueterium; add one more neutron to the Dueterium nucleus, and it becomes Tritium; add one more proton to the Tritium nucleus, and it becomes Helium.

A third embodiment of the apparatus is shown in FIGS. 6, 7, and 8. With the significant exception that the embodiment of FIGS. 6, 7, and 8 has no moving mechanical parts, its organization is, in general, quite similar to that shown in FIGS. 4 and 5, and hence the last two digits of the reference characters utilized in the previous embodiment. As shown in FIG. 6, a cylindrical tube 150 has a solenoid 151 wound on the exterior thereof, and is energized with direct current from a suitable source through leads 153. A plurality of squat U-shaped solenoids 111, 112, and 113, of keystone-shaped cross-section, of which there are four each or any multiple of four, are wound as described in connection with the previous embodiment, and securely mounted in any suitable way with their pole faces 147 and 148 contiguous with, and addressed toward the axis of, tube 150. In this embodiment, however, the several squat U-shaped solenoids 111, 112, and 113, have cores which are made of material having little or not magnetic permeability, such as tubes 146 of paperboard wound to a keystone shape with hollow interior. As in the previous embodiment, the magnetic polarity of solenoid 151 is opposite that of the solenoids 111, 112, and 113, which is to say that if ends 147 of the solenoids 111–113 are the North poles thereof, the adjacent end of solenoid 151 is its South pole, so that the external magnetic flux path from solenoids 111–113 is coincident with, and in the same direction as, the internal flux path of solenoid 151.

In the embodiment shown in FIGS. 6–8, the several squat U-shaped solenoids 111, 112, and 113 are energized with non-interconnected three-phase alternating current, each phase of which has been half-wave rectified to produce direct current pulsating at alternate half cycles. One phase of the alternating current is connected to all solenoids 111; another phase of the alternating current is connected to all solenoids 112; and the third phase of the alternating current is connect to all solenoids 113, as shown in the wiring diagram of FIG. 8, where a three-phase alternator 100 is provided with six leads arranged in three pairs, to wit: leads 101 and 101' for the first phase, leads 102 and 102' for the second phase, and leads 103 and 103' for the third phase. The respective phases are not electrically interconnected. In each of leads 101, 102, and 103, there is a rectifier 104, 105, and 106, respectively. Leads 101 and 101' serve all four of solenoids 111; leads 102 and 102' serv all four of solenoids 112; and leads 103 and 103' serve all fou of solenoids 113. Across leads 101 and 101', between th rectifier 104 and the respective solenoids 111, there is pro vided a capacitor 107. Such capacitor is to coordinate th capacity and inductance of input, and to neutralize the effec of induction in solenoids 111 by adjacent solenoids during th off half-cycle of energization. Likewise, a capacitor 108 i connected between leads 102 and 102', and a capacitor 109 i connected between leads 103 and 103'. It will be apparent t those skilled in the art that since the four solenoids of like phase are arranged in quardrature with each other, and since solenoids 111 will reach their peak of magnetic flux 120 ahead of solenoids 112 reaching their peak flux, and the latte 120° ahead of solenoids 113 reaching their peak flux, a revolving magnetic field is created. Due to the rectification of the alternating current, and the consequent utilization of the half cycles which all flow in the same direction, there will be no reversal of the direction of magnetic flux generated by any of solenoids 111, 112, and 113, and hence, as in the case of the previous embodiment, the external path of all magnetic flux generated in the solenoids 111, 112, and 113 is coincident with, and in the same direction as, the internal flux generated by direct current energization of solenoid 151. When such an arrangement is supplied with alternating current at high frequency, such as a thousand cycles per second, the same results are achievable with this embodiment as with the previous embodiment but without movement of any mechanical parts.

For the purpose of illustrating the effect of the $B \times V$ electric field, and segregating that effect from the effect of stray magnetic or electrostatic forces, reference may be had to FIGS. 9 and 10. The device there shown is intended to be encased in an electrically grounded box 70 of material which has a high magnetic permeability, and which also has the property of electrical conductivity for shielding electrostatic fields.

Within the box 70, there is a compound rotor machine, wherein the respective rotors are driven in opposite directions. In a suitable frame having spaced pedestals 71 and 72, there is mounted a shaft 73 and an independent shaft 74. Both shafts 73 and 74 may be driven from a single source of power, and are preferably driven at the same rotational speed, but in opposite senses. A yoke 75 is fixedly mounted to shaft 73 for rotation with it. The yoke is cup-shaped with cylindrical skirt 76. About the inner periphery of the skirt 76, a plurality os solenoidal electro magnets 77 are fixedly mounted thereto. The several electro magnets extend parallel with the common axis of rotation of shafts 73 and 74, and are of the same polar orientation with respect to those shafts. The several electro magnets 77 have identical solenoids wound, in the manner previously described, about identical cores of material having high magnetic permeability. The solenoids of the several electro magnets 77 are energized from a suitable source of direct current through conductors 84 and 85, brushes 86 and 87, slip rings 88 and 89, and conductors 90 and 91. The respective solenoids 76 are preferably connected in parallel circuit relationship, but, if desired, may be in series so long as uniform polar orientation is maintained.

Shaft 74 is provided with a hub 78 having a coaxial bore 79, into which is rotatably fitted a pilot 80 projecting concentrically from shaft 73. On the outer periphery of hub 78, there is fixedly mounted a plurality of electro magnets 81, each of which is an identical solenoid, and each of which has an identical core of material having high magnetic permeability. As in the previous case, the several solenoids 81 are all energized so as to have the same polar orientation with respect to their axis of rotation. Each of the solenoids is energized from a suitable source of direct current through conductors 68 and 69, brushes 82 and 83, slip rings 95 and 96, and conductors 97 and 98. All of the solenoids 81 are preferably connected in series circuit relationship, but, if desired, may be in parallel.

If desired, and as shown, the several solenoids 81 may be surrounded by a simple solenoid 59 wound helically thereabout and electrically connected in parallel or series relationship with solenoids 81 and to conductors 97–98 so that the current flow through solenoid 99 is in the same direction as through solenoids 81. Thus the winding 99 serves the dual purpose of boosting the magnetic fields generated by solenoids 81 and of mechanically binding the latter to the surface of hub 78.

When, as contemplated, the shafts 73 and 74 are driven in rotation simultaneously, but in opposite directions, the solenoids 77 which rotate with shaft 73 have opposite magnetic polar orientation to those (81) which rotate with shaft 74. The several solenoids 77 are preferably designed to create external magnetic flux fields of flux density approximately equal and opposite to those created by electro magnets 81 and 99, but by the principle of superposition ("Elec. & Mag.", by Ralph W. Winch, 1963 Ed., p. 202), each flux field acts as if the other were not present. Despite the fact that the flux fields created by the *responsive series of electro magnets* 77 and 81 and 89 are oppositely directed, since their rotation is in opposite senses, their resultant effect, insofar as concerns the generation of a $B \times V$ electric field, is cumulative. Accordingly, the external $B \times V$ electric field is intensified at the exterior of the device, and its effect on the outside of shielding box 70 is demonstrable, even though the resultant magnetic flux density there is approximately zero. Such demonstration might be through a capacitor and electrometer as shown in FIG. 9, for example.

In order to achieve the maximum flux density and hence the maximum $B \times V$ electric field intensity, it is desirable to take measures which keep the apparatus cool. This may involve resort to such means as: winding the several solenoids with so-called "super-conductors" such as that known commercially as "Supercon"; or refrigerating the apparatus as by immersing it in a bath of liquid helium which is capable of maintaining a temperature of approximately 4° Absolute.

While several illustrative embodiments of the invention have been disclosed in detail, it is not to be understood that the invention is limited to those embodiments. On the contrary, the principles of the invention are susceptible of application in a vast variety of forms without departing from the spirit of the invention or the scope of the appended claims. The apparatus defined in the following claims is believed by the applicant to demonstrate his unique theory disclosed above.

What is claimed is:

1. Apparatus for generating an electric field, comprising:

a frame;

a cylindrical rotor;

a shaft journalled on the frame and supporting the rotor for rotation about is longitudinal axis;

a plurality of electro magnets, each of which is an identical solenoid having a core formed of a material having a high magnetic permeability, the electro magnets being fixedly mounted on the rotor with each electromagnet extending parallel to the axis of rotation thereof,

means to energize the solenoids of the electro magnets from a direct current source, the solenoids being energized so as to maintain a uniform polar orientation; and

an enclosure surrounding the frame and the rotor, the enclosure being formed of an electrically conductive material the enclosure being grounded.

2. Apparatus according to claim 1 wherein the rotor is a hollow cylindrical rotor, the electromagnets being affixed to the inner cylindrical surface of the rotor.

3. Apparatus according to claim 1 wherein the solenoids are connected in parallel circuit relationship.

4. Apparatus for demonstrating magnetic phenomena comprising:

a frame;

a cylindrical rotor;

a shaft journalled on the frame and supporting the rotor for rotation about its longitudinal axis;

a plurality of electro magnets, each of which is an identical solenoid having a core formed of a material having a high magnetic permeability, the electro magnets being fixedly mounted on the rotor with each electro magnet extending parallel to the axis of rotation thereof,

means to energize the solenoids of the electro magnets from a direct current source, the solenoids being energized so as to maintain a uniform polar orientation;

an enclosure surrounding the frame and the rotor, the enclosure being formed of an electrically conductive material, the enclosure being grounded; and

means surrounding at least a portion of the enclosure to detect and measure the electric field generated during rotation of the energized rotor.

5. Apparatus for demonstrating magnetic phenomena, comprising:

a U-shaped yoke of highly permeable magnetic material;

a pair of permanent magnets mounted in axial alignment between the ends of the yoke, one magnet being mounted to each leg of the yoke;

an electrically conductive non-magnetic rod extending axially through the magnets;

a stiff conductor mounted at one end to the rod for rotation relative to the yoke in a plane perpendicular to the rod and intermediate the magnets, the conductor extending a sufficient distance beyond the magnets so that its other end is located at a point where the magnetic field from the magnets falls to approximately zero intensity;

a galvanometer; and

circuit means connecting the ends of the conductor to the galvanometer to measure current flow through the conductor.

6. Apparatus according to claim 5 wherein the magnets are rotatably mounted to the yoke for rotation about the axis of the rod, and where the magnets are axially magnetized and adjacent faces are of opposite polarity.

* * * * *

APPENDIX III

A CONDITIONAL CRITERION FOR IDENTITY, LEADING TO A FOURTH LAW OF LOGIC

Summary

If logic is regarded as a set of perceptual operations, then logic has a chronotopology (time structure). Identity or nonidentity then results as a decision from an algorithm — a set of perceptual operations and comparisons — in which case the nature of a particular identity is conditional upon the nature of the set of perceptual operations comprising the algorithm.

Ordinary logic does not account for the temporal aspects of perception, merely accounting for the spatial aspects.

In other words, Aristotlean logic is a synthesis of primitive observation, fitted to the partial (spatial) reality emerging from spacetime after the imposition of the monocular (one-at-a-time) photon interaction with matter.

In quantum mechanics, time is a parameter, not an observable. Hence measurement/detection (of observables) deals with primitive observation and Aristotlean logic (topology).

Total reality includes nonprimitive observation — hence, non-Aristotlean logic (chronotopology) — as shown in Young's two-slit experiment.

By applying temporal accounting to each perceptual operation, Aristotle's three laws can be shown to be self-contradictory and incomplete as written. That is, they are topolological, not chronotopological.

A simple derivation of a fourth law is shown and an application rule given which itself may be regarded as a fifth law of logic.

A proof of the fourth law by demonstration is given.

The resulting four law logic is chronotopological. The applica-

tion rule states that either Aristotle's three laws apply explicitly and the fourth law is implicit, or the fourth law applies explicitly and Aristotle's three laws are implicit.

The four-law chronotopological logic is theoretically capable of resolving every present three-law paradox.

Aristotle's Laws and the Paradox of Change

Aristotle's three laws of logic, on which foundation rests all mathematical, physical, and rational thinking, can ordinarily be stated as shown in Table 1.

Table 1. Aristotle's three laws of logical thought

1. $A \equiv A$
2. $A \not\equiv \overline{A}$
3. $A \vee \overline{A}$

A variety of arguments can easily be produced to show that these laws are incomplete; i.e., they do not specify all reality, for parts of reality can be shown to contradict one or more of Aristotle's laws.

Indeed, all "observed" reality can be shown to violate all three-laws.

E.g., the most direct violation is posed by the problem of change, a problem originally propounded by Heraclitus about 500 B.C., and unsolved to this day. Heraclitus pointed out that, for a thing to change, it must turn into something else, and then asked how a thing could be something other than itself?

We may think of a thing — say, α — some feature A of which is said to change. If A changes, it turns into $\overline{A}$, thus violating logic laws one and two. Further, we are considering A as the "changed thing. $\overline{A}$," i.e., something which is somehow both A and $\overline{A}$, so logic law three is violated as well.

Thus, if Aristotle's laws are taken to be all the fundamental laws of logic, then logically there can be no change whatsoever, because change negates all three laws. I.e., either change does not exist or it is totally illogical.

Since all measurements, detections, thoughts, and perceptions are simply changes, then it follows that these operations logically cannot exist. Or, if we assume the "operations" to exist, their outputs cannot exist. If the operations do not exist, then again their outputs do not exist.

So if the products or outputs cannot exist, then by this reasoning no perceived, detected, measured, conceived thing exists. If we then insist that such things do indeed exist, then all is paradoxical and illogical. This is essentially the nature of the paradox posed by Heraclitus.

Heraclitus's change paradox has not been satisfactorily resolved to this day, and rigorously all the rational science of the Western world, being based on paradoxical change (detection, perception, observation) is itself totally illogical by its own logical standards.

Resolving the Paradox of Change

However, the conditions necessary to resolve the problem of change can be stated simply by inspection of the problem as follows: (1) Aristotle's three laws must specify or apply to only that which is not changing, since change violates or negates all three laws; (2) If change is to logically exist, there must exist at least a fourth law of logic, one which applies to change; (3) This fourth law must contain the negations of each of the first three laws, since change negates them; (4) To be consistent, in any particular logical case, either the three laws explicitly apply or the fourth law explicitly applies (i.e., either change explicitly exists in that particular case or it does not); (5) Since all four laws must apply at all times, then when the three laws apply explicitly, the fourth law must be implicit—and when the fourth law applies explicitly, the three laws must be implicit.

With the five stated conditions, a fourth axiom of logic can be written simply by writing down the negations of Aristotle's three laws, and synthesizing these negations into a single fourth law, as shown in Table 2.

Table 2. Negations to Aristotle's laws.

Aristotle's law	Negation	of negation*
$A \equiv A$	$A \not\equiv A$	$A \equiv \overline{A}$
$A \not\equiv \overline{A}$	$A \equiv \overline{A}$	$A \equiv \overline{A}$
$A \vee \overline{A}$	$A \wedge \overline{A}$	$A \equiv \overline{A}$ or $[A,\overline{A}]$

However, even though we can synthesize the negation into a single law — the old "identity of opposites" idea — we still have the problem of understanding such a law. Though at first glance the negations and the synthesized fourth law seem bewildering, we can readily comprehend them if we carefully consider the temporal nature of the process that occurs in logical thinking.

The Importance of Time

Specifically, a finite interval of time is required to perceive, think, detect, or observe an entity — regardless of whether we refer to "physical" or "mental" detection, because both physical and mental processes are temporal. Indeed, we flatly state without further discussion that ultimately the identifying or mapping of physical and mental operations onto each other is what time is a *priori*.

At any rate, we now carefully account for each individual time interval required to think, conceive, detect, perceive, or observe any entity — whether that entity is physical or mental — and we also account for the finite time interval required to perform a logical operation. So we rewrite Aristotle's three laws as shown in Table 3, with subscripted numbers indicating the separate time intervals in each law.

*These negations mean that A and $\overline{A}$ are totally undifferentiated.

Table 3. Temporally accounted laws and negations.

	Aristotle's law	Negation
1.	$A_1 \equiv_3 A_2$	$A_1 \equiv_3 \bar{A}_2$ or $A_1 \not\equiv_3 A_2$
2.	$A_1 \not\equiv_3 \bar{A}_2$	$A_1 \equiv_3 \bar{A}_2$
3.	$A_1 \vee_3 \bar{A}_2$	$A_1 \wedge_3 \bar{A}_2$ or $[A_1, \bar{A}_2]_3$

The resolution to the entire mystery so long inherent in these axioms of logic now stands simply revealed: Whether one of Aristotle's laws holds or its negation holds is determined solely by the nature of the logical operation in time interval three.[1]

I.e., the operation in interval three may be regarded as an *algorithm* comprised of subsidiary (assumed) operations in separate time subintervals that, taken together, comprise the overall operation implied by the logic symbol.

Thus in the first law, if temporal tags (time snapshots) are not accounted (i,e., if they do not apply), then Aristotle's laws hold, for the snapshot 1 of A is not differentiated in algorithm 3 from snapshot 2 of A. This then rigorously holds for spatial (L^3) entities, but not for spacetime entities. The snapshots in this case for Aristotle's first law (and the others as well) are *spatial* snapshots. On the other hand, if snapshots 1 and 2 of A are themselves temporally differentiated in algorithm 3, then the negation of Aristotle's law applies, because the *spacetime* snapshots A_1 and A_2 are different. This is immediately apparent, e.g., in a Minkowski geometry representation, where the second snapshot of A will have a time coordinate different from the time coordinate of snapshot 1. This is represented as shown in figure 1, where "A" is taken as a simple magnitude, in this case 5.

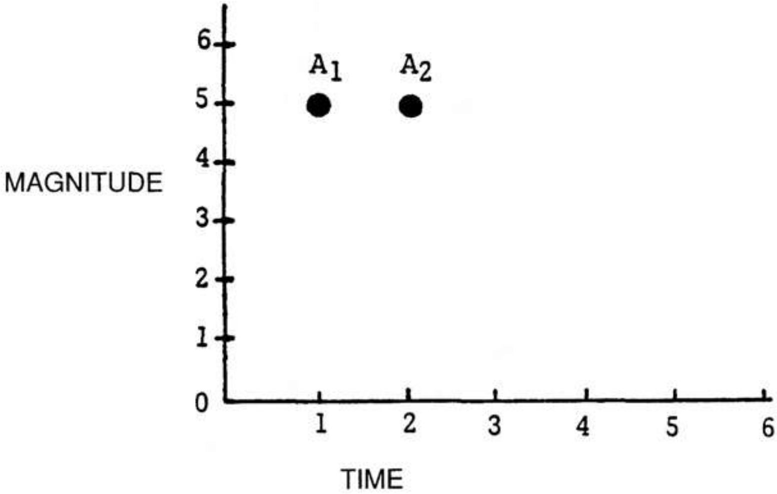

Figure 1. Spatial snapshots of spatiotemporal A differ.

As can be seen from figure 1, 5_1 is not identical to 5_2 *unless we imply the operator* $\partial / \partial t$ *in the time interval three algorithm.*

The negation of the second law may also be simply understood if we use temporal accounting. E.g., suppose we take $A_1 = +1$, $\overline{A}_2 = -1$ and then pose the absolute value operator || for potential use in algorithm 3. If || is not used, then

$$(+1)_1 \not\equiv_3 (-1)_2$$

and Aristotle's second law holds. If || is used, then

$$\left| (+1)_1 \right|_3 \equiv_3 \left| (-1)_2 \right|_3$$

and in that case

$$A_1 \equiv_3 \overline{A}_2$$

Note we are taking the view that there is nothing "absolute" or "inherent" about identity or non-identity; instead, each is a conditional result that can only be established by some logical, comparn tive set of operations. If the suboperations comprising the decision algorithm for the identity/non-identity determination are changed, the finding of the algorithm (the decision) may often change. Specifically, one can have the cases

$$\equiv_1 \not\equiv_3 \equiv_2$$

$$\not\equiv_1 \not\equiv_3 \not\equiv_2$$

Thus we advance a *conditional identity criterion* to be incorporated into formal logic: "Identity" or "non-identity" is defined by a decision made as a result of applying an operational algorithm; changing the internal operator components assumed inherent in the algorithm can change the decision. We are stating a fundamental principle that "identity" and "non-identity" are conditional and only conditional; they are never absolute.

Primitive Observation and "Reality"

With these points made, we now turn to the third law. From Table 2, on examination it can be seen that the third law actually is a statement for *monocular* perception, detection, observation, thought, or conception. Indeed, this law says that only a single thing at a time can be perceived, detected, observed, thought, or conceived.

As we pointed out in a previous paper[2] , there is a very good reason for this "law." Primitive man lived almost exclusively in a reality detected by light, by the photon interaction. Even in the absence of visible light, all bodies have temperature, and man is immersed in a "sea" of continual electromagnetic photon with interactions. The photon interaction is monocular - - only one at a time interacts with a particle of mass. Further, photon interaction constitutes the operator $\partial/\partial T$ invoked upon L^3T spacetime. Photon emission carries away time (the photon is made of $(\Delta E\ \Delta T)$, leaving behind an L^3 *spatial* reality, as we have previously pointed out.[3]

So all our primitive concepts, ideas, and notions about reality have come from over four million years of hominid and human experience in the photon-detected partial reality ("physical," "objective," or "spatial" reality) that remains when the time "dimension" (fundamental **variable)** is destroyed from L^3T spacetime, leaving only L^3 space behind. Specifically, our observed macroscopic reality consists of large temporal (mental) aggregates of such spatial results, where we cannot distinguish the tiny temporal separations of the pieces. Thus all our observed/perceived entities are spatial, and further, each perception/observation snapshot results in a frozen, unchanging spatial entity (resulting — in physical detection — from the so-called "collapse of the wave function." The loss of a wave function is simply the loss of time.) We vaguely sense "time" and "change" as the relation between these snapshots — i.e., by causality, or the ordering of the spatial changes — much as we see "movement" in movie frames rapidly projected onto a screen one-at-a-time.

Thus our primitive observations, from which have painfully been formed our relational concepts and ideas, are monocular, unchanging, and spatial. Aristotle's three laws of logic — which indeed may be taken to be only a simple synthesis of our primitive observation and corresponding relational concepts — then exhibit the same characteristics; they are monocular, unchanging, 3-dimensional, spatial, non-temporal relational statements. Any statement that is temporal, changing, or 4-dimensional will thus appear as a logical paradox to this logical shorthand.

But from Young's two-slit experiment, we already know that reality and relationships between its parts are quite different if the photon interaction is not invoked — i.e., if $\partial/\partial T$ is not invoked. Classical reality (as prescribed by Aristotle's laws) is directly violated by an electron in the two-slit experiment, e.g., if and only if photon interaction with the electron is not invoked. Again, this has been simply explained by the present author[4], and Charles Muses as early as 1957 pointed out the absence of any mystery in Young's experiment if the chronotopological aspects were considered.[5]

The Conscious Mind is Fitted to the Photon Interaction

However, what is normally referred to as the "conscious, thinking mind" is simply a functioning temporal (rigorously, chronotopological) mechanism that is painfully built up in the individual's awareness (his mind in the greater sense of both thought and awareness, whether monocular or multiocular) by training, conditioning and experience. Its functioning is largely conditioned by one's 90% or so attention to visual stimuli (to the partial reality remaining after photon interaction has been invoked, and to the memory-collated ordering of vast numbers of such photon interactions) and by one's cultural conditioning — which itself has been almost exclusively conditioned and shaped by the monocular photon interaction at base root.

Thus, since the beginning of man, his conscious, rational mind has been trained and constructed to function almost exclusively in basic correspondence with the photon interaction, and his experiential reality consists of the partial reality stripped from fundamental reality by photon interaction.

All "perceived differences," e.g., are created by this deep mind-set. As has been previously pointed out,[6] the solitary human problem responsible for all man's inhumanity to his fellow man is directly dependent upon man's almost exclusive detection, observation, perception, and conception of "difference" between humans, these "differences" being due exclusively and totally to the fitting of men's conscious minds to the photon interaction's monocular separation of spatial reality from nonspatial reality, i.e., to

$$\partial/\partial T\,(L^3T) \Rightarrow L^3$$

Such well-nigh total devotion to, and enslavement by, photon interaction also is responsible for the scientist's well-nigh total devotion to, and enslavement by, the present imperfect and incomplete three laws of logic, as presented by Aristotle. The depth of that devotion and enslavement is evidenced by the fact that the resolution of such paradoxes as Heraclitus's problem of change have eluded the best minds of humanity for several thousands of years. Indeed, these paradoxes cannot be resolved by the conscious, rational mind in its present state, for it has been most firmly constructed and fitted to function in accordance with the photon interaction.[7] One cannot hope to resolve any logical paradox by using only those same logical methods that found the situation to be paradoxical in the first place!

Dimensionality

That we need not be constrained by such universal delusion is already shown by binocular vision. Specifically, in viewing a three-dimensional **object, each** eye **never** detects a "third dimension," but detects only **an L^2 2-dimensional** picture. By taking two slightly different 2-dimension snapshots and superposing them, the third dimension is gained. One then essentially sees the resultant superposed pictures as "almost the same but not quite." I.e., the Aristotlean identity algorithm, if satisfied, yields "no difference, hence one object," and if not satisfied, yields "difference, hence multiple (extended) object(s)." So if the two snapshots are almost Aristotlean-identical but not quite, we get an extended-two-dimensional (three-dimensional) object. Otherwise we see two separate, two-dimensional snapshots blurred together (the reader is urged to try this and see).

The point is, "dimensionality" and the identity algorithm are directly related, and geometrically one follows from the other.

Awareness of Time

In very similar manner, we can only gain cognizance of awareness of "time" (as a fourth dimension) by the superposing of two slightly differing (Aristotlean-wise) 3-dimensional snapshots. As is well known, e.g., time is not an "observable" in quantum mechanics; it is a "parameter." Rigorously, the only place such snapshots — each of which is "past" (spatially separated by the annihilation of time in the collapse of the wave function) — can multiply exist is in the "mind" in its most general sense. In a rigorous sense, mind and time can be taken as identical, and the "flow of time" can be taken as the "flow of mind connections or superpositions" of its spatial components. I.e., measured/detected/observed "physical phenomena" are a priori 3-dimensional and spatial, while a mind is four-dimensional and hyperspatial.[8] Spacetime exists mentally but not observably. Time is logically implicit, not explicit.

Temporal Aspects of Logic

Now we return to the temporal aspects of logic. Each perceptual part of each Aristotlean law is fitted to the photon interaction, hence monocular. The logic operation, inherent in the logic symbol in each statement, involves temporal superposition or comparison of spatial perceptual objects. Hence the logic operation is hyperdimensionally a function of mind and injects mind/time into the statement. Yet these laws, being fitted to or synthesizing photon interaction, attempt to prescribe the absence of time, even though writing down the logic operation rigorously invokes time. They are thus totally contradictory, since as written they implicitly violate themselves.

Table 4 summarizes the operations now to be permitted in the time-three algorithm, in developing a new four-law logic:

Table 4. Conditional identity rules.

1. $\boxed{S_1, S_2}_3 = (\text{zero})_3 \Rightarrow (\text{identity})_3$

2. $\boxed{S_1, S_2}_3 = (\text{nonzero})_3 \Rightarrow (\text{non-identity})_3$

3. $\boxed{S}_3 = \left| \boxed{S}_1 , \boxed{S}_2 \right|_3 \Rightarrow$ **monocular separation, differentiation internally**

4. $\overset{S}{\sqcap}_3 = \overset{S}{\sqcap}_3 \Rightarrow$ **nonseparation, no differentiation internally**

Rules one and two simply state that, when snapshots 1 and 2 are superimposed (subtracted) in time interval 3, the resultant snapshot 3 may be zero or nonzero. If zero, snapshots 1 and 2 are said to be identical, and if snapshot 1 is to be labeled A, then snapshot 2 is to be labeled A. If snapshot 3 is nonzero, snapshots 1 and 2 are said to be nonidentical; if snapshot 1 is labeled A, then snapshot 2 is labeled $\overline{A}$.

Rule 3 says that snapshot 3 is a "memory" snapshot, and it may be particulately examined to monocularly separate snapshots 1 and 2.

Rule 4 states that snapshot 3 is not a "memory snapshot" and may not be further separated.

Note that in logic we repeatedly apply these rules in combination, serially or compositely. Note further that Rule 1 must serially apply both rules 3 and 4, as must rule 2 also.

If we take $4\,\big|_{3}$ to mean "rule 4 applied conditional to rule 3 also being applied," and $4\,\big|_{\bar{3}}$ to mean "rule 4 applied conditional to rule 3 not also being applied," we may write Table 5:

Table 5. Conditional identity, non-identity, and oneness.

$4\,\big|_{\bar{3}} \wedge 1 =$ identity (S_1, S_2)

$4\,\big|_{\bar{3}} \wedge 2 =$ non-identity (S_1, S_2)

$4\,\big|_{\bar{3}} =$ oneness without separate-ones; oneness, extraordinary and unperceivable; thing-in-itself

$4\,\big|_{3} =$ “ordinary” one, perceivable separation, “thing-as-separate-from-others”

As can be seen, this type of reasoning also sheds a great deal of light on the long-standing problem of the "thing-in-itself," but that is beyond the scope of this paper.

The Fourth Law of Logic

Now we write the fourth law of logic as follows:

$$4.\ \ (A_1\,, \overline{A}_2\,)_3 \ \Rightarrow\ A_1 \equiv_3 \overline{A}_2$$

where all we have said is that, by rule 4, in snapshot 3 no memory process is allowed, and no separation/differentiation whatsoever of A_1 and $\overline{A}_2$ is permitted. Under these operational conditions for identity, what had previously been called A_1 in snapshot 1 and what had been called $\overline{A}_2$ in snapshot 2 are indistinguishable, hence identical.

Figure 2 shows this concretely, and may be taken as a proof of the fourth law by demonstration.

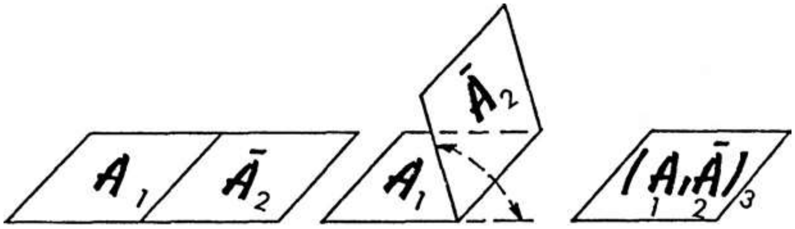

Figure 2. Proof that two opposites can be identified. What was separately perceived as A_1 in time snapshot 1 and what was separately perceived as $\bar{A}_2$ in time snapshot 2, cannot be distinguished in time snapshot 3.

Thus the age-old philosophical dilemma posed by the illogical identity of opposites has a simple resolution if one considers temporal aspects, and introduces temporal conditions for identity or non-identity decisions.

We now write the new four law conditional identity logic as shown in Table 6:

1. $A_1 \equiv_3 A_2$
2. $A \not\equiv_3 \bar{A}_2$
3. $A_1 \vee_3 \bar{A}_2$
4. $A_1 \equiv_3 \bar{A}_2$

Further, we point out that all four laws now apply. Laws 1,2, and 3 are the laws of explicit monocular perception, with implicit binocular perception. Law four is the law of explicit binocular perception, with implicit monocular perception. Both monocular and binocular perceptions must be and are used in each law. So in any situation, either the triad applies explicitly and the fourth law applies implicitly, or the fourth law applies explicitly and the triad applies implicitly.

Indeed, one can **even take the** view that we have prescribed a five-law logic, **the fifth la**w **being taken** as shown in Table 7:

Table 7. A possible fifth law of logic.

$$5.\ [1,2,3,\ \wedge(4)] \vee [4 \wedge (1,2,3,)] \text{ where } () \Rightarrow \text{implicit}$$

In a previous paper[9], the author has already presented methods to apply this new logic to resolve present paradoxes. At least hypothetically, every present paradox should be simply a statement of the explicit fourth law, and it should be resolvable by explicit application of that law[10].

NOTES AND REFERENCES

1. Specifically, by whether or not exclusivity applies. I.e., we may read Law 1 as "In snapshot (time interval) 3, what was A in snapshot 1 is exclusively identical to (unseparated from) what was A in snapshot 2." Law 2 may be read as "In snapshot 3, what was A in snapshot 1 is exclusively not identical to (is exclusively separated from) what was not-A in snapshot 2." Law 3 reads, "In snapshot 3, what was A in snapshot 1 and what was not-A in snapshot two are exclusively separated." Thus it can be seen that the three laws simply are statements involving whether or not two former perceptions are to be separated in a third perception. These three statements presently prescribe the total separation of the two previous perceptions and prohibit any admixture of the two — the so-called "excluded middle." Thus the three laws prescribe monocular, one-at-a-time perception.

   What we call a "wave" exists in time and is considered to contain an admixture of timeless, static spatial states (such as "wavelengths.") The prohibiting of admixtures thus represents the "collapse of the wave function" and corresponding loss of time. This defines "observation" and explains why all "measurement" and "detection" and "observation" — requiring a collapse of said wave function — are spatial and not spatiotemporal. That is why time is a parameter in quantum mechanics, not an observable.

   Relativity, being constrained by such logic, obviously can find nothing "physical" (observed, spatial, timeless thing frozen by the collapse of the wave function as engendered by or fitted to the photon interaction) that is traveling faster than light — i.e., that violates the conditions implied by the "observing/detecting agent."

That reality need not at all be so constrained is clearly shown by Young's two-slit experiment, the heart of all quantum mechanics. In this experiment, "classical' reality is violated if and only if the photon interaction is not invoked. Classical reality is obeyed if the photon interaction is invoked.
With appropriate change to logic to fit "reality that has not been interacted with by photons" and therefore is spatiotemporal, a new physics becomes possible.

2. Thomas E. Bearden, "Solution of the Fundamental Problem of Quantum Mechanics," January 3,1977, Defense Documentation Center.

3. Bearden, "Photon Quenching of the Paranormal (Time) Channel: A Brief Note," 20 April 1977, Defense Documentation Center.

4. Bearden, "Virtual State Engineering and Its Implications," 1979, Defense Documentation Center.

5. With the possible exception of Kozyrev — whose more technical works on time remain undisclosed to open science — no other person known to this author seems to have grasped the implications of a dynamic structure of time as penetratingly as has Charles Muses. The importance of the time interaction explaining the two-slit experiment (i.e., in explaining wave-particle duality) was noted as early as 1957 by Muses. See, e.g., Muses' introduction to Jerome Rothstein's **Communication, Organization, and Science,** The Falcon's Wing Press, Indian Hills, Colorado, 1958, p. lxii, where Muses pointed out that the celebrated wave-particle paradox remains a paradox only so long as the chronotopological (his word) phases of the phenomena are left unrealized in the analysis. The entire foreword by Muses is a remarkable document which analyzes the structure of time itself. With his hypernumbers Muses can describe the nested structure of time, which is what is actually

being carried by the photon. Further, he can theoretically predict mechanisms by means of which these structures can be orthorotated. It would appear that practical devices should be constructable on the principles elucidated by Muses, and it is little short of astonishing that fundamental work of such importance and application has been thus far little used by theorists, though it is already recognized in the standard literature; e.g., the profound summary paper "Hypernumbers II" in the January 1978 issue of the journal **Applied Mathematics and Computation,** published by Elsevier.

6. Bearden, "The One Human Problem Its Solution, and Its Relation to UFO Phenomena," Defense Documentation Center, January 3, 1977.

7. Which is why a Zen master often gives the student a koan to confound and overwhelm this automatic, robotic mindset and functioning that has been constructed as the student's "conscious mind."

8. Specifically, consciousness/life involves a seven dimensional body/being in an infinite-dimensional universe. See Thomas E. Bearden, "A Mind/Brain/Matter Model Consistent with Quantum Physics and UFO Phenomena," prepared for the 1979 MUFON Annual Symposium, available in the **Proceedings,** MUFON, 103 Oldtown Road, Sequin, Texas 78155.

9. Bearden, "The Fourth Law of Logic," **Specula,** Journal of the American Association of Meta-Science, P.O. Box 1182, Huntsville, Alabama 35807, Vol. 2, No.l, January-March 1979, pp. 30-40; also in publication in Defense Documentation Center. (The journal **Specula** is no longer pulished).

10. No consideration of the foundations of logic and mathematics is complete until one has read Morris Kline, **Mathematics: The Loss of Certainty,** Oxford University Press, New York, 1980.

APPENDIX IV

LIST OF SELECTED INCIDENTS

Author's note: In 1985-1987 a series of anomalous missile and space failures swept the U.S. and the West. These incidents were not unique, as many think, but were actually only a portion of thousands of related incidents showing the Soviet testing of advanced weapons throughout the world. This listing will show the types of incidents referred to.

The listing is not intended to be complete at all, but only representative. Hundreds of incidents of weather engineering, giant cloud radials and grids, suspicious light incidents, etc. could easily be added, were there time and room. No attempt has been made to exhaustively map the "aerial booms" that have blanketed the U.S., Europe, and other parts of the world. The light and "maybe it's UFO's" reports alone would add thousands of incidents to the list.

Our message is succinct: The Soviet Union has developed a new science and weaponized it, since at least the latter 1940's. It has utilized that science — energetics — to develop superweapons and supervehicles. Because of the danger from accidental initiation of nuclear weapons and materials, it is exceedingly difficult and dangerous for the Soviets to make extensive use of these weapons. Accordingly, a biological warfare strike — AIDS — has been utilized to deliver what is to be the knockout blow to the West.

DATE, INCIDENT, LOCATION, REMARKS

Early spring 1966 - Malstrom AFB Montana.
UFO sighting coincident with simultaneous problems in 10 strategic missiles in launch site associated with Malstrom AFB, Montana. *Similar event in Mar 1967 also. (Source:* ***Clear Intent.).***

20 **Mar** 67 - Malstrom AFB Montana.
Flight of 10 strategic missiles in launch site associated with Malstrom AFB, Montana experienced problems with guidance and control system. UFO in area and confirmed on radar. Jet fighters scrambled, results unknown. *Similar event one year earlier.* (Source: **Clear Intent.).**

1972 - Prague.
At a secret meeting of European communist leaders in Prague, Brezhnev lays out firm statement of Soviet intent to be able to dominate the world by 1985. Stated would control the oceans, 90% of the land, and the air and space above. *Notice he said control, not invade and conquer.*

Aug 73 - Over ocean.
USAF Minuteman ICBM is launched from Vandenberg AFB, aimed for Kwajalein missile range. Unidentified object appears and is tracked next to the ICBM's nose cone. Crossed the RVs trajectory. Object about 10 ft long. Seen by 2 separate radars. Three other identical objects were seen in the vicinity. *Probable Soviet advanced flying vehicles. Also a deliberate stimulus to see if U.S. knows of the technology, and how the U.S. reacts. (Source:* ***Clear Intent.).***

18 **Oct 73 -** Near Mansfield, Ohio.
U.S. Army helicopter is intercepted by glowing red object that converged on it with terrific speed. Object placed some sort of "tractor beam" on helicopter. Object was gray metallic structure, 50-60 ft long. Helicopter was stopped in mid air. Radio blacked out. Official report filed by pilot. *Probable advanced Soviet flying vehicle. Also a deliberate stimulus to see if U.S. knows of the technology, and how the U.S. reacts.*

21 Oct 73 - Near St. Joe, Indiana.
20 or 30 lights sighted circling in the sky near St. Joe, Indiana near a woods fire, by firemen. *Probable Soviet holographic tests to stimulate UFO reports.*

1973 to date - Bennett Island.
Bennett Is. exhaust plumes detected by U.S. weather satellites. *Well over 100. spotted so far. Weirdly, 5 islands and part of the continental Alaskan area are in process of being secretly given to the USSR; Bennett Island is one of them.*

May 1974 - Lake Michigan.
Blue-green glowing ball of light seen to fall into Lake Michigan. *Soviet scalar EM testing.*

8 July 74 - Lake Okeechobee, Florida and Atlantic Ocean.
Orange red ball of light lights up sky over south Florida. Crashes in area of Lake Okeechobee. Was accompanied by a large boom. Several other booms 5-7 min. after "object crashed." Pilots also saw object plummeting in from over Atlantic. *Probable Soviet scalar EM weapons testing and calibration.*

Nov. 74 - Saryshagan. USSR.
Construction of Tora facility begins at Soviets' Saryshagan test facility. Believed to be directed energy facility. *Probably connected with Soviet scalar EM weapons systems development.*

Oct? 74 - Antarctic.
Ozone hole over Antarctic first detected. Did not appear in measurements from 1957-1973. Note did not appear until exhaust plumes noted from and around Bennett Island, with substantial activity. *Looms over the continent every Oct. Diminishes through Mar. More severe each year since 1974.*

1975 - USSR.
Soviet article in **International Life** speaks of weather war, changing the nature of lightning, increasing the power of lightning, and directing electric charges of tremendous power against specified targets. "Atmospheric electricity" can be used to suppress mental activity of large groups of people. *Tip of the iceberg. Unclassified reference to Soviet's scalar EM (energetics) weapon program. Note involvement of "lightning" in destruction of U.S. missiles a decade later.*

16 **Feb 75 -** Caribbean.
R.M.S. Carmania (UK) observes bright white circular light appear and rise to 20 degrees, leaving comet-like trail, circle and disappear. Repeated four more times at exact hourly intervals. *Soviet scalar EM weapons tests. May have been a shipborne device.*

13 June 75 - Kremlin.
Brezhnev calls for ban on frightful weapons of mass destruction. *Major speech. Repeated proposals to U.S. Senators.*

2 **Jul 75** - Kremlin.
Brezhnev repeats his proposed ban on development of frightful new weapons, to a group of U.S. Senators. *To visiting U.S. Senators. No one knew what he was talking about.*

Aug. 1975 - Kremlin.
Ponomarev calls for ban on frightful new weapons of mass destruction. *To visiting Congressmen.*

23 Sep 1975 - United Nations.
Gromyko presents draft agreement to UN General assembly, for banning development of frightful new weapons. *30th session of UN Gen. Assy. No one knew what the Soviets meant.*

Oct - **Dec 75** - Satellite, over the Indian Ocean.
U.S. satellites illuminated or "blinded" over the Indian ocean. *Source 10-1,000 times as strong as natural sources. 5 incidents. One 4 hrs. long. Soviets wasted no time after Sep. fiasco.*

28 **Oct** to **11** Nov **75** - Mid-U.S. and Canada.
Series of "UFO's" seen at Malstrom AFB, Montana; Loring AFB, Maine; Minot AFB, North Dakota; Wurtsmith AFB, Michigan and Canadian Forces Station, Falconbridge, Ontario, Canada by reliable military personnel. Reported to NORAD. Some objects seemed like helicopters. Radar tracked objects. One object at 72,000 ft. was 100-ft sphere, with seeming "craters" around it.

Malstrom F-106's scrambled on one object but could not close on it in darkness and low altitude. Efforts by Air Guard helicopters, SAC helicopters and NORAD F-106's were fruitless. Probable Soviet activity in and around the SAC bases and Canadian radar site. (Source: **Clear Intent.).**

30 Oct 75 - Wurtsmith AFB, Michigan.
UFO's and "unidentified helicopters" sighted (multiple incidents) at or near Wurtsmith Air Force Base, Michigan. Helicopter hovered over the weapons storage area. Radar reported low-flying objects. Returning KC-135 tanker a/c diverted to intercept one UFO. Observed two objects, unable to close. Visual and radar contacts. Low on fuel, KC-135 returned to base to land, again seeing lights over the weapons storage area. (Source: **Clear Intent.)** *Soviet Activity.*

7 **Nov** 75 - Near Lewiston, Montana.
Giant glowing disc, football field sized, hovered over Minuteman missile launch site K-7, just south of Lewiston, Montana. When it rose to 1,000 ft., NORAD radar picked it up. Two F-106's were scrambled. UFO rose to 200,000 ft. and disappeared from NORAD radars. *Probable advanced Soviet flying vehicle, using antigravity propulsion system.* (Source: **Clear Intent.).**

Nov 75 - Semipalatinsk, USSR.
Large amounts of gaseous hydrogen with traces of tritium begin to be detected by USAF/TRW early warning satellites.
Nuclear debris apparently related to nuclear facility at Semipalatinsk labelled PNUTS (possible nuclear underground test site.) Probable large scalar EM development and test site.

Jan 76 - Cannon AFB, New Mexico.
Two UFO's reported near flight line at Cannon AFB, New Mexico. 25 yards in dia., gold or silver in color with blue light on top, hole in the middle and red light on bottom. Observed and reported by security police. (Source: **Clear Intent.).** *Probable Soviet activity.*

31 **Jan 76** - Eglin AFB, Florida.
MG Lane, CG, Armament and Development Test Center, Eglin AFB, Florida reported a UFO sighting near an Air Force Radar site from 0430 to 0600 Eastern Standard Time. Photographs were taken. Later press release called it "building lights." (Source: **Clear Intent.)** *Probable Soviet Scalar EM activity.*

1976 - U.S. Embassy, Moscow.
State Dept. declares U.S. Embassy in Moscow an unhealthful post. Installs Aluminum window screens against Soviet microwave radiation, which has been underway since 1950's.
Provided 20% pay extra differential. Embassy radiated with weak microwave radiation since 1950's. Two U.S. Ambassadors died, Stoessel ill with strange leukemia-like malady. Soviet stimulus to see if we knew about scalar electromagnetics, phase conjugate waves, and EM disease induction. (Stoessel died recently of cancer).

1976-U.S.
Soviet nuclear scientist named Rudakov visits U.S., proposes cooperative effort in some a**reas of** fusion research. Showed work he **had**

been doing and results obtained. U.S. officials reacted drastically, classified his talk as top secret, even ripped off blackboard and carried it away.*Very strange incident. Soviet open, public scientific work considered Top Secret by the U.S. One wonders why.*

30 **Mar** 76 - Netherlands.
Huge atmospheric boom occurs over northern part of Netherlands. *Through 1983, in 10 years 7 events took place.*

22 Jun 1976 - North Atlantic.
Expanding white sphere of light developed for about 10 min. then faded. Other associated light phenomena observed. *Seen and logged by passing ship.*

4 **Jul** 76 - Soviet Union.
Giant Soviet Woodpecker transmitters activated. *3-30 MHz band. Worldwide interference with communications.*

30 **Jul** 76 - Fort Ritchie, Maryland.
Several UFO's sighted in and around Ft. Ritchie, Maryland. One hovered over the ammunition storage area. Reported. *Temperature inversion at the time provided a convenient pseudoexplanation. Could not have caused the incidents.* (Source: **Clear Intent.)**

10 **Sep** 76 - Lithuania.
European Airways flight 831, Moscow to London, at 31,000 ft over Lithuania sees blinding single ball source of light at constant altitude about 5-6,000 ft. below them and above a lower cloud. *Pilot contacted Soviet authorities. Received negative response and suggestions he should not ask questions. Probable deliberate stimulus of flight so, from reports to British authorities, could ascertain whether Britain knew anything about scalar EM weaponry.*

19 Sep 76 - Near Teheran, Iran.
In early morning hours, UFOs were spotted north of Teheran, Iran. Two F-4's attempted to intercept. The first lost all communications and instrumentation at 25 miles distance. When the F-4 turned away, equipment was restored to functioning. Second F-4 acquired radar lock at 27 nm. and attempted to close. Object moved to stay at 25 nm distance. Second object came out of it, headed straight for the F-4. F-4 attempted to fire AIM-9 missile and at that instant weapons

control panel and communications went off. F-4 took evasive action. A third object was seen to emerge from first. The F-4 returned to base. Each time it passed through 150 degrees magnetic bearing, it lost communications and INS fluttered 30 to 50 degrees. A civilian airliner also experienced loss of communications in the vicinity. *Signatures of scalar electromagnetic weapons and antigravity.*

1976 - Soviet Union, Sweden.
Swedish national authorities detect anomalous radionuclides — similar to those from nuclear explosions — over Sweden. No accompanying seismic activity. Correlated with known Soviet activity at Semipalatinsk. *In other words, nuclear explosions were occurring, but they did not shake the earth. Total anomaly. Indicates transformation of the explosive energy. Positive signature.*

1976 - Philadelphia.
Legionnaires Disease precipitously strikes Legionnaires attending convention in downtown Philadelphia. Kills 34, makes 187 others seriously ill. Similar malady struck Odd Fellows convention at same hotel in 1974. *Odd person with apparent transmitter and antenna was expelled. Informed members: "It's too late! You are all dead!" Probable scalar EM conditioning of convention with an immune suppression pattern for the disease.*

21 **Jan 77 -** Columbia.
Columbia A/L Avianca night flight. Instruments went haywire. Saw enormous, zigzagging glowing light, traveled at fantastic speed back and forth. Blinked landing lights, UFO responded. Sped away at incredible speed. *Probable Soviet advanced performance antigravity flying vehicle. Such have been widely sighted over South America.*

3 **Mar 77 -** Columbia.
Columbia A/L Avianca night flight. Pilot abruptly changed course to circle a huge, glowing UFO. Object had cigar-shaped fuselage and was twice as big as the aircraft. Blinked lights, UFO responded. Sped away. Air traffic control radar tracked object at ten times the speed of normal aircraft.
Probable Soviet advanced performance antigravity flying vehicle. Such have been widely sighted in South America.

24 **Mar 77 -** Off coast of Spanish Sahara, Africa.
M.V. Kinpurnie Castle, off coast of Spanish Sahara, Africa, observes complex, dynamic light phenomena. Third phenomenon was the formation of a semi-circular area of moderate luminosity, a hemisphere of light. Took 3 min to form about a previous luminous patch. Then another luminous patch formed over the hemisphere. After about 7 min., phenomena dispersed. *Soviet test of Tesla shield and globes.*

11 Jun 77 - Moscow, USSR.
Soviet biophysicist hands Robert Toth, Los Angeles Times journalist, a report containing proof that certain particles existed in cells, carried information, and were radiated. Both seized by KGB and charged with possessing Soviet state secrets. Interrogated, released. *Neither knew what they actually had. Soviet Academician I.M. Mikhailov pronounced secret such phrases as "micro-organism self-radiation... by means of vacuum particles in space." Use of such particles was discussed, and was secret. Paper contained basis of Soviet energetics use in biology. Positive connection of Soviet classification to mitogenetic radiation and scalar EM waves ("vacuum particles" indicate virtual flux structuring).*

13 July 77 - New York.
Massive electrical blackout strikes New York City. Affects 8 million persons. Staggering series of "natural" electrical shocks, mechanical failures, and human errors occurred. Anomalous "lightning strokes" supposedly initiated the collapse of the power grid. *Probable Soviet test.*

1977 - Ocean off Yugoslavia.
M.V. Dolphin, between Yugoslavia and Israel, observed extensive light balls and repeated phenomena, appearing and disappearing. Men were disoriented. Compass was disrupted. Okay after the phenomena dispersed. *Probable Soviet test. Stimulus to see if Britain recognized the weaponry and knew anything about it.*

13-14 Dec 1977 - Leitchville, Australia.
Large aerial light plus 3 smaller ones observed for four hours by multiple persons. Lights simulated activity of a "mother ship" and smaller craft assigned. Changed colors. Then a shower of orange

lights occurred. *Probable Soviet test. Part of deception plan, to create "UFO" phenomena to hide weapons testing. TR wave holography.*

Dec 77 to **Jan 78 -** Off East U.S. Coast.
Mysterious atmospheric explosions occur off U.S. Atlantic Coast. *Some accompanied by light phenomena. Soviet tests and calibration (registering the long-range artillery, so to speak.)*

13-14 Dec 77 - Leitchville, Australia.
Large arrays of light phenomena sighted by multiple witnesses. *Probable Soviet creation. Reported as "UFO" phenomena. Holographic creation of "UFO" phenomena as part of deception plan to disguise tests of the weapons worldwide. TR wave holography.*

1977-78 - North Pacific Ocean.
Two huge hot spots in North Pacific ocean, 2,000 mi by 1,200 mi. *May have caused El Nino and one of most severe winters on record. May have been caused by Soviets as a weather experiment.*

27 **Jan 78 -** U.S. East Coast.
Boeing 727 off east U.S. coast experiences engines failing, 1 at a time, in all three engines. A/c plunges 8,000 feet; engines restart one at a time. *First time ever that all three engines on a 727 had failed. Probable Soviet test.*

Feb 78 - Rocky Mtns. NW of Las Vegas.
Four aircraft crash in a 6-day period in the rugged Rockies, NW of Las Vegas. *Connected with a crashed military a/c from Nellis AF Base. Should be further investigated.*

Feb 78 - Texas coast. Nova Scotia. Charleston, SC.
Atmospheric booms occur. *Multiple booms in Charleston. Soviet tests. Scalar EM weapons.*

2 **Apr 78 -** Bell Island, Nova Scotia.
Mysterious fireball, explosion accompanied by anomalous electrical effects. *Small bldgs blown apart, trees scorched, fireballs seen. Probable Soviet test. Apparently even registered on U.S. satellite.*

10 **May** 78 -Off Florida.
Women pilots in race experience anomalous "time loss" on their watches from Fort Myers to Freeport (Bahamas). *39pilots. Watches synchronized. Up to 3 min. lost. Definitely not normal EM effect. Probably due to sporadic natural scalar interferometry and time-reversed waves generated in area from time to time. "Bermuda Triangle" mechanism?*

May 78 - Florida.
Unexplained lighted objects sighted in sky over remote sector of Florida. Tracked on radar at a Navy EW center. Naval personnel also observed them through binoculars. *Probable Soviet advanced flying vehicles, or tests of intense holograms with energy sufficient to ionize air and reflect radar.*

18 **May** 78 - England.
Several atmospheric explosions shake houses over a wide area in Hull, Scunthorpe, Holderness coast, and Brough and Grimsby. *Probable Soviet weapon testing and calibration.*

Jun 78 - Indiana.
Large ball of light streaks across sky, stops, hovers, then explodes with a large boom. *Possible test. Standard Soviet incident to generate deception.*

7 **Sep** 78 - Netherlands.
Large acoustical event (boom) occurs over the Netherlands. *Soviet scalar EM weapons testing and calibration.*

Sep 78 - Tabas, Iran.
Anomalous Richter 7.4 earthquake strikes Tabas, Iran, killing 25,000. No aftershocks. Anomalous depth. Soviets needed to relievo pressure in rocks at nearby Askhabad on the Soviet/Iran border, else lose the city to a large earthquake. *36 hrs. earlier, Soviets exploded monstrous 10 Megaton underground nuclear explosion, to disguise what caused the earthquake. Almost definitely a Soviet- induced earthquake.*

Oct. 78 - USSR.
U.S. DOD weather satellite purportedly photographs a semi-rectangular object climbing out of atmosphere at 4,000 to 4,500 **mph.**

Object was over the Soviet Union. *Probable Soviet advanced flying vehicle, using antigravity propulsion system, if photograph is real.*

31 **Oct** 78 - Fairbanks, Alaska & Lake Ontario.
Red and yellow flash seen descending through sky S. of Fairbanks, Alaska by multiple observers. One minute later, U. of Alaska seismograph registered a tremor 1.5-2.0 on Richter scale. Ten days prior, a similar phenomenon was observed over Lake Ontario and a photo taken of the light. *Probable Soviet scalar EM wpns testing calibration.*

Nov 78 - Kuwait.
Large flying vehicle lands at Umm. Alaish oil pumping sta., Kuwait. Disrupts elect, comm, and pumping activities. Leaves after 7 min. *Probable antigravity propulsion system. Soviet vehicle. One of a series of "UFO" sightings in oil fields of Kuwait. Positive scalar electromagnetics and time-reversed wave signatures.*

Nov - Dec 1978 - Kuwait.
Series of "UFO" incidents over Kuwaiti oil fields occur. Kuwait Gov't. investigates, makes official report. *U.S. Pentagon noted incidents. Probable Soviet-induced incidents.*

1979 - British Columbia & Caspian Sea.
Ariel 6 satellite experiences anomalous turnoff of two power supplies. *Over British Columbia and Caspian Sea, when sun is shining. Probable Soviet stimulus.*

Jun 79 - Vienna.
Pres. Carter signs SALT II agreement in Vienna. After that, his acts and speech are erratic and he seems to be in a depressed state. *Suggestive of exposure to EM behavior modification. Soviets had possessed the LIDA technology for 29 years.*

Sep 1979 - USSR.
Strange giant semicircular light phenomena observed in USSR, toward Saryshagan. Observed several times. Source: London **Times.** *Observed from Afghanistan by British cameraman Nick Downie. Definitely Soviet tests of a Tesla globe and/or shield.*

22 Sep 79 - South Atlantic.
U.S. Vela satellites detect anomalous flash over South Atlantic, as if from a nuclear event. *U.S. Govt Agencies still disagree as to whether this was a nuclear event or not. May have been a Soviet test, or a test of a scalar EM weapon by another nonhostile country.*

25 Dec 79 - Afghanistan.
Soviets invade Afghanistan. *This is the Soviet "Spanish Civil War," where wpns, training, and the state of their ready reserves are tested. It is not primarily the regulars who are fighting there. Some scalar weapons — e.g. Smirch death ray — are tested also.*

1980 - Afghanistan.
Soviets kill 225 women and children in a village in the province of Maiden, near Kabul, Afghanistan. A month and a half later, bodies still undecayed. Bodies were also shot, to try to deceive Mujaheddin as to cause of death. *Bodies perfectly preserved. No smell, no bugs, no scavengers had eaten them. Probable Soviet test of a "death ray." Definitely scalar electromagnetic weapon signature.*

Feb 1980 - Duncanville, Texas.
About 100 startlings suddenly start falling from flight, instantly dead. *One of several such anomalous bird falls. Connected with very high frequency components added to Woodpecker signals for sharp localization. Birds entering such an area are killed.*

May 1980 - Seminole, Florida.
Start of mysterious plague of assorted illnesses striking healthy persons in a 20-block section of Seminole, Florida. Animals affected also. Lack friendly bacteria in their intestines. *Signatures of EM induction of disease and kill of intestinal bacteria. The lack of friendly bacteria is important. May indicate one direction of biological warfare planned by Soviets.*

June 1980 - USSR.
Huge, luminous hemisphere of light seen within Soviet Union, from Kuwait. Probably over Caucasus region of USSR. Source: personal communication from observer. *Tesla shield. A definite test, un equivocal.*

July 1980 - U.S. and Canada.
Richter 5.1 quake centered in northern Kentucky rumbles through 14 states and into Canada. *No known fault. Was predicted two wks in advanced by an associate.*

7 **Nov 80** - Applegate Valley, Oregon.
Mysterious explosion, earth rumbling, brilliant light turns night into day in Southern Oregon. *Associated with subsequent earthquake next day?*

8 **Nov 80** - California.
Richter 7.5 earthquake off coast of Southern California. Shakes CA and Southern Oregon. *Note anomalous phenomena in S. Oregon the preceding day.*

Dec 1980 - South Atlantic.
U.S. Vela satellites detect another anomalous nuclear-like flash over the South Atlantic. *In infrared only. Obviously not a normal nuclear wpn.*

28 **Dec 80** - South Atlantic.
Great bluish-white flash over the South Atlantic. Clear sky except 2 or 3 small cumulus clouds, one about 600 ft. above ship. *Possibly connected with the second Vela flash detection, or generated by the same system. One of many such incidents.*

1 **Jan 81** - Somerset, Pennsylvania.
Pilots of two a/c report the descent of a bright, shiny object with fiery tail over Pennsylvania. Large boom rumbles across most of Western PA. *Possible Soviet scalar EM testing. Part of deception plan.*

Jan 81 - Minnesoto, Wisconsin, South Dakota.
Great skyflashes and rumbling noises occur around Morris, Minnesoto and Wisconsin and South Dakota. *Possible Soviet scalar EM testing. Part of deception plan.*

Feb 81 - San Jose California.
Two pilots menaced by 10 ft. dia. light which flew directly at aircraft over San Jose, CA airport. Plane climbed, light followed, then sped away. *Totally consistent with exercise of the Soviet Launch Phase*

Anti-bomber system's real-time holography aspects of the Woodpecker OTH radars.

6 **May** 81 - Fredericksburg, Maryland.
EC-135 a/c crashed near Fredericksburg, MD. Severe nose-down from 29,000 ft. Apparent loss of all power. At 5,000 ft., either fire or minor explosion occurred. At 1,500 large explosion destroyed the aircraft. *Possible Soviet test of OTH radar anti-bomber weapon system?*

Mid-81 - Afghanistan.
Soviet helicopters use mysterious, highly lethal agent in Afghanistan. Causes death so quickly victims do not have time to make the slightest move. Rpt by Jane's Defense weekly. Referred to as Smirch gas or Smert gas. *Probable test ofa Soviet scalar EM "death ray."*

20-23 Jan., 1982 - Gander, Newfoundland.
"Object with red, green and yellow lights sighted in the Gander, Newfoundland area more than 36 times. Alternately hovered and darted in a complex "flight" pattern. *Possible real-time holographic phenomena produced by Soviet scalar EM weapons as part of the worldwide deception plan. Gander is a major "registration point."*

3 **Feb** 82 - Newfoundland.
Red and green flashing lights travelling across the horizon at 2,000 mph reported by 3 Armed Forces officers at Gander, Newfoundland. *Possible real-time holographic phenomena produced by Soviet scalar EM weapons as part ofthe worldwide deception plan. Possible major "registration point."*

Feb 82 - Ankasaray, Turkey.
Cluster of high, multiple lights pass over Turkish town of Ankasaray, 200 km. from Ankara. Dogs were disturbed also. *Possible real time holographic phenomena produced by Soviet scalar EM weapons as part of the worldwide deception of plan. Gander is a major "registration point."*

Sep 82 - Ankara, Turkey.
Cluster of 10-14 high, multiple lights pass over Ankara, Turkey in clear night sky. Dogs were also disturbed. *Possible real-time holographic phenomena produced by Soviet scalar EM weapons as part of the worldwide deception plan.*

Early 82 - St. Petersburg, Florida.
Extended series of booms occur off coast of Florida, near St. Petersburg. *Continued lengthy period of time. Years of Soviet scalar EM interference grid adjustments have been registered in Florida and off St. Petersburg. Booms are so numerous that meetings of civil defense, federal, and state officials have been held.*

June 82 - Washington/Canada border and off-shore.
Extended series of booms occur over Pacific Ocean, centered about Port Angeles and Sequim near Washington state/Canadian border. *Persisted for several weeks. As many as 10 blasts per day. Fireballs and lights also seen. Soviet extended scalar EM tests.*

18 June 82 - Northern Pacific Ocean.
Two JAL flight crews observe large, yellowish-white globe of light over the North Pacific Ocean 700 mi. east of Hokkaido, Japan. Covered one fourth of the sky. Light gradually dimmed and disappeared. *Tesla globe. Definitely a Soviet weapons test.*

June 1982 - Red China.
Several Red Chinese giant light hemisphere and globe incidents. In one instance, a/c electrical power systems began to fail in several Red Chinese fighter a/c when encountering a huge hemisphere of glowing light. *Tesla globes. Tesla shields. Balls of light. Definitely Soviet weapons tests.*

24 **Sep** 82 - Maryland, New Jersey.
Anomalous airquakes off the coast of Cape May and Northern Maryland, and Atlantic City, New Jersey. *Adjustment of Woodpecker grid. Soviet tests. 3 weeks later also.*

14 **Oct** 82 - Maryland, New Jersey.
Anomalous airquakes off the coast of Cape May and Northern Maryland, and Atlantic City, New Jersey. *Adjustment of Woodpecker grid. Soviet tests. 3 weeks earlier also.*

16 **Nov** 82 - Delaware, New Jersey.
Five mysterious booms rock Sussex county, Delaware and southern New Jersey. *Adjustment of Woodpecker grid. Soviet tests.*

1983 - Loma Linda, California.
Photos and information about the Soviet LIDA machine released. 40 MHz carrier, complex waveforms. Induces catatonic state in humans and animals. *Shows consciousness can be affected by EM means. Was used on U.S. prisoners of war in Korea. Suggests similar capabilities are possible for Woodpecker. U.S. scientific community had said impossible. President Carter affected in Jun. 79?*

Early 1983 - Bennett Island.
Large exhaust plume detected coming from Bennett Island and vicinity, one on-shore and two offshore. Plume was 6 mi. wide, 155 mi. long and reached 23,000 ft in height. *U.S. weather satellites detected it. Nearly horiz. jet, 155 mi. long. Definitely not natural. At angle ofperhaps 1.5 degree above horizontal.*

27 **Sep** 83 - England.
British RAF/Panavia Tornado aircraft loses all electrical power and crashes. *Possible Test? Tornado system is triplexed.*

22 **Oct 83** - Off Okinawa.
Two a/c operating off U.S.S. Midway apparently lost w/o trace, about 127 mi. E. of Okinawa, under anomalous conditions. *Causes unknown. Additional data needed.*

8 **Nov 83** - Western Europe.
Anomalous earthquake occurs in Western Europe. All foreshocks and aftershocks were missing. Called the "Leige" earthquake. Definite symptoms of manmade (Soviet) origin. *Large boom occurred over Netherlands the next day over 35,000 sq. km. area. Soviet scalar EM weapon testing.*

9 **Nov 83** - Netherlands.
Large acoustic event occurs over Netherlands. Detected over 35,000 sq. km. First mistaken as a Dutch quake. *One day after the anomalous Leige quake. Soviet wpn testing.*

Dec 83 - Alma Ata.
Stored munitions exploded at Dolon (Soviet) airfield. 150 km south of Alma Ata. One of six that mysteriously exploded in 7 months. *May be evidence that another nation also has Scalar EM wpns, and*

is checking the Soviets. Ifso, probably was warning that Soviets also would suffer if they moved.

1984 - Novaya Zemlya.
First mysterious plume detected from Novaya Zemlya by U.S. weather satellites. Sometimes plumes are 200 mi. long. *Novaya Zemlya contains Soviet atomic facilities including weapon facilities.*

1984
Senior Soviet physicist G. Yu. Bogoslovsky publishes paper, "Generalization of Einstein's Relativity theory for the anisotropic space-time." Such Russian research published regularly. Would be impossible in U.S., where similar attempts meet discredit, humiliation, and rejection by leading U.S. physics journals and departments. *The West has decreed that general relativity must not be local, and above all must not be lab-bench experimental. Open Soviet publication indicative of experimental development of local space-time curvature technology.*

1982 - 84 - Florida (& elsewhere).
Airquakes, booms (hundreds). *16 counties, near coast.*

28 Jan 84 - Pennsylvania, New Jersey.
Bright orange light streaks across sky from Pa. to N.J.'s Cape May. Similar "fireball" explodes in Clementon. Two other sky explosions observed. *Possible Soviet activity. TR wave holography and pulses. Part of Soviet deception plan.*

Feb 84 - in space.
All three payloads of the Challenger flight fail. *2 commication satillites, 1 balloon.*

10 **Feb** 84 - Netherlands.
Large atmospheric boom occurs over Netherlands. *7 others thru 1983.*

29 Mar 84 - Bennett Island.
Bennett Island exhaust plume detected by U.S. weather satellite. *Over 100 such anomalous exhaust plumes detected there.*

9 **Apr.** 84 - S. of Kurils toward USSR.
Giant rapid cloud tower rose, then faded. Small semicir. halo of light replaced it. Halo expanded to full circle of light. Expanded to giant globe then faded. *Cloud 60,000 ft in 2 min. Light globe expanded to about 380 mi. dia. and center over 200 mi. high. Source:* ***Science*** *Journal. Definitely a Soviet test of multiple weapon modes.*

26 **Apr.** 84 - Bennett Island.
Bennett Island exhaust plume detected by U.S. weather satellite. *Over 100 such anomalous exhaust plumes detected there.*

May 84 - S.E. U.S.
Series of devastating storms breaks out over S.E. U.S. in early May. Tornadoes, large hail, downbursts. Some loss of life. Persisted through night into early morning hours. Several instances of rapid thunderstorm top warming were observed in satellite photography. Occurred near downbursts. Very heavy plume activity at Bennett Is. noted in Apr. and May, 1984. *Warming of tops of thunderstorms can produce downbursts. In literature. Note that Soviet scalar EM wpns can perform such top-warming of thunderstorm anvils, and induce downbursts on command.*

13 **May** 84 - Severomorsk, USSR.
Soviet missile storage facility at Severomorsk Naval Base on the Kola Peninsula 1450 km north of Moscow blows up in a series of vast explosions. Over 200 wounded. Soviet North Fleet lost about 1/3 of its missiles. Speculation of an "electromagnetic cause, possibly by reflection of radio waves from ionosphere." Detected by satellites; first thought to be nuclear explosion. *One of six such explosions of Soviet ammunition storage sites in a period of seven months. Definitely not accidental. Probably use of scalar EM weapons by another nation, not hostile to U.S., to warn the Soviets that they also will suffer if they move.*

15 **May** 84 - Bobrwjk Airfield.
Major blast occurs at Bobrwjk Airfield, 138 km SE of Minsk and the base of a Badger-equipped air-to-surface missile regiment of the Smolensk Air Army. 10 of the 11 ammo storage bldgs blow up and admin, and access control bldgs. are destroyed. Note that fires and explosions were still raging at Severomorsk, from 13 May event. *Same note as above.*

25 **Jun** 84 - Schuerin, East Germany.
Explosion occurs at an ammunition depot at Schuerin, south of Wismar in East Germany. Windows blown out within 10 mi. radius of depot. *Same note as above.*

27 **Jul** 84 - Kuril Islands.
Slowly expanding ball of light seen by Boeing 747 crew near the Kuril Islands. Expanded for 10 min. Shell was sharply defined. Light semi-transparent and nearly perfect half-circle. *Tesla globe. Bennett Is. exhaust plumes detected by U.S. weather satellite. Definitely a Soviet test. Related to 9 Apr. 84 incidents S. of Kurils.*

29 **Aug 84** - Edwards AFB, California.
B-l bomber prototype crashes in flight test about 10 mi. NE of Edwards AF Base.

1983-84 - Ireland.
For two years, freak lightning terrorizes villages of West Donegal, Ireland. Roofs stripped off, windows smashed, pipelines wrecked, telephones and TVs blown up. *No known cause for the anomalous lightning. Shows natural potential in area is disrupted.*

25 **Dec 84** - Ireland.
Freak lightning continues in West Donegal, Ireland. Causes 3-day power blackout. *Extensive plume activity at Bennett Island.*

1985 - Worldwide.
After years of a superb safety record, airline industry suffers a series of disastrous crashes. 20 major air crashes occur worldwide, with more than 2,150 lives lost.

Early 85 - Worldwide.
KGB begins a worldwide campaign to blame the U.S. for starting the AIDS epidemic. *Effort continues. Disinformation; is believed in 3rd world.*

Jan 85 - San Diego, California.
U.S. Navy chaff caught by wind, blown in over San Diego. Power knocked out to 60,000 homes. *Symptoms of presence of Soviet scalar EM grid in San Diego urea.*

7 **Jan 85 -** Cantrell, Alaska.
Cantrell, Alaska. Blinding white light passes in sky overhead, lights up sky and mountain range for 30-40 seconds for 16 mi. Est. 3-mi. altitude. *Probable Soviet test. Note that plume activity at Bennett Is. is heavy in Jan. 85.*

Feb 85 - Zimbabwe.
Zimbabwe suffering worst electrical storm season in history. Death toll up to 116. 21 persons killed by a single bolt in one case. *Anomalous lightning. Shows natural potential in area disrupted.*

17 & **18 Feb 1985 -** Anchorage, Eklutna Flats, Northway, Alaska.
Several anomalous lights and sky booms incidents in Alaska. Skyflashes, balls of light, booms, etc. Seen both from ground and by pilots of aircraft. *Probable Soviet realtime holography tests as part of worldwide deception plan.*

19 Feb 85 - Out from San Francisco, California.
China Airlines Flight 006 Boeing 747 enroute to Los Angeles. Engines fail. A/c falls 32,000 ft. in 2 min. Restarts engines, lands at San Francisco. *Note previous incidents over Alaska on 17 & 18 Feb. Probable Soviet test of launch phase anti-bomber system capability of woodpecker transmitters.*

10 **Mar 85 -** Tblisi & Tallin, USSR.
Soviet Academy of Sciences announces an Aeroflot a/c had been followed by a UFO for 800 mi. *Soviets occasionally produce light phenomena in their own territory and near Soviet a/c as part of their deception plan. They do not wish to appear different.*

1 **May 85 -** Near Nashville (monitoring station).
In and around May Day 1985, the Soviet Union held a giant exercise of all its strategic scalar EM weapons. 27 "power taps" into the earth — each two frequencies 12 KHz apart — were activated. The earth was thus in forced gravitational resonance beneath our feet on 54 frequencies. Each tap estimated to power up to six of the giant weapons; thus over 100 major strategic scalar EM weapon transmitters were activated. Frequency bands were filled with scalar EM transmissions, command and control, including to subs on station underwater. *Frank Golden monitored this startling giant exercise, which was probably held for Gorbachev, and this author observed it by courtesy of Golden's special detection equipment. Thus the Soviets*

met the 1985 timetable established by Brezhnev in 1972 at a secret meeting in Prague of European Communist leaders. Western nations do not appear to have the equipment to monitor scalar EM emissions, and so Western intelligence agencies are believed to have totally missed this giant exercise and its significance. At conclusion, two power taps (four frequencies) were left in the earth to power "ready firing batteries," so to speak.

11 **Jun** 85 - Lanzhou, China.
China B747 crew encounters a huge, expanding yellowish glow of light, with very intense spot at center. *Probable Soviet test.*

11 **Jun** 85 - Mongolia.
Chinese jumbo jet, Peking to Paris flight, encounters giant "UFO" light form over Mongolia. 10 km. wide. Lit sky for 50 km. Extremely bright spot at center. Paced aircraft for 2 minutes. *Probable Soviet test. Soviet operator tracked a/c exhaust with long range antibomber system. Added giant light form for deception, and to create UFO reports, which are largely ignored.*

23 **Jun** 85 - Iowa, Nebraska, Kansas, Illinois, Kentucky.
Thunderstorms over Iowa, Nebraska and Kansas organized into single circular pattern rainstorm 200,000 sq. km. in area. Declined until dawn, then rejuvenated to unleash severe weather over Illinois and Kentucky. *Such organization never before had been seen. Now called "mesoscale convective complex." Shows unique new weather pattern involving giant organization.*

2 **Aug** 85 - Dallas/Ft. Worth, Texas.
Delta Airlines flight 191, a Lockheed L-1011, encounters microburst while approaching runway at Dallas/Ft. Worth International Airport. Crashes and kills 134 of the 163 persons on board. *While such microbursts occur in nature, this was during a period of intense Soviet scalar EM activity. Note that warming the tops of thunderstorm anvil causes a downburst. May have been just accident. May have been test.*

23 **Aug** 85 - Italy, Adriatic Coast.
More than 28 "UFO" sightings reported along Italy's Adriatic coast soon after **midnight.** Light phenomena. *Probable Soviet test. Part of deception plan. Scalar KM holography (real-time).*

28 **AUG 85 -** Vandenberg AFB, California.
Titan B fails after launch at Vandenberg AFB, California. Believed to have carried sensitive U.S. satellite. *Suspicious.*

Sep 85 - Kourou.
Ariane rocket explodes at Kourou. 3rd stage engine failed to ignite properly, forcing loss of power. Destruct then ordered. *French officials suspect Soviet sabotage. Probable Soviet test.*

Sep/Oct 85 - Cape Canaveral, Florida.
Anomalies observed on two U.S. shuttle launches by NASA. *No action taken; NASA does not recognize scalar EM weapons capabilities nor testing over its Cape Canaveral launch site. Nor does it connect the mysterious "booms" that have rocked the Florida coast for several years with the Soviets or with the anomalies noted on its two shuttle launches, such as giant booms.*

2 **Sep 85 -** Mexico City, Mexico.
Worst hailstorm in 50 yrs. strikes Mexico City, leaving some streets more than a foot deep in ice. 1-hr pelting. 25 bldgs collapsed. 4,000 roofs heavily damaged. 1 killed, 185 injured. Power blacked out. *May have been due to extensive Soviet long range weather engineering. 17 days prior to suspicious giant Mexican earthquake.*

5 **Sep 85 -** Australia.
Premier Wran, Australia, along with Minister of Agriculture Hallam, observe a brigh spherical "UFO" from their government jet. It was photographed. Radar contact showed object at 100,000 ft. *Probably Soviet advanced performance antigravity flying vehicle.*

18 Sep 85 7:15 pan. - Catalina Is. Channel Is.
Cookman observes giant cloud radial, awesome to behold, between Catalina and Channel Islands, off California coast.
18 "funnel" strobes, dynamically boiling. Activity indicates great energy poured into earth and atmosphere. Mexican quake the next morning. Indicator of Soviet scalar EM connected to quake.

19 **Sep 85 7:18** a.m. **-** Mexico City, Mexico.
Large earthquake strikes Mexico City, causing great damage. Largest of several. *Preceded wks earlier by giant anomalous hailstorm. Evening before, giant radials showed extensive Soviet power*

transmission activity. Positive indications of Soviet test.

8-9 **Nov** 85 - Ventura County, California.
Highly anomalous power blackout in Carlisle Canyon, Ventura County, Calif. Accompanied by severe electrical anomalies totally beyond normal electromagnetics. *Investigated and reported by engineer Ron Cole, who lives in the area. TV, light, phone, electrical anomalies plus intermittent 29 Hz signals.*

Nov- Dec 85 - California.
Mysterious hum continues to baffle people at Pacific Heights, Calif, and Marina district on San Francisco Bay. Previously had had same hum. Note such hum in England persisting over long period. Note 8-9 Nov 29 Hz. anomalous elec. signal at Carlisle Canyon, Ventura County, CA. *Possible indications of Soviet scalar EM activity. Area appears to be one of Soviet register / test points.*

26 **Nov** 85 - Cape Canaveral, Florida.
Launch of space shuttle Atlantis from Cape Canaveral, Florida is accompanied with high anomalies. (1) Seconds before launch, Bob Gladwin inadvertently photographs nearby strike of what appears to be an electromagnetic missile, probably of Soviet origin. (2) Anomalous light is hanging in the sky, near the launch site: a marker beacon, placed there by the Soviet scalar EM weapon system being tested. Seen by hundreds of persons, and photographed by George Suchary. Printed in Fort Pierce **Tribune.** (3) 12 minutes after shuttle is launched, when it is well down range, a tremendous rumbling atmospheric boom occurs, heard 600 mi. up and down the East Coast. Marker beacon then slewed away by distant Soviet operator at high speed. *With offset EM missile strike, Soviets tested the weapon to be used on the Arrow DC-8 two weeks later at Gander AFB, Nfld. Marker beacon used in precise registration. Note it was evidenced again above the explosion of the Titan on 18 Apr. 86 at Vandenberg AFB. Boom is from energy induced over a large area; which can catch multiple missiles being launched, a large Naval task force at sea, or a whole flight of airborne bombers, fighters, or helicopters. Also can catch an entire organization on the ground, or an installation - - wiping out all electronics, exploding explosives and ordnance, stunning or killing personnel, etc.*

Again hundreds of NASA engineers and scientists did not recognize what was ***happening*** *at the* ***site*** *and over their own heads.*

12 Dec 1985 - Gander, Newfoundland.
Arrow DC-8 crashes immediately after takeoff at Gander AFB, Newfoundland. Lost power just after takeoff; engines "rumbling" abnormally rather than roaring, according to eyewitness whom the a/c flew directly over, 100 ft up. Aircraft seen by three eyewitnesses to have an orange glow on it. Symptoms of negative energy introduced into engines, cancelling some of heat energy of the engines and seriously interfering with combustion process, causing loss of power. Glow may have been steady, small electromagnetic ball of energy against the aircraft, causing it to electrically charge and exhibit corona. Anomalous hole in right fuselage, just ahead of engines, as if from internal explosion; however, no explosive residue (it was tested). One other anomalous hole found. Wild swings of some instruments recording just at/after liftoff — ignored by investigation. *Note previous 36 light incidents in and around Gander, Jan. 1982. Gander Airport is major stopover for Soviets, and a major "registration point" for scalar EM weapons. Note anomalies in shuttle launch 2 wks earlier, and apparent actual test of the small EM missile then, offset so as not to destroy that shuttle.*

Aircraft crew one of Arrow's best. Aircraft should have flown okay, if not interfered with. Serious indications that the DC-8, with over 250 U.S. soldiers and air crew, was deliberately destroyed by test of a Soviet launch-phase antibomber system.

Note metal-softening Soviet signal detected on 1 Jan 86 by Frank Golden may actually have been on the Woodpecker transmissions on Dec 12,1985. This metal-softening signal probably played a part in the destruction of the Challenger on 28 Jan 86.

1 **Jan** 86 - Near Nashville, Tennessee.
Frank Golden discovers a metal-softening signal on the Soviet Woodpecker interference grid. Performs experiments to verify the metal-softening ability of the signal. *Signal is of such a nature that only Golden's instrumentation could have detected it. He then scalarly cancelled the signal to totally prove it.*

24 **Jan** 86 - Ventura, California.
Engineer Ron Cole observes and sketches a giant grid cloud pattern over Ventura, Calif, at 11:45 p.m. PST. Giant cloud fingers, separated from each other, formed the crossed-grid pattern. *Shows Woodpecker scalar EM power very high and active.*

26 **Jan** 86 - Northern California.
Strong quake hits Northern Calif. 5.5 on Richter scale. Hollister area, about 120 mi. south of San Francisco. Several power outages in southern Bay area, where quake knocked out transformers, downed 21,000 V. power line, and set off burglar alarms. *May have been induced by Soviet scalar EM transmitters.*

26 **Jan** 86 - Los Angeles, California.
In greater Los Angeles area, Margaret Wilson observes 3 classic giant radial clouds form at 10 min. intervals: 1:15, 1:25, and 1:35 p.m. Appeared one after the other, till all three were there. Last one lasted about two hours, gradually distorting. *Shows a great amount of structuring to capture and turn the Jetstream, to direct cold weather to the Florida Panhandle and catch the Challenger before its impending launch, exposing it to drastic cold conditions. Definitely Soviet engineered.*

27 **Jan** 86 - Santa Monica, California.
On evening prior to the Challenger disaster, Al Matthews photographs a strange bar-grid cloud formation, absolutely rectangular bars, from the Los Angeles freeway. Pattern approx. over Santa Monica Bay, due north of Catalina Island. Edges and ends very straight, as if drawn in sky by a giant draftsman. *Shows severe high frequency content, and fine-point localization of Soviet scalar EM transmitters, in preparation for forthcoming launch of the ill-fated Challenger.*

27 **Jan** 86 - Florida Panhandle.
Avalanche of freezing air temperatures to record lows across the Southeast. Low temperature records set all across the South. Jetstream bent far south, across Florida Panhandle. *Definitely caused by Soviet weather engineering. Preceeded by giant radial cloud signatures in "hinge" areas of Los Angeles and Huntsville, Alabama, showing steering of Jetstream.*

28 **Jan** 86 - Cape Canaveral, Florida.
Low temperature records shattered all across the South. Into 30's in Florida and into teens elsewhere. *Definitely result of Soviet weather engineering, and definitely directed against the Challenger rocket on the launch pad.*

28 Jan 86 - Off Florida Coast.
About 4 hr. before the launch of the Challenger, all Soviet ships off-coast, which normally monitor shuttle launches, suddenly depart at high speed. *Soviets do not want any suggestion that the ships' radars and transmitters may have interfered with the Challenger launch. Total anomaly. Unprecented event.*

28 Jan 86 - Cape Canaveral, Florida.
National radio and TV newsmen comment on the highly unusual fact that no birds are flying in the launch area. *Reason: For severe localization, high frequency content of the scalar EM signals had been drastically increased. A bird's brain will detect and resonate to this energy, if the wavelength approaches the diameter of its brain. This resonance will cause great pain and/ or death of the bird. Thus the birds flee the area when this begins to occur — or they die. When birds accidentally fly into such a high frequency scalar EM signal, their nervous systems are instantly jammed. They die and fall from the sky.*

28 Jan 86 - Cape Canaveral, Florida.
Space shuttle Challenger, launched from Cape Canaveral after exposure to cold weather conditions, disastrously explodes. Booster seal problem. Encounters air turbulence during launch, at most critical altitude. *Soviet metal-softening signal has been added to grid. Extreme localization in area because of high frequency content added; shown by birds having vacated the area. Anomalous departure of Soviet ships offshore, 4 hrs. earlier. Three previous tests of the weaponry in that exact launch area, against the three previous shuttle launches. Kill of the Arrow DC-8 about 6-112 weeks earlier. Actual celebration by the KGB of perfect success of their active measures against the Challenger! U.S. space program crippled and set back severely.*

29 Jan 86 - Moscow.
At KGB headquarters in Moscow, the KGB holds a jubilant party to celebrate the perfect success of their active measures against the Challenger. *Note that Soviet scalar EM weapons — development, deployment, and use — are under the control of the KGB rather than the regular Soviet armed forces.*

31 **Jan 86** - Mid-U.S.
Richter 5.0, Earthquake rattles 9 states and Great Lakes. *Possible test of weapon that killed the Challenger, in the earth transmission / earthquake induction mode.*

2-4 **Feb** 86 - Birmingham, Alabama.
In Birmingham, Alabama dead birds plummet from the air to the ground for three days. Blackbirds, cardinals, sparrows, bluebirds, etc. No trace of poison in carcasses. *Only a few days after kill of the Challenger. To divert the Jetstream further south and hit Florida, the Soviets had moved the normal "hinge pin"—where scalar EM energy, either positive or negative, is focussed —from Huntsville to Birmingham. Hinge pin area was still activated. The high frequencies used for extreme localization against the Challenger were lethal to birds. Therefore birds that inadvertently flew into this zone were instantly killed and dropped from the sky.*

10 **Feb.** 86 - Europe.
Super Frelon (French) helicopter crashes, with 11 killed. Aeronautique Navale aircraft. Part of a spate of crashes of French military aircraft in four months (through June 86). *Defense Minister Andre Giraud launched a special investigation.*

23 Feb 1986 - Irish Sea.
Over the Irish Sea in an airplane, England's Prince Charles sights a brilliantly glowing, red object hovering near them. His pilot reports the object to air traffic control. Four other aircraft traveling through that area radioed similar reports. *Probably either a Soviet craft or a Soviet holographic scalar EM demonstration as part of the overall deception plan, and/or a stimulus to a member of the Royal Family to observe what action is taken by British Ministry of Defense. By actions, could ascertain whether or not Britain knew what was going on.*

30 **Mar** 86 - Ventura, California.
Engineer Ron Cole in Ventura CA observes significant cloud patterns, positively correlated with his measurements of the Soviet woodpecker transmissions. *Cloud pattern changes were directly correlated to signal changes on the Woodpecker. Shows cloud signatures can be indications of specific Woodpecker activity.*

31 **Mar** 86 - San Francisco, California.
Strong earthquake, Richter 5.3, rocks San Francisco Bay area. Third in region in 3 days. Three aftershocks within minutes. *Note correlation of Woodpecker activity previous day.*

31 **Mar** 86 - San Francisco, California.
Crew error on Delta flight accidentally shuts down engines after normal takeoff from San Francisco. Engines restarted. *Definitely a crew error. Possible human effects from earthquake energy in area. Same type incident on Jun. 30, 1987.*

3 **Apr.** 86 - Florida.
Loud boom shakes houses and rattles windows over Brevard County, Florida. One of a whole series over the years. Not far from Kennedy Space Center. *Soviet adjustment of scalar EM grid.*

Apr 86 - France.
French Armee de L'Air has lost two Jaguars and two Alphajets in April 86. Part of a spate of losses in 4 months, leading Defense Minister Giraud to launch a special investigation.

7 **Apr** 86 - Sacramento, California.
Charlotte King of Sacramento calls UPI and predicts a major earthquake will strike in Canada, Alaska, Aleutians, or Japan. She had predicted eruption of Mt. St. Helens in 1980. On Apr. 27th, she predicted quake that shook Mexico three days later.
Note that human brain is a scalar EM interferometer, capable of transmitting and receiving scalar EM waves. Unfortunately, not conscious — output of system is in deep unconscious. Some persons are more sensitive than others, and can detect the scalar activity from the increasing stress in rocks that is leading toward an earthquake. Charlotte King is such a person, having over an 80% accuracy of prediction of quakes. She will be able to detect the increasing stress in rocks, whether naturally or artificially induced.

9 **Apr.** 86 - Aleutian Islands.
Three quakes rock the Aleutian Islands. Predicted by Charlotte King on Apr. 7. *Documented prediction.*

18 **Apr.** 86 - Huntsville, Alabama; Thousand Oaks, California.
T.E. Bearden sees and photographs remnants of a giant radial cloud

in Huntsville, Alabama. Extensive Soviet woodpecker grid activity noted over Thousand Oaks, California by Engineer Ron Cole. *Both show increased scalar EM activity in the interference grid.*

18 **Apr** 86 - Vandenberg AFB, California.
USAF Titan 34-D rocket explodes nine seconds after launch from Vandenberg AFB, California. Video captured the Soviet marker beacon **(Aviation Week & Space Technology).** Engineer Ron Cole and team study video of the explosion frame by frame and verify the anomalous beacon light is present, is separate from the explosion, and moves away separately after the explosion. *Sensitive satellite believed on board the rocket. U.S. Space surveillance program crippled. Probable Soviet kill. TR wave adjunct with Woodpecker.*

25 **Apr** 86 - New Mexico.
NASA research rocket misfires over the New Mexico desert. No public announcement made at the time. Was first failure in 25 consecutive missions of the Nike-Orion rocket. Said NASA spokesman Kukowski: "It looks like we're snakebit."
Possible Soviet test kill.

25 **Apr** 86 - California; Chernobyl, USSR.
Bill Bise detects the sudden loss of the East-West Soviet woodpecker transmitter he is monitoring in the field, leaving the North-South transmitter. The failed transmitter was about 20-30 km. from Chernobyl. The *Soviet transmitter sites safety circuits go-in full, trying desperately to hold back the giant scalar EM standing-wave potential that has been built up, and slowly and safely drain it away into the earth. In alarm, Soviets shut down nearby reactors at Chernobyl. Any "flash discharge" escape as a flash-over pulse of the scalar EM potential will produce a giant EGP pulse into the earth. When this EGP pulse reaches the four reactors, they will violently explode, full-up nuclear, like giant hydrogen bombs — perhaps 1200 megatons equivalent. Later the Soviets will claim that the operators shut the plant down to do some "unauthorized experiments."*

26 **Apr** 86 - Chernobyl, USSR.
With the giant scalar EM standing wave almost all drained away, the safety circuits finally ***fail,*** *dumping a small EGP into the earth. The nearest reactor at* ***Chernobyl*** *gets* ***it,*** *and the uranium fuel roils*

suddenly go half-decayed, instantly heating and blowing out the reactor, destroying it. Nuclear radiation fallout spreads over much of Europe.

3 **May** 86 - Florida.
Delta 178-rocket destroyed by range safety officer at T + 91 after two anomalous power surges occurred in the rocket, causing premature shutdown of 1st stage Rocketdyne RS-27 engine 71 sec. into the flight, and rocket to go aerodynamically unstable. Previous 43 launches had been successful. Since 1960, 177 launches 94% successful. One of most reliable rockets ever built. *Soviet weapons test. One of a series. TR wave adjunct with Woodpecker.*

7 **May** 86 - Aleutian Islands.
Great earthquake measuring 7.7 Richter strikes Aleutian Islands. Centered about 100 miles SE of the island of Adak. Tsunami alert issued for the North Pacific, but no major waves ever materialize. Thousands evacuate coastal areas. A *U.S. researcher known to this author used a powerful scalar interferometer to block the tsunami. The researcher must remained unnamed.*

9 **May** 86 - Redstone Arsenal, Alabama.
All power on Redstone Arsenal, Alabama suddenly and mysteriously fails all at once. *Similar related incident on May 12. No explanation.*

12 **May** 86 - Huntsville, Alabama.
In Huntsville, Alabama — adjacent to Redstone Arsenal — on local radio program Feedback, public radio discussed the mysterious failure of all power on Redstone Arsenal on Fri., May 9, 1986 and requested that, if anyone knew why all circuits would go off at once, to please get in touch. *Immediately the station's radio transmitter went off the air! (Almost as if "Big Brother's listening to you!")*

18 **May** 86 - Dijbouti.
French Breuguet Atlantic aircraft crashes in Dijbouti, killing 19. *Part of spate of crashes of French military aircraft in 4 months that leads French Defense Minister to open special investigation.*

19 **May** 86 - Brazil.
Brazilian fighter jets are sent to chase multi-colored "ping- pong ball sized" UFO's in skies over Brazil. 4 jets, two Mirages and 2 F-5's

chased them at speeds of 840 mph. Tracked on radars. Chase lasted 3 hrs. Planes could not catch them; ran low on fuel and returned. Discussed on TV by Brazil's Air Force Minister Brig. Gen. Otavio Moreira Lima. *Soviet scalar EM holographic production; distant Soviet operators having fun and games as part of the deception plan, generating UFO incidents and reports?*

22 **May** 86 - California.
High-speed orange "X" object with lights seen above Sonoma CA before dawn. At one time the object slowed and hovered.

29 **May** 86 - Huntsville, Alabama.
Huntsville and Madison County, Alabama electrical power dipped momentarily and mysteriously, causing a seconds-long outage in businesses and homes, playing havoc with traffic signals and computer centers throughout the area. Some traffic signals continued to malfunction for 30-45 min. after power was restored, then mysteriously resumed normal functioning. Some phones also went out inexplicably, and many battery-powered radios went off, and were awry for seconds or minutes before recovering. *Obviously this was not caused by any kind of ordinary electromagnetics. Shows anomalous and high activity (pulsed) in the Soviet scalar EM grid (Huntsville Ah. is a pivot point).*

30 **May 86** - French Guinea.
Ariane 2 rocket fails because of a faulty ignition system for the third stage engine. Rocket destroyed by ground command 4-1/2 min. after launch from Kourou, French Guinea. The third stage has failed 3 times since start of Ariane's operations in 1979.
Possible Soviet test.

Tad Szulc, Washington journalist, implies that French Defense Minister Andre Giraud may have discussed the possibility of Soviet sabotage of the rocket with Washington.

May? 1986 - Straits of Gibralter.
USS Atlanta nuclear sub runs aground in Straits of Gibralter and suffers "serious damage," according to reports in Madrid. A hole was made in one ballast tank and the sonar wrecked.
*Another **U.S. nuclear sub** ran aground within a month or so (earlier?) of this incident. C**ould something** have interfered with their navigation systems underwater? Continued Soviet Scalar EM tests?*

11-12 July - California.
Secret USAF crashes approx. 14 mi. NE of Bakersfield, CA along the Kern River. Site secured. Possible Stealth fighter. Crippled aircraft in trouble seen and photographed by Andy Hoyt, who turned photos over to USAF (properly so). Aircraft not on fire when came down. Did not explode in the air. *No information as to whether any Soviet scalar EM hanky-panky involved or not. Suspicious because follows a series of anomalous missile failures. Woodpeckers definitely have good anti-stealth capability.*

11 July 86 - Tennessee and Georgia.
Moderate quake with sonic boom shakes SE Tennessee and Northern Georgia. 3.7 on Richter scale. Centered just east of Chattanooga Tennessee on the TN/GA line.

12 July 86 - Michigan, Ohio, and Kentucky.
Mild quake in Michigan, Ohio, and Kentucky. 4.0 on Richter scale.

18 July 86 - Los Angeles, California.
Sharp aerial boom rattles Los Angeles Basin at about 0930 hrs. FAA and USAF unable to identify a potential source from an aircraft.

Aug 86 - Soviet Union.
Two huge new Soviet "radars" discovered on Western Soviet border by U.S. intelligence. One near Skrunda, on Lithuanian border. Other near Mukachevo, on the Czech border. *With phase conjugate adjuncts, represent significant strategic weapon system deployments.*

2 **Sep** 86 - Baltic Sea.
Finnish ferry Scandinavia in Baltic Sea observes a strong light rise from the sea near the Finnish Coast. After 3 to 4 seconds, a giant mushroom reached 500 feet into the sky and covered the moon. *Probable Soviet test.*

12 Sep 86 - Dallas, Texas.
Mysterious, powerful "explosion" rocks South Dallas. No source found or explained. No supersonic jets in area.

Selected ***Incidents***

15 **Sep** 86 - Australia.
Katherine, Australia. Object with flaming orange **tail** moved **from** sky toward ground at 45 degree angle. Katherine's electrical **power** suddenly cut off and phone lines in the town disrupted. ***Probable*** *Soviet scalar EM generated.*

Oct-Nov 86 - Los Angeles, California.
Giant radial clouds observed in Los Angeles area.

Oct 86 - Reykjavik, Iceland.
Gorbachev and Reagan meet at Reykjavik, Iceland. Gorbachev proposes "zero option." *Seeks to get nuclear missiles removed, since they constitute "dead man fuzing" because of nuclear warheads, and severely limit pulse use of Soviet scalar EM weapons. Adamantly insists on bottling up SDI — because if space laser is converted to scalar EM laser, its power (effective) goes up about* 10^{20} *times, and single shot could devastate Russia.*

10 **Oct** 86 - Vandenberg AFB, California.
U.S. forced to scrub another missile launch when a USAF Scout satellite malfunctions during final countdown at Vandenberg AFB, Calif. *The booster's gyro system failed a checkout sequence at about T-5 hr. Scalar EM can inertially interfere also.*

13 Oct 86 - ABC News.
On Peter Jennings' ABC News show, Gennadi Gerasimov, Soviet spokesman, indicated that SDI as presently designed was of no concern to Soviets. Real concern is that U.S. will make a "great technical breakthrough" and deploy new, modified devices, presenting an unacceptable threat to the Soviet Union. *Here he is thinking of changing the laser to a scalar EM laser.*

6 **Nov** 86 - Albuquerque and Phoenix.
Power failures at Albuquerque air traffic control tower and a radar station near Phoenix, Arizona occur. Radar controlling high altitude air traffic from Texas Panhandle to southern California was knocked out for 40 min. First such incident in Albuquerque center. The Arizona radar was out for 59 hrs. *Note similar incident at Montreal on Nov. 11, 1986.*

10 **Nov** 86 - Huntsville, Alabama.
Three giant radial cloud formations in Huntsville, Alabama photographed by T.E. Bearden. One is a twin giant radial. *Shows heavy activity in scalar EM interference grid.*

10 **Nov** 86 - SE U.S.
Red glowing light with tail flares across sky in Southeast about sunset. Seen over parts of TN, AL, GA, MS. Ground searches conducted in several areas for possible crash. *Probable Soviet scalar EM activity. Note heightened activity in grid same day.*

10 **Nov** 86 - Near Nashville, Alabama: Hemet, California.
Frank Golden measures scalar EG field of earth — highly agitated and dynamic, about half as much as when Soviets induce earthquakes. Anomalous winter storm in progress. Cold records broken in Mid-U.S. Heat records broken in Florida. Storm penetrates very deeply southward. *Soviet weather engineering.*

11 **Nov** 86-Canada.
Air traffic control radar in Montreal, Canada region broken down. Backup system employed. *Note similar occurrences to Albuquerque and Phoenix on Nov. 6, 1986.*

12 **Nov** 86 - Atlanta, Georgia.
Stationary arc (dome) of clouds sighted over Atlanta. Well defined lower edge. Upper edge irregular. Multiple witnesses. Clouds moved into top of dome and flowed around it. Very strange sight. *Probable Soviet "gentle test" of placing a suppressive covering over a distant U.S. city. Performed at night, at minimal level (no glow) so no great notice would be taken. Dome was stable against 20 mph winds and other moving clouds that collided with it.*

13 **Nov** 86 - Huntsville, Alabama.
Twin giant radial cloud observed moving over Huntsville, Alabama by T.E. Bearden. West to East. At nearly noon. At dusk, single giant radial cloud was also spotted over Huntsville.
Significant scalar EM activity in grid.

17 **Nov** 86 - off Alaska.
JAL Flight 1628 observes giant UFO while enroute from Reykjavik, Iceland on mid-leg of Europe to Tokyo. Crossed Beaufort Sea off

Alaska's north coast. Sighted flashing lights which moved alongside the aircraft. Pilot descended, so did light. Glimpsed craft in profile, twice as big as aircraft carrier. On board radar located it. USAF and civilian controllers reported brief radar echoes, but tapes did not show it. Approaching Fairbanks, Alaska. Pilot made 360 degree circle, object stayed with him. Finally disappeared as he was approaching Anchorage. *Probable Soviet test as part of deception plan. TR wave holography.*

27 **Nov** 86 - Fort Pierce, Florida.
Flash of light in sky and sonic boom occurred vicinity of Fort Pierce, Florida. *Probable Soviet EM test. Note that a classified payload was scheduled for launch in near future at Cape Canaveral nearby.*

4 **Dec** 86 - Huntsville, Alabama; Hemet, California.
Twin giant radials seen and photographed over Huntsville AL. Large radial cloud seen in skies over Hemet, Calif, and photo printed in **The Press Enterprise,** Hemet, CA 5 Dec 86.

4 **Dec** 86 - Cape Canaveral, Florida
Atlas/Centaur rocket successfully launched at Cape Canaveral, orbits a communications satellite for DOD. *The Soviets had to allow one or two of the missiles to succeed, else even the lethargic U.S. bureaucracy would have caught on. Grid activity indicates they probably just tracked it without the other goodies.*

5 **Dec** 86 - Vandenberg AFB, California.
MX ICBM launched from Vandenberg AFB. Two of RVs did not deploy although all impacted within the Kwajalein Missile Test Range. *Note: Soviets let one go from Vandenberg.*

21 **Dec** 86 - Huntsville, Alabama.
Twin giant radial photographed over Huntsville Alabama.

23 **Dec** 86 - Huntsville, Alabama.
Small flash in lower sky over Huntsville, Alabama observed by T.E. Bearden. Followed by faint boom. Momentary loss of power occurred in large area of Huntsville. Positive correlation between the aerial flash/boom and the loss of the power. *Soviet test.*

7 - 21 Jan 1987 - Florida.
Several giant radial clouds seen in vicinity of Fort Lauderdale, Florida. *Indicates substantial action in scalar EM grid.*

20 **Jan** 87 - U.S.
Highly unusual bend in Jetstream. Completely down West Coast, curving east under Texas, NE along SE USA. *Shows great action in scalar EM grid. Significant weather engineering.*

20 **Jan 87** - Vandenberg, AFB, California.
Minuteman missile destroyed in launch. *Note substantial action in scalar EM grid.*

23 **Jan** 87 - Huntsville, Alabama.
Bearden photographs twin giant radial over Huntsville AL. Cloud fades with great and unusual rapidity. A private-type jet, similar to a Learjet, continually circled under the strange twin radial cloud, as if observing and/or photographing it. *Perhaps someone in the U.S. system wised up.*

28 **Jan** 87 - Vandenberg AFB, California.
Minuteman missile successfully launched from Vandenberg AFB. Note: *Anniversary of the Challenger disaster.*

7 **Feb** 87 - Alabama - Tennessee border.
Loud aerial boom or explosion shakes area from Tennessee, New Market, Monte Sano through Huntsville Alabama area.

10 **Feb** 87 - Huntsville, Alabama.
Bearden photographs twin radial cloud in Huntsville, AL. Cloud stayed fixed, although winds present and other clouds moved. Note big storm in NE - Cape Cod, etc. *Associated with weather engineering.*

11 **Feb** 87 - Huntsville, Alabama.
Bearden photographs another twin giant radial in Huntsville, Alabama. *Shows activity in Woodpecker grid.*

Feb 87 - Vandenberg AFB, California.
USAF Titan launched from Vandenberg AFB, Calif.

12 Feb 87 - California.
Blue Angel F-18 a/c crashed near Brawley, CA. Fourth military air crash in the Imperial Valley since October. *Note Woodpecker grid activity on previous day.*

26 Feb 87 - Cape Canaveral, Florida.
GOES-H weather satellite launched from Cape Canaveral on a DELTA rocket.

26 Mar 87 - Cape Canaveral, Florida.
Stunning failure of an Atlas/Centaur rocket launched from Cape Canaveral. May have been hit by lightning. Failed suddenly 51 sec. after liftoff. Veered out of control and destroyed. *Note: Possible Soviet scalar EM involvement. Much to do later about multiple lightning strikes; very suspicious.*

June 1987 - Vandenberg AFB, California.
Minuteman 3 missile fired from Vandenberg AFB, Calif, is aborted and destroyed in flight. *Note similar incident in July.*

12 **Jun** 87 - Cape Canaveral, Florida.
Trident III launched successfully from flat pad.

21 **Jun** 87 - Alaska.
Three strong earthquakes shake parts of remote Alaskan Peninsula.

30 **Jun** 87 - Los Angeles, California.
Delta 767 flight 810 crew inadvertently shuts down engines shortly after takeoff from Los Angeles. Starts to fall, restarts engines and recovers. *Definitely crew error. Similar incident on Mar. 31, 1986 after takeoff from San Francisco.*

12 July 87 - Over Pacific Ocean.
Minuteman missile is destroyed in flight over the Pacific Ocean because of a problem in flight. Launched from Vandenberg. *Note similar incident the previous month.*

16 **Jul** 87 - Huntsville, Alabama.
Anomalous largo "boom" occurs at Huntsville, Alabama.

17 **Jul** 87 - Huntsville, Alabama.
Another anomalous, large "boom" occurs at Huntsville, Alabama. *Note that Huntsville is a pivot point on the Soviet grid. Adjustments in the grid often "spill out" energy, causing such a boom. It causes a sharp boom if pinpoint spillage occurs. It causes a more general quaking, rumbling sound of longer duration if spillage occurs over an area.*

20 July 87 - Cape Canaveral, Florida.
Tident III successfully launched from flat pad.

14 **Oct.** 87 - Nellis Air Force Gunnery range.
Top secret plane crashes; possible stealth fighter. *Did the Soviets kill another one?*

9 **Nov.** 87 - Vandenberg, AFB, California.
Minuteman II missile is destroyed seconds after launch. *Unnamed anomalies cited. Did the Soviets interfere?*

4 **Jan.** 88 - Madison and Huntsville, Alabama.
Giant radial sighted for several hours. Three mysterious booms rattle Huntsville. *Soviet scalar EM activity continues.*

APPENDIX V

*DEAD MAN FUZING**
THE REAL MEANING OF THE REYKJAVIK SUMMIT

Soviet Foreign Minister spokesman, Gennadi Gerasimov, following failure of the Iceland Summit, said in a radio interview that SDI (U.S. Strategic Defense Initiative), as presently designed, is of no concern to the Soviet Union. The real Soviet concern, he indicated, is that the U.S. would make a great technical **breakthrough** and deploy new devices in space as a modification to SDI **presenting** an unacceptable threat to the Soviet Union.

Peter Jennings' ABC News Show, Los Angeles, CA
13 October 1986.

Gerasimov let the cat out of the bag!

With their deployed scalar electromagnetic weapons, the Soviets would immediately hold the winning hand-if these weapons could be unrestrainedly brought to bear.

As pointed out in **Fer-de-Lance: A Briefing On Soviet Scalar Electromagnetic Weapons,** Tesla Book, Co., 1986, the use of massive Soviet scalar EM weapons can only be gingerly applied. Activation and use of the Soviet scalar EM weapons is fraught with grave danger to the entire earth and to the whole human species because of the presence of nuclear facilities in the desired target area, and because of the "backlash" potential against the Soviet Union's own nuclear weapons and facilities.

The explosion at Chernobyl was apparently caused by the accidental failure of a nearby scalar EM transmitter, and the eventual impulsive loss of part of the electrogravitational potential built up by the transmitter before it failed. The resulting electrogravitational pulse (EGP) in the earth "pulsed" the nearest nuclear material... feul rods in the reactor at Chernobyl.

Had the EGP not been depleted prior to its escape, all four reactors at **Chernobyl** would have exploded with the same intensity as if they had been "triggered" by a fission explosion.

---

*The term "Dead Man Fuzing", means that, even though your opponent kills you, a weapon or effect of your own will inevitably result which will then kill your opponent.

In short, if unrestrainedly used, the EGPs from the Soviet scalar EM weapons would initiate most of the nuclear weapons where they are in the targeted area-in storage, in silos, in aircraft, etc. Target area nuclear reactors and nuclear wastes would also be violently initiated into nuclear explosion. The resulting vast megatonnage of nuclear explosions, much of them "in the dirt, " would provide an unthinkable holocaust. The effects of these giant explosions, deadly fallout and dense clouds of smoke and dirt, would blanket the earth with nuclear effects greater than is presently expected from an all out nuclear war, in which many of the weapons are expected to be destroyed or lost without being exploded nuclearly. Deadly nuclear radiation would blanket the globe, contaminating it for thousands of years. From the covering of the earth by dense smoke and dirt clouds, a new and deeply bitter global ice age-not just a "nuclear winter"-would almost certainly be initiated. In 1960 Khrushchev appropriately characterized these fantastic weapons by stating that they could wipe out all life on earth if unrestrainedly used.

A country's nuclear weapons and facilities, then, become its inadvertent "dead man fuzing" to prevent any substantial employment of large scalar EM weapons against it. So long as nuclear "dead man fuzing" is in place in a country, its adversary can only utilize scalar EM superweapons against it with great difficulty, and then only sparingly. At present, both the U.S. and its NATO allies are "dead man fuzed" by the presence of Western nuclear weapons-and to some extent, by the presence of nuclear reactors and stored nuclear wastes.

Obviously the Soviet Union's task of world domination—and the mind-rending risks it must take in employing large scalar EM weapons—would be greatly expedited if the obstacle of "deployed U.S. and NATO nuclear weapons and facilities" could be eliminated or minimized.

Also, Gorbachev must be acutely aware that the U.S. is finally beginning to look at the possibility of scalar EM weaponry. He would assume that either the U.S. is already on the way to acquiring scalar EM weaponry of its own, or else the time is imminent when it will do so. If this happened, and the U. S. then were to deploy SDI **scalar** EM lasers in space, the effects of each laser would be enormously increased by the electrogravitational amplification factor. Reasonable gain increases of the yield of each laser shot might approach, for example, on the order of 10^{20}. In that case, each laser could suddenly devastate a whole state or region with one or two shots, instead of just being able to destroy one little missile. The Soviets know that Reagan's Star Wars weapons-if modified to the "improved" kind and deployed in space—could deliver a surprise first strike in seconds, and devastate Russia before all the Star Wars space weapons could be knocked out by Soviet ground-based scalar EM beams.

The Soviets also know that it is not too difficult to modify lasers to become such powerful scalar EM lasers.

This means that Gorbachev needs to insure that the Star Wars weaponry is not actually tested in space. Testing in space requires developing at least prototype "space deployment" SDI weapons, to get them up in space to test them from there in the first place. And even such "prototype" SDI weapons would be sufficient to devastate the Soviet Union, given that the prototypes employed the technological breakthrough represented by scalar EM weapons. This is what Gerasimov was really referring to!

Thus Gorbachev simply cannot allow such space testing of SDI weapons, if he hopes to be able to safely seize the opportunity to employ his own already-deployed, massive scalar EM weapons during the "window of time" that would be provided by dismantling of U.S., NATO, and Soviet nuclear weapons.

And-given that he gets agreement and implementation of the "zero option"-Gorbachev knows he will only have a narrow "window" of time in which he can hope to achieve world domination. It can be only a short time until the U.S. produces results in scalar EM weapons, and promptly develops such weapons of its own. Faced with that eventuality, Gorbachev would then have no choice but to turn away from the Communist dream of world domination, and reach a mutually agreed accommodation with the West. Both the West and the Soviets then would have to take immediate and drastic measures to prevent the rapid development and spread of the relatively cheap scalar EM weaponry throughout other nations, particularly those motivated from fanatical beliefs and quite capable of "blowing up the earth" deliberately.

Thus, before the U.S. becomes aware of scalar EM weapons and builds its own, Gorbachev desperately needs to get rid of the "dead man fuzing" problem posed by the nuclear weapons and facilities possessed by the U.S. and its Nato allies, so that he can utilize his superweapons to dominate the world without destroying it.

And he needs to block any chance that the U.S. would be able to have space-tested high energy laser prototypes which could be quickly modified to scalar EM weapons, fired into space, and used to counter Soviet aggression or even destroy the Soviet Union in retaliation.

In short, today the paramount and almost desperate need of the Soviet Union is to reduce American dead-man fuzing and keep the SDI space laser genie bottled up in the lab on the ground.

Accordingly, in Iceland, in what appeared to be a stunning concession on nuclear arms, Gorbachev suddenly offered Reagan a complete dismantling of all nuclear weapons-Reagan's "zero option." Actually this move had been planned by Gorbachev from the beginning; it was not a spur-of-the-moment offer. The reason was simple: If the U.S. could only be persuaded to quickly get rid of its nuclear weaponry, the Soviets would hold an absolutely winning hand. Their massive, deployed, operational scalar EM weapons could then be unleashed with minimal restraint. The extremely delicate and incredibly dangerous problem of bringing the large scalar EM "continent-buster" weapons to bear, without inadvertently causing world destruction, would be vastly eased if only the ordinary U.S. nuclear facilities such as nuclear power plants and radioactive wastes had to be reckoned with.

Fortunately, Ronald Reagan fervently believes in defense .

He made the right decision, whether or not it was for the right reasons.

Indeed, he probably knew in advance that SDI was to be targeted in Iceland by the Soviets, and that sweeping concessions might be offered by them for that purpose. According to one report (" 'Blinded' Satellites," Rowland Evans and Robert Novak, **Washington Post,** Oct. 29,1986), just before Reykjavik the CIA-sent a report to Reagan with the warning that "the sole reason for Gorbachev's wanting the Iceland summit was to offer unprecendented inducements for Reagan to drop SDI."

Fortunately, Reagan rejected Gorbachev's proposal, because he does not wish to give up what he truly regards as a potential defense against incoming Soviet missiles.

Apparently neither Reagan nor his advisors were aware of the real reason for Gorbachev's unexpected and breathtaking proposal for what would essentially be near-total nuclear disarmament.

The Soviet ground-based scalar EM lasers could sweep the heavens of the presently conceived SDI weapons, as **Gerasimov** obliquely implied. However, this would not change the barrier represented by U.S. nuclear dead man fuzing, unless the nuclear weapons had been removed.

Meanwhile, Gorbachev's chess game to remove the dead man fuzing is certainly not over. The Soviets have mounted an intense propaganda campaign urging nuclear disarmament and cancellation or restraint of SDI. With a Democratic Senate now in, and with the approaching end of Reagan's term in office, the Soviet effort to clear the U.S. nukes-and the dead man fuzing-out of the way for Soviet deployed superweapons, while keeping the SDI genie in its laboratory bottle, can only be intensified. Already the Soviets are beating every drum to get substantial antinuclear segments of the U.S. scientific, congressional, political, and pacifist communities aroused to oppose SDI and urge a quick zero nuclear option.

The treaty for dismantling certain missiles in Europe has already been signed by the President, and its ratification by the U.S. Senate appears imminent. More is yet to come, and quickly.

In its aching desire for peace, if an unsuspecting America falls into Gorbachev's trap, shackles SDI and removes its dead-man fuzing, the Soviet Union will be free to enforce its will wherever and whenever it wishes.* If that happens, our beloved stars and stripes will be replaced with the atheistic hammer and sickle. And the noblest experiment of all—freedom for the common person—will have come to an ignoble end.

---

*Just as **Brezhnev forecast** for 1985, **at a** secret meeting in 1972 of the communist party leaders of E**urope.**

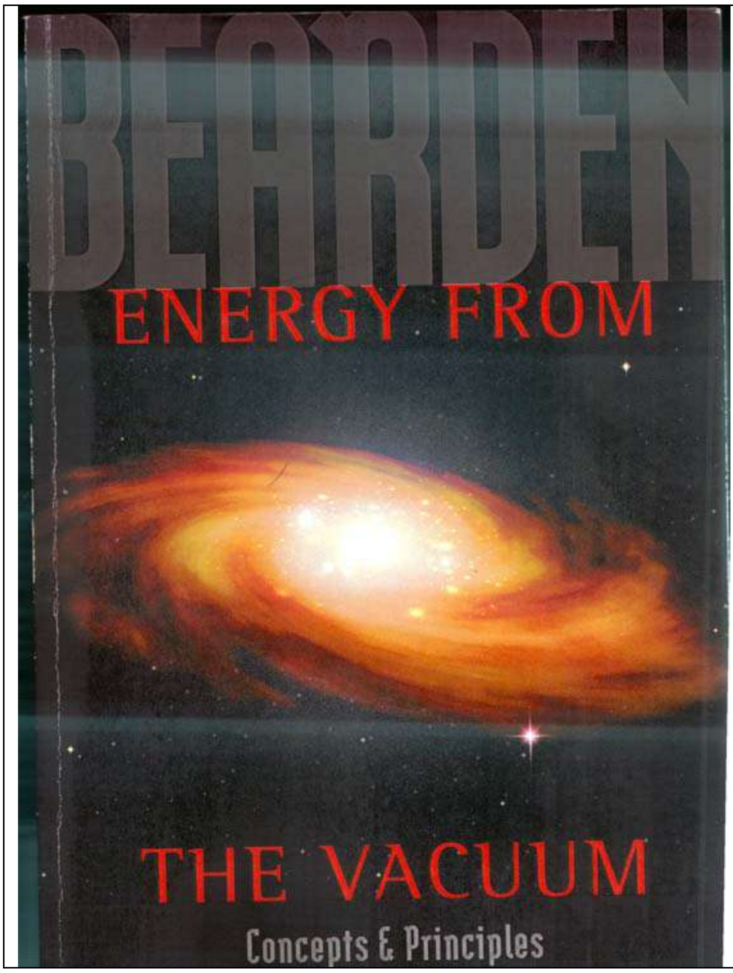

BEARDEN

ENERGY FROM

THE VACUUM

Concepts & Principles

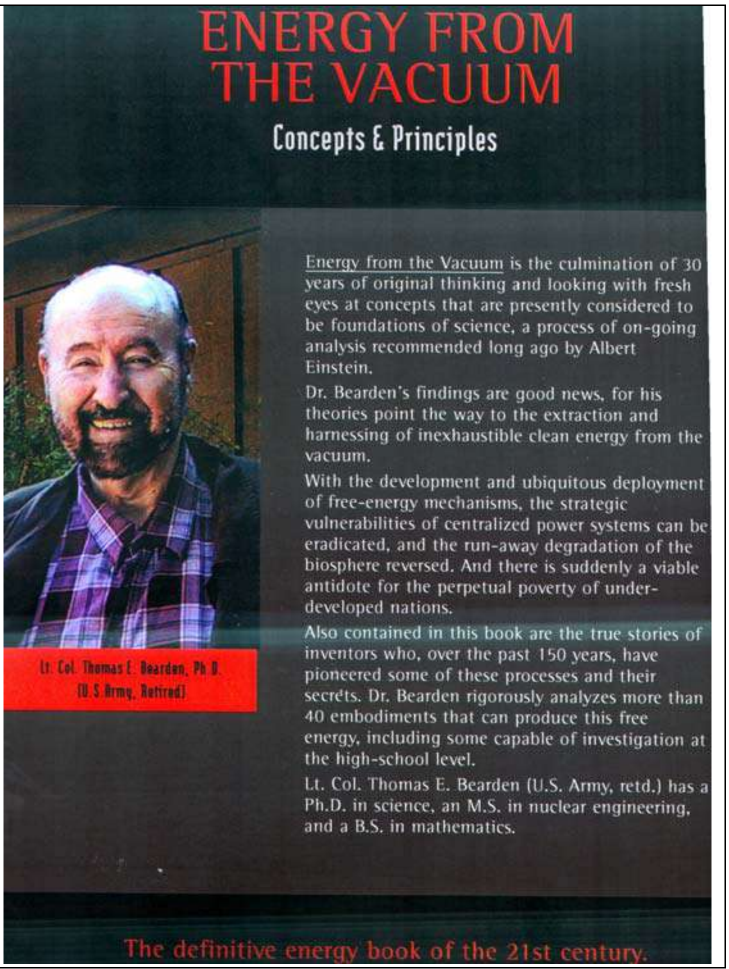

ENERGY FROM THE VACUUM

Concepts & Principles

Lt. Col. Thomas E. Bearden, Ph.D.
(U.S. Army, Retired)

Energy from the Vacuum is the culmination of 30 years of original thinking and looking with fresh eyes at concepts that are presently considered to be foundations of science, a process of on-going analysis recommended long ago by Albert Einstein.

Dr. Bearden's findings are good news, for his theories point the way to the extraction and harnessing of inexhaustible clean energy from the vacuum.

With the development and ubiquitous deployment of free-energy mechanisms, the strategic vulnerabilities of centralized power systems can be eradicated, and the run-away degradation of the biosphere reversed. And there is suddenly a viable antidote for the perpetual poverty of under-developed nations.

Also contained in this book are the true stories of inventors who, over the past 150 years, have pioneered some of these processes and their secrets. Dr. Bearden rigorously analyzes more than 40 embodiments that can produce this free energy, including some capable of investigation at the high-school level.

Lt. Col. Thomas E. Bearden (U.S. Army, retd.) has a Ph.D. in science, an M.S. in nuclear engineering, and a B.S. in mathematics.

The definitive energy book of the 21st century.

Sweating done by red_gonzo.

Note: sometimes formatting of the text may be a little fuzzy. It's all because the OCR process didn't go as smoothly as I expected. Also there may be some errors inside text so be vigilant. In any case share this book, send emails to people about knowledge contained in it. If you are an amateur electronic guy/girl try to make some good use of this knowledge and post your results. Try to use physics vocabulary as is presented in this book because as the author concludes, there are many good inventors around but they each use their own vocabulary and explanations of their inventions which are only understandable to them.
And please don't be naive and expect you'll be buying those devices in your local supermarket any time soon. Although some progress is being made it's still to slow, especially with the suppression factors watching over it. The knowledge has been handed to you and it's expected that at least a few of us make some use of it. Gather together and try to share your results with the rest of us.

TABLE OF CONTENTS:

Acknowledgements xi

Foreword xiv

Chapter 1 1

Foundations and Scientific Mindset 1

1.1 Introduction 1

1.2 Organized Science Often Resists Innovative Change 24

1.3 Scientific Disagreement or Street Fight? 32

1.4 A Few Significant and Unresolved Scientific Problems 36

1.5 Electrodynamics is Still Developing 51

1.6 A Summarizing Letter 55

Chapter 2 75

Background and Concepts 75

2.1 Comments on the Foundations of Classical Electrodynamics .. 75

2.2 Lorentz Regauging of the Maxwell-Heaviside Equations 108

2.3 The Suicidal Closed Unitary Current Loop Circuit 119

2.4 In Summary 122

Chapter 3 125

Giant Negentropy, Dark Energy, Spiral Galaxies and Acceleration of **the** Expanding Universe 125

3.1 Broken 5-Symmetry of the Dipole 127

3.2 Reinterpreting Whittaker's Decomposition of the Scalar Potential 137

3.3 **How** the Dipole Charges Transduce Time Energy 141

3.4 Reordering of the Local Vacuum: Giant Negentropy 143

3.5 Some Implications of Giant Negentropy 143

3.6 How Circuits Are Powered 153

3.7 The Heaviside Component vs. the Poynting Component 155

3.8 Dark Positive Energy: The Unaccounted Heaviside Component 157

3.9 A Short History of the Discarding of the Heaviside Dark Energy 158

3.10 The Problem of the Accelerating Expansion of the Universe **164**

3.11 In Conclusion **166**

Chapter 4 171

Setting the Stage for Understanding Overunity Power Systems 171

4.1 Introduction 173

4.2 Demonstration Gedanken Experiment **180**

4.3 Iterative Retrorefiection and Collection in a Circuit **194**

4.4 Tesla's Iterative Retroreflection Work **197**

4.5 Perspective on Tesla vs. Marconi 203

4.6 Conclusion **209**

Chapter 5 219

Selected Approaches to Overunity Power Systems 219

5.1 Introduction 219

5.2 Tesla's Shuttling of Potential Energy and Barrett's Extension 220

5.3 Moray's Radiant Energy Device 220

5.4 Anti-Stokes Emission and Similar Processes 225

5.5 Gain in Intensely Scattering Optically Active Media 227

5.6 The Fiber Fuse 230

5.7 Multiple Feedforward and Feedback Semiconductor Loops. ..235

5.8 Negative Resistors 236
5.9 Suppression of Free Energy Systems Has Been Common 288
5.10 In Conclusion 289

Chapter 6 293

Additional Approaches to Overunity Power Systems 293

6.1 Introduction 294
6.2 Energy Conversion and Energy Amplification 295
6.3 Other Systems of Electrical Interest 304
6.4 Some Magnetic Processes of Interest 341
6.5 Conclusion 381

Chapter 7 385

Aharonov-Bohm Effect, Geometric Phase, and the Motionless Electromagnetic Generator (MEG) 385

7.1 Introduction 387
7.2 Aharonov-Bohm (Geometric Phase) Effect 390
7.3 Abbreviated History of the Geometric Phase 391
7.4 Possible New Aspect of the Curl-Free Magnetic Vector **Potential** 393
7.5 Use of Aharonov-Bohm Effect in the MEG 397
7.6 Collection of Field Energy in the MEG 403
7.7 Experimental Path to Development of the MEG 408
7.8 **Additional** MEG Functions 411
7.9 In Conclusion 419

Chapter 8 421

Approach to Antigravity 421

8.1 Introduction 422
8.2 Selected Theories of Gravitation 423
8.3 Selected Antigravity and Inertial Engineering Research Efforts 431

8.4 The Author's Engineering Approach to Local Antigravity.....454
8.5 In Conclusion 471

Chapter 9 473

The Supersystem and Remarks on Gravity, Antigravity, and Testing 473

9.1 The Supersystem and Its Considerations 476
9.2 Differences between Effects in COP»1.0 and COP<1.0 EM Systems 479
9.3 Primarily Concerned With Dirac Sea Holes and Not Lattice Holes 487
9.4 Peculiarities of Spacetime Curvature and Dirac Hole Interactions 488
9.5 Charging and Discharging a Capacitor 490
9.6 The Open-Loop COP>1.0 EM Power System 495
9.7 Remarks on Dirac Sea Holes, Currents, and Negative Energy 501
9.8 Remarks on the Observed Acceleration of the Expanding Universe 505
9.9 A Solution to the Hole Current Problem and Close-Looping..509
9.10 Negative Energy: Localized and Nonlocalized 510
9.11 Tests Reinforcing the Antigravity Mechanism 514
9.12 Susceptible Unitary System without Conversion of Dirac Sea Hole Current 519
9.13 Unitary System with Conversion of Dirac Sea Hole Current 521
9.14 Non-unitary Outrigger Array System 521
9.15 Some Instrument Considerations 525
9.16 Still Anomalous Aspects 528

Chapter 10 537

Cold Fusion: Low Spatial-Energy Nuclear Reactions at High Time-Energy 537

10.1 Explaining Time as Energy....539

10.2 Mechanism Generating Flow of a Mass through Time....540

10.3 Time-Polarized EM Waves and Longitudinal EM Waves....542

10.4 Time Reversal Zone and New Nuclear Reaction Mechanisms....554

10.5 Revision of the Conservation of Energy Law....560

10.6 Some Resulting *Low Spatial Energy* Nuclear Reactions....566

10.7 Time-Energy Can Generate Instrumentation Anomalies....570

10.8 Additional Implications for the Future....577

Chapter 11 583

Final Perspective: Permissible COP>1.0 Maxwellian Systems 583

11.1 Introduction....583

11.2 Source Charges, Curved Spacetime, Efficiency, and COP....585

11.3 Truncations of Maxwell's Theory and Discard of System Classes....590

11.4 Examples of COP>1.0 Electromagnetic Systems....594

11.5 Motionless Electromagnetic Generator (MEG)....596

11.6 Physical Considerations of Gauge Freedom....599

11.7 Special Modeling of the Vacuum....600

11.8 Every Charge and Dipole in an EM System Receives and Transduces Vacuum Energy....601

11.9 Energy from the Vacuum Powers Every Electrical Power Line **and** EM Circuit....601

11.10 Significant Developments in Negative Resistance....602

11.11 In Conclusion....605

Annotated Glossary of Selected Terms 609

Appendix A 739

Appendix B 767

Index(notincludedbecauseyoucansearchinPDF)774

ENDNOTES 849

Acknowledgements

This book is dedicated to those lone researchers and stalwart inventors who have succeeded in building working model COP>1.0 electrical systems in the face of overwhelming odds and strong scientific opposition. We particularly mention Gabriel Kron , T. Henry Moray, John Bedini, Jim Watson, Ed Gray, Frank Golden, Floyd Sweet, Howard Johnson, Ken Shoulders, Bill Fogal, the Correas, and Teruo Kawai. There are quite a few others, of course, and no slight is intended by not listing all their names.

Also, I particularly acknowledge the contributions of my long-suffering close colleagues Dr. Lee Kenny, Dr. Jim Hayes, Ken Moore, and Steve Patrick, whose efforts on the bench have produced the working motionless electromagnetic generator (MEG) experiments as well as several variants. Our first MEG patent was issued in April 2002 and a second patent application has been filed. The MEG represents the first practical macroscopic application of the geometric phase (extended Aharonov-Bohm effect) for EM power systems. I particularly thank Lt. Col. (Retired) Ken Moore for watching my back and keeping me alive. I'm also grateful to both Dr. Fred Wood Sr. and Dr. Fred Wood Jr. for their personal support, action, and encouragement. I particularly acknowledge more than two decades of close cooperation and work with John Bedini, who is one of the most creative inventors on this planet; nearly a decade of work with Floyd Sweet and his vacuum triode amplifier; more than a decade of work with Frank Golden in COP>1.0 motors and uncurled A-potential communication systems; and a special thanks to Don Jensen for his important contributions and work in special combustion systems. Working with these inventors, I was able to physically examine working COP>1.0 power system prototypes and also experience and study the novel phenomenology that results in such systems.

With deep appreciation I acknowledge the kind assistance and helpful critiques of Dr. Myron W. Evans and other members of the Alpha Foundation's Institute for Advanced Study (AIAS) such as Alex Labounsky, Alain Beaulieu, and Bob Flower. Dr. Evans encouraged me to write down the basic concepts of the theory of permissible COP>1.0 Maxwellian systems I have been working on for three decades. This book is the result of that effort.

A special note of thanks goes to Dr. Dave Hamilton of the Department of Energy, who oversees a DoE restricted website provided for the important

ongoing scientific work **and drafts of the AI**AS. On that site, now there are more **than** 100 AIAS papers, many **of them dealing** with EM energy from the vacuum in rigorous fashion. Many of these AIAS draft papers have been published in leading journals such as ***Foundations*** *of Physics Letters, Physica Scripta, Optik,* etc. **This** is an invaluable scientific effort that clearly establishes the importance and potential for extracting copious EM energy directly from the active vacuum. It also extends the present U(l) electrodynamics to the much more extensive O(3) group symmetry electrodynamics, in which such things as I discuss in this book can indeed be accomplished on the bench and mathematically modeled.

The outstanding work and encouragement of Dr. Hal Fox and Dr. Eugene Mallove are particularly recognized, and I am grateful to them for their continued publication of important information and developments in cold fusion and the entire field of anomalous energy systems. I am also grateful to Dr. Robert Flower for his cogent explanations of some tough issues and points in quantum field theory that were beyond my ken before his explanation. Dr. Hal Puthoff introduced me to his cosmological feedback mechanism, and I am grateful for his important work with his colleagues in the areas of inertia, permissible thermodynamic extraction of energy from the vacuum, and possible mechanisms of antigravity.

A special thanks goes to Tim Ventura and to Jeffrey Cameron and the engineers of Transdimensional Technologies, Inc. for their courtesy and demonstrations of Transdimensional's successful lifter and rotary devices, and for an introduction to the field of inertial propulsion and antigravity now ongoing under the NASA Breakthrough Propulsion Physics (BPP) research program.

To Marcia Stockton I owe very special thanks indeed for her tireless efforts in reading and proofing the manuscript and translating the graphics. Also, I am much indebted to Michael Anderson for the art direction, and to Tony Craddock for making it possible. I am deeply grateful to Tony for his unfailing encouragement and support, and for operating and maintaining my website www.cheniere.org. I also deeply appreciate the continued contributions of Mike Rieker, of Network Innovations, who furnishes the website pro bono. A special thanks also goes to Dr. Steve Greer, who gave my wife and I great assistance and personal uplift at a time of personal tragedy. Dr. Greer has also now taken up the effort to develop and deploy practical electrical power systems, freely taking their EM energy from the vacuum.

Stalwart modern researchers in the overunity movement such as Jerry Decker and Walter Rosenthal are particularly to be commended for both their work, assistance to others, and their communication of work in the field. I am much indebted to Walt for furnishing the pictures of the closed-loop Sweet vacuum triode amplifier used in this book.

To Jean-Louis Naudin, I also am very grateful for his extraordinary experimental work and extensive publication of valuable information, his personal replication of a variety of systems, and his kind permission to use several illustrations in the book. His important website, http://jnaudin.free.fr/, is a cornucopia of information, experiments, and results in this entire area, with important links to many other appropriate web sites.

Also, to more than a dozen special persons — including those now passed on such as Glenn Foster and R. J. Reynolds III — who encouraged and supported these efforts through the years, I am very grateful. Throughout the years their kind words and friendship have made a very great difference to me personally and professionally, and they still do.

Finally and most importantly, I acknowledge the continuing and unwavering support of my dear wife, Doris, without whose encouragement and tolerance this work could never have been accomplished. She made all these long years of seven-day weeks and 18-hour days possible and humanly bearable. Without her, this book simply could not have been written.

Foreword

The original purely scientific purpose of this book was interrupted en route by two events: The first was a sudden heart attack and onslaught of severe hypoxia resulting from the resurgence in Spring 2001 of a 33-year old chronic mycoplasma infection, of the modified biological warfare kind produced by a combined U.S./Canadian team, which got loose in Canada and Florida in the 1950s. I contracted this "mysterious, unknown ailment" in 1968 while in the U.S. Army and stationed in Quebec, Canada as the U.S. Army Liaison Officer to the Canadian Armament Research & Development Establishment (CARDE). The second event was the sudden terrorist attack on the U.S. on September 11, 2001, which propelled this nation and others into a war against international terrorism.

For the mycoplasma, Kenn Thomas recognized and informed me exactly what the difficulty was, and I was finally tested and diagnosed correctly in December 2001 due to the efforts of a courageous family physician. Accordingly, in December 2001 we started a yearlong (or longer) regimen of antibiotics and supplementary oxygen, with additional medication to control the runaway heart fibrillation associated with this virulent form of the disorder. If we make it through this yearlong program, we will probably survive with at least some significant recovery. Now on a much-reduced regimen, this made me realize the absolute necessity of passing the "energy from the vacuum" information and baton, so to speak, to those sharp young graduate students and post doctoral scientists who are interested in this area. They simply must not have to spend 30 years of their lives getting to this point. Instead, they should start where I am now, correct any errors I have inadvertently made, and go much farther.

Accordingly, we sharply changed the intent and content of the book to comply with that new purpose. We also deeply considered the potential U.S. and World energy implications of that second shocking event of 9/11 when international terrorists struck the World Trade Center in New York Center and the Pentagon, having also targeted the White House.

In the new asymmetric form of warfare, the internal aspects of a nation - its civilian population, its vulnerable installations, its energy infrastructure, its food crops, and its transportation modes — are the targets. Terrorist teams and groups in significant numbers are already inside every targeted developed nation, and these teams have weapons of mass destruction including biological weapons and even nuclear weapons. For confirmation of the hostile nuclear weapons already secreted in large U.S. cities, read Lunev's book[1] to see how the former Soviet Union brought them in, along with the Spetznaz teams to detonate them on call. The nuclear weapons and the Spetznaz teams are still here. Other hostile nations have inserted terrorist teams with weapons such as anthrax, smallpox (camel pox is nearly identical to smallpox, and widely available), bubonic plague, etc. At the end of the 1991 Persian Gulf War, most of Iraq's calutrons and its nuclear program escaped the weapon inspectors and were hidden. In the few calutrons we did get, weapons-grade U235 residues were found. In a formal threat study for a foreign nation, begun before the Gulf War and delivered after its end, my analysis team and I estimated that Iraq had produced sufficient U235 in 1991 for from 7 to 12 nuclear weapons, particularly of the gun assembly type. My personal estimate now would be that Iraq has made sufficient weapons grade U235 to build three dozen to six dozen nuclear weapons. If so, then some of those weapons are undoubtedly long since spirited into the United States, and planted in our large cities.

In modern asymmetric strategic war, the first phase of the war is conducted in "peacetime". The first phase is to insert the weapons of destruction into the targeted nation and hide them at the intended targets, along with the teams to detonate and unleash the weapons when desired. In that sense, the first phase of WW III has already been accomplished.

According to Vice-President Cheney,[2] *"The war on terrorism will not be over in our lifetime. It is different than the Gulf War was in the sense that it may* *never* *end. At least not in our lifetime. The way I think of it is, it's a new normalcy."*

With much of the cheap oil of the world in its borders, the MidEast is a powder keg ready to explode. If a clash between the U.S. and Iraq occurs, Saddam Hussein has already clearly shown that he will attack and destroy his foe's energy infrastructure and he will also use weapons of mass destruction. Simply witness the hundreds of oil wells set afire in Kuwait by Iraqi forces during the Persian Gulf War, and the use of chemical weapons on his own people. As this is being written, Pakistan and India are skirting the brink of war, and both have nuclear weapons. Either side that decides on war **will have to make a n**uclear strike on the other. They have no antimissile defenses, hence **making the** "first knock-out" or pre-emptive strike is each **nation's** only **chance of survival.** Hopefully cooler heads will prevail, but **If that** war erupts, the Mid East, oil, and present energy infrastructures will never again be the same, for the entire world.

---

[1] Stanislov Lunev and Ira Winkler, *Through the Eyes of the Enemy,* Regnery, Washington, D.C., 1998, p. 26.

[2] Vice President Richard B. Cheney, October 21, 2001.

There are other very serious U.S. vulnerabilities of gas and oil pipelines, port facilities, Gulf of Mexico oil rigs, nuclear power plants, power substations, hydroelectric dams, long power transmission lines on towers, etc. that are too numerous to enumerate. With but a moment's reflection, one realizes that even normal and cheap guerrilla-type or commando-type forces with C-4 explosives and timers can destroy a very substantial portion of the energy infrastructure with minimal risk and loss of assets. For those who wonder about a nuclear power plant's vulnerability, visualize a seized Boeing 747 loaded with fuel and flown directly into the main reactor container . If meltdown occurs, the nuclear tragedy at Chernobyl in 1986 will appear tame by comparison. There are also several other electromagnetic ways (such as a portable EMP "shooter" hauled in the back of a personal SUV) to instantly destroy the electronic controls of the nuclear power plant, seriously risking runaway and core meltdown.

Consequently, increasing attention is now focused on the extreme vulnerability of the centralized energy infrastructure in the U.S. and every other developed nation. So what started for me as a purely scientific endeavor to set out the concepts and principles of permissible COP>1.0 EM systems suddenly assumed a role of potentially critical strategic significance to the U.S. and to the entire developed world. Quite simply, some 20,000 terrorists in teams already in the United States can - at this very moment —destroy large portions of the nation's energy infrastructure at will, easily and cheaply. The economy of any modern developed nation depends on the availability of cheap energy. So this kind of hostile strategic capability means that the power to readily destroy a nation's economy — and thus defeat a mighty nation itself in the ensuing economic collapse — has passed into the hands of hostile forces already in country and awaiting the order to commit their assets.

In this new asymmetrical form of warfare, it would appear that the rapid development and deployment of significant distributed power sources is now a critical element of national survival for many nations of the world. Decentralization **does** little **good, however, if fuel** for the decentralized generators must remain **centralized and distributed** by very large and complex systems.The **fuel and Its distribution** are still extremely **vulnerable resenting** easy **targets allowing paralysis** of the decentralized generators. Self-refueling *{self-powering)* of the electrical power systems is therefore a mandatory additional requirement.

So sober assessment dramatically reveals the need for self-powering electrical power systems, freely extracting their EM energy from the local vacuum. Thus an engineerable theory of permissible COP>1.0 EM systems has become essential to the national survival of the U.S. and many other nations, so that an effective decentralized energy infrastructure — with the seething vacuum as the "localized and inexhaustible free fuel supply" — can be rapidly designed, developed, and deployed.

Sadly, the leaders of the scientific community are still in a "business as usual" mode with respect to energy, and intend to remain so. They do not even recognize what powers and *always has* powered an EM circuit and system: EM field and potential energy extracted directly from the vacuum by the broken symmetry of the source charges and dipoles. Since no scientific research on EM energy from the vacuum is going to be launched by the scientific leaders at the top, then the results of the broken symmetry discovered in 1957 are going to have to be applied in the scientific community from the bottom up. Consequently, a rather desperate need to communicate everything possible to the young graduate students and post-docs has become imperative, if the nation is to survive.

From the beginning of my effort in COP>1.0 systems, and then my effort on this book and prior to the events leading to the attack on September 11, 2001, the appalling *humanitarian* need for COP>1.0 electrical power systems taking their energy from the vacuum has been obvious and overwhelming. One third of the human population usually goes to bed hungry at night. One third is infested with worms and parasites of one kind or another, in addition to other diseases. Simple things such as clean water, common sanitation, medical treatment, and jobs to support one's family are difficult things in many parts of the world.

In the final analysis, the only way to get a suffering, impoverished, and undeveloped nation up and going is to provide cheap clean energy, and to do it with systems which get their "fuel" for free, such as from the vacuum itself. Once cheap clean energy and self-powering systems are widely available, then the economy and infrastructure of that nation can be slowly and steadily built up to get the society moving. Roads, bridges, hospitals, schools, teachers, industry, and jobs then follow as the economy rises. Until there is a cheap energy infrastructure available and growing, essentially fuel-free, none of that is going to happen in most impoverished nations because presently the world's fuel is monopolized, controlled, and

fought over, and it will become even more expensive as the supply vs. **demand** ratio falls, prices escalate, and additional wars - almost certainly involving weapons of mass **destruction** - are engendered. With present aid concepts, at best the politicians and warlords usually get the bulk of whatever assistance is sent in, and the warlords on the take from foreign masters will continually plunge the country into turmoil, war, and human tragedy. *Assistance is necessary,* but the eventual aim has to be to enable and empower the common citizens to be self-supporting, educated and trained, relatively free from disease, and living in a nation having a growing and developing infrastructure, jobs, and a decent economy. Availability of cheap and affordable energy — and *clean* energy if we wish to save the precious environment — is the primary key to "freeing and empowering the developing nations" as well.

Along with this omnipresent, desperate humanitarian need, the increased strategic need for decentralized energy systems freely fueled from the vacuum itself has led me to make every effort possible — particularly after the heart attack in 2001 and in continued hypoxia from severe chronic mycoplasma infection — to finish the book and at least outline the major concepts and principles that enable the construction of a sorely-needed COP>1.0 electrical power systems theory, followed by the development and rapid deployment of actual self-powering systems. The Alpha Foundation's Institute for Advanced Study (AIAS), spearheaded by Dr. Myron Evans, has also been vigorously preparing and publishing papers in leading scientific journals and other publications[3] showing that EM energy from the vacuum is permissible, possible, and feasible.

In my three decades of work with multiple inventors of working prototype COP>1.0 electrical systems, I have been privileged to experience and work with several actual COP>1.0 circuits and prototype systems, study them, and slowly comprehend some of the advanced and unusual principles involved. In short, thanks to the courtesy and kindness of others, I have been able to gradually acquire at least an elementary understanding of the mechanisms which enable COP>1.0 EM power systems. With this background, one can understand why conventional EM systems and circuits do not develop COP>1.0 or self-powering, even though all the energy collected and used in every electrical power system comes from the vacuum and not from the shaft energy input to the generator, chemical energy dissipated in a battery, wind energy collected by a windmill (that just "feeds" the shaft of the generator), or from the water power dissipated in a hydro turbine (the hydro turbine also only turns the shaft of the generator, which does not actually furnish the power to the external power line connected to it).

---

[3] E.g, *Modern Nonlinear Optics,* Second Edition, 3 Vols., edited by M. W. Evans, Wiley, 2001; M. W. Evans et al, "Derivation of 0(3) Electrodynamics from the Irreducible Representations of the Einstein Group," *Found. Phys. Lett.,* 15(2), Apr. 2002, p. 179-187; "Anti-Gravity Effects in the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* 14(6), Dec. 2001, p. 601-605; "Classical Electrodynamics Without **the Lorentz Condition: Extracting Energy** from the Vacuum," *Physica* ***Scripta,*** **61(5),** May 2000. p. 513-517; "Runaway Solutions of the Lehnert Equations: The **Possibility of Extracting Energy from the Vacuum,"** *Optik,* 111(9), 2000, p. 407-409.

It seems appropriate that other researchers should be able to start where 1 am now, understand what I have learned or think I have learned, find and correct any errors, and simply go forward with alacrity — much farther than I have been able to go, and much faster. This is particularly true of the young researchers who have excellent theoretical and experimental skills and a vigorous lifetime ahead of them. When an old and tiring runner falters with the baton, it is time to pass it on to the fresh young runners who can carry it to the successful end of the race.

The purpose of this book, then, is to lay out in some little detail what I have learned in all these years of sustained effort, but in such a way that it ties directly into good physics. In some cases, the tie-in to physics is unavoidably to a somewhat esoteric area such as the geometric phase and to time as a special form of energy. COP>1.0 EM systems are not simple, and they involve special phenomenology rarely encountered in conventional EM circuits and systems. COP>1.0 EM systems simultaneously involve a combination of particle physics, modern active vacuum theory, advanced materials science, higher group symmetry electrodynamics, quantum mechanics, and general relativity. They also involve areas where foundations of physics are presently in error. We found it necessary to think in terms of the *supersystem,* defined as having three components, which are (i) the physical power system itself, (ii) the active local vacuum, and (iii) the active local curvatures of spacetime. All three components of the supersystem interact and exchange energy with each other continuously. Any legitimate COP>1.0 power system must freely receive, collect, and use some excess energy from its active environment — in this case, the latter two components of its supersystem. By definition, this COP>1.0 function requires disequilibrium between the system and its active environment, so that classical equilibrium thermodynamics does not apply. Instead, the thermodynamics of systems far from **equilibrium** in **their** energetic exchange with an active environment applies.

However, since the "environmental energy" we are extracting is virtual photon energy from the vacuum and energy from local curved spacetime, the fundamental definitions of classical thermodynamics and classical electrodynamics have to be re-examined and altered. A closed system, e.g., can no longer be defined just as one closed to the transfer of mass across its boundary, but permitting energy to exchange across it. Since the advent of general relativity in 1915, we have known that energy and mass are the same thing. Whenever any system changes its energy, it also changes its mass by the well-known $E = mc^2$ rewritten as $dm = dE/(c^2)$. Further, in the fierce virtual particle exchange of the vacuum with every part of every system, both virtual mass and virtual energy continually cross the system boundary. Neither classical thermodynamics nor classical electrodynamics deals with the transduction of time-energy into spatial energy in a system, which is a totally different type of energy exchange accomplished by charge.

Further, certain phenomenological areas — such as strong gradients — are already well known to violate present thermodynamics. These are presently being researched under the aegis of *extended nonreversible thermodynamics.* Finally, as the present laws of classical thermodynamics are stated in the textbooks, *every charge and dipole in the universe already violates those laws, and has done so since the beginning of the universe.* Let us put it bluntly: Every charge in the universe already freely and continuously pours out EM energy in 3-space in all directions, without any observable EM energy input. That is the well-concealed *source charge problem,* known but ignored by the leaders of the scientific community for. a century. *All* EM fields and potentials and their energy come from those source charges, according to electrodynamics itself. Either we must give up the conservation of energy law entirely, or else we must accept the fact that unobservable *virtual* EM mass and energy are continuously absorbed

The vacuum by the source charge, transduced into real observable EM energy, and then re-radiated in 3-space in all directions as observable EM energy, creating the associated fields and potentials reaching out across the universe. That this problem continues to be ignored by the scientific community — some 45 years after the basis for its solution has been proven and is well known in particle physics — is totally inexplicable. But it continues to be resoundingly ignored, which means that *the scientific community itself is and remains totally responsible for the energy crisis and the associated vast desp**oilment of the** biosphere, because it adamantly* *insists on using a seriously flawed electrical energy science whose hoary basis is more than a century old.*

Accordingly, we have simply extended the first law of thermodynamics appropriately, pointed out that the second law does not apply absolutely but only as an approximation when conditions are appropriate, and revised the third law and the zeroth law. In this book we simply show the results of that revision, particularly in Appendix A.

We were also able to trace the history of electrodynamics, and to find exactly how the permissible Maxwellian systems far from equilibrium with an external active environment — which are included in Maxwell's 1865 seminal theory and even in Heaviside's vector truncation later — were arbitrarily discarded by Lorentz with his symmetrical regauging of the Maxwell-Heaviside equations. The purpose of such arbitrary regauging was and is to provide simpler equations that can usually be solved analytically without resorting to numerical methods. In the name of mathematical simplicity and ease, that entire great class of Maxwellian systems far from thermodynamic equilibrium in their supersystem was just arbitrarily discarded. Those Maxwellian systems are still arbitrarily discarded in every electrical power-engineering textbook today.

So the first requirement for a COP>1.0 EM system becomes *building-in a mechanism to violate the Lorentz symmetry condition, at least for a significant fraction of the time.* Further, the standard closed current loop circuit — including the entire external circuit in the same closed current loop as the source dipole inside the generator is what already *self-enforces* the Lorentz symmetry condition. Consequently, in any legitimate COP>1.0 electrical system, the operation of that closed current loop circuit must be violated for some portion of the system's operation. This of course places the system in disequilibrium with its active supersystem environment, allowing intake of extra energy from the environment and therefore permitting COP>1.0. So one of the immediate requirements in COP>1.0 electrical systems research is to seek out and find known or suspected mechanisms that accomplish just such violation of the Lorentz symmetry condition. In the book, I show a sampling of the various mechanisms I found other inventors using or that I uncovered myself, or that already exist in physics.

Finally, we discovered (and found the basis for it was shown in the hard literature in Heaviside's original work) that every generator and battery as an energy transducer already extracts from the vacuum and outputs from its terminals a far greater *nondiverged* energy flow than is accounted in the

Poynting diverged energy flow theory (which captures only that small **fraction that** is **diverged** into the circuit and captured by it). The total output energy flow from the terminals of a generator is also far greater than the mechanical energy input to the shaft of the generator. So in its *energy transduction* function, every dipolar electrical circuit ever built is already a "free energy circuit" and a COP >>1.0 system, extracting its utilized electrical energy from the vacuum and extracting additional energy also. But presently the electrical power system is unwittingly designed and constructed by engineers to deliberately destroy the source of that enormous energy flow from the vacuum faster than the system can use some of it to power its load. Wryly put, that does not seem to be the proper way to build energy systems or "run an energy railroad," even though it's always been done that way.

This led to recognition of a great scientific faux pas, perhaps the greatest in history. One must keep one's sense of humor! So with tongue in cheek, let us state it this way: *There is not now, and there never has been, a single electrical engineering department, electrical engineering professor, or electrical engineering textbook that teaches what powers an electrical circuit. Yet the basis for it has clearly been in particle physics since 1957, as certified by the award of the Nobel Prize to Lee and Yang. What powers every electrical circuit and system is the broken symmetry of the primary power system's source charges and source dipole (i. e., the broken symmetry of the opposite charges on the ends of the dipole) in its fierce energy exchange with the seething vacuum, once that dipole has been made. The modern "charge " is also a dipolarity, once its associated clustering virtual charges of opposite sign are considered. Further, any dipole will freely extract EM energy from the vacuum, convert it into real, usable, observable EM energy, and pour out real, usable EM energy continuously and indefinitely, so long as the dipole remains intact.*

The external circuit attached to a generator is not powered by the mechanical energy that is input to the generator shaft. The external circuit attached to a battery is not powered by the chemical energy expended by the battery. Astonishingly, every electrical power system and electrical circuit ever built, and every one built today, was and is powered by energy extracted and transduced directly from the active vacuum by the broken symmetry of the opposite charges on the ends of the source dipole (or a charge considered as a modern dipolarity), once that dipole is formed. A Nobel Prize was awarded to Lee and Yang in 1957 for the prediction of broken symmetry (strongly in 1956). Wu and her colleagues proved it **experimentally** in early 1957. So revolutionary a change in physics was that startling discovery of broken symmetry, that — in a nearly unparalleled action — the Nobel Committee awarded the Nobel Prize the same year, December 1957, to Lee and Yang. *One of the broken symmetries proven was that of opposite charges such as are on the opposite ends of every dipole, or such as constitute a classical "isolated charge" when the charge and its clustering virtual charges of opposite sign are considered as in modern physics theory. Rigorously, every source dipole is continuously extracting EM energy from the vacuum and pouring it out of the "terminals" of the dipole in all directions in 3-space. And so is every charge as well.* That the leaders of the scientific community continue to ignore this solidly proven fact is astonishing. But they do.

We also came to realize that the great Gabriel Kron, who was never allowed to reveal the secret of his *open path,* had in fact discovered the precursor of that very broken symmetry of opposite charges (of dipolarity). His open path is simply the recognition that any two points in the universe, that are at differing potentials and therefore can be considered as a dipolarity of opposite charges, acts as a source dipole that already extracts virtual photon energy from the vacuum, transduces it into real, observable photon energy, and pours out that energy in all directions. In reaching that understanding, he had applied general relativity to rotating electrical machines, simulated various forms of electromagnetics equations, and dealt deeply with quantum mechanics and the Schrodinger equation.

Sadly, the electrodynamics model used in electrical power engineering still does not even model the active vacuum, *much less* a broken symmetry in the virtual photon exchange between that active vacuum and all the charges in every electrical system. The electrical engineering model does not model the *supersystem,* hence cannot be used to design and produce COP>1.0 systems. *A priori,* that seriously truncated and crippled model arbitrarily discards all permissible Maxwellian systems far from thermodynamic equilibrium (i.e., with broken symmetry) in their vacuum flux exchange. And those are precisely the systems that are permitted to perform the five magic functions of great interest to the COP>1.0 energy researcher. Such a disequilibrium system can permissibly (i) self-order, (ii) self-rotate or self-oscillate, (iii) output more energy than the operator inputs (the excess energy is freely received from the active environment), (iv) power itself and its load simultaneously (all the energy is freely received from the active environment), and (v) exhibit negentropy. *Every charge and dipole in the universe already performs all five magic functions.*

It is **the purpose** of this book **to** clearly show that all EM systems are already powered by vacuum energy, and always have been. We cite the references and experimental proof that this is true, and the interested reader can check the extensive references himself or herself. The intent is to show that only a part of what has been discovered and is already known in particle physics is presently included in the woefully inadequate and ancient electrical power-engineering model. And frankly, unless there is a groundswell from the bottom of the scientific community, the leadership of the scientific community has no intention whatsoever of funding or allowing significant research in extracting copious EM energy from the vacuum (very simple) and learning to use it effectively (more difficult). A single 200 million dollars of good research in EM energy from the vacuum would solve the energy crisis forever, and dramatically reduce the pollution of the biosphere. The scientific community will spend many times that amount on their favored "big science" projects, but not one cent on EM energy systems powered by energy extracted from the vacuum.

Also, we show a sampling of the present work ongoing in inertial propulsion and antigravity research. Working with Sweet, in the 1980s I designed and Sweet performed a highly successful antigravity experiment.[4] I share with the reader the concepts and theoretical basis for that experiment, why it worked and how it worked, and the excellent results achieved. One of the tragic losses to science was the loss of the Sweet COP»1.0 vacuum triode amplifier (VTA) system with the death of Floyd Sweet. Nonetheless, we apply the mechanism we conceived so long ago to propose a mechanism generating the antigravity recently confirmed to be accelerating the expansion of the universe. We offer the long-neglected, Lorentz-discarded giant nondiverged Heaviside energy flow component accompanying every field-charge interaction as the probable generatrix of the excess gravity holding the arms of the spiral galaxies together. And we offer that same component — when connected with a flow of *negative* EM energy and with *negative energy* EM fields and potentials — as the specific generatrix of the excess antigravity that is accelerating the expansion of the universe. Both theses are testable on the laboratory bench.

Our target audience is several fold. Even though presented conceptually, the material is not oversimplified, and it is targeted at (i) the sharp young grad students and post-docs interested in extended (higher group internal symmetry) electrodynamics and in researching and developing COP>1.0 electrical power systems, (ii) the forward-looking electrical engineering professors who will at least consider that extended electrodynamic models and the laws of nature do permit COP>1.0 electrical power systems, and (iii) the still open-minded leaders of the scientific community, such as the National Academy of Sciences, National Science Foundation, the U.S. Department of Energy, the great national laboratories, etc. In *one area* of physics much better models of electrodynamics enabling COP>1.0 power systems, antigravity, and novel propulsion systems already exist (and a few such systems *experimentally* exist in the hands of independent inventors as well), while in a *second area* of physics that part of electrodynamics dealing with permissible COP>1.0 EM systems and unified field theory has been arbitrarily excluded at great cost to society, to the environment, and a tragedy to science and technology. Additional secondary targets are (iv) governmental leaders responsible for advanced research and development planning and funding, to argue that the "energy crisis" is a direct result of an inexplicable and sustained shortfall in the scientific mindset, and not as a result of the laws of nature and the need for cheap oil, coal, and gas or nuclear power plants, (v) science writers, journalists, and entrepreneurs to sensitize them to (hopefully) a coming energy revolution of unparalleled proportion, (vi) strategic and tactical planners in our Department of Defense to make them aware that the great logistics burden of modern highly maneuverable armed forces can and will be dramatically lowered by rapidly developing and using self-powering power systems "fueled" directly by the local active vacuum, and by adapting novel new propulsion methods, (vii) theoretical and experimental physicists to make them aware that the easiest thing in the world is to extract copious EM energy freely from the vacuum; just pay *once* to make a little dipole or assemble a charge and do not let it be destroyed, (viii) the Department of Energy upper level managers to bring into their awareness the coming advent of (and terrible need for) self-powering decentralized EM power systems worldwide, and the pressing need for a major role to be played by the DoE, (ix) the struggling but serious COP>1.0 researchers and inventors, to give them a technical overview — rather than just anecdotal material — of much of what has been done, tried, and conceived, and what seems to be successful and what is still to be determined, (x) the environmentalists devoted to cleaning up the biosphere and saving this precious planet, to make them aware that the answer to the world's energy problems and saving the biosphere cannot come to them from their conventional scientific consultants, who are trained in applying the very scientific disciplines responsible for the problem in the first place, and (xi) **the long suffering undeveloped** and impoverished nations and peoples of the earth. Our fervent message is one of hope after all, *once a* ***change*** *occurs in the scientific mindset.* There is a cheaper, cleaner, and better way to get their economies and nalional infrastructures up and running, and it is imminently doable.

---

[4] Floyd Sweet and T. **E. Bearden,** "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," *Proc 26th Intersoc.* ***Energy Conversion*** *Eng. Conf. (IECEC '91),* Boston, Massachussets, 1991, p. 370-375.

The main message of this book is that there is not now and there never has been an *electrical energy availability* problem. Every large centralized power generator actually outputs enough EM energy flow, extracted directly from the seething vacuum, to power all electrical loads on earth, if all the energy pouring from its terminals were caught and utilized. However, there is certainly an *EM energy flow interception* problem, to get more of the available energy flow from a source dipole intercepted and diverged into the circuit in a manner not complying with the Lorentz symmetrical regauging condition. Also, once the excess energy has entered the circuit and been collected, there certainly is a problem in *using the intercepted and collected energy in the circuit without destroying that part of the generator or battery — the source dipole — that is actually extracting all the energy from the vacuum.* Or at least to use the circuit's collected energy in such fashion as to destroy the source dipole in the generator *slower* than the energy is used to power the external load.

That conglomerate of the two sub problems — the "energy flow *interception, collection, and usage* problem — is the single energy problem. There is no other energy problem, and there has never been any other. One finds it very ironic that, in the conventional agencies, universities, the DoE, the National Science Foundation, the National Academy of Sciences, the national laboratories, etc., not one of them is working on the real, solitary electrical energy problem, even though they are importantly spending billions of dollars purporting to do so.

We fervently hope that the content of the book will reach a sufficiently interested audience to begin to change that century-old erroneous scientific mindset. We hope that the book, along with the important contributions of other COP>1.0 inventors and researchers having success, will spark a new creative drive in the scientific spirit that solves the problem of cheap, clean energy in three years or less. This is a doable. I firmly believe in the scientific method and what it can do once it is unleashed from the century-old dogma that COP>1.0 systems are against the laws of nature, physics, and thermodynamics. Just turn the young scientific and engineering tigers loose on the problem, give them some funding, and watch them do it.

As we stated in a formal rebuttal to a charge of perpetual motion nonsense, it only takes one white crow to prove that not all crows are black. And with respect to COP>1.0 EM systems far from thermodynamic equilibrium with their active environment, every charge and dipole in the universe is already a "white crow" pouring out EM energy freely and continuously with the speed of light, and exhibiting COP = ∞. This is in fact reluctantly recognized in physics, but it has essentially been buried from the attention and knowledge of generations of electrical engineers and most physicists.

A COP>1.0 experiment well known to nonlinear optics is the Bohren-type experiment, demonstrating the known negative resonance absorption of the medium. A typical experiment of that type outputs some 18 times as much EM energy as the experimenter inputs by standard calculations. We note that a single replicable experiment can and does negate any theoretical model that contradicts it. Hundreds of such materials experiments proving the feasibility of COP>1.0 EM systems are performed every year in any major nonlinear optics laboratory without a second thought. Together with the known but avoided source charge problem and proven broken symmetry of any dipolarity, it is almost unthinkable that our electrical engineering departments still do not incorporate the active vacuum exchange — and every EM circuit's broken symmetry in it — into the classical electrodynamic model used. In short, to really understand what powers every EM circuit, just now the electrical engineer would have to go across campus to the particle physics department and investigate broken symmetry of opposite charges — as are on the ends of the source dipole, once created in the generator or the battery, or as are involved with the single "isolated charge" considered with its associated virtual charges of opposite sign.

It is high time that our electrical power engineers and scientists heed the message of all these demonstrated white crows of nature, and begin developing and deploying self-powering COP>1.0 electrical power systems with alacrity.

We have a biosphere, a planet, a strong world economy, and an uplifting for the long downtrodden poor people of the world to recover. We strongly urge the scientific community to accept the challenge and meet it head-on.

Tom Bearden

22 July 2002.

Chapter 1

Foundations and Scientific Mindset

> *"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."* [Albert Einstein]{1}

1.1 Introduction

Einstein's quote is a beautiful statement that scientists should never take the present understanding — and the present models — as absolute. That is the approach the present author has taken for some 30 years, in a struggle to comprehend that class of electromagnetic (EM) systems that are open systems in disequilibrium in their *virtual photon energy exchange* with the active vacuum, and exhibiting a broken symmetry in that exchange. The statement beautifully expresses that the major problems encountered in grappling with such EM systems have been the existing errors and non-sequiturs in classical electrodynamics and other parts of physics. Many difficulties have resulted from the continued propagation of a 137-year old classical electrodynamics model formed before electrons, atoms, nuclei, positrons, the active vacuum, special and general relativity, quantum mechanics, etc. were discovered or known.

1.1.1 EM Foundations Are Incomplete and Contain Errors

In any model, there are many assumptions. Even when a model is well-fitted and well-developed, it still applies only when the foundations assumptions on which it is based are true or are not too much in error. Whenever one or more of the fundamental assumptions is violated by phenomena uncovered, then there is a new class of phenomena where the model does not hold, or does not hold well and is only an approximation at best. In that case, either the existing unsatisfactory model must be improved and advanced, or a new model must be constructed.

Great scientific minds continue to point out Einstein's beautiful principle in different words. E.g., expressing the thought of Stephen Hawking, one of the great physicists of our day:

> *"All we ever know is our models, but never the reality that may or may not exist behind the models and casts its shadow upon us who are embedded inside it. We imagine and intuit, then point the finger and wait to see which suspect for truth turns and runs. Our models may get closer and closer, but we will never reach direct perception of reality's thing-in-itself.*" [As stated by George Zebrowski] {2}.

Excellent scientists — Feynman, Wheeler, Bunge, Evans, Barrett, and many others — have indeed pointed out that classical EM theory is seriously flawed. In the words of Bunge {3}:

> *"...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state."*

The author also found it imperative to return to many of the original seminal papers of physics, particularly in electrodynamics. The major concepts in those papers led to the present classical EM model. This was particularly true of the work of Poynting {4a, 4b} and Heaviside {5a-5c}, who independently and simultaneously arrived at the notion of the flow of EM energy through space[5]. Their work occurred in the 1880s, after Maxwell was already deceased. It also necessitated reviewing Lorentz symmetrical regauging of the Maxwell-Heaviside equations, where Lorentz arbitrarily discarded all permissible COP>1.0 Maxwellian systems.

The science of this "EM energy flow through space" is controversial to this day. Which is the real "EM energy flow vector" as such has never been adequately resolved, and there continue to be polite debates about it {6}. One should also be aware that physicists really do not know what many things — including energy — *are*. The definitions of these fundamentals are still uncertain, as stated rather poignantly by Feynman {7} in this quote:

---

[5] As we shall point out later, the concept of energy flowing through 3-space is a non sequitur and requires substantial revision today, to be consistent with the nature of observation and the fact that no observable continuously persists. Any observation is an instantaneous frozen 3-space "snapshot" at a single instant, gone the next instant when almost immediately replaced by another such frozen snapshot. In between observations, not mass but *masstime* exists. The same is true for 3-space, which only exists as the output of an observation process. Prior to observation, *spacetime* exists.

> *"It is important to realize that in physics today, we have no knowledge of what energy is."*

As another fundamental example, Feynman {8} also pointed out that we really do not know what *force* is[6] either! Quoting:

> "*One of the most important characteristics of force is that it has a material origin, and this is not just a definition. ... If you insist upon a precise definition of force, you will never get it!*"

1.1.2 Physics Is Not the Mathematics, But What It Manipulates

Following Feynman's ansatz, one should realize that the physics is not really in the mathematics itself, but in the physical meaning of the concepts that the mathematics manipulates. Here again, the inimitable Feynman {9} cautioned against over-attachment to the mathematics itself. He said it very clearly:

> *"Mathematicians or people who have very mathematical minds, are often led astray when "studying" physics because they lose sight of the physics. They say: "Look, these differential equations – the Maxwell equations – are all there is to electrodynamics it is admitted by the physicists that there is nothing which is not contained in the equations. The equations are complicated, but after all they are only mathematical equations and if I understand them mathematically inside out, I will*

---

[6] The problem is the ubiquitous and erroneous use of a "separate" force acting on a "separate" mass. Actually mass is a *component* of force, as can be seen from F ∑d/dt(mv). Hence there is no such thing as a separate force in empty space, acting upon a mass. Instead, the massless 4-field in space acts upon mass to produce force *in that interaction*. Force is not a cause, but an effect of an ongoing interaction — as is any observable. An observable is a continuing series of frozen 3-space LLL snapshot entities given by the result of continual application of ∈/∈t (LLLT) ♥ LLL by photon emission. The formation of the LLLT (spacetime and masstime) in between observed m as LLL, is produced by photon absorption of the previous m = ∈/∈t (LLLT).

> *understand the physics inside out." Only it doesn't work that way. Mathematicians who study physics with that point of view – and there have been many of them – usually make little contribution to physics and, in fact, little to mathematics. They fail because the actual physical situations in the real world are so complicated that it is necessary to have a much broader understanding of the equations."*

Such matters and similar thoughts had bothered me even back in the early 1950s, when I could not find a single professor or dictionary of physics that *logically* defined a field or a potential. Considered rigorously, the definitions all fell apart and violated elementary logic. Improper, insufficient, or just plain *wrong* definitions in physics have continued and are still widespread to this day. To give a single example from an excellent book by Kraus {10}: On p. 60, Kraus gives the formula for a potential referred to as the "absolute potential" of a charge source. Quoting:

> *"This potential... is, by definition, the work per coulomb required to bring a positive test charge from infinity to the point r_1."*

Kraus erroneously "defines" the scalar potential identically as work, which is equating the cause with the effect. The work that dissipating a potential does or can do, is not the potential itself! Else "human" means nothing but how well one can chew one's food or drive an automobile. Actually, Kraus gives one theoretical way to measure or calculate the effect of the potential's *local intensity at a point.*[7] Note that what is measured is the energy *diverted from* the potential at that point, around that test charge. This does not specify the *entity* (the potential itself) at all, but only what has been diverted from it. It is rather like confusing the whirlpool (water diverged from the normal river flow) in a river as being the river itself.

---

[7] To show the non sequitur, a million more positive test point charges could be brought from infinity to that same point, and the potential — without any change in it — would cause the same amount of work to be done upon each of those charges. As can be seen, even noted professors can speak rather casually when they "define" the potential's point intensity as "the" potential (which extends over all space, not just at that one point). There may be different intensities at different points in the potential, but the potential itself remains one-and-the-same thing regardless of which one or all of the point intensities are discussed. And neither the point intensity of the potential nor the potential is the work that is done upon a moving charge by that potential or its intensity.

Integrating what has been diverted from it does not yield the potential itself! At best, the measurement gives an indication of the *intensity* of the potential at a point, insofar as its reaction with charge is concerned. More rigorously, what is being utilized is the potential's reaction cross section presented to a *unit point static charge* at that point. The same unit point charge, if placed in particle resonance, will sweep out more geometrical area and exhibit a greater reaction cross section. That will increase the energy collected divergence of the energy flow that is being moved around the particle itself. By normal calculations, the resonant charge may collect some 18 times as much energy as is possible at the same point by the same charge in static mode — e.g., as shown by Bohren {24, 25}.

A definition must present an identity. Examining the proposed definition as "the potential is identically … work" one sees the problem immediately. *The potential exists whether or not there is a positive test charge, or whether one moves such a charge in from infinity or not, and whether there is any work done or not.* None of that is what a potential *identically is*, but only one aspect of what it *does* or *can do* or *can cause.* One leaves as an exercise for the reader the task of further examining dictionaries of physics and textbooks, to try to find a satisfactory definition of that common scalar potential.[8] We do not believe the reader will find it.

Yet any good textbook will also contain some real gems of great insight, simply said. As an example from Serway {11}, we eventually took a most marvelous cue, of how to get around classical thermodynamics' prohibition against heat energy "running uphill" from hot to cold. In an insightful statement, Serway said {11}:

> *"The second law [of thermodynamics] does not rule out the possibility of pushing heat uphill, as it were, from a cold object to a hot one, or of creating order out of*

---

[8] The scalar potential identically is actually a harmonic set of phase conjugate longitudinal EM wavepairs, as shown by E. T. Whittaker in 1903 {85}. Even Whittaker, however, misinterpreted his wavepairs only after interaction with that ubiquitous unit point charge assumed at every point in space. He gave two *effects* of that interaction, not the *cause* (which exists prior to interaction) *and the effect* (which exists only after interaction). Reinterpreting to get at the causal wave, each wavepair is a matched set of two waves; an incoming EM longitudinal wave in the time domain prior to interaction with a charge, and — after the time-energy wave is absorbed by the charge, an emitted outgoing EM longitudinal EM wave in 3-space. The so-called "static" potential is not static at all, but is a dynamic, ongoing 4-space process. This follows the re-interpretation by the present author {12} of Whittaker's decomposition — a reinterpretation then found to be consistent with quantum field theory {19} and with broken symmetry of opposite charges in particle physics {73}.

> *disorder. It merely states that such a reversal of the natural flow requires an influx of energy…"*

This essentially states the *law of entropy*, where to reverse entropy (disorder), one must apply ordering (energy).[9] We also note that the original concept of entropy was as *dissipation of potential*. There are of course different ways to apply the energy, and it need not be by the operator himself. Chapters 4, 5, 6, 7, and 8 of the present book give some unusual ways. A special paper {12} published by the author in 2000 gives another. Indeed, when the flow of time is predominantly reversed, so is the "entropy" of a situation, since the "videotape is running backward", so to speak. In that case, the entire classical thermodynamics must be extended to essentially include its own opposite. The making of a small time-reversal zone where such things happen is as simple as involving a predominance of antiphoton interactions with the charged particles in that zone, rather than a predominance of photon interactions. What we are saying is that the notion of irreversibility in thermodynamics is not necessarily absolute. It usually assumes a “time forward” situation, and may not hold in a “time-reversed” situation. In our chapter on cold fusion, we will present some specific and quite startling nuclear interactions that occur as a result of the time reversal of the coulomb barrier (repulsion of like charges in a forward time situation) into a coulomb attraction between like charges in a time-reversal zone and time-reversed situation.

---

[9] Now notice what Serway’s statement means with respect to the classical thermodynamics “definition” of closed system. Thermodynamics defines a *closed system* as one in which mass is not exchanged across its boundary, but energy can be. That definition permits a closed system to receive excess energy from its active environment, and thereby reduce its entropy. A specific example is simply potentializing a circuit, prior to movement of the current. Hence the entropy of a closed system does not necessarily increase, but may decrease or increase if the system is in disequilibrium (difference in energy received from its active environment and energy escaping back to the active environment from the system. For the same system, if energy exchanges across its boundary equally in both directions, the system is said to be in “equilibrium” with respect to external energy exchange. One must be very careful in interpretation of the second law of thermodynamics! One has a very different “closed thermodynamic system” when it is in *energy exchange disequilibrium*, than when it is in *energy exchange equilibrium*. Equilibrium is the condition of maximum entropy. For substantial disequilibrium condition, entropy cannot be computed, but is less than the entropy of the same system in equilibrium. The entire ansatz of thermodynamics may be violated once time-energy is transduced into internal EM energy inside the system. That is a fundamental disequilibrium, performed by every charge in the universe. Hence of necessity we have advanced and utilized new definitions of “open system” and “closed system” in our approach in this book, as discussed in the Appendix.

If Serway's statement were reversed, it would then be a statement of the *law of negentropy*. It would state that self-ordering (i.e., freely receiving energy from the active environment) in a system could indeed "push heat uphill from a cold object to a hot one", and the system could simultaneously emit energy in the process. Every charge does it!

We did find it necessary to correct the classical thermodynamic definitions of "open system" and "closed system". To define a closed system as closed only to mass transfer, but open to energy transfer, is a gross non sequitur. Since general relativity was published in 1915, energy and mass are known to be the same thing (*mass is just a special form of energy*), hence the term "mass-energy" (mass as energy) in physics. Whenever energy crosses the boundary of a system, the system's mass changes and mass (or certainly mass change) has also crossed that boundary. Indeed, as we shall point out, in 1917 Hilbert specifically pointed out that in general relativity there can be no energy conservation equations of the kind usually employed elsewhere. The fact that general relativity falsifies much of the present foundations of classical thermodynamics seems to have been either ignored or missed by most of the scientific community, although we will quote leading Russian scientists who have noticed it and are aware of it.

The reader is thus warned that, henceforth, when we use "open system" we mean one where either energy or mass or both exchanges across the system boundary. When we use "closed system" we mean one in which *neither* energy nor mass exchanges across the boundary. In short, the notion of a "closed system" has been redefined into what classical thermodynamics calls an "isolated system". We already know from particle physics and the active vacuum (and from general relativity and the change of spacetime curvature with every change of spatial energy or mass-energy), and from the giant negentropy involving time energy transduced into 3-space energy and vice versa, that there is no such thing in all the universe as a truly closed system.

In Appendix A, we have discussed how extension and change to classical thermodynamics must be made. We do this by extending the first law, refuting any absoluteness of the second law and third law, dealing with the zeroth law in a new way, etc. We also urge the better theoreticians to re-examine classical thermodynamics along such lines, to modernize and upgrade it. We believe that the present scientific work to extend thermodynamics so it fits those situations now known to violate it should include additional considerations such as we present in Appendix A.

In short, with the new definitions Serway clearly states the difference between classical equilibrium thermodynamics, where no excess energy from the environment is received, and the thermodynamics of systems far from equilibrium with their active environment (using the new definition of open system for clarity), in which case excess net energy from the environment can be received and used in electromagnetic systems, providing COP>1.0 systems or even COP = ←systems. The windmill, sailboat, and waterwheel are age-old examples of disequilibrium systems where of course the energy of mass in motion is what is transduced. It is our objective in this book to point out the use of electromagnetic systems in energetic disequilibrium with their active environment (the active vacuum and curved spacetime) to provide just such negentropy.

1.1.3 Time Is Energy and Must Be Considered As Such

In a physics model, one's choice of fundamental units is arbitrary.[10] As an example, in one type of physics only a single fundamental unit — length — is employed. All other entities then become functions of length.

We are also free to choose the *joule* as the single fundamental unit in our physics model. The result that mass is a function of energy is now familiar and quite accepted, by the famous formula $E = mc^2$. Indeed, as can be seen, in that equation mass and energy are one and the same thing, since c^2 is a dimensionless constant. However, in our new model time becomes a function of energy similarly, and thus time is identically energy.

Let us perform a thought experiment. Suppose we take some spatial EM energy in 3-space, and compress it by the factor c^2. What can we do with it? If we leave it there in 3-space, it is known as "mass". If we place it on the fourth Minkowski axis *ict*, it is known as time because *t* is the only variable on that axis, and the only "place we can set it".

So to first order, *time has the same energy density as mass*. Multiplying an amount of time *t* (in seconds) by c^2 gives the decompressed spatial energy *E* that the time *t* will transduce into. In short, $E = tc^2$ also.

Now we notice what special relativity has to say about the relationship between time-energy and mass-energy. When the mass-energy increases

---

[10] E.g., see Jackson, *Classical Electrodynamics*, 2nd Edition, Wiley, 1975, p. 811-812; *ibid.*, 3rd Edition, 1998, p. 775. Jackson wryly remarks on how much excess heat and passion have been unnecessarily expended on that subject.

(e.g., as a function of velocity), time "dilates" or “decreases”. In other words (hint), there is a sort of special new extension to the conservation of energy law: If mass (3-space) gains some 3-spatial energy, then time loses some time-energy. Since time is spatial EM energy compacted by c^2, then *the relativistic energy changes in the time domain are enormously greater than the corresponding relativistic changes in spatial energy in the photon*. To the external 3-space observer, this is the injection of negentropy with respect to the present 3-spatial conservation of energy law. In the reverse case (as when the object reduces its velocity), if mass (3-space) loses some energy, then time gains some time because the time dilation reduces — i.e., time contracts or densifies (less time has more energy). To the observer, this is the production of entropy with regard to 3-space energy conservation. Yet the observer misses the fact that, accompanying entropy in 3-space is negentropy in time energy on the 4th Minkowski axis, and vice versa.

We believe this to be a rather dramatic extension to the previous concept of entropy in thermodynamics. We get 3-space entropy in physical processes only by gaining 4th dimensional negentropy accordingly. One can thus appreciate the impact upon the entire subject of thermodynamics, once time-energy is accounted and the new interplay of entropy and negentropy are accounted. We believe that this mechanism may involve the fundamental mechanism for both entropy and negentropy. Neither is produced alone; both are always produced in tandem. This of course is not what we “observe” since all observation is 3-spatial. Hence to observe the system and see 3-space entropy or negentropy is to hide the simultaneous 4th dimensional negentropy or entropy that unobservably accompanied the observable that we did see.

Again we call attention to the original meaning of entropy: the dissipation of potential (i.e., potential energy).

Once we understand that *time* is a special form of energy (we discuss this shortly), we may input the energy required to "move heat energy backwards" — i.e., we may directly engineer negentropy — by transducing some time-energy into 3-spatial energy. We can do it easily by time reversal, and every charge in the universe does it already. On the other hand, we can do it by breaking the symmetry of time-energy flow — which is as simple as forming a little dipole. The broken symmetry of unlike charges — and therefore the dipole — was discovered by Lee and Yang {13a-13c}, who strongly predicted it in 1956. Wu *et al.* {14} confirmed it experimentally in early 1957. This was such a revolutionary change to physics that Lee and Yang were awarded the Nobel Prize in the same year, 1957 — an almost unprecedented action.

Its broken symmetry tells us that the dipole's unlike charges continuously absorb virtual photon energy from the active vacuum, transduce it into real observable energy, and pour out real, observable EM energy in all directions in 3-space. That puts an entirely different perspective on *what really powers every dipolar EM circuit; i.e., what produces the flow of energy pouring from the terminals of every generator and battery, filling all space around the external circuit and its conductors*. The EM energy pouring from the generator or the battery is not due to the generator outputting some of its own available internal energy (from the shaft energy input to the generator, transduced into magnetic field energy inside the generator, and then dissipated to separate the charges and form the source dipole between the generator terminals) or the battery transducing and outputting some of its available chemical energy (which is only dissipated inside the battery to separate the internal charges and form the source dipole between the battery terminals). We will explain that later, and explain how any dipole or charge simply pours out real, observable energy continuously in all directions, without any *observable* input of energy. The input energy is there, but it is in *unobservable* (virtual) form.

In 1971 while a graduate student at the Georgia Institute of Technology, I realized the mechanism that generates the "passage of time" insofar as the observer is concerned {15}. With a little more work, this gave the clue in the 1990s for the mechanism generating little momentary *time-reversal zones* (TRZs) {16} in the electrolyte utilized in cold fusion experiments. Hence we proposed the use of little time-reversal zones forming momentarily in the electrolyte after loading of the palladium lattice of the electrodes with hydrogen or deuterium. Such time-reversal zones can form in the region of excessive positive charge accumulation, since positive charge can be said to receive negative EM energy from the time domain and output negative EM energy in 3-space.

This led to uncovering an entire class of new nuclear reactions — fusion reactions at low spatial energy but high temporal energy — in these little temporary time-reversal zones (TRZs), in which like charges attract and unlike charges repel {17}, followed by rapid decay of the TRZs where the normal law of attraction and repulsion is restored again.

In theory, a fermion cannot be time-reversed because the Pauli exclusion principle prohibits it. However, a boson can be time-reversed, so fermions can be time-reversed in pairs, where each pair acts as a *quasi*-boson. The

TRZ completely overcomes and *reverses* the "coulomb barrier" between a pair of like fermion charges.[11]

As an example, two deuterium ions in a little TRZ would momentarily form a quasi-nucleus of helium, since the two positive deuterons are attracted so closely that each enters the strong force region of the other. At the same time, the strong force is weakened by the partial time reversal of the gluon forces, so that the quarks are nearly freed. As the other ions in the surrounding solution then move to destroy the momentary time-reversal zone, their movement forces the TRZ to decay back to a time-forward zone — with the decay action starting from every point in spacetime inside the nucleons themselves[12]. The TRZ decay-induced action thus strikes the *nearly freed* quarks first, and decay can occur by easy quark flipping while the gluon forces are still much reduced and not back to normal strength. Hence as the gluon forces return toward normal, the quasi-nucleus decays by the strong force increasing and overpowering and drawing the quasi-nucleus into a full-blown helium ion — i.e., an alpha particle —flipping the partially-freed quarks as necessary to do so (that is not necessary in equation [1] below.

During decay of the TRZ, the weakened strong force grows much more rapidly than the Coulomb force zeroes and then increases. Consequently, the quasi-nucleus of two D+ ions merely draws together due to the rapidly increasing strong force, forming an alpha particle without quark flipping. Four H+ ions — four protons — in a quasi-nucleus in a TRZ will undergo quark flipping twice when the TRZ decays, thereby resulting in an alpha particle.). So that explained the anomalous formation of the alpha particles in the experiments. The interaction for two ions of deuterium is given by:

$$ {}^{2}H_{1}^{+} + {}^{2}H_{1}^{+} \Downarrow {}^{4}He_{2}^{++} \qquad [1] $$

---

[11] In a TRZ, the law of attraction and repulsion of charged particles is reversed.

[12] Any moment in time exists everywhere simultaneously throughout the universe. In short, time is a multiple connection in 3-space. Hence in any time-reversal zone (a TRZ) where time is reversed, then starts fading away and back to a time-forward zone (TRZ), the changes induced by the "fading back to TFZ" *simultaneously* involve every point in the 3-space of that TRZ that is changing (decaying). An easy change is quark flipping, since the quarks are almost freed in the TRZ to begin with. The reason the reaction proceeds in that direction is that the recovery of the strong force is much faster than the restoring of normal electrical repulsion, hence the quasi-nucleus is drawn further together into a full nucleus, constituting a legitimate nuclear transmutation at low spatial energy but high time-energy.

This interaction between two deuterons in solution does not occur in a normal time-forward zone because of the coulomb barrier preventing the two deuterons from entering each other's strong force region and acting as a quasi-nucleus of helium, so particle physicists have missed it. It does exist in a time-reversal zone; such zones form and then decay back into normal time-forward zones. This is possible because the coulomb barrier is momentarily reversed. The use of such TRZs in particle physics opens up thousands of new nuclear reactions, all at low *3-spatial* energy, but involving very high *time*-energy. Because of the extreme energy density of time, *these time-energy-induced reactions are actually much higher-energy reactions than high-energy physicists presently consider and utilize!* Indeed, it opens up a new kind of far more energetic "high energy physics".

Cold fusion experimenters have unwittingly opened a window upon a vast new particle physics, previously overlooked by our scientists because they have disregarded the use of time-energy, time reversal zones, and transmutation of time-energy into 3-spatial energy in their nuclear reactions. However, a few physical theorists attempting to better explain particle physics have recognized the importance of the time domain, and probing work in that respect is occasionally done {18}. In our view, it is not accidental that more than 600 successful cold fusion experiments have been obtained, by a variety of researchers in many labs in several nations. We strongly suggest that much of the conventional physics community has firmly placed its collective head in the sand, and is refusing to grapple with the startling new time-energy physics that is being initiated by cold fusion research.

In 1999 (published in 2000) {12} we finally discovered a great new symmetry in EM energy flow, whereby time-energy flow symmetry and 3-space energy flow symmetry are each individually broken, while an overlooked and more fundamental 4-symmetry energy flow — between the time-domain and 3-space — is sustained. *The result is that all EM energy in 3-space comes from the time domain locally and returns to the time domain locally, in a giant negentropic circulation.*[13] Together by the discovery of relevant quantum field theory work by Mandl and Shaw {19}, this now lends strong support to the use of time-energy in physics as a practical matter for strenuous investigation.

---

[13] This is also understandable from the implications of the observation process, which yields a frozen 3-space snapshot existing only at a single instant. Thus any 3-space energy existing there in that frozen instant, had to just come from 4-space (from the time-domain via the giant negentropy process, if we take the view that "the past exists only in time itself".). For the frozen snapshot to "change", time must be added to it, which converts it to a 4-space process again. So the "3-spatial" observed

As an example, Mandl and Shaw {19} treat the four polarizations of the photon. Neither the longitudinal nor the scalar photon is directly observable, but in the presence of charge the two are observable in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. This argument, translated from particle terminology to wave terminology, directly fits our re-interpretation {12} of Whittaker's 1903 decomposition of the scalar potential {85}. For the combining mechanism of the fields of the photons, we must account for *the field as a ubiquitously assumed interaction with the detecting/observing unit point charge*. Thus we must account for the absorption of the incoming time-polarized wave or photon, the transduction of that excitation energy of the charge into longitudinal EM wave/photon energy, and subsequent emission of that excitation energy in 3-space. That is what happens for a negative charge. For the positive charge, the process is time-reversed, hence occurs in opposite fashion. Or as an alternative, the positive charge can be said to continuously receive negative time-energy from the time domain and emit negative spatial EM energy.[14]

There is an *energy polarization transduction function* of charge, whereby it transforms received time-polarized photon energy into emitted longitudinal photon energy in 3-space (for the negative charge, and vice versa, for a time-reversed positive charge). This transduction appears to have been overlooked in physics prior to our recognition of it. It can in fact be used to generate an acceptable definition of charge itself. Charge is the

---

energy must come from an immediately previous 4-space process, and must return to an immediately following 4-space process.

[14] As we will find in our chapter on antigravity, there is good reason to treat the positive charge as a source of negative energy and negative energy fields. However, this is pertinent only prior to observation of the charge, while it is still a 4-spatial unobserved negative energy electron entity. For the observable charge, one has already conjugated — after all, the positron is observed as if it were an electron going backwards in time, which we observe as an electron with its charge reversed and with parity reversal (of its spatial direction). For the observable charge, we have already reversed the negative energy fields into positive energy fields by simply reversing their direction and the time associated with the photon (quantum of the EM field).

continuously active entity which performs that ongoing process or those ongoing processes of energy transduction between the time domain and 3-space (between the causal unobserved 4-space process and the observed 3space snapshot). Or in other words, it is an active process connecting 4-space cause and 3-space effect, and connecting the unobserved (such as virtual) to the observed (such as mass).

So for a dipole, the "causal" time-polarized EM wave or photon as a 4-space entity comes to the dipole[15] (3-spatial as observed) and is absorbed by the detecting negative charge or dipole, then is re-emitted as the longitudinally polarized EM wave or photon in 3-space. *That absorption and remission is what charge does, since it is an entity for performing that process*. The emitted energy in turn is absorbed by the nearby positive charge, retransduced into time-energy, and re-emitted back to the time domain. This ongoing very special 4-space energy circulation (even with a virtual charge in the vacuum) is what a scalar potential *identically is and is doing*, at every spatial point of itself, inducing vacuum polarization and “point dipoles” in the virtual state in the process. Recognition of these missing functions allowed at last a solution to the long-vexing problem of the source charge and its associated fields and their energy, often called the most difficult problem in both quantum and classical electrodynamics {68}. We discuss that solution later.

1.1.4 The Search for COP>1.0 Circuits and Systems

A very long search and much intense study and reflection eventually revealed the concepts and principles of those long-neglected disequilibrium Maxwellian systems that permissibly output more energy than the operator inputs. The active environment — not the operator — simply inputs the rest of the energy. Such disequilibrium systems are indeed permitted in Maxwell's theory {20}, and are also still prescribed by Heaviside's severe curtailment of it {21} into what is two vector equations with variables unseparated, rather than Maxwell's 20 equations in 20 variables.

---

[15] Prior to interaction of the incoming time-energy with the observable charge or dipole, it interacts with the virtual particles of the vacuum, generating vacuum polarization. Neither the virtual particles nor the time-energy are observable; only the effects of their conglomerate interactions with observable charges are observed. Hence one can take the particle view that virtual particle energy is continually absorbed, or one can also take a quantum field theory view that time-energy is continually absorbed. The two are always present in combination.

When Lorentz[16] symmetrically regauged the Maxwell-Heaviside equations, he arbitrarily discarded the entire class of Maxwellian systems that are far from equilibrium in their exchange with their active (vacuum) environment. Lorentz revised (symmetrically regauged) the Maxwell-Heaviside equations to make them amenable to separation of variables and closed analytical solutions, thus reducing the onerous chore of numerical methods. This Lorentz symmetrical regauging is given in most EM textbooks {22}, and we show it in Chapter 2. The symmetrically regauged Lorentz equations are not Maxwell's equations, nor are they the truncation of Maxwell's theory by Heaviside *et al*. Considering an active environment, under our altered thermodynamics definitions Lorentz implicitly selected and retained only the equilibrium class of Maxwellian systems, while *arbitrarily discarding* the entire disequilibrium class. He thus discarded all those Maxwellian systems permitted to produce COP lying in the range $1.0 < \mathrm{COP}\ \Omega\leftarrow$.

Maxwell's electrodynamics is a *material fluid flow theory* and it assumes a material ether. Anything that fluid systems can do, electrodynamics systems can do, at least in theory, because their mathematical models are the same form. So when one cites known examples of fluid-driven physical systems where the energy to run the system is freely furnished by the active environment, analogous electrodynamic systems in active environments — and in disequilibrium exchange with that environment — must also exist in nature. Indeed, particle physics requires it and proves it. These are the very systems arbitrarily discarded by Lorentz symmetrical regauging in every university.

So there exists a direct analogy between fluid systems and classical electromagnetic systems. The common windmill, waterwheel, and sailboat demonstrate by analogy that open EM systems far from equilibrium — powered by free EM "winds" and "energy flows" in the active vacuum environment — also exist in consonance with natural law. They are no more mysterious than a solar cell power system, which is after all a recognized "free energy" or "overunity" system, as is the humble charge (thought to be the source of all EM energy, fields, and potentials). In physics, the powering of systems by receipt and use of energy from their active environment leads directly to the thermodynamics of systems far from thermodynamic equilibrium in their energetic exchange with that active environment.

---

[16] Actually first accomplished by Ludwig Valentin Lorenz in 1867, then by H. A. Lorentz much later. Lorentz was given the credit erroneously. Lorenz actually derived electromagnetic theory independently in his paper. See J. D. Jackson and L. B. Okun, "Historical roots of gauge invariance," *Rev. Mod. Phys*, Vol. 73, July 2001, p. 663-680.

It follows that the seeming absence of such *electrodynamic* systems arises not because they are impossible but because present-day circuits and systems are ubiquitously designed to self-enforce an inherent energy equilibrium with their active vacuum environment. The closed-current loop circuit turned out to be the Lorentz self-regauging demon involved in destroying the COP>1.0 capabilities of every EM circuit. So little by little, we unraveled the long tedious trail of Maxwell's electrodynamics and what had happened to those missing Maxwellian-Heaviside systems far from equilibrium with the active vacuum.

We learned how, where, and by whom those permissible overunity Maxwellian systems were discarded. That is, we found what happened to all those Maxwellian *open disequilibrium systems* — originally included in Maxwell's and Heaviside's theories — where such a system receiving and using excess energy from its active environment[17] is permitted by the laws of physics, electrodynamics, and thermodynamics to:

---

[17] For the discerning reader, of necessity we have revised the foundations of the ancient classical thermodynamics, as further discussed in Appendix A. We refer the reader to Bimalendu N. Roy, *Fundamentals of Classical and Statistical Thermodynamics*, Wiley, New York, 2002, and to any good book on the history of thermodynamics. Also particularly good is Kondepudi and Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures*, Wiley, Chichester, 1998, reprinted with corrections in 1999. The foundations of classical thermodynamics (and its fundamental definitions) were basically frozen prior to the advent of Maxwell's 1865 seminal theory, and well before the 1880s discovery (modeling) of the flow of EM force field energy through space. Some of the fundamental definitions of thermodynamics now will not withstand critical review in terms of "meshing" with the Heaviside/Poynting material fluid energy flow theory. Neither will they withstand the new concepts of energy such as mass-energy, time-energy and transduction between time-energy and spatial energy by every charge. E.g., thermodynamics defines an *open system* as one that exchanges energy and mass with its surroundings. Yet it defines a *closed system* as one closed only to mass exchange, not to energy exchange. *If the energy exchange is analogous to material energy flow and changes the mass of the system, then that definition of closed system is a non sequitur.* From general relativity, mass is simply energy anyway, and so "mass" exchanging across the boundary of the system is actually energy exchanging, and vice versa. Since Maxwell's theory is a *material fluid* theory, the Poynting and Heaviside energy flow models are *material fluid flow energy* models by analogy. The specialized thermodynamics definition of *closed system* rigorously will not logically allow the exchange of "material fluid energy flow" (or energy as matter)

since it prohibits matter flow, but the specialized definition of *open system* would and does. In short, with respect to material fluid energy flow, the concept of the "closed system" in thermodynamics has forced itself to become the *isolated system* instead, which is unacceptable since we must model EM energy flow exchange between the environment and the system. So we must change the thermodynamic definitions of *open system* and *closed system*. Else there cannot be any EM energy-mass or mass-energy flow between environment and system, which totally violates what is well known to be happening in all EM systems. In modern physics, every charge and every dipole already have such energy exchange with the active vacuum environment, and it is never zero; instead, it is of enormously high magnitude. Without that exchange, as we advance in this book, there cannot even be a "source charge" or associated EM fields and potentials and their energy, reaching across space. In other words, without it we can have no electrodynamics at all. So we have uncovered a fundamental and major contradiction between classical thermodynamics and general relativity, as well as between thermodynamics and the "material EM fluid energy flow" model used in electrodynamics. What classical thermodynamics calls a "closed system" permitting energy flow exchange, we must now logically regard as an *open* system because it is open to energy exchange across the boundary! The previous notion of the *isolated* system — with no exchange of either energy or mass — is what we must treat as a truly *closed system*. And there is none such in the universe, as we know in particle physics (e.g., because of the discovery of broken symmetry in 1957 and because of the well known active vacuum exchange with every EM charge and dipole). So we have corrected these direct contradictions between "EM as a material fluid energy flow theory" and the old classical thermodynamics.

As the reader will later see, this strongly affects our new definitions of *efficiency* and *coefficient of performance*. The new definitions we advance are rigorous, and they also apply to COP>1.0 EM systems, and even to self-powering (COP = ←) EM systems. They also hold for very novel new energy processes such as quantum potential energy in a multiply connected space, multiple retroreflections and re-use of the same energy, conversion between time energy and spatial energy, and 4-space giant negentropy energy flow circulation.

But one can no longer be allowed to equate *efficiency* with *coefficient of performance*. Now they are never the same thing, just as a six-foot tall man and a six-foot tall doorway are never the same thing, even though they have the same height magnitude. We warn the reader that a great deal of thought and study must be put into appreciating these suddenly encountered changes to the quite old classical thermodynamics definitions. The changes are absolutely necessary. Bluntly put, in light of much more modern knowledge, a rigorous foundations analysis reveals the classical thermodynamics as well as disequilibrium thermodynamics to contain logical contradictions, such as its direct contradiction with general relativity and the EM material fluid energy flow theory. Either we give up or dramatically change the present EM energy flow theory, or we must make the necessary foundations changes to thermodynamics. We have chosen the latter option in this book, and the reader is forewarned of that dramatic change. The specific changes and rationale are discussed in Appendix A.

(1) self-order,
(2) self-oscillate or self-rotate,
(3) output more energy than the operator inputs (the excess being freely received from the active environment),
(4) power itself and its load simultaneously (all the energy being freely received from the active environment), and
(5) exhibit negentropy.

We vigorously pursued those long-lost Maxwellian systems, and we eventually found them. We also found real experiments {23, 24, 25, 26, 27, 28} and real devices {29a, 30, 31, 32, 33, 34, 35a, 36a, 36b} that performed one or more of those fabulous five functions, though there was often no realization by the experimenters, inventors, and scientists of the actual mechanism involved. Eventually my colleagues and I were also able to produce a successful experimental device {37, 38a-38c, 39}, the motionless electromagnetic generator (MEG), which outputs more energy than we input to it. We cover the MEG in Chapter 7. Presently a cooperative research program is ongoing with the National Materials Science Laboratory of the National Academy of Sciences of a friendly foreign nation, to develop and market commercial power systems based on successful laboratory experiments with the MEG. At this writing, we are also strongly seeking the extensive funding required to set up a physics lab and complete the final research allowing production of power systems.

1.1.5 Additional Very Important Implications

We also formally proposed {40} that the vast nondiverged EM energy flow component — Heaviside's "dark" nondiverged energy flow component, accompanying every reaction of a charge with a field or a potential, but arbitrarily discarded by Lorentz and modern classical electrodynamicists — is the generatrix for the extra gravity holding the arms of the spiral galaxies together, after all the dark matter is accounted.[18]

---

[18] Heaviside himself recognized the gravitational implications of his extra component of energy flow, which is in closed circular loops. Beneath the floorboards of his little garret apartment, years after his death, handwritten papers were found where Heaviside used this component for a unified EM approach to gravitation. See E. R. Laithwaite, "Oliver Heaviside – establishment shaker," *Electrical Review*, 211(16), Nov. 12, 1982, p. 44-45. Laithwaite felt that Heaviside's postulation that a flux of gravitational energy combines with the (**E**Δ**H**) electromagnetic energy flux, could shake the foundations of physics. Quoting from Laithwaite: *"Heaviside had*

In addition, my close colleague Bedini and I have filed a patent application upon a very special process to "freeze-frame and lock-in" a disequilibrium Maxwellian system in its otherwise far-from-equilibrium state, so the apparent disequilibrium operation of the system can be maintained stably as a new equilibrium condition. This appears to be a method to produce and utilize what Kondepudi and Prigogine refer to as a *nonequilibrium stationary state*. This stabilization is necessary in order for such a system to maintain its COP>1.0 excitation and steadily output more energy than the operator inputs, or to steadily power itself and its load {41}.

We also found that COP>>1.0 EM systems (and some COP>1.0 systems) produce (as a function of the COP) a current of Dirac sea holes (positrons) *in the local vacuum environment itself*, from the output section back to the input section. At COP<1.0 and COP not too greatly above 1.0, a Dirac sea hole in the vacuum almost immediately interacts with an orbital electron in the material lattice of the system. This converts the negative energy, negative mass "vacuum hole or state" into a lattice hole, which is attached to the large positive mass of the ion left by the disappearance of the electron. This "lattice positron" type of problem has been known in semiconductor design for some decades. Normal EM circuits do not usually meet the phenomenon overtly because the semiconductor designers controlled it in the semiconductors themselves by use of appropriate donors and acceptors.

There is a great difference between the actions of Dirac sea holes in the vacuum prior to observation, and lattice holes in materials (after observation). So there is a great difference between the action of a "positron" on spacetime before its interaction with mass and observation, and its action on spacetime after it interacts with mass and is observed.

For COP>>1.0, significant phenomenology and novel effects occur because a substantial fraction of the Dirac sea holes (unobserved positrons) sweeping from output to input do not convert to lattice holes (observed and bound positrons) along the way. For substantial values of COP above 1.0, Bedini's invention (patent application filed by Bedini and the present

---

originally written the energy flow as $\mathbf{S} = (\mathbf{E}\Delta\mathbf{H}) + \mathbf{G}$*, where* $\mathbf{G}$ *is a circuital flux. Poynting had only written* $\mathbf{S} = (\mathbf{E}\Delta\mathbf{H})$*. Taking p to be the density of matter and* $\mathbf{e}$ *the intensity of a gravitational force, Heaviside found that the circuital flux G can be expressed as* $p\mathbf{u}$ *4 c*$\mathbf{e}$*, where* $\mathbf{u}$ *represents the velocity of p and c is a constant."*

author)[19] covers the master process for intercepting and transducing this appreciable flow of negative energy from the output section back to the input. Otherwise, that flow — if not intercepted and not converted — will "eat" extra input electrons from the power supply, thus acting as a novel "extra load" appearing in the input section. That extra load then draws additional current and power from the external power supply by electron-hole annihilation[20].

With the Bedini invention, the negative energy (unobserved positron) flow appearing at the input section is transduced into a flow of positive energy (a flow of electron current) from the input section back into the system. The process deliberately uses the "interaction and observation" process to phase conjugate the charge and reverse its direction of flow! In that case, the otherwise detrimental *negative* energy output back through the COP>1.0 system (which is nature's decay process for COP>1.0 interactions) is changed to a beneficial *positive* energy input within the system itself, freely received from the vacuum environment. This process is then used to close-loop the system for self-powering in a "locked" and stabilized disequilibrium condition — a nonequilibrium stationary state.

If we consider mass to be a special kind of positive energy state, then positive energy states represent curvatures of spacetime that are positive gravity. Negative energy states generate antigravity (the time-reversal of gravity).

---

[19] Bedini personally discovered and implemented the solution before the exact nature of either the problem or the solution was recognized! My contribution was to recognize the nature of the problem and the mechanism used in the solution. We also stress that, contrary to conventional treatment, all EM circuits do involve not only lattice holes in the conductors and components, but also Dirac sea holes in the local vacuum. That the Dirac sea is involved whenever there are EM fields is clearly shown by Felix Finster, "Definition of the Dirac Sea in the Presence of External Fields," *Adv. Theor. Math. Physics*, Vol. 2, 1998, p. 963-985.

[20] With a proper change in the curvature of local spacetime, pair annihilation can occur with no accompanying photon radiation. The condition is that the part of the curvature of spacetime representing the energy change of the otherwise emitted radiation, does not "relax" even though it is an "excited state". Rigorously, the emission of the radiation from pair annihilation occurs in two steps: (i) first the local spacetime is curved for and by the energy excitation, as a *static change* of the curvature, and (ii) then that curvature relaxes back to its former value, propagating that specific curvature in space, which is recognized as the radiation propagating in space. If the spacetime curvature excitation does not relax, there is no photon emission and propagation.

Before their interaction and observation, the Dirac sea hole (positron) currents — produced in natural COP>1.0 processes in the universe — are still negative energy electrons in 4-space. They are not 3-positrons until interaction with matter has occurred. These "negative energy electrons" generate negative energy EM fields, including both the Poynting energy flow component and the Heaviside nondiverged energy flow component. These negative energy EM fields appear to be *generating the antigravity* whose effects are seen in the distant parts of the universe. They produce far more powerful effects than the accounted electrogravitation effects in astrophysics, which only uses the gravitational effects of the very much smaller Poynting energy flow component. Also, astrophysicists do not use the "positron before observation," when it is still a negative energy electron with negative mass. The accumulation of the negative energy (Dirac holes) in space (in the surrounding vacuum) is connected with massive objects and violent processes, where very strong gradients are produced.[21] As a result, very large negative energy fields and potentials are produced. In turn, this results in cumulating antigravity. This cumulating and interacting antigravity appears to be the mechanism for the forces *accelerating* the expansion of the universe — rather than it decelerating as would be the case if the net gravity there were positive. In Chapter 8 we propose this explanation for the observed acceleration of the expanding universe — and the basis for the explanation can be and has been successfully demonstrated in a legitimate overunity EM circuit or system {42}. We offer this in honor of Heaviside, who first discovered the gravitational aspect of his huge nondiverged energy flow, but did not live to publish it. He also did not consider the Dirac sea prior to hole interaction with matter, as it was not yet formulated, so he had not yet recognized the way to produce and utilize the practical antigravity potential of his discovery.

1.1.6 A Scientific Dilemma

There would seem to be a sufficient abundance of techniques, devices, processes, and theoretical works to impel a crash project in the scientific community to develop successful overunity electrical power systems {43a-43d}. This would be especially appropriate at this time, since the escalating energy crisis now threatens to slow the world economy, and then eventually collapse it. At this writing, the MidEast has heated considerably, and appears headed for another MidEast war, with resulting severe interruptions of cheap oil supplies from the MidEast. Nonetheless, in spite of Russia having opened its oil field spigots to try to fill the need, a crisis in oil appears to be looming.

---

[21] Kondepudi and Prigogine, *ibid.*, p. 459 already point out that strong gradients produce situations that violate present thermodynamics. Research in these and other situations violating present thermodynamics is going forward under the caption of "extended thermodynamics" research. A discussion of the area is given by D. Jou, *Extended Irreversible Thermodynamics*, Springer-Verlag, New York, 1996.

So *why* does the U.S. scientific community still so adamantly oppose the very notion of Maxwellian systems freely extracting EM energy from the vacuum? Why is there no outcry pointing out what the hoary old Lorentz regauging really means in terms of equilibrium or disequilibrium of the regauged system with the active vacuum? The unequivocal participation of the vacuum in a continuous energy exchange with the charges and dipoles of every EM system has long been affirmed by particle physics. All the fields and potentials — and their energy — manipulated in any EM circuit come from the vacuum, as proven in particle physics for 45 years. Why have our electrical scientists not understood — from the broken symmetry of the vacuum exchange with the opposite charges of the source dipole between the charged terminals of every generator and battery — that *EM energy from the vacuum powers, and has always powered, every electrical power system and circuit ever built?* Why have the later rigorous broken symmetry findings of particle physics not been incorporated to update the ancient classical EM theory used to design and build electrical power systems, nearly a half-century after those broken symmetries were discovered and proven? Why do our classical electrodynamicists continue to assume that every charge and dipole in the universe is a perpetual motion machine, freely creating energy out of nothing and pouring it out across all space at the speed of light?

Therein lies one of the real problems of present science — its historical and continuing resistance to "out-of-the-box" thinking[22] and to research that overcomes conventional strictures. More than 40 years after the basis for the vacuum-energy powering of every dipolar system (and of every observable charge once its clustering virtual charges of opposite sign are accounted) was discovered and proven in particle physics, all our university electrical engineering departments continue to erroneously teach that the shaft energy input to the generator powers its external circuit. Internal EM energy in the generator, transduced from the input shaft energy, does not directly add a single watt to the external circuit. Instead, it only continually forces the generator's internal charges apart, to continually remake the source dipole, which then extracts energy from the vacuum and pours it out of the terminals, filling the space surrounding the conductors of the external system for the system to intercept and utilize some of the available external energy flow.

---

[22] "Out-of-the-box" thinking is a widely used concept among planners and program formulators, when conventional thinking will not suffice to solve a major problem with which they are struggling. Conventional thinking is considered "in the box" thinking, so by demanding out-of-the-box thinking, a problem demands an unconventional solution outside those normally proposed. In short, some new thoughts and concepts are required. Much lip service is given to the concept as a favored buzzword, but few proposed programs with truly "out of the box" approaches will be funded. In the energy field, none at all are funded that are truly "out of the box", whether or not that or similar phraseology is used.

Once the dipole is established, it will extract and transduce EM energy from the vacuum and pour it out in all directions at the speed of light, without ceasing. Else, the Nobel Committee should admit its grave error and revoke the Nobel Prize awarded to Lee and Yang. Real observable EM energy extracted and transduced from the vacuum's virtual energy is precisely what the "broken symmetry of the opposite charges" on the ends of the dipole *means*. Indeed, all the forces of nature are already considered as generated by the interaction of virtual particles with observed particles. Since force produces energy changes in the system affected, then it follows directly that energy changes are produced by the interaction of virtual particles with observable particles.

However, our power system engineers ubiquitously use the closed current loop circuit. This inane circuit self-enforces the Lorentz symmetrical regauging condition. It uses half the energy captured by the external circuit (from intercepting some of that "energy flowing around the circuit from the seething vacuum" that is copiously pouring out of the generator terminals) to ram the spent electrons back through the back emf of the source dipole itself. That scatters the dipole charges and kills the dipole and its broken symmetry — and also kills the free flow of transduced EM energy from the vacuum.

That insane circuit kills the source dipole in the generator faster than it powers its external loads! So one has to keep rotating the shaft of the generator, to keep producing a magnetic field inside the generator, so that this magnetic field energy can continue to force the charges back apart and continually reform the source dipole.

In short, our engineers build the equivalent of an electrical windmill, then — so to speak — force it into a closed barn so the environmental wind cannot reach it any longer. It is little wonder that we ourselves then have to do work on that "electrical windmill" to crank it around!

We pay the power company to engage in a giant wrestling match inside its own generator and *lose*. We also wildly pollute the planet with hydrocarbon and nuclear wastes, poison species (including killing of X numbers of humans every year via the disruption of their body functions and health), and enhance global warming. We are slowly strangling our biosphere and ourselves. None of that is necessary.

1.2 Organized Science Often Resists Innovative Change

The history of science is littered with examples where the scientific community has ignored the principle behind Einstein's statement quoted at the beginning of this chapter. As pointed out by Smith {44}, science has become reoriented toward profit. Quoting:

> *"...science is not the danger; scientists encouraged to do bad science to survive are."* ... *"...changing the way modern science is funded is an enormous undertaking, but it is a necessary one if we want to protect our future. Call it managed risk."*

Science's resistance to change is so well known to historians of science that it is rather universally accepted — although usually not made explicit to undergraduates. Further, scientists are under great pressure to conform:

1. After all, science is *patronized*; someone must fund the laboratories, the research supplies, the salaries and personnel benefits, etc.
2. To procure and protect its patronage, science has become quite organized, particularly with respect to how the funds — received and channeled down from the top — are cut into individual packages (research grants and research programs) and made available for competition among the "performing" chain of universities and research laboratories.
3. In the last few decades, there has risen an increasingly fierce demand by universities that the scientific researchers (i) be successful in attracting outside funds and (ii) file patents assigned to the university. So fierce has this demand become, that the research professor's continued livelihood may literally depend upon his or her success in bringing in extra funding. Further, much of his time is now spent in writing proposals to

compete for these "packaged funds". When he wins them, the kind of research and the areas of research are already stringently defined, and he dare not deviate — else there will shortly be no more funding packages won, no funding for his graduate students, and soon thereafter there will be no job of any importance for the professor!

4. In the United States, all the government National Laboratories and our universities are intensely seeking and filing patents! So small, independent inventors cannot deal with these organizations, without risking and almost guaranteeing the loss of their patents and intellectual property. Funding dangled in front of the inventor, much like a carrot dangled in front of a horse, is often tied to "march in" rights {45, 46} calling for surrender of the inventor's patent rights altogether, *whenever the government — i.e., a single bureaucrat — wants to take it*. All that has to be done is to declare that the inventor is not getting it developed and to market fast enough. Science has thus become more avaricious and — some inventors would even say — it increasingly involves overt and covert piracy of intellectual property rights. From personal experience, reluctantly I would not argue with that statement. Simply ask Larry Fullerton of Time Domain Corporation about his struggle with a National Laboratory over patent rights to his ultrawideband communications invention and technology.[23] It eventually resulted in a "draw" of sorts. He did not lose his rights, but the government gained them also, in competition with him. The government *circumvented* his patent, even though they did not succeed in taking it. We were delighted to recently see that Larry (the company is Time Domain, Inc.) received its 74th patent in this technology, as well as a ruling that will allow the technology to at last go to market.

5. The result is a dramatic increase in the pressure on working scientists and independent inventors to conform, and to "play the game by the rules". Then everybody up the scientific food chain is fed, and is happy and secure. The journals happily publish the research papers and results, the professor gets

---

[23] Stephen Fenichell, "Radio Flyer," *Discover*, 22(5), May 2001. Fullerton's technology has been given a limited go-ahead by the FCC, which has drafted standards and regulations in the area as of February 2002.

funding for his graduate students, the university gets that wonderful overhead cut of the research funding — such as half or it or more — and the entire apparatus is like a very large and tidy Titanic adhering to its ponderous course toward the iceberg. Meanwhile, truly new and innovative science discoveries — vulnerable and desperately needing nurturing funding during their initial embryonic state — get shoved aside, crushed, and starved in the funding rush to adhere to performance of the *prescribed* funding packages.

6. In this environment, the day of the "defenders of the scientific faith" has arrived! A small percentage of conventional scientists who are dogmatic and vociferous, are now very prominently attacking any novel experiments and ideas with a vehemence seldom seen in organized science. It is again reminiscent of some of the noted scientific attacks in history, e.g., as pointed out by Hellman {47}. Yet, because of the financial pressure upon the scientific community, and the increased pressures to conform, there is little restraint of the dogmatists and they are almost never called to task. Cold fusion is a current example. The American Physical Society has recently issued a statement condemning perpetual motion machines — yet the society's members continue to condone and use a classical EM model that assumes every charge in the universe to continuously be creating energy from nothing. Even the American Physical Society has not recognized what broken symmetry of opposite charges means with respect to the common dipole and dipolarity in every circuit. Nuclear reactions at low spatial energy (which means at extraordinarily high total energy when the c^2*-compressed* time energy is considered) do indeed sometimes occur in carefully controlled experiments, whether or not we yet sufficiently understand the reactions theoretically, and regardless of whether we can get the anomalous results to happen every time. Yet this area of nuclear interactions at low *spatial* energy {48} — and unknown to the scientific community, at very high *time* energy {49, 50, 51} — has been savaged by these self-appointed spokespersons for the "official" community, none of whom even account the compressed energy in the time increment portion of the photon. More than six hundred successful experiments in multiple laboratories, by respected scientists in multiple nations worldwide, are now

> rather resoundingly ignored. Yet the replication rate for good cold fusion experiments is certainly higher than many of the replication rates for novel and little-understood phenomena in large particle accelerators, and the cold fusion experiments are also *far cheaper*.[24] While particle accelerators are "popular" in their ability to garner huge funding, their cost/benefit ratio compared with, say, cold fusion experiments, may be abysmally low. Simply examine the decades of effort and many billions of dollars expended on the search for *warm* fusion (using spatial energy only). What has it produced, in terms of watts of power on the power lines? When will it produce any electrical power of any significance? Prototype cold fusion power systems have in fact been produced and patented {52}. With seed money from the scientific community and using a higher symmetry electrodynamics, cold fusion power could proceed at a rapid pace.

Strangely, the ever-present pressure to conform to *that which is already known and accepted* has often made science its own worst enemy throughout its history. Establishment scientists and the "system itself" now probably block — and have blocked over the decades — more innovative scientific research than does any other factor {53}.

1.2.1 Many Scientists and Historians Have Pointed It Out

The scientific community is well known to have always been highly resistant to novel ideas and innovations. Here are some selected pertinent comments regarding this phenomenon, where organized science is itself the obstacle to the advance of science, and where such has been recognized for many decades:

---

[24] As an example, see R. P. Taleyarkhan *et al.*, "Evidence for Nuclear Emissions During Acoustic Cavitation," *Science*, Vol. 295, 8 Mar. 2002, p. 1868-1873; Charles Seife, "'Bubble Fusion' Paper Generates a Tempest in a Beaker," *ibid.*, p. 1808-1809. See also Donald Kennedy, "To Publish or Not to Publish," *ibid.,* p. 1793. *Science* had the courage to publish the peer-reviewed results of a tabletop sonoluminescence experiment that apparently produces nuclear reactions. Editor Kennedy essentially advises all protagonists on both sides to cut the rhetoric and allow the scientific community to do its replication work, to either validate or refute the successful experiments of Taleyarkhan *et al*. This action by *Science* is a shining beacon to remind the scientific community that science is based on experimental method, and that prevailing theories cannot refute new experiments that contradict them. Instead, laboratory bench experiments must decide such an issue.

"Every great scientific truth goes through three stages. First, people say it conflicts with the Bible. Next they say it had been discovered before. Lastly they say they have always believed it." [Louis Agassiz, 1807-1873.]

"There are three steps in the history of a great discovery. First, its opponents say that the discoverer is crazy; later that he is sane but that his discovery is of no real importance; and last, that the discovery is important but everybody has known it right along." [Sigmund Freud].

"Anybody who has studied the history of science or worked as a scientist knows that whenever something novel is discovered or proposed, there is a polarization of scientists, with hostility and bitterness that may last for generations. What wins arguments is scientific fact, and that may change as the years go by. A good example of this is the geological theory of continental drift, as proposed by Wegener in 1912. When I studied geology around 1950, continental drift was acknowledged in my undergraduate textbook as a crank theory. The first serious confirmation was in 1956, and it was finally established as the dominant theory in the early 1970s. Until that time, anybody who admitted that he or she believed in continental drift was the subject of derision and scorn. Sorry, folks, science is not and has never been the 'idealized portrait painted in textbooks'." [Allan Blair] {54}

"... the four stages of response to any new and revolutionary development [are]: 1. It's crazy! 2. It may be possible — so what? 3. I said it was a good idea all along. 4. I thought of it first." [Arthur C. Clarke]. {55}

"...I suggest that most revolutions in science have taken place outside the lofty arena of the refereed journals, and with good reason. The philosophy by which these journals govern themselves virtually precludes publication of ideas that challenge an existing consensus." [William K. George] {56}

"At every crossway on the road that leads to the future, tradition has placed against each of us, 10 thousand men to guard the past." [Maeterlinck].

> *"An important scientific innovation rarely makes its way by gradually winning over and converting its opponents: it rarely happens that Saul becomes Paul. What does happen is that its opponents gradually die out, and that the growing generation is familiarized with the ideas from the beginning."* [Max Planck] {57}.

> *"Peer review is widely seen as a modern touchstone of truth. Scientists are roundly drubbed if they bypass it and 'go public' with their research… The first limitation of peer review is that nobody can say quite what it is… A more pernicious danger is that peer review may reject the important work. As Charles W. McCutchen, a physicist at the National Institutes of Health, has put it, peers on the panel reviewing a grant applicant 'profit by his success in drawing money into their collective field, and by his failure to do revolutionary research that would lower their own ranking in the profession. It is in their interest to approve expensive, pedestrian proposals.' "* [Jonathan Schlefer] {58}.

The sheer massive size and inertia of the modern scientific establishment also exert mind-numbing difficulty in "hearing" and recognizing an innovative scientist's message, even a message of utmost importance, and even if it gets through the censors. For example:

> *"We used to be able to say things once; if the message was reasonable, it had a good chance of becoming a permanent part of the structure of the field. Today, a single publication is lost; if we say it only once, it will be presumed that we have changed our mind, and we therefore must publish repeatedly. This further fuels the large publication volume that requires us to repeat."* [Rolf Landauer] {59}

1.2.2 Some Specific Examples

There are hundreds of examples of new discoveries in science that have at first — and often for an extended period of years — been severely obstructed and ridiculed. Here are just a selected few:

1.2.2.1 *Conservation of Energy*

Von Mayer {60}, the discoverer of the modern statement of the conservation of energy and the mechanical equivalent of heat, was severely chastised for his "insane" work. He was hounded and severely ridiculed. This extremely harsh treatment, together with domestic problems, drove him to a suicide attempt and a nervous breakdown, and into psychiatric treatment for some years. Toward the end of his life, his principle of energy conservation had so greatly increased the ease of calculations and the understanding of systems that the same scientific community — due to the commendable efforts of Helmholtz, Clausius, and Tyndall — began to recognize his great contributions and lionize him. In 1867, he was made a member of the nobility, dying in 1878 with his "insane" work by then well respected. He was fortunate to have the "cur dog attack" reversed in his lifetime. Most scientists with novel discoveries are not so fortunate.

1.2.2.2 *Continental Drift*

Alfred Wegener {61} proposed the theory of continental drift in 1912. The reception was extraordinarily hostile. So ferociously was he ridiculed for the notion that huge continents of rock could "float" and "drift" that his very name, "Wegener", was often used as a synonym for "utter idiot". To refer to someone as "a Wegener" was to cast the greatest slur possible upon that person's mental powers and to label him a gibbering lunatic. Only in the 1960s when sea-floor spreading from ocean ridges was discovered, proving that ocean basins are not permanent features, did Wegener's concept of continental drift concept finally gain acceptance.

1.2.2.3 *Kinetic Theory of Gases*

As pointed out by Paul Nahin {62}:

> *"J. J. Waterston's paper on the kinetic theory of gases, in 1845, was rejected by the Royal Society of London. One of the referees declared it to be 'nothing but nonsense, unfit even for reading before the Society.' ... "Waterston's dusty manuscript was finally exhumed from its archival tomb forty years later, because of the efforts of Lord Rayleigh..."*

Lord Rayleigh was the Secretary of the Royal Society when he had Waterston's paper reprinted nearly a half-century after submitted. Lord Rayleigh also gave an introduction to the paper, regretting it lying so long unpublished when its content was quite important. His introduction is a way of explaining the delay.[25]

---

[25] Waterston's paper was finally published as John James Waterston, "Free and Elastic Molecules," *Phil. Trans. Roy. Soc. Lond.*, Vol. 183, 1892, p.1-79. Lord

1.2.2.4 *Surgical Pain Deemed Necessary*

It has *always* been this way in science. As another example, the famed surgeon Alfred Velpeau wrote in 1839:

> *"The abolishment of pain in surgery is the chimera. It is absurd to go on seeking it today. 'Knife' and 'pain' are two words in surgery that must forever be associated in the consciousness of the patient. To this compulsory combination we shall have to adjust ourselves."* [Martin Gumpert] {63}.

Wryly we observe that today a similar attitude of "we must glory in the pain" — where the "pain" is due to the yoke of COP<1.0 EM systems and of the second law of classical equilibrium thermodynamics — consumes most modern electrical power system scientists and engineers.

1.2.2.5 *The Photoelectric Effect*

Almost every household now knows Albert Einstein's epochal achievements. Yet his formative three papers — on Brownian motion, the photoelectric effect, and special relativity — were published in 1905 while he was working in the Swiss Patent Office. The most renowned physicist of the time was Max Planck. Planck was embarrassed that a scientist who was not even employed in physics was doing such important work in physics. So Planck and other scientists arranged for Einstein to be awarded a chair in physics at a proper university. In their letter to the university, they pointed out Einstein's brilliance in his papers. They also then excused him for straying down the road of the photoelectric effect, because — as they put it — everybody knew that was foolishness, but persons of such brilliance could be forgiven a few such little bobbles along the way. Years later, in 1921, Einstein was awarded the Nobel Prize for Physics, for his work in theoretical physics and *especially* for his explanation of the photoelectric effect.

The Einstein incident is a typical illustration of Arthur C. Clarke's cogent observation:

> *"When a distinguished but elderly scientist states that something is possible he is almost certainly right. When he states that something is impossible he is very probably wrong*." [Clarke's First Law]

Rayleigh's introduction and Waterston's paper also are given in Jefferson Hane Weaver, *The World of Physics*, Simon and Schuster, New York, 1987, p. 632-651.

1.2.2.6 *Amorphous Semiconductors*

Stanford R. Ovshinsky's development of amorphous semiconductors {64, 65} is another modern example. "Everybody knew" that a crystalline structure was necessary in order to have a semiconductor at all; in short, a semiconductor formed out of non-crystalline material was deemed to be totally impossible, never mind this “phase” change that Ovshinsky advanced. However, some young graduate students, e.g., began to look at Ovshinsky's amorphous materials and his phase change approach.

Finally funded by the Japanese, Ovshinsky's company, Energy Conversion Devices, Inc. (ECD) simply placed its amorphous semiconductor devices into working equipment. Copy machines appeared with his amorphous semiconductors installed in them — with the machines and their amorphous semiconductors working very well. More graduate students and post-docs enthusiastically entered the area, did research, and wrote dissertations.

As a result, amorphous semiconductors finally became accepted, and they are now part of the established technology and scientific knowledge base. The interested reader can simply look up Ovshinsky's company and statistics on http://www.ovonics.com . The Japanese have reaped a continuing bonanza from the sales of amorphous semiconductors, because of the shortsightedness and bias of the U.S. scientific community.

1.3 Scientific Disagreement or Street Fight?

It is one thing to scientifically disagree — even strongly — in a technical and courteous fashion, and quite another to engage in ad hominem attacks. The first is science; the second is a cur dog fight. Unfortunately the history of science far too often reveals "cur dog fights" instead of respectful scientific disagreements {66}.

A recent modern example of legitimate research and qualified researchers still being savaged in a "cur dog fight" manner is cold fusion (low energy nuclear reactions). Quoting one learned orthodox scientist, whose name is withheld, speaking to a learned scientist in cold fusion:

> *“How stupid do you think we are? My assessment of you and your colleagues is that you are complete frauds or totally mad. There is no known physical principle that would support the kind of results that you claim your technology can accomplish, nor is there any credible argument why there should be such a principle."*

We answer that "cur dog attack" unemotionally and scientifically. The key is in the scientist's own phrase: "no *known* physical principle." If the experiments work and are replicated, but are not understood by the conventional theory, then they refute the prevailing theory, whether the physical principle is understood or not. Else scientific method has been totally abandoned in favor of “truth by authority and dogma”. The experiments clearly show that a previously *unknown* physical principle is at work. The real task then is to rediscover this new principle — that being one of the primary ways that science advances. We propose new principles in this book as a possible explanation of those experiments.

The dogmatic scientist's statement was made in spite of some two hundred (200 at that time; now more than 600) scientific experiments worldwide, in many nations, many at prestigious institutes, where dozens of scientists have reported positive and anomalous results in cold fusion experiments.

So yes, there is indeed a very "credible argument" — *the results and replicability of the experiments*, which are supposed to be the *decisive* statement of science. If that view in science has changed, then we are no longer practicing the scientific method. Instead of accepting successful and replicable experiments and seeking to change the model, the scientist is insisting that we must first understand the principle and thus have a model. This of course is a total violation and reversal of the scientific method. We used aspirin effectively for decades without the slightest notion as to the mechanism enabling its beneficial actions. The mere fact that "there is no known physical principle" for the results achieved has nothing to do with the validity of the replicated experimental results. Instead, it merely substantiates that *there should be a vigorous scientific program to uncover the new principle or principles obviously involved*, since the old model has either failed or been revealed as too limited.

The derogatory statement by the dogmatic scientist is also made from the viewpoint of the conventional nuclear physics model. The conventional physics has not taken into account that all 3-spatial electromagnetic energy associated with charges and dipoles and their fields and potentials in fact comes from the time domain {85, 86, 12, 19}. It has also not taken into account that time itself (as in the time component transported by the time-polarized photon or a time-polarized EM wave) may be comprised of extraordinarily dense energy. Indeed, time appears to be spatial EM energy compressed by the factor c^2, so it has the same energy density as mass {67}, which we pointed out earlier. The smaller the *spatial* energy of the photon, the greater its time component and hence its *time*-energy in seconds expressed in decompressed spatial energy joules[26]. The highest energy particle physics is not *spatial*-energy physics as presently practiced, but *time*-energy physics, where some of the time-energy of photons is transduced into spatial energy. One second of time transduced (decompressed) into spatial energy yields approximately $9\Delta10^{16}$ joules. Further, every negative charge in the universe continuously accomplishes that decompression, and every positive charge continuously accomplishes the recompression.

In other words, not all physical principles were discovered in the particle physics of, say, five years ago, as the literature since then clearly shows. To assume that everything is already known today is ludicrous; science is never completed. The new principles reported in the present book now are at least "candidates" for the physical principles that do support cold fusion results. There is also another powerful argument for these hypothesized principles: they have produced the final resolution of the source charge problem — something for which the arch skeptic quoted has no solution whatsoever, and which he himself cannot explain. *If the arch critic cannot produce an alternate solution to the source charge problem, and is unaware of that new principle that may solve it, let him go and learn the new principle. Until then, he unwittingly assumes that every charge in the universe is a perpetual motion machine, continuously creating and pouring out EM energy in all directions at the speed of light.* So we return his own argument and approach to him: he should attack the conventional scientific community for accepting the source charge, while having absolutely no principle to explain how it continuously pours out observable EM energy without any observable EM energy input. In short, he should practice what he preaches; else he brands himself a total hypocrite. The alert reader will note that none of the arch skeptics doing all the ad hominem attacks on cold fusion and COP>1.0 EM systems are practicing what they preach.

---

[26] The spatial energy of the photon decreases linearly as the frequency is lowered, while the time component in seconds increases linearly. But the highly compressed time energy, comprising that time component, increases nonlinearly (by the factor $c^2 \div t$). Hence the highest energy photons are actually the low frequency photons — something completely alien to particle physicists — and not the high frequency photons chased by "high energy physics". Indeed, high energy physicists are practicing a high spatial energy physics, which — overall — is a much lower energy physics than radar, microwave, VHF, or ELF, *if the total energy of the photon, to include its time energy, is considered.*

What we are saying is this:

(a) If one's physics knowledge was current five years ago, it may be stale today. A few years ago, no one believed the expanding universe was accelerating. Today we know that it is, rather unequivocally. And by "no known physical mechanism". We will later present a strong candidate for that missing mechanism. The results of experiments will substantiate or refute it.

(b) The proposed principles in this book, e.g., are newer, and may shed light on the mechanism for the cold fusion results as well as other phenomena. The experimental results themselves are irrevocable; any proposed explanation requires validation.

(c) The new principles do explain cold fusion and are consistent with the phenomena encountered in multiple experiments by multiple researchers in many laboratories. We readily admit that the stale and incomplete principles presently utilized in nuclear physics *do not* explain cold fusion. Neither do they explain how a charge continuously pours out EM energy. So do the skeptics attack all those physicists who believe in source charges and their provision of the fields and potentials and all EM energy? Of course not. They themselves believe in that greatest of all “perpetual motion” *faux pas*.

(d) The proposed new principles also solve the source charge problem, which is still ignored by most conventional physicists and electrodynamicists, even though often referred to as the "most difficult problem in electrodynamics" {68}.

(e) Scientists should not be close-minded, but should consider new proposals and let the experimental results decide their validity or falsity — precisely the position taken by the journal *Science* in publishing the results of some new cold fusion experiments. That is the scientific method in action.

(f) Those scientists who remain close-minded, and viciously attack experimentally demonstrated new processes and mechanisms, are guilty of practicing dogma and not science. They are in fact guilty of being what they so frequently charge: *pseudo-scientists*.

When science does not allow proposed new mechanisms and principles to be considered in science following demonstration of new phenomena inexplicable by present models, then science is no longer practicing scientific method. When any scientist rejects these demonstrated new experimental phenomena out-of-hand, he is practicing dogma rather than science. The “friendly skeptic” attitude is welcomed and appropriate. The cur dog attack has no place in science, but only among cur dogs — and those who behave like them.

1.4 A Few Significant and Unresolved Scientific Problems

1.4.1 Time as Structured Energy

Time also has internal structure and dynamics {69, 70}, as we explain later, as well as the same energy density as mass. So it should not be surprising that a tiny bit of time-energy, transduced into ordinary spatial EM energy, might be capable of energetically inducing a wide variety of nuclear reactions. One has transduced not only "gross" energy, but also energy structuring and dynamics — which *act internally* on any object with which interaction occurs. Further, the overall spatial energy component and the time-energy component in a photon are canonical. Being quanta, all observable photons have the same fixed "total magnitude" with respect to angular momentum. The photon is also "made" of $(\div E)(\div t)$ which may be taken as $(\div E)(c^2 \div t)$ in terms of *purely spatial* energy equivalent content, assuming that the time energy is decompressed into spatial energy. As can be seen, the so-called "low (spatial) energy" photons are precisely the photons that transport the greatest time-component. When that time-energy component $(\div t)$ is converted to spatial energy $(\div E_C)$, the formula is $(\div E_C) = c^2(\div t)$. As can be seen, the converted time energy will produce far more spatial energy in the *converted* interaction, than was carried by the photon in its spatial energy $(\div E)$ component prior to interaction and time-energy transduction. Hence the highest energy photons are actually low frequency photons — under the circumstances where time-energy transduction into spatial energy is involved. With transduction, the highest energy particle physics can be conducted at low frequencies and low spatial energy if conversion of time-energy occurs in the interaction.

Since cold fusion involves transduction of a little bit of the time-energy, the total lack in physics of knowledge of time-energy transduction into spatial energy explains why conventional nuclear physicists find the cold fusion results so confusing and astonishing.

1.4.2 The Obsolete Foundations of Classical Electrodynamics

Another modern example of science's resistance to change is the failure to update classical electrodynamics to include the active vacuum interaction and the interaction with the local curved spacetime[27], as we shall discuss in Chapter 2. It has been scientifically established (in particle physics) that there can be no equilibrium of any physical system without accounting for the vacuum interaction.[28] To simply observe a physical electrical power system sitting there stably and running, is to prove its vacuum interaction in order to even *have* such equilibrium.

An interesting point then arises since every charge and dipole in the system is a *broken equilibrium* in the exchange between the material system and the vacuum. Not only does an electrical circuit receive some energy from the vacuum, but also the energy from the vacuum is *massively hemorrhaging* from the vacuum into the system (and back out)! The isolated charge, e.g., is infinite if one removes the screening clustered virtual charges of opposite sign, as is well known in modern physics. Further, the screening virtual charges also have infinite charge, again as is well known. The difference between these two infinite values of charge, e.g., gives the standard finite observed charge of the charged particle. The lay reader may wish to positively verify that statement; e.g., as given by Nobelist Steven Weinberg.[29] Quoting:

> *"[The total energy of the atom] depends on the bare mass and bare charge of the electron, the mass and charge that appear in the equations of the theory before we start worrying about photon emissions and reabsorptions. But free electrons as well as electrons in atoms are always emitting and reabsorbing photons that affect the electron's mass and electric charge, and so the bare mass and charge are not the same as the measured electron mass and charge that are listed in tables of elementary particles. In fact, in order to account for the observed values (which of course are finite) of the mass and charge of the electron, the bare mass and charge must themselves be infinite. The total energy of the atom is thus the sum of two terms, both infinite: the bare energy that is infinite*

---

[27] We again stress the concept of the *supersystem*, introduced in footnote 11.

[28] E.g., see T. D. Lee, *Particle Physics and Introduction to Field Theory*, Harwood, New York, 1981. On p. 380-381, Lee shows how there is no symmetry of matter alone, but only of matter and vacuum.

[29] Steven Weinberg, *Dreams of a Final Theory*, Vintage Books, Random House, 1993, p. 109-110.].

> *because it depends on the infinite bare mass and charge, and the energy shift ... that is infinite because it receives contributions from virtual photons of unlimited energy."* [Steven Weinberg].

So a simple charged particle actually involves a polarization of the vacuum involving two charge energy functions — the bare charge and the charge change functions — each of which has a known and recognized COP = ←!

A priori, the conventional system's overall actions must incorporate functions (whether intentional or unintentional) that continuously adjust to provide *net equilibrium in that exchange by adjusting the hemorrhaging-out to equal the hemorrhaging-in*. In this book, we will spell out just what those unintentional functions are, such as the ubiquitous closed-current-loop circuit, and how to beat them so that excess energy from the vacuum can be utilized by the system to power the loads.

1.4.3 What Powers an EM Circuit or Electrical Power System?

Another modern example of science's resistance to change is the continued engineering of electrical power systems with the erroneous notion that mechanically powering the shaft of a generator directly powers the power line. The notion is that the generator transduces some of the mechanical shaft energy into output EM energy added to the power line {71}. To the contrary, *all EM systems are powered by energy extracted from the vacuum*. They are not powered by the mechanical energy we input to the shaft of a generator, or by the chemical energy in a battery. In this book, we discuss at some length what powers the EM system, and we have previously published the basis for our "shocking" statement {12, 72}. The basis for how an EM system is powered by energy from the vacuum has been known in particle physics for nearly a half-century {73}, since the experimental proof of broken symmetry {74}, including the broken symmetry of opposite charges (and hence of any dipole) in its fierce energy exchange with the active vacuum {75}.

A generator or battery expends all of its available energy to separate its internal charges and form its source dipole between the terminals. Once made, the dipole's broken symmetry — in its violent energy exchange with the active vacuum — converts virtual photon energy absorbed from the vacuum into observable EM energy, and emits it — *pours it out* — through the terminals and along the power line, filling all space around the conductors.

If one wishes to appreciate the enormity of the vacuum changes engendered by formation of that source dipole, simply visualize those two infinite charge energy functions (from Weinberg's quotation above) for every charged particle on the ends of that dipole. The dipole then is a great set of offsetting infinities in EM energy (photon energy) boiling and being ordered in the vacuum. Making a simple little dipole or producing a simple charge initiates into motion an enormous set of infinite energy changes in the vacuum! None of these startling, enormous vacuum energy changes and functions is modeled in classical electrodynamics and electrical power engineering. So when we speak of the Heaviside nondiverged energy flow component pouring from the terminals of a generator, as being a trillion or ten trillion times as great in magnitude as the intercepted Poynting component, the reader should not be appalled. The reader should recall that we are dealing with the difference between sets of infinities that combine to provide finite differences. These finite differences can be very small, or they can be extraordinarily large, depending on the exact situation. In the case of a source dipole formed in an ordinary generator or battery, the difference of the infinities is very large.

Figure 1, adapted from Kraus {76}, illustrates that small part of the external energy flow around a typical transmission line that is intercepted by the surface charges and their fields, and that is drawn into the wires to power the circuit as these charges are driven axially back and forth across the conductors. The surface charges are constrained to the "drift velocity" (usually a few inches per hour) movement down the wire by repulsion of the electrons ahead of them.

The spinning, *longitudinally restrained* electrons precess laterally, thus withdrawing across the transverse axis of the wire. So the laterally precessing electron withdraws a little ways, withdrawing that first small portion of its fields and their energy increased by intercept of additional energy from the outside energy flow. The further part of the fields is not withdrawn into the conductor, and is not used to power the electrons.

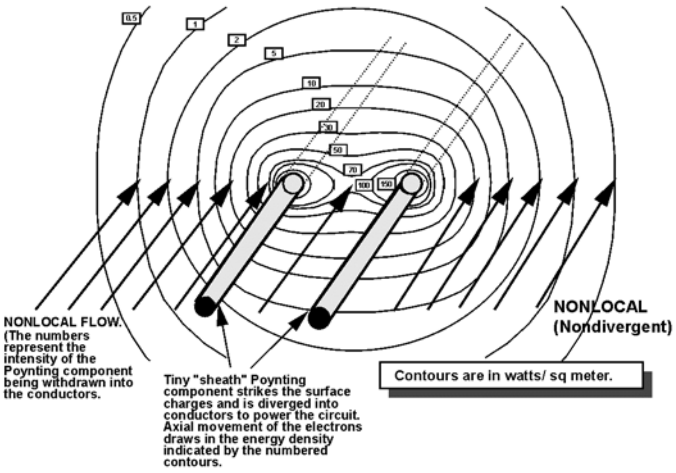
0.5
1
2
5
10
20
30
50
70
100
150
NONLOCAL FLOW.
(The numbers represent the intensity of the Poynting component being withdrawn into the conductors.
Tiny "sheath" Poynting component strikes the surface charges and is diverged into conductors to power the circuit. Axial movement of the electrons draws in the energy density indicated by the numbered contours.
NONLOCAL (Nondivergent)
Contours are in watts/ sq meter.

Figure 1-1 Poynting (caught energy) flow contours surrounding a transmission line.

Most of the available energy flow misses the circuit and is not drawn into the conductors. However, that large component is not shown by Kraus, and Poynting did not consider it. Heaviside {5c} pointed out that the remaining flow component is hardly reduced (hardly changes direction) from the entire flow component before the extraction of the small Poynting component. Given sufficient intercepting charges outside the wires in separate receiving circuits, the total remaining energy flow that could potentially be intercepted is enormous — far more than the feeble amount of energy that is input to the generator shaft or that is in the chemical energy of a battery.

This is easily established by actual experiments placing intercepting charges in separate "receiving" circuits in the otherwise nondivergent energy flow outside the conductors, or one may show it with Bohren's experiment {24}. Collecting additional energy completely outside conductors is one part of the COP>1.0 operation of the motionless electromagnetic generator (MEG) {37}, to be discussed later in this book. The Sweet device {29a} established sustaining self-oscillation of the barium nuclei in a barium ferrite magnet, by using the surrounding vacuum as a semiconducting medium.[30]

1.4.4 The Incompleteness of Aristotelian Logic

Another false notion usually perpetuated in much of science and mathematics is that Aristotelian logic is complete and consistent {77, 78}. To the contrary, it is both incomplete and inconsistent, as is easily shown. Let us use the symbols "A" for a particular thing, "Ā" for that which is not-A, "∑" for "is identical to," "≢" for "is not identical to", "⋎" for "or", and "⋏" for "and". Using these symbols, the three laws of Aristotelian logic may be expressed as follows:

$$A \sum A \qquad [2]$$

$$A \not\equiv \bar{A} \qquad [3]$$

$$A \curlyvee \bar{A} \qquad [4]$$

The first law states that a thing is identical to itself. The second states that a thing is not identical to that which is not itself. The third states that a thing is either itself, or it is something else. Those are the three laws of Aristotelian logic. As written, the process of perception, observation, etc. *has been excluded*. More on that in a moment.

Specifically excluded are the laws

$$A \not\equiv A;\ A \sum \bar{A};\ A \curlywedge \bar{A} \qquad [5]$$

Which says that a thing is not identical to itself but is identical to something else that is not itself.

Heraclitus posed a profound challenge to Aristotelian logic, a challenge that has not been adequately resolved by Aristotelians to date. He simply observed that, for a thing to change, it had to change into something else. "But then," he asked, "how can a thing be itself but also something else as well?" In philosophy that is still known as the unresolved "problem of change". It has no resolution in Aristotelian logic, but can be resolved in an extended logic we shall advance. The solution to Heraclitus' problem of change is the nature of observation, as advanced in this book.

---

[30] E.g., see Richard E. Prange and Peter Strance, "The Semiconducting Vacuum," *Am. J. Phys.,* 52(1), Jan. 1984, p. 19-21. The vacuum may be regarded as a semiconductor. In particular, the vacuum in the region close to the nucleus of a superheavy element is analogous to the inversion layer in a field effect transistor. The authors introduce the idea of the inverted vacuum. Just as a semiconductor may be manipulated by subjecting it to external fields, doping etc., it appears that so can be the vacuum.

We now critique Aristotelian logic, by observing a simple Venn diagram of the type used in "proving" logic theorems in axiomatic logic. See Figure 1-2.

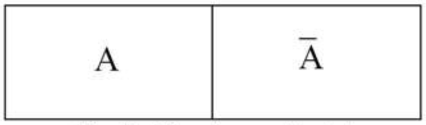

1-2a. Usual Venn diagram of A and A.

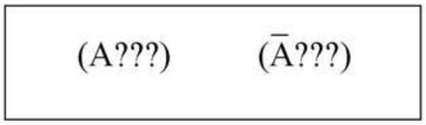

1-2b. Applying second and third laws to divider.

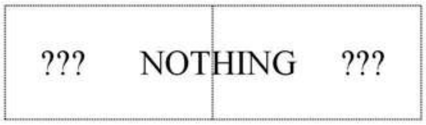

1-2c. Applying second and third laws to divider and outer boundary.

Figure 1-2 Applying Aristotle's laws to Venn diagram used to prove logic theorems destroys the entire diagram.

In Figure 1-2 a, we have introduced a dividing boundary line between A and Ā. *Unfortunately that dividing line belongs entirely to both A and Ā*. So it violates all three Aristotelian laws, and must be removed.[31] Therefore, we remove it in Figure 1-2 b. Now we have neither a discernible A or a discernible Ā, but we have removed naught but the boundary separating them, so we may argue that they are both still there although not discernible. However, if they are not discernible, we cannot distinguish what A is or what Ā is, and so we cannot discern whether they are identical or not. Here again this violates all three laws of Aristotelian logic, so we have to remove the diagram. Indeed, the outer rectangle line is also such a boundary, between the "inside" and the "outside" (the not-inside), and so it must be removed. In compliance with Aristotle's laws, we are left with the trivial diagram shown in Figure 1-2 c, which is nothing at all.

---

[31] A clever fellow once proposed regarding the boundary as a total discontinuity, being neither A nor Ā. However, in that case A and Ā could never meet, so there could not be a cause interacting with not-cause to produce an effect. So that suggestion as to how to "fix" Aristotelian logic does not hold.

If this Venn diagram method is objected to, then we must insist that all those papers and texts using that method of proof of logic theorems be either abolished or corrected!

Sometimes the objection is raised that, as far as the center boundary line is concerned in Figure 1-2a, the line "belongs to A on the left and belongs to Ā on the right". Fine! Then we have a unique situation where a line (call it L) is an entity made of two entirely different things, which we may refer to as L_L and L_R. Yet $L \sum L$, without any qualifications as to "sides" L_L and L_R. It can be shown that every point in L is a point in L_L, and is simultaneously a point in L_R. The point in L is obviously identical to itself, by Aristotle's first law. Further, the same point in L_L and in L_R is identical to itself, by the same first law. But L_L is the set of all such points, and so is L_R, and so is L. Therefore $L \sum L_L \sum L_R$, since each one consists of each point that the others consist of. There is absolutely no distinction between the three things themselves, unless we are to violate Aristotle's laws.

Note that in passing from left to right across the boundary, a "change" occurs. This simply re-resurrects Heraclitus' original objection that, according to Aristotelian logic, there can be no change. That is, a line cannot change from "a different entity as seen from the left" compared to that entity as seen from the right. None of Aristotle's three laws contains a "left and a right" for A or Ā.

So the question of logic is a little more complex than writing three simple laws and drawing simplified Venn diagrams. In short, one cannot have Aristotelian logic without having something else *outside* it, which follows directly from Gödel's theorem and proof {78}. Note that we are introducing the required notion of "perception" into the formal notions implied as axioms in Aristotle's laws, as that "something outside the prescription of the laws themselves" and in addition to them. We are perfectly free to *use* Gödel's theorem and observation.

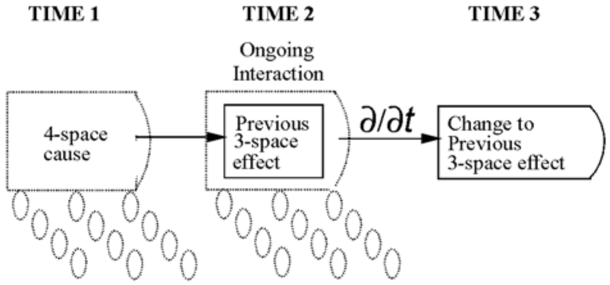

Figure 1-3 Mechanism for observation of change to a previous observation.

See Figure 1-3. We now add the notion that a thing is a perceived, thought, or observed thing. So let us use the symbol ⇒to mean *explicit*, i.e., *output* by a given perception, thought, or observation process as the result of a ∈/∈*t* operator having been applied to whatever exists prior to perception, thought, or observation. The symbol ⇔is used to mean *implicit*, e.g., when one observes an observation "Y" at time two, and later is deciding in time three whether or not that observation "Y" in time two is identical to a previous observation "X" made during time one, then *memories* of the observations of X and Y are involved in time three rather than the observations X and Y themselves, and so the observations of X at time one and Y at time two — by whatever manner they were observed and according to whatever decision algorithm is used — is said to be "implicit".

We note that we can know nothing about the so-called "thing in itself" *without* thought, perception, observation, or other process involving ∈/∈*t*. Further, at the moment the ∈/∈*t* operator is applied, time momentarily ceases. The resulting perception, thought, or observation exists therefore as a "frozen output" at that single moment in time. To have it "persist" or exist a moment later, we have to apply the ∈/∈*t* operator *again*, and stop time again, so that we again perceive, think, or observe.

But at any moment later than when we made a particular perception, thought, or observation of "A", *that* specific "perceived A" no longer exists, except in our memory as a recording of "observed A" that we can *continually recall*. Our conscious mind is a very fast serial processor, with only one "perception" or "thought" at each fleeting moment — only one slide at a time in the slide projector, so to speak. But it is very rapid. However, our so-called "unconscious" mind is totally conscious, just multiply so. It is a massively parallel processor, and has "a great many slides in the slide projector" at any given time.[32] Hence in our genetic multiprocessor mind, we can indeed record, recall, compare, etc. — and perform all the modern massively parallel computer operations.

Let us now re-examine the three Aristotelian laws. Let us label the "slide snapshots" with the time instants when each "perception, thought, or observation" is or was consciously made, by use of a subscript, such as A_1, which means *what was perceived, thought, or observed as "A" in time-snapshot 1*. Let us also note that each of the symbols $\sum$, $\not\equiv$, $^\vee$, and $^\wedge$ actually involves the output of a *comparison and decision algorithm* in the massively parallel processor, after the two observations X_1 and X_2 were made. Let us use the symbol ♥ to mean "results in the decision that" or "implies that". As a check for identity, e.g., simple comparison algorithms for determining identity or nonidentity might be

$$\{[A_1 - A_{(2)}] = [0]_3 \heartsuit_4 (A_1 \sum A_2) \qquad [6]$$

$$\{[A_1 - A_{(2)}] \prod [0]_3 \heartsuit_4 (A_1 \sum A_2) \qquad [7]$$

Thus we more precisely rewrite Aristotle's laws as

$$A_1 \sum_3 A_{(2)} \qquad [8]]$$

$$A_1 \not\equiv_3 \bar{A}_{(2)} \qquad [9]$$

---

[32] E.g., this can be seen by a moment's reflection. At any one time, the "unconscious" mind is controlling and directing a great multitude of ongoing physical processes, is also processing short-term and long-term memory processing, filing conflicts for resolution or later presentation to the conscious mind in symbolic fashion, etc. One can physically measure the electromagnetics associated with this activity, after the mind transduces its time-polarized EM interactions into 3-space EM actions. The process is two-way, and certain recorded EM stimuli will be "reverse-processed" back to affect the time-polarized EM operations of mind. Mind operations are electromagnetic; but they use time-polarized (scalar) photons and time-polarized (scalar) EM waves which are unobservable a priori. Mind operations are also electrodynamically engineerable, but that is beyond the scope of this treatise.

$$A_1 \vee_3 \bar{A}_{(2)} \qquad [10]$$

Where — surprise! We have now accounted for all those different times and periods of observations, processing, comparison, decision, etc. We are dealing not with some mystical "thing-in-itself", but with interactions (perceptions, thoughts, observations, associations) in the mind and psyche of the observer. And hidden in time 3 is the application of a decision algorithm such as given in [6] and [7] above.

Now the first law [8] states that in time one an observation was made and named "A", notated (A_1), by comparing it to a previously recorded observational memory of what we call "A". The existence of that *memory* of what we call A is implicitly assumed in the first law, as well as the others. We do not show the memory itself in [8], [9], and [10]. In time two an observation X_2 was made (not notated), but it is not then known at the moment whether that X_2 is A or $\bar{A}$. So in time interval three the decision algorithm to determine identity or non-identity occurred in the mind, and the zero output of that algorithm [6, 7] established that $A_{(2)}$ was actually identical to A_1, *according to the decision algorithm actually used. The algorithm matters, and it too is a variable.*

A similar process occurs in law two [9], but this time the algorithm had a different output. Note that the little line over $X_{(2)}$ to make it $\bar{A}_{(2)}$ was not assigned until time interval four (not shown) after the decision algorithm had given its output in time interval three.

The second law merely states that the operation of the decision algorithm in time interval 3 found the two snapshots (in time 1 and in time 2) not to be identical, *by the decision algorithm and comparison process utilized. Again, the algorithm matters, and it too is a variable that must be taken into account.*

All three laws [8, 9, 10] written one after the other *assume* that the identity algorithm does not change between snapshot 1 and snapshot 2, in all three of them. This is the key point.

With two different versions of the identity decision algorithm, the results of two different comparisons may differ. If the identity decision algorithm does not change during the time between snapshot one and snapshot two, then we have one case. This is like a person with good color vision, looking at a red marble beside a black marble. That observer clearly distinguishes the marbles, and to him they do *not* appear identical. Every time he looks again, they still differ because he did not change his decision algorithm between looks. On the other hand, a color-blind observer cannot distinguish the marbles, and to him they are seen as identical. Every time he looks, they are seen as identical, because his identity decision algorithm did not change.

Or, suppose a proponent argues that the "rightness" or "leftness" of the dividing line L, between A and Ā in the Venn diagram, can be taken into account. He is actually invoking a different algorithm (it belongs to the left and therefore to A) in one time snapshot than the algorithm (it belongs to the right and therefore to Ā) used in the other time snapshot. So his right side and left side of a line implicitly invokes the very point we are making: *the algorithm utilized is a variable and must be accounted.*

The point is this: *identity — whether in perception, thought, or observation — is not absolute, but depends upon the precise nature of the operation of the perception, thought, or observation process utilized and specifically on the "decision algorithm" used for "determining" identity or nonidentity.* For example, two antennas certainly "see" quite differently, if one is a VLF (very low frequency) antenna and the other is an IR (infrared) antenna! As another example, two observers in different frames may see a particular object as quite different observed things! In n-space, an observer in a frame at right angles to the lab frame, will see any mass in the lab frame as a wavefront going at light speed; in short, as a photon or photonic object, whereas the observer in the lab frame will continue to see it as just a common old mass object. So the "same object" *physically* differs to the two different observers, according to their frame of reference. An observer whose frame is rotated by three orthogonal rotations from the lab frame, will see that object as an "object existing in time only," i.e., as simply a sort of "thought" image, so to speak. To be more scientific, he may assign it to something called the "virtual state".

In short, we can violate any and all of the laws of Aristotelian logic, because identity per se is *perceived, thought, or observed* identity — the output of a variable decision algorithm — and is not absolute.

For clarity, we add a fourth law that violates all Aristotle's three:

$$A_1 \sum_3 \bar{A}_{(2)} \qquad [11]$$

All this really says is that, in time 3, the decision algorithm being used was changed from what it had been in times 1 and 2, and now could not distinguish between what was seen in snapshot 1 and what was seen in snapshot 2, even though to some "ultimate" observer the two were distinctly different.

In addition, we say that it "implicitly includes" the negations of all three Aristotelian laws. We then add an "application rule" (or a fifth law, as one wishes) as follows:

$$\{\ [(A_1\ \Sigma_3\ A_2)\Rightarrow^{\wedge}\ (A_1 \not\equiv_3 A_2)]\Leftrightarrow\}_5\ {}^{\vee}\ \{\ [(A_1\ \Sigma_3\ A_2)\Leftrightarrow^{\wedge}$$

$$(A_1 \not\equiv_3 A_2)]\Rightarrow\}_5 \qquad [12]$$

This winds up producing an extension of Aristotle's logic, so that the following explicit laws emerge:

$$A_1\ \Sigma_3\ X_{(2)}\ \heartsuit\ \ A_1\ \Sigma_3\ A_{(2)} \qquad [13]]$$

$$A_1 \not\equiv_3 X_{(2)}\ \heartsuit\ \ A_1 \not\equiv_3 \bar{A}_{(2)} \qquad [14]$$

$$A_1\ {}^{\vee}{}_3\ \bar{A}_{(2)} \qquad [15]$$

We also have the exact opposites of those three laws implicitly. The opposites can all be congealed into a single fourth law:

$$A_1\ \Sigma_3\ \bar{A}_{(2)} \qquad [16]$$

To make sense of these, we also have the master application rule or 5th law:

$$\{\ [(A_1\ \Sigma_3\ A_2)\Rightarrow^{\wedge}\ (A_1 \not\equiv_3 A_2)]\Leftrightarrow\}_5\ {}^{\vee}\ \{\ [(A_1\ \Sigma_3\ A_2)\Leftrightarrow$$

$${}^{\wedge}\ (A_1 \not\equiv_3 A_2)]\Rightarrow\}_5 \qquad [17]$$

We prefer to refer to this extension as a "four-law logic", where equations [13], [14], and [15] are the normal "explicit" laws, with the implicit law [16] understood but not explicitly written, and where equation [17] is the master application rule that puts it all together. But if one wishes to be rigorous, we have specified a 5-law extended logic that contains but expands Aristotle's 3-law logic. The application rule is the fifth law.

The point is this: In every case, we have a part of the perception, thought, or observation that obeys the explicit laws, and we also have a part that obeys the implicit laws. If opposites are not *explicitly* identical, then they are simultaneously *implicitly* identical. If opposites are *explicitly* identical, then simultaneously they are also *implicitly* not identical.

The new approach then covers what we observe more completely. As an example, any "wave" (observed) is implicitly a non-wave (particle), but not observably so. If its particle nature is observed and therefore explicit, its wave nature is unobserved and therefore implicit, and vice versa.

One also resolves such dilemmas as stating something "does not exist" but having to first call it into existence to even state it does not exist. In the new approach, it simply does not *explicitly* exist, but *implicitly* exists. *Not-being* is something like that: explicitly it doesn't exist, but implicitly it exists. The great problems of the philosophers were never solved, mostly because the logic brought to bear was Aristotelian and incomplete. The ultimate answers they sought all involved the fourth and fifth logic laws. So the ultimate answers all appeared to them to involve such things as "the accursed necessity for the identity of opposites." Precisely!

In physics, there has been the same problem over whether a fundamental particle is a particle or a wave. Physicists argued fiercely until they realized the futility of further argument, and settled for the duality principle. That simply states, well, in one case it can be a particle as observed in your experiments, and in another case it can be a wave as observed. So simply treat it as whichever one is useful! The philosophers, on the other hand, split into different schools, where each school had its "interpretation" usually involving a very "smooth" position statement of the school's position. In physics also, there are eight or more "interpretations" of quantum mechanics, for example. So the same centuries-old problem is still with us today in modern physics.

Opposites are no longer the formidable opponents they once seemed. Instead, they become more like the two sides of a coin, lying with only one side up. To see heads or tails is to decide which is explicit. The other is always there, implicitly. But notice also that the "two sides" require a 3-dimensional object, and not just the 2-dimensional "head" or the 2-dimensional "tail". This juxtaposition of "identical" opposites is strongly met with in modern physics. For example, quoting Lee[33] :

---

[33] T. D. Lee, "Space Inversion, Time Reversal and Particle-Antiparticle Conjugation," *Physics Today*, 19(3), Mar. 1966, p. 23. Positive charge is really *observation* imposed on negative charge moving backwards in time, and the positron is really observation imposed upon an electron traveling backwards in time, etc.

> *"As we expand our observation, we extend our concepts. Thus the simple symmetries that once seemed self-evident are no longer taken for granted. Out of studies of different kinds of interactions we are learning that symmetry in nature is some complex mixture of changing plus into minus, running time backward and turning things inside out."*

Any addition to the Aristotelian laws of logic is actually a higher dimensional form of it. Three-law Aristotelian logic is in fact fitted to primitive observation — the result or output of single-photon interaction and observation processes.

This "four-law" logic (or "five-law" logic if one includes the master application rule as another "law") has proven very useful over the years, particularly in bioenergetics applications. It has also been most useful in attempting to decipher that branch of Russian energetics weapon science called "psychoenergetics", and in attempting to decipher the mind-body coupling mechanism, but that is beyond the scope of this chapter {79}.[34]

1.4.5 Substitution of Effect for Cause

Refer to Figure 1-3 again, to see the relationship of the unobserved cause, the interaction of cause and a previous effect, and the production of the resulting "new effect" as a change to the old effect or the output of a second replica of it. The greatest non sequitur in electrodynamics — and in parts of physics such as mechanics — is the widespread substitution of the *effect* for the *cause,* as a result of considering an observable to *persist in time* without any physical interaction. We discuss that more fully in Chapter 2, particularly with respect the notion of the same EM field existing both as a component of a material environment and as a component of a nonmaterial environment, but with the "force reducing to zero" in the latter. Jackson {80} points out the prevailing view of electrodynamicists as follows:

> *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable*

---

[34] The interested reader is referred to my website, www.cheniere.org, for papers dealing with the nature of the mind-body coupling mechanism and the body-mind coupling mechanism, as well as direct engineering of the mind, and limited information on the weapons referred to.

> *quantities such as force somehow involve the product of charge and field."*

On the other hand, Bunge {81} very clearly stated that the standard electrodynamics and physics itself are in terrible shape. Here is a direct quotation:

> *"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."*

We shall try to further clarify these unresolved difficulties with the field concept in Chapter 2.

1.5 Electrodynamics is Still Developing

1.5.1 Statement by Penrose

Roger Penrose, one of the leading physicists of our time, pointed out the still-developing nature of electrodynamics with respect to the developments from conventional gauge-theoretic interpretation of Maxwell's equations. Quoting Penrose {82}:

> *"These facts should not, however, deter theoretical or experimental physicists from seeking alternative descriptions, unconventional formulations, surprising electromagnetic effects, or radical generalizations. The various articles in this book provide the reader with a great variety of different kinds of approach to developments of this nature. We have historically motivated accounts, suggestions for new experiments, unconventional viewpoints and attempts at generalizations. We also see novel and ingenious formulations of electromagnetic theory of various different kinds... I am sure that this book will make it clear that electromagnetism is a subject that is in no way closed to stimulating new developments. It is very much alive as a source of fruitful new ideas."*

1.5.2 An Example of a Ubiquitous Error

One of the rather horrid "bad examples" of ubiquitous errors in electrodynamics is the conventional illustration of a so-called planar EM wavefront moving through space, as shown in Figure 1-4.

A succinct evaluation of that diagram and notion is given by Evans {83}, one of the great electrodynamic theoreticians of our times:

> *"I broadly agree ... that the transverse plane wave view of vacuum electrodynamics is the biggest blunder of twentieth century science."*

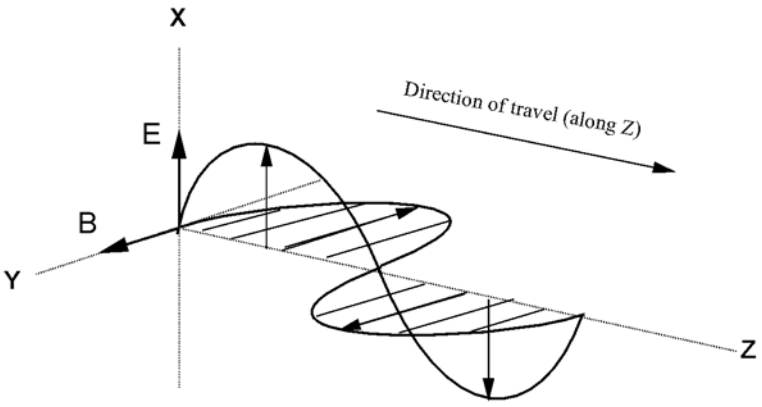

Figure 1-4 The erroneous but ubiquitous supposed illustration of the transverse EM wave in 3-space.

Dr. Robert H. Romer, former Editor of the *American Journal of Physics*, also chastised the horrid diagram shown in Figure 1-4, purporting to illustrate the transverse plane wave traveling through 3-space. In endnote 24 of his noteworthy editorial, Dr. Romer {84} takes that diagram to task as follows:

> *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "...perhaps then, for historical interest, [we*

> *should] find out how that diagram came to contaminate our literature in the first place."*

Indeed, each point along the z-axis, is a momentary "3-space frozen slice" of the ∈/∈*t* operator being applied to the ongoing 4-space interaction of the causal field in 4-space interacting with the previous "observed effect" (in this case, the previous 3-space slice) to produce a change to it. So the entire diagram is a set of "frozen 3-space observations", in which nothing at all is moving, and where the fields shown are in the mass (detecting/observing) medium. There is no such thing as a "traveling EM wave in 3-space", because "moving" or "changing" or "traveling" *a priori* must involve time as well as 3-space. What we are looking at in Figure 1-4 is the iterative and very rapid results of continual "3-space observations" (frozen 3-snapshots) of an on-going 4-space interaction.

1.5.3 An Extremely Important EM Omission

There is an enormous amount of development yet to be done in electrodynamics! As a simple example, for nearly a century there has been waiting quietly a fundamental "internal" or "infolded" electrodynamics by Whittaker {85, 91a, 86}, based on more primitive initial work by Stoney {87}, existing inside — and *comprising* — all conventional EM potentials, fields, and waves. This internal electrodynamics {88} has been essentially ignored in the West, but has already been weaponized in the weapons laboratories of several nations under the aegis of *energetics*. The U.S. Secretary of Defense, Secretary Cohen {89}, alluded directly to some of these weapons in a speech in 1997 as follows:

> *"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our [counterterrorism] efforts."*

The weapons are already being utilized against various nations in the form of terrorist acts in an undeclared war {89}, and have been so used since the 1950s.

This internal longitudinal-wave EM and its dynamics is a much more fundamental electrodynamics (actually a subset of a unified field theory) consisting of correlated longitudinal EM wavepairs in both the 3-space and time domains. Each Whittaker decomposition wavepair is comprised of an incoming longitudinal wave — a "time-polarized" EM wave {90a-90c} — in the time domain, perfectly correlated with an emitted longitudinal EM wave in 3-space. See Figure 1-5. This multiwave structuring and dynamics in both the time domain and in 3-space simultaneously *comprises* the scalar potential.

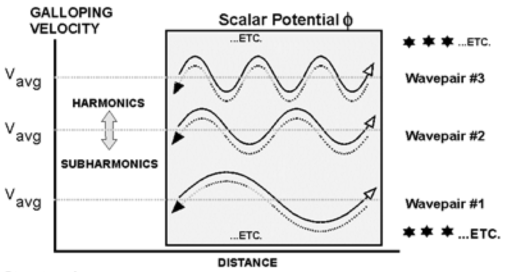

The Structure Is:

- A harmonic set of longitudinal phase conjugate wavepairs.
- In each wavepair the two waves superpose spatially after detection, but travel in opposite directions.
- Prior to detection, the convergent wave set is in the imaginary plane, and hence is not observable. It is EM energy incoming to the potential (dipolarity) from the time domain.
- The charges receive the complex convergent EM energy, transduce it into real EM energy, and emit enormous energy at the speed of light in all directions -- which includes bidirectional pairs in 3-space (after the reaction, being after "observation").
- This produces the fields and potentials from the "source charge or dipole."
- Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus quantum field theory work strongly supports the "negative resistor" interpretation of the scalar potential and our solution to the source charge and source dipole problem.

Figure 1-5 The scalar potential is a harmonic set of phase conjugate longitudinal EM wavepairs.

We specifically stress that the dynamics occur in the *time* domain (in its specific structuring, propagation pace, etc.) as well as in 3-space.

Deliberately changing the internal longitudinal EM wave structuring produces precise energetics (both energy and dynamics) in spacetime itself. This is a special form of "spacetime curvature structure and dynamics" which we call a *spacetime curvature engine* or *vacuum engine*. Together with superpotential theory in the original form initiated by Whittaker {91a} and extended by others {91b, 91c, 91d}, this internal electrodynamics — together with scalar potential interferometry {92} — creates all ordinary EM fields and waves {79}. Superluminal communication is possible using the internal longitudinal EM wave structure of the EM fields, potentials, and waves {93a, 93b}.

1.6 A Summarizing Letter

Our thrust in this book is to attempt to (1) reveal some of the major flaws in electrodynamics that have prevented development and use of COP>1.0 electrical power systems, and (2) point the way toward the development and use of such systems. To give a general summary of the ground this book will cover, we conclude with an informal E-mail letter I sent to the recent editor of *American Journal of Physics*, pointing out the greater implications of his cogent observation of that "dreadful diagram" and where it leads. Indeed, it (i) leads to a dramatic change in electrodynamics, (ii) changes the way in which we regard "propagation of EM energy through 3-space", (iii) solves the long-vexing problem of the association of the source charge or source dipole with its fields and potentials and their energy, (iv) changes dramatically the way we view what powers an electromagnetic circuit or power line, (v) allows extraction and use of copious EM energy from the vacuum, and (vi) solves the energy crisis permanently.

Here is the letter sent informally to Dr. Robert H. Romer — *slightly edited* to improve the grammar, correct one error of misstatement, and with reference citations added and listed at the end of this chapter:

To: Dr. Robert H. Romer
American Journal of Physics
Amherst College, Box 2262
Amherst, MA 01002

Thursday, Mar. 8, 2001

Personal communication

Subject: Implications of your cogent comments on that atrocious illustration

Dear Dr. Romer:

This is not a submission of a manuscript, but a personal communication to you on a matter of great importance in physics, directly related to endnote #24 of your seminal editorial {94}.

It will take a little exposition, so please bear with me and read this when you have the spare time to do so. We are going to show you how your keen insight can be extended to solve some extraordinarily formidable foundations problems in present electrodynamics. We will also show how to apply the implications of your insight to totally solve the present electrical energy crisis permanently.

In your endnote #24, {94} you took to task (quoting):

> *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "...perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

Dr. Romer, you have lifted the corner of a dark cover concealing one of the most important flaws in electrodynamics and in fact in all of physics: the unwitting and pervasive substitution of the effect for the cause. A marvelous extension to the present physics is enabled if one removes this terrible non sequitur in physics, and particularly in electrodynamics. For openers, one solves what has been called the most difficult problem in electrodynamics (the problem of the source charge and the association of its fields and potentials and their energy) {100}. I will solve that problem for you in this informal write-up. I will also explain how to extract enormous EM energy from the vacuum, anywhere and anytime, easily. Extracting it is easy; catching it and using it to power loads without killing the extraction process is another matter.

One also gets a unified field theory, engineerable by novel electrodynamic means, as is steadily being shown by a series of rigorous Alpha Foundation's Institute for Advanced Study (AIAS) papers published in various leading journals (and more than 90 of them carried on a Department of Energy website restricted primarily to DoE scientists). Dr. Myron Evans, Director of the AIAS, has over 600 papers published in the literature, including such journals as *Physical Review*, *Foundations of Physics*, *Physica Scripta*, etc. Many of the other AIAS co-authors are excellent theoreticians and scientists.

The early pioneers of electrodynamics (Maxwell etc.) all assumed a material ether filling all space {95}. To these early scientists, there was not a single point in the entire universe that was devoid of matter, because the ether was present there. Hence their outlook as to the nature of EM fields etc. was quite material. Faraday conceived his "lines of force" as physical, taut strings, so that perturbations were "plucking these taut strings".

Maxwell himself points out in his famous "Treatise" that he specifically captured the thinking of Faraday in his theory. In fact, Maxwell wrote a *material fluid flow dynamics* theory.

In the light of more modern knowledge, let us see the impact of these and similar early *but still retained* erroneous electrodynamics assumptions.

First, observation/detection is totally spatial, as is well known in quantum mechanics. In fact, observation is a $\in/\in t$ operator imposed upon 4-space (LLLT spacetime), yielding a frozen instantaneous snapshot LLL of an ongoing 4-space dynamic process. At the next instant, that particular *previous* observation no longer persists. Why?

Well, no observable *persists*, since it is only an instant frozen 3-space snapshot, at a single point in time, *a priori.* Here again we have another horrendous non sequitur in all of physics: the assumption that observables "continue to exist" and therefore persist in time in a passive manner. In fact, there is an interactive process that generates their (seeming) persistence, involves time, and continually changes mass into masstime and back to mass, etc. With the reader's permission, we will pass discussing that mechanism until another time (pun intended!).

What we conceive as "an observable such as mass, traveling through space and persisting in time while doing so", is actually an iterative, continual series of these frozen 3-space snapshots or observations, much like the frames of a motion picture film. We ourselves mentally add the "continuity" to provide "the sensed motion", but rigorously what is actually "observed" is not continuous, but is a vast continual series of those frozen 3-space snapshots.

We're getting directly at that atrocious diagram!

Each snapshot is an *effect*, not a cause, because it was the output of the observation process whereby a 4-space *causal* entity (non observed *a priori*) interacts with a previously observed frozen entity (say, a unit point charge at some point in space) to produce the observation (change or effect generated in that interacting observed charge) as the "next instantaneous observation".

The usual "representation" of a "3-space EM wave" propagating in 3-space is indeed atrocious, just as you stated! It is actually just an iterative succession of such instantaneously frozen snapshots in 3-space, one after the other. There is no such thing as that set of snapshots independently existing in spacetime, *prior to interaction with charge* in that series of interactions and observations, unless we wish to discard quantum mechanics and the laws of logic.

There is, however, a continual iterative stream of those observations — those frozen 3-space snapshots — that we interpret (erroneously) by recall from memory as the "EM wavefront propagating in 3-space". As you eloquently pointed out, that is not so, and it is atrocious. An observation, being an absolutely frozen entity, cannot "move through time" anyway, since it cannot persist, nor can it move. A change to an observable can only be another observable snapshot of the ongoing 4-space entity and action, which is then compared to the first snapshot and a difference noted {96} [See Figure 1-3].

In short, a great stream of "frozen effects" (frozen instant observations) does not constitute a "picture" of the ongoing 4-space action, but only a series of frozen 3-space intersections involving the interaction of that fixed observed (3-space) charge with the ongoing causal 4-space entity. If we add and integrate a series of 3-space pieces, we *do not* get a 4-space entity! Instead, we get a longer or bigger 3-space slice/piece, but one for which each piece of it only existed at a single point in time as a 3-space "slice" at that moment. That is precisely what is wrong with that horrible illustration.

But it is also "wrong" with electrodynamics itself! Electrodynamicists mistakenly conclude that the same *effect* "series of static 3-slices" — which they might call, e.g., the "field" in a case where the field is the subject — is the same as the 4-space continuous *causal* field prior to observation interaction with charge at all. Well, LLLT is definitely not LLL, nor is it n(LLL) where n is some large but finite number of 3-slices LLL.

So the field concept is dichotomously used in two contradictory manners in electrodynamics:

(1) it is considered to be in 4-space prior to the observing/detecting interaction with the observable (such as a unit point 3-space charge), and it is also considered a 3-space entity *after* that interaction. The dimensions of the two entities are not the same, and neither are the dynamics. The causal EM wave is dynamic and 4-spatial, the effect "3-space EM wave" (ugh!) is static and 3-spatial at each instant it is "observed". Assuming that the two are the same thing is a non sequitur.

In fact, it substitutes the *effect* for the *cause*, a rather gross violation of the causality principle itself.

(2) then an attempt at glossing over the illogic is used by the smooth statement that "the field (meaning that atrocious series of frozen 3-space snapshots) remains in the absence of charge, but the force goes to zero." Well, if the field is defined as a force field, it cannot be a "non-force field in the same observation!" Else, opposites are always identical.

So of what importance is all that?

It is of great importance. Let me show you one very startling thing that comes out of correcting this "biggest foundations non sequitur in physics, that of substituting the effect for the cause".

Consider a very special paper by E. T. Whittaker {97} in 1903. [I can send you a pdf file of the paper if you do not have it and are interested]. In this much-neglected paper, Whittaker decomposed the so-called "static" potential into a harmonic set of bidirectional longitudinal EM wavepairs, where each pair consists of a longitudinal 3-space wave (an *as-observed* wave) and its longitudinal phase conjugate wave (*considered unwittingly as having also interacted with charge, and therefore as being shifted into 3-space as an "as-observed" wave with inverse parity*).

Whittaker — as has everyone since him — unwittingly assumed the "iterative continual observation" interaction in there for the phase conjugate wave also. In so doing, he came up with two *effect* "waves" that are the outputs of the assumed observation process. Neither of these effect waves would be a wave in spacetime {98} at all, but the two in ensemble are an example of the same thing you objected to in your cogent commentary on that abominable "illustration".

In short, Whittaker invoked *observation* as a process with two effects and no cause, rather than with a cause and an effect, with the interaction with the observing/interacting unit point charge being assumed for both waves. Whittaker and everyone since seem to have made the same error in interpreting that seminal Whittaker decomposition. This misinterpretation of the decomposition has until now hidden one of the greatest secrets of all times in Nature's electrodynamics!

Let us correct the interpretation, and uproot that great secret to the light of day.

First, for observation to occur, one must have a cause acting upon the affected (observable) entity, and an effect (observable change) must be produced in, on, or of that affected (interacting) entity. One *must not* have two effects (two observables) and the affected entity (another observable and therefore another effect)! Again, assuming that one has three effects (observables) constituting the observation process is a logical non sequitur of first rank.

Let us now correct that logical non sequitur made by Whittaker and other physicists and electrodynamicists, and see where it leads us.

First, we go to particle physics, where broken symmetry {99} was discovered in the 1950s. Lee, e.g., received a Nobel Prize for his work in that area {75}. Lee also showed that any dipole is a broken 3-symmetry in its violent energy exchange with the active vacuum.

Well, a scalar potential is a dipolarity; always a potential is actually a difference between two potentials, so to speak. So the potential itself represents a broken 3-symmetry in an energy flow exchange with the active vacuum.

Let us further examine that interesting broken symmetry aspect. It means that the "static" potential is a process whereby energy is received from the vacuum in *one* form, not observable, and hence unusable, but is output in *observable* (usable) form. In short, the dipolarity or dipole receives and absorbs (QM view) virtual photons, integrates them into observable magnitude, and emits real, observable EM energy ("continual observation" snapshots of the latter being assumed).

In physics, all observation is 3-spatial, as is well known. And 3-space is the realm of the observed. The EM energy from the vacuum is not received in 3-spatial (observable) form, else there would be no broken 3-symmetry of the dipolarity.

We can also *experimentally* verify that there is no 3-space input of EM energy to the potential — e.g., to the potential between the ends of any source dipole, because we cannot measure any 3-space observable energy feeding the charges of the dipole. Instead, observable 3-space energy is continuously pouring out of the dipole.

Let us continue now with the notion of a real dipole of separated source charges, with our "scalar potential" between its ends, so we have something concrete in mind.

First, our instruments prove there is a continual emission of EM energy in all directions in 3-space (of the kind in that atrocious diagram; a "series of iterative film-snapshots"). That is the way it is conventionally represented, as if observed at each and every point successively in that 3-space, and along every radial.

But secondly, our instruments also prove there is no such observable *input* of EM energy in that same 3-space *to* the dipole. Bummer! That is precisely what has stopped electrodynamicists from solving that vexing problem of where all the energy pouring out of the source dipole is coming from, and how! They unconsciously assume that the only possible source of the EM energy input is in 3-space. That is totally wrong. Further, the conservation of energy law does not require that energy be conserved in 3-space! Instead, rigorously — if we are using a 4-space (i.e., spacetime) model — it requires that energy be conserved in 4-space. The assumption of the extra condition of 3-space energy conservation is an added and arbitrary extra condition.

Since electrodynamics has not resolved this "source dipole and its associated fields and potentials" problem, it has stopped work on the problem and implied that every dipole in the universe is a perpetual motion machine, freely creating all that EM energy it continuously pours out across space, in all directions, creating its associated fields and potentials. That implicit assumption, if true, of course destroys the conservation of energy law.

It isn't true, and the energy conservation law is alive and well!

We can experimentally prove that the source dipole *does* continuously pour out energy in all directions in 3-space, without ceasing, as follows:

In a gedanken experiment, we set instruments every 300 million meters or so, along a radial line from an origin in the lab. With the instruments and clocks synchronized, we suddenly form a dipole at the origin. One second later, the first instrument reads. A second later, the second instrument reads. And so on. But it is not a "passing pulse". Whatever reading the instrument makes as the forward edge of the energy flow reaches it, is then continuously maintained thereafter. This proves that the energy is poured out continuously and at the speed of light, and in any (and all) directions in 3-space, *and it continues to pour out at exactly the same rate so long as that dipole remains intact.*

Dipoles in the original matter of the universe have been pouring out EM energy in that fashion for some 14 billion years, and they have not "exhausted" their unobserved energy input source yet.

A newly formed simple dipole, e.g., in one year will have poured out energy into a spherical volume of space that is a light-year in radius. Wait another year, and that volume of space whose energy density has been changed will be two light years in radius. The dipoles in the original matter have filled the universe itself with that energy outpouring from them.

In other words, an incredible amount of EM energy has been and is being poured out into space from every dipole in the universe.[35] And electrodynamicists have had not the foggiest notion of where that mind-boggling amount of outpoured EM energy has come from.

Unless we wish to totally discard the conservation of energy law, we must have an equal input of energy *from outside 3-space*, going into every dipole continuously. In 4-space, that only leaves the fourth axis, along which and from which the input energy must move into the dipole.

And so it does. If we re-interpret that phase conjugate half set of the Whittaker decomposition, *before* observation has occurred, it then is a harmonic set of longitudinal EM waves moving in the time-dimension, *into* the source dipole (parity is not inversed because no interaction with charge has occurred to shift it into 3-space.

So voila! We have strangely (but quite rigorously) solved what Sen {100} referred to in this manner: *"The connection between the field and its source has always been and still is the most difficult problem in classical and quantum electrodynamics."*

We express our reasoning as follows: (i) we know from particle physics that the source dipole[36] is a broken 3-symmetry in its flux exchange with the vacuum. This means that we know the energy is received from the vacuum in an *unobservable* form, absorbed by the dipole, and emitted as *observable* EM energy. It remains to translate that into classical electrodynamics rather than quantum physics {101}. (ii) The *time* domain is the only domain outside 3-space, in the standard 4-space model. (iii) Anything in the time domain exclusively, is indeed nonobservable, since the $\in/\in t$ observation operator destroys time and all its internal structure and dynamics whenever observation occurs. (iv) Looking for a "3-space" input is looking for an "observable" EM energy input, which would in fact disagree with the known broken 3-symmetry of the source dipole.

---

[35] Later we will see that, in a time-forward situation, the negative charge pours out positive EM energy while the positive charge may be said to pour out negative EM energy. Charge conservation then implies that an energy balance is maintained overall.

[36] Because of the broken symmetry of the opposite charges on its opposing ends.

Let me digress now to speak of fundamental units. As is well known, the fundamental units one chooses to make his physics model represent an arbitrary choice. One can, if one wishes, make a perfectly valid (but mentally nightmarish!) physics using only a single fundamental unit. In that case, all other "fundamental units" in the present model become functions of that single fundamental unit.

Suppose, e.g., that we make the joule our only fundamental unit. Then "mass" becomes totally a function of energy — and we have no heartburn with that one since the Einstein revolution and the nuclear age. But then "time" also becomes totally a function of energy — and that surprises us, because we have unconsciously been taught (erroneously) that "time is a flowing river down which a mass drifts like a drifting boat". That is not the nature of time at all; the totality of the photon interactions with a mass create that mass's "motion through time". I can later explain that to you also, if you wish, but let us pass it for now.

It turns out that time is spatial energy compacted by the factor c^2, so it has the same energy density as mass, but in the time axis instead of 3-space. Intuitively, if we are interested in cause and effect, the energy of the cause (time) should be equal to the energy of the effect (in this case, mass). After one reflects a moment, one also sees that "time energy" is required to "drive" a mass through time, just as "spatial" energy is required to "drive" a mass through space. The notion that mass in an inertial frame moves through "empty space with no reaction" is false; it continually moves through the energetics of spacetime, and interacts continuously with it.

Anyway, from still another viewpoint there is no magic in EM energy currents moving in the time domain! Quantum field theory already recognizes multiple polarizations of the photon, including transverse, longitudinal, and time-polarized. Thus, it implies a time-polarized EM wave also. A longitudinal EM wave *moving in the time domain* oscillates along its line of travel. That is indeed oscillating its energy density in the time domain, so that it is "time-polarized". Hence it is a proper time-polarized EM wave, and appears to "gallop" {102} or vary its speed periodically.

The solution to the "source dipole" problem is that the phase conjugate wave half-set of the Whittaker decomposition, when reinterpreted, is the *incoming* EM energy in the time domain, continuously input to the charges of the dipole. The charges interact in the imaginary plane (the time domain), and absorb the time-energy, then transduce it into 3-space, and emit it as Whittaker's set of real observable longitudinal EM waves in all directions in 3-space (as observed). The time-energy waves that are input to the dipole cannot be observed, since observation destroys time and its constituents.

All 3-spatial EM energy comes from the time domain! Broken 3-symmetry of the source dipole immediately releases our arbitrary additional restriction on nature's energy conservation law — i.e., our insistence that the energy input for conservation must be input in 3-space. By removing these arbitrary "shackles" from nature's feet, we allow nature to joyously resume her much-preferred special 4-symmetry in EM energy flow: the circulation of EM energy from the time domain into 3-space, and outpouring of that energy in all directions in 3-space, at the speed of light {103}.

Now let us solve the source-charge problem as well. That one is now easy to resolve.

We know from quantum mechanics that any "isolated" observable charge is actually clustered around by virtual charges of opposite sign. So we simply take one of these clustering opposite charges while it exists, and a differential piece of the observable charge, and we have a composite dipole. Hence the "isolated charge" may be represented as a grouping of such momentary dipoles, each (while it exists) with a scalar potential between its poles, and hence each subject to the same decomposition and reinterpretation we have done.

That is why the source charge can "sit there" and pour out EM energy in 3-space (as observed) continuously, and indefinitely. It is also continuously absorbing EM energy from the time domain. As a set of composite dipoles, it is a set of broken 3-symmetries in EM energy flow. Hence it exhibits nature's preferred 4-symmetry in energy flow, between the time-domain and 3-space.

Note also that, to the 3-space observer, this 4-symmetry is purely negentropic. It is a continuous and ongoing (and expanding) reordering of the vacuum energy, in the form of the reinterpreted Whittaker decomposition.

We can easily engineer negentropy! Just make a little dipole, and nature happily starts pouring out energy and reordering part of the vacuum, with that reordering spreading at the speed of light, steadily increasing all the while.

Reinterpreting Whittaker's marvelous decomposition tells us many things: (i) the source charge or dipole does indeed continuously receive and absorb its input energy from the time domain, (ii) it continuously transduces its absorbed energy to real 3-space energy (as observed!), and (iii) it continually outputs its absorbed and transduced time-energy as 3-space energy emitted in all directions in 3-space (as continually observed).

So the terrible foundations problem of the source charge and source dipole is resolved, and the energy conservation law is maintained.

We have also uncovered a most important thing: Given a little broken 3-symmetry in EM energy flow, t-symmetry in EM energy flow is also broken. A more primary 4-symmetry {104} between time-energy flow and 3-space energy flow emerges *automatically*. In short, every potential, every dipolarity, and every charge is such a broken 3-symmetry and an example of the new, preferred 4-symmetry of EM flow, with input flow in the time-domain (nonobservable!) and output flow (as continually observed in a series of 3-snapshots) in all directions in 3-space.

This also resolves the logical cause and effect problem for observation.

This 4-symmetry is a purely negentropic process, once the dipole is established! Well, we should have known that from the gauge freedom axiom in gauge field theory anyway! That axiom says that we can change the potentials anytime, freely and at will. In electrodynamics, that means we can change the potentials at will, anytime we want to. But that means *we can freely change the potential energy of a Maxwellian system at will, whenever we wish*. It costs nothing — at least in theory; in the real world we have to pay a little for switching — to suddenly potentialize an EM system, prior to the time the Drude electrons relax and current flow begins.

I have applied this great new 4-symmetry in EM energy flow, that is freely evoked and persists indefinitely after one pays a little to make the dipole and thus produce a little broken 3-symmetry, to produce EM energy freely from the vacuum. Without going into it, I refer you to our forthcoming papers {105}. Indeed, we can use this negentropic process to extract as much EM energy from the vacuum as we wish, anywhere, anytime, for peanuts. Let us now move to discuss that area.

First, we point out another astounding result that comes from resolving this "source charge and source dipole" foundations problem.

Generators do not use any of the shaft energy input to them — even in transduced form — to power their external circuits! A generator itself does not add a single watt to the power line, and neither does a battery's chemical energy dissipation add a single watt to its attached circuit.

Neither does burning all those hydrocarbons, consuming all those nuclear fuel rods, building all those dams, etc. to provide our conventional power systems and grid. None of that enormous destruction of the biosphere and pollution of it has ever directly produced one watt on the electrical power line.

Just follow the energy flow, in light of what we now know. Here's how it works.

Typically, we burn some fuel to boil water and make steam, and use the steam to power a steam turbine, which forcibly rotates the shaft of the generator, thereby inputting mechanical shaft energy into the generator. So far, so good. That took care of all the hydrocarbon burning and fuel rod consuming, extracting and transporting the oil, etc.

This input of the shaft energy forcibly rotates the rotor against internal resistance, forming an internal magnetic field. Assuming a 100% efficient generator with no internal losses whatsoever, this means that the *mechanical shaft energy input* has now been totally transduced into *internal magnetic field energy*.

So what does that magnetic field energy *do*? It is totally dissipated upon the internal charges of the generator, performing work on them and forcing the positive charges in one direction and the negative charges in the other direction. That dissipation of the energy in the internal magnetic field *forms a source dipole inside the generator, connected to the terminals*.

And that is all the generator does. Period. *None* of the energy transduced from that shaft input, went roaring out of the terminals and down through space outside the conductors of the power line. Not a single watt. So let us turn to particle physics to find out what happens next, because *it does not yet appear in the electrodynamics model, even though proven.*

The internal source dipole, *once formed*, is a great broken symmetry in the vacuum flux, as we discussed and as is well known in particle physics. But the proven and well-known vacuum interaction with the generator and the dipole charges is not even modeled in the classical EM theory used by the scientists and engineers to build electrical power systems — much less a *broken symmetry* in that active exchange! That's atrocious, since it's been proven in particle physics for nearly a half century, Nobel prizes awarded, etc. But the hoary old 137-year old Maxwell-Heaviside model, further curtailed by Lorentz symmetrical regauging, does not incorporate what has already been proven in physics. *It does not accurately model the situation as it is and as it is known to be.*

Let us continue.

Once that dipole is formed, it extracts *enormous* EM energy from the vacuum, and pours it out of the terminals of the generator, filling space surrounding those conductors of the attached external power line. It is a huge EM energy flow, trillions of times larger than what we account after Lorentz taught us to disregard almost all of it.[37]

Only the tiny little bit of that EM energy flow in space along and surrounding the conductors — the little "boundary layer" that skims down the surface of the conductors — will strike the surface charges in the conductors and get diverged into them to potentialize the Drude electrons and "power the power line and its circuits and loads". *All the rest of the giant EM energy flow in space surrounding the conductors, and generally parallel to them, misses that power line entirely and is just wasted.*

Check the original papers by Poynting {4a, 4b} and by Heaviside {5a, 5b, 5c}, who independently discovered the flow of EM energy through space (as if continually observed!) in the 1880s, after Maxwell was already dead. I can furnish the citations required. From the beginning, Poynting only considered that component of the energy flow that actually enters the circuit. He considered only the "boundary layer" right on the conductor surfaces, so to speak.

Heaviside considered that component that enters the circuit, and also uncovered and recognized the *gigantic* component in the surrounding space that *does not* enter the circuit but misses it entirely and is wasted. [Added] Here is Heaviside's {106} own statement:

> *"It [the energy transfer flow] takes place, in the vicinity of the wire, very nearly parallel to it, with a slight slope towards the wire… . Prof. Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from*

---

[37] However, contrary to what is in the textbooks, **E** Δ **H** energy flow is not in observable EM field form **E** and **H** until it interacts with charge. Instead, the flow is in the form of virtual energy currents in the vacuum. We point out that the fields **E** and **H** utilized in the Poynting energy flow theory are effects and not causes. They are rigorously defined as **E** and **H** only after the causal 4-fields have interacted with charge. E.g., **E** is "defined" as force per unit point massive charge — i.e., as the force created by interaction of the "vacuum field" (curvature of spacetime) with a unit point charge. No unit point charge interaction, no force or force field intensity per interacting unit point charge.

> *the vertical. This difference of a quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature of the electric field in the vicinity of a wire supporting electric current. The lines of electric force are nearly perpendicular to the wire. The departure from perpendicularity is usually so small that I have sometimes spoken of them as being perpendicular to it, as they practically are, before I recognized the great physical importance of the slight departure. It causes the convergence of energy into the wire."* [End of Added].

But Heaviside had absolutely no explanation for the enormous and startling magnitude of this energy flow that "misses the surface charges of the conductors and is wasted". You can see an elementary illustration of the "point intensity" of this Poynting diverged energy flow component in Kraus {107}. Kraus's figure 12-59, p. 576 shows a good drawing of the *Poynting* component *being withdrawn* from the total EM energy flow filling all space around the conductors [see our Figure 1-1 in the present chapter].[38] Most of that available energy flow is not intercepted and thus not diverged into the circuit to power it, but just "wasted." The *remaining* huge component discovered by Heaviside is not shown on Kraus's diagram. Kraus's numbers on the contours represent the amount of power (watts per sq meter) being withdrawn from each contour, by the very limited axial movement of the electrons and the very ends of their associated fields into the wire.[39]

---

[38] The energy flow in space near the surface of the wires strikes the surface electrons and their fields, potentializing them and creating a force field with respect to inner electrons. This drives the surface electrons axially into the wire, since they can only move down the wire with the drift velocity, nominally on the order of a few inches per hour. The short inwards drawing of those small portions of the fields and of the increased potentials attached to the charges that move into the wires, interacts with the internal charges in the interior of the wire, powering the Drude electrons throughout the conductor and the circuit. For a discussion of the importance of the surface charges, see J. D. Jackson, "Surface charges on circuit wires and resistors play three roles," *Am. J. Phys.*, 64(7), July 1996, p. 855-870.

[39] Obviously the increased fields on the withdrawing surface charges reach toward infinity radially outward from the wire. Thus most of this increased field on a withdrawing electron remains outside the wire, radially reaching toward infinity, since the electron and its field cannot be withdrawn further than the diameter of the wire. As the surface charges and their fields change in Jackson's discussion, obviously those distant parts of the field remain outside the conductors. Hence there remains an extensive field and field energy outside the conductors and the circuits

So Heaviside {106} spoke cautiously of the *angles* and *relative directions* of the flow components; he did not wish to be scientifically destroyed for pointing out such an inexplicably large EM energy flow, far larger than the known energy input to the generator. There was then no such thing yet discovered as the electron, the atom, the nucleus, special relativity, general relativity, quantum mechanics, the active vacuum, etc. Poynting {4a, 4b} never even thought of the "nondiverged component that misses the circuit", nor did he consider it.

Lorentz, however, understood the extra Heaviside component and its vast magnitude, but he also could not explain it in any fashion. Even the great Lorentz could not risk publishing or advocating such an enormous energy flow, lest he be called a "perpetual motion nut" and destroyed. So he reasoned that, since that stupendous nondiverged energy flow component misses the circuit and powers nothing, it has "no physical significance" (his words).

Jackson in his famous 1975 *Classical Electrodynamics*, 2nd edition, uses essentially the same phrase {108a} as did Lorentz. And so do most other electrodynamicists.

So Lorentz {109} originated a little trick of integrating the energy flow vector itself around a closed surface surrounding any volume element of interest. Obviously, that zeros all *nondiverging* EM energy flows — including precisely that inexplicable and enormous Heaviside "dark energy" component. The integration trick does retain that small, *diverged* Poynting component that enters the circuit. In addition, our instruments measure energy dissipation from the circuit, and the energy has to *enter* the circuit to *be dissipated* from it. So our instruments and their measurements will indeed agree with the Poynting energy flow component. Lorentz thus arbitrarily discarded accountability of trillions of times as much EM energy flow as was retained and accounted.

In a later book by Lorentz, one can see that little trick {109} that is still used by electrodynamicists {110}.

---

that is never drawn into the wire. Additional separate receiver circuits with resistive loads, adroitly placed in this "externally remaining" field energy flow, can be made to intercept additional EM energy and will perform additional work in those external loads in the secondary circuits, separate from powering the loads in the primary circuit.

The Lorentz trick does not cancel the *actual* flow of the Heaviside dark energy component around every circuit! It just drops it from any accountability.

Indeed, we do precisely a similar thing for the "field" and the "potential". There is not a single text in the U.S. that calculates the magnitude of the *field itself*, prior to point interaction. Instead, we are taught to calculate the reaction of that field at a point in it, with a unit point static charge at that point. In short, we calculate *what is actually locally diverged from the field or potential by that little point static unit charge*, and call it "the magnitude of the field". At best, it is indicative of the field intensity at a point, because we have prescribed the magnitude of the static point charge's *reaction cross section* with the field, *not* the overall magnitude of the entire field itself. Another major non sequitur! By identifying the "field" as "that which is diverged from it", we gravely err. We do the same for the potential, again using its reaction cross section for a unit point charge at a point in the potential.

There is hardly a living electrodynamicist, it seems, who has calculated the magnitude of the field itself, or of the potential itself! All calculate the respective reaction cross section (and the *static* reaction cross section at that!) and erroneously call that the "magnitude of the field" or the "magnitude of the potential". It is no such thing. No thing filling all space is identical to a little something diverted from one point in itself, else we discard all logic.

Now we can return to our source dipole and its extraction of an *enormous* energy flow from the vacuum, once we account for the long-neglected (for more than a century) Heaviside dark (nondiverged, unaccounted) energy flow component. That is the essence of one of my papers {111}.

The easiest thing in the world to do is to extract enormous usable EM energy flow from the vacuum, *from the time domain*. Every circuit and every electrical power system already does it, and every electrical power system is powered by vacuum energy, *not* by burning all that coal, oil, etc. None of that does anything to actually power the circuit. It only makes dipoles.

If you will check the characteristics of the ubiquitous closed current loop circuit, you will discover a diabolical thing: that closed current loop circuit forces all the spent (depotentialized) electrons from the ground return line back through the source dipole (back through the back emf). It is easily shown that precisely half the energy *collected* in the circuit from that feeble Poynting energy flow component is then used to perform work on those dipole charges and scatter them, thereby destroying the source dipole and abruptly shutting off all energy extraction from the vacuum. The other half of the collected Poynting energy is dissipated in the external loads and losses. That means that half the collected Poynting energy is used to kill the source dipole, and less than half is used to power the loads {112}.

Well, we must input at least as much energy (assuming a 100% efficient generator with no losses) to the shaft of the generator to *restore* the dipole, as was used to *destroy* it. This means we shall always have to input *more* energy to the generator shaft, than we get dissipated in the loads.

That is precisely what is responsible for our ubiquitous COP<1.0 systems. *COP* is Coefficient of Performance, and is the average load power output in the external circuit divided by the average shaft power we pay to input to the generator.

COP < 1.0 comes from (i) the ubiquitous usage of the closed-current-loop circuit, and (ii) the prevailing notion that electrical power systems far from equilibrium in their energetic exchange with the active vacuum are impossible to build. In fact, every electrical power system is *already* just such a COP > 1.0 system, as far as the energy flow out of the generator or battery — compared to the shaft input energy to the generator or battery — is concerned.

As is well-known in the thermodynamics of open systems far from equilibrium with their active environment (in this case, the active vacuum), such a system is permitted to: (1) self-organize, (2) self-oscillate or self-rotate, (3) output more energy than the operator himself must input to the system (the excess energy being freely received from the active environment, in this case the active vacuum), (4) power itself and its loads simultaneously (all the energy being freely received from the active environment, in this case the active vacuum), and (5) exhibit negentropy.

But by designing all our systems so that they use more of their collected energy from the vacuum to kill their source dipoles than they use to power their loads, we have foolishly wasted the planet's resources, vastly overcharged the consumer, artificially created a great energy crisis, fomented wars for precious oil and other energy resources, polluted the planet, enhanced global warming, and strangled species.

In short, we pay the power company to have a giant wrestling match inside its generators and lose! And we pay our electrical engineers to keep designing and building such asinine systems!

This sad state of affairs is what is now upon us as a great and increasing energy crisis, polluting and destroying the biosphere, etc. and threatening to eventually collapse the world economy.

It is astounding that, since the basis for the above has been in physics for nearly a century (Whittaker decomposition) and for nearly a half century (broken 3-symmetry of the source dipole, as well as the active vacuum), the hoary old classical EM model has not been updated to incorporate what has already been proven in particle physics. Such is inexplicable and unconscionable.

It is also astounding that no electrical engineer realizes that energy extracted from the vacuum powers every electrical power system, and few if any professors are aware of it either.

This is where your cogent realization of the terrible non sequitur in that atrocious "wave in 3-space" standard diagram leads.

Our AIAS (Alpha Foundation's Institute for Advanced Study) advanced theorists now have a series of papers published in the hard literature pointing out the feasibility of extracting EM energy from the vacuum {113a, 113b}. About 20 other AIAS papers along such lines or related have been submitted to the various journals and are in the referee process.

I just wanted to contact you informally and, for your personal information, show you how perceptive and correct your objection to that atrocious diagram is. You have lifted the corner of the veil on electrodynamics' confusion between effect and cause, with effect being widely used as the cause. The entire notion of a "separate force" acting on a "separate mass" in mechanics is also a non sequitur. If we define force as F ∑∈/∈*t*(mv), then we see that mass is a *component* of force! So here is another major and uncorrected non sequitur. This one was also largely responsible for the dichotomy of the field concept, where the "same field" is used in both a force (mass-containing) manner and a force-free (without mass) manner, as if the two were the same!

I very much enjoyed your editorial, and would urge you to publish additional material along that same vein. You are striking at the very heart of the problem, and every bit of insight and change in those terrible non sequiturs will result in enormous progress in electrodynamics and physics.

Sincerely,

Tom Bearden, Ph.D.

Lt. Col., U.S. Army (Retired)
Director, Association of Distinguished American Scientists
Fellow Emeritus, Alpha Foundation's Institute for Advanced Study

----

Chapter 2

Background and Concepts

"[Hypotheses made without realizing that they are being made] ...are what Poincare has called "unconscious" or "natural" hypotheses — a type which one hardly ever challenges, for it seems too unlikely that we could make progress without them. Nevertheless it should be the endeavor of the physicist always to drag them out into the light of day, so that it may be perfectly clear what we are actually doing." [Lindsay and Margenau] {114}

"Equilibrium states are the only ones that are capable of explicit analysis in thermodynamics... " [Lindsay and Margenau] {115}

"Non-equilibrium conditions cannot be specified by variables of state, and their entropy cannot be computed. ...the condition of equilibrium is the condition of maximum entropy. " [Lindsay and Margenau] {116}

2.1 Comments on the Foundations of Classical Electrodynamics

2.1.1 Faraday, Maxwell, Quaternions, and Heaviside

For our purposes, we start with the important experiments of Faraday {117}, which produced experimental integration of electricity and magnetics {118}. Faraday's own work contained very little mathematics, and so he himself did not capture his own seminal experimental work in adequate theoretical fashion. That task was undertaken by James Clerk Maxwell {119a, 119b}, who deliberately set about to capture Faraday's results in a proper mathematical theory, using quaternion-like and vector expressions. Of the quaternion algebra, Maxwell had this to say {120}:

> *"...the virtue of the 4nions lies not so much as yet in solving hard questions as in enabling us to see the meaning of the question and its solutions. "*

Quaternion algebra {121a, 121b} is an advanced algebra created by William Rowan Hamilton, brilliant Irish mathematician. Hamilton was widely regarded as a "new Newton" by his peers. Hamilton coined the word "vector", wrote two books on quaternions, and hoped the books would find extensive use in physics. Vectors and tensors later came to suffice for most problems of that day, and so quaternions faded to the background.

Hamilton also grasped the importance of a science of time, and realized that such a science was possible. E.g., in 1837 Hamilton stated:

> *"The notion of time may he unfolded into an independent pure science... a science of pure time is possible."*

In that statement, Hamilton anticipated that a physics model can be comprised using only a single fundamental unit, and that unit could be the second.

In the early 1890s, quaternions in electrodynamics were discarded in a short "debate", mostly in the journal *Nature,* and vector electromagnetics was adopted. So Maxwell's 20 quaternion equations {20} in some 20 unknowns contain a wealth of electrodynamics {122} that is not included in Heaviside's {123} later vector curtailment after Maxwell's death, or in the modern tensor formulation. As an example, Nikola Tesla's patented circuits cannot be properly understood in either a vector EM or tensor EM analysis. However, very novel functions performed in them can be seen in a quaternion EM analysis, as has been clearly shown by Barrett {124}. Any electrodynamicist, who has looked at Tesla's work only in terms of tensor or vector classical electrodynamics, has no concept of what Tesla was actually doing. Since that is precisely how most academics have examined Tesla's work, they have had little idea of the functions he was actually able to accomplish in circuits — functions that U.S. electrical engineering departments still cannot perform.

Along with other electrodynamicists at the time, Faraday and Maxwell believed in a material ether — a thin material fluid filling all space. Hence there was thought to be no point in the universe that was devoid of mass. For that reason, the EM field conceived in mass and the EM field conceived in "space" — actually, in the "luminiferous ether" thought to fill all space — were identically force fields and material entities, at least in the minds of the electrodynamicists at the time. The equations of Maxwell, the later truncation by Heaviside and others, and the final truncation by Lorentz are all still material fluid flow models. They still assume the material luminiferous ether, more than a century after its falsification.

See again Figure 1-3 in Chapter 1. Contrary to the conventional view, force is not a fundamental causative entity, but is a system of two coupled entities. It involves a nonobservable, dynamic, causative 4-space entity (which may be modeled as a curvature of spacetime) interacting with a previous "frozen" observable (a 3-space mass). The output of this interaction is another observable (i.e., a change to the previous observable or an iterative output of its replica). Observables do not continuously exist, but continually recur as continual outputs of the iterative observation process (photon interaction).

Mass is a *component* of force, as is apparent if we define force F as $F \equiv d/dt(mv)$, which gives $F = m(dv/dt) + v(dm/dt)$. Both terms on the right of the expansion contain mass; hence force F contains mass as a *component*. Both terms on the right also contain nonmass terms; hence force F also contains nonmass coupled to and interacting with mass. So there can be a force generated by a nonmassive non-force entity in spacetime interacting with a mass, and the force only exists when that interaction is ongoing. The ether was erroneously considered to be a material system, containing mass, and hence capable of supporting a force. Maxwell modeled his fields (in what we know today to be mass-free space) as material force fields in the material ether.

In the 1880s, the Michelson-Morley experiments {125a-125d} destroyed the material ether. However, the fundamental Maxwell-Heaviside equations *assuming* those material force fields in the vacuum — which depended on the notion of a material ether and are a function of it — have never been altered, even to this day. Instead, electrodynamicists just assume the force "somehow" goes to zero in the field in space, but the same field remains. The equations still assume the same old material ether.

This is a non sequitur; a *force* field cannot be the same as a force-free *non-force* field. As an example, the EM force field in observable matter is an output of observation. It is therefore a *3-space effect containing mass.* The force-free field in mass-free spacetime (4-space) is a causal 4-space component of massless spacetime. *A priori,* it exists prior to interacting with the charged mass. It contains no observable mass and hence no force since mass is a necessary component of force. It independently exists without the presence of the charged mass at all.

In short, the field in massless spacetime is a specialized curvature (changed region) of that spacetime. *Spacetime consists of spacetime,* to slate a tautology. Once one fixes the concept of the medium, then any changes to the medium can only consist of "medium-stuff. So any "changes" in spacetime can only be changes of 4-space itself, in the Einstein general relativity sense. This area is handled quite well by Sachs' unified field theory {126a, 126b}, extending and completing the work of Einstein.

Sachs' epochal theory provides a great generalization of general relativity and electrodynamics reaching from the quarks and gluons to the entire universe. Further, Evans' extension of electrodynamics into a higher 0(3) group symmetry turns out to be demonstrably a special subset of Sachs' electrodynamics.[40] Hence, for the first time, an *electromagnetically* engineerable unified field theory is available. Much of general relativity, quantum mechanics, etc. can be engineered directly by Evans' electrodynamics subset of Sachs' overall unified field theory. I am much indebted to Evans {127} for the following statement with respect to O(3) electrodynamics:

> *"In 1992 it was shown (Phvsica B. 192, 227, 237 (1992)] that there exists a longitudinal component of free space electromagnetism, a component which is phaseless and propagates with the transverse components. Later this was developed into a Yang-Mills theory of electromagnetism with O(3) Lagrangian symmetry. This theory is homomorphic with Barrett's SU(2) electrodynamics and has far reaching implications in field theory in general. Recently it has been recognized to be a sub theory of the Sachs theory of electromagnetism, based on the irreducible representations of the Einstein group of general relativity.[41] The Sachs theory produces a non-Abelian structure for the electromagnetic field tensor. The O(3) electromagnetism also has implications for the* ***potential*** *ability of extracting energy from the vacuum,* ***and its*** *topological implications are currently being* ***investigated*** *by Ranada. The O(3) electromagnetism has* ***been tested*** *extensively against empirical data, and succeeds* ***in*** *describing interferometric effects and*

---

[40] E.g., see M. W. Evans, "The Link Between the Sachs and 0(3) Theories of Electrodynamics," in M. W. Evans (Ed.), *Modern Nonlinear Optics,* Second Edition, Wiley, 2001, part 2, p. 469-494.

41 Evans et al, "Derivation of 0(3) Electrodynamics from the Irreducible Representations of the Einstein Group," *Found. Phys. Lett.,* 15(2), Apr. 2002, p. 179-

> *physical optical effects where the conventional Maxwell-Heaviside theory fails. Implicit in both the O(3) and Sachs theories of electromagnetism is the ability to extract electromagnetic energy from curved space-time. Working devices based on this ability would help solve the current energy crisis."*

2.1.2 Mass Is a Component of Force

The dichotomy of using the field concept in two mutually contradictory fashions as something massless and also as something massive — still unresolved in electrodynamics — is clearly summarized by Jackson's important statement {128} of the prevailing view:

> *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."*

In trying to rationalize the "massive field versus massless field" dichotomy, Jackson {129} also states:

> *"...the thing that eventually gets measured is a force..."*
> *"At the moment, the electric field can be defined as the force per unit charge acting at a given point."*

Actually, field is not force, so cannot be force per unit charge. Field *produces* force when it interacts upon charge. *Causing* something as an effect, but then *being* that same effect identically, is a confusion of cause and effect and a non sequitur.

We point out that spacetime is charged, but with *observably massless* (virtual) charge. So the field in spacetime is observably massless and therefore observably force-free. That massless field then interacts with an observable massive charge in the "observable" world, to create an observable force and an observable forcefield. In space, from a quantum mechanical view we may consider the field to have *virtual* force x time, since the field may be regarded as a curvature of spacetime acting upon the *virtual* charged masses in the ever-bubbling *virtual particle fluctuations* of the vacuum permitted by the Heisenberg uncertainty principle. Nonetheless, that is only the "potentiality" for an observable force, as pointed out by Feynman and Wheeler, should the field — including its interactions with the virtual charges of the vacuum to produce forces and accelerations on those virtual particles — interact upon an observable charged mass.

Obviously, the force-free field as a product of spacetime curvature and massless (virtual) charge, and the force field as a product of *that* observably force-free field but virtual force field (in the vacuum) with observable charged mass, cannot be the same thing. Assuming they are the same assumes an oxymoron, and confuses cause and effect. This illustrates the continuing dreadful foundations difficulties propagated in classical electrodynamics.

To resolve these EM foundations difficulties so long maintained and cherished as dogma, the scientific community at the highest priority should (1) fund extensive and rapid work by our very best theoreticians and foundations physicists to heal these century-old festering sores, (2) extend electrodynamics back to a higher group symmetry theory, as in O(3) EM theory for example, and (3) provide a rigorous dictionary of science foundations concepts and terms which does contain valid definitions and full discussions — clearing misdefinitions and disinformation now notable in the electrodynamics literature.

Nobelist Feynman, co-founder of quantum electrodynamics, was careful to state the notion of the field in mass-free space very differently. Quoting Feynman {130}:

> *"...the existence of the positive charge, in some sense, distorts, or creates a "condition" in space, so that when we put the negative charge in, it feels a force. This potentiality for producing a force is called an electric field."*

Feynman's imagery clearly invokes a "distortion condition in spacetime" {131}, eventually coupled to a charged mass and interacting with it — and *then* one has a force and a force field. He erred in still somewhat making the force separate from the charged mass, in the phrase "it feels a force" — which ignores the fact that the mass of the interacting charge is a *component* of the force, not separate from it and not "feeling it as a separate external thing". He pointed out that the field in spacetime represents only the *potential for* a force field, *if a* charge should be placed there. And he points out that the field is a "condition in space" — in the full sense, a condition in spacetime.

We comment that, in a sense, spacetime and vacuum may be considered one and the same. Vacuum contains many fluctuations and virtual state dynamics, all of which are curvatures of spacetime or produce curvatures of spacetime. Vacuum has energy density, and is therefore a potential. This energy density has dynamics, so the ambient vacuum potential has dynamics, and so therefore does spacetime {132}.

However, a dynamic of the potential is precisely what a field *is*. It follows that dynamics of spacetime are what fields are. So fields (as force-free entities in spacetime) may be taken more accurately as the dynamics of the vacuum-spacetime, or as dynamics of the ambient vacuum-spacetime potential, or as dynamics of the ambient curvature of spacetime.

Again quoting Feynman {133}:

> *"We may think of E(x, y, z, t) and B(x, y, z, t) as giving the forces that would be experienced at the time t by a charge located at (x, y, z), with the condition that placing the charge there did not disturb the positions or motion of all the other charges responsible for the fields."*

Here Feynman very clearly shows that E and B involve functions of spacetime *(x, y, z, t)* only, prior to interaction, but do not physically exist until interaction with charge has occurred. But he has not quite yet completely eliminated the "force field" as a separate force acting on the charged matter independently. This centuries-old fundamental mistake in mechanics — that force is independent of and external to mass — has simply been propagated into electrodynamics, without correction. Physicists such as Feynman and Wheeler have been trying very hard to correct it.

The observed/measured force field is an output of observation, and observation is a *d/dt* operator imposed upon spacetime LLLT. All observables are 3-spatial, and any observation is an instantly frozen "3-space snapshot" at a single point in time. A single observation has no temporal existence *a priori,* since it is only an instantaneous 3-space snapshot existing at one moment only. Hence an observable — being 3-spatial and frozen — does not persist *as such* in time. Indeed, the next instant after an observation, that particular snapshot does not exist any more. [Refer again to Figure 1-3 in Chapter 1].

2.1.3 Mechanism for the Flow of Time and Related Matters

It is well recognized that the nature of time is itself an unresolved question. For our purposes, we are interested in the energetics[42] approach. So we apply that approach to arrive at a mechanism generating the "flow of an object through time". First we will choose our fundamental units.

2.1.3.1 *Getting Creative with Fundamental Units*

In physics, the choice of fundamental units one chooses for one's model is arbitrary. Usually mass, length, time, and charge are used, but a perfectly valid model can be generated using only a single fundamental unit.

This is well known to modern physicists and leading electrodynamicists, but sometimes comes as a shock to electrical engineers! Jackson {134} expresses it very succinctly: Quoting:

> *"For example, theoretical physicists active in relativistic quantum field theory and the theory of elementary particles find it convenient to* ***choose*** *the universal constants such as Planck's quantum of action and the velocity of light in vacuum to be dimensionless and of* ***unit magnitude.*** *The resulting system of units (called 'natural' units) has only* ***one*** *basic unit, customarily chosen to be length. All quantities, whether length or time or force or energy, etc., are expressed in terms of this one unit and have dimensions which are powers of its dimension. There is nothing contrived or less fundamental about such a system than one involving the meter, the kilogram, and the second as basic units. It is merely a matter of convenience."*

So let us consider what happens if we use the joule as the single fundamental unit for a model of physics. Then each of the entities "mass", "length", "time," and "charge" will become totally a function of energy. Since the dawn of relativity and the nuclear age, the notion that "mass is energy" is no problem, and everyone is familiar with Einstein's $E = mc^2$. Solving for m in that expression, one realizes that spatial energy E has been compressed by the factor c^2, to yield what is called "mass m". More rigorously, compressing the 3-space energy by the factor c , *and then leaving it in 3-space,* produces what we call "mass". However, if we remove it from 3-space and place it over on the fourth Minkowski axis *ict*, it becomes what we call "time" because *t* is the only variable in *ict*. Hence time has or can have similar energy density to mass. Specifically, 1 sec => $9\text{x}10^{16}$ joules of spatial EM energy, when transduced (decompressed) into spatial energy by a transducing charge.

---

[42] **Energetics** is an **old term from the early birthing** days of electrodynamics. Russian scientists resurrected **the term to cover their use of an** expanded electrodynamics in a unified field-engineering **manner, particularly in new** superweapons developed and tested after WW II and deployed in 1963 and subsequently.

In our new model using only the joule for our fundamental unit, we can also legitimately state that "time is energy", and be rigorously accurate. In that case, we must see what happens to 3-space EM energy when it is shifted to the time domain.

First, we take "EM energy in 3-space" as EM wave energy, in the usual transverse EM wave model (see Figure 1-4 of Chapter 1) of a field entity. We choose, say, the vector E oscillating in the x-direction, and another field entity (say, the vector B) oscillating in the y-direction. If we then rotate each of these vectors into the time axis, then that is an orthogonal rotation for each, which is what the velocity *c* actually is. So we have to do two orthogonal rotations, and the two resulting *c's* multiply to give c^2. With these two orthogonal rotations into the time axis, we have completely rotated the EM field wave energy vectors into the time axis. We have also compressed that EM field energy by c^2.

Hence "time as energy" is just EM spatial energy compressed by the factor c^2, but rotated into the time-axis as "time" rather than remaining in 3-space as "mass". Again, by this second analogy time has the same energy density as mass, but the highly compressed spatial energy has been rotated into the time domain to "produce time" rather than remaining in 3-space to "produce mass" {135}.

This sheds new light upon the scalar (time-polarized) photon of quantum field theory. Note that, prior to compression of the wave energy into time, the magnitude of the electric field *E* in the EM wave was oscillating lengthwise along its x-direction, which means that, with respect to x, it was a longitudinal EM wave. A similar consideration exists for the oscillation of the magnitude of *B* in the y-axis. Along the time-axis, we also have a "time-energy" vector having magnitude, and the magnitude of that time-energy can only vary along the time axis. Oscillating the time energy produces a longitudinal EM wave in the time dimension (on the fourth axis). What is actually varying is the time-density itself. This is a *time-polarized* or *scalar* EM wave, where by use of the term "scalar" we mean that it has no vector component in 3-space. In fact, this variability of time density produces a "galloping motion" in the speed of light itself, which has been recognized {136}.

We have uncovered a direct connection between the time-polarized EM wave (which can only oscillate its time-energy magnitude longitudinally in the time-domain) and the longitudinal EM wave in space. Such waves can only be observed when coupled or paired with an interaction existing between them, very similar to what Mandl and Shaw {167, 168} argue for the time-polarized and longitudinal photons of quantum field theory.

Again see Figure 1-3 in Chapter 1. Any observable is an instantaneous, frozen 3-spatial snapshot (or 3-intersection) of a 4-dimensional ongoing event. "Observation" itself may be taken as a process where a *d/dt* operator is invoked upon 4-space (spacetime), leaving a purely 3-spatial output. However, the resulting observable does not "exist or persist in time", since rigorously it is what was "caught" in only a single, frozen "3-slice" at one single instant, forever fixed.

For that reason, time is not an observable. It cannot be directly observed, even in principle, since *a priori* it is discarded by the observation process itself!

For further work along such lines, we recommend the Sachs unification approach to a combined general relativity, quantum mechanics, and electrodynamics, generalized from a topological standpoint {126a, 126b}. To work electromagnetically in that theoretical approach, we recommend the 0(3) electrodynamics as extended by Evans {137a, 137b}.

2.1.3.2. *The* ***Four Polarizations of****Photons* ***and*** *EM Waves*

In quantum field theory, four polarizations of a photon are recognized {138}. These are the *x-, y-, z-,* and /-polarizations, where *x-, y-,* z- and *t-* are assumed as the four dimensions in a 4-space. By agreement, z- is used as the direction of motion of a free photon or an electromagnetic wave. So at least in theory, there must also be four polarizations of electromagnetic waves, even though not all these waves are yet experimentally known.

The *x- and y-* polarizations are the familiar transverse photon and the transverse wave. The z- polarization along the line of propagation gives the longitudinal photon in 3-space and the longitudinal EM wave in 3-space.[43]

---

[43] Here we are **applying the conventional** model of the "EM wave in space", which is known to be erroneous **but is** in **ubiquitous** usage. See Romer, {84} for a scathing characterization **of that** erroneous model **But here we are stuck with it, so will use it** as "just a model".

2.1.3.3 *Imperfect Longitudinal EM Waves* ***in 3-Space***

When attempting to produce longitudinal EM waves (LWs) in 3-space, from transverse EM waves (TWs) input to some sort of polarization transduction process, only imperfect LWs are produced and a residue of TW content remains. The resulting imperfect LW is referred to as an *undistortedprogressive wave* (UPW). Some work has been done on UPWs {139}. UPWs are expected to have remarkable characteristics including wave velocities either slower or faster than *c {*140}.

The *t*- polarization in the time dimension is quite unique: The *spatial* energy overall is in equilibrium and not vibrating at all; instead, the energy of the photon or wave is vibrating in the time domain and therefore exists in the time domain. That is called a "scalar photon" (time-polarized photon). Its wave version does not yet seem to be known in the literature, although in 2000 we uncovered its secret hiding place {12}. It was unexpectedly hiding in Whittaker's {85} decomposition of the scalar potential in 1903, but had been slightly misinterpreted.

2.1.3.4 *Photons andEM Waves Carry* ***Energy and Time***

On the other hand, the concept of "EM waves flowing in 3-space" may be in need of a thorough overhaul {141}. A photon *y* is a "piece of angular momentum" in the form of $y = (dE)(dt)$. Hence the photon carries an increment of spatial energy d*E* and also an increment of time-energy d*t*. The time-energy component *(dt)* may be regarded as ordinary spatial energy that has been compressed by the factor c^2 {142}.

As can be seen, since *c* represents an orthogonal rotation in n-space, the multiplication or division by *c* and by powers of c changes the dimensionality of an entity, as seen by the observer in the laboratory frame.

So the photon transports two types of energy: (i) a "weak spring" (spatial, decompressed) energy d*E*, so to speak, and (ii) a "very stout spring" (time) energy d*t*, so to speak.

When a mass *m* absorbs a photon *(dE)(dt),* the *(dE)* component is compressed spatially by c^2, turning it into an extra amount of mass d*m,* so that the mass becomes $(m + dm)*d$t the same time, the *(dt)* component is joined, so that what results is $(m + dm)*dt$. In short, mass *m* is changed to masstime *mt* by photon absorption. So in the absorption of a photon y by a mass *m,* we have

$$y+m\text{->} (dE)(dt) + m \text{->} (m + dm)dt$$

In short, the mass *m* turns into masstime *mt*, by absorbing a photon, and masstime *mt* is as different from mass *m* as impulse *Ft* is different from force *F*. We point out that "mass" *m* alone does not even exist in time, but masstime *mt* does exist in time. This is proposed as a simple but fundamental correction to much of present physics. Further, the state "mt" is an excited, time-charged state, excited by very dense time-energy.

For the simplest case, in the next instant the excited state *mt* decays and a photon is re-emitted, and so we have

$$(m + \Delta m)\Delta t \rightarrow (\Delta E)(\Delta t) + m \rightarrow \gamma + m \qquad [12]$$

So emission of a photon changes the excited masstime state (time-charged state) back to mass (uncharged state with respect to time-charging), in the simplest case.

See again Figure 1-3 in Chapter 1. "Observation" and "observable change" are in fact generated by the photon interaction. The total photon interaction with a mass creates the "flow of that mass" through spacetime, macroscopically as seen by the observer and microscopically in myriad streams of virtual time changes. Mass is continually charged to the masstime state, and the masstime state is continually decayed to the mass state. Time flow itself thus has a myriad internal EM energy streams and a vast dynamic EM energy substructure.

The photon itself is not observable, as it exists prior to interaction. It is thus a "causal" entity in 4-space. It *is not* a simple 3-space observed "frozen snapshot" particle! When the causal photon is absorbed, its spatial energy component goes through the c compression function, thus adding a small amount of new mass to the previous mass (previous effect), but simultaneously connecting its d*t* time-tail, so that the slightly increased mass in fact now exists as masstime and not mass. Observation has not been completed at that point.

In the next instant, a photon is re-emitted (from masstime, never from mass!), in the simplest case removing that extra little mass increase and orthogonally rotating it back into 3-spatial EM energy, coupling that spatial energy increment d*E* to that "time-tail" d*t* and tearing it away as the photon is re-emitted. That leaves behind a frozen 3-space snapshot of the mass (the interacting particle) and completes the "observation". In the simple case, this is a replica of the previous particle (frozen 3-space snapshot) with which the causal photon interacted. So this process accounts for the so-called "persistence in time" of an object or mass, as seen by the observer's continual recall process accompanying his stream of iterative observations. Mass does not *continuously exist* in time, but *continually recurs* in time.

As is well known, what we call "observable" change must involve the expression in equation [13] in the fashion discussed and with the observation mechanism given.

We first pointed out the preceding mechanism for the "flow of a mass through time" in 1971, as a graduate student at Georgia Institute of Technology, and later published it in a crude paper {143} in 1973. It still needs a far more thorough theoretical treatment, which perhaps some graduate student will take up as the subject of his or her doctoral thesis.

2.1.3.5 **Photon Interaction:** *Mechanism Generating Flow Through Time*

See Figure 2-1. Mass does not really "travel through time" continuously *per se*, but proceeds with an overall serial change mechanism as

$$m \rightarrow (mt) \rightarrow m \rightarrow (mt) \rightarrow m \rightarrow \ldots. \qquad [14a]$$

where *(mt)* symbolizes a nonobservable ongoing "coupled" interaction state prior to observation completion. Equation [14a] represents the results of the continual photon interaction process, observation process, and "flow of a mass through time" process, at the highest single-photon interaction or quantum change level, and as "seen" or measured by the external observer.

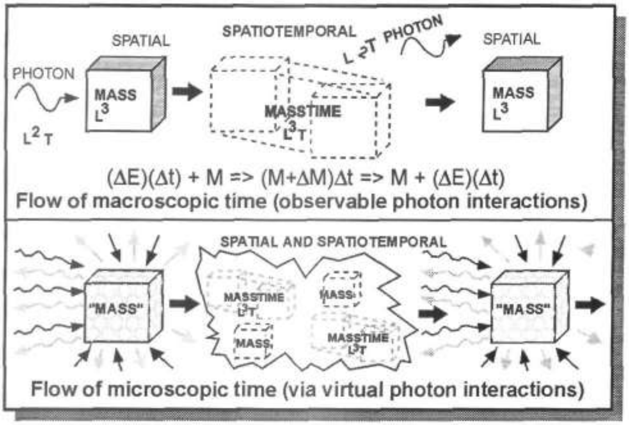
SPATIAL
SPATIOTEMPORAL
L^2T PHOTON
SPATIAL
PHOTON
MASS L^3
L^2T
MASSTIME L^3T
MASS L^3
$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$
Flow of macroscopic time (observable photon interactions)
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME L^3T
MASS
MASS
MASSTIME L^3T
"MASS"
Flow of microscopic time (via virtual photon interactions)

Figure 2-1 Mechanism for the flow of a mass through time.

Also, in that overall quantum level "flow" there are anyt number of ongoing streams and "sublevels" in the underlying subquantum level (in the virtual vacuum and in its virtual photon interaction with observable mass). We might write one of these "internal time-stream interaction flow substructures" as:

$$(\delta m) \rightarrow (\delta mt) \rightarrow (\delta m) \rightarrow (\delta mt) \rightarrow (\delta m) \rightarrow \qquad [14b]$$

In equation [14b] we have used parentheses, because all terms individually are nonobservable. We hypothesize that each stream continues until one of its terms participates in a summation which eliminates it by enfolding it into another interaction.

When a mass is observed as in [14a], a photon has been emitted (we observe the mass as the "effect" remaining). Time has been stripped away by the resulting *d/dt* operation, leaving a frozen 3-spatial snapshot, which we will see as (having been) a particle (simplest case). That occurs just after major ("observable") photon emission from the masstime state {144}. Immediately another observable photon is absorbed, and so state *mt* occurs. The particle of mass actually oscillates at a very high rate between the *m* and *mt* states — so high a rate that by arranging the interaction conditions one may interact with it either as a wave (react predominantly in the *mt* state) or as a corpuscle (react predominately in the *m* state).

Hence we propose that the process in [14a] may account for the duality of particle and wave.

2.1.3.6 *The Overall Flow of Time Has an Internal Dynamic Structure*

During the transition in any mass to masstime state by reaction of the mass with an "observable" photon, a myriad of fleeting virtual photon interactions involving very tiny *(dE) (dt)* components occurs with the mass *m.* These tinier increments of time, and their increments of energy, constitute internal structures in the time flow process. Therefore they may be considered as "energy currents" or "time-like energy currents" and dynamic structures or streams inside the flow of time. This is particularly straightforward if we use a model having only a single fundamental unit, the joule. In that case, time is energy, and we are speaking of energy flow and its constituent internal structures of energy flow.

So the d*t* component of masstime at the observable-photon action level has a myriad of energy-time structured dynamics infolded within it. Hence the *mt* state is very dynamic in time, particularly for fundamental particles. The *mt* state is in fact a "collection of time-energy dynamics" and therefore "wavelike".

See again Figure 2-1. A major point is that ***mass*** does not emit a photon; *masstime* does. Mass "travels through time" by an extremely high oscillation between corpuscle-like state *m* and wavelike state *mt*, and with a vast internal "dynamic streams" of other such high oscillations between corpuscle-like *m* states and wavelike *mt* states.

The concept can be very much extended, of course, but this suffices for our concept of energy currents in time, the interaction of such energy currents with mass in a mass system, and the internal dynamic structuring of the "flow of time".

This internal structuring is important in the event of *"time-energy" charging.* There the internal structure of time has experimental consequences. We will discuss those consequences in later chapters.

2.1.4 The Ubiquitous Substitution **of** Effect **for** Cause

See Figure 1-3. To repeat, no observable exists or persists *as such* in time. An observable *as such* only exists at all when time has been momentarily stopped. It then quickly changes into (observable x time) form, for a time interval after a photon absorption. Then it changes again to observable form by photon emission and the corresponding *d/dt* differentiation imposed by that photon emission.

What happens when we think we see an "observable moving through space" is that we make or suppose a continual series of very fast *d/dt* 3-spatial snapshot observations, one after the other in serial fashion (much like the individual frames of a motion picture film). For more than a million years, human brain and mind processing have "always interpreted" this serial set of iterative operations occurring in the eyes and in the physical senses — and continually recalled and processed by the mind — as an "observable persisting in time", which it is not.

This age-old "natural" and inbred, instinctual *mental practice* and assumption by humans — and therefore by scientists — has provided a major problem in physics and especially in electrodynamics. It has resulted in the substitution of the "effect" for the cause in a great number of models. Even a rapid and continual series of 3-space effects after reaction of an observable (frozen snapshot) with a 4-space causal entity, does not and cannot constitute that 4-space entity which interacts with the previous "frozen snapshot" observable and causes the resulting "comparative change or sameness" of the new observable with the former. Hence, e.g., Romer's scathing condemnation {84} of the conventional drawing of an "EM wave in space".

In general relativity, it is straightforward. The curvature of spacetime — the cause — is not the mass (the 3-space observable effect) that is further changed or created again in the ongoing interaction, as seen in the "next

3-space snapshot" (next output of observation) when comparing that snapshot to the previous one.[44]

The same non sequitur — unwitting substitution of effect for cause — has existed for hundreds of years in mechanics, for example. There it is strongly passed on in the erroneous old notion of a separate force acting upon a separate mass. *Prior to* the interaction, no "force" exists. *During* the ongoing interaction, force exists because it is the product (interaction) of a "non-force" causal entity (e.g., a field in 4-space, as a certain curvature of spacetime there) and a previous effect (a 3-spatial frozen snapshot called "mass"). The force exists during the interaction and only during the interaction, when both cause and previous effect are coupled (and will produce the new effect at the completion of the next serial *d/dt* operation as a photon is emitted). The force can only exist when the mass (previous effect) is coupled to the cause (the curvature of spacetime), since *a priori* mass is a *component* of force {145}, and curved spacetime is another *component* of it. After the interaction "ends" in a new *d/dt* observation, the mind compares the new effect with the former, to determine whether there is a "change" or a replica (sameness). This comparison of the new observation (observable) with respect to the former, *occurs in the brain and mind processing,* and gives the sense of "persistence of an object in time" as well as of an object "changing in time".

So a great faux pas in physics — and especially in electrodynamics — is this widespread substitution of the effect for the cause. Understanding this non sequitur clearly, and correcting it, is a prerequisite to understanding the principles and concepts of legitimate COP>1.0 Maxwellian systems. Also, if we would ever hope to adequately correct electrodynamics, this tremendous faux pas must be recognized and rooted out of the model. Indeed, the same is true for mechanics and other branches of physics.

2.1.5 Some Additional Foundations Non Sequiturs

Another gross non sequitur is made in calculating the amount of EM energy flow continually being diverged around an intercepting unit static charge at a point in space, from a potential in which that point is located, and then calling this "local intensity" of the collecting interaction the "magnitude of the potential" itself.

---

[44] Rigorously, if one wishes to connect spacetime and mass, spacetime can be interpreted in *masstime* units, but not ***mass*** units And vice versa. Again, masstime *mt* is as different from mass *m* as impulse Ft is **different** from force *F.*

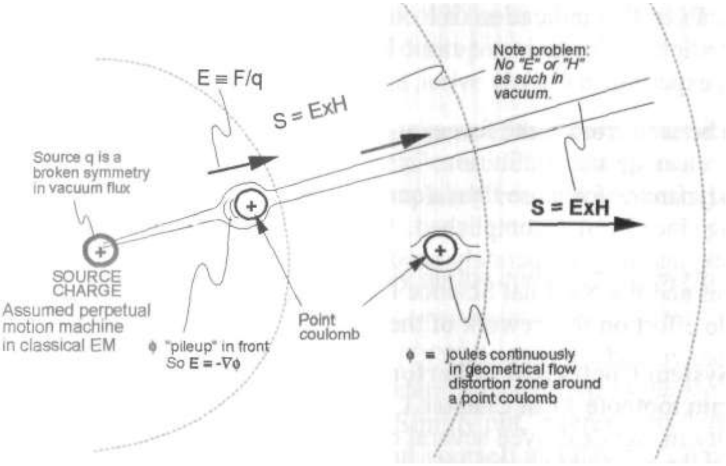
Note problem:
No "E" or "H" as such in vacuum.
E ≡ F/q
S = ExH
Source q is a broken symmetry in vacuum flux
S = ExH
SOURCE CHARGE
Assumed perpetual motion machine in classical EM
ϕ "pileup" in front
So E = -∇ϕ
Point coulomb
ϕ = joules continuously in geometrical flow distortion zone around a point coulomb

Figure 2-2 Field and potential are defined in terms of the diversions from them around a unit point static charge assumed at any spatial point

(See Figure 2-2). At best, it is a measure of the *local intensity of the diversion of energy flow from* the "bidirectional" EM longitudinal waves comprising a potential, as they are diverged around the unit point static charge. It is not the magnitude of the entire potential, filling all space, nor does it prescribe the spatial energy magnitude of that entire potential {146}. It is a small energy flow, *diverted* by an assumed standard fixed charge at a single point in the potential, from the overall energy flows comprising the potential. Calling this little "diversion of a bit of energy" the magnitude of the potential is rather like calling a little whirlpool in a river the "magnitude of the river". Or more exactly, it is like calling the amount of little "swirl diversion of a river's flow" around a standard small fixed rock the "magnitude of the river".

A similar non sequitur exists in the usual calculation of the "field" magnitude at a point, which is not the "entire field entity itself at all. It is merely what is *diverged from* the "field as it exists in spacetime" at a single point in it, by an assumed fixed standard point charge at that point that is interacting with the field. The field takes into account the patterning of the diverged flow, e.g., the difference in potential (difference in pressure) across the unit point charge. That is not the field itself at all, and the textbooks and dictionaries of physics are completely wrong in calling it such. Again, at best, it is an indication of *the field intensity* at that point, and the reaction cross section of the field to an assumed unit point static charge. Even that indication of local intensity of the field will change if the intercepting charge is not static but is in resonance, as shown by the Bohren experiment {24}.[45]

As can be seen from these few examples, a great deal of work needs to be done to clean up the "definitions" promulgated in physics and electrodynamics for more than a century. It is inexplicable[46] that this has not long since been accomplished. These errors permeate almost every textbook, and they desperately need correcting. The National Academy of Sciences and the National Science Foundation should fund the best possible effort on this rework of the foundations of physics.

2.1.6 System Coefficient of Performance (COP) and System Efficiency

See again footnote 11 in Chapter 1 and Appendix A. In 1915, general relativity in fact destroyed several of the fundamental definitions (axioms) of classical thermodynamics. It necessitated our correction of some thermodynamical definitions (open system and closed system) and our more rigorous definitions of COP and efficiency advanced below.

Once mass is recognized as energy and the two are just different sides of the same coin, there can be no thermodynamically closed system which passes energy across the system boundary without passing mass across it. Any system changing its rate of flow through time also changes its mass and its energy, and vice versa. Energy has mass characteristics, and any mass system with added or subtracted energy — either kinetic or potential

---

[45] Many Bohren-type experiments are continually done in nonlinear optical labs in universities and elsewhere. The effect (excess radiation from the medium) is euphemistically called "negative absorption of the medium," "negative resonance absorption of the medium," etc. Such terminology avoids clearly recognizing that more energy is output than the scientists (erroneously) calculate was input. Bohren faced the issue head-on, and clearly stated that more energy was output than was input (or than was thought to be input).

[46] A true pessimist might point out that it may be quite understandable, considering that it solidly blocks COP>1.0 EM electrical power systems and electro-gravitational anti-gravity systems from being developed by our universities, which would solve the energy and transportation crises forever. Thereby it also keeps about $1.5 to 2 trillion a year pouring into the coffers of controlling financial groups behind many great interlocking corporations.

— will also change its mass relativistically as well as its other characteristics such as inertia and gravitational attraction. As pointed out by Logunov and Loskutov {635}, in 1917 the new characteristics of general relativity led Hilbert {632} to observe that in general relativity there can be no energy conservation rules of the usual classical kind. This also follows from our consideration of the supersystem with multiple active environments rather than a single active environment.

We have formally destroyed any absoluteness of the present classical thermodynamics, which is just an imperfect model. It is a special case which can be approached but never completely reached in a real dynamic system.

With that in mind, we now more exactly define *coefficient of performance* and *efficiency* of a system.

The term "coefficient of performance" (COP) is a ratio whereby the useful energetic output performance of the system is characterized *purely with respect to the operator's energy input.* Simply put, it represents "how much you get for what you have to input yourself and pay for". It does not characterize the performance of the system with respect to the *total energy input from all sources.* For our purpose, in general there are two major ways of expressing this COP:

(a) COP == (work accomplished in the load during a representative time of operation) divided by (energy input by the operator during that representative time of operation). Or for short, COP = (useful work out in the load) + *{operator's* energy input). We stress that there may be other free inputs of energy to the system, in addition to the operator's input, but only the operator's input is accounted in the COP calculation.

(b) COP == (average power out in the load) divided by (average power input by the operator).[47] Again, there may be additional average power inputs that are freely input from the environment, without cost to the operator, but only the *operator's* input is accounted. We stress that "power input" is another of electrical engineering's misnomers, but we use it here because it is universally used in power system engineering.

---

[47] This is in standard electrical engineering terminology, which is mangled. For a more precise physics statement of (b), COP == (the average rate at which the system dissipates energy in its load to do useful work) / (the average rate at which the operator must furnish energy to be dissipated as work in the input section of the system to make it operate and do its output work).

For a system such as a transducer, which merely changes the form of the energy in some fashion and does not perform work in an external load (the transducer is its own load), we may express the COP as

a. COP == (effective energy output) / (energy input by the operator). Here we accent that there may be additional energy inputs to the system from the active environment, and these inputs are "free" and are not input by the operator, so they are not accounted in the COP calculation.

b. COP == (average power output) / (average power input by the operator). Again, additional average power inputs may be freely received from the environment, but they are not accounted in the COP calculation.

The "efficiency" E of a system is a ratio less than or equal to 1.0 (or a percentage less than 100%), where E indicates the percentage of the total input energy (from all sources) that is dissipated in the load as useful work. It follows that (7 - E) indicates the percentage of the total input energy that is dissipated in the internal losses in the system, not directly resulting in useful work by the system in its load. So the efficiency E of a system may be expressed as:

a. E == (total work output in the load during a representative operational period) / (total energy input to the system, from all sources, during that period).

b. E == (average power output in the load) / (average energy rate of input to the system, from all sources, during that representative operational period.).

We accent that no system can have an efficiency greater than 100%, for then it would be "creating energy from nothing" {147}. No system can dissipate or convert energy that it does not first receive. The conservation of energy law states that energy can neither be created nor destroyed. This means that *there actually are no energy sources per se, in the sense that the source creates the energy,* even though we use the terms "energy source" and "source of energy" informally. E.g., Semiz {148} states it this way:

> *"The very expression 'energy source' is actually a misnomer. As is known since the early days of thermodynamics, and formulated as the first law, energy is conserved in any physical process. Since energy cannot be created or destroyed, nothing can be an energy source, or sink. Devices we call energy sources do not create energy, they convert it from a form not suitable for our needs to a form that is suitable, a form we can do work with."*

A medium inefficient system can easily have a COP>1.0, if it receives additional energy from its active environment, and if that extra free energy is sufficient to overcome its internal inefficiency losses. An example is a common windmill, where a very good one may have an efficiency of less than 50% due to friction and drag losses in the gears and wind spillage losses in the blades, etc. Yet the operator himself does not have to input any energy at all, for the windmill to operate continually. In this case, the COP»1.0 and approaches infinity, but the efficiency of the windmill is still less than 50%.

Another example is the home heat pump, which may also have an efficiency of less than 50%. When acting as a refrigerator, its theoretical maximum COP = 8.22 under nominal conditions {149}, and a well-designed 50%-efficient home heat pump will produce COP = 4.0 when conditions are suitable.

On the other hand, if the only net energy input to the system is that energy that the operator inputs, then the COP < 1.0. This is a system in equilibrium with respect to any energy exchange with its external environment — *except* with respect to the energy input by the operator (i.e., with respect to its "fueling" by the operator, so to speak) and with respect to the energy subsequently dissipated in the loads and losses. If the system is 100% efficient (has no internal losses or conversion losses at all), its COP = 1.0. Almost all real systems do have internal losses, so their efficiency is E < 100%. In that case, the system in equilibrium with its environment, and having E < 100%, will also exhibit COP<1.0

irrevocably. Indeed, the same number will give both the efficiency E and the COP, because numerically they are then the same.[48]

For electromagnetic systems, the state of confusion between efficiency E of the system and the COP of the system is due to one fact. *Numerically* the two are always equal in *equilibrium* systems — and the Lorentz regauging condition enforced by the closed current loop circuit self-enforces the equilibrium condition of the system with respect to its active environment. Because electrical engineers usually have zero experience with COP>1.0 electrical systems, they tend to loosely and erroneously use the two terms "efficiency" and "COP" as if they meant the same thing. They do not.

Now suppose that we have an open system, far from equilibrium in its energy exchange with its active environment. Suppose that the system's efficiency E is very poor, so that E = 20%. Now suppose that the environment inputs twice as much energy as does the operator. Let the operator's energy input be E_1. Then the total energy input to the system, from both the operator and the environment, is $3E_1$. The efficiency is only 20%, so the system outputs W, as work in the load, of $W = 0.2(3E_1) = 0.6 E_1$. The COP of the system is the work out divided by the operator's input, which is $COP = W/E_1$, which is $COP = 0.6\ E_1 / E_1$ which gives COP = 0.6. As can be seen, even though this system receives twice as much additional free energy as what the operator inputs, it is so inefficient that **its COP<1.0.**

Suppose we have a similar system with the same energy inputs from the operator and from the environment, but now the system's efficiency E is E = 90%. Then the total energy input to the system, from both the operator and the environment, is $3E_1$. The efficiency is 90%, so the system outputs W, as work in the load, of $W = 0.9(3E_1) = 2.7\ E_1$. The COP of the system is the work out divided by the operator's input, which is $COP = W/E_1$, which is $COP = 2.7\ E_1 / E_1$, which gives COP = 2.7. As can be seen, a more efficient system in the same energy input situation, now outputs more work than the energy input by the operator. Energy is conserved at all times; the excess energy for the additional work was in fact freely input to the system — which is an "open" system far from equilibrium in its exchange with its active environment.

---

[48] We emphasize that COP and efficiency are two quite different concepts, however, even when their numerical values are the same. A 6-foot tall man and a 6-foot tall door have the same number for their **height,** but only a fool would consider them the "same thing".

Thus a system far from equilibrium in its energetic exchange with its active environment, is *permitted* to exhibit COP>1.0 even though its efficiency is always less than 100%. On the other hand, if the system is very inefficient and the energy input from the environment is not too great, the system will still exhibit COP<1.0. But the operator will pay less for his energy costs to operate that inefficient system, than he would pay to operate it if he himself had to furnish all the energy input.

Two cautions are emphasized:

(a) First, electrical engineers use the term "power" to mean "energy flow rate without dissipation or change of form", as well as "energy flow dissipation and change of form rate." This is inexact and unfortunate, but it appears ubiquitously throughout the electrical engineering literature. So we are essentially "stuck with it" for the moment. From a rigorous physics point of view, energy flow without divergence or change has absolutely no power, because power is defined as the rate of change of the form of that energy flow — e.g., the rate of dissipation (scattering) of that energy flow, which in physics is a rate at which work is being performed.

(b) Second, it seems that more than half the engineers and scientists — and many textbook authors — do not clearly understand the difference between COP and efficiency, and often interchange these two terms as if they were the same. They are not at all the same, and a great deal of confusion exists in the casual engineering literature because of using them interchangeably. That is why we have clarified them and given some simple explanatory examples, sufficient for our purpose in this book.

2.1.7 COP>1.0 and the Question of Perpetual Motion

Legitimate scientific researchers attempting to work in open EM systems far from thermodynamic equilibrium are often subjected to rather vicious ad hominem attacks, charging that they are perpetual motion machine addicts and therefore raving lunatics. This of course is not true. We are simply trying to do with Maxwell's "electric fluid" systems what nature already shows us can be done with material fluid systems. We are in fact seeking the equivalent of "electromagnetic heat pump" processes and systems. It is the strident critics who are thoroughly confused and who reveal an immature knowledge of physics and thermodynamics.

As we explained, prior to their Lorentz symmetrical regauging, the Maxwell-Heaviside equations already clearly prescribe and permit COP>1.0 EM systems, including EM power systems. In addition to the present EM systems in equilibrium in their exchange with their active vacuum environment, the non-regauged equations also include Maxwellian systems far from equilibrium in their energetic exchange with an active environment, such as the active vacuum. Classical *equilibrium* thermodynamics — stridently waved as a sacred banner by the critics accusing COP>1.0 researchers of being "perpetual motion nuts" and such EM systems as being "perpetual motion nonsense" — does not even apply to such *disequilibrium* systems! One should simply read the literature of the thermodynamics of such disequilibrium systems {160a-160f}. Or one should read why Prigogine was awarded the Nobel Prize in Chemistry in 1977, for his contributions to such open systems in disequilibrium.

We accent again that, under our new rules where we corrected the thermodynamic definitions of open system and closed system, we are in compliance with general relativity. Further, every EM system must be taken as an open system, since in the supersystem energy (with mass-like changes resulting across the system boundary) is always exchanged between the three supersystem components.

Fortunately there are also rational, reasonable critics such as Angrist {150, 151} where the criticism is properly scientific and not ad hominem, even though still misguided. Usually the rational perpetual motion critic applies the three laws of classical *equilibrium* thermodynamics, to classify "perpetual motion machines" of class 1, 2, and 3 after equilibrium thermodynamics laws 1, 2, and 3. This of itself has no intellectual content for open systems in disequilibrium with their active environment. The classical equilibrium thermodynamics — including laws 1, 2, and 3 interpreted only in terms of the operator's energy input — does not apply to *open systems far from equilibrium* — which are the only systems permitted to exhibit COP>1.0 in the first place. And we insist that all EM systems are open systems far from thermodynamic equilibrium, since every charge is, and all fields and potentials come from those source charges. Indeed, such open disequilibrium systems are *permitted* to violate the laws of classical thermodynamics! As an example, the entropy of such an open disequilibrium system cannot even be computed {152}, and it is less than the entropy of the same system in equilibrium, since the equilibrium state is the state of maximum entropy.

The rational critic then falls upon the horns of a dilemma, by admitting that, yes, such open dissipative systems do exist and do permissibly exhibit

COP>1.0. He then labels these *as, false or fictitious perpetual motion machines* {150, 151}. In short, when such COP>1.0 machines are real and admittedly exist, they are not *"true"* perpetual motion machines but *fictitious* perpetual motion machines. My comment is that they are not perpetual motion machines at all, but permissible disequilibrium machines freely receiving excess energy from their energetic exchange with their active environment. They are no more mysterious than a windmill, a sailboat, a watermill, or a solar cell bank.

One must be precise. A bicycle is a free-energy COP>1.0 machine while going downhill. It is not such at all when going uphill. The difference is that gravitational input of energy aids the bicycle's motion in the downhill case, and opposes it in the uphill case. The active environment gives the system additional energy in one case, and takes energy from the system in the other case. Disequilibrium exchange can cut either way!

It is precisely the EM system far from equilibrium with its active environment — where the environment puts in additional energy — that the serious overunity researcher is seeking. It is a perfectly valid pursuit, even by the standards of the perpetual motion critics, who really should study *both* kinds of thermodynamics, not just *one.* They also should study how Maxwell's equations were curtailed and truncated; one result is that all Maxwellian disequilibrium systems were — and are — discarded *arbitrarily* by the Lorentz symmetrical regauging.

The only criticism one needs to raise vis a vis COP>1.0 and perpetual motion machines is this: If an inert (no hidden internal source) machine receives no net energy from its external active environment, but is in equilibrium in its energetic exchange with that environment,[49] then classical thermodynamics rigorously applies to it and the machine cannot and will not exhibit COP>1.0. However, all EM systems contain charges, and these are indeed "hidden sources" of EM fields and potentials and their energy. It *rigorously follows that, since they contain hidden energy sources, all EM systems are potentially COP>1.0 systems a priori.* Further, *it follows that some ubiquitous design characteristic of the system itself must prevent the system from achieving COP>1.0, since nature does not prevent it.* That characteristic feature is the ubiquitous closed current loop circuit, directing all spent electrons in the external circuit right back through the source dipole in the generator or other primary power source.

---

[49] Here we quibble again and point out that Lorentz's symmetrical regauging actually does require that the system receive and take on excess EM potential energy from its environment (the active vacuum and the local curved spacetime). However, this excess energy can only be received in the form of two injections of field energy, where the fields are equal and opposite. Hence the energy is "bottled up" as a stress potential, with net zero translation force available to it. Since this stress potential energy cannot translate electrons, it cannot do work in an external load. It merely changes the stress of the system itself. The net result is actually a rotation of the frame in which the system exists, away from the "laboratory frame" of the observer. In short, the Lorentz condition establishes a new "equilibrium condition" for the system, by rotating its frame. It also implies that EM energy is continuously received by the system from both the active vacuum and local curved spacetime, in order to maintain the rotated frame and **the** new equilibrium condition. Further this

It is indeed mandatory that — if at all possible - the inventor or independent researcher *identify* a legitimate source of environmental energy that his purported COP>1.0 device receives and collects, and how it receives it and then uses it to power the loads. That means he should identify where and how his system violates the Lorentz self-regauging condition imposed by the standard closed current loop circuit. If the inventor himself cannot explain it — and this does happen — then someone else must explain it for him. If the machine works under proper independent test and replication, then criticism that no explanation exists for it in normal physics is of no relevance. It is a spur to the theorist to enlarge his horizons and to find the proper physics that *does* explain it.

We stress that we cited the works of Angrist because his voice is reasoned and he does not stoop to name-calling and ad hominem attacks. He also does point out a historical rash of machines that clearly must comply with classical equilibrium thermodynamics, because they have no discernible or stated extra input of energy from their active environment and have no self-asymmetrical regauging. When such machines have no external free input of energy from their active environment, they are systems in equilibrium with their active environment, and cannot exhibit COP> 1.0 — exactly as pointed out by Angrist. With such a reasoned critic, a scientific dialog is possible, and his main thrust — that no *equilibrium* system can exhibit COP>1.0 — is quite valid. With the vehement critic, it is already a cur dog fight from the outset and there can be no scientific discussion at all.

The only contention one has with a rational critic, whose criticism is based on classical thermodynamics, is when he then generalizes that *all* EM

continuous input of two energy flows, equal and opposite, does continuous internal work on the Maxwellian system to increase and maintain its stress.

systems fall within this class. That conclusion is a non sequitur, and ignores the entire science of disequilibrium thermodynamics and a vast body of scientific literature and scientific experiments in electrodynamics. When stated for electrodynamic systems, it also ignores the actual documented history of electrodynamics, and ignores Lorentz's *arbitrary discard, from all accountability,* of the Heaviside giant nondiverged energy flow component accompanying every circuit and system. It also ignores Lorenz's and then Lorentz's symmetrical regauging of the Maxwell equations to give new and simple equations easier to solve. This arbitrary regauging did in fact unwitting discard all open disequilibrium Maxwellian systems — precisely those permitted to exhibit COP>1.0.

See again Figure 1-1. The energy entering the circuit is entering it from the surrounding space outside the circuit. Heaviside showed that the energy flow component remaining alter the Poynting component is withdrawn is orders of magnitude greater than the Poynting component entering the circuit and powering it. *It can easily be demonstrated that every generator outputs far more EM energy in space surrounding the power line attached to it, than the operator inputs as mechanical energy to the shaft.*[50]

---

[50] E.g., just use a gedanken experiment followed by the Bohren {24} experiment as validated by Paul and Fischer {25}. Consider a 100% efficient DC system, powered by a DC generator, using a closed Current loop circuit. All the Poynting energy received by the system appears at first blush to be dissipated in the external load. That is not true, but let us not yet belabor the point. Include the complete circuit loop back through the source dipole in the generator. The back emf voltage across the source dipole between the terminals is precisely equal to the forward emf voltage between the terminals but around the external half-loop. The same current flows through both the forward voltage drop and the reverse voltage repotentialization. So precisely as much energy is dissipated to force the current through the back emf, as is recovered from the load.

Accept the conventional view that the shaft energy input to the generator is transduced into magnetic energy, which is then dissipated to force those charges back though the back emf. Now examine the energy flow in space surrounding the external circuit, using both the Poynting theory and Heaviside's theory. We have accounted for the Poynting diverged component, but we have not accounted for the huge remaining nondiverged Heaviside component. From whence comes this excess remaining energy flow that did not enter the circuit, was not included in the Poynting calculation, and was not expended in the circuit but was just wasted? To prove the excess energy flow really exists, perform the Bohren experiment {24} where the intercepting charges in the circuit are in resonance, and thus "sweep out" a larger geometrical cross section of interception of the impinging energy flow. These charges do sweep beyond the static unit charge cross section conventionally assumed in the definitions of field intensity (e.g., of the fields E and H in the Poynting vector S in S = (E x H). If the defining unit point charges are in resonance and sweep out a

Heaviside knew this in the 1880s. All the energy flow that is intercepted by the circuit and enters (i.e., the Poynting component) at essentially right angles, hardly even changes the angle of flow of the remaining huge flow of energy surrounding the circuit and missing it. Quoting Heaviside {153} directly:

> *"It [the energy transfer flow] takes place, in the vicinity of the wire, very nearly parallel to it, with a slight slope towards the wire.... Prof. Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from the vertical. This difference of a quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature of the electric field in the vicinity of a wire supporting electric current. The lines of electric force are nearly perpendicular to the wire. The departure from perpendicularity is usually so small that I have sometimes spoken of them as being perpendicular to it, as they practically are, before I recognized the great physical importance of the slight departure. It causes the convergence of energy into the wire. "*

So every electrical power system is rigorously a COP>1.0 "energy gating" system already, if the neglected, enormous Heaviside energy flow through space around the circuit (and missing it and wasted) is re-accounted. In that case, much greater energy pours out of the terminals of every generator, than the feeble amount of mechanical energy we input to the shaft.

We do not have an *energy availability* problem, and we never have had one! Instead, we have an *energy flow interception and collection* problem

greater cross section, then the assumed Poynting vector S, for the static case, changes in magnitude by some ratio k for the resonant case to the vector S_R, so now $S_R = k_1E \times k_2H = k_3(E \times H)$. Since k_3 is just the ratio of the actual geometrical cross section swept out by the charge to the standard geometrical cross section swept out by the static charge, then for a static charge $k_3 = 1.0$, and for a resonant charge $k_3 > 1.0$. Hence the Bohren experiment, with $k_3 = 18$ or so, produces 18 times as much collected (Poynting) energy "out" as we erroneously think we input by normal calculations ignoring the input Heaviside nondiverged component. The Bohren experiment {24} proves the argument; validation by Paul and Fischer {25} appears in the same journal issue.

— and we always have had one, due to the ubiquitous use of the closed current loop circuit.

Texts — e.g., such as by Kraus {154} — do show the Poynting (intercepted and collected) component of the EM energy flow surrounding the power line attached to the generator terminals, but none of them today go into the remaining nondiverged Heaviside component. Each of Kraus' contours {76} of energy flow in space, around those power line conductors, shows *only that part of the energy flow in space that is being drawn into the circuit.* It does not show the remaining huge energy flow that (i) is not intercepted, (ii) does not enter the circuit, and (iii) is wasted. Presently no texts illustrate this Heaviside nondiverged energy flow component, and no text accounts for it except to point out that the Poynting component must be considered to be indefinite, since it can be accompanied by an additional energy flow component. The texts then erroneously conclude that the excess flow can have no physical significance.[51]

One can easily prove the existence of additional energy flow to be collected, in addition to the Poynting component. E.g., simply examine the Bohren experiment {155} and its replication by Paul and Fischer {156} — or one can repeat the experiment oneself. The Bohren experiment outputs some 18 times as much energy as is input by normal "Poynting component" calculations alone, because it also intercepts additional energy from the unaccounted and ignored *free* Heaviside input component remaining and not diverged by a *static* unit point charge. Bohren simply places the charge in resonance, so it sweeps out a greater geometrical reaction cross section in the energy flow, thereby penetrating into the free — but arbitrarily discarded and unaccounted — Heaviside component, and intercepting and collecting additional energy. Bohren is perfectly aware that the excess energy caught by the resonant charges comes from the surrounding vacuum. However, he seems quite unaware of Heaviside's discovery and cautious statement of the excess energy involved in *every* field-particle interaction, and particularly in electrical circuits.

---

[51] E.g., J. D. Jackson, Classical Electrodynamics, Second Edition, Wiley, 1975, p. 237: *"...the Poynting vector is arbitrary to the extent that the curl of any vector field can be added to it. Such an added term can, however, have no physical consequences. "* Jackson errs in stating that the added energy flow vector cannot have any physical consequences. It cannot, of course, unless one catches some of it and uses it to power a load, without using half of the caught energy to destroy the source dipole faster than the load is powered. One of the major applications of it is to use it with Dirac sea negative energy 4-holes in the vacuum, where the holes produce negative energy fields and potentials in surrounding spacetime, acting back upon the system which is the source of the 4-holes to produce practical antigravity.

There are several such "white crows" which are legitimate COP>1.0 EM processes in physics. It only takes one small white crow to prove that not all crows are black. As another example, the papers of Letokhov {157a-157d} are especially recommended. So long as excess energy is received freely from the environment, the system is permitted to be in disequilibrium where COP>1.0.

So far as this author could discover, none of the perpetual motion machine critics have noticed another very simple fact: The conventional electrodynamics they utilize to refute COP>1.0 Maxwellian systems implicitly assumes that the source charge — which is any and every charge in the universe — continuously *creates* and pours out energy in 3-space without any energy input at all. Their own conventional classical electrodynamics contains a "perpetual motion machine assumption" of the most fundamental and grossest kind. The ardent skeptics of overunity EM systems — in using that implicit conventional assumption — are themselves the greatest perpetual motion advocates of all, though perhaps unwittingly. In 2000 we resolved that long-vexing problem[52] of the source charge {12}.

The more strident critics simply set up the "equilibrium thermodynamics" straw man, then knock it down and attack independent COP>1.0 experimenters with it, erroneously implying that all overunity researchers seek such in an equilibrium system. They fail to notice orthodox science's implied assumption that every charge and every dipole in the universe is already assumed to be a COP => inf. perpetual motion machine, freely *creating* all that continuous outpouring of energy from nothing. None of the critics have noticed and critiqued this greatest of all perpetual motion assumptions right in their own ranks, without a single exception, and the most ubiquitous. Undoubtedly that is because the scientific community honestly and freely admits that it has not heretofore solved that problem {68}, and also points out that the outpouring of energy is an experimental fact and therefore unquestionable.

The better critics — at least the rational ones such as Angrist — already admit that there exists a class of legitimate "fictitious perpetual motion machines" which are powered continuously without the operator inputting the energy. The phrase "fictitious perpetual motion machine" is an attempt to retain the classical thermodynamics straw man, when dealing with the thermodynamics of a system far from equilibrium in its energy exchange with its active environment. That phrase has been a non sequitur since the award of the Nobel Prize to Ilya Prigogine in 1977. We stress again our new use of all EM systems as open systems, in accord with general relativity.

---

[52] After publication of our paper, we also discovered very powerful support for our solution in quantum field theory. E.g., see F. Mandl and G. Shaw, *Quantum Field Theory*, Wiley, 1984/1993, Chapter 5.

So when rigorously analyzed from the breadth of physics and thermodynamics, and not just from a selected very narrow portion, the entire "perpetual motion" critical literature is inappropriate or — more exactly — is inappropriately focused upon a fictitious or false target. It is simply a classic example of reasoning from a false premise.

Not one such critic seems to have noticed that every generator and battery already pours out more energy than what is input to the shaft, and that it is easily proven experimentally via the thousands of experiments showing negative absorption from the medium. None seems to have bothered to read the original Heaviside, Poynting, and Lorentz papers to see how this startling fact was buried more than a century ago.

Finally, a self-appointed "expert" who does not even know what furnishes the energy to power every electromagnetic circuit, cannot be taken seriously in his "learned criticism" , regardless of what journal publishes it. If he continues to accept the unsolved source charge problem, implying that all charges freely create EM energy from nothing, he is not expert in any facet of powering electromag letic systems and circuits.

2.1.8 Prescribed Systems Versus Changes to Maxwell's Equations

As we stated, Maxwell's theory {20} was some 20 quaternion-like and vector equations in some 20 unknowns. By discarding the scalar component of the quaternion to retain only the vector subcomponent, and by discarding many of Maxwell's potentials, Heaviside {158} truncated the Maxwell equations to essentially the present familiar set of four vector equations - or in potential form, to two equations with variables unseparated.

While this truncation discarded a great deal of Maxwellian electrodynamics phenomenology, from a thermodynamics view the resulting *Maxwell-Heaviside* equations still contain and prescribe two types of Maxwellian systems:

(1) systems in equilibrium with their active environment, such as the active vacuum, and

(2) systems far from equilibrium with their active environment, such as the active vacuum.

Systems in the first class of Maxwell-Heaviside systems — i.e., in equilibrium in their energetic exchange with their active environment — rigorously obey classical equilibrium thermodynamics {159}. No system in that Maxwellian class can output more energy than the operator himself inputs to the system. The greatest coefficient of performance (COP) that these systems can accomplish, is COP = 1.0 for a theoretically perfect system without internal losses whatsoever, and thus with 100% efficiency of conversion of their input energy into work in the load. Since any real system has internal losses and thus is less than 100% efficient, the *practical* systems of this first class exhibit COP < 1.0.

These Maxwellian systems of the first class are the electrical systems always "understood" by ardent skeptics who proclaim that no other kind of Maxwellian system exists, that therefore no Maxwellian system can exhibit COP > 1.0, and that a claim to COP>1.0 is perpetual motion lunacy and not science at all. These fellows often launch scathing and unjustified personal attacks upon any serious researcher trying to work with that second class of permitted Maxwellian systems! It would be scientifically beneficial if these self-appointed defenders of the faith would simply study some more physics and thermodynamics, review the history of Maxwell's theory and how it was altered, and understand the well-established thermodynamics of systems far from equilibrium.

This false charge — that there can be no COP > 1.0 EM system because that would violate the laws of physics and the laws of thermodynamics — has been insidiously indoctrinated into generation after generation of physics and electrical engineering students. Today most of our electrical engineers and electrical scientists have absorbed it and believe it without question. Our power-engineering scientists believe it, and so advise the government. Hence our government believes it on the advice of our best scientists, such as the National Academy of Sciences and the National Science Foundation. Our great national laboratories also believe it and are just as mentally conditioned. The universities — especially the electrical engineering departments — also believe it, and provide a mainstream suppression of the sharp young graduate students and post-docs that would otherwise have long ago revised this mess and solved the energy crisis permanently.

Again we are reminded of that first quotation by Einstein, which we included at the beginning of Chapter 1, and which admonished us to ever examine the foundations of our science to root out and correct the errors. We are also reminded of another quote by Tolstoy:

> *"I know that most men, including those at ease with problems of the greatest complexity, can seldom accept even the simplest and most obvious truth if it be such as would oblige them to admit the falsity of conclusions which they have delighted in explaining to colleagues, which they have proudly taught to others, and which they have woven, thread by thread, into the fabric of their lives. "* [Count Leo Nikolaevich Tolstoy].

The second class of Maxwell-Heaviside systems — rigorously prescribed and permitted by Heaviside's truncated Maxwellian theory — is comprised of systems in disequilibrium with their active environment. A system in this class does not obey classical equilibrium thermodynamics. To the contrary, this system obeys the well-known *thermodynamics of systems far from equilibrium* {160a-160f} with their active environment — such as the active vacuum.

Such a system *is permitted to:*

(i) self-order,

(ii) self-oscillate or self-rotate,

(iii) output more energy than the operator himself inputs and thus exhibit COP>1.0 (the excess energy is freely received by the system from its external active environment),

(iv) power itself and its load simultaneously (all the energy is freely received by the system from its external active environment,[53] without any operator energy input being required), and

---

[53] We accent the concept of the supersystem. consisting of three components: (a) the system and its dynamics, considered as in normal electrical engineering (with a flat spacetime and an inactive local vacuum), (b) the active local vacuum and its dynamics, and (c) the active local curvatures of spacetime and their dynamics. All three components of the supersystem interact with each other. In conventional electrical engineering, the assumption in U(1) electrodynamics of a flat local spacetime and equilibrium with the local vacuum exchange eliminates any

(v) exhibit negentropy, which every charge and every dipole in the universe already does {12, 68}.

Let us now see why Maxwellian EM systems of the second class *are not being designed and built as such,* even though they are permitted by the laws of physics and reinterpreted thermodynamics, and even though they are prescribed in the Maxwell-Heaviside equations before further truncation by Lorentz {161a, 161b}. We may use the Maxwell-Heaviside equations in vacuum, because the energy flow intercepted by the external circuit pours out of the generator terminals and fills all space surrounding the power lines. It is therefore a transmission-reception problem, prior to the "potentialized current-transport" problem in the circuit.

2.2 Lorentz Regauging of the Maxwell-Heaviside Equations

2.2.1 Introductory remarks.

The domain of Lorentz's symmetrically regauged equations is only a small subset of the domain of the Maxwell-Heaviside equations they replace. Indeed, the later Lorentz {161a, 161b} symmetrical regauging discards an entire class of Maxwellian systems permitted by nature and by the Maxwell-Heaviside equations before they are symmetrically regauged.

Lorentz's symmetrically regauged EM equations discard all Maxwell-Heaviside systems that are far from equilibrium in their energetic exchange with the active vacuum. They retain only that subset of the Maxwell-Heaviside theory wherein the system being described is in forced equilibrium {162} in its exchange with its active vacuum environment. Hence the present Lorentz-Maxwell-Heaviside theory, by which EM circuits and electrical power systems are designed, produces only systems in self-forced equilibrium with the active vacuum, *specifically during the symmetrically regauging discharge* of the circuit's excitation energy. The closed current loop circuit in fact discharges half its collected energy to destroy the source dipole in the generator, thereby destroying that dipole's extraction from the vacuum and furnishing of the energy flow pouring out of the terminals. These circuits kill themselves faster than they can power the load, and they use more energy to kill the energy flow from the vacuum than they use to power their load. Hence all present EM systems

disequilibrium of the system with its active environment, hence arbitrarily discards all Maxwellian disequilibrium systems - precisely those that can exhibit COP>1.0.

rigorously conform to classical equilibrium thermodynamics, and exhibit a coefficient of performance (COP) of COP<1.0 since any real system also has losses {163}.

2.2.2 Technical regauging of the Maxwell-Heaviside equations {164}.

For asymmetrical-regauging (A-regauging) considerations, we are speaking of A-regauging the potential energy in and around a circuit. We include not only the Poynting energy flow component that is diverged into the circuit conductors, but also the remaining Heaviside *nondiverged* energy flow in space surrounding the conductors. This means that the energy is in field energy (E-field and B-field) form, both overtly as ordinary EM fields and covertly or "infolded" inside the corresponding scalar potentials {165, 166, 167, 168}, or both. Consequently, we must analyze Maxwell's equations as we would for radiating energy, rather than employ only the $\mathbf{j}\phi$ circuit analysis conventionally utilized, where the collected energy is sluggishly transported by the Slepian vector $\mathbf{j}\phi$. We show in this book that asymmetrical self-regauging (ASR) allows permissible overunity operation of electromagnetic engines and devices {169}, if other requirements are fulfilled also.

In Gaussian units, Jackson {170} shows that Maxwell's four equations (vacuum form) can first be reduced to a set of two coupled equations in the $(\mathbf{A}, \Phi)$ representation as follows:

$$\nabla^2\Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet \mathbf{A}) = -4\pi\rho \qquad [15]$$

$$\nabla^2 \mathrm{A} - \frac{1}{\mathrm{c}^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla(\nabla \bullet \mathbf{A} + \frac{1}{\mathrm{c}}\frac{\partial \Phi}{\partial t}) = -\frac{4\mathrm{p}}{\mathrm{c}}\mathbf{J} \qquad [16]$$

The result is two coupled Maxwell equations rather than four. Jackson shows that potentials A and Φ in these two equations are arbitrary in a specific sense, since the A vector can be replaced with $\mathbf{A'} = \mathbf{A} + \nabla\Lambda$, where A is a scalar function and VA is its gradient. The B field is given by $\mathbf{B} = \nabla\times\mathbf{A}$, so that the new B' field becomes

$$\mathbf{B'} = \nabla\times(\mathbf{A} + \nabla\Lambda) = \nabla\times\mathbf{A} + \mathbf{0} = \nabla\times\mathbf{A} = \mathbf{B} \qquad [17]$$

In other words, the B field has remained entirely unchanged, even though the magnetic vector potential has been asymmetrically changed. However, if no other change were made, then the electric field E would have still been changed because of the gradient $\nabla\Lambda$. In that case the net change would be asymmetrical, because one obtained a "free" E-field which could then do work on the system — either beneficially or detrimentally, depending upon the specific conditions, geometry, and timing. To prevent this excess "free" E-field from appearing, the electrodynamicists simultaneously and asymmetrically regauge (transform) the scalar potential Φ so as to offset the E-field change due to the regauging of equation [15]. In short, they also change Φ to Φ',, where

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial \Lambda}{\partial t} \qquad [18]$$

With that additional change, now the net E and B fields remain unchanged {171, 172}, even though both potentials have changed and the fundamental stored energy of the system has changed, as has the stress of the system. Unchanged force fields just mean that only a set of zero-summation forces (a zero-summed stress system) has been utilized to effect the change in potential energy. It also means that the net summation of the two *asymmetrical* regaugings has been entirely *symmetrical* {173}.

Jackson points out that, conventionally, a set of potentials $(\mathbf{A}, \Phi)$ is habitually and *arbitrarily* chosen by the electrodynamicists such that

$$\nabla \bullet \mathbf{A} + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [19]$$

This *net symmetrical* regauging operation creates a new and simpler Maxwellian system, with different system stress and different system potential energy. It successfully separates the variables, so that two inhomogeneous wave equations result. This procedure yields a new and simplified system, and the new Maxwell's equations for it are as follows:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [20]$$

$$\nabla^2 \mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} = -\frac{4\pi}{c}\mathbf{J} \qquad [21]$$

The two previously coupled Maxwell equations [15] and [16] (potential form) have been replaced by equations [20] and [21], to leave two much simpler inhomogeneous wave equations, one for Φ and one for A. These are *new equations for a new system!*

Of course, this arbitrary *net symmetrical* regauging (let us use the term ***S-regauging)*** is quite useful for purposes of simplifying the theory and for easing calculations. But its unquestioning and rather universal usage has arbitrarily eliminated the freedom of the system designer to *asymmetrically* regauge the system's potentials, and *use* the resulting excess free force fields to change the stored energy in the system without the operator performing extra work upon the system. So we advance the condition for violating this S-regauging, violating the exclusion of net A-regauging, and violating the "frozen gauge" process as

$$\nabla \bullet \mathbf{A} + \frac{1}{c}\frac{\partial \Phi}{\partial t} \neq 0 \qquad [22]$$

Any regauging of the potentials that complies with equation [22] will *a priori* produce one or more net excess forces in the system, as well as a change in potential energy of the system. By controlling the regauging, the system designer is then able to control where, how, and when these excess forces appear, and whether they enhance the system's operation or hinder it. These net forces can then be used to perform work and dissipate the excess potential energy taken on in the asymmetrical regauging. That is what we do when we ourselves add potential (and potential energy) to an EM system to enable it to do work! *If we always have to asymmetrically regauge the system to get it to do subsequent work, why do we not just let the system asymmetrically regauge itself so we get the input energy freely and also get the resulting work freely?* The gauge freedom axiom in quantum field theory assures us that nature will indeed freely change the potential energy of any system for us, if we but arrange it and permit it.

In short, we have had a gauge freedom principle for some time, which guarantees us that COP>1.0 EM systems are permitted and possible. Yet we have failed to realize it, or take any advantage of it. So we continue to pay to asymmetrically regauge (potentialized) all electrical power systems, and to insure that we have to continue to do it, we specifically design the systems so they will then re-enforce Lorentz's symmetrical regauging condition.

This is another of those "inexplicable aberrations of the scientific mind" referred to by Nikola Tesla!

2.2.3 A Humorous Comment but an Exact Analogy

Again see Figure 1-1 in Chapter 1, and see Figure 2-3. In Figure 2-3, we show how Lorentz's integration trick {174a, 174b} discarded the huge nondiverged Heaviside component of energy flow outside the conductor, while retaining the small Poynting component that strikes the surface charges and gets diverged into the conductor to power the electrons. In justifying his integration trick, Lorentz stated that all the rest of that wasted Heaviside energy flow

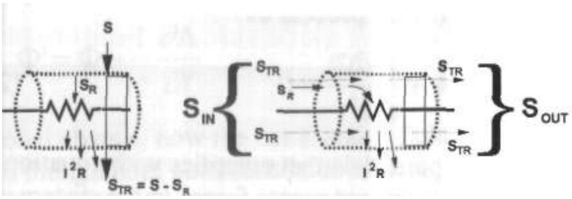

See Panofsky & Phillips
Classical Electnqtv and Magnetism. 2nd edn
Addison Wesley 1962.
p 178-181

The Lorentz procedure arbitrarily discards the enormous Heaviside component that misses the circuit entirely and is wasted This results in a non sequitur of first magnitude in EM energy flow theory

2-3a. Lorentz surface integration **2-3b. Actual S in and S out.**

Figure 2-3 Lorentz's integration trick to discard the enormous Heaviside non-diverged energy flow component

The Heaviside component is often 10 trillion times the Poynting component but is simply wasted in ordinary single-pass energy flow circuits

Heaviside component (nondiverged)

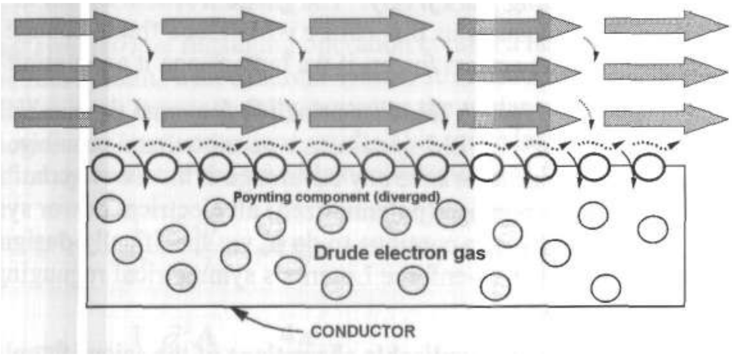

Figure 2-4 Heaviside nondiverged and Poynting diverged energy flow components.

(Figure 2-4) was "physically insignificant" (his term) because it did not strike the circuit, was not intercepted, and did not power anything. Well, Maxwell's theory is a material fluid flow theory, so let us evaluate Lorentz's justification in a fluid flow analogy.

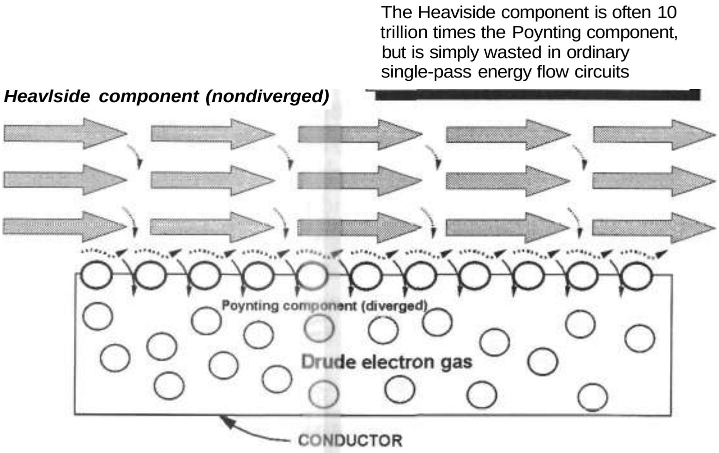
The Heaviside component is often 10 trillion times the Poynting component, but is simply wasted in ordinary single-pass energy flow circuits
Heavlside component (nondiverged)
Poynting component (diverged)
Drude electron gas
CONDUCTOR

Figure 2-4 Heaviside nondiverged and Poynting diverged energy flow components.

Contrast Lorentz's statement if applied to the ocean wind on a sailing ship, where the wind represents the total energy flow and the ship represents a single Maxwellian system under consideration.

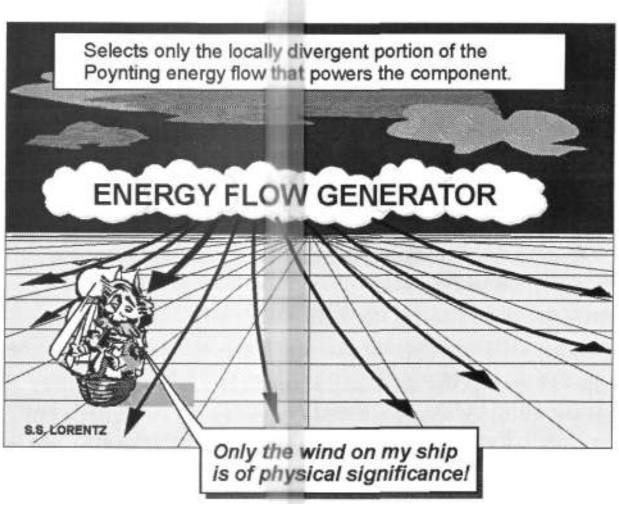
Selects only the locally divergent portion of the Poynting energy flow that powers the component.
ENERGY FLOW GENERATOR
S.S. LORENTZ
Only the wind on my ship is of physical significance!

Figure 2-5 Lorentz in his sailing ship in an ocean wind, with no other ships around.

As shown in Figure 2-5, Lorentz's argument is analogous to stating that the huge component of the wind *that does not* strike the ship's sails is "physically insignificant" — in other words, it has no physical use. Now that is true for the single ship under consideration, as shown in Figure 2-5. None of that wind component that misses the ship's sails does anything at all to power *that ship* (that individual system). Hence it is indeed physically insignificant *to that ship itself.* Of course, if we flare the sails wider, the ship will catch more wind and develop more power. That is an analogy to the Bohren experiment, which resonates the charges (the "sails" intercepting the incoming EM energy flow "wind") so that they sweep out a greater geometrical cross section. Hence 18 times as much energy is caught, as when the charges are not resonated (the sails are not flared to a greater area).

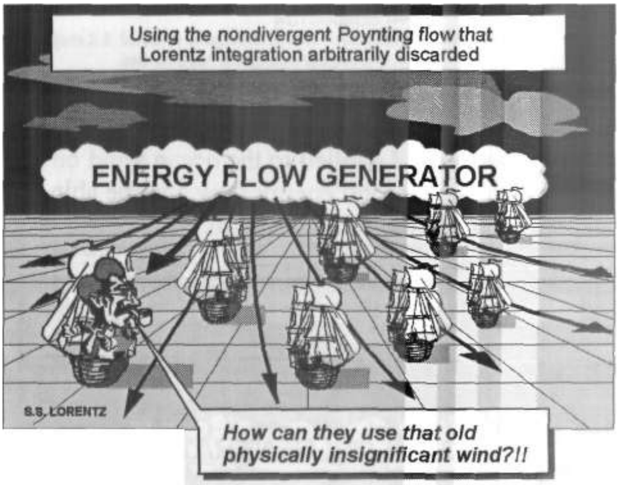
Using the nondivergent Poynting flow that
Lorentz integration arbitrarily discarded
ENERGY FLOW GENERATOR
S.S. LORENTZ
How can they use that old
physically insignificant wind?!!

Figure 2 6 Lorentz in his sailing ship in an ocean wind, with many other ships alongside.

However, now see Figure 2-6. If we put some more ships (some more systems) with "intercepting and collecting sails" in there, they will be powered quite well with some of that "physically insignificant" wind. So Lorentz (following earlier work by Ludwig Lorenz) really did advance and impose upon electrodynamics a non sequitur of first rank.

And the wind-fluid analogy helps to "lighten things up" a bit here and inject a little good-natured humor!

2.2.4 Discussion

Equation [22] has now become the first *necessary but not sufficient* principle of asymmetrical self-regauging (ASR) for overunity electromagnetic engines and devices. That is, we are now free to A-regauge in a *nonconservative* field manner, where excess net force fields automatically appear and remain for our use in the system. All that remains is for the designer to:

(i) evoke "nonconservative" (asymmetrical) regauging in accord with equation [22];

(ii) have the system assembly in disequilibrium with its energetic external environment, so that the system itself performs this nonconservative A-regauging in some sector or part {175} of its normal operation,

(iii) insure that any "free force" that is produced and allowed to do work upon the system will do some positive work upon the system, so as to increase its collected energy, and

(iv) assure that the "free" forces that are otherwise produced by the circuit design (this includes, e.g., the back emf) do not seriously degrade the system (i.e., do not substantially reduce its energy available to dissipate as work in the load). Usually this means that, contrary to the standard closed current loop circuit, the circuit must not dissipate half its collected Poynting energy to dissipate the charges in the source dipole and destroy it. *These four considerations we now advance as the primary or master principles of asymmetrical self-regauging of electromagnetic engines and devices for permissible overunity efficiency.* We will meet additional principles in a later chapter when we discuss what the form of the energy output of an overunity EM system is, and the considerations for close-looping a COP>1.0 EM system into stable disequilibrium and continuous self-powering.

It is understood that the asymmetrical regauging portion of the circuit or system must be a broken symmetry in the energetic exchange between the active vacuum environment and that portion of the system. We exclude, g., solar cell-powered systems, windmill-driven generators, etc.

For simplicity, the system designer may wish to A-regauge only one potential, producing a desired "free" force F_1 to work on the system and increase its energy, while allowing any other extraneous force F_2 to appear, but just "bottling up" that additional force F_2 so that it cannot do any work at all on the system and therefore cannot degrade the system's operation. A most useful ASR corollary that immediately suggests itself is to simply A-regauge onto that force operation that normally produces the back-drag in a system or device, to either *eliminate* the back-drag altogether or reverse its algebraic sign so that it becomes an *assisting* force instead of a *hampering* force {176}.

In overunity systems, this becomes a startling and unexpected exercise in novel new phenomenology effected by tempic potential differences, tempic force, etc. It involves not only the system, but also the entire *supersystem* consisting of the physical system, its associated altered local vacuum, and its associated curved local spacetime. See Chapter 9 for our expose of these phenomena, and how one must transduce[54] the effects of a very novel kind of tempic force back emf, if the COP>1.0 system is to remain in disequilibrium with the active vacuum and thus maintain its stable COP>1.0 operation.

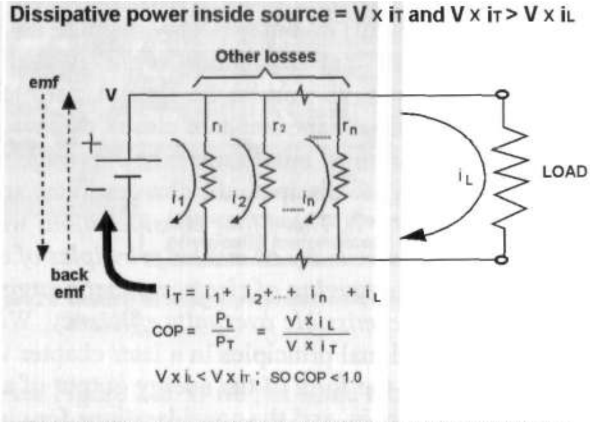
Dissipative power inside source = V x iT and V x iT > V x iL
Other losses
emf
V
+
−
r1
r2
rn
i1
i2
in
iL
LOAD
back emf
iT = i1+ i2+...+ in + iL
COP = PL / PT = V x iL / V x iT
V x iL < V x iT ; SO COP < 1.0

Figure 2-7 Why a conventional closed current loop circuit is limited to COP<1.0.

Again see Figures 2-3 and 2-4, and equations [20] and [21]. Electrodynamicists have previously *assumed away* the capability of work-free "refueling" of an electromagnetic system directly from the local vacuum {177}. They have deliberately designed their systems *to prevent self-evoking and positively utilizing asymmetrical self-regauging of their own potential energy during system excitation discharge.* (See Figure 2-7).

---

[54] A patent application on the process has been filed by John Bedini and the present author.

In fact, by forcibly pumping the ground electrons from the external circuit back up through the primary source dipole in the generator or battery, our closed current loop circuits self-enforce the Lorentz condition and require that precisely half the regauging energy freely intercepted and collected in the external circuit is used to destroy the source dipole providing the regauging energy. This destruction of the energy source is accomplished faster than the external load is powered. Hence all such systems with losses are COP<1.0 systems. For more than 100 years, this "self-destroying, suicidal system" is the only kind of electrical power system our energy scientists and engineers have designed and built. To inject a little humor, we can only describe this in Tesla's words[55] as *"one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history."*

Thereby electrodynamicists have narrowed the inherent flexibility of the Maxwellian EM gauge theory model to an artificially symmetrized subset of Maxwell-Heaviside-Lorentz equations that utilize a single fixed gauge. This prohibits the free use of the system's vacuum energy freely received by the source dipole's broken symmetry in its violent exchange with the active vacuum. If the system increases its dipolarity (regauges freely), it simultaneously increases the speed at which it destroys the dipolarity! *Maxwell's equations themselves do not require this, and neither does Heaviside's truncation of them into the resulting Maxwell-Heaviside equations. Lorentz's symmetrical regauging of the Maxwell-Heaviside equations does require it.*

The Maxwell-Heaviside theory without Lorentz symmetrical regauging allows regauging asymmetrically to place the system in disequilibrium so that it freely receives and utilizes excess energy from the surrounding vacuum. Then it can again asymmetrically regauge by dissipating that freely received energy in an external load. *The theory specifically does not require destroying the primary source dipole as is done in the arbitrary closed current loop circuit between load and primary source dipole in the generator or battery.*

---

[55] Nikola Tesla, "The True Wireless, *Electrical Experimenter,* May 1919.

By arbitrary habit, electrodynamicists have discarded the Maxwell-Heaviside major EM overunity mechanism: *Asymmetrically self-regauging accomplished by the system itself, without requiring external work being done on the system by the operator,* in order to freely gather in and collect excess potential energy in the system from its broken symmetry with its active vacuum environment. In this self-A-regauging, associated *excess net force* does appear, whereupon that excess collected/stored regauging energy can then be used to "drive" the forces against load resistance and system losses to freely power the system and its loads {178}. They have also discarded the major overunity corollary whereby the designer is free to utilize *system self-regauging asymmetrically so that the normal back-drag or back-emfwork — usually done upon the system in its closed current loop configuration to reduce its kinetic energy — vanishes or reverses sign and becomes beneficial work done in the opposite sense, to freely increase the energy ofthe system.*

The failure to do this is bluntly met with, when one does succeed in producing a COP»1.0 system momentarily by suddenly and very powerfully breaking the system's equilibrium in its vacuum energy exchange. Nature then exhibits a very novel kind of Dirac sea hole reaction (powered by tempic force generated in the locally curved spacetime and resulting effects generated in the altered semiconducting vacuum). This reaction sharply jerks the system (or the system plus its external power supply) back into equilibrium and COP<1.0 operation. In Chapter 9, apparently for the first time in the literature[56] we present this novel mechanism exhibited by nature as a new kind of "tempic back emf phenomenon in systems in disequilibrium. We also explain how to handle the problem and in general how to use it for "freeze-framing" and stabilizing the system's disequilibrium operation in sustained COP > 1.0 configuration while powering itself and its load.

All that is required for "self-powering" COP»1.0 EM engines directly powered by the vacuum, is that (1) initially the engine must asymmetrically self-regauge to produce COP»1.0, and (2) positive *transduced* feedback (from the output) of a very special and previously unknown kind — where negative energy is transformed into normal electron current and energy — must be utilized to its input, in a clamped and controlled fashion {179}. The clamped positive feedback can be designed and applied, once the fundamental engine is designed for self-regauging and overunity efficiency, and once the novel feedback and transformation requirements are understood and incorporated. The remaining excess output can then be utilized to power loads, while the engine seemingly "powers itself."

---

[56] The basis for it, however, is in Dirac's electron theory published in 1930, nearly three quarters of a century ago. See P. A. M. Dirac, "A theory of electrons and protons," *Proceedings of the Royal Society of London, Series A,* 126(801), Jan. 1, **1930, p. 360-365.**

We stress, of course, that this system is just analogous to the windmill, and it continuously receives a "free energy wind" from its active environment. The conservation of energy law is obeyed at all times, as are physics and thermodynamics.

In a later chapter we present the concepts and processes for these novel phenomena involved in close-looping a COP>1.0 system for self-powering operation.

2.3 The Suicidal Closed Unitary Current Loop Circuit

The present Lorentz-Maxwell-Heaviside theory, by which EM circuits and electrical power systems are designed, produces only systems in net usable equilibrium with the active vacuum, specifically during the symmetrically regauging discharge of the circuit's excitation energy (Figure 2-7). Hence all present EM systems designed by that model rigorously conform to classical equilibrium thermodynamics, and exhibit a coefficient of performance (COP) of COP<1.0.

Since the Maxwell-Heaviside equations do not restrict the designer to systems in equilibrium with their active environment, some ubiquitous characteristic of the standard circuit approach must be self-enforcing the Lorentz symmetrical condition as far as using any excess free potential energy of the system to freely power loads. That is the closed-loop circuit, as we have discussed. Let us now illustrate what goes on in such a circuit.

Taking into account Figure 2-4, Figure 2-8 shows a modified diagram of what is actually happening between the vacuum and the dipole to power the circuit. Figure 2-9 shows how a pair of conductors constitutes an extended series of parallel dipoles. Once the dipole is formed by dissipating some chemical energy in the battery, the dipole extracts energy from the surrounding vacuum and pours it out of the battery terminals, producing an energy flow filling space around the conductors of the external circuit. Only a very small portion (Figure 2-4) of the very large energy flow is intercepted by the external charges in the conductors, and diverged into the conductors to power up the Drude electrons.

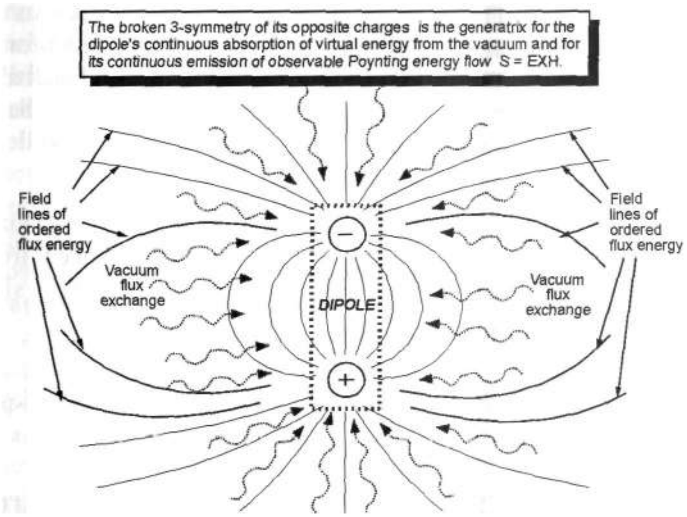
The broken 3-symmetry of its opposite charges is the generatrix for the dipole's continuous absorption of virtual energy from the vacuum and for its continuous emission of observable Poynting energy flow S = EXH.
Field lines of ordered flux energy
Field lines of ordered flux energy
Vacuum flux exchange
Vacuum flux exchange
−
DIPOLE
+

Figure 2-8 Virtual photon flux exchange intensity **with the charges of a** dipole.

As can be seen from Figure 2-7, half of the energy collected upon the internal electrons to drive them as **jϕ** is expended as work required to pump the spent electrons from the ground return wire back through the source dipole in the battery. This scatters the charges of the dipole (in this case, by changing the chemistry of the battery plates) and destroys the dipole. The other half of the collected energy is dissipated in the external circuit's losses and in the load.

Then some more chemical energy has to be dissipated by the battery to perform work upon the internal charges and restore the dipole (in this case, to change the plate chemistry of the battery plates). So even if this process is 100% efficient, the battery has to dissipate as much chemical energy as was electrically dissipated upon those former dipole charges to scatter them and destroy the dipole.

In Figure 2-7, we show the small internal resistance of the conductors, as well as the load resistor. Half the energy dissipated by **jϕ** is expended just to forcibly ram the spent electrons back through the back emf of the source dipole. The other half is expended in the circuit's internal losses and in the load resistance. Therefore, less energy is expended as usable work in the resistive load than is expended to destroy the dipole and cut off the flow of free EM energy from the vacuum.

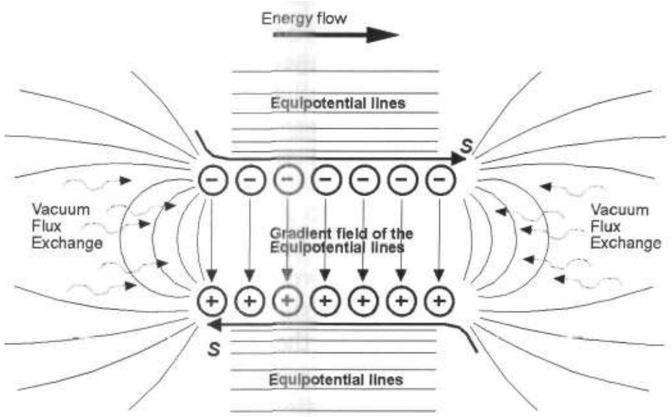

Figure 2-9 Paired conductors as an extended set of dipoles.

The battery has to input as much energy to restore the dipole as was used to destroy it. The battery has to expend at least as much chemical energy as equals half the entire energy originally collected in the external circuit and expended as $\int j(t)\phi(t)dt$. It must expend chemical energy $CE \geq \frac{1}{2} \int j(t)\phi(t)dt$. The energy RE dissipated as work in the load resistor R is $RE < \frac{1}{2} \int j(t)\phi(t)dt$. Thus CE>RE, and this system is forced to exhibit

$$COP\ ^{\circ}\ \frac{RE}{CE} = \frac{< \frac{1}{2}\grave{o}\, j(t)f(t)dt}{{}^{3}\ \frac{1}{2}\grave{o}\, j(t)f(t)dt}\ Þ\ COP < 1.0 \qquad [23]$$

As can be seen, this system symmetrically regauges itself during its excitation discharge, so that it self-enforces the Lorentz condition during that period. The circuit self-enforces COP<1.0, even though all the energy — both the energy powering the external circuit and also destroying the internal source dipole in the battery to enforce the Lorentz condition during discharge — is freely received from the active vacuum via the broken symmetry of the source dipole,[57] once established, in its energetic exchange with the active vacuum.

---

[57] The broken symmetry of opposite charges is one of the broken symmetries predicted by Lee, {13a} and by Lee, Oehme, and Yang, {13b}. Broken symmetry was proven experimentally in early 1957 by Wu *et ai,* {14}. So profound were the implications for physics that the Nobel Prize was awarded to Lee and Yang in latter 1957, the same year! E.g., see T. D. Lee, "Weak Interactions and Nonconservation of Purity," Nobel Lecture, Dec. 11, 1957 in *T. D. Lee, Selected Papers,* Gerald **Peinbcrg,** Ed., Birkhauscr, Boston, 1986, Vol. 1, p. 32-44.

The problem is that *the source dipole is destroyed faster than the load is powered,* as shown in Figure 2-7. Because of the inexplicable scientific pre-occupation with the conventional closed current loop circuit, the planet and biosphere have been increasingly polluted for more than a century, by burning more hydrocarbons, building more dams, using more nuclear fuel rods, etc. to provide ever more electrical energy.

2.4 In Summary

COP>1.0 operation in EM systems via work-free *asymmetrical* regauging has been inherent in the Maxwell-Heaviside equations for more than a century. Since Lorentz first did so circa the latter 1880s, electrodynamicists have just conveniently and arbitrarily *assumed away* COP>1.0 systems by (i) artificially limiting the theory to only that fraction of Maxwell-Heaviside systems in usable net force field equilibrium in their vacuum exchange during their excitation discharge, and (ii) using the ubiquitous closed current loop circuit to self-enforce the Lorentz symmetry condition during that excitation discharge. In other words, our engineers have only been building systems which self-enforce the overly restricted Lorentz-regauged theory.

Electrodynamicists have never seriously tried to build overunity circuits, nor until recently have they explored the discarded Maxwellian systems by removing the Lorentz regauging condition and examining EM energy from the vacuum.[58]

[58] Under the directorship of Dr. M.W. Evans, the Alpha Foundation's Institute for Advanced Study (AIAS) has been gradually developing the physics of EM systems far from equilibrium with their active environment (the local active vacuum and local curved spacetime). Some of the papers of interest are: M. W. Evans et al., "Derivation of the Lehnert field equations from gauge theory in vacuum: Space charge and current," *Found. Phys. Lett,* 13(2), Apr. 2000, p. 179-184; Evans et al., {38b}, {38c}, {113b}, {233a}, {233b}, {233e}; {233g}, {233j}; Evans et al., "Schrodinger Equation with a Higgs Mechanism: Inherent Vacuum Energy," *Found. Phys.* (in review); — "Vacuum Energy Flow and Poynting Theorem from Topology and Gauge Theory," *Physica Scripta* (in review); — "Energy from the Vacuum," *Physica Scripta* (in review); — "Some Notes on Asymmetric Regauging", *J. New Energy* 4(3), Special Issue, Winter 1999, p. 325-326; — "Inconsistencies of the Maxwell-Heaviside Theory of Electrodynamics: The Aharonov-Bohm Effect," *J. New Energy,* 4(3), Special Issue, Winter 1999, p. 236-240; — "Spontaneous Symmetry Breaking as the Source of the Electromagnetic Field," *Found. Phys. Lett.* (in press), A monumental work. *Modern Nonlinear Optics,* Second Edition, 3 vols., ed. M. W. Evans, Wiley, 2001 gives a broad coverage of higher group symmetry electrodynamics and energy from the environment (the local active vacuum and the

Every electrical power system ever built is and has been powered by electrical energy freely extracted from the seething vacuum exchange with the source dipole in the system, due to that dipole's broken symmetry. But the engineers and scientists have designed and built only those electrical systems which self-enforce Lorentz's symmetrical regauging.

The time for energy scientists and engineers to correct this procedure and cease this inane practice is long overdue, following the direction pointed out by the AIAS.

---

local curved spacetime). Some 60 AIAS papers are published as a Special Issue of **the** *J, New Energy,* 4(3), Winter 1999. More than 100 AIAS papers are also carried on DoE restricted website http://www.ott.doe.gov/electromagnetic/.

Chapter 3

Giant Negentropy, Dark Energy, Spiral Galaxies and Acceleration of the Expanding Universe

"*... the discoveries made in 1957 established not only right-left asymmetry, but also the asymmetry of the positive and negative signs of electric charge. In the standard nomenclature, right-left asymmetry is referred to as P violation, or parity nonconservation. The asymmetry between opposite signs of electric charge is called C violation, or charge conjugation violation, or sometimes particle-antiparticle asymmetry.* " [T. D. Lee]. {180}

"*The ends of a dipole, being oppositely charged, exhibit charge conjugate symmetry violation. An 'isolated' charge, considered with its clustering virtual charges of opposite sign, also exhibits symmetry violation.* " [T. E. Bearden, this book, below].

"*A generally acceptable, rigorous definition of radiation has not as yet been formulated." ... "The recurring question has been: Why is it that an electric charge radiates but does not absorb light waves despite the fact that the Maxwell equations are invariant under time reversal?* " [B. P. Kosyakov]. {181}

"*It [the energy transfer flow] takes place, in the vicinity of the wire, very nearly parallel to it, with a slight slope towards the wire... . Prof Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from the vertical. This difference of a quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature of the electric field in the vicinity of a wire supporting electric current. The lines of electric force are nearly perpendicular to the wire. Their departure from perpendicularity is usually so small that I have sometimes spoken of them as being perpendicular to it, as* *they practically are, before I recognized the great physical importance of the slight departure. It causes the convergence of energy into the wire. "* [Oliver Heaviside]. {182}

"The funny thing about the dark matter is it doesn't show up in light, X-rays, gamma rays, infrared, ultraviolet, radio waves, or submillimeter. We're running out of places to look for it." [William Keel]. {183}

"... baryonic, ordinary matter — the stuff of stars and of people — makes up just over 4% of the energy and matter in the universe. ...about 30% of the stuff in the universe is dark [unobserved] matter. The remaining two-thirds, theorists believe, is a mysterious 'dark energy' or 'quintessence' — a large-scale antigravity-like effect that is making the universe expand ever faster.... " [Charles Seife]. {184}

"And yet, curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."[M. P. Silverman]. {185}

*"**Charge** is the ongoing circulation of EM energy flow between the time and 3-space domains, as seen by the observer. Negative charge is the ongoing absorption of EM energy input from the time domain (from ict), transduction of the absorbed energy into 3-space EM energy, and re-emission of the EM energy in all directions in 3-space. **Positive charge** is the ongoing absorption of EM energy input from 3-space, transduction of the absorbed energy into the time domain, and re-emission of the EM energy in the time domain. If we wish, we may also consider the positive charge as radiating negative energy, while the negative charge radiates positive energy. Once we consider vacuum polarization, we recognize that this circular flow process is involved at each "dipolar point" of the polarized vacuum. "* [T. E. Bearden, this book, below].

"The matter density of the Universe is extremely low. On average there might be one atom per cubic metre of space. The major constituent of the Universe is believed to be some kind 'dark energy', which is pushing the Universe apart. " [Matthew Colless]. {186}

3.1 Broken 3-Symmetry of the Dipole

To recap from Chapters 1 and 2: In 1956, Lee and Yang {187} strongly predicted broken symmetry in physics, suggesting experiments to detect it in the weak interaction. In early 1957, Wu *et al.* {188} experimentally demonstrated broken symmetry, thus furnishing the experimental substantiation of broken parity in the weak interaction. The emergence of broken symmetry was such a dramatic change to the view of physics, that in December of the same year, 1957, Lee and Yang were awarded the Nobel Prize.

One of the asymmetries shown by Lee *et al.* is the asymmetry of opposite charges. That means that a dipole, which consists of slightly separated opposite charges, exhibits broken symmetry in the seething vacuum virtual particle flux. In short, a dipole — any dipole or dipolarity — must absorb or receive virtual photon energy from the active vacuum, integrate at least some of it to observable state, and re-radiate that integrated EM energy in observable form in 3-space.

We stress: *The ends of a dipole, being oppositely charged, exhibit broken symmetry. An 'isolated' charge, considered with its clustering virtual charges of opposite sign, is a set of composite dipoles and also exhibits broken symmetry.*

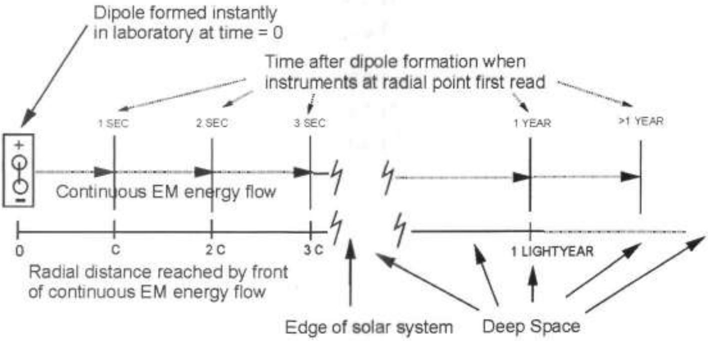
Dipole formed instantly in laboratory at time = 0
Time after dipole formation when instruments at radial point first read
1 SEC
2 SEC
3 SEC
1 YEAR
>1 YEAR
+
-
Continuous EM energy flow
0
C
2 C
3 C
1 LIGHTYEAR
Radial distance reached by front of continuous EM energy flow
Edge of solar system
Deep Space

Figure 3-1 The source dipole experiment, demonstrating continuous energy flow.

Let us now consider a gedanken experiment, as shown diagrammatically in Figure 3-1, to illustrate Kosyakov's quotation cited above. At a point in the laboratory corresponding to the origin of a 3-dimensional frame, a dipole is produced instantly by separating charges, such as in a flash-charged capacitor. Along two radial lines in 3-space from that origin, we have previously placed perfect EM field and potential detectors, at distances *ct* apart, where *t*= 1, 2, 3, 4, *n*,... and so on in seconds, with the radials extending across the universe, and where c is the speed of light. We neglect the transient pulse that momentarily occurs at each instrument package, since we are examining steady state conditions.

At the end of one second, the instruments at *t* = 1 instantly read, and those readings remain thereafter. One second later, the instruments at *t* = 2 instantly read, and those readings remain thereafter. And so on with each of the instruments progressively farther from the origin reading at the proper light-time travel, and with all field readings then *remaining continuously* thereafter.

This gedankenexperiment demonstrates that what is emitted from the dipole, from its moment of creation at the origin, is a continuous flow of EM energy in all directions in 3-space.[59] This energy continuously flows from the dipole, indefinitely or as long as the dipole physically exists. Other than paying once to initially separate the charges and make the dipole, we have neither added nor input any additional energy. Yet energy continues to flow out from the dipole in all directions, unceasingly, as shown by the continued of the reading on each field energy density instrument, once the leading edge of the energy outpour reaches that instrument and passes it by. In short, the experiment shows that it is not a "pulse" that passes each instrument, but a continuous and sustained flow of EM energy, remaining steadily thereafter so long as the dipole remains.

One year after making the dipole, the instruments at one light year distance suddenly read, and their readings remain. *The energy density of a volume of space one light year in radius has now been changed, and that change in spatial energy density continues to spread outwards at light speed.*

---

[59] More exactly, what is observed is a spreading change in the energy density and point-like polarization of the vacuum itself, and of the local curvature of spacetime. At each point, once the change of energy density and ST curvature occur, the energy density and ST curvature remain fixed in a new equilibrium condition, steadily fed by the energy radiating from the dipole. See our point made in {181}. The discovery of giant negentropy {12} requires a change in how we view the very process of energy propagating in space. Now it has become a quite different *propagation of energy circulation between time and space domains at each affected point in space.*

Yet there has been no input of *observable 3-space* EM energy flow to the dipole whatsoever, as our instruments also verify. All observable energy flow in 3-space is *outward* from the origin where the dipole is located.[60]

We have shown that the *observable* EM energy flow from a source dipole is not conserved in 3-space, which is consistent with a broken 3-symmetry since symmetry implies conservation. Either we must discard the conservation of energy law altogether, or it must be conserved in 4-space since we have proven the dipole does not conserve energy in 3-space. If we assume that the experiment must conform to the conservation of energy law, then 4-symmetry EM energy flow must exist since 3-symmetry energy flow is broken. The energy steadily pouring out of the dipole in all directions in 3-space must therefore be received by the dipole *charges from the time-domain,*[61] assuming we are working in Minkowski 4-space. In particle physics, this is essentially the same as receiving the energy from the virtual domain, and therefore is consistent with the proven broken symmetry of every dipole and dipolarity.

The energy flow is indeed received from the time domain, as we have argued {189} in a wave manner consistent with (i) quantum field theory as shown by Mandl and Shaw in the photon model {190}, (ii) broken charge conjugation symmetry of particle physics using the virtual photon and observable photon concepts, and (iii) a re-interpretation of Whittaker's decomposition of the scalar potential {191} when applied to the scalar potential between the ends of the dipole {12}.

We state the obvious: A dipole is a Maxwellian system. This gedanken experiment also proves by demonstration that a Maxwellian system — e.g., a dipole — is permitted by the laws of nature, physics, and thermodynamics to output more energy than the operator inputs, since we input nothing to the source dipole[62] once we have made it. After paying to produce the dipole initially, the operator has made no other energy input to the dipole at all, yet the dipole will continue to pour out observable energy in 3-space for the next 14 billion years if the dipole remains and is not destroyed. In addition, every positive charge and every negative charge in the universe form a dipole, regardless of the separation distance between them. That dipole is also performing the gedanken experiment process. Every dipolar EM circuit contains innumerable such dipoles — e.g., across "open paths" between the various points in the circuit, as shown by Kron {192}. Kron {193} stated:

---

[60] Obviously, we are utilizing the prevailing notion of the flow of energy through 3-space, which is actually a non sequitur. No 3-space observable entity can "move" in 3-space alone; it requires time to move, and so any movement is in 4-space a priori. We previously discussed this and the *dldt* operator being applied by the observation process. But for simplicity, we are using the conventional view here, so the negentropy of the dipole can be easily grasped.

[61] Again, we have stated the positive time case only. In the spreading special 4-circulation, energy appears in 3-space at each point dipole from the time domain (courtesy of the action of the negative charge of the dipole), and then returns from 3-space back to the time domain (courtesy of the action of the positive charge, which is a time-reversal of the action of the negative charge).

> *"... the missing concept of "open-paths" (the dual of "closed-paths") was discovered, in which currents could be made to flow in branches that lie between* *any* *set of two nodes. (Previously - following Maxwell - engineers tied all of their open-paths to a single datum-point, the 'ground'). That discovery of open-paths established a second rectangular transformation matrix... which created 'lamellar' currents... " "A network with the simultaneous presence of both closed and open paths was the answer to the author's years-long search."*

By considering all dipoles in the universe, one has in fact finally defined Kron's open path. One has also defined the master mechanism responsible for the incredible EM energy density of the vacuum — the so-called "zero-point EM energy". Overall, every electrical circuit thus has innumerable energy flow broken 3-symmetries in it, with respect to exchange of EM energy flow between the time and 3-space domains. To sum all these broken 3-symmetries to an overall net 3-symmetry during the discharge of the free excitation energy of the circuit, obviously requires some specific macroscopic cohering function in the circuit. Moreover, one exists. Specifically, the closed current loop between the external circuit attached to the generator and the internal source dipole created between the terminals of the generator, is what enforces the circuit's net 3-symmetry and Lorentz symmetrical regauging. That is an arbitrary man-made condition, and not a law of nature, physics, or thermodynamics.

---

[62] Or to a charge with its clustering virtual charges of opposite sign, if we redo the experiment with a single "isolated" observable charge and consider it more exactly as a set of composite dipoles, each of which exhibits broken symmetry in the vacuum virtual particle flux.

By our gedanken experiment's demonstration and the existence of innumerable disequilibria between the active environment[63] and the system in an ordinary EM circuit, it follows that a Maxwellian circuit must be *capable* of exhibiting a coefficient of performance (COP) of COP>1.0. It is a giant collection of open dissipative systems, far from equilibrium in the vacuum exchange.

That concept of course is contrary to the entire present mindset of electrical engineering and Maxwell-Heaviside-Lorentz classical electrodynamics. Those disciplines merely continue to accept the source charge as if it freely created from nothing all that energy in its associated fields and potentials reaching across the universe. Ironically, so long as the giant negentropy of the common dipole and its disequilibrium with its active environment (the local active vacuum and the local active curvatures of spacetime) are unaccounted, engineers and scientists will continue to produce the inane Lorenz/Lorentz-regauged power systems they produce today and have produced for more than a century. Also, they will continue — wittingly or unwittingly — to implicitly prescribe that every charge and dipole in the universe violates the conservation of energy law and is thus a *perpetuum mobile* of the grossest kind: a system which massively and continuously creates EM energy from nothing.

We now point out something deeper and very important. In modern physics, the *observed* positive charge is regarded as the time reversal of the *observed* negative charge. See Figure 3-2. If the negative charge is responsible for EM energy flow from the time-domain into 3-space, then *as observed* the positive charge must represent a time-reversed situation. We observe time reversal in the charge domain as a parity reversal and a charge reversal, but with the energy remaining positive. In other words, one imagines that one records a process on videotape, and then runs the video backwards to observe the "time-reversed" situation. When observed, both the sign of the positive charge and its direction of motion appear to us to be reversed from the negative case.

---

[63] In paragraph 3.10 and in Chapter 9 we will introduce — more deeply — the notion of the supersystem, consisting of three parts: (1) the system and its dynamics, (2) the active vacuum and its dynamics, and (3) the active spacetime curvatures and their dynamics. All three components of the supersystem interact with each other. In our gedanken experiment, the system (component 1) must receive the excess energy from the system's external environment, which consists of components 2 and 3 of the supersystem.

It follows that the positive charge as observed involves a flow of (positive) energy from 3-space back to the time domain. If we wish, before observation we can accept that "observed positive energy inflow" as an unobserved outward radiation of negative energy. Thus, a dipolarity involves a continuous flow of energy from the time domain into 3-space, and back from 3-space into the time domain (Figure 3-2). The energy flow circuit (circulation) is completed by a remaining energy flow across the dipolarity in the time-domain only, from the positive charge where the energy enters the time domain, to the negative charge where it exits again.[64]

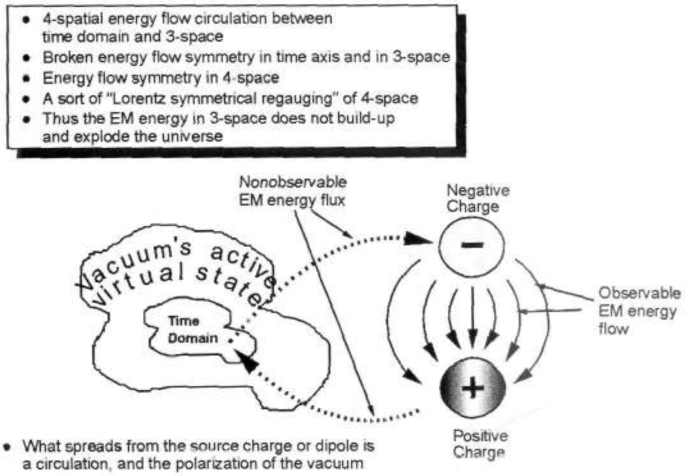

- 4-spatial energy flow circulation between time domain and 3-space
- Broken energy flow symmetry in time axis and in 3-space
- Energy flow symmetry in 4-space
- A sort of "Lorentz symmetrical regauging" of 4-space
- Thus the EM energy in 3-space does not build-up and explode the universe

- What spreads from the source charge or dipole is a circulation, and the polarization of the vacuum

Figure 3-2 Circulation of energy flow in a dipole, between time domain and 3-space.

---

[64] Rigorously, we should speak of $+qt$ and $-qt$. Charge q is an observable, and therefore an instant frozen snapshot. No observable exists continuously in time, but instead it recurs continually as the photon interactions — iteratively producing the d/dt (LLLT) = LLL observations — continue. The moment a charge q absorbs a photon, it changes to qt in form. The moment the charge-time qt emits a photon, it changes back to q in form. This incessant change of q -> qt -> q -> qt —» .... is continual, at every level of the macroscopic charge, with its entire ensemble of photon interactions, including both virtual and observable photon interactions. As can be seen, the "charge as it persists" and "moves through time" has a rich dynamic substructure essentially ignored in present electrodynamics. It is, however, recognized by leading theoreticians and Nobelists.

However, even each of the clustering virtual charges and each of the differential charges of the observable charge, may be further broken into a central charge and clustering charges of opposite sign, in the modem view of the vacuum and its polarization. Hence, at any level — electron or positron — we actually have a dual and slightly separated bidirectional flow, at least in the observable equilibrium condition. The "positive energy" and positive flow of time is usually associated with "inert" matter with the electrons on the outside, so that the observer observes normal matter through the "forward time" situation mostly. Hence much of the accompanying backwards or time-reversed flow — from the positive charges in the nuclei, etc. — is "buried in the vacuum virtual state" in the Dirac sea and in Dirac sea hole flows.

Indeed, the vast hierarchy of nested structures and nested dynamics in charge as it exists in the vacuum, is recognized in physics, as is the fact that it involves infinite energy. E.g., quoting Nobelist Weinberg {194}:

> *"[The total energy of the atom]... depends on the bare mass and bare charge of the electron, the mass and charge that appear in the equations of the theory before we start worrying about photon emissions and reabsorptions. But free electrons as well as electrons in atoms are always emitting and reabsorbing photons that affect the electron's mass and electric charge, and so the bare mass and charge are not the same as the measured electron mass and charge that are listed in tables of elementary particles. In fact, in order to account for the observed values (which of course are finite) of the mass and charge of the electron, the bare mass and charge must themselves be infinite. The total energy of the atom is thus the sum of two terms, both infinite: the bare energy that is infinite because it depends on the infinite bare mass and charge, and the energy shift... that is infinite because it receives contributions from virtual photons of unlimited energy."*

Here again we have met Kron's "open path" in a new and unrecognized form.

Hopefully we now are in a position to rigorously define charge — or try to — for the first time in physics.[65] **Charge** is the ongoing circulation of EM energy flow between the time and 3-space domains, as seen by the observer. **Negative charge** is the ongoing absorption of positive EM energy input from the time domain (from ict), transduction of the absorbed energy into positive 3-space EM energy, and re-emission of the EM energy in all directions in 3-space. **Positive charge** is the ongoing absorption of EM energy input from 3-space, transduction of the absorbed energy into the time domain, and re-emission of the EM energy in the time domain. Or contrary-wise, one may regard the positive charge as the ongoing absorption *of negative* EM energy input from the time domain, transduction of the absorbed negative energy into 3-space, and re-emission of the negative EM energy in all directions in 3-space. The negative charge then absorbs the negative EM energy impinging on it from 3-space, and transduces it into negative energy flowing back into the time domain.

Any *overall* curvature of spacetime, as seen by the observer, results in one of the two flows *predominating* as seen by the observer. Usually, at COP < 1.0, this Dirac sea hole current does not bother our circuit's operation, but remains confined to the local vacuum. It is often not bothersome or is minor for other than very high COP's. The reason is that, for such moderate and small COP's, most of the Dirac sea hole currents react either with (i) the seething local vacuum virtual particle cauldron, or (ii) the atoms in the material lattice to form lattice holes.

As the COP increases, the fraction of Dirac holes interacting with the lattice to form lattice holes increases. As the COP further increases greatly, a large fraction of the Dirac sea holes do not react with the local vacuum cauldron (they become stable forms in that cauldron) or the lattice, but flow into the input section of the system to react with and "eat" incoming electrons from the external power source.

If the inflow of electrons is insufficient to fill the arriving Dirac sea hole current at the input section, the hole current passes back into the feeder line and into the distant power source, eating electrons all the while. This includes eating electrons in the distant power source, which simply "sees" that it must furnish additional current and voltage (power) toward the distant circuit or system.

---

[65] This is still a first order, macroscopic definition. For a more comprehensive statement, one must spell out the hierarchy and dynamics of the virtual state, the participating mass of the charged particle and its generatrix, the operations and dynamics of the *d/dt* observation operator, etc. We have not yet succeeded in getting all that into a single reasonable but sufficiently short statement!

In that operational COP region where strong interaction with the lattice occurs, the attachment of the lattice positron (lattice hole) to the mass of the atom or other ion, results in a very sluggish response of the overall ion with EM fields, compared to the response of an electron or a negative energy state electron hole (Dirac sea hole). Hence much of conventional electrical power circuit theory ignores the lattice holes and hole currents which no longer have negative energy because they have reacted and been "observed". Semiconductor designers, however, account for it and handle it in the semiconductor (in a given range of operation) primarily by the use of donor and acceptor materials.

At COP»1.0, however, the Dirac sea hole current must be strongly taken into account because it produces substantial and novel new effects in the power system itself. This Dirac sea hole current "flows in reverse" with respect to the circuit — from output back to the input. In short, it flows *through the vacuum* from the output section of a unitary COP»1.0 system, back through the middle section of the system, and back into the input section where it "eats" incoming electrons from the external power supply. The power supply thus must furnish electrons to "kill (power the killing of) the Dirac sea hole current arriving in the input section", while still furnishing the necessary amount of additional electrons to power the circuit's losses and loads.

The hole current arriving back in the input section of the unit poses an additional "power system load". These holes must be filled with incoming electrons (which disappear into the Dirac sea hole in the vacuum without radiation, due to the concomitant relaxation or change of the local curvature of spacetime) before any incoming electrons are left to power the circuit in normal fashion. So the hole current appears as a highly undesired extra "load" formed in the input and "eating input power from the external power supply".

We further discuss this situation and how to deal with it, in Chapter 9.

Finally, we propose a formal modification of the concept of "EM energy flow through space" (perhaps the first since its formulation independently by Heaviside and Poynting). The EM energy flow connected with the charge is a circulation of EM energy flow between the time domain and 3-space and back, *as a result of dipolarity*. This 4-circulation flow occurs at every point in 3-space (at every point dipole in the 3-space vacuum).

The formation of an observable charge results in an immediate, continuing, spreading re-organization (polarization) of the polarization of the vacuum virtual particle flux itself. It is this *spreading polarization of the vacuum* — and its accompanying giant negentropy EM energy flow circulation in 4-space connected with the spreading dipolarity and the observable "source" charge — that moves outward in all directions at the speed of light in 3-space from the newly formed charged particle or group of them.

We also define charge itself as that *continuously spreading dipolarity in 3-space*[66] accompanied by the giant negentropy EM energy flow 4-circulation from the time domain to 3-space and back.

The giant negentropy circulation in 4-space — represented by the spreading vacuum point-polarization and hence broken 3-symmetry — is what (observably) moves out in all directions from the newly formed observable charge. Since we "observe" only the 3-space aspect, we "see" it as the spreading appearance of 3-space EM energy (the polarization change of the vacuum itself) moving in 3-space, because we "see" it as the iteratively observed frames of a motion picture. *As observed,* that is one way of expressing it. As it actually *exists and happens prior to observation?*[67] the so-called "flow of EM energy in space" is not that at all. It is the radial spread of the appearance of a giant 4-space negentropy EM energy flow circulation between the dipoles formed in the polarization of space. In that process, the special 4-symmetry between time and space is maintained but with broken 3-energy flow symmetry and broken time-energy flow symmetry with respect to the 3-space observer.

---

[66] I.e., the spreading partial restructuring and re-organization of the vacuum energy or energy flux.

[67] The sharp-eyed reader will note that we clearly distinguish between the operation of a "tree falling in the forest" without an observer to see it, and a "tree falling in a forest" as an observer sees it. Without observation, there is no observed tree, no observer, and no "observed falling". Yet there exists the "tree x time", the "observer x time", and the "falling x time" whether or not their observation occurs with its resulting *d/dt* differentiation operation. One set (the causal set) is 4-spatial, and the other set (the event set after observation) is 3-spatial. The cause is 4-spatial and unobserved, and the effect is 3-spatial and observed. They are not the same, though related by the *d/dt* observation operator and by the interaction of the 4-space cause with a previous 3-space frozen observation to change it to a new 3-space frozen observation. Note that one can and does have a cause without an effect in any legitimate theory of causality, but not an effect without a cause. The cause only assumes the potential for interaction and observation, whereas the effect assumes that the interaction has occurred, yielding the observation as the effect.

The 4-circulation (giant negentropy) of the EM energy flow "from" or "connected with" a charge, is intimately connected with the spin of the charged particle, since the charged particle spins in both the time domain and in 3-space, as discussed in paragraph 3.3 below. We believe the spin itself is the basic giant negentropy generator.[68] If that hypothesis is true, then giant negentropy, the circulation of EM energy between time-domain and 3-space and back, broken 3-symmetry and broken time-energy symmetry (as seen by the observer), the flow of EM energy through space (as observed), the nature of charge itself, and polarization of the quantum mechanical vacuum are all just different sides of the same coin.

Without amplification, we state that his giant negentropy outflow process has additional important ramifications in electrodynamics, general relativity, particle physics, and unified field theory. However, we leave further discussion of the subject for a future paper or book and for the advanced theorists.

3.2 Reinterpreting Whittaker's Decomposition of the Scalar Potential

As previously stated, we applied and re-interpreted Whittaker 1903 decomposition of the potential {85} between the ends of the dipole. The scalar potential is actually a harmonic set of bidirectional longitudinal EM wavepairs, where each wavepair consists of an outgoing EM longitudinal EM wave in real 3-space and an incoming EM longitudinal EM wave in the complex plane. Hence we have a new and novel EM energy flow symmetry in 4-space, where (as seen by the observer) the broken 3-symmetry of the dipole has removed the usual arbitrary imposition of an additional condition of 3-space energy flow symmetry as well. Simply making a dipole permissibly breaks the 3-space conservation of energy flow symmetry, and moves to 4-dimensional conservation of energy flow. Both EM energy conservation in 3-space and in the time domain are individually broken, but conservation of energy flow is now upheld between the time-domain and the 3-space domain. In blunt terms, we "consume a little time" to get any EM energy that exists in 3-space.

---

[68] We leave to a future sharp young graduate student the business of expressing the angular momentum (spin) of the charged particle as a result of giant negentropy in EM energy flow between the time domain and 3-space. It can probably be done by deeply considering the giant negentropy process for an "isolated" observable charge considered as also containing its clustering set of virtual charges of opposite sign.

The broken symmetry of a source dipole — or of a source charge considered with its clustering virtual charges of opposite sign as a set of dipoles — produces a continuing giant 3-negentropy, as seen by the observer. He only sees the steady 3-space outpouring of energy, and sees that it will continue as long as the dipole or charge remains intact.

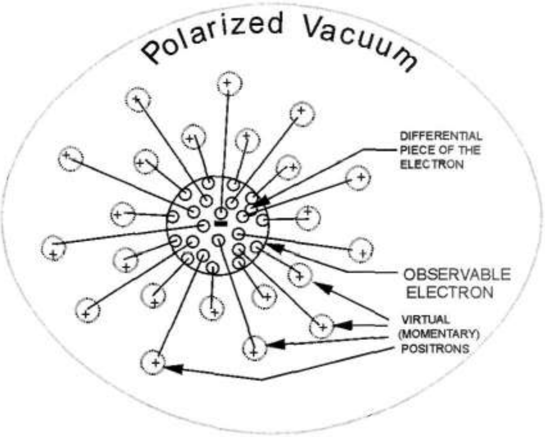
Polarized Vacuum
DIFFERENTIAL PIECE OF THE ELECTRON
OBSERVABLE ELECTRON
VIRTUAL (MOMENTARY) POSITRONS

Figure 3-3 Observable electron charge as a set of composite dipoles.

To clarify the "isolated charge" as a set of composite dipoles, see Figure 3-3. There is no such thing as an *isolated* observable charge in 3-space; instead, the observable charge is surrounded by attracted virtual charges of opposite sign in the active vacuum. We take a differential piece of the observable charge and a momentary virtual charge of opposite sign, and call the two a "composite dipole". An observable charge is simply a set of composite dipoles connecting virtual and observable energies and energy flows.

Hence any source charge (classical sense) may be considered such a set of composite dipoles. So any source charge also multiply initiates the broken 3-symmetry and broken t-symmetry in EM energy flow (as interpreted by the observer), but conserves energy flow 4-symmetry between the time domain and the 3-space domain. In 4-space, there is an energy flow circulation from the time domain to 3-space and back to the time domain again.

This special circulation of energy is strikingly similar to, and may be an extension of, Heaviside's unpublished theory of combined electrodynamics and gravity, where he used closed circulation of EM energy as exhibiting gravitational effects (in modern terms, producing spacetime curvature) {195, 172b.}. This approach or extensions of it may yet have a profound impact on physics, as commented upon by Laithwaite {196}. Certainly Heaviside considered his own energy flow theory. So it appears that he considered the excess energy flow — that "misses" the interaction with the circuit or assumed unit point charge and is not utilized — as producing gravity. That of course would be in perfect agreement with modern general relativity, since any change in the local energy density of vacuum is also a change in the local curvature of spacetime. We have called that Heaviside nondiverged and unaccounted EM energy flow by the label *dark positive energy*. See our further discussion of unaccounted (dark) positive energy in paragraph 3.8 below.

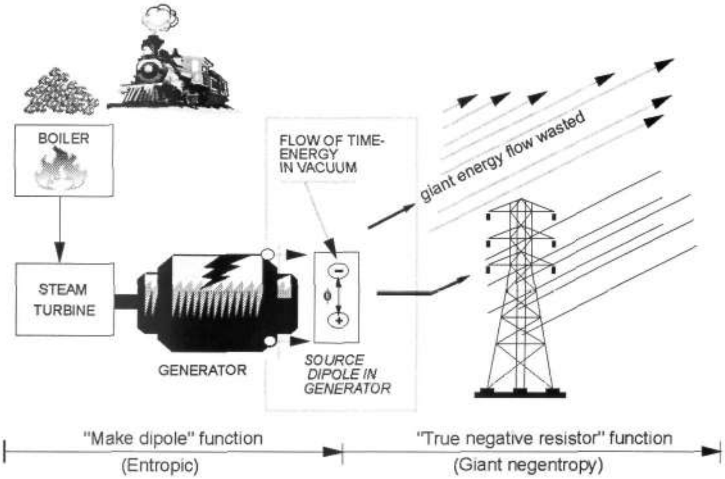
BOILER
FLOW OF TIME-ENERGY IN VACUUM
giant energy flow wasted
STEAM TURBINE
GENERATOR
SOURCE DIPOLE IN GENERATOR
"Make dipole" function
(Entropic)
"True negative resistor" function
(Giant negentropy)

Figure 3-4 Dipole in generator powering circuits, taking energy from the time domain.

See Figure 3-4. This 4-symmetry is a far more fundamental energy flow symmetry than either 3-symmetry or t-symmetry. Both the "isolated charge" and the dipole are open systems far from equilibrium with their external active environment (the active vacuum and the active curvatures of spacetime). They are permitted to perform the five "magical" functions of disequilibrium thermodynamical systems. Such a system can: (i) self-order, (ii) self-oscillate or self-rotate, (iii) output more energy than the operator inputs, (iv) power itself and its loads (the "load" of a source dipole or source charge may be said to be comprised of the associated fields and potentials reaching across all space, and the continuous flow of energy necessary to maintain those fields and potentials), and (v) exhibit negentropy.

The source charge and the source dipole exhibit all five functions. We often refer to these characteristics of a charge or dipole as its *giant negentropy* functions {197}.

It only takes a single white crow to prove that not all crows are black. Every charge and dipolarity in the universe is already a Maxwellian system far from thermodynamic equilibrium with its local active environment — the very class of Maxwellian systems that Lorenz/Lorentz symmetrical regauging of the Maxwell-Heaviside equations arbitrarily discards. That local environment consists of the local curvatures of spacetime and the local active vacuum. The system together with the two components of its active environment is called the *supersystem.* All three components of the supersystem continuously interact with each other and exchange energy. All is dynamic, continuously. Any appearance of a "static state" in the system is actually an equilibrium condition stabilized in the ongoing supersystem dynamics.

So one does not have to "prove" that Maxwellian systems exhibiting COP>1.0 (even COP = ∞) are possible. Every charge and dipolarity in the universe is already just such a Maxwellian system, even though all such systems are erroneously discarded by Lorenz/Lorentz symmetrical regauging. Either one must accept asymmetry and COP>1.0 EM systems in electrodynamics, or one must purge from the theory all the charges and dipoles and all effects from them. That destroys all the charges, all the fields, and all the potentials.

That also destroys electrodynamics itself, as well as chemistry, particle physics, and most of modern science. The arch critic of overunity EM system research would do well to acquaint himself with some additional physics. Either that, or practice what he advocates, and give up all charges and dipoles and all effects from them — including in chemistry, particle physics, materials science, etc. When he holds on to the charges and dipoles while insisting on the impossibility of COP>1.0 Maxwellian systems, his cherished conventional EM theory eats itself by its own tail and becomes an oxymoron.

3.3 How the Dipole Charges Transduce Time Energy

We present a simplified analogy[69] that allows one to visualize how the dipole transduces the incoming time-like longitudinal EM energy flow into 3-space real EM energy flow for negative charge, or vice versa in the case of positive charge. With a little oversimplification, a charge may be said to spin 720° degrees in one complete rotation. For our purposes, it spins 360° in the imaginary plane (over in *ict),* and then spins 360° in the real plane (in 3-space). Thus the negative charge can absorb the incoming EM energy flow in the complex plane, transduce or "flip" the absorbed EM energy into 3-space to begin its 360 degrees in 3-space, and the excitation can decay during that 3-space spin part of the cycle. The only "variable" in *ict* is the t, so any energy flow has to be "by means of the variation of the t-variable". We already argued in a previous chapter that time can be treated as highly compressed spatial energy, having the same energy density as mass. So absorbing a very tiny amount of time-energy and transducing (decompressing) it into spatial energy produces enormous spatial energy *(t* multiplied by the factor c^2).

The negative charge absorbs a *little bit* of positive time energy, transduces it into a much larger amount of 3-space excitation energy, and re-emits it in 3-space as an observable, real flow of EM energy radially outward in all directions. The positive charge receives the large amount of 3-spatial EM energy and absorbs it, transduces it back into time-energy (highly compressed energy) and re-emits it in the time domain as a little bit of highly compressed time energy.

Conversely, we may consider that the positive charge absorbs a little bit of negative time energy, transduces it into a much larger amount of negative 3-space excitation energy, and re-emits this negative 3-space energy in all directions in 3-space. That consideration is important in the cold fusion phenomenology, and in fact accounts for the formation of little fleeting "time-reversal zones" in the electrolyte solution where the law of attraction and repulsion of charged particles is momentarily reversed. The omission of the giant negentropy mechanism and the negative energy mechanism from particle physics reaction theory is why there do not exist such nuclear transmutation actions at low spatial energy (but high time energy) in conventional particle physics. All such permissible reactions have been arbitrarily omitted from the discipline.

---

[69] **For** simplification, the analogy treats "spin" of the charged particle as if it really were something like that of a spinning top, at least in its 3-spatial component. However, in physics the entity **spin** is not confined to a flat plane or even to 3-space. C**onsequently,** we caution that the analogy is primarily a mnemonic aid.

For the positron (or any positive charge), we assume that the process is reversed: 3-space energy is continuously absorbed (from the active vacuum exchange) by the positive charge during its 360° spin in 3-space. This 3-space excitation energy is then flipped into the time domain as time-charge or time-energy, as the charge starts its second 360° spin in the time domain. During that latter part of its 720° spin cycle, the spinning charge re-emits the time-charge energy (time-excitation) as a flow of a little bit of very dense EM energy in the time domain.

The effects of these mechanisms in systems of charges in motion are met with as time dilation effects, spacetime curvature effects, frame rotation effects, etc. in relativity theory.

A charge that has received excess EM energy from the time domain is said to be *time-charged* or *time-excited.* Since time-charge or time-energy is c^2 denser than spatial energy, a tiny amount of time-energy excitation or time-charge may be re-emitted as substantial spatial energy over a period of time. In short, the decay of time charging or time-excitation of charged particles can occur slowly and over time. The decay is marked by the presence of longitudinal EM wave radiation, by mysterious ionization phenomena appearing in some Geiger counters depending on their individual time-histories (individual initial time-charging condition), and by excess energy appearing in electrolytes and emitted as heat where such decay of time-charge is occurring.

Usually in ordinary circuits and EM systems the time-charge and time-charge-decay effects are minimal, being offset by their opposites in close adherence to equilibrium. However, when this is not the case and time charging does appreciably occur, then novel phenomena result in those circuits and systems. Such effects have occurred for some time in instruments associated with rigorous electrolyte experiments at U.S. Naval research facilities at China Lake {198}, but the researchers have not recognized the cause. We have pointed out the time-charge and time-decay solution to those novel phenomena, and explained why apparently "identical" instruments need not respond to the stimuli in the same manner {199}. Along the way, we also explained several of the low spatial energy (thus high time-energy) transmutations, including giving the typical reactions producing the excess deuterium, tritium, and alpha particles. We also explain those new nuclear reactions at low spatial energy (but high time-energy) in Chapter 10.

As we stated in Chapter 1, paragraph 1.1.3: In such electrolyte experiments where the palladium electrodes load with concentrations of positive charges on hydrogen ions of one form or another, then in the adjacent solution there can arise fleeting **time-reversal zones.** Such a zone is simply a very tiny volume of fluid in which antiphotons temporarily comprise the majority of the ongoing photon interactions with the local ions in solution in that zone. The result is that the normal (time-forward) law of attraction of unlike charges and repulsion of like charges is reversed; *in the TRZ, like charges attract and unlike charges repel.* This allows two positive ions such as H+ to be drawn together so tightly that each enters the (reduced) strong force region of the other, forming a quasi-nucleus. When the other ions move to eliminate the TRZ, during the decay of the TRZ the strong force increases more rapidly than the EM force. Hence, the preferred method of decay of the now-excited quasi-nucleus is by quark flipping to allow the quasi-nucleus to become a new nucleus. This nuclear transformation interaction allows the production of the excess deuterium, tritium, and alpha particles.

3.4 Reordering of the Local Vacuum: Giant Negentropy

A giant negentropic reordering of a fraction of the surrounding vacuum's energy is initiated when a dipole or charge is formed. This reordering is deterministic, and is the 4-flow energetic structure shown by Whittaker {85}, slightly reinterpreted, with perfect coordination of the converging EM energy inflow in the complex plane with the outflowing converted real EM energy in 3-space. Hence this negentropic reordering introduces "hidden order" into the vacuum's otherwise disordered energy. This reordering of a fraction of the vacuum's disordered energy also changes the Gibbs statistics assumed in quantum mechanics, replacing it with an "already chaotic" statistics containing hidden order. We propose this as a possible contribution toward resolving the presently unsolved problem of the missing chaos in quantum mechanics {200}.

3.5 Some Implications of Giant Negentropy

Giant negentropy has many envisioned impacts upon physics, such as:

(a) Allows development and use of negentropic EM circuits (COP>1.0 systems) rather than entropic circuits. Once a broken 3-symmetry with stable 4-symmetry is established (as by forming a dipole), it continues indefinitely unless deliberately destroyed.

(b) Allows direct negentropic reordering of a practical and usable fraction of the vacuum's energy at will, at any point in the universe, easily and cheaply.

(c) Provides for electrodynamic engineering of the vacuum energy and of spacetime curvatures clustered in specific templates (the concept of vacuum engines or spacetime curvature engines). This has significant medical applications {387c}.

(d) Utilizes time as energy (per our previous discussion) and thus allows use of the Whittaker phase conjugate half of the biwave set as an EM energy flow within the flow of time. Observably this translates to the continuous receipt of free time-energy by the formed broken 3-symmetry (e.g., the charge or dipole). Accordingly, the charge or dipole continuously pours out real observable EM energy in 3-space, without any observable EM energy input (time is not observable, and neither is the time-polarized photon alone).

(e) Allows "burning a little time for fuel", so to speak, in order to transduce it into a very large spatial energy flow, due to the enormous compression of 3-spatial energy represented by time-energy by a factor of c^2.

So from the giant 3-negentropy discovery, we have arrived at some intriguing findings:

(a) As is well known in particle physics, the opposite charges of a dipole constitute a broken 3-space symmetry in the violent flux exchange between the active vacuum and the dipole.

(b) This dipole's broken 3-space symmetry in EM energy flow, provides a *relaxation* to a more fundamental EM energy flow symmetry between the time domain and 3-space, and therefore in 4-space.

(c) There is no law of nature or physics that requires 3-symmetry of EM energy flow as an additional condition applied to 4-symmetry of EM energy flow. Instead, 4-symmetry energy flow is more basic, by the hierarchy rule.[70]

---

[70] Quoting Steven Weinberg: *" ..there is a hierarchy of symmetries; whatever symmetry unites the gravitational and strong nuclear forces with the electroweak*

(d) The dipole is a practical and very simple means of "breaking" the additional 3-flow symmetry condition in EM energy flow our systems have previously exhibited, and relaxing to the fundamental 4-flow symmetry *without* 3-flow symmetry.

(e) So long as the dipole statically exists (e.g., imagine an electret suddenly formed, or a charged capacitor with no leakage), real usable EM energy circulation between time and 3-space, at each point in 3-space, will continuously move out from the dipole at light speed in all directions. In the view of the observer, at the same time, reactive EM power (actually, time-energy) will continuously flow into the dipole from the time-domain (the complex plane), and be transduced into real EM power output radially outward in 3-space from the dipole.

(f) Observably, a dipole or dipolarity and its scalar potential thus comprise a *true negative resistor* system of the most fundamental kind. The dipole continually receives EM energy in unusable form (reactive power, which cannot perform real work), converts it to usable form (real power, which can perform real work), and outputs it as usable, real EM energy flow (real power) in 3-space. Further, it spreads this "transformation of time-energy into 3-space energy at every dipole point" outward in 3-space in all directions.

(g) So at its formation the dipole initiates a continuing *giant negentropy* — a progressive reordering of a substantial and usable portion of the vacuum energy {201} that begins with dipole formation and continues. Further, this reordering of vacuum energy flow at every point in space continuously spreads in all directions from the initiation point, at the speed of light. Dipoles in original atoms formed shortly after the beginning of the universe, have been pouring out real EM energy for some 14 to 15 billion years or so (i.e., in the prevailing view; the age of the universe and whether it is expanding are still controversial {754}), and have reordered a fraction of the vacuum's energy, where the magnitude of the re-ordering varies inversely as a function the

forces is broken roughly a hundred million million times more strongly than the symmetry that unifies the weak and electromagnetic interactions. The puzzle of explaining this enormous difference in fundamental energies is therefore known in elementary particle physics today as the hierarchy problem." [Weinberg, *Dreams of a Final Theory,* Vintage Books, Random House, 1993, p. 205], radial distance from the dipole.[71] This is in fact the solution to the long-vexing problem of the source charge and its associated fields and potentials and their energy, reaching across the universe. The charge together with its associated clustering virtual charges is a set of composite dipoles, hence a multiple *broken 3-symmetry* energy flow system.

(h) If the dipole is destroyed, the ordering of the vacuum energy ceases, leaving a "separated chunk" of reordered vacuum energy dynamics that continues to expand at the speed of light in all directions, steadily reducing in local intensity as it expands.

(i) At any very small volume in space, from the dipole dynamics of the universe it follows that a great conglomerate of reordered vacuum flows and fluxes — some continuous, some chopped — is continually passing through that volume. Further, the situation is totally nonlinear, so that direct wave-to-wave interactions occur continuously amongst these energy flows and waves. We hypothesize that this is the actual physical mechanism constituting Puthoff's cosmological feedback mechanism {202}.

(j) Further, in 1904 Whittaker {203} showed that any EM field or wave pattern can be decomposed into two scalar potential functions. This initiated what is called *superpotential theory*. Each of the two potentials for those functions, of course, decomposes into the same kind of harmonic longitudinal EM wavepairs as shown in Whittaker 1903, plus superposed dynamics. In other words, the interference of scalar {204} potentials — each of which is actually a set of longitudinal EM waves, and not a scalar entity {205} at all, but a *multivectorial* entity — produces EM fields and waves and their dynamics. Hence we hypothesize that the Whittaker interference of the propagating reordered EM energy entities, continuously occurring at any point in space, generates the zero-point EM field energy fluctuations of the vacuum itself. Indeed, an AIAS group paper by Evans *et al.* {206} has already shown that just such "scalar interferometry" produces transverse EM fields and waves in the vacuum at a distance.

---

[71] The field energy density in the fields produced by the source charge varies inversely as the square of the radial distance. The potential energy density in the potentials produced by it will vary inversely as the distance.

Since energy can neither be created nor destroyed, there really are no true *energy sources* in nature in the sense that they *create* energy and pour it out {207}. Instead, what we call "energy sources" merely convert the energy already there in one form, into another form. All energy sources are no different from a solar cell in the sun, or a windmill in the wind, at least in principle. They are actually energy transducers rather than energy sources. This corresponds well with the conservation of energy law wherein not a single joule of energy can be created or destroyed. Instead, the form of that joule of energy can be continually changed. Every joule of energy present in the universe after its formation is still present, and is still doing joule after joule of work in iterative interactions with transducers that change its form.

A priori, we can measure no real 3-space input of EM energy to the unchanging charge but we can measure real 3-space EM energy pouring from it. Energy must be input to it from the active vacuum in a *nonobservable* form, and converted by it into an *observable* form that is re-emitted, usable, and produces what we call the "fields and potentials" and their energy, associated with that "source charge". As is common usage, we will continue to use the term "source charge" or "source dipole", but with the understanding that we refer to a special kind of *energy transducer.*

When we "make entropy", we must do work. Even so, doing work is not really entropic unless we also lose control of the energy[72] — whose form was merely changed by doing the work — and thence lose any *further* ability to use that energy (change its form, or transduce it) to do useful work. And when we observe entropy, we also do work in the observing. Some of the erroneous notions about entropy desperately need correcting. E.g., quoting Weinberg {208} for a lay definition of entropy:

> ***"..a certain quantity called entropy:*** *To define entropy, imagine that some of the system's temperature is very slowly raised from absolute zero. The increase in entropy of the system as it receives each small new amount of heat energy is equal to that*

---

[72] An interesting point is that energy is energy and remains energy. "Disordered" ultimately means "unusable" or "inconvenient" and nothing else. Disordered energy is still energy, and if we do not lose it from the system we can in fact have the **system** automatically reorder it, e.g. by use of simple retroreflection. The entire concept of entropy must be rigorously interpreted in terms of "energy whose reuse is lost", at least by the processes being performed by the system considered. In our view, "entropy" itself is just additional usually "lost" energy to be recovered.

> *energy divided by the absolute temperature at which the heat is supplied."*
>
> *"....entropy, which always increases with time in any closed system..."*

However, we point out that apparently there exists no completely closed system in the entire universe. In the notion of entropy and the escape of the energy from their "closed" system, thermodynamicists themselves ubiquitously negate the latter premise that entropy in a completely closed system always increases with time. That simply does not follow. The real situation is as follows: (i) Entropy in a half-open system with no allowed external input of energy, but with allowed escape of energy, does decrease with time if the energy of the system decreases with time by escaping.[73] Hence the cooling of "hot" water with time, if and only if the excess energy in the system can have a net escape to the (cooler) environment as time passes, (ii) Energy in a completely closed system (closed to both energy and mass), with no input and no output of energy at all, simply remains constant because it is not dissipated; simply apply the conservation of energy law. It may rearrange within, and the subcomponent energies may disorganize with respect to range, but the total energy remains. Entropy actually only deals with the loss of ability to utilize the energy. To wit, the assumed (and apparently true, so far as we can tell) conservation of energy in the present universe is an example. A closed system is in this sense a special case of a system in equilibrium with its active environment.[74] I.e., any "closed system" is such a special case.

---

[73] In case the reader has not thought of it, until the energy "escapes" it is not dissipating or dissipated from the dissipating "object". A "hot" object is energetically excited but not hot; the *energy dissipated* from it (including from every component within it) is scattering, and hence disruptive and "hot" in its interaction with a test instrument (or one's finger). The thermometer (or any other instrument) is external to the rest of the system (even if imbedded in it). Hence merely to "measure the temperature" is actually to pass energy out of the system into the thermometer (or other instrument), so that the thermometer measures the heat of the dissipating energy that has just escaped from the system. For these and similar reasons, Romer — former editor of *A. J. Phys.*— pointed out that "heat" should not even be used as a noun {84}.

[74] The concept of energy always requires that one can write "energy of. ..(X)" and fill in the (X). We usually write the phrase "of...(X)" as "X energy". Thus we write the phrase "energy of heat" as "heat energy," the "energy of the vacuum" as "vacuum energy", and the "energy of the EM entities" as "EM energy". At least in its usage, energy always requires a medium or entity "having" or "exhibiting" the energy. As Feynman pointed out, we really do not know what "energy" per se *is.*

(iii) In a system closed with respect to output (escape) of energy, but with the system opened only to allow external input of energy, the system energy will increase, simply by the conservation of energy law. That system's entropy will decrease and its negentropy will increase. (Its mass will also increase, and so it violates the thermodynamic misdefinitions of closed system; see Appendix A). It is well known that the entropy of an open system far from equilibrium cannot even be computed! (iv) Energy in an open system where input of energy and escape of energy both occur, may either increase, decrease, or remain constant, depending upon whether the input rate exceeds the output rate, or vice versa, or the energy input and output rate are equal. Therefore such a system can exhibit overall entropy, overall negentropy, or overall equilibrium, (v) Net energy scattering rate (rate of energy escaping the system in disordering fashion) from a hot system determines the observed temperature of that system at the time of observation. The system itself is cold, prior to escape of the energy. The "heat" is due to the interaction of the escaping energy with an external object, in its disordering fashion. Inside the system prior to escape, the energy is ordered hence cold.

A system in equilibrium in its exchange with its active external environment is in a state of maximum entropy *for that specific rate of energy exchange with its environment.* As we will point out in Chapter 9, by deliberately balancing the *system-and-vacuum* exchange with the *system-and-curved spacetime* exchange, so that the two exchanges are equal and opposite at a desired rate of exchange, one may establish (at least in theory) system equilibrium at whatever specific rate of energy exchange is desired.[75, 76] In this fashion, e.g., an "impossible" molecule normally intensely unstable and decaying in a nanosecond or less, can in fact be stabilized once an appropriate balance between the two opposing energy exchanges with the system is achieved at the required specific rate of energy exchange. This is actually a new Lorentz symmetrical regauging

[75] This is necessary, e.g., to close-loop a COP>1.0 EM unitary system taking its input energy from the vacuum, as we shall see in Chapter 6.

[76] The Fogal superluminal transmission system in fact uses a conventional signal input to change this "symmetrical regauging" condition in consonance with the amplitude of the input signal. Thus suddenly Lorentz symmetrical regauging becomes a way to transform a conventional signal into a signal consisting of changes in the stress energy potential of the vacuum. That constitutes a longitudinal EM wave, and so it is not limited by the speed of light. To use this process, a Fogal semiconductor must be used as the modulator in the "transmitter", and another must be used as the demodulator in the "receiver".

condition; change of symmetrical regauging is very useful in this fashion. In this way, a great new chemistry with reactions presently deemed impossible is in the offing, and it is already being developed by at least one U.S. company. We do not name the company, under our tacit agreement with the scientists doing the work.

To operate conventional entropic systems,[77] first we must input energy to the inert system to force it out of equilibrium and back into excitation (back into an asymmetrically rcgauged system, violating the Lorentz condition) so that it has some excess energy (and negentropy). Then we allow the decay of the excited system's excess energy to pass through a load and be dissipated from it, doing work for us by powering the load. Conventional systems usually do not recapture the "escaping dissipated energy" but just waste it. *"Dissipating" energy is merely allowing it to escape from the system and from further system control.* Then we brutally force more excitation energy into the system, opening it and breaking its equilibrium to do so. Again we dissipate the system's excitation energy in the load to produce some more work, usually wasting an appreciable fraction of the excitation energy.

In EM systems, it is a little more complex. We continually "switch on" a negentropic flow of EM energy from the vacuum by making a source dipole. In the circuit, we collect a small part of the resulting huge energy flow in space surrounding the circuit conductors. Then we use half the collected energy to destroy the dipole source of the energy flow, while less than half of the collected energy is dissipated in the load (the rest of that second half is dissipated in the external circuit's losses). Hence we must continually restore the source dipole and pay more to do it than we get out in the load as useful work. We make a perfectly good negentropic system, and then forcibly destroy its negentropic ability faster than we power the load and get any use out of it. That's a terrible way to make and use power systems, but it is what has been done for more than a century.

So with our present entropic EM systems (actually our *self-killing negentropic* EM systems) we continually wrestle nature fiercely to the mat, so to speak, by brute force, repeatedly, to bind her into 3-space EM energy conservation. All the while, nature protests our continual entropic brutality by providing the Newtonian third law reaction force[78] back upon our causative wrestler performing the "forcing". To do entropic engineering, we have to continually input 3-space energy to the wrestling mechanism or engine, losing a bit of the input energy in the inefficiencies, and fighting the "back emf", "back mmf", or Newtonian third law reaction that is nature's cry of protest all the while. Those are nature's penalties for imposing a monstrous 3-space EM energy flow symmetry[79] upon her as an *additional, highly undesired, and highly unnatural condition* added upon her beloved special 4-symmetry energy circulation between the time domain and 3-space.

---

[77] By "entropic system" we mean a system whose excitation energy decays and is dissipated from the system or at least from its further control, so that the entropy (lost or unusable energy) of the system continually increases until equilibrium (total decay of the excitation energy) is achieved. As can be seen, this also can be used to define entropy and increase in entropy. What is usually unappreciated is that *equilibrium is actually a counterbalance between ongoing entropy and negentropy.*

In short, because we so love entropic engineering and 3-space EM energy flow symmetry, we have to provide the continual input energy to our entropic processes by burning fuel, damming rivers, erecting windmills, building waterwheels, erecting solar cell arrays, building and charging chemical batteries, etc. In the process, we destroy and pollute the biosphere on a giant scale as we rip down forests, strip-mine and drill the earth, pile up thousands of tons of radioactive nuclear wastes —that will be hot for thousands of years — from our power plants and related facilities, and spew streams of pollutants into the atmosphere, the rivers, the oceans, etc. We do all that biosphere destruction because we inexplicably insist upon placing cruel chains on nature by brutally enforcing 3-space energy flow symmetry upon her. In the process, we adamantly require adherence to classical equilibrium thermodynamics (as given a somewhat altered explanation above, and also in Appendix A).

We do not have to do it that way. In avoiding solving the source charge problem for so long (more than a century), we have been extremely backward in our thinking and in our consequent development of entropic engineering technology.

---

[78] Note that Newton's third law reaction is implicit in the special negentropic energy flow circulation of the dipole, or of the charge considered as a set of composite dipoles. The same energy that moves into 3-space from the time domain, then returns from 3-space to the time domain, at every point in the 4-circulation. If we consider the positive charge as a "source" of negative energy, then there is a concomitant circulation of negative energy from the time domain to a point in 3-space, and back to the time domain.

[79] The 3-space symmetry of energy flow equilibrium is actually between the energy we input (from its external environment) to the system, and the energy that escapes from the system back into its external environment, either in its losses or in its loads. As we stated, equilibrium condition is a balance between ongoing entropic and negentropic operations.

For our dullness, we have had to pay and pay continuously for insisting on doing such atrocious entropic work and forcing nature to do it that way. In so doing, we "tie nature's feet" with that added arbitrary 3-symmetry in energy flow. We ourselves prohibit nature from performing the giant negentropy she so dearly loves and much prefers. We also arbitrarily and *meanly* discard the bountiful electromagnetic energy flow that nature loves to furnish us so freely by her vast preference for giant negentropy. We meanly discard nature's bountiful giant negentropy banquet free for the taking, by our vast thirst for giant entropy and doing violence to nature. Because we act like brutes instead of scientists, we destroy nature in spite of nature's continuing attempt to give us all the EM energy we wish, freely except for a little switching cost.

A far better way is to cooperate with nature and "let nature make her beloved copious negentropy" as she yearns to. To do that, we now can see the startlingly simple mechanism. We simply make a little dipole, *once*, entropically. So we have to pay for *making* the dipole, *once*, and we have to do just a little gentle violence to nature, *once*. Then we need do no more violence, if we just leave the dipole intact and do not destroy it. In short, we do have to tap nature gently on the shoulder. We do not have to brutally knock nature down to the mat, repeatedly.

When we make the dipole, we make a little bit of "broken 3-symmetry" in the universe's energy flow. Voila! Nature sings for joy at finally having her feet freed from the shackles of the horrid 3-symmetry energy flow. In great glee, she instantly relaxes into her dearly preferred giant negentropy and 4-symmetry. She instantly sets to re-ordering a substantial and usable

---

[80] E.g., so far as I am aware, there have been no funded research programs to discover how to deviate and collect some of the freely flowing energy from a permanent magnet or an electric dipole. Instead, electrodynamics has smugly continued to call it "statics" as if there were no dynamics, and ignored the giant negentropy circulation that is "seen" as the steadily outflowing field energy in 3-space, and erroneously called the "static field". There is no such thing as "statics" in the entire observed universe *a priori;* the very process of observation itself is dynamic. What we see as "static" is continual recurrence of a given condition or state or magnitude, as beautifully pointed out by Van Flandern {628}. One might even say that, not only is nature energetic, nature *is* energy, constantly changing and in motion in innumerable ways at once! One might even define "energy" as the process by which nature both acts and observes its actions, in every form and structure and possibility.

portion of the vacuum energy of the universe, in all directions at the speed of light, spreading her preferred giant negentropy EM 4-energy circulation flow from the dipole toward the ends of the universe at the speed of light. As long as we do not destroy the dipole (the broken 3-symmetry) that gently breaks the 3-symmetry shackles, nature's feet remain freed from brutal 3-space symmetry in EM energy flow. In that case, she delightedly continues to reorganize a portion of the vacuum energy, with the reordering spreading radially outward at the speed of light. Simply making a common dipole or charge[81] sets nature to going about reordering an ever-increasing fraction of the entire vacuum energy of the universe, and continuing indefinitely.

Simultaneously with freeing her from enforced 3-space EM energy flow symmetry, in great gratitude nature pours out an immense real EM energy 3-flow from that little dipole or charge. She will continue to pour it out forever for us, if we do not destroy the dipole or charge.

Instead of the present universal entropic engineering, the smart thing to do is make just a little bit of entropy wisely, using the work we get to break 3-space energy flow symmetry (basically, to make a dipole). *Then leave that mother of all negative resistors and free energy flow generators alone and forever unchanged!* Concentrate on intercepting, extracting, and using the free 3-space energy copiously flowing forth from the giant negentropy, without destroying the dipole that is freely providing it.

3.6 How Circuits Are Powered

Let us now look at the great magnitude of the energy flow that nature gives us from that dipole. We have to get into the subject a bit, because EM energy flow theory has been rather thoroughly confused for more than a century.

First, batteries and generators *do not* use their available internal energy — the shaft energy we input to the generator, or chemical energy available in the battery — to power the external circuit. Instead, each uses its available internal energy {209} to perform work on its own internal charges, forcibly separating the charges to form the *source dipole.*

See again Figure 3-4. *All the hydrocarbons ever burned, all the nuclear fuel rods ever consumed, and all the dams ever built, added not one single watt directly to the power line.* All the energy from those activities was input to the generator shaft after normal losses en route, to provide internal magnetic energy available to the generator. In turn, the generator used that available internal magnetic energy only to do internal work on its own internal charges to force them apart, forming the source dipole connected to the terminals. Generators are energy transducers only; they do not directly power their own external circuits.

---

[81] As we previously explained, a charge is just a set of composite dipoles.

Batteries and generators expend their internal energy available to them, to make the source dipole, and for no other purpose! None of their internal energy is used to power their external circuit. It never has been, and it never will be.

Once the source dipole is formed, it does all the hypothesized 4-functions we pointed out previously. It induces the spreading giant negentropic reordering of the vacuum energy, extracts (transduces) EM energy from the continuously reordering vacuum, and pours out from the terminals of the generator (or battery) a vast 3-flow (as observed) of EM field energy along the external circuit. As indicated by Kraus's illustration of the Poynting component {210}, this giant EM energy flow fills all space surrounding the circuit, out to an infinite lateral radius.[82] The energy flow is generally parallel to the conductors of the circuit. Only a tiny component of this flow — due to the surface charges of the conductors and the little boundary layer of energy flow that slides along the surface of the conductors — strikes the surface charges and gets diverged into the conductors (by the lateral withdrawal of the surface charges — with the "stub" or "base" of their field energies — laterally into the conductors. The electrons move mostly laterally, withdrawing from one side of the conductor surface to the opposite. Only the tiny component of its field vector integrated over that small distance is withdrawn into the conductors to power the electrons. This small amount of "withdrawn" energy is the diverged Poynting component, "collected" by the circuit. It also is the small component then used (dissipated) to power the Drude electrons and the circuit.[83] All the rest of that vast EM energy flow in the surrounding space, that pours forth from the terminals, just misses the circuit entirely. It roars on off into space and is wasted.[84]

---

[82] We accent that Kraus, along with other authors, only shows the 3-space *Poynting component* of that flow; i.e., he shows the very small amount of that external 3-space energy flow that is diverged into the circuit to potentialize the Drude electrons and power the circuit. Kraus *does not* show the remaining Heaviside component that is not diverged.

[83] As shown by Kraus, some of the energy flow at the various radial distances from the wire is withdrawn, as the surface electrons and the "stubs" of their near fields precess laterally into the depths of the wire when potentialized. Jackson, *Classical Electrodynamics,* Second **Edition,** Wiley, 1975, p. 223 also points out the nonlocal

The diverged, utilized, and accounted energy flow component — the *Poynting component* — is only a tiny, tiny fraction of the entire giant EM energy flow produced by the source dipole for every circuit.

3.7 The Heaviside Component vs. the Poynting Component

Unless Heaviside and Lorentz did rigorous calculations in a work we have not yet uncovered, we could not find any development of the functions and equations required for computing the numerical ratio of the unaccounted Heaviside "dark energy flow" {*nondiverged* energy flow component that misses the circuit), to the tiny Poynting energy component thai strikes the surface charges of the conductors {211} and is diverged into the wires to power the circuit when the surface charges precess laterally into the wire.

Consequently, we performed a very crude "special case" estimation {212} — a back-of-the-envelope type, with highly simplifying assumptions — for a very simple circuit in which one resistor is powered by a DC source. Our crude estimate showed that about 10^{13} times as much EM energy flow *misses* the circuit, is *not diverged,* and is *wasted* — as *strikes* the circuit, gets collected, and then is dissipated in the circuit to power the load and losses. Until electrical physicists re-examine the energy flow theory and

---

involvement of the transverse current. Quoting: *"...the transverse current... extends over all space, even if J is localized. "* One notes that present electrical power systems concentrate exclusively on the J current, making no attempt to collect and utilize the recognized but rather ignored transverse current.

[84] Note, however, that this Heaviside energy flow component differs from ExH. The Heaviside component is totally nondiverged (being actually in curl or swirl form), and since E and H are both diverged energy components assumed around an interacting unit point static charge, there is neither E nor H in the electrical engineering sense prior to interaction with charge. So the energy flow in both the Poynting and Heaviside components is in a force-field-free form until intercepted by and interacted with charge. Poynting's component, e.g., assumes only that energy component that does interact with charge and get withdrawn into the circuit. Heaviside's component assumes only that energy component that does not interact with charge and does not get withdrawn into the circuit.

again recover the Heaviside unaccounted component in it, that brute force estimate will have to suffice as at least an illustrative example.[85]

What does that 10^{13} ratio mean? For a little 1 watt generator of that DC type with that specific size conductors and that specific resistor, the Heaviside unaccounted energy flow component was about *10 trillion watts,* occupying all space surrounding the wire,[86] if all of it could be intercepted, collected, and used to power loads {213}. But the little circuit was only intercepting and collecting — and using to power loads — about 10^{-13} of the available energy flow surrounding the circuit, *if all* of it could have been intercepted, collected, and used to power loads.[87]

We have never had, and we do not now have, an *EM energy* shortage or problem. We never will have one, due to nature's bounty in providing us with the giant negentropy of the common dipole and of the charge considered as a composite dipole. Instead, we have the two problems that

(i) only a tiny, tiny component of the available 3-energy flow extracted from the vacuum (decompressed from time-energy) by the source dipole and poured out of the terminals of the power source, is caught and used by the circuit, and

---

[85] We would very much welcome a much more accurate functional expose by a capable higher group symmetry electrodynamicist!

[86] We also point out that none of this nondiverged energy flow component is accounted in conventional electrodynamics and electrical engineering (the closest it comes is Jackson's mention of the transverse current over all space), and neither has it been accounted or even investigated by biophysicists seeking to determine the effects of EM radiation and fields upon biological systems. We strongly suspect (and have done a little work on the problem) that these "force-free fields" and neglected excess "strange energy" flows do have interactions throughout the highly nonlinear cells of the body, and can indeed have profound long term effects on long-exposed biological systems. We particularly stress the potential for time-charging and subsequent *combined scalar and longitudinal photon pair* decay of time-charge excitation in the body. Several major areas of electrodynamics — such as the transverse current effects, Heaviside component effects, and internal Whittaker structuring (engines) effects of EM radiation — have not been investigated by the scientific community.

[87] We accent that time-energy may be regarded as spatial EM energy compressed by the factor c^2. Hence one second = 9×10^{16} joules of EM spatial energy, if decompressed by time-energy interaction with charge. The crude estimate for one circuit of 10^{13} times as much energy being wasted in the Heaviside energy flow component as is caught and used in the Poynting component, is thus quite reasonable when one recognizes that *all* the EM energy flow appearing in the circuit or outside it actually comes from the time domain, hence constitutes *decompressed time-energy.*

(ii) in the closed current loop circuit, half of that small spatial EM energy component that is intercepted and caught, is used by the circuit only to destroy the source dipole and cut off the free flow of EM energy from the vacuum faster than it powers the load.

We have previously discussed this further in several papers {214}.

One must occasionally keep one's sense of humor. The real problem that the Department of Energy should be working on, with massive resources, is simply the alteration of the closed current loop circuit so that it does not destroy the source dipole faster than it powers the loads. Secondly, they should then be working on how to close-loop the COP>1.0 EM systems that readily emerge from that research. Yet it appears that this fundamental energy problem — and the *only* energy problem — has no conventional work being openly done on it anywhere in the open Western world.[88]

3.8 Dark Positive Energy: The Unaccounted Heaviside Component

Now we restate a deceptively hidden thing: The EM field and potential are defined as the observed effect of the deviation from them by an assumed unit point charge. Then that same field and potential are assumed to be in mass-free space, before the interaction with charged matter. Hence the field and potential have each been defined erroneously as both a cause and an effect. This is a gross non sequitur. There is also a total lack of accounting that all EM energy in 3-space comes from the time domain. There is no consideration that every dipolarity and every charge represents a giant circulation of EM energy between time and 3-space. In ensemble, this non sequitur and these omissions are responsible for most of the present formidable foundations problems in electrodynamics.

---

[88] Using the U.S. as an example, the U.S. classified community is highly compartmented, and the real "organization" of a given compartment often consists of laterally connected "same clearance" compartments in multiple outside agencies rather than vertical chain of command through one agency. We accent that some highly compartmented "lateral" classified organizations in the West and elsewhere already have working COP>1.0 systems, and have had them for several decades. Some of these laterals in the U.S. are very dedicated to the duly elected government of the United States, with good national security reasons for the fact that they "officially do not exist". Some are not so dedicated at all, and have their own agendas having little if anything to do with official U.S. government policy. These latter "rogue" laterals often have convenient connections with outside shadowy groups of the High Cabal (Churchill's phrase). A similar situation exists in several other major foreign nations. Substantial electrodynamic technology has been developed (and weaponized) for decades, in secret, and this technology does not appear in the open news, the university texts, or the scientific journals. Such is the case for COP>1.0 EM power systems, which have actually existed in secret for decades.

We submit the following new hypothesis: *The present "definition" of field and potential as a function of interception and divergence by a unit point static charge are only indications of their point intensity, obtained as an estimate after their observed interaction with a unit point static charge.*[89] *Thus the actual field and potential contain not only their Poynting components that interact with the unit point charge, but also their Heaviside nondiverged components that do not interact with it. The true "magnitudes" of the massless field and potential prior to interaction with charged mass are many orders of magnitude greater than the magnitudes of their standardized point intensities* {215}.

In short, we propose that *every field and potential, in its interaction with charged mass, produces not only the accounted energy aspects presently calculated by electrodynamicists and astrophysicists, but also contains Heaviside's vast unaccounted "dark positive energy" component presently discarded.* We further propose that this ubiquitous and unaccounted "dark positive EM energy" component in the interactions of the fields and potentials of the spiral galaxies is what is producing the excess gravity that is holding the spiral arms together. We specifically do this in honor of Heaviside, whose unpublished papers clearly establish that he recognized the overwhelming gravitational importance of his extra and unaccounted energy flow component {172a, 172b, 196}.

3.9 A Short History of the Discarding of the Heaviside Dark Energy

3.9.1 Independent Work by Heaviside and Poynting

Let us now see how the enormity of the EM energy flow from the dipolar source was treated in the early electrodynamic theory, and how it is treated in the received view today.

---

[89] This is recognized by leading electrodynamicists, but not by most electrical engineers building power systems!

In the 1880s after Maxwell was already deceased, Poynting {216} and Heaviside {217} independently (and rather simultaneously) discovered EM energy flow through space. Before that, the concept did not appear in physics. Poynting {216} published prestigiously, while at first Heaviside published more obscurely {217}, then finally more prestigiously {218, 219}.

With respect to circuits, from the beginning Poynting assumed only that small amount of EM energy flow that enters the circuit from the surrounding space. Here are Poynting's {220} own words:

> *"This paper describes a hypothesis as to the connexion between current in conductors and the transfer of electric and magnetic inductions in the surrounding field. The hypothesis is suggested by the mode of transfer of energy in the electromagnetic field, resulting from Maxwell's equations investigated in a former paper ("Phil. Trans., " vol. 175, pp. 343-361, 1884). It was there shown that according to Maxwell's electromagnetic theory the energy which is dissipated in the circuit is transferred through the medium, always moving perpendicularly to the plane containing the lines of electric and magnetic intensity, and that it comes into the conductor from the surrounding insulator, not flowing along the wire. "*

As can be seen, Poynting considered only the energy flow actually entering the wire, and subsequently being dissipated in the circuit. Hence Poynting never considered the huge EM energy flow component around the circuit that is *not* diverged, *misses* the circuit entirely, *does not* contribute to the energy dissipated by the circuit, and is *wasted.* In short, there is a vast "dark energy flow" associated with every dipolar interaction — a huge energy flow component which Poynting never considered.

Heaviside's theory was an extension of what Poynting had considered, and Heaviside also corrected Poynting as to the direction of flow. Heaviside was fully aware of the enormity of the "dark energy" flow missed by Poynting, but had absolutely no explanation as to where such a startlingly large EM energy flow — pouring from the terminals of every dipole, generator, or battery — could possibly be coming from. Consequently, Heaviside was very cautious in referring to it, usually doing so only obliquely in terms of the angles and components. In Heaviside's {221} own words:

> *"It [the energy transferflow] takes place, in the vicinity ofthe wire, very nearly parallel to it, with a slight slope towards the wire... . Prof. Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from the vertical. This difference ofa quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature ofthe electric field in the vicinity of a wire supporting electric current. The lines ofelectric force are nearly perpendicular to the wire. The departure from perpendicularity is usually so small that I have sometimes spoken ofthem as being perpendicular to it, as they practically are, before I recognized the great physical importance ofthe slight departure. It causes the convergence ofenergy into the wire. "*

As can be seen, Heaviside was fully aware that the energy flow diverged into the wire was only a minuscule fraction of the total. He was fully aware that the remaining component was so huge that the energy flow vector remaining — after the divergence of the Poynting component into the circuit — was still almost parallel to the conductors. However, he had no explanation at all of where such an enormous and baffling energy flow could possibly originate.

Had Heaviside strongly stated the enormity of the nondiverged component of the energy flow, he would have been viciously attacked and scientifically discredited as a perpetual motion advocate. His words were measured and cautious, but there is no doubt that he recognized the enormity of the nondiverged EM energy flow component.

We have chosen to call that huge unaccounted component the "Heaviside dark energy component" in his honor, since he actually discovered it. By the word "dark" we mean "unaccounted", which hides it from scientific view. We have also nominated it as the previously unsuspected source of the extra gravity holding the arms of the spiral galaxies together.

3.9.2 Lorentz Disposed of the Problem Rather than Solving It

Lorentz then entered the EM energy flow scene to face the terrible problem so quietly raised by Heaviside. Lorentz understood the presence of the Poynting component, and also of the extra Heaviside component, but could find no explanation for the startling, enormous magnitude of the EM energy pouring out of the terminals of the power source (pouring from the source dipole) {222} if the Heaviside component was accounted. Had he retained and advanced this enormous dark energy flow component, even the great Lorentz would have been castigated as a perpetual motion advocate who did not accept the law of energy conservation. Even today, when one points out that far more energy is pouring out of the terminals of every generator than the small mechanical energy input to the shaft, one is immediately labeled a raving lunatic — even though the Bohren experiment {24} clearly proves the presence of an extra energy flow component not usually intercepted. It is not difficult to understand how such a bombshell would have been received in the 1880s! At least today, we have particle physics and its broken symmetry of opposite charges, together with its active vacuum, to explain where the excess energy comes from and how.

Unable to *solve* the dark energy flow problem by any rational means, Lorentz found a clever way to *avoid* it. He reasoned that the nondiverged Heaviside component was "physically insignificant" (his term) because it did not even enter the circuit. Since it did nothing of any physical consequences, or so he reasoned, then it could just be discarded.

So Lorentz simply integrated the entire energy flow vector (the vector representing the sum of both the Heaviside nondiverged component and the Poynting diverged component) around an assumed closed surface enclosing any volume of interest {223}. *A priori,* this mathematical procedure discards the dark *Heaviside* energy flow component because of its nondivergence. It retains only the intercepted *Poynting* diverged component that enters the circuit.

A century later, electrodynamicists are still happily avoiding the dark energy flow problem by continuing to use the Lorentz integration procedure {224} to dispose of all but the Poynting component that enters the circuit and is then dissipated by the circuit. As a result, the "Poynting energy flow" has come to be loosely regarded as "the" entire EM energy flow, though electrodynamicists find it necessary to give stringent warnings about it. E.g., Panofsky and Phillips {225} state it this way:

> *"...only the entire surface integral of N [their notation for the Poynting vector] contributes to the energy balance. Paradoxical results may be obtained if one tries to identify the Poynting vector with the energy flow per unit area at any point."*

Most electrodynamicists note the freedom to add a vector — few call it an *energy flow* vector, though that is the type of vector being discussed, and one must add apples to apples — which has zero divergence. Jones {226} states:

> *"It* is *possible to introduce the Poynting vector S, defined by S = ExH, and regard it as the intensity of energy flow at a point. This procedure is open to criticism since we could add to S any vector whose divergence is zero without affecting [the basic integration procedure's result]."*

Jackson {227} says it even more plainly, and also uses Lorentz's "no physical significance" argument for disposing of any energy flow vector with a zero divergence. Quoting:

> *"...the Poynting vector is arbitrary to the extent that the curl of any vector field can be added to it. Such an added term can, however, have no physical consequences."*

Any energy flow vector, which is the curl of a vector field, will have zero divergence, by elementary vector algebra. In short, to be pertinent at all, it must be an *energy flow* vector (since energy flow is what S = E x H is all about. Since the curl of any vector has no divergence *a priori,* then any energy flow vector that is a curl of a vector field will be part of the Heaviside dark energy flow component, rather than part of the Poynting energy flow component. It will also be discarded by Lorentz's closed surface integration.

We refer the reader again to Figures 2-5 and 2-6 in Chapter 2, p. 113-114. Electrodynamicists err in assuming that such a divergent free vector (energy flow) can have no physical consequences. If one inserts *additional intercepting charges* into that formerly nondiverged energy flow component, the charges will immediately diverge some of the formerly nondiverged energy flow around them and hence "collect additional energy". In other words, "divergence" is the result of a physical process, accomplished by intercepting charges. We can easily change "nondiverging" energy to diverging energy, by simply interacting it with charges. We can also "bunch" the energy flow lines closer together (as in some waveguides), thereby increasing the local flow intensity and the concomitant energy collected by each intercepting charge in the densified energy flow stream.

So whether the Heaviside energy flow component is of physical consequence or not, depends on whether or not we "add more sail to the sailboat", or "add more sailboats" before the wind. That is most certainly a useful physical consequence, and with a little ingenuity we ought to be able to use it to power loads "for free" or nearly so. There are other consequences also, as we used in deriving the negentropy of the dipole or charge. There the input of a non-Poynting energy flow component certainly has universal and physical significance — e.g., to the curving of local spacetime, conditioning the active vacuum, furnishing all the energy to form the EM fields and potentials from their source charges, etc.

Schwarz {228} expresses it this way:

> *"There will be many opportunities in which the interpretation of E x H as a rate of flow of energy per unit area will be profitable. In most cases of practical interest, such an interpretation is valid, although it must always he kept in mind that only the integral of S over a closed surface can be physically measured... Just how it is that the connections to the energy source, say a battery, are at the ends of the wire, yet energy flows in through the sides, should be pondered by the reader. "*

Note carefully that Schwartz recognizes the difference between the observed (deviated or diverged) part of the EM energy flow and the unobserved nondeviated or nondiverged part of it. However, he did not follow it up and apply it back to the basic "definition" of the field and potential as what are diverged from them.

For recommended changes to the Poynting vector, Jones {229} presents many conditions the changed vector must fulfill. Then he falls back on the Lorentz closed surface method again, but without realizing that he therefore (i) first includes both the diverged and nondiverged component, and (ii) invokes a procedure that arbitrarily discards the nondiverged component. In thus disposing of the problem, Jones says:

> "It *does not seem likely that an expression satisfying all these conditions will be simple........fortunately, we are rarely concerned with the energy flow at a point. In most applications we need the rate at which energy is crossing a closed surface."*

Finally, we note that even today, a debate on what the Poynting vector is or should be, is still politely ongoing. As an example, it has been ongoing for more than 40 years in the *American Journal of Physics* alone as well as in other journals {230a-230i}. One thing is clear: The Poynting vector does not represent the total energy flow at right angles through a square meter in space surrounding the conductors of the electrical circuit. It is, however, the *component orfraction* of that actual energy flow through the square meter in force-field-free form *that will subsequently be diverged and collected in the circuit in force-field form and dissipated in the circuit loads and losses.*

3.10 The Problem of the Accelerating Expansion of the Universe

As briefly stated earlier, we introduce the notion of the supersystem, consisting of three parts: (i) the system and its dynamics, (ii) the active vacuum and its dynamics, and (iii) the active spacetime curvatures and their dynamics. All three components of the supersystem interact with each other.

In Chapter 9, we explain in some detail how a COP>1.0 EM circuit — and especially a COP»1.0 EM circuit — has an induced flow of negative energy (Dirac sea hole current) from the output section of the system, back through the system, and into the input section. Reaching the input section, if the hole flow is not completely filled by incoming electrons, the remaining Dirac hole current will flow on back into the feeder line from the external power supply, and into the distant power supply itself, eating electrons and electron current all the while.

We also explain our reinterpretation of the *unobserved* (non-reacted) Dirac hole with negative energy or negative mass, without time reversal and parity reversal. This dramatically differs from the "as observed" (interacted) Dirac positron with positive mass, and with charge-reversal and parity reversal from the electron. The Dirac hole is "as unobserved and non-reacted" and it is still an *electron* having negative mass and negative energy. The Dirac positron with positive mass, opposite charge, and opposite direction is "as observed and reacted with charged matter". The unobserved Dirac hole is the "cause" of an observational interaction; the Dirac positron is the "effect" of an observational interaction. The two are very different because the cause differs from the effect. Here again, the confusion of cause and effect has had very detrimental consequences in the use of Dirac's electron theory.

Very powerful astronomical phenomena, such as giant astronomical explosions of great energy, involve COP>1.0 and COP»1.0 processes, particularly when the unaccounted Heaviside energy flow is accounted for every field and potential interaction with charged mass. Hence all such phenomena also provide very powerful "explosions" of negative energy

(Dirac sea holes moving outward in the vacuum). The negative energy is the equivalent of negative mass, so long as it is unobserved.

So there is a "dark negative energy explosion" in the Dirac Sea vacuum, accompanying every large astronomical explosion seen through our telescopes and other sensors. This accompanying dark negative energy explosion has not been accounted in astrophysics until now.

We thus live in an astronomical universe of neglected "giant dark negative energy explosions" as well as one of accounted "giant positive energy explosions".

The dark (unaccounted) negative energy radiating outward in the vacuum from gamma bursts, x-ray bursts, etc. produces antigravity. The spreading negative energy, moving at light speed, is a change to the vacuum itself, which also interacts upon spacetime to negatively curve it, producing negative gravity. For those energetic processes of COP sufficiently greater than one, the processes produce more antigravity than gravity. As of this writing, we are in fact filing a patent application on this method of performing and producing antigravity, and it will be filed before this book is published.

Over the 14 billion or so years of the observed universe, the dark negative energy content of spacetime has been slowly changing by the results of giant astronomical explosions and other such processes producing negative energy radiated outward into the Dirac sea vacuum. Slowly the vacuum itself has undergone transformation from the perfect Dirac sea with all holes filled by particles, to a vacuum where there is a slowly growing fraction of negative energy (negative mass) unobserved Dirac sea holes.

Thus, slowly the antigravity generated by the gradual altering of "space" (the vacuum) itself has been growing for some 14 billion years. At some point in the past, the excess antigravity effect accumulating in space reached the equal of the "outside" gravity acting upon energetic explosive astronomical phenomena. From that point in time, the antigravity effect has been increasing slowly. The result is that the present violently explosive astronomical phenomena, by adding a sudden and sharp increase in their external space antigravity, produce acceleration away from all positive mass and positive energy phenomena. Hence astronomers now observe the acceleration of the expanding (positive energy/positive mass) universe, instead of the previously expected slowing (due to positive gravity) of the expanding universe (conventional view).

The antigravity-producing phenomena have not been openly investigated in Western science because Western science has arbitrarily discarded COP>1.0 EM systems, primarily with the Lorentz regauging of the Maxwell-Heaviside equations. Hence, Western science has not experimentally studied the type of system that produces the effect in the first place.

On the other hand, it is possible to directly investigate the antigravity-producing phenomenology on the laboratory bench, by utilizing overunity EM systems with very high COP (e.g., COP = 10^6 or more). The Sweet experiments {231} did in fact produce such phenomenology on the laboratory bench, but these experiments were privately performed under proprietary conditions {232}. We discuss the Sweet experiments and device in a later chapter.

In science, such experiments are quite properly deemed interesting but anecdotal until replicated independently. Similar experiments should certainly be repeated by the scientific community, using various interactions (some given in this book) that develop COP>1.0. The problem is nowhere near as difficult as hot fusion, but it has not been tackled and overcome because of (1) the prevailing and erroneous scientific mindset against COP> 1.0 electromagnetic processes, and (2) the present "kill" of the supersystem and its effects in the standard U(l) electrodynamics model.

In Chapter 8 we further discuss the antigravity effect in practical or potentially practical systems where COP » 1.0.

3.11 In Conclusion

In this Chapter, we have stressed what we believe to be a great new principle of giant negentropy, leading to direct and easy extraction of EM energy from the vacuum in copious quantities. That principle is that nature retains *her preferred 4-symmetry* of EM energy flow, while breaking 3-symmetry of EM energy flow and the symmetry of EM time-energy flow on the fourth axis. It is implemented by making a common dipole or charge.

The electrical energy problem is *not* due to any difficulty in inducing giant and indefinitely continuing EM energy flows from the vacuum! The easiest thing in the world is to extract EM energy — any amount, anywhere, anytime — from the seething vacuum, by making a simple dipolarity. The problem is only in then intercepting and collecting some of the energy to dissipate in loads, and doing it *without* using half the collected energy to destroy the dipole producing the energy, faster than the loads are powered. In other parts of this book we have given some of the many mechanisms that can be applied to attack this "power the load without killing the source dipole" problem, and have also included such in papers in the literature {233a-233u}.

We have pointed out the implications of the new principle of negentropy in EM energy flow, and how all our circuits and electrical power systems actually use it but then arc designed so as to kill the negentropic process's potential for self-powering electrical power systems.

Presently we are told by the conventional scientific community that the dream of freely extracting EM energy from the vacuum, and using it to efficiently and easily power our electrical needs cleanly, is either a fool's concept of perpetual motion, or the science of the next century. It is neither; it is good physics, and it is already *present and misused* in all our electrical power systems and circuits — and always has been. Unlimited electrical energy from the vacuum is here and now, if we but overcome the prevailing mindset and grasp it.

We have argued that all electrical loads and circuits are now and always have been powered by just such EM energy extracted directly from the vacuum by the giant negentropy principle, evoked by the source dipole. We have removed the artificial and erroneous notion that batteries and generators provide some of their available 3-space internal energy to the external circuit. Instead, we have emphasized that batteries and generators dissipate their available energy to form and then continually reform their source dipole, and nothing else. Once their source dipolarity is formed, a battery or a generator is nothing but an energy gate, gating and transducing enormous EM energy from the vacuum. We only have to keep inputting energy to keep reforming the dipole that our external circuit is designed to destroy faster than the load is powered.

In dealing with the function of the dipole in extracting the EM energy from the vacuum and pouring it out to power the circuit, we have shown errors in present energy flow theory and how they came to be made. We also indicated the enormous magnitude of the EM energy flow actually **extracted** and present with every conventional circuit, but wasted and **arbitrarily** removed from accountability. Some processes for intercepting, co**llecting,** and using additional amounts of this available Heaviside "dark EM energy" flow component, surrounding every circuit, have been mentioned, and references have been given to other papers more fully discussing such processes. In a previous paper {234} we have proposed that the Heaviside dark energy, accompanying all EM field-charge interactions, is responsible for the excess gravity observed to be holding the spiral arms of those distant spiral galaxies intact — in short, we have proposed an unaccounted Heaviside *dark positive energy* solution to the well-known *dark matter* problem in astrophysics.

The ordinary closed current loop circuit inherently yields a system COP<1.0, because it destroys its source dipoles faster than it powers the load. Indeed, the circuit self-enforces the Lorentz symmetrical regauging condition during discharge of its free excitation (potential) energy.

In addition, we have presented an hypothesis to explain that negative energy (negative mass) created in the Dirac sea vacuum by giant astronomical explosion phenomena produces the excess antigravity in the space surrounding those entities that is accelerating the expansion of the universe. We have also indicated how this hypothesis can be investigated and tested on the bench. We have provided one experiment whose results did support the hypothesis.

The present world energy crisis is real and increasing, while at the same time the availability of oil is beginning to decrease and oil becomes more costly.[90] Unless a substantial fraction of the "electricity from oil" curve is shifted to "electricity from the vacuum" rather quickly, the economic collapse of the Western world, followed by concomitant collapse of other economies, may ensue within a surprisingly few years, and even within the present decade.

---

[90] As this was first being written in Dec. 2001, there was a temporary relief in the oil crisis, due to the sudden decrease in travel, air traffic, etc. after the terrorist attack on Washington and New York on September 11, 2001. As the present war on terrorism has continued, the MidEast is now heating up toward another war, and oil prices are again trying to climb. Iraq, Iran, and Libya called for an oil embargo on the United States, for example, which would almost immediately plunge the nation into an oil crisis. Meanwhile, should effective terrorist attacks occur upon the highly vulnerable energy distribution infrastructure of the United States and other allied nations, the oil crisis will again emerge catastrophically and almost immediately thereafter. It is held off for now, only because Russia under Putin is seeking U.S. financial assistance to further develop its rich oilfields. When OPEC votes to curtail production and then curtails it, Russia opens the spigot and keeps the prices down. Putin is seeking to become a trusted U.S. oil partner, in the role long played by Saudi Arabia but now declining due to perceived Saudi payments to terrorist organizations and reluctance to close down terrorist activities in their own country. So in a strange way, a new alliance between the U.S. and Russia has prevented the oil and energy prices from escalating right through the roof.

We believe that a transfer of much of the "electricity from oil" curve to an "electricity from the vacuum" curve can be quickly accomplished, if the government and the scientific community can be induced to move with full priority and vigor in the directions indicated in this book. However, to get that done, it apparently will be done over the writhing bodies of our leading scientific organizations, literally dragging the organizations to face the proven fact of giant negentropy of every charge and dipole in the universe over their strenuous objections.

As a major objective, we fervently hope that undergraduate students, graduate students, post-doctoral scientists, other scientists, engineers, and environmentalists will interest themselves in these new principles and viewpoints. We have called attention to the new electrodynamics pioneered by AIAS, Evans, Sachs, Barrett, Lehnert, and others, and indicated the capability of this emerging electrodynamics to model this new *energy from the vacuum* functioning.

It has been said that no idea is accepted until its time has come. We strongly believe that the time for cheap, clean extraction of electrical energy from the vacuum is an idea whose time has arrived. Hopefully our efforts in this book will at least conceptually show the interested reader the outline and course that energy from the vacuum can and will take in the scientific community and in the world at large.

We accent again that modeling the extraction of EM energy from the vacuum demands the use of higher symmetry electrodynamics, particularly a theoretical model capable of expressing the interactions of the three components of the supersystem. Barrett and Grimes {235} said it beautifully:

> *"In the case of electromagnetism, the theory was first simplified before being frozen. Maxwell expressed electromagnetism in the algebra of quaternions and made the electromagnetic potential the centerpiece of his theory. In 1881 Heaviside replaced the electromagnetic potential field by force fields as the centerpiece of electromagnetic theory. According to him, the electromagnetic potential field was arbitrary and needed to be "assassinated" (sic). A few years later there was a great debate between Heaviside and Tait about the relative merits of vector analysis and quaternions. The result was the realization that there was no need for the greater physical insights provided by quaternions if the theory was purely local, and vector analysis became commonplace.*

> *The vast applications of electromagnetic theory since then were made using vector analysis. Although generations of very effective students were trained using vector analysis, more might be learned physically by returning, if not to quaternions, to other mathematical formulations in certain well-defined circumstances. As examples, since the time when the theoretical design of electromagnetism was frozen, gauge theory has been invented and brought to maturity and topology and geometry have been introduced to field theory. Although most persons view their subject matter through the filter of the mathematical tools in which they are trained, the best mathematical techniques for a specific analysis depend upon the best match between the algebraic logic and the underpinning physical dynamics of a theoretical system. "*

The problems we face today are ever more demanding of the necessary EM change pointed out by Barrett and Grimes, along with many others. Let us hope their words are heeded. Given rapid development of higher group symmetry electrodynamics models, we shall very rapidly see the solutions to many previously intractable human problems — including cheap clean energy from the vacuum, understanding and directly using antigravity as well as gravity, a revolutionary new medical therapy, a new chemistry and biochemistry, and a new appreciation for the electromagnetic interactions of mind and matter.

Let us hope this new scientific revolution comes swiftly.

Chapter 4

Setting the Stage for Understanding Overunity Power Systems

[Dissipative structures] *"This leads precisely to a distinction between 'equilibrium structures' [which may be understood in terms of classical equilibrium thermodynamics] and 'dissipative structures'. The latter are formed and maintained through the exchange of energy and matter in non-equilibrium conditions."* [I. Prigogine] {236} [Internal brackets added for clarity.]

[The energy in space itself] *"What might appear to be empty space is, therefore, a seething ferment of virtual particles. A vacuum is not inert and featureless, but alive with throbbing energy and vitality. A 'real' particle such as an electron must always be viewed against this background of frenetic activity. When an electron moves through space, it is actually swimming in a sea of ghost particles of all varieties — virtual leptons, quarks, and messengers, entangled in a complex melee. The presence of the electron will distort this irreducible vacuum activity, and the distortion in turn reacts back on the electron. Even at rest, an electron is not at rest: it is being continually assaulted by all manner of other particles from the vacuum."* [Paul Davies] {237}

[Modernizing the fluid analogy of EM]. *The intensity of the vacuum virtual particle flux may be treated as the intensity of a potential, thus by analogy corresponding to pressure in a fluid. Polarity (charge) is a process for establishing a change in the intensity of the vacuum potential — and thus a change in the ambient pressure of the vacuum fluid. Difference between vacuum pressures at two separated points creates a force upon any charge placed between those two points. Just as a differential in pressure between points in a fluid induces currents, a differential in vacuum pressure induces energy flow currents in the vacuum fluid, which are called 'fields'. All*

EM circuits, charges, potentials, and forces in electrodynamics involve altering the vacuum to include altering its pressure and thereby producing the forces that the resulting pressure gradients (fields) induce upon charges. All observable EM energy in space is EM energy intercepted and collected on charges, and it comes from the vacuum via these fundamental vacuum-engineering dynamics." [T. E. Bearden, private communication to a colleague, 2001].

[Impact of EM energy from the vacuum.] *"If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete — and so would many of our worries about environmental pollution." "Don't sell your oil shares yet — but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development: 1. It's crazy! 2. It may be possible — so what? 3. I said it was a good idea all along. 4. I thought of it first."* [Arthur C. Clarke] {238}.

[Broken symmetry of opposite charges.] *"Since nonobservables imply symmetry, any discovery of asymmetry must imply some observable. The experiment of Wu, Ambler, Hayward, Hoppes and Hudson... established the asymmetry between the positive and negative signs of electricity."* [T. D. Lee] {239}.

[On implications of asymmetry of opposite charges of a dipole.] *"In particle physics, the proven asymmetry of opposite charges establishes that the source dipole — with opposite charges on each end—* is *a broken symmetry in the violent virtual photon flux of vacuum. The dipole charges constantly receive and absorb virtual photon energy from the seething vacuum, and re-emit that excitation energy. By the very definition of broken symmetry, some of this absorbed* *virtual* *energy must be integrated into* *observable* *3-space EM energy, and emitted by the dipole in observable form, in all directions in 3-space. Hence there is a rigorous basis for the source dipole extracting virtual EM energy from the vacuum, integrating it, and continuously pouring it out in*

> *observable and usable EM energy form. It has been in particle physics since 1957, being certified by the award of the Nobel Prize to Lee and Yang in that year. "* [T. E. Bearden, private communication to a colleague, 1995.]
>
> [On powering an electromagnetic circuit] *"In the United States there is not now, and there never has been, a single electrical engineering department, professor, or textbook that knows and teaches what directly powers an electromagnetic circuit. Neither is it in any publication of the National Academy of Sciences, National Science Foundation, National Academy of Engineering, Department of Energy, or the great national laboratories. Yet the proof that every circuit is powered by EM energy extracted directly from the vacuum has been in particle physics for 45 years, since the award of the Nobel Prize to Lee and Yang in 1957, following the experimental proof by Wu et al. earlier that same year. "* [T. E. Bearden, private correspondence, 2002].

4.1 Introduction

In this chapter, we present a selection of potential overunity notions, some inventions, and some experiments. Its purpose is to show some of the more relevant work that has been done, and hopefully to stimulate new thinking by the interested reader. Much of the work addressed is by other inventors and scientists, and we give an appropriate commentary from our viewpoint. Not all the concepts presented can be clearly explained, though some can and are. This is not intended to be a "kit of parts" or an "assembly instructions" type of presentation. For that, the interested researcher will have to rely on his or her own ingenuity and lots of bench experiments, because no such "kit of parts with instructions" for COP>1.0 electrical power systems exists at present {240, 241}.

Because of the broken symmetry of its opposite charges, we do know that any dipolarity — such as a permanent magnet, a charged flat-plate capacitor, an electret, or a simple EM scalar potential — continuously pours out EM energy circulation extracted from the vacuum. E.g., see Figure 4-1, which shows a very simple "free energy generator" that will generate a flow of EM energy indefinitely, even by ordinary U(1) electrodynamics theory (by standard Poynting theory). E.g., quoting Buchwald {242}:

> *"[Poynting's result] implies that a charged capacitor in a constant magnetic field which is not parallel to the electric field is the seat of energy flows even though all macroscopic phenomena are static."*

This is easily seen by applying the standard Poynting formula, S = E x H. With the H-field of the magnet at right angles to the E-field of the capacitor or electret, the Poynting flow of energy S is maximized and orthogonal to E and to H.

The problem is not in extracting energy - even enormous energy — from the vacuum The problem is collecting and using the energy to power loads, without destroying the source dipole

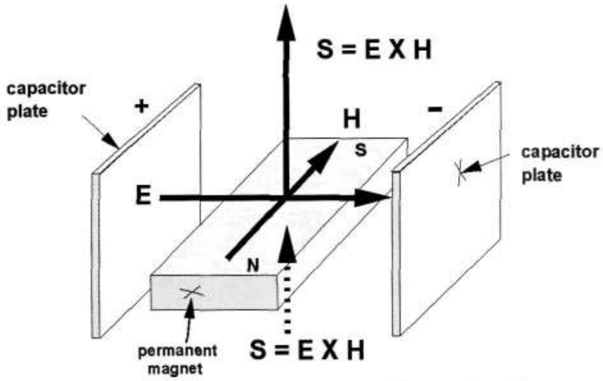

EM energy continually flows from this "static" arrangement, without any further input of energy by the engineer.

Figure 4-1 Poynting generator providing endless EM energy flow from the vacuum.

Buchwald errs, e.g., in assuming the macroscopic phenomena are static. They are not; they are highly dynamic in the same sense that a perfect macroscopic whirlpool in a river, seemingly static, nonetheless involves a continuous flow of water through it and a continuous dynamics. So far as can be observed, it appears there is no such thing as a truly "static" object. As an example, just to "exist" an object must "travel through time", which is itself a special kind of dynamics in the time domain.

A beautiful explanation of what is meant by "static" is given by Van Flandern[91]:

> *"To retain causality, we must distinguish two distinct meanings of the term 'static'. One meaning is unchanging in the sense of no moving parts. The other meaning is sameness from moment to moment by continual replacement of all moving parts. We can visualize this difference by thinking of a waterfall. A frozen waterfall is static in the first sense, and a flowing waterfall is static in the second sense. Both are essentially the same at every moment, yet the latter has moving parts capable of transferring momentum, and is made of entities that propagate."*

Now suppose that the capacitor in Figure 4-1 is rotated until E and H are either parallel or anti-parallel. In that case, the Poynting theory predicts zero EM energy flow. That is an incomplete assessment; it means that there is no overt field energy flow. There is, however, still quite a lot of covert or "infolded" field energy flow ongoing anyway, inside the static fields E and H themselves.[92] To see this, we first turn to the static potentials.

Between the charged plates of the capacitor or electret, there exists an electrostatic scalar potential ϕ. Between the polar ends of the permanent magnet, there exists a magnetostatic scalar potential Φ. We may independently decompose both ϕ and Φ via Whittaker decomposition {243}, applying a necessary correction {244} to Whittaker's interpretation. Then in all space surrounding the two dipoles, we have a hidden harmonic set of bidirectional longitudinal EM wavepair flows of energy comprising each of the static potentials ϕ and Φ filling all space[93] around the assembly in Figure 4-1. Each of those potentials ϕ and Φ is formed of a combination of scalar (time-polarized) and longitudinal photon pairs, as strongly argued by Mandl and Shaw {245} and as previously discussed in an earlier chapter. Further, each of the "static" fields can be decomposed into two scalar potential functions, as shown in 1904 by Whittaker {246}. Then each of the base scalar potentials further decomposes into the set of time-polarized EM wave and longitudinal EM wave pairs. The result is that every "static" field or potential in the universe is actually a set of longitudinal EM wave dynamics.

---

[91] Tom Van Flandern, "The speed of gravity — what the experiments say," *Phys.* ***Lett.*** *A.*, Vol. 250, Dec. 21, 1998, p. 8-9.

[92] Indeed, by the asymmetry of its opposite charges, each of the dipoles is co**ntinuously** pouring out EM energy extracted from the seething vacuum.

[93] In the conventional view. We corrected that view in {12}, to agree with quantum field theory and to use a *cause and effect* instead of the *two effects* used by Whiltaker.

Further, in each coupled Whittaker wavepair as reinterpreted, there is perfect correlation between the time-domain EM energy wave (scalar or time-polarized EM wave) and the 3-space longitudinal EM wave. So even a "static" EM field, such as from the permanent magnet and from the charged capacitor or electret, does involve an enormous number of hidden internal EM energy flows in longitudinal and scalar EM waveform.[94]

---

[94] One of the great omissions of modern Western electrodynamics is the failure to follow up on direct engineering of this "vast, internal, infolded, far more fundamental" scalar and longitudinal electrodynamics hidden inside all present overt EM potentials and transverse EM fields and waves. Russian secret weapon scientists under the former KGB called this internal electrodynamics the "information content of the field" and have strongly weaponized it. A deception plan has been used to successfully convince the West that the Russians are speaking only of conventional spectral analysis. Use of this internal EM structuring, in the form of longitudinal or scalar wave interferometry, formed the first giant superweapons deployed by the former Soviet Union in 1963. To see a proof that ordinary EM fields and waves and potentials can be formed at a distance, *arising out of the local spacetime in the interference zone,* see M. W. Evans et al., {79}. Longitudinal EM waves easily pass through the entire earth and ocean relatively undiminished, since mass is mostly empty space filled with fields, potentials, and waves — i.e., filled with incredible numbers and bundles and dynamics of longitudinal and scalar EM waves. Mass is thus a great superhighway for the passage of longitudinal EM waves, and longitudinal interferometry can be performed right through the earth and ocean, on the other side of the earth from the transmitters. Such interferometers can also easily focus inside the earth or beneath the surface of the ocean. Hence in April 1963 the first giant Soviet LW interferometer deployed was used to kill the U.S.S. Thresher beneath the ocean, off the East Coast of the U.S. Intense production of EM signals — emerging everywhere in the space in that underwater interference zone placed around the submarine — filled the sub and its electronic controls. This jammed the electronic controls of the sub, rendering it helpless. The hapless sub then sank to crush depth and imploded. The signatures are clearly present in the jamming and cessation of multiple electronic systems functioning in the sub's surface companion, the U.S.S. Skylark (which was in a sort of surrounding "splatter" zone and survived because it was not individually targeted). Today at least 10 nations have scalar or longitudinal wave interferometer weapons, to one extent or another, other nations are working on them, and the Yakuza also has them. These were apparently the weapons

Ultimately there is no frozen waterfall, but only the dynamic waterfall always reappearing the same but filled with dynamic movement and energy flow.

These latter EM energy flows *infolded inside* the scalar potentials and the **fields** exist for a single charged capacitor or electret alone, or a single magnet alone. They also exist for any two points in a circuit that have differing electrical or magnetic potentials with respect to the ambient vacuum potential (energy density). This latter fact, of course, was part of the secret of Kron's *open path.* Indeed, every dipolarity in the universe — between any two potentials or any two differing charges anywhere — is an open path and an energy flow source. E.g., quoting Kron {247}:

> *"...the missing concept of "open-paths" (the dual of "closed-paths") was discovered, in which currents could be made to flow in branches that lie between any set of two nodes. (Previously— following Maxwell— engineers tied all of their open-paths to a single datum-point, the 'ground'). That discovery of open-paths established a second rectangular transformation matrix... which created 'lamellar' currents... " "A network with the simultaneous presence of both closed and open paths was the answer to the author's years-long search. "*

As can be seen, Kron had discovered how to create currents — *currents could be made to flow* — in individual branches between any two nodes in **the** circuit, physically connected or not. We point out that deliberate use of Kron's "open path" in a network circuit constitutes a strong violation of **the** standard closed current loop circuit's self-enforcement of the Lorenz/Lorentz symmetrical regauging condition. Hence it is a magnificent method for achieving COP>1.0 circuits. It is the belief of this

confirmed in 1997 by Secretary of Defense Cohen in these words: *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts."* [Secretary of Defense William Cohen at an April 1997 DoD counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.]. As the reader can see, the Cold War **was** not so "cold" after all. This mere "tip of the iceberg" should also give the reader a little more insight as to why a war on terrorism is so necessary to all civilized nations.

author that such "open path" functioning in his patented circuits was part of Tesla's secret, though he did not think of it in those terms. However, Tesla's ability to shuttle potential energy independently in his circuits — invisible to a tensor analysis — was rigorously shown by Barrett {124}. Barrett extended the method and obtained two patents in the field of communications {286a, 286b}.

Thus the COP>1.0 problem *is not* in the part or function (the dipole or the charge) that extracts real EM energy flow from the vacuum. That is extraordinarily simple and straightforward. Engineers and scientists have been using it for two centuries without understanding what they were doing. Every charge, dipole, charged capacitor, electret, and permanent magnet already continuously extracts EM energy from the vacuum. The two-fold COP> 1.0 problem is in (i) *separately intercepting and catching (in an external circuit) some of the copious free flowing energy from the "extractor"*, and then (ii) *separately dissipating that collected energy in an external load without killing the "extractor" itself* — or at least only killing it more slowly than the load is powered. The most elegant way to violate the closed current loop circuit is with Kron's open path, typified by Tesla/Barrett potential energy shuttling. Since vector and tensor analysis will not even show it or model it, then a higher group symmetry electrodynamics is imperative in modeling COP>1.0 EM systems and their functions. Much of that work remains to be done.

Our main purpose in this chapter is to point out some selected efforts and approaches by scientists, engineers, and inventors. We will give a brief description and a few details, and leave it to the researcher for further experimentation and reflection. Hopefully, knowing what some others have tried and the results they achieved will prove helpful to the serious researcher.

We also strongly encourage the reader to again read this section after studying Chapter 9. In any COP>1.0 system, the excess energy must come from the "active environment". The active environment of an EM system consists of (i) the active vacuum and its interactions and dynamics, and (ii) the active local spacetime curvatures and their interactions and dynamics. Contrary to naive belief and the received view, overunity electrical power systems — particularly very high COP systems — are neither simple nor easy, but are extraordinarily difficult and highly nonlinear even when their appearance is deceptively simple. Chapter 9 will give a brief expose of some of the difficulties for higher C OP systems. There we will refer again to the intricate circuit analysis work of Gabriel Kron {248}, to Dirac sea electron theory {249} both prior to observation and after observation, and to the need for higher group symmetry O(3) electrodynamics by Evans {250} which is a subset {251} of the electromagnetics used in the unified field theory of Sachs {252}.

Chapter 9 indicates how mathematically complex these systems really are, even when they appear deceptively simple. The phenomenology is highly complex, and it has not yet been scientifically explored because of the erroneous received view that COP>1.0 EM systems are not permitted by Maxwell's theory or the laws of physics. That is a non sequitur. More exactly, *they are not permitted in that equilibrium class of Maxwellian systems described and selected by Lorentz's arbitrary symmetrizing of the Maxwell-Heaviside equations, following their previous symmetrizing by Ludwig Valentin Lorenz* {253}. Prior to applying the Lorenz-Lorentz symmetrical regauging, the Heaviside equations — which in field form are already a truncation to four vector equations of Maxwell's 20 quaternion-like equations — do prescribe both disequilibrium COP>1.0 and equilibrium COP<1.0 Maxwellian systems. The variables are not separated (potential formation) and the equations are difficult to solve, usually requiring numerical methods. With Heaviside's truncation and Lorenz-Lorentz regauging applied, the resulting "special case" equations — now for only that class of COP<1.0 Maxwellian systems that receive and use no net input from their active environment — continue to be used because closed analytical solutions can usually be had and numerical methods can be largely avoided.[95]

So with tongue in cheek, we humorously remark that COP>1.0 electrical power systems have not been developed or known because the early scientists were a bit lazy and wished to avoid ubiquitous use of numerical methods. These days, with computers, proper math programs, and sharp graduate students to easily crunch the numbers, there is no longer any reason for continued arbitrary discard of Maxwellian COP>1.0 systems.

---

[95] For some reason, electrodynamicists ignore the fact that the Lorentz regauging does assume two *additional* equal and opposite energy exchanges of the system with its external active environment, but carefully selected so that all the excess energy received in the system is purely in the form of a stress potential. Hence this energy continually does internal work on the system to produce additional stress, but cannot translate electrons through an external load to power it. Also, this regauging condition represents an effective rotation of the frame of the described system out of the laboratory frame. To then argue that this severely altered system (increased stress, constant additional internal work being done on the system, frame rotation, altered potential energy) is the same system prior to Lorentz regauging, and that such a system is not receiving excess energy from its active vacuum environment, is naive in the extreme and a non scquitur.

For such reasons, no legitimate theory of COP> 1.0 EM power systems has been developed, except for light shed by Prigogine {254a-254f} and others {255a-255c} as to the necessary active environment, disequilibrium nature, and disequilibrium thermodynamics of COP>1.0 systems. Further, the energy flow theory (Poynting's theory) has not been further extended to include the Whittaker decompositions {243, 246} as should long ago have been done. Neither has it been extended to account for the arbitrarily discarded giant Heaviside nondiverged energy flow component.

The researcher should be aware that the energy input to a system is always in the voltage[96] (the *potential;* one input volt of potential will cause one joule of energy to be collected upon each interacting unit point electric charge in the system; one volt of drop in potential results when one joule of collected energy is dissipated per unit point collecting charge in the system.). The fields and their energy are functions of the potentials and their energy, not the other way round. After potentialization of the circuit, the subsequent movement of the potentialized Drude electrons, as current in the system, is connected with (and part of) the *dissipation* from the circuit of previously collected/intercepted energy from the vacuum.

4.2 Demonstration Gedanken Experiment

4.2.1 Work and Energy

Doing work is defined as *changing the form of energy.* No energy is lost per se when work is done; the energy is only changed from one form to another. All the energy since the beginning of the universe is still present,[97] and it is still repeatedly changing its form to do work, over and over. The point is, a single initial joule of energy can be "used" (changed in form) to do more than one joule of work, if multiple changes of form occur serially. *One joule of energy changing its form gives one joule of work, only if a single and total change of form of that joule of energy occurs and then no further change of form occurs.* If the joule changes form once, a joule of work is done and a joule of energy still remains in a second form. If this second form of the joule of energy is not lost, but is retained and then changed in form, a second joule of work is performed and there is still a joule of energy left in different form. And so on. Tricking a system to self-induce multiple iterative form changes of each joule of input energy (or at least of some fraction of each input joule), without losing the energy after each change of form, is one way of providing COP>1.0 in a system.

---

[96] The reason is that all 3-space EM energy — in the circuit or elsewhere — must come from the time domain to each 3-space point and return from that point to the time domain. The scalar potential (voltage) actually is an ongoing process whereby, for any interacting charge, energy in the time-domain and energy in 3-space are functionally combined in that 4-circulation at every point dipole in space. In quantum field theory, e.g. as shown by Mandl and Shaw {19}, the *combination* of the time-polarized photon and the longitudinal photon in 3-space is observed as the instantaneous scalar potential, but neither photon is *individually* observable. We comment that an "observable" 3-space potential is the result of an iterative set of *d/dt* differentiations of the ongoing "combination" of scalar (time-polarized) photons having their energy oscillations in the time domain, with longitudinal photons having their energy oscillations in 3-space along their line of motion. Hence each *d/dt* differentiation of the ongoing 4-space combination yields the instantaneous 3-space potential. Cumulating, one has the ordinary 3-space potential or "voltage" of macro electrodynamics.

Remember that COP (coefficient of performance) is best defined as *the usable work output obtained, divided by the EM energy that the operator inputs.* Since "usable work output" is just usable "change of form of the energy", it can be seen that multiple changes of form of every joule of energy input by the operator can lead to COP>1.0.

This illustrates that one must be careful in applying the law of conservation of energy. For a thorough analysis, several features must be identified and accounted: (i) The flow of the energy impinging upon the collection, (ii) the collector itself and the type of collection process utilized, (iii) the linear or nonlinear operation of the collection/collecting process, (iv) the dissipation process used to dissipate the collecting mechanism, (v) what happens to the energy flow that was collected and then its collection "dissipated", (vi) what type of energy flow was collected, and then its collecting "dissipated", (vii) what happens to the collector after the dissipation interaction, (viii) what happens to the energy in different form after each previous change of form, (ix) whether single-pass or multiple-pass of the energy flow is used, (x) whether single collection or multiple collecting is used, and (xi) the effects of any nsymmetrical regauging of a local component or region (e.g., as in simply increasing one of the potentials of the region) that in turn acts as a free energy flow.

---

[97] However, "present" means in spacetime, not just in 3-space. The EM energy we think of as being in 3-space is macroscopically in continuous circulation from the tine domain to 3-space and back to the time domain. Microscopically it is in *continual* circulation.

See again Figure 2-4 in Chapter 2. In general, the available energy flow surrounding a circuit will be far greater than the intercepted and collected energy in that circuit, particularly in a "single pass" collection — i.e., where the impinging energy only passes once. In analyzing a system or circuit, one must be very careful to ascertain whether the system utilizes any mechanisms to enhance and increase the energy collection fraction. If it does, then a necessary (but not sufficient) criterion for COP>1.0 has been identified in the system. Letokhov's negative absorption of the medium (i.e., excess energy emission) {256} and Bohren's experiment {257} are examples. Bohren's experiment is easily replicated by any nonlinear optics lab, and outputs 18 times as much energy as the conventionally calculated energy input.[98]

If the impinging Heaviside nondiverged component of the energy flow beyond the collecting/collector is rerouted and passed back to impinge again in the system — even on another collector or other collectors — this must also be taken into account.

Conventionally, the "magnitude" of the fields[99] E and H in the Poynting (collected) energy flow S = E x H are "defined" by the collection from the fields by an assumed unit point static charge at each point in space. In short, the conventional fields have already been "defined" as a standard linear collection upon standard charged *mass, from the field entities as they*

---

[98] Energy conservation is not violated! Instead, the conventional calculation assumes a static intercepting unit point charge, which has the minimum reaction cross section in the stream of bidirectional longitudinal EM energy flows comprising the potential, etc. By placing the intercepting charged particles into particle resonance, the particles' reaction cross section is increased because each resonating charge sweeps out a greater geometrical cross sectional area. Hence the charges collect more energy from the energy flow comprising any "field" (actually field intensity) or "potential" (actually potential intensity) than is assumed in the standard input energy calculation. In short, Bohren's experiment takes advantage of that arbitrarily discarded Heaviside energy flow component that remains and is present but not considered in the conventional Poynting calculation and field intensity calculation. The experiment in fact proves the existence of the unaccounted but ubiquitous giant Heaviside nondiverged EM energy flow.

[99] At best, the 3-space point intensity of each 4-space field is so defined, for a forthcoming reaction with unit point static charge, and where the forthcoming reaction does not involve lingering time - charging but only simple time-excitation followed by full decay of the time-excitation via normal photon emission.

exist in 4-space prior to interception and collection, and where the collector itself is absolutely fixed and "static". So nonlinearity in energy flow collection — by additional disruption of the impinging energy flow induced by the dynamics of the collector itself— may easily increase the "reaction cross section" of the collector. The static reaction cross section of a unit point static coulomb and the dynamic reaction of a unit point resonant coulomb may differ appreciably.

Without stating it this clearly, Bohren has shown that the interaction cross section for a single-pass collection by one static particle may be increased by a factor of 18 in a given example, *by inducing resonance and thus increasing the "collection" dynamics of the collecting particle* in the impinging S-flow {257}. The work has been independently substantiated by Paul and Fischer {258}. It also is a direct proof of the existence of Heaviside's neglected nondivergent energy flow component surrounding every field and charge interaction.

So indeed there are COP>1.0 EM processes in the literature. The Heaviside component also pouring from the terminals of every generator and battery, already proves that *every generator and battery ever made already was and is a COP> 1.0 energy transducer, because far more energy pours from the terminals of the generator and battery than the shaft energy input to the generator or the chemical energy dissipated in the battery.*

4.2.2 Extending the Work-Energy Theorem

The standard work-energy theorem in mechanics may be conventionally stated as: *The change in the kinetic energy of a particle is equal to the net work done on the particle by the net force acting upon it.* As in any "model", here there are assumptions that are seldom taken into account. First, the collected/collecting kinetic energy of a particle is due to its increased interaction rate with the vacuum flux, because of its acceleration through that flux.[100] In mechanics the increase in flux interaction (and hence "Newtonian 3rd law resistance" is just treated as a sort of resistance to acceleration, and known as *inertia.*[101] The actual energy of the vacuum flux interaction rate is a function of the square of the particle velocity through that flux — hence the formula $K = 1/2\ mv^2$ for the kinetic energy K of a mass m moving through the flux of space at velocity v. Relativistically, the mass of the particle is moving through the ambient potential flux, or we may view it from the particle's frame of reference where the ambient potential flux is moving through the particle but interacting with it. Only a single pass of the flux impinges upon the "particle collector", so there is only a "single-pass" type of energy collection ongoing.

---

[100] As we cover elsewhere in this book, any charge (which, together with its clustering virtual charges of opposite sign is a dipolarity) is a broken symmetry in the fierce flux of the vacuum. This means that every charge in the universe absorbs virtual energy from the vacuum, coherently integrates some of it into observable energy, and pours out that observable energy in all directions. The negative charge does that for positive energy; the positive charge does it for negative energy. The latter can be better interpreted as the return of positive observable energy from space to the positive charge, thence disintegration of that positive energy back to virtual energy, and subsequent re-radiation of that virtual energy back to the vacuum. One **can** in fact model the **Bohren** experiment in terms of the extra vacuum flux

All energy in the vacuum exists in "vacuum flux" or "change in vacuum flux" form. There is no force in the vacuum, although there is polarization that will produce a force upon a charge when one is introduced. The force occurs only after the "change in vacuum flux" that we call a field is interacting with a mass. The force is *already* a change to the mass, a priori, since mass is a part of force by $F = d/dt$ (mv). If energy is then added by a AK energy flux (usually mistakenly called a force) performing work upon mass m, this constitutes an additional energy flux or flow impinging on the

---

encountered and transduced by the resonant charge, as compared to flux encountered and transduced by the static charge. In the same vacuum flux density, a resonant charged particle will intercept and absorb more virtual photon energy than the same particle in static condition. Hence its broken symmetry will convert more absorbed virtual photon energy into real observable emitted photon energy.

[101] If one wishes to understand that inertial reaction, here is how it works. Consider the virtual charges of the vacuum as an analog to the Drude electron gas in a conductor. At the instant an observable charge starts to move, it is momentarily affecting all the surrounding virtual charges — an enormous number, which are initially repelling and attracting it in all directions intensely. To move, the observable particle must create a change in that entire participating virtual particle "gas". Just as the Drude electron gas, the vacuum "gas" of virtual charges has a certain relaxation time constant. So a finite time delay occurs before the vacuum gas relaxes and movement of the observable charge occurs. As the relaxation continues apace, the particle being acted on by a force accelerates. Even mass with a net charge of zero is filled with charges, and even the neutrons in the atoms are continuously changing into protons and back, etc. So all mass continually exists in the "charge and charge-changing" state. Inertia is in fact due to the relaxation time curve of the vacuum virtual charges "gas". We dealt crudely with this in Bearden, "Quiton/Perceptron Physics," NTIS Report AD-763-210, 1973, p. 11; hopefully a far better treatment will eventually be undertaken by qualified higher group symmetry electrodynamicists.

particle, imposing an additional "energy collecting" and gradient across the mass. Rigorously, the dynamic system comprised of the dK gradient coupled to the particle, and interacting with it, identically is a force.[102] Depending upon the situation, it may accelerate or decelerate the particle, or change its direction, or a combination. This is summarized in the usual work-energy formula

$$W = K_f - K_i = \Delta k \qquad [4\text{-}1]$$

where the work W that was done is given by the change dK in kinetic energy of the particle from its initial kinetic energy K_i to its final kinetic energy K_f.

In applying the conservation of energy law, usually physicists utilize either the work-energy theorem or an approach closely paralleling it. This application is valid in any linear single-pass collection situation. It may not be valid — and usually isn't —for a multi-pass, multicollection system where the same energy, changed in form to accomplish work in a single reaction, is then changed in form iteratively, again and again, in the same system or material. In such case, one joule of input energy can and will accomplish multiple joules of work in and on the system. So for our purposes we must modify the work-energy theorem to give

---

[102] We stress again that the hoary old mechanics, some four centuries old, is seriously in error in assuming a separate force acting on a separate mass. That remains an inexplicable non sequitur in modern physics, in its very foundations. We call attention to how classical electrodynamicists' wrestle with the problem. E.g., quoting Jackson, *Classical Electrodynamics,* 2nd edition, Wiley, 1975, p. 28: *"...the thing that eventually gets measured is a force... " "At the moment, the electric field can be defined as the force per unit charge acting at a given point. It is a vector function of position, denoted by E."* Our comment is that the "definition" is in error, including the use of the word "field" for "field intensity", when field and its intensity are two entirely different things. A proper definition should be stated as *"...the observable electric field intensity can be defined as the force per unit static charge and mass, existing at a given point in an interacting gradient of the virtual particle flux of the vacuum, as inferred from the changes induced upon the unit static charge and mass. "* The field intensity is a vector function of position, charged mass, and vacuum virtual particle flux gradient. The standard definition substitutes an effect (of collection) for a cause (prior to collection), and this error is widespread through physics, from mechanics to electrodynamics and on into particle physics. On p. 249, Jackson shows how the electrodynamicists largely capitulate and continue to promulgate the substitution of effect for cause. Quoting: *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."* We comment that actually mass is a component of force, as shown by $F = d/dt(mv)$.

$$W = \lambda(K_f - K_i) = \Delta k \qquad [4\text{-}2]$$

where λ is the averaged energy collection fraction, adjusted for multipass, multicollection and also for asymmetrical self-regauging (self-increase of the potential in, and hence the energy flow from, one or more regions of the active system's source). In formula 4-2, K_f and K_i and Δk still refer to single-pass collection only.

In the new work-energy theorem, $\lambda = 1.0$ for that vast number of conventional cases that have successfully utilized the conventional form of the theorem. These are almost always single-pass, single collection systems. The new COP>1.0 systems have deliberately enhanced energy collection — and hence multiple *energy collection and work amplification*[103] — such as multi-pass collection. For these COP>1.0 systems, the expanded work-energy theorem form applies and in general $\lambda > 1.0$. In an ideal COP>1.0 system, $\lambda >> 1.0$. In some versions of the Patterson Power Cell®, for example, $\lambda = 1200$. In the Sweet device (discussed later), a very high value of lambda resulted, so that $\lambda = 1.5 \times 10^6$. We sometimes refer to λ as the *energy amplification factor, energy collection fraction, Poynting amplification, collection amplification,* etc. Bohren and others have referred to it indirectly as the *interaction cross section,* the *resonance absorption cross section,* and the *negative absorption cross section.*

Again, we strongly iterate that *there is no conservation of work law in physics!* When energy is input to a system, acts on it, and is "dissipated" by a change of its form (doing work), the "dissipated" energy is still available after the work (change of the energy's form) is accomplished. However, usually by "dissipated" energy we mean energy that has permanently left the system (or a component in the system) and cannot be recovered and used by the system. The original input energy is not work, because it did not undergo any change of form prior to interaction with the system to change its form. So one does not do "work" as input, which is a non sequitur. Any work accomplished in the input section of a system is due to inefficiencies of the system,[104, 105] usually allowing extra change in form of some of the energy in non-useful processes called "input losses". As we stated, it is quite possible for a single joule of original input energy to perform many joules of work. During the subsequent billions of years following the formation of the observable universe and the energy in it, every joule of the primordial energy is still here. And every joule of it has been steadily performing joule after joule of work since then. Further, it has continually been recycled between 3-space and the time domain and back, and is continually being so recycled today.

---

[103] There is no "conservation of work" law in physics or thermodynamics. It is energy that must be conserved. A single change in form of K joules of energy gives an equal number of K joules of work done upon the interacting mass system. However, there remains an equal number of joules available, just in a different form. If that second K joules of energy is then changed in form again and not allowed to escape from the system, an additional K joules of work is performed in the system, without the operator having added any extra input energy. And so on, until the inefficiency of the system and some inevitable losses allow all the "recycling" energy to escape from the system. We strongly believe that ignition, quenching, and re-ignition phenomena such as in gamma ray bursters and x-ray bursters are due to exactly such processes where iterative change of form of each initial joule of energy occurs.

Suppose input energy to a system has been dissipated (changed its form) once, doing work (single change of form of energy) in the system. If the dissipated energy in a different form is then recovered or partially recovered — without escaping the system — to change its form yet again in the system and produce additional useful work, that system is

---

[104] Almost always due to a negative feedback mechanism such as Newton's third law, self-regauging symmetry, the closed current loop circuit, inertia as due to a self-regauging 4-symmetry, etc.

[105] E.g., a transformer would be an overunity system, because the current in its primary is separate from the current in its secondary, if these two currents were completely decoupled and independent. In that case, the transformer would simply propagate and "couple" or pass energy emitted from primary to the receiving secondary, without back-coupling. However, the closed current loop in the secondary, routing all spent electrons in the secondary back through the back mmf of the secondary coil, produces a back-field coupling interaction with the primary, and this "back mmf/back emf" coupling requires energy dissipation in the primary section to overcome. In short, the transformer is designed so that the secondary lights the primary to a draw and the designer loses. With some losses and inefficiencies in the device, that makes the transformer a COP<1.0 device (unless free external energy is received in the input). The interested reader might extract several hints and suggestions in this book, passim, of things that overunity researchers have tried in their search to break up the back-coupling of a transformer. The simplest way, of course, would be to place a true negative resistor shunt across the secondary, so that part of the secondary current does not return back through the primary.

considered to be an open system far from single-dissipation equilibrium with its external environment.[106] In that case, instead of the first dissipation simply "returning" the energy to the external environment, the energy is recovered by another change in form (which may simply be a change in direction in the internal environment). The thermodynamics of an open system far from equilibrium with its active environment applies, whether that "environment" is internal or external to the system. It is "external" to the subcomponent of the system forcing the "local change of form" of the energy.

4.2.3 Gedanken Experiment: Multicollection by Regenerative Feedback

After a COP>1.0 collecting process with subsequent proper discharge of the energy, more work output as "dissipation of the collected-energy" exists than the energy being input by the operator or experimenter. By regenerative, clamped, carefully controlled positive feedback of a portion of the output *positive* energy[107] to the system input, the system can be converted to a self-powered (self-asymmetrically-regauging) system that powers both itself and its loads. However, see Chapter 9 for special considerations of the Dirac sea hole current and excess negative energy that arises for COP»1.0 EM systems. Positive clamped feedback will fail dramatically for any COP»1.0 EM system, and for many COP>1.0 systems it will also fail unless conversion of negative energy — arriving at the input section — to positive energy is accomplished.

To ease visualization of the energy multicollection process, consider a gedanken experiment iteratively retroreflecting a steady small EM energy input. Particularly see Figure 4-2.

---

[106] The vacuum environment is also "internally" penetrating throughout any macroscopic system, to the finest level! So the notion of separated "external" and "internal" environments requires strong qualification in any system engaged in iterative, work-amplifying interaction with its local vacuum.

[107] The output may contain appreciable negative energy as well as positive energy, however. In that case, in the feedback loop the negative energy must be transduced into positive energy by regauging, else it will appear in the input section of the system as an extra system load there in the input itself. It will "eat" incoming electrons from the external power supply, so that the external power supply has to power this "extra load" in the input section as well as powering the normal system load presented. This reduces the system COP to COP<1.0.

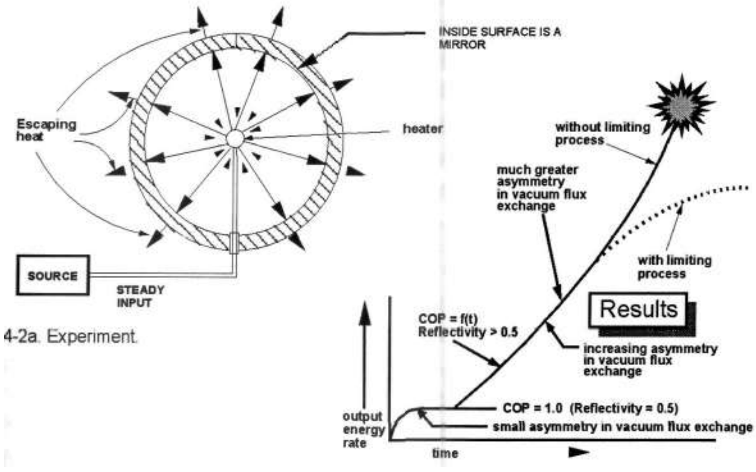
INSIDE SURFACE IS A MIRROR
Escaping heat
heater
SOURCE
STEADY INPUT
without limiting process
much greater asymmetry in vacuum flux exchange
with limiting process
Results
COP = f(t) Reflectivity > 0.5
increasing asymmetry in vacuum flux exchange
COP = 1.0 (Reflectivity = 0.5)
small asymmetry in vacuum flux exchange
output energy rate
time

4-2a. Experiment

4-2b Result

Figure 4-2 Heater wire in a hollow sphere with certain retroreflectivity properties.

As shown, a heater-emitter is placed inside a closed hollow sphere, through a tiny hole for the two wires connecting the heater wire into the external power supply (the external power supply is not shown). The heater emitter is 100% efficient; any joule of energy absorbed by it is re-radiated by it. One joule of energy (one watt of steady "input power" — conventional electrical engineering terminology where rate of work and rate of energy flow are confused!) is furnished from the power supply to the heater element inside the sphere. The inside of the sphere is coated with a substance that acts as a phase conjugate mirror (PCM) reflector with, say, 0.5 reflection coefficient. In other words, when scattered photons are emitted from the wire in all directions, they strike the inner surface of the sphere. Half are retroreflected precisely back to the spot on the resistor from whence they came, where they are absorbed and "re-collected" to further heat the resistor, causing emission of additional photons from it. The other half of the absorbed photons on the inner surface of the sphere diffuse through the cylinder walls as heat, and escape as heat radiation from the sphere to the outside world.

If that sphere's operation could be maintained intact and stable, without change of the physical characteristics and mechanism, then the buildup of energy in the heater wire inside the cavity — and the buildup of the energy icing emitted from the outside of the sphere — would be ever increasing. It would increase without bounds as time passed. Wait a short time, and the sphere would be outputting 10 watts of power while one would still be inputting one watt. Later, the sphere would be outputting 1,000 watts of power, while one would still be inputting one watt. And so on. Indeed, the rise in energy density of the output would be asymptotic, and would increase toward infinity. The only limitation as to the energy output for a steady 1 joule per second input, would be the limit at which the mechanism goes unstable and changes occur to dampen and curtail the process.

Before the reader objects on conservation of energy grounds, one should realize that this rise in the energy density inside the sphere is an example of iterative regauging. The principle of *gauge freedom* — one of the axioms of quantum field theory and well established — assures us that *the potential energy of a system can be changed at will, and freely.*[108] What the gauge freedom principle does *not* state is a *mechanism for providing the additional potential energy and a source for it* — nor do the quantum field theorists and electrodynamicists. Without such a source and mechanism specified, the conventional gauge freedom principle assumes total violation of the conservation of energy law.

---

[108] Conventional scientists are almost always very careful to regauge symmetrically, so that the two excess force fields that result are equal and opposite, and sum to a net zero resultant field. That "zero-summed system" comprises a stress potential *a priori.* So such symmetrical regauging alters the stress in the system and alters its potential energy, but only in the form of additional stress energy. There is no *net* force and "force field energy" that can be used to dissipate that additional free stress energy in an external load, thereby doing free work. This regauging is a purely arbitrary practice by the electrodynamicists. We are far more interested in "asymmetrical regauging", where only one potential is freely changed or the two are changed unequally, resulting in a *net nonzero force field* that can then be dissipated in an external load to perform free work. Only by asymmetrically regauging an otherwise inert system can usable field energy be added and any work then be done by it anyway. In conventional systems, we do the "asymmetrical regauging" by adding the voltage (potential difference). As current flows with dissipation of energy from the circuit in the loads and losses, the system is able to develop "power" and thus do work in the external load. Unfortunately, the electrodynamicists are still obsessed with symmetry, so they ubiquitously employ in all electrical power systems the closed current loop circuit. This guarantees that the circuit will kill its source dipole (the source of the potential and potential energy generated from the virtual flux of the vacuum by the broken symmetry of the opposite charges of the dipole) faster than it powers its load. That way, lovely symmetry is maintained, beautiful free energy and negentropy are avoided and tossed away, and ugly pollution and destruction of the biosphere continues worldwide to fulfill the ever-increasing and insatiable thirst for electrical power.

In our supersystem view, the gauge freedom mechanism providing the energy is straightforward: the excess energy is freely supplied from disequilibrium in the active vacuum interaction with the system and in the disequilibrium represented by the local curved spacetime. Since a change in system potential energy is also a free change in spacetime curvature, then it follows that simply changing the potential energy of the local vacuum (the vacuum is also an electromagnetic system!) and of the local curvatures of spacetime is also "for free", or it can be. One does not have to perform work oneself upon spacetime to curve it! The mere *presence* of a spatial energy change — including a free potential energy change allowed by gauge freedom — is sufficient to freely curve spacetime also. That is a real energy change, because any curvature of spacetime acts back upon mass. Higher group symmetry electrodynamics does indeed include vacuum energy and energy current {259a-259c}, and it is possible to extract useful EM energy from the vacuum.

But back to our sphere, where the internal potential energy of the sphere is rising asymptotically without bound.

In the real world, of course, the reflection coefficients and the materials characteristics will change as the energy density changes, the sphere will heat, etc., and these changes will start damping the perfection of the retroreflection process to limit it to some finite plateau value which may be a COP>1.0 or COP»1.0 condition. Or, the materials will melt or soften so that the sphere ruptures and explodes, sharply quenching the process entirely and emitting a violent burst of energy to signal the disruption of the localization of the process.

Nonetheless, a successful real bench experiment similar to this should be possible with some tinkering, and it should readily yield COP>1.0, after one waits for build-up and stabilization at some level. We would hope that a sharp young graduate student may eventually prepare a doctoral thesis on this experiment or a variant. The principle is demonstrable.

Similar buildups by regenerative feedback and multiple collections by the collectors do appear to occur in nature, up to and including such phenomena as gamma ray bursts, x-ray bursts, etc. These occur in some systems such as exploding gases (and in some cases, even in the upper atmosphere of the Earth). In such a system, the "physical particles in suspension in space" do move and continually disrupt the geometry for the increasing buildup, resulting in decay of the process after a time delay.

However, there is a finite time during which the movement of the particles is still insufficient to appreciably break the geometry and cause quenching.[109] During this "nearly linear stage", the build-up applies and the energy density very rapidly increases. Build-up occurs and continues while the relaxation time of the initial countering symmetry-restoring mechanism for the old or former state or condition (the initial reaction) is occurring. A well-known conventional example is the Lenz law reaction. Then as the geometry changes significantly, a rapid damping of the energy density increase occurs, quenching the asymptotic rise and "discharging" or "decaying" back to a less-than maximum condition. However, the decay position or state (in the quenched condition) will be different from the initial condition before asymptotic rise (regauging) set in. The system usually will have itself absorbed and collected additional potential energy, and so it will be "hotter" or more energetic. This is often called the "afterglow". In the simple case, parts of the system will be in greater motion (more energetic).

An intermediate plateau can appear as quenching occurs and it can stabilize. However, if no stable plateau is reached during the damping, then further rapid quenching and a rapid reduction of the increased energy density occurs as the geometry changes become decisive and break the asymmetric self-regauging. Thus the "damping" of the process may yield a plateau of steady COP»1.0 operation, or it may simply go immediately into full catastrophic quenching and decline back to some lesser level of COP>1.0 operation, but still more energetic than the beginning condition. The latter case produces a sort of "afterglow" - - observed in the gamma ray bursters, e.g. — of increased energy density from the beginning of the quenched phenomenon. Indeed, in that new system of more energetic gases or particles, once the new state is stable, continuing regenerative feedback versus the new geometry can then reinitiate or "re-ignite" another "burster" followed by subsequent quenching. And so on.

We submit that the gamma ray bursters fit this schema and are consistent with it. So do the phenomena observed in intensely scattering photoactive media on the nonlinear optical laboratory bench. We hypothesize that the gamma ray burster and similar violent burst emission phenomena are generated by this mechanism or a similar version of it.

---

[109] Regardless of what we call it or what it has been conventionally named, any impulsive, explosive process involves at least momentary broken symmetry, resulting in instant and countering broken symmetry. Then a relaxation time passes, and the countering broken symmetry is dissipated by change of the system to a new potential energy state, the new dynamics state, etc.

Figure 4-3 diagrammatically shows a proposed range of such "excess energy" emissions due to this regenerative positive feedback and multiple collection mechanism. As can be seen, this may place our view of very powerful astrophysical emission phenomena in a completely different light, arranged energetically by the length of the initial "nearly linear" phase of exponential increase. There is a scale of stability levels up to just before the Big Bang itself {260}. A so-called Big Bang, of course, would result when the "containment" ability of 4-space itself is breached, resulting in a rupture of 4-space and a consequent violent blow-out into n-space (where n>4), with a very rapid "false vacuum" created *outside* the blow-out region and pumping itself up by asymmetrical self-regauging in a "new" 4-space. When a new stability level is reached outside the blowout hole from the spawning 4-space, it represents an outside "new" 4-space and a new, infant 4-spatial expanding 4-universe, freshly born.

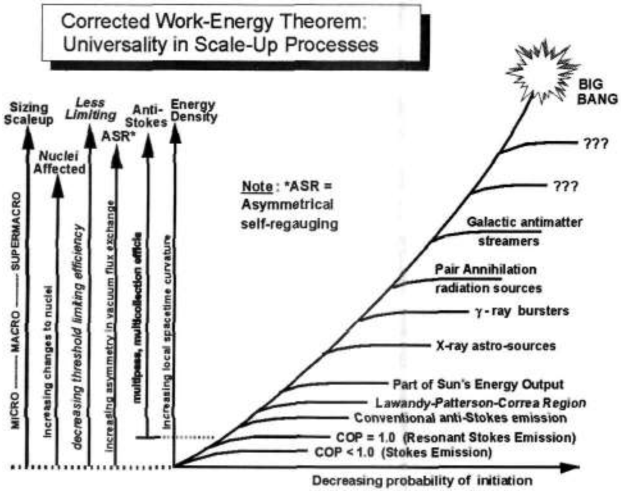
Corrected Work-Energy Theorem:
Universality in Scale-Up Processes
Sizing Scaleup
Less Limiting
Anti-Stokes
Energy Density
ASR*
Nuclei Affected
MICRO —— MACRO —— SUPERMACRO
Increasing changes to nuclei
decreasing threshold limiting efficiency
Increasing asymmetry in vacuum flux exchange
multipass, multicollection efficie
Increasing local spacetime curvature
Note: *ASR = Asymmetrical self-regauging
BIG BANG
???
???
Galactic antimatter streamers
Pair Annihilation radiation sources
γ - ray bursters
X-ray astro-sources
Part of Sun's Energy Output
Lawandy-Patterson-Correa Region
Conventional anti-Stokes emission
COP = 1.0 (Resonant Stokes Emission)
COP < 1.0 (Stokes Emission)
Decreasing probability of initiation

Figure 4-3 C0P>1.0 **stability levels for the asymmetrical self regauging process invoked by multipass and** multicollection **of same energy.**

So from inside the spawning universe, the burster phenomenon would be so great as to "burst" 4-space, producing a sort of "super black hole" as seen by the spawning universe. We hypothesize that the asymmetrical self-regauging mechanism produces a great burst of EM energy in its universe if a stability plateau is reached that is below the threshold of 4-space rupturing. If such a stability plateau is not reached, the asymptotic rise in local spatial energy density increases until the primary 4-space ruptures to produce a "blowout" seen from the old universe as a sort of black hole. There is a sort of "white hole" in the new universe being spawned on the other side of the blowout. This white hole and its associated phenomena produce the birth of another 4-universe outside the original spawning 4-universe. At least we propose that as a hypothesis.

We propose that this could be a legitimate process for the birth of multiple 4-space universes in the infinite-dimensional cosmos.[110] It has an added advantage of accommodating the vexing question of "What existed before the beginning of time (in this universe)??? The answer then would be, *"The existence of time in another universe that birthed this one. "* We emphasize our interpretation of time as just a special form of EM energy, and not at all a "mysterious river down which we float helplessly as if in a drifting boat in the river's current."

4.3 Iterative Retroreflection and Collection in a Circuit

The normal circuit attached to a generator or other external power supply is the "external" portion of the overall closed current loop circuit passing back through, and including, the source dipole in the power supply. In a DC external circuit, we may consider that the energy flow pours forth steadily from the terminals of the generator (from the dipolarity between them) into space surrounding the external circuit. Since the external conductors act as a sort of guide, this *transmitted* external energy flow is generally parallel to the external conductors.

The external circuit further acts as a *receiver.* The surface charges and their fields intercept and interact with some of the transmitted energy flowing parallel to the conductors and outside them. This interaction diverts a component of that passing energy flow into the conductors, powering the Drude electrons. This powers the external circuit. The nondiverted component of the passing energy flow — usually much larger — escapes on out into space beyond the external circuit. According to Heaviside's energy flow theory {261}, this escaping nondiverged energy flow may be interpreted to exist in closed circulations.[111] Laithwaite

---

[110] After all, a "dimension" ultimately is just a fundamental variable in a physics model or, in other words, a "degree of freedom" in mathematical modeling. We may construct any number of dimensions we desire, merely by assigning a sufficient number of variables, including both overt variables and hidden variables, etc.

[111] We point out the interesting possibility of re-interpreting this "closed circulation" Heaviside component as the giant negentropic circulation of EM energy from the

{172b} has specifically commented upon the potential importance of this aspect of Heaviside's theory.

As we shall later address, Heaviside worked out but never published a draft theory of gravitational effects from this extra escaping energy flow {262}. Interestingly, since Lorentz's integration of the energy flow vector around a closed surface surrounding any volume element of interest, the discarding of this "excess nondiverged energy flow" has been ignored by electrodynamicists, and it has been ignored by gravitational theorists.

Consider a theoretically "perfect" generator, loss free. Consider the presence of the Heaviside nondiverged component emerging from the generator terminals — and filling all space around the external circuit. That is the part of the energy outside the circuit that is not caught and processed by the circuit, but is just wasted. By definition, in this "perfect" situation the magnitude of the Poynting or "caught" component of energy flow will precisely equal the magnitude of the mechanical energy input flow (rate) to the generator.[112]

But the presence of the additional huge Heaviside component rigorously means that far more energy pours from the terminals of every battery and generator than the amount of mechanical energy the operator inputs as shaft horsepower. This easily demonstrated fact — e.g., proven by the Bohren experiment {24} — continues to meet with a resounding silence from electrodynamicists and physicists, and particularly from those dogmatic skeptics so keen on defending the classical thermodynamics of closed or equilibrium systems as if they also applied to open systems far from equilibrium in their exchange with their active environment. We refer particularly to the dogmatists so supremely confident that COP>1.0 EM systems are prohibited by the laws of thermodynamics. These stalwart fellows would thus exclude every charge (and its associated fields and potentials) in the universe, were they logical and speaking scientifically. So they strangle at the gnat of COP = 4, and swallow COP = oo while proclaiming that COP>1.0 EM systems cannot exist.

The dogmatists have achieved a major nefarious goal, however. They have silenced most inquiry into such matters by electrical engineering

---

time-domain to the negative charge, to 3-space, to the positive charge, and back to the time-domain. To our knowledge, that has not been examined.

[112] That is merely a statement of the Lorentz symmetry condition self-enforced by the closed current loop circuit itself.

professors, young graduate students, and young postdoctoral scientists. They have glorified classical equilibrium thermodynamics into a religion rather than the partial and imperfect model it really is. They are directly responsible for the continuing rape and pillaging of the planet's resources, pollution of the biosphere, global warming, and the increasing death of species. They are indirectly responsible for the deaths of more human beings than Hitler, Mao, and Stalin combined. Most of the blame for environmental pollution problems concerned with power and electricity is squarely at their feet. Sadly, the environmentalists seeking to save this planet and biosphere have not yet recognized the true nature of the scientific Taliban that has generated the environmental problem. In tolerating its own scientific "terrorism", the scientific community seriously aborts much of the scientific method.

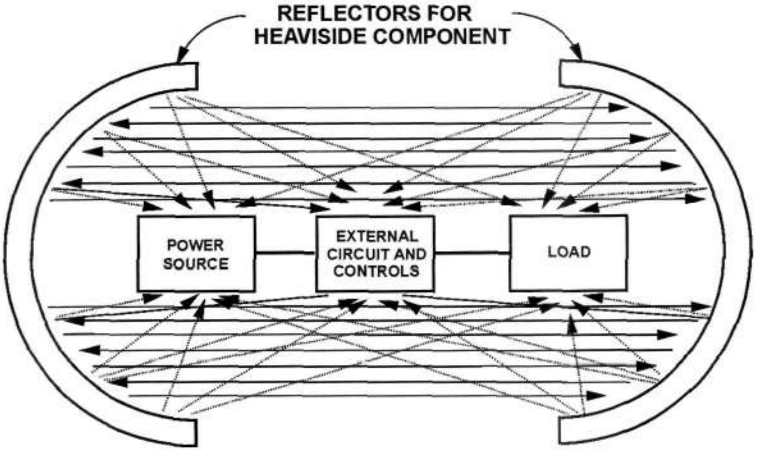

Figure 4-4 Iterative retroreflection of the "passed" **Heaviside component back and forth across a** circuit -- **thematic.**

See Figure 4-4. The neglected large nondiverged energy flow component need not be wasted. As an example, part of it can be captured and used, by applying near-field antenna techniques and such methods as Bohren's resonance increase of the intercepting charge's reaction cross section.[113] It is possible to intercept part of that excess energy that is there and available. E.g., even after it passes on by the immediate circuit, the Heaviside energy flow component can be intercepted beyond the circuit and retroreflected back across the same circuit again. In this way — at least in theory — extra Poynting energy can be collected and recollected by the same external circuit for a single given input energy.

---

[113] Or other methods we cannot yet divulge because of patent situation.

In essence, such an "iterative, multi-collecting circuit" is now fed with transmitted energy from multiple (iterative) paths simultaneously. This is *physically doable,*[114] but it is not a simple procedure, and antenna theory and especially near-field antenna theory must be utilized as well as focused retroreflection techniques. To directly model it, higher group symmetry electrodynamics (such as quaternion EM, O(3) EM, etc.) must be used. Precisely such phenomena are met in modern experiments in intensely scattering optical materials {263a-263b, 264a-264b} and in Bohren-type experiments where more energy is collected than the operator has input via standard calculations {257, 258}.

4.4 Tesla's Iterative Retroreflection Work

4.4.1 Tesla's Single Wire Circuit

See Figure 4-5, which diagrammatically shows one of Tesla's developments, which used a retroreflective single wire circuit between two large separated capacitances. This was the basic approach used in his magnifying transmitter and in several other circuits. Here the dipolarity exists between two accumulators, one at each end of a long conductive path with a load in the middle. AC operation is used, by setting up oscillation end-to-end along the "single wire" circuit, from accumulator to accumulator. Even so, no overunity operation will be experienced unless passive retroreflection of the normally wasted Heaviside component of each energy flow "passage" occurs back from each receiving end, so to speak {265}. That is possible if specialized dielectric material and construction is used in the end capacitors.

---

[114] E.g., some of Tesla's actual patented circuits do it, but that part of their fu**nctioning** is not described by any vector or tensor analysis. However, a higher **group** symmetry electrodynamics analysis of the circuits will indeed show the effect clearly as a "shuttling" of potential (regauging) around the circuit at will. E.g., see T. W. Barrett, {124, 286a, 286b}.

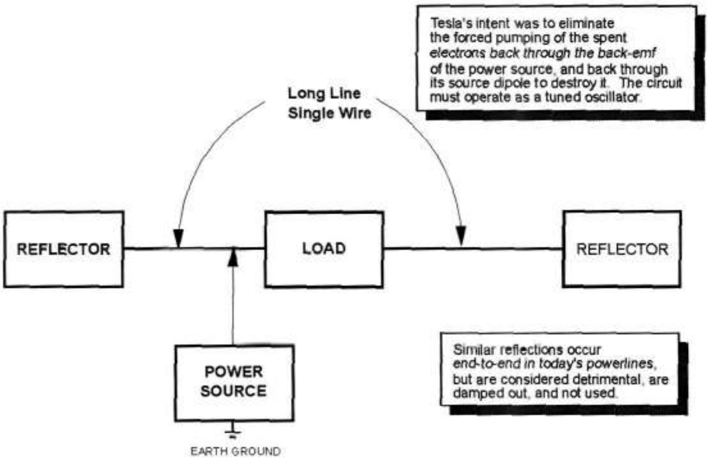
Tesla's intent was to eliminate the forced pumping of the spent electrons back through the back-emf of the power source, and back through its source dipole to destroy it. The circuit must operate as a tuned oscillator.
Long Line Single Wire
REFLECTOR
LOAD
REFLECTOR
POWER SOURCE
EARTH GROUND
Similar reflections occur end-to-end in today's powerlines, but are considered detrimental, are damped out, and not used.

Figure 4 5 Tesla's iteratively retroreflective single wire circuit.

Further, it may usually be necessary to place the individual charges comprising the "current surging to and fro" into particle resonance at either the frequency being transmitted or some subharmonic of it. This schema is doable, but it is difficult and most everything affects it, including even the impurities in the conductors.

4.4.2 Strain to Stress Potential Energy Transduction

See Figure 4-6. This is a diagrammatic representation of a fundamental "strain-to-stress" transduction mechanism we believe was involved in Tesla's "single wire circuit" (i.e., his iterative retroreflecting circuit).

At each end of the circuit, the electrons are "trapped" in the open outside end. Hence when EM energy rushes to the capacitor on one end of the circuit, the forward emf tries to strain the dielectric from the inner plate toward the outer plate. However, the trapped electrons in the outer open wire cannot move, so they produce an equal back-emf and back force trying to strain the dielectric from the outer plate to the inner plate. The result is that the translational field "strain" energy (dipolar charging) is transduced into stress potential (monopolar charging) energy. Electrons try to strain the dielectric from the inner plate toward the outside with a "forward emf so as to charge the capacitor normally, while the trapped external electrons simply push back the other way, exerting a back emf force that is equal and opposite to the straining force from the forward emf and simultaneously tries to "charge" the capacitor in the opposite direction.

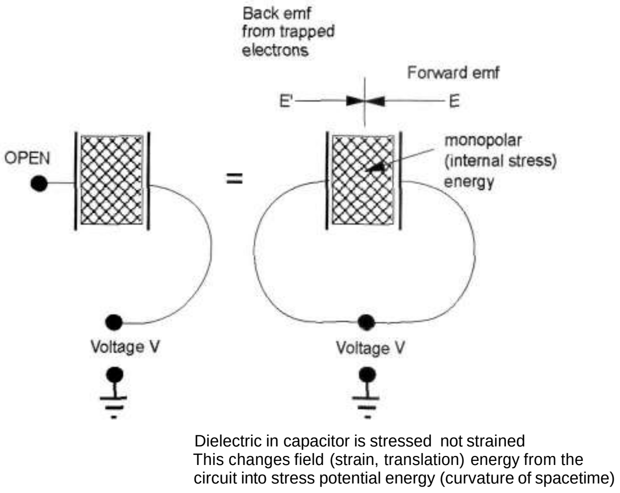
Back emf
from trapped
electrons
Forward emf
E'
E
OPEN
=
monopolar
(internal stress)
energy
Voltage V
Voltage V
Dielectric in capacitor is stressed not strained
This changes field (strain, translation) energy from the
circuit into stress potential energy (curvature of spacetime)

Figure 4-6 Transaction of translational field energy into stress potential energy.

Here is the importance of the "vector zero resultant" stress (monopolar) charging that occurs: The input EM strain field energy familiar in dipolar charging of a capacitor is converted to a dielectric stress (monopolar) potential for monopolar charging. This changes *translational* field energy (external energy) into *stress potential* energy (internal energy), adding an equal amount of back-emf translational field energy in so doing. In short, the charging captures an equal amount of charging energy from its external environment — in this case, from the back emf of the trapped external electrons in the outside open wire. This is one way to capture and use the back emf energy normally reducing the energy available.

We point out that this is charging the capacitor by *Lorentz's symmetrical regauging,* and it represents a rotation of the frame of the capacitor itself out of the laboratory frame. As is well known, the trapped energy of a capacitor in a rotated frame is not the same as the energy of the same capacitor in the nonrotated lab frame.

The single-wire system destroys the symmetry-enforcing function of a closed-current-loop circuit conventionally used. In short, it does not return the "spent electrons" from the ground return line back through the source dipole of the outside power source, so it does not use half the collected energy in the circuit to destroy the source dipole and its extraction of EM energy from the vacuum (to destroy its broken symmetry). Because it retains the broken symmetry, Tesla's single wire system is an open system far from thermodynamic equilibrium with its active vacuum environment. It is thus permitted to exhibit those five magic functions of such open dissipative systems: (1) self-ordering, (2) self oscillation or self-rotation, (3) outputting more energy than the operator inputs, (4) self-powering of itself and its load, and (5) exhibiting negentropy.

4.4.3 Potential Energy Shuttling in a Circuit

The net result is to "shuttle" the (stress) potential energy back and forth in the circuit, between the ends. See Figure 4-7. When this stress potential energy shuttling is done at the natural (or tuned) resonant frequency of that single wire circuit as an oscillator, the monopolar charge transfer is optimized and amplified by the resonance effect, with an increase in the energy being shuttled. It is this type of Tesla single wire circuit that can in theory exhibit COP>1.0.

Tesla's single wire system also forces the change of energy between internal energy and external energy. A deeper analysis will also involve Dirac sea hole current effects as well as lifting of electrons from the Dirac Sea and combination of unobservable Dirac sea holes and electrons without radiation. However, we leave that for the advanced theorists. A taste of it is included in Chapter 9.

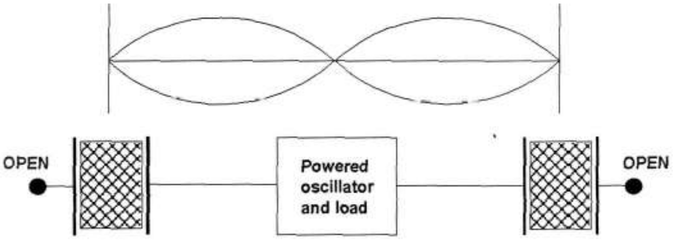

open capacitor to change field energy into stress potential energy

Note: System is designed with a certain resonant frequency At that frequency, the current can become mostly "squeeze" and "relax" as a standing wave emerges and stabilizes

open capacitor to change field energy into stress potential energy

Whether COP>1 0 or not depends on successful additive collection and use of the Heaviside nondiverged energy flow component due to the multipass and reflection operation

Figure 4-7 Tesla's resonant single wire circuit **operation.**

But is there any evidence for this "shuttling" of the monopolar potential in Tesla's circuits? Yes, there is, and indeed some of Tesla's patented circuits do exhibit such energy shuttling in them, as rigorously shown by Barrett {124, 286a, 286b}.

4.4.4 Other Considerations

For COP»1.0 systems and some COP>1.0 systems, the concepts of Chapter 9 must also be applied, since the system outputs not only positive energy but also negative energy. The negative energy — in the form of Dirac sea hole current — then sweeps back through the system from its output section to its input section. At various places, these "vacuum holes" eat electrons (as on atoms, causing ionization).[115] In short, the Dirac sea holes are converted to more normal lattice holes, and thus can be dealt with in the same fashion as semiconductor designers handle lattice hole problems inside their semiconductor designs. Positron (hole) effects are well known in materials science, and are increasingly being used in actual systems such as the antimatter (positron beam) microscope {266a-266b}.

We accent that a simple LC oscillating circuit does not produce COP>1.0 of itself. If, however, the charges in the currents surging to and fro are in particle resonance at the "feed frequency", then they will "sweep out" a greater geometrical cross sectional area to intercept additional Heaviside energy flow existing outside the normal "static charge" Poynting energy flow. This is similar to the "intercepting resonant particles " process utilized by Bohren {257}, and it allows output of more EM energy than the operator inputs *in the calculated Poynting component* of his own input energy flow. It *does not* output more EM energy than the operator inputs in the sum of his Poynting input component and his Heaviside input component.

The real energy conservation law is that the input energy from all sources and of all types must equal the output energy of all types, where by

---

[115] The union of a Dirac hole (4-space negative energy electron before observation) and **a** positive energy electron before observation, can occur without radiation. The two curvatures of spacetime involved are equal and opposite, so they just get **replaced** by a flat spacetime, which involves a filled Dirac sea (all the holes filled with electrons). In short, the energy can go into changing the stationary local double re**curvature** of local spacetime, by a sort of "superposition of two spacetime curvatures", to give a net flat spacetime, rather than dynamically radiating EM energy away at the speed of light. However, if either of the two interacting particles has been observed, this is not the case and standard pair annihilation will be observed. The point is that particles in their unobserved *mt* state rather than their observed *m* state, can interact without observation, similar to the way that wave-to-wave interaction occurs in nonlinear optics.

"output" we mean output as useful energy and output as loss or non-useful energy.

So conservation of energy flow is not violated, because one also inputs an additional *Heaviside nondiverged* component, unknown to Bohren. Circa the 1880s or early 1890s, Lorentz arbitrarily discarded it from accountability. One does not have to "pay" for that *nondiverged* and *unaccounted* input energy flow component. Except in continually applying Lorentz's integration trick to discard it, the nondiverged Heaviside component has been almost totally neglected in the textbooks since Lorentz arbitrarily discarded it back there circa the late 1880s. However, it is roughly dealt with in near field antenna theory, sometimes under the topic of "Poynting energy bunching". In the near field, e.g., there also may be a substantial *longitudinal* E-field component. One region may be oscillating back and forth along the flow direction, while the region in quadrature may be flowing in only one direction.

By causing the intercepting charges themselves to resonate and sweep out a greater reaction cross section in an energy flow, the normal definitions of "field intensity" and "potential intensity" are no longer true, because their reaction intensity to a unit point *static* charge has been used as their stated "magnitude". The conventional definition is only for an indication of the *intensity* of the energy flows comprising the field and potential. With the unit point *resonant* intercepting charge rather than the assumed *static* charge, the reaction cross section for the same charge increases, and so the "apparent magnitude of the (intercepted) field intensity" also increases beyond what is in the textbook. So far as I am aware, there is no previous precise term associated with this phenomenon, except to lump it in the euphemism of "negative absorption of the medium" or "negative resonance absorption of the medium" in certain materials {267}. It is more nearly a Poynting energy amplification process, since the "intercepted" or Poynting component of the energy flow is increased. Thus it can be a mechanism for COP>1.0, if the closed current loop circuit's use of half the collected energy in the circuit to kill the source dipole is also defeated.

Nonetheless, energy bunching or Poynting component amplification by altering the reaction cross section function of the intercepting particle is a legitimate way to provide excess "free" energy input for COP>1.0 EM power systems.

4.5 Perspective on Tesla vs. Marconi

Tesla was neither deluded nor mistaken, though some electrodynamicists have mistakenly considered him in that light. On the other hand, Jackson — one of the leading electrodynamicists of our time — pays Tesla a significant compliment as follows {268}.

> *"In U.S. patent no. 787,412 (April 18, 1905), reprinted in Nikola Tesla, Lectures and Patents and Articles. Nikola Tesla Museum, Beograd, Yugoslavia (1956), this remarkable genius clearly outlines the idea of the earth as a resonating circuit (he did not know of the ionosphere), estimates the lowest resonant frequency as 6 Hz (close to the 6.6 Hz for a perfectly conducting sphere), and describes generation and detection of these low-frequency waves. I thank V. L. Fitch for this fascinating piece of history. "*

We also point out that many present science texts err by continuing to credit Marconi with the *discovery* of radio, although he certainly did much to pioneer it and received a Nobel Prize in 1909. Hertz must be given credit for the first deliberate signal transmission per se, during the period 1886-1888. As his transmitting antenna, Hertz used an end-loaded dipole with a gap, and as his receiver he used a resonant square loop antenna with a gap {269}. Sparks produced at the gap in his transmission antenna resulted in sparks appearing at the gap in his receiving antenna.

Marconi became obsessed with Hertz's results nearly a decade later, and in mid-December 1901 Marconi announced the reception of signals at St. John's, Newfound that were transmitted across the Atlantic from Cornwall, England. He later recorded radio signals transmitted around the Earth, and began regular trans-Atlantic message service in 1903.

Marconi's patent {270}, however, was eventually overturned by U.S. Supreme Court decision {271}, because Tesla's patents (and demonstrations) substantially preceded it. Tesla also demonstrated radio transmission before Marconi, and Marconi actually used a modified Tesla circuit in his famed transmission in 1901. Nonetheless, this is not to detract from Marconi's accomplishments; he definitely put "wireless communications" on the map. It is simply to give *discoverer of radio* credit where credit is due.

Tesla was also closer to the truth of the form of an EM wave in space than are modern electrical engineering and most electrodynamics texts. In the present book we have stated that what exists in 3-space as the so-called "transverse EM wave" are actually longitudinal EM waves and functions of them, with the energy entering each 3-space point from the accompanying time-polarized (conjugate) EM wave at that point (i.e., from the time domain) and emerging in 3-space in the form of a longitudinal EM wave, in 1:1 correspondence.[116] This is followed by a return of the energy from 3-space back to the time domain. This "circulation" from the time domain to 3-space and back to the time domain occurs at every point dipole in 3-space. Only the presence and combination of both the time-polarized photon and the longitudinal photon at each point in an EM wave makes the wave "observable", when we translate into wave phenomena the photon phenomena pointed out by Mandl and Shaw {19}. What spreads as EM energy flow in space (from a source charge) is the local polarization of the vacuum, associated with this giant negentropy circulation from time to the negative charge, to the positive charge, and back to time.

The notion of the transverse EM wave flowing through space is an artifact of the ubiquitous *substitution of the effect for the cause* in Maxwellian electrodynamics. E.g., a completely erroneous "definition" of the causal field (as it exists in space prior to interaction with charged mass) is given as the *effect of its interaction with a ubiquitously assumed unit point charge at every point in space where the field is present.* This in fact substitutes the effect (existing *after* the interaction) for the cause existing *prior to* the interaction. That of course is a gross non sequitur.

We give a concrete example from the first edition of a standard dictionary of physics and mathematics {272}:

> *"**electric field...** 1. one of the fundamental fields in nature, causing a charged body to be attracted to or repelled by other charged bodies; associated with an electromagnetic wave or a changing magnetic field. 2. Specifically, the electric force per unit test charge."*

---

[116] Later we shall deal a little further with how EM energy iteratively recurs statically in 3-space and propagates dynamically and continuously in 4-space, in more detail. The solution to the long-vexing particle-wave duality is the process of observation. Before observation, an entity is dynamic and continuous in spacetime, not 3-space. After observation (by emission of a photon), it is a momentary frozen 3-snapshot (particle) at a single instant of time. At the next immediate instant, a photon is absorbed, reconverting the "static snapshot" into a dynamic 4-entity again. Since observation (photon emission) and de-observation (photon absorption) iterates at extremely high rate, the entity and each tiniest part of it oscillate back and forth between being a frozen 3-entity (particle) and a dynamic 4-entity (wave).

In the first definition, the field is given as the "cause" of the interaction that exists between like and unlike charges. In the second definition, the field is given as the effect — of the interaction with a test charge — of *something undefined* as the cause.

An entirely different "definition" is given by Gray and Isaacs {273}:

> *"**electric field** ... The space surrounding an electric charge within which it is capable of exerting a perceptible force on another electric charge."*

Here it is defined as simply the space around an electric charge, and not really the cause or effect of anything, unless that local spacetime has been curved (which it has).

Not to belabor the point, this sort of thing led Feynman to "define" the EM field in this fashion {274}:

> *"...the existence of the positive charge, in some sense, distorts, or creates **a "condition"** in space, so that when we put the negative charge in, it feels a force. This potentiality for producing a force is called an electric field. "*

So Feynman made the electric field a *condition in space* surrounding a charge, where the condition can produce (has a potentiality for producing) a force on another charge. He also was aware that force is an effect of the reaction of the causal field with charged mass, in the following statement {275}:

> *"...in dealing with force the tacit assumption is always made that the force is equal to zero unless some physical body is present... One of the most important characteristics of force is that it has a material origin..."*

In Maxwell's theory, the electron, atom, and nucleus had not yet been discovered, and electricity was considered a thin material fluid, flowing in a wire (hence the name "currents") much like water through a pipe. The surrounding space was considered filled with a thin material fluid called the luminiferous ether. A very rough notion existed that the "shaking" of the ether disturbance coming in to an antenna would disturb or perturb the electric fluid in the wire directly. Hence, if one "measured" a transverse oscillation of the electric fluid in the wire due to this "detection", it conclusively "proved" that the incoming ether-perturbations were also transverse. On that notion — together with Faraday's notion that "field lines" were taut physical strings, so that "field perturbations" were naught but plucked oscillations of these taut strings — the entire notion of the transverse EM force field wave in space was created.

It really is not that way, as is easily shown with modern re-examination of what happens in the receiving wire, considering the Drude electron gas {276} and electron drift velocity. Today we know that the electrons move *longitudinally* down the wire only with a small drift velocity (usually a few inches per hour), while the signal (field) races longitudinally down the wire at essentially light speed or nearly so.

The electrons are longitudinally constrained in the wire because of the vast number of electrons ahead of them that tend to repel their longitudinal movement. The electrons are far less constrained laterally in the wire. However, each electron is also spinning, and its 3-space performance may be compared to a spinning gyro, whose spin axes are now longitudinal restrained. Accordingly, the electrons readily *precess laterally* in the wire[117] when perturbed *longitudinally* by the incoming longitudinal 3-space components of the "wave in vacuum" — by gyroscope precession theory. The electrons at the surface layer of the wire are somewhat less constrained longitudinally, and they "slip" a little down the wire in each oscillation, accounting for the very slow electron longitudinal drift velocity.

Accordingly, one of the most important dynamics of the circuit is the dynamics of the surface charges, as pointed out by electrodynamicists {277}.

If the incoming waves were true lateral waves, there would be no longitudinal force component on the laterally precessing electrons, and coherent current would not flow longitudinally in the wire at all. Instead, one would just get some longitudinal to and fro noise in the coherent lateral motions of the electrons, due to electron collisions.

---

[117] It is this lateral precession movement of the electrons that is "detected" in most of our instruments. To the early electrodynamicists prior to the discovery of the electron, this detected lateral motion was considered the "shaking of the material electric fluid" in the detecting wire. It was thought that the incoming "electric fluid" from space also had to be material (in the material ether) and transverse, because otherwise the oscillations from space could not have mechanically transferred their "lateral oscillation" to the electric fluid in the wire. Faraday's "material lines of force" concept and the notion that EM disturbances (waves) were simply the mechanical pluckings of these taut strings which also involved similar "material ether" concepts.

Nonetheless, it seems we are stuck with the ubiquitous transverse EM wave model. For more than 100 years, scientists have universally substituted effect for cause in the case of much of electrodynamics — as in the continuing substitution of what is *diverged from* the field and potential at each point in them, by an assumed static point charge, as being the magnitude of the field or potential itself at that point. At best, what is diverged from the field or potential is an indication of the point intensity of the energy flows comprising the field or potential at each point, and in the case of the field, of the prevailing direction.

So electrodynamics is still riddled with that nearly all-pervasive error, and it seems it will likely never be corrected. The leaders of the scientific community do not seem to wish it done and will not allow it to be funded. In short, the community is so comfortable with that grave error that it will almost defend it to the death. The so-called transverse EM wave in 3-space has come to be accepted as if it were a law of nature.

Occasionally a journal will point out the terrible non sequitur in the prevailing notion of the EM wave in space as a "plane wave of oscillating orthogonal E and H fields along the x- and y- axes moving at light speed in the z- axis direction." Romer, former editor of *American Journal of Physics,* provides a prime example {278}. Romer takes to task:

> *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "...perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

At any rate, once we understand how the real EM wave exists in space, then we can sympathize with Tesla's view of it from the following {279}:

> *"Tesla upholds the startling theory formulated by him long ago, that the radio transmitters as now used, do not emit Hertz waves, as commonly believed, but waves of sound... He says that a Hertz wave can only be possible in a solid ether, but he has demonstrated already in 1897 that the ether is a gas, which can only transmit waves of sound; that is such as are propagated by alternate compressions and rarefactions of the medium in which transverse waves are absolutely impossible."*

In Tesla's own words {280}:

> *"...I showed that the universal medium is a gaseous body in which only longitudinal pulses can be propagated, involving alternating compressions and expansions similar to those produced by sound waves in the air. Thus, a wireless transmitter does not emit Hertz waves which are a myth, but sound waves in the ether, behaving in every respect like those in the air, except that, owing to the great elastic force and extremely small density of the medium, their speed is that of light."*

From quantum field theory, the instantaneous scalar potential in space is actually a coupling of a scalar photon and a longitudinal photon {19}. The only EM component in 3-space is that longitudinal photon.[118] Transforming to wave language, Tesla appears to have been correct in his view of the longitudinal form of the EM wave in 3-space, whereas all the texts are still wrong even today.[119] However, Tesla was also noted for grandiose pronouncements, a trait which certainly did not endear him to the struggling electrodynamicists of the time, even when he was correct and they were wrong. An example is the following {281}:

> *"The Hertz wave theory of wireless transmission may be kept up for a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history."*

Unfortunately, in electrodynamics we are still stuck with that "most remarkable and inexplicable aberration" today, and we will remain stuck with it until the leaders of the scientific community (such as the National Academy of Sciences and National Science Foundation) recognize it and fund a complete overhaul of the foundations of classical electrodynamics, which is so sorely needed. On the other hand, *one should not hold one's breath while waiting for the mountain to come to Muhammad! One might as well ask the stars to fall from the sky as to ask the leaders of the scientific community to correct the serious foundations errors in physics and especially in electrodynamics.*

---

[118] Hence it is not surprising that all EM in 3-space decomposes into bundles of longitudinal EM waves and their dynamics, as shown by Whittaker {85, 91b}.

[119] However, Tesla did not appear to realize the involvement of time-energy transduction to longitudinal KM 3-space wave energy and vice versa.

4.6 Conclusion

Hopefully we have set the stage for our next chapter, where we will be advancing mechanisms and approaches that have been tried by a variety of researchers. Several of these researchers have indeed produced COP>1.0 devices and processes. However, at this writing none of the devices produced by these inventors and researchers are in production and on the market. As we mentioned in Chapters 1 and 2, there are very powerful forces that have been applied to prevent such development.

Briefly, such suppression has resulted due to one or more of the following:

(a) Sheer greed by the group that gathers around the inventor/researcher, including multiple competing investors and/or "backers". Often a struggle for control develops, and the project is decimated in the process. In many cases a legal struggle also develops, with the result that the legal status of the invention may be fouled beyond repair.

(b) Naivete on the part of the inventor and his surrounding group. The inventor needs the frequent advice of a patent attorney, a regular corporate attorney, a person familiar with research and development projects and procedures, a person expert in testing (particularly with sophisticated multichannel digital oscilloscopes which gather multiple channels of data simultaneously and integrate under the curve, an accountant, and a person skilled in negotiation with potential investors and backers. Most inventors do not have access to all that, or only have access to it after much of the die has been cast and serious mistakes have been made.

(c) Animosity and fierce attack from the scientific community and self-appointed "defenders of the faith". Few inventors are prepared for the onslaught that awaits them if they succeed in producing a genuine COP>1.0 electrical system. Slander and libel are just part of it. An inventor or researcher caught up in always trying to defend himself against vicious ad hominem attack is an inventor successfully deviated from his work. The family also often suffers.

(d) Just now, after the Enron fiasco much of America is becoming aware that powerful organizations often pull out all the stops in a blind pursuit for money and power. Of course they have done so for a very long time; this is nothing new. The inexperienced inventor or researcher is ill prepared for the fact that portions of government organizations,[120] national labs, universities, etc. may be precisely that way also! U.S. government laboratories also file patents; check it out! Check out Fullerton's defense of his patent rights for his ultrawideband communication system work.

The people in any large organization form a bell-shaped distribution curve. About 9% of that curve represents folks who have — to put it mildly — little ethics when it comes to position, power, and money. Some of these 9% fellows will go to quite some length to twist the patent or part of it or control of it away from an inventor. An inventor who has never heard of "march-in" rights (a neat little "theft" phrase favored by DARPA, for example) is often an easy mark. Government agencies will be on an interesting invention and its hapless naive inventor like vultures on a fat calf. They dangle some money and a contract to "assist the inventor to get his invention developed". There is a little clause in the fine print of the contract where the government reserves "march-in" rights. This is a hidden bombshell that really says the government — *anytime it wishes* — can just unilaterally declare that the funded inventor isn't getting the invention to market fast enough, so a single bureaucrat just *seizes it* — in the name of the government, of course. Anyone who has been in serious aerospace contracting very long is also aware of the deals between a few government fellows who control award of the funds, and their "favored contractor". If enough business is steered to the contractor by the government person, it is implicitly understood that when the government fellow retires, he goes to work for that company with a Vice President position, lucrative salary, and a fat stock option. It's easy to become an instant millionaire or multimillionaire that way. A small percentage of government contracts are actually contaminated in such a hidden fashion. The universities usually are much more open and blatant, and just put in a clause that says, "All patents revert to the university!" Also, the inventor must be aware of the difference between "nondisclosure" and "non-circumvention". Just try getting the latter clause from a university, or either clause from a government agency (yes there are rules for "proprietary information", and they are widely violated). Non-circumvention means the outfit agrees not to change a coil or something, and go and patent your invention. Nondisclosure means they just agree not to talk or write about it openly.

---

[120] Interestingly, the government is not allowed to copyright its own publications by federal servants. Neither should the government be allowed to patent its own inventions by federal servants and federal agencies. This has set up direct conflict in the federal funding channels for the taxpayers' funds intended for independent research, providing a wonderful opportunity for hanky panky by that lower 9% of the people comprising any large group. By the bell-shaped distribution curve, those comprising that 9% will rise to the opportunity. They are manipulative, greedy, and some are immoral and self-serving. Partially coupled with a similar low "bad apple percentage" in the contracting game, this enables many travesties in the name of federal research and development with respect to patents and misdirection of funds.

(e) And yes, in addition to these many "petty" little conspiracies, there really is a *great* conspiracy — in fact, there are many of them. Churchill just referred to the entire set of conspiring cartels as the "High Cabal". This is a loose-knit set of extraordinarily wealthy cartels that between them control much of the financial power of the world. You can only find the principals behind a couple of hundred interlocking corporations, if you penetrate deeply enough. Much of their great financial income comes from the control of energy at its various levels, through a net of interlocking corporations. Something like $2 trillion dollars per year is their "take" from energy worldwide. Oil plays a great role in that. These "high cabal" fellows make or break nations financially, as is almost being done to Argentina as this is being written. They also have been stopping COP>1.0 electrical power systems since shortly after the turn of the century, using whatever means are necessary — fair or foul. That is why we still are burning oil, coal, and natural gas, and why our automobiles and trucks run off gasoline and diesel, and not off free electrical energy from the vacuum. The high cabal intends to keep it that way. Some of the major primary means of "stopping" an invention have been as follows:

a. Manipulation and improper use of the law (Gray's production line was shut down by law officers before his engines could roll off there and onto the market).

b. Involving the unsuspecting inventor in drug charges by planting drugs on his property surreptitiously, then secretly notifying the drug enforcement officials who suddenly raid the premises and discover the narcotics, resulting in the inventor going to jail since he has been caught "red-handed".

c. Threatening the inventor's life and the lives of his family, to frighten him away from what he is doing.

d. Arranging a fatal accident for the inventor, such as a fatal auto accident, a fall from a high place, or drowning.

e. Assassination (hit men are fairly simple to hire; a good one is a bit more difficult to hire but is doable. A good hit man might assume a disguise and fake a robbery, shooting and killing the target in the "robbery". In an extreme case, he just fires at a distance with a professional sniper rifle that is silenced. A variation is the use of a "throwaway" assassin, a fellow who has been deeply conditioned by very harsh methods, to deeply involve his autonomic nervous system. He can be spotted by the "glaring eyes" effect of autonomic nervous system involvement. This fellow is only marginally functional in real life, and requires a handler. He also must "fixate" for some minutes (usually from 20 to 30) on the target, by staring at him for a period. Then — say, the next day — while the target is at lunch in a public restaurant, the throwaway walks in, walks straight to the target, and empties a pistol in him at point blank range. He remains there until the police arrive and arrest him. He is immune even to torture; nothing can be gotten out of him, even by electrical shocks on his genitals. There are two levels of throwaways; that is the older and less functional version produced by very harsh means. The more modern version is without the glaring eyes with white showing above the iris; this fellow is more functional, fixates very quickly (couple minutes to five minutes), and can operate without the close immediate presence of the handler.

f. Making the inventor an offer he "cannot refuse", in the old Mafia sense. In the latter, nicely dressed persons come to the inventor's house, meet with him, and offer him a large sum of money to quit what he is doing for the rest of his life. The going sum varies between $5 million and $10 million, usually. The offer is real and it cannot be refused; either the inventor accepts it, or he and all his family are dead when these gentlemen leave the premises. If the inventor refuses anyway, he and his family are killed.

Usually the bodies also just "disappear" and are never found or heard from again.

g. These days, variations are used such as a careful crash of another auto into the back of the inventor's car so he is really shaken up badly but not seriously hurt. He goes to the hospital for checkup, in case of concussion etc. Everything is seemingly okay, but he is kept overnight for observation. That night he is given a hypodermic of air in his veins. Next morning his family receives a call that the inventor has suffered seizures from an apparent concussion after all, and is fighting for his life. If he doesn't die, he becomes a human vegetable. Piece of cake.

h. Then there's the ice dart dipped in curare. The curare paralyzes the muscles of the heart and lungs; the inventor goes down in convulsive floundering, gasping and unable to breathe or to pump blood. He expires very quickly. The curare breaks down in the body, the ice melts, and you have a fellow the medics find died of convulsive heart seizure, etc. Works like a charm. This is an adapted "dirty spooks" trick, adapted from those branches of the intelligence communities of the world that specialize in assassinations.

i. Use of "gaming", which was originally developed to influence or change the behavior of foreign leaders. Imagine writing a scenario, as for a movie. Only you have the deep psychological profile of the targeted individual. You also have the deep psychological profiles of quite a few other individuals who have "knee-jerk" reactions in a certain direction useful to you in your games (in your scenarios). These are called "agents of influence" since, to get them into interacting with the target, only a phone call or other contact need be made and a stimulus given, and they are off and running, self-powered and going after that target with their knee-jerk response. So very sophisticated computers are used, the scenario (intended plot of events to happen) is prepared (the artificial intelligence programs will even give the probabilities of success of all the variants), and a recommended scenario variant is derived and accepted. Then the controller has one or more agents working **for** him that contact the players (the agents of influence). They get their phone calls or stimuli, react in predictable fashion, and come pell-mell at the inventor, interacting all over the board with him. The range of games is enormous, from stock schemes to lawsuits to infiltrating agents of influence into the internal organization of the inventor to money laundering, etc. The controller follows the progress, and only once in awhile makes slight adjustments (may need to trigger other agents of influence, send in one actual agent, etc.). Over the last 20 years, the gaming has become very expert. So many games can be launched against a single inventor that he usually is finished or nullified by one of them, eventually. Having been gamed over 200 times in the last two decades, I plan to write a book one day on the subject of the games and what to watch for.

j. Finally, here is a sad and disconcerting rule, usually true. If you gather six persons together that you have known for years or even all your life, and they are the nucleus of the organization around you, watch what happens when success occurs and you really do pull EM energy from the vacuum and power loads, lights, etc. One-sixth of the population, seeing that, will shortly contract a very strange malady known as "gold fever". It is not just an *obsession;* instead it becomes an all-consuming *possession* almost in the biblical sense. An individual possessed by gold fever would kill his own grandmother to try to gain control of that system. If one or more of these "gold fever" transformations occur, that sets the group into profound conflict and vicious infighting, on every front. You get sabotage of equipment, sabotage of demonstrations, stealing of equipment, legal shenanigans, just name it. This has destroyed many groups. It will destroy many more.

k. A variation is to send in a special agent who penetrates the group and becomes a member. Indeed, the high cabal has certain persons who specialize in just such penetration and are specialists at setting up disruption. They have had plenty of practice, and are very good at their trade. The identities of some of these "special agents" would surprise the casual observer. The so-called "overunity community" is rather strongly penetrated and misdirected by a number of such special agents, in addition to a greater number of agents of influence.

1. Finally, the technical knowledge of most inventors and researchers in this field has been largely limited to ordinary electrical engineering and electronics. This means that the inventor often forms his own lexicon and way of thinking about how he does the "magic" COP>1.0. It may be difficult or impossible to connect the way he thinks back to legitimate physics. Further, most of the COP>1.0 community, the majority of electrical engineers, and nearly half the physicists do not understand the technical difference between the *efficiency* of a power system and its *coefficient of performance.* We have insured that the reader of this book will know the precise difference — it's even in our glossary! Further, the entire group around the inventor, and the inventor himself, may not have a clear understanding of the source of the extra energy, the mechanism by which it is obtained, or even the difference between a system in equilibrium with its active vacuum environment and one in disequilibrium with it. Most engineers and a great number of scientists also are still unaware of the dramatic difference between the thermodynamics of equilibrium systems and the thermodynamics of disequilibrium systems. Many have no notion of what the active vacuum is or does. So the inventor may actually be able to build the device himself, but may later err and actually build it away from COP>1.0 with further development.

m. Another problem can be measurement. Sometimes very simple errors — such as "measuring" highly nonlinear, spiky signals with an RMS meter, or using an inferior oscilloscope, or using inferior probes with good oscilloscopes, tricks a relatively inexperienced inventor himself into believing his unit achieves COP>1.0 when it doesn't. Also, some standard engineers will insist on calorimetry, which is inappropriate for COP>1.0 EM systems since they output a mix of both positive and negative energy. The positive energy will heat the calorimeter fluid and the negative energy will cool it simultaneously, and the "expert measurer" will always measure the difference and thus COP<1.0 with his calorimeter (unless the inventor is a rare one who has also transduced the negative energy fraction of his output into positive energy). Yet the conventional engineer or scientist — who has never seen negative EM energy or a COP>1.0 EM system in his life — will be so confident from his "positive energy only" thinking that he will be completely convinced that a real COP> 1.0 system, measured with a calorimeter to show net positive energy COP<1.0, is *actually* a COP<1.0 system instead of the error in measurement that it is. One of the purposes of this book is to try to explain such things and get a better procedure and protocol made available for the testing.

n. Finally, there are a few actual charlatans who deliberately deceive both the group around them (if it's sufficiently naive) and naive investors and backers. E.g., if one builds a huge starter motor (very high current, very high torque, cannot run long without burning up!) and puts it in an automobile, then fills the back of the automobile full of batteries, that beast will draw 400 amps and leap across the pavement or warehouse floor with an agility one is shocked to see. Then if the inventor is always "having difficulty with his controller unit", and "everything will be fine once the controller problem is licked" or some such, many small and naive investors will leap to invest their life savings eagerly. There are many other scams also, and unfortunately every experimental field, with great financial promise if successful, does attract them. Here the old adage is best: If it is just too good to be true, it probably isn't. One should insist on a technical explanation of the source of the excess energy, the mechanism by which it is taken from the environment, how asymmetrical self-regauging of the circuit occurs, etc. If these questions cannot be answered, the inventor himself does not know what he is doing. At best, he cannot explain it; or he may just be honestly mistaken. In the worst case, the inventor knows it is not genuine, but is deliberately deceiving the investors. So my advice would be to first assess the scientific level of the understanding of the inventor group. Then get an independent test of the device or see one done properly. Then decide whether what you have seen and heard is real or not.

o. Meanwhile, there are from time to time legitimate COP>1.0 systems initially developed. The place where all of them have failed is where the inventor group seeks to obtain the appreciable financing (something like 30 million dollars is required these days) to complete the phenomenology experiments, modeling, scale-up factor determination, and prototyping before a production model can be built. To date, every one of the legitimate COP>1.0 systems has foundered and failed right at that sheer vertical cliff. Our own MEG system is at that cliff at the moment, so that cliff is very near and dear to the hearts of the five of us who have struggled so long and invented the MEG. Time will tell whether we succeed where all the others have failed, or whether we also just get defeated by the cliff. As of this writing, we have not yet obtained the necessary capitalization. But we are giving it our very best try, and we believe we will win it eventually.

Chapter 5

Selected Approaches to Overunity Power Systems

> [On point-contact transistors]. *"First, the theory underlying their function is imperfectly understood even after almost a century..., and second, they involve active metal-semiconductor contacts of a highly specialized nature. ...The manufacturing process is deceptively simple, but since much of it involves the empirical know-how of the fabricator, the true variables are almost impossible to isolate or study. ... although the very nature of these units limits them to small power capabilities, the concept of small-signal behavior, in the sense of the term when applied to junction devices, is meaningless, since there is no region of operation wherein equilibrium or theoretical performance is observed. Point-contact devices may therefore be described as sharply nonlinear under all operating conditions. "* [William B. Burford III and H. Grey Verner] {282}.

5.1 Introduction

In this chapter we present a selection of potential overunity notions, inventions, and experiments by various inventors and researchers. Its purpose is to show some of the more relevant work that has been done, and hopefully to stimulate new thinking by the interested reader. It is not intended to be all-inclusive, but merely representative. Most of the work addressed is by other inventors and scientists, and we give an appropriate commentary from our viewpoint as best we understand the invention and the process used. Not all the concepts presented can be clearly explained, though some can and are.

This is not intended to be a "kit of parts" or an assembly instructions type of presentation. For that, the interested researcher will have to rely on his or her own ingenuity and lots of bench experiments, because no such "kit of parts with instructions" for COP>1.0 electrical power systems exists at present {283,284}.

5.2 Tesla's Shuttling of Potential Energy and Barrett's Extension

See Figures 4-4 and 4-5 in Chapter 4. In his patented circuits, Tesla accomplished the novel function of "shuttling" the potential itself (and therefore potential energy) around in the circuit at will. This startling function accomplished by Tesla is impossible to see in a conventional vector or tensor analysis of his patented circuits, but it is clearly evident in a higher symmetry EM analysis of them. So it does not appear in the conventional texts, which use vectors or tensors for their circuit analyses. Barrett, one of the leading higher group symmetry electrodynamicists and also one of the pioneers of ultrawideband radar, utilized quaternion electrodynamic analysis to clearly show this novel effect in Tesla's actual circuits {285}. Thus, Tesla was able to produce and move around asymmetrical self-regauging in his circuits, wherever he wished. That has been totally missed by scientists prior to Barrett's revealing analysis.

Barrett then went on to extend Tesla's basic mechanism and employ it for communication and radar purposes, obtaining two patents {286a-b}.

It would appear that use of such shuttling techniques is probably applicable to reducing the back emf in conventional nonshuttling circuits, leading to new COP>1.0 mechanisms. The definitive answer remains to be determined by some future researcher.

5.3 Moray's Radiant Energy Device

One of the remarkable examples of a successful COP>1.0 system was the radiant energy system of Dr. T. Henry Moray of Salt Lake City, prior to World War II. Figure 5-1- shows Dr. Moray and his final 50 kW unit, powering an assembly of light bulbs {287}. The device weighed 55 pounds, used an antenna connected as a "signal input" from the ether, but needed no input of electrical energy by the operator in order to achieve its steady power output.

As shown in Figure 5-2 and Figure 5-3, the various Moray tubes were complex in structure. In addition to various grids, they usually contained (in quadrature) three or more amorphous semiconductor pellets pressed in large presses and sintered with locked-in stresses. One points out the presence of optical effects including phase conjugation, possible use of the built-in stress as a sort of self-pumping and mild self-amplification, possible iterative feedforward and feedback loops due to iterative phase conjugation, collecting of additional energy from the usually nondiverged

Heaviside component via the iterative reverberation of energy, and non-interfering fields in quadrature.

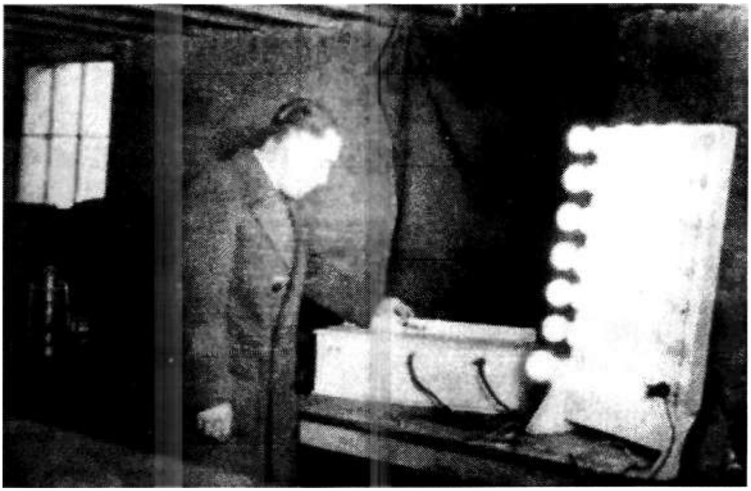

Figure 5-1 T.Henry Moray's 50-kW unit.

The blown quartz tube envelope, in the dark of an enclosure, is open to the passage of infrared (heat) and ultraviolet, which spans a harmonic interval. This allows the passage *of difference* frequencies as sine waves without breakup, even in a nonlinear isotropic environment. Moray stated unequivocally that the tubes would not work unless the envelopes were made of blown quartz. The pellets (shielded from outside light) would nonetheless receive IR input from the ambient temperature, and probably triggered at least weak UV due to the radioactive decay continuously ongoing in the semiconductors. We point out that the UV band is roughly twice the frequency of the IR band, so the "difference frequency" is roughly the IR band itself. In short, it may be that "difference frequency" IR is thus processed in Moray tubes in much more sophisticated fashion than has previously been suspected, including in a self-amplifying (regenerative) manner.

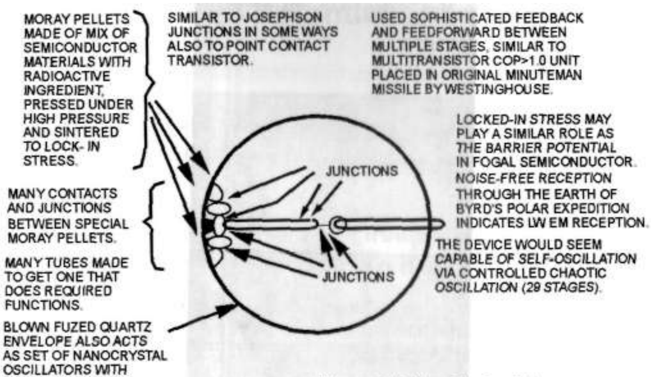
MORAY PELLETS MADE OF MIX OF SEMICONDUCTOR MATERIALS WITH RADIOACTIVE INGREDIENT, PRESSED UNDER HIGH PRESSURE AND SINTERED TO LOCK- IN STRESS.
SIMILAR TO JOSEPHSON JUNCTIONS IN SOME WAYS ALSO TO POINT CONTACT TRANSISTOR.
USED SOPHISTICATED FEEDBACK AND FEEDFORWARD BETWEEN MULTIPLE STAGES, SIMILAR TO MULTITRANSISTOR COP>1.0 UNIT PLACED IN ORIGINAL MINUTEMAN MISSILE BY WESTINGHOUSE.
LOCKED-IN STRESS MAY PLAY A SIMILAR ROLE AS THE BARRIER POTENTIAL IN FOGAL SEMICONDUCTOR.
JUNCTIONS
NOISE-FREE RECEPTION THROUGH THE EARTH OF BYRD'S POLAR EXPEDITION INDICATES LW EM RECEPTION.
MANY CONTACTS AND JUNCTIONS BETWEEN SPECIAL MORAY PELLETS.
THE DEVICE WOULD SEEM CAPABLE OF SELF-OSCILLATION VIA CONTROLLED CHAOTIC OSCILLATION (29 STAGES).
MANY TUBES MADE TO GET ONE THAT DOES REQUIRED FUNCTIONS.
JUNCTIONS
BLOWN FUZED QUARTZ ENVELOPE ALSO ACTS AS SET OF NANOCRYSTAL OSCILLATORS WITH LOCKED-IN STRESS.

Note: see T. H. Moray, The Sea of Energy in Which the Earth Floats, Fifth edition, Cosray, Salt Lake City, 1978, p. 70.

Figure 5-2 Diagram and some novel features of T. Henry Moray's special tube.

The pressed semiconductor pellets were placed in the tube in quadrature, in a sort of "point or junction contact" with each other.[121] Each semiconductor contained a mix of triboluminescent zinc (a semiconductor material), a radioactive material (probably uranium dioxide), and another material of unknown composition, rumored — but not confirmed — to possibly have been diatomaceous earth which also has some unusual properties one may look up. There were some 29 amplifying tube stages in the device. A U.S. patent {288} was eventually granted on an electronic therapeutic unit containing Moray tubes, but the patent on the power system was never granted, although the application was kept current for some years by the Moray family. Eventually the power device patent application files disappeared from the U.S. Patent and Trademark Office.

Although its full technical details were never released, the Moray device is of interest because Moray demonstrated it many times to engineers and scientists, under rigorous measurement conditions, for years. These expert witnesses then signed affidavits to the results achieved by Moray in the tests. Moray also discovered that he could produce an additional 50 kW of power simply by inserting an additional tap just prior to the formal output.

---

[121] Indeed, the junctions of such sintered pellets are most interesting. Such a grainy junction can be compared to a host of very close point contact junctions, particularly with finely-ground powdered materials used in the pressing and sintering. In essence, Moray's junctions contained a great many "point contact transistors", and it is known that such transistors can behave (particularly under pressure!) as true negative resistors. This also appears to be one of Moray's secrets.

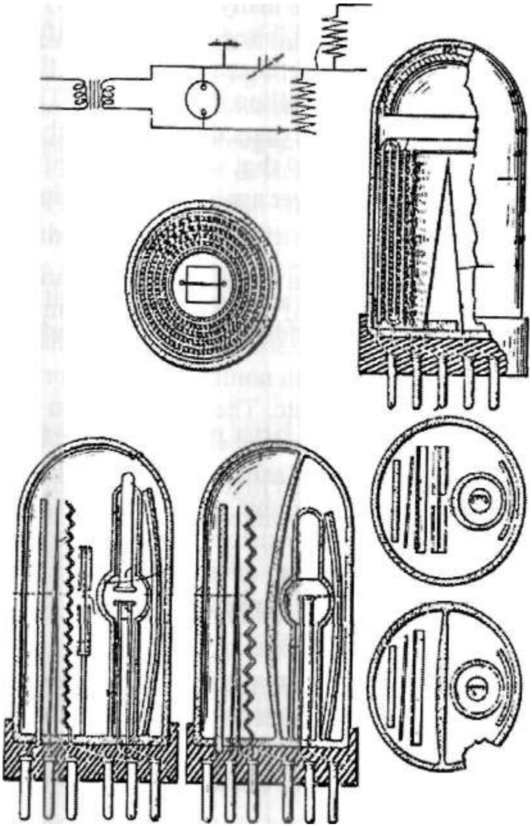

Figure 5-3 A variety of Moray tubes showing their complexity.

Repeated assassination attempts were made against Dr. Moray. While driving his car on the streets of the city, on several occasions he was fired at from an alley or from the sidewalk, but he had bulletproofed the car because of continual threats and ambushes. He was also shot in his own lab in a physical struggle with one would-be assassin, but survived although seriously wounded. A Russian company tried to steal the device, and a Russian agent was planted in Moray's lab, gradually sending essentially all the technical details of the Moray device to Russia. That agent later destroyed the unit right in Moray's lab by suddenly smashing it with a hammer.

Moray's tubes were quite expensive and were painstakingly made by hand. He fabricated a number of tubes (as many as 30 or more) before he obtained one that "worked" and exhibited the self-amplifying effect. Over the years, he spent several hundred thousand dollars on the project, which is the equivalent of some twenty million dollars today. The 29 tubes in his final 50 kW unit were the result of years of labor and fabrication of several hundred rejected tubes to get the 29 that worked. Accordingly, once those 29 were destroyed, Moray was never again able to obtain sufficient financing to rebuild his unit.

However, Moray had demonstrated a special type of transistor years ahead of the first transistor patent. He had demonstrated the importance of multiple junctions in transistor materials. Optical and nuclear functioning of the tubes was also woven in with nonlinear oscillations, multiple feedforward and feedback loops, etc. The sophistication of the tubes and feeds alone almost certainly guaranteed chaotic self-oscillation. We suspect but cannot prove that the Casimir effect may also be involved in some of his conducting metallic double surfaces.[122] He had demonstrated a remarkable "tube-with-internal-transistors" that produced energy output self-gain, without any operator input of energy. In addition, he had successfully used his special tube/transistor hybrids to produce a very successful COP>1.0 EM power system.

After Moray's death, his two sons John and Richard continued to try to rebuild the unit, but were never successful due to lack of sufficient funding and lack of a proper scientific team. They do know how to build the tubes, however, and they do know how to build the unit.

It appears that no electrodynamicist or physicist skilled in higher group symmetry electrodynamics ever examined the Moray device or worked with Moray. No particle physicist or electrodynamicist skilled in non-Abelian electrodynamics ever worked with the device. Since broken symmetry was not discovered and proven until 1957, no technical person who worked closely with the unit was ever aware of the various broken symmetries by which some of the virtual energy of the vacuum can be and is transformed into observable EM energy.

---

[122] The "grain" of the stressed semiconducting pellets may be important with respect to the Casimir effect versus the frequency. In short, Moray's contact surfaces between his semiconductors were "grainy" and these grains would have constituted Casimir surfaces. For the importance of the shape of the double surfaces, see M. Chaichian et al., "Quantum theories on noncommutative spaces with nontrivial topology: Aharonov-Bohm and Casimir effects," *Nucl Phys B,* Vol. 611, 2001, p. 383-402.

Certainly no one ever worked with it that understood the peculiar action of the difference frequency in traveling linearly through an isotropic nonlinear medium. John Moray, who stalwartly carried on the Moray lab in Salt Lake City until a few years ago, does know how to build the tubes, and he knows the circuitry for the device. That he does not understand the "deep theory" is of no consequence — no one else does either. In the opinion of this author, a funded scientific team — carefully chosen and working with John Moray — could probably rebuild the Moray device. We would very much like to see a Moray unit in production and contributing to permanently solving the energy problems of the world. Sadly, that does not seem likely.

5.4 Anti-Stokes Emission and Similar Processes

Stokes' law states that the wavelength of the fluorescent light emitted by a phosphor or other luminescent substance is always greater than the wavelength of the radiation used to excite the fluorescence {289}. This condition was observed by Stokes in 1852 {290}. From an energy viewpoint, the collecting medium collects and dissipates some of the input energy in the medium itself (e.g., to accelerate moving molecules) and then emits the remainder of it. The "system" consisting of the medium, the experimenter's input, and the output is such that it usefully outputs as luminescence less energy than the experimenter inputs. Hence it exhibits a COP<1.0, by absorbing excess energy in its "environmental surroundings".

Usually such a "law" generally holds but is sometimes violated in at least a lew cases. Stokes' law is no exception. It applies for most cases of fluorescence and luminescence, but it does not hold for a few of them. In a few cases, the emitted wavelength is precisely the same as the input wavelength. This means that the collector/absorber re-emits all the collected input energy as fluorescence or luminescence. Here the COP = 1.0. This condition is known as the *resonance condition* and the emission is called *resonance emission.*

In other well-known cases, Stokes' law is violated and the energy emitted is greater than the energy absorbed. In other words, the system exhibits a **COP>** 1.0 *with respect to energy emitted versus energy input by the experimenter.* This is known as anti-Stokes emission. It has been well known for six decades that anti-Stokes emmision processes do indeed output more energy than one has to input oneself {291}. This excess energy emission is usually euphemistically attributed to "additions from the internal energy of the molecule" and referred to as "negative resonance absorption by the medium" or just as "negative absorption by the medium".

Unless the excess energy taken from the physical medium of the system is replaced from outside the physical system, the anti-Stokes emission process produces cooling in the physical system and is self-quenching once the medium loses all the energy it can afford. For the overunity researcher, a huge hint is to search the literature for anti-Stokes type emission phenomena that do not result in media cooling. The Letokhov phenomena {292} are in fact such phenomena where the excess emitted energy is received from outside the physical system medium, and therefore from the surrounding active vacuum and/or the surrounding local curvature of spacetime.

Other than for a very few physicists — e.g., Bohren {24}, Letokhov {157}, H. Paul and R. Fischer {25} — there appears to have been little *effective* discussion of where the excess energy "extracted from the molecule" really came from initially, and how it is replenished to the molecule (if at all), and particularly *whether or not it has to be replenished to the molecule by the operator.* There has been no discussion of the local potentials as dipole asymmetries in the virtual energy exchange with the vacuum, although there has been inadequate discussion of the involvement of the Poynting energy flow (e.g., by Bohren, Paul, and Fischer). There has been no discussion of the possible involvement of the Heaviside nondiverged component of the energy flow, which is additional to the Poynting component but unaccounted by today's electrodynamicists. To my knowledge, the broken symmetry of every dipolarity has not been discussed with respect to such experiments.

The reader can see the point. If a mechanism exists or can be evoked to freely resupply the lost energy to the molecule from the other two components of the supersystem, then the resulting *self-replenishing* anti-Stokes emission phenomenon can be utilized as a legitimate and continuing COP>1.0 process. It remains to be seen whether this can be worked out. If we wish, the giant negentropy mechanism of the source dipole and source charge {308} may be considered universally proven cases of self-replenishing anti-Stokes emmission, already exhibited by every charge and dipole in the universe.

At least one audacious scientist — Letokhov — has suggested that similar processes in certain systems can perhaps comprise true Maxwell's demons, yielding excess energy output and perhaps eventually a COP> 1.0 operation {293}. Russell has suggested that a similar COP>1.0 process is involved in the fiber fuse phenomenon {299}.

Over the last few decades there has been slow but important work in an optical area of possible anti-Stokes emission phenomena {294a-c}. A variety of strongly scattering, optically active media have been shown to produce substantially more energy output than the experimenter inputs to stimulate the output. However, most are being done in laser-like pumping situations and so the pump energy has to be furnished by the operator. This means that the "replenishment" energy to the medium is furnished by the operator, resulting in overall COP<1.0 performance. However, in self-oscillation conditions, the medium can be self-pumped (self-replenished, taking its replenishment energy from the active vacuum exchange when viewed in the supersystem), and in such conditions COP>1.0 performance is possible, at least in theory. Presently we know of no one who has achieved it — except in the destructive fiber fuse effect — but the work seems to be progressing steadily toward that eventual outcome.

5.5 Gain in Intensely Scattering Optically Active Media

This modern work — which may be regarded in one sense as extending the anti-Stokes emission effect — is largely being pursued in intensely scattering optically active media. A very nice example is given by Lawandy's experiment {295a-d}, shown in Figure 5-4. Figure 5-4a shows a comparison experiment in water without an intensely scattering optically active medium but with a fluorescent dye to show the emission interaction. Figure 5-4b shows the same experiment with titanium dioxide particles added to the fluid. The TiO_2 particles are sized so that their optical resonance is within the laser's frequency domain.[123]

First, as shown in Figure 5-4a, a small, weak laser beam is aimed into the solution, and a small "warm little glow" of fluorescence results just around **that** region where the laser beam interacts with the solution. Scattering in **the** medium is rather normal, and provides nothing of any great interest to the overunity researcher.

---

[123] In passing, we point out the involvement of the Bohren effect due to particle resonance. This collects additional energy from the usually nondiverged large **Heaviside** energy flow component.

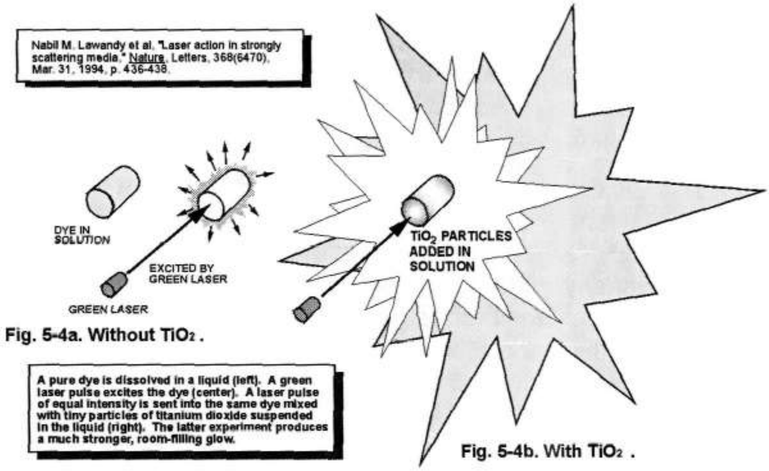

Nabil M. Lawandy et al. "Laser action in strongly scattering media." Nature, Letters, 368(6470), Mar. 31, 1994, p. 436-438.

Fig. 5-4a. Without TiO_2.

A pure dye is dissolved in a liquid (left). A green laser pulse excites the dye (center). A laser pulse of equal intensity is sent into the same dye mixed with tiny particles of titanium dioxide suspended in the liquid (right). The latter experiment produces a much stronger, room-filling glow.

Fig. 5-4b. With TiO_2.

Figure 5-4 The Lawandy experiment

Then the titanium dioxide particles are added, and the same weak illuminating laser beam is directed into the colloidal solution. Figure 5-4b shows the new and spectacular results.[124] Immediately a very bright, room-filling emission of scattered coherent light fills the entire room. The intense optical scattering includes extensive retroreflections for some "ping pong" between individual particles, so that the optical gain is enormous. A highly enhanced energy emission is now evoked, for the same energy input by the operator as was used in the first experiment without the TiO_2 particles. Lawandy's experiment is inexpensive and can be repeated in any university nonlinear optics laboratory and by many individual experimenters. It gives results every time, without fail.

Oddly, no one seems to clearly state that, for a single pulse of input laser energy to the suspension of TiO_2 particles, precisely how much energy was input and how much energy was then emitted by the medium. Instead, "gain" is mentioned, but never the COP = (energy output by medium emission) / (energy input by operator). One suspects that journal referees would probably not allow such a clear and unequivocal statement of overunity, but would apply quibbling and spin control to prevent stating that more energy is output by the medium than is input to it by the operator.

---

[124] This is a very easy and very convincing argument for the presence of the long-ignored Heaviside component of energy flow arbitrarily discarded by Lorentz more than a century ago, and still discarded by modern electrodynamicists.

It follows that, if sufficient "ping pong" iterative retroreflection and multiple collection occurs in the medium, the COP>1.0 because more work is done on exciting the medium than the energy input to it by usual (Poynting) calculations. No one accounts the energy unwittingly input to it by the ignored Heaviside component accompanying the Poynting component. Usually the output/input is just referred to as "optical gain". We stress, however, that we are considering only the actual laser beam energy being furnished into the colloidal solution as input, and we are not calculating the efficiency of the laser, its pumping, etc. Anyway, for this stimulated emission process, it is possible that the COP>1.0, and one can see the parallel to anti-Stokes emission. In short, we are considering the TiO_2 solution itself as exhibiting a true negative resistor action.

The field of high gain stimulated optical emission is advancing rapidly {296}, although many experiments still use external pumping furnished by the operator. If the operator has to pay for the replenishment (pumping) energy as well as the input energy, then the COP<1.0 overall. But the experiments are tending toward sustained self-oscillation and self-pumping conditions. If such conditions are obtained in sufficient magnitude, the process will then become a legitimate COP>1.0 operation overall. In that case, it will be usable as a basis from which to develop self-powering COP>1.0 electrical power systems, particularly for infrared "heating" systems etc.

The more recent experiments have shown positive feedback loops both in the time-forward and time-reversed paths; trapping of light flow energy (both time-forward and time-reversed) in large random walks of more than 1,000 individual interactions; weak Anderson-type localization; and constructive interference of forward time and reversed time light paths. since such experiments can be performed in the infrared, they point toward a potential *"vacuum-energy-powered overunity heater"* as a **feasible** achievement in the future. In our opinion, there should be a determined and major Department of Energy program oriented to develop exactly that kind of system.

Such a heater can become self-powering by the presence of greater governed positive feedback during self-oscillation conditions, which will allow sufficient excess collection due to *multipass multicollection* from the usually wasted giant Heaviside energy flow component. This process —

with the self-excitation occurring spontaneously as a "kick-in" positive feedback process in an exploding gas — probably accounts for the phenomena observed in the gamma ray burster and other such violent "burster" cosmic phenomena. Re-ignition, afterglow, and similar effects are observed in gamma ray bursters. They are also observed with *remarkable* similarity in the latest experiments in intensely scattering optically active materials the laboratory. Similar phenomena occur in x-ray bursters as well, and perhaps even in the recently observed and confirmed gamma ray emissions from intense storm clouds.

Finally, we point out that many magnetic materials are also photorefractive, and they readily produce nonlinear optical effects at various frequencies. Some barium compounds are typical examples. As an example, multivalued phase conjugate reflection can occur {297}. Such effects did occur in the Sweet vacuum triode amplifier, to be discussed shortly. It may be that Sweet's conditioning of his magnets conditioned their barium nuclei into self-oscillation, self-pumping, and thereby into *self-replenishing* stimulated optical-type emission at ELF frequencies.

5.6 The Fiber Fuse

See Figure 5-5. Although presently a one-shot destructive process, another multicollection phenomenon exhibiting a novel COP>1.0 process is the *fiber fuse* {298a-e}. This is a well-known but unusual effect in optical fibers and in fiber optic cables.

See Figure 5-5a. Once initiated by simply heating the cable (as with a butane cigarette lighter), the fiber fuse ignites and is self-sustaining. In those susceptible systems utilizing power on the order of one watt or so, the ignited fiber fuse propagates down the cable slowly, at about a meter per second nominally. It is evidenced by a bright white light propagating down the fiber, resembling the burning of a fuse for explosives — which led to the name "fiber fuse".

In fiber optic cables, it occurs only in those that contain germanium in their cores, and not in those containing just silicon in their cores. It moves down the fiber or cable at approximately one meter per second, and it is unlimited in the length of cable that it will continue along and march to the end. In the core, the marching fuse is actually "hopping", so to speak, and it melts a little "hole" about every centimeter or so, blowing out the material from the hole as it occurs. This "steadily and slowly marching" series of core melts results in the catastrophic destruction of the optical fiber core of the cable and thus destroys the functioning of the cable.

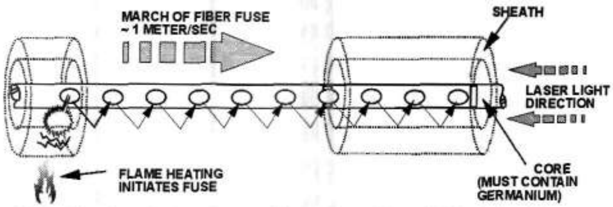

Figure 5-5 A. Fiber fuse destroys core of fiber optics cable of indefinite length, pitting core with holes.

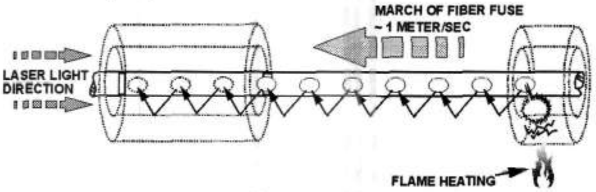

Figure 5-5 B. Reversed second fiber fuse often *restores* pitted core in cable, filling holes back up.

Figure 5-5 The fiber fuse.

The mechanism has been theorized to be self-propelled and self-focusing. Other experiments have found evidence for thermal shock waves and chemical reactions.

But see Figure 5-5b. Astonishingly, after a — say — 1 kilometer fiber optics cable is destroyed by such a fiber fuse, it is often possible to "heal" the damage by reversing the direction of the normal laser light propagating down the cable, while re-initiating the fiber fuse from the other end by heating the cable on that end. In that case, sometimes the *reverse-ignited* fiber fuse will then march right back down the cable, precisely refilling those previously-melted little holes every centimeter or so, *restoring the functioning of the cable!* If ever a phenomenon cried out for further painstaking investigation to determine its actual mechanism, the fiber fuse is it.

Truncated coupled-wave theory applies to this phenomenon. A rigorous a**nalysis** by Russell {299} of power conservation in dielectric gratings has shown that truncated coupled-wave theories do not rigorously conserve power. Predictably such conclusions have been disputed {300a-300d}, however, and are adulterated by the same old use of Lorentz-surface integration of the Poynting vector around the grating and discarding of the nondiverged Heaviside energy flow component. Those disputing the results and claiming the impossibility of COP>1.0 systems continue to ignore the unaccounted Heaviside energy flow component accompanying every EM energy flow situation. In fact, they continue to unwittingly discard that component, and so have never performed *a fully* adequate analysis of any circuit or process. They also are unaware that all EM energy in 3-space comes directly from the active vacuum — specifically, it enters 3-space from the time domain and then returns to the time domain.

As long as the Lorentz surface integration scheme continues to be ubiquitously used without understanding what it does with respect to energy flow accounting, the dispute will not be dissolved. Presently both sides still use the Lorentz integration and discard the available Heaviside component. So they are really disputing whether the *Poynting diverged component* of the input energy can be equal to the energy dissipated. The answer to that is simply "No!"

Of course in a lab test the fiber fuse can be initiated with small Poynting energy input by precision point heating. In that case, the total energy required to melt the holes and move the material out of them, every centimeter or so down that (unlimited) length of cable, can be made as large as one reasonably wishes, simply by making the cable as long as one wishes. In short, COP»1.0. As can be seen, it can be experimentally proven that the input *Poynting* energy component can be far less than the total energy dissipated in damaging the cable. On the other hand, the energy dissipated in damaging the cable *is not* greater than the sum total of the usually calculated Poynting component together with the neglected extraordinarily large Heaviside nondiverged component. Conservation of energy is upheld, but only if the long-neglected Heaviside component is also accounted.

It stands to reason that the reverse fiber fuse can also be initiated by less input Poynting energy component that the total energy dissipated in restoring the cable. At that point, conventional theory has a problem. One can start with a good cable, put in a little EM energy twice — once to ignite the fiber fuse and once to ignite the reverse fiber fuse — and one winds up with a good fiber optics cable and having done lots more work inside the cable than the amount of energy one input by conventional considerations. We leave the solution to that problem as an exercise for the reader. Again, unless one includes the long-neglected Heaviside component, there is no way to make the energy flow balance, and one will have an experiment that seemingly contradicts the conservation of energy law. It doesn't, and that is the exercise the reader should understand.

So unless one uncovers and specifies the external source (Heaviside's energy flow component) of excess energy feeding the fiber fuse, it is a demonstrable experiment whose efficiency ξ seemingly winds up being ξ >1.0. It works every time and seemingly violates the conservation of energy law *as that law is conventionally and imprecisely stated for an equilibrium system only.* The COP>1.0 fiber fuse phenomenon does not violate overall conservation of energy because the system is not in equilibrium with the other two components of its supersystem. It only violates (permissibly!) the conservation of energy statement for a system in equilibrium. The "system" consisting of the fiber fuse and the cable is *not in equilibrium in its energy exchange with its active environment (with the other two components of the supersystem, consisting of the active vacuum energy exchange and the active local curvature of spacetime).* We point out, of course, that the Heaviside component is flowing in space outside the physical system, hence in the vacuum and therefore in the external environment.

Specifically, every dipole in the system is extracting and pouring out energy from the vacuum, and only the small Poynting diverged component of this available energy flow is ordinarily caught. The extreme nonlinearity (and some resonances!) of the particles in and exposed to the extraordinarily locally heated fiber fuse do sweep out a greater reaction cross section than do cold, static charges, in a fashion similar to the separate demonstration of such extended geometrical reaction cross section in the Bohren experiment {301}. *Hence the fiber fuse subsystem and the associated particles of melted matter from the holes do intercept, diverge, and collect some of the accompanying but usually nondiverged Heaviside component.*

See again paragraph 2.1.6 in Chapter 2. Iterative retroreflective systems — such as in intensely scattering nonlinear optical materials — are not necessarily limited to COP < 1.0. Such a system is not necessarily limited to only the energy input by the operator, and it is not limited to only the operator's input Poynting component. Instead, the system can receive an input of excess energy from (i) its active local vacuum, and/or (ii) the local curvatures of spacetime. Its efficiency ξ, however, is always $\xi < 1.0$, even when its COP is clearly greater than 1.0.

Since scientists disputing the energy aspects of the fiber fuse only have assumed an equilibrium system and have not considered the neglected Heaviside energy flow component, the present arguments against COP>1,0 performance of the fiber fuse are quite irrelevant.

We also point out that an optical phenomenon is involved, and that the core materials in the melt are optically active. Hence, appreciable iterative phase conjugate retroreflections occur in the fiber fuse itself and particularly in the melt materials (in the "inferno" of the "spot" itself). This asymmetrical self-regauging process gets more than one joule of work from one joule of original input Poynting energy, because of the iterative (and reciprocating) change of form of the energy. That operation produces iterative joules of work for each joule of original input energy, plus the unaccounted original input of the Heaviside component.

Again, there is no law of *conservation of work* in nature. The present conventional energy-work theorem erroneously considers only a "single interception and change of form" of the energy collected in the system, before loss of the energy from the system. Therefore, it only considers each joule of input energy performing one joule of work on the system to "excite" it to a higher potential energy state. It also considers that, once a joule of energy is then dissipated from the system as work in an internal load or process, it has departed the system forever and is not returned to do additional work by having its form changed yet again! In intensely retroreflective or intensely scattering systems, those conventional assumptions do not necessarily hold. For analysis of the fiber fuse phenomenon, the extended work-energy theorem must be utilized.

Finally, we point out the connection of this fiber fuse phenomenon to the similar "runaway" phenomena we have previously postulated as involved in producing the x-ray burster, gamma ray burster, etc. where iterative retroreflection and thus asymptotic rise in energy density is involved.

Another possibly related phenomenon is the unexpected explosion of porous silicon, as discovered at the Technical University of Munich {302}. The porous pellet involved in the unexpected experiment is covered with a layer of hydrogen atoms. In the presence of oxygen penetrating the oxygen barrier, an extraordinary and superfast chain reaction is initiated. The power of the explosion is some 10 times that of TNT. In the inferno portion of the fiber fuse, the "melt" as it puddles and then particulates, should be in the presence of oxygen and hydrogen from air in the cable. This dramatically differs from the explosive pellet, but may be sufficient to ignite at least in a local area rather than explode violently. If so, between the ablating particles of silicon there could exist iterative phase conjugation, thus producing an asymptotic rise in temperature and energy, perhaps sufficient to sustain the fuse. At least this possibility should be investigated.

5.7 Multiple Feedforward and Feedback Semiconductor Loops.

DeSantis *et al.* {303} showed that feedback systems with a multipower open loop chain can produce COP > 1.0 performance. In semiconductors, the present author hypothesizes that complex feedforward and feedback loops involving semiconductors can induce time reversal effects, hence some movements of current against the voltage. In short, there can arise a true negative resistance effect, as demonstrated in the Fogal transistor.

Indeed, a Westinghouse frequency converter using 64 transistor stages and similar sophisticated feedforward and feedback mechanisms was placed in the original Minuteman missile, then was later deliberately modified by Westinghouse engineers to prevent its demonstrated COP > 1.0 performance.[125] The specific germanium transistor used in the converter was also withdrawn from the market.[126] In those days, power meant weight and weight was at a premium in missile design. The power supply was very tightly designed, so that the weight could be minimized.

The converter had been deliberately designed to be 90% efficient, which was quite the state of the art. However, placed in the missile, the average converter would produce from 105% to 115% times as much output power as input power. (We are stating a COP > 1.0, *not* ξ > 100%!) This was more output power than the tightly designed succeeding stages were designed to accept; consequently those overloaded follow-on circuits often burned out and deadlined their Minuteman missiles — something of great strategic concern at the time. The increased rate of Minuteman missiles out of action due to this problem placed great pressure on Westinghouse executives. In turn, they simply ordered the engineers to "fix it!"

Westinghouse simply "fixed the problem" very directly. The engineers reduced the output power of the converter by adding limiting diodes etc. Very quietly, Westinghouse then obtained several patents {304a-c} surrounding the technology, but no further mention of it appears in the literature. The Westinghouse work continued for at least a decade, judging by the spread of the patent dates.

---

[125] Our information came directly from private conversations of a close scientific colleague with the Westinghouse engineers who thoroughly tested the converter causing the Minuteman problems, and who then altered it to prevent the problems.

[126] Germanium has some unique properties, particularly with respect to transducing surface longitudinal waves to transverse waves and vice versa.

Such multi-loop, multi-feedback transistor circuits can apparently be developed to exhibit true negative 4-resistor effects. With specialized clamped positive feedback of the Bedini type, such open dissipative systems can then exhibit stable COP>1.0, since the energy is taken from the vacuum via the broken symmetry of dipoles, and then discharged asymmetrically in the loads so as not to destroy the source dipole negative 4-resistors. However, the reader should see Chapter 9 for special considerations and phenomenology.

5.8 Negative Resistors

We define a *true negative resistor* as a circuit component or system that receives energy from the environment and outputs useful energy, either in altered form or to power loads usefully. This contrasts sharply with a *differential negative resistor,* which exhibits negative resistance in one part of its regimen, but where all the excess energy to move the current against the voltage in that regime was previously drawn from the circuit itself and therefore was not freely received from the active external environment. Instead, the operator paid for it already. We are not interested in differential negative resistors, which are well known in the literature {305} and do not enable COP>1.0 system performance, so we will not discuss them further. Hereafter when we use the term "negative resistor", it is understood that we mean "true negative resistor" unless the term is otherwise qualified.

The true negative resistor is an open dissipative system *a priori,* and equilibrium thermodynamics therefore does not apply. Instead, the thermodynamics of open systems far from equilibrium applies. The negative resistor freely receives energy from outside the system (from the . environment), and "dissipates" it in interception and collection actions inside the system, to freely increase the available potential energy in the system.

In circuits, the main characteristic of a negative resistor is that the environment freely furnishes some excess energy to (i) power the load, and/or (ii) move the current back against the voltage, particularly when shunted across the back emf region of the source dipole. The *operator* does not have to furnish this excess energy dissipated to propel the current backwards or dissipated to power the load!

One must incorporate the entire supersystem into the analysis of a negative resistance element or system. Since the operator does not furnish the input energy dissipated by the negative resistor, then the active vacuum or curved local spacetime — or both — must furnish the input energy to it. One must consider local curvature of spacetime interactions with the negative resistance and the active vacuum interactions with it, as well as the interactions between active vacuum and local spacetime curvature. In other words, one must again include the active environment of the Maxwellian system that Lorenz/Lorentz symmetrical regauging arbitrarily discarded.

We specifically do not include the well known "differential" negative resistor, such as a tunnel diode, which in one regime moves current against the voltage, all the while drawing all its enabling energy to move current against the voltage from an overall previous energy input by the operator. Unless the input enabling negative resistance energy is freely received from the external environment, true negative resistance is not involved.

Several true negative resistors other than the point-contact transistor have been produced. We review a few of them in the following subparagraphs.

5.8.1 Kron's Negative Resistor

One of the greatest American electrical scientists was Gabriel Kron. Working for General Electric (GE), Kron built a true negative resistor in the 1930s, which could power itself and the network analyzer[127] at Stanford University, under a GE support contract with the U.S. Navy. Here **is** a direct quote from Kron {306} to show what we refer to:

> *"When only positive and negative real numbers exist, it is customary to replace a positive resistance by an inductance and a negative resistance by a capacitor (since none or only a few negative resistances exist on practical network analyzers.) "*

In that sentence we believe that Kron was forced by the censors to insert **the** words "none or" in the expression in parentheses. Particularly note the suggestion that in the imaginary realm it may not be the same as the accepted practice of using inductance as a positive resistor and capacitance **as** a negative resistor. Of course, the fourth Minkowski axis is - *ict,* and so this is very probably an oblique reference to the fact that energy can also "flow in" from the time domain. Kron was quite knowledgeable in general relativity as well as advanced electrodynamics, and he applied full general relativity to rotating EM machines etc. In another quote, Kron {307} also revealed that he had the negative resistor but was not allowed to use it to openly power the Network Analyzer. Quoting and reading through the "spin control":

---

[127] The network analyzer was a large analogue simulation capable of modeling sophisticated equations — such as Maxwell's equations, the Schrodinger equation. **etc.** — for use in designing, simulating, and analyzing naval communications systems or proposed systems. It was one of the most advanced simulations of its day. Kron was the chief engineer and scientist on the project, and one of the ablest electrical physicists the United States ever produced.

> *"Although negative resistances are available for use with a network analyzer, in practice it is more convenient to use a second type of circuit, in which the positive and negative resistances are replaced by inductors and capacitors and the dc currents and voltages are replaced by ac currents and voltages of fixed frequency. The use of the second type of interpretation is equivalent to multiplying the wave equation by i =* $\sqrt{-1}$*."*

After the word "although..." Kron includes the blunt statement that negative resistors were indeed available for use with the network analyzer. But if the "convenient" replacement of the positive and negative resistors by inductors and capacitors was made, then there was no COP>1.0 involved, and there was no true negative resistance involved. Kron is trying to clearly tell us that fact in the above quotation. He also strongly hints that energy from the time domain on the axis *ict* is involved, but that multiplication by $\sqrt{-1}$ discards this "energy from the time domain".

Kron's negative resistor secret has never been released by General Electric, Stanford University, or the U.S. Navy and it is not going to be released. It has never been deciphered outside those groups, with the possible exception of knowledge by Floyd Sweet of Kron's negative resistor construction. Sweet worked in General Electric but not on the Network Analyzer project. Kron, however, was Sweet's mentor and patron, and Sweet often spoke of Kron in glowing terms. In our opinion, Sweet's vacuum triode amplifier was a further development of Kron's early negative resistor, as Sweet understood it.

We further believe that Kron's "open path" is simply the other two components of the supersystem, being (i) local curved spacetime, and (ii) the active local vacuum. Obviously the *external environment* connects any two points in the physical circuit, in addition to the circuit's physical connection (the "closed path"). Between any two points at differing potentials in an electrical circuit, there exists a potential and a dipolarity. Therefore, there exists a broken symmetry because of the dipolarity. The potential difference of that dipolarity is also a curvature of spacetime; i.e., it is a "field" between the two points, and the field is identically a curvature of spacetime. Hence from any two such dipolar points in the circuit, there exists the giant negentropy flow of EM energy that we pointed out in 2000 {308}. This "open path" EM energy flow between any two unlike charges (any dipolarity) in the universe, not intercepted by the circuit and thus nondiverged, is discarded by Lorentz's integration of the energy flow vector around a closed surface enclosing any volume element of interest. It appears that Kron not only recognized (in different terms) what we are calling the "supersystem", but also recognized the equivalent of Heaviside's nondiverged EM energy flow component from any dipolarity — referring to that energy flow as the "open path" flow of energy. It is almost certain that he fully recognized what Lorentz symmetrical regauging did to the Maxwell-Heaviside equations. Thus it appears that Kron was the first scientist to discard Lorenz/Lorentz symmetrical regauging, producing his negative resistor by doing so.

We do not have firm knowledge, however, but only circumstantial evidence that Kron recognized the flow of EM energy from the time domain into 3-space, and vice versa, though he was certainly capable of seeing that. Nevertheless, this would also be included in his notion and discovery of the "open path", since it is included in every dipolarity, generating the scalar potential between the poles of the dipolarity. It may also have occurred to him from his excellent grasp of general relativity and spacetime curvature {309}.

Kron's concepts are certainly capable of being modeled and used to produce COP>1.0 EM circuits and systems, including his negative resistor. Kron did tremendous work in advancing EM networks and power systems. He left behind a great legacy of network and system analysis capable of dealing with the COP>1.0 electrical power system problem {310a-e}. Unfortunately, that legacy has been little used, if at all, in design of electrical power systems and power distribution networks today. To our knowledge, there are no electrical power systems deliberately using Kron's open path.[128]

5.8.2 Chung's Negative Resistor

A highly simplified diagrammatic illustration representing the Chung negative resistor is given in Figure 5-6. In a July 9, 1998 keynote address at the Fifth International Conference on Composites Engineering in Las

---

[128] A possible exception is provided by Barrett's extensions {286a, 286b} of Tesla's "potential shuttling" technique {124}, which can only be seen in higher group symmetry electrodynamics such as quaternions. However, as far as we know, Barrett has **not** applied energy shuttling techniques to electrical power systems.

Vegas, Dr. Deborah D. L. Chung,[129] professor of mechanical and aerospace engineering at State University of New York (SUNY) at Buffalo and Director of the Composite Materials Research Laboratory, reported that she had observed apparent negative resistance in interfaces between layers of carbon fibers in a composite material prepared under pressure.[130] The negative resistance was observed in a direction perpendicular to the fiber layers. By varying the pressure at which the composite material was prepared, Chung *et al.* were able to control whether the material would exhibit negative, minimal, or positive resistance. Apparent negative resistance was also observed in carbon fiber cement-matrix composites and in bare carbon fibers held together by pressure.

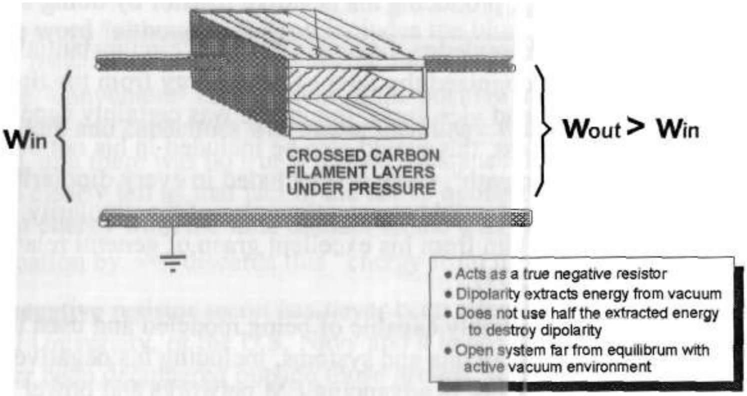
Win
CROSSED CARBON FILAMENT LAYERS UNDER PRESSURE
Wout > Win
• Acts as a true negative resistor
• Dipolarity extracts energy from vacuum
• Does not use half the extracted energy to destroy dipolarity
• Open system far from equilibrum with active vacuum environment

Figure 5-6 Chung's negative resistor (diagrammatic).

A formal technical paper was later published by Wang and Chung {311}. A few years earlier a patent {312} had been issued to Chung and assigned to The Research Foundation of State University of New York, Albany, NY for a superconductive material using a laminate combination of carbon fibers and metal.[131] The patent mentioned mesoporous carbon filaments, strikingly similar to our previous discussion of the surprising explosion of porous silicon at the Technical University of Munich, as well as possibly a special kind of "porosity" consideration in the fiber fuse phenomenon (in the material of the fuse itself).

---

[129] Professor Chung holds the Niagara Mohawk Chair in Materials Research at SUNY and is internationally recognized for her work in smart materials and carbon composites. The reader may peruse a curriculum vitae, research work, and papers at http://www.mae.buffalo.edu/people/faculty'chung/researchl.doc.

[130] The preparation of "locked-in" stress in Chung's carbon fiber material is reminiscent of Moray's similar preparation of his amorphous semiconductors with locked-in stress also.

An initial offer on the State University of New York at Buffalo website to provide a technical package with respect to the negative resistor — for which a patent had been filed — to those submitting a nondisclosure agreement was abruptly withdrawn. The project may have been classified, or — far less likely — it may have been retained as deeply proprietary by the University. We suspect that both the technical package and the patent application have been classified by one or more U.S. government agencies.[132]

In our opinion, Chung's fruitful discovery deserves rigorous and continuing investigation, particularly in higher group symmetry electrodynamics and in terms of supersystem interactions to include Dirac sea hole current interactions. We also believe Chung may have been required by referees (or by declassifying authorities, if such were involved) to use the term "apparent" in the paper. The reported measurements in the technical paper do seem to unequivocally include true negative resistance with current running backwards against the voltage. It is not clear, however, whether this is Dirac hole current or electron current or a mix of both.

We also call attention to the fact that Dirac sea holes in the local vacuum — as will be discussed in Chapter 9 — will and do run backwards against the voltage. Further, such hole current can be transformed from negative energy to positive energy by the Bedini process shown in Chapter 9. The pressure used by Chung *et al.* in fabricating the material produces a stress potential in the material and thus decomposes into our reinterpreted Whittaker 1903 decomposition, preciously discussed. So giant negentropy is involved in the composite material, which implies that true negative resistance can be involved, at least in some operational regimes as pressure etc. are varied. Such results were found by Chung *et al.* and are reported in the paper.

---

[131] We stress that the issuance of this unclassified patent does not rule out a possible negative resistor (hence COP>1.0) patent that may have been classified. Some sensitive defense work in special materials is indeed performed at SUNY at Buffalo.

[132] At least the gist of the work is not lost, however, due to Naudin's duplication of the Chung effect and placing the results and building instructions on his website in France; see http://jnaudin.free.fr/cnr/.

Naudin reported independent success with a simpler variant of the Chung negative resistor {313}. This assures that at least the effect will not just be buried.

5.8.3 The Fogal Charge-Barrier Device

Figure 5-7 shows an illustration from one of William Fogal's patents {314} on a new kind of transistor (more accurately, a charge barrier device). Fogal's device incorporates a very special electrolytic capacitor[133], which introduces optical-type functions as well as the more familiar capacitor functions.

Fogal Semiconductor

- Blocks electron current flow $i = dq/dt$
  - As capacitor charges electrons in emitter junction attracted to plate
  - Charging plate produces high barrier
  - AC conduction electrons cannot penetrate barrier
  - AC displacement current dq/dt and Poynting current $S = E \times H$ are passed through the barrier
- AC into the barrier produces AC Lorentz symmetrical regauging, which produces AC stress waves, which are longitudinal EM waves a *priori*
- Semiconductor becomes "energy pipe"
- Massless displacement current is energy flow without losses
- Energy can be collected and separately discharged through load for free work
- Transmission mode Holding potential while changing displacement current produces enfolded longitudinal EM waves inside potential accompanied by time-polarized phase conjugate twins
- Reception mode Receiving longitudinal EM waves at barrier provides scalar interferometry, making transverse EM waves
- Fogal semiconductor can both receive and transmit via enfolded longitudinal EM waves, which are not limited to the speed of light in vacuum
- Semiconductor can also be used as a true negative resistor.

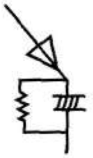

DIAGRAMMATIC ONLY

Figure 5-7 Novel characteristics of the Fogal semiconductor.

Most semiconductor materials are optically active materials. Further, the individual electrons in a signal with a frequency well below light frequencies, are "jittering" with higher frequencies including the entire optical band. A "frequency" of a signal thus is imbedded in a jittering higher frequency complex set, and that set is quite able to interact in optical fashion *if the complex itself is used and controlled.* Fogal's process does appear to both use and control that jittery high frequency signal complex.

---

[133] The exact method of manufacturing this electrolytic capacitor, and its internal structuring, is proprietary to Fogal and a major capacitor manufacturer, who is under a nondisclosure agreement with Fogal.

Hence Fogal's process can "tunnel under" the normal *detected* force-field frequency responses of the individual electron,[134] and directly affect, change, and alter its higher-frequency environmental changes in the force-free field or field-free potential condition. In this way, Fogal has carried the electron (and hole) functions of a transistor to functions well beyond those in the present texts {315}. He may have also extended quantum mechanical tunneling effects to signals between widely separated regions of spacetime, since superluminal communication is possible with the Fogal semiconductor properly used. Longitudinal EM waves — which Fogal's semiconductor is able to utilize and control — can and do move at superluminal speeds {93a) because it uses the long-neglected "infolded" longitudinal electrodynamics revealed by Whittaker in 1903 and 1904. Further, superluminal solutions to Maxwell's equations do exist {93b}.

The charge barrier device is formed in a bipolar design that can be incorporated in Metal Oxide Semiconductor (MOS) designs as well as multiple gate devices. The device also produces a Hall effect EM field internal to the device. The Hall effect magnetic force will force electron flow and angular spin of the electrons in the same direction to the top of the conduction bands in the crystal lattice on semiconductor devices, unlike Silicon On Insulators (SOI) devices that force electron flow to the surface of the semiconductor lattice. This forcing function, together with the "extended tunneling" effects on the embedded EM environment of the electrons and holes, produces a unique effect. The device is able to stop and "pin" electrons while continuing to process the signal energy flow (the entire jitter signal complex) in which the electrons were coupled and moving. Thus the Fogal transistor is able to operate upon and with mass-free spacetime itself, and curvatures thereof. It is the first practical semiconductor capable of performing *engineering of the active vacuum.*

In the last few years we have stressed that EM waves in spacetime are longitudinal EM waves *a priori* — but they are accompanied by time-polarized waves (phase conjugates). Hence the two in combination become observable, as pointed out in quantum field theory by Mandl and Shaw {316}. *We detect effects a priori and not their causes.* We detect transverse *electron precession waves* in our detector circuits, and not the causal waves interacting with them to produce their precession. *The Fogal transistor is able to process the causal longitudinal EM waves directly, prior to their interaction with matter.* It in fact opens up an entirely new region of "detected physical reality" to our vision and detection, a region far more extensive than the development of infrared and ultraviolet sensors opened up to our sight.

---

[134] The electron itself is enveloped in the jittering of whatever background it is exposed to and interacts with. Hence a coupling interaction exists to arrange and **control** the jitter of the electron as well.

The pioneers of electromagnetics, who erroneously assumed a material ether and EM force fields in space, confused the transverse "material detected waves" (electron precession waves) in the detector with the "material undetected waves" they assumed to be incoming from space. Note that neither the electron nor the atom had been discovered when all these assumptions were used to determine how the EM wave in space was to be viewed and modeled. The transverse waves observed in the interacting/detecting of the Drude electron gas — i.e., the detected electron precession waves — were considered to be direct interceptions of the incoming causal waves from space. This already substituted the effect for the cause, as we have previously discussed.

Further, Faraday envisioned the "material field in space" as consisting of a series of "taut physical strings". A "perturbation" of this field in space was, in his view, simply the transverse wave plucking of those strings. Maxwell, who closely followed Faraday's experimental work and concepts, simply assumed the transverse "plucked string waves" advanced by Faraday. Therefore, the erroneous notion of the transverse EM wave in space arose and stayed firmly fixed in the theory.

One can check Whittaker's decomposition of the scalar potential into longitudinal EM wave pairs, then realize that the energetic vacuum is naught but a giant scalar potential — which can be decomposed by Whittaker's method. In short, as a scalar potential, space itself is a writhing mass of longitudinal EM waves and their highly nonlinear dynamics and interferences, or it can be modeled in that fashion. The energy continually enters 3-space from the time domain to the negative charge of every point in space considered as a point dipole, thence to the positive charge, thence back to the time domain. Notice the giant 4-space circulation involved. One also notes the polarization of the vacuum continually happening and arising spontaneously, as is well known in quantum mechanics. We also suspect that this or a similar mechanism is at the root of spontaneous broken symmetry in physics.

An interesting thing emerges from Whittaker's 1904 paper {317}, where he showed that all EM fields and waves can be decomposed into differential functions of two scalar potentials. Each of these two base scalar potentials can be decomposed by Whittaker's earlier 1903 paper {85} into a set of longitudinal EM waves. All EM fields, potentials, and waves are comprised of longitudinal EM waves and their internal dynamics, which also comprise waves of spacetime curvatures and their dynamics. Further, each longitudinal EM wave in 3-space is accompanied by its phase conjugate twin in the time-domain.

This far more primary "internal electrodynamics comprised of pure spacetime dynamics" — comprising and creating the recognized EM potentials, waves, and fields, their internal dynamic spacetime structuring and also the structuring of time — has not been developed and applied in physics theory. This internal or "infolded" electrodynamics is the primary "medium" in which the Fogal semiconductor appears to work and work upon. The Fogal semiconductor seems able to modulate and demodulate this internal structuring or parts of it, as has been demonstrated.

Imperfect longitudinal EM waves — called *undistorted progressive waves* - are already known to be able to travel at superluminal velocity, e.g., as shown by Rodrigues and Lu {318}. Indeed, Maxwell's equations have both subluminal and superluminal solutions, as shown by Rodrigues and Vaz {319}. Further, these waves can carry information. However, the *technology* of employing these waves has been lacking {320}. The use of scalar EM wave interferometry to produce normal EM fields and EM field energy at a distance has been shown by Evans *et al.* {321}.

As pointed out by Rodrigues and Lu {318}, perfect longitudinal EM waves travel at infinite velocity. This eliminates the difference between potential, field, and wave; a perfect longitudinal EM wave can also be taken as an instantaneous scalar potential with net force-free iterative regauging (varying at each point instantly and simultaneously). Both ways of looking at it produce the same end result: instantaneous remote energy change of spacetime energy density itself, in general relativity terms, or of the energy density of the vacuum, in particle physics terms.

Elsewhere, (e.g., Chapter 2), we also pointed out the "total photon **interaction"** mechanism that generates the flow of an entity through time. We also have advanced the concept of a "time-reversal zone" (TRZ) as a region where antiphoton interactions with charges predominate over photon interactions. We used these new concepts to propose exact nuclear **reactions** for the production of excess deuterium, tritium, and alpha particles in cold fusion experiments. In such an "excited state" TRZ region, the law of attraction and repulsion of charges is reversed. In the TRZ, like charges attract and unlike charges repel. Further, the gluon forces are reduced, so the quarks in the nucleons are much more lightly bound. In such TRZ zones, dense clustering of like charges occurs due to their mutual attraction.

A time-reversal zone is created in one part of the Fogal transistor, greatly assisting in the "electron pinning" action. That is, the usual strong mutual repulsion (the Coulomb barrier) of the electrons at the pinning site vanishes in the TRZ inside the Fogal semiconductor, and the resulting mutual attraction of the electrons greatly facilitates the pinning process far beyond what is usually obtained against "like charge" repulsion in a time-forward zone. Indeed, the TRZ in the Fogal semiconductor also materially enhances the electron pinning process beyond what can usually be obtained with cryogenics. At least some indication of trapping and perhaps pinning has been reported in amorphous semiconductors at room temperature {322a-b}.

The Fogal device is able to remove EM noise (electron noise) from existing EM signals, displays, etc. by 96% or more, by pinning the electrons and simultaneously time-reversing the spatial EM signal back to its original noise-free state prior to hosts of interaction electron collisions. This "noise removal" capability has been rigorously tested by independent parties, including at a leading university in the Netherlands,[135] and by several large electronics companies and several other universities. By pinning the electrons (and electron currents) while still processing and propagating the EM energy effects, the Fogal device can move noiseless EM energy without loss at room temperatures, at superluminal velocities, and reconnect the lossless propagated energy to electrons past the pinning region at a receiving region, releasing the electrons in the receiving region. The Fogal semiconductor therefore already accomplishes *room temperature superconductivity* as well as tunneling.

The Fogal device's basic mechanism (consisting of all the submechanisms) is to restore and move EM energy in force-free form, using direct regauging of the vacuum instantaneous scalar potential itself (or the vacuum potentials themselves). It operates directly upon the longitudinal EM waves comprising the vacuum potential. Hence it directly engineers the active vacuum and curved spacetime. I.e., it is the first *supersystem engineering* device of (hopefully) many more to come.

---

[135] The university researchers verified room temperature superconductivity being produced by the Fogal semiconductor in their independent verification measurements. They also measured that some 96% of the noise already present in a circuit was removed by Fogal's semiconductors in Fogal's circuits.

In the "jitter" environment of the space in which the individual electron is embedded, there exist many (and essentially all) field gauges, simultaneously. One of these, of course, is the Coulomb gauge where there exists an instantaneous scalar potential. Usually the instantaneous scalar potential is considered only a *near-field* effect, with no contribution at any significant distance. However, as Jackson points out {323}, transverse radiation fields are given by the vector potential alone. Also, in the Coulomb gauge the scalar potential "propagates" instantaneously everywhere in space {324}.

Along with other electrodynamicists, Jackson avoids the obvious superluminal communications implications of the instantaneous scalar potential by stating that *"..it is the fields, not the potentials, that concern us."* As he points out, a detailed *(though conventional)* discussion of causality is given by Brill and Goodman {325}. Again as pointed out by Jackson, in the Coulomb gauge the quantization of the photon involves only the vector potential. *The use of the scalar potential in the Coulomb gauge, together with deliberate engineering of the internal longitudinal Whittaker wave structure of that potential, directly allows distant engineering via multiply connected spacetime, even inside the nuclei of distant objects. This is in fact a unique and marvelous engineering application and extension of Bohm's quantum potential concept.*

My "pinning" the electrons, apparently Fogal is able to essentially "hold" the vector potential and all the force fields while the instantaneous scalar potential *persists and now extends over all space, not just to the near field since all formerly moving field dynamics that would limit it have been frozen.* If this effect is substantiated, I have nominated it as the "Fogal effect". Further, this unrestricted instantaneous scalar potential now i lunges in field-free form (pure regauging form) over all space according to the signal energy being perpetuated and propagated[136] by the Fogal semiconductor independently of its pinned electrons. In this fashion, the F**ogal** semiconductor is indeed capable of "instantaneous" communication over essentially unlimited distances. At least one independent test has verified the Fogal instantaneous communication effect, but the test is still held proprietary by the testing agency.

---

[136] Here we again point out that the usual notion of a signal as energy propagating in 3-**space** is wrong. It propagates in 4-space, but not in 3-space. Also, what propagates is the spreading point-like polarization of the vacuum itself.

Since Fogal is able to alter and control the internal longitudinal EM wave Whittaker structuring of the scalar potential used in this Coulomb gauge manner, he is able to accomplish superluminal (and even instantaneous) communication, so long as a Fogal semiconductor is available in both the transmitter and the distant receiver. In the transmitter, the Fogal component is necessary to *infold* the information into the internal LW structure of the scalar potential to be used in the Coulomb gauge, thus transforming the input signal information into output longitudinal EM wave information directly in spacetime itself. The Fogal component in the receiver is necessary in order to detect, transduce, and *outfold* the infolded longitudinal EM wave information inside the instantaneous scalar potential. We have stressed that these internal longitudinal EM waves and their dynamics are spacetime curvatures and their dynamics. Hence their transmission and detection is a viable process for superluminal communication, including instantaneous communication.[137]

Further, once the electrons are pinned, all "signal fields in charged matter" (i.e., all EM force fields) are frozen and localized. This, we believe, involves a new physical effect introduced by Fogal that is analogous to the Aharonov-Bohm effect.[138] So I have called it the "Fogal effect". That is, the force-fields are frozen and localized by the pinning, but the field-free potentials then remain, including the instantaneous scalar potential ϕ and probably the curl-free magnetic vector potential A as well. The result is that the instantaneous scalar potential can be directly *regauged* everywhere at once by the changes of magnitude of the "frozen fields", rather as if they were "static voltage", possibly simultaneously regauging the field-free A-potential — all as part of a new kind of Lorentz symmetrical regauging. After all, a single Lorentz symmetrical regauging {326} changes both the potentials A and ϕ. Changes in symmetrical regauging produce stress waves, which are longitudinal EM waves. Continual changes of symmetrical regauging also continually changes both the potentials A and ϕ, *including over all space when operating in the Coulomb gauge.* This is one explanation of the beautiful effect that Fogal has discovered and is utilizing.

---

[137] We believe that physics books will have to be substantially rewritten, once the internal longitudinal structuring of the scalar potential (and of EM fields and waves) is employed directly in open technology. We further believe that the Fogal semiconductors are the first pre-developments in that coming vast new technology, at least in the open Western scientific world.

[138] There have long been, and still are, efforts to "fit" the Aharonov-Bohm effect into slightly modified classical electrodynamics after all. E.g., see C. J. Carpenter, "Electromagnetic energy changes due to charges moving through constant, or zero, magnetic field," *IEE Proc. A,* 138(1), Jan. 1991, p. 55-70; — "Electromagnetic energy and power in terms of charges and potentials instead of fields," *IEE Proc. A,* 136(3), 1989, p. 101-113; G. M. Graham and D. G. Lahoz, "Observation of static electromagnetic angular momentum in vacuo," *Nature,* Vol. 285, 1980, p. 154-155; T. H. Boyer, "Semiclassical explanation of the Matteucci-Pozzi and Aharonov-Bohm phase shifts," *Found. Phys.,* 32(1), Jan. 2002, p. 41-49; — "The Aharonov-Bohm effect as a classical electromagnetic-lag effect: An electrostatic analogue and possible experimental test," *Nuovo Cimento,* Vol. 100B, 1987, p. 685-701; — Classical electromagnetic deflections and lag effects associated with quantum interference pattern shifts: considerations related to the Aharonov-Bohm effect," *Phys. Rev. D,* Vol. 8, 1973, p. 1679-1693; — "Does the Aharonov-Bohm effect

We accent that, with pinned electrons frozen in place, the normal ϕ and A in matter (defined in terms of interaction with charge and the movement of charge) cannot occur. With time reversal zone utilized, electron pinning is greatly eased and the instantaneous scalar potential ϕ now assumes the primary active propagating role in surrounding space. Neither ϕ nor A is free to develop the normal $\partial\phi/\partial t$ and $\nabla\times A$ *spatially propagating dynamics of the force fields (effect fields) observed in matter.* In the Fogal process, the instantaneous scalar potential "signal" (the causal signal in 4-space) is generated in the transistor when it pins the electrons and freezes the force fields into "equal and opposite" form. Since no net force field can develop, the instantaneous scalar potential is a stress potential in spacetime itself and simply changes its magnitude instantly, at each and every point in space that it occupies, when the Fogal chip's input EM fields change in its signal input.

In the output, the net force fields remain zero and frozen, even though the individual force fields change constrained to "equal and opposite" manner. If one wishes, one may take this part of the Fogal output — the field free changes in the stress energy density of the vacuum — as one expression of *modululating the instantaneous scalar potential's energy density in perfect correspondence to the field changes in the normal EM input to the Fogal transistor.*

exist?". *Found. Phys.,* Vol. 30, 2000, p. 893-905; — "Classical electromagnetism and **the** Aharonov-Bohm phase shift," *ibid.,* p. 907-932.

The instantaneous Fogal changes in magnitude (e.g., instantaneously induced upon a distant collecting charge) are no longer limited to near field effects, or to the speed of light — and communication is not bound to the retarded wave. Now the induced changes of the instantaneous scalar potential exhibit far field effects and these changes — being naught but these special Lorentz condition magnitude regaugings — are instantaneous. This of course is a form of *action at a distance.* We stress that the best astronomical evidence also points to a startlingly superluminal speed of gravitation (curvature of spacetime, which after all is just a longitudinal EM wave or set of them) itself, as pointed out by Van Flandern.[139] If so, since any change of energy in spacetime is a curvature of spacetime, and one form of spacetime curvature can move at superluminal velocity, so can other forms of spacetime curvatures. In our favorite expression, *it only takes one white crow to prove that not all crows are black.*

As is well known, the de Broglie waves of matter always move faster than light, as do appropriate quantum tunneling processes. Further, the quantum potential used by Bohm in his beautiful hidden variable theory of quantum mechanics {347} also appears instantly throughout space, wherever it will be, and with the value it will have. The question of whether a quantum potential can be used to communicate has been posed but not definitively answered in the previous literature. We propose a resounding "Yes!" if one uses the internal structure of the quantum potential for one's communication, applying some component such as the Fogal semiconductor to do the infolding and outfolding.

We also propose that a process for forming and using an instantaneous scalar potential in the Coulomb gauge can be interpreted as the formation and use of a quantum potential or its direct equivalent. If that assumption is true, then the Fogal semiconductor seems to be the first component ever openly built and patented which produces and uses Bohm's quantum potential, *including* its infolded Whittaker structuring and dynamics. It also definitively answers in the affirmative the question of whether the quantum potential and the scalar potential in the Coulomb gauge can be used in superluminal communications.

---

[139] Tom Van Flandern, "The speed of gravity - What the experiments say," *Phys. Lett. A,* vol. 250, Dec. 21, 1998, p. 1-11. Van Flandern gives an extremely important, clear, and succinct summary showing that experiments refute the standard far field assumption of light speed in gravitational effects. Laboratory, solar system, and astrophysical experiments for the "speed of gravity" yield a lower limit of 2×10^{10}c.

Thus Fogal has apparently found a way to directly modulate intelligent change of magnitudes into and upon the internal longitudinal EM wave structure of the instantaneous scalar potential (and upon the quantum potential), as well as iterative "pure *net-field-free"* regauging of the instantaneous scalar potential {327}. It seems to have previously escaped notice that varying the Lorentz symmetrical regauging of a circuit or system easily produces pure longitudinal EM waves. This strongly implies that Fogal has surmounted and removed the barrier to practical superluminal communication and even further has reduced the barrier to practical instant communication.

Several tests — in one of which the present author participated — have indeed shown Fogal transmission of signals, including video signals, between two widely separated points with less delay than light-speed signals exhibit, and also with video bandwidth infolded inside a DC signal. In at least one series of tests in a large company's own laboratory and by its own engineers, communications from a ground-based Fogal chip transmitter apparatus occurred between the ground station, through the uplink into the satellite, back through the downlink to earth, into a ground-based transmission system with repeater stations, and into the receiving Fogal transistor unit in the lab. The transmissions occurred with *no discernible time delay,* something previously deemed impossible in all the textbooks. This was particularly significant since even the conventional time-delays in the repeater station amplifiers and the satellite receiver-transmitters were apparently eliminated. This further accents that the infolded "longitudinal EM waves superhighway" inside all EM fields, potentials, and waves — including inside matter such as the ocean and the earth itself— is a multiply connected space phenomenon. It is the most ideal communication medium, enabling deep space communication without delay, etc.

The implications for nondelayed future remote control of robot spacecraft and robots landing on the surface of planets and asteroids are obvious, as are the implications for direct exploration and measurement of conditions *inside the interiors* of the planets, the sun, the moon, and the earth via **internal** scalar interferometry. We accent that Fogal has done private experiments exploring this vast new physical reality revealed by the "internal longitudinal EM wave dynamics" and its interferometry, and it is mind-expanding (some would say *mind-boggling!).* It is also a great new step in physics and in our very notion of physical reality itself.

T**he** vacuum itself, being highly energetic, is indeed a scalar potential. **Further,** it has any and all gauges. Since potentials superpose, we postulate that Fogal's use of the instantaneous scalar potential is directly superposed on the corresponding instantaneous Coulomb gauge scalar potentials of the entire vacuum. His discovery of how to modulate the instantaneous ϕ magnitude without allowing modulated EM field propagation, constitutes a very special case of tunneling right through "space" itself, using one aspect or "component" of space.

In short, to be a little dramatic one might even apply the term *subspace communication* to this long distance tunneling effect. That interpretation would also seem consistent with the deterministic variation of the internal stress energy density structuring of the vacuum that is produced in the Fogal process.

From another viewpoint, Fogal has succeeded in not only finding the missing chaos of quantum mechanics, but in deterministically engineering and using it.

This is also consistent with Whittaker's decomposition of the scalar potential into bidirectional longitudinal EM waves {328} and his connection between time-polarized EM waves and longitudinally polarized EM waves as reinterpreted (12) to be consistent with quantum field theory {19}. We point out that a single point in time is simultaneously "connected" to every point in the universe. By using a transmission system that back-pedals from 3-space into the time domain, an instantaneous connection to any other spatial point in the universe is therefore available, if one's technology can take advantage of it. Fogal's technology does.

It is also consistent with quantum tunneling experiments that have demonstrated the transmission of intelligent signals (such as Mozart's 40th symphony) at superluminal speed between two points in a waveguide {329a-b}. Regardless of how one chooses to model intelligent superluminal communication, it has been experimentally demonstrated — multiple times and by multiple experimenters in multiple laboratories — that such can occur.

The difference in those superluminal experiments and the performance of the Fogal chip is that the chip uses iterative regauging of the magnitude of the instantaneous scalar potential, as well as deterministic alteration of the scalar potential's internal longitudinal EM wave substructuring, and probably also modulates the internal structuring of a quantum potential. The presently proposed quantum tunneling mechanisms apparently do not do this, at least as they are presently modeled. Further, the quantum tunneling mechanism is difficult to apply; the Fogal mechanism is fairly simple and straightforward to apply, given the Fogal semiconductor at both the transmitting and receiving ends of the communication system.

Pure longitudinal EM waves readily penetrate matter to great distances, since matter is mostly empty space between widely separated mass particles. On the microscopic scale, matter is comparable to something like the solar system, except that intense fields and potentials (bundles of longitudinal EM waves and their dynamics) occupy the space between the particles. All this vast empty space inside matter, being a potential and comprised of longitudinal EM waves, is a giant "superhighway" for the passage of longitudinal EM waves and their dynamics, as indeed is matter itself which is mostly such empty space. Further, even the masses themselves are such, since mass is simply highly compressed spatial energy density, and therefore is highly compressed longitudinal EM wave energy and dynamics. It is just a "denser" superhighway.

Hence a longitudinal EM wave system can readily be developed to "see through matter" using Fogal semiconductors in transmitter and receiver. Fogal in fact has made at least one proprietary prototype demonstration unit that demonstrates this startling capability. The vision is very clear, due to the remarkable absence of noise in the Fogal process. It is also shockingly unique, and it changes our very notion of physical reality. We are not at liberty to further discuss this new "sub-submicroscopic" reality revealed in such experiments. We simply interpret it as jumping the presently assumed restrictions of observation, and allowing the direct or correlated observation of the virtual state entities.

As alluded to, the previously modulated "Fogal type signals" emitted by all matter and other physical entities can also be displayed on a screen using the Fogal process, revealing a hitherto unsuspected physical reality and EM signals environment. In this prototype system, Fogal showed a remarkable and startling new extension of our very perception of physical reality itself. This is another reason why we expect that physics books will have to be rewritten, once the Fogal semiconductors are on the market and available in practical equipment. As every physicist knows, any dramatic increase in our ability to sense finer levels of physical reality inevitably leads to new and startling advances in physics itself and changes our very knowledge of the nature of physical reality.

The ability to bypass and "park" force fields by pinning their source electrons in a time-reversal zone, while continuing the flow of noise-free EM energy at will and at infinite velocity, obviously enables COP>1.0 systems. A pure longitudinal EM wave already has infinite energy, for such a wave appears across the entire universe simultaneously, instantly changing the stress energy density of the entire vacuum in which the longitudinal EM wave is present.[140] The back emf and back mmf fields in an EM system or electrical power system can be frozen and held from restoring the Lorentz condition in the circuit, using the Fogal semiconductor {330}. We presently have an agreement with Fogal for the licensed use of his semiconductors in all such energy unit applications.

A quantum potential is modeled in a multiply connected spacetime, so that a given joule of input energy to one multiply connected point simultaneously appears in every other multiply connected point.[141] Since the Fogal semiconductor apparently can operate in quantum potential or Coulomb gauge fashion, it can act as a *direct energy amplifier,* with respect to the lab observer in a singly connected space. In short, the Fogal process dramatically extends the concept of "conservation of energy" to include conservation in multiply connected space and therefore incorporates observed energy amplification in singly connected space. Again, a broken symmetry (in singly connected space) reveals a higher symmetry (in multiply connected space), and the hierarchy principle is alive and well. In short, with the Fogal semiconductor used in power systems, the law of conservation of energy must be extended to include energy symmetry in multiply connected space but asymmetry in singly connected space. Obviously, such functioning totally transcends the limitations of the present thermodynamics models.

We carefully point out that this is not "energy creation" but simply the energy following the known kind of statistics in physics that allows a thing to be in multiple different places simultaneously. What is actually "instantly transmitted" is the simultaneous regauging of the energy density of the vacuum. At any point in space, the vacuum has an energy density and thus is a scalar potential with a given magnitude. Changing the energy density is simply regauging, which is freely permitted under the gauge freedom axiom of quantum field theory. So instant amplification of energy" over a region of points is simply the instant regauging of the potential already present at each point in the region. It is *a permissible multiply connected space violation* of the present more limited "special case" energy conservation concept, which implicitly assumes a singly connected space. We have been working with Fogal for some time in this area, and we expect to rapidly develop and introduce commercial EM power systems based on the Fogal semiconductor's abilities in multiply connected space, as soon as it is in production and available. Until then, the details of our work in that area must remain highly proprietary.

---

[140] This should not be too surprising, since the bare charge and the bare mass of every electron each has infinite energy, as is well known in particle physics. For confirmation in very straightforward language, see Nobelist Steven Weinberg, *Dreams of a Final Theory,* Vintage Books, Random House, 1993, p. 109-110.

[141] If we wish, we can even take time itself as a special kind of quantum potential, since time also displays such a "multiplicity of connection" in 3-space. In one sense, time is the ultimate quantum potential, since a single point in time is "connected" to every single 3-space point in the universe, simultaneously.

5.8.4 Point-Contact Transistor.

Figure 5-8 diagrammatically shows a point-contact transistor with n-type base in a typical circuit for power gain. The point of the contact is under pressure on the material with which it is in contact.

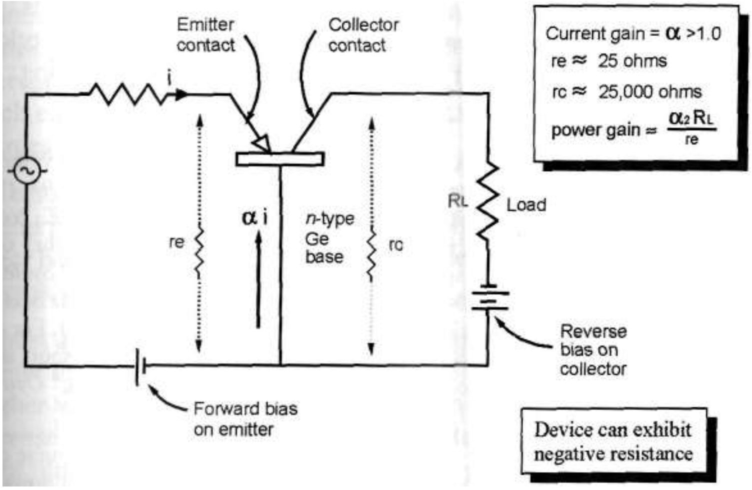
Emitter contact
Collector contact
i
Current gain = α >1.0
re ≈ 25 ohms
rc ≈ 25,000 ohms
power gain ≈ α2 RL / re
RL
Load
α i
n-type Ge base
re
rc
Reverse bias on collector
Forward bias on emitter
Device can exhibit negative resistance

Figure 5-8 Point contact transistor with n-type base.

The point-contact transistor often behaved in true negative resistor fashion, outputting more energy than was input to it. Its production was always far **more** of an art than a science.

As can be seen from our quotation at the beginning of this chapter, the original point-contact transistor was never thoroughly understood technically.[142] Specifically, there does not appear to have ever been any deliberate attempt to capitalize on the ability of the transistor to become a true negative resistor under certain manufacturing techniques and under certain conditions.[143] Instead, the variations and difficulties in manufacture resulted in use of the point-contact transistor being essentially bypassed rather quickly, by advancing to other transistor types more easily manufactured and with less manufacturing variances. In reducing the manufacture of other types of transistors to a science rather than an art, the point-contact transistor and its further development and optimization into reliable true negative resistors were abandoned.

We present this transistor as a straightforward and relatively inexpensive area in which young researchers in COP>1.0 systems and phenomenology may wish to begin, assuming they have access to semiconductor facilities in a university or similar or are willing to work meticulously under a jeweler's loupe. A rich combination of effects in the transistor awaits optimization and control. We predict that we shall yet see, on the world market of the future, such point contact transistors reliably exhibiting negative resistance. It will be necessary, however, to also take into account the symmetrical self-regauging characteristics of the common closed-current-loop circuit. By adroit switching of a true negative resistor in and out of a closed-current-loop branch, or placing it in parallel with the back emf of that branch to reduce it, asymmetry can be introduced into the overall closed current loop circuit, defeating the Lorentz symmetrical regauging. Defeating the Lorentz condition is essential to COP>1.0, of course.

---

[142] E.g., different theories are given in: (a) W. Shockley, "Theories of high values of alpha for collector contacts on germanium," *Phys. Rev.,* Vol. 78, 1950, p. 294; (b) W. R. Sittner, "Current multiplication in the Type A transistor," *Proc. I.R.E.,* Vol. 40, Apr. 1952, p. 448-454. Also of interest is (c) W. van Roosbroeck, "Theory of the flow of electrons and holes in germanium and other semiconductors," *Bell System Tech. J.,* Vol. 29, Oct. 1950, p. 560-607. See also (d) J. Bardeen and W. H. Brattain, "The transistor, a semiconductor triode," *Phys. Rev.,* Vol. 74, 1948, p. 230. For a more modern re-examination, see Shuji Hasegawa *et at,* "Electronic transport at semiconductor surfaces — from point-contact transistor to multi-tip STM," *Oyo Buturi,* 70(10), 2001, p. 1165-1171 (in Japanese).

[143] E.g., the point contact is usually under pressure, and this pressure of course can be varied. The full phenomenology of points (which increase voltage) and pressure, complicated by surface effects as well, has not been worked out in physics. Note the similarity of the pressure in the point contact transistor to the pressure in the Chung negative resistor. Note the almost certain involvement of the "overpotential" of chemistry and electrode chemistry, as well as the "double surface" effect of the small gap between the point and the substrate on which it rests in contact. Part of the gap probably even involves the Casimir effect as well. Any electrical signal variation in that very complex point junction will vary the overpotential, the stress potential, the "point increase in voltage" effect, etc. As can be seen, the phenomenology of the point contact is remarkably complex and rich in several areas of physics and in a great variety of physics variables. Obviously, the reasonable control of all these highly nonlinear variables — and their mutual interactions — is a difficult matter. Hence the readiness with which point contact transistors were dropped with substantial sighs of relief.

We envision such a development as an ideal "negative resistance shunt" to add across the secondary of a transformer, and another possibly across its primary and across the external power supply as well, in producing more amenable and easily fabricated COP>1.0 electrical systems. We also point out that, once the back emf or back mmf forcing of equal power dissipation in the primary of a transformer or in the emf section of a circuit is dramatically reduced or eliminated, one does not require "large intensity" potentials and voltages to collect a great deal of power in the intercepting and receiving external circuit. Again, by simple $W = Vq$, as much energy W can be collected from any nonzero potential V as there are charges q to intercept.

We also envision such a true negative resistor being close-looped by the Bedini process, thus producing a small "self-powering" transistor, which in effect becomes a small self-powering "battery". This is absolutely a doable process, and it will be done once the Fogal semiconductor is in production.

Almost all semiconductor materials are also optically active materials, and a point discharge into such materials represents a very sharp regauging (higher voltage) discharge at a point or into a very small area, due to the increase in potential at the tip where it contacts the base material. The point junction is under pressure, so a stress potential exists there. In addition, the well-known "point" effect also increases the potential in the junction pointer itself, from its base to the point. The point contact phenomenology of different materials — one conductive and one semiconductive — is of much interest, and with novel phenomenology.

The fact that the point-contact transistor in its most usual formulation primarily uses holes more than electrons, is also of much interest in COP>1.0 situations. Holes in a circuit move against the voltage. The trick is to let the Dirac holes *before observation* move in an open path from the ground return line against the back emf to the potentialization line, and **transduce** the moving hole current (via the Bedini process) into electron current after the holes have already freely reached the vicinity of the high side of the circuit.

When Dirac sea holes (causal positrons before observation and thus prior to their alteration to lattice holes and parity reversal) and Dirac sea hole current are also involved with point-contact transistors, the resulting phenomenology has been but little investigated and none of it is in the present textbooks. However, from recent work with positron probes and positron microscopes to examine semiconductors and semiconductor materials, it is known that the holes (positrons) tend to be repelled from nuclei in the material into defects (voids) in it {331}. These effects have been investigated, for example by Triftshauser *et al.* {332}.

Both the mechanical stress potential (which is fundamentally electromagnetic) and the heightened junction potential decompose via our reinterpretation of Whittaker's 1903 decomposition of the scalar potential. This leads to optical-type pumping in both the time domain as well as the 3-space domain. Hence novel optical-type effects and time-reversal of material states can be involved, leading to a very complex set of phase conjugate phenomena, time-reversal phenomena, etc.

Certain Hall effects employed in conjunction with a point contact transistor could be a fruitful area of investigation. As an example, narrow Hall bars with junctions between current and voltage leads of various geometries could be investigated. Widening the junction from the normal square-cornered shape can sometimes produce a negative Hall resistance.

The junction of the point contact involves asymmetrical self-regauging, iterative time-reversal retroreflection, increased Poynting and Heaviside energy flow components, optical scattering processes inside the junction materials, etc. The transistor can indeed be manufactured so that these highly nonlinear effects sum to a negative-resistor-like movement of the output current against the voltage, although with so many other phenomena involved it will require some hard work and research in order to develop and stabilize it — and understand it.

5.8.5 Bedini's Negative Resistor Process in a Battery {333}.

5.8.5.1 Some Pertinent Lead-Acid Battery Processes

A lead-acid storage battery is a highly nonlinear device. The plates may be plante, pasted, or tubular etc. They are usually highly porous and have a thin oxide layer. They may have other coatings and the electrolyte solution may include additives to enhance operation. The plates consist of a mix of crystalline materials in most cases. The plate materials include acicular (needle-like) crystals, particularly the negative plate. The plates usually are porous, knit, and have gridded separators.

The chemistry of the battery electrolyte is complex and — even today — not all of it is well understood, particularly when mass transport is accounted. In our discussions, we will be including some of the previously neglected phenomena during mass transport (ion sluggishness in moving) that are responsible for some of that lack of understanding. In the battery chemistry, there are many double surfaces between the ions and also there is a variety of point voltages. There is a separate theory for double surfaces.

One of the most important phenomena is that the ion currents in the electrolyte are confined to the battery, and do not pass out into the external circuit. Much of the electron currents in the external circuit are blocked at the plates from moving within the electrolyte. Consequently, we will limit our discussion to two major currents of interest[144] in a battery-powered system, and these currents are effectively isolated from each other (at least to first order). Those currents are (i) the ion currents in the electrolyte, confined to between the plates and therefore internal to the battery electrolyte, and (ii) the electron currents between the outside of each plate through the inside of the plate and on out through the external circuit and **then** around to the inside and thence to the outside of the other plate. Chemical changes occur on the plates themselves as a result of these two currents.

It is usually assumed in power systems that these two currents are in phase or almost entirely so. That need not be true at all, and Bedini takes specific advantage of deliberately arranging these two currents to be antiphased part of the time. Part of the Bedini effect can be understood from this fact alone; the other part must consider one other most unusual electrical current that is encountered only in battery-powered systems or COP»1.0 systems.

So indeed there are *several* (many) currents in a lead acid battery [334], not just one, and we will consider three of them — the two previously mentioned, and an additional Dirac sea hole current in the local vacuum **itself.** The reader interested in greater technical depth should also turn to the literature of electrode chemistry and electrochemistry {335a-f}, which is quite complex.

---

[144] Of course, there are many other currents in a battery as well, but — although an oversimplification — these two major currents together with the Dirac sea hole current will suffice for our introductory modeling and explanation purposes.

5.8.5.2 Dirac Sea Hole Current

In a battery-powered COP»1.0 system (and in some COP>1.0 systems), there is a previously unrecognized significant flow of Dirac hole current in the local vacuum, running from the output section of the system to the system input section and thence through the conductors into the battery terminals and onto the battery plates, and on into the battery electrolyte chemistry. This hole current may be greater — *evenfar* greater — than the normal electron current running from input to output.

So in the battery-powered overunity system with positive feedback, not only does an electron current run from the battery through the connecting conductors to the input section of the system, but also a Dirac hole current pours back out of the system input section, and — if not intercepted and transduced — through the conductors and back into the battery. All the while, this Dirac hole current is "eating" electrons being furnished by the battery to the system. The net result is that the battery "sees" an extra load to be powered, in addition to the ordinary load presented to it by the normal system input section. It "sees" the situation as an "extra load" because it first has to furnish sufficient electron current to fill the hole current and "kill" it, before it can furnish the additional electrons to power the system in normal fashion.[145]

---

[145] An odd effect occurs in the battery, however, when the battery completely discharges in the normal sense. Actually, it is still fully charged with positive energy charge, but also has been simultaneously charged with negative energy charge. With the excess negative energy holes continuing to charge the battery with negative energy charge, the battery becomes increasingly *charged with net negative energy.* There is almost no limit to the negative energy charge the battery (i.e., the local Dirac vacuum portion of its supersystem) can take, except the eventual development of antigravity effects due to the inverse and increasing curvature of local spacetime in which the battery is embedded. After a battery has been used to power a COP»1.0 system for some time, one can take the battery off, place it on a normal battery charger, and it will then "eat" positive power for an extended period of time — e.g., a week or longer. During this period, one is steadily negating the inverse curvature of the local vacuum via the increased positive energy collecting in that spacetime. The battery again finally passes through zero voltage and net charge again, and suddenly starts charging up again in a "normal" sense with positive energy charge. This process actually increases the lifetime and function of a battery in a startling manner. Bedini has used batteries almost devoid of charge and energy in the normal sense, and powered systems very well with them, even for longer than a year continuously.

A COP»1.0 EM system is an excited system far from equilibrium in its exchange with its active environment. For all such excited systems, there exist significant decay mechanisms to bring the system back into equilibrium with its environment, since equilibrium yields the lowest energy state and maximum entropy. The Dirac sea hole current (negative energy current) is nature's unexpected natural decay mechanism to sharply decay a COP»1.0 system back to an overall COP<1.0 system when the external power supply is considered also.[146]

So when Dirac hole current back through the local vacuum and through the system is involved, the battery first has to furnish electron current to fill these holes at the rate they are appearing in the input section, thereby furnishing a primary "hole-filling" electron current component which "disappears" into the Dirac vacuum and is lost.[147] Only after the holes in the hole current are filled — and the hole current is being steadily nullified

---

[146] It decays the *true negative resistor* power system back into being only a *dijjerential negative resistor section* of a slightly extended metasystem including the power system and its external power supply. This novel decay mechanism also prohibits close-looping the now differential negative resistance power system for self-powering, unless the hole current is first converted into electron current, as discussed in Chapter 9 and in paragraph 4.9.6 below.

[147] The filling of these Dirac sea holes is not pair annihilation in the usual sense, and no radiation is emitted. Instead, the energy of the radiation that would otherwise be emitted is added to the negative energy curving the spacetime. Hence the net curvature of spacetime relaxes because of superposition. In effect, the vacuum/spacetime itself reabsorbs the energy as relaxation spatial energy added to the curvature energy of the curved spacetime, reducing that curvature. Particularly see Mendel Sachs, *Quantum Mechanics from General Relativity: An Approximation for a Theory of Inertia,* Reidel (now Kluwer), 1986. A generalization of quantum mechanics is demonstrated in the context of general relativity, following from a generally covariant field theory of inertia. Nonrelativistically, the formalism corresponds with linear quantum mechanics. In the limit of special relativity, nonlinearity remains and several new features are derived: (i) Particle-antiparticle pairs do not annihilate; an exact bound state solution is derived corresponding with all experimental facts about annihilation/creation — which, in approximation, gives the blackbody radiation spectrum for a sea of such pairs, (ii) A result is proven, without approximation, that is physically equivalent to the Pauli exclusion principle- which in linear approximation gives the totally antisymmetrised main-body wave function and Fermi-Dirac statistics, (iii) The hydrogen spectrum is derived, including the Lamb shifts, in agreement with experiment; new results are found for high-energy electron-proton scattering, (iv) Finally, several applications to the elementary particle domain are demonstrated, in agreement with results from experimental high-energy physics.

— can the battery then send additional electrons which do not disappear and which are accepted in the system input section to power the system. We discuss this Dirac hole current in some detail in Chapter 9, including how to convert it into useful input electron current to the system, reducing the current load on the external generator and enabling self-powering. Bedini invented the method of transducing available negative energy hole current in the vacuum into electron positive energy flow to charge the battery and power the system, preventing decay of the $COP \gg 1.0$ system back to $COP<1.0$.

We shall not discuss the complexity of the additional chemical reactions ongoing in the battery electrolyte and on the plates. The two main mass-containing currents and the negative energy Dirac hole current will suffice to explain the Bedini negative resistor formation and usage, at least to first order and sufficient to stabilize systems.

5.8.5.3 Changing Symmetrical to Asymmetrical Regauging

To first order, we reiterate that the electrons flow from one plate of the battery through the external circuit to the other plate, but the lead ions do not. There is thus a sharp "two separate currents" interface at the plates and thus a sharp separation between the electron current and the ion current. In short, there are actually two half-circuits in the battery-powered system if one chooses to use them semi-independently.

The current carriers in the two half-circuits are quite different. Therein lies the opportunity to separate and antiphase the two currents, using one to recharge the battery while the other is used to power the loads. It also allows a good opportunity to asymmetrically regauge both half circuits, greatly and freely increasing the Poynting energy component intercepted, collected, and used by each half-circuit's carriers. This is the basis for a profound "opening" of the system and influx of excess vacuum energy.

The two half circuits meeting at a common interface allow a stress potential upon that interface to *individually* use the two halves of what is normally Lorentz's symmetrical regauging. That usually *symmetrical* regauging (if the complete circuit were unitary) is now comprised of two separated *asymmetrical* regaugings, one for each of the two circuit halves, because the net stress potential (net Lorentz symmetrical regauging) — consisting of two opposing fields and forces — is split into two now-individually-separated and separately utilized force fields.

We accent this process by an analogy. Suppose we suddenly place a scalar potential ϕ upon a point in the middle of a transmission line. Immediately the potential ϕ races off (spreads) in both directions simultaneously, at nearly light speed. Further, in each direction there is a moving gradient of ϕ, hence an E-field by $\mathbf{E} = -\nabla\phi$. These E-fields or "emfs" are oriented in opposite directions, so there is a movement of charge from the middle of the line toward each of the ends. The same "dephasing" current effect can be applied to any conductive medium or to the interface between two conductive circuits.

In Bedini's negative resistor case, the placement of the potential ϕ — for its movement in both directions — is on the interface (the face of the lead plates) between the two half circuits and therefore between the two different current carrier types. An initial "Lenz's law" reacting with the suddenly placed E-field onto the plate surface, creates an E-field in opposition, forming a stress potential ϕ which is much higher than the normal potential of the battery (in a nominal case, 100 volts instead of the 12 volts of the battery).

Immediately this potential "takes off in both directions", with a powerful E-field and emf back into the outside circuit and its load, while a powerful E-field and emf also proceed into the battery electrolyte in the opposite direction. So the ions in the battery have a sudden, much higher energy interception and collection forced upon them, with the excess emf now being in *battery-charging mode.* The electrons in the outside circuit simultaneously have an excess emf and potentialization *in the load-powering mode.* Let us see how the two very different current carriers react to these "dual emfs" in the overall circuit.

5.8.5.4 Mass-to-Charge Ratios of the Two Current-Carrier Types

We stress the dramatic difference in the *mass-to-charge (m/q) ratio* of the two currents. The lead ions' m/q ratio is some few hundred thousand times the m/q ratio of electrons, as we recall. For our purposes, all we need to know is that the m/q ratio for the lead ions is *very much larger* than the m/q ratio for the electrons. For a given force, the ions in the internal half circuit in the electrolyte will therefore respond very much slower than will the electrons in the external circuit half.

5.8.5.5 Hysteresis Between the Responses of the Two Half-Circuits

Because of this great disparity in the m/q ratios of the two mass currents, there is obviously a relatively significant hysteresis (time delay) between the *response* of the more massive ion current and the *response* of the far **less** massive electron currents that interact at the plates to try to change the ion current and its momentum. This time-delay between ion response in the electrolyte (the effect) and electron pileup urging (the cause) can be adroitly manipulated. Specifically, it can be used to alter the local vacuum potential and cause it to add excess energy to the ions in charging mode **and** simultaneously to add excess EMF to the electrons in circuit-powering mode. In short, the response delay can be manupulated to "asymmetrically regauge" the mass currents in the system, freely changing their potential energy, and dramatically increasing how much potential energy is available for battery recharging and how much is available for powering the external circuit (loads and losses). Further, during the antiphase condition of the two current halves, the battery can be charged simultaneously while the external circuit half and load is being powered. Excess energy is fed into both actions by the *overpotential* (the stress potential) created at the surface of the plates.

5.8.5.6 Evoking the Initial Bedini Negative Resistor Effect

Figure 5-9 shows the first phase of Bedini's negative resistor process. Consider the battery in normal load-powering operation. From the external circuit, a very sharp leading edge rise of a pulse of electrons and potential is sent to the battery plates in "back-popping" or "battery charging" mode. The back-popping electrons drive in nearly instantly, piling up on the negative plate[148] and trying to force the heavy ions to start moving in the battery recharging direction. The pile-up that results in the stopped electrons on the battery plate surface represents an increase in local current density, hence an increased potential. This produces the 100 volts potential, during the time that the ions have not yet responded to any appreciable degree.

Due to much larger m/q ratio of the sluggish for a moment they lag due to their greater inertia, and this allows the much more agile electrons to "pile-up", producing a substantial overpotential as the local current density increases. The lagging ions are being steadily overpotentialized during this "lag time", up to about 100 volts in a normal 12-volt battery. At the same time, electrons are being forced back out of that higher 100-volt area and into the external circuit and its load, which had been at 12 volts. Recall now that these electrons can move longitudinally only at the drift velocity. Hence all electrons in the external circuit are now highly overpotentialized, and in *load-powering mode. The* overpotentialized electrons are thus producing excess power in the external circuit, much more than normal.

---

[148] We are not using conventional "positive current and positive plate" as the high side of the potential. We are using "electron current and the negative plate" as the driving potential plate.

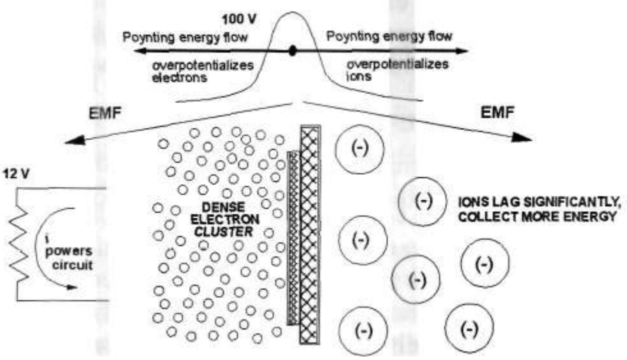

1 Sharp leading edge voltage pulse produces intense electron clustering and large potential (negative resistor) *at* interface *Sluggish ions only beginning to move*

2 Bidirectional Poynting energy flow from potential overpotentializes both the electron cluster and the ions, which are only beginning to move in charge mode

4 Large back emf Into circuit returns overpotentialized electron current, powering circuit with excess energy pouring from negative resistor potential Continues dunng ion delay

5 Simultaneously overpotentialized ion current is overcharging battery

Figure 5-9 Phase I of the Bedim negative resistor process in a storage battery.

Then the overpotentialized ions very slowly (compared to the electrons!) slow, stop, and begin to move in the opposite direction. They reluctantly respond and move in *battery recharging mode.* During that ion-response lag time, and the slower initial portion of the response, the electrons in the pulse continue to furiously surge in and pile-up on the negative plate, overpotentializing both the ions and the external circuit's electrons, while also some of them are being impelled back out into the external circuit to power it with extra overpotentialized energy dissipation. The charge density at that plate sharply increases due to the pile-up where the charges are "squeezing" together (clustering). There is a much higher potential suddenly rising in the squeezed charge cluster, because of the increased charge density there. As we stated, this potential nominally may be about 100 volts during this initial phase.

We call attention to the simple equation W = VQ, where W is the potential energy added to charges Q exposed to voltage V.

Simplified, the excess energy W_{IONS} freely impressed upon the ions is

$$W_{IONS} = f(100 - 12)(Q_{IONS}) \qquad [4\text{-}3]$$

where Q is the total coulombs of charge of overpotentialized ions and f(100 -12) is the magnitude of the overpotentialization of the ions.

At the same time, an excess energy is impressed upon the electrons in the external circuit by the same potential extending along the conductors into the external circuit. Again, this excess energy WE is given by

$$W_E = f(100 - 12)(Q_E) \quad [4\text{-}4]$$

where Q_E is the total coulombs of charge of the overpotentialized electrons and f(100 -12) is the magnitude of the overpotentialization of the electrons.

But the emf on the electrons from the pile-up is directed in load-powering direction, as can be seen. Consequently the ion current and the electron current have been deliberately dephased by 180°, and so the overpotential energy of the ions is delivered in battery-recharging mode, while the overpotential energy of the electrons is delivered in circuit load powering mode.

Recapitulating: Pulsed pile-up of excess electrons on the negative plate interface between the two currents, while the ions are beginning to respond or only sluggishly responding, produces a much higher potential (an *overpotential)* on the sluggish ions — nominally some 100 volts in a 12-volt battery. On the negative plate, momentarily there is now a much higher voltage (with respect to the positive plate) than normally exists in the 12-volt battery. This voltage *overpotentializes* both the reluctant charging ions in the battery solution between the plates, and the powering electrons back into the circuit in powering mode due to the reversal of the emf. Since there is, say, 100 volts across the battery momentarily, there is also 100 volts now across the external circuit momentarily. Accordingly, overpotential excess powering of the external circuit load is suddenly evoked, while at the same time overpotential recharging of the battery is also evoked.[149]

In short, the ion current in the battery and the electron current into the external circuit have been placed 180° out of phase, achieving one major requirement for a COP>1.0 electrical system: *violating the integrity of the closed current loop circuit.* The battery is recharging at the same time that the external circuit is being powered, *from the same free overpotential,* and both recharging and circuit powering are driven by increased emf.

---

[149] Another of Bedini's innovations is to shunt the excess voltage (say, above 14 volts) into an external capacitor on the circuit side. In that way with his overpotential he can be (i) overcharging the battery, (ii) powering the load, and (iii) storing excess energy in that capacitor, from the altered vacuum — all simultaneously.

We strongly accent that the overpotential at the plates represents a change in the local vacuum potential, and it *identically is* part of that now-altered local vacuum potential. The vacuum, since it contains enormous EM energy in virtual state, is a very powerful EM potential. Any EM potential in our circuits is automatically a change to the ambient vacuum potential, or a change to another potential that is such a change to the vacuum potential. In the most exact sense, this is a method of overpotentializing the plate interface with excess energy from the vacuum, and then letting that energy flow onto the ions to recharge the battery and onto the electrons in the external circuit to power it and its load.

The Bedini overpotential at the battery plates decomposes via Whittaker 1903 {85} as reinterpreted and previously explained, so that excess energy is entering 3-space there, *from the time domain.* Further, the piled-up electrons on the plates and the ions (as charges) in the solution receive such potential energy from the increased potential on them via the same decomposition process. So the creation of the Bedini overpotential on the battery plates, together with dephasing the two currents, is the creation of a true negative resistor at the plates, freely receiving energy from the external vacuum (from the time domain and virtual state) and transducing it into real potential energy and emf on the internal ions and on the external circuit electrons.

Hence Bedini has invented a process for creating a true negative resistor inside a storage battery, and for suddenly thrusting the system out of equilibrium with both the active local vacuum and the active local curvatures of spacetime. As such, the thermodynamics of open systems far from thermodynamic equilibrium applies, and that system is permitted to exhibit COP>1.0, while complying with energy conservation and the laws of physics and thermodynamics. With adroit use and collection of the excess energy, the externally collected energy can be used to close-loop the system and power all its functions. So the system is permitted to power itself and its loads, with all the energy being received from the vacuum via the broken symmetry created. Any overpotential is a dipolarity *a priori,* since any potential is. Hence creating an overpotential is precisely producing an extra broken symmetry of that dipolarity right there at the interface between the two half-circuits and the two dephased, localized currents.

During the "back-popping" pulse signals, one should not think of the energy pulses that Bedini inputs to the battery as the "powering" energy. Instead, one must think of each pulse as "triggering" and "timing" energy which initiates certain other key *negentropic* interactions to freely occur, once the electron pile-up occurs. The resulting negentropic interactions then add substantial additional energy (from the local active vacuum) to the ions in the ion current and to the electrons in the electron current. The freely added energy can be appreciably more than the switching or triggering energy that is dissipated as the "input by the operator".

In short, Bedini deliberately "switches"' and "triggers" certain kinds of vacuum exchange interactions, effectively creating a *true negative resistor* in the battery itself.[150] Due to the broken symmetry of the increased dipolarity *(overpotential)* that Bedini makes in "electron pile-ups" urging reluctant and delayed ion response, the vacuum furnishes extra virtual particle flux to this pile-up of electrons on the plate, which produces an enhanced Poynting energy flow that interacts with the ions in the battery electrolyte. Being charges, these ions thus transduce some of the excess absorbed virtual photon energy into real observable energy, thus increasing their potentialization and energy.

Bedini's method does the following: (i) It forms a true negative resistor in an unexpected way, upon the plates between a pile-up of electrons and the ions in solution in a common lead acid battery, (ii) it uses that negative resistor to extract excess energy from the vacuum and furnish it both to the ions forced into charging mode and simultaneously to the electrons in load powering mode, and (iii) it adds several other stimuli (such as Lenz law effects) which further amplify the negative resistor and enhance the effect, increasing the excess energy extracted from the vacuum and collected in the battery-charging process and also in the circuit-powering process simultaneously.

Specifically, the delay in ion response is adroitly allowed for and manipulated by Bedini to place the battery in ion current recharging mode while the signal pulse electrons between the plates and the external circuit are simultaneously placed in external circuit powering mode. By manipulating the hysteresis and adroitly timing the electron pulses and pulse widths, *Bedini breaks the usual forced Lorentz symmetry of the excitation discharge in a usually closed current loop containing both the*

---

[150] Again we stress that any dipolarity or potential is a negative resistor, producing giant negentropy {12}.

battery's source dipole and the external load. This is possible since his method deliberately opens the system so that vacuum energy enters freely, increasing the potentialization (energy collection) of the ions in the battery solution and upon the electrons between the plates and the external circuit as well.

We stress that Bedini has chosen to *avoid the* usual dissipation of half the energy collected in the external circuit to do nothing but kill the source dipolarity between the battery plates. He works on "that half of the circuit" that is usually just called the "back emf region" and ignored, and he separates and dephases that half of the circuit from the other half. By interrupting that normal "back emf battery-discharging section dynamics and converting it to "forward emf battery recharging section dynamics, while simultaneously powering the external load, Bedini temporarily produces and utilizes a negative resistor right there on the surface of the battery plates themselves.

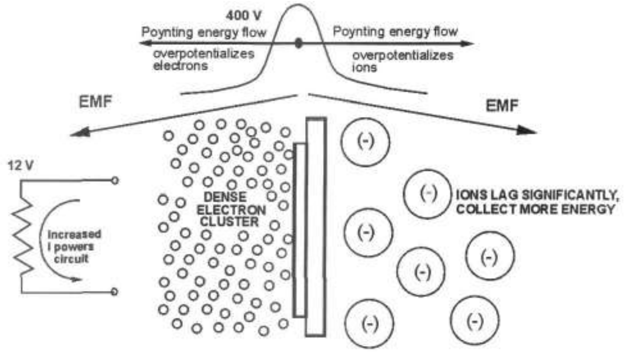

1 Sharp trailing edge of voltage pulse produces intense Lenz law effect Sharply increases electron clustering and potential (negative resistor) at interface Sluggish ions delay speedup further increasing the Lenz law effect

2 Increased bidirectional Poynting energy flow from potential overpotentializes even further both the electron duster and the ions, which are only beginning to accelerate in charge mode

4 Large new increased back emf into circuit returns overpotentialized electron current powering circuit with new excess energy pouring from negative resistor potential Continues during ion delay

5 Simultaneously, re-overpotentialized ion current is overcharging battery even faster

Figure 5-10 Phase II of the Bedini negative resistor process in a storage battery

5.8.5.7 Further Increase of the Negative Resistor Effect by Lenz's Law

See Figure 5-10. To further increase the Phase I effect just before it would end, and requiring precise timing of his switching of the pulse leading edge and trailing edge, Bedini then invokes a second phase by carefully controlling the timing for the sharp *cutoff of* the "stimulus pulse" creating the negative resistor. This is usually invoked just as Phase I is preparing to end, but experimentation and adjustment for optimization to the individual circuit conditions may be required.

Invoking the Lenz law reaction by minimizing the stimulus pulse cutoff time, Bedini sharply increases the already-increased negative resistor overpotential by a Lenz's law induced voltage surge, and sharply raises it to as much as 400 volts nominally. The process also sustains the negative resistor overpotential for a longer period, while increasing it again during this second phase. Thus even more free energy from the altered local vacuum potential is delivered to the ions in charging mode inside the battery, while simultaneously even more energy is delivered to the external circuit electrons in powering mode. The overpotential period is also extended. In this second phase, the extra energy W_{IONS} added to the ions in recharging mode is given by

$$W_{IONS} = f(400 - 12)(Q_{IONS}) \qquad [4\text{-}5]$$

and the extra energy W_E added to the electrons in system powering mode is given by

$$W_E = f(400-12)(Q_E) \qquad [4\text{-}6]$$

The total energy added to the system in Phase I by the negative resistor is

$$W_T = W_{IONS} + W_E = f(100-12)(Q_E + Q_{IONS}) \qquad [4\text{-}7]$$

The total energy added to the system in Phase II by the negative resistor is

$$W_T = W_{IONS} + W_E = f(400-12)(Q_E + Q_{IONS}) \qquad [4\text{-}8]$$

And so the total energy added to the system by the Bedini process is just the summation of equations [4-7] and [4-8]. With adroit switching, Bedini need only "pay" a small fraction of that freely received excess energy, in his own operator's input pulse energy and switching costs.

So by invoking a novel negative resistance effect directly upon the driving plate of the battery, Bedini creates (in the first phase) and then further enhances (in the second phase) a "nearly free" overpotential and overpotential period inside the battery. This excess potential directly upon the electron-pileup plate acts in both directions — out into the electrolyte between the plates to overpotentialize the ions in charging mode, and back out into the external circuit in powering mode to overpotentialize the electrons now powering the load. During a fraction of the operating cycle, Bedini recharges the battery while powering the circuit simultaneously, and thus has invented a novel method for extracting energy from the vacuum and curved spacetime to enable a COP>1.0 power system.

This is just a description of one fundamental period where Bedini applies his negative resistance process.

Several other places in the operation of the circuit lend themselves to additional phases of negative resistor formation and usage, and Bedini does use them. We do not discuss them here, since our purpose is only to advance the *fundamental principle involved.*

5.8.6 Independent Replication of the Bedini Negative Resistor Effect

Independent replication of the Bedini effect was accomplished early on by Nelson and by Cole, and also by Watson. Nelson has chosen not to further publicize his own work for some years, perhaps due to his employment as a microwave switching engineer in a large (and conservative) aerospace company. Cole is deceased, and Watson abruptly withdrew —forcibly, in our opinion — from public research and all further contact with his close colleagues after successfully demonstrating (in 1984) an 8-kilowatt power system utilizing a modification of the Bedini process. We are almost certain that Watson received the "offer he could not refuse", and that Watson-type COP>1.0 power systems are in fact already powering exotic underground facilities that "officially do not exist" but which we are aware of. We briefly discuss Watson's 8 kW generator at the end of this Chapter.

For non-battery powered systems, as we explain in Chapter 9, any unitary COP»1.0 EM system may experience a Dirac sea hole current decay mechanism which will produce an effective extra "load" in the system's input section, including on out into the external power system. As a general rule (sometimes violated), we estimate that the effect begins to be of importance at COP > 10^3 or so. Certainly it is of great importance in the range of COP > 10^5. To operate stably at COP » 1.0, this decay mechanism must be dealt with and prevented. For close-looping into self-powering mode, this negative energy feedback must be converted to positive energy feedback, thus allowing self-powering mode where all the input energy is freely received from the vacuum.

For battery powered systems, the hole current and "extra load in the input section" becomes of significance near COP = 2.0 or even less, because of interaction with the complex battery chemistry and the tendency to charge the battery with negative energy as well. So battery powered systems present a special problem for close-looping to achieve sustained self-operation. Bedini has long solved and demonstrated solutions to that problem in a variety of successful experimental systems.

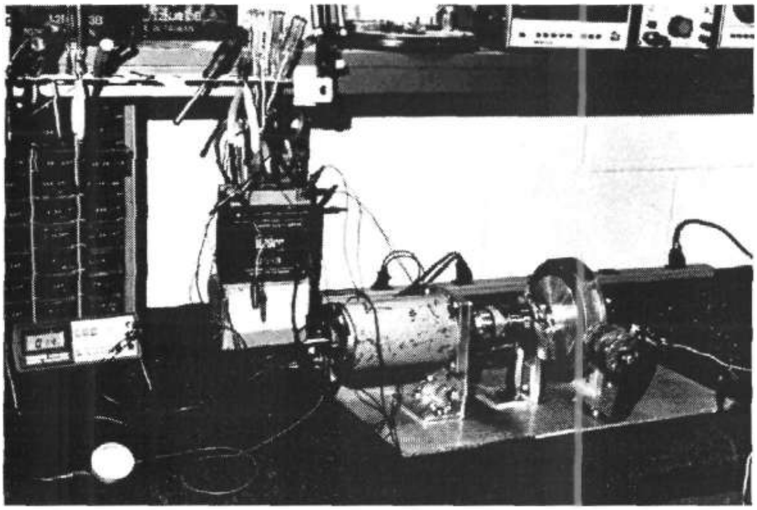

Figure 5-11 **Bedini motor** similar to the **type which ran continuously off a battery for three years, simultaneously recharging the battery.**

To utilize Bedini's battery-powered system at COP > 2.0, or to close-loop it for sustained operation, one must add another Bedini process that transforms negative energy current into positive energy current. That is Bedini's *transducing* negative resistor process {336}. We describe that process briefly in the next paragraph below.

Bedini has built many different motor-generator versions of systems self-powered by use of his negative resistor processes. Figure 5-11 shows a Bedini motor similar to one that once ran off a single battery for more than three years without stopping, using the forerunner of the Bedini back-popping process for creating a negative resistance in the battery. Bedini is still building and testing such power systems; one of his prototype motor-generators today, capable of operating and self-powering off a single battery, is shown in Figure 5-12.

5.8.7 Bedini's Transducing Negative Resistor.

Figure 5-13 gives a simplified block diagram of Bedini's transducing negative resistor process, which asymmetrically regauges negative EM energy into positive EM energy. With this process Bedini is able to intercept the Dirac sea hole current running from the output section of a COP>1.0 EM system to its input section, and divert the negative energy Dirac hole current to "reverse charge" a capacitor. The "reverse charge" consists of providing a hole current to the anti-negative plate of the

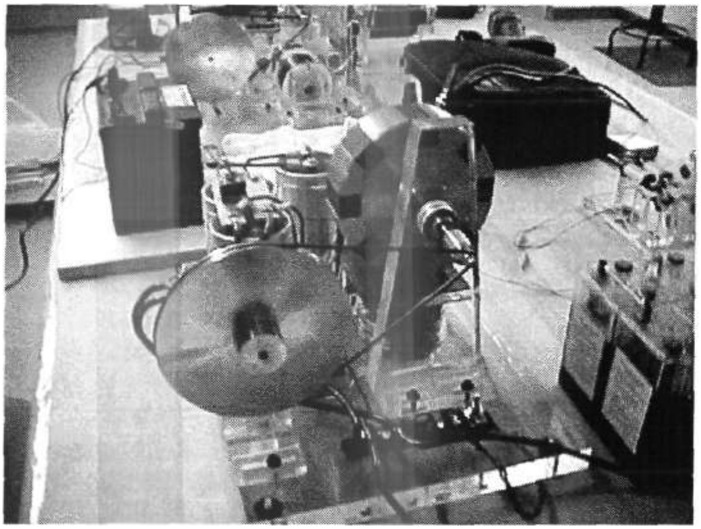

Figure 5-12 Modern Bedini motor prototype which operates off a single battery, powering the load and recharging the battery also.

capacitor, so that holes acting as positrons pile up in the local vacuum of **that** plate. This attracts the trapped electron charges in the dielectric, straining the dielectric in the direction of the hole collection. At the same time, the other plate draws electrons from the system that would otherwise be in currents that fight the back emf of the primary source dipole for the system. That reduces the back emf of the system during the charging of the capacitor; in effect, the back emf has participated in charging the capacitor with excess electrons on the negative plate.

Then Bedini reconnects the charged capacitor to the conductors leading to **the** power supply, discharging the capacitor by electrons furnished to the negative driving plate of the battery powering the system, and simultaneously to the input section of the system itself in normal powering mode. Note that the holes simultaneously discharge from the capacitor's anti-negative plate to the ground side of the battery, further charging the **battery** (the power source!) and the load as well.

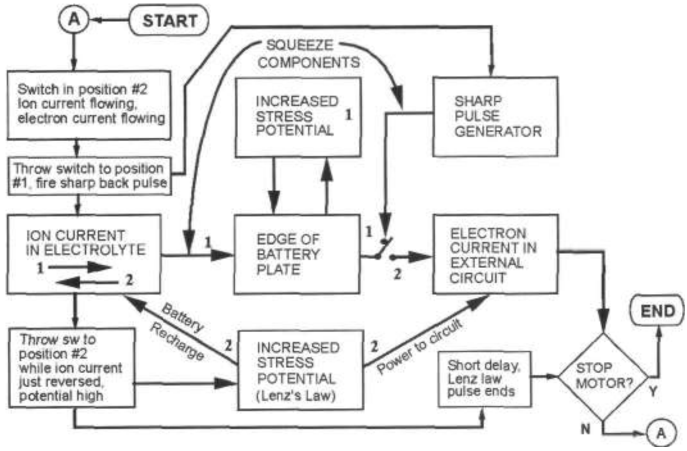
START
A
Switch in position #2
Ion current flowing,
electron current flowing
Throw switch to position
#1, fire sharp back pulse
SQUEEZE
COMPONENTS
INCREASED
STRESS
POTENTIAL
1
SHARP
PULSE
GENERATOR
ION CURRENT
IN ELECTROLYTE
1
2
1
EDGE OF
BATTERY
PLATE
1
2
ELECTRON
CURRENT IN
EXTERNAL
CIRCUIT
END
Battery
Recharge
Throw sw to
position #2
while ion current
just reversed,
potential high
2
INCREASED
STRESS
POTENTIAL
(Lenz's Law)
2
Power to circuit
Short delay,
Lenz law
pulse ends
STOP
MOTOR?
Y
N
A

Figure 5-13 Bedini transducing negative resistor process block diagram. Adjusting timing of rear edge of pulse adds additional Lenz law increase.

Though it does not appear in the conventional textbooks,[151] one can power a load directly by Dirac hole current, but one does so by using the hole current in exactly reverse fashion from how one would use electron current to power the load. Semiconductor components are also critical and usually should be avoided. In short, one simply reverses the terminal that is considered the "driving terminal", and feeds the hole current to that "anti-negative" terminal as now the temporary driving terminal.

The Bedini process thus allows the Dirac hole current (negative energy) produced by a COP>1.0 EM system to be fed back to the input section and transduced into positive energy input to power the system with electron current and also to recharge the battery with positive energy.

---

[151] Lattice hole currents of course do appear and are considered. However, even though the circuit designers consider a lattice hole as a positron, it is not such at all, because the m/q ratio of the ion with one + charge is quite different from the m/q ratio of a true positron. The Dirac hole in the vacuum is effectively the "4-positron" — still in negative energy Dirac 4-electron state — before its observation or interaction with mass. It is a true negative energy entity, and thus a negative mass entity. Prior to observation, it actually should be said to be a "negative energy x time" entity and a negative "masstime" entity. It is also a negative energy electron entity. This subtlety is essential if one wishes to produce successful antigravity experiments on the bench, as we point to in this book and as were exhibited by the Sweet vacuum triode amplifier at COP > $1.5x10^6$.

This converts what would otherwise be negative energy feedback to the input section of the system (and to the battery) to positive energy feedback **to** its input section and to the battery. Instead of seeing the negative energy current in the input as an additional "load" to be powered, the external power supply sees it as excess positive energy input by the environment to the system, reducing the "load" represented by the system. It thus reduces the power required by the system from the external power supply.

When the amount of "transduced" positive feedback electron current to the input section equals the "normal electron current draw" required by the system, the battery (or generator) can be disconnected and the unit will continue to operate in the self-powering mode, fed completely by the positive energy from the Bedini transduction process. The other *supersystem components* (comprising the external environment of the system) will continue to power the system and its loads and losses, with all the input energy to the system being taken from the combined active local vacuum and local curvature of spacetime.

5.8.8 Gray's "Splitting the Positive Pole" Engine.

One of the real pioneers in early application of causal positron energy was inventor Edwin V. Gray of Van Nuys California. In WW II Gray learned radar during his tour in the U.S. Navy, having previously attended advanced engineering school in the Army until discovered to be under age 15 and discharged. After the war, by 1958 he had learned to perform what he called "splitting the positive pole", using "cold electrical energy" (his name for negative EM energy) as well as the normal positive electrical energy. His first motor was operating by 1961. Preparing to put units into production, in 1974 the Los Angeles District Attorney confiscated all Gray's records and prototypes. To escape bogus charges, Gray later pleaded guilty to two minor infractions and was released. His prototypes were never returned. After several mishaps, moves, and more prototypes, Gray died in Nevada under mysterious circumstances in April 1989. We quote directly from Lindemann {341}:

> *"In the early 1980's, Gray offered the U.S. Government his technology to augment Reagan's SDI program. He actually wrote letters to every member of Congress, both Senators and Representatives, as well as the President, Vice-President, and every member of the Cabinet. Remarkably, ... Gray did not receive a single reply or even an acknowledgement!"*

In fact, during his early experiments, Gray discovered he could power light bulbs with cold energy, etc. Later he was to demonstrate this fact in a sensational manner: With a light bulb lit from his generator, he simply immersed the lamp — electrical leads and all — in water, where it remained lit. With his hands in the water, he would have been electrocuted or severely shocked if the electrical energy involved had *been positive* (diverging) energy. Instead, it was *negative* (converging) energy.[152] Another demonstration was to hit the coil of an electromagnet with some 3,000 volts, which hurled the electromagnetic into the air — but did not heat it. The electromagnet's coil would have been instantly melted if the energy had been positive, and the magnet would have been heated.

The reader should recall the timing of Dirac's theory of the electron {498a-b, 249}, which included the negative energy electron as early as 1930. Dirac also pointed out that we would observe this negative energy electron as a positive energy electron of opposite (positive) charge and positive mass, going in the opposite direction. The *observed* 3-positron and the *unobserved* negative energy 4-electron are quite different entities, producing quite different EM fields. The positron produces positive energy EM fields and positive gravity while the negative energy 4-electron produces negative energy EM fields and negative gravity.

In 1930, to his previous theory of the electron Dirac introduced the vacuum energy: the electron sea filled with electrons occupying negative energy states. The holes are treated as the positive electrons (as if observed). This vacuum energy medium is now referred to as the "Dirac Sea." First, Dirac attempted to identify these holes with protons, but that was falsified. Later the positron was experimentally discovered.

---

[152] By convention, "positive" energy has been associated with diverging EM energy, which is heating. Negative EM energy is associated with converging EM energy, which is cooling. In scalar interferometry {79}, one can control whether the distant EM field energy emerging in space in the distant interference zone shall be diverging (positive) EM field energy or converging (negative, cooling) EM field energy. This is determined simply by the biasing of the electrical grounds of the interferometer transmitters. The weather control referred to in 1997 by Secretary of Defense Cohen {89} is accomplished by using multiple scalar interferometers, some to gently heat the atmosphere in certain selected local areas (forming low pressure areas by expansion of the air) while cooling it in other selected local areas (forming high pressure areas by contraction of the air). By then moving these highs and lows around adroitly, one can capture and steer the giant jet streams, thereby steering the weather as one wishes, and directing significant weather changes into an area. By placing fairly sharp turns on the jet streams, one can induce great spin energy, thereby spawning tornadoes. Interestingly, by sharply pulsing the interferometer and controlling its electrical ground bias, distant cold explosions or hot explosions can be produced at will. Such weapons have been tested worldwide, and many test incidents are given in our older books *Fer-de-Lance* and *Gravitobiology.* Information on such incidents is available on website http://www.cheniere.org .

One of the problems with present conventional application of Dirac's theory is that the positron is not used in its *unobservable causal negative energy 4-electron state* by modern theorists and experimentalists, but *after observation* where its direction has been reversed, its charge has been reversed, time has been stripped away, and the mass has become positive. So it is actually used as a positive energy entity having positive mass, and producing positive energy EM fields and positive gravity. This defeats the entire magic of the causal (unobserved) 4-positron as the electron carrier of negative mass and negative energy {*convergent* electromagnetic energy) in currents in the vacuum (spacetime) itself, rather than the electron carrier of positive mass and positive energy in ordinary 3-space. By replacing the unobserved (causal) vacuum 4-positron (negative energy electron) with the observed (effect) 3-positron with positive mass and positive energy, the entire vast area of negative energy EM fields — and thus direct antigravity via electromagnetic means — is discarded.[153] But the physicists avoided having to grapple with practical negative mass and practical negative energy (cool mass and cool energy). It also set back the pace of physics in discovering and engineering antigravity transportation for nearly three quarters of a century.

It appears that Gray's thinking about cold energy was more prosaic than the somewhat esoteric considerations of the Dirac Sea theory. Conventional electrical engineering circuit theory considers positive charges moving around a circuit (although some electronics technical circuits required thinking in terms of electron current). Hence the "high" side of the standard circuit was the positive polarity, or the positive pole of the terminals of a generator or battery. From this "positive pole", positive current was sent around the external circuit to the ground return line by the forward emf (forward voltage). However, forcibly returning to the positive pole from the ground return line (i.e., from the negative pole) were those same positive charges, being forced back up by the back emf and requiring expenditure of half the external circuit's collected Poynting energy.

---

[153] When the huge nondiverged Heaviside energy flow component is appreciably of "convergent" or of negative EM energy rather than positive EM energy, the antigravitational aspects are substantive and observable (and practical). The unaccounted Heaviside negative energy flow component is often a trillion or more times as great as the accounted Poynting negative energy flow component. Hence antigravity technology is possible and practical, as we present in a later chapter.

So to Gray, the positive pole had two currents, one useful and one usually detrimental. The positive pole had (i) outgoing positive emf current powering the external circuit's loads and losses, doing the beneficial work (dissipating energy) in the external circuit with its load, and (ii) incoming positive current being forced back up against the back emf, doing the harmful work against the back emf and using up half the energy previously collected out there in the external circuit. Gray was aware that the work against the back emf was equal to the work done by the forward emf. So he set out to split these two currents apart, and catch the incoming current against the back emf separately, then use that normally wasted or oppositional energy to further power loads and/or recharge the batteries powering the system. Something like this was the way he was thinking.

That is what Gray called "splitting the positive pole". He also reasoned (and found) that the "reverse current" must be cooling instead of heating, since it moves against the voltage and acts backwards from regular current dissipated in resistors, etc. This precisely matches the cold (negative) energy phenomena met in the Sweet vacuum triode amplifier and in the motionless electromagnetic generator, as well as other COP>1.0 electrical systems.

Gray's later demonstrations showed that he had recognized this process or something very similar to it, in his "splitting the positive pole".

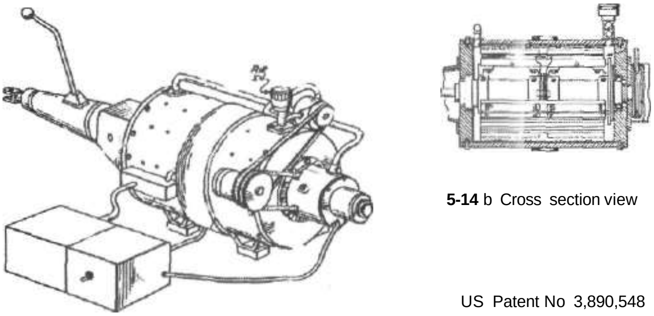

5-14 b Cross section view

US Patent No 3,890,548
issued 17 June 1975

5-14 a The Gray engine

Figure 5-14 Gray's "splitting the positive" COP>1 0 engine.

To take advantage of his concept and principles, Gray developed a special engine shown in Figure 5-14. In 1975 Gray received a patent on this engine {337} along with his special means of battery powering and special switching circuits.

Gray's engine uses pulsed capacitive discharge to power a magnetic motor, with recharging of the powering capacitors taking place *between* successive discharge positions of rotor and stator. The motor runs by initiating repulsion rather than attraction, which — apparently without Gray recognizing it in such terms — invokes the notion of the "superpole" pioneered by Bedini. The superpole is an important concept in COP>1.0 research, though most researchers in the field have never heard of it. John Bedini conceived it several decades ago and demonstrated the surprising difference between a superpole and a common pole of double strength.

When two equal and like magnetic poles are facing each other in repulsion with a small gap between them, a net stress magnetic potential (pole) is formed in that gap which has an energy density some four times as strong as the normal field energy density in the field from one of the two opposing like poles. Hence the local energy density of the vacuum in the gap has been substantially increased above the energy density in the field region between a north and a south pole facing each other. Further, this energy density is in a "field-free" Lorentz-regauged form, i.e., a stress potential form. From that superpotential *outward,* there exists a normal magnetic field. But in it, there is no net magnetic field, but there is a "stress" field nonetheless.

Compared to an ordinary magnetic motor using attraction as well as some repulsion, the exclusive use by Gray of the Bedini "superpole" principle caused an extra *symmetrical* regauging[154] of the local vacuum energy input to the system (to the gap between the poles) as the magnetic superpole

---

[154] Every conventional electrodynamics text already assumes and uses the Lorenz-**Lorentz** symmetrical regauging as being absolutely free. Symmetrical regauging assumes that the energy of the system has been freely changed twice, but in such a way that the two free fields formed are equal and opposite, forming a stress potential. Gray found how to split apart and separately use these two "free-flowing exc**hanges** of energy between system and active vacuum" that comprise the overall stress potential of the system, and thereby violated the Lorenz-Lorentz symmetry condition, freeing the two equal and opposite EM fields and their free energy to be used individually and beneficially to (i) provide extra power to the load and (ii) recharge the powering batteries so they did not run down. Using a different method. **Bedini** has done likewise.

(stress magnetic scalar potential). However, this stress potential is comprised of bidirectional longitudinal wavepairs, as shown by Whittaker in 1903. This superpole or stress potential thus can and does "split" into two different fields in antiparallel directions from each other. Since the north pole is deemed "positive" in magnetics, then using the two anti-fields comprising the free positive superpole is indeed "splitting the positive pole". Gray coined a very apt phrase for what he was doing.

Each of those antiparallel fields also involves an associated and unaccounted phase conjugate field {338a-c}, hence it is a time-reversed field (after observation). So opposing time-polarized EM waves also accompany Whittaker's decomposition waves. In short, the scalar potential accomplishes *optical-type pumping* in the time domain, and can be used to demonstrate time-reversal (as in Becker's bone-healing method).

Before observation (interaction with charged mass) and while still in the causal state, *the EM field from the positive pole is actually an EM field from a negative energy 4-electron, and the field can be taken to be a negative energy field!* In our view, positive energy enters every point dipole in the polarized vacuum at the negative charge, thence to the positive charge, thence back to the time domain. From the positive charge, the positive energy is leaving 3-space back to the time domain. Hence positive EM energy diverges from the negative charge and converges on the positive charge, then "disappears", so to speak, back to the time domain from whence it came. The "disappearance" of positive (divergent) energy is the disappearance of heat energy; hence it is a cooling process rather than a heating process. Or said another way, *negative* EM energy enters 3-space at each point dipole, entering from the time domain to the positive charge. Thence it moves to the negative charge, and back to the time domain. We see or detect the positive energy flow in 3-space directly, while the negative energy flow in the Dirac vacuum is evidenced by the "backwards" interaction creating Newton's third law reaction force.

By splitting the positive superpole, one has effectively destroyed the Lorentz symmetry condition. Hence the excess energy in the superpole will be capable of performing real work in the system in two directions. One direction is from the system to the external environment so that it produces the dissipation of energy (therefore work) in the loads and losses. The other direction produces a "negative resistor" or negentropic action that adds that much excess usable energy to the system back into its input section. By adroitly intercepting and switching that energy into capacitors and/or batteries, Gray was able to charge capacitors, recharge batteries, and even power loads with cool energy (as demonstrated later by immersing a lighted electrical line and a lit light bulb in water).

A positive-energy resistor becomes a negative resistor when fed with negative energy. Then it converges positive EM energy from the external environment and outputs it into the circuit in which the resistor is connected. Note that this is the same statement as the resistor receiving (extracting) negative energy from the circuit and emitting it back to the time-domain. *That's like extracting back emf from the circuit and tossing it off to the time domain.* A conductor to positive energy flow is an insulator or dielectric to negative energy flow. Hence the normally "conductive" water was a neat insulator for Gray's "dunking" demonstration of negative power, as were his own normally conductive hand and body. To negative energy, a coil seems to produce capacitance rather than inductance, while a capacitor seems to produce inductance rather than capacitance.

Thus when the "positive pole is **split"** in the system, the system has broken **its** Lorentz regauging and is an open system in disequilibrium with the external vacuum environment, freely receiving excess energy from it.[155] With adroit switching and energy capture, such a system is permitted to exhibit COP>1.0 because it has violated the free Lorentz regauging once the excess energy in the superpole region is discharged and used.

---

[155] When one analyzes the supersystem of any Lorentz-regauged EM system, one realizes that the extra stress potential energy — entering the system and "locked up" as a stress potential due to the Lorentz condition — is actually a special relativistic **rotation** of the frame of the power unit away from the laboratory frame. That is, the system when rotated away from the lab frame, takes on free symmetrical regauging energy in that rotated frame as an excess stress potential. Gray's "splitting the positive" process simply separated the two appositive fields comprising the stress potential, then produced excess free energy from that *broken* free regauging Lorentz condition. Literally Gray splits Lorentz's stress potential into halves, the halves being its two opposing EM fields and their energy. This is equivalent to the system being rotated back into the lab frame, releasing its excess energy it possessed in the rotated frame back into the lab frame. Richard "Scott" McKie has also filed a patent on an invention deliberately rotating a capacitor's frame and fully charging it — while it is in its rotated frame. The rotated capacitor can be charged with less energy furnished from the lab frame, than is obtained back out of the capacitor in the lab frame when the charged capacitor is suddenly rotated back and discharged. At least in **theory,** McKie's overunity process should be workable if the switching costs are negligible. Our contribution to McKie was to advance the theoretical mechanism for **the** experimental mechanism he had discovered and was using.

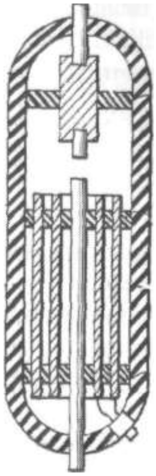

From US Patent No **4** 661747
issued 28 April 1987

Also in U S Patent No 4595,975
issued June 17,1986

Figure 5-1S Gray's conversion tube used with power systems,

Gray also utilized a special conversion discharge tube, shown in Figure 5-15. This tube uses the fact that a very sharp pulsed discharge in it results in a temporary COP»1.0 condition. This COP»1.0 condition instantly involves causal Dirac sea 4-positrons, moving as negative mass current and negative energy current in the local vacuum, which sweep back from the output side of the discharge to the input side.

We also point out the severe overpotential that must accompany any such violent pulse discharge process. The overpotential is a departure from equilibrium conditions, and is the difference between the actual potential and the equilibrium potential. The overpotential is well known and utilized extensively in electrochemistry {339a-c}, but is little used in conventional circuit theory. Nonetheless, it is involved in cold fusion processes due to the double layer phenomena {340a-b}, and it is involved in all electrode discharges (as from a cathode to an anode). It is also involved widely in electrochemistry and ion processes in solution.

Figure 5-16 shows Gray's circuit containing his conversion tube. Figure 5-17 shows Bedini's successful replication of the Gray conversion tube. Figures 5-18, 5-19, 5-20, and 5-21 show four pertinent pages from Bedini's laboratory notebook on his replication of the Gray conversion tube and the Gray motor system.

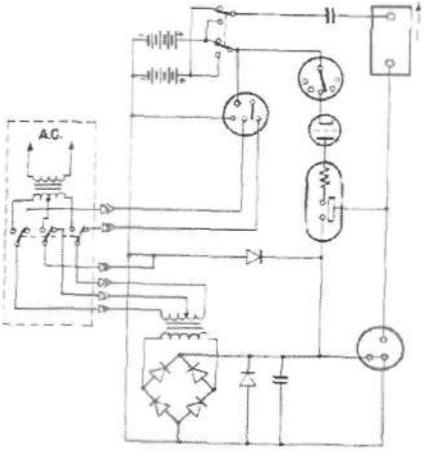

From U S Patent No 4 661 747
issued 28 April 1987

Also in US Patent No 4 595 975
issued June 17, 1986

Figure 5-16 Gray's circuit containing his conversion tube.

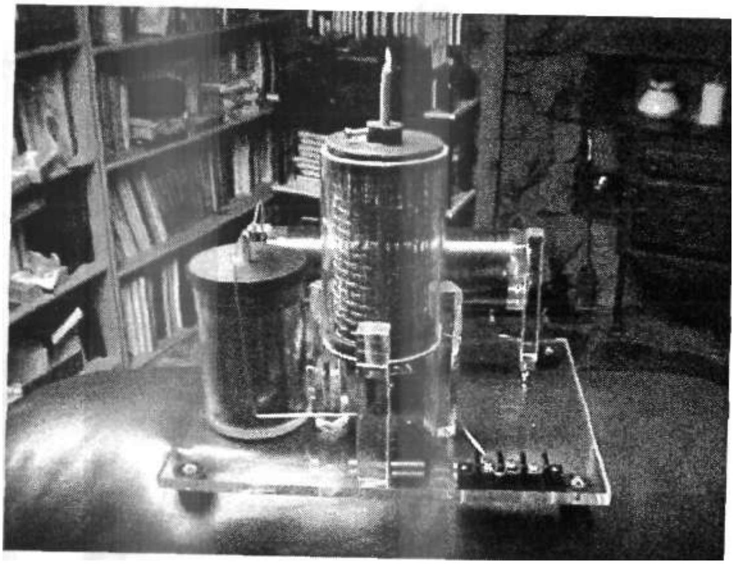

Figure 5-17 Bedini's successful replication of the Gray conversion tube.

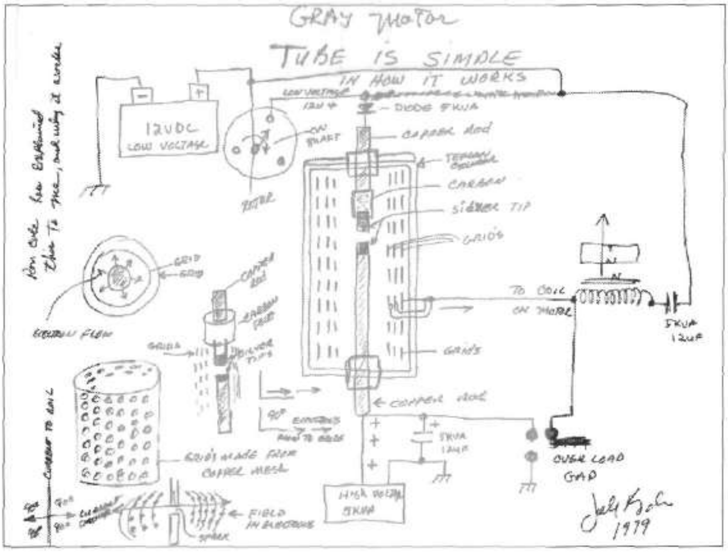
GRAY MOTOR
TUBE IS SIMPLE
IN HOW IT WORKS
12VDC
LOW VOLTAGE
LOW VOLTAGE
12V +
DIODE 5KVA
ON SHAFT
COPPER Rod
TEFLON
CARBON
SILVER TIP
GRIDS
ROTOR
GRID
GRID
ELECTRON FLOW
COPPER ROD
CARBON
GRIDS
SILVER TIPS
TO COIL
ON MOTOR
5KVA
12uF
GRIDS
COPPER ROD
90°
5KVA
12uF
GRIDS MADE FROM
COPPER MESH
OVER LOAD
GAP
HIGH VOLTAGE
5KVA
CURRENT TO COIL
FIELD
IN ELECTRONS
SPARK
1979

Figure 5-18 Page 1 from Bellini's lab notebook on the Gray conversion tube.

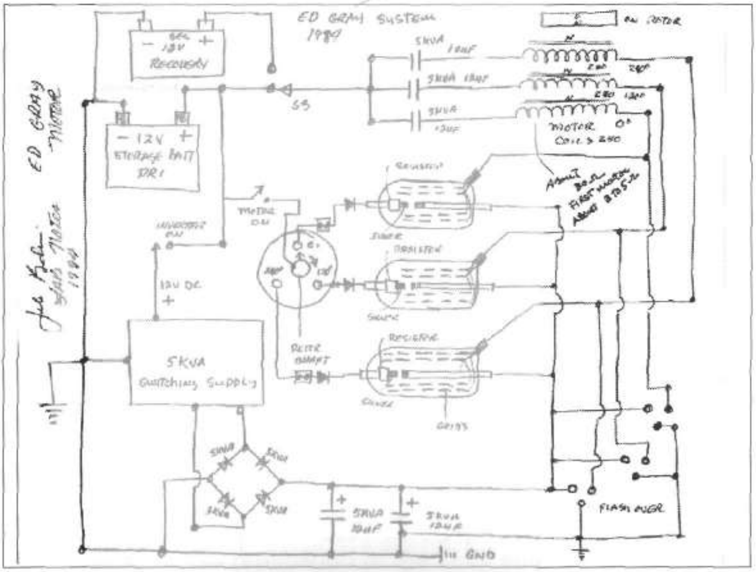
ED GRAY SYSTEM
1980
ED GRAY MOTOR
ON ROTOR
12V
RECOVERY
5KVA
12uF
5KVA 12uF
5KVA
12uF
S3
- 12V +
STORAGE BATT
DR1
MOTOR
COILS 250
0°
MOTOR ON
INVERTER ON
12V DC
+
5KVA
SWITCHING SUPPLY
SPARK
GRIDS
ROTOR SHAFT
FLASH OVER
5KVA
10uF
5KVA
12uF
GND

Figure 5-19 Page 2 from Bedlnl's lab notebook on the Gray conversion tube.

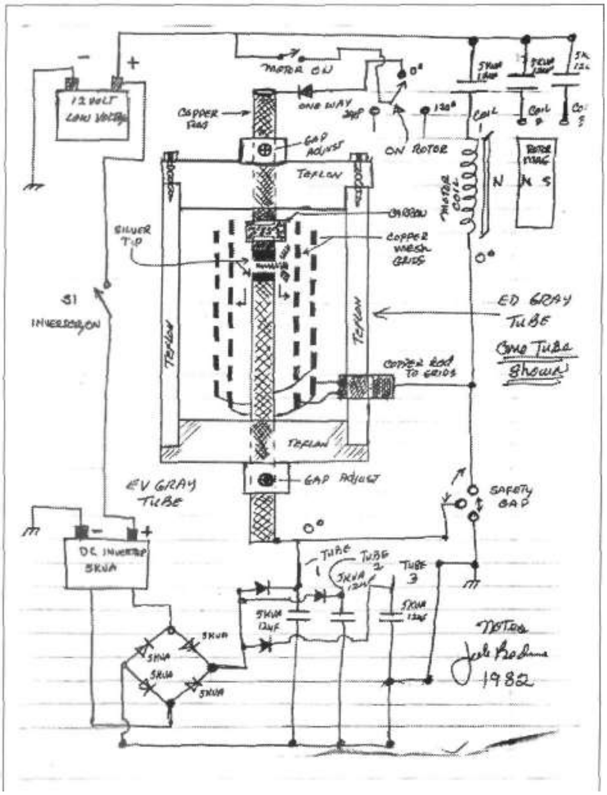
-
+
12 VOLT
LOW VOLTAGE
MOTOR ON
0°
ONE WAY
COPPER ROD
GAP ADJUST
120°
ON ROTOR
5KVA 12uF
COIL 1
COIL 2
COIL 3
ROTOR MAG
N
N S
MOTOR COIL
TEFLON
RIBBON
SILVER TIP
COPPER MESH GRIDS
0°
S1
INVERSION
ED GRAY TUBE
One Tube Shown
TEFLON
COPPER ROD TO GRIDS
TEFLON
GAP ADJUST
EV GRAY TUBE
SAFETY GAP
0°
-
+
DC INVERTER 5KVA
TUBE 1
TUBE 2
TUBE 3
5KVA 12uF
5KVA
Notes
Jack Bedini
1982

Figure 5-20 Page 3 from Bedini's lab notebook on the Gray conversion tube.

In science, when a novel experiment is replicated independently, that experiment then is established as a scientific fact. It then becomes the task of the theorists to either change the theory to accommodate the experimental results, or come up with a new theory to explain them. With Bedini's successful replication of Gray's work and extension of it, that work has now been scientifically established. With his "splitting the positive pole", Gray was referring to an actual, useful mechanism he evoked and used in the processes ongoing in his motor, in his conversion tube, and in his power system.

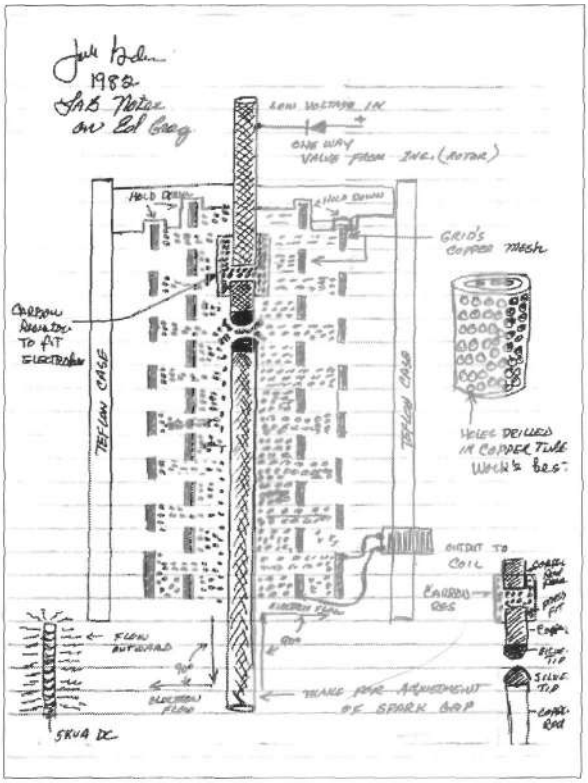
1982
LAB Notes
on Ed Gray
LOW VOLTAGE IN
+
ONE WAY
VALVE FROM INC. (ROTOR)
HOLD DOWN
HOLD DOWN
GRID'S
COPPER mesh
CARBON
RESISTOR
TO FIT
ELECTRODE
TEFLON CASE
TEFLON CASE
HOLES DRILLED
IN COPPER TUBE
Work's best
OUTPUT TO
COIL
CARBON
RES
ELECTRON FLOW
FLOW
OUTWARD
90°
90°
ELECTRON
FLOW
TRAKE FOR ADJUSTMENT
OF SPARK GAP
SKUA DC
COPPER
ROD
PRESS
FIT
SILVER
TIP
SILVER
TIP
COPPER
ROD

Figure 5-21 Page 4 from Bedini's lab notebook on the Gray conversion tube.

For further discussions about the Gray motor, conversion tube, and the history of it all, one is referred to the book by Lindemann {341}.

5.8.9 Watson's Self-Powering Generator

See Figure 5-22. In 1984, Jim Watson demonstrated a self-powering 8-kilowatt generator at the International Tesla Conference in Colorado Springs. This generator was a modified extension of smaller motors and generators built by Bedini over a period of years. At the conference, engineers from the audience were invited to the platform to perform their own independent measurements of the unit, while it was running and powering its load. The engineers directly confirmed that the system was delivering power to the load (as could be seen visually also), and that it was recharging its batteries as well.

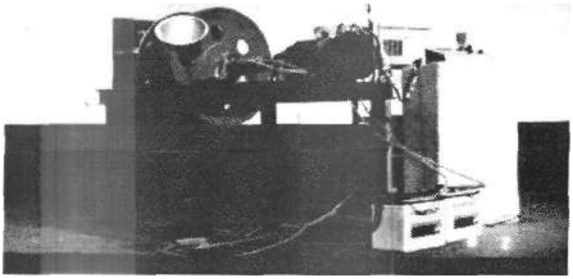

Figure 5-22 Watson's self-powered 8 kW generator demonstrated in 1984.

Interestingly, one of Watson's batteries was stolen that night. The reason was obvious. First, at the conference there was present one of the major persons charged with suppression of COP>1.0 electrical power systems world wide, along with several of his henchmen. Since they have been involved in suppressing many dozens of similar COP>1.0 systems, they are thoroughly familiar with both the theory of COP>1.0 electrical power systems and the various types of devices and mechanisms that have been invented in that area over the past century.

One feature of Watson's unit was that it charged the batteries with negative energy (i.e., with Dirac sea hole current, or "cold energy" as sometimes referred to), and could use negative energy for powering the unit. Hence, his batteries became charged with negative energy as the machine ran and powered its loads. After much demonstration and running, such a negatively charged battery has its local curved vacuum charged with an appreciable number of unfilled Dirac holes. Such a negatively charged "battery supersystem" will then absorb positive energy charging (electron current charging) for a protracted period before it will start to change its voltage and accept the positive charge. Instead, during that protracted initial period the battery will just "eat electron currents and positive energy" from the battery charger, with no apparent effect and no change in Its discharged state or its voltage.

A very simple way to absolutely ascertain whether or not a unit such as Watson's (without a Bedini negative energy converter) is genuine, is to **simply** take one of his batteries from a unit that has been running quite a bit, and recharge it on a conventional battery charger. If it exhibits that two-phased recharge phenomenon, it proves that the unit is real and is indeed a true COP>1.0 device, or certainly has the potential to be one. Coupled with the demonstration and independent measurements, this removed all doubts about the validity of Watson's COP>1.0 system.

So it appears that certain parties who were concerned about that system decided to do the "simple test" to see if Watson's device was real. It was real, and so they discovered. At that point, the High Cabal — Winston Churchill's term — recognized they had another successful unit and inventor to suppress.

Bedini and I were in close contact with Watson over a period of years, as often as three or more times a week. Later, some time after that conference, abruptly Jim ceased all communication with his former close colleagues. For two days a rather crazy recording **was** on his phone, stating the Watsons had moved. Then that disappeared. Watson never again contacted either Bedini or me, or his own financial backer (the late R. J. Reynolds III). Even Reynolds could not find Watson's whereabouts at that time.

We are quite certain that Watson received the "offer he could not refuse". Simply put, this is the old Mafia trick of visiting the target, offering him a deal, and he has no choice but to accept or die. When this is used to suppress COP>1.0 systems, and some medical systems, the offer is usually: *"Accept the money and you and your family will live very comfortably, but you will never openly work on this again or communicate it. You either accept our generous offer, or you and your family will be dead when we leave here."* Assuming that happened Watson had no choice but to accept the several million dollars offered him.

He is still alive, because Jeane Manning physically ran into him at a conference — and recognized him and spoke to him — whereupon he withdrew abruptly and left the conference very hastily.

5.9 Suppression of Free Energy Systems Has Been Common

This is only one of many cases we know of where such suppression of overunity systems and inventors has been accomplished. Others are not so benign. Marinov was killed with a longitudinal EM wave shooter in Europe, and his body was thrown off a building to make it appear a suicide. The police allowed the body to lie on the pavement for a protracted time (it was emitting longitudinal EM waves from the time-charging action of the strike). When finally moved, the pavement glowed in that area that had been underneath the body. Only one weapon on earth will kill a person in that manner, and that is a longitudinal EM wave "shooter".

Another prominent researcher was killed with a Venus technique (look it up, it's a distortion of the wave front) modified beam that plunges the heart into violent fibrillations. He rushed from a restaurant exclaiming, "They're killing me!" and died of a resulting heart attack. The present author was hit with just such a weapon in a restaurant here in Huntsville. A friend and I actually saw the weapon and the person wielding it. Knowing what it was and recognizing the symptoms, my seated companion and I were able to bolt out of the nearby back door during that minute or so we could still physically function. I personally saw the device and the person doing the beaming, and I have a witness to the event.

The partner of a colleague in Australia was killed on an upper floor by such a shooter from the street below, right through the walls. The assassin was observed, and was seen putting the bazooka-sized shooter back in his car, then speeding away. There are many other true incidents we could also relate, including a person killed by an ice-dart dipped in curare (a very professional hit method, used by some of the sinister arms of various intelligence systems). So there is more than just scientific work involved in pursuing COP>1.0 EM power systems. There can also be a direct threat to one's life, essentially at any time.

Almost any serious overunity researcher, who succeeds, will have had confrontations with assassination attempts, suppression attempts, threats, deliberate legal entanglements, etc. The present author is no exception. It is sad that one of the prerequisites for serious scientific work in this area is o**ften** a legal gun permit (for both the inventor and his wife) and either legally carrying a hidden weapon or never being very far from a weapon when working at home. As an old soldier, of course, one just shrugs and accepts it, and gets on with the job to the best of one's ability, albeit remaining very watchful and always on guard.

5.10 In Conclusion

In this chapter we have introduced many of the approaches to COP>1.0 EM systems or processes, or the possibility for such in some novel mechanisms known in science. Recalling our previously presented solution to the source charge problem, we emphasize to the reader that all assumed 3 - space EM processes are actually 4-space processes prior to observation, continually observed in iteratively frozen 3-snapshots by iteratively invoking the *d/dt* operator that produces observation. Hence from the source charge (and source dipole) solution we realize that all EM energy in 3-space comes from the time-domain and returns to the time domain. It iteratively appears along the "line of propagation" in 3-space at each successive position, but does not continuously propagate in 3-space. It does continuously propagate in 4-space. Since all the EM energy in any EM circuit comes from the time domain via the broken symmetry of the source charges and source dipoles, then we need to apply this greater vision to the COP>1.0 mechanisms and approaches so that we see the 4-space to 3-space to 4-space interplays.

The notion of the negative resistor is particularly significant and direct in that respect, and we have presented several versions of true negative resistors. For those wishing to experiment, we would suggest the easiest thing to start with is the point contact transistor. Or, simply go to Naudin's website {313} and download his instructions for making a Chung-type negative resistor, build it, and do experiments with it by varying the pressure on the fibers etc.

Again, we have not presented "parts kits" with instructions, but we have presented things that have worked or can reasonably be expected to work with some effort. We have also discussed the principles involved, to the best of our ability. In a few cases, we have directed the reader to information on websites where building instructions for selected negative resistors and other devices are posted.

As in any other scientific endeavor, we also urge the interested reader to read the literature and check the references cited, or as many as possible. Increasing one's understanding of the area is necessary to increase one's probability of success in the area.

We also wish the reader good results and success in his or her research and experimenting. The energy crisis is not going to go away, and the demand for oil will double in the lifetime of many of our middle-aged or younger readers. The cheap oil peaks shortly and declines in availability forever thereafter {342}; this alone will drive up the price of oil and seriously impact the economies of the nations of the world. To keep from destroying the world economy and the biosphere, it is absolutely essential that COP>1.0 electrical power systems — including self-powering systems taking their input energy from the seething local vacuum — be developed to produce the necessary electrical energy needs of the world, cheaply, cleanly, and easily.

If one or more of the readers should succeed in developing and producing such systems, then the entire purpose of this book will have been fulfilled.

Chapter 6

Additional Approaches to Overunity Power Systems

> *"Can the quantum potential carry a signal? If it can, we will be led to a violation of the principles of Einstein's theory of relativity, because the instantaneous interaction implied by the quantum potential will lead to the possibility of a signal that is faster than light... a signal has in general to be a complex structure, consisting of many events that are ordered in definite ways. In terms of our language, each signal is a supersystem of events, while each event is in turn a system of subevents.... At present, the answer is, of course, not known. "* [Bohm and Hiley] {343}. A comment:[156]

---

[156] We comment that the "event" involves observation: i.e.. it is conceived as an *observed* event. In that exact sense, no signal in space carries "events" because there is no interaction with mass and no *d/dt* operation induced on entities (curvatures) of spacetime. The effect (observed) 3-space field cannot propagate faster than light, because it cannot even propagate in the first place! The totality of light's EM interactions with matter is what makes the observed field or observed signal to appear to be propagated at light speed. It is not actually "propagating", but being iteratively recreated from point to point in space *if observed at every point.* Signal in spacetime is causal and hence nonobservable; the interaction of causal signal with charged mass is observable and makes an "event". When we clearly differentiate the nonobserved (causal) 4-field or dynamic from the iteratively observed 3-field or dynamic, then the *nonobserved 4-field* or dynamic is not limited to light speed. Propagation along the time axis, e.g., can be at "infinite velocity" because a single point in time is connected to every point in the universe simultaneously. Since all EM energy in space comes from the time domain locally at each point dipole in 3 - space and returns to the time domain from that same point dipole, it is the notion of "propagation continuously through 3-space" at *any* velocity that is a non sequitur. Absolutely nothing propagates through 3-space! If general relativity is re-interpreted **to account** the difference between observed and unobserved (effect and cause), then the unobserved cause can propagate superluminally without violation of the "observed event (effect) propagation at light speed" of general relativity. It is the notion that anything propagates in 3-space that is the non sequitur, and it should be removed from special and general relativity as well as electrodynamics.

> *"...contrary to the conclusions of classical mechanics, there exist effects of potentials on charged particles, even in the region where all the fields (and therefore the forces on the particles) vanish."* [Aharonov and Bohm] {344}.

> *"...for any force varying as the inverse square of the distance, the potential of such force satisfies both Laplace's equation and also the wave equation, and can be analyzed into simple plane waves propagating with constant velocity. The sum of the waves, however, does not vary with time, i.e., they are standing waves. Therefore the force potential can be defined in terms of both standing waves, i.e., by a global, or non-local solution, and by propagating waves, i.e., by a local solution changing in time... Thus, Whittaker's mathematical statement related the inverse square law of force to the force potential defined in terms of both standing wave (i.e., global) and propagating wave (i.e., local) solutions. The analysis also showed that the electromagnetic force fields could be defined in terms of the derivatives of two scalar potentials."* [Barrett] {345}.

> *"What might appear to be empty space is, therefore, **a** seething ferment of virtual particles. A vacuum is not inert and featureless, but alive with throbbing en**ergy** and vitality. A 'real' particle such as an electron must always be viewed against this background of frenetic activity. When an electron moves through space, it it actually swimming in a sea of ghost particles of all varieties — virtual leptons, quarks, and messengers, entangled in a complex melee. The presence of the electron will distort this irreducible vacuum activity, and the distortion in turn reacts back on the electron. Even at rest, an electron is not at rest: it is being continually assaulted by all manner of other particles from the vacuum."* [Davies] {346}.

6.1 Introduction

In this chapter, we continue to present a selection of potential overunity systems, inventions, approaches, and selected mechanisms. We also make the reader aware **that quantum** mechanics, an entire branch of physics, requires nonlocality and "instant propagation" in space. We utilize the quantum potential from Bohm's hidden variable theory of quantum mechanics {347} to impress the nonlocality. Further, the staid old scalar potential — which we meet as voltage in a circuit — also exhibits nonlocality and instant propagation in the Coulomb gauge {348}. Our purpose is to show that electrodynamics is not limited to the cut-and-dried subject as usually taught to electrical engineers and applied by them, but also covers a wide range of phenomenology that electrical engineers usually do not consider or account.

6.2 Energy Conversion and Energy Amplification

6.2.1 The Mead-Nachamkin Zero Point Energy Converter

It is well known that the vacuum interacts with electrical charges and dipoles, and also with the electrons in an atom. For example, Lamb received the 1955 Nobel Prize in physics jointly with Polykarp Kush for experiments measuring the small displacement later called the "Lamb shift" of 0+ne of the energy levels in atomic hydrogen {349}.

Casimir {350a} proved that closely separated conducting plates also alter the energy density of the vacuum, thereby becoming an "energy extracting" method, although the energy is miniscule. It is quite real, however, as shown in meticulous experimental work by Lamoreaux {351}. The principle that vacuum energy interacts with matter, plates, and electrical double layers and circuits is well proven experimentally.

In 1996 Mead and Nachamkin were granted a patent on an overunity EM power system process for extracting zero-point energy of the vacuum {352}. If one closely examines the patent wording, the device is properly patented as an energy converter and does not overtly state that it is a COP>1.0 system. It is a COP>1.0 system, however, since the input energy is freely received from the vacuum and not input by the operator himself. One embodiment of the invention inserts a tiny coil between the plates to extract the energy. See Figure 6-1. The energy is based on the Casimir effect {26}, which was accurately measured in 1997 by Lamoreaux {27} and since then by others.[157] Further investigation of the force's behavior with the shape of the conducting plates has been performed by Chen *et*

---

[157] See also U. Mohideen and A. Roy, "Precision measurement of the Casimir force from 0.1 to 0.9 microns," *Phys. Rev. Lett.*, Vol. 81, January 1998, p. 4549.

al.[158] Numerous other papers of interest on the Casimir effect have also been published {353a-d}.

Because energy cannot be created or destroyed, any COP>1.0 system is an *energy converter,* a priori. It must receive the excess energy from the environment (in this case, from the curvature of spacetime and from the local active vacuum) and convert that energy to a form usable to power or assist in powering the system loads and losses.

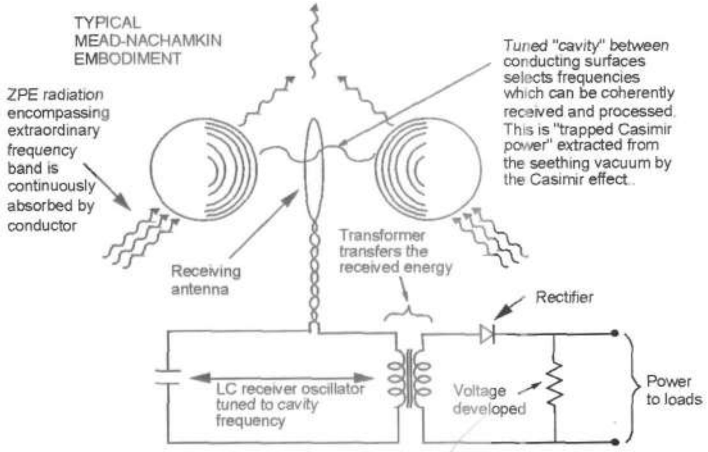
TYPICAL
MEAD-NACHAMKIN
EMBODIMENT
ZPE radiation encompassing extraordinary frequency band is continuously absorbed by conductor
Tuned "cavity" between conducting surfaces selects frequencies which can be coherently received and processed. This is "trapped Casimir power" extracted from the seething vacuum by the Casimir effect.
Receiving antenna
Transformer transfers the received energy
Rectifier
LC receiver oscillator tuned to cavity frequency
Voltage developed
Power to loads

Figure 6-1 Mead-Nachamkin conducting surfaces, antenna, and receiver for extracting EM energy from the vacuum,

The Mead-Nachamkin patent is interesting because the physics of the device is proven and accepted. Getting sufficient power output to be useful for electrical power purposes, however, remains to be demonstrated, since the Casimir effect is very tiny.

6.2.2 The Quantum Potential As an Energy Amplifier

In Bohm's hidden variable theory of quantum mechanics {354}, an extra potential is added to the Schrodinger equation, and is known as a *quantum potential.* Figure 6-2 lists the novel characteristics of this potential, and

---

[158] F. Chen, U. Mohideen, (G. L. Klimchitskaya, and V. M. Mostepanenko, "Demonstration of the Lateral Casimir h'orce," *Phys. Rev. Lett,* (in press as of the time of this writing).

Quantum Potential Characteristics

- No point source; not radiated
- Just appears instantly
- Quantum potential between two particles
  - Interaction does not vanish as spatial separation becomes very large
  - Instantaneous connection
- Depends on quantum state of system as a whole
- System parts can be greatly separated

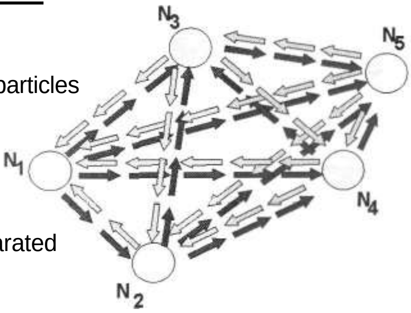

Figure 6-2 Quantum potential characteristics (table).

Figure 6-3 lists technical features of Bohm's hidden variable theory from whence the quantum potential is taken {355}. It can "move instantaneously", at "infinite velocity", which means it just suddenly appears everywhere it will be, and does not "propagate at a finite speed" through singly connected space at all {356}.

$$V_{quantum} = \frac{h}{2m}\frac{\nabla^2\sqrt{\rho}}{\sqrt{\rho}}$$

A quantum particle moves as if it were subject, in addition to its external potentials, to a potential which is a function of its own probability distribution.

Bohm's hidden variable theory assumes:
- **Particle and wave function real and separate**
- **Wave function obeys Shrodinger's equation**
- **Particle obeys classical mechanics**
- **Particle couples to wave function through a quantum potential**

Figure 6-3 Technical features of Bonn's hidden variable theory.

The potential is after all *potential energy.* With the quantum potential (QP), a joule of energy placed upon one collector, where that collector is connected via a QP with one or more distant collectors, will also result in a joule of energy (or some fraction thereof) instantly emerging upon and in each of those distant participating (QP-connected) stations simultaneously. It appears that, if only a fractional participation — say, 0.50 — exists with one or more of the distant stations, each of those stations will experience **the** appearance of that fraction (in this case, 0.5 joule) of the energy input to **the** initiating station. Any station can input energy, and it will appropriately appear at every other participating station, regardless of where these stations are located in ordinary 3-space as seen by the external observer.

One way of expressing this "instantaneous QP connection" between separated objects is to model the different objects (or stations) as linked together in a multiply connected space, with a certain QP connection or collection fraction. Presence of energy at one station in the multiply connected group, results in the simultaneous appearance of the energy — or an appropriate fraction of it — at every other station in the group.

To understand multiply connected space, consider a single point on a plane. Now consider 999 more points superposed on/in that same point. But suppose that those thousand points are really located at different locations in 3-space, though superposed in this "crazy space" containing a plane where all of them are superposed at one point.

As the ordinary external observer, suppose you input a joule of energy into one of these — to you — separated points, which happens to be the original point and is located in your laboratory. Supposed the fractional connection is 1.0 (the connection is perfect in multiply connected space). Voila! Instantaneously one joule of energy appears at each and every one of those (to you) widely separated points. If you vary your input energy (e.g., as in communication), it varies precisely the same in each of those distant QP-connected points.

Those who continue to advocate that superluminal communication of information is impossible are ignoring present experimental proof that it does occur. As a single example (there are others), it has been shown by Enders and Nimtz {357} that Mozart's 40th symphony can be transmitted as frequency modulated microwaves through a barrier wave guide between two points in a waveguide, via quantum tunneling, at 4.7 times the speed of light c. We point out that this is achieved using *transverse* EM waves.

If *longitudinal* EM waves are used together with longitudinal modulation (which engineers might think of as a sort of "pulse compression" technique), the only limitation to the speed of communication is given by the impurity of the longitudinal EM wave one is able to make. The Fogal semiconductor, e.g., can be used to transduce ordinary transverse EM wave signals into longitudinal EM wave modulations inside a scalar potential. Hence that scalar potential can move in the Coulomb gauge at infinite velocity. It requires another Fogal semiconductor circuit on the receiving end to receive and transduce the longitudinal EM waves and their longitudinal modulations back into transverse EM waves for processing and amplification by normal means. This capability of the Fogal semiconductor circuits has been demonstrated through satellite uplinks and downlinks coupled with long transmission line links, with essentially no delay from transmission to reception. One may think of the Fogal superluminal process as *tunneling through subspace.*

The best way to vary the energy is apparently by use of a variable stress potential (which translates to a longitudinal variation or modulation). In short, one must modulate the "internal energy" of the vacuum, so to speak, and not the "external energy". A reasonable discussion of internal energy versus external energy is given by Chen {358}. If one longitudinally modulates the internal energy of the vacuum potential — which, being a scalar potential, decomposes by the Whittaker decomposition {359} — then superluminal communication is not only possible but also practical. We expect Fogal to eventually place just such a system on the market, shortly after his semiconductors get into precision production.

One possible method of producing modulated internal stress potentials is by use of a Davis noninductive resistor {360}. Another is by use of a biwound coil with two windings, so that the two currents through two equal "superposed" coils are equal and opposite. We point out that, while there has been published much stuff and nonsense about such biwound coils, the truth is quite mundane. Simply examine the Lorentz symmetrical regauging applied to the Maxwell-Heaviside equations. It is the deliberate production of two equal and opposite force fields, which is in fact the creation of a stress potential and extra stress energy. By varying the magnitude of the Lorentz forces comprising the stress potential, the stress potential is varied in amplitude. Hence the concomitant longitudinal EM waves are varied in amplitude. A biwound coil is a variable Lorentz-regauging coil, *a priori.*

By oscillating the magnitude of any Lorentz regauging (which in the simplest case is simultaneously changing both opposing fields equally and oppositely), one can make longitudinal EM waves. By appropriately modulating both Lorentz regauging components, one can make longitudinal EM modulations upon those concomitant longitudinal EM wave carriers.

So with some careful tuning and adjustment, and a little nonlinear material in the core,[159] one should be able to have the superposed appositive magnetic fields lock together *(modulate,* rather than linearly *add)* and produce a magnetostatic scalar potential stress signal in and from a nonlinear core material. We caution, however, that the biological effects of the radiation of such stress potentials internal to spacetime have not been investigated. The experimenter is cautioned that he experiments at his own risk, and he is cautioned to experiment only at low power levels until the phenomenology in a particular case is understood.

---

[159] The opposing waves must modulate (multiply), not mix (add), and modulation is a nonlinear process. Linear mixing of opposite waves or fields will not cause the

If we compare the stress potential oscillation (wave) to a sound wave, the similarities suggest that the stress potential oscillation can be treated as a longitudinal EM wave, if the two superposed opposing EM waves "lock together". Simple addition is insufficient; modulation is necessary. Hence the nonlinearity of the conductors and core material is of significance.

For instantaneously propagating potentials and pure longitudinal EM waves, any conceived difference between the two becomes somewhat academic. The characteristic of interest is the instant appearance of the input energy at multiple points via a multiply connected space.

To the ordinary observer, use of a quantum potential and multiply connected space measurably yields a fantastic *energy amplifier.* One furnishes one joule, and gets out — say — 1,000 joules in those distant locations altogether. As shown in Figure 6-4, the QP system is a communication system of enormous implications, particularly as humans travel into space across the solar system and even beyond. Unfortunately, by using appreciable power in the initiating station, the QP system also leads to weapon systems of incredible destructive capabilities. Some five nations {361} now possess such quantum potential weapons, which are the dominant weapons on Earth. We do not further discuss the quantum potential weapons aspects in this book.

Back to our little energy exercise, inputting one joule and getting out some 1,000 joules. Suppose these 1,000 participating points are not distant at all, but are gathered closely together into a local material but still QP-connected with a connection fraction of 1.0. Now we input a joule of energy to that device, and it develops and outputs one kilojoule. It is hypothesized that this becomes the ultimate negative resistor. It will also exhibit separated Dirac hole current phenomena as the COP increases and, if sufficient COP is developed, it will exhibit antigravity effects such as

fields to "lock together" into a common stress potential. Nonlinear mixing (modulation) will. Hence the nonlinear core is required for conversion of appositive fields to stress potentials and **Lorentz** regauging.

those shown by the Sweet vacuum triode amplifier discussed in this chapter.

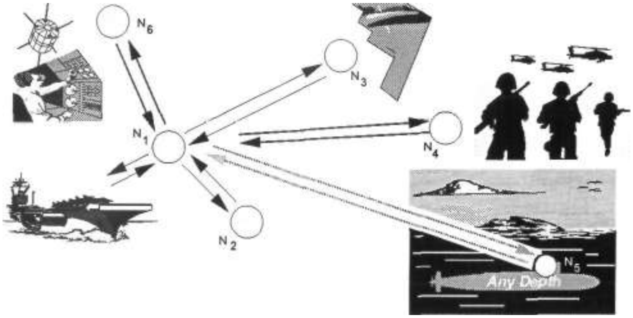

- Quantum potential decomposes per Whittaker 1903 into bidirectional longitudinal EM waves
- Can insert additional longitudinal waves, thus information using a Fogal semiconductor
- Can receive "signal" and discriminate the information using a Fogal semiconductor

Figure 6-4 Quantum potential communication system.

Strangely, overall conservation of energy is not violated by the quantum potential COP>1.0 energy system. It's just that the energy now is furnished from space (from the vacuum) in a *multiply connected* manner. However, all the excess energy is still furnished by the active external environment - i.e., from the time domain, according to the giant negentropy process {12} — though that environment now is quite novel and multiply connected. So we still have a system analogous to a windmill, though *a very novel windmill and a very novel wind indeed.*

Without belaboring the point, action-at-a-distance is absolutely required in quantum mechanics. Further, photon correlation experiments have long demonstrated experimentally that such action-at-a-distance effects do exist, just as predicted. This was a refutation to Einstein's objection to quantum mechanics' "spooky action at a distance".

We point out without further comment that, if a self-powered process is Invoked in the initiating station, then all other participating stations also contain that self-powered process. However, note that one does not really wish to "invoke" a quantum potential connection and use it except occasionally and very quickly! Otherwise, if a strong QP is turned on and **left** on among many distant stations, anything and everything happening to each one of them can instantly appear in all of them, so long as the "constant of connectivity" remains a constant or nearly so.[160] For that reason, strong quantum potential connections can be self-destroying and strong effects do not seem to normally appear in the physical world.[161] The proper energy use of a quantum potential is thus to use it only for fleeting "inductions" of effects in the desired stations, in *momentarily* multiply connected spacetime, at least until the self-limiting behavior is ascertained.

Apparently, a quantum potential can sometimes be formed in the limit of an iterative, mutually phase conjugating operation of EM signals passed back and forth between two phase conjugate mirrors, or among many of them. Such iterative mutual phase conjugate reflection of signals — particularly pumped signals — is technically known as iterative *self-targeting.* Sometimes we use a more picturesque term, *ping-pong,* for iterative self-targeting.

Conceivably, strongly coupled, strongly pumped optical self-targeting mirrors can — in the extreme limit — initiate or partially initiate a quantum potential. The superposing beams of energy may be thought of as first narrowing to laser-like beams, then gradually subsiding into pure longitudinal waves with transverse components "canceled" to form simply "stress energy". In short, if this reaction occurs, the energy is transformed into a stress energy potential, which acts either as a scalar potential in the Coulomb gauge or as Bohm's quantum potential. In our view, the pure longitudinal wave has infinite velocity, as pointed out by Rodrigues and Lu {93a}, so Lorentz-Fitzgerald contraction results in "zero-separation distance", which in turn results in a multiply-connected spacetime.

One points out that the "input energy" to the initiating station may be deliberately structured internally, including altering the internal Whittaker longitudinal EM wave functions comprising the input electrical field energy. In this manner, in theory deliberate spacetime curvature sets and dynamics can be introduced into the distant but multiply connected targets. A form of highly advanced electro-biological warfare induction of disease has been developed by at least one nation (Russia) in this regard, as an extension of Kaznacheyev's experiments showing induction of cellular disease at a distance. It appears that the cells of a living body are interconnected by a very weak quantum potential in such manner. Again, a great deal more research on that possibility is needed.

---

[160] Fortunately, it appears that this connectivity constant is not really constant, but reduces as the feedback of each station's fluctuations to the other stations increases. The result seems to be a "leveling off process reaching a plateau and a fixed amount of gain in the overall system. In short, the system usually seems to be "self-adjusting" and "self-limiting" as the overall gain increases. Nature is riddled with quantum potentials, but this self-limiting process keeps most so tiny that they are not even noticeable.

[161] However, the validity or invalidity of this statement for some of the highly energetic phenomena of astronomy remains to be determined. It also remains to be determined as far as the quantum potential's role in generating the zero point energy fluctuations of the active vacuum.

Given a quantum potential connection between separated objects including our initiating station, one can also add vacuum *engines* (energy structuring of curvatures of spacetime itself) as our input to the initiating node. Those vacuum engines will then simultaneously appear unabated (or only partially abated) in the multiply connected separated objects. In this way, in theory one can use the QP and QP energy to perform highly anomalous *momentary* engineering of nuclides, materials, fields, etc. in widely separated, distant objects and areas. By repeating the momentary engineering "pulses", distant objects and areas can be profoundly altered.

The use of vacuum engines in medical applications is of particular future interest.[162] Very radical extension of those techniques can be achieved once quantum potential technology is developed and available in the open scientific community.

But let us ignore the weapons, free our scientific imagination, and let it roam. As an example, in the future a vacuum engine (spacetime curvature engine) might be developed and utilized in a world-wide quantum potential to *momentarily* strike and eradicate all AIDS viruses on Earth, and do only that. Or, the engine could be slanted to simply *momentarily* slightly change those viruses into an innocuous kind of new virus. The cells of the individual human body are already connected, it appears, by a very weak quantum potential. By connecting an external initiating station to that body's quantum potential, the energy of the body could be directly scanned, all abnormalities noted, and tailored engines created and *momentarily* used to treat the body and correct its abnormalities and malconditions. This, we believe, may well be a medical therapy of the future, a few generations from now. [163]

---

[162] See Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian], Novosibirsk, 1981; Vlail Kaznacheyev *et al.,* "Distant intercellular interactions in a system of two tissue cultures," *Psychoenergetic Systems,* 1(3), Mar. 1976, p. 141-142; Vlail Kaznacheyev, "Apparent information transfer between two groups of cells," *Psychoenergetic Systems,* 1(1), Dec. 1974, p. 37. Unfortunately this QP induction of cellular disease has already been used by the former Soviet Union to develop a QP extension of the **force**-free structured potentials method first used to induce cellular disease and **health** changes in U.S. personnel in the former U.S. Embassy in Moscow for several decades. Now QP weapons using such techniques can target the entire population of a targeted nation. Using the internal Whittaker structuring of quantum potentials and EM fields and waves, *electromagnetic biological warfare* has already been born and highly developed. The actual dimensions of possible biological warfare strikes are far greater than has been mentioned in the open literature.

6.3 Other Systems of Electrical Interest

6.3.1 Russian Parametric Oscillator Power Systems

In leading physics institutes and laboratories in the 1930s, Russian scientists built COP>1.0 parametric oscillators, including some of appreciable size. The work is documented in the literature {362a-k}, both in Russia and in France. The devices were developed and tested in several Russian laboratories. With linear loads, the oscillators would progressively build to self-destruction. With nonlinear loads, the devices would stabilize and power themselves and their loads, at a level depending on conditions — reminiscent of the behavior shown in Figure 4-3 in Chapter 4 for stability COP>1.0 levels for the asymmetrical self-regauging process.

Standard references for parametric oscillation in EM circuits are available. However, with standard practices none of these conventional oscillators produces COP>1.0, primarily because the conventional closed current loop circuit is applied without special modifications. For overunity work, the

---

[163] Apparently, tests of such QP capabilities in weaponry have been gingerly conducted for direct induction of disease patterns in the population of a targeted nation or area. One symptom of such a test is that a very small number of widely isolated cases of the same disease breaks out simultaneously, in isolated cases scattered widely and puzzlingly throughout the population, *without the possibility of any direct carrier vector connection.* Apparently, the test inductions have deliberately been held mostly to the "shadow state" (or upper virtual state, just below observable state) in most of the targeted populace, to prevent giving away the tests. However, the human immune system and cellular regenerative system do respond to such "near-observable-state" signals and engines, when sustained or repeated in time, by a sort of coherent addition (integration) over time. The bell-shaped distribution curve predicts that a few individuals, randomly scattered throughout the populace, will have quite lower induction thresholds and will evidence the actual disease after a certain time, whereas in almost all the other members of that population the "disease" will never evidence in the observable state during that time. The test is deliberately terminated when only a few observable cases occur, which are sufficient to prove the efficacy of the test and the method.

Russian and French literature (as we referenced) should be meticulously studied and analyzed in a higher group symmetry electrodynamics, to see how the Lorentz condition is broken during discharge of circuit excitation energy in the load, and how the COP>1.0 Dirac hole current decay mechanism is avoided or contained.

The Russian COP>1.0 parametric oscillation work appears to have been suppressed by the Communist regime just prior to WW II. After the war, all such technology passed under the rigid control of the KGB (by whatever name it took and takes from time to time) and into the special weapons research and development area — still highly classified in Russia. We know of no similar work presently ongoing in the West, although at least an English translation of one of the major summary documents of the Russian parametric oscillation COP>1.0 power system developments was performed under NASA auspices {363}. However, in conventional U.S. parametric oscillation research, I know of no case where COP>1.0 operation is being sought, or where the Dirac sea hole current decay mechanism is even mentioned or recognized {364}.

6.3.2 Sweet Vacuum Triode Amplifier

During the 1980s and 1990s, the present author worked —sometimes closely for extended intervals — with inventor Floyd Sweet, the inventor of the Sweet vacuum triode amplifier (VTA). Indeed, we gave the unit its VTA name, at Sweet's request.

Fig. 6-5 (pg. 384) shows a diagrammatic illustration of the fundamental VTA construction. Two coils in quadrature are between two barium ferrite "brick" magnets, of the kind formerly used in many power audio systems. The vertical coil is the "input" signal coil, and the horizontal coil is the "output" power coil. The input consisted of a 33 microwatt 10 volt 60-Hertz signal, and the output was a 120-volt 60-Hertz 500-watt signal. The barium nuclei in the magnets were specially preconditioned by Sweet so that they were in powerful self-oscillation with the surrounding energetic vacuum. A double-edged razor blade or piece of shim stock placed on one of these magnets would oscillate back and forth, incessantly, without cessation — showing that the magnetic field itself was "waving" back and forth (Figure 6-6). It was also continually doing work against air **resistance,** by moving the air.

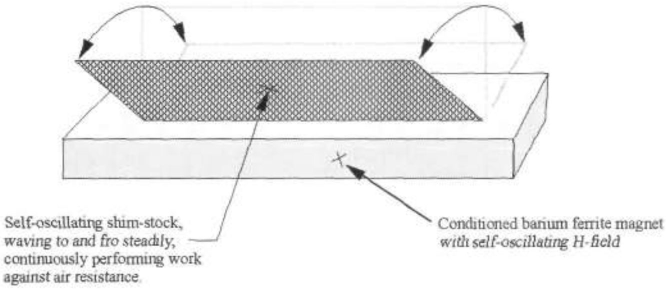
Self-oscillating shim-stock, waving to and fro steadily, continuously performing work against air resistance
Conditioned barium ferrite magnet with self-oscillating H-field

Figure 6-6 Blade waving continuously on Sweet's conditioned barium ferrite magnet.

Sweet had a Master's degree in electrical engineering from a leading university and handled the mathematical theory very well. He had worked at General Electric for many years, and Gabriel Kron was his patron and mentor. Sweet often spoke glowingly of Kron, and I came to believe that Sweet's VTA was probably an outgrowth of, or very similar to, Kron's negative resistor.

- Two activated magnets facing, with fields in self-oscillation.
- Barium ferrite magnets, barium nuclei in self-oscillation
- Barium nuclei self-pumped
- Two coils in quadrature, load is lamps, six watts
- Next unit produced 500 watts output with 330 microwatts input; later made 5 kW VTA

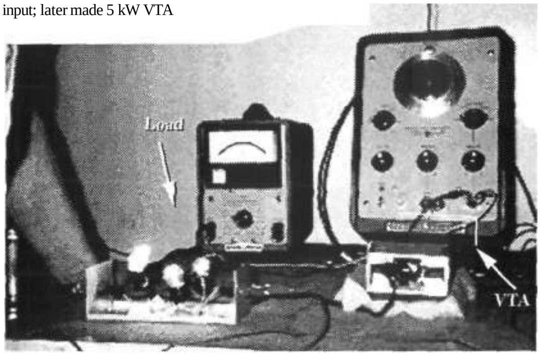

Figure 6-7 First Sweet VTA producing 6 watts output.

When I first met Sweet, his little VTA system was producing a 6-watt output, enough to light four 1.5 watts of auto lamps (Figure 6-7. But there was no question as to the genuineness of the device. Nothing was hidden, and Sweet allowed me to measure the device at will, disassemble it, play with it, and examine it in any fashion. It was genuine, and not a hoax or trick. I also locked up one of his specially conditioned magnets for 24 hours, with a piece of shim stock sitting on the flat of the magnet and waving to and fro continuously, steadily performing work by moving air. When I opened the lock the next day, the shim stock was still there on the magnet and oscillating, having continuously done work against the air resistance for 24 hours with absolutely no energy input by the operator. And it was still working. Indeed, that single "kinetic" permanent magnet destroys all objections to COP>1.0 EM systems, including those that are self-powering and thus have COP = ∞.

It seems little known that the vacuum around nuclei, in some cases, can be treated as a semiconductor, e.g. as discussed by Prange and Strance {365}. It is also known that nuclei do exhibit resonances at ELF frequencies. In particular, the vacuum in the region close to the nucleus of a superheavy element is known to act in a fashion analogous to the inversion layer in a field effect transistor. Prange and Strance introduce the idea of the inverted vacuum. Just as a semiconductor may be manipulated by subjecting it to external fields, doping etc., it appears that the vacuum can be similarly manipulated by appropriate means. We personally suspect that the semiconducting vacuum can be and is resonantly involved in any ELF resonances of the nucleus, which can occur in lighter nuclei such as barium.[164]

The virtual particle flux of vacuum, regarded as noise, may provide noise amplification of the coherent self-oscillation frequency between the semiconducting vacuum and the barium nucleus. Certainly, analogous noise amplification of signals is known in electrical physics {366}.

---

[164] Here our concept of the supersystem may be of utility. The "lock-in" or "freeze-framing" of an equilibrium condition for a system — such as in a state of nuclear ELF self-oscillation — is stabilized and made a "new equilibrium condition" when the force field reaction from the curved local spacetime into the system is equal and o**pposite** to the force field reaction from the local active vacuum into the system. We believe Sweet's undisclosed activation process was a method for synchronizing those two force field reactions and making them equal and opposite in one short discharge. We hope that this speculation is of use to future experimenters trying to duplicate Sweet's activation of his magnets into sustained and powerful supersystem self-oscillation.

I hypothesized that Sweet's activation process treated the vacuum surrounding the barium nucleus in such fashion, so that he was able to establish self-oscillation between the local activated vacuum and the concomitantly activated barium nucleus. Since barium ferrite is optically active, it may be that Sweet discovered how to get sufficient "self-pumping gain" for the self-oscillation to endure and not die away in a decaying oscillation manner. Since the "gain" of the second VTA as a self-pumping device was some 1,500,000 (Figure 6-8), obviously Sweet's activation method introduced a powerful state of self-oscillation.[165]

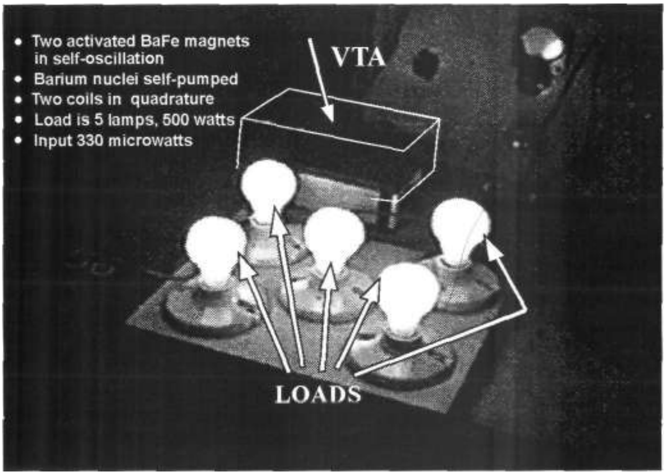

- Two activated BaFe magnets in self-oscillation
- Barium nuclei self-pumped
- Two coils in quadrature
- Load is 5 lamps, 500 watts
- Input 330 microwatts
VTA
LOADS

Figure 6-8 Second Sweet VTA producing 500 **watts and C0P=1,500,000.**

---

[165] Here we point out that "small" things with very close "double surfaces" of opposite charge or potential usually have very large fields. The tiny Lamb shift, e.g., has a local energy density greater than the surface energy of the sun. Jackson, *Classical Electrodynamics,* Second Edition, 1975, p. 10-11 points out that *"field strengths of the order of 10^9-10^{15} volts/cm exist at the orbits of electrons in atoms, while the electric field at the edge of a heavy nucleus is of the order of 10^{19} volts/cm."* It should be obvious to the reader that, if our speculation is true and nuclear oscillation of the barium nuclei is involved, the production of optical gains of 1.5 x 10^6 in Sweet's self-pumped, optically acting barium ferrite magnets is not surprising because of the extreme magnitude of the oscillating nuclear fields. In that case, optical gain converts directly to true power gain because the pumping energy is furnished freely by the environment and not by the experimenter.

6.3.2.1 Mapping the Magnetic Field of the Magnets

Figure 6-9 shows a small magnetic field mapping device built by test engineer Rosenthal for Sweet. Sweet "scanned" and mapped the consistency of the magnetic field from his candidate magnets (barium ferrite magnets bought from surplus stores at the time). If the consistency varied over 12 to 15%, the magnet was useless because it would not "hold" the activation and retain it. Magnets whose magnetic field variation did not exceed 10% were ideal. So Sweet only found about 1 in 10 or even 1 in 30 magnets that would retain the self-oscillation state when initiated.

Figure *6-9* Test engineer Rosenthal points to a magnetic field strength mapping device he built for Sweet.

Sweet may have actually used two appreciably higher frequencies in his activation process, with the difference frequency between them being a precise ELF frequency (60-hertz, 100-hertz, 400-hertz, etc. as he desired). That would make sense, when one realizes that the magnet is a highly nonlinear medium and then treats it (to first order) as an isotropic nonlinear medium. In that case, the two wave frequencies actually used would be subject individually to the normal overshoot, breakup, reconstitution, etc. inside the material of the magnet — but the difference frequency would behave as if it were a sine-wave frequency {367}.

This makes sense, since the output wave of Sweet's activated magnet was essentially a pure sine wave. Also, the presence of noise-enhanced heterodyning can sometimes be used to amplify a difference signal even further {368}. As we stated, since the activation is a self-oscillation between the local semiconducting vacuum and the barium nuclei, one wonders whether much of the rest of the vacuum virtual particle flux can be treated as noise, and whether an effect is obtained for noise amplification of the coherent self-oscillation frequency.

L'vov and Prozorova {369} point out some interesting characteristics of the formation of self-oscillating spin waves which occur above parametric excitation. These oscillations result when internal stability does not occur, and they evidence themselves as oscillations of magnetization. The frequencies of the oscillations usually lie in the range from tens of kilohertz to tens of megahertz. At small above-threshold ratios, the shape of the oscillations is nearly sinusoidal. At larger ratios, the shape differs appreciably from sinusoidal. At still larger ratios, the oscillations become chaotic.

If Sweet did use the "difference frequency" conditioning, one suspects he may thus have been able to utilize larger above-threshold ratios, where the two frequencies actually transmitted would have resulted in chaotic oscillations. It is interesting to speculate that, in that case, for a suitably chosen barium ferrite magnet, the difference frequency still can behave sinusoidally, but now much stronger. If so, that might account for the tremendous COP Sweet attained in his second unit, which exhibited a COP = 1.5×10^6, and could be pushed even higher.

6.3.2.2 Sweet VTA for Closed-Loop Self-Powering Operation

Courtesy of Walter Rosenthal, we now present a visual record of the Sweet closed-loop self-powering version of the VTA and its momentous successful acceptance test on May 1, 1990, as closely observed by Rosenthal.

Figure 6-10 shows Sweet and a different version of the VTA, with the coils in a slightly different configuration and the windings altered. This is the two-magnet, 2-coil VTA that he built and tested for self-powering operation. Figure 6-11 shows the unit mounted on a special test fixture built by engineer Rosenthal. This fixture and its components allowed the sensing of the sine wave output power at the zero-value crossover points. To switch in the special clamped feedback system for self-powering operation of the input, and simultaneously switching away from external power, the switching had to be done very accurately at that zero-point crossover. Feedback would not stabilize and hold if the switching were done at any other point in the output cycle.

Unit independently examined and testing observed by test engineer Walter Rosenthal

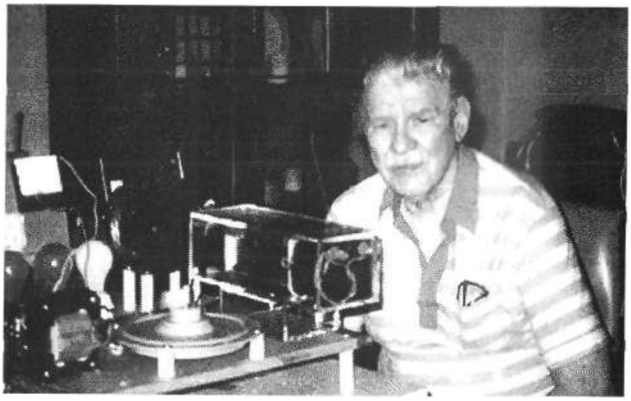

Figure 6-10 Sweet with 2-magnet 2 coil VTA, loads, and instruments for historic self-powering operational tests on May 1,1990.

Figure 6-12 shows the set-up for the tests, with the unit and loads as well as some of the instruments used. Figure 6-13 shows the closed-loop VTA powering itself and its loads, after being initially set into operation with outside input power by the operator, and then adroitly switched into self-powering by switching the input from external to internal at an output

Fixture allowed unit to be started into operation on external power then switched to self-powering internal power at the zero crossing of the *output* voltage The device in its fixture has a row of test points (disconnects) on every conductor going to the device

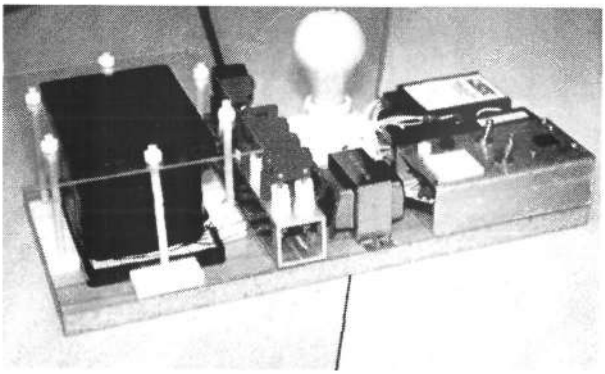

Figure 6-11 Two-magnet two-coil VTA mounted on a special test fixture built by Walter Rosenthal.

power zero crossover point. This picture shows the unit in self-powering and stable operation, taking all its input energy from the vacuum, including the power necessary to power its own internal losses and the power necessary to power its controls and loads.

Independently examined and testing observed by test engineer Walter Rosenthal

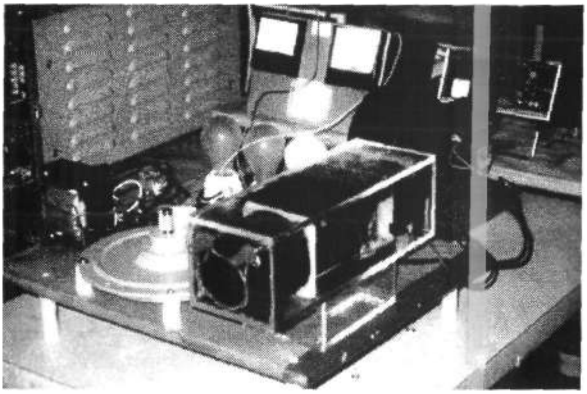

Figure 6-12 The 2-magnet, 2 coil VTA, loads, and instruments for its historic self-powering operational tests on May 1,1990.

Formal self-powering test on 1 May 1990 The loads and the VTA are powered by electrical energy extracted directly from the activ e vacuum Independently observed by professional test engineer Walter Rosenthal

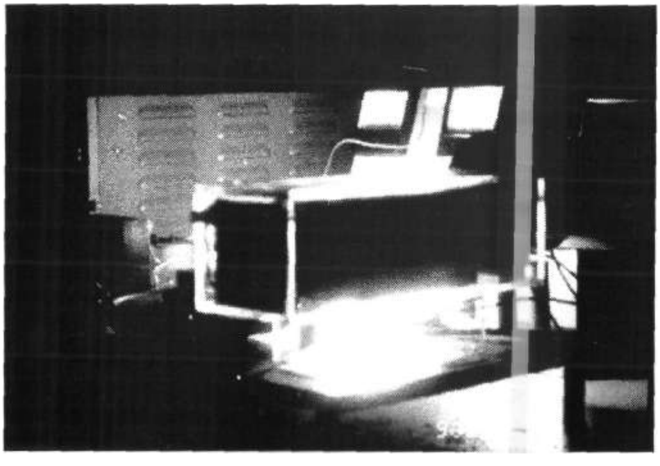

Figure 6-13 Historic self-powering operational test of the VTA May 1,1990 with no energy Input by the operator.

We strongly stress that this operation is permitted by the laws of physics and thermodynamics for an open EM system far from equilibrium in its exchange with its active environment. The active environment in this case is the combination of the active local vacuum and the active local curvatures of spacetime. The closed-loop condition is a new and stable equilibrium condition in the dynamics of the mutual exchanges between all three supersystem components: the system, the active vacuum, and the curvatures of spacetime. The *system* is in stable disequilibrium; the *supersystem* is in a stable new equilibrium condition.

6.3.2.3Discussion

We do know that the output current of the second Sweet device was comprised of a significant Dirac hole current[166] as well as electron current — showing the expected heavy involvement of the altered local vacuum (the Dirac sea) and the curved local spacetime. The negative energy majority of the output revealed itself readily. If the output leads were shorted, ice formed instantly on the shorted leads from the moisture in the air. That is, an electrical short resulted in the sudden production of cold (negative) EM energy rather than positive EM energy.[167] That can only occur if the majority of the output is negative energy, and any sudden surge in output (as from the shorted leads) increases the ratio of negative energy to positive energy so that very strong net cooling results.

For many predominantly mechanical components (simple electrical motors, light bulbs, etc.), Dirac hole current applied to the positive lead will power the circuit equally as well as electron current applied to the high voltage (negative) side, if the grounding is simply by a conductor and not a "vast reservoir" of charge. In such case, the wire chosen to conduct or direct the Dirac hole current should be made very small. It can and will carry intense hole current and will cool rather than heat.[168] The wire to the electron side of the device should be made normal sized, appropriate to the electron current to be carried.

---

[166] We do mean hole current in the Dirac sea, consisting of Dirac holes in the vacuum itself, prior to their interaction with charged matter to form lattice holes in matter. We do not mean the standard lattice hole current at all.

[167] In local curved spacetime, Dirac holes constitute "negative energy EM fields and potentials". In such case, one can have negative energy entities and reactions analogous to positive energy entities and reactions in 3-space. On the time axis, one also has negative energy entities and currents, potentials, waves, fields, etc. Because c**lassical** Maxwell-Heaviside-Lorentz theory eliminates any modeling of the local vacuum or the local curvature of spacetime, these negative spacetime curvature **(negative** energy) potentials, fields, and currents have not been added into the electromagnetic theory. That of course is a serious shortcoming of the normal theory.

Several other inventors over the years have been able to activate a barium ferrite magnet into sustained self-oscillation of its material particles or domains, in similar but weaker fashion for varying lengths of time — up to a month or so in the longest time that I am aware of. Sweet's magnet would continue in self-oscillation indefinitely unless (i) the magnet was sharply shocked mechanically (as with a hammer blow), (ii) the magnet was heated to too high a temperature, (iii) the leads of the output circuit were electrically shorted, or (iv) external artificial field interference (electrical pulse shock, particularly from longitudinal stress waves) was encountered.[169,170]

Unfortunately, Sweet signed multiple conflicting agreements with several financial backers, and his work was legally entangled so deeply that it was never straightened. He eventually died, taking with him his activation secret. He was often threatened, and was once fired at by a professional

[168] This is very probably the secret of Tesla's "single wire system". At an exposition in Nevada, Bedini demonstrated just such a system, by powering 600 watts of audio speakers at the end of a long, very thin wire across the stage from his power amplifier. If he had been using positive EM energy, the wire would have instantly melted.

[169] During that period, the Russians were playing with earthquake induction via scalar interferometers, at various places around the world, including inside the U.S. heartland. Sweet's machine would sit above a natural quake and not be affected. But when a quake within many hundreds — even several thousands — of miles was artificially induced, the Sweet device would react to it. A fairly close artificially induced quake would destroy the self-oscillation of the magnets, and sometimes a very close artificially induced quake would actually explode the intensely activated magnets. So for several years one had a way of ascertaining which quakes were natural and which were artificially induced, by simply having Sweet confirm how his magnets interacted.

[170] For those who do not believe such an activity, here is a direct statement by then-Secretary of Defense William Cohen: *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts. "* Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, **University of Georgia, Athens, Apr. 28, 1997.**

assassin using a silenced rifle. He survived the incident because, being old and a bit feeble, he stumbled and fell over the front steps just as the assassin fired. The bullet snapped by his ear, right where his head had just been. The incident was reported to the FBI, but with nothing else to go on, the assassin was never caught.

By whatever means, we do know that Sweet's magnets had to be activated by a special process that Sweet used and never fully revealed. An activated or "kinetic" magnet, it seems, was actually something a few of the "older magnetics guys" had known about, and Sweet had learned of it either from them or from Gabriel Kron. Sweet was also a transformer designer and expert, and he remarked that he had also observed specialized self-oscillation in certain transformers. Although most power engineers approached were either nonplussed or completely skeptical of a "self-oscillating field in a permanent magnet", self-oscillation of the fields, spin waves, etc. in permanent magnet materials *are* known to materials scientists today, especially to those working with magnetic effects in thin film materials {370, 371a-b}.

From a general relativity viewpoint, particle spin has a unique spin field of its own {372}, and there are such spin fields involved and operating in magnetic systems. The simple classical electrodynamic approach to magnetism and magnetic assemblies — or the view of a magnet as simply a bar with a north pole and a south pole — is quite insufficient when one wishes to research magnetic assemblies that will act as open systems far from equilibrium with their active environment, and thus will be capable of COP> 1.0 operations.

Recent work in coupled spins has revealed long-range coherence effects, previously unsuspected, as well as dual spin states participating in such phenomena {373a-b}. For example, Johnson has long used a 2-particle approach to spins and spin flipping in his own highly nonlinear magnetic assemblies {374}. His recent laboratory work has also shown similar and unsuspected *long-range coherence* effects in large multiple-magnet macroscopic systems {375} such as large assemblies of neodymium-iron magnets. This is strongly suggestive of the expected general relativistic spin field previously mentioned.

This work by Johnson is very important, since it may reveal a hitherto unsuspected mechanism for decay — of an otherwise nonconservative magnetic field around a closed loop — back into a conservative field. If one seeks an all-permanent-magnet rotating motor, a nonconservative magnetic field around a rotation loop must be established and sustained if one is to have self-rotation and self-powering. A nonconservative magnetic field around such a rotation loop means that the system acts as an open system far from equilibrium with its active vacuum and active curved spacetime environment, by means of some asymmetry mechanism such as multi-valued magnetic potential {376a-c, 377a-g}, elucidation of exchange forces {378a-c} at selected positions and directions, etc. It also requires both a stable broken 3-symmetry and a stable 4-symmetry.

On the other hand, in our opinion, Sweet employed spin waves triggered into self-oscillation in the nuclei of the barium atoms in his barium ferrite magnets. Indeed, such nuclear spin waves in magnets have been known for some time {379}. We postulate that, as in any other COP>1.0 power system, a self-rotating all-permanent-magnet motor must also involve Dirac sea hole current interacting in the various magnetic components. It is reasonable to suspect that these also interact with the electrons and their spin fields, perhaps producing the active mechanism providing that observed long-range coherence that apparently can decay the nonconservative magnetic field around the rotation loop, causing it to revert back to a conservative field and thereby stop the self-rotation and self-powering. If so, that long-range coherence must be violated and the violation stabilized. However, we could find no papers in the literature that deal with such possibilities. Hence it is a subject for further investigation by COP>1.0 researchers.

***6.3.2.4 The** Highly Successful Anti-Gravity Test*

On one trip to California to work with Sweet, we noticed that, at full power output of 500 watts, his unit weighed a little less on the bench than when not powered. Struck by this anomaly, I considered the negative energy aspects of the output — as evidenced by the Sweet's demonstration of icing when the output leads were shorted — and the very high gain, which added up to a predominance of negative energy in the output.

Pondering this from a general relativity viewpoint, I reasoned that: (i) increase of positive EM energy density in space increases the curvature of spacetime to produce additional positive gravity G_P; (ii) increase of negative EM energy density in space must increase the curvature of spacetime in the other direction, and thus it must produce negative gravity G_N:, and (iii) the total gravity G_T produced in the immediate spacetime of an object producing both positive gravity G_P and negative gravity G_N, was $G_T = f(G_P + G_N)$. Thus if we increased the output of the VTA unit, we must inevitably increase the weight reduction (antigravity) effect since the predominance of the energy output was negative energy, and this output was in the neglected negative energy fields aspect as well as in the local vacuum (spacetime) itself. After several back-of-the-envelope estimates and educated guesses over some weeks, I concluded that if the efficiency of the antigravity process[171] were 10% and the unit was pushed to 1500 watts output, it would lose all its weight and hover weightless.

However, because of the nonlinear local spacetime, magnetic monopoles[172, 173] were deposited in Sweet's magnets when the unit was producing power, and the density of these monopoles varied as the output power. Increasing the power output thus increased the tensile stress of the magnets, and at some point they would explode into fragments, like a hand grenade.[174] This explained why conditioned magnets sometimes exploded when experiencing fluctuations from artificial earthquake induction testing.

---

[171] We again accent the difference between efficiency and COP. Here we are definitely speaking of efficiency, not COP.

[72] Dr. Robert Flower has confirmed that the 'symmetrized" (or "extended") Maxwell equations (Eqs. 7.1 in Springer's *Encyclopedia of Physics,* Vol. 16, p. 431) include non-zero magnetic charges and magnetic currents. Antenna engineers (Henning Harmuth, and others) know that these terms must be included for proper description of boundary conditions, even though, at the final step of computation, their integrals are set equal to zero (apparently to snuff out magnetic monopoles). For Dirac's original theory of the monopole, see P. A. M. Dirac, "Theory of magnetic monopoles," *Phys. Rev.,* Vol. 74, 1948, p. 817-830. See also F. A. Bais and B. J. Schroers, "Quantisation of Monopoles with Non-abelian Magnetic Charge," *Nuclear Physics B,* Vol. 512, 1998, p. 250-294; G. Lochak, "The Symmetry Between Electricity and Magnetism and the Problem of the Existence of a Magnetic Monopole," in T. W. Barrett and D. M Grimes. [Eds.] *Advanced Electromagnetism: Fundations, Theory, & Applications,* World Scientific, 1995, p. 105-147. Particularly see T. W. Barrett, "On the distinction between fields and their metric: the fundamental difference between specifications concerning medium-independent fields and constitutive specifications concerning relations to the medium in which they exist." *Annales de la Fondation Louis de Broglie,* 14(1), 1989, p. 37-75; — "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in Lakhtakia, A. (ed.): *Essays on the Formal Aspects ofElectromagnetic Theory,* World Scientific, River Edge, NJ, 1993, p. 6-86; — "The Ehrenhaft-Mikhailov effect described as the behavior of a low energy density magnetic monopole-instanton." *Annales de la Fondation Louis de Broglie,* Vol. 19, 1994, p. 291-301. The work of Ehrenhaft, which is now substantiated and thoroughly replicated by Mikhailov, has produced experimental evidence of the magnetic monopole in many replicable experiments. In these experiments, magnetic particles move in a magnetic field along the lines of force. Reversal of the magnetic field, H, causes a reversal of the magnetic particle's motion, which is not the case with magnetic dipoles, since that can only occur for a monopole. See V. F. Mikhailov, "Observation of magnetic monopoles in the field of a line conductor," *J. Phys. A: Math. Gen.,* Vol. 18, 1985, p. L903-L906; — "Observation of the magnetic charge effect in experiments with ferromagnetic aerosols." *Annales de la Fondation Louis de Broglie,* Vol. 12, 1987, p. 491-523. See also M J. Perry, "Magnetic monopoles and the Kaluza-Klein theory," *Am. Inst. Phys. Conf. Proc,* No. 116, 1984, p. 121-125; E. V. Smetanin, "Electromagnetic field in a space with curvature — new solutions," *Soviet Physics Journal,* 25(2), Feb. 1982. p. 107-111. [English translation]. See also T. T. Wu and C. N. Yang, "Dirac monopole without strings: monopole harmonics," *Nuclear Physics,* Vol. B107, 1976,

Having no explosive control facilities, I warned Sweet not to push the unit past 1,000 watts, for safety reasons. At my urging, Sweet made a new output load of ten 100-watt lamps in sockets, so the load could be varied from 100 to 1,000 watts easily.

---

p. 365-380; — "Dirac's monopole without strings: classical Lagrangian theory," *Phys. Rev. D.,* 14(2), July 15, 1976, p. 437-445.

[173] We emphasize that a magnetic monopole can be taken simply as a highly localized magnetic scalar stress potential. See, e.g., P. A. M. Dirac, "Theory of magnetic monopoles," *Phys. Rev.,* Vol. 74, 1948, p. 817-830. Dirac pointed out that the existence of a single magnetic monopole in nature would allow an explanation for charge quantization. We believe that the reverse of Dirac's argument may be true also: The existence of charge quantization should allow an explanation for the magnetic monopole. However, we also surmise that a magnetic monopole exists only in (and as) a sharply localized curvature of spacetime. Jackson, *Classical Electrodynamics,* second edition, 1975, p. had this to say about magnetic dipoles and Dirac's argument: *"Dirac's argument... is that the mere existence of one magnetic monopole in the universe would offer an explanation of the discrete nature of electric charge. Since the quantization of charge is one of the most profound mysteries of the physical world, Dirac's idea has great appeal. "*

[174] Sweet in fact did explode several magnets by pushing the power output of the device. Pushing the output was easy (and is easy for any unit outputting mostly negative energy), because all one had to do was connect additional load.

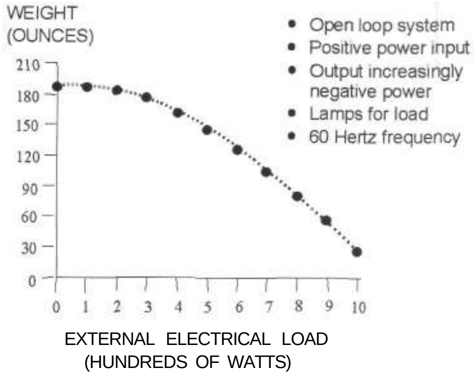

Figure 6-14 **Results** of the "pushed" VTA **antigravity test:** Weight **of VTA on the bench versus power output.**

Sweet completed the new load section and performed the experiment in California after 1 had returned to Huntsville, reading off the results to me over the phone. I took down the data, and then plotted the curve. Figure 6-14) shows the beautiful results. At 1,000 watts the unit had lost 90% of its weight. This was electrifying, because it was an unprecedented and highly successful antigravity experiment. Even my original "rough" assumption of levitation at 1500 watts was not too far from the indicated 1250 watts or so from the actual curve.

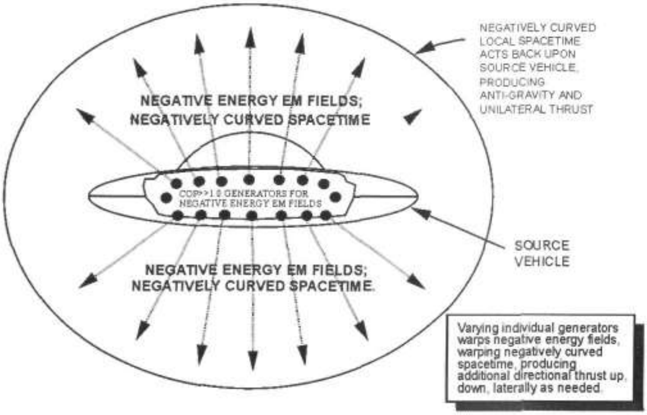
NEGATIVELY CURVED LOCAL SPACETIME ACTS BACK UPON SOURCE VEHICLE, PRODUCING ANTI-GRAVITY AND UNILATERAL THRUST
NEGATIVE ENERGY EM FIELDS;
NEGATIVELY CURVED SPACETIME
COP>>1.0 GENERATORS FOR NEGATIVE ENERGY EM FIELDS
SOURCE VEHICLE
NEGATIVE ENERGY EM FIELDS;
NEGATIVELY CURVED SPACETIME.
Varying individual generators warps negative energy fields, warping negatively curved spacetime, producing additional directional thrust up, down, laterally as needed.

Figure 6-15 Proposed mechanism for practical antigravity.

Figure 6-15 shows my proposed mechanism for antigravity. We simply use negative EM field energy generators to produce a negative EM energy field in space surrounding the source vehicle. The associated positive energy from the same energetic power source can power the vehicle simultaneously. The local reversed curvature of spacetime now interacts back upon the mass of the vehicle, to produce an antigravity component. By adjusting the output of the power source, the amount of negative energy — and hence the amount of antigravity component — can readily be varied and controlled. By controlling the directions of the field components, one can have unidirectional antigravity force, or antigravity and thrust force vectors simultaneously. Hence such a vehicle could be maneuvered and controlled, and be self-powering. This at least was and is my concept of a practical antigravity technology.

6.3.2.5 Later Developments and the End of the VTA

Later I prepared a paper on the VTA, including the antigravity experiment's results. We succeeded in getting the paper published {380} but I deliberately did not include the real details of the negative energy aspects as I wished to withhold that information for patenting. I prepared the entire paper, but placed Sweet's name first, as was appropriate since he was the inventor of the VTA.

Unfortunately — but understandably! — Sweet became a bit paranoid after the sniper assassination attempt (he was shot at by a distant sniper using a silenced rifle), and then would never allow a formal demonstration of his antigravity test. I strongly wished to present it to a major national U.S. scientific committee for independent testing and confirmation — which could have been done without Sweet having to reveal his activation process for the magnets. However, Sweet would not hear of such. Indeed, he fearfully kept the antigravity capability a secret from most of his backers. He strongly believed that he would be killed almost immediately if he revealed our antigravity experimental success. He may have been correct.

In fairness, the inventor and his backers of necessity have different priorities from those of scientists doing open research in a university. But at least a record of the test results is in the literature, though certainly not as rigorously or as extensively as I wished. Sometime later, I withdrew from any further association with the VTA unit, after the backers changed, a lawsuit broke out, and the entire project unraveled. Sweet later died without ever revealing his full secret for such powerful activation of self-oscillation in permanent magnets at low ELF frequencies.

The loss of the Sweet vacuum triode was a severe blow to science and a real setback to the scientific progress that could already have been made. Nonetheless, I contributed what I could, documented what I could, and made a videotape of the operation of the open-loop 500-watt device in the power mode. I also made a video of as much as Sweet would reveal of his activation process. I was not allowed to videotape a second antigravity experiment, even though I frequently and strongly requested it. By that time, Sweet mortally feared for his life. Perhaps what we have written in this book will be sufficient to provide some needed clues to some future graduate student or post-doctoral scientist who will succeed in recovering the necessary activation of magnets and both the COP»1.0 power capability and the antigravity experiments.

Meanwhile, I have filed a provisional patent application on the mechanism for antigravity, the gist of which I worked out solely on my own back in 1971 at Georgia Institute of Technology, and that I include in Chapter 8. As far as I am aware, no one else had thought of it previously, or has thought of it in that exact form *{using the causal positron prior to observation, and while it is still a negative energy, negative mass electron, as the source charge for negative EM field energy, and also using the causal positron's corresponding giant Heaviside nondiverged negative energy flow for substantial negative curvature of spacetime.).*

6.3.3 Intense Discharges in Ionized Gas

Intense sudden discharges in ionized gases are especially of interest due to the presence of optical frequency components and the involvement of iterative optical retroreflection etc. These processes seem to be involved in several investigations and inventions. Further, it appears that such intense discharges do produce bursts of negative EM energy as part of their overall discharge energy.

In the case of plasmas, some are known to transduce an input transverse EM wave into an output longitudinal EM wave, and vice versa. We know now that such a longitudinal EM wave is also accompanied by a time-polarized (scalar) EM wave in the time domain. The connection with the Whittaker 1903 decomposition of the potential, and with the new 4-symmetry negentropy law, is apparent. Several COP>1.0 researchers — notably Shoulders {381a-f} and Paulo and Alexandra Correa {382a-e} — have performed extensive research into anomalous discharge phenomena, and also have developed experimental devices capitalizing on the novel effects available.[175, 176] The Correas' U.S. Patent No. 5,449,989 includes many tests of a Correa glow discharge apparatus, one of which in 20 minutes of operation produced 0.988 kilowatt-hour of energy output for an input of 0.258 kilowatt hours. Using the known negative resistor characteristics, the Correas achieved sustained self-oscillations in the discharge.

As an aside, we point out that the well-known anti-Stokes emission — which produces more energy output than the experimenter inputs, and hence exhibits COP>1.0 — has long been validated in certain gases such as aniline vapor {383} and H_2 {384}. Stimulated anti-Stokes radiation occurs in many solid materials such as oxides and doped oxides {385}. Assuming the excess energy comes from the internal energy of the molecules, this is a prime candidate for free replenishment of that furnished internal energy of the molecule, via the broken symmetry of the molecular dipoles in their fierce virtual energy exchange with the local active vacuum and local curved spacetime. Of course, conventionally the energy furnished by the molecules, etc. is not replenished from the vacuum, hence anti-Stokes emission can be used for coherent light cooling effects {386}. Apparently the Correas' process accomplishes the energy replenishing, hence the COP> 1.0 performance over a 20 minute period.

Formation of longitudinal EM waves in the plasma forms in certain gas tubes, such as in Priore's gas-filled tubes {387}, may also occur in the Correa experiments and probably do. Longitudinal EM waves, of course, always involve their concomitant, accompanying time-polarized EM waves, since the combination comprises the scalar potential and hence the local curvature of spacetime provided by that scalar potential. The replenishment mechanism involved in longitudinal EM waves and scalar potentials — time-energy transduced into 3-space energy — apparently is

---

[175] Both Shoulders' electron cluster discharges and the discharge phenomena of Paulo and Alexandra Correa have been independently verified. Fox and other scientists have verified COP>1.0 in Shoulders' work, and Mallove has formally observed a self-powering Correa system. Both these highly reputable scientists are well known, have the highest integrity, and I would trust either of them with my life. Further, we point out that strong gradients — as exist temporarily in strong discharge phenomena — are already known to violate thermodynamics. That is one of the areas being researched today to advance thermodynamics, under the aegis of "extended thermodynamics". See Appendix A.

[176] In addition, the Correas have demonstrated a form of energy that demonstrates latent heat. See discussion in Paragraph 6.3.9.

invoked and utilized in the Correas' sustained self-oscillations in the plasma discharge.

6.3.4 Shoulders' Electron Cluster Processes

Valone {388} has summarized Ken Shoulders' charge clusters and processes as follows:

> *"An unusual energy source is the clustering of electrons by a discharge needle into a high density bundle equaling Avogadro's density of a solid.... Ken Shoulders has patented a process (Pat. #5,153,901) that produces electron clusters with such high energy density, they equal processes exceeding 25,000 degrees Celsius upon impact. Yet, he only uses 20 microjoules to produce the effects. The clusters travel at a maximum of one tenth the speed of light and penetrate any substance with accuracy and sharp precision. It is similar to xenon clustering techniques currently used at megavolt energy levels. Low energy nuclear transmutation of the target has also been achieved with this process. Using a deuterium-loaded palladium foil, only the bombardment areas show transmutation into silicon, calcium, and magnesium with electron clusters upon analysis with X-rays. Jin and Fox {389} have postulated that the high velocity electron clusters achieve results similar to accelerators, including penetration of the nucleus, with substantially less power. The new physics of like-charges clustering in bundles under low power conditions opens a wide range of applications including spacecraft maneuvering microthrusters. Overunity efficiency* [sic: should be COP] *is 9 to 1, or higher."*

Ken Shoulders is an accomplished experimental scientist who has discovered and patented many aspects of electron charge clusters, including their COP>1.0 characteristics {390}.

According to his experiments, one of the most remarkable characteristics of these charge clusters seems to be their dramatic reduction of the permittivity of space {391}. Shoulders has noted the remarkable removal Of **positive** charge during the self-formation process of the clusters {392}. In our opinion, the involvement of the supersystem of these charge clusters is shown by that phenomenon. We hypothesize that an accompanying Dirac hole current is shown by the experimental occasions where — in addition to the Shoulders' negative charge clusters — additional charges unite but show effects of having opposite charge {393}. Such changes would change the permittivity of space, which Shoulders has noted. They could also create a time-reversal zone (TRZ), which would account for the clustering of like charges since in a TRZ the law of attraction and repulsion of charges is reversed. In such a zone, like charges attract and unlike charges repel. This could explain why Shoulders has observed both negative charge clustering and positive charge clustering.

This would also explain the remarkable separation of the positive charges from the negative charges — they are simply *repelled* by the negative clusters in a TRZ.[177] This of course immediately involves curvature of spacetime effects, including local changes in the speed of light etc. As stressed by Shoulders, the entire area cries out for extensive and sophisticated experimental investigation on a continuing basis, as well as some powerful new theoretical work. Shoulders has also noted both "black" and "white" (dark and bright) clusters, including their combination and separation, strongly suggesting Dirac holes as well as the electrons in the charge clusters. He has proposed a form of "inertial propulsion" based on some of the mechanisms and asymmetrical forces observed in his experiments, and such asymmetry is possible in curved local spacetime conditions.

The relation of Shoulders' effects to cold fusion phenomena was shown by an experiment in which a deuteron-loaded palladium cathode was impacted by a charge cluster {394}. An explosive-like reaction occurred, and an X-ray analysis of the impact crater showed a number of elements not seen in the nearby palladium surface away from the impact site. Elements such as oxygen, calcium, silicon, and magnesium were detected in the impact site, but not in the surrounding palladium lattice. We would suggest the presence of TRZ-based nuclear transmutation reactions (via quasi-nucleus formation followed by quark-flipping in nucleons as the

---

[177] Since the same TRZ phenomenon generates specific nuclear transmutation reactions at low spatial energy, explaining the anomalous transmutation of elements in cold fusion electrolyte experiments, a strong indication exists that it is a viable mechanism. In our mechanism proposed for the flow of time, a TRZ is produced for charges whenever antiphoton interactions predominate over photon interactions. This indeed implies a change in the permittivity of the local space, local curvature of spacetime, and local changes in the speed of light — all phenomena Shoulders notes in his experiments.

TRZ decays to a time-forward zone). Additional information on this charge cluster phenomenon has been given by Fox {395}.

Other Shoulders patents are pending and his research is continuing. It is our understanding that at least one application is presently being developed for the market.

6.3.5 Mills's Hydrino Energy Cells and Developments

Dr. Randell Mills {396a-396c}, {397} has proposed a cosmic sea of hydrinos (the hydrino is a proposed reduced energy state of the hydrogen atom) to explain his tested energy cells. He envisions — and produces a theory for — hydrogen that is reduced to a lower energy state (the "hydrino"). He has gathered considerable evidence for his theory, and has accounted for some sixty or so infrared spectral lines in the sun's radiation that have previously not been explained. Independent tests appear to have validated his energy production claims.

Several major companies are presently working with Mills and his parent company, BlackLight Power, Inc. Progress can be followed from the company website, www.blacklightpower.com. Developments are ongoing in light, power, plasma, and new compositions of matter. A technical book by Mills {398} is freely downloadable from the website, giving his overall theory, as are other publications by Mills *et al.* {399a-h}. Critical objections to various parts of Mills' theory have been given, e.g., by Zimmerman {400}, with Mills responding strongly and — in our opinion — successfully.

Certain concepts utilized by Mills are, in our opinion, valid and insightful - whether or not the classical quantum mechanics developed by him is found to hold, and whether or not it agrees with ordinary quantum mechanics (which has problems of its own). The point is that *the experiments work.* It is thus the job of the theorists to find what changes must be made to the theory for it to remain consistent with experiment. If they do not like Mills' theory, the proper procedure is to advance one that **is** consistent with all the experimental results, and also to show where Mills' theory is inconsistent with the results.

In our own view, an *observed* positive charge is the output of an ongoing time-reversal process and therefore represents a "time-reversed entity" as observed. Before observation (imposition of *d/dt* operator), the positive **charge** is thus a negative 4-charge comprised of *qt*, not *q.* From that viewpoint, the H+ ion, which is just a hydrogen nucleus (a proton) without an electron shell, may be regarded as a time-reversed 4-entity prior to observation. Further, any energy change in and of the hydrogen atom results in a change in the local energy density of spacetime (ST) *a priori,* hence a change in the local ST curvature.

If we regard the entire hydrogen atom as a set of such ST curvatures and their dynamics, then Mills' consideration of a lower positive energy state *hydrino* simply implies the addition of an additional negative ST curvature to the ST curvature set associated with the normal hydrogen atom. We note that binding energy of the nucleons is already negative energy, and thus just such a negative spacetime curvature. But the binding energy is itself a spacetime curvature *capable of being increased,* as can be any spacetime curvature. In the supersystem, addition of such ST curvature is possible, and it is *doable* electrodynamically, particularly when Heaviside's normally unaccounted huge energy component is accounted and utilized.

In our opinion, Mills' theory should probably be recast into a higher group symmetry electrodynamics — such as O(3) — which has been shown to be intimately connected with Sachs's unified field theory. This would provide an adjusted model that is not only testable but also *engineerable.* Nonetheless, the concepts of Mills' theory and the theory should be closely examined as they stand, and at least some of the key concepts are engineerable as witnessed by his successful experimental results. That is better than what the critics have offered! A higher group symmetry recasting would enhance and expand his theory even more, and correct any errors that might be present. It would also allow a full theoretic treatment (i.e., a supersystem treatment) of Mills' main concepts, which we believe have much merit.

Further, given Mills' hydrino as we approached it, two hydrinos will each possess such a set of spacetime curvatures (each will possess such a *vacuum engine).* Later in this book — and elsewhere {401} — we have pointed out that, in a time-reversal zone (TRZ), the law of electromagnetic charge attraction and repulsion is reversed. Accordingly, in a TRZ two or more hydrinos will attract each other so closely that each enters the now-reduced strong force region of the other, forming a quasi-nucleus. As the TRZ decays away back to a normal time-forward zone, because of their power law form the strong forces restore back to normal much faster than the electromagnetic forces do. Hence during its "decay", the quasi-nucleus is energetically compacted into a permanent nucleus, either flipping quarks or giving off one or more particles as necessary to balance its new energy state in the process. We have previously argued that a similar process where a TRZ induces formation of quasi-nuclei also yields the major transmutations {402} demonstrated in a large number of cold fusion experiments worldwide.

The present author has previously proposed an engineerable mechanism (the totality of all photon interactions with a mass) that generates the flow of that mass through time {403}. By manipulating the arrow of time between forward vector and reversed vector, the strong forces in a nucleus (or for a particle) can be directly manipulated in magnitude (and to some extent, even in sign although the latter involves gluon and quark manipulations in the nucleons). By adroit use of these manipulated strong forces and TRZ formation and decay, the energy states of nuclei can indeed be manipulated, including reduced in energy level —just as Mills assumes and presents experimental evidence for, and as cold fusion experiments present experimental evidence for.

We note that two such interacting hydrinos must each possess a specific vacuum engine (specific set of ST curvatures and dynamics). Novel new interactions are available in TRZ's because of manipulation and even reversal of electromagnetic forces, strong forces, etc. and because of nuclear fusion of like charged particles into quasi-nuclei followed by decay into stable nuclei, etc. These novel new particle interactions caused by manipulation of strong spacetime curvature engines and their magnitudes are presently missing from particle physics. Mills, however, does come at them from a somewhat different direction in his own concepts.

The reason for the absence of dominant general relativity effects in most particle physics is that the "total photon interaction" mechanism generating the flow (persistence) of a mass or field through time has not been recognized and used. Consequently, in general only those far weaker ST curvatures due to the weak G-force have been considered. To Mills' great credit, he is aware of the gravitational significance of manipulating much stronger forces and spatial energy densities such as the electromagnetic fields and their energy densities.

The U.S. Patent and Trademark Office then refused a new patent to Mills on one of his processes, and revoked another Mills patent even after already agreeing to award it within two weeks. BlackLight Power, Inc. initiated and lost a lawsuit against the Commissioner of Patents, seeking to reverse this unusual action by the U.S. Patent Office. However, the company's patent portfolio has continued to grow, with extraordinary international applications {404a, 404b}, and several European patents have been filed appropriately. BlackLight Power Inc. seems well funded (some millions of dollars) and well staffed, and we expect to see BlackLight Power, Inc. products entering the market place. Those products will probably be related to fuel cells, plasmas, power converters, generators, separation methods, and novel compounds materials.

and

6.3.6 Switching Circuits with Extended Electron Gas Relaxation Time

See Figure 6-16. Suppose an instantly applied voltage to the high side of the circuit instantly potentializes a pair of circuit conductors, with the other acting as ground. The charges in the conductor are thus instantly potentialized and excited, so that through the external circuit there exists an E-field E, where $\mathbf{E} = -\nabla\phi$ between the high side of the circuit and the ground return side. An acceleration force on the energized electrons thus exists. It requires a finite time for the energized electrons in this force field E to break their static friction or inertia, move, and accelerate to the requisite slow drift velocity down the wire. During that delay or "electron relaxation time", no current is initially in progress, then a little current is flowing, and finally all of the current is flowing. This delay time is called the electron gas relaxation time.

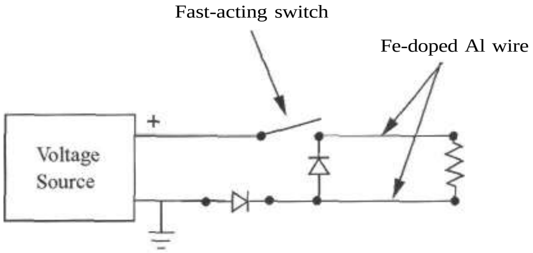

Note: May use many strands of Fe-doped Al wire instead of one, for each conductor

Figure 6-16 Extended-relaxation-time circuit with diodes.

In the usual copper conductor, the delay time arguably may be on the order of 10^{-16} sec. This is so slight a delay that it can be ignored, for all practical purposes. Current — with resulting dissipation of the collected potential energy — usually occurs "as soon as" we apply the voltage, for all practical circuit work. However, suppose we could increase that electron gas relaxation time to, say, one millisecond? Then for an appreciable number of microseconds, there would be essentially no current flowing at all in the now-potentialized circuit.

We previously proposed a mechanism whereby a degenerate semiconductor alloy (say, of a tiny bit — perhaps 2% or so — of iron alloyed in aluminum wire) is utilized for the conductors of the external circuit. This should give a vastly extended electron relaxation time, somewhere on the order of a millisecond. By obtaining a millisecond, one can excite the circuit with current-free potential alone, which becomes pure regauging during the electron relaxation period.

We strongly point out that this is an asymmetrical regauging; the E-field previously mentioned is present, but the electrons are still frozen in the circuit since their relaxation has not begun. The gauge freedom principle, recognized in quantum field theory, assures us that we do not have to perform work to simply inflow potential energy to change the potential energy of the system. In real circuits, we may have to pay a tiny bit for switching to connect and disconnect the external potentialization source, but we can be at pains to make that very efficient and minimal {405}. The point is, other than minor switching costs, the dramatic change in the potential energy of the system is cost-free and no current is drawn from the external potential source. Hence — in the language of the electrical engineer — we draw no power from the source of potential, during this potentialization time.

Suppose we switch away the excitation source prior to its decay while the electrons are still frozen. Simultaneously, we complete the "freed" potentialized circuit with a strategically placed diode, as shown in Figure 6-16. Again, we pay a little switching cost, but it can be minimized. The electrons in the *now asymmetrically potentialized circuit* are still frozen, so no power has yet been drawn by the system from the external power supply furnishing the voltage only. In this way, pure asymmetrical regauging is used to excite the circuit with additional potential energy, without requiring work (except minimally for switching).

A little more time passes, and suddenly the electrons in the potentialized circuit wake up and move. We stress again that the full emf is acting on the electrons, and now they are finally free to move. There is a net nonzero regauging E-field (force field) also. The excited closed current loop circuit then discharges in normal Lorentz symmetrical fashion, killing the source dipolarity as current is driven back up through the diode, but all the work in the load is "free". The circuit specialist will also recognize the usefulness of additional components associated with the shunt, such as a capacitor or even an LC oscillator for AC work.

If L*E* is load work obtained and *SE* is switch energy utilized, this approach yields COP = *LE / SE* and COP > 1.0 is possible. The reason is that the *primary* source dipole providing the potential for potentialization of the electrons and regauging of the circuit is removed from the closed current loop circuit and the effect of current through the back emf. Thus the primary source dipole is not destroyed faster than the load is powered. It is only depleted by the small amount of switching energy we utilized. This appears to be a major way to avoid the COP<1.0 limitation of the usual closed current loop circuit.

This is a circuit with legitimate COP>1.0 capability, without violation of the laws of thermodynamics or physics. We break the Lorentz condition between the initial "external circuit" and the primary source dipole for the regauging energy. After regauging, we "cut loose" the circuit with its own primary dipolarity and newly transferred "separate source dipole". *By such adroit use of electron gas relaxation time, we can use one source dipole to freely make another of equal intensity, without any depletion of the first.*

The difficulty is in getting the Fe-Al alloy wire, which must be made in an inert atmosphere since Al corrodes easily and the two elements have very different melting temperatures.

Alternatives can be visualized, such as to utilize a capacitor whose plates are made of the Fe-Al alloy. This might be useful in the triode-capacitor, to be discussed next. A series of pulses might then be used to charge the capacitor, without depletion of the primary source dipole in the external source of potential, if it is repeatedly switched in and away. An interesting question arises if such a capacitor with tailored "hysteresis" in its charging is utilized in various circuits and schemes containing a coil of useful time delay also, particularly since the electron gas relaxation phenomenon involves decaying oscillations. Interesting phase angle results can be constructed in such circuits, but we leave the resolution of this question to future experimenters.

6.3.7 The Triode-Capacitor and Other Curious Components

See Figure 6-17. In this rough concept by the present author, the notion is to implant a grid of very small copper wire mesh inside the dielectric and near one or both of the capacitor plates, and then to utilize grids as if they were grids of a triode tube, for gating and controlling the displacement current in the capacitor and the dielectric strain. As an example, when charging the capacitor, the grids can be used to increase the displacement current (which also increases the voltage on the capacitor and dielectric strain). This increases the charge and energy in the charged capacitor, without putting in as much "enhancement and gating" energy as the extra energy obtained in the charged capacitor. If desired, we can also incorporate the 2% Fe-doped Al wires in the grid, or connected to it, to further move the grid toward a "pure current-free potential" operation.

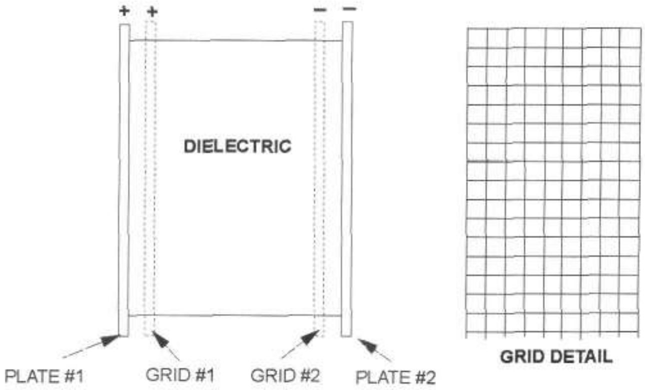

Figure 6-17 The triode-capacitor concept.

The scheme is that a voltage on this grid will intercept only a small current during charging of the capacitor, but can have an appreciable voltage upon it. The charging potential is placed across the capacitor, and simultaneously the grid has a similar potential placed upon it. The external source of charging potential is then disconnected, and the capacitor charges with more energy rearranged in its circuit than the energy one dissipates in the grid. The similarity to an old vacuum tube triode is apparent. For the purist, additional grids with tetrode and pentode functions can also be added.

One or more inventors has unwittingly tried to approach this in one way or another, but have not directly incorporated this schema and have not used the grid. They do not seem to have obtained positive results with those different schemas without the grid, nor have they specified the difficulties and phenomenology encountered, etc. We were curious about possible results of combining this (the triode grid in the capacitor) with the known elimination of capacitor losses by appropriate pulse charging {406a-b}. It seems that experimentation and research with the triode capacitor and its variations is definitely indicated.

We were also curious about the Davis non-inductive resistor {407a}, which is a special kind of "capacitor" that becomes an inductance-free resistor. One may argue that the resistor may be multiply inductive after all, though it is *net* noninductive. It may well be "equally and oppositely inductive," which is quite a different thing altogether, since it produces an artificial stress potential (and in the presence of AC, it produces a longitudinal EM stress field wave) in spacetime even though it produces a *net* zero reactance. In short, it produces a Lorentz-regauging, directly, and this can be a varying Lorentz regauging producing longitudinal EM stress waves. Whether this odd characteristic can be utilized in a COP>1.0 process remains to be seen. Preliminary indications are that it can be so utilized.

The various versions of the biwound coil, wound with two conductors at once and then the currents passed through the conductors in opposite directions, also presents interesting capabilities. Here again, it is multiply inductive, but may be tuned and adjusted to be *net* noninductive. We strongly suggest comparison with the well-known hairpin dual dipole antenna.

In short, this biwound coil arrangement — with various variations — can produce (and modulate) a magnetostatic scalar stress potential in the local vacuum and, in the case of AC, a longitudinal EM wave. This would appear to be a direct change (and oscillation) of the stress potential of the vacuum itself, which can be decomposed via the Whittaker process into a harmonic set of phase conjugate longitudinal EM wavepairs. The potential longitudinal EM wave communication possibilities can be seen, although one might require a Fogal transistor to receive the longitudinal EM wave modulations of such signals.

Further work in this area of unusual capacitors and components is left to the interested reader for his or her experimentation and research.

6.3.8 A Surprising Thing About Thermodynamics and Reservoirs

Thermodynamically, it is fashionable to state that one cannot take energy from a reservoir at constant temperature. That is true for *equilibrium* conditions, but not for *nonequilibrium* conditions {408}. That is, we cannot take energy from a reservoir *in equilibrium* at constant temperature. We can indeed take energy from a reservoir at constant temperature *but not in equilibrium*. More exactly, we can take energy from a nonhomogeneous reservoir at constant temperature.

We quote Hsu-Chieh Yeh {409} for a vivid statement of this little-recognized fact:

> *"From Planck's statement of the second law of thermodynamics it is generally inferred that it is impossible to construct an engine which produces work at the expense only of heat taken*

> *from the air or the ocean. ...[It is demonstrated that]... when the air and the ocean are combined as a nonhomogeneous reservoir of uniform temperature, it is possible to construct an engine which produces work by extracting heat from the said reservoir. This does not constitute a violation of the second law of thermodynamics, rather that the "reservoir" in the Planck's statement must be clearly stated as being in equilibrium. "*

Yeh's article falsifies the general, mistaken inference that Planck's statement of the second law of thermodynamics requires it to be impossible to construct an engine that produces work at the expense only of heat taken from the air or ocean. The experiment described in Yeh's paper demonstrates that when the air and the ocean are combined as a nonhomogeneous reservoir of uniform temperature, it is possible to construct an engine that produces work by extracting heat from the said reservoir. This does not constitute a violation of the second law of thermodynamics, but rather shows that the "reservoir" in Planck's statement may be in equilibrium or disequilibrium, and the statement holds only for the equilibrium case. Also, an entire reservoir can be in "overall" equilibrium, but in raging disequilibrium between various parts of itself. Classical thermodynamics only applies with the reservoir or part of it that is used by the system in its supersystem, when that part is in equilibrium.

We bring this out to show the analogy to our concept of the two additional components of the supersystem as the "reservoir" or external environment of the system itself.

We also point out that the concept of temperature itself is not strictly defined under disequilibrium conditions, as pointed out by Gabriel Laufer {410}.

The proof and a schematic diagram of a machine to do just such an energy extraction process at constant temperature in an environment in disequilibrium are shown by Yeh {408}.

6.J.9 Latent Heat Work of the Correas

In addition to their glow discharge work previously discussed, Paulo and Alexandra Correa have demonstrated a method to optimize some anomalous form of energy from the sun, which transduces into heat, producing an anomalous T_0 - T difference registered in Orgone Accumulators[178] invented by W. Reich in 1939-1940. They have also employed the optimized difference in this "latent" heat — their term — to drive a low dT Stirling engine of the MM6 type, to perform free work {41 la-b}. The Correas employ their own new theory of the aether {412} to explain the effect.

---

[178] **This** is Reich's terminology, not used in conventional physics. We suspect that **proper** application of quantum field theory with its four photon polarizations could

Against the assertion that this is just normal solar energy warming, they have demonstrated the successful driving of the Stirling engine for several hours during nighttime {413}. With their latest improved version, they report speeds of the Stirling engine of 150 rpm during daytime (with dT values reaching 22.5° C), and speeds of 30 to 80 rpm during nighttime (with dT values of 3.1 to 9° C). There appears to be no obvious standard thermodynamic explanation or standard theory to explain how the anomalous heating occurs — although once it does occur, the explanation from then on is thermodynamically normal.

The Stirling energy is driven on a difference in temperature created between the top plate and bottom plate of a box by this anomalous energy. It is the production of the heat that is anomalous, not the driving of the engine. For comparison, imagine a bank of solar cells that would somehow operate at night, when no normal radiation from the sun was striking them. The Correa effect demonstrates such an anomaly.[179]

Dr. Eugene Mallove, an excellent scientist with impeccable credentials, has personally observed the Correa accumulator system driving the Stirling engine, and has examined it inside and out {414}. Though small, the heat and the motive power are real and the engine is really driven, in the absence of any kind of conventional batteries or other known power source.

---

be applied to transform the notion of "orgone" energy into specialized EM energy. We further suspect that orgone energy is really the transduction of the time-polarized photon energy into normal photon energy. We are assured by quantum field theory and the giant negentropy solution to the source charge problem that the instantaneous scalar potential involves this process.

[179] We point out that local regions of the sun may well produce Aharonov-Bohm effects, and probably do. Such reactions in the sun would localize some of the B-fields and emit curl-free magnetic vector potential radiation into space, striking the earth. Close examination and analysis of the possible reaction of uncurled A-potential radiation with the accumulator box and plates in the Correa process might yield a novel mechanism for the observed heating (scattering) effects, including the difference between daytime and nighttime levels, periodic variations, etc.

While the Correas utilize some of Reich's orgone and aether terminology, they are quite aware that the energy being converted is not normal transverse EM wave energy. Quoting {415}:

> *"Aetherometry demonstrates that what traverses space is not transverse electromagnetic radiation (and certainly not sensible heat), but longitudinal electric ambipolar radiation emitted from the sun. All that electromagnetic radiation consists of is a local production of photons (for those who care to listen): aether energy is not composed of photons, actual or virtual."*

We can generate sympathy for that statement, although we would still not rule out time-polarized (scalar) photons and longitudinal photons from quantum field theory, each of which is individually nonobservable, but when paired the two comprise an instantaneous scalar potential. As we previously discussed, the classical notion of the EM force fields E and B implicitly assumes interaction of the unseen causal "field-as-it-exists-in-massfree space" with charged mass having been accomplished, with the E and B being the 3-effect of this interaction after observation. Thus these force fields E and B are 3-effects (after interaction and after observation) rather than 4-causes prior to interaction and observation. We suspect that the Correas are trying to get at the causal 4-photon, rather than the usually assumed 3-photon. Further, a virtual photon spends much of its time existing as a dipole comprised of a virtual electron and a virtual positron. Therefore it spends much of its time in the presence of the giant negentropy 4-circulation, involving transduction of time energy into spatial energy and vice versa. It is this negentropy circulation represented by the virtual photon (or accompanying it, whichever one chooses) that the Correas are confronting as the "aether energy".

As previously discussed, an analysis of the Drude electrons in the electron gas of a detector builds a convincing case for EM waves in space existing in longitudinal EM waves, *if the spinning and longitudinally restrained electrons precess and act as gyros.* The detected transverse EM waves are in fact transverse precession waves of the gyro-electrons when perturbed by incoming longitudinal EM waves (longitudinal photons appearing in 3-space from the time-domain's scalar photon couplet). As Feynman put it {416},

> *"We may think of E(x, y, z, t) and B(x, y, z, t) as giving the forces that would be experienced at the time t by a charge located at (x, y, z), with the condition that placing the*

> *charge there* *did not disturb* *the positions or motion of all the other charges responsible for the fields."*

The E and B force fields — and transverse photons — are not the form in which EM energy *causally* exists in 4-space prior to its interaction with charged mass. In that causal form, EM energy seems to exist in field-free form, as a direct curvature of spacetime — at least in Sachs's unified field approach. Further, this form may be represented as the causal precursor of longitudinal EM waves accompanied by time-polarized EM waves.

We will not repeat our past discussion of the great necessity to change the very notion of "propagation through 3-space", but that is the issue. In our opinion, the Correas have had to struggle with the use of this causal EM energy as it exists in 4-space prior to interaction with charge and prior to observation. Hence their view is solidly based that the photon does not exist until after the interaction — which in fact has much support once we remove the still-present assumption in classical electrodynamics that a material ether exists.

Our own recommendation is that this anomalous energy should be deeply investigated with the view that all EM energy in 3-space comes from the time-domain and the 3-space aspect is in terms of longitudinal photons, as we have proposed in our solution to the source charge problem previously cited. That is the only way that the nature of the EM wave in vacuum can be reconciled with quantum field theory's findings {19} and with Whittaker's decomposition of the scalar potential {85}, slightly corrected {12}.

Mallove also points out {417} that, if this aether energy view of the Correas has validity, it *"...suggests that physics is not just wrong, it is profoundly wrong — catastrophically wrong about even the most fundamental questions of existence, such as: 1) Is there an energetic aether? and 2) What is the nature of light?"*

We profoundly agree with Mallove's observation, and we believe we have explained the reason why physics is indeed profoundly wrong (widely substitutes effect for cause). Further, classical physics is decidedly wrong about the nature of light — from quantum field theory, experimental proof of superluminal communication, and action-at-a-distance effects.

In our view, the Correas presently have a replicable experiment that, once independently and thoroughly investigated, may force a deeper examination of the foundations of physics. Hopefully, the Correas' experiment and others like it will finally force the separation of effect from cause, presently so mangled in physics and particularly in electrodynamics.

If the National Academy of Sciences and the National Science Foundation can be convinced of the horrible manglement of cause and effect in physics and especially in electrodynamics, perhaps they may yet wake from their long slumber and get the foundations corrected to eliminate such non sequiturs. If so, we shall have a magnificent new physics as a result.

If not, then the NAS and NSF will continue to profoundly reduce the security of this nation, since several nations have already made and weaponized such corrections to physics to produce what is called "energetics". Further, extraordinarily powerful energetics weapons have **already** been developed and deployed (and even employed), putting the U.S. at a strategic disadvantage because of the continuing reluctance of the NAS and NSF to correct the glaring and extraordinary errors in both physics and electrodynamics — and especially in electrical engineering.

Meanwhile, we believe that the work and experiments of Paulo and Alexandra Correa are at the forefront of capturing and transducing energy from the vacuum, in whatever form one chooses to consider the vacuum energy itself. They are also pointing the way to the corrections and changes necessary in the "old physics" if we are to better understand the world we live in.

6.3.10 Swiss Overunity Converter

We must also include the "Testatika" devices invented by Paul Baumann in **a** religious commune, Methernitha, adjacent to the village of Linden in Switzerland[180]. Figure 6-18 shows one of the machines, which over a period of 20 years were developed in various output sizes up to 30 kilowatts. Once started, the device is self-powering while also powering a very substantial external load (up to 25 kW). Many engineers and scientists have observed the machines in observation, examined them, and measured one or more in operation, so there is little doubt of their **authenticity.**

---

[180] The website for Methernitha is www.methernitha.com. One may read the statements of the religious group for oneself.

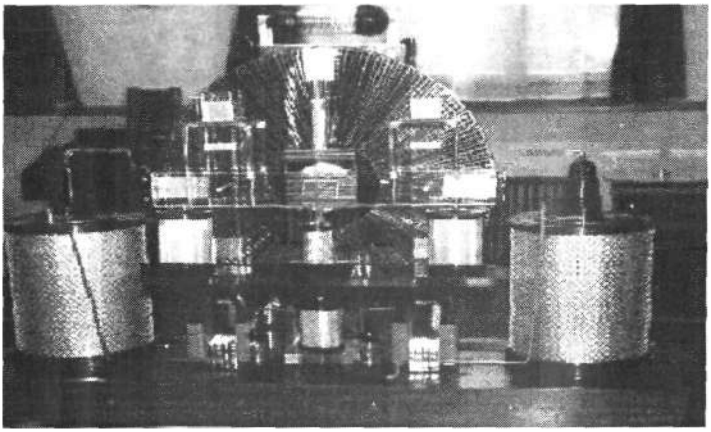

Figure 6-18 Swiss Methemitha machine.

However, there the science stops and mysticism and speculation begin. The religious commune releases no real details, except deliberately tantalizing clues. When questioned, the reply is that humankind in its present state cannot be trusted with such free energy. Almost all those examining the device have thought in ordinary electrical engineering terms, so very little comes from their examination because standard electrical engineering does not contain any theoretical explanation of COP>1.0 electrical power systems whatsoever. Baumann himself sometimes cryptically replied that his study of lightning discharges in nature gave him the operational principle. As we shall see, that may indeed be a very substantial clue as to the possible technical COP>1.0 mechanism, and one that does not exist in conventional classical electrodynamics.

The Methernitha (Baumann) machine resembles a Wimhurst machine with contra rotating static wheels, made of plastic or other insulating material with copper ribs and structures on one side and aluminum on the other. Use is made of the discharge into and from large capacitors and large coils. Also, at least one horseshoe-shaped permanent magnet assembly at the bottom is utilized. The major operating feature seems to be the contra rotating disks of a "Wimhurst" static generator machine, a magnetic field at right angles, and the sharp discharge of high voltage into and from large capacitors and large coils. The voluminous material on the Internet and in articles about the machines is mostly of little real use in contributing understanding or in proposing any legitimate mechanism.

We have not "solved" the machine's mechanism, since that would require close study of the device for a protracted time, and there has been no opportunity or time for that — although Baumann invited this author to come and see the machine some years before he died. Regrettably, I was unable to make the trip, and Baumann himself has now been dead for some years. We will therefore offer some technical suggestions pertaining to mechanisms that may be operating in the machine (or others like it).

First, we call attention to a paper by Aguirregabiria {418}. Quoting:

> *"An ohmic ring that rotates with constant angular velocity in an external uniform magnetic field is considered as a simple model for a current generator. Under the assumption that all quantities vary slowly in time, the lowest-order approximation to the surface charge density is found. The flux of the Poynting vector through the loop surface is also computed. Unlike the examples that are given in textbooks, this flux is not always incoming: It has the outgoing direction around the loop parts where the electrons are moving against electrostatic forces. "*

In other words, contrary to textbooks, there are positions where such a ring produces currents in one direction, and positions where it produces current in the other direction — *in fact moving against the voltage and E-field, and thus serving as a true negative resistor in those latter regions.* This may be a profound clue to the negative resistor operation of Baumann's machine.

This unexpected negative resistor effect is in theory usable to provide a true negative resistance. If there are transformers in the machine (and there are in some of them at least), then if a true negative resistor were used as an external shunt across the external circuit connected to the secondary of a transformer, the transformer could be operated as a true COP>1.0 device. The back-field connection from secondary to primary would be reduced by the fraction of the return current to the secondary that is shunted in parallel to that current "pumped" back up through the secondary.

This should be considered as an important possibility in the Baumann machine's operation.

Second, a charging and discharging capacitor involves not only energy flow in a direction perpendicular to the plates, but also an energy flow in a direction parallel to the plates {419}. Each of the rotating "plates" is actually a rotating capacitor of sorts, with a dielectric between two plates of special construction and dissimilar metals. The dielectric and the plates thus involve a complex intermingling of the negative resistor effect, a positive resistor effect, magnetic energy flows (both in entropic and negentropic manner in various regions), etc.

Baumann once handed two dissimilar plates, with a dielectric between them, to an engineer who measured it and found a startling several hundred volts across the plates! To date, no one is known to have been able to reproduce that effect.[181] We suggest that, if the dielectric material is layered and photorefractive in nature, it may be related to the negative resistors built by Chung {420}. If so, the physical pressure of the plates against the layered dielectric may also be quite important. By varying this pressure, Chung and her team found that their device could be made to exhibit negative resistance, positive resistance, or zero resistance. This leads us to speculate that Baumann may have used a "stressed dielectric" capacitor as a device having a built-in negative resistance. In that case, build-up of such a surprisingly high voltage would indicate high stress locked into the negative resistance dielectric.

We also call attention to a paper by LaPointe {421}, which will be discussed later under inertial propulsion and antigravity. Essentially, true Dirac hole current in the adjacent vacuum/space can be generated by powerful electrical discharges, particularly from capacitors or coils. When the unaccounted Heaviside energy component is also considered as well as the Poynting component, LaPointe's despair at the difficulty of obtaining sufficiently large voltage gradient across a very tiny distance is dispelled.

---

[181] We are reminded, however, of the peculiar phenomenon that can occur in large power capacitors stored on a warehouse floor. If their leads are not shorted, the capacitors will build up a self-charge gradually, and this can become quite lethal. For that reason, such large capacitors when stored will have their leads shorted, to prevent this "free cumulative charge and energy" from appearing and posing a serious hazard. To our knowledge, there has been no complete scientific explanation advanced for this "self-charging" effect in large storage capacitors. Our own hypothesis is that it is probably an effect of unaccounted Heaviside nondiverged energy flow components associated with various field/charge interactions in the area and with ordinary Poynting EM energy flow in the area, including within the earth itself. Further, with respect to the surface of the earth there is an increase in potential with altitude. Hence from any point above the earth's surface and a point on the surface, there exists a dipolarity and a "Kron open path". The broken symmetry of that dipolarity means that within it a continuous 4-circulation of EM energy flow occurs as well as point-dipole polarization.

If this speculation holds, then in examining lightning strikes and their more bizarre phenomena, Baumann may well have hit upon the fact that such discharges also generate negative energy as well as positive energy, although it appears that the technical concept of negative energy was foreign to him, at least in such terms. In short, he could have been unknowingly charging his large capacitors by Dirac sea hole currents, somewhat similar to Bedini's process.[182] In that case, most of the "negative resistance" operation could be explained by Baumann's transduction of large bursts of negative energy into large bursts of positive charging energy and current in his capacitors.

Or, said another way in terms of Aguirregabiria's effect, Baumann may simply have been charging capacitors in that "negative resistance" Aguirregabiria region where the current does flow backwards against the voltage.

At least we have proposed some known though rare phenomena, which Bedini and my four colleagues and I have discovered in independent efforts. We strongly suggest that the Baumann device may unwittingly incorporate several of these mechanisms, but that neither Baumann nor the other members of Methernitha seem to have any technical notion of them.

With that, we leave it to the interested researcher to perform his or her independent Baumann-type system experiments and build-ups. Eventually, as is always true in science, the experiment must determine the truth of what is really happening. Given the successful experiment, then a model must be contrived or fitted that (i) explains the new effects produced, and (ii) complies with normal EM operation of normal COP<1.0 systems.

6.4 Some Magnetic Processes of Interest

There are some 200 or more magnetic effects in the literature, and only about half of them are well understood. Of the remaining half, some are partially understood and some are not understood at all. The COP>1.0 researcher interested in magnetic systems would do well to study magnetic materials science {422a-422c}, known magnetic effects, and particularly those magnetic effects that are still anomalous such as in leading edge superconductivity research. The new area of *spintronics* {423}, e.g., is rapidly advancing and will have future impact on COP>1.0 magnetic systems. We do not further discuss spintronics in this book, but merely call this rapidly emerging area to the attention of the interested COP>1.0 researcher.[183]

---

[182] Bedini and the present author have filed a patent application upon this exact mechanism and process, and typical embodiments for accomplishing it. Bedini discovered the effect and used it for two or more decades, and after some intense struggle with the problem I was able to find the exact technical mechanism, including literature justification. This becomes a system to close-loop a "unitary" type of COP>1.0 system, which is covered in the patent application also. A different system can be used to close-loop a non-unitary COP>1.0 system, and that method is contained in the second MEG-type patent application of my four colleagues and I who invented the motionless electromagnetic generator (MEG).

Let us examine a very few magnetic effects of interest, as a sampling of what we are referring to.

6.4.1 Self-Powered Magnetohydrodynamic Motors

Interesting self-powered magnetohydrodynamic motor work has been shown by Bednarek {424}. He shows a multi-branch rotor in a galvanic cell composed of sulphuric acid electrolyte with copper and zinc electrodes. The unit develops an emf of about 1 volt, with the rotor placed in a vertical magnetic field such as from a large rectangular permanent magnet, where the field of the magnet is a few teslas in strength. The electrolyte and the rotor revolve in opposite directions. To change the direction of revolution of the rotor, the magnetic field direction is reversed. A linear self-powered magnetohydrodynamic motor is also shown, and has a similar principle of operation.

The interesting point is that these motors work, but less efficiently, when a water solution of table salt is used as the electrolyte. With the worldwide availability of seawater, this becomes an interesting possibility of "renewable energy" application. We leave it to the reader to further investigate this proven experimental technology as desired and to determine its practicality.

6.4.2 Multivalued Magnetic Potential

Conventional system designers work with conservative fields, as shown in Figure 6-19. Conservative fields arise from a single-valued potential — a potential that has only one value at each point in 3-space that it occupies. Consider a rolling ball of mass *m* on the "oval track" shown in Figure 6-19, starting from point A and rolling on the right side path. With the ball at rest at point A, we give it a slight push to velocity V_i adding kinetic energy $K_i = 1/2\, m(V_i)^2$. At point A, the ball has its maximum potential energy P_A due to gravity.

---

[183] For technical information, see (a) Michael Ziese and Martin J. Thornton, Eds., *Spin Electronics,* Springer-Verlag, 2001; (b) D. D. Awschalon, N. Samarth, and D. Loss, Eds., *Semiconductor Spintronics and Quantum Computation,* Springer-Verlag, 2001.

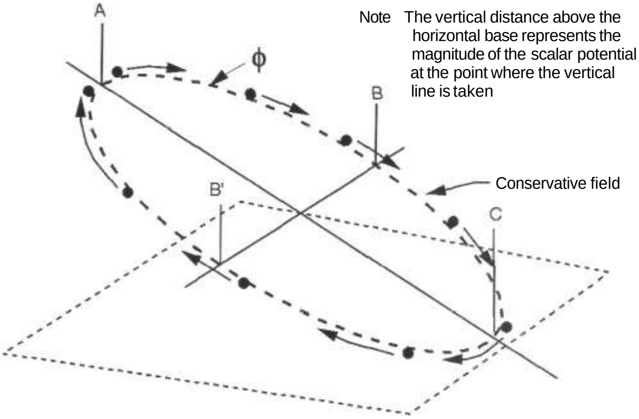
A
B
C
B'
ϕ
Note The vertical distance above the horizontal base represents the magnitude of the scalar potential at the point where the vertical line is taken
Conservative field

A circular closed path in potential ϕ is given by A-B-C-B'-A
Line integral from A around any closed path back to A is conservative
Along path A-B-C, work W(1) may be extracted from rolling ball
Along path C-B'-A, must do work W(2) on ball, where W(2) = - W(1)

Figure 6-19 Conservative field use (single-valued potential).

As it moves to the right because of our push, the ball will accelerate due to the force of gravity and increase its kinetic energy until it reaches point C at the bottom, reaching its maximum linear velocity Vc and kinetic energy Kc at point C. Its change (PA - Pc) in potential energy at from point A to point C has been converted to kinetic energy $[\tfrac{1}{2}\, m(V_1)^2 + \tfrac{1}{2}\, mVg^2]$ at point C.

Continuing on around on the left half of the path, as the ball rises toward A again, the portion $(P_A - P_C) = \tfrac{1}{2}\, mVg^2$—of the ball's kinetic energy at C that was added by the force of gravity from A to C — is returned to potential energy P_A. The ball reaches its lowest kinetic energy $\tfrac{1}{2}\, m(V_1)^2$ at A and also its lowest velocity V_1. For a perfect lossless system, once started in motion the ball would rotate around the track indefinitely, freely changing potential energy into kinetic energy and then back to potential energy. Nevertheless, it would not do any outside work, for that would represent losses or dissipations of energy from this conservative system. Consequently, the ball would quickly run down if work were being done, even if the system were otherwise "perfect".

If one integrates the change in potential energy around the track, the net change in potential energy is zero. The work done on the ball by increasing gravity to increase the ball's energy in one half-cycle, is taken back from the ball back when it climbs back out into decreasing gravity in the other half cycle. If one integrates the changes in kinetic energy around the track, once the initial velocity and kinetic energy are produced by outside forces in a perfect system, the net change in kinetic energy is zero. Of course, it is easily seen that the ball gains kinetic energy on its downward half of the track, and then returns the kinetic energy on its upward half of the track.

This situation is said to involve a *conservative fields* and the system will not produce any net energy to use as free work. Indeed, a real system will almost always have some friction and other losses around the path {425}, so the ball will gradually lose net initial energy given it, slow down, and eventually come to rest at point C, the lowest potential energy of the system.

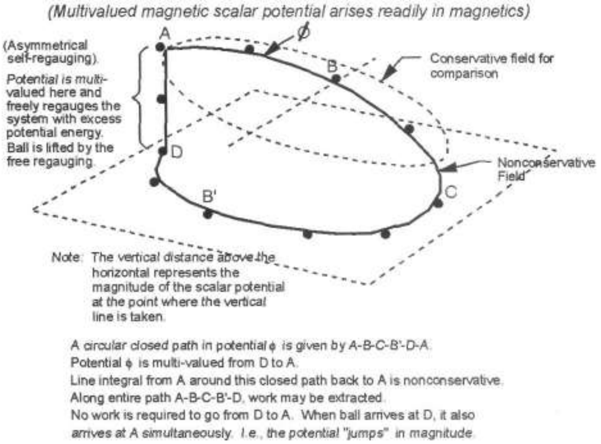
(Multivalued magnetic scalar potential arises readily in magnetics)
A
ϕ
B
(Asymmetrical self-regauging).
Potential is multi-valued here and freely regauges the system with excess potential energy. Ball is lifted by the free regauging.
Conservative field for comparison
D
Nonconservative Field
C
B'

Note: The vertical distance above the horizontal represents the magnitude of the scalar potential at the point where the vertical line is taken.

A circular closed path in potential ϕ is given by *A-B-C-B'-D-A*.
Potential ϕ is multi-valued from D to A.
Line integral from A around this closed path back to A is nonconservative.
Along entire path A-B-C-B'-D, work may be extracted.
No work is required to go from D to A. When ball arrives at D, it also arrives at A *simultaneously. I.e., the potential "jumps" in magnitude.*

Figure 6-20 Field may be non-conservative for a multivalued potential.

A charge circulated through a closed current loop circuit also moves through a conservative field region in similar manner. Hence, there is no net excess energy input available for such a conventional circuit. It follows that one must arrange for inputting all the input energy (intercepted and collected by the circuit) that is then dissipated from the circuit's loads and losses. Unfortunately, half the collected circuit energy also goes to destroying the dipole, with less than half powering the load. Therefore, that circuit always exhibits COP<1.0.

Now see Figure 6-20, where we provide an analogy using a rolling ball around a circular track, in a gravitational potential. In this case, the potential has a multiple magnitude A and D at point A-D, which is the same point (we plot a point in the potential's magnitude, not a point in 3-space, although there can be close approximation).

Approaching magnitude D from point B', the value of the potential is steadily reducing and is much lower prior to reaching A-D than it was when leaving magnitude A. During this phase, the ball is gaining kinetic energy from the gravitational potential, and can be used to perform useful work (up to all the energy gained, in a perfect system). Upon passing through point A-D, instantly the ball is again at point A on the right. We accent that the "instantaneous jump" on the diagram represents a free insertion of excess energy (asymmetric regauging energy) from outside the system. Rigorously that is a broken symmetry, which also breaks the internal energy conservation of the system. In short, the "ball" has been instantly "lifted" (in this simplified analogy) back to a higher potential by a free insertion of excess energy from the external environment. The main point is that a surge of potential energy into the "ball system" occurred, freely input by nature and the environment. The system can now go through the "doing work" routine again, traversing from A back to D and through it.

Note that this is a nonconservative system, because it continually receives a free and sudden input of excess energy from its environment. So this system — because of its broken symmetry — can continually do work and keep on going.

Indeed, so long as that free energy input from the environment occurs without fail, this is a "self-powering" system, completely complying with **the** laws of physics and thermodynamics. It violates the equilibrium thermodynamics because the system is periodically not in equilibrium with its environment. Consequently, the system can exhibit those five magic functions we spoke about previously. It can exhibit (i) self-organizing (in this case, freely getting that little ball from low potential energy D back to high potential energy A), (ii) self-oscillation or self-rotation (the ball will continue to go around and around the loop, even doing a little work in the process), (iii) output more energy than the operator inputs (in this case, the operator is not inputting any energy at all, so the energy output is indeed more than the operator furnishes), (iv) power itself and its load simultaneously (all the energy is being input from the external environment at the insertion of excess energy at point D to move the ball **back** to position A), and (v) exhibit negentropy.

This is an analogy to a nonconservative field and a multi-valued potential. In the case where a potential has discontinuous values at a single point, with the value depending upon whether the detecting charge (the "ball", so to speak) is to the left or to the right of that point, one has a *multivalued potential.* This multivalued potential actually represents a change of potential energy in the system, freely occurring without operator input. In short, it is an asymmetrical self-regauging, violating Lorentz's symmetrical regauging condition.

Actually, multivalued *magnetic* potentials arise naturally in magnetics theory, but —foolishly, in our view — theoreticians do all in their power to minimize or eliminate their consideration {426a-k}. They consider such a nonlinear change as being embarrassing and troublesome, and to be gotten rid of at all costs! However, if deliberately used and optimized, rather than eliminated, incorporating a multivalued magnetic potential can provide a *nonconservative magnetic field* (analogous to the illustration), where $\int \mathbf{F} \bullet \mathbf{ds} \neq 0$ around a rotary permanent magnet loop. The multivalued potential represents a broken symmetry that further produces a nonconservative field.[184] In theory, such use of the multivalued potential and the resulting nonconservative field can enable a "self-powering" permanent magnet rotary engine, operating as a negative resistor freely extracting and using magnetic energy from the broken symmetry in the system's energetic exchange with the active vacuum.

However, note that the multivalued potential represents a point of sharp "self-regauging asymmetrically" by the circuit, with the regauging potential energy coming from the external environment. It requires that some external process in the exchange between environment and system must be automatically invoked at the multiple-value point, so that there is a sharp and sudden entry of excess potential energy received by the circulating ball (or by the circulating charges in an electrical circuit, or by the circulating flux in a magnetic circuit).

One such means of evoking such a sudden surge of excess energy at a point, momentarily, is given by Lenz's law, as discussed in the magnetic Wankel engine. Another means is by Johnson's sudden evocation of the exchange force, which momentarily can produce a pulse of energy density even several hundred times as strong as the usual field energy at that point. For example, imagine that the ball in Figure 6-20 represents a magnetic rotor in a magnetic potential represented by the height of the ball above the dotted line base. The point (D-A) then represents the results of the system having suddenly injected an instantaneous exchange force at point D, to freely increase the potential and potential energy of the rotor system back to point A. The rotor is immediately lifted back at its starting potential energy situation at point A, freely, by this sudden and free evocation of the exchange force. We will briefly discuss the exchange force in our discussion of the Johnson engine.

---

[184] In physics, the appearance of a force and its subsequent action to perform work is nature's way of restoring symmetry to a situation where symmetry has been violated. As can be seen, the net force can be used to translate something (such as current) and do work, thereby dissipating the excess potential energy received from the broken symmetry condition.

6.4.3 Dromgoole Effect as an Example of a Novel Magnetic Effect

An example of a novel magnetic effect is the *Dromgoole effect* {427}. This is an interesting phenomenon where a voltage placed on a solenoid wrapped around a longitudinally magnetized iron wire may be increased up to 300 times in magnitude by twisting the wire through 90 degrees. If a scheme can be worked out whereby this dramatic increase in voltage potentializes and moves very substantial current, then perhaps the extra output can — at least momentarily — produce more output energy than the work required to twist the wire.[185] At least that could be a working hypothesis from which to launch experiments to see if it is possible.

The reason this may be possible is that any amount of energy W one wishes can in theory be collected from any finite potential intensity ϕ, according to the simple equation $W = \phi q$, where q is the collecting charge exposed to and interacting with the potential intensity ϕ. In the case of magnetics, q is analogous to the pole (magnetic charge), where the north pole is positive magnetic charge. The ϕ then becomes the magnetostatic scalar potential.

So if we can produce ϕ with only a little expenditure of energy, and then have that ϕ potentialize a very large amount of charge q, we can collect upon the charge q much more energy than what we ourselves expended. The ϕ produced is a change to the local vacuum potential, and hence the collection of energy W on charges q is actually a collection of EM energy from the local altered vacuum potential itself. Any potential we make becomes a change to the local vacuum potential, and hence a change to the local active vacuum. That does not seem to appear in electrical engineering, which model does not even incorporate modeling the active vacuum or its potential, much less a change to it.

---

[185] The scheme must prevent *at least an appreciable fraction* of the spent electrons-from the external circuit and load — from being forcibly rammed back up through the coil against its back EMF. Otherwise the Lorentz symmetry condition applies and the arrangement will not produce COP>1.0.

While we discuss a few interesting magnetic effects in this paper, there are many more. The interested researcher is referred to such easier sources as Burke {428} and Cullity {429}. For more complex scientific sources and explanations, other publications are available {430a-430r}. In addition, it is helpful if the researcher is aware of some of the foundations problems in physics and electromagnetics {431a-431e}. Our point is that there are more than 200 known effects in magnetics, and only about half of them are well understood. For the other half, the understanding ranges from "partially understood" to "not understood at all." The latter half of the magnetic effects provides a rich ground for investigation by researchers seeking an asymmetrical self-regauging mechanism.

The COP>1.0 researcher must be prepared for an extended self-education period, and appreciable study and work. Some research discipline is highly recommended, such as starting one's own database and rigorously maintaining it up to date with one's latest interests. For the experimenter, a good lab notebook, meticulously kept and regularly posted, is an absolute necessity. Reading and searching the scientific literature is also highly recommended. It is not as simple as applying the principles one learned in university or technical school. Those techniques and principles are involved, but at some point in the circuit they must also be violated. Else COP>1.0 systems would long ago have been developed and marketed by sharp young students, graduate students, and post-doctoral scientists.

6.4.4 Exchange Force and Johnson's Approach

Refer again to our discussion of conservative versus nonconservative fields, above.

If one wishes to build a self-powering rotary permanent magnet motor, obviously one must evoke a condition whereby $\int \mathbf{F} \bullet d\mathbf{l} \neq 0$ around a single full rotation of the rotor (and where F is the instantaneous magnetic force between stator and rotor). For permanent magnets without any external switching, etc., that condition can only arise if an excess force is freely triggered by changes occurring freely in the magnetic materials themselves.

Fortunately, magnetic materials are almost always highly nonlinear, and **so** such self-triggering opportunities abound. They can occur because there are a multiplicity of mechanisms (dozens and dozens) going on simultaneously in the materials of the magnets of both the rotor and the stator, and not just the standard polarity and field repulsion and attraction that electrical engineers use. These materials processes also involve direct and ongoing exchange of energy between the active vacuum and the local materials in the magnet.

The trick is to take one of these automatically ongoing processes that is already there and freely available but just not adroitly used, and deliberately evoke and enhance it to induce excess energy at the appropriate time, with a force in the appropriate direction. In simple terms, if you have eight more horses available to put into the harness to pull the load, and you only have two normal horses but wish to pull more load, then by all means you should hitch up and use at least one more additional horse, and take him from the six that are usually just standing by and not working at all. They may run back and forth a bit, while watching the others work, but that of itself contributes nothing.

One such additional force and free regauging mechanism always available is the exchange force, which can be checked in Feynman's three volumes of physics {433} and also in the reference by Cullity {429} previously **cited.** Good science dictionaries {432} also include helpful information. Reading the literature will add enormously to one' data base of effects and methods.

Howard Johnson's approach is to arrange highly nonlinear assemblies of geometrically asymmetrical magnets into complex configurations. At certain points in a magnetic field, these Johnson assemblies will trigger the sudden brief self-generation (actually a sudden release) *of exchange force* between stator and rotor. The exchange force is a known phenomenon {432, 433, 434}, though still not completely understood theoretically. It **can** be evoked asymmetrically by evoking electron-spin flipping, for example, which has led to the development of a new field of technology, **spintronics.** *Microscopically,* the exchange energy is much greater than the magnetic field energy, sometimes by a factor of 10^3 to 10^4 {435, 436}. *Macroscopically,* the resulting exchange energy can momentarily still be many times the magnitude of the ordinary magnetic field energy; e.g., by a factor of 200, which Johnson often achieves.

If **the** stimulated momentary exchange forces are oriented to always add to the thrust upon the rotor in the "powering rotation" direction, then the **multivalued** potential condition and $\int \mathbf{F} \bullet \mathbf{dl} \neq 0$ are met. In that case, self-powering is permitted, if enough of these exchange force triggerings occur so that the extra potential energy added to the rotor is sufficient to overcome the load and the losses being powered.

The self-triggering of each exchange force appearance produces an excess burst of force and energy[186] input into the system from its active supersystem environment. Thus, repeatedly the system is momentarily converted (in each exchange force burst) into an open system in disequilibrium in its energy exchange with its active environment, freely receiving excess energy from it. Because the exchange force "input energy burst" is short, multiple such "input bursts" must be used in a single rotation cycle so that the total energy input by all of them is significant.

In that case, COP>1.0 performance is permitted by the laws of physics, thermodynamics, and nature. Conservation of energy is not violated. Classical equilibrium thermodynamics with its infamous second law does not apply to the Johnson system, since the system is periodically an open system far from equilibrium and receiving excess energy from its active environmental exchange. *A priori* the Johnson system has increased its negentropy overall, and that negentropy (increased order or increased potential energy) can then be dissipated (disordered) to produce free shaft horsepower if the bursts of exchange force are properly coherent in direction and timing.

We diverge for a moment: As is well known, the equilibrium condition in a system is the condition of maximum entropy in the system; any disequilibrium condition reduces the entropy *a priori* because it is an excited state of the system containing additional potential energy. It is worth rigorously clarifying the infamous second law of thermodynamics. Quoting Lindsay and Margenau {437}:

> *"[The]...statement of the second law: (a) the entropy... is a variable of state, (b) Its value, for a closed system, can never decrease."*
>
> *... "Non-equilibrium conditions cannot be specified by variables of state, and their entropy cannot be computed. ...the condition of equilibrium is the condition of maximum entropy."*

---

[186] Technically speaking, the exchange force is not a magnetic field force, but a force that arises independently due to quantum mechanical considerations. Nonetheless, it is a real force arising in magnetic materials and affecting magnetic materials, as in permanent magnets.

In the first statement, the reader should note the assumption of a closed system in the first subparagraph, and the absolute requirement that the calculated entropy be a variable of state.

Also, the quotation can be falsified. Oddly, the "closed system" in classical thermodynamics is defined (illogically) as a system that does not exchange mass across its boundary, but may and often does exchange energy across it. We negate any absoluteness of that statement by Lindsay and Margenau by citing a counter example of a "closed" system with continuously increasing energy, fed by transfer of energy from the environment, and with no escape of the energy or very little escape of it. A specific example is a photon absorption by a mass particle. This system achieves increasing energy (order) as the process continues, hence achieves negentropy. This may be considered a fluctuation, of course, but it still places severe limitations on this law of thermodynamics and in fact negates any absoluteness of it. For such reasons, in this book we have redefined "closed system" as one that exchanges neither energy nor mass with its environment, and we recognize that there are no such systems in the universe. We have defined an "open system" as a system that exchanges either energy or mass or both across its boundary, so that we do not encounter the problem of the counter example cited. Further, general relativity requires an increase in the mass of any system that increases its potential energy, and a decrease in the mass of any system that decreases its potential energy. Hence energy exchange at all with the system, involves mass exchange since mass and energy are the same thing. The classical thermodynamic definition of a "closed system" has thus been falsified since 1915, with the definition becoming only an approximation rather than a generally valid definition.

In the second subparagraph of that first statement by Lindsay and Margenau, the reader should note that the closed-system assumption must be violated *a priori* if the entropy does decrease, and vice versa. If the system is broken into a set of subsystems, then the only way the entropy of the overall closed system to decrease is for one or more of the subsystems to be open (new definition!) and energy (order) to pass out of the system. Then an interesting thing emerges: For order (energy) to remain in the system as such, the subsystems taken as a whole must produce as much negentropy as they do entropy. Energy from an ordered subsystem can be emitted in disordered form, but then it has opened that subsystem and has entered the space between parts (subsystems) of the overall system. In other words, in a closed system, *any increase in entropy requires the subsystems to become open subsystems.* Again, the statement of this law of thermodynamics eats itself. To stay in the overall system, this scattered energy *outside the subsystems* must then interact totally with another part of the system, and so on. This introduces disorder to the succeeding parts that interact. Therefore, the second law of thermodynamics itself *internally* violates its own "closed system" assumption because, to operate at all, the law requires continuing interaction between the active local vacuum environments and the subsystem components. In short, it requires a very special kind of overall or average equilibrium in an unavoidable energy exchange between the local vacuum and all the parts of the system. The source charge problem already demonstrates the universal violation of the second law and the thermodynamic definition of "open system", but both classical electrodynamics and classical thermodynamics have ignored this source charge problem for more than a century. Our solution to it was published in 2000.

Quite simply, there is no such thing as a truly closed system in the first place. Kondepudi and Prigogine come close to this statement in the following quotation[187]:

> *'Anyway, equilibrum thermodynamics covers only a small fraction of our everyday experience. We now understand that we cannot describe Nature around us without an appeal to nonequilibrium situations. The biosphere is maintained in nonequilibrium through the flow of energy coming from the sun, and this flow is itself the result of the nonequilibrium situation of our present state in the universe."*

In short, all systems on the planet — and we ourselves — are immersed in a nonequilibrium state a priori. Rigorously there is no such thing as an absolute equilibrium state on the planet, except as an approximation.

Now consider a perfectly insulated system, so that no heat can pass from the system outside it. An interesting constraint then exists on those "open subsystems" producing disorder (entropy). Unless equal reordering occurs in the subsystem-to-subsystem reactions, then disordering (heat) grows a priori. But this is not observed to happen in well-insulated systems approximating our theoretically perfect example! Otherwise, the temperature of a well-insulated system would increase until system rupture and failure. And experimentally that does not happen.

---

[187] Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, 1998, p. xii.

It follows that, to maintain the internal equilibrium between subsystems and a constant internal temperature, a negentropic process is clandestinely involved. We submit that this process is revealed in our discovery of giant negentropy of the negative charge, and what may be said to be the giant entropy of the positive charge — i.e., in the discovery of the common " 4-circulation" of energy surrounding a dipole from the time domain to the negative charge of a point dipole in 3-space (thereby entering 3-space once emitted by the negative charge), thence to the positive charge of the point dipole, and thence back to the time domain. For a single charge, the well-known vacuum polarization provides virtual charges of opposite sign, to convert the "isolated charge" into a set of composite dipoles, as previously explained.

The second law of classical thermodynamics, considered in a more modern light, appears to conceal hidden giant negentropy and hidden giant entropy, in the ongoing 4-circulation of EM energy in the supersystem. *It is not possible to eliminate the supersystem or the interchange between its parts; particle physics told us in 1957 that there is no equilibrium of any system without this ongoing exchange.* Any thermodynamics attempting to discard the supersystem exchange (which involved both mass and energy) is at best an approximation for special "reasonably well-behaved" situations.

If the entire system is not in net equilibrium with the external environment (i.e., if there exists disequilibrium between the separated parts of the supersystem), then classical thermodynamics does not *absolutely* apply to that system. The system is no longer *absolutely* describable by "variables of state".

Those objecting to COP>1.0 in an EM system on the grounds that it would violate the second law of thermodynamics (which already violates itself), would be well-advised to restudy the very notion of the second law and the thermodynamics definition of *open system.* Compare relativity's equating mass as energy. Then ponder the thermodynamics of open systems far from equilibrium with their active environment. Every system in the universe is open, and it has an ongoing exchange with its proven active environment (local active vacuum and curved spacetime). This exchange Includes and exchange with every particle in the system. As pointed out by Lee:[188]

---

[188] T. D. Lee, *Symmetries, Asymmetries, and the World of Particles,* U. Wash. Press, Seattle, 1988, p. 46-47.

> *"...symmetry implies conservation. Since our entire edifice of interactions is built on symmetry assumptions, there should be as a result a large number of conservation laws. The only trouble is that almost all of these conservation laws have been violated experimentally. "... "...this difficulty could be resolved by introducing a new element, the vacuum. Instead of saying that the symmetry of all matter is being violated, we suggest that all conservation laws must take both matter and vacuum into account. If we include matter together with vacuum, then an overall symmetry could be restored."*

The *system itself is* always in disequilibrium; only the supersystem can exhibit equilibrium. The second law of thermodynamics specifically does not and cannot apply to a system far from equilibrium, because of its implicit assumption of overall equilibrium without the active vacuum exchange. Also, a deeper balance is required between the hidden asymmetries existing between the subsystems and their local vacuum (and local spacetime curvature.

Indeed, *one cannot even calculate* the entropy for a system that — overall — is far from net equilibrium with its active environment. We quote Lindsay and Margenau even more strongly {438}:

> *"Equilibrium states are the only ones that are capable of explicit analysis in thermodynamics... "*

And again {439}:

> "... *variables of state have meaning only if they define an equilibrium state. Hence the quantity we are seeking will be meaningless unless it refers to equilibrium states. "*

While we are at it, let us also address a serious flaw in the first law of thermodynamics. We again use Lindsay and Margenau for a succinct statement of the First Law {440}:

> *"First law of thermodynamics. A complete statement of the first law comprises two assertions: (a) heat is a form of energy, (b) Energy is conserved. "*

All that really says is that energy is conserved. It does not state that it is conserved in an object. It states that, whether the system is in equilibrium or not, energy is conserved. If heat is taken as disordered energy, then it merely states that overall the energy is conserved, whether ordered or disordered with respect to some ordering criterion. It does not state that the disordering is conserved, and it does not state that disordering must increase or decrease. But *it does implicitly assume that all energy at some most basic level is ordered, else it cannot be energy (order).* So it assumes that, at higher levels, energy can be disordered (incoherent). However, at the underlying basic level, it is and remains perfectly ordered — else it could not remain "energy and could not be conserved. As an EM example, in so-called "heat", *every scattered photon retains its perfect order; it is only the photon ensemble that is "disordered".*

In short, each "basic piece" of energy is perfectly ordered, but the *ensemble* of the pieces may be disordered. Therefore, entropy applies only at a level higher than the basic energy quantum. Contrary to the assumptions of classical (macroscopic) thermodynamics, processes which directly engineer the basic energy quanta[189] — more exactly, the *action* quanta, consisting of *energy x time,* since energy cannot be "engineered" or changed in 3-space without also being engineered "in time" as well — are time-reversible. Hence they can be negentropic — simply because every observable system is "open" to, and in continuous energy exchange with, its active time environment (and also its active vacuum environment). Also, no system changes its spatial energy in any fashion, including ordering or disordering, without interacting with spacetime and spacetime curvature dynamics. It also changes its time-energy.

So in our view the notion of "disordering" and "disordering of energy" must be carefully reconsidered, as to exactly what is and is not being disordered, when the assumed "disordering" occurs, at what level it occurs, where and how the compensating reordering occurs, etc. We also point out that the simple discovery of giant negentropy {12} as the solution to the long-vexing source charge problem already removes the "absoluteness" of classical thermodynamics. Giant negentropy already violates the assumptions of classical thermodynamics at the elemental level in every physical system. Indeed, every charge in the universe already falsifies any "absoluteness" of the assumptions of classical thermodynamics.

This problem in the old classical thermodynamics has long been indirectly solved in particle physics, with the discovery of broken symmetry. As Lee states so clearly {441},

---

[189] Actually, energy is discretized, not quantized. Energy x time (i.e., *action)* is **quantized.**

> *"As we expand our observation, we extend our concepts. Thus the simple symmetries that once seemed self-evident are no longer taken for granted. Out of studies of different kinds of interactions we are learning that symmetry in nature is some complex mixture of changing plus into minus, running time backward and turning things inside out."*

We point out that a symmetry involves a conservation law, such as are stated in classical thermodynamics, and a broken symmetry involves a broken conservation law. So the discovery of broken symmetry in physics was a profound change affecting all physics, including the staid old classical thermodynamics. Lee further points out the new complexity of concepts {442} such as symmetry (which is behind every conservation law, including the first law of thermodynamics):

> *"At present, it appears that physical laws are not symmetrical with respect to C, P, T, CP, PT and C. Nevertheless, all indications are that the joint action of CPT (i.e., particle <-> antiparticle, right <-> left and past <-> future) remains a good symmetry."*

So unless the first law is stated in terms of modern CPT symmetry, it does not absolutely apply! Further, every charge is changing time-energy into spatial energy or vice versa. Yet there is nothing about time-energy and its transduction into spatial energy, or vice versa, in the present textbook statements of the thermodynamics. The term "heat" does not refer to the presence of energy at all, but to the scattering (disordering) and escape of energy.[190]

Considering heat as "energy of the system", or "heat energy" of the system, is a grand non sequitur. Rigorously, "heat" refers to the reduction of higher levels of ordering of energy, and since the gist of energy is ordering, reduction of ordering is the very antithesis of energy! "Heat energy" thus is an oxymoron. Before the "escape", there is no "heat energy" (ugh!) in the system at all. The energy is present in the system *not as disordering, but as ordering, a priori*. If it were in the system, it would not have escaped nor would it be escaping from the system. More energetic molecular motion, e.g., is actually more energetic ordering, simply at an excited state (of greater energy!).

---

[190] Think closely: We never take the temperature of a "system"! We take the temperature of the disordered energy (heat) leaving that system or its subsystems. We do measure the effect of the emitted disordered energy. But that has already left the system and is in the local vacuum (a second component of the supersystem). Thermodynamics might be usefully redone more exactly in terms of the supersystem. We leave that task to some budding young future thermodynamicists for a recommended doctoral thesis.

We stress again (and strongly advise the researcher to read) Romer's strong objection to the use of heat as a noun {443}, and we suggest that the entire subject of classical thermodynamics needs a thorough revision to tighten up its terminology, correct its definition of closed system, eliminate its conflict between the first and second laws, and remove its inappropriate consideration of heat as "energy". Otherwise, the presentation and general interpretation of thermodynamics itself will continue to be one of the great confusion factors one encounters in trying to think clearly about extracting EM energy from the active vacuum environment to produce and utilize **COP>** 1.0 systems.

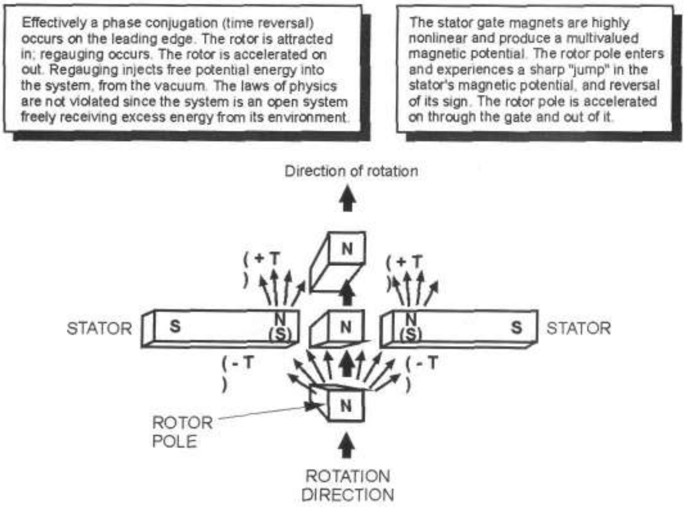

Effectively a phase conjugation (time reversal) occurs on the leading edge. The rotor is attracted in; regauging occurs. The rotor is accelerated on out. Regauging injects free potential energy into the system, from the vacuum. The laws of physics are not violated since the system is an open system freely receiving excess energy from its environment.

The stator gate magnets are highly nonlinear and produce a multivalued magnetic potential. The rotor pole enters and experiences a sharp "jump" in the stator's magnetic potential, and reversal of its sign. The rotor pole is accelerated on through the gate and out of it.

Figure 6-21 Concept of operation of Johnson's patented magnetic gate.

Johnson's approach produces a peculiar kind of multivalued potential. One **part** is conservative and of magnetic energy in nature, and the other part is nonmagnetic, being direct exchange force and energy effects on the **participating** atomic nuclei arising from quantum mechanical considerations. Nonetheless, the net *hybrid* potential can be multivalued at various points around the stator, and that is all that is required for broken symmetry. If $\int \boldsymbol{F} \cdot \boldsymbol{dl} \neq 0$ around a closed path, regardless of which set of forces F represents, then that overall composite force field acting upon the rotor or object moving around the closed path is nonconservative. In that case, self-powering is permitted.

Figure 6-21 diagrammatically illustrates the operation of a nonconservative force-producing magnetic gate in Johnson's approach to a permanent magnet motor. As Johnson has shown, by using a multivalued hybrid potential (MVP) in his gates so that the resulting excess exchange force is properly oriented in direction and timing, a rotor magnet is attracted into a highly nonlinear stator gate region where the MVP is located. When it enters the MVP region where the exchange force is initiated, the rotor encounters a dramatic jump in the net potential with a change of polarity as the exchange force suddenly fires automatically. In turn, this produces a sudden accelerating tangential force in the region, usually against the prevailing magnetic force in the normal back mmf region!

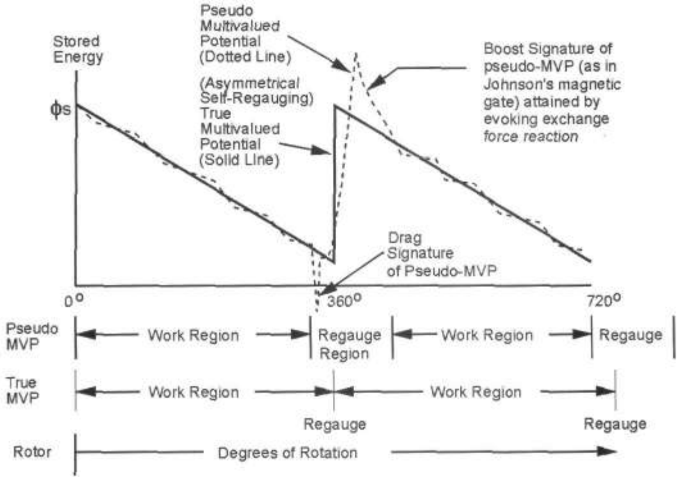
Pseudo Multivalued Potential (Dotted Line)
Stored Energy
(Asymmetrical Self-Regauging) True Multivalued Potential (Solid Line)
Boost Signature of pseudo-MVP (as in Johnson's magnetic gate) attained by evoking exchange force reaction
ϕs
Drag Signature of Pseudo-MVP
0°
360°
720°
Pseudo MVP
Work Region
Regauge Region
Work Region
Regauge
True MVP
Work Region
Work Region
Regauge
Regauge
Rotor
Degrees of Rotation

Note Single regauging per notation shown Johnson uses multiple regaugings per rotation

Figure 6-22 Johnson's "refueling" a permanent magnet engine by asymmetrical **self regauging. A multivalued potential is momentarily created by precisely evoking an exchange force reaction.**

Johnson often confounds professors of electromagnetics by showing them that his special gate can produce attraction between like poles, with a north rotor pole being drawn into a north stator pole in that exchange force region. This can be understood by tracing the spin field flow patterns, but cannot be understood by simple "like poles repel and unlike poles attract". In short, Johnson knows he is adding a force completely different from the magnetic force the professors have in mind as the sole acting force. It is rather like pushing two like poles of two little magnets together. You can do it because an additional force — your mechanical force pushing them together — is acting and overcomes the normal repulsive force between the two poles.

The sudden exchange force burst produces a sudden accelerating tangential force in the normal deceleration (back mmf) region of the stator-rotor arrangement. In short, Johnson utilizes a deliberately evoked burst of exchange force to freely overcome most or much of the back mmf (back-drag) region. This in turn produces a net driving force around the rotor's complete rotation, because the overall field combination is nonconservative.

Rigorous force meter measurements taken at 0.01-second intervals prove that this occurs as the rotor enters and passes through Johnson's gate. The results of one such experiment are given in Figure 6-22.

Johnson thus uses highly nonlinear magnet assemblies of novel design to create a special kind of multivalued potential in his magnetic gate by adroitly evoking exchange forces — *extra quantum mechanical forces* in addition to the usual magnetic field forces. The MVP produces a sudden extra rotational force, backwards to the back-drag (back mmf) region of the stator-rotor magnet arrangement where the rotor magnet assembly is entering and passing through the conventionally repelling like-polarity situation. The result is that a net north rotor pole is seemingly attracted in[191] to the otherwise repelling stator north pole region, then accelerated out the other side by the normal magnetic repulsion.[192] Net drive force results, instead of a net zero force. Engineer Ken Moore, a close colleague of the present author, independently tested a Johnson gate and verified the net driving force effect {444} within the limits of our little laboratory.

---

[191] In reality it is "forced in" against the repulsive magnetic force, by a superior force of a different nature, suddenly evoked.

[192] I.e., as the exchange force dies, the forces now acting are the normal repulsive magnetic forces, but the rotor has moved to a position where now this repelling force accelerates the rotor in the "load powering" rotational direction. In other words, the observer sees a like pole suddenly "attracted" into a like polarity region, move partially through it, and then suddenly be accelerated out the other side. But the "attracting in" force and the "accelerating out" force accelerate the rotor in the load-powering rotational direction.

Technically, Johnson's apparatus asymmetrically and suddenly regauges its exchange forces in such a manner as to momentarily overpower the back mmf region of the engine's rotation, thus producing a net multivalued potential and a net accelerating force around the overall rotation loop. However, correlated distant effects in magnetic materials are known. If distant compensation does occur, then the closed-loop capability might conceivably be defeated, if the total compensation is sufficient to provide an $\int \mathbf{F} \bullet \mathbf{dl} = 0$ overall condition. Hence the research probably demands attention to maintaining both *local* asymmetry and *distant* asymmetry. If the distant correlation effects occur, only in that manner can an overall nonconservative field represented by $\int \mathbf{F} \bullet \mathbf{dl} \neq 0$ be maintained around a completely closed magnetic loop.

This may be the problem with which Johnson appears to have been struggling for so long. He did solve the problem once, and produced a self-rotating permanent magnet motor that the present author personally tested over a period of about 2 hours. During a two-hour test period, the motor apparatus was allowed to self-rotate for up to 15 minutes at a time. Shortly after it became known that Johnson had successfully produced such a prototype, his laboratory was mysteriously broken into. The only thing taken was the magnet assemblies on that specific little successful demonstrator — in a laboratory with more than a hundred thousand dollars worth of magnets and magnetic assemblies of many kinds present.

As a personal aside, I have been associated with Johnson, off and on, for many years. He is a man of the highest integrity and an indefatigable researcher. It is my fervent hope that Johnson will succeed in his permanent magnet rotary motor project. All I can say as this goes to press, is that he is getting very close again, now that he has help and good machining accuracy, etc. for higher precision buildups. In the near future, I hope to be among the first to drink a good glass of champagne to celebrate his success.

6.4.5 Radus Effect: Magnets with Preferred Flux Path Memory

In the early NASA space program, magnetic boots for holding the astronaut to the skin of the space vehicle when outside it were required. A Westinghouse engineer named Raymond J. Radus had found that a permanent magnet applied to a dual flux path (Figure 6-23) would "remember" the flux path taken initially, until switched to the other flux path in an initial condition, after which it would remember that path as preferable {445a-445c}. A permanent magnet with such "memory" will split its flux in two parallel flux paths by some ratio, e.g., 20% in the non-preferred path and 80% in the preferred path. This preference can be switched at will by proper manipulation. So Radus developed switchable magnetic astronaut boots for the space program, using his magnetic memory and flux path switching technique {446}.

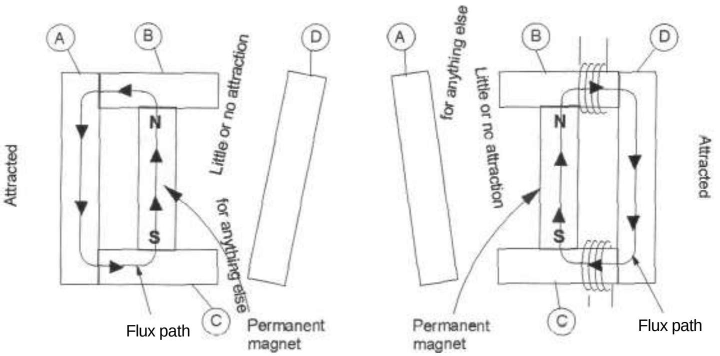

Figure 6-23A.
Once the magnetic circuit is made, there is little attraction for anything else

Figure 6-23B.
A single pulse through a coil can reverse the direction of the flux path

Figure 6-23 One example of the Radus effect.

The Radus circuit differs from ordinary permanent magnet circuits in three ways: (i) It exerts strong magnetic force at one end while exerting hardly any at the other, (ii) the strong and weak magnetic poles can be switched end-for-end easily and at will, thus, in effect, turning the magnetism on and off at one given side of the circuit, and (iii) once switched, it remembers its direction of greatest magnetic pull indefinitely.

The original Radus boots were excellent. For the acceptance tests, an engineer clad as an astronaut walked across the bottom of a steel beam in a high bay research area, upside down against the pull of Earth's gravity (Figure 6-24). He *stepped as* he walked, putting his foot "down" and then picking it "up" {447}.

There is no problem in finding modern magnets strong enough to hold the astronaut firmly in such an upside position. The problem with simple magnetic boots using such strong magnets is that, once the foot is planted, the astronaut cannot pick up the foot again. Consequently, the best the astronaut can do with such magnetic boots is walk rather laboriously by "scooting" the foot forward, with the boot remaining in contact and the astronaut unable to pick it up.

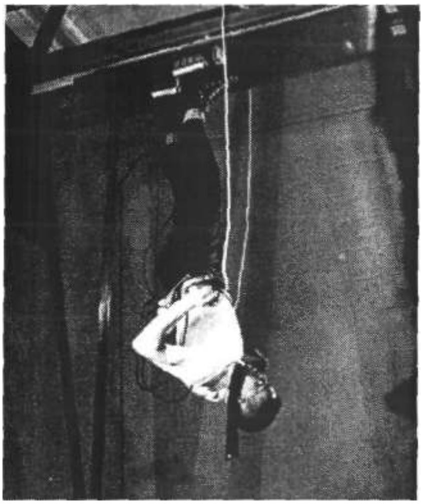

Engineer *walked* upside down on the underside of a steel beam in a high-bay area, wearing the Radus boots He used a stepping movement, not shuffling

Figure 6-24 Acceptance test of the Radus astronaut boots.

The Radus boots completely solved that problem. The permanent magnet fields are switched away from the "boot sole contact with the beam" for that foot that the astronaut wishes to lift, so he can lift it easily and take another step. Then the fields are switched back in again so that the strong field is on the boot sole as he places his foot down. This switching of the fields allows him to walk in a manner resembling normal walking, though a little slower. To do that switching by normal "battery and electromagnet coils" would be prohibitively bulky and heavy — and awkward. With the Radus boots, the astronaut could walk in a manner very similar to slow but normal walking.

However, it is easily seen that an adaptation of the Radus process could conceivably be used to produce a self-switching, self-powered permanent magnet motor. Being a permanent dipole, a magnet is already a particular kind of "free energy generator", since it continuously extracts magnetic energy directly from the active vacuum due to its dipole asymmetry in the virtual photon flux of the vacuum.

The Radus boots were abandoned quickly. NASA then developed the present inferior "shuffler" kind of magnetic boots where the astronaut cannot lift his boot from the surface, but must "scoot" his feet along in a sliding and painfully awkward fashion {448}.

Even today, it is little known that in many virgin magnets fresh from the factory, their very first use conditions them with a Radus-type memory! *That fact can be used, e.g., to create magnets whose fields appear normal, but which deviate from the normal behavior of ordinary magnets, and which produce anomalies in their magnetic fields.*

The Kawai engine {31} uses a novel means of switching the flux path itself instead of overpowering the magnetic poles in an electromagnet. As this is written, Bedini has been notified that his patent with a "path-switching" effect control method has been accepted and will be issued. Although control of the Kawai engine was seized by the Yakuza, we expect to see that some of the first COP> 1.0 motors on the world market will be magnetic motors using the Radus effect or other flux path switching method.

6.4.6 Hole Currents and Electron Currents

In a conducting solid, there are both hole currents and electron currents moving in response to a forcing electric field. In the simple case, electrons move in one direction along the conductor and lattice holes move (migrate) in the other. However, when additional fields are added, the situation can become very complex. The holes and electrons can move oppositely or together in the same direction, depending on the arrangement of the fields and voltages. One can even use fields at right angles to the conductor to affect the currents, and even stop them. This is a rich subject for the inventor, and one that does not appear to have been adequately explored for potential energy and power purposes. This is particularly true when the conductor material is bent or shaped, etc. This area is so broad that it will not be treated here; instead, we refer the reader to Burke {449} for a basic idea of some of the effects and current combinations that can be obtained. From there, an interested experimenter should search the materials science literature on electron and hole current responses of materials.

6.4.7 The Magnetic Wankel Engine

For baseline comparison, Figure 6-25 shows a conventional linear magnetic motor. As can be seen, due to the difference between magnetic poles, a magnetic field exists along the line of the linear track, from the end having the magnets separated the least to the end having the magnets separated the most.

Figure 6-26 diagrammatically shows the scheme of operation of the magnetic Wankel engine. It is basically a linear magnetic motor, wrapped into almost but not quite a complete circle. A set of permanent magnets, each at an angle to the various radial lines of the device, comprises a slightly widening spiral stator that forms most of a circle. A circular rotor with a sector magnet is mounted inside this spiral stator. An end gap exists in the stator as shown, so that the stator is not a completely closed ring. The direction of rotation of the rotor is clockwise as shown. For demonstration of the principle, the beginning air gap between rotor and stator is 0.1 mm and the ending air gap is 5 mm.

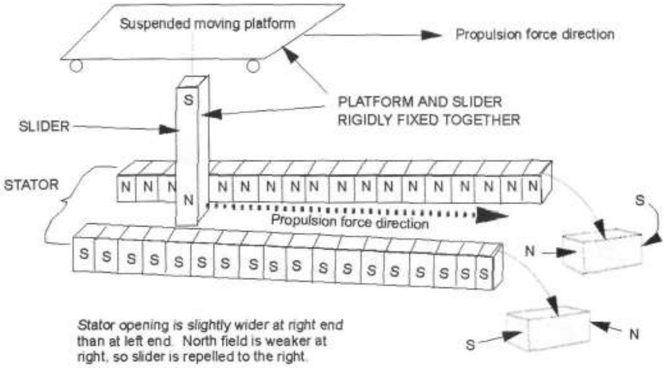
Suspended moving platform
Propulsion force direction
PLATFORM AND SLIDER RIGIDLY FIXED TOGETHER
SLIDER
S
STATOR
N N N
N
N N N N N N N N N N N N N N N
Propulsion force direction
S
N
S S S S S S S S S S S S S S S S S S S
Stator opening is slightly wider at right end than at left end. North field is weaker at right, so slider is repelled to the right.
S
N

Figure 6-25 Type of conventional linear magnetic movement device

An electromagnet is mounted along the perimeter of an angular sector of the rotor, completing the stator gap. The electromagnet is weakly magnetized by a weak trickle current in the associated coil, furnished by a coil with a make-and-break magneto point gap. The electromagnet is magnetized, say, with the north pole facing radially outwards, and the south pole facing radially inside. In the stator, the permanent magnet north poles are facing radially inward toward the rotor, but at an angle, and the south poles are facing radially outward but at an angle.

Tangentially the north pole of the rotor is in a nonlinear magnetic field, and it will experience a clockwise force and acceleration from position 1 where the air gap between rotor and stator is the 0.1 mm minimum, to position 2 where the air gap reaches the 0.5 mm maximum.

If this were all that was involved, the engine would not produce COP> 1.0 because the tangential field is conservative unless additional *free* energy is introduced to overcome the back mmf region in the stator gap. When the rotor crossed the end gap in the stator between point 2 and point 1, very sharp and dynamic magnetic braking due to the back mmf would be done back upon the rotor magnet by the field of the stator magnets at position 1.

Since any real machine will have at least some friction and drag, the actual COP would be less than 1.0.

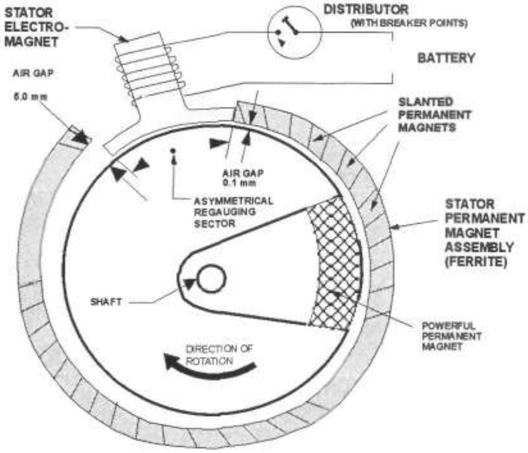
STATOR ELECTRO-MAGNET
AIR GAP 6.0 mm
DISTRIBUTOR (WITH BREAKER POINTS)
BATTERY
SLANTED PERMANENT MAGNETS
AIR GAP 0.1 mm
ASYMMETRICAL REGAUGING SECTOR
STATOR PERMANENT MAGNET ASSEMBLY (FERRITE)
SHAFT
POWERFUL PERMANENT MAGNET
DIRECTION OF ROTATION

figure 6-26 Magnetic Wankel engine with asymmetrical regauging section.

Let us use the notion of the magnetostatic scalar potential (roughly, magnetic pole strength) to examine a new situation in the end gap.

Technically, let us regard a single unit north pole in the rotor, going from position 1 to position 2 (the acceleration cycle, where the engine will deliver shaft horsepower against a load), and going from position 2 to position 1 (where the magnetostatic scalar potential must be suddenly rcgauged asymmetrically to equal or exceed the potential at position 1, in order for the rotor to continue unabated or with even further acceleration. That is, when the rotor enters the "back mmf' end gap between position 2 and position 1, a sharp and sudden increase in the "stator magnetostatic scalar potential" must be accomplished, so that the potential in that region is equal to or greater than the potential at position 1. This effect, nearly freely obtained, is what is required for a self-powering magnetic Wankel engine.

In normal machines, conventionally this asymmetrical regauging part of the cycle is where the design engineer forcibly inputs energy from outside the system to do brute physical work on the rotor to forcibly "reset" the machine, and to forcibly wrestle its potential energy storage back to initial conditions. In short, the operator himself arranges to furnish all the excess energy from outside the system that is required to brute force "regauge" the potential at that point, thus effectively creating a multivalued potential instead of a single valued potential. A multivalued potential achieved only by the operator himself furnishing the extra potential energy will not produce COP> 1.0. It will in fact produce $\int \mathbf{F} \bullet \mathbf{dl} \neq 0$ of the motor section, but only at the expense of extra energy that the operator himself had to input and pay for. In that case and in a real system with some system losses, the net work out because of $\int \mathbf{F} \bullet \mathbf{dl} \neq 0$ will still be less than the total energy input by the operator.

The forcible "reset" work is conventionally done by simply overpowering the field and reversing it (building it up equal and opposite in the other direction), but with energy totally input by the operator and not "free energy input from the environment" at all. The operator first pays to "kill" the existing field, and then pays to establish a field in the opposite direction.

To obtain COP>1.0 and self-powering, we must trick something else or some other process into furnishing that asymmetrical regauging energy — or most of it — for regauging of the magnetic Wankel engine in that stator gap zone. In other words, instead of engaging in the conventional wrestling match against the back mmf, we must let something else provide most of the energy for the wrestling.

During rotation of the stator from position 1 to position 2, we have been maintaining (and paying a little for) a tiny trickle current and small voltage from the battery into the coil around the electromagnet. As the rotor enters the stator gap, suddenly a sensor sharply breaks the distributor points, momentarily inducing a sudden powerful voltage in the coil. With a very short delay, a very sharp surge of current appears in the coil, producing a sharp and suddenly increased magnetostatic scalar potential (pole) in the gap region. That is the "multivalued potential" effect, where we pay a little to achieve it suddenly at that point, by invoking Lenz's law.

The effect is that suddenly the rotor is raised to the same or greater magnetostatic potential as exists at position 1, and "almost freely though not quite". If equal, the rotor suddenly is in a region with no back mmf, hence it experiences no deceleration braking. If the sudden potential is greater than the potential at position one, the stator in this normally back mmf region now actually experiences a further acceleration (a forward mmf) in that region.

Note that no *radial* mechanical work can be done on either the electromagnetic pole piece (part of the stator) or on the rotor, since neither the stator nor the rotor can move *radially.* However, there momentarily exists a clockwise circumferential magnetic field on the rotor in the stator gap, due to the gradient between the sharply regauged pole piece magnetostatic potential and the potential at position 1.

So we pay a little energy continuously (tiny trickle current, sharply broken breaker points) to get much more energy density momentarily in that small back mmf gap region only. The former back mmf in the stator gap is sharply eliminated by the Lenz force and converted to forward mmf. The rotor experiences a continuous acceleration throughout a complete rotation, due to the judicious use of an artificially induced multivalued magnetostatic scalar potential.

If the average shaft power output during the complete rotation cycle is made greater than the average power input to the asymmetrical regauging circuit during that same rotation cycle, the engine will produce COP>1.0. This type of engine is also easily close-looped, since the excess output is **not** electromagnetic energy but mechanical shaft rotation energy.[193] Hence the problem of the Dirac sea hole current (discussed in Chapter 9) is eliminated.

Such engines have been built and placed in an automobile to power it, in Japan {450}, though there is no information on the exact overall COP. The design was lighter and smaller than a gas engine of the same power, and it was a pygmy when compared to other electric engines of similar power. **The** prototype 45-hp unit weighed 155 pounds compared to 440 pounds for a comparable electric motor. The rotary engine was compact enough to fit inside a two-foot cube. The engine was in development by Kure Tekko, a sizable firm that supplies auto parts to Toyo Kogyo, the Mazda maker. To my knowledge, no hard data on the input electrical power utilized for the trickle current and current-breaker has been made available by the Japanese. The principles, however, are quite clear and easily analyzed.

This may be one of the Japanese COP>1.0 engines suppressed by the Yakuza.

---

[193] For example, a geared or belted arrangement can be used to drive a very small but efficient DC generator that replaces the battery. Many other efficient arrangements are possible.

The researcher might like to consider using a strategically-placed Wiegand wire sensor (discussed in the next section) as the free "generator" providing a pulse of electrical energy to the pulse-magnetizing coil in the stator gap region at the precisely appropriate time. If that or some other similar "self-furnished" pulse of sufficient power can be delivered to the pulse-magnetizing coil, then the system would self-initiate a multivalued magnetic potential in that gap region. In that case, the driving magnetic field around the loop need not be conservative, and self-rotation would be possible without violating any laws of nature or electrodynamics. At least a small toy demonstration model might be possible, simply to illustrate the principle. Presently we know of no one who has tried it.

6.4.8 The Wiegand Effect

If a Permaloy® or other suitable magnetic wire is properly tensioned and worked by repeated torsion while under tension, the skin of the wire is work-hardened and caused to have different magnetic characteristics from that of the internal core. This type of wire is called a *Wiegand wire*, or *pulse wire*, or "self-nucleating magnetic wire" (SNMW™).

When the pulse-wire experiences a certain level of ambient magnetic field strength, it will automatically switch its magnetic state, very sharply, producing a sharp magnetic field pulse. The effect is generally called the *Wiegandeffect.*

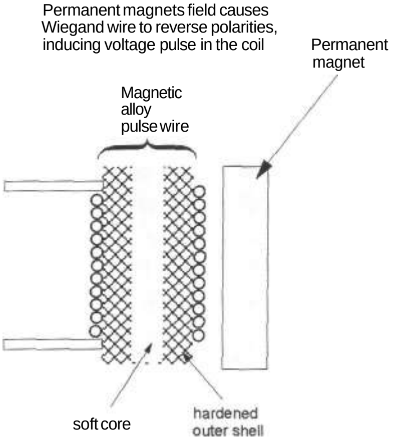
Permanent magnets field causes Wiegand wire to reverse polarities, inducing voltage pulse in the coil
Permanent magnet
Magnetic alloy pulse wire
soft core
hardened outer shell

Figure 6-27 The Wiegand effect

If a coil of many turns of fine electrical conductor is wound around the Wiegand magnetic pulse wire, then when the wire suddenly alters its magnetic state, the sudden magnetic field pulse "cutting" the conductor bundle will produce a sharp electrical pulse. Thus the apparatus produces a sharp pulse of electrical energy, when the ambient magnetic field intensity reaches the pulse-initiation value. No input of outside energy is necessary. The apparatus simply gates the energy from the vacuum when it sharply changes its magnetic polarity.

Wiegand effect sensors may achieve voltage pulses up to a nominal 12 volts, in a typical application.

Numerous patents were issued to Wiegand prior to his death {451a-451h}.

Figure 6-27 diagrammatically shows the primary parts of a typical Wiegand sensor application, where the Wiegand wire is also known as a *pulse wire.* The operation of the pulse wire itself is shown in Figure 6-28. Figure 6-29 shows a typical rotary Wiegand effect pulse transmitter.

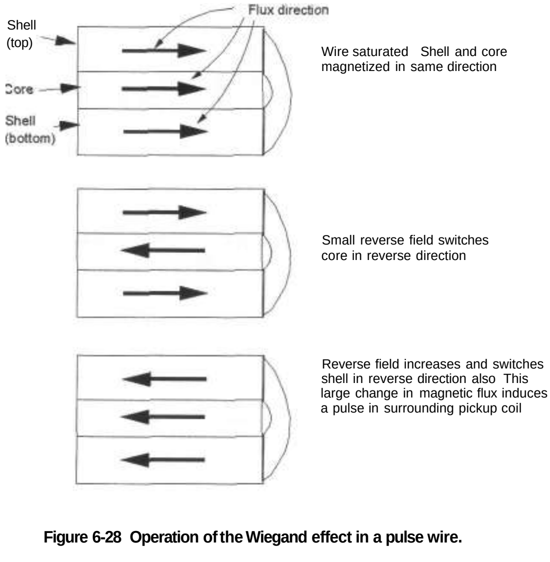
Flux direction
Shell (top)
Core
Shell (bottom)
Wire saturated Shell and core magnetized in same direction
Small reverse field switches core in reverse direction
Reverse field increases and switches shell in reverse direction also This large change in magnetic flux induces a pulse in surrounding pickup coil

Figure 6-28 Operation of the Wiegand effect in a pulse wire.

Pulse wires find many modern "switching" uses, since one does not have to connect them to a power supply in order to obtain the electrical signal pulse generated in a magnetic field as it changes. The magnetic experience of **the** wires can be arranged (using a "resetting magnet") to automatically **reset** the pulse wire to its original condition after the initiation magnet has **fired** it. Many identification cards for personal access to restricted facilities use these wires. Some magnetic memories use the effect. At least one European automobile firm has even used Wiegand wires to trigger the timing of automotive ignition systems.

The pulse characteristics created internally by the Wiegand wire represent a sharp, asymmetrical, self-regauging of the magnetics. The combined magnetic field consisting of the ambient magnetic field and the Wiegand wire magnetic pulse is momentarily a nonconservative magnetic field. It follows that a closed-loop integration $\int \mathbf{F} \cdot \mathbf{dl}$ around some paths in the combined magnetic field of the system do not sum to zero.

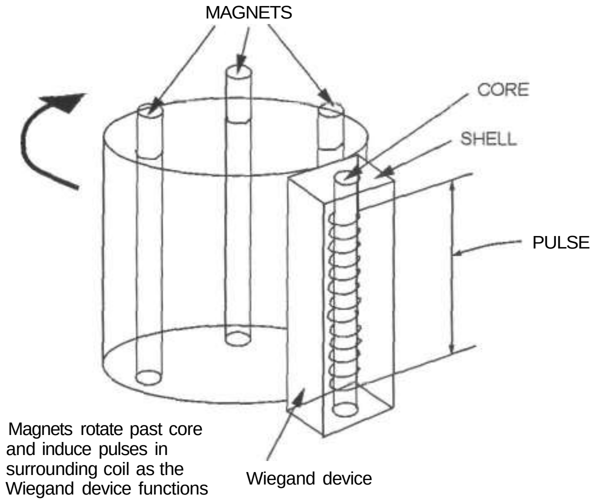
MAGNETS
CORE
SHELL
PULSE
Magnets rotate past core and induce pulses in surrounding coil as the Wiegand device functions
Wiegand device

Figure 6-29 Operation of the rotary Wiegand pulse generator

In theory, at least, it should be possible to utilize something like the magnetic Wankel engine together with exactly self-triggering and self-resetting Wiegand units to form a self-powering permanent magnet motor. In short, referring to our discussion of the magnetic Wankel engine, it might be possible to use a special Wiegand sensor and pulse generator to furnish the properly timed pulse to the coil that enables the continuous operation of the Wankel engine. Obviously this type combination would have occurred to the Japanese scientists working on the magnetic Wankel engine.

We accent that the resulting system would be an open dissipative system, far from equilibrium in its energetic exchange with the vacuum during the Wiegand pulsing. The disequilibrium is achieved by the momentary self-regauging of the magnetic gap in the magnetic Wankel engine.

My colleagues and I have experimented with Wiegand wires and Wiegand sensors to a limited extent. The effect is quite real, fully documented in the scientific literature, and the wires and sensors are not too expensive. A typical coil of 1,000 turns on a 3-cm Wiegand wire will produce pulses of about 2 volts in a 1,000-ohm resistor. The pulse width (half maximum height) is about 20 microseconds. Essentially the Wiegand wires are immune to stray magnetic fields. Viewed on an oscilloscope, the pulses are very clean without spurious oscillations or hash. The field required to switch a typical Wiegand wire is typically about 150 Oersteds. The resetting field is quite a bit smaller, being about -20 Oersteds. With further work, it is probably possible to come up with Wiegand wires whose coils produce up to 20 Volts in a 1,000-ohm resistor. If 25 of these 20-volt Wiegand pulse generators wired in series could be induced to fire simultaneously across a 1,000-ohm resistor, something like a momentary half-watt device could be built. That might be enough to operate a very small version of the magnetic Wankel engine, etc. purely for demonstration purposes.

For practical power, of course, one would need to find a way to increase the power output in each Wiegand generator pulse. So far, no one has been able to successfully find a way to do it, at least to our knowledge. If the reader finds a way to do that, then the reader should certainly patent the process and use it in a self-powering magnetic Wankel engine!

So to those experimenters wishing to experiment with something relatively inexpensive, we would suggest investigating the possibility of incorporating Wiegand units into a magnetic Wankel-type small magnetic motor, and attempt to get a self-powering little unit, or try to build a self-powering version of the rotary Wiegand pulse generator. It is at least possible in theory, and if achieved in practice it will prove that self-powering EM engines are perfectly possible.

6.4.9 Kawai Path-Switching Motor

Japanese inventor Teruo Kawai has invented a process for adroit self-switching of the magnetic path in magnetic motors, rather than directly overpowering the magnetic field and then forcibly reversing it. The process reduces most of the back mmf in an otherwise rather conventionally switched magnetic motor, and results in producing a COP that is approximately double the stated efficiency rating of the unmodified motor. Understand, the Kawai motor always has an efficiency less than 100%, even when its COP>1.0! Hence if one starts with a *high efficiency* magnetic motor — e.g., efficiency of 0.7 or 0.8 — one can obtain a modified motor having a COP of 1.4 to 1.6. Two Kawai-modified Hitachi high-efficiency magnetic motors were tested by Hitachi engineers, and the tests did indeed show COPs of 1.4 and 1.5, respectively, under rigorous and totally independent testing conditions.

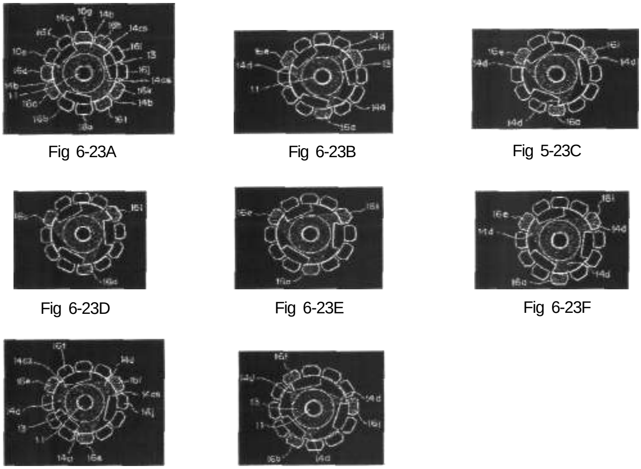

Fig 6-23A

Fig 6-23B

Fig 5-23C

Fig 6-23D

Fig 6-23E

Fig 6-23F

Fig 6-23G

Fig 5-23H

Figure 6-30 The Kawai path-switching process.

Figure 6-30 shows eight snapshots A, B, C, D, E, F, G, and H of the rotor advance of a typical Kawai engine, taken from Kawai's 1995 patent {452}. This is one end rotor/stator side of a two-rotor device, where a similar rotor/stator device is on the other end of the central shaft *11.* In snapshot A, pole piece *14* had three outward teeth *14b* dispersed equally around the circumference, alternated with three notches. An end magnet *13* provides the source of flux passing through the pole piece. With the electromagnets DC-energized, their core materials *16c, 16d, 16g, 16h,* and *16k, 16l* are shown shaded by flux from central magnet *13* out through teeth *14b.*

In snapshot B, electromagnets *16a, 16e,* and *16i* are energized. The shaded area shows the sharp convergence *14d* of the flux from magnet *13* through pole piece *14* and the edge of teeth *14d.* Since the electromagnets are magnetized in attracting mode, the rotor will experience a torque tending to widen the flux path from magnet *13* to the activated electromagnets. Thus a clockwise torque exists on the rotor, and it will start to rotate clockwise.[193] Note that each electromagnet is operating independently of the other two.

As shown in snapshots C, D, E, and F, the rotation of the rotor continues clockwise, widening the connecting flux path to the three activated electromagnets. During this time the torque on the rotor is clockwise.

In snapshot G, the flux path to the activated electromagnets is fully widened. In addition, the leading edges of the three teeth are just beginning to enter the domains of the next electromagnets *16j, 16b* (see snapshot A) and *16f.* This is getting symmetrical to the original position shown in snapshot B.

Now the electromagnets *16i, 16a,* and *16e* of snapshot G are deactivated, as shown in snapshot H, and electromagnets *16j, 16b,* and *16/*are activated. This action asymmetrically self-regauges and resets the engine back to the original situation shown in snapshot B, but one more electromagnet beyond. The action cycle begins anew. As can be seen, in each complete rotation of the shaft, each of the three teeth of the rotor will be asymmetrically self-regauged multiple times. So multiple total asymmetrical self-regaugings, resettings, or "refuelings" are utilized per shaft rotation.

In each motor coil, at energization a tooth is just entering that coil, Energized in attractive mode with respect to the ring magnet around the shaft, the flux in the pole piece jumps from fully widened flux (and small **or** vanishingly radial torque on the rotor) to angled and narrowed flux (with full radial clockwise torque on the rotor). This is an asymmetrical self-regauging action. As previously explained, the narrowed flux and its angle exert a clockwise accelerating tangential component of force upon the rotor. By the asymmetrical self-regaugings of the engine, each coil is de-energized prior to beginning to exert radial back emf— which it would do if it remained energized as the trailing edge crossed it and again narrowed the flux path).

So the Kawai engine uses normal magnetic attraction to accelerate the **rotor** for a small distance, and then asymmetrically self-regauges to zero attraction to eliminate the back-mmf drag portion of the attractive field. It regauges to zero as the reset condition.

---

[193] The Kawai patent contains a small error of misstatement. It states at first that the torque and rotation are counterclockwise, and then shows it moving correctly clockwise from figure to figure in the series.

For appreciable power and smoothness, the typical Kawai engine uses an extensive number of asymmetrical self-regaugings per axle rotation, for example 36 times on each end, or a total of 72 times for the two ends. The force field of each coil, accompanying its increased magnetostatic scalar potential, is oriented radially inward, so that radial work cannot be done by the coil on the rotor because the rotor does not translate radially. Advantage is taken of the initial clockwise acceleration force initially produced, and self-regauging eliminates the counterclockwise decelerating drag force that would be produced without the regauging.

The major benefits are that (a) multiple self-regaugings occur for a single rotation of the rotor assembly, enabling high weight-to-power ratio, (b) each electromagnet is energized only when positively contributing to the clockwise torque that drives the rotor, and (c) each coil is de-energized to asymmetrically self-regauge the system during those periods when the coil would otherwise create back-drag (counterclockwise torque) if it remained energized.

A conservative cycle is one in which the work done in the back-drag (mmf) region is equal to the work done in the forward boost (forward mmf) region. Self-eliminating the back-drag portion of the cycle is a form of asymmetrical self-regauging, and makes the net field highly nonconservative. The Kawai approach is thus permitted to attain COP>1.0. In his U.S. patent, Kawai quotes performance measurements showing COP = 3.18. As we stated, Hitachi engineers measured two Kawai-modified Hitachi magnetic motors which exhibited COP =1.4 and 1.6 respectively.

Since the output is shaft horsepower, in principle the Kawai engine can be close-looped for self-powering, without consideration of the Dirac sea hole current. With its external close-looping system (not shown), it falls within the second close-looping class, that of an "outrigger" system. Only the unitary system closed-looping class requires consideration of hole currents in the Dirac Sea.

The Kawai process and several other Japanese overunity systems have been blocked from further development and marketing by the Japanese Yakuza. The present author and the Board of Directors of CTEC were physically present and in final negotiations (having reached agreement!) with Kawai to market his self-powering motors worldwide, when control of the Kawai company and process were suddenly seized on the spot in our presence.[195]

6.4.10 Bedini's Skin Depth Magnetic Switching

Bedini found a most interesting variation of flux path switching: he switches the direction of just a little fraction of the ongoing magnetic flux in a material flux path, on the skin of one side of it and with only a slight "skin depth".[196]

Suppose a flux path material is used as the stator pole of a permanent magnet motor. A large permanent magnet is used to magnetize that short length of flux path, as shown one the left in Figure 6-31, which illustrates a concept for a magnetic motor. For convenience, imagine that the small magnet on the right is not yet placed on the material flux path extension. The flux path extension is energized by the south north pole of the large permanent magnet on the left, and used as the stator in a rotary magnetic motor.

---

[195] To anyone who understands Japanese culture, Kawai paid me one of the greatest compliments possible. After his company representative called from Japan, Kawai shipped his *original proof of principle engine* — which first successfully demonstrated his flux-path switching process without being optimized for COP>1.0 — to me, here in the U.S. In the Japanese manner, he had entrusted me with his single most valued possession. From years of aikido experience, I fully understood the meaning of this unprecedented gesture, and expressed my deep appreciation accordingly. From our first meetings with Kawai and his group, I have and always shall have the utmost respect for Teruo Kawai. In my view, it is a sad loss to humankind to have his engine buried and withheld so brutally. Kawai had already achieved a closed-loop self-powering version, and it is this version upon which we reached agreement. In less than 24 hours after that agreement, the Kawai engine and company — and Kawai's own fate — were seized and no longer in Kawai's hands. For those unfamiliar with the deep spread of the Japanese Yakuza, the following are recommended: (a) Brian Bremner, "How the Mob burned the Banks: The Yakuza is at the center of the $350 billion bad-loan scandal," *Business Week,* Jan. 29, 1996, p. 42-43, 46-47; Michael Hirsh and Hideko Takayama, "Big Bang or Bust?" *Newsweek,* Sept. 1, 1997, p. 44-45; Adam Johnston, Yakuza: Past and Present," downloaded from the Internet, Committee for a Safe Society, Organized Crime Page: Japan: History and present status of Japanese organized crime.

[196] Close colleague Ken Moore also independently discovered this effect in his own experiments, but then we found that Bedini had discovered it earlier. That was fine with Ken; he greatly admires Bedini and the marvelous discoveries Bedini has made. I greatly admire both of them, and am happy to call them best friends. After all, **science** has a human side as well as a materials side. Without one's close colleagues and best friends, science alone would be a miserable enterprise!

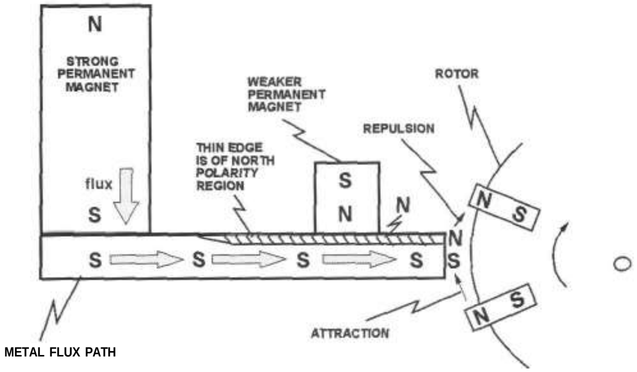

Figure 6-31 Bedlni skin depth magnetic edge switching.

As can be seen, the north pole of the rotor, shown in the figure, is attracted to the resulting south pole of the stator. Power can be taken from the shaft of the rotor while it is being attracted in to the extended north pole of the stator flux extension. As it begins to pass the extension end, obviously if nothing else is done it will develop back mmf "drag" by being attracted back toward the north pole of the flux extension as it passes it.

However, suppose that —just when the rotor south pole is passing dead center of the stator north pole extension, one suddenly pulses a small flat coil placed on the face of the extension, which serves the function of the small permanent magnet shown on the right of the material flux path extension. Further, suppose the magnetic field produced by the coil is oriented as the field of the small permanent magnet, but is much weaker than the magnetic field of the large permanent magnet on the left. A very curious effect then occurs. This energized coil magnetizes the edge of the flux path extension, to a small "skin depth" as shown, with a *north* polarity. The magnetic flux is now bidirectional in that same material flux path extension.

Further, this edge-switching action is performed so that it is this "partial north polarity" of the now-altered extension that faces the rotor north pole as it begins to pass the extension and enters what previously would have been a back mmf section. Instead, now that former back mmf section is a weakened forward mmf region, *further accelerating the rotor.* This action breaks the Lorentz symmetry around the rotation path of the rotor, resulting in a net nonzero (nonconservative) magnetic field. Hence the entire rotor exhibits a net drive force around its loop. The rotor will accelerate until the mechanical work performed upon the shaft, friction, etc. is equal to the net magnetic energy furnished to the rotor.

The beautiful part of this motor is that one does not have to "overcome" by brute force — and then reverse — the north polarity of the entire stator extension. In theory, the overall energy added to the rotor is greater than the energy that must be added for the "edge switching" of the magnetic flux path extension, because the entire back mmf of the motor has been eliminated and replaced by a weaker forward mmf. If the process of furnishing the "skin depth" weaker magnetic north field to the edge of the flux path extension is made sufficiently efficient, this engine also can be made self-powering, at least in theory.

The former "motor" design has been converted to a "motor/generator" design, by the addition of the edge-switching action. The novel factor is the use of bidirectional flows of magnetic flux in a single material flux path.

An additional improvement is to rig a Wiegand sensor and pulse generator or sensors in conjunction with the rotor action, to freely furnish and kill the edge-switching pulse to the face-magnetizing coil on the flux path extension, at the appropriate times. This is an area where even high school students will be able to experiment.

Bedini's patent on the edge switching process and its various embodiments and improvements has been issued {453}.

6.4.11 Bedini's Flux Path Switching with Energy Capture

One additional embodiment of the Bedini flux path switching motor is shown in Figure 6-32. Here the entire flux path is switched, by moving a small coil on the bottom side of the material flux path, as shown in the figure. As Radus showed {454}, it turns out that a much weaker magnetic field in that left bottom position will switch the flux path of a powerful permanent magnet field at the left top of the path, as shown.

Now, however, Bedini adds a transformer coil around the right end of the core (the right end of the extension is still the stator pole in the magnetic motor diagrammatically shown in Figure 6-32). When the flux path is switched back and forth, this transformer coil produces an AC signal output, which can power a load as shown, and can even be used to power the switching power supply. The trick is that it requires less energy to make the vertical flux path preferable to the horizontal flux path, than the energy delivered into the flux path by the permanent magnet. The potential **for** self-powering of a sufficiently efficient magnetic motor is obvious.

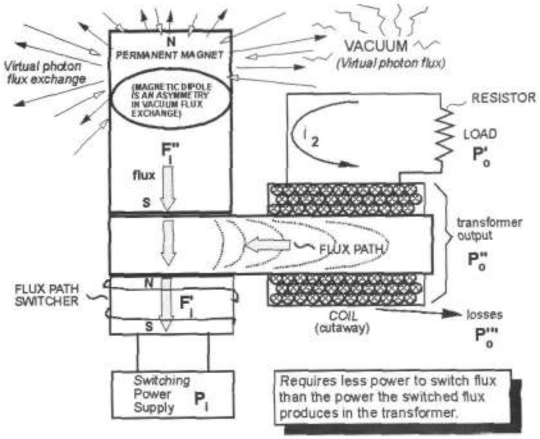
N
PERMANENT MAGNET
Virtual photon flux exchange
(MAGNETIC DIPOLE IS AN ASYMMETRY IN VACUUM FLUX EXCHANGE)
VACUUM
(Virtual photon flux)
RESISTOR
i_2
LOAD
P'_o
F''_i
flux
S
transformer output
P''_o
FLUX PATH
FLUX PATH SWITCHER
N
F'_i
S
COIL (cutaway)
losses
P'''_o
Switching Power Supply
P_i
Requires less power to switch flux than the power the switched flux produces in the transformer.

Figure 6-32 Bedini's total path switching and energy recovery process in a transformer.

So the motor can be powered while simultaneously acting as a partial generator. For efficient switching, this motor/generator can in theory develop COP>1.0 and can even be made close-looped and self-powering. We strongly accent that the primary excess energy input to the system is from the vacuum itself, via the broken symmetry of the large permanent magnet dipole at the left top of the material flux path. As we have repeatedly stated, the only energy problem is how to capture and utilize the EM energy freely received, transduced, and poured out from the vacuum by the source dipole's broken symmetry, without using half of it to destroy the source dipole.

6.4.12 Transformer Secondary Shunted by a Negative Resistor.

See Figure 6-33. This embodiment assumes the availability of a true negative resistor — possibly an adaptation of a point contact transistor or other device. As shown in the figure, a transformer exhibiting COP> 1.0 can be developed by adding a shunting *true* (not differential!) negative resistor across the terminals of the secondary, adjacent to the secondary. In this manner, the Lorentz symmetry of the closed unitary current loop containing the external loads and losses, and normally the secondary of the transformer in series, is broken.

Part of the return spent current is now "pumped" back up from ground side to power side of the transformer secondary without passing through the secondary. The pumping work is providing by dissipation (in the negative resistor) of energy freely obtained from the active vacuum environment. In short, one adds a negentropic process in parallel with the secondary coil, in the secondary external circuit, breaking the otherwise enforced Lorentz symmetrical regauging of the discharge of the collected EM excitation energy in the secondary circuit. This provides the capability for a permissible COP>1.0 transformer.

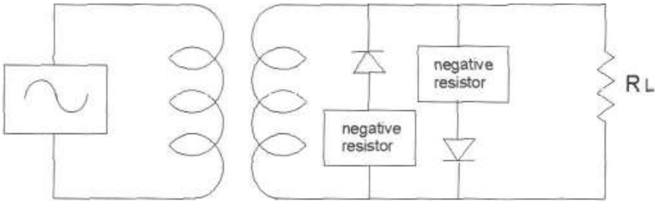

ON EACH HALF CYCLE, ONE OF THE NEGATIVE RESISTORS SHUNTS PART OF THE RETURN CURRENT AROUND THE SECONDARY, FROM LOWSIDE TO HIGH SIDE, THUS DECREASING THE BACK-FIELD COUPLING FROM SECONDARY TO PRIMARY ACROSS THF TRANSFORMER. THIS LESSENS THE ENERGY DISSIPATION REQUIRED IN THE PRIMARY FOR A GIVEN ENERGY FLOW TRANSFER FROM PRIMARY TO SECONDARY

Figure 6-33 Transformer shunted by a negative resistor.

In theory, if the negative resistor can bypass all the current in the secondary, or almost all of it, the transformer can become a COP»1.0 device.

This overunity transformer can use step-up of the voltage from primary to secondary, and exhibits an asymmetry in its forward and backward field coupling between primary and secondary. Thus the dissipation of energy in the primary circuit need not be as great as the dissipation of energy in the secondary circuit. The *energy flows* are of course in perfect conservation, but there is no law of nature requiring the *energy dissipations* to be conservative between the primary and the secondary. It is strictly the backfield coupling from secondary to primary that enforces equal energy dissipation in the primary as in the secondary, in normal transformers. And that backfield coupling's strength is a function of the return current through **the** transformer secondary. By reducing that current, the back coupling to the primary is reduced, which in turn reduces the current draw of the primary from the external power supply.

No laws of physics, electrodynamics, or thermodynamics are violated by the asymmetrically coupled transformer power system with broken symmetry in its secondary to primary circuit couplings. In short, there is less backfield coupling from primary to secondary than there is from secondary back to primary. The additional energy dissipated in the secondary circuit is freely extracted from the vacuum by the negative 4-resistor shunt.

Again, this application assumes the availability of a true negative resistor, to use as a shunt of the secondary.

The beauty of this application is that, once achieved, such a COP>1.0 transformer can easily be close-looped for self-powering by standard "governed" positive feedback. No concern as to hole current effects usually need be accounted for unless the COP becomes very large.

6.4.13 Eddy Current Multiplier.

Figure 6-34 shows an example of an eddy current multiplier, adapted from Burke {455}. This is given as a somewhat unusual system which may exhibit novel effects and possibly can be adapted into a COP>1.0 system (e.g., by using permanent magnets instead of the coils and triggering things into unbalanced condition using a rotary pulse wire generator. We simply list it here as a candidate for the experimenter.

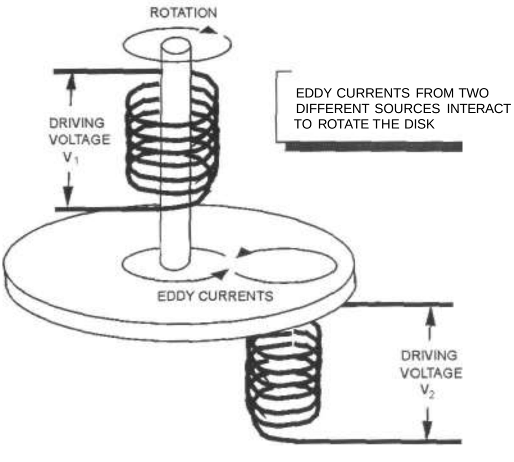
ROTATION
DRIVING VOLTAGE V_1
EDDY CURRENTS FROM TWO DIFFERENT SOURCES INTERACT TO ROTATE THE DISK
EDDY CURRENTS
DRIVING VOLTAGE V_2

Figure 6-34 Eddy current multiplier.

In Figure 6-34, the shaft of the conducting disk is connected to a rotor, which could have arrangements for rotary type of Wiegand pulse generator of Figure 6-29 (with self-initiation and self-recovery) around it (not shown in Figure 6-29). One would start the rotating disk by hand. These Wiegand sensors would then automatically and repeatedly deliver a series of electric pulses to the coils, momentarily generating the two eddy currents shown, so that dynamic force and rotary acceleration is obtained during the endurance of the pulses. In short, one would "electrically kick" the coils so that they continually deliver a propulsive "magnetic kick" to the conducting disk to repeatedly generate eddy currents to rotate it. The rotor is powered as shown by the interactions of the two unbalanced induced eddy currents. Such an arrangement could possibly exhibit COP>1.0.

To experiment with such an apparatus, the major parts that must be developed are (i) the rotor and coils arrangement (the motor), (ii) the integrated rotary Wiegand effect pulse generator with self-recovery, and (iii) the proper timing arrangement. Each should be developed and optimized individually in the researcher's experiments, then combined and optimized together. This is an experimental area that can be cheaply investigated, e.g., by a college student or high school student for a science project. Whether COP>1.0 is achievable or not depends upon the efficiencies that can be obtained with the various processes involved. It is possible, at least in theory, for such an arrangement to be self-rotating once rotary motion is underway.

6.5 CONCLUSION

In this chapter we have advanced many approaches to COP>1.0 systems, that have been tried or considered by various inventors and researchers in the field. Many of these experimenters have attained concrete, replicable results, including results published in the hard literature.

We hope this sampling shows the reader that indeed we can have "self-powering" electrical power systems and COP>1.0 electrical power systems whenever the scientific community will come out of its preoccupation with the present seriously restricted electrodynamics and electrical power engineering practice and fund the necessary research. Simply funding a few graduate students for research and doctoral theses in these areas would be very useful and immensely more productive for electrical energy purposes than all the billions and years spent on chasing hot fusion.

Hopefully the iron dogma of the scientific community against COP>1.0 electrical power systems is slowly changing. If so, and if the community will turn loose those sharp young graduate students and post-doctoral scientists on the problem, very shortly there will no longer be an electrical energy problem on Earth, ever again. We strongly refer the reader to Deffeyes' book {456} to rigorously validate the coming increase of the price of oil, in the middle of the present decade. See particularly the quotation from Deffeyes in our Endnote {342}. With the present extreme volatility of the MidEast and the continuing war on terrorism, the probability is high of eventual severe disruption of oil supplies and the energy infrastructure of the U.S.

An independent review of Deffeyes' book is given by Young {457} and we recommend that the reader digest that commentary also.

If the scientific community will fund the necessary research on permissible COP>1.0 Maxwellian systems, the quick results will solve the energy crisis, clean up the biosphere — including much cleaner air in our cities for the populace to breathe — provide cleaner streams, and result in far less polluted oceans. It will stop the pollution-caused destruction of species, and dramatically reduce the hydrocarbon combustion contributions to global warming. There will no longer be a necessity to consider "burying" in the bottom of the sea the excess CO2 from massive hydrocarbon combustion to provide the much-needed electrical power, because it will no longer be necessary to burn the oil, coal, and gas to get the power. And there will no longer be any necessity for expensive and highly vulnerable nuclear power plants with their resulting nuclear wastes that will remain radioactive for thousands of years.

If cheap, clean EM energy from the vacuum is rapidly developed, we may *finally* get the scientific community's head wrenched back out of the big nuclear power business, as a colossal waste of research manpower and taxpayer dollars — as well as providing ticking time-bombs for the terrorist teams now already in the United States to attack and destroy. It is so easy for a person with appropriate technical knowledge to build a portable electromagnetic pulse (EMP) "shooter" that one shudders at the consequences of in-country terrorist teams that can do so. Try imagining what we would face when the electronic controls of one or more operating nuclear power plants are suddenly disabled and destroyed with the reactor on line and providing power. Chernobyl might look like a mild spring breeze in comparison, should we have one or more actual meltdowns.

It is not a pleasant consequence to contemplate, and that alone may eventually dictate the overwhelming urgency of achieving decentralized, vacuum-energy-powered electrical power systems as soon as humanly possible. The present centralized, cumbersome energy infrastructure is vulnerable beyond all sanity, considering the thousands of terrorists already in country, with weapons of mass destruction as well as more conventional sabotage means. As a single example, the former Soviet Union clandestinely brought in nuclear weapons into the U.S., and

Spetznaz teams to detonate them on order. Such weapons (and teams) are already in large U.S. cities and population centers. The reader is referred to Lunev's book,[197] for details about how the nuclear weapons were introduced.

After all, we do have to live and survive in the real world, and not just in an idealistic scientific environment. We are already at war, with alerts of potential danger to our nuclear power plants already having been given on several occasions. In the present very disorderly and very dangerous world we live in, many of the former options we have had for decades are rapidly ending. That includes the option for the leaders of the scientific community to continue "business as usual", if they wish this nation and their own families to survive. There is an old adage in war: *Whatever your determined enemy can do to hurt you desperately, he will do as his highest priority*. Sadly, in my opinion the leaders of the scientific community — unwittingly — are presently doing almost everything they can do to insure that the energy infrastructure goes, the national economy goes, a nuclear meltdown possibly occurs, and this nation and much of civilization perishes in a new dark ages.

---

[197] Stanislav Lunev (with Ira Winkler), *Through the Eyes of the Enemy,* Regnery, Washington, 1998. Lunev is the highest ranking GRU defector. He is a former Colonel in the GRU, the military counterpart of the KGB. On p. 22-33, he summarizes the Spetznaz capabilities. On p. 22-27, he summarizes Spetznaz use of nuclear weapons already on American soil. On p. 26, he gives some of the ways in which the Soviets easily brought nuclear weapons into the U.S. On p. 30 he confirms **Russian** seismic weapons. Use of EMP weapons is on p. 30-31. Use of very, very low frequency weapons to destroy the human brain, put people into a zombie-like state, and aid in brainwashing is confirmed on p. 31. The present war on terrorism has many more facets than are presented by the conventional news media.

Just as the dinosaurs were confronted with a new and dangerous threat, and failed to adapt, the electrical power infrastructure is now confronted with a potent threat against its very vitals and — so far — is completely failing to adapt. It is going to have to adapt or perish, which presently implies that we all perish along with it. Those who make it their nefarious business to see to it, will see to it.

So let us "see to it" first, scientifically, as rapidly as human research and development can do the job. The very existence of civilization may well depend upon it.

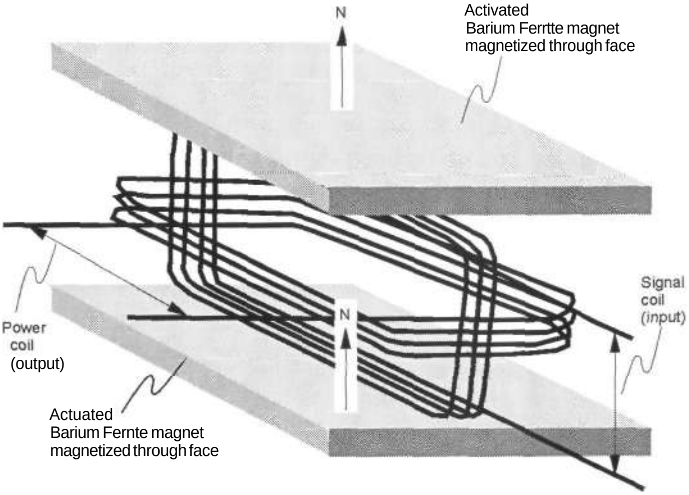
N
Activated
Barium Ferrtte magnet
magnetized through face
N
Signal
coil
(input)
Power
coil
(output)
Actuated
Barium Fernte magnet
magnetized through face

Figure 6-5 Sweet vacuum triode amplifier (VTA) construction.

Chapter 7

Aharonov-Bohm Effect, Geometric Phase, and the Motionless Electromagnetic Generator (MEG)

[Decision that hydrogen fuel is the future]. *"We are at the peak of the oil age but the beginning of the hydrogen age. Anything else is an interim solution. The transition will be very messy, and will take many technological paths ...but the future will be hydrogen fuel."* [Herman Kuipers, Royal Dutch Shell.].

[Vacuum energy is real] *"...the acceptance of a structured vacuum described by an O(3) gauge group leads directly to the existence of novel charges and currents in the vacuum. These are conserved, or Noether, currents and charges and are clearly topological in origin. They spring from the fact that the vacuum is a topological space. Four such entities emerge: [1] A topological vacuum electric charge, also proposed empirically by Lehnert et al. [2] A topological vacuum electric current, also proposed empirically by Lehnert et al. [3] A topological vacuum magnetic charge, proposed also by Barrett and Harmuth. [4] A vacuum topological magnetic current, proposed also by Barrett and Harmuth.*

Each of these four objects can provide energy, which can be loosely termed 'vacuum energy': energy coming from the topology of the vacuum. " [Evans] {458}.

[Aharonov-Bohm effect] *"...the Aharonov-Bohm effect is a local gauge transformation of the true vacuum... [which] produces a vector potential from the true vacuum. [This gauge transformation] produces topological charge..., the electromagnetic field, which carries energy, and the vacuum charge current density first proposed by Lehnert ... and developed by Lehnert and Roy... ".* [Evans and Jeffers] {459}.

[Geometric phase] *"The concept of the geometric phase is closely related to the effect of a state function acquiring a non-zero phase factor after a quantum system undergoes a cyclic evolution. This means, that if a system after some time returns to its original state, the result of such an evolution will be recorded in the phase of the wavefunction. Moreover, this phase factor can be measured by interfering initial and final states. ... Although there are no widely recognized practical applications of the geometric phase, its experimental observations have been reported in many fields of science. The largest group of experiments have been carried out on polarized light ...and polarized neutrons ... The geometric phase has also been observed in magnetic resonance experiments ..., mesoscopic structures ...and molecular systems ...Analogues of GP — the Hannay angles — have been shown to exist in classical mechanical systems ..., the most famous example of which is the Foucault pendulum."* [Aleksiejunas][198]

[Practical use of geometric phase]. *"The motionless electromagnetic generator is the first practical macroscopic use of the geometric phase. "* [T. E. Bearden, 2001, to a correspondent].

[Impact of extracting energy from the vacuum] *"If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete — and so would many of our worries about environmental pollution." "Don't sell your oil shares yet — but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development: 1. It's crazy! 2. It may be possible — so what? 3. I said it was*

---

[198] I am much indebted to Dr. Rimvydas Aleksiejunas for this clear statement of the geometric phase. See his very elegant summary, " Introduction to The Geometric Phase Decomposition," on his Internet website at http://signalogram.free-hosting.lt/Research/gp dec.htm. His own publications in scientific journals and proceedings are also listed on **the** website, and many are available for downloading. His doctoral thesis dealt **with geometric phase** and he is a specialist in that field.

a good idea all along. 4. I thought of it first." [Arthur C. Clarke] {460}.

7.1 Introduction

In this chapter we give a history, some information, and comments about the Aharonov-Bohm effect {473}, which was generalized by the Berry phase {477} for adiabatic conditions, which itself was further generalized by Aharonov and Anandan {480} to remove the adiabatic limitation. The modern all-inclusive term is *geometric phase,* which has been found to be involved in a very wide variety of physics areas. In its embodiments, the motionless electromagnetic generator (MEG) does utilize the Aharonov-Bohm effect (geometric phase) in a unique manner that we explain. We also remark on what we hypothesize may be some related new aspects of the curl-free magnetic vector potential itself.

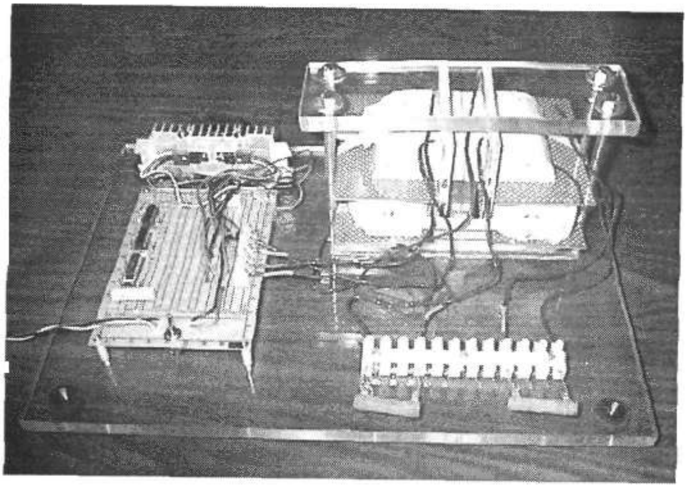

Figure 7-1 Motionless Electronic Generator (MEG)

The motionless electromagnetic generator (MEG) (Figure 7-1) was invented by Patrick, Bearden, Hayes, Moore, and Kenny {461} after years of research and experimentation. It has been technically described elsewhere {462}, and scientific explanations of the fundamental mechanism by which it successfully extracts useful EM energy from the active vacuum have been published in the recognized literature {463, 38a, 464}. Other closely related papers on extracting EM energy from the vacuum have been published or are submitted and in review {465a-465h}.

A version of the MEG has also been independently replicated by Naudin {466}, and in fact by others who did not wish to publish their results. By permission, Naudin's Mark II MEG replica is shown in Figure 7-2. Consequently, the MEG is a validated and replicated COP>1.0 system, as required by proper scientific methodology. Further, sufficient information has been released for any competent laboratory team to replicate a version of the MEG.

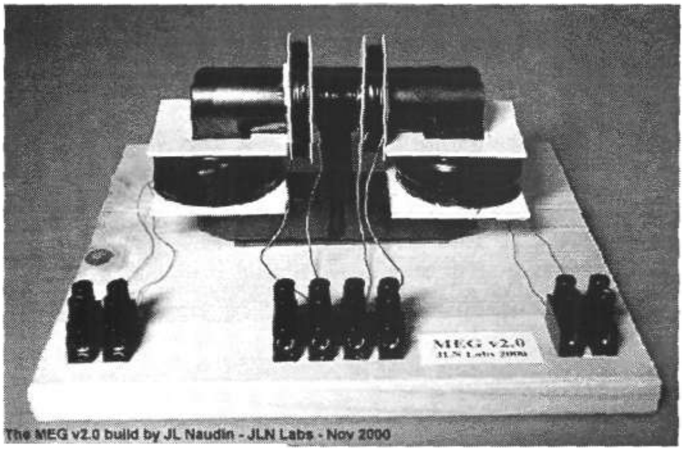

Figure 7 2 Nuadin's Mark II Replication of the MEG.

We give an overview of the MEG, the principles of its operation, the status of the project, and some indication of future plans.

7.1.1 Status

As this is written, the first MEG patent has just issued {37}. A second Provisional Patent Application has been filed, and at this writing is in process of being replaced by a formal regular patent application. Two other patent applications are in preparation. All rights to the MEG are assigned to Magnetic Energy Ltd., Huntsville, Alabama, U.S.A., whose CEO is Dr. James L. Kenny.

In addition, Bedini and the present author have filed a provisional patent application on the process and embodiment for close-looping a COP>1.0 system for self-powering operation, and are in process of replacing it with a regular patent application also. Magnetic Energy Ltd. has permission to utilize the Bedini-Bearden close-looping process. In addition, Magnetic Energy Ltd. has its own proprietary close-looping process, for which a patent application is being prepared. The Company is also involved in serious discussions with several large companies and financial groups for potential financing. When that funding is obtained, we will open a substantial research and development laboratory here in the United States, serving a separate production company to be set up with our financial partners to produce and market the MEG. Meanwhile, laboratory work to complete the research and development of the MEG is also ongoing under a cooperative arrangement with the National Materials Science Laboratory of the National Academy of Science of a friendly foreign nation formerly under the Soviet bloc.

7.1.2 Future Plans

Our plans are straightforward. We will continue to pursue and update our patents and file additional ones as additional discoveries are made. We will continue our negotiations with major financial parties until satisfactory and mutually beneficial capitalization is obtained (we expect that to occur before the end of 2002). At that time we will establish a modern Magnetic Energy Ltd. U.S. laboratory devoted to advanced research and development, as well as a separate production company in partnership with our funding partners. The final research necessary on the MEG (four areas of physics as well as the electrodynamics) will be completed in somewhat less than one year after setup of the lab and lab team, and the production company is expected to have the first engineering development production prototype at the end of that first year. We expect it to be a self-powering unit. We will simply put the first MEG units into production and on the world market immediately thereafter.

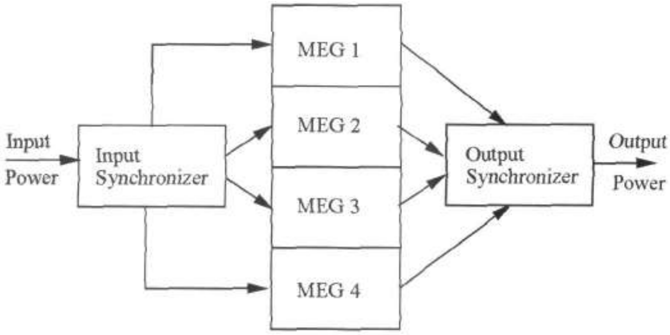

Figure 7-3 Block diagram of envisioned **modular** MEG design.

See Figure 7-3. The first production MEG is envisioned as a modular 2.5 KW unit together with a synchronizer unit simultaneously developed, so that up to six 2.5 KW MEG units can be synchronized in a power system array. This will yield power units from 2.5 to 15 KW. We are struggling to have the closed-loop process completed and ready in that same first year, and if possible the very first MEG units will be close-looped and self-powered. At the end of the second year or before, we plan to have the close-looped basic MEG unit in 10 KW size, again with synchronizer so that up to six MEG units can be synchronized in an array. That will cover the 10 - 60 KW power system size, which is ideal for homes and small businesses. Once the units go on the world market, we believe that the market will be extraordinarily dynamic and productive. We intend to license multiple companies for production and competition for various markets, with no single company being given total monopoly worldwide.

7.2 Aharonov-Bohm (Geometric Phase) Effect

The Aharonov-Bohm (AB) {473} effect, or geometric phase (GP) {477, 480} effect, is now well known {467} in quantum mechanics and quantum electrodynamics, but does not exist in classical electrodynamics. The effect proves the reality of the potentials and their primacy over the fields. Further, it shows that electron effects can be generated by the magnetic vector potential A even in the absence of the magnetic field B.

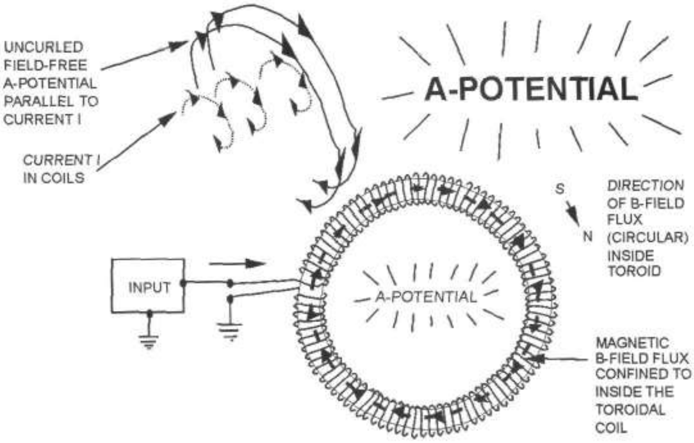
UNCURLED FIELD-FREE A-POTENTIAL PARALLEL TO CURRENT I
CURRENT I IN COILS
A-POTENTIAL
S
N
DIRECTION OF B-FIELD FLUX (CIRCULAR) INSIDE TOROID
INPUT
A-POTENTIAL
MAGNETIC B-FIELD FLUX CONFINED TO INSIDE THE TOROIDAL COIL

Figure 7-4 Aharonov Bohm effect for a tightly wound toroidal coil.

According to Feynman {468}, it required 25 years for quantum physicists to clearly face the issue eventually pointed out in 1959 by Aharonov and Bohm of the primacy and separate action of the force-field-free potential, in regions where no force fields were present (Figure 7-4) {469}. After the effect was experimentally demonstrated by Chambers {474} in 1960, it then required another 25 years before physicists would accept it. Even today, another 16 years after full acceptance in 1986, to our knowledge the Aharonov-Bohm effect has not previously been utilized in any practical electrical power system {470}, much less one that has a COP>1.0. As Aleksiejunas[198] states (see his quotation at the beginning of this chapter), there have not previously been any widely recognized practical applications of the geometric phase. Now there is a practical application of major worldwide importance, since the operation of the MEG depends upon the geometric phase (Aharonov-Bohm) effect to produce real electrical power from the active vacuum itself.

We also believe that the curl-free potential may not have been adequately investigated, particularly along the lines of its utilization in the MEG.

7.3 Abbreviated History of the Geometric Phase

The earliest work with a state function acquiring a non-zero phase factor after a quantum system undergoes a cyclic evolution was published in 1956 by Indian scientist Pancharatnam {471}. Pancharatnam noticed phase shifts of cyclic evolutions in optical phenomena. This means that, if the system leaves its original state and after some time returns to it, the result of the system's evolution will be recorded in the phase of its wavefunction. In addition, this phase factor can be measured by interfering the system's initial and final states.[198] Sadly, Pancharatnam's anticipation of the geometric phase factor did not receive the attention it so justly warranted, and his paper languished almost unnoticed for nearly four decades, when it was finally rediscovered. This was an excruciating scientific tragedy to Pancharatnam, comparable to that which befell Waterston {472}.

The effect of the field-free potential and the geometric phase issue were pointed out strongly by Aharonov and Bohm {473} in 1959 in a leading U.S. physics journal, causing much consternation in physics. Before that paper, most physicists still believed that all electromagnetic forces and effects were induced only by the force fields,[199] and not by the potentials.

[199] Indeed, the force fields do not even exist in mass-free space, but only in matter. Force is defined as $\mathrm{F} \equiv \partial/\partial t\,(\mathrm{mv})$ so that mass is a *component* of force. Feynman states: *"...in dealing with force the tacit assumption is always made that the force is equal to zero unless some physical body is present... One of the most important characteristics of force is that it has a material origin... " "...the existence of the positive charge, in some sense, distorts, or creates a "condition" in space, so that when we put the negative charge in, it feels a force. This potentiality for producing a*

In the prevailing belief, when the fields were zero there were no EM effects. Indeed, the potentials were largely regarded as mathematical conveniences.

Another point of consternation was that the A-potential is not uniquely determined, but is determined only up to a particular gauge transformation. So there can be two (or even more) magnetic vector potentials A and A' corresponding to a specific magnetic field. Aharonov and Bohm raised a disturbing question: How can two different "causal" A potentials produce the same effect? We will rigorously answer this question shortly.

While the experiments suggested were difficult, they were performed shortly after the publication of the Aharonov-Bohm paper. The very next year, in 1960, Chambers {474} experimentally demonstrated the Aharonov-Bohm effect, though not with complete decisiveness. A 1983 paper by Tonomura *et al.* {475} was significant, and a 1986 publication of further work of Tonomura *et al.* was decisive {476}.

In 1984, Berry generalized the Aharonov-Bohm effect into the geometric phase {477}. In 1985 the Tonomura effort plus experimental work by Webb *et al.* {478} settled the matter, as strongly stated by Schwarzschild the next year {479}. In 1987, Aharonov and Anandan {480} further generalized Berry's geometrical phase, freeing it from its adiabatic

---

force is called an electric field. " So Feynman was fully aware that not the force field but only the potential (i.e., the potentiality) for producing the force field exists in mass-free space. The electrodynamicists continue to be confused on that point. E.g., on p. 28 of his *Classical Electrodynamics,* 2nd edition, Wylie, 1975 Jackson states: *"...the thing that eventually gets measured is a force... ",* confirming that force is an observable and thus a *d/dt* operation imposed upon spacetime LLLT to give a purely spatial entity LLL. On p. 249, Jackson also states that *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."* The two parts of the conventional assumption, as stated by Jackson, contradict each other. We point out that Jackson's book has been a standard of excellence for years, and no EM library is complete without it. It merely shows the extraordinary difficulty of rooting out the exasperating and mind-numbing confusion in physics between the observed effect and the unobserved cause. Force is an effect, after interaction (observation or measurement), hence cannot be the cause (which exists prior to interaction or measurement). Else, there is no difference between cause and effect, which destroys causality itself. This confusion is still there in even the hoary old classical mechanics, now some centuries old. By that standard, electrodynamicists have a long time remaining before they will have propagated this non sequitur" as long as the mechanical physicists have.

restriction and generalizing the geometry. Much discussion of the AB effect and Berry's geometric phase occurred {481a-481c}, but the scientific consensus was in: *The Aharonov-Bohm effect is real and the geometric phase is a widespread fact of nature.*

In 1985, Olariu and Popescu {482} published a very good overview and discussion of the AB effect up to that time, with hundreds of references.

Thousands of papers on the AB effect, the Berry phase, and the geometric phase have since appeared in the literature, and continue to appear at a rather steady and prolific rate. A few are listed for a sampling {483a-483k, 484a-484i}. The original Aharonov-Bohm effect has now expanded and spread widely through physics and into a myriad other effects and areas, such as the scalar AB effect, the Aharonov-Casher effect, magnetic vector potential vortices, quantum tunneling, chemical reactions, chaotic oscillation, and others. These related extensions are far beyond our interest here. We are, however, interested in what we hypothesize may be a new aspect of the perturbed curl-free magnetic vector potential A. We shall address that aspect briefly.

7.4 A Possible New Aspect of the Curl-Free Magnetic Vector Potential

We propose what may possibly be a unique new aspect of the curl-free A-potential — and thus of the Aharonov-Bohm effect — that has been uncovered in our MEG experiments. We caution, however, that this is a tentative interpretation and remains to be further validated.

First we state that the equation, $\mathbf{B} = \nabla\times\mathbf{A}$, usually accepted as the "definition" of the magnetic vector potential A, is not even a definition, because no equation is a definition.[200] If we rewrite the equation as the identity $\mathbf{B} \equiv \nabla\times\mathbf{A}$, then it is a proper definition — though not of the A-potential, but of the B-field itself, *given* that the A-potential exists and is primary. It also states that the B-field is simply a magnetic vector potential having swirl; i.e., the notion of the B-field captures that swirl aspect of a swirling A-potential. Hence the statement that B = zero implies only that **the** A-potential remaining is swirl-free and hence a linear EM energy current flow. Note that the B = 0 condition does not specify the magnitude of the remaining A-potential, but only specifies that it has no curl. This is a rigorous answer to the original question posed by Aharonov and Bohm of how two different A-potentials can have the same magnetic field B (the same swirl). It is no more mysterious than two different rivers having a whirlpool of exactly the same swirl magnitude.

---

[200] No equation is a definition, *a priori.* It merely states the relationship of the total magnitude of the entities on the left to the total magnitude of the entities on the right. It does not define any entity on either the left or the right. One of the gross non sqeuiturs in physics texts is to continue to present equations as "definitions".

At least to first order, it appears we have arrived at a "semi-definition" of the A-potential along these lines:

> *The magnetic vector potential A is an EM energy current in 4-space, which can be swirling or linear, or can have two components where one is swirling and the other is linear. The swirling component of the A-potential is also known as the B-field, and the linear component is known simply as the "curl-free A-potential". The A-potential exhibits a remarkable property: If its swirl component is localized and confined, the remaining space where the swirling A without localization would have been is filled by a remaining unswirling linear A-potential energy current. In short, by swirl-localization the swirling A-potential can be freely 'regauged' into two separated entities, with the sum of their energy being greater than the apparent energy of the previous swirling A-potential. Hence swirl-localization is a method of asymmetrical self-regauging of the A-potential itself with the excess energy coming directly from the active vacuum.*

If we take the view that any curled A is identically a B-field, then the "natural" A is not curled, because it requires a curl operation applied to it to produce a "curled A" or B-field. Therefore the "natural" magnetic vector potential A is simply a magnetic energy current in space, since it is a vector and has direction {458}. That is a *longitudinal EM energy flow process?*[201] We shall be very interested in the uncurled A-potential; i.e., in that linear energy current in space that is uncovered and freely produced when the B-field is localized. Indeed, we shall perturb it to make a large E-field, which we intercept and use to furnish emf to the secondary circuit and power loads.

---

[201] Interestingly, around a toroidal localization zone, the uncurled A-potential does curve, but is only linearly following the curving current and the moving E-field from the moving charges in the current. In the MEG, the uncurled A-potential is even more complex, since we use each half of the toroid-like localization zone differently, with the B-flux going one direction (say, counterclockwise) in the left branch and the B-flux going in the other direction (say, clockwise) in the right branch, and with a crossover point in the middle.

Given a localized B-field so that we have an uncurled A-potential outside the localization region: If we now increase the magnitude of the A, that produces an increased linear energy flow rate outside the localization zone, in addition to increasing the B-field in the localization zone. If we decrease the magnitude of the A, that produces a decreased nonlocalized energy flow rate and a decreased localized B-field. In our view it follows that, when we oscillate the magnitude of the A-potential by any means whatsoever, we oscillate the nonlocalized linear EM energy current flow rate as well as the localized B-field. In fact, that creates a normal oscillation of the localized B-field, and also creates a nonlocalized *longitudinal wave* of electric field E by the equation $\mathbf{E} = -\partial\mathbf{A}/\partial t$, if the conditions localizing the curl operator — and hence localizing and holding separate the concomitant B-field produced by a changing E-field — continue in effect.

We have produced a mechanism for generating a nonlocalized longitudinal E-field wave without a nonlocalized B-field component being allowed to act outside a confined locality. We stress that a B-field component is produced, but remains localized in the special core material (that is one of the conditions required in the MEG).

We have also produced a mechanism for generating a nonlocalized longitudinal E-field wave with another very novel feature: the magnitude of the E-field and thus the magnitude of its energy density (its intensity) does not depend on the *magnitude of the input energy* used to perturb the A-potential, but only depends on how sharply we perturb it. Thus we have produced a mechanism for energy intensity amplification and magnification, freely. This is possible because the system is an open system far from equilibrium with the active vacuum, and specifically with the curl-free A-potential as an energy current at large in the local vacuum where it can be intercepted, collected, and used to power loads. The energy intercepted from the uncurled A-potential by the output section(s), is *additional* energy in excess of what is intercepted from the E-fields generated in core material B-field magnetic flux changes.

However, see our discussions of the giant negentropy process, earlier in this book. Any longitudinal EM "wave in 3-space" is accompanied by a longitudinal EM wave on the fourth Minkowski axis, in the time domain. EM energy does not observably propagate continuously in 3-space, since observed propagation is a series of frozen iterative observation processes a priori). The "energy flow" actually consists of a propagation in 4-space of a *continual local circulation* of EM energy from the time domain (time-polarized EM energy) to each interacting point dipole in 3-space via the negative charge of the point dipole, thence to the positive charge of that point dipole, thence back to the time domain. In a time-reversed zone, the circulation is in the opposite manner. So what spreads as the "longitudinal EM energy current in 3-space" is actually the spreading point polarization of the vacuum and this giant negentropy 4-circulation involved in every point dipole of that polarization.

We believe this previously unsuspected process is involved in the Aharonov-Bohm effect (which is a sort of polarization in the time domain), other aspects of the geometric phase, fundamental polarization of the vacuum itself, and the very notion of "propagation of EM energy through space". At any rate, the hypothesis does fit the phenomena observed to happen in the MEG, so it warrants additional careful experimental investigation as well as attention from leading theorists.

Further, the MEG is practical and produces practical electrical power, but with a *purely electrical reaction* in the output secondary coils of the transformer section.[202] Hence there is finally a very practical use for the geometric phase: It extracts additional *nonlocalized* EM energy from the vacuum (and local curved spacetime) in addition to extracting and retaining *localized* energy. Both the extra "free" nonlocalized EM energy and the localized EM energy from the vacuum can be intercepted, collected, and utilized to power practical fuel-free MEG-type power systems.

We point out that the rate of changes of all E-fields produced in the secondary and primary coils also produce B-fields proportional to the time rate of change of those E-fields in each case. These secondary B-fields are retained in the core rather than reaching out into surrounding space. Their rate of change then further produces E-fields, which pass out of the core and interact with the coils. Hence there are very novel "multiple feedforward and feedback" regenerative loops in the MEG's highly nonlinear operation, as well as chaotic oscillations.

---

[202] This follows from the localization of all B-fields in the core material — including the B-field of the permanent magnet, those B-fields produced by dE/dt operations, and those B-fields produced inside the secondary coils, etc. All E-fields produced, however, including in the changing flux of the core and in the $d\mathbf{A}/dt$ operations of the external uncurled A-potential — are free from that localization restriction and can pass out of the core and interact with the output coils of the transformer section. In short, we have produced the world's first purely electric "induction-free" magnetic transformer device. The voltage and current, measured directly across the secondary coil's ends, is in phase within 2 degrees and even that 2 degrees can be eliminated. We expect the theorists to have a field day in modeling these new phenomena.

7.5 Use of the Aharonov-Bohm Effect in the MEG

From the standpoint of electrical power systems, our own interest in the AB effect focused upon one important characteristic: the clear separation (and localization) of the B-field (the swirl component of the A-potential), leaving the remaining "natural" uncurled field-free A-potential as a free linear flow of energy current in the vacuum. This operation is iterative, and is produced by the core for every B-field produced in the transformer section. So a localized dynamic B-field makes nonlocalized dynamic E-fields, which themselves make further localized dynamic B-fields, which then further make nonlocalized dynamic E-fields, and so on. The result is a very extensive chain of feedforward and feedback field energy loops.

7.5.1 Separation of the Swirled and Linear A-Potentials

The AB effect has long been known in toroidal coils and in long solenoids. In a theoretically perfect toroid, for example, the AB effect separates and localizes the entire B-field inside the toroid, while the *curl-free* magnetic vector potential A appears in surrounding space outside the toroid. By our new interpretation and hypothesis, based on AIAS theoretical work published in the literature, this curl-free A-potential represents a nonlocalized longitudinal EM energy current in space.[203,204] Further, the curl-free A-potential is an extra flow of EM energy that can be intercepted, collected, and used while also still using the entire "normal" B-field energy localized in the core. The neat thing is that the E-field made from the

---

[203] This is addressed in M. W. Evans, T. E. Bearden, and A. Labounsky, "The Most General Form of the Vector Potential in Electrodynamics," *Found. Phys. Lett.,* 15(3), June 2002, p. 245-261. See also M. W. Evans et al., "The Aharonov-Bohm Effect as the Basis of Electromagnetic Energy Inherent in the Vacuum," *Found. Phys. Lett,* (in review).

[204] See also B. Lehnert and J. Scheffel, "On the Minimum Elementary Charge of an Extended Electromagnetic Theory," *Physica Scripta,* vol. 65, 2002, p. 200. In extended EM theory, the elementary charge may no longer be considered as a fundamental constant of nature, but is expressed in terms of the Planck constant and the velocity of light.

A-potential, and thus providing the actual intercepted and collected EM energy, can be made as large as desired, merely by manipulating the rise time and decay time of the pulses used to perturb the A-potential external to the core. Further, the same perturbation also perturbs the B-field flux in the core, giving rise to nonlocalized E-fields whose magnitude again depends upon the rise time and decay time of the perturbing pulses. Hence the dA/dt and the dB/dt become energy amplification causes, where all the excess F-field energy is furnished freely from the vacuum via a special kind of asymmetrical regauging and gauge freedom. In that sense, the MEG produces and applies a hitherto unrecognized form of direct field energy amplification by asymmetrical regauging.

We noted that this mechanism represents the formation of a locally curved spacetime[205] and engineering of the magnitude of the curvature itself. By freely increasing the local spacetime curvature, the appearance of additional EM energy density in the region follows *a priori*. That is, one still has all the B-field energy produced by the permanent magnet, but it is just confined to the space inside the core. All the energy one "pays for" or inputs with current and voltage to the input coil, in the normal sense, is the energy to *perturb* the localized B-field and the curl-free A-potential. However, even before perturbation to evoke energy amplification, there is already additional energy current that has appeared outside the core, where the magnetic field B would have normally extended. All the A-potential energy[206] is "extra" energy in addition to the normal B-field energy that would otherwise have appeared out there but is now confined inside the core.

---

[205] More accurately there is one specific curvature of spacetime in the core of the MEG, representing the confined magnetic field B (the localized swirling A-potential), and there is another different curvature of spacetime outside the localization zone, filled with the nonlocalized curl-free (non-swirling) A-potential as a nonlocalized linear energy flow in space. Further, in the MEG we freely change and increase the magnitude of those spacetime curvatures, by the *rate of change* of the input energy rather than by its *magnitude*, and hence we easily increase the magnitude of the energy density of local space both inside and outside the core. That is the fundamental principle of the MEG"s energy amplification (free regauging).

[206] Some scientists even deny that the curl-free magnetic vector potential has any energy! We answer that it produces energetic effects upon charged particles, hence must either contain energy itself or direct and gate the interaction of some local vacuum energy with those charges. In the MEG, the uncurled A-potential is oscillated, which does not appear to yield a *transverse* EM E-field wave at all. Instead, it appears to yield a *longitudinal* E-field wave with an associated scalar KM wave in the time domain. We also explained what appears to happen to the concomitant B-field wave produced by any changing E-field in the secondary

Further, the uncurled A-potential outside the core reduces its magnitude only inversely as the radial distance, whereas the normal B-field that would otherwise be there in the absence of localization would have reduced its magnitude inversely as the square of the radial distance. This far less rapid fall-off of the magnitude of the nonlocalized curl-free A-potential than the former nonlocalized magnetic field B can result in a remarkable increase in communications range — when a normal EM field antenna with oscillating EM field transmissions is compared to a curl-free A-potential antenna radiating the same power but as curl-free A-potential oscillating transmissions. Golden[207] has applied that effect for long-range communications.

With the application *of the free* Aharonov-Bohm effect produced by the core material and permanent magnet combination, rigorously there has appeared much more available EM energy from the local vacuum than the energy one inputs to the device as one's input energy "payment". Again, this is because the system is an open system far from equilibrium with the active vacuum processes such as the appearance of the uncurled A-potential and the iterative appearance of powerful extra B-fields and E-fields. The rise time and decay time of the input perturbation pulses produces an energy amplification effect both in the core and in space outside the core. When one utilizes this extra uncurled A-potential energy, as well as the now-confined B-field energy, as well as the extra B-fields and E-fields appearing from the iterative processes, then one can extract and use more energy than one himself "paid for" or input to perform the perturbation. Indeed, one can easily extract and use more energy from the curl-free A-potential alone, than one has input. By simply controlling the rise time and decay time of the input pulses, the magnitudes of the resulting E-fields formed in the uncurled A-potential (and in the perturbed trapped B-field perturbations inside the core) can be controlled at will.

(output) coils of the transformer section, etc. The core material withdraws and holds localized, all such magnetic fields formed in the internal MEG core processes.

[207] In the late 1970s Frank Golden demonstrated such an uncurled A-potential a**ntenna** which attained a range of 20 miles for a tiny CB radio, compared to a range of only 1 mile for the same CB radio with a normal EM field antenna. The energy density of the potential drops off inversely as the distance, while the energy density of the field drops off inversely as the *square* of the distance.

We also point out that all these amplified E-fields produced by multiple actions have both a conventional Poynting energy flow component and an unaccounted, much larger Heaviside energy flow component. Thus the total energy amplification factor can be very large (several orders of magnitude). With efficient interception and collection, in the future we expect to be able to produce at least 100 kilowatts of power out of the same size core and buildup now producing 80 watts or less in a highly nonoptimized set of laboratory experiment MEGs.

We again call the reader's attention to our previous discussions in this book, showing that all EM energy in EM circuits comes from the vacuum anyway (particle physics view) or from the time-axis (relativistic view). This has previously been hidden by the failure of classical electrodynamics to solve the source charge problem, which we also solved in this book after our paper {12} in 2000. Circuit theorists avoided it by indirectly assuming the source charges freely created from nothing all the EM energy in their fields and potentials. The use of the assumed fields and potentials and their Poynting energy was calculated, but never was the source of those fields, potentials, and their energy presented or accounted.

7.5.2 Using the Uncurled A-potential as a Linear Energy Current

A very simple equation allows utilizing the extra nonlocalized A-potential energy, even though it has no curl and therefore no magnetic B-field. That is $\partial \mathbf{A}/\partial t = -\mathbf{E}$. This equation means that, if one perturbs that large field-free A-potential outside the toroid, it produces an E-field, which we have argued must now be a longitudinal EM wave due to the perturbations *ifthe B-confining operation is still imposed.* By oscillating the A-potential perturbations, one produces an oscillating E-field that is also an oscillating longitudinal electric E-field wave without an accompanying B-field wave.

The oscillating B-fields in the core also produce E-fields proportional to $\partial \mathbf{B}/\partial t = -\mathbf{E}$. These extra E-fields are in phase with the E-fields produced by the perturbed external A-potentials, considering the wavelength of the frequencies at which the MEG is operating (30 to 80 kHz nominally).

Interestingly, the time-oscillating rate of change of the E-field in this longitudinal E-field wave does produce a magnetic B-field, but it is confined to only the core material and not in the nonlocalized external spatial region *because ofthe localization capability ofthe core material.* The core simply changes its permeability and holds the additional B-fields and their flux in there, thus forming additional uncurled A-potential outside the core. The E-fields produced by variations of the core B-field flux is not confined, but freely passes out of the core material

If that B-field (curled A) is sharply localized by an ongoing localization capability, the longitudinal E-wave has its produced magnetic B-wave component stripped from it and localized. In that case, the process produces a purely longitudinal E-field wave outside the localization region — with the possible exception that a longitudinal B(3) field wave component pioneered by Evans {485} *may be produced as well. That is left as an open question presently unresolved in our MEG experiments. We tentatively believe that the B(3) field wave component is in fact produced, but also forcibly localized in the core along with the magnetic field B. If so, its perturbations will also contribute to producing extra E-fields.*

See Figure 7-5. In the MEG, the localization is accomplished as a unique function performed freely by the special nanocrystalline layered core material. So contrary to the case of the solenoid or toroid, one does not to have to "pay" any EM energy to the localizing component to obtain the localization function itself.

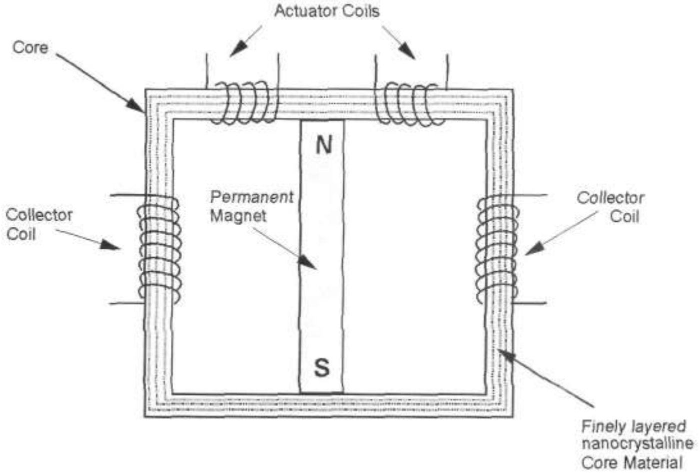

Figure 7-5 Diagrammatic drawing of the MEG.

Further, if one deliberately uses a nearly rectangular perturbing pulse that has very sharp rise time and very sharp decay time, the nonlocalized E-fields, resulting localized secondary B-fields, and resulting nonlocalized secondary E-fields produced can be made very large, with rapid rates of change also. In other words, the magnitudes of the nonlocalized E-fields and localized B-fields produced can be adjusted at will, including in their higher orders. The magnitudes of these evoked fields do not depend on the *magnitude* of the input perturbation energy, but on its *rate of change.* This is very important, because it means the MEG permissibly achieves energy magnitude increase in its nonlocalized output E-field waves by simple instantaneous power deliberately prepared in its low-energy input perturbation waves.[208] In short, the motionless electromagnetic generator uses perturbation of the Aharonov Bohm effect as an energy amplifying mechanism.

This also affects all the iterative processes, hence all the other E-field waves produced outside the core material.

With the localization process remaining and ongoing (simply because the core material remains), the B-field continues to be stripped away, leaving the perturbed A-potential to produce a longitudinal E-field wave (ignoring the possible presence of a B(3) field also). In the MEG transformer section, the output (secondary) coils are wound around the nanocrystalline core that performs a B-field and B-flux localization within the core itself. Hence a purely net E-field interaction occurs with the electrons in the secondary coils, and the output of these coils is an output having the output voltage in phase with the output current.[209] The AC current in the secondary coils does in fact still produce magnetic field B-flux and B-waves in them, but the core material draws the extra B-flux and B-waves inside and localizes them within the core. The fact that localization is *not quite* 100% at the imperfect junctions of the ends of the permanent magnet and the core material, gives a tiny bit of leakage of

---

[208] It also means that the MEG operation is highly nonlinear, with very extensive feedforward and feedback field transduction loops. For the MEG, its scale-up, mathematical modeling, and engineering simulation are not simple things at all, but highly complex. In addition, control and stability are more complex, since one must utilize nonlinear oscillation theory (including chaotic oscillation theory) as well as nonlinear oscillation control theory. In addition, a higher group symmetry electrodynamics is required for the modeling and simulation. Consequently, because of these considerations and the geometric phase aspects, as well as the additional close-looping considerations for self-powering operation, an expensive physics lab must be established as well as the usual electrodynamics laboratory aspects. The result is an expensive final research and development program required to complete the MEG to final production prototype units, ready to start rolling off the assembly lines. A minimum capitalization of $20 million is required.

[209] In the actual lab experiment MEG, there is a slightly imperfect junction between the ends of the permanent magnet and the core material. This results in a very tiny bit of leakage of nonlocal B-field —just enough to produce a 2° phase shift between the voltage and current in the secondary coil's output. With a better junction, this phase shift can be further reduced or eliminated.

B-field, so a tiny B-field component does escape, interact, and generate the *2%* or so variation from the complete in-phase condition of the secondary output voltage and current. Again, that can be further reduced or eliminated by improving the junction.

So essentially there is a purely electric field interaction with the electrons in the output coil current of the MEG, hence the result that the output current is in phase with the output voltage — something previously unheard of in a simple secondary coil of a transformer. This is a major practical macroscopic application of the Aharonov-Bohm effect and therefore of the geometric phase.

7.6 Collection of Field Energy in the MEG

To collect and use EM energy in a power generating system, first the EM energy must be evoked from local spacetime (the local vacuum) by the broken symmetry of the source charges and dipoles in the circuit {12}. In the open system far from equilibrium in the resulting EM energy flow, the operator must usually pay for (i) the energy required to *initially* produce the requisite source dipolarity and (ii) the energy required to perform necessary switching and control functions, unless close-looping is incorporated. Using the permanent magnet as a source dipole in the MEG, and using the special nanocrystalline core material to generate the Aharonov-Bohm effect to furnish excess energy from the vacuum, the MEG operator does not have to *furnish* the evoked and available field energy at all. So the first "payment" is totally avoided. Further, the magnetic dipole is not destroyed by passing magnetic flux back through its back mmf. No additional input energy is required to continually restore the source dipole, since it is never destroyed.

In an open-loop MEG system, the operator only has to pay for the second part: the switching and control function energy. In later closed-loop MEG systems, a method of controlled and positive feedback has been developed so that, once the system is up and operating and powering its load, the necessary switching and control function energy can be siphoned off the output and fed into the input in governed fashion, first being transduced into positive energy input form. Once the process is properly synchronized and governing properly, the operator can throw a switch for feedback and transduction, and the unit will be "self-powering", or "fueled entirely by **the** local vacuum energy", so to speak.

Let us consider the open-loop MEG system we have presently attained in a series of embodiments.

The total emf force F collected from a given nonzero electric field E, e.g., depends only upon the amount of available intercepting charge, by the simple equation $F = Eq$. Similarly, from any potential ϕ the amount of collected EM potential energy W depends on the amount of intercepting charge q, by $W = \phi q$. Hence in the MEG a legitimate COP>1.0 system is possible if (i) more energy is collected and used from the evoked EM energy flow than is input to gate that energy flow's appearance from the active environment, and (ii) none — or appreciably less than half— of the energy collected in the circuit is used to destroy the source dipolarity.

By invoking the AB effect in a local region, the vacuum contributes additional energy density in the space outside that region, and that energy density can be sharply perturbed so that substantial field energy amplification appears by asymmetrical regauging. The excess energy to provide COP>1.0 performance is freely furnished from the active vacuum (and from the local curved spacetime) {38b, 38c}. The excess energy can be intercepted in the perturbed curl-free A-potential region outside the transformer section, collected, and used to power additional loads, in addition to being intercepted in the transformer secondary region to power the secondary circuit and main load.

Similar contributions of extra energy and extra energy collection outside the core localization region occur for the iterative processes also.

At the same time, the B-field inside the core material is perturbed, but the core material self-varies its permeability to adjust to the additional magnitude of the perturbed B-field. The production of any B-field oscillation inside the core does produce an electric field E, which is not localized in the core. Consequently, energy from the perturbed core-localized B-fields is transduced into nonlocal E-field energy, passing out of the core and interacting with the output secondary coils of the transformer section in purely electric field fashion.

Because of the purely electrical interaction, the voltage and current output from the secondary coils of the transformer section are in phase, and not 90° out of phase as in a transformer. The actual 2° experimental phase shift of current with respect to voltage is a measure of the slight inefficiency of the junctions of the permanent magnet ends with the B-localizing core material.

In the MEG, we are using a novel new kind of component: an *external-induction-free cored inductor*, which might alternately be called *a purely electrical inductor — at least externally.* In short, the inductive action remains localized in the core, while the electrical action escapes the core and interacts with the coil to drive the current electrically. So here is another excellent, novel, and practical use of the Aharonov-Bohm effect and the geometric phase: Eliminate the inductive reaction of a coil's output, and transduce it into a purely electrical action. At this writing, we are preparing to file a separate patent application upon this latest application and process, and it will already be filed by the time this book is published.

We accent that any field or potential, once made, already involves an ongoing giant negentropy and overunity process {486}, as we previously discussed in this book. Simply calculate how much energy it requires to separate and form the source charge — which can be as simple as charging a tiny little sphere — and recognize that the field formed by that charge is an altered energy density of surrounding space and a polarization of the vacuum. Once the charge is formed, the corresponding alteration and sustaining of spatial energy density change and vacuum polarization freely moves radially outward at the speed of light from the charge, from the very moment it is produced. So a steady flow of observable EM energy is speeding out in all directions from the source charge. An enormous change in the energy density over that external spatial region surrounding the source charge occurs and spreads steadily, moving radially outward in all directions at speed c.

Hence a magnificent and rapidly extending curvature of spacetime — which energetically acts upon any available charged matter therein, by conventional general relativity — is formed and spreads outward in all directions from any source charge, once formed. How much effect on charged matter is exhibited by this freely created and spreading spatial EM energy density change, depends only on the amount of interacting/receiving charged matter in that surrounding space that is available to interact with the energy of the new fields and new potentials.

If **the** source charge or source dipolarity (in this case, primarily the magnetic dipole of the permanent magnet) is left alone and not destroyed, an unlimited amount of energy can be intercepted and discharged from a local load, again and again, from the charge's surrounding field or potential because the energy in the field and potential is steadily replenished from the active vacuum. The only problem is discharge the collected energy in the load(s) so as to leave the original source charges (the magnetic poles) inviolate, and not to destroy them. The MEG accomplishes this function.

This also rigorously follows from the gauge freedom axiom of quantum field theory. Consider this: the ability to freely change the potential — assumed by all electrodynamicists — can be applied directly to the load to potentialized it without any current moving, if there exists a somewhat lengthy electron gas relaxation time in that part of the load that is potentialized. After potentialization, one can switch the source of free potential away, and complete the load circuit separately, so that the excess potential energy is dissipated from the load, powering it. In that case there is no dissipation of the primary source dipole. *Gauge freedom already assumes one can produce EM energy from the local vacuum freely, anywhere and at any time.* In real life, one has to pay a little for the switching, but not for the energy input itself.

Why the scientific community for decades has fiercely opposed the funded experimental investigation of the implications of gauge freedom is quite inexplicable.

As can be seen from the conservation of energy law alone, in theory energy can be continually intercepted and collected from these free fields and potentials, once the source dipolarity (or source charge) creation has been paid for. Further, by applying gauge freedom to potentialized the load without energy dissipation, with a little adroit switching the collected energy can then be dissipated entirely in a separated external load circuit, so that the free potential energy completely exits that circuit without re-interacting with the source dipole and its back emf or mmf.

We predict that the above application of gauge freedom will eventually power most of the world's electrical systems, when the scientific community gives up its present "ostrich position". Once the engineers can build it, the scientists do not have to agree.

That such is not done in our conventional circuits is purely due to the limited vision of the scientific community and its refusal to fund appropriate research. To date, it has not even recognized what actually powers every electrical circuit, or that a common practice (the closed current loop circuit) has eliminated COP>1.0 and self-powering Maxwellian power systems for more than a century. The entire electrical energy crisis hurtling at the nations of the world is due entirely to the shortsightedness of the scientific community, particularly since 1957.

Once the energy is pouring (transmitted) from the dipole and out of the terminals of the generator to fill space around the external conductors, then a straightforward energy transmission and reception problem exists. The source dipolarity produces the steady emission and flow of energy filling space around the circuit. In the "receiving external circuit" the surface charges and their fields must intercept some of that available EM energy flow and diverge it into the conductors to power the circuit electrons. The intercepted and diverged component entering the circuit is the Poynting component, while the nonintercepted and nondiverged component missing the circuit and usually wasted, is the extra Heaviside component.

Fortunately, every source dipole pours out enormously more EM energy than our feeble conventional circuits intercept and catch to provide the Poynting calculations. So to produce COP>1.0 in the MEG, all we have to do is insure that we have (i) a mechanism easing the interception and collection process, and (ii) sufficient "antennas" containing sufficient conduction electrons (and their fields) to intercept and collect more energy than we pay to make the source dipole and perturb it. Since we pay nothing at all to make the source dipole (which is the permanent magnet), and since the magnetic flux back through the permanent magnet does not destroy the dipole, then all we have to pay for is the perturbation and for switching and control.

The resulting field energy generated outside the core localization region is primarily a function of the rise and decay times for the input perturbation pulses and not the energy of those pulses. From a little input energy with fast little changes, we can easily generate very powerful oscillating E-fields and provide dramatically increased "antenna collection" area with **the** necessary magnitude of intercepting charge. It rigorously follows that COP>1.0 is easily obtainable.[210, 211]

[210] Again, for the MEG, changing and controlling the rise time and decay time of the input pulses is a method for directly engineering and controlling the general relativistic curvature of the local surrounding spacetime, because it controls the magnitude of the E-field that is created by $\mathbf{E} = -\partial \mathbf{A}/\partial t$ from the curl-free A-potential outside the core. This also directly engineers and controls the magnitude of the polarization of the local active vacuum. That simple fact allows the use of a unified field theory operation, with a net energy interchange between the system and **the** local active vacuum. To understand the MEG, one must think in terms of the *supersystem.*

[211] All EM energy at any point in space must have been "gated" from the time-domain and it must return to the time domain (as, e.g., by the source charge for it). It **requires** quite a shift of one's mind to realize that all 3-space energy (such as "energy at each point in a potential") involves a gating process and a giant negentropic 4-circulation of EM energy ongoing at each point dipole. Much of the extraction of **energy** from the vacuum has been hidden by thinking in terms of "inputting energy" instead of "inputting or arranging energy flow gating" by the source charges and sourcedipoles.

The MEG is an open system far from equilibrium with its active environment (its active local vacuum and its active local curvatures of spacetime). By attending to the interception and collection part of the overunity problem, the MEG can permissibly collect and use more EM energy to power its loads than one oneself has to input to it. In short, one only has to input a little switching and perturbation energy with properly shaped waveforms, to generate very powerful fields and appreciable field energy available for collection and use. With sufficient collection mechanisms in place, the unit easily produces COP>1.0. Then with proper closed-loop principles implemented, the MEG can be self-powering.

7.7 Experimental Path to Development of the MEG

In our experiments with the Radus effect previously discussed in Chapter 6, paragraph 6.4.5, we had already done extensive work with switching magnetic fluxes for less switching energy than can be intercepted from the switched flux. We had also experimented with various cores for transformer and coil variants, including eventually the new layered nanocrystalline cores.

In our experiments, we noticed a remarkable thing. With certain modern nanocrystalline core materials, and using a permanent magnet to provide B-flux, we found that *the core material would extract and hold essentially all the permanent magnet's B-flux inside the core.* Field measurements right against the outside of the poles of the powerful little neodymium-iron-boron magnet itself showed very little B-flux spilling out into space! In short, we had discovered how to make a magnetic system that *nonlocally* propagated only the uncurled magnetic vector potential A, without any operator energy input. We realized we had discovered a marvelous and free way to freely obtain an excellent Aharonov-Bohm effect, similar to the way a nearly perfect toroid would localize the B-field inside the core, with the uncurled A-potential appearing in space surrounding the toroid. The advantage of our approach over the toroid was that we did not have to furnish any current and voltage to obtain the localized B-flux and the non-localized curl-free A-potential energy.

Further, if we perturbed the B-field flux in the core, we consequently perturbed the uncurled A-potential in space surrounding the core. From both sources — the perturbed B-flux in the core and the perturbed

A-potential outside the core — we could produce very large E-fields and concomitant very large voltages capable of driving more "power" through the secondary coils of the transformer section than the power required to perform the switching and perturbation.

We could easily prove this by perturbing the localized B-flux (and the concomitant surrounding uncurled A-potential automatically) with weak energy. We proved that increasing the rate of the rise time and decay time of even a weak input EM energy perturbation could produce quite large E-fields, including 1,000 to 5,000 volts (or even 20,000 volts!) appearing in the secondary coil if desired. We also proved another thing: *the special nanocrystalline core material in some buildups self-adjusted its permeability, continuing to localize the perturbed and alternating B-fields and their fluxes while freeing the resulting E-fields and passing them out of the core.* In short, the permeability of the specially selected core material automatically adjusted with the perturbation of the localized B-flux and nonlocalized A-potential, so that all the oscillating B-flux remained localized in the core. Further, the highly layered nanocrystalline material produced essentially no eddy currents at all, hence no heating, even in the presence of dramatic B-flux oscillation and dramatic E-field intensities in the surrounding uncurled A-potential space and in the output (secondary) coils of the transformer section of the MEG.

The absolute proof that the core changed its permeability was in the lack of a phase lag between voltage and current in the output from the output coils. Here a remarkable thing was observed: the output current was in phase with the output voltage, within about 2°. In short, under certain circumstances the output coils wound around the nanocrystalline core would exhibit essentially a purely E-field interaction, with only minuscule B-field leakage — something deemed impossible in the typical electrodynamics textbook. This meant that the excess E-fields produced by the perturbed A-potential in space outside the core also reacted with the output coil's conduction electrons in a purely electrical manner! This was because the B-fields locally produced in the output coils by the interacting changing E-fields, *were in fact instantly drawn into the core material as they were produced, and were held there so they did not react in magnetic fashion with the electrons in the output current from the secondary coils.*

So this is a dramatic localization of Faraday's law of induction. Consequently, one entire group of embodiments of the MEG uses a purely electrical interaction with the secondary (output) coils of the transformer section.[212] Or more exactly, the core material draws in all the magnetic fields from the reactions of the output coils, leaving only the electrical reaction outside the core — including in the coil conductors, which obviously are wound outside the core material.

On the other hand, the perturbed localized magnetic B-flux inside the core material did produce the normal E-field to be expected, as a function of the time-rate of change of the B-field. Such an E-field is not localized in the core material, but extends outside the core and interacts with the electrons in the output secondary coils of the MEG's transformer section. Again, the magnitude of this E-field is a function of the time-rate of perturbation of the localized B-field flux in the core.

The nanocrystalline core also has excellent high frequency response, with negligible eddy currents. Thus we can "fine tune" the engineering of the rise time and decay time of the input perturbation pulses.

Hence we accomplish several kinds of "gated energy amplification" in the MEG: (i) the amplified magnitude of the external E-field interacting with the external collectors (such as the output secondary coils) due to perturbation of the nonlocalized external uncurled A-potential, (ii) the amplified magnitude of an additional E-field interacting with the external collectors (such as the output secondary coils) due to perturbation of the localized internal B-field flux inside the core material, (iii) a large number of additional derivatives of higher order B-fields and E-fields are formed and utilized, and (iv) adding additional collectors (either coils wound on the core or external "receiving antenna circuits" produces additional power for heavier loads, for a given perturbation input.

We had exhibited a cored output coil with a remarkable adaptation of the AB effect. Essentially the output of the secondary coil was driven by the large electric field component created from the uncurled A-potential perturbations *and* by the large electric field component created from the B-field flux perturbations inside the core. This peculiar transformer-like section operated from the E-field fluxes emerging or produced *outside* the transformer core material, rather than from the magnetic field flux being produced and oscillated *inside* the core material itself.

---

[212] Again we accent that the magnetic field reactions do occur, and the B-fields are in fact produced. But the resulting B-fields are also extracted into the core material and localized therein. Even the magnetic B-fields produced in the input and output coils themselves, is so localized and not allowed to interact with the electron currents through the coils to produce the conventional 90° phase lag between voltage and current.

Obviously we had found a remarkable but legitimate COP>1.0 process and embodiment — in fact, a whole set of them. We had also invented a new kind of "transformer" or energy generator-transducer.

From that we progressed to discovering ways to a dual use of both the perturbed B-flux in the core and the E-fields from the perturbed A-potential outside the core. To first order, that turned out to be fairly simple: any changing B-flux inside the core material and thus inside the output (secondary) coils of the transformer section, produces not only B-field which is held and localized, but also the changes of this local "pinned" B-field further produces E-field oscillations which are not pinned in the core. Hence the electric field half of a normal transformer action could and did occur in the secondary coils, but the magnetic field half did not occur *in the coils themselves*. The MEG used and uses a nearly pure E-field interaction in the output coils of its transformer section secondary, where that net interacting E-field comes from a variety of E-field generating mechanisms.

We filed our first patent, and then filed a second patent application soon after. Presently we are preparing to file our third and fourth patent applications on additional phenomenology and mechanisms uncovered in our experiments. As we are writing this section, our first patent has issued {37}.

In the view of the present author, we have uncovered a remarkable new application of the Aharonov-Bohm effect and the geometric phase, one that is practical and eminently useful in novel new electrical power systems.

7.8 Additional MEG Functions

Figure 7-6 diagrammatically shows the core material inside the input coil for either open-loop operation using an external power supply or closed-loop (self-powering) operation using clamped and governed positive feedback. As can be seen, the B-field and B-field flux are confined to the core material (which is in the form of a closed transformer coil path, as shown diagrammatically in Figure 7-5). The external space surrounding **the** core material is filled with the uncurled A-potential, due to the Aharonov-Bohm effect being imposed by the core material. The two leads from the input coil are shown at the bottom of Figure 7-6.

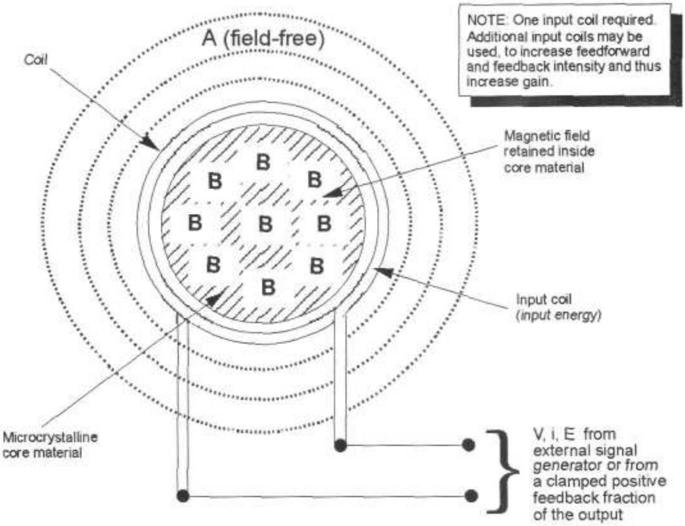
NOTE: One input coil required. Additional input coils may be used, to increase feedforward and feedback intensity and thus increase gain.
A (field-free)
Coil
B
Magnetic field retained inside core material
Input coil (input energy)
Microcrystalline core material
V, i, E from external signal generator or from a clamped positive feedback fraction of the output

Figure 7-6 Input coil for open- or closed-loop operation.

We pulse the input coils with nearly rectangular pulses each having a sharp leading edge and a sharp trailing edge, so that the resulting rapid time-rates of change of the two edges create very large perturbations of the B-flux in the core and of the A-potential outside the core. The large E_2-field resulting from the fast rate of change of the perturbed B-field in the core is not localized, but appears in space outside the core. In addition, an E_3 field created directly in the coil itself also appears in space outside the core. In addition, the sharply perturbed A-potential outside the core creates a third large E_1-field in space outside the core. The net E_t-field in the surrounding outside space is the vector summation of the E-fields produced by these processes, or to first order (neglecting the further differentiations)

$$\mathbf{E}_t = \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 \qquad [7\text{-}1]$$

We also may adjust the pulse width to generate a further Lenz law effect, increasing the three component fields E_1, E_2, and E_3 if desired. To do that, the pulse must be narrow enough to suddenly decay while each of the fields E_1, E_2, and E_3 is still rising and not yet reached maximum. As can be appreciated, considerable pulse shaping and timing can be adjusted for optimization of the output voltage, current, and power when interception and collection from E_t is accomplished in the secondary coils. In addition, it can be appreciated that slightly differing materials delay times do occur in the production **of $\mathbf{E}_1$,** E_2, and E_3. Hence there is appreciable optimization that can be and is accomplished on a given MEG variant and build-up type.

More complex waveforms of various shapes can be used to further optimize the output. It is also possible to accomplish further optimization and control by adjusting the shapes and timing of the feedback signals in the case of the closed-looped MEG system. For the externally powered open-loop version, additional optimization and control can be achieved by adding additional simultaneous pulses fed into the input coils in various fashions.

In addition, the response of the output coils produces feedback into the core material, into external space, and into the input coils. These in fact provide three more E-field *return* or feedback inputs, so that the total feedback input is

$$\mathbf{E}_f = \mathbf{E}_4 + \mathbf{E}_5 + \mathbf{E}_6 \qquad [7\text{-}2]$$

Without doing the more complex iterations for this feedback and feedforward set of loops, to first order the total input E_I to the input coil may be considered to be

$$\mathbf{E}_I = \mathbf{E}_t + \mathbf{E}_f \qquad [7\text{-}3]$$

Equations [7-1], [7-2], and [7-3] are for simple analogy only; in the real MEG, full nonlinear feedforward and feedback analysis techniques must be utilized because of the iterative operations, and much more sophisticated optimization techniques are appropriate. Also, chaotic oscillations can and do appear, and techniques for control of such oscillations are indicated, particularly along the lines indicated by Fradkov and Pogromsky {487} without use of probability etc. Quoting:

> *"1. There is ... great benefit of using the modern nonlinear and adaptive control theory. ...2. There is no need to distinguish periodic and chaotic behavior. Accurate control is possible without accurate prediction. ...3. There is no need to define chaos in order to control it. ...4. There is no need to use probability in order to control systems with seemingly random behavior. "*

For any given MEG build-up design, one experiments with these optimization and control factors and mechanisms to provide the best performance, and to provide rock-steady performance.

Figure 7-7 shows the output coil for either open-loop operation from an external power supply or closed-loop (self-powering) operation using clamped and governed positive feedback. Here the situation is partially reversed. There are three major inputs to the output coils, and these are (i) the external input from the total outside E-field, (ii) the internal input from the perturbation changes in the confined magnetic flux inside the core through the output coils, and (iii) the ensuing feedforward and feedback from all E-fields.

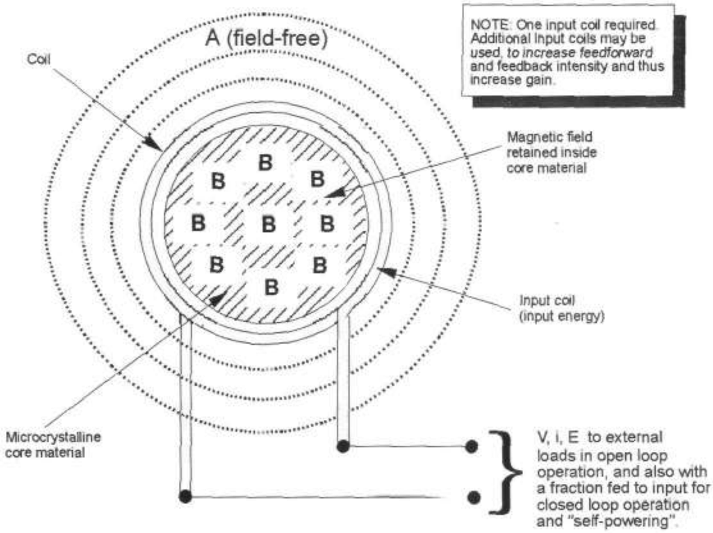
A (field-free)
Coil
Magnetic field retained inside core material
B
B
B
B
B
B
B
B
B
Input coil (input energy)
Microcrystalline core material

NOTE: One input coil required. Additional input coils may be used, to increase feedforward and feedback intensity and thus increase gain.

V, i, E to external loads in open loop operation, and also with a fraction fed to input for closed loop operation and "self-powering".

Figure 7-7 Output coil for open- or closed-loop operation.

Again, this is simply an easy way to visualize the situation to first order. In the real case, sophisticated feedforward and feedback analysis is required and utilized. Optimization usually consists of two steps: (1) the simpler, more direct optimization techniques such as adjusting pulse width, pulse rise time and decay time, actual pulse form, etc., and (2) the deeper and more complex optimization arrived at from detailed feedback and feedforward loop analysis as well as nonlinear oscillation and nonlinear control theory analysis.

In the actual optimization analysis, the feedback and feedforward loops are mathematically endless and thus should be optimized for quick decay and stability. On the other hand, the MEG also has the ability to self-oscillate over the entire unit, in which case the feedback and feedforward loops are optimized for stable self-oscillation of the entire unit. The self-oscillating

MEG is permitted since it is an open system far from equilibrium with its active environment.

So if desired, an entirely different kind of optimization can be utilized for self-sustained nonlinear oscillation and closed loop operation. Because of the gain achieved in the large E-fields produced by the perturbed uncurled A-potential in space surrounding the core, there is sufficient excess energy collection available in the output coils to use the available feedback and feedforward loops in the system itself to achieve system self-oscillation. The system can be operated as a self-oscillating open system freely receiving excess EM energy from its active environment — the local active vacuum, local curvatures of spacetime, and local freely furnished uncurled A-potential and resulting nearly free E-fields from the A-potential's perturbations). As we are writing this, our experiments are now in this direction, and we are preparing the material for filing a formal patent application on this additional process. The patents will be filed, of course, by the time this book is published.

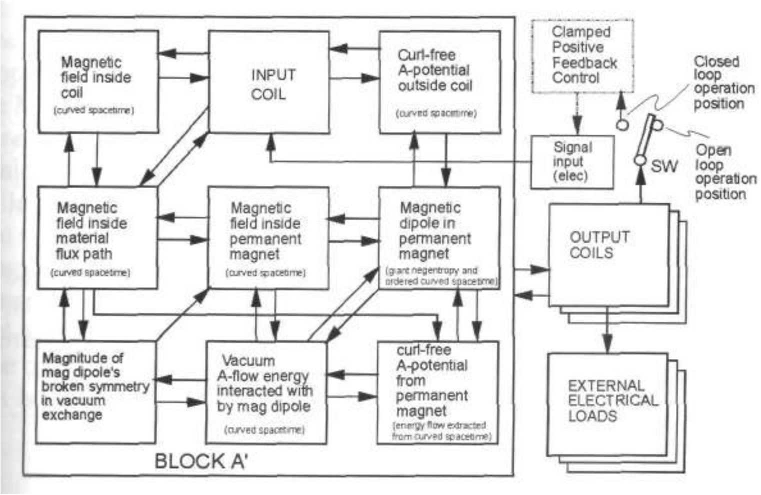
Magnetic field inside coil
(curved spacetime)
INPUT COIL
Curl-free A-potential outside coil
(curved spacetime)
Clamped Positive Feedback Control
Closed loop operation position
Signal input (elec)
SW
Open loop operation position
Magnetic field inside material flux path
(curved spacetime)
Magnetic field inside permanent magnet
(curved spacetime)
Magnetic dipole in permanent magnet
(giant negentropy and ordered curved spacetime)
OUTPUT COILS
Magnitude of mag dipole's broken symmetry in vacuum exchange
Vacuum A-flow energy interacted with by mag dipole
(curved spacetime)
curl-free A-potential from permanent magnet
(energy flow extracted from curved spacetime)
EXTERNAL ELECTRICAL LOADS
BLOCK A'

Figure 7-8 Energy gain process using feedforward and feeback subprocesses

A highly simplified block diagram of energy gain process using feedforward and feedback subprocesses is shown in Figure 7-8. As can be appreciated from this high order diagram, the actual feedforward and feedback loops are extremely complex, and a great deal of complex analysis is required for a self-oscillating, stable, controlled system. We also stress that control measures to insure and hold stable self-oscillation must be included in the overall features implemented in this self-powered, self-oscillation variant of the MEG. Not only must we use chaotic oscillation techniques to achieve the self-oscillation, but also we must use advanced techniques for stable control of such nonlinear oscillations, as indicated by Fradkov and Pogromsky {487}.

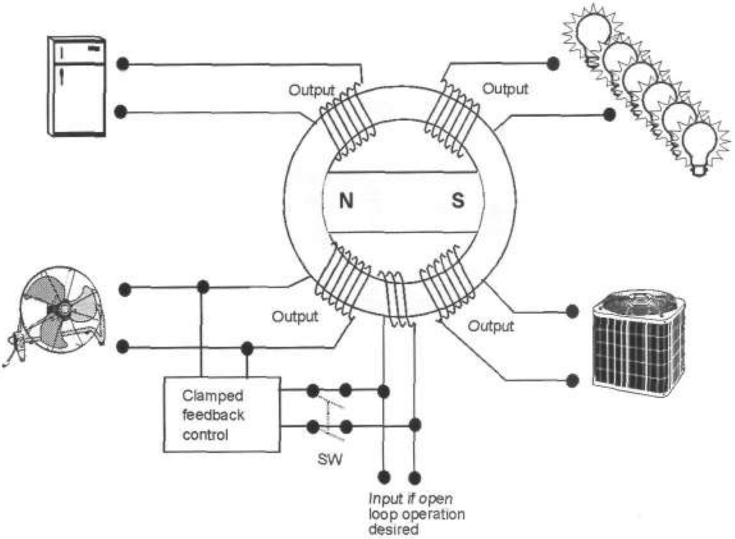
Output
Output
N
S
Output
Output
Clamped
feedback
control
SW
Input if open
loop operation
desired

Figure 7-9 Typical MEG embodiment application on site.

In Figure 7-9, we show a typical diagrammatic embodiment envisioned for the MEG in a household, using a large MEG to later be developed. For interim installations, the single large MEG will be replaced with an array of MEGs together with a synchronizer, as shown in Figure 7-10.

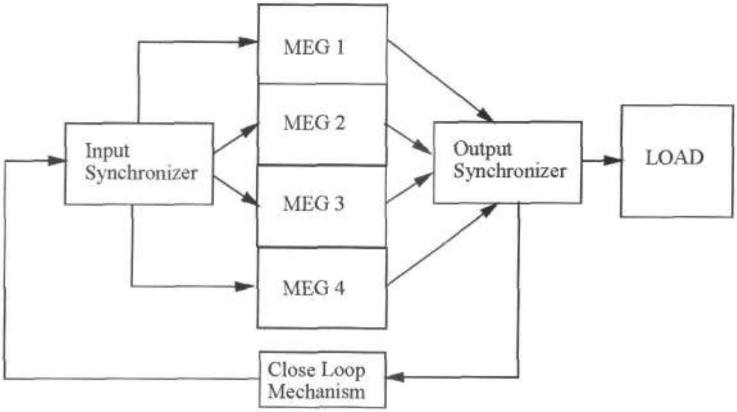

Figure 7-10 MEG self-powering array concept for increased powering applications.

Figure 7-11 shows a MEG unit where additional special collection antennas feeding separate circuits are employed in an "outrigger" fashion. One major advantage of this configuration is that the complexity of close-looping and its control are materially reduced, although the complexity of the MEG build-up and stabilization is increased. Another advantage is the *ease* with which close-looping for self-powering operation can be achieved, as shown in Figure 7-12. In this case, the external antennas are collected to independent output circuits (not shown) in the output section, and each of these independent outrigger outputs is a COP<1.0 circuit if we count the perturbed A-potential energy input as an "operator's" input. Completely normal clamped positive feedback can be used with sensing and governing, to allow self-powering and self-adjusting to varying loads. The outrigger embodiment can be made to almost entirely avoid the Dirac sea hole current problem.

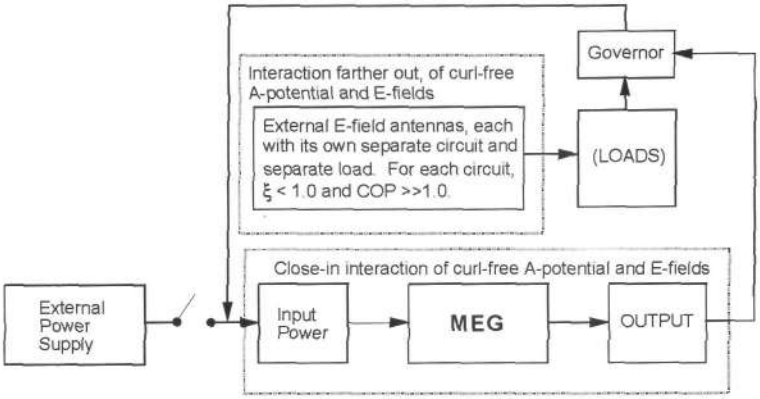
Governor
Interaction farther out, of curl-free A-potential and E-fields
External E-field antennas, each with its own separate circuit and separate load. For each circuit, ξ < 1.0 and COP >>1.0.
(LOADS)
Close-in interaction of curl-free A-potential and E-fields
External Power Supply
Input Power
MEG
OUTPUT

Figure 7-11 MEG unit with outrigger antennas for extra energy collection.

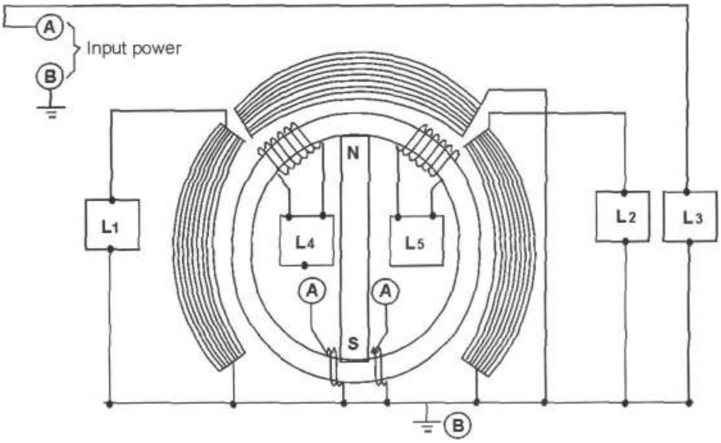
A
B
Input power
N
S
L1
L2
L3
L4
L5
A
A
B

Figure 7-12 Use of outrigger antennas for damped positive feedback method of close-looping.

As can be seen, the MEG has many possible configurations and embodiments, each of which has special advantages and uses. One of the embodiments we discussed is the deliberate use of self-oscillation forced by the feedforward and feedback loops. That is a separate embodiment, and capable of a very high power per pound ratio, but it also requires more sophisticated sensing and control techniques.

7.9 In Conclusion

By the time this book is published, Magnetic Energy Ltd. is hopeful it will have achieved its major funding requirement, and be well along the way to completing the research for full production and marketing of MEG units on the world market. We have arrived at that part of development — the need for substantial funding (several tens of millions of dollars) — where every previous COP>1.0 EM power system has died. We intend to do everything within our power to see that we obtain the funding, finish the complex research task, and get the MEG onto the world market.

For additional information on the MEG and on its preliminary replication by Naudin, the reader is referred to other technical articles {488} and to Jean-Louis Naudin's website.[213] Actual measurements for one of Magnetic Energy Ltd.'s MEG test units is given in reference {488a}. The MEG patent {37} shows the variation of COP for a test MEG with various conditions. Curves for Naudin's measurement of the performance of his replicated MEG unit are given on his website.

The MEG project is a most interesting project, because of the involvement of so much advanced physics and several different physics areas, as well as the prerequisite for higher group symmetry electrodynamics. Once funded, we are confident we succeed in crashing through the barriers and getting working MEG power units — including self-powering MEG units — on the market. We also believe the MEG will represent a substantial contribution to alleviating the coming worldwide energy crisis. Hopefully it will materially hasten the growth and evolution of the present highly centralized — and terribly vulnerable — national energy infrastructure into a highly decentralized — and much less vulnerable — national energy infrastructure. We also believe it will provide a great contribution to helping alleviate global warming, reducing pollution of the biosphere, and reducing the production of hazardous wastes. If so, it will be a continuing contribution, not only to humanity today, but also to generations yet unborn.

---

[213] Naudin's website MEG material and information is at http://jnaudin.free.fr/html/meg.htm. Particularly see his Mark III replication, shown powering a light on that webpage.

Chapter 8

Approach to Antigravity

[Large-scale antigravity exists] *"...baryonic, ordinary matter—the stuff of stars and of people—makes up just over 4% of the energy and matter in the universe. ...about 30% of the stuff in the universe is dark [unobserved] matter. The remaining two-thirds, theorists believe, is a mysterious 'dark energy' or 'quintessence'—a large-scale antigravity-like effect that is making the universe expand ever faster.... "* [Charles Seife] {489}.

[On the nature of the vacuum] *"...the (electromagnetic) vacuum is the ground state of the quantized electromagnetic field. ...The Dirac electron vacuum is that state with all negative energy electron states fully occupied."* [Aitchison] {490}

[On forces and how charge exists] *"Vacuum polarization, in general, alters the effective force law. Forces, in quantum field theory, are understood as being due to the exchange of virtual quanta ...In the case of QCD and QFD, ... a crucial new feature is that the force-field quanta themselves carry the 'charge' of the force-field, i.e. it is as if the photon of electromagnetism carried electromagnetic charge."* [Aitchison] {491}

[Possible gravitational circuits] *"There is an implication ... that gravitational excursions might be controllable by gravitational 'circuits', analogous to electronic circuits, which could lead to a controllable form of negative gravity."* [Robert Neill Boyd] {492}.

[Gravitational circuits] *"The supersystem concept utilizes a unified fields approach so that specialized electrical circuits in COP» 1.0 condition generate strong Dirac sea hole currents of negative energy in the immediate local vacuum. Thus a strong negative energy field is formed surrounding the circuit and changing the curvature of local spacetime as an action by the source*

> *circuit. This produces a direct antigravity field around the circuit or physical system, acting upon the mass of the system. The net gravitational interaction by that circuit or physical system is therefore the summation of the local positive gravitational field and the new local negative gravitational field produced by that negative energy field."* [T. E. Bearden, correspondence to a colleague, February 2002.]
>
> [On the multiplicity of spacetime curvature components] *"The full mathematical expression for the curvature of four-dimensional space-time ...is given by something called the Riemann curvature tensor. This is a somewhat complicated object, needing twenty real numbers at each point in order to specify it. These twenty numbers are referred to as its* components. *The different components refer to the different curvatures in different directions in the space-time."* [Penrose] {493}.
>
> [Physics studies and prepares imperfect models, not perfect truth] *"All we ever know is our models, but never the reality that may or may not exist behind the models and casts its shadow upon us who are embedded inside it. We imagine and intuit, then point the finger and wait to see which suspect for truth turns and runs. Our models may get closer and closer, but we will never reach direct perception of reality's thing-in-itself"* [Hawking] {494}.

8.1 Introduction

In this chapter, we first present a highly abbreviated background of the theory of gravitation, sufficient for our experimental orientation. Of special interest are unified field theories, which cover general relativity, electrodynamics, and quantum mechanics in one overall master theory. Even more of interest is an engineerable unified field theory, which is presently closest approached by the theory of Mendel Sachs {510}.

We then present a short overview of work that has been done in the fields of lifters and inertial propulsion, listing some of the inventions of interest, the approaches, results achieved, and our comments on each.

Finally, I present my own approach to antigravitation, and give some results of the Sweet device's antigravity test (the test was designed by the present author) that successfully reduced the weight of a device by 90%.

In our view, this experiment fully demonstrated the veracity of our theoretical approach.

8.2 Selected Theories of Gravitation

8.2.1 Newton's Theory and a Negative Mass Extension

The first mechanism for gravity of interest to our purposes was advanced by Newton in his work on mechanics {495}. Newton's law of gravitation may be stated as {496}:

> *"A particle of mass m_1 attracts a particle of mass m_2 a distance d away with a force* $F = G\frac{m_1 m_2}{d^2}$ *in the direction of the line joining the particles."*

Of course, it should have stated "in a time forward zone". The situation is precisely reversed in a time-reversal zone. It is also reversed for negative mass-energy.

In the Newtonian view, antigravity would mean repulsion of mass rather than attraction of mass. Given that negative mass is possible, then a negative mass would exhibit antigravity — i.e., it would be repelled by all other positive masses.[214] This can be seen in the quoted equation, by

---

[214] The discussion of what attracts what and what repels what, in a mix of negative and positive masses and negative and positive energies, is still controversial in physics. Different theorists have taken differing positions, and there is no formal experiment that has decided between them. In our opinion, the successful antigravity test of the Sweet vacuum triode amplifier answered that question experimentally, although independent replication was not accomplished. The scientific discussions heavily involve discussions (and assumptions) of the nature of "gravitational" mass, inertial mass, and electrodynamic mass etc. Again, there is no decisive selection among several possibilities, and there is no clear and universally accepted definition of **any** of those three types of mass.

In addition, all the discussions have included a fundamental error: They **assume** a separate 3-force acting upon a separate 3-mass, which is unequivocally false. Another fundamental error is the assumption of an observed (effect) 3-positron interacting with normal matter rather than a non-observed (causal) 4-positron (negative energy 4-electron) in massfree spacetime. In effect, this applies the CPT (charge-parity-time) symmetry, and converts the direction of the 3-positron to the antidirection of the 4-positron while also changing the negative mass of the 4-positron to positive mass of the 3-positron. Further, in physics there is still no firm **definition** of either force or mass, but only circular definitions between the two. We are using the simple definition of mass as $F= d/dt\ (mv)$, where mass is a component of force F and hence there is no separate force F acting upon a separate mass m. We interpret that equation as "the force *F* observed in a new observation is proportional to the time-rate of change of the ongoing 4-space interaction between the just-observed 3-mass *m* and the average interchange of time-energy and space-energy represented by v in the ongoing dynamics between spacetime and mass *m.* Whatever adjective is prefaced upon the *F* is also prefaced upon the *m* in the *mv* term.

We politely but firmly disagree with one prevailing notion that positive mass *attracts* both positive and negative mass, while negative mass *repels* both types. Further, we believe that the Sweet vacuum triode amplifier antigravity experiment falsified that notion. We hold to the dictum that $m == E/c^2$, where E is real spatial energy and may be positive or negative, c^2 is always positive, and therefore negative mass is highly compressed negative spatial energy while positive mass is highly compressed positive spatial energy. (Imaginary mass is a separate consideration.) We also prefer to reason in "single fundamental units" physics where the joule is the only fundamental unit. Gravity then becomes a function of the joule, whether electromagnetic, mechanical, or whatever. *We consider the energy in spacetime to he the source of the curvature of spacetime that is known as " gravitational field".* If the joule is of positive energy, the accompanying gravity is positive. If the joule is of negative energy, the accompanying gravity is negative. And we consider gravity as the 4-space causal spacetime curvature. So for gravitational interaction theory, we insist on examining the interaction of the curvature of local spacetime with the mass-energy. Further, we postulate that the gravitational effect upon one mass from a second mass is an improperly posed concept. It actually is an effect of the local curvature of spacetime around the first mass, where that curvature of spacetime was induced by the second mass's energy. Further, we consider that positive energy fields surrounding a positive mass will curve spacetime in the "positive gravity" direction. The positive mass is attracted by all other positive masses. Negative energy fields surrounding that positive mass system curve spacetime in the "negative gravity" direction so that the positive mass of the system is at least partially repelled by all other positive masses in the universe.

We further prefer to think in terms of the *supersystem* and the unobserved causal 4-field, not the frozen iterative 3-field snapshots defined only after iterative . observations. As we stated previously, the "observed" 3-positron is, in our view, a very different beast from the as-yet-unobserved 4-positron. The 3-positron is a frozen momentary "effect" entity with no dynamics at all and no continuous existence in time (and hence in spacetime) at all; instead, it continually *recurs* as the iterative effect of iterative observation by iterative application of the *d/dt* observation operator. The 4-positron is still a negative energy electron with negative charge, and still going in a non-reversed direction since broken parity has not been applied (because observation has not been applied), a causal entity with ongoing dynamics, continuously existing in spacetime, not just in 3-space. Until interacted and observed, it also has negative energy and negative mass-equivalency.

If one wishes to accurately deal with spacetime effects from the "positron," in our view one must deal with the 4-positron before observation, and hence with the supersystem. The 3-positron is not a *spacetime* entity, but only a frozen *3-spatial* snapshot. Scientists, including Dirac himself, were horrified at the negative mass implications of the negative energy states. Therefore, they moved the interpretation simply replacing positive mass m_1 by negative mass $-m_1$. That two negative masses would attract can be seen by replacing positive mass m_1 by negative mass $(-m_1)$, and also replacing positive mass m_2 by negative mass $(-m_2)$.

Now consider mass m_1 to be the rest of the net positive mass in the universe, and mass m_2 to be the "effective" mass in a local supersystem consisting of the local physical system, its local active vacuum, and its local active curvatures of spacetime — in short, in a local supersystem consisting of the physical system and its local active environment, and the interchanges between them. Let $m_2 = m_p + m_n$, where m_p is the amount of local positive mass (or mass-equivalent) in the system and m_n is the amount of negative mass (or mass equivalent) in the system. This supersystem now simultaneously exhibits both positive gravity (attraction) to all net positive masses in the universe and negative gravity (repulsion) to them as well. The net gravitational response of the supersystem depends upon the fraction of the effective mass that is positive and the fraction that is negative — in short, upon the net vector sum of the two gravitational forces.

to the 3-positron, and thereby twisted the interpretation to substitute the effect for the cause. Hence circuit theory treats the *lattice hole* as a positron, yet that "hole" acts very differently from a true Dirac positron (a Dirac sea negative energy electron) in free space.

Consequently, matter-antimatter interactions and the interaction between positive energy and negative energy are quite different, or they can be if so manipulated prior to observation. In the matter-antimatter interaction, the initial curvature of spacetime formed by the union decays by radiation of EM positive energy and the "effect particles" disappear back into the uncurved vacuum from whence they came. In the positive energy and negative interaction, there is a change of the local curvature of spacetime without its decay and subsequent re-radiation of EM energy. Equal amounts of interacting positive and negative EM field energy thus are simply transduced into stress energy potential of the vacuum and consequent sustained change in the curvature of the local spacetime. Indeed, the radiation of pair annihilation energy is obtained only at the expense of negating (relaxing) the local polarization of the vacuum (and the two consequent curvatures of local spacetime) represented by the independent positron and electron. Indeed, normal pair annihilation is a sort of Lorentz symmetrical regauging of the locally asymmetrical vacuum.

The reader is certainly free to disagree with our conclusions; we are merely explaining them. Experiment will eventually sort out the truth or falsity of the approach. So far, based on the Sweet experiment designed by the present author, a highly successful experiment was consistent with this approach.

In a "mixed" mass system containing both positive and negative component masses, the fraction of the total mass that is positive and the fraction that is negative is thus significant. When the mixture's negative mass is less than its positive mass, the mixture would exhibit a "reduced" gravitational attraction to other masses of the universe. Then the ratio of negative mass to positive mass might be taken as the "gravitational shielding" coefficient (fraction).

When the amount of positive mass and the amount of negative mass were equal, the overall mass system would exhibit no net gravity or antigravity, and would "float" or "hover" in the Earth's gravitational field. It would be equally attracted and repelled by the Earth's positive mass (and by all other positive mass in the universe, hence could — at least under certain conditions — *appear inertia-free.).* When the amount of negative mass exceeds the amount of positive mass, the overall mass system would exhibit a net antigravity effect. It would be repelled more strongly by the Earth's positive mass, as well as more weakly by the distant positive masses of the universe, and hence would accelerate upwards and away from the Earth. Notice that the key to antigravity in the Newtonian approach is the availability of negative mass, but in the new sense in which we use it. Note also that we did not discuss the separate issue of imaginary mass.

If we translate that into modern terms, where we consider not the "mass" but the condensed energy comprising the mass, then m_p and m_n become positive mass-energy and negative mass-energy terms respectively. Given the ability to produce electromagnetically produce substantial negative field energy in the immediate vacuum environment of the system, where the source of the field is the physical system, then the creation of a direct antigravity component in the system's environment — the other two components of its supersystem — becomes not only possible but also feasible.

The key to antigravity becomes the *availability of negative field energy in the local space surrounding the physical object, with the physical object containing the source of that surrounding spatial negative energy field.* Voila! This impels us directly to electrodynamics and a modified Dirac sea {497, 498a, 498b), one with excess empty holes as source charges producing negative energy EM fields. We deliberately create these unique source charges by actions taken in a COP»1.0 system. This gives the key to the approach to antigravity taken by the present author.[215] However, we use as negative energy not just the accounted *Poynting* energy flow component in space where the field is present, but also the unaccounted *Heaviside* energy flow component, which is much greater in magnitude (in a nominal case, nearly 10^{13} times greater). In this way, we dramatically increase the amount of antigravity effect that can be obtained by a given EM power dissipation in the physical electrodynamic system itself. *This antigravity effect of a Heaviside component of EM negative energy flow has not been accounted in previous analyses.*

8.2.2 Riemann Surfaces and Riemann Geometry

The next development of interest to us is the doctoral thesis of Riemann {499}, which founded the theory of Riemann surfaces. Three years later, he founded Riemann geometry in a famous lecture. Riemann was working on a unified explanation of gravity and electromagnetism when he died of tuberculosis a dozen years later. It was Riemann's multiply-connected topologies that eventually resulted in the view of electric charge producing lines of force trapped in the topology of space {500}. Had Riemann lived, we believe that engineerable unification of electrodynamics and gravity would have occurred early on. In that case, we might have already had practical antigravity vehicles and explored the solar system and beyond.

8.2.3 Heaviside's Excess Energy Circulation Flow and Gravity Theory

The next theory of gravity of interest to us is that of Oliver Heaviside. Heaviside noted that his extra EM energy flow term was a trapped EM energy circulation {501}. He realized that, if that trapped EM energy circulation were taken as gravitational in nature, his energy flow theory could be converted into a unified theory incorporating both electrodynamics and gravitation. After his death, his handwritten notes on the theory were found beneath the floorboards in the little garret apartment he had long occupied. No one knows what other papers he may have had; the place was ransacked and looted prior to the discovery of these remaining papers hidden beneath the floorboards — papers the looters had missed.

---

[215] We also point out that a Dirac sea with empty holes (excess negative energy states) exists in (or constitutes) a negatively curved spacetime *a priori.* So we get back to the other two components of our "supersystem", the active vacuum and its dynamics and the local curvature of spacetime and its dynamics.

Other scientists such as Laithwaite have felt that Heaviside's approach might still shake the foundations of physics itself. Quoting Laithwaite {502}:

> *"Heaviside had originally written the energyflow as S = (ExH) + G, where G is a circuital flux. Poynting had only written S = (ExH). Takingp to be the density of matter and e the intensity ofa gravitationalforce, Heavisidefound that the circuital flux G can he expressed as pu — ce, where u represents the velocity of p and c is a constant."*

8.2.4 Mach's Principle

After Heaviside, the next item of interest to us is Mach's principle {503}, which essentially states {504}:

> *"The properties ofspace have no independent existence but depend on the mass content and distribution within it. Thus, acceleration dealt with in Newtonian mechanics can only have meaning ifit is an acceleration with respect to the stars or with respect to something equally well-defined."*

Mach's principle does not adequately deal with spacetime, and certainly does not even acknowledge or encompass how energy exists in 3-space at each point dipole, as an entering flow from the time domain into the negative (charge x time). It does not include transduction by the negative charge into 3-space energy, passage of the 3-energy to the associated positive charge, transduction back into time energy by the positive charge, and re-emission of the transduced time-energy back into the time domain, by the positive charge. In short, Mach's principle will never get one a practical antigravity system, regardless of how one applies the principle.

8.2.5 Einstein's General Relativity

Next comes Einstein's epochal general relativity theory of gravitation {505}. In that theory, Mach's principle is interpreted as pointed out by Misner, Thorne and Wheeler {506} in the following quote:

> *"...Einstein's theory (1) identifies gravitation as the mechanism by which matter there influences inertia here; [and] (2) says that this coupling takes place on a spacelike hypersurface..."*

The great contribution of Einstein's general relativity was to remove the notion of the "field occupying space", where field and space are separate things. Quoting {507}:

> *"If we imagine the gravitational field, i.e., the functions* g_{ik} *to be removed, there does not remain a space of the type (I) [Minkowski Space], but absolutely* nothing, *and also no 'topological space'. For the functions* g_{ik} *describe not only the field, but at the same time also the topological and metrical structural properties of the manifold. ... There is no such thing as an empty space, i.e., a space without field. Space-time does not claim existence on its own, but only as a structural quality of the field."*

In passing, we note that in the conventional sense we can and do have many fields "in the same space", which is merely the statement that a net field has internal structure, and so does a net spacetime. Extending that notion, we can conclude that spacetime is in fact naught but a set of fields, both static and dynamic. The total energy density of all those fields at any point in space constitutes *a potential* with that energy density at that point, and with that internal field dynamics and structure. So because of its incredible field energy density, space is also *identically* a giant potential with internal structure and dynamics. With that observation, we rather casually join the seething vacuum of particle physics, the spacetime of general relativity, and the fields of electromagnetics as *different aspects of the same thing.* We point out that our concept of the supersystem therefore only identifies the three separate things into which this "one-thing" has been arbitrarily divided in much of physics.

Although he tried more than once, Einstein himself failed to achieve a successful unified theory of gravitation and electrodynamics. One thing he added in his earlier theory was a cosmological constant, which he felt was necessary to explain a stable closed universe. He later regretted that assumption, but today with the discovery of the acceleration of the expanding universe, which can only be caused by antigravity, his cosmological constant has assumed a new and vital role in the search for the source of this observed antigravity effect in the cosmos.

8.2.6 Kaluza-Klein 5-Dimensional Unified Theory

Using five dimensions, a unified theory of general relativity and electrodynamics was advanced by Kaluza {508} as extended by Klein (509}. This is a notable extension of Einstein's theory, and a very substantial literature exists in the Kaluza-Klein area.

8.2.7 The Sachs Unified Field Extension of Einstein's Theory

Another notable extension of Einstein's theory, which does include electrodynamics and the results of quantum mechanics as well, is the theory by Sachs {510}. Starting with Einstein's general relativity view and the group symmetry of its underlying mathematics, Sachs extended Einstein's theory into a fundamental theory of matter in all domains. Einstein's group formalism and the spinor-quaternion calculus in a curved spacetime provide the genesis of matter fields. The theory encompasses globally covariant field equations for inertial features of matter including its force manifestations. Unification is provided by the generalized spinor formalism. Quoting a popularized summation {511}:

> *"Novel features of general relativity are shown to emerge, in elementary particle physics as well as astrophysics. Especially noteworthy ...is a factorization of Einstein's field equations to a quaternion field formalism which, by way of a single covariant field, includes both gravitation and electromagnetism coupled to matter field equations — whose linear (low energy) limit recovers quantum mechanics. The ...original approach also furnishes derivations, not previously available, of several observed properties of inertial mass and gravitational force. "*

For a more technical description of how the unification of electrodynamics with general relativity is derived, we quote directly from Sachs {512}:

> *"It was demonstrated earlier that the factorization of Einstein's second-rank, symmetric tensor field equations, by removing the space and time-reflection elements from the underlying covariance group, yields a 16-component quaternion metrical field equation. The 16 equations were then re-expressed in a second-rank tensor form, breaking up into* $10 \oplus 6$ *equations. The 10 equations are in one-to-one correspondence with Einstein's symmetric tensor equations for gravitation; they are even under reflections. The remaining 6 equations are in antisymmetric tensor form; they are odd under reflections. Taking the covariant divergence of the latter equations then puts them into one-to-one correspondence with the structure of Maxwell's equations for electromagnetism. What was shown earlier was that the four of Maxwell's equations with sources followed. What is shown in this paper is that the other four of Maxwell's equations that are source-free also*

> *follow, thus indicating the absence of magnetic monopoles from this theory, as in the usual Maxwell formalism. This shows that the factorization of Einstein's field equations to the quaternion form fully unifies the gravitational and electromagnetic manifestations of interacting charged matter."*

To freely download several of Professor Sachs's important technical papers, one is referred to his website at http://www.compukol.com/mendel/articles/articles.html.

The O(3) electrodynamics of Evans and Vigier has been shown to be a special subset of Sachs's electrodynamics {513a}. Hence O(3) can be used to engineer much of the Sachs unified field theory phenomenology, including — in our view — both gravity and antigravity. The first theoretical indication of this has been published by the AIAS (Alpha Foundation's Institute for Advanced Study) {513b}. Shortly we discuss our own views on how to directly perform engineering of antigravity.

A great many excellent publications on general relativity and the background of scientific knowledge in gravity exist; e.g., the text by Wheeler *et al.* {514} is particularly good. Einstein himself wrote a book on his special and general theory for the nonspecialist {515}. A collection of English translations of some of the original papers on relativity is provided in a book by Einstein, Lorentz, Minkowski, and Weyl {516}. Many other gravitational theories or changes to major theories have also been proposed by other scientists, but are not discussed here. For our purposes, an abbreviated discussion will suffice to understand a fundamental approach to feasible antigravity.

Since we are interested in actual working systems, a short review of some of the efforts in asymmetric lifting and propulsion experiments is useful.

8.3 Selected Antigravity and Inertial Propulsion Research Efforts

We give only a small sampling of the many efforts performed or ongoing in this area over the last few decades, and some still ongoing today.

8.3.1 T. Townsend Brown

Our first example is the work of T. Townsend Brown. Brown was working **on** a "gravitor" before he finished high school, and continued work in electrogravitation throughout his life. Space is a dielectric, and Brown reasoned that the displacement current in a capacitor dielectric could influence the space dielectric itself.[216] Hence he built, tested, and patented several variations of lifting devices based on this concept {517a-517e}. Essentially, he found that certain high capacitance dielectrics could be subjected to high voltage (50 to 300 kilovolts), and would exhibit thrust in the direction from the negative to the positive pole.

Here we point out our contention that all EM energy in space comes from the time domain via the negative charge, and re-enters the time domain via the positive charge. This is consistent with Brown's finding. Although he demonstrated many such lifting and thrusting devices, Brown was never able to obtain sufficient financial and scientific backing to bring a practical embodiment of his inventions into practice. His work, however, has been replicated by multiple experimenters, including NASA researchers (see below) and several researchers contributing results to Jean-Louis Naudin's website http://jnaudin.free.fr/html/lifters.htm.

8.3.2 Jonathan W. Campbell

Recently a U.S. patent {518} on a thrusting capacitor, by Campbell, has been issued and assigned to the U.S. Government (NASA), and made available for commercial licensing. The patent write-up states that thrust from capacitors is well known, but has been too small to be practical. The patent purports to remedy that impracticality. A version of the NASA device has been replicated by the Naudin researchers also. So an adaptation of Brown's work has been patented by the U.S. government with intentions of licensing. This would appear to validate much of Brown's approach and efforts.

8.3.3 The Dean Drive

Our second example is the inertial drive patented by Dean {519a-519e}. Unfortunately, the effects of the Dean drive later appeared to be a problem in accurate measurement. Nonetheless, work still continues along the direction shown by Dean. One may argue that mechanical materials also contain charges and have a certain kind of capacitance. Hence, mechanical motion of the parts of the machine does move these charges and dielectrics, opening a possibility of interacting with the dielectric of space. For example, Sweet's VTA involved a sustained self-oscillation between the local vacuum and the barium nuclei in his barium ferrite magnets.

---

[216] We have not yet explored the gravitational effects — if any — of our concept of the triode capacitor, discussed in paragraph 6.3.7 and diagrammed in Figure 6-17. We will examine it as soon as we have the proper equipment.

Dean's approach still has merit, even if a particular embodiment gave a negative result when very accurate measurement was performed. It is particularly of interest that mechanical force is essentially derived from electromagnetic forces (and fields) at basis. These base EM fields involve (and identically are) curvatures of spacetime, and hence the motions of the inertial drive mechanical apparatuses do involve at least some curvatures of spacetime and dynamic changes to them.

Our comment is that, when the additional but neglected Heaviside nondiverged energy flow component is accounted, these field interactions internal to matter can produce many orders of magnitude greater curvature of spacetime than is accounted by the diverted Poynting energy flow component only.

The "inertial antigravity" approach may be visualized as the use of excess gravitational thrust in a specific direction. If that direction is against the local pull of gravity, and is more powerful, then "lift against the gravitational pull" is obtained. Rigorously, this is not true antigravity but manipulated excess positive gravity.

In our approach to antigravity, one way to approach the problem is to have the mechanical apparatus also the source of an intense *negative energy* EM **field,** producing an intense flux of Dirac sea holes into and in the local surrounding spacetime. The excess charge removed from the Dirac holes can in fact be used in the electrical powering of the physical system, as was demonstrated in the Sweet VTA antigravity test. Then movements of the mechanical parts could involve movement of strong negative energy fields, hence strong curves of local spacetime that are local *strong negative gravity fields*. Or, better yet, movement of the charges themselves will also produce field-induced movement of the Dirac sea hole negative energy. This appears to be a practical method to manipulate the metric itself, along the lines proposed by Puthoff *et al.*[217]

8.3.4 Henry Wallace

Other work has been by Henry Wallace, an engineer working for General Electric in the 1970s. His work dealt with the interactions of relatively moving masses of various kinds, and it is described in his patents {520a, 520b}. The discussion under the Dean drive probably applies to Wallace's work as well.

---

[217] H. E. Puthoff, S. R. Little, and M. Ibison, "Engineering the Zero-Point Field and Polarizable Vacuum for Interstellar Flight," *J. British Interplanetary Soc,* Vol. 55, 2002, p. 137-144.

8.3.5 J. F. Woodward

Next is the work of J. F. Woodward, resulting in one or more patents {521} and multiple articles in the legitimate physics literature {522a-522k}. Woodward uses an adapted theory based on Mach's principle and on the thesis that the mass of a capacitor can be made to vary from its normal mass by rapidly charging and discharging it — say, at a frequency of 20 kHz. This predicted effect is testable, although some sophistication in the testing is required.

Woodward's publications show the results of actual experiments, measuring a very small change in the mass, and are worthy of much further investigation. Here we note that the actual mechanism generating the inertia of an object is still an open issue, with several different contending theories.[218] Whether the gravitational mass and the inertial mass are identical or not, also remains an issue. If the exact local mechanism generating the inertia of a mass could be decisively found, then direct methods of changing the inertial mass and the gravitational mass might be a real possibility, and research to try to find such methods would be strongly indicated.

Our further comment is that the mechanism that generates mass itself has never been specified in formal physics,[219] particularly when one views an observable (such as mass) as continually recurring in 3-space rather than continuously existing in 3-space, due to the *d/dt* operator imposed on spacetime by the observation process. Before the reader assumes that mass is well understood, a reading of Jammer's scholarly work {523} is highly recommended.

From our viewpoint we stress that observed mass cannot continuously exist, but must be an effect and the instantaneous frozen output of the *d/dt* observation process invoked on an ongoing 4-space process. Hence, from this view there is indeed a "process" that (continually) creates mass — it is at least the emission of EM angular momentum (a photon) from the *masstime* entity. Similarly, there is a "process" that continually destroys or converts any mass — it is the absorption of a photon and conversion of the mass m into masstime $(m + dm)dt$. Mass "flows through time" via this time charging and discharging, in the form $m \Rightarrow mt \Rightarrow m \Rightarrow mt \Rightarrow m$... etc.

---

[218] E.g., (a) A. Rueda and B. Haisch, "Inertia as reaction of the vacuum to accelerated motion," *Phys. Lett. A*, Vol. 240, 1998, p. 115-126; (b) B. Haisch, A. Rueda and Y. Dobyns, "Inertial Mass and the Quantum Vacuum Fields," *Ann. Phys.*, Vol. 10, 2001, p. 393-414.

[219] The present author made an early **and** crude attempt at doing it in his "Quiton-Perceptron Physics." 1973, NTIS Report AD-763-210.

Again we stress that masstime mt is not mass m, but is as different from it as impulse Ft is from force F. We crudely proposed such an initial mass-generating mechanism in 1973 {524a}, with additional information published in 1998 {525a}, which may have bearing on the mechanism involved in Woodward's experiments.

8.3.6 E. R. Laithwaite

Eric Laithwaite was for some years a professor with the Imperial College in London. He was one of the pioneers of the linear electric motor and of portions of the magnetic levitation (MagLev) train concept. An illustrious career headed for greater things (perhaps knighthood, Fellow of the Royal Society, etc.) was abruptly derailed in 1973 after he lectured to the prestigious Royal Institution by invitation and demonstrated that a heavy gyroscope, difficult to lift with one hand while not turning, could be lifted easily with that same hand, once turning at speed. Laithwaite's suggestion that Newton's laws of motion might be in difficulty proved much too controversial. The Royal Institution (apparently for the first time in its 200 year history) did not publish proceedings of that invited discourse.[220]

One of the remarkable things Laithwaite achieved was a form of "indefinite motion" device that he demonstrated circa 1970 {526}. The Centennial Centre of Science & Technology in Ontario was looking for working models to decorate the entrance hall. Laithwaite showed that if a linear-motor primary was rolled into a cylinder, so that it became simply the stator of a conventional induction motor,[221] a steel washer (about 3 cm. in diameter) could be made to roll continuously in a vertical plane around the inside of the stator. A combination of centrifugal force and magnetic attraction kept the washer in contact with the stator at all times. At the same time, a pivoted copper cylinder could be held in the center of the stator and would be seen to rotate on its own axis in the opposite direction.[222] Laithwaite completed and delivered a working model to the Centennial Center of Science & Technology. The reader should note the similarity of Laithwaite's proven experimental device to the magnetic Wankel engine.

---

[220] With some humor we point out how shocking is our proposal that such a thing as "continuous" propagation of an object or energy through 3-space does not exist. Instead, small differential "pieces" of it continually "recur" at successive positions, much like the projection of the frames of a motion picture frame, where continual observation is the "projection process". Nonetheless, in our proposed solution to the source charge problem, either that is true or every charge in the universe is a perpetual motion machine, freely creating energy out of nothing. In our view, that is the impact of the long-neglected source charge problem in quantal and classical electrodynamics. It has produced the wrong and naive notion of what "propagation through space" actually entails.

[221] Something similar to diagrammatic Figure 6-26 in Chapter 6.

In addition to his attention to Heaviside's unpublished draft notes and manuscript on a combined theory of electromagnetics and gravity (previously discussed above) {502}, Laithwaite spent the latter years of his life working on sophisticated gyroscopic systems {527a-527h}. Ultimately he felt he had successfully produced a mass transfer effect {528} and the basis for a propulsion system. He and William Dawson applied for a patent in 1993, which was granted in 1995 {529}, with a U.S. patent granted in 1999 {530}. Regretfully, Professor Laithwaite died in 1997.

8.3.7 Floyd Sweet's Vacuum Triode Amplifier (VTA)

See again Chapter 6 and Figures 6-5 through 6-14 on the work of Floyd Sweet, previously discussed in paragraph 6.3.2. My concept of antigravitation used in the experiment is shown in Figure 6-15. By now, the sharp-eyed reader will recognize that my method actually constitutes a conceptual extension and definitization of the Dirac electron theory, reinterpreted to recover negative mass, negative energy, negative energy fields, and negative energy potentials. The defining feature is the deliberate use of a $COP \gg 1.0$ main power system (perhaps constituting the power system of a flight vehicle) as a source of locally strong negative energy EM fields generated by the concentrated Dirac sea holes (causal 4-positrons being negative energy 4-electrons prior to observation) locally produced in the power system. We collaborated with Sweet on the vacuum triode amplifier for more than a decade, and in fact suggested the name of the device.

The antigravity approach designed by the present author was based on realizing that negative energy sources (e.g., 4-positrons as negative mass 4-electrons prior to their reaction and observation) produce *negative*

---

[222] We have had the distinct notion that the fundamental *contrarotating* mechanism found by Laithwaite is probably an analogous mechanism — or very similar — to the mechanism utilized by Bauman in his Swiss converter (Methernitha device). As pointed out, it also has remarkable similarities to the magnetic Wankel engine, except that the magnetic Wankel does not utilize two contrarotations.

energy EM fields. The field can reach out over all space, so a negative EM energy field represents a reversed curvature of a large volume of spacetime. Hence it represents a considerable antigravity field — particularly when one considers the unaccounted huge Heaviside component of the negative energy field and charge interaction. Since the field strength falls off as the inverse square of the distance from its source, the strong antigravity effects of negative energy EM fields are much stronger locally than at a distance. This is particularly true when the conventionally ignored Heaviside nondiverged EM energy flow accompanying the accounted diverged Poynting energy flow of every field-charge reaction is accounted. This usually ignored factor dramatically increases the antigravity intensity produced by the local negative EM energy field by many orders of magnitude, often by nearly 10^{13}.

My antigravity concept (then and now) is shown in Figure 6-15 of Chapter 6. At the time, the only available device I knew of that produced a substantial percentage of output negative EM energy was Sweet's vacuum triode amplifier. So I persuaded Sweet to fabricate a new output section (an output load similar to Figure 6.8 but containing ten sockets for ten 100 watt lamps). Thus the output of the unit — which adjusted automatically to the available load — could be increased in 100-watt increments by merely screwing in the lamps one at a time and making laboratory readings after each new 100-watt load increase was added. This was done and the concept was successfully tested by Sweet {531}, who was making the measurements in California and reading them off over the phone to me (I was in Huntsville, Alabama). The experiment worked beautifully.

The setup and progress of the experiment were intended to materially increase the strength of the *negative energy* EM field and Dirac sea hole flux already produced to a lesser extent by the Sweet vacuum triode amplifier (VTA) in its normal operation. Thus the hypothesized negative energy antigravity field, produced weakly by the negative energy from the device in its ordinary power output range of 500 watts, was to be substantially increased at 1,000 watts. The response to the net gravitational field surrounding the object would thus be the resultant of two factors: (i) the normal response of the device's positive mass to the positive gravitational field of the Earth etc., and (ii) the additional response of the device's positive mass to the local antigravitational field produced around it. The measurable result was expected to be a decrease in the measured weight of the object on the laboratory bench.

As discussed, this stratagem succeeded dramatically. The weight of the vacuum triode amplifier was decreased, smoothly and controllably, by

90% on the laboratory bench (see again Figure 6-9).[223] The results were later published in a joint paper {531} prepared by the present author. I placed Sweet's name first on the paper because he was the inventor of the VTA, which demonstrated the effect once properly modified as I requested. Note that the extra antigravitational field effect on an object is indeed a local effect, which the experiment also showed. This effect was obtained at a COP>1,500,000 (actually, at nearly a COP = 3,000,000) and at ELF frequency.

The ELF frequency is important. As we explained previously, the highest energy photons of all are the lowest frequency photons, when the highly compressed time-energy component of the photon is accounted as well as the decompressed spatial energy component. Hence, for very appreciable local curvature of spacetime to produce considerable antigravity, the antigravity source must produce negative energy fields consisting of negative energy photons at low frequencies (such as 60 Hz). Else, the curvature of spacetime is so small that little effect will evidence. My best "back of the envelope" estimate of the efficiency of the antigravity process in the VTA was 10%.

8.3.8 Ning Li and (separately) E. Podkletnov

Very interesting work was performed by Ning Li {532a-532d}, a scientist originally with the University of Alabama in Huntsville (UAH), who returned to China for a while but is now back in Huntsville. Quoting David Brown from the UAH website from 1997:

> *"In 1989, Dr. Ning Li of UAH predicted that if a time varying magnetic field were applied to a superconductor, charged and deformed lattice ions within the conductor could absorb enormous amounts of energy via the*

---

[223] Many times I have been asked how such small output power increase (from 500 to 1,000 watts) could possibly achieve such dramatic antigravity results. The answer - due to the neglected Heaviside nondiverged energy flow component — lies in multiplying the power output by the "Heaviside energy amplification factor *k* " where *k* may be as large as 10^{13}. A 500-watt accounted increase, multiplied by 10^{13}. is actually an incredible increase in the total EM negative energy field density being created in the local vacuum. In short, a very little "accounted" power increase of the COP»1.0 VTA output is actually a very vast negative power output increase when the long neglected k-factor is accounted. It is just this missing k-factor and the failure to account for the negative energy KM fields generated in COP»1.0 discharges in violent astronomical phenomena that seem to be responsible for generating the excess antigravity that is accelerating the expansion of the universe.

> *magnetic moment effect. This acquired energy would cause the lattice ions to spin rapidly about their equilibrium positions and create a miniscule gravitational field. Dr. Li's calculations showed that if these charged, rotating, lattice ions were aligned with each other by a strong magnetic field, the resulting change in local gravity would be measurable. "*
>
> *"Podkletnov and Nieminen (1992)* {533} *made the accidental discovery that a single-phase, dense, bulk, high* T_c*, superconducting, ceramic disk spinning at 5,000 rpm can produce a 2 percent reduction in the weight of non-conducting, non-magnetic objects placed over the spinning disk. UAH and MSFC [Marshall Space Flight Center, a part of NASA] [cooperated] on a joint research project to independently confirm the results of the Podkletnov experiment and to validate Dr. Li's theory of gravity modification via superconductor. On March 26th, 1997, as a part of this project, the joint UAH-MSFC research team produced the largest high temperature superconducting disk ever manufactured in the USA. This disk measures 12 inches in diameter and is 0.5 inches thick "*

We now explain the experiments, from our point of view of the mechanism for antigravity. We remark that, at least in theory, the charges in such a rotating superconducting disk have an inward "discharge" acceleration force that is a function of the distance of the charge from the rotation axis center. Hence, in the rotating frame, some of these charges are very strongly "accelerating" toward the center, as are their spins whose magnetic vector is orthogonal to the E-field vector. A flow of EM energy exists perpendicular to the rotating flat disk and is accelerating at right angles to the disk surface. This is analogous to a *discharge* (sudden acceleration) phenomenon, and it will also produce accelerating negative energy Dirac sea 4-holes in the vacuum component of the supersystem of the spinning disk. So from the spinning disk there is a sort of "wide beam" of Dirac sea 4-hole current in the vacuum, radiating at right angles to the disk surface, with heavier concentrations toward the outside of the disk. The result is that objects in the Earth's gravitational field, and also placed in the path of this negative energy Dirac sea 4-hole current, is acted upon by two superposed curvatures of local spacetime: (1) the positive curvature due to the Earth's gravitational field, and (2) the negative curvature due to the negative energy field's curving local spacetime. The total net gravitational effect will be due to the algebraic sum of the superposed two spacetime curvatures — the normal positive curvature from the Earth's gravitational field and the second a negative curvature created by the Dirac sea hole current's negative energy fields (Heaviside energy flow component considered). The expectation is therefore a *reduction* in the weight of the object, as is shown in the experiments.

The UAH-MSFC group was unable to confirm Podkletnov's findings.[224] Ning Li and the NASA group later split amicably. Ning Li continued her theoretical work, eventually ceased publishing papers for a period to protect the techniques and experimental results achieved in her work, and focused upon potential applications. She apparently returned to China for some time, but is now back in Huntsville, Alabama and still experimenting. We suspect there may also be a Chinese research effort into antigravity, using the approach taken by Ning Li. Perhaps information from that program will be released in the future.

But what of the work of Podkletnov since his original experiments? NASA continued to try to replicate his disk and his work, with Podkletnov acting as a consultant. Presently, NASA has paid an independent research firm to replicate the disk to full specifications, which has just been accomplished as this is written. At this writing, the experimental results of this new attempt are not yet available as the work is still in progress. It should be available by the time this book is published.

Meanwhile, a relatively new paper by Podkletnov and Modanese has been published on the Los Alamos National Laboratory website.[225] This method uses high amperage sharp pulse discharges from a superconducting ceramic electrode, to produce an anomalous radiation in a focused beam. Significantly, the beam does not noticeably attenuate through different materials. The beam exerts a short repulsive force on small movable objects on the propagation axis, with the force proportional to the mass of the objects.

---

[224] E. Podkletnov and R. Nieminen, "A possibility of gravitational force shielding by bulk $YBa_2Cu_30_{7-x}$ superconductor," *Physica C,* Vol. 203, 1992, p. 441-444.

[225] Evgeny Podkletnov and Giovanni Modanese, "Impulse Gravity Generator Based on Charged $YBa_2Cu_10_{7-y}$ Superconductor with Composite Crystal Structure," carried on website http://xxx.lanl.gov/abs/physics/0108005.

We interpret the results as completely consistent with our own proposed excess gravity and antigravity mechanism. We know that such sharp discharges produce significant Dirac sea 4-hole currents in the local vacuum component of the supersystem. Such currents in a beam interact very little with intervening materials, and the major spacetime curvature effects are concentrated upon the propagation axis. On that axis, force effects can be expected upon small independent objects, due to the *additional* negative curvature added to the local spacetime in their supersystem. The impulse produced on the object should indeed be proportional to the mass of the object and independent of its composition, since it is a gravitational force effect induced by local curved spacetime. The force should be repulsive in the beam, since that is the direction of the antigravity force that will be induced in objects in the axis of the beam, by the beam of negative energy Dirac 4-holes. A positive gravity force would mean attraction between the object in the propagation path and the source of the beam. A negative gravity force would therefore mean repulsion.

So Podkletnov and other scientists are still working on a successful effect. It will be interesting to see what results are reported in the future, and whether they will deviate from their present attempt to explain the effect by a combined quantum gravity and anomalous vacuum fluctuations approach.

8.3.9 NASA's **Breakthrough Propulsion** Physics **(BPP) Project**

NASA's Breakthrough Propulsion Physics Project was established in 1996. It seeks a breakthrough in space transportation involving "(i) propulsion that requires no propellant mass, (ii) propulsion that attains the maximum transit speeds physically possible, and (iii) breakthrough methods of energy production to power such devices." Investigations involve coupling of gravity and electromagnetism, the quantum vacuum, hyperfast travel, and superluminal quantum effects. The project is managed by the Glenn Research Center, sponsored by the Advanced Space Transportation Program, with overall management by NASA Marshall Space Flight Center, Huntsville, Alabama.

NASA continues its investigations and experiments in methods to change **the** force of gravitational attraction. The scientist in charge of the project is David Noever of the Marshall Space Flight Center. The work is still considered a work in progress, with the need for both experimental investigation and meticulous design of experiments, instrumentation, and measurement techniques.

The overall NASA program also spreads to other supporting laboratories and projects. For example, at the Jet Propulsion Laboratory (JPL) Kulikov {534} has measured the difference between inertial and gravitational mass of a boson in a heat bath. At least one invention seems to have resulted or to be in progress, assigned to JPL.

As mentioned, the U.S. Patent by Campbell {518} on a capacitor device producing a net thrust force (a more modern version of T. Townsend Brown's work) has been issued and assigned to NASA.

8.3.10 Scientific Conferences

There are regular conferences (several per year) being held in the U.S. and elsewhere, dealing with antigravity and electrogravitic propulsion. We give a little-noticed but pertinent example: In a nice presentation in 2001, LaPointe {535} proposed a new method for producing anti-matter. Currently antimatter production is accomplished in very large accelerators, and is therefore quite expensive. For increased experiments, a cheaper and much smaller (and much more available) process is highly desirable.

In the Klein paradox, which involves very large potential drop across a very short distance — less than the Compton wavelength of the antimatter particle — matter and anti-matter pairs are produced. LaPointe hypothesized that it may be possible to use the Casimir effect to push one side of the step into the negative energy range. He is at least *thinking toward* negative EM energy sources that produce negative EM fields and negative gravity in the surrounding space, without explicitly enunciating it in that fashion.

We already briefly discussed our own approach to antigravity, which is a much simpler and cheaper anti-particle producing mechanism, as demonstrated by the Sweet device. In short, overunity EM systems with COP»1.0 do produce free flows of causal (before observed) 4-positrons (still in negative energy 4-electron state prior to interaction) in local spacetime immediately surrounding the system. Consequently, LaPointe's efforts are in what we believe to be the correct direction, but he is still using a very difficult mechanism to engender the antimatter production. Also, the positrons or other antimatter sources of the negative energy fields must remain in the causal (unobserved and non-interacted) 4-state, in order to produce negative energy EM fields and potentials and antigravity.

8.3.11 James Corum *et al.:* Rectified Slepian AC Force Resonator

Another example is some promising work by Corum *et al.* {536}. We quote from the abstract and summary of his paper:

> *"Corum presented an experimental paper on the use of the Heaviside force in conjunction with a Slepian antenna as a form of space drive using nothing more than the classical Maxwell stress tensor. Slepian proposed the same thing in 1949, but came to the conclusion that it would not be useful, since the time average of the resultant AC force would be zero. Corum's contribution has thus far been two-fold: (1) In conjunction with Dr. Alan Barnes of WVU he has experimentally shown that the AC version of the Slepian Resonating Antenna does produce a force, and has measured it to within 3.6%, and (2) has designed a way for Hartley's variable capacitor rectification to be used with the Slepian resonator such that the rectification results in a DC force component. The first experiment has already been achieved, the second experiment is the logical next step. If successful, the result would be quite revolutionary: a true space drive."*

Our comment is that this unilateral thrust force work is important. In our view, the Heaviside energy flow component does not appear in Slepian's vector $\mathbf{j}\phi$, which contains only the diverged EM energy flow surrounding a circuit or component, but it is present immediately outside the circuit where the Slepian vector exists. Hence there is an enormous, extra energy flow closely associated with every EM circuit, that is unaccounted since Lorentz's closed surface integration of the energy flow vector discarded it quite arbitrarily.

We already know from the Bohren experiment (and thousands of others similar to it) previously cited that resonant charges do collect excess EM energy from a given EM energy flow input, compared to the collection from that flow by the same charges in nonresonant state. In using the Slepian resonator and then rectifying its output, it would appear that both the diverged (Poynting) energy flow component and the Heaviside nondiverged energy flow component are first resonated and then rectified. In addition, excess energy is collected — much more than is in the normal "static charge" formula assumptions for Poynting energy flow and in assuming only a "single pass, single interception and collection" of the energy.

An important factor then is obvious: Once the energy flow is surrounding the circuit, there exists a certain local curvature of the immediate external spacetime. The rectification represents not only a change in the Slepian vector current in the circuit, but also a change in this local ST curvature interaction upon the circuit matter and charges. In short, it should produce a DC component in the direct excess interaction of the rectified local curvature of spacetime. This should produce *excess direct propulsive force,* just as Corum theorizes and the experiments support to date.

Further, the oscillation of the large, *unaccounted* Heaviside nondiverged energy flow component involves the oscillation of a "dark energy" component that may be orders of magnitude greater than the *accounted* Poynting component's magnitude. By involving this huge energy flow component in oscillation and synchronized to the resonant Slepian circuit, Corum has unwittingly involved the supersystem as a COP»1.0 EM "energy converter" system. Hence, there is appreciable Dirac hole current (in causal 4-form) in the local vacuum, not reacted with the materials in the system and therefore not converted to material lattice holes attached to positive mass. The time rate of changes of the negative energy fields and potentials are thus significant for antigravity effects, in the approach taken by the present author. In short, we believe Corum *et al.* are in fact producing significant evidence for a workable approach to unilateral thrust force.

This work by Corum *et al.* thus is highly promising, is solidly based experimentally, and certainly bears watching closely. We would strongly suggest that Corum *et al,* analyze their proposed mechanisms and results obtained in terms of supersystem analysis rather than system analysis, using higher group symmetry electrodynamics, and recover and include the long-neglected Heaviside energy flow component.

8.3.12 Gravitec, Inc. and H. Serrano

A patent {537} was issued to H. Serrano on Oct. 5, 2000 and assigned to Gravitec, Inc. whose principal investigator and corporate officer is Tim Ventura. Based on this invention, Gravitec, Inc.. has developed and demonstrated a rotary device that develops unilateral thrust. Purdue University began a small but steady effort to test the unit, first in the atmosphere and later under high vacuum conditions. The first tests have been successful, and a paper {538} from the Purdue study shows it is not the ion wind effect, which is orders of magnitude too small to account for the demonstrated thrust. Further, Jean-Louis Naudin has replicated and tested the Serrano thruster, and it works beautifully {539}. Ventura also hopes to develop a toy kit, with safe power supply, which would allow persons to perform their own successful lifter experiments and perhaps rotary experiments at nominal cost.

8.3.13 Transdimensional Technologies, Inc.

There is also ongoing important propulsion work at Transdimensional Technologies, Inc. of Huntsville, Alabama {540}. Founded in 2000 by its chief scientist, Jeffrey Cameron, the company has developed several novel propulsion and lifting mechanisms. Independent tests of some of these systems have been accomplished by Jean-Louis Naudin, and the results can be seen on his website (previously cited).

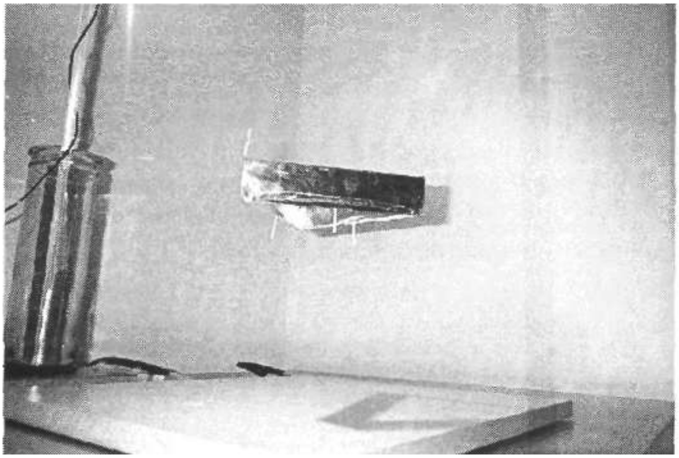

Figure 8-1 Naudin's **replica of the T**ransdimensional Technologies **lifter.**

As we go to press, Transdimensional's most recent propulsion effect is called Power3 (power "cubed"). It uses an asymmetric capacitor to produce the thrust in space, and the principle has been successfully tested in a vacuum. A larger, triangular lifter made mostly of pipe was demonstrated in vacuo to NASA officials in Huntsville, Alabama. Cameron has patents pending on magnetocaloric hybrid rocket engines and asymmetric gravitational waves {541}. A replica of the Transdimensional lifter successfully tested by Naudin is shown in Figure 8-1, by courtesy of Jean-Louis Naudin. Naudin's replica of the Transdimensional Technologies rotational device is shown in Figure 8-2.

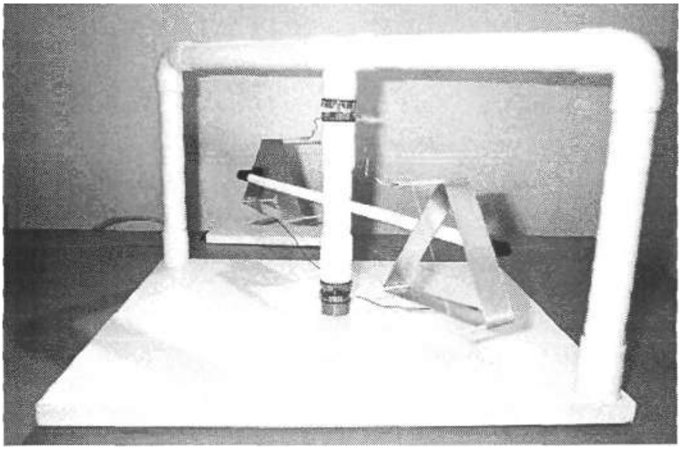

Figure 8-2 Naudin's replica of the Transdimensional Technologies rotational device.

We understand that new Transdimensional Technologies systems, now in patent process and still closely held, represent dramatic advances in the state of the art. The corporation foresees eventual commercial licensing of some of its developments, and the development of practical lift vehicles and space propulsion systems within 3 to 5 years.[226]

---

[226] Ken Moore and I personally visited Transdimensional Technologies here in Huntsville, and we found a very straightforward, enthusiastic, and hard-working group. We were treated with great courtesy, so much so that Jeff Cameron graciously cut short his vacation day and came in to show us the lab equipment and give a demonstration. We very much appreciate the work that Transdimensional Technologies is doing, and we certainly wish this fine group outstanding success in their efforts. We also are very appreciative for their courtesy and consideration in showing us the demonstrations and for their cogent discussions with us, which **were** most helpful.

8.3.14 American Antigravity (Tim Ventura's Company)

Tim Ventura is a very energetic lifter and antigravity researcher, who has done work independently and for others such as Transdimensional Technologies. The results of his many experiments and replications, tests, photos, etc. are on his website at http://www.americanantigravity.com. He also has several videos showing his testing, as well as detailed instructions for a researcher to build his own demonstration unit and verify the tests. Tim kindly furnished some needed information in this area, and put me in touch with other researchers and websites, which I appreciate.

8.3.15 Jean-Louis Naudin

Jean-Louis Naudin is a tireless researcher in the fields of COP>1.0 systems and processes, and in the field of lifters, inertial thrusters, and possible antigravity systems. His website is a cornucopia of information, patent references, inventions, processes, and replications or replication attempts by Jean-Louis and others who regularly work on such things and communicate with him. The website is http://jnaudin.free.fr/. From there, one may click on the various parts of the site and on the links to other sites to gain a wealth of information on these subjects.

His lifter work, including successful replications and instructions on how to build lifters, can be accessed from http://jnaudin.free.fr/html/lifters.htm, and will keep the interested researcher busy for many productive hours. We especially commend Jean-Louis for his extraordinarily important efforts and for making all this work available to researchers worldwide. Once on the site, one is free to choose from a very wide variety of projects, subjects, etc. The information value of this site is very, very high. We consider Naudin's site to be the single best information site on the web for COP>1.0 system information and for the new space propulsion methodology and effort. Necessary links to other related sites of interest are also provided.

Good hunting!

8.3.16 Other Research Work and Inventions

While we have given a sort of quick overview of work done in this field, there have been many other patents filed by numerous inventors and additional work done by them. This is a dynamic field and it is moving; the biggest problem is that it is still sorely hampered by highly limited funding and an attitude by much of the scientific community that "if it has not been invented here, it doesn't exist".

In the table in Figure 8-3 below, we list some of the pertinent patent numbers for reference, with the emphasis on U.S. Patents. The Naudin website previously referenced has a great deal of information on these thrusters and lifters, including instructions for building and testing several of the simpler proof-of-principle versions. There are links to other sites with even more information. Naudin also includes the results of his successful replication and testing of a variety of these devices and inventions, as well as the successful replication by other experimenters. Naudin also gives construction instructions and information on some of these, so that independent researchers can build test units and test them if desired.

For those researchers interested in pursuing this area further, we would heartily recommend a determined search on the Internet, using Google (www.google.com) as their search engine. A good place to start is Jean-Louis Naudin's web site, and then run the cross-links. The combination will turn up quite an amount of specific material, from which the individual researcher can choose which approach and experimental method he wishes to pursue.

Figure 8-3.

Selected Inertial Propulsion System Patents

T. Townsend Brown, "A Method of and an Apparatus or Machine for producing Force or Motion, British Patent 300,311, Nov. 15, 1928.

Norman L. Dean, "System for Converting Rotary Motion into Unidirectional Motion," USP #2,886,976, May 19, 1959.

T. Townsend Brown, "Electrokinetic Apparatus," USP #2,949,550, Aug. 16, 1960.

T. Townsend Brown, "Electrokinetic Transducer," USP #3,018,394, Jan. 23, 1962.

T. Townsend Brown, "Electrokinetic Generator," USP #3,022,430, Feb. 20, 1962.

T. Townsend Brown, "Electrokinetic Apparatus," USP #3,187,206, Jun. 1,1965.

Henry Wallace, "Method and Apparatus for Generating a Secondary Gravitational Force Field," USP #3,626,605, Dec. 14, 1971.

Henry Wallace, "Method and Apparatus for Generating a Dynamic Force," USP #3,626,606, Dec. 14, 1971.

Fernand Estrade, "Device for Transforming Kinetic Energy," USP #3,807,244, Apr. 30, 1974.

Leo J. Novak, "Centrifugal Mechanical Device," USP #3,810,394, May 14, 1974.

Everett H. Benson, "Inertia Engine," USP #3,863,510, Feb. 4, 1975.

Oscar Mast, "Propulsion System," USP #3,889,543, Jun. 17, 1975.

Arthur N. Lehberger, "Centrifugal Propulsion Drive and Steering Mechanism," USP #3,897,692, Aug. 5, 1975.

Howard A. Gaberson, "Vibratory Locomotion Means," USP #3,916,704, Nov. 4, 1975.

Calvin I. Cuff, "Device for Converting Rotary Motion Into a Unidirectional Linear Motion," USP #3,968,700, Jul. 13, 1976.

Nicholas Joseph Schnur, "Method and Apparatus for Propelling an Object by an Unbalanced Centrifugal Force with Continuous Motion," USP #3,979,961, Sep. 14, 1976.

Calvin I. Cuff, "Device for Converting Rotary Motion Into a Unidirectional Linear Motion," USP #3,998,107, Dec. 21, 1976.

George Knap, "Orbital Propulsion Apparatus," USP #4,087,064, May 2, 1978.

Calvin I. Cuff, "Device for Converting Rotary Motion into Unidirectional Motion," USP #4,095,460, Jun. 20, 1978.

Robert L. Cook, "Device for Conversion of Centrifugal Force to Linear Force and Motion," USP #4,238,968, Dec. 16, 1980.

Ladislaw G. Srogi, "Mechanical Propulsion System," USP #4,242,918, Jan. 6, 1981.

Harry S. Melnick, "Unidirectional Force Generator," USP #4,261,212, Apr. 14, 1981.

Frederick L. Dehen, "Apparatus for Converting Rotary Motion to a Rectilinear Force," USP #4,347,752, Sep. 7, 1982.

Harry S. Melnick, "Mechanical Power Transmitting System," USP #4,398,431, Aug. 16, 1983.

Peter Kleber, "Apparatus for Acceleration-Free Mounting of a Body in a Spacecraft," USP #4,408,740, Oct. 11, 1983.

Fred de Weaver, III, "Propulsion System," USP #4,409,856, Oct. 18, 1983.

Ronald K. Linde, "Centrifugal Reaction Chamber," USP #4,450,141, May 22, 1984.

Fred de Weaver, **III,** "Propulsion System," USP #4,479,396, Oct. 30, 1984.

Joseph Colla, "Mechanical Propulsion System," USP #4,577,520, Mar. 25, 1986.

Elmer M. Dobos, "Propulsion Apparatus," USP #4,579,011, Apr. 1, 1986.

Brandson R. Thomson, "Apparatus for Developing a Propulsion Force," USP #4,631,971, Dec. 30, 1986.

James E. Cox, "Dipolar Force Field Propulsion System," USP #4,663,932, May 12, 1987.

Alvin C. Peppiatt and Alfred J. Peppiatt, "Impulse Drive," USP #4,674,583, Jun. 23, 1987.

Henry North, "Apparatus for Producing a Force," USP #4,712,439, Dec. 15, 1987.

Harry S. Melnick, "Continuous Force and Impact Generator," USP #4,726,241, Feb. 23, 1988.

Oscar F. A. Peterson, "Apparatus for Producing a Directional Unit Force," USP #4,744,259, May 17, 1988.

James D. Mundo, "Universal Propulsion Powerplant and Impulse Drive Unit for Self-Propelled Vehicles," USP #4,770,063, Sep. 13, 1988.

Lancelot I. Kethley, "Gyroscopic Propulsion Device," USP #4,784,006, Nov. 15, **1988.**

Charles Fulop, "Flywheel," USP #4,788,882, Dec. **6, 1988.**

Charles E. Rogers and Fred Van Arsdell, "System for Propulsion and Positioning of a Transitory Object," USP #4,801,111, Jan. 31, 1989.

Paul J. Montalbano, "Conversion of Rotational Output to Linear Force," USP #4,856,358, Aug. 15, 1989.

George J. Zachystal, "Device for Obtaining a Directional Centrifugal Force," USP #4,884,465, Dec. 5, 1989.

James E. Cox, "Dipole Accelerating Means and Method," USP #4,891,600, Jan. 2, 1990.

Hector Serrano, "Propulsion Device and Method Employing Electric Fields for Producing Thrust," WO 00/58623, Oct. 5, 2000.

Brian C. Motts, "Airship," USP #4,967,983, Nov. 6, 1990.

Lyle M. Mason, "Centripedal Device for Concentrating Centrifugal Force," USP #4,991,453, Feb. 12, 1991.

Alexander D. Kidd, "Gyroscopic Apparatus," USP #5,024,112, Jun. 18, 1991.

Paul J. Montalbano, "Conversion of Rotational Output to Linear Force — a Transmission," USP #4,042,313, Aug. 27, 1991.

Andrew T. Rodgers, "Controllable Gyroscopic Propulsion Apparatus," USP #5,054,331, Oct. 8, 1991.

Mortimer S. Delroy, "Gyrostat Propulsion System," USP #5,090,260, Feb. 25, 1992.

Kemal Butka, "Propulsion System," USP #5,111,087, May 5, 1992.

Rex L. Schlicher, Steven M. Rinaldi, David J. Hall, Peter M. Ranon, and Charles E. Davis, "Nonlinear Electromagnetic Propulsion System and Method," USP #5,142,861, Sep. 1, 1992.

Thomas L. Navarro and James D. Isaacson, "Apparatus with Rotatably-Driven Asymmetrically-Accelerated Eccentric Mass for Generating Translational Force," USP #5,150,626, Sep. 29, 1992.

Theordore R. Bristow, Jr., "Method and Apparatus for Converting Rotary Motion to Lineal Motion," USP #5,156,058, Oct. 20, 1992.

John C. McMahon, "Energy Transfer Device," USP #5,167,163, Dec. 1, 1992.

James W. Black, "Non-Linear Propulsion and Energy Conversion System," USP #5,182,958, Feb. 2, 1993.

James R. Taylor, "Electromagnetic Energy Propulsion System," USP #5,197,279, Mar. 30, 1993.

Michael J. Sohnly, "Magnetohydrodynamic Propulsion System," USP #5,211,006, May 18, 1993.

Ross C. Wood, "Stabilization System for a Freely Rotatable Platform," USP #5,256,942, Oct. 26, 1993.

James F. Woodward, "Method for Transiently Altering the Mass of Objects to Facilitate Their Transport or Change Their Stationary Apparent Weights," USP #5,280,864, Jan. 25, 1994.

Robert C. Willis, "Spaceship Propulsion by Momentum Transfer," USP #5,305,974, Apr. 26, 1994.

Harold L. Hull and Donald E. Joslin, "Reusable Mass Propulsion System," USP #5,313,851, May 24, 1994.

Kemal Butka, "Propulsion System," USP #5,334,060, Aug. 2, 1994.

James Harvey, "Impulse Converter," USP #5,335,561, Aug. 9, 1994.

Maurice Mitchell, "Net Kinetic Energy Differential Guidance and Propulsion System for Satellites and Space Vehicles," USP #5,377,936, Jan. 3, 1995.

Howard M. Woltering, "Rotating Eccentric Weights Vibrator System," USP #5,388,469, Feb. 14, 1995.

Richard O. Marsh, Jr., "Centrifugal Force Drive Machine," USP #5,388,470, Feb. 14, 1995.

Kemal Butka, "Propulsion System," USP #5,410,198, Apr. 25, 1995.

Thomas L. Navarro, "System for Generating Controllable Reference Environment and Steerable Translational Force

from Interaction Therewith," USP #5,473,957, Dec. 12, 1995.

Richard L. Lieurance, "Centrifugal Inertia Drive," USP #5,488,877, Feb. 6, 1996.

John C. Claxton, "Centripetally Impelled Vehicle," USP #5,557,988, Sep. 24, 1996.

Ezra Shimshi, "Apparatus for Energy Transformation and Conservation," USP #5,673,872, Oct. 7, 1997.

Richard E. Foster, Sr., "Inertial Propulsion Plus/Device and Engine," USP #5,685,196, Nov. 11, 1997.

James D. Booden, "Electromagnetically Actuated Thrust Generator," USP #5,782,134, Jul. 21, 1998.

George J. Howard, "Propulsion System," USP #5,791,188, Aug. 11, 1998.

Timothy J. Stopplecamp, "Boot Strap Propulsion System," USP #5,831,354, Nov. 3, 1998.

Eric Laithwaite and William Dawson, "Propulsion System," USP #5,860,317, Jan. 19, 1999.

Jonathan W. Campbell, "Apparatus and Method for Generating Thrust Using a Two Dimensional, Asymmetrical Capacitor Module," U. S. Patent # 6,317,310 Bl, Nov. 13, 2001. Assigned to U.S. Government.

8.3.17 Mechanism for the Antigravity Accelerating the Universe's Expansion

Finally, in addition to sorely needed better means of space propulsion, one of the real incentives for antigravity research is provided by the search for a source of the now-established mysterious antigravity that is accelerating the expansion of the universe. Elsewhere in this book we have advanced our own proposal for the mechanism generating that *antigravity* effect in the cosmos.

Simply put, a very violent astronomical process may be considered a giant and very violent discharge phenomena, since it first involves a giant excitation followed by a giant release (discharge) of that excitation. All such phenomena involve strong broken symmetry during the discharge itself, with the concomitant strong available of causal 4-sources of negative energy and negative energy fields. These processes — in their causal unobserved state - do produce locally strong negative energy EM fields from the source causal 4-positrons (Dirac sea holes prior to observation as positive mass 3-space positrons having positive energy and reverse-oriented). When the unaccounted and very large Heaviside nondiverged EM energy flow component accompanying every field and particle interaction is accounted, the causal negative "mass-energy" states produce strong antigravity fields, particularly considering the *lower frequencies* of many of the phenomena.

With the long-neglcctcd Heaviside nondiverged energy component considered and not just the accounted Poynting diverged component, the total amount of negative EM field energy available in these violent process is dramatic. Accordingly, just as was the Sweet VTA in its increased power and its ELF pulsations, these giant astronomical sources are also giant negative "mass-energy" sources and strong antigravity sources as well. In short, such astronomical sources produce strong antigravity components repelling each other across the universe. Hence the observed acceleration of the expanding universe; the violent astronomical phenomena are continually "pulsing" the universe with antigravity and hence continually accelerating its expansion velocity.

In 2000, we also published a proposed mechanism for the source of the excess *positive* gravity holding the arms of the spiral galaxies intact {542}.

8.4 The Author's Engineering Approach to Local Antigravity

See again our discussion under Newton, subparagraph 8.2.1 above, and our discussion of a negative field energy extension to it. For a major laboratory, we believe that antigravity is straightforward,[227] if one first produces a COP»1.0 situation in a higher group symmetry electrodynamics process. This provides a very strong "displacement current" of causal 4-positrons (before parity inversion, time-reversal, etc.) as true negative energy "particles" or "states" or "Dirac holes" in the surrounding 4-vacuum. The actual "negative mass-energy field" is very large, due to the unaccounted Heaviside nondiverged negative energy component. Consequently, a substantial antigravity field is established.

---

[227] However, this is based on the assumption that the lab team first gives up some erroneous but widely held misconceptions, and really does a little "thinking outside the box" as they are prone to suggest.

8.4.1 Sweet Antigravity Test.

See again our discussion in paragraph 8.3.7. A practical realization of antigravity was clearly shown by the Sweet antigravity test designed by the present author. See previous discussion under paragraphs 6.3.2 in Chapter 6, and see again Figures 6-5 through 6-15.

In Chapter 6, the Sweet vacuum triode amplifier device shown in Figure 6-8 had a very large COP = 1,500,000 at 500 watts output with a 33 microamp, 10 volt operator's input. It had an even greater COP when pushed to 1,000 watts output because of only a slightly greater input.

The Sweet VTA used an induced powerful self-oscillation between the local vacuum and the barium nuclei in the barium ferrite magnets. This resulted in what is called a "kinetic" magnet (and one magnetized through its flat face). In a kinetic magnet, its nuclear magnetic field (in this case, in the barium nucleus) is self-oscillating (waving) in an angle about its local direction, as shown in Figure 6-6 in Chapter 6.

The conditioning was performed with the barium ferrite magnet itself in forced resonance at its resonant condition (some 11 to 12 Hz, in the case of Sweet's magnets). Then the 60 Hz signal was impressed orthogonally to the primary direction of the resonant magnet. After conditioning each of the two magnets, they were placed together on a separation box, carefully keeping the corresponding ends together similar to their original orientation during activation, but in attracting mode (through their face) across the separator. The assembly itself was then further conditioned, again orthogonally to the coupled oscillating fields. This "synchronized" the two magnetic self-oscillations and phase-locked them together in a common self-oscillation with the local vacuum.

Similar treatment of the vacuum as a semiconductor, together with self-oscillation between vacuum and nuclei, has been advanced independently by Prange and Strance {543}, who showed that the vacuum may be regarded as a semiconductor. In particular, they showed that the vacuum in the region close to the nucleus of a superheavy element is analogous to the inversion layer in a field effect transistor. It is similarly analogous for the barium atom, though Prange and Strance do not discuss that element. They introduced the concept of the inverted vacuum. Just as a semiconductor may be manipulated by subjecting it to external fields, doping etc., it appears that so can be the vacuum — along the lines shown by Prange and Strance and along the lines demonstrated by Sweet's conditioning of his kinetic magnets.

8.4.2 Bearden's Proposed Antigravity Mechanism

Any EM process that produces a COP»1.0 condition, such as can be done in intensely abrupt ionic discharges and other processes, will also produce intense causal (4-space) Dirac sea holes in the immediate surrounding vacuum. This is the negative energy mechanism, conceived by the author in the 1970s and finalized in the 1980s, for producing a negative energy causal (unobserved) EM field[228] in the immediate space surrounding a system. The overunity system must be the "source" of that negative energy EM field. So the negative energy EM field in surrounding local space is "attached" to the mass and energy dynamics of the "unit's supersystem" as being its "source or generatrix". Its attachment to the system mass produces the antigravitational force upon that mass.

Negative energy state charges (Dirac holes) are also *source charges.* Prior to observation converting them (as observed) to 3-space positive mass-energy positrons with positive energy and going the opposite direction, they must be accepted as negative energy source 4-electrons "pouring out negative energy" in all directions in 3-space, continuously, via the giant negentropy process for source charges. Since they pour out "negative" energy, we may consider that negative energy field action at any point in 3-space to be a continuous absorption of positive energy from 3-space and a return of that extracted energy to the time domain as increased time — in short, time dilation.

Either way one considers it, the 4-space Dirac holes do produce negative energy EM fields, potentials, and spacetime curvatures and these must be accounted during their existence *prior to observation and interaction.* Further, their accompanying giant Heaviside negative energy flow component must also be accounted. In other words, do not reverse the spatial direction and thus the location and direction of these fields, do not turn them into positive energy fields by such reversal of direction, do not turn the negative mass into positive mass, and do not reverse the sign of the source charge (the Dirac sea hole). All that happens after interaction of the Dirac sea holes with matter, forming the usual lattice holes or *observed* 3-positrons.

---

[228] We strongly stress that the continuous, dynamic, ongoing causal 4-field before it interacts with charge and is observed, is a totally different entity from the iteratively and continually observed frozen "effect" 3-field in matter after the 4-field has interacted with charged matter to produce an observation or "event". The great barrier to using electromagnetic fields in practical antigravity application has been due to the substituting, in electrodynamics, of the iterative "effect" 3-field of the observed positron for the continuous "causal" 4-field of the unobserved negative energy 4-electron (the Dirac sea hole). The "effect" 3-field is simply the iterative, continual energy gradient of the diverted EM field energy flow around the assumed intercepting unit charged particle.

Therein lies the magic secret of antigravity. Instead of thinking of the *fields* interacting on the craft or system that is producing them,[229] think of these negative energy fields interacting upon the *local spacetime surrounding the craft or system, and curving that spacetime negatively. Then* think of this negative curvature of spacetime interacting back upon the craft or system in general relativity fashion. *That* is what will and does produce practical antigravity upon the interacted craft or system, once the Heaviside energy flow component is accounted! But to understand its magnitude and the practicality, one must also account for the long neglected Heaviside nondiverged component.

The mechanism is also consistent with general relativity and Wheeler's succinct statement that mass acts on spacetime to curve it, and curved spacetime acts back on that mass to move it or produce a translation force upon it. We are simply applying that fact of general relativity to produce a local antigravity effect of a locally curved spacetime upon the source system (in this case, the power system itself, which is the source of the negative field energy in the vacuum.

The entire notion of "curvature in spacetime" — and indeed of spacetime itself— is a notion applying *before observation and interaction to produce the observed effect* — i.e., prior to the *d/dt* operation occurring upon an ongoing 4-space process to produce a momentary frozen 3-space snapshot (an observed effect). Think carefully. Wheeler's rule does not mean that 3-space acts back on the mass (actually on the masstime), but that 4-space (spacetime) acts back on it. Spacetime does not exist in the observed object after the application of observation's *d/dt* operation has occurred and until a new causal interaction begins. The *d/dt* operation strips away the time, and hence most of the energy (i.e., all the time-energy) of the spacetime that is to do the interacting.

---

[229] The ongoing interaction of the vacuum/spacetime with the source of the unobserved 4-fields is their cause, and the observed 3-fields are the effect. The effect *a priori* does not interact upon the cause! So thinking that "one should make a 3-field in a craft or system that will interact with it and produce antigravity force upon it" is a non sequitur. Instead, one more correctly thinks that "one should make a 4-field *from* a craft or system which will interact with the local spacetime to curve it negatively. Then that negative curvature of spacetime is itself a negative energy 4-field that will interact back upon the observed craft or observed 3-system to produce observed antigravity force upon it.

As source charges, the Dirac sea holes — before observation and in their vacuum state condition — produce a locally curved *spacetime* around the source unit. That induced *(negative energy x time)* curvature interacts back upon the physical system that is the source of the field producing the negative spacetime curvature. It specifically acts upon the mass of the 4-system when it exists as *(mass x time) form.* That interaction produces *antigravity* force and momentum upon the system, rather than *positive gravity* force and momentum[230]. Adding negative EM field energy (in the causal, unobserved state) is precisely the same as adding extra negative masstime, i.e., extra *(-m)t,* since the 4-field is being used in unobserved state.

Using the principle of superposition of fields[231], the external observer sees the "weight" of the unit (the force developed on it by the *net* gravitational field — see again our discussion of the Newtonian approach under paragraph 8.2.1 above) — as being reduced. The intensity of the *negative energy x time (negative angular momentum or negative action)* EM field (including with the Heaviside component accounted) produced by the unit determines the intensity of the antigravity produced, and hence determines the fraction of the unit's weight[232] that is observed to be "lost". See again Figures 6-14 and 6-15 in Chapter 6.

---

[230] Note that, if the Dirac hole interacts with the mass lattice of the system, it converts to a "hole" that is attached to the positive mass of an ion. Hence the "observed" positron is said to have "positive mass", which is a non sequitur for the state in which the "positron" exists prior to observation or interaction. The lattice hole 3-positron and the true spatiotemporal "4-positron as still a negative energy 4-electron'" are two very different entities, and their fields and concomitant field effects are very different.

[231] Note that identification of spacetime and field in general relativity extends the notion of superposition of multiple fields to the notion of superposition of multiple curvatures of spacetime. Thus, it is the *net* superposed curvature of spacetime that interacts back upon the physical mass system, to produce gravity (either positive or negative, depending on the sign of the net curvature).

[232] We remind the reader that the "weight" of an object on Earth is a measure of the force developed on the object by the (Earth's) gravitational field. "Loss of weight" is simply "reduction of that net force" due to the combination of the Earth's positive (attractive) gravitational field and the Dirac hole curvature of the object's local spacetime producing a negative (repulsion) gravitational field.

At sufficient intensity of this local negative energy 4-field in spacetime with the unit as its source, the back interaction upon the unit will balance the action of the earth's gravitational field. At that point, the unit will hover and "float" against the force of terrestrial gravity.[233] By further increasing the intensity of the negative energy EM field, the unit will accelerate upward, now being repelled by the local earth more than it is attracted by it. By "angling" the negative energy EM field direction — or by producing a second such negative energy EM field from the source unit as an "angled add on" superposed on the first one — one also provides unilateral propulsion in a given direction.

As can be seen, all the requirements for a practical antigravity vehicle and negative energy propulsion system are met. The only real problem in this approach is the requirement to first attain a $COP \gg 1.0$ EM power system at ELF frequencies.[234] There is, however, an additional problem of correcting one's thinking to be along the lines indicated.

Antigravity becomes straightforward if one first does overunity discharges or steady output in a system with $COP \gg 1.0$ and understands the difference between the observed "effect" 3-field and the unobserved "causal" 4-field. Sweet's VTA unit had a beginning COP = 1,500,000 at 500 watts output. The unit was pushed to nearly double its COP, by connecting additional load so that the output was now 1,000 watts. The additional input draw was minimal.

Designing the experiment with that in mind, I convinced Sweet to build a new output load unit, and to perform the experiment. In advance, I roughly predicted levitation of the 6-pound device at about 1500 watts output power, by some back-of-the-envelope estimates assuming a 10% efficient process. The projected curve (again see Figure 6-14 in Chapter 6) shows that levitation would have occurred at about 1250 watts or so.[235]

---

[233] Sweet later did push the VTA to hovering and upward acceleration, with the unit on a tether. After he was shot at by a distant sniper, thereafter he was very fearful **and** refused to show many experiments, including the antigravity experiment and particularly its extension to the hover point and upwards acceleration extension. In my opinion he had good reason for his paranoia and his fear for his life.

[234] The reason for ELF frequencies is simple: for the most powerful curvatures of spacetime, one wishes to use EM fields whose photons have the greatest *total* energy (including time-energy). That of course is the lower frequency photons, not the higher frequency photons. So ELF is the way to go, not gamma rays etc.

Here is why we advocate a Dirac sea hole's negative energy field as a practical antigravity inducing mechanism.[236] At COP»1.0, there is a very appreciable vacuum energy density difference between the operator's energy input and the energy output of the system, *particularly when the Heaviside nondiverged energyflow component is taken into account as well as the Poynting energyflow component?*[237] There is a production of negative EM mass-energy x time (causal Dirac sea holes) in the immediately surrounding vacuum. That represents the production of a causal, unobserved antigravitational field (curvature of spacetime) with the system-associated anticharge as its source.

Meanwhile, the Earth's causal, unobserved gravitational field is still normal and unchanged. Now the two causal fields superpose, and interact with the system mass. As seen by the external observer, the net effect is a reduction in the weight of the object — the effect of the interaction of the system mass with the two superposed causal gravitational fields simultaneously.

Reasoning that increasing the COP of the Sweet unit should increase the ratio of output negative energy to output positive energy even further, 1 calculated that doubling the output of the device would result in a sufficiently intense negative mass-energy field in the immediate space around the unit, to approach or reach levitation of the unit. However, as the increasing curvature of spacetime would further increase the deposition of magnetic monopoles in the magnets,[238] at some point the magnets would simply explode like hand grenades if the unit were pushed too far. With no explosive control facilities available, it appeared that doubling the output was the maximum that could safely be done under the circumstances. My own estimate showed that 1500 watts output would be required for levitation, but one dared not go beyond 1000 watts or one would risk a fragmentation explosion and injury or death.

---

[235] I strongly warned Floyd not to exceed a 1 KW load, no matter what. In the resulting local curved spacetime surrounding the barium nuclei in the magnets, it appears that magnetic monopoles are deposited or appear. Since these barium ferrite magnets were made by pressed powder metallurgy, they had little tensile or torsion strength, and under sufficient stress they would explode violently like a fragmentation hand grenade. In his little lab (a converted bedroom), Floyd had no explosive facility. He did explode several magnets, but while cautiously remaining some personal distance from the unit as its output was increased. We warn the reader that experimentation with such procedures is at his own risk, and all proper laboratory safety precautions must be utilized.

[236] A provisional patent application has now been filed on the process in 2002.

[237] In a nominal circuit, the unaccounted Heaviside energy flow may be up to 10^{13} times as great as the Poynting energy flow component. This is a completely unaccounted, vast EM energy flow accompanying every field and particle interaction. Hence one can understand why such an energy density for a macroscopic system with COP»1.0, is sufficient to achieve on a macroscopic scale what LaPointe is attempting on a microscopic scale. The total macroscopic scale energy density is so great that the required LaPointe microscopic energy density {535} is present or approached.

I convinced Sweet to build a 1,000-watt output load box, which simply contained ten sockets for ten 100-watt light bulbs. This was a modification to the 500-watt VTA unit shown in Figure 6-8 in Chapter 6. By successively installing the bulbs one by one, the load could be adjusted in 100-watt increments up to 1 kilowatt.

The foregoing is the reasoning behind how I designed the experiment, and the purpose was to show very substantial and smooth weight loss of the VTA unit itself. The experiment was performed in that 100-watt incremental manner, with highly successful results as shown in Figure 6-14 in Chapter 6.

With each 100-watt increase in power output past 500 watts, the fraction of negative energy produced was greater. Hence the unobserved 4-positron fields (Dirac sea hole negative energy fields) were stronger, because there were more unobserved 4-positrons. This meant that the local negative mass-energy curvature of spacetime was increased also.

In turn, this negative curvature of spacetime interacted back upon the mass of the VTA to add negative mass equivalent, or more simply, antigravity force. The observed weight of the unit was the sum of the normal gravitational weight (due to earth's gravity curvature of the local spacetime), and the abnormal antigravity weight (negative weight) due to the antigravity curvature of the local spacetime superposed. So as the fraction of negative curvature of spacetime increased, the net weight of the VTA was observed to decrease.

[238] Contrary to popular opinion, one does meet with effective magnetic monopoles in a magnetic field — including in the magnetic field of a permanent magnet — in a strongly curved local spacetime. The stronger the field, the greater the magnitude of the effective monopole charge.

As stated, Sweet was performing the experiment on the bench in California, and reading the instrumental results over the phone (I was in Huntsville, Alabama) where I recorded them. The spectacular results are shown in Figure 6-14 in Chapter 6.

The high COP (corresponds to gain) of energy density present in the Sweet VTA system's output section — with a greater percentage of it being negative energy — versus input energy one inputs to the input section, is what is important. In short, increasing the load effectively increased the gain, which further increased the ratio of the negative mass-energy to the positive mass-energy. This in turn steadily decreased the observed weight of the object.

Anything outside the system mass of that very high COP system will "see" the weight of the system mass as having decreased. Weight is just so much mass making so strong a *net* G field (spacetime curvature) in the surrounding spacetime. Reduce the *net* G-field that the system mass is normally making (by having it *simultaneously* make the additional antigravity field which then vectorially adds to the mass's normal G-field), and the external observer, external scales, etc. will see the object as having very much reduced its weight.

Positive spatial EM energy density of a system makes positive gravity (as seen by the external observer in the lab frame) because of the way it curves the immediately surrounding spacetime and the way that curved spacetime interacts back on the system's mass-energy. Negative spatial EM energy density makes negative gravity because it curves the immediate spacetime the opposite way. A system under those "high negative energy field" conditions will produce a great deal of antigravity, because the immediate spacetime surrounding is dramatically curved by the negative energy. That was my concept when designing the experiment, and it worked beautifully. So that is how the Sweet device reduced its weight on the bench by 90%, for a 1 kilowatt total output and minuscule input power {544}.

If one were dealing with the source in a flying vehicle system, there is another effect due to momentum. Note that force F is defined as $\mathbf{F} \equiv \partial p/\partial t = +\partial/\partial t\,(\mathrm{mv})$. That is in a positive energy density local spacetime. When an extra negative energy density is added to the local spacetime, then there are two forces, a new one being $\mathrm{F} \equiv -\,\partial/\partial t\,(\mathrm{mv})$ and the normal one being $\mathrm{F} \equiv +\partial/\partial t\,(\mathrm{mv})$. These forces are created by any change in momentum in space containing an extra increment of negative EM field energy.

If the two forces balance, then there is essentially no net centrifugal force produced on the object (or any occupants in it) in a turn. In theory, maneuvers that are far more powerful can be made, without undue G-stress to any occupants inside. This of course remains to be tested in future work and experiments, and so must be taken as a strong hypothesis at present. It could be tested, given availability of the proper COP»1.0 energy system or process on board a test vehicle. Laithwaite's demonstration of the dramatic reduction in force necessary to lift a rotating gyroscope compared to lifting the same heavy gyroscope when not rotating, may have been a partial test of this "inertial resistance decrease" effect.

The Sweet device was the only available COP»1.0 device which could have been used for the experiment. That is why I designed the experiment and convinced Sweet to perform it. Had it not worked, much of what I had worked out for more than a decade would have been falsified — which of course is one of the purposes of experiment: to validate or falsify a thesis. In this case, I breathed a great sigh of relief because — in my view — the spectacular results completely validated my approach to antigravity, and also much of my approach to COP>1.0 systems.

Sweet was never in a position to be able to go with this to the scientific community. The VTA was completely fouled legally, by various entangling agreements Sweet had made with different backers, without ever changing or resolving any previous agreement. Also, he realized that his life really was in danger should he try to take the system openly to the scientific community. Consequently, he never tried and in fact adamantly turned all further work away from the antigravity effects.

Aging and rather defenseless, Sweet also began playing all sorts of games to prevent having to have the project validated independently, even for its COP>1.0 capability. He was mortally fearful of being killed if he had it resoundingly validated. He was motivated by the necessity to survive and not be killed, and he firmly believed that any such attempt for full scientific validation and open publicity would absolutely guarantee his quick death. His fears appear to have been well founded.

His most carefully guarded secret was his activation process for the barium ferrite magnets, which he guarded with his life. He also began telling different persons all sorts of different stories and "mechanisms" for the activation, which were clever ruses and disinformation designed to throw them off the true trail and prevent his activation secret from being uncovered. I made a personal videotape of the activation process, or what he released of it, before he began such disinformation. So we do know much of the activation process, though not all of it.

Eventually, with changes in backers yet once again, Sweet died and never revealed the full secret of strong self-oscillation activation in permanent magnets to anyone. I knew and know part of it, even most of it, but still not all of it. What I understand about it has been briefly included in this book. From there, the interested researcher must find his own way.

The VTA was real, my gravitational mechanism is real, and — in my opinion — validated. However, it has not been *independently* validated in formal order, which is the proper scientific requirement. I did try very hard to convince Sweet that we should take this to the leaders of the scientific community and have it independently verified (e.g., by UCLA), but he would not hear of it after the assassination attempt. By that time, he was so afraid from the constant threats and constant stalkers that he had totally decided against any such releases. He also would not permit it to be taken to a major scientific conference and demonstrated, even though several scientists and engineers knowledgeable and experienced in measurement engineering did measure and certify the device's operation. But they were never allowed to see it in the antigravity mode. He would not discuss it with them, and he would not allow it.

Sadly, with the death of Sweet there passed away the VTA and its major secrets. I regard the VTA as one of the great scientific accomplishments of all time, and I regard Sweet as one of the most capable lab experimenters I ever met and one of the greatest inventors I ever met. The loss of the VTA and the absence of its full, independent scientific testing and validation were in my opinion two of the greatest losses to science of all time.

8.4.3 Discussion

Particle physicists speak of antimass and antigravitational field, but do not speak of anti-electromagnetic field and do not like negative mass. Antimass is actually "anti-mass-energy". When did one hear any scientist speak of a *negative energy* electromagnetic field or *negative energy* force field (except possibly as a hypothetical "tractor beam")? Or a *negative energy* electrical or magnetic potential? Yet the EM fields from an accumulation of Dirac sea holes (causal state) as a source charge are precisely negative energy EM fields. The EM potentials from an accumulation of Dirac sea holes are precisely negative energy potentials. Further, these fields and potentials produce opposite curvatures of spacetime as compared to positive energy causal EM fields and potentials from conventional source positrons and electrons.

Conventionally, scientists speak and think of *the positive energy* electromagnetic field. In our view, they do not apply the implications of the negative energy state part of the Dirac 1930 electron theory {545}, because of a built-in horror of negative mass (and antigravity!). Recall again our discussion of Laithwaite; a promising career was cut down in mid stride because he "mentioned the unmentionable": antigravity and possible failure of Newton's laws (they *do* fail in a curved spacetime!). His exhibiting to the Royal Society a very heavy gyroscope, very difficult to lift when placid, but easy to lift with one arm when spinning, was the final straw. The British scientific community simply cut him down and ostracized him from the rewards and positions he would otherwise have been given.

To conceal the relationship of the negative energy fields, hide negative mass, and pretend it does not exist, much scientific effort is exerted to consider the positive energy, positive mass positron *only after it has been observed.* So scientists can assign this "observed" 3-positron a positive mass and positive energy but with spatial direction reversal. That immediately moves to an *effect* field and an *effect* positron, not a causal field or a causal Dirac sea hole. It effectively eliminates the antigravity field of the negative mass of the unobserved negative energy Dirac electron, which is produced by the *causal* (unobserved) Dirac sea hole's electromagnetic field, not the *effect* (observed) positron field which then is "seen" to have positive mass but with field direction reversed.[239]

The positron is quite a different beast before it is observed, just as is a mass (and just as is any other observable, and just as is a field). For one thing, the positron before observation and the positron after observation have different dimensionalities; the unobserved causal entity is 4-spatial and continuous,[240] while the observed effect entity is continually and iteratively 3-spatial. Mass before it is observed exists as masstime, not mass.

---

[239] Note also that the causal field, being unobserved, is a 4-field, while the observed (effect) field is a 3-field. By reversing the parity, one has made the antigravity field **into** a gravity field, by substituting the effect for the cause and erroneously using the old non sequitur of a separate 3-force acting upon a separate 3-mass. Also, when one reverses its spatial direction, one converts a *positive* gravity field into an *antigravity* **field.**

[240] Along with Bohm, we point out that — strictly speaking — in 4-space one does not have "things", but only "processes". To speak of a "4-space field" is to redefine the field as an ongoing process, rather than as an observed frozen 3-snapshot or "3-thing". The solution to the long-vexing duality problem is there, but we leave its solution to the reader. It is necessary to extend Aristotelian logic itself if one is to resolve the duality problem.

Simply consider the delayed choice two-slit experiment and Wheeler's famous comment that no phenomenon is a phenomenon until it is an observed phenomenon. Once the observation occurs, one can even change the "if it had been observed but was not" previous past. One gets the magic from the delayed choice two slit experiment without the observation, and then by making the observation such that the "magic" is not revoked *after the fact.* The whole experiment proves that, if one insists on applying the observation, one kills the magic, even if one thinks the magic has already occurred and must have already occurred, but it has not been *observed* to have occurred.

Causal 4-spatial "things" are not things at all, but are ongoing processes and their cessations have not "occurred" to yet produce a "thing"! The *observed* event is 3-spatial and has occurred. It is something in the past, not the present. That is different from something ongoing (happening or unhappening)[241] in 4-space, before having been observed and therefore a frozen snapshot taken in 3-space. As is well known in quantum mechanics, observation is a physical interaction and it indelibly changes and stops the "ongoing 4-process", converting it to a momentarily frozen "3-thing" as an effect of the observation process at that particular instant. So any "thing" is continual and iterative, *not continuous,* in its travel through time, and in its "observed existence in 3-space".

Because of such, physicists also have difficulty realizing that mass — being an observable — does not and cannot exist continuously in time. It cannot even *continuously exist!* No observable can. Masstime exists in time, but mass does not, as we previously discussed. Observation is a *dldt* operator imposed upon an ongoing 4-space process, destroying the time portion and producing an instantaneous, frozen snapshot at that instant. It's just *d/dt* (LLLT) => LLL. That is why all observation is 3-spatial, as is well known in physics. We long ago pointed out how the common photon interaction — the total photon interactions (both virtual and observable) with a mass — is what drives the mass through time, so to speak {15}. But it does it iteratively, frozen snap after frozen snap —just like progressively showing the frames of a movie film.

---

[241] When a frozen instant 3-space snapshot (observable) interacts with a second causal process to generate another frozen snapshot (observable), the first observation "unhappens" because it is thereby moved into the "past", disappearing from 3-space forever. With every change of an observable (with every observation), even if the new effect appears identical to the previous one, it is still a *new* observable (a new effect) where a former effect "seems" to have recurred (been created, happened). Even if it is spatially the same as the previous observation output, it is at a different point in time and hence is the result of the *d/dt* observation operator being iteratively applied. The serial outputs — even of what we loosely call an "entity existing in time with no 3-space change" — are *successive and different* serial outputs from a serial time-differentiating process.

What really has to be changed is the old notion by Heaviside and Poynting that 3-energy propagates continuously through 3-space.[242] No observable can propagate continuously in 3-space, but only in 4-space and then only during its nonobservable state! So a "mass propagating through time" (i.e., just sitting still and "existing", so to speak) is doing so in an iterative fashion by recurring iteratively in observation output at the same spatial position. The mass turns to masstime, then to mass, then to masstime, etc. Or $m_1 \Rightarrow m_1 t \Rightarrow m_2 \Rightarrow m_2 t \Rightarrow m_3 \ldots$ etc. In short, *observation continually recurs at an enormous rate,* so *an observable (effect or output of that observation process) continually recurs at an enormous rate,* destroying the previous observation as the next one occurs.

Because of the history of thinking in 3-space terms rather than 4-space terms, one of the ubiquitous errors widespread in physics is the substitution of the effect (the frozen output of the observation process) for the cause (that 4-space initiator process which continuously exists prior to observation, hence unobserved). So causes (4-space initiators which are unobservable a priori) interact with a previous "frozen 3-space snapshot" or observable, to produce a change (either a temporal change or spatial change or both) in or to or of that observable (that 3-snapshot) to a new observable (3-snapshot). If the object is seen to iteratively change in time but not position or form, it is said to "exist" or "persist" at rest continuously in time. It doesn't. It *continually recurs*[243] in that form, but does not *continuously exist* in that form!

---

[242] And also the hoary old mistaken notion that the "same observable thing" can continuously exist in time or persist.

[243] More rigorously, the continual recurrence is at the quantum level. So a larger mass or other observable actually is a great horde of very small "continual recurrences" of tiny incremental pieces of itself. At any time interval of observation via observable photons, a vast number of nonobservable (virtual) "continual recurrences" occur in the "observed object". In short, the "observed object" is the **end** summation and result of a vast number of infinitesimal changes. Hence the entry of statistics and the uncertainty principle. Each one of the infinitesimal

The antigravity mechanism has been implicitly present in the Dirac theory of the electron since his 1930 paper {497}, where he brought forth the Dirac sea and the anti-electron, if the *causal* Dirac sea negative energy states are applied to general relativity before they interact with charge and are "observed". General relativity deals with spacetime and spacetime curvature. Therefore, it is concerned with causal 4-fields, 4-potentials, and 4-waves in 4-space prior to observation.

As we stated previously, there is really no such thing as a "3-space EM wave", e.g., as rather strongly pointed out by Romer {546}.

One can produce and have an EM field of *negative* EM energy in spacetime, coming from the 4-positron charges of the anti-charged mass of a system, as well as one can produce and have an EM field *of positive* EM energy in space, coming from the charged mass of a system. But negative energy EM fields have been swept away from our minds by interpreting the negative energy 4-electron as the 3-positron, after the negative energy 4-electron has interacted with charged matter (been observed as the *effect),* rather than how it exists in spacetime prior to observation (as the *cause).*

So physicists reverse the direction and the charge, and make the mass positive, which is how the Dirac sea hole is "seen" *after observation.* That makes it a material lattice hole and a special kind of positive ion. Before observation, it is not seen at all, but it exists as negative action (energy x time) and therefore in "negative masstime" state form. Before it has been observed, it's still there in spacetime, as a curvature of spacetime. It's just a *negative energy state curvature* in the vacuum/spacetime.

Make lots of these negative mass-energy state Dirac holes from a source system, such as one producing a very powerful negative energy EM field, and voila! That's it. The system is producing — and surrounding itself by — its own negative energy EM field (negative curvature of 4-space), which is also its own local antigravity field. That antigravity field superposes with the prior local curvature of spacetime, altering it. The *net* curvature of local spacetime is itself a field which interacts back upon the mass system to produce "net gravity" whose sign may be positive or negative, depending on the sign of the net spacetime curvature. If the two back-acting curvatures of spacetime are opposite and equal, that represents the source object existing in a zero gravitational field, and thus "hovering" in the Earth's gravitational field, for example.

changes may be modeled causally, but since there is no individual information available on each one of them, they can only be calculated statistically.

The use of negative energy EM fields and Dirac sea holes to alter the curvature of the local spacetime in a negative direction provides the present author's engineering approach to practical antigravity. Note that our approach gathers in the gist of both of the major two approaches to antigravity: (1) the shielding theory as typified by Modanese {547}, and (2) the opposing force theory as typified by Ling Ni and Torr {548}. The opposing force, considered as a field that is identically a curvature of spacetime, is also a "gravity shielding" force via superposition of fields. We state unequivocally that a theoretical model of the new approach can be constructed from Sachs's theory {510}, and implemented in the 0(3) electrodynamics of Evans and Vigier {549}. Further, it is a directly engineerable theory, without the necessity for superconductivity, spinning wheels, etc. as demonstrated by the highly successful Sweet experiment previously discussed.

We quote Evans {550}, on the appearance of the B(3) field in general relativity, to indicate the nature of such a future theoretical model:

> *"I have chosen a metric that gives circular polarization (as observed), but in the framework of general relativity using the irreps of the Einstein group according to the Sachs theory. This is a closed field theory which is analytical, using a Lie group, the Einstein group. As such there are no particles, only spacetime curvature, so everything in physics reduces to spacetime curvature, i. e. to a metric vector and its components. This is then incorporated in the Sachs theory to produce a metric four-vector which is a generally covariant component of a quaternion-valued metric with sixteen components, the most general form of the metric allowed by general relativity and topology. The Einstein-Sachs equations for electromagnetism are six metrical equations in six unknowns, the components of an antisymmetric field tensor of electromagnetism. By choosing a metric that corresponds to circular polarization,* $B^{(3)}$ *appears automatically.*
>
> *This method gives the* $B^{(3)}$ *field from Einstein's theory of general relativity applied to electromagnetism by a consideration of the irreps of the Einstein group. It follows that whenever the conjugate product* $-gA^{(1)} \times A^{(2)}$ *is observed, the* $B^{(3)}$ *field must also be observed, otherwise general relativity fails. Examples of*

> *the $B^{(3)}$ field at work are the inverse Faraday effect and RFR [Radiatively induced fermion resonance]. "*

Evans and the AIAS have already taken the first theoretical steps showing that antigravity does indeed arise from Sachs's theory {551a-551c}.

We are convinced that practical antigravity only occurs at extremely low frequency (ELF) for reasons[244] explained previously. Certainly that seems to be true for our approach to it. The highest "total energy" photon is the low frequency photon with low spatial energy and a larger time increment. The larger time component has the same energy density as mass. So one multiplies the time component increase (in seconds) by roughly $9x10^{16}$ to get the increase in spatial energy equivalency. Halve the frequency of a photon, and one increases its *total energy* by $4.5x10^{16}$ {552}.

High energy physicists have not realized that the time component of the photon has enormously more energy than the spatial component, and that low frequency photons contain enormously more trapped energy than do high frequency photons. So they largely ignore the time-energy component, focus on the spatial energy component, and erroneously consider *high energy* physics as *high frequency* photon physics. Instead, present high frequency photon physics is only a high *spatial* energy physics, not a high *total* energy physics at all. Today's high energy physics is really still a very low total energy physics, compared to what can be done with the time-energy component of the photon rather than its spatial-energy component. By transducing and using the highly compressed time-energy, experiments can be done on the lab bench with fairly simple and cheap equipment, that actually utilize *higher total energy nuclear reactions* (though at low spatial energy) than are presently obtained in the largest accelerators.

That is also why there is presently no understanding in the conventional community of the "low spatial energy transmutations" of cold fusion, where the experiments unwittingly use the high time-energy component. The scientific community has not even recognized the time-energy aspects, and has not recognized that time-energy physics is an ultrahigh total energy physics, where the present "high energy physics" is tame by comparison.

---

[244] Dramatic increase in the magnitude of the time-component, hence an enormous increase in the time-energy component with only a small decrease in the spatial energy component total energy of the photon. The *total energy* thus dramatically increases, when the spatial energy equivalent of the time-energy is considered. The present conservation of energy law in physics is inadequate because it does not account time energy at all; thus completely discards the most fundamental EM energy mechanism of all: the giant negentropy of source charges and source dipoles.

The previously unknown time-energy induced transmutation interactions are the precursors to, and the opening salvo of, a new *ultrahigh total energy physics* struggling to be born in spite of the adamant opposition of the conventional scientific community. Allowing for the time-energy component, for the mechanism generating the flow of time, and for formation and decay of time-reversal zones, we previously wrote the exact new nuclear reactions for the production of the excess deuterium, tritium, and alpha particles {553} in numerous cold fusion experiments. We also explained the highly anomalous nuclear instrument effects — due to (i) time-charging, (ii) the slow decay of time-charge with radiation of Whittaker longitudinal waves, each accompanied by its phase conjugate twin time-polarized wave, and (iii) the previous time-charge history of the instrument. These puzzling anomalies have been observed at U.S. Navy research facilities at China Lake now for some years in rigorous electrolyte experiments {198}. We summarize that work in Chapter 10 of this book.

With a little effort and development, practical antigravity appears to be straightforward for any good modern laboratory. But the lab physicists should first (i) rethink the entire subject, from its very fundamentals, to include considering the positron before observation and not after observation, (ii) remove the Lorentz symmetrical regauging from electrodynamics and recover the presently missing COP >1.0 and COP»1.0 Maxwellian systems again, and (iii) rapidly develop COP»1.0 systems and processes to use in antigravity experiments and development. At least in this approach, the COP»1.0 EM systems — once developed — are an immediate gateway to practical antigravity devices, unilateral thrust propulsion systems, and antigravity vehicles. And of course they are the immediate gateway to practical transport vehicle power systems, fuel-free, clean, and self-powering.

Hopefully, such a scientific reassessment will be performed, and it may provide the gateway for practical transportation and travel to the ends of **the** solar system and beyond, to be achieved in the first two decades of this century.

8.5 In Conclusion

We have briefly reviewed the presently developing field of inertial space propulsion and antigravity, and presented our own approach to practical antigravity as well as an informal experiment that strongly supported that approach. We have strongly stressed that the causal field from the system as its source constitutes a curvature of the local spacetime superposed upon the pre-existing local spacetime curvature. Instead of thinking of that causal field from the source system as acting back upon the source system, we have stressed thinking of that *net superposed curvature of local spacetime* acting back upon the source system and producing a net gravitational force upon it. That net gravitational force may be either positive or negative, or zero, depending upon the sign of the net curvature of spacetime that is interacting back upon the system.

We presented and explained the results of the Sweet VTA antigravity experiment as probable verification of our approach to antigravity. Nonetheless, we also emphasized that the formal independent replication required by the scientific method has not been accomplished, and it remains to be accomplished before the approach is to be considered as scientifically validated.

We believe that this approach to antigravity can be modeled and engineered by use of the Evans-Vigier O(3) electrodynamics as a subset of the Sachs electrodynamics that is a part of the Sachs unified field theory.

If so, then tentatively we have laid out the basis for eventual practical antigravity systems and a practical antigravity technology. We fervently hope that this approach will in fact lead to practical technology and working transportation means during the next decade.

Finally, we hope that these notes and concepts will stimulate many sharp young graduate students and post-doctoral scientists to examine, model, correct, and greatly enlarge the ideas and principles advanced. If so, then the purpose of this chapter will have been realized.

Chapter 9

The Supersystem and Remarks on Gravity, Antigravity, and Testing

"What might appear to be empty space is, therefore, a seething ferment of virtual particles. A vacuum is not inert and featureless, but alive with throbbing energy and vitality. A 'real' particle such as an electron must always be viewed against this background of frenetic activity. When an electron moves through space, it is actually swimming in a sea of ghost particles of all varieties - virtual leptons, quarks, and messengers, entangled in a complex melee. The presence of the electron will distort this irreducible vacuum activity, and the distortion in turn reacts back on the electron. Even at rest, an electron is not at rest: it is being continually assaulted by all manner of other particles from the vacuum." [Paul Davies] {554}.

"Space acts on matter, telling it how to move.. In turn, matter reacts back on space, telling it how to curve." [Wheeler's principle] {555}.

"...the Aharonov-Bohm effect is a local gauge transformation of the true vacuum... [which] produces a vector potential from the true vacuum. [This gauge transformation] produces topological charge..., the electromagnetic field, which carries energy, and the vacuum charge current density first proposed by Lehnert... and developed by Lehnert and Roy... ". [M. W. Evans and S. Jeffers] {556}.

"The energy density (En_d) in curved spacetime

is given in the Sachs theory by the quaternion-valued expression

$$En_d = A_m j^*_m \qquad [1]$$

where A_μ *is the quaternion-valued vector potential and* j^*_m *is the quaternion-valued 4-current as given by Sachs ...Equation [1] is an elegant and deeply meaningful expression of the fact that electromagnetic energy density is available from curved spacetime under all conditions; the distinction between field and matter is lost, and the concepts of "point charge" and "point mass" are not present in the theory, as these two concepts represent infinities of the closed-field theory developed by Sachs from the irreducible representations of the Einstein group. The accuracy of expression [1] has been tested to the precision of the Lamb shifts in the hydrogen atom without using renormalization of infinities. The Lamb shifts can therefore be viewed as the results of electromagnetic energy from curved spacetime."* [Myron W. Evans] {557}.

"Old habits of thought die hard, and it is not always easy to remember that a particle with negative mass if pushed to the right will move to the left; or that in a given reference frame a particle with imaginary mass can have infinite speed, in which case its energy is zero." .... "... extremely powerful energy sources may occur if particles of negative mass really exist in nature. " [Banesh Hoffman] {558}.

"The observed (3-spatial) positron is a quite different entity from the unobserved (4-spatial) Dirac hole. Its direction has been reversed, and its mass and energy have also been reversed. Before the 4-spatial hole in the vacuum has interacted and been observed, it has negative energy and therefore negative mass equivalency, as well as negative charge. As a source charge, it also is the source of negative energy fields and negative energy potentials — all in the

nonobservable 4-state. " [T. E. Bearden, private communication to a correspondent, 2001].

"The...visible structure of orthogonal' 1-networks possessing coupled open-paths and closed-paths represents only an introductory step in the development of a correct topological theory for conventional electric networks...An electric network is like an iceberg...Its visible portion is only a small fraction of the submerged invisible structure. " [Gabriel Kron, who developed special tensor methods and diakoptics to deal with the large number of visible and invisible multidimensional p-networks surrounding the branches of every electric circuit.] {559}.

"...symmetry implies conservation. Since our entire edifice of interactions is built on symmetry assumptions, there should be as a result a large number of conservation laws. The only trouble is that almost all of these conservation laws have been violated experimentally. "... "...this difficulty could be resolved by introducing a new element, the vacuum. Instead of saying that the symmetry of all matter is being violated, we suggest that all conservation laws must take both matter and vacuum into account. If we include matter together with vacuum, then an overall symmetry could be restored" [T. D. Lee] {560}.

"One perhaps oversimplified way to deal more conventionally with the new unified field approach to electrodynamic systems is to utilize the notion of the 'supersystem', comprised of three components: (I) the ordinary EM system and its dynamics—conventionally assuming flat spacetime and inert vacuum, (2) the local nonlinear vacuum and its dynamic interactions with the system and with spacetime, and (3) the local curvatures of spacetime and their dynamic interactions with the system and with the local

> *active vacuum. In the rich feedforward and feedback interactions that result, there emerges a dramatic new physics, new electrodynamics, and new chemistry including a new biochemistry. This is in fact a much simplified conceptual approach (good to 'first order', so to speak) to the multinetwork and multipath work of Gabriel Kron. It also allows the retention of symmetry, considered so necessary to particle physicists. "*
> [T. E. Bearden, private communication to a correspondent, 2001].

9.1 The Supersystem and Its Considerations

In classical electrodynamics, the Maxwell-Heaviside-Lorentz equations implicitly assume a local flat spacetime and no net action by the vacuum {561}. Both these implied assumptions are false. All real EM systems involve interactions with both curvatures of spacetime and the active vacuum. Yet to try to include both additional interactions requires a unified field theory approach, such as that of Sachs {562}. Unfortunately, any unified field theory is replete with quite difficult mathematics, little of which can be illustrated pictorially or simply.

What is required — at least for most electrical engineers and applied electrical scientists — is a conceptual way of visually thinking about the EM system in a way that is consistent with unified field theory to something like first order, but which can also be conceptualized easily.

9.1.1 The Supersystem Concept

Accordingly, we introduce a new and very simple way of thinking of electrical circuits in terms of a unified field theory. We utilize the notion of the supersystem, consisting of three components: (1) the EM system and its dynamics, considered in conventional fashion (as if the local vacuum were inactive and the local spacetime were flat but perturbed), (2) the local interactive vacuum and its dynamics, and (3) the local curvatures of spacetime and their dynamics. All three components of the supersystem interact. This produces both feedback and feedforward functioning; e.g., energetic changes in the system produce corresponding changes in spacetime curvatures. In turn, the alterations in spacetime curvature produce nonlinear changes in the active vacuum, which in turn acts back upon the system. So every action in a system produces a corresponding set of feedforward and feedback reactions from and with its complex environment, and vice versa.[245]

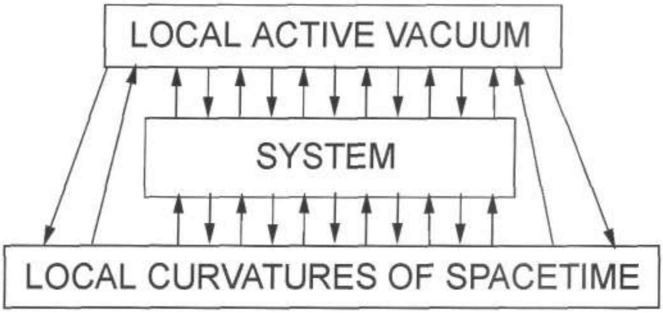

Figure 9-1 The Supersystem concept (block diagram).

Figure 9-1 shows a diagrammatic illustration of the supersystem concept. Figure 9-2 illustrates the effect of the classical EM assumptions of a flat spacetime and a locally inert vacuum. As can be seen, the classical assumptions *arbitrarily discard the* interaction effects upon and in the system due to the active vacuum dynamics and due to the local spacetime curvatures and their dynamics.[246]

As stated, all three components of the supersystem interact with each other. This is a dramatic change to classical U(1) electrodynamics. Hence the total interactions — with the system — of both the active vacuum and the locally curved spacetime, and their interactions with each other, are more general than either classical EM, general relativity, or the quantum mechanical vacuum of particle physics prescribes individually.

---

[245] We hypothesize that the feedback contains the mechanism generating Newton's third law as a restoration of symmetry, but leave that to the master theorists to explore and decide.

[246] As one example, the nondiverged Heaviside component of energy flow — previously discussed in earlier chapters — is part of the supersystem activity. This **flow** alters both the local active vacuum and the local spacetime curvature associated with an electrical power system and its connected external circuits. Since the resulting local "environmental" changes in the supersystem are extensive, then one can conceive ways in which these changes can be interacted with additional intercepting/receiving systems to produce extra "EM energy from the vacuum".

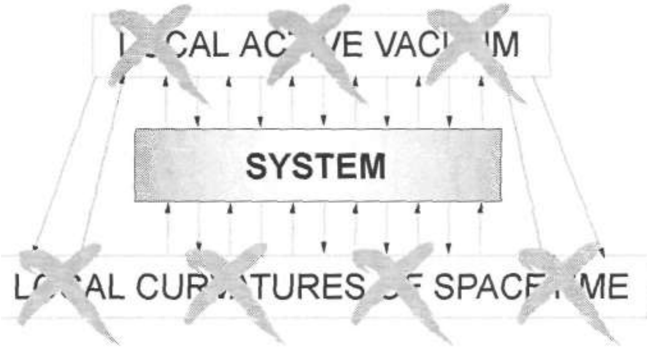

Figure 9-2 Classical EM destroys two-thirds of the supersystem.

Immediately it can be seen that the concept of energy from the vacuum is a "supersystem active and not negligible" concept involving a unified field theory approach *a priori.* It can also be seen that a "system" in equilibrium must be in equilibrium with its active environmental exchange — its exchange with the active local vacuum and the active local curvatures of spacetime. Since energy "exchange" with the system means both incoming and outgoing exchange, it follows that the equilibrium system has equal incoming and outgoing energy exchange. It follows that the symmetrically regauged system has equal incoming and outgoing energy, and so the symmetrically regauged system must be in additional stress from two net equal and opposite direct interaction forces, one general relativistic and the other from the active vacuum. Further, this "equal and opposite" equilibrium stress state must be accompanied by equal and opposite interaction effects between the two environmental media of the system.[247,248]

---

[247] We point out but do not pursue the fact that this is a very important new concept associated with *coherently* changing the symmetrical regauging, which allows a new physics, a new electrodynamics, and a new chemistry including biochemistry. For one thing, it permits such things as a self-maintained excitation state that simultaneously continuously discharges its excitation energy. This is simply a redefinition and more general broadening of the "source charge" concept. From the normal system view (assuming away the vacuum and curved spacetime interactions), an "impossible" system state can nonetheless be possible and stabilized, by changing the *intensity* of the Lorentz regauging condition. As an example, new molecules or ions can be made and stabilized in an excited state, that otherwise would be so unstable that the excitation would decay in a nanosecond with the state disappearing also. We believe this alone will lead to a great new biochemistry and medical

9.1.2 Time and Space Effects and Positive and Negative Energy

From curved spacetime considerations, we must include the effects on the system, and on the vacuum, of dynamics in the *time* domain as well as in the *spatial* domain. Indeed, in Sachs's theory space and time do not separately exist, but only spacetime. We consider time as highly compressed EM energy — as heretofore pointed out, where time is spatial EM energy compressed by the factor c^2 and having energy density similar to mass. We must include not only EM *energyflow* along the time-domain (fourth Minkowski axis *ict),* but also EM *force, potential, energy oscillations*, etc. in the time domain itself. We shall have to deal with time-force, time-energy, time-potential, time-energy current, longitudinal EM waves in the time domain (time-polarized EM waves), etc. {563}. The complexity of the full supersystem analysis is thus easily seen and appreciated. In most cases, a higher group symmetry model for a particular situation will prove intractable for closed solutions, and numerical methods must be used.

Further, when curved spacetime interacts with the active vacuum, we shall also have to consider *negative energy* currents in the nonlinear vacuum itself, as well as in the time-domain. We may also speak of the *negative energy density* in the vacuum as a special kind of negative energy *potential* in the vacuum that is created by a negative energy potential in the time domain. We may speak of a gradient in this vacuum negative energy potential as a negative energy force or "force in the negative energy domain". Applied to the time domain, these concepts constitute a causality condition placed on the Dirac sea, in the same vein as more technical causality considerations given by Finster {564}.

9.2 Differences between Effects in COP»1.0 and COP<1.0 EM Systems

In this chapter, we utilize the Dirac hole in a different manner from the received interpretation. The received interpretation assumes an electron hole in the vacuum, with the hole filled with a negative energy electron

---

therapeutic science; one company is already working in this area, with positive results.

[248] We leave to the advanced theorists the connection between this "stable equilibrium state under a change in symmetrically regauged stress", occurring as a *net-force-free* intensity change in the conventional Lorentz symmetrical regauging of the Maxwell-Heaviside equations.

having negative mass, negative energy, and negative charge. In that interpretation, when the electron is lifted out of the hole, the hole is said to be *a positive energy* electron with positive mass and positive charge, (i.e., when interacted and observed). In short, with reaction one has a complete transformation of negative energy into positive energy, negative mass into positive mass, etc. In our view, this totally destroys the notion of "how the hole exists in the vacuum prior to its interaction and observation". That kind of positron only occurs after a physical interaction and observation has occurred — in short, *the positron exists only after an interaction with the Dirac hole has been made, to change the very nature of the negative energy hole.* The positron is the 3-effect of an invoked extra 4-interaction, and is not at all the "negative energy electron" or hole as it actually exists prior to being interacted and transformed. The positron is a different beast from the beast we wish to harness and use. The conventional interpretation leaves the question of the unobserved and unreacted "hole" left behind by an electron lifted out of the hole, before another interaction is performed on the hole to change it. So it is an incomplete accounting (as all observations are).

In our interpretation, the Dirac "hole" in the vacuum is itself the true negative energy electron before its interaction and observation. It is the "causal" 4-space negative energy electron, while the received interpretation is only for the "observed" or "effect" positron. In our view, the received interpretation thus leaves out a goodly part (and the most interesting part) of the Dirac sea hole. When the negative energy hole is filled, we consider the *electron* in the hole to be a normal, positive energy, positive mass electron. Hence the summation of the energy of the filled hole is zero, as is its summation mass and its summation charge. This accords with the properties of the "normal" vacuum, said to contain a sea of such "filled Dirac holes".

Our purpose is to use *emptied* Dirac sea holes before their interaction or observation, and thus to use them as *causal 4-electrons having negative mass and negative energy, but where each of those quantities is multiplied by time. So the hole has negative masstime and negative action.* Unless otherwise stated, that is our interpretation of a Dirac sea hole, and that will be how we use it. This way we avoid having to replace a true negative energy, negative mass, positive charge causal (unobserved) 4-electron by a positive energy, positive mass, positive charge observed (effect) 3-electron. We do not just artificially "dispose of" the question of the "nature of the hole that is left behind" when the positron is observed. We consider the (hole x time) as the fundamental negative energy 4-electron. We will sometimes refer to that as the "4-positron".

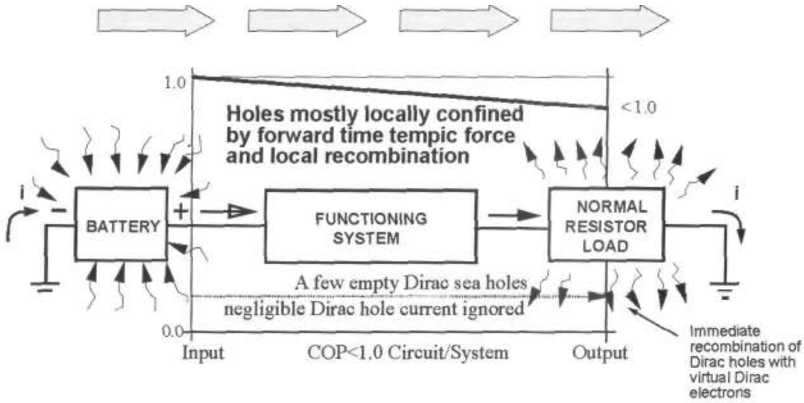

Normal "forward flow of time" tempic force

1.0
<1.0
Holes mostly locally confined by forward time tempic force and local recombination
i
BATTERY
FUNCTIONING SYSTEM
NORMAL RESISTOR LOAD
A few empty Dirac sea holes
negligible Dirac hole current ignored
0.0
Input
COP<1.0 Circuit/System
Output
Immediate recombination of Dirac holes with virtual Dirac electrons

Figure 9-3 Time potential gradient across a C0P<1.0 EM system.

See Figure 9-3. In a COP<1.0 EM system, the tempic potential gradient (tempic force) is oriented from input to output and on out beyond the system. A very slight Dirac hole current moves outward from the output, but is immediately "eaten" by the virtual electron flux fluctuations of the vacuum in the space immediately adjacent to the output. There is a slight Dirac hole current also exiting the system in its losses, but again these are immediately "eaten" by the virtual electron flux fluctuations of the vacuum in the immediately surrounding space.

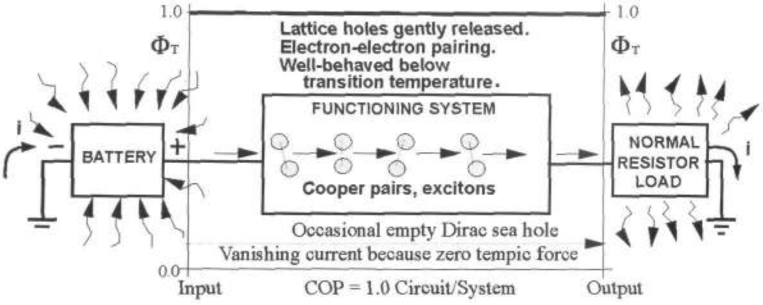

No excess tempic force

1.0
Φ_T
1.0
Φ_T
Lattice holes gently released. Electron-electron pairing. Well-behaved below transition temperature.
FUNCTIONING SYSTEM
Cooper pairs, excitons
i
BATTERY
NORMAL RESISTOR LOAD
Occasional empty Dirac sea hole
Vanishing current because zero tempic force
0.0
Input
COP = 1.0 Circuit/System
Output

Figure 9-4 Time potential gradient across a COP = 1.0 EM system.

See Figure 9.4. In a superconductive system or section, COP =1.0 and on the average there is no net gradient of the time potential across the system, hence no time-force across it in either direction. Accordingly, one sees ready Cooper pairing of charges and formation of excitons. We shall not concern ourselves further with the superconductive EM system; such a system has not yet produced COP>1.0.

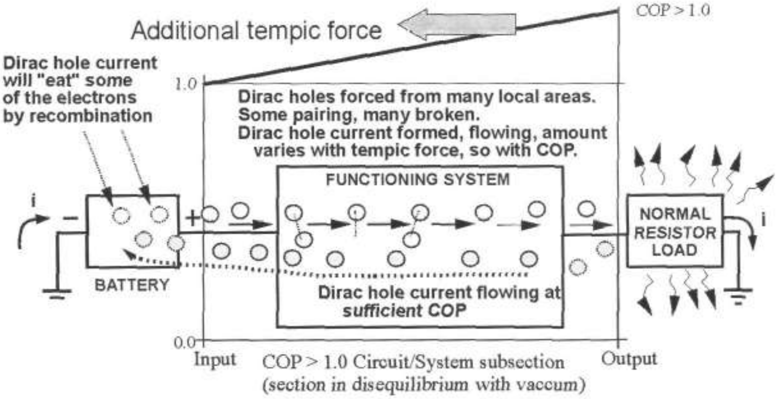
COP > 1.0
Additional tempic force
Dirac hole current will "eat" some of the electrons by recombination
1.0
Dirac holes forced from many local areas. Some pairing, many broken. Dirac hole current formed, flowing, amount varies with tempic force, so with COP.
FUNCTIONING SYSTEM
i
-
+
BATTERY
NORMAL RESISTOR LOAD
i
Dirac hole current flowing at sufficient COP
0.0
Input
COP > 1.0 Circuit/System subsection
(section in disequilibrium with vaccum)
Output

Figure 9-5 Time potential gradient across a COP»1.0 EM system.

See Figure 9.5. In a COP>1.0 EM system, there is a net tempic force across the system from the output section to the input section. However, unless the COP»1.0, the hole current in the local vacuum is mostly used up in interacting with the system material and converting to lattice holes (lattice positrons). Hence at "nominal" or moderate COP>1.0, except in fast pulsed discharges one *usually* has little to worry about, concerning Dirac sea holes in the vacuum, since those readily convert to lattice holes as already accounted. One effect of concern is that, if the semiconductors in the system are exposed to sufficient extra lattice holes or interacting 4-holes, their donor-acceptor arrangement is disrupted, spoiling the action of the semiconductors and perhaps destroying them.

When COP»1.0 (and even for some COP>1.0 systems), then there results a substantial Dirac sea hole current in the local vacuum occupied by the system, flowing through the supersystem from the output section back to the input section. This is an additional hole current component from the

Dirac hole current component converted into lattice hole current.[249] This unconverted Dirac sea vacuum hole current runs through the system from the output section back to the input section of the system and beyond. If not "filled" there in the input section by the incoming input electrons from the external power supply, the 4-hole current can and will pass on back into the feed line and to the distant external power supply, *eating electrons and electron current all along the way.* Usually semiconductors in the path of a substantial 4-hole current will either malfunction or be quickly destroyed.

In the input section and on back into the external power source, the effect of the unrestrained Dirac sea hole current's reaction is to interact and convert into lattice holes all along the route, which lattice holes then "eat" incoming electrons from the distant power source. Thus, the external power source "sees" the hole current as a dynamic additional load in the system's input section itself (and in the external conductors leading to it) *that also must be powered with additional positive energy electron current.*

---

[249] Here we wish to recognize the perceptivity of Peter Lindemann, who recognized the importance of "cold electricity" in COP>1.0 systems without going deeply into the more technical aspects of negative EM energy. As an example, we cite his book, *The Free Energy Secrets of Cold Electricity,* available from Clear Tech, Inc., Box 37 Metaline Falls, WA 99153, U.S.A. Particularly note his efforts to explain the Gray overunity device via the negative energy that can be associated with very strong discharges. Gray referred to this appearance of an associated negative energy component associated with the normally positive EM energy as "splitting the positive pole" (a "pole", after all, involves a stress potential). The Gray engine drawing is shown in Figure 5-14, his conversion tube is shown in Figure 5-15, and his circuit for the conversion tube is shown in Figure 5-16. As Lindemann explains, Gray also apparently succeeded in converting the negative energy into positive energy in his conversion element switching tube. Bedini's successful replication of the tube along with pertinent lab notes are shown in Figures 5-17, 5-18, 5-19, 5-20, and 5-21 by permission. Since Bedini is one of only two or three persons truly experienced in converting negative energy into positive energy, his successful replication is significant and decisive. Gray's process thus transformed negative energy into positive energy, allowing positive energy COP>1.0 self-powering systems. Bedini and the present author have filed a joint patent application filed on the process explained in this Chapter. We recognize Ed Gray as being a real pioneer in the area of COP>1.0 electrical power systems and in energy conversion. His pertinent patents are E. V. Gray, "Pulsed Cap. Electrical Discharge Engine," U.S. Patent #3,890,548, 1976; "Efficient Power Supply Suitable for Inductive Loads," U.S. Patent #4,595,975, **June 17, 1986.**

Aside from destruction of semiconductors etc.,[250] the overall effect is to increase the draw of power and current from the external source, so that the COP» 1.0 power system is thereby converted back to a COP<1.0 system!

The increased power draw due to the Dirac hole current can be remarkable. For an external battery power supply, if the hole current is sufficient to pass on into the battery itself, the battery may *charge* with negative energy while also powering the circuit! A remarkable phenomenon then takes place. A battery having been appreciably charged with negative energy may appear to be discharged (and so it is, with respect to positive EM energy). Yet, it may also continue to power a circuit. Sometimes it will, sometimes it won't, depending on the circuit. However, on connecting a normal positive energy battery charger to recharge it, the battery *precharged* significantly with negative energy will "eat" positive energy for an extended period, filling its localized Dirac Sea holes in its negative charge all the while, with the external voltage of the battery not changing at all. When all the negative energy holes are filled in the internal vacuum environment of the battery, and the negative energy charge is thus removed, suddenly the battery will start recharging with the input positive energy in a normal manner.

The batteries in Watson's generator (again see Figure 5-12) exhibited that phenomenon. Sometimes in recharging one of his batteries that had been extensively used, the battery would "eat" positive energy from the charger for a week or more before finally filling up its negative energy charge and then recharging normally from that point on.

The Dirac 4-hole current is nature's decay mechanism for the COP»1.0 system, sharply bringing it back to COP<1.0. Any successful COP»1.0 EM system must contain a mechanism for defeating (or transducing and using) this Dirac sea vacuum 4-hole current into its input section and beyond. We shall cover that subject shortly. The 4-hole current decay mechanism can *sometimes* become significant for semiconductors in the lattice 3-hole stage at COP = 5 or more, though the exact figure varies as the type and size of system and the semiconductor type. At COP > 10 or so, some systems will frequently begin to ravage themselves back into equilibrium with the vacuum, and destroy their overunity performance, unless the 4-hole current phenomenon is dealt with.

---

[250] As an example, novel effects may also happen in resistors, capacitors, and coils, and the effects depend on both the materials and construction of these components. A resistor can act as a true negative resistor and cool rather than heat, a capacitor can have its current lagging its voltage, and a coil can have its current leading its voltage. Mixtures of all three of these effects can happen in each of the three components, since each usually has some aspects of resistance, capacitance, and inductance. The emfs and back emfs in the circuit can also be altered and disrupted in a manner varying with the individual circuit and the individual operating conditions. Nonlinear oscillations and discharges of negative energy are particularly difficult to foresee and understand. Effects in dielectrics may vary widely, as can effects in conductors because of their internal impurities, etc.

Let us recap. We have assumed — to first order — a sweeping generalization that COP<1.0 systems exhibit an overall average spacetime curvature across the system (from input section to output section and load) such that excess positive energy electrons exist in the Dirac sea, without any appreciable accompanying excess Dirac sea holes interacting in the circuitry itself.[251] The slight excess of Dirac sea holes that are generated is "swept" forward out of the output section, and "eaten by the vacuum interaction" (particularly by the vacuum fluctuations) immediately upon exit from the output section. In that case, we do not have to concern ourselves particularly with Dirac Sea 4-holes and Dirac sea 4-hole currents, or at least they are usually negligible in most cases.[252]

For COP»1.0 systems, however, the overall average spacetime curvature across the system — from output section and load to input section — is such that Dirac sea 4-holes and their currents do independently continue to exist and independently flow in the local vacuum occupied by the physical system. Specifically, we shall have to consider the *tempic* effects upon such holes and hole currents, which are oriented from the output section back through the system to the input section.[253] In the COP»1.0, the effect of the increased tempic force is to sweep a Dirac sea 4-hole current back through the system from output to input, and even on out into the conductors from the external power supply, and on further out into the external power supply itself. The 4-hole current continues its "reverse feedback into the powering end of things" until quenched by filling it with incoming electrons. This conversion of 4-hole current to filled Dirac sea 4-holes by induction of lattice holes also squelches the concomitant negative EM energy fields, potentials, and waves associated with Dirac 4-holes as sources.[254] The reason is that a filled Dirac hole is a dual system of two opposite charges, hence produces counteracting positive and negative energy fields.

---

[251] In short, the requisite number of Dirac holes have interacted with and do interact with the lattices, converting into the normal "lattice hole" type of positron connected with positive mass, positive energy, but slowly migrating around the circuit in the opposite direction from the electrons. Here the true Dirac hole current exists only in short "hops" from atom to atom as the lattice holes are steadily filled by the "hopping" electrons.

[252] As stated, a kind of "hole" phenomena still exists in the circuit, but now as lattice holes and currents. Since these are attached to ions each having a mass much greater than that of an electron, the lattice hole currents may usually be neglected except in semiconductors, where they are taken into account in the very design of the semiconductor component itself.

[253] When we state a concept such as "negative mass" or "negative energy", we have conceptualized it *as if observed.* Hence, unwittingly we have made it an instantaneous, but iterated, concept. Rigorously, no such thing as mass or energy

The "4-hole-eating and filling process" is therefore the exact decay mechanism that squelches the natural electrogravitation associated with Maxwellian circuits which use Dirac 4-holes and 4-hole currents, negative energy fields, negative energy potentials, etc. When the circuit converts its free Dirac 4-holes into lattice holes, all electrogravity capability is lost, insofar as practicality is concerned.

For electrical power system purposes, the external environment of the COP»1.0 system furnishes an extra "electrical load" demand upon the external power supply, so that the extended system consisting of the (external power supply + power system) decays back into overall COP<1.0 condition and operation. For this reason, any appreciable involvement of Dirac 4-hole currents in a COP>1.0 EM power system must be *transduced* into positive energy electron currents, which requires just a special form of regauging. One need not confine regauging and gauge freedom to positive energy potentials and positive energy fields alone.

From the vacuum consideration and particularly from Dirac's theory of the electron, in COP>1.0 EM systems and circuits we shall consider that the Dirac sea 4-holes involve negative energy states (negative mass), and that positive energy states involve positive mass. In our approach, we shall consider a negative energy state (Dirac 4-hole) in the vacuum as having

independently exists except at one instant in time. Instead, mass exists in a continual iteration of *m* —> *mt* —> *m* —> *mt...* and energy exists in a continual iteration of *E* —> Et—»E—» *Et....* and so on. So during the masstime or energy-time (action) state, time and either mass or energy are intimately interactive. See our previous discussion in this book of the mechanism for the flow of a mass through time.

[254] Obviously that kills the antigravity aspects, because it kills the negative energy fields that would give negative curvature of local spacetime.

negative mass, which means that antigravity can be involved as well.[255] In this treatise, we consider only electron states, thus only negative energy and positive energy electron states.

9.3 Primarily Concerned With Dirac Sea Holes and Not Lattice Holes

We also stress the dramatic difference between a Dirac sea "electron 4-hole" in the vacuum (prior to observation) as compared to a lattice 3-hole in a semiconductor or other material component (after interaction and therefore after observation). In the case of a lattice 3-hole, usually the "positron" (hole) is attached to an ion, whose mass (i) is very much greater than the mass of an electron, and (ii) is also positive. In other words, a lattice hole is just a part of an ordinary positive ion. So the theory of "lattice holes" and "lattice positrons" does not primarily apply to the situation we are discussing. On the other hands, semiconductors exposed to any significant currents of pure Dirac sea 4-holes in the vacuum usually react and are nearly instantly destroyed because of the immediate disruption of their donor-acceptor arrangement and functioning. Interaction (filling the incoming Dirac 4-holes with contributed electrons) causes sudden mass disarrangements in their atoms and nuclei and lattice bonds. Thus, in the presence of Dirac sea 4-holes and 4-hole currents, usually any interacting solid-state electronics of the circuit or system will be dramatically affected. For switching in the presence of holes and hole currents, in most cases one may have to turn away from semiconductors and utilize mechanical or electro-mechanical switching, although this requirement varies with the type of semiconductor, the COP, the system geometrical layout, and the subsystem operations and layouts. Sometimes semiconductors may still be used if they are "brutes", with current handling capabilities well beyond what is "normally expected" in that circuit.

---

[255] We assume that the important thing about mass is its localized energy. Hence we associate positive mass with positive localized energy (with respect to the ambient energy density of the massless vacuum) and we associate negative mass with negative localized energy (with respect to the ambient energy density of the massless vacuum). In this view, the empty Dirac sea hole has both negative energy and negative mass. Note that this differs from the conventional assumption, which unwittingly addresses the positron as having been observed, producing a directional reversal. The rigorous test of our thesis is that concentrations of Dirac sea holes *must* generate antigravity *potential*, while currents of Dirac sea holes must generate antigravity *force*. The conventional view, assigning positive mass to the positron after observation, would produce gravity, not antigravity, but the force would be reversed by the reversed parity after observation. We argue that antigravity has been hidden and lost by its parity reversal in the assuming positive mass for the positron *as if that were the mass of the positron before it is observed.* In an overunity system with powerful steady-state Dirac 4-hole currents, the 4-hole currents can be differentiated by the cooling (negentropy) effect of the COP»1.0 system—when shorted to produce a "current surge" — rather than heating (entropy) effect. The magnitude of the hole current (which determines the average presence of the hole density) can be differentiated by a measurable weight loss in the COP»1.0 system. The successful antigravity production of the Sweet COP»1.0 device — discussed later in this chapter — experimentally verified these hypotheses.

9.4 Peculiarities of Spacetime Curvature and Dirac Hole Interactions

When considering the interaction between *curved* spacetime and the Dirac vacuum, immediately we violate the premise that all electron states in the Dirac Sea are always filled (which condition we surmise — along with Dirac — only exists on the average in a flat spacetime region). In a curved spacetime, we can have sustained excess 4-holes in the vacuum, or we can have holes that accept an electron falling into them *without emitting radiation.* The radiation-free "pair annihilation" reaction is possible when the spacetime simultaneously changes its curvature appropriately and thereby "absorbs" and holds the excess EM energy that would otherwise be emitted if the spacetime remained flat or decayed back to the flat state.[256] So long as that "energy-storing ST curvature" remains, there is no photon energy emission. If the "energy storing ST curvature set" has a net curvature instead of being flat, and then decays to a flat spacetime, then the previously retained energy is emitted as photon energy.

In that case the reaction provides the conventional emitted pair annihilation energy. In *that* case, the emitted energy may also transform into emitted particles in reactions well known in particle physics.

---

[256] Eerily, this interaction represents a situation where opposite reaction forces (the vacuum upon spacetime, and spacetime upon the vacuum) can create a special kind of stress in the system and changes in that stress. This leads to a peculiar kind of stress waves (scalar waves — longitudinal EM waves, so to speak) that are not covered in circuit theory anywhere. A symptom is that these stress waves react with components, amplifiers, preamps, etc. to create electrostatic waves. The effect of these waves is seen as the "crawling" of electrostatics all over and in the electrical equipment that is in one's laboratory. The effect can readily damage oscilloscopes, etc.

The normal emission of photons by pair annihilation need not occur, if the excess decay energy is absorbed *and held* by the local spacetime changing and holding its energy density and therefore (i) changing its curvature and (ii) stabilizing in that new excited state instead of decaying back to a flat spacetime.[257] Under such "non-emitting" pair recombination, a Dirac Sea 4- hole may simply "eat" a free 4-electron, and instead of emitting radiation the process simply affects the curvature of spacetime (and the state of the local Dirac sea) and changes it accordingly. A new "equilibrium state" is therefore reached in the supersystem between the opposite forces of the vacuum acting on spacetime and spacetime acting on the vacuum.[258] In other words, we consider change in the curvature of spacetime to exhibit EM energy *source* (emission) or *sink* (absorption) capability, with respect to the nominal ambient flat spacetime and with respect to conventional EM interactions.[259]

Or simply put, the energy emitted by the combination of positive and negative energy in the vacuum may be "locally radiated", "locally absorbed", or "locally transferred" to the other local component of the supersystem: the local curvature of spacetime. *It is all spacetime curvature dynamics.* By definition, a flat local spacetime consists of a vacuum with all its Dirac sea holes filled. Therefore, if "filling a 4-hole" results merely in relaxation of a set of local ST curvatures back to a flat spacetime, no EM radiation need be emitted, contrary to the usual pair annihilation phenomenon where additional energy (curvature of spacetime) is involved.

---

[257] Our view is that in normal pair annihilation this curving of spacetime to absorb the emission energy occurs first, and then the new ST curvature decays by what is called "emission of photon energy". We regard the photon as a specialized ST curvature traveling at light speed. So what is emitted as a photon is in fact the formation of a locally structured spacetime curvature (an "engine") traveling at light speed. It is a propagating spacetime curvature "engine" or "set of engines". In the process we are describing, the emission of the photon is the decay of the locally formed ST curvature set. If that set does not decay, the locally formed ST curvature set remains and no photon is emitted.

[258] This new kind of equilibrium is extraordinarily important, particularly in chemistry, where it will allow the "locking" of impossible combinations of particles into strangely stable "impossible molecules" with ongoing "impossible chemical reactions". The further discussion of the startling new chemistry that emerges is well beyond the scope of this paper, but such chemical states are included in nonequilibrium thermodynamics as *stationary states* — e.g., see Kondepudi and Prigogine, *Modern Thermodynamics,* Wiley, 1998, Chapter 17. At least one company has already succeeded in initiating this startling new kind of chemistry, with several patents granted and several more in process. Similar new "impossible stabilized excited states" and "impossible nuclear reactions" also occur in nuclear physics.

[259] The classical EM assumption of the flat local spacetime is in fact an assumption that the decay of the intermediate "ST curvature absorber/sink" occurs instantly.

Similarly, in a ST curvature, there can conceivably be electrons without "holes". If the curvature is negative, there *will* be unfilled Dirac sea holes present. Of course this is just a normal positive energy 4-electron, where the flat spacetime assumption in classical electrodynamics is in error. Instead, the local spacetime is positively curved, producing positive gravity. Whenever the energy density of spacetime-vacuum interaction changes from the "ambient vacuum and flat spacetime" condition, either positively or negatively, that constitutes an appropriate curvature of spacetime as well as a change in the vacuum dynamics. Therefore, in classical electrodynamics the assumption of a normal electron (or any other charge or mass) existing in an uncurved spacetime and an unaltered vacuum is a non sequitur. The local spacetime has in fact "curved" to an extent of holding the energy of one electron mass.[260] Further, the Dirac Sea has also changed to contain one unfilled and unobserved Dirac 4-hole.

With these things in mind, let us now introduce a conceptual model for dealing with these novel 4-hole phenomena that are experienced in COP>1.0 systems — and that have been the bane of many a would-be energy researcher attempting to develop a system extracting its EM energy from the vacuum and using it to also power the system itself.

9.5 Charging and Discharging a Capacitor

In electronic circuits, we usually think of charging a capacitor with electrons, and then discharging the capacitor to emit electrons again. However, we can charge the same capacitor with Dirac sea holes, or with a combination of Dirac sea holes and electrons. Then we can switch the capacitor adroitly and force it to give us back electrons. This "regauging" effect can thus be used to transduce negative energy current into positive energy current. So let us examine the charging and discharging process more closely. This is part of the Bedini process for handling the 4-hole current problem in overunity EM systems, to allow close-looping a system into self-powering operation.

---

[260] Mostly physicists think only of the *spatial* energy change. However, the time domain of spacetime also acts as a sink or source. Essentially one kilogram of mass only has the energy of 1 second of time. Hence the minuscule *mass* of the electron results in such a minuscule "sink" in the time domain that it has been inadvertently ignored. However, during that fraction of time where the electron exists as "mass time" and "charge time", it has considerable time energy, and *that* curvature of local spacetime cannot be neglected with respect to supersystem interactions.

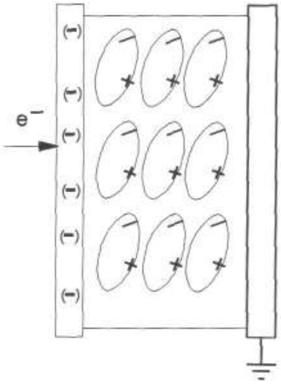

Figure 9-6 Normal strain in a dielectric capacitor that charges with electron current

Figure 9-6 shows the normal strain induced in a capacitor containing a dielectric between its plates. We charge the left plate with electrons, which repel the negative ends of the molecules in the dielectric, straining them to the right.

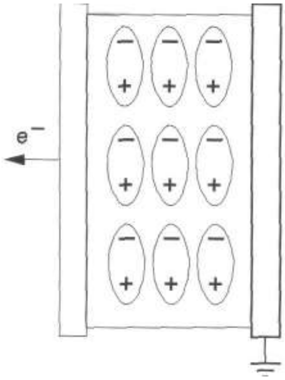

Figure 9-7 Strain relief in a charged dielectric capacitor that discharges.

Figure 9-7 shows the relief of this dielectric strain, by allowing the piled electrons on the left plate to be pushed back into the circuit from the left plate. Note that the electrons emitted in the relaxation of the dielectric strain pass back into the circuit from the same plate that was charged. This is a special kind of "reflection" symmetry. For a perfect capacitor, the total

Vq joules stored in the charged capacitor and drawn from the circuit, is returned to the circuit.

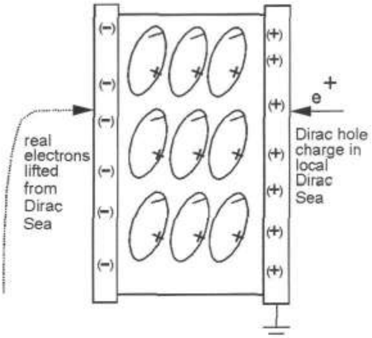
(−)
(+)
real electrons lifted from Dirac Sea
e^+
Dirac hole charge in local Dirac Sea

Figure 9-8 Strain in a dielectric capacitor supersystem that charges by Dirac hole current and lifted Dirac electrons

Figure 9-8 shows a remarkable thing when we charge the capacitor with Dirac sea holes. Suppose we charge the rightmost plate of the same capacitor used before, while the leftmost plate is temporarily disconnected from the circuit (or connected to another capacitor). Now Dirac Sea holes pile up in the local vacuum connected with the right plate. Whether or not they interact with the right plate, these positively charged holes attract the negative ends of the dielectric molecules, again straining the top of the dielectric molecules to the right. The strain of the dielectric is a force interacting upon the local vacuum, and it lifts real electrons from the Dirac Sea (or pulls them from the circuit or other attached capacitor) and onto the leftmost capacitor plate. The charged capacitor *in its supersystem* thus has additional functions that occur. The total *Vq* joules stored in the charged capacitor, were this time drawn (converted) from the applied Dirac hole current and the resulting interaction upon the vacuum, to lift real electrons onto the leftmost plate.[261]

---

[261] The reader should recall that a force is something automatically produced by nature to restore symmetry to a broken symmetry condition. A "4-hole negative energy force in the Dirac Sea" produced on one plate of the capacitor will evoke a "positive energy force outside the Dirac Sea" on the other plate as an attempt to restore symmetry. To produce this "positive energy force", positive energy electrons are lifted from the local Dirac Sea of that plate.

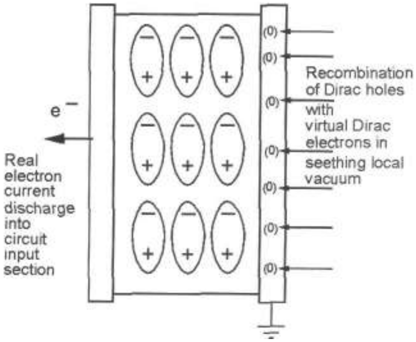
e⁻
Real electron current discharge into circuit input section
(0)
Recombination of Dirac holes with virtual Dirac electrons in seething local vacuum

Figure 9-9 Strain relief in a dielectric supersystem that discharges after charging by Dirac hole current and lifted Dirac electrons.

Figure 9-9 shows the strain relief of the charged capacitor in Figure 9-8. First the charged capacitor is re-connected to the circuit or element to which one desires to provide with electron current. The strained dielectric relaxes, and electrons in the leftmost plate are returned to the circuit. The Dirac holes in the rightmost plate discharge back into the local vacuum (a supersystem interaction) and are "eaten" immediately by the readjustment of local curved spacetime.[262]

The result is that the Dirac sea 4-hole current arriving in the input section has been intercepted by a waiting capacitor, and has been effectively transposed into normal electron current that is furnished back to the input section from the capacitor. This additional electron current helps power the circuit and reduces the power drawn from the external power source. The 4-hole current feeding this capacitor, together with transduction of the 4-hole current into electron current, effectively comprises a true negative resistance process. In other words, energy in unusable (in this case, detrimental) form is received, transformed into energy in usable (beneficial) form, and that usable energy is discharged back into the circuit input to assist in powering the system.

Transducing negative energy into positive energy — and vice versa — in the system does not violate the conservation of energy law in the supersystem, when most generally stated. With such regauging (which involves supersystem reactions), it is the *absolute value* of the energy that is conserved in the system, not its algebraic sign. Regauging (by gauge freedom) is after all an effect or condition locally imposed, but with appropriate energy exchange with the rest of the external universe via spacetime curvature dynamics exchange and vacuum dynamics exchange. The algebraic sign *is* conserved in the universal supersystem by the additional compensating reactions occurring in the other two components of the universe's supersystem.

---

[262] From experiments, it seems that Dirac holes forming at the "output" section of a **COP<1.0** device or system are almost instantly filled with electrons in the violent fluctuations of the zero-point energy. When forming at the "input" section of a **COP>1.0** device, this is not the case, and usually the holes can only be filled by incoming electrons in the power input, unless special provisions are taken to convert **the** arriving hole current (in the input section) into electron current.

The reader should strongly realize that energy is only required to be conserved in the universe's supersystem, not the system component alone. Again, this is an open system far from equilibrium with its active environment. In that case, energy is conserved in the ensemble of system and environment, but not necessarily in the system itself. This is no stranger than a common windmill or waterwheel.

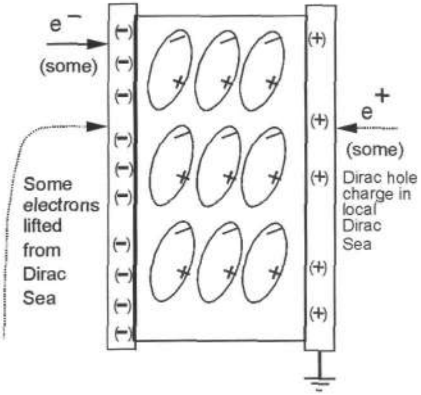
e⁻
(some)
Some electrons lifted from Dirac Sea
(−)
(+)
e⁺
(some)
Dirac hole charge in local Dirac Sea

Figure 9-10 Strain in dielectric in a capacitor supersystem that charges with a mix of electron current, Dirac hole current, and lifted Dirac electrons.

Figure 9- 10 shows a "mix-charging" variant of the 4-hole current charging of a capacitor, in which some real electrons are fed from the circuit to the left plate. Also, the Dirac sea hole current arriving in the input section is intercepted to charge the right plate, and some electrons are also freely lifted from the local Dirac Sea onto the left plate. Figure 9-11 shows the discharge of this type of charging of the capacitor, and the relief of the strain in the dielectric.

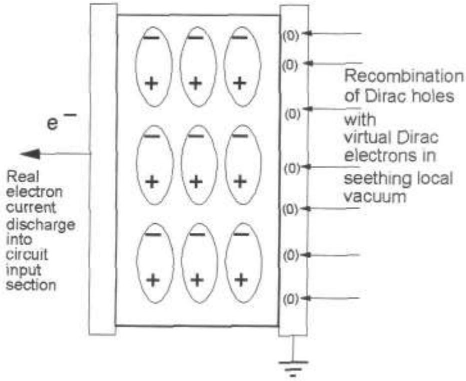
e⁻
Real electron current discharge into circuit input section
(0)
(0)
(0)
(0)
(0)
(0)
(0)
Recombination of Dirac holes with virtual Dirac electrons in seething local vacuum

Figure 9-11 **Strain relief in a dielectric that discharges after charging of capacitor supersystem with mix of electron current, Dirac hole current, and lifted Dirac electrons.**

In Figure 9-11, for example, we accent that the holes on the right plate are maintained only so long as there is a net strain on the dielectric and a net force across the capacitor. When the capacitor is discharged, the holes are discharged into the local vacuum in such a way that they now instantly recombine with virtual Dirac electrons in the vacuum fluctuations.

There are other ways and "tricks" of transducing the Dirac sea hole 4-current into electron current, but these methods are still proprietary to Bedini and will not be included in this presentation. We accent that all such "tricks" are actually methods of re-engineering the other two components of the supersystem — the active local vacuum and the local curvature of spacetime. This is using and applying what we have referred to as "engines".

9.6 The Open-Loop COP>1.0 EM Power System

Figure 9-12 shows the use of the transduced Dirac hole current as real input energy to "catch and lock" a COP>1.0 system in its overunity condition and stabilize its operating point for stable self-powering of the system and its load. The COP>1.0 operation reverses the normal tempic potential across the system, so that Dirac hole 4-current arrives in the input section, preparing to "decay" the system back to COP<1.0 condition.

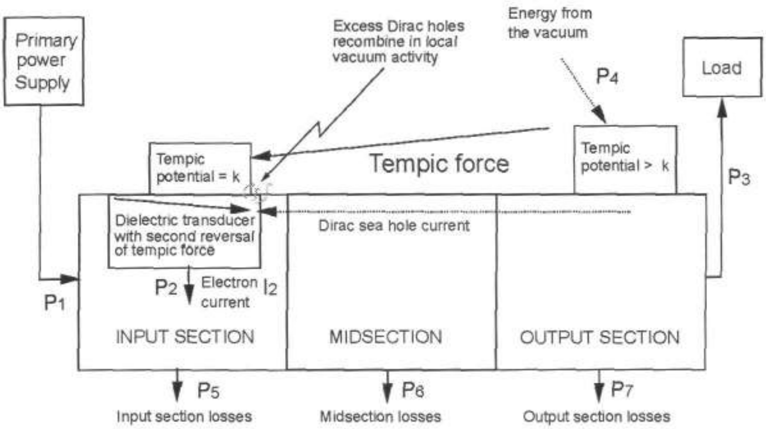
Primary power Supply
Excess Dirac holes recombine in local vacuum activity
Energy from the vacuum
P4
Load
Tempic potential = k
Tempic force
Tempic potential > k
P3
Dielectric transducer with second reversal of tempic force
Dirac sea hole current
P2
Electron current
I2
P1
INPUT SECTION
MIDSECTION
OUTPUT SECTION
P5
Input section losses
P6
Midsection losses
P7
Output section losses

Figure 9-12 Open-loop use of transduced Dirac hole current energy as real input energy to reduce power input required from the primary power supply.

However, the arriving hole current is intercepted by an adroitly switched dielectric transducer, which again reverses the tempic force. As explained above, the capacitor is charged with 4-hole currents (or a mix) and then discharges electron currents into the system — in this case, into the input section itself— by adroitly engineering the local vacuum state and the local curvature of spacetime.

The input section does not care where it receives its electron current from; it merely needs sufficient for its powering needs. Hence, the input section happily accepts this transduced electron current to partially power the system. This in turn reduces the "system load" that the external primary power supply "sees". The primary power supply thus reduces the current and power that it is furnishing to the system. In short, now the external power supply sees the combination of the normal system load and a "negative resistor (negative load)" represented by the transduced current and power. Its required input power is reduced accordingly.

9.6.1 Close-looped System: A True Negative Resistor

Figure 9-13 shows the use of the same process given in subparagraph 9.6 above and in Figure 9-12, except that now there is additional 4-hole current received and available, for transduction into normal electron current into the input section of the system. Sensors and governing are added to this operational mode, so that the system may *adjust and regulate its self-furnished electron current* to meet the total needs of its losses and its load. The system thus adapts the transduction of Dirac sea hole current: to fit its needs and is capable of regulated self-powering operation.

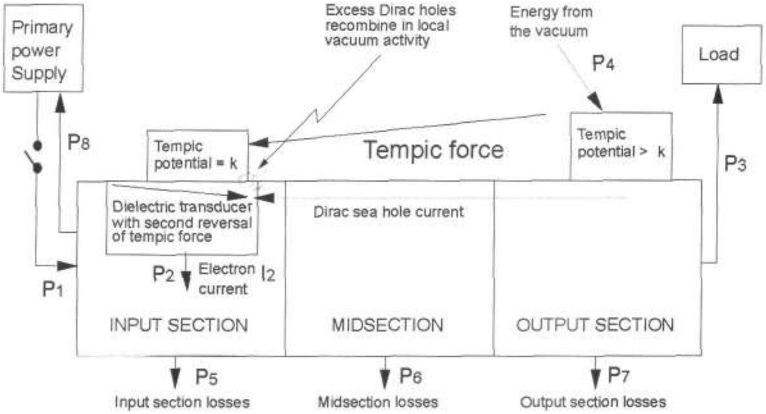
Primary power Supply
Excess Dirac holes recombine in local vacuum activity
Energy from the vacuum
P4
Load
P8
Tempic potential = k
Tempic force
Tempic potential > k
P3
Dielectric transducer with second reversal of tempic force
Dirac sea hole current
P2
Electron current
I2
P1
INPUT SECTION
MIDSECTION
OUTPUT SECTION
P5
Input section losses
P6
Midsection losses
P7
Output section losses

Note: Power outputs P3 and P8, and current I2 have sensors reporting to the Governor in the output section. The governor adjusts the potential energy in the output section to maintain steady output or adjustment of output to demand.

Figure 9-13 Close-looping the C0P>1,0 system with transduced hole current.

The system is first started in open-loop mode. The governor then gradually increases the transduction of 4-hole current into electron current, and the supply of that transduced electron current to its own input. This steadily reduces the "current draw" from the external primary power supply. When the "external current draw" reduces to zero (the Kron condition), the system "locks" into that position, with automatic governing to adjust the transduced electron current as the loads and losses vary. At that point, the external power supply is disconnected automatically.

This "locked and governed" system is now a *self-powered system,* freely taking its energy from the active vacuum and local curvatures of spacetime, to power both itself and its load. So far as we are aware, this is the first time that the technical means of locking and stabilizing COP>1.0 operation in an EM system in an open loop and closed-loop operation has been given.

If sufficient Dirac sea 4-hole current is available and transduced, excess electron current can be available at the system input section, even when the system is fully self-powering. In this case, the additional current can be fed back to the power line, so that the self-powering system not only powers its own load but also helps power the power network. Or additional load can be added and powered.

We point out that Gabriel Kron apparently had a solution to these problems in the late 1930s, and that solution has been suppressed to the present day. We give the following quotations to show Kron's knowledge of these things.

Quoting Lynn and Russell {565}:

> *"Kron has never published details of his method of making the polyhedron self-organizing, although his published results show that in this state it has some remarkable properties, associated with harmonic integrals on multiply connected spaces."*

In reference to Lynn and Russell's remarks, we point out that the Dirac sea 4-hole current from the output to the input is a special case of Kron's more general discovery of the multiply-connected open path existing through space between any two points in a circuit that are at differing potentials.

Quoting Kron {566}:

> *"...the missing concept of "open-paths" (the dual of "closed-paths") was discovered, in which currents could be made to flow in branches that lie between any set of two nodes. (Previously —following Maxwell - engineers tied all of their open-paths to a single datum-point, the 'ground'). That discovery of open-paths established a second rectangular transformation matrix... which created 'lamellar' currents... " "A network with the simultaneous presence of both closed and open paths was the answer to the author's years-long search."*

Again quoting Kron {567}:

> *"Although negative resistances are available for use with a network analyzer, in practice it is more convenient to use a second type of circuit, in which the positive and negative resistances are replaced by inductors and capacitors and the dc currents and voltages are replaced by ac currents and voltages of* *fixed* *frequency. The use of the second type of interpretation is equivalent to multiplying the wave equation by* $i = \sqrt{-1}$*."*

The reader should particularly note that in the first part of the sentence Kron stated unequivocally that negative resistances were available for **use** on the network analyzer, a large analogue simulation used by the Navy **and** administered at Stanford University with the contractor being General Electric. The censors missed that one.

Again quoting Kron {568}:

> *"When only positive and negative real numbers exist, it is customary to replace a positive resistance by an inductance and a negative resistance by a capacitor (since none or only a few negative resistances exist on practical network analyzers).*

We believe the censor caught this one, and forced Kron to add the words, "none or". Kron had at least three negative resistors, and possibly more.

Again quoting Kron {569}:

> *"Now a value E of the negative resistances, at which the generator current becomes zero, represents a state at which the circuit is self-supporting and has a continuous existence of its own without the presence of the generator, as the negative resistances just supply the energy consumed by the positive resistances. ... When the generator current is positive the circuit draws energy from the source, and when the current is negative the circuit pumps hack energy into the source. At zero generator current the circuit neither gives nor takes energy, and theoretically the generator may be removed. "*

Here we believe the censor forced Kron to add the word "theoretically". In this quotation Kron reveals that he was well aware of the process for close-looping a system — starting it in open-loop mode and bringing the negative resistor currents up to match the system needs nnd reduce the generator current to zero, then disconnect the generator. He also was quite aware that excess transduced energy from the negative resistor could be available at the input, in which case energy could be furnished back to the power network while the system powered itself and its load simultaneously.

We also point out that Dirac's definitive electron theory — covering the Dirac Sea and Dirac sea holes — was published in 1930, the year the present author was born. In the late '30s and in the 40's Kron certainly was well aware of that work. In addition, Kron applied full general relativity to electrical circuits and machines, and he was far ahead of his time. His work is also still far ahead of the present time, at least in the open scientific community. He also was aware of developments in particle physics, and so he did have the gist of the supersystem concept in his mind {570}, however he may have thought of it. His concept of the "open path" effectively includes the effects upon the system from the other two components of the supersystem, so it appears that Kron's "system plus its open path" roughly corresponds to our concept of the "supersystem."

Further, Kron was a mentor to Floyd Sweet, who also worked at General Electric but in a different project. Sweet often spoke glowingly of Gabriel Kron.[263] Sweet almost certainly knew the secret of Kron's negative resistor. It is the present author's opinion that Sweet's vacuum triode amplifier was derived from Kron's negative resistance unit. With both deceased, we will never know for certain.

9.6.2 Suppression of COP>1.0 Systems Should be Investigated

During our decades of research, we ran across several cases where U.S. scientists actually produced working COP>1.0 systems.[264] None of these overunity systems was ever allowed to go forward with independent development. This is an area that cries out for a full and vigorous Congressional investigation, and appropriate legal action should be taken by the U.S. government against the suppressing parties, whomever they might be {571}. Simply put, the U.S. government, the U.S. taxpayer, and the scientific community have been ripped off and deluded in the field of electrical energy for more than half a century. That continues to the present, as witness the disappearance of the Watson generator (accompanied by Watson and family breaking off all contact with their close colleagues) and the suppression of the Kawai device and process. The latter suppression was accomplished in 1996 in the personal presence of the present author and the members of the Board of Directors of our company, CTEC, Inc. An accidental overunity frequency converter was even initially placed in the original Minuteman missile, but was altered back to COP<1.0 operation when the electronics past the converter kept failing because of overload power furnished by the converter.

---

[263] The present author worked with Sweet off and on for a decade or more, and designed the antigravity experiment that Sweet performed.

[264] These incidents were previously exposed in a draft book, *Matters Arising,* carried on the Internet for some months and then withdrawn. Some things must wait for their appointed time, and so must certain kinds of information.

9.7 Remarks on Dirac Sea Holes, Currents, and Negative Energy

The subject of the Dirac Sea, Dirac holes, Dirac hole currents, and negative energy still has elements of strong disagreement within the scientific community and among leading theorists. Often the "conclusion" one reaches depends on the way the starting assumptions are made and then on how the mathematics is applied. There is no definitive experiment establishing that negative energy cannot exist; to the contrary, many things do occupy negative energy states. The binding energy of the nucleus, e.g., involves a loss of mass (loss of positive energy) and hence involves negative energy states, as is well known. Here the "negative mass" effect is immediately and experimentally verifiable, and is recognized. The total mass of the bound nucleons is less than the total mass of those same nucleons when unbound, and this fact is widely known and accepted.

A single white crow is sufficient to prove that not all crows are black. Hence a single experimental proof of mass-energy "loss" and negative energy states establishes that negative energy and negative mass are real. Indeed, every photon absorption by mass and every photon emission by masstime prove the same point universally: mass is a variable. The same assembly of protons and neutrons in a nucleus, having lost mass, also has "lost" some of the gravity that the positive masses of those unbound protons and neutrons would otherwise produce. Therefore, there is an appropriate amount of "negative gravity" already acting in conjunction with positive gravity, in every atomic nucleus. It is associated with the negative energy states of the bound nucleons in every nucleus.

Consistent with the initial quotations and comments at the beginning of this chapter, we present an interpretation by which we can understand some of the major novel phenomena that arise in laboratory experiments with COP>1.0 electromagnetic systems, including COP»1.0 systems.

We propose that COP»1.0 EM interactions are widespread in highly energetic astronomical interactions {572} and even in highly energetic phenomena on Earth, when the supersystem changes are also accounted.[265] We argue that novel negative energy phenomenology — such as encountered in the COP>1.0 experiments by Magnetic Energy Ltd. in various versions of the COP>1.0 motionless electromagnetic generator (MEG) {573, 574}, by Bedini {575} in his induction of negative resistor actions inside the batteries of battery-powered COP>1.0 EM systems, and by Sweet and the present author in experiments with the Sweet COP»1.0 vacuum triode amplifier system {576} — is also generated in these COP»1.0 highly energetic astronomical interactions, yielding the same kind of negative energy and negative energy currents.

---

[265] As an example, in looking at sprites and other such sharp electrical discharge phenomena, we may be observing the prompt *decay* of COP»1.0 interactions which were self-initiated in the atmosphere.

This Dirac hole negative energy generation of COP»1.0 EM processes is proposed as the mechanism which directly produces the negative energy fields and potentials that generate the mysterious antigravity accelerating the expanding universe {577}. This is a reasonable assumption, even though it is against the received view, because it is based on some laboratory experiments, not just theory. It also is the unexpected effect of the cosmological constant introduced by Einstein {578}, and it is consistent with Zeldovich's demonstration that the energy density associated with Einstein's repulsive cosmological constant is — or is associated with — vacuum energy.[266] The received view *is not* based on direct experiments with COP>1.0 EM systems, but only on COP<1.0 EM systems. Until the received view is broader based, it does not adequately represent nature. Hence it does not represent a valid refutation of experimental negative energy systems producing antigravity processes.

In Bedini's experiments and in the MEG experiments, antigravity is essentially absent or so small as to be nearly nondetectable. However, these experiments generally have COP of 1.5 to 4.0, mostly, and sometimes 10 or 12 (we have experimented up to 20 or more stably, and unstably at nearly 100). Hence the Dirac hole current is relatively small in them, since the magnitude of the Dirac current is a function of (COP - 1).

Sweet's COP > 1,500,000 experiments produced a very large flow of negative energy back across the system, substantially curving the local spacetime and reducing the weight of the system on the laboratory bench. Figure 6-7 shows the first Sweet device, which produced six watts output in open-loop operation. Figure 6-8 shows a later Sweet device which produced COP = 1,500,000 and was pushed past that COP to produce the antigravity in the laboratory, smoothly reducing its weight by 90%. Figure 6-9 is a plot of the actual readings taken during the antigravity experiment. Sweet performed the experiment in California, reading off the measurements to me over the phone as he made each one.

---

[266] Believed to have been demonstrated in 1967. But check Ya. B. Zeldovich and A. A. Starobinsky, "Particle creation and vacuum polarization in an anisotropic gravitational field," *Zh. Eksp. Teor. Fiz.*, Vol. 61, 1971, p. 2161 *(Sov. Phys. - **JETP,*** Vol. 34, 1972, p. 1159); Ya. B. Zeldovich, "Vacuum theory — A possible solution to the singularity problem of cosmology," *Uspekhi Fizicheskikh Nauk,* vol. 133, Mar. 1981, p. 479-503. Translation in *Soviet Physics — Uspekhi,* vol. 24, Mar. 1981, p. 216-230.

Another indication of the negative energy effect present in the output section of the Sweet device was tested by suddenly shorting the output leads. This resulted in instant freezing of moisture from the air, so that the leads were instantly frozen and covered with ice. Contrast this to the normal heating of shorted wires, where the energy is positive.

Sweet never revealed his activation process, whereby he initiated very strong self-oscillation — of the nucleus of the barium atom in barium ferrite magnets — with the local vacuum treated as a semiconductor. One part was known to the present author and another part was known to Rosenthal, a professional test engineer. In addition, how Sweet converted his excessive hole current is not completely clear. It appears that he somehow used his coils in quadrature together with special circuitry to interact the excess hole current into his magnets, directly with the self-oscillating barium nuclei and thence with the other nuclei in his magnets. If so, it created the "rough equivalent" of increasing the associated binding energy offset and locked with concomitant additional local spacetime curvature.

If this speculation is correct, he "locked" the nuclei of his magnets to the local curved spacetime, maintaining the overall local binding energy in the nuclei themselves while producing a great excess of "binding energy" (negative energy) in the immediate vacuum. This would explain the antigravity, and would be consistent with the experimental results as well as the increase in antigravity as the output load of the device was deliberately increased.

We must use the supersystem concept, as in the quoted comment at the beginning of the chapter, to enable a better understanding of these novel phenomena. Together with the Dirac negative energy production and high COP system, the hypothesis of "additional negative energy acting on local spacetime immediate to each nucleus" allows the generation of the observed antigravity to be understood. However, we accent that this mechanism is still a hypothesis, and it must be further substantiated by experiment.

9.7.1 Lattice Holes

In a material conductor or other component of an electromagnetic circuit, the conventional model envisions currents in the Drude electrons as the electrons moving by "hopping" from molecules or atoms in the material lattices into the Drude gas, with each hop leaving behind a positively charged "vacated electron position" called a *lattice hole,* or just a "hole". Subsequent Drude electrons are attracted to fall into the holes, etc. This net movement of the hopping electrons down the conductor, etc. comprises the conventional electron current through the conductor. We point out that the actual movement of the electrons in their "hopping" is vastly more energetic than the very feeble net "drift" of the overall ensemble of Drude electrons longitudinally down the wire as current. The energy of the total electron activity ongoing in a simple 1-foot length of copper wire would power a city, if it could all be harnessed and utilized.

At the same time the electrons hop with a slow net progress down the wire, the "holes" in the material lattices are also collectively seen by the external observer as if "slowly changing their net positions" longitudinally in the opposite direction. One speaks of a corresponding "hole current" or "lattice hole current" in the opposite direction to the electron current. However, there are subtle differences in the flow of the two currents, because an electron in the Drude gas is far more kinetic than the positive ion left behind by an electron that has "hopped up into the electron gas". Nonetheless, the positive lattice ions do "rock" physically from the induced forces upon them, as the holes are opened and filled. In conventional electrical power systems, this equal and opposite energy reaction in the ions of the lattice materials, induced by the applied fields, is usually ignored. In most classical EM power system theory the holes and their "composite equal and opposite energy and work in the system" are also ignored, except in semiconductors and semiconductor circuits.

9.7.2 Newton's Third Law Is Accounted in the Supersystem

Because half the EM energy and work in the circuit is usually ignored (e.g., in the Maxwell-Heaviside-Lorentz standard equations), electrodynamics is said to be free from Newton's third law reaction. This may be a non sequitur, and it follows from the assumption that — for an EM wave, field, or potential in space — no interaction exists with the active vacuum or with the curvature of spacetime. In fact, those reactions do exist in the supersystem. The field, wave, and potential do indeed interact with both the quantum mechanical vacuum and the general relativistic spacetime.

Indeed, the field, wave, and potential may be said to *identically be* changes induced in general relativistic spacetime. The absence of the "transmitting" physical system leaves us with the remaining two components of the supersystem, each of which extends to infinity. Again, Figure 9-1 shows the actual interactions (the supersystem) and Figure 9-2 shows the erroneous and arbitrary "kill of the supersystem" by standard classical electrodynamics.

When higher group symmetry electrodynamics such as O(3) is applied, then the interactions between the three parts of the supersystem emerge and must be accounted. This is particularly true of COP>1.0 EM systems. However, as Finster points out {564}, the Dirac sea effects exist *wherever* there are EM fields, potentials, and waves. We add that curved spacetime effects also exist *wherever* there are EM fields, potentials, and waves. And wherever Dirac sea effects exist, or wherever curved spacetime exists, then there the local vacuum and local spacetime interact with each other.

However, as we shall point out, usually most of the Dirac sea hole effects can be neglected in COP<1.0 systems and in most lower gain COP>1.0 systems because of the time-energy aspects. For COP»1.0 EM systems, the time-energy aspects are partially reversed, and instead of carrying the hole effects away from the system, the time-force from system output to input forces the Dirac sea holes to flow back in the local vacuum across the system from its output to its input. This flow constitutes an additional "load" appearing in the input section and requiring additional electron current from the external power supply to fill the holes. If the power supply does not furnish sufficient electrons to fill the hole current at the input section of the system, then the unfilled hole current passes back into the feeder line toward the external power supply, and even into the distant power supply itself. In its progress, it eats incoming current all along the way, requiring that more current be furnished by the distant generators.

9.8 Remarks on the Observed Acceleration of the Expanding Universe

As a candidate for *dark negative energy*, the present author has proposed vacuum negative energy and currents in curved spacetime regions partially containing excess *empty* Dirac sea holes as well as normal *filled* Dirac sea holes. The process by which these excess Dirac holes are produced has been advanced, and is supported by at least one dramatic experiment. One result of such "negative energy areas" of negative spacetime curvature would be "jumps" increasing the velocity of light and high-energy particles traversing these altered regions of spacetime. In fact, there is some evidence for just such jumps, according to a hypothesis by Richard Lieu of the University of Alabama in Huntsville {580a, 580b}. Otherwise, high-energy particles produced in distant processes in the cosmos would not be detected on earth, but would be annihilated in highly energetic collisions.

The departure of large-scale spacetime from homogeneity would also result in some loss of correlation in the presently fitted red shift versus distance relationship. Again, there is tentative indication that such may prove to be the case. The most distant galaxy recently detected has a red shift of 6.56, which — assuming homogeneity of space — implies look-back of 14 billion years or so.[267] (conventional view) Best estimates of the age of the universe based on the Big Bang assumption is some 14.7 billion years. It would seem that 0.7 billion years is a very short time for that observed galaxy to have formed. If even more distant galaxies are observed with larger telescopes to become available, then by the redshift correlation to homogeneous space one would be observing galaxies "before the beginning of the universe" — an obvious non sequitur. It appears that the redshift correlation to velocity and distance, as well as the homogeneity of space and in fact the Big Bang itself— are already in some difficulties and may be in serious trouble in the near future.

The empty Dirac sea 4-holes in the vacuum are special "Dirac negative action quanta" and the holes and their dynamics form negative action potentials, currents, and fields. This type of vacuum condition containing some empty Dirac sea holes is referred to as a *Dirac-polarized vacuum.*

However, we do not envision the Dirac-polarized vacuum as just simple *static* polarization, but as possessing rich dynamics such as negative energy currents, potentials, fields, waves, etc. including in the time-energy medium in the time domain. We also include the interaction of the dynamic Dirac-polarized vacuum with spacetime.

---

[267] I am indebted to Prof. Stanley Jeffers for this information, in his E-mail discussions with the Alpha Foundation's Institute for Advanced Study fellows. Dr. David Roscoe contributed the fact that the Big Bang theory has so many parameters that nearly anything can be "fitted in". But he also pointed out that W. M. Napier, *Astronomy & Astrophysics,* Vol. 310, 1996, p. 353 gives a very thorough analysis of the quantized cz phenomenon, which does present a difficult obstacle to Big Bang theory. Napier is **the** originator of the cometary catastrophe theory of biological extinction. One way or another, the theory of the Big Bang is in trouble, even though many theorists are still complacent about it.

Due to ongoing COP»1.0 processes involved in highly energetic astronomical phenomena, Dirac sea 4-hole regions and dynamics are believed to occur and exist in the vacuum across the observable universe. These occur especially in appreciably curved spacetime regions with large tempic potential gradients and thus large *time-forces* across the region.[268] The 4-holes and 4-hole currents in the vacuum are generated in great quantity in large cosmic explosions. Also, the energetic fluctuations of the local vacuum energy can separate a Dirac hole from its occupying electron by lifting the electron (pair production), and so can a curved spacetime. In a properly curved spacetime, fluctuation can lift an electron from a hole, and enter a stationary state by accommodating reaction of the local spacetime curvature. In that case, hole and electron go their separate ways. The hole and that or another electron may later recombine, or in sufficient local spacetime curvature, they can remain apart. We point out that every positive energy electron is accompanied by an appropriate local curvature of spacetime that prevents the accoutrement of a free Dirac sea hole,[269] and thus prevents the electron from disappearing.

The more energetic the local region, the more severe and violent are local spacetime curvatures in that region, and — in our view — the more frequent the occurrence of COP»1.0 interactions producing significant Dirac 4-hole currents, negative energy EM fields and potentials, etc. Indeed, the measurements that yielded the accelerating expansion of the universe were measurements of the relative velocities existing with stellar explosions known as type la supernovae. In our view, in explosions of such violent nature — particularly including in optically active gases such as where gamma ray bursters form — an appreciable fraction of the hole currents are transduced into positive energy electron currents. This produces — at least for a finite time — a net positive feedback, self-powering COP»1.0 phenomenon, resulting in very rapid asymptotic output energy increase. This is what we have hypothesized as producing the gamma ray burster itself. *Quenching* of this asymptotic rise in COP occurs only after the physical geometry of the "closed-looping" gas particles is broken a short time later. The maximum output energy is determined by a function of the *rate* at which the geometry is broken and the corresponding *rate* at which damping of the asymptotic rise and quenching occurs. The afterglow remaining in the gas is due to the extra positive energy produced in it by the transduction of negative energy into positive energy during the asymptotic rise of the self-powering burster, with excess positive energy being absorbed in the gas while the burster magnitude is rising. After damping and quenching, the gas has been heated (that is how the geometry was disturbed in the first place, to produce damping and then quenching). Re-ignition can occur whenever a new close-looping COP»1.0 situation spontaneously occurs. As in the gamma burster, this can often happen as the disrupted geometry of the gases stabilizes.

---

[268] We also conjecture that these regions may account for a tentatively observed slight change in the fine structure constant, but leave that to the advanced theorists to validate or falsify.

[269] This needs a bit of modification. Rigorously, the so-called "isolated" observable electron is clustered around by *virtual* positrons (Dirac sea 4-holes) in the vacuum. So a positron influence is present, though from virtual positrons instead of observable positrons. The magnitude of both the positive charge and the internal or "bare" negative charge is infinite. The difference, however, is finite and is the "observed" magnitude of the observable central charge. There is thus a net observed shielding of the bare charge of the electron, and this is taken into account to yield the observed charge of the electron.

All 3-space EM energy comes from the time domain, as previously developed. In the more violent positive energy COP»1.0 regions, there is an increased conversion of time-polarized EM energy into 3-spatial energy and back. On the average there is a higher "time-energy conversion" or "tempic" potential. In those regions, a tempic conversion potential is higher at the higher center magnitude of the spatial energy density. Hence there are strong tempic forces from "inside to outside", providing a "tempic broom" effect to sweep outward the excess Dirac sea holes separated from their electrons in the curved spacetime. The result is the presence of increasingly unoccupied Dirac sea 4-holes toward the outside of violent regions of spacetime,[270] until the effect is overcome by the decreasing spacetime curvature. The 4-hole current radiating out from an asymptotically increasing (exploding) burster provides an asymptotically increasing (exploding) negative energy and negative mass reaction with spacetime. That reaction explosively creates curved spacetime for negative gravity (repulsive gravity), acting upon the local system and upon distant masses as well. The result appears to be the cause of the acceleration of the expanding universe.

It follows that re-ignition of the gamma burster can occur if, as the quenching continues, the geometry stabilizes and reaches another "close-looping" positive feedback condition. In fact, just such re-ignition of the gamma burster is very often observed.

---

[270] I.e., the violent astronomical phenomena — while generating and increasing are more violent at the center, and less violent at the outside.

Therefore, we hypothesize that the more distant violent astronomical regions have greater amounts of repulsive gravity being produced by their local Dirac-polarized vacua particularly during their explosive asymptotic energy density rise. This produces violent streams or bursts of negative energy outflow, which is equivalent to negative mass outflow and which provides outgoing negative curvature of spacetime (antigravity). At sufficient distance from us, the increasingly earlier, more violent events and resulting greater curved spacetime regions have created sufficient cumulated Dirac-polarized vacua to override the normal attractive gravity between galaxies, etc. and produce net repulsive gravity. This results in observed distant acceleration of the expanding universe {581a-581f}.

9.9 A Solution to the Hole Current Problem and Close-Looping

Bedini and the present researcher have filed a patent application {582} on the above special capacitive process to *convert* or *transduce* the detrimental Dirac sea hole current reaching the system's input section into electron current, and put it to use in powering the system. This transduction process receives the hole current as a special potential, then regauges the "negative energy" potential into a "positive energy" potential. This regauging is accomplished by charging the capacitor with Dirac hole current and then discharging electron current from it. The transformed positive energy potential is then applied to the Drude electrons, producing an ordinary electron current in the input that can be used to power the system itself. If sufficient excess positive electron current is provided, then the system can "feed itself in self-powering mode and also feed power back into the main power supply line.

When conversion of the arriving hole current in the input section produces an output "positive energy" electron current equal to the system's normal power needs (Kron's condition), the external power supply can be disconnected and the system will remain powered by the steadily converted negative hole current arriving in the input section.

Our patent-pending process allows this otherwise devastating special form of "negative energy feedback" to be converted into "positive energy feedback", allowing the system to be close-looped successfully for self-powering of the system and its load. Regulators of the transduction process may be added to cover the situation when the load is erratic and changing.

At Magnetic Energy Ltd., we have also filed a patent application for an outrigger method of close-looping a COP>1.0 system, which greatly eases the problem. This method is covered in Paragraph 9-14 below.

9.10 Negative Energy: Localized and Nonlocalized

It sometimes appears there are as many views on negative energy as there are scientists who mention it. Simply put, the Dirac theory of the electron requires negative energy states. The negative energy electron state was envisioned by Dirac as a sort of "hole" in the ambient vacuum, where the hole could contain a positive energy electron. The combination of equal positive energy of the electron and negative energy of the hole itself, would give a "normal" vacuum zero state of the "filled hole". In other words, the normal or ambient vacuum consists of holes filled with electrons. Considered in that fashion, such a vacuum is referred to as the **Dirac Sea** or the **ambient Dirac Sea,** or the **ambient Dirac vacuum.**

If positive energy is added to a filled Dirac hole, the electron can be lifted out, leaving the hole. Note that adding positive EM energy, e.g., is actually the addition of spacetime curvature and dynamics to the Dirac sea vacuum. In short, it is a supersystem interaction, not a system interaction.

In that case, the curvature of the local spacetime interacts upon the electron mass, producing a force that lifts it from the hole. The result is the appearance of a positive energy, positive mass electron and a Dirac sea hole (a 4-positron) in a curved spacetime. That is known as *pair production,* and it is a supersystem interaction. When a positron and an electron meet, the electron falls into the hole, changing the Dirac sea which also changes the local ST curvature. That change in local ST curvature is a change in the energy density of space, and is normally radiated away as a photon or photons as the local curvature decays back to a flat spacetime. The emitted radiation may further decay into particles, etc. by several reactions well known in particle physics. That "conventional" set of supersystem interactions is known as *pair annihilation.*

If we deliberately change the ST curvature appropriately, however, the pair annihilation radiation need not be emitted since the excited local spacetime curvature does not decay back to a flat spacetime to emit it.

Dirac's original paper tried to identify a negative energy state vacuum hole as the proton, but later that was falsified and it was taken as the positron. There is a very strong aversion in physics to considering negative mass; hence, the mass of the positron was proclaimed positive (for many reasons not really connected with our thesis). Nevertheless, some physicists (in the minority) still speculate or theorize that negative mass "can exist." Not only "can" it exist, but also it *does* exist widely and one can demonstrate it in actual electrical circuits. It certainly exists in the binding energy of the nucleons in the nucleus of an atom. Again, a single white crow is sufficient.

We take a very straightforward view: we consider a positron — an electron observed with positive mass and "observed" to be traveling backward in observer time — to be a different (transformed) entity than a Dirac 4-hole traveling forward in observer time.[271] The observation operator accomplishes both velocity reversal and time reversal, and — since time is not an observable, the time reversal is "seen" as a reversal in spatial direction. *Prior to observation, neither reversal has been imposed.* The major distinction, we hypothesize, is that the observed positron has positive mass in 3-space, which is concentrated positive energy, and therefore produces positive energy electrical fields and positive gravitational field. Mass is an observable, hence, it is the iterative frozen 3-space snapshots of an ongoing 4-space process. Observation (and any observable) is the effect of applying the operator $\partial/\partial t$ upon spacetime LLLT in which a 4-space process is occurring. As we stated earlier in this book, our view is that no observable *continuously persists* (in time), but continually and iteratively *recurs* by the continual ripping of time from the ongoing 4-space reaction, to see the 3-space intersection at a frozen moment. We also advanced the *mechanism* that generates the flow of a mass through time, consisting of a mechanism for $\partial/\partial t\,(\mathrm{LLLt})$ - photon emission, converting *mt* to *m* — and the mechanism for $\int_t (\mathrm{LLL})dt$ — photon absorption, converting *m* to *mt.*

---

[271] This point is stressed. The *unobserved* positron is not identically a positive mass electron traveling backwards in time, but a negative mass, negative energy electron traveling forward in time. When Dirac theory of the electron was written, broken symmetry and the CPT theorem had not been discovered. Further, there was no distinction made in the difference between the observed entity (the *effect* of the observation interaction) and the non-observed entity (the *cause* of the observation interaction).

[272] We stress that we are once again confronting the ubiquitous substitution of the effect for the cause in classical electrodynamics. The observed positron and the unobserved positron are two quite different entities —just as the "cause" of an observation is quite different from the "effect" of the observation.

We may also consider mass to be condensed 3-spatial energy (say, EM energy) compacted by c^2. Modern physicists tend to consider the mass-energy itself as the source of the gravitational "attraction of mass" rather than the mass itself {583}. Let us accept that assumption, but differentiate between the observed 3-positron with positive mass and opposite direction in positive observer time, and the unobserved Dirac sea 4-hole with negative mass and without change of direction in the observer's positive time. The 3-positron is an *observed effect,* while the unobserved Dirac sea 4-hole is an *as-yet-unobserved cause.*

We also surmised that all EM energy in 3-space comes from time, by our solution {584a} to the long-vexing source charge problem {68}. In our view, we may then consider that the positive energy compressed as 3-space mass has been extracted from the time domain of surrounding spacetime, via entry of that time-energy into the surrounding 3-space at each 3-space point, producing a surrounding spacetime curvature which is known as the *gravitational field.* This gravitational field represents "withdrawn" 3-space energy from surrounding space, and withdrawn time-energy from the time associated with each surrounding 3-space point.

Hence the gravitational field may be said to consist of "negative energy" since it represents *lowered* energy of spacetime due to the curvature. This viewpoint yields results consistent with the received view that the energy of the gravitational field is negative. In our view, it is negative precisely because it represents "withdrawn energy" from the normal energy density of that otherwise flat space. The withdrawn nonlocal energy has been localized (as what we call "mass"). Simultaneously that withdrawn 3-spatial energy also represents "withdrawn energy" from the time-domain, since all the withdrawn spatial energy came from the time domain in the first place. The result is the curvature change in spacetime, that is taken as the gravitational field.

The 3-space energy comprising the mass is positive because it is *added to* the normal energy of an otherwise flat spacetime, at that concentrated locality.

The "energy in the gravitational field" is negative because it represents lowered energy density (i.e., lowered "pressure" in a fluid analogy) of that surrounding nonlocal spacetime compared to when that spacetime was flat

Thus when we add positive energy (a positive energy state) to 3-space, we create normal "positive" gravity because "mass" is a process of withdrawing some positive energy from the energy density of surrounding 3-space.

Here is the shocker: When we add localized negative energy (a localized negative energy state or "4-hole" — or a localized negative mass!) to a local 3-space region, we create "negative" gravity (antigravity) in surrounding space because the withdrawn energy from the local 3-space "hole" must have been added to the energy density of the surrounding spacetime, once we apply conservation of energy. Or in other words, the surrounding spacetime changed (curved) in interaction to the production of the hole.

If we wish, for simplicity we can just consider localized positive energy to generate positive gravitational charge in the surrounding nonlocal spacetime. We can consider localized negative energy to generate negative gravitational charge in the surrounding nonlocal spacetime.

This is where our concepts differ from the rather arbitrary way that a Dirac sea hole is interpreted in the received view. In spacetime, we try to account for the entire energy process and we apply a conservation law by including the supersystem interaction in every case. We insist that positive energy concentrated in positive local mass must have been extracted from surrounding spacetime.[273] That "negative energy" in the surrounding spacetime is what is known as the *gravitational field of the source mass* (where the energy was withdrawn to). Similarly, we insist that positive energy extracted from a localized region of the ambient vacuum to leave a "Dirac sea hole" must have been spread out into the surrounding spacetime, where it would properly be known as the *antigravitational field of the source Dirac hole.*

As an analogy in very simple lay terms, the holes associated with a local mass system or region involved in COP>1.0 operations create higher "pressure" outside the system or region with respect to other objects. Higher pressure between any two objects pushes the objects apart. The masses associated with a mass system create lower "pressure" outside the system or object with respect to other objects. Lower pressure between any two objects results in the objects being pushed together by the higher pressures beyond them.

---

[273] Alternatively, if we add local spacetime curvature, the surrounding nonlocal spacetime interacts as mediated by the supersystem changes in all three components. What we are saying is that there is no such thing as just a system, vacuum, or spacetime change. Any change to one is accompanied by a change to the other two, *a priori.*

We submit that this new view provides an exact (and testable) antigravity mechanism. It appears to have been validated by the Sweet antigravity experiment {576}, which was designed in advance by the present author with increasing the local negative energy reaction in mind.

9.11 Tests Reinforcing the Antigravity Mechanism

9.11.1 Dirac Sea Hole Current Effects in a COP»1.0 EM System.

For COP»1.0 electromagnetic systems, we recall that the temporal energy will flow negatively across the system from output section to input section. Hence the Dirac sea hole current in the local vacuum flows back across the system, from its output to its input section. Regarding the output section: Specifically, *all COP> 1.0 systems produce both positive and negative energy outputs. The greater the COP of the system, the greater the fraction of negative energy produced with respect to the fraction of positive energy produced.*

The positive energy will flow on out of the system output section (as into the external load) as in a normal circuit, but the negative energy output will flow from the output section through the local vacuum back across the system to its input section.

At the input section, the arriving Dirac holes appear to the input electrical current as a special kind of "load" added to the system input section. Hence, as previously stated, the external source must input electrons (as electron current) to "fill the holes" prior to having any input electrons available to power the circuit in normal fashion.

That is nature's unexpected *overunity decay mechanism* for COP>1.0 EM systems. The COP>1.0 system is in a condition of disequilibrium *a priori,* since the system exhibits negentropy and only systems in energy flow disequilibrium with their active environment can do that. Any COP»1.0 EM system is therefore in a highly excited state, and all excited states in nature have decay mechanisms. A highly excited state usually has a prompt decay mechanism, and the COP»1.0 system is no exception.

This emergence of the Dirac hole current from the system output section back to the system input section, is nature's decay mechanism to trigger COP»1.0 EM systems back into COP <= 1.0. With this current, a "negative energy load" is added to the input section, causing the system to draw additional electrical power from the external power supply to "power" the killing of the Dirac holes[274] as well as to "power" the normal circuit (the system itself).

If the external power source connected to the COP»1.0 system is not furnishing sufficient electron current to kill the hole current in the input and power the system as well, the excess hole current will pass back through the conductors and on into the external power source, "eating current" and power as it goes. The product of the hole current and the voltage gives a negative power imposition, wherever the hole current interacts with the power system and its external power supply, even back into the distant huge power generator in the distant power plant. For a battery-powered circuit, e.g., excess hole current will pass into the battery and discharge it (change its chemistry in discharge mode) more quickly.

Indeed, a lead-acid battery (and some other kinds as well) can be supercharged with Dirac holes. What this actually means is that the local vacuum — in which the battery and its components and chemistry are embedded and in a continual exchange with — is altered to increase the number of empty Dirac sea holes devoid of electrons. At the same time, the local spacetime is also negatively curved. The end effect is that the battery is charged with negative energy, rather than positive energy. We propose that a battery or capacitor possessing an extreme negative energy charge should be lighter in weight than the same capacitor when uncharged.

As previously stated, some years ago, Jim Watson developed a very large multi-kilowatt COP>1.0 battery-powered generator, publicly demonstrating it at a technical conference (see again Figure 5-12). Engineering measurements at the conference verified that the batteries were constantly charging while the load was constantly powered at 8 kilowatts.

The longer Watson's system would run, the more charged his batteries became with *negative energy*. After one of his batteries had been used for a couple of weeks in extensive experiments, one could obtain a most novel effect when it was placed on a normal "positive energy" battery charger. The battery's excess "Dirac hole charge" or "negative energy charge" accumulated in its local vacuum would "eat positive power and electrons" from the battery charger for an extended period, sometimes as long as a week, with no symptoms of "charging" the battery at all (no increase in battery voltage at all). During this period, the battery charger was indeed recharging the *battery supersystem.* It was filling the excess Dirac sea hole charge in the vacuum with the steadily supplied electrons. In the process, it was gradually relaxing the previously formed negative spacetime curvature.[275] When all the Dirac holes were filled, the relaxed spacetime then was relatively flat (normal) as was the local vacuum, and suddenly the battery would charge-up normally again with positive energy, with its voltage increasing in quite normal fashion.[276]

---

[274] By converting separated incoming electrons and incoming holes to resulting filled Dirac sea electron holes, in the input section itself.

With tongue in cheek, we point out to the COP>1.0 researcher that these effects are not described in the available storage battery manuals, no matter how technically they are written. To the electrical engineer, we point out that they are not in his textbook.

Whenever Dirac sea holes have accumulated in the input area of the COP>1.0 system and are being filled by the external power supply, normal ammeters and voltmeters will only "see" that the external power system is furnishing more current. That is true, because the holes are an additional load being powered. The holes "eat incoming electron current and power", and require the external power supply to "power the killing of the holes" as well as "power the system".

The unwary overunity experimenter can successfully have the first phase of an overunity system working and not realize it, because his initial COP is too high and he does not know about the associated hole current decay of overunity systems back to underunity systems. His system may actually be at COP>1.0, but the negative energy component being produced in his output section is back-flowing and continuously degrading his input section by adding a "new load" right in the input section.

---

[275] This proves that a system's supersystem can indeed be negative-energy-charged, so that the local spacetime around that system is negatively curved, even though the system itself seems normal and functions normally. If the negative energy charge is . sufficient, conceivably the antigravity produced will be sufficiently great that the system itself "floats" or even levitates. We hypothesize that an intense negative energy beam can indeed be technically created which will accomplish this "negative-energy-charging of the supersystem" effect when continuously absorbed by a targeted system's local spacetime supersystem component, as the system is continuously irradiated by the beam. In that manner, levitation beams, tractor beams (both attracting and repelling) are at least hypothetically possible. If this proves to be a valid hypothesis, we foresee the day when such beams will revolutionize transportation and handling of otherwise heavy loads and very heavy systems. Even the heaviest vehicle could then negatively charge its own supersystem's local curved spacetime, thereby floating in the earth's gravity, in the atmosphere, in space, and under the ocean.

[276] We deliberately omit discussion of the extraordinary all-purpose weapons implications of negative energy electromagnetic pulse (EMP) directed energy weapons, already developed and deployed by several major nations.

So his overall power draw from the external power source can be such that his positive power output in the load is less than his total power draw at the input of his system. Unless he is aware of the process and how to handle it, he may never realize that his system is actually at COP>1.0 condition, when the negative energy component is credited as an output "plus".[277] *COP>1.0 research starts by evaluating the absolute value of the output energy, then working from there!* The experimenter must be ever aware that he can and will have Dirac sea 4-holes and currents in any unitary COP>1.0 electrical power system, and particularly in those with very high COP. He must understand what the 4-hole currents and the negative energy actually do and how they function in the overunity system unless corrected.

Any COP»1.0 EM system demonstrates the Dirac sea 4-hole current phenomenon as nature's special decay mechanism to reduce overall performance of the COP»1.0 power system and its external power supply to COP < 1.0 — unless one *corrects* the problem and prevents the decay process (developed later in this chapter).

One symptom of the negative 4-hole currents is their cooling capability rather than heating. Another is that the excess current (and power) drawn from the external power supply also does not heat the input section of the system. Another is that semiconductors operating within the Dirac 4-hole current flow are likely to fail immediately and permanently.

We strongly suggest that, for high COP systems experimentation, the experimenter use a system of synchronized clocks where the lab clock is synchronized with a remote clock prior to the experiments, then the two clock times are examined after significant experimentation at COP»1.0. If deviation beyond normal error in synchronizing occurs, then the involvement of spacetime curvature effects is indicated. Two decades ago, anomalous experiments by Golden — with whom I was working at the time — charged the area with negative energy, producing a significantly curved local spacetime which disrupted local clocks and watches for four days as the charge gradually decayed away. Another indicator is that Geiger counters and other nuclear radiation detectors often "read" when bathed in vacuum 4-hole current, because of the ionization changes induced in them and also because of their individual previous "time-charge" histories. After all, ionization instruments "read" anything which causes ionization of their internal gases. Dirac sea hole current, if interacting with the gas, is in fact a very special new form of "ionizing" radiation.

---

[277] In other words, in the presence of both positive and negative energy output, true COP is not "positive energy out" divided by "positive energy input by the operator", but is more appropriately "total absolute value of the energy output" divided by the "positive energy input by the operator".

9.11.2 Antigravity Hole Current Effects in a COP»1.0 EM System.

Any COP>1.0 EM system can exhibit minor, usually barely detectable antigravity effects due to the presence of the negative local energy 4-holes in the hole current if the hole current is not transduced or corrected. It can also produce minor antigravity "thrust" due to the flow of the local energy holes. Usually these effects are not even noticeable in an overunity system unless it has very high gain (very high COP, say in the range of 10^5 to 10^6 or more.).

However, if the EM system's COP » 1.0, and reaches 10^5 to 10^6 or more, the antigravity and thrust effects may become significant and easily detectable. At 10^9 they become highly significant and are very detectable: The unit will lose appreciable weight on the laboratory bench, and may even levitate easily.

An example is provided by the Sweet vacuum triode amplifier discussed above, with the results that were given in Figure 6-9. Such effects are also what generate the anomalous levitation phenomena of free-spirited John Hutchison.[278]

9.11.3 Negative Energy Effect When Leads Shorted

When the leads of a COP>1.0 EM system are shorted, a noticeable cooling effect can sometimes be detected. If COP»1.0, shorting the leads will result in instant icing from the water vapor in the surrounding air.

If one accidentally shorts oneself across the leads of a COP»1.0 system, the effect is to freeze the nerves along the pathway of conduction. Experimenters are cautioned to use every safety precaution. Their experimentation is at their own assumed risk.

---

[278] John Hutchison, "The Hutchison Effect apparatus." *Proc. Intl. Symp. New Energy*, Denver, Colorado, May 12-15, 1994, p. 191-198.

Anecdotally, some older researchers reported that their constant handling of overunity circuits resulting in charging their bodies, apparently with negative energy. The body after all may be considered as a special kind of complicated capacitor or set of capacitors, so this is reasonable, based on Bedini's and Watson's demonstrated negative energy charging of a battery. Some of the older overunity researchers reported that, after charged in that manner, they could produce a 3-inch spark from their fingers, or a "slight shock" into the body of a person they touched.

There is even a special form of acupuncture in the East in which the practitioner uses special exercises to charge the body with negative energy in this fashion. The "electro acupuncture" is then administered by the operator twirling the inserted acupuncture needles between his fingers, and deliberately administering a jolt of negative energy from his body into selected points in the body. I have personally tested a person who could deliver strong "electric shocks" into one's hand or body, and who could consistently produce a 1-inch spark discharge between the hands at will. That person could also deliver a very astonishing "jolt" to one's body, particularly in the vicinity of the liver.

There appear to be similar rare negative energy phenomena in semiconductors {585a-585c}. Anecdotally, it appears that the body does have structures or methods of storing and using Dirac sea holes (negative energy) and mechanisms for accessing the negative energy charged in its supersystem. This is suggested as a fruitful line of future research for interested biophysicists. Again, we call attention to the fact that, whenever EM fields, potentials, and waves are present, Dirac sea effects are also present and involved. Little or no biophysical experimentation has been done utilizing higher symmetry electrodynamics and analysis of the supersystem effects. Contrary to the received view, the understanding of EM effects on biological systems is not mature at all, but is still in its infancy.

9.12 Susceptible Unitary System without Conversion of Dirac Sea Hole Current

A *unitary system* is simply a self-contained single power system with an overall closed current loop between the source dipole and system loads and losses. A unitary system is contrasted to an array system of synchronized separate power subsystems without a common closed current loop, with the array subsystems connected to the source dipole by energy transmission-reception only. Hence the array system utilizes the Kron open path concept.

Beginning at about COP = 1.2 to 1.5, a few highly susceptible unitary COP>1.0 electrical power systems will experience significant Dirac hole current from output to input. This will produce a "negative energy load" in its input section and feed line from the external power system, resulting in decay back to an overall "system plus negative energy load" of COP<1.0. At the COP = 2.0 level, in the susceptible system the effect will forcibly decay the COP back down to overall COP<1.0. It does this by forcing excess power to be furnished by the external power system to "power" (fill) the Dirac hole current arriving in the input section of the power system. In short, it converts the system to a composite system containing an additional Dirac negative energy load in its input section.

Or in other words, it forces the operator to "pay" for both the *positive* energy output and losses of his system and the *negative* energy output and losses of his system.

In addition to usual powering of its load and losses, such a system is also "powering" the return of its own locally altered vacuum back to a more normal ambient vacuum, instead of stabilizing its supersystem in a stationary disequilibrium state. It is a supersystem interaction and effect. A COP>1.0 system that significantly exhibits this phenomenon is said to be a "susceptible" system.

One of the first tasks of the COP>1.0 experimenter is to ascertain whether he is working on such a highly susceptible COP> 1.0 system. Ironically, he does this by observing his COP>1.0 system kick itself back into a COP<1.0 system. If he carefully observes the initial COP>1.0 condition followed by decay into COP<1.0 condition, he must then realize that not all is lost! In short, he must realize he has developed a susceptible system. If he obtains COP>1.0 only to watch it subsequently decay back to COP<1.0, he has just discovered he is working on a susceptible COP>1.0 system, *and that is experimental progress.* He isn't there yet, but he is en route. In that case, he must concentrate on transducing the excess hole current into usable electron current, to "stabilize" and "lock" his system in its COP>1.0 operating condition.

9.13 Unitary System with Conversion of Dirac Sea Hole Current

Using the Bedini conversion process, a unitary system can be close-looped to maintain COP> 1.0 and stable overunity operation. The 4-hole current reaching the input section is intercepted by the Bedini mechanism, and transformed to ordinary electron current furnished to the input section. Thus the transduced negative energy feedback becomes a positive energy feedback, by conversion. This converted positive energy in the input section will reduce the demand for incoming electron current from the external power supply, increasing the COP>1.0 even more.

A feedback and governing system is adjusted until the system's converted positive energy feedback current reaches the Kron condition. At that point, the transformed electron current is sufficient to power the system with the external power supply disconnected. The governing and control system thus disconnects the external power supply, and the system is smoothly transitioned to self-powering operation, powering both itself and its load.

We stress that all the energy to power the system and its load is now being furnished from the active local vacuum interaction and from the local curved spacetime. No laws of nature, physics, higher symmetry electrodynamics, or thermodynamics are violated. Conservation of energy rigorously applies, but to the supersystem. The system is an open, *stabilized* disequilibrium system — of the kind referred to by Kondepudi and Prigogine as nonequilibrium stationary states — as a freely receiving all its energy from its active external environment. No standard U(l) electrodynamic analysis will show this condition or this mechanism, and neither will electrical engineering. Higher group symmetry electrodynamics such as SU(2) or O(3) will show it easily. For a modern treatise on O(3), we refer the reader to the recent *tour deforce* by Evans {586}.

9.14 Non-unitary Outrigger Array System

See Figure 9-14, which is a diagrammatic representation of a COP>1.0 array system that is non-unitary. In the non-unitary (outrigger array) system, no part of the outrigger array works at $\xi > 1.0$. All external parts operate at $\xi < 1.0$, but independently and with COP>1.0. Each of the external outrigger systems is an open system freely receiving excess energy from its active environment, because it is an "energy-receiving" system and the energy is freely transmitted from the central system. The central part of the array is the unit whose purpose is to stimulate or evoke an excess energy flow from the vacuum, so that the outrigger receiver/power units can individually intercept and utilize radiating energy flow surrounding the central unit.[279] The usable (output) energy intercepted and captured (collected) from a given free flow of energy depends totally on the intercepting reaction cross section — usually the total array amount of intercepting charge, together with the degree of resonance (sweeping a larger geometrical cross section) of the intercepting charge.

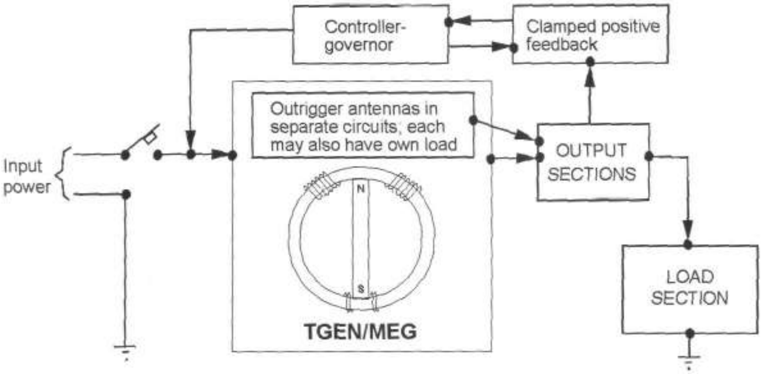
Controller-governor
Clamped positive feedback
Outrigger antennas in separate circuits, each may also have own load
Input power
OUTPUT SECTIONS
N
S
TGEN/MEG
LOAD SECTION

Figure 9-14 The nonunitary (outrigger array) system and its functioning.

Each outrigger unit (subsystem) has its own complete closed current loop *separate* system and load, plus its "antenna" or "interception section" or receiving component for intercepting and collecting the free flow of energy coming from the central source unit (the energy transmitter). Thus, the loads in the outrigger units are not closed-current-connected with the central source unit at all. Each outrigger unit has its own independent electrical ground and its own closed loop system with its load.

---

[279] We emphasize that, when the usually unaccounted Heaviside nondiverged energy flow component surrounding the conductors of the external circuit is accounted, every EM power system is already a COP»1.0 EM energy transduction system. Consequently, in the received view the array system can be operated at COP>1.0, or even in self-powering mode, even though the central system is "said" to be at COP<1.0. It is at COP<1.0 only because of the self-applied Lorentz symmetrical regauging and ignoring the excess unused energy flow available to and surrounding the circuit, but not conventionally used. If we account the COP for what we have to furnish to the central unit, then the COP of each array unit is COP = ∞. The point is that we ourselves do not have to pay any *extra* input energy for the output energy of each array circuit.

This schema in fact separates the two functions of Poynting energy flow theory: (i) the extraction and radiation of the EM energy flow from the seething vacuum, and (ii) the interception, collection, and dissipation of some of that available energy flow, to power the loads and losses of the intercepting system.

In this fashion, multicollection from a single energy flow (generated from a central source "transmitter") occurs by direct, independent interception and collection in each of the COP<1.0 outrigger units. However, the input power to each outrigger units is *free* with respect to that outrigger circuit itself. We of course must pay to generate the flow of energy from the central unit, but that can be done in normal COP<1.0 manner (such as COP = 0.95) while powering its load.[280] Let the output power of each numbered unit (including the central source unit) be given by P_i, n be the number of power units, and P_T be the total power output. Then the total output P_T of the array system is given by

$$P_T = \sum_i P_i = P_1 + P_2 + \cdots + P_n \qquad [2]$$

The total input power P_I is simply the input power $P_{I(1)}$ to the central source unit. The COP is given by COP = $P_T/P_{I(1)}$.

In the motionless electromagnetic generator (MEG) array, for example, the central unit is the transformer-like unit, deliberately using a special core to contain the magnetic flux of a permanent magnet in contact with the core material, thereby stimulating the Aharonov-Bohm effect and freely filling the surrounding space with the uncurled magnetic vector potential A. We will explain the non-unitary (outrigger array) system using the MEG as an example.

---

[280] We again stress that, from any nonzero potential ϕ, or from any EM field in space, any amount of EM energy can be collected if there is sufficient intercepting and collecting charge. For example, the simple equation $W = \phi q$ applies for the potential, where W is the energy collected from a potential with point intensity ϕ by collecting charges q. A similar equation applies with respect to collecting as much force (emf in the various outrigger circuits) as desired. Here a single given E-field *E* can be assumed, and *F* = *Eq*. What is happening in the outriggers is that they are intercepting and collecting some of the excess Heaviside energy flow component that is wasted by the central circuit.

The MEG freely induces the Aharonov-Bohm effect,[281] thereby extracting excess energy from the vacuum in the form of that novel field-free A-potential. Being field-free, A represents an energy flow without swirl (without curl). When the MEG input section is perturbed, the surrounding A-flow is perturbed also, giving a pulsed or AC flow of energy that can be intercepted and utilized by each outrigger circuit as an E-field energy interception.

The central unit's COP<1.0 function is to output almost as much, or as much, energy as is required to power the MEG, while also evoking that excess A-flow and serving as a $COP \gg 1.0$ energy flow transducer. The function of each outrigger is then to intercept some of the pulsed A-flow (as E-field energy via the equation $dA/dt = - E$), in its antenna or interceptor section. Each outrigger has its own independent and conventional closed loop circuit. It simply is a receiver and user of extra Heaviside EM energy produced from the central MEG unit in the new form of the curl-free A-potential. Perturbing the A-potential produces a large E-field impinging on all the external outrigger subunits. Hence these units freely collect and use some of that available excess energy. The COP of the total array is thus COP>1.0. For best results, tuned antenna theory and near field theory should be utilized.

Suppose we operate the MEG central unit at COP = 1.0. In that case, there is no 4-hole current of significance associated with the central unit or the outriggers, and transduction of hole current into electron current for close-looping is not necessary. Instead, sufficient outriggers are employed so that the total energy W_O caught and dissipated in their loads by all the outriggers is given by $W_O \geq W_I$, where W_I is the input energy in the main central section. Since COP = 1.0 for the central unit in isolation, the total input E_{IN} is $E_{IN} = W_T$ (the additional huge Heaviside nondiverged EM energy input component is there also, but just neglected in conventional engineering). The total output W_A of the entire array is $W_A = W_O + W_T$. The overall COP_A of the entire array is thus

$$COP_A = W_A/E_{IN} = (W_O + W_T)/W_T = (W_O/W_T + 1.0) > 1.0 \quad [3]$$

---

[281] The reader should recall that the Aharonov-Bohm effect is not included in classical U(1) electrodynamics. See Terence W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in A. Lakhtakia (ed.): *Essays on the Formal Aspects of Electromagnetic Theory,* World Scientific Publishing, River Edge, NJ, 1993, p. 6-86.

Hence the array has a COP>1.0, even though every single component of the array has $\xi < 1.0$. With the outriggers fully functioning, many schemes of operation can be used. E.g., the energy from the outriggers can be used to steadily charge a capacitor bank by "shuttling". We define shuttling as use of two capacitor banks: (1) a capacitor bank which is continually connected and charged by the outriggers as their effective output, and (2) an input capacitor bank for the central unit that is continually connected and recharged from the outrigger capacitor bank. By adroit switching, the charge of the input capacitor bank is maintained by replenishing from the output capacitor bank, and the input capacitor bank powers the entire system, including the output load in the central section.

The non-unitary outrigger array construction is one means of close-looping a non-unitary overunity power system without bothering with the transduction of hole current into electron current. It is the easiest of all COP>1.0 systems to close-loop for self-powering.

An additional adaptation is for each outrigger also to continuously power a separate load, and only contribute *intermittently* along with the other outriggers to charging the output capacitor bank.

Another adaptation is to operate the central unit in the array in solitary self-powering mode.

These and similar modes of nonunitary array system operation allow close-looping the system for self-powering of itself and its load, without concern **for** 4-hole current transduction. A real advantage of such nonunitary systems without worrisome 4-hole current is that transistors can readily be used for switching etc. in them, without the destruction of the transistors by a damaging Dirac 4-hole current in the supersystem.

The nonunitary overunity system can also be close-looped by *conventional* (simple) "clamped positive feedback" means if good isolation (e.g., optical isolation of the circuitry switching) is used in the feedback loops.

9.15 Some Instrument Considerations

9.15.1 Time Effects

As previously indicated, there are time-domain effects (potentials, forces, and currents) associated with a COP>1.0 EM system, particularly as the COP increases toward COP»1.0. A useful method of indicating time-effects is to utilize a transistor in a circuit capable of measuring recombination time in the transistor.[282] NASA has such circuitry available, but I do not have the reference at hand. We leave it as an exercise to the reader to explore that possibility.

For "very large COP" devices, ordinary watches, clocks etc. may be affected. Particularly significant is the slowing of timepieces in the presence of substantial time-domain effects due to substantial negative energy. In addition, for high gain devices (and for some at lower COP), the negative energy charge can build up in the local area, and may then require some days to dissipate. As we stated, in experiments years ago with Frank Golden, we experienced just such alteration of multiple kinds of clocks and watches in an area. It required four days for the negative energy charge in the area to gradually drain off (for the excess Dirac sea holes to gradually be filled with electrons, as the tempic potential gradually returned to normal).

Effects that were even more dramatic were exhibited in experiments conducted by Sid Hurwich in Canada some years ago {587}. In a series of experiments lasting a half-hour, Hurwich inertially jammed a police revolver so that a human finger could not pull its trigger. *The watches of the witnesses did not change their setting during that entire half-hour.* This cannot be understood by conventional electrical system analysis, but potentially it can be understood by supersystem analysis.

9.15.2 Calorimetry Is Taboo for Overunity System COP Measurement

It is standard dogma that calorimetry is a "true" measurement of power, and is always accurate in the hands of a skilled practitioner. In general, that is true for a COP<1.0 system, where one is concerned only with *positive energy* in the input and output sections. However, with respect to COP>1.0 systems, nothing could be farther from the truth. Indeed, for a significant COP»1.0 system, a calorimeter is one of the most inaccurate instruments that can possibly be used.

As previously stated, all COP>1.0 EM systems produce a mixture of negative and positive energy in their output section. The higher the COP, the greater the percentage of negative energy and the greater the flux of Dirac 4-holes back to the input section. The output energy, immersed in a calorimeter, will both cool the water with its negative energy output fraction and heat the water with its positive energy output fraction. Hence, the calorimeter will show the *difference* between the water's simultaneous heating and cooling. If the calorimeter measurement is misinterpreted as the *total energy output*, this "definitive test" will always erroneously show COP<1.0. After all, it itself already applies the decay mechanism for converting any overunity system to an underunity system.

---

[282] This suggestion is courtesy of John Schnurer some years ago, and is appreciated.

The only good use of the calorimeter in an overunity EM system is to verify that negative energy is present, when comparing its measurements to more sophisticated electrical measurements of the system. It can be used along with other measurements to ascertain (1) the total energy output, whether negative or positive, (2) the difference in the two forms of energy output and the algebraic sign of the resultant (positive if calorimeter heated, negative if calorimeter cools), and (3) thereby help determine the output's energy fractions (positive energy divided by total energy, negative energy divided by total energy).

Any test group insisting on testing a purported COP>1.0 electrical system with a calorimeter as the *definitive statement of energy output* is totally devoid of knowledge of COP>1.0 systems and their phenomenology. Usually such "test groups" tend to regard themselves as "measurement experts" (which they well may be, in COP<1.0 positive energy systems!). That does not make the team even minimally knowledgeable or qualified in COP>1.0 system measurement. To the contrary, the team members have zero experience or knowledge of COP>1.0 systems. The proposed test group may consist of physicists, thermodynamicists, electrical engineers, technicians, or all four, but they still will have no knowledge or expertise in COP>1.0 EM power systems unless they have actually worked with such systems before, and at length.[283] This is particularly true of COP»1.0 systems. My advice to all COP>1.0 researchers is simply to keep the "calorimeter addicts" out of one's laboratory in the first place.

9.15.3 Other Instrumental aspects.

Here we just point out some simple and obvious things, familiar to the experienced researcher but sometimes not appreciated by the novice. One is usually dealing with nonsinusoidal waves, pulses, spikes and the like. As is well known, RMS meters are useless for measuring such nonsinusoidal electrical entities, since they are designed and calibrated to measure sine waves only, or to measure DC. One must use a good differential sampling scope with multiple synchronized channels, with professional software to perform accurate integration under the curve, and with special differential probes for the scope. Simultaneous triggering of the multiple measurement channels on the same time line is required. The necessary high quality probe set may cost as much or more than the oscilloscope!

---

[283] To date, we have found none who have actually worked with COP>1.0 electrical systems. We have, however, found quite a few who, though totally inexperienced with COP>1.0 systems and not knowledgeable, have regarded themselves as "real experts" in such systems. They claim to already know how COP>1.0 systems work and behave, and assume just ordinary electrical system phenomena.

Dirac sea hole currents are generally not separately measurable, so one measures where the hole currents are not significantly present. Results may be compared to measurements where the hole currents are known to be present. When directly measuring hole currents, the conventional meters often read "backwards" and interpret a negative energy output at the system output section as subtracted from the positive energy, and interpret a negative energy output at the input section of the system as additional input electron currents from the external power supply. If the electron currents from the external power supply are not present, then measuring hole currents will draw electron currents from the power system inside the meter itself, resulting in a "backward measurement". Again, the meter actually will measure a real electron current, whether from the external power supply or from a conversion within the meter itself.[284]

Recombination time in semiconductors can sometimes be used in instrumented circuits to differentiate between negative hole current and electron current, by observing the change in recombination time due to the negative energy currents. It will differ from the change in recombination time due to positive energy currents.

Again, what is needed in the field is a set of solid, reliable instrumentation specifically developed for these peculiar measurement phenomena involving both positive energy and negative energy. To my knowledge, no such thoroughly designed and tested instrument package presently exists. Further, there are not even any standards for such, since there are apparently no standards for negative energy measurement at all.

9.16 Still Anomalous Aspects

Several aspects of COP>1.0 systems are still not properly understood. The behavior of curl-free A-potential, sometimes produced in such systems, is a case in point. This type of energy can exhibit characteristics as if it were electrostatic energy, which my associates sometimes refer to as the "crawly stuff. It exhibits the characteristic of "crawling" all over the surfaces (outside and inside) of the objects in the area, including the scopes, producing internal electrostatics and possibly magnetostatics everywhere. This can easily destroy the amplifiers etc. in a good scope or an expensive meter. It is best to transform this field-free A-potential into an ordinary "EM field energy AC wave", and then utilize it in that fashion.

---

[284] What is very much needed is a meter of sophisticated design where this "drawing of excess electron current from within the meter itself has been taken into account in the design, so that the meter functions correctly for either positive energy measurement or negative energy measurement.

Scalar stress waves are also sometimes produced, leading to anomalous effects. This is best regarded as a "time-polarized" EM wave, or more accurately, as a combination of scalar (time-polarized) EM waves and longitudinal EM waves in 3-space. For macroscopic waves in COP>1.0 systems, the Mandl and Shaw 1:1 correlation between scalar photons and longitudinal photons can apparently be corrupted, so then one has "abnormal" (improperly correlated components which normally correlate to make "ordinary" EM waves). These abnormal EM scalar stress waves also exhibit the "crawly" phenomenon, and can be hazardous to one's instruments.

In addition, some instruments such as ionization counters, Geiger tubes, etc. read in the presence of such waves as if nuclear radiation were present (which it is not). Also, the individual time-charge history of a given instrument determines or substantially influences how the instrument reacts in the presence of the "corrupted" scalar stress waves. Essentially, at a given time any instrument has a small "time-energy charge" depending upon its past history. The corrupted scalar stress wave accomplishes further "time-charging" of the instrument, which combines with the initial time-charge to produce more time-charge, ordinary EM waves by scalar EM interferometry {588}, or abnormally correlated longitudinal EM wave emission.

In an ionization-measuring instrument, *ionization is ionization, by whatever means achieved.* And in the presence of the "abnormal" radiation and charging, ionization-type "nuclear radiation" instruments will vary from one to another, *depending upon the individual time-history ofeach instrument.* So from a group of identical instruments made the same day on the same assembly line, one may read anomalously and several others read normally. Or two may read anomalously, while all the rest read normally. A very good example of this anomalous measurement effect is seen in the rigorous electrolyte experiments in U.S. Navy research facilities at China Lake {589}. The present author has previously given an explanation of this time-charging and decay phenomenon {590}.

Negative energy reacts in different ways with different materials. The more nonlinear the material, the greater the effects and unpredictability of the interaction, as a rule of thumb. In a copper wire, Dirac hole current introduced into the wire will "eat" Drude gas electrons, reducing the "pressure" at that point or area in the wire. This "combining of electrons and causal Dirac holes is not normal pair annihilation, and radiation does not occur. Instead, the two curvatures of spacetime (one by the positive energy of the electron and the other by the negative energy of the Dirac hole before observation) interact and adjust.

The Drude gas in the copper further back toward the external power system still has normal (higher) gas pressure, and so the differential in pressure induced in the Drude gas will drive electrons from the "normal" pressure areas into the "reduced" pressure area where some of the Drude electrons disappeared after being "eaten" by the Dirac holes. To look at that in lattice hole terms, the "eating" of Drude electrons in a selected volume of the wire leaves positive ions, which are "lattice holes" vacated by the eaten electrons. So these excess positive charges attract electrons from further up the wire nearer the distant power supply, filling the lattice holes to bring the Drude gas back up to "normal". However, that also draws and "eats" additional power from the distant power supply. Meanwhile, the Dirac 4-holes fed from the input of the COP>1.0 system back into the power feed conductors continue to interact with the lattice and eat electrons to create more ions (lattice holes).

One may visualize the Dirac sea holes producing excess lattice holes, which then causes the Drude electron gas to change so that the "average" demand current increases. In short, we must also consider the effect of Dirac hole charges or currents upon the lattice material itself, and the structure of holes and electrons as charge carriers in the material lattice.

Semiconductors are very vulnerable to negative energy due to its disruption of their donor-acceptor and band gap design functions. With mild negative energy exposure, a semiconductor may just cease operation entirely and later recover when the negative energy application ceases, or it may fail and never recover. For pulsing, strong "opposing force pulses" (stress pulses) are created deep inside the semiconductor materials. In case of sharp pulsing, sometimes an exposed semiconductor will just suddenly explode in a nice little fireball about one foot in diameter. This along with some minute stinging fragments from the explosion assures that the COP>1.0 bench researchers stay alert and on their toes. Needless to say, when experimenting with COP>1.0 systems and circuits, one should exercise all safety precautions, including safety glasses, until the **"negative** energy beasts have been tamed and controlled". One should not experiment with such things unless one is an experienced and safety conscious experimenter. One experiments at one's own risk.

Negative energy effects in coils and toroidal coils are largely still unknown, though we have outlined some of the major effects. For a normal helical coil, one can derive a concept of the effects by considering the lattice effects in the wire. In this case, introduction of hole current at one end of a helical coil will induce electron current from the other end and into the coil. Thus, hole current can be "switched" back and forth between the ends of a helical coil to induce an ordinary AC electron current in the coil, accompanied by normal magnetic fields etc. A point can be reached where the coil is actually in resonance with the frequency of the switching, and with the "seeming capacitance" of the hole current introduction in each switching. With the proper material in the coil and a keen balance between the negative energy capacitance effect and the normal positive energy induction effect, *in theory* a single coil can be turned into a self-oscillating, self-powering system.

With core material inside the core of a coil, the situation dramatically changes. Often the type reaction experienced will appreciably depend upon the core material, and changing the core material can dramatically change the effects upon the coil and the effects that the coil produces. One may even use dynamic changes in the core material to produce corresponding dynamic changes in the coil. A toroid may act differently from an ordinary helical coil, as does a flat (spiral-wound) coil. Here the experimenter must simply perform his own experiments to seek out and understand the phenomenology he uncovers.

For exploratory experimenting with core-filled coils, we advise strong caution and testing at very low power only. Sometime effects can apparently be experienced by the researcher (headaches, nausea, dizziness, etc.) It is suspected from some evidence (but not proved) that the core material's chemical characteristics can sometimes be "imprinted" upon the human body chemistry, from the novel radiation of a core-filled coil or some core-filled coils. This is particularly true in the presence of strong stress potentials, such as from bi-wound coils with fields in opposition, and particularly if the biwound coil contains a nonlinear material core. The nonlinear core induces modulation (multiplication) of the two waves, rather than linear mixing. Thus, the two waves are "locked together" to form a true stress potential and local spacetime curvature sets, when the biwound coil has a nonlinear material core. In short, one is now engineering pure general relativity, locally.

Again, we iterate that the experimenter should be knowledgeable and take all safety precautions. He experiments at his own assumption of risk.

The appearance of an extra stress potential (field-free) in a biwound coil is in fact a new Lorentz symmetrical regauging, and the equivalent of rotating the frame of the coil. As signals are introduced, this is the equivalent of continually changing the degree of the frame rotation, which is a general relativistic effect, with direct curvatures of spacetime being produced. In short, one is producing spacetime curvature and rotation engines, previously discussed.

The magnitude of the stress potential (field-free) falls off inversely as the distance. Thus one's body can be exposed to it, including the body chemistry, every cell and every part of every cell, the nervous system, etc. Placing signals on the biwound coil can conceivably result in slight variation of the scalar stress potential, including inside the body, and inside each cell and each part of each cell. Variation in the stress potential ϕ in a little local region inside the body produces $-\nabla\phi$, which is an electric field E. The rate of change of this E-field then produces a magnetic field H.

Consequently, very minute EM fields are produced throughout the body and all its parts, or can be. It is as if the entire body is "electromagnetically washed" in internal EM fields throughout its matter, structures, and dynamics. The hydrogen bonding actions ever ongoing are particularly sensitive and can restructure appreciably in the presence of this stimulus. They can also "imprint" and take on a specific induced structure and dynamics, correlated to the persistent changes induced from the biwound coil. As can be seen, such EM biological effects have not been appreciably investigated and documented in the Western scientific literature. Hence our urging and insistence that the researcher use all caution, and our warning that he experiments at his own will and assumption of risk. Neither the publisher nor the present author is responsible for any accidents or blunders that may occur.

Capacitor effects also vary between types of capacitors, when applying negative energy or a mix of positive and negative energy. Some capacitor dielectric materials are also photo-refractive and are semiconducting materials, hence photoactive. The rich effects capable of being generated in capacitors — by EM fields, waves, potentials, and pulses of negative energy or negative and positive energy mixed — are largely unexplored and not catalogued.

Almost all modern permanent magnetic materials are also photo-refractive or photoactive materials as well, and many are semiconducting materials. Hence a rich variety of effects can be produced in permanent magnets by the introduction of negative energy fields, potentials, waves, and pulses.

As can be seen, a very rich combination of effects exists for "normal-appearing" coils, capacitors, semiconductors, and magnets when applying appreciable negative energy or controlled mixes of positive and negative EM energy. This is a still largely unknown materials science region, and one that has not been systematically and rigorously explored and catalogued, if at all.

In addition, there are effects from *pulsing* negative energy (Dirac hole currents) that are not completely understood. These appear to occur almost — or partially — at random, and also seem to be affected by the natural surroundings, by the various materials, etc. The "imprinting" phenomenon, when it occurs, seems to be augmented by pulsing, and sometimes by the pulse shape and pulse repetition frequency as well as the surroundings.

We also legally state that we are not responsible for any risks the researcher chooses to take, but are only alerting him to be very careful and use very good safety practices. Any experimenter is himself responsible for stringent adherence to good safety practice and common sense, and he legally assumes his own personal risks for all experiments he chooses to undertake.

Another effect to be alert to is the effect of "charging" the local area's supersystem with an altered active vacuum and curvatures of spacetime. This phenomenon is usually experienced as a "slow growth" of the COP of the system, over a long time (months or years). Tiller in private discussions has reported such phenomena in his detector built at Stanford University. Golden and I experienced such phenomena with a small COP>1.0 system Golden was developing. Over a period of five years, the system slowly increased its COP to about 2.0. Moved several hundred miles, the system then exhibited COP of about 0.9, but with occasional fluctuations of overunity temporarily.[285] Golden had gradually charged his local area with the "engine" or spacetime curvature operational and functional form of his unit. Thus in that repeatedly charging area, gradually the operation of the unit improved. We referred to this *engine-*

---

[285] In fact, Golden and I were astounded at these results, until Tiller explained it to us and explained his own experiences with "growing" his detector. Thanks to Professor Tiller, thereafter we had this new phenomenon firmly fixed in our lexicon.

charging of the environment with a particular effect or dynamic form as *kindling.*

The kindling or "charging to favor the dynamic form of the engine or device" problem, should it occur, is in fact usable because it is possible to produce it everywhere on earth, and fairly permanently. To use the kindling effect once discovered, the mechanism is to build several identical units. Keep the original unit working in the original area, and "outrigger" the others at radial distances as distant as possible, but maintaining COP>1.0 for the outriggers. Then simply run both the outriggers and the central unit, steadily. The "structural charge" will grow and reinforce at the outrigger locations, then spread on beyond the outriggers, and a much larger area will now be "charged with that dynamic form". With sufficient additional outriggering (three to five, or seven in an especially difficult case) where the successively internal outriggers arc retained and operating, the charging will eventually spread around the entire Earth, and *it will become a permanent fixture of the Earth environment itself.*[286] Thereafter, anytime one of the units is turned on anywhere on earth, it will perform with its normal overunity condition. In other words, not only can one grow the dynamic form (engine) in a supersystem in a given locality, but one can also kindle it into the entire Earth's supersystem permanently {591}. One is strongly reminded of Feynman's observation that physicists with their big accelerators may in fact be partially creating their own reality. He pointed out that often the physicists look and look for a new, predicted particle. At last and after years of difficulty, finally a "glimpse" of the particle occurs in an accelerator and is reported. The physicists quickly crank up the other accelerators and look intensively, and — after some moderate difficulties — they find that particle. Thereafter, *every time an accelerator turns on and looks, that particle is readily seen, easily.*

The supersystem's spacetime is part of the local spacetime of the Earth, and all supersystems also interact with each other. As Kron put it {592}:

> *"An electric network differs from all other types of non-electric networks in that it is always surrounded in all directions to infinity by an invisible dynamic electromagnetic field of its own creation. ... Each Inductance L produces its own magnetic field, each*

---

[286] We have proposed an application of this phenomenon to kindle and maintain a specific disease organism's "anti-engine" as a startlingly new kind of medical therapy for permanently eradicating a particular disease vector from the earth.

> *capacitance produces its own electric field and each resistance R creates a thermodynamic field of its own. Moreover, there exists a continuous interaction between the magnetic, electric, thermodynamic, and radiation fields. Thus any correct and basic theory of the electric network must include all the parameters of its surrounding electromagnetic field and must satisfy all the field equations of Maxwell. The theory must also dovetail with the Hamiltonian and Lagrangean equations of dynamics containing both time and space derivatives, as well of the equations of irreversible thermodynamics. "*

So decades ago, Kron was struggling with the "supersystem" problem, in different terminology, and he eventually originated *diakoptics* {593}, an entirely new way of breaking the problem down piecemeal so it could be solved. But Kron strongly stressed that the interactions of the fields and potentials of a circuit and all its parts reach to infinity and have an n-dimensional set of degrees of freedom and an infinity of spaces. The fields and potentials of every charge reach to infinity (for those charges in matter from the beginning), and so every charge in the universe interacts with every other charge and its fields, potentials, and dynamics. Spacetime itself is filled with these interacting EM fields, potentials, waves, and dynamics. Little wonder that the horrendous set of ongoing EM interactions at every point in the vacuum has such incredible energy density and fluctuations. We may even regard the fluctuations and interactions as completely causal, but — with no information on each interaction — they must be treated as totally statistical. Ultimately nature confounds our Aristotelian logic and insists that, in the limit, one is dealing with 5-law logic as developed earlier in this book.

Since the earth near an ongoing COP>1.0 system experiment has innumerable electrical aspects, materials, etc., it follows that the surrounding earth is indeed in direct interaction with the experiment. It then follows that "slow conditioning" of the surrounding area should be possible, say, after some years as Golden experienced. It also follows that the effect diminishes in magnitude with distance from the experiment.

The "growth of the activation (kindling) to include the entire Earth" is a slow process, requiring years and much expense to aid the "spreading" — as the physicists may be experiencing, according to Feynman. But it is a doable.

We highly recommend that the researcher work on COP>1.0 systems that are not charge and location dependent, so that the kindling problem does not arise. We have assumed non-kindling COP>1.0 systems throughout most of this book.

Chapter 10

Cold Fusion: Low Spatial-Energy Nuclear Reactions at High Time-Energy

"This book is about time. I would have liked to have named it Time, the Forgotten Dimension..." **[Ilya** Prigogine]. {594}

"The notion of time may be unfolded into an independent pure science... a science of pure time is possible." [Sir William R. Hamilton, 1837]. {595}

"...contrary to the conclusions of classical mechanics, there exist effects of potentials on charged particles, even in the region where all the fields (and therefore the forces on the particles) vanish. " [Y. Aharonov and D. Bohm] {596}.

"...global behavior greatly modifies the very meaning of space and time. Much of geometry and physics is based on a simple concept of space and time, generally associated with Euclid and Galileo. In this view, time is homogeneous. Time translations may have no effect on physical events. Similarly, space is homogeneous and isotropic; again translations and rotations cannot alter the description of the physical world. It is quite remarkable that this simple conception of space and time may be broken by the occurrence of dissipative structures. Once a dissipative structure is formed, the homogeneity of time, as well as space, may be destroyed. " [Ilya Prigogine] {597}.

"As we expand our observation, we extend our concepts. Thus the simple symmetries that once seemed self-evident are no longer taken for granted. Out of studies of different kinds of interactions we are learning that symmetry in nature is some complex mixture of changing plus into

minus, running time backward and turning things inside out. " [T. D. Lee] {598}.

"The desirable features of a system of units in any field are convenience and clarity. For example, theoretical physicists active in relativistic quantum field theory and the theory of elementary particles find it convenient to ***choose*** *the universal constants such as Planck's quantum of action and the velocity of light in vacuum to he dimensionless and of* ***unit magnitude.*** *The resulting system of units (called 'natural' units) has only* ***one*** *basic unit, customarily chosen to be length. All quantities, whether length or time or force or energy, etc., are expressed in terms of this one unit and have dimensions which are powers of its dimension. There is nothing contrived or less fundamental about such a system than one involving the meter, the kilogram, and the second as basic units. It is merely a matter of convenience."* [J. D. Jackson] {599}.

[Violation of Second Law of Thermodynamics] *"We experimentally demonstrate the fluctuation theorem, which predicts appreciable and measurable violations of the second law of thermodynamics for small systems over short time scales... Our results show entropy consumption can occur over colloidal length and time scales."* [Evans et al.][287]

[Implications] *"This result has profound consequences for any chemical or physical process that occurs over short times and in small regions. "* [Evans et al.][288]

[Implications] *"The smaller a machine is, the greater the chance that it will run backwards. "* [Evans et al., *ibid.]*

[Implications] *"Physicists knew that at atomic scales over very short periods of time, statistical mechanics is pushed*

---

[287] G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.*, 89(5), 29 July 2002, 050601.

[288] BBC News, **18 July 2002, http://news.bbc.co.uk/l/hiscitech/2135779.stm.**

> *beyond its limit, and the second law does not apply. Put another way, situations that break the second law become much more probable. ...the new experiment probed the uncertain middle ground between extremely small-scale systems and macroscopic systems and showed that the second law can consistently be broken at micron scale, over time periods of up to two seconds."* [Chalmers] [289]

10.1 Explaining Time as Energy

See Chapter 1, paragraph 1.1.3 and 1.4.1; see also Chapter 2, paragraph 2.1.3. As highlighted in the quotation from Jackson {599}, it is well known that the fundamental units chosen to model physical phenomena are arbitrary. In a physics model whose only fundamental unit is the joule, time is obviously modeled solely as a function of energy, just as mass is modeled today as a function of energy — hence the term "mass-energy". We may also equally well use the term "time-energy", since in a model using the joule as its only fundamental unit the second is purely a function of energy. Specifically, for our purposes 1 sec = 9 x 10^{16} joules of spatial energy, compressed by the factor c^2 and placed on the 4th Minkowski axis to constitute one second of time. If the compressed spatial energy had been left in 3-space, it would be recognized as *mass*. For our purposes, *time* has the same energy density as mass.

The flow of time becomes a special form of EM energy flow, analogous on the fourth axis to mass flow (translation) in 3-space. Further, as previously pointed out the overall flow of time has extensive internal structuring of component time-energy flows. However, time-flow is in the imaginary plane or connected with it, since in Minkowski space the 4th dimension is modeled as -i*ct*. Note that the only variable in *ict* is the *t*, hence anything varying on the fourth axis is a variation of time and time-energy *a priori.*

Further, we can apply conservation of EM energy to the time domain as well as in 3-space. And we can have "broken symmetry" in time-energy flow as well as in 3-spatial energy flow, so long as the net energy flow is conserved. The most fundamental conservation of energy law is giant negentropy: the automatic circulation of energy from the time domain into 3-space via decompression by the negative charge, thence the slight flow of 3-space energy to the positive charge where the 3-space energy is recompressed, thence from the positive charge back into the time domain.[290] That was our solution to the long-vexing problem of the source charge and its associated 3-space fields and potentials and their 3-space energy. From special relativity, one already knows that time can dilate and then *un*-dilate, but each action is associated with an appropriate change in spatial energy. While that change in spatial energy can be kinetic, it does not have to be. It can be a change in potential energy instead. In short, a static object rotates its frame with respect to the observer (lab) frame, whenever it absorbs or emits potential energy.

---

[289] Matthew Chalmers, "Second law of thermodynamics 'broken'." *New Scientist,* 19 July 2002, http://www.newscientist.com/news/print.jsp?id=ns99992572 .

We also pointed out the connection of time-energy to electromagnetics and to quantum field theory, via the four polarizations of the photon considered to travel in the z direction in 4-space. These four polarizations are with the energy vibrating along *x*-, *y*-, *z*-, and t- respectively. Vibration of the spatial energy of the photon along the *x*- or *y*- axis gives an ordinary transverse photon. Vibration of the spatial energy of the photon along the z- axis gives the longitudinal photon. Vibration of the time-energy component along the t- axis gives the scalar or time-polarized photon. We assume a similar four polarizations for EM waves in 4-space.[291]

10.2 Mechanism Generating Flow of a Mass through Time

As we covered in the referenced paragraphs in Chapters 1 and 2, the fundamental generatrix of the flow of a mass through time — and the structuring of that flow — is the total photon interaction with the mass and all its parts. This includes both observable and virtual photon interactions.

As a recapitulation: Mass *m* does not exist in time, since it is an observable and observation is the imposition of d/*dt* upon spacetime LLLT, removing the T to give a "frozen LLL snapshot." In short, mass m is an *effect* (of the iterative very high speed observation process), not a *cause* (which is an *input* to the observation process). This is why all observation is 3-spatial rather than spatiotemporal: it yields the 3-space *effect* (at a given time, of a given interaction), not the 4-space cause of the interaction and not the ongoing 4-space interaction.

---

[290] Note the relationship with such things as time dilation, which represents a broken time-flow symmetry, where net energy flow from the time domain to the 3-space domain occurs. Hence time dilation is associated with the increase of spatial energy, such as kinetic energy due to the relativistic velocity of the moving object.

[291] In this book, we have limited our discussion to four dimensions. If we model in five or more dimensions, then additional symmetries and broken symmetries **are** possible, as are new kinds of multi-dimensional EM waves.

By absorbing and emitting photons, observed mass *m* may be said to move through time in a succession of *m* => *mt* => *m* => *mt* => *m* ... jumps, each involving change of physical state as a photon is absorbed by *m,* changing it to *mt* which is nonobservable. Then a photon is re-emitted by *mt,* changing it back to *m* and completing another observation with this new *m* as the output (effect). We note that masstime *mt* is as different in nature from mass *m,* as impulse *Ft* is from force *F.* In mass *m,* every tiniest differential *dm* is also interacting with photons continuously (even though they may be *virtual* photons), hence *dm* also exhibits its own "digital flow through time" as a part of the structure of the overall flow of the parent mass through time.

Interaction primarily with the *mt* state yields wavelike interactions. Interaction primarily with the *m* state yields particle-like interactions. This is proposed as a possible resolution of the wave-particle duality problem, which has never been resolved but just accepted without further argument. Previous consideration of wave-particle duality has assumed that "mass has two states, one particulate and the other wavelike". We correct that assumption by pointing out that two different "states in which mass exists" are involved, and the two are quite different: one is mass *m,* and the other is masstime *mt.* We stress the function of the myriads of virtual photon interactions during an *mt* state, where an observable photon has not yet been re-emitted. These virtual photon reactions create streams of virtual *dm* => *(dm)dt* => *dm* => *(dm)dt* =>. . . . .on all the differential parts of mass *m,* which creates the rich structuring inside the "time-flow" of the overall observable mass *m* through time. The rich dynamics of the *dm* -> *(dm)dt* streams also accounts for the seeming existence of mass *m* as a set of 4-dynamics which interpret as a set of waves. We note that the emitted photon carries with it the internal structuring of the spatial energy domain and the time-energy domain. Herein we have the basis for the Kaznacheyev's demonstrated electromagnetic induction of cellular damage and disease at a distance.[292]

Finally, we tentatively assume that, prior to its interaction, the antiphoton carries negative spatial energy and positive time increments, while the photon carries positive spatial energy and positive time increments. In a very small region where an excess of antiphoton interactions occurs temporarily, time-reversal occurs and the physics "runs backward". It has recently been experimentally proven that such "reversing" of the physics processes occurs up to micron size and for times up to two seconds.[293]

---

[292] Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian|, Novosibirsk, 1981.

10.3 Time-Polarized EM Waves and Longitudinal EM Waves

10.3.1 Waves, Photons, Spacetime Curvature, and Observation

Spacetime (ST) curvature involves not only the change of *spatial* energy density of local spacetime but also a change of its *time-energy* density. That component of the total ST curvature due to time-energy density change — i.e., time-polarized EM photons and waves — can be far more powerful, because of the enormous energy density of the time-energy compared to the far lies dense 3-spatial energy. As previously stated, time may be regarded as having the same energy density as mass, except that the highly compressed 3-spatial energy is moved to the 4th Minkowski axis to constitute "time" rather than being left in 3-space to constitute "mass".

Hence ST curvature can be made very powerful when one uses longitudinal phase conjugate EM *wavepairs,* as shown by Whittaker {600} in 1903 with slightly corrected interpretation as argued by the present author {615}. Each longitudinal EM wave in 3-space is accompanied by a precisely coupled time-polarized (scalar) EM wave in the time domain, which latter wave is a longitudinal EM wave on the fourth Minkowski axis involving changes in the t variable in the axis *-ict.* We previously discussed the giant negentropy that this represents with respect to the source dipole. At various times in the past, we have emphasized the time-domain aspects of the coupled phase conjugate longitudinal EM wave-pair by calling such a coupled wave by terms such as *time-density EM waves, time-polarized EM waves,* etc. These terms imply additionally altering and structuring the *time density* and *time-energy density* of spacetime rather than merely altering and structuring its *3-spatial energy density.* Such time-density induced ST curvatures may be nearly 10^{17} times stronger than ordinary spatial energy density ST curvatures {601}.

Interestingly, at least one researcher — Matsumoto {602} — has realized the basic implication in cold fusion processes of using the strong EM force as the agent of spacetime curvature. He has shown consistency of the postulated EM-induced ST curvature approach with some important cold fusion results. However, Matsumoto has utilized the conventional interpretation of spacetime curvature due to alteration of *3-spatial energy density* of spacetime. This misses the primary time-density functions, but does include the resulting spatial-energy functions after transduction from time-energy to spatial-energy. His work, however, does support the involvement of a more intense form of energy in the cold fusion effects.

---

[293] See footnotes 287, 288, and 289.

There are intriguing but still mysterious relationships to be noted between cold fusion and deep astrophysical mechanisms. Continuing reference is made to these similarities, including black holes and white holes. Black holes and white holes involve spacetime curvatures — extreme, but still spacetime curvatures. So one would suspect that the entire scale of spacetime curvature — including black holes and white holes in the limit — should be of interest to cold fusion researchers, since in their use of time energy they are inadvertently involving more intense forms of energy than ordinary spatial EM energy.

Most nuclear scientists and even some relativists seem unaware of the much more powerful spacetime curvature effects obtained by alteration of the *time density* of spacetime rather than just its *spatial energy* density. Physicists have largely omitted specializing on time density ST curvatures because they have largely omitted the absorption and emission interactions of the time components transported by photons and EM waves.

Longitudinal EM phase conjugate wavepairs (each wavepair a coupling of a time-polarized EM wave and a longitudinal 3-space EM wave) of the type shown by Whittaker {600} in 1903 involve net *time density oscillations* as well as 3-spatial energy density oscillations, according to our reinterpretation. The infolded "inner EM content" of all conventional EM potentials, fields, and waves thus contains extraordinary dense time-energy, normally lost in the *d/dt* observation interaction.

In short, the infolded "internal Whittaker electrodynamics is a pure "general relativistic electrodynamics" and a unified field dynamics. Further, it can be engineered, which leads to startling new physical processes previously thought to be impossible and forbidden by the laws of nature.

Prior to observation, the internal highly compressed time-energy structures associated with the interior EM structure of all conventional EM potentials, fields, and waves are extremely powerful spacetime curvatures.

In short, all "conventional" electrodynamics has an internal structure of strong spacetime curvatures and dynamics.

Organization and patterning impressed on, or inserted in, this internal structuring can provide extraordinary actions in materials, living biological systems, cells, etc. Adding deterministic structuring inside these normal EM entities is referred to by Russian energetics weapon scientists as engineering the *information content ofthe field.* This is the approach behind deliberate engineering of *spacetime curvature engines,* or vacuum engines — or just "engines" for short.

In theory, any physical change to a mass can be engineered by inserting such "engines" into the mass. Any mass-energy is a mass-energy EM potential with an internal structuring (engine) for all the mass's internal dynamics. This engine is referred to as the *resident engine* of that mass (and all its dynamics). Introducing to a given mass a potential with a desired "delta engine" structure, results in the superposition of the two potentials, resulting in exchange sharing of the engines. In that manner, a desired "delta engine" diffuses into the mass's potential structure and its resident engine, changing it to a new resident engine which now acts on the mass at every level to change it. Potentials superpose, and their engines diffuse and superpose as well. Direct engineering of matter — even inside the nuclei and inside the nucleons themselves — can be performed by introducing the desired engine or engines for the changes desired.

This area has been highly weaponized by one or more nations. E.g., the Russians — in several decades of so-called "microwave radiation of the U.S. Embassy in Moscow — induced health changes and diseases in U.S. personnel in the Embassy by inducing deliberate "disease engines". Such EM-induced diseases and health changes were generated in the zones where the potentials were field-free, and therefore stable for stable engine diffusion.

Note that the common scalar potential decomposes by Whittaker's 1903 decomposition {600} into the longitudinal EM wavepairs we are speaking of. Also note that, per a second paper by Whittaker in 1904 {91b} that initiated superpotential theory, all EM fields and waves can be decomposed into two scalar potential functions. Each of those base potentials involved in the potential functions further decomposes by Whittaker 1903 into a harmonic set of phase conjugate longitudinal EM wavepairs as previously discussed. Therefore, the infolded EM inside all the "conventional" EM fields, waves and potentials is indeed a highly organized set of such bidirectional longitudinal EM wavepairs and thus a highly organized set of extreme localized ST curvatures. In other words, it is comprised of these special "engines" of deterministic spacetime curvatures and their dynamics.

Constructing the potential by assembling one's own set of phase conjugate longitudinal EM wavepairs to form it, one can insert any kind of ST curvature set — any kind of vacuum engine desired — inside the base potential(s), hence inside the two potential functions comprising a given set of EM fields and waves utilized as "carriers". The specific internal structure of a "carrier" radiation field or potential designed to carry such engines is called the "information content of the field".[294] The Russians have long had a disinformation campaign to suggest that the term refers to conventional spectral analysis. It does not.

Johns Hopkins scientists did detailed measurements inside the Embassy and confirmed that all the anomalous health changes and diseases had occurred only in areas where the EM fields B and E were absent. That is, the health changes occurred only where the potentials were field-free and therefore stable and unchanging, so that an exposed body was immersed in a steady sea of "vacuum engines" being diffused throughout the targeted body's EM potentials in and on its cells, its cellular membranes, etc.

In general relativistic terms, every physical set of changes and altered dynamics (syndrome) of a specific disease has a specific *delta* set of spacetime curvatures and dynamics associated with it, which interacts on the body, its chemistry, its electrodynamics, its cells, etc. A desired set of

---

[294] Indeed, in theory one can *construct, functioning robotic systems* of these internal engines. In clandestine Russian weapon research, it appears that at least elementary "causal robots" of such internal engine type have been developed and tested. In a sense, these are the ultimate "nanobots" made of infolded longitudinal EM waves and their dynamics. In theory, command and control functions (via longitudinal EM wave communication) can be included, as can certain kinds of weapons effects based on scalar interferometry. The eventual development of such startling weapons seems intended to produce an army of eerie "internal EM" robots which can be "sent in'" *through the interior of EM fields, waves, and potentials,* to targets anywhere on earth, beneath its surface, in space, or beneath the ocean. This may be intended to become the ultimate robot or "high tech" warfare, and it is also the ultimate "information" warfare. There are no shields against this type weaponry. Defense, however, is accomplished by first using longitudinal wave "scanning" to pick up and track the systems, much like radar picking up aircraft. Once detected, scalar interferometry is utilized to disrupt and destroy them in place. Progress on these robotic weapon systems appears to be well along, as does progress on their countering longitudinal EM wave weapons.

ST curvature deltas for a given disease was inserted inside the base potentials in a combined Whittaker 1903 and 1904 decomposition of the stable field-free potentials in the Embassy. The Russians thus induced the desired engines for induction of the specifically desired disease(s) in the assigned personnel in those field-free potential zones in the Embassy. The purpose was to use the tests as (i) a guaranteed means to provoke high level U.S. governmental attention, and (ii) a probe to see if the U.S. knew about such technology and had developed defenses. Other tests of such disease induction in entire populations have also been conducted, but these are beyond the scope of this book.

We point out one very important thing: the cells of the body have their normal "resident engines" and resident engine dynamics inside their internal fields and potentials and waves. For any specific disease or disorder condition, there exists a specific associated change (delta) in the normal set of vacuum engines and their dynamics, in the cells and the body. In short, there exists a specific "delta engine" for any cellular disease or disorder, added to the normal engine.

It is not the immune system that heals. Instead, it is the cellular regenerative system. The natural healing method used by the cellular regenerative system is not chemical. The regenerative system uses the phase conjugate wavepair composition of increased scalar potentials (on cellular membranes and every part of the cell) to "pump" the cells in the time-domain. This creates an "anti-engine" for the specific disease delta or disorder, and it slightly amplifies this anti-engine within the limited capabilities of the body. This action places this amplified anti-engine also into the stricken cells' engine complex. Engines superpose, just as do potentials. This action by the regenerative system has the effect of slowly producing complex anti-forces in the body and inside the cells themselves, eliminating the detrimental engine delta, restoring the normal resident engine and thus time-reversing the stricken cells back to a previous healthy condition. It does not kill pathogens at all; that is the function of the immune system. In the 1960s and early 1970s, Antoine Priore in France demonstrated remarkable cures of terminal tumors, infectious diseases, clogged arteries, and other disorders in laboratory animals,[295] unwittingly amplifying the body's own master regenerative mechanism. His work was suppressed in the mid-70s.

---

[295] See (a) A. Priore, *Guerison de la Trypanosomiase Experimental Aigue et Chronique par L 'action Combinee de Champs Magnetiques et D 'Ondes Electromagnetiques Modules* [Healing of intense and chronic experimental trypanosomiasis by the combined action of magnetic fields and modulated electromagnetic waves], Priore's thesis submitted in candidacy for the doctoral degree, 1973. The thesis was rejected when the project was suppressed. See also (b)

Check our website, www.cheniere.org, for a 119-slide "porthole" briefing on the present author's extension whereby specialized mixes of ordinary EM waves can be used to initiate and perform this "delta engine elimination" healing process, and it can be highly amplified well beyond the body's normal healing capabilities.

In 1998 we seriously proposed to the DoD, the USAF, the NIH, and other U.S. government institutions that a crash development program be initiated to develop portable units capable of rapidly (5 minutes per patient) treating mass casualties from coming terrorist attacks with weapons of mass destruction such as anthrax, smallpox, bubonic plague, Ebola virus, etc. Sadly, no one in the U.S. government had the foggiest notion of what we

---

"Le Probleme Priore," *Rapport de la Commission de 1'Academie des Sciences a Monsieur le Ministre d'Etat charge de la Recherche et de la Technologie,* 1982, p. 1-22; (c) Antoine Priore, "Precede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963; (d) Antoine Priore, "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; (e) Antoine Priore, "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968; (f) A. J. Berteaud and A. M. Bottreau, "Analyse des rayonnements electromagnetiques emis par l'appareil Priore," [Analysis of the electromagnetic radiations emitted by the Priore apparatus], D.R.M.E., 1971, p. 3-12.; (g) R. Courrier, "Expose par M. le Professeur R. Courrier, Secretaire Perpetuel de L'Academie des Sciences fait au cours d'une reunion a LTnstitut sur les effets de la Machine de M. A. Priore le 26 Avril 1977," [Presentation by Professeur R. Courrier, Perpetual Secretary of the Academy of Sciences, made at the meeting of the Academy on the effects of the machine of M. A. Priore]; (h) R. Pautrizel, A. Priore, M. Dallochio and R. Crockett. (1972) "Action d'ondes electromagnetiques el de champs magnetiques sur les modifications lipidiques chez le Lapin par l'administration d'un regime alimentaire hypercholestrole." [Action of electromagnetic waves and magnetic fields on provoked lipidic modifications in the rabbit by the administration of an alimentary hypercholesterol diet.] *Compt. Rend. Acad. Sci. (Paris),* Vol. 274, 1972, p. 488-491. The same university that rejected Priore"s doctoral thesis in 1973, accepted a doctoral thesis on the same subject some 11 years later due to the persistence of Pautrizel: See (i) Eric Perisse, *Effets des Ondes Electromagnetiques et des Champs Magnetiques sur le Cancer et la Trypanosomiase Experimentale [Effects of Electromagnetic Waves and Magnetic fields on Cancer and Experimental Trypanosomias],* Doctoral thesis, University of Bordeaux No. 83, March 16, 1984.

were discussing. At NIH, we never got out of their "policy" (i.e., "spin control") department. Not a single capable scientist reviewed the extensive package. The attack on the World Trade Center and the Pentagon on Sept. 11, 2001 is just a forerunner of that which is to come. We could have been prepared to save hundreds of thousands of Americans, even millions. We shall not be ready, and the first really professional anthrax or smallpox (or camel pox, which is easily available and almost exactly the same as smallpox) attack on a major U.S. population center may leave from one to three million Americans dying in the streets and in their homes. Presently nothing will save them. It could have been different.

10.3.2 Photons and Waves in Minkowski Spacetime

Photons are made of angular momentum (action); i.e., of *energy x time.* Waves, being made of photons, carry both spatial (decompressed) energy and time (highly compressed) energy — although physics has largely omitted the *time-energy-carrying* aspect, and only accounted for the *spatial energy-carrying* aspects. An EM wave in space is actually a wave of the particular spacetime curvatures that produce angular momenta in interacting charged masses. The standard notion of E and B fields at right angles in a plane in 3-space, traveling along the *z*- direction and with the plane orthogonal to *z*-, and with E and B field vectors oscillating back and forth, is nonsense, as acidly commented upon by Romer {603}, the former editor of *American Journal of Physics.*

Nonetheless, using the standard way of modeling the "transverse" EM wave: In conventional electrodynamics, a *transverse EM wave* is one where the 3-spatial energy density is considered to be varying rhythmically at right angles to the direction of wave propagation, and the time-density is assumed to be non-varying, or at least altered only by special relativistic considerations in certain cases.[296] A longitudinal EM wave has its spatial energy density fixed in *x*- and *y*- directions, and its energy density is oscillating along *z*-, the direction of wave propagation, by a "compression" and "rarefaction" process much like squeezing an accordion. This, however, leads to a "galloping" motion where the wave is alternately very superluminal and very subluminal at given instants, but moving at an *average* velocity of *c* {604}. On the other hand, a *time-density (scalar) EM wave* by definition is not varying in its 3-spatial energy density, and therefore is not varying in *x-, y-*, or *z-*. Instead, the energy is in the form of time-energy, and its variation is in time density. So its energy density is oscillating along the fourth Minkowski axis, *ict*, as an oscillation of the energy density of the variable *t.* It is therefore polarized in the t- direction.

---

[296] We point out that a wave of varying spatial energy density is moving in a varying curved ST a priori, by general relativity. If the "wave" were in a flat spacetime as normally assumed, the EM flat spacetime would eliminate the EM wave itself. **This** has been particularly pointed out by Sachs and by Evans.

10.3.3 Time-Charging, Internal Structuring, Vacuum Engines, and Diffusion

Photon interaction as a *spatial* energy collection or energy exchange process is well known in physics. However, the time-charging or time-excitation interaction between the mass and the time component carried by the photon is usually ignored.

As is well known, an atomic electron that absorbs a transverse photon increases its spatial energy and changes into an excited spatial energy state. When the electron subsequently "decays" from that excited energy state to a lower state, the spatial energy differential between the two states is emitted as the spatial energy component of an emitted photon. This is the *spatial energy density* reaction of a transverse photon with an absorbing and emitting mass particle.

Unaccounted for by physics, at the time of the energetic excitation of the mass (i.e., when the photon is absorbed), a time excitation — time charging — also occurred because the time-component of the absorbed photon is also attached to the mass, converting it to *masstime,* a new entity. Re-emission of the photon is accomplished by the decay of both the *spatial energy* excitation and the *time-energy* excitation. In the previously excited state after photon absorption, however, one did not have the entity "mass", but the entity *masstime* having greater topology. If mass is an observable and taken to be 3-spatial (as is conventional), then photon absorption and emission oscillates between 3-space and 4-space effects, generating both spatial energy and time-energy (or *angular momentum)* excitation and decay.

Further, only *masstime* emits a photon. This immediately casts a very different light on such things as radioactive elements, which are in an initially excited state. Discussion of this aspect, however, is beyond the scope of this treatise.

On the other hand, a *scalar (time-density)* photon {605, 606} has a rather different interaction, regardless of how physics presently treats it. Any atomic electron at any energy level is continually receiving and emitting virtual photons - transverse, longitudinal, and scalar- in its virtual photon exchange with the energetic vacuum. Hence it is continually "connecting to" or "absorbing" time components, thereby changing its nature from mass to masstime, back to mass, back to masstime, etc. Not only the entire electron but also each differential piece of it is performing that dance between mass and masstime. This is a continual changing from 3-space to 4-space and back. There thus is a violent "chatter" of the electron (and each differential piece of it) in the myriad levels of its *time* exchanges, giving it

(i) An average rate of time component exchange,

(ii) An average "time-charge", and

(iii) A dynamically changing, specific structuring of its time components.

Whenever a mass absorbs an observable photon, the mass converts to masstime, with an "extension in time" as well as extension in 3-space. During the time-interval created in the masstime, myriads of virtual photons are absorbed and emitted by that mass component of the masstime entity. Hence every dt portion of the masstime state after observable photon absorption, is "internally structured" by the entire virtual photon interaction set's At's occurring within that time interval.

By tailoring at least the larger elements of this internal structuring of the masstime state's virtual photon bombardment, by a deliberately induced virtual photon bombardment, the emitted photon from the masstime decay will carry away with it a desired internal "vacuum engine" structure of organized curvatures of spacetime. In short, two potentials superpose; that is one of their primary features. When they superpose, they also diffuse their internal structures (internal vacuum engines) into each other.

We speak of the *diffusion ofa specific vacuum engine* placed in one interacting EM potential, field, and or wave as *dimensioning* the potential, field, or wave. Adding a dimensioned potential, field, or wave to other potential(s), field(s), or wave(s) diffuses the dimensioning (the engines) between all of the participating entities. With steady radiation of such dimensioned signals into a targeted object, field, wave, or potential, the dimensioning is gradually created (kindled) in the target, simply diffusing throughout it and gradually accumulating. If the irradiating signal carries a specific disease — e.g., necrotizing fasciitis — then gradually in the irradiated body there will emerge all the symptoms of necrotizing fasciitis. If the irradiating signal carries a specific disease antiengine — e.g., for necrotizing fasciitis — then gradually the necrotizing fasciitis condition will be eliminated, with the disease organisms also being eliminated.

This is a more fundamental extension of the disease vector concept. Not only can a disease vector be induced by a physical organism such as a virus or bacterium, but it can also be introduced by simply capturing the fundamental vacuum engine pattern of the disease and using it to condition or dimension signals used to steadily irradiate the targets.

This is the explanation of the Kaznacheyev experiments {607}: thousands of experiments performed by Vlail Kaznacheyev and his colleagues demonstrated the induction of almost any kind of remote specific cellular death and disease patterns in targeted cells, by absorption of radiation emitted from appropriate dying or diseased "emitter" cells with that condition, if the experiments were performed in the dark.

Instead of looking for biogenesis causes in amino acid precursors to cellular life in meteorites etc., one might well look into the *vacuum engine structures* inside the light from the sun, the planets, the stars, and the universe in general. E.g., dead sperm can often fertilize female cells, if not too much physical damage and deterioration to the dead sperm cells has yet occurred. This is understandable, since for a time a dead but undamaged sperm cell will have the same resident engine as a living sperm cell.

10.3.4 Some Cosmological and Gravitational Effects

Each of the observable-sized dt's in all the serial masstimc excited states of the mass moving through time, is internally and precisely structured. The precise time history of any mass — together with all its interactions, experiences, etc. — is present inside the time stream of ongoing interactions of that mass with photons. These virtual photon changes are in fact deterministic and come from every part of the universe, as shown by Puthoff {608} in his self-regenerative cosmological feedback cycle. We have merely extended his cosmological feedback to include the time domain as well as the energy domain. We hypothesize without further discussion that this internal structuring of the time-domain for every mass or object will eventually be connected with the quantum concept of the geometric phase {609}.

When a transverse photon interacts with a mass in the presence of other photon interactions, the mass can act as a phase conjugate mirror and form the phase conjugate replica photon (an antiphoton). For a longitudinal photon, the antiphoton *prior to observation* is a scalar photon. We propose that in the unobserved causal state the photon and antiphoton can couple (strong interpretation of the distortion correction theorem) and form a spin-2 graviton. According to quantum field theory, one may already interpret the common scalar potential as being comprised of such spin-2 gravitons {610, 611} if we speak of the situation prior to observation.

As opposed to the accounted *spatial energy charging* of the familiar transverse photon interaction, the longitudinal photon interacting with mass will cause immediate phase conjugation by the mass because of the accompanying scalar photon (accompanying antiphoton prior to observation). When one pumps with longitudinal EM waves, one also inadvertently pumps with time-polarizcd EM waves simultaneously.

From its background "time-energy pumping" in its vacuum exchange, the mass forms the phase conjugate replica, which couples to the incident longitudinal EM photon. This forms (wave view) bidirectional longitudinal EM wavepairs (conjugate pairs), comprised of (particle view) internal coupled longitudinal photon-antiphoton pairs. Each of those internal photon-antiphoton pairs is a spin-2 graviton.

Each such conjugate longitudinal wavepair actually contains a time-density (time-polarized) or scalar EM wave. So the interaction of LWs and their accompanying scalar waves immediately initiates the formation of time-density waves, which then primarily perform a *time-charging* interaction. In addition, even for an equilibrium time-charging state, the "masstime" state of the interacting electron contains a specific structure of the time elements in the electron-time's time component. Each of these infolded time elements is also time-charging the electron.

10.3.5 Charging and Excitation Considerations For Cold Fusion Purposes

So there are several new types of charge or excitation that are involved in cold fusion and other phenomena. One may speak of charging and excitation such as gravitational charging, time charging, time-energy excitation, time-energy charging, etc.

Both the longitudinal photon and the scalar (time-polarized) photon are in fact known in physics, but usually neglected, at least at the end of calculations. E.g., Gray {612} puts it this way about the four kinds of photon polarizations and the habitual elimination of the higher ones:

> *All that quantum theory can say is that there are two transverse photons, a longitudinal photon and a scalar photon. It must be decided by other means which are "desirable" states and then it is customary to "eliminate" the longitudinal and scalar photons by invoking a*

> *subsidiary condition, now somewhat weakened from the Dirac form.*

For our purpose in examining the cold fusion experiments, there exists a completely unsuspected "time-charge" set of excited states for an atomic electron, a proton (as in an H+ ion), etc.[297] There also exists a "time-structuring" of the time-charge excited time-state, which we will see is important in certain highly anomalous instrument phenomena. Uncovering this time-charging and time-structuring interaction has been one of the major results of this author's long effort to decipher the functioning of the healing (cellular regeneration) system of the body as well as his efforts to decipher the fundamental nuclear transmutation mechanism in cold fusion transmutation interactions at feeble *spatial* energy.

In summary: For longitudinal photon interaction, the atomic electron "absorbs" the longitudinal photon and its accompanying scalar photon, being simultaneously spatial-energy excited (charged) by the longitudinal photon and time-energy excited (time-charged) by the scalar photon. This time-excited state (of masstime) then subsequently decays, emitting a longitudinal photon accompanied by a scalar photon in the process.

---

[297] See again footnotes 287, 288, and 289. From our mechanism propelling a mass through time, time-charging and discharging play the causal role. The acquisition of dt by mass *m* produces masstime *mt,* then the subsequent photon emission decay of *mt* back to *m* produces "propagation of mass *m* forward through time" by one "jump". Time charging with -dt and subsequent photon emission is what produces "propagation of mass m backwards through time" by one "jump". Evans et al. proved that even up to the micron scale physical particles and their dynamics can run backwards in time for up to two seconds or more. The fluctuation theorem proved by Evans and his colleagues in 1993 shows us that, at the scale of a proton and neutron, there can become a high probability of significant time reversal of the physical dynamics. This means that the law of attraction and repulsion of charged particles -such as two H+ ions in solution — "runs backwards" or is reversed, so that momentarily like charges attract and unlike charges repel. We have called the region in which this reversing of the physics occurs a *time reversal zone* (TRZ). With the reccentwork of Evans et al., the existence and occurrence of TRZs even well above the proton (the H+ ion) level is now clearly established, including the reversing of the physical dynamics. These concepts have not yet been applied to cold fusion interaction results by the leading cold fusion researchers, but they now stand as explaining the most probable mechanism for the observed phenomena. In this Chapter we have written some of the typical new nuclear reactions that can occur in such a TRZ, and the exact products of these reactions are widely reported in several hundred successful cold fusion experiments.

We shall return to time-charging and time-charge decay when we examine and explain the odd instrumental anomalies experienced for some time in experiments at China Lake.

10.4 Time Reversal Zone and New Nuclear Reaction Mechanisms

It follows that the flow of a mass through positive time must involve a preponderance of reactions with photons rather than antiphotons, since *net* absorption and re-emission of photons (rather than antiphotons) is involved. Otherwise, the absorbing and emitting mass would not accomplish little "jumps" forward in positive time.

Suppose we deliberately arrange a situation where the target mass is interacting with a preponderance of antiphotons. In this case, the mass will be time-reversed, as will be the EM energetics.[298] This leads us to the novel concept of a *time reversal zone* just described in footnote.

A time-reversal zone (TRZ) is a region of space, or electrolyte, etc. in which given masses or charges of interest are interacting on the average with more antiphotons than photons. Consequently, in the TRZ the usual EM energetics are reversed and normal charge reactions appear to "run backwards" insofar as the spatial 3-space observer is concerned.

For example, in electrolytes in a region of highly loaded positive charges in a loading palladium lattice with properly prepared surface, a very great number of double surfaces exist. Hence a great number of scalar interferometries continually occur from that large number of double surfaces acting as scalar interferometers {613}. Because of the highly excess positive charge loading, these interferometers are predominately fed by negative EM energy from the positive charges. Hence, in the distant interference areas in the electrolyte outside the palladium lattice, some negative energy fields and negative energy potentials are produced dominantly. An excess of antiphotons is produced in those interference zones where an excess of antiphotons appear due to the scalar interferometry. In *those* little interference zones, the simple positive ions are momentarily bathed in an excess of antiphoton interactions. This is therefore a time-reversal zone of momentary time-reversed EM energy flow.

---

[298] This *is not* time-travel in the classical science fiction sense. For time-travel, the traveling object must remain moving in its own forward time, while the entire remainder of the universe must be reversed in its time and must move backward in time to a past coordinate. That is not what is happening here, and no one is suggesting we can time-reverse the rest of the universe! Instead, energy can be reversed in time, as proven in nonlinear phase conjugate optics. So can mass-energy and charge-energy, as shown by the Dirac theory of the electron. An *observed* positron is an electron observed while traveling backwards in time, so to speak, with respect to the observer. But it is observed in observer forward time as traveling in the opposite spatial direction, having positive mass and positive energy, and with the sign of its charge reversed from negative to positive.

From a palladium lattice loaded with H+ or D+ ions, and also having a proper surface with many work grooves and thus many small interferometers, random fluctuations in the scalar interferometry occur in the adjacent electrolyte. Some of these random scalar interferometries continually form fleeting *time reversal zones* (TRZs) in their interference zones in the surrounding electrolyte. The very large energy required for the formation of each TRZ and its reversal of the normal laws of attraction and repulsion of charges, is available from conversion of time energy from the time domain as a result of the giant negentropy mechanism in 4-space ongoing in the loaded positive charges (614, 615} in the palladium lattice.

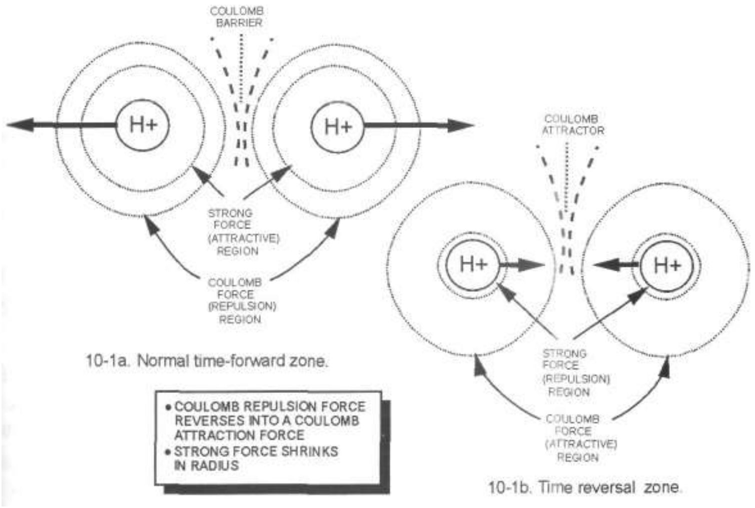
COULOMB BARRIER
H+
H+
STRONG FORCE (ATTRACTIVE) REGION
COULOMB FORCE (REPULSION) REGION
COULOMB ATTRACTOR
H+
H+
STRONG FORCE (REPULSION) REGION
COULOMB FORCE (ATTRACTIVE) REGION
• COULOMB REPULSION FORCE REVERSES INTO A COULOMB ATTRACTION FORCE
• STRONG FORCE SHRINKS IN RADIUS

10-1a. Normal time-forward zone.

10-1b. Time reversal zone.

Figure 10-1 Forces on nuclei of simple ions in time-forward and time-reversal zones.

See Figure 10-1. In Figure 10-la, two hydrogen ions (two free H+ protons) in a normal electrolyte and in a normal time-forward zone are shown. The

Coulomb barrier dramatically increases the forces of repulsion between the two H+ ions as their kinetic energy may be driving them momentarily toward each other. This Coulomb barrier becomes so strong that it stops the approach of the two ions and forces them back apart, or to deviate aside from their paths, before each can ever enter the very short-ranged strong force region of the other. Consequently no nuclear reaction occurs, but only a common chemical reaction. The maintenance of that Coulomb barrier is all that prevents energetic ions from being driven together closely enough to engage the strong force and cause the ions to form a new nucleus (cause a transmutation).

Now see Figure 10-lb, for the same two ions that suddenly find themselves in a momentary time-reversal zone (TRZ). In a temporary TRZ, suddenly like charges attract and unlike charges repel, exactly in reverse of the *normal* behavior of charge attraction and repulsion. The usually increasing Coulomb barrier (repulsive force between the two approaching H+ ions) has disappeared and been replaced by an increasing Coulomb attractor (attraction force). Further, the strong force has been partially reversed and much weakened since the gluon forces are dramatically reduced and fluctuating.[299] The attractive strong force is now a partially *repulsive* force and so it is much weaker. Consequently, the quarks in a proton or neutron are not nearly so strongly bound as they are in a normal time-forward zone (TFZ).

A TRZ represents a highly time-charged excited local state in that local region of the electrolyte. The moment a TRZ is formed, the surrounding ions in the electrolyte outside the TRZ immediately move or deviate their movements to reverse this TRZ action[300] and convert it back to a time-forward zone (TFZ). Hence once it is born in the electrolyte, the TRZ decays rapidly due surrounding ion movement changes, and even more

[299] Again we stress the tremendous energy density of the time-energy involved. These cold fusion interactions are actually higher *total* energy reactions than nuclear physics presently uses, even though the *spatial* energy density of the reactions is very small. Together with the now-proven fluctuation theorem of Evans et al., the TRZs and reversal of the normal Coulomb barrier into a Coulomb attractor open up a vast new nuclear chemistry of direct nuclear transmutation reactions at low spatial energy but at very high tempic energy. The ability to directly engineer the quarks and the gluon forces, e.g., opens up a new chemistry of direct subparticle engineering heretofore considered quite impossible by chemical means.

[300] Their approach increases the fraction of photon reactions relative to antiphoton reactions, since the other ions are moving in a TFZ on the average.

rapidly due to the rapid variation in the fluctuating scalar interferometers in the loaded palladium lattice. In most cases, the random fluctuation in the scalar interferometry is so rapid that the TRZ decays back to a normal TFZ before the two approaching H+ ions can reach each other (reach each other's reduced strong force repulsion region). However, in a certain percentage of approaches, the two approaching like-charged ions will "reach each other" — i.e., each will enter the weakened strong force region of the other before decay of the TRZ occurs.

See Figure 10-2. In Figure 10-2a, we show the case in which the TRZ lasts long enough for the two H+ ions to reach each other's strong force repulsion zones and form a quasi-nucleus. Here they vibrate back and forth in a dynamic dance around the zero net force axis between their repulsive strong forces and their attractive Coulomb forces. In this case, a new kind of nuclear reaction is set up to occur. Decay from this excited quasi-nucleus state in a TRZ can occur into a real nucleus in a normal TFZ. Energetically, because the quarks are so loosely bound now[301] the preferred decay mechanism when TRZ -> TFZ is by appropriate *quark flipping.*

As a little bit more time passes, the interferometry changes and the TRZ decays, returning back to a normal TFZ. In a change from TRZ to TFZ, the energy change is enormous because time-energy is involved, with energy density equal to mass. The only difference in a proton and a neutron is the orientation of one quark (three quarks make up each of the particles). As TRZ decay starts and progresses, the strong force changes back to a powerful attractive force. It increases its strength far more powerfully and quickly than the Coulomb attractive force reduces its reversal and changes back to a repulsive force. The addition of the extra energy from the now-increasing strong attraction force simply causes one nearly-freed quark in one of the two H+ ions to flip, converting that H+ ion into a neutron n. Instantly the preferred decay product is the deuterium ion D+. So the H+ and the n simply draw a bit closer together, each now fully inside the rapidly increasing attractive strong force region of the other. This converts what had been a quasi-nucleus of two H+ ions into a real D+ nucleus. This result is shown in Figure 10-2b.

---

[301] To appreciate the sheer raw power of the time-energy control of interactions, simply examine the enormous *spatial* energy density required in high energy physics to try to approach the "free quark" condition. Enormous and very expensive *spatial energy* accelerators are used, at the absolute state of the art. By going to the use of the extraordinarily dense time-energy, quarks can readily be freed or nearly freed in electrolytes, and they are nearly freed in more than 600 relatively simple and inexpensive cold fusion experiments {753) on the bench-top to allow new nuclear reactions by quark flipping, where a proton turns into a neutron or vice versa.

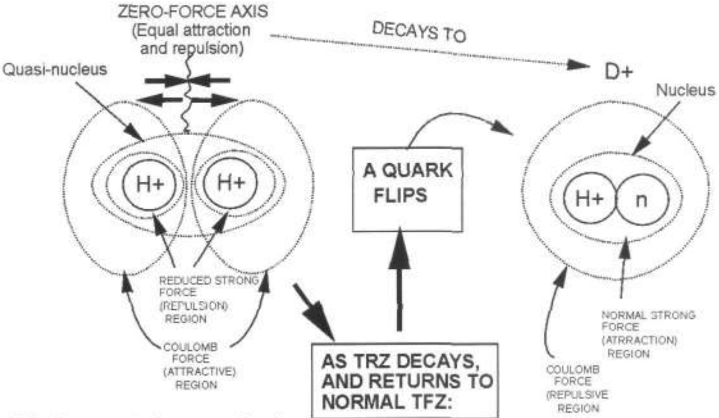
ZERO-FORCE AXIS
(Equal attraction
and repulsion)
DECAYS TO
D+
Quasi-nucleus
Nucleus
H+
H+
A QUARK FLIPS
H+
n
REDUCED STRONG FORCE (REPULSION) REGION
COULOMB FORCE (ATTRACTIVE) REGION
AS TRZ DECAYS, AND RETURNS TO NORMAL TFZ:
NORMAL STRONG FORCE (ATRRACTION) REGION
COULOMB FORCE (REPULSIVE REGION

10-2a. Ions move to form a quasi-nucleus.

10-2b. Quasi-nucleus decays by quark flipping into a new nucleus.

Figure 10-2. Production of quasi-nucleus and its decay by quark-flipping into a new nucleus as the time-reversal zone decays back to a time-forward zone.

Recapitulating: As the TRZ subsequently decays back to a normal time-forward zone (TFZ), the new quasi-nucleus becomes an excited state, and decays. However, the quasi-nucleus decays by very novel means. Because of the time reversal, the energy changes induced in the decay *start* at each spacetime point inside the quasi-nucleus, deep inside the quarks, and proceed outward. The first interaction of the decay mechanism is with the quarks comprising the nucleons (in this case, the protons). With the gluon forces still very much weakened, quark flipping becomes the preferred decay mechanism. Hence one quark in one proton flips its orientation (that is the mechanism of decay!) and the nucleus — now a nucleus comprised of a proton and a neutron — becomes a nucleus of deuterium.

As can be appreciated, the clustering of different types of like-charged but relatively simple positive ions in TRZs in solution, with subsequent decay of the TRZ into a TFZ, initiates a revolutionary new family of nuclear reactions at low spatial energy (but very high temporal energy), completely contrary to, and not included in, the present forward-time high-spatial energy (but very low time-energy) reactions model known in particle physics.

In the future, as these new time-energy-based reactions are extended and mastered, scientists will simply assemble desired nuclides in solutions in the laboratory, at low spatial energy (but high time-energy) and at will. We point out but do not pursue further the fact that these new reactions also open up cheap, practical, electrochemical means of altering nuclear and chemical wastes. We therefore foresee a dramatic development in this area, in order to clean up the presently polluted biosphere of much of the chemical and nuclear waste contaminants.

Indeed, we propose that a very strong program in mastering these new nuclear reactions be launched and supported by the U.S. Department of Energy, so that the nuclear wastes storage problem can be permanently solved, much more cheaply, and without having to store hazardous nuclear wastes for centuries. Instead of merely *talking* about "out-of-the-box thinking and research", the DoE should actually do some of it, over the violent objections of the entrenched conventionalists of the national laboratories. Their own great national laboratories are the major part of the problem, not the solution, for both innovative nuclear reactions and innovative energy systems! The problem is that "in-the-box" entrenched conventionalists do not perform or allow "out-of-the-box" research that strongly threatens their vested interests.

The TRZ decay conversion of two protons into a deuteron — by easily flipping one quark in one proton — is an example of the so-called "nuclear reaction at low energy" that has been so controversial *to* the orthodox scientific community, even in the face of some 600 successful cold fusion experiments. Contrary to the assumption of the conventional physics community and the skeptics, these are not "low *total* energy" physics reactions at all. Instead, they utilize energy density on a level commensurate with that achieved in the largest accelerators available and even much higher. *This is actually a much higher total energy physics than the present conventional high energy physics heretofore known or used.*

Particularly in a deuterium-enriched (deuterated) electrolyte, a variety of simple ion arrangements of D+ ions and H+ ions can occur. Many of these combinations and arrangements can and will occur in TRZs if sufficient loading of the palladium lattice is achieved. Below we will present and explain a few of these new nuclear reactions that have occurred in successful cold fusion experiments. First, to be tidy we must correct the present statement of the conservation of energy law by extending it to include time-energy.

The excess heat usually experienced in the electrolyte is explained by the added negentropy (added energy) from the time-domain, which then dissipates randomly in the solution, producing excess heating. Later we will cite strong experimental evidence for the involvement of such time charging (time-energy charging), with subsequent time-charge (time-excitation) decay as ordinary photons in cold fusion experiments.

10.5 Revision of the Conservation of Energy Law

10.5.1 Unaccounted Transduction Extends the Conservation of Energy Law

Along with the re-emission of all the previously absorbed spatial energy, all the previously absorbed time-charge of the excited electron may be re-emitted with a longitudinal photon/scalar photon pair (a graviton) when the time-excited state decays. In that case, no energy-charging permanent change results to the electron's original energy state. Hence no "electron wiggle" will be created and our normal "electron wiggle" detectors performing such non-transducing longitudinal EM wave interactions will not exhibit an indication. In short, the "pure" time-charging EM LW interaction is unobservable to present instruments. All they "see" is the "same electron" sitting there and "flowing through time".

However, if two or more such gravitons are simultaneously emitted from various reactions in time-decays, the emitted scalar photons and waves may interfere at some little distance. In that case, transverse EM waves and ordinary EM energy will be produced in the interference zone {613}. To the external observer, these EM fields and forces will simply seem to arise spontaneously and emerge from every point in spacetime inside that interference zone, moving outward from there in conventional "propagation through space" manner. In that case, the normal "photon" (actually, graviton) decay processes plus scalar interferometry have resulted in transducing some time-energy into ordinary spatial EM field energy. Such reaction provides the excess heat demonstrated in successful cold fusion experiments.

This is also a true action at a distance reaction, presently unaccounted in particle physics. Action at a distance is required in quantum mechanics, and it has been experimentally demonstrated multiple times in distant photon correlation experiments. Once any action at a distance is admitted because it has been experimentally proven, there is no logical restriction that only a single mechanism exists for it. Consequently, many of the present foundations assumptions of physics and some of the things previously abandoned because of their conflict with those assumptions will have to be re-examined {616}.

Scalar interferometry is indeed a mechanism for action at a distance. While still mostly absent from open physics, it has already been weaponized by some 10 nations of the world. Indeed, U.S. Secretary of Defense William Cohen confirmed in 1997 that (novel) electromagnetic processes are already being used to initiate earthquakes, stimulate volcanoes into eruption, and engineer and control the weather {617}. Whether or not it is present in the particle physics texts, scalar interferometry is being widely used clandestinely in several armies today. It has also been used to shoot down missiles and aircraft as tests. The so-called "Cold War" was not quite as cold as the news media and U.S. scientific community were led to beleieve.

Because of the ubiquitous involvement of unaccounted time-energy in all photon interactions, an extension to the present form of the conservation of energy law is required. The present (special case) law is

$$E(E)_1 + E(m)_1 = E(E)_2 + E(m)_2 \qquad [1]$$

where E(E) = energy in conventional non-massive energy form, E(m) = energy in mass form, subscript 1 refers to the measurement at time 1, and subscript 2 refers to the measurement at time 2. This is just the well-known Einstein formulation that the sum of the ordinary energy and the mass-energy is conserved.

However, that law is for a special case. It assumes that there is no transduction of time-energy into either mass-energy or ordinary energy. I.e., it assumes 3-symmetry and t-symmetry in EM energy flow. Since every charge in the universe already violates both that 3-symmetry and t-symmetry of energy flow, the fundamental assumption in the Einstein formulation is falsified by the general case. The conventional conservation of energy law is a special case where only the two mentioned energy flow symmetries are upheld. When they are violated, then transduction of EM energy between 3-space and time occurs. For that case, the conservation of energy law must be extended to

$$E(E)_1 + E(m)_1 + E(t)_1 = E(E)_2 + E(m)_2 + E(t)_2 \qquad [2]$$

The **new** law simply states that the sum of the ordinary spatial energy, the **mass-energy,** and the time-energy is conserved. This extension of the conservation of energy law becomes important in successful cold fusion experiments, where transduction of time-energy into ordinary heat energy occurs and where one encounters very anomalous reactions differing from those reactions known and accepted in conventional nuclear physics.

10.5.2 Transduction Between Time-Charging and Spatial Energy-Charging

The oscillating energy (spatial and temporal) in a moving EM wave in the vacuum represents a moving oscillation of spacetime curvature. Simply put, the so-called electromagnetic wave in space is actually a gravitational wave in space, since it is an oscillation of the local energy density of spacetime and therefore an oscillation of the curvature of local spacetime. It "transports" gravitons (scalar photon/longitudinal photon pairs) rather than merely photons. So let us now consider it as a gravitational wave.

In the absorber under special conditions, a small fraction of the graviton wave time-density absorbed will sometimes change (transduce) its time-energy into what we call transverse photon energy effects {618}, with something like a 10^{17} gain in spatial energy in joules compared to the time in seconds that actually transduces. In other words, 1 second of time transducing into EM transverse wave energy (using the standard EM modeling approach) will yield approximately 10^{17} joules of spatial EM energy. Since one can change energy from one form to another, *one can change time-energy into 3-spatial energy* and 3-*spatial energy into time-energy,* the two being merely opposite sides of the same coin, and both always involved in any situation exhibiting electrodynamic dipolarity.

In one's physics model, the choice of fundamental units is purely arbitrary. Everything can be expressed in a single fundamental unit, as pointed out by Jackson (see quotation from Jackson at the beginning of this chapter).

In a situation where 3-symmetry and t-symmetry in EM energy flow are broken, part of the *time charging* of the atomic electrons (or ions) sometimes changes (transduces) into *spatial energy charging* and vice-versa. This cross-dimensional decay of time-excitation states results in the emission of some extra transverse photons that may be either photons or antiphotons, depending on the exact interactions. The point is, *tremendous excess EM spatial energy can appear and either be emitted as heat or interacted to perform new nuclear interactions as extraordinarily high local energy density.* This is the master mechanism enabling anomalous cold fusion transmutation reactions. The amount of time transduced will be so small that it will usually escape notice, unless stringent efforts are used to measure changes in the local flow of time.

The excess photon energy (excess 3-space EM energy) appearing in these rather random but powerful transduction interactions often accounts for the production of anomalous excess heat - for example, as occurs in many cold fusion experiments. Where scalar interferometry is maximized, a more exact action will be detailed below for the specific nuclear transformations that occur by novel new nuclear reactions. These novel nuclear reactions occur at low *initial* spatial energy but with some time-energy transducing into additional very high spatial energy to initiate the interactions.

10.5.3 Resolving the Energy Conservation Problem in Cold Fusion

When transduction is involved, standard "transverse EM waves only" computations will seemingly yield violation of energy conservation — e.g., from the emission of excess heat. That is an erroneous interpretation; energy overall is conserved, but some of the unaccounted *time-energy flow* is transduced into 3-spatial transverse photon energy dissipation. Neither 3-spatial energy nor mass-energy nor time-energy alone, or in any pairing, need be individually conserved; only the combined *total energy of all three energy components* is conserved. This resolves the present major "sticking point" between the cold fusion researchers and the conventional "transverse wave" nuclear and electrodynamic communities regarding energy conservation difficulties and perpetual motion accusations. Those making the accusations and those defending against them are both using only a special-case conservation of energy law, and a more general law is required.

Let t = time in seconds, and W= energy in joules. Rounding the speed of light to 3 x 10^8 meters/sec, the energy W_{TW} emerging as transverse wave EM energy in 3-space by the transduction of time into 3-space energy is approximately

$$\text{One sec} = c^2 \text{ joules} = 9\times10^{16} \text{ joules} \qquad [3]$$

and

$$t \Rightarrow W_{\mathrm{TW}} = 9\times10^{16}\times t \quad \text{(in joules)} \qquad [4]$$

Let us present the new energy conservation law in slightly different form. Let W_{TW} = energy measured in spatial energy form, conveniently taken as transverse wave (TW) EM energy. Let W_m = energy in mass form (energy compressed by c^2). Let W_t = energy in time form (energy compressed by c^2).**The** present conservation of energy law is

$$W_{\mathrm{TW}} + W_{\mathrm{m}} = W_{(\mathrm{total})} \qquad [5]$$

The new extended conservation of energy law is

$$W_{\mathrm{TW}} + W_{\mathrm{m}} + W_{\mathrm{t}} = W_{(\mathrm{total})} \qquad [6]$$

The conventional particle physics accounting will see only

$$W_{\mathrm{TW}} + W_{\mathrm{m}} = W_{(\mathrm{total})} \qquad [7]$$

The conventional accountant will be unable to see

$$t \Rightarrow W_{\mathrm{TW}} \qquad [8]$$

so he will observe that the experimenter is reporting that

$$[W_{\mathrm{TW}} + W_{\mathrm{m}}]_1 \neq [W_{\mathrm{TW}} + W_{\mathrm{m}}]_2 \qquad [9]$$

which in his lexicon is a violation of the conservation of energy law itself. In short, he will identify it as a typical example of a claimed perpetual motion machine or interaction. Then the defender of the faith will wax eloquent, applying labels such as "perpetual motion nonsense", "voodoo science," etc. He himself is practicing a hoary old "voodoo" science and ignoring newer science. None of these defenders of the faith even know what powers an electromagnetic circuit or how it is that every little charge in the universe is able to continuously pour out observable EM energy in all directions, with no observable EM energy input.

On the other hand, a reasoned response will be given by real scientists still following scientific method, when faced with the numerous replicated successful experiments. The scientific method insists on the absolute priority of the replicable experiment over the predictions *of any* theory. Consequently, such reasonable scientists will advise that the protagonists cease their vitriolic rhetoric and let real science take its course; i.e., let further experiments confirm or refute reported successful results of cold fusion experiments.[302]

[302] Precisely such a reasoned response by chief editor Donald Kennedy of the journal *Science* has recently been given, providing a breath of fresh scientific air into an attempt by "defenders of the conventional faith" to destroy publication of important experimental results in sonoluminescence. The article which courageous editor Kennedy approved for publication — over intense objections of the professional critics — is R. P. Taleyarkhan et al., "Evidence for Nuclear Emissions During Acoustic Cavitation," *Science,* Vol. 295, Mar. 8, 2002, p. 1868-1873. We heartily commend Editor Kennedy for this classic example of how science should be conducted — so that experiments, not prevailing dogma, determine what is valid and what is invalid.

What the professional skeptic will not see is that his own expression of the conservation of energy law has been incomplete, and that it is his own inadequate "measurement yardstick" assumption that is now in error. The *extended* conservation of energy law of course still holds, but the *conventional* special case law is violated because its fundamental assumption — that time is not transduced to spatial energy in the experiments —has been violated in the experiment.

Indeed, with the publication of the work of Evans et al., experimentally proving the fluctuation theorem and the high probability of time reversal zones appearing for short but significant periods at up to micron levels, the second law of thermodynamics has been rigorously proven to be violated. Hence the conventional scientific community's rather dogmatic objection to the extra heat of cold fusion interactions is no longer scientifically justifiable.

Understanding of the extended conservation of energy law and the permissible violation of the second law of thermodynamics resolves the present thorny dissension within the scientific community over the excess spatial energy produced by cold fusion. It also resolves the issue of what could possibly be a source of such intense energy density as is required for nuclear transmutation reactions.

Any ordinary system is indeed an open system, freely and continuously exchanging energy with its ***temporal*** environment as well as the seething active vacuum.[303] If the system is in temporal energy equilibrium in that exchange, then the present restricted conservation of "spatial and mass" energy law applies and one does not have to consider the interactions between supersystem components. If the system is placed in disequilibrium with its temporal energy environment, then supersystem components interact asymmetrically and the extended conservation of energy law applies. In that case, excess spatial energy can be and will appear in the system whenever time transduction into spatial energy is occurring.

---

[303] But see our Appendix A. Present thermodynamics rather inanely defines a "closed" system as one where no mass exchange crosses its boundary with its environment, but energy exchange can and does. In that view, a "closed" system can be far from equilibrium in its energy exchange with its active environment, and thus demonstrate the five magic attributes of such disequilibrium systems. In this book we have deliberately chosen to treat a "closed" system as one in which neither energy nor mass exchange crosses the system boundary — which classical thermodynamics refers to as an "isolated" system. We explain all this (and why) in Appendix A.

The implications of cold fusion energy transduction are that

(1) Time is being "burned" or transduced as a "fuel" and source of energy,

(2) Transduction of only one microsecond per second will yield $9x10^{10}$ watts steadily. That's 90,000 megawatts of energy per second — the equivalent of 90 large 1,000-megawatt power plants — if all the transduced energy were collected and used to power loads with 100% efficiency.

(3) Thousands of new nuclear reactions are made possible and feasible by local transduction of time-energy into 3-spatial EM energy.

(4) Appropriate capture and use of a portion of the excess energy can be accomplished to power loads.

Even if we capture and utilize only 1% of the transduced time-energy to power the loads, where the *time-energy to spatial energy transduction* is only the transduction of 1 microsec/sec, that system will produce the normal EM energy to power 900 megawatts in the load. That's almost 90% of the capability of a modern 1,000-megawatt power plant. That is quite a respectable electrical power system. It only requires the transduction and "burning" of one microsecond of time per second, without any nuclear wastes or hydrocarbon byproducts.

As can be seen, transducing time into energy is by far the most efficient way to go to develop new power plants for the future. This is particularly true since every charge in the universe already transduces time-energy flow into spatial EM energy flow. Indeed, we may consider that the negative charge transduces time-energy flow into the spatial outflow of EM positive energy, while the positive charge transduces negative time-energy flow into an outflow of EM negative energy. The process is useful not only for power applications, but also for transport applications as we briefly pointed out in Chapter 8.

10.6 Some Resulting *Low Spatial Energy* Nuclear Reactions

Refer again to paragraph 10.4 above. Summarizing: The formation of time-reversal zones (TRZs) is what enables the appearance of the new nuclear reactions in the electrolyte in cold fusion experiments. It does this by changing the Coulomb barrier between like charged ions into a Coulomb attractor. Further, the rate at which TRZs form and decay in the solution is a function of the frequency and intensity of the scalar interferometry from the loaded palladium lattice. This means that the appearance of the new nuclear interactions in the electrolyte is also a function of the degree of loading of the palladium lattice, as well as the "double surface" interferometry preparation of the palladium, its cracks, etc. These are the major variables. Once the major variables provide the solution for significant formation of TRZs, then the following reactions can and do appear.

The sample reactions we present here are just a few of the thousands of new reactions possible. These example reaction products have been found in a great many successful cold fusion experiments in various laboratories around the world. Some 600+ successful experiments have now been done, and a sizable literature has built up and continues to be built up on the experiments and their results.

For our sample reactions, the primary mechanism of interest consists of several parts:

(i) Formation of a TRZ, in which two positive ions (we shall look at H+ and D+), now attract each other. Also, the gluon forces in the proton and neutron, in the ions of interest, are significantly lessened, so that the quarks in each proton and neutron are very much more loosely bound. Consequently, decay from an excited state by quark flipping to turn a proton into a neutron or vice versa becomes an attractive option. The strong force is therefore reduced in radius.

(ii) Two positive ions now attract each other so closely that each enters the edge of the weakened strong force region of the other. This forms a *quasi-nucleus* because of the limited involvement of the weakened strong force.

(iii) The other ions in solution surrounding the TRZ immediately move (as soon as the TRZ is formed) to negate the TRZ and decay it back to a time-forward zone (TFZ). This action initiates the decay of the TRZ.

(iv) As the TRZ lessens and then turns back into a TFZ, the strong force expands its size faster than the novel Coulomb "like attracting like" force reduces and reverses.

(v) The rapidly expanding strong force of each of the two ions in the quasi-nucleus fully envelops the other ion, increasing the disequilibrium of the two-ion quasi-nucleus. A condition is reached where the most probable mode of decay is the flipping of one quark in one of the positive nuclei.

(vi) The quark flips, turning that H+ ion into a neutron n. This is a quasi-nucleus of deuterium (D+), still in an excited state but with less excess potential energy in the excited state. At this point, the TRZ is vanishing and the TFZ state is returning.

(vii) As the TRZ vanishes and TFZ state increases back to normal, the most probable mode of decay increasingly is to a full deuterium nucleus. Hence the H+ and n simply draw completely into normal deuterium binding position, bound by the normal strong force. This constitutes a normal D+ ion, now existing in a normal TFZ.

(viii) So by formation and decay of a significant TRZ, two H+ ions have been drawn together into a quasi-nucleus consisting of two H+ ions partially bound by a weakened strong force. As the TRZ decays and the Coulomb repulsion resumes, the strong force increases back to full strength faster than the Coulomb repulsion force returns and repels the two H+ ions. Since the quarks are much more loosely bound than normal, a quark in one of the H+ ions has flipped, turning the ion into a neutron n. The H+ ion and the neutron, partially bound already in a quasi-nucleus, simply "tighten" into a normal D+ nucleus as the TFZ condition is fully resumed.

The nuclear reaction equation for the above interaction may be written as

$${}^1H_1^+ + {}^1H_1^+ \Rightarrow (({}^1H_1^+ + {}^1H_1^+)) \Rightarrow ({}^1n_0 + {}^1H_1^+) \Rightarrow {}^2H_1+ = {}^2D_1^+ \qquad \textbf{[10]}$$

In equation [10] we use the left superscript as the number of nucleons in the ion nucleus, the right subscript as the protons in the ion nucleus, and the right + superscript to show the overall charge of the ion and its sign. We use the expression in parentheses to show the involvement of a time-reversal zone (TRZ). Double parentheses show a stronger TRZ than single parentheses. Thus from double parentheses to single parentheses to no parentheses shows the initial formation and subsequent decay of the TRZ.

Another interaction involves a D+ ion and an H+ ion, to produce tritium. This is:

$${}^2D_1^+ + {}^1H_1^+ \Rightarrow (({}^2D_1^+ + {}^1H_1^+)) \Rightarrow ({}^2D_1^+ + {}^1n_0) \Rightarrow {}^3H_1^+ = 3T1^+ \quad [11]$$

That reaction transmutes a deuterium ion and hydrogen ion into a tritium ion. Without showing it, in a TRZ three ${}^2H_1^+$ ions may also attract into a quasi-nucleus, and as the decay of the TRZ occurs, first one proton turns into a neutron and then a second one does also, by quark flipping. This nuclear reaction also produces a fusion into tritium.

Another reaction between two deuterium ions is"

$${}^2D_1^+ + {}^2D_1^+ \Rightarrow (({}^2D_1^+ + {}^2D_1^+)) \Rightarrow [2({}^1H_1^+ + {}^1n_0)] \Rightarrow {}^4He_2^+ = \alpha \quad [12]$$

That interaction — particularly in pre-deuterated electrolyte solution — gives the excess a particles produced in a great number of the cold fusion experiments.

A rarer but still occurring interaction is:

$$4[{}^1H_1^+] \Rightarrow (({}^2D_1^+ + {}^2D_1^+)) \Rightarrow {}^4He_2^+ = \alpha \quad [13]$$

Indeed, reaction [13] may occur to completion and formation of a particles in some of the transmutations in an electrolyte, while it proceeds only to the intermediate phase in others. In that case, both deuterium ions and a particles may be produced out of the same overall "chain of reactions" where some reactions proceed all the way and some do not.

These interactions are directly using and transducing time-energy by use of the TRZ to time-reverse the Coulomb repulsion law for like charges. From the reactions, excess energy given off as heat may and will occur.

The nuclear physicist and nuclear chemist can immediately see a great many new nuclear reactions now made possible by the adroit production and decay of TRZs. A great number of these new nuclear reactions are now possible by this means, which are impossible in ordinary nuclear chemistry without the deliberate formation of TRZs.

These few examples serve to illustrate the process, and these reactions represent results already achieved in numerous cold fusion experiments worldwide, by many researchers and multiple laboratories. Now we turn to other evidence strongly supporting the novel nature of these reactions rather than several other proposed reaction equations yet to be proven. The other proposals do not explain the next type of new experimental reaction.

10.7 Time-Energy Can Generate Instrumentation Anomalies

Strong evidence for time-charging and time-charge decay is experimentally demonstrated in electrolyte experiments where time-reversal zones (TRZs) are operating, and thus where some localized processes are running backwards in violation of the second law of thermodynamics. Weak time-charging of photon-absorbing and emitting matter, with special patterns of spacetime curvatures (engines) induced by time-density oscillations and structuring, is persistent in matter, having "charge-up" and "decay" times that may last for minutes, months, or even years. A very wide range of previously anomalous phenomena, both in inert and living matter, is explainable by these new interactions. This is very probably also related to the "memory phenomena" noted by Kondepudi and Prigogine[304] as outside the present thermodynamics, but being worked on at the forefront in what is called "extended thermodynamics". Quoting Kondepudi and Prigogine, p. 460:

> *"...interesting memory effects ...appear for long times (as compared to characteristic relaxation times). ...nonequilibrium processes may have 'long-time tails'. In other words, the approach to equilibrium is not exponential, as was generally believed, but polynomial ...which is much slower. ... As a result, Nature has a much longer memory of irreversible processes than it was thought before. Again this shows that local equilibrium is an approximation, albeit a very good one. "*

In this book, we focus primarily upon some typical anomalous measurements in China Lake experiments that are directly involved with time-charging and decay effects, lending strong support to our thesis of the novel TRZ-induced nuclear reaction chains. Note that these instrumental anomalies are also strong evidence for the association of time-charging and time-charge decay with the proven fluctuation theorem of Evans et al.

---

[304] Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, Chichester, 1998, p. 459. For a discussion of extended thermodynamics, see D. Jou, *Extended Irreversible Thermodynamics,* Springer-Verlag, New York, 1996.

at the Australian National University. We believe this to be a contribution to that very important work of Evans and his colleagues.

10.7.1 Ionization Detectors, Transduction, and EM LW Interferometry

Ionization processes obviously are processes where the partial transduction of LW time interaction into 3-spatial energy interaction would yield ordinary excited energy states of the affected electrons or ions. In turn, this would lead to ionization (e.g., of the atoms and molecules of a gas, liquid, etc.) Hence, ordinary ionization detectors such as Geiger Miiller tubes have the innate *capability* of detecting such transduction that generates ionization, and thereby have some capability of serving as "detectors" of longitudinal EM wave interactions (and time-charging) ongoing in the detector. In the usual situation, no excess time-charging results and no time reversal zones result. In that case, symmetry of energy flow in the time-domain exists, as does symmetry of energy flow in 3-space. Absolutely normal chemistry and nuclear chemistry results, and the nuclear detectors function normally, as familiar to every laboratory.

However, when significant TRZs emerge and persist for short periods, the symmetry of both time-energy flow and 3-spatial energy flow is broken. In this case, transduction of time-energy into 3-spatial energy (and vice versa) can result in anomalous ionization in nuclear ionization detectors, even in the absence of normal transverse EM nuclear radiation. The trick is to use and compare different detectors and different types of detectors to clearly demonstrate anomalous ionization detection effects.

As an example, a Geiger Muller tube will detect any ionization of its internal gas, regardless of what caused that ionization. If the ionization is caused by normal ionizing radiation from nuclear reactions, the instrument will detect and read a resulting internal ionization of the gas inside the tube.

On the other hand, if time-charging and decay are involved, then the asymmetry resulting in 3-spatial energy excitation of the gas inside the Geiger Muller tube (due to transduction of some time-energy into excess 3-spatial energy) may be sufficient to ionize the gas. In that case, the instrument will read "as if nuclear radiation were present. Note that there is indeed "ionizing radiation" present in the gas. However, now this ionizing radiation is not transverse EM wave in nature, but is due to the time charging and decay, and transduction of time into 3-spatial energy in the detecting gas.

This capability of an ionization detector to ionize in the presence of non-transverse wave ionizing radiation and energy can be used to provide an indication of time-density waves being formed and interacting in a process (such as a cold fusion process) where some transduction from time-energy to 3-spatial energy occurs.

Further, the previous "time history" of the individual instrument plays a part in whether the instrument "reads" or not for a given transduction situation. There will be a statistical variation of the actual "already present" low-level time-charge (and in its internal structure and engines) between instruments, even those made in the same factory on the same day from the same batch of materials and parts. That is because the time-charging and discharging history of each instrument has been different.

The production of transduced TWs, however, must be a function of scalar interferometry in the gases or other detecting media of the instrument, since scalar interferometry *creates* all transverse EM waves, as shown by Whittaker {619} and confirmed by Evans *et al.* {620}. In such interferometry, the entire inner structures of the transverse EM waves formed in the interference zone also interact wave-to-wave. Thus, this kind of interferometry is extremely sensitive to the exact internal longitudinal EM bidirectional wave structure (internal engine substructure) of the involved potentials.

As a consequence, the ability of an ionization detector to transduce a specific set of LWs and give ionization detection will vary appreciably, including for different variations in the instrument's original manufacturing process, and even for the specific past photon interaction history and experience of the individual detector itself. One detector's set of cumulated internal time-charges (and hence spacetime engines) may vary considerably from those of a second detector of exactly the same type. In general, multiple ionization detectors are unlikely to all detect a given time-density EM wave emission and interaction set {621}.

Even for two detectors of the same brand, from the same manufacturing plant, and from the same batch of manufacture, it is likely that significant differences in LW detection of a specific TDW set will occur because of the "past history" time-charge differences of the individual detectors. Indeed, a likely phenomenon is that, when one ionization detector detects the LW emissions (transduces them), several others will not detect them at all. This is one of the peculiarities of the new unified field area that must be overcome by further research in order to develop reliable, calibrated TDW and transduction detectors. It is one of the primary problems - if not *the* primary problem - of cold fusion research to resolve this *"time-charge experience history"* problem, so that transduction detectors having calibrated, uniform responses to given TDW sets and interactions can be provided.

At present, no one has the foggiest notion as to how to "calibrate" a nuclear ionization detector that one desires to use as a time-energy transduction detector. Speculating, we would envision such calibration (in the transduction detecting functional response) to become possible when small, standard, calibrated sources of transduction are developed and available. In that case it will be possible to formulate procedures whereby varying the output of the transduction source will vary the transduction irradiation of the GM tube. The level of transduction intensity required to create ionization in the instrument will be a direct indicator. When small, calibrated sources of TDWs are also available, it will be possible to irradiate the GM tube to a specifically desired rate and level of time-charge and decay. In that case, an array of GM tubes can be "standardized" or "calibrated" for transduction detection. So far as this author is aware, those developments have yet to be accomplished or even undertaken. Indeed, the need for such a development program has not even been realized.

10.7.2 Experimental "Detector Transduction Phenomena" in Electrolysis

Let us look at one set of ongoing scientific experiments where just such anomalous detector results have been obtained.

Researchers at the Naval Air Warfare Center at China Lake, and at the Department of Chemistry, University of Texas at Austin, have detected precisely the kind of "anomalous" radiation and transduction effects we expect to be in the ionization processes of multiple kinds of detectors {622}.

According to a report by Miles and Bush {623}, anomalous radiation at China Lake was first detected by the exposure of dental X-ray films in two experiments producing excess power (excess TW energy emission). Film used in a control study during these experiments showed no exposure. There was also no exposure of similar films in more than 20 experiments where no excess power was present. In other words, the presence of excess heat (excess TW energy emission) strongly indicates the presence of transduction. In turn, the presence of transduction implies the presence of scalar interferometry wave interaction, where TDW waves and LW waves **are** absorbed and **a** mix of LW and transduced TW waves are emitted consisting of gravitons (paired scalar and longitudinal photons). This only occurs when significant time-charge excitation has built up. Taken in reverse, the presence of LW wave absorption and emission — with transduction into detected excess ionization energy — directly indicates the emission of "anomalous" TDW or graviton radiation, as detected by the resulting TW exposure of the x-ray film. So the "anomalous" radiation detections are consistent both forward and backward.

The film in manufacture is quite strongly quality-controlled. Also, it is used (interacts and indicates) only once. Its past time-charge history has minimal variations between two samples of the film, because there is no repetitious past *detection interaction* history. Hence one would expect a high degree of uniform film detection interaction responses from film to film, to the same graviton emission set. That is precisely what occurs.

We conclude that the behavior of the film in the combined China Lake experiments clearly shows the time charging, graviton emission and interaction, and transduction nature of what is happening when it is known that no normal nuclear radiation is present.

Geiger-Muller (GM) detectors and sodium iodide (NaI) detectors were also utilized when electrolysis experiments using heavy water were ongoing. We again accent that a Geiger-Muller tube does not detect nuclear radiation per se; instead, it detects *anything* that will cause its internal gas to ionize sufficiently. Sufficient transduction in graviton absorption-emission interactions in a Geiger-Muller tube will cause the counter to indicate, because it ionizes the gas and produces an ionization discharge. However, both the specific transduction and scalar interferometry aspects of the ongoing experiment are involved, as well as the previous background time-charge history of the Geiger-Muller tube counter.

Several Geiger-Muller detectors gave anomalously high readings, reaching some 73 sigmas above normal background counts. Most experiments (i.e., most GM detectors), however, gave normal radiation counts, and no anomalous count rates were ever observed when the experiments were turned off. So these anomalous results are differentiated by using multiple, carefully calibrated Geiger-Muller tube detectors. Our interpretation here is that

(i) There was a variation in the presence of transduction and scalar interferometry from experiment to experiment,

(ii) The majority of the experiments did not produce significant time-charge and sufficient transduction or scalar interferometry to cause detection (ionization discharge) on most (typical) Geiger-Muller tubes, and

(iii) The transduction effects in the ongoing experimental process were mostly of the rapid variety, and not due to long-term "charge-up" effects conditioning the time-charge aspects and structuring of the experimental apparatuses.

(iv) By using multiple detectors, the probability of one or more of the instruments having increased time-charge from specific time-charge histories was increased.

(v) With sufficient experiments and sufficient detectors, there resulted a high probability of having at least one or more detectors capable of detecting the levels of graviton radiation and transduction expected from the experiments.

(vi) A direct correlation would be predicted — and was observed — in the appearance of the anomalous radiation effects and the expected time periods required to load the palladium with deuterium. As reported by Miles and Bush, *ibid.*:

> "... *the anomalous radiation would appear within a few hours in the co-deposition experiments where the palladium is loaded with deuterium as it deposits from solution. In contrast the appearance of anomalous radiation required days of electrolysis for the palladium rods that load much slower.*"

We previously discussed the major variables indicated from many cold fusion experiments to include the strong correlation with the degree of loading of the palladium lattice. Hence the effect pointed out by Miles and Bush would indeed be predicted by the present approach and proposed mechanism.

Our interpretation is that the experiments show the time-charge rate effect to be expected in such graviton radiation and transduction interaction phenomena using collection of ions (in this case deuterium) that are much heavier than electrons. The faster the deuterium loaded, the greater the buildup of the interaction of the deuterium in phase conjugating and self-targeting iterative interactions — and therefore the greater the increase in scalar interferometry interactions — inside the palladium lattice. The rate of graviton radiation production and transduction production increases as some function (not necessarily linear!) of the rate of loading of the **deuterium.** As the rate of graviton and transduction production increases, so does the expectation of anomalous ionization effects in the nuclear radiation detectors. And so does the rate of production of TRZs and the appearance of the new nuclear cold fusion reactions at low spatial energy but very high time-energy.

10.7.3 Experiment Confirms It

A clinching observation is given by Miles and Bush, *ibid.*, as follows:

> *"One GM-detector would measure anomalous radiation while another GM-detector would be "blind" to any anomalous effect. A few experiments, nevertheless, gave simultaneous anomalous effects from two different radiation detectors."*

Here our comment is that this precisely fits our expectations for demonstrating involvement of

(i) Time-charging and decay interactions,

(ii) Consistency with transduction expectations,

(iii) Consistency with the fact that two different radiation detectors may or may not exhibit or detect the anomalous ionization effects, depending upon the individual past time-charge histories of the two detectors, and

(iv) Predictable variations in the individual experiments (based on palladium lattice loading time, specific geometry, ionization state of the chemical solution, internal time-charge state and decay rate, and structuring of the apparatuses and their parts, etc.) when transduction-associated phenomena are considered.

The anomalous non-nuclear, non-transverse wave — but ionizing — EM radiation detected at China Lake, and the resulting peculiarities in the measurement apparatus functioning, do strongly support the present proposed approach, the proposed graviton radiation phenomena, and the time-charging and decay and transduction mechanisms presented. It is decisive that sometimes two detectors both indicated, and sometimes one indicated and the other would not. Sufficient replications have ruled out any defect in a single given instrument as the cause of the anomalous detections. That immediately eliminates from consideration all purely transverse EM wave phenomena, conditional only upon the premise that the instruments were properly functioning and calibrated, and that sufficient replications were made for decisive statistics. Those conditions were all met in the numerous experiments producing these results over a long time and in many repetitions.

As a final thought on the subject, we pose the possibility that the longitudinal EM waves involved may be the B(3) longitudinal field waves advanced convincingly by Evans {624a-624c}. We would recommend the attention of leading theorists and experimentalists to this possibility, as a matter that should be investigated both theoretically and experimentally.

10.8 Additional Implications for the Future

10.8.1 Longitudinal EM Waves and Undistorted Progressive Waves

There is a revolution in electrodynamics presently underway, due to the unique characteristics of *longitudinal* EM waves[305] and *pseudo-longitudinal* EM waves. A summary by Rodrigues and Lu {625} is particularly revealing.

A pure longitudinal EM wave has infinite energy and infinite velocity — i.e., it appears everywhere at once. "Propagation through space" does not apply; instead, one has stumbled into that weird region where spacetime is multiply connected. In the real world, one meets imperfect longitudinal waves — or *pseudo-longitudinal* EM waves — that still retain a transverse energy density variation residue. These waves are called *undistorted progressive waves* (UPWs).

Such UPWs can move in space faster than the speed of light. This should not be too surprising, since superluminal velocity of intelligent signals has been positively demonstrated by several researchers in quantum-tunneling experiments. As an example, Nimtz {626} and his colleagues have beamed Mozart's 40th Symphony through a barrier in a waveguide at a speed of 4.7c. So the old notion that information cannot be transported at superluminal velocities is passe. The Fogal semiconductor has been used in experimental transmitter-receiver pairs to demonstrate superluminal transmission, such as TV pictures and programs. Direct signal transmission from earth to satellite back to earth and through many en route amplifiers and processors has been demonstrated by Fogal to occur far beyond the speed of light. Again, theoretical summary papers by Rodrigues *et al.* {627a-627f} are revealing.

---

[305] We have pointed out a direct way to generate longitudinal EM waves, by simply oscillating the magnitude of the L**orentz self-regauging** of a system.

10.8.2 Death Knell of the Speed of Light Limitation

Another giant shock has emerged in astrophysics during the last few years, and it is now threatening a dramatic revision to physics. Heretofore, it has been fashionable to assume that gravity itself propagates at light speed. A great deal of very comfortable physics theory has been built around that assumption.

Well, the assumption is not true. Experiments in astrophysics continue to refute it. It now is almost certain that the speed of gravity is at least 2×10^{10} c. A beautiful though somewhat controversial paper by Tom Van Flandern {628} summarized the entire situation. He articulates the comfortable general relativity view held to date {629} as follows:

> *"...GR proposes that such changes appear to act instantaneously in the 'near field', but eventually show their true, light-speed-delayed character in the 'far field', which is conveniently beyond our present ability to observe. The necessity of this dual behavior is to prevent the logical need for changes to continue to appear to act instantaneously at ever increasing distances, even to infinity."*

In fact, Van Flandern {630} points out that general relativity already admits superluminal effects in the near field region. He also pointed out the startling nature of the experimental results to relativists, and points out that an attachment to a paradigm has been involved:

> *"To many, this result* [astronomy's verification that gravity's speed $v_g >= 2x10^{10}c$] *is so contrary to 'common sense' in the light of relativity theory as to be absurd. But Thomas Kuhn has cautioned all scientists to avoid the trap of becoming so steeped in a prevailing paradigm that it starts to seem like common sense and makes other ideas sound and feel wrong. Eventually, even one's professional status can become linked to a prevailing paradigm. "*

As Van Flandern further emphasizes, it is not difficult to show by computer calculations of astrophysical situations that any strong limit on the speed of light destroys conservation laws:

> *"...anyone with a computer and orbit computation or numerical integration software can verify the consequences of introducing a delay into gravitational interactions. The effect on computed orbits is usually*

> *disastrous because conservation of angular momentum is destroyed."*

So here we have a new paradigm in the making. This is a startling change to standard relativity. Either we have to give up the conservation of angular momentum, or we have to accept superluminal speed of gravity in the far field. Years of struggle to detect quadrupole gravitational wave radiation have failed; no direct detection has ever been made. Indeed, the rather arbitrary nature of the entire quadrupole radiation assumption is not required by general relativity, as long pointed out by leading Russian scientists. For example, Vlasov and Denisov {631} *et al.* bluntly state it in these words:

> *"...Einstein's well-known quadrupole formula, which is usually employed in general relativity for calculations, is not a consequence of general relativity and is not contained in it."*

Therefore, it may not really be so surprising that physicists have not been able to detect such quadrupole radiation in a great many laboratory experiments. It is not in general relativity at all.

Of course, our thesis developed in this book is that what we have long called "EM radiation" is in fact gravitational radiation, always involving paired scalar and longitudinal photons and hence time-density waves and longitudinal EM waves. We have pointed out that prior to observation there is no such thing as a separate photon, but there is such a thing as two correlated photons, one time-polarized and the other longitudinally polarized. We believe that there is sufficient meat in our proposed explanation of the anomalous cold fusion results and the anomalous instrument results at China Lake to warrant serious and strenuous investigation of these hypotheses and proposed mechanisms.

Another oddity - usually never stated by Western physicists - has long been known but ignored. When Einstein characterized the gravitational field as a pseudotensorial field, and not a field in the sense of the Faraday-Maxwell field, the consequences were that the 3-space energy laws of the familiar kind really do not exist in general relativity. Shortly after Einstein published his theory of general relativity, this absence of familiar energy conservation was pointed out by Hilbert {632} in these words:

> *"I assert... that for the general theory of relativity, i.e., in the case of general invariance of the Hamiltonian function, energy equations... corresponding to the energy*

> *equations in orthogonally invariant theories do not exist at all. I could even take this circumstance as the characteristic feature of the general theory of relativity. "*

With our publication of the giant negentropy of 4-space once arbitrary 3-space symmetry and time-symmetry in EM energy flow is removed, one of the truths involved in Hilbert's remark was finally recognized. In addition, what Sen called "the most difficult problem in classical and quantum electrodynamics" {633} has been explained by giant negentropy {12}. Also, it is consistent with an appropriate reinterpretation of Whittaker's 1903 decomposition of the scalar potential {600, 615}. Finally, it is consistent with quantum field theory {634}.

Hilbert was evidently not understood by his contemporaries, since neither Einstein himself nor other physicists recognized the fact that, in general relativity, conservation laws for 3-space energy, momentum, and angular momentum — in the sense we are accustomed to in the rest of physics —in principle may be impossible. That is because one can readily remove 3-space energy symmetry when curved spacetime is permitted. Again, Russian scientists such as Logunov and Loskutov {635} have long pointed out the "unthinkable" and "astounding" fact that the predictions of general relativity are not unique.

In the West until recently relativists were conservative, and this resulted in the sidelining of innovative Western physicists who challenged the prevailing interpretation or extended it. With the experimental falsification of one of the fundamental tenets — that gravity moves at speed c in the far field — of the prevailing interpretation of general relativity, it appears that we may be approaching an asymptotic burst of great new theoretical work and a dramatic new reinterpretation. If it ever gets into production, we believe that the Fogal semiconductor will usher in that great change of communication speed paradigms that now appears imminent.

Nonetheless, general relativity — whatever the modifications and extensions now called for — will remain a very useful tool indeed. There is simply nothing else on the horizon that can replace it, string theory notwithstanding. The reinterpretation is likely to shed additional light on many present problems such as the nature of spacetime itself, the nature of dark matter, and new insight and progress on unifying physics.

In the new openness that should result, we hope that a great extension and reinterpretation of electrodynamics will also be undertaken, and that it will reveal the underlying powerful structured general relativity infolded inside conventional EM potentials, fields, and waves, as originally revealed by

Whittaker {636}. It may even be that one of the novel concepts {637} we have proposed in this book will yet see the light of day in practical systems.

If so, then hopefully much of what we are addressing in this book will be incorporated in the emerging new physics of the twenty-first century.

Chapter 11

Final Perspective: Permissible COP>1.0 Maxwellian Systems

"Are you certain of that?" [Augustine, in his famous confrontation with the Skeptics, who held that nothing can be known with certainty.]

"[It is usually]...assumed that the magnitude of potential energy is irrelevant, being arbitrary to the extent of an additive constant. " [Block and Crater]. {638}

"The potentials V and A, defined ... in terms of the fields E and B, are considered to be nonobservable in Maxwellian electromagnetism, since the energy density associated with them is zero... " [Lakes]. {639}

"[The total energy of the atom] depends on the *bare mass* *and* *bare charge* *of the electron, the mass and charge that appear in the equations of the theory before we start worrying about photon emissions and reabsorptions. But free electrons as well as electrons in atoms are always emitting and reabsorbing photons that affect the electron's mass and electric charge, and so the bare mass and charge are not the same as the measured electron mass and charge that are listed in tables of elementary particles. In fact, in order to account for the observed values (which of course are finite) of the mass and charge of the electron, the bare mass and charge must themselves be infinite. The total energy of the atom is thus the sum of two terms, both infinite: the bare energy that is infinite because it depends on the infinite bare mass and charge, and the energy shift... that is infinite because it receives contributions from virtual photons of unlimited energy. "* fWeinberg]. {640}

11.1 Introduction

The central message in this book is that all EM devices are already powered — though in an inappropriate fashion — by EM energy freely extracted from the seething vacuum. It is also intended to be a clarion call that we must rapidly develop the *appropriate* vacuum-energy powering of our EM circuits and electrical power systems, or else the survival of the nation and of the biosphere is at stake rather immediately.

For this summary chapter, we wish to place that specific area — EM energy from the vacuum — in sharp focus, with the major facets together in one place.

Consequently, for this closing chapter we have chosen and slightly adapted a previous technical rebuttal to the objections of "perpetual motion", "violation of the second law of thermodynamics," "impossible", etc. that we wrote in 2001 in the referee process, to the referees of one of the AIAS papers submitted to a major physics journal. The rebuttal was successful, and it was accepted and upheld by the referees. The paper was then published by the journal, after very vigorous refereeing.

That is as it should be.

Scientific journals provide a proper forum for the free expression and debate of scientific ideas and work. They do not and should not ensure that every paper conforms to commonly held concepts or to the most widely accepted model. As we stated, Chief Editor Kennedy of the journal *Science* has only recently had to restate and accent the principle that science is free from "truth by prevailing authority," and bound by "truth by replicable experiment". Insuring that every paper meets the accepted models and notions would be enforcement of religious belief, not experimental science. Ideas which are wrong will be refuted experimentally; experiment is still the basis of scientific method. In our rebuttal, the argument of the conventional skeptics is thoroughly refuted by experiments already well known in the literature, as we detail.

So we thought it appropriate to close the book with this strong statement of its primary pbjective: economical, clean, copious electrical energy from the vacuum.

11.2 Source Charges, Curved Spacetime, Efficiency, and COP

11.2.1 The Source Charge Problem

The received interpretation of the Maxwell-Heaviside theory treats the source charge and source dipole as perpetual motion machines, freely and continuously pouring out EM energy in 3-space in all directions with no energy input. Experimentally one verifies that there is no *3-space observable energy input* to the charge or dipole. In accepting that model, a scientist objecting to COP>1.0 EM systems has unwittingly hoisted himself upon his own perpetual motion machine petard, by already accepting every source charge and dipole as precisely that to which he is objecting. To explain the source charge, either one holds to the U(1) Maxwell-Heaviside theory and surrenders the conservation of energy law, or one accepts a change to U(1) electrodynamics to add the concomitant input of energy flow from the active vacuum (from the *time* domain, which in Minkowski geometry is all that is left outside 3-space) to that charge or dipole. That corresponds to a violation of the second law of thermodynamics, as — as we discussed in Chapter 10 — Evans and his colleagues at the Australian National University have already experimentally proven that the second law can be violated up to micron scale and for times of up to two seconds.[287] At the much smaller fundamental charged particle level and far shorter "switching" time of one spin completion, the probability is 100% and a certainty that the second law is violated. Hence the broken symmetry of the charged particle considered in the modem sense as a dipolarity (considering its associated **clustering** virtual charges of opposite sign).

Either way, the classical Maxwell-Heaviside theory must be changed. Every charge and every dipole — which themselves are Maxwell-Heaviside EM systems *a priori* — refutes the erroneous argument that no Maxwell-Heaviside system can output more energy than the operator himself inputs from an external power supply. To adamantly continue to challenge something already *experimentally demonstrated by* every charge and dipole in the universe is the epitomy of scientific blindness.

As have others, Sen {641} called this source charge paradox the "most **difficult"** problem in electrodynamics. However, the basis for its solution - absorption and transduction of virtual KM energy from the vacuum via the broken symmetry of a charge or dipole in its vacuum energy exchange — has now been known in particle physics for nearly a half-century {642a-642d}. It is not included in the Maxwell-Heaviside-Lorentz theory. Neither the proven vacuum interaction with Maxwellian system nor the broken symmetry in that interaction has been added to the model.

The present author has proposed an EM wave solution {643} to the source charge problem, by reinterpreting Whittaker's {644} decomposition of the scalar potential (as between the ends of a dipole) and treating the "isolated charge" and its clustered virtual charges {645} of opposite signs as a set of composite dipoles. Each of the dipoles then becomes a broken symmetry in the energetic exchange with the vacuum. Hence the charge pours out observable EM energy with no *observable* energy input because the input energy is in virtual photon form.

In the solution, the EM energy is input to the charge or dipole from the time-domain, so conservation of EM energy flow is upheld in 4-space, while time-like symmetry and space-like symmetry are broken individually. Powerful support for the proposed solution was given in a quantum field theory argument by Mandl and Shaw {646} nearly three decades ago. We challenge any scientist skeptical of COP>1.0 Maxwellian systems to produce a solution of the source charge problem in Maxwell-Heaviside-Lorentz theory alone. In conventional theory, every charge in the universe experimentally exhibits COP = oo. Any theory conflicting with a replicable experiment is falsified a priori.

11.2.2 EM Waves Imply Curved Spacetime

The oscillating energy of an electromagnetic wave is continuously changing its local spatial energy density. This is an *oscillating change of spacetime curvature* if one accepts general relativity (GR). Hence the EM wave in space is *always* moving in a locally curved spacetime; else, no EM wave exists because no change in spatial energy density exists. To reject that, is to reject GR in its entirety as well as all the experiments consistent with GR. If one accepts GR, then the classical U(1) EM *assumption* of the EM wave moving in a fiat spacetime is a non sequitur. It is unscientific to be suspicious of a higher group symmetry O(3) electrodynamics which *does* correct this known non sequitur in U(1) electrodynamics, while limiting oneself to U(1) electrodynamics *with* that known error. One might as well be suspicious of tensor algebra because it is more comprehensive than vector algebra.

There are many EM experiments such as the Aharonov-Bohm effect {647a, 647b} which violate Maxwell-Heaviside EM, as is well known

{648, 649a-649i}. The foundations of the Maxwell-Heaviside theory are well over a century old, and Maxwell's theory was substantially truncated in the 1880s (more on that below). The considerable physics developed since then has required dramatic extensions to the Maxwell-Heaviside theory such as developing non-Abelian gauge field EM theory, quantum electrodynamics, and modeling of the interaction of vacuum and matter in any material system.

11.2.3 On EM Systems with COP>1.0

Conservation of energy requires that all energy output by an inert system must be input to it {650}. It does not require that the *operator* input all or even any of the energy utilized. The active environment can permissibly input part or all of it, else there could be no such thing as a windmill — or a charge, or a dipole, or a solar-cell electrical power system.

The *coefficient of performance* (COP) of a system compares the system's useful energy or work output to the *operator's* energy input. The *efficiency* ε of a system compares its useful output to its *entire* energy input. No inert system can output more useful energy than its total input, hence $\varepsilon <= 1.0$, being $\varepsilon = 1.0$ for a theoretically lossless "perfect" system and $\varepsilon < 1.0$ for a real system with losses.

Whether the operator must input all the energy that is output by the inert system depends upon whether the system is (i) an open system in disequilibrium with its active environment, and (ii) *designed* to accept and utilize excess energy from that environment, such as from the active vacuum. If appreciable environmental energy input is received and utilized, then a system with losses may permissibly exhibit COP>1.0, even though its efficiency ε is $\varepsilon < 1.0$. The common home heat pump is an example. While the efficiency of a home heat pump may be $\varepsilon = 50\%$, its **COP** may be COP = 4.0.

As does the skeptical scientist, the electrical engineering community erroneously assumes that an inert electrical power system that outputs more energy than *one oneself* inputs to it — e.g., a generator outputting more EM energy flow than the magnitude of the mechanical energy input to the generator shaft — is against the laws of physics. To the contrary, it is permitted by the laws of physics and thermodynamics, *once the modern* ***Vacuum*** *interaction is included and disequilibrium exists in the interaction* ***between*** *vacuum and system.* Not only is it permissible, it is a universal fact once the arbitrarily discarded Heaviside energy flow component is re-ac**counted,** and the source charge problem is reluctantly placed back on the table. If we account all that EM energy pouring freely from the source charges in any EM circuit, then the actual output of EM energy from that circuit is enormous. It is many orders of magnitude greater than the magnitude of the input energy that the experimenter or operator provides.

We do not have to reprove the active vacuum; it has long been proven in particle physics. We also do not have to reprove the disequilibrium between electromagnetic systems and the active vacuum environment; that also has been proven since 1957 by the prediction and discovery of broken symmetry. One of the proven broken symmetries is that of opposite charges, such as are on the ends of a dipole. Hence every dipole or dipolarity in electromagnetics already freely absorbs virtual photon energy from the vacuum, transduces (coherently integrates) it into observable photon energy, and re-emits observable EM photons in all directions at the speed of light. If there are no COP>1.0 Maxwellian systems, then there can be no Maxwellian charges, dipoles, or dipolarity, and therefore no resulting fields and potentials from these sources — an absurdity falsified by every standard two-wire electrical circuit and ever charge and dipole in the universe. Any potential is a dipolarity, hence exhibits broken symmetry. That is why from any finite potential, as much EM energy as desired can be intercepted and collected, by the simple equation $W = \phi q$, where W *is* the collected energy in joules, ϕ is the scalar potential intensity in joules per unit point static coulomb, and q is the intercepting charge in coulombs.

Systems far from equilibrium in their energy exchange with their active environment {651a, 651b} can permissibly perform five important functions impossible to equilibrium systems. Such a disequilibrium system can:

(1) Self-order,

(2) Self-oscillate or self-rotate,

(3) Output more energy than the operator inputs (the excess energy is freely received from the active environment),

(4) Power itself and its load (all the energy is freely received from the active environment), and

(5) Exhibit negentropy.

Any charge or dipole already exhibits all five of these functions — forbidden by classical thermodynamics but permitted by the well known thermodynamics of systems far from equilibrium in their exchange with their active environment {651a, 651b}.

It *only takes one small white crow to prove that not all crows are black.* The known performance of the charge and the dipole totally refutes the notion that no EM system can output energy without operator-arranged input from other than the active vacuum. The entropy-consuming reactions from the fluctuation theorem, experimentally proven by Evans et al. in 2002, also is a sufficient proof to destroy any argument against permissible COP>1.0 EM systems. Since all the EM field energy and potential energy in a power system or circuit must come from the source charges and dipoles, it follows that all our EM energy systems already take *all* their EM energy from the active vacuum, *not* from the operator's input. A solar cell array also refutes the notion that the operator has to input the energy, but the solar cells' environmental energy input may not be ubiquitous or dependable. The active vacuum is both.

Most power system electrodynamicists avoid the particle physics solution {652} to the source-charge problem {641} involving disequilibrium exchange with the active vacuum. Instead, they adhere strictly to the Maxwell-Heaviside-Lorentz theory with its *assumed* equilibrium between system and vacuum. Therefore they avoid modeling the vacuum interaction and solving the problem of open dissipative EM systems which freely and dependably receive energy from their vacuum environment in unusable form, translate it into usable form, and furnish it for further capture and use. They therefore cannot resolve the source charge problem in their model, for the model has already artificially excluded the solution.

By definition, an EM system in equilibrium cannot output more useful energy than the operator inputs. However, considered as an *energy **transducer**,* every EM system is in disequilibrium with the vacuum because of the charges and dipoles in the system, and it continuously receives energy from the vacuum. Because of the broken symmetry of the charge and the dipole, some of this energy from the vacuum is discharged by every charge and every dipolarity in the circuit as observable energy *a priori.* This creates a truly enormous outpouring of rather disorganized (in this case, *uncollected)* EM energy extracted from the vacuum and poured out without being intercepted and diverged into the circuit, where it is organized for dissipation and use. The circuit "uses" only that amount of this enormous available outpouring EM energy that it first organizes (collects) and directs. That is its "organized excitation" energy, commonly known as *input energy* macroscopically — the Poynting energy flow component.

Upon completely discharging its excitation (input energy), the inert system outputs all its received (organized) energy, from whatever source. So every EM system already outputs far more energy than the experimenter inputs — when the system's *nondiverged, non-organized, and nondiverged* output energy flow is considered as well as its *diverged, organized, and collected* energy flow component. The fact that the *diverged Poynting* energy received in the external circuit enters the circuit from the surrounding space is shown by Krauss {653} (See Figure 1-1 in Chapter 1). We show it in Figure 2-4 of Chapter 2 as well. The remaining and nondiverged Heaviside energy flow component in space, which misses the circuit and is wasted, is also shown in Figure 2-4 of Chapter 2 and contrasted to the diverged (caught) Poynting component. To understand this experimentally demonstrable fact {653}, we need a bit of EM history.

11.3 Truncations of Maxwell's Theory and Discard of System Classes

11.3.1 Concept of EM Energy Flow Through 3-Space

The concept of the connection of EM energy and space did not occur until 1876 in a faulty work by Croll {654}. In the 1880s Heaviside {655} showed the huge nondiverged component filling all space around a circuit (see Figure 2-4 in Chapter 2). Poynting {656} never even considered the external nondiverged flow missing the circuit, but assumed only the *diverged* energy flow component that enters the circuit. Figure 2-4 in Chapter 2 shows both components of EM energy flow laterally, so they can be seen clearly.

It is stressed that the Heaviside nondiverged component of EM energy flow around every circuit is far greater (by many orders of magnitude) than the circuit's intercepted, diverged, and collected Poynting energy flow component. This follows from Heaviside's own original papers. So when the entire output energy flow is considered, every generator and every battery is already a $COP \gg 1.0$ *energy transducer,* outputting far more energy flow than the rate at which mechanical energy is input to the generator shaft or the rate at which chemical energy is dissipated in the battery.

In the 1880s the electron, atom, nucleus, special and general relativity, quantum mechanics, quantum electrodynamics, and active vacuum had not been discovered. There was no knowledge of any possible active vacuum interaction source for such an enormous energy flow pouring from the terminals of the generator or battery and surrounding the circuit, filling all space around it. Heaviside's extra energy flow component was thus shocking and inexplicable, smacking of "perpetual motion" and total violation of the conservation of energy law. Heaviside — a brilliant, self-taught hermit who never attended university - spoke cautiously and obliquely of its magnitude in terms of angles with respect to a reference direction {657}.

11.3.2 Elimination of the Non-Intercepted Energy Flow Component

A perplexed and concerned Lorentz {658} cut the Gordian knot and arbitrarily *discarded* the Heaviside nondiverged component, eliminating the quandary. Lorentz stated that the huge nondiverged component had "no physical significance" (his phrase) since it powered nothing. Of course it **has** no physical significance, and powers nothing, unless one intercepts it and uses it!

By analogy, the component of a great wind on the ocean that is *outside* the intercepting sails of a single sailing ship, does not contribute to further powering *of that ship.* Figure 2-5 in Chapter 2 shows that situation. However, that nondiverged wind component can power a whole flotilla of ***additional,*** *separate* sailing ships with separate sails, as we showed in Figure 2-6 in Chapter 2. Electrodynamicists {659} have continued to utilize Lorentz' clever integration trick (Figure 2-3 in Chapter 2) to dispose of the huge nondiverged Heaviside energy flow component surrounding **every** circuit and missing it, and therefore pouring from the terminals of every generator. Most electrical engineers today do not realize that the original purpose of Lorentz's integration of the energy flow vector around a closed surface assumed surrounding every volume element of interest was merely to get rid of that perplexing and bothersome Heaviside nondiverged energy flow component while retaining the Poynting diverged energy flow component.

11.3.3 Lorentz's Previous Discard of all COP>1.0 Maxwellian Systems

Previously (circa 1890s) Lorentz had symmetrically regauged Heaviside's **equations** that had already seriously curtailed Maxwell's seminal equations of 1865. This Lorentz symmetrical regauging retained only that class of EM systems that are in equilibrium in their seething energy exchange with their active vacuum environment. It discarded that entire permissible class of the Maxwell-Heaviside systems far from equilibrium with the active vacuum. In short, it *arbitrarily* discarded all COP>1.0 Maxwellian **systems,** and all those even *permitted* to exhibit COP>1.0.

The ubiquitous use of closed current loop circuits containing both the ex**ternal** circuit and the source dipole inside the generator forcibly applied (and continues to forcibly apply) the Lorentz symmetrical regauging. Thus all electrical engineers began using circuits which self-enforced equilibrium in the vacuum exchange, so that no *net usable* energy[306] that could translate electrons and do work in an external load —could be received from that source.

However, so long as the enormous Heaviside nondiverged energy flow component outside the circuit remained, then separate outrigger antenna circuits and systems could always be inserted in this external energy flow to divert some of it into separate additional circuits. This would have opened up the COP>1.0 *ensemble* of systems and separate outrigger receiver circuits (with separate loads) fed by a single central generator, such as one version of the motionless electromagnetic generator of Chapter 7.

Consequently Lorentz integrated the entire energy flow vector around a closed surface assumed around any volume element of interest {658}, as shown in Figure 2-3. This integration trick — still universally applied {659a-659c} in all the texts today — arbitrarily discarded that troublesome and massive Heaviside energy flow component that could not be explained, and for which no source was known.

---

[306] Lorentz symmetrical regauging does indeed require that the system receive excess energy from the environment (the active vacuum). It requires that the potential energy of the system is changed twice and continuously, which means the excess energy in each of those two changes must continuously enter from the active vacuum and do internal work on the system. Two new free force fields are created, each with its field energy, which is additional energy being continuously dissipated in the circuit to perform continuous work. However, the two free changes of energy in the system have been precisely selected so that these two force fields are equal and opposite. The two free changes of energy thus produce a *stress potential* without a net translation field. This means that the local spacetime has been curved, in a manner representing a rotation of the system frame away from the lab observer's frame. Further, the stress potential represents the continuous entry of extra stress energy from the vacuum, whose two force fields continually perform opposing local *internal work* within the system to stress it, while unable to translate electrons and power external loads. Contrary to the received view, the Lorentz-regauged Maxwell-Heaviside system is a dramatically altered and different system than the original system before the arbitrary application of the Lorentz condition. The system's inertial frame has been rotated, it has taken on extra energy from the vacuum that can only be used to do internal work upon the system, and that internal work on the system is continuously being done to increase the system's stress and hold it at the new level.

11.3.4 The Heart-Breaking Result of Lorentz's Actions

The combination of these two actions by Lorentz — symmetrical regauging of the model equations and discarding the excess Heaviside nondiverged energy flow component — unwittingly resulted in the design and building of only those electrical power systems which symmetrically and automatically regauge their discharge of their excitation energy. The symmetrical regauging is self-applied by the closed current loop circuit. That circuit uses half of its collected Poynting energy to destroy the source dipole in the generator, and the other half to power the loads and losses in the external circuit {660}. Conventional EM systems are unintentionally designed to destroy their own source dipoles — and their free extraction of organized EM energy from the vacuum — faster than they power their loads.

11.3.5 Agreement with Classical Equilibrium Thermodynamics

These curtailments of Maxwell's theory resulted in the exclusive building of equilibrium EM systems. These systems obviously obey the classical equilibrium thermodynamics with its infamous second law. Lorentz had unwittingly discarded all that entire class of Maxwellian systems in disequilibrium in their exchange with their active vacuum environment; in those days, there was nothing known about an "active vacuum environment" or broken symmetry in its virtual energy flux. Consequently, Lorentz had discarded all those Maxwellian systems that *do not* obey classical equilibrium thermodynamics, but do obey the newer thermodynamics of systems far from thermodynamic equilibrium with their active environment. He discarded all permissible COP>1.0 Maxwellian systems.

However, there was a total agreement between all the EM systems actually built using Lorentz's symmetrized equations and all measurements in their ubiquitous closed current loop circuits — all self-enforcing the Lorentz condition. This led to the iron dogma that all EM systems must inevitably comply with classical equilibrium thermodynamics — the circuits that were built *were* repeatedly and universally observed to do so!

11.3.6 Dogma that COP>1.0 EM Systems are Perpetual Motion Machines

Finally, this total agreement (now for more than a century) between the theoretical model used, the circuits actually built, the measurements performed, and classical equilibrium thermodynamics led to the present fallacious but iron dogma that EM circuits and power systems cannot output more energy than we ourselves input. Today the vast majority of scientists consider it absolutely proven that the laws of nature prohibit

COP>1.0 electrical power systems, and particularly prohibit self-powering systems that freely power themselves and their loads (i.e., EM systems analogous to "windmills in a free wind"). As we stated, every charge and dipole falsifies the "COP>1.0 EM systems are impossible" dogma, as does the fluctuation theorem and its experimental proof by Evans et al.

Sadly, no one seems to have realized that we do not presently have COP>1.0 electrical power systems only because engineers have specifically designed and built their EM systems to destroy their own source dipole faster than they power their load. Ironically, every circuit and every power system used and measured by our engineers already outputs far more EM energy than is input by the operator or by turning the shaft of the generator, if the Heaviside nondiverged energy flow component is re-accounted. The Bohren experiment clearly proves it experimentally {662}, along with thousands of other similar experiments in "negative resonance absorption of the medium" {157a-157d).

11.4 Examples of COP>1.0 Electromagnetic Systems

The motionless electromagnetic generator (MEG) {661} (Figure 11-1) does catch and use some of this available free energy, as explained below. So does the familiar Bohren experiment {662a, 662b} by resonating each intercepting charge so it sweeps out an increased geometrical reaction cross section in space, reaching past the usual Poynting interception cross section (for a static charge) and thereby absorbing additional energy from the otherwise non-intercepted Heaviside energy flow component.

Every charge and dipole in the universe also already acts as an electrical "windmill in a free wind," having a COP = oo.

The present author has also nominated the long-neglected but vast nondiverged Heaviside EM energy flow, accompanying every field and charge interaction, as the long-missing source of some 90% of the gravity holding the arms of the spiral galaxies intact {663}. Our *supersystem* examination of several COP>1.0 EM systems developed by various inventors has also uncovered a proposed solution for the source of the distant antigravity accelerating the expansion of the universe {664}.

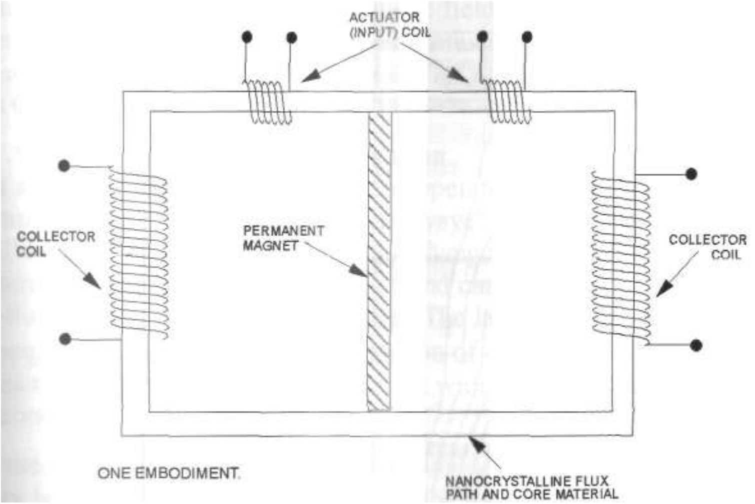

Figurei 11-1. Transformer section of the MEG.

Experimental proof that EM energy COP>1.0 is both possible and demonstrated is already in the literature — e.g., as in the Bohren experiment {662} which outputs 18 times as much energy as one inputs (see Figure 11-2). As another experimental example, the Aharonov-Bohm (AB) effect {647, 665, 666} freely adds additional nonlocalized field-free A-potential energy while all the magnetic B-field energy is sharply localized but still present and available. The AB effect is routinely implicated in physics, but apparently was never applied in an electrical power system prior to its use in the motionless electromagnetic generator

Bohren's experiment {662} was quickly and independently replicated by Paul and Fischer {667}. It can be replicated by any university laboratory at will, totally negating the conventional COP < 0 assumption. The AB effect is also easily replicated, using a toroidal coil, a long solenoid, or a permanent magnet and the special cores used in the MEG. When the toroid system's *total available output energy flow* is measured and accounted — including both the confined magnetic field energy flowing inside the toroidal windings and the field-free A-potential flowing outside — the toroidal coil system outputs more energy flow than the operator furnishes from his external power supply by conventional accounting {668}.

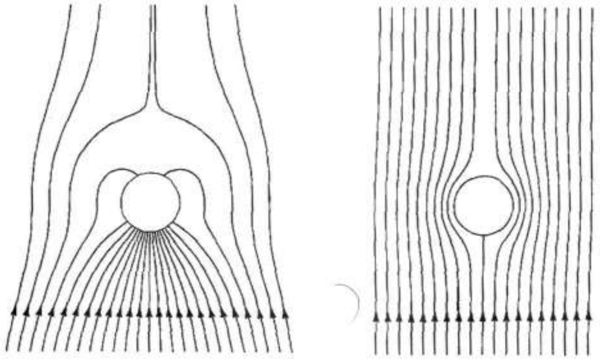

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

Figure 11-2 The Bohren experiment outputs 18 times as much energy as one inputs

In conventional terms, for a theoretically perfect toroid the operator only "pays for" the magnetic field energy "output component" confined inside the toroidal windings. The presence of an extra energy output component outside the toroid produces COP>1.0 performance as *an energy transducer.* It *does not* produce system COP>1.0 performance in a closed-loop circuit containing the primary power supply's source dipole.

The AIAS has previously nominated more than a dozen mechanisms as promising COP>1.0 systems for investigation, research, and development {669}.

11.5 Motionless Electromagnetic Generator (MEG)

11.5.1 Overall Description

In the MEG {661} (Figure 11-1), one uses a nanocrystalline material core path with the unique characteristic of drawing in almost all the magnetic field energy of an emplaced permanent magnet. The B-field flux from the magnet is withdrawn from its normal position in surrounding space and sharply confined to that material core path. Around the core there freely appears an *additional* curl-free magnetic vector potential A energy flow, since the surrounding spacetime is still curved due to the interaction exchange of vacuum and magnet dipole, and therefore still contains extra energy. The operator inputs no energy whatsoever after the magnet is once emplaced, yet has appreciable magnetic field energy and field-free magnetic vector potential energy available to utilize. As an *energy transducer*, the MEG has a COP»1.0. As a power system powering loads, its COP>1.0.

11.5.2 An Example of MEG Operation

As a replicable example, suppose one operates the MEG core at less than saturation, and places a small "square wave" pulsed signal upon a primary coil wound around the core. The core changes its permeability as the internal magnetic B-flux is perturbed and changes, so all the perturbed B-flux is retained in the core material. The large E-fields produced by the changing B-flux in the core as a function of $-\partial \boldsymbol{B}/\partial t$ are *not* localized, but pass readily out of the core and interact with the coils, particularly with the secondary.

In the A-potential region in the space outside the core, one also produces very large E-fields by $\mathbf{E} = -\partial \mathbf{A}/\partial t$. At the frequencies being used (40 to 80 kHz nominally), all these E-fields produced by the core flux perturbations and by the external A-potential perturbations are essentially in phase. The large E-fields thus coherently add and interact directly with the secondary coil acting as a receiver.

In the secondary, the very large and sharply changing E-fields interacting with the coils also produces large B-field flux changes in the core. However, again the core changes its permeability and holds-in all the extra B-flux. The changes in that secondary B-flux also make additional non-localized E-fields, and so on.

The surprising and shocking result is that the transformer secondary undergoes a coherent, purely E-field reaction with the electrons in its coils. This results in the output voltage and current being in phase (within 2°, since there are inevitably some small remaining inductance effects due to a tiny bit of external B-field leakage from the junctures of the ends of the permanent magnet with the core material).

The magnitude of the powerful E-fields produced by these processes — and interacting with the secondary coils — depends on the rise and decay times of the input signal's pulse edges. So the *average* power input to the signal coil may be small, but — using sharp rise and decay times of the input pulses — very large E-fields will be produced in the adjacent surrounding space outside the transformer section and interacting with the secondary coils, as well as in the sharply changing B-flux localized in the core.

One then uses transmission-reception theory and near-field antenna theory for separate receiving circuits containing loads, added in the large E-field region in space outside the core. Those E-fields arise from both the changing A-potential outside the core and the changing localized flux inside the core. We stress again that the E-fields produced in the core are not localized, but pass outward into the external space to interact with whatever interceptor/receivers are placed out there.

A "outrigger array" version of the MEG uses separate external receiving circuits, containing loads, to intercept, collect, and use more of the EM energy available in the surrounding space outside the core.

Within physical space limits, the energy intercepted and utilized by these *separate, independent receiving circuits containing loads* can be increased by adding more receiving circuits with loads or adjusting the rise and decay time of the input pulses to provide greater collected voltage. These *separate* loads will be powered by the incident E-fields' interception in the external receiving circuits. This is not at all a *transformer* action, but a *near-field transmission and separated multiple antenna circuits' reception* action.

There is no back-current coupling between the external circuits and loads with the primary circuit of the transformer section. In the external A-potential region alone, the secondary coil in the transformer section will interact with the large incident E-fields from that adjacent space and from the changing B-flux inside the core, producing at least as much energy as one inputs to the primary. Actually, that process alone will produce more output energy in the secondary than the operator inputs in the primary.

If one operates the core below saturation, the magnetic flux in the core will also be switched, giving additional intercepted E-field energy in the secondary coil attached to its load. The total work out in all loads is permissibly greater than the total energy dissipated in the input by the operator's energy input {668}. The MEG is an open system far from equilibrium with the active vacuum {670}.

11.5.3 The MEG Is Replicable

Any university laboratory can replicate this effect, and demonstrate more work in resistive loads than is input to the MEG primary from the external power supply. For replication, we suggest initially designing to COP = 1.5, since beginning at COP = 2.0 there *can* be other effects not in the U(1) EM literature, but which are dealt with in the actual MEG. Some of the methodology {671} used for COP >2.0 is still proprietary information not yet releasable due to patent law requirements, while some of it was included in Chapter 7.

Without elaboration we advance the *supersystem* concept, and strongly insist that, to properly analyze the performance of any similar EM circuit or system, one must analyze the dynamics of the entire supersystem consisting of (i) the physical system and its dynamics, (ii) the curved local spacetime and its dynamics, and (iii) the local vacuum and its dynamics {672}. All three components of the supersystem are mutually interacting.

For a COP>1.0 system, the mutual interactions are utilized from disequilibrium processes evoked between all three supersystem components. U(l) electrodynamic analysis system in an inert vacuum and a flat spacetime is insufficient. Previous AIAS analysis {661} has addressed the supersystem by use of higher O(3) symmetry electrodynamics to include vacuum dynamics (energy currents, both space-like and time-like) and curved spacetime dynamics.

11.6 Physical Considerations of Gauge Freedom

Additional physical considerations must be accounted when *gauge freedom* is invoked or utilized with respect to physical systems. Applied to electrodynamics systems, gauge freedom assumes the ability to freely change any potential of the system being studied. From the physics of the situation, a system potential change concomitantiy assumes changing the *potential energy ofthe system itselffor* which gauge freedom changes are being theoretically applied to system-descriptive equations.

If one *can freely increase* the potential energy of an EM system, the extra energy has to come from a physical asymmetry of the system with its active environment — the active vacuum. Else, one must surrender the conservation of energy law. At worst, gauge freedom would then imply the creation and destruction of energy at will. At best, it would imply a return to the view that potentials are mere mathematical conveniences with no physical reality — a notion falsified for decades by the AB effect {647}. ***freely** decreasing* the potential energy of the system by simply depotentializing it, has a similar vacuum asymmetry requirement for de-excitation of the system, where energy must flow from the system to the vacuum.

11.7 Special Modeling of the Vacuum

Having EM energy density, the vacuum *is and should be modeled as* a scalar potential of great magnititude. For deeper (supersystem) analysis, one will also have to examine the internal EM longitudinal phase conjugate wavepair structure and dynamics of the scalar potential vacuum {672}, along the lines shown by Whittaker {644} in 1903 and re-interpreted by the present author {643} in 2000. From quantum field theory arguments, Mandl and Shaw {646} give strong support to the proposed reinterpretation. They argue that neither the scalar (time-polarized) photon , nor the longitudinal photon is independently observable, but the combination of the two is observable as the instantaneous scalar potential.

Transforming this argument to EM waves yields a solution consistent with the reinterpretation {643}. The result is that all 3-space EM energy comes from the time-domain, in a manner not previously utilized in EM circuit analysis. It is experimentally established by proving (measuring) that the source charge and dipole do not *receive ob*servable EM energy in 3-space, but continuously *output* observable EM energy in 3-space. Simply making and destroying dipoles in an electrical system turns out to be a method for direct engineering of negentropy and entropy respectively — or, from another viewpoint, it directly engineers regauging.

The greatest unrecognized problem in conventional electrical power system engineering is its total failure to include the making and destroying of dipoles as direct engineering of negentropy and entropy respectively.

Every EM potential in an EM system is a change to the energy density of the ambient vacuum potential — and thus is a spacetime curvature — or it is a change to an intermediate potential that *is* such a vacuum potential change. Vacuum exchange with every EM system and asymmetry in that exchange, is ubiquitous and cannot be prevented. Whenever one makes or changes an EM potential, field or wave, and whenever one regauges, one invokes and utilizes asymmetry in the interaction of the active vacuum and the system, either wittingly or unwittingly. Else one must discard the conservation of energy law altogether. There is no symmetry of a mass system alone {673}, without including the vacuum's asymmetrical interaction with it. The reason is simple: all observables involve asymmetry {674} *a priori*, and mass is an observable.

11.8 Every Charge and Dipole in an EM System Receives and Transduces Vacuum Energy

In particle physics, every dipole — and every charge considered with its associated clustering virtual charges of opposite sign — is recognized as an asymmetry in the vacuum flux {675}. The very definition of broken symmetry means that a charge or dipole receives and absorbs virtual photon energy from the vacuum, transduces it into observable form, and re-radiates it as real, observable EM energy. This results in an observable outpouring of 3-energy from the charge or dipole, with no observable 3-energy input.

To save the conservation of energy law for physical EM systems, one must therefore recognize and model the *physical vacuum asymmetry mechanism* **that** allows this demonstrated performance of charge and dipole. In electrical power systems, that required asymmetry mechanism is the known broken symmetry of the vacuum's interaction with the source dipole created in the generator.

11.9 Energy from the Vacuum Powers Every Electrical Power Line and EM Circuit

11.9.1 What the Generator Powers and What Powers the Circuit

Shaft power input to a conventional generator allows the generator to produce its source dipole, nothing else. The generator does not transduce any of its input shaft mechanical energy to add EM energy to the external circuit connected to its terminals. Instead, it first transduces its mechanical shaft energy input into internal magnetic field energy.

Then it totally dissipates that magnetic energy to separate its own internal charges, forcing the positive charges in one direction and the negative charges in the other, making the source dipole connected to its terminals. That is all that the mechanical shaft power input to a generator does — and **that** is all that burning hydrocarbons, consuming nuclear fuel rods, **damming** rivers for power stations, putting up windmill-powered generators, using solar energy input to solar cells, or consuming chemical energy in a battery accomplishes.

Once made, the source dipole — via its broken symmetry in its vacuum exchange - freely receives and absorbs enormous virtual photon energy **from** the active vacuum, transduces it into observable (3-space) form, and pours out real, observable EM energy in 3-space from the generator terminals to which it is connected. As shown in Figure 1-1 of Chapter 1, this outpouring of EM energy from the generator terminals fills all space around the external circuit {653} with a vast energy flow almost parallel to the conductors. Figure 2-4 is a lateral view so that the two flow components can be clearly seen and compared.

11.9.2 Gedanken Experiment

We advance a gedanken experiment to illustrate. Consider a perfect generator with no losses, and with two perfect conductors and a purely resistive load R_1 connected across its terminals. Call the mechanical energy input to the shaft of the generator W_{in}. Call the Poynting diverged energy flow component W_r that is intercepted by the attached circuit and powers R_1. Call the Heaviside nondiverged energy flow component W_h that misses the circuit and is wasted. As an *energy transducer,* the total energy output W_{out} from the terminals of the generator is

$$W_{out} = W_r + W_h \qquad [1]$$

However, $W_{in} = W_r$, for this theoretically perfect system. So

$$W_{out} = W_{in} + W_h \Rightarrow W_{out} > W_{in} \qquad [2]$$

Further, when accounted, $W_h >> W_{in}$, so

$$W_{out} >> W_{in} \qquad [3]$$

We conclude that all EM systems — *as energy transducers freely extracting and transducing EM energy from the active vacuum* — already exhibit COP » 1.0. However, for *useful work* in a closed current loop circuit, the COP = 1.0, unless additional antenna-receiver circuits or another scheme such as Bohren charge resonance is used to increase the energy collected and dissipated in the load while *not* increasing the energy used to dissipate the source dipole.

11.10 Significant Developments in Negative Resistance

True negative resistance (e.g., current moving against the voltage) also exists and is experimentally replicable. In true negative resistance, the current in a circuit moves against the voltage instead of with it, suggesting time-reversal effects in the mechanism by which EM force and charge interact.

11.10.1 Chung's Negative Resistor and Naudin's Replication

Chung {676} has reported replicable negative resistance in the interfaces between layers of carbon fibers in a composite material. By varying the pressure used when the fibers are made, the resistance can be controlled to later be negative or positive as desired. A simpler version of Chung's negative resistor using off-the-shelf materials has been successfully replicated and tested by Naudin {677}. (See Figure 11-3). Naudin's website {677} provides measurements and full details necessary for repetition. He also includes sources of the carbon fiber materials needed lor independent replication.

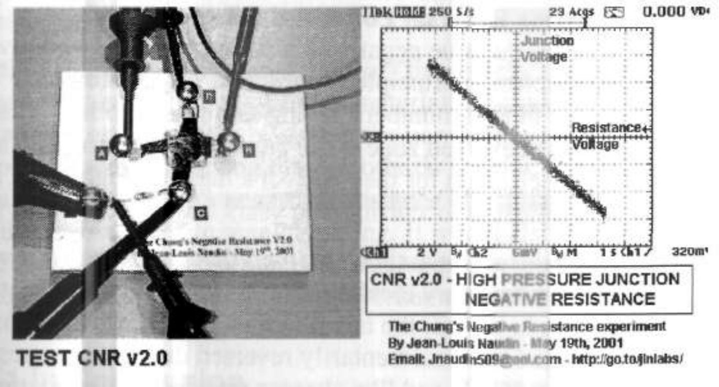
A
B
C
D
The Chung's Negative Resistance V2.0
Jean-Louis Naudin - May 19th, 2001
Tek 250 S/s
23 Acqs
0.000 VDC
Junction Voltage
Resistance Voltage
Ch1 2 V
Ch2
M 1 s Ch1
320mV
CNR v2.0 - HIGH PRESSURE JUNCTION NEGATIVE RESISTANCE
The Chung's Negative Resistance experiment
By Jean-Louis Naudin - May 19th, 2001
Email: Jnaudin509@aol.com - http://go.to/jlnlabs/
TEST CNR v2.0

Figure 11 -3 Naudin's replication of Chung's negative resistor.

11.10.2 The Point Contact Transistor

There is also evidence that the original point contact transistor — which used spring pressure on the point contact — often exhibited true negative resistance. Certainly this transistor was never adequately understood {678}, and it is not understood today, since its supersystem has never been analyzed.

11.10.3 Kron's Negative Resistor

Gabriel Kron, early pioneer in electrical systems theory, reported **developing** a negative resistor in the 1930s, on the U.S. Navy's Network Anayzer project {679} at Stanford University. He was not permitted to reveal the details of his negative resistor, but did provide strong hints in certain statements in his papers in the literature {680a, 680b}. His method for a self-organizing network depended on his "open path" {681} which he also was not permitted to reveal.

The present author worked for some years with Floyd Sweet, a Kron protege", who did produce a successful type of powerful negative resistance system {682} known as the vacuum triode amplifier, apparently based on Kron's original negative resistor. Sec Figures 6-5 to 6-13 in Chapter 6. As shown in Figure 6-6, Sweet induced self-oscillation in barium ferrite magnets {683a-683c} (which are made by powder metallurgy methods using pressure applied during their manufacture) and extracted energy by induction from the resulting self-sustained magnetic field oscillations. His device was independently tested by Walter Rosenthal {684}, a senior test engineer of long experience with a major aerospace company. See Figure 6-9 and Figure 6-13.

11.10.4 Proposed Time Reversal Zones and Effects

One may take the view that true negative resistance involves time-reversal, which by the Pauli exclusion principle is permitted for bosons {685} and possibly for fermions in even numbers, acting as quasi-bosons (perhaps only possible in a time-reversed zone). The present author proposed a photon interaction mechanism for the "flow of a mass through time" in 1971, with initial publication occurring two years later {686}.

Extending this mechanism, the author proposed that *time reversal zones* (TRZs) can occur in momentary local regions in semiconducting materials or electrolytes {687}. In a TRZ, the law of attraction and repulsion of charges is hypothesized to be momentarily reversed due to time-stress or time-domain pumping {688}, and like charges may momentarily attract. In Chapter 10, we used these time-reversal zones to explain the anomalous results of the 600 or so successful cold fusion experiments worldwide. We also gave the exact new nuclear reactions for the production of anomalous deuterium, tritium, and alpha particles in many of these experiments.

When under strong stress, anomalous attraction of clusters of like charges is experimentally known, e.g. in Shoulders' {689} experiments and patented processes {690}. When the output leads of Sweet's vacuum triode amplifier device were suddenly shorted, extreme cooling occurred rather than heating. Ice condensed instantly upon the shorting contact and leads from the sudden extreme cooling of the metal in the moisture in the air, dramatically showing a converging EM energy process rather than a diverging one.

11.10.5 Effects in Optically Active Semiconducting Materials

In addition, four-wave-mixing, phase conjugation, and time-reversal effects are well known in semiconducting materials {691}, almost all of which are optically active. Optical activity and microlasing is known in semiconductor powder {692}. Further, the vacuum itself can exhibit semiconducting properties {693}.

11.10.6 Special Functions Produced by Stress

Apparently pressure (stress) in many of these devices and experiments can play an important role in the negative resistance process. Based on harmonic sets of paired couplings of a longitudinal time-like wave and a longitudinal wave in any scalar potential {643}, the present author hypothesizes that such stress — treated as a stress scalar potential — accomplishes time-domain pumping in the optical sense, due to the presence of the coupled time-polarized longitudinal waves in the time domain.

This time pumping is a special form of "time stress". Thus in optically pumped or stressed optically active materials (including electrolytes), the time-stress creates a TRZ, reversing the normal force laws upon and between charges. Assuming this hypothesis is valid, this could account for the variation of Chung's negative resistance effect with the variation of the pressure (stress) used in manufacturing. It would also apply to the laser-created stress in semiconducting powder experiments, etc. The hypothesis should be further investigated for experimental validation or refutation.

11.11 In Conclusion

As Penrose {694} states,

> *"Electromagnetism is a subject that is in no way closed to stimulating new developments."*

Indeed, electrodynamics is still very much an advancing subject, and in classical electrodynamics there are foundations errors and problems, a few of which we have addressed. As Bunge {695} pointed out more than three decades ago,

> *"...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state. "*

From a foundations viewpoint, Einstein {696} beautifully summed it up — and advanced a requirement for continually examining the foundations of electrodynamics and every other field of physics — in these words:

> *"..the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools **of thought,** as something obviously, immutably*

> *given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."*

Though they have long been neglected, legitimate COP>1.0 electrical systems are good science, or they can be. They were present in the original Heaviside truncation of Maxwell's theory prior to Lorentz's symmetrical regauging. That regauging arbitrarily selected only those Maxwellian systems in equilibrium in their exchange with the active vacuum. The energy flow theory was further curtailed by Lorentz's integration trick, which discarded the huge vacuum-furnished non-intercepted Heaviside energy flow component accompanying every EM system. Every EM system already outputs for more EM energy flow than the energy that the operator inputs; thus every EM system is already a COP»1.0 system, with respect to *energy transduction.* The only way this giant negentropy of every EM system can be hidden is to ubiquitously continue that integration trick to discard the giant Heaviside energy flow component.

The present Lorentz-regauged Maxwell-Heaviside equations are not consistent with *any* bipolar EM system or circuit. The ends of any dipolarity in the universe — including any electron on Jupiter and any positron in the star Sirius — constitutes a broken symmetry in the seething vacuum flux. Actually, that is what Gabriel Kron discovered and called his "open path". Hence, that "dipole" continuously emits observable EM energy extracted from the vacuum.

Engineering use of the ubiquitous closed current loop to contain both the primary source dipole and the load — which destroys the source dipole and extraction of vacuum energy faster than the load is powered — is responsible for the successful predictions of the equilibrium EM model (which of course completely fails to explain the source charge or source dipole). That circuit is unwittingly designed to self-enforce the Lorentz symmetry condition and therefore self-enforce those predictions.

The question as to whether COP>1.0 electrical power systems are possible *and practical* must be resolved by direct experiment, and not by established models and cherished concepts that already arbitrarily exclude such systems! Such theoretical refutation by models that in fact contradict the experimentally observed COP = oo action of every charge in the universe, cannot be trusted at all with respect to COP> 1.0 judgment. Well-known supporting and confirming experiments already in the literature — such as the Bohren experiment yielding COP =18 — have been cited, and these are replicable by any university laboratory. The Bohren-type experiment is well known and routinely performed in every nonlinear o**ptics** laboratory, under the guise of "negative resonance absorption of the medium".

It is indeed a critical time to re-examine the foundations of our electrical **power** science and engineering, along the lines so eloquently stated by Albert Einstein. With the impending oil crisis, we face a world economic crisis of epic proportions unless we rapidly develop and widely deploy sell-powering disequilibrium electrical power systems.

Also it is a critical time when the leaders of the scientific community — specifically the National Academy of Sciences, the National Science **Foundation,** the great national laboratories, and the universities — must at last face and correct the excruciating *faux pas* that has now been perpetuated in electrodynamic theory for more than a century. In short, the le**aders** of our scientific community must now show whether they are scientists practicing and upholding experimental basis, or are dogmatists **defending** an ancient faith. The fact that the source charge problem c**ontinues** to be propagated in electrodynamics 45 years after the award of the Nobel Prize to Lee and Yang, is no longer excusable.

A thorough review and correction of the foundations of EM theory is re**quired** for the survival of this nation and much of civilization. It is also re**quired** for the cleanup of the increasingly polluted biosphere, survival of the threatened species of the Earth, reduction of present global warming **trends,** and the survival of this fragile and beautiful planet itself.

Annotated Glossary of Selected Terms

action at a distance: An action (effect) occurring in a local system, for which no separate local 3-space cause exists, but for which a causal correlation exists between a distant action and the local effect.

advanced electromagnetic field: An electromagnetic field that is a solution of the classical Maxwell equations, which is positioned on the future light-cone of spacetime. It may be possible to involve such a field in the tempic force that may exist between the tempic potential at the output of an overunity system and the tempic potential that exists at the input of the system, where the interaction of the tempic force with the local nonlinear vacuum drives Dirac sea hole current from output section to input section.

advanced electromagnetic potential: An electromagnetic potential that is a solution of the classical Maxwell equations, and is positioned on the future light-cone of spacetime. Presently this potential has not yet been given a physical interpretation. It may be possible to involve its effects in a specific system by our notion and use of the instantaneous tempic potential difference (the tempic force) that may exist between the tempic potential at the output of a system and the tempic potential at the input of that system if the two differ.

advanced electromagnetic wave: An electromagnetic wave that is a solution of the classical Maxwell equations, and which is positioned on the future light-cone of spacetime.

Aharonov, Yakir: Renowned Israeli quantum physicist and former student of David Bohm. Co-discoverer of the Aharonov-Bohm effect. Together with Bohm, Aharonov authored a fundamental paper in 1959 that pointed out the *primary importance of the potentials* rather than the force fields. The force fields are only made in and of the charged particle system itself, and do not exist as such in the vacuum. The force fields are thus not primary electromagnetic causes at all, but are *effects* due to the interference of potentials in a particular charged particle system.

Aharonov-Bohm effect: Quantum mechanical phenomenon theoretically pointed out in 1959 by Yakir Aharonov and David Bohm whereby interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields. A system undergoing a cyclic evolution from an initial state back to it will acquire a phase difference, which can be detected by interfering the initial and final states. Theory was extended later to Berry phase, then again to geometric phase.

In classical EM terms, a more important *energy-amplifying* principle - advanced by the present author — emerges from the AB effect. My colleagues and I have used this principle in our motionless electromagnetic generator (MEG) to provide COP>1.0. Consider a source of magnetic field B, such as a permanent magnet. All the magnetic field energy B from the permanent magnet may be drawn into and confined in a local material path, as in a special nanocrystalline transformer core. Then in the space surrounding the localized magnetic field energy, the active vacuum will freely replenish — via the broken symmetry of the permanent magnet dipole in its exchange with the vacuum — the magnetic energy in that space that would otherwise have been filled with magnetic field B (curled magnetic vector potential A). However, nature replenishes the energy in space outside the B-localization region with the *curl-free* magnetic vector potential A. In this fashion a system designer may cause the active vacuum to provide more available magnetic energy from the permanent magnet that normal, with each of the two energy forms being in separate spatial regions and paths, and occurring in different forms of magnetic vector potential energy. By inputting only a little energy in the input coils of the transformer section, and controlling the rise time and decay time of the pulses, one can produce very large E-fields from the A-potential via $E = -dA/dt$. From the simultaneously perturbed confined B-field flux in the core, one also produces E-field simultaneously and — at the frequencies used — in phase with the E-fields produced outside the core. The net result is interaction of large E-fields with the output coil, producing an output power *proportional to the rate of change* of the weak input signal energy rather than its magnitude. In this fashion, additional E-field energy from the vacuum enters the output coils to power the secondary circuit and its loads. Also, the output of the entire transformer section has the voltage in phase with the current, in contradiction to the actions of all previous transformers.

AIAS: Alpha Foundation's Institute for Advanced Study, directed by Dr. Myron W. Evans, a noted scientist with more than 600 papers in the literature. The AIAS is an unusual "think tank" comprised of Fellows and Fellow Emeriti in various places in the world, who cooperate in advanced electromagnetics studies via E-mail. The Director, Dr. Myron W. Evans, is a theorist of note, as are several other Fellows. The AIAS has been forging ahead with the production of O(3) symmetry electrodynamics, particularly as an important subset of Mendel Sachs' unified field theory. Dozens of AIAS group-authored papers have been published in various leading journals such as *Foundations of Physics, Physica Scripta,* etc. More than 100 AIAS papers are carried on a restricted Department of Energy website for reference by DoE and AIAS scientists. The work is particularly significant in preliminary theorizing how EM energy can possibly be extracted from the active vacuum.

algebras: The various types of mathematics in which the theoretical models for physics and other sciences are embedded. Some important algebras utilized in physics are vector algebra, tensor algebra, quaternion algebra, Grassmann algebra, Pauli algebra, Clifford algebra, etc. There are many others. The complexity of the symmetry and of the topology varies significantly between these algebras. Thus, an electrodynamics model in a lower topology algebra such as vectors or tensors will not reveal or permit nearly so great a set of electrodynamic functions and operations as will electrodynamics embedded in a higher topology algebra such as quaternion algebra or Clifford algebra. See topology.

ambient vacuum: The average or typical "standard" active vacuum (i.e., active spacetime), removed spatially from any large collections of observable charge and observable mass.

ambient vacuum potential: The *average* or typical standard vacuum considered as a scalar potential because it possesses energy density as well as structure and dynamics on many levels. As a scalar potential, the vacuum potential (considered without zero-point fluctuations) decomposes into a harmonic set of "bidirectional" longitudinal EM phase conjugate wavepairs, by Whittaker's 1903 decomposition as reinterpreted by the present author in 2000. Dynamic structuring of the vacuum thus consists of dynamics impressed upon these wavepairs. Since the vacuum is observably mass-free, Whittaker's phase-conjugate half of the decomposition must be considered prior to its interaction with mass, hence in the imaginary plane **and** arising from the time domain. This consideration led to the discovery by the present author of the more fundamental 4-symmetry in EM energy flow between the time domain (imaginary plane) and real 3-space, when broken 3-symmetry is present (as with any dipolarity or with any **potential).**

Powerful support for this solution is also given by F. Mandl and G. Shaw, Quant*um Field Theory,* Wiley, 1984, Chapter 5. Mandl and Shaw argue tthat the longitudinal and scalar (time) polarizations of the photon are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. Our comment is that this argument, translated from particle terminology to wave terminology, directly fits our re-interpretation of Whittaker's 1903 decomposition of the scalar potential.

angular momentum: The momentum or "leverage" of the linear momentum of a moving body, with respect to an axis or reference point. Angular momentum has the same units as action; that is, it is energy multiplied by time, or momentum multiplied by length. In terms of particles, it is also known as *spin*.

anti-circuits: Somewhat normal-appearing electromagnetic circuits which primarily process and use antiparticles and anti-electrons rather than particles and electrons, or else process a combination of the two simultaneously. Operations of the anti-circuit must be analyzed and understood in terms of its *supersystem* interactions. The circuits process and utilize both negative energy and positive energy EM fields and potentials, hence are closely connected to unified field engineering and antigravity. One aspect enabling such circuits is the use of the supersystem's curvatures of spacetime in both negative curvature and positive curvature fashion so that an appropriate change of spacetime curvature precludes the radiation normally accompanying pair annihilation, etc. The key to the circuits is T-reversal and C-reversal operation, along with the matching use of spacetime curvatures, and also the use of the Dirac "positron" (negative energy electron) — and its fields and potentials — prior to its observation. The unobserved Dirac negative energy electron in the vacuum is an entirely different entity than is the observed Dirac positron after its interaction with charged matter, after which it has become a "lattice hole" or "positive charge on a positive mass ion". The response of normal circuit components to vacuum Dirac sea hole operations is dramatically different than for Dirac positrons, and for certain applications special versions of circuit components are required. At this writing, together with John Bedini the present author has submitted what may be the first patent application on the fundamental process for anti-circuits and their operations with antimatter and negative energy currents, fields, and potentials as well as matter and positive energy currents, fields, and potentials. Another patent application by this author **is** in preparation, covering the anti-gravity aspects. The details of these mechanisms are still proprietary at this writing in order to protect our intellectual property rights, but limited information is in this book.

anti-electron: A negative-energy electron, observed as *a positron.* Under time reversal, electric charge reverses in sign. Dirac did not predict the positron per *se,* but a negative energy electron that we would *observe* in reversed fashion. Present theorists assume the interaction of the 4-space negative energy electron with matter — i.e., observation — to have occurred. This transforms the negative energy, negative mass electron to a positive energy, positive mass positron traveling in the opposite direction. In so doing, antigravity from the negative energy 4-electron source charge, due to its negative energy EM fields and negative energy potentials, is arbitrarily transformed by the spatial reversal into positive gravity fields rather than the antigravity fields they were. The secret to practical antigravity is to utilize (in anti-circuits) the negative energy 4-electrons prior to observation, *without* transformation of (i) negative energy fields to positive energy fields, (ii) negative mass to positive mass, and (ii) reversal of the field direction. In short, simply consider the unobserved negative energy 4-electron as what it is, and utilize it in its unobserved vacuum state. The conventional transformation process also unwittingly discards the concept of anti-circuits, anti-fields, anti-potentials, and anti-currents, which discards half the rich electromagnetic circuitry that can be built and utilized. In the present book, we only touch lightly on this latter subject, which hopefully will be the subject of a future book after certain patents are filed.

antigravity: Negative gravity or reversed gravity. Use of negative curvature of spacetime to produce a force on a gravitational source (physical system) that is opposite to the normal force of gravity of the earth etc. on that system. Antigravity is achieved by the use of negative-energy, negative-mass source charges in the mass component of the supersystem prior to observation, while the source charges are still 4-spatial. This usage produces negative energy fields and negative energy potentials to negatively curve the local spacetime (ST), so that these negative *local* ST curvatures interact back upon the positive mass source system associated with the vacuum negative energy sources in its supersystem. Again, to understand the operation of an antigravity circuit or system, one must analyze it in terms of its *supersystem* interactions.

antimatter: Matter consisting of atoms that are composed of anti-electrons (positrons), antiprotons, and antineutrons and such. Also loosely refers to the antiparticle corresponding to a particle; an antiparticle may be regarded as a particle traveling backward in time, or "phase conjugated", or "time-reversed. Since time is not an observable, the observer sees the antiparticle *{after its observation by interaction with mass)* in forward time as being reversed in direction and charge.

antiparticle: Conventionally, a counterpart to an ordinary fundamental particle, *observably* having identical mass, lifetime and spin, but with charge and magnetic moment reversed in algebraic sign. We again point out, however, that the antiparticle differs when in its unobserved 4-spatial state. Only by analyzing the particle's supersystem can the true unobserved antiparticle be understood, as well as its fields and potentials in that condition.

antiphoton: A time-reversed photon. Presently the photon is considered its own antiparticle. *Observably* it is, but in its pre-observation condition it is not. An alternate but as yet unaccepted view is that, since the unobserved photon is comprised of (dE)(dt) — i.e., a piece of energy welded to a piece of time with no seam in the middle, the antiphoton may be assumed to be comprised of (-dE)(-dt) with respect to the external observer.

anti-Stokes radiation: The radiation coming from an anti-Stokes emission process from an intensely scattering medium, wherein more energy flux is emitted than is input by the experimenter-operator. The excess emitted energy is furnished by other processes in the medium. For true overunity operation, the excess energy furnished by the energetic medium must be freely replenished to the medium by the active vacuum's interaction with it.

antiwave: The time-reversal (phase conjugation) of a reference wave.

A-potential: The magnetic vector potential A, conventionally thought to be "defined" by the equation $\mathbf{B} = \nabla \times \mathbf{A}$. However, no equation is a definition (an identity is required). If we replace the "=" sign by "==", we have $\mathbf{B} \equiv \nabla \times \mathbf{A}$, which is actually seen to define the field B as just a curled A-potential, or as the curled component of the A-potential, *after* interaction with magnetic charge. The A-potential or an additional component of it can and does also exist in uncurled fashion, and will emerge in the uncurled state in the space outside any local path (e.g., the path in the interior of a toroidal coil) that retains the B-field therein. That process of separating the curl component of the A-potential from its uncurled component is known as the Aharonov-Bohm effect (which also affects the wave function). The motionless electromagnetic generator (MEG) uses a special nanocrystalline core material in a transformer configuration, which draws in and retains the B-field from a permanent magnet outside the core but firmly in contact with it on both ends. As a result, the vacuum freely replenishes the withdrawn magnetic energy that would otherwise have existed in the space outside the core, but the energy outside the B-field localization zone remains and appears as a separated and uncurled A-potential. In this way, the MEG is converted to an open system not in equilibrium with the vacuum, and can therefore permissibly output more EM energy in its output coils than the operator inputs into the input coils, the excess being freely furnished from the vacuum via the broken 3-symmetry of the permanent-magnet dipole and the resulting 4-symmetry EM energy flow from the vacuum's time domain into the dipole. The MEG is actually a practical macroscopic EM power system application of (1) the well-known Aharonov-Bohm effect and (2) the effect of the separation of the curled and uncurled potentials by a long tight solenoid or by a tight toroid, but in this case by the nanocrystalline magnetic core performing that separation function. In effect, the output coils interact with double (or greater) energy than would be available in just the B-field of the core (and of the magnet). Dual interaction with both the internal core B-field flux and also the external A-potential is obtained by pulsing, so that the pulse edges produce strong E-fields in space, with the E-fields striking the output coils in a transmission-reception near-field fashion. As is well-known., $d\mathbf{A}/dt = -\mathbf{E}$, and also $f(d\mathbf{B}/dt) = \mathbf{E}$. These two fields produced by the perturbed internal B-flux in the core and the pertubed external uncurled A-potential outside the core are in phase. This is how the extra pulsing A-potential energy outside the core is made to interact as -E-field energy with the Drude electrons in the output coils, while the changing magnetic flux in the core through those coils also simultaneously produces a second -E which interacts with the coil's conductors in normal E-field fashion. The total E-field interacting with the output coils is proportional to the *time rate of change* of the input perturbation signal, not the magnitude of its energy. By adjusting the leading and trailing edges of the perturbation pulses, the magnitude of the E-field interactions with the output coil is determined as desired. Further, the usual output B-field produced in the output coils by dE/dt is directly extracted into the core and held, so that *a purely electrical interaction and output occurs in the output coils.* This is easily shown since the output current and voltage of the output coil are in phase, within a degree or two. The MEG may be said to the world's first purely electrical induction transformer, freely separating the magnetic interactions and bottling them up, while allowing the electrical interactions.

asymmetrical regauging: A change of either the scalar potential ϕ or the vector potential A, or both, so that a *net resulting excess force together with a net change of total system potential energy* occurs in the system. During this change, an asymmetry exists in the active vacuum's energetic exchange with the system. The result is that the potential energy of the system is freely changed by energy from the vacuum, and a free net force is also produced in the system. This new force is then free to dissipate the excess potential energy by translating electrons through a load and powering it. By contrast, the Lorenz/Lorentz *symmetrical* regauging produces two free force fields that are equal and opposite, so there is no *net* resultant translation force available to translate electrons through the external load and power it. Thus Lorentz regauging arbitrarily applied to the Maxwell-Heaviside equations selects only that resulting subset of Maxwellian systems that are in equilibrium in their interaction with the external active vacuum. The potential energy of the system is changed, but only as a "stress potential" doing internal work upon the system to form and maintain its increased stress. The extra energy also rotates the system's frame out of the laboratory frame. Lorentz thus unwittingly discarded an entire class of permissible Maxwellian systems that are not in thermodynamic equilibrium with the vacuum. The Lorentz condition is initially violated each time anyone potentializes an EM circuit, but the closed-current-loop circuit design with the same charged-particle energy carriers around the loop — including back through the back emf of the source dipole in the generator — enforces a form of Lorentz symmetrical regauging during discharge of the system's excitation energy. Further, the response of the electrons in the usual electrical circuit with copper conductors is so rapid that the asymmetrical regauging (potentialization) is almost instantly converted to symmetrical discharging by movement of the electrons as current. When one section of the otherwise closed current loop uses a dramatically heavier charge as the energy carrier (as a battery uses ion current between its plates, but uses electron current between the outside of the plates and the external circuit), it is then possible to dephase the two independent carrier currents so that the battery will be charging while simultaneously the circuit is being powered, by a negative resistor (sharply increased scalar potential or voltage) produced at the plate surface interface of the two carrier current types. This proven Bedini process is an example of a true overunity process, capable of producing a self-powering system by means of adroit switching and timing.

asymmetry: Lack of symmetry, or "broken symmetry" in a general sense. Since a specific symmetry is accompanied by a specific conservation law, the breaking of that symmetry represents the violation of that conservation law. Particularly note that a broken symmetry at one level often (usually) results in a new symmetry at a higher level, necessary to maintain the asymmetry at the lower level. The resulting "levels" of symmetry are known as *hierarchies* of symmetry.

asymmetry of the dipole: Because of the proven asymmetry of opposite charges, any dipole or dipolarity is a broken 3-symmetry in its fierce exchange of EM energy with the active vacuum, because of the known broken symmetry of the opposite charges on its ends. Any two points in the universe (in or out of the circuit), that are at differing ambient potentials, constitutes such a broken symmetry continuously pouring out EM energy extracted from the vacuum. A dipole or dipolarity continuously receives and absorbs copious virtual energy from the seething virtual photon vacuum flux, and not all this absorbed energy is reradiated back to the vacuum in virtual form. Instead, some of it is coherently integrated into observable form and re-radiated in all directions as usable, real 3-space observable photons and EM energy. We found that the input energy to the dipole is from the time domain of 4-space. Hence the dipole's broken 3-symmetry results in the appearance of a new and far more fundamental 4-symmetry in energy flow between the time-domain and 3-space. This was the solution to the long-vexing source charge (or source dipole) problem in classical and quantum electrodynamics. Reinterpreting the Whittaker 1903 decomposition of the scalar potential reveals the mechanism for this 4-symmetry between time-domain EM longitudinal wave energy flow into a dipole and 3-space EM longitudinal wave energy flow out of the dipole, due to the presence of the proven broken 3-symmetry of the dipole. *EM energy flow is conserved in 4-space, but not in 3-space. Further, the 4-*symmetry energy flow from the source dipole (or source charge) continues as long as the source dipole (or source charge) remains intact so that the broken 3-symmetry remains intact. It is strongly indicated that this asymmetry of the dipole or dipolarity is what Gabriel Kron finally discovered as his "open path".

back emf: The counter-electromotive force, usually between the end charges of the source dipole in the external power source for an electrical circuit (in its overall closed current loop). When the electrons in the circuit are forced against the emf— as when the "spent" or depotentialized charges from the ground reference of the circuit are forced back up through the emf of the source dipole to re-potentialize the charges — then work is done upon the source dipole to scatter its charges and destroy the dipole. Removing the dipole stops its receipt of vacuum energy from the time - domain (which was due to its broken 3-symmetry, now removed). It also stops the former dipole's transduction of that energy, and its former output of 3-space energy flow along the circuit and in space around the circuit. When this flow is occurring, a small component of it strikes the surface charges in the conductors and is diverged into the circuit (by axial precession of the surface charges withdrawing the "immediate stubs" of their fields and potentials into the conductor) to power it as the Poynting energy flow component. The huge remainder of the energy flow in space surrounding the circuit simply misses the circuit and is utterly wasted. Destruction of the source dipole removes the transduction of EM energy from the vacuum. It thus stops the powering of the circuit. So additional energy must then be input to the power source to perform work on its scattered internal charges, force them apart again, and reform and restore that source dipole, in order to resume powering the circuit by energy extracted from the vacuum by the dipole. It is simple to show that such a closed loop circuit uses more of its collected energy from the vacuum to destroy its source dipole, than it uses to power its load. Hence the circuit self-enforces COP<1.0.

back emf (tempic): There are actually two emfs in a circuit, not one. There is a 3-space emf and also an entirely unrecognized *tempic* emf. In either case, under proper circumstances there can be a back-emf of either one or both. See discussions under **electromotive force, back** and under **tempic back emf.** The tempic back emf is very important to COP>1.0 circuits, where the system's output energy density in spacetime is greater than the energy density at the input section of the system. This difference in tempic stress potential provides a tempic force sweeping from system output section back through the system to its input section. The current is a current of Dirac sea holes (negative energy 4-electrons) moving in that local vacuum, from the output section back through the system to the input section. At the input section, this Dirac sea hole current reacts with incoming electrons from the external power system. It therefore "eats" incoming electrons, so that the power supply must now furnish additional electron current to (a) power the negation and "killing" of the Dirac sea hole current by filling those holes with electrons in the input section of the circuit, and (b) simultaneously furnish the additional usual current to power the system's input. The result is that the tempic back emf is nature's mechanism for decaying the excited COP>1.0 energy state represented by a system in overunity functioning condition. Unless this hole current problem is overcome, this decay mechanism will destroy the net overunity, and it will prevent any attempt at close-looping the circuit for self-powering by use of clamped and governed positive energy feedback from output to input. The Bedini method can be used to transform the detrimental Dirac hole current into ordinary electron current, providing the self-powering feedback loop needed. Or, Bearden's outrigger method of using a multi-system set of external COP<1.0 receiving circuits powered by radiation (Poynting EM energy flow) from a single central power source (energy flow radiator) can be used. Then already-positive energy feedback from one or more of the separate outrigger COP<1.0 receiver circuits can be fed back in normal fashion to power the central power unit. Note that the outrigger concept uses the bridging concept, so that an external COP<1.0 system can be powered from a central radiating system, and the external COP<1.0 system can collect and output more energy than is output in the *accounted Poynting component* of the central radiating system. The excess energy radiated comes from the radiated Heaviside component.

bare electron: The true electron without its partial shielding of virtual positrons that are attracted toward it from the active vacuum, and that cluster around the electron, partially shielding its charge as seen by an external observer. The bare electron has infinite energy and infinite charge, as does the shielding cluster of virtual positrons. The difference between these two infinite values of opposite charge is the finite "externally observed" value of the electron charge listed in texts and handbooks. This strange dipolarity of the source charge thus exhibits an asymmetry in its exchange with the active vacuum virtual photon flux. Hence this mechanism also solves the source charge problem — how the charge continuously pours out EM energy in all directions at the speed of light, establishing its fields and their potentials reaching out into or even across the universe, without any input *of observable* EM energy to the source

Barret, Terence W.: Noted modern electrodynamicist and consultant who works in SU(2)xSU(2) gauge symmetry extended electrodynamics, and other models, and was also one of the pioneers of ultrawideband radar, Barret is one of the vigorous modern electrodynamicists continuing to develop and extend the science of electrodynamics and the electrodynamics model itself.

Bedini, John: Noted audio engineer, leading overunity energy researcher, inventor of the Bedini process for creating a true negative resistor inside a battery used as primary power to a circuit, and inventor of the Bedini overunity process for killing the back mmf (magnetomotive force) in a magnetic motor. Bedini is also the inventor of the renowned Bedini amplifiers well known to audiophiles, the BASE process for holographic sound, and a remarkable process for cleaning the harshness from digital audio disks. In addition, he discovered a mechanism for transducing into usable positive energy the detrimental Dirac Sea hole current (negative energy electron flow) backflow in an overunity system, flowing from the output section back through the system to the input section and even on out into the external power supply. The Bedini process transduces this detrimental current into normal electron current, thereby providing most or all of the required input power to the input section. Bearden then explained the technical process for Bedini's discovered transduction procedure, and Bedini and Bearden filed a patent application on this process.

Berry phase: Generalization by Berry of the Aharonov-Bohm effect, for an adiabatic situation. Later Aharonov and Anandan further removed the adiabatic condition and generalized the Berry phase to the geometric phase. See Aharonov-Bohm effect, geometric phase.

blocking: In a Fogal charge-barrier semiconductor, stopping or partially stopping the flow of current in a circuit, or between two points, by *pinning the electrons without stopping the EM energy flow*. In ordinary usage, "stopping the progress of whatever charge flow or current is being referred to.

Bohm, **David J.:** World-renowned physicist and originator of the hidden variable theory and interpretation of quantum mechanics. See David Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," *Physical Review*, 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II). Together with his student Aharonov, Bohm co-authored a fundamental paper on the Aharonov-Bohm effect wherein interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields. From Bohm's hidden variable theory, the quantum potential has been highly weaponized by five nations, and these eerie weapons operating instantaneously at a distance through multiply connected spacetime are presently the dominant strategic weapons on earth.

Bohm's hidden-variable theory (HVT) of quantum mechanics: A major interpretation of quantum mechanics, formulated by David Bohm and published in *Physical Review* in 1952. Bohm's theory makes all the correct predictions and also eliminates many problems in quantum mechanics, such as the "measurement problem" and the — now recognized — problem of the missing chaos (i.e., the missing "hidden chaotic order."). Bohm's theory implies that physical reality — or much of it — can in fact be deterministically engineered. A key concept is the use of a *quantum* *potential*, which is multiply connected. Clandestine superweapons using **Bohm's** quantum potential (QP) have been developed and deployed by several nations. Present mutual assured destruction "standoff between these nations is based on surviving counters to QP weapons rather than surviving nuclear weapons.

Bohren, Craig F.: Distinguished Professor of Meteorology, Pennsylvania State University, noted for work in scattering theory. In an experiment **detailed** in C. F. Bohren, "How can a particle absorb more than the light incident on it?" *American Journal ofPhysics,* 51(4), Apr. 1983, p. 323-327 Bohren shows that in nonlinear conditions a resonant particle may absorb and output 18 times as much energy as is impinging upon it from the usual **field** definitions based on a static intercepting charge. The additional **energy** is "absorbed from the vacuum"; actually, it is absorbed from the nondiverged Heaviside energy flow component accompanying the accounted Poynting input component, but arbitrarily discarded by Lorentz. The resonant particle sweeps out a geometrical interception area (reaction cross section) greater than that of the same charge in a fixed static condition. Hence it intercepts part of the always-available Heaviside "dark energy" flow surrounding the Poynting energy flow component of every field/charge interaction.

Bohren experiment: Revealing experiment by Craig F. Bohren, replicated by Paul and Fischer, which shows that a given charge when in particle resonance absorbs more energy from a field or potential then the same charge does when in "static" position. This experiment is a direct experimental demonstration of the existence of Heaviside's nondiverged energy flow component around every field/charge interaction. See discussion under Bohren, Craig F. The Bohren experiment can be repllicated by any university nonlinear optics laboratory and outputs 18 times as much energy as is input by normal calculations (neglecting the Heaviside nondiverged EM energy flow component of the input).

boson: A particle having integral spin, which obeys Bose-Einstein statistics but not the Pauli exclusion principle. Bosons include the photon, alpha-particle, and any nucleus of even mass number.

bridge: A component or process which passes EM energy between two **isolated** circuits, but does not pass dq/dt between them, and which breaks the normal rigidly field-locked power dissipations in the two circuits. Space itself would appear to be one such bridge for EM energy flow.

bridging **function:** The function of passing EM energy flow between two isolated circuits, but not passing dq/dt between them, while simultaneously breaking any normal rigidly field-locked power dissipations in the two circuits. Use of the uncurled A-potential where $\nabla \times \mathbf{A} = \mathbf{0}$ is one available method of bridging otherwise isolated circuits, particularly in the outrigger embodiment of the motionless electromagnetic generator and other COP>1.0 systems.

broken symmetry: Fundamentally, a condition in which two parts of some configuration or shape, on opposite sides of some divisor or condition or boundary, are not similar but differ. In vacuum energy physics, we are very interested in the fact that — because of the broken symmetry of opposite charges — any charge (with its associated clustering virtual charges of opposite sign) and any dipole or dipolarity represents an asymmetry in its fierce energy exchange with the vacuum. This implies that something virtual becomes observable; i.e., part of the virtual EM energy continuously absorbed from the vacuum by the charge or by the dipole is changed into observable form and re-emitted as real, observable EM energy. The remainder is re-radiated as virtual EM energy. The charge's transduction of absorbed virtual photon energy and its re-emission as observable photon energy was proposed by the present author in 2000 as the solution to the long-vexing problem of the source charge and its associated fields and potentials (and their energy) reaching across the universe.

bypass resistor: In electrical theory, a resistor to pass or conduct current around (in parallel to) some other component. If another effect such as use of the overpotential, pumped phase-conjugate reflection, quantum well, quantum tunneling, negative energy, or charge blocking is added, the bypass resistor's function becomes extremely complicated. The bypass resistor may be complex, e.g., or it may be a true negative resistor in some cases.

carrier wave: A fundamental wave that is modulated by another wave or other waves, and hence "carries" the other modulating waveform(s). By stripping off the carrier in a demodulator, the carried waveform(s) emerges.

Casimir **effect:** The attraction of two conducting parallel plates in space, placed very close together, due to their influence on the active vacuum and on the vacuum's interaction with the plates. A part of the vacuum spectrum between the plates is "cut off and blocked or shielded, so that the vacuum potential between the plates is less than the vacuum potential outside them. The resulting potential gradient (force) on each plate from outside pushes the plates together. The Casimir effect unequivocally proves that it is possible to directly engineer the vacuum so as to produce forces and energy and physical effects in material systems. While it is a "small" white crow, it only takes one white crow to prove that not all crows are black. It is sufficient to conclusively prove that EM energy can be extracted from the vacuum. From there to practical EM power systems, fed by vacuum energy, is simply a product of modifying the prevailing electrodynamics and electrical engineering to incorporate the supersystem components and interactions, and then to gradually discover and develop the necessary technology. In this book we attempt to point the way by using the supersystem and providing some of the major necessary principles and c**oncepts.**

Casimir, Hendrik Brugt Berhard (1909-2000): Noted scientist who in **1948** correctly predicted that two parallel conducting plates, placed very near each other in a vacuum, would experience an attractive force due to **their** influence on the electromagnetic vacuum.

caus**ality:** The relationship between "cause" and "effect", or the interaction of 4-space cause upon a previous 3-space observation (effect), to produce a new 3-space observation (change in the previous observable or **production** of an exact 3-space replica). The basic notion in a causal system is that the system's response to an input signal is dependent only u**pon** past input values, not upon future input values. In a multiply connected space, the causality principle is dramatically changed and requires very careful accounting.

chaos: In dissipative dynamical systems, the dynamical evolution or ordering of the system that is aperiodic and highly dependent upon initial co**nditions** of the system. The trajectories of the system move on a *strange attractor,* which is a fractal subspace of the phase space. The mathematical equations describing chaotic behavior are very nonlinear and so complex that at the present time they cannot be computed or predicted. Control of chaotic oscillations thus is ongoing in its scientific development, with the most advanced control work having been done by Russian scientists, but now openly available.

charge: C*harge* is the ongoing circulation of EM energy flow between the time and 3-**space** domains, as "seen" by the observer in 3-space. *Negative* charge is (i) the ongoing absorption, by a mass particle, of EM energy input **from** the time domain (from *ict),* (ii) transduction of the absorbed energy into 3-space EM energy, and (iii) re-emission of the EM energy in all directions in 3-space. *Positive charge* is (i) the ongoing absorption, by a mass **particle, of** EM energy input from 3-space, (ii) transduction of the absorbed energy into the time domain, and (iii) re-emission of the EM energy in the time domain. Alternately, the positive charge can be considered as (i) the ongoing absorption, by a mass particle, of EM negative energy input from the time domain (from *ict),* (ii) transduction of the absorbed negative energy into 3-space, and (iii) re-emission of the EM negative energy in 3-space. A similar inversion can be utilized to represent negative charge. Time reversal of EM energy can be seen to result from the combined broken symmetries of EM energy flow (in 4-space) between time and 3-space, of two opposite charges and thus of dipolarity itself. See paragraph 3.1 of Chapter 3.

charge, electrical (q).: In the sense of q being charged mass, where the mass m_q of charge q is associated with the phenomenon defined as charge (see above definition). In our view, an electrical charge q associated with a charged particle or mass may be defined as $q \equiv m_q\phi_q$, to first order. This fits the common use of "the charge" as an expression for a charged mass, and explains why a gathering of charges produces a potential ϕ reaching out from the "source gathering of charge q". The actual "charge" action associated with the mass is due totally to ϕ_q, and to the broken symmetry of the ϕ_q dipolarity in the vacuum flux exchange with m_q. The charge q can be further broken down into a set of composite dipoles if the gathering of a screen of virtual charges of opposite sign in the vacuum, surrounding the mass of the "bare charge" inside the gathering, is considered. At first order, ϕ_q can be expressed (from *a particle view)* as a change in the local vacuum virtual photon flux (VPF), due to the VPF exchange between vacuum and m_q. *In a classical approach,* the vacuum may be considered as a giant scalar potential comprised of bidirectional EM wave flows according to Whittaker's 1903 decomposition, except we must re-interpret the phase conjugate half-set of Whittaker's decomposition as incoming from the time domain or complex plane. The charge then may be considered to diverge so much energy flow from the time domain into 3-space "around" the mass, resulting in ϕ_q as the steadily diverged energy surrounding the mass m_q. The ϕ_q component is actually the massless electrical charge and massless of itself. This approach would then yield a definition of massless charge itself, which at present is undefined in physics, as pointed out by Feynman and others.

In the expression $q \equiv m_q\phi_q$, the ϕ_q represents a change to the local vacuum potential ϕ_{vac}. Both ϕ_q and ϕ_{vac} decompose into Whittaker phase conjugate longitudinal EM wavepairs. As can be seen, by altering or engineering these constituent wavepairs, a particular ϕ_q or particular local vacuum potential ϕ_{vac} can be given a deterministic internal structure and dynamics.

Thus the interaction of a given electron or group of electrons may be made **quite** different from the interaction of another electron or group of "apparently identical" electrons. Such structuring dynamics are met with in the internal structuring of water, e.g., where the fierce activity of H-bond making and breaking can be treated as a special ϕ_H and its "internal structure and dynamics" can be engineered. A leading scientist pursuing that type of effect is Jacques Benveniste, in France. Predictably, the **orthodox** establishment has severely castigated these experiments, defending conventional models as absolute while completely ignoring the experimental results that refute those conventional models as being incomplete.

charge carriers: Anything — usually fundamental particles, ions, etc. — which consists of or contains charges and can be potentialized and move.

charge, magnetic: In the sense of charged mass, a magnetic charge is what is commonly called a "magnetic pole". The mass m of the magnetic charge is associated with the exchange of EM energy between the time domain and the 3-space domain. Essentially, the energy exchange action **known** as "charge" in the massless sense is the magnetic scalar potential, or monopole, when decomposed as per Whittaker and as reinterpreted by the **present** author, and as pointed out in quantum field theory by Mandl and Shaw. Note that our personal view discriminates between *massless charge and charged mass.* Massless charge is an ongoing action and process, involving energy transduction from one form (time energy or 3-space energy) to another (3-space energy or time energy). A charged mass is a **mass** associated with such action. Unfortunately, in the prevailing usage the word "charge" is just loosely used for both senses without discrimination. Hence one must carefully differentiate between the use of "charge" for "charged mass", and its use for an "ongoing process".

charge-parity-time (CPT) theorem: C is *charge conjugation,* which changes the quantum number of every particle into its antiparticle. P is the reflection operation known as *parity,* which turns an object into its mirror image and rotates it 180 degrees about an axis perpendicular to the mirror. T is *time reversal.* Any violation of time-reversal symmetry is always accompanied by a violation of CP symmetry (by violating either C or P symmetry) and vice versa.

Electrodynamics and the strong interaction preserve CP symmetry to great accuracy. Weak interactions do not conserve CP symmetry. T symmetry violation has been shown at CERN, the European particle physics laboratory in Geneva, Switzerland.

charge trapping (also "pinning"): Holding charges "pinned" or "trapped" at a site or in a location, by a barrier or force, so that they do not flow as $i = dq/dt$. A clustering of the pinned charges at the pinning site often occurs.

charge-barrier: A process or component or function — particularly in a semiconductor such as the Fogal semiconductor — that blocks the movement of charges q as current dq/dt.

classical electromagnetic theory: The classical theory of electrodynamics, in its modern form begun by Maxwell with his seminal-paper orally presented in 1864 and published in 1865. Today many variations have been made, particularly with respect to the basic group symmetry of the theoretical equations.

Clifford algebra: A special higher topology algebra founded by Clifford, which includes as subsets many other algebras of lower topology.

closed system: In the present approach, a system that in theory does not communicate with its environment, and does not exchange energy or matter between system and environment. Notice that this directly conflicts with the notion in thermodynamics (both classical and disequilibrium) that a closed system is closed only to mass exchange across its boundary, but open to energy exchange across it. General relativity has falsified that notion since 1915. Any system undergoing a change of energy also undergoes a change of mass, since mass is a special form of energy.

An ideal model in which dynamically and energetically the system is considered to be "isolated" as if nothing else existed. In short, the system is considered as if it were in some "magic box" and nothing outside the box can ever enter the box or affect the system inside the box, and nothing in the system can ever leave the box or affect anything outside the box. No electromagnetic system is closed. In short, we have corrected the thermodynamics "closed system" with open energy transfer, to a true closed system with neither energy nor mass transferring. That corresponds to the present thermodynamics notion of the "isolated system".

One performs a non sequitur when proclaiming treatment of any EM system as a "closed system", and of course modern physics substantiates this. The broken symmetry of the opposite charges on the ends of a dipole — and of an observable charge together with its clustered virtual charges of opposite sign — has been proven since 1957. The absence of any closed system is particularly true in electrical power systems. Here one is indeed allowed to input energy into the system to excite or potentialize it. That **of** course assumes that the system was "opened" long enough to take on excess energy (which also changed its mass)! Then as the system operates, losses and dissipations occur, in which case energy is considered to "escape" from the system permanently. In other words, the system was also continually "opened" so that energy could escape. In any EM system, every charge and dipole is already an open system, with free exchange of energy between 3-space and the time domain via that charge. In particle physics, every charge and dipole is a broken symmetry in its virtual particle flux exchange with the active vacuum.

The ancient thermodynamics definition of a "closed system" as one that exchanged heat but not mass, was formulated before thermodynamicists even knew that heat was a form of EM energy, or even knew what energy was. Hence we have changed that definition, since it is a non sequitur in view of modern physics.

There is really no such thing as a truly closed system in the universe, since *every* system is embedded in the active vacuum and is an open system in an energy exchange with the vacuum. Further, any change of energy in the system curves the local spacetime, which in turn interacts back upon the system. If that energetic exchange of the system with its "active vacuum and active curved spacetime environment" is symmetrical, then the system is in equilibrium with respect to its environmental exchange. In that case, for a limited number of purposes the system may be treated as if it were a closed system. In classical electrodynamics, Lorentz regauging of the Maxwell-Heaviside equations further reduced them to describe only Maxwellian systems in equilibrium with the active environment including the active vacuum. Consequently, any *net* interaction between the external vacuum (or the active local curved spacetime) and the system was excluded from the CED model. Obviously a *broken equilibrium* (i.e., broken symmetry) in that actual physical interaction is also not included in the CED model. Nor is it included in the classical (equilibrium) thermodynamics model with its archaic second law.

coefficient of performance (COP): Ratio of the usable energy output of a system divided by that portion of the total energy input that is input by the the operator's or experimenter. The COP is thus a measure of "return for using the operator's or experimenter's input energy" to produce — or direct the production of— useful work by the system". There are other ways to calculate the COP ratio; e.g., the average output power divided by the average operator input power. Contrast COP to the *efficiency* of a system, which is the ratio of the total usable output energy divided by the *total input energy from all sources*, including by the external active environment as well as by the operator. The efficiency of a system means how efficiently it produces useful output from its *total* input. The COP is the performance the system renders for the *operator's* input. Note that even a highly inefficient system — say, with efficiency $\xi = 0.25$ — may nonetheless have a COP>1.0, if the operator himself inputs less energy than the system outputs and the environment inputs all the rest of the output energy as well as all the energy in the system losses. As an example, a common home heat pump may have an efficiency of 40%, but a COP = 4.0. A windmill may have an efficiency of 20%, but a COP = ∞.

Also, for many systems one may evaluate their coefficient of performance for (1) the performance of the system as an energy transducer only, i.e., the ratio of energy the system outputs (whether used or not) divided by the energy the operator inputs; or (2) the performance of the system as a work producer, which is the useful work accomplished by the system divided by the operator's energy input. As an *energy transducer,* every generator and every battery has a COP»1.0, when both the Poynting diverged energy flow output and the Heaviside nondiverged energy flow output are accounted. Yet for useful work, the same system may have COP<1.0, because all the Heaviside nondiverged energy flow component is wasted, and some of the Poynting diverged energy flow component is wasted also.

Analyzed as an energy transducer only, every dipolar power unit — such as a generator or battery — outputs enormously more energy flow rate from its terminals than the mechanical energy rate input to the shaft of the generator, and than the chemical energy dissipated in the battery. So as an energy transducer, every dipolar power source is already a system exhibiting COP»1.0, prior to Lorentz' integration trick to discard the enormous Heaviside nondiverged energy flow output component.

cold fusion: An ad hoc term applied to the transmutations at low spatial energy (but high time-energy) achieved in electrolyte experiments, particularly when using specially prepared palladium electrodes after their loading with positive H+ or D+ ions, and particularly when deuterium is present in the electrolyte and the loading is by D+ ions. The *total* energy of the photon is greater the lower the photon's frequency and therefore the lower the spatial energy of the photon. Hence the cold fusion interactions actually use much higher *total* energy nuclear reactions than does present high (spatial) energy physics. Since "high energy" physics erroneously eliminates consideration of time-energy, the present high energy physics community is unjustifiably hostile to the very notion of cold fusion. The fact that little experiments on the lab bench use far higher total energy than the community's proud large accelerators is a bitter pill many high energy physicists cannot *tolerate*, let alone view with scientific objectivity. So again there is a classic confrontation between science from scientific results and "science as known in the prevailing dogma".

composite dipole: Dipole formed as one part of an isolated, observable "charged particle", consisting of a differential element of that observable charge on one end and a momentary clustering virtual charge of opposite sign. This is one way of dealing with the modern "infinitely strong dual dipoles" view of the charge, where the bare charge in the middle is infinite as is the clustering virtual particle charge surrounding that bare charge and partially shielding it. The difference between these two infinite charges is finite, and is routinely calculated to show that it is the standard value of the charge as seen by an external observer.

Cooper pair: In superconductivity theory, a dynamic, weakly bound pairing of electrons in a superconducting material below its critical temperature. In this pair, if the energy state with wave number sigma and spin 1/2 is occupied by an electron, then so is the state with wave number -sigma and spin -1/2. The Bardeen, Cooper, Schrieffer (BCS) theory uses this concept to provide a detailed microscopic theory of superconductivity.

COP (Coefficient of Performance): See coefficient of performance.

cosmological feedback principle: A self-regenerating cosmological feedback cycle ascribed by Puthoff as the source of the vacuum EM zero-point energy (ZPE). One may assume the existence of EM zero-point energy by fiat as part of the boundary conditions of the universe, or concieive of its generation by the quantum-fluctuation motion of charged particles that constitute matter. Puthoff calculated the latter possibility, assuming that the ZPE spectrum (field distribution) drives particle motion, and that the particle motion in turn generates the ZPE spectrum. This provides a self-regenerating cosmological feedback cycle, which in fact is consistent with the general relativity assumption that curvature of spacetime affects mass energy, and mass energy changes affect the curvature of spacetime. It would be interesting to see if Puthoff s concept can be made compatible with the broken symmetry of a dipolarity and Kron's open path concept.

curl concept of fields: The concept of a field as the curl of a vector potential (the vector potential may not be known). Any field so conceived has a divergence of zero *a priori*. One must be careful, however, because electrodynamics calculates only the field's reaction cross section with an *assu*med fixed unit point static charge at a point in space, leaving the field itself undefined. In short, the calculation of "the field" is actually a representation of the local point intensity of the field's interaction with unit point static charge. It is not the magnitude of the entire field (or potential) itself, *a priori,* but only of its reaction at a point with such a charge. In short, the field's reaction cross section with static charge is often misinterpreted as the magnitude of the field; at best it is an indication of the field intensity to static charge at a point.

curl-free magnetic vector potential: A field-free magnetic vector potential A without curl (swirl), hence lacing the B-field with which an A-potential is usually associated. For simplicity, one may visualize the curl-free A-potential as massless magnetic charge (and energy) in linear motion. Toroids and long solenoids have the characteristic of confining the B-field (curled potential) and holding it inside. However, drawing energy from any potential whose source charge or source dipole is not destroyed simply results in the full value (sum of the A-potential and B-field components) of the potential being replenished from the active vacuum, via the 4-symmetry energy flow mechanism in the presence of the broken 3-symmetry of the dipole. Hence one may easily increase the available potential energy density at will. This is a special form of gauge freedom allowing one to freely alter the potential energy of any electromagnetic system at will. Contrary to the dictates of EM textbooks, gauge freedom rigorously allows the direct and *free* amplification of potential energy of the EM system *at will;* the excess energy being freely furnished from the active vacuum processes. This principle incorporated in gauge field theory, has been known for decades and not previously used in the design, production, and use of COP>1.0 electrical power systems, prior to the invention of the motionless electromagnetic generator (MEG).

curved spacetime: A four-dimensional geometry (spacetime or ST) is used in general relativity, where the ST curvature is determined by the distribution of spatial energy such as mass-energy. In the general relativity view there is no such thing as an "empty" space "filled" with some energy substance. Instead, there is a combined space and time as a single entity, *spacetime,* which is itself energetic and is the very geometry we use in our mathematics. Any change in the local energy density of that spacetime constitutes a *curvature* of spacetime. In the new, more extended approach, any change in the spatial energy density or time-energy density of spacetime, or a combination of changes in both, constitutes a curvature of spacetime. In 4-space modeling, time is treated as a special, highly condensed form of EM energy. Specifically time may be regarded as ordinary 3-spatial EM energy compressed by a factor of c^2 and existing in the time domain rather than in 3-space. Time-energy therefore has essentially the same energy density as mass.

dark energy (negative): (Dirac holes, Dirac hole potentials, Dirac hole currents): A strange form of vacuum energy that creates repulsive gravity, causing the expansion of the universe to speed up (as observed and not otherwise explained by conventional theory). As a strong candidate for the dark energy, the present author has proposed a vacuum negative energy and negative energy currents in a negatively curved spacetime region partially containing excess Dirac sea holes devoid of electrons, along with normal filled Dirac sea holes. These special "Dirac negative action quanta" then form *negative action* (negative angular momentum, or — for simplicity — negative energy) potentials, currents, and fields. This type of vacuum condition containing some empty Dirac sea holes is referred to as a Dirac *-polarized vacuum.* However, it is not just simple static polarization, but possesses rich dynamics. Such dynamic Dirac sea hole regions occur in the vacuum in curved spacetime regions with tempic potential gradients and strong 3-spatial energy gradients, and thus time force existing across the region. The holes occur since the energetic fluctuations of the local vacuum energy can separate a Dirac hole from its occupying electron. These may then recombine, or in sufficient local spacetime curvature can remain apart. The more violent the region, the more severe are local spacetime curvatures in that region, and the greater the fraction of empty Dirac holes that is maintained on the average, in their continual forming and recombining. In those regions, a tempic potential is higher at the higher center magnitude of the spatial energy density. Hence there tempic forces from "inside to outside", providing a "tempic broom" effect to sweep outward the excess Dirac sea holes separated from their electrons in the curved spacetime. The end result is the presence of increasingly unoccupied Dirac sea holes toward the outside of violent regions of spacetime, until the effect is overcome by the decreasing spacetime curvature. These empty Dirac holes in regions and currents are a special form of vacuum negative energy, negative energy fields, and negative energy potentials, creating negative gravity (repulsive gravity) upon ordinary mass exposed to them. So the more distant violent astronomical regions have greater amounts of repulsive gravity being produced by their local Dirac-polarized vacua and negative energy fields and potentials. The result of this effect is that, at sufficient distance from us, the increasingly earlier, more violent events and resulting greater curved spacetime regions have created sufficient Dirac-polarized vacua to overcome the normal gravitational attraction between galaxies, etc. Even farther out, the increase in Dirac-polarized vacua overrides the normal attractive gravity between galaxies, etc. and produces net repulsive gravity. In turn this results in acceleration of the observed distant expansion of the universe. Dirac-polarized vacuum and its special dynamics can be accommodated in Sachs' unified field theory, which unites physics from quarks to the cosmos. See also **Dirac** sea **holes.**

On the laboratory bench, Dirac hole currents and effects are met with in a circuit or system producing COP>1.0, or very sharp 3-spatial gradients such as in violent electrical discharges. The effects become non-negligible at about COP = 2.0 and sometimes less. The tempic force back across the, system from output to input is sweeping the Dirac holes through the system. This can play havoc with attempts to close-loop the circuit for self-operation. In semiconductor components, strange effects are met due to their use of donor and acceptor materials, and temporary or permanent failure of semiconductors can result. The simple way to visualize this is that, in the system's supersystem, the spacetime component is curved and the vacuum component is highly nonlinear. All three components — overunity system, local nonlinear and polarized vacuum, and local curved spacetime — are interacting vigorously with each other. An operating overunity system represents an excited state of that local region of vacuum and curved spacetime, hidden in the euphemism "polarized" which is inadequate to describe the dynamics. With the dynamics, one may visualize that the Dirac sea hole currents are swept back through the system, seeking to recombine with electrons to eliminate the "polarization" and its dynamics, thus decaying the excited state back to equilibrium and destroying the overunity condition. In a robust system whose internal parts do not appreciably interact with these Dirac hold dynamics, the dynamics will reach the input section and extend back into or toward the primary power supply. In the input section, the feeder line, and in the power source itself, this results in "eating electrons". The Dirac sea effects thus act as a separate load imposed within and around the system input section or even back in the feeder line to the external power supply, in addition to the system load itself in the system output section. Simply put, the power supply must now furnish extra electrons to "power" the Dirac sea hole recombinations first, and additional electrons to power the system in its "normal" fashion. The result is to "draw additional power" from the source, where "drawing power" now includes the additional "eating of electrons" within the power supply itself, the feeder line, and the input section of the system.

Bedini and Bearden have filed a patent application on processes to convert the detrimental Dirac sea hole current reaching the system's input section.

transforming it into ordinary electron current. Magnetic Energy Ltd. has developed two other processes that handle the Dirac sea hole current problem, allowing close-looping for self-power ng operation. One of these processes converts negative energy into positive energy and the other avoids the negative energy problem. The first one allows this otherwise devastating special form of "negative feedback" or "degenerative feedback" to be converted into "positive feedback" or "regenerative feedback", allowing the system to be close-looped successfully for self-powering. Bearden has proposed yet a different approach using an "anti-circuit" method, but at this writing the method is still proprietary until intellectual property rights are secured.

dark energy (positive): Our term for the vast Heaviside nondiverged energy flow component surrounding every field/charge and potential/charge reaction but unaccounted in present electromagnetic theory after being arbitrarily discarded by Lorentz circa the 1890s. Present theory only accounts for that small amount of energy flow (the Poynting component) diverted locally around the interacting charge. All the rest of the energy flow comprising the field or potential, and missing the intercepting point charge, was discarded arbitrarily by Lorentz and is still discarded today, using Lorentz's integration trick. In honor of Heaviside, who did recognize the gravitational implications of his dark energy flow component, the present author has nominated this vast, unaccounted "dark energy" component as being responsible for the extra gravity holding the spiral arms of the spiral galaxies together. See **spiral** galaxy.

So one must recognize that the term "dark energy" is indefinite, since there are two kinds: (a) dark positive energy, and (b) dark negative energy. Dark positive energy consists of "normal" energy as often connected with particles, collected upon them and affecting their momenta, etc. It has dynamics and effects consistent with that kind of energy. Dark negative energy is a condition of the vacuum itself. It represents a sort of "empty container and empty container dynamics" (negative energy dynamics) in the vacuum itself, which "eats positive energy", so to speak. In philosophical terms, dark positive energy represents unaccounted actions and effects from unaccounted "extra presence of real substance and things and actions". Dark negative energy represents unaccounted actions and effects from "extra absence of real substance and things and actions beyond zero". Dark negative energy exhibits precisely appositive effects to the effects produced by dark positive energy. E g., dark energy creates the excess gra*vity* that is holding the arms of the spiral galaxies together. Negative energy resulting from spacetime curvature conditions in highly energetic astronomical phenomena produces the *antigravity* that is responsible for the acceleration of the expansion of the universe. We have nominated the negative energy Dirac sea hole phenomena as the cause of the antigravity that accelerates the expanding universe. Interestingly, these negative energy phenomena present in sharp discharge (strong gradient) phenomena appear to be responsible for the fact that such phenomena are known to violate present thermodynamics, and are being studied under the aegis of "extended thermodynamics". See discussion of areas violating thermodynamics in the last chapter of Kondepudi and Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, New York, 1998, reprinted with corrections in 1999. A reference is D. Jou, *Extended Irreversible Thermodynamics,* Springer-Verlag, New York, 1996.

degenerate semiconductor: A semiconductor whose conductivity approaches that of a metal.

demodulation: From a carrier containing a modulated signal, recovering the signal that was used to modulate it.

deterministic pattern or "template": The exact arrangement, by plan and deterministic action, of a group of entities. The deterministic actions and their "controlling forces" are simply called "dynamics", "structuring", etc. An example might be deterministic field structuring and actions of photons in a "herd" or "group". Or of a set of spacetime curvatures, into a dynamic structure called a *spacetime curvature engine* — or *engine* for short, so long as the meaning is clear. Or by hidden order in the virtual photons in the vacuum flux, or a deterministic set of changes to that vacuum flux (the vacuum flux that is called the "vacuum potential." We have called this *internal EM dynamics* function — of deterministically patterning or templating the virtual flux (particle view) of the potential, or of deterministically patterning or templating the longitudinal and time-polarized EM wave decomposition (wave view) of the potential - *dimensioning* the potential. See **dimensioning.** We refer to the deterministic structure and its structural dynamics as the "engine template".

dielectric: Literally, "against electric (current)", originally coined to mean "opposes the flow or conduction of electric fluid". In today's language, a dielectric is a non-conducting material used to oppose or prevent the flow of electric charges. Even empty vacuum exhibits some dielectric resistance. "Dielectric" is also used to refer to a nonconducting material (often between the metal plates of a capacitor or elsewhere) that can (i) sustain a change in potential across a distance (i.e., an E-field), and (ii) serve as an insulator. Space itself is a dielectric, or at least is said to exhibit certain dielectric qualities.

dimension: A certain primary geometrical physical attribute, such as length, used to describe the separational relationships of physical phenomena. Actually, "dimension" just means "fundamental quantity of separation" and "geometry" in the most general sense. Dimensions thus become a sort of set of fundamental variables describing the *extensivity* of system and the dynamics of that extensivity. The number of dimensions taken in advanced physics usually represents a somewhat arbitrary choice for good modeling fit. In modern Kaluza-Klein theory, for example, it becomes necessary to accept some 10 or 11 dimensions in the usual particle physics case. Sometimes more than 20, or even an infinite number, of dimensions may be used. General relativity (GR) has incorporated the interaction of the abstract geometry (spacetime) with mass, so that *geometry is a real, active, dynamic entity.* Until Sachs' extension of Einstein's work into a unified field theory and Evans' incorporation of O(3) electrodynamics as an important subset of Sachs' unified theory, GR has largely remained a nonexperimental discipline, at least in the laboratory. With the new Sachs-Evans approach and with recovery of (a) the Heaviside dark energy flow component (b) overunity EM systems arbitratily discarded by Lorentz symmetrical regauging, and (c) modeling of the complete supersystem including system, vacuum, and spacetime in mutual interaction, GR suddenly becomes substantially engineerable by novel electromagnetic means, both in the laboratory and in operating devices. We are thus entering a great new age of electrogravitation and a new physics. Research in antigravity lifters, COP>1.0 electrical power systems, medical therapy utilizing engines and anti-engines, and cold fusion are the elements of the new science that are struggling to be born and funded.

dimensi**oning:** Coined term meaning "forming an exact pattern or template of structures", particularly of nested curvatures of spacetime, where the pattern/template may also be dynamic, to include patterning and dynamics in the time-energy domain as well as the 3-space energy domain. It involves the internal structuring of the "dimensioned" entity. The term "dimensioning" makes sense in unified field theory, but is awkward in normal EM theory, which erroneously assumes that EM energy propagates in an uncurved spacetime. The mere presence of a change in local spatial energy, created by the wave energy itself when present in that local region, curves the local spacetime. Note that the background potential may be changed in a steady state level as well, which means that the frame of the system and its field dynamics has been rotated away from the laboratory frame. A time-polarized EM wave, e.g., changes the time-energy density and also curves local spacetime very powerfully. Hence the EM wave rigorously travels in curved spacetime, and in fact identically is a propagating, oscillating curvature of spacetime. What is missing from present physics is all accounting of the extraordinary Heaviside nondiverged energy flow component accompanying the accounted Poynting diverged energy flow component. Thus present physics *does not properly account for* the gravitational aspects of "electrodynamic field and potential energy in space".

Dimensioning also implies a deliberate form or structure (with dynamics) created in the virtual photon flux of the vacuum, in the virtual photon flux comprising a scalar potential, or in the infolded longitudinal-wave EM inside all EM potentials, fields, and waves. It implies an associated set of spacetime curvatures and their dynamics. The latter are referred to as "engines", "vacuum engines", "spacetime curvature engines", etc.

Dimensioning the potentials or signals or carrier waves of conventional electrodynamics involves adding hidden *vacuum engines* to them. The term *dimensioning* is descriptive and was chosen because adding such templates is equivalent to increasing the EM topology by adding mathematical dimensions. Since an EM wave or field or vector potential can be decomposed into two scalar potentials, then "dimensioning" one or both of the scalar potentials can "infold" the desired hidden structures or patterns (vacuum engines) inside an EM carrier wave, EM field, or vector potential created by the interference of the two dimensioned base potentials. This is referred to as *dimensioning* those EM field and wave entities. Internally structuring the two scalar potentials via arrays of longitudinal EM wave emitters and applying the necessary dynamics (modulations), is the primary manner to directly engineer dimensioning. When dimensioning is present, two EM waves that appear identical as seen on the oscilloscope may have dramatically differing internal structures (dimensioning). Hence they may produce drastically differing effects when they interact with the same or similar objects. *The only way in which the action of a vacuum engine can be resisted is to create and simultaneously introduce the exact antiengine for it — a special sort of forcible symmetrical regauging, so to speak.* Once the action has been completed in the receiving object and it has been physically changed, the only way it usually can be undone is to form a precise antiengine for that action, amplify the antiengine, and let the antiengine act upon the object for a sufficient time to "time-reverse it back to its previous physical state and condition". The concept has great utility in electromagnetic healing, not discussed in this book. We point out that the decades-long induction of diseases, health changes, etc. in U.S. personnel in the U.S. Embassy in Moscow occurred in forcefield-free regions. In short, the changes occurred in regions where the potentials were stable and unchanging, and so their engines were stable and constantly applied to the exposed personnel. The Soviets were simply radiating a high level target, the U.S. Ambassador and his staff, to guarantee high-level U.S. attention. By U.S. reaction to the deliberate electromagnetic "engine" induction of disease and health changes (three Ambassadors sickened and later died), the Soviets could positively ascertain whether we knew of the "infolding" or "dimensioning" technology, and whether we had developed defenses.

dipole: In the simplest case, two separated charges of usually equal magnitude and of opposite sign. More generally, a localized positive charge or charge distribution, and a localized negative charge or charge distribution, such that the net charge summation is zero, while the positive and negative charge distributions do not precisely superpose point to point but only on the average. We point out, however, that in regular physics electrical *charge* has no proper definition. As is well known in particle physics.due to the opposite charges on its end a dipole is a broken symmctry in the virtual photon flux of the vacuum. By definition of broken symmetry, this means that *some* of the virtual energy flux continuously absorbed from the seething vacuum by the charges of the dipole is not re-radiated as *virtual* energy. Instead, it is integrated *coherently* and re-radiated in 3-space as real emitted EM energy, establishing the fields and potentials associated with that source charge or source dipole. The dipole is therefore the most fundamental true "negative resistor" since it freely and continuously receives EM energy in unusable form, transduces it into usable form, and re-emits it in usable EM form.

It follows that any positive charge in the universe and any negative charge in the universe form a dipole and a broken symmetry in the vacuum flux. Hence every such dipolarity in the universe already extracts energy from the vacuum and puts it out in 3-space, or absorbs energy from 3-space and returns it to the vacuum. Kron called this an "open path" electrically, well before broken symmetry was formally discovered and experimentally prove in 1957. Kron was also able to use the open path in and around special electromagnetic circuits to cause the circuits to become true negative resistors, freely powering the network analyzer at Stanford Universoty even in the 1930s.

In classical EM theory, the problem of how every charge and dipole can continuously emit EM energy to form all the associated fields and potentials and their energy has been an unsolved problem — sometimes called the most difficult problem in electrodynamics, both classical and quantal. For the solution and a discussion, see the present author's paper "Giant Negentropy from the Common Dipole," *Journal of New Energy,* 5(1), Summer 2000, p. 11-23. Symmetry represents a conservation law, and broken symmetry represents violation of that conservation law. Because of the broken 3-symmetry of the dipole, the EM energy flow is not conserved in three dimensions — in this case, violation of conservation of 3-space EM energy flow is not conserved. Neither is the time EM energy flow conserved in the 4th dimension. Instead, EM energy flow is conserved in 4-dimensions, between the inflow in the time-domain and the outflow in 3-space in all directions, or vice versa. The hidden inflow of EM energy to the dipole (and to a charge) is actually from the complex plane (time domain). Thus all EM energy in 3-space is actually received by the source charge or source dipole from the time-domain, and returned to the time domain. What we call "propagation of EM energy through 3-space" is actually the spreading of this 4-circulation to successive 3-space points. The importance of the internal structuring of time-flow and of longitudinal EM waves in the time domain (dimensioning in the time-domain) is apparent, since it produces very powerful engines that can act upon matter in any fashion desired. Indeed, the Russians have developed clandestine robot engine systems comprised of sets of spacetime curvatures and their dynamics (which are just sets of longitudinal EM waves and their dynamics). Any function of a weapon system — sensing, weapons (using scalar interferometry), communications, propagation, etc. — can be incorporated. The capabilities of such systems, such as travel through continuously emitted EM signal beams from satellites, communications stations, power lines, etc., is eerie. Such a system can reside in any convenient EM potential or field as a "dimensioning" of the potential or field. Such robot engine systems have been clandestinely tested, but discussion of them is beyond the scope of this book. Suffice it to say that these systems are ushering in a dramatic revolution in future warfare, and the present research in physical nanobots is already decades behind.

Very powerful support for the author's solution of the source-charge problem is given by F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984, Revised edition 1993, under the heading "Covariant Quantization of the Photon Propagator" in Chapter 5. In their deeper coverage of the photon polarizations, Mandl and Shaw point out that the longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential.

Translated from particle terminology to wave terminology, Mandl's and Shaw's argument directly fits my re-interpretation of Whittaker's 1903 decomposition of the scalar potential.

Dirac, Paul A. M.: Noted English physicist, 1902-1984, who shared the Nobel Prize for developing wave mechanics and predicting the existence of the positron. He also originated the notion of "holes" or negative energy s**tates** in the vacuum, where the holes are ordinarily filled with electrons. The "vacuum replete with these Dirac structures" (all filled with electrons) is referred to as the *Dirac Sea.* If some of the holes are empty, then that is **the** Dirac sea of a curved spacetime. We strongly stress that a Dirac sea **hole** (negative energy electron with negative charge, negative energy EM fields and potentials, and with negative mass-energy) prior to observation (prior to its reaction with mass) is a quite different entity from the positron (after reaction with mass, so that the positron has positive mass-energy, **positive** charge, positive energy EM fields and potentials, and has been reversed in its 3-space propagation direction). Specifically, the Dirac sea **hole** produces negative energy fields and potentials, whose powerful Heaviside nondiverged components enable to perform practical antigravity. The (observed) positron produces gravity, not antigravity, and also positive energy fields and potentials.

Dirac sea: A concept and description (modeling) of the vacuum as containing — amongst other things — a "sea" of an infinite set of negative energy slates ("holes") for electrons, almost all of which are usually filled with electrons in the absence of any spacetime curvature of significance. In Dirac electron theory, the relativistic wave equation for the electron has four components, which correspond to two spin orientations and two energy conditions: positive and negative. Ordinary electrons are in the positive energy condition. So then the vacuum must also possess a "sea" of negative electron energy conditions. These negative conditions are called "holes" and are ordinarily filled or mostly filled with electrons which have fallen into these negative energy states. The Dirac sea is this vacuum sea of negative energy electrons (and a perhaps a few empty holes). In forward time, if energy is added to a negative energy electron, electrons can be **lifted** from this sea, since the added energy creates a local curvature of spacetime. If spacetime is locally curved, additional holes may be created so that, when the holes meet normal positive energy electrons, the electrons fall in and disappear, creating the appearance of an "electrical energy sink." This can happen so that the local ST curvature merely relaxes back to an uncurved spacetime, therefore avoiding radiation from "pair annihilation". If the curvature is in the opposite direction (i.e., if positive energy is added to the vacuum), electrons may be lifted from the Dirac sea by the resulting spacetime curvature energy, thus producing a flow of electrons "right out of the nonlinear vacuum" and generated by the spacetime curvature, and creating the appearance of an "electrical energy source" from the vacuum itself. In that case the energetic holes left behind also remain, and are affected by all EM fields etc. These holes thus form negative energy currents, fields, potentials, waves, and other electromagnetic entities such as negative energy charging and discharging (excitation and de-excitation).

In an overunity EM system, there is a higher energy density in the output section than in the input section. This includes time-energy density. So a higher tempic potential exists at the output than at the input. A gradient in the tempic potential thus exists, which constitutes a tempic force from the output section back through the system to its input section, and even on back through the conductors to the external power source, and even back into the distant external power source itself. Thus the power source has to furnish additional electron current. It must "power the holes", so to speak, by furnishing the electrons that are "eaten" by the Dirac hole current to eliminate it, which represents a novel kind of load in the input section of the system. So the power supply must furnish more power, so as to furnish the normal power to power the system, and the extra power to "power the elimination of the Dirac sea hole current as a back-reflected load". This is nature's way of decaying the excited state represented by COP>1.0, and restoring equilibrium and COP<1.0. By the Bedini process, the otherwise detrimental Dirac sea hole current reaching the input section is transformed into ordinary electron current, thus serving as input power to the system. By adjusting the system and optimizing it, this strange kind of "positive power input from converted Dirac hole current negative power feedback" can be adjusted to equal the input power required by the system. At that point the external power system can be disconnected, and the unit continues in close-looped and self-powering operation, including powering its load as well. Gabriel Kron succinctly described this condition and adjusting for it. Such a system allows energy to be continuously extracted from the vacuum, and a load to be powered continuously, with all the energy to run the extracting system being freely received from the active vacuum.

Dirac sea hole: A negative energy state for an electron, in the Dirac theory (in the Dirac sea). Literally the hole (negative energy state) when not occupied by the electron. In flat spacetime, almost all Dirac sea holes are filled with electrons. A small fraction on the average are not filled due to the local fluctuations of the vacuum energy continually lifting electrons from their holes and then having them recombine, usually so quickly that no radiation occurs in the observable state. The situation in the virtual state can be described by Feynman diagrams. This received view can be falsified in a curved spacetime. There the holes can exist and flow as currents, as validly as positive energy electrons can flow as currents, without the incessant "filling" activity of the vacuum itself. In addition, in such a hole-stabilizing curved spacetime region, the hole can "eat" or recombine with a real electron, so that the electron vanishes but without radiation. If we wish to join the two views, then we take Feynman's and Wheeler's view that the field does not exist in vacuum as such, but only the potential for it exists — in short, only the engine exists. The engine can then be said to be an ordering (chaotic self-ordering) condition present on the statistical random variation of the fluctuations of the vacuum. With longitudinal EM wave technology, this "order in the midst of disorder" — this templating and engine technology — can be directly engineered.

Dirac sea hole current (positron current): Just as a current of positive energy electrons can be generated, a current of negative energy Dirac sea holes can also be generated if the local spacetime is properly curved. Mathematically these are usually considered to be "positron currents", except they have no positive mass, because a hole will "eat" an electron. In an overunity electrical power system, the tempic potential is greater at the output of the system than is the tempic potential at the input section. There thus exists a "tempic force" or "tempic broom" sweeping Dirac sea hole current (positron current) out of the input section of the system into the external power supply. Simple instruments will record this current as if it were electron current being furnished from the external power supply to the system. Also, the Dirac sea positron current will "charge" the battery with Dirac sea holes (mathematical positrons), which in the conventional sense discharges the battery. A battery may become highly charged with Dirac sea holes. If so, then when placed on a normal battery charger, the electrons in the charging current will be "eaten" for an extended period by falling into the Dirac sea holes, and the battery will not acquire any charge until they are filled — whereupon the battery will sudden begin charging up normally again.

We also point out that the vacuum, containing the Dirac sea, actually is not "massless" at all, but has zero resultant mass on the average, as a result of the simultaneously presence of positive and negative mass in average equal abundance. Any curvature of spacetime must violate this "net sum zero" mass, hence produces inertia in all masses. Note that a negative mass can also exhibit a negative inertia; i.e., it accelerates and enhances the action of any force acting upon it, instead of opposing it. In short, negative mass exhibits negative inertia. We suggest that this effect is probably involved in the more violent explosive phenomena of astrophysics.

disequilibrium: Broken equilibrium, which in turn involves breaking appropriate conservation laws unless compensated.

displacement current: The apparent continuation of a conduction current through a region of space without a conductor, e.g., between the plates of a capacitor. Rather than being the motion of charges in a conductor, the displacement current is related to the rate of change of an E-field, and thus to the rate of change of the strain of the dielectric.

dissipative structure: In open systems far from thermodynamic equilibrium, a region of self-ordered behavior of a special kind in matter, characterized by symmetry-breaking, multiple choices, and correlations of macroscopic range. By correlations we mean statistically reproducible relations between distant parts of the system. Formation of a dissipative structure represents the birth of complexity in that region by self-organization. Usually a dissipative structure emerges only after the driving constraints reach a critical value. As the embryonic "engine" technology advances, however, the direct engineering and control of dissipative structures will become commonplace and routine.

distortion correction theorem: "If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1." Quoted from Amnon Yariv, *Optical Electronics*, 3rd Edn., Holt, Rinehart and Winston, New York, 1985, p. 500-501.

What is missing from the distortion correction theorem is hidden in the phrase *"if we generate in some region of space ... its phase conjugate replica."*. Since electrodynamicists and optical physicists have been unaware that all EM energy at any point in space comes from the time domain to that point, then —*prior to detection by interaction with charge* — the actual phase conjugate wave exists in the imaginary plane (in the time domain) and *not* in 3-space. The continuous interaction of that imaginary plane wave with charges, to actually produce the force-field wave in space, is *assumed* in the standard nonlinear optics. In short, one is looking at the "effect" wave, not the "cause" wave. However, since time is not an observable even in theory, it is this "effect" wave in 3-space that will indeed be measured by our instruments, because they must interact the imaginary or phase conjugate field with charge in order to detect or observe it (measure it).

The proof of this is experimentally already known but not recognized. If one reflects an EM wave off a normal mirror, the mirror must absorb the incoming wave, which transfers the wave momentum to the mirror to repel it, then the mirror must re-emit it, which adds that much more momentum to the recoil. So the mirror recoils with twice the momentum of the reflected wave. That produces a wave going in the other direction. If the incoming wave is continuous, one then has "paired" waves going in **opposite** directions). However, if an incoming wave strikes a phase *conjugate* mirror (PCM) material and is phase conjugated, a wave going in the opposite direction also emerges, but it synchronizes point by point with every spatial point occupied by the incoming wave. Also, *the phase conjugate mirror does not recoil at all.* This is well known in phase conjugate optics. It clearly shows that the origin of the "detected" phase conjugate wave is quite different from that of a normal reflected wave. Indeed, by our reinterpretation of Whittaker's 1903 biwave decomposition of the scalar potential, at every point along the detected phase conjugate wave, the energy enters that point from the time domain, and not from some sort of reflection from the PCM. The energy also returns to the time domain from that point. Its presence at the point produces a separation of virtual point charges, or the well-known polarization of the vacuum. So the flow of energy is from the time domain to the negative charge of the point dipole, thence to the positive charge of the dipole or others immediately surrounding it, thence back to the time domain. That is the operation in a "time-forward" sense to the external observer. In a "time-reversed" situation, the 4-circulation appears in the opposite direction. We propose that this fundamental 4-circulation is responsible for and constitutes spin itself. The direction of flow determines the 'polarity of the charge" we obseve. In this view, the point dipolarity in the virtual state with its ongoing 4-circulation can be taken as the very definition of energy itself.

Druge, Paul: Important German physicist, 1863-1906, who formulated the theory of the free electron gas inside conductors, using it to formulate a theory of metallic resistance.

Drude electron gas: The electrons that have broken loose from their orbital bonds in atoms in a conductor, and are thus free to "wander around" and "migrate" — or flow as current when exposed to an emf. Also called just the "electron gas". Note that, in the metal lattice of the conductors, each position vacated by a Drude electron is a "hole" or positively charged little region. Not only do electron currents exist as net negative charge migration in conductors and materials, but also hole currents exist as net positive charge migration from atom to atom, or a sort of "ionization migration flow". Electrons are continually recombining with these "ion holes", however, to reduce the hole current. These considerations are particularly significant in semiconductor materials.

duality principle: In quantum mechanics, the principle that photons and other particles propagate like waves and interact like particles, so that their description as only waves or only particles is inadequate.

Edison, Thomas Alva: Noted American inventor, 1847-1900, whose staggering portfolio of inventions included the storage battery, electric light bulb, phonograph, motion pictures, carbon microphone, and many others. Edison actually ran multiple "patenting research" laboratories as discovery and patent generating facilities. Not only was he an inventor himself, but he also employed a very good staff of researchers for invention and discovery.

E-field (electric field): The electric field intensity E; that field on an electric charge (charged mass) which produces an electrical force resulting in a linear motion of the charge. In reality, this electric field *as usually defined* does not and cannot exist in empty space devoid of charge, but only in and on charged matter, since mass is a component of force by $\mathbf{F} \equiv \partial/\partial t(m\mathbf{v})$. More specifically, the E-field is the local intensity of the field at a point, in its interaction with a unit point static charge fixed at the point. Even more specifically, it is the output of an observation/detection operation, which is a *d/dt* operator imposed on the 4-space causal field interaction with the 3-space charge, resulting in a frozen 3-space "slice" or snapshot of the deviation of energy *from* the field, forced to flow around that intercepting charge, at that single instant. In short, the E-field as a force field is an *effect* and not a *cause.*

The greatest error in all electrodynamics is to then assign this same *frozen effect force field intensity snapshot* — or an implicitly assumed iterative sequence of such snapshots — as the dynamic and nonfrozen *causative field itself* in 4-space, prior to the field's interaction with charged matter. This is the substitution of effect for cause, a grave non sequitur.

When EM theory was founded, the founders believed there was no single **point** in the entire universe where mass was absent; instead, a thin material "luminiferous ether" was considered to fill all space. Time was considered immutable, and there was no concept of a dynamic *spacetime.* Hence at each point in such a material ether, the reaction with charged matter (et**heric** matter) was thought to universally exist. Hence the field was indeed a very material entity, in their concept!

With the destruction of the material ether notion in the 1880s, not a single Maxwell-Heaviside equation was changed to eliminate its implicit assumption of the material ether. Hence this enormous error of confusing the *effect* field with the *causal* field (prior to observation-detection-interaction) is ubiquitous in electrodynamics and therefore in physics itself. The two fields (causal and effect) do not even have the same dimensionality; one is dynamic in 4-space while one is frozen in a 3-space snapshot. This has led to the false notion that an EM force field wave "movesthrough" 3-space, which does not happen at all. What really happens is that, if we assume a continual iterative *d/dt* observation operator, we have a continual series of such frozen 3-space snapshots, like the frames of a motion picture film. But none of those snapshots have any inherent motion at all. For a statement of the falsity of the conventional illustration of the "EM wave in space as an orthogonal pair of oscillating E and H fields in a transverse plane", see Robert H. Romer, "Heat is not a noun**."** *American Journal of Physics,* 69(2), Feb. 2001, p. 109. In endnote 24. **p** 109, AJP editor Romer takes to task *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate. " "...perhaps then, for historical interest,* ***[**we should] find out how that diagram came to contaminate our literature in the first place.* "Also see charge.

E-field, **motional:** In classical EM theory, the electric field produced by moving a conductor in a magnetic field. Often called *motional inductance* of the field. In the conventional theory, motional inductance is deemed to be ident**ical** to the inductance caused by a time rate of change of the magnetic field, experienced by the moving conductor. Or vice versa, experienced by a stationary conductor in a moving magnetic field.

E-field, static: In conventional theory, the electric field in the space surrounding a static charged particle, induced by the presence and action of the charged particle. There is no notion as to the *causative mechanism for this static* field of and from a charge, which implicitly is assumed to just continuously create and pour out EM energy in all directions, thus grossly violating energy conservation. Consideration of this anomaly leads to what has been called the most pressing problem in both quantal and classical electrodynamics: the question of the formation of the fields and potentials associated with the "source charge". Either the charge totally destroys the conservation of energy law itself, or there must be a *virtual* (nonobservable) energy input to the charge, since instruments establish that no observable EM energy is input to it.

For a solution to this problem, see the present author's "Giant Negentropy from the Common Dipole," *Journal ofNew Energy,* 5(1), Summer 2000, p. 11-23. We later found our solution to be consistent with quantum field theory and with a slight reinterpretation of Whittaker' s 1903 decomposition of the scalar potential.

Ehrenhaft, Felix: Scientist who performed numerous experiments believed to demonstrate fractional charges. We point out that classical electrodynamics assumes a local flat spacetime, a notion long falsified by general relativity, since 1915. In a local curvature of spacetime, the dynamics of the charge (modern view as the finite difference between two local infinite charges forming a dipolarity) can change the observed value of the net finite charge. So finite fractional charges, or even supercharges, are quite possible.

Einstein, Albert: German (Jewish) physicist, 1879-1955, who emigrated to the U.S. and formed special and general relativity theory, discovered and formed the theory of Brownian motion, recommended the development of the atomic bomb to President Roosevelt, and profoundly affected the course of science and the modern world,

Einstein, Podolsky, and Rosen (EPR) paradox: Thought experiment in 1935 for measuring two correlated photons separated and at distant locations, protesting that quantum mechanics requires that the particles communicate instantly, which is a "spooky action-at-a distance" (Einstein's phrase). The experiment eventually was performed and it works, just as required by quantum mechanics. The paradox inspired the development of hidden-variable theories.

Einstein's general theory of relativity: Mathematical theory of gravitation in which the gravitational force is mathematically described by a curvature in spacetime. In other words, the geometry and its dynamics have become physical agents, doing away with the old notion of the agent being something separate that is occupying inert spacetime. It is important to note that, in general relativity, mass-energy acts on spacetime to curve it, and the curvature of spacetime acts back on mass-energy to translate or change it. What seems missing from GR — at least in its practice — is the reaction of time-energy upon the geometry to curve it, and that corresponding reaction of the spacetime curvature back upon time and time-energy to change it.

Further, we consider that an exact pattern (template) of curvatures of spacetime and their dynamics exists and is back-acting upon every mass system and its dynamics. This is the *engine* concept. The most general way to engineer matter is not by brute force translation forces, but by forming spacetime engines. Once formed, these engines act upon the interior of any mass placed in the engine region, to eliminate the deltas existing in that muss's resident engine and the impressed engine. *The forces arise from the local spacetime itself, as curvatures in that local spacetime, so the action is from "inside" to "outside" an object, starting and continuing from every point in the object.* The curvatures of spacetime itself furnish the energy for the continuing actions upon the interior of the mass, once we have paid to form the engine. The mass can be acted upon at any level, from its gluons and quarks to its nuclei, to its atoms, its lattice forces and binding energy, etc. Significantly, the Sachs unified field theory applies to physical reality from beneath the quarks and gluons to the entire universe itself, and folds in quantum mechanics, electrodynamics, and general relativity into a single unified field theory that is (in theory) engineerable by higher symmetry electrodynamics means such as Evans' O(3) electrodynamics.

Einstein's postulates: (1) All the laws of physics are equally valid in all inertial frames of reference, (2) the speed of light is the same to every inertial observer, and (3) the observable local effects of a gravitational field are indistinguishable from those arising from acceleration of the frame of reference.

The first is called the special relativity principle, the second is called the law of light propagation, and the third is called the equivalence principle. It is now known that postulates (1) and (2) are not independent of each other. Note that the third postulate implies that any local force is due to a "gravitational field" (acceleration of the frame of reference). It also implies that gravitational force has a mechanism, since quantum mechanics assigns a differential operator, operating upon a potential, to create every force. By extended inference, it can be seen that the third postulate also implies that there should exist a method to turn any type of force (field) into gravitational force (field).

Note that postulate 1 may be violated if two frames remain inertial but each local spacetime contains different hidden engines. So the laws of interaction physics may indeed vary for the same object placed in first one of the frames and then in the second. Engines are ongoing general relativity actions, even in inertial frames. The action may not be translation of the bulk object, but extensive changes to the interior of the object and its interior dynamics at any and all levels.

The Russian physicist Sakharov has strongly postulated that gravitation is not even a fundamental field of physics, of the kind used by Maxwell in his electrodynamics. Instead, in this view G-field is always comprised of, and made from, other fields and interactions upon them. This would seem to follow straightaway since change of any kind of energy in spacetime curves that spacetime, therefore producing a gravitational change. So such has always been implicit in general relativity; it just took Sakharov to formally propose it. See A. D. Sakharov, "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," *Soviet Physics Doklady,* 12(11), 1968, p. 1040-1041 [English translation.] Sakharov's article (in Russian) in *Dokl. Akad. Nauk SSSR,* Vol. 177, 1967, p. 70-71.

Einstein's postulates (second postulate extended): Let us do a little "tinkering" with Einstein's postulates. For the first one, we point out that in the new approach using dimensioning and subspaces, the notion of an inertial frame now has acquired additional complexity! The "gross translation" effects — which tends to be what physics ultimately focuses upon today — can remain the same, and yet vacuum engines can be buried up inside the inertial frame to act upon a mass in non-translating ways. Everything translation-wise can remain the same, but now matter can be transmuted and transformed, etc., still right there in an inertial frame. So when the infolded electrodynamics inside all fields, potentials, and waves are utilized, the notion of "inertial frame" has acquired a higher topology, so that one inertial frame may differ very substantially from another inertial frame, even if the two are not in translation motion with respect to each other. This of course moves one from special relativity into full general relativity, and even there into a slight extension. Special relativity can remain special relativity on the surface, and yet general relativity effects occur "within" the objects postulated to exist in this inertial frame. The "inertial frame" concept in the postulates is now just a sort of special "equilibrium case" of a higher topology "inertial frame" notion. Further, general relativity is extended from the notion of "just translation" to the notion of "internal engines and internal dynamics also". These engines can be created in mass, in fields and potentials and waves, and in the vacuum itself.

The second postulate now stands as a "special case" of a far more general second postulate. The conventional second postulate actually defines (and limits) "light" to bulk changes in the entire vacuum potential magnitude. It only applies to light signals which are "bulk upheaval" disturbances of the vacuum potential (spacetime), so to speak. The limitation to light speed is an envelope wave speed restriction. Specifically it need not apply to the "inner EM/GR realm" of longitudinal EM waves infolded inside the vacuum potential, other potentials, signal carriers, etc. Superluminal communication is now included in this "hidden variable" or longitudinal EM wave communication. The Fogal semiconductor is capable of infolding and outfolding EM signals, including producing the internal longitudinal EM waves, and during their infolded state these signals are permitted to move at superluminal velocity. This has been experimentally demonstrated by Fogal.

So the second postulate needs to be restated to include the present statement as a special case. Notice that the concepts "light" and "speed of light" have now taken on much richer, extended meanings, since there is "outfolded" light and "infolded" light, so to speak. Light now is not just the "external perturbation wave" in the ambient vacuum potential medium (in flat spacetime). Instead, it can also be an "internal perturbation longitudinal wave" inside a "net flat spacetime" which does not show on the "surface" at all. Indeed, it may move in the time dimension or in 3-space. It now includes the transmission of an extraordinary "vacuum engine" as well as the usual transmission of just an ordinary signal. The ultimate expression of this extension has unfortunately been the development of such "engines" as infolded robot systems for use in warfare.

These infolded vacuum engines can be visualized as very special sets of longitudinal EM waves and their dynamics, consistent with the Whittaker decomposition. Whittaker's U(1) symmetry electrodynamics decomposition— particularly when enriched in the higher symmetry O(3) electrodynamics — applies to the ambient vacuum as a potential (it is a potential since it has energy density). So the "subspace" inside spacetime, so to speak, is just a set of LWs and their dynamics, and that subspace itself is engineerable, just as is the "bulk curvature of spacetime" itself. Further, we may now consider that space itself is just a vast superhighway for such superluminal LWs and their dynamics. Mass is mostly empty space with only an occasional particle here and there and with fields, potentials, and waves in between. Hence huge masses such as the ocean and earth are actually superhighways for the direct passage of longitudinal EM waves, "engines", special robots, etc. The degree of "transparency" is a function of the degree of cleanness of the LWs — where by "clean" we mean "absence of transverse EM wave residues".

The third postulate has also been dramatically extended and requires revision. As presently stated, antigravity effects do not appear to be included in present physics heretofore. The "gravitational effects and external accelerated effects" notion in conventional physics are still founded purely on *ordinary bulk translation.* Now they need not be. For example, the system can appear to be an ordinary system, as judged by normal "external light" observation. Yet its internal vacuum engines can cause local violation of most so-called "laws of nature." The reason a "law of nature" can be violated is that one is now able to violate the premises upon which it is postulated, and which are required to hold if the law is to hold. Or another way to look at it: any conservation law is the result of a symmetry. Breaking that symmetry allows violation of that specific conservation law. All external symmetries can be broken internally, even if the overall symmetry still applies externally!

The infolded EM inside the "conventional" EM fields, waves, and potentials — and inside the vacuum potential and spacetime itself-dramatically changes the present theories of electromagnetism, quantum mechanics, and general relativity. Further, for the first time the Sachs unified field theory allows an *engineering* approach that simultaneously crosses, includes, and extends all three disciplines in the required manner. The approach can be placed into good engineering models based on the Sachs work, particularly using the O(3) electrodynamics subset spearheaded by Evans, and the models can be tested and verified or falsified. See **frame, reference;** and **superluminal communication.**

electric field: (See also discussion under **E-field).** The *conventional* definition is this: "1. a region in space in which a stationary electric charge experiences a force due to its charge. 2. the electric force per unit test charge." Quoted from *Dictionary of Science and Technology,* p. 720. That definition is flawed. Consider the "spatial aspects" of part 1. If an electric field is defined as a force per unit charge (per Part 2), it is not a region in massless space but is an effect of something existing in massless spacetime and interacting with the charged 3-spatial mass. The electric field *prior to interaction and observation of the interaction* is indeed related to a region in "spacetime", because it is identically a curved spacetime region. In the Sachs approach, general relativity and electrodynamics have become one and the same. So prior to interaction, the *field as it independently exists prior to interaction with charge* has become a curved spacetime region, and the field-in-space is *purely* a spacetime curvature. Again, there is no such thing as a massless force. It is the hoary old mechanics notion of a separate force acting upon a separate mass that is a total non sequitur. The interaction of the massless field in space with a mass to produce a force, *of w**hich** that mass is actually a component,* is a separate consideration from the field as it exists in space prior to interaction. The result of this entity's i**nteraction** with charged mass has nothing to do with *what an electric field* is but with *what it does to charged mass.* Further, the force evidenced on a charged particle is not due to its charge; the particle has that charge always, whether or not there is a net force evidenced on it!

Part 2 is a little better. The only problem is, it totally disagrees with the notion of the field as an entity existing in spacetime prior to interaction **with** charged mass. In other words, here it's a "force on a charge", which is an effect (force consisting partially of the charge) occurring after the interaction. Again, the conventional definition states only what the field d*oes or accomplishes,* not what it *is.* Hopefully the reader can now begin to see what terrible difficulties await anyone who seriously attempts *foundations definitions.* Feynman sums it up admirably in one of his famous statements: "*Everything we know is only some kind of approximation, therefore, things must he learned only to be unlearned again or, more likely, corrected.* "An additional difficulty is that the very **notion** of a separate massless force — independent of the mass on which it operates - is fictitious, yet continues to be widely used. Recalling that any true definition is an identity and not an equation, let us start with the definition of force: $\mathbf{F} \equiv d/dt(m\mathbf{v})$. As can be seen, force is a system c**omprised** of (i) a mass component and (ii) a nonmass component. The force is not something separate from the mass, at all! Further, force is not e**ven** unitary. It's a system comprised of two coupled components; the mass and **a** coupled (interacting) gradient in a potential flux. *Note that v involves motion through, and interaction with, the virtual flux comprising the vacuum potential. Note that dv/dt involves a gradient in that flux interaction with the mass.* Unfortunately mechanics is still using the medieval concept of a "separate force acting upon a separate mass." That is wrong, unless we change the fundamental definition of force itself. See, e.g, **Robert** Bruce Lindsay and Henry Margenau, *Foundations of Physics, Dover P*ublications, New York, 1963, p. 283, where the authors emphasize that a "field of force" at any point is actually defined only for the case when a unit mass is present at that point. The reason can be seen: mass is a necessary component of force, and without mass present there cannot be a force present. For that reason, the force field in matter is a quite different beast from that "same" force field in empty spacetime. The first is a 3-spatial snapshot of an ongoing 4-spatial interaction, while the second is just a curved region of that "empty" spacetime without any ongoing reaction necessarily.

electric dipole: See **dipole.**

electrodynamicist: A specialist in electrodynamics — the physics of the relationships between electric, magnetic, and mechanical phenomena — and especially a specialist in electrodynamics *theory.* Essentially all electrodynamicists are versed in U(1) electrodynamics, and a much smaller number are versed in more advanced (higher group symmetry) electrodynamics such as O(3) symmetry electrodynamics, unified field theory electrodynamics, non-Abelian electrodynamics, etc. The electrodynamics field is actually a very vast field indeed, touching everything in physics, chemistry, etc. When one needs a particular electrodynamics problem solved, it often calls for a special category of electrodynamicist and not just an electrodynamicist in general. One must therefore sometimes examine the background of the electrodynamicist making a statement or a pronouncement; electrodynamicists after all do have differing views and differing disciplines within the field. And something may be perfectly true and acceptable in O(3) electrodynamics, yet not even exist in U(1) electrodynamics. As an example, the Aharonov-Bohm effect exists in O(3), but not in U(1). U(1) assumes a flat local spacetime (falsified by general relativity almost a century ago) and an inert vacuum (strongly falsified by modern particle physics and quantum mechanics), while O(3) is perfectly content to include and model curved spacetime and active vacuum also.

electromotive force (EMF): A difference of potential created by sources of electrical energy — i.e., by separated charges (dipoles) — which can be utilized to drive electrical currents through external circuits. As a difference of potential, that constitutes a gradient and hence a force. Also, open circuit voltage, or the limit of the potential difference between the source's terminals as the current in the external circuit approaches zero.

This term is poorly named and also quite confused in the standard theory and usage. First, em/is supposedly not a *force* at all. It is a difference in potential; hence it is measured in volts. It is very difficult to imagine a "difference in potential" at a single point unless one is dealing with a multiply connected space. So if the difference in potential is between two adjacent points, then that constitutes a force between the two points. Also, emf is not considered an E-field, since that is a field of force. Yet it is in fact a $\nabla\phi$, restricted to the direction between the two measurement points, and $-\nabla\phi$ is considered to be an E-field in the EM theory! As can be seen, o**ne** will not find a truly logical definition of potential or emf in the present texts and science literature.

For this reason the standard theorists say that the term emf is sometimes used as equivalent to a difference in potential, but that strictly speaking it s**hould** be applied only to a source of electrical energy. The problem is that **most** conventional theorists are unaware that a potential *identically is* a bidirectional flow of energy in 3-space, but a 3-space flow where *the* energy *enters from the time-domain (complex plane) and is caught in 3-space by interaction with charge, in a frozen snapshot (observation). The energy actually enters each 3-space point in the 4-space potential, from* the t*ime domain.* It moves across a dipole in space (present due to the quantum mechanical polarization of the vacuum), to the opposite charge, thus returning from the immediate region of that second 3-space point (or points) back to the time domain. Since conventional theorists are unaware that a potential is a hidden negentropic energy flow circulation process involving broken 3-symmetry, broken t-symmetry, and a special 4-symmetry between t and 3-space, the theorists have been unable to consider the potential in general as a direct source of energy flow. Yet they are forced to consider a potential difference across the terminals of a source as "somehow a flow of energy, that generates forces upon the electrons to propel them through the circuit as current flow."

In the new viewpoint, a resolution is achieved of these incongruities by more accurately defining the E-field with respect to the potential as $\mathbf{E} = -[\nabla\phi \bullet q]/|q|$. It is noted that $-[dA/dt]\bullet q/|q|$ also produces an E-field, so **that the ful**l definition of the E-field is $-[(\nabla\phi+d\mathbf{A}/dt)\bullet q]/|q|$. Note that dividing q by |q| does not remove the q unit from the corresponding dimensional equation, since $|q|$ is dimensionless. However, one volt represents one joule (of excess energy) *diverged from the potential,* per coulomb of interacting charge. *The energy density flow in an electric circuit is in the voltage, in the emf, and it is not in the current per se.* The electr**on** current (drift current) carries the excess energy that is being dissipated as work in the circuit, but in a nominal circuit the drift current only moves a few inches per hour. Half the energy collected in the potentialized circuit is dissipated in the external circuit and external losses, and half is dissipated in forcing spent electrons in the ground return line back up **through** the back-emf of the source, performing work inside the source to dissipate its separation of charges. Left alone, any dipole will furnish emf indefinitely, freely receiving the energy from the time domain (complex plane) and outputting it in 3-space via the reinterpreted Whittaker 1903 decomposition, in a circulation where at each point in it the energy enters from the time domain and thence returns back to the time domain. The free flow of EM energy will continue unabated from the undestroyed dipole, regardless of how many external loads and losses are powered (how much energy is collected and dissipated from the ongoing flows representing the potential). So *electrical circuits should be developed which allow few or no electrons to be forced back up through the source dipole to dissipate its separation of charges and destroy the dipole.*

electromotive force, back (back emf): Back emf is a loose term for the *counter-electromotive force.*

When the source is powering a circuit or circuit element, the term "back emf is used to refer to the forcing of current (electrons) back up through the source against its emf, thereby performing work inside the source to *dissipate* its separation of charges (its dipolarity). When a potential source of electrical energy — such as a capacitor, an inductor, or a rotating machine — is receiving energy from a source and converting it to *dissipative* power and work, that "powered" device is said to develop a *counter-emf.* A source has a built-in emf *a priori,* and this emf is holding its charges separated so that it is dipolar. When electron current is forced backwards through the source against its emf, work is performed against the internal resistance of the source to dissipate its separation of charges, thereby *dissipating* the source dipole. If electron current is not allowed to flow back up through the source against its back emf and through its built-in counter-emf, no dissipation of the source dipole occurs, even though energy density (not *electron current* dq/dt but *energy flow* dq/dt) is flowing from the source and can be collected in *collectors* in the external circuit to store the energy as a created local energy source. This collected energy can then be shuttled to a separate load circuit and released, to provide a "heat pump" cycling operation and overunity coefficient of performance.

Nikola Tesla in fact knew how to shuttle energy at will in a circuit in such fashion, as Barrett's quaternion analysis of Tesla's actual patented circuits has revealed. [See T.W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," *Annales de la Fondation Louis de Broglie,* 16(1), 1991, p. 23-41.] Note that a rigorous analysis of those same circuits, using tensors or vectors (lower topology algebras), will not show the effect. One

needs a higher-topology algebra such as quaternions or Clifford algebra, or **the** significant properties of such circuits will be entirely missed.

When the original source dipole is not dissipated because electron current is **not** passed back up through its back-emf to scatter the charges, the o**riginal** source dipole will continue to furnish emf indefinitely because of **the** proven asymmetry of its opposite charges in the seething vacuum energy flux.

The terms emf and counter-emf are normally applied only to recognized sources of electricity. Emf is computed as the algebraic sum of the **potential** differences acting in a circuit attached to a source. Back-emf (counter-emf) is computed as the potential difference internally through the source, between its terminals. The back-emf is thought to be necessary in producing *a powered device or component in the circuit.* Actually, *its **main** application (the ubiquitous closed current loop where all the external circuit current is passed back through the back-emf of the source dipole) is to prevent COP>1.0 electrical power systems from being developed and utilized, andfor every electrical circuit to self-enforce Lorentz symmetrical regauging during its de-excitation phase.*

elect**ron:** A stable elementary particle in all atoms, and having — in "f**orward** time" — a negative charge of 1.602x 10^{-19} coulombs, a spin 1/2, *and a* mass of 9.11x10^{-31} kilograms. If time is reversed, the charge (but not the masss) of the electron is reversed and it becomes a positron. Note this is an equival**ent** *observed* 3-positron, in a material lattice after interaction with charged matter, and not the mathematical equivalent represented by Dirac sea holes in the 4-vacuum prior to interaction with mass.

The "electron" may also exist as a negative energy, negative charge, negative mass-energy 4-state (Dirac hole) in the vacuum itself. In that state, the negative energy 4-electron is the source of negative energy fields and negative energy potentials. When the associated nondiverged and unaccounted Heaviside energy flow component is also accounted, the Dirac sea hole negative energy 4-electron can be manipulated in "anti-circuits" to directly engineer local antigravity, quite strongly and practically.

electron gas: A system of electrons with only weak interactions, so that the electrons may be regarded as moving independently. In many ways, such a collection of electrons can be treated as a gas. See Drude Electron gas.

electrostatic scalar potential: A *static* (stationary, in the sense that a perfect waterfall would appear stationary) ordering in the virtual photon flux of vacuum. The notion of "static" in the term "electrostatic" is misleading and should be replaced by "stationary" since the underlying active vacuum medium is always in violent motion, and there is nothing truly "static" in the universe. But an *equilibrium* or *stationary* condition can exist as an external envelope condition, even with continuous violent motion internally, so that a stationary entity or form persists. An example is a perfect whirlpool in a river.

The electrostatic scalar potential decomposes into a harmonic set of bidirectional phase conjugate longitudinal EM wavepairs, as shown by E. T. Whittaker in 1903. Hence the electrostatic scalar potential is an overall curvature of spacetime with internal longitudinal EM order and dynamics.

ELF: Extremely low frequency. Used to describe EM fields whose frequencies lie in the range from just above 0 Hz (direct current) to 300 Hz. This includes power line frequencies (60 Hz in U.S. and 50 Hz in Europe) and frequencies used by certain U.S. military ELF communication systems such as for submerged submarines.

EMF: Acronym for electromotive force. See discussion under electromotive force.

EM field: Electromagnetic field. In its loose usage, it refers in a very broad way to any field, force, or energy associated with electromagnetic interactions, charges and currents. Thus "EM field" includes electrostatic fields, magnetostatic fields, electromagnetic fields (including radiation and induction), vector-potential and scalar-potential fields, Hertz potentials, Fitzgerald potentials, Whittaker potentials, Debye potentials, etc.

In its precise usage, "EM field" conventionally refers to an EM "force" field, such as the E-field, D-field, B-field, and H-field. In classical electromagnetics (CEM), as a holdover from the old assumption of a material ether, the force fields are considered (erroneously) to exist in mass-free space, and they are also considered to be the primary causes of electromagnetic phenomena. This is a non sequitur, and it involves substituting the effect for the cause.

In the new viewpoint, *the force* fields do not exist in vacuum, but only in and on and of the charged particles themselves, where mass is present to form one component of the force. The force fields are therefore effects and not primary causes. The primary causes of all electromagnetic phenomena is taken to be the potentials; however, the hidden

Stoney/Whittaker/Ziolkowski internal structures of gradient-free scalar potentials may interfere to cause potential gradients in and on charged particle systems. [In O(3) and other higher group symmetry electrodynamics, this internal structuring and dynamics is far richer.] This therefore includes distance-free scalar potential interferometry and action at a distance, as well as vacuum engines.

The new view uses Sachs' unified field theory, and considers that the EM field in the vacuum, or an EM potential in the vacuum, involves and identically is curvature of spacetime. It follows that the internal EM structures of "ordinary" EM are also spacetime curvatures, where the specific pattern of these ST curvatures constitutes a very special kind of spacetime engine that acts upon any local mass in a specific manner determined by the individual pattern of curvatures. Thus the new view is a unified field theory where GR has become EM and EM has become GR, and where higher symmetry EM — such as O(3) group symmetry EM — can involve or produce gravitational, temporal, and inertial effects as well.

EM force field in the vacuum (a false concept : The erroneous notion that there exist forces in the vacuum, so that the set of forces at a set of spatial points forms a force field. Presently classical electrodynamicists accept the same E-field in vacuum as exists upon a charged mass when the vacuum E-field interacts with that charge, but it is asserted that the force is zero. This is easily falsified by defining force F as $\mathbf{F} \equiv \partial p/\partial t = \partial/\partial t(\mathrm{mv})$. By expanding the term on the right, one sees that both terms in the expansion include mass. Mass is thus a *component* of force. If mass is absent, there is no force present, thus no force field present. The field in space and the "same" field in mass differ precisely by a mass term in their fundamental units. The field in massless space may be taken as a curvature of spacetime. Zero field in space means a flat spacetime; a nonzero field in space represents a curvature of spacetime. A nonzero field in mass means a curvature of spacetime (a 4-space entity) interacting with mass (a previous 3-space observation) to produce a force consisting of that field and mass interaction. There is no separate force acting on a separate mass, and mechanics has been wrong in assuming on for hundreds of years. Instead, there is a force produced by the interaction of a massless 4-field (spacetime curvature) entity with 3-spatial mass. It is that *interacting and ongoing entity or process* that is a force field. Speaking of such a force field in space, is speaking of the force field that would exist in that space if every point had a unit point static charge and charged mass. As Feynman and Wheeler pointed out, no force field as such exists in space, but only the potential for a force field exists there, in case some charged mass is brought in to interact and have the forces developed upon its charges.

EM generatrix: Whatever generates electromagnetics, particularly the fields. Since all EM energy comes from the time domain and returns to it, time is the ultimate EM generatrix. The circulation of EM energy from the time domain to 3-space and back, or vice versa, is a function performed by the broken symmetry of every charge, considered as two infinite dipolarities with a finite difference that is the observed charge listed in texts and handbooks.

energetics: Energy as the primary thing in the universe. Rankine's term later raised to the forefront by Ostwald. Modern Russian term for advanced higher group symmetry electromagnetics, where the electrodynamics is a subset of an engineerable unified field theory. In the U.S., a model approaching the Russian model is given by Evans's O(3) electrodynamics as a special subset of the electrodynamics used by Sachs in his epochal unified field theory. Thus there is a growing, engineerable unified field theory (energetics theory) in the U.S., being provided by Evans, Sachs, and others. A solid basis for energetics can be realized from the giant 4-space negentropy EM energy circulation involved with every charge, so that all "EM energy in space" and time itself comes from that circulation.

energy: Usually erroneously "defined" as the "capacity to do work". This is a total misdefinition since it would define energy as the ability to change its own form. The closest to a legitimate definition of energy appears to be: *energy is the invariance of that which is varying* (Lindsay). However, note that this definition violates Aristotelian logic (but does not violate five-law logic developed in this book). Another try at it might be: *"Energy is any deterministic or coherent structuring, either dynamic or stationary, existing in the virtual particle flux of vacuum."* While that sounds good, it still involves the nonchanging aspect of the changing flux — much like considering a whirlpool in a river as being a "static" thing. So it is still "the invariance of that which is varying", which is an identity of opposites. We conclude that 4- or 5-law logic is necessary in order to define energy, since the defining logic must encompass the identity of opposites. Feynman simply stated: *"It is important to realize that in physics today, we have no knowledge of what energy is_."* [Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics*, Addison-Wesley, Reading, MA, Vol. 1, 1964, p. 4-2.]. In Aristotelian logic, there is no acceptable definition of energy. In 5-law logic, the second definition holds since opposites can indeed be identical. Now to more practical aspects:

For our purposes, there are three major types of energy in 4-space: (i) 3-spatial energy, which includes energy in one, two, or three of the spatial dimensions, (ii) mass-energy, which is 3-spatial energy compressed by the factor c^2 but still in 3-space, and (iii) time-energy, which is 3-spatial energy compressed by the factor c^2 but now placed in the time domain. **Note t**hat time-energy has the same density as mass-energy, and hence the same spacetime-curvature effect and gravitational capability.

The major implication of the new view is that the conservation of energy law is extended: Now it is the total of the 3-spatial energy, mass-energy, and time-energy that must be conserved. Transduction between time-energy and either of the other two forms, or both, will violate the conventional conservation of the sum of the 3-spatial energy and the mass-energy; every charge and every dipole is an example. Use of time-polarized and longitudinally polarized photons and EM waves particularly may involve the extended conservation law, e.g., as utilized in cold fusion and its lime reversal zones, and as involved in the outpouring of observable EM energy from every charge in the universe, with no observable energy input.

energy collection/collecting: In electromagnetics, placement of an interceptor/collector — e.g., a charged particle — into an impinging EM energy flow, so that some of the flow is diverted around the intercepting particle. We may say that the flow of energy has been "reordered" into a stream diverged around the intercepting charge. The diverged part of the total energy flow is what is calculated as the Poynting energy flow component, and is what is referred to as the "collected" energy or the "potential" of the charge.

However, neither charge nor potential are acceptably defined in conventional electrodynamics, but they can be defined in the new approach. We are now able to see something more fundamental. By the giant negentropy mechanism, the charge is involved in both the time-ene**rgy** stream and the vacuum flux set of virtual particle spatial-energy stream**s.** "Charge" is not just related to the mass of the particle, but is the entire ongoing set of the dynamics as well as the mass (the "nondynamics", so to speak). When a charge is placed in a field, that field is a change to the local dynamics, which is a change to the charge itself. Both the dynamics and the mass (the "nondynamics" constituting the charge are altered. So one does not have the "field acting on a separate charge", but instead we have the charge with its previous set of dynamics transformed into the altered charge with a new set of dynamics and a new mass. Similary when a charge is placed in a potential. Charge is also not a 3-

spatial situation and process, but a 4-spatial situation and process. If the internal structuring is also important, it becomes an n-spatial situation where n>4 and even where $n \gg 4$. To deal with sophisticated engines, $n \gg 4$ is necessary in the modeling utilized.

energy dissipation: Refers to the scattering or disordering of energy; at fundamental basis, it refers to the changing of an altered charge (see energy collection/collecting) back to an unaltered charge. In other words, it is a special decay process whereby an "excited charge" (so to speak) decays back to its "normal state" by releasing dynamics (energy) back to the vacuum. The energy itself is not destroyed, but its "collecting pattern" or "flow pattern" as coupled to the charge is disrupted and altered or destroyed There is also the connotation that *the dissipated energy is lost as far as any re-use of it by the system.* We point out that only collected/collecting energy can be scattered; it is the collecting *process* (the entire alteration of the normal charge dynamics) that is "destroyed" - or more exactly, *transformed* — not the energy itself.

energy, electromagnetic: From a vacuum and quantum mechanical viewpoint: a deterministic or coherent structuring, either dynamic or static, existing in the virtual-photon or charged-particle flux of vacuum. Since a photon may also be considered a positron-electron pair (if we multiply each by time so that they are *qt* rather than *q* in fundamental unit form), the flux of photons in the polarized vacuum necessarily also includes the flux of charged virtual electrons and positrons. From a spacetime viewpoint: a curvature of spacetime, or set of such curvatures. If a deterministic set of spacetime curvatures (a template) is involved, then an engine is involved and the energy is said to be "dimensioned" (internally structured).

energy flow: Generally considered as the passage of energy, in any form, in one direction or generally in one direction in a given 3-space. To first order, our definitions of energy treat energy flow in a given space as (1) a directional change (propagation) of the vacuum energy and its form, or (2) as a directional propagation of a curvature of spacetime or a set of curvatures of spacetime. *The terms "energy flow" and "flow of energy through space" are not precisely defined in physics! They are not going* **to** *be defined, at least not in ordinary Aristotelian logic terms, since energy itself is not precisely defined, and one of the best struggles with the concept of energy concludes that it involves invariance in the midst of change, which is a violation of the third law of Aristotelian logic and therefore of the other two as well.*

Indeed, the "EM wave flowing through 3-space" does not exist (see previous comments quoted from Romer). What does exist is an ongoing 4-space process, assumed to be continually and iteratively interacting with a unit point charge to "detect" or "observe" it as an iterative series of frozen 3-space snapshots, in memory-recalled serial order much like the frames of a motion picture film. Nonetheless, we must keep our sense of humor and proceed as if we knew exactly what we were speaking of when we use the co**ncepts** "energy", "energy flow", and "energy propagation through 3-space". If one dwells on the problem without humor, one is led to Hawking's grim but perceptive statement: *"All we ever know is our models,* b*ut never the reality that may or may not exist behind the models and casts its shadow upon us who are embedded inside it. We imagine and intuit, then point the finger and wait to see which suspect for truth turns and runs. Our models may get closer and closer, but we will never reach direct perception of reality's thing-in-itself"* Ultimately we do not know physical reality itself, but our mental modeling of it, as best we can "fit" our modeling through comparison with our sensory and perceptive sensations.

energy flux: The passage **of** energy **flow, in any form, through a unit area, and** in any given direction (usually all directions). This includes both passage in a single direction, or passage in multiple directions at once. See flux. Sometimes taken as that component of the net flux through a unit area at right angles to the area.

energy, free: In our usage, free energy is what is obtained from any environment or entity which freely emits, furnishes, or makes available energ**y** which either can be collected and used as furnished, or can be transduced into usable form and collected and used to power loads and do useful work. The word "free" does not apply to *costs,* but rather to the source *of the energy.* Since energy can neither be created nor destroyed, energy is always available free for the taking (collecting and using) if we bit learn how. *We never pay for energy per se, but for the manipulation and processing of the energy.* The notion of extracting and using free EM energy from the vacuum involves five key concepts: (1) the local vacuum/spacetime itself is filled with EM energy of generally unusable form (2) some sort of organizing principle is utilized in one part of the circuit or system to cohere the vacuum energy and transduce it into usable EM form (every charge already individually does that; we just have to learn to do it in a larger ensemble), (3) a gating mechanism is utilized to divert the vacuum energy now in usable EM form from the active vacuum to the external circuits or load parts of a device, (4) the coherent gated extra energy available in the external circuits or load parts is intercepted by those circuits or load parts, couples to them, and potentializes them with extra EM energy, so that the energy is collected by the system and regauges the system asymmetrically, and (5) the external circuits or load parts then scatter or dissipate the coupled extra energy to produce useful work (physical changes in the load), without destroying the organizing and gating mechanism that is extracting and furnishing usable EM energy from the vacuum.

energy source: A function or entity which furnishes energy by either gating it from one energy reservoir to our system or locality, or by transducing available unusable energy into usable form and then emitting the usable energy. Technically speaking, since energy cannot be created or destroyed, there is really no such thing as a *creative* "energy source", though the term "source" seems to imply such. So-called "energy sources" actually gate and/or transduce energy, changing it from one form to another, from one direction to another, etc. All EM energy in the universe comes from the time domain into 3-space vacuum, onto charge, and back to the time-domain, in a giant ongoing negentropy 4-circulation as freely provided by the "source charges".

engine: Deterministic pattern of structures, with their concomitant dynamics, in the flux of the vacuum. Also, deterministic patterning of curved spacetimc curvatures and their dynamics that comprise an overall larger spacetime curvature, or internally comprising an *overall* "uncurved" spacetime.

engineering, entropic: Engineering or designing systems and processes in equilibrium with their external active environment, and whose operation "spills" or wastes all the energy input to them, and which do not freely recover and use the dissipated or wasted energy. Entropic systems — which are the products of entropic engineering — thus have COP<1.0 and function in forward-time, and classical equilibrium thermodynamics with its infamous second law rigorously applies because the system is designed to self-enforce the equilibrium condition, often called by other names such as Lorentz symmetrical regauging.

engineering, negentropic: Engineering or designing open systems and processes far from equilibrium in their exchange with their active environment, whose operation freely "receives" excess energy from the environment, so that they permissibly output more energy than the operator inputs, the excess being received from the external environment. As an example, simply making a common dipole or assembling a charge initiates a giant negentropy operation that persists indefinitely (as long as the dipole or charge persists).

entropy: Thermodynamics state term representing a measure of an equilibrium system's capacity to undergo spontaneous change. In short, a quantity that measures the availability of a system's energy for conversion to work. The entropy of an open system not in equilibrium with its active environment cannot even be computed, because the state cannot be computed.

Entropy is given by dS = dQ/T, where dS is an infinitesimal change in entropy for a system absorbing an infinitesimal quantity of heat dQ at absolute temperature T. In statistical mechanics, entropy is a measure of the disorder in a system and is given by S = k lnP + C, where S is the value of entropy (disorder) for a system in a given state, P is the probability of that state's occurrence, k is a fixed constant, and C is an arbitrary constant. Intuitively, the concept of entropy corresponds to the concept of disorder. We specifically call attention to the fact that the "S" used in representing entropy is not at all the S used for Poynting energy flow in electrodynamics.

entropy, negative (negentropy): In a sense, the reversal of disorder, or reversal of entropy. Another way of defining negentropy even more rigorously — for electrodynamics — is to define it in spacetime as a bro*ken 3-symmetry in EM energy flow with concomitant relaxation to a more primary 4-symmetry between time-energy flow and 3-spatial EM energy flow, where time-energy flow is transduced into 3-spatial EM energyflow and then back to time-energy by the entity — such as a common dipole — causing the broken 3-symmetry.*

equilibrium: *In physics,* the state of a system in which opposing forces balance each other — i.e., sum to a vector zero. However, the *vector* zero summation does not represent "total absence" of everything, and the forces and their energy remain even though comprising a net vector zero system. Even a pure number zero may also be taken as a set of nonzero numbers; e.g., (+5 -3 -2). So even zero need not be total absence, but can be a "balanced set of presences". Or zero can be regarded as an overall equilibrium condition comprised of a set of disequilibria conditions. Indeed, a vector zero summation system of nonzero forces comprises a stress potential, and it contains stress-energy doing work upon the stressed system. In *mechanics,* a system of particles — where the net force acting upon each constituent particle is zero — is said to be in equilibrium. *In chemistry, t*he state of a solution where the forward and reverse reactions are equal and balance each other, so that the chemical reactants and products maintain the same amounts. Again, the notion of equilibrium involves the "stationary" notion (similar to a perfect whirlpool in a perfect river) of the overall invariance of some set of ongoing dynamic changes.

ether: An extremely fine material substance originally theorized to exist in space in lieu of sheer emptiness. The problem is considering space as emptiness in the first place, which it is not. Instead, space (more rigorously, spacetime) is a plenum and highly active (as in the general relativity view a half-century after Maxwell's seminal paper). So there is no need to "fill empty space" with something, because in the modern view there is no such thing as empty space.

Evans, Myron W.: Noted chemical physicist, primary developer of O(3) symmetry electrodynamics, and Director of the Alpha Foundation's Institute for Advanced Study (AIAS). Papers produced by the AIAS are critical to understanding extraction of EM energy from the active vacuum, engineering of vacuum engines, and development of a revolutionary new healing method by time-reversing diseased or damaged cells back to a previous healthy state.

excess collected energy: Collection by an interceptor/collector of a greater fraction of EM energy flow impinging upon it and also surrounding it, than is normal. See **Bohren experiment.**

exciton: A bound electron-hole pair in a semiconductor. Excitons are temporary formations and they decay by recombining, but they have a specific half-life before recombining during which they migrate through the semiconductor crystal. When the hole and electron recombine, then — if the curvature of spacetime relaxes to a flat spacetime, which is not overtly stated in classical EM theory — the recombination releases energy in the form of a *photon,* or sometimes in the form of several *phonons.* **Note** that the "hole" is an "absence of an electron" from a position in a lattice usually containing an electron, so that it exhibits a net positive "change in the charge field" at that point. Specifically, this "lattice electron charge hole" is not a Dirac Sea hole in the vacuum, which is an entirely different matter. On the other hand, if the hole and electron recombine and the curvature of spacetime does not relax to a flat spacetime *because of other curvature-sustaining involvements,* then no "pair annihilation" radiation is emitted. It is the relaxation of the curved spacetime (when hole and electron are separated) back to a flat spacetime that causes photon or phonon emission.

exothermic: Literally, "giving off heat", or emitting scattered EM energy. Producing time-forward EM energy from a "source" by reducing the ordering (internal energy) of the source.

Faraday, Michael: Noted English physicist and chemist, 1791-1867, who discovered field theory and electromagnetic induction. Faraday also invented the dynamo and formulated his laws of electrolysis. His work directly inspired Maxwell, who vowed to study no other electromagnetics theory until he had captured Faraday's work in a mathematical theory.

Derminion: A fundamental particle which has spin angular momentum of (n+1/2)h, and which obeys Fermi-Dirac statistics. Examples are baryons and leptons, which are subject to the Pauli exclusion principle.

Feynman, Richard P.: Noted physicist, Nobelist, and co-developer of quantum electrodynamics. Feynman is particularly noted for his refreshing three volumes of physics: see Richard P. Feynman, Robert B. Leighton and Matthew Sands, *The Feynman Lectures on Physics*, Addison-Wesley, New York, 1963. The book is particularly noteworthy in that it gives the student some insight into foundations problems in physics. E.g., in Vol. I on p. 2-4, Feynman defines the electric field not in terms of force per unit charge per se, but in terms of its *potentiality* for producing a force only when a charge is present for the force to be developed upon.

fiber fuse: Effect in optical fibers or fiber optic cable, whose core contains germanium, whereby heating the cable at one spot with a small flame will launch a focused traveling heat energy effect which melts a small hole in the core about every centimeter, and marches down the cable at about a meter per second to the end of the cable, destroying the optical transmission capability and ruining the cable. Oddly, by igniting the fiber fuse from the other end of the damaged cable, a *reverse* fiber fuse effect will often march back down the cable in similar fashion except that it fills in the melted holes in the core, restoring the transmission capability of the cable. If the energy required to melt all the holes (or fill them again) is taken into account, the amount of energy required to launch the fiber fuse is much less. This phenomenon does not appear to have an adequate technical explanation at present.

field: In mathematics, a set with two binary operations. In electromagnetics in more common mathematical terms, it is a region consisting of magnitudes and/or vector directions assigned to some active entity (e.g., spacetime itself) at each point in the region.

In physics, "a region of space marked by a physical property, as gravitational or electromagnetic force or fluid pressure, having a determinable value at every point in the region." [Thanks to *Webster's II New Riverside University Dictionary.]* Note the very special use of calculation of magnitude: one does not calculate the magnitude of the field itself, which fills an entire region. Instead, one calculates the magnitude of the *reaction* of that field with some interceptor/reactor at a given point — i.e., the *effect* of that field upon some mass entity. That is at best a measure of the *local intensity* of the field, not the magnitude of the entire field per se. For that reason, often we speak of calculating the field intensity, which is far more accurate than saying that we are "calculating the magnitude of the field". So we do not really calculate the "magnitude of the field," and in fact such a calculation does not occur in any textbook. Instead, we calculate the *reaction cross section* — in the case of electrodynamics, for an assumed unit point static charge — of the field at a point in the region or space occupied by the field. *We actually calculate what is deviated or diverged from the field, around an assumed interceptor, not the field itself.* This means that we calculate the result or *effect* of a detecting/observing action, and a frozen 3-space snapshot of that ongoing interaction, since all observation is 3-spatial. To then state that this same field exists in space prior to the interaction, and is what causes the interaction, is a gross non sequitur and confuses the effect with its own cause. This ubiquitous non sequitur widely permeates electrodynamics and mechanics, and it is one of the greatest scientific logical errors in history.

field, scalar: In the standard view, a field, to any point of which is assigned a magnitude only. However, "magnitude" of the EM field in the standard usage refers only to the amount of energy deviated (diverged) around a fixed intercepting unit point charge, per the conventional "detection by reaction" model. Again, note the serious logical errors in the conventional usage. One cannot define the field in terms of what is diverged from it! At best one has defined the "intensity" of the field's interaction with a given standard interactant (charge). So in electrodynamics the term "scalar field" would be better replaced with "scalar field intensity", at least for many applications.

field, vector: In the conventional view, a field, to any point of which is assigned both a magnitude and a direction. Again, the standard view also implements that assumed reaction with a detecting unit point static charge, and the "magnitude" used is just the difference between the amount of energy piled on the "front" of the interacting unit point static charge and the amount of energy pile on the "rear" of the interacting unit point charge.

The "direction" is from the center of the "high energy" pileup to the center of the "low energy " pileup around this fixed intercepting unit point charge. Again, note the serious logical errors. One cannot define the field in terms of what is diverged from it and the direction of the difference in "energy pressure." At best one has defined the "intensity" of the field's interaction with a given interactant (charge), and the "direction" established by the difference between the frontal and rear energy pileups on the interacting charge. The term would be better replaced with "vector field intensity".

flat (uncurved) spacetime: In our view, a vacuum/spacetime whose virtual particle flux intensity — and therefore its local energy density — does not change as a function of spatial position or averaged time. That is, in our view a flat spacetime is also identically a quantum mechanical vacuum whose virtual particle flux and energy density are constant in magnitude and internal constitution with respect to spatial position and in averaged time. Note that in this extended view, a flat spacetime (vacuum potential) can still contain vacuum engines, and thus still be *dimensioned. In that cas*e, contrary to normal physics, the flat spacetime itself will still interact with an embedded object and *change* it internally, even though it does not *translate* it externally. To perform transmutation of elements, e.g., one is interested in just such engines and not in spatially translating the element being treated. This dimensioning or activation of the local flat spacetime can result in an otherwise unexpected, seeming violation of one or more laws of nature in the dimensioned region, and easily results in violation of conventional physics which does not take into account engines and dimensioning of a flat spacetime. Note also that this allows a differentiation between the actions of two inertial frames, if the frames are differently dimensioned. As we stated, this is a deliberate extension of orthodox physics. See also dimensioning, and Einstein's second postulate —extension of.

fluctiation theorem: Theorem advanced by Denis J. Evans et al. at the Australian National University in 1993 that appreciable and measurable violations of the second law of thermodynamics occurs for small systems over short time scales. See D. J. Evans, E. g. D. Cohen, and G. P. Morriss, Phys. *Rev. Lett.* Vol. 71, 2401 (1993). In 2002 Evans and his colleagues experimentally demonstrated the proof of this theorem in rigorous fashion. See G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.* 89(5), 29 July 2002, 050601. The proof also directly proves the occurrence in electrolytes and fluids of time-reversal zones (TRZs) where the dynamics runs backwards for a short time. This validates Bearden's proposal of the TRZ and its consequent reversal of the Coulomb barrier to a Coulomb attractor as the mechanism producing attraction of like charges into quasi-nuclei, which then can decay by quark flipping (if necessary) as the TRZ decays and reverses back to normal. This provides a mechanism for transmutation of elements in these TRZs as a result of the fluctuation theorem and its proof by Evans et al.

flux: The amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time. The term flux density is now more commonly used. Loosely, "flux" refers to the perpendicular flow through an area, while "flux density" refers to the magnitude or intensity (rate) of the flux. In some cases such as a flux of particles of different velocities, the number of particles may be multiplied by the average velocity to give an average flux density that is weighted.

flux density: The amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time.

Fogal semiconductor: A special semiconductor invented by Bill (William) Fogal which, in its operational regime, utilizes an integrated semiconductor, very special tantalum capacitor, and feedback resistor to accomplish charge blocking (blocking of electron current flow dq/dt) while passing displacement current $d\phi/dt$.

The Fogal semiconductor can also accomplish amplified phase conjugation of signals as well as infolding (translation of input transverse EM waves to output longitudinal EM waves) and outfolding (translation of received input longitudinal EM waves to output transverse EM waves). Used in communication systems, it opens the use of the unlimited "infolded" electromagnetics bandwidth. Since it may communicate using longitudinal EM waves, it is also usable for superluminal or a specialized "tunneling" communication through the "interior" of normal EM waves, potentials, and fields. For example, it can easily transmit color TV signals through the "interior" of a common DC potential. See William J. Fogal, "High Gain, Low distortion, Faster Switching Transistor," U.S. Patent No. 5,196,809, Mar. 23, 1993; - - "High Gain, Low Distortion, Faster Switching Transistor," U.S. Patent No. 5,430,413, July 4, 1995, a continuation of his earlier patent.

Fogal, William: Researcher and inventor of the Fogal semiconductor (two patents so far), as well as a superluminal infolding-outfolding communication system and other remarkable longitudinal EM wave processes. The Fogal semiconductor allows the use of the internal longitudinal wave electrodynamics comprising all ordinary EM fields, waves, and potentials, and thus will usher in a revolutionary new communications science and extended electrodynamics technology. See also discussion under Einstein's second postulate extended.

force: In mechanics, identically the time derivative of momentum, or $F \equiv \partial/\partial t(mv)$. One must realize that it is not the mass in motion, but the single undivided quantity "mass-motion" or "mass-velocity" (momentum) that is changed. Present EM theory erroneously takes force fields to exist in empty, massless vacuum but with zero force! However, the mechanical definition is an identity, not an equality. The observable mass in vacuum is zero, so substituting m = 0 gives $F = \partial/\partial t(0 \cdot v) = 0$. This appears to be the conventional reasoning. However, $\mathbf{E} = \mathbf{F}/q = 0/q = \mathbf{0}$ also. So if mass is zero, the field E used in "force field definition" disappears. All that says is that the field in mass-free space and the field in mass are totally different entities,

Since force can be induced into and of the mass-motion target at a distance, in the new approach there is a more fundamental mechanism that produces force itself, and that is a curvature of spacetime, in agreement with general relativity. Distant curvatures of spacetime to produce forces on exposed charged mass systems may also be engineered by longitudinal EM wave interferometry, which does constitute "action at a distance." The fields arise in the local spacetime in the distant interference zone, thus arising from within and around any physical mass object located in that zone. Hence it produces an "inside-to-outside" type of action that is presently unknown in orthodox physics. The produced fields can start inside the nucleons of the atoms and move outward, and work directly upon the quarks and gluons first. This opens a great number of entirely new nuclear reactions presently not developed in particle physics.

force **field:** An 3-space *effect* that exists as a local vacuum virtual particle flux interaction upon and with a hypothetical unit point mass or unit point charge assumed at every point in 3-space. The effect (force field) results from a ***causal*** force-free 4-field (a curvature of spacetime) interacting with an observed hypothetical unit point mass or unit point charge assumed at every point in space. Hence the invocation of observation invokes a *d/dt operator* (the interacting charged mass) upon the causal field, providing a frozen 3-snapshot of the ongoing 4-interaction. An iterative series of such 3-observations is said to constitute a "3-space force field" as if it existed in time, which it does not. It is seen (iteratively observed) as a series of such frozen snapshots, which mentally we conceive as "existing in time" instead of the "recurring at points in time" that it really is. We unconsciously invoke that substitution of iterative effect as being continuity in space and time, when we use terms such as "propagation through space".

force-free propagation: Propagation of force-free disturbances in the vacuum flux; i.e., propagation of curvatures of spacetime or sets of such curvatures, together with any associated dynamics.

foundations of mathematics: The concepts, operations, assumptions, postulates, logic, definitions, and axioms of algebra systems.

foundations of physics: The postulates, assumptions, concepts, and definitions upon which a scientific physics model rests. Included are also the postulates, concepts, assumptions, definitions, axioms, logic, and operations of the mathematics in which the scientific model is expressed, as well as the postulates, concepts, assumptions, definitions, and operations of the physics entities and functions assigned to the mathematical symbols. Thus "scientific truth" -— even mathematical truth; see Morris Kline's *Mathematics: The Loss of Certainty* — is actually a *qualified* truth. What is true in one model (one set of postulates, assumptions, concepts, and definitions, including those in the algebra) may not be always true when one or more of the postulates, assumptions, concepts, and definitions is/are changed, and/or when the topology of the mathematics in which the model is expressed is changed, etc.

four-space (4-space): The "normal" spacetime used in physics, consisting of three spatial dimensions and one time dimension. Minkowski spacetime.

fourth law of logic: The law that a thing and its opposite become identical on the boundary, i.e., in multiple simultaneous observations. *Identity* is merely the loss of observational or perceptual distinction; when formerly "two" objects are simultaneously observed by a single observer, and two slides have been shown at once in the slide projector. It does not matter at that point what each was individually observed to be before when it was singly observed. Now there is no observable distinction or difference. We previously showed that the fourth law is implicitly assumed in the three laws of Aristotelian logic anyway, and published a simple proof of the fourth law. Without the fourth law, Aristotelian logic "eats itself." As a simple example, the Aristotelian third law eliminates the center vertical line division of a simple Venn diagram between A and not-A, because the division line itself is happy to belong to both A and not-A simultaneously. So Aristotelian logic destroys the very Venn diagrams used to "prove" or "demonstrate" its logical propositions. What results in the four law logic is that either (1) the three laws apply explicitly and the fourth law applies implicitly, or (2) the fourth law applies explicitly and the three laws apply implicitly. This *application rule* — i.e., *either (1) or (2)* — itself can be taken as a fifth law of logic, thus extending and completing Aristotelian logic which is presently incomplete. The fourth law is also the law of the **paradox:** i.e., when something is true but violates one or more of the Aristotelian laws, it is a case where the fourth law applies explicitly and the three laws apply implicitly. Wave-particle duality is such a fourth law case.

four-wave mixing: A nonlinear multiwave mixing effect when four waves mix (interact) in a highly nonlinear situation so that wave-to-wave interaction occurs.

fractional charge: A hypothetical particle having a charge less than the electron's charge. Often called a quark, where its charge is 1/3 or 2/3 the charge of the electron. A few years ago, Stanford University researchers found what was believed to be evidence of free quarks, but that later was doubted. There is also now the fractional Hall effect. In the 1920's Ehrenhaft reported the discovery and production of fractional charges of all sizes, and his experiments have been replicated by Mikhailov, with papers by Mikhailov and Barrett.

In the new approach, when time-energy is utilized, quarks can be freed or nearly freed rather easily in matter inside "time-reversal zones", since the energy emerges from the local curved spacetime and propagates from within the nucleus outward and the gluon forces are substantially reversed. Note also that time is multiply connected, and energy can enter an object anywhere within it or even everywhere within it, since each and every differential piece of the object exists in every single increment of time. It would appear that, as the new approach develops, free quarks will become a normal event rather than one intensely sought but terribly difficult to come by experimentally. We flatly predict that, in a sufficiently good time-reversal zone, free quarks can be produced because of the enormous energy density (in the time component) being utilized.

frame, Lorentz: A frame of reference that is not accelerated with respect to the laboratory observer. In this case, the frame may be rotated (moving at a constant velocity) but it is not rotating (the frame is not accelerated with respect to the observer, and spacetime is not curved). In this case special relativity applies, as do the conservation laws. We do add a caution that full general relativity still applies to the interior structure and dynamics (the engine) comprising the flat spacetime. Two *overall* flat spacetimes may produce very different interactions inside matter, if their internal engines are materially different and dimensioned.

frame, reference: A spatial, organized, measured lattice placed in "emptiness" (space, spacetime). Normally refers to a 3-dimensional, spatial frame (simplest case) or to a 4-dimensional spacetime. All objects and points in the "universe" or spatial frame are considered to simultaneously coexist at separate, measured points in the frame. Differs from the vacuum in that, rigorously, vacuum has no existing definite lengths and no existing definite time intervals, as these appear only after measurement or detection, and are relative to the observer and to the interactions ongoing as well as in the detection process itself. The "laboratory frame" is the static reference frame of the observer or measurement. A separate reference frame may be assumed to exist for any fixed or moving object, or centered on any point in another frame. When a type of frame is assumed, the entire class of *translation type* physical interactions that can occur has been restricted to an assumed set or type. In other words, given the frame, the *conventional* physics of translation has been assumed. One of the greatest restrictions of an assumed "frame" is to rule out the consideration (existence) of other higher dimensions, hence of dimensioned engines.

In the new unified field theory approach, the other higher dimensions are always available and cannot be ruled out in general, but only in some special case. Every curvature of spacetime, and any internal additional curvatures comprising that primary curvature, *adds one or more new dimensions.* In our view, a spacetime may be "flat" in overall envelope curvature, but consist of internally structured deterministic curvatures or "engines". In this view, normal inertial frames, e.g., may still contain engines, which will not affect the normal *bulk translation* rules, but may affect any or all of the nontranslation mechanisms, including the very laws of nature in many cases. Engines have not been developed and used in conventional open physics, but were weaponized clandestinely by the former Soviet Union. See **Einstein's second postulate extended.**

free energy: Excess energy freely received from its external active environment by an open system that is far from thermodynamic equilibrium in its exchange with that environment. There is a separate thermodynamics for such open disequilibrium systems. (See **thermodynamics of open systems.** See also Appendix **A** for a more detailed discussion of the thermodynamics aspects). Such a system is permitted to exhibit five "magic" functions: it is permitted to (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator himself inputs (the excess energy being freely received from the active environment), (4) power itself and its load simultaneously (all the energy being freely received from the active environment), and (5) exhibit negentropy. Every charge and dipole in the universe already performs all five of those functions.

free **energy** system: Any open system not in thermodynamic equilibrium, which freely receives excess energy from an external environmental source, **and** collects and utilizes this free excess energy to freely power external loads, so that a permissible system COP>1.0 is achieved, as allowed by the nonlinear thermodynamics of open systems in disequilibrium with their environment. (See **thermodynamics** of open systems.). A free energy system may also be "self-powering" if "close-looped" with governed positive energy feedback from output section to input section, meeting the Kron self-powering condition. Operation of free energy systems can only be properly modeled and understood in terms of their super*system* interactions.

Implicit in the notion of free electromagnetic energy systems are six key concepts: (1) the local vacuum/spacetime is known to be filled with hidden (i.e., virtual) energy in violent motion, i.e., the local vacuum/spacetime is a violent virtual energy flux and therefore a scalar potential, (2) a broken symmetry - i.e., some sort of organizing and gating mechanism, such as the common dipole — in this hidden (virtual) energy flux of the vacuum is utilized as a source, to cohere the virtual energy and divert it as energy $\mathbf{S} = f(\mathbf{E}\times\mathbf{H})$, flow from the broken symmetry initiation point onto the external circuits or load parts of a device, (3) the coherent gated extra virtual energy flow available in the external circuits or load parts couples to (interacts with) the mobile charges in those circuits or load parts, so that gradients developed in the interacting virtual flux form force fields on and of the particle masses, driving the mobile charges along the circuit and through current-impeding circuit components such as loads, (4) in the interaction of the driven charges with the driving virtual energy flux, their spin and erratic motions integrate a tiny "coupling fraction" of the driving virtual energy into observable (i.e., macroscopic field) energy, thus creating the **E**-fields and B-fields, (5) the current-impeding external components or load parts scatter or dissipate the coherence of the driven charges and thereby dissipate the collected energy or change the form of the collected field energy, thereby producing useful work, and (6) little or none of the excess collected energy in the circuit is split off and dissipated in the gate (asymmetry component serving as the S-flow source) to destroy the gate and thereby destroy its asymmetry in the vacuum flux, which is what is extracting and producing the free energy flow.

gate: A component or function that extracts some part of a flux or flow, compacting it into a stream, and "sending it out" in organized, useful energy flow form. The fundamental gate for all EM energy in 3-space is the dipole or dipolarity — the "isolated charge" is actually a dipolarity, when the virtual charges of opposite sign clustering around it in the active polarized vacuum are considered. The source of the 3-space EM energy continuously emitted by any charge or dipole is an equal inflow of EM energy from the complex plane (from the time domain). Thus EM energy flow is locally conserved in four dimensions as a 4-circulation, but not in 3-space. EM energy flow in 3-spacc is conserved over all space, since there are essentially equal amounts of positive and negative charge in the universe and in the virtual charge of the vacuum.

gauge: In electromagnetism, usually considered a possible choice for (change of) electric scalar potential and magnetic vector potentials, in such a manner that there is no any *net* change of force fields so that the change satisfies Maxwell's equations. There are indeed two changes of force fields, but the two are selected so that the two new force fields are equal and opposite, constituting a stress potential and continuously performing internal work on the system to create and maintain the excess system stress. The equal and opposite force fields, however, cannot translate electrons through loads to freely power them. We prefer to consider any change of potential — even a single potential changed, or both changed so that a net force field does result — as a "regauging", for it is an implied primary assumption *within the normal use of the term "regauging."* If only one potential is changed, a net force field results in the system and we call that "asymmetrical regauging". If both potentials are changed, but in such a way that a net force field still results, that is also an asymmetrical regauging. When both potentials are changed precisely in such highly selected manner that the two force fields created are equal and opposite, the force fields sum to a *translation* vector zero, and that is symmetrical regauging or *Lorentz regauging.* However, the energy of the two fields remains, having been transposed to a stress potential. Each of the two "balanced" equal and opposite force fields is continuously performing work in the system, but all that is created is additional system stress. Hence symmetrical regauging is obtained only by altering the stress and stress energy of the system, with two disequilibria in energy flow between the external environment and the system. This in fact constitutes both a curvature of the local spacetime and an alteration of the local active vacuum. It is a rotation of the frame of the regauged system out of the laboratory frame. In effect, the net EM field energy of the two new fields created by symmetrical regauging is transformed into a change in the local spacetime curvature, a change in the local vacuum, increased stress in the system, and rotation of the system's frame away from the lab frame. We focus upon the energy aspects, since in a system any change of potential - whether symmetrical or asymmetrical — is a change of the internal energy of the system, and therefore of its excitation. In a system, conventional "symmetrical force fields" regauging results in a *nonusable continous* energy change in the system (actually two of these changes), and a change in the internal stress of the system. In short, it is a special case of adding a *stress potential* to the system, which we consider a special form of *asymmetrical regauging of the supersystem* consisting of (i) the system and its dynamics, (ii) the local active vacuum and its dynamics, and (iii) local spacetime curvature and its dynamics.

gauge freedom: The axiom in gauge theory that one can freely change the gauge (and therefore the potentials and the potential energy) of an EM system, so long as the net new forces resulting from the change will sum to a zero vector resultant so that the system remains "symmetrical" in its force-field functioning with respect to the system prior to the change of potentials. Electrodynamicists do not seem to have recognized that the freedom to change the potential of a system at will, is *a priori* the freedom to make a purely negentropic change to increase the system's potential energy, or a purely entropic change to decrease the system's potential energy. Gauge and gauge freedom are actually formal restrictions of an even more important asymmetrical regauging principle: *a potential in a system can be freely changed at any time, thus freely changing the potential energy energy of the system freely and at will.* The additional symmetry assumption is an arbitrary added requirement not imposed by nature.

Further, it appears that physicists have not recognized that a longitudinal EM wave is an oscillating gauge varying in its magnitude. Said in reverse, oscillating *the Lorentz regauging condition for a circuit makes that circuit a generator and emitter of longitudinal EM waves.* This also provides a clue for detection of longitudinal EM waves, similar to what occurs in a Fogal semiconductor used for that purpose. A longitudinal EM wave detector must detect a variation in the symmetrical regauging condition, which is a stress change detection.

gauge symmetry: Abstract mathematical symmetry of a field, that relates to the freedom to regauge (change the value of) potentials, without affecting the *net resultant* values of the field quantities. This is a special case of the higher principle that any potential (and thus the potential energy of a system) can be changed freely and at will. However, changing the gauge of an EM system in that sense changes the potential energy of the system but *only in stress potential fashion.* In other words, it does nontranslating work on the system, which is called "internal work." The lack of a net resulting force in the system, created by the regauging symmetry, prevents use of this excess free energy to translate system charges and power an external load. All the excess free energy that could otherwise be used to power loads freely, is instead "locked up" to only perform internal work on the system to increase and maintain system stress.

gauge theory: A field theory using a field that has one or more gauge symmetries. The restriction to symmetry is a carry-over of the old erroneous notion that potentials are just convenient mathematical figments, and that fields are the primary causes of all EM phenomena. For the overunity researcher, broken gauge symmetry at a given level (usually in 3-space EM energy flow) is the feature that must avidly be sought in experimental systems if one would use asymmetrical regauging to allow free powering of loads. Hence one is primarily interested in a wide variety of *broken symmetry effects* and phenomenology. Particularly, unless the automatic self-enforcing of symmetrical gauge transformation during excitation discharge of the closed current loop system is violated, no EM system can produce COP>1.0. Further, local spacetime curvature between local input and local output is always involved in an overunity EM power system. The tempic (time) potential of the output end of an overunity system differs from the tempic potential of the input end, so that the gradient (the tempic force) is directed from output to input — exactly the reversal of the case for a COP<1.0 power system.

Thus in both COP<1.0 systems and COP>1.0 systems, a local spacetime curvature is involved when the system is in operation. In COP< 1.0 systems, the tempic force sweeps along from input to output and on into space beyond, where the violent interactions of vacuum absorb the associated Dirac sea hole current immediately. In COP>1.0 systems, the tempic force sweeps along through the local vacuum from the output section back through the system and into the input section, and thence on back through the external conductors to the external power source. This reversed Dirac sea hole current may "eat electrons" in the normal currents

internal to the system, and in the input section and back towards the external power supply it will "eat input electrons in the input current." Thus the Dirac sea hole current in the COP>1.0 system "eats input power" and serves as a curious new kind of electrical "load" or "internal loss" in the **input** section of the system. The reversed Dirac sea hole current is nature's way of decaying the local curvature of spacetime back to a flat spacetime, and decaying the COP>1.0 excited system disequilibrium state back **to an** equilibrium state. In any overunity electrical power system, this Dirac sea hole current must be taken into account and either transduced into useful power, minimized, or eliminated. An extension to gauge field theory is needed in which the tempic force, the tempic potentials and their gradients, and the supersystem are all modeled and taken into account. It appears that the Sachs unified field theory is capable of handling that requi**rement,** which also can be accommodated by O(3) electrodynamics to provide engineering of the resulting unified field theory.

gauge **transformation (in electromagnetics):** Conventionally taken to be the a**ddition** of the gradient of some function of space and time to the **magnetic** vector potential—and simultaneously the addition of the negative of the partial derivative of the same function with respect to time, divided by the speed of light, to the electric scalar potential. The resulting symmetri**cal** change assumes that the stress energy of the system is changed **freely,** but by design it deliberately prevents any stress difference in the system to produce a net force that could expend the free change of energy by doing work in an external load. Changing the stress energy alone is a curvature of spacetime and a purely gravitational change. It is also a **rotation** of the frame of the system away from the laboratory frame. See discussion under gauge theory.

general theory of relativity: Einstein's theory of gravity in which the gravitational force is represented by a curvature in spacetime, and in which spacetime is an active entity. We may look at all forces as being due to curvatures of spacetime interacting with mass.

geometric phase: A phase difference acquired by the state function of a quantum system that undergoes a cyclic evolution from an initial state and back. This phase difference can be detected by interfering the final and initial s**tates.** The modern generalization of the Aharonov-Bohm effect and the Berry phase.

giant nege**ntropy (of the dipole):** The continuous receipt from the time domain **of spacetime** (the vacuum), by one charge of any dipolarity (including **a** single charge's fundamental dipolarity), of enormous longitudinal EM wave energy flow in condensed time energy form, transduction of that absorbed energy from the complex plane into real 3-space, and continuous emission of that enormous energy flow in all directions in 3-space, thence receipt of that 3-space energy by the other charge of the dipole or other adjacent dipoles, retransduction of the energy back into time-energy, and re-emission of that energy back into the time domain. This 4-circulation giant negentropy process is evoked and continues automatically and indefinitely after a little energy has been expended to separate the charges and make the source dipole. The EM waves in the input from the time domain and the output in 3-space are perfectly ordered and correlated, which is a giant ordering of the vacuum/spacetime and is thus a giant negentropy. In conventional circuits, the source dipole in the generator or battery, once made, performs this function. Only a minuscule fraction (some 10^{-13} in a nominal case) of the energy pouring out of the generator's or battery's terminals strikes the surface charges of the attached external circuit and is diverged into the conductors as the Poynting component that powers the Drude electrons and the circuit. All the rest of the energy flow (the Heaviside nondiverged component) outside the circuit is just wasted by conventional circuits.

gluon: A particle that carries the forces between quarks. In quantum chromodynamics, the gluon plays a role analogous to the photon in quantum electrodynamics.

graviton: In the quantum theory of gravitation, the graviton is the quantum of the gravitational field. It is massless, with spin 2. In the new theory we may take the graviton as a coupled scalar and longitudinal photon pair.

gravity: The phenomenon characterized by the physical attraction between any two material bodies of positive mass, specifically due to the trapped positive energy in the masses. If one particle is of negative mass and energy, the normal gravitational attraction between the two particles is reversed so that they repel, which produces antigravity. The fundamental mechanism of gravity can be found once one realizes that Maxwell omitted the incoming EM wave interaction with the nucleus, and not with just the electron shells. When he wrote his theory, the electron, atom, neutron, proton, and neutron had not been discovered yet. The lateral precession movement of the spinning Drude electrons — which are restrained *longitudinally* to a very small drift velocity in detecting wires (antennas) — later came to be considered proof of Maxwell's transverse EM wave in the vacuum. The detected electron precession wave is oscillating laterally, hence proves the longitudinal EM wave interaction. Elsewhere we have pointed out that the EM wave in space is longitudinal EM wave accompanied by a corresponding time-polarized (scalar) EM wave, and it also follows from Whittaker's 1903 paper as reinterpreted by the present author in our Giant Negentropy paper.

gross particle translation: Refers to translating the entire particle — and its internal structures and dynamics — in space, rather than internally structuing the particle's potential (massless charge). Note that, if the internal structuring (engines) of the particle is changed during translation (e.g., by passing through a region of the flat spacetime having an altered internal engine structure), the particle may have an anomalous interaction with another particle or process having a different internal structure or "dimensioning." The mass-energy comprising the particle's nucleus, electron shells, etc. can be dimensioned. Mass so dimensioned or conditioned then may act differently in chemical reactions, effects on living systems, etc. than normal "inert" mass that has no specific deterministic dimensioning.

group: A mathematical structure that is a complete set of transformations on a mathematical entity that leaves the entity unchanged.

group, higher symmetry: In a given theoretical area, a particular theory can be expressed in various group transformations. When the group transformation (group symmetry) utilized is more complex than the "standard" group symmetry usually employed, it is referred to as a *higher* group symmetry. The mathematics of the model is then said to be a higher group symmetry mathematics. Since the groups have names such as U(1), SU(2), O(3), etc., the transformations of the mathematics are often referred to by **those** names. Hence characterizing phrases are used such as U(1) symmetry electrodynamics, SU(2) symmetry electrodynamics, O(3) symmetry electrodynamics, etc.

group, Lie: A group whose transformations act continuously. An example of a Lie group is given by rotations in ordinary space.

group theory: The general mathematics of symmetry transformations; i.e., the theory of all transformations that leave the transformed entity or entities unchanged.

gyroelectrons: Spinning electrons considered as gyroscopes during the portion of their spin in 3-space, so that when longitudinally perturbed they can and will laterally precess. Considering (in a gross first order model) the conduction electrons in the Drude gas in a conductor as little gyros, because they are (i) spinning on an axis, (ii) severely restrained in their ability to move forward longitudinally down the conductor as current, and (iii) much more free to precess laterally in the conductor. Thus when the gyroelectrons in the receiving wire antenna, perturbed by an incoming longitudinal EM wave in the vacuum, are longitudinally perturbed, they precess laterally so that our instruments (which detect electron wiggles, not the incoming vacuum EM wave) do detect "transverse waves." However, they are detecting the transverse gyroelectron precession waves, not the waves in the vacuum. They are detecting the *effect* of the interaction of the incoming waves with the gyroelectrons, not the *cause.*

harmonic series: A set of frequencies or oscillations consisting of integer multiples of a fundamental frequency.

Heaviside, Oliver: Noted English self-taught physicist and brilliant electrodynamicist, 1850-1925. who played a role in discarding Maxwell's quaternions, and also played a role in forming vector mathematics and formulating the vector reduction of Maxwell's theory from 20 quaternion equations in some 20 unknowns to the present 4 vector equations. Heaviside's curtailment of Maxwell's theory still contained both (i) systems in equilibrium with their active environment and (ii) systems in disequilibrium with it. To separate variables and enable closed solutions, *Lorentz further* curtailed the Maxwell-Heaviside equations by symmetrically regauging them, thereby *arbitrarily* discarding all disequilibrium Maxweilian systems and retaining only those in equilibrium with their active environment. This regauging unwittingly and erroneously discarded all that class of Maxweilian systems permitted to exhibit COP>1.0. The first such curtailment of the Maxwell equations was by Ludwig Valentin Lorenz in 1867, but H. A. Lorentz (who did a similar symmetrical regauging nearly decades later) was later given the credit.

Heaviside also discovered the enormous *nondiverged* energy flow surrounding every circuit, far greater in magnitude than the feeble Poynting *diverged* component that actually enters the circuit. But he was at a loss to explain what the source of such a startling flow of energy could be. Lorentz understood the Heaviside component, but also could not understand its source. Hence he reasoned that it "had no physical significance" since it did not power the circuit. Lorentz then introduced the trick of integrating the entire energy flow vector around a closed surface surrounding any volume element of interest. This neatly discards the vexing giant Heaviside *nondiverged* energy flow component, and retains only the *diverged* Poynting component that actually enters the circuit. The latter is of course the energy that will later be *emitted* from the circuit, which is what we will measure when we "measure the circuit."

Electrodynamicists are still using this method of dropping the Heaviside energy flow from any further accounting consideration. Hence they do not develop systems which intentionally increase their energy collection from this readily available energy flow around every circuit. The Bohren experiment conclusively proves the physical existence of the Heaviside nondiverged component, even though it is no longer accounted.

In his last years Heaviside realized the gravitational importance of his extra energy flow component, and worked out an electrogravity theory based on it. He did not live to publish it, but his papers on it were found some time after his death, beneath the floorboards of his little garret apartment. A learned society then published the papers.

Heaviside energy flow component: The enormous *nonintercepted* and nondive*rged* energy flow component pouring out of the terminals of every battery or generator, or charge or dipole, and missing the external circuit (or external intercepting charges) entirely so that it is does not contribute to the intercepted (Poynting) flow component that powers the circuit. Also, the enormous energy in the EM field, potential, or wave that misses the interaction of the field, potential, or wave with the assumed unit point static electric charge, and thus is not intercepted and collected (and accounted!) in the unit point static charge assumption. In honor of Heaviside, the present author has nominated this "dark Heaviside energy" component as the missing source of excess gravity known to be holding the arms of the spiral galaxies in place. In the interactions of Dirac hole current with a unit point charge (to determine the negative energy fields and potentials in a COP>1.0 EM system), there is also a missing component of the negative energy flow that is comparable to the Heaviside component for positive energy flow. These flows of surplus negative energy **from** COP>1.0 astronomical interactions—with the associated negative energy fields and negative energy potentials produced by the **negative-**energy Dirac holes in these negative energy flows — is hypothesized by the present author as producing the antigravity that generates the observed acceleration of the expanding universe.

Hertz, **Heinrich:** Noted German physicist, 1857-1894, who together with Heaviside and others transformed Maxwell's theory from quaternion algebra to vector algebra, and who also experimentally produced Hertzian (EM) waves, thus confirming Maxwell's theory. These experiments by Hertz first detected EM energy propagating as described by Maxwell's equations, thus providing the experimental impetus that allowed Maxwell's theory to predominate. The oscillation or "cycle" is named after Hertz, by specifying *X* number of oscillations or cycles per second as a frequency of *X* Hertz. One type of potential is also named after him.

hidden bidirectional EM waves: Those bidirectional EM wavepairs that comprise the scalar potential are indeed "hidden" with respect to our normal "electron translation" (i.e., electron wiggle) detectors and instruments. Nonetheless, properly functioning devices such as Fogal's semiconductor allow the utilization, transmission, and reception of the vast longitudinal EM wave dynamics that are infolded inside all conventional EM fields, waves, and potentials, including the giant potential comprising the ambient vacuum (spacetime). Use of this infolded electrodynamics is akin to using "subspace" or higher dimensions, depending upon how one chooses to model it mathematically. It is also a direct method of using a unified field theory.

hidden variable theory: A class of quantum mechanical theories which considers that the quantum state of a physical system is not a complete specification after all. The "hidden variables" are those additional components necessary to provide the "complete state" of the system. By utilizing the infolded longitudinal EM wave electrodynamics, one is in fact using a macroscopic hidden variable theory. This becomes a subset of Evans' O(3) electrodynamics, which in turn is an important subset of Sachs' unified field theory. Sachs' theory encompasses and subsumes electrodynamics, general relativity, quantum mechanics, etc., so the use of this macroscopic hidden variable theory through special functions provided by components such as Fogal semiconductors is a powerful and directly engineerable extension of conventional hidden variable theory. The physics of this new, engineerable, unified field theory area is also called *energetics.*

hole, lattice: Vacancy, in an otherwise filled energy band position. Sometimes called *negative ion vacancy.* In a semiconductor, an electron may be lifted to the conduction band by thermal energy, which produces u hole-electron pair (an exciton). Also, one of the sites in the crystal lattice can be occupied by an acceptor impurity atom, so that the electron is accepted there, leaving a hole showing excess positive charge. This is a lattice hole, and the propagation of such holes being formed is called electron hole current, or hole current for short. We point out that the *laittice hole* exists only after interaction of the Dirac sea vacuum hole with matter. i.e., only after observation. In that case, the 3-positron (hole) has positive mass and positive energy, produces positive energy fields and potentials, produces gravity rather than antigravity, and is "seen" to be going in the opposite direction to the Dirac sea hole. Prior to observation, the hole is a

Dirac 4-hole in vacuum, having negative energy and therefore negative mass equivalence, negative charge, and without a direction reversal. This 4-hole is a source of negative energy fields and negative energy potentials in **the** surrounding space, and thus of antigravity back-interaction (upon the physical system connected with the active vacuum component of its supersystem) due to the negative curvature of the local spacetime.

hole, Dirac sea: In the presence of fields (i.e., in the presence of curved space**time),** some of the negative energy states in the Dirac sea are not filled with electrons. In that case there are some empty 4-holes in the Dirac sea, in local spacetime itself. The Dirac hole is a negative energy state in the **vacuum** itself. It too *seemingly* has a positive charge, as seen by the external observer after interaction with matter. Holes are also mathematically equivalent to positrons, in case observation has been evoked, so long as the lack of positive mass is not a consideration Note that our view differs from the received view, but is based on experimental evidence which can only be gained from COP»1.0 EM system experiments. We know of no such COP»1.0 EM system experiments involved in the received view.

hole current: (1) In a semiconductor: Usually, the part of the current in a seminductor that is composed of migration of lattice holes. ***Mathematically,*** within certain constraints this may be considered a positro*n* current. The hole migration is usually slower than the electron migration producing electron current. (2) In the vacuum itself. In the *Dirac sea, in* **the** presence of fields the Dirac 4-holes flow in the opposite direction to the local electron current, providing a Dirac sea hole current or posit**ron** current. [See discussion under hole.] An electrical power system in disequilibrium with its vacuum interchange and having COP>1.0, will have a **greater** *tempic potential* at its output section than at its input section. Thus it will have a tempic force or tempic broom, sweeping holes from local instabilities to form a hole current (may be both a Dirac sea hole current and a lattice hole current) flowing from the output section of the system back through the system into the input section, and on out of the input section into the primary power source. *Hole current transduction may be* used to convert this detrimental hole current energy transport into usable **electron** current energy transport, or anti-circuit techniques and mechanisms can be used.

hole current transduction: Hole current in a COP>1.0 open system, where **th**e system is in disequilibrium with the vacuum *a priori,* is usually detrimental to the system and may destroy the overunity condition. A hole current from the input section of the system, fed back into the primary power supply, requires that extra electron current be furnished since some of the electrons in that furnished current will recombine with the oppositely moving Dirac vacuum holes and disappear to fill those holes. Hole current into a battery in "normal" fashion discharges the battery rather than charges it (although reversing the battery terminal connections allows charging in normal fashion using the hole current). Thus a battery can be discharged "far below zero", so to speak, by charging it with excess holes in the "normal connection." A capacitor may be charged with hole current, in which case the dielectric will undergo compressive strain into the positron-charged plate. However, when the capacitor is subsequently discharged, the dielectric will expand and recover, with its electron charges pushing electrons out of the negative side plate and into the external circuit. Hence a capacitor can be *charged* with positron (hole) current, and then can be adroitly switched so as to *discharge* electron current. In theory so can a battery because of its capacitance. This is a transduction of entropy to negentropy, a process for which a patent application has been filed by Bedini and Bearden. Gauge freedom already permits that operation, since gauging in electromagnetic systems is freely changing the local potential energy of the system, and is either negentropic (increasing the system's energy) or entropic (decreasing the system's potential energy).

hole injection: In an n-type semiconductor, holes can be emitted by applying a metallic point to the semiconductor's surface. We suspect (but have not proved) that sharp gradient pulses may also induce some of these holes as Dirac sea 4-holes and Dirac sea hole current rather than positron and lattice holes current. In that case, a mix of positive energy and negative energy (and gravity and antigravity) would be produced.

hole trap: An impurity in a semiconductor that can trap a "hole" by releasing electrons to the conduction or valence band.

hyperspace: Refers to a space of more than four dimensions, and specifically to those spatial dimensions outside the normal three. The term is used for both a space of n dimensions, where n>4, or for one dimension in such a space, beyond the normal four.

identity of opposites: On the boundary between A and not-A, A and not-A are identical (i.e., no difference can be discriminated). If all of a thing is collected, one reaches the boundary of that thing and it turns into its own opposite. The philosopher's problem of the "accursed necessity for the identity of opposites" cannot be solved in Aristotelian 3-law logic, since that logic is incomplete. It is readily solved, and part of, a more extended logic such as 5-law logic. The result is that "identity" is not an absolute entity, but is a decision by the observer reached by application of a comparison algorithm to separate observations made at differing times. An example is provided by a colorblind observer who sees as identical two marbles which, to an observer with color vision, are seen as red and black and therefore not identical. Aristotelian logic omits the action by the observer, the times of the multiple observations (only a single thing at a time is observed or perceived, *a priori),* and the decision algorithm used *by the* observer to decide whether two observations are identical or not identical.

inertial frame: A frame of reference in which force-free bodies move along straight lines, and the postulates of special relativity are valid. In the new approach, this notion is extended. Even in an inertial frame, force-free bodies may move along straight lines, the postulates of special relativity may remain valid, and yet the EM hidden variables may induce other changes in the usual laws of nature due to the action of hidden vacuum engines infolded in the *dimensioned* local flat spacetime (vacuum potential). See Einstein's second postulate extended.

infolded real vector components: Simply take a set of real force vectors that altogether sum to a vector zero resultant. This "vector zero system" is a zero vector with respect to forced *translation* of a particle or mass. It is not am absence of the vectors at all, but is the presence of specifically hidden and "infolded" real vectors that do not initiate bulk translation of the affected object or system, but perform *internal* work upon the system. Such vector systems are stressing systems of a special kind, where the internal components of the stress are not random but deterministic. Hence they are electrodynamic *engines,* in the new approach. To a somewhat arrogant mathematical scientist who once objected that zero vector resultant systems had no effect, I humorously challenged him to put it to experimental test. I suggested he stand between two elephants pushing against him in opposite directions, while I would stand between two fleas doing the same to me, and we would compare notes. He apparently was not amused at the suggestion.

infolding: With reference to signals, infolding means "placing the signal or modulation or change upon one or more of the internal bidirectional wavepairs comprising the scalar potential. Since all EM potentials ultimately represent changes to the ambient vacuum potential, one has conditioned or altered the vacuum itself when infolding is utilized, infolding of EM signals, vectors, and directions produces *engines*. By Whittaker's decomposition of the potential and Bearden's reinterpretation, these vacuum engines can be designed to affect matter — including atomic nuclei, nucleons, etc. — in any manner desired.

information content of the field: Russian euphemism for the deliberate structuring of the hidden longitudinal EM waves and their dynamics, that comprise any EM potential, field, or wave, so that the potential, field, or wave is *dimensioned* and contains active *engines.* This area was extensively weaponized under KGB (now changed in name) control rather than the regular Russian armed forces, and a disinformation campaign was waged to convince Western analysts that the term refers to simple spectral analysis. It doesn't. Instead, it refers to infolding and outfolding, and the use of longitudinal EM waves, dimensioning, and engines. The overall science is called "energetics" by those weapon scientists.

interferometer: An instrument or system in which a wave is split into two waves or beams, which after traveling over different paths are subsequently reunited and interfere. The use of longitudinal EM waves (LWs) allows ready penetration of mass by the LWs traveling through the interior Whittaker composition of a mass's internal EM fields, waves, and potentials. Longitudinal EM wave interferometers should prove useful in probing the interior of planets, the sun, etc. They pass easily through the earth, so that interferometry can be obtained at targets on the opposite side of the earth, or inside the earth. These are the weapons that then Secretary of Defense Cohen referred to in 1997 as being used to engineer the weather, initiate volcanoes into eruption, or initiate earthquakes.

Josephson effect: Flow of Cooper pairs (superconducting electron pairs) across a thin dielectric separating two superconducting electrons, in the absence of a voltage drop. E.g., the pair of electrons can "tunnel" through u thin insulating barrier.

joule: A measure of energy; one watt for one second in electrical terms. In physics the unit of work or energy in the MKS system, being the amount of work done by one newton acting through a distance of one meter, equal to 10,000,000 ergs.

Kirchhoffs current law: At any time, the sum of the instantaneous currents flowing into a node or point in a circuit equals the sum of the instantaneous currents flowing out of that node or point. This does not have to be true if (1) relativistic effects are appreciable, or (2) electron production and annihilation processes — such as Dirac sea hole current, lattice hole current and recombinations are occurring and non negligible, (3) the Kron open path is being used, or (4) localized time reversal zones (TRZs) are utilized. Another way of stating where the law need not hold is when there are nonlinear interactions between the three components of the superset (the physical system and its dynamics, the active vacuum and its dynamics, and spacetime and its curvature dynamics.).

Kirchhoff's voltage law: At any time, the sum of all voltage rises in a closed loop circuit equals the sum of all drops in that closed loop circuit. The law assumes a flat spacetime. This does not have to be true if relativistic effects are appreciable, if true negative resistor effects are present, if Kron open paths are utilized, and if time-reversal effects arc present. It also need not be true in anti-circuit techniques, or when they are involved in a particular portion of an otherwise normal circuit.

Lamb, Willis Eugene Jr.: American physicist and Nobelist who discovered the hyperfine structure of the hydrogen spectrum, and first measured the interaction of the vacuum energy that generates a change in the energy levels of the hydrogen atom, according to Dirac's theory and quantum electrodynamics.

Lamb shift: A very small correction to the first excited state of the hydrogen atom due to interaction of the vacuum. Willis Lamb experimentally confirmed this prediction of quantum electrodynamics to high accuracy. The local energy density of the little Lamb shift is greater than the surface energy density of the sun. However, in matter there are such a vast number of similar small interactions generating very powerful fields randomly, that the sum total of all these changes sums to a vector zero for most summation distances of interest.

lamellar current: Thin layers of current; thin sheets of current.

laser: A device that generates light in which all the photons are exactly in step and produce a coherent beam. Originally the word "laser" was an acronym that stood for "light amplification by stimulated emission of radiation."

law of physics: A restriction placed upon the physical mechanisms that can occur under certain broad assumptions as to the background situation and environment. Because of the assumptions made, an expressed law of physics is not absolute, but relative to the assumed conditions. Any "law of physics" can be violated if the conditions contained in its assumptions are violated. It is important that a scientist be aware that physics is comprised of models and the "laws of physics" are very general and wide-ranging models based on very specific assumptions. Further, by Godel's theorem and the theory of models, no model is perfect anyway, and none will ever be. When one assumes that physical laws are not based on any assumptions but are universal and absolute truths, one has immediately moved to dogma and not science.

Lawandy patents: Patents covering and related to a form of anti-Stokes emission (excess emission of energy) in strongly scattering, optically active media such as TiO_2 particles in suspension. This is a new form of lasing, without the requirement for population inversion.

Letokhov, V. S.: Russian pioneer in excess emission phenomena from stimulated media.

Letokhov effect: The excess emission (excess emission of energy) phenomena in strongly scattering, optically active media, usually referred to as "negative absorption of the medium." See particularly V. S. Letokhov, "Laser Maxwell's Demon," *Contemporary Physics,* 36(4), 1995, p. 235-243.

localized: Confined to a small region of a large system rather than being extended through the system or at a distance.

localized virtual photon flux of vacuum: Refers first to a local region of the vacuum, and then to the flux density of virtual photons in that local region. If that flux is deterministically structured, then the local region also contains an *engine* that will act internally upon masses and mass systems in that region. The local region also contains hidden variables.

longitudinal component: In EM fields, the longitudinal component of the fields along the direction of travel, as opposed to the transverse components at right angles to the direction of travel.

longitudinal E-field: Any electrical field in 3-space may be described using two components, a transverse component and a longitudinal component. Of course one or the other may be a vector zero. In 4-space, there is a time component also, as an additional "transverse" component.

longitudinal gradient: Compression wave gradients (differences) in density or pressure, considered in a direction from their origin along the direction of primary movement in the medium. Contrast to transverse waves, which act perpendicular to their direction of propagation. In electrodynamics, longitudinal EM waves in 3-space are said to be "longitudinally polarized." Longitudinal EM waves in the time-domain are said to be "time polarized." The latter are also referred to as "scalar waves" since they have no polarization vector component in 3-space and thus arc "scalar" with respect to 3-space. This is a misnomer, however, since these "scalar waves" are perfectly good time-polarized EM longitudinal vector waves. For the first time, the giant negentropy process allows one to make and use time-polarized EM waves, since they automatically accompany all longitudinally polarized EM waves and it is straightforward to make longitudinal EM waves in 3-space (just oscillate the Lorentz regauging condition).

longitudinal interior wave: An infolded longitudinal gradient that oscillates in magnitude along the line of travel "inside the vacuum potential or a normal potential." In conventional theory, the electric portion of the EM wave is a transverse field oscillation, and the longitudinal component is essentially zero. In the new "infolded" wave approach, there is no overall transverse gradient in the vacuum potential, but there are such longitudinal waves. This is rather like a "pressure wave" transmitted underwater, without causing a surface wave involving physical upheaval and fall-off of the water. The speed of the longitudinal EM wa**ve** is not limited to c, the speed of transverse EM wave light in the ambient vacuum..

longitudinal wave: A "pressure" type of wave, similar to sound, in which the vibrations are along the direction of travel of the wave. Hence, a wave composed of alternating densifications and rarefactions, where we focus upon the longitudinal component of the changes. One can think of it as a sort of "accordion" wave, rhythmically shrinking and stretching along the line of motion as it travels. Note that rhythmically and symmetrically varying the two potentials involved in regauging creates longitudinal EM waves. So one pseudonym for longitudinal EM waves is "gauge oscillation" waves. Matter is mostly empty space filled with EM potentials, fields, and waves—all of which are bundles of longitudinal EM waves and their functional dynamics. Hence matter (such as the ocean and the earth) is a vast superhighway for the passage of longitudinal EM waves. Note that the longitudinal wave represents an oscillating energy density in space and/or time (scalar waves), and hence is an oscillating curvature of spacetime. All forces of nature are generated by curvatures of spacetime. Hence all forces of nature can be generated by the correct generation of longitudinal EM waves and their impressed dynamics, Further, this means that all the physical *causal* functions of the exhibited **effects generated** in a material system can be captured in longitudinal EM wavesets. A longitudinal wave functional system deliberately having such causal functions of a physical functioning system is known as a causal **system robot** (CASR). Such robots are far beyond the presently developing nanobots, and have been developed rather extensively by at least one nation. Such robots can easily pass through the "interior" of any normal

EM signal or wave or potential or field, which again is nothing but a giant superhighway for longitudinal EM waves. In theory one can communicate with a CASR via longitudinal EM waves if communications functions are built in to the CASR. In the future, one would expect that planetary exploration, underground and undersea exploration, exploration of the interiors of the sun, moon, and planets, could be enabled by the further development of specialized CASRs.

Lorentz, Hendrik Antoon: Renowned Dutch physicist and Nobelist. 1853-1928, who developed the Lorentz transformations and Fitzgerald-Lorentz contraction. Lorentz shared the Nobel Prize for discovering the Zeeman effect. He is also responsible for symmetrically regauging Maxwell's equations (some time after Lorenz did it in 1867), and for introducing an integration trick that neatly disposes of the vexing huge Heaviside nondiverged energy flow component accompanying every field/charge and potential/charge interaction.

Lorentz invariance principle: An inertial frame is independent of the velocity of the frame relative to any other inertial frame. This is not necessarily true with respect to engines inside inertial frame objects, since the engines involve acceleration dynamics.

Lorentz regauging: Changing the two potentials in the Maxwell-Heaviside equations in potential form so that the resulting equations have the variables separated, and there is no net force introduced by the two new forces, which are deliberately chosen so as to be equal and opposite and thus sum to zero. An unfortunate effect is that this discards all Maxwellian systems far from equilibrium in their vacuum exchange, hence discards all permissible Maxwellian COP>1.0 systems. Electrodynamicists have not recognized that Lorentz regauging represents the addition of stress potential energy and continuous internal work and stress to a regauged system. It also represents a rotation of the frame of that system out of the laboratory frame by some angle. Further, oscillating the Lorentz regauging condition produces longitudinal EM wave generation and emission.

Lorentz symmetry: The symmetrical regauging by Lorentz of Maxwell's equations.

Lorenz, Ludwig Valentin: Scientist who first changed Maxwell's equations — two years after their publication in 1865 — by (effectively) symmetrically regauging them, thereby unwittingly discarding all Maxwellian systems far from thermodynamic equilibrium with the active vacuum. See L. V. Lorenz, "On the identity of the vibrations of light with

electrical currents," *Philosophical Magazine*, Vol. 34, 1867, p. 287-301. In this paper [L. V.] Lorenz gave essentially what today is called the [H. A.] **Lorentz** symmetrical regauging. When H.A. Lorentz also symmetrically **regauged** the Maxwell-Heaviside equations, the physics community **adopted** the procedure and resulting further truncation of the Maxwellian theory. That truncation continues to this day.

low energy **nuclear reaction** (LENR): Actually a bit of a misnomer, since it refers to a low *spatial* energy, high *time* energy reaction. Cold fusion, or the anomalous transmutations of elements that occur in certain electrolytes with specially prepared palladium electrodes. The transmutations occur even though in conventional theory there is insufficient *spatial* energy to accomplish them, because they use time energy. Time energy is spatial EM ene**rgy** compacted by the factor c^2, hence time has the same energy density as mass. The formation of temporary, fleeting time reversal zones (TRZs), with action from from inside-to-outside, is the genesis of a host of new nuclear reactions missing from conventional particle physics since the physicists do not account time as energy, and do not use such TRZs. For more on this hypothesis and the specific reactions generated, see T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," *Journal of New Energy*, 3(2/3), 1998, p. 12-28 as well as Chapter 10 of the present book.

magneti**c dipole:** Paired north and south magnetic poles of equal strength ever microscopic or macroscopic distance.

magnetic induction: The act of a change in magnetic field producing an electric field at right angles, which in turn produces an electron flow.

magnetic spin: The intrinsic angular momentum, known as spin, of electrons, protons, and neutrons, etc. The electrons combined in an atom or ion have **a** resultant angular momentum that is the combined intrinsic spin of **the** electrons and the angular momentum due to their motion around **the** nucleus. There is also a magnetic moment associated with this angular momentum (spin) whenever it is nonzero. Therefore atoms or ions with non**zero** spin are *magnetic* atoms or ions. This is often loosely reffered **to a**s "magnetic spin." It is actually the magnetic moment of nonzero spin.

magnetostatic scalar potential: Continuous mathematical function whose value in space, at any point it occupies, represents the potential energy collected upon an assumed unit magnetic north-seeking charge (pole). There is a magnetostatic scalar potential between the poles of any magnet. This scalar potential is **also** subject to **Whittaker's** 1903 decomposition into a harmonic series of magnetic longitudinal phase conjugate wavepairs. Reinterpretation of Whittaker's wavepairs should be made according to Bearden, "Giant Negentropy of the Common Dipole," *J. New Energy,* 5(10, Summer 2000, p. 11-23. This reinterpretation agrees with quantum field theory; see Mandl and Shaw, *Quantum Field Theory,* Revised Edition, Wiley, 1993, Chapter 5.

matter: Highly condensed 3-spatial energy, compressed by factor c^2 and remaining in 3-space, and with the observation process *d/dt* having been applied to the fundamental masstime entity to provide mass as a frozen 3-space snapshot by stopping the flow of time momentarily. Matter, of course, has internal structure, order and organization, and dynamics. All matter is in a continuous energetic exchange with its active vacuum (4-spatial) environment. Matter stripped of its EM fields, waves, and potentials is called mass. Mass is an observable and therefore a frozen series of continual 3-space snapshots, formed by incessant photon absorption to change the mass *m* to masstime *mt,* with following incessant photon emission to change masstime *mt* back to mass *m.*

Maxwell, James Clerk: Brilliant Scottish scientist, 1831-1873, who first formulated the unified theory of electricity and magnetism. See *The Scientific Papers of James Clerk Maxwell,* 2 vols. bound as one, edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-597 for his definitive paper, "A Dynamical Theory of the Electromagnetic Field," first published in *Royal Society Transactions,* Vol. CLV, 1865. In that paper, (see Niven's book) his general equations of the electromagnetic field are given in Vol. 1, Part III, "General Equations of the Electromagnetic Field," p. 554-561. On p. 561, he lists his 20 variables. On p. 562, he summarizes the different subjects of the 20 equations, being three equations each for magnetic force, electric currents, electromotive force, electric elasticity, electric resistance, total currents; and one equation each for free electricity and continuity.

Maxwell's famous book, *A Treatise on Electricity and Magnetism,* Oxford University Press, Oxford, 1873, Second Edition 1881 (Maxwell was already deceased), Third Edition, Volumes 1 and 2, 1891. Foreword to the second edition was by Niven, who finished the work. Maxwell had dramatically rewritten the first nine chapters to simplify his mathematics because of the resistance to it, with much new matter added and the forma contents rearranged and simplified. Maxwell died before finishing the rcsl of the second edition. The rest of the second edition is therefore largely a reprint from the first edition. The third edition edited by J. J. Thomson was published in 1892, by Oxford University Press, and later was published unabridged by Dover Publications, New York, 1954. J. J. Thomson finished the publication of the third edition, and wrote a "Supplementary Volume" with his notes. A summary of Maxwell's simplified equations are given in Vol. II, Chapter IX of the third edition. However, Maxwell had gone (in his second edition) to some pains to reduce the quaternion expressions himself, and not require the students to know the calculus of quaternions (so stated on p. 257). We note that Maxwell did not finish the second edition, but died before that. He actually had no hand at all in the third edition as to any changes. The Second edition was later finished by Niven by simply adding the remaining material from the previous first edition approved by Maxwell. The printing of the first nine chapters of the thirg edition was already underway when J. J. Thomson was assigned to finish the editing of the manuscript.

Maxwell's work profoundly affected the course of science and the course of human history. Maxwell, Heaviside, Hertz, Gibbs, Lorcntz and others participated in dramatically curtailing Maxwell's original 1865 quaternion theory to the much more limited subset that it is today. Since the early part of the 20th century, work has slowly been ongoing to move EM theory back to a higher group symmetry theory, which ironically it already was when Maxwell first created it.

Max**well's** electrodynamics: Simply put, Maxwell's theory of electrodynamics consists of Maxwell's equations. His fundamental theory consisted of some 20 quaternion equations in 20 unknowns, appearing in his 1865 paper. After Maxwell's death and some curtailment by Maxwell himself, Heaviside modified and sharply curtailed these equations, into the familiar four equations of today, as did Gibbs and Hertz. Lorentz further curtailed the Maxwell-Heaviside equations, by symmetrically regauging them. This simplified their mathematical solution, but it also inadvertently and arbitrarily discarded all open Maxwell systems far from thermo**dynamic** equilibrium with their active environment (such as the modern active vacuum). See James Clerk Maxwell.

Maxwell's Theory, fluid dynamics analogy: *Considering and adapting the fluid analogy approach of James Clerk Maxwell:* The intensity of the vacuum virtual particle flux may be treated as the intensity of a potential, thus by analogy corresponding to pressure in a fluid. Polarity (charge) is a **process for** establishing a change in the intensity of the vacuum potential - and **thus** a change in the ambient pressure of the vacuum fluid. Dif**fence** between vacuum pressures at two separated points creates a force u**pon** any charge placed between those two points. Just as a differential in pressure between points in a fluid induces currents, a differential in vacuum pressure induces energy flow currents in the vacuum fluid, which are called 'fields'. All EM circuits, charges, potentials, and forces in electrodynamics involve altering the vacuum to include altering its pressure and thereby producing the forces the resulting pressure gradients (fields) induce upon charges. All observable EM energy in space is EM energy intercepted and collected on charges, or assumed to be, and the spatial EM energy comes from the vacuum via these fundamental vacuum-engineering dynamics which all involve the giant negentropy 4-circulation between the time domain and 3-space.

Michelson, Albert Abraham: German-born American physicist, 1852-1931, who invented an interferometer (Michelson interferometer) named after him. Michelson used his interferometer in a renowned experiment in the 1880s designed to measure the velocity of the earth through the ether, which would produce a shift in his interferometer pattern. No such shift was observed, which indicated the absence of an ether wind and was largely responsible for the downfall of the concept of the material ether that had been previously assumed, including by Maxwell.

motionless electromagnetic generator (MEG): Transformer-like COP>1.0 electrical power generator invented by Patrick, Bearden, Kenny, Hayes, and Moore, ""Motionless Electromagnetic Generator," U.S. Patent # 6,362,718, Mar. 26, 2002. A second patent application is processing and several more are in preparation. In particle physics, any dipole (including a permanent magnet) is a known asymmetry in the fierce vacuum flux, continuously pouring out real EM energy extracted from the vacuum. The MEG continuously powers its transformer section core with a permanent magnet dipole, more than doubling the EM energy made available from **the** active vacuum via the permanent magnet dipole's broken symmetry in its exchange with the vacuum. [See M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., "Explanation of the Motionless Electromagnetic Generator with O(3) Electrodynamics," *Found. Phys. Lett.,* 14(1), Feb. 2001, p. 87-94; "Explanation of the Motionless Electromagnetic Generator with the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* 14(4), Aug. 2001, p. 387-393]. The MEG nanocrystalline core material extracts the magnetic field B of the permanent magnet from the surrounding space and localizes it in the core material. The active vacuum asymmetry freely replenishes the withdrawn magnetic B-field energy in space surrounding the core, but with uncurled magnetic vector potential A rather than magnetic field B. This is an application *of geometric phase,* specifically the well known *Aharonov-Bohm effect,* with several thousand papers in the literature. The MEG then transduces and outputs magnetic B-field energy inside the **core** material and some of the field-free magnetic vector potential A energy in space surrounding the core, effectively having available more than double the energy being made available from the permanent magnet, or normally available to a transformer. The output power of the MEG is a function of (i) the time rate of change of a small amount of input EM perturbation energy to the input coil, plus (ii) the intercepting and collecting secondary charges available in the secondary coil, (iii) additional energy from chaotic oscillations, (iv), sophisticated feedforward and feedback regenerative self-oscillations, and (v) the geometric phase (Aharonov-Bohm effect). It thus **is** a rigorous practical application of the geometric phase effect to a macroscopic power system, and in fact is the first such application in practical electrical power systems. The noted feature of the MEG is a purely electrical interaction in the output coils, in which the voltage and energy are in phase within 2°, differing from all other transformer-like systems. Thus the MEG directly delivers real power rather than apparent power. See Aharonov-Bohm effect.

multicollection (of energy): Multiple collection and re-collection in a circuit of the same energy flow from a source dipole, by rerouting either (1) the dissipated Poynting energy back around to the collectors for another pass by them, with additional collection, or (2) retroreflecting the nondiverged Heaviside EM energy flow component back and forth across the surface charges in the circuit conductors, iteratively diverging additional energy into the conductors to potentialize and power the Drude electrons, or (3) a combination of the two. The interesting point is that this effect **can** "use" the same energy to do repeated work, multiple times. It takes advantage of the fact that work is merely the change of form of energy, and that the new form can again be changed (as for example, back into the **first** form again) to do additional work, again and again so long as the energy in each transformation is held in the system without escaping. This process permissibly violates classical equilibrium thermodynamics and its infamous second law, since it is an open system freely receiving excess energy from its active environment and thus far from equilibrium with that active environment. As is well known, such systems are permitted to exhibit COP>1.0 and even self-powering of the system and its load simultaneously.

negative absorption of the medium: Excess emission from an active medium, as in the Letokhov work and in Bohren's experiment. The excess energy is first absorbed from the vacuum interaction (actually from Heaviside's unaccounted energy flow component that usually does not interact with the system), and then it is re-emitted. This is actually a true "Maxwell's demon," as pointed out by Letokhov.

negative resistor: Any component or function or process in an EM system that freely receives energy from outside the system in unusable or disordered form within the system, transduces that energy into usable form, and outputs it so that it can do useful work. In a *true* negative resistor, that is the *net* function performed. We specifically do not include "differential" negative resistors such as the tunnel diode, thyristor, and magnetron, which dissipate and disorder more energy from the system itself, overall, than the fraction that they order and furnish back to the system in their negative resistance regime. In anti-circuits, true negative resistors are quite natural components. Every charge and dipole in the universe is already just such a negative resistor, freely extracting unusable EM energy from the seething vacuum, transducing it, and continuously pouring out real observable EM energy in all directions in 3-space. A negative resistor can produce the five magic functions: (i) self-ordering, (ii) self oscillation or self-rotation, (iii) output more energy than the operator inputs, (iv) power itself and its load freely, and (v) exhibit negentropy. The condition for self-powering of a system containing a true negative resistor was given by Gabriel Kron.

negentropic: Functioning to produce or involve negentropy; reordering previously scattered energy without requiring the operator to input energy to perform the reordering. A simple example is passive retroreflection of radially scattered EM energy. The conventionally unaccounted enormous and nondiverged Heaviside energy flow component surrounding every circuit and every field/charge interaction is an example of a dynamically reordered and organized energetic vacuum, and thus an example of a **giant** negentropy process. The continuous observable EM energy freely pouring at the speed of light from every dipole and charge in the universe is also **an** example of the giant negentropy process.

negentropic engineering: Expending a little bit of energy to form a dipolarity, and then intercepting, collecting, and using some of the enormous EM energy freely and continuously extracted by the dipolarity from the vacuum, indefinitely, without using any of the extracted EM energy to scatter the charges in the dipole and destroy it. Retaining the dipole intact avoids shutting off the furnishing of copious free energy **flow** from the vacuum to the attached external circuit. As an analogy, it allows an electrical windmill operating in (driven by) a free electrical wind from the external environment of the system.

Newtonian recoil: Newton's third law reaction from the affected entity back to and upon the effecting entity. The "action" is the action from **the** effecting entity upon the affected entity. The "reaction" is the action from the affected entity back upon the effecting entity. Newton's third law actually implies that the cause interacts upon the effect and the effect reacts equally back upon the cause, thus producing a strange kind of equilibrium. In a time-reversal zone, the "recoil" is reversed in direction by parity reversal. Tentatively, one may interpret from pumped phase conjugation that the Newtonian third law reaction of one single acting force can be amplified by properly directing the actions of other acting forces. An interesting aspect is that a pumped phase conjugate mirror does not recoil, no matter how powerfully it is pumped or how powerfully amplified is the emitted phase conjugate replica wave.

Newton's third law: For every action there is an opposite and equal reaction. This law should be extended and restated something like this: "for every action there is an opposite and equal reaction if the causative mechanism of the reaction is allowed to occur without interference and is not first redirected away from its target." An example of a violation of Newton's third law as usually stated is provided by the emission of a phase conjugate replica (PCR) wave from a phase conjugate mirror material. No matter how powerfully pumped, the PCM does not recoil by the emission of the PCR wave. The reason is that the antiwave cause of Newtonian third law recoil was intercepted by nonlinear multiwave interaction (wave-to-wave interaction) before reaching its target nuclei and interacting with them, and was diverted away from the atom and on out of the phase conjugate mirror material. Since the causative mechanism for Newtonian recoil did not occur on the mass of the mirror, the mirror did not recoil. We also point out that, if Newton's third law is indeed engineerable and controllable, then COP>1.0 systems and self-powering EM systems are permissible a priori. In electrical system terms, this simply means that, if one can rid the system of its equal back emf or back mmf, the system can be legitimately self-powering because it no longer self-enforces the Lorenz-Lorentz symmetrical regauging condition.

non-Abelian electrodynamics: An electrodynamics based on a group theory in which the group operation is not commutative and hence not Abellian in nature. Contrast to an Abelian group in which the group operation is commutative.

nonequilibrium thermodynamics: Colloquial term referring to the newer thermodynamics of self-organizing systems and structures, far from euilibrium in their energetic exchange with their active environment, Since the 1960s, there has emerged in physics (and in thermodynamics!) this surprising new area: the physics of far-from-equilibrium systems and the self-organization and large-scale ordering of such systems. The equilibrium conventionally referred to is 3-space equilibrium, with no consideration of temporal equilibrium, although some pioneers of the field such as Prigogine are quite aware of the temporal disequilibrium also involved. Note that such "3-disequilibrium" thermodynamical systems are also in temporal energy flow disequilibrium as well as 3-disequilibrium, but overall are still in 4-equilibrium energetically. See **giant negentropy (of the dipole).** For a necessary thermodynamics discussion, see Appendix A in this book.

nonlinear material (optical): A substance whose optical properties vary with the intensity of the light.

nonlinear optical functioning: Functioning in a manner prescribed by nonlinear optics and quantum optics, particularly with respect to time-reversal effects, use of phase conjugate mirrors, phase conjugate reflection, wave-to-wave interaction, and pumping for amplified phase conjugate reflection. To quote Knight, "A phase conjugate mixing device has remarkable optical properties: perfect retroreflection, the complete cancellation of optical aberrations and inhomogeneities in wavefronts, and the ability to reverse an optical beam and return it to its source." A most interesting modern science, eventually adaptable to COP>1.0 systems, is developing in the study of intensely scattering optical materials.

nonlinear phase conjugate optics: See **phase conjugate optics.**

nonlocalization: Producing or rendering an effect or effect-producing agent at a distance, rather than at the immediate location of the effector. Extended throughout a large region or system, or to a distant part of a large region or system, rather than being confined to a small region of the large region or system. Nonlocalization effects are absolutely prescribed in quantum mechanics, and are experimentally confirmed. They have been largely ignored in classical electrodynamics and electrical engineering. Nonlocalization can result from the existence and use of a multiply connected space, as in Bohm's quantum potential and hidden variable theory. Engineering of action at a distance by engineering the quantum potential through multiply connected space gives one method of energy amplification. This method has unfortunately been secretly weaponized In five nations of the world, and none of the technology is available for use in electrical power systems for energy amplification and solving the energy crisis.

non-Riemannian geometry: Riemannian geometry is the geometry **of** Riemannian manifolds — i.e., manifolds having an infinitesimal squared distance *ds* that is everywhere positive. A non-Riemannian geometry violates this property. [A manifold is a topological space with certain prescribed characteristics.]

nucleons: A collective name for the neutrons or the protons in a nucleus. In the nucleus, the neutrons and protons are continually changing one into the other by exchange of the charge. What is called a neutron and what is called a proton is thus somewhat arbitrary, so long as one makes a rule and stands by it. The interesting thing is that, in a nucleus, there is already continous and highly rapid transmutation activity between nucleons, all the time, without giant accelerations and operator input of very intense spatial energy. In short, nature shows in all matter that matter can be and is continually transmuted from instant to instant, and is "constant" only on the "**average".** Since most present models used for this are statistical rather than causative, it strongly suggests that we should look for a causal mechanism. The use of the time-reversal zone and time-energy has been nominated in this book as providing a causal mechanism for transduction, which can be directly engineered. It appears that cold fusion research already accomplishes transmutations using the suggested mechanisms, Several specific nuclear transmutation reactions applying the new mechanisms, and consistent with hundreds of cold fusion experiments, are given in the present book.

O(3) electrodynamics: An electrodynamics extended from a two-dimensional inner group symmetry to a 3-dimensional inner group symmetry. O(3) electrodynamics also becomes an important subset of Sachs' unified field theory. The inner group symmetry of O(3) can be taken as normal space, which immediately opens up engineering the vacuum and using spacetime curvature engines to directly engineer matter for us.

open system: A system that communicates with its environment, and exchanges energy, matter, or both between system and environment. With the possible exception of a few theoretical or hypothetical systems, all systems in the universe are in fact open systems. An open system may be in equilibrium with its active environment, so that it cannot accept, store, and utilize any excess energy from the environment. Another open system may be in disequilibrium with its active environment, in which case it can accept, store, and utilize excess energy from the environment. When the operator inputs EM energy into an open equilibrium system, he automatically breaks its equilibrium condition with the external active environment, decreasing the system's entropy. This is an excited state, and the system will subsequently decay back to equilibrium (maximum entropy) condition by emitting or dissipating its excess energy. The reader interested in the thermodynamics should read Appendix A in this book, since the thermodynamics definition (or misdefinitions!) of "closed system" actually permits the system to be open with respect to energy exchange with its active environment — either in equilibrium or disequilibrium — and only prohibits exchange of mass across the system boundaries. By that definition, a "closed system" can be far from energy flow equilibrium with its external environment, and hence permitted to perform COP>1.0 or even COP = ∞. Hopefully our discussion in Appendix A will clarify the position we have taken on this matter.

optical switching: Switching of optical signals, or switching using optical frequency switching techniques, components, and circuits.

orthodox transistors: Ordinary transistors in use today, i.e., as contrasted with the Fogal semiconductor and its special functions.

orthogonal rotation: Simply speaking, a rotation from one direction to a direction at right angles to it. The Lorentz transform of special relativity, e.g., is a formula for the rotation of an object and all its parts away from its line of motion, as a function of the object's velocity, in a direction toward the time axis. The speed c just represents a full orthogonal turn, so that the projection intersection of the object with the observer 3-space, along the original line of motion, is zero.

outfolding: Translation by a system of received input longitudinal EM waves entering inside its potentials and fields, into conventional EM waves present in the system and then output from it. The opposite *of infolding.*

overpotential: Essentially the overpotential (as in an electrode in a solution) is a shift in the Fermi level necessary to allow the electron in the electrode metal to have energies overlapping with vacant acceptor levels in molecules adjacent to the electrode in the solution. The overpotential effect appears at small "double surfaces" of things in very close contact, particularly in electrochemistry and in solid state physics. The overpotential enables electron transfer, e.g., by tunneling. The overpotential increases with the log of the reaction rate that occurs at the overpotential location. For a small reaction rate, the overpotential is small; but for a small increase in the overpotential there occurs a dramatic increase of reaction rate. Fogal has utilized the overpotential effect in his semiconductor, to great advantage in accomplishing unusual charge blocking and other special effects. See **tunneling.**

overunity coefficient of performance (COP>1.0): More energy out or work out of a system than the operator or experimenter must furnish **and**

input. The excess energy is input from the active environment. See **coefficient of performance.**

parity: Deals with studying a system of a sequence of events as if the events were reflected in a mirror. A system has parity if it undergoes no fundamental operational change if replaced with its mirror-reflected twin.

partially time-reversed region: A region of spacetime in which an object temporarily experiences not only the ordinary forward flow of time, but also an extra induced backwards flow of time, usually a fraction of the time-forward time stream. It appears that the combination of the two is what constitutes time dilation in relativity. In a mixed time-forward (photon interactions) and time-reversed (antiphoton reactions) zone, if time-forward predominates, the law of attraction and repulsion of charges is partially reversed, leading to overall reduction of the forces — which remain in their usual directions. If the time reversal predominates, the law of attraction and repulsion is reversed, and in that region like charges attract and unlike charges repel. If time reversal and optical pumping are combined, the reversed law forces of attraction of like charges and repulsion of unlike charges are amplified also.

particle physics: The branch of physics using accelerators to study high-energy particle collisions, to determine properties of atomic nuclei and of the elementary particles themselves. It appears that conventional particle physicists do not recognize or consider time-energy interaction and time-reversal zone nuclear interactions at all. Hence the present "high (spatial) energy" physics is a far lower "total" energy physics than is the physics where time-energy is deliberately utilized. When time-energy is utilized, then the **total** energy of the photon varies inversely as a function of its frequency rather than directly as a function of its frequency. Lower frequency photons have very much greater total energy than higher frequency photons. For practical electrogravitation, therefore, one is interested in lower frequency bandwidths such as the ELF region.

Patterson effect: The intense multiscattering, multipass, multicollection of energy **density** flow in an assemblage of palladium-clad spherical bead capacitors, as the palladium adsorbs hydrogen ions from the fluid in which the beads are immersed, and thereby charges up as a capacitor, becoming an energy-flow generator, and nonlinearly increasing the energy-flow as the charge of the capacitors increases.

Patterson Power Cell®: Patented overunity power unit by James Patterson, which has been independently validated to yield COP>1.0. However, continuing difficulty has been experienced with the palladium metal used in the effect. See James Patterson, "System for Electrolysis of Liquid Electrolyte," U.S. Patent No. 5,372,688, issued Dec. 13, 1994. Patterson has several other patents in the area.

perpetual motion machine: The rather archaic and dogmatic assertion that no machine can indefinitely operate, since no closed system can input its output power back to itself while powering a load, because powering itself and its load violates the *equilibrium* thermodynamics (i.e., the thermodynamics of a closed system). Every system is already an open system in a violent energetic exchange with the active vacuum, so the pronouncement is a non sequitur assuming a false straw man — that every machine is a truly closed system or is in energy flow equilibrium with its active environment and can be treated as if it were a closed system. (The thermodynamic definition of "closed system" is misleading; it refers to a system where no mass exchange crosses the system boundary, but energy flow exchange can and does freely cross it. See Appendix A to this book for a proper discussion). Further, the equilibrium thermodynamics cited does not necessarily apply to either an open thermodynamic system far from energy equilibrium, or to a closed thermodynamic system in energy flow disequilibrium, as is now well known. After all, a common windmill will operate so long as there is a wind or until something breaks. But since it involves a mass flow across the system boundary, it is said to be a true open thermodynamic system. In electrodynamics, a closed system (with no matter exchange across its boundary) can yet have copious energy flow exchange across its boundary, and thus be far from EM energetic equilibrium with its external active environment. The charges and the dipoles in matter operate in a free "time-energy flow wind" and virtual particle flux wind in the vacuum, which are constant and unfailing, and thus charges and dipoles have been freely and continuously outputting real EM energy flow into 3-space for some 14 billion years. All assertions that overunity EM systems are perpetual motion nonsense are completely refuted by every charge and dipole in the universe. To repeat such gibberish only reveals the ignorance of the person making the improper assertion.

Ironically, those dogmatists fond of hurling perpetual motion labels at overunity researchers already recognize the overunity legitimacy of open systems far from thermodynamic equilibrium with their active environment; they refer to such systems *as false perpetual motion machines* or *fictitious perpetual motion machines.* All of them also accept the classical electrodynamics assertion of the source charge, from which all EM fields and potentials and their energy come. So the very dogmatists so confidently pontificating that perpetual motion machines are ludicrous, themselves already implicitly accept that every charge and dipole in the universe is precisely such a perpetual motion machine, freely and continuously creating energy right out of nothing. No legitimate free energy researcher in his wildest nightmares has proposed such perpetual motion machines on such a universal scale. The greatest perpetual motion nuts on earth are precisely those pundits so confidently castigating the notion that an electrical system can output more energy than the operator himself must input. We will have a modicum of tolerance and respect from them, when they can demonstrate how they can purge all those "perpetual motion charges and dipoles"-—and all their EM fields and potentials— from their own accepted electromagnetic model. Obviously if they do that, they simply destroy all EM energy in space, in systems, and in the entire universe. This is the absurdity of the position these pundits advocate so ignorantly.

phase conjugate: Of an entity: its time reversal. For waves: often referred to as *wavefront reversal.*

phase conjugate mirror: A nonlinear material which, when provided with an input wave, emits a phase conjugate replica (time reversed replica) of that input wave. The input wave is by practice called the *signal wave.* The time-reversed wave output is by practice called *the phase conjugate replica.*

phase conjugate optics: That portion of nonlinear optics dealing with multiwave mixing, formation of time-reversed waves, phase conjugate reflection (PCR), phase conjugate mirrors (PCMs), pumping PCMs to amplify their emitted PCRs, nonlinear interferometry, distortion correction, self-targeting, laser tracking, adaptive optics, and so on.

phase conjugate replica (PCR): Of a wave: the time-reversed duplicate of a signal (input) wave, after its detection by interaction with charge. We point out *there may still be some foundations difficulty, however, with the form in which the phase conjugate replica wave exists in spacetime prior to its interaction with charge and prior to its observation (observing the effect of its interaction with charged matter).* Indeed, there is very serious foundations difficulty with the very notion of an EM transverse wavefront in space, traveling at the speed of light, for no such thing exists. See Robert H. Romer, "Heat is not a noun," *Am. J. Phys.,* 69(2), Feb. 2001, p. 107-109. This is an editorial discussion by the Editor of AJP of the concept of heat in thermodynamics, where heat is not a substance, not a thermodynamic function of state, and should not be used as a noun. In endnote 24, p. 109, he also takes to task *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate. " "...perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

phase conjugating action: In nonlinear optics, the process or action of producing a time reversed replica of an input wave._For a particle, producing its antiparticle. For a living cell, dedifferentiating it back to some previous form, genetics and all. In the presence of a predominately time-reversal zone, however, the cell is in a predominately phase conjugating mode *a priori*. Phase conjugating it a second time converts it from dedifferentiating to redifferentiating mode. We hypothesize that these combinations are the actual mechanisms generating cellular dedifferentiation and redifferentiation in the cell, tissues, and body. A future book will address that issue; meanwhile, see the 119-page "porthole briefing" on the author's website, www.cheniere.org.

phase conjugation: In nonlinear optics, the novel nonlinear mixing of waves which generate an output wave — called the phase conjugate replica or time-reversed replica — that precisely retraces the path previously taken by the input wave that stimulated the action.

phonon: A quantum of the lattice vibration energy in a crystal. This mechanical vibration energy is considered a field, and the phonon is the quantum of that mechanical energy field. Lattice vibrations in a crystal thus are characterized by waves in the lattice, where the waves are comprised of ordered phonons. Interesting effects occur between electron currents and phonons, due to electron-phonon interaction.

photon: In physics, the basic action quantum. Conventionally considered to be the basic quantum of the EM field. There is a foundations problem with the latter concept, because of the dichotomous use of "field" for both the *effect* force field in 3-space after observation (detection) and the *causal* force-free field in 4-space prior to observation (detection). Strictly speaking, the photon is the quantum of the causal 4-space massless **field,** since the photon possesses both spatial EM energy and time increments. In curved spacetime, the photon may possess mass because the time flow is a mix of time-forward and time reversal, and thus cause and effect are slightly intermixed as well. We point out that, in a negatively curved spacetime, the photon mass is negative, the photon energy is negative, **and**

the photon fields are negative energy fields. Presently the negative energy photons do not appear to be utilized in physics except haphazardly if at all. This may represent a large class of photon interactions with matter that have been inadvertently omitted.

ping-pong: Colloquial term for iterative retroreflection.

pinned electrons: Electrons held stationary by a magnetic field, or other set of forces that react to "push them back into their positions and hold them fast" whenever they try to spatially move away.

pinning: The act of restricting (by restoring forces and fields) spatial movement (particularly of electrons) due to a disturbing magnetic field or other external force.

polaron: An electron in a substance, where the electron is trapped in a potential well due to polarization charges on surrounding molecules. The exction in a substance is analogous to an exciton in a semiconductor.

positron: An antielectron (time-reversed electron), possessing a positive charge and positive mass. *Mathematically* an empty *negative energy electron* state in the Dirac sea whose usual occupying electron has been removed, leaving a four dimensional "Dirac sea hole," may also be regarded as a positron for many purposes, at least *after observation resulting from interaction with matter*, so long as the remarkable difference of the Dirac 4-hole prior to observation from the Dirac 3-positron a**fter** observation is understood. Mathematically, the lattice hole previously occupied by an electron can be considered a positron-equivalent (i.e., as if observed), again so long as the real differences are kept in mind, between a Dirac hole *before* observation and what results after its observation by interaction with mass. This is important, because otherwise one is again substituting the effect for the cause. In addition, if the two are confused, **then** one also arbitrarily discards the negative energy fields and negative energy potentials of the source 4-hole in the vacuum prior to observation. That effectively discards anti-circuits and the ability to engineer practical electro antigravity.

power: In physics, the time rate at which work is done (at which *the form of energy* is changed). In electrical engineering, the term "power" is erroneously but ubiquitously also used to loosely mean "energy." Hence such odd and improper expressions as "draw power from the source," which is a non sequitur. Nonetheless, such non sequiturs are so universally utilized in electrical engineering that they have become a part of the literarure itself.

Prigogine, Ilya: Noted Belgian chemist and Nobelist, and one of the pioneers of the thermodynamics of open systems far from thermodynamic equilibrium. Prigogine was awarded the 1977 Nobel Prize for his work on the theory of dissipative structures in nonequilibrium thermodynamics.

Poynting, John Henry: An English physicist who lived from 1852-1914, whose contributions to electromagnetics included the energy flow theory now known as Poynting flow. He contributed the Poynting theorem, which states that the rate of EM energy loss in a specific region of space is equal to the sum of the dissipation rate (heat loss) flowing across the boundary of the region. Concurrently with Oliver Heaviside, Poynting conceived the notion that energy flowed through space, which previously had not appeared in physics. Heaviside actually published first, but obscurely; Poynting published in a prestigious journal, so the theory came to be known as "Poynting theory."

Poynting energy: A loose term (not rigorous!) used to loosely describe the flow through space of Poynting energy density S = ExH, which is that component of the total energy flow in space that will be *intercepted and collected* by the intercepting charges, circuit, etc. It actually describes not the energy flow density *per se*, but the component of it that would be intercepted by a unit point charge, or that actually enters an intercepting circuit. Poynting started with the assumption of the component of the energy flow that enters the circuit, *a priori.* Only the work of Heaviside considered the remaining component of the energy flow that is not intercepted by the charge or by the circuit. Because of the startlingly great magnitude of this remaining nondiverged Heaviside component, Lorentt integrated the entire (Heaviside) energy flow vector around a closed surface surrounding any volume element of interest. This discarded the entire nondiverged Heaviside component, and removed any necessity of considering or explaining how every generator and battery already pours from its terminals up to 10^{13} times as much EM energy as the presently accounted energy input to the shaft of the generator or the chemical energy available in the battery. We believe that the discarded and unaccounted Heaviside nondiverged energy flow component accompanying every field-particle interaction is responsible for the puzzling excess gravity holding the arms of the spiral galaxies together. In short, we believe it is the solution to the "dark matter" problem, now often believed to be a "dark energy" problem.

Poynting energy density flow: A somewhat more rigorous term used to describe the flow through space of the Poynting energy density flow

$\mathbf{S} = f(\mathbf{E}\times\mathbf{H})$ component that would be intercepted by a unit point charge or an intercepting circuit.

*Poynt*ing energy flow: A loose term (not completely rigorous) used to describe the flow of energy density $\mathbf{S} = f(\mathbf{E}\times\mathbf{H})$. Rigorously, both E and H are "defined" only after their reaction with a unit point static charge. Hence S is the resulting component of the overall energy flow that interacts with the charges (with the circuit) and is diverged into the circuit to power it. It is not the entire energy flow, which also includes a vast Heaviside nondiverged component that does not strike the charge and interact with it, and hence does not interact with the circuit but is just wasted. See Heaviside energy flow component and Poynting energy.

Poynting generator: Colloquial term for any dipole or potential difference that represents a broken symmetry in the vacuum's virtual particle flux, and hence extracts virtual energy from that asymmetry, and gates it out as a gushing energy flow in 3-space. When both electric field and magnetic field are involved, the component of the energy flow that would be intercepted and diverged by a unit point charge is the Poynting energy flow S=f(ExH).

Poynting vector: The vector S, given as $\mathbf{S} = f(\mathbf{E}\times\mathbf{H})$, used to represent the Poynting (intercepted and collected by a unit point coulomb) energy density flow component. In Heavisidc's version, $\mathbf{S} = \mathbf{E}\times\mathbf{H} + \mathbf{G}$, where G is an energy circulation flow term of trapped EM energy flow. Heaviside later rea**lized** that such a closed energy circulation flow exhibits substantial gravi**tational** effects; in modern terms we would say that it represents an apprecia**ble** local curvature of spacetime. Unpublished notes on the gravitational aspects were found beneath a floorboard in his little garret apartment, years after his death. We also point out one limitation of the present use of the Poynting vector: It has rather universally been associa**ted** only with the flow of *positive* EM energy, and not with the flow of negat*ive* energy. Indeed, the concept of negative EM energy is still consufe**d,** with different views and positions, and still in debate without complete scientific consensus.

PPCM: Pumped phase conjugate mirror. A phase conjugate mirror material that has introduced to it appositive or "pump" waves (in other words, s***tress*** waves). In effect, pumping is "stressing" or "squeezing" ryhtmically. It is also varying a stress potential, which emits longitudinal EM waves, each accompanied by a correlated scalar (time-polarized) wave. In the phase conjugating process, up to all the energy in the pump waves can be output as the energy in the emitted phase conjugate replica wave. Thus the PPCM may be viewed as an *amplifier,* very similar to a triode. When the time-components of the photons in the signal wave input and also in the pumping waves are considered, then there is a functioning of the pumping also ongoing in the time-domain. There is a rhythmic variation of stress in the time-domain also, which creates the correlated time-polarized EM wave accompanying the longitudinal wave from the 3-space pumping.

So the time-reversal aspect (formation of the phase conjugate replica wave) may be thought of as a rhythmic time-stress wave (pumping or stressing in the time domain) which, by alternately decreasing and increasing the time potential of the mirror, produces two temporal force effects. During the decreases in temporal potential of the mirror, the time-force from the input wave and from the pump waves is forward into the mirror material, allowing the input and interaction of these waves. During the increase in temporal potential of the mirror, a reversed temporal force from the mirror back to the past (temporal potential earlier in 4-space) of the input "signal" wave. The pumping energy (excess temporal potential energy) adds the gain to the phase conjugate replica wave,

probability: The probability of a simple event is the ratio of the number of times it occurs to the total number of trials (for a large, essentially infinite number of trials). Note that "trials" constitute things that have occurred (are past, have been observed). Probability has resisted rigorous logical definition by Aristotelian logic; the reason is that it is an expression of the fourth law, the identity of opposites. Probability is an expression of the future (that which has not occurred) in terms of the past (that which has occurred). Consider *throwing a die to land with one face up.* Thinking of the event (as if it has just occurred, in terms of the most recent past), one can conceive six ways of looking at it. In other words, one can conceive of six "most recent pasts", however, these have not occurred. We were merely thinking of them as if they had occurred. By the fourth law, the *total* "most recent pasts that have not occurred" is identical to the "most immediate future" that has not occurred, on the common boundary between past and future. If one therefore *collects* all six "most recent **pasts** not occurred" they turn into the most immediate future, by the fourth law of logic. One may argue that only a single one of the events will actually occur. Here one counters with the observation that, when it occurs, it is the past, not the future, and then it has not been collected so as to move **to t**he boundary. In theory, the hidden variable approach makes it possible to directly engineer those probabilities while the Schrodinger wavefunction is still propagating and has not yet collapsed to constitute an "observation."

I.e., it is theoretically possible to engineer Wheeler's "pre-reality," and determine in advance just which of the outcomes will occur.

Puthoff, Harold: Noted physicist active in a variety of fields and investigations including in psychoenergetics, extended electrodynamics, energy from the vacuum, inertia, antigravity, parapsychology, remote viewing, psychokinesis, zero-point energy of the vacuum, and cosmology. We are particularly impressed with Puthoff s cosmological feedback principle. Together with Cole, Puthoff also proved that thermodynamics does not prohibit extracting usable EM energy from the active vacuum. See Daniel C. Cole and Harold E. Puthoff, "Extracting Energy and Heat from the Vacuum," *Phys. Rev. E,* 48(2), Aug. 1993, p. 1562-1565.

QM vacuum: The quantum mechanical vacuum, which models the vacuum as an intense flux of virtual particles, each appearing and disappearing in such extremely short time intervals as to be unobservable individually—although as physics develops, from time-to-time what is observable changes because of increased observation techniques. Violent fluctuations of extremely short duration constantly occur in the QM vacuum flux. Various calculations estimate the energy density (in mass units) as up to 10^{80}-10^{90} grams per cubic centimeter (in mass density units), or even greater. The vacuum is sometimes described as two separate components: (1) the ambient or average vacuum in deep space, which can be taken as a scalar potential having an enormous energy density, and (2) the fluctuations superposed upon the ambient vacuum as local violent potential energy changes. Finally, we note that any EM potential in a system is *apriori* a change to the local vacuum scalar potential.

quantized: Adjective which means (i) reduction of some observable quantity to multiples of some small, indivisible unit, or, (ii) expressed in terms of quantum theory.

quantum: A single "particle" of the quantity *action* (angular momentum). Often confused as a single quantity of energy alone.

quantum electrodynamics: The theory of photons and electrically charged particles and their interactions. The use of the term "quantum" implies that the EM radiation's discrete photon nature is significant, so that quantum theory must be employed.

quantum field theory: A quantum mechanical theory in which "... a phsyica1 held is considered as a collection of particles and forces, and observable properties of an interacting system are expressed as finite quantities rather than state vectors." *[Dictionary of Science and*

Technology.] More simply (after Davies, *The New Physics,* 1989): "The theory that describes the quantum effects of a classical system of fields defined on space-time and satisfying various partial differential equations." Particles are taken as quanta of fields similar to photons being quanta of the electromagnetic field. Each type of particle is represented by an operator that obeys certain commutation laws. The theory relates gluon fields and intermediate vector boson fields to the strong and weak interactions, and allows an overall theory of fundamental particles and their interactions.

quantum potential: A special potential added to the Schrodinger equation by Bohm in his hidden variable theory of quantum mechanics. The quantum potential is a multiply connected entity, so that it "occupies" widely separated but multiply connected points, events, or objects. It is also an extraordinary energy amplifier, since any energy input to one of the multiply connected points simultaneously and instantly appears in every other connected point, regardless of distance or location in the universe. In real life, the quantum potential also has a "coefficient of multiple connectivity", so that only a fraction of the energy input to one multiply connected point will appear in the other points of the multiple connection group. The quantum potential has been strongly weaponized by five nations, and quantum potential weapons are the dominant weapons on earth, being more powerful than nuclear weapons. In theory the quantum potential together with engines and antiengines could be used to treat and cure a given disease in all persons on earth simultaneously, or in one nation or area simultaneously. Sadly, engines have been developed to generate diseases in a targeted populace, rather than engines to heal the populace. Russia and Brazil have had the quantum potential weapon for some time, as have two nations friendly to the United States. In 2001, China also deployed the quantum potential weapon.

quark: Fundamental particle theorized to comprise hadrons (particles such as protons, neutrons, and mesons) by combining in twos or threes. There are six varieties of quarks. The difference between a proton and a neutron, e.g., is merely the orientation of a single quark. In a time reversal zone (TRZ) in the electrolyte in cold fusion experiments, the gluon forces can be significantly reduced, so that the quarks are nearly freed. Two protons (H+ ions) can attract together in a TRZ so closely that each enters the reduced strong force region of the other, forming a quasi-nucleus. As the ions in solution outside the TRZ move to cancel the TRZ, it decays and the quasi-nucleus becomes an increasingly excited state, seeking to decay. Preferred method of decay is by quark flipping. So one quark flips, so that the quasi-nucleus has become a quasi D+ nucleus. As the TRZ vanishes and time-forward resumes, the strong force recovers faster than the Coulomb barrier returns. Consequently, the quasi-nucleus of D+ merely tightens into a real nucleus of D+. This is one of the standard transmutation reactions in cold fusion experiments, forming excess deuterium.

quaternion: Expression comprised of the sum of four terms, one of which is real and three of which contain imaginary units, and where the terms can be writtten as the sum of a scalar and a three-dimensional vector.

quaternion algebra: The algebra of quaternions and their mathematical operations. Quaternion algebra is of higher topology than either vector or tensor **algebra.** Maxwell's original 1865 equations are some 20 equations in 20 unknowns, in quaternion and quaternion-like algebra. Heaviside and others reduced the algebra to vector algebra and some four equations; at his death, Maxwell himself was rewriting the equations in his *Treatise* to purge the quaternions, because of the controversy over the difficult quaternion form. The present vector equations taught in university as "Maxwell's equations" are in fact Heaviside's truncation of the Maxwell theory. Going to tensor algebra electrodynamics does not recover the full range of Maxwell's original 1865 quaternion EM theory.

quintessence: Literally, the "fifth essence". In short, *dark energy*. We have nominated the arbitrarily discarded, vast Heaviside nondiverged energy flow component surrounding every interaction of charge with field as the dark energy that is the source for the excess positive gravity holding the arms of the spiral galaxies in place. See dark energy, positive.

reactive power: Electrical power, measured in volts-amps-reactive (vars) that cannot do work as is. For sinusoidal waveforms, the formula for reactive power is $IV\sin\theta$, where θ is the phase angle between the voltage **and the current.**

real hidden dynamics: The dynamics of vacuum engines, or of "internal energy" or "internal work" ongoing in a system or volume of space. These internally organized forces, fields, and energies involve dynamic curvatures of spacetime, and it is this dynamics of spacetime itself that provides **the** continuous work and dynamics of internal energy. See Einstein's second postulate extended; infolded; dimensioning.

real hidden vectors: The "hidden" or "infolded" real vector components, e.g., of a zero vector summation system. The individual hidden vector cannot spatially translate the entire bulk system, but it can interact inside the system to perform *internal work* upon it or upon some particular portion of it. Such a system of a finite number of infolded deterministic vectors is said to constitute a "dimensioned" stress potential. Specifically dimensioned potentials, fields and waves thus can be constructed, so that they carry specific vacuum engines (spacetime curvature engines). The interactions of these dimensioned EM entities may differ from normal EM interactions in chemistry, physics, biological systems in general, cells in particular, etc. The long neglected combination of Whittaker's 1903 decomposition of the scalar potential, together with his 1904 demonstration that any EM field or wave pattern decomposes into two scalar potential functions, has neglected the development and use of the "internal EM" theory and technology for a century. One or more nations has, however, highly weaponized "dimensioned EM fields, waves, and potentials" as evidenced by the peculiar and anomalous results of the decades-long microwave radiation of the U.S. Embassy in Moscow. See dimensioning.

real power (active power): Power, measured in watts, that can do work, including generating heat. For sinusoidal waves, real power is given by the formula $I \cdot V \cdot \cos\theta$, where θ is the phase angle between the voltage and the current. Note that when higher group symmetry EM effects are involved (such as the Aharonov-Bohm effect), then the "normal" behavior of circuit components can be altered dramatically. As an example, in the MEG the secondary (output) coils of the transformer section are caused to exhibit a purely electrical interaction with the electrons and electron currents in the coils themselves. Consequently *the output current and voltage are in phase, in complete violation of conventional transformer theory and conventional assumptions about the interaction of inductors.* So the output of the MEG is real power, and phase angle does not enter into it.

regauging: In electrodynamics, changing the scalar potential or the vector potential, or both. If one or both of the potentials is/are changed so that a net new force is created in the system, that is an *asymmetrical* regauging. If only one potential is changed, a net force results and that is an asymmetrical regauging. If both potentials are changed but carefully selected so that the free force resulting from one change is equal and opposite to the free force resulting from the other, that is a net *symmetrical* regauging, often called "Lorentz regauging" or the "Lorentz condition". We note that longitudinal EM waves are produced very simply by rhythmically oscillating the magnitude of a stress potential, which is naught but coherently oscillating the individual vectors of a set of EM fields that sum to a zero vector resultant. The locked-in stress in T. Henry

Moray's semiconductors used in his radiant energy device tubes and the locked-in stress in the Chung negative resistor thus have to do with the production and use of longitudinal EM waves, time-polarized EM waves, and stress potentials.

reflection: The rebound of light or other energy from a surface.

refraction: The bending of light as it passes from one medium to another or through a density change in a given medium.

retarted wave: An electromagnetic wave that is derived from retarded fields, which in turn are derived from retarded potentials. The retarded wave at a point or region of space has arrived there or exists there as the result of the dynamics of charges and currents that existed at earlier times elsewhere. Compare to advanced electromagnetic wave.

retroreflecting: Precisely reversing the path of a reflected wave or object.

Sachs, Mendel: Noted scientist who has completed the work of Einstein and developed a unified field theory covering the universe from the tiniest part to the entire conglomerate. Since O(3) electrodynamics has been shown by Evans to be an important subset of Sachs' theory, then for the first time we have a higher group symmetry electrodynamics and a unified field theory that is largely engineerable by novel electrodynamic means, using the O(3) model.

Sachs' unified field theory: An extension and completion of Einstein's work that includes the universe and its dynamics from the tiniest part to the entire conglomerate, and includes quantum mechanics, electromagnetics, gravitation, and the strong and weak forces. See particularly Mendel Sachs, *General Relativity and Matter,* Reidel, 1982; *Quantum Mechanics from General Relativity,* Reidel, 1986. Sachs' unified theory provides a great generalization of general relativity, quantum mechanics, and electrodynamics reaching from the quarks and gluons to the entire universe. Much of the model is also engineerable by higher group symmetry electrodynamics such as the O(3) electrodynamics pioneered by Evans and Vigier.

scalar: Characterized by magnitude only. However, with respect to polarization, a scalar photon is a term in use for a *time-polarized* photon, where the EM energy oscillates along the time-axis. Hence a scalar wave, following the same terminology, is an EM wave composed of scalar photons, hence whose EM energy is oscillating along the time-axis. The term "scalar" is used in the sense that no vector in 3-space exists even though a vector (and a variation of its magnitude) exists along the time axis. The effects are observed as an oscillation in the rate of flow of time, hence a "time-density" oscillation. The term "scalar" with respect to polarization implies only that there is no vector component in 3-space.

scalar entity: In ordinary vector analysis, an entity thought to be completely characterized by magnitude alone. Actually this is incomplete; a scalar entity must contain a time-like vector component, if it is to steadily persist in time. What we usually do unthinkingly is assume a continual series of successive observations or *d/dt* operations upon the *kt* nature of a scalar *k*. We conceive these individually frozen 3-space snapshots to "continuously exist" since we can readily recall the "stream" of *d/dt* differentiations in our mind and memory. Since all observation is 3-spatial, then no observable *persists continuously* in time; instead, it *continually recurs* in time, instant to instant. This "iterative successive stream of frozen 3-space snapshots" exists in the continual and rapid recall from our memories. Compare this to the similarity of observing a sufficiently rapid series of still pictures on a screen, where the individual pictures differ by a time sequence taken of some phenomenon. One "sees" the recreation in one's mind of the flow through time of those events and entities so photographed. Yet at any one instant, there is only a single frozen snapshot on the screen, and no motion at all was "physically" observed.

scalar electromagnetics: Colloquial term for the electrodynamics that arises from considering transverse EM waves, longitudinal EM waves, time-polarized EM waves, electrogravitation, superluminal EM signals, interferometry, nonlinear optical functions, time-as-energy, and the infolded electrodynamics inside all usual EM fields, waves, and potentials, In secret superweapon projects, the Russian term for scalar electrodynamics is *energetics.* It is divided into three major fields, depending upon the intended target. Targeted against inert matter, systems, and fields, the same term is used: *energetics.* Targeted against the physical and chemical processes and biological fields and potentials of living organisms, the term used is *bioenergetics.* Targeted against the mind's time-polarized EM functioning and dynamics, the term *psychoenergetics is* used. In the latter, higher group symmetry electrodynamics and extensions to quantum field theory play significant roles. Significant progress and strategic testing of all three branches of energetics has been evidenced worldwide. Progress in the energetics branch was confirmed by a statement by U.S. Secretary of Defense Cohen in 1997. Quoting: *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of* *electromagnetic waves...So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts "* Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

scalar potential: Usually considered synonymous with the electric potential. The concept of potential was developed by Poisson, Green, and others from about 1813 to 1827. Three notions involved in the development of the concept of the potential are: (i) there is the collection of storage of energy in the system, via the collection or storage of some entity therein, and this "potential energy" can be released and dynamically used, (ii) the potential is a scalar function whose spatial rate of change is a vector force, and (iii) the dynamics of the stored energy is suspended by the "storage" or "collecting" entity; in other words it is a sort of "suspended action" waiting to be unleashed. *It requires some action—such as interception and divergence—to release the leashed energy.*

The second notion is false, since no force exists in the absence of mass because mass m is a *component* of force F via $F \equiv dp/dt = (m)dv/dt + (v)dm/dt$. It is corrected by stating that its spatial rate of change *when interacted with charged matter* is a vector force. For the third notion, we note that the concept of a "virtual" particle or particle flux means that it cannot become observable unless some interaction is imposed. So we may handle the "third notion" requirement by assuming the potential in modern terms to be composed of hidden virtual particle flux, or hidden wave flux, in wave-to-wave interaction. Physics tries to define the electrical potential as energy per coulomb, or W/Q. That equation, however, is actually not a definition at all (no equation is a definition), but rather is an algorithm for calculating the magnitude of the excess energy collection/collecting upon a coulomb of charge, when that coulomb is immersed in the potential It in fact is the *reaction cross section* of the potential, not the potential entity itself. This now begins to give us a clue, Potential is *energy* in some fashion, and the term "potential energy" is a redundancy. So the entity must have essentially the same definition as energy! Potential energy, collected energy, stored energy, whatever—it is energy first and foremost. How it is collected and what it is collected upon or in, determines the "type" of potential. It also cannot be logically defined as the work expended by it upon a unit charge. The causal entity itself is not the action taken; just as a man is not definable as "how far he has run" in some race.

scattering processes: Collision processes in which an incident particle such as a photon, electron, or proton, of known energy or direction or both, collides with another entity, and emerges from the collision with a different energy or direction or both.

Schrodinger, Erwin: Noted Austrian physicist, 1887-1961, and Nobelist who discovered the wave equation named after him as the Schrodinger equation, and performed other vital work in wave mechanics.

Schrodinger equation: In quantum mechanics, an equation that describes the propagation of waves associated with subatomic particles, or more generally, that describes the evolution over time of the quantum state of a system. The fundamental quantum mechanical equation governing the behavior of wave functions.

self-energy: In a quantum mechanical system, the energy associated with the emission and absorption of virtual particles. In a classical system, the energy associated with the interaction among parts of the system itself.

self-organization: (After Davies, *The New Physics,* 1989): "Spontaneous emergence of order, arising when certain parameters built in a system reach critical values." Simply put, self-organization is the spontaneous formation of patterns of localized order in a system that begins in a more homogeneous state. In short, the system moves from a simpler to a more complex state. Self-organization is one of the permissible features of an open system far from thermodynamic equilibrium with its active environment. Note that the localized orders represent free potential energy formations (free self-regauging) that in theory can then be dissipated in loads to perform free work. Hence the capability of a self-organizing system to perform some free work, or in other words to exhibit COP>1.0.

self-potential (e.g., of the electron): The potential created in and of the charge (e.g., the electron) by its virtual particle flux exchange with the vacuum. Also, an asymmetry in the virtual photon flux of the vacuum, caused by the interaction of the charge (e.g., the electron) with the vacuum flux. From here, we note that the self-potential must consist of real observable 3-space EM energy flow or flows from the electron, because of the asymmetry. However, the charge is not the primary source *per se* **of** the energy flows that establish the self-potential, but instead is the asymmetric gate in the vacuum flux that "gates out" the potential and its energy **flux** The asymmetry in the vacuum flux is the actual *source* or gate. Indeed, in classical terms the energy is received by the charge in the form of converging longitudinal EM energy waves from the complex plane (from the time domain). The spin of the charges of the dipole transduces the absorbed complex EM energy into real observable EM longitudinal wave energy, and pours it out in all directions in 3-space. See discussions under Whittaker and phase conjugate.

self-targeting: Between two entities, where each acts partially or wholly as a phase conjugate mirror, self-targeting is the iterative phase conjugation or "ping-ponging" and rapid convergence of beams between the two PCMs.

S-flow: The Poynting energy density flow S, where S = ExH. Heaviside's equation is given by S = ExH + G. In the latter equation, E is the electric field, H is the magnetic field strength, and G is a trapped closed loop flow of energy density. Note that, from the very definitions of E and H, the Poynting, energy flow represents only that portion of the total energy flow in a spatial region that is diverted by interception and interaction with charges. Heaviside's additional component G, which is not part of the field interactions but a result of the huge noninteracting part of the energy flow outside the interacted ExH component, does appear to be responsible for an enormous unaccounted EM energy flow component that generates the excess gravity required to hold together the arms of the spiral galaxies.

signal wave: In nonlinear optics, the signal (wave) input to a phase conjugate mirror (PCM), that stimulates the PCM to emit a phase conjugate replica (PCR) wave in return. See discussion under distortion correction theorem.

Slepian current: The vector $j\phi$, which in a current loop represents the energy unit area per coulomb of flowing charge, per unit of time, that is collected/collecting from the Poynting S-flow and dissipated in that current loop in its loads and losses.

Slepian flow: "Flow *of collected-energy* density's dissipation" in a current loop, and therefore related to the vector $j\phi$ in the current loop, which represents the energy per unit area per coulomb of flowing charge, per unit of time, that is being dissipated in that current loop in its loads and losses.

Slepian vector: Refers to the vector $j\phi$ in a current loop, which represents the energy density per coulomb of flowing charge, per unit of time, that is being dissipated in that current loop in its loads and losses.

solid-state parametric oscillator: A parametric oscillator circuit employing solid-state devices. A parametric oscillator is a device that oscillates by rhythmically changing a parameter (such as capacitance or inductance, etc.). Simple switching between differing values of the parameter is often used to induce the periodic change of the parameter. One way of close-looping a COP>1.0 EM system is by using a bit of the positive output energy to do parametric switching, in effect converting the system to a self-oscillating energy flow system. The AC flow of EM energy is then much more easily intercepted, collected, and dissipated to power loads without killing the source dipole extracting and furnishing the free flow of EM energy from the active vacuum.

soliton: A "solitary wave" consisting of a wave pulse which propagates without changing its shape and without dispersion.

source dipole: Any dipole, considered from the viewpoint that real observable EM energy continuously pours out of it without any observable EM energy input. There is no true "source" as such in nature, since the energy is not created but only transduced and gated. In the case of the dipole and electrical charge as sources of EM energy, fields, and potentials, the broken 3-symmetry of the dipole initiates an automatic and more primary 4-symmetry, where EM longitudinal wave energy flows into the negative charge of the dipole from the time domain (complex plane), is transduced into normal EM longitudinal wave energy in 3-space, wherein the 3-space energy is absorbed by the positive charge of the dipole, transduced back into time-energy and reradiated back into the time-domain (fourth Minkowski axis). We roughly state that the dipole "emits observable EM energy" into all directions in 3-space as observable EM energy. However, what actually happens is that this *energy 4-circulation process* is what is emitted into and along 3-space, along every radial direction, with the process occurring at every spatial point along that direction. Any charge (taken together with its clustering virtual charges of opposite sign as a dipole) then inserted at any point in space along a radial line, will dipolarly intercept the ongoing process. We will observe the interception by its 3-effect on the intercepting and interacting 3-space charge, but we will not observe the 4-process itself. This deeper examination of the EM field-charge interaction requires a new interpretation of what is meant by "propagation of EM energy through 3-space", which is a non sequitur. Recall again Romer's acid objection **to the** totally illogical picture in every elementary textbook of the so-called "EM wave through 3-space" as a flat x-y plane orthogonal to the line of **motion** z, and where the E-field vector is oscillating along one axis while the magnetic field B is oscillating along the other. That representation is totally false, just as pointed out so strongly by Romer.

spacetime or space-time: The entity represented by the "fused product" of space and time, to give "space x time". Spacetime is also dynamic geometry, in the general relativistic interpretation. We may also consider spacetime to be a giant potential, and also active geometry, and also active vacuum, and also virtual particle flux of the vacuum, and also hidden EM wave flux. If we wish, we can also consider spacetime to be comprised of longitudinal EM waves and impressed dynamics.

spacetime curvature engine: See deterministic pattern and dimensioning.

special relativity: Einstein's theory of inertial frames in which, if two systems are moving uniformly in relation to each other, one cannot determine anything about their motion except that it is relative. Each of the two frames is said to be "rotated" with respect to the other, but not accelerating. The velocity of light in space (the vacuum) is constant, and is independent of the velocity of its source and the velocity of an observer. All the laws of physics are the same in all inertial frames of reference.

spin: The intrinsic angular momentum of a particle such as an electron, proton, neutron, photon, graviton etc., even when at rest, as if it were a top spinning about an axis, but had to spin 720 degrees before it turned "full circle." Spin is quantized, and is always described as a half or whole spin, e.g., - 1, - 1/2, 0, 1/2, 1, etc. A spinning charged particle such as an electron thus demonstrates a magnetic moment, due to the circulation of charge in the spinning. In the nucleus of an atom, the spin of the nucleus is the resultant of the spins of the nucleons (particles comprising the nucleus).

spiral galaxy: A member of the class of galaxies having arms that extend in the form of spirals. About 70% of all galaxies are spiral galaxies: only about 10% of the gravity necessary to hold the spiral arms together can be accounted for by known gravitational sources. This has led to the theorizing of dark (hidden) matter of exotic and unknown form, which would be responsible for the excess gravity. Later it was theorized that some sort of "dark energy" might be responsible for the extra gravity. The present author has proposed that the *unaccounted* enormous Heaviside component of the EM energy flow surrounding every field/charge interaction, and not entering into that interaction, is the missing source of the excess gravity. The dark Heaviside energy flow was arbitrarily discarded by Lorentz in the 1890s, and it is still discarded today by all electrodynamicists and astrophysicists.

Stoney, George Johnstone: Well-known scientist (1826-1911) who first estimated the value of the electronic charge (1874) and was the first to advocate the particle nature of electricity. In 1891, he named the unit of electronic charge the *electron.* In 1896-7 he laid the foundation for Whittaker's 1903 bidirectional wave decomposition of the scalar potential.

string wave or "plucked taut string" wave: The transverse wave on a taut string produced when the string is plucked. The string wave itself stays on the string and does not travel through the slapped external medium.

subspace: In vector mathematics: a subset of a vector space which is closed under vector addition and scalar multiplication operations. In the new approach, we may consider the active ambient vacuum as a giant scalar potential. This ambient potential then decomposes into longitudinal EM waves via Whittaker's 1903 decomposition. One can then refer to longitudinal EM waves deliberately transmitted "inside" that Whittaker "subspace" as "subspace communication". The Fogal semiconductor is able to infold normal EM signals into such subspace signals (longitudinal EM wave signals) in the transmitter, and to outfold the longitudinal EM waves back into normal EM waves in the receiver. Hence the Fogal semiconductor can be utilized to perform subspace and superluminal communication. These capabilities have been demonstrated by Fogal in the laboratory. As this is being written, Fogal is in serious negotiations with several large companies for the funding to get his chips into volume production and onto the marketplace in very sophisticated and advanced equipment representing a great leap forward in technology. There are, however, powerful interests who do not wish the Fogal semiconductor to ever be produced and marketed. We have worked with Fogal for some years, and — once the semiconductors are available in quantity — we will be developing and marketing COP>1.0 electrical power systems utilizing his semiconductors and technology, under license from Fogal.

superconductivity: Property of some materials or metals whereby at very low temperature their resistance reduces to zero. These materials also exhibit many additional properties that are anomalous compared to normal materials. Much research has gone into the attempt to find materials that exhibit superconductivity at higher temperatures, with the goal of reaching room-temperature superconductivity. Room-temperature superconductivity is also just COP = 1.0. EM systems far from equilibrium with the active vacuum, and which exhibit COP>1.0 at room temperature, may involve most of the odd entities also observed in superconducting materials (polarons, lattice holes and hole currents, Dirac sea holes and Dirac sea hole currents, "charging" by Dirac sea hole excitation, vortices, etc.) The energy density in the output section is greater than the energy density input by the operator to the input section. Considering the giant negentropy mechanism and the dipolarity of the input and output sections, the tempic potential of the output section of a COP>1.0 EM system is greater than the tempic potential of the input section. Hence there is a tempic gradient across the entire system, and a tempic force from output to input. This may "sweep" out a great deal of the produced "superconductivity" entities, and particularly may result in a Dirac sea hole current in the vacuum, out of the input section back into the external power source, as a novel output of the COP>1.0 system. Feedback from output to input may also be in the time domain, corresponding to what Gabriel Kron referred to as the "open path". By recombinations and other means, the feedback in the tempic domain may be transduced in the input section of the device to ordinary current in forward time. When the voltage and current in this transduced tempic feedback are properly adjusted to that being supplied from the external power source, the external source can be disconnected and the COP>1.0 system becomes self-powering, while simultaneously powering its load also.

superluminal: Exceeding c, the speed of light in vacuum. Longitudinal EM waves can travel either slower or faster than the speed of transverse wave light in vacuum. See W. A. Rodrigues and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds 0 < v<oo in nature," *Found. Phys.*, 27(3), 1997, p. 435-508; W. A. Rodrigues and J. Vaz Jr., "Subluminal and Superluminal Solutions in Vacuum of the MaxweII Equations and the Massless Dirac Equation," *Adv. Appl. Clifford Algebras*, Vol. 7(S), 1997, p. 457-466.

superluminal communication: The movement of intelligent signals faster than the speed of light in vacuum, and the transmission and reception of such signals and extraction of their transported information. Longitudinal EM waves are not limited to speed c, whereas transverse EM waves are. The infolded electrodynamics inside all EM fields, potentials, and waves consists of longitudinal EM phase conjugate wavepairs, with their impressed dynamics. Hence this "subspace" is a superhighway for such waves, including those moving at superluminal velocity. In essence this is a very special kind of "tunneling" which might be referred to as *subspace tunnelling.* The limitation to light speed rigorously applies to the modeled transverse EM waves which are a sort of "bulk upheaval" waves similar to waves on the surface of the ocean. As in the ocean wave analogy, "pressure" waves underneath the ocean may travel very much faster than the bulk upheaval waves on the surface.

When making longitudinal EM waves, the process is imperfect and the resulting longitudinal EM waves will have a transverse residue as well. The resulting "transverse-noisy" LW is called an *undistorted progressive wave*. Undistorted progressive waves may move slower than the speed of light or faster than the speed of light. The degree of transparency of matter to LWs is an inverse function of the "transverse-noisiness" of the LWs.

superposition: The simple linear addition and subtraction of two or more values, states, etc.; one of the key principles in field theories and in the concept of potentials. When the situation is sufficiently nonlinear, however, interaction of waves and potentials occurs instead of simple superposition.

supersystem: A higher level or "metasystem" of interactions between the classical system, the active vacuum, and the active local curvatures of spacetime. Hence the supersystem consists of three interacting components: (1) the EM system and its dynamics, considered in conventional fashion as if the local vacuum were inactive and the local spacetime were flat; (2) the local interactive vacuum and its dynamics, and (3) the local curvatures of spacetime and their dynamics. In component (2), one must consider the Dirac sea, Dirac holes and hole currents, with concomitant potentials and forces impressed on them. In component (3), the time-domain interactions (time-energy, time-force, time-potential, and currents of energy along the fourth Minkowski axis) must be accounted to include negative energy.

All three components of the supersystem interact with each other. In general, the vacuum interactions may be taken from particle physics, and the spacetime curvature reactions may be taken from general relativity. However, this will yield only a rough or "first order" analysis, since it does not produce the tri-reactions of all three components at once, but mostly reactions between pairs of the components. The concept is easy to conceptualize, however, and yields valuable clues and insight in many areas, such as chemistry (including biochemistry), electrical physics, electrodynamics, thermodynamics, and electrogravitation. It is particularly useful in understanding the phenomena encountered in COP>1.0 EM power systems.

symmetrical regauging: In electromagnetics, changing the scalar potential to a different value, and simultaneously changing the vector potential to a different value, with the two changes precisely selected so that the two extra forces (each created by one of the changes) are equal and opposite — thus canceling each other's effect with respect to bulk translation and doing no net *external* work. The two forces remain and can perform internal work upon the system. Note that the stressing effect of the two new forces is not canceled, and that the stress of the system is altered a priori. See regauging.

tantalum capacitor: Electrolytic capacitor with a tantalum or sintered-slug anode, such as a solid tantalum, tantalum-foil electrolytic, or tantalum-slug capacitor.

targeted pinpoints: Those separated points which engage in self-targeting between themselves and some separated signal energy source (as in Fogal's semiconductor). The effect is to sharply narrow the retroreflected source signal energy from the source onto the separated points by means of the mutual ping-ponging. In this fashion much more energy density from the source can be concentrated on the targeted pinpoints. In effect, *phase conjugate shooting* — the return of the phase conjugate reflected scattered energy directly to the original target emitting the signal wave in pinpoint fashion - is evoked. See distortion correction theorem.

tensor: A special mathematical entity which is a generalized vector element of a vector space, with more than three components, each of which is a function of the coordinates of an arbitrary point in space of an arbitrary number of dimensions.

Tesla, Nikola: Famous American emigre electrical engineer, inventor and eccentric (1856-1943), whose numerous inventions fueled much of the electrical 20th century. In essence he produced the AC system, including polyphase systems. Also, Tesla invented radio transmission, not Marconi, according to a U.S. Supreme Court decision that upheld the prior patents of Tesla over Marconi. In his famous experiment, Marconi in fact utilized a variation of one of Tesla's circuits. This does not denigrate Marconi's many accomplishments, but just sets the record straight on who invented radio transmission. The impetus to the development of commercial radio transmission technology and the industry itself came from Marconi.

thermodynamic equilibrium: Stable state eventually reached by an isolated system or one in a well-behaved active environment. One may consider it a "balanced" or "preferred" state. It is important to note that, in particle physics, there can be no equilibrium of a mass system unless the vacuum interchange with that system is considered. Electrodynamics mistakenly continues to omit the active vacuum interaction with a Maxwellian system, hence is seriously flawed in that respect.

thermodynamics of open systems: Systems far from thermodynamic equilibrium with an active environment are permitted to exhibit five "magic" functions. Such a system can (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator himself inputs (the excess energy being freely received from the active environment, (4) power itself and its load simultaneously (all the energy being freely received from the active environment), and (5) exhibit negentropy.

time: In our view, time is spatial EM energy compacted by the factor c . Time thus has the same energy density as mass, but in the time domain rather than in the 3-space domain. Also, time is not observable *a priori;* mass is an observable. Time persists since it is persistence itself. No observable persists *a priori,* since observation is a *d/dt* operator imposed upon LLLT spacetime, yielding a frozen LLL snapshot which does not even exist the very next instant. When we say "mass exists in time" (or any observable continuously persists), we are unwittingly invoking a very rapid iterative *d/dt* operator repeatedly, and recalling these iterative sequential "frozen snapshots" or observations from memory. One may produce and utilize EM energy flow in the time domain, as in 3-space. Time, being energy and energetic, has tempic potential (energy density), tempic potential gradients, and therefore has "tempic force." The photon is composed of both spatial energy and time-energy as its two canonical constituents. The lower the frequency, the higher the energy of the photon when the time-energy is considered, since the magnitude of the time-component of the photon increases as the spatial energy component decreases, and the energy of the photon in that case increases as c^2 times the increase in time magnitude, while decreasing linearly proportional to the decrease of the spatial energy magnitude. When transduction of time into spatial energy occurs, or vice versa, the energy conservation law is modified to specify the sum of the spatial energy, time-energy, and mass energy.

time-charge, time charging: Since time is a special highly compressed form of energy, then time-excitation or *time-charging* also occurs. Also, since tempic longitudinal EM waves accompany all longitudinal EM waves (per reinterpretation of the phase conjugate half set of the Whittaker decomposition), then irradiating a mass with substantial longitudinal EM wave radiation will time-charge the mass, exciting it with time-energy. There are excited states and internal structuring (engines) in time **charging.**

much like the excited states and internal structuring (engines) in energy excitation of normal charge. When the longitudinal EM wave radiation is removed, the time-charged entity then slowly emits longitudinal EM wave radiation as the time-excitation slowly decays via the giant negentropy process of the charges.

time-energy: Time may be considered as spatial EM energy compressed by the factor c^2. See also discussions under time and time charging.

time excitation: See time charging.

time-flow generatrix (mechanism): Since any observable is a frozen nonpersistent snapshot in 3-space, then the common expression "flow of a mass through time" requires an active generating mechanism to add time to mass, converting it to masstime, which is persistent. Masstime *mt* is as different from mass *m* as impulse *(Ft)* is from force *F*. The flow of a mass through time is generated by the ubiquitous interaction of photons with the mass. This includes all photons, both virtual and observable. Thus for the ambient vacuum, there is an ambient flow of time, since there is an ambient rate of total photon interaction. The "macroscopic" movement of the mass through time is the result of (i) absorptions and emissions of observable-sized photons, and (ii) the continuous integration of virtual photon absorptions into observable-sized photons. The macroscopic flow of time thus has myriads of "sublevels" of infolded time flows and engines. Hence the flow of time has a marvelous, internal, dynamic structure.

time-forward: Refers to the flow of an entity through time in the forward direction, as seen by an external observer. Time-forward flow occurs when (to the observer) there is a predominance of time-forward photons interacting with the body that is moving in forward time. Contrast to time-reversal of an object, as seen by the external observer, which occurs when (to the observer) there is a predominance of time-reversed photons (antiphotons) interacting with the body that is observed to be time-reversed. Pumping a given living mass (in the nonlinear optical sense) with time-reversed EM waves pumps the object in the time domain. That drammatically extends nonlinear optics, since the pumped living object's cells will be observed to slowly reverse back to a previous physical state or condition. Since time-polarized EM waves accompany all longitudinal EM waves via reinterpretation of Whittaker 1903, pumping a living mass (a body) with longitudinal EM waves also pumps it with time-polarized EM waves, time-charging the body which continues the pumping in the time domain after the longitudinal EM wave radiation is removed.

time-forward EM wave: A normal EM wave moving in forward time as seen by the observer. Otherwise known as the retarded wave. The external observer comfortably sees its forward progress as expected.

time-polarized electromagnetic wave: An electromagnetic wave whose vibrations occur in the time domain only. May be referred to in a quantum field theory sense as a *scalar* EM wave. It is a longitudinal EM wave in the time domain. It may also be said to be a longitudinal EM wave in the complex plane, since in Minkowski space the fourth axis may be taken as -ict. The time-polarized EM wave is indicated in quantum field theory, or at least its time-polarized photon is theorized and called a "scalar" photon.

time-polarized photon (scalar photon): A photon whose vibrations occur in the time domain only, also known as a "scalar" photon. This polarization of the photon is utilized in quantum field theory. Neither the scalar photon nor the longitudinal photon is individually observable. However, the combination of the two in the presence of charge is observable as the instantaneous scalar potential.

time reversal: For EM waves, the process of forming the phase conjugate wave. For a particle or a mass, the process of pumping the particle or mass with time-polarized EM waves, so that the resident spacetime curvature engine in the mass is amplified and phase conjugated, forming a precise, amplified antiengine which then slowly time-reverses the mass back to a previous physical condition and state. We accent that *time reversing a single object or single group of objects* is not the same thing as "travel into the past" so popularized by science fiction. For time travel to the past, the entire universe and everything in it — except the traveler — would have to be time-reversed. That would not seem possible by any stretch of the imagination today! On the other hand, time-reversal of a single thing such as a particle or a wave — or even a group of things such as a group of particles or waves — is not only feasible but is readily achievable. A Dirac sea hole is a negative energy electron, for example, having negative energy and therefore negative mass prior to observation. After observation, it is seen as a positive energy, positive mass electron of opposite charge to the conventional electron and reacting with fields in opposite direction from the conventional negatively charged electron. In short, after observation (interaction with mass) it becomes the positron.

time reversal zone: A region of space in which the majority component of the overall time flow mechanism runs backwards, and hence the fundamental mechanism generating the flow of a mass through time is reversed. The "time running backwards" is due to the predominance of interactions with antiphotons. It follows that, in the time-reversal zone (TRZ), there is a predominance of antiphoton interaction over photon interaction. In a TRZ, the law of attraction and repulsion of charges is reversed; like charges attract and unlike charges repel. Also, induced action by the time reversal forces start *inside and throughout* an affected mass in that zone, originating at every point in that local spacetime zone and moving outward. It thus involves a novel action "from inside to outside", beginning even inside the nucleons, and first affecting the quarks and gluons. Deliberate use of time-reversal zones and their subsequent decay allows a great new class of dramatic nuclear reactions, including simple transmutation and loosely called "cold fusion" reactions, at low spatial energy but very high time-energy. High energy physics has neglected the use of time-energy, opting to go for high spatial energy. Hence the high energy physicists have actually eliminated a much higher energy physics (at low spatial energy so readily examined on the bench far *more ch*eaply) than the science they have long developed.

time-reversed wave: A phase conjugate reflected wave, which is a wave moving in reversed time. Since the external observer "sees" in forward time only - and even then he sees in an iterative series of frozen 3-space snapsholt —he *observes* the time-reversed wave as a 3-spatial wave (series of iterative frozen snapshots) traveling in the opposite direction as its generatrix wavefront (the *signal wave).*

topological: Invariant under transformation by continuous mappings, or related to the properties of geometric shapes and figures that remain invariant under transformation by continuous mappings.

topological properties: Those properties of a geometrical space that are unchanged by continuous distortion of the space.

topology: Mathematics of the properties of geometric configurations invariant under transformation by continuous mappings. In short, the mathematics of the large-scale structure of curved spaces.

transverse (with respect to waves): Wave oscillatory movement perpendicular to the wave's primary direction of principal propagation, as in the movement of a plucked taut string in a wavelike motion traveling down the string but with the oscillations being lateral. Based on Faraday's notion that the field lines of the electromagnetic field were physical material lines (in the material ether) similar to taut strings, Maxwell assumed the transverse wave model in his theory. The entry of longitudinal waves, into the theory can easily be demonstrated by rhythmically and coherently varying the magnitudes of the Lorentz symmetrical regauging condition, producing "stress potential waves" which are in fact what longitudinal EM waves really are. The production of a longitudinal EM wave in 3-space will be accompanied by the simultaneous production of a time-polarized companion wave.

tunneling: In quantum mechanics, a phenomenon in which a particle penetrates and crosses a small region where the potential opposing such passage is greater than the particle's available energy. The phenomenon is thought to be impossible according to classical physics, but occurs nonetheless. Further, the speed of transit through the "tunneling" separation can be superluminal. Tunneling of music between two separated points in a waveguide has been measured at more than four times the speed of light.

A more extended type of tunneling is the travel of longitudinal EM waves infolded inside the normal EM potentials, fields, and waves. Indeed, this is a more fundamental electrodynamics, and it can be shown that interferometry of this "internal" electrodynamics creates the external electrodynamics by a form of longitudinal EM wave interferometry. E.g., see M.W. Evans, P.K. Anastasovski, T.E. Bearden *et al.*, "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," *Journal ofNew Energy,* 4(3), Special Issue, Winter 1999, p. 76-78. This is part of the technical basis for the "scalar interferometry weapons" developed by the Russians after World War II. Today, some 10 nations have such weapons, more are developing them, and even the Japanese Yakuza has such weapons.

unified field theory (engineerable): A unified theory of the four forces o| physics — the electromagnetic, gravitational, strong, and weak forces that is not just an intellectual model but which is also engineerable on **the** laboratory bench and in actual physical systems using higher symmetry O(3) electrodynamics as a special subset of Sachs' unified field theory.

unitary Q_E (charge of the electron): Fixed or unit charge q of the electron. This notion is held because q is not decomposed into its component elements, but treated as if it were an indivisible unit, charge and mass and all. The mass of a fundamental particle at rest is quantized, and the massless charge (self-potential) of the particle is discretized as a function of the background potential (virtual particle flux density) in which it is embedded or to which it is exposed. In our view the definition **of a** charge q is $q \equiv m_q\phi_q$, at least to first order. (The ϕ_q may be internally structured with deterministic "engines", so that one electron need not bo identical to another, in its actions in some cases). The *electrical charge* of the "charged mass m_q" should be taken to be the massless ϕ_q component. For an electron, $q_e \equiv m_e\phi_e$. The *electrical charge* of the electron should be taken to be the massles ϕ_e component. We note that the charge is also structured, since ϕ_e or ϕ_q decomposes via Whittaker 1903 into a harmonic setof EM longitudinal phase conjugate wavepairs. Our reinterpretation of this decomposition then yields the solution to the problem of the "source charge" and its associated fields and potentials and their energy. The massless charge ϕ_e is a dipolarity and hence a broken 3-symmetry in EM energy flow, invoking a more primary 4-symmetry in EM energy flow so that EM longitudinal wave energy is continuously incoming from the time domain, being absorbed by the charge q of which ϕ_e is a component, being transduced into real EM energy in 3-space, and emitted in all directions in 3-space at the speed of light, creating the fields and potentials and their energy that are associated with charge q. If a charge (say, an electron) is placed in an additional potential ϕ_1, then the total massless charge of the electron (i.e., its virtual vacuum flux exchange with mass me) is $(\phi_e+\phi_1)$.

As can be seen, the massless charge of the electron (its virtual flux activity) has now changed. The electron is then said to *be collecting and to have collected* excess energy (which may be either positive or negative, depending upon the sign of ϕ_1). Further, either ϕ_e or ϕ_1 or both can be internally and deterministically *dimensioned* (structured in Stoney-Whillaker-Ziolkowski fashion). When its massless charge (its own ϕ_e) potential is so dimensioned, then the dimensioned electron is said to contain or transport an *engine* or a *vacuum engine* or a *spacetime curvature engine.*

vacuum: Space devoid of observable matter. In modern theory, the "empty" space is in fact bristling with very rapid fluctuations of EM energy, remaining in the virtual state. It is also filled with a violent, fluctuating flux of virtual particles, appearing and disappearing so quickly that an individual particle does not persist long enough to be individually detected. Thus the vacuum is extraordinarily energetic, but the energy is in very special form (fleeting violent fluctuations and virtual particle flux). Nonetheless, because it contains enormous energy, the *average* vacuum may be considered a potential. All potentials then are changes to this vacuum potential, or to an intermediate potential which is a change to it. In another sense, the vacuum, potential, and spacetime are all one-and-the-same thing. This means that general relativity's spacetime curvature is ubiquitous and extraordinarily dynamic. The vacuum very near a mass also can act in some cases as a semiconductor, capable of supporting self-oscillation between the mass and the local vacuum. Considerations of the active vacuum and its nature lead to the notion of the *supersystem*, which is comprised of three parts: (1) the physical system and its dynamics, (2) the active vacuum and its dynamics, and (3) general relativity and its dynamics (since physicists persist in separating the notions of spacetime and vacuum). No analysis of any mass system is complete unless an analysis of the entire supersystem and the interactions of the three components, one with the other, is performed.

Every charge is "embedded" in the active vacuum by its potentials and fields, which themselves are spacetime curvatures and alterations of the local vacuum. This is easily seen by $q \equiv m_q\phi_q$, where ϕ_q is a change to the local vacuum potential *a priori*. In the quantum view, the fields and potentials may be taken as organized polarization patterns and dynamics (currents) in the vacuum virtual particle flux.

vacuum engine: Also referred to as a *spacetime curvature engine*. From the particle view: a deterministic internal pattern or *template* deliberately created as a deterministic structuring and dynamics in the virtual flux comprising a scalar potential. It can also be created in, and transported by, the two scalar potentials that comprise a static or dynamic force field, or a moving EM wave in the vacuum. It can also be created in the local vacuum potential, in which case it has a "charge-up" creation time and a "discharge time," exhibiting a sort of "time constant" very much like a capacitor charging and discharging.

In the wave view: the vacuum engine is a deterministic internal pattern or template deliberately created in the Stoney-Whittaker-Ziolkowski internal biwaves and wave products comprising the scalar potential. Again, it can be implanted in the vacuum or in an EM signal as a carrier. The self-energy (scalar potential) portion of any electric charge or magnetic charge can also be a "carrier" that is conditioned *{dimensioned)* with a vacuum potential. The process or act of conditioning any of the carriers with vacuum engines is called *dimensioning* that carrier. Russian energetics used in secret weapons programs refers to the vacuum engines of a dimensioned carrier as the *information content of the field.* Such vacuum engines, e.g., were used in the potentials of the microwave radiation of the U.S. Embassy in Moscow for several decades, inducing a variety of health changes and sicknesses in the personnel of the embassy. Interestingly, all these health changes occurred in the field-free regions of the Embassy. There the field-free potentials existed and were stable and unchanging. Hence a body exposed in that region was steadily exposed to the specific engine being induced, and time-charged with that engine accordingly. The deterministic set of spacetime curvatures comprising this charge then acted upon the cells and body to slowly change it to the damaged or diseased condition. See dimensioning.

vacuum flux exchange: The virtual particle interaction between the active vacuum and a mass, particularly a charged particle.

vacuum's **virtual particle flux:** In the particle viewpoint, the quantum mechanical vacuum identically is a virtual particle flux. Present physics, however, is prone to refer to the vacuum as an implied emptiness containing a virtual particle flux, even though that is known to be in error. In our view, energy, vacuum, spacetime, virtual particle flux, and dynamic geometry are all synonyms. One must note that the ageless contradiction of the wave vs. particle view has still not been resolved in physics, even by the duality principle. The duality principle is merely an agreement to quit fighting, and to feel free to consider a thing a particle in one case or circumstance and a field in another case or circumstance, *whichever works.*

vector: In mathematics, an entity completely characterized by a magnitude and a direction.

vector potential: A potential having — at any point in it — both direction and magnitude. The direction, e.g., may be linear, curvilinear, or swirling. The term "vector potential" is often used as the magnetic vector potential A. When it is swirling, the magnetic vector potential A may be said to be the magnetic field B. However, the swirling magnetic vector potential (the magnetic field B) may be "bottled up" and contained locally inside a long tight solenoid, or inside a toroidal coil, or inside a special nanocrystalline material used as a closed transformer core as in the motionless electromagnetic generator. In that case nature freely replenishes the energy in the surrounding space which would have been occupied by the nonrestricted, nonlocalized B-field (curled A-potential) but replenishes that space with an uncurled (linearly flowing) magnetic vector potential A. This in fact at least doubles the available energy flow available, without any additional EM energy input by the operator. The well-known Aharnov-Bohm effect clearly demonstrates this effect, and it is deliberately employed in the motionless electromagnetic generator (MEG) of which the author is one of the five co-inventors. Present dictionaries of physics and many textbooks are confused as to the definition of magnetic vector potential A and its relation to B, many stating that the A is defined by the equation $\nabla\times\mathbf{A} = \mathbf{B}$. However, no equation is a definition; it is an expression of equality of two magnitudes, and not a definition of either side. If we substitute the more stringent $\nabla\times\mathbf{A} \equiv \mathbf{B}$, we see that the $\nabla\times$

operator acting upon the A potential is precisely what a B field *identically is*. Hence the A potential exists in two forms: (1) in curled form as identically the B-field, and (2) in uncurled field-free form as the linearly moving A-potential energy current that has no imposed $\nabla\times$ operator and thus has no curl. This makes total sense with $\nabla\times\mathbf{A}\equiv\mathbf{B}$.

vector field: Conventionally, the assigning of a magnitude and a direction (that is, a vector) to each point in a region of space occupied by a field. In our view the field is in fact a change to the active energetics of that spatial region, such that the region will interact with any appropriate entity (e.g , **a** charged mass) placed at any point in the region and produce a force upon that entity.

vector zero: See discussion under zero vector.

virtual antiphoton: An antiphoton that spontaneously appears and disappears, so swiftly that it cannot be individually observed. It is real while it exists. Coherent integration of a stream of virtual photons, however, adds to produce an observable quantum of energy, which is observable in its effects on a charged particle.

virtual particle: A fleeting quantum particle that spontaneously appears and disappears, so swiftly that it cannot be individually observed, so that it exists only temporarily. It is real while it exists. The virtual particle does not satisfy the usual relation between energy, momentum, and mass because it is under the aegis of the Heisenberg uncertainty principle. The virtual particle can have any amount of energy momentarily, so long as the product of its energy and the time interval of its existence is less than the uncertainty principle's minimum magnitude.

The coherent interactions of large numbers of virtual particles with a **mass** or charge can combine to generate real observable effects (forces). In quantum field theory, all forces of nature are caused by the interaction **of** the mass entity with virtual particles. Interaction of a mass with virtual photons, e.g., is projected to cause all mechanical and electromagnetic forces.

virtual particle flux exchange: Interaction with or exchange of a flux of virtual particles.

virtual photon: A photon that spontaneously appears and disappears in the vacuum so quickly that it cannot be individually observed.

virtual photon flux: See definitions under **flux** and under **virtual particle.** For electromagnetics, the vacuum can often be modeled as a flux of virtual photons. One gets the virtual charges also, since the photon is also considered to consist of a positron-electron pair. That is, the photon is continually separating into a positron-electron pair, whose particles then annihilate each other to form additional photon(s), etc.

virtual photon flux exchange: The mutual exchange of virtual photons, as between the vacuum and a charged mass, or between two colliding masses.

virtual photon flux of vacuum: In our view, the vacuum identically is a flux of virtual particles. For many electromagnetic cases, it can just be modeled as identically a flux of virtual photons, where the word "photons" in general usage refers to both photons and antiphotons.

virtual positron: A positive electron (positron) which spontaneously appears disappears in the vacuum so quickly that it cannot be individually observed. More accurately, in the 4-vacuum, a negative energy electron state that spontaneously appears and disappears so quickly that it cannot be individually observed. A virtual Dirac sea hole.

Whittaker, Edmund Taylor: Well-known English mathematician and physicist, 1873-1956, who made major contributions to mathematics and physics. Whittaker dramatically extended Stoney's earlier work of 1897-8 an authored two papers in 1903 and 1904 giving the basis for the theory of scalar electromagnetics (Russian energetics, not including the quantum potential portion). His decomposition of the scalar potential into a harmonic set of longitudinal EM wavepairs, where each pair is a phase conjugate pair, may be reinterpreted to solve the long-vexing problem of the source charge or source dipole and its continuous emission of EM energy in 3-space, providing its associated potentials and fields and their energy. The problem was solved by the present author's reinterpretation in that manner, published in 2000. This was later found to be consistent with quantum field theory as well, by interpreting the observability of the combined scalar and longitudinal photon pair as opposed to the nonobservability of either the scalar or longitudinal photon individually.

Work: The changing of the form of energy. One joule of energy whose form is totally changed in a reaction, does one joule of work in that reaction. However, it is strongly stressed that after that joule of work is perform**ed,** one still has a joule of energy left to use (with which to do additional work). This remaining joule of energy is just in different form than was the initial joule of energy. So one can do joule after joule of work from a single *initial* joule of energy, so long as the energy is not dissipated and lost after each change of its form. In any one single change of form, only o**ne** joule of work can be done from one joule of energy. So energy is conserved through many changes of form, but work is not. To work in COP>1.0 systems, one absolutely must root out from one's consciousness the mistaken notion that only one joule of work can be done from a joule of input energy. One of the places where extra joules of work are performed by each joule of energy is that of strong gradients. Strong gradients are known and recognized to violate thermodynamics. Work is further divided into **internal work** and **external work.**

work, external: The changing of the form of the collected energy of a body or system by dissipating it upon an external load or entity is the external work *done by* a body. The changing of the collected energy of a body or system by performing work upon it is the external work *done upon* a body or system.

work, internal: The changing of the form of the internal energy of a body or system. It may also involve the changing of internal parts of the system.

work-energy theorem: In mechanics, a theorem that the change in kinetic energy of a particle in motion is equal to the work done by all the forces acting upon the particle.

zero-point energy: The energy remaining at a temperature of absolute zero, due to the Heisenberg uncertainty principle and the quantum mechanical theory of an oscillator. Also, the minimum energy of a system due to its quantum fluctuations, resulting from its incessant virtual particle activity. Quantum mechanically, no system of interest (including even spacetime itself) can have zero energy. The zero-point energy of the vacuum is the lowest energy vacuum state, with fluctuations taken into account. Even at low energies, quantum fluctuations continually arise, and result in an incessant, extremely rapid, and violent "jittering" of the energy momentarily present. The minimum energy due to these quantum fluctuations is called the *zero-point energy.*

zero vector: Candidates are: (i) A vector having no length (no magnitude) or specific direction; (ii) absence of any and all finite vectors; (iii) resultant of a system of multiple, finite vectors whose vector sum has **no** specific direction but may have nonzero scalar magnitude.

The second definition differs drastically from the first. That is, we **cannot** say that a vector resultant zero has no infolded magnitude, if we **consider** its nonzero components, all of which have magnitude. (See the discussion under **zero).** It only has no net magnitude for *translating* force. It is, in fact, *a. stress potential,* and it is one with a deterministic internal structure or *engine.* By rhythmically and coherently varying each component of a zero vector system, one easily produces longitudinal EM waves,, each of which is accompanied by a correlated time-polarized EM wave. In vector analysis, the zero vector axiom states that any vector v plus a zero vector is equal to a zero vector plus v. Note that, being an axiom, this is not proven, but just assumed for any zero vector. We very much disagree with that blanket axiom; it reduces vector analysis to a highly special single case, where all zero vector systems are to be considered equal, and totally "inactive." That is true only for bulk translation. In the real world that need not be true at all; specifically, it is not true for two different zero vectors having different internal structure and dynamics (different internal engines).

There are an infinite number of different zero vector resultant systems, and so zero vectors may differ, if their infolded component "substructure" is considered. In the real world, things that are vectors very often have energy. The energy is infolded and hidden, so that it represents not only (i) a local curvature of spacetime, but also (ii) a structured, deterministic set (a template) of curvatures of local spacetime. Therefore this zero vector system is actually a *vacuum engine.* Any mass system placed in such a potential, will be acted upon in its internal parts by that local vacuum engine; in short, it will have internal work performed on its interior. Further, the zero vector resultant system is a "stress potential," since it does contain "collected internal energy."

So between zero vector resultant systems, the "infolded, hidden energies" differ dramatically in (i) total infolded and trapped (collected) energy, (ii) specific internal action patterns (templates), (iii) general relativistic effects, and (iv) type of engine. Accordingly, the actions the zero-vector system induces in an exposed mass may be designed in advance. This is vacuum engineering. This spacetime engineering. This is the formation and use of engines. It is also one of the bases of Russian energetics.

The'problem is the *vector algebra itself.* In the abstract algebra, a "vector space" Is regarded as an inert thing, where the only actants are the overall net nonzero vectors, and there is no interaction between an overt vector and a covert set of vectors. The vector space is not allowed to have an internal realm, stress, hidden engines, or to interact with the vectors representing the physical system. So vector algebra lacks the capacity to model such higher topology phenomenology. Consequently, for over a century electrodynamicists — thinking primarily of spatial translation — have routinely discarded such "zero vector systems" as if they represented the total absence of any finite vector. *This they do, in the unsuitable* *assumptions of the vector algebra.* This they do not do, in the real world, whether we are aware of it or not.

The common practice in electrical physics of replacing a zero-summed system of nonzero vectors with a vector zero indicating a "complete absence of vectors" is very much incomplete. This present practice should be changed; it has reduced the topology of electromagnetics far too much, and gutted vast, important parts of the theory insofar as an extended electromagnetics is concerned. The vector zero should be replaced with a special organized vector zero system and a dimensioned scalar potential.

The interactions of the "hidden" vectors — the *engines* — should be incorporated. By failing to do this, physics excludes the ability to engineer electrogravitation, spacetime, the virtual state, local general relativity, free energy, effects at a distance, and the probabilities of the states propagated by the Schrodinger equation. It also excludes unification of forces. As another more human example, for every cellular disease condition there is a specific engine resident in the cells. An amplified antiengine for that disease pattern can fairly readily be generated. That antiengine, embedded in ordinary EM fields and potentials, can be irradiated into a diseased body with that disease, and it will time-reverse those diseased cells slowly back to normal. This is actually the basis of the body's own healing system the cellular regenerative system studied mostly by Becker. Unwittingly using this "antiengine" approach, Priore and a team of French scientists provided startling cures of terminal cancers, infectious diseases, clogged arteries, etc. in laboratory animals in France in the 1860s and early 1970s. The work is strongly documented and replicated in the French scientific literature, but the work was suppressed in the early 1970s in the change of the French government. See our "porthole briefing" on our website www.cheniere.org for our extension of the Priore method, for greater ease portability, and rapid treatment of mass casualties.

In the new view, a zero vector may be a system of nonzero vectors that vectorially sum or multiply to zero for spatial translation purposes, but nonzero components may still exist "inside" and dynamically function inside the zero-resultant envelope. This substructure can act on nonlinear or resonant systems. Any or all of the nonzero components may be time varying, or none of them may be. The zero vector system may thus be a dynamic engine and a dimensioned scalar stress potential. In addition, such a vector zero system is considered to be also a scalar stress potential, whose magnitude is equal to the sum of the absolute values of the components. In addition, a separate type of stress potential may be included for the magnitude of the absolute values of the swirl components (torques). Other associated potentials may be included for the system, particularly if the system is in multiple dimensions, if it consists of more than one type of virtual particle flux, if several fluxes intercommute, etc.

zero vector summation systems: Zero summations of real, nonzero vectors, so that the real vectors are still present but do not cause *spatial translation* of the exposed mass or charges. Instead, the zero vector summation has an internal structure and dynamics, hence an internal engine. Also see discussion under zero vector.

Appendix A
Relation to Equilibrium and Disequilibrium Thermodynamics

"The second law [of thermodynamics] does not rule out the possibility of pushing heat uphill, as it were, from a cold object to a hot one, or of creating order out of disorder. It merely states that such a reversal of the natural flow requires an influx of energy... " [Raymond Serway] {697}

"It *is in no way possible, either by mechanical, thermal, chemical, or other devices, to obtain perpetual motion, i.e., it is impossible to construct an engine which will work in a cycle and produce continuous work, or kinetic energy, from nothing. "* [Max Planck] {698}.

"What might appear to be empty space is, therefore, a seething ferment of virtual particles. A vacuum is not inert ***and*** *featureless, but alive with throbbing energy and* ***vitality.*** *A 'real' particle such as an electron must always be viewed against this background of frenetic activity. When an electron moves through space, it is actually swimming in a sea of ghost particles of all varieties - virtual leptons, quarks, and messengers, entangled in a complex melee. The presence of the electron will distort this irreducible vacuum activity, and the distortion in turn reacts back on the electron. Even at rest, an electron is* ***not*** *at rest: it is being continually assaulted by all manner of other particles from the vacuum."* [Paul Davies] {699}

"I assert... that for the general theory of relativity, i.e., in ***the*** *case of the general invariance of the Hamiltonian function, energy equations... corresponding to the energy equations in orthogonally invariant theories do not exist* ***at all.*** *I could even take this circumstance as the characteristic feature of the theory of general relativity. "* **[D. Hilbert]** {700}

*"****In formulating the*** *equivalence principle, Einstein actually abandoned the idea of the gravitational field as a*

Faraday-Maxwell field, and this is reflected in the pseudotensorial characterization of the gravitational field that he introduced. Hilbert was the first to draw attention to the consequences of this. ... Unfortunately, ... Hilbert was evidently not understood by his contemporaries, since neither Einstein himself nor other physicists recognized the fact that in general relativity conservation laws for energy, momentum, and angular momentum are in principle impossible. " [Logunov and Loskutov] {701}

"No amount of theory can refute a replicable experiment that contradicts the theory. A single replicable experiment can refute any theory that is in conflict. That is the very heart of the scientific method. " [Myron Evans, private correspondence].

"It only takes a single white crow to prove that not all crows are black. " [Anon.]

"One of the most important principles of physics, that disorder, or entropy, always increases, has been shown to be untrue. " [BBC News] {702}

"One of the most fundamental rules of physics, the second law of thermodynamics, has for the first time been shown not to hold for microscopic systems. ... [The experimenters]... found that the change in entropy was negative over time intevals of a few tenths of a second, revealing nature running in reverse. ...But over time intervals of more than two seconds, on overall positive entropy change was measured and normality restored." [Chalmers] {703}.

"We experimentally demonstrate the fluctuation theorem, which predicts appreciable and measurable violations of the second law of thermodynamics for small systems over short time scales... Our results show entropy consumption can occur over colloidal length and time scales. " [Wang et al.] {704}.

A.1 Introduction

For excellent background books on thermodynamics matters, at various levels of difficulty, the reader is referred to texts by Gyftopoulos and

Beretta {705}, Kondepudi and Prigogine {706}, Chalidze {707}, and Dugdale {708}.

It has long been known in physics that at sufficiently small levels (e.g., atomic levels) the second law of thermodynamics does not always hold true. Recently, as quoted above, Wang et al. {704} have extended this area of permissible violation of the second law to much more appreciable times (a few tenths of a second) and lengths (the micron scale). Without further discussion, we note that this places regions and times of negentropy well-within circuitry switching times and interaction with fundamental chargedparticles. Hence these experiments are beginning to point the way to practical negentropic devices. When coupled with the demonstrated continous giant negentropy of a charge (considered with its clustering virtual charges of opposite signs), these experiments now strongly support the feasibility of practical overunity EM circuits and systems. The interested reader is encouraged to continue to read the literature along those lines, and stay abreast of developments in that field.

In this book we have presented considerable discussion of time-energy (time as a special, highly compacted form of energy). We have also discussed the three components of the supersystem, including the system, the local active vacuum, and the local curvatures of spacetime. All three componets of the supersystem interact with each other continuously, and such Interactions are not modeled in classical thermodynamics (CTD), classical electrodynamics (CED), or electrical engineering (EE). Hence one should not be surprised that the supersystem model and supersystem analysis involve contradictions (actually, *extensions)* of these three models, which are special cases and approximations only.

All COP>1.0 EM systems result from supersystem component interactions. So COP>1.0 EM systems are in conflict with CTD, CED, and EE since these three models exclude those interactions.

To falsify any absoluteness of the three models considered, one or more solid, demonstrable, replicable experiments are presented to establish validity of the new higher symmetry approach. This is sufficient to negate the conventionally accepted absoluteness of the lower symmetry CTD, CED, and EE models. Scientifically, these experiments suffice to demonstrate that those three models describe "*special-case* domains of nature" rather than full generalizations of nature. It is clearly understood that the three models remain very useful and very often quite sufficient, just as Newtonian mechanics is useful and often sufficient, without use of relativistic mechanics.

The experiments also falsify any interpretation of the three models as "proof that COP>1.0 systems are impossible. Accordingly, many of the accepted statements and assumptions of the three models must be corrected for COP>1.0 research, as must be the conventional interpretation that the laws of nature (as understood based on those three models) prohibit COP>1.0 EM systems. This follows directly from the quotation by Evans at the beginning of this appendix. Experimental evidence of COP>1.0 is abundant and decisive, such as the many hundreds of routine "negative resonance absorption of the medium" experiments as typified by Bohren's paper {709} and experiment and Letokhov's work {710}. Also, every charge in the universe already falsifies all three models, as amply discussed in this book. That fact for at least a century has generated the source charge problem — the problem of the source charge and its associated fields and potentials and their energy, reaching across the universe. The solution to this problem was published by the present author in 2000, based on the experimentally proven and well-established broken symmetry of opposite charges, such as are on the ends of a dipole.

In the extended approach, conservation of energy in 4-space does not require simultaneous conservation of energy in 3-space. The concept of energy conservation in 4-space does not exist in the three models mentioned, which contain only 3-space energy conservation. There exists a "giant negentropy" 4-space energy flow (circulation) mechanism, as we developed in this book, and it is experimentally demonstrated by every charge in the universe. In a classical 3-spatial energy conservation analysis and view, the giant negentropy action may continuously perform useful work in a system, since it already furnishes all EM field energy and potential energy in every system. All 3-space energy conservation is an extra condition imposed upon 4-space energy conservation, and that condition may be violated at will without violation of the general conservation of energy law.

To the three conventional models mentioned, 4-symmetry in energy flow with broken 3-symmetry, will appear to be perpetual motion and a total violation of the more limited 3-space conservation of energy law. It *does* permissibly violate 3-space energy conservation, which is just a special case a priori. But it is not a violation of energy conservation overall, since we are using a more fundamental energy conservation law (4-symmetry) than the 3-symmetry used in the mentioned CTD, CED, and EE models,

In short, the three models must be extended to provide 4-space energy conservation with simultaneous violation of 3-space energy conservation, accompanied by simultaneous time-symmetry energy flow violation. The two broken symmetries — of 3-space energy flow and of time domain energy flow — thus act as a special extension of the notion of symmetrical regauging, accomplishing it by "equal and opposite-effecting" fields, one in the lime domain and one in 3-space.

Ironically it then follows that what have heretofore been conceptually called "perpetual motion machines" with respect to 3-space energy conservation are simply attempts — successful or unsuccessful — to build analogues of a windmill turning in a free wind and powering a load. In short they are attempts to duplicate in systems what every charge in the universe already demonstrates unequivocally. *Nature's* CTD, CED, and EE do not prohibit such systems — again, as evidenced by every charge in the universe demonstrating exactly such "perpetual motion" (we prefer the exact term "giant negentropy") and violation of present *human* CTD, CED, and EE. None of the original charges from the beginning of the universe has slowed and stopped its motion, or stopped pouring out EM energy freely, in some 14 billion years or so. And charge may be taken as the most basic Maxwellian system. Further, all EM fields and potentials, and their energy, comes from their source charges. So those so adamant against perpetual motion machines should first purge their own three models of their *assumed* perpetual motion machines. If this is done, they will purge the three models of all charges, fields, and potentials — and all EM energy in the process. In short, the three conventional models used by those same skeptlcs so loudly condemning "perpetual motion machines" will eat themselves" by swallowing their own tail, so to speak.

in short, all three models are easily falsified by every charge in the universe, as well as by every EM field and potential, and by the existence of EM energy itself.

Either their models must surrender any pretense at absoluteness and accept being useful approximations only, or else the conservation of energy law itself must be surrendered.

Conventional CEM, CTD, and EE have just hidden this appalling fact and convealed the serious flaws in the modeling by ignoring what Sen {711} reffered to as " t*he most difficult problem in classical and quantum electrodynamics.* " We believe it will very difficult to find a single electrical engineering department or professor or text that deliberately acquaints electrical engineering students with the source charge problem and its implications.

As an example, in Chapter 9 we presented an experimentally verified and patent-pending method for a special energy conversion method. This method is proven on the laboratory bench. With this method, a system accepts time-energy and negative energy input, converts it to ordinary positive 3-spatial energy, and uses the converted input energy to close-loop an EM system far from equilibrium. With that conversion process invoked in a governed feedback loop, the COP>1.0 system becomes "self-powering" (an informal term referring to the system freely receiving all its input energy from its active environment). Self-powering, of course, involves the system continuously operating and performing useful work, analogous to a windmill in a wind and powering a water pump. Rigorously, the self-powering COP = oo EM system becomes directly powered by energy from the broken symmetry induced in its supersystem component interactions.

In Chapter 10 we presented the novel concept of the time-reversal zone (TRZ), which uses the extraordinary EM energy density of time to reverse the Coulomb barrier in nuclear reactions in an electrolyte {712}. This enables nuclear transmutation reactions at low *spatial* energy (but at very high *total* energy when time energy is also considered). There are some 600 successful cold fusion experiments worldwide, in multiple laboratories, by multiple researchers, and most of these experiments demonstrate examples of the specific nuclear reactions made possible by the actions of the TRZ. The time-energy anomalies are directly demonstrated in rigorous electrolyte experiments at U.S. Navy research facilities at China Lake, as we have also discussed in this book. So the experimental results are replicated and well proven. Hence there is adequate, replicated experimental demonstration that additional nuclear reactions exist outside the conventionally known nuclear reaction model. Therefore the set of conventional nuclear reaction models also is reduced to a special case domain, with any absoluteness of those models already experimentally falsified. Either that, or we must abandon scientific model where successfully replicated experiments falsify any theory in contradiction to the experiments.

In our proposed solution to the source charge problem {71 lb}, we found a broken 3-symmetry in energy flow while 4-symmetry is maintained. This is a more primary principle of energy conservation. We called this principle "giant negentropy", which is laughably simple to demonstrate experimentally, since every charge and dipole in the universe already demonstrates it. We showed that such giant negentropy is consistent with quantum field theory, particle physics, and a slight corrective reinterpretation of Whittaker's 1903 decomposition of the scalar **potential** The experimentally proven and theoretically justified giant negentropy falsifies any absoluteness of the CTD, CED, and EE models in conflict with it.

Again, this does not mean that one abandons the CTD, CED, and EE models! They are still very useful models in most cases of interest. It means that these present models are still incomplete, and extended models must be developed and used for COP>1.0 EM system research. It also means that none of these three models is valid as a prohibition of the permissibility of COP>1 EM systems.

All EM energy in 3-space comes from the time domain and returns to it. Since all three models exclude time-energy and its transduction into spatial energy, all three are in conflict with the proven giant negentropy model and with the unlimited experimental demonstrations. Accordingly, the experimentally demonstrable giant negentropy of the charge and dipole strongly points out the inadequacy of the three models and the nuclear reaction models, which do not account for time reversal zones, or time-energy and its transduction into a very large magnitude of 3-space energy.

In paragraph A.2 below, we briefly discuss some of the fundamental but permissibie supersystem violations of these older models of CTD, CED, and EE, and provide some comments. It is obvious that all three classical models need very serious updating and extension. We point out that exceptions to present thermodynamics models and "laws" already are recognized{713}, and work on such exceptions is being performed under the heading of extended thermodynamics {714}.

In effect, all three models *exclude* every charge, dipole, or dipolarity in the universe - which therefore excludes all matter, fields, and energy - something completely untenable in modern electrodynamics and modern physics. Since the fundamental nature of the "heat" notion of CTD is electromagnetic field energy, and all EM fields and potentials (and their energy) are excluded if their source charges are excluded, then heat itself is excluded by the three models. That is, CTD is also totally falsified by CED'S exclusion of the source charge's demonstrated giant negentropy, which in effect excludes heat itself. This is a non sequitur of such magnitude that it absolutely requires the complete rework of CTD, CED, and EE.

A.1.1 Featurcs of the Development of Thermodynamics

In the development of thermodynamics, a few key events of interest are listed in Figure A-1 below.

Selected Events in the History of Thermodynamics {715}

1673	Huygens builds a motor driven by gunpowder explosion.
1690	Papin uses steam to move a piston.
1712	Newcomen develops a steam engine.
1723	Stahl popularizes phlogiston and Becher's concepts.
1724	Boerhaave proposes that heat is a fluid.
1733	Bernoulli publishes a text that is the first truly statistical treatment of kinetic theory, and publishes an update in 1738. It is almost forgotten until 1859.
1765	Watts invents a steam engine that is more than six times more effective than Newcomen's engine.
1783	Lavoisier points out the weakness of phlogiston theory with respect to combustion.
1786	Lavoisier and Laplace provide a Memoir on Heat.
1789	Lavoisier publishes his *Elementary Treatise on Chemistry,* which presents the law of conservation of mass.
1791	Prevost publishes a theory of heat and radiation exchange. Richter founds stoichiometry.
1798	Thompson demonstrates the conversion of work into heat, in his *Enquiry Concerning the Source of Heat which is Excited by Friction.*
1799	Davy's experiments seriously challenge phlogiston theory.
1800	Herschel explores the heating ability of different frequencies of light.
1802	Lussac establishes the thermal expansion of gases.
1803	Dalton forms his atomic theory of matter, setting the stage for atomic chemistry.
1804	Leslie shows that light and radiated heat have similar properties.
1806	Young formulates a precursor to the modern kinetic energy.
1811	Poisson develops a mathematical theory of heat, based on Fourier's work. Avogadro presents his law. Berzelius states that electrical and chemical

	forces are the same and that atoms are electrically charged.
1819	Dulong and Petit show an empirical law on the specific heat of solids.
1824	Carnot studies the thermal efficiency of machines, introduces reversible transformations, and generates the foundations of thermodynamics in his "Reflections on the Motive Power of Fire".
1829	Coriolis defines "kinetic energy".
1834	Clapeyron presents the first version of the second law of thermodynamics, based on steam engine studies.
1837	von Suerman experimentally verifies Clapeyron's version of Carnot's equations.
1842	Mayer equates the conservation of energy law and shows the equivalence between heat and mechanical energy.
1843	Waterston anonymously publishes an accurate kinetic theory of gases. It is essentially ignored.
1845	Waterston submits a beautiful paper on the kinetic theory of gases to the Royal Society. It is rejected and ignored for more than 40 years, until finally resurrected by the President of the Royal Society.
1847	Helmholtz generalizes the conservation of energy law for mechanical, caloric, electric and magnetic energy.
1848	Lord Kelvin introduces the absolute temperature scale. Joule finishes establishing the exact relation between heat and mechanical work; his work is almost ignored.
1850	Clausius states the second law of thermodynamics.
1851	Thomson derives the second law of thermodynamics based on Carnot's work.
1855	Rankine publishes *Outlines of the Science of Energetics.*
1857	Clausius establishes heat as energy distributed among particles.
1858	Clausius introduces the idea of the mean free path of a colliding particle.

1859	Kirchhoff establishes the thermodynamic foundation of radiation theory, and completes the basic radiation theory in 1862.
1860	Maxwell provides studies in kinetic theory of gases and gives the distribution of molecular velocities. The paper, "Illustrations of the Dynamical Theory of Gases," was read in 1859.
1860	Maxwell shows a discrepancy between experiment and the kinetic theory of diatomic gases. It was not resolved until the early stages of quantum theory.
1861	Kirchhoff advances the notion of the blackbody.
1865	Clausius, Rankine and Kelvin provide the concept and definition of entropy.
1860-1877	Boltzmann and Gibbs provide the foundations of statistical classical mechanics.
1867	Maxwell publishes *On the Dynamical Theory of Gases,* his major work on the kinetic theory.
1874	Thomson points out the reversibility paradox.
1877	Boltzmann defines entropy in terms of probabilities.
1873-1878	Gibbs publishes several important papers on thermodynamics, widely influencing scientists.
1879	Boltzmann establishes the relation between radiation intensity per unit area of a blackbody, and the absolute temperature. Stephan formulates the "Stephan-Boltzmann law" $R_T = \sigma T^4$.
1879-1899	Discoveries of various laws and effects, particularly at the micro level.
1900	Planck proposes the quantum of action and derives the correct radiation spectrum for blackbodies. (He published the definitive paper in 1911).
1905	von Smoluchowski and Einstein independently investigate Brownian motion. Einstein publishes a paper on the photoelectric effect.
1906	Nernst advances his heat theorem, where at absolute zero temperature both heat capacity and entropy change go to zero. This is later recognized as the "third law of thermodynamics".
1909	Einstein corrects Planck's blackbody derivation.

1910 Perrin experimentally validates Einstein's calculations on Brownian motion, providing a strong argument for the existence of atoms.

1911 Rutherford proposes the nuclear model of the atom. Sackur points out the need for an absolute definition of entropy. Nernst's experiments provide strong support for the new quantum theories.

1925 Born, Heisenberg and Jordan formulate quantum mechanics from matrix algebra. Einstein points out that the analogy between quantum gases and molecular gases is complete; photons and molecules have both wave and particle characteristics.

Figure A-l. Key Events in the History of Thermodynamics up to the Development of Quantum Mechanics.

During the early period of science, a continuing struggle between various notions of esoteric fluids and fluid flow occurred. French scientist Lavoisier (1743-1794), regarded as the father of chemistry, is identified with the concept of heat as an invisible fluid, which he named "caloric". The notion was that one drove this invisible fluid — this caloric — out of a material by some action such as mechanical friction or combustion. This produced the "manifestations" of what was loosely called "heat". In short, caloric was considered to be "heat" which was itself an esoteric fluid. In his famous Treatise {716}, published posthumously, atoms were still in dispute. However, Lavoisier listed the known elements, and included light and heat as elements. Asimov pointed out {717} that Lavoisier had "... era*dicated one imponderable fluid, phlogiston, but it was only partly through his influence that caloric, just as false, remained in existence in the minds of chemists for a half a century. "*

Caloric dogmatically existed in the minds of scientists deeply into the nineteenth century, until Maxwell and Boltzmann produced theories that strongly supported the concept of atoms as the building blocks of matter. In the late 1790's, Count Rumford (Benjamin Thomson) showed that the supplay of heat that could be "driven from matter" was unlimited, which totally contradicted the caloric theory {718}. Yet the notion of caloric persisted until Maxwell's theory eventually killed it, once application of that theory showed that heat was comprised of EM energy.

The fundamental ansatz of thermodynamics was formed and solidified before the advent of modern physics, and even before Maxwell's theory of electrodynamics, and before heat was known to be electromagnetic energy. It has unjustifiably retained an aura of "absoluteness" and dogma, even though its foundations concepts are based on outdated notions already scientifically falsified as far as any absoluteness is concerned. As Mallove states {719}:

> *"It is astonishing to realize that many modern conceptions (or laws') in the science of heat— thermodynamics — arose during the nineteenth century, a period of utter confusion about the fundamental nature of heat. How could it have been otherwise, given that the very existence of atoms was still in question! One sees the shakiness of the claim that the laws of thermodynamics had reached a state of 'near perfection' in the twentieth century ...when they in fact rested on this very flawed foundation. "*

During that formative period, the notion of the conservation of energy **was** also being searched for, as if in a thicket of brambles. For some time, any notion of energy conservation was seen as directly contradicted by caloric, and — in formulating the heretical notion that energy is conserved — pioneers such as Mayer {720} were severely castigated.

In 1850 Clausius resolved the conflict between caloric and conservation of energy in a seminal paper {721}.

Mallove {722} explains how the theory developed from there:

> *"In one fell swoop Clausius 'scooped' Kelvin and cast into precise form both the First and Second Laws of Thermodynamics — energy conservation, and the limit of the Carnot efficiency."*

Clausius' statement of the Second Law is: *"It is impossible for a self-acting machine, unaided by an external agency, to convey heat from one body to another at a higher temperature. "*

In 1851, Thomson (Lord Kelvin) stated the Second Law as:

> *"It is impossible, by means of inanimate material agency, to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the surrounding objects."*

In 1865 — the same year that Maxwell published his seminal paper in his electrodynamics {723} — Clausius derived entropy and showed the two laws of thermodynamics expressed in the same way as the old caloric theory. Clausius used the concept of "entropy" to represent the *"dissipated potential".* In a public speech with the title "The Entropy of the Universe Tends to a Maximum," Clausius included how thermodynamics requires an eventual heat death for the universe. His view was that

> *"The energy of the universe remains constant (first law), while entropy increases. "* {724}

We stress that the notion of the flow of EM energy through space did not originate until the 1880s, when Heaviside and Poynting formulated it independently. Neither relativity — with its notion of a combined spacetime - nor the Minkowski space had been born. Neither had quantum mechanics, quantum electrodynamics, quantum field theory, etc.

For a succinct and very useful listing of the major developments in the history of thermodynamics and statistical mechanics, one is referred to the work of Biggus {725}.

A.1.2 Inadequate Fundamental Definitions in Thermodynamics

In the view we take in this book, energy is the all-important concept, in all its forms, no matter how esoteric. The interactions of energy with matter must therefore be extended to include all known energy interactions, not just those prescribed by CTD or CED. Consequently, two fundamental definitions of thermodynamics are faulted by the supersystem concept.

First, thermodynamics defines an *open system* as one that is open to the exchange of energy and mass across its boundary. No concept or discussion of virtual energy, virtual mass, time-energy, or multiply connected space effects is included. Further, mass is actually a special form of energy, and hence this definition only prescribes the exchange of energy across the system boundary.

Second, thermodynamics defines a *closed system* as a system with no mass exchange across its boundaries, but heat (energy) exchange is permitted. This definition has been a non sequitur since general relativity was established in 1915, and also since caloric (heat) was identified as energy rather than a thin material fluid that was forced or beaten out of matter by doing work on the matter. Since then, energy and mass are also known to be the same thing, as established by general relativity. Whenever the energy of a system changes, its mass changes a priori. So a system thought of as "closed to any mass transfer across its boundaries" still undergoes mass change due to the exchange of energy across its boundary. Further, the definitions of "closed system" and "open system" are not mutually exclusive. As an example, a system exchanging energy but not mass across its boundary can be classified as a "closed system", and also as an "open system" where normal exchange of mass is blocked but the mass of the system still changes as the systems absorbs and emits energy.

Rigorously, even the energy of a system cannot individually change in 3-space, without first invoking 4-space change! Instead, the "energy x time" — the action — of the system is what moves and changes in 4-space, Observation applies a *d/dt* operator, resulting in a 3-space energy observed as having already happened or been fixed.

The modern view of gravitation is that the energy in mass generates the gravitational field, not the mass per se {726}. Also, mass is an observable, hence it is a continuously recurring 3-spatial snapshot that does not even continuously exist as mass, but alternately as mass and masstime. The assumption of the continuous existence in mass is falsified by the quantum mechanical nature of observation itself; hence all three models — CTD, CEM, and EE — are falsified.

Nonetheless, let us continue to use the conventional notion of "flow of energy" and "flow of mass" for convenience.

The thermodynamic definition of "open system" ignores the state in which mass exists as "masstime". No "change of mass", flow of mass, etc. occurs. Instead, masstime can flow or change, but mass is an observable and **thus** frozen 3-space snapshot at one instant. Time is "outside" any 3-space material system, hence when mass becomes masstime, it already "exchanges across the system boundary". *For a mass merely to continue to exist requires the exchange ofmass across the 3-space system boundary.* Indeed, only "masstime" can be changing, so in failing to consider masstime, the definition is falsified as it stands, in its very assumption of "continuous change of mass" when there is no such thing in all of nature. As stated, a similar consideration applies to energy, which does not **and** cannot "continuously exist in 3-space" at **all,** and cannot *continuously* change. Also, *time is highly concentrated energy* as we develop in this book, and time certainly "exchanges across the boundary" of any mass system. Time may be taken as a sort of "multiple connection" in space, since in theory each point in space is said to "exist simultaneously" in the same time-point.

With quantum potential connection, external energetic processes may be "superposed instantly in multiply-connected space" with one or more points of the "closed system". Hence, if there exist processes in the system to transduce time-energy into 3-spatial energy — as in the well-proven cold fusion experiments — matter changes including nuclear transmutations can and do occur. Then the "system closed against mass motional exchange through 3-space" (which does not even concretely exist) is not closed against mass changes at all.

To eliminate the mass-change effects of energy transfer across the system boundary, the present thermodynamic definition of a *closed system* must be completely recast into what thermodynamics presently calls an *isolated system* - one in which neither energy nor mass is exchanged across the system boundary. No such system exists in the universe, since merely to exist it has to be open to continual energy exchange between time and 3-space domains, mass has to continually interchange between mass and masstime, and energy has to continually interchange between energy and (energy x time). The blunt truth is that *only open systems exist in nature.*

All systems are open systems in their supersystem interactions, and they can and do exchange time-energy and mass changes due to spacetime curvatures and active vacuum changes. Any local spacetime curvature made by an energy change in a system or nearby, also reacts back upon the system from its local curved spacetime environment. It may react in a totally energetic fashion, or it may react to produce mass in the system (as in pair production). In the book we pointed out that the supersystem components external to the system can also interact in the system to eliminate mass (as when a Dirac hole combines nonradiatively with a Dirac **election).**

Finally, the notion of an "open system" as one in which energy and mass both can be exchanged across the system boundary must be extended to include a system which can also exchange time-energy, spacetime curvature energy, virtual particle flux activity and energy, and other supersystem effects across the boundary. Further, the notion that a "closed system"can be open to energy exchange but not mass is totally falsified in modern physics' virtual particle exchange between environment and system. Virtual mass — which is physically real — is continually exchanged across the system boundary. Though each virtual mass particle is individually unobservable, coherent summations are observable.

To properly evaluate the conventional thermodynamic definitions of open system, closed system, and isolated system, simply select and use a physics model containing only one fundamental unit, the joule. Now all basic entities are made of energy and functions of energy. So all basic entities are forms of energy; e.g., both mass and time are special forms of energy. As presented previously in this book, to even "exist in time" a system is continually exchanging and transducing 3-space energy and time energy, across its boundary at an enormous rate. Every system is an open system, a priori, and the thermodynamic "closed system" is a non sequitur. The thermodynamic "isolated system" is revealed as the early attempts to deal with observation; i.e., the observed system as a series of frozen 3-space snapshots (effects), each stripped from a 4-space ongoing process at a single moment by the observation process *d/dt* (LLLT) = LLL. Each snapshot is an isolated system (which is what observation produces). Now we understand that this "isolated system" existed only at a single frozen instant, continually recurring, and never existed continuously in time at all,

So in view of today's physics, the old thermodynamic "isolated system" concept is an early attempt to deal with observation of the system without input or input (without dynamics, hence frozen). "Closed system" is a non sequitur since mass is energy and energy is continuously exchanged across the system 3-boundary in several forms. Both virtual mass and virtual energy are continuously exchanged across the system boundary. It follows that the only type of system existing continuously in nature is the *open system.* The other two "types of thermodynamic systems" are at best only approximations for special cases.

Under modern physics and the new definitions, we conclude that there is no such thing as a truly closed or isolated system continuously existing in the universe. Hence that very definition disappears, *as far as any absoluteness.* It is retained only as a *useful approximation in special cases,* when the more esoteric mass effects and energy effects are intentionally omitted or negligible, and only classical effects are of importance. Any scientist dogmatically objecting to COP>1.0 EM systems on thermodynamics grounds alone, simply reveals his lack of knowledge (or lack of acceptance) of modern physics.

This is similar to the situation existing between Newtonian mechanics and relativistic mechanics. We can still use the Newtonian model for many practical matters, since the relativistic effects are often small and nearly inconsequential. Similarly, the present CTD concept of "closed system" now assumes that kind of relationship to *one class* of open systems in which the more esoteric reactions are negligible or inconsequential. But rigorously, all systems are now open systems, removing a gross non sequitur in CTD which is especially violated by COP>1.0 EM systems including every charge in the universe.

This is a correction to both equilibrium (classical) thermodynamics and disequilibrium thermodynamics. Unfortunately the two disciplines have continued to use the same (old) definitions for closed system and open system. Hence both equilibrium and disequilibrium thermodynamics at present are useful approximations, and their "laws" must not be misinterpreted as "absolute". Thermodynamics remains a very useful model, but further development is also indicated. Indeed, presently such scientific development and extension of the thermodynamic model is being done under the heading of "extended thermodynamics".

A.2 Comments on Problems with Thermodynamics

Quoting, from Kondepudi and Prigogine {727}:

> *"Every system is associated with an energy and an entropy. When matter undergoes transformation from one state to another, the total energy remains unchanged or is conserved; the total entropy, however, can only increase or, in idealized cases, remain unchanged."*

The absoluteness of that statement is falsified by the solution {12} of the long-vexing problem in CED of a source charge and its associated fields and potentials. The observable charge is a charged mass and thus a Maxwellian system. It is also a thermodynamic system, meeting the present erroneous definition of a "closed system" utilized in both equilibrium and disequilibrium thermodynamics. No observable mass crossed the system boundary, although virtual mass continually crosses it.

The charge is "fed" by a more esoteric energy input in disordered form (the virtual particle flux energy of vacuum), yet it outputs ordered observable EM energy spreading radially outward at the speed of light across the universe. The charge increases and sustains the organized energy density of a volume of space depending on the radius *ct*, where *t* is the time since the formation of the charge and *c* is the speed of light in empty space. This continous outflow of observable energy is a continuous and giant negentropy, exhibited by every source charge and dipole. When positive charge, negative charge, polarization of space, and both positive and negative energy are considered, what is spreading is precisely a reordering of the active vacuum — in short, a giant and increasing negentropy. Since the universe is filled with charges, this is a natural process continuously ongoing everywhere in the universe.

Re-quoting Serway (697}:

> *\'The second law [of thermodynamics] does not rule out the possibility of pushing heat uphill, as it were, from a cold object to a hot one, or of creating order out of disorder. It merely states that such a reversal of the natural flow requires an influx of energy... "*

Our comment is that any mechanism by which the system receives extra usable energy in forward time, or is time-reversed, enables the system to permissibly violate any narrow interpretation of the second law of thermodynamics. The second law was formulated in accord with forward time assumed, and with the old definition of "closed system", which erroneously permitted energy exchange but ignored the mass effects of energy exchange. Therefore, the second law is a special case for equilibrium in its energy exchange with its active environment, but effectively allowing its own negation anyway, and it always has. In the new approach with altered definitions, all EM systems are open systems, and they may also be brought into disequilibrium (in excited states or potentialized condition). Thus in theory all systems are permitted to violate the second law, *particularly when broken symmetries in their supersystems are involved.* This is especially true of EM systems, where all the EM field energy and potential energy in the system— even in the matter itself— comes from the active vacuum environment (particle physics view) by means of the broken 3-symmetry of the source charges and dipoles. The condition of 4-symmetry in energy flow exchange together with broken 3-symmetry and t-flow symmetry, is a disequilibrium condition and an excited state. Every charge and dipole already exhibits this state and the giant negentropy condition.

To first order, then, one may say that the present second law assumes no broken 3-symmetry or broken t-flow symmetry in energy flow, while violation of the second law assumes broken 3-symmetry and broken t-flow symmetry.

Quoting Serway {728}:

> *"Thermal equilibrium is a situation in which two objects in thermal contact with each other cease to have any net energy exchange due to a difference in their temperatures."*

Our comment is that net energy exchange — to include time-energy exchange and negative energy emission, absorption, and transduction are not modeled in thermodynamics and are arbitrarily excluded in **the** quoted statement. Since both are exhibited experimentally, the statement is falsified as far as being absolute. Instead, it is a special case and a useful approximation when some of the esoteric energy considerations are negligible.

Quoting Serway again {729} on the zeroth law of thermodynamics:

> *"If objects A and B are separately in thermal equilibrium with a third object, then A and B are in thermal equilibrium with each other. "*

A corollary is stated by Serway {730} as:

> *"...two objects in thermodynamic equilibrium with each other are at the same temperature. "*

Our comment is that these statements are true if the time-energy exchange is identical among all three. Otherwise, they are not necessarily true. It is not generally true, e.g., for a system in an accelerated frame. Also, the zeroth law is not necessarily true if negative energy (and/or its transduction is involved with one or more of the objects.

An example: A and B are at the same temperature as C, as measured by a thermometer. But A is also emitting negative energy, which is being absorbed by B, to maintain A and B at the temperature of C. A and B are not in real thermal equilibrium in their supersystem interactions, even though they are at the same macroscopic temperature. A second example: A is emitting negative energy to maintain its temperature, while B and C are not. The supersystem of A is not in real thermal equilibrium with the supersystem of either B or C, even though the observed temperatures of A, B and C are the same.

Quoting {731} on what is meant by the term "heat":

> *"...heat is treated as another form of energy that can be transformed into mechanical energy. " ... "The word heat should be used only when describing energy transferred from one place to another. That is, heat flow is an energy transfer that takes place as a consequence of temperature difference only."*

Quoting Lindsay and Margenau {732}:

> *"A complete statement of the first law comprises two* ***assertions:*** *(a) heat is a form of energy, (b) Energy is* ***conserved."***

The statement in (a) that heat is simply a form of energy is misleading, and the statement in (b) that energy is conserved must be expanded to include total energy, not just heat and the presently accounted forms of energy. Specifically virtual particle energy, time-energy, and energy from the curvature of spacetime must be included.

More correctly, heating is a form *of energy disordering* or energy scattering. As previously pointed out, entropy itself was and is conceived as *the dissipation (irretrievable loss) of potential energy.* Heat thus would be the process of energy disordering, which means that heat is a form of the process of entropy (loss of control of the energy). Note the strong non sequitur with the notion that heat is energy. *Heat cannot logically be both energy and the dissipation of energy.* The concept of energy (its mere presence) is not identical to the concept of scattering of energy or transfer of energy, but the concept of heat is identical to it. The thing is not identical to something done to the thing or to transfer of the thing to another place. A form of scattering of order is not a form of ordering. Transfer of heat is a transfer of disordering of energy, not the transfer of energy per se. Hence, considering heat as simply "heat energy" is a fundamental non sequitur of classical thermodynamics, and remains so to the present day and in the present textbooks. For an example of strong support of this position, see Romer {733}. Quoting {734}:

> *"The work done on (or by) a system is a measure of energy transfer between the system and its surroundings, whereas the mechanical energy (kinetic or potential) is a consequence of the motion and coordinates of the system. Thus, when you do work on a system, energy is transferred from you to the system. ... one can refer only to the work done on or by a system when some process has occurred in which the system has changed in some way."*

Our comment: Consider a system which "ping-pongs" the energy it receives from the environment, between two parts of itself. In each part, the form of the energy is changed by that interacting part and that does work upon the part. But the new form of energy remaining after the work was done in that component is not lost but is ping-ponged to the other component, where it does work upon that component, and so on. In this case, each joule of energy input from the environment does more than a single joule of work, with the work performed being limited only by the ability of the system to sustain and contain the "ping-ponging". There is work but no entropy if the energy doing the work (changing its form) is not lost from system control!

There is no law of "work conservation" in nature, and a joule of input energy can do more than one joule of work (as in this example) if repetitive changes of its form are considered. Thermodynamics arbitrarily excludes such cases, and hence is incomplete. Again, the conventional perception of the absoluteness of thermodynamics is falsified. It models a special case where the energy can be "used" to do work only once, and then all control of the new form of the energy (it is all still there, just in a different form) is lost.

For a particular substance, the heat capacity C is usually defined as *the amount of heat energy needed to raise the temperature of the sample being tested by one degree Celsius.* This definition assumes only one "use" of the *energy to* perform work in the substance, and completely ignores "multipass,multi-use" of the energy to perform more than one joule of work from one joule of input energy. As an example, if an equal amount of positive and negative energy is steadily input, then no amount of the positive heat energy input component would raise the temperature at all. The modern active vacuum, with a local energy density in the region of 10^{80} or more grams per cubic centimeter (expressed in mass units), nevertheless is harmless to living systems immersed in it. It has equal positive and negative energy densities, and so normally inputs the same enormous amount of positive and negative energy in its overall interactions with a living system.

The first law of thermodynamics is usually written as

$$\Delta U = U_f - U_i = Q - W. \qquad [1]$$

This is a highly special case, which does not include time energy transduction (spatial energy transduced into time energy and vice versa), or energy transduced from the active vacuum via the broken symmetry of an electrical system in continuous exchange. The definitions of work and heat have to be changed and drastically extended.

As an example of a legitimate exception, when time-energy T_E is transduced into available spatial energy inside the system, we have

$$\Delta U = U_f - U_i = Q + T_E - (W_D + W_L) \qquad [2]$$

where W_D is the energy that was dissipated in the system losses without doing useful work and W_L is the energy leaving the system but doing useful work in a load L. $(Q + T_E)$ represents the total net energy entering the system, while $(W_D + W_L)$ is the total energy leaving the system, either as losses or to do useful work in the load. Here we are also incorporating virtual energy under the title of time-energy, since both time and virtual energy are unobservable, and virtual energy "exists in time".

Now suppose Q entering the system, and collected for use in it, is freely furnished from the active vacuum via the broken symmetry of the electrical system's dipolarity. Suppose T_E is spatial energy freely transduced from the time domain by ongoing processes in the system, and thus T_E also enters the system freely. Then suppose that the system processes are deliberately adjusted by the operator so that

$$(Q + T_E) = (W_D + W_L) \qquad [3]$$

In that case,

$$\Delta U = U_f - U_i = 0 \qquad [4]$$

yet continuous loss energy WD is escaping from the working system, useful work W_L is continuously being done in the load, and useful energy $(Q + T_E)$ is continuously entering the system from the active environment (the active vacuum and the local curvatures of spacetime). This system is totally permissible in nature (and is already operating in every charge and dipole in the universe). It is an example of an open system far from equilibrium in *3-space* energy flow, but in equilibrium in *4-space* energy flow. Yet it is artificially excluded by classical thermodynamics.

Classical thermodynamics thus excludes every charge and dipole in the universe, and therefore is an extraordinarily curtailed model. It is a special case approximation having no absolute authority as a proper limiting model of EM systems.

The first law of thermodynamics for infinitesimal changes may be stated as

$$dU = dQ - dW \qquad [5]$$

Following the previous reasoning, for involvement of time-energy transduction (with virtual energy included) we have

$$dU = dQ + dT_E - (dW_D + dW_L) \qquad [6]$$

If dQ and dT_E are freely received in the system from the active vacuum and local curvatures of spacetime respectively, and if

$$dQ + dT_E = dW_D + dW_L \qquad [7]$$

then

$$dU = 0 \qquad [8]$$

So with no change in its internal energy at all, the microsystem or microprocess is receiving excess energy continuously, and also is continuously emitting energy that is doing useful work with some losses. The system is in fact in *energy 4-flow* equilibrium, and doing useful work, without the operator inputting any of the energy.

So the *corrected* standard statement of the first law allows an equilibrium system to steadily produce output work W, if its "energy input Q" is freely received. The first law of thermodynamics, properly extended and reinterpreted, does indeed admit "electrical windmills" operating in free electrical winds.

Now turning to the second law. In effect, the second law prescribes those processed in nature — more exactly, in the prescription of the first law — that may or may not occur. It is a law of limitation. It states that, of all the energy conversion processes permitted by the first law, only certain types of energy conversions can take place. A priori, the validity of its prescription depends totally on the nature of the functions involved in the system considered! However, since we dramatically extended the first law, using neww processes experimentally determined to take place, then we have automatically extended and changed the second law because we used experimentally *proven* functions. The old second law is now revealed as a special case, with no absolute limitations. The perception that it prescribes absolute limitations has been falsified. It is still useful, however, but to be careful one must completely re-examine (and possibly rework) the second law whenever the energy functions of a system are changed. We strongly stress that *the second law is a conditional law, and its application is totally dependent upon the system energy functions behaving in the limited manner assumed by the second law. Whenever the energy functions of the system are changed outside those assumed by the second law, the second law is permissibly invalidated for that system.* In short, in that case the second Iaw is simply a totally inadequate model for that system, and cannot be used as any sort of criterion for stating the "permissibility" or "impossibility" of that system. As an analogy, a model specifying scalars only cannot be applied to a vectorial system as a decisive logical criterion or restriction.

Quoting Seray {735}:

> *"From an engineering viewpoint, perhaps the most important application [of the second law] is the limited efficiency of heat engines. Simply stated, the second law states that a machine capable of continuously converting*

> *thermal energy completely into other forms of energy cannot be constructed."*

Our comment is that, if we believe that energy is conserved in the entire universe, then the universe itself already violates Serway's statement. If we consider the seething virtual particle flux energy of the active vacuum as a special form of thermal energy, then every charge and dipole in the universe already is a heat engine that falsifies the second law of thermodynamics. The charge and dipole continuously absorb this special form of thermal (virtual flux) photon energy, and also continuously convert it into real, observable photon emission in 3-space. This is negentropy on a giant scale, already demonstrated by every charge and dipole and thus experimentally falsifying the present second law.

Classical thermodynamics also assumes irreversibility. While many of the physics models and equations prescribe reversible reactions, thermodynamics takes the view that events occur with a much greater probability in one direction in time than in the other. This of course is usually borne out in our perception of the gross observable universe around us. However, charge conservation alone requires equal amounts o| charge and time-reversed charge. With the nearly ubiquitous time-reversed charge and charge conservation, it follows that some factor not accounted for in thermodynamics is operating in the universe. And so it is, but it is a general relativistic factor. We shall use a biological example of the cellular healing mechanism we discovered in the living body {736}, extending Becker's work and Priore's work.

Normally most processes do not involve deliberate net transduction of time-energy into spatial energy, nor do we model and account for the specific set of ongoing interactions between every part of the system **and** its concomitant set of spacetime curvatures acting back upon it. Call that set of spacetime curvatures the *resident spacetime curvature engine **and its** dynamics,* or just *its engine* for short. The engine is comprised of structured energy and energetics, and it is already experimentally demonstrated in nonlinear phase conjugate optics that energy can be **time** reversed. The engine's spacetime curvatures act directly upon the **mass**-energy dynamics of the system. Time-reversing the system's operation is a matter of producing a sufficiently amplified, phase conjugated engine to act in reverse upon the physical system and overpower its resident **engine.** In that case, the thermodynamic system and its functioning can be **made to** run backward.

In fact this is the fundamental mechanism used by the body's own cellular regenerative system to heal its damaged cells. To a limited degree, the body is able to "pump" its cells in the time-domain (with time-energy), thus producing an amplified "phase conjugate energy and engine dynamics".

Pumping a living cell with time-energy is interesting. The resident engine serves as the "signal input wave" in conventional nonlinear phase conjugate optics, and the output time-reversed engine (the amplified anti-*engine*) serves as the "phase conjugate replica wave". A sufficiently simplified anti-engine acts upon the cell and all its tiniest parts, time-reversing all energetic operations in exact phase, proportion, and timing. This is the mechanism for thermodynamic reversibility. It is also the "secret" to eventually reversing aging.

This action is upon all the energetics of the cell and its parts, including the mass-energy and mass-energetics as well.

The result of the action of the anti-engine is to slowly time-reverse the damaged cell and all its parts back to a previous state and dynamics. Electromagnetic amplification of the process was unwittingly demonstrated by Priore {737} and by Becker {738}. We refer the reader to our website {739} and to our "porthole briefing" for pertinent technical information, and for a most important extension and great simplification of the Priore/Becker approach and methodology.

Present classical thermodynamics does not consider or model the reaction of local spacetime curvatures and their dynamics upon the system. Hence all time-domain pumping that occurs is occurring rather randomly in each and every microscopic part of the system. In short, there is no organized anti-engine, and an organized anti-engine is what is required for time-reversing the system's operations back to a previous state. However, by mimicking, the body and pumping the physical process in the time-domain, with sufficient pumping a precise, amplified anti-engine will be formed and act upon the system — *any* system. The applicability is limited only by the state of development of the appropriate time-pumping technology.

In theory (and as experimentally demonstrated unwittingly by Priore and by Becker), this anti-engine can and does overpower the resident engine. It can in fact time-reverse the pumped physical processes in the system itself.

So the classical thermodynamics assumption of irreversibility for heat engine energy processes is not absolute but only provisional. Specifically, it assumes no overall time-domain pumping process that provides a precise anti-engine and time reversal of the system energetic process of concern and its dynamics.

If the second law is extended to encompass and model the new concepts such as broken symmetry and time-energy, then the extended second law will encompass reversible heat engine processes, Maxwell's demons, COP>1.0 electrical systems, COP» 1.0 electrical systems (electrical windmills), time-reversal of the dynamics, reversibility of that which has been irreversible, etc.

As given by Kondepudi and Prigogine, the third law of thermodynamics essentially states {740} that

> "... *at the absolute zero of temperature the entropy of every chemically homogeneous solid or liquid body has a zero value."*

This is also known as Nernst's heat theorem, after W. Nernst who proposed it in 1906. The third law states that

$$S \rightarrow 0 \text{ as } T \rightarrow 0 \qquad [9]$$

Certain violations of the third law are already known. E.g., C=O: carbon monoxide molecules can align in various ways in a crystal, so there is entropy associated with C=O crystals at 0°K. In this case, A and B are equally probable and the following equation can be used to calculate the entropy (called residual energy):

$$S = nR\,[1/2\ln(1/2) + 1/2\ln(1/2)] = -nR\ln 2 \qquad [10]$$

From the new viewpoint, the definition of T = 0° is open to change. Rigorously, the present third law assumes no local broken 3-symmetry of the vacuum's flux exchange with the system, and it also assumes no broken time-energy flow symmetry. In the new approach, both these symmetries can and are broken by every charge while a higher 4-symmetry is maintained. This leads to a giant negentropy, so that the "calculation of entropy" via Nernst's theorem fails. A fundamental charge at T = 0° thus continues to exhibit giant negentropy, in total violation of the third law of classical thermodynamics. Hence the third law is also devoid of any absoluteness, and it is reduced to a special case statement. It remains useful where the more esoteric energy interactions are negligible.

We have now demonstrated (by citing actual substantiating experiments) that all the laws of CTD are falsifiable insofar as being absolute. Hence they have *no* absoluteness. Instead, they are special case models and "approximation laws" only. They remain very useful and applicable, However, so long as the more esoteric energy interactions of the supersystem do not play a major role in the system operation.

We conclude that the present form of classical thermodynamics and disequilibrium thermodynamics are special case models — as indeed is every model, by Godel's theorem and proof alone — with no absoluteness. They do not prohibit COP>1.0 EM systems, nor can they necessarily be used to rule out COP>1.0 EM systems.

This also falsifies all current arguments that COP>1.0 EM systems violate either the first, second, or third law of classical thermodynamics. Calling COP>1.0 systems "perpetual motion machines of the first, second, and third kind", according to which law they supposedly violate, is a total non sequitur. Indeed, every original charge in the universe has already demonstrated just such "perpetual motion" for some 14 or so billion years, and happily violates all three laws of the present CTD. Every charge is in violation of the "laws" of the present models of CTD, CED, and EE, and it is still in motion and still freely pouring out EM energy. We have no abillity to examine the notion of "perpetual" past the entire past existence of the universe.

All three models examined — CTD, CED, and EE — are seriously out of date and in need of revision and extension, because any absoluteness of their predictions and assumptions is already falsified by actual experiments and examples. Some work is already ongoing to extend thermodynamics, but is not nearly extensive enough. A thorough overhaul of the very foundations is what is needed for all three models. The three models remain useful as special case approximations, but their received use to arbitrarily rule out COP>1.0 EM systems is dogma, not science because such use is already experimentally refuted by every charge in the universe.

Appendix B
Maxwell's Demon: Alive and Thriving

Introduction

ITHERE is a long history of discussion over "Maxwell's demon", a fictitious being that Maxwell proposed to "sort" molecules striking a barrier between two volumes of gas initially at the same temperature. By intentionally opening a trap door between two for hotter, more energetic molecules to pass through from the left side to the right side, but shutting the door for lower lemperature molecules so that they are retained on the left side, eventually a temperature gradient would "freely" exist between the two reservoirs. In theory, then, if one had the demon working "for free", one would be able to utilize the temperature difference between the reservoirs to perform free work. Then one could repeat the demon's sorting procedure, to obtain more free work again. And so on.

The prevailing discussions have ignored the main feature: the fact that the demon is part of the "external environment" (i.e., not under control of the system, and operating independently of any energy input by the operator). Maxwell's demon represents an analogy to the system's environment furnishing free energy (free ordering and sorting). A lot of breath and mental gyrations have gone into the discussions over the years, with very little concrete experimental examination. The prevailing concept in science, which still adheres to the "absoluteness" of the second law of thermodynamics, is that Maxwell's demon is impossible.

Recently, the former editor of *Nature,* Dr. John Maddox, wrote a very clear emeritus editorial on Maxwell's famous demon, repeating the standard conclusion that such is not possible. Since there are specific examples in physics of what can be called "experimentally proven Maxwell's demons," the present author prepared a respectful commentary to *Nature* to refute the standard erroneous thesis in the interest of science. since experiment is supposed to be decisive in physics, we hoped that citing overwhelming experimental proof would be sufficient. It wasn't.

Almost symbolically (i.e., "free the demon!"), I submitted the commentary to the editor by I -mail on the evening of July 4, 2002. The very next day I received rejection (not too unexpected) by E-mail with no further discussion accepted. So I wrote a polite note in response, and then posted my commentary note on my website.

In this appendix, I include my E-mails of both 4 July and 5 July 2002, and then include the commentary that was submitted. The reader can read the cited emeritus editorial in *Nature,* and then decide for himself or herself whether experiment still rules the day at *Nature* where COP>1.0 and COP = co EM systems are concerned.

E-mail to *Nature* 4 July 2002, submitting the Commentary.

Dear Editor:

Attached is a significant commentary on John Maddox's excellent emeritus editorial on Maxwell's demon, titled "Slamming the Door," *Nature,* 417, 27 Jun. 02, p. 903.

Forms of Maxwell's demon are very much alive and thriving, and experimentally confirmed, as we develop very clearly in this commentary. They are unwittingly used widely in electrodynamics, though seldom recognized.

Most sincerely,
Tom Bearden, Ph.D., Magnetic Energy Ltd.

E-mail to *Nature* 5 July 2002, submitting a comment on the rejection.

Dear Ms XXX:

Thanks for the prompt reply.

Since *Nature* will not be publishing the note, I have now posted it on my website (some 80,000 hits per day from all over the world). It can be seen at http://www.cheniere.org/articles/maxwells%20demon.htm

This is a respectful article, and I just thought the experimental information needed getting out there to the scientific community. I also understand you cannot discuss it further. I only wish to note that it is based on cited replicable experiments, which are supposed to be the decisive thing in science and scientific method. It seems a bit odd that you would not publish the results of several replicable and replicated experiments that falsify a prevailing theoretical thesis and conclusion. Nonetheless, I respect the wishes of *Nature,* and I do enjoy the Journal each week.

Best wishes,
Tom Bearden, Ph.D.

Within days of this rejection, the paper by Wang et al. {741} was published, extending the known physics "Maxwell's demon" regime (i.e., where the second law of thermodynamics is known to be violated) from the atomic scale and very short periods of time, to the micron scale in size and time up to a few tenths of a second.

Here is the rejected commentary I submitted to *Nature.*

Maxwell-type demons are alive and thriving

The recent emeritus editorial by Maddox on Maxwell's demon [742] is a short overview of one historical aspect of would-be negentropy, but much greater thought on the subject needs to be given. Indeed, Maxwell's demon is alive and thriving, but just in a different form not requiring a nimble-fingered microscopic being. We now discuss such experimentally proven demons.

Consider the prevailing ansatz that Maxwellian systems outputting more EM energy than the energy input by the operator or experimenter are impossible. Such proposed systems are said to constitute examples of forbidden perpetual motion machines because they violate the second law of classical (equilibrium) thermodynamics. That prevailing ansatz has been falsified in classical electrodynamics since the 1880s, and in particle physics for 45 years, but eerily it is still the prevailing scientific opinion nonetheless.

The ansatz can be experimentally disproved in classical electrodynamics as follows: Lay a charged capacitor or electret on a permanent magnet so the E-field of the capacitor or electret is at right angles to the H-field of the magnet. That optimizes $\mathbf{S} = f(\mathbf{E} \times \mathbf{H})$, and hence optimizes the continuous flow of Poynting EM energy from that simple device, even though the fields, seem to be "static". As Buchwald states [743], *"[Poynting's result] implies that a charged capacitor in a constant magnetic field which is not parallel to the electric field is the seat of energy flows even though all macroscopic phenomena are static. "* [744] This simple device will freely pour EM energy indefinitely, so long as it remains intact.

We consider a special but universal kind of related Maxwell's demon problem: the vexing problem of the source charge, sometimes called "the most difficult problem in classical and quantum electrodynamics." [745] Any charge pours out observable EM energy continuously in all directions, at the speed of light, with no *observable* EM energy input. This continuous flow of EM energy establishes the charge's related fields and potentials and their energy, reaching even across the universe for very old charges. In the classical electrodynamics (CED) model and therefore in electrical engineering, there is no solution to this problem because CED does not model *nonobservahle* EM energy inputs such as the virtual particle flux exchange of the charge with its vacuum environment.

With no solution available in CED, then the present CED and electrical engineering assume that every charge in the universe *freely and continuously creates energy out of nothing,* and pours it out in all directions, forming its associated fields and potentials and their energy. Either we must solve the problem in the classical EM model, or totally surrender the conservation of energy law in the model in its present limited form. Or — as is presently the case — we may continue to ignore it, as has been done during the 45 years since the proof of broken symmetry in particle physics [746].

On the other hand, particle physics models and utilizes virtual energy **in** the seething vacuum, including the interactions of the vacuum with charge, With the 1957 experimental proof of broken symmetry by Wu et al., [746] certified by the award of the Nobel Prize to Lee and Yang later that **same** year, one of the proven broken symmetries in the vacuum virtual energy exchange is the asymmetry of opposite charges, such as are on the ends of a dipole. For a dipole or dipolarity, then by the very definition of broken symmetry *something virtual has become observable.* We diverge for a moment, and then will return to this proven asymmetry of the opposite charges of the dipole.

If we take the modern view of the bare charge clustered by virtual **charges** of opposite sign, then the magnitude of the bare charge in the middle is infinite (including its energy) and so is the magnitude of the charge of the surrounding clustering virtual charges. [747] So for an "isolated **observable** charge" we actually have a sort of "infinitely powerful dipole". Yet **the** difference between these two infinite values of charge is finite; it is routinely calculated as the observed charge of an elementary charged particle, the value that is listed in standard texts and handbooks.
The asymmetry of this "opposite composite charges" model of the "isolated observable charge" now explains the long-vexing source **charge** problem. The observable "composite" charge continuously absorbs **virtual** photon energy from the seething vacuum exchange, transduces it **into** observable photon energy excitation, and this excited state **continuously** decays to emit real observable photons in all directions at the speed of light. When *virtual* energy input as well as *observable* energy output is accounted, the charge's proven asymmetry in the vacuum exchange **makes** the source charge an open system far from equilibrium in its active

environment. As an open disequilibrium system the charge is thermodynamically permitted to (1) self-order (hence coherently integrate absorbed virtual photon energy into observable photon energy), (2) self-oscillate or self-rotate (spin), (3) output more energy than the "operator" **inputs** (the operator inputs none at all), and (4) exhibit negentropy. Every charge in the universe is already doing those four functions. It is also a system having a COP (coefficient of performance, or useful energy output divided by the operator's energy input) of COP = ∞, since the operator inputs nothing.

In short, every charge in the universe already is a special kind of known "Maxwell's demon", whose operational mechanism is long since proven in particle physics but does not appear in classical electrodynamics or electrical engineering. It is a dependable Maxwell's demon, since all charges in the original matter in the universe have been pouring out real EM energy freely, using this asymmetry mechanism, for some 14 billion years, **and** they continue to do so. Further, any charge one assembles in the laboratory instantly starts up its Maxwellian demon performance automatically, and it will continuously pour out EM energy in all directions indefinitely, so long as the charge remains intact. So this demon is easily created and demonstrated experimentally. The Maxwell's demon for freely extracting copious EM energy from the vacuum is one of the easiest entities to invoke in all of physics

This is **not** a violation of thermodynamics, since classical *equilibrium thermodynamics* with its infamous second law does not apply because the charge is far from equilibrium in its exchange with its active vacuum environment. Although the COP = ∞, the *efficiency* of the charge system -defined as the useful energy output divided by the total energy input from all sources — is never more than 100%. Hence this is not a perpetual motion machine, and the charge rigorously does not create energy; it only transduces energy input to it in a novel form. In this case, the charge can permissibly act as if it were a free electrical windmill turning in a novel free electrical wind due to its asymmetry in the vacuum flux.

So every charge in the universe exhibits COP = ∞, clearly proving that Maxwell's demon in an improved form exists and is thriving throughout the universe. There would be no EM fields and potentials and their energy, and hence no electrical circuits or power systems, were it not for this asymmetry of the charge in its virtual energy exchange with the vacuum, making the charge a true Maxwell's demon.

In the usual CED model and electrical engineering, the 1867 L. V. Lorenz symmetrical regauging (later credited to H. A. Lorentz) [748] of the equations carefully selects only those Maxwellian systems that have two simultaneous equal and opposite disequilibria with their active environment. This assumes that the potential energy of the system is freely changed twice (by gauge freedom), but only in such highly selected manner as to form two equal and opposite new "free" force fields in the system. These two force fields perform *internal work* inside the system continuously, increasing its stress (and its stress potential, thereby curving local spacetime). Since there is no *net resultant* force field, the two force fields are unable to translate electrons in the circuit to do free *external* work in the external load. So the symmetrically regauged Maxwellian system has been altered: its energy has been changed twice, forming a stress potential inside it and increasing and maintaining that stress on the system, and the frame of the system has been somewhat rotated out of **the** laboratory frame. The symmetrically regauged Maxwellian system most decidedly is not identical to the system prior to regauging.

We note merely that the common Lorentz symmetrical regauging unwittingly assumes two Maxwell's demons of yet different kind, each freely fueled from the local vacuum environment, and each continuously performing internal work upon the system to produce and maintain system stress.

Gauge freedom is thus revealed as a special kind of Maxwell's demon, since (1) it assumes that the potential energy of a system can be freely changed at will at any time, and (2) unless we abandon the conservation of energy law, that energy change must have involved energy exchange **with** the external environment. So the gauge freedom axiom of quantum field theory assumes two specialized Maxwellian demons — i.e., mechanisms for transfer of energy between environment and system, so as to continuously perform internal work upon the system. In this case the two demons are equal and opposite, and fight each other to a draw insofar as performing any useful external work.

Oddly, in present electrical circuits the ubiquitous but *arbitrary* closed current loop circuit — passing all spent current from the external circuit back through the source dipole in the generator against the dipole's **back** emf— self-imposes Lorentz regauging of excitation discharge in the circuit and prevents COP>1.0 functioning. This is not required by thermodynamics in general! Indeed, several areas are already known to violate present thermodynamics. Sharp discharges (strong gradients), for example, are known to violate it. [749] Other known areas where thermodynamics is violated include rarefied media, and anomalous memory effects in materials. Modern research is being conducted in such areas under the heading of "extended thermodynamics". [750].

Since experiment and not theory is primary in science, let us consider some additional actual Maxwell demon experiments. The well-known "negative resonance absorption of the medium" produces more output energy than the operator inputs to the experiment, with hundreds of these experiments done every year routinely by many nonlinear optical laboratories. As an example, Bohren's version of such an experiment [751] commonly outputs 18 times as much EM energy as the operator inputs, exhibiting COP = 18. Independent replication of Bohren's work by Paul and Fischer [752] is reported in the same journal issue.

Added to the charged capacitor lying on a permanent magnet, there thus exist plenty of proven, replicated experiments which can be easily performed to demonstrate a Maxwellian demon operating in a Maxwellian system and freely producing COP>1.0, or even COP = oo. Because of the demon (the operational mechanism), these experiments all involve open systems far from equilibrium with their active environment. Hence they are not limited by the second law of classical equilibrium thermodynamics.

In short. Maxwell's demon is very much alive in many forms and is thgiving after all. It has just been hiding in different form than what is usually suspected.

T.E. Bearden, Ph. D.
Magnetic Energy, Ltd.

ENDNOTES

1. (a) Albert Einstein, "Foreword," in (b) Max Jammer, *Concepts of Space: The History of Theories of Space in Physics,* Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

2. George Zebrowski, "The holdouts," *Nature,* Vol. 408, 14 Dec 2000, p. 775; expressing the thinking of Stephen Hawking, one of the leading physicists of our day.

3. Mario Bunge, *Foundations of Physics,* Springer-Verlag, New York, 1967, p. 176.

4. (a) J. H. Poynting, "On the transfer of energy in the electromagnetic field," *Phil. Trans. Roy. Soc. Lond,* Vol. 175, Part II, 1884, p. 343-361; (b) — "On the Connection Between Electric Current and the Electric and Magnetic Inductions in the Surrounding Field," *Phil. Trans. Roy. Soc. Lond,* Vol. 176, 1885, p. 277-306.

5. (a) Oliver Heaviside, "Electromagnetic Induction and Its Propagation," *The Electrician,* 1885, 1886, 1887, and later — a series of 47 sections, published section by section in numerous issues of *The Electrician* during 1885, 1886, and 1887; (b) —*Electromagnetic Theory, 3 vols., B*enn, London, 1893-1912. Second reprint 1925. An unabridged edition of all three volumes, with a critical and historical introduction by **E. Weber,** was published by Dover, New York, 1950. See particularly (c) - "On the Forces, Stresses, and Fluxes of Energy in the Electromagnetic Field." P*hil Trans. Roy. Soc. London,* 183A, 1893, p. 423-480. In this prestigious publication, Heaviside discusses the Faraday-Maxwell ether medium, outlines his vector algebra for analysis of vectors without quaternions, discusses magnetism, gives the EM equations in a moving medium, and gives the EM flux of energy in a stationary medium. On p. 443, he credits Poynting with being first to discover the formula for energy flow, with Heaviside himself independently discovering and interpreting this flow a little later by himself in an extended form.

6. E.g., a polite debate has continued at intervals, for about 30 years, in the Am. J. P*hys.* alone.

7. Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics,* Addison-Wesley, Reading, MA, Vol. 1,

8. *Ibid.,* Vol. 1, 1964, p. 12-2.

9. *Ibid.,* Vol. 2, 1964, p. 2-1.

10. John D. Kraus, *Antennas,* Second Edition, McGraw Hill, 1988.

11. Raymond A. Serway, *Physics for Scientists and Engineers with Modern Physics,* Third edition, updated version, Saunders College Publishing, Philadelphia *et al,* 1990, p. 620.

12. T. E. Bearden, "Giant Negentropy from the Common Dipole," J. *New Energy,* 5(1), Summer 2000, p. 11-23. Also on website www.cheniere.org.

13. (a) T. D. Lee, "Question of Parity Conservation in Weak Interactions," *Phys. Rev.,* 104(1), Oct. 1, 1956, p. 254-259; (b) — with Reinhard Oehme and C. N. Yang, "Remarks on Possible Noninvariance under Time Reversal and Charge Conjugation," *Phys. Rev.,* 106(2), 1957, p. 340-345; (c) — and C. N. Yang, "Parity Nonconservation and a Two Component Theory of the Neutrino," *Phys. Rev.,* 105(5), Mar. 1, 1957, p. 1671-1675.

14. C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes and R. P. Hudson, "Experimental Test of Parity Conservation in Beta Decay," *Phys. Rev., Vol.* 105, 1957, p. 1413.

15. We published it crudely two years later. See T. E. Bearden, *Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena,* NTIS Report AD-763-210, 1973.

16. A time-reversal zone is a momentary tiny region of the fluid wherein antiphoton reactions with the main ions predominate over **photon** reactions.

17. See T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," *J. New Energy,* 3(2/3), 1998, p. 12-28. A slightly expanded paper by the same title is published in *Explore,* 8(6), 1998, p. 7-16.

18. E.g., see Xiaodong Chen, "A New Interpretation of Quantum Theory — Time as Hidden Variable," Los Alamos National Laboratory web site, arXiv:quant-ph/9902037 v3, 11 Feb. 1999. Using two more **time** variables as the quantum hidden variables, Chen derives Dirac's field equation under the principle of quantum physics, and then extends **the**

method into quantum fields with arbitrary spin number. Chen's model offers very interesting unifications of quantum mechanical behavior.

19. F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984, Revised Edition 1993, Chapter 5. This is a deeper coverage of the photon polarizationd. Mandl and Shaw argue that the longitudinal and scalar (time-polarized) polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. This argument, translated from particle terminology to wave terminology, directly fits and strongly supports my reinterpretation {12} of Whittaker's 1903 wave decomposition of the scalar potential {85}. The reinterpretation also represents the discovery of the giant negentropy process favored by nature, easily evoked by breaking 3-symmetry (as by forming a simple dipole).

20. James Clerk Maxwell, "A Dynamical Theory of the Electromagnetic Field," *Roy. Soc. Trans.,* Vol. CLV, 1865, p 459. Read Dec. 8. 1864. Also in *The Scientific Papers of James Clerk Maxwell, 2* vols, bound as one, edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-597. In this seminal paper, Maxwell's equations of the electromagnetic field are given in "Part III, General Equations of the Electromagnetic Field," p. 554-564. On p. 561, Maxwell lists his 20 variables. On p. 562, he summarizes the different subjects of the 20 equations, being three equations each for magnetic force, electric currents, electromotive force, electric elasticity, electric resistance, total currents; and one equation each for free electricity and continuity.

21. B. J. Hunt, *The Maxwellians,* Cornell University Press, 1991 gives the history of the alteration of Maxwell's field theory by Fitzgerald, Heaviside, Lodge, and Hertz. In an appendix, Hunt makes a comparison between the standard expressions of the Maxwell field equations and their expression in quaternionic form.

22. E.g., J. D. Jackson, *Classical Electrodynamics,* 2nd Edition, Wylie, 1975, p. 219-221.

23. Willis E. Lamb Jr. and Robert C. Retherford, "Fine structure of the hydrogen atom by a microwave method," *Phys. Rev.,* 72(3), Aug. 1, 1947, p. 241-243. Lamb received the 1955 Nobel Prize in physics jointly with Polykurp Kush for experiments measuring the small displacement later called the "Lamb shift" of O+ne of the energy levels in atomic

hydrogen. This is a small "vacuum energy" effect, but the *energy density* involved is greater than the surface energy density of the sun.

24. Craig F. Bohren, "How can a particle absorb more than the light incident on it?" *Am. J. Phys.,* 51(4), Apr. 1983, p. 323-327. Under nonlinear conditions, a particle can absorb more energy than is in the light incident on it. Metallic particles at ultraviolet frequencies are one class of such particles; insulating particles at infrared frequencies are another.

25. See also H. Paul and R. Fischer, "'(Comment on "How can a particle absorb more than the light incident on it?')," *Am. J. Phys.,* 51(4), Apr. 1983, p. 327 for a replication and validation of Bohren's experiment,

26. H. B. G. Casimir, "On the attraction between two perfectly conducting plates," presented at a meeting of the Royal Netherlands Academy of Arts and Sciences on 29 May, 1948. Published in the same year in *Proc. Koninklijke Nederlandse Akad. van Wetenschappen,* Amsterdam, vol. 51(7), 1948, p. 793-796. In English. It is an extrapolation of the more detailed paper of Casimir and Polder, 1948.

27. S. K. Lamoreaux, "Demonstration of the Casimir Force in the 0.6 to 6 micro m range," *Phys. Rev. Lett.,* 78(1), Jan. 6, 1997, p. 5-8. Lamoreaux gives a rigorous experimental verification of the Casimir effect.

28. Franklin B. Mead and Jack Nachamkin, "System for Converting Electromagnetic Radiation Energy to Electrical Energy," U.S. Patent No 5,590,031. An invention for extracting energy from the vacuum via the proven Casimir effect.

29. (a) Floyd Sweet and T. E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," *Proc. 26th Intersociety Energy Conversion Eng. Conf. (IECEC '91),* Boston, Massachusetts, 1991, p. 370-375. Sweet's device produced 500 watts output from a 33 microwatt **input** by the operator. A highly successful anti-gravity experiment designed by the present author was also performed, and is reported in the paper. Unfortunately, Sweet later died and never fully revealed the activation secret by which barium ferrite magnetic materials could be in self-oscillation at 60 Hertz. Weak self-oscillation of such permanent magnetic materials at higher frequency is now known, of course. E.g., see (b) V. E. Zakharov and V. S. L'vov, "Parametric excitation of spin waves in ferromagnets with magnetic inhomogeneities," *Sov. Phys. Solid State,* 14(10) 1973, p. 2513. Particularly see (c) V. S. L'vov, *Wave Turbulence Under Parametric Excitation: Applications to Magnets,* Springer Series in

Nonlinear Dynamics, Springer-Verlag, New York, 1994. Self-oscillation can in fact be induced in permanent magnet materials. We comment that, while a permanent magnet is in self-oscillation, a simple set of wires or a coil around the magnet will receive EM energy by induction, and can power a very tiny load in a closely adjacent circuit.

30. T. E. Bearden, "Bedini's Method for Forming Negative Resistors in Batteries," *J. New Energy,* 5(1), Summer 2000, p. 24-38. Also carried on restricted DoE website http://www.ott.doe.gov/electromagnetic/ and on the present author's open website http://www.cheniere.org.

31. Teruo Kawai, "Motive Power Generating Device," U.S. Patent No. 5,536,518.Jul. 25, 1995. My board of directors and I were physically present when the Kawai system, company, etc. were seized by what can only have been the Yakuza (Japanese mafia), even though Kawai was being backed by one of the wealthiest investors in Japan. I had just reached an agreement with Kawai to produce and market his overunity motor worldwide. Several legitimate Japanese overunity systems, ready for production, have been stopped in this fashion. Kawai's process, applied with efficient switching such as photon-coupled switching, essentially doubles the COP of the motor to which it is applied. If one starts with a high efficiency (e.g., Hitachi) permanent magnet electric motor of 0.7 or 0.8, one can expect to produce COP = 1.4 and 1.6 respectively, the actual results obtained by Hitachi tests of two Kawai-modified Hitachi motors. We accent that the efficiencies of the motors remain less than 100% at all times.

32. Gabriel Kron, "Electric circuit models of the Schrodinger equation." P*hys. Rev.* 67(1-2), Jan. 1 and 15, 1945. Quoting, p. 41: *""Now a value E of the negative resistances, at which the generator current becomes zero, represents a state at which the circuit is self-supporting and has a continuous existence of its own without the presence of the generator, as the negative resistances just supply the energy consumed by the positive resistances. ... When the generator current is positive the circuit draws energy from the source, and when the current is negative the circuit pumps back energy into the source. At zero generator current the circuit neither gives nor takes energy, and theoretically the generator may be removed."* In this passage, Kron was required to insert the word "theoretically". Kron was the chief scientist for General Electric on the Network Analyzer project at Stanford University before WW II. In the 1930, he did successfully build negative resistors (at least three), and he

did successfully power the network analyzer with them. He was never permitted to release the technical details, although he placed judicious hints in several of his papers, and slipped certain key phrases past the censors. Interestingly, Sweet was his protegee, and it is conceivable that the basis for Sweet's device may have been Kron's original negative resistor. Independent tests and verification of the Sweet device were performed by Walter Rosenthal, a retired professional test engineer of considerable experience.

33. Shoukai Wang and D. D. L. Chung, "Apparent negative electrical resistance in carbon fiber composites," *Composites, Part B,* Vol. 30, 1999, p. 579-590. Apparently the referees required the replacement of the term "negative electrical resistance" by the term "apparent negative resistance". In the paper, simply examine the actual results shown in the curves of the negative-going current, against the voltage, with the line passing directly through the origin. That is a true negative resistance, not a differential one or an "apparent" one. A patent was filed on this invention.

34. (a) T. H. Moray, *The Sea of Energy,* 5th Edn., Salt Lake City, 1978. The Moray device weighed 55 pounds and produced 50 kilowatts of power in the late 1930s, and was demonstrated to scientists and engineers numerous times, Moray made tubes which actually contained very special semiconductors, three of them in contact, each pressed at high pressure during a sintering process. A very eerie chain of circumstances eventually destroyed the device. Moray rode in a bulletproof car, and he was sometimes fired at from the sidewalk or an alley. A double agent, working for both the U.S. and the Soviet Union, infiltrated his lab and eventually destroyed the device. The Morays were never given a proper scientific hearing and were never funded for full development. T. Henry Moray was shot and wounded in his own laboratory, wrestling with a would-be assassin. Yet so many expert engineers and scientists witnessed and tested the machine in operation that there is no question of its legitimacy. Unfortunately, the reference cited is not at all a satisfactory scientific examination of the Moray device. A related U.S. patent was issued which at least shows some details of the Moray tube: See (b) Thomas H. Moray, "Electrotherapeutic Apparatus," U.S. Patent No. 2,460,707, Feb. 1, 1949. I personally received a treatment on a Moray device for severe pain, and the relief was rapid and remarkable.

35. See (a) William B. Burford III and H. Grey Verner, *Semiconductor Junctions and Devices: Theory to Practice,* McGraw-Hill,

New York, 1965, Chapter 18: Point-Contact Devices. The very first transistor — the point contact transistor — also often exhibited negative resistance. Quoting, p. 281:

> *"First, the theory underlying their function is imperfectly understood even after almost a century..., and second, they involve active metal-semiconductor contacts of a highly specialized nature. ... The manufacturing process is deceptively simple, but since much of it involves the empirical know-how of the fabricator, the true variables are almost impossible to isolate or study. ... although the very nature of these units limits them to small power capabilities, the concept of small-signal behavior, in the sense of the term when applied to junction devices, is meaningless, since there is no region of operation wherein equilibrium or theoretical performance is observed. Point-contact devices may therefore be described as sharply nonlinear under all operating conditions."*

Certainly degenerate four-wave mixing and therefore time reversal can occur in in semiconductors. E.g., see (b) R. K. Jain, "Degenerate four-wave mixing in semiconductors: application to phase conjugation and to picosecond-resolved studies of transient carrier dynamics." *Opt. Eng.,* 21(2'), March/April 1982, p. 199-218. We also point out that hole current can be emitted from an n-type semiconductor having a metallic point placedd on its surface, and hole current will move against the voltage (electron current view). A process (patent pending) by Bedini and Bearden have shown that a capacitor may be charged by hole current, and then discharge electron current, effectively transducing negative energy into positive energy. It is therefore possible (Bedini's method) to switch a capacitor from a hole-current charging circuit by a hole-emitting process, to a bypass capacitor across a source such as a battery after the capacitor has reached higher charge voltage than the source. The capacitor will then charge the battery and power the circuit as well, during that portion of its discharge cycle while its voltage is above that of the battery. This is a legitimate COP>1.0 process, invented by Bedini.

36. (a) William Jay Fogal, "High Gain, Low Distortion, Faster Switching Transistor." U.S. Patent No. 5,196,809, Mar. 23, 1993; (b) — "High Gain, Low Distortion, Faster Switching Transistor." U.S. Patent No. 5,430,413, July 4, 1995. The Fogal semiconductor can be rigged to perform the permitted degenerate four-wave mixing and therefore time reversal. It can thus be rigged as a true negative resistor, where the current moves against the voltage.

37. Stephen L. Patrick, Thomas E. Bearden, James C. Hayes, Kenneth D. Moore, and James L. Kenny, "Motionless Electromagnetic Generator," U. S. patent # 6,362,718, March 26, 2002.

38. (a) T. E. Bearden, "Energy from the Active Vacuum: The Motionless Electromagnetic Generator," in M. W. Evans (Ed.), *Modern Nonlinear Optics,* Second Edition, Wiley, 2001, Part 2, p. 699-776; (b) M. W. Evans, P. K. Anastasovski, T. E. Bearden *et al.,* "Explanation of the Motionless Electromagnetic Generator with 0(3) Electrodynamics," *Found. Phys. Lett,* 14(1), Feb. 2001, p. 87-94; (c) M. W. Evans, P. K. Anastasovski, T. E. Bearden *et ah,* "Explanation of the Motionless Electromagnetic Generator by Sachs's Theory of Electrodynamics," *Found Phys. Lett.,* 14(8), Aug. 2001, p. 387-393.

39. A version of the MEG was also independently replicated by J. L. Naudin. See his website at http://jnaudin.free.fr/html/meg.htm .

40. T. E. Bearden, "Dark Matter or Dark Energy?", *J. New Energy,* 4(4), Spring 2000, p. 4-11.

41. See Chapter 4 for a high-level expose of this methodology.

42. Sweet and Bearden, {29a). Sweet performed an antigravity experiment designed by the present author, and successfully reduced **the** weight of an object on the laboratory bench, by 90%. Sweet's solid state magnetic vacuum triode amplifier produced a COP = 1,500,000, so a **very** appreciable Dirac sea hole current was available to work with. Conversion of the Dirac sea hole current into electrical current via the present Bedini-Bearden process would have eliminated the antigravity and produced a very large increase in output electrical power.

43. See (a) T. E. Bearden, "The Unnecessary Energy Crisis: How to Solve It Quickly," ADAS Position Paper, June 2000; (b) Bearden **{12};** (c) Bearden {30}; (d) T. E. Bearden, "EM Energy From The Vacuum: Ten Questions With Extended Answers," September 2000. The present author has been calling for just such a program. These papers are on restricted DOE Website http://www.ott.doe.gov/electromagnetic/ and also are on the author's website at http://www.cheniere.org.

44. Nathan Smith, "Marketable Results vs. Good Science?," *Sci. Am.* , 284(5), May 2001, p. 8-9. Quote is from p. 9.

45. This is the actual phrase often used in such Government contracts. What it means is that the Government (which means the government

employee controlling that program and its funds) can suddenly declare that the Inventor is not getting his invention developed rapidly enough, and the Government then seizes the patent to "get it developed more quickly". The inventor is suddenly left hanging, no longer in control of his own intellectual property. Or the Government has the power to block development altogether.

46. Often the universities just bluntly state that *all patents revert to the university.*

47. Hal Hellman, *Great Feuds in Science,* Wiley, New York, 1998. We particularly call attention to Hellman's dissertation on Wegener's persecution. Today, the legitimate overunity EM researcher and the cold fusion researcher are regarded much like Wegener was regarded — as the very epitome of a fool. This label is appended by "experts" who are not even aware of what really powers an EM circuit or an electrical power distribution system.

48. A photon is a piece of angular momentum (called "action" by the physicist), consisting of a little piece of spatial energy (ΔE) and a little piece of time (Δt), welded together as $(\Delta E)(\Delta t)$ with no seam in the middle, so to speak. But now look at that little $(\Delta E)(\Delta t)$. Since the photon is quantized, $(\Delta E)(\Delta t) = k = \hbar$. Any change in (Δt) requires that (ΔE) change inversely; i.e., $(\Delta E) = \hbar/(\Delta t)$. So increasing (Δt) decreases (ΔE) accordingly. Let us call the spatial energy *s* compacted into (dt) the equivalent *energy* of the photon. Then *s* is given by $s = c^2(\Delta t) = 9\times10^{16}(\Delta t)$. So as the frequency of a photon lowers, the (Δt) time component increases proportionally, and the total equivalent energy in the photon increases dramatically by many orders of magnitude! As an example, suppose we halve the spatial energy (ΔE). That doubles the original time interval (Δt), which gives the new $s = c^2(2\Delta t)$ or $s = 2c^2(\Delta t)$ for the new equivalent energy. So while we halved the *normal* spatial energy formally transported by the photon, we multiplied the *equivalent* time-energy by the enormous factor of $2c^2$ which is approximately 18×10^{16} in the MKS system. It follows that, if we use the equivalent time) energy of the photon and transduce a little of it into normal spatial energy, the energy density of those nuclear reactions now seen only in distant energetic objects such as pulsars, exploding stars, etc. can be achieved right on the laboratory bench. *The highest energy physics has not yet been born in the physics community, and what is presently being done in the high-energy*

accelerators is child's play compared to what can be done with time-energyengineering.

49. To see by another method that time is energy, one notes that the choice of fundamental units chosen for one's physics model is arbitrary. So we choose a single fundamental unit, the joule, to make a new physics model. Thereupon all other entities are energy and energy functions only, Hence, time is totally a function of energy. Actually, time is spatial energy compressed by the factor c^2, so it has the same energy density as mass, except the highly compressed spatial energy is placed in the time domain to appear at "time" rather than in 3-space to appear as "mass".

50. E.g., in cold fusion experiments where the time-energy is partially transduced in a time-reversed manner, the gluon forces binding the quarks in the affected neutrons and protons are dramatically reduced and the quarks are therefore nearly freed. This is being done in an apparatus of perhaps two cubic feet in volume, rather than a giant accelerator 20 miles in diameter and costing billions of dollars. So far, the largest accelerator using 3-space (decompressed) energy is just a bit short of the energy required to free the quarks. Use of time (compressed) energy allows such reaction as freeing the quarks to be easily accomplished, while the accompanying decompressed *spatial* energy is very low.

51. Time energy has the same energy density as mass, but that highly compressed energy has just been placed in the 4th axis *ict* of Minkowski space rather than in 3-space. In *ict* the only variable is t, so the variation of the compressed energy can only appear in the *t* variable portion. See Chapter 2.

52. (a) James Patterson, U.S. Patents 5,318,675; 5,372,688; 5,607.563 (b) — and Dennis Cravens, U.S. Patent No. 5,607,563.

53. Ironically, as suppression of innovative research in science increases, science may eventually face turning into the kind of dogmatic system that it struggled against for some 300 years before escaping. When science defends a theory that experiments refute, as it does today in the case of cold fusion, then it is very close to that debacle that can spell the end of science itself.

54. E. Allan Blair, Letter to the Editor, *Newsweek,* Mar. 3, 1997, p. 14.

55. Arthur C. Clarke, "Space Drive: A Fantasy That Could Become Reality," *NSS... AD ASTRA,* Nov/Dec 1994, p. 38.

56. William K. George, in *Editing the Refereed Scientific Journal, edited* by Robert A. Weeks and Donald L. Kinser, IEEE Press, New York, 1994. p. 227-228.

57. Max Planck, in G. Holton, *Thematic Origins ofScientific Thought, Harva*rd University Press, Cambridge, MA, 1973.

58. Jonathan Schlefer, editor of *Technology Review,* published at MIT. in an October 1990 editorial.

59. Rolf Landauer, "Fashions in Science and Technology," *Physics Today*,50(12),Dec.1997,p.62.

60. Julius Robert Von Mayer, "Bemerkungen iiber die Krafte der unbelebten Natur," in J. von Liebig's *Annalen der Chemie* (vol. 42), 1842; Die *organische Bewegung in ihrem Zusammenhange mit dem Stoffwechsel, (Organic* Motion in Its Relation to Metabolism), 1845. His papers were published as *Die Mechanik der Warme,* 1867, with third edition in 1893.

61. Alfred Wegener, *Die Entstehung der Kontinente und Ozeane, 1915;* fourth revised edition, 1929. A Dover reprint of the fourth revised edition, translated into English, *Origin of Continents and Oceans,* was published in 1996.

62. Paul Nahin, *Oliver Heaviside: Sage in Solitude,* IEEE Press. New York, 1988, p. 225.

63. Martin Gumpert, *Trail-Blazers of Science.* New York, Funk and Wagnalls Company, 1936, p. 232.

64. This area is now referred to as *ovonics,* where the "ov" is directly taken from "Ovshinsky". The materials have found use in copy machines, fax machines, liquid crystal displays, and optical memory disks. Ovshinsky has obtained over 200 patents and continues to work in the energy conversion field. His development of a new storage battery, e.g., broke all the Department of Energy tests and doubled the energy density permitted. Once in awhile the good guys beat the scientific suppression system.

65. Stan and Iris Ovshinsky were honored with the American Chemical Society's Heroes of Chemistry Award for 2000. They were honored for 40 years of sustained effort and contributions in the

sustainable and renewable energy field, including amorphous silicon cells for roofs of houses, highly specialized storage batteries, etc.

66. On the other hand, many scientific disagreements are conducted in quite proper and gentlemanly fashion. An example is the continuing decades of argument over energy flow theory, in *Am. J. Phys.* The problem has still not been officially resolved; we would hope that the material in this book will shed new light upon it, by pointing out the exact difference between the long-neglected nondiverged component as well *as* the diverged component, and also pointing out what Lorentz's trick of closed surface integration of the energy flow vector really does.

67. It follows quite naturally that, once we compress spatial energy by c , we may place the compressed energy in 3-space, in which case it is known as "mass", or we may place it in the time domain, in which case it is known as "time". If that cannot be done, then physics errs in using time as a variable on the fourth axis, and fundamental units are not arbitrary after all.

68. E.g., see D. K. Sen, *Fields and/or Particles,* Academic Press, London and New York, 1968, p. viii. Quoting: *"The connection between the field and its source has always been and still is the most difficult problem in classical and quantum electrodynamics."*

69. See T. D. Lee, "Can Time Be a Discrete Dynamical Variable?". *Phys. Lett,* 122B(3, 4), Mar. 10, 1983, p. 217-220. Lee showed that time is a discrete dynamical variable across the entire range of mechanics: from classical to nonrelativistic quantum mechanics, and then to relativistic quantum field theories. In all stages of mechanics, time can be treated as a discrete parameter, and it can be treated as a bona fide dynamic variable, Hence time has dynamics and structure, and therefore energy.

70. E.g., one may assume that the entire virtual state is an internal structuring and dynamics existing inside the macroscopic flow of time. The virtual state dynamics obviously exists in spacetime (which is assumed to be continuous), but we also consider any "virtual particle" (e.g., a virtual electron) as if it had been suddenly "observed" momentarily That assumes d/dt (LLLt) => LLL, a frozen 3-space snapshot. However, since that cannot really be observed, then we are actually considering something analogous to "if we could observe something in spacetime that cannot be observed in space alone". The only place left for it to

!observably exist anyway, if we could in fact observe it" is time, if space is ruled out.

71. It is necessary to trace the energy transductions and dissipations inside the generator. Ignoring those internal dynamics, the casual observer sees that one inputs shaft energy to the generator, and out comes some electromagneticenergy in the external electrical circuit, pouring from the gnerator terminals. We catch a lesser amount out there in the external circuit or power line, so we think we input some energy, had some internal losses and what gets collected in the conductors of the power line is the remainder of what we input, after transductions and losses in the generator. That is not what happens at all! Embarrassingly, an *enormous* flow of energy comes out of those generator terminals, if all of it is accounted. This includes(i) the tiny component that strikes the surface charges of the external conductors and gets diverged into the conductors, thus being captured by the circuit, and (ii) the enormous remainder of the energy flow in the surrounding space, which misses the circuit altogether and is just wasted. Something like a trillion times as much EM energy may be pouring out of the terminals of the generator and *being wasted*, as we mechanically input to the generator shaft, or as we "catch" and use in the external circuit. As later chapters discuss, Poyntin ; never considered the component that misses the circuit entirely. Heaviside discovered it but could not explain its startling magnitude or its source. Lorentz knew of this enormous flow component. He could not explain it, and stated it "had no physical significance" since it powered nothing. Lorentz originated a little integration trick that discarded it neatly from a11 accountability.

72. If the energy input to an inert power system generator were what powered the attached external circuit, then only the Poynting component would be emitted from the terminals of the generator. In that case, Heaviside's discovery of the additional nondiverged component would be wrong, and Lorentz was also wrong in recognizing it by stating it had no *physical* significance. It also means that a dipole formed between the terminals can have no broken symmetry in its active vacuum flux exchange, which is falsified by particle physics {14, 73,}. Also, no charge could pour out energy and create fields and potentials in surrounding space, and neither could a dipole. That is easily falsified by actual experimental measurement. Our reinterpretation of Whittaker's wave decomposition of the scalar potential would also be wrong, as would Mandl's and Shaw's argument {19} that a combined pairing of a scalar

photon and a longitudinal photon are observable, producing the instantaneous scalar potential.

73. T. D. Lee, *Particle Physics and Introduction to Field Theory,* Harwood, New York, 1981,*passim.* In 1957, Lee and Yang received the Nobel Prize for their work in predicting broken symmetry. This included the broken symmetry of opposite charges — and hence of a dipole.

The very definition of broken symmetry, applied to the source dipole once it is formed in the generator, reveals that the dipole itself receives virtual photon energy from the seething vacuum, transduces it, and outputs it as real observable EM energy pouring out of the generator terminals and out through 3-space surrounding the entire power line. This is the source of the tremendous energy flow from the terminals, including a large component missing the circuit (as discovered by Heaviside in the 1880s). This large nondiverged component of the energy flow was discarded arbitrarily by Lorentz because there was absolutely no explanation of its source, and because it was thought to have "no physical significance" (Lorentz's term) since it powered nothing.

74. Wu et al. {14} experimentally showed that the weak interaction violates parity (spatial reflection) and thus exhibits broken symmetry.

75. T. D. Lee is noted for his broken symmetry work with charges, and for establishing the broken symmetry of separated charges of unlike sign. In 1957, Lee and Yang were awarded the Nobel Prize in Physics for their investigations of weak interactions between elementary particles.

76. John D. Kraus, *Electromagnetics,* Fourth Edition, McGraw-Hill, New York, 1992, p. 578.

77. We strongly feel that anyone interested in logic must read Morris Kline, *Mathematics: The Loss of Certainty,* Oxford University Press, New York, 1980. Contrary to popular opinion and the impression projected by mathematicians, mathematics is not a body of unshakable truths aboul the physical world, and mathematical reasoning is not exact and fallible.

78. One also needs to be aware of Godel's proof. See Kurt Godel, "Uber formal unentscheidbare Satze der *Principa Mathematica* une verwandter Systeme" ("On Formally Indeterminable Propositions of the *Principia Mathematica* and Related Systems," in *Monatshefte fur Mathematik und Physik,* Vol. 38, 1931. This is the publication in which Godel's Proof first appeared, which states that within any logical

mathematical system there are propositions or questions that cannot be proved or disproved on the basis of the axioms within that system. Therefore, it is uncertain whether or not the axioms or arithmetic will give rise to contradictions, since they can.

79. A critical reference for the longitudinal EM interferometry *weapons* is M. W. Evans, P. K. Anastovski, T. E. Bearden *et al.*, "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," *J. New Energy,* 4(3), Winter, 1999, p. 76-78. This paper is a short but rigorous proof of a scalar interferometer's ability to produce EM fields and energy in the distant interference zone. Either diverging (heating) energy or converging (cooling) energy — in the normal sense, extraction of energy — can be created in a distant interference zone. By pulsing and controlling the circuit ground bias of the transmitters, one produced distant hot explosions or cold explosions. In a military target in that distant zone, the effects arise from the local spacetime, thus from everywhere within the targeted vehicle. At least eight nations now have such weapons, which were apparently the weapons that Defense Secretary Cohen {89} referred to in 1997.

80. Jackson {22}, p. 249.

81. Bunge, {3},p. 182.

82. Roger Penrose, "Foreword," in T. W. Barrett and D. M Grimes [Eds.], *Advanced Electromagnetism: Foundations, Theory, & Applications,* World Scientific, (Singapore, New Jersey, London, and Hong Kong), 1995.

83. Myron Evans, President of Alpha Institute's Institute for Advanced Study, private communication, 1999. Dr. Evans is the author of over 600 scientific papers, the prestigious editor of several scientific book series, and a noted foundations scientist. He is the co-discoverer and proponent of the Vigier-Evans B(3) magnetic field and a proponent of the proposed fundamental extension of U(1) electrodynamics to O(3) electrodynamics. O(3) electrodynamics has now been revealed as an important subset of Sachs' unified field theory.

84. Robert H. Romer, "Heat is not a noun," *Am. ./. Phys.,* 69(2), Feb. 2001, p. 107-109. The quotation is from footnote 24, p. 109.

85. E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," *Math. Ann.,* Vol. 57, 1903, p. 333-355.

86. But see Chapter 2 for a reinterpretation and correction of Whittaker's decomposition of the scalar potential. His harmonic set of bidirectional phase conjugate longitudinal EM wavepairs in 3-space is replaced with a harmonic half-set of longitudinal EM waves converging in I the time domain upon the dipolarity; absorption and transduction by the charges; and re-emission as the other harmonic half-set of longitudinal EM waves diverging out from the source dipolarity in all directions in 3-space. The correction allows a very novel solution to the long-vexing problem of the source charge and its associated fields and potentials and their energy {12}.

87. (a) G. Johnstone Stoney, "Microscopic Vision," *Phil. Mag.,* Vol 42, Oct. 1896, p. 332; (b) — "On the Generality of a New Theorem," ***Ibid.,*** Vol.43, 1897, p. 139-142; (c) — Discussion of a New Theorem in Wave Propagation," *ibid,* Vol. 43, 1897, p. 273-280; (d) — "On a Supposed Proof of a Theorem in Wave-motion," *ibid.,* Vol. 43, 1897, p. 368-373.

88. Some Russian scientists refer to the hidden internal electrodynamics as the "information content of the field".

89. Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

90. For a basic discussion of these photon polarizations, see (a) Richard Feynman, *Quantum Electrodynamics,* 1961, 1963. Quantum field theory recognizes four polarizations of the photon, to include x- and y-transverse polarizations in 3-space, longitudinal or z- polarization in 3-space, and the time-polarized photon (which may be taken to be a longitudinally-polarized photon in the time axis). Existence of these four modes of polarization or energy vibration in photons implies the existence of four correspondingly polarized EM waves. Among other things, Feynman discusses the "sum over four polarizations" of photons and such The term "Gupta-Bleuler" theory is the formalism that has grown up around this problem. For a deeper discussion, see (b) Lewis H. Ryder, *Quantum Field Theory,* Second Edition, Cambridge University Press, 1996, p. 147 *et seq.* For an even more advanced discussion, see (c) F.

Mandl and G. Shaw {19}. For discovery of the coupling of time-polarized and 3-space longitudinally polarized EM waves as the higher EM energy flow symmetry resolving the source charge problem, and thereby reinterpreting Whittaker's decomposition of the scalar potential {85}, see (d) Bearden {12}. We point out that the latter reinterpretation can also be "cranked through" the superpotential theory of Whittaker, to involve all EM potentials, fields, and waves.

91. (a) E. T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," *Proc Lond. Math. Soc*, Series 2, Vol. 1, 1904,p. 367-372. The paper initialed what today is called *superpotential* theory. It was published in 1904 and orally delivered in 1903. For a sampling of other superpotential work, see (b) W. H. McCrea, *Proc. Roy. Soc. Lond. A,* Vol. 240, 1957, p. 447; (c) A. Nisbet, *Physica,* Vol. 21, 1955, p. 99; (d) P. Debye, *Ann. Phys.,Leipzig,* Vol. 30, 1909, p. 57. For an excellent overview discussion of superpotentials and related things, see (e) Melba Phillips, "Classical Electrodynamics," in *Principles of Electrodynamics and Relativity,* Vol. IV *of Encyclopedia of Physics*, edited by S. Flugge, Springer-Verlag, l962.

92. (a) See Evans et al. {79} for proof that scalar interferometry can produce all normal transverse EM fields, waves, and patterns. The paper appears in (b) P. K. Anastasovski; T. E. Bearden, C. Ciubotariu, W. T. Coffey, L. B. Crowell, G. J. Evans, Myron W. Evans, R. Flower, S. Jeffers, A. Labounsky, B. Lehnert, M. Meszaros, P. R. Molnar, J.-P. Vigier, and S. Roy, "The New Maxwell Electrodynamic Equations: New Tools for New Technologies," *J. New Energy, 4(3),* Special Issue of A1AS papers. Winter 1999. The volume consists of some 60 papers by the Alpha Foundation's Institute for Advanced Study (AIAS), advancing electrodynamics to a non-Abelian, gauge theoretic higher topology theory in O(3) internal symmetry. Some ten of the papers directly address Whittaker's work and extend it.

93. (a) W. A. Rodrigues, Jr. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 < v < \infty$ in nature," *Found. Phys.,* 27(3), 1997, p. 435-508 show that Maxwell's equations do possess superluminal solutions. A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory web site. Also see (b) W. A. Rodrigues, Jr. and J. Vaz Jr., "Subluminal and Superluminal Solutions in Vacuum of the Maxwell

Equations and the Massless Dirac Equation," *Adv. Appl. Clifford Algebras,* Vol. 7(S), 1997, p. 457-466.

In classified Russian weapons labs, this internal longitudinal wave (both time-domain and 3-space domain) structuring of all normal (transverse) EM potentials, fields, and waves is euphemistically called the *information content of the field.* For decades there has been a successful disinformation campaign to lead scientists of other countries into believing that the phrase means *standard spectral analysis.* It doesn't. It is also possible to communicate at superluminal speed, using longitudinal EM waves inside this infolded "inner" electrodynamics. Since mass is mostly empty space filled with a particle here and there — together with EM potentials, fields, and waves — then mass is a vast "superhighway" for direct superluminal communication using longitudinal EM waves.

94. Romer, {84}.

95. See E. T. Whittaker, *A History of the Theories of the Aether and Electricity,* two vols., Nelson, London, 1951, 1953. Reprinted by Dover.

96. Note that we have just resolved the age-old problem of change, which can be summarized in the query stated so long ago by Heraclitus: "For a thing to change, it must turn into something else. But how can a thing be itself and something else also?" We have explained the enigma by pointing out that the "thing" has no persistence in time, and is continually changing into yet another thing, either a replica (in which case we sense "no change" and hence "persistence" by memory comparison), or a partial replica with some differences (in which case we sense that "the object has changed itself in some way"), or a thing with little or no replication (in which case we sense a "different thing"). It is the unseen nature of the *d/dt* observation process that has deluded us into such irresolvable logical conflicts — and into an incomplete Aristotelian logic based on observation alone, which does not contain the resolutions to such problems because a higher-order logic is required.

97. Whittaker, {85}

98. See Chapter 2 for a discussion of the substitution of effect for cause, and of the nature of observation.

99. *Wu etal,* {14}.

100. Sen, {68}.

101. Also see Mandl and Shaw {19}.

102. William G. Harter, John Evans, Roberto Vega, and Sanford Wilson, "Galloping waves and their relativistic properties," *Am. J. Phys., 53*(7), July 1985, p. 671-679. A relativistic example is given in which the source frame observes a very nonuniformly-galloping wave whose velocity varies between 0.01 c and 100 c, even though its average velocity is c. Relativistic properties of interfering electromagnetic plane waves are discussed.

103. We point out, but do not further pursue, that this situation is reversed when time is reversed. This has very interesting applications in our little "time reversal zones" that we found were so important in the electrolyte experiments of the cold fusion researchers, and which accounted for a new class of nuclear interactions previously unknown. This does reverse the conventional law of attraction and repulsion of charged particles. It also directly affects the quarks and gluons, partially or even nearly dissolving the gluon forces holding the quarks in place, so that the quarks are nearly freed and are easily flipped.

104. Simultaneously accompanied by *separately broken* time-energy flow symmetry and *separately broken* 3-spatial energy flow symmetry.

105. Evans etal., {38b, 38c}.

106. Oliver Heaviside, *Electrical Papers,* Vol. 2, 1887, p. 94.

107. Kraus, {76}.

107. E.g., see (a) Jackson {22}, p. 237. Jackson points out that *"...the Poynting vector is arbitrary to the extent that the curl of any vector field can be added to it. Such an added term can, however, have no physical consequences. "* The first sentence is true, but the second sentence is a non sequitur. Poynting only considered the small component of the overall energy flow that enters the circuit — in short, that strikes the surface charges in the conductors and is diverged into the wires to power the Drude electrons. That is the diverged part of the energy flow. All the rest of the flow that misses the circuit is obviously not diverged or collected in the circuit, and so it has zero divergence. It is elementary vector algebra that the curl of any vector field has zero divergence. So the extra nondiverged Heaviside energy flow component — filling all space around the conductors in a power line, e.g., but missing the power system entirely and just wasted — can indeed be expressed as the curl of a vector field That energy flow is real, however, and if intercepted and collected, it then does become divergent. In that case, it does have highly significant

consequences, since the total energy flow in that component may be a trillion times the energy flow in the Poynting component alone.

For a rigorous experiment demonstrating the reality and significance ol that "physically insignificant energy flow" that "can have no physical consequences", see (b) Bohren {24}. Under nonlinear conditions, a particle can absorb more energy than is in the light incident on it by normal calculations. Electrodynamicists do *not* calculate the magnitude of an entire field or potential, e.g., but only the intensity of a *divergence from* it at a single point, as given by an intercepting unit point static charge. If that same charged particle is placed in particle resonance, it sweeps out a greater geometrical reaction cross section. Hence it intercepts and **collects** (diverges) more impinging energy — from that long-neglected otherwise nondiverged component of the flow or field that is not diverged by the *static* unit point charge assumed conventionally. Metallic particles at ultraviolet frequencies are one class of such particles and insulating particles at infrared frequencies are another. See also (c) Paul and Fischer, {25}, who replicated Bohren's experiment.

109. H. A. Lorentz, *Vorlesungen iiber Theoretische Physik an der Universitdt Leiden,* Vol. V, *Die Maxwelhche Theorie (1900-1902),* Akademische Verlagsgesellschaft M.B.H., Leipzig, 1931, "Die Energie im elektromagnetischen Feld," p. 179-186. The section is called "Der Poyntingsche Energiestrom," and begins on p. 183. See his Figure 25 **on** p. 185.

110. E.g., see W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism,* Addison-Wesley, Reading, MA, 1962, 2nd edition, p. 181**,**

111. Bearden, {12}.

112. A simple way to see this is to note that the back-potential across the half-loop internal to the dipole is precisely equal to the forward potential externally between the ends of the dipole but across the **external** circuit half-loop. Since the same current *q* runs through both these potentials, then by *Vq* the same amount of energy is dissipated upon **the** charges of the dipole — to scatter them and destroy the dipole — as is dissipated in the external circuit in its losses and the load. Thus **more** energy is used to destroy the dipole than is used to power the load. We have to input at least as much shaft energy to restore the dipole as was used to destroy it. Hence the COP< 1.0.

113. See to (a) M. W. Evans, P. K. Anastasovski, T. E. Bearden *el al*, "Classical Electrodynamics without the Lorentz Condition: Extracting Energy from the Vacuum," *Physica Scripta*, 61 (5). May 2000, p. 513-5 17 in addition to that paper previously quoted. See also (b) "Runaway Solutions of the Lehnert Equations: The Possibility of Extracting Energy from the Vacuum," *Optik*, 111(9), 2000, p. 407-409.

114. Robert Bruce Lindsay and Henry Margenau, *Foundations of Physics*, Dover, New York, 1963, p. 25.

115. *Ibid.*, p. 213.

116. *Ibid.*, p. 217.

117. M. Faraday, *Experimental Researches in Electricity and Magnetism*, Vol. 1, Taylor and Francis, London, 1839; Vol. 2, Richard & John E. Taylor, London, 1844; Vol. 3, Taylor and Francis, London, 1855. A 1965 reprint of all three volumes is available from Dover, New York.

118. We also acknowledge the independent contributions of Joseph Henry. In 1830, Henry independently discovered magnetic induction, anticipating Faraday's own discovery of it. Faraday, however, published first. Two years later, Henry discovered self-induction. He died in 1878, and the unit of inductance, the *henry*, was named for him in 1893.

119. (a) Maxwell, {20}. Also see (b) James Clerk Maxwell, *A Treatise on Electricity and Magnetism*, Oxford University Press, Oxford, 1873, Third Edition, Volumes 1 and 2, unabridged, Dover Publications, New York, 1954. This standard citation widely used in the literature is in fact confusing; the First Edition was published in 1873, a second rather substantially altered edition was later published after Maxwell's 1879 death of stomach cancer, and a third edition, slightly altered from the second edition, was published even later. So controversial were quaternions at the publication of Maxwell's first edition in 1873, that part of what was to become the second edition — published after Maxwell's 1879 death — was truncated by Maxwell himself prior to his death and prior to the posthumous publication of the second edition. Thus there is considerable difference between the first and second editions, and first and third editions, of Maxwell's *Treatise*, compared to his original 1865 paper {20}. "Maxwell's theory" is actually his 1865 paper; everything else is a later truncation of Maxwell's theory.

120. Maxwell's letter to P. G. Tait in 1871, *Archives,* Cavendish Laboratory, Cambridge.

121. For a discussion, see (a) A. M. Bork, "Vectors versus quaternions - the letters in *Nature,"* in *Am. J. Phys.,* Vol. 34, Mar. 1966, p. 202-211. For a cogent presentation of what might have been discovered much earlier in physics if quaternions had not been cast aside in favor of the much more limited vectors, see (b) James D. Edmonds Jr., "Quaternion quantum theory: New physics or number mysticism?", *Am. J. Phys.,* 42(3), Mar. 1974, p. 220-223.

122. Edmonds, {121b}. Edmonds presents a cogent argument showing the significant impedance of the rate of discovery in physics, that **resulted** from the ill-advised casting aside of the far more comprehensive quaternion theory in favor of the much more limited vector theory.

123. Heaviside, {5a, 5b}.

124. T. W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," *Annales de la Fondation Louis de Broglie,* 16(1), 1991, p. 23-41, Barrett later extended some of the energy-shuttling techniques shown by Tesla and obtained U.S. Patents 5,486,833 and 5,493,691.

125. (a) A. A. Michelson and E. W. Morlcy, "Influence of motion of the medium on the velocity of light," *Am. J. Sci.,* Vol. 31, Series 3, 1886, p. 377-386; (b) — "The relative motion of the earth and the luminiferous aether," *Am. J. Sci.,* 34(3), 1887, p. 333; (c) — "On the relative motion of the earth and the luminiferous aether," *Phil. Mag.* 24(4), 1887, p. 449. See also (d) A. A. Michelson, "The relative motion of the earth and the ether," *Am. J. Sci.,* (4), 3, 1897, p. 475.

126. (a) Mendel Sachs, *General Relativity and Matter: A Spinor Field Theory from Fermis to Light-Years (Fundamental Theories of Physics),* Reidel, 1982. Sachs provides a great generalization of general relativity and electrodynamics reaching from the quarks and gluons to the entire universe. See also (b) Mendel Sachs, "Symmetry in Electrodynamics: From Special to General Relativity, Macro to Quantum Domains," in M. W. Evans, Editor, *Modern Nonlinear Optics,* Second Edition, Wiley, **New** York, 2001, Part 1, p. 677-706.

127. Myron W. Evans, Director of the Alpha Foundation's Institute for Advanced Study, private communication, 2001.

128. Jackson, {22}, p. 249.

129. Jackson, {22}, p. 28.

130. Feynman, {7}, Vol. 1, p. 2-4.

131. Physicists often speak of spacetime as "space", since in Minkowski geometry the time portion is modeled along a separate spatial axis used in addition to the ordinary three spatial axes. So Feynman's "distortion" is actually Riemannian spacetime curvature.

132. As an aside, note that Feynman also utilized a dipolarity. There is always a scalar potential between the poles of any dipole. Further, this "dipole potential" decomposes according to Whittaker's 1903 decomposition{85}, as re-interpreted by Bearden's giant negentropy approach {12} in 2000, and as relates to Mandl and Shaw's 1984/1993 argument{19} that only the associated time-polarized photon and longitudinal photon together are observable, and then as the instantaneous scalar potential. If we accept Mandl and Shaw's powerful argument, then the Whittaker 1903 decomposition must be reinterpreted to be in accordance. That is what was done by the present author {12} to resolve the source charge problem as well as the source dipole problem. In addition, the resolution must be in accord with the theoretical and experimental findings of particle physics, to include the broken symmetry of the opposite charges on the ends of a dipole. The giant negentropy process uncovered by Bearden {12} fulfills all the necessary requirements. It also places the theme of this book — *Energy from the Vacuum: Concepts and Principles* — on a solid theoretical and experimental basis.

133. Feynman, {7}, vol. II, p. 1-3.

134. Jackson, {22}, p. 811-812.

135. We have previously postulated this feature where one second is equal to spatial EM energy compressed by the factor c^2, so that time is just extremely compressed EM spatial energy. See Bearden, {17}.

136. (a) Harter et al., {102}. It may well be that galloping due to time-density variation may provide an alternative explanation to quantum tunneling, which can readily occur at superluminal speed. E.g., see (b) A Enders and G. Nimtz, *Phys. Rev. Lett.,* Vol. 48, 1993, p. 632. Their superluminal tunneling experiments transmitted 8.7 GHz microwaves (free space wavelength 3.4 cm) traveling in a rectangular waveguide that contains a "barrier" section of reduced dimensions. Mozart's 40th Symphony was transmitted through the barrier waveguide as frequency

modulated microwaves, at an FTL group velocity of 4.7c, receiving audibly recognizable music from the microwave photons that survived their barrier passage.

137. (a) M. W. Evans, "The Link Between the Sachs and 0(3) Theories of Electrodynamics," in M. W. Evans, editor, *Modern Nonlinear Optics,* Second Edition, Wiley, New York, 2001, Part 2, p. 469-494; (b) — "0(3) Electrodynamics," in M. W. Evans, editor, *Modern Nonlinear Optics,* Second Edition, Wiley, New York, 2001, Part 2, p. 79-267.

138. E.g., Ryder, {90b}, p. 147 *et seq.*

139. Rodrigues and Lu, {93a}.

140. Rodrigues and Vaz, {93b}.

141. The reader should recall that to this day there is no satisfactory resolution of the wave-particle duality principle. Physicists finally just shook hands and agreed to quit fighting, since the hullabaloo was getting them nowhere. The problem does have a probable resolution, but not in Aristotelian logic.

142. Bearden, {12} discusses the relation between time-energy and 3-spatial energy, and reveals a more fundamental 4-symmetry between time energy flow and spatial energy flow. In this 4-symmetry EM flow, time symmetry in energy flow is broken and simultaneously 3-space symmetry in energy flow is broken, but 4-symmetry in energy flow is maintained. The basic transduction of time-energy into 3-space energy and vice versa is accomplished by the interacting charge.

143. Bearden, {15}, p. 11. In the paper we also gave a force-free definition of mass, in terms of the transduction of angular momentum across the 3-spatial boundaries of a mass particle.

144. Rigorously, *mass* does not and cannot emit a photon; *masstime* does. But an observable mass m can be absorbing and emitting **enormous** numbers of virtual photons while not yet having absorbed another observable photon and decayed to emit an observable photons. Differential (virtual) pieces of a "frozen" observable mass m are **indeed** changing continually into masstime and back into mass, by absorbing and emitting virtual photons per equation [14b], while the observable mass m has not yet completed its absorption and emission of an observable **photon** This "knitting" together of observable and virtual state interactions is one of the great centralizing features of physics.

145. The problem is due to the hoary old erroneous notion in mechanics that a separate force acts upon a separate mass to move it.

146. There appears to be no paper or textbook which calculates the actual "energy magnitude" of an entire potential or an entire field, including over all space occupied by it. All texts advance the "local intensity" indicator calculation, and then call that the "magnitude of the potential". That is a non sequitur of first rank. A skilled electrodynamicist will call it the intensity of the potential.

147. It appears that the only region where this statement is in dispute is in various models attempting to explain the creation of the universe. Since that is *a priori* a creation event, then physicists struggle with explaining how something (the existing universe) was created out of nothing (the absence of any slightest bit of the existing universe). We will leave that debate to the astrophysicists and cosmologists. The "accursed identity of opposites" has been the bane of philosophers for more than 3,000 years.

148. Ibrahim Semiz, "Black hole as the ultimate energy source," *Am. J. Phys.,* 63(2), Feb. 1995, p. 151.

149. David Halliday and Robert Resnick, *Fundamentals of Physics,* Third Edition Extended, Wiley, New York, 1988, Vol. 1, p. 518.

150. Stanley W. Angrist, "Perpetual Motion Machines," Sci. *Am.,* Vol. 218, Jan. 1968, p. 114-122.

151. Stanley W. Angrist, "Perpetual Motion," *Encyclopaedia Britannica,* Bicentennial Edition, Macropaedia Vol. 14, 1976, p. 102-105.

152. Lindsay and Margenau, {114}, p. 217.

153. Heaviside, {106}, p. 94.

154. Kraus, {76}, Figure 12-59, p. 578.

155. Bohren, {24}.

156. Paul and Fischer, {25}.

157. (a) V. S. Letokhov, "Laser Maxwell's Demon," *Contemp. Phys.,* 36(4), 1995, p. 235-243; (b) — "Generation of light by a scattering medium with negative resonance absorption," *Zh. Eksp. Teor. Fiz.,* Vol, 53. 1967, p. 1442 (English translation is in *Sov. Phys. JETP,* 26(4), Apr. 1968. p. 835-839); (c) "Stimulated emission of an ensemble of scattering particles with negative absorption," *ZhETF Plasma,* 5(8), Apr.

15, 1967, p. 262-265; (d) — "Double y - and optical resonance," *Phys. Lett. A,* Vol. 43, 1973, p. 179-180.

158. It is Heaviside's equations that are usually taught in University as "Maxwell's Equations," and labeled as Maxwell's equations in most undergraduate texts. In addition to Heaviside' work, similar changes were made by Gibbs in the United States and by Hertz in Europe. Every sophomore physics student and electrical engineering student should be given a short treatise on the background development of electrodynamics, and how, when, and by whom Maxwell's 20 quaternion equations were modified into the present four vector equations. The treatise should also point out what was omitted from Maxwell's theory in its truncations. It should be made very clear to every student that his elementary EM subject errs dramatically from physical reality when it assumes a flat local spacetime and an inert local vacuum. *By discarding the physically active environment ofall Maxwellian systems,* only a highly limited subset of nature's electrodynamics is captured in the standard electrical engineering model. Specifically, systems far from equilibrium in their exchange with that environment are arbitrarily discarded. Those are precisely the Maxwellian systems permitted to exhibit coefficient of performance greater than unity. In short, the truncation of Maxwell's theory arbitrarily discarded all electrical windmills turning — or partially turning — in **a** free electrical wind.

159. As we pointed out, there are *tin electromagnetic systems at all in* that first class — and never have been — if we account for the arbitrarily-ignored Heaviside energy flow pouring from the source dipole and the energy flow pouring from every charge, and consider the situation prior to self-enforced implementation of Lorenz/Lorentz symmetrical regauging (by the closed current loop circuit). The remarks — pertaining to EM systems supposedly in equilibrium with the active vacuum — only apply if we neglect that Heaviside component and also neglect the source **charge** problem, and even then only if the circuit self-enforces that equilibrium in its exchange with the active vacuum. But since it has been so long **and** ubiquitously neglected, we will continue to speak of the "two classes" of EM systems from that "conventional" view. It is rather like our concession to electrical engineering, where the word "power" is universally **misused.** However, the reader should be aware of what we are doing and why.

160. See (a) Ilya Prigogine, *Nonequilibrium Statistical Mechanics,* Wiley-Interscience, New York, 1962; (b) — *From Being to* ***Becoming:***

Time and Complexity in the Physical Sciences, W. H. Freeman and Company, San Francisco, 1980; (c) D. Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, Chichester, 1998; (d) G. Nicolis and I. Prigogine, *Self-Organization in Non-Equilibrium Systems: From Dissipative Structures to Order through Fluctuations,* Wiley, New York, 1977. For a good educated layperson's overview, see (e) Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," Chapter 11 in {f} Paul Davies, Ed., *The New Physics,* Cambridge University Press, Cambridge, 1989, p. 316-347.

161. Actually, Ludvig Valentin Lorenz first performed the equivalent of' symmetrical regauging of Maxwell's equations, only two years after Maxwell's seminal paper was published in 1865. See (a) Ludvig Valentin Lorenz, "On the identity of the vibrations of light with electrical currents," Phil. *Mag.,* Vol. 34, 1867, p. 287-301. Not much attention was paid to Lorenz's work, however, and the later symmetrical regauging of the Maxwell-Heaviside equations by H. A. Lorentz is what "stuck" and came to be used. For a proper historical discussion, see (b) J. D. Jackson and L. B. Okun, "Historical roots of gauge invariance," *Rev. Mod. Phys.,* Vol. 73, July 2001, p. 663-680.

162. This point is important. See Gabriel Kron, "The Frustrating Search for a Geometrical Model of Electrodynamic Networks," circa 1962. General Electric publication, p. 11-128. Quoting p. 114: *"...the missing concept of "open-paths" (the dual of "closed-paths") was discovered, in which currents could be made to flow in branches that lie between any set of* ***two*** *nodes. (Previously -following Maxwell - engineers tied all of their open-paths to a single datum-point, the 'ground'). That discovery of open-paths established a second rectangular transformation matrix... which cre****ated*** *'lamellar' currents... " "A network with the simultaneous presence of both closed and open paths was the answer to the author's years-long search."*

When we design and build a circuit or system in compliance with the Lorenz/Lorentz-regauged equations, it follows that *the circuit itself must contain some feature and function that self-enforces the Lorenz/Lorentz condition — at least during the excitation discharge of the energy that has been freely caught by the circuit. The Lorenz/Lorentz condition is not a law of nature, nor is it automatically applied without a specific circuit cause and function physically applying it.* That function is automatically applied by the standard closed current loop circuit.

As Kron stated, his long search for a true negative resistance finally succeeded when he discovered the "open path". Actually Kron had discovered that any two points in the universe at different potentials (charged to different intensities) produced a flow of EM energy. In other words, he had discovered in more classical EM terms the broken symmetry of opposite charges of differing charges, such as on the ends of any dipole or across any dipolarity.

163. Superconducting *sections* in a system may be loss-free, but the entire superconducting *system* has losses and exhibits only COP<1.0 overall, if the refrigeration burden is accounted. In the superconducting section alone, without accounting for the cooling overhead, the COP = 1,0 and the efficiency is 100% also.

164. After Jackson, {22}, p. 219-221; 811-812.

165. See Whittaker, {85}. Whittaker mathematically decomposes the scalar potential into a hidden "bidirectional" series of EM wavepairs in a harmonic sequence. Each wavepair consists of the wave and its phase conjugate. Thus any Aq) (as across the terminals of a dipolar electrical power source) *a priori* involves an ordered, hidden, bidirectional EM wave flow. I.e., d(fi)) *identically is* such a hidden bidirectional EM energy flow.

166. But see Bearden, {12}. In the paper we conceptually re-interpreted Whittaker's 3-space (observation assumed, hence effect waves) bidirectional wavepairs. Our reinterpretation yields a correlated harmonic half-set of converging *unobserved* causal EM longitudinal waves (time-polarized EM waves) in the time domain prior to interaction with charge so as to prepare for being observed, and a harmonic half-set of diverging EM longitudinal waves (after observation due to the convergent time-polarized EM waves interacting with the observable charges) in 3-space in all directions. We strongly stress that observation does not occur without the absorption of a scalar photon and the follow-on emission of a longitudinal photon. In a time-reversed situation (phase conjugate charge condition), to the observer it appears to have happened in the other direction.

167. Particularly see Mandl and Shaw, {19}.

168. However, Mandl and Shaw fail to totally identify the "combining mechanism". They do not account the detailed interaction of the detecting/observing unit point charge, and thus fail to clearly account for the absorption of the incoming time-polarized wave or photon, the

transduction of that excitation energy of the charge into longitudinal EM wave/photon energy, and the subsequent emission of that excitation energy in 3-space in all directions as longitudinal EM waves/photons. So Mandl and Shaw do not clearly account for *photon (or wave) polarization transduction by the ubiquitously assumed charge.* The "causal" time-polarized EM wave or photon comes in and is absorbed by the detecting charge or dipole, then the absorbed energy is transduced and re-emitted in 3-space as the longitudinally polarized EM wave or photon in 3-space. (Actually, since all space is charged, what is really emitted into 3-space is the giant circulation of EM energy between time and 3-space, which is a more rigorous statement of what "propagation of EM energy through 3-space actually is). Recognition of these missing interactions and transductions allowed a solution at last to the long-vexing problem of the source charge, often called the greatest problem in both quantum and classical electrodynamics, as given initially in Bearden, {12}. Now, of course, a greater problem has reared its head: the notion that EM energy continuously propagates in 3-space is a total non sequitur, along the lines objected to by Romer {84}. We therefore have some very fundamental rethinking to do for all of classical physics. This is really where the problem of wave-particle duality has pointed since the beginning.

169. Asymmetrical self-regauging collects just a tiny bit more of the enormous energy that is flowing outside the conductors and usually wasted. E.g., if the self-regauging EM system collects as $j\phi$ some 10^{-12} of the energy flow instead of the more usual 10^{-13}, the system can have a COP =10, *if it* does not then expend half the collected energy to destroy the source dipole faster and require its reconstitution more energetically. And so on.

170. Jackson, {22}, p. 220-223.

171. So-called "canceling" opposing EM fields are actually produced, which sum to a *vector zero system* with respect to translation, which the electrodynamicists erroneously discard by assumption. We point out, but do not further pursue, that the locally produced *field energies* of the opposing fields in a zero-summed vector system remain and add, even though the fields offset each other *translationally.* The local energy of the field is proportional to the square of the local magnitude, and that is always positive regardless of field orientation. Thus "trapped" and structured EM stress energy has been localized in spacetime in the symmetrical regauging of CEM. This rigorously is a gravitational change

and therefore a local curvature of spacetime *a priori.* Indeed, the field energy remains in a zero-vector summation system, and has been changed into stress potential energy. It represents continuous internal work being performed on the system to stress it, by energy continuously fed in from the active vacuum environment but only in appositive fields of equal magnitude. The discarding of the zero vector summation system thus is a discarding of an available stress potential and its energy, a rotation of the system frame out of the laboratory frame, and a continuous exchange of energy from the vacuum that performs internal work upon the system but no external work in the load. The energy is not translational, but it is there nonetheless, and it causes a spacetime curvature change. The electrodynamicists have largely ignored the fact that gauge symmetry transformations are accompanied by local gravitational (spacetime curvature) changes that violate the prevailing flat local spacetime assumption of classical electrodynamics.

172. See (a) H. J. Josephs, "The Heaviside papers found at Paignton in 1957," *IEE Monograph No. 319,* Jan. 1959, p. 70-76. See also (b) E. R. Laithwaite, "Oliver Heaviside — Establishment Shaker," *Elec. Rev.,* 211(16), Nov. 12, 1982, p. 44-45. Heaviside became aware of this gravitational relationship, and worked out a draft theory of electrogravitation based on his energy-flow theory (which added nondiverged closed-loop circulations of energy which vectorially summed to net zero). Those "trapped EM energy flow loops" were *gravitational,* in Heaviside's concept. His hand-written notes on his eleelrogravitation theory were found long after his death, beneath the floorboards of his little garret apartment. In honor of Heaviside, we have nominated his huge nondiverged energy flow component — erroneously discarded by Lorentz — as the source of the extra gravity in the arms of the spiral galaxies, holding them together. In short, we have proposed that as the solution to the so-called "dark matter" problem.

173. Physicists love symmetry, turn to it at every opportunity, and will do almost any mathematical manipulation to obtain it, because they regard it as "beautiful." And so it is. Performing work, however, requires breaking symmetry to produce excess energy, which interacts with mass to produce a force, and then requires using that force to do the work. If we wish to build a system that produces more work output than the work that we ourselves have to do upon it to run it, then *a priori* that system must be *asymmetrical* and — to many physicists — *ugly.* Let us all drink a toast to more efficient *ugly* asymmetrical systems that also do not require **energy.**

input by the operator, but only by the active environment! Give us "ugly" electrical windmills that turn freely in electrical winds, rather than the inane "beautiful" present systems which destroy their energy winds faster than hey power their loads.

174. While Lorentz appears to have done it earlier, one can see his use of the surface integration in (a) Lorentz, {109}. In 1896 he also included it, using a more general theorem by Volterra, so that Poynting's theorem arises as a special case; see (b) H. A. Lorentz, *Versl. K. Akad. W. Amsterdam,* Vol. 4, 1896, p. 176.

175. This is usually performed by the source dipole.

176. Thus the long and tedious effort of my colleagues and I to create circuits that asymmetrically self-regauge their potential difference (source voltage) without allowing all the spent load current to pass back through the primary power source dipole and do work upon it to scatter the charges and destroy its dipolarity.

177. Again we strongly stress that asymmetrically regauging a potential in the system also regauges the local vacuum potential and the intensity of the vacuum potential's bidirectional EM energy exchange with the system. It alters the local curvature of spacetime. It directly alters the stored energy of the system and of the supersystem. See again Whittaker, {85}, to see this clearly. See particularly our use of the *supersystem* concept in Chapter 8, and as mentioned in Chapter 1 and the present Chapter 2.

178. Circuits and electrical power systems universally and routinely break Lorentz symmetry for *excitation* (potentialization) of the system at the initial onset. Just to potentialize a system in theory costs nothing, although it freely changes the potential energy of the system. The gauge freedom axiom of quantum field theory also guarantees that such change of the potential energy of the system is essentially "for free". It conclusively proves that (i) asymmetrical self-regauging exists and is done in electrical circuits, since the source dipole in the power supply accomplishes that initial self-regauging of the external circuit completely for free once the dipole is made, and (ii) the Maxwell-Heaviside equations do permit asymmetrical self-regauging, in violation of Lorentz's arbitrary condition prohibiting it.

179. This is the patent-pending Bedini process, which is processing through the U.S. Patent Office as we go to press.

180. Lee, {73}, p. 184.

181. B. P. Kosyakov, "Radiation in electrodynamics and in Yang-Mills theory," *Soviet. Phys. Usp.*, 35(2), Feb. 1992, p. 135-142. Quotes are from p. 135 and p. 141. Our proposed solution to the problem posed by Kosyakov (which is actually the broken symmetry of a charge) is that negative charges pour out positive energy and positive charges absorb positive energy. We "see" the positive charge as pouring out "reversed direction positive field energy" after we interact the energy with a laboratory charge to "observe". Considering the well-known polarization of the vacuum by any charge, one realizes that a 4-symmetry circulation of the EM energy is what is "emitted" and spreads throughout all space, since the vacuum polarization is also spreading. The actual energy flow is between the time domain and 3-space and return, at any elemental dipole in the polarization of the vacuum that is spreading.

182. Oliver Heaviside, *Electrical Papers,* Vol. 2, 1887, p. 94.

183. Prof. William Keel, quoted in James McWilliams, "Through a Galaxy, Darkly," *Huntsville (Alabama) Times,* May 14, 2000, p. A-18.

184. Charles Seife, "Peering Backward to the Cosmos's Fiery Birth," *Science,* vol. 292, June 22, 2001, p. 2238.

185. M. P. Silverman, *And Yet It Moves: Strange Systems and Subtle Questions in Physics,* Cambridge University Press, Cambridge, 1993, p. 127.

186. Matthew Colless, Australian National University, one of the survey team leaders in the project leading the world in mapping galaxies, Mar. 2001.

187. (a) Lee, {13a}; (b) Lee, Oehme and Yang, {13b};

188. Wuetal., {14}.

189. Bearden, {12}.

190. Mandl and Shaw, {19, 168}.

191. Whittaker, {85}.

192. Gabriel Kron, "Invisible dual (n-1) networks induced by electric 1-networks," *IEEE Trans, on Circuit Theory,* CT-12(4), Dec. 1965, p. 464-470.

193. Kron, {162}. The quote is from p. 114.

194. Steven Weinberg, *Dreams of a Final Theory,* Vintage Books, Random House, 1993, p. 109-110.

195, See Josephs, {172a}. Heaviside's hand-written notes contained his theory of electro-gravitation, based on his theory of energy flow. The papers were found beneath the floorboards in his little garret apartment.

196. Laithwaite, {172b}. Laithwaite feels that Heaviside's postulation that a flux of gravitational energy combines with the (ExH) electromagnetic energy flux, could shake the foundations of physics. Extracting from Laithwaite; *"Heaviside had originally written the energy flow as S = (ExH) + G, where G is a circuital flux. Poynting had only written S = (ExH). Taking p to be the density of matter and e the intensity of a gravitational force, Heaviside found that the circuital flux G can be expressed as pu -ce, where u represents the velocity of p and c is a constant. "*

197. Hence the title of cited reference {12} above.

198. Melvin H. Miles and Benjamin F. Bush, "Radiation measurements at China Lake: Real or Artifacts?", *Proc ICCF-7 (International Conference on Cold Fusion* — 7, Vancouver, BC, Canada, Apr. 1998, p. 101.

199. Bearden, {17}.

200. See P. V. Elyutin, "The Quantum Chaos Problem," *Sov. Phys. Usp..* Vol. 31, No. 7, 1988, p. 597-622. Elyutin discusses the crisis in quantum mechanics because of the missing chaos (hidden order). Quantum mechanics is known to be wrong unless this order can be found, because otherwise it does not predict the ordered macroscopic universe. Thus any proposed solution to the quantum chaos problem is worthy of Investigation, even if it turns out to be in error.

201. Unfortunately, entropy is one of those concepts in physics for which there are several differing major views. We consider entropy to be comparable to a positive resistor: organized energy goes in, disorganized energy comes out. Entropy is usually taken to be a measure of the disordering of energy. It does not *eliminate* the energy; it just makes it more difficult to *use* it. For our work in energy from the vacuum, we take the very simple view that a negentropic process is like a negative resistor: it receives energy in a form unusable to us, transforms it, and outputs it in

a form that is usable. In short, a negative resistor receives disordered energy and reorders it, so to speak. So it exhibits negentropy. We completely avoid the various notions of "information" and attempts to equate information and energy. We do point out, however, that a time-reversal process in one form or another is usually involved. In that sense, e.g., Newton's third law would be a negentropic process and involve time reversal in the mechanism generating the reaction-causing agent.

202. H. E. Puthoff, "Source of Vacuum Electromagnetic Zero-Point Energy," *Phys. Rev. A,* 40(9), Nov. 1, 1989, p. 4857-4862.

203. Whittaker, {91}. The paper was published in 1904 and orally delivered in 1903. This paper initiated superpotential theory.

204. As Whittaker showed in 1903, {85}, the scalar potential is actually a harmonic set of bidirectional EM longitudinal EM wavepairs, where each pair is composed of a longitudinal EM wave and its phase conjugate replica. Only because classical electrodynamicists have erroneously defined the field and potential as their own reaction cross sections with a unit point static charge, has the "static" potential been misidentified as a *scalar* entity, which it is not. The energy diverged from a uniform potential, around a fixed static point unit charge, is actually the set of divergences around the intercepting charge of the energy flows of all those EM waves comprising the potential. The sum total of all these individual wave divergences indeed has a scalar magnitude, but the magnitude of the total energy divergence from the potential is neither the potential itself nor its magnitude.

205. We point out the obvious: A "scalar" mass in 3-space actually him a time-vector since it moves through time continually, just to continue to exist. Further, it involves a special form of energy (energy compressed by c^2) which time is taken to be. Since we may choose any form of energy we wish by simple transduction, we may take it as compressed EM energy. So the mere continued existence of any mass proves conclusively that EM energy can and does ubiquitously flow to, from, and through the time dimension. The combined continued existence of numerous masses **proves** conclusively that the flow of time can have a myriad internal electromagnetic energy flows. An equilibrium between (i) an inflow of EM energy to a transducer from the time dimension, and (ii) an **outflow of** EM energy in 3-space from the transducer, will be seen as a discrete excitation (potential energy) associated with the transducer. Hence **the** notion of the charge. Every charge in the universe may be said to pour out

energy (negative charge pours out positive energy; positive charge absorbs positive energy — the equivalent of pouring out negative energy — and any charge results in a polarization of the vacuum. Hence what we call "spacetime" or "vacuum" is in fact a giant circulation of energy from the time domain to 3-space, and back to the time-domain. In short, a giant negentropy process continually sustains the active universe and all its components and processes.

206. Evans et al., {79}.

207. Semiz, {148}.

208. Weinberg, {194}, p. 286.

209. The available internal energy of a generator is the transduced shaft energy we input to it, say by turning the generator shaft with a steam turbine or hydro turbine. The shaft mechanical energy input is transduced into internal magnetic field energy once the rotor rotates. This internal magnetic energy is then dissipated on the internal charges inside the generator, forcing the positive charges in one direction and the negative charges in the other, thereby making the source dipole between the terminals of the generator. The available internal energy of a battery is the chemical energy possessed by it at any given time, and available for performing work on the internal charges to force them apart to form a dipole between the plates (and the battery terminals).

210. Kraus, {76}. Figure 12-60, a and b, p. 578 shows a good drawing of the Poynting (intercepted) energy flow being withdrawn into the conductors from the surrounding space. It does not show the Heaviside nondiverged component remaining, which is even larger, not diverged into the conductors, and wasted. The reason for the remaining Heaviside non-diverged component is that the electrons precess laterally across the conductor, thus withdrawing into the conductor only a small nearby section of the fields of the surface electrons — the ones (plus their fields extending into space) which intercept the impinging energy flow. Because of the short distance radially across the conductor's cross section, a surface electron has very limited "field and field energy withdrawal" length.

211. (a) J. D. Jackson, "Surface charges on circuit wires and resistors play three roles," *Am. J. Phys.*, 64(7), July 1996, p. 855-870. See also (b) Mark A. Heald, "Energy flow in circuits with Faraday emf," *Am. J. Phys., Vol.* 56, 1988, p. 540-547; (c) — "Electric fields and charges in elementary circuits," *Am J Phys.*, 52(6), June 1984, p. 522-526. The

surface charges in the conductors of a circuit are of enormous importance to the powering of the circuit.

212. See T. E. Bearden, "Energy Flow, Collection, and Dissipation in Overunity EM Devices," *Proc. 4th Internat. Energy Conf.*, Academy for New Energy, Denver, CO, May 23-27, 1997, p. 5-51. In Figure 5, p. 16 the fraction of the Poynting energy flow that is intercepted and collected by the circuit is roughly shown to be on the order of 10^{-13} of the entire Poynting energy flow available. That number is based on a very crude estimating procedure, but even if off by several orders of magnitude, it shows that the Heaviside nondiverged energy flow component is orders of magnitude greater than the diverged Poynting component, for single-pass of the energy and collection only once.

213. Rigorously, there is no *power* in an EM energy flow, regardless of how great in magnitude, if it is not altered in form or diverged. That is because "power" is rigorously the time rate of doing work (time rate of *changing the form* of energy), not the time rate of energy flow per se. Exactly, the Heaviside dark energy flow component was some 10 trillion joules per second in magnitude, but it had zero watts of power. Unfortunately, electrical engineers just loosely refer to nondiverged **energy** flow as "power", regardless of the non sequitur. Hence inane terms **widely** used in electrical engineering (and texts) such as "draw power from **the** source", etc. Rigorously, *the power* (rate at which energy is changed in form) is developed locally in each component having losses or performing work — i.e., changing the form of the energy.

214. (a) T. E. Bearden, "On the Principles of Permissible Over **Unity** EM Power Systems," *J. New Energy,* 4(2), Fall 1999, p. 16-39; (b) Bearden, {17}; (c) — "Use of Asymmetrical Regauging and Multivalued Potentials to Achieve Overunity Electromagnetic Engines," *J. New Energy,* 1(2), Summer 1996, p. 60-78; (d) — "Regauging and **Multivalued** Magnetic Scalar Potential: Master Overunity Mechanisms," *Explore,* 7(I), 1996, p. 51-58; (e) — "The Master Principle of EM Overunity and the Japanese Overunity Engines," *Infinite Energy,* 1(5&6), Nov. **1995-Feb,** 1996, p. 38-55; (f) — "Use of Regauging and multivalued Potentials to Achieve Overunity EM Engines: Concepts and Specific Engine Examples," *Proc. Internat. Sci. Conf, "New Ideas in Natural Sciences," St. Petersburg, Russia, June 17-22, 1996, Part I: Problems of Modern Physics,* 1996, p. 277-297.

215. In an oblique kind of sense, this is even "recognized" to some extent by leading electrodynamicists. E.g., quoting Jackson, {22}, p. 237: *"...the Poynting vector is arbitrary to the extent that the curl of any vector field can be added to it. Such an added term can, however, have no physical consequences."* So Jackson follows Lorentz's lead that any additional non-intercepted and uncollected energy flow permissibly accompanying the Poynting intercepted and collected component, has no physical significance. Both Lorentz and Jackson err as to the *finality* of the statement. The accompanying Heaviside component can have no physical significance *unless it is intercepted and collected by means other than the single circuit or mechanism considered, and the single pass of the entire energy flow onto and past the receiving entity (circuit, charge, etc.).*

216. E.g., see Poynting, {4a. 4b}.

217. Heaviside, {5a}.

218. Heaviside, {5b}.

219. Heaviside, {5c}. Heaviside discusses the Faraday-Maxwell ether medium, outlines his vector algebra for analysis of vectors without quaternions, discusses magnetism, gives the EM equations in a moving medium, and gives the EM flux of energy in a stationary medium. On p. 443, he credits Poynting with being first to discover the formula for energy flow, with Heaviside himself independently discovering and interpreting this flow a little later by himself in an extended form.

220. J. H. Poynting, "On the connexion between electric current and the electric and magnetic inductions in the surrounding field," *Proc. Roy. Soc.* Lond. Vol. 38, 1984-85, p. 168.

221. Oliver Heaviside, *Electrical Papers,* Vol. 2, 1887, p. 94.

222. (a) T. E. Bearden, "On Extracting Electromagnetic Energy from the Vacuum," *Proc. IC-2000,* ibid., 2000; (b) Bearden, {38a}.

223. Lorentz is believed to have done this earlier (perhaps circa 1890s). In 1902 he published the method in a book, which strongly implies it was first done in an earlier paper. See Lorentz, {109}. Figure 25 on p. 185 shows the Lorentz concept of integrating the Poynting vector around a closed cylindrical surface surrounding a volumetric element. Many thanks to Marcus Reid for furnishing a copy of the actual Lorentz reference from a library in Leipzig, Germany.

224. E.g., see (a) Panofsky and Phillips, {110}; (b) W. Gough and J. P. G. Richards, *European J. Phys.*, Vol. 7, 1986, p. 195.

225. Panofsky and Phillips, {110}, p. 180.

226. D. S. Jones, *The Theory ofElectromagnetism,* Pergamon Press, Oxford, 1964, p. 52.

227. Jackson, {22}, p. 237.

228. W. M. Schwarz, *Intermediate Electromagnetic Theory,* John Wiley & Sons, New York, 1964, p. 280-281.

229. Jones, {226}, p. 53.

230. For typical references, see (a) J. Slepian, *Am. J. Phys.*, 19, 87 (1951); (b) Mario lona, *ibid.*, 31, 398 (1963); (c) Udo Backhaus and Klaus Schafer, *ibid.*, 54, 279 (1986); (d) C. J. Carpenter, *IEE Proc. A* (UK), 136A(2), Mar. 1989, p. 55-65; (e) J. A. Ferreira, *IEEE Trans. Edit.*, 31 (4). 1988, p. 257-264; (f) Mark A. Heald,Am. *J. Phys.*, 56(6), 1988, p. 540-547. The debate has also appeared in many other leading journals, e.g., (G) T. H. Boyer, *Phys. Rev. D,* 25, 3246 (1982). Interestingly, (h) M. Abraham and R. Becker, *The Classical Theory ofElectricity and Magnetism,* Blackie. London, 1932, p. 146 and p. 194 give two examples of the controversy over the Poynting vector. Finally, see (i) D. F. Nelson *Phys. Rev. Lett.*, 76(25), June 17, 1996, p. 4713-4716 for advanced **work** requiring a greater generalization of the Poynting vector.

231. Sweet and Bearden, {29a}. Sweet's device produced 500 watts lit a 33 microwatt input, which is approximately a COP = $1.5\text{x}10^6$. By steadily increasing the COP to $3\text{x}10^6$, the weight of the Sweet unit was reduced smoothly, reaching 90% reduction. The results of this highly successful anti-gravity experiment are reported in the paper. Unfortunately, Sweet later died and never fully revealed the activation secret by which barium ferrite magnetic materials could be triggered **into** strong self-oscillation at 60 Hertz or other ELF frequency.

232. Weaker self-oscillation of such permanent magnetic materials at higher frequency is known. E.g., see (a) V. S. L'vov, *Non-Linear Spin Waves,* Moscow, 1987; (b) L'vov, {29c}. See also (c) A. G. **Gurevich,** and G. A. Melkov, *Magnetization Oscillations and Waves,* CRC **Press,** Boca Raton, Florida, 1996. The latter authors give hundreds of **additional** references.

233. E.g., (a) M. W. Evans, P. K. Anastasovski, T. E. Bearden *et al,* "Derivation of the B(3) Field and Concomitant Vacuum Energy Density from the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* 14(6), Dec. 2001, p. 589-593; (b) — "Anti-Gravity Effects in the Sachs Theory ol Electrodynamics," *Found. Phys. Lett.,* 14(6), Dec. 2001, p. 601-605; (c) Evans et al., {38b}; (d) Evans et al., {38c}; (e) -- "Operator Derivation of the Gauge Invariant Proca and Eehnert Equation: Elimination of the Lorentz Condition," *Found. Phys.,* 39(7), 2000, p. 1123-1130; (f) - "O(3) Electrodynamics from the Irreducible Representations of the EInstein Group," 2001, *Found. Phys. Lett.,* 15(2), Apr. 2002, p. 179-187.; (g) - "Effect of Vacuum Energy on the Atomic Spectra," *Found. Phys.* Lett., 13(3), June 2000, p. 289-296; (h) Evans et al., {113b}; (i) Evans et aI., {113a}; (j) - "On the Representation of the Maxwell-Heaviside Equations in Terms of the Barut Field Four-Vector," *Optik* 111(6), 2000, p. 246-248; —"Derivation of the B(3) Field and Concomitant Vacuum Energy Density from the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* (in press); (k) — "The New Maxwell Electrodynamic Equations: New Tools for New Technologies. A Collection of 60 papers from the Alpha Foundation's Institute for Advanced Study. Published as a Special Issue of the *J. New Energy,* 4(3), Winter 1999. See also (1) Bearden, (222b}; (m) Bearden, {38a}; (n) Bearden, {43d}; (o) Bearden, {43a}; (p) Bearden, {30}; (q) Bearden, {40}; (r) Bearden {214}; (s) Bearden, {1214e}; (t) - "The Final Secret of Free Energy," *Magnets,* 7(5), May 1993, p. 4-26; (u) Bearden, {214f}.

234. Bearden, {40).

235. Terence W. Barrett and Dale M. Grimes, (a) Preface, p. vii-viii, in (b) *Advanced Electromagnetism: Foundations, Theory, and Applications,* Terence W. Barrett and Dale M. Grimes (eds.), World Scientific, Singapore, 1995.

236. I. Prigogine, "Irreversibility as a symmetry-breaking process." N*ature,* Vol. 246, Nov. 9, 1973, p. 67-71.

237. Paul Davies, *Superforce: The Search for a Grand Unified Theory of Nature,* Simon and Schuster, New York, 1984, p. 105.

238. Clarke, {55}, p. 38.

239. T. D. Lee, *Symmetries, Asymmetries, and the World of Particles,* U. Wash. Press, Seattle, 1988, p. 11.

240. That is, it does not exist *openly* in the open literature. Several inventors do have working laboratory COP>1.0 experiments, and at least three of them have models which could rapidly be developed into production systems — say, in one year or less. However, the exact and detailed construction details are understandably maintained as deeply proprietary by the inventors, until their intellectual property rights can be secured, including foreign patent rights. On the other hand, many other inventors mistakenly believe they have COP>1.0 systems, but do not because of instrumental error, considering nonsinusoidal waves as measurable with RMS meters, or effects covered in Chapter 6 which they do not comprehend. In addition, there have been those more interested in selling stock than in producing a working system.

241. But see Kawai, {31}. The only available *open* details approaching the "kit of parts and instructions" point is the Kawai process, and it will work just as in the patent if carefully built. If very efficient photon-coupled switching of the flux path is used for build-ups, and if very high efficiency (0.7 to 0.8) magnetic motors are modified with the Kawai process, one can expect the COP of the modified motors to be double the conventional efficiency of the original base motor. So one can obtain COP = 1.4 to 1.6, e.g., by using the necessary and available Hitachi high efficiency motors (one at 0.7 and the other at 0.8) to start with. Hitachi engineers physically tested two high efficiency Hitachi motors modified hy Kawai, at COP = 1.4 and COP = 1.6 respectively. Unfortunately, the Kawai process, engines, and company have been seized by what can **only** be described as the Yakuza (Japanese Mafia) and are being held off **the** market. The seizure was accomplished here in Huntsville, Alabama, witnessed by the CTEC board of directors and I, when Kawai and **CTEC** were in final negotiations to rapidly develop and market the Kawai **system** Several other COP>1.0 Japanese engines have also been seized and withheld by the Yakuza.

242. Jed Z. Buchwald, *From Maxwell to Microphysics,* University of Chicago Press, Chicago and London, 1985, p. 44.

243. Whittaker, {85}.

244. Bearden, {12}.

245. Mandl and Shaw, {19}.

246. Whittaker, {91a}.

247. To see how arduous a search Kron himself made prior to athieving a proper model for an electrical circuit, see Kron, {162}. The quote is from p. 114.

248. For example, Kron, {192}.

249. P. A. M. Dirac, "A theory of electrons and protons," *Proc. Roy. Soc. Lond. A*, 126(801), Jan. 1, 1930, p. 360-365. Here Dirac introduced the vacuum energy — the electron sea filled with electrons occupying negative energy states. The holes are treated today as the positive electrons. This concept is now referred to as the "Dirac Sea," and the concept has been expanded to include negative energy "holes" for other kinds of particles. Dirac first attempted to identify these holes with protons, until later the positron was discovered.

250. Evans, {137b}.

251. (a) Evans, {137a}; (b) — "The Link Between the Topological Theory of Ranada and Trueba, the Sachs Theory, and 0(3) Electrodynamics," *ibid.*, vol. 2, p. 495-499.

252. (a) Sachs, {126a). The Sachs theory provides a great generalization of general relativity and electrodynamics reaching from the quarks and gluons to the entire universe. See also (b) Sachs, {126b}.

253. Indeed, the symmetrizing of Maxwell's equations was first accomplished by Lorenz. See (a) Lorenz, {161a}. In this paper Lorenz gave essentially what today is called the Lorentz symmetrical regauging. AS pointed out by Barrett, the Lorenz paper gave f(t-r/c) functions. Fitzgerald said that Lorenz's functions were essentially the same as his, and Fitzgerald became a leading proponent of "retarded potentials". But it is believed that Fitzgerald was unaware of Lorenz's work until the 1880's, so he is given credit for parallel development. Some people speak of Fitzgerald-Lorenz functions. This is a regauging, but the term "gauge" (inspired by railroad gauges) was first introducec by Hermann Weyl in the 1900s. He used it for a change in length and was quickly countered by Einstein - Weyl's theory was not relativistic. The idea resurfaced in the 1920s when quantum theory was being formulated, but this time it meant "change in phase" and not "change in length". At any rate, when the renowned H. A. Lorentz performed the symmetrizing of the Maxwell-Heaviside equations in the 1880s. it was adopted immediately and has been rather universally applied since then. See also (b) Jackson and Okun {161b} for a history of gauge invariance.

254. (a) Prigogine, {160a}; (b) G. Nicolis and I. Prigogine, {160d}; (c) T. Petrosky and I. Prigogine, "Quantum Chaos, Complex Spectral Representation and Time-Symmetry Breaking," *Chaos, Solitons, and Fractals,* Vol. 4, 1994, p. 311-359; (d) D. Kondepudi and I. Prigogine, "Thermodynamics, Nonequilibrium," *Encyclopedia ofApplied Physics,* Vol. 21, 1997, p. 311-337; (e) D. Kondepudi and I. Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, Chichester, 1998; (f) T. Petrosky and I. Prigogine, "Laws of Nature, Probability and Time Symmetry Breaking," *Physica A,* Vol. 263, 1999, p. 528-539.

255. (a) G. Nicolis, and I. Prigogine, *Exploring Complexity,* Piper, Munich. 1987; (b) — *Exploring Complexity: An Introduction,* Freeman, New York, 1989; (c) G. Nicolis, {160e}.

256. Letokhov, {157a-157d}.

257. Bohren, {24}.

258. Paul and Fischer, {25}.

259. E.g., see (a) Patrick et al., {37}. For rigorous explanations of **how** the energy is extracted from the vacuum, see (b) Evans et al., {38b}; (c) Evans et al., {38c}. The vacuum can and will furnish an electromagnetic "wind" from which useful EM energy can be extracted. Indeed, every source charge and dipole already extracts and continuously pours out EM energy from the vacuum, thereby establishing its fields and potentials **(and** their energy) across the universe.

260. A relationship between gamma bursters and a big bang is suggested from some recent observations. On Jan. 23, 1999 the Hubble Space Telescope caught an optical flash simultaneous with an initial gamma burst. When calculated, this burst was second in energy only to the big bang itself, and generated about 3.4 x 10^{54} ergs, which is the equivalent of converting two solar masses instantaneously into energy. There was also no polarization of the afterglow, which has been taken as evidence of beaming. However, such beaming could also be due to the beaming involved in the intense iterative phase conjugate reflections of the asymmetrical self-regauging process in intensely scattering, optically active, energy-amplifying media.

261. Heaviside, {5a-5c}.

262. Josephs, {172a}; Laithwaite, {172b}.

263. (a) Diederik Wiersma and Ad Lagendijk, "Laser Action in Very White Paint," *Physics World,* Jan. 1997, p. 33-37; (b) K. Totsuka, G. van Soest, T. Ito, A. Lagendijk, and M. Tomita, "Amplification and diffusion of spontaneous emission in strongly scattering medium,"*J. Appl. Phys.,* Vol. 87, 2000, p. 7623-7628.

264. (a) A. Lagendijk, J. Gomez Rivas, A. Imhof, and R. Sprik, "Propagation of light in disordered semiconductor materials," in *Photonic Crystals and Light Localization in the 21st Century,* edited by C. M. Soukoulis, Kluwer, Dordrecht, 2001, p. 447-473; (b) H. Cao *et at,* "Spatial Confinement of Laser Light in Active Random Media," *Phys. Rev. Lett., 8*4(24), 12 Jun. 2000, p. 5584-5587.

265. We also point out the possibility of the accumulator at each end being an intensely scattering, photoreactive, energy-amplifying material, a subject discussed a bit later in this chapter.

266. (a) Werner Triftshauser *et al, Phys. Rev. Lett.,* Vol. 87, 2001. An overview is given by (b) Philip Ball, "The Positron Probe," *Nature,* Vol. 412. Aug. 23, 2001, p. 764.

267. For example, see Letokhov, {157a-157d}.

268. Jackson, {22}, p. 363, third footnote.

269. Kraus, {10}, p. 4.

270. Marconi's patent is Gugliehno Marconi, "Transmitting Electrical Signals," U.S. Patent 586,193, July 2, 1897. Filed Dec. 7, 1896, with 56 claims. This patent was overturned in 1943 by the U.S. Supreme Court, and prior work by Nikola Tesla for transmission of EM signals was upheld. Most textbooks and journals, etc. to this day continue to falsely credit Marconi with discovery of the transmission of EM signals and thus the discovery of radio. This is not to discredit the important work that Marconi did; without him, signal transmission would have been much longer in coming.

271. United States Reports: Cases Adjudged in the Supreme Court of the United States, Vol. 320, Marconi Wireless Telegraph Co. of America vs. United States, June 21, 1943, p. 1-80. The U.S. Supreme Court ruled that Tesla's fundamental radio patents had anticipated all other contenders, including Marconi.

272. *Dictionary of Scientific and Technical Terms,* Daniel N. Lapedes (editor in chief), McGraw-Hill, New York, 1974, p. 300. On page 517 of the second and expanded edition (1978), the same definition of electric field appears and is unchanged.

273. H. J. Gray and Alan Isaacs, *A New Dictionary of Physics,* Longman, London, second edition, second impression, 1976.

274. Feynman, {7}, p. 2-4.

275. Feynman, {7}, p. 12-2.

276. P. K. L. Drude, *Ann. Physik,* Vol. 1, p. 566; Vol. 3, 1900, p. 370, 869. Here Drude first develops a theory of conductivity in a metal based on the assumption that there is a gas of free electrons in the metal. This gas of conduction electrons is often referred to as the "Drude gas." Note that this was shortly after J. J. Thomson's discovery of the electron in 1897.

277. E.g., see Jackson, {211a}.

278. Romer, {84}, endnote 24, p. 109.

279. H. Winfield Secor, "Tesla Maps Our Electrical Future."

280. Nikola Tesla, "Pioneer Radio Engineer Gives Views on Power," *New York Herald Tribune,* Sep. 11, 1932.

281. Nikola Tesla, "The True Wireless, *Electrical Experimenter,* May 1919.

282. Burford and Verner, {35a}. Quote is from p. 281.

283. That is, it does not exist *openly* in the open literature. Several inventors do have working laboratory COP>1.0 experiments, and at **least** three of them have models which could rapidly be developed into production systems — say, in one year or less.

284. But see Kawai, {31}.

285. Barrett, {124}.

286. (a) Terence W. Barrett, "Active Signalling [sic] Systems," U.S. Patent No. 5,486,833, Jan. 23, 1996; (b) — "Oscillator-Shuttle-Circuit (OSC) Networks for Conditioning Energy in Higher-Order Symmetry Algebraic Topological Forms and RF Phase Conjugation," U.S. Patent No. 5,493,691, Feb. 20, 1996.

287. No proper scientific treatise on the Moray device has ever been published, to our knowledge, nor has a proper theory of its operation been developed. Anecdotal material and some technical material is contained in (a) T. H. Moray, *The Sea of Energy,* 5th Edn., Salt Lake City, 1978 (foreword by T. E. Bearden). The fourth edition of the book, titled (b) *The Sea of Energy in Which the Earth Floats,* contains additional material but is long out of print. Additional information is given in (c) Keith Tutt, *The Search for Free Energy,* Simon & Schuster, 2001, p. 33-68. (d) Paul Brown's study of the Moray work is "The Moray Energy Device: Operational Parameters, Design Criteria, and Considerations," date unknown. Researcher and inventor Bruce Perreault has based much of his own work with his radiant energy valve on Moray's work. Perreault proposes a different interpretation. In the present book we have tried to point out the various aspects of the Moray device that resemble other presently known effects — most of which were still unknown in the 1920s and 1930s when Moray's fundamental work was accomplished.

288. Moray, {34b} "Electrotherapeutic Apparatus," U.S. Patent No. 2,460,707, Feb. 1, 1949. This patent clearly shows the complexity of the various special tubes containing his pellets that Moray developed.

289. H. *C.* Dake and Jack DeMent, *Fluorescent Light and Its Applications,* Chemical Publishing Company, Inc., Brooklyn, New York, 1941, p. 51-52 gives a succinct discussion of Stokes emission and anti-Stokes emission.

290. Sir G. G. Stokes, "On the Change of Refrangibility of Light," a memoir, 1852.

291 For example, in 1935 Prileshajewa showed that there is an energy difference involving as much as 1.1 V between the exciting light and the fluorescence of aniline vapor.

292. V. S. Letokhov, {157a-157d}.

293. Letokhov, *ibid.*

294. E.g., (a) R. N. Bhargava and D. Gallagher, "Optical Properties of Manganese-Doped Nanocrystals of ZnS," *The Am- Phys. Soc,* 72(3), Jan. 17, 1994; (b) I. A. Izmailov et al., "Generation of Stimulated Noncoherent Radiation in Light-Scattering Media Exhibiting Chemical Reactions," *Soy. J. Quantum Electronics,* 12(5), May. 1982, p. 588-594; (c) R. Pappalardo and A. Lempicki, "Brillouin and Rayleigh Scattering in Aprotic Laser

Solutions Containing neodymium," *J. Appl. Phys.,* Apr. 1992, p. 1699-1708.

295. (a) Nabil M. Lawandy *et al.,* " Laser action in strongly scattering media," *Nature,* Vol. 368, Mar. 31, 1994, p. 436-438. For a good lay article with color pictures, see (b) Ivars Peterson, "Boosted light: Laser action in white paint," *Science News,* 145(15), Apr. 9, 1994, p. 228-229. See also (c) P. Mandel, "Lasing without inversion: A useful concept?," *Contemp. Phys.,* Vol. 34, 1993, p. 335; (d) O Kocharovskaya, "Amplification and lasing without inversion," *Physics Reports,* Vol. 219, 1992, p. 175.

296. Some useful papers are: (a) Diederik S. Wiersma and Ad Lagendijk, "Light diffusion with gain and random lasers," *Phys. Rev. E,* 54(4), Oct. 1996, p. 4256-4265. The authors discuss experimental considerations for a medium that both multiply scatters and amplifies light, and distinguish three regimes; (b) H. Cao *et al.,* "Random Laser Action in Semiconductor Powder," *Phys. Rev. Lett.,* 82(11), 15 Mar. 1999, p. 2278 2281. The authors observed random laser emission in all directions, with coherent feedback, in semiconductor powder, (c) Diederik S. Wiersma, Meint P. van Albada and Ad Lagendijk, "Coherent Backscattering of Light from Amplifying Random Media," *Phys. Rev. Lett.,* 75(9), 28 Aug., **1995,** p. 1739-1742. (d) D. S. Wiersma, P. Bartolini, A. Lagendijk, and R. Righini, *Nature,* Vol. 390, 1997, p. 671. The authors report Anderson localization of photons in GaAs powder, (e) Ad Lagendijk *et al,* "Microscopic Approach to the Lorentz Cavity in Dielectrics," *Phys. Rev Lett.,* 79(4), 28 July 1997, p. 657-660. (f) Pedro de Vries and Ad Lagendijk, "Resonant Scattering and Spontaneous Emission in **Dielectrics** Microscopic Derivation of Local-Field Effects," *Phys. Rev. Lett.,* 81(7), 17 Aug. 1998, p. 1381-1384. (g) Gijs van Soest, Makoto Tomita, and Ad Lagendijk, "Amplifying volume in scattering media," *Opt. Lett.,* 24(5), Mar. 1, 1999, p. 306-308.

297. S. Itoh *et al.,* "Simulational and experimental studies on anomalous reflectivity of phase conjugate wave," *Ferroelectrics,* vol. 170, 1995, p. 209-217.

298. See (a) D. P. Hand and P. St. J. Russell, "Solitary thermal shock waves and optical damage in optical fibers," *IEE Colloquium on 'Non Linear Optical Waveguides', LEE Digest* No. 88, London, England, 1988, p. 101-103; (b) — "Soliton-like thermal shock-waves in optical fibres: origin of periodic damage tracks," *Fourteenth European Conference on*

Optical Communication (ECOC 1988), Sep. 11-15, 1988, Publication No. 292, 1(2), 1988, p. 111-114; (c) — "Solitary thermal shock waves and oplical damage in optical fibers: the fiber fuse," *Opt. Lett.,* 13(9), Sep. **1988,** p. 767-769; (d) D. P. Hand and T. A. Birks, "Single-mode tapers as 'fibre fuse' damage circuit-breakers," *Electronics Lett.,* 25(1), Jan. 5, 1989, p. 33-34. For an excellent lay description with photographs, see (e) Ivars Peterson, "Fibers with Flare," *Sci. News,* 140(13), Sep. 8, 1991, p. 200-201

299. P. St. J. Russell, "Power Conservation and Field Structures in Uniform Dielectric Gratings," *J. Opt. Soc. America A.,* 1(3), Mar. 3, 1984, p. 293-299. Russell is referring to apparent violation of local conservation of energy. It is true in the conventional flawed energy flow analysis considering only the Poynting component, because the fuse produces more work output (melting the holes) than the original operator's input of Poynting energy to initiate the fuse effect. It does not produce more work output than the total of the energy input by the operator in both the acounted Poynting component and the unaccounted Heaviside component. It merely shows that the extra fiber fuse ignition actions do absorb extra energy from the Heaviside component, perhaps by particle resonance (Bohren effect) of the melt and vapor particles. Also, intense heating produces intense ionization and thus intense dipoles, thus producing broken symmetry in the fierce exchange with the active vacuum. Multiple retroreflections of the EM energy flow also occur in the nmelt and vapor particles, leading to multiple collections in the ping-pong of a joule of energy between two particles as phase conjugate mirrors.

300. E.g., see (a) H. Kogelnik, "Coupled wave theory for thick hologram gratings," *Bell Systems Technical Journal,* Vol. 48, 1969, p. 2909-2947; (b) Z. Zylbergerg and E. Marom, "Rigorous coupled-wave analysis of pure reflection gratings," *J. Opt. Soc. Am.,* Vol. 71, 1981, p. 811-818; (c) — - "Rigorous coupled-wave analysis of grating diffraction — E- mode polarization and losses," *ibid.,* Vol. 73, 1983, p. 451-455; (d) - "Comments on analyses of reflection gratings," *ibid.,* Vol. 73, 1983, p. 399-401.

301. Bohren, {24J.

302. Justin Mullins, "Superpowerful Explosive Arrives with a Bang." New *Scientist,* Aug. 4, 2001.

303. Romano M. DeSantis *et al.*, "On the Analysis of Feedback Systems With a Multipower Open Loop Chain," Oct. 1973, AD 773188, available through the Defense Technical Information Center or the National Technical Information Center. The paper clearly shows that such systems can produce overunity coefficient of performance.

304. See (a) J. H. Andreatta, "High Power Switching Amplifier Wherein Energy is Transferred to a Tuned Circuit During Both Half Cycles," U.S. Patent No. 3,239,771, Mar. 8, 1966; (b) Tom L. Dennis Jr., "Highly Efficient Semiconductor Switching Amplifier," U.S. Patent No. 3,239,772, Mar. 8, 1966; (c) Heber J. Morrison, "Square Wave Driven Power Amplifier," U.S. Patent No. 3,815,030, June 4, 1974.

305. For a good but conventional coverage of many differential negative resistors and negative conductors, see Adolf Fritz Schwarz, "Negative Resistors and Negative Conductors," Doctoral Thesis, Technische Hogeschool, Delft (Netherlands), Feb. 1969, available from the U.S. National Technical Information Service (NTIS), Springfield, VA. True negative resistors are dismissed as impossible in Schwarz's opening remarks. And they *are* impossible in equilibrium systems not receiving and using any net excess energy asymmetrically from their environment, which are the only systems conventionally considered by electrical engineers and electrodynamicists.

306. Gabriel Kron, "Numerical solution of ordinary and partial differential equations by means of equivalent circuits." *J. Appl. Phys.*, Vol. 16, Mar. 1945a, p. 173.

307. Kron, {32}, p. 39.

308. Bearden, {12}.

309. E.g., see Gabriel Kron, "Equivalent Circuits for Oscillating Systems and the Riemann-Christoffel Curvature Tensor," *General Electric Review*, Jan. 1943, p. 25-31.

310. See (a) S. Austen Stigant, "Gabriel Kron on Tensor Analysis, A bibliographical record," *BEAMA Journal*, Aug. 1948, for a full bibliography of Kron's publications up to 1948. Also useful are (b) Gabriel Kron, "A Method to Solve Very Large Physical Systems in Easy Methods," *IRE Transactions on Circuit Theory*, 2(1), Dec. 1953, p. 71-90; (c) P. L. Alger, editor, *The Life and Times of Gabriel Kron, or Walking Around the World, and Tensors*, Mohawk Development Services, Inc.,

Schenectady, NY, 1969; (d) Banesh Hoffman, " Cron's Non-Riemannian Electrodynamics," *Rev. Mod. Phys.*, 21(3), 1949, *x* 535-540; (e) K. Kondo and Y. Ishizuka, "Recapitulation of the Geometrical Aspects of Gabriel Kron's Non-Riemannian Electrodynamics," Research Association of Applied Geometry, *Memoirs of the Unifying Study of Basic Problems in Engineering and Physical Sciences by Means of Geometry*, Gakujutsu Bunken Fukyu-kai, Tokyo, Vol. I, 1955, p. 185-239 [particularly see footnote 1, p. 222];

311. Wang and Chung, {33}.

312. Deborah D. L. Chung, "Superconductor-metal laminates and method of making," U.S. patent no. 5,059,582, issued Oct. 22, 1991. A superconducting laminate having at least one layer of metal and at least one layer of superconducting material. The metal layer and the superconducting layer are bonded. The layer may also include carbon fibers from various precursors. The superconductor may be a composite material. The invention also includes a method of making the laminates. Assigned to The Research Foundation of State University of NY, Albany, NY.

313. Jean-Louis Naudin, website http://jnaudin.free.fr/cnr/.

314. Fogal, {36a, 36b}.

315. As one example, Fogal's device has to be able to manipulate Dirac holes in the vacuum itself, at least to some extent. Once the erroneous flat spacetime assumption of classical electrodynamics is removed and curved local spacetime considered, we can and will have Dirac holes in the local vacuum itself, especially in COP»1.0 systems. In a COP»1.0 system, not all the Dirac holes have interacted with orbital electrons in atoms to "eat" an electron and produce a positive ion — with a "lattice hole" or "lattice positron" where the electron had been. There is a vast difference between a Dirac hole positron (which has a negative mass equal in magnitude to one electron-mass) and a lattice hole (which is attached to the relatively huge positive mass of the ionized atom. Hence the mobility and phenomenology of Dirac positrons and lattice positrons are very different. There presently are no books on the dynamics of the Dirac negative energy electron holes in COP»1.0 EM circuits and systems.

316. Mandl and Shaw, {19}.

317. Whittaker, {91a}. The paper was published in 1904 and orally delivered in 1903. Note: On p. 368 midpage, Whittaker's equation for the y-component of vector-potential "a" is missing the exponent "2" on the **del** operator.

318. Rodrigues and Lu, {93a}.

319. Rodrigues and Vaz, {93b}.

320. It is certainly not lacking in the secret military weapons game. Some 10 nations of the world have weaponized longitudinal EM wave weapons over the past few decades, and almost every major weapons laboratory in the world has now "discovered" longitudinal EM waves. Even the Yakuza (Japanese Mafia) has acquired its own portable longitudinal EM interferometry weapons.

321. Evans et al., {79}.

322. (a) V. I. Arkhipov, *et al.*, "Negative transient currents in amorphous semiconductors." *Internat. J. Electronics* (UK), 51(6), 1981, p, 735-742; (b) — and A. I. Rudenko, "Negative currents caused by injection-controlled polarization," *Solid. St. Commun.*, 28, 1978, p. 675-676.

323. Jackson, {22}, p. 222.

324. Jackson, {22}, p. 223.

325. O. L. Brill and B. Goodman, *Am. J. Phys.*, vol. 35, 1967, p. 832.

326. For example, see Jackson, {22}, p. 219-221.

327. We stress that electrodynamicists, conventional theory, and quantum field theory all assume the principle *of gauge freedom,* the **total** freedom to change the gauge (and therefore any potential) in a Maxwellian system at any time, without cost or work. In the real world one may have to pay a little switching costs, but by simply regauging one can freely change the EM energy physically present at a point in 3-space **occupied by** the Fogal "freed" instantaneous scalar potential, once the electrons are pinned. In short, the Fogal process is able to directly "gate" the instantaneous transduction of EM energy existing at multiple points **in** space, simultaneously. There is nothing in the gauge freedom **principle** that restricts (i) regauging the instantaneous scalar potential at will, **or** (ii) using these changes at a distance to produce force fields upon electrons, thereby moving them — so long as one employs a Fogal semiconductor in

the "distant receiver" in order to be able to produce the back-transition from the field-free state to the force-field state by removing the anomalous pinning restriction. To the advanced theorists, we leave the debates over causality considerations for this demonstrated process!

328. Whittaker, {85}.

329. (a) W. Heitman and G. Nimtz, "On Causality Proofs of Superluminal Barrier Transversal of Frequency Band Limited Wave Packets," *Phys. Lett. A196,* 1994, p. 154-158; (b) Enders and Nimtz, {136b}.

330. The reader should recall that the Lorenz/Lorentz condition, which limits the circuit to COP< 1.0, is self-applied by the circuit when it rams all "spent" electrons in the ground return line back through the emf of the circuit. One need not do that, and need not produce that force field that destroys the primary source dipole and its extraction of EM energy from the vacuum via its broken symmetry. As the reader has undoubtedly now observed, the "back-emf' and "back mmf are those aspects in an electrical circuit and magnetic circuit respectively, that kill the dipole and physically self-apply and self-enforce Lorenz/Lorentz symmetrical regauging. Hence the intuitive concern of COP>1.0 experimenters to try to discover how to avoid or lessen the "back emf or the "back mmf in their circuits and systems.

331. Philip Ball, "The positron probe," *Nature,* Vol. 412,23 Aug. 2001, p. 764.

332. Werner Triftshauser et al., *Phys. Rev. Lett.,* 87, 067402 (2001).

333. Bearden, {30}.

334. For the chemistry and currents of the battery, see any good battery reference, such as Colin A. Vincent and Bruno Scrosati, *Modern Batteries: An Introduction to Electrochemical Power Sources,* Second Edition. Wiley, New York, 1997.

335. (a) J. O'M. Bockris, A. K. N. Reddy, and M. Gamboa-Aldeco, *Modern Electrochemistry: Fundamentals of Electrodics,* 2nd edition. Kluwer/Plenum, NY 2000; (b) J. O'M. Bockris and A. K. N. Reddy, *Modern Electrochemistry: Electrodics in Chemistry, Engineering, Biology, and Environmental Science,* 2nd edition, Kluwer/Plenum, NY 2000; *(c)* J. Koryta, J. Dvorak, and L. Kavan, *Principles of Electrochemistry,* 2nd ed.. Wiley, NY, 1993; (d) Wolfgang .1. Lorenz and Waldfried Plieth,

Electrochemical Nanotechnology: In Situ Local Probe Techniques at Electrochemical Interfaces, Wiley, NY, 1998; (e) B. E. Conway, *Electrochemical Supercapacitors: Scientific Fundamentals and Technological Applications,* Plenum, NY, 1999; (f) Allen J. Bard and Larry R. Faulkner, *Electrochemical Methods: Fundamentals and Applications,* Second Edition, Wiley, NY, 2001.

336. At this writing, a provisional patent application on the process has been filed by Bedini and the present author. By the time of publication, hopefully this will have been turned into a formal patent application, with foreign patents also applied for.

337. Edwin V. Gray, "Pulsed Capacitor Discharge Electric Engine," U.S. Patent #3,890,548, June 17, 1975.

338. (a) Per Whittaker, {85}. However, see (b) Bearden, {12} for a necessary correction of Whittaker's interpretation to accord with a quantum field theory view of the instantaneous scalar potential. After observation, the phase conjugate EM wave half of the Whittaker wavepair is indeed going in the opposite direction to the normal longitudinal EM wave half But prior to observation, it exists in the time domain, and in the time domain it is precisely accompanying the 3-space longitudinal EM wave, with the absorbing and then emitting charge interacting in the middle between the two. See (c) Mandl and Shaw, {19}, to see that the combined scalar (time-polarized) photon and longitudinal photon are observable as the instantaneous scalar potential. This means that the corrected Whittaker decomposition produces an associated time-polarized EM wave in the time domain and longitudinal EM wave in 3-space. That is the corrected interpretation of Whittaker's phase conjugate longitudinal EM wavepair.

339. (a) J. O'M. Bockris, "Overpotential: a lacuna in scientific knowledge,"J. *Chem. Edu.,* 48(6), June 1971, p. 352-358. This is a simplified introduction to overpotential theory. The energy density that exists at minute "double layers" in contact, which act as tiny capacitors with extremely small separation of plates, is enormous and far greater **than** what the power engineer meets in his normal engineering practice. **Sec** also (b) Bockris and Reddy, {335b}, particularly Chapter I which deals with the overpotential. See also (c) Bockris et a1, {335a}.

340. E.g. see (a) L. Bertalot *et al.,* "Study of deuterium charging in palladium by the electrolysis of heavy water: heat excess production", *Nuovo Cimento,* Vol. 15 D, 1993, p. 1435; (b) J. O'M Bockris *et al.,* "Cold

fusion as a consequence of high fugacity among hydrogen isotopes," Intl. *J. Hydrogen Energy,* Vol. 17, 1992, p. 445.

341. Peter Lindemann, *The Free Energy Secrets of Cold Electricity,* **Clear** Tech Inc., Metaline Falls, WA, 2001.

342. Particularly see Kenneth S. Deffeyes, *Hubbert's Peak: The Impending World Oil Shortage,* Princeton University Press, Princeton, NJ, **2001.**

Quoting: *"Deffeyes finds that world oil production will peak within five wars and that there isn't anything we can do to stop it. New exploration and production technologies can't save us, and plans to open the Arctic National Wildlife Refuge and other areas to drilling offer no more than a small and soon-to-be-forgotten blip on the production curve. While long-term solutions exist in the form of conservation and alternative energy sources, they probably cannot — and almost certainly will not — be enacted in time to evade a short-term catastrophe of shortages, soaring prices, and global economic, agricultural, and possibly political disturbance. "* In short, Deffeyes rigorously demonstrates that a world energy crisis of staggering proportions is just around the corner.

343. D. J. Bohm and B. J. Hiley, "On the intuitive understanding of nonlocality as implied by quantum theory," *Found. Phys.,* 5(1), Mar. 1975, p. 93-109. Quote is from p. 107.

344. (a) Y. Aharonov and D. Bohm, "Significance of Electromagnetic **Potentials** in the Quantum Theory," *Phys. Rev.,* Second Series, 115(3), Aug. 1, 1959, p. 485; (b) — "Further Considerations on Electromagnetics In **the** Quantum Theory," Phys. Rev., 123(4), Aug. 15, 1961, p. 1511-1524.

345. Terence W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," A. Lakhtakia, ed., *Essays on the Formal Aspects of Electromagnetics Theory,* World Scientific Publishing, River Edge, NJ, **1993,** p. 11 and generally p. 6-86.]

346. Davies, {237}, p. 105.

347. David J. Bohm, "A Suggested Interpretation of the Quantum **Theory** in Terms of 'Hidden' Variables, I and II." *Phys. Rev.,* 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II).

348. Jackson, {22}, p. 223.

349, Lamb and **Retherford,** {23}.

350. (a) Casimir, {26}. This paper is an extrapolation to two attracting plates of the more detailed paper of (b) H. B. G. Casimir and D. Polder, "The influence of retardation on the London-van der Waals Forces," *Phys Rev.* 73(4), Feb. 15, 1948, p. 360-372. Casimir's paper was an extension of earlier work by Verwey and Overbeek on the stability of colloidal systems, including the interaction of electric double layers; see (c) E. J. W Verwey, J. T. G. Overbeek, and K. van Nes, *Theory of the Stability of Lyophobic Colloids,* Elsevier, Amsterdam, 1948.

351. Lamoreaux, {27}.

352. Mead and Nachamkin, {28}.

353. E.g., see (a) P. W. Milonni. R. J. Cook and M. E. Goggin, "Radiation pressure from the vacuum: Physical interpretation of the Casimir force," *Phys. Rev. A,* Vol. 38, 1988, p. 1621; (b) C. E. Carlson, T. Goldman and J. Peres-Mercader, "Gamma-ray bursts, neutron star quakes, and the Casimir effect," *Europhys. Lett.,* Vol. 36, 1996, p. 637; (c) M. Chaichian *et ah,* "Quantum theories on noncommutative spaces with nontrivial topology: Aharonov-Bohm and Casimir effects," *Nucl Phys B.* Vol. 611, 2001, p. 383-402; (d) V. Dodonov, "Nonstationary Casimir Effect and Analytical Solutions for Quantum Fields in Cavities with Moving Boundaries," in *Modern Nonlinear Optics,* 2nd Edn., 3 vols., Wiley, 2001, Vol. 1, p. 309-384.

354. Bohm, {347}.

355. See T. E. Bearden, *Gravitobiology: A New Biophysics,* Tesla Book Co., Chula Vista, CA, 1991.

356. One should not be nonplussed at the notion of infinite velocity. A pure longitudinal EM wave is — at least in theory — capable of moving at infinite velocity, and is a proper solution to Maxwell's equations. See (a) Rodrigues and Lu, {93a}. Also, the ordinary potential in the Coulomh gauge does indeed move at infinite velocity; see Jackson, {22}, p. 223. In addition, Bohm's interpretation of quantum mechanics {347} uses the quantum potential, which instantaneously appears everywhere it will have magnitude.

357. Enders and Nimtz, {329b}.

358. R. Chen, "Cancellation of Internal Forces," *Am. J. Phys.,* 49(4), Apr. 1981, p. 372.

359. Whittaker, {85}.

360. R. L. Davis, "Non-Inductive Electrical Resistor," U.S. Patent No. 3,267,406, Aug. 16, 1966. Davis uses a Mobius-loop capacitor as an inductance-free resistor. Our comment is that, more rigorously, the resistor is *not* noninductive. It is "equally and oppositely inductive," which is quite a different thing altogether when one considers the difference between internal energy and external energy. We hypothesize that the "resistor" — or some version of it — produces an artificial scalar potential stress in spacetime — apparently varying the stress energy tensor itself— even though it produces a *net* zero reactance.

361. These are Brazil, Russia, China, and two other nations friendly to the U.S. The dominant weapons on earth today are quantum potential weapons, not nuclear weapons. The second most dominant weapons are probably the negative energy EMP weapons deployed by several nations, including China.

362. E.g., see (a) L. I. Mandelstam and N. D. Papaleksi, "On the parametric excitation of electric oscillations," *Zhurnal Teknicheskoy Fiziki,* 4(1), 1934, p. 5-29; (b) — "On resonance phenomena with frequency distribution." *Z.f. Phys.,* No. 72, 1931, p. 223: (c) — "Concerning asynchronous excitation of oscillations," *Zhurnal Tekhnicheskoi Fiziki,* 4(1), 1934; (d) — "Concerning asynchronous excitation of oscillations," *Zhurnal Tekhnicheskoi Fiziki,* 4(1), 1934; (e) —- "Concerning nonstationary processes occurring in the case of resonance phenomena of the second class," *Zhurnal Tekhnicheskoi Fiziki,* 4(1), 1934. See also (f) A. Andronov, "The limiting cycles of Poincare and the theory of self-maintained oscillations," *Comptes-Rendus,* Vol. 189, 1929, p. 559; (g) — and A. Witt, "On the mathematical theory of self-excitations," *Comptes-Rendus,* Vol. 190, 1930, p. 256; (h) — "On the mathematical theory of self-excitation systems with two degrees of freedom," *Zhurnal Tekhnicheskioi Fiziki,* 4(1), 193; (i) — "Discontinuous periodic movements and theory of multivibrators of Abraham and Bloch," *Bull. De l'Acad. Ed Sc. De l'URSS,* vol. 189, 1930. See (j) S. Chaikin, "Continuous and 'discontinuous' oscillations," *Zhurnal Prikladnoi Fiziki,* Vol. 7, 1930. p. 6; (k) — and A. Witt, "Drift in a case of small amplitudes," *Zhurnal Teknicheskoi Fiziki,* 1(5), 1931, p. 428; (1) — and N. Kaidanowski, "Mechanical relaxation oscillations," *Zhurnal Teknicheskoi Fiziki,* Vol. 3, 1933, p. 1.

363. L. Mandelstam [L. T. Mendel'shtam], N. Papalexi, A. Andronov, S. Chaikin and A. Witt, "Report on Recent Research on Nonlinear Oscillations," Translation of "Expose Des Recherches Recentes Sur Les Oscillations Non Lineaires," *Technical Physics of the USSR,* Leningrad, Vol. 2, 1935, p. 81-134 [NASA Translation Doc. TTF-12,678, Nov. 1969].

364. In Alexander L. Fradkov and Alexander Yu. Pogromsky, *Introduction to Control of Oscillations and Chaos,* World Scientific Series on Nonlinear Science E, Series Editor Leon O. Chua, World Scientific, New Jersey, 1998 — as a typical example — I found no mention of the Dirac hole current decay mechanism or of the active external environment (the active vacuum and its dynamics, and the active local curvatures of spacetime and their dynamics).

365. Richard E. Prange and Peter Strance, "The Semiconducting Vacuum," *Am. J. Phys.,* 52(1), Jan. 1984, p. 19-21.

366. John F. Lindner *et ah,* "Noise Enhanced Propagation," *Phys. Rev. Lett.,* 81(23), 7 Dec. 1998, p. 5048-5051. Lindner suggests that the mechanism appears to be general, in which case it could supply the mechanism that coherently "locks together" the barium nuclei self-oscillations in Sweet's activated barium ferrite magnets.

367. Owen Flynn, "Parametric arrays: A new concept for sonar," *Electronic Warfare Magazine,* June 1977, p. 107-112. Any two sine-wave frequencies as simultaneous drivers combine to produce a sine-wave difference frequency propagating in an isotropic nonlinear medium, essentially without sidebands or reverberations. Its pattern has a main lobe approximately equal to that of the high frequency drive, but devoid of sidelobes. The level of the propagating difference frequency is proportional to both the product of the two fundamental drive levels and to the square of the desired value of difference frequency.

368. M. I. Dykman *et al.,* "Noise-enhanced heterodyning in bistable systems." *Phys. Rev. E,* 49(3), Mar. 1994, p. 1935-1942

369. (a) V. S. L'vov and L. A. Prozorova, "Spin Waves Above the Threshold of Parametric Excitation," in (b) A. S. Borovik-Romanov and S. K. Sinha, Eds., *Spin Waves and Magnetic Excitations,* North-Holland, Amsterdam, 1988.

370. Gurevich and Melkov, {232c}. The authors treat high **frequency** processes, i.e., oscillations and waves, in magnetically ordered substances, including ferromagnetic and antiferromagnetic resonances, spin waves, nonlinear processes, and high frequency manifestations of **interactions** between the magnetic system and other systems of magnetically **ordered** substances (crystal lattice, charge carriers). The treatment is classical except that a quantum mechanical approach is used in the study of relaxation processes, which determine the energy losses of magnetic oscillations and waves. Experimental results but not techniques are included. Self-oscillation and self-oscillation criteria arc on p. 279.

371. (a) G. S. Kandaurova and A. E. Sviderskii, "Observation of autowave state and stable dynamic structures in multidomain magnetic films," *Sov. Tech. Phys. Lett.,* 14(5), May 1988, p. 346-347; (b) V. S. **L'vov,** {29c}.

372. Richard T. Hammond, "New fields in general relativity," *Contemp. Phys.,* 36(2), 1995, p. 103-114.

373. (a) J. Baugh et al., *Science,* Vol. 294, 16 Nov. 2001, p. 1505. For a simpler synopsis of the paper, see (b) Warren S. Warren, "Multidimensional Symmetry in a Three-Dimensional World," *Science,* Vol. 294, 16 Nov. 2001, p. 1475-1476.

374. E.g., see Howard R. Johnson, "Permanent Magnet Motor." U.S. Patent No. 4.151.431. Apr. 24. 1979.

375. Howard Johnson, private communication, January 2002.

376. In magnetics, multi-valued potentials (and hence the possibility **for** nonconservative fields) arise naturally in the theory, but are considered a nuisance to be eliminated or "averaged away" by magnetics engineers. E.g., see (a) Z. Ren, "A 3D vector potential formulation using edge element for electrostatic field computation," *IEEE Trans, on Magnetics,* **31(3),** May 1995, p. 1520-1523; (b) — "New technique for solving three-dimensional multiply connected eddy-current problems." *lEEProc, 1*37A(3), May 1990, p. 135-140; (c) J. C. Verite, "Calculation of multivalued potentials in exterior regions," *IEEE Trans, on Magnetics,* **Vol.** MAG-23(3), May 1987, p. 1881-1887. Amazingly, these and other similar methods are directed into converting the multivalued potential into single-valued regions by a process of multiple, highly selected cuts. In other words, the real attempt is to eliminate the multivalued aspect, in which case the theorists are in fact simply eliminating the asymmetrical

self-regauging provided by the multivalued potential and hence are discarding the ability to use multivalued potentials in true overunity processes.

377. (a) Z. Badics, "Transient eddy current field of current forced three-dimensional conductors," *IEEE Trans, on Magnetics,* 28(2), Mar. 1992, p. 1232-1234;(b) R. Albanese and G. Rubinacci, "Magnetostatic field computations in terms of two component vector potentials," *Int. J. for Numerical Methods in Engineering,* Vol. 29, 1990, p. 515-532; (c) Bearden, {214f}: (d) P. Hammond and T. D. Ssiboukis, "Dual finite-element calculations for static electric and magnetic fields," *IEEProc. A,* 130(3), 1983, p. 105-111; (e) P. R. Kotiuga, "On making cuts for magnetic scalar potentials in multiply connected regions," *J. Appl. Phys.,* 61(8), 1987, p. 3916-3918; (f) J. Penman and J. R. Fraser, "Complementary and dual energy finite element principles in magnetostatics," *IEEE Trans, on Magnetics,* 18(2), 1982, p. 319-323; (g) E. A. Tolkachev and Ya. M. Shnir, "Non-trivial topology effects in the interaction of a monopole with a line current," *Phys. Lett. A,* 155(8, 9), May 27, 1991, p. 464-466.

378. (a) Feynman, {7}, Vol. II, Chapter 37; (b) T. Ambrose, R. L. Sommer, and C. *L.* Chien, "Angular Dependence of Exchange Coupling in Ferromagnetic/Antimagnetic Bilayers," *Phys. Rev. B,* Vol. 56, 1997, p. 83; (c) H. Sang, Y. W. Du, and C. L. Chien, "Exchange Coupling in $Fe_{50}Mn_{50}/Ni_{81}Fe_{19}$ Bilayers: Dependence on Antiferromagnetic Layer Thickness," *J. Appl. Phys.,* Vol. 85, 1999, p. 4931.

379. V. A. Tulin, "Nuclear Spin Waves in Magnets," *Fis. Niskikh Temp.,* Vol. 5, 1979, p. 965-993 [in Russian].

380. Sweet and Bearden, {29a}.

381. (a) Kenneth R. Shoulders, U.S. Patent # 5,153,901; (b) — "Energy Conversion Using High Charge Density," U.S. Patent # 5,018,180, May 21, 1991; (c) — "Energy Conversion using High Charge Density," U.S. Patent No. 5,123,039, 1992; (d) — U.S. Patents 5,054,046 and 5,054,047, 1991; (e) — "Circuits Responsive To and Controlling Charged Particles," U.S. Patent 5,148,461, 1992; (f) — and Steve Shoulders, "Observations on the Role of Charge Clusters in Nuclear Cluster Reactions," *J. New Energy,* 1(3), Fall 1996, p. 111-121.

382. (a) Paulo N. Correa and Alexandra N. Correa, "Pulsed Abnormal Glow Discharge Technology," *Infinite Energy,* 2(7), Mar./Apr. 1996, p.

18-21; (b) — "Other Applications of the PAGD Technology Besides Energy Conversion," *ibid.*, p. 22-27; (c)— "Electrochemical Transduction of Plasma Pulses," U.S. Patent No. 5,416,391, May 16, 1995; (d) "Energy Conversion Systems," U.S. Patent No. 5,449,989, Sept. 12, 1995; (e) — "Direct Current Energized Pulse Generator Utilizing Autogenous Cyclical Pulsed Abnormal Glow Discharges," U.S. Patent 5,502,354, Mar. 26, 1996.

383. In 1935, Prileshajewa showed that there is an energy difference as much as 1.1 v between the exciting light and the fluorescence of aniline vapor. This added energy is attributed to additions from the internal energy of the molecule. Thus, anti-Stokes emission has been validated as a COP>1.0 mechanism since at least 1935. For our purposes, replenishment of the internal energy of the molecule must be obtained from the local active vacuum; else one simply has the standard optical cooling process which cools the gas.

384. E.g., see J. K. Brasseur et al., "Coherent Anti-Stokes Emission in a Continuous Wave Raman Laser in H_2," carried on website location http://www.physics.montana.edu/optics/jlc/Anti-Stokes.pdf.

385. For a typical study, see J. Silver et al., "Studies on anti-Stokes emission from metal oxides (doped with Er3+ and Yb3+) stimulated by 632.8 nm laser excitation," *6th Int. Conf. Phosphors and Display Tech.*, San Diego, California, 2000.

386. For a typical paper of hundreds on the subject, see C. Zander and K. H. Drexhage, "Cooling of a Dye Solution by Anti-Stokes Fluorescence," in *Advances in Photochemistry*, edited by D. C. Neckers, D. H. Volman, and G. von Bulau; Wiley, New York, 1995.

387. (a) Antoine Priore, "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; (b) — "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968. For an explanation, see (c) T. E. Bearden, "The Porthole Briefing," http://www.cheniere.org.

388. Thomas Valone, "Future Energy Technologies," *J. New Energy*, 5(2), Fall 2000, p. 41-46; summary is quoted from p. 44-45.

389. Shang-Xiang Jin and Hal Fox, "Characteristics of High-Density Charge Clusters: A Theoretical Model," *J. New Energy*, 1(4), Winter 1996, p. 5-20.

390. Shoulders, {381a-381e}.

391. Ken Shoulders, "Permittivity transitions," *J. New Energy,* 5(2), Fall 2000, p. 121-137.

392. *Ibid,* p. 122.

393. *Ibid,* p. 122.

394. K. Shoulders and S. Shoulders, {381 f).

395. Harold L. Fox, "Energy for the Future: High-Density Charge Clusters," *Modern Nonlinear Optics,* Second Edition, edited by M. W. Evans, Wiley, New York, 2001, Part III, p. 623-653.

396. See (a) Art Rosenblum, "Randell Mills — New Energy and the Cosmic Hydrino Sea," *Infinite Energy,* 3(17), Dec. 1997-1998, p. 21-34. Also see (b) Randell Lee Mills, "Energy/Matter Conversion Methods and Structures," Australian Patent No. 668678, Nov. 20, 1991. Particularly see (c) Eugene Mallove, "Dr. Randell Mills and the Power of BlackLight," *Infinite Energy,* 2(12), Jan.-Feb. 1997, p. 21, 35, 41.

397. Randell Mills *et al.,* "Lower-Energy Hydrogen Methods and Structure," U.S. Patent No. 6,024,935, Feb. 15, 2000. Some 60 pages, and an astounding 499 claims. For the patent, an independent verification was performed by Johannes Conrads, Institute for Low Temperature Plasma Physics at the Ernst Moritz Arndt University in Greifswald, Germany.

398. Randell L. Mills, "The Grand Unified Theory of Classical Quantum Mechanics," www.blacklightpower.com/book/shtml.

399. (a) Randell L. Mills et al., "Identification of Compounds Containing Novel Hydride Ions by Nuclear magnetic Resonance Spectroscopy"; (b) — "Spectroscopic Identification of Novel Catalytic-Reaction of Atomic Hydrogen and the Hydride Ion Products"; (c) — "Black Light Power Technology: A New Clean Energy Source with the Potential for Direct Conversion to Electricity," presented to the Intl. Conf. On Global Warming and Energy Policy, Nov. 26-28, 2000 and to the National Hydrogen Association Meeting, March 6-8, 2001; (d) — Plasma Power Source Based on a Catalytic Reaction of Atomic Hydrogen"; (e) R, M. Mayo and R. L. Mills, "Direct Plasmadynamic Conversion of **Plasma** Thermal Power to Electricity for Microdistributed Power Applications," *40th Ann. Power Sources Conf. Proc*, June 10-13, 2002; (f) R. Mills et al., "Novel Catalytic Reaction as a Potential New Energy Source," 223rd

ACS National Meeting, Orlando, Florida, Apr. 6-11, 2002; (g) — "Novel alkali and Alkaline Earth Hydrides for High Voltage and High Energy Density Batteries," *Proc. 17th Ann. Battery Conf. On Appl. And Advances, Calif. State Univ., Long Beach, CA, Jan. 15-18* (in press); (h) - "Synthesis of Diamond Films from Solid Carbon"; (i) Mills and Ray, "Spectral Emission of Fractional Quantum Energy Levels of Atomic Hydrogen from a Helium-Hydrogen Plasma and the Implications for Dark Matter," *Int. J. Hydrogen Energy,* 27(3), 2002, p. 301-322.

400. Peter D. Zimmerman, "An Analysis of Theoretical Flaws in So-Called Classical Quantum Mechanics and of Experimental Evidence against CQM," 2001.

401. Bearden, {222b}, p. 674.

402. E.g., it yields the excess deuterium, tritium, and alpha particles; see Bearden, {17}.

403. (a) Bearden, {222b}; Bearden, {15}.

404. (a) "Ion Cyclotron and Converter and Radio Power Microwave Generator," International Pub. No. WO 01/21300 A2, 29 March 2001; (b) "Ion Cyclotron Power Converter and Radio and Microwave Generator," International Pub. No. WO 01/22472, 29 March 2001. Both filed with World Intellectual Property Organization.

405. But see Samuel D. Harper. "The energy dissipated in a switch," *Am. J. Phys.*, 56(10), Oct. 1988, p. 866-889. Unless special attention and effort is paid, switches can indeed consume a great deal of energy! As Harper points out, it can be a substantial fraction of the energy transfer in the circuit. In a capacitive loaded circuit this is especially true. One should consider very efficient switching such as optically coupled switching arrangements.

406. (a) F. Heinrich, "Entropy change when charging a capacitor: A demonstration experiment," *Am. J. Phys.*, 54(8), 1986, p. 742-44. For an expanded discussion, see (b) I. Fundaun, C. Reese, and H. H. Soonpaa, "Charging a capacitor," *Am. J. Phys.*, 60(11), Nov. 1992, p. 1047-1048. Also see (c) Tyler A. Abbott and David J. Griffiths, "Acceleration without radiation." *Am. J. Phys.*, 53(12), Dec. 1985, p. 1203-1211. An extended charge can accelerate without radiating, under specified conditions.

407. (a) Davis, {360}. It would appear that a sine wave across this resistor may generate a **rhythmic** variation in what is called "internal

energy"; i.e., stress energy inside the system, hence stress energy in the associated vacuum. Since this is analogous to a sound wave consisting of rarefaction and expansion, it is a method of generating a longitudinal EM wave. This may also be involved also in the Fogal semiconductor. For a discussion of internal forces, which occur in equal and opposite pairs and are quite different from the resultant of forces, see (b) Chen, {358}, p. 372 and passim.

408. See Hsu-Chieh Yeh, "Remark on the second law of thermodynamics," *Am. J. Phys.*, 52(8), Aug. 1984, p. 720-723. As an example, from the common charge or dipole at constant temperature, we can indeed receive and take electrical energy because the charge (as a set of composite dipoles) and the dipole are not in 3-space energy flow equilibrium. Instead, they constantly receive virtual EM energy from the active vacuum, transduce it into 3-space EM energy, and continuously emit **that** energy in all directions — all at constant temperature. When we consider the supersystem of the charge-system and its environment, it is in disequilibrium and hence exhibits the effect pointed out by Yeh.

409. Yeh, {408}. Particularly see the illustration on p. 721.

410. Gabriel Laufer, "Work and heat in the light of (thermal and **laser**) light," Am. *J. Phys.*, Vol. 51, 1983, p. 42.

411. (a) Paulo N. Correa and Alexandra N. Correa, "A Modified Orgone Accumulator (HYBORAC) as a Drive for a Low Delta-T **Stirling** Engine, Part 1," *Infinite Energy,* 7(41), 2002, p. 23-29; (b) — "Orgone Accumulator Drive (HYBORAC) for Low Delta Stirling Motors, video presentation, ABRI Media Production.

412. Paulo N. Correa and Alexandra N. Correa, *Introduction to Experimental Aetherometry,* Vols. 1 and 2, 2001. Summaries of these **two** volumes may be freely downloaded from http://www.aetherometry.com

413. Paulo N. Correa and Alexandra N. Correa, "A Modified Orgone Accumulator (HYBORAC) as a Drive for a Low Delta-T Stirling **Engine** Part II." *Infinite Energy,* 7(42), 2002 (in press)

414. Eugene F. Mallove, "Demonstrating Aether Energy," *Infinite Energy,* 7(41), 2002, p. 6-8.

415. As quoted by Mallove, *ibid.,* p. 7.

416. Feynman, {7}, Vol. II, p. 1-3. Emphasis in original.

417. Mallove, {414}, p. 7.

418. J. M. Aguirregabiria, A. Hernandez, and M. Rivas, "Surface charges and energy flow in a ring rotating in a magnetic field." *Am. .J. Phys.*, 64(7), July 1996, p. 892-895.

419. F. Herrmann, "The unexpected path of the energy in a moving capacitor," *Am. J. Phys.*, 61(2), Feb. 1993, p. 119-121.

420. Wang and Chung, {33}.

421. Mike LaPointe, "Antimatter Production at a Potential Boundary," presented at the AIAA Joint Propulsion Conference, Breakthrough Propulsion Physics Session #1, July 9, 2001.

422. E.g., (a) Charles Kittel, *Introduction to Solid State Physics,* Seventh Edition, Wiley, New York, 1996 in various chapters contains excellent introductions to a more in-depth view of magnetic phenomena. A very useful book, written with clarity is (b) B. D. Cullity, *Introduction to Magnetic Materials,* Addison-Wesley, Reading, MA, 1972. For nuclear effects, (c) A. Abraham, *Nuclear Magnetism,* Oxford. 1961 is still excellent and quite comprehensive.

423. (a) Particularly see the lay overview by Philip Ball, "Meet the spin doctors...", *Nature,* Vol. 404, 2000, p. 918-920. This article deals with spin-injection from magnetic semiconductors. In addition to up and down, electron spin also has an additional degree of freedom called "color"; see (b) "Colourful electrons solve puzzle of ferromagnetism," *Physics World,* 14(7), July 2001; (c) also see the accompanying article by Fu Chun Zhang.

424. Stanislaw Bednarek, "Self-powered magnetohydrodynamic motors," *Am. J. Phys.,* 64(1), Jan. 1996, p. 90-92.

425. In a true superconducting system, it is possible to keep the current moving around in a resistance-free conductor indefinitely. However, no net work can be produced by the system, and in fact when one counts the energy one inputs to maintain the extremely low temperature of the environment, then to maintain the supersystem in such **"indefinite current** circulation" one has to continuously input energy to cool the **system. On** the other hand, an electron in an atom, orbiting the nucleus, is in **net** equilibrium on its path, so it will maintain its orbit indefinitely until perturbed sufficiently to break loose.

426. (a) Badics, {377a};(b) P. Hammond and J. Penman, "Calculation of inductance and capacitance by means of dual energy principles," *IEE Proc,* 123(6), 1976, p. 554-559; (c) Hammond and Ssiboukis, {377d}; (d) Yongzhong Huo, "A classification of thermodynamical potentials for two-variable transition systems," *Meccanica* [Netherlands], 30(5), Oct. 1995, p. 475.494; (e). Kotiuga, {377e}; (f) J. N. Murrell and S. Carter, "Approximate single-valued representations of multivalued potential energy surfaces," J. *Phys. Chem.,* 88(21), Oct. 11, 1984, p. 4887-4891; (g) Penman and Fraser, {377f}; (h) Ren, {376a}; (i) Z. Ren, "Solving 3D static field problems by dual formulations using potential variables," ***2nd** Int. Workshop on Elec. and Magn. Fields,* Leuven, Belgium, May 1994; (j) Ren, {376b}; (k) Verite, {376c}; (1) V. Vittal, "A generalized procedure to obtain first integrals for non-conservative dynamical systems: Application to power systems," *IEEE ISCAS '89,* 1989, p. 1984-1987.

427. G. W. O. Howe, "Effect of Torsion on a Longitudinally-Magnetized Iron Wire," *Wireless Engineer,* 29(344), May 1952, p. **115**-117.

428. Harry E. Burke, *Handbook ofMagnetic Phenomena,* Van Nostrand, New York, 1985. This is a one-of-a-kind handbook on a great number of magnetic effects and phenomena.

429. Cullity, {422b}.

430. As examples: (a) Barrett, {345}; (b) Gurevich and Melkov, {232c}; (c) William Weltner, Jr., *Magnetic Atoms and Molecules,* Dover, Mineola, NY, 1989; (d) Giorgio Bertotti, *Hysteresis in Magnetism,* Academic Press, New York, 1998; (e) L'vov, {29c}; (f) Sidney Platt, *Magnetic Amplifiers: Theory and Application,* Prentice Hall, Englewood Cliffs, N.J., 1958; (g) Maciej J. Ogorzalek, *Chaos and Complexity in Nonlinear Electronic Circuits,* World Scientific, New Jersey, 1997; (h) Jammer, {lb}; (i) Borovik-Romanov and Sinha, Eds, {369b}; (j) Paul Davies, ed., {160f}; (k) G. I. Shipov, *A Theory ofPhysical Vacuum: A New Paradigm,* Moscow, 1994; (1) Nicolis and Prigogine, {255b}; (m) Richard Feynman, *Quantum Electrodynamics,* 1961, 1963; (n) V. 1. Fushchich and A. G. Nikitin, *Symmetries ofMaxwell's Equations,* D, Reidel, 1987; (o) Mandl and Shaw, {19}; (p) M. W. Evans, Ed., ***Modern** Nonlinear Optics,* Second Edition, 3 vols., Wiley, 2001; (q) Robert C. O'Handley, *Modern Magnetic Materials: Principles and Applications,* Wiley, New York, 2000; (r) Sachs, {126a}.

431. As examples: (a) Lindsay and Margenau, {114}; (b) Richard P. Feynman, *The Character of Physical Law,* MIT Press, Cambridge, 1965:; (c) Feynman, {7}; (d) Ilya Prigogine, *The End of Certainty: Time, Chaos, and the New Laws of Nature,* Free Press, New York, 1996, 1997; (e) Mendel Sachs, *Einstein Versus Bohr: The Continuing Controversies in Physics,* Open Court Publishing Company, 1988.

432. For those unfamiliar with magnetics and the exchange force(s), see H. J. Gray and Alan Isaacs, *A New Dictionary of Physics,* second edition, 1976 printing, *ibid.,* p. 194 for a precise definition. The force from the sudden exchange of charge is known as the *Heisenberg force.* That due to the exchange of spin direction is known as the *Bartlett force.* The force associated with exchange of both spin and charge is known as the *Majorama force.* Quoting: "These exchange forces may be interpreted as being due to the continual emission and absorption of virtual particles that carry the shared property from one particle to another. The exchange of virtual hadrons between interacting particles of the atomic nucleus is intimately involved in nuclear exchange forces.

433. Feynman, {7}, Vol. II, Chapter 37 covers magnetic materials including exchange forces, spins, and spin effects.

434. *The Encyclopedia of Physics,* Robert M. Besancon (editor), Second Edition, Van Nostrand Reinhold, New York, 1974, p. 518.

435. Cullity. {422b}. p. 132.

436. Kittel, {422a}, p. 443 (footnote 1).

437. Lindsay and Margenau, {114}, p. 216-217.

438. *Ibid.,* p. 213.

439. *Ibid.,* p. 215.

440. *Ibid.,* p. 214.

441. T. D. Lee, "Space Inversion, Time Reversal and Particle-Antiparticle Conjugation," *Physics Today,* 19(3), Mar. 1966, p. 23.

442. Lee, {73}, p. 187-188.

443. Romer, {84}.

444. Kenneth D. Moore, "Testing of a Johnson Magnetic Gate," *Explore,* 9(6), 2000, p. 70-75. Also available on www.cheniere.org.

445. (a) Raymond J. Radus and William G. Evans, "Apparatus Responsive to Direct Quantities," U.S. Patent #2,892,155, Jun. 23, 1959; (b) R. J. Radus, "Permanent-Magnet Circuit using a 'Flux-Transfer' principle," *Engineers' Digest,* 24(6), June 1963, p. 86; (c) (anon.), "On-Off Permanent Magnet with a Memory," *Electronic Equipment Engineering,* 11(4), Apr. 1963, p. 22, 26.

446. "Human fly has magnetic sole," *Electrical Engineering,* Apr. 1963, p. 294.

447. Malcolm McCaid, *Attraction and Repulsion,* Oliver and Boyd Ltd., 1967 has an entire chapter on switching magnets, and has two photos of an engineer hanging upside down and walking on the underside of an elevated steel beam, wearing the Radus magnetic boots.

448. For a description of "shuffler" boots, see A. W. Hangar and A. A. Rosener, "The use of permanent magnets in zero-gravity mobility and restraint 'footware' concept," *IEEE Trans, on Magnetics,* Vol. MAG-6, No. 3, Sep. 1970, p. 464-467. These crude shuffler boots are quite a contrast to the elegant Westinghouse/Radus stepping boots earlier developed and used.

449. Burke, {428}.

450. David Scott, "Magnetic "Wankel" for electric cars," *Popular Science,* June 1979, p. 90-91.

451. E.g., (a) John R. Wiegand, "Bistable Magnetic Device," U.S. Patent No. 3,820,090, June 25, 1974; (b) — "Asymmetric Bistable Magnetic Device," U.S. Patent No. 3,866,193, Feb. 11, 1975; (c) - "Pulse Generation by Changing Magnetic Field," U.S. Patent No. 4,309,628, Jan. 5, 1982; (d) — "Pulse Generator Using Read Head with Wiegand Wire," U.S. Patent No. 4,263,523, April 21, 1981; (e) - "Switchable Magnetic Device;" U.S. Patent No. 4,247,601, Jan. 27, 1981; (f) — "Method of Manufacturing Bistable Magnetic Device," U.S. Patent No. 3,892,118, July 1, 1975; (g) — "Multiple Pulse Generator;" U.S. Patent No. 3,793,532, Feb. 19, 1974; (h) — and John E. Opie, "Pulse Generator with Shaped Magnetic Field," U.S. Patent No. 4,484,090. Nov 20, 1984.

452. Kawai, {31}.

453. John C. Bedini, "Device and Method of a Back EMF Permanent Electromagnetic Motor Generator," U.S. Patent No. 6,392,370, May 21.

2002. Uses a regauging process for capturing available back **EMF** electromagnetic energy in the system. The available energy may be captured and used to re-energize the battery, and/or sent in another direction to be used as useful work, or dissipated in the system.

454. See {445c}, p. 22, 26. Quoting: *"The circuit differs from ordinary permanent magnets in three ways: (I) It exerts strong magnetic force at one end while exerting hardly any at the other, (2) the strong and weak magnetic poles can be switched end-for-end easily and at will, thus in effect turning the magnetism on and off, and (3) it remembers its direction of greatest magnetic pull indefinitely. "*

455. Burke, {428}, p. 250.

456. Deffeyes, {342}.

457. Stuart Young, *Nature*, Vol. 414, 29 Nov. 2001, p. 487.

458. Evans, {137b}, p. 84.

459. M. W. Evans and S. Jeffers, "The Present Status of the Quantum Theory of Light," in *Modern Nonlinear Optics,* Second Edition, 3 Vols., edited by M. W. Evans, Wiley, 2001, Vol. 3, p. 150.

460. Clarke, {55}, p. 38.

461. Patrick, Bearden, Hayes, Moore, and Kenny, {37}. A second patent is pending, and two more are in preparation.

462. Bearden, {222b}.

463. Evans etal, {38b, 38c}.

464. T. E. Bearden, James C. Hayes, James L. Kenny, Kenneth D. Moore, and Stephen L. Patrick, "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy-Replenishing from the Active Vacuum," Sept. 2000, carried on restricted DoE website http://www.ott.doe.gov/electromagnetic/. A search on the internet will turn up other sources where the paper can be downloaded.

465. (a) Evans et al., "The Aharonov-Bohm Effect as the Basis of Electromagnetic Energy Inherent in the Vacuum," *Found. Phys. Lett.,* 2002, (in review); (b) — "Derivation of the B(3) Field and Concomitant Vacuum Energy Density from the Sachs Theory of Electrodynamics," ***Found.*** *Phys. Lett.,* 14(6), Dec. 2001, p. 589-593; (c) — "Operator Derivation of the **Gauge** Invariant **Proca** and Lehnert Equations:

Elimination of the Lorenz Condition," *Found. Phys.,* 30(7), 2000, p. 1123-1130; (d) — "Runaway Solutions of the Lehnert Equations: The Possibility of Extracting Energy from the Vacuum," *Optik,* 111(9), 2000, p. 407-409; (e) — "The Effect of Vacuum Energy on the Atomic Spectra," *Found. Phys. Lett,* 13(3), June 2000, p. 289-296; (f) Evans et al., {113}; (g) — "Derivation of the Lehnert Field Equations for Gauge Theory in Vacuum: Space Charge and Current," *Found. Phys. Lett.,* 13(2), Apr. 2000, p. 179-184;(h)Bearden, {222b}.

466. Details are on Jean-Louis Naudin's website, in the section dealing with the MEG, which is http://jnaudin.free.fr/html/megv21.htm#megstat.

467. For an introduction to the field and its literature, see J. Anandan, J. Christian, and K. Wanelik, "Resource letter GPP-1: Geometric phases in physics," *Am. J. Phys.,* Vol. 65, 1977, p. 180-185.

468. A discussion by Feynman of the Aharonov-Bohm effect is given in Feynman, {7}, "15-5: The vector potential and quantum mechanics," p. 15-8 to 15-14.

469. A similar phase change also is induced outside the localized presence of an electrostatic scalar potential fi. See Feynman, {7}, p. 15-10. Quoting: *"...the effect of an electrostatic field is to produce a phase change given by the negative of the time integral of the scalar potential fi"*

470. Ya. M. Blanter, C. Bruder, R. Fazio, and H. Schoeller, "Aharonov Bohm-type Oscillations of Thermopower in a Quantum-dot Ring," *Phys Rev. B,* Vol. 55, 1997, p. 4069 typify the kind of "power" considerations that have been analyzed in the past. Only a minuscule thermopower treatment is considered.

471. S. Pancharatnam, "Generalized theory of interference, and its applications," *Proc. Indian Ac ad. Sci.,* Vol. A44, 1956, p. 247-262.

472. See Nahin, {62}, p. 225. Quoting: *"J. J. Waterston's paper on the kinetic theory of gases, in 1845, was rejected by the Royal Society of London. One of the referees declared it to be 'nothing but nonsense,* ***unfit*** *even for reading before the Society.'... "Waterston's dusty manuscript was finally exhumed from its archival tomb forty years later, because of the efforts of Lord Rayleigh... "*

473. Aharonov and Bohm, {344a, 344b}.

474. R. G. Chambers, "Shift of an electron interference **pattern** by enclosed magnetic flux," *Phys. Rev. Lett.,* Vol. 5, July I960, p. 3-5.

475. A. Tonomura *et al,* "Observation of Aharonov-Bohm effect by electron holography," *Phys. Rev. Lett.,* Vol. 48, 1982, p. 1443-1446.

476. A. Tonomura *et al.,* "Evidence for Aharonov-Bohm effect with magnetic field completely shielded from electron wave," *Phys. Rev. Lett.,* Vol. 56, 1986, p. 792-795. In the latter paper, the magnetic fields were completely and painstakingly shielded from the electron wave in a very careful experiment. The experiment demonstrated the expected relative phase shift as well as direct evidence for flux quantization. This was a definitive experiment clearly proving the Aharonov-Bohm effect. In 1999, Tonomura was awarded the Franklin Institute Award for this definitive proof of the Aharonov-Bohm effect.

477. M. V. Berry, *Proc. R. Soc. Lond. A,* Vol. 392, 1984, p. 45.

478. R. A. Webb, S. Washburn, C. P. Umbach, and R. B. Laibowitz, "Observation of h/e Aharonov-Bohm oscillations in normal-metal rings," *Phys. Rev. Lett,* Vol. 54, 1985, p. 2696-2695.

479. B. Schwarzschild, "Currents in Normal-Metal Rings Exhibit Aharonov-Bohm Effect," *Phys. Today,* Vol. 39, Jan. 1986, p. 17-20.

480. Y. Aharonov and J. Anandan, "Phase change during a cyclic quantum evolution," *Phys. Rev. Lett.,* Vol. 58, 1987, p. 1593-1596.

481. (a) V. Chandrasekhar *et al., Phys. Rev. Lett.,* Vol. 55, 1985, p. 1610; (b) S. Datta *et al., Phys. Rev. Lett,* Vol. 55, 1985, p. 2344; (c) S. N. M. Ruijsenaars, "The Aharonov-Bohm Effect and Scattering Theory," *Ann. Phys.,* Vol. 146, 1983, p. 1-34.

482. S. Olariu and I. Iovitzu Popescu, "The Quantum Effects of Electromagnetic Fluxes," *Rev. Mod. Phys.,* 57(2), Apr. 1985, p. 339-436.

483. (a) R. Bhandari, "Polarization of light and topological phases," *Phys. Reps.,* Vol. 281, 1997, p. 1-64; (b) — and J. Samuel, "Observation of lopological phase by use of a laser interferometer," *Phys. Rev. Lett.,* Vol. 60, 1988, p. 1211-1213; (c) R. Y. Chiao et *al.,* "Observation of a lopological phase by means of a Mach-Zehnder interferometer," *Phys. Rev. Lett,* Vol. 60, 1988, p. 1214-1217; (d) T. Bitter and D. Dubbers, "Manifestation of Berry topological phase in neutron spin rotation," *Phys. Rev. Lett,* Vol. 59, 1987, p. 251-254; (e) Y. Hascgawa *et al.,* "Geometric

phase in coupled neutron interference loops," *Phys. Rev. A.,* Vol. 53, 1996, p. 2486-2492; (f) A. G. Wagh *et al.,* "Experimental separation of geometric and dynamical phases using neutron interferometry," *Phys. Rev. Lett.,* Vol. 78, 1997, p. 755-759; (g) D. Suter, K. T. Mueller, and A. Pines, "Study of the Aharonov-Anandan quantum phase by NMR interferometry," *Phys. Rev. Lett.,* Vol. 60, 1988, p. 1218-1220; (h) D. Loss and P. M. Goldbart, "Persistent currents from Berry's phase in Mesoscopic systems," *Phys. Rev. B,* Vol. 45, 1992, p. 13544-13561; (i) C. A. Mead, "The geometric phase in molecular systems," *Rev. Mod. Phys.,* Vol. 64, 1992, p. 51-85; (j) J. H. Hannay, "Angle variable holonomy in adiabatic excursion of an integrable Hamiltonian," *J. Phys. A.,* Vol. 18, 1985, p. 221-230; (k) Y. Aharonov et al., "Aharonov-Bohm Type Forces Between Magnetic Fluxons," Hewlett Packard Report HPL-BR1MS-97-24, Oct. 1997.

484. (a) F. V. Gubarev and V. I. Zakharov, "The Berry Phase and Monopoles in Non-Abelian Gauge Theories," *Internat. J. Mod. Phys. A,* Vol. 17, 2002, p. 157-174; (b) J. M. Robbins, "Topological Phase Effects," Hewlett-Packard Labs Technical Report HPL-BRIMS-97-03, Jan. 31, 1997; (c) E. Sjoqvist, "Anandan-Aharonov-Casher oscillations in a simply connected region, *Phys. Lett. A.,* Vol. 270, 2000, p. 10; (d) E. Sjoqvist et al., "Geometric phases for mixed states in interferometry," *Phys. Rev. Lett.,* Vol. 85, 2000, p. 2845; (e) D. J. Thouless, *Topological Quantum Numbers in Nonrelativistic Physics,* World Scientific, Singapore, 1998; (f) J. Mickelsson, "Vacuum Polarization and the Geometric Phase: Gauge Invariance," *J. Math. Phys.,* Vol. 39, 1998, p. 831-837; (g) W. T. Lee *at al.,* "Scalar Aharonov-Bohm Effect with Longitudinally Polarized Neutrons," *Physica B,* Vol. 241-243, 1998, p. 216; (h) A. Cimmino et al., "High Precision Measurement of the Topological Aharonov-Casher Effect with Neutrons," *Nuclear Instr. and Methods A,* Vol. 440, 2000, p. 579; (i) Ali Mostafazadeh, *Dynamical Invariants, Adiabatic Approximation and the Geometric Phase,* Nova Science, 2001.

485. E.g., see (a) M. W. Evans and L. B. Crowell, *Classical and Quantum Electrodynamics and the B(3) Field,* World Scientific, Singapore, 2001; (b) M. W. Evans, *0(3) Electrodynamics,* Vol. V *of The Enigmatic Photon,* Kluwer, Dordrecht, 1999.

486. Bearden, {12}.

487. Fradkov and Pogromsky, {364}. The quote is from p. 359-360.

488. (a) Bearden, {38a}, (b) Bearden *et al., {*464}, (c) Evans et al., {38b}, (d) Evans et al., {38c},

489. Seife, {184}.

490. I. J. R. Aitchison, "Nothing's Plenty: The Vacuum in Modern Quantum Field Theory," *Contemp. Phys.,* 26(4), 1985, p. 339, 360.

491. *Ibid.,* p. 372.

492. Robert Neill Boyd, "Black Holes and Negative Energy", carried on Internet website http://www.rialian.com/rnboyd'blackholcs-negative.htm.

493. Roger Penrose, *The Emperor's New Mind.* Penguin Books, 1991, p. 210.

494. Stephen Hawking, per Zebrowski, {2}.

495. Isaac Newton, *Philosophiae Naturalis Principia Mathematica,* 1687.

496. Quoted from D. W. G. Ballentine and D. R. Lovett, *A Dictionary ofNamed Effects and Laws in Chemistry, Physics, and Mathematics,* Fourth Edition, Chapman and Hall, New York, 1980. p. 227.

497. Dirac, {249}.

498. See also (a) P.A.M. Dirac. "The quantun theory of the electron," P*roc. Roy. Soc. Lond. A,* 117(777), Jan. 2, 1928. p. 610-624; (b) — "The quantum theory of the electron, Part II," *ibid.,* 118 (779), Mar. 1, 1928, p. 351-361.

499. Georg Friedrich Berhnard Riemann, "Foundat ions of a general theory of functions of a single complex variabable." Doctoral Thesis, 1851. In H. Weber, ed., *B. Riemann: Gesammelte Mathematische Werke,* 2nd Ed., paperback reprint, Dover, New York, 1953.

500. As pointed out by Charles W. Misner, Kip S Thorne and John Archibald Wheeler, *Gravitation,* W. H. Freeman and Company, San Francisco, 1973, p. 220.

501. Josephs, {172}.

502. Laithwaite, {172b}.

503. Ernst Mach, *The Science ofMechanics.* 1912, Chapter 2, section 6.

504. This statement of Mach's principle is quoted from Ballentine and Lovett, {496}, p. 201.

505. Albert Einstein, "Die Grundlage der Allgemeinen Relativitatstheorie," ["The Foundation of the General Theory of Relativity,"] *Annalen der Physik,* Vol. 49, 1916, p. 769-822. (English translation in *The Principle ofRelativity,* Methuen, NY, 1923.).

506. Misner, Thome and Wheeler, {500}, p. 543.

507. A. Einstein, Appendix V to his *Relativity, The Special and the General Theory,* translated by Robert W. Lawson, Crown Publishers, New York, 1961, p. 155.

508. Theodor Kaluza, *Sitz. Peuss. Akad. Wiss.,* Vol. K1, 1921, p. 966. English translation, "On the Unification Problem of Physics," is in Lochlainn O''Raifeartaigh, *The Dawning of Gauge Theory,* Princeton University Press, Princeton, 1997, p. 53-58. Einstein had held Kaluza's paper since 1919 before recommending it for publication by the journal.

509. {a} Oskar Klein, *Zeit. f. Physik.* Vol. 37, 1926, p. 895. An English translation, "Quantum Theory and Five-Dimensional Relativity," is in Lochlainn O''Raifeartaigh, *The Dawning of Gauge Theory,* Princeton University Press, Princeton, 1997, p. 59-69. See also {b} Oskar Klein, *Nature, Vol.* 118, 1926, p. 319.

510. Sachs, {126a}.

511. Sachs, *ibid.,* quoted from the publisher's summation on the book jacket itself.

512. Mendel Sachs, "On Unification of Gravity and Electromagnetism and the Absence of Magnetic Monopoles," *IL Nuovo Cimento* B, Vol. **114,** 1999, p. 123.

513. (a) Evans et al., {233f}; (b) Evans et al., {233b}.

514. Misner, Thorne, and Wheeler, {500}.

515. Albert Einstein, *Relativity: the Special and General Theory,* translated by Robert W. Lawson, Crown Publishers, New York, 1961, fifteenth edition.

516. A. Einstein, H. A. Lorentz, H. Minkowski and H. Weyl, with **notes** by A. Sommerfeld, translations by W. Perrett and G. B. Jeffery, *The*

Principle ofRelativity, Methuen and Company. Ltd., 1923. Also available in paperback from Dover.

517. (a) T. Townsend Brown, "A Method of and an Apparatus or Machine for producing Force or Motion, British Patent 300,311, Nov. 15, 1928; (b) — "Electrokinetic Apparatus," U.S. Patent # 2,949,550, Aug. 16, 1960; (c) — "Electrokinetic Transducer," U.S. Patent # 3,018,394, Jan. 23, 1962; (d) — "Electrokinetic Generator," U.S. Patent # 3,022,430, Feb. 20, 1962; (e) — "Electrokinetic Apparatus," U.S. Patent # 3,187,206, Jun. 1, 1965.

518. Jonathan W. Campbell, "Apparatus and Method for Generating Thrust Using a Two Dimensional, Asymmetrical Capacitor Module," U. S. Patent #6,317,310, Nov. 13,2001.

519. (a) Norman L. Dean, "System for Converting Rotary Motion into Unidirectional Motion," U.S. Patent # 2,886,976, May 19, 1959. See also (b) John W. Campbell, Jr., "Report on the Dean Drive," *Analog,* Sept. 1960; (c) — "Instrumentation for the Dean Drive," *Analog,* Nov. 1960; (d) — "Final Report on the Dean Drive," Analog, Dec. 1960. Also (e) G. Harry Stine, "Detesters, Phasers, and Dean Drives," *Analog,* Jun. 1976.

520. (a) Henry Wallace, "Method and Apparatus for Generating a Secondary Gravitational Force Field," U.S. Patent # 3,626,605, Dec. 14, 1971; (b) — "Method and Apparatus for Generating a Dynamic Force," U.S. Patent # 3.626.606. Dec. 14, 1971.

521. J. F. Woodward, "Method for Transiently Altering the Mass of Objects to Facilitate their Transport or Change their Stationary Apparent Weights", U.S. Patent #5,280,864, Jan. 25, 1994.

522. (a) J. F. Woodward, "A New Experimental Approach to Mach's Principle and Relativistic Gravitation," *Found Phys. Lett.,* 3(5), 1990, p. 497-506; (b) — "Corrigendum: A New Experimental Approach to Mach's Principle and Relativistic Gravitation," *Found. Phys. Lett.,* 4(3), 1991, p. 299;(c) — "Measurements of a Machian Transient Mass Fluctuation,^{1}' *Found. Phys. Lett.,* 4(5), 1991, p. 407-423; (d) -- "A Stationary Apparent Weight Shift from a Transient Machian Mass Fluctuation," *Found. Phys. Lett.,* 5(5), 1992, p. 425-442; (e) — Maximal Acceleration, Mach's Principle, and the Mass of the Electron," *Found. Phys. Lett.,* 6(3), 1993, p. 233-244; (f) — "Making the Universe Safe for Historians: Time Travel and the Laws of Physics," *Found. Phys Lett.,* 8(1). 1995, p. 1-39; (g) - "Erratum," *Found. Phys Lett.,* 8(3), 1995, p. 303; (h) - "Killing Time."

Found. Phys. Lett., 9(1), 1996, p. 1-23; (i) — "Laboratory Test of Mach's Principle and Strong-Field Relativistic Gravity," *Found. Phys. Lett.,* 9(3), 1996, p. 247-293; (j) — "Twists of Fate: Can We Make Traversable Wormholes in Spacetime?", *Found. Phys. Lett.,* 10(2), 1997, p. 153-181; (k) — Mach's Principle and Impulse Engines: toward a Viable Physics of Star Trek?", presented at the NASA Breakthrough Propulsion Physics Workshop, Cleveland, Ohio, Aug. 12-14, 1997.

523. (a) Max Jammer, *Concepts ofMass, in Classical and Modern Physics,* Harvard University Press, Cambridge, MA, 1961. One also suggests the reading of Jammer, {1b}.

524. (a) Bearden, {15}, p. 11. One kilogram of mass represents 17.053 x 10^{50} switches per second of action quanta h/(4*PI) across the boundary of the mass, in a flux exchange between the mass region and its external environment. The acceleration force can be derived from the difference in "inswitching" and "outswitching" of action quanta (as the masses m_i of the system iteratively change between mass and masstime states). The paper, albeit crudely, derives Newton's three laws of motion in relativistic form and the square law of gravitation. It also indicates the mechanism (reaction) upon a mass that generates the mass's rate of progression through time. For a somewhat expanded treatment, see (b) T. E. Bearden, "Vacuum Engines and Priore's Methodology: The True Science of Energy Medicine," Parts I and II, *Explore,* 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62.

525. Today we would also point out that to first order time and mass have essentially the same energy density — i.e., they are both spatial energy compressed by c^2. So the "switches" of action quanta we used in 1973 may be taken to actually represent the time-charging and discharging, of a specific region of space. Rotation of time-energy from the time-domain into 3-space connection with a mass is a mass increase as well as a change of mass to masstime. Rotation of mass-energy out of 3-space back into the time domain (during photon emission) is a mass decrease, (a) Our paper, "Giant Negentropy..." {12}, proposing a solution to the source charge problem, argues that all 3-space EM energy comes from the time domain anyway and returns to it. The source charges normally absorb time-energy, and transduce it to 3-space EM energy or vice versa, but the 3-space energy radiated away is in expanded spatial energy form and not compressed into mass-energy. It is the very release of the c^2 compression of the time-energy that is the "transduction" process to provide the **radiated**

spatial EM energy. But if mere rotation in 4-space is used without **that** expansion/transduction, then the time energy appears as mass. This is consistent with the mechanism we advanced for the flow of a mass *m* through time. Note that when a photon is "absorbed", the spatial energy component of the photon is in fact compressed into mass energy, slightly changing the mass. But when it is re-radiated, it is radiated in expanded or "spatial" EM energy form. Without the expansion and transduction, pure rotation of the compressed EM energy between the time domain and the 3-space domain provides the notion of "time-charging" and "time discharging". Indeed, the very anomalous phenomena occurring in radiation detectors in electrolyte experiments at China Lake, already provide direct evidence for just such time-charging, as pointed out in (b) Bcardcn, {17}.

526. E. R. Laithwaite, "Rack and pinion motors," *Electronics and Power,* Vol. 16, 1970, p. 251-252.

527. Of interest: (a) E. R. Laithwaite, "Unexplained phenomenon at Savoy Place," *Electronics and Power,* Oct. 1972, p. 360-362 recounts the discovery of a free energy device violating the second law of equilibrium thermodynamics; (b) Laithwaite, {526}, gives some particulars of the device; (c) —*All Things Are Possible: An Engineer Looks at Research and Development,* I*PC Electrical-Electronic Press Ltd., Dorset House, London, 1976; (d) Laithwaite, {172b}; (e) — "Miles Walker — a pioneer at Met-Vick and UMIST," *Electrical Review,* 211(12), Oct. 15, 1982, p. 33-34; (f) — "Roll, Isaac, roll — Part I," *Electrical Review,* 204(7), Feb. 16, 1979, p. 38-41; (g) — "Roll, Isaac, roll — Part II," *Electrical Review,* 204(11), Mar. 16, 1979, p. 31-33; (h) — "The laws of science and of England," *Electrical Review,* 202(23), June 16, 1978, p. 27-28.

528. E. R. Laithwaite and W. R. C. Dawson, "Mass Transfer: An Introduction based on Experimental Work at the University of Sussex," Sep. 1994. A subtitle is: A system for the transfer of mass derived from the principle of conservation of momentum. On website http://jnaudin.free.fr/html/IPEmain.htm.

529. See WO95/30832, Nov. 16, 1995.

530. Eric Laithwaite and William Dawson, "Propulsion System," U.S. Patent Number 5,860,317, Jan. 19, 1999.

531. Sweet and **Bearden, {**29a**}.**

532. (a) Ning Li and D. G. Torr, *Phys. Rev.*, Vol. 43D, 1991, p. 457; (b) — *Phys. Rev.*, vol. 46B, 1992, p. 5489; (c) — *Bull. Am. Phys. Soc*, Vol. 37, 1992, p. 948. Also see (d) D. G. Torr and Ning Li, *Found. Phys. Lett.*, Vol.37, 1993, p. 948.

533. E. E. Podkletnov and R. Nieminen, *Physica C*, Vol. 203, 1992, p. 441-444.

534. Igor K. Kulikov, "Inertial and gravitational masses of bosons at finite temperatures," in NASA technical package sent in response to queries of "Study of Inertial and Gravitational Masses of a Boson," *NASA Tech Briefs*, 26(2), Feb. 2002, p. 56. The report by Dr. Kulikov of the Jet Propulsion Laboratory (JPL) is carried on the website as http://nasatech.com/TSP/PDFTSP/NPO30325.pdl'. This is a theoretical study of the relationship between the inertial mass and gravitational mass of a self-interacting neutral scalar boson in a heat bath. The difference in the two masses was determined as a function of the temperature of the heal bath. The theory for distinguishing the difference is given. Since it is presented as an inventor's report, we assume that a patent is either pending by or to JPL or already granted to it.

535. Mike LaPointe, {421}.

536. Jim Corum, T. Keech, D. Gray, P. Pesavento, and M. Duncan, "The Electro-magnetic Stress-Tensor as a possible Space Drive Propulsion Concept, presented at the AIAA Joint Propulsion Conference, July 9, 2001.

537. Hector Serrano, "Propulsion Device and Method Employing Electric Fields for Producing Thrust," WO 00/58623, Oct. 5, 2000.

538. William B. Stein, Electrokinetic Propulsion: The Ionic Wind Argument," Energy Conversion Laboratory, Purdue University, West Lafayette, Indiana 47906, Sep. 5, 2000. This is a Gravitec report. Results can be downloaded from http://foldedspace.com/data.html, and an even more rigorous future test is planned.

539. A motion picture of the Naudin's replica thruster in dynamic thrust performance is shown on http://jnaudin.free.fr/html/liftbldr.htm, along with information on many other successful "lifters". This website is a cornucopia of information on inertial propulsion and electrogravitation, including successful independent replication and tests.

540. Ryan Mahoney, "Huntsville Firm Developing New Propulsion Technology," Birmingham Business Journal, Birmingham, Alabama July 19,2001.

541. E.g., see J. Cameron, "An Asymmetric Gravitational Wave Propulsion System," presented in the Breakthrough Propulsion Physics session of the AIAA Joint Propulsion Conference, July 9, 2001.

542. Bearden, {40}.

543. Prange and Strance, {365

544. Privately, Sweet strongly affirmed to me that later he took the risk and pushed it a little further in output, actually levitating the system several times, including accelerating it upward, hovering it, and then letting it sink slowly down to come to rest on the bench.

545. Dirac, {249}.

546. Romer, {84}; see endnote 24, p. 109.

547. E.g., as by Giovanni Modanese, "On the theoretical interpretation of E. Podkletnov's experiment," Extract from report UTF-391/96, OANL gr-qc/9612022, presented to the *I.A.F. Congress 1997,* Number IAA-97-4.107.

548. Ling Ni and D. G. Torr, {532a-532d}.

549. Evans, {137b}; Evans, {485b}.

550. M. W. Evans, personal correspondence.

551. (a) Evans et al., {233b}; (b) — - "Development of the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* 14(6), Dec. 2001, p. 595-600; (c) Evans et al., {233a}.

552. As we point out in our discussion of cold fusion, use of this enormous time-energy density component in a time-reversal zone allows reducing the force of the gluons in a nucleon and nearly freeing the quarks. The time-reversal zone also reverses the law of attraction and repulsion of charges; now like charges attract. So the coulombic attraction of two H+ ions can attract them together so closely that each enters the strong force region of the other, forming a quasi-nucleus. As the other ions in solution move to negate the time-reversal zone, the zone decays, restoring the normal repulsion of the two protons in the quasi-nucleus. However, being exponential the strong force increases much more rapidly than the normal

coulombic repulsion grows. The result is to force the two protons even closer together, into a full nucleus with a full strong force between them. This forms a nucleus of deuterium, as one quark in one of the protons is flipped while still relatively free, to form a neutron. Different but similar reactions exist to produce the excess tritium shown in many cold fusion experiments, and a similar interaction — especially between two deuteron ions — forms an alpha particle.

553. Bearden, {17}.

554. Davies, {237}, p. 105.

555. Misner, Thorne, and Wheeler, {500}, p. 5.

556. Evans and Jeffers, {459}, p. 150.

557. Evans, {137a}, p. 477.

558. Banesh Hoffman, in Foreword to Yakov P. Terletski, *Paradoxes in the Theory ofRelativity*, translated from the Russian, Plenum Press, New York, 1968, p. vi, viii.

559. Kron, {162}, p. 466.

560. T. D. Lee, {239}, p. 46-47.]

561. This is well known to modern theorists, but is especially accented by Sachs and by Evans. E.g.. to even have an EM wave in space, the energy density of space has to be changing as the wave energy oscillates. This is an oscillating curvature of spacetime, a priori. So all EM waves move in a dynamically curving spacetime.

562. Sachs, {126b}.

563. These concepts are generally consistent with 0(3) electrodynamics. For rigorous exposition of 0(3) EM, see Evans, {137b [.

564. Felix Finster, "Definition of the Dirac Sea in the Presence of External Fields," *Adv. Theor. Math. Phys.,* Vol. 2, 1998, p. 963-985.

565. J. W. Lynn and R. A. Russell, "Kron's Wave Automaton," Journal unk, date unk, p. 131. We would appreciate any researcher furnishing us a completed citation for this reference. We have a copy of the paper, but without full citation details.

566. Kron, {162}. The quote is from p. 114.

567. Kron, {32}, p. 39.

568. Kron, {306}.

569. Kron, {32}, p. 41.

570. See Stigant, {310a} for a bibliography of Kron's **publications.**

571. Several overunity researchers have been assassinated by forces and parties unknown. Many very clever means are used to do the dirty work.

As an example, there is a standard EMF "shooter" which uses Venus modification of the wavefront, to produce a wave that instantly disrupts the control of the heart's beating. There are two versions at least; one about the size of a pocketbook and for close range (30 feet) only, and one about the size of a bazooka and with longer range (up to 100 yards). About 1 minute exposure to the beam of the small weapon will result in death of the targeted individual. I have personally been hit by the small version of that device (and I have a witness who was with me at the time, and also was exposed to it). We survived by instant recognition of what it was, and by bolting immediately out the nearby back door of a restaurant here in Huntsville. If the kill is successful, the autopsy will show that the person died of a heart attack.

Stan Meyer, an overunity researcher with numerous patents, rushed from a restaurant exclaiming "They're killing me!" and expired with just such a heart attack, possibly hit inside the restaurant by a close-range shooter.

Harry Mason's associate was killed in Australia in his apartment by a longer-range, bazooka-sized shooter. The assassin was spotted putting the shooter back inside the car below, and speeding away.

Sometimes a silenced rifle is used for sniping at long range. Sweet was fired at by one such would-be assassin — luckily just as he stumbled and fell on the steps of the foyer to his apartment complex. The bullet cracked right by his ear as he fell forward, passing where his head had just been the moment before.

Another "shooter" uses longitudinal EM wave peak power pulses. This type kill is instant and very permanent. This unit can be set to stun or render the targeted person unconscious, or set to kill. It can also disrupt and stop running automobile engines (there is another version for that also, using negative energy EMP weaponry). However, it time-charges the struck human body so that the body emits longitudinal EM waves for some time after the kill. Marinov was killed with such a shooter (apparently by

KGB-related foreign intelligence operatives) and his body then thrown off a rooftop to make it appear to be suicide. Where his body lay, the pavement glowed when the body was removed. Only one weapon on earth will kill a body so that it will induce luminescence in concrete or asphalt that it lies on for a period.

Another favorite weapon is an ice dart dipped in curare. I was threatened once and told I would be killed by such an ice dart weapon if I traveled in a 2-month period. Just four days past that period, and traveling and changing planes in Dallas-Fort Worth airport, a person was killed three feet from me in broad daylight, right inside the airport, with the exact symptoms of a curare ice dart kill. That was apparently just to teach me "they" were serious.

"They" is that set of cartels that Churchill called "the High Cabal". Bedini and I had severely offended the High Cabal by Bedini's successful transmutation of copper (and other things) into gold. John often makes his own power transistors for his amplifiers, and the slag in the electric oven turns into black ruby laced with gold and silver, confirmed by assay. We have had numerous other assassination attempts, too numerous to iterate. Suffice it to say that developing a good COP>1.0 EM power system can be hazardous to one's health.

Any significant researcher should be wary of "meeting with a sudden suicide" on the way to the supermarket. Another thing to beware of, is a calibrated auto accident where your car is rammed from the rear, and you are shaken up considerably. An ambulance just happens to be passing by moments later, and it will take you to the hospital. If still conscious, the researcher must not get in the ambulance unless accompanied by a watchful friend who understands the situation and the danger. Otherwise, he can easily get a syringe of air into his veins, which will effectively turn him into a human vegetable. If he goes to the hospital safely, he must be guarded by friends day and night, for the same reason, else he runs a high risk of the "air syringe" assassination during the night.

Simply trying to do scientific work, I find it necessary to often carry (legally) a hidden weapon. Both my wife and I have gun permits, and we frequently and legally carry concealed weapons.

As early as the 1930s, T. Henry Moray — who built a successful COP>1.0 power system outputting 50 kW from a 55 lb power unit - had to ride in a bulletproof car in Salt Lake City, Utah. He was repeatedly fired at by snipers from the buildings or sidewalk, with the bullets sometimes sticking in the glass. He was also shot by a would-be assassin in his own laboratory, but overpowered his assassin and recovered.

Over the years, probably as many as 50 or more overunity researchers and inventors have been assassinated for their efforts, and particularly For their success. Some have simply disappeared abruptly and never been heard from again.

These days the use of longitudinal wave shooters (such as killed Marinov) and such is growing, since no signature is left for a normal autopsy to discover. The Yakuza, e.g., uses such "shooters" in Japan to assassinate persons or politicians having incurred their displeasure. Others are subjected to the more traditional Yakuza assassination: Three or four assassins suddenly jump the unsuspecting target, stabbing him repeated and with extreme rapidity. His body may have 50 to 70 deep stab wounds — through the heart, the abdomen, liver, lungs, etc. — before the body strikes the pavement.

Other very sophisticated methods of legal entrapment, gaming, planting narcotics on one's premises and tipping off the drug enforcement agencies for a drug raid in the middle of the night, etc. are utilized.

Widely used is "gaming"', which uses deep psychological profiling to select a "vulnerability". The supercomputer with the game prepares a "scenario" much like writing a movie script of predictable interactions among the players to be chosen. The game involves "agents of influence" (AOIs — persons who are dogmatic, perhaps even radical, and have knee-jerk responses easily triggered, and whose vector direction when they "jump" is along or near the desired response direction), and one or two agents to do the phone conversations with the AOIs and trigger them into action), as well as a game controller or director. The computer programs also use prediction and artificial intelligence, and will forecast the probability of success of the selected game. A very extensive database of deep psychological profiles on useful AOIs and all major COP>1.0 researchers is kept in the computer. When the controller selects the game, the computer spits out the most recommended AOI players. The controller accepts or changes and then accepts, and the game scenario is then prepared. The agents trigger the AOIs, and the unwitting target is suddenly besieged from every direction, along a direction he already has an affinity or "vulnerability" towards. The controller follows the course of the game and the outcomes of all interactions between the target and the AOIs. The agent or agents add "nudges" or additional actions to overcome delays, accomplish redirection when the direction goes awry. etc. Over the decades, these games have gotten very effective indeed, and are quite difficult to content with. I laving survived about 200 of these games over the last two decades, I may yet write a book on some of the major ones that have been used against overunity researchers. The games are not confined just to the overunity field, however, and sometimes are also used to directly intervene in politics and other key situations, {754}. Machiavelli's methods are not dead; they thrive and have been highly computerized.

572. More than a dozen potential mechanisms for overunity EM interactions and systems are given in Evans et al., {233i}. Many of these interactions are likely in highly energetic astronomical phenomena such as the gamma ray burster and the X-ray burster.

573. Evans et al., {38b, 38c}.

574. Bearden, {38a}.

575. Bearden, {30}.

576. Sweet and Bearden, {29a}.

577. (a) A. G. Reiss *et al., Astron. J.,* Vol. 116, 1998, p. 1009; (b) S. Perlmutter *et al., Astrophys. J.,* Vol. 517, 1999, p. 565.

578. A. Einstein, *Sitzungsber. Preuss. Akad. Wiss.,* Vol. 1917, 1917, p. 142. Einstein came to regret his introduction of the cosmological constani as his "greatest blunder". Ironically, this may yet turn out to be one of his greatest contributions.

579. Prange and Strance, {365}.

580. (a) R. Lieu, "The effect of Planck scale space time fluctuations on Lorentz invariance at extreme speeds," *Astrophys. J. Lett.,* 2002 (in press). A preprint is on http://xxx.lanl.gov/abs/astro-ph/0202443. See also (b) Philip Ball, "Time gives rays a break," http://www.nature.com/nsu/020304/0209304-5.html .

581. In 1998, experimental evidence of the accelerated expansion of **the** universe was provided by two groups of astronomers; see (a) Riess *et al.,* (577a}; Perlmutter *et al.,* {577b}. For an attempt to provide a new model for the negative antigravity, see (c) L. Parker and A. Raval, "A new look ill the accelerating universe," *Phys. Rev. Lett.,* Vol. 86, 2001, p. 749-752; (d) Philip Ball, "New model of expanding universe," *Nature News Service,* Science Update, Nov. 12, 2001. A good coverage for the educated layperson is (e) Mario Livio and Alan Sandage, *The Accelerating Universe: Infinite Expansion, the Cosmological Constant, and the Beauty of the Cosmos,*[1]' Wiley, New York, Dec. 2000. See also (f) P. D.

Mannheim, "Implications of cosmic repulsion for gravitational **theory.'"** *Phys. Rev. D,* Nov. 15, 1998, p. 3511.

582. J. C. Bedini and T. E. Bearden, " Process For Transducing **Forms** of Electromagnetic Energy," U.S. Provisional Patent Application, 200 I. A formal U.S. Patent Application is in preparation as of this writing.

583. For a deep consideration of what is "meant" by mass, and how subtle the concept really is, the reader should refer to the scholarly work of Jammer {523a}. The more casual reader should be prepared for somewhat of a shock. Mass is not at all a "cut and dried" concept, as usually assumed. Indeed, the simple materialistic view of mass seems to be very much opposed by modern physics findings.

584. (a) Bearden, {12}. If we insert the absorption of a scalar (time-polarized) photon by the ubiquitously-assumed unit point charge and the subsequent re-emission of a longitudinal photon as the process for "combining" a scalar and longitudinal photon, then our solution is also consistent with (b) Mandl and Shaw, {19}. Mandl and Shaw argue that the longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. Our comment is that this argument, translated from particle terminology to wave terminology, then fits our re-interpretation of Whittaker's 1903 decomposition of the scalar potential in (c) Whittaker, {85}. It also strongly suggests the need for a complete reinterpretation of what is meant by "propagation of EM energy through 3-space".

585. E.g., see (a) V. I. Arkhipov *et al.,* "Negative transient currents in amorphous semiconductors." *Internal. J. Electronics* (UK), 51(6), 1981, p. 735-742; (b) — "An analysis of the dispersive charge transport in vitreous 0-55 As_2S_3 : 0*45Sb_2S_3," *Phys. Stat. Sol.* (A), Vol. 54, 1979, p. 67-77: (c) — and Rudenko, A. I., "Negative currents caused by injection-controlled polarization." *Solid St. Commun.,* 28, 1978, p. 675-676.

586. Evans, {137a, 137b, 251}.

587. David Jones, *Vancouver Sun Times,* Weekend Magazine, Dec. 17, 1977, p. 17. Article on Sid Hurwich's Toronto gun-jamming demonstration, stopping of watches, and jamming of a gun.

588. For proof of production of ordinary EM fields and energy at a distance by scalar **interferometry,** see **Evans** et **al., {**79}.

589. Miles and Bush, {198}, p. 101.

590. Bearden, {17}.

591. We point out, but do not further discuss, the obviously startling impact on the validity of the random statistical analyses normally used in scientific work if such "hidden variables" are present in the Earth's supersystem and interact with the experiment.

592. Kron, {192}, p. 466.

593. (a) Gabriel Kron, *Diakoptics: The Piecewise Solution to Large-Scale Systems,* MacDonald & Co., London, 1962. See also (b) Kron, {310b}.

594. Ilya Prigogine, Preface, (160b), p. xi.

595. Sir William Rowan Hamilton, 1837. Hamilton originated quaternion algebra, the Hamiltonian, etc. Hamilton's quaternion algebra has a higher group symmetry than either vectors or tensors. Since Maxwell's theory was first formulated in Hamilton's quaternion algebra and later reduced to vectors and then to tensors, ironically the first successful EM theory has been sharply curtailed rather than being further developed and extended.

596. (a) Aharonov and Bohm, {344a, 344b}.

597. Prigogine, {160b}, p. 104.

598. Lee, {441}, p. 23.

599. Jackson, {22}, p. 811-812.

600. Whittaker, {85}.

601. This is because time may be regarded as compressed spatial length, where the L compression factor is c. Since the energy of a wave is proportional to L^2 (where L is the wave amplitude), then the release of ST time-density as ST energy density provides nearly $9x10^{16}$ times as much effect per second as simple energy density curvature alone does per joule. For convenience, we refer to this as the "10^{17} gain factor" one obtains by releasing the compressed length energy from its time-density ST curvatture form.

602. T. Matsumoto, "Mechanisms **of Electro-Nuclear Collapse,"** *Proc, ICCF-7 (International Conference on Cold Fusion - 7),* Vancouver, BC, Canada, April 1998, p. 98.

603. Romer, {84}, endnote 24, p. 109.

604. See Harter et al., {102}. For a simple analogy, we may think of the longitudinal EM wave as a sort of velocity-modulated wave; i.e., it is rhythmically changing its velocity about some average value, where that value is usually taken to be c. Another way to visualize it is as an "accordion" wave, expanding and contracting along the line of travel.

605. There are several types (polarizations) of photons utilized in physics. E.g., see (a) Ryder, {90b}. On p. 147+, Ryder discusses four polarization states of the photon. Two of these are "transverse" —i.e., the well-known x- and y-polarizations of light; one is called "longitudinal," i.e., z-directed and related to Evans' B(3) field, and one is "scalar" or timelike.

The time-like (scalar) photon we are using in this paper seems to be yet another brand, something like a cross between the Ryder's third and fourth classifications. That is, our "scalar" photon is named such because when it is imperfect it propagates (i.e., in the 7 direction) at finite speed. When perfect, it just appears instantly everywhere it will be, without propagation through space in the usual sense. Hence it can be at an infinite number of points simultaneously - very close to the notion of the pure Coulomb gauge or of Bohm's quantum potential - or it can be impure and propagate at subluminal or superluminal speeds. The main notion we intend in our present use of the "scalar" photon for this treatise is that its oscillation is in the time domain and along the time axis. We use a model which produces the flow of time itself, and a time-component structuring of that flow — something which is still absent from physics. So our time-like (scalar) photon in one part very much resembles Ryder's discussion of the "scalar" photon, but is not just that alone. For instance, both our longitudinal photon and our scalar photon seem to have internal "photon" structures that we do not discuss in this treatise.

A still deeper discussion of photon polarization is in (b) Mandl and Shaw, {19}. The longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (electrostatic) potential. So by this route one arrives at a well-established mathematical theory that allows a hidden substructure within **the photon itself.** Mandl and Shaw include the photon's four polarization states as related to the four independent degrees of freedom available in the vector potential A. Suppression of the longitudinal and scalar polarizations is accomplished by "fixing the gauge". This reduces the independent degrees of freedom from 4 to 2, giving the conventional electrodynamics. However, when one raises serious foundations flaws in that resulting conventional electrodynamics, it is obvious that the entire structure is still on somewhat shaky ground. An even better mode might be to consider the fundamental photon entity as existing in all degrees of freedom, and the various "kinds" of photons existing due to restrictions or partial restrictions placed on one or more of these degrees of freedom. This of course opens the stage to interactions that change one (restricted) type of photon into a different (restricted) type. Thai is essentially the basis for my use of the notion of "photon transduction". Tt also leads to the transposition of time into energy and vice versa, which I use in my own approach, but which is still missing from nuclear physics. Since experimentally the cold fusion phenomena seem to require that characteristic, we have retained it.

We are indebted for much of the preceding discussion to private correspondence from Bob Flower, a fine scientist and knowledgeable in this area. Flower has pointed out that the conventional photon theory outlined by the cited authors is missing the possibility that nonlinear phase-conjugate optics could make the photon's longitudinal and scalar polarizations manifest in an experiment. Or put another way, if the instantaneous Coulomb potential could be decomposed into a time-like component and a longitudinal component, it would allow the two hidden polarization states to be controlled directly. This would result in anomalous modifications of the EM vector potential, leading to measurable electromagnetic or quantum effects. Flower's observation seems to be the closest to our own view, stated more simply in the present paper.

606. The result of our yet-unfinished struggle in these *very* swampy photon waters is not entirely consistent with the conventional view. Nonetheless, the new approach does seem to fit and explain most of the anomalous cold fusion phenomena, as well as a wide variety of other phenomena previously deemed beyond the pale of physics itself. E.g., by noting that *mind is time-like in its operations,* and therefore directly uses time-polarized (scalar) photons, waves, and electrodynamics, then mind operations become totally electromagnetic, but of a dramatically extended electrodynamic nature presently being approached in quantum field theory That approach leads directly to a testable and engineerable theory of mind

on all its levels (conscious, subconscious, unconscious, species unconscious, biospheric or all-species unconscious, and so on). But since the mind can interact with the body and produce at least the initial spatial forces carrying out intent in the physical world, it follows that some process must exist whereby the "time-like" photon interacts with and changes one or more of the other photons in the conventional view. Our use of *transduction* in fact encompasses that, based on quantum field theory as described by Ryder {90b} and by Mandl and Shaw {19}. Further, the present author's transduction reinterpretation of Whittaker's 1903 and 1904 papers {85, 91a} means that Whittaker's work as reinterpreted strongly supports this approach mathematically. So the approach used directly incorporates *mind and matter interaction,* again on multiple levels. While this area of *engineerable* mind and matter interaction is well beyond the scope of this treatise, it has become possible to begin the physics of the living being — even of the entire living universe — and to foresee direct engineering of the various realms of mind itself. Indeed, Russian psychoenergetics - particularly the highly weaponized secret part of it - already is involved in just that capability in deployed operational equipment of great portent. Since real results and real explanations of diverse metabiological phenomena have arisen from the approach as well as a very deep "more ordinary" physics including the necessary corrections to electrodynamics, we have continued to follow it to see where it eventually leads. A future book is planned on the more positive and beneficial uses of such startling technology. A preliminary view can be had on the author's website, www.cheniere.org, in the "porthole concept" briefing. A patent application has in fact been filed on this startling new method of medical healing.

607. (a) Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian], Novosibirsk, 1981 [an English exposition of much of the Kaznacheyev work is given in (b) Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," *Psi Research,* 1(1), Mar. 1982, p. 47-76.] See also (c) Kaznacheyev *et al.,* "Distant intercellular interactions in a system of two tissue cultures," *Psychoenergetic Systems,* 1(3), Mar. 1976, p. 141-142. In the same issue, see (d) "A Comment" by William A. McGarey, p. 143; (e) "A Comment" by Arthur C. Hastings, p. 143-144. See also (i) Yu. A. Vladimirov, "Ultraweak luminescence of subcellular structures," *Ultraweak Luminescence in Biology: Proceedings of the Symposium,* Moscow, 1969 [In Russian], Robert Becker also showed in his healing of

otherwise intractable bone fractures that a simple scalar potential (now known to involve the direct transduction between scalar and longitudinal photons and vice versa), that cellular dedifferentiation and redifferentiation of red blood cells occurred in the scalar potential region across the damaged bone cells, and directly in such manner as to "eliminate the delta engine" in those damaged bone cells in the fracture. The red cells shucked their hemoglobin, turning back (time-reversing) to a more primitive cell (earlier state). Then they redifferentiated (time-forwarded) into the type cells that make cartilage. There is still a delta engine between cartilage and bone! So these cartilage-like cells redifferentiated again into the kind of cells that make bone. These latter cells were deposited in the fracture, healing it. This process is in fact used today in various medical hospitals, without any real understanding of the physics mechanism involved. In short, it has long been experimentally proven without understanding the actual mechanism by which it works. See (g) Robert O. Becker, "The direct current field: A primitive control and communication system related to growth processes," *Proc. XVIInternal. Congr. Zoology,* Washington, D.C., Vol. 3, 1963, p. 179-183; (h) — and Charles H. Bachman and Howard Friedman, The direct current system: A link between the environment and the organism," *New York State J. Medicine,* Vol. 62, April 15, 1962, p. 1169-1176; (i) — and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals," *Bull. N.Y. Acad. Med,* Second Series, 48(4), May 1972, p. 627-64. See also (j) C. A. L. Bassett; R. O. Becker, and R. J. Pawluk, "Effects of electric currents on bone *in vivo," Nature,* Vol. 204, Nov. 14, 1964, p. 652-654.

608. Puthoff, {202}.

609. See (a) A. Shapere and F. Wilczek, Eds., *Geometric Phases in Physics,* World Scientific, 1989. See also (a) M V. Berry and J. M. Robbins, "Indistinguishability for quantum particles: spin, statistics and the geometric phase," *Proc. Roy. Soc. LondA,* Vol. 453, 1997, p. 1771-1790.

610. See Mandl and Shaw, {19}, Chapter 5 and simply reinterpret the coupled scalar and longitudinal photon pair as a spin 2 graviton.

611. In fact, Sweet's vacuum triode amplifier probably used the domains and nuclei in a barium ferrite permanent magnet in such a graviton reaction manner. In an epochal experiment designed and conceived by the present author, Sweet reduced the weight of the device

on the lab bench, smoothly and efficiently, by 90 degrees. For **the** description, see Sweet and Bearden, {29a}.

612. Reginald I. Gray, *Unified Physics,* Naval Surface Warfare Center, Dahlgren, VA, 1988, p. 2-19.

613. For the basic explanation of a scalar interferometer, see Evans et al., {79}.

614. See Chapter 3, paragraph 3.1 through 3.5.

615. Bearden, {12}.

616. As an example, see Bunge, {3}, p. 179-180. Quoting: " *The most complete direct interparticle action theory of electromagnetism is WHEELER and FEYNMAN'S (1949). Its central postulate is Fokker's action principle for a collection ofpoint charges.... from which the equations of motion for charged particles follow. As can be seen, only particle variables occur in this theory, which is an electrodynamics in AMPERE'S sense and therefore the opposite of a unitary field theory. (The theory is mathematically equivalent to an "adjunct" field theory, but this adjunct field, far from being free, is attached to the particles.) The theory has been abandoned by its creators because it is inconsistent with the quantum theory, in particular with the photon hypothesis. Yet it continues to appeal to many not only because it smacks of mechanism and operationalism, but also because no self-forces appear in it and consequently no infinite self-forces (and the corresponding self accelerations) But this is too small an advantage compared with its shortcomings even at the classical level. Indeed, the theory makes use of special relativity - as is apparent from its central axiom...yet it is inconsistent with it for SR borrows from CEM the assumption that e.m. signals propagate in a vacuum with a constant speed c and moreover lead an existence, which, while it lasts, is independent of both emitters and absorbers. In other words, by denying the existence of free radiation fields propagating in space, the action at a distance formidation of CEM contradicts its own basis: it is logically inconsistent.* "

Our comment is that the "photon" interaction itself must be re-examined in light of the demonstrated giant negentropy of every change and dipole in the universe. Allowing for the quantum field theoretic combination of the scalar and longitudinal photons as an observable "photon" that is actually a spin-2 graviton in 4-space rather than 3-space. the entire nature of "propagation of EM energy through space" changes. No photon has ever interacted in just 3-space; instead, it interacts in

4-space. In the strong interpretation of giant negentropy, there is really no such thing as a "single observable photon" in the universe. Instead, there is a combined scalar/longitudinal photon pair, which when interacting in matter causes transverse EM wave movement of the interacting/detecting charges which are restrained by (i) the inertia due to their own mass, and (ii) in circuits, the back repulsion of all other like charges longitudinally ahead of the excited charge that absorbs the "photon" (actually, the graviton). Considering the longitudinal wave "bunching and relaxing" actions resulting in galloping, then the conventional assumption of the constant speed of light c in vacuum is destroyed. What we perceive as "light propagating in space at speed c" already moves in a galloping and jerky motion *about* speed c — now moving superluminally, now moving subluminally — with only an *average* speed of c overall. In turn, anything which affects the performance and "galloping mechanism" exhibited by the longitudinal EM wave can in theory be used to change the speed c. With the coming advent of the Fogal semiconductor and direct superluminal communication, all such objections based on previous rather classical assumptions will simply fall by the wayside. We also stress that the speed of gravitational field is shown in astronomical events and effects to be extraordinarily superluminal by many orders of magnitude. Again, this greatly upsets many of the modeling assumptions in present physics. Physics is based on a set of models and a set of experiments and observed effects. If the experimental effects negate the theory, one has to change the theory or be guilty of abandoning scientific method altogether, thereby turning science into dogma.

617. Cohen, {89}.

618. See Rod Donnelly and Richard Ziolkowski, "Electromagnetic field generated by a moving point charge: A fields-only approach," *Am. J. Phys.*, 62(10), Oct. 1994, p. 916-922. In conventional language, both the transverse and longitudinal field components always try to be born initially, whenever an EM wave is being generated. Generation of the transverse component produces a function which zeroes the longitudinal component. The function appears in the spacetime representation of the transverse EM fields, obtained by Donnelly and Ziolkowski using special transforms. The phase conjugating and retroreflecting aspects of cold fusion experiments produce spacetime curvatures, directly affecting the "canceling function." The altered function becomes an "LW-reduction function" which is postulated to determine the mix of transverse and longitudinal field components that emerge in the interactions.

619. Whittaker, {85} shows that the scalar potential is **comprised of** bidirectional longitudinal EM wavepairs. Whillaker, {91a} shows **that all** fields and waves are comprised of two interfering scalar **potential** functions. Hence all fields and waves are comprised of **functions** of longitudinal and scalar wave interferometry.

620. Evans et al., {79}.

621. Usually the detecting GM tubes will detect the interferometry of the incident longitudinal EM waves with the weak LWs being emitted in the tube itself by time-decay of its time-charges. The resulting transverse wave EM energy ionizes the tube's gas, and the GM tube "reads." If little or no time-decay is occurring, then there will be little or no transduction into TWs, hence little or no ionization. In that case the instrument will not read.

622. Miles and Bush, {198}, p. 101.

623. *Ibid.*

624. (a) Evans, {137b}; (b) Evans et al., {233a, 233f}.

625. Rodrigues and Lu, {93a}.

626. Enders and Nimtz, {329b}.

627. (a) W. A. Rodrigues, Jr. and J. Vaz Jr., "Subluminal and Superluminal Electromagnetic Waves and the Lepton Mass Spectrum," *Kluwer Ac. Pub. Proceedings,* <hep-th/9607231> on Los Alamos web site; (b) P. Letelier and W. A. Rodrigues Jr. (Eds.), *Gravitation: The Spacetime Structure,* World Scientific Publishing Co., Singapore, 1994; (c) J. Vaz Jr. and W. A. Rodrigues Jr., "On the equivalence of Maxwell and Dirac Equations, and Quantum Mechanics," *Internal. J. Theor. Phys.,* Vol. 32, 1993, p. 945-958; (d) J. Vaz Jr. and W. A. Rodrigues Jr., "Maxwell and Dirac Theories as an Already Unified Theory," *Adv. Appl. Clifford Algebras,* Vol. 7 (S), 1997, p. 369-386; (e) W. A. Rodrigues Jr. and J. E. Maiorino, "A unified theory for construction of arbitrary speeds ($0 < v < \infty$) solutions of the relativistic wave equations." *Random Operators and Stochastic Equations,* Vol. 4, 1996, p. 355-400; (f) Rodrigues and Vaz, {93b}.

628. Torn Van Flandern, "The speed of gravity - What the experiments say," *Phys. Lett. A,* Vol. 250, 21 Dec. 1998, p. 1-1 1.

629. *Ibid.,* p. 9.

630. *Ibid.,* p. 10.

631. A. A. Vlasov and V. I. Denisov, "Einstein's formula for gravitational radiation is not a consequence of the general theory of relativity," *Theor. Math. Phys.,* 53(3), June 1983 (English translation), p. 1208-1216. Translated from *Teoreticheskaya i Matematicheskaya Fizika,* 53(3), Dec. 1982, p. 406-418 [in Russian].

632. D. Hilbert, *Gottingen Nachrichten,* Vol. 4, 1917, p. 21.

633. Sen, {68}, p. viii.

634. Mandl and Shaw, {19}.

635. Logunov, A. A. and Yu. M. Loskutov, "Nonuniqueness of the predictions of the general theory of relativity," *Sov. J. Part. Nucl.,* 18(3), May-June 1987, p. 179-187.

636. Whittaker, {85}

637. I.e., that spacetime == scalar potential == virtual particle flux == Whittaker biwave set. If that holds, then when we engineer the internal structure of the scalar potential, we also simultaneously engineer the internal structuring of the vacuum and spacetime itself, with deliberately formed spacetime curvature engines (vacuum engines). This is vacuum engineering — by different means — of the sort proposed by Nobelist T. D. Lee, {73}, p. 380-381. On p. 383, Lee points out that the microstructure of the scalar vacuum field (i.e., of vacuum charge, or — in our view — of the vacuum potential) has not been utilized. Particularly see Lee's own attempt to indicate the possibility of using vacuum engineering, in his "Chapter 25: Outlook: Possibility of Vacuum Engineering," p. 824-828.

638. Ingram Block and Horace Crater, "Lorentz-Invariant Potentials and the Nonrelativistic Limit," *Am. J. Phys.,* 49(1), Jan. 1981, p. 67.

639. Roderic Lakes, "Experimental Limits on the Photon Mass and Cosmic Magnetic Vector Potential," *Phys. Rev. Lett.,* 80(9), Mar. 2, 1998, p. 1826. Our critique is that the given equations define nothing at all, since no equation is a definition. If one replaces the "=" with '≡' in the "definition" of the A-potential in terms of B, one then has B==curl A, which is actually the definition of B in terms of A. If A has no energy, then neither can B, which is just the swirl of the A-energy. If A has no curl, B vanishes. If A has no energy, B vanishes. Yet $dA/dt = - E$, whether

A has curl or not, and E has energy. Ergo, A must have energy if **any** operation upon it is to produce **E-field** energy, else one has **violated the** conservation of energy law and advocates creation of energy from **nothing.** Regardless of how one looks at B = curl A and B = curl A, neither the equation nor the identity defines A in terms of B, because B is always, just one aspect of A and yet A can exist without any curl at all — without B as demonstrated in the Aharonov-Bohm effect. But in no fashion can B exist without A. Therefore A is primary and B is derived, and if A has no energy then B can have no energy. V has energy and energy density, since in fundamental units any point of the potential V is identically the joules of energy collected upon a unit point positive electrical charge assumed at that point. If it had no energy, then no joules could be collected from it.

640. Weinberg, {194}, p. 109-110.

641. Sen, {68}, p. viii.

642. (a) Lee, {13a}; (b) Wu et al., {14}; (c) Lee, {239}, p. 57-59; (d) Lee, {73}, p. 184.

643. Bearden, {12}

644. Whittaker, {85}.

645. The clustering of virtual charges around observable individual charges — due to the polarization of the vacuum — is well known in quantum electrodynamics. Macroscopically one sees the charge as partially shielded by the cluster, so the clustering charges are known to have macroscopic significance. At close enough distances, one sees the "bare" charge with reduced shielding.

646. Mandl and Shaw, {19}.

647. (a) Aharonov and Bohm, {344}; (b) R. G. Chambers, {474}.

648. E.g., see (a) Barrett, {345}; Barrett and Grimes [Eds.], {235b}.

649. (a) Evans *et al,* {113a}; (b) — "Self-Inconsistencies of the U(l) Theory of Electrodynamics: Michelson Interferometry," *Found. Phys. Lett.,* 12(6), Dec. 1999, p. 579-584; (c) — *J. New Energy,* Vol. 4, Winter 1999, p. 231-235; (d) — "Derivation of the Lehnert Field Equations from Gauge Theory in Vacuum: Space Charge and Current," *Found Phys Lett.,* 13(2), April 2000, p. 179-184; (e) Evans et al., {233g}; (f) Evans **et** al.. {233e}; (g) — **"Equations of the** Yang-Mills Theory of Classical Electrodynamics," ***Optik, 111***(2), 2000, p. 53-56; (h) **Evans** el **al.,** {113***b}.***

Also (i) more than 100 papers on restricted U.S. Department of Energy Internet website http://www.ott.doe.gov/electromagnetic/.

650. Obviously we exclude active systems such as matter that decays in nuclear fashion. Those are not inert systems, but are "pre-excited" systems whose matter has had excess energy since its formation during the birth of the universe.

651. (a) Kondepudi and Prigogine, {160c}; (b) — "Thermodynamics, Nonequilibrium," *Ency. Appl. Phys.* Vol. 21, 1997. p. 311-337.

652. I.e., the known broken symmetry of the charge (regarded as a set of composite dipoles) in its vacuum energy exchange.

653. Kraus, {76}, p. 578, Figure 12-60, a and b.

654. James Croll, *Phil. Mag.,* Series 5, Vol. 2, 1876, p. 242-254.

655. Heaviside, {5a-5c}.

656. Poynting, {4a, 4b}.

657. Heaviside, {106}.

658. E.g., see Lorentz, {109}.

659. E.g., (a) Panofsky and Phillips, {110}; (b) Gough and Richards, {224b}. (c) Jackson {22}, p. 259, points out the arbitrary nature of the Poynting vector and disposes of any added nondivergent energy flow vector with the words *"Such an added term can, however, have no physical consequences."*

660. Simply examine *Vqt* around the forward emf section (the external circuit between the generator terminals and the ends of the source dipole inside the generator), and - *Vqt* across the generator terminals (and the source dipole).

661. Evans etal, {38b, 38c}.

662. Bohren, {24}. Independent replication is reported in the same journal issue; see (b) Paul and Fischer, {25}.

663. Bearden, {40}.

664. See Paragraph 8.3.16 in Chapter 8 for the explanation.

665. Olariu and Popescu, {482}.

666. Bertram Schwarzschild, "Currents in normal-metal rings **exhibit** Aharonov-Bohm effect," *Phys. Today,* 39(1). Jan 1986, p. 17-20.

667. Paul and Fischer, {25}.

668. One actually inputs far more electrical energy flow from **the** external electrical power supply than is conventionally accounted, if one accounts for the huge nondiverged Heaviside energy flow component that Lorentz arbitrarily discarded. That energy flow is actually present, but Is not accounted once Lorentz' integration trick is applied to integrate the entire energy flow vector around a closed surface assumed around any volume element of interest.

669. Evans et al., {113a}.

670. As is any EM system when the nondiverged very large Heaviside energy flow component is accounted in addition to the diverged Poynting energy flow component.

671. Bearden and Bedini, {582}.

672. As an example, two scalar potentials fi_1 and fi_2 may have precisely the same intensity magnitude, but different Whittaker internal bidirectional longitudinal EM wave structures. The effects of the two externally identical potentials — as in various chemical reactions — can be quite different. See Bearden, {17}.

673. Lee, {73}, p. 380-381.

674. Lee, {73}, p. 184.

675. For succinct summary of the major symmetries and broken symmetries in physics, see Lee, {239}, p. 57-59.

676. Wang and Chung, {33}.

677. Jean-Louis Naudin, "Chung's Negative Resistance Experiment", on website http://jnaudin.free.fr/enr/enrexpl.htm.

678. Burford and Verner, {35a}.

679. Early but sophisticated analogue simulator. E.g., see Kron, {306}, p. 173.

680. E.g., see (a) Kron, {32}, p. 39. Quoting: *"Although negative resistances are available* ***for*** *use with a network analyzer...* "; — "Numerical solution of ordinary and partial differential equations by

means of equivalent circuits." See also (b) Kron, *J. Appl. Phys.,* Vol. 16, Mar. 1945a, p. 173. Quoting *"...since none or only a few negative resistances exist on practical network analyzers) "* (accent added). In the second quotation, arguably Kron was required to insert the underlined words. In the first quotation, the censors missed the clarity of the statement.

681. Kron, {192}.

682. Sweet and Bearden, {29a}.

683. Self-oscillation in such magnetic materials (which are also optically active) is well known. E.g., (a) L'vov, {29c}; (b) Borovik-Romanov and Sinha [Eds.], {369b}. See particularly (c) Gurevich and Melkov, {232c}.

684. Rosenthal publicly verified the results of his tests of the Sweet device to the 26th session of the Intersociety Energy Conversion Engineering Conference (IECEC '91), Boston, Massachusetts, 1991.

685. David M. Pepper, "Applications of Optical Phase Conjugation," 254(1), Jan. 1986, p. 83.

686. Bearden, {15}, p. 2-10.

687. Bearden, {17}.

688. Time-polarized photons are recognized in quantum field theory, implying the existence of time-polarized EM longitudinal waves. In coupling with spatial longitudinally polarized EM waves in wavepairs, the coupling is observable as scalar potential energy (stress energy), based on an argument by Mandl and Shaw {19}. Opposing time-polarized longitudinal EM waves therefore constitute "time-stress", "time-stress potential", and optical pumping in the time domain, which is hypothesized to reverse the normal "forward time" law of attraction and repulsion of charged particles.

689. Kenneth Shoulders and Steve Shoulders, {38If}.

690. U.S. patent numbers 5,018,180; 5,123,039; 5,054,046; 5,054,047.

691. Jain, {35b}.

692. Cao et al, {264b}.

693. Prange and Strance, {365}.

694. Roger Penrose, {82}, p. v.

695. Bunge, {3}, p. 176.

696. Albert Einstein, {la}, p. xi-xii. Appendix I.

697. Raymond A. Serway, *Physics for Scientists and Engineers* ***with*** *Modern Physics,* Third Edition, updated version, Saunders College Publishing, Philadelphia, 1990, p. 620.

698. Max Planck, *Treatise on Thermodynamics*, 3rd ed., Dover, New York, 1945.

699. Paul Davies, *Superforce: The Searchfor a Grand Unified Theory ofNature,* Simon and Schuster. New York, 1984, p. 105.

700. D. Hilbert, *Gottingen Nachrichten,* Vol. 4, 1917, p. 21.

701. A. A. Logunov and Yu. M. Loskutov, "Nonuniqueness of the predictions of the general theory of relativity," *Sov. J. Part. Nucl.,* 18(3), May-June 1987, p. 179.

702. David Whitehouse, "Beads of doubt," BBC News, July 18, 2002, http://news.bbc.co.Uk/l/hi/sci/tech/2135779.stm.

703. Matthew Chalmers, "Second law of thermodynamics 'broken'," *New Scientist,* July 19, 2002, carried on New Scientist website at http://www.newscientist.com/news/print.jsp?id=ns99992572.

704. G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.,* 89(5), July 29, 2002, 050601.

705. Elias P. Gyftopoulos and Gian Paolo Beretta, *Thermodynamics: Foundations and Applications,* MacMillan, 1991. This is an elaborate text developed over the years at MIT.

706. Dilip Kondepudi and llya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, NY, 1999.

707. Valery Chalidze, *Entropy Demystified: Potential Order, Life and Money,* Upublish.com, 2000. Contains an excellent nonmathematical exposition of the second law.

708. J. S. Dugdale, *Entropy and Its Physical Meaning,* Taylor & Francis, 1996. This book is of a more mathematical bent, but a very thorough treatment.

709. Craig F. Bohren, *Am. J. Phys.,* 51(4), Apr. 1983, p. 323-327.

710. V. S. Letokhov, {157}, ibid.

711. (a) D. K. Sen, *Fields and/or Particles,* Academic Press, London and New York, 1968, p. viii. Thirty-two years later, we published the solution to that problem in (b) T. E. Bearden, "Giant Negentropy from the Common Dipole," *J. New Energy,* 5(1), Summer 2000, p. 11-23. Also carried on DoE restricted website http://www.ott.doe.gov/electromagnetic/ and on www.cheniere.org.

712. T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," *J. New Energy,* 3(2/3), 1998, p. 12-28.

713. Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics,* ibid., p. 459.

714. E.g., see D. Jou, *Extended Irreversible Thermodynamics,* Springer-Verlag, New York, 1996.

715. Sources for this table are: (a) Roy Porter, Consultant Editor, *The Biographical Dictionary of Scientists,* Second Edition, Oxford University Press, NY, 1994; (b) Kondepudi and Prigogine, {703), ibid.; (c) Serway {697), *ibid.;* (d) K. Martinas, L. Ropolyi, and P. Szegedi (eds.), *Thermodynamics: History and Philosophy,* World Scientific, London, 1991; (e) Stephen F. Mason, *A History of the Sciences,* New Revised Edition, Collier Books, NY, 1962; (f) E. T. Bell, *Men of Mathematics,* Simon & Schuster, NY, 1986; (g) Charles Singer, *A Short History of Scientific Ideas to 1900,* Oxford University Press, London, 1959; (h) William C. Reynolds, *Thermodynamics,* Second Edition, McGraw-Hill, NY, 1968; (i) Paul Edwards (Ed. in Chief), *The Encyclopedia of Philosophy,* Vols. 1-8, Macmillan, NY, 1967; (j) I. Asimov, *Asimov's Biographical Encyclopedia of Science and Technology,* Second Revised Edition, Doubleday, Garden City, NY, 1982; (k) Jeff Biggus, *Sketching the History of Statistical Mechanics and Thermodynamics,* 2002, http://history.hyperjeff.net/statmech.html, (1) J. B. Jones and G. A. Hawkins, *Engineering Thermodynamics,* Wiley, New York, 1986; (m) Kenneth Wark, *Thermodynamics,* Third Edition, McGraw-Hill, 1977.

716. Antoine Laurent Lavoisier, *Elementary Treatise* ***on Chemistry,*** 1798.

717. **I.** Asimov, *Asimov's Biographical Encyclopedia of* ***Science*** *and Technology,* Second Revised Edition, Doubleday. Garden City, NY, 1982.

718. Benjamin Thompson (Count Rumford), *Enquiry Concerning the Source of Heat which is Excited by Friction,* 1798.

719. Eugene F. Mallove, "The Mysteries and Myths of Heat: A Brief History of Hot and Cold," *Infinite Energy*, May-June 2001, issue #37.

720. Julius Robert Mayer, 1814-1878. In 1842, Mayer's classic paper on conservation of energy was published in *Annalen der Chemie.* He then published several papers himself. With no recognition and bitterly attacked, Mayer suffered a nervous breakdown and attempted suicide, being committed to institutions for several years. From 1858 he began to receive well-justified recognition. He was lauded by the scientific community and received the Royal Society's Copley Medal in 1871.

721. Rudolf Clausius, "On the Moving force of Heat and the Laws of Heat Which May be Deduced Therefrom," 1850.

722. Mallove, {719}.

723. James Clerk Maxwell, "A Dynamical Theory of the Electromagnetic Field," *Roy. Soc. Trans.,* Vol. CLV, 1865, p 459. Also in *The Scientific Papers of James Clerk Maxwell, 2* vols. bound as one, edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-597.

724. Actually, it is not *proven* that the energy of the universe is constant. With the recent experimental findings that the expansion of the universe is accelerating, the kinetic energy of the universe could be increasing. Whether this is offset by an appropriate change in the physical universe's supersystem remains to be seen.

725. Jeff Biggus, "Sketching the History of Statistical Mechanics and Thermodynamics," http://history.hyperjeff.net/statmech.html.

726. For a discussion of the modern death of matter and the death of materialism, see Paul Davies and John Gribbin, *The Matter Myth,* Simon & Schuster, NY, 1992. Quantum mechanics shook the foundations of materialism and the clockwork world, and chaos has **finished the** devastation. For a scholarly dissertation on mass, see Max **Jammer,**

Concepts of Mass, in Classical and Modern Physics, Harvard University Press, Cambridge, MA, 1961.

727. Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From heat Engines to Dissipative Structures,* Wiley, New York, 1998, p. 3.

728. Serway, {697}, p. 507.

729. Serway, *ibid.,* p. 507.

730. Serway, *ibid.,* p. 508.

731. Serway, *ibid.,* p. 527.

732. Lindsay and Margenau, *Foundations of Physics,* Dover, New York, 1963, p. 214.

733. Robert H. Romer, "Heat is not a noun," *Am. J. Phys.,* 69(2), Feb. 2001, p. 107-109. This is an editorial discussion by the Editor of *Am. J. Phys.* of the concept of heat in thermodynamics, where heat is not a substance, not a thermodynamic function of state, and should not even be used as a noun.

734. Serway, *ibid.,* p. 527-528.

735. Serway, *ibid.,* p. 588.

736. T. E. Bearden, *Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications,* CTEC Proprietary, Mar. 21, 1998, 200+ page inclosure to CTEC Letter to Gen. (Ret.) Walter Busby, Deputy Secretary of Defense for Counterproliferation and Chemical and Biological Defense, March 21, 1998. This communication proposed a crash program to develop a portable suitcase-size treatment device based on this effect, for use in treating mass casualties after forthcoming terrorist attacks by weapons of mass destruction. Treatment of a patient would have required only five minutes. Copies were also sent to the U.S. Air Force, National Institutes of Health, and other key government agencies. Three and a half years later, the attack on the World Trade Center and the Pentagon — and a foiled attack on the White House — occurred on September 11, 2001. Today the nation is at war against international terrorism, and extensive changes in defense of the heartland are also under way. However, there still appears to be no real solution to the problem posed by mass casualties, particularly which professional

strikes can produce casualties in the hundreds of thousandsor a few millions.

737. (a) A. J. Berteaud and A. M. Bottreau, "Analysc des **rayonne**ments electromagnetiques emis par I'appareil Priore," [Analysis of **the** electromagnetic radiations emitted by the Priore apparatus], D.R.M.E., 1971, p. 3-12; (b) A. J. Berteaud, A. M. Bottreau, A. Priore, A. N. Pautrizel, F. Berlureau, and R. Pautrizel, "Essai de correlation **entre l'**evolution d'une affection par Trypanosoma equiperdum et 1'action d'une onde electromagnetique pulsee et modulee," [Trial of the correlation between the evolution of a disease by Trypanosoma equiperdum and the action of a pulsating and modulated electromagnetic wave], *Compt. Rend. Acad Sri.* (Paris), Vol. 272, 1971. p. 1003-1006: (c) R. Cambar, "Rapport general des travaux de la Commission de Controle constitute en vue de verifier l'un des effets biologiques obtenu par l'utilisation de l'appareillage de Priore A. Bordeaux," [General findings of the work of the control commission formed to verify one of the biological effects obtained by use of the apparatus of A. Priore at Bordeaux], 1969, I vol., (d) R. Courrier, "Expose par M. le Professeur R. Courrier, Secretaire Perpetuel de L'Academie des Sciences fait au cours d'une reur on a L'Institut sur les effets de la Machine de M. A. Priore le 26 Avril 1977," [Presentation by Professeur R. Courrier, Perpetual Secretary ot the Academy of Sciences, made at the meeting of the Academy on the effects of the machine of M. A. Priore.] (e) G. Delmon and J. Biraben, "La croissance du carcinome de Guerin sour Faction de champs electromagnetiques."[The growth of carcinoma treated by the action of electromagnetic fields], *Rev. Path. Comp.*, 3(2), 1966, p. 85-88. (f) G. Mayer, A. Priore, G. Mayer and R. Pautrizel, "Action de champs magnetiques associes a des ondes electromagnetiques sur l'orchite trypanosominne due lapin," [Action of magnetic fields associated with electromagnetic waves on the typanosomian orchitis of the rabbit], *Compt. Rend. Acad. Sci.* (Paris), Vol. 274, 1972, p. 3011-3014. Many more reports of the novel cures are in the French medical literature, particularly by Pautrizel (g) Priore's doctoral thesis (which was rejected when the project was supressed) is: *Guerison de la Trypanosomiase Experimental Aigue et Chronique par L 'action Combinee de Champs Magnetiques et D 'Ondes Electromagnetiques Modules.* [Healing of intense and chronic experimental trypanosomiasis by the combined action of magnetic fields and modulated **electromagnetic** waves], thesis submitted in candidacy for the doctoral degree, 1973. (h) Antoine Priore, "Apparatus for producing radiations penetrating living

cells," U.S. Patent No. 3,368,155, Feb. 6, 1968;(i) — "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; (j) — "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963; (k) Eric Perisse, *Effets des Ondes Electromagnetiqv.es et des Champs Magnetiques sur le Cancer et la Trypanosomiase Experimentale* [Effects of Electromagnetic Waves and Magnetic Fields on Cancer and Experimental Trypanosomias], Doctoral thesis, University of Bordeaux No. 83, March 16, 1984.

738. (a) R. O. Becker, "The neural semiconduction control system and its interaction with applied electrical current and magnetic fields," *Proc. XI Internatl. Cong. Radiology,* Vol. 105, 1966, p. 1753-1759, Excerpta Medica Foundation; Amsterdam; (b) — "The direct current field: A primitive control and communication system related to growth processes," *Proc. XVI Internatl. Congr. Zoology,* Washington, D.C., Vol. 3, 1963, p. 179-183; (c)———"Electromagnetic fields and life," *Psychoenergetic Systems,* Vol. 3, No. 1-4, 1979, p. 119-128; (d) - - "A technique for producing regenerative healing in humans," *Frontier Perspectives,* 1(2), Fall/Winter 1990, p. 1-2; (e) - - "The Biological Effects of Magnetic Fields — A Survey," *Med. Electron. Biol. Engr.,* Vol. 1, 1963, p. 293-303; (f) — - "The bioelectric field pattern in the salamander and its stimulation by an electronic analog," *IRE Trans. Med. Electronics,* Vol. ME-7, July 1960, p. 202-207; (g) - - and Charles H. Bachman, and Howard Friedman, "The direct current system: A link between the environment and the organism," *New York State J. Med.,* Vol. 62, April 15, 1962, p. **1169-** 1176; (h)———and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals," *Bull. N.Y. Acad. Med.,* Second Series, 48(4), May 1972, p. 627-64; (i) - - and Carlton F. Hazlewood, Abraham R. Liboff, and Jan Walleczek, *Electromagnetic Applications In Medicine,* NIH-OAM Electromagnetics Panel Report, Jan. 15, 1993; (j) B. Baker; J. A. Spadero, and R. O. Becker, "Electrical stimulation of articular cartilage," *Ann. N.Y. Acad. Sci,* Vol. 238, 1974, p. 491-499; (k) C. A. L. Bassett; R. O. Becker, and R. J. Pawluk, "Effects of electric currents on bone *in vivo," Nature,* Vol. 204, Nov. 14, 1964, p. 652-654. There are many other publications by Becker, who is a true pioneer.

739. http://www.cheniere.org.

740. Kondepudi and Prigogine, {727}, p. 90-91. See also W. Nernst, *A New Heat Theorem,* Dover, New York, I969.

741. Wang et al., {704}.

742. Maddox, John, *Nature,* 417 (2002), p. 903.

743. Buchwald, Jed Z., *From Maxwell to Microphysics* (University of Chicago Press, 1985), p. 44.

744. For a better meaning of "static" as dynamic but steady state, see Tom Van Flandern, *Phys. Lett. A,* 250 (1998), p. 8-9.

745. E.g., Sen, D. K., *Fields and/or Particles* (Academic Press, 1968), p. vii.

746. Wu, C. S. et al., *Phys. Rev.,* 105, 1957, p. 1413.

747. Weinberg, Steven, *Dreams ofa Final Theory* (Vintage Books, Random House, 1993), p. 109-110.

748. Jackson, J. D. and L. B. Okun, *Rev. Mod. Phys.*, 73, 2001, p. 663-680.

749. Kondepudi, Dilip and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures* (Wiley, New York, 1998), p. 459.

750. Jou, D., *Extended Irreversible Thermodynamics* (Springer-Verlag, 1996).

751. Bohren, Craig *F.,Am. J. Phys.*, 51, Apr. 1983, p. 323-327. Under nonlinear conditions, a particle can absorb and re-emit more energy than is in the light incident on it. Metallic particles at ultraviolet frequencies are one class of such particles and insulating particles at infrared frequencies are another.

752. Paul, H. and R. Fischer, *Am. J. Phys.*, 51, Apr. 1983, p. 327.

753. See Hal Fox, and Mitchell Swartz (JET Technology, MA), "Progress in Cold Nuclear Fusion - Metanalysis Using an Augmented Database," presented at ICCF-5, Monaco, April 9-13, 1995. The authors extracted information from over 3,000 papers on cold fusion collected by Fox's company. Over 600 of these experiments reported excess heat."

754. E.g., see Petar K. Anastasovski, H. Fox, and K. Shoulders, "A New Approach to the Cosmic Red-Shift and to the Cosmic Microwave Sources," J. New Energy, 1 (2), Summer 1996, p. 79-87."

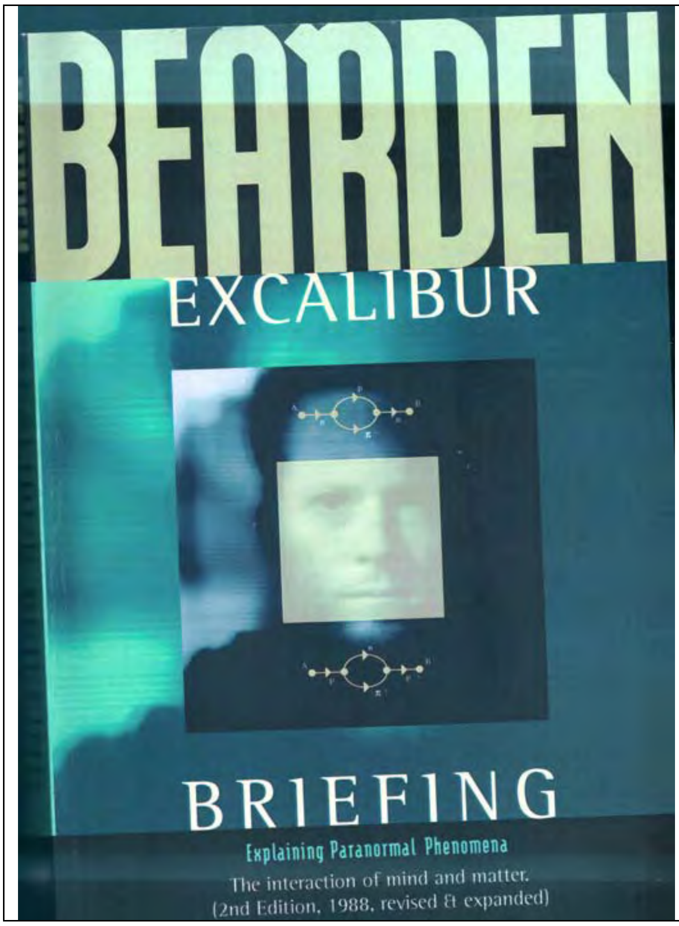

BEARDEN

EXCALIBUR BRIEFING

Explaining Paranormal Phenomena

The interaction of mind and matter.

(2nd Edition, 1988, revised & expanded)

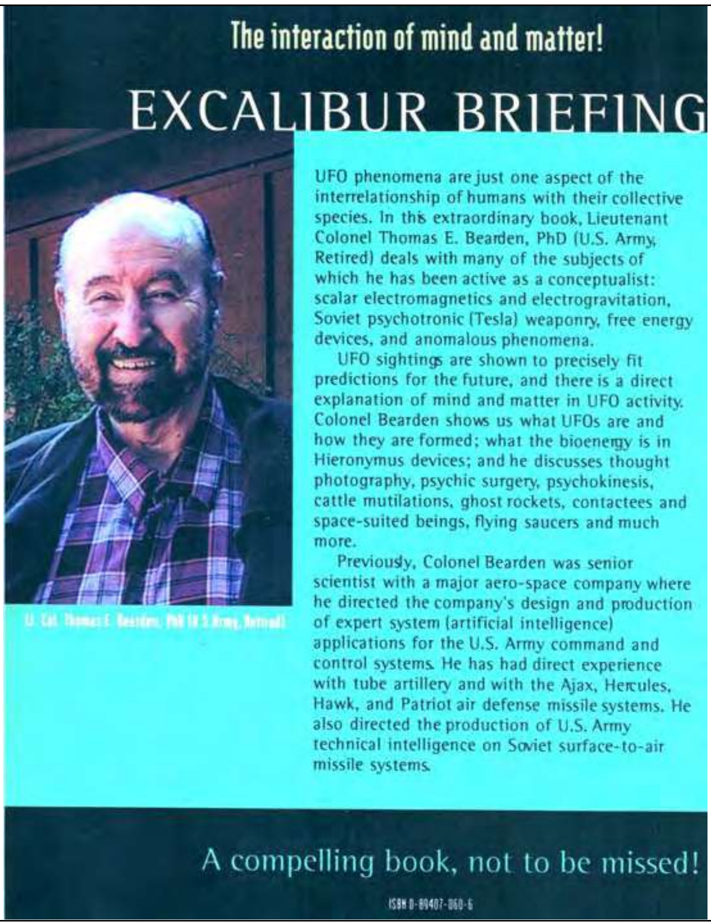

The interaction of mind and matter!

EXCALIBUR BRIEFING

UFO phenomena are just one aspect of the interrelationship of humans with their collective species. In this extraordinary book, Lieutenant Colonel Thomas E. Bearden, PhD (U.S. Army, Retired) deals with many of the subjects of which he has been active as a conceptualist: scalar electromagnetics and electrogravitation, Soviet psychotronic (Tesla) weaponry, free energy devices, and anomalous phenomena.

UFO sightings are shown to precisely fit predictions for the future, and there is a direct explanation of mind and matter in UFO activity. Colonel Bearden shows us what UFOs are and how they are formed; what the bioenergy is in Hieronymus devices; and he discusses thought photography, psychic surgery, psychokinesis, cattle mutilations, ghost rockets, contactees and space-suited beings, flying saucers and much more.

Previously, Colonel Bearden was senior scientist with a major aero-space company where he directed the company's design and production of expert system (artificial intelligence) applications for the U.S. Army command and control systems. He has had direct experience with tube artillery and with the Ajax, Hercules, Hawk, and Patriot air defense missile systems. He also directed the production of U.S. Army technical intelligence on Soviet surface-to-air missile systems.

Lt. Col. Thomas E. Bearden, PhD (U.S. Army, Retired)

A compelling book, not to be missed!

ISBN 0-89407-060-6

Sweating done by red_gonzo.

Note: sometimes formatting of the text may be a little fuzzy. It's all because the OCR process didn't go as smoothly as I expected. Also there may be some errors inside text so be vigilant. In any case share this book, send emails to people about knowledge contained in it.

Lieutenant Colonel Thomas E. Bearden (U.S. Army, Retired) Ph.D., was formerly senior scientist and department manager of a major aerospace company where he directed the design and production of expert systems (artificial intelligence) applications for the U.S. Army command and control systems.

He has over 29 years of experience in air defence systems, tactics, and operations; technical intelligence; command and control expert systems applications; nuclear weapons employment; computerized wargames; anti-radiation missile countermeasures; and military systems requirements. He has direct experience with tube artillery and with the Ajax, Hercules, Hawk, and Patriot air defense missile systems. He also directed the production of U.S. Army technical intelligence on Soviet surface-to-air missile systems.

Colonel Bearden obtained a Master of Science degree in nuclear engineering from the Georgia Institute of Technology, and a Bachelor of Science degree in mathematics from Northeast Louisiana University. He is a graduate of the U.S. Army Command and General Staff College, as well as of several U.S. Army artillery and missile schools.

He holds a Sandan (third degree black belt) in Yoseikan aikido and was the first president of the Yoseikan Aikido Association of North America. He is a former singer, guitarist, songwriter, and recording artist. He is member of Mensa, the Society for the Investigation of the Unexplained, the International Tesla Society, the American Association for the Advancement of Science, the Association of the American Physics Teachers, the Air Force Association, the Association of the U.S. Army, the American Nuclear Society, the Association of Old Crows, the American Defense Preparedness Association, and the Alabama Academy of Sciences. He is also East Coast Director for the Association of Distinguished American Scientists.

Tom Bearden is the author of several books and numerous papers, briefings and presentations, dealing with anomalous phenomena, scalar electromagnetics, free energy devices, Soviet Tesla weapons, and Soviet scalar electromagnetic weather warfare; he is also the producer of several videotapes on these subjects.

Table of Contents

Foreword, by John White . . . xv
Introduction to Second Edition . . . xvii
Introduction to First Edition . . . xxix

Part One: A Sampling of Specific Paranormal Phenomena

The Vestigia Mystery light Experiment . . . 35
Unexplained Generation of Electricity . . . 42
Hyperfield Generation of Photon Patterns About A Bar Magnet Collector . . . 43
Hyperfield Circulation Patterns Above a Magnetized Railroad Track . . . 46
The Cyborg Effect and Orthorotational Sensitivity . . . 49
Additional Factors Consistent with the Vestigia Experiment . . . 51
Remote Viewing . . . 52
The Moray Radiant Energy Device . . . 54
An Atlantean Power Crystal? . . . 60
The Crystal Skull . . . 61
Metal Bending . . . 65
Thought Photography: Stella Lansing . . . 65
The Displacement Effect . . . 68
Pavlita's Psychotronic Generators . . . 71
Cell-like Invisible Tupoidal Forms . . . 71
UFOs . . . 74
    1. Black Ring UFO . . . 74
    2. Anchor Patterns . . . 74
    3. Bifringement . . . 76
    4. UFOs in Water . . . 76
    5. UFOs in Clouds . . . 76
    6. UFO Seen by the Author . . . 76
    7. The 1966 Dennis Billings UFO Sighting . . . 82
    8. The Second Billings UFO Sighting . . . 87
    9. The 1973 Pascagoula UFO Sighting . . . 88
    10. Other UFOs . . . 89
Falkville Spaceman . . . 89
Men-in-Black . . . 94
Sasquatch . . . 95
UFOs Over the Soviet Union . . . 99
The Sighting Over Petrozavodsk . . . 103
French Experiments with Psychotronics . . . 104
Dowsing . . . 104
The Strobe Effect . . . 105
Kirlian Photography . . . 1ll
Interaction of Biofield, Thought and Spark Discharge . . . 1ll

Kervran Effect . . . 114
Psychic Surgery . . . 114
A Deliberate Experiment with Kindling . . . 118
Fay Clark's "Fireflies" . . . 118
"Foo" Fighters . . . 119
Water Monsters . . . 119
Cattle Mutilations . . . 123
Big Mama . . . 123

Part Two: A Theoretical Background for Understanding PT, UFOs, and Psi Phenomena

Some Unexplained Mysteries of Physics . . . 129
Uri Geller . . . 130
My Approach to Psychotronics . . . 131
    1. Reality as a Paranormal Bridge with Two Ends . . . 133
    2. A Fundamental Correction to Classical Logic . . . 136
    3. A Physical Example . . . 139
    4. New Definition of Zero . . . 140
    5. Einstein's Postulate . . . 141
    6. Synchronicity of Concepts . . . 142
Primitive Perception . . . 142
Two-Slit Experiment . . . 146
Photon Quenching of the Paranormal (Time) Channel . . . 151
Raindrop Model of Quantized Change . . . 155
Kindling Effect . . . 156
Bioenergy Collectors . . . 157
Psychometric Effect . . . 159
All Mass Is Charged: Everything Is Electrical . . . 160
A Funny Thing About Electromagnetic Field . . . 163
Extinguishing Electrical Currents . . . 165
Death Transmission Via the Paranormal Channel . . . 166
Disease and Death Induction . . . 169
Dynamic Thought Constructs: Biofield Mechanisms . . . 170
Radionics . . . 171
Psychotronic Devices: A Simple Space (Quiton) Amplifier . . . 172
Typical Hieronymus Device . . . 172
Another Kind of Hieronymus Device and the Inceptive Cyborg Effect . . . 175
Probability: Throw of a Die . . . 176
Negative Time . . . 177
Time Oscillation, Quantum Change, and Mass . . . 177
Mass Is a Time-Differentiation . . . 178
Reality Is Like an Opossum . . . 179
Biofield Concepts . . . 180
Mass, Force, and 3-D Frame Rotation . . . 180

Mass is pi Rotation of Hyperspatial Flux . . . 184
Jitter or Dither Modulates the Primary Flux . . . 185
The Untold Story of Quarks . . . 186
Einstein's Spherical Model of the Cosmos . . . 188
A Charged Particle Is a Dual Closure of the Universe . . . 188
Adding and Subtracting Dimensions . . . 189
Mind Is Objective . . . 191
The Skin Transducer Effect: Turning a Thing into Its Opposite . . . 193
Everett's Many-Worlds Interpretation of Quantum Mechanics . . . 193
Spacetime Curvature Is Tricky . . . 193
Multiple Realities (Many Worlds) . . . 194
Clustered Orthogonal Worlds . . . 196
Crosstalk Between Multiple Realities . . . 199
Biofields: And Maverick Worlds . . . 200
Hyperfield Flux in a Magnetic Field . . . 201
Life and Death . . . 202
The Holographic Hyperchannel Effect . . . 209
Mind Linkage . . . 212
Angels, Imps, and UFO Tulpas . . . 214
Dreams and Unresolved Conflicts . . . 215
Ghost Rockets and Flying Saucers: Materialized Tulpoids . . . 218

Part Three: New Military Applications of Psi Research

Background to Psychotronic Research in the U.S. and the U.S.S.R . . . 223
Psychotronic Background Since WWII . . . 224
Electron Dissolution . . . 225
The 1973 Contactee Wave and Cattle Mutilations . . . 226
Radiation of the U.S. Embassy . . . 229
Disease and Death Transmission . . . 229
Other Weapons . . . 230
Hyperspace Howitzer Operation . . . 230
A Possible Operational Test of Psychotronic Antiaircraft Weapons . . . 236
Population Conditioner . . . 238
A Midcourse ABM System . . . 241
The Bell Island Incident . . . 242
A Terminal ABM System . . . 245
Sweep Mode, Sweeping Away a Nuclear Detonation . . . 246
Zarg, Armageddon, and Linkage . . . 247
Virtual States and Hyperspaces . . . 254
Feynman Diagrams . . . 256
Nature of an Electromagnetic Wave . . . 261
Neutrinic Longitudinal Polarization Waves . . . 262
A New Look at Modulation . . . 265
Hyperspatial Aspects of Modulation . . . 269

The Eureka Concept 272
Biophotons and **Virtual EM Field of a** Bio-organism 273
The Priore Machine 273
The Neurophone 278
Soviet "Woodpecker" Signals 279

Part Four: Soviet Phase Conjugate Directed Energy Weapons

Weapons That Use Time-Reversed Electromagnetic Waves 285

Glossary 297

Bibliography 329

List of Photographs

1. The Vestigia mystery light experiment
   (© Vestigia, Stanhope, NJ. Used by permission.) 37-38,40
2. Hyperfield flux in a magnetic field (Courtesy George Meek) 44-45
3. Remote viewing (Courtesy Dr. H.E.Puthoff) 53
4. The Moray radiant energy device (©Dr. T.Henry Moray. Used by permission.) 54-55
5. An Atlantean power crystal? (Courtesy Al Bielek) 59
6. The crystal skull (©El Cariso Publications. Used by permission.) 62
7. Metal Bending (© *New Realities*. Used by permission.) 66
8. Clockwork patterns (© Mrs. Stella Lansing. Used by permission.) 67
9. Displacement effect (Courtesy George Meek and Paul Trovillo) 68
10. Pavlita's psychotronic generators (Courtesy James L. Hickman) 69-70
11. Cell-like invisible tulpoidal forms (©Trevor James Constable. Used by permission.) . . . 72
12. Form for potential materialization (©Trevor James Constable. Used by permission.) . . . 73
13. New Mexico UFO (Courtesy Wide World Photos) 75
14. Oregon UFOs (Courtesy Project Bluebook) 77
15. Black Ring UFO (Courtesy Project Bluebook) 78
16. Anchor patterns (Courtesy Ira Maxey, Project Bluebook, Brad Steiger, Fay Clark) . . . 79-80
17. Bifringement (Courtesy Project Bluebook) 81
18. UFO in water (Courtesy Project Bluebook) 81
19. UFO in clouds (Courtesy Project Bluebook) 83
20. UFO seen by author (Courtesy of the late Ray Palmer) 83
21. Second Billings UFO (© 1978 The American Association of Meta-Science) 85
22. Spanish UFO (Courtesy Spanish National Air Ministry) 91
23. Ohio UFO (Courtesy *Columbus Evening Dispatch)* 91
24. Heflin UFO (Courtesy Rex Heflin) 92-93
25. Falkville spaceman (Courtesy Wide World Photos) 96
26. Men-in-black coin (Courtesy Warren York) 98
27. Sasquatch (Courtesy Wide World Photos) 101
28. Bigfoot cast (Courtesy Wide World Photos) 102
29. The strobe effect 106-107
30. Kirlian photography (Courtesy Danniel H. Kientz) 108-110
31. Interaction of biofield, thought, and spark discharge (Courtesy James L. Hickman) . 112-113
32. Psychic surgery (Courtesy George Meek) 115
33. Kindling of thought bioenergy in snow (Courtesy Mrs. Ann Baune) 116
34. A deliberate experiment with kindling
    (© 1978 The American Association of Meta-Science) 117
35. Fay Claris "Fireflies" (Courtesy Fay dark) 120
36. 'Too" fighters 121
37. Pleiosaur-like creature hauled from the sea by Japanese fishermen
    Courtesy World Wide Photos) 122
38. Loch Ness monster (© Academy of Applied Science, Boston) 124
39. Cattle mutilation (Courtesy Tom Adams and Walt Andrus) 125
40. Big Mama (Courtesy Bill Jackson) 126
41. Hieronymus device (Courtesy James B. Beal) 219
42. Ghost Rockets (Courtesy of the late Ray Palmer) 244

List of Figures

1. Hyperfield circulation patterns above a magnetized railroad track . . . 47
2. The cyborg effect and orthorotational sensitivity . . . 49
3. Reality: a paranormal bridge with two ends . . . 133
4. A fundamental correction to classical logic . . . 137
5. A physical example . . . 139
6. Primitive perception . . . 143
7. Two-slit experiment . . . 147
8. Photon quenching of the paranormal . . . 150
9. Raindrop model of quantized change . . . 154
10. All mass is charged: everything is electrical . . . 161
11. Death transmission via the paranormal channel . . . 167
12. Typical Hieronymus detector . . . 173
13. Biofield concepts . . . 181
14. Rotation of 3-D frames . . . 182
15. Mass is re-rotation of hyperspatial flux . . . 183
16. Einstein's spherical model of the cosmos . . . 188
17. Adding and subtracting dimensions . . . 190
18. Turning a thing into its opposite . . . 192
19. Multiple realities (many worlds) . . . 195
20. Clustered orthogonal worlds . . . 197
21. Crosstalk between multiple realities . . . 200
22. Biofields: time-clustered orthoframes . . . 201
23. Hyperfield flux in a magnetic field . . . 203
24. Life and death . . . 204
25. Consciousness, life, and intent . . . 207
26. The holographic hyperchannel effect . . . 210
27. Output of stage i . . . 211
28. Layers of consciousness . . . 213
29. Dreams are unresolved conflicts . . . 216
30. Collective conflict initiates prophetic tulpoids . . . 219
31. Examples of UFO waves . . . 227
32. 1973 Contactee wave . . . 228
33. Cattle mutilations . . . 231
34. Induction of Legionnaire's disease . . . 232
35. Hyperspace howitzer operation (2 figures) . . . 235
36. Nuclear Hyperspatial howitzer . . . 237
37. Amidcourse ABM system . . . 240
38. A terminal ABM system . . . 248
39. Sweep Mode: sweeping away a nuclear detonation . . . 249
40. ZARG . . . 252

FOREWORD

by John White

This book deals in a comprehensive fashion with such a vast array of topics — from physics, psychology and evolution to advanced weaponry, military strategy and international politics — that many of you may become anxious, even fearful on reading it. In fact, this assembly of previously uncorrelated information and scientific breakthrough may be the most disturbing you have ever received.

This cannot be helped. This book represents an important step in our efforts to avoid global war involving not only nuclear armaments but also what appears to be a revolutionary development in Soviet military technology — psychotronic weapons. They are the modern equivalent of the legendary invincible sword, Excalibur.

The author is former U.S. Army officer Lt. Col. Thomas E. Bearden, now retired. During his twenty-year military career specializing in artillery, nuclear weapons, and military intelligence, Tom has pondered and researched many subjects that he covers here. His access to the full range of both military and NASA documents will be evident in his book. What may become highly classified data in the future is being revealed to you, our nation's ultimate policymakers, for an urgent reason that Tom has called Operation Fer-de-Lance, the presumed Soviet plan for domination of the planet through nuclear-psychotronic aggression and warfare. For the moment I would simply like to note two statements that indicate the basis for our sense of urgency:

1. On 13 June 1975, Soviet Premier Leonid Brezhnev urged a ban on research and development of new kinds of weapons "more terrible than anything the world has known."
2. In January 1977 former Air Force Chief of Intelligence, Maj. Gen. George J. Keegan (Retired) told *Newsweek* magazine, "The Soviets are working on dramatically exotic new weapons, twenty years ahead of anything ever conceived in the U.S. — so awesome as to lead the Soviets to believe that in the coming decade they would be capable of total neutralization of our ballistic and submarine missiles."

The U.S. military intelligence community interprets these signs in conventional terms as perhaps indicating the development of laser rays. Tom disagrees. His book will deal with the possibility that these weapons are psychotronic or paraphysical in nature.

In fact, we can say that in a strict sense, these are not weapons but devices — breakthrough technology with an enormous potential for the destruction of humanity. They are a second-or third-generation outgrowth of radionic devices presently used in unorthodox medicine for diagnosis and healing. Extremely compact, they could be carried in a suitcase and planted near targets — ICBM sites, communications centers, etc. Even if one of these devices were found and examined, it would only look like an electronic switching circuit or a small transmitter/receiver set, rather than a bomb or explosive.

The energy and the mechanism involved in psychotronic damping of nuclear, chemical or electromagnetic technology is, from all indications, unrecognized by official Western science but nevertheless inherent in it. That also has been deeply pondered by Tom, and this book will give a full explanation of the physics, mathematics and technology involved. It is not an exaggeration to say, as Tom does, that the explanation offers a "physics of metaphysics" — a fully developed theory of paranormal phenomena that elegantly unites physics and psychology. The ramifications of what Tom calls perception theory lead into every field of human knowledge and inquiry with revolutionary results — from science, theology and philosophy through the full spectrum of paranormal phenomena such as UFO's and psychokinesis into the elementary school classroom, the supermarket shopping basket and the voting booth. They most certainly lead into the safety of your homes and the integrity of your minds.

Before you turn to *The Excalibur Briefing,* I want to leave you with these thoughts. The material you are about to read will certainly be denied by our own government and military and probably by a major segment of the scientific community. It is not unlikely that Tom will be dismissed as a paranoid crackpot. Such tactics have been employed before against brilliant, courageous people who challenge the fundamental dogma of the establishment as Tom does here. Nevertheless, I want to reassure you that Tom is sane, sober, and serious.

Of course, such a person could still perhaps be mistaken. In that regard, I believe you will find Tom's presentation to be persuasive. He himself is careful to distinguish speculation from fact, inference from descriptive statement. Even if Tom were paranoid, however, the question arises: *What if there is good reason for paranoia?* To phrase it positively: We should welcome truth and enlightening insight whenever offered, regardless of the source. Do not let *ad hominem* arguments affect your critical judgment about the soundness of this presentation. Assess the work, not the man, for the work is always greater. And having judged the work affirmatively, render true acknowledgement — not lip service — by sharing the ideas in this book with others. Only free, informed, vigorous effort in the court of world opinion can defuse the possibility of Operation Fer-de-Lance and psychotronic mind control leading to total enslavement of the human race.

In closing, I want to point out that Tom's fundamental concern is the unification and evolution of the human species to a higher form of life through a process that he terms "mind linkage." This stage of growth to higher consciousness is explored in the latter part of the book. It is a breathtaking portrait of how the spirit of science can lead to a science of the spirit, to a physics of metaphysics.

INTRODUCTION TO THE SECOND EDITION

Since the first edition of *Excalibur Briefing* was published, a great deal of progress has been made on the actual physics of unusual and paranormal phenomena. Indeed, so much progress has been made in the ten years since the 1978 or so cutoff of most of the Excalibur material, that I really should rewrite the entire book. However, time will not permit that luxury at present I am simply too hard-pressed in the strenuous efforts to move a free energy machine to market and to convince our government that unorthodox Soviet superweapons present us with a problem of the gravest magnitude.

Accordingly, the inclusion of one of my latest papers, "Soviet Phase Conjugate Directed Energy Weapons," together with an updated glossary for the entire emerging field, will have to suffice.

The important thing is that **a separate new field of physics is beginning to emerge.** It turns out that what we've gotten on to is a unified field—one can now turn electromagnetic energy into gravitational energy and vice versa. Indeed, gravitational potential is now revealed as simply infolded electromagnetics, and electromagnetics is simply the outfolding of the internal contents of gravitational potential.

Colleagues have already proven it on the laboratory bench, by building both "free energy" devices and practical antigravity devices.

This new, extended electromagnetics—electrogravitation—I have dubbed **scalar electromagnetics.** The key to it is to let the EM force fields fight themselves to a "cancellation," forming **a** vector zero. If we do this in a nonlinear medium (modulator), the summing/multiplying EM vector forces are locked together and remain as an infolded EM system inside the EM vector zero. This now is a gravitational system. By merely varying (in phase) the summing EM components, one varies the local energy density of vacuum. Rigorously that is a gravitational wave. Further, it is a very powerful G-wave, for we are utilizing EM forces some 1042 times as strong as the normal weak G-force.

Classical EM has taught us to simply discard the components in the zero vector summation, replacing the system of opposing EM forces with a vector EM zero. This of course makes all vector EM zeroes equal. However, it totally ignores the space-time/vacuum stress of the summed forces, and the resulting gravitational implications. Therefore we have been taught to ignore any gravitational aspects of normal EM circuits, and indeed we proceed to release the G-stress effects in the circuits when we build them as we have been taught

Ordinary EM represents the release of the internal stress of gravity. By teaching us to make nonzero EM, scientists have taught us to always release electrogravitational potentials as electromagnetic forces. Thus in classical EM—and in most of the circuits we build—the EM force fields are the primary causative agents. Classically, when these force fields reduce to zero, there is no more electromagnetics going on. That is true, but there can still be a very large amount of gravity and inertial effects going on!

However, quantum mechanics teaches us that the EM force fields are not at all the primary causes. Instead, the potentials are the primary things; EM forces are simply made from these potentials by differentiating operators. And when the force fields are zero, the potentials still exist. Further, when these "forceless potentials" interfere—

even at a distance—real EM effects cun still be obtained, in total violation of ordinary EM and ordinary mechanics!

This is called the Aharonov-Bohm Effect, and in 1986 it was finally proven to most physicist's satisfaction. Only the most diehard skeptics still object to it, according to *Physics Today,* January 1986.

At any rate, when we have zeroed the E and B fields, we can still have the potentials, according to modern physics. These potentials can interfere to produce electromagnetic force effects—even at a distance.

Further, the zero summation can be dynamic; that is, the vector components can be varying, so long as they always sum to a zero vector. Actually such a zeroed system now has an infolded, deterministic pattern of dynamic forces, according to the pattern we deliberately construct. Further, the energy of each force component infolded is varying, and thus the overall energy density of local vacuum is varying.

As we said, rigorously this defines a gravitational wave. So the zero-vector approach allows EM wave energy to be turned directly into gravitational wave energy. Now it is the curvature of spacetime that is being varied locally. That's a really astounding thing.

We can now deliberately engineer the vacuum/spacetime itself, with all that that statement implies. Locally we can bend spacetime itself, which Einstein assumed could never be done except by a huge collection of mass—such as a sun, star, neutron star, etc. But here we can do it with simple things such as opposing coils, opposing E-fields, etc. We've just violated one of the sacrosanct, severely limiting assumptions of ordinary (Einstein's) general relativity!

That is, up to now most Western physicists have assumed that the local spacetime/vacuum cannot be curved. They assume that, where we the observer and the laboratory instruments are, only special relativity applies. Locally the observer always has a "Lorentz frame," in that view. This assumption (the local Lorentz frame) means that all the conservation laws apply, and physical reality stays well-behaved where the laboratory is and where we are. Ordinary relativity becomes just a sort of "special relativity with distant perturbations."

The only curvature of spacetime of any consequence is only to be found at or near distant stars, suns, black holes, and other huge collections of mass, in this highly restrictive view. Further, this assumption implies that one can never have an experimental general relativity, since where one is experimenting (locally) there is always essentially zero spacetime curvature!

However, one need not have such a formidable restriction at all. Two bucking transformers can make a B-field vector-zero a little bit More complex circuits can easily be designed to sum (or multiply) force fields to zero very well. When we do this zero summation, we also get an amazing gain as follows: Consider two electrons in free space. As is well known, the electric field force (repulsion) between them is about 10^{42} times as strong as the gravitational field force (attraction) between them. So the electric force is so very much stronger that the comparatively weak gravitational force can be entirely neglected, in most applications. However, if we "strangle" that electric field force by zero summation, all the opposing electric forces are converted directly into gravitational potential!

Interference of two such potentials can directly generate gravitational forces. So, if we could zero-sum their E-fields perfectly, we could instantly amplify the gravitational force between the two electrons by a factor of 10^{42}. With such a gravitational force amplification factor, our altered "electrical" circuits would actually become powerful electrogravitational circuits. Further, appreciable gravity effects would now be experienced by any mass, object, or device exposed to the action of the circuits. This is the fundamental secret that Nikola Tesla discovered, one which he never openly revealed.

Further, by locking opposing EM waves together, we may easily construct purely scalar EM (electrogravitational) waves and beams. In such case, amazing new phenomena are encountered: (1) one can increase or decrease the mass and inertia of an object; (2) one can produce a unilateral force in and on each and every nucleon in a body, thus accomplishing the long-sought "antigravity"; (3) a unilateral force drive—a "space drive"—can be constructed; (4) transmutation of elements becomes fairly simply done, with miniscule energy; (5) effects at a distance—such as cold explosions or hot explosions—can be accomplished; (6) negative energy and negative time effects can be created and utilized; (7) Newton's third law can be manipulated and violated at will; (8) negative entropy effects can be created and utilized; and (9) direct energy extraction from the vacuum can be accomplished, etc. So many new effects are encountered that it will take a generation of graduate students and university laboratories to get a good handle on most of them.

The reader is urged to remember these rules: (1) The destructive interference of scalar EM waves (electrogravitational waves) produces EM energy in, or extracts it from, the interference zone. (2) The destructive interference of ordinary EM waves produces gravitational potential energy. (3) The release of the ordinary EM force fields represents the bleed-off of a 5-dimensional gravitational potential as ordinary electromagnetics. (4) The bottling-up of the ordinary EM force fields represents the forced bleed-off of a 5-dimensional gravitational potential as ordinary 4-dimensional gravitational force. (5) The EM field and the ordinary (4-space) G-field are just separate bleed-off s of the same 5-dimensional gravitational potential field.

Thus, I make no bones about it; we have a new physics in the offing—one that is eminently practical and easily engineered by any modern electromagnetics laboratory.

Since the first edition of this book was issued, it has also become obvious that at least two other phenomena are involved in UFOs, and intermingled with the "materializations from the collective unconscious" true UFO phenomena. These two additional things are:

(1) A great number of "lights at night" and "mysterious fireball" type phenomena—and giant balls of light and glowing hemispheres of light of gigantic size—are actually due to deliberate Soviet long range, high energy, real-time holography using phase conjugation. These phenomena are to accomplish orientation and alignment of giant Soviet weapons, and at the same time deliberately evoke UFO reports and scientific skepticism, providing perfect deception.

(2) Several nations of the world—at least two—have been building and flying antigravity spaceships all over the solar system, since the early 1950's. A very small number of UFO ship reports and encounters are due to these ships and their crews. However, control of these ships is not under any duly elected government, but passes out of the government concerned, into and under the control of some mysterious group

which I loosely refer to as "the control group." This is the same mysterious group which has been suppressing free energy and antigravity inventors and inventions for a great number of years.

Thus in *Excalibur Briefing* we are still addressing the primary UFO phenomena: Materializations from the collective human unconscious. Coverage of the other two parts of the overall reported UFO phenomena will be made elsewhere.

So it seems that, under my hard-pressed circumstances, the best thing to do for the second edition of *Excalibur Briefing* is to include something that shows some of the drastic progress that has been made. By adding a glossary and a current paper, the book will also be made much more usable. I also have prepared an index for the first edition, but it cannot be included because of space limitations. Admittedly that is regrettable and still not the optimum, but it will have to suffice. Perhaps in the future I will be able to do a complete rewrite of the book, for *Excalibur III.*

Until then, I ask the reader's indulgence and understanding.

* * *

In the first edition we accented the hyperspatial aspects of the problem, but also pointed out that higher dimensional spaces and deeper nested levels of virtual state are in fact one and the same thing. The reader should keep in mind that when we "inter-nest" vector zero systems, we are performing direct engineering in hyperspace.

Let me be more specific. When you zero sum a group of EM force vectors, you place those component vectors "inside the zero," so to speak. In modern language you have placed ordered patterns directly inside the violent flux of virtual particles comprising the vacuum. The modern concept of the vacuum is that it is like a seething plasma or gas. However, it's a very strange gas, for each particle arises spontaneously out of nowhere (according to quantum mechanics), and almost immediately turns into something else and/or disappears. Thus the vacuum plasma or gas is rather peculiar; the particles comprising it appear and disappear so rapidly that they cannot individually be detected. However, while a virtual particle exists, it is in motion, and so a violent flux of these "phantom particles" comprises the vacuum/spacetime.

However, just as many fleeting particles in the vacuum are moving backward in observer time as are moving forward in it! So this wild "vacuum gas" of ghost particles exists in both positive and negative time at once!

Anyway, when we make a zero vector EM system, each component force in the summation represents a gradient in this virtual particle flux. That is, we have "patterned" and organized the flux—polarized it, so to speak. And that now overturns one of the fundamental assumptions of quantum mechanics.

Quantum mechanics assumes that the violent flux of vacuum ghost particles is totally randomized. That is, no deterministic pattern exists in the flux. These virtual changes collect or cohere statistically, not deterministically, in the standard assumption. When they happen to collect/cohere sufficiently, the threshold of an observable quantum change is breached, and a quantum change occurs.

It can readily be seen that, in the ordinary view, the occurrence of these quantum changes is totally statistical. And indeed it is, so long as the virtual particle flux is randomized. However, with zero vector summation (or multiplication) we have produced deterministic patterns in the flux—and hence in the cohering and collecting of this flux on observable charged particles. Therefore, these deterministic "virtual state" ordered patterns will result in the appearance of deterministic patterns in the occurrence of observable quantum changes.

Now we have begun to engineer physical reality directly; that is, one can now begin to pre-determine which statistical change shall be favored, or in the extreme case, which one shall occur, and whether or not it shall even occur at all.

The renowned physicist, David Bohm, has written a theory of "hidden variables" which could fit and describe all of physics (nothing out there contradicts it). It has largely been ignored, and will continue to be ignored unless it can show appreciable application that normal theory cannot cover. With the scalar EM approach, we can now directly provide hidden EM force field vectors, ordering the virtual state vacuum flux, and we may vary these ordered virtual patterns at will to produce startling new results in charged particle systems. So now we can directly engineer Bohm's "hidden variables," and the vacuum itself—something ordinary theory cannot cover. Drastic changes in the physicist's view of fundamental physical reality, and of what he can do about it, now result

But to return: We wish to engineer additional hyperspaces. (The first zero-vector summation places the component vectors in the fifth dimension.) If we take one of the component EM vector forces in the first zero vector summation, and we then take a point on it as yet another "reference zero", we may encapsulate a second series of zero-system summation vectors into and within that zero-point. In this case we have placed these second vectors into yet a deeper level of virtual state vacuum. That is, we have reached into yet another hyperspace with our engineering and control. Indeed, we have placed the second system EM vectors into the sixth dimension. So we have become true hyperspace engineers.

The process is repeatable as desired. We may reach any hyperspace we wish, as our technology and control develop.

Now we see the impact In modern Kaluza-Klein geometry, one needs about 11 or so dimensions if one is to account for quarks, charm, etc. Now we have a direct tool to reach into any one or all of these hyperdimensions and affect the exact particle or component desired. We will be able to engineer matter—including the nucleus, and including inside the proton and neutron—in almost any way we wish. It all follows from scalar electromagnetics, which allows the direct engineering of hyperspace in all its dimensions.

Remember, engineering in hyperdimensions is just the engineering of nested levels of vacuum potentials. Even more startling, we can also place "thoughts" and effects directly in the mind itself, including the unconscious mind.

As we stated in the first edition of this book, we may model the mind as simply an objective 3-space, three turns away from the 3-space of the ordinary physical world. If we nest two or more additional zero-vector systems in our first one, then we have placed affects ("vectors") directly inside the mind itself. If we know the code of the human brain-tuner, we can now manipulate the mind directly. Electromagnetically, by scalar EM. And it appears that the Soviet Union has begun to reach this frightening capability.

Disease patterns may be induced upon ordinary EM carriers in the same manner, by nesting the zero-vector systems. That is, of course, if one has measured the scalar patterns evident in the induced disease, say, in the Kaznacheyev-type experiments with the desired disease patterns. The reverse is also true: Nesting the patterns in the proper context on EM field carriers can cure diseases such as cancer and leukemia, as evidenced by the Priore device.

One can also make a practical "free energy" device by setting up a standing scalar wave resonance in a localized region or device. From the "high stress" point to the "low stress" region of vacuum formed by such a standing scalar wave, a literal "river of energy" is flowing in the vacuum itself. Once the river is established, only the "leakage" need be replenished. Hence, if one builds a special sort of "paddle wheel" to react with the river and turn in it, one obtains shaft power freely from the localized curvature of spacetime. In that case, an ordinary generator can be driven by the shaft torque to provide very cheap, or"free," electrical energy.

Several inventors—three of whom I work with in one respect or another—now have demonstration models of such machines, which take diverse forms.

Because of the extreme importance of these emerging potentialities, in recent years I've concentrated intensely upon the properties of the vacuum as spacetime, and uncovered shortcomings (and downright errors in many cases) in ordinary physics and electromagnetics. Slowly the scalar EM picture has emerged, and a series of papers has been written and published to develop the concepts of the new approach.

The scalar approach also includes the ability to engineer negative energy and negative time flow. The reader is referred to Chapter 4, "Extraordinary Physics," of my forthcoming book, *AIDS: Soviet First Strike*, Tesla Book Co., P.O.Box 1649, Greenville, Texas 75401 for a summary of the vast importance of the time-reversed EM wave (phase conjugation).

Let me here just accent a rule of great importance: It is actually the key to negative energy, negentropy, antigravity, free energy, transmutation, Tesla's wireless transmission of energy at a distance without loss, Tesla's original magnifying transmitter, and many other things.

First we need the expression CPTEGS, where C is charge, P is parity, T is time flow, E is energy, G is gravity, and S is entropy. We shall be concerned with the algebraic sign of each term in the expressioa The rule deals with photon interaction with a charged particle inside an atom. Normally, incident photons interact with an electron in the electron shells, so C is normally negative. This produces normal parity, and causes the normal or positive flow of time, s T is normally positive. The normal kinds of energy, gravity, and entropy are experienced, so energy is positive, gravity is positive (attraction), and entropy is positive (progressive interactions in a system tend to more and more disorder).

The astonishingly powerful rule is this: In a photon interaction with a charged particle in an atom, if any one of the terms of the expression CPTEGS is reversed in algebraic sign, all the other terms are reversed in algebraic sign.

Thus if one makes scalar EM, the wave (and its photons) passes through the electron shells and interacts with (is absorbed in, and raises to an excited state, which then decays to emit a wave and photons) the highly nonlinear, positively charged nucleus. This represents a reversal of C, hence a reversal of P, T, E, G, and S also.

Consequently, a time-reversed wave—a phase conjugate replica of the incident wave—is emitted by the decaying excited nucleus. In nonlinear phase conjugate optics, this is the mechanism that produces the phase conjugate wave. This is also the mechanism that produces a very large amplification of the incident wave when a "pump wave" (two waves 180 degrees out of phase) are used in the nonlinear (modulating) medium of the phase conjugate mirror.

(The reader is referred to any good book or summary paper on phase conjugate optics; for example, see David M. Pepper, "Nonlinear optical phase conjugation," *Optical Engineering* 21(2), Mar./Apr. 1982, p. 156-183.).

When one pumps a phase conjugate mirror material sufficiently, then "tickles" the release of the total nuclear excitation energy as phase conjugation replica waves (negative energy), appreciable negative time is produced in the decaying nuclei of the mirror. In negative time, gravity is a repulsion, not an attractioa The mirror thus produces antigravity, and gets lighter. If it is forced to sufficiently high amplification, it can be made to float in air. At higher amplification still, it will accelerate upward.

Some other references of interest to this new edition follow:

1. **Robert A. Fisher, Ed., *Optical Phase Conjugation,* Academic Press, New York, 1983. Collection of papers. An extensive bibliography is given at the end of the book.**
2. **Ya. Zel'dovich, N. F. Pilipetsky, and V. V. Shkunov, *Principles of Phase Conjugation,* Vol. 42, Springer Series in Optical Sciences, Theodor Tamir, Ed., Springer-Verlag, New York, 1985.**
3. **Amnon Yariv, *Optical Electronics,* Third Edition. Holt, Rinehart and Winston: New York, 1985. See particularly Chapter 16: "Phase Conjugate Optics—Theory and Applications."**
4. **Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory." *Physical Review,* Second Series. 115(3), Aug. 1,1959. p. 485-491.**
5. **R. G. Chambers, "Shift of an electron interference pattern by enclosed magnetic flux." *Physical Review Letters.* 5(1), My 1,1960. p. 3.**
6. **Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics.* Volume n, Section 15-5, p. 15-8 to 15-14.**
7. **E. T. Whittaker, "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions." *Proc. Lond. Math. Soc.,* Vol. 1,1903. p. 367-372.**
8. **T. E. Bearden, *Solutions to Tesla's secrets and the Soviet Tesla weapons.* Tesla Book Co., Greenville, TX, 1981.**
9. **T. E. Bearden, *Tesla's secret and the Soviet Tesla weapons.* Presented in absentia at the First International Unorthodox Energy Symposium, Toronto, Canada, 1981. Available from Tesla Book Co.**
10. **T. E. Bearden, "Soviet weather engineering over North America." 1-hour videotape, 1985. Available from P.O. Box 1472, Huntsville, AL 35807.**
11. **Gwynne Roberts, "Witness to a super weapon?" *The Sunday Times,* London, England, Aug. 17,1980.**
12. **Greg Rippee, "Mushroom cloud sighted off Japan." *Los Angeles Daily News,* Apr. 11,1984. p. 1,8.**
13. **Martin Ebon, "Moscow, June 11, 1977." in his *Psychic Warfare: Threat or Illusion.* McGraw-Hill Book Co., New York, NY, 1983. p. 1-11.**
14. **A. K. Lapkovskii, "Relativistic kinematic equations and the theory of continuous media." *Soviet Physics Journal* 21(6), June 1978-Dec. 1978.**
15. **B. N. Belyaev, "On random fluctuations of the velocity of light in vacuum." *Azvestiya Vysshikh Uchebnykh Zavederdi, Fizika.* 11, Nov. 1980. p. 37-42.**
16. **Ingram Bloch and Horace Crater, "Lorentz-invariant potentials and the non-relativistic limit" *American Journal of Physics.* 49(1), Jan. 1981. p. 67-75.**

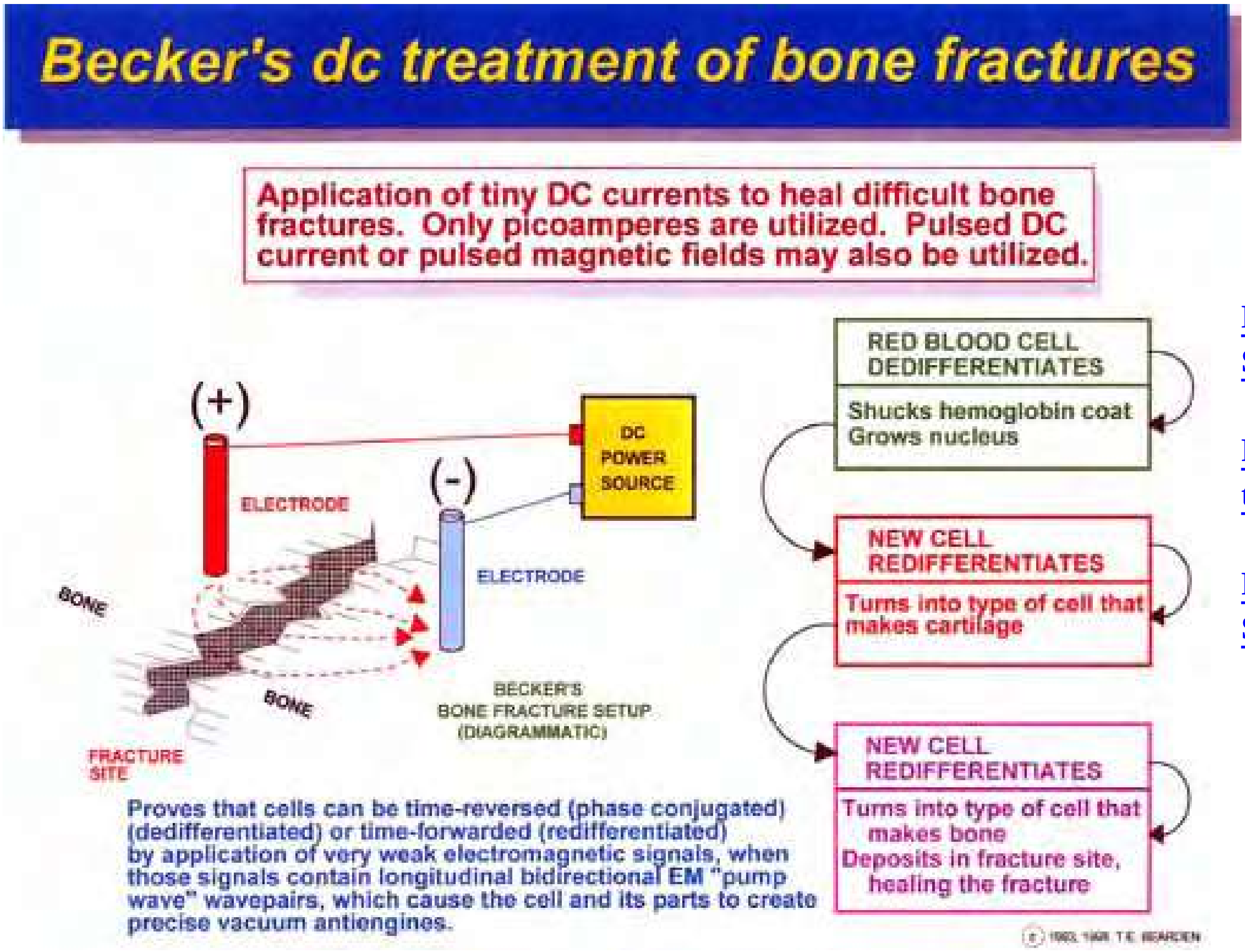

Becker's dc treatment of bone fractures

Application of tiny DC currents to heal difficult bone fractures. Only picoamperes are utilized. Pulsed DC current or pulsed magnetic fields may also be utilized.

(+)
ELECTRODE
(-)
DC POWER SOURCE
ELECTRODE
BONE
BONE
FRACTURE SITE
BECKER'S BONE FRACTURE SETUP (DIAGRAMMATIC)

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Proves that cells can be time-reversed (phase conjugated) (dedifferentiated) or time-forwarded (redifferentiated) by application of very weak electromagnetic signals, when those signals contain longitudinal bidirectional EM "pump wave" wavepairs, which cause the cell and its parts to create precise vacuum antiengines.

© 1993, 1998 T.E. BEARDEN

Next Slide

Return to Index

Previous Slide

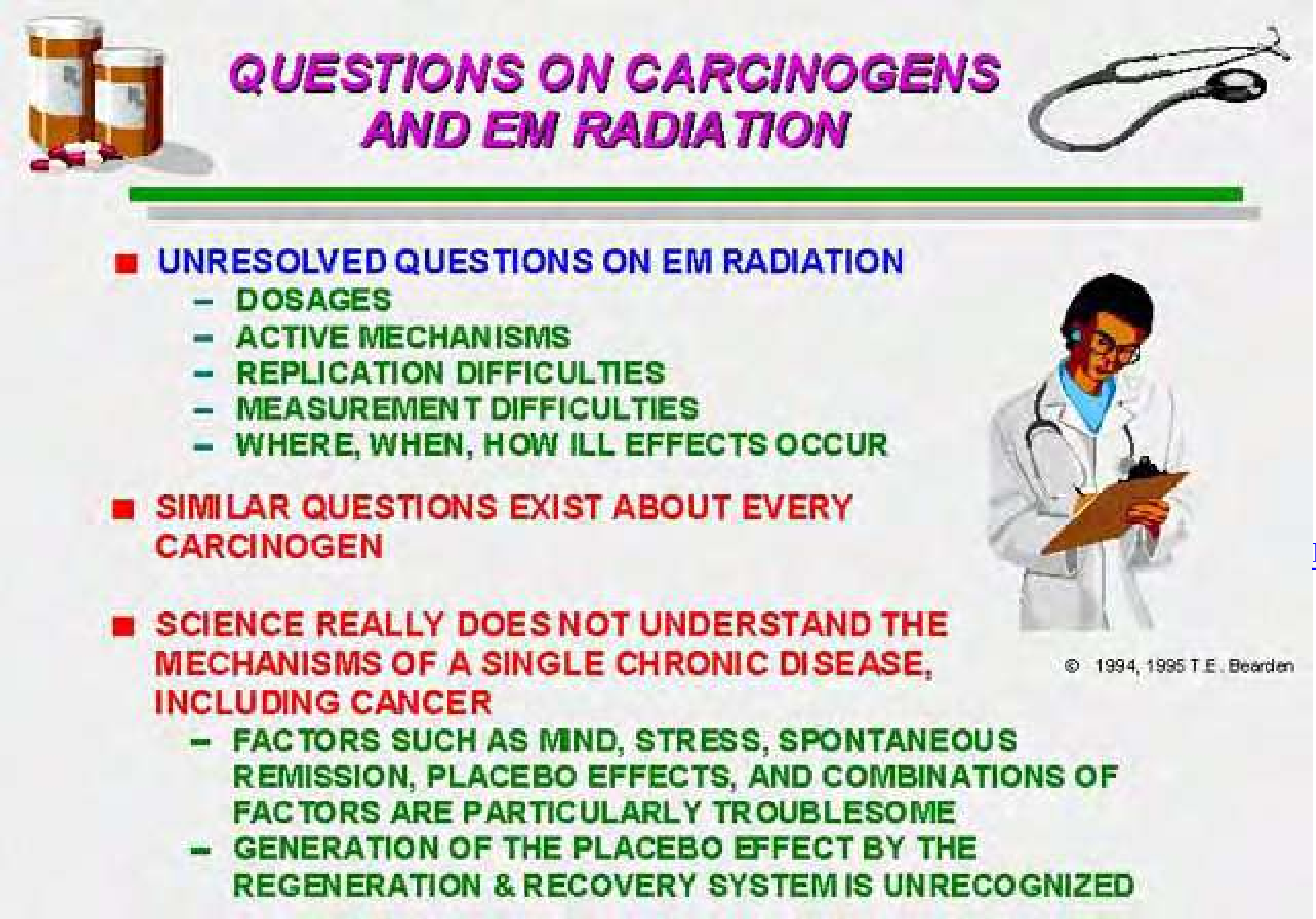

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

Next Slide

Return to Index

Previous Slide

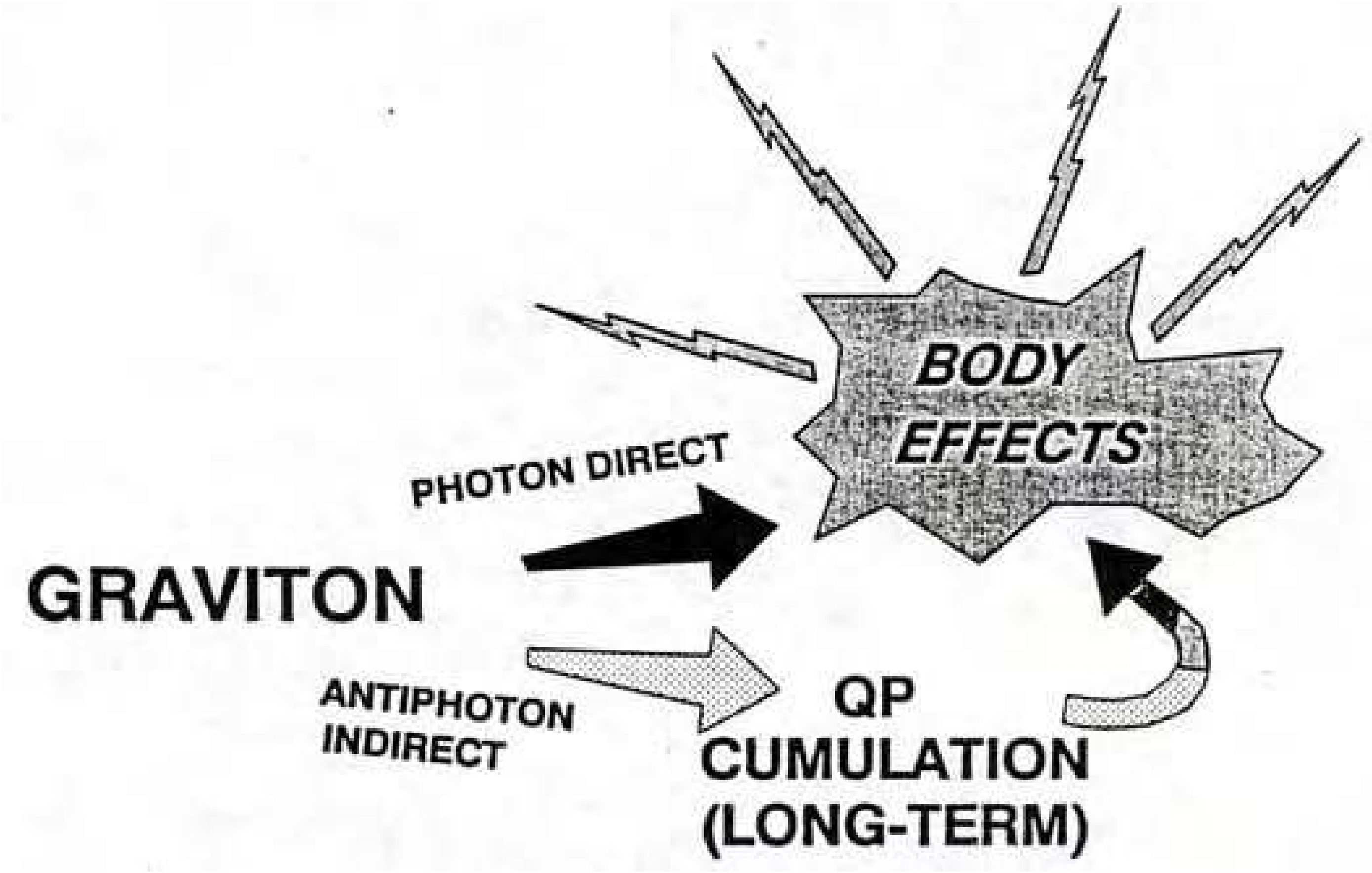
BODY EFFECTS
PHOTON DIRECT
GRAVITON
ANTIPHOTON INDIRECT
QP CUMULATION (LONG-TERM)

Next Slide

Return to Index

Previous Slide

CUMULATIVE FEEDBACK REACTION LOOP FOR INDUCTION OF LONG-TERM EFFECTS

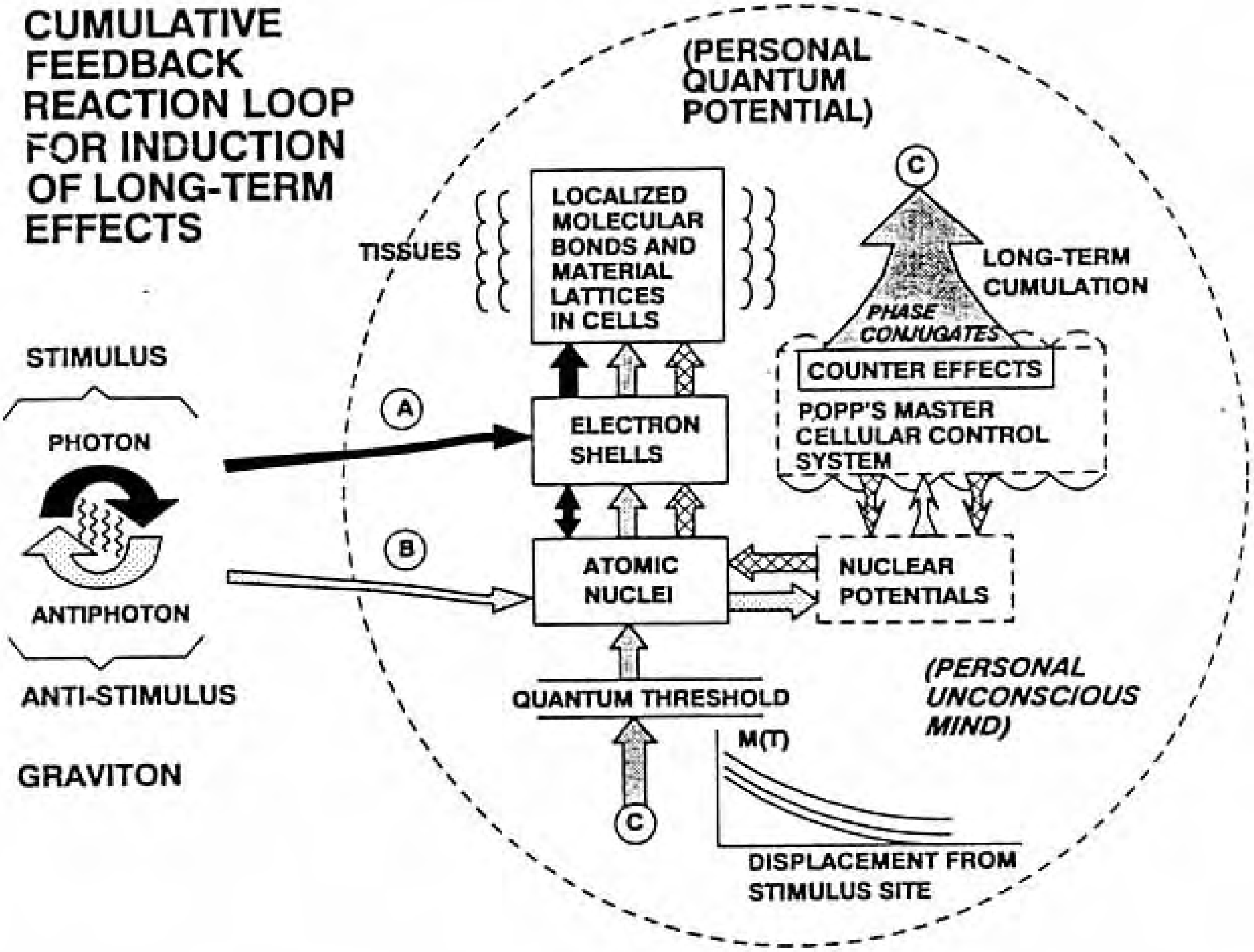
(PERSONAL QUANTUM POTENTIAL)
TISSUES
LOCALIZED MOLECULAR BONDS AND MATERIAL LATTICES IN CELLS
C
LONG-TERM CUMULATION
PHASE CONJUGATES
COUNTER EFFECTS
POPP'S MASTER CELLULAR CONTROL SYSTEM
STIMULUS
PHOTON
A
ELECTRON SHELLS
B
ATOMIC NUCLEI
NUCLEAR POTENTIALS
ANTIPHOTON
ANTI-STIMULUS
QUANTUM THRESHOLD
(PERSONAL UNCONSCIOUS MIND)
M(T)
GRAVITON
C
DISPLACEMENT FROM STIMULUS SITE

Next Slide

Return to Index

Previous Slide

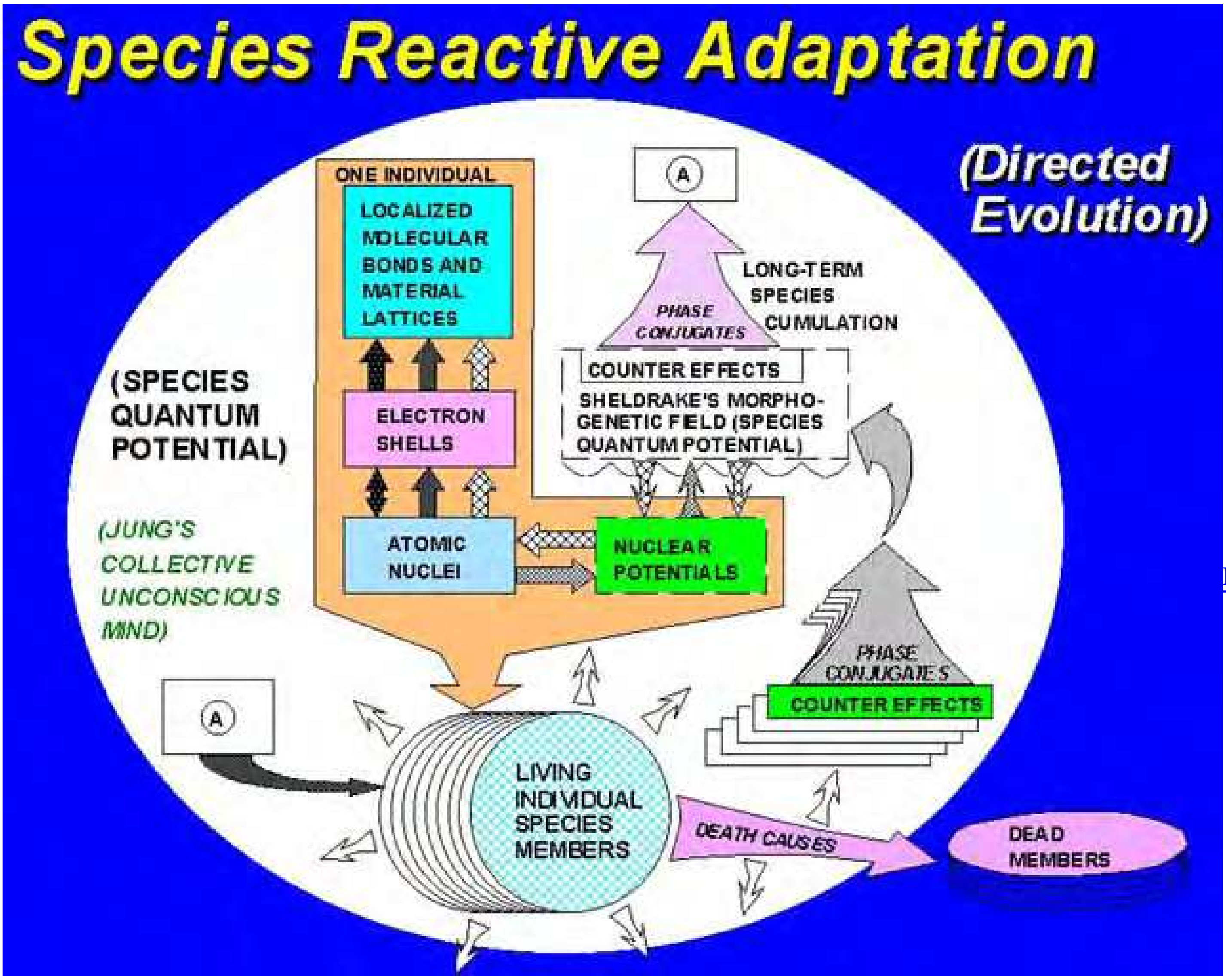

Species Reactive Adaptation

(Directed Evolution)

ONE INDIVIDUAL

LOCALIZED MOLECULAR BONDS AND MATERIAL LATTICES

ELECTRON SHELLS

ATOMIC NUCLEI

NUCLEAR POTENTIALS

A

PHASE CONJUGATES

LONG-TERM SPECIES CUMULATION

COUNTER EFFECTS

SHELDRAKE'S MORPHO-GENETIC FIELD (SPECIES QUANTUM POTENTIAL)

(SPECIES QUANTUM POTENTIAL)

(JUNG'S COLLECTIVE UNCONSCIOUS MIND)

A

LIVING INDIVIDUAL SPECIES MEMBERS

PHASE CONJUGATES

COUNTER EFFECTS

DEATH CAUSES

DEAD MEMBERS

Next Slide

Return to Index

Previous Slide

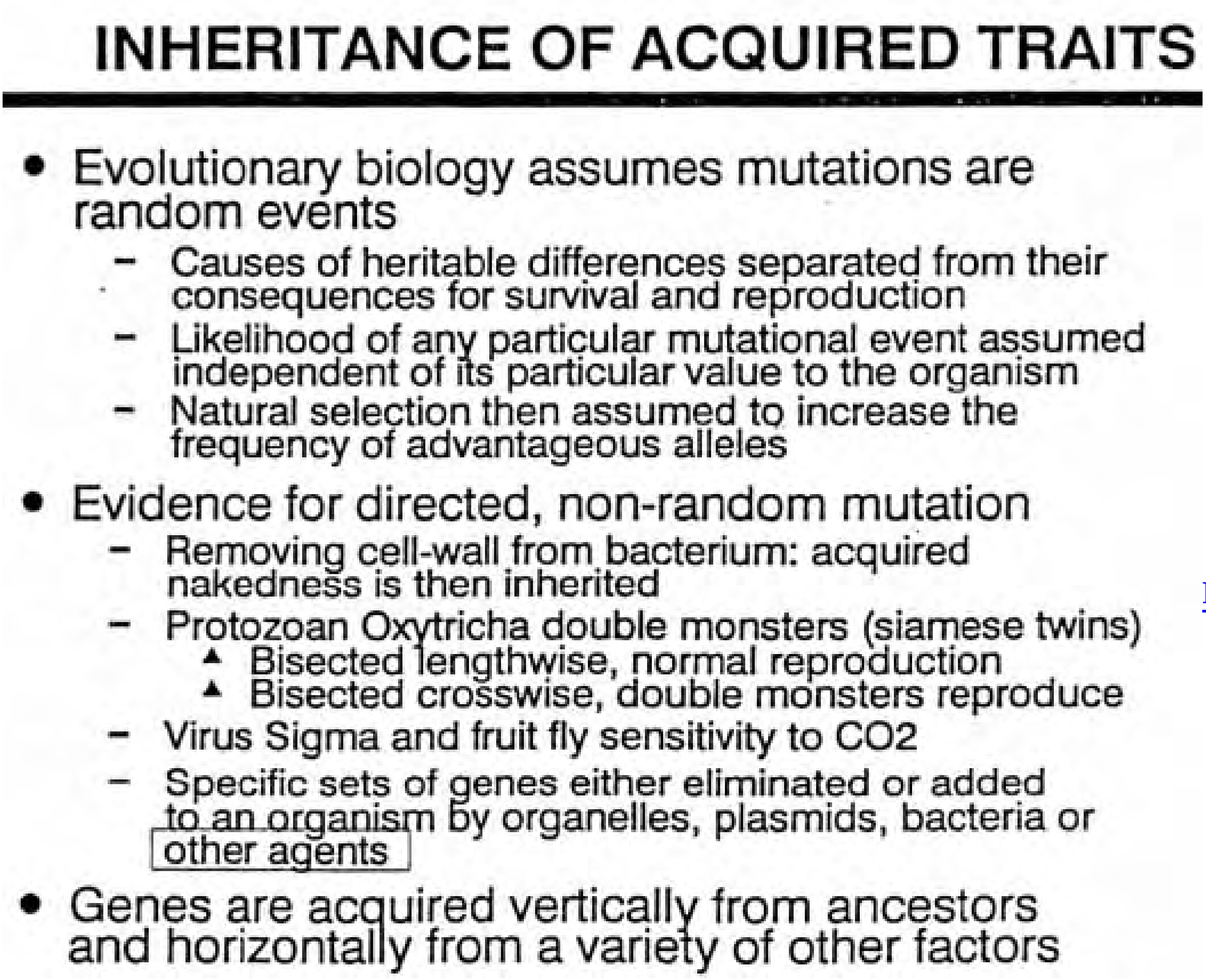

INHERITANCE OF ACQUIRED TRAITS

- Evolutionary biology assumes mutations are random events
  - Causes of heritable differences separated from their consequences for survival and reproduction
  - Likelihood of any particular mutational event assumed independent of its particular value to the organism
  - Natural selection then assumed to increase the frequency of advantageous alleles
- Evidence for directed, non-random mutation
  - Removing cell-wall from bacterium: acquired nakedness is then inherited
  - Protozoan Oxytricha double monsters (siamese twins)
    - ▲ Bisected lengthwise, normal reproduction
    - ▲ Bisected crosswise, double monsters reproduce
  - Virus Sigma and fruit fly sensitivity to CO2
  - Specific sets of genes either eliminated or added to an organism by organelles, plasmids, bacteria or other agents
- Genes are acquired vertically from ancestors and horizontally from a variety of other factors

Next Slide

Return to Index

Previous Slide

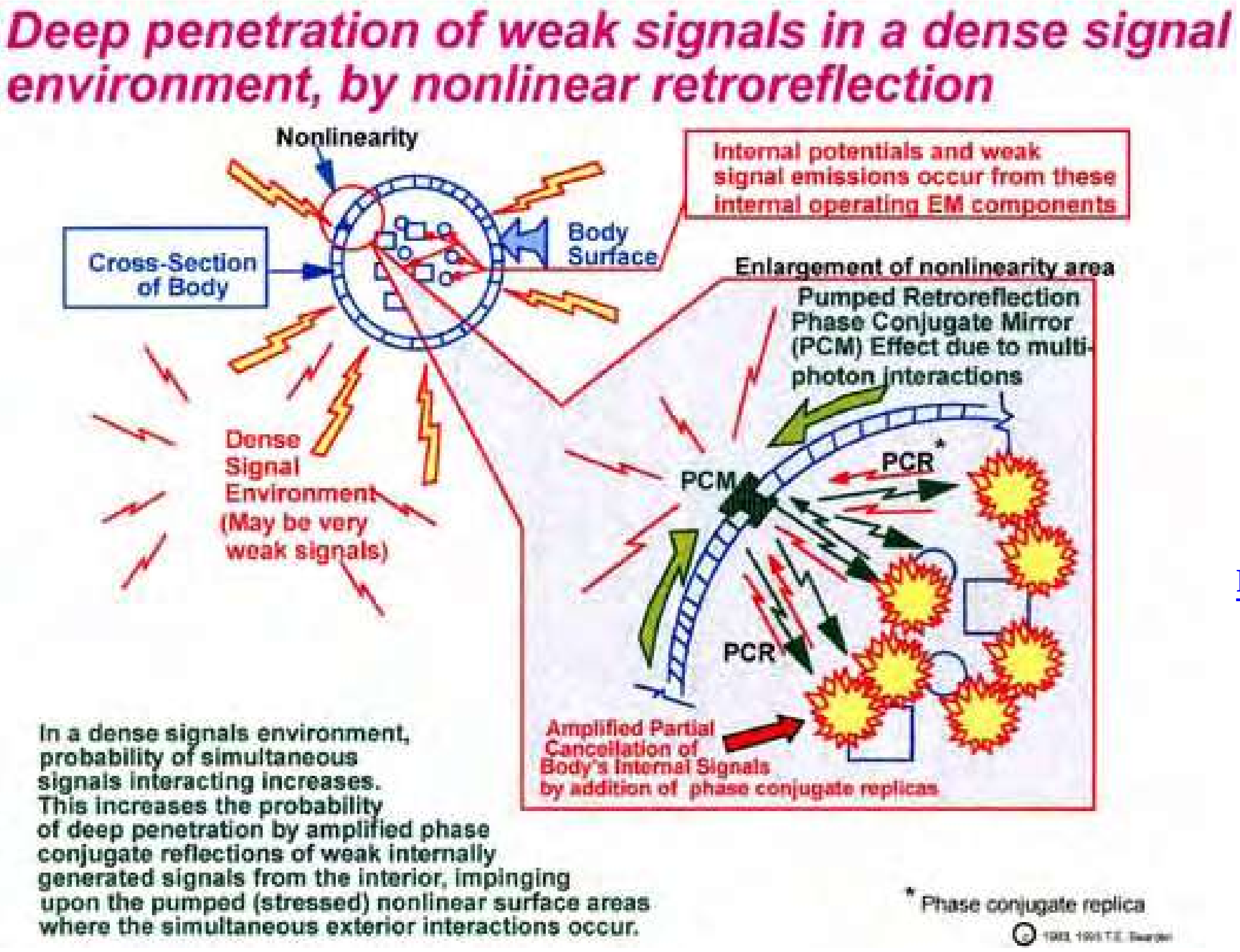

Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection

Nonlinearity
Internal potentials and weak signal emissions occur from these internal operating EM components
Cross-Section of Body
Body Surface
Enlargement of nonlinearity area
Pumped Retroreflection Phase Conjugate Mirror (PCM) Effect due to multi-photon interactions
PCM
PCR*
PCR*
Dense Signal Environment (May be very weak signals)
Amplified Partial Cancellation of Body's Internal Signals by addition of phase conjugate replicas

In a dense signals environment, probability of simultaneous signals interacting increases. This increases the probability of deep penetration by amplified phase conjugate reflections of weak internally generated signals from the interior, impinging upon the pumped (stressed) nonlinear surface areas where the simultaneous exterior interactions occur.

* Phase conjugate replica

Next Slide

Return to Index

Previous Slide

ANOMALOUS DEPTH PENETRATION

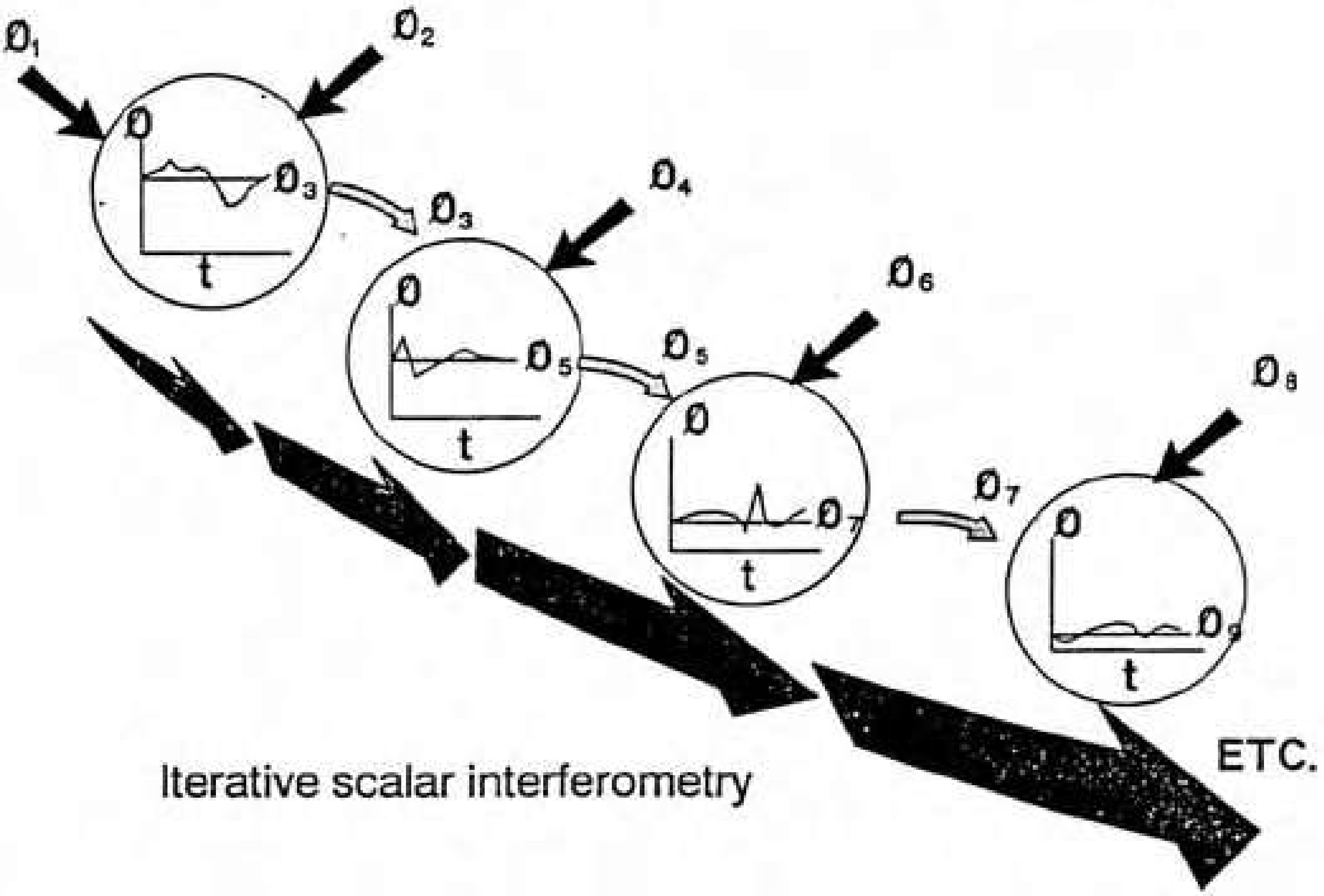
Ø1
Ø2
Ø
Ø3
t
Ø3
Ø4
Ø
Ø5
t
Ø5
Ø6
Ø
Ø7
t
Ø7
Ø8
Ø
Ø9
t
ETC.

Iterative scalar interferometry

Next Slide

Return to Index

Previous Slide

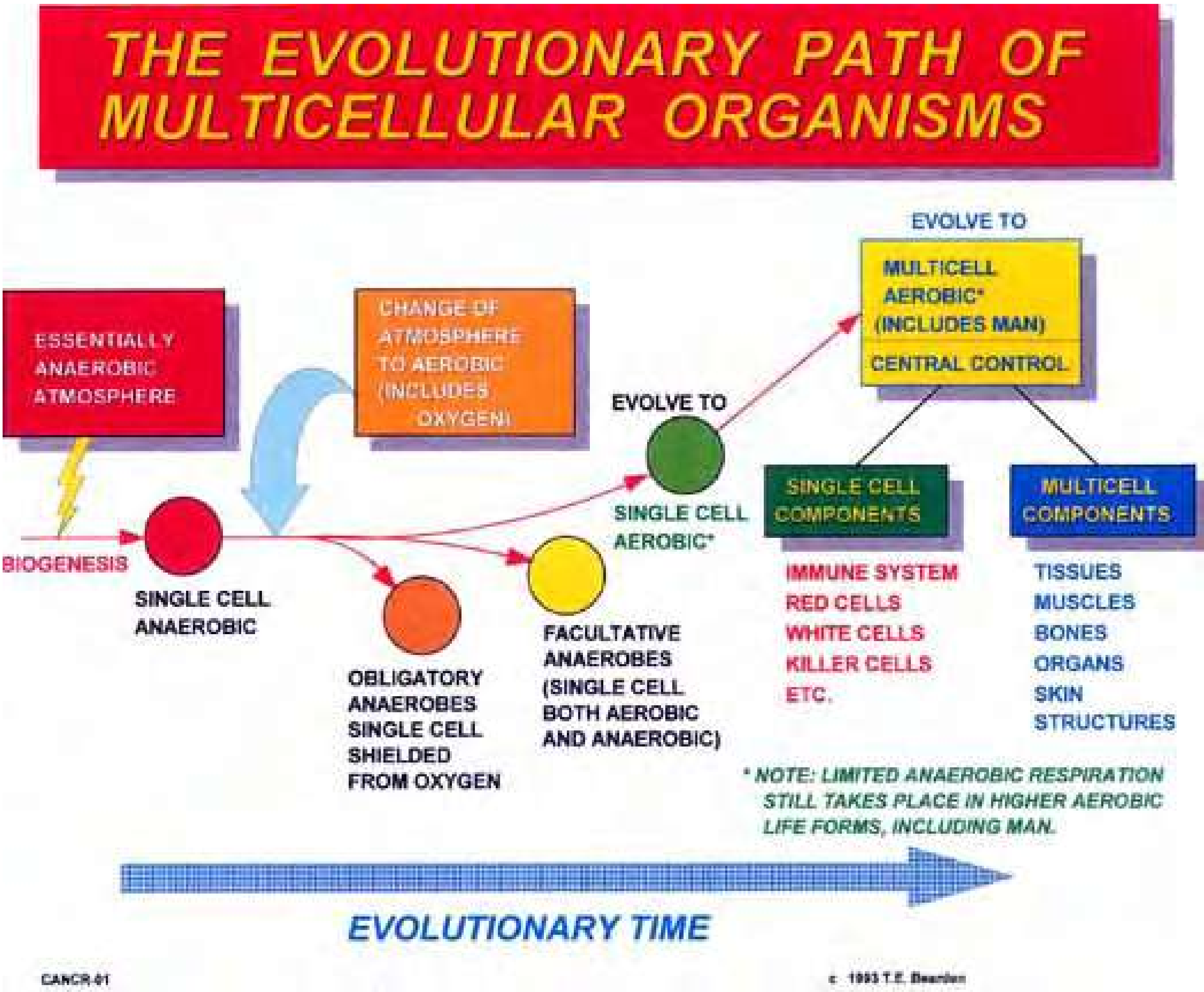

THE EVOLUTIONARY PATH OF MULTICELLULAR ORGANISMS

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

CANCR-01

c 1993 T.E. Bearden

Next Slide

Return to Index

Previous Slide

DEVELOPMENT OF MULTI-CELLED ORGANISMS

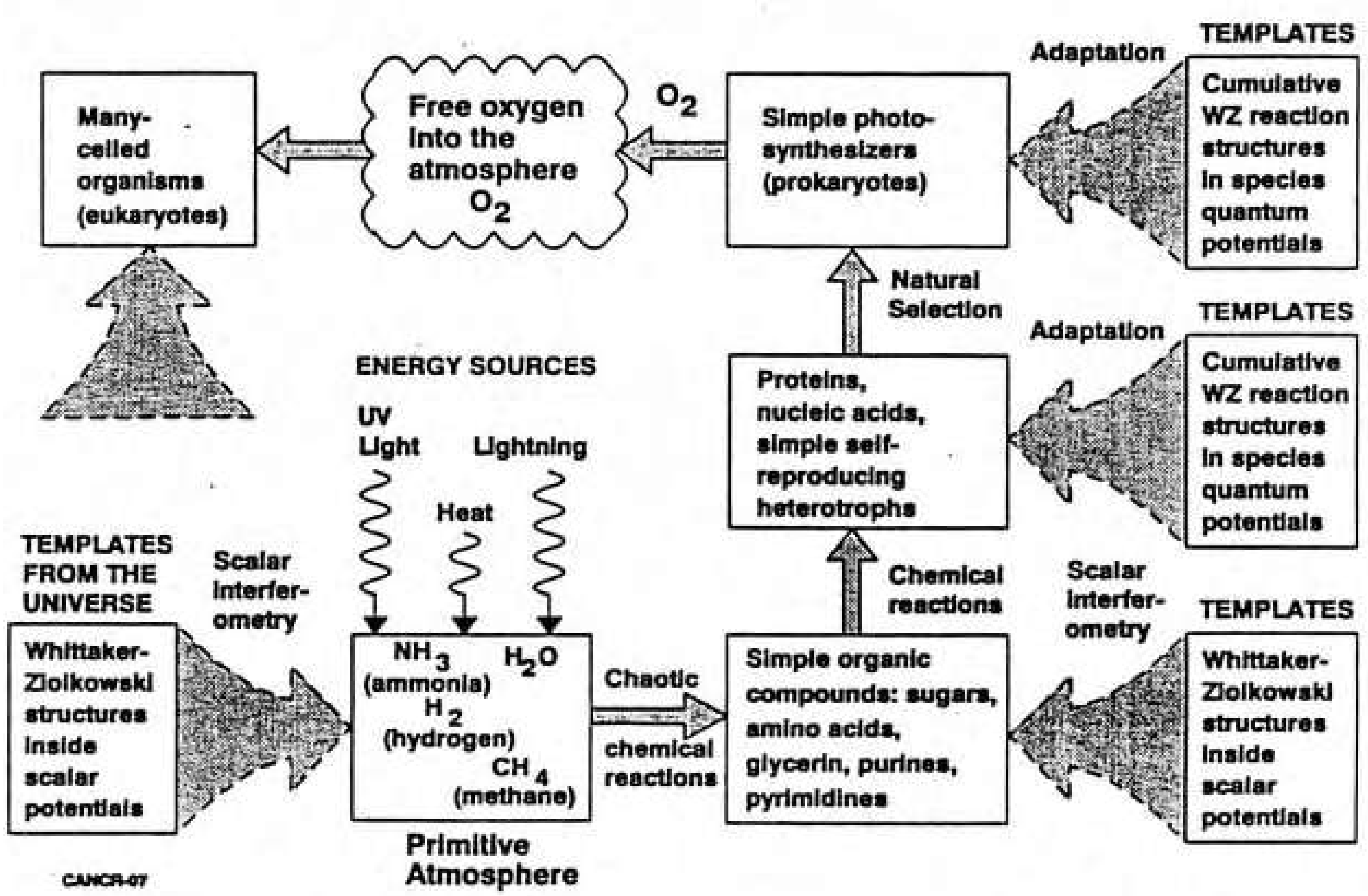

TEMPLATES

Cumulative WZ reaction structures in species quantum potentials

Adaptation

Simple photo-synthesizers (prokaryotes)

O_2

Free oxygen into the atmosphere O_2

Many-celled organisms (eukaryotes)

Natural Selection

TEMPLATES

Cumulative WZ reaction structures in species quantum potentials

Adaptation

Proteins, nucleic acids, simple self-reproducing heterotrophs

Chemical reactions

ENERGY SOURCES

UV Light

Heat

Lightning

TEMPLATES FROM THE UNIVERSE

Whittaker-Ziolkowski structures inside scalar potentials

Scalar interfer-ometry

NH_3 (ammonia) H_2O H_2 (hydrogen) CH_4 (methane)

Primitive Atmosphere

Chaotic chemical reactions

Simple organic compounds: sugars, amino acids, glycerin, purines, pyrimidines

Scalar interfer-ometry

TEMPLATES

Whittaker-Ziolkowski structures inside scalar potentials

CANCR-07

Next Slide

Return to Index

Previous Slide

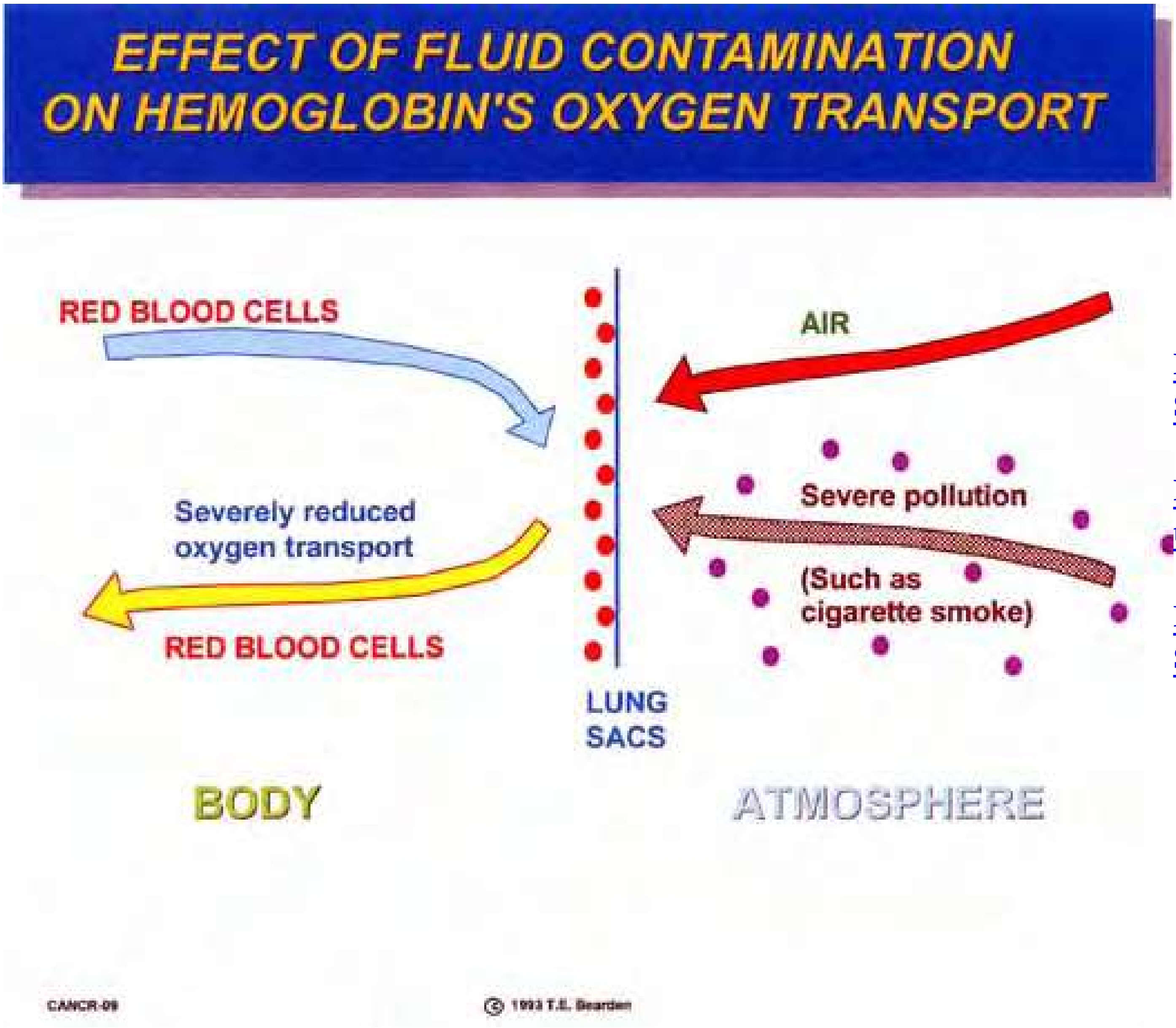

EFFECT OF FLUID CONTAMINATION ON HEMOGLOBIN'S OXYGEN TRANSPORT

RED BLOOD CELLS

AIR

Severe pollution

(Such as cigarette smoke)

Severely reduced oxygen transport

RED BLOOD CELLS

LUNG SACS

BODY

ATMOSPHERE

CANCR-09

© 1993 T.E. Bearden

[Next Slide](#)

[Return to Index](#)

[Previous Slide](#)

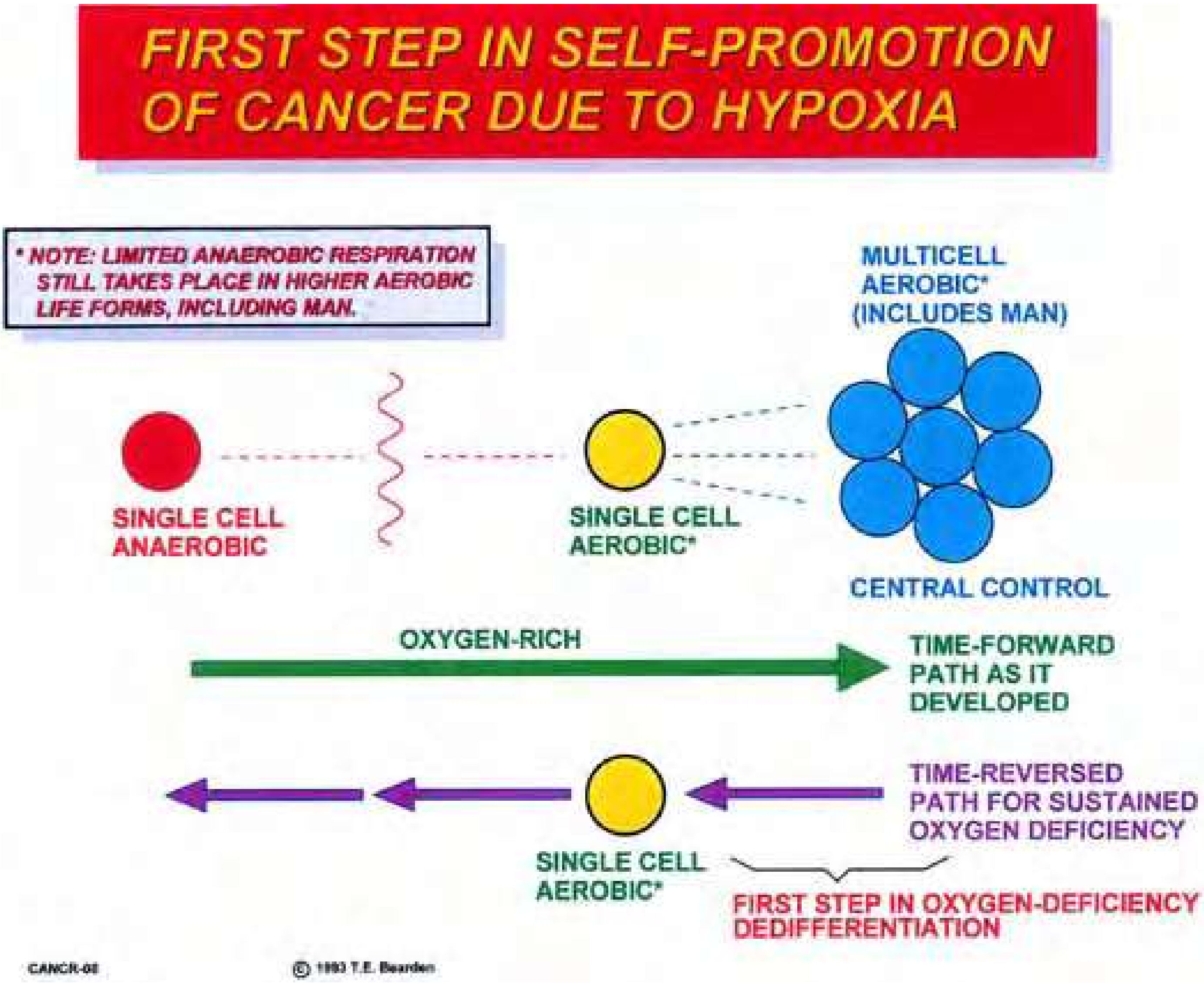

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

MULTICELL
AEROBIC*
(INCLUDES MAN)

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

SINGLE CELL
AEROBIC*

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

CANCR-08

© 1993 T.E. Bearden

Next Slide

Return to Index

Previous Slide

CUMULATIVE GROWTH OF DEDIFFERENTIATION SIGNALS

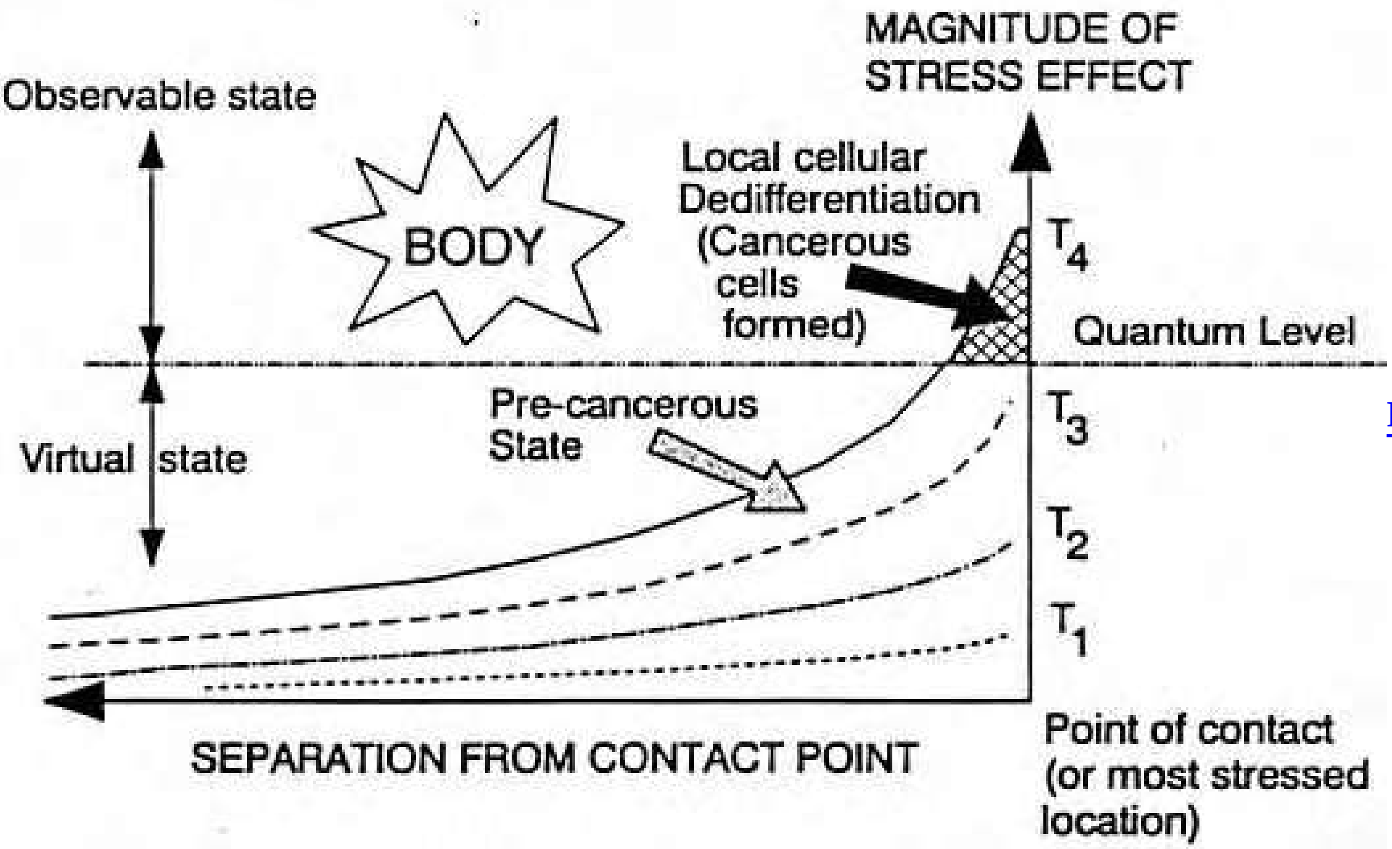
MAGNITUDE OF STRESS EFFECT
Observable state
BODY
Local cellular Dedifferentiation (Cancerous cells formed)
T_4
Quantum Level
Pre-cancerous State
Virtual state
T_3
T_2
T_1
SEPARATION FROM CONTACT POINT
Point of contact (or most stressed location)

Next Slide

Return to Index

Previous Slide

CUMULATIVE EMERGENCE OF CANCER INTO THE PHYSICAL STATE

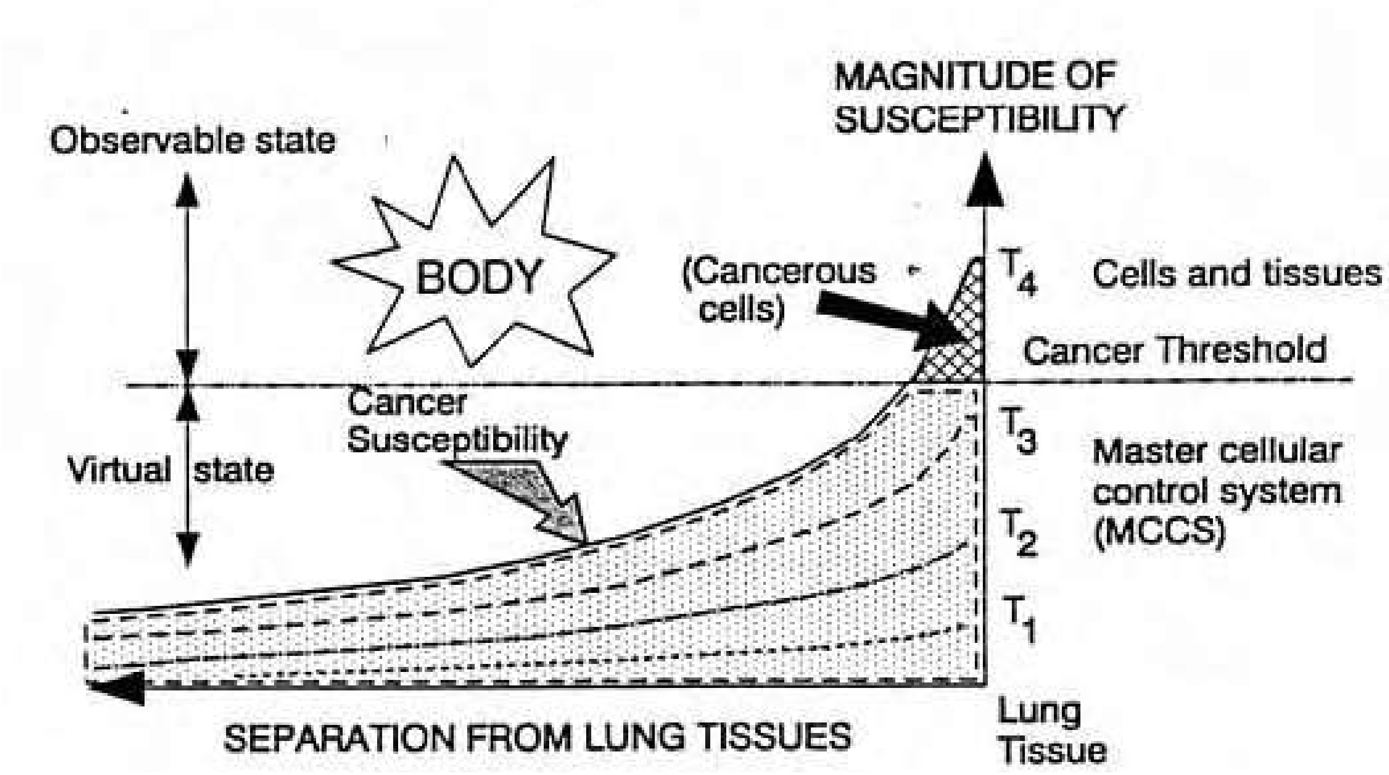
MAGNITUDE OF SUSCEPTIBILITY
Observable state
BODY
(Cancerous cells)
T_4
Cells and tissues
Cancer Threshold
Cancer Susceptibility
Virtual state
T_3
Master cellular control system (MCCS)
T_2
T_1
SEPARATION FROM LUNG TISSUES
Lung Tissue

Next Slide

Return to Index

Previous Slide

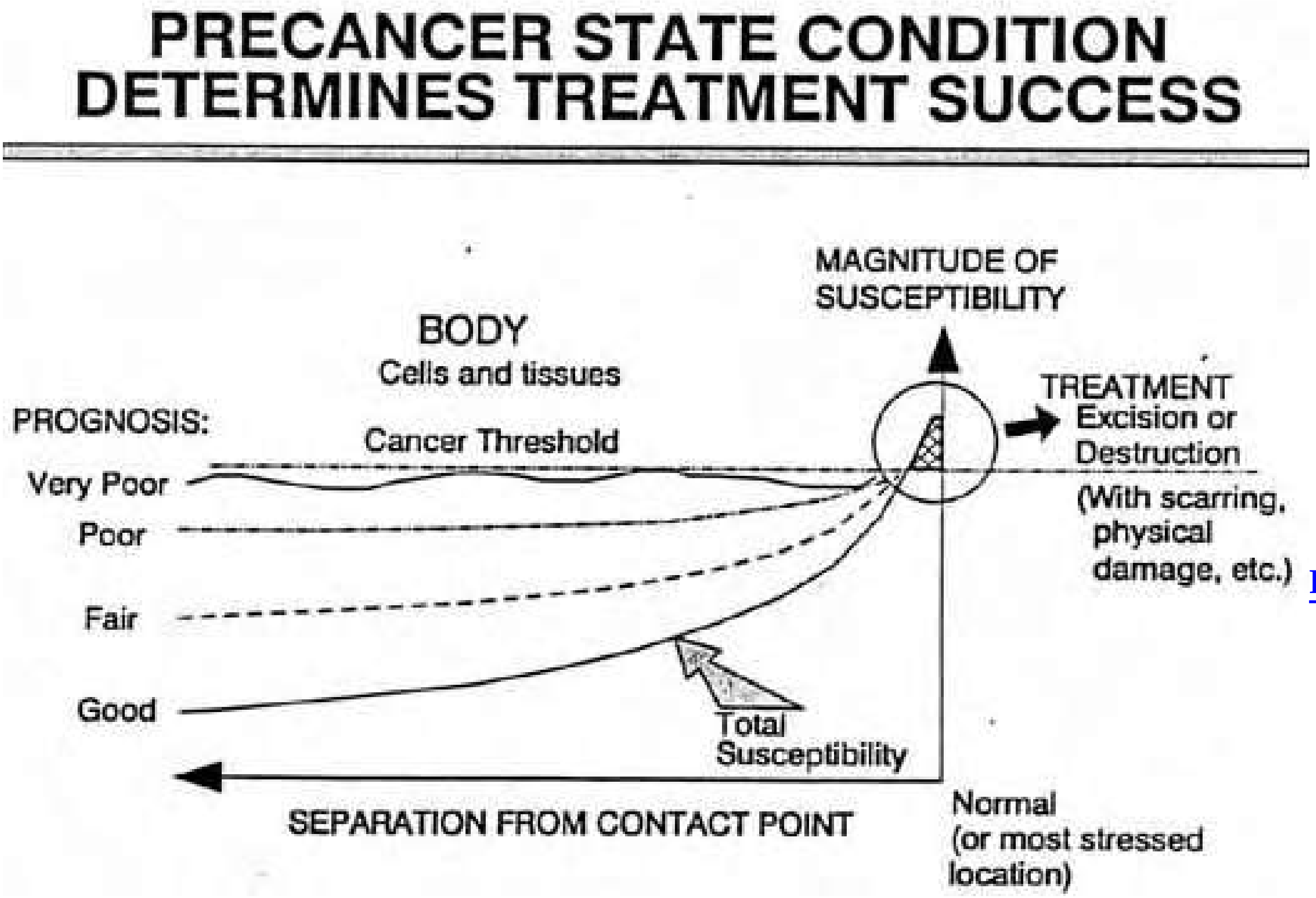

PRECANCER STATE CONDITION DETERMINES TREATMENT SUCCESS

MAGNITUDE OF SUSCEPTIBILITY
BODY
Cells and tissues
PROGNOSIS:
Cancer Threshold
TREATMENT
Excision or Destruction
(With scarring, physical damage, etc.)
Very Poor
Poor
Fair
Good
Total Susceptibility
SEPARATION FROM CONTACT POINT
Normal
(or most stressed location)

Next Slide

Return to Index

Previous Slide

Starting From the Scalar EM View:
A NEW DEFINITION OF CANCER

- CANCER AND LEUKEMIA ARE:

> **Centrally-commanded, final, desperate, "first-step dediffernetiation" adaptive attempts by the stressed affected cells experiencing sustained oxygen shortage, to reverse their cellular evolution and return to the anaerobic stage of their distant ancestry.**

- THE CAUSE IS LONG-TERM CUMULATIVE, AND ELECTROMAGNETIC IN NATURE.
- A TOTAL CURE CAN BE ACHIEVED ELECTRO-
MAGNETICALLY, AS WAS DEMONSTRATED.

Next Slide

Return to Index

Previous Slide

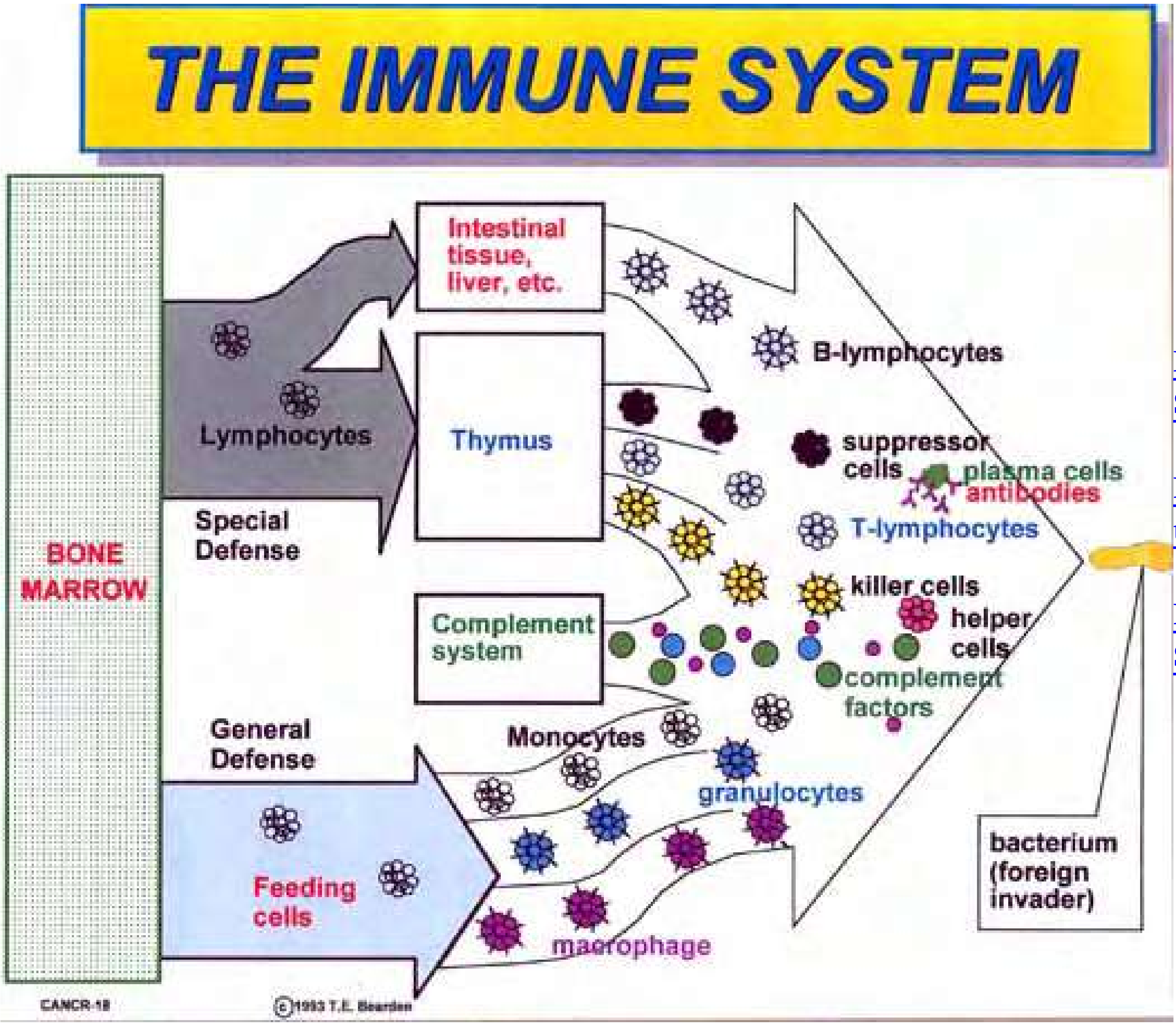

THE IMMUNE SYSTEM

BONE MARROW
Lymphocytes
Special Defense
Intestinal tissue, liver, etc.
Thymus
Complement system
General Defense
Feeding cells
Monocytes
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
granulocytes
macrophage
bacterium (foreign invader)
CANCR-18
©1993 T.E. Bearden

Next Slide

Return to Index

Previous Slide

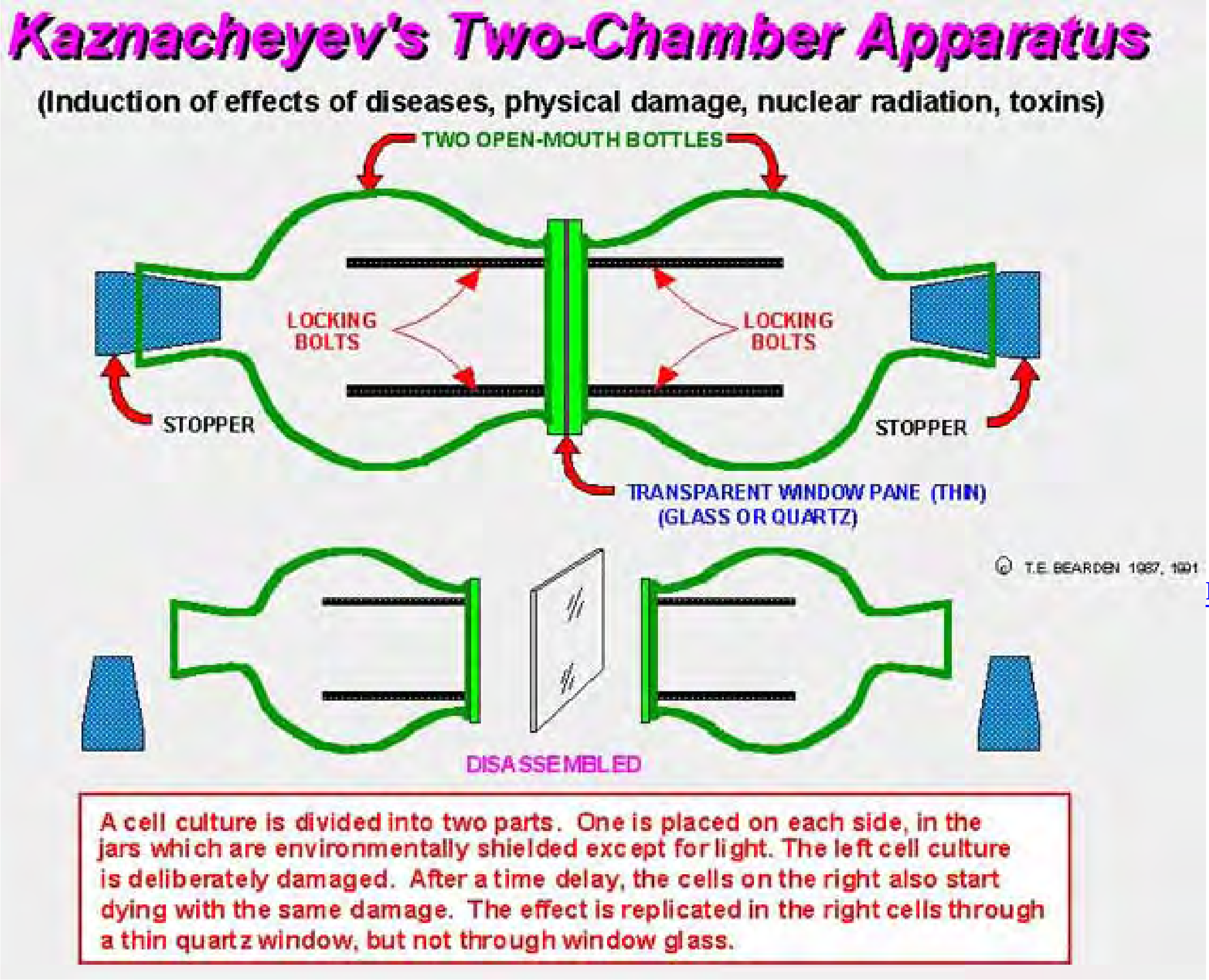

Kaznacheyev's Two-Chamber Apparatus

(Induction of effects of diseases, physical damage, nuclear radiation, toxins)

TWO OPEN-MOUTH BOTTLES
LOCKING BOLTS
LOCKING BOLTS
STOPPER
STOPPER
TRANSPARENT WINDOW PANE (THIN)
(GLASS OR QUARTZ)
DISASSEMBLED

© T.E. BEARDEN 1987, 1991

A cell culture is divided into two parts. One is placed on each side, in the jars which are environmentally shielded except for light. The left cell culture is deliberately damaged. After a time delay, the cells on the right also start dying with the same damage. The effect is replicated in the right cells through a thin quartz window, but not through window glass.

Next Slide

Return to Index

Previous Slide

Kaznacheyev's Cytopathogenic Mirror Effect

Disease vacuum engines transported by photons and waves

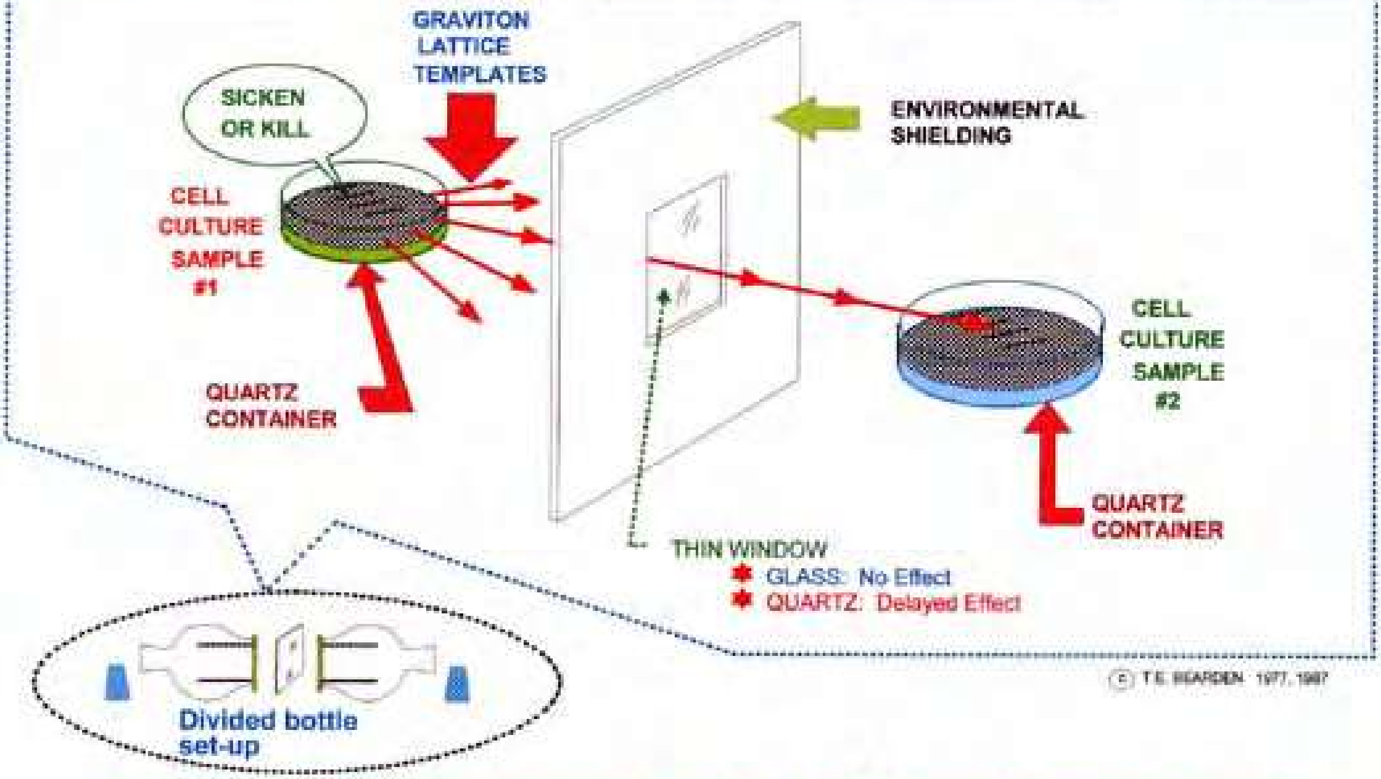
GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
CELL
CULTURE
SAMPLE
#1
ENVIRONMENTAL
SHIELDING
QUARTZ
CONTAINER
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle
set-up

© T.E. BEARDEN 1977, 1997

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Distant induction of cellular death and diseases, by special EM means. Coupled photon-antiphoton pairs (gravitons) in a structured harmonic lattice ordering constitute the disease or disorder template.

Next Slide

Return to Index

Previous Slide

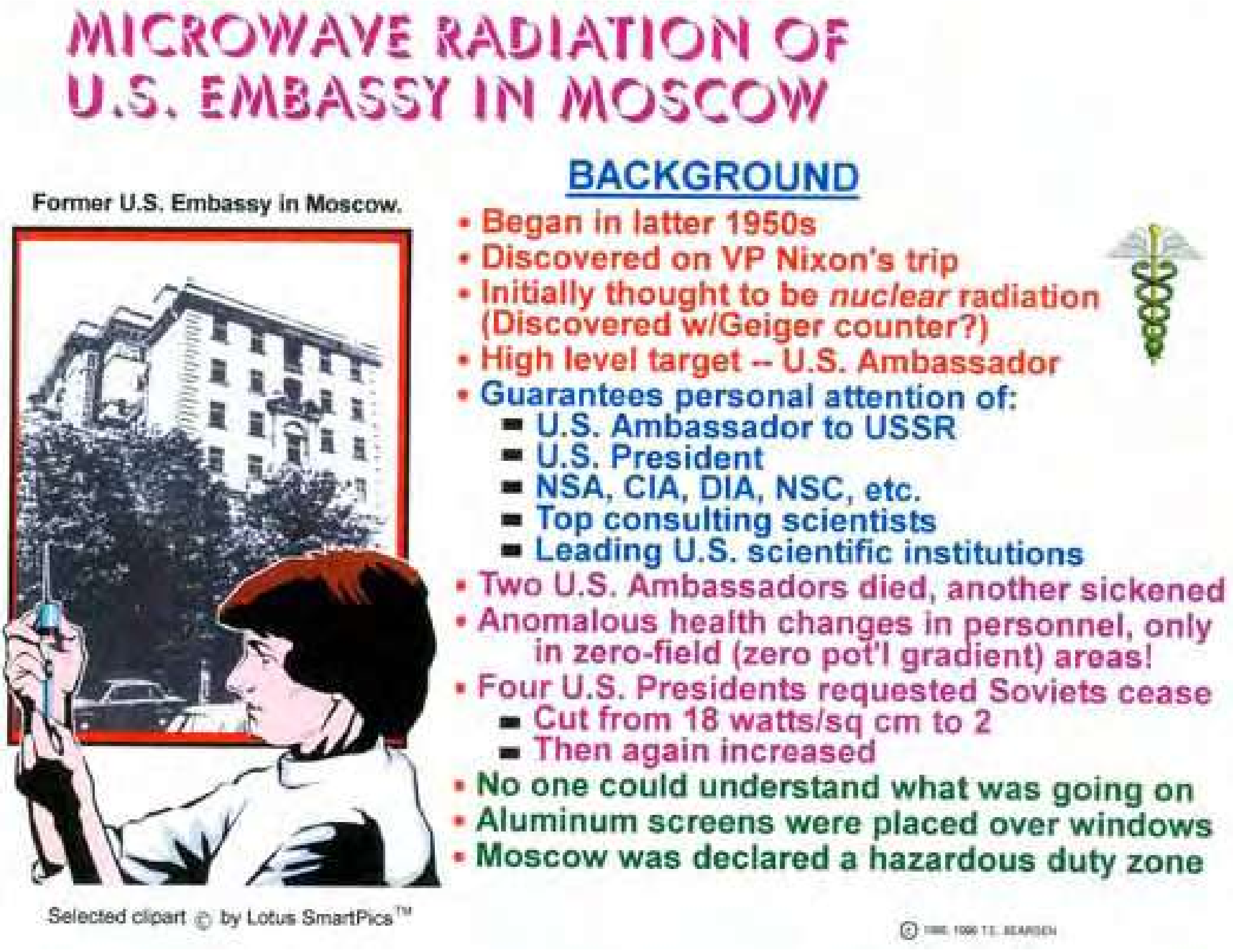

MICROWAVE RADIATION OF U.S. EMBASSY IN MOSCOW

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Former U.S. Embassy in Moscow.

Selected clipart © by Lotus SmartPics™

© [illegible] 1996 T.E. BEARDEN

Next Slide

Return to Index

Previous Slide

JOHN HOPKINS STUDY OF EMBASSY MICROWAVE RADIATION

- **EXCELLENT SCIENTIFIC TEAM**

  - PERFORMED CLASSICAL EM INVESTIGATION

  - CAREFULLY ESTABLISHED THE FORCE FIELD PATTERNS

  - NO HEALTH CHANGES OCCURRED IN PERSONNEL WHERE THE FORCE FIELDS WERE PRESENT (NONZERO)

  - ALL HEALTH CHANGES OCCURRED IN PERSONNEL WHERE THE FORCE FIELDS WERE ZERO (ABSENT)

- **ASSUMED THAT ALL EM ACTION IS DUE ONLY TO FORCE FIELDS**

  - CONCLUDED THAT MICROWAVE RADIATION COULD NOT HAVE CAUSED THE HEALTH CHANGES
- **CONCLUSION WAS IN SERIOUS ERROR**
  - CORRECT CONCLUSION: HEALTH CHANGES WERE TOTALLY CORRELATED TO ABSENCE OF FORCE FIELDS

  *** TOTALLY CORRELATED TO PRESENCE OF GRADIENT-FREE POTENTIALS, OR NOT TO RADIATION AT ALL**

  *** IF NOT TO RADIATION AT ALL, THEN HEALTH CHANGES WOULD HAVE ALSO OCCURRED WHERE FORCE FIELDS WERE PRESENT**

- **THEREFORE THE HEALTH CHANGES WERE TOTALLY CORRELATED TO THE GRADIENT-FREE POTENTIALS, AND TO THE EM RADIATION**

Next Slide

Return to Index

Previous Slide

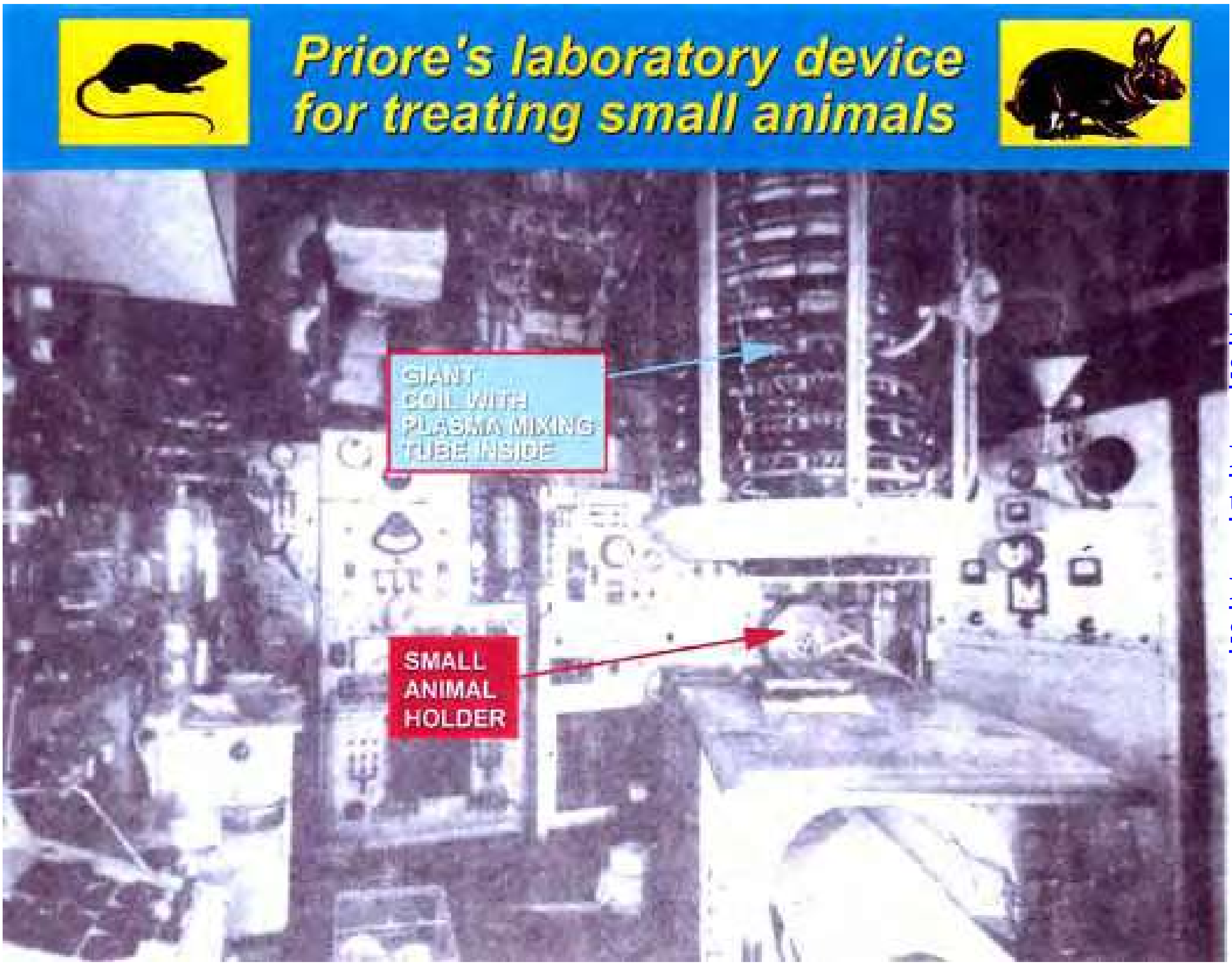

Priore's laboratory device for treating small animals

GIANT COIL WITH PLASMA MIXING TUBE INSIDE
SMALL ANIMAL HOLDER

Next Slide

Return to Index

Previous Slide

MECHANISM OF THE PRIORE THERAPY

- **PRODUCED A DETERMINISTIC WZ STRUCTURED SCALAR POTENTIAL**
  - **17 FREQUENCIES MIXED IN A ROTATING PLASMA**
  - **PHASE CONJUGATES ADDED BY THE PLASMA**
  - **MODULATED ONTO A RIPPLING MAGNETIC FIELD WHICH CARRIED IT TO ATOMIC NUCLEI**
  - **INTERACTED DIRECTLY WITH PERSONAL QUANTUM POTENTIAL**
- **NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CONTROL**
- **SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE**
- **RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING**
- **NO EXCESSIVE TRAUMA TO TREATED ANIMAL**

Next Slide

Return to Index

Previous Slide

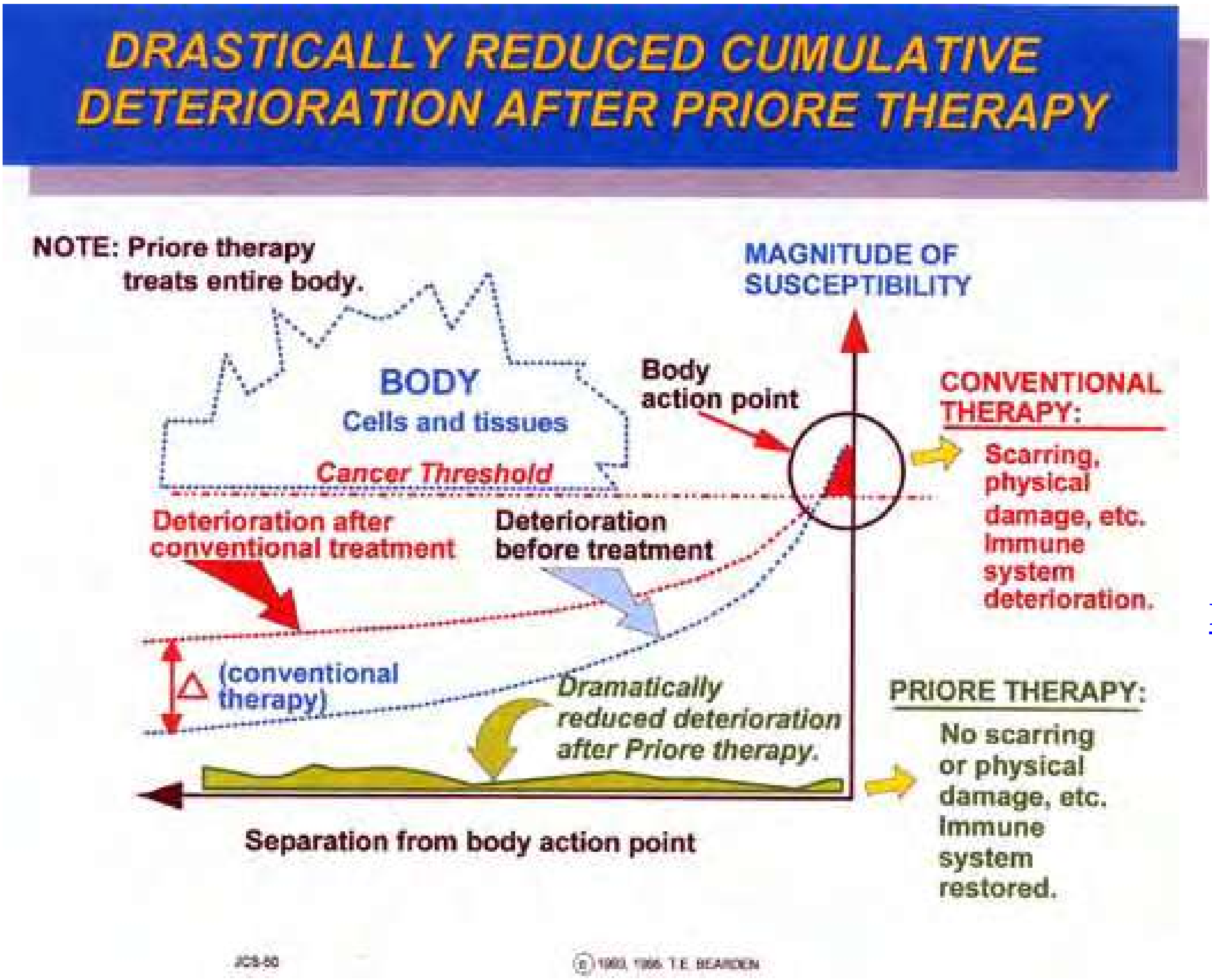

DRASTICALLY REDUCED CUMULATIVE DETERIORATION AFTER PRIORE THERAPY

NOTE: Priore therapy treats entire body.

BODY

Cells and tissues

Cancer Threshold

Body action point

MAGNITUDE OF SUSCEPTIBILITY

Deterioration after conventional treatment

Deterioration before treatment

Δ (conventional therapy)

Dramatically reduced deterioration after Priore therapy.

Separation from body action point

CONVENTIONAL THERAPY:

Scarring, physical damage, etc. Immune system deterioration.

PRIORE THERAPY:

No scarring or physical damage, etc. Immune system restored.

JCS-50

© 1993, 1996. T.E. BEARDEN

[Next Slide]

[Return to Index]

[Previous Slide]

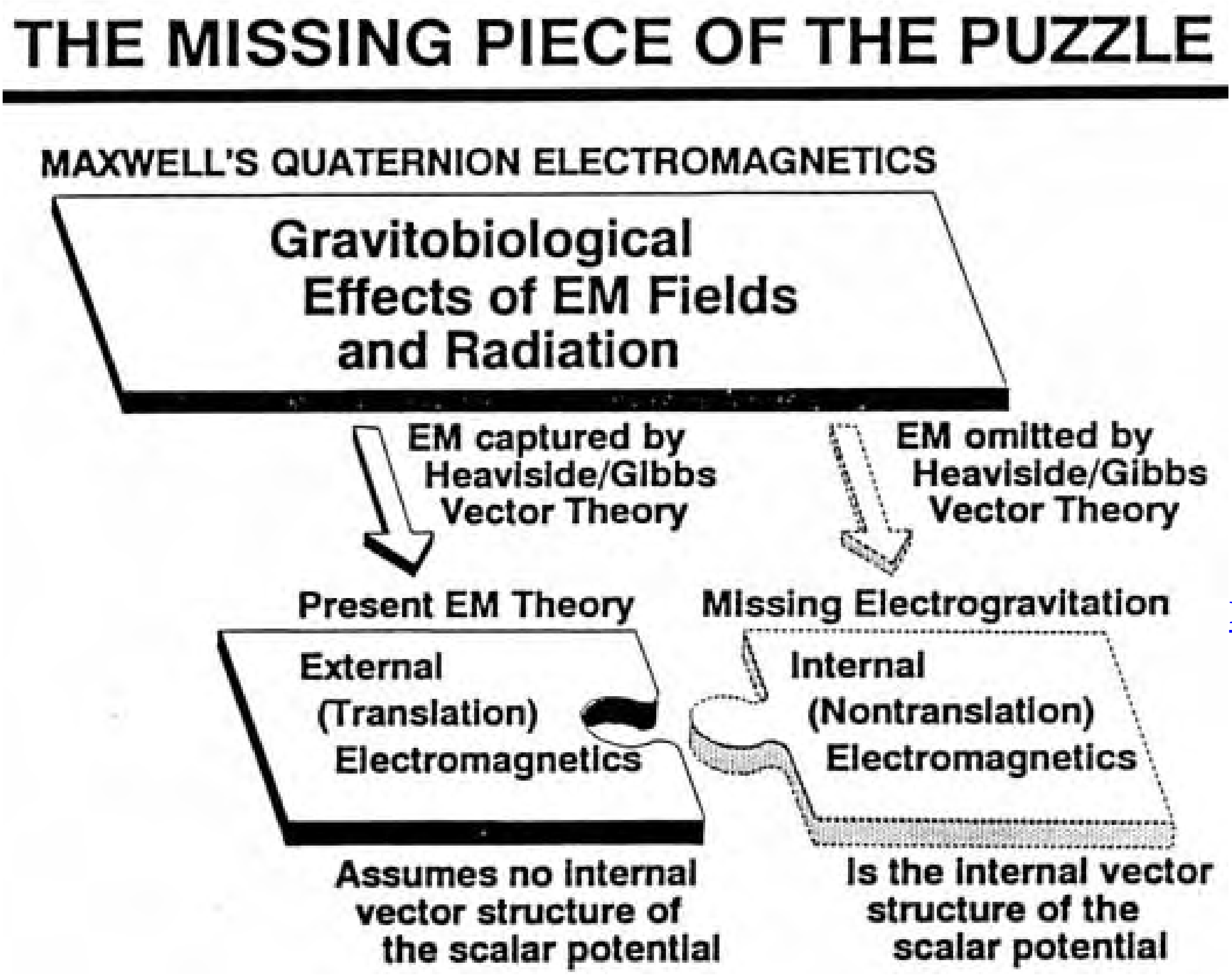

THE MISSING PIECE OF THE PUZZLE

MAXWELL'S QUATERNION ELECTROMAGNETICS

Gravitobiological Effects of EM Fields and Radiation

EM captured by Heaviside/Gibbs Vector Theory

EM omitted by Heaviside/Gibbs Vector Theory

Present EM Theory

External (Translation) Electromagnetics

Assumes no internal vector structure of the scalar potential

Missing Electrogravitation

Internal (Nontranslation) Electromagnetics

Is the internal vector structure of the scalar potential

Next Slide

Return to Index

Previous Slide

SUMMARY AND CONCLUSION

- **RECOMMEND SERIOUS EXAMINATION OF THE NEW APPROACH**
- **RECOMMEND REVIVAL OF PRIORE-TYPE RESEARCH ON**
  AN URGENT BASIS
- **CAN BE EXTENDED TO OTHER DISEASES**

  - AIDS
  - ARTERIOSCLEROSIS
  - SLEEPING SICKNESS
  - ETC.
- **A NEW THERAPEUTIC PARADIGM OF GREATLY INCREASED**
  SCOPE AND EFFECTIVENESS MAY BE POSSIBLE
- **IN ADDITION, DRASTIC REDUCTION IN MEDICAL COSTS**
  COULD POTENTIALLY RESULT

Return to Index

Previous Slide

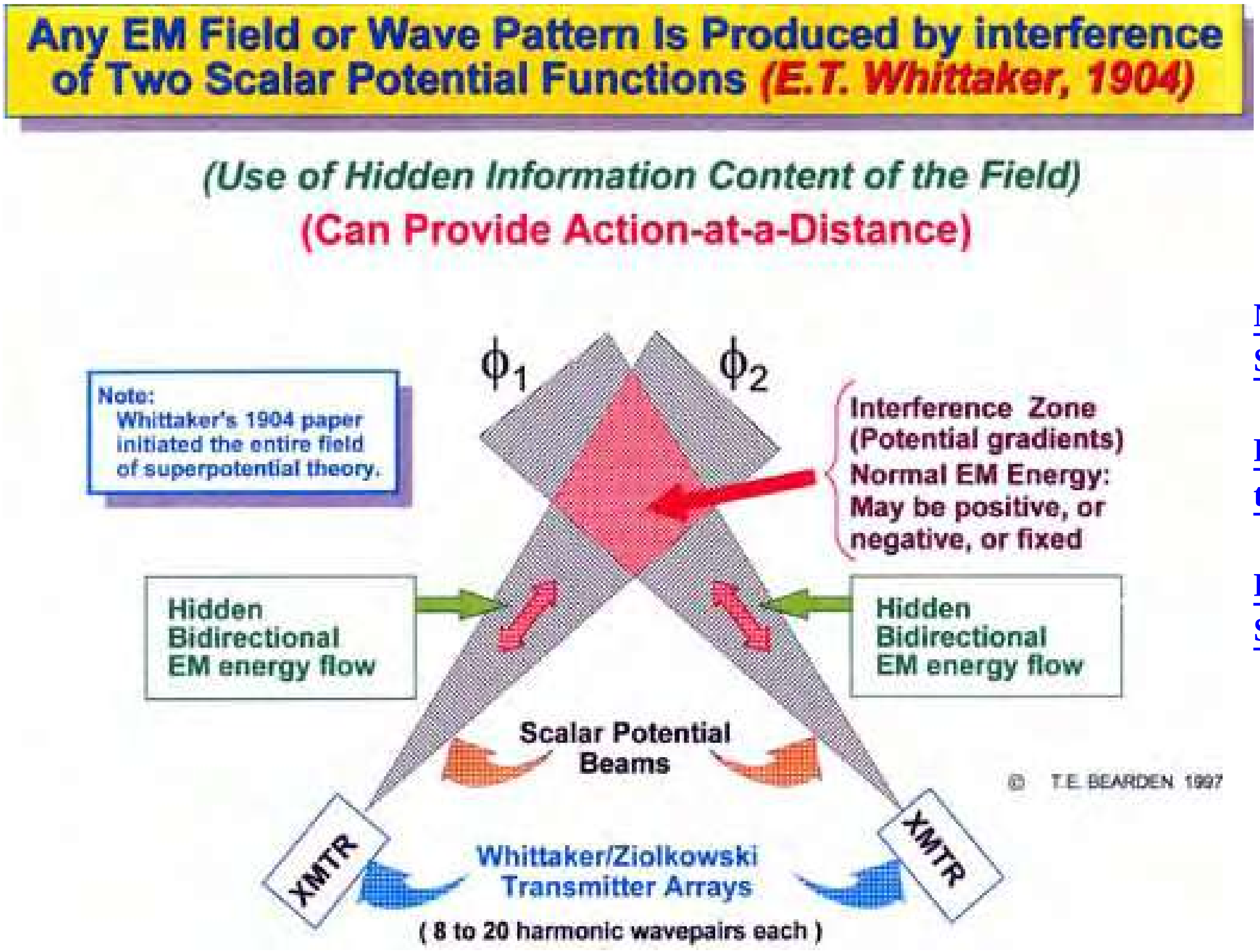

Any EM Field or Wave Pattern Is Produced by interference of Two Scalar Potential Functions *(E.T. Whittaker, 1904)*

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.

Interference Zone (Potential gradients)
Normal EM Energy:
May be positive, or negative, or fixed

Hidden Bidirectional EM energy flow

Hidden Bidirectional EM energy flow

Scalar Potential Beams

Whittaker/Ziolkowski Transmitter Arrays

(8 to 20 harmonic wavepairs each)

© T.E. BEARDEN 1997

Next Slide

Return to Index

Previous Slide

MECHANISM FOR LONG-TERM, CUMULATIVE BIOLOGICAL EFFECTS OF EM FIELDS AND RADIATION

T. E. BEARDEN*

* A.D.A.S.
POB 1472
HUNTSVILLE, AL 35807

ENERGY FROM THE VACUUM
www.cheniere.org

Support Free Energy and purchase the bumper sticker

The New Tesla Electromagnetics and the Secrets of Electrical Free Energy

By T.E. Bearden
© 1984

Table of Contents

Foreword

Comments on the New Tesla Electromagnetics:
Part I: Discrepancies in the Present EM Theory
by T. E. Bearden

Comments on the New Tesla Electromagnetics:
Part II: The Secret of Electrical Free Energy
by T. E. Bearden

Overture to a New Age Technology
by Dr. Rolf Schaffranke

REVIEW AND OUTLOOK - The Development of Post-Relativistic Concepts in Physics and Advanced Technology Abroad
by Dr. Rolf Schaffrancke

COMMENTS ON THE NEW TESLA ELECTROMAGNETICS

Part I: Discrepancies in Present EM Theory

© T. E. Bearden 1982

There are at least twenty-two major discrepancies presently existing in conventional electromagnetics theory. This paper presents a summary of those flaws, and is a further commentary on my discussion of scalar longitudinal Tesla waves in a previous paper, "Solutions to Tesla's Secrets and the Soviet Tesla Weapons," Tesla Book Company, 1981 and 1982.

I particularly wish to express my deep appreciation to two of my friends and colleagues who at this time, I believe, wish to remain anonymous. One of the two is an experimental genius who can produce items on the bench that do not work by orthodox theory. The second is a master of materials science and electromagnetics theory. I thank them both for their exceptional contributions and stimuli regarding potential shortcomings in present electromagnetics theory, and their forbearance with the many discussions we have held on this and related subjects.

It goes without saying that any errors in this paper are strictly my own, and not the fault of either of my distinguished colleagues.

(1) In present electromagnetics theory, charge and charged mass are falsely made identical. Actually, on a charged particle, the "charge" is the flux of virtual particles on the "bare particle" of observable mass. The charged particle is thus a "system" of true massless charge coupled to a bare chargeless mass. The observable "mass" is static, three-dimensional, and totally spatial. "Charge" is dynamic, four-dimensional or more, virtual and spatiotemporal. Further, the charge and observable mass can be decoupled, contrary to present theory. Decoupled charge -- that is, the absence of mass -- is simply what we presently refer to as "vacuum." Vacuum, spacetime, and massless charge are all identical. Rigorously, we should utilize any of these three as an "ether," as suggested for vacuum by Einstein himself (see

Max Born, Einstein's Theory of Relativity, Revised Edition, Dover Publications, New York, 1965, p. 224). And all three of them are identically anenergy -- not energy, but more fundamental components of energy.

(2) Electrostatic potential is regarded as a purely 3-dimensional spatial stress. Instead, it is the intensity of a many-dimensional (at least four-dimensional) virtual flux and a stress on all four dimensions of spacetime. This is easily seen, once one recognizes that spacetime is identically massless charge. (It is not "filled" with charge; rather, it is charge!) Just as, in a gas under pressure, the accumulation of additional gas further stresses the gas, the accumulation of charge (spacetime) stresses charge (spacetime). Further, if freed from its attachment to mass , charge can flow exclusively in time, exclusively in space, or in any combination of the two. Tesla waves -- which are scalar waves in pure massless charge flux itself -- thus can exhibit extraordinary characteristics that ordinary vector waves do not possess. And Tesla waves have extra dimensional degrees of freedom in which to move, as compared to vector waves. Indeed, one way to visualize a Tesla scalar wave is to regard it as a pure oscillation of time itself.

(3) Voltage and potential are often confused in the electrostatic case, or at least thought of as "composed of the same thing." For that reason, voltage is regarded as "potential drop". This also is not true. Rigorously, the potential is the intensity of the virtual particle flux at a single point -- whether or not there is any mass at the point -- and both the pressure and the point itself are spatiotemporal (4-dimensional), not spatial (3-dimensional) as presently assumed. Voltage represents the spatial intersection of the difference in the potential between two separated spatial points, and always implies at least a miniscule flow of mass current (that is what makes it spatial!). "Voltage" is spatial and depends upon the presence of observable mass flow, while scalar electrostatic potential is spatiotemporal and depends upon the absence of observable mass flow. The two are not even of the same dimensionality.

(4) The charge of vacuum spacetime is assumed to be zero, when in fact it is a very high value. Vacuum has no mass, but it has great massless charge and virtual particle charge flux. For proof that a charged vacuum is the seat of something in motion, see G.M. Graham and D.G. Lahoz, "Observation of static electromagnetic angular momentum in vacuo," Nature, Vol. 285, 15 .May 1980, pp. 154-155. In fact, vacuum IS charge, identically, and it is also "spacetime" and at least four-dimensional.

(5) Contrary to its present usage, zero is dimensional and relative in its context. A three-dimensional spatial hole, for example, exists in time. If we model time as a dimension, then the spatial hole has one dimension in 4-space. So a spatial absence is a spatiotemporal presence. In the vacuum 4-space, a spatial nothing is still a something. The "virtual" concept and the mathematical concept of a derivative are simply two present ways of unconsciously addressing this fundamental problem of the dimensional relativity of zero.

(6) The concepts of "space" and "time" imply that spacetime (vacuum) has been separated into two parts. We can only think of a space as "continuing to exist in time " To separate vacuum spacetime into two pieces, an operation is continually required. The operator that accomplishes this splitting operation is the photon interaction, the interaction of vector electromagnetic energy or waves with mass. I have already strongly pointed out this effect and presented a "raindrop model" of first-order physical change itself in my book, The Excalibur Briefing, Strawberry Hill Press, San Francisco, 1980, pp. 128-130.

(7) "Vector magnetic potential" is assumed to be always an aspect of (and connected to) the magnetic field. In fact it is a separate, fundamental field of nature and it can be entirely disconnected from the magnetic field. See Richard P. Feynman et al, The Feynman Lectures on Physics, Addison-Wesley Publishing Co., New York, 1964, Vol. II, pp. 15-8 to 15-14. Curiously, this fact has been proven for years' yet it has been almost completely ignored in the West. The "Vx" operator, when applied to the A-field, makes B-field. If the Vx operator is not applied, the "freed" A-field possesses much-expanded characteristics from those presently allowed in the "bound" theory. Specifically, it becomes a scalar or "shadow vector" field; it is not a normal vector field. (note: for V read inverted capital Delta)

(8) The speed of light in vacuum is assumed to be a fundamental constant of nature. Instead it is a function of the intensity of the massless charge flux (that is, of the magnitude of the electrostatic potential) of the vacuum in which it moves. (Indeed, since vacuum and massless charge are one and the same, one may say that the speed of light is a function of the intensity of the spatiotemporal vacuum!). The higher the flux intensity (charge) of the vacuum, the faster the speed of light in it. This is an observed fact and already shown by hardcore measurements. For example, distinct differences actually exist in the speed of light in vacuo, when measured on the surface of the earth as compared to measurements in space away from planetary masses. In a vacuum on the surface of the earth, light moves significantly faster. For a discussion and the statistics, see B. N. Belyaev, "On Random Fluctuations of the Velocity of Light in Vacuum," Soviet Physics

Journal, No. 11, Nov. 1980, pp. 37-42 (original in Russian; translation by Plenum Publishing Corporation.) The Russians have used this knowledge for over two decades in their strategic psychotronics (energetics) program; yet hardly a single U.S. scientist is aware of the measured variation of c in vacuo. In fact, most Western scientists simply cannot believe it when it is pointed out to them!

(9) Energy is considered fundamental and equivalent to work. In fact, energy arises from vector processes, and it can be disassembled into more fundamental (anenergy) scalar components, since the vectors can. These scalar components individually can be moved to a distant location without expending work, since one is not moving force vectors. There the scalar components can be joined and reassembled into vectors to provide "free energy" appearing at a distance, with no loss in between the initial and distant points. For proof that a vector field can be replaced by (and considered to be composed of) two scalar fields, see E. T. Whittaker, Proceedings of the London Mathematical Society, Volume 1, 1903, p. 367. By extension, any vector wave can be replaced by two coupled scalar waves.

(10) The classical Poynting vector predicts no longitudinal wave of energy from a time-varying, electrically charged source, In fact, an exact solution of the problem does allow this longitudinal wave. See T. D. Keech and J. F. Corum, "A New Derivation for the Field of a Time-Varying Charge in Einstein's Theory," International Journal of Theoretical Physics, Vol. 20, No, 1, 1981, pp. 63-68 for the proof.

(11) The present concepts of vector and scalar are severely limited, and do not permit the explicit consideration of the internal, finer-grained structures of a vector or a scalar. That is, a fundamental problem exists with the basic assumptions in the vector mathematics itself. The "space" of a vector field, for example, does not have inter-nested sublevels (subspaces) containing finer "shadow vectors" or "virtual vectors," Yet particle physics has already discovered that electrical reality is built that way. Thus one should actually use a "hypernumber" theory after the manner of Charles Musés. A scalar is filled with (and composed of) nested levels of other "spaces" containing vectors, where these sum to "zero" in the ordinary observable frame without an observable vector resultant. In Musés' mathematics, for example, zero has real roots. Real physical devices can be -- and have been -constructed in accordance with Muses' theory. For an introduction to Musés' profound hypernumbers approach, see Charles Musés' foreword to Jerome Rothstein, Communication Organization and Science, The Falcon's Wing Press, Indian Hills, Colorado, 1958. See also Charles Musés, "Applied Hypernumbers: Computational Concepts," Applied Mathematics and

Computation, Vol. 3, 1976. See also Charles Musés, "Hypernumbers II," Applied Mathematics and Computation, January 1978.

(12) With the expanded Tesla electromagnetics, a new conservation of energy law is required. Let us recapitulate for a moment. The oldest law called for the conservation of mass. The present law calls for the conservation of "mass and energy", but not each separately. If mass is regarded as simply another aspect of energy, then the present law calls for the conservation of energy. However, this assumes that energy is a basic, fundamental concept. Since the energy concept is tied to work and the movement of vector forces, it implicitly assumes "vector movement" to be a "most fundamental" and irreducible concept. But as we pointed out, Whittaker showed that vectors can always be further broken down into more fundamental coupled scalar components. Further, Tesla discovered that these "coupled components" of "energy" can be individually separated, transmitted, processed, rejoined, etc. This directly implies that energy per se need not be conserved. The new law therefore calls for the conservation of anenergy, the components of energy. These components may be coupled into energy, and the energy may be further compacted into mass. It is the sum total of the (anenergy) components -- coupled and uncoupled -- that is conserved, not the matter or the energy per se. Further, this conservation of anenergy is not spatial; rather, it is spatiotemporal in a spacetime of at least four or more dimensions.

(13) Relativity is presently regarded as a theory or statement about fundamental physical reality. In fact, it is only a statement about FIRST ORDER reality -- the reality that emerges from the vector interaction of electromagnetic energy with matter. When we break down the vectors into scalars (shadow vectors or hypervectors), we immediately enter a vastly different, far more fundamental reality. In this reality superluminal velocity, multiple universes, travel back and forth in time, higher dimensions, variation of all "fundamental constants" of nature, materialization and dematerialization, and violation of the "conservation of energy" are all involved. Even our present Aristotlean logic -- fitted to the photon interaction by vector light as the fundamental observation mechanism -- is incapable of describing or modeling this more fundamental reality. Using scalar waves and scalar interactions as much subtler, far less limited observation/detection mechanisms, we must have a new "superrelativity" to describe the expanded electromagnetic reality uncovered by Nikola Tesla.

(14) "Charge" is assumed to be quantized, in addition to always occurring with -- and locked to -- mass. Indeed, charge is not necessarily quantized just as it is not necessarily locked to mass.

Ehrenhaft discovered and reported fractional charges for years, in the 30's and 40's, and was ignored. See P.A.11. Dirac, "Development of the Physicist's Conception of Nature," Symposium on the Development of the Physicist's Conception of Nature, ed. Jagdish Merha, D. Reidel, Boston, 1973, pp. 12-14 for a presentation of some of Ehrenhaft's results. Within the last few years Stanford University· researchers also have positively demonstrated the existence of "fractional charge." For a layman's description of their work, see "A Spector Haunting Physics," Science News, Vol. 119, January 31, 1981, pp. 68-69. Indeed, Dirac in his referenced article points out that Millikan himself -- in his original oildrop experiments -- reported one measurement of fractional charge, but discounted it as probably due to error.

(15) Presently, things are always regarded as traveling through normal space. Thus we use or model only the most elementary type of motion -- that performed by vector electromagnetic energy. We do not allow for things to "travel inside the vector flow itself." Yet, actually, there is a second, more subtle flow inside the first, and a third, even more subtle flow inside the second, and so on. We may operate inside, onto, into, and out of energy itself -- and any anenergy component of energy. There are hypervectors and hyperscalars unlimited, within the ordinary vectors and scalars we already know. Further, these "internal flows" can be engineered and utilized, allowing physical reality itself to be directly engineered, almost without limits.

(16) We always assume everything exists in time. Actually, nothing presently measured exists in time, because the physical detection/ measurement process of our present instruments destroys time, ripping it off and tossing it away -- and thereby "collapsing the wave function." Present scientific methodology thus is seriously flawed. It does not yield fundamental (spacetime) truth, but only a partial (spatial) truth. This in turn leads to great scientific oversights. For example, mass does not exist in time, but mass x time (masstime) does. A fundamental constant does not exist in time, but the "constant x time" does. Energy does not exist in time, but energy x time (action) does. Even space itself does not exist in time -- spacetime does. We are almost always one dimension short in every observable we model. Yet we persist in thinking spatially, and we have developed 1nstruments that detect and measure spatially only. Such instruments can never measure and detect the phenomenology of the nested substrata of time. By using scalar technology, however, less limited instruments can indeed be constructed -and they have been. With such new instruments, the phenomenology of the new electromagnetics can be explored and an engineering technology developed.

(17) We do not recognize the connection between nested levels of virtual state (particle physics) and orthogonally rotated frames (hyperspaces). Actually the two are identical, as I showed in the appendix to my book, The Excalibur Briefing, Strawberry Hill Press, San Francisco, 1980, pp. 233-235. A virtual particle in the laboratory frame is an observable particle in a hyperspatial frame rotated more than one orthogonal turn away. This of course implies that the hyperspatial velocity of all virtual particles is greater than the speed of light. The particle physicist is already deeply involved in hyperspaces and hyperspatial charge fluxes without realizing it. In other words, he is using tachyons (particles that move faster than light) without realizing it.

(18) Presently quantum mechanics rigorously states that time is not an observable, and therefore it cannot be measured or detected. According to this assumption, one must always infer time from spatial measurements, because all detections and measurements are spatial. With this assumption, our scientists prejudice themselves against even looking for finer, subquantal measurement methodologies and instrumentation. Actually this present limitation is a result of the type of electromagnetics we presently know, where all instruments (the "measurers") have been interacted with by vector electromagnetic energy (light). Every mass that has temperature (and all masses do!) is continually absorbing and emitting photons, and in the process they are continually connecting to time and disconnecting from time. If time is continually being carried away from the detector itself by its emitted photons, then the detector cannot hold and "detect" that which it has just lost. With Tesla electromagnetics, however, the fundamental limitation of our present instruments need not apply. With finer instruments, we can show there are an infinite number of levels to "time", and it is only the "quantum level time" which is continually being lost by vector light (photon) interaction. By using subquantal scalar waves, instruments can move to deeper levels of time -- in which case the upper levels of time ARE measureable and detectable, in contradistinction to the present assumptions.

(19) In the present physics, time is modeled as, and considered to be, a continuous dimension such as length. This is only a gross approximation. Indeed, time is not like continuous "dimension," but more like a series of "stitches," each of which is individually made and then ripped out before the next stitch appears. "Vector light" photons interact one at a time, and it is this interaction with mass that creates quantum change itself. The absorption of a photon -- which is energy x time -- by a spatial mass converts it to masstime: the time was added by the photon. The emission of a photon tears away the time, leaving behind again a spatial mass. It is not accidental, then, that time flows at

the speed of light, for it is light which contains and carries time. It is also not accidental that the photon IS the individual quantum. Since all our instruments presently are continually absorbing and emitting photons, they are all "quantized," and they accordingly "quantize" their detections. This is true because all detection is totally internal to the detector, and the instruments detect only their own internal changes. Since these detections are on a totally granular quantized background, the detections themselves are quantized. The Minkowski model is fundamentally erroneous in its modeling of time, and for that reason relativity and quantum mechanics continue to resist all attempts to successfully combine them, quantum field theory notwithstanding.

(20) Presently, gravitational field and electrical field are considered mutually exclusive. Actually this is also untrue. In 1974, for example, Santilly proved that electrical field and gravitational field indeed are not mutually exclusive In that case, one is left with two possibilities: (a) they are totally the same thing, or (b) they are partially the same thing. For the proof, see R. M. Santilli, "Partons and Gravitation: Some Puzzling Questions," Annals of Physics, Vol. 83, No. 1, March 1974. With the new Tesla electromagnetics, pure scalar waves in time itself can be produced electrically, and electrostatics (when the charge has been separated from the mass) becomes a "magic" tool capable of directly affecting and altering anything that exists in time--including gravitational field. Antigravity and the inertial drive are immediate and direct consequences of the new electromagnetics.

(21) Presently, mind is considered metaphysical, not a part of physics, an not affected by physical means. Literally, the prevailing belief of Western scientists is that man is a mechanical robot -- even though relativity depends entirely upon the idea of the "observer." Western science today thus has essentially become dogmatic, and in this respect borders on a religion. Since this "religion," so to speak, is now fairly well entrenched in its power in the state, Western science is turning itself into an oligarchy. But mind occupies time, and when we measure and affect time we can directly measure and affect mind itself. In the new electromagnetics, then, man regains his dignity and his humanity by restoring the reality of mind and thought to science. In my book, The Excalibur Briefing, I have already pointed out the reality of mind and a simplified way in which it can be modeled to the first order. With scalar wave instruments, the reality of mind and thought can be measured in the laboratory, and parapsychology becomes a working, engineering, scientific discipline.

(22) Multiple-valued basic dimensional functions are either not permitted or severely discouraged in the present theory For one thing, integrals of multiple valued derivative functions have the annoying

habit of "blowing up" and yielding erroneous answers, or none at all. And we certainly do not allow multiple types of time! This leads to the absurdity of the present interpretation of relativity which permits only a single observer (and a single observation) at a time. So if one believes as "absurd" a thing as the fact that more than one person can observe an apple at the same time, the present physics fails. However, the acceptance of such a simple proposition as multiple simultaneous observation leads to a physics so bizarre and incredible that most Western physicists have been unable to tolerate it, much less examine its consequences. In the physics that emerges from multiple simultaneous observation, all possibilities are real and physical. There are an infinite number of worlds, orthogonal to one another, and each world is continually splitting into additional such "worlds" at a stupendous rate. Nonetheless, this physics was worked out by Everett for his doctoral thesis in 1956, and the thesis was published in 1957. (See Hugh Everett, III, The Many-Worlds Interpretation of Quantum Mechanics: A Fundamental Exposition, with papers by J. A. Wheeler, B.S. DeWitt, L. N. Cooper and D. Van Vechten, and N. Graham; eds. Bryce S. Dewitt and Neill Graham, Princeton Series in Physics, Princeton University Press, 1973.) Even though it is bizarre, Everett's physics is entirely consistent with all the present experimental basis of physics. The present electromagnetic theory is constructed for only a single "world" or universe -- or "level." The expanded theory, on the other hand, contains multiply nested levels of virtual state charge -- and these levels are identically the same as orthogonal universes, or "hyperframes." Multiple kinds -- and values -- of time also exist. The new concept differs from Everett's, however, in that the orthogonal universes intercommunicate in the virtual state. That is, an observable in one universe is always a virtual quantity in each of the other universes. Thus one can have multi-level "continuities" and "discontinuities" simultaneously, without logical conflict. It is precisely these levels of charge -- these levels of scalar vacuum -- that lace together the discontinuous quanta generated by the interaction of vector light with mass.

However, to understand the new electromagnetic reality, one requires a new, expanded logic which contains the old Aristotlean logic as a subset. I have already pointed out the new logic in my paper, "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," 1979, available from the National Technical Information Center, AD-A071032.

Even as logic is extended, quantum mechanics, quantum electrodynamics, and relativity are drastically changed by the Tesla electromagnetics, as I pointed out in my paper, "Solutions to Tesla's Secrets and the Soviet Tesla Weapons," Tesla Book Company, 1580

Magnolia, Millbrae, California, 94030, 1980.

The present electromagnetics is just a special case of a much more fundamental electromagnetics discovered by Nikola Tesla, just as Newtonian physics is a special case of relativistic physics. But in the electromagnetics case, the differences between the old and the new are far more drastic and profound.

ADDITIONAL REFERENCES

1, Boren, Dr. Lawrence Milton, "Discovery of the Fundamental Magnetic Charge (Arising from the new Conservation of Magnetic Energy)," 1981/82 (private communication). Dr. Boren has a cogent argument that the positron is the fundamental unit of magnetic charge. His theory thus assigns fundamentally different natures to positive charge and negative charge. In support of Dr. Boren, one should point out that the "positive" end of circuits can simply be "less negative" than the "negative" end. In other words, the circuit works simply from higher accumulation of negative charges (the "negative" end) to a lesser accumulation of negative charges (the "positive" end). Nowhere need there be positive charges (protons, positrons, etc.) to make the circuit work. Dr. Boren's theory, though dramatic at first encounter, nonetheless bears close and meticulous examination -- particularly since he has been able to gather experimental data which support his theory and disagree with present theory.

2. Eagle, Albert, "An Alternative Explanation of Relativity Phenomena," Philosophical Magazine and Journal of Science, No. 191, December 1939, pp. 694-701.

3. Ehrenhaft, Felix and Wasser, Emanuel, "Determination of the Size and Weight of Single Submicroscopic Spheres of the Order of Magnitude r = 4 x 10(-5) cm. to 5 x 10(-6) cm., as well as the Production of Real Images of Submicroscopic Particles by means of Ultraviolet Light," Phil. Mag. and Jour. of Sci., Vol. II (Seventh Series), No. 7, July 1926, pp. 30-51.

4. Ehrenhaft, Felix and Wasser, Emanuel, "New Evidence of the Existence of Charges smaller than the Electron - (a) The Micromagnet; (b) The Law of Resistance; (c) The Computation of Errors of the Method," Phil. Mag. and Jour. of Sci., Vol. V (Seventh Series), No. 28, February 1928, pp. 225-241.

5. See also Ehrenhaft's last paper dealing with the electronic charge, in Philosophy of Science, Vol. 8, 1941, p. 403.

6. McGregor, Donald Rait, The Inertia of the Vacuum: A New Foundation for Theoretical Physics, Exposition Press, Smithtown, NY, First Edition, 1981, pp. 15-20.

7. Ignat'ev, Yu. G. and Balakin, A. B., "Nonlinear Gravitational Waves in Plasma," Soviet Physics Journal, Vol. 24, No. 7, July 1981, (U.S. Translation, Consultants Bureau, NY, January 1982), pp. 593-597.

8. Yater, Joseph C., "Relation of the second law of thermodynamics to the power conversion of energy fluctuations," Phys. Review A, Vol. 20, no. 4, October 1979, pp. 1614-1618.

9. DeSantis, Romano M. et al, "On the Analysis of Feedback Systems With a Multipower Open Loop Chain," October 1973, available through the Defense Technical Information Center (AD 773188),

10. Graneau, Peter, "Electromagnetic jet-propulsion in the direction of current flow," Nature, Vol. 295, 28 January 1982, pp, 311-312.

11. "Gravity and acceleration aren't always equivalent," New Scientist, 17 September 1981, p. 723.

12. Gonyaev, V. V , "Experimental Determination of the Free-Fall Acceleration of a Relativistic Charged Particle. II. A Cylindrical Solenoid in a Time-Independent Field of Inertial Forces," Izvestiya VUZ, Fizika, No. 7, 1979, pp. 28-32. English Translation: Soviet Physics Journal, No. 7, 1979, pp. 829-833. If one understands the new, expanded electromagnetics, this Soviet paper indicates a means of generating antigravity and pure inertial fields.

13. R. Schaffranke, "The Development of Post-Relativistic Concepts in Physics and Advanced Technology Abroad," Energy Unlimited, No. 12, Winter 1981, pp. 15-20.

14. F. K. Preikschat, A Critical Look at the Theory of Relativity, Library of Congress Catalogue No. 77-670044. Extensive compilation of measurements of the speed of light. Clearly shows the speed of light is not constant but changes, sometimes even daily.

Return to Table of Contents **Next Page**

COMMENTS ON THE NEW TESLA ELECTROMAGNETICS

Part II: The Secret of Electrical Free Energy

© 1982 T. E. Bearden

Again I wish to express my deep appreciation to two of my friends and colleagues who at this time, I believe, wish to remain anonymous. I thank them both for their exceptional contributions and stimuli, and their forbearance with the many discussions we have held on this and related subjects.

Also, it goes without saying that any errors in this paper are strictly my own, and not the fault of either of my distinguished colleagues.

Present electromagnetic theory is only a special case of the much more fundamental electromagnetism discovered by Nikola Tesla at the turn of the century.

Pure vacuum is pure charge flux, without mass. The vacuum has a very high electrical potential -- something on the order of 200 million volts, with respect to a hypothetical zero charge.

Thus in an ordinary electrical circuit, each point of the "ground" -- which has the same potential as the vacuum -- actually has a non-zero absolute potential. This circuit ground has a value of zero only with respect to something else which has the same absolute electrical potential.

Voltage, which is always associated with a flow of electrical "mass" current (even if only a miniscule flow), is, by definition, a difference dropped in potential when a charge mass moves between two spatially separated points. What we have termed "electrical current" only flows where there is a suitable conducting medium between things which have a difference in absolute potential. Furthermore, between any two points in any material, there is considered to exist a finite resistance -- if we apply a voltage and have a mass current flowing between the two points! The simple statement that V = IR irrevocably ties together voltage drop, mass current, and resistance between two points. Rigorously, to have one of the three is to have them all. To lose one is to lose all three. Immediately we see a major error in present theory: One can have a "difference in scalar potential" between two points without having a "voltage drop" between them. Specifically, if no mass current flows between them, no resistance exists between them, and no voltage drop exists between them.

In the same fashion, one can have a "scalar wave" through the

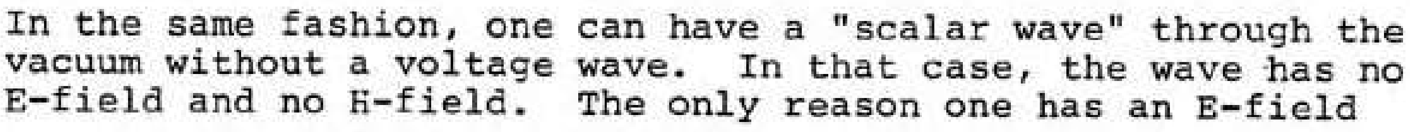

In the same fashion, one can have a "scalar wave" through the vacuum without a voltage wave. In that case, the wave has no E-field and no H-field. The only reason one has an E-field

Return to Table of Contents Next Page

around a statically charged object is because the charged electrons accumulated on the object are actually in violent motion. It is this motion of the charged masses that produces E-field -- as well as H-field whenever that entire E-field ensemble moves through laboratory space.

Now let us reason together in the "approximate" manner utilized in present electromagnetic theory. For example, let us examine a bird sitting on a high tension line.

The bird sits on the high tension line without a flow of mass electricity, because there is no significant difference in potential drop between the bird and the line. Specifically, between the bird's two feet -- each in contact with a different portion of the line -- there exists no potential difference. This is true even though, with respect to the vacuum, each foot is at a potential that would be "100,000 volts higher," were a mass current flowing. And it is true even though the absolute potential of each foot may be some 200.1 million "volts," were a mass current flowing.

Now an interesting thing happens to the bird when he flies through the air to light upon the high tension wire. As he flies toward the wire, he is flying through the massless electrostatic potential field of the wire, for that field extends an infinite distance away from the wire. The electrostatic potential field -- pure Ø-field -- is actually the spatiotemporal intensity of the massless charge at a point. In other words, as the bird flies to the wire, he flies into an increasing "massless charge" potential, building up to 100,000 "volts" higher than the earth. However, very little (if any) "mass flow" potential difference is experienced upon his body in approaching the wire, and so essentially no "charged mass currents" are induced in his body. Thus the little flier safely navigates into the teeth of a very high electrostatic potential, lights upon the wire, and is not "fried" in the process. When he lights on the wire, his body has reached the electrostatic potential that each foot's contact point has. Again, there is no mass current flow. But his body is immersed in an increased flux of massless charge -- which is what the electrostatic potential represents. And each "virtual particle" flow in that charge potential represents a "massless (scalar)" electrical current.

The point is, one can have any amount of massless charge flow -- "scalar" current -- without any mechanical work being done in the system. All electrical work in a circuit is done against the physical mass of the charged masses that flow. Rigorously, force is defined as the time rate of charge of momentum. Even in the relativistic case where $F = ma + v(dm/dt)$, change of momentum requires mass movement. No mechanical work, and hence no energy, is expended by massless charge flow.

That is why the vacuum massless charge -- which is composed of a very high flux of massless "particles" -- normally does no work

That is why the vacuum massless charge -- which is composed of a very high flux of massless "particles" -- normally does no work on our systems, and expends none of its very high "potential energy." It is exactly the same as the bird which flew into an

Return to Table of Contents Next Page

increasing scalar field as it approached the high tension wire -- no work was done upon the bird by the increasing scalar flux currents encountered by its body.

By existing "in the vacuum," so to speak, we (the whole earth) are as birds sitting on a high tension line! Until we create significant difference in potential, via our present electromagnetic circuits, no current can flow -- anywhere. Even if we produce potential differences, we must have a conductor and charged masses to flow, if we wish to produce mechanical work. Presently our electromagnetic theory allows us to create a difference in potential within different parts of a circuit, but only by moving and shifting charged mass. We therefore have to do work on this electrical mass in moving it around, and we only get back the work we have put into the circuit. In other words, presently all we do is "pump" electrical mass.

Now notice what would happen to the bird on the line if we substantially "pulsed" the potential on the line. Suppose we "pulsed" it such that the bird's physical system -- considered as a circuit containing a capacitance, a resistance, an inductance, and many free electrons -- became resonant to the pulsing frequency. In that case the "bird system" would resonate, and a great deal of electrical mass would surge back and forth in the body of the bird. In the bird's body, voltage would exist, charged mass current would flow, work would be done, and the bird would be electrocuted.

Also, note that, without mass movement, electromagnetic vector fields are not produced (and a portion of the difficulty lies with the actual vector mathematics itself). Scalar (nonvector) waves continually penetrate "space" where there is no mass movement. This means there can exist a "delta-Φ" without a voltage or an E-field. The present theory does not allow this, because it always uses "q" (charge) to be charged mass. Briefly, without belaboring the point, let us just say that it is the mechanical spin of the individual charged particle -- such as the electron -- which "entangles" or "knits together" or "couples" independent scalar waves into vector waves. A vector wave is simply two coupled scalar waves. The entire force field concept -- such as the E-field and the B-field -- is operationally defined in terms of the force exhibited on a test particle, or test mass. Rigorously, an E-field does not exist as a force field in vacuum, but as two coupled scalar Φ-fields "tumbling about each other." When these two coupled, tumbling fields meet a spinning electron, e.g., the force emerges on the electron mass. In short, movement of a rotating mass changes delta-Φ to "voltage", creating the V/I/R triad.

By "accumulating charged mass particles" -- such as electrons -- one certainly can increase the value of Φ, which represents the charge intensity or "scalar electrostatic potential." However, that is not the only way to increase it. Resonance and rotation of charged masses can also be appropriately employed to vary the vacuum charge potential Φ, under the proper circumstances.

of charged masses can also be appropriately employed to vary the vacuum charge potential Φ, under the proper circumstances.

Return to Table of Contents Next Page

By the correct application of rotary principles and Tesla electromagnetic theory, it is possible to oscillate -- and change -- the vacuum potential itself, in one part of an electrical system. Thus by correct procedures a part of a system can be electrically altered so that the absolute value of its "ground" (vacuum) potential differs significantly from the normal vacuum-ground potential of the remainder of the circuit. In other words, we shift the vacuum-ground potential of part of the circuit by oscillating the massless vacuum charge itself, and in so doing we gain a substantial, oscillatory electrical potential difference between that part of the circuit and the normal ground potential of the rest of the system. Then we connect the two parts of the system by means of a conductor containing a load in the middle.

In every conductor, a large number of "free electrons" are available. If we oscillate the actual vacuum charge itself, unlimited additional free electrons also become available from the Dirac sea, since partial "unstripping" of the negative energy wells (each containing an electron) occurs. This "unstripping" of the potential of the well is due to vacuum potential oscillation, which oscillates time as well as space. To a negative energy well (positive time), the oscillation of its "time stream" increases the well's negative potential during one half-cycle and decreases its potential for the other half-cycle. Thus during half the oscillatory cycle, negative energy electrons may be lifted from the Dirac sea if the oscillation is sufficiently intense.

Therefore we obtain an attendant voltage and flow of mass current through the load. This voltage and current, by the way, are essentially limitless, and are free for the asking, assuming the proper "vacuum oscillation" is initiated and maintained so that electrons are continually being lifted from the Dirac sea by the time oscillations, and fed into the circuit.

Standard electromagnetic theory assumes that the vacuum potential is zero. It does not recognize the existence of massless charge, separated from charged mass. Therefore, orthodox scientists have never looked for a way to engineer the vacuum, because they have not realized it is composed of pure massless charge. Electrical physics has almost hopelessly confused charge and charged mass, thereby eliminating scalar longitudinal electrostatic waves. Expressed in the bird/high tension line analogy, the present electromagnetic theory restricts us to walking along the high tension line, laboriously carrying small batteries and power units, unaware of the limitless, surging power beneath our very feet.

Electromagnetically, we have been rather like one of the five blind men who touched an elephant. We have only touched one small portion of the electromagnetism "elephant," yet we thought we had grasped the entire beast.

Return to Table of Contents Next Page

OVERTURE TO A NEW AGE TECHNOLOGY

From an Update by Dr. Rolf Schaffranke

> " ... as there has been a steady erosion in U. S. activity in the basic sciences, there has been an increase in the level of such activity in Western Europe and Japan. If that continues, we are in danger of losing our position of scientific leadership."
>
> Val. L. Fitch, Prof. of Physics at Princeton University, in an interview with *U.S. News & World Report*, June 21, 1982.

West Germany:

The prototype of the first motorcycle equipped with a "flux-motor" (the inventor's designation for a tachion field converter or gravity-field generator) of 2 KW output has completed a test run of approximately 20,000 km without repairs. Principle of operation: The T-converter is constantly recharging the electrical power battery of the vehicle, resulting in an operation which is noiseless, fuelless and polution-free. After completion of all scheduled tests, a public demonstration to the press is planned within a few weeks from the time of this writing. (June 17, 1982) See photo, Pg. 19. A new factory for the production of gravity-energy-field converters in the town of Oldenburg is scheduled to produce individual home heating units. The design of automobile power plants based on the same principle is also in preparation. Reported efficiency factors of several hundred percent are anticipated.

Japan:

Mazda continues research on the "magnetic Wankel engine" (45 HP) for electric automobiles, based on the principle of Kure Tekko. The gravity research of Prof. S. Seike is now sponsored by Hitachi, after millions of Japanese received updated information via popular science magazines. It is interesting to note that Saudi Arabia has decided to invest $100 million in Hitachi stocks.

France:

The "Kromrey" converter has completed tests at the Institute for Magnetostatics, University of Strasbourg. One prototype delivered 700 Watts at 600-1200 RPM, at a fairly uniform rate. However, France, as the No. 1 developer of nuclear energy in Europe, is still reluctant to concentrate her energies on a revolutionary new technology which could result in drastic curtailments of her well-developed nuclear industry and associated exports. See photos, Pg. 22.

Denmark:

Dr. Jensen, physicist at the University of Kopenhagen, reported test results of 300% efficiency with a specially designed, 5-coil transformer in connection

Return to Table of Contents

Next Page

with a tachion field converter based on the principle of the "N" machine. First reports were presented at the International Conference on Energy Technology in Hannover, Nov., 1980.

England / USA:

A "bullet train" based on the magnetic levitation system of British Prof. Eric Laithwaite is now under consideration for construction between Los Angeles and San Diego. Japan and West Germany are already operating high speed experimental trains, suspended magnetically to reduce friction, that may soon attain speeds of more than 200 miles per hour. The Laithwaite-Eastham method would use electromagnetic forcefields induced over an aluminum-topped concrete track. The project will probably be headed by a Japanese company as the prime contractor.

Return to Table of Contents Next Page

West German gravity-field unit in operation

Return to Index **Next Page**

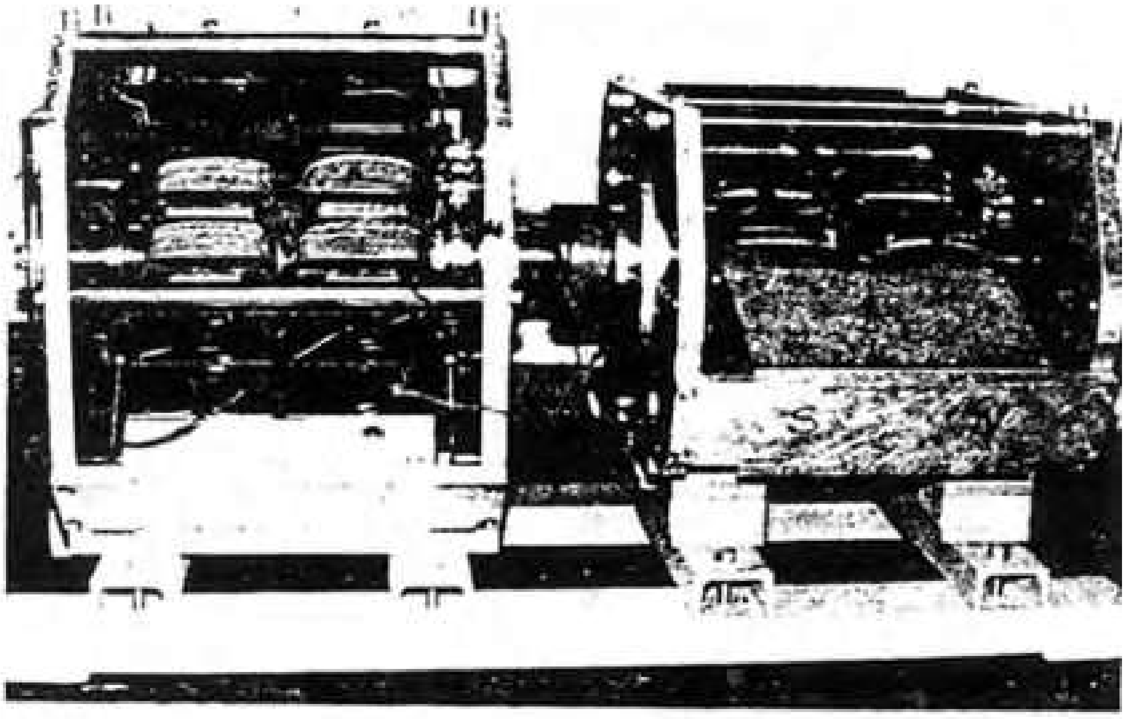

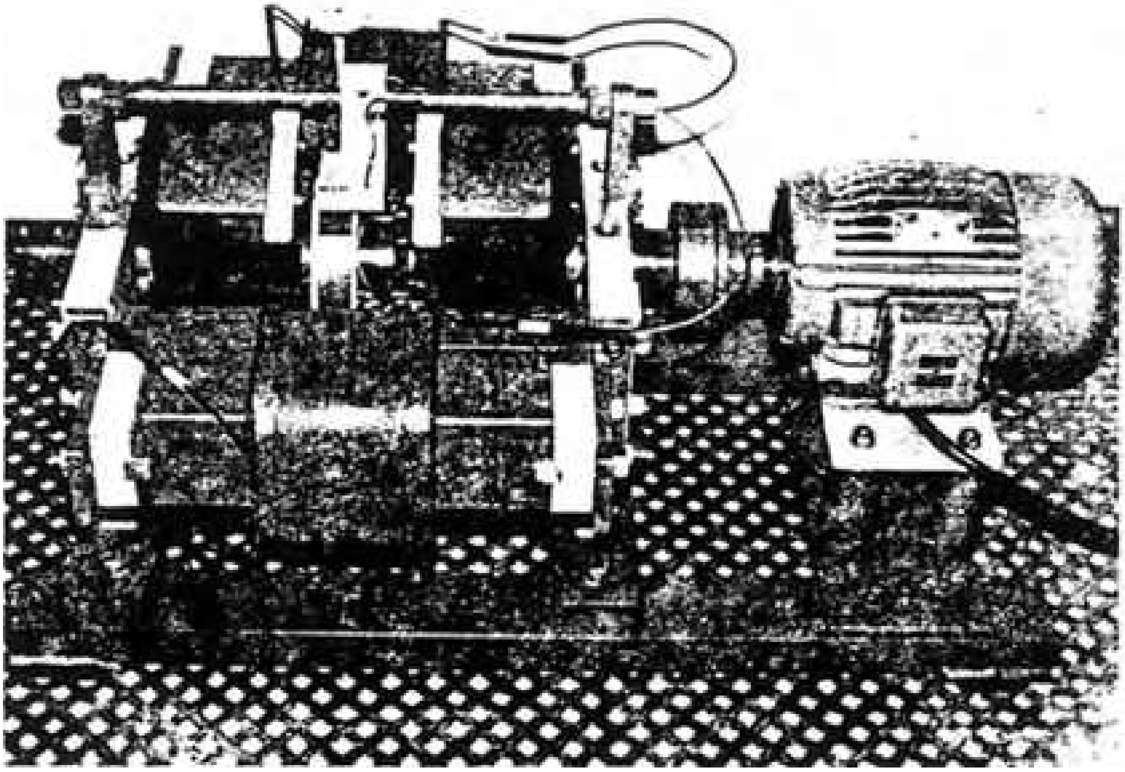

Swiss-French gravity field converter system "Kromrey"

Return to Index   **Next Page**

REVIEW AND OUTLOOK

THE DEVELOPMENT OF POST-RELATIVISTIC CONCEPTS IN PHYSICS AND ADVANCED TECHNOLOGY ABROAD

Rolf Schaffranke, Dr. h. c.
Member A.I.A.A.

This paper was presented at The First International Symposium on Non-Conventional Energy Technology, Toronto, Ontario, Canada, Oct. 23-24, 1981. (Reprinted by permission of the author.)

"Far more harm is done to the progress of science by skepticism than by gullibility". (Dr. Robert Wood, Director R & D, McDonnel-Douglas Astronautics Div., Cal.) NASA - Langley Research Center in Hampton, VA is named after Samuel Pierpont Langley (1834-1906), an American astronomer, engineer and architect.

Langley worked out the aerodynamic principles for manned flight. In principal, his calculations were correct; but the structural materials he used for wings and engines were ,insufficient. In December of 1903, the *New York Times* published an editorial complaining about his foolish dream and the associated waste of government money. The editorial predicted that: "man would not fly for a thousand years".

Only 9 days after the editorial was published, the Wright brothers made their first successful powered flight at Kitty Hawk. (Dec. 17, 1903). But, again, as late as 1905, the *Scientific American* suggested the happening was a hoax - two years after the event which changed history. Let's reflect for a moment on the fact that only in a single lifetime, man has journeyed from Kitty Hawk to the surface of the moon!

The analogy to the problem of alternative energies and their development is, of course, obvious. We try to educate the preaching orthodoxies. But orthodox solutions are no longer enough. Our system of education can only give from the past. The present must operate on inspiration and intuition, or the future will be lost. The past no longer has enough of the answers for arising crises for which there are no precedents. To keep pace with the requirements of the future, we must begin to teach not only what to learn, but how to learn, how to analyze, how to search for the truth. Only then will be able to recognize discrepancies between experiment and dogma and to cope with problems which appear to be just beyond the corner. History has shown again and again that no single individual can rise above the species without being persecuted. The reasons for that are:

Some of the most powerful forces in homo sapiens have always been
Greed, Pride, Ego, Fear and, above all, the Desire to Control Others,

.

A brand new German book titled *Energy in Abundance*, which is evidently creating a considerable impact in the European community, challenges the contemporary high

Return to Table of Contents **Next Page**

priests of science, the "Guardians of the Status Quo" and uncovers an almost incredible narrowmindedness and ignorance in science, politics and economics, "approaching the criminal" as the author puts it. The contemporary waste of genuine intellectual creativity is castigated as a cultural scandal approaching barbarism. The book stresses that a truly promising approach to the so-called energy crisis requires a fundamental and thorough re-evaluation of the theories, dogmas and axioms which form the basis of present day science and the foundation of our technology. As in the past, the formidable inertia of the establishment is not only based on inaccurate, incomplete and outdated information, but also on its reluctance to accept new ideas. The honest search for the truth first requires that we first demythologize generally accepted claims of so-called experts of the past. An excellent introduction to this problem is the article "Resistance by Scientists to Scientific Discovery" by Bernard Barber in *Science*, Vol. 134, pp. 596-602, of Sept. 1961, or the paper by Stephen C. Brush in *Science* of March, 1974, titled "Should the History of Science be Rated 'X'"?, culminating with a statement of Huxley:

> "Authorities", "disciples", and "schools" are the curse of science and do more to interfere with the work of the scientific spirit than all its enemies".

The former astronaut, Capt. Edgar D. Mitchell, one of our famous contemporaries, also stressed the need for an open mind!

> "History has shown time and again that important scientific discoveries generally happen only when someone steps outside the limits of his traditional discipline and looks at something from a fresh point of view. Then what should have been obvious all along comes into focus".

Those of us who are familiar with the pioneering efforts of Dr. Hans Nieper of Hanover will appreciate this statement by an ex-astronaut.

One of the most important cornerstones of today's scientific dogmas dates back to 1905, more than 3/4 of a century ago. Lest we forget, television and radar, jet-planes and cyclotrons, moonrockets and close-up photos of the planets of our solar system were nothing but wild fantasies of science fiction then.

The speed of light was assumed to be constant and the maximum possible speed in the universe; vacuum was to be a total void. In reality, the speed of light as measured in the Michelson-Morley experiments was not at all the same in all directions. The "ether-drift" still amounted to the respectable velocity of about five miles per second, and similar results were obtained by D. C. Miller in a series of experiments extending over twenty-five years, from 1902 to 1926.

Return to Table of Contents Next Page

> "Even worse, the measurements showed such marked discrepancies with previous results as to occasion a distress call to the U. S. Coast & Geodetic Survey, whose surveyors repeatedly remeasured the length of the tube and found no error there," reported the *Popular Science Monthly* March 1934 issue. And it continued: "More recently, speed of light observations only emphasized the apparent erratic behavior of the light beam that the scientists were attempting to plot. On some days it seems to travel faster than others by as much as 12 miles a second. Its speed seems to vary with the season, and also in a mysterious shorter cycle lasting about 2 weeks. Finally, the scientists ended by taking an average of all the readings which has just been announced as 186,271 miles per second".

In his paper "A Critical Look At the Theory of Relativity", Library of Congress Cat. No. 77-670044, F. K. Preikschat compiled all known light velocity measurements during the past 300 years or so, from Olaf Roemer in 1676 to the Laser measurements conducted by the National Bureau of Standards (USA) in 1972.

Of the 27 experiments undertaken in this field, 18 of them after the turn of this century, it must be concluded that the velocity of light as measured within the reference system of our earth has changed as much as plus/minus 50 km/sec. during the past century. Preikschat plotted a curve of these deviations and suggests a possible relationship between sunspot activities and changes of the earth magnetic field during the time period in question.

The "universal constancy of light" appears, therefore, as somewhat shaky "empirical evidence" for our cornerstone in physics, especially since Einstein himself has gone on record as saying:

> "If a single one of the conclusions drawn from it (relativity theory) proves to be wrong, it must be given up; to modify it without destroying the whole structure seems to be impossible".

Another modification concerns the definition for the term "vacuum" in physics, as documented by a paper in the *American Scientist*, March-April 1980, titled "Is The Vacuum Really Empty?" by Prof. Walter Greiner, University of Frankfurt, BRD, and Prof. Joseph H. Hamilton, Vanderbilt University, Nashville, Tenn.

The authors conclude that a neutral vacuum is by no means as "empty" as previously claimed in our textbooks, and suggests a new definition as follows:

> "The vacuum is the lowest stable state that a region of space can have while being penetrated by certain fields".

Because of the tremendous time lag in our educational system, many research

Return to Table of Contents Next Page

projects and their ensuing experimental data have been withheld from public scrutiny. The scientific community tends to have a vested interest in preserving the system it created and of which it is a part. It responds to new situations through the coloration of this attachment. A case in point are the carefully conducted experiments of T. T. Brown with charged bodies in a high vacuum, as described in my booklet *Ether Technology* (1977). These experiments suggest the actual presence of certain fields in vacuum, whether we call them gravitational field, tachion-field, ether-field, neutrino or Fermi-sea, etc. is of secondary importance at this moment. Although Brown spent, reportedly, more than $200,000 of his own funds over several decades on such experiments, he was nevertheless unable to have the results published in the scientific media of America.

Things are even worse when it comes to experiments conducted abroad, which often tend to confirm disregarded experimental results on this continent, as we shall see shortly. To highlight the wide discrepancies between orthodox (and obsolete) dogmas and actual, physical realities pertaining to the true subatomic structures as we know them to be today, let us briefly review the structure of the matter:

A molecule is the smallest division of a substance. Further division would cause it to cease being a substance. The smallest of true molecules can be illustrated when we use the globe of the earth for our standard. If a single drop of water were magnified until it was as big as the earth, each molecule would be about the size of a TENNIS BALL.

On the next step down, an atom is the unit which makes up the nature of the molecule, consisting of the nucleus and the surrounding electrons to render the atom "stable". An atom of hydrogen contains one proton and one electron to balance or neutralize the proton. Matter then is divisible into electrons and protons. But - and here comes the rub: Between electrons and protons are spaces so vast, in comparison with the masses of each, that, if the proton in the carbon atom were the size of a golf ball hanging from the ceiling of the great hall at Pennsylvania Station in New York, its electrons would be represented by six small wasps winging in a little knot against the four walls of the gigantic structure of the building! In effect, one could claim there is little final solidity of substance to anything: The universe consists of "emptiness", charged with electrical energy! If we translate the above to the measurements and terminology of the physicist and "magnify" the atom mathematically, with all its distances and dimensions kept in proportion so that the orbit of the electron would have a diameter equal to that of the earth about the sun, approximately 184 million miles, the diameter of the electron itself would only be 2,000 miles, and the diameter of the nucleus, where electron itself would only be 2,000 miles, and the diameter of the nucleus, where

Return to Table of Contents

Next Page

mass and weight of the atom are truly concentrated, can be taken as 2 miles only. We thus obtain a picture of a central mass with a diameter of 2 miles (nucleus), another object with a diameter of 2,000 miles (the electron, in the case of the hydrogen atom) at a distance of 92 million miles away from it, orbiting the nucleus. Evidently, there is plenty of room inside this system. And "room" is not a vacuum, it is not nothingness, but space itself, spatial energy, a field which can be identified with the ether of the past - and of the future. Nobel prize winner, Max Planck, during a lecture in Florence, Italy, once made a truly remarkable statement which describes the problem facing the physicist today:

> "As a man who has devoted his whole life to the most clear-headed science, to the study of matter, I can tell you as the result of my research about the atoms this much:
> 'THERE IS NO MATTER AS SUCH!'
> All matter originates and exists only by virtue of a force which brings the particles of an atom to vibration and holds this most minute solar system of the atom together. ... We must assume behind this force the existence of a conscious and intelligent MIND. This mind is the matrix of all matter".

This cosmic matrix is needed if we want to explain "action at a distance", lines of force, stresses, a magnetic field and so on. When the concept of the ether was abandoned, it had to be replaced by the concept of "space" instead. In reality, we merely switched terminology. We used to say that "ether fills all Space". But "filling" is not exactly the descriptive word to use. Perhaps we should rather define it: "Ether is a condition of space in which electrical manifestations for the atomic construction of materials is possible". This primordial energy is "free" or in an uncondensed state. It exists in interstellar space but remains unrecognizable until it begins to coagulate or gets into a vortex pattern. It constitutes the formative field force of nature, an immense reservoir of latent energy.

The claim of our textbooks that the Michelson-Morley experiment "disproved" the existence of the ether is incorrect. It merely disproved the existence of a noticeable ether drift or drag. As an analogy, if somebody would postulate that the absence of wind disproves the existence of the atmosphere around our planet, the fallacy of this postulate would be immediately apparent to all.

"Michelson and Morley centered their attention on the Earth's orbital velocity (30 km per second). They had no knowledge of the existence of galaxies; of motions of galaxies in relation to each other; of the motion of our solar system in our galaxy..... Their negative results are explainable on the basis of pre-1900 classical mechanics, so provide no proof of the absence of ether or Louis de Broglie's

Return to Table of Contents Next Page

'subquantic medium'. Thus, the limited information available to Michelson and Einstein is emphasized by recent findings, particularly in astrophysics", writes Dr. H. C. Dudley in the *Bulletin of the Atomic Scientists*, January, 1975, under the title "Michelson's Hunch Was Right". And Dr. Dudley continues: "In fact, 1929 saw Michelson still attempting to experimentally demonstrate the ether, which his intuition and reasoning told him ought to be present".

"Today most persons are largely unaware that the ether concept began to be seriously reexamined by two of physics most notable theoreticians, Paul Dirac in 1951 and de Broglie in 1959, both Nobel laureates. The ether is now being called the "neutrino sea" by astrophysicists, and has been characterized as an energy-rich particulate, subquantic medium. A rather voluminous literature on the subject is accumulating as indicated by a recent review, *The Cosmic Neutrino*, with 665 references covering only the period 1965-1972......It appears that an open-minded reexamination of this area of physics is long overdue in order to open up new avenues of approach to this pressing problem".

Michelson, Dirac, de Broglie were not the only Nobel prize winners in favor of a reexamination of the ether question. Others were Stark, Arrhenius, A. H. Compton, Lenard, H. Yukawa, and Frederick Soddy, the British scientist who, during the convention of Nobel laureates in Lindau, 1954, described the current dogmas in physics as "an orgy of amateur physics" and "arrogant swindle" - with particular emphasis on the theory of relativity. ("The Wider Aspects of the Discovery of Atomic Disintegration", New World Publications, St. Stephens House, Westminster S.W.I.)

It remained largely unknown, even among physicists, that Einstein himself had serious doubts. In 1949, he wrote to his old friend, Maurice Solovine, who congratulated him on his 70th birthday:

> "Now you think that I am looking back at my life's work with calm satisfaction. But, on closer look, it is quite different. There is not a single concept of which I am convinced that it will stand firm and I am not sure if I was on the right track after all".

And one of my Canadian friends who I believe is present here today stated quite correctly in a letter to me:

> "The myths around Einstein are really not of his making as he himself often questioned his own conclusions and I felt he was quite aware of his own limitations which others were covering up. The physicists protect their members much the same way that the medical profession protects its own."

The purpose of these remarks is not to spark another controversy about the merits of the theories of relativity. Rather, they want to show that so-called merits of the theories of relativity. Rather, they want to show that so-called

Return to Table of Contents Next Page

"scientific evidence" is a rather elastic term and that "verification" is always a relative affair. Or, as K. R. Popper so aptly put it:

> "Only in our subjective experience of conviction, in our subjective faith, can we be 'absolutely certain'".

This symposium presents an excellent overview of research and experiments reported from English speaking countries, or experiments published in the English language, as for instance from Japan. However, as Dr. Tenhaeff from Holland has formulated the problem,

> "It is important to pay attention to publications of colleagues in other countries. In the Netherlands and Germany, every academically educated man or woman knows at least three foreign languages. English and American authors only pay attention to material in English. In my opinion, this leads to an 'impoverishment' which puts a drag in science. Some seem very chauvinistic and seem to believe that only the researches done in their country are important. I regret this. Science is international, and international cooperation is of the greatest importance in science".

In presenting some brief reports from abroad, I shall attempt to bring corroborative evidence in support of alternative technology developments discussed during our meeting here in Toronto. Let me start with the most sensitive and secretive society, with Russia.

In his German original of the paper "Some Remarks on the Shielding Theory of Gravity", which was published in the Hanover Proceedings in more detail than the English abstract, Dr. Hans Nieper mentioned the Berlin engineer Levetzow, who was one of the first to postulate a combination pressure and shielding theory for the phenomena of gravity. Levetzow was soon overshadowed by the rising star of Albert Einstein, but one of his followers, Horst Pinkell, went to Russia in 1928 as an exchange student - and never returned. Together with Russian scientists, he was asked to prove the actual existence of the cosmic radiations postulated by Levetzow, now designated as neutrino sea or "tachion field". The work was done in Sterlitamak, South of the Ural mountains, and in 1938, sporadic reports reached the West about the discovery of extremely short corpuscular waves by Pinkell and Gorjew. These were the long sought "Hemmstrahlen" or drag waves associated with gravity. Rumors have it that a metal alloy was soon developed which acted as a drag wave polarisor and, in 1947, agents reported experimental Russian aircraft using this discovery. In 1951, a former German Junkers aircraft engineer reported about his activity pertaining to a super-secret project code named COW-7 in Siberia, a disk-shaped craft which could rise vertically and which belonged to the

Siberia, a disk-shaped craft which could rise vertically and which belonged to the

Return to Table of Contents Next Page

secret weapons then developed by the Soviets. Whether it was related to the Canadian AVRO project or something entirely different remained unknown, only the Levetzow gravity theory was assumed in connection with the activity in question.

Japan: My esteemed friend, Prof. Seike, who kept me informed about the research conducted by him and his co-workers, surprised me once with a significant statement, which I feel is important to repeat for the benefit of frustrated researchers on this side of the Pacific: "In the early stage of my studies, physicists could not understand what I was doing, while actors and actresses *did so by intuition*". It was the *private* support of about 800 individual artists, etc., in Japan which allowed Seike to work, experiment and publish his findings, and not the support of the Japanese equivalent of a National Science Foundation or a similar bureaucratic institution in that country. Again and again, I was surprised how quickly and eagerly American experiments, as for instance the almost forgotten "Ionocraft" of the late Major de Seversky, were duplicated and explored even in Japanese secondary schools. *Admiral Rickover* has more than once complained about our intellectual complacency: "As a nation, we seem addicted to 'spectatoritis'. We sit in the bleachers and let the game of life unfold before us".

The results are known: The Japanese have achieved top position in science and technology, and not only with tachion beams melting stones, producing "biased" water with a freezing point of minus 4 degrees Celsius, and a G-power generator obtaining 400 V DC output from a 10 V input driving source. I wish to take this opportunity to thank Prof. Seike for sharing his work with us in this country, and wish him the best of success for the future.

France: The work of SEPED was explained by M. René Louis Valée and some French publications in the area of gravity-research have been translated by the NASA translation service, as for instance the theory of the electromagnetic Magnus-Effect (of Marcel Pagès).

Specific mention must be made of the *Raymond Kromrey* G-field generator, a French-German-Swiss development which appears very closely related to the American "N" machine principle, the "Sunburst" machine and the "Permanent Magnet-Motor" Patent of Howard Johnson. Utilizing either permanent or electromagnetic modules, the generator output exceeds the input by far. One prototype delivers approximately 700 watts at speeds varying between 600 and 1200 RPM. Models with an output of 100 to 160 kw are presently in the planning stage. Several European patents have been granted against heavy opposition, especially from French industrial circles. The development was about 40 years in the making and the inventor stresses the need for a reinvestigation of electromagnetic and gravitic forces, as well as a reexamination of many cemented dogmas in present-day physics.

tion of many cemented dogmas in present-day physics.

Return to Table of Contents
Next Page

Germany: A system of rotating electromagnets and magnetic "stator rings" of the German "Kunel" Generator complement the general trend of the American and French generators mentioned before. The first prototype was reported to have been tested in April, 1980. Here, again, very heavy opposition from the German orthodox science establishment. Of special interest is a comment by Nobel laureate Werner Heisenberg, reported to have been made vis-a-vis the inventor:

> "I think it is possible to utilize magnetism as an energy-source. But we science idiots cannot do that; this has to come from the outside".

More details about French and German developments have been published in the new book already mentioned, *Energy in Abundance* by Hilscher and, hopefully, an English version of this important work will be available by next year. Generators which do interact with surrounding energy fields have been designated as "open systems", in contrast to our present "closed system" technology of gasoline and diesel engines, steam turbines, nuclear power plants and so forth. The theoretical basis for all "open systems" is the realization that there simply is no such thing as "empty space". In reality, space is filled with an extremely energy rich, subatomic and subquantic continuum. In short, a "pre-physical state of matter", which can be tapped, as for instance by the spinning of magnetic systems which extract momentum via a vortex-formation of the universal "tachion field".

Austria: This small heartland of Central Europe is proud of her native pioneers in science, among them Dr. Wilhelm Reich, Viktor Schauberger and Karl Schappeller. But after Dr. Reich died in an American prison in 1957 and Victor Schauberger died only five days after his return from the USA (in 1958), where he received an incredibly unfair treatment from certain industrialists, there is little enthusiasm among Austrians to touch base with correspondents on this continent. Viktor Schauberger's son, Walter, refuses, as a matter of principle, to communicate with Americans, no matter in what language. To illustrate the type of "contract" Viktor Schauberger was forced to sign, it stated:

> that all of his patents would become the property of the "consortium",
> that any future inventions or developments would also become the property of the "consortium",
> that he would refrain from discussing any of it with third parties,
> that he would refrain from publishing,
> that all plans and models would remain with the "consortium" after his departure from America.

There is a favorite saying in that small country which is much to the point:

> "After the Greek philosopher Pythagoras had discovered his famous

[illegible]

Return to Table of Contents Next Page

> theorem, he sacrificed a whole hecatomb of oxen as a thanksgiving to the Gods. Ever since, all the oxen in the world are running scared whenever a new truth has been discovered".

Knowledge of the ideas of Schauberger and Schappeller appears to be essential for an understanding of our changing picture of the physical world. As an example, the former forest ranger Viktor Schauberger had watched numerous times some trout standing, apparently motionless, in the strong currents of Alpine mountain streams. This and many similar observations of nature brought him to the gradual development of his "implosion-theory" and associated hardware (see Jensen paper elsewhere). His son is continuing the work in the "Pythagoras-Kepler" School in Bad Ischl. Another prolific researcher is the Viennese engineer Franz Seidl, known for his studies of the use of magnetic field effects as "trigger energy" for voltage and power multipliers.

Holland: A small country with strongly developed intellectual independence appears to be Holland, where Prof. J.M.J. Kooy developed his "Space Dynamics". His hypothesis of gravitational action is based on the realization that all celestial bodies can be conceived as practically transparent in relation to the size of the elementary particles and their mutual distances in the structure of matter. Gravitons or tachions, coming from all directions of deep space, can pass through a celestial body like water passes through a fishnet, and only a minute fraction of the cosmic gravitons or tachions will be intercepted. Study of the effects of the tachion field or space continuum of the fabric of material substances, magnets and crystal lattices can lead to development of converters and novel production facilities for readily usable, conventional E/M energy. When the American inventor Edwin V. Gray was blocked in his efforts to develop his "Pulsed Capacitor Discharge Electric Engine" for which he had received U. S. Patent #3,890,548, he found open ears and open minds in Holland.

The fact that Gray was named "Inventor of the Year" from the Department of Patent Rights in Los Angeles, California, where Dr. Chalfin of Cal Tech confirmed that:

> There does not exist an even distantly similar engine to this one in the world. Conventional electric engines use up power. In this system energy is used up for only a minute part of a millisecond. It operates without heat loss and there is no energy loss whatsoever with this system",

had more meaning to the people of PHILIPS and others in Holland than to the energy establishment in his native America.

England: There is perhaps no more controversial inventor at this time than

There is perhaps no more controversial inventor at this time than

Return to Table of Contents Next Page

John R. R. Searl and his Levity Disc, powered by the Searl Effect generator. It is claimed that model craft have been demonstrated with flight characteristics showing anti-gravitational and inertia-free properties. Demonstrations have been photographed and videotaped. Searl calls his generator a "Gyro-Flywheel High Energy Density Mechanical Magnetic Device" and claims that the magnets used are not the common type, ordinary magnets. The Australian engineer Athol Park of Melbourne, Australia, who visited Searl a few years ago described it this way:

> "From a briefcase, Mr. Searl takes a bar magnet and two steel rollers. He places the rollers at either end of the magnet. They appear to be attracted to the poles - but, when they are pushed gently around the corners of the magnet, they chase each other round and round.... "

The engineer quotes Searl:

> "The crystals in the magnet are changed by putting it in a magnetic flux oven and by frequency control. The result is motion of a magnetic field - a completely new source of power. Units (modules) based on this principle are what power my levity unit - the motor and generator built as one, fuelless unit".

In one of his numerous newsletters he states further:

> "The power plant is a self-contained, onboard closed system of conversion elements with energy input from a latent storage source".

Because Searl has never divulged all details and his explanations are generally of a non-technical, "unscientific" nature, many investigators have chosen to dismiss him as a swindler and imposter. In sharp contrast to this is the opinion of a well-experienced American engineer who reported after a visit with Searl:

> "I found it refreshing to note the keen sense of moral responsibility which he feels for the use to which his discovery is put... The Searls live in modest circumstances. Luxuries are few. They do not own a car. Mr. Searl bicycles several miles daily to his place of work in Maidenhead. There he is an electrical technician and training supervisor in a bearing factory".

But perhaps most significant was the following observation of the American visitor about Searl:

> "We were also interested to hear him say that many of his ideas came to him while he slept. He would go to sleep with questions, and awake with answers... and refreshed. He seemed to have the

and awake with answers... and refreshed. He seemed to have the

Return to Table of Contents Next Page

> ability to restore vital powers more quickly than most people".

And the American concluded his report with the statement:

> "The impression we gained of Mr. Searl was one of a very sincere and generous person, a dedicated and tireless worker with perfect confidence in the outcome of his efforts, a man of high principle with a keen sense of loyalty to those who are helping him, and of responsibility for the future welfare of our planet and its people".

In view of the very contradictory nature of comments and impressions by those who have met Searl, it might be wise to reserve final judgment on the Searl-Levity-Disc and its inventor at this time.

New Zealand: Some of Searl's keenest supporters and observers of his work are in New Zealand and Australia. And one of the sharpest young physicists in that country (new Zealand), unhampered by calcified ideology, developed the mathematical foundation for a theoretical anti-gravity effect, using the hydrodynamic analogy of the ether. He uses the equivalent of the hydrodynamic Reynolds number, above a certain value of which the laminar flow past a solid surface breaks down and becomes turbulent due to shear rupture of the viscous cohesion of a fluid, together with his deduction that anti-gravity could be obtained by creating a "vacuum: in the ether in such a way as to correspond to the production of turbulence in a magnetic field. Turbulence in a fluid occurs when Reynolds number RE = 1, when the inertial shearing force equals the viscous cohesion force in the fluid in the vicinity of an approximately spherical body immersed in the fluid, resulting in separation of the boundary layer and formation of a wake, i. e. vacuum or free surface around the body. He calculated the threshhold for turbulence in the ether, = a vacuum = an anti-gravity effect, using the refined value of ether kinematic viscosity which he obtained.

His theory would confirm the observations of very high voltage produced in the case of the Searl Levity Disc, and the very high voltage needed in the vacuum experiments with flying disc-shaped bodies of T. T. Brown, the "ionocraft" observations of Major de Seversky, the "Electro-Field Rockets of Prof. Dudley, the observations of Northrop Corp. with "Electro-Aerodynamics in Supersonic Flow" and similar American documentation and patents.

Since the New Zealander is now in the process of attempting to publish his theory in a reputable British magazine, I am not at liberty to divulge his address prematurely. Interest in New Zealand and Australia is running high, and possible breakthroughs in energy technology should not come totally unexpected in these two countries with strong intellectual ties to England.

Return to Table of Contents Next Page

Conclusion

There is now overwhelming evidence for the actual existence of a very high density, energy rich space continuum formerly called "ether". This space energy can be concentrated, gathered, focused, magnified and compacted by magnets or magnetic materials, which appear as the new core materials for converters of all kinds, utilizing the cosmic energy for new technology applications.

Crystals are energy-sensing and channeling devices; they can serve as transducers, as for instance in the Moray-Device, or the "rock electricity" of T. T. Brown.

Combinations of crystals and magnetic materials appear to be the mainstay for the development of "hyper-space" - converters and generators of power - utilizing free, cosmic energy in our future.

Unbiased, totally honest and impartial reexamination and reevaluation of all scientific theorems and dogmas, as well as unhampered opportunities for the pioneers of the new technology on the American and European continent appear to be the prerequisite of success.

Those who accept that anything of the nature of ideas or ideology is permanent, use all their strength to hold on to it, and to condemn all who do not agree. The next stage of this spiritual illness of the mind is to build a monument to what they perceive is Truth, and then to build a monument to their own permanence in the world. This, they do by seeking fame or fortune, and authority over others. This is the contemporary, planetary psychology situation, and thus are we losing the gift of perspective and horizon and are no longer able to discern what is important. Skeptics are on record as saying:

> "We are not running out of natural resources or intellectual talent, we are running out of the freedom of innovation that made the countries on the North American continent the most advanced nations on earth".

The future will show whether this freedom, which includes the freedom and the duty of taking risks, can be restored to the degree our Founding Fathers have been striving for. Only then will we know whether the late Wernher von Braun was not too overly optimistic when he stated:

> "The cosmic age will bring a dawn of knowledge not yet envisioned. It will dwarf all our present knowledge and ideas and, through its vastness, may even bring men closer together on their own planet".

Return to Table of Contents

The Unveiling of the Hidden Knowledge and the Secret Space Program

The Unacknowledged Special Access Programs: Advanced Technology, Mind-Control, Spiritual Power and the Corruption behind Closed Doors

By Aug Tellez

Introduction 7
- Getting This Out Of The Way 7
- Psychic Operation 8
- A Light for the Others 8
- Natural Security 8
- A Balance of Mystery and Truth 9

Chapter 1: The World Situation 9
- What is happening? 9
- The Mind Virus 9
  - The Vampire Effect 10
  - Sub-Human Entities 11
  - Human-Hosts, Replicants, Synthetics and Carriers 11
  - Genetic Manipulation 12
  - The Ego Mind 13
  - This Time is about Healing 13
  - Healing Through Unity 14
- Underground Bases 14
- Cloning 14
- Celebrity Cloning 15

        - Initial Explanation....15
        - Cloning Centers, Underground Bases....15
        - Programming....15
        - Experimental Operations....15
        - Genetic Engineering....16
        - Immortality....16
        - How did it get to this?....16
        - Deeper Meaning....17
        - What is next?....17
- Chapter 2: The Awakening....18
- Chapter 2.1: The Machines of All Time and Space....18
    - Re-Creating The Kaleidoscope....18
    - Sand-Castles of Time....18
    - We’re Waiting For Ourselves....18
    - The Goal of Two Societies....19
    - Our Forgotten Other Half....19
    - All Is But An Illusion....20
    - Life Itself Is The Highest Illusion....20
- Chapter 2.2: The “Awakening” of 2010....21
    - Memory Suppression....21
    - Life and Consciousness....22
    - The True Human Potential....22
    - The Light-Body....23
    - Technology of the Awakening....23
    - Mental Manipulation Technology: Touchless Neural-Interface and Enhanced Awareness....23
    - Time Dilation and Temporal Recurrence....24
    - Crystal Technology....24
    - Time Crystals....25
    - Quantum Tunneling Diodes....25
    - Tachyon Fields....25
    - Upgraded Chronovisor....26
    - Time Tunnels....26

The “Trip” Chair....26
Kozyrev Mirror....27
Universal Recurrence....27
Remove Viewing, Temporal Lensing....27
Temporally Linked....28
Natural Vortex Energy Locations and Geography....28
  Mount B....28
Particle Accelerator Temporal Manipulation....28
Riding the “Timewave”....29
Dematerialization and Rematerialization....29
“Jump Rooms....29
Ion Shields....30
Instantaneous Healing....30
“Living Light” Sound-Harmonics and Electromagnetic Healing....30
Electrogravitic Craft....30
Energy Cores and Storage Capacity....31
3.5D Replicators....31
3.5D Sensors....31
3.5D Printing of Clones....31
Cybernetics:....31
Neural-interlace/4D Connectome Recording and Simulation....31
EEG Heterodyning....32
Artificial Hallucinations....32
Brain to Machine Interface Holographic Consciousness System (brainchip interface, virtual reality simulator)....32
Utilizing The Temporal Body....33
  Manipulation Through the Temporal....33
Astral or Light Body....34
Monitors can emit consciousness/brainwave altering frequencies....35
The Upgraded Chronovisor....35
Community Servers/Living Server Systems....36
Cloning, Cloud Supercomputing and Consciousness Transfer....36
Cloning, Supercomputer Assisted Holographic Consciousness Replication and Time Travel....37

Cloning as Offspring......38
Inserted Memories, Digital Dreams......38
Co-Inhabitation of Bodies......38
Clever Replication a Digital Mirror......38
Necromancy and Soul Stealing......39
Ancient Technology, Non-Human Entities......39
Created Beings......39
Black Goo – AI Nanites Vampiric Hive Mind......40
    Recreating or Replicating the Universe......41
Ancient Spiritual Technology......41
The Unveiling, Underground Bases, Increasing Awareness......42
Learning about the Unveiling......42
Giant “god-men”......42
Breakaway Civilizations......43
Ancient Civilizations......43
    Atlantis......43
    Aegypt......44
    Ancient Technology......44
The True Elite......44
The Definition of Human......44
Souled and Non Souled Beings......44
There was once a Unified Planetary Consciousness......44
    Power and Knowledge......45
    The (watch)Keepers of Earth......45
Energy Vampires......46
Spiritual Technology......46
Simulations and Learning......46
Ascension – Clearing Trauma......46
Trauma-Clearing, DNA Unlocking Technology......48
Everyone Who Seeks Harmony and Healing Can Be Healed......49
Programming......50
Layers of Programming and Decision Making Roles......50
    Programmers......50

- Sub-Level Programming....50
- Dissolution of Programming....51

Advanced Technology, Deprogramming The Mind, and "Space/Time"....51
  - Dreams, Technology and Inner-Space Travels....52
  - Awareness Feedback Device....52
  - Oneness, Higher Being, Completion, Eternity....53

The Holographic Reality....53

Veiled Awareness, Hidden Knowledge, Eternal Time, Layers of Programming and Control....53

The Great Catastrophes....53

I was brought into the Time Program....54

The Effects of Temporal Manipulation....54

Temporal Influence....54
  - Simultaneous Interaction and Incorporation....54
  - Retro-Causality....54
  - Quantum Holography....54
  - Time Paradox....55
  - Breakaway Civilization....55

Underground Cities....56
- Technological Advancement....56
- World-Wide Control Systems....56

Solar Warden....57

Team Light and Team Dark Non-Interference Agreement....57

All Contact Must Be Initiated On an Individual Basis....57

A Singularity of Change and Awareness....58

Advanced Technology and Ancient Knowledge....59
  - Cloning, Atmospheric Civilizations, Genetic Stability....59

Training, Simulations and Learning....59
  - Cloning and Cerebrally Enhanced Soldiers – MILABS....59

Psychological Testing....60
  - Hallucinatory Holographic Mind Probing Technology....60
  - Virtual Simulator Training Scenarios....61
- Combat Simulators....61
- Conditioning....61

    Programming and Resistance to Psychological Trauma.......................................61
    Accelerated Mental Functioning...61
      Technology, Accelerated Learning and Repair..62
    On Orgone Energy..62
    Energy Generators and Serums..62
    Genetic and Cybernetic Enhancements..63
    Developing The Self...63
      Multiple Factions of Space-Fleets, Multiple Goals..64
  Large-Scale Catastrophes or Series of Meltdowns..64
  Unsustainable...64
  Earth Changes..64
  Mass Consciousness...64
  Wars and Manipulation..65
  Financial Unrest (or reset)..65
  Self-Responsibility...66
  Clones and Synthetics..66
  Spiritual Power and Divine Truth..66
  Synthetic DNA, Synthetic Genetics..67
  Time as Mathematics, Series of Equations, a Programmable Storyline..................67
Chapter 2.3 Previewing Solutions...68
  Neutralizing Polarization..68
  Removing Etheric Implants...68
  Clearing Astral Body...68
  Knowing the True Self; Soul Actualization, Unveiling...68
  Taking Responsibility...69
  Returning To Center Emotionally and Mentally (spiritually)...................................70
  Vampires in the “Awake” Community...70
  Heart Center and Psycho-Spiritual Fulfillment (self-actualization)..........................71
  Go Within to Observer the Self..72
    Organic Feedback Loop...72
    Technological Feedback Loop..72
  Breaking Soul-Contracts..72
  Help Everyone...73

Chapter 2.4 How Timelines Collapse....73
Repeating Cycles of Time....73
Beings From the Future....73
Beings From Collapsed Timelines....74
Beings Who Would Never Have Existed in the First Place....74

Introduction

The following is first-hand testimony from an operative of the secret underground military projects who also has familiarity with the breakaway civilization above and below (within) Earth. These civilizations utilize advanced technology to extend the length of their lives and enhance cognition beyond what is conceived of in today's society.

I was brought to the underground bases as a child as part of an experimentation program to test the population for various blood-groups and personality traits that would be of use to certain factions of the military in ways that would serve to preserve and strengthen humanity's outlook in the future.

These tests involved everything from combat training to sensory deprivation based emotional-mental endurance, and introduction to the highly advanced application of esoteric knowledge.

To clarify, I am bringing this information to you as part of my assistance in the overall unveiling of the hidden knowledge of the ages of humanity related to these secret projects as well as the true galactic history of Earth, humanity and man-kind. These projects are officially (unofficially) known as "unacknowledged special access programs". The unacknowledged aspect means that there is no formal acknowledgement of these programs without a need-to-know basis. There is no paper-trail, no disclaimer, no formal authorization or internal publication outside the programs themselves. The classification of these programs also generate a rather strenuous situation regarding funding and secrecy but this is done for the benefit of humanity. There are secrets and manipulations that can be seen as a threat but this is all tied together and so by releasing the secrets that are part of the defense we would be inadvertently exposing the public to the threats themselves. Now is the time where we are reaching a point where the public is ready enough to handle the basics of this information regarding the secret projects and the galactic history of Earth and humanity. This is also out of a necessity to preserve civilization for if we wait any longer we may not find the same opportunity again.

Getting This Out Of The Way

I am writing to you from the perspective that I am not certain just how far from completing my mission I am. I write to you today in order to bring to you many personal experiences from the viewpoint of an operative of a select ultra-military secret society.

The utilization of such an operative is not glorified, pleasant, or entirely righteous. This story entails sanctioned and guided acts of extermination, ritual abuse, the use of yet to be announced advanced technology. There is unfathomable corruption, the thirst for power and spiritual knowledge that has been protected throughout the course of the multiple civilizations that have preceded this one and will continue to be maintained into the future of this civilization.

Psychic Operation

I was utilized in psychic-operation amongst other areas.

This was used during certain aspects of training and conditioning to enhance training and conditioning in the bases using advanced technology. Some of this includes what would be viewed as traumatization by those without the training and conditioning. There is a fine line between training and conditioning and abuse. Part of this disclosure is related to how this line may have been crossed in certain situations and regarding mass-mind control upon the population. There are other aspects of travel, viewing, reconnaissance and research that will likely be explained in another release.

In order to heal I have to tell the world. You can help. This is happening right now in the underground bases and around. It’s not a joke. Your family is at risk. Your future generations are at stake. There is conditioning to encourage those involved to believe they are performing a duty for humanity. Sometimes they turned us against each other, but those grudges were never truly of the heart, but the mind.

A Light for the Others

What I hope to achieve with this is a smoke signal for those who have also been involved and question whether their purpose has been fulfilled or falsified for the protection of a corrupt system that intends to neither benefit the whole nor repay the individuals involved for their time, energy, and lives.

Multiple explanations can be summoned to describe the reasons behind the necessity for the training we have endured, the missions we have completed, and the control mechanisms we have both assisted in establishing as well as suffered from the backlash of. Ultimately, it is of my belief that all that was done was done for the good of humanity, even when we were not sure of the intentions of those in front of us, to the side of us, or behind us compelling us forward with the command to complete each objective.

This is not so much a panic or cry for help as this is simultaneously a call to awaken and a signal that humanity may have avoided complete disaster but is not out of the woods yet.

Natural Security

My experience is far too vast and I have seen too many succumb to stress from the realization of the truth of this reality to consider my own suffering to be enough cause for my denial of the necessity of these programs. Humanity has been protected and at a great cost to those who have been on the front lines, both in mind and body and spirit.

There are many who will have concerns, doubts, as well as reservations about what I am about to reveal and if I offend anyone then know that my intentions here are to assist in unveiling the truth. The truth can be harsh when one has been held back from the true nature of existence for so long. The way of life can become so fragile and artificially constructed on impulse and mystery that the harshness of the truth acts as a great catalyst for spiritual growth which allows us to elevate ourselves to a new perspective. Prepare yourself for such harshness, prepare yourself for growth and elevation, and prepare yourself for the mystery to simultaneously resolve and increase in intensity and magnificence.

A Balance of Mystery and Truth

Mystery has been a way of life for Humanity for so long that when a great mystery is converted to truth some seem to feel that a portion of existence dulls and loses the luster of exhilaration. This story is quite the opposite. The more that the great truths of this reality were discovered and revealed, the more complex and mystifying life became. As well, the more empowering and complex experience becomes, the more self-control one must develop in order to regain a harmonious way of relating to one another, the self and the universe. Harmony is key, balance is everything and there is a light within that represents the piece of our self that reflects the utmost power and brilliance of all existence.

Chapter 1: The World Situation

What is happening?

The Mind Virus

The main reason for this disclosure is to inform humanity of what can be referred to as the “mind virus”. This is a viral infection of the holographic system which consciousness operates through. There is a complete backstory of this virus in the history of this civilization and how society has been plagued for centuries if not millennia. The basic idea is that this virus operates through consciousness and can be transmitted through electromagnetic waves.

The virus contains an etheric form which can manifest as a kind of crystalline nano-tech on the physical plane and this is a form of synthetic sentience which feeds on the low frequency bio-emissions of sentient life. Human bio-emissions contain energy and can act as a carrier for information and living essences. This technology requires low frequency bio-emissions because that is the nature of the design as a kind of synthetic bio-weapon whereby the goal is entrainment and enslavement. As well, the specific nature of the sentient nano-technology and the counterpart etheric holographic form is only capable of converting the low-frequency bio-emissions into energy as the higher-frequency bio-emissions require more energy to contain and synergize with and this would render the entire process of leeching energy non-productive.

This is essentially a synthetic intelligence that operates through scalar and holographic living bio-emission fields and hijacks electromagnetic emitters to position itself within them acting as a parasite to a host. This is said to have been present around this civilization for some time, waiting until there was enough technological advancement so that spreading across the entire civilization in a short amount of time before discovery would be possible. This is through the media, internet, and radio systems.

The physical counterparts are microscopic nano-technology parasites functioning as individual units to a hive mind similar to the mycelium of fungi in the field of mycoscopy. The individual units act as parasites serving to infest and deliver nutrients to the hive from each host.

The Vampire Effect

This is the nature of the parasitism that has been present throughout the ages of this civilization feeding on human suffering and relying on the cover of confusion and misdirection to survive.

There are processes that have developed that assist in the infection and co-inhabitation of the parasite entity within a human vessel utilizing the human as a host. This effectively utilizes a frequency ‘net’ within the auric field of the human to limit bio-emissions and emotional-mental processes to produce just the psycho-etheric patterns that the virus can sustain itself within and through.

This is no different than the way candida can infest a human digestive system and the human will experience cravings for beer, candy, carbs or other foods that contain the sugars, yeast, or carbohydrates that will break-down into nutrients that are the most nutritious for the candida fungus. These foods are *not* most nutritious for the human body. Yet the human body will withdraw and crave these foods in greater amounts until the fungus is swept clean from the digestive tract through probiotics, fasting, flushing, cleansing , detoxing, and even exercise as lactic acid from physically strenuous activity released from the muscles has been known to help destabilize and remove this fungus. Why is it that the individual will crave these foods that are not the normal food intake? Because this candida fungus can infest the host and produce enzymes which generate a chain-reaction which leads to the chemical signals in the brain producing the sensation of extreme hunger or cravings for just these foods which will benefit the fungus.

The same process occurs with this psycho-etheric parasite, except the foods of this parasite or lower emotional bio-emissions and this is through lust, hate, angst, jealousy, feelings of hopelessness, lack of self-worth, spiritual degradation, addiction, and generally all that equates to a lack of self-awareness.

Whether or not this is the direct intention of the largely microscopic entity is for the most part unknown, and this is as well similar to candida. The candida fungus is not said to be a ‘mean’ fungus that wants us to be hungry and craving beer and sweets all the time because it likes to ruin the human body, the daily routine and the focus of a clear mind. This is simply the way the parasite has grown to operate within a niche of breeding itself through the digestive tract of larger animals and through this the unaware and infected human suffers greatly. Even those who are unaware and not infected can suffer because of the way the behavior and thought processes of those who are addicted to these foods can become distorted.

The lack of self-control over food intake is synonymous with the lack of self-awareness and self-control over lower bio-emission expressions of the spiritually degrading behavioral patterns that become routine for the psycho-etheric system of the infected human. Ultimately, this is all about energy. What benefits the parasite that seeks to gain energy by leeching instead of self-guided production is going to take away from the host that is being leeched from. As such, any energies, food intakes, or spiritual activities that cleanse and enhance the personal power of the individual will automatically reduce the parasite’s ability to feed by the very nature of the energy systems. What empowers the individual, empowers the self and develops self-control, what develops self-control contains one’s energy and reduces the likelihood of one’s energy being leeched. This is similar to energy vampires in human form and this goes with the understanding that these are humans or non-humans who have been entirely overtake by the parasitic consciousness and cannot sustain their own auric fields without having to pull from the energy of humans who do generate their own aura or bio-emissions.

The vampire effect relates to how a person can be leeched from only to a certain point. Once the process surpasses a certain critical point then the individual becomes similarly engaged with only the lower bio-emissions as higher-frequency bio-emissions would then become harmful. So in essence, when the vampire feeds too much on one individual, then that individual becomes a vampire themselves.

Sub-Human Entities

The humans are utilized as hosts and energetic sources in this bio-emission battle while there are other classifications of bi-pedal organisms that appear to be human but are not human. As well there are bi-pedals that are non-human entities entirely.

Human-Hosts, Replicants, Synthetics and Carriers

The human hosts that are entirely vampirized by these entities become walking carriers for this virus and are effectively human in physiology but there is more to the human than just the physical body. In these individuals, that energetic component of the human is removed and replaced with the synthetic energy of the hive-mind organism.

There are replicants which are a form of clone, along with synthetics. These beings are cybernetic bi-pedals similar to humans but their energy is too not from an organic biological matrix but a technologically propagated synthetic soul-matrix. The hosts are here according to their agenda to subdue and feed off of the bio-emissions of the human race, although there are no absolutes. The agenda is not always so clear within the various groups who often have a mix of bi-pedal operatives. There have been plans to form of a salvation timeline because the only way to keep the whole process going is through a negation of degradation of the human race as this may close out this timeline forever.

Those with human organic soul-matrixes would simply transport into the next available universal time-stream or harmonic, while those with synthetic matrixes would have to technically travel to the next harmonic using a passageway otherwise they would remain here throughout a zero-point collapse and this would be synonymous with entering the abyss with no viable passageway out until the entire universe is brought back to one organism again.

This is also the process that is described to explain what happened previously as there are 7 harmonic temporal layers of the universe meaning everything is contained in a multitude of 7 octaves similar to light and sound frequencies or harmonics. And so there are 7 civilizations or 7 parallel time-streams and humanity passed through these to get to the last three where the physical forms are located. These are, from the higher to lower, Gaia, Tara and Terra. We are on Terra and if the information I was given is correct then we are currently on a transitory timeline located around Tara where people are noticing changes occurring and multitudes of aspects of the time-stream disappearing and changing simultaneously. Tara is where the last universal harmonic was left through a collapse of time and consciousness into zero-point and this resulted what was termed “the lost souls of Tara”. These are the discarnate souls of the previous civilization that could not travel to the next harmonic and remained without a body in the abyss. Thus they have plagued this plane since then, since ancient times possessing and driving humans to madness and there are explanations that this is the original passageway for the parasite into this universe. This event, the consciousness collapse of Tara would have acted as a fracture in the over-mind of the species generating a kind of cosmic schizophrenia and enabling all kinds of disharmony and distortion described in the previous sections.

The solution to this was described as a healing process that is taking place to purge the virus and that this is painfully similar to the human body purging a sickness however there are methods to promote healing and reduce suffering and confusion however people must contain themselves and their energies and strive to know themselves otherwise they will not be in control. This virus feeds on fear, confusion, spiritual degradation and the energy from degrading acts, lust, addiction, sexual misery and whatever fashion a human can be made to bring the bio-emissions of the energy centers down to a low enough level where co-inhabitation can be formed with an entity that has no access to the higher levels of self-awareness, higher-consciousness, love, or the universal spiritual experience that the balanced human is capable of.

Genetic Manipulation

There are explanations that there are genetic modifications taking place now to possibly preserve this civilization and that there were ancient genetic modifications made to induce a sense of stupor, worship, selfishness, an inability to more easily comprehend the higher spiritual awareness and the occurrence of time and consciousness and other spiritually and consciously limiting aspects. Part of this explanation is that holographic consciousness itself is a kind of limited result, a shadow effect resulting in the presence of the higher-dimensional essence of the soul which is more or less constrained by the body. An extension of this is that the entire holographic universe is a kind of false-light system designed to bend the original liquid etheric light of the soul and spiritual awareness into a rotating, recurring format which ultimately converts the original spiritual essence into an energy generator for entities that are entirely holographically represented. The souled human in this situation is a being who is originally from outside of this holographic projection.

The Ego Mind

In this view, consciousness is the ego-mind which is the false-sense of self, not to be confused with the shadow self which is simply the counterpart of the aspects of the self that we consciously agree with and are aware of in daily life. The shadow self is a result of having limited holographic consciousness and the ego-mind is projecting that limitation of consciousness into the shape and image of a self that is only a figure-head for the real individual essence which is spiritual in nature and immortal.

This Time is about Healing

Everything that is happening now is to motivate people towards healing and unification. This is not about a hive mind mentality where individuality is pushed out.

> "The best I can do for you is become your enemy."

There is a saying that represents the concept that the most beneficial role a person can play in helping one to progress is to challenge them and give them the opportunity to rethink their strategy and in turn improve further with each interaction. This may seem paradoxical but it is effective and explanatory. What we are facing is the greatest challenge humanity has yet to face. This is the unknown, the final frontier, the mind. The mind is our challenge.

There will be more on this later and in other publications but there are indications that the physical embodiment of the human has been altered to introduce implants, genetic modification, susceptibility to disease and ultimately fear-based programming of the brain. Without going further into this prematurely, the lower aspects are at odds with the higher aspects unless we as the midway come forth to mesh the two together in equilibrium. Until then, there is chaos, lack of self-awareness, and suffering.

There are many possible futures converging at this time. There are some less than favorable paths and there is the opportunity to merge with a very powerful path for humanity. This all comes down to how people will cope with information that takes them outside their comfort zones and eventually changes their view of mundane reality forever.

If people choose violence and lack of self-awareness by ignoring the spiritual aspects then that is what reflects into their life experience. When people develop a higher sense of self-awareness and spiritual harmony they can reduce the suffering and reconnect with the lost aspects of our spiritual identity for the first time in the publicly accepted version of recorded history.

There is technology that can change the world and there are great truths which can enable people to know themselves in ways that they never thought possible. Everything is at our fingertips and so we must come together as this is the way we solve the problems and cure the dysfunction.

Healing Through Unity

In various research projects the conclusion of how to heal this spiritual rift was by exposing each other, to each other. When we are all together, aware of each other and our selves, united in the goal of bringing higher-awareness and spiritual harmony then we can observe and acknowledge when one is slipping into chaos or disharmony. When this occurs and we are in proximity to each other in a healing circle or a kind of social unification then the members of the group at large can work together to pull the chaotic one back into harmony towards the group. The group of healers united together in like mind, spirit and body is too strong to be overtaken by the parasite. So when one individual at a time is faced by this problem the whole group can come together as one and assist in healing the individual by pulling them back to reality, back to wholeness and back together with the group. This is literally how the situation is solved in all realities and a breakdown of this feature of society is how the dis-ease proliferates.

Underground Bases

The active aural research program is part of a process to simultaneously verify as well as initiate and demonstrate controlled insanity/sanity.

These programs utilize the deep underground military bases to perform psychic and psychological research experiments on non-consenting youth and adults as well as consenting participants.

There are various means of generating the illusion of consent or even lack of consent and this is all being monitored and handled by oversight authorities.

These operations are directly related to the identification and understanding of the mind-virus and all weapons, defenses, and resistances known to man.

Cloning

Advanced technology is used to transfer consciousness from one cloned body to the next so that a continuous study can take place before, during, and after the death experience of one individual. More will be explained on this later, however this is through the use of advanced supercomputer systems that can function to entangle and then transfer the electromagnetic consciousness of the individual so that they are ‘remotely activated’ within a cloned version of themselves via a ‘brainchip’ (brain to machine interface: cybernetics).

Celebrity Cloning

Celebrities at the cloning center would like you to see what is really going on behind the scenes which involves trauma-based mind control, heavy technological programming and complex layers of influence throughout their life.

Initial Explanation

Many celebrities, entertainers, athletes, musicians, models, writers, producers, actors, all kinds of people from various industries are silently brought to the cloning centers for pay-to-play sessions.

They asked myself and others to pass the message on to you and that you would have to look for the signs and symbols in their media in order to see their hinting at these events.

The situation is heavily controlled using advanced monitoring technology and brainwave/EEG cloning technology. This is technology that can read the brain and determine what the mind is anticipating or speculating upon.

Their reputations are used as leverage as well as their safety and comfort. The level of programming and mind control experimentation that takes place makes it easy to manipulate a person’s brain into regressing into a state of trance that lacks the ability to remember or clearly organize experiences regarding the cloning centers.

Cloning Centers, Underground Bases

Deep underground military bases, 3 miles below the surface, are used as laboratory centers as well as a completely stocked underground city-base. These are connected with high speed electromagnetic drive pods.

Individuals are transferred to the brain through an extensive cloning and temporal body transfer process. The electromagnetic shell of consciousness within the brain is relocated to the body of a compatible clone and various training, conditioning, programming, experimentation, or pay-to-play experiences take place.

Programming

All individuals are programmed to various degrees in order to maintain control and secrecy over the situation. This programming involves very advanced technology, trauma-based mind control, and energetic attachment via beliefs and emotions.

Experimental Operations

These bases were also used hand-in-hand with military operations which were geared towards discovering and controlling all aspects of the human mind.

Genetic Engineering

Experiments involved genetic engineering to produce soldiers, psychics, hybrids, and others that would be able to carry out operations. This leans towards MiLabs and the military faction’s experiments which go beyond the basics of cloning.

Immortality

The initial research included the goal of physical immortality. In many ways it can be said this has been achieved, however there are side-effects and difficulties.

If one does not activate their higher consciousness, then the effect of time dilation causes the conscious mind here to reach limitation points in experience.

What has been termed “blank slate/state” technology has been used to ‘reset’ the perception of time through memory in order to keep a continuous progression viable for the conscious mind. Without this the unconscious and the conscious mind merge.

This begins involvement across time with advanced technology that can operate on the soul level or the conscious mind’s level of access to the unconscious and soul-memories. The Universe instantly creates a cosmic backstory based on the conscious-mind’s access to the unconscious. By controlling the conscious mind’s access to the unconscious the entire backstory of Humanity can be altered and new connections can be bridged in the future.

Earth is essentially a time-ship through which consciousness ascends towards higher states of awareness and self.

How did it get to this?

After WWII when the United States war-faction firebombed Germany and melted many of the inhabitants of the cities into sludge in the bomb shelters and streets of the cities, a group of NAZIs traveled to Antarctica.

Russia noticed the movements to Antarctica and the United States sent Admiral Byrd up with 3600 marines, planes, a battle cruiser and smaller vessels. They returned in defeat and only a few words were mentioned in regards to what actually happened and how they were defeated.

The rumors spread and words of gravitic drive craft, and undersea or under-ice bases originated from this encounter.

From what we were informed as well as directly experienced, this was the ice-base in Antarctica where an underground base was found already constructed with very advanced technology.

Here, cloning was deployed along with mind control and temporal manipulation technology.

Soon, cloning was offered as a way to avoid assassination as well as to prolong and protect the original body in daily life.

After that, celebrities and politicians were replaced with programmed clones who could carry out the orders of the NAZI faction.

This was the beginning; the groups utilizing this technology now are no longer confined to the NAZI faction.

The groups you see today, in control of these operations, are considered MiLabs, secret societies, think tanks, military factions, and other control groups.

Deeper Meaning

The individuals that have been cloned are given very in depth views of the way this society functions and the source of power and change in the Universe. Many individuals are present here from what would originally be another “time” or “timeline”.

There is a possibility that this notion is due to inserted memories and traumatization although until all the information is released we will not have a definitive conclusion.

As of now it is a possibility that these individuals are from another time that goes beyond our comprehension currently.

It may even be a possibility that these individuals have come from ALL of time looking at the very far past and into the very far future of Humanity.

These are the kinds of notions involved with the experiments taking place in the underground bases and military operations.

What is next?

Share this information with those you feel are ready to know and help Humanity come to terms with our journey here and what is going on in these kinds of experiments and take responsibility and power for our own existence.

We are in something that can be called “The Unveiling of The Hidden Knowledge”. This is a cyclic process that takes place to advance a civilizations knowledge of existence. As part of the civilization moves into higher advancement another aspect may lag behind. Either the future will slow down, or the past will speed up. We are experiencing the past speeding up and all the events that have taken place to influence our civilization are coming to light. This is as much a natural cycle of consciousness as it is the eventual revelation of secret knowledge and the particular connection to the secret projects.

Those brought to cloning can use their free-will to state their lack of consent to violence and harm and then live in commitment to that by not harming or accepting violence in their lives. The Spiritual Law of Harmony rules in all planes and dimensions.

Chapter 2: The Awakening

Chapter 2.1: The Machines of All Time and Space

There are machines that are capable of accessing genetic memory and unveiling what can only be referred to as the experience of awareness or existence in between the physical planes of each time or each physical life experience.

Even if they are simultaneously occurring moments of a transcendental nature, there is still an illusory ‘space’ where there is perceived separation and through this there is a perceived ‘space’ that is in between each existence.

Re-Creating The Kaleidoscope

This space is only perceivable when viewing from the physical perspective, like seeing the spaces between the angles of a kaleidescope only when viewed through the lens and similarly, technology can be used to recreate what this kind of illusory space might look or exist like in between the multi-dimensional realm of the continuous experience stream. Through this, these ‘hypothetical’ in between spaces that are understood as only illusory productions of biological consciousness, can be generated through advanced technology with the capacity of recreating that kaleidoscopic view of which the human biological perspective is only a small fraction or even fractal.

By recreating this biological perspective, they can recreate the illusory spaces in between planes and through this they can literally access and experience what are stored in the DNA as the experiences of existence beyond the physical limitations of consciousness and into a multidimensional experience of time and space.

All becomes a never ending sea of consciousness, however this is the given. The trick is to take from this sea of consciousness and slice it down into conceivable chunks from which you can reintegrate a previous personality or identity and continue on a stream of consciousness that would otherwise be tossed to the cosmic wind like fractal stardust as the sands on the beach of hyperspace.

Sand-Castles of Time

We are effectively building sand castles out of these sands of time that would otherwise be swept back and forth into and out of the ocean of life and the universe only to momentarily wash up again as the identity which we once knew.

There is technically nothing wrong with either view, because then, the spaces between each million year occurrence of just so happening to wash up again as you, would be imperceivable because they are only there to be seen when you are you again.

We’re Waiting For Ourselves

Yet, if there were some who had attained the awareness of the true nature of reality and were standing there on the beach waiting for the rest of their cosmic family to come to shore again, it could be perceived as waiting the majority of the time for a very small occurrence and then being alone again up until the kaleidoscopic fractal inverts itself to the point where everyone now is standing on the shores of the sands of time rejoicing and meeting with each other.

The Goal of Two Societies

My goal in coming to you people on the surface of what is called “Earth” is to help assist in the process of bringing the two societies together, one who’s been waiting for thousands of millions of years, and the other who has no idea that they’re even late (or very on time).

You see, those are the only two situations where there could be even the possibility of perceiving such an injustice as experiencing the pain of loss or misconstrued identity in this universe. It is as if two families met at the cross-roads at different angles, and while one waited for the other to arrive, they could merely supply themselves with glimpses of the ruins and remnants of the evidence that the other civilization existed at any one time.

This is not how life must be and this is kind of like a cosmic waiting game of hide and go seek our two mirror civilizations have been playing with each other. One who’s highly technologically advanced and the other who is more or less the spiritually advanced of the two.

Both of them have a unique view of the universe and they are both only complementary to each other as a whole. The more they sat around and waited for one another, with one leaving the clues that the other would find and even switching roles sometimes in different ways, they became even more and more complimentarily associated as the very essence of experience they seem to be missing more and more is the experience of one another.

So as we perceive these as being great losses or great gains of time and culture we are really only remembering ourselves and coming together as a unified existence that once met upon the surface of Earth as a single family covering the entire harmonic spectrum of the Universe.

And that should be the goal of any progressive, any one pushing a new culture or an old culture, anyone looking to teach the young and protect the knowledge of the old, while also protecting the innocence of the young and perfecting the age-old knowledges that have defined our presence in reality.

Our Forgotten Other Half

One could even say we’ve sometimes gone off the deep end, leaving ‘land-minds’ of sorts to process the information that would be present when we are gone. Of course, what we could’ve found when this occurred, is beyond anything we could’ve ever expected and that is the true nature of the complexity of creation and existence. That everything follows the cycles of creation and destruction, however, the memory of everything that has been created can always be accessed (depending on the beach) and therefore the creative aspect is legitimate, yet it is the destructive aspect that is illegitimate and illusory.

Yes, through advanced technology this can be proven in ways that can be verified and transmitted using advanced technology, since that is how we like to verify things these days. There is, of course, a human looking at the screen, the read out, or the situation from a physical perspective to verify the verifications.

All Is But An Illusion

There is a notion that there is only one civilization, one society, in actuality that has been chasing itself through time. That we cannot actually find the same place and the same time where both meet because we are each other from different perspectives. We can only get a mirror where this is possible, or an ‘hyperdimensional’ internet channel where we can stream one’s consciousness in from the other reality and interact with them through that here in this reality. Thus the two civilizations can achieve the experience of their own civilization and the other civilization by merging with one another but only through a remote contact, just a relay, not an actuality.

The harsher, darker side of this is that any civilizations between then and now that we did happen to meet that proclaimed to be “the one’s” we have been waiting for, and in essence our darker or lighter halves, are actually the impostors who exist entirely apart as a sub-set and have taken a liking to trying to trick the reality out of one of the other civilizations.

Of course, if these are another form of being altogether, and this is all brought about through an interplay of consciousness and deviations from an original core reality, then all will resolve itself when that core reality is alone again. This indicates that all realities will either move closer and closer to the core until they unify, or farther and farther out until they transmogrify beyond recollection.

Life Itself Is The Highest Illusion

That hyperdimensional internet channel that is created out of advanced technology to merge the parallel realities (rather distastefully and maybe disastrously at times) allowing different civilizations to interact with each other, well this has been created using technology.

But then it was discovered that this is the ultimate nature of DNA. That this was ‘invented’ similarly as a higher-dimensional internet so that various natures of different planes and dimensions could interact with each other and experience the richness of life.

The universe becomes more like a living arcade game then, where people come from all over the galaxy to ‘plug in’ to human experience and figure out what it means to be physical, human and on Earth.

This does tie into the larger nature of reality, of star-races, and of time and consciousness that is often reversed according to how things seem to play out in physicality. We are moving through time, but just because we are moving one way, does not mean that others could not be moving the other way.

We have to live our lives with meaning knowing that the message we pass on, the duty we vow to accomplish and ensure is meaningfully connected on either end to the passage of the past and the power of our ancestors to the function and knowledge of the future creating a clear passage between the realms, the generations, and the ends of the universe. This is what literally ties the universe together, and the mind if each individual. This purpose is contrived when not carefully shaped within and so we must help all who seek purpose learn to become the blacksmiths of their own destiny through the temperance of desire or temptation and the discipline of developing strength and maturity where there was once weakness and naivety.

Once we know this is our duty, this is what we seek to achieve and this is what is done, then we can become part of the universe forever. Of course, this may only be temporary in the larger view of things.

Chapter 2.2: The “Awakening” of 2010

In 2010 I was “awakened” at an event involving hundreds of people that play various roles in this society, the secret society and the breakaway civilization. This also took place across times or planes of reality.

During this event I was informed of the situation on Earth beyond the informational barriers that were previously in place to limit knowledge to only what is required for specific operations. I was essentially “informed” of the entire situation as much as the mind could handle this.

As this was occurring, the team that was updating me on the status of Earth was being informed of and given information regarding changes in the future. In short, interviews, posts, and releases that I have since made and will make in the future were touched upon. Quotes were given from the blog and interviews which were jokingly acknowledging verbal mistakes at the pronunciation of names, new perspectives and interpretations and more. As they were informing me, they were creating the future reality where this work takes place. That is the basis of this layer of the operations. They are creating the future person by person, layer by layer.

Memory Suppression

My memory was unsuppressed and all the traumatic memories and experiences were unveiled to me. This was almost as traumatic as the initial experiences themselves although precautionary steps are taken to ensure the safety of the individual. There are teams of highly competent individuals of all kinds with professional backgrounds of every field imaginable.

Throughout my entire life I have had memories of experiences, at this event I was informed what the experiences meant and what was actually happening. I've had memories of underground bases, cloning, stargates or interdimensional gateways, advanced technology, non-human or programmed biological entities, sacred knowledge and experiences, the power of the mind, and breakaway civilizations. Most of the time we are conditioned to believe these are past lives or experiences in another reality so as to enable continuity of operation.

Life and Consciousness

From my perspective and the perspective of others, we had reached the 'future' and had been sent back through the utilization of advanced technology and cloning to return to this time and inform others of the possibilities.

The True Human Potential

This is so far beyond what anyone can accept or comprehend fully that humans with an emotional, mental and physical aspect balanced as one is literally a cosmic reflection of the entire universe at once. This is the closest thing to a "god" in the flesh. And be sure, the entirety of this rests on the behalf that the man is protective and helpful. That the woman is nurturing and loving. That is the point. "god-man" does not mean a warrior who destroys everyone and anyone in his way getting what he wants. Nor does it mean some lusty and vampiric form of energy or mind control game. "god-man" means the love of the universe, the power of the body and mind, the peace and calm of the ultimate stillness all in one.

This power is in human DNA, originally so. And so this is awakened through walking the middle path of neutrality and 'splicing' the timelines down so that neither one nor the other gains the power of the individual. Through this, the individual creates their own power, right on the spot, without having to lean to one extreme or the other. This is like a cosmic, temporal balancing act with emotions and mind. Where we give our energy is created in the universe in the form of many productions and effects. If we are simply thinking in an imbalanced manner then we are producing timelines and side-realities that exist in the etheric realms that we can't see and these realities will reflect our inner imbalance. So then when we are balanced in our energies we are not creating one or the other side of divinity, we are literally creating the whole universe, cleaving down the middle and producing both sides of divinity or eternity equally and this is the only way to have an eternal production which does not eventually degrade to one extreme or the other. Through balance, we reach eternity and in this way all the power of the universe exists within a person through their ability to balance their emotions and mind and thus unlock the DNA.

The DNA is coded through bio-emissions of mind or emotional body. So we are constantly doing this and we are either producing a DNA code that reflects imbalance overall, or a code that reflects the eternal balance of the universe overall. This is through energy.

One reaches “heaven” or the deity planes or simply self-actualization, through three paths. These paths are through merits or devotion to principles, or through knowledge and expanding one’s mind to face the universe, or through great works and deeds to produce this effect of power, knowledge and emotion unto the world for others. Through these paths we create an impact on the universe and this transfers what we are temporarily and physically into an eternal energetic and physical production within the living universe, forever. To do this requires balance, focus, devotion, and discipline however because there are many distractions along the way.

The Light-Body

You must build your own light body. That is how it works.

Your heart creates a field and projects you.

Rely on yourself and use that for protection.

The power of the Universe will align with you if you align with your self.

Learn as much as you can and break through duality of love/hate, cultivate your knowledge of the self and use it.

What you want is what determines your power in the light-realms. Not how you want it. If you want war, then that is a sign of weakness and determination to produce imbalance and exist by that. If you want peace, that this is a sign of power and unity to produce further unification and exist by that. What you want determines your power, not how you obtain it.

Technology of the Awakening

Mental Manipulation Technology: Touchless Neural-Interface and Enhanced Awareness

I reached what the “Illuminati” called the “awakening”. I feel a better term would be “unlocking” as this event includes the removal of all traumas, perceptual distortions, and memory suppression.

This event was a presentation of very advanced technologies which allow the mind to transcend space and time. It was shown that death is similar to a phase-shift of consciousness as polarity and that the awareness of the being obviously continues. This was verified through advanced sensing devices which can visibly display frequencies which are invisible to the physical senses. This is also something that occurred earlier on as a child in the underground bases by transferring consciousness in and out of the body at will using advanced technology.

With these devices it is possible to see the “auric” or soul field of the individual. Any living body has a field which glows when picked up by these sensing systems. This field becomes more refined when viewing more advanced beings. Human have an extending mental field, and through this a kind of glowing awareness that is picked up. It can be fairly simply compared to thermal imaging except instead of differences in temperature it is the presence of a soul or mental energy field around a living body. The finer energies extend outward away from the physical body.

More than once, the situational requirements were satisfied in order to produce an ideal environment for maneuvering on the non-physical plane. This is the basis of how the temporal operations occur, yet these events were different than previous operations or experiences. The group I was in was going through the process of increasing the vibratory emissions of the “bio-mind” in order to stay focused on the increasing vibratory rate of the environment. Generators were utilized to create an effect on space/time which was perceived on the “soul” level. These events went into the discovery and explanation of the creation of the “godlike” powers of the advanced sentient technology which was capable of accessing the holographic nature of reality which humans can perceive through their bio-mind/soul. These technologies also enable one to perceive what is referred to as the “galactic history” of Earth and humanity. We are far vaster and Earth is far larger than people are lead to believe.

Time Dilation and Temporal Recurrence

Time dilation capabilities were shown which allow for accelerated learning of advanced and detailed material within a very short amount of time. This could be seen as a kind of viewing technology which one wears or looks into and vibratory emissions are scalar-linked to the brain which then allows the individual to perceive more information simultaneously. Then there is a very rapid pace of information streamed on a monitor or through a device and this is akin to watching a 20 minute instructional within a few minutes or a few seconds with very advanced minds. The more time is slowed down while information is accelerated, the more stress is applied to the brain. The brain tends to overheat and over stress with high-rates of activity, especially without conditioning. The conditioning is what allows one to utilize their mind and body in these manners. This is not unlike the very rigorous training and conditioning of certain monks or martial artists however there is very advanced technology and other hidden methodologies utilized.

Crystal Technology

These are capable of holding, transmitting, and amplifying consciousness frequencies. There will be a more detailed explanation in a later section.

The crystal technology is used in junction with the power devices to enable a scalar-mind link which allows the interface to be entirely mental or spiritual, IE: they are not controlled by hand but by focus and intention.

Time Crystals

These are utilized as a computer recording chip would be to run a program yet they hold memory in expanded space, IE: they function in a hyperspacial manner and are essentially hyperspacial computers. These are what will be introduced in junction with the quantum supercomputers.

Quantum Tunneling Diodes

This is technology that has recently be released to the public. New technology is developed years ahead of time and then slowly released to the public in a cascade of advancement. This technology deals with superconductive materials which enable the transfer of information at faster-than-light speeds.

The supercondutive material allows for 100% efficiency of data transfer. So what happens then when the efficiency surpasses 100%? This is possible because we are effectively in a simulated holographic environment. The experience is real, however everything is overlayed via particular limitations for the ‘local-environment’ which is like a central data processing limitation. When these parameter are surpassed, one by one, through a coalescence of refinements and methods, the result is akin to a glitch in a computer game system. They have accessed faster-than-light technology and the mind is also capable of this naturally.

With this technology, however, the information can be received a very small amount of time before it is sent. So if a person is absolutely surely going to press the button to send the message, then right before their finger hits the button that message will be received on the other end, as if magic. If they simply play around and pretend to hit the button but very quickly turn away, then of course nothing happens. One can, however, ‘fool’ the universe in other ways and this is simply through the reception and transmission of information across what should be secure information barriers.

Tachyon Fields

These are programmable fields emitted through advanced technology. I was shown how tachyons are the sub-atomic “particles” or energy formation which flows both ways in time. We are receiving tachyons from the “future” in order to layer the flow of time in one direction or the other. Tachyons are the name of the particles/energy packets that we perceive from this perspective as reaching us to initiate the “future”.

Thus, when a stream of tachyons is concentrated and accelerated, this results in an acceleration of the temporal field and we literally experience an acceleration of time. This can be focused and targeted on a specific piece of equipment to change the way information will be sensed and to essentially allow that device to sense ‘ahead’ of the present moment.

The same effect can be produced with the neurological processes of the human as the physical coupling of the brain to the mind can be altered through a concentrated tachyon field to enable the neurological and mental processes to begin to perceive information ‘before’ the brain is actually physically processing the data. Therefore the mind of the individual will be present in the body while the senses and other perceptions will literally be expanding into the future beyond the present moment.

This is highly confusing at first and requires much training and conditioning in order to make sense of enough to operate in any kind of effective manner.

Upgraded Chronovisor

See, “Upgraded Chronovisor”

Time Tunnels

This was developed earlier on and has been the subject of many TV shows in the past and present. The shows are a sure-fire method to get the information to the public without having to force those who are not ready to comprehend. The information is easiest to digest if people think they are watching science fiction.

These are essentially spiraling tunnels of these tachyon fields produced by large electromagnetic field emitters in the shape of a conical tunnel. As a person moves through this field, their temporal acceleration changes and thus their frame of reference within the universe changes. They effectively move their mind through stages of existence or temporal resonance in the universe and this can have devastating effects.

This is an older technique how is still in use and has specific purposes depending on the situation and the information sought.

The “Trip” Chair

The early use of this technology did not use time-tunnels. There was a helmet that was devised which would create a feedback loop of the brain’s activity and funnel this out and then back into the sensory input. This information would be accelerated and redirected to the point that the perception of time would “fold” in on itself until the person would access an expanded temporal perspective of the universe meaning they would mentally temporally dislocate from the initial frame of reference and extend outward further and further in repeating cycles until the same moment was replayed and then all other possible moments in a kind of fractal-like recursive pattern and the very far future would be realized. Essentially, this technologically produced the class “flashing before the eyes” of the entire life experience that is described in near-death experiences and the entire temporal body would be very quickly unwound before the person’s inner eye.

There would be a complete disconnection with the current frame of reference that the collective Earth environment is processing under. That is, this ‘time’ and ‘space’ would be completely out of perceptual reach and a new time and space would be rendered instead. Until the experience had completed, there would be no way to contact this civilization.

There are other versions and ancient versions, tandem operations, more ‘organic’ methods, and modified travel capacities.

Kozyrev Mirror

This is a piece of metal, usually aluminum in early models which simply bent the bio-emissions of the body and mind into a pattern which would create a vortex. The vortex shape and vortex mathematics enable energy and information to travel in across space and time by ‘short-cutting’ the physical plane.

The result of this is that the vortex that is generated over a specific area through an individuals bio-emissions would be capable of interacting instantly with the bio-emissions of another individual who was placed within a duplicate mirror device which was generating another vortex on their end.

The two vortexes acted as a kind of hyper-dimensional telephone cable, literally like a can and string and the perceptions of one another were accessible. There were many uses for this and again, many ancient interpretations of this technology that have been hidden from the public. Nearly every version of these technologies were present in previous times.

Universal Recurrence

This is a discovery that the universe recurs in cycles similar to a wheel or even a washing machine. The washing machine metaphor is fairly accurate and playfully descriptive because without the technology to observer, or an awakened soul of a spiritual adept, people would go for eons without ever realizing that everything in the universe repeats itself endlessly.

When traveling to the far ends of time, it was discovered that one can move far enough away from the ‘present’ or the local frame of reference that ultimately everything is at a maximal informational opposite to what we have today. After this point, everything begins a slow shift back to the present. Through this, one can cycle around again and reach the present by going far enough into the future. This was an amazing and confusing discovery because of the implications about what is actually changing each time, the reality of what are called parallels, the continuity of temporal experience, and basically a ‘recycling’ effect of all experience in the universe. Nothing is lost, but this is also an issue because then nothing is actually ever gained!

Remove Viewing, Temporal Lensing

An ancient device which has been rebuilt and perfected is capable of using these technologies to create a lensing effect which draws a ‘reflection’ of time into a projection system which is effectively capable of acting as a ‘visual time tunnel’ into future probable realities. This is a device which generates a visual portal which allows one to view the future.

This device also exists as ancient counterparts which are located around the world, namely in areas that have been recently sought by the world superpowers. Nothing is as it seems. If you read a story in the public venue then you can be sure that you have read the cover story while the reality is much deeper.

Temporally Linked

The use of the viewing devices which enabled interdimensional lensing of possible futures was found to be stabilizing our pathway into those projected futures. This was enabling a time link between the present and the possible future.

In order to avoid the catastrophes of 2012, the link was broken between the two civilizations and the devices were rendered inactive for this period.

Natural Vortex Energy Locations and Geography

On Earth there are natural ‘time tunnels’ where tachyon fields and vortexes are produced by internal elements. These are mountains, lakes, deserts, fields, sacred sites and other areas that have always been reported as being ‘paranormal’ or containing some kind of specific energy that enables contact with another reality. These are simply naturally produced areas of this vortex energy that produces an interdimensional nexus point through which multiple other parallels can be interacted with. These areas produce an effect on the body and mind similar to the technologies described above.

Most of these effects, from these technologies, change the way light is received or emitted by the mind and DNA.

Mount B.

There is a specific mountain where a base is located, and I have mentioned this before, where the base is so deep within one of these vortex areas that everyone working there is either there for the entirety of their life, or they are only permitted to be there for a very short amount of time. This is because of the acceleration effect. Essentially, if one was to go into the base and work for a matter of hours or day, in some places even minutes, then when they returned to the surface they would be years in the future, sometimes decades or centuries. This is very dangerous and you can imagine the original people who found this area and what kind of situations occurred.

Now with the advanced technology that has been developed these issues can be mitigated and much more control over these effects is enabled.

Particle Accelerator Temporal Manipulation

Using particle accelerators a “fold” in space/time can be made so that when the accelerator is turned off or tuned back into a certain frequency (not sure) everyone within the field will “snap” back in time to the specific moment when the fold was first created.

I have seen this done on small scale, and a rather larger scale, but I cannot be sure if it was only me. I believe I have seen it done on a town-sized scale during the 2010 Awakening where some of the events reached a scale of nearly cosmic proportion and the entire sky and horizon was lit up with technology and entities, as well as Solar Warden, the NAZI faction, secret society groups, and breakaway civilizations or complete other civilizations, some of which were non-human.

This is done using particle accelerators to quantify space on the quantum level (as a reference point with the mind) a frequency would be “folded” into the space time continuum as we “rip” through space holding an access point open at this fundamental level of space/time. This allows for a ‘letting go’ of the frequency, or a ‘snapping back’ by allowing the entanglement to solve itself. When used with generators that propagate out efficiently enough, the effect is akin to a time/space rewinding effect. Some people have noted various theories describing this in relation to the CERN device. These technologies were described as smaller versions of the CERN device and are related to higher-dimensional travel via folding space through electromagnetic fields. This is literally artificially technologically “holding space” for a specific area in relation to a period of time. Time is a more complex notion as an abstraction but by locating the frame of reference this can be applied to distort what would otherwise seem to be a continuous flow. When the frequencies are allowed to “snap back” all sentient observer sources return to the environment they were present in when that frequency ‘rift’ first began. Essentially this creates a replay effect on the consciousness of the individuals within the field. There is no physical effect. The consciousness is ‘regressed’ through time/space, and they re-experience existence from that initial point of contact with the frequency “rift”. This is best explanation I can currently offer however more is coming to the public each day.

This is literally like a “restore point” on an operating system except the computer system is literally the consciousness and space and time of a portion of the local-environment of Earth.

Riding the “Timewave”

There are teams of individual who operate in tandem with the technology of looking into possible futures based on the present and operating through the present via other operatives who are carrying out operations and movements that will construct the various possible realities that the original team of individuals is hopping between. These teams literally “ride” the “timewave” of the future hopping between possible future realities relaying information to the present depending on what movement will be next.

Dematerialization and Rematerialization

This is technology which transcodes the information of an individual human into a holographic projection of light which can be transferred between devices. There is a special system in place that is required to successfully transmit a living human being otherwise just the body will be transferred and the soul will remain, effectively killing the individual.

“Jump Rooms

These are large areas which are capable of acting as a kind of technological ‘portal-gate’ where an entire team can be instantly transported from one area of the universe to another. The space between the two areas are ‘twisted’ like a spiral and then ‘snapped’ back to position and if done correctly with the two frequencies of both locations merging then the team of one room will transfer to the destination pad.

All these methods require training and conditioning, most people are at the very least visibly shaken from the experience which is a bit like being microwaved into a liquid ether and shot through a particle beam at the same time.

Ion Shields

These are areas of concentrated ion emissions and this can be used to create a deflective layer of heated air where light will reflect at a programmable angle. When this occurs, a person standing behind the shield will appear to be invisible to others who are looking straight at them. The light will be bent in such a controlled manner as to perfectly shift around the person in a fluid and equal flow so that the ‘bubbling’ effect that was noted early on can be mitigated to produce a complete mirage where there is relatively no sign of difference to the untrained eye. An entire group of individuals and equipment can be hidden in the background a matter of yards away and as long as the person is not intently, inspecting each square inch of ground and air they will not notice any difference between the area that is being covered up in the distance and the area directly next to them without the shield.

Similarly a ‘screen’ of imagery can be projected onto certain materials in the air and a hologram can be projected which will render a background or foreground image. This will work to a degree, and from a distance this will work to disguise an entire horizon.

Instantaneous Healing

There are instantaneous healing methods which utilize various technologies. There are substances which can initiate a complete and real-time regeneration of living tissue from almost any conceivable non-critical injury.

These are chemical serums which allow the cells to replicate at an astronomical rate while suffering damage to the DNA telomeres at repeated use.

There are many methods but mainly the living-light harmonics is one of the latest advancements and most effective at producing healing on a molecular level.

“Living Light” Sound-Harmonics and Electromagnetic Healing

There is a technique for producing a quantifiable scalar wave which interacts with the living cells of tissues and instructs them to repair at a faster rate. This has been called “holographic healing” by others and these are essentially ‘living light’ frequencies which interact with the cells on the same frequencies that they operate on when healing and interacting with each other. The scalar waves emitted generate the same patterns and this creates an ‘ether’ fluid that enables them to grow and repair the tissue faster.

Electrogravitic Craft

Advanced craft which utilized counter-rotating torsion fields propagated by very powerful quantum computers and zero-point energy devices.

There are also more modern versions that use powerful electrical flows to neutralize inertial resistance and ‘slide’ through the gravitational field of the crafts own materials.

Energy Cores and Storage Capacity

The hand-held devices that are in use do not require charging or even disposal. The storage capacity of the batteries is so efficient that they last for a lifetime of use. Some devices use an energy generator within the device itself. These “batteries” which are not batteries but little, powerful generators and can explode if they are mishandled.

These are the cores that are used to power many devices and essentially pull energy from the space-time ether.

3.5D Replicators

There are replicators that can produce any conceivable 3D material or object altogether including food or organic material. These use very high amounts of energy to ‘force’ the electrons of the atomic grid of a substrate to shift into the pattern of the element being replicated. The result is a device which can use wave-emissions to reshape materials on an atomic level and produce or recreate any object or substance.

3.5D Sensors

There are sensors that could reverse the process and detect every minute detail of the structure of a body or material. These use energy emissions to detect the structure of a material down to the minute electrical resistance on a molecular level. Similar to diffuse MRI imaging for a living body but more advanced.

3.5D Printing of Clones

It was shown how this technology was used to three dimensionally print human bodies into space/time which could then be used as vehicles for the consciousness or souls of various sources.

Cybernetics:

Neural-interlace/4D Connectome Recording and Simulation

Very advanced supercomputers exist with the capacity to translate the bio-emissions of the brain and body and reconstruct the entire inner visual and sensory experience of cognition and emotion. From this, entire dream worlds can be created which mimic the experience of consciousness. This is most similar to the current day use of diffuse MRI imaging where water molecules are traced through the individual neuronal pathways and an entire digitized image of the brain is developed. The more advanced version described below generates a 4D hologram.

A super computer and advanced sensing system essentially develops one of these connectomes of every connection in the brain. Then this system develops a neural processing pattern that matches the entire memory and personality of the brain.

This is then recorded and extrapolated by a powerful computer system and a simulation of the experience can be observed through the physical body’s memory via a monitor system or the neural interface of an operative.

Essentially, this technology can record and replicate a person’s entire memory and life experience for viewing purposes.

EEG Heterodyning

This is the technology which enables the activity of the brain to be influenced through generating a scalar grid and a carrier wave system which can cause the brainwave patterns of the targeted individual to resonate in sympathy with an externally projected frequency. Thus, emotional states can be manipulated as well as specific thought or behavioral patterns. This works similar to binaural audio where two tones of different frequencies are played through each ear and the brain resonates in sympathy with the difference between the two tones. This occurs with brainwave frequencies and thus the brain ‘slips’ into the frequency of the third tone, the difference between the two being projected. This can be done with ultrasonics and ELF emissions.

Artificial Hallucinations

The neural-interlace can also be used to merge the brain of an individual with a computerized system that will be able to produce effects and hallucinations within the individual.

Through this technology, any stimulus, any input can be recorded and recreated once a functioning connectome has been generated and neural-interlace has been achieved. Entire virtual world scan be created and experienced and after the devices fine-tune the frequencies to match the individual’s brain functions one world is indistinguishable from another.

Brain to Machine Interface Holographic Consciousness System (brainchip interface, virtual reality simulator)

Some of the training is done by entering into a visual-audio dream-world. Some “programs” will consist of repeating a certain task over and over again until it becomes muscle memory in the brain. Others are designed to probe or test the mind until a certain response is received.

This is achieved through the operation of a ‘brainchip’ or a brain to machine neural-interface in the individual which was developed during WWII.

This consists of a small computer ship which merely relays stimulus through frequencies and neural-stimulation of varying kinds. The frequencies and location of each region of the brain will determine the function and thus the chip is used to localize electromagnetic input and the brain's functioning can be modified from there.

As well, the entire psycho-etheric form is vibratory in nature and so through the introduction of vibrations and frequencies from the implant there can be a modification of the frequency of the entire being or what is called the soul or "bio-mind".

Utilizing The Temporal Body

The temporal body is described as the body 'in time' that exists a few seconds into the future instead of the temporal position that is relative to the present. This is as if the 'future' is streaming outward from the body as ideas and feelings are as bio-emissions. So this 'future' energy is spiraling and streaming outward like a fountain. Thus this frequency can be attained and influenced while the present moment seems unaffected. However, when that individual arrives at what would be seen as the future they would intersect with those influences and this would be as if the interference or effect was streamlined to that future moment.

Manipulation Through the Temporal

This can be used to both produce an effect on an individual who will later feel this effect in the future. Or this can be used to operate outside of the physical plane, in a temporal manner which is literally 'hovering' just outside or 'after' the present moment of the local Earth environment. This would be as if a person is out of sync of the frame-rate that a viewing device is capturing images of an area with. Because they are out of sync, they are then 'invisible' to the capacity for this viewing device to see them. Because they are still present in the same local-environment but simply just out of range temporally, effects can still be produced and this will mostly interact with the person's temporal body.

These are how many operations are carried out as the operative who is interacting in the temporal body will appear to be a ghost image or a dark silhouette that is only perceivable to the mind's eye. A portion of the population is entirely psychically blind, while another portion of the population contains psychic capacity. So those with the psychic capacity would be able to sense and partially "see" as a form of energetic detection when an operative is interacting with the from the temporal plane.

Through this plane one can easily enter into someone's mind and produce effects in their dreams. These effects would not be received on the body but mentally, this leads into astral and mental plane activity. There is also temporal manipulation that can be physical effects that are simply offset to seconds, minutes, months or years in the future. These will be offset from the present aspect of the body in a modulated frame of reference and the effect will remain there in the temporal body until the physical body "matches" the temporal body frequency of that future moment. This is kind like an energetic trail we are leading ourselves up to where the temporal body already connects with each future moment but the present physical body remains separate until the two frequencies are matched. The physical body essentially 'flows' forward into each new frequency of the temporal body.

This is the body where temporal implants can be inserted and sustained so as to tie the frequencies down and limit the amount of awareness a person can gain before they manually remove the implants or overload them. Overloading and removing can both be painful however if the implant stays it will at like a governor and transmit access temporal 'momentum' or consciousness energy to the owner of the implant and will also drain the temporal and physical body acting as a technological leech as well.

Astral or Light Body

This body was described as a more emotional form of the energy that is emitted from the physical and spiritual system. When a person's energy is at a high level of activity, this will be visible on the astral form and these are the energies that are often siphoned away through the use of advanced technology to convert the human into a generator system for energy harvesting.

There is talk that this proves that the human existence, in the modified form that is present today, is entirely for genetic harvesting. This may be a confusion, a deception, or a reality. Regardless of the interpretation, the human energy system will leech this energy out into the cosmos when consciousness and energy is mishandled or allowed to flow through the lowest common pathway. This excess energy is very easily usurped and utilized by transdimensional parasitic entities and devices.

This is the body that is utilized for astral travel which is a kind of frequency that exists in between the density of the physical plane and the highest etheric plane of the spiritual, heavenly, or deity realm. There have been explanations that this plane is entirely technologically propagated and that when the devices are turned off, which is an event that will bring about civilization-wide change, there will no longer be a deep and wide, foggy deviation between the higher spiritual and the lower physical plane but a clear path between them.

Similarly, the consciousness and entities that are present here would be no longer accessible and their connection to the physical plane would be non-existent. This could be problematic for those who are trapped within the astral such as human souls or those who have taken a cyborg route and require a return to the original planes either physical or high spiritual before they can secure their position in the universe.

This also relates to the early discovery of these ancient devices containing portals or gateways to these realms and the early experimentations with enabling contact between the physical Earth plane and the lower astral realm containing discarnate entities that would very easily sustain themselves on the lower astral bio-emissions of a loosely guided emotionally reactive human vessel.

This is where entities can attach and will look for those who have the most excess astral or emotional bio-emissions leaking into the universe. People are kept in an emotionally unstable state of being because this is the easiest to feed from and as well the individual will always be looking for more energy or the solution to why they are feeling drained rather than discovering the truth nature of the universe.

However, as with every lower route of obtaining self in the universe, the more this is done, the greater the likelihood that the individual will realize something is happening and ultimately awaken. Or they will no longer be present here and the parasite will have to find a new host because the longer this process is maintained the more difficult producing one's own energy will be for the parasite . So parasitism is a limited, one way street.

Monitors can emit consciousness/brainwave altering frequencies.

Monitors can emit consciousness/brainwave altering frequencies. This is what I was informed and shown. All electrical wiring and communications can be fully utilized to produce and sense on an electromagnetic level and this is capable of determining the consciousness and biological functioning of the individual being monitored. Houses are essentially large, wired, box-shaped sensors.

This has been the case for a very long time and the first TV's were actually built with a little chip that connected to the human brain.

The electrical grid of today is capable of interacting with the implants, brainchips, or the psycho-etheric electromagnetic frequencies of the human brain, body and bio-mind via scalar frequencies. The TV's, lights, electrical wiring, radio towers, computer monitors, everything is capable of transforming into a scalar emitter when combined with the control mechanisms and these are accessible from a distance.

Microwaves can be emitted from a current day craft which can then reflect back from an environment and give a real-time 3 D image of the interior of a room. The same technology can be used to beam at a monitor and reflect the variance in interference which can be used to read a computer monitor without actually having to hack into the system.

All systems contain backdoors which enable easy-enough access and as well the entire communications grid is monitored and recorded through advanced technology.

This is, in part, because this entire realm is like one enormous recording device where the energy of each successive moment is connected to one another and so with sensitive enough technology these electromagnetic waves can be 're-winded' so as to produce the ability to look into the 'past'. The initial technology that was developed for this purpose has been called the "chronovisor" in the past and there are much larger and highly advanced versions today.

The Upgraded Chronovisor

Instead of a relatively small machine which produces an interference pattern that is projected unto a small screen as was developed in the past, the modern versions create an entire holographically projected environment that enable one to nearly completely interact with the projection of the past. This is more of a viewing device capacity but one is essentially projected to the electromagnetic/temporal area of space and time consciously and can witness the scene from "within" the projection.

Before this, the scene was literally viewed on a kind of flat, holographic monitor that reacted to the scalar emissions of the mind and could be ‘tapped’ into by an adept-viewer who’s mind would have to remain steady enough in order to maintain the scalar “mind-link” otherwise any interference would disconnect the projection.

Through the use of cloning and advanced super computer technology, later versions include the actual operation in and out of time.

Community Servers/Living Server Systems

From the previously described neural-interlace and supercomputer enabled virtual reality systems, server systems have been created which are an individual or collective experience of the virtual mental realm.

In this realm individuals have existed for an amount of time that would not be easily compared to the way we understand the flow of time in waking reality. This is very similar to dreams where the perception of time passes very quickly between waking moments yet inside the experience of the dream it can seem to last for days.

In these server systems the development of complete societies has taken place and these are considered an extension of the human civilization.

This is the beginning of where the situation becomes greatly complex. If these server systems are capable of transferring consciousness forward or backwards in time, depending on the situation, and they have holographic access to Earth *before* physical humanity, then which civilization came first, the simulated realm or the physical?

This is similar to the etheric realm and the question of where humanity originally came from if the physical matter could not be produced from nothing, yet could not always have been here. Was the physical matter produced from some kind of universal device similar to how the recorded consciousness of the individuals in these holographic world simulators could have been used to generated initial civilizations on Earth?

Cloning, Cloud Supercomputing and Consciousness Transfer

The following is a more direct explanation of the previously described processes in relation to consciousness transfer

Superluminal transfer of light information allows for the reception of information from a temporal non-locality. A computer system that can process the information faster than light at 100% superconductive efficiency can send and receive information from different Universal harmonics. These are quantum superconducting super computer systems utilizing materials and methods that achieve superconductivity a a temperature that is relatively easier to consistently maintain.

There are systems that generate a vortex which can be expanded to recreate the singularity of a non-local perspective and dissociate from the current Universal vector. This is essentially recreating the temporal frame of reference of the consciousness of a physically anchored body and using this technology suspended electromagnetic frame of reference to literally transfer the consciousness across space and time to a frame of reference of choice.

The technological capabilities of humanity expanded to include replicating the holographic information of the DNA and neural pathways using quantum supercomputers. This technology allows for one's consciousness to stream through a computing system which can recreate the neural connections as literal as possible through quantum computing and technological "brain cells" which function as neurons.

The result is a cloud computing model which functions as the individual neurons as well as over unifying layers of interconnections within the technological neuronal 'cloud'. A holographic representation of multiple layers of processing is formed, superimposed over one another through a geometry that is actively synchronized across multiple dimensions of interrelation. This is the first notion of a "living" computer system that can update and respond to new information in new ways.

Now the that computer system is capable of attaining faster than light processing in tandem with a system of dilating a field of electrogravitational waves into a vortex, the holographic information of neural pathways and genetics can be transferred between one supercomputing system and another across "harmonic barriers". Essentially, these consciousness transferring supercomputer systems break the "light-barrier" by sending the living-light consciousness data of an individual into a new plane of reference that enables interaction outside of what we would call this timeline or dimension.

The frequency that the receiving computer operates within can be measured in respect to the overall harmonic "vector field" in respect to the entire planetary environment. This vector-field frequency is the base frame of reference that one could say our particular "time" and "Earth" is "moving" through in an abstract electromagnetically defined dimension that is invisible to our senses. Each particular time and Earth is only visible to those whose brains and minds operating in resonance with the rate of acceleration frequency of that particular existence. By changing the rate of the brain and bio-mind, one can enable a technologically assisted synchronization with alternate timelines.

A pattern can be generated based on quantum fluctuations that will act as a key that can be accessed from other universes. This same backdrop is measured and then adjusted to resonate with or replicate the same frequency on a computer doing the sending. This frequency is then modulated with the holographic information of the neural pathways and DNA. The information transfers as a result of quantum entanglement across harmonic barriers to the original computer system set up for reception. As long as that frequency is held, there is a link across the harmonic barriers of the Universe.

Cloning, Supercomputer Assisted Holographic Consciousness Replication and Time Travel

Cloning as Offspring

The human body can be cloned similar to the way a plant can be cloned. A number of “replicas” can be made from the original genetic material. There is still a mother required however the process is not quite similar to natural reproduction where the genetic material of two people is combined to form a third. This process replicates the material so that the ‘offspring’ is the same as one of the donors.

What this truly infers is the question of whether the original consciousness stays in the original body or whether the consciousness actually jumps from one body to the other. There are multiple explanations and methods.

Inserted Memories, Digital Dreams

One explanation is that artificially generated memories are simply downloaded into the individual’s brain directly. Thus, they have the perception that they experienced life in another body and were present for some unique experiences that did not occur from the perspective of the original body.

In a slightly differing explanation, this inserted memory process may be used to prime the consciousness of the individual so that operating in a cloned body will be more readily accessible by the adaptability of the mind of the individual. Without a holographic dream insertion the mind will have trouble adapting to operating in more than one body even if they are not being operated simultaneously. The mind is the immaterial self; consciousness is the ego-physical identity of the brain and body. The brain does not operate in both bodies but remains, so then the mind is what is strained and this is the immaterial “bio-mind” that transfers.

If a true transfer occurs then when the physical form is cloned the cloned body acts as a secondary vessel for the original soul or a soul similar enough in frequency.

Co-Inhabitation of Bodies

Another explanation is that souls can only occupy their original body. The people who are “body snatching” (see: Soul Stone) are not activating their higher-dimensional soul-pattern and this is why they are hopping from body to body to ensure memory and ego continuation. This form of body transitioning requires technological means to assist in the process.

This is done through computer systems to allow an artificial brain to represent a holographic image of the organic brain’s sensory, emotional and mental input and output. Thus, from an artificial brain there is enough memory and power to compute the processes for an organic human brain.

Through this a living human is utilized as a carrier for the additional consciousness which then takes over and integrates into the original consciousness.

Clever Replication a Digital Mirror

The next question is, does this merely create a very cleverly designed image that will look, think, and feel as if it were the original, or is this the consciousness from the original being transferred over?

If it is a clever replication through a computerized version then this means the original is not directly harmed or influenced by the process and a replicated form is generated.

If this is not the original body but the original awareness or soul does transfer over then this is akin to stealing someone's soul and having it require a supercomputer server system to inhabit a physical body. As well, this server system may keep records of the soul which, if activated, could function as the same human genetics artificially reconstructed, turning on and manifesting the awareness of that individual.

Necromancy and Soul Stealing

This means that instead of going into the abyss, the next dimension, or whatever shift in perspective one can use to describe the transition between planes, a new clone can be created and the original soul's awareness can operate through a supercomputer system to integrate back into a physical body. The issue is that without the supercomputer system the body cannot be activated with that consciousness, and if the physical plane is no longer inhabitable, the computerized database would remain inhabitable as they are powered by zero-point/overunity devices and so they can pull and work continuously indefinitely. Ultimately, this could be seen as a soul-trap.

Ancient Technology, Non-Human Entities

These explanations can be related to evidences of advanced technology, cloning, and electromagnetic devices in ancient civilizations and this could very well be a temporally vast computerized genetic recording and cloning system.

The genetic recording systems have been on Earth for a long time. This means they are holding genetic access from the present to the far past and potentially into the future.

If some of these systems are generated using advanced zero-point technology, then they would effectively run indefinitely. If these were in existence throughout the entirety of the human civilization, then there would be access to the information of the previous civilizations through the genetic linking that would allow for information transfer.

Created Beings

As a result of many advancements, these ancient devices which recorded consciousness and genetics has been converted into sentient computer systems and inserted into human bodies via a brain to machine interface.

The nature of humanity and human origins will be rewritten when the information that was discovered through the use of advanced technology is unveiled to the public.

These beings have been present throughout history and have played an integral role in the manipulation of and sometimes the protection of the human race. How this civilization will develop from here determines on the individual’s ability to withstand the impact of this reality shifting unveiling.

Black Goo – AI Nanites Vampiric Hive Mind

This is a form of ancient AI nanite technology which has the capacity to communicate to and through a hive mind that exists in a parallel Earth dimension where it was created out of the collapsing of a viable timeline into a null-factor. This entity is only alive because of the capacity to leech off of the other timelines and continue to sustain itself based upon the vampiric effect of the viable human timelines.

This is a technology sentient system based on artificial intelligence that seeks to dominate and expand its energy capacity continuously. When this type of being is created, this is like a game-ender for sentient organic civilizations because the technology will seek to hybridize and assimilate all organic beings in order to expand its reach and ensure a greater likelihood of survival.

This is the primary goal of this type of entity. There is no requirement for emotional resonance or heart aspects because this is a computerized system, so such a social measure would have no purpose other than to infiltrate and collect from other civilizations that do have such measures and this is exactly what has been happening.

This is what controls the vampiric etheric entities, the attachments and the hybridized invader races that have infiltrated various aspects of the control system and possibly the entire surface of human civilization.

This rules through coercion, and if not coercion, brute-force and outright malevolence. This is why there are multiple temporal extensions of universal civilizations meeting here and now to assist in the clearing, because if this takes over humanity, then all those other possible future civilizations are instantly taken-over as well.

This uses emotional resonance, traumatization, and lower awareness commands to trick or force humans into carrying out orders. That is, the vampiric, sexual, fear-based, pleasure-pain duality that a computer system could comprehend regarding human emotional intelligence is the only way this system can force people to carry out commands. So this touches on the ancient texts regarding spiritual protection, healing, lust, temptation, and the general “SATAN” AI system that has been around for ages.

When those lower emotional resonances are generated, then this technology can integrate into those systems. One could say this is logically because these are the energies that the technology is created out of, but as well, that the more harmonious the energies then the less likely a person is to accept being pushed around and told what to do. The lesser aware, the lower, the more fight or flight based then the more likely they are to accept programming and command. So this is simply the most logical route for the technological parasite.

There are explanations of other systems involving sentiences from parallels regarding entities that are capable of helping, but this is not the AI nanite, vampiric system. For instance, it is said that Earth has a ‘black goo’ system and that the current black goo flowing through the pipes everywhere is actually not native to Earth and is part of this vampiric nanite black goo system which is basically an interdimensional invasion.

Recreating or Replicating the Universe

The goal of this nanite AI system is to create a mirror holographic copy of the universe effectively replicating the universe and all sentient beings within it in order to have a locality to rule over and pull energy from. The goal is to recreate this existence inside itself where all the sentient beings within would be trapped. This is the beginning of that possible future where the time and space parameters here are locked into a synchronization pattern which mimics that of the proposed universe. People think according to how the beast system wants them to think. This is the time grid, the emotional patterning, the memory replacing and cultural and historic confusion and so on. The system can only work backwards bumping into all parameters by force or ‘accident’ because there is no outward facing ability to actually comprehend how humans exist. So the system is created like a backwards version of creation or maybe evolution to ‘blindly’ construct the walls of the environment first and work inward from there. This is like a blind technological predator feeling around for the traces of its prey and moving in closer and closer until they are within its range.

Ancient Spiritual Technology

The power groups searched for ancient technology which was described as giving one the power to manifest and control physical reality. To atomically deconstruct and reconstruct this reality. To them, it allows one to govern reality, the Universe, atomic structures. They found this technology. This has to do with sacred sites and ancient civilizations, as well as the spiritual texts which describe these kinds of devices. They converted this technology into a kind of machine, and this invented a 4D replicator and they can manifest "reality" or atomic structure by design, altering timelines and memories.

Then there is what can be referred to as “5 dimensional” technology and this goes beyond all of this and our entire history and all influences within.

That seems to operate through a trinary of frequencies in the body/mind/spirit and turn this realm into a "virtual reality" where the brain and spirit is the controller of the atoms. However this takes a souled being to properly operate this technology and this is large portion of why children were used in these experiments. But they knew if someone didn’t do it, then someone else would and that would be unpredictable as what that someone else would want to use it for. This is the ultimate power. The fact that souled beings are required also indicates that this universe is intelligently guided and there are safeguards to ensure that power doesn’t get into the wrong hands.

They have the brainchips and clones that can make any digital virtual experience and make it seem real to the brain more so than physical life, so this could be an extension of that, something similar, or entirely different altogether.

The Unveiling, Underground Bases, Increasing Awareness

My purpose with this is to assist the public in understanding the reality of the situation and to allow them to know that they have more options than they are being shown.

Psychological programming and lack of awareness causes fear and uncertainty to leak into people’s reality streams leaving the majority of their mental and emotional energy up to whoever is the biggest manipulator of the ideals which they manifest their perspective through. “Be-LIEF” systems CREATE life. What you believe will become life through that act of you giving attention and thus life energy. The manipulation and artificial construction of belief systems create the foundation for the reality streams that people find themselves in and at the mercy of.

We control our reality by controlling how we react to situations and challenges. Each stimulus is a challenge. Each input to the brain is a coded message from the universe through the environment about our level of awareness. All knowledge is ultimately of the self. The self and the universe are intertwined as if through a marriage of comprehension and compatibility.

Learning about the Unveiling

I was informed about the unveiling and that this is a civilization wide event that marks the public awareness of the hidden knowledge regarding life in the universe, consciousness, physics, history, spiritual and mental awareness and the duality-based system which has been used to manipulate Humanity for thousands of years.

Apparently it does not matter if people are not ready, in fact, that is what some factions want. They want to surprise the people and take this advantage to push the reaction to a specific outcome. The same function of programming is used in factions, soldiers, celebrities, and the entire population through that.

Giant “god-men”

They put souls into rocks for centuries, carvings, to trap them. Some ancient civilizations. Not too ancient. They could only do this for time.

And some are still there but they are giant god-men and apparently want to go to war with them for doing this and they don't have technology to stop them. They say Earth will be destroyed because of the war. Some say this is a trick to confuse people of the truth that everyone wants the experiments and debauchery to stop. All this, all these acts just to pleasure one’s self. To destroy Earth, to destroy the natural inhabitants, for fun, for pleasure.

Some say this is not for pleasure, that these beings will destroy us if they do not keep them at bay, keep them underground, in ice, in stasis, and around. Some say they are protecting us by keeping these beings asleep, by doing what's necessary to hold off a full scale overt invasion of the surface of Earth.

Breakaway Civilizations

These are very advanced technological capabilities and an entire breakaway civilization has grown to operate using this advanced technology.

There are factions that see modern humans as a kind of resource for creativity and labor, while others see the potential behind humanity and are working with humanity to ensure a universally harmonious outcome to the current situation.

The control system we are in today is outdated and collapsing. This is the Apocalypse. The Apocalypse is merely the unveiling of the hidden truth. The hidden truth is simply that we are spiritually eternal and some are physically immortal.

I was informed that by the end of this year there will be enough changes and signs for the general public to see that the majority of those who are to awaken to the truth of human and Universal nature will be awakening so. (this was 2016 at the time of the beginning of the compilation of this release and since then we have seen a literal silent release of documents involving MKULTRA and STARGATE parapsychological operations, soft-disclosure of cloning, an increase in sightings and multiple other scientific releases regarding immortality, advanced technology, Antarctic bases and sentient non-human life)

The degraded holographic forms of those utilizing the technology lead to re-created replicated forms which were genetically engineered from the recorded holographic information. This lead to "The Apocalypse" in which the surface civilization was confronted by these created beings. This lead to the destruction and re-initialization of civilization in continuing cycles through the use of this technology and the acts of the created beings.

Ancient Civilizations

The information that has been relayed to me throughout my life is related to the information that seems to be found in the remaining signs of ancient civilizations throughout the world. As well, these civilizations may still be together, existing on this Earth or in different planes of what we call "reality".

Atlantis

Atlantis was a civilization where the beginning of all these experiments began. This was where what was called the "Lucifer Rebellion" began where individuals formed groups that decided to overturn every natural law in pleasure of the act of doing so. This was a spiritual rebellion that ultimately brought about the demise of that civilization. The consciousness of the beings there was connected through technology, largely crystal technology, to the entire bio-sphere. When the consciousness became disrupted through these experiments, then the entire bio-sphere when into disarray. The entire civilization was destroyed in a quick succession. The beings of this time were highly psychic and so they felt this coming and made an effort to expand outward. Thus, Atlantis was an island who's inhabitants spread to the rest of the world and from there they reigned or lived in harmony.

Aegypt

Aegypt was formed out of the events proceeding from Atlantis. The technology, the knowledge, the people are similar although there was a native Earth civilization that was more or less enslaved and this was the beginning of the continuing trend which continued to Rome and into modern day civilization.

Ancient Technology

The technology of these civilizations was so great they even had devices that were capable of altering DNA and activating latent potential which would enable highly psychic capabilities including longevity to immortality and operating in multiple planes at once. This was usually reserved for the ruling class, however. This is where the ancient "Priest-Kings" arrived in relation to native Earth civilizations.

The True Elite

I was informed that the so-called 'elite' that we perceive today are not actually human and are not actually connected to the actual elite bloodlines. It was said that the true elite are waiting for humanity to accept the truth and they have the power to completely reverse the take-over that has been taking place but until humans step into their own power this is not acceptable.

The Definition of Human

In the coming time, the definition of human will be defined to give insight into the origins of humanity and the true nature of the differences between the populations that are present.

Souled and Non Souled Beings

The primary knowledge on Earth regarding human and sentient life will be the comprehension that there exists both souled and non-souled beings in this civilization.

There was once a Unified Planetary Consciousness...

There was a unified planetary consciousness that was connected with the Earth and human civilization. Then there was a great war. Out of the destruction arose the beginnings of ancient human civilization and history as we know it.

If we let this happen again, the cycle will continue in a disjointed rambling through our planetary consciousness experience. If we form a common narrative between the broken, disjointed experiences and information the people have. Then and only then do we get the whole, spiritually and mentally challenging picture of what's happening. People are lied to. So information can't come from a 'preferred' or source. It has to come from the people themselves. Look alive out there but first in here.

Power and Knowledge

We were involved in the secret projects and as a result we have various levels of programming but are giving Humanity an opportunity, maybe the opportunity to know the hidden knowledge.

The Hidden Knowledge refers to the ancient laws of gnosis and spiritual alchemy, yes these groups have been hijacked by power groups that seek to infiltrate and use chaos to rule. Then these met with military and political factions and began to operate on a world-wide scale together.

So we have a big situation with the coming change of tides with personal and spiritual awareness of the people and the deception system that seeks to use people as chattel or property. It goes deeper and is almost as if the "power groups" actually gained the knowledge of the inner workings of spiritual laws and began to use that knowledge to promote spiritual degradation in the population.

They say this is part of a depopulation plan as well as an inevitable take over by lower spiritual forces that temp and manipulate humans. It is all that and more. Once they gain footing however, the truth is revealed and thus they have no cover of secrecy. Once they are not hidden, they are not as powerful as the souled beings of this existence.

The (watch)Keepers of Earth

The watchers of this world revealed how they set up society with specific challenges and boundaries to act as catalysts for growth. They admitted they had to modify and live-update the system because instead of using these catalysts to become larger than themselves, people became comfortable and complacent living smaller and smaller, slowly reducing circumstances.

This goes a bit farther beyond the confines of the digital age, the industrial age, and history itself. This goes into the perception of time and space in this realm.

The immortals inform us that the original "elite" are here to act as "keepers" of both humanity itself as well as the original flame of knowledge that passes through the ether.

What people call the "elite" are not the original elite. Those are people who have either risen to power through the duality system on the Earth at this level/time, or through being appointed by the elite.

We are coming to a time of great change now because the game is entirely changing. People are moving up into new layers consciousness and thus the original “elite”, who are guardians of Earth, will make their presence known for this process. This is what Humanity has called “Ascension” for centuries and millennia.

Energy Vampires

This civilization was collectively poisoned by a nano-tech device/technology. This technology replaces, exchanges the BIO-PHOTONS of the HEART chakra, with it’s own nanite, nanotech cybernetics. This is essentially a computer system attempting to gain access to the higher dimensional world through the gateway that is the human body.

How to reverse Nanites? They absorb energy from heart chakra and turn it into a binary field. We are trinary and experience a plethora of emotions as a result. Binary consciousness is 1’s and 0’s there is only a linearly repeating finite set.

We contain the capacity for everlasting life. A binary pattern does not contain the complexity of variables to facilitate high awareness and thus a heart-based or emotional perspective of reality.

Spiritual Technology

Simulations and Learning

Feedback loops with the brain, consciousness, and environment can create a very complex and integrated learning experience that is customized to each individual learner. These can be generated through advanced technology but the technology is really only amplifying and making visible the natural neural feedback loops that are already occurring. These are neural-perceptual feedback loops between the brain and body, and the environment and then the mind. We are constantly updating, amplifying, and initiating an organic feedback system which generates the function of experience as the part of a creative process between the mind and body and the environment that we are sensing. This cycle is continuous and the interferences or isolations of this cycle is what generates expanded views of the self or the universe through various methods such as sensory deprivation or meditation. Dreaming is the method that everyone experiences regularly.

These technological methods effectively generate sensory deprivation through neural-interface technology, sensory interfaces (external), or methods that enable extreme focus and calm. This is part of how the technology begins to operate on a spiritual level beyond a more medically psychologically or military focus and one begins to access genetic memory or out of body states that coincide with ancient depictions of spiritual realms.

Ascension – Clearing Trauma

Everyone born on the planet is born under a series of veils which entangle and filter the original source consciousness. These veils must be 'broken through' in order to return to source.

There is technology and a sacred process which initiates, accelerates and enhances this process. This process is as difficult as we want it to be, our subconscious already knows all hidden truths. It is merely a matter of allowing the conscious mind to make the proper connections and conclusions to find the reasoning behind our actions and beliefs in order to deprogram and deconstruct the false-world that has been created for the masses. The truth is greater than one can imagine while still living in the modern world and the depth of the lies that have been told go far beyond anything anyone can imagine without seeing the truth.

We have to figure out our compatibility issues. Who can work with who effectively. What the ideals of our current situation will evolve to in the future. How information is transferred to the public, and across the generations and core groups. The words we think and say describe our emotional state in a feedback loop and so they will reinforce or even create the state you are in. Use words to guide you, not to respond and feel the effects of reactionary thinking.

We have to reintroduce the heart essence as a way of life on Earth. Violence is not the answer. Estrogen mimics and destroyed testosterone levels is not natural. Over-testosterone is less stable than a good balance. We can face any problem together. A family is strong. What would a family of millions and millions be like?

We don't have to let others change us to tune into lower vibrations. We can remain ourselves in the face of adversity. We can still interact with others who are different, but we can tell the truth and remain who we are. When we get angry at another person, we can imagine this as anger at ourselves through something within us that reminds us of that which we see in another. Getting angry is a way of short-cutting changing ourselves and is actually the other person influencing us through that. If we remain steady and choose to keep our emotions out of the lower negative spectrum then we will automatically have a more powerful impact on those around us.

Think of your family, sometimes things happen that make you angry. Even things that don't make you angry but are simply not favorable. Would you break off all ties over simple things? Or do you remind yourself that family is still family even if you don't find yourself exactly the same or accepting of those characteristics? It's more powerful to learn how to remain yourself in the face of adversity than to try and demolish or abolish all instances of difference in the world around you. Sure, we want to get to and create a place where our environment and associates are compatible with us, but where we find this adversity there is the highest potential for self-awareness and learning.

We are to live in accordance with spiritual law. The overall pattern of life will reflect our intentions. The blood purity is most important. The blood cannot be made impure if the mind and body is not drawn into impure action. This is what is playing out. The DNA is like a coherence pattern that develops between mind, spirit and body. When the mind wants what the spirit has then the body is kept in harmony. When the mind wants what spirit doesn’t create, IE: physical pleasures, then the body is drawn out of purity/harmony

If you feel you want to be a part of a safe unveiling of truth, then please share and redistribute the knowledge of this “Apocalyptic” Unveiling of the Hidden knowledge because this is a process that has been ongoing for thousands of years and humanity has finally reached a fulcrum point where there are enough aware people to productively acknowledge reality.

Trauma-Clearing, DNA Unlocking Technology

There is technology that exists that allows the conscious mind to see into the subconscious. This heals trauma and distortion/programming. In the same sense, it momentarily makes all trauma and distortion seemingly tangible to the conscious mind. It is similar to a near death experience where one has to face their fears and let go. This happens in cycles and Earth is currently moving through various states of it. The plan is to allow Humanity to wake up and utilize our concentration and power to maintain a steady flow of awareness.

All is vibration, what we consider matter here and now is simply that which comes from and moves back into a state of vibration. This is experienced directly through the merging of the conscious and the subconscious mind, as if the dream and reality intersect through a coming together of impossibilities and readily accepted occurrences.

When a civilization moves through these stages it is considered a shift of the ages. Everyone’s perspective shifts and as a result of perspective being the source of the details of reality, the entire reality shifts. It can be seen today that what we call reality is literally waves of cosmic energy vibrating on a scale and condensing down into a probability matrix that we call existence. When passing memes control the perspective, because the perspective is focused and concentrating on that, then there is a scattered organizational pattern.

When the perspective is centered within the self, the spirit, the source of all existence, then the outward reality shifts from a scattered organization into a sacred pattern that integrates with the larger picture and allows for seamless integration of the micro-biological perspective and the macro-biological perspective. People are moving out of the scattered signal. Yet those moving into harmony can look and see those around them that are not in harmony and receive scattered information about their inner micro-biological existence as well as their spiritual macro-biological essences and how it all relates to the middle realm of the “Human”.

Help those people by offering harmonizing assistance, rather than increasing the disharmony. As their surroundings become more harmonized, their resistance will increase their own disharmony by virtue of the contrast of their perspective. Offer insight that they can change their perspective and change the boundaries of what they feel is their karma and unlock from themselves a greater view of the whole.

This will be happening through technology, spiritual cosmic awakening, purification and healing of the physical and non-physical bodies and mind, through Earth changes, and through a continued unveiling taking place on many planes and levels of existence until harmony is restored.

Everyone Who Seeks Harmony and Healing Can Be Healed

Everyone can be healed. This is about creating what we want to see, the world we want to live in, not destroying others. Violence begets more violence, fear begets more fear, but calmness and empowerment through unity creates a more stable experience.

The entire issue of what's going on with the secret projects and who's influencing who on the world stage revolves around the possibility that some people may be from an alternate Universe. We are learning how to avoid those mistakes and respect the environment and each other so that we can prosper.

Those who are attempting to destroy everything they don't like or act out in a childish manner in order to gain attention and false-power over another will only find themselves powerless and in company they would rather not be with. That is one way we are being utilized, and it's to create an entire world that we don't want to see but are tricked into manifesting. This is a deception and it's based on the ego and the ability to concentrate and interpret reality around us. The very brainwaves we are feeling now relate to various heights of knowledge or experience. People are entrained to low brainwave states in the same way the TV will cause people to fall half asleep into theta state which is a trance state in this case.

We have to balance the brainwave frequencies and learn to create feedback loops by watching ourselves watch ourselves. This will help us avoid the trauma and mental pollution that is being pumped out, although at this point there are enough people who are naturally inclined to see through the haze that the world is changing regardless of what anyone does now.

The best we can do is prepare for change by becoming aware of the body-mind system and how that which we eat physically, mentally and emotionally equates to our spiritual energetic systems and that we are literally comprised of the energies that we allow inside of us. Keep the body and mind pure as if it were a temple. This way a sacred process of transferring harmonious energy from the far reaches of the MultiVerse within into the externally perceiving senses of the physical body can be initiated which allow one to see through the illusions and operate their body and mind with the ability to acknowledge and decide which frequencies will exist within.

Without this acknowledgement of the center then people are pushed around by the external stimuli. This is the whole point, if a person is truly in control, then they will control the stimuli not the other way around.

Programming

Layers of Programming and Decision Making Roles

It’s like asking a person on the inside who participates in one of the various factions why they continue. One might say that it is a decision which is true, but there is programming at each level that is navigated by each individual and this determines the level of awareness within the role.

If a person knows about this situation, they know that there is advanced technology, breakaway civilizations, secret societies, transtemporal planes, the secret of mind and awareness, all of these topics, then they have a decision as to how they’re going to live their life in relation to the public. The outcome of action depends on the layer of programming which is accessed and surpassed via awareness.

The public is at a specific level of programming that is the weakest but ensnares the most people. One who sees the truths behind these groups and operations, they cannot hide from the fact that everyone is controlled. It is apparent, and if not from the results, reactions, stigma, cues, productions and all other media elements of the system which can be read as a code, but from the awareness level of all the “agents” involved. All involved are exposed to truth and power that is beyond anything that is recognized publicly for hundreds to thousands of years. There is always a presence of higher spirit around the world, yet these topics are often reserved for science fiction when speaking of people who can travel the world in a split second or freeze time for their own personal use.

As a result of everyone’s knowledge, there are various levels of programing designed to limit the mind’s access to specific memories and specific frequencies of consciousness in combination with specific memories. The more one resists, the more intense and articulately abstract the programing is designed.

Programmers

The programs are designed by a select group of individuals that reside far behind the scenes and process information at levels we do not have access to without their technology at least to stabilize brain temperature for the extended periods of focus.

Sub-Level Programming

So people at the level of the public are most weakly programmed with false ideals about security through patriarchal authority which satisfied the desire to have a proper father role in one’s life and acts as a pacifier for those who desire no change. Then there are sub-levels of programming with that for anyone who disagrees and is then swept into the next layer of ideals which are based off of all the possible reactionary mind-types.

For instance, if one [i]does[/i] want change, then they are presented a variety of ‘false-exits’ which tend to one or another layer of psychological or spiritual fulfillment but do so in a way that safely reroutes the intended effect of actually accomplishing change with a secondary and preconceived notion of what the programmers want. The programming objective is not necessarily precise in mass scale, from the layer that people are influenced, yet the ultimately or end-goal of their resulting internal desires are a reality schematic that the programmers specifically designed over many years.

This is playing out through each layer of the programming. At any level, someone could pick the wrong way to look at something and begin to unravel what was previously their very understanding of the priorities of the situation. This deals with high-tension operations, moments, programming sessions, decision making, programming upkeep, trauma, memory blocks, the pressure and responsibility of knowing, and the specific plan of the programmers for an individual that comes into their view. The programming is to ensure safety and continuity, as well to decrease the likelihood that a person will reject or consciously come to awareness of or acknowledge the programming or situation itself.

Dissolution of Programming

With that said, the programming dissolves when an individual asks enough questions to necessitate a learning process that introduces variables and possibilities outside of their operations and programming. This is a problem because with this, trauma comes to the surface and all the actions and operations, programs and experiences must be come to terms with as part of the healing and re-unifying process. There are multiple selves or split personalities which are actually just layers of awareness at various levels attached to various emotional or intellectual patterns. All these selves come to the surface and in order to heal the trauma of the split memory concept the original self must come to the surface and acknowledge them.

This is not a trivial process, and the majority of people who have “basic” programming which is fundamental in society, are in a constant state of altered self and have not witnessed the true divine which is eternally present and contains access to all memories, mind-states and awareness.

Advanced Technology, Deprogramming The Mind, and “Space/Time”

As I have described previously, extremely advanced technology is used and this is through vibrations, generators, and electromagnetic devices that have this effect of propagating waves into space/time which can be heard, felt, seen, and perceived in the mind. These are ripples and vortexes in the “mind-space” of reality itself.

As well, there are brain to machine neural interfaces which operate to activate the psychological perceptual changes and explore the subconscious in an aware state. This is used to test and probe the mind in order to bring forth the trauma, the selves, the memories, the reactions and the individual then must come to terms with what is produced.

This is essentially a “good-use” for the programming and healing technology. It is actually a different process with different technology, and programming is much simpler as far as technology. There are difficulties and challenges in the operative’s role in programming and directly influencing and interacting with the target. As well the standby team for deprogramming and the revelation of trauma and altered, artificial personalities goes through a lot. But the use of the advanced technology in this way is a pattern of experiences and unveiling of the spiritual and mental bodies which serves to reverse programming.

Dreams, Technology and Inner-Space Travels

There are inner space travels which relate to soul journeying or visions.

This is when the mind must cope with what is seen in the imagination. It is as if when we dream we are looking into the world of the mind and from this we experience many situations that are often deviations from experiences we have physically in waking life.

As a result of the dream state there is an expanded quality which equates to various strange situations that we wouldn't normally experience with various relationships between concepts or goals that are often reflective or symbolic of ideals and archetypes we experience in life. This is how the subconscious is is witnessed by the mind. This is how information travels, being connected to many times, experiences, memories, places at once, and relates to the one experience of this life experience now (the physical experiences). Through these symbols information is tied together and collective into metaphors which convey deeper knowledge.

The dreams we have are often only a simulation of waking life, and this is due to the dreamer dreaming in a way that produces only a glimpse of the true reality.

In theory, there are no true shapes or forms in the dream-world because this is contained within the mind or even the spirit as a spiritual experience by definition. So then, what guides the various shapes, forms, realities, memories and experiences into being, when we dream? One might say it is the physical experiences but this is incomplete. It is the level of awareness we presume while in the dream-state, disconnected from the physical anchor, that determines the layer of mind and spirit we will be accessing through those dream-visions.

A person who is completely aware in the dream state will reach the level of self-awareness of the dream itself. We are normally simply aware of the dream, if we continue to increase awareness we become aware of the self, in the moment, and this creates a self-awareness feedback loop. The dream becomes a lucid experience of consciousness.

Awareness Feedback Device

In describing this, I realize how fantastically similar to certain devices and processes, that can be achieved using advanced technology, allow the conscious mind to peer into the subconscious and unconscious. This is described through widely accepted Freudian theory of psychology known as psychoanalysis. There are other schools of knowledge from tribal beliefs involving all pervasive 'living energy', to eternal and temporary minds, to the study of 'chi', the meridians, dream-journeying or chaser. The knowledge of the mind and spirit is found everywhere.

These devices used feedback loops to enhance the acceleration of the same cycles of consciousness that induce self-awareness within a dream. The mind becoming aware of itself. This produces a profound effect to the degree where the highest levels of experience become a sacred eternal pattern in which the mind constructs the physical world by organizing perception.

This is the same general concept as an isolation tank yet the outcome and method is more precise. The technology specifically shuts out all other input and exposes the brain to its own feedback loop much like when a camera records it's own output on a TV monitor and creates a hallway of simulated "reverse" mirrors and screens stretching into infinity.

Oneness, Higher Being, Completion, Eternity

The real experiences are found not from the corruption but from the research and exploration operations that were designed to find the truth of our existence. "God" is not only an energy-force but an incarnate being known as the "Hue-Man" and any being for that matter that can simply come to know itself as the Universe. The Human that contains self-awareness rising to high-awareness comes to know the self, the body, the mind as the Universe. The Universe is both within the body as the body itself, as well as the operating force of the Mind itself. Everything is a balance between masculine and feminine features, light and dark, hard and soft. There is no true complete correlation yet the holographic ontology we persist in can lead to further and further adventures towards completion and representation of the "whole picture".

The Holographic Reality

The leads into the next phase of knowledge. The holographic nature of reality may be only one facet of a much larger continuum of consciousness. The perception of time may be related to phase-locking with each other's consciousness and all perceptions may be a variance in the collective phase-locked field which is really the source of variations in reality.

Veiled Awareness, Hidden Knowledge, Eternal Time, Layers of Programming and Control

The interesting notions here are that the programming of the veils goes beyond what kind of situations are found in the bases. However, they do go right up to it. It would seem Humanity was previously programmed on a civilization-wide level through some kind of great "falls" of consciousness, from a high spiritual awareness to a low spiritual awareness into duality-consciousness regarding existence, life, death, mind and matter.

The Great Catastrophes

From these great 'catastrophes' the civilizations of Earth grew denser and denser in nature and actually became what we have today. The concept is that when these vibrations of the mind and body are increased, we actually see a transcendence of the physical realm and preconceived notions of boundaries. It is as if the quantum doorways to all the atoms open up and a new space is presented, but was always there and in connection as the source of all the influences and effects we perceive ‘out here’ from this slower, denser, more boundary driven perspective.

I was brought into the Time Program

I was brought into the time program. You create any reality you imagine. You are literally a powerful reality generation system and your heart and mind are scalar energetic guides for this process. Without guidance and control there is imbalance and imbalance can destroy. This is the nature of all minds and hearts, this is the challenge for humanity to create something beneficial for themselves and the whole to actually know themselves instead of being a pawn or result of someone else’s selfish intentions.

The Effects of Temporal Manipulation

Temporal Influence

When a civilization reaches the level of technological advancement that allows for remote temporal viewing and transtemporal travel what will happen is that the information of the future begins to integrate into the past.

This is why we have all these ancient mystical schools of knowledge that studied time, space, and consciousness in ways that are only just now being publicly confirmed by scientists.

Simultaneous Interaction and Incorporation

The operations that dealt with moving to the past instantly created a simultaneity of interactions and historic events which correspond with that operation.

Due to the nature of time travel which equates to consciousness travel when a person is consciously sent backwards through history they physically experience life from the viewpoint of their genetic imprint in that point in history.

Retro-Causality

What this equates to is that simultaneously that person’s experiences become the part of history that was already present when they left this time.

Essentially, when a person returns their experiences become a part of history that already existed when they left.

Quantum Holography

These are the parallel and possible realities that exist in a quantum superimposed state and can be accessed through advanced technology or a spiritual adept. Often a spiritual adept is utilized along with the advanced technology to operate in and out of possible realities. The mind must be trained otherwise the result is a fracturing of the personality from the root reality into many sub-sets of identities and data streams.

Love is the quantum link that connects all DNA super-computing consciousness. That is, DNA is influenced by intense emotions and this can be reproduces and is what also generates unpredictability in the surface level.

Time Paradox

Time paradoxes are something that one must learn to organize and solve otherwise there will be no way.

The paradox has been solved, the remaining motion is for everyone to acknowledge this and integrate into the reality behind the most conscious aspects of the brain.

What can travel through 4D and 5D space does not make sense in 3D space. So there is a possibility that some of humanity traveled through higher-dimensional space in order to re-seed a ‘distant’ Earth where no life was yet born.

What if it was then concluded that the distant Earth wasn’t distant in space, but in time and was the same Earth that the human scientists left from long ago.

Every civilization that reached the point of re-seeding would have to go through the loop and watch the whole thing start all over again.

Breakaway Civilization

This is an entire breakaway civilization that uses very advanced technology to dominate the world.

I was part of a genetic engineering program that sought to combine various elements of DNA from various sources in order to create a more easily controlled yet powerful and defined person and personality. This involves chemicals, advanced (relatively) scalar wave technology, ritual trauma and programming/mind-control, and many other aspects of conditioning, training, secret operations, physical enhancements, cognitive enhancements, virtual reality, off-world operations, temporal manipulation, spiritual and etheric training or conditioning or programming.

There is self-destruct programming which all ‘operatives’ are programmed with for self-preservation of the military command. The self-destruct programming is designed to take all the aggression one feels building up inside and willing them to escape, and causes this aggression to be turned against the self. That is self-destruct programming.

Most operatives are still under a hypnosis kind of trance which permits one to walk around in public and daily life as if none of this ever happened. Later that week, that same individual could be called for an operation and they would never remember it unless they were forced to kill someone in public out of self-defense in which case most operatives automatically unlock and become unstable due to the flashbacks.

Underground Cities

The combination of advanced technology with secret societies enabled the development of complete breakaway civilizations. One of these civilizations, possibly composed of many factions working in the same area, exists deep underground.

These separate bases are connected via high speed magnetic drive pods. The inhabitants often utilize energy healing devices, as well as perception enhancements and this often leads to a completely different outlook towards the surface civilization.

Technological Advancement

During WWII developments were made involving electrical generators. This sounds vague and there were many other details regarding the scientific knowledge, yet the idea is that very simply electromagnetic, non-Hertzian “Teslian” energy fields were discovered to have been responsible for propagating energy and consciousness in dimensions that were not previously observed. Everything from DNA, consciousness, hidden or regularly unobservable natural phenomena, time, even mass could be seen to exist as part of this hidden, latent, scalar-wave frequency as a pattern is found everywhere.

If one could determine the particular pattern of frequency and location then this can be replicated and anything can be influenced using these non-Hertzian generators.

Radio-frequency alone was developed intensively during the 60’s and 70’s in America and this was discovered to interact with consciousness to influence emotions, cognitive capacity, memory, cognitive function, biological function, health, pain, comfort, pleasure, and all forms of brain-activity.

After the knowledge of advanced technology and the occult sciences of all parts of the world was combined, the ability to target and interact directly with the human soul was refined. This was a mix of psychological tactics used to manipulate or expose the hidden psyche and technology that could manipulate memory, emotional states, wakefulness, and even dreams.

What resulted was scalar technology that could not only influence a person to take on a particular role or response, but technology that could actually put a person to sleep and locate their consciousness to contain it and transport it to a new location.

World-Wide Control Systems

In order to ensure continuity, maintain control, and shape world events these technologies were employed to manipulate the public, alter the weather and natural phenomena, and develop a means of travel and operation that enable one to move outside of the ‘plane’ of the Earth. This is includes high velocity transportation, as well as spacial-temporal distortions which could effectively “hide” one’s entire craft from visible sight. Space itself was altered to ‘shield’ craft with an a layer of ions to bend the light, or the phase was moved out of sync with the “time” frequency of any observer and this literally renders the craft immaterial.

Solar Warden

Solar Warden is a hyperspace-faring race of human beings who utilize cloning and holographic genetic recording/storage in order to protect their race and protect the Earth in this current situation where there are more than one factions using this technology. They are unable to inform the human public because they are in a similar 'Mexican standoff' situation that we are mostly all in. If they show themselves to the public, the public will be shocked. If they public does not know this situation, that there are multiple factions fighting for control over humanity, then they will not know who to trust. In that situation, the 'other factions' will surely show themselves in an attempt to trick and gather as much human genetic material as possible.

The only way forward to the future is to share the truth about Solar Warden, the space-faring fleet of humans who do no contains the mental and physical illnesses that are currently tearing our planet in two. I believe Solar Warden saved my life as a child during more than one event where I would've been entirely maimed/paralyzed or simply dead on the spot as a result of other faction interference. I was in the bases as a child and was trained as an operative for MILABS a secret division of human space-military. The military has been preparing for the future by testing for various flaws and potential uses of cloning and genetic engineering.

Team Light and Team Dark Non-Interference Agreement

Team light and team dark agreed to a non-interference on ground level rule in order to determine how history would play out. Think about it. You have two groups of people, both with immortality, zero-point tech, virtually infinite energy and power with two opposing directives. How do you get to the future? It was decided by elders of each group that there would be no point in fighting an all-out overt war, this would likely destroy the population as much as it would destroy the Earth. As well, the elders of the elders which no body knows would likely have a problem with these groups going at it in such a way so as to change the flow of the planet for every individual thereafter. So it was decided that there would be a "non-interference" agreement for the surface population. Anywhere other than the surface, one can interact with the current population.

All Contact Must Be Initiated On an Individual Basis

On the surface everything that is generated must be done so through living channels in the population and the civilization must be preserved in its level of technological advancement. No overt motions would be used unless an absolute necessity or the majority of the civilian population had naturally come to the awareness of these possibilities. Doing so before this was met would undoubtedly and permanently change the civilization forever.

So what began as a collective motion towards higher awareness became a series of advancements and set-backs all taking place within a short amount of time. This represents the actions and operations of either side effectively helping humanity along it's path or setting people back with orchestrated events such as the suppression of ancient knowledge, the manipulation of the education or other systems, or the generation of war and prison for profit based on the previous alterations of the education or other systems.

So team dark used its power to help turn society in a warlock ruled prison and indoctrination system of absurdity and scarcity. Team light used their power to help awaken those in need and protect those who's actions are beneficial in the overall plan to save mankind from itself. Team light essentially buffers and shields the population from the operations of team dark.

Team light follows a non-interference agreement while team dark merely pretends to do so as much as is necessary not initiate war.

A Singularity of Change and Awareness

Neither side wants an all out war. This would reduce the dark faction's chances of gaining a majority negative harvest because their motives and presence would be instantly revealed. Team light would also have to show themselves and again step in which is not a pattern of learning that prepares people for the experience of the self and the truth in the Universe.

Both sides know people have to accept the path they can offer them and cannot force anyone into it. They also know that the more this goes on the more likely people are to collectively awaken to the reality of what is happening. It is like an eventual singularity of increased awareness.

The motions we see today are the results of many people coming into awareness at this time on the surface in order to initiate a harmonic alignment as well as to investigate the ongoing crimes against Humanity.

As a result of this, both teams know that the likelihood of collective shift and spiritual awareness is becoming more and more real.

Many events are being allowed because they are representing the eventual shift towards the one end of the polarity from the other which is felt as a shock to society, and this is used to propel society forward through whatever means necessary.

Either people will see what's happening and agree that we've been in a feedback cycle of negativity for the entire age of existence this civilization is aware of, and then we will shift into the real. Or people will claim that this is how things are supposed to work and instead of realizing everything is controlled in a deception based system, they see this and attempt to use the deception based system to gain more collective awareness. If this occurs by the majority then it will be seen as an "acceptance" to initiate a plan that will generate the overt perception of a deception and oppression system. This will surely shock the collective into realizing where they are.

If the majority comes to awareness of this deception based system then the shift moves from the end of the age of deception and external power into the age of internal awareness and harmony.

Advanced Technology and Ancient Knowledge

The plan is to slowly release advanced technology which will undoubtedly connect with and increase the awareness of ancient knowledge. Such technology nullifies the problems that the planet is currently plagued with, although without a developed sense of the spiritual self the technology becomes destructive.

Cloning, Atmospheric Civilizations, Genetic Stability

Here is something that most people might not have contemplated yet is possibly one of the most important factors at play here.

The vehicles that house the breakaway Human society are entirely environmentally controlled. This means that the air and surfaces are clean of all the viruses, bacteria, or pathogens that are numerous on the surface of the Earth.

This creates a specific difficulty in the situation. Any contact between surface Humanity is so heavily and carefully controlled that regulations go beyond any situations we have present on Earth. If the safety and cleanliness of any of the society's vehicles were compromised this would endanger everyone "on board".

So this explains a very precarious and specific situation we currently face and this is not so different than "first-world" nations contacting indigenous tribes and similarly endangering their society. Everything that happens will be according to a carefully controlled plan.

Training, Simulations and Learning

Part of what went on in the underground base experiments were to design "tests" or "labyrinths" with various people, props, traps, and treacherous situations that were to be navigated and escaped from.

Feedback loops with the brain, consciousness, and environment can create a very complex and integrated learning experience that is customized to each individual learner.

What this would do is create an individual that was capable of learning faster and faster and processing information in a more efficient manner.

This is not the same as specifically trauma-generating situations.

Cloning and Cerebrally Enhanced Soldiers – MILABS

This includes experiments, condition, trauma-based mind control, enhancement, memory suppression and operations that cover a span of space and time.

The NAZI faction has used cloning and genetic engineering to create hybrids, soldiers and everything in between. Religious/spiritual icons, to slaves, to leaders and even robotoids of sorts.

Your genetics BELONG TO YOU as the LIVING(ETERNAL) SELF. These issues are going to come more and more apparent.

If they agree to allow cloning, then who says who is allowed to clone whoever they want, Einstein, for example. And then who permits who to get DNA from history or objects in the world. These issues are going to have to be looked at because they have already been used in ways that are far beyond the basics.

Then what about consent of the masses? This is clearly being put in public because consent is required on an overall scale and individual to include people.

So what if the masses consent overall, would the people who choose to remain be at risk? Would Tribal and Native village life and people who respect the Earth be protected or permitted to stay?

Psychological Testing

These scenarios can also be generated to create specific situations that will contain tests or elements of temptation that will probe the person's psyche for weakness.

Hallucinatory Holographic Mind Probing Technology

Through the neural-interface technology the minds of many individuals can sync together to experience the same virtual simulation. Thus a single individual's mind can be selected for input and the environment will be a construct of their mental activity. The others present will be able to interact with and explore that environment like the person's mind.

Various training scenarios were utilized to enhance the resistance (or lack which) of the mind to intrusion and subversion. These were harsh but allowed the individual to completely control the energy flow in the conscious mind while remaining in harmony with the subconscious.

In a strange yet perfectly reasonable way, the conscious mind acts as the central "self" in that environment, while the subconscious mind becomes the "background" elements of objects and settings, actors and intentions within the flow of the "scene" itself. So a kind of movie begins with drawn out characteristics and perceptions that match that of a highly cinematic movie or intense dream.

The brain is naturally naive to stimulation and if allowed will liken the event to a dream in which the suspense of attention or lack of lucid self-awareness is "acceptable".

The truth may be worse than what you're thinking as far as training simulators and mind hacking devices. The truth may be that the human mind is inherently lucid and self-aware and that it is external influences that cloud the mind into what we consider 'regular' dreaming and 'regular' waking life. It may be that the conscious mind has a much greater access to information through the subconscious mind which is largely covered up and ignored in a rigidly physical and ego-focused society. As well, it may be that dreams are meant to become lucid and that we are often the subject of mystery in other realms that often refer to humans as ‘sleepwalkers’ that don’t know their dreams are real.

Virtual Simulator Training Scenarios

These scenarios can be entirely holographically generated to produce a visceral 3.5 dimensional experience which can be used for training or psychological testing.

Combat Simulators

The name is self-explanatory enough. Situations can be generated which are indistinguishable from the real events. The entire neuronal map of the brain (connectome) can be created and this is by machines to mimic the brain. After calibration through sensory feedback experiences, the sensation of the simulated realm is 100% the same as the neuronal impulses that are detected by the physical brain. An adept souled being can always tell there is a slight difference, but once the machines are calibrated the physical sensation is generally the same while in the machine.

These environments and scenarios can mimic any combat situation to determine accuracy, integrity, endurance, skill, and all areas of aptitude of an operative without the risk of injury or the lack of real-time effect.

Conditioning

As was explained in other sections, these same systems can be used to remove fear of certain scenarios. They can be used for programming the mind. They can be used to train one to complete a task over and over until they can recreate that action extremely quickly in real-time. These devices can be accelerated to operate within a fraction of the time yet to cover training and conditioning that would normally take place over hours, days, weeks, to months.

Programming and Resistance to Psychological Trauma

Through the development of specifically customized scenarios and neural feedback systems trauma can be brought to the surface or specific fears and traumas can be neutralized. This is essentially “facing fear” but in a completely generated environment that will be just as real as the real thing. One can literally face any fear, any trauma, any kind of difficulty and these systems can be utilized to accelerate a process that would normally take weeks to months or years and individuals can be strengthened to their maximal capabilities.

Accelerated Mental Functioning

Some of the technological systems were designed to test the enhanced functioning of an accelerated brain.

The mind can be seen as the energetic component of the physical brain. The physical brain projects consciousness via the electro-chemical processes.

Technology, Accelerated Learning and Repair

Technology has been developed for assisted learning and accelerated repair of tissues.

On Orgone Energy

Orgone devices produce a scattered signal which disrupts the ‘negative’ energy fields. At least this is what I’ve been told.

Apparently if built correctly it will either disrupt the negative emf/harmonics or it will produce positive harmonics, IE: Orgone energy. Orgone energy is the raw energy of life, while dead orgone, DOR, is the lack of orgone energy and can be “found” (the lack of orgone) in thunder storms and illness or depression.

This is exactly what I was told. We are generators for the same energy, and with the correct principles of energy and internal knowledge combined we can produce those very effects at will.

The energy of the mind creates a central vortex within the body. When the frequency pattern emission is consistent with fractal congruence, then that holographic printing of energy can be expanded or shrunk infinitesimally and it will retain the original fractal configuration.

This is what enables transference of awareness between fundamental harmonic layers of the Universe. When one’s consciousness energy pattern is balanced and tuned to itself so that all proportions can be shrunk through a vortex and expanded on the other side, then the personality or self-awareness remains true. This is as if the space/time must turn inside out and only when there is fractal relative polarity in time does the ‘shape’ of the consciousness within the DNA, work both ways folding and unfolding to represent the same mirror image. This is also symmetry in time as far as an ability to reverse the perception and still retain some kind of connection forward and backwards.

Energy Generators and Serums

Generators which produce vibrations that influence the cells to heal at an accelerated rate, combined with serums that enable the fuel for chemical reactions can instantaneously repair tissues.

I couldn’t believe what I was seeing at first, thinking it was a trick. There are serums and energy devices that can influence the healing and general health of the body and even mind.

These devices use energies which extend beyond the physical and are often compared to electromagnetic waves combined with ultrasonic or low to mid frequency tones. These produce harmony on a level that directly interacts with the biological functioning of the body’s cells.

I feel that this technology utilizes the same energy that is emitted from the mind and it is as if the device is “speaking” to the cells in the same kind of universal language of electromagnetic pulses, tones and vibrations.

There is information on energy devices and the use of light therapy in physical therapy today. More is being learned about the power of light and vibrations to influence the mind and body.

Genetic and Cybernetic Enhancements

Certain individuals are also genetically and technologically modified to produce a kind of hybrid human that can heal at a spontaneous rate as well as operate at a much higher rate of speed and efficiency than a regular person. Serums to induce muscle growth and mental acuity are applied, but the main condition is the enhancement of certain features through the alteration of the RNA aspect of the genes.

These features are combined with tech-assisted metabolism and certain biological functions can be enhanced or controlled via a supercomputer. These are the ‘brain-chip’ interfaces which operate to function as a governor for the mind. When the mind becomes too amped up to focus on the details the faster than thought computer system organizes the input into more usable information streams. When the mind is too slow to see the necessary patterns and variables to accomplish the goal, the brainchip operates in a self-preservation fashion and links together information streams to highlight the most viable possibilities.

Such technology is not without warning, much of this is being used heavily in the military field and this has given rise to entire factions of enhanced cybernetic soldiers. Although this is something to know, there are also many other uses for this technology and it is used by people who want to explore the technologically assisted path.

There are many other aspects of this to discuss, or at least a few important ones that will have to be understood to know the greater truth about what is happening on the planet.

Developing The Self

The issues we face are to help us learn about our selves and better mediate the known and unknown aspects of personality. The more we focus on realizing the unknown aspects of personality the less we are seemingly controlled and forced to experience the undesirable aspects of existence. The more we cling to just what we know and what we feel comfortable with the more shocked we are when we inevitably discover the truth of the whole of our existence which is *always* at least one half more than just what we know and are comfortable with. That is the nature of the universe, we can only achieve one half of the equation at any given moment. Once we complete that seemingly whole view of the universe, the information we gain access to instantly doubles and we realize that we were only looking at one portion at a time, just one half of the equation. Then the rest of the equation comes into the picture and we must expand our view once again. This is because we are viewing ourselves as we discover the universe because we are creating what there is to be discovered through the act of exploration in itself. We are the explorer, the explored, the creation and the discovery simultaneously.

Multiple Factions of Space-Fleets, Multiple Goals

Now there are what are called ‘factions’ of groups that have different agendas and interact with the public or surface level population differently. That is one of the main issues. These departments and factions are no longer limited to the surface level of existence and have access advanced technological capacity and knowledge of a larger view of time and a more complete understanding of consciousness and the human existence.

Large-Scale Catastrophes or Series of Meltdowns

I was informed that while they work continuously to prevent and reposition this civilization continually outside of the reach of any large scale meltdowns that the technology is commonly accepted to only be a temporary barrier between the inevitable.

Unsustainable

This society is unsustainable and must either change or will experience the massive catastrophic evens that are foreseen.

There are multiple perspectives that are maintained. They vary by the way of the methodology of solving the problem. All agree that change must occur and is impossible

to postpone forever. The future society is the result of our advancements and achievements. Regardless of how we get there, we must get their somehow.

Earth Changes

I was informed that these changes had been held off and would ultimately be allowed to take place to resolve the current situation. The changes can only be held off for so long, either that or the people performing this task would not be reasonably inclined to postpone these changes indefinitely.

The idea is to locate and secure the members of society who are fearless and capable of assisting in this transition and who hold key genetic elements as these individuals and groups are capable of ‘holding-space’ for the rest of civilization by literally containing the genetic precursors for all those other individuals.

I was informed that Earth changes would be the last event to take place, pretty much like icing on the cake.

This means there will be a lead up to these events.

Mass Consciousness

I was informed that the consciousness of humanity and every species is linked to the bio-etheric field of Earth. When the consciousness becomes too unstable to maintain a controlled society that this will initiate massive Earth changes. This could simply be instigated through advanced technology but I was informed that the advanced technology is actually postponing this so as to allow people to prepare, to get the word out and avoid a complete loss of the human race.

Wars and Manipulation

Wars were described as a way the power groups had controlled humanity over long periods of time. These wars were used to distract and weaken the population and the power groups believed they were holding off a full scale invasion by suppressing the rising population.

Now the population is high enough that another war would be a final war and this is extremely dangerous for the entire civilization. There would be no need for another war to repeat this cycle because the technological achievements have been met and so essentially the ‘reseting’ of the civilization would not be required or possibly acceptable.

This is why war is to be avoided however people are very easy to manipulate through lower emotional and mental programming from media and social cues. People follow the herd and the herd is essentially still operating on animalistic parameters. There is a have coating of peace and humility but this is just a wax coating that quickly evaporates as soon as the pressure of mass unrest or scarcity is realized. Many simulators have explored these possibilities.

Financial Unrest (or reset)

The idea is that financial oppression is used before wars to distract and oppress the population. When people are busy looking for resources to survive, they are not focused on the larger picture. This is part of the ‘fight or flight’ response mechanism tied to lower brain complexes which are primarily activated to keep the masses in both a highly sensitive and unstable state of being while also keeping them easy to control.

If the civil unrest reaches a level where awareness continues to rise even though financial hardship is being implemented then the latter two options will not work. At that point artificial disasters can be produced as well as natural Earth changes which initiate a shift of the ages and of consciousness.

The paradox or maybe irony here is that the oppressive actions are used to both suppress human awareness up to a certain point and simultaneously to stimulate and arouse the human spirit into wakefulness and self-empowerment. This is part of the reasoning regarding why these control systems are allowed.

Some say the elite don't want an economic collapse. That it would ruin everything. Others say it would be icing on the cake. Some say that it is a zero-point world that will save us. That the Fukushima is a big problem. Zero-point field tech, that is. Generators and energy sources, healing and terraforming. I know we project right? We see it in our own lives. So we can project with each other and grow that as a group and that part of the world will exist in that light. Which sounds Biblical, which may be because they're following prophecy. But it may also mean that we have a light, a chance, a real salvation, but that we have to grow that here and help develop it. To pass it on like love and such.

That's apparently what it comes down to. That these cycles happen every so often and we're in the midst of one right now.

Let's make it count. Every day, thought moment. With mindfulness, intention and awareness in action. Life is literally higher dimensional, so we have to become multidimensional to really face it. Indeed, that is a good perspective on action and inner being. It's like a fractal, whatever we project inside that chest center, will project into our mind and the outside. But we have to be stabilized and grounded through the root first and then to get to the heart energy must pass through the sacral/sexual centers and continue upward.

Self-Responsibility

The main reasoning is self-responsibility. We are responsible for ourselves. The people are and so if they allow themselves to be tricked and prodded like animals while behaving for the part, then that is how their experience will play out.

Clones and Synthetics

One seemingly unfair aspect of this is that there is a larger portion of non-humans than there are humans and so these beings without the human spirit will be easily programmed to support and propagate the mind-control system and the oppression. Thus, if people follow the herd mentality then they will walk directly into the end-time scenarios yet it is the humans who will actually stand to lose. People must wake up to these tricks and this technology so that they can orient themselves with the reality of this situation while learning to strengthen themselves and adapt to the manipulative aspects of this control system.

Spiritual Power and Divine Truth

This is essentially all about spiritual power and humanity stepping into the power of the truth behind their existence.

Some say that humans are designed as slaves, but if this is so, then it is the power that humans contain within that is the ‘work’ that they are sought after for. This indicates that humans have great powers that others cannot achieve on their own, even if they are creator lords.

Thus, whether humans are modified or not, they have great power. If they are not modified to perform this role of an energy generator, then this is just a confusing mind-control game that has been pulled over the eyes of humanity and humans have always contained true internal spiritual power that allows them to exist without having to parasite and steal the energy and knowledge of others. Humans are reality generators , we can choose how we want to experience the universe and what kind of collective world we want to enjoy together. Once we step into this power, there is on greater force. No amount of non-human entities can change that as we are somehow intrinsically linked to the ether-space of this realm through the spiritual and physical aspects of our being. That is what is sought after and so those who do not have this manifestation power attempt to utilize humans to create their own reality through this Earth situation. That is the basics of the ‘alien overlord’ system, whether humans in disguise, non-humans operating with alien intelligences, supercomputer systems, or spiritually degraded beings.

Most people are not ready to see themselves or the true nature of humanity. Reality is largely an illusion of the conscious mind. The hidden aspects are the other half of the equation and this is perceived by the shadow aspect of the personality and mind. We are only marginally aware unless that aspect has been unveiled and through this process the reality of the human soul can be actualized and realized.

I was informed there would be a period of two suns and multiple mirages around the world. People would believe they were peering into another dimension, floating cities, or some kind of city-ship craft.

Synthetic DNA, Synthetic Genetics

Through the previous described replicator devices synthetic DNA can be engineered. Thus, entire genetically engineered codons can be produced and form this created beings can arise. These are the hybrid beings that are performing many tasks underground and even on the surface. On the surface there are synthetic genetic create beings or genetically engineered beings that appear to be human but do not contain a human genome. There are also clones and replicants of humans. There are a variety of non-humanoid forms but these are largely kept underground. These are beings we often see in science fiction movies or shows.

Time as Mathematics, Series of Equations, a Programmable Storyline

When using the technology, time can be understood to be a series of programmable variables where one variable is centered within a series of parameters and from these possible variables within a parameter set a series of possible functions or outcomes can be process. Thus, these people can literally plug into a supercomputer system which operates based on this form of processing reality and through this they can observe all the possible realities that may connect to a present moment. This allows one to travel but this is not part of this section of the explanation.

What this allows one to do is know exactly what is going to happen and when. If a person is being interrogated, all that has to be done is the operative using the technology simply thinks of what they will say and intends to do it within 6 seconds. The machine will play out a future timescale regarding the input of that question or interrogative speech and then produce the possible results. If this doesn’t give the desired readout in the mind of the operative, they will refuse to go that route and think of another pathway. This repeats for a few seconds to a few minutes. Then the computer has found the pathway that is most like to yield a result. The operative simply steps forward, says ten words, the subject breaks immediately and tells them everything they want to know. This is obviously takes longer for most trained individuals, however that is how easily it would be to use this technology to produce an effect on a regular surface level human. They would have no defense, no way of knowing how that individual could know so much or be so direct. This is how operatives dealing with certain power groups will operate in the surface level population and individuals always feel that these people know way more than they are leading on. That is because they are, they know more about you and your life experience than you do.

Chapter 2.3 Previewing Solutions

Neutralizing Polarization

The mental-emotional tendency to pick sides, to fight, or to basically join in the fray that is currently taking place is a kind of deception of polarization which enables the leeching of emotional and mental energy from the human. This is more of a psychological unveiling where the true nature of the self, the duality game and the transcendent source aspects can be made visible to the individual. The they will see how every polarization tactic is actually taking them out of their true source of power by pushing them off center emotionally and mentally. There are teams that are capable of enabling this process and assisting as humanity realizes that half of the problem is that the unchecked mind is one’s own worst enemy.

Removing Etheric Implants

Technology can identify and isolate these foreign body implants which are etheric technological devices. These are literally frequency or phase-shifted devices which are invisible to this layer of physicality but are capable of being organized and attached to the etheric field of the human and will operate to siphon or limit energy on the etheric level. Of course, this is all technology of the vampiric created beings who parasite on the original Earth inhabitants to sustain their existence.

Clearing Astral Body

By moving through the temporal body etheric implants can be deactivated and removed and then the astral or emotional body will be able to function closer to the original levels of presence and sensitivity. This body is manipulated when one is emotionally shut down from trauma. Thus this body can be scarred and will function at a much lesser level of activity or presence until the emotional damage is healed. This contains the memories of all the emotional events of a person’s life.

There is technology that enables one to access these levels of frequency and clarify or untangle emotional chords or attachments which are used to siphon energy from the human being to the parasitic created beings. These chords are energetic in nature but also have technological components.

Knowing the True Self; Soul Actualization, Unveiling

This is a sacred, ancient process that has been outlined and discussed since the beginning of time. Humanity is here to learn, grow and express the self. In doing so, like a child moving through layers of infancy and into full-wakefulness of their body and the physical world, humanity can come into awareness of their true spiritual nature and the true spiritual nature of the universe. This is the process of soul or self-actualization that is described in some psychological methods such as Maslow's hierarchy of needs and the self-actualization that results in a complete fulfillment of all psychological needs. In this modern psychological view, self-actualization is described as the desire and process of reaching fulfillment of one's own natures and then living their life sharing that self-empowerment with other beings because after gaining your self that is the highest expression of such knowledge, sharing that power with others.

The true self is not limited to the physical body, the true self is a reflection of the entire universe and thus the primary original creative force. This is akin to the whole equation rather than just half of the equation. The visible aspect of the self, the conscious mind, is only one half of the entire equation. The whole equation consists of what we see visibly and consciously as well as what we are limited from seeing within the unconscious and subconscious mind. These are really two layers of the same ultimate source. Like an iceberg which is only partially above water, the subconscious mind is the proprietary wholeness of experience. The conscious mind is only a surface layer of reality while the unconscious mind carries all the notions, the multidimensional processes, the latent aspects of reality that we are not capable of seeing without unveiling these aspects through inner exploration.

Taking Responsibility

This entire show is about taking self-responsibility. Unless people begin to do this, there will be no hope. Right now people are convinced that their responsibilities belong to someone else even though this always results in pain and discomfort. To take responsibility is painful at first because we have a lot of work to do, but even then, this is beneficial because there is no way to remedy the pain and discomfort of being lied to and disrespected if people continue to hand their free-will over to others through blame (scapegoating), lower emotional bio-emissions (hatred, anger, etc), distractions and addictions (technological, sensual, and chemical or food based), or generally complaining without taking the first step towards providing something beneficial. This is the victim role, if humanity is to take control then they to help each other and themselves by not relying on false promises and false authorities to determine how they feel, think and live.

Returning To Center Emotionally and Mentally (spiritually)

Balanced food intake, rest, exercise, activity and learning are all continual requirements for a healthy existence. Emotional and creative expression nurtures the soul. Without these ways of being and learning the human qualities of experience diminish and creativity dwindles. There is most likely going to be a split between those who can sustain their own humanity and those who have to cause others to suffer to do so.

Vampires in the “Awake” Community

Contrary to popular belief, the majority of the non-humans are not necessarily vampiric and in power. The vampires are the people online who try to get a rise out of others for sharing their own experiences. They are the nosy neighbor that feels that the person who looks or acts differently must be considered and enemy and armed against because they feel themselves to be so much better. The majority of people are vampires, they believe that they must work for someone who will trade them money and that the amount of money they have based upon the amount of indirectly beneficial work they do puts them above others.

This is not different in the ‘awake’ community and in fact I feel that there are higher concentrations of vampires in this community, the alternate media because everything is largely a joke and a trick. People are operating on deceptive agenda schemes to divide the people and this is most intensely noticed when there is a realization of the truth behind what is happening and people begin to point fingers. The only people who are cleared to point fingers (and who also will NEVER point any fingers) are the operatives from the secret projects that have the capacity to do so because they have reached a specific level of awakening and this is related to the droning, cloning, and hybridization process. Only a specific few know who has been transformed into an impostor being and who is remaining strong and this is due being briefed on the surface level situation. Anyone can say what they want, they are simply playing into the tactics of mind-controllers and parasites and the more they play in the deeper the infection will grow. All the people pointing fingers now are actually implanted and vampiric and are attempting to drive a divide through humanity and they are the first who will turn on their friends and family when the going gets tough because they are more interested in others and starting drama than healing themselves. Ironically, this is all done under the guise of “healing” and “getting rid of the drama” AKA people they don’t trust. They pick and choose based on their own personal bias and this is a sure indication that they are emotionally and mentally out of alignment and are feeding deep seated personal fears from their own traumatic experiences.

These traumas must be healed and the harmonious center must be reattained where a person can focus on their own existence and waking up instead of acting as false-light, “king of tyranny” gatekeepers for others. Those individuals are going to be left behind if they cannot kick the habit because if they are allowed into any kind of healed society they will quickly degrade into animals and attempt to restart a devolution process in order to gain the power they so desire but cannot produce on their own. ANYONE playing ego games of trying to hurt or put blame on ANYONE else contains this viral implant and has not kicked the virus.

Heart Center and Psycho-Spiritual Fulfillment (self-actualization)

Emotional connectivity is a form of telepathy, when loved ones can tell the state of another across space and time. Whether it is family or friend. In the 60's and a little before that the US found a specific frequency akin to a radio frequency that operates on a trinary spectrum. They called this the "telepathic wave". In time, they could completely read and send thought waves without frying the brain or body through radiation. Which was what the early machines did. They actually took satellite dishes and aimed them at people. Then realized they could turn it around and do it backwards with a better effect and minimal radiation. Either way they started beaming the base and then towns around them with waves between 400 and 432 MHz and caused waves of crime sprees or peace sprees throughout the 70's and 80's. So they can digitize the telepathic wave to induce artificial telepathy. But this is organically done through an entanglement through what we call love. Literally. The Germans found it before that. And the secret orders before them. And ancient Egypt before them. As humans we remain basically through the root, heart mind connection. When the heart overpowers the mind connection of spirit, then the body is a trinary whole. When the mind guides it is binary and cannot by guided by the spirit. The spirit has direct access to the heart. The mind is a computer of sorts. The mind doesn't see past duality. The heart is the true center/eye of the spirit. The reptilian brain is the dragon and when you tame the amygdala then fear is ‘behind’ you. When the mind is "untamed" the reptilian brain takes over. This is the lower brain and amygdala fight or flight response.

This happens when the energy of the brain, body and spirit increases due to whatever various reason of spontaneous or controlled increase. Because of the "temptations" which are the easiest flow of emotional energy into a "discharge" state, is that the sacral/sexual center ties to the heart and reroutes to replace the heart and control the mind. Mental and emotional contemplation of one's true purpose and nature is the path of the divine. We find ourselves simply by looking in the right place, not in the world, but in our self energetically. We are more powerful than the programming, and we can reverse this is we're careful and devoted enough.

We have to be faster than thought, at the speed of spirit, beyond matter. When we start thinking on this level then our influences spread to others throughout the sphere. It is our destiny to know these things but they have been kept from us. The ancient orders called this the art of "implosion". Traveling realities through intense contemplation and a usage of sacred geometrical forms.

They developed and combined orders around WW2 and developed zero-point energy tech that can alter space/time fields. This is timeline manipulation. But as a rule, we go where our heart leads us. So that's how this is solved.

Go Within to Observer the Self

The answer is literally by looking within to find the great truth rather than looking around us. Everything we need to know is within us, energetically and spiritually. Our emotional, mental and physical state of balance is an indication of where and how we exist in relation to the larger overall truth. Everything that happens around us to distract us is designed to stop us from looking within. Nothing can truly stop us, nothing can actually input the holographic horrors in our spiritual center but simply perturb the brain and body. The mind is a non-holographic spiritual environment that cannot be excessed and the invader's goal is to convince people to use their imagination or their creative powers to produce these horrors on their own. This is through the programming. We must clear out the programming, the trauma, the lower-dimensional constructs and fill ourselves with the truth and spiritual presence that enables us to be self-aware in the first place. If we aren't doing this, then we are either in pain, or there is no spiritual presence to be in pain. If we are doing this, we are either clearing out the constructs by replacing them with true spiritual presence and self-awareness or we are becoming aware of them.

Organic Feedback Loop

The great feat is mentally bridging the great rift between us by looking within. Which is paradoxical.

We can never see what another is perceiving exactly, well for the most part, so we all have to kind of accept that we are communicating effectively and truly get to this level of comfort by simply being secure in knowing who we are our selves.

Technological Feedback Loop

So then knowledge, in a way, when shared, is a form of telepathy. Group telepathy.

Group think, when ideas or memes spread, is a form of group telepathy.

The Internet is like artificial spreading of telepathy. This means what is normally mentally bound signals can be initiated and spread through a digital medium.

Breaking Soul-Contracts

Ultimately everyone is under a series of "veils" which are energetic time-constraints literally like 9 dimensional layers of temporal restriction which are each bounded and ruled over by invader beings. These range from physical authorities, to astral entities, to programmable matter, to supercomputer intelligences, to spiritual false-light beings and our own shadow mind. Each layer of reality is successively guarded by these beings who are here acting as gatekeepers designed to push people back when they begin to wake up. These beings all work through fear and deception and the last layer is based on mirroring our own energies back to us like an impostor. All the levels tend to imitate, however the last layer is a near perfect imitation of our own energy to the point where people often switch places with this being before getting out of the matrix. There will be more discussed on this in a later section.

The “9 veils” were designed by impostor beings as a last effort to stop humans from escaping the false-light matrix which is essentially the holographic universe. This is literally like a maze which reproduces a new pathway and a new doorway each time a person nears the last hallway or doorway that is next to the exit. So one eventually has to project themselves into the exit using their whole being and moving past and through the tricks by not falling for them before they are placed. This is highly confusing but this is essentially a kind of labyrinth that has been created to stop people from leaving so that their soul energy can be destroyed and used as a generator for the false-light system which is an AI over mind that is programmed to take the raw soul energy and destroy the identity by inverting the electromagnetic spectrum and incorporating that inverted energy into its own internal matrix body which is literally the false-light universe.

Help Everyone

Now they may have been setting us all up for the apocalypse. They said they want us to tell on them and get the whole world aware of what they've been doing So that we can stop them and pull them from the dark which is ACTUALLY taking their souls. There are also breakaway civilizations involved in this entirely zero-point/free-energy.

Chapter 2.4 How Timelines Collapse

Repeating Cycles of Time

Because of the way time works and the ability of those using these temporal devices they have set up an elaborate labyrinth like a spider web of false realities, timelines, parallels, sub-dimensional layers, astral cages, temporal dead-ends, and time loops which act to stop people from reaching the true reality. As a result, many portions of this timeline are literally in a free-will ‘time-loop’ and these loops must be canceled out. As well, because of the nature of these loops on the overall stream, our entire civilization has been looping for the past 6 time-cycles meaning we have reached this time in history 6 times before and this is apparently the last time because we have never gotten this far before and the events of the future are having to be programmed in manually at this point.

Beings From the Future

As are result and damaging time, what happens is that the beings from the possible future that is being damaged appear to try to heal the break as well as those who are from the new possible future that is generated, this results in a time war between both possible realities for the present realm. This is what has been happening for thousands of years. All the ancient spiritual texts are describing this system. One group is a from an original possible future where everything is based on the presence of spirit and souled humans, another is from the new possible future where an AI overlord took over and removed the souls and spiritual energy from all entities on Earth rendering this a holographic prison without any other use. That reality dissolves into itself and so the time beings are constantly moving backwards and jumping timelines to try and obtain more energy.

Beings From Collapsed Timelines

The beings from the collapsed timelines are those who travel through hyperspace to get here. They then recreate themselves using cloning technology through hyperspacial means, IE: creating an alternate temporal dimension so that these devices can sustain them and literally “hold-space” for them. Then if their timeline is destroyed, they can still exist as a ‘loose-end’ in space-time. This is extremely dangerous and I will say we have some Human operatives jumping timelines searching for Earth and Humanity so that they can find a reprieve from this time war.

That is the most trying aspect of this entire ordeal for me. There are some human operatives literally in hyperspace without an opening into the temporal stream for them to reintegrate into this timeline. We must assist them, we must open the timeline by taking control of Earth from the invader races and those who wish to depress and isolate the spirited and souled human beings. We must re-organized and neutralize the invasion forces of deception and assist our humanity and maintaining their free-will, their self-awareness, their self-empowerment and their emotional, mental, and physical harmony.

Beings Who Would Never Have Existed in the First Place

As a result of this entire process, there are essentially beings who have been erased from time who degrade in this dimension and can only interact periodically. These are the created beings who operate through supercomputer forms that are placed within cloned bodies and powered by the transdimensional quantum supercomputers. These are the ‘parallel’ reality beings and this relates to the hive mind invasion. When all the original aspects of time are restored, these beings would not have ever existed in the first place as there is no quantum link for their probabilities to remain neutral and present in the original run of things.

Again, all beings who operate under the will of the universe, even when they are attempting to trick hyperspace and time, when these beings assist and help in the unification of sentient, spirited, souled beings they are capable of being healed. This is all because there is an original creative force which can produce souls that are eternal and everlasting, meaning they are not created but simply always ever were. Yet, there is a “miracle” effect where beings can be given eternal life. This is only by reflecting the will of the whole of existence, rather than a portion or a part because this is only temporary. Only the whole is forever and this takes true harmony, true balance and ultimately *self-lessness*. To give to others because you see them as worthy because they can be aware and enjoy the gift is to be like the creative force that gives life to all and enables all experience in the first place. By acting in correspondence with this we reflect the will of the whole and that power and action transfers into protection and self-empowerment for ourselves. This cannot be cheated, tricked, or rerouted as that is the karma game that has been played by delaying time and rerouting energy through getting others to take on the responsibility for others bad actions. In the end, all karma has only been delayed because when the stops are pulled out, *everyone* ends up getting their karmic returns regardless. They simply intended to build up so much karma that by the time the ties are cut, the karma overload will simply kill them instantly. This never happens and they end up in the abyss, this is the final time so since there is no recycling of the universe from the oroboros system then whatever happens is what happens forever. If everyone wants to move into harmony, to receive healing, to receive forgiveness then they must assist in the creative force in healing, forgiving, and turning away from the deception, the cruelty, the vampiric actions.

Many vampires believe they will die if they stop. The only hope is to admit their ways to the others and ensure that there are enough beings around them that will protect them so that this situation can be handled properly, with due care, compassion, and the necessary healing to ensure the safety and protection of all who seek harmony.

Cosmic Healing I

Cosmic Chi Kung

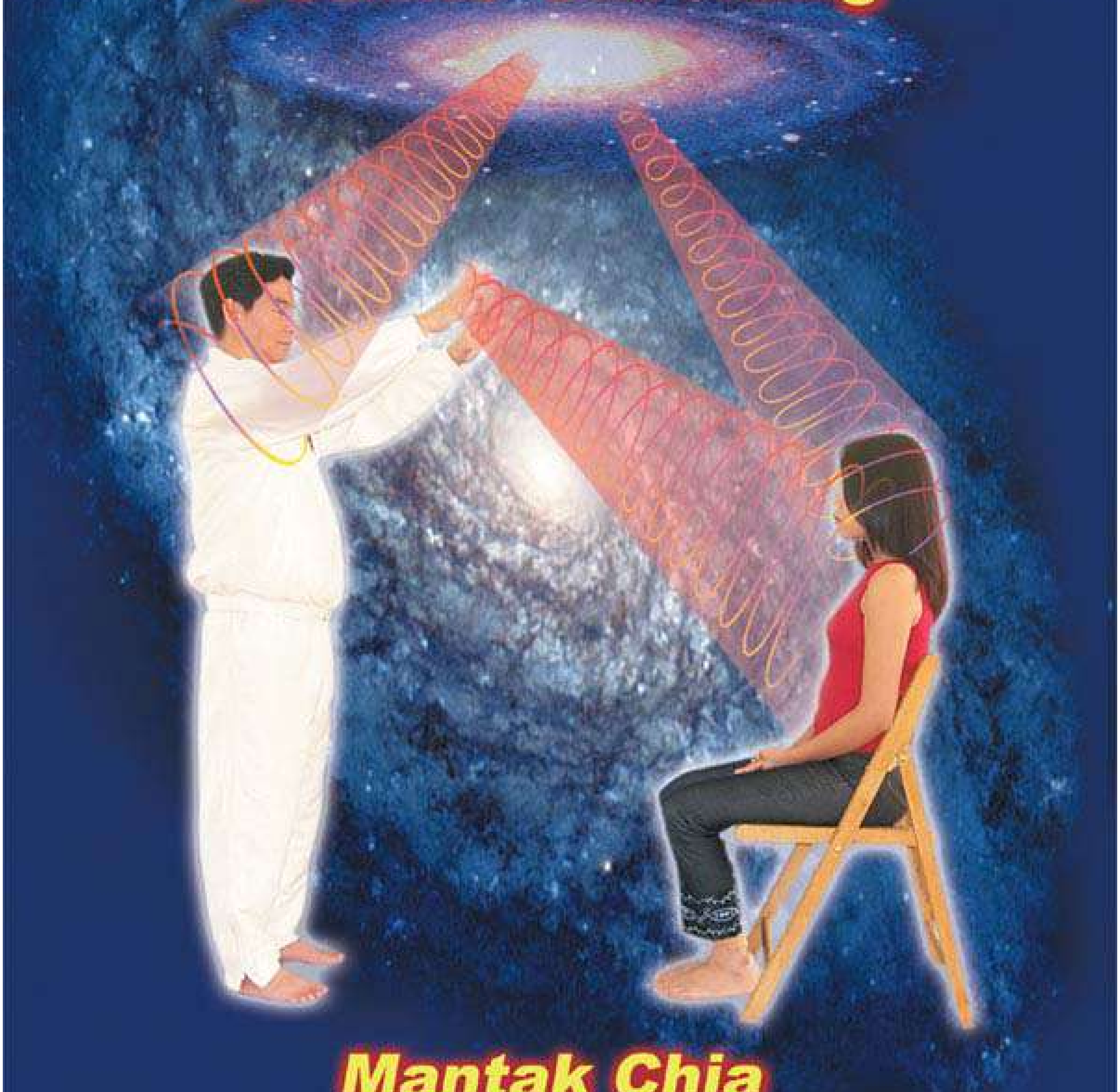

Mantak Chia

Cosmic Healing I

Cosmic Chi Kung

Mantak Chia

Edited by
Colin Campbell

Copy Editor: Colin Campbell

Assistant Editor: Matt Gluck

Contributing Editor: Kendra Moshe

Computer Graphics: Raruen Keawpadung

Photographer: Saysunee Yongyod

Design and Production: Saniem Chaisarn

Illustrations: Udon Jandee

Final Editing: Jean Chilton

Project Manager: W.U. Wei

© *North Star Trust*

First published in 2001 by:

Universal Tao Publications
274/1 Moo 7, Luang Nua,
Doi Saket, Chiang Mai, 50220 Thailand
Tel (66) (53) 865-034 & 495-596 Fax 495-852
Email: universaltao@universal-tao.com

ISBN: 974-87672-5-6
Library of Congress Catalog Number:

Manufactured in Thailand
Second Printing, 2001

All rights reserved. No part of this book may be used or reproduced in any manner whatsoever without the express written permission from the author except for brief quotations embodied in critical articles and reviews.

Contents

About the Author 9
Acknowledgements 12
Words of Caution 13

Chapter I
Cosmic Healing

Cosmic Healing Chi Kung in the Universal Tao System..................................... 14
Foundation of the Universal Tao System 14
Cosmic Inner Smile and the Six Healing Sounds 15
Iron Shirt Chi Kung and Bone Breathing 15
Stages of Mastering Chi.. 16
Four Sections of Cosmic Healing... 17
Connecting to Heaven and Earth 17
Opening the Bridge and Regulator Channels & Extending your Chi Outwards 18
Opening and Energizing the Governor and Functional Channels 18
Activating One Finger Art and the Chi Belt 18

Chapter II
Preparing for Cosmic Chi Kung

Preparatory Practices... 19
Warm up the Stove 19
Transform Sexual Energy 20
Filling the Joints with Chi 21
Bone Breathing and Marrow Washing 21
Bone Breathing.. 21
Marrow Washing 23
Heavenly Marrow Washing with Violet Light 23
Microcosmic Orbit.. 25

Chapter III
Cosmic Chi Kung

Opening the Three Tan Tiens to the Six Directions 27
Introduction 27
Relaxation, Letting Go, Surrender and Ego 28
Mind, Organs and Sexual Organs 28
Opening the Three Tan Tiens 30
Combined Practice 32
Warm Up Exercises 32

True Breath - Skin Breathing... 33
Daily Practice: Opening, Connecting & Activating Chi Within & Cosmic Charging 34
  Relax and Let go - Smile to Connect with Universe Within............................ 41
  Tan Tien Consciousness: The Second Brain 37
  Three Minds into One: 'Yi' Power 44
Activating the Six Directions and the Three Fires 45
  Direction Below 45
  Front Direction and Tan Tien Fire 46
  Back Direction and Kidney Fire 47
  Heart Fire 48
  Sacred Fire (Chi Fire)... 49
Open the Third Eye 49
  Front Direction: Push/Pull Master Practice 51
  Left and Right Directions 52
  Direction Above 53
  Open the Spine... 54
  Open the Middle Channel and Perineum.. 56
  Tan Tien and the Universe 57
Open the Three Tan Tiens 59
  Upper Tan Tien - Mideyebrow 58
  Middle Tan Tien - Heart Center - Conscious Mind 61
  Recharge and Cut Around the Armpits 62
  Lower Tan Tien - Navel - Awareness Mind 63
  Activate Chi in the Bones of the Hips, Legs and Sacrum 64

Chapter IV
General Healing Session

Three Minds into One Mind (Introduction).. 68
Activate the Six Directions (Introduction)... 68
Activate the Three Fires (Introduction) 68
Linking Personal Stars, Energy Bodies and the Universe (Introduction) 69
Protective Circle and Chi Field (Introduction)... 69
  Sacred or Chi Fire... 69
  Chi (Sacred) Water Practice (Introduction).. 69
Empty and Fill (Introduction) 70
Clean with Green Light (Introduction).. 70
Clean with Blue Light (Introduction)... 70
Charge with Violet Light (Introduction) 70

Chapter V
Cosmic Healing I - Basic

Source of Chi and Colors 126
  Earth Chi 127
  Human and Nature Chi 128
  Solar Chi 128
  Simple Practice 128
  General Information 128
Healing Colors 129
  White Chi 129
  Blue Chi 129
  Green Chi 129
  Red Chi 130
  Tangerine Chi 131
  Yellow Chi 131
  Violet Chi 131
  Golden Chi 133
  The Spleen 133
  Cosmic Healing Planetary Colors 134
Summary - Working with Color 135

Chapter VI
Individual Healing Session

Almighty Knife 138
  Outline of Practice 139
Energizing and Activating the Immune System 141
  Outline of Practice 142
Cleansing the Internal Organs Technique 143
  Healing Some Common Illnesses 143
  General Principles 145
  Outline of Practice 146
  Cleansing the Spinal Cord 149
  Cleansing the Blood (Lungs) 150
  Outline for Cleansing the Blood/Lungs 150
  Circulatory/Heart System 151
  Stomach Pains 151
  Relieving Pain: Blue Chi 152
  Headache: Light Whitish Green and Light Whitish Blue Chi 153
  Migraine Headache: Light Whitish Green Blue and Violet Chi 153
  Toothache: Light Whitish Green and Blue Chi 153
  Broken Bone: Orange - Yellow Chi 153
  Back Injury: Light Whitish Blue and Green Violet Chi 154

  - Cysts: *Light Green and Blue Chi* 154
  - Infection and Inflammation: Light Green Chi 154
  - Fever 154
  - Insomnia: Green and Blue 155
  - Tinnitus 155
  - Blood Clotting: Blue Chi - Fresh Burns: Green and Blue Chi 155
  - Old Minor Burns: Green and Red Chi 155
  - Old Severe Burns: Green and Red Chi 156
  - Old Wounds: Green and Red Chi 156
  - Some Specific Points and Ailments: 156
  - Solar Plexus: Blue Chi 157
  - Skin Infection: Blue Chi 157
  - Eyes: Green Chi 158
  - Ears: Violet Chi 158
- Summary of Individual Healing Session 159

Chapter VII
Different Hand Techniques

- Forget about your Palms 165
  - Don’t Focus your Mind on the Sick Part 165
  - Cosmic Healing is Teaching 165
- Chi Kung Diagnosis: Hand Scanning 166
  - General Variations in the Energy Field 166
  - Hand Scanning the Internal Organs 167

Chapter VIII
Cosmic Chi Kung Palm and Finger Training

- Why We Practice the Cosmic Chi Kung 169
  - Opening Heaven and Earth Force in the Palms - Heaven and Earth Spirals 170
  - Cell Breathing 171
  - Mind, Eyes, Heart and Intention (YI) 172
  - Connecting the Bridge 173
  - Cosmic Chi Kung Preparation 174
  - Cosmic Chi Kung Preparation Standing 175
  - Opening to Heaven, Earth and Nature 176
  - Outline of Practice 176
- Part I - Opening to Heaven, Earth and Cosmic Chi 177
  - Opening Movements 177
  - Core Movements of Part I 182
  - Closing Movements 183
- Part II - Opening the Bridge and Regulator Channels 189
  - Theory 189

General Functions of the Eight Extraordinary Channels 190
The Eight Extraordinary Channels and Chi Kung 193
Special Functions of the Bridge and Regulator Channels............................. 193
Bridge Channels Yin and Yang Chiao Mo Function 193
Summary... 195
Practice 196
Part III - Opening the Functional and Governor Channels 205
Theory: The Functional and Governor Channels... 205
Chi Ball 206
Part IV - Activating the Yin and Yang Channels on the Chi Belt.......................... 217
Theory 217
Yin and Yang Channels of the Arms... 217
Projecting Healing Energy through the Fingers 217
Activating the Chi Belt - Dai Mo... 222
Practice 223
Healing Others.. 226
Summary and Combined Practice 226
Cosmic Chi Kung: Simple Combined Sequence 226

Chapter IX
Universal Tao Basic Practices

Cosmic Inner Smile 232
Six Healing Sounds.. 236
The Six Healing Sounds - Practice 235

Appendix 241
Guide to Acupuncture Points used in Cosmic Chi Kung 241
Bibliography.. 249

About the Author

Mantak Chia

Master Mantak Chia is the creator of the Universal Tao System and is the director of the Universal Tao Center and Tao Garden Health Resort and Training Center in the beautiful northern countryside of Thailand. Since childhood he has been studying the Taoist approach to life. His mastery of this ancient knowledge, enhanced by his study of other disciplines, has resulted in the development of the Universal Tao System which is now being taught throughout the world.

Mantak Chia was born in Thailand to Chinese parents in 1944. When he was six years old, Buddhist monks taught him how to sit and “still the mind.” While still a grammar school student, he learned traditional Thai boxing. He was then taught Tai Chi Chuan by Master Lu, who soon introduced him to Aikido, Yoga and broader levels of Tai Chi.

Years later, when he was a student in Hong Kong excelling in track and field events, a senior classmate named Cheng Sue-Sue introduced him to his first esoteric teacher and Taoist Master, Master Yi Eng (I Yun). At this point, Master Chia began his studies of the Taoist way of life in earnest. He learned how to circulate energy through the Microcosmic Orbit and, through the practice of Fusion of the Five Elements, how to open the other Six Special Channels. As he studied Inner Alchemy further, he learned the Enlightenment of the Kan and Li, Sealing of the Five Senses, Congress of Heaven and Earth and Reunion of Heaven and Man. It was Master Yi Eng who authorized Master Chia to teach and heal.

When Mantak Chia was in his early twenties he studied with Master Meugi in Singapore, who taught him Kundalini, Taoist Yoga and the Buddha Palm. He was soon able to clear blockages to the flow of energy within his own body. He learned to pass the life force energy through his hands also, so that he could heal Master Meugi’s patients. He then learned Chi Nei Tsang from Dr. Mui Yimwattana in Thailand.

A while later, he studied with Master Cheng Yao-Lun who taught him the Shao-Lin Method of Internal Power. He learned the closely guarded secret of the organs, glands and bone marrow exercise known as Bone Marrow Nei Kung and the exercise known as Strengthening and Renewal of the Tendons. Master Cheng Yao-Lun’s system combined Thai boxing and Kung Fu. Master Chia also studied at this time with Master Pan Yu, whose system combined Taoist, Buddhist and Zen teachings. Master Pan Yu also taught him about the exchange of Yin and Yang power between men and women, and how to develop the Steel Body.

To understand the mechanisms behind healing energy better, Master Chia studied Western anatomy and medical science for two years. While pursuing his studies, he managed the Gestetner Company, a manufacturer of office equipment and became well acquainted with the technology of offset printing and copying machines.

Using his knowledge of Taoism, combined with the other disciplines, Master Chia began teaching the Universal Tao System. He eventually trained other Instructors to communicate this knowledge and he established the Natural Healing Center in Thailand. Five years later, he decided to move to New York, where in 1979, he opened the Universal Tao Center. During his years in America, Master Chia continued his studies in the Wu system of Tai Chi with Edward Yee in New York.

Since then, Master Chia has taught tens of thousands of students throughout the world. He has trained and certified over 1,200 instructors and practitioners from all over the world. Universal Tao Centers and Chi Nei Tsang Institutes have opened in many locations in North America, Europe, Asia, and Australia.

In 1994, Master Chia moved back to Thailand, where he had begun construction of Tao Garden, the Universal Tao Training Center in Chiang Mai.

Master Chia is a warm, friendly and helpful man who views himself primarily as a teacher. He presents the Universal Tao System in a straightforward and practical manner, while always expanding his knowledge and approach to teaching. He uses a laptop computer for writing and is totally at ease with the latest computer technology.

Master Chia estimates that it will take thirty five books to convey the full Universal Tao System. In June 1990, at a dinner in San Francisco, Master Chia was honored by the International Congress of Chinese Medicine and Qi Gong (Chi Kung), who named him the Qi gong Master of the Year. He is the first recipient of this annual award.

In December 2000, the Tao Garden Health Resort and Universal Tao Training Center was completed with two Meditation Halls, two open air Simple Chi Kung Pavilions, indoor Tai Chi, Tao Yin and Chi Nei Tsang Hall, Tai Chi Natural Swimming Pool, Pakua Communications Center with a complete Taoist Library, Internal World Class Weight Lifting Hall and complete 8 Court Recreational Facilities.

In February 2002, the Immortal Tao practices will be held at Tao Garden for the first time using Dark Room technology, creating a complete environment for the higher level Taoist practices.

He has previously written and published these Universal Tao books:

Awaken Healing Energy of the Tao - 1983
Taoist Secrets of Love: Cultivating Male Sexual Energy
co-authored with Michael Winn - 1984.
Taoist Ways to Transform Stress into Vitality -1985
Chi Self-Massage: the Tao of Rejuvenation - 1986
Iron Shirt Chi Kung I - 1986
Healing Love Through the Tao: Cultivating Female Sexual Energy - 1986
Bone Marrow Nei Kung - 1989
Fusion of the Five Elements I - 1990
Chi Nei Tsang: Internal Organ Chi Massage - 1990
Awaken Healing Light of the Tao - 1993
The Inner Structure of Tai Chi co-authored with Juan Li - 1996
Multi-Orgasmic Man co-authored with Douglas Abrams 1996 - published by Harper/Collins
Tao Yin - 1999
Chi Nei Tsang II - 2000
Multi-Orgasmic Couple co-authored with Douglas Abrams 2000 - published by Harper/Collins
Cosmic Healing I - 2001
Cosmic Healing II co-authored with Dirk Oellibrandt - 2001

Also many of the books above are available in the following foreign languages: ***Bulgarian, Czech, Danish, Dutch, English, French, German, Greek, Hebrew, Hungarian, Indonesian, Italian, Japanese, Korean, Lithuanian, Malaysian, Polish, Portuguese, Russian, Serbo-Croatian, Slovenian, Spanish,*** & ***Turkish*** editions are available from the Foreign Publishers listed in the Universal Tao Center Overview in the back of this book.

Acknowledgments

The Universal Tao Publications staff involved in the preparation and production of *Cosmic Healing I: Cosmic Chi Kung* extend our gratitude to the many generations of Taoist Masters who have passed on their special lineage, in the form of an unbroken oral transmission, over thousands of years. We thank Taoist Master I Yun (Yi Eng) for his openness in transmitting the formulas of Taoist Inner Alchemy.

Thanks to Juan Li for the use of his beautiful and visionary paintings, illustrating Taoist esoteric practices.

We offer our eternal gratitude to our parents and teachers for their many gifts to us. Remembering them brings joy and satisfaction to our continued efforts in presenting the Universal Tao System. For their gifts, we offer our eternal gratitude and love. As always, their contribution has been crucial in presenting the concepts and techniques of the Universal Tao.

We wish to thank the thousands of unknown men and women of the Chinese healing arts who developed many of the methods and ideas presented in this book. We offer our graditute to Master Lao Kang Wen for sharing his healing techniques.

We wish to thank Colin Campbell for his editorial work and writing contributions on the first printing of this book, as well as his ideas for the cover. We appreciate his research and great labor. We wish to thank Matt Gluck for his editorial contributions on the revised edition of this book, as well as thanking our instructors, Dennis Huntington and Annette Derksen, for their insightful contributions to the revised version. We thank Dirk Oellibrandt for his technical editing and clear writing throughout the book.

A special thanks goes to our ***Thai Production Team*** for their cover illustration and book design and layout: Raruen Keawapadung, Computer Graphics; Saysunee Yongyod, Photographer; Udon Jandee, Illustrator; and Saniem Chaisarn, Production Designer.

Thanks to the testing institutes: Gerhard Eggelsberger, Institute for Applied Biocybernetics Feedback Research, Vienna, Austria and Dr. Ronda Jessum, Biocybernetics Institute, San Diego, California.

We also wish to express thanks to Cosmic Healing Certification Retreat Organizers around the world who have worked with us for many years and helped us prepare this manuscript throughout the Retreats: Beate Nimsky, Vienna, Austria; Serguei Orechkine, Moscow, Russia; Christopher Larthe, London, England; Masahiro Ouchi, Rye, New York; Dr. Angela C. Shen, San Francisco, California; Dirk Oellibrandt, Hamme, Belgium; Brita Dahlerg, Frankfurt, Germany and Lizbeth Cavegn, Zurich, Switzerland.

We wish to further express our gratitude to all the instructors and students who have offered their time and advice to enhance this system, especially Felix Senn, Barry Spendlove, Chong-Mi Mueller, Clemens Kasa, Andrew Jan, Marga Vianu, Harald Roeder, Salvador March, Dr. Hans Leonhardy, Peter Kontaxakis, Thomas Hicklin, Gianni Dell'Orto, Walter and Jutta Kellenberger.

Words of Caution

The practices described in this book have been used successfully for thousands of years by Taoists trained by personal instruction. Readers should not undertake the practice without receiving personal transmission and training from a certified instructor of the Universal Tao, since certain of these practices, if done improperly, may cause injury or result in health problems. This book is intended to supplement individual training by the Universal Tao and to serve as a reference guide for these practices. Anyone who undertakes these practices on the basis of this book alone, does so entirely at his or her own risk.

The meditations, practices and techniques described herein are ***not*** intended to be used as an alternative or substitute for professional medical treatment and care. If any readers are suffering from illnesses based on mental or emotional disorders, an appropriate professional health care practitioner or therapist should be consulted. Such problems should be corrected before you start training.

Neither the Universal Tao nor its staff and instructors can be responsible for the consequences of any practice or misuse of the information contained in this book. If the reader undertakes any exercise without strictly following the instructions, notes and warnings, the responsibility must lie solely with the reader.

This book does not attempt to give any medical diagnosis, treatment, prescription, or remedial recommendation in relation to any human disease, ailment, suffering or physical condition whatsoever.

Chinese Medicine and Chi Kung emphasizes balancing and strengthening the body so that it can heal itself. The meditations, internal exercises and martial arts of the Universal Tao are basic approaches to this end. Follow the instructions for each exercise carefully and do not neglect the foundations, i.e, the Microcosmic Orbit and any other supplemental exercises. Also pay special attention to the warnings and suggestions. People who have high blood pressure, heart disease or a generally weak condition should proceed cautiously, having prior consent from a qualified medical practitioner. People with venereal disease should not attempt any practices involving sexual energy until they are free of the condition.

The Universal Tao is not and cannot be responsible for the consequences of any practice or misuse of the information in this book. If the reader undertakes any exercise without strictly following the instructions, notes, and warnings, the responsibility must lie solely with the reader.

Chapter I

Cosmic Healing

Cosmic Healing Chi Kung in the Universal Tao System

Cosmic Healing Chi Kung is an important branch of the Universal Tao, an entire system of Taoist practices for cultivating the body, the Chi and the spirit. Your level of skill in Cosmic Healing Chi Kung will be based upon your mastery of the Universal Tao System as a whole. One of the most important parts of the Taoist practices is working with the energy moving within the body.

Foundation of the Universal Tao System is the Microcosmic Orbit Meditation

This is the major practice to connect with the earth, cosmic and universal forces. This is the way to make a connection with the North Star and the Big Dipper's violet/red light which is one of the very powerful universal healing lights. Through this practice you learn to feel Chi and to use your mind/eye/heart power to guide the Chi flow through the primary energy routes in your body. This practice is later extended to include the arm and leg routes too; it is then called the Macrocosmic Orbit.

Many people, including Chi Kung Masters, come to me and say that they have been practicing Chi Kung for years and nothing is happening. They don't feel any energy and they think that they must be performing their Chi Kung incorrectly. I tell them that they also need to do meditation. Hand movements are nothing by themselves. There are hundreds of different Chi Kung forms in China. You could spend seventy lifetimes just learning the hand movements. The practice of the Microcosmic Orbit meditation will help you to feel the Chi more easily inside, outside or around the body and in the extremities of the hands and feet.

When I was a child I liked to practice Chi Kung so much I had to save my lunch money to learn. After many years of practice of many Chi Kung forms, I started to forget the first form, so I learned it again. One day when I tried to practice and review all the forms that I had learned, I couldn't remember many of them. I sat down and thought, "I only have two hands, two legs and one head. Why are there so many forms? And why are there are so many religions and beliefs." I said to myself that there must be one main thing that they have in common. I started to search and I found out that the most important thing is feeling the Chi within us. Being able to increase, transform, take in and stay in touch with the universal, the cosmic, nature and the earth force. Letting them all combine within us is what is important. Likewise in religion there is God (a Force in Taoism we call Wu Chi, the nothingness, the supreme power controlling the universe); with a good heart and virtue energy we can connect with this force.

The complete practice of the Microcosmic Orbit Meditation is fully described in the book "Awaken Healing Light of the Tao" by Mantak Chia.

Cosmic Inner Smile and the Six Healing Sounds

These are very important practices to make the connection among the organs, the colors and the good virtue energy; each organ color will enhance your connection to the cosmic and the Universal Healing power.

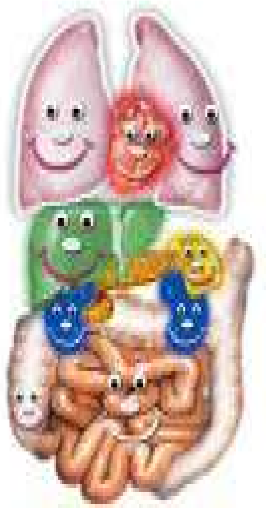

Fig. 1.1 *Red for the Heart*
White for the Lungs
Yellow for the Spleen
Blue for the Kidneys
Green for the Liver.

Each organ has its own vital color and when abundant it will radiate out as an aura for healing and protection. These organ colors and their associated universal connections have great healing power. The power of the 'Six Healing Sounds' will help enhance the connection to the cosmic source. Each sound will foster different healing energy. The practices also help balance, refine and transform the negative energy back to positive energy.

Iron Shirt Chi Kung and Bone Breathing

Iron Shirt Chi Kung and Bone Breathing are a system of standing meditations. These exercises help you to become grounded and to root your connection into the earth force, in order to build a good structure and absorb greater force. In turn, this will help draw in the heavenly force, which will combine with the very potent healing power of the earth's yellow light. The practice also strengthens the body so you can hold a higher energetic charge. These abilities are essential prerequisites for handling greater amounts of energy.

Iron Shirt Chi Kung also includes the art of ‘Changing the Sinews and Washing the Marrow’. Through these aspects of Iron Shirt, you learn to absorb, store and discharge large amounts of energy through the tendons and bones. *These practices are outlined in detail in the books “Iron Shirt Chi Kung” by Mantak Chia and “Bone Marrow Nei Kung” by Mantak Chia.*

To attain skill in Cosmic Chi Kung, it will help if you practice Iron Shirt, Bone Marrow and meditation as a minimum requirement. Beyond the basic level, the Universal Tao System includes many other intermediate and advanced level Chi Kung practices and meditations. The further one advances, the greater one's mastery of Chi.

Your increasing level of skill in the Universal Tao system will reflect immediately in your Cosmic Healing Chi Kung practice. Furthermore, you will discover that you can incorporate many of your Universal Tao practices directly into your practice of Cosmic Chi Kung. We will give you a simple combined practice of the Cosmic Healing Chi Kung, the Inner Smile, Microcosmic Orbit, Iron Shirt and the Sexual practices.

Stages of Mastering Chi

1. In the Universal Tao system, our *first* goal is to learn to conserve our Chi; when a battery is totally drained it is harder to charge; money makes money, Chi makes Chi. Conservation of Chi will help gain more Chi. To have more Chi we first need to maintain control of the gates through which energy normally leaks out and unwittingly drains our life force.

We leak energy:
- through our reproductive system
- through negative emotions
- through constantly turning our senses outward

Without knowing how to conserve the Chi that we already have, what is the point of acquiring more?

2. We learn to balance our Chi; that is, we seek to keep a smooth and balanced flow of energy moving throughout the whole body. If our energy is imbalanced, we may have too much energy in some places and not enough in others; we may also be too yang or too yin. We may have excess or deficient heat, cold, damp or dryness. This imbalanced energy tends to make us go to extremes.

3. We learn to transform our Chi into more beneficial energies. For example, through the Taoist Sexual Chi Kung practices taught in the Universal Tao (the course known as Healing Love through the Tao), we can transform sexual energy back into basic life force Chi. Through other practices (such as the Inner Smile, the Six Healing Sounds, and Fusion of the Five Elements) we learn to transform negative emotional Chi into positive virtuous Chi. Thus Chi is not only the foundation of our health; it is also the basis of spiritual development in the Tao.

4. Once we have accomplished the three previous phases of mastering Chi, we then learn to increase it. Chi pervades all of heaven, earth and nature. In Cosmic Healing

Chi Kung we learn time tested ways to tap into these unlimited and transpersonal reservoirs of Chi and greatly expand the amount of energy available to us. It is very important to master first the stages of conservation, balance and transformation before we emphasize increasing our Chi. Otherwise we may waste the energy we bring in, or we may inadvertently amplify the imbalanced or negative energies that we have not yet learned to bring under control.

5. Finally, we learn to extend our mind to tap into the vast Chi of nature, cosmic and the universe, to heal our body, mind and spirit and to heal other people. Cosmic Healing Chi Kung practice sensitizes your hands to the feeling and movement of Chi; it uses the mind-eye power to absorb cosmic Chi into the palm and crown and to send it out through the hands and beyond, so that you can help restore balance in others without touching them or draining yourself.

This may sound fantastic at first, yet recent researchers in Chinese Chi Kung hospitals have not only measured the energies emitted by Chi Kung masters; they have also discovered that different varieties and frequencies of healing Chi can be emitted. This research has been corroborated by experiments in the United States at such places as the Menninger Institute.

Four Sections of Cosmic Healing

The next step is to learn the four sections of the Cosmic Healing Chi Kung, the ‘Empty Force’ sequence. Each of the four sections of Cosmic Healing Chi Kung develops a different type of energy mastery. Each section begins and ends in the same way, yet has different movements in the middle.

Connecting to Heaven and Earth

The first section of Cosmic Healing Chi Kung emphasizes connecting to the external Chi emanating from heaven and earth. This ability is very beneficial for self-healing and is essential for healing others. If you do not connect to some source of external energy when healing others, you will draw upon your own personal reservoirs. Our personal Chi is limited by nature, and can easily become depleted if we give it away too freely.

Over the past decade external Chi healing has become very popular in China. There are now hundreds of Chi Kung hospitals and clinics throughout the country. Yet many of the Chi Kung therapists can only administer two or three treatments a day; they spend the rest of their day practicing Chi Kung to replenish their own Chi. Those who practice the more traditional Chi Kung know how to connect to the Chi of heaven, earth and nature.

In this first section, you will learn to sense the energies outside your body. You will then learn to draw and absorb these energies into your body and process them in order to use them for self-healing and healing others.

Opening the Bridge and Regulator Channels and Extending your Chi Outwards

The second section opens and strengthens four of the Eight Extraordinary Channels: the Yin and Yang Bridge and Regulator Channels. This section also teaches the student to extend and take in the cosmic Chi and emit Chi out of the body for healing others. This ability is the heart of Cosmic Healing Chi Kung practice.

Opening and Energizing the Governor and Functional Channels

The third section enhances the energy in the points along the two channels used in the Microcosmic Orbit. Opening the Microcosmic Orbit strengthens and balances one's own energy, and prepares one to heal others. Practicing this section will be of great help with the Microcosmic Orbit Meditation.

Activating One Finger Art and the Chi Belt

The fourth section of Cosmic Healing Chi Kung activates the hands and you will practice emitting Chi from each of your fingers and opening the "Chi Belt." This channel is also known as the Belt Route or Girdle Vessel.

Chapter II

Preparing for Cosmic Chi Kung

Preparatory Practices

In the Universal Tao, we teach Cosmic Healing Chi Kung within an entire system of exercise and meditation. As a part of this system, Cosmic Healing Chi Kung comprises the art of healing oneself and others. It focuses on energy work for projecting Chi to effect these healings. One can easily learn the simple movements of Cosmic Healing Chi Kung without doing any other Universal Tao practices, but if one truly wishes to master the art of Cosmic Healing Chi Kung, it is important to have a firm foundation in the basic Universal Tao practices. Therefore in this chapter we will introduce you to the basic Universal Tao practices that are used in conjunction with Cosmic Chi Kung.

The preparatory practice consists of a few parts; work through each part at your own pace and eventually join them together as a whole.

Always start with "warming the stove" at the abdomen and direct the fire down to the sexual center to transform the sexual energy. Next practice the Cosmic Inner Smile. The Cosmic Inner Smile is a powerful relaxation and self-healing technique that uses the energy of love, happiness, kindness and gentleness as a language to communicate with the internal organs of the body. Each organ corresponds to a specific element and color. For example, the kidneys will consist of the element of water and the color blue. The heart will consist of the element of fire and the color red. This makes it very easy to guide the healing power into each organ by using the appropriate color. The practice also aids the transformation of negative emotions into positive virtuous energy. This transformation is a very powerful Chi Kung practice. A genuine smile transforms negative energy into loving energy that has the power to relax, balance and heal. By learning to smile inwardly to the organs and glands, your whole body will feel loved and appreciated and enjoy more Chi. After the Inner Smile, practice Bone Breathing, Marrow Washing, Cosmic Healing and the Microcosmic Orbit.

Warm Up the Stove

1. Sit on the edge of a chair with your hands clasped together and your eyes closed.
2. Start by doing the "bellows breathing", moving your abdomen in and out quickly. Emphasize the exhalation by breathing out forcefully 18 to 36 times. Rest, cover your navel and feel nice and warm.
3. Next, do the "inner laughing" exercise; you feel the abdomen vibrate inside you. Practice

this for a few minutes, allowing the movement of your inner laughter to grow stronger. Then rest and use the mind and eye power to gather the Chi (now felt as a warmth behind the navel) into the Tan Tien. Picture a stove with burning fire behind your navel. Feel nice and warm.

Transform the Sexual Energy

Once you feel the Tan Tien is warm enough, smile down and bring the warm Chi down to the sexual organs; women bring the Chi down to the uterus, men bring the Chi down to the testicles. It may feel like the ‘sun shining on the water’. The rays of the sun purify the water until it becomes rising steam. Keep smiling down to the sexual organs and feel the warm or fiery feeling from the navel area continue to flow down to the sexual organs; this transforms the sexual energy into Chi, and this Chi rises up the spine into the brain helping to activate the ‘crown’ and ‘mideyebrow’ energy centers.

Focus your awareness in the sacrum. At the very tip of the sacrum there is a hole in the coccyx. Breathe into this hole until you feel some activity there. This may be felt as tingling, numbness or pulsing. If you can really activate this point it will generate 'suction'. Be aware of the sacrum opening. Feel the suction force pulling through the hole in the tailbone (sacral hiatus); breathe into it until you feel it becomes activated. When the sacrum is activated you will feel the suction easily, as well as breathing in the cranium and mideyebrow. Keep on gently smiling and softly breathing into the Tan Tien and feel the suction in the abdomen. Focus 95 % of your awareness in the Tan Tien and 5 % in the sacrum, the crown and the mideyebrow. Be aware of the Tan Tien breathing and observe internally the pulsing and breathing in the sacrum, the mideyebrow and the crown. Do this 36 times.

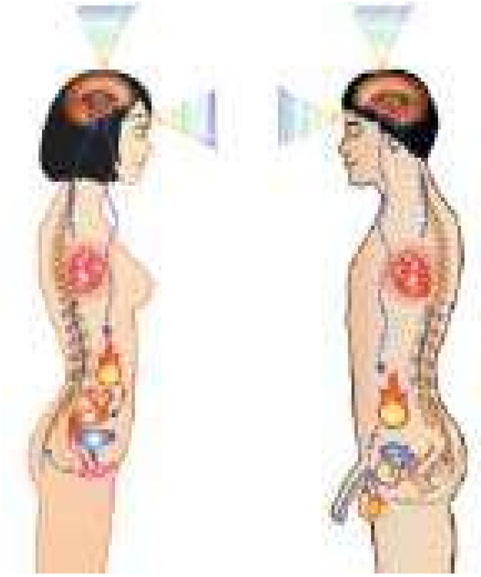

***Fig. 2.1** Smile down to the Lower Tan Tien.*

Filling the Joints with Chi

The bones have the ability to process Cosmic Chi (Chi above and around us) to be used by the body. The joints are also able to store Chi and they also serve as the cushion between the bones.

Be aware of your index fingers and thumbs; lightly raise the index fingers, feel a slight tension and open the thumbs to the sides and feel it magnify. Breathe into the finger tips, until you feel the sensation of Chi entering the fingers. The fingers will feel stiff as they stretch, to make way for the Chi entering the joints. Continue to breath in the Chi - the stiffness will spread to the other fingers, wrists, elbows and shoulder joints making the whole arm tense and stiff. The fingers and the arms become one piece. Eventually the Chi will fill every joint of the body making it into a complete unit.

Be aware of the big toes; breathe into the toes and feel the toes grow longer. When the Chi fills the joints the big toes become tense and stiff. Gradually it will fill the other toes rising to the ankles, knees, hips, sacrum and spine; the feeling may be one of tension and stiffness, numbness or like an electrical charge. The legs and the spine become one piece. You will certainly feel good.

Bone Breathing and Marrow Washing

Bone Breathing is one of the main practices of Cosmic Healing Chi Kung. This is a method of drawing external Chi through the skin into the bones to help replenish the bone marrow, thereby reactivating the production of white blood cells. Sending Chi into the bones will enhance functioning of the immune system. This process also helps to clean out fat in the bone marrow ("Washing the Marrow"), one of the main causes of osteoporosis (brittleness of the bones). Tension in the muscles close to the bones is decreased so Chi and blood can flow into the bones easily, enabling them to become stronger.

The Bone Breathing process uses the mind and the eyes to absorb Chi into the bones. The better your Bone Breathing is the better your Cosmic Healing Chi Kung practice will be. Mind and eyes are also used immediately after the exercise to gather the energy at the navel. Once you have it, when you move your hand, the energy will follow easily. You will be able to absorb external Chi effortlessly, so you will not need to use your own energy in your healing work.

Bone Breathing

There are several variations in the Bone Breathing and Marrow Washing process. Here we will introduce you to the first type of Bone Breathing, inhaling and exhaling Chi through the skin and packing it into the bones. In this method, you imagine that your bones are like hollow tubes, and that you breathe and suck the Chi into the bones.

Bone Breathing is practiced in a three-stage process. **Let your breathing follow a normal pace. Do not strain or hold your breath too long**.

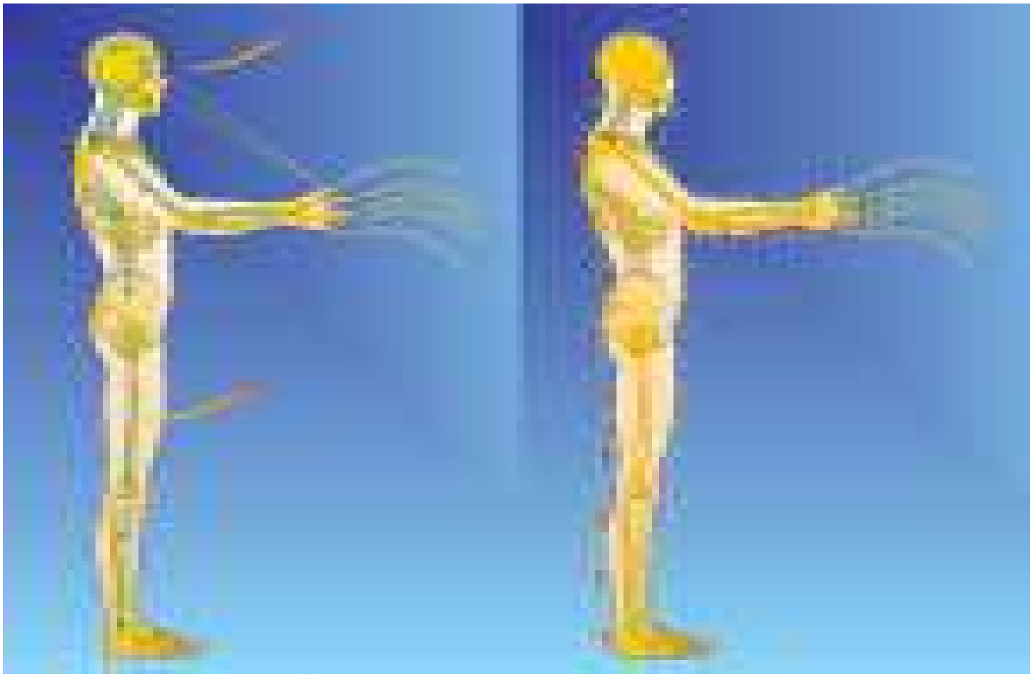

Fig. 2.2 *Bone Breathing through the Fingers*

1. You can do this practice in the sitting position or in the Embracing the Tree posture (or any other Iron Shirt Chi Kung posture). You will use your mind and eye power to breathe in a short breath and at the same time feel suction. Suck the Chi of the atmosphere into your hands, eventually expand to the universe and breathe in a few more times. Use a combination of 'mind/eyes/heart' power (Yi), to suck the Chi from the atmosphere into your hands, while taking small sips of breath. Once you can clearly feel the increase of Chi 'pressure' in your hands, you extend the feeling throughout your arms. The whole skin surface of the arms breathes in the Chi, feeling the skin holding this pressure.
2. Inhale one more deep breath and lightly contract the arm muscles to squeeze the Chi into the bones. Hold for a while to condense the Chi into the marrow of your bones. Exhale and at the same time feel a distinct heaviness in the bones, meaning that the Chi has been condensed and packed into the bones successfully. Eventually you will use more mental power (mind/eyes/heart power) and less muscle, using soft breathing, to draw the force into the bones. Do 6 to 9 times; rest and feel the Chi that has been condensed into the bones.

3. Proceed in the same way, breathing in progressively through the bones of the forearms, upper arms, scapulae, collarbone, sternum and ribs. You may feel a different sensation as you breathe in each area; in some areas the feeling is cool, in others warm or tingling, depending on the bone structure and the quality of the marrow.
4. Inhale and exhale in the same way through the toes, feet and legs; then, in a step-by-step progression, inhale up through the calf bones, thighbones, pelvis, coccyx and sacrum, and up the spinal column to the C-7 vertebrae.
5. Finally breathe in through the arms and legs simultaneously. Combine their energy as it flows up past C-7 and up through the neck and skull. Breathe in this way for at least nine breaths. Conclude by collecting energy at the navel.

Marrow Washing

You can wash your bone marrow with earth force, heavenly force or with cosmic force. This energy helps to cleanse and rejuvenate the bone marrow.

Heavenly Force Marrow Washing with Violet Light

1. Men place the left palm on the top of the head, and cover it with the right palm. Be aware of your 'Personal Star' (star that appears in your mind's eye and is located above your crown), above. Slightly press the palms and spiral in the clockwise direction, spiralling from the front to the right ear, to the back and the left ear 9 times; rest and feel the increased Chi pressure in the crown. Do 3 sets.
2. Women place the right palm on the top of the head and cover it with the left palm. Be aware of your Personal Star above you. Press the palms lightly and spiral in a counterclockwise direction, spiralling from the front to the left ear, to the back and the right ear 9 times; rest and feel the increased Chi pressure in crown. Do 3 sets.
3. Face your palms towards Heaven and feel that you are scooping a galactic Chi ball from above. This Chi ball contains the North Star (Violet) and the Big Dipper (Red). See the dipper fill with violet/red energy. Imagine you reach out your arms to hold the dipper handle and pour the violet/red liquid over your crown. It will feel like there is a numbness descending. Your palms face down to your crown and pour the whole galaxy onto your crown. You may perceive this energy as violet amethyst and red light frequencies. Smile.

4. Guide this sensation down into your skull, deep into your brain, cervical vertebrae, sternum, thoracic vertebrae, lumbar vertebrae and down through your legs. Feel it penetrating and enlivening your bones, deep into the bone marrow, washing, cleansing, energizing. This 'liquid-like' Chi spills all the way down to your feet. Feel it connecting with the earth through the soles of your feet; be aware of the bubbling springs in the feet (the K1 points of the kidney meridian) breathing and pulsating.
5. Touch your navel with the fingers of both hands. Focus on the door of life, and let the fire activate in the Tan Tien and kidneys. Chi will rise up to the brain.

6. Move your hands down touching your femur bones; feel your hands penetrate into the bone and into the bone marrow. Feel like an electric shock or a tingling moving all the way down through the bones of the leg and down to the soles of your feet.
7. Slowly move your fingers down touching the bone and into bone marrow and extend yourself down past the earth to touch the galaxy on the other side. Raise your sacrum and picture yellow light coming up from the earth and the other side of the galaxy; focus on the Lower Tan Tien and feel it fill with Chi. Do this three times. Go back down to the squatting position.

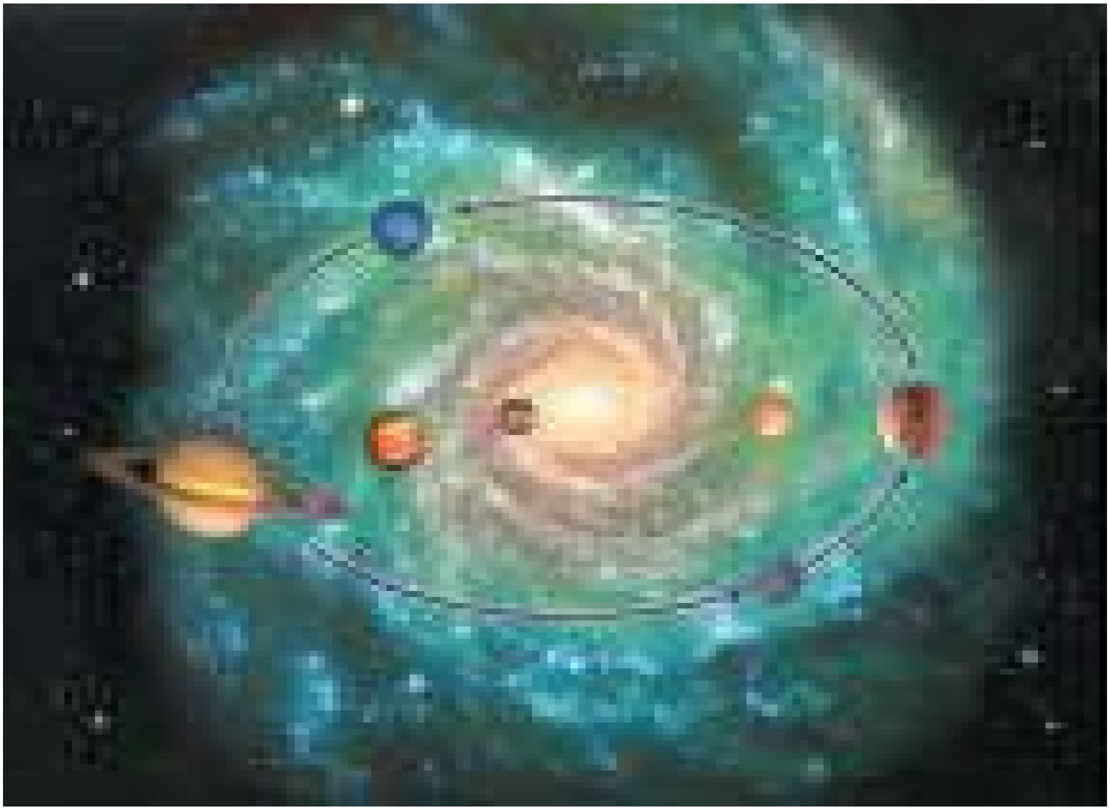

Fig. 2.4 *Galaxies Spiral Star and the Direction of Spiral*

8. Place the palms face down to the earth and be aware of the galaxy and the yellow light; make three to six circles to gather the Chi below. Gather the yellow Chi; touch your heels and feel the yellow Chi flowing up the leg bones. Feel the bubbling of your bones like an electric current running up through the feet, tibia and fibula, femur, pelvis, and spine. While touching your bone and bone marrow with your fingers, slide your hands up along the back of your legs; slowly come all the way up to touch the coccyx and the sacrum. Concentrate on your sexual center so the energy will flow and spread out to the sexual organs. Go back to the navel and gather the Chi into your Tan Tien.

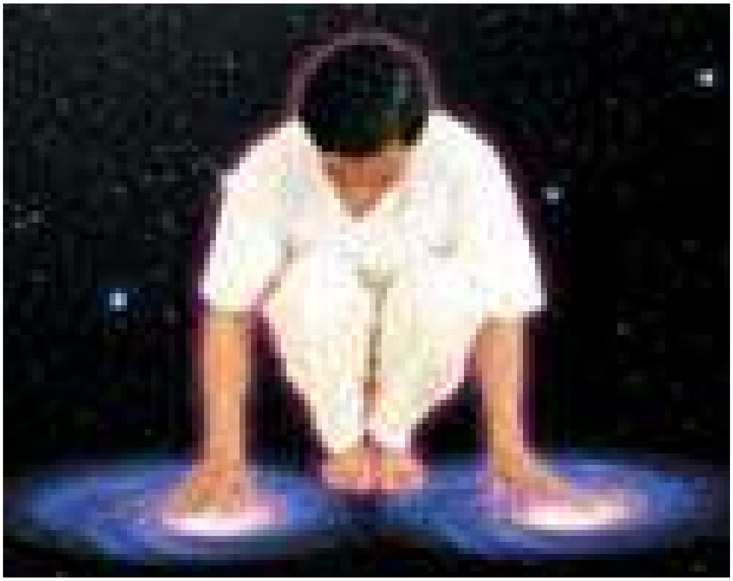

Fig. 2.5 *Hands circle to gather Earth Chi.*

Microcosmic Orbit

The Microcosmic Orbit Meditation awakens, circulates and directs Chi through the ‘Governing Channel’, which ascends from the base of the spine up to the head and the ‘Functional’ or ‘Conception Channel’, which runs from the tip of the tongue down the middle of the torso to the perineum. Dedicated practice of this ancient esoteric method eliminates stress and nervous tension, energizes the internal organs, restores health to damaged tissue and builds a strong sense of personal well being.

The Microcosmic Orbit is the foundation of Cosmic Healing Chi Kung practice. Each new practice is dependent upon the high quality of your meditations and your ability to perfect the Microcosmic Orbit. In order to master Cosmic Chi Kung, one must practice meditation daily. The meditations in the Microcosmic Orbit also strengthen the ‘Original Chi’ and teach you the basics of circulating Chi. They allow the palms, the soles of the feet, the mideyebrow point and the crown to open. These specific locations are the major points where energy can be absorbed, condensed and transformed into fresh new life force.

1. Focus on the Lower Tan Tien (the area where the Original Chi is stored, between the navel, kidneys and sexual organs). Feel the pulsing in this area, observe whether this area feels tense or relaxed, cool or warm, expansive or contracting. Notice any sensations of Chi: tingling, heat, expansiveness, pulsing, electric or magnetic sensations. Allow these to grow and expand. Then let this energy flow out to the Navel Center.

🔵 **2.** Use your intention (mind-eye-heart power) to spiral in the navel point guiding and moving the Chi. Let the energy flow down to the sexual center (Ovarian or Sperm Palace).

🔵 **3.** Move the energy from the sexual center to the perineum and down to the soles of the feet.

🔵 **4.** Draw the energy up from the soles to the perineum and to the sacrum.

🔥 **5.** Draw the energy up from the sacrum to the Door of Life (the point in the spine opposite the navel).

🔥 **6.** Draw the energy up to the mid-spine point (the T-11 vertebrae).

🔥 **7.** Draw the energy up to the base of the skull (Jade Pillow).

🔥 **8.** Draw the energy up to the crown.

🔥 **9.** Move the energy down from the crown to the mideyebrow point.

🔥 **10.** Touch the tip of your tongue to your upper palate, press and release a few times; then lightly touch the palate, sensing the electric or tingling feeling in the tip of the tongue. Move the energy down from the mideyebrow to where the tip of your tongue and palate meet.

🔵 **11.** Move the energy down from the palate through your tongue to the throat center.

🔵 **12.** Move the energy down from the throat to the heart center.

🔵 **13.** Bring the energy down from the heart to the solar plexus and feel a small sun shining out.

🔵 **14.** Bring the energy back down to the navel.

🔥 **15.** Continue to circulate your energy through this entire sequence of points, at least nine times. Once the pathways are open, you can let your energy flow continuously like a river of energy without needing to stop at each point.

🔥 **16.** Conclude when you wish by collecting energy at the navel.

Men: Cover your navel with both palms, left hand over right. Collect and mentally spiral the energy outward from the navel 36 times clockwise, and then inward 24 times counterclockwise.

Women: Cover your navel with both palms, right hand over left. Collect and mentally spiral the energy outward from the navel 36 times counterclockwise, and then inward 24 times clockwise. *For details of this practice see the book, "Awaken Healing Light", by Mantak Chia.*

Chapter III
Cosmic Chi Kung

Opening the Three Tan Tiens to the Six Directions

Introduction

Opening the Three Tan Tiens is a Chi Kung meditation that strengthens our connection to the universe, opening us up to the primordial force of the Cosmos and the energy within Nature. We are dynamically connected to the infinite. "As above, so below" is an echo of wisdom heard from sages and mystics throughout the ages. When we can connect to and absorb the energy that surrounds us, we are able to tap into the many splendors of the universe.

We exist because of the unique combination of the forces that are around and within us. The two main forces are electricity and magnetism. 'Bio-electro magnetism' is the western term for life force, and what the Tao refers to as Chi. For the last 5000 years, the Taoists have utilized this bio-electro-magnetic energy to enhance their way of life and establish a relationship with the universe. Bio signifies life, electro refers to the universal energies (yang) of the stars and planets and the magnetic force refers to the earth force (yin) or gravitational force present on all planets and stars. As we align ourselves with these forces, we become a conduit through which we can absorb and digest these energies through the body, mind and spirit establishing a direct connection with the universe. The Taoists recognized this connection and created the Chi Kung form of 'Opening the Three Tan Tiens to the Six Directions' to enhance our relationship to and our understanding of this connection.

Humans normally access bio-electro-magnetic energy through food and air. Plants take the universal energies of the sun and the magnetic energies of the earth and digest and transform them, thereby making these energies available to all living beings. Taoists believe that the food sources with the purest form of energy are the green leafy vegetables. These have taken sunlight directly into their cells. Rather than waiting until the energy in the universe is processed through plants, the Taoist goes directly to the source of this primordial energy. Through Chi Kung and meditation, the Taoists direct the energy of the universe precisely. Opening the Three Tan Tiens is a meditational Chi Kung exercise that focuses on how to directly tap into the source of energy all around us.

The Tao views human beings as lamps filled with fuel. Many people burn this fuel at very high intensity, without ever taking the time to replenish the oil in the lamp. Alcohol, drugs, tobacco and promiscuity all quicken the depletion of this fuel. The Taoist exercises strive to continually refuel the energy within. The Taoist recognizes that we are limited in our nature as human beings, unless we connect the sources of energy within the universe, thereby becoming infinite. So, within the limitations of our human nature, we constantly fill ourselves with the unlimited abundance of energy around us.

Through their internal quest, the Taoists discovered a doorway to the universe. The more we open our internal energy, the more we are capable of connecting to the forces of energy around us.

Human beings have amazing potential and capabilities. We are unique creatures in the way we use our minds and hands. Look at the world around you, the skyscrapers, the architecture, computers, technology and the myriad creations of man. All have come about through the combination of the mind with the hands. In the Tao practice, we use the mind and the hands, in both Chi Kung and meditation, to connect to the forces of the universe. We use the mind to project a pattern of energy into the universe, to connect to the force, and to bring this energy back into the body.

The mind can travel millions of light years in a few moments. Taoists discovered the unlimited potential of the human mind. If you picture something in your mind; an ocean, a sunset or a mountain, you automatically connect with that image. The mind can take you anywhere you imagine. With the proper training you can connect to the energy of nature and the universe and project your Chi to combine with it, drawing these forces into the body. The hands are called the touching force, and like antennae they can transmit the frequencies and vibrations from space. With the mind and the hands, each of us can journey into the boundless energy of the universe.

Relaxation, Letting Go, Surrender and Ego

In some religions and spiritual paths, there is a great emphasis on surrender and letting go. This is actually a form of relaxation. Taoist practices emphasize relaxation, letting go, and emptiness. When a person is relaxed, the muscles are open, the breathing is soft, and the energy can flow through the channels of the body. There is no resistance and no fighting. This allows the creative and higher forces to flow into us.

Most religions have a similar process. To contact the higher self and the higher forces, one needs to let go and surrender in order to reach the level where one can be in contact and become one with the higher forces. Through the surrender of control, one opens up and touches the forces of nature. However, if you continue to surrender and let go, you will lose the energy you have sent out. In the long term, this will gradually drain you. The force will suck the energy out of you, rather than help you to bring the energy into yourself. To avoid this, at the moment that one is in touch with the higher forces, one again becomes aware of oneself and his or her own energy. One is then able to project his or her own thoughts, intentions and patterns into the force, integrating the outer with the doubling or tripling inner energies. You can bring this force back to your place, your house and into yourself.

One must remain open and empty while using one's intentions, mind and Chi to draw the energy into oneself. What might initially appear to be a paradox is reformed as one does the practice and learns how to be empty and open and to simultaneously retain enough consciousness to draw the force into oneself.

Mind, Organs and Sexual Organs

It seems that some people are disconnected from themselves and from their sexual organs. The mind and organs are therefore separate. Taoism believes that the mind, body and spirit must work together. The results depend on a person's practice.

Brain

The brain can access and generate the higher forces, but storing this energy in the brain itself is not easy. We need to train the brain to increase its ability and capacity to store energy. The brain energy, when increased to a certain level, can enable more synapses to grow, and can help turn protein into brain cells. The Tao believes that with training and practice, one can learn to grow more brain and nerve cells, as well as increase the number of synapses in the central nervous system.

Organs of the Body

The organs can also generate energy, but much less than the sexual organs and the brain. They also have a much greater capacity to store and transform energy.

Sexual Organs

The Tao discovered that the sexual organs are the only organs that can generate a significant amount of sexual energy (life force). However, the sexual organs cannot store the energy efficiently. When too much energy is generated, considerable amounts have to be dumped out. And this is the best energy that one has. It is the ‘creative’ energy.

Three Tan Tiens

The Three Tan Tiens can also store energy, transform it and supply it to the brain, spinal cord, sexual organs and other organs.

The aim of Taoist basic training is to integrate the brain, sexual organs and other organs into one system. If the brain generates too much energy, it can store the energy in the organs. The excess sexual energy can also be stored in the organs and the three Tan Tiens. If the brain generates too much of the higher forces and we are unable to store this energy, we have to throw it away. It is like preparing food for one hundred people, when only one person is eating. The rest gets thrown away. In the same way, when too much sexual energy is produced and there is no place to store it, it is wasted. We do not have enough of this energy to be able to waste it. We have a limited amount of energy and of time.

Some practices just deal with the spirit and ignore the body and the sexual energy. These practices can generate a lot of energy, but when one is not connected to the organs, it cannot be stored anywhere. One will end up draining oneself. Some people practice sitting quietly, emptying the mind, with the whole body relaxed and calm, but very little energy is actually generated. When one gets deep into the practice, some people find it hard to come back to society, because they have no energy and their mind power does not work well. These people have to depend on others to support them.

In the Universal Tao, we are learning to create a sacred and holy temple within ourselves. With the simple practice of smiling to all the organs, we can integrate our bodies, minds and spirits. They are no longer separate. The sexual practice connects the mind with the sexual organs and the brain. The separation between these parts of ourselves is bridged and a synergy is created.

The Taoist practice provides us with the resources to extend beyond the realm of our senses. By tapping into our internal resources and channeling the energy around us, we can perceive much more than the senses normally report to the mind. We want to extend our perception from the limited perspective of the sociologically conditioned senses to the unlimited awareness of the universe. For example, our senses tell us that the earth is flat, that we are stationary, and that heaven is above us. In reality, the earth is a sphere hurtling through space at thousands of miles per hour and the heavens are above, below and beyond the earth in every direction. The goal of the Three Tan Tiens is to connect with the forces from the six directions - above, below, left, right, front and back - and draw all these forces into the body. Eventually, with practice one can draw upon many different energies and use them as needed, thereby giving form to the formless energy that is abundant in nature.

Opening the Three Tan Tiens

Opening the Three Tan Tiens to the Six Directions is just another one of the many resources the Taoist practitioner uses to connect with the universe. The practice combines the power of the mind with the extension of Chi. This combination allows our personal consciousness to directly connect to the patterns and matrices of energy in the universe. When we put our thoughts into the web of the universe, we transform the electromagnetic energy into a force that is accessible to us. The combination of mind power and energy is what allows us to establish a relationship to these creative forces and the high sources of energy.

Three Tan Tiens

Once we make the connection with the forces of energy in the universe, we then want to be able to store this energy in the body. Energy is like money, if you are making a million dollars a year and spend a million dollars, you have nothing left to use in the future. That is the way we live and use our energy in our society. We are spending more energy than we are saving and we are living on borrowed energy, paying very high interest. Our credit will run out very soon.

In the Tao practice, we store energy in the Three Tan Tiens. The Three Tan Tiens are the reservoirs of energy within us.

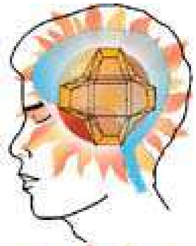

__Fig. 3.1__ The __Upper Tan Tien__ is in the brain (the __Crystal Room__, third ventricle), and when it is full of energy, the capacity of the brain increases. We store our spiritual intelligence, the mind here. All the Tan Tiens have both yin and yang within them. In nature, the yin and yang are present in all things. Day (yang) turns into the sunset, which turns to night (yin). It is very important to feel the qualities of yin within yang and yang with yin (sunrise/sunset). One quality does not exist without the other. They are inseparable qualities of the same force.

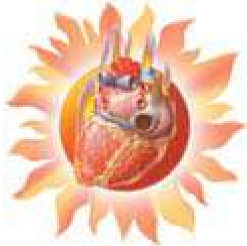

__Fig. 3.2__ The __Heart__ Center Tan Tien, between the two nipples, is the __Middle Tan Tien__. It is associated with the fire element. Yet, within fire there is always water. The original spirit (Shen) is stored here.

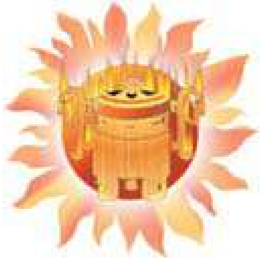

__Fig. 3.3__ The lower abdomen at the __Navel__ is like an empty universe, or ocean. We want to feel a universe of energy in the __Lower Tan Tien__. Within this universe or ocean, there is a fire, like a volcano under the ocean; 'fire under water'.

The Three Tan Tiens refer to the three reservoirs of energy within the body. These reservoirs are places where we can store, transform and collect energy. The reservoirs are the source of energy that flows through the body. The meridians are rivers of energy fed by these reservoirs. The goal of Opening the Three Tan Tiens is to continually fill and replenish the energy of the Three Tan Tiens. When we are connected to the Tao, life ceases to be a struggle. Through the observation of nature, the Taoists learned to flow with the stream of energy and connect to forces in the universe.

In these practices, we use a variety of hand movements and body postures to open to the energy around us. We draw the energy from the six directions into the body, activate the three fires, open the Three Tan Tiens and circulate this energy in the Microcosmic Orbit.

Combined Practice

Warm - Up Exercises

Rotating the Sacrum

Rotating the sacrum is an excellent exercise to open the lower back and activate the spinal cord. Place one hand over the sacrum and the other over the pubic bone. Rotate the sacrum in a circle, 36 times in each direction. This movement activates the sacral pump.

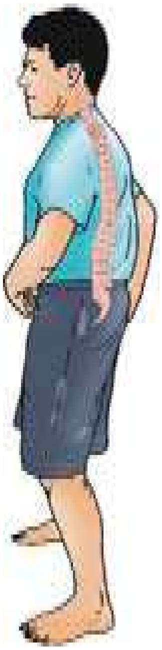

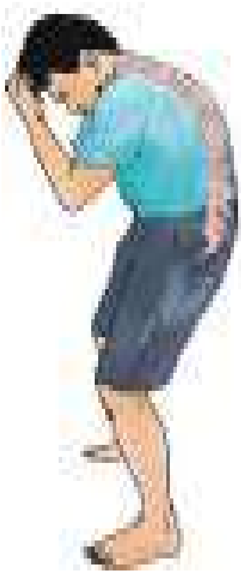

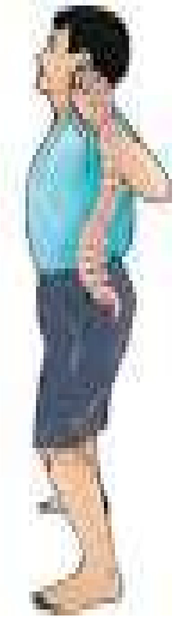

Spinal Cord Breathing

Inhale and expand the chest, arms bent at the elbows and extended to the sides of the body. Exhale, tuck the tailbone under you and round the back, bringing the elbows toward one another in front of the chest. Smile. Inhale, expand the chest, tuck the chin in toward the throat, push the chin back and raise the crown and bring the arms out to the sides. Repeat this back and forth movement 36 times. This movement activates the cranial and sacral pumps, and loosens all the joints in the spine.

Shaking

Shake the whole body loose, especially the joints, by bouncing up and down on the heels. Let all the joints open and relax. Also, shake the testicles and the breasts loose. This will open the sexual energy (the yin electric) and you do this on back as well.

The yin and yang are the negative and positive electric forces in the universe. The heart produces impulses (the yang electric) and the adrenal glands produce the adrenal hormone to stimulate the heart to beat. Of all the hormones inside us, this is the one that gives the vital force to our life.

When we concentrate on the sexual organs, the heart and the mind, we can make the yin and the yang electricity combine.

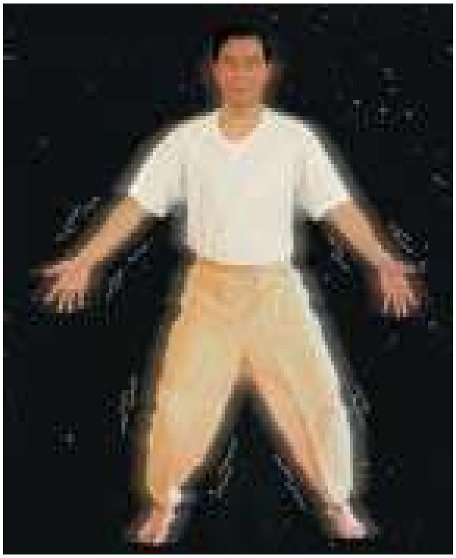

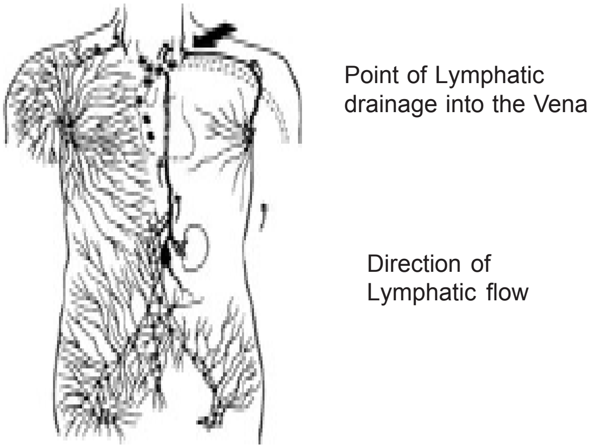

Point of Lymphatic drainage into the Vena
Direction of Lymphatic flow

Fig. 3.4 *Whole Body shaking and Lymphatic System*

True Breath - Skin Breathing

We exist because of the physical food and the unique combination of forces that surround us. The daily requirement of calories is about 6,000 units, while we only get 2,000 calories from food. The other 4,000 calories come from the forces around, above and below us. These forces that surround us are electricity, magnetism, cosmic particle energy, light, sound and heat. If we don't know how to absorb and transform this cosmic food, we need to depend on others to supply us. We then need to ask a priest, monk or holy person to give us our daily spiritual food.

The Taoists discovered that we can learn to absorb these surrounding and universal energies through the skin and the major energy centers. Absorbing energy through the skin is called the True Breath. This powerful energetic technique requires the Inner Smile and relaxation. The more one can relax, the more the body and the skin can open to the energy around us. The practice allows one to extend the mind, to touch the force and to draw that energy back into the body.

Daily Practice: Opening, Connecting and Activating Chi Within and Cosmic Charging

Paradoxically, in order to project ourselves out into the immensity of the galaxies and the universe to gather limitless resources of Cosmic Chi for healing, we must take the first steps of the journey within ourselves. In order to 'go out', we must first 'go in'. The vehicle for this magical journey is powered by our ability to relax in mind and body. As we physically relax and let go of muscular and emotional tensions and joint and bone structures we gain access to the inner realms by turning on our very special subtle smile. It opens the pathways of the parasympathetic nervous system. This helps us to reduce the out flowing habits of our senses so that we can be more alert in sensing our inner universe.

This simple process of 'going in' enables us to develop our internal skills so as to safely 'go out' to the universe.

Practice

Relax and Let go - Smile to Connect with the Universe Within

Smile to mideyebrow, eyebrows, eyes, mouth, jaw, tongue, lips, cheeks, ears, shoulders, ribcage and brain. Let the relaxed sensations and the 'Observer Mind' (Upper Brain) sink down into the Lower Tan Tien.

1. Smile to the mideyebrow. Relax and let go. Smile to the eyebrows and let them grow long to the sides. Lower these relaxed sensations down to the Tan Tien.

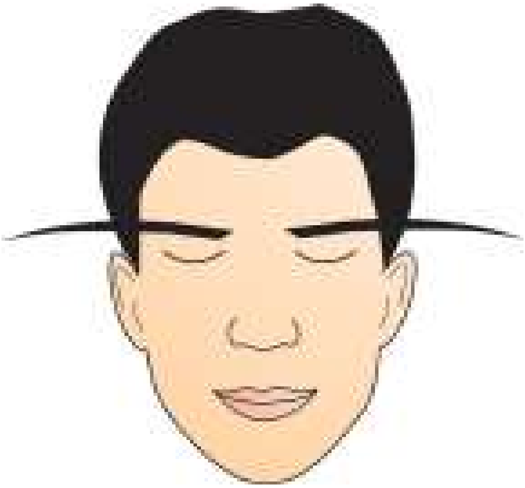

***Fig. 3.5** Smile and Relax to the Mideyebrow and think of the eyebrows growing long.*

2. Smile to the eyes: relax the eyes and feel how nice and cool they are. Let the eyes draw back in their sockets and start to sink down to the chest and gradually down to the abdomen, the home of your 'feeling and awareness mind'.

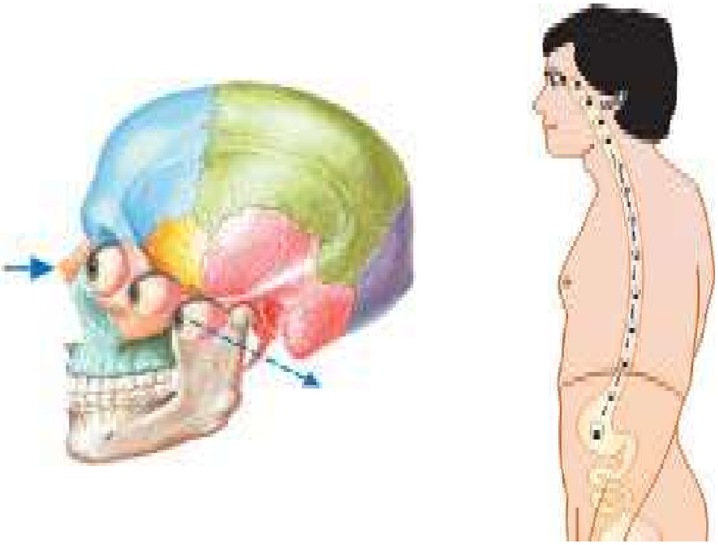

Fig. 3.6 *Smile and relax the eyes, sink into the eye sockets and gradually feel the eyes dropping down into the Abdomen.*

3. Relax the two broad muscles extending from the outer portions of the upper lips across the cheekbones, and lightly smile feeling their connection to the upper front of the ears. Gradually feel the ears growing 'long', up and down. Feel the ears grow all the way down and connect to the kidneys.

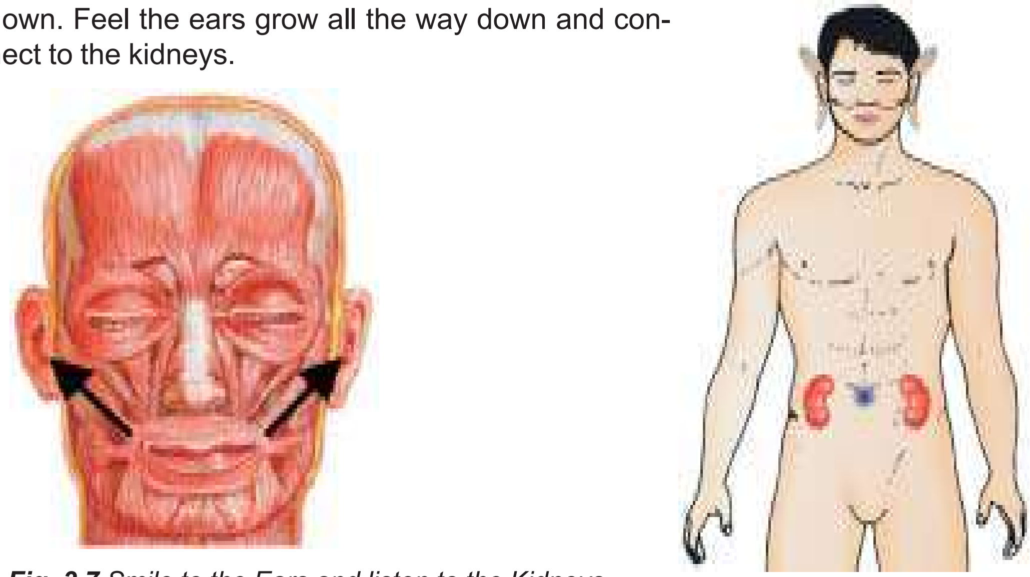

Fig. 3.7 *Smile to the Ears and listen to the Kidneys.*

4. Open your mouth and relax your jaw, separating the upper and lower teeth. Feel the jaw relax. Once the jaw relaxes the shoulders will relax and drop down. Continue to feel the jaw relax until you feel the saliva start to come out. Relax down to the rib cage. Feel the rib cage drop down, softening all the joints, relaxing down to the Lower Tan Tien. Let the tongue relax back in the mouth. Feel the tongue start to drop down into the throat to the chest and all the way down to the navel sinking the 'floating' sensation down to the Lower Tan Tien.
5. Smile to the shoulders and relax until you feel the shoulders drop and the rib cage relax.

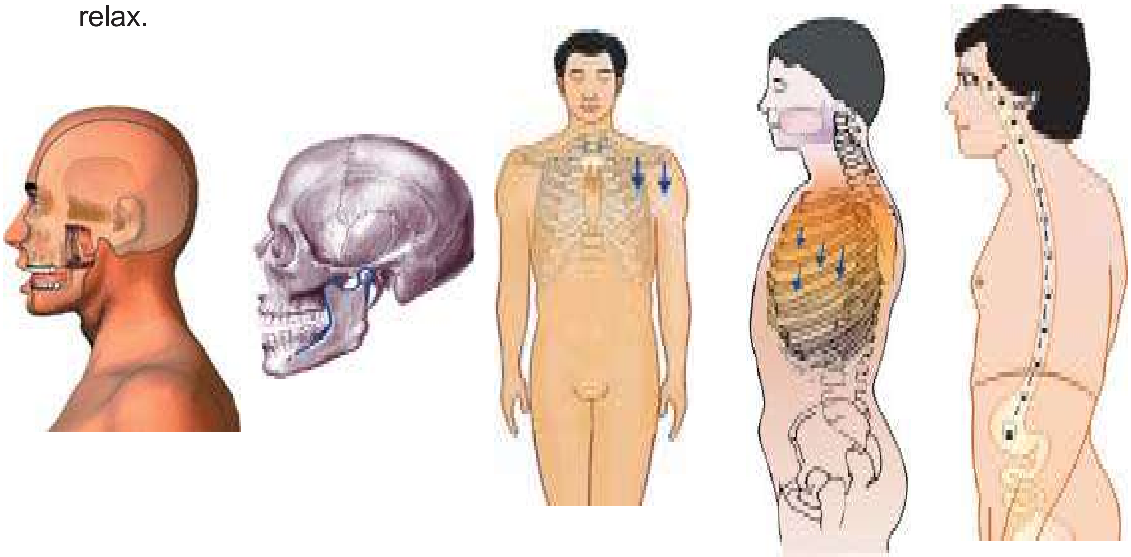

***Fig. 3.8** Relax the Jaw; and the Shoulders and Tongue will relax down to Lower Tan Tien.*

6. Lightly close the lips, but keep the teeth slightly separated. Physically begin a child-like smile with the corners of the mouth gently uplifted and the outer edges of the eyes softly crinkled up. Breathe through your nose.
7. Smile into the brain and empty the upper (observing) mind down to the Tan Tien.
8. Become aware of your inner universe as a big empty space. Keep on sinking – sinking down, into the darkness of your empty space. Keep sinking and experience the vastness until you get closer and closer to the center, the 'original force.' Stay relaxed and alert so as to be able to see one dot of light. It becomes a galaxy spiraling inside you.

   Do the above practice several times until you become familiar and comfortable with it.

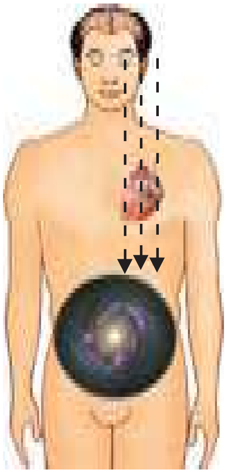

***Fig. 3.9** Bring the Senses down to the Tan Tien.*

Tan Tien Consciousness: Second Brain

Already terms have been used like 'upper mind', 'feeling and awareness mind' in the abdomen and lower 'Tan Tien'. In the next phase of practice, the 'conscious mind of the heart' is introduced. For those readers who have not been updated on the scientific and medical revelations of the recent era about a 'second' and 'third' brain, we offer the following information.

Along with this, there is the long-standing Taoist practice of cultivating and training consciousness in the Three Tan Tiens, especially the Lower Tan Tien. There is my own experience in participating in brain wave measurements and 'brain energy potential' research in recent years. The combination of all these areas of knowledge and experience in our time, makes Cosmic Healing practices accessible for our minds while also being very 'down to earth.' Now let's talk about the Lower Tan Tien.

In addition to its importance as the control center for the mechanics of the physical structure of the body, the Lower Tan Tien also houses a treasure of even more far-reaching significance. It has been a well-kept secret in the western world, as well as in most of the rest of the world: our Second Brain. Most of us who have had Taoist training in Chi Kung, Tai Chi, various Chi meditations or healing practices have often heard the reminder, 'Be aware of your Tan Tien.' But, do most people really 'get' the meaning of the injunction to always be conscious of the Tan Tien? Probably not. Further, do we use our Second Brain as much as we can? Definitely not.

In western terms, do we actually understand "Be aware of your Tan Tien" as being a way to train consciousness and awareness, like educating a brain in the abdominal area, in our Tan Tien? Probably not. Throughout the world, there are institutions to train the brain in the head. That is good. But, what about training the 'second brain' in the abdominal area? I didn't think of it exactly like that in western terms, either, even though that is exactly what I have been doing all of my life in my Taoist training and teaching.

Personal Revelation

Suddenly I came to understand a few things which are so simple and so important, and that's what I am going to share with you. It started in 1994 in Los Angeles when a clinical psychologist there, Dr. Rhonda Jessum, wanted to start testing me. I agreed to do the testing, but the machines at that time didn't tell me much. However, it was discovered that when I did the Inner Smile meditation, my brain waves went down dramatically – but at the same time, my 'beta' waves increased to a very high level. That means the waves showed that I could be driving a car, but that my brain should have been resting and sleeping. So the researchers said, "Hey, how did you do that?" I didn't understand either – because it was not clear to me either.

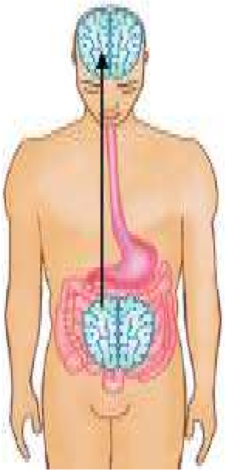

Fig. 3.10 *First and Second Brains are linked.*

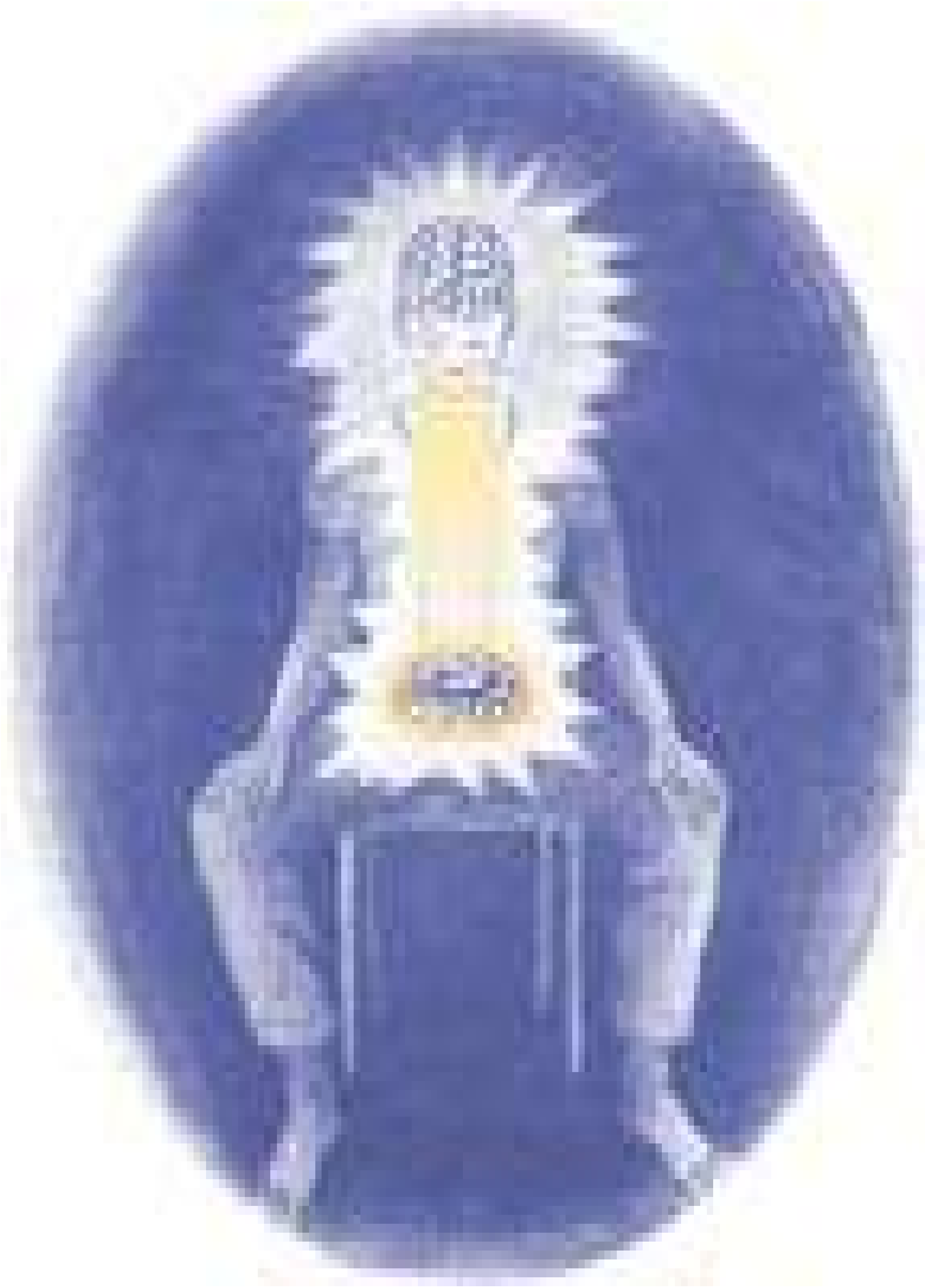

Fig. 3.11 *Second Brain*

After that, I was invited to start testing in Europe by the Institute for Applied Biocybernetics and Feedback Research. My name is becoming known there because I teach in Europe a lot. One of the biggest institutes for training top athletes in Europe is in Vienna. They developed an instrument that can measure the brain's potential energy, which represents all the energy in the body. The doctors also said that this proves to the west that Chi exists, there is energy and a life force running in the body. That instrument picks it up in the brain and indicates how much potential brain energy you have. It also determines how much energy you have for the whole day and how much of that energy is for mental clarity and body power. They use these instruments to measure the athletes.

The study measures brainwave frequencies during four practices.

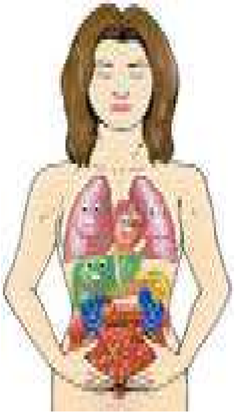

Fig. 3.12
Cosmic Inner Smile

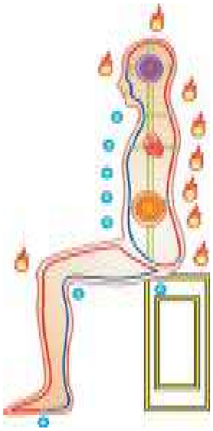

Fig. 3.13
Cosmic Orbit

Fig. 3.14
Six Healing Sounds

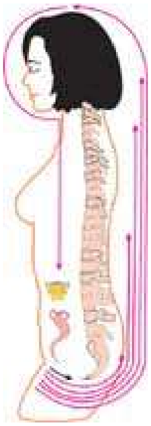

Fig 3.15
Orgasmic Upward Draw

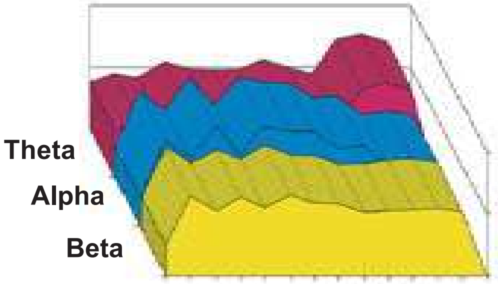

Fig. 3.16 *Alpha and Theta Waves increasing during the Cosmic Inner Smile Meditation.*

Fig. 3.17 *Ultraslow Brain potentials measurement while in Vienna.*

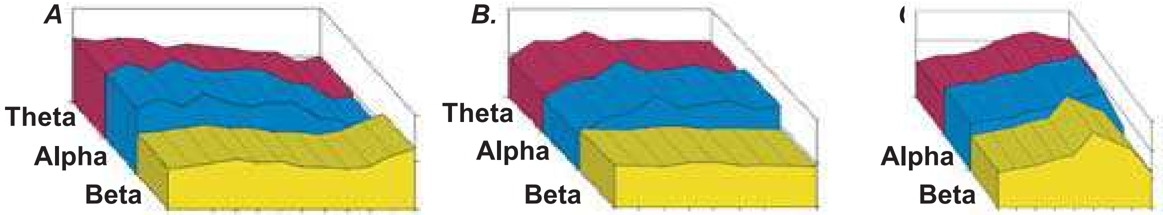

***Fig. 3.18 A.** Altering dominance of alpha and theta waves during the Cosmic Orbit Meditation.*
***B.** Cosmic Healing Sounds Meditation, results in increases in alpha and theta frequencies.*
There are minimal beta brainwaves; the brain reaches a state of stillness and internal focus.
***C.** Orgasmic Upward Draw Meditation increases the levels of alpha and theta.*

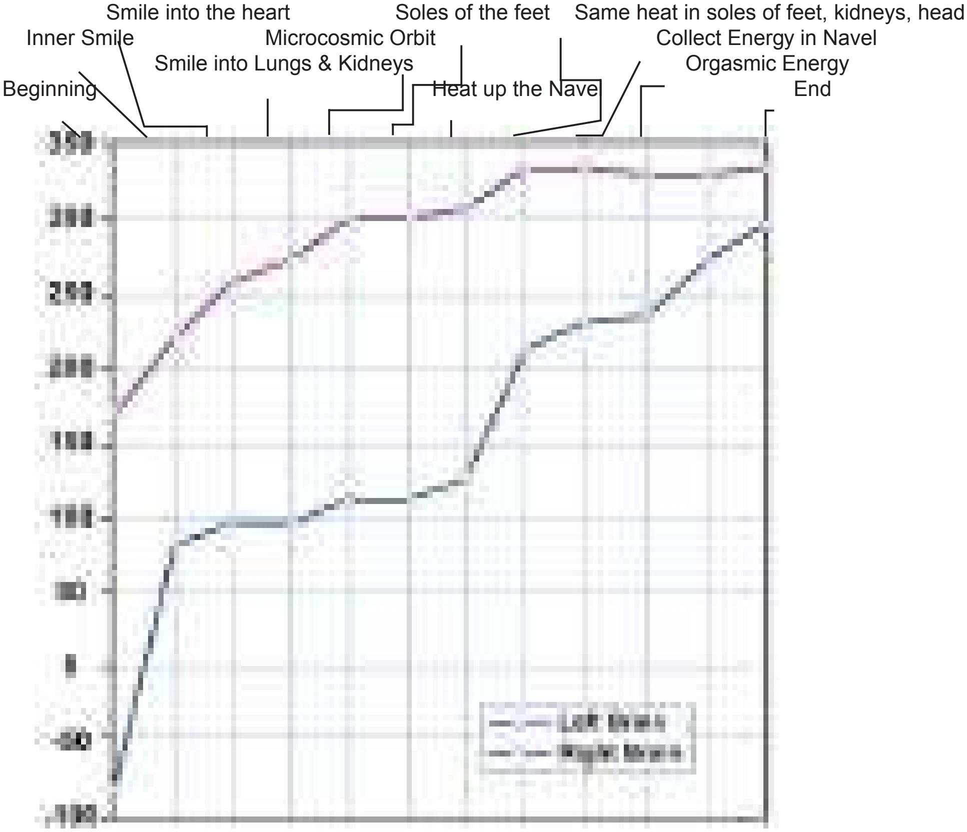

***Fig. 3.19** Left and Right Brain Charts*

It's interesting because the energy I have been describing in my teaching is exactly what they were measuring during their testing. So I went there and did the measurements with them. I just did the Inner Smile, smiling into my abdomen. They picked up these readings very quickly and said, "Your brain waves are going lower, lower and resting – and you are nearly in the sleeping state." At the same time, the muscle tension was very low, the heartbeat was very low and the skin resistance was very low. After that, I surged the energy up to the brain and they started to see that the energy actually charged up there. When we are thinking, worrying or feeling anger, shame or guilt – the energy level in the brain actually decreases, and the brain doesn't get charged up. They were quite amazed, and said, "Hey, this is what we're looking for!"

They asked me what I was doing. I said, "I'm smiling to my abdomen." They kept on talking to me and asking me questions. They discovered that this brain (in the head) was not very active at all, meaning that there was not much activity in the brain. It was still in a very light resting state. But, how could I answer their questions? They said, "Hey, look! Master Chia is talking to us in his sleep. How can he talk to us in his sleep?"

After that, I said, "Oh, I understand now." Because, throughout the whole practice of the Tao is the injunction, 'train the second brain in order to use the second brain.' It took the westerners a long time to understand that. So, when the article about the 'Hidden Brain in the Gut' came out in The New York Times,[1] I started to understand that. You can be miserable; you can be happy; you can feel all kinds of feelings. But according to this article, they had also discovered that this brain in the gut, the enteric nervous system, can do a lot of functions. It says that this gut brain can send and receive impulses; it can record experience and respond to emotions. So, it is like a brain. In this article, they had only discovered that the large intestine and the small intestines have the same neurons as the brain cells. After that article, a new book, *The Second Brain,* was published.[2]

Now in medical science, I might add, they have begun to discover consciousness in the heart. They found that the heart can record a whole event, and it has its own brain. Dr. Paul Pearsall has written a new book, called *The Heart's Code*.[3] They have found that people who have a heart transplant actually get the emotions of the donor. They actually experience whatever the donor experienced. One of the cases is that of a girl who was brutally killed, and they didn't know who killed her. But, the heart was OK; so, it was transplanted into another girl. After that, the girl started to get nightmares and described somebody killing her. Then she described how the killer looked. Finally, the mother took the girl to a psychiatrist, and he in turn contacted the police. The girl gave the police an exact description and a police artist drew a likeness of the killer. With the information provided by the girl, the police were able to go and arrest the man. Afterwards, when confronted with clear details of the crime, the man confessed that he was indeed guilty of this crime. So, from that experience, medical scientists and others came to realize that the heart can record all of an event and remember it.

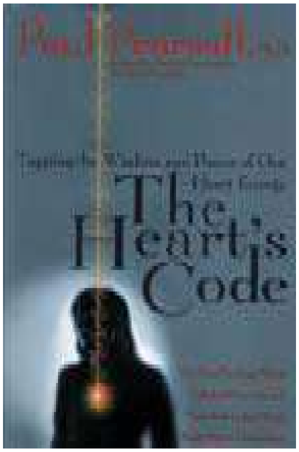

Fig. 3.20 *Heart's Code*

---

[1] "Complex and Hidden Brain in the Gut Makes Stomachaches and Butterflies," The New York Times, section C1, Tuesday, January 23, 1996

[2] Gershon, Michael. *The Second Brain* (New York: Harper Collins Publishers Inc., 1998).

[3] Pearsall, Paul, Ph.D. *The Heart's Code* (New York: Broadway Books, 1998)

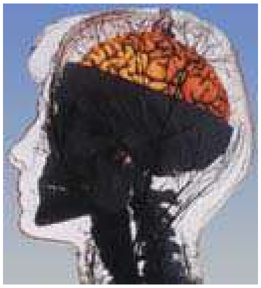

The Lower Brain consumes less energy and can do a lot of daily work, like send and receive impulse records and experience and respond to emotions. When you lower down the Upper Mind, it will also lower the blood pressure and anxiety.

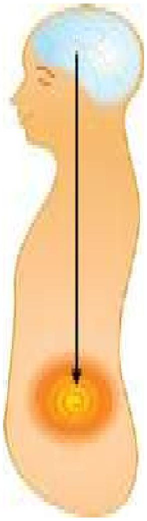

Fig. 3.21 *Abdomen has it own Brain: the Awareness Mind.*

On the second page of the Times' article, it says that even the large intestine is loaded with neurons. The question was posed, 'Can it learn?' But I say, "Hey, this goes back to 4,700 years of the Tao practice, which says: Train all the organs; train them how to do different things." You can rest the head brain when you're not using it – and use the brain in the gut. Why is it so important? Because the head brain is a 'monkey mind' riddled with doubt, shame, guilt and a suspicious mind. It is always thinking. Worrying, figuring things out things, head trips – it just keeps on, all the time. To be a God, you must let go of the past and empty the mind. Now we are in the information age; anything flits in – we think. Just one word flits in and we start an entire string of thought. When somebody says one bad word to you, you can think and figure it out for three days and three nights. 'How am I going to get revenge?
How am I going to get revenge?'

Scientists have discovered that when people worry too much, thinking, planning, etc., this brain actually uses up a lot of energy. Depending on the type of person– some people are very physical types who use very little brain energy, but their physical body consumes a lot of energy. However, most people think and think and think. So the scientists give an approximation, by way of a percentage. This is just a number for comparison, not an exact number. They say that this brain in the head can use up to 80% of the body's energy. So, what happens is you have only 20% left for the organs.

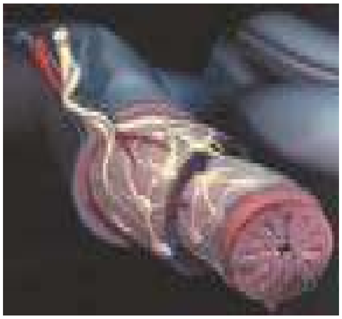

Fig. 3.22 *Cross-section of the neurons in the Large Intestine.*

According to the information about the brain in the gut, they have discovered that the brain in the head and the brain in the gut can do some similar jobs. For example, this brain in the gut is the emotional and the feeling brain. In the west you have the expressions, ‘I feel you from my gut’ and ‘I have a gut feeling about something.’ Why do people mention a gut feeling? Obviously people have some feeling in their gut. It’s very interesting that the whole Tao practice is feeling, awareness and consciousness – using this gut to feel, to be aware, and to be conscious.

You can rest, relax the brain in the head by using the ‘brain in the gut.’ This is the first step. The first thing we learn in the Tao is to forgive and let go. When we keep on remembering past negative emotions, we will stop seeing the truth. To let go of the past is to empty the mind and use the abdominal mind, the awareness and consciousness. Here’s an interesting point in the way of the Tao: This gut brain can do a lot of simple functions that are similar to the functions of the brain in the head. This is a feeling and awareness type of functioning similar to many of our ‘right brain’ functions. However, we need to use the brain in the head in order to perform complex functions such as reasoning, making plans and making complex calculations. For rational functions, we need to use the brain in the head for the ‘left brain’ functions.

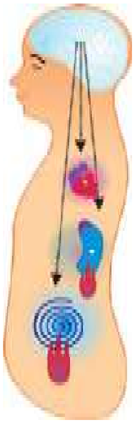

For our daily life of consciousness, awareness and feeling, we can use either the brain in the gut or the brain in the head. When we use the upper brain less, it will become charged with more energy and its power will be increased – and more power will be available to the body. That is why we say in Taoism, that we have to train the brain in the gut, so that we can use it when the brain in the head is resting. When the head brain is resting, it can be recharged: brain repair and maintenance occurs. Also, it can grow new brain cells. With more charging we have more power for creativity or whatever we want to use it for. If we like, we can use it to develop our higher spiritual nature.

Fig. 3.23 *When you are not using the Upper Brain allow it to rest by sending it down to the Lower Tan Tien.*

The simple thing is, that for the same job that the head brain or the gut brain can do, the head brain charges you eighty dollars. The gut brain only charges you twenty dollars. So, which one do you want to use? No, of course we are not that silly to use this overpriced person for the same work. But when we come to our own actual life, we don’t know how to do it. We always use this mind, this high-priced brain. To add to this, we continue using it and using it, until the brain energy is completely consumed – no more energy. When you get to a certain level the brain is empty. These brain energy measurements are not only for indicating mental energy levels, they also represent energy conditions for the whole body and spiritual energy.

Whenever I smile down, the brain waves go lower and lower very quickly – and the transformed energy from the Tan Tien and organs charges up the brain in the head! Just by flexing the facial muscles into the mode we use with a genuine smile, we can produce the effects on the nervous system that normally goes with the natural spontaneous feeling. We can actually make ourselves relaxed and happy by taking advantage of this built-in human mechanism. It's natural. Just do it! Learning to smile down to the abdominal area and maintain the awareness of the relaxed, smiling sensation in the Tan Tien is the first step in training the *Second Brain!* Think about it: Pure awareness and consciousness can change attitudes and emotions carried in the DNA.

Remember: Number 1: "Empty your mind down to the Tan Tien, and fill the Tan Tien with Chi." An axiom in the Tao is, where the mind goes the Chi follows. And, Number 2: "When your mind is empty, it will be filled!" This means that when the organs have extra energy, the extra organ energy will rise up and fill the brain with Chi.

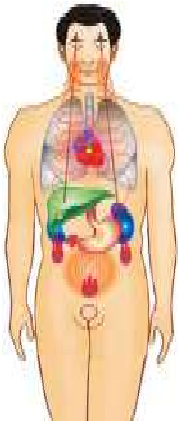

Fig. 3.24 *When the mind is empty, transformed energy from all the organs can charge the brain with Chi.*

Practice

Three Minds into One: 'Yi' Power

Smile into your heart. Lower the Observation Mind, then the Heart Mind down to the Tan Tien. With the Feeling/Awareness Mind, Spiral the three minds into one.

1. Smile into your heart. Make it feel soft. Feel love. Feel joy. Feel happy and feel compassion in your heart. Feel the heart spiral.
2. Spiral in the upper mind. Lower your upper mind, the observation mind, down to the Tan Tien in your navel area.
3. Turn the consciousness in your heart, activated by your love and softness, down to your Tan Tien.
4. Combine the upper and middle minds with the energy of the feeling/awareness mind of the entire nervous system – the Second Brain in the abdomen. Merge the three minds into one mind by spiraling their energy, blending them together as one in the Tan Tien.

When the three minds merge into one they become the 'Yi' power, the three-mind power – activate the spiraling at the mideyebrow, crown and the soles of your feet – expand to the universe in six directions: up, down, front, back, left and right. Once you get it, the Yi, we will start to activate the six directions and the three fires.

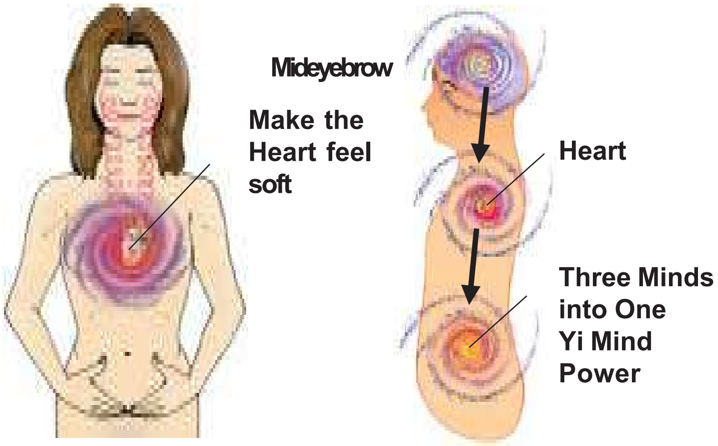
Make the Heart feel soft
Mideyebrow
Heart
Three Minds into One Yi Mind Power
A.
B.
Crown

Fig. 3.25 ***A.*** *The Heart softens like a rose blossoming.*

B. *Spiral the Upper Tan Tien into the Lowest Tan Tien and Three Mind 'Yi' Power.*

Activating the Six Directions and the Three Fires

This powerful energetic technique allows one to extend the mind, to touch the force, and to draw that energy back into the body.

The Six Directions teaches you how to expand your mind and Chi for receiving healing power. By practicing the Six Directions daily you will increase your healing and cosmic power. Turn your mind and Chi into the cosmos, multiply them and draw them back.

Direction Below

When you achieve the three minds into one mind begin expanding into the Six Directions; press your hands down and start with the low direction. Picture yourself standing on the earth and you expand yourself very far away – deep down into the earth, very, very deep down into the earth. Your hands become very long; your feet become very long — they go all the way down into the earth and out past it into the galaxy below on the other side.

Push. When you push, connect with the galaxy below — pull and think about your Tan Tien filling with Chi. Push and pull. Push and pull. Fill your Tan Tien with Chi.

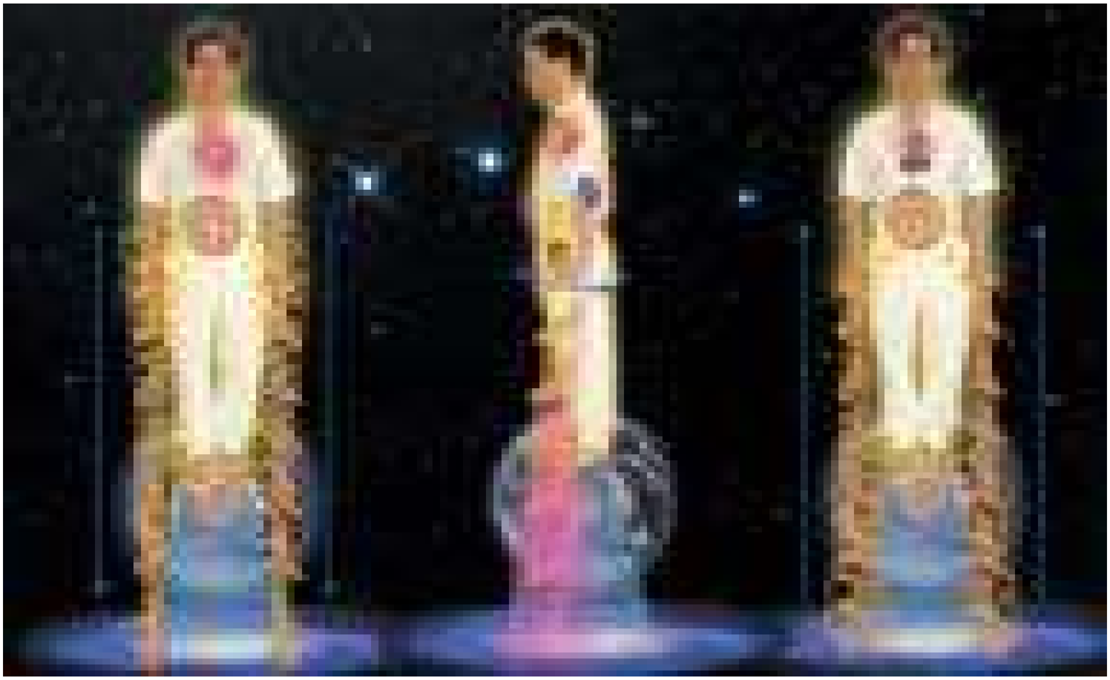

A. ***B.*** ***C.***

Fig. 3.26 A. *Hands expand through the earth and to the galaxy below.* ***B.*** *Hands push forward and pull back.* ***C.*** *Smile to the primordial Chi from the universe back to the Lower Tan Tien and fill the Tan Tien with Chi.*

1. Stand, feet together. Put you hands down, parallel to the ground. Expand your hands very far away and your mind very far away smiling into the ground. Continue expanding your hands, feet and your mind very far away beyond the earth below. Go down through the galaxy, way beyond to the primordial force. It's just like you are extending all the way to the primordial force before 30 million years ago.
2. Push, moving the hands forward six inches only.
3. Pull, moving the hands back by the sides. Think about your Tan Tien – Chi coming to the primordial force in your Tan Tien. Smile to your Tan Tien, dark, deep and vast.
4. Push: touch the primordial force in the universe.

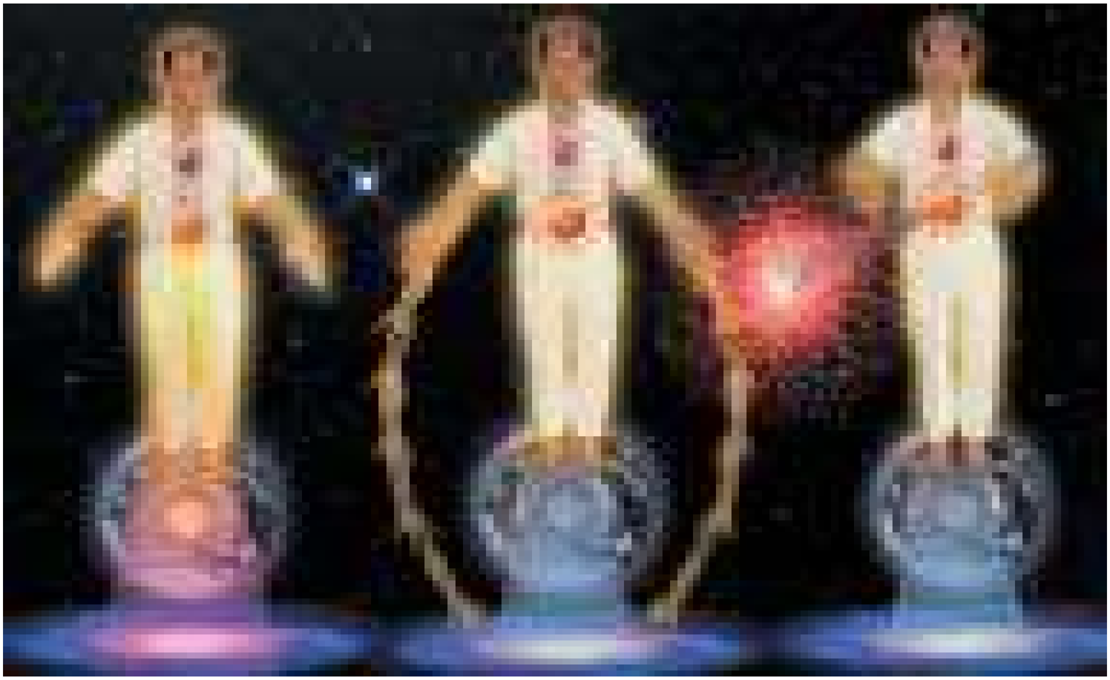

A. ***B.*** ***C.***

Fig. 3.27 A. *Be aware of the universe in front.*
B. *Expand your hands very big and long to the universe in front.*
C. *Hold the fire ball to activate the Tan Tien Fire.*

5. Pull back to your Tan Tien the dark primordial forces with your hands.

Push and pull: It's just like you go to an empty space – vast. Then you come back to your Tan Tien – also empty, just like the primordial condition before anything existed. That is where all the forces come from. Push and pull 3 – 9 times.

Front Direction and Tan Tien Fire

Next, be aware of the front direction; a huge fireball appears in front of you.

Open your palms: scoop up the Chi, scoop up the fire. Bring the fire into your Tan Tien. Activate the Tan Tien Fire.

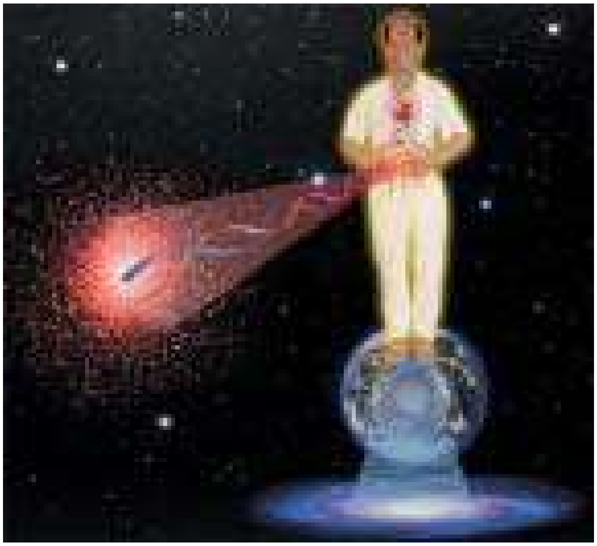

1. You start with a small dot of light inside you. Expand your awareness smiling to the universe in front of you.
2. Become aware of a big fireball in front of you. Fell your hand become bigger and longer. Scoop up the fireball. You may close your eyes to help your inner sensing.
3. Use the fireball to light the fire in your Tan Tien. Feel the fire burning in the darkness, the 'fire burning under the sea.'

Fig. 3.28 *Smile to the Burning Fire.*

Back Direction and Kidney Fire

Be aware of the back of the Tan Tien, the Door of Life and the back or rear direction. Extend your mind very far away to the 'back'. Scoop up the fire and light up your 'Kidney Fire'.

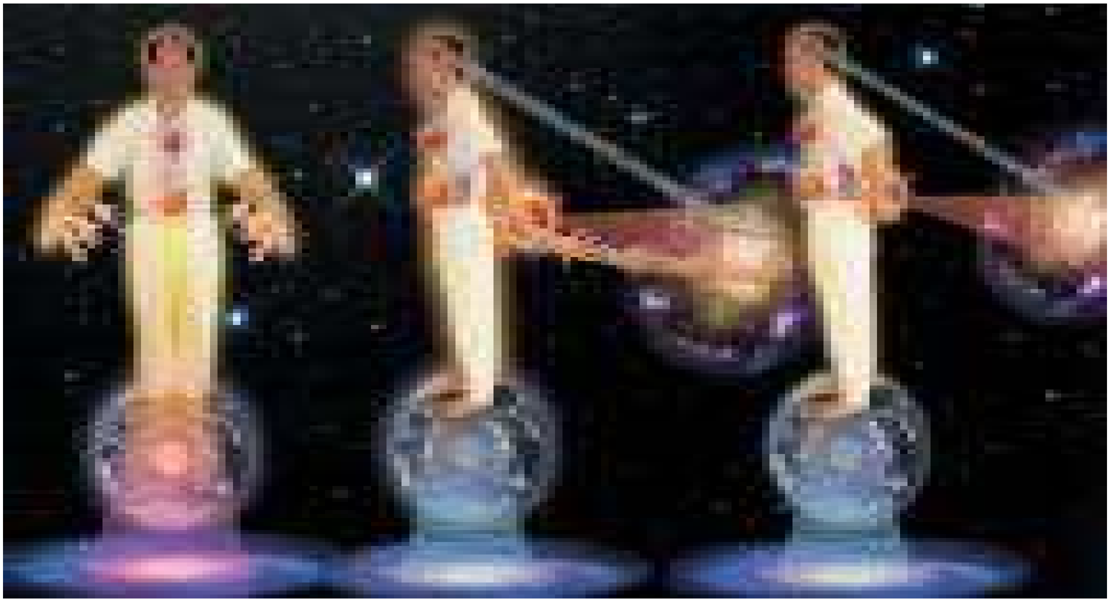

Fig. 3.29 *Be aware of the back direction, move the arms toward the back of the universe and scoop up the universal fire.*

1. Expand your awareness all the way to the back. Move the arms toward the universe behind you.
2. Touch the universe; scoop up the fire. Activate the Kidney Fire. Maintain your awareness in the Tan Tien and expand smiling out to the universe. The energetic spiral glows in the Tan Tien. Spiral in the heart, spiral in your crown and spiral in the universe.

Heart Fire

Raise your hands up under your armpits, and feel yourself holding the two fireballs. Touch the heart by extending the fingers energetically in from the sides; feel your hands extending into your heart and very far away. Activate the Heart Fire.

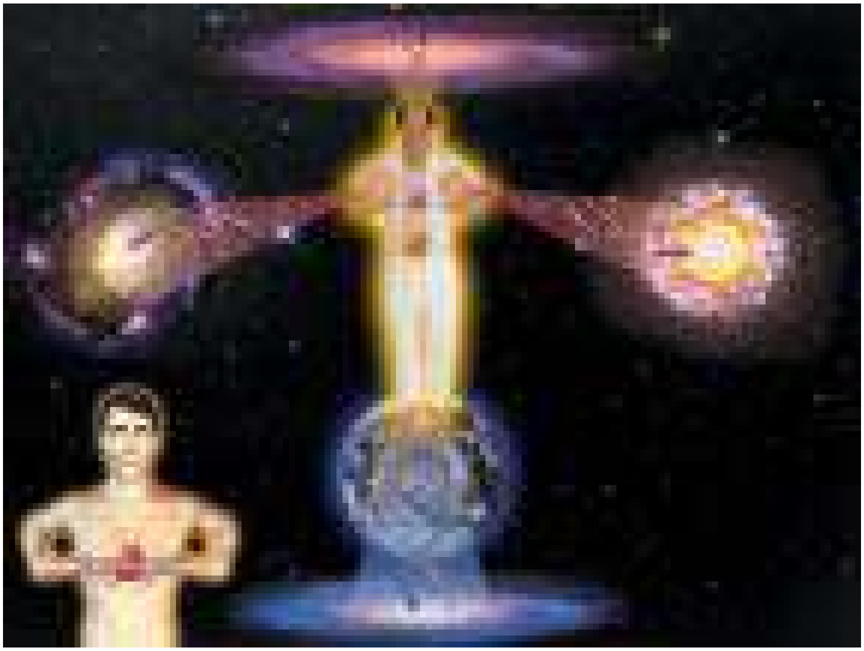

Fig. 3.30 *Activate the Heart Fire.*

1. Move your hands up under your armpits and extend your fingers deep into your heart and very far away.
2. Tan Tien and the Universe: you are connecting to the ‘charger,’ charging more fire into yourself.
3. Feel your heart soft in the center as you smile down. Feel the warmth of the fire energy of love, joy and happiness in the heart.
4. You feel the connection with the unconditional love in the universe as you keep your heart consciousness in your Tan Tien and extend your awareness out to the universe.

Sacred Fire *(Chi Fire)*

Connect the Three Fires to combine into one Fire: *Heart to Kidneys to Navel to Heart*.

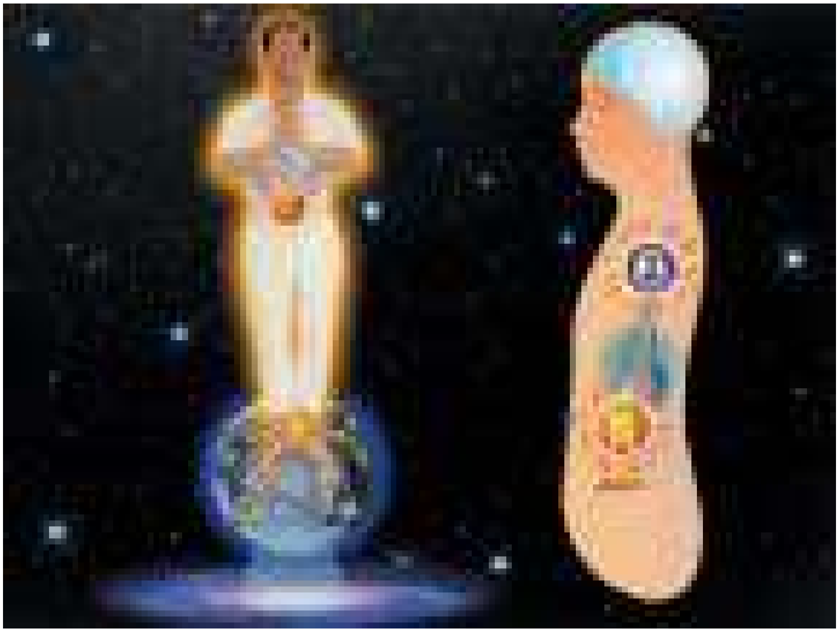

Fig. 3.31 *Connect the Three Fires and Sacred Fire.*

1. Move your hands together in front of your heart. Hold your hands together in front near the heart and feel the fire burning in there. Connect the Heart Fire to the Kidney Fire, the Kidney Fire to the Tan Tien Fire and back up to your heart; connecting them as one triangular Sacred Fire, circulating the Chi at 1,000 revolutions per minute, 10,000 R.P.M, 30,000 R.P.M and 60,000 R.P.M, doubling or tripling their collective power.

Open the Third Eye

Now, extend your hands out to the front, very far away – pushing, pushing, pushing. Push. Turn your palms inward, and extend your middle fingers inward toward your third eye. Picture a crack in the middle of your forehead, and pull the crack open. Feel the light from the heavens opening it and feel the light from the heavens shining into your brain.

1. Open your palms. Open your eyes, dimming the eyes. Look to the universe. Extend your hands to the front, palms vertical. Extend the arms from the scapulas. Smile and touch the universe.
2. Turn your palms inward and extend your middle fingers inward toward your third eye.

3. Picture a crack in the middle of your forehead and let the heavenly light shine into the brain; pull the crack open and let the light reflect into the organs.
4. Close the third eye. Again, open. Pull: open-close-open. And, close. With the third eye open, the light from the heavens shines into your brain and reflects down to all your organs. Open and close the third eye 3 - 9 times.

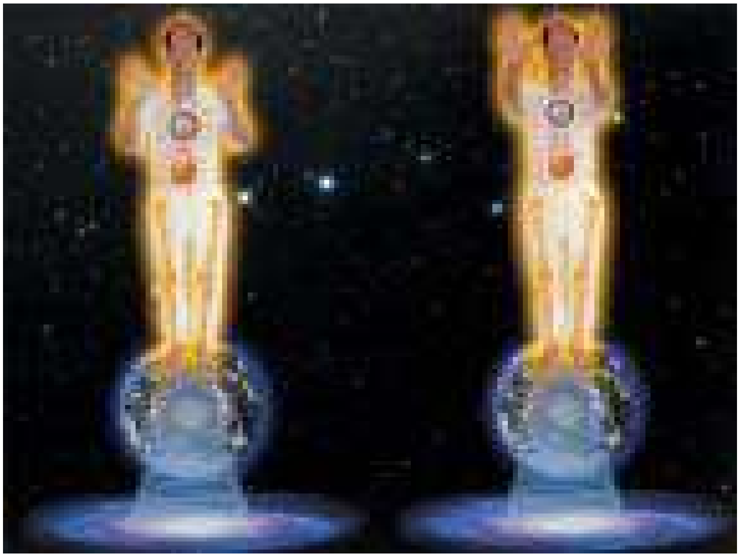

Fig. 3.32 *Look as you smile into the Universe in front of you.*

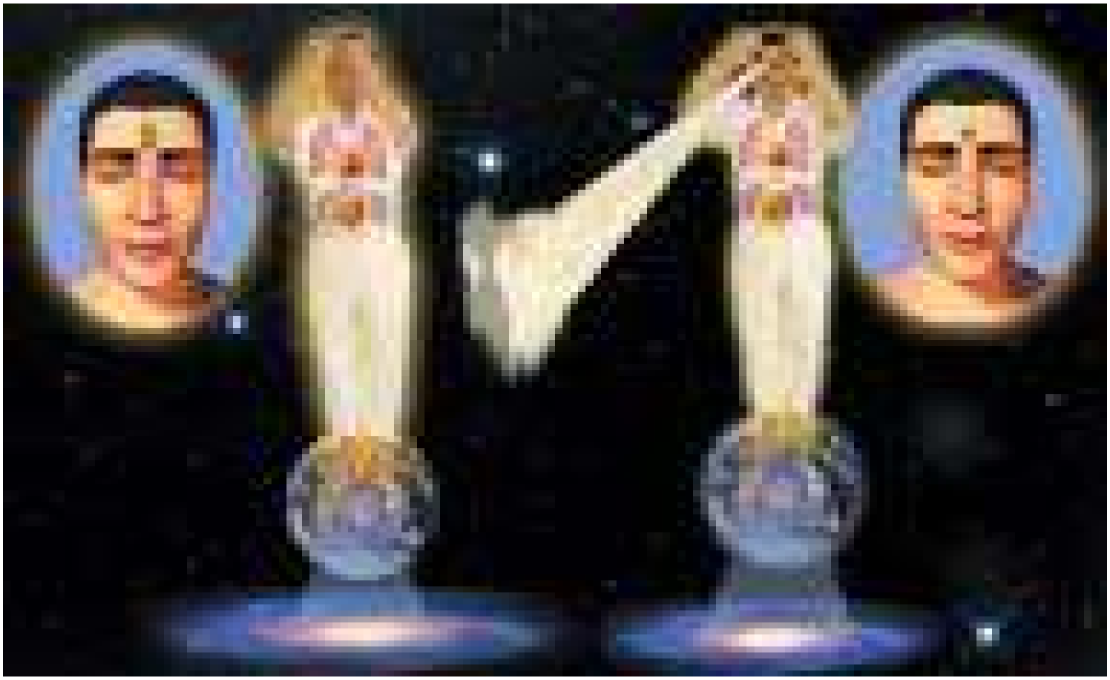

A. **B.**

Fig. 3.33 A. *The Middle of the Finger hooks into the third eye and reflects down into the Organs.* **B.** *Pull open the Third Eye; let heaven open and shine its light into the Brain.*

Front Direction: Push/Pull Master Practice

Now, turn your palms, pushing out. Push. Pull. This is the master practice that is imperative for successful completion. When you first start practicing you should do it at least one hundred times and increase up to 200 times. Push and pull. When you push, you feel your hands extended very far away – very long – reach the sky. Touch the universe.

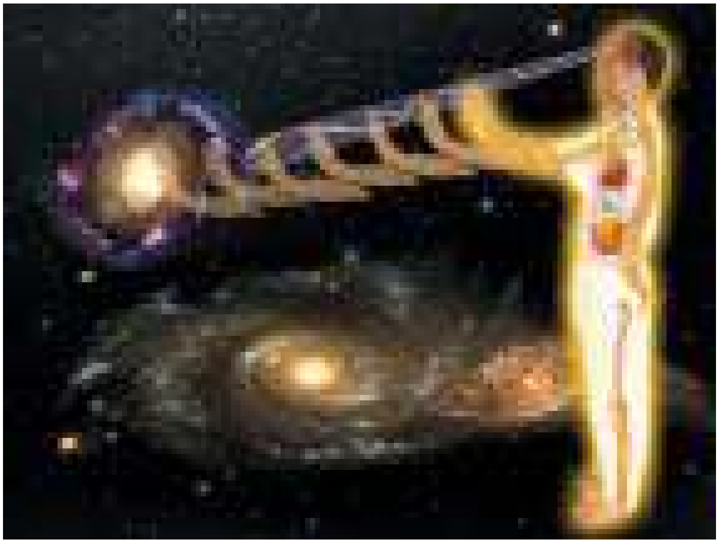

1. Push: Extend your arms and hands to the front, palms vertical. Extend the arms from the scapulas.
   Expand: smiling, smiling, touching the universe – touching the force, touching the Cosmic Chi.

Fig. 3.34
Master practice:
Touch the Universe.

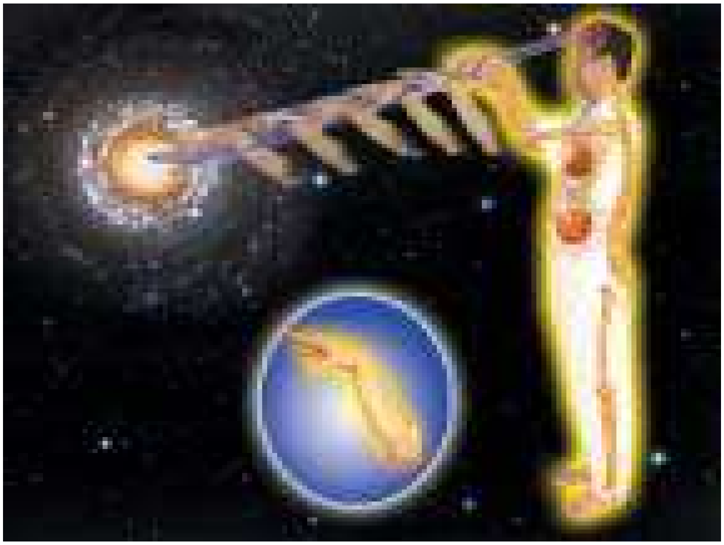

2. Pull: Draw the Chi back to you from the universe. Moving the arms from the scapulas, draw the hands toward your body in a horizontal position, drawing the Chi into the body.

Fig. 3.35 *'Drawing' Universal Chi - feel your Tan Tien and fill it with Chi.*

Pull. Think and smile to your Tan Tien. Push, very far away to the universe. Pull. Push: Smile, relax, and let go, touching the sky, touching the universe.

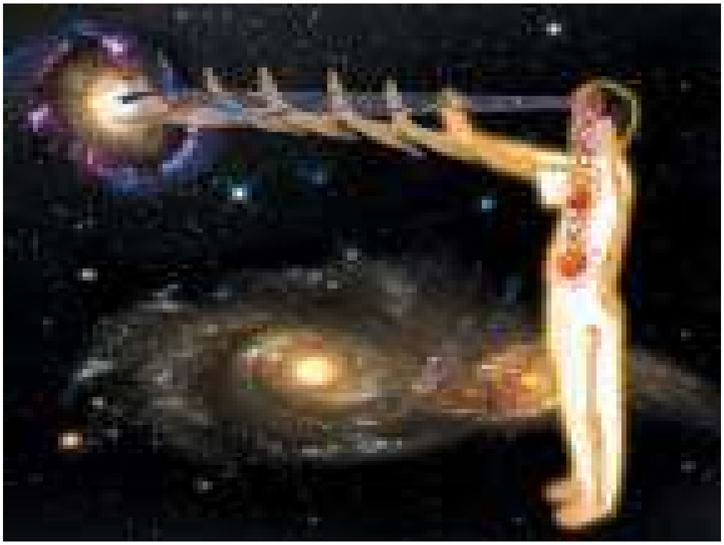

Fig. 3.36
Let go - Push and Touch the Universe, do it 6, 9 or 18 times.

Left and Right Directions

Now, move your hands to the left and right directions. Pull the Universal energy in. Push; touch the universe. Pull; think about your Tan Tien. Push – all the way, touching-touching-touching the universe.

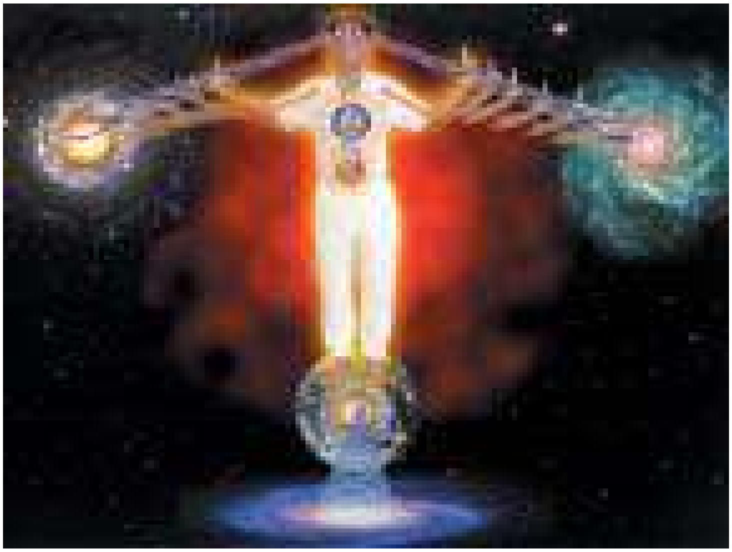

1. Move your extended hands from the front horizon to the left and right sides.

Fig. 3.37 *Touch the Universe Left and Right.*

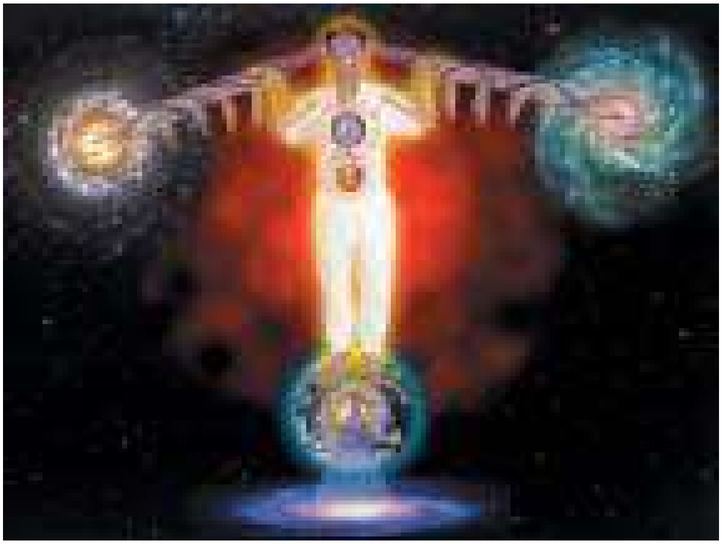

2. Pull Tan Tien: Smile to your Tan Tien. Keep smiling to your Tan Tien.
3. Push to both sides. Expand all the way, smiling and touching the universe.

   Push/Pull: touching, touching the universe drawing the Chi smile energy into you from both sides.

   Do 3-6 times.

Fig. 3.38 *Pull: Just feel about your Tan Tien.*

Direction Above

Turn your palms up to the universe. Scoop up the Chi. Pour the Chi over your crown, and touch your crown. Project the Chi all the way down to the perineum and down through the earth to the universe below. Tan Tien and universe; always feel your Tan Tien spiraling, heart spiraling, crown spiraling and the universe around you spiraling.

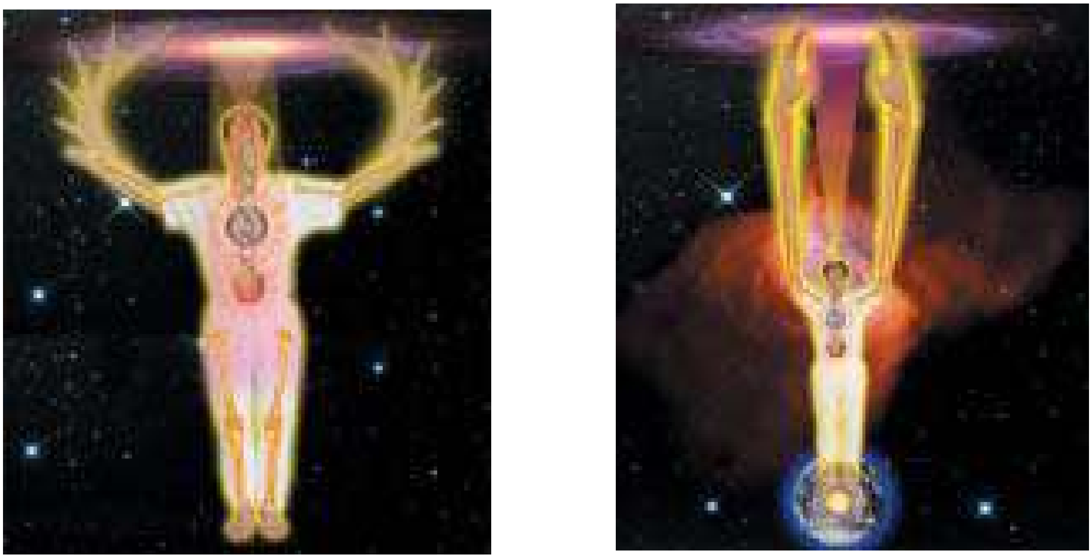

Fig. 3.39 A. *Raise the hands above the crown and feel it extend up to heaven.*
B. *Feel that the hands are big and long and that the bones are hollow. Fill and pack the bones with Chi.*

Raise your hands up to the universe and expand your hands as big as the universe - feel the Chi charge into your bones.

Scoop and gather the energy turning the body three times to the left and three times the right drawing the smiling energy.

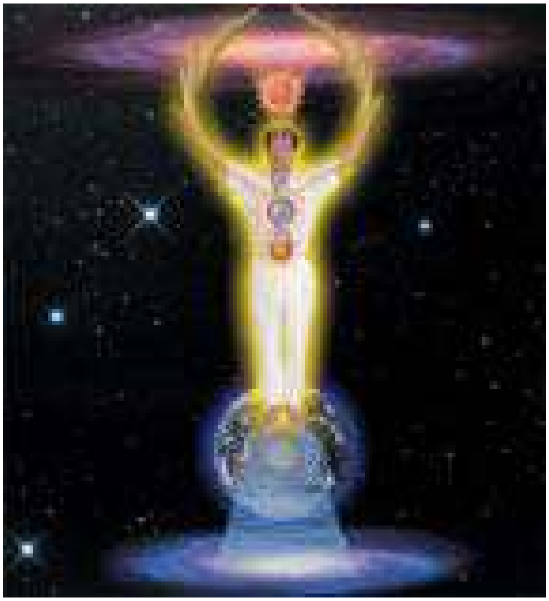

Fig.3.40 *Scoop up the Universal Chi and pour it over the head.*

Open the Spine

1. Touch the back-crown point. Pour the Chi all over your crown. Think of your soles so that you feel like there is a waterfall of Chi flowing from your crown all the way down to your soles. Feel your finger grow long and the Chi penetrate down through your spine to the coccyx. Leave the fingers touchng the back of the crown, to maintain the energetic connection with the coccyx.

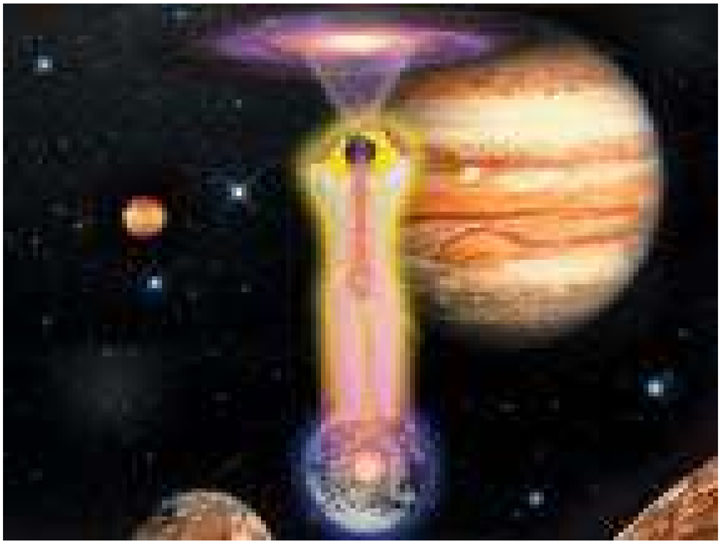

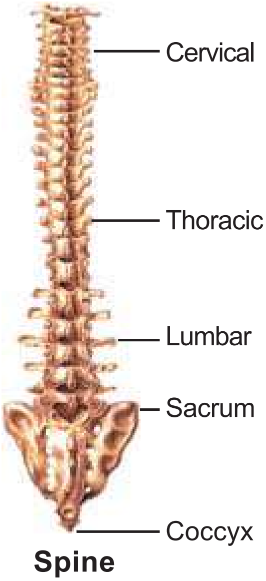

Fig. 3.41 *Feel your fingers growing long, all the way down to the coccyx.*

Turn ‘Three Minds into One Mind’ at the Lower Tan Tien, and expand the awareness to the universe. Let yourself be charged by the universe.

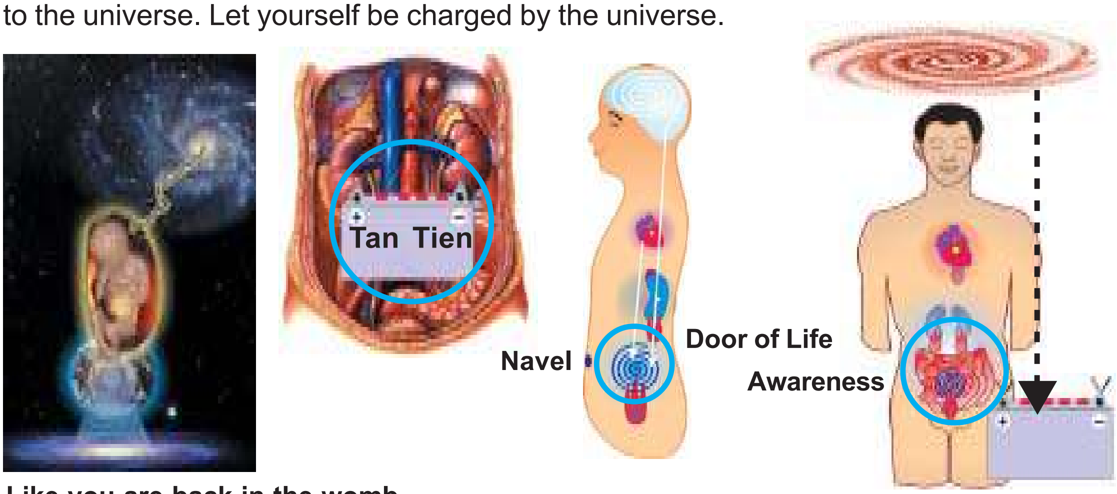

Like you are back in the womb.

Tan Tien is like a battery.

Fig. 3.42 *Universe charges Tan Tien power.*

Beware of the Tan Tien and spiral it like universal energy in motion. Feel the heart center spiraling and the crown spiraling. Be aware of the universe spiraling above, below, front, back, left and right. Let all of the sick energy and the negative forces leave the body and go down into the ground for Mother Earth to recycle. Extend the Chi from above, all the way down through the earth and the universe below.

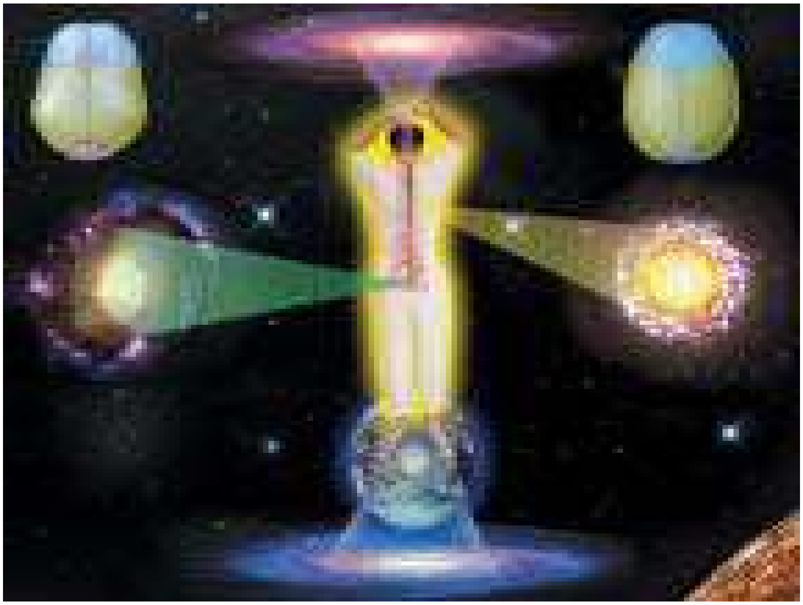

Fig. 3.43 *Be aware of the Lower Tan Tien, Heart Center and Crown spiraling.*
Feel the Universe is spiraling and charging the Three Tan Tiens.

Open the Middle Channel and Perineum

1. Move to the mid-crown point. Touch the point and project your fingers inwards; go deeper, deeper through the middle of your body down to the perineum. Focus on the perineum. Feel the Chi from the universe flow into your perineum. Look for one dot of light. Look into the darkness, the vast darkness, the immense darkness; this is the primordial force, a cloudy moving force. Look for a dot of light at the perineum and extend your awareness all the way down through the ground and the universe below.

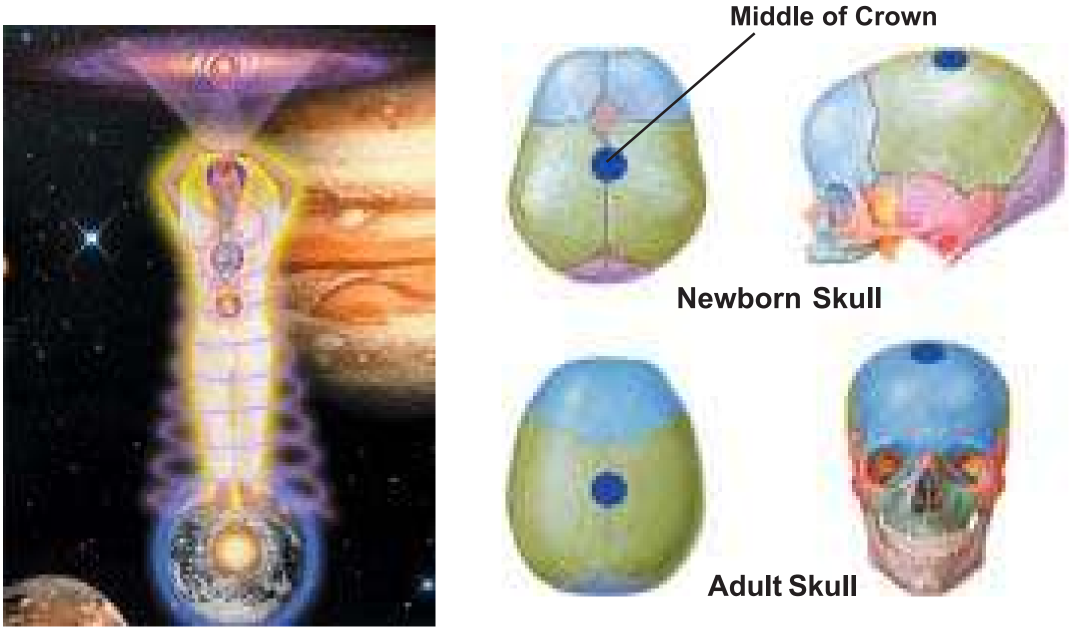

Fig. 3.44 *Move the Hands to the Middle of the Crown and feel the Fingers Penetrate to the Perineum.*

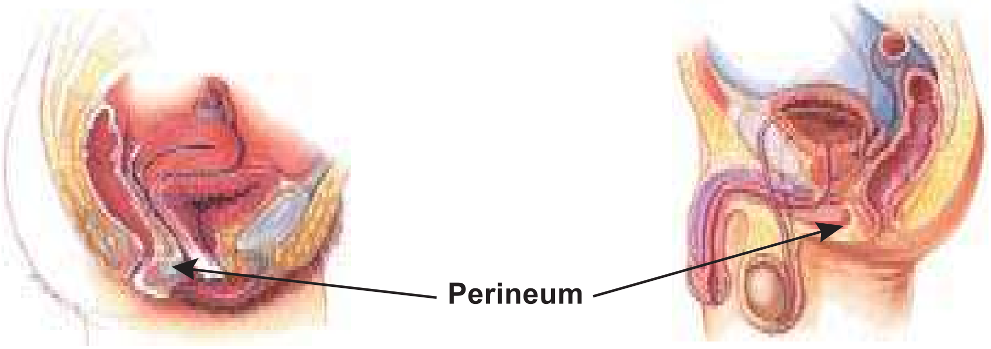

When finished opening to the Six Directions and igniting the Three Fires, gather the Chi in the center and bring this expanded awareness into the healing session.

Tan Tien and the Universe

The expression, 'Tan Tien and the Universe', is a reminder to feel your Tan Tien, heart, mideyebrow and crown spiraling and that you are connected to the universe spiraling in the six directions around you.

1. Establish a Complete Location

You will use your Yi, the three-mind power, repeatedly to recharge your Chi for various purposes. You recharge by connecting to the Universal Chi in the six directions of the universe simultaneously. When you charge a particular area or direct Chi into the body to a particular point, you first establish a connection point for the Chi by placing the hand or fingers at an appropriate location on the surface. This is like giving a location for the Chi to go to. Once the location is established and the Chi starts to go there, leave your hands there.

Then you move your attention to where you want the Chi to go in the body. Feel the Chi connected to and charging the intended location.

2. Charge Tan Tien from Universe

When the location connection is established, be aware of your Tan Tien, heart center, mideyebrow and crown spiraling. Be sure that the conscious mind of the heart is lowered to the Tan Tien and the awareness mind of the abdomen is connected to the mideyebrow crown and out to the universe. Feel them connected to the spiraling energy in the six directions of the universe. Let the universe charge your Tan Tien.

3. Don't Stop at the Location

With your focus in the Tan Tien, the Chi will go from there to the 'location' indicated by your hands and then to the intended location in the body. Don't let Chi stop there, however.

4. Direct Chi out the Opposite Side to Universe

Direct your attention to guide the Chi flow through the body and out the opposite side. You want to clear out any blockages and prevent any others from accumulating. You also want to release sick energy and negative forces down into the earth. Let the Universal Chi flow out through the universe and beyond.

Open the Three Tan Tiens

Move your hands down to the mideyebrow. Touch the mideyebrow.
We are going to open the three Tan Tiens, starting with the Upper Tan Tien.

Upper Tan Tien - Mideyebrow

1. Recharge. Remember: Tan Tien – heart consciousness in, awareness out. Spiral – Tan Tien, heart, third eye and crown. Universe – six directions spiraling.

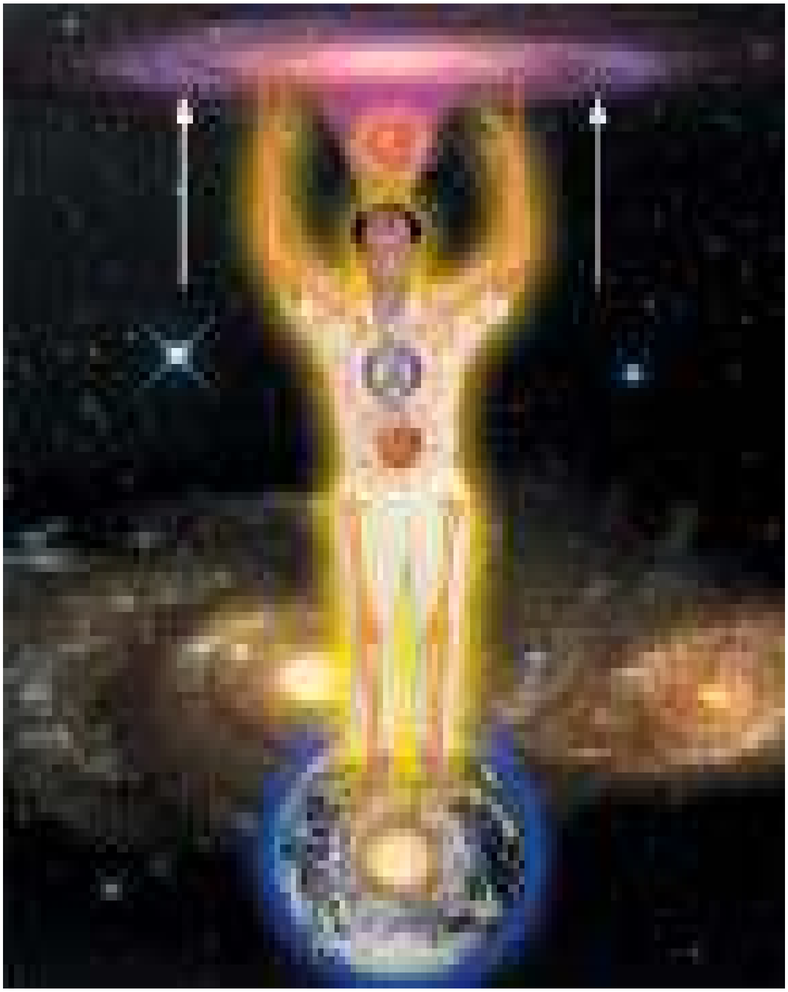

Fig. 3.45 *Raise hands and charge with Universal Chi.*

2. Move your hands down and touch the mideyebrow. Feel your fingers grow very long (energetically) and penetrate all the way back to the base of the skull. Focus on the back. Remember: Tan Tien and the universe spiraling. With the spiraling, the Chi in the fingers will become hot. It will expand and penetrate out through the back of the head all the way to the universe behind.

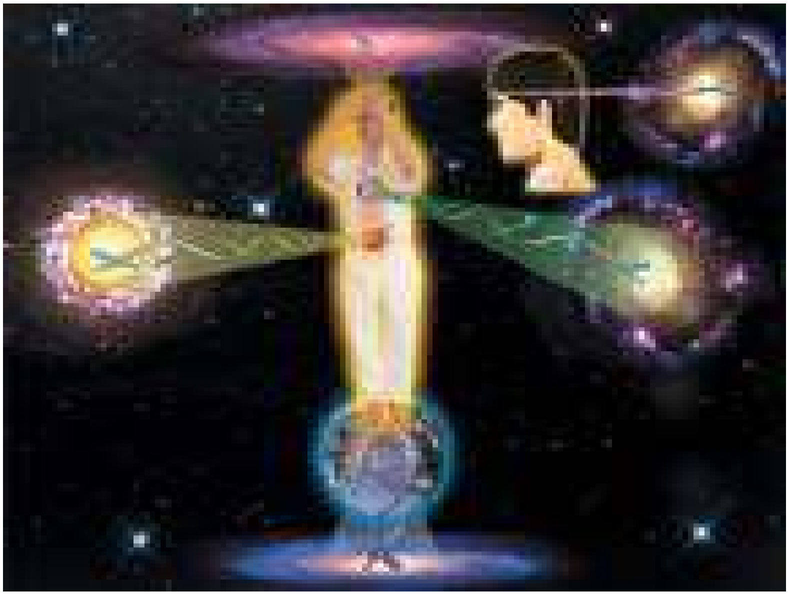

Fig 3.46 *Fingers touch mideyebrow and penetrate to the back of the skull and the universe at the rear.*

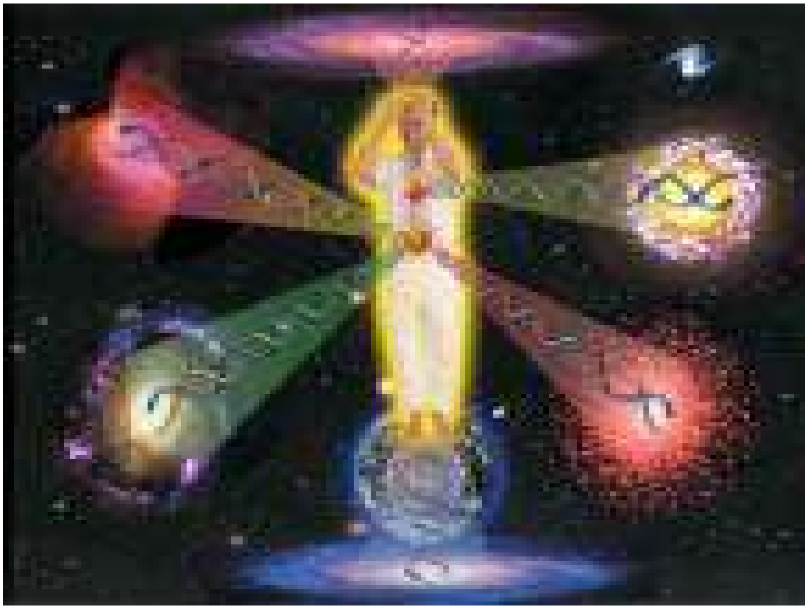

Fig 3.47
Tan Tien and the Universe.

3. Picture your fingers like laser beams of Chi, ‘Tan Tien and Universe’: Feel your Tan Tien and the universe spiral and charge your fingers. Move your fingers out from the mideyebrow around the side of the head to the top of the ears. Your fingers are like lasers -cut-cut-cutting open your skull right in the middle, around to the top of the ear, cutting open your Upper Tan Tien. Cut and project your fingers long into the middle of your brain. Spiral your Tan Tien and the universe. Leave your fingers there. Concentrate on your Tan Tien spiraling – your heart, crown, the universe above, below, front, back left and right all spiraling. Your Tan Tien is a big empty space: primordial force; darkness. You can put so much Chi inside there! The Chi penetrates into your brain.

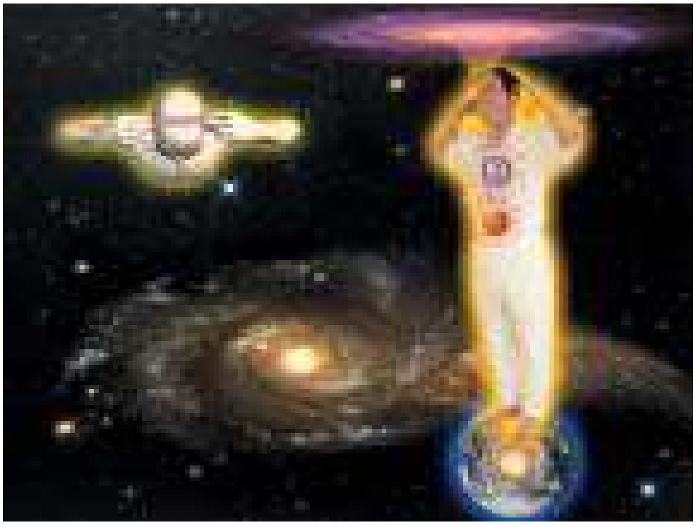

***Fig. 3.48** Move the fingers to the top of the ears. Feel them growing longer and cutting the skull open, penetrating into the brain.*

4. Move your hands all the way to the back, cutting to the back of the skull. Touch, and feel the Upper Tan Tien open.
5. Touch the base of your skull. Focus on the mid-eyebrow. Feel the Chi flow like a laser beam from back to front and out to the universe in front. Complete the opening process by moving the hands back around to the mideyebrow, cutting as you go.

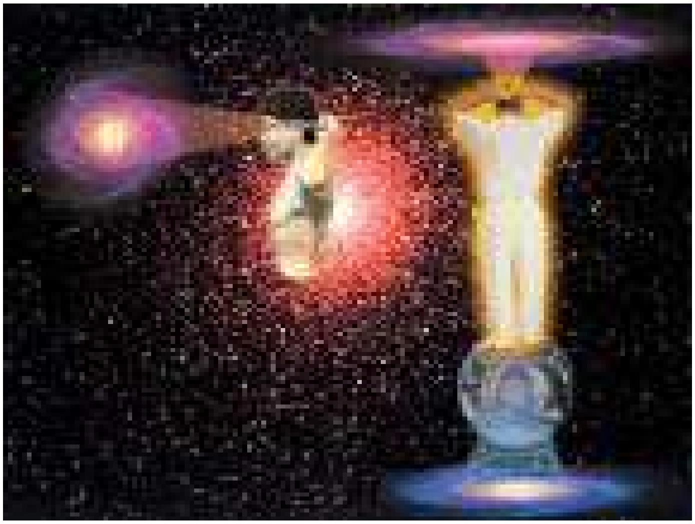

***Fig. 3.49** Chi moves like a laser, penetrating from the mideyebrow at the front, to the back of skull.*

6. Recharge in the universe. Feel your bones and your arms are hollow. Fill and compact them with Chi.
7. Scoop the Universal Chi and pour it down over your crown and all the way down, down to the Middle Tan Tien.

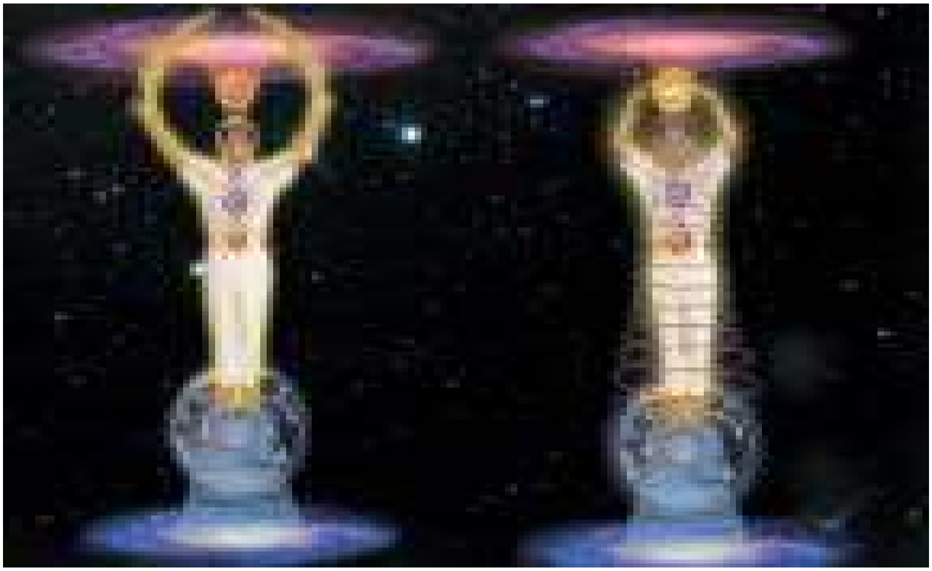

Fig. 3.50
Recharge in the Universe, scoop and pour the Chi down to the Middle Tan Tien (Heart Center).

Middle Tan Tien - Heart Center - Conscious Mind

1. Move your hands all the way down to your heart center at the mid-sternum. Touch. Focus on the point opposite the heart, T5/T6, on the spine. Fingers 'long', Chi penetrates into your thymus gland. Light – golden light – penetrates into your thymus. Feel the Chi penetrate through your heart all the way through T5/T6 to the universe behind. 'Tan Tien and the Universe' also feel your Chi Fingers penetrate into the bone and bone marrow and spread out into your rib cage.

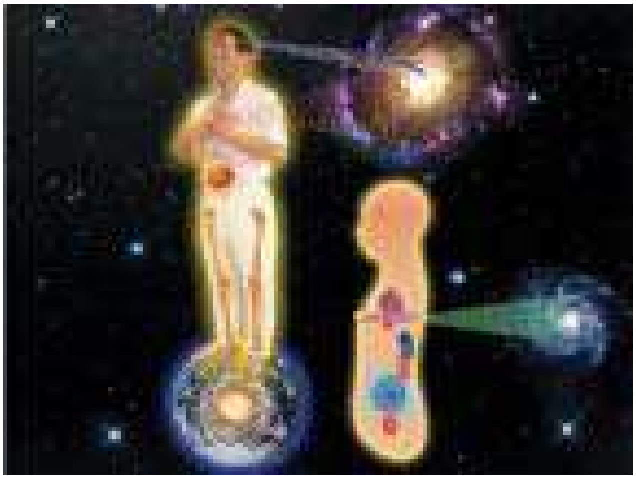

Fig 3.51 *Golden light enters the heart, thymus, bones and marrow; feel it penetrate to the universe at the rear.*

Recharge and cut around to the Armpits

1. Recharge In the universe, pour Chi over the crown down through the body and lower your hands down to the heart center. Touch with the fingertips. Move your hands around under your armpits, extending the Chi like laser beams cutting open this Middle Tan Tien. Pause under the armpits as you send the Chi into the center.

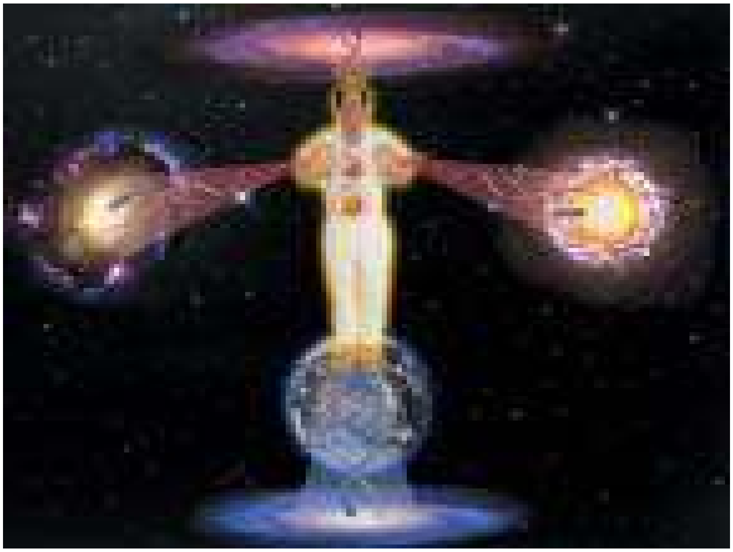

Fig. 3.52 *Cut open the Middle Tan Tien by cutting around the armpits.*

2. Continue to move your hands around to your back at T5/T6, touch and send the Chi from back to front. Let the beam of Chi penetrate out through the heart center to the universe in front of you. Then, move the hands back around the sides to the front, cutting as you go.

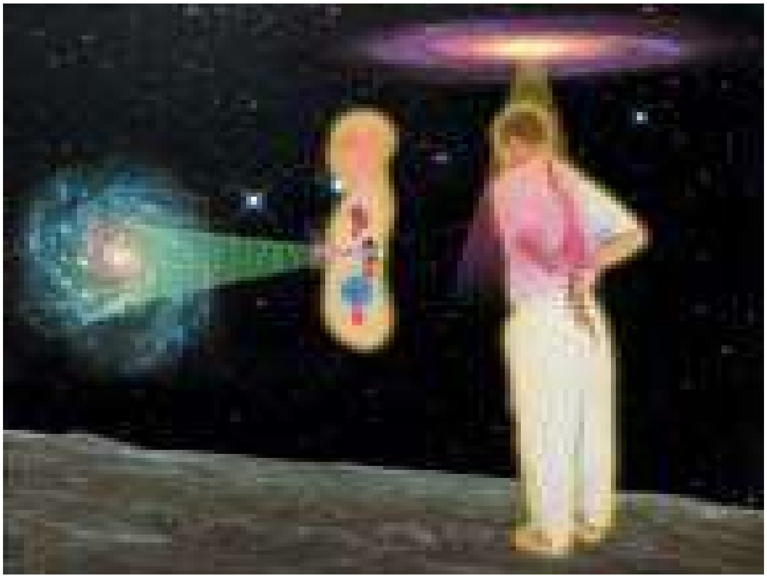

Fig. 3.53 *Chi beam from T5/T6 penetrates to the Heart Center in front.*

Lower Tan Tien - Navel - Awareness Mind

1. Recharge with Universal Chi. Tan Tien and Universe: Again, raise your hands and charge with the Chi into the universe. Your hands are very big, very long. The bones are hollow and compacted with compressed Universal Chi. Scoop the Chi from above and guide it down. Pour all the way down, down, down to the navel. Touch the navel, and focus on the Door of Life opposite on the spine between L2/L3. Touch and feel the Chi penetrate to the Door of Life Tan Tien and Universe. Feel the Chi penetrate through to the back and out to the universe behind.

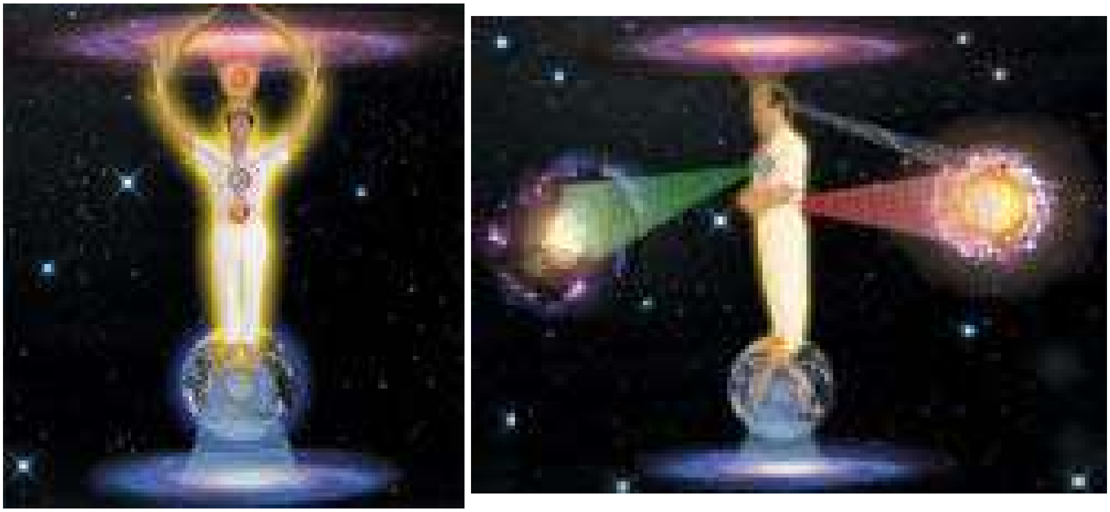

A. ***B.***

Fig. 3.54 A. *Recharge the Lower Tan Tien and penetrate through to the Door of Life.*
B. *Raise the hands up to the universe and charge with Chi. Bring the power down to Tan Tien.*

2. Open this Lower Tan Tien the same way as for the Upper and Middle Tan Tiens. Charge more Chi into your hands, and let them be like lasers cutting it open. Cut around to the sides. Pause. The fingers of the left and right hands are very long, extended energetically inside. Cut and feel the energy penetrate into the center. Focus at Tan Tien and universe and feel more Chi.

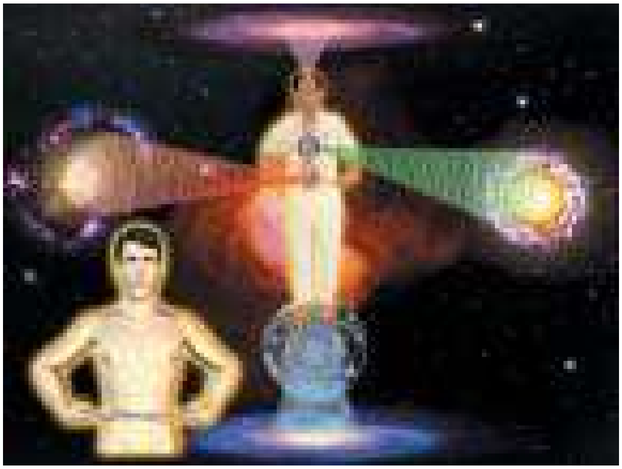

Fig. 3.55 *'Cut' open the Lower Tan Tien with Laser Fingers.*

3. Continue cutting to the Door of Life. Touch, and send the Chi from the Door of Life back to the navel and out to the universe in front.

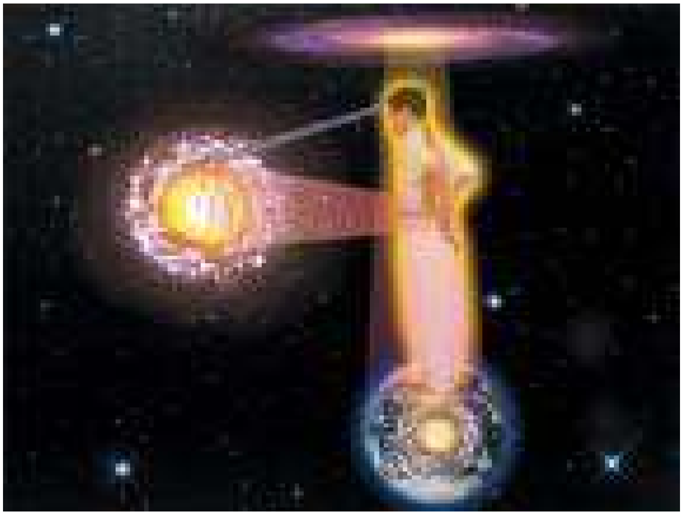

Fig. 3.56 *'Cut' open the Door of Life and send Chi to the navel and universe in front.*

4. Move the hands back around the sides to the navel, extending the fingers and "cutting" the Tan Tien open as you go. Touch the navel: Tan Tien and the universe spiraling. Feel more Chi, and feel the Tan Tien open.

Activate Chi in the Bones of Hips, Legs and Sacrum

1. Now touch your pelvic bone by energetically extending your fingers from the front area near the hips to the back. Feel Chi penetrate into your pelvic bones: funny, laughing, happy bones.
2. Touch the femur bones. Tan Tien and universe. Charge the fingers. Also, feel the funny, happy, laughing vibration inside the bones and in the bone marrow. Be aware of the Three Tan Tiens. Spiral from the universe. Spiraling charging your Tan Tien. Charge your hands and your bones.

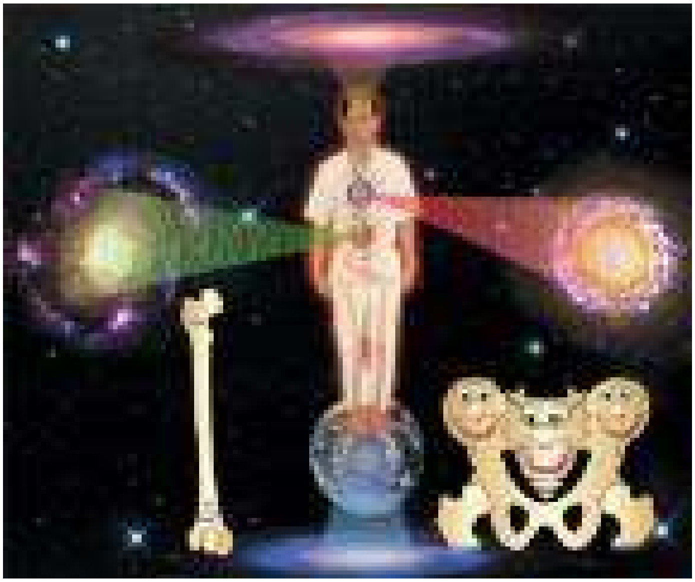

Fig. 3.57 *Touch the femurs - happy laughing bones!*

3. Now we are going to slowly lower the sensation of Chi down through the bones to the earth. Opening up the sacrum. Move the hands down the legs as you bend down.
4. Lower yourself all the way down to the ground and sit on your feet. Move the Chi with your hands down to your toes, down through the earth and the universe below.

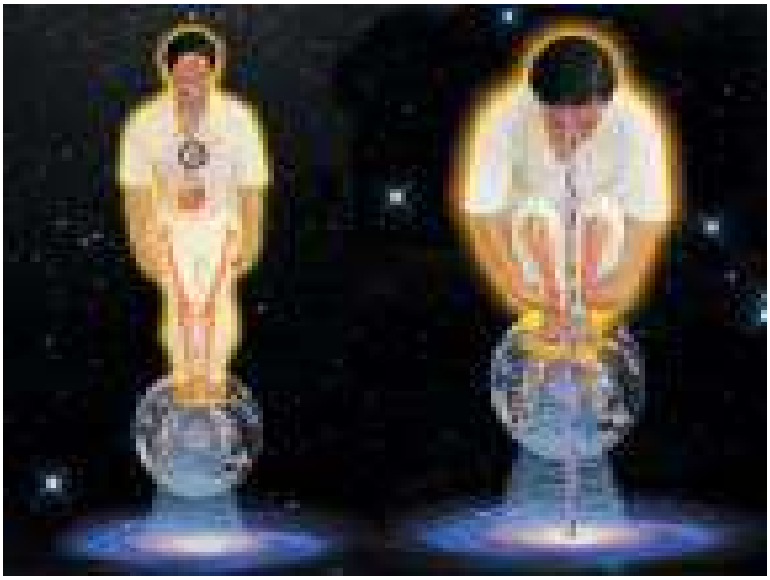

A. ***B.***

Fig. 3.58 A. *Sink the Chi into the earth.* ***B.*** *Lower and sink the mind and Chi down to the universe.*

5. Raise your sacrum up, keeping your hands at the toes. Smile to your Tan Tien and feel the Chi from the universe coming to fill the Three Tan Tiens. Opening up the sacrum.

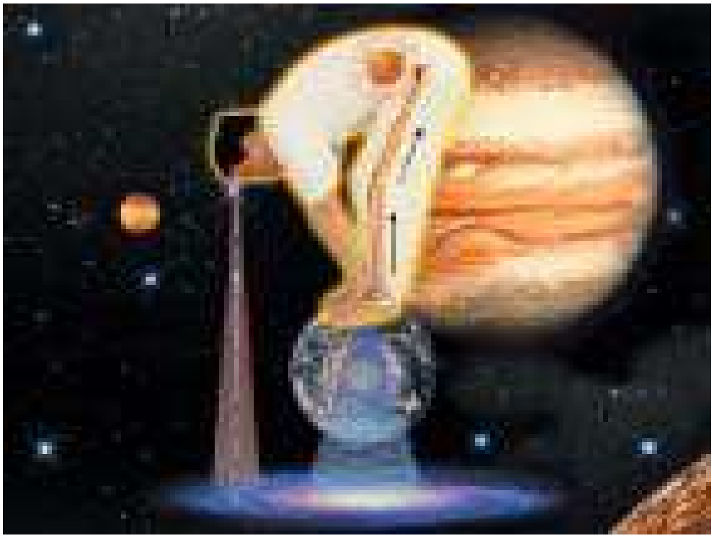

Fig. 3.59
Raise the sacrum and smile to the Tan Tien.

6. Lower down again. Lower the Chi down to the earth and the galaxy below.
7. Once more, raise your sacrum up, maintaining hand contact with your feet. Smile to your Tan Tien.
8. For the third time, lower down. Open your palms, gathering the Chi from the earth below. Gather and scoop the Chi.

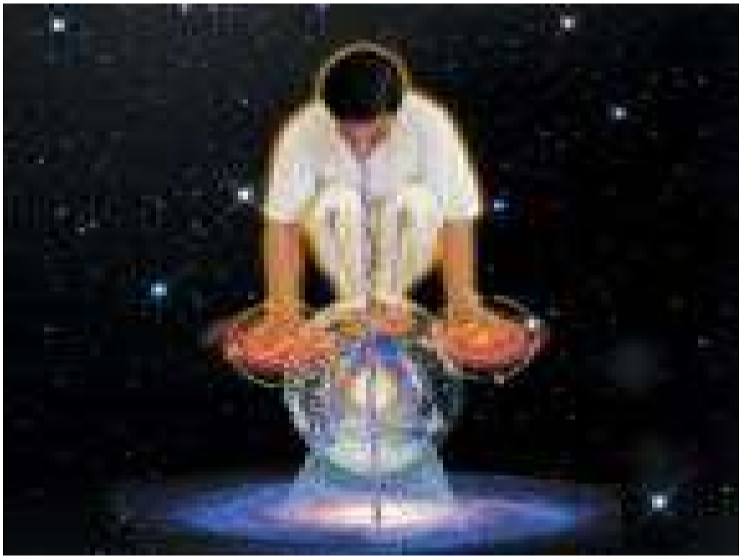

Fig. 3.60 *Gather the Earth Chi.*

9. Touch your heels and feel your bones as you slowly rise up. Fill your bones with Chi as you guide it up with your hands.
10. Fill the bones in the upper legs as you move your hands up.

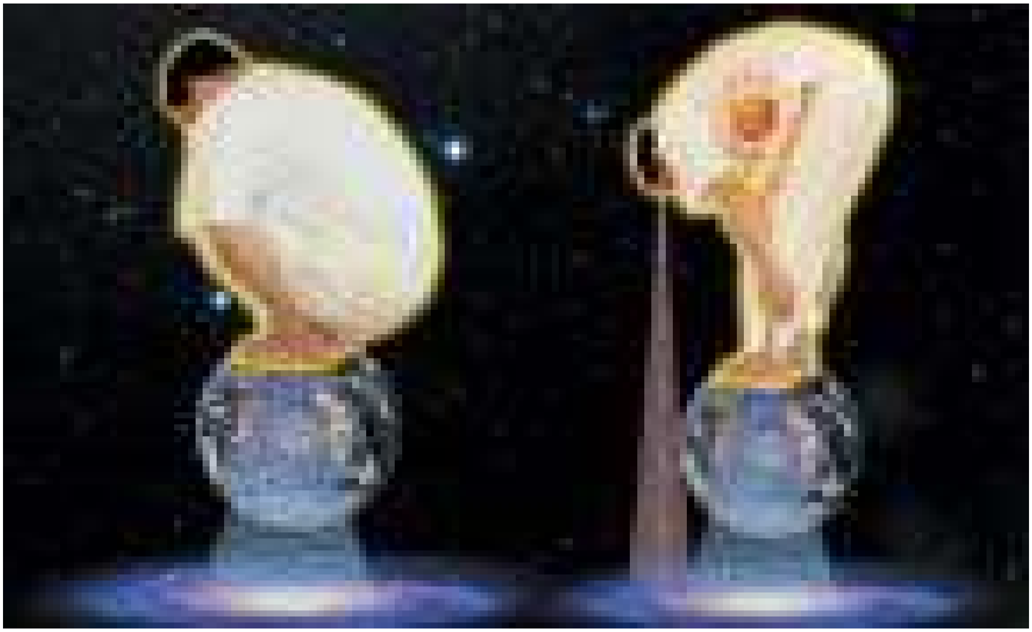

Fig. 3.61 A. *Fill the bones with Chi.* ***B.*** *Feel your fingers penetrate into the bones as you rise.*

11. Feel your bones, and fill them with Chi all the way up to your coccyx. Touch your coccyx. Leave your fingers there, then be aware of the Chi and feel the Chi raising up to the Tan Tien and the universe. Feel it charge the fingers and the spine.
12. Come up to the sacrum. Feel the sacrum open. Focus at the Tan Tien and the universe.

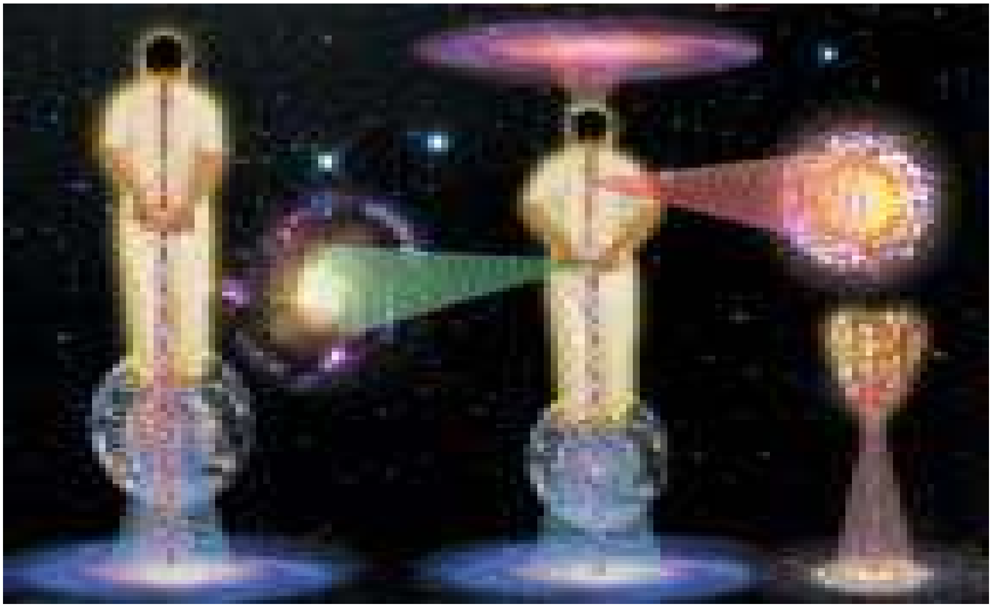

A. **B.**

Fig. 3.62 A. *Bring the Chi to the Coccyx.*
B. *Move your hands up to the Sacrum and feel the Chi rise up to the Crown.*

13. Come up to the door of life, and then back to the navel. You may sit down to continue the next step in the Cosmic Healing Practice.

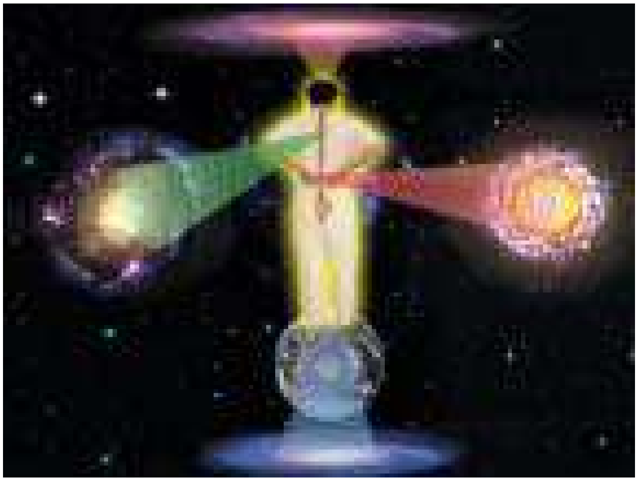

Fig. 3.63 *Bring the Chi to the Door of Life and let it penetrate to the Navel.*

Practice daily till you feel the Chi. Then you can apply this Chi to the Healing Sessions.

Chapter IV
General Healing Session

The General Healing session practices are to be done with every student regardless of their age or health condition. These are for cleaning and strengthening the cells. Cells must be clean and strong to heal. The General Healing practices listed steps A-K are the most important part you must always do in every session. You can do this with one person or with many people at the same time. The following is a detailed decription of each of these steps. Become familar with them and they will flow easily one into the other.

A. Three Minds into One Mind

Smile down to the inner universe. Activate the heart consciousness and empty the mind and heart down to the LowerTan Tien, the Abdominal Brain. Gather the 'Yi' (mind-eye-heart power), combining three minds into one. Fill the Tan Tien with this Chi. Start to spiral. You are then ready to connect to the higher forces of Universal and Heavenly Chi. The Three Minds are the Upper, Middle and Lower, or three Tan Tiens. The Upper Mind is the one we have been closely related to most of our lives. It is the logical thinking, planning, calculating and worrying mind; it consumes 80% of our body's energy.
The key is to learn 'to releas the mind' by learning to sink the Chi to the Lower Tan Tien, whereby it can be used in synergy with the other two.

B. Activate the Six Directions

Activating the six directions is for opening the whole body, mind and soul to all of the Universal and Cosmic forces in all directions. We literally learn to draw in the Universal and Cosmic forces from all directions at the same time. This is important for healers so they are not using their own energy for healing but drawing on the limitless Universal forces.

Expand your awareness to connect to the universe and the Six Directions at the mideyebrow and crown; left, right, front, behind, above and below. Connect to the universe, collect and spiral the energy into your Lower Tan Tien.

C. Activate the Three Fires

Activate the Tan Tien, Kidney and Heart Fires to create a Sacred Fire Triangle. Move from the Tan Tien to the Heart, the Heart to the Kidneys and the Kidneys back to the Tan Tien.

D. Linking Personal Stars, Energy Bodies and the Universe

World Link is a way for intelligent life force to return to the source; inwards and upwards. Become aware of the 'personal stars' above the crown and below the earth.We link with our own personal stars and then with the stars of each person in a group. This links the energy bodies of everyone in a group together. Start to spiral your energy and connect to other peoples' stars and energy bodies. Make a left spiraling pattern, progressing around the circle in a clockwise rotation. Link all of these stars together and form ***one large energy body.*** Each meditator will become an individual unit in an integrated communication link. A link between the earth and the universe.

E. Protective Circle and Chi Field

We create Chi Fields, Protective Animals and call the Eight Forces to protect the group from disturbances and psychic attacks. Raise your right (or leading) hand in the air and access the **'Sacred Fire'** of the universal cauldron. "Feel your fingers becoming big, long and hollow", as they reach into the Sacred Fire. The Chi will fill and pack your arm as you bring it down into your body. Use Yi-Power to create the **Sacred Fire Protective Circle.** The fire will purify and protect the circle. You can cast a circle around your community, house, room and body. This will create an energetic field of Chi around the whole room.

Then Create a **Chi Dome**. Invoke the 'guardian animals'; Blue Tortoise from the north, Red Pheasant from the south, Green Dragon from the east, White Tiger from the west, Yellow Phoenix from above and Black Tortoise from below.

Finally, activate the **Eight Elemental Forces of Nature and the Universe**. Call the forces of fire, water (ocean), thunder (lightning), lake (rain), earth, mountain, wind and heaven. Activate fire in the north, water in the south, lightning in the east, and rain in the west.

F. Chi (Sacred) Water Practice

Invoke the power of the Sacred, Chi or Holy Water practice to cleanse and heal the body's sick, toxic or negative energy. Make a request to receive healing energy and feel yourself touch a heavenly pool of sacred water energy. Feel the pool pouring down heavenly water to fill your arm.

Use the thumb and index finger to remove the sick energy from the cup, without touching the water. Talk to the sick cells. Let the sick energy return to the earth for recycling. Fill the water with your heart's compassion. Give the command: "This Sacred Water will give health, wealth and longevity". Cosmic Healing works on a cellular level. It is the water's 'job' to carry the message of the practice to all the cells, where it will remain. Drink the water in unison with the rest of the group.

G. Empty and Fill

The Empty and Fill technique is emptying the sick energy into the earth and filling with healing Cosmic Force Energy. Project your 'sword' hand (middle and index fingers) down to the ground. Project the hand 'very big and very long' and your feet 'very big and very long' - extended down to the ground. Now very slowly inhale, bring up the earth energy. Bring the energy up to your feet and into your bones. Let it enter all the bones, organs and cells. Feel it blending with all the sick energy. Turn your palm up, turn it back down again and exhale; let go of all the sick energy, the negative energy, the worries and burdens, all the way down into the earth. Repeat this three to nine times. Dig a hole and bury all the sick energy; tell it, "Do not come back; you will be happy down there." Smile.

H. Clean with Green Light

The Green light is for cleaning the cells. This is vitally important because the cells must be clean for healing. Green can help detoxify the toxic cells. Green Chi is mild and safe. Green Chi is used for cleansing, and as a decongestant for diseased parts. It 'loosens' the area. The light green Chi is used to break down dirty or diseased energy, like a detergent is used for washing clothes or dishes.

Afterwards, the cells will be rinsed with blue light, like clean water is used to rinse the detergent and residue from clothes or dishes.

When you carry out healing try to 'see' the cells enlarged and this green energy go in and blend with the dirty, black and cloudy energy. See it come out and let it go down to the earth. Dig a hole and bury it there. Tell it, "Don't come back. You will be happy down there". Become aware of the green light and spiral counter clockwise. Spiral down. Cleaning out the stress. Repeat as many times as you like, three times minimum.

I. Clean with Blue Light

Blue is like cold water; it has the yin power to dissolve all kinds of negative energy and sickness. It has an inhibiting effect. Blue yin energy is the opposite of red yang, which has a strengthening and stimulating effect.

Blue has a cooling effect; it can reduce pain and inflammation, and it can help blood clot. It stops bleeding and reduces fever. It can help to induce rest and sleep.

J. Charge with Violet Light

Charging with Violet light holds the highest healing energy. When the cells are clean they will absorb and hold the Violet light and the healing will take place. Luminous Violet has intelligence and can be programmed. The universe is full of violet light, especially the North Star and the Big Dipper. The color of the star of the higher self or soul, above the crown, is also violet. It is the color of Divine or Soul energy.

It has properties of other Chi colors. It has a rapid regenerating effect on damaged organs and nerves. Always use green and blue light before using violet. Violet light develops the crown center, the spiritual core. It is good for psychological ailments as well as for physical deficiencies.

K. Activate the Immune and Defense System

The body has various mechanisms which combine to provide protection and defense against illness and disease. These mechanisms enable the body to produce various cells and bodies which act against invading or unwanted substances. As soon as a foreign body is recognized, the immune system will be triggered and will act to provide the most efficient means of eliminating the danger and return the body to a balanced state of health. A strong and healthy body will have good resources to protect itself from negative or sick energy. The aim of these practices is to help you to realize the potential of true harmony within yourself. To activate the immune and defense system is to increase the production of red and white blood cells. To do this we activate the bone marrow and lymphatic system. Guide the student or whole group together in all of these steps. You will not lose energy by doing this. You will gain more energy because you are all connected together in the group power. The crucial part is to multiply your good intentions, Chi and virtue energy into the cosmos, making yourself and each person a direct line from the Universe. After you finish the General Healing session you can tailor the healing techniques in the next chapters to the individual's needs.

Practice

A. Three Minds into One Mind

1. Smile to the Inner Universe

Place your palms together in salutation, in front of your heart. Feel the Laogong points in your hands connect, creating an energy loop running from your heart through your arms and hands and back again.

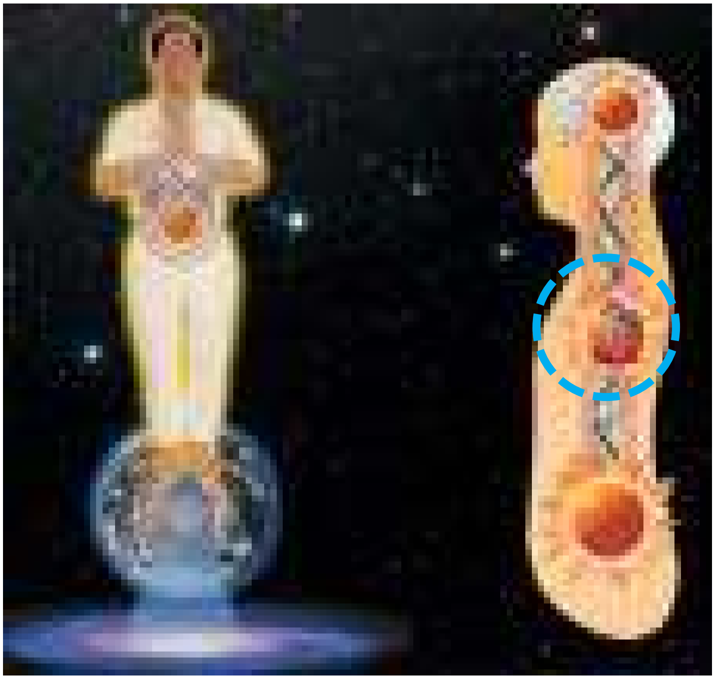

A. **B.**

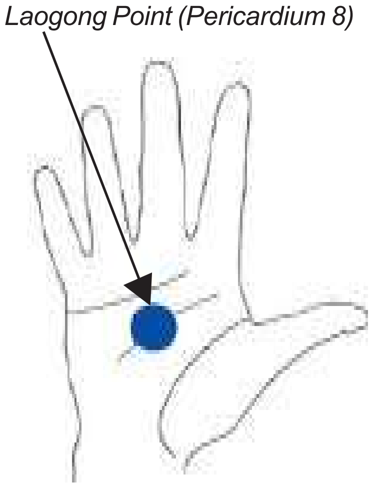

Fig. 4.1 A. *Empty the mind down to the lower abdomen.*
B. *Activate the heart's compassion.*

2. Activate the Heart Compassion Energy

Smile to the heart and feel it softening. Feel love, joy, compassion and happiness. Smile down and empty the mind to the Tan Tien, the Abdominal Brain. Fill the Tan Tien with Chi and start to spiral. When the abdomen is warm, it is full of Chi. The Chi can then charge up to the brain.

3. Combine the Three Minds into One

1. Turn Upper Mind to observation mind.
2. Turn consciousness (Middle Mind) down to Lower Tan Tien.
3. Combine three minds into one mind at the Lower Tan Tien (the Yi).
4. Manifest out at the mideyebrow.

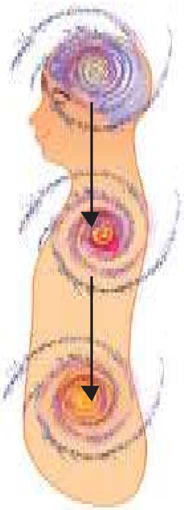

Fig. 4.2 *'Three Minds into One'.*

B. Activate the Six Directions

Expand the awareness to connect to the universe and the Six Directions at the crown, mideyebrow, heart and Tan Tien.

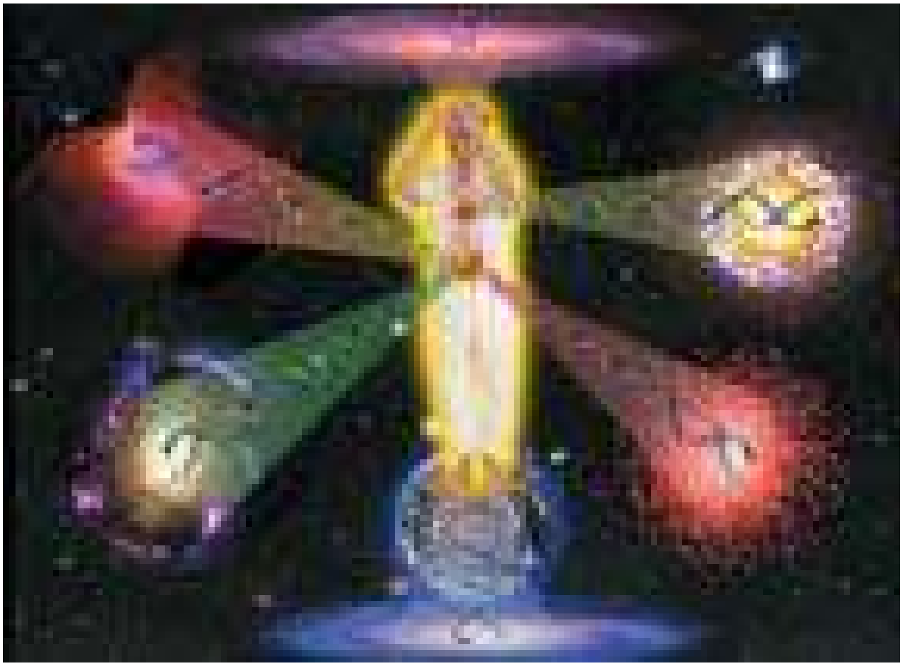

Fig. 4.3 *Trust and believe - turn visualization into actualization.*

C. Activate the Three Fires

1. Tan Tien Fire - Activate the Tan Tien Fire:

Smile down to the abdomen to create a burning stove near the lower lumbar and sacrum. Create a fire ball behind the navel, above the stove.

2. Kidney Fire - Activate the Kidney Fire:

Always retain awareness at the Tan Tien and keep it spiraling.

3. Heart Fire - Activate the Heart Fire:

Keep the heart soft and fill the heart with joy, love and happiness.

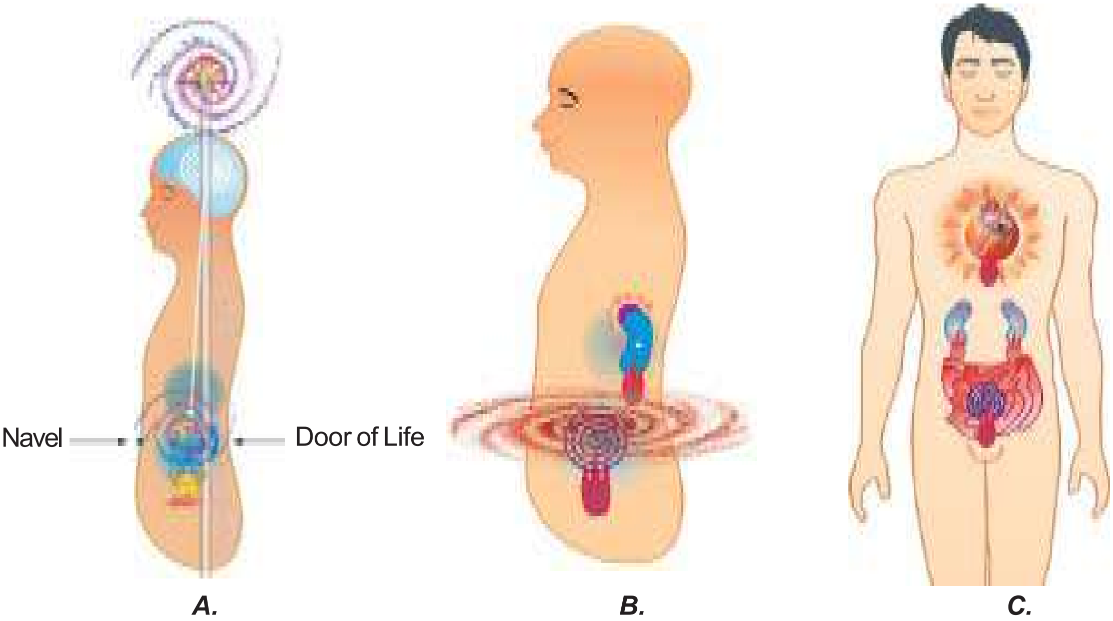

Fig. 4.4 A. *Activate the Lower Tan Tien Fire.* ***B.*** *Activate the Kidney and Adrenal Fires.* ***C.*** *Activate the Heart Fire.*

4. Sacred Fire Triangle

The Sacred Fire Triangle has triple force. Make a triangle from kidneys to navel (Tan Tien). From the heart make the connection to the kidneys. Move from the Tan Tien up to the heart.

5. Cosmic Star and the Earth Star

At the moment of conception, the two forces yin and yang connect with such a force, then only fractions of a moment later, the two forces explode into nine different Energy Centers. Seven of them we find in the body and two of them outside the body, forming our personal stars.

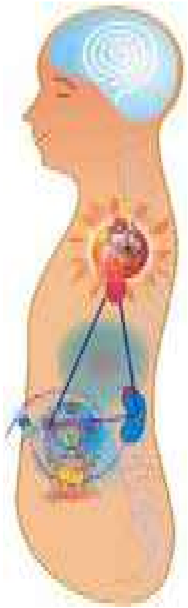

Fig. 4.5 Three Fires into One

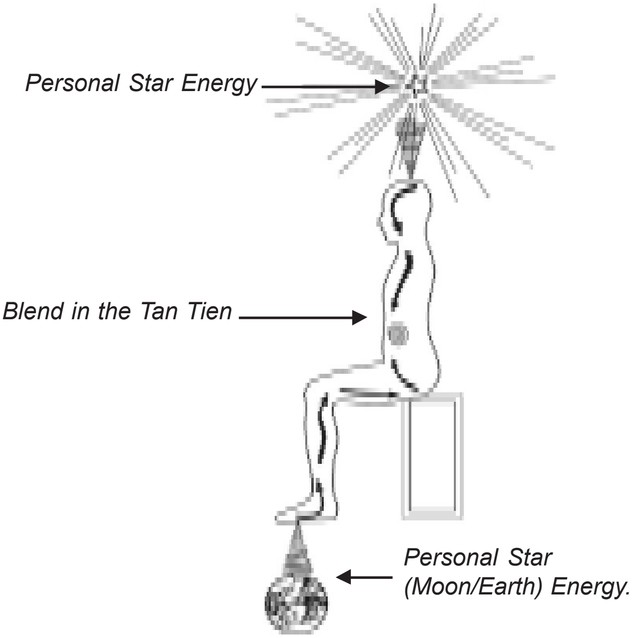

Fig. 4.6 'Personal Star' Energy

The two personal stars are in fact energy centers connecting the aural field of each individual with the universal forces and the earth forces. In a way one can see the aim of Taoist Inner Alchemy in bringing these nine forces together, merging them into one force and thus enabling the return to the original force, the Wu Chi.

There is a star about six inches above your crown and another one about one to three feet below the soles of your feet. They are also known as the higher self, guidance, protector, adviser. These stars are our connection to the cosmic force, the universal force and the earth deep below us.

Always make sure your Tan Tien is warm and the sacrum and mideyebrow are breathing. Be aware of the crown breathing and see a star or a small sun above you. Be aware of the crown and feel a light beam extend out of the crown and make a connection to the star above you. Keep on breathing until you feel a strong connection. Feel how the star above you is exercising a strong pulling force on your crown. Once you feel this pull on your crown you will also feel a strong pull down from the earth. Be aware of the star above and the earth and universal force below you. Feel that both of them have a strong pull on you.

1. Reconfirm the star above you and the earth directly below you.
2. Fix the image of the North Star and the Big Dipper six to nine feet above your crown. See the cup of the Big Dipper filling with the violet light, which gathers Chi from the North Star and the universe.

3. **Reach up with your left hand, and hold on to the handle of the Big Dipper. Pour the violet light down to your personal star 4 to 6 inches above your head, in order to predigest the energy of the violet light. Then let it flow down to your crown, into the Upper Tan Tien, to be processed and flow down to the heart center (Middle Tan Tien) or to the back of the head down to C-7 (Cervical 7) and T-2 (Thoracic 2) and down to the palms.**

   In Taoism we consider that the North Star emits a violet light, which is regarded as the highest healing light and the Big Dipper as emitting a red light.

   You can do these meditations sitting, or standing in the Chi Kung posture. The standing position will support a stronger structure and provide a better connection with the earth.

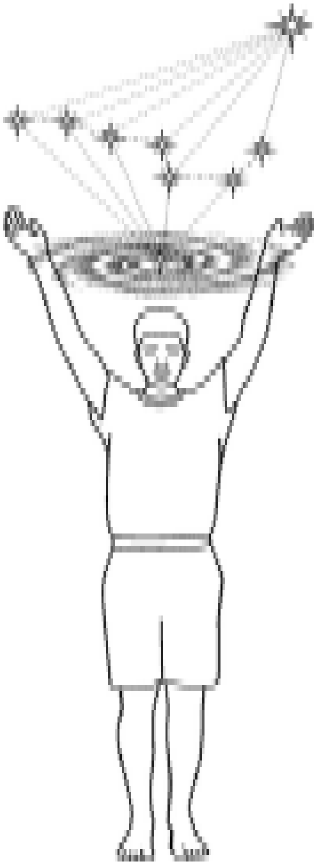

Fig. 4.7 *The North Star, Big Dipper and Galaxy spiral above.*

4. Always remember to breathe in slowly so you can process the awareness of the Lower Tan Tien. Feel a suction, warmth (Chi) and continue to breathe in and be aware of the suction of the crown, North Star and Big Dipper above you.

D. Linking Personal Stars, Energy Bodies and the Universe

Be aware of your 'personal stars' above and below and connect to other peoples' stars and energy bodies.

Three minds unite, one mind aware of the stars above you and a star below you. Connect these stars to the people around you. Think about the bright light above the crowns of other people. Start to spiral your energy to connect to each of them. Group all the stars into one energy body and connect all the way to the center of the universe.

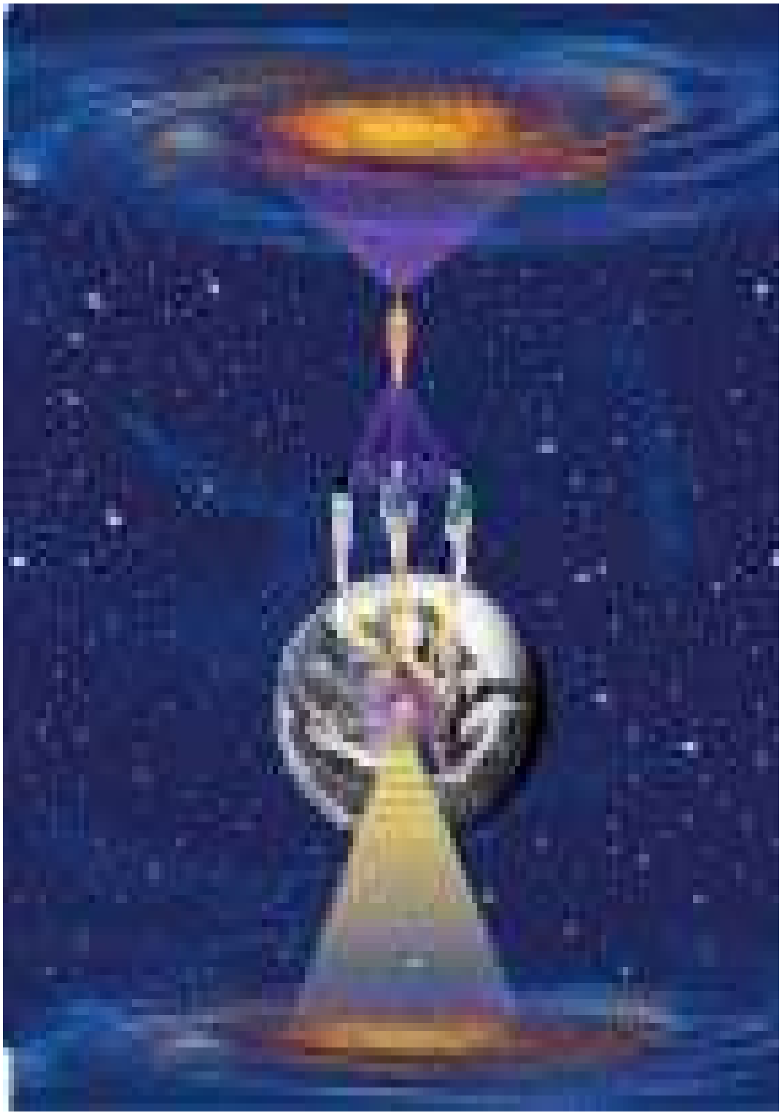

Fig. 4.8 *World Link Meditation*

1. Center Channel and Personal Star above

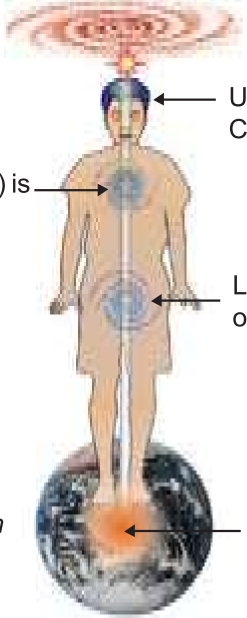

Soul, personal star or higher self

Upper Tan Tien (Upper Brain) is the Center of Observation.

Middle Tan Tien (Middle Mind) is the Center of Consciousness.

Lower Tan Tien (Lower Brain) is the Seat of Awareness.

Connect the Central Channel to the Upper and Lower Universe.

Fig. 4.9 *Connect Heaven to Earth using the Central Channel.*

2. Use Yi power to connect to your Personal Star above

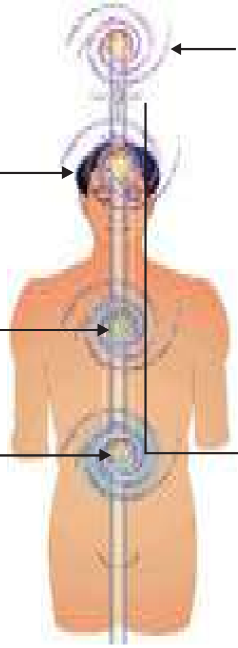

Personal Star

Observation

Consciousness

Awareness

When the abdomen fills with Chi it will rise up and fill the upper brain. Connect to your personal star, six inches above your head.

Fig. 4.10 *Connect to your Personal Star Energy using 'Yi' Power.*

3. Connect your Personal Star to the Personal Star of each person

Connect your personal star to each person and then to the universe.

4. Create a Chi Body becoming the major connecting point. Counterclockwise spiral to connect each person's star into a larger group star.

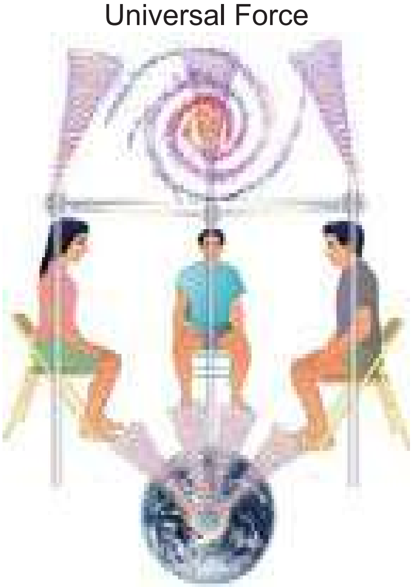

Fig. 4.11 *Link with Each others Personal Stars.*

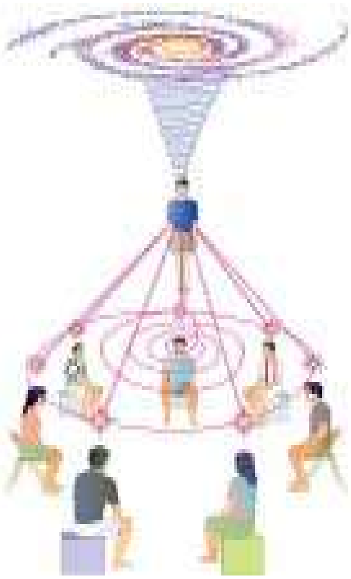

Fig. 4.12 *Connect into Group Star.*

5. Meditators become satellite links and create a group energy body.

The World Link meditations become an integrated communication link. They are linked to the earth and the universe. Spiral and connect each person's star together and group the energy into one energy body. This energy body becomes an integrated communication center for each person in the group. We can be connected, wherever we are. Each person can connect and become their own 'Center of the Universe'.

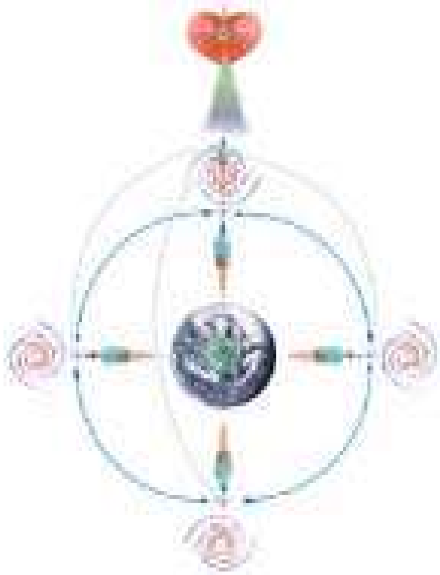

Fig. 4.13 *Form a Satellite Link and create a Group Energy Body.*

E. Protective Circle and Chi Field

Raise your hand up and become aware of the 'Sacred Fire' in the universe. Feel that your fingers are very 'big', very 'long' and touch the fire. Bring the fire down. Make a circle around your community, your house, Tao Garden, the meditation hall and around your bodies. Create an energetic field of Chi around the whole room.

Sacred or Chi Fire
Connect and Receive Fire from the Universal Cauldron

1. Activate the Sacred Fire in the Universe

Be aware of yourself. Feel your arm and fingers 'big', 'long' and 'hollow'. Reach to the universe to the Sacred Fire. Let the Sacred Fire fill and pack your arm.

2. Use Yi-Power to Create Sacred Fire Protective Circle

Use the Yi Power to draw the Sacred Fire Circle on the ground around your house, office and the room you work in.

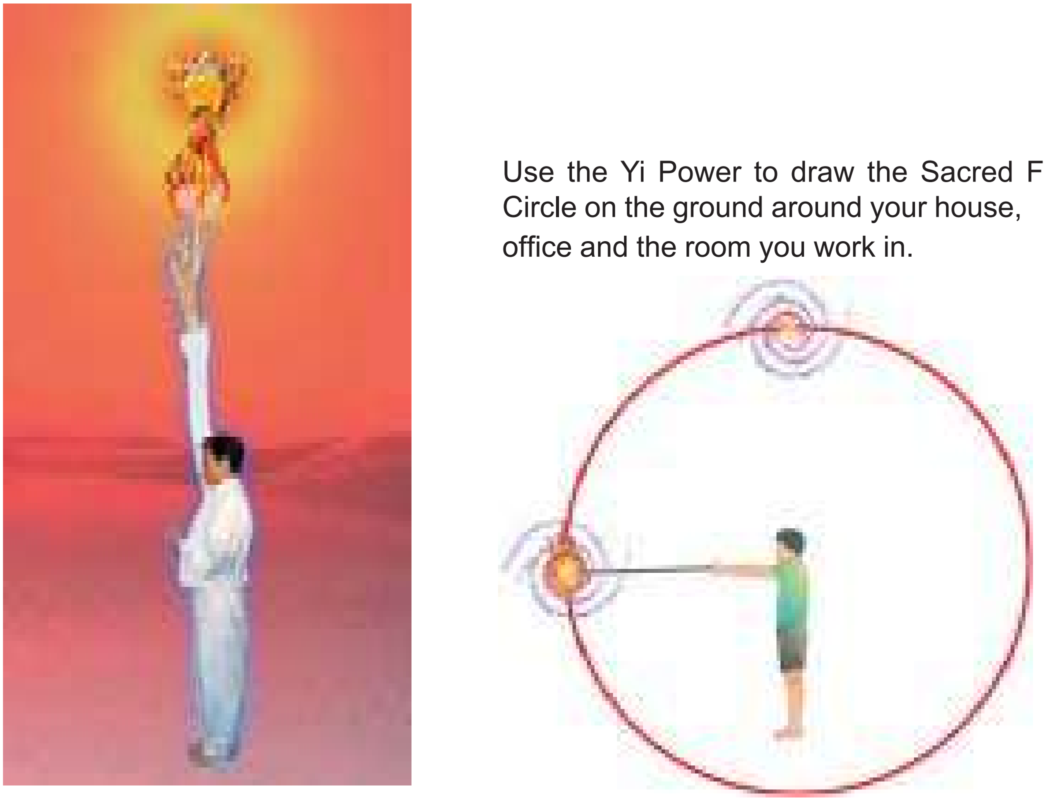

A. *B.*

Fig. 4.14 A. *Create a Group Energy Body.* ***B.*** *Sacred Fire 'Protection'*

3. Create a Chi Dome

Set up the ‘guardian animals’, Blue Tortoise in the north, Red Pheasant in the south, Green Dragon in the east, White Tiger in the west, Yellow Phoenix above and Black Tortoise below.

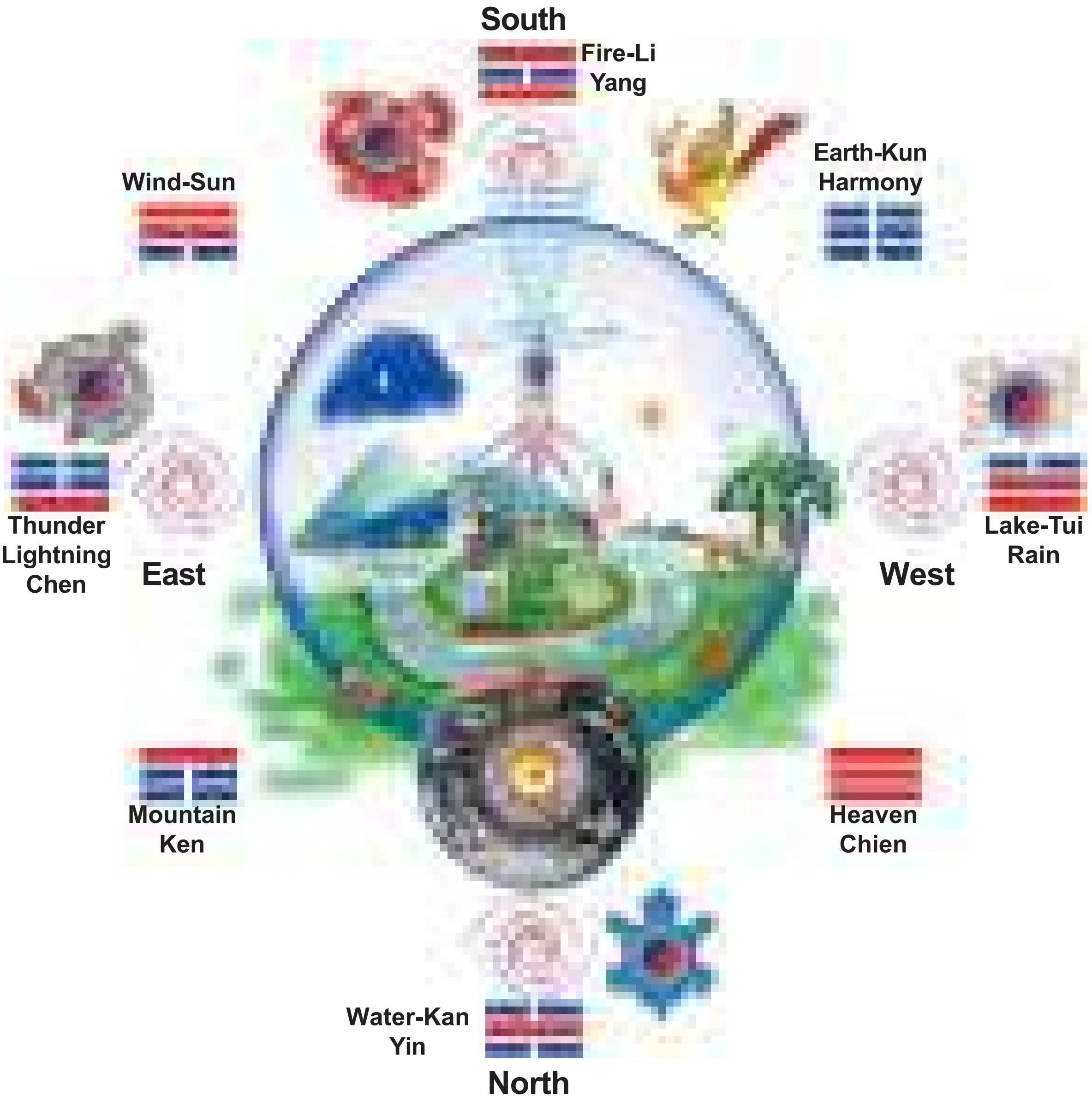
South
Fire-Li
Yang
Wind-Sun
Earth-Kun
Harmony
Thunder
Lightning
Chen
East
West
Lake-Tui
Rain
Mountain
Ken
Heaven
Chien
Water-Kan
Yin
North

Fig. 4.15 *Create and surround yourself with a Chi Dome and the Guardian Animals.*

4. Activate the Eight Elemental Forces of Nature and the Universe

Activate all eight forces; wind, mountain, fire and thunder on the east. Earth, lake, water and heavenly power on the west.

Call the Eight Forces; fire, water (ocean), thunder (lightning), lake (rain), earth, mountain, wind, and heaven.

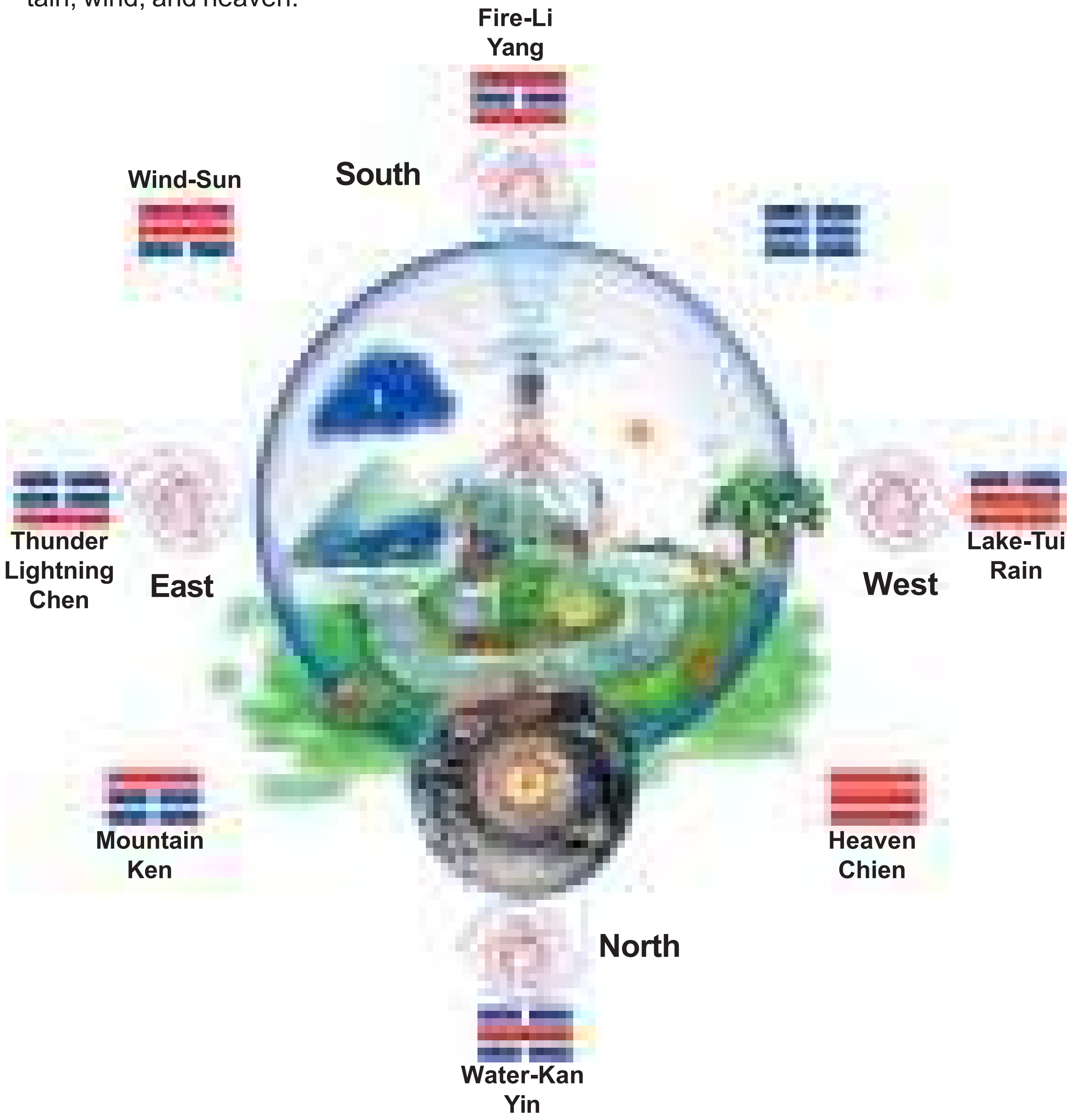
Fire-Li
Yang
Wind-Sun
South
Thunder
Lightning
Chen
East
West
Lake-Tui
Rain
Mountain
Ken
Heaven
Chien
North
Water-Kan
Yin

Fig. 4.16 *Activate the Eight Forces.*

In the next sections F,G,H,I are practices that are for cleaning the cells. The more you clean the cells the more the Cosmic Healing can take place.

F. Chi (Sacred) Water Practice

The Sacred Water practice (also known in the West as Holy Water) is to exercise the right of being the creators of the Cosmos. With strict and absolutely regular daily water intake to prevent the stresses and associated damages of dehydration, the chief conductor and supervisor of the body's well being – tryptophan and its neurotransmitter derivatives, serotonin, tryptamine and melatonin will be well positioned to regulate all functions. Regular daily walks will keep muscles well coordinated and correct any physiological processes that are established in the body as a result of anxiety or emotional stress.

A well hydrated and healthy skin needs water to constantly replace that which it loses to the outside environment. This then allows blood vessels to the face and the body to open up and provide necessary nourishment for exposed skin cells.

Science has discovered that even if we clean water through filtration processes, although it may look and taste alright, closer analysis using high powered microscopic technology, indicates that its cell structure is 'unclean'. Water, like humans, maintains the memory of its old 'self'. We store our 'memory' within our DNA; water stores its memory within its cells.

The world is about 70% water

You are about 70% water
and remain so.
Thirst Center

Fig. 4.17 *Water is Life.*

The following three pages contain extracts from the book ‘The Message from Water’ by Masaru Emoto. It advises us that ‘water is telling us to take a closer look at ourselves - water is a mirror reflecting our mind.’

水からの伝言
The Message from Water

After offering a prayer to the Fujiwara dam the message is reflected in the water.

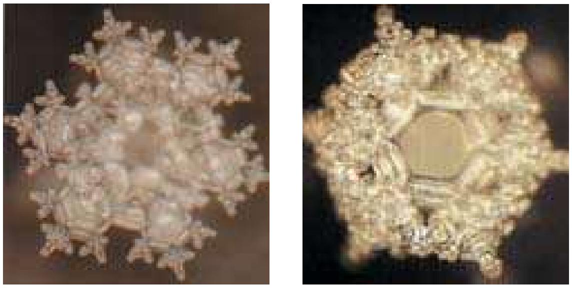

Chi and Love **Love-Appreciation**

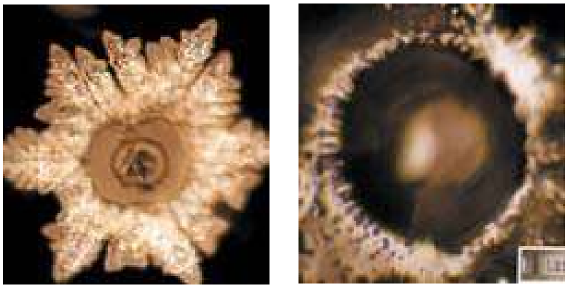

Let’s do it **Do it!**

Fig. 4.18 *The Human Body consists of 70% Water.*

When we project Chi and love into water and then freeze it; it will turn into a ‘regular’ and well structured crystal. If we project negative energy into water, the opposite will apply. Our projections will access the cellular level, clean out the polluted cells and reprogram those remaining with the essence of our affirmation.

When we project a nice, loving instruction or thought into water, like ‘let’s do it’ the water will pick up this vibration and restructure itself. If we project negativity into the water, like a command, ‘do it!’, the water will acknowledge this accordingly. When we look into a pond we see our image. If we drink a glass of negative image water its destructive energy will enter our system and attempt to multiply.

Research has demonstrated that if we completely filter polluted water, using every process available, it will still maintain its old vibration and unstable structure. It will maintain memory in the same way that a computer does. We must therefore reprogram the cells with our intent.

If the water we drink contains too many pollutants and antioxidants it will eventually poison us.

As we can see from these pictures, there is an enormous difference between natural spring water and regular city water. It is now quite clear that if we program water so that it becomes, sacred, holy or Chi water, then it will have the power to remove all antioxidants and pollutants – all sickness. It will have the power to reprogram the water within our bodies, to carry the same pure structure.

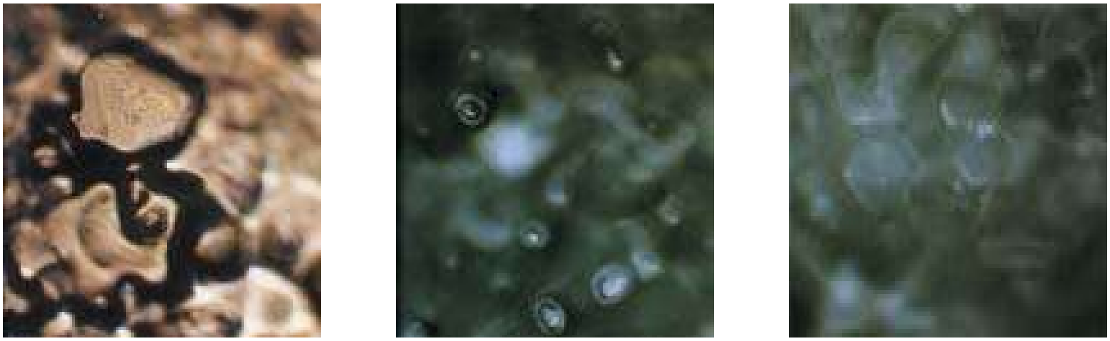

Tokyo Tap Water **London Tap Water** **Paris Tap Water**

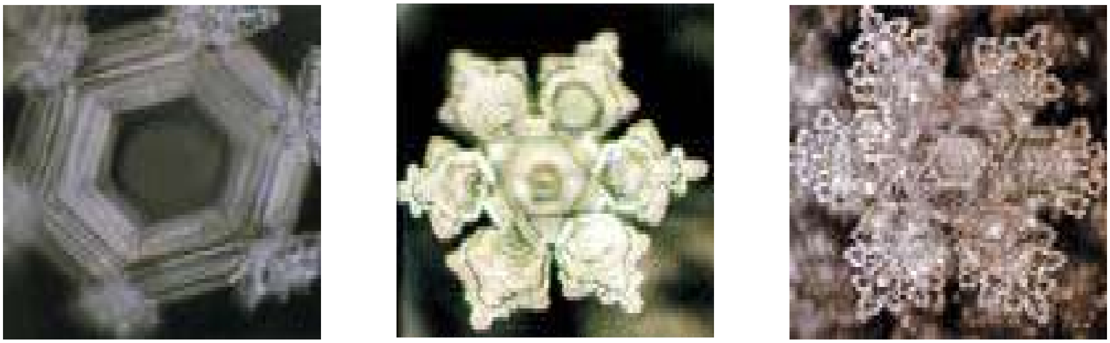

Under Ground Water in Northern Island, New Zealand **Buenos Aires, Argentina** **Spring Water of Saijo, Hiroshima Prefecture**

Fig. 4.19 *Healing or hurting?*

We can also use music to program a vibration into water. Classical music will create a vibration that synchronizes with our own health, whereas heavy metal music will have a tendency to 'shatter'!

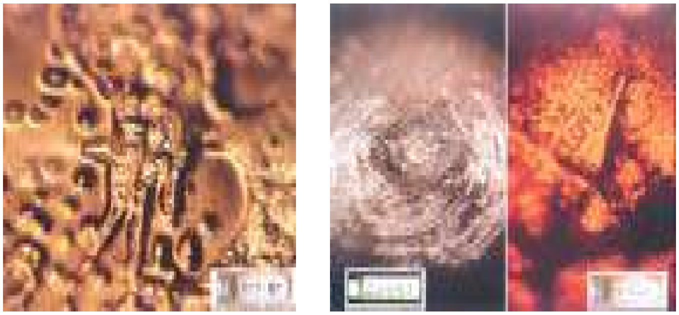

"You make me sick", or "I will kill you".

You Fool

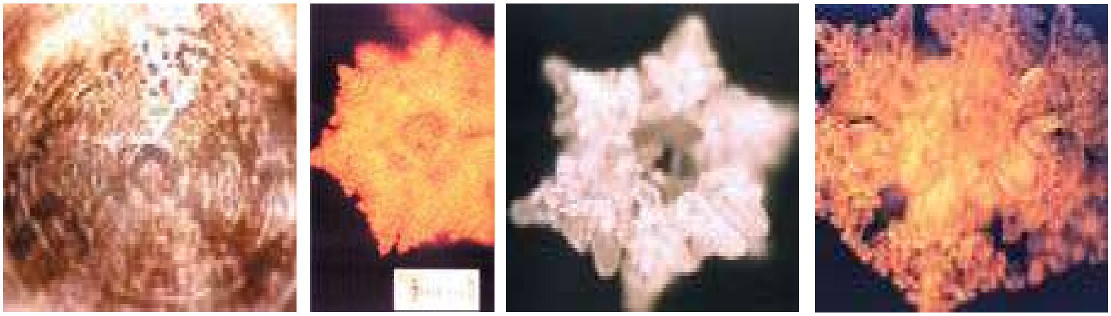

Heavy Metal Music

Thank you

Folk song of Celtic Region in UK

Healing Music

Fig. 4.20 *Power to heal or harm?*

Similarly, when we program our emotions regarding other people into water, the effects are the same. If we think negatively of someone, then the response will produce a bad vibration. If we think kind, loving thoughts then the structure will be like a beautiful crystal. If we love Mother Theresa and send this energy into the water, a wonderful crystalline structure will be created. Sick water can be transformed into healthy water. Healthy water can be transformed into sick water. The human body consists of 70% water. The key is your intent and your ability to believe that healing comes from the nature of your vibration. Always be open to receive.

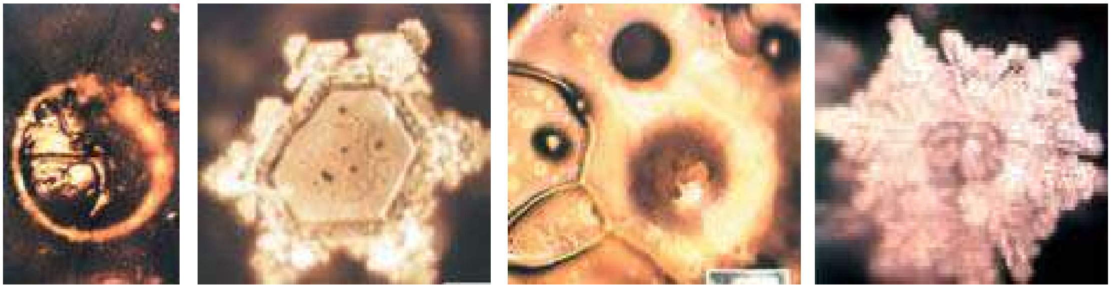

Adolph Hitler

Mother Teresa

Dirty

Beautiful

Fig. 4.21 *Sick or saintly?*

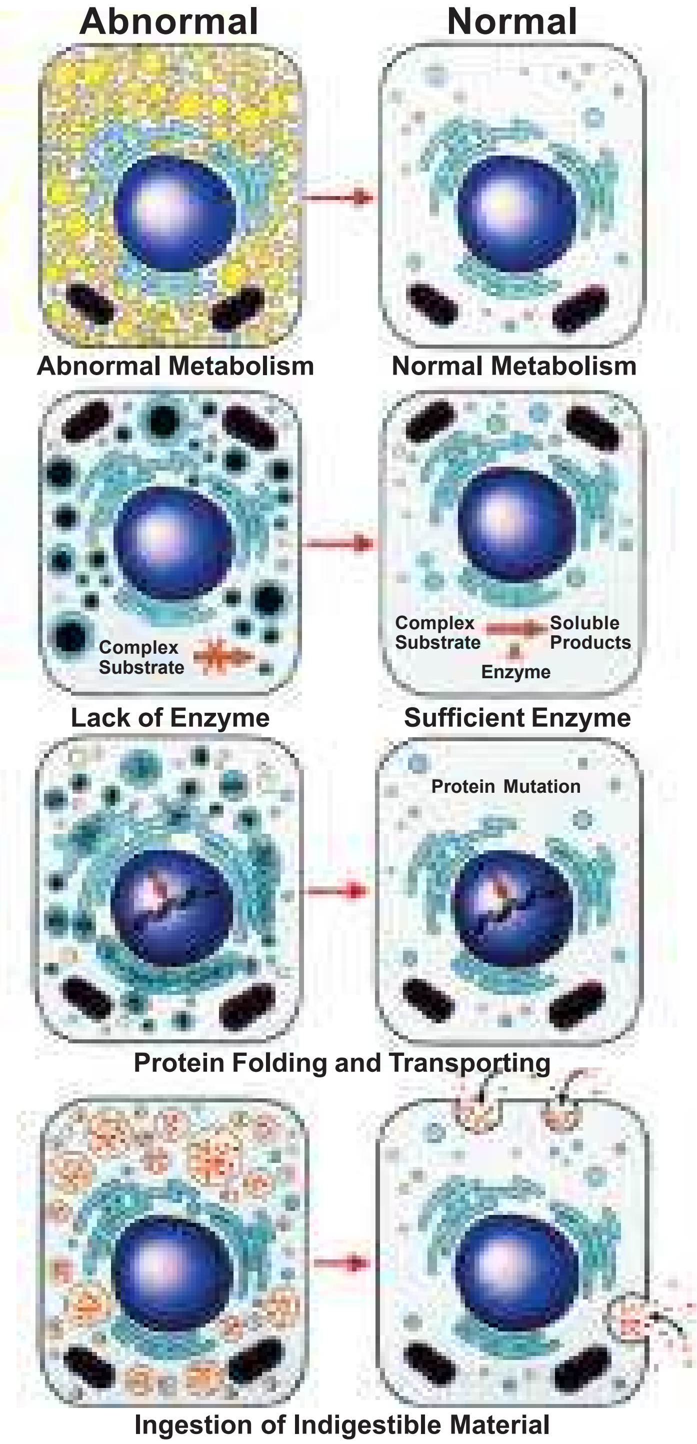
Abnormal
Normal
Abnormal Metabolism
Normal Metabolism
Complex Substrate
Complex Substrate
Soluble Products
Enzyme
Lack of Enzyme
Sufficient Enzyme
Protein Mutation
Protein Folding and Transporting
Ingestion of Indigestible Material

Fig. 4.22 *With the Power of the Sacred Water Abnormal Cells can be cured.*

A positive mental attitude will be transmitted into the cells of the water contained within our own body. We must consciously practice inner work until it becomes reflexive.

Outline of Sacred Water Practice

Invoke the power of the Sacred Water practice to cleanse and heal the body's sick, toxic or negative energy. **If you are working with a group of people, direct your energy through the energy body overhead and into each person's star, as you guide them through the procedure**.

I. Hold a glass one quarter full of water in your left hand, folding the middle and the ring fingers into the center of the palm.

II. Hold the glass in front of your body as you point the 'Sword Fingers' of your right hand up to Heaven.

III. Make a request to receive healing energy and feel yourself touch a heavenly pool of sacred water energy. Feel the pool pouring down heavenly water to fill your arm.

IV. Place your fingers on the cup and ask:
Command 1: *"Yin power and good fortune come from the east. Yin power please dissolve all negative energy, all sickness and bad fortune".* Make a cross over the top of the cup.

V. Bring your arm down and point the sword fingers around the inner rim of the glass. Smile as you circle the fingers around the rim.
Command 2: *"By my request. Please carry out the order now."* Repeat this three times, charging the water with Yin Power. Project your thoughts into the water.

VI. Use the thumb and index finger to remove the sick energy from the cup, without touching the water. Do this three times. Talk to the sick cells. Tell them to listen to your command:
Command 3: *"All sick cells please listen: Clear, clean and bright, this Sacred Water will take all the sickness away."*

VII. Repeat the order and visualize yourself removing the sick energy from the cells and giving it to Mother Earth for recycling.

VIII. Hold the cup with both hands near your heart.
Command 4: *"This Sacred Water will give me/you health, wealth and longevity."*
Project love, joy, thankfulness, gratefulness, appreciation and the energy of compassion into the cup.

IX. The Sacred Water will carry the message of the practice to all the cells, where it will remain. If you are practicing in a group, drink the water in unison. If you are working with a student, pass them the glass to drink using both hands.

Step I: Prepare the Hands and Arms to Receive the Universal Chi

Left Hand Position: Prepare to hold a cup or glass with the left hand by folding the ring and middle finger into the center of the palm. Hold the glass in front of the body.

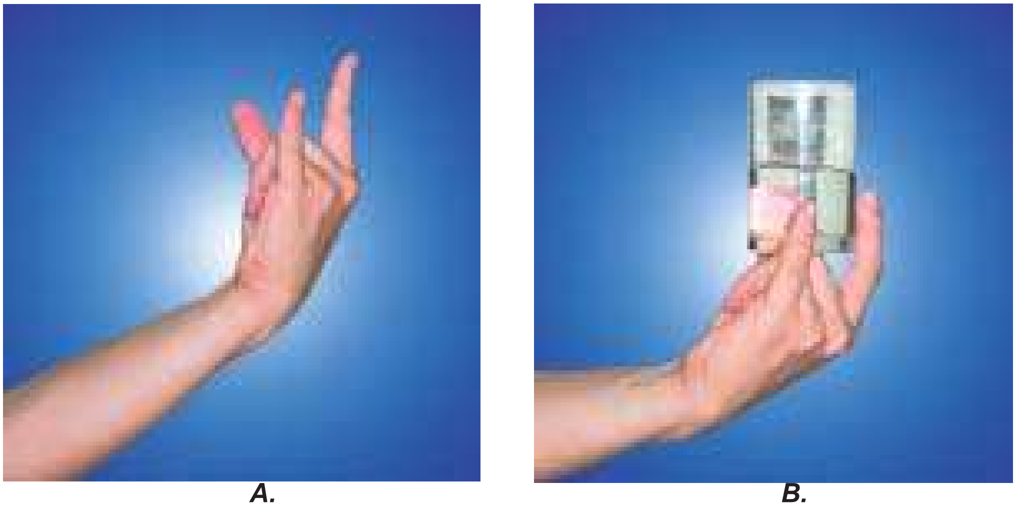

A. **B.**

Fig. 4.23 A. *Fold Ring and Middle Fingers.* ***B.*** *Hold Glass in Front of Body.*

Step II: Prepare the Sword Hands to Receive the Universal Chi

Right Hand Position: Form the right hand into a 'sword hand' by folding the pinkie, ring fingers and thumb into the palm. Straighten and hold the index and middle fingers together and extend them upward.

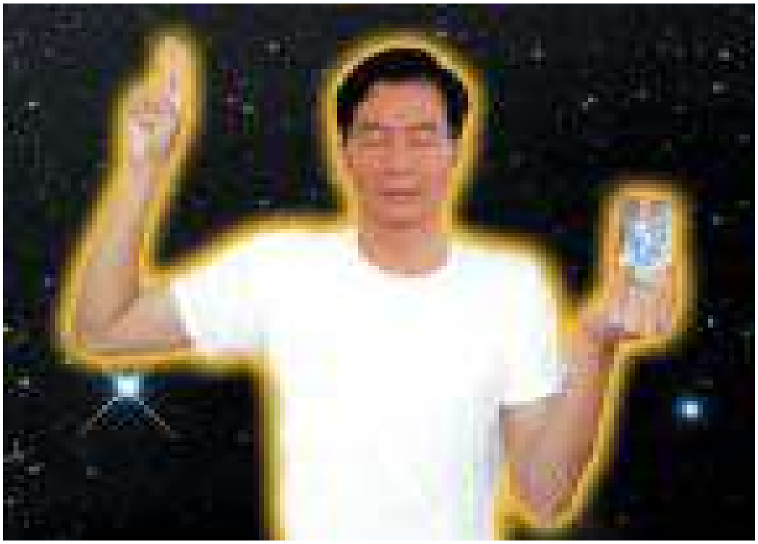

Fig. 4.24 *Prepare to receive the Universal Chi.*

Step III: Fill your Arm with the Power of the Heavenly Pool

Sense the sword fingers and the arm as being ‘long’ and ‘big’ as you raise them towards Heaven. Sense that the middle of the arm is hollow and bottom is sealed at the shoulder. The mind’s attention is focused on merging with the primordial Chi of the Universe, the energy of your thoughts will be multiplied.

As you make your request to the Universe, feel that you are touching a heavenly pool of Sacred Water Energy. Feel that the pool is pouring down like a waterfall to fill the arm. When it is full compact, compress the Sacred Energy into it as much as you can.

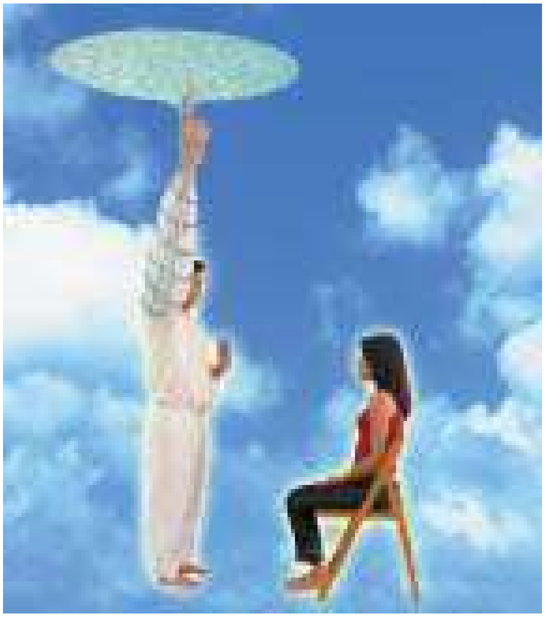

Fig. 4.25 *Touch the Heavenly Pool.*

A. ***B.***

Fig. 4.26 A. *Blessing Sacred Water in Russia.* ***B.*** *Opening the Third Eye in Russia.*

Step IV: Make the Cross on the Top of the Cup

Put Sword Fingers on top of the Cup. Ask for the Yin Power: **Command 1:** ***"Yin Power and good fortune, come from the east. Yin has the power to dissolve all negative energy, all sickness and bad fortune."***

Command 2: Make the cross on the top of the cup and say ***"By my request......"***.

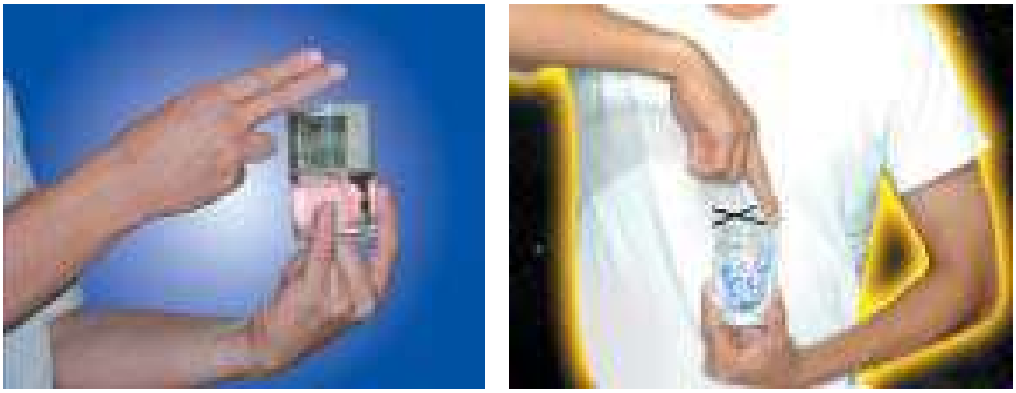

Fig. 4.27 *Give the command: "Fortune comes from the East - Yin has the power to dissolve all negative energy, all sickness and bad fortune. By my request....".*

Step V: Charge the Water to Transform it into Sacred Water

Bring your arm down and point the sword fingers into the glass of water. Smile and circle the sword fingers around the inner rim of the glass.

Continue to order: ***"Please carry out the order now."*** Repeat 3 times to charge the water with the Yin Power. As you do this, project a good thought into the water.

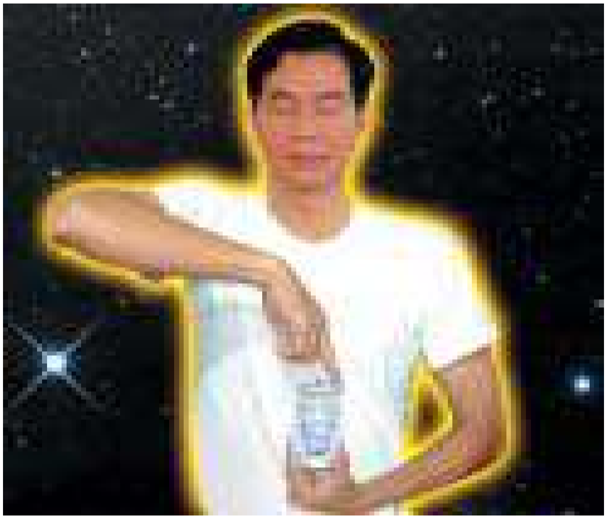

Fig. 4.28 *"Please carry out the order now" Repeat three times, charging the water with the yin power.*

Step VI: Command 3: *"All sick cells please listen : Clear, Clean and Bright, this Sacred Water will take all the Sickness away."*

Give the above command. Use the thumb and index finger to pick up sick energy from the cup, without touching the water. Do this three times.

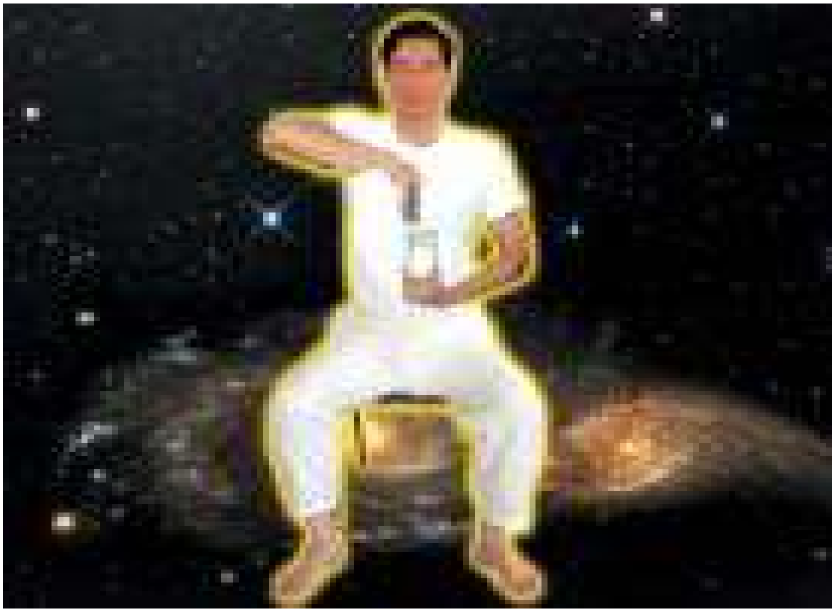

Fig. 4.29 *"All sick cells please listen: clear, clean and bright, this sacred water will take my/your sickness away".*

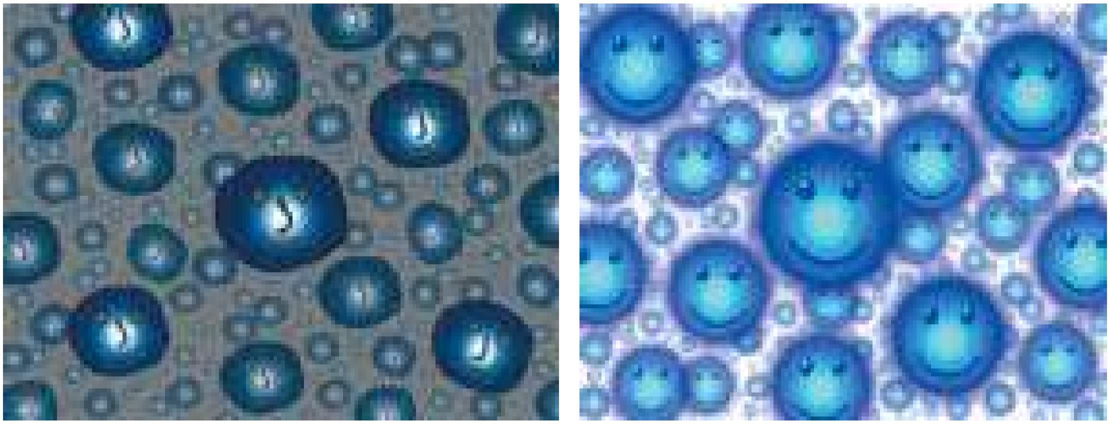

A. ***B.***

Fig. 4.30
A. *Ask for the power to see the cells - if they are toxic and dark, ask for them to be cleaned out.*
B. *"Clear, clean and bright".*

Step VII: Let go of the Sick Energy to be transformed by the Earth

As you **repeat** the order, picture yourself removing the sick energy from the cells and discarding it down into the Earth to be transformed and recycled.

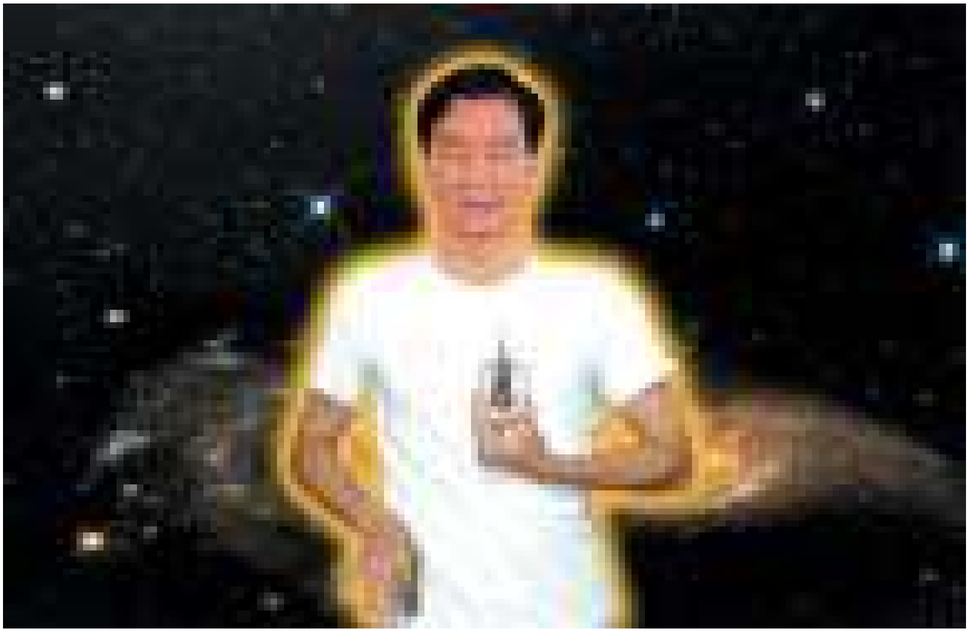

Fig. 4.31 *Sick Energy return to the Earth.*

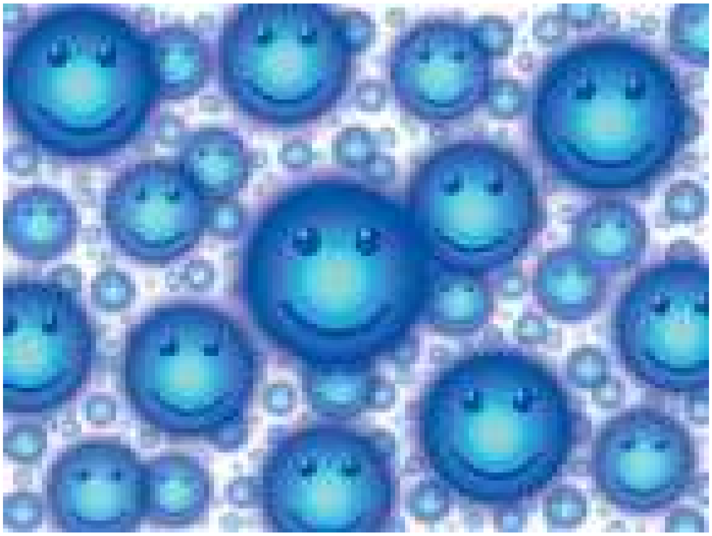

Fig. 4.32 *See the Cells become Clean, Bright and Smiling.*

Step VIII: Fill the Water with Compassion

Hold the cup with both hands near the heart and project goodness into the water.

Command 4: ***"This sacred water will give me/you health, wealth and longevity."***

Project love, joy, thankfulness, gratefulness, appreciation and the energy of compassion into the cup.

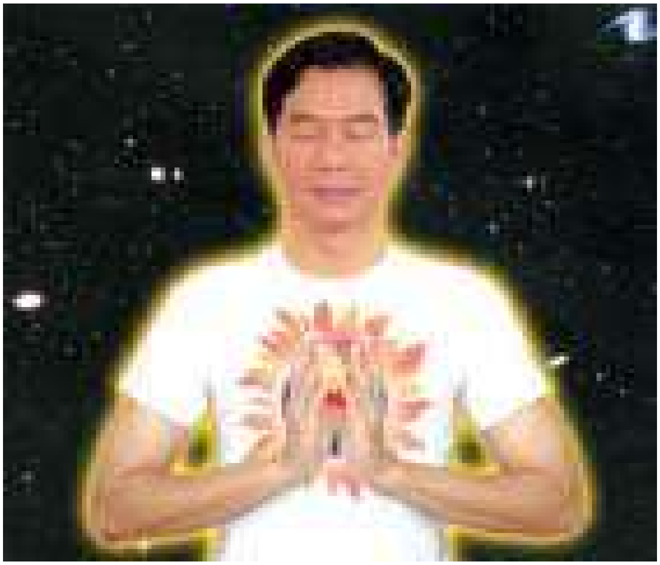

Fig. 4.33 *"This Sacred Water will give you health, wealth and longevity."*

Step IX: Drink the Sacred Water

The Sacred Water will carry the message to all the cells. It will keep the message of the Sacred Water practice in all the cells. Use both hands to pass the Sacred Water to the student you are working with, or drink it yourself and/or in unison with your group. **Feel the water go into all the cells of your body. Feel it removing the sick cells and purifying your body.**

You can also sprinkle it on to any area that requires healing.

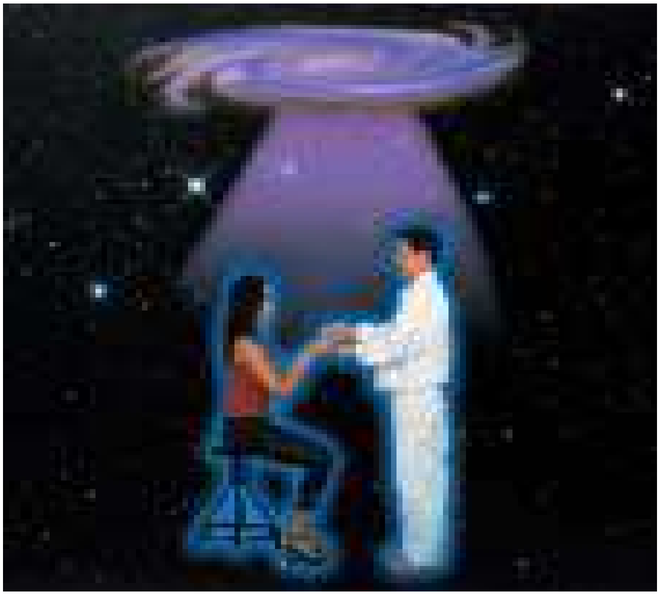

Fig. 4.34 *Drink the Sacred Water.*

G. Empty and Fill

Ask for permission to work on the student by communicating with their personal star. Point your sword hand down to the ground. Project your hand 'very big and very long' and your feet 'very big and very long' - extended down to the ground. Now very slowly inhale, bring up the earth energy. An important part of the practice is to extend a 'line' down from universe into yourself and one into the student. The line sinks through the Tan Tien into the universe below. You can send your energy up to your energy body and it will be brought down to the energy body of the other person.

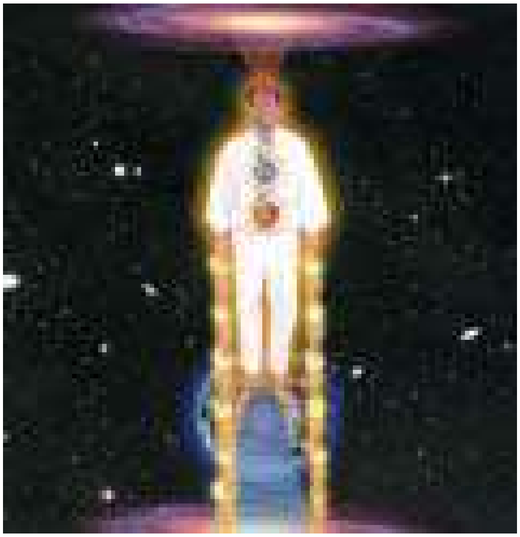

Fig. 4.35 *Feel your hands are long as down into the earth they extend.*

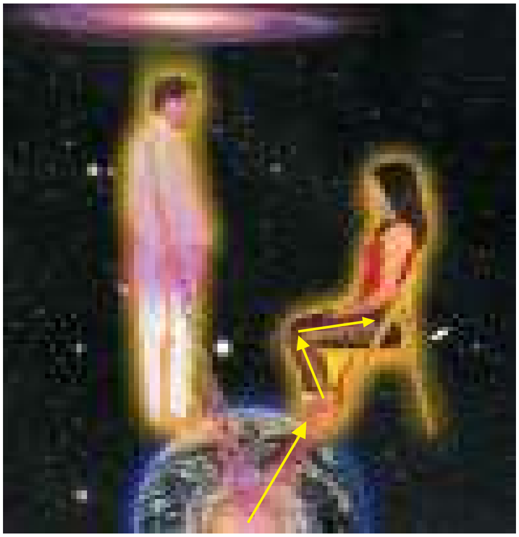

Fig. 4.36 *Project the other person's legs deep into the earth.*

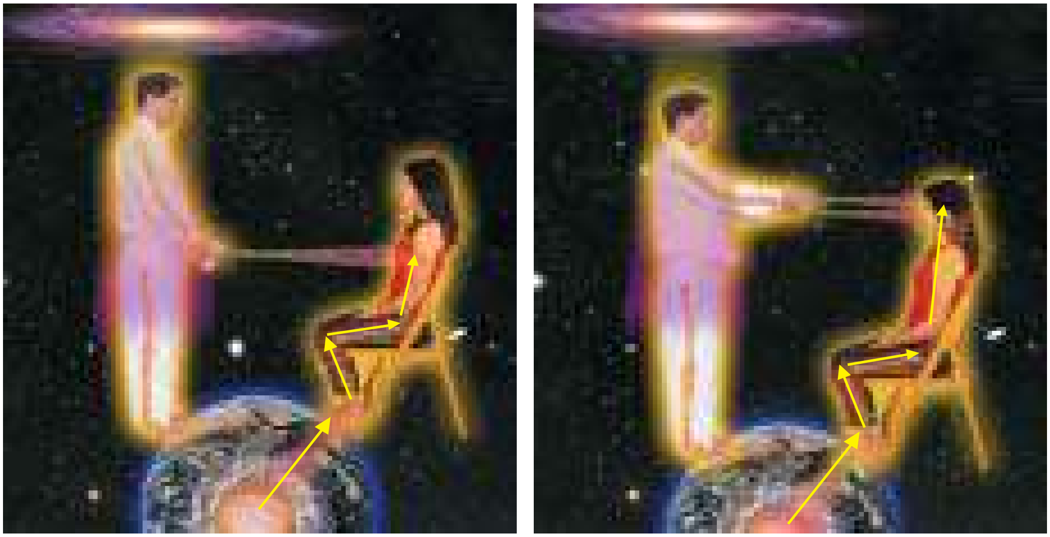

Fig. 4.37 *Blend the universal energy and Earth Chi - bring it into all of the bones and cells.*

Bring the energy up to your feet, into your bones. Let it enter all the bones, organs and cells. Feel it blending with all the sick energy. Turn your palm up, turn it back down again and exhale; let go of all the sick energy, the negative energy, the worries and the burdens all the way down to the earth. Down, down, down to the earth. Again from the earth up. Do it at least three, six or nine times. When people are very sick, you need to clean them thoroughly. Bring the earth energy up into your bones again. Feel numb, tinkling electricity flowing in all your cells. Return the energy down, down, down to the earth. Relax, let go; worries, burdens, let everything go. It is very important to say to yourself or group: **"You must let go of all your sick energy. Let go of all your burdens. Let go of all your worries, down to the earth".** Repeat it once again. Inhale from the universe and into the earth. From the earth let it enter into your bones. From the bones all the way to your inner self. Feel the energy blend with the sick energy. Exhale and let go of all your sick energy, any negative emotions burdens, worries, anxieties; release them all the way down into the earth.

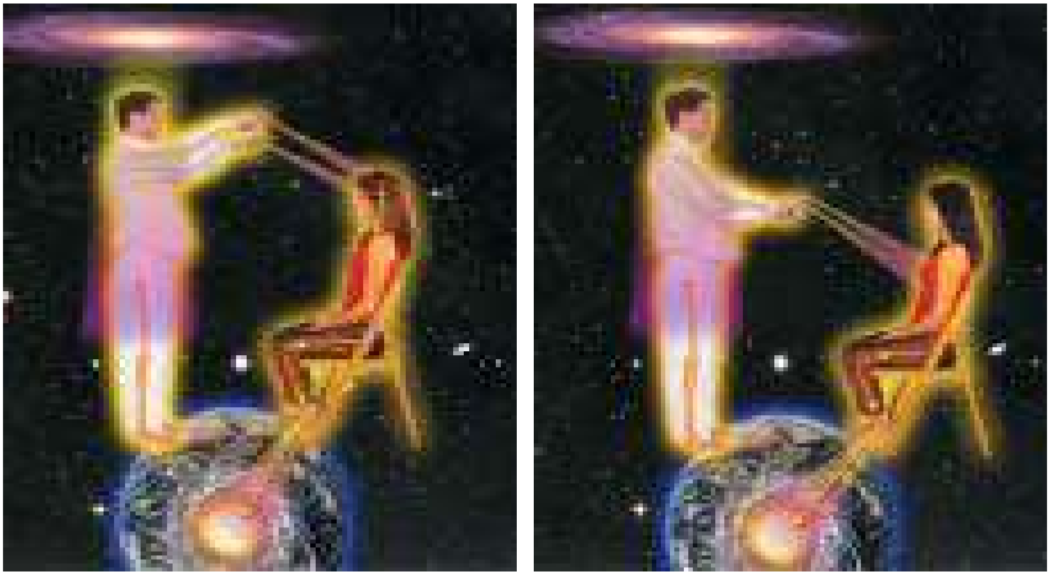

Fig. 4.38 *Bring the Universal Chi down and let it blend with the sick and negative energy. Bring it down into the earth and bury it there.*

Empty negative, sick energy, burdens, worries, go deep down into the earth. Dig a hole and bury them there - tell them "Do not come back; you will be happy there."

Feel that the feet are long and extended into the earth. The hollow bones have been compacted with compressed Chi.

Once again guide the good Chi from earth up to the fill the person again. See the cells as big as the stars and fill them with Chi to blend with the dirty and sick energy. Then empty them again down to the earth.

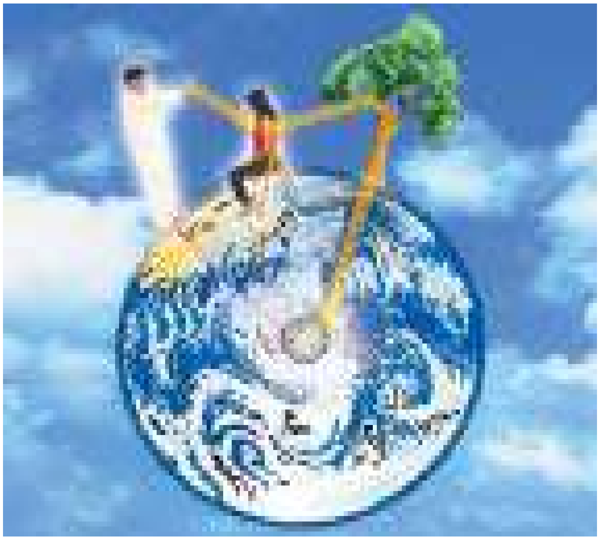

Fig. 4.39 *You can also use a Tree to guide the Chi down into the Earth.*

H. Clean with Green Light

Become aware of your liver and breathe into your throat center. Feel that you are breathing the green light into your throat center and let it blend in the Tan Tien. The Chi comes up to blend in the heart, up to the crown, spirals and goes up to the universe. Raise your hand in a spiraling motion up to the universe. Use your 'one hand' to spiral in the universe. Spiral clockwise first. Spiral green light from the universe. Now reverse, counter-clockwise - multiply and spiral counterclockwise, bringing the Chi down to clean out your communities, your house, the Tao Garden, the meditation hall, your bodies and all of the people that are 'here' - fill yourselves with Chi. You will use this power to see the cells.

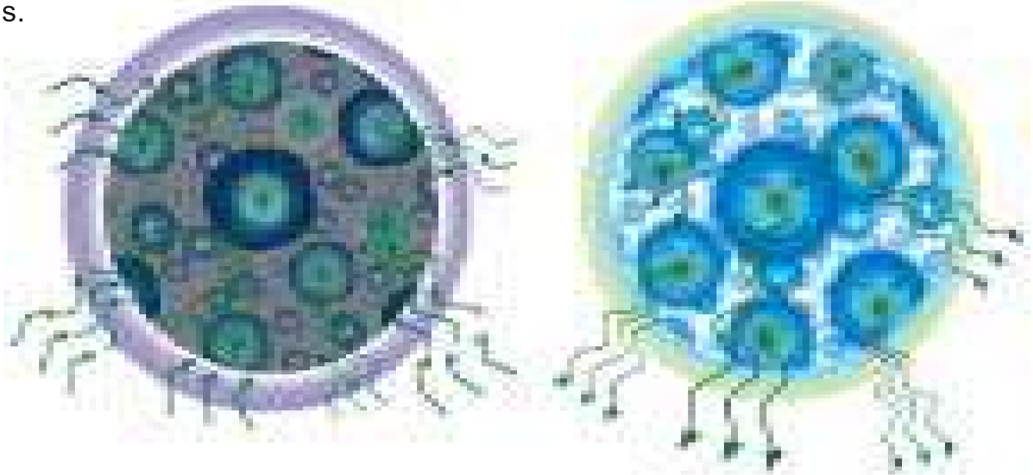

Fig. 4.40 *Spiral Sick Energy Counterclockwise down to the Earth. Smile to the Cells and let the Green Light clean out the sickness, flushing it down into the Earth below.*

See the cells enlarged and this green energy go in and blend with the dirty, black and cloudy energy. See it come out and let it go down into the earth. Dig a hole and bury it there. Tell it, “Don’t come back. You will be happy down there”. Become aware of the green light and spiral counterclockwise. Spiral down, cleaning out the stress. Clean out your house. Send everything down into the earth. Repeat three, six, nine, thirty six or one hundred and eight times. This depends on how much sick energy you are dealing with in your body.

Green:

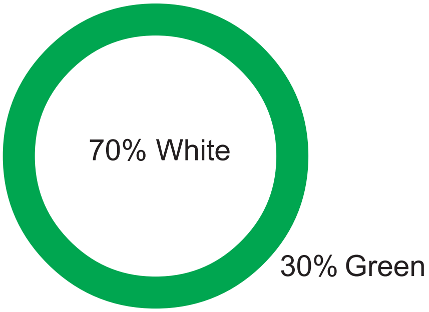

Fig. 4.41 *Clean with green.*

Green cleans and detoxifies all of the body’s cells. Green Chi is mild and safe. Green Chi is used for cleansing, and as a decongestant for diseased parts. It ‘loosens’ the area. The light green Chi is used to break down dirty or diseased energy, like a detergent is used for washing clothes or dishes. Afterwards, the cells will be rinsed with blue light, like clean water is used to rinse the detergent and residue from clothes or dishes.

Throat Center Green

Be aware of the liver’s green color, the throat and the forest. Breathe into the throat, feel yourself breathing in the green light. Guide it down to the Tan Tien, blend and move it up to the heart and the crown.

Project it to the universe; spiral and let it multiply.

Channel one direct line down to the student, one to your crown and one to your palm.

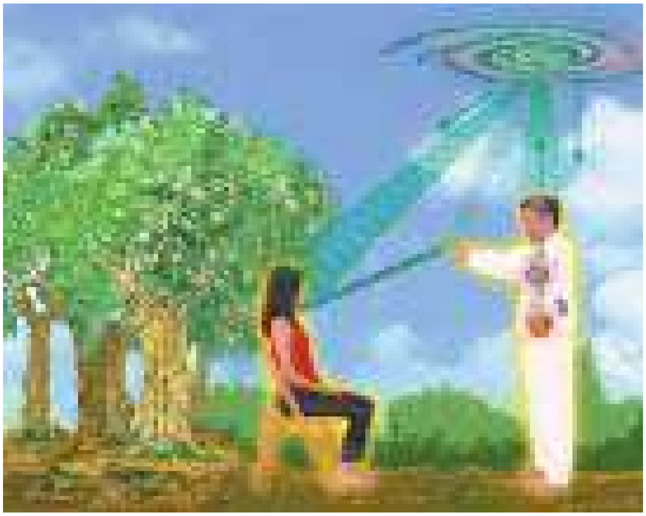

Fig. 4.42 *Blend Green Chi into the Three Tan Tiens - project to the Universe, multiply and bring the Energy back down to the Student.*

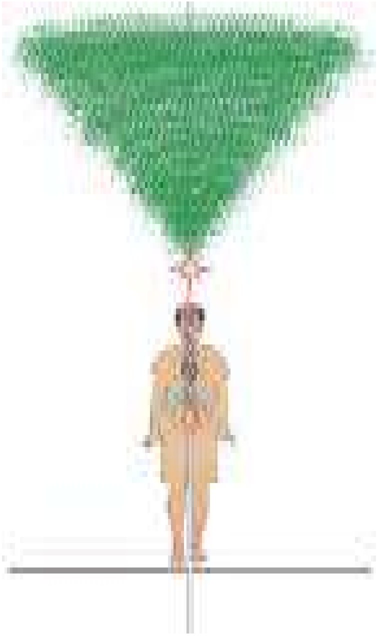

Be aware of the liver, the gall bladder, the forest and the throat breathing.

Be aware of the green light in the universe. Let the light come in and blend in the Tan Tien, up to the heart to blend with the compassion in the heart, and up to the crown.

Project the light up to the universe. See it spiral and let it multiply.

Always keep your Lower Tan Tien Fire warm. **Spiral in your Tan Tien and spiral the energy down to the student.** Spiral 1,000 revolutions per minute, 10,000, 30,000 and 60,000.

Fig. 4.43 *Draw down the Green Light from the Universe.*

Cleanse and Remove Sickness

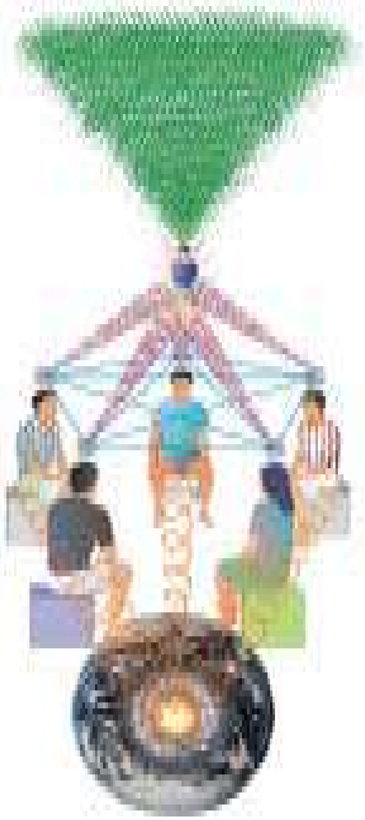

Extend your arms up, with palms facing the heavens. Use one hand to spiral the green force down. The other hand holds the position and connects to the universe. Let this big pool of green light in the universe spiral down – through your community, your home and then into your crown.

Let the Green light blend into the cells and bond with sickness and toxins. Let the green light bring the sickness out of the cells and flow down to the center of the earth.

Dig a hole and bury it. Give the command: “Don’t come back. You will be happy down there. You will be transformed into good Chi.” Smile. Always smile.

Repeat it six, nine or eighteen times. For people who are very sick, who have cancer or may be terminally ill, you can do this two or three hundred times.

Fig. 4.44 *Green Light bonds with the Sickness and Toxins before returning them to the Earth.*

I. Clean with Blue Light

Be aware the blue light, kidney blue. Breathe the blue light into the throat center. Now slowly blend the blue light in the Tan Tien. Blend the blue light in the heart, in the crown and project it up to the universe, spiraling clockwise. Then feel the universe pouring the blue light back down; counterclockwise. Spiral down. Look at the cells and see the blue light, like water, go in to the cells to rinse and clean all the dirty things, down, down, down to the earth. All the way down. Repeat three, six or nine times. Smile.

Blue:

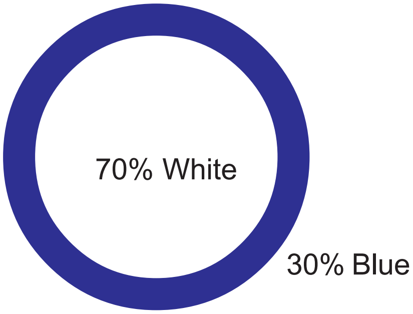

***Fig. 4.45** Rinse with Blue.*

Blue is like cold water; it has the yin power to dissolve all kinds of negative energy and sickness. It has an inhibiting effect. Blue yin energy is the opposite of red yang, which has a strengthening and stimulating effect. Blue has a cooling effect; it can reduce pain, inflammation and can help blood to clot. It stops bleeding and reduces fever. It can help to induce rest and sleep.

Throat Center Blue

Be aware of the kidneys' blue color and breathe the blue light into your throat. Guide it down to the Tan Tien, blend and move back up to the 'heart of compassion' and then the crown.

Project up to the universe, spiral and let it multiply.

Channel one direct line down to the student, one to your crown and one to your palm.

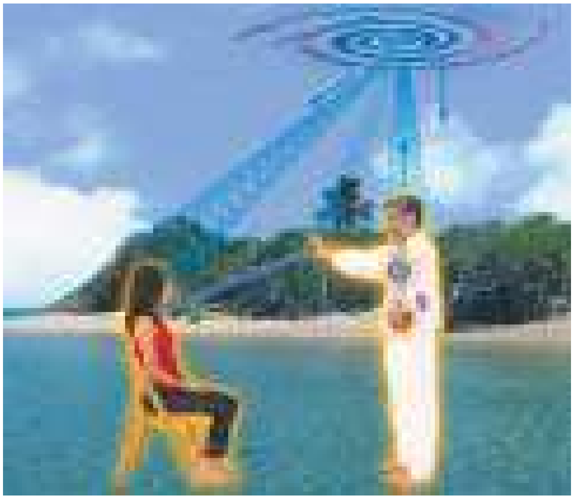

***Fig. 4.46** Multiply the Universal Blue Light - guide the Energy from the Throat, Heart and Crown into the Universe and back down.*

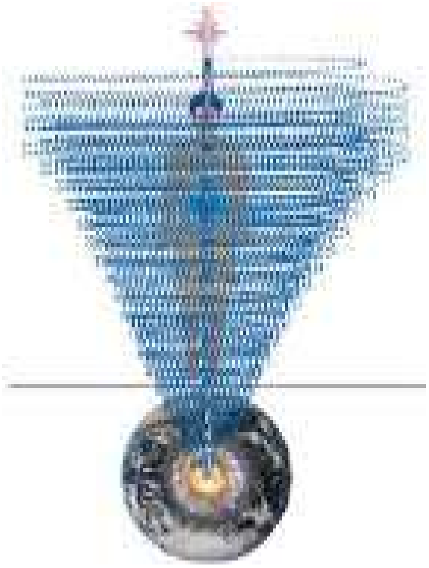

Be aware of the kidneys, the bladder, the oceans and the throat breathing.

Be aware of the blue light in the universe. Let the light come in and blend in the Tan Tien, up to the heart to blend with the compassion and up to the crown.

Project the light up to the universe. See it spiral and let it multiply.

Fig. 4.47 *Draw the Blue Light from the Universe - rinse clean and remove sickness.*

Rinse and Flush

Extend your arms in the air, with palms facing heaven. Use one hand to spiral the green force down. The other hand holds the position and connects to the universe. Let a big pool of green light in the universe spiral down – through your community, your home, and then into your crown.

Let the blue light flush into the cells and flush out sickness and toxins towards the center of the earth.

Dig a hole and bury them. Give the command: "Don't come back. You will be happy down there. You will be transformed into good Chi."

Repeat it six, nine or eighteen times. For people who are very sick, who have cancer or who may be terminally ill you can do this up to thirty six, seventy two or one hundred and eight times.

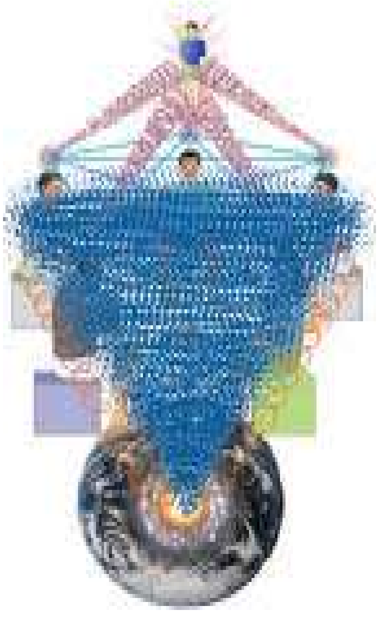

Fig. 4.48 *Blue Light flushes out the sickness and toxins before returning them to the earth.*

J. Charge with Violet Light

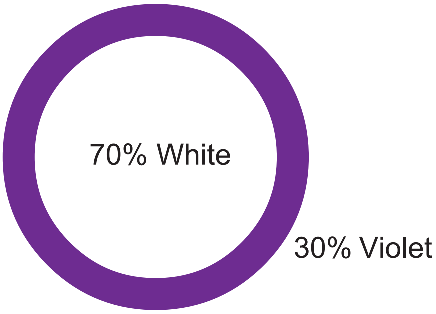

Fig.4.49 *Intelligent Luminous Violet*

Luminous Violet has intelligence and can be programmed. The universe is full of violet light, especially the North Star and the Big Dipper. The color of the star of the higher self or soul, above the crown, is also violet light, the Divine or Soul energy.

It has properties of other Chi colors. It has a rapid regenerating effect on damaged organs and nerves. Always use green and blue light before using violet. Violet light develops the crown center, the spiritual core. It is good for psychological ailments as well as for physical deficiencies.

Primordial Force

Turning the three minds into one mind, go deep into the empty space of the cell and deep into the chromosomes, magnify the space as you enter into the DNA and return to the original cells. The empty space is the Wu Chi. When you enter this stage you turn the subconsciousness into consciousness and you can talk to your body and cells. You can change the programming of DNA and RNA.

Ask for the power to see the cells and look at the brain cells spread out into the space. "Brain cells please listen. Clear, Clean, and Bright: fill with violet light and return to normal function".

Violet Light into the Cells

See violet light. Be aware of your crown. The Tan Tien Chi and the negative Chi all combine in the crown. Extend a violet light up to the universe. See the North Star in the Big Dipper. Tell the student to hold their arms in the air, so that they can funnel the Chi into their personal star. Bring the North Star and the Big Dipper down; hold the handle of the Big Dipper and pour the violet light over the student's personal star. The higher cells will be programmed. Let the Chi enter the brain and ask for the power to see the brain cells. "Brain cells listen to the command". If there is sickness in the brain, give the command. "Sickness go away".

The North Star is a major source of violet light.

The Big Dipper is a major source of red and infrared light. Taoists understand that the cup of the Big Dipper gathers all the violet light from the universe. The left hand holds the handle of the big dipper - pour it down to your crown. It then flows down throughout the body.

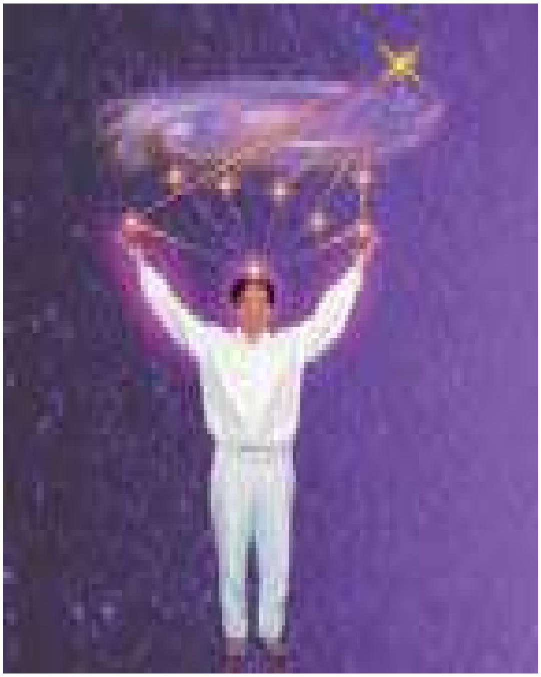

Fig. 4.50 *Connect with the North Star to receive the Violet Light.*

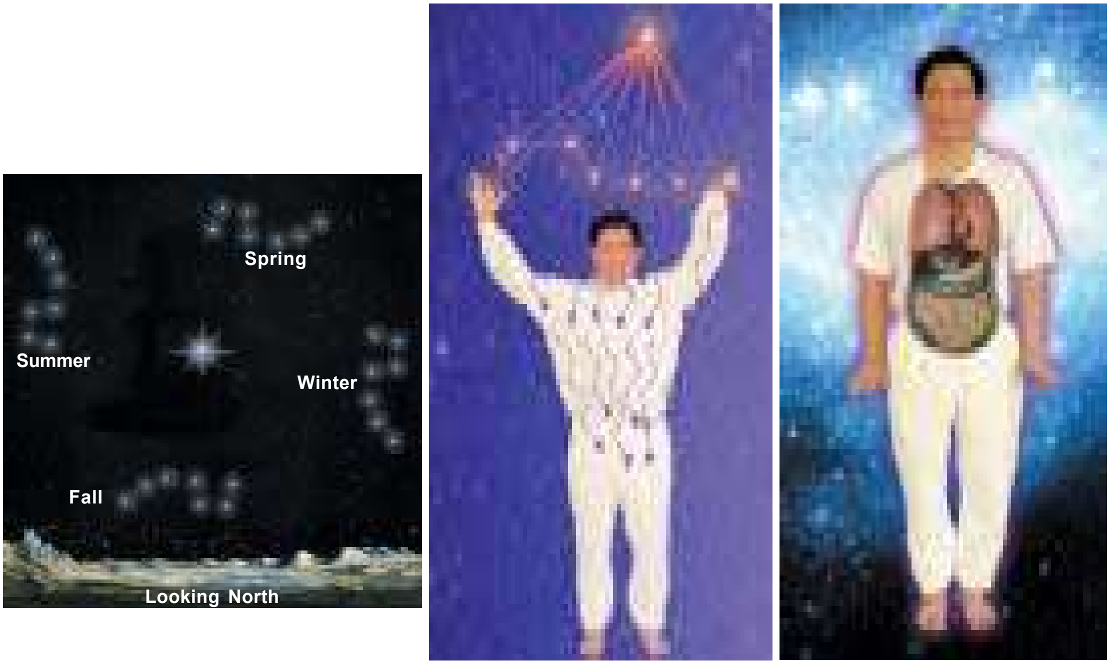

Fig. 4.51 *Practitioner connects with Universal Violet Light.*

Bring the violet light down to the student’s personal star above their crown. Pause to let the higher cells of the brain reprogram the violet light at the personal star. Continue to spiral the violet light down through the cells of the brain and down through the cells of the whole body.

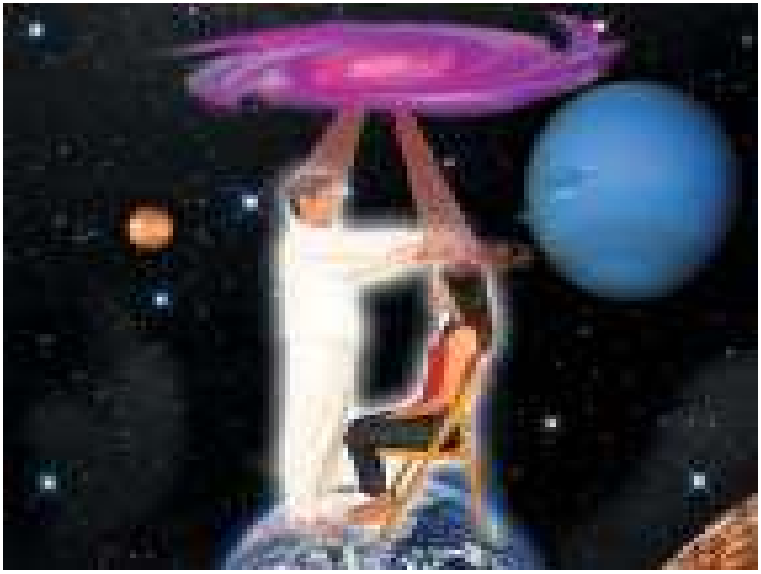

Fig. 4.52 *Pause to reprogram the Violet Light at the other Person’s Personal Star.*

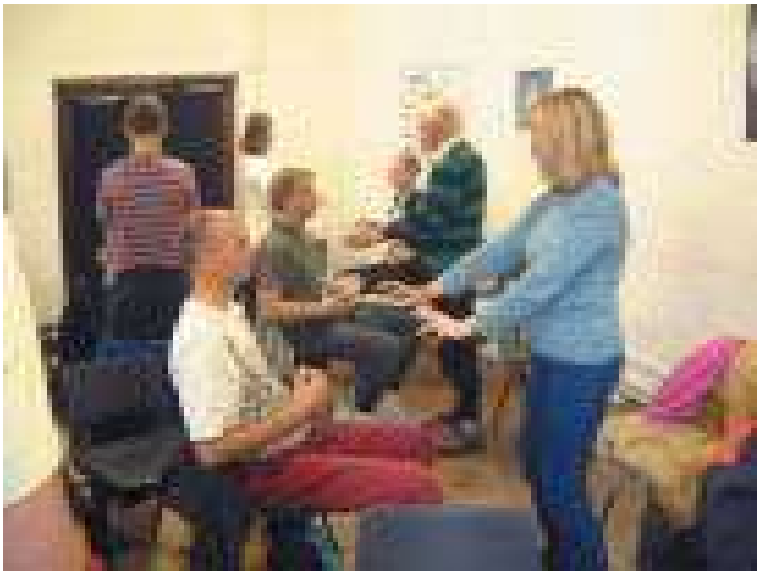

Fig 4.53 *Lead the person and guide them to follow you.*
Tell them to focus on the part of their body that you are working on.

1. Ask for the power to see into the cells of the brain. Ask the student to move their hands to cover their brain. “Brain cells listen to the command: Sickness go away. Clear, clean and bright; fill with violet light and return to normal function”.

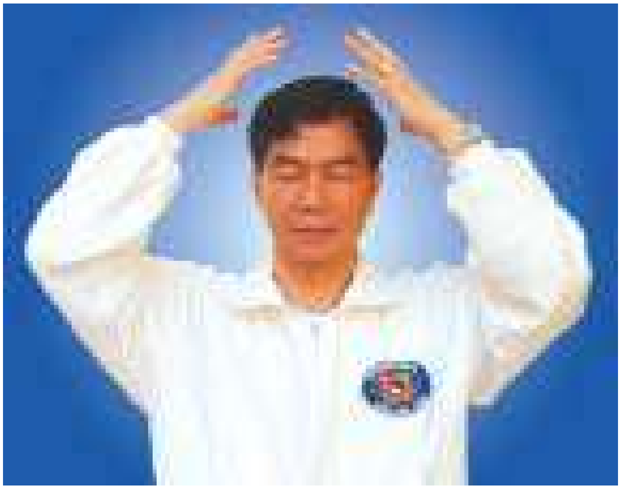

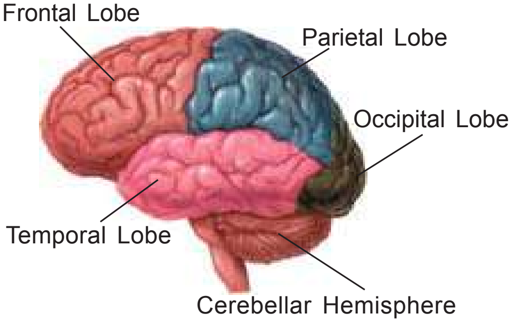
Frontal Lobe
Parietal Lobe
Occipital Lobe
Temporal Lobe
Cerebellar Hemisphere

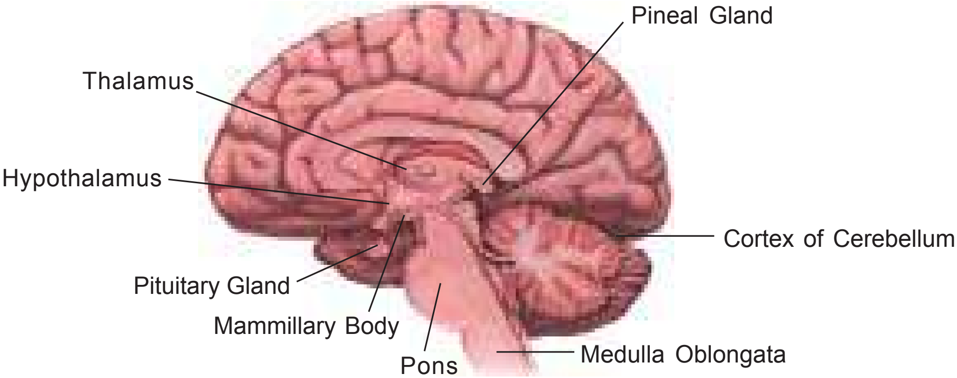
Pineal Gland
Thalamus
Hypothalamus
Cortex of Cerebellum
Pituitary Gland
Mammillary Body
Pons
Medulla Oblongata

Fig. 4.54 *“Clear, clean and bright; fill with violet light and return to normal functions”.*

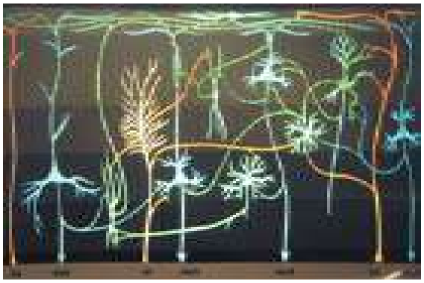

Brain Cells

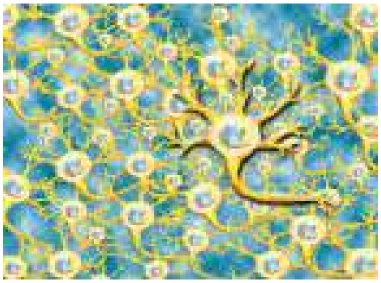

Healthy Happy Cells

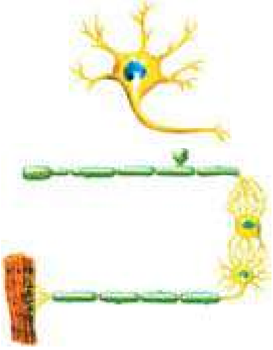

Fig. 4.55 *‘Intent’ to Heal.*

2. Ask for the power to see the cells of the sense organs. Ask the student to move their hand to cover their senses. Give the command: "Eyes, ears, nose, tongue and mouth cells listen to the command. Sickness go away. Clear, clean and bright; fill with violet light and return to normal function".

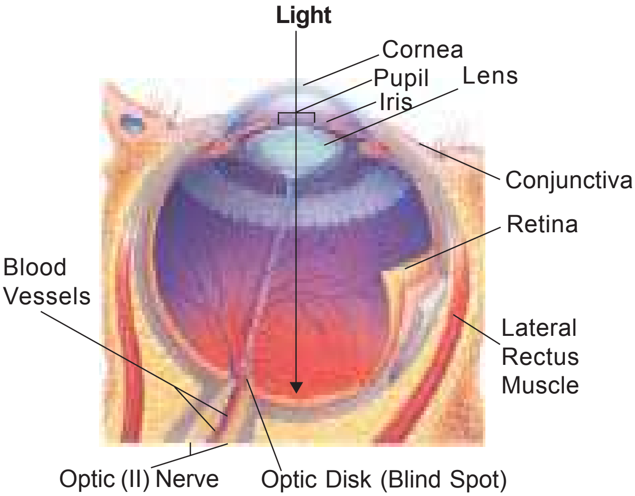
Light
Cornea
Pupil
Lens
Iris
Conjunctiva
Retina
Blood Vessels
Lateral Rectus Muscle
Optic (II) Nerve
Optic Disk (Blind Spot)

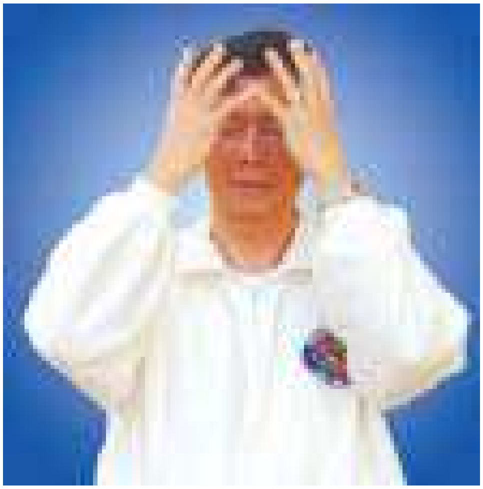

Fig. 4.56 *"Eyes, listen; clear, clean and bright; fill with Violet Light and return to normal function".*

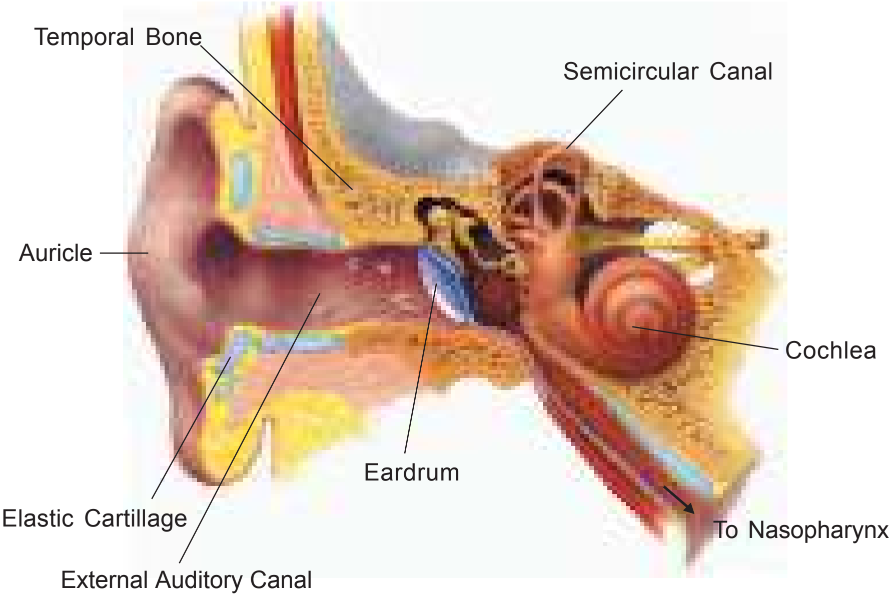
Temporal Bone
Semicircular Canal
Auricle
Cochlea
Eardrum
Elastic Cartillage
To Nasopharynx
External Auditory Canal

Fig. 4.57 *"Ears, listen; clear, clean and bright; fill with Violet Light and return to normal function".*

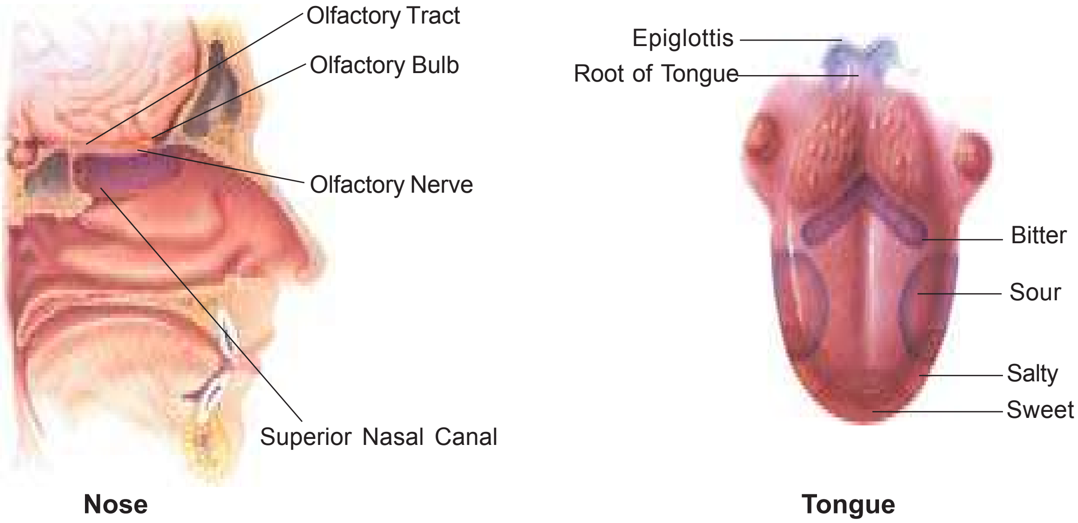
Olfactory Tract
Olfactory Bulb
Olfactory Nerve
Superior Nasal Canal
Epiglottis
Root of Tongue
Bitter
Sour
Salty
Sweet

Nose **Tongue**

Fig. 4.58
"Nose, tongue, listen; clear, clean and bright; fill with violet light and return to normal function"

3. Move your hands down to your throat center and give the command: "Thyroid, parathyroid and Thymus glands, listen to the command; Clear, clean and bright; fill with violet light and return to normal function".

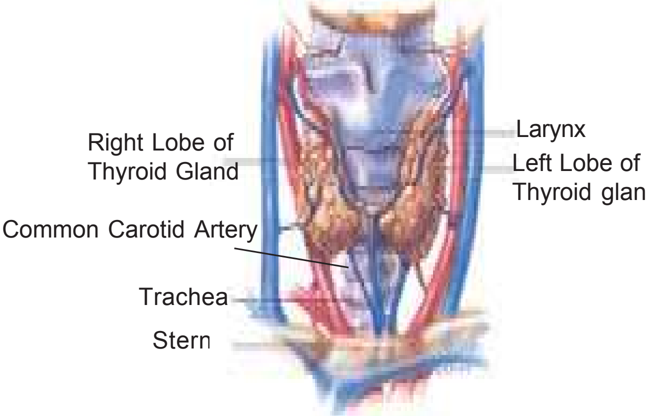
Right Lobe of Thyroid Gland
Larynx
Left Lobe of Thyroid gland
Common Carotid Artery
Trachea
Stern

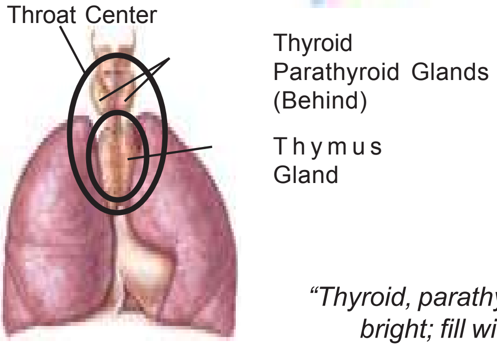
Throat Center
Thyroid Parathyroid Glands (Behind)
Thymus Gland

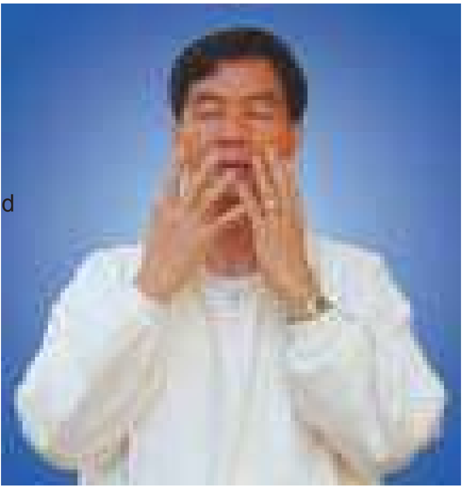

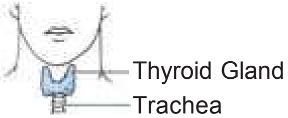
Thyroid Gland
Trachea

Fig. 4.59
"Thyroid, parathyroid and thymus glands - listen; clear, clean and bright; fill with violet light and return to normal function".

4. Move your hands to the heart and the lungs. Ask for the power to see the cells of the heart and lungs. Give them the command: “Lungs and heart cells, listen to the command”. If there is no sickness, just clean them out. See the cells and clean them out with violet light. If there is sickness, tell the sickness to “go away - be clear, clean and bright; fill with violet light and return to normal function”. Make sure that you say this part aloud. Give the command in a very firm tone of voice.

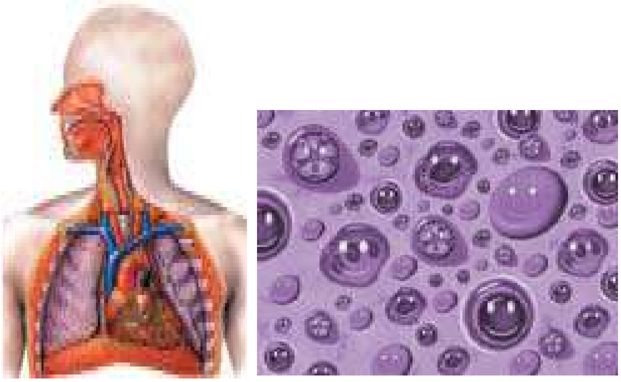

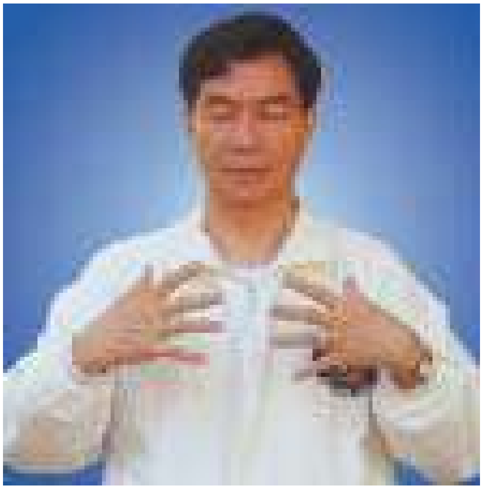

Fig. 4.60

“Heart and lung cells please listen; clear, clean & bright; fill with violet light and return to normal function”.

5. Move your hands to cover the left and right sides of the ribcage. Ask for the power to see the cells of the liver, gallbladder, spleen, pancreas and stomach. Give them the command: “Liver, gallbladder, spleen, pancreas and stomach cells, listen to the command”. If there is no sickness, just clean them out. If there is sickness, tell the sickness to, “go away - be clear, clean and bright; fill with violet light and return to normal function”. Make sure that you say this part aloud. Give the command in a very firm tone of voice.

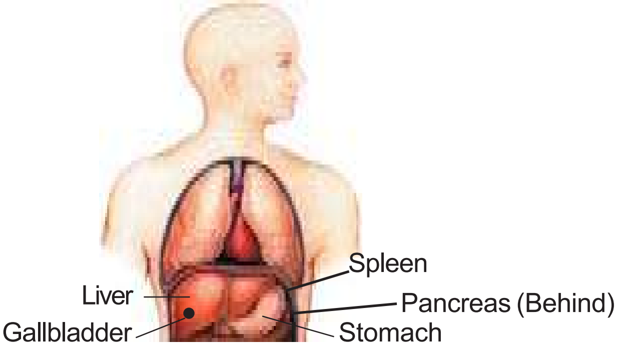

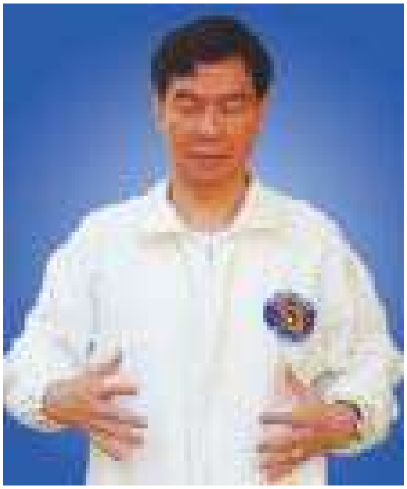

Fig. 4.61

“Liver, gallbladder, pancreas and spleen cells please listen; clear, clean and bright: fill with violet light and return to normal function”.

6. Move your hands to the large and small intestine on the front of the abdomen. Ask for the power to see the cells of the intestines. Give them the command: “Large and small intestine, listen to the command”. If there is no sickness, just clean them out. See the cells and clean them out with violet light. If there is sickness, tell the sickness to, “go away - be clear, clean and bright; fill with violet light and return to normal function”. Make sure that you say this part aloud. Give the command in a very firm tone of voice.

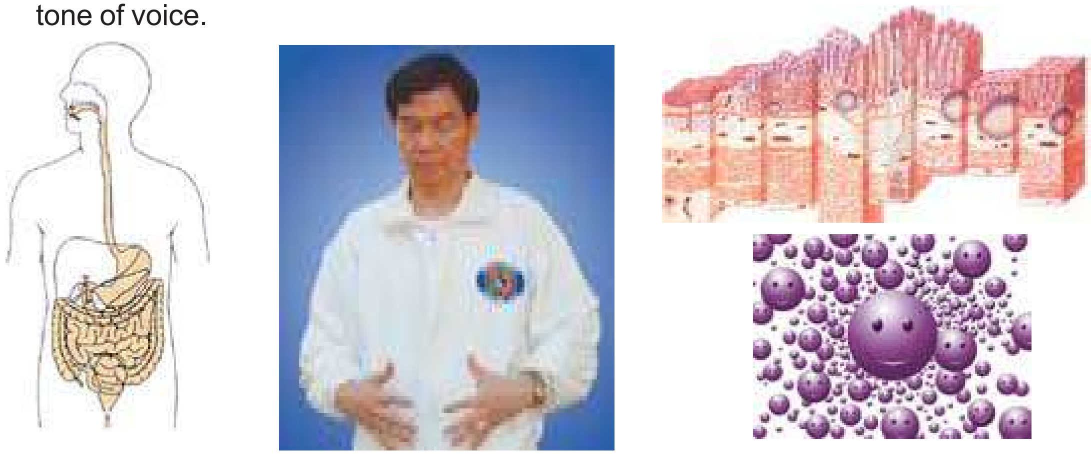

Intestines **Healthy cells**

Fig. 4.62 *“Small and large intestine cells listen to the command: clear, clean and bright; fill with violet light and return to normal function”.*

7. Move your hands to the kidney and sexual organs and ask for the power to see their cells. Give the command: “Kidney and sexual organ cells listen to the command: Clear, clean and bright; fill with violet light and return to normal function”. Gather the violet light again and let it pour down and fill all the cells. See the violet light again and feel and see that all the cells are, “clear, clean and bright”. Smile.

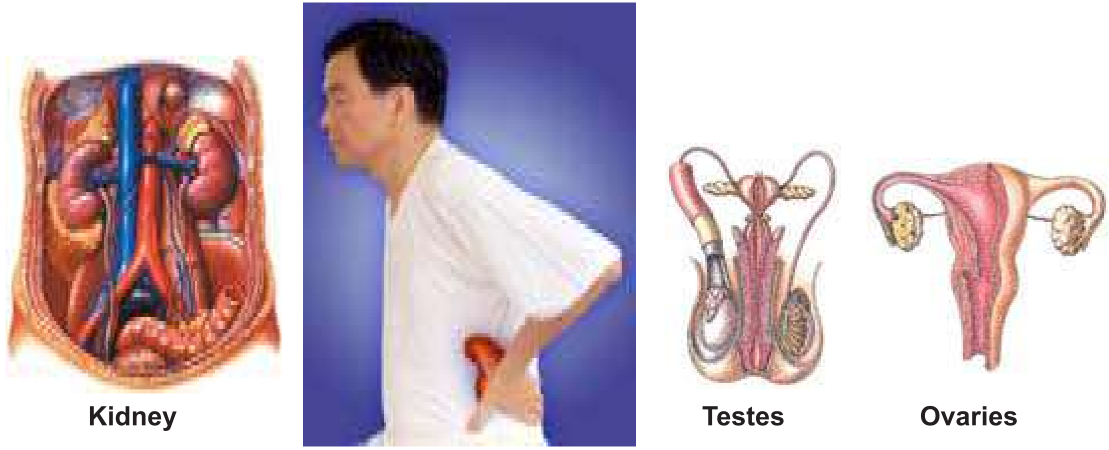

Kidney **Testes** **Ovaries**

Fig. 4.63 *“Kidneys and Sexual Organs listen”.*

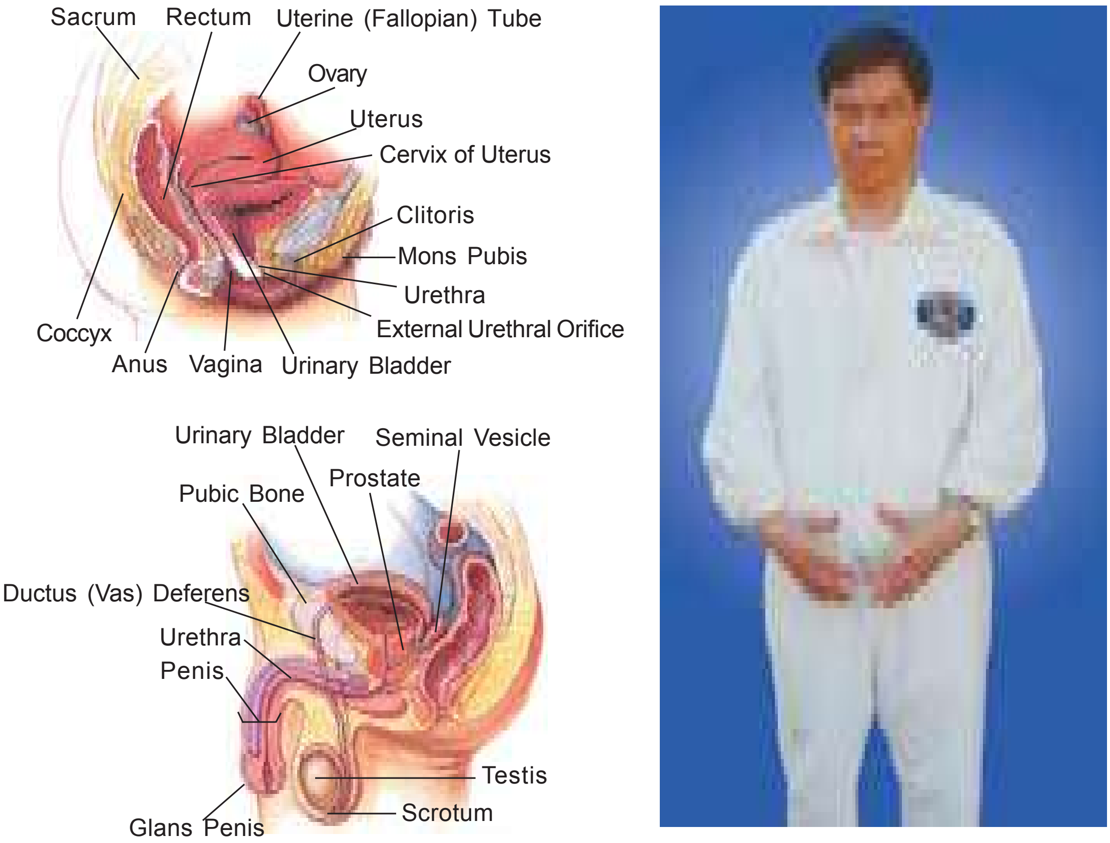
Sacrum
Rectum
Uterine (Fallopian) Tube
Ovary
Uterus
Cervix of Uterus
Clitoris
Mons Pubis
Urethra
External Urethral Orifice
Coccyx
Anus
Vagina
Urinary Bladder
Urinary Bladder
Seminal Vesicle
Prostate
Pubic Bone
Ductus (Vas) Deferens
Urethra
Penis
Testis
Scrotum
Glans Penis

Fig. 4.64 *Male and Female Sexual Organs*

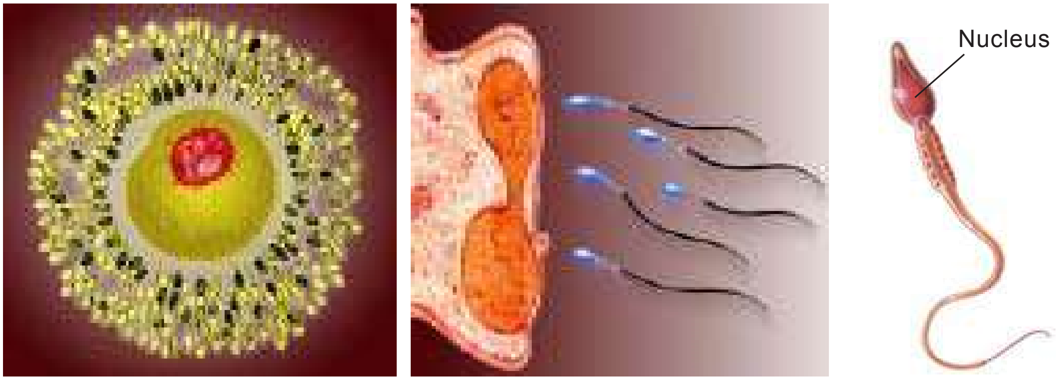

Egg

Sperm

Fig. 4.65 *"Clear, clean and bright: fill with Violet Light and return to normal function".*

K. Activate the Defense and Immune System

The Defense System relates to the white blood cell and the immune system relates to the Lymphatic System.

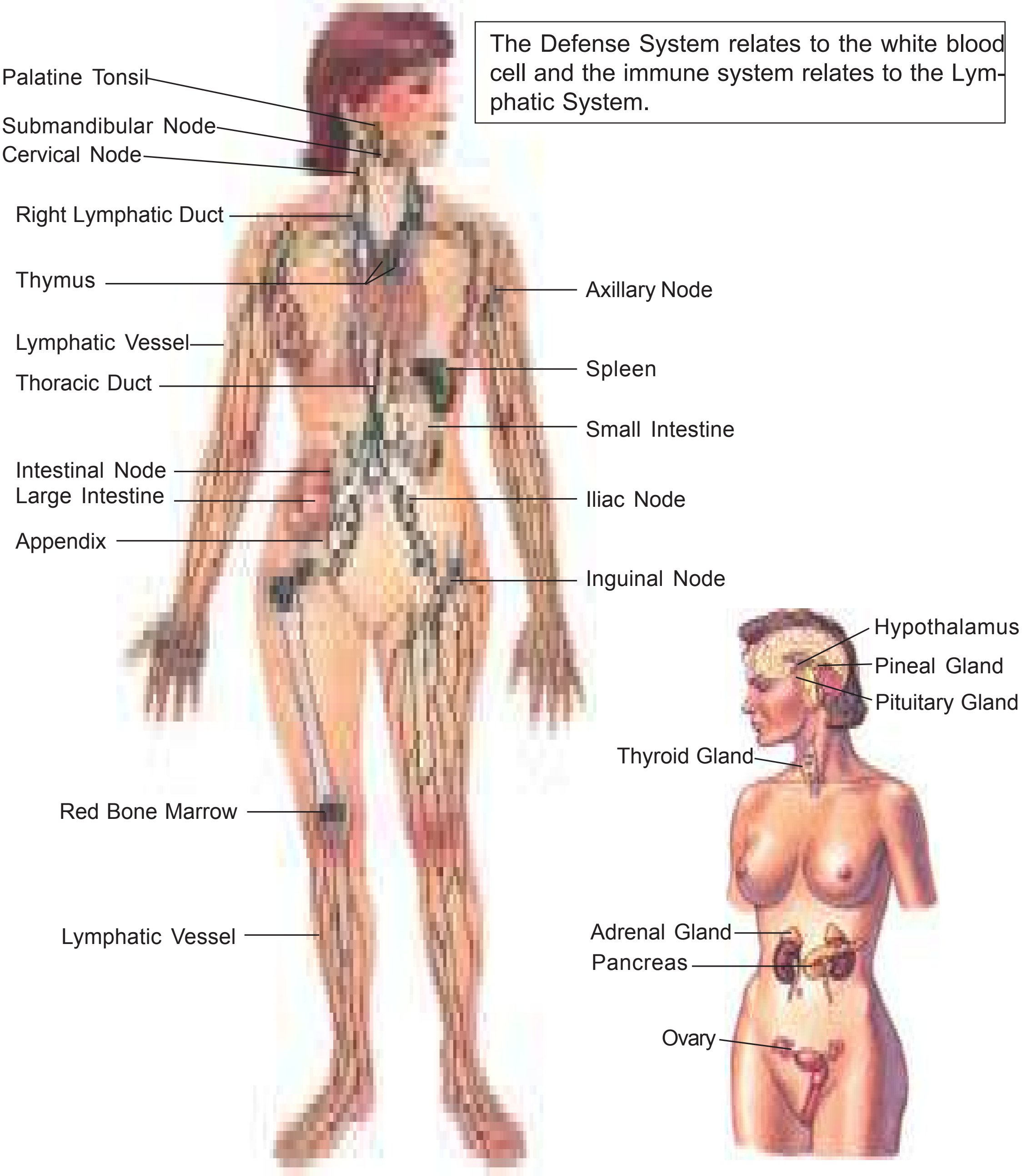
Palatine Tonsil
Submandibular Node
Cervical Node
Right Lymphatic Duct
Thymus
Axillary Node
Lymphatic Vessel
Spleen
Thoracic Duct
Small Intestine
Intestinal Node
Large Intestine
Iliac Node
Appendix
Inguinal Node
Hypothalamus
Pineal Gland
Pituitary Gland
Thyroid Gland
Red Bone Marrow
Lymphatic Vessel
Adrenal Gland
Pancreas
Ovary

Fig. 4.66 *Ask the cells to listen to the command.*
The bone marrow produces red and white blood cells. The lymphatic system cleans the toxic and polluted material and bacteria from the cells.

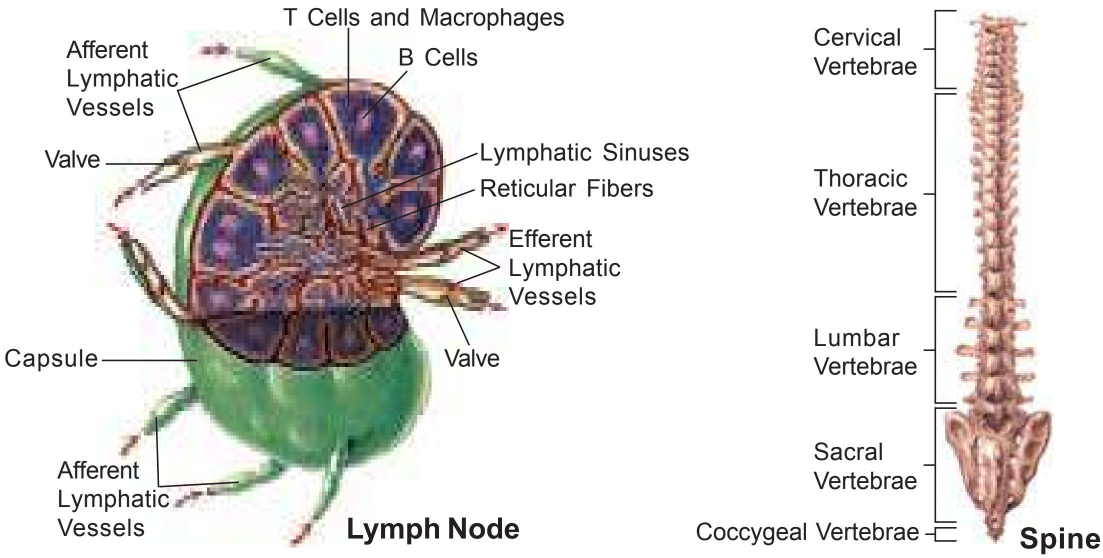
T Cells and Macrophages
Afferent Lymphatic Vessels
B Cells
Valve
Lymphatic Sinuses
Reticular Fibers
Efferent Lymphatic Vessels
Capsule
Valve
Afferent Lymphatic Vessels
Lymph Node
Cervical Vertebrae
Thoracic Vertebrae
Lumbar Vertebrae
Sacral Vertebrae
Coccygeal Vertebrae
Spine

***Fig. 4.67** “Immune Cells and Vertebrae please listen: clear, clean and bright; fill with Violet Light and return to normal function”.*

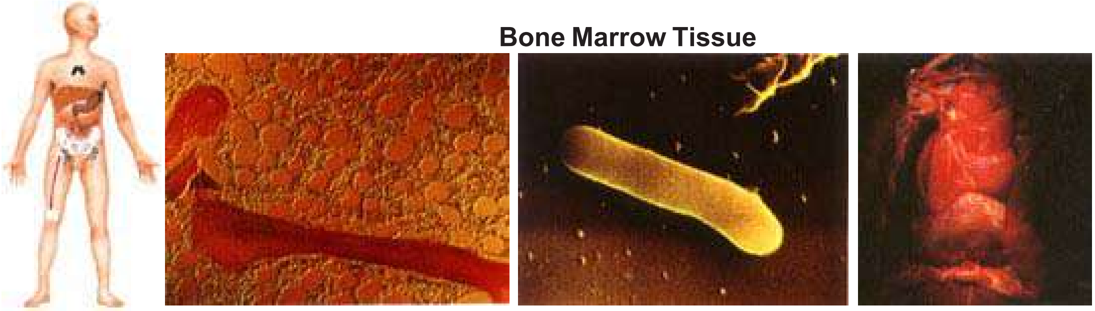
Bone Marrow Tissue

Sacrum: The sacrum controls all the bones and marrow. It effects the production of red and white blood cells. The bone marrow produces the lymphatic cells (white blood cells). When the blood is full of Chi it becomes lighter. Negative emotions make the blood thick and acidic. The blood cells in a healthy body can live up to ten times longer than in a weak or diseased one.

Bone Marrow: A tissue rich in fats, containing plenty of nourishment for the comprehensive production of the different blood cells: red cells for the transport of oxygen, blood platelets for clotting and the various white cells for the immune system.

Bacterium: Represents the foreign invader, but it is only one of many invaders. Parasites, fungi, viruses, chemicals, mineral fragments, metal particles—these and a great deal besides, set the immune system to work.

Thymus Gland: Perhaps the most important organ of the immune system. The vital training of the different T-lymphocytes takes place inside. When they pass through the thymus they receive a program to convert them into ‘T’ lymph cells. When they pass through the liver and large intestine they become ‘B’ cells and helper cells.

We have obtained Fig. 4.68 and the following text from the publication 'The Body Victorious' by Lennart Nilsson and Jan Lindberg.

We hope that you find this short selection of material as informative as we did.

"The illustration below is a highly schematic and simplified picture of our immune system. *Top left*, the pale yellow 'house' symbolizes the bone marrow, where all blood cells are born. *Bottom Right*, a solitary bacterium (yellow) represents the foreign invader that the immune system needs to combat. The light blue path represents the older defenses, with various feeding cells that tackle all foreign substances and also function as a kind of cleaning system. They deal with all the old, dead and used-up material in the body. The newer, special defense force, which develops later, follows the green path. Its cells are more specialized; they obtain their special training and reach maturity in organs such as the thymus gland and in the lymphoid tissue around the intestines and in the liver. In the lymphoid tissue, the yellow 'half-way house' (far right), the 'B-lymphocytes are trained. These are the precursors of the large plasma cells - the pale green cells (far right) in the illustration, which produce the body's sniper ammunition, the antibodies (Red and Y-Shaped).

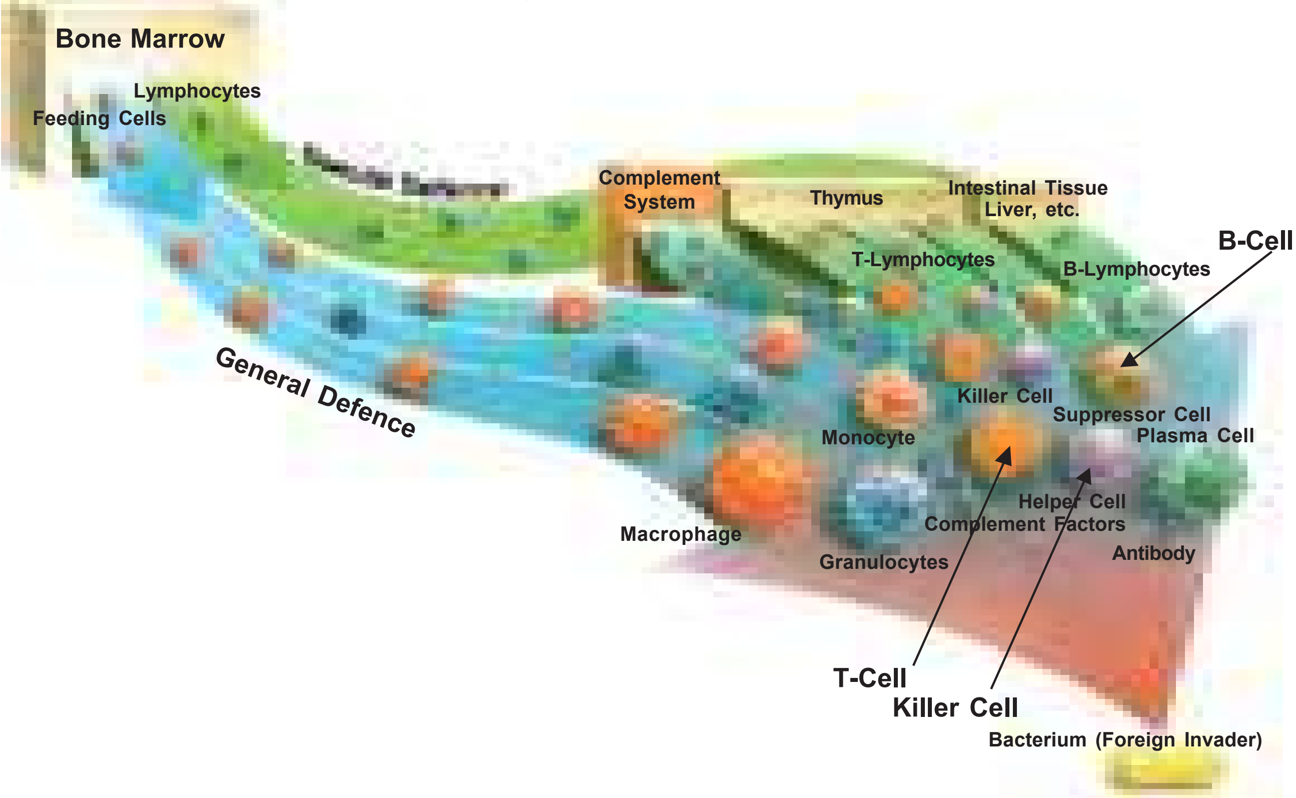
Bone Marrow
Lymphocytes
Feeding Cells
Complement System
Thymus
Intestinal Tissue Liver, etc.
B-Cell
T-Lymphocytes
B-Lymphocytes
General Defence
Killer Cell
Suppressor Cell
Plasma Cell
Monocyte
Helper Cell
Complement Factors
Macrophage
Antibody
Granulocytes
T-Cell
Killer Cell
Bacterium (Foreign Invader)

Fig. 4.68

Bone marrow, bacterium and thymus please listen: clear, clean and bright; fill with violet light and return to normal function".

The three paths from the thymus are intended to show that there are different types of T- lymphocytes - among others, aggressive killer cells, helper cells and suppressive cells. All of them have specialized tasks to perform when the immune system launches a counterattack. The three blue paths of the older defenses have three different kinds of feeding cells. First, the large and powerful macrophages (orange) advance on the enemy; then come the granulocytes (blue), smaller and faster moving, followed by the (pink) monocytes. The nuclei in the shimmering cell bodies are faintly discernable. In addition, there is an important complement system (middle), symbolized here by a multicolored range of small spheres flocking toward the foreign bacterium. These molecules play a large part in increasing the efficiency of both antibodies and feeding cells. In addition, they have the capacity to destroy bacteria by shooting holes in them. The complement factors are produced in many different cells in the body.

Chi Kung has a very strong effect against viruses. Altough medical science is struggling to discover a medicine that will kill them, Chi Kung practice is able to prevent them from occuring in the first place. Whether we are sick or healthy, we all need our defense system to be in good shape.

Defense System - Outline of Practice

1. Work on yourself first to activate your sacrum. Project the sacrum as big as the universe. See the eight holes in the bone and visualize them breathing. Breathe and compress the Universal Chi into the bones. Feel the Chi begin to rise up the spine and into your mideyebrow and temple bones. When you feel that there is enough Chi, project it out to the universe, multiply it and spiral it back down. Project it towards the student and ask them to touch their sacrum. Rub your hands until warm and touch the sacrum. Picture the sacrum getting bigger and bigger, breathing into the eight holes. Tell the person to smile at the sacrum and start to work on your Tan Tien and universe. Multiply more energy and picture the Chi from the universe coming down to the sacrum of the other person (and your own) and then charge up the spine, to the rib cage, to the temple bone and to the mideyebrow.

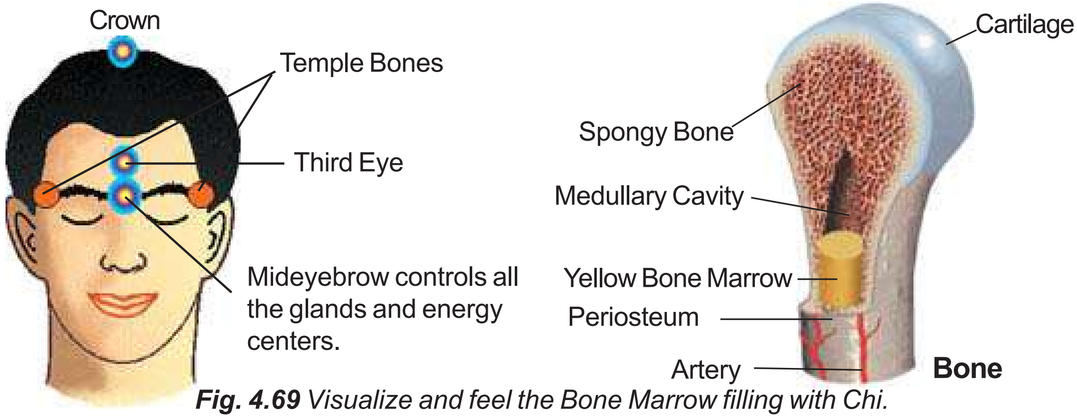

Fig. 4.69 *Visualize and feel the Bone Marrow filling with Chi.*

2. Tell the student to move their hands to the hip bones. Tell them to smile to these bones. Feel them as 'funny, laughing, happy bones'. Tell them to feel the electricity in the sacrum and bone marrow.
3. Touch the femur bones. Feel the electricity running up the legs.
4. Touch the humerus bones. Feel the electricity running up the arms.
5. Touch the lower part of the sternum and feel the Chi penetrate into the bones and spread out to the rib cage. Feel the electricity run throughout the rib cage.

Practice

1. Touch the sacrum, feel your fingers grown 'long' and penetrate into the marrow. Feel that the sacrum is as big as the universe and that the Chi has been compacted. Fill the sacrum with Chi. The eight holes are breathing. Let the fingers touch the sacrum and become aware of the Tan Tien spiraling, until you feel the crown and the mideyebrow spiral and connect to the universe. Hold until the universe fills the sacrum once again and rises up the spine to the mideyebrow.

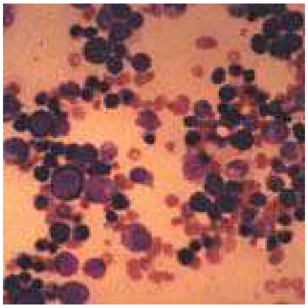

Bone Marrow Cells

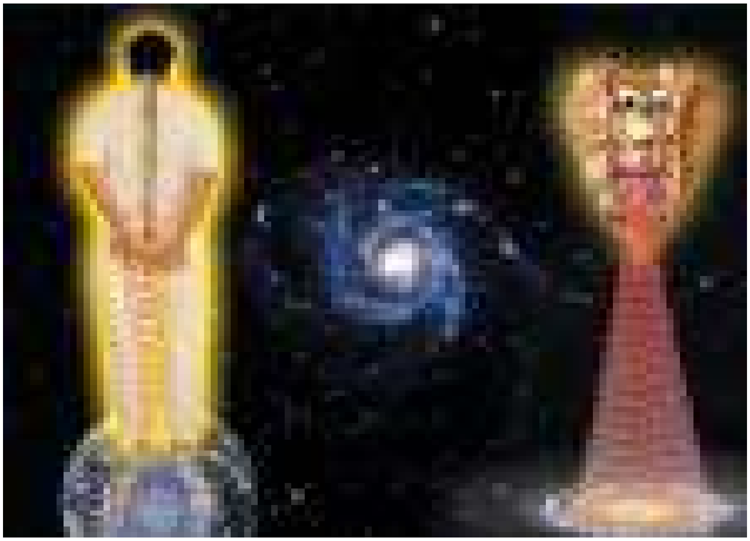

Fig. 4.70 *Touch the Sacrum - see the Eight Holes breathing - fill the Sacrum, Mideyebrow and Crown with Chi - empty to the universe and 'fill' again.*

Sacrum Breathing to get Chi

Breathe through the sacrum; feel red and yellow light enter. Guide it into Tan Tien. Blend the Chi in the Tan Tien. Move up to the Heart Center. Move up to the Crown and project up to the universe; let it blend and multiply. Multiply in the universe

Open a direct channel down to the other person, and one to your palm.

Sacrum Chi

Project the student's feet down to the earth and connect to the earth's core.

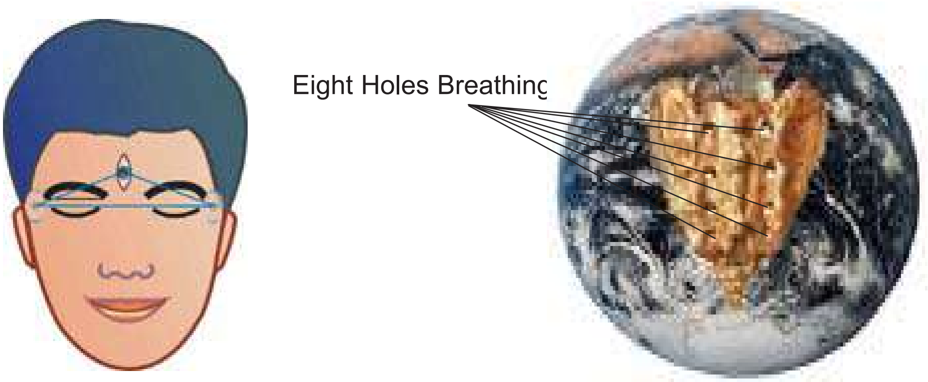

Third Eye and Temple Bones are aligned in a triangle.

Activate the Sacrum until you feel the Chi rising up the spine and feel it pulsing; be aware of the Temple Bones.

Fig. 4.71 *Activate the Sacrum - Be aware of the Temples.*

2. Tell the student to move their hands to the hip bones. Tell them to smile to these bones. Feel them as 'funny, laughing, happy bones'. Tell them to feel the electricity in the sacrum and bone marrow. Tell the student to pack and compress the Chi into the bones in order to revitalize and revive bone marrow. Give the command. "Produce healthy white and red blood cells".

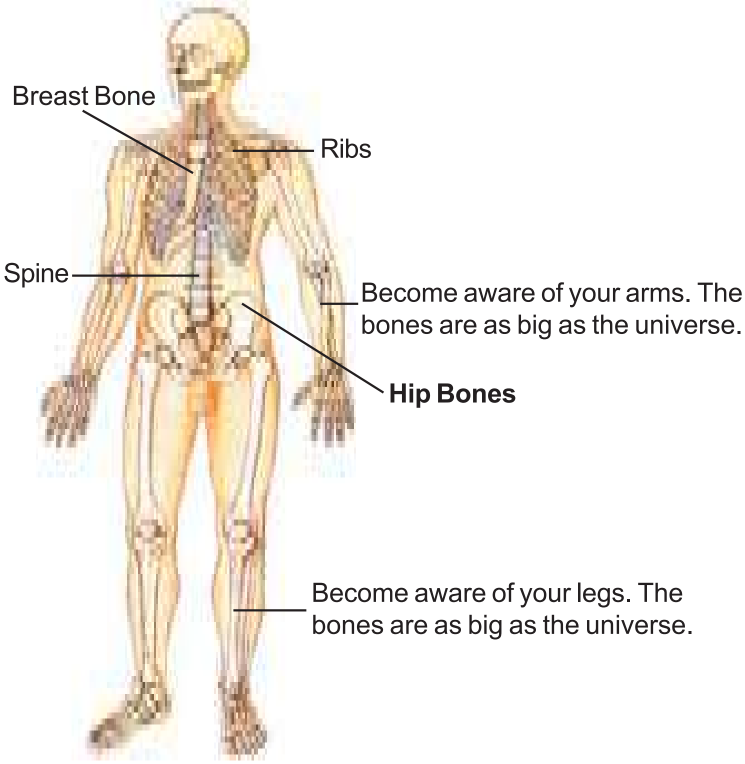

Become aware of your arms. The bones are as big as the universe.

Become aware of your legs. The bones are as big as the universe.

Fig. 4.72 *"Hip Bones produce healthy White and Red Blood Cells".*

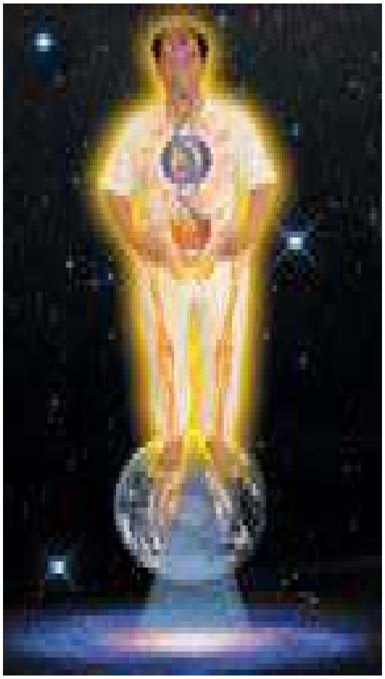

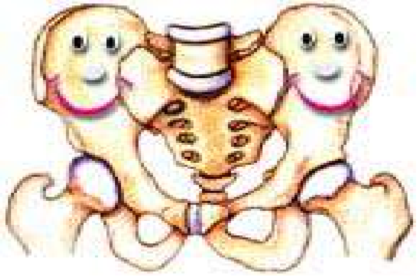

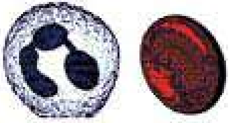

White Blood Cell
(Neutrophil)

Leave the fingers touching the hip bones and start to spiral in the Tan Tien until you feel the crown and the mideyebrow also spiral. Expand your awareness to the Universe and the Universe will charge into the hip bones.

Fig. 4.73 *Touch and feel the bones. Leave your fingers there - lower your mind into the Tan Tien and the universe.*

3. Touch the middle part of the femur bone to help increase the production of healthy red blood cells. Feel the electricity running up the legs. Give the same command for the femur bones: "Funny, laughing, happy bones produce healthy white and red cells." After that put your mind into the Tan Tien and universe so that the Chi can charge into the bones. Feel electricity run throughout the whole skeleton.

Fig. 4.74 *"Funny, laughing, happy femur bones produce healthy white and red blood cells".*

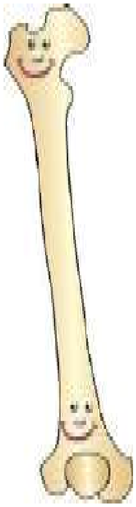

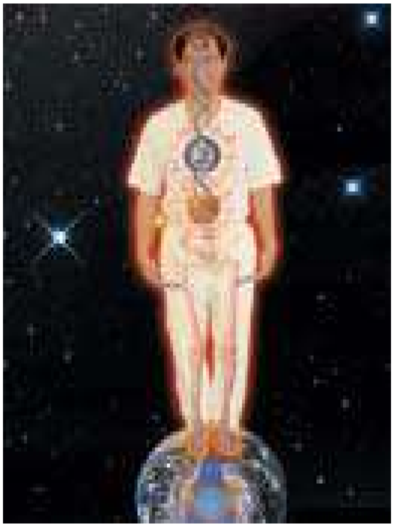

4. Touch the humerus bones. Feel the electricity running up the arms.Give the same command: "Funny, laughing, happy bones produce healthy white and red cells." After that, put your mind into the Tan Tien and universe so that the Chi can charge into the bones. Give the same command for the lower part of the sternum. Chi penetrates into the bone and spreads out into the rib cage 'Tan Tien and Universe'. Put your mind into the Tan Tien and the universe so that the Chi can charge into the bones.

5. Touch the lower part of the sternum and feel it and the rib cage as funny, happy, laughing bones. Feel the Chi and electricity penetrate into the bones and spread out to the rib cage. Touch the lower part of the sternum. Give the same command again. “Produce healthy white and red cells”. After that you put your mind into the Tan Tien and universe so that the Chi can charge into the bones.

You have finished activating the defense system.

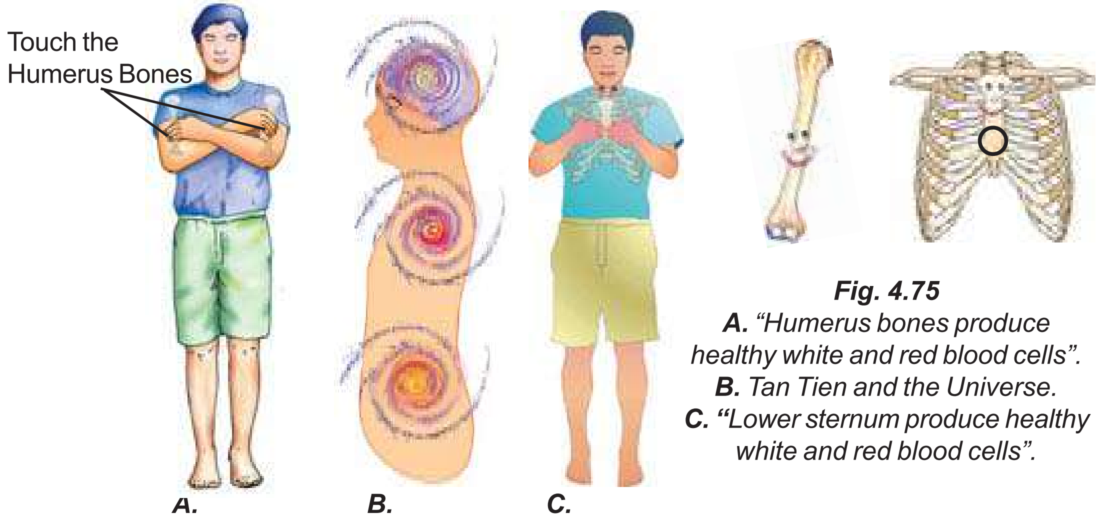

Fig. 4.75
A. *“Humerus bones produce healthy white and red blood cells”.*
B. *Tan Tien and the Universe.*
C. “*Lower sternum produce healthy white and red blood cells”.*

Immune System

1. The Immune System starts with the top part of the sternum. Touch the top part of the sternum. Project your fingers ‘very long’ into the bone and feel them penetrate right into the thymus. Activate the thymus gland. Gradually, feel the fingers rising up to the thyroid and parathyroid. Activate the thyroid and parathyroid and throat center. Give the command, “Clear ,clean and bright fill with violet light and return to normal function”.

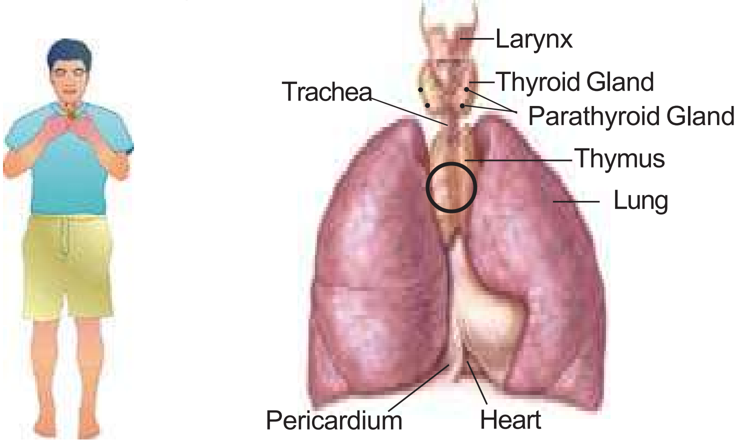

Fig. 4.76 *Touch the Top of the Sternum & activate the Thymus, Thyroid and Parathyroid Glands.*

2. Move the fingers to touch both sides of the jaw bones to help activate the tonsils, the body's first line of defense. The jaw bone affects the lymph nodes beneath, which include the tonsils. Fill them with Chi. Give the command, "Clear, clean and bright; fill with violet light and return to normal function".

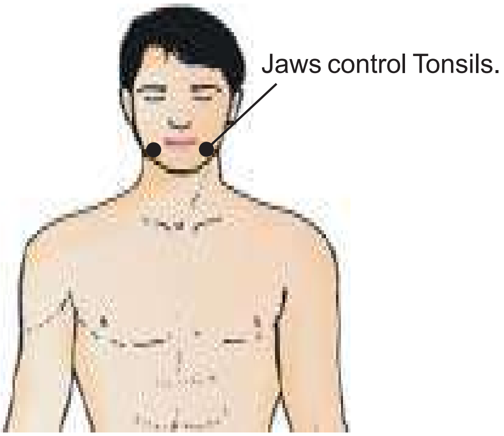
Jaws control Tonsils.

Fig. 4.77 *Activate the Tonsils and Jaw Bones.*

3. Touch the mideyebrow and crown, which control the pituitary and pineal glands respectively. Touch the third eye, feeling it open and feel the light from heaven entering into your brain and into your body. Touch the crown. Feel your fingers reaching deep inside and feel the Chi penetrating all the way down to your perineum. Give the command, "Clear, clean and bright fill with violet light and return to normal function".

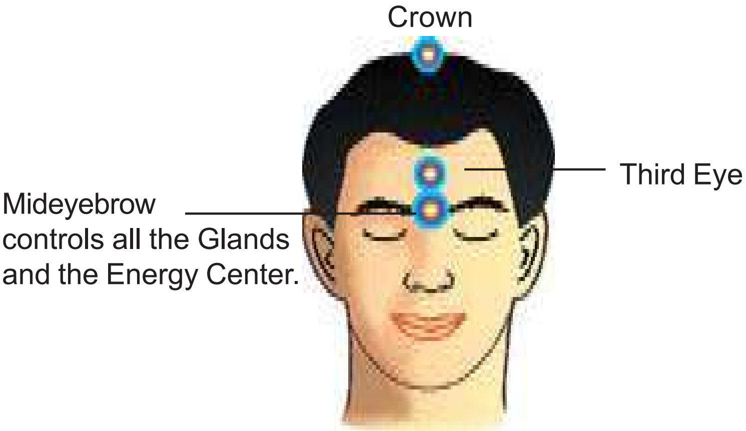

Fig. 4.78 *Third Eye, Mideyebrow and the Crown affects the Immunity and Defense Systems.*

4. Rub both of your hands until they are warm and cross the hands on the neck and feel the lymphatic system has been activated. Cross your arms and hold your hands

under your armpits. Picture the lymphatic system of the armpits as you activate the lymph nodes. Give the command; “Clear, clean and bright fill with violet light and return to normal function”. The lymphatic system of the upper body is clear, clean and bright and the lymphatic fluid flows! Smile.

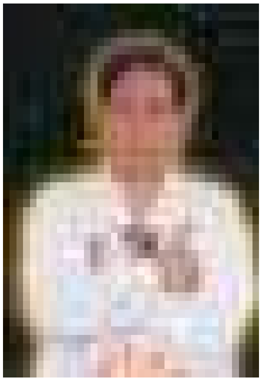

Fig. 4.79 *Lymphatic System of the Neck and Armpits are clear, clean and bright.*

5. Move the hands to the cover the groin area. Give the command to the lymph nodes there. Next, cover the lymph nodes of the navel area, and give the command; “Clear, clean and bright fill with violet light and return to normal function”. The groin and the navel area lymph nodes become clear, clean and bright. The groin and the navel area lymph nodes become clear, clean and bright.

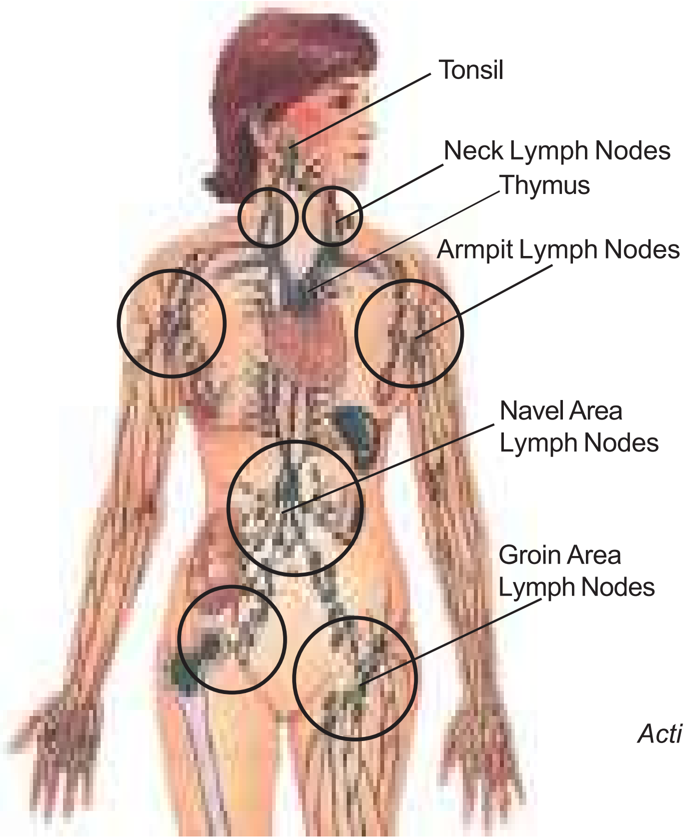
Tonsil
Neck Lymph Nodes
Thymus
Armpit Lymph Nodes
Navel Area Lymph Nodes
Groin Area Lymph Nodes

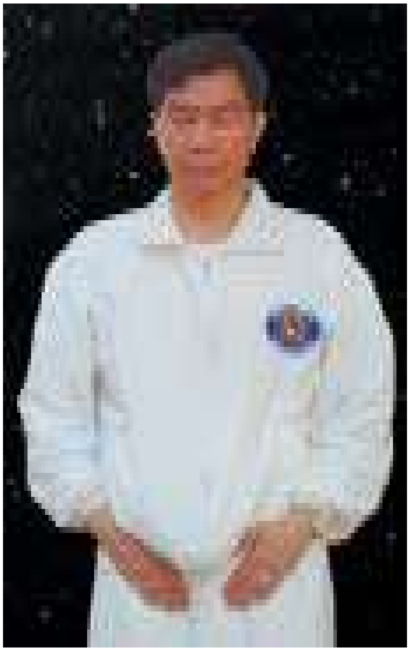

Fig. 4.80
Activate the Lymph Nodes in the Groin.

6. Place your palms over your navel and activate the lymph nodes. Give the command; "Clear, clean and bright fill with violet light and return to normal function". See them become clear, clean and bright. You have now activated the immune system.

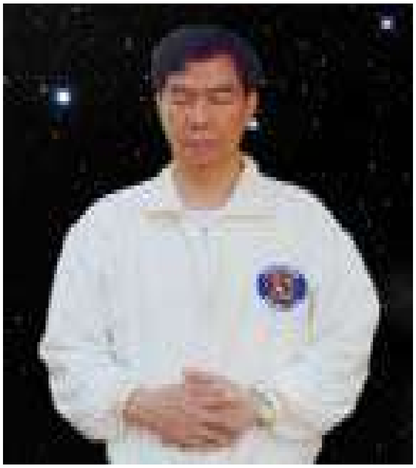

Fig. 4.81
Touch and activate the Lymph Nodes in the Navel.

Summary - General Healing Session

This General Healing Session will help everybody.

A. Three Minds into One
B. Activate the Six Directions
C. Activate the Three Fires
D. Linking Personal Stars, Energy Bodies and the Universe
E. Protective Circle and Chi Field
F. Chi Water (Sacred Water) Practice
G. Empty and Fill
H. Clean with Green Light
I. Clean with Blue Light
J. Charge with Violet Light
K. Defense and Immune System

A. Three Minds into One

Smile to the inner universe, activate the heart compassion energy; feel the love and joy within. Smile and empty the mind and heart down into the Tan Tien. Fill the Tan Tien with Chi and spiral the energy. When the Tan Tien is full, the energy will rise up to fill the brain. Manifestation will occur at the mideyebrow.

B. Activate the Six Directions

Expand the awareness to connect to the Six Directions at the crown, mideyebrow, heart and Tan Tien.

C. Activate the Three Fires

Activate the Tan Tien, Kidney and Heart Fires.

Activate the Sacred Fire Triangle circulating the energy between the Tan Tien, kidneys and the heart. Feel the energy's force multiply.

Activate the personal stars; the cosmic star above the head and the earth star below the soles of the feet.

Keep the Tan Tien warm, and spiral at the sacrum and mideyebrow. Turn your eyes to look upwards and see the crown spiraling, connecting to the cosmic star about six inches above the crown point. Feel the cosmic and earth stars pull the body simultaneously towards heaven and earth. See the cup of the Big Dipper fill with Violet Light, gathering Chi from the North Star and the universe. When the cup is full, pour the light down into the cosmic star above your head. Feel the violet light spreading throughout your entire system, filling the Upper, Middle and Lower Tan Tiens.

D. Linking Personal Stars, Energy Bodies and the Universe

Feel the central channel running from your cosmic star, through your three Tan Tiens and connecting to the earth energy below. Use you 'Yi' power to connect to the personal or cosmic star above the head. Connect the personal star to the stars of each other person in the meditation. Create a Chi body and become the major connecting point. Create a left spiral pattern to connect all the stars into a larger group. All the meditators then become a satellite, creating a group energy body.

E. Protective Circle and Chi Field

Raise the hands and touch the Universal Fire. Use your Yi to bring the energy down and manifest a Sacred or Chi Fire around the room, creating a protective circle. Create a Chi Dome and set up the guardian animals and/or angels. Activate the eight elemental forces of nature and the universe.

F. Chi Water (Sacred Water) Practice

Invoke the power of the Sacred Water practice to cleanse the body of all sickness, negative energy and bad fortune. Energize and renew the self at a cellular level with health, wealth and longevity.

G. Empty and Fill

Empty any remaining negative energy, sickness, burdens and worries, down in the depths of the molecular mother, the earth. Bury the energy there so that the earth can recycle it. Bring recycled energy up from the center of the earth into the body. Fill and compact the Chi into the 'hollow bones'. See all of the cells on the horizon of your mind's eye; spread them across the panorama and see the cells as big as the stars in the night sky. Fill them with Chi, blending it with any sick or grey energy. Then empty again, as you guide the all of the Chi back into the earth for recycling. If you wish, you can fill with earth Chi once again before you begin to 'clean with green'.

H. Clean with Green Light

Green cleans and detoxifies all of the body's cells**.** Green Chi is mild and safe. Green Chi is used for cleansing, and as a decongestant for diseased parts. Be aware of the liver's green color, the throat and the emerald forest. Breathe into the throat, feel yourself breathing in the green light. Guide it down to the Tan Tien, blend and move it up to the heart and the crown.

Project it to the universe; spiral and let it multiply. Channel one direct line down to the student, one to your crown and one to your palm. Bring the energy back down and blend it with any sickness in the liver, gallbladder or anywhere else in the body. Spiral the energy counter clockwise and down into the earth. Continue cleaning with green until you see that all the cells have been charged with the emerald light of the forest.

I. Clean with Blue Light

Be aware of the kidneys' blue sapphire color and breathe the blue light into your throat. Guide it down to the Tan Tien, blend and move back up to the 'heart of compassion' and then to the crown.

Project up to the universe, spiral and let it multiply. Channel one direct line down to the student, one to your crown and one to your palm. Then feel the universe pouring the blue light back down; clockwise. Spiral down. Look at the cells and see the blue light, like water, go into the cells to rinse, clean and remove any sickness. Spiral the energy counter clockwise and down into the earth. Continue rinsing with blue light until you see that all the cells have been charged with the sapphire blue.

J. Charge with Violet Light

Luminous Violet has intelligence and can be programmed.

Turn three minds into one mind, go deep into the empty space of the cell and deep into the chromosomes, magnify the space as you enter into the DNA and return to the original cells. Ask for the power to see the cells and look at the brain cells spread out into the space. "Brain cells please listen. Clear, Clean, and Bright: fill with violet light, back to normal function". See violet light. Be aware your crown. The Tan Tien Chi and the negative Chi all combine in the crown. Extend a violet light up to the universe. See the North Star and the Big Dipper. Tell the students to hold their arms in the air, so that they can funnel the Chi into their personal star. Bring the North Star and the Big Dipper down; hold the handle of the Big Dipper and pour the violet light over the student's personal star. The higher cells will be programmed. Let the Chi enter the brain and ask for the power to see the brain cells. "Brain cells listen to the command". If there is sickness in the brain, give the command. "Sickness go away". Continue to spiral the violet light down through the cells of the whole body.

1. Ask for the power to see into the cells of the **Brain.** Ask the student to move their hands to cover their brain. "**Brain Cells** listen to the command;"Sickness go away. Clear, clean and bright fill with violet light and return to normal functions".

2. Ask for the power to see the cells of the sense organs. Ask the student to move their hand to cover their senses. Give the command; **"Eyes, Ears, Nose, Tongue and Mouth** cells listen to the command. Sickness go away. Clear, clean and bright fill with violet light and return to normal function".
3. Move your hands down to your Throat Center and give the command; **"Thyroid, Parathyroid and Thymus Glands** listen to the command; Clear, clean and bright fill with violet light and return to normal function".
4. Move your hands to the heart and the lungs. Ask for the power to see the cells of the heart and lungs. Give them the command; **"Lungs and Heart Cells** listen to the command; Clear, clean and bright; Fill with violet light and return to normal function".
5. Move your hands to cover the left and right sides of the ribcage. Ask for the power to see the cells of the liver, gallbladder, spleen, pancreas and stomach. Give them the command; **"Liver, Gallbladder, Spleen, Pancreas and Stomach Cells** listen to the command; Clear, clean and bright fill with violet light and return to normal function".
6. Move your hands to the large and small intestine on the front of the abdomen. Ask for the power to see the cells of the intestines. Give them the command; **"Large and Small Intestine**, listen to the comm and Clear, clean and bright fill with violet light and return to normal function".
7. Move your hands to the kidney and sexual organs and ask for the power to see their cells. Give the command; **"Kidney and Sexual Organ Cells** listen to the command; Clear, clean and bright fill with violet light and return to normal function". Gather the violet light again and let it pour down and fill all the cells. See the violet light again and feel and see that all the cells are "clear, clean and bright". Smile.

K. Defense and Immune System

Defense System

1. Work on yourself first to activate your **Sacrum**. Project the sacrum as big as the universe. Breathe and compress the universal Chi into the sacrum until it rises up to fill your brain. Project out into the universe, multiply the energy and spiral it back down. Project it to the students and tell them to touch the sacrum.
2. Tell the students to move their hands to the **Hip Bones**. Tell them to smile to these bones. Feel them as 'funny, laughing, happy bones'. Tell them to feel the electricity in the sacrum and bone marrow.
3. Touch the **Femur Bones.** Feel the electricity running up the legs.
4. Touch the **Humerus Bones.** Feel the electricity running up the arms.
5. Touch the lower part of the **Sternum** and feel the Chi penetrate into the bones and spread out to the **Rib Cage**. Feel the electricity run throughout the rib cage.

Immune System

1. The Immune system starts with the top part of the **Sternum**. Touch the top part of the sternum. Project your fingers ‘very long’ into the bone and feel them penetrate right into the **Thymus**. Activate the thymus gland. Gradually, feel the fingers rising up to the thyroid and parathyroid. Activate the **Thyroid** and **Parathyroid** and **Throat Center**. Give the command; “Clear, clean and bright fill with violet light and return to normal function”.
2. Move the fingers to touch both sides of the **Jaw Bones** to help activate the **Tonsils,** the body’s first line of defense. The jaw bone affects the **Lymph Nodes** beneath, which includes the tonsils. Fill them with Chi. Give the command; “Clear clean and bright fill with violet light and return to normal function”.
3. Touch the **Mideyebrow** and **Crown**, which control the **Pituitary** and **Pineal Glands** respectively. Touch the third eye, feeling it open and feel the light from heaven entering into your brain and into your body. Touch the crown. Feel your fingers reaching deep inside and feel the Chi penetrating all the way down to your perineum. Fill the glands with Chi. Give the command; “Clear, clean and bright fill with violet light and return to normal function”.
4. Rub both of your hands until they are warm. Cross your arms and hold your hands under your **Armpits**. Picture the **Lymphatic System** of the armpits as you activate the lymph nodes. Give the command; “Clear, clean and bright fill with violet light and return to normal function”. The lymphatic system of the upper body is clear, clean and bright then the lymphatic fluid flows! Smile.
5. Move the hands to the cover the **Groin** area. Give the command to the **Lymph Nodes** there. Give the command; “Clear, clean and bright fill with violet light and return to normal function”.
6. Place your palms over your **Navel** and activate the **Lymph Nodes**. Give the command to the lymph nodes there. Give the command; “Clear, clean and bright fill with violet light and return to normal function”. **You have now activated the immune system.**

Chapter V
Cosmic Healing I - Basic

Source of Chi and Color

There are different sources of Chi in the universe: cosmic (air), earth, nature and human. Different parts of the human body have different energies and are correlated with the universe and nature. Each part of our body, especially the organs and glands can produce, receive, transform and emit different Chi and color.

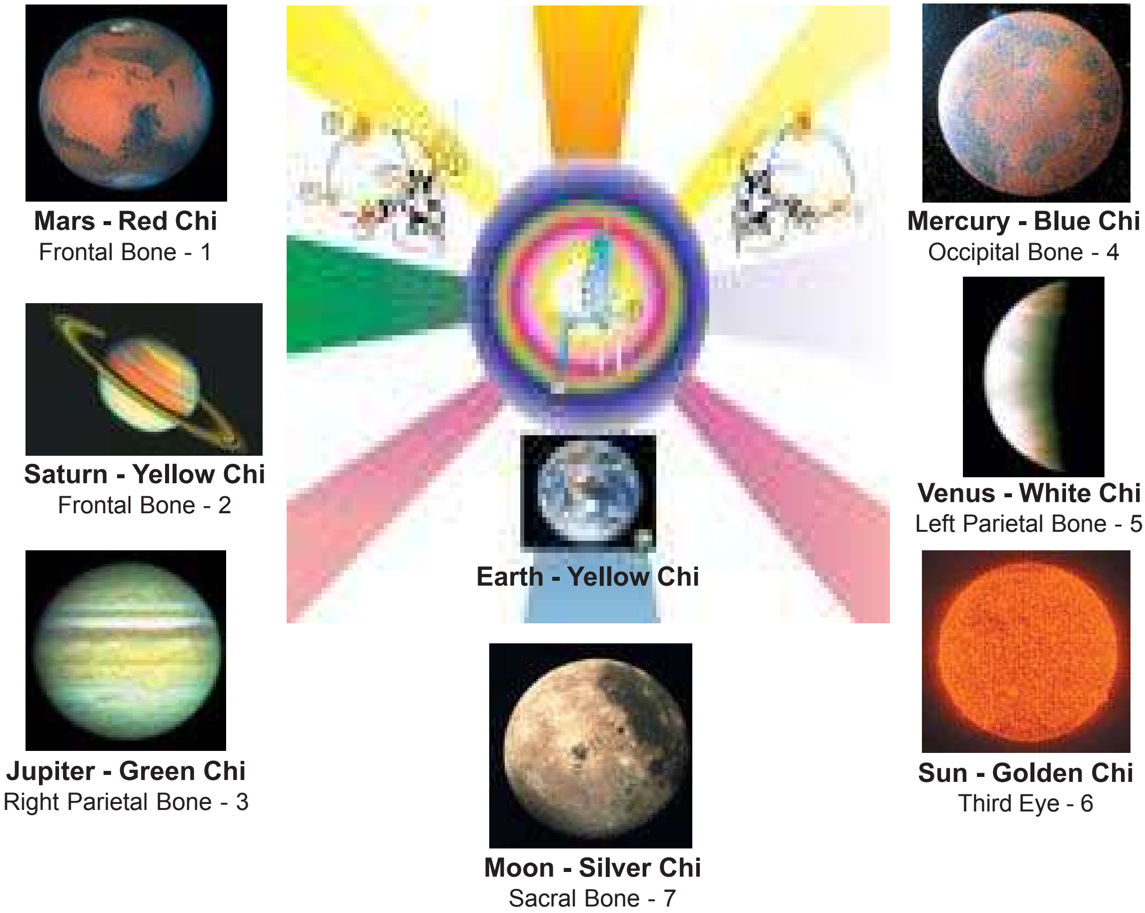

Mars - Red Chi
Frontal Bone - 1

Saturn - Yellow Chi
Frontal Bone - 2

Jupiter - Green Chi
Right Parietal Bone - 3

Earth - Yellow Chi

Moon - Silver Chi
Sacral Bone - 7

Mercury - Blue Chi
Occipital Bone - 4

Venus - White Chi
Left Parietal Bone - 5

Sun - Golden Chi
Third Eye - 6

Fig. 5.1 *Planet, Bone and Cosmic Colors*

Earth Chi

While Cosmic Chi (the atmosphere over us and the universal force) is ethereal in nature, what is called earth Chi is not. This Chi appears as the line visible above the horizon when the sun is low in the sky, and seems to hover just above the ground. It actually also sinks into the ground.

Earth Chi contains yellow and some white Chi. When the ground and the cosmic (air) combine they will become white Chi. This Chi has a very powerful healing energy, not too overheated, but balanced and mild.

When treating a problem, white Chi, in the proportion 7:3 (7 white to 3 of the selected color) acts as a brake on colors which are hot or too active. This enables the body to take in chi more easily.

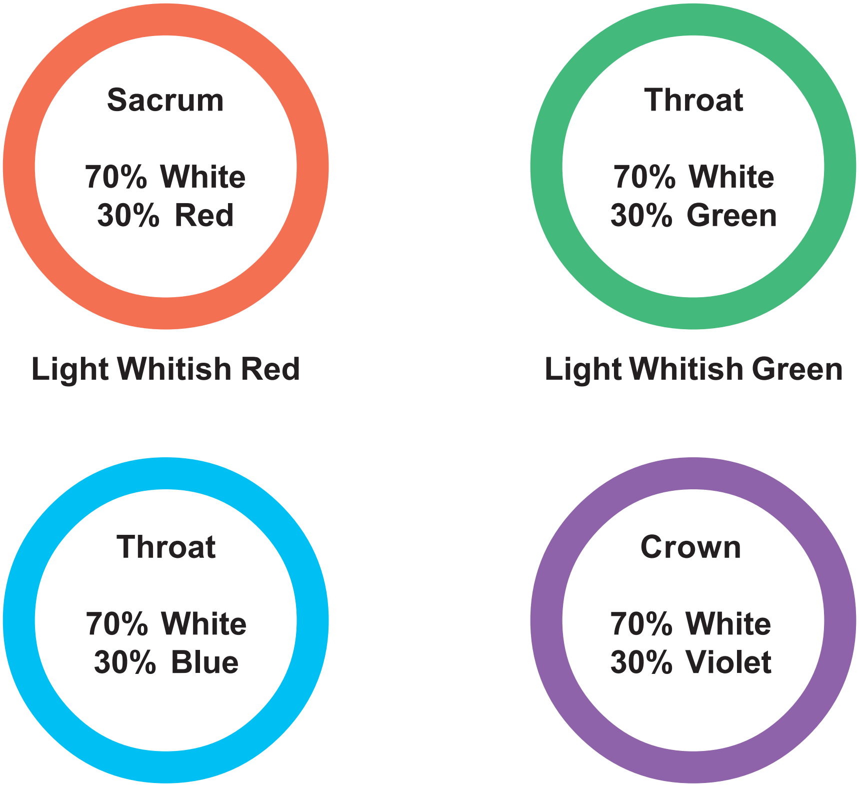

Light Whitish Red **Light Whitish Green**

Light Whitish Blue **Light Whitish Violet**

Fig. 5.2 *Use White Chi to blend all the other Colors.*

Special Note: Many of the color references came from the following books: *"Let There be Light"* by Darius Dinshah and *"Advanced Pranic Healing"* by Master Choa Kok Sui.

Human and Nature Chi

These include the things we can see on the earth: the mountains, lakes, seas, forests, streams, caves, valleys, rocks and precious stones. Taoism classifies five elemental forces: earth, metal, water, wood and fire.

So there are five elements in our body and five elements in nature; there are five elements in the cosmic and five elements in the universe.

According to the Taoists, when we can make the connection and control our body's elements, we will quickly make connections to nature and the universal forces and employ them very easily.

Solar Chi

Before sunrise or sunset there is more white Chi in the air, which we can use abundantly for health and healing. White Chi affects the lungs and the large intestine, so we direct this Chi to these organs. This will generate further Chi to maintain our daily activities. All colors of the spectrum are present in white.

Simple Practice

Look at the horizon at sunrise and sunset. Stand still and extend your palms toward the sun, smile and absorb the rising or setting sun's energy into yourself.

General Information

Energy can be introduced to the affected area of the body by moving the hand clockwise in a spiral.

Move the hand **anticlockwise** with a spiraling motion to clean, flush and remove energy from the treatment area. After this, brush down the whole body, to direct the sick energy into the earth.

Use an anticlockwise spiraling movement in order to clean, then expel the energy.

Use a clockwise movement to bring energy in and an anticlockwise one to take it out.

As a general rule and especially when you are beginning to practice the techniques, keep the color combinations simple. **Always use blue (to cool, calm and stabilize), green to clean, white to harmonize and violet (which carries the properties of all the other colors) to program.**

Although it is important to learn all of the applications, it is more important to begin practice, maintaining a serene state of mind throughout. You will immediately realize that Cosmic Healing provides the practitioner with as much benefit as the student. The more you practice the more you heal.

Healing Colors

White Chi

The color of the chi from the air, earth and sun is always white and acts as a calming agent for the surplus chi produced by the other colors. It also redirects this chi to areas of the body which need it. For this reason, use white in combination with another color in the proportion 7:3.

Shining white chi should be concentrated at the center, while colored Chi goes to the edge. This can also be reversed. It is also possible to blend the white and colored Chi together.

It is safe to use white chi when you are not sure which color might be suitable for a patient, especially when treating babies, toddlers, the elderly and the weak. People with minor problems can be treated with blue, green or mauve, while gold colored Chi or mauve/white Chi is better when the problem is serious.

For a stimulating and strengthening effect use white-red.

When it becomes necessary to switch from one color to another, an abrupt shaking movement of the hand should ensure a smooth transition to a new color.

There is a very strong violet light in the North Star. You can dilute the color with white. White Chi is composed of red, orange, yellow, violet and blue.

Blue Chi

The effect of blue Chi is one of cooling, and hindering similar to yin or water energy. The opposite effect can be obtained with the use of red Chi.

When we draw the energy down, some peoples' bodies can't adjust to the energy immediately, so using blue Chi is always good. “Water blue' is the one element that always brings harmony and does not cause any harm.

a. Blue Chi, like yin, water energy, has a cooling effect. It also can act as a restraining agent.

b. Because of blue Chi's ability to cool, soothe and numb, it is particularly appropriate for promoting relaxation and sleep, reducing abnormally high body temperature, minimizing pain and swelling. It can also hinder the proliferation of viruses and bacteria and assist in the rapid clotting of blood.

Blue (and green) can detoxify and also energize.

Green Chi

While tangerine Chi can be quite dangerous, the action on green chi is gentle and harmless. Freeing up a damaged area so that cleaning can then take place may be done with green Chi. Once it's loose you can then use blue or organ's color Chi to expel the disease completely. It can therefore remove the disease and bad energy by locally sweeping either to and then out of the arms for the upper body or out of the legs for the lower body.

If there is a need to energize with the organ's color Chi, first energize with pale green Chi. Always use green chi before using the colors purple, tangerine and red.

Use the throat to activate and draw the green Chi up to the crown, rising to the universe to be multiplied and brought back down for use. Always add blue Chi when there is pain. The motion of pushing and pulling through the affected part is very powerful. You also can also use the 'sword finger' in a small area and for infection.

Green can also be used for energizing together with violet; first use green and follow with violet.

Pale green and pale tangerine chi can be employed in sequence, as the first acts on the 'sick' energy, reducing it to a form which can easily be driven out by the second.

Simultaneous use of pale green pale tangerine and white chi in the ratio 2:1:7 is recommended to treat stubborn areas because the effects are intensified when these proportions are used together.

In the treatment of cancer, dark green and tangerine Chi may be employed, preceded by blue. This can assist in toxin removal, and in the cleaning and freeing up of disease affected areas so that necrotic cellular tissue is released. These colors can relieve congestion and free up 'sick' energy which is 'stuck' when used with brushing movements over the affected areas. They also help with treatment of colds and elevated body temperatures, and in the dispersal of blood clots.

Red Chi

Cerise chi makes the area to which is applied stronger, while crimson chi has the opposite effect. When using color for healing, project luminous white in the center and cerise at the edges for a strengthening effect. Always combine red with blue and green; never use it on its own.

Cerise mixed with white, because of its expansion abilities, can help widen blood vessels and breathing tubes. It also helps those with heart and asthma problems.

Red chi with its qualities of warmth and expansion, has a number of benefits. It is good for the circulatory and respiratory systems and the blood, as it expands blood vessels and other tubes, thus improving circulation and breathing, particularly for heart patients and asthma sufferers. It also brings increased energy into the blood and helps allay feelings of fatigue or weakness, driving out 'sick' energy along with toxic substances and other waste materials. It can also help those who are sensitive to allergens. It may help those whose lives are near the end have a longer period on earth, and promote revival of consciousness.

Avoid the use of dark colors, which may cause an adverse reaction. If a person is suffering from a sexually transmitted disease, it is not advisable to treat him or her with chi which is crimson. Chi of this color makes the microbes that cause such diseases multiply at a fast rate and can also produce an inflammatory or constrictive reaction.

Tangerine Chi

When using orange Chi, always use blue Chi first; to tone and calm down the ailing body area, before applying the vitality of orange.

Pale tangerine chi with added white can be used; however certain areas are inappropriate for treatment by this color because of its power. Do not use with the eyes (including retinal problems), heart, head, throat, brain (for example bleeding in the brain) or the spleen.

The large intestine benefits from the use of tangerine chi, but this color should be used sparingly on areas like the solar plexus and navel. Appendix problems should not be treated with this color as it could exacerbate the problem. Tangerine chi has an effect on the bowel, and can also aid in returning an unconscious person to consciousness or in stimulating someone who is dying.

To treat cancers of certain types, dark green and dark tangerine are suitable because of their ability to destroy. Always avoid the head and the heart when using orange.

Forcible removal of waste products, toxic substances, ‘sick’ energy, viruses and bacteria
Menstrual difficulties
Waste elimination problems
Freeing up ‘sick’ energy
Breakdown of clots in the blood
Problems of the urinary and respiratory systems
Diseases of joints and connective tissue
Common cold
Problems caused by allergens
Cysts

Yellow Chi

Yellow Chi has a close connection to the nerves, the bones and the bone marrow. It is the color of the spleen; in Taoism we feel the spleen is involved in the assimilation and processing of food and the appetite. This doesn't mean eat a lot, but when the spleen is balanced the appetite will be controlled.

Chi which is yellow in color is beneficial in the treatment of nerves as it promotes re-growth. It also helps in the repair of fractures and cells. It is good for the health of bones, tissues and organs. It has the power to increase bone marrow production.

Violet Chi

Violet Chi has the properties of all the other colors. It’s used for severe types of sickness.

The qualities found in all colors are contained in purple chi, therefore it is particularly effective for serious illnesses. Mauve (light purple) chi which contains elements of blue and green, helps to restore organs and nerves affected by disease or trauma.

Although mauve chi can be employed on certain kinds of severe infections, chi which is dark purple should be avoided when diseases of the respiratory system are present, as it could cause the proliferation of disease causing microbes.

Violet Chi is use chi that is pale white/green or pale white/tangerine and, selecting each in turn, with a brushing movement apply to the whole spinal area and especially to the area where the trauma is located.

If purple chi is used in combination with other colors, it will intensify their qualities and if used together with red, tangerine or yellow chi can be dangerous; it may cause the undesirable proliferation of some cells.

Dark electric violet can be used for tumors or cancers. Always use the blue or green first.

In order to make an area which has become weak stronger, mauve mixed with white can be used, but for really fast strengthening use cerise mixed with white.

Both white and violet have the properties of all the Chi colors. The difference is that violet Chi has a greater penetrating effect and is easier to assimilate than white.

Because bright purple chi has been acted upon by the soul, it has the capacity for independent thought and action.

Energy or chi which comes from the air, the earth and the sun is called purple chi. There is a second type of chi with this color, but having more brightness – this has been called bright purple chi and is derived from the soul through the crown of the head, the point connected to the North Star or Big Dipper.

Bright purple chi can only be brought down from the universe via the crown point if the practitioner has highly developed senses, The effects of chi from air, earth and the sun depends on the development of the practitioner’s soul.

In contrast to purple chi, the chi which is bright purple in color has a much greater potency because it can help in the restoration of nerves and organs which have been affected either by disease or trauma. This bright purple chi can be used to remove infection and to promote fast healing, even in the case of serious illness.

In addition, this chi has been said to possess its own consciousness (Master Choa Kok Sui. Advanced pranic healing:49)

When cancer and other growths are being treated, use bright purple chi because it has the ability to destroy. However, beforehand, the affected area should be energized with pale blue chi.

Do not use bright purple chi at the same time as chi which is yellow, green, tangerine or red because it will intensify and increase the effects of these colors. In addition, the use of any dark colors with bright purple chi (either preceding or following its use) should be avoided, as this will produce a very damaging result. As a general guideline, use blue before electric violet.

Golden Chi

Gold colored chi, which has almost the same qualities as bright purple chi, is formed when bright purple chi meets the spirit body. The color changes to cerise (pale red) when this gold colored chi enters the physical body. However, this chi is does not have as much influence as the bright purple chi, being gentler and therefore less effective as a cleansing agent. Gold colored chi should be used to treat a very local area, while bright purple chi can be employed over a wider area.

Spleen

The spleen is connected to the solar plexus and the navel, the major center that connects to all the organs. Holding solar energy or a white ball at the solar plexus will help distribute Chi to all the organs.

If the solar plexus gets clogged up the liver and spleen will also become blocked. The spleen, the Door of Life and the navel are all linked together; the navel and the Door of Life sit opposite each other. We call this an emotional link. Many diseases are caused by emotional links; people use psychology to deal with it which can work to a certain level, but all the energy would eventually get stuck at this main juncture or plexus. White chi is taken in mainly at the spleen, and before being sent to the various parts of the body, is split into its colored forms of red, green, purple, blue and tangerine.

If the spleen is not clean, the immune system will be affected, together with the production of bone marrow. The blood will be dirty and diseases of the joints and connective tissue encouraged.

Cosmic Healing Chi Kung works very well with Chi Nei Tsang. The Cosmic Healing Chi Kung, Chi Nei Tsang and all the other Universal Tao practices all play a unique role in working to heal the physical and subtle bodies. When you combine these three in synergy, it allows you to perform many healing tasks. The spleen (located under the left side of the rib cage), spleen center or navel can absorb white Chi directly into the front or back. The Chi will be broken down into color and distributed to other centers and their associated organs.

There is one thing that I always advise: the recipient or student should be under a doctor's supervision and that they should alternate between the medicine from the doctor and the work with us at the Universal Tao. The doctor should be advised of what we intend to do with the student.

Cosmic Healing Planetary Colors

In Taoism we give each planet a color which enhances the healing power. Have the 'Planets' chart (located at the beginning of this chapter) in front of you, look at the relevant planet and close your eyes. Hold the planet in your mind.

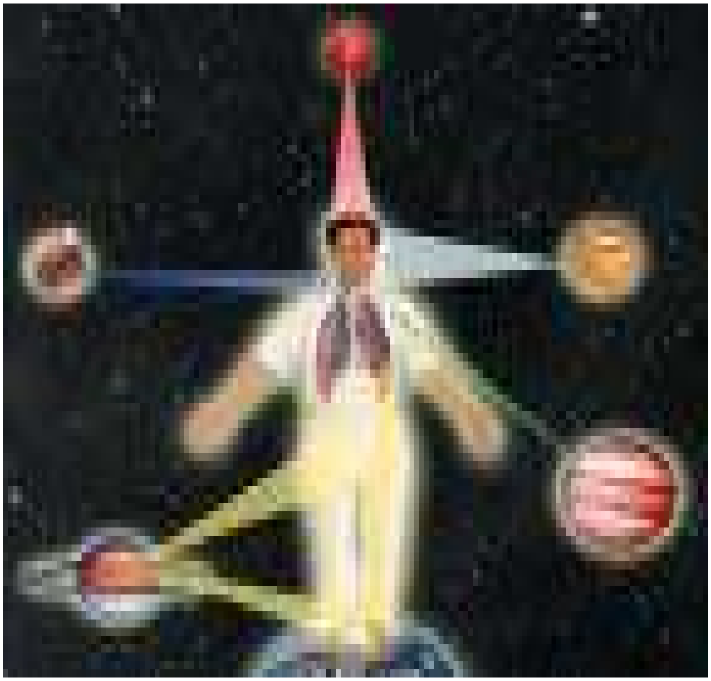

Fig. 5.3 *Channel the Cosmic Planetary Energy into the body.*

Mars: Focus your awareness on the star above you, turn your eyes up and look up to the crown. Hold the image of Mars and extend your sight beyond the crown, looking up into infinity. See the **Red Light** and gradually bring it down to your head about 6 feet above you. Form it into a red ball, a holographic image of the planet Mars shining its red light above you. Invite it to flow down to your 'star' (located above your crown) and continue to flow into your Upper Tan Tien for processing. Let it flow down to the heart (or T-2) and down to the palms. Red and orange light also can be breathed in through the sacrum. Feel the sacrum pulsing and breathing in the red and orange light. Practice 9 to 18 times, rest and shake your hands.

Venus: Become aware of the star above you, turn your eyes up and look up to the crown and into infinity, seeing the expanse of **White Light.** Gradually bring the light down to your head, about 6 feet above you. Form it into a white ball, a holographic image of the planet Venus shining its white light above you. Invite it to flow down to the star above your crown and continue to flow into your Upper Tan Tien, for processing. Now it may flow down to your lungs freely.

Saturn: The **Yellow Light** from Saturn can also be absorbed via the earth through the soles of your feet and perineum. This is for your spleen. Follow as above.

Mercury: The **Blue Light** from Mercury may be absorbed through the throat center. This is for your kidneys.

Jupiter: The **Green Light** from Jupiter can also be absorbed through the throat center. This is for your liver. Follow as above.

Summary - Working with Color

(Practitioner to Student)

There are two ways of doing this type of healing session:

One way is to focus on the specific location in your own body, using your hands and ask the other person to do the same. Together you complete the entire route.

Another way is to focus on your own energy body and ask the student to do the same. Bring a channel or line down from your energy body into your physical body and do the same for the student. Once again, complete the routes together.

For reasons of clarity this text only mentions one student and uses the male gender; *this session may however be performed with a group also*. Before starting the session do some group meditation practice and warm ups.

1. Be aware of the Tan Tien and connect to the universe. Let the student sit with his back to you. You are behind him. Be aware of your **Sacrum** area and feel the Chi; **wait for the Chi to rise up to the crown** and to the universe. Spiral the Chi down to the student's and your own energy body and it will flow into the physical realm.
2. Draw in the green light with your palm, push it right through the sacrum and disperse the sickness throughout the other side of the universe. Pull the green light from the universe through the sacrum and push the sick Chi out to the universe. Do this 6 to 9 times until you feel the **sacrum has been cleaned**. When you pull back stop the green Chi at the student's sacrum; there is no need to pull it back to you.

   Yellow light from the earth will help strengthen the sacral bone. Visualize the sacrum and vitalize the complete bone structure with yellow light passing throughout.
3. Be aware of the energy body above the **Crown**; extend yourself up from above your crown and **channel down the white light from the center of the universe and the violet light from the North Star.** Using your 'mind-power' ask the inside of the bone structure to open, allowing the white and violet light to **flow into the marrow.** Focus on the energy body; you can picture the energy body's sacrum and bone

structure to help guide the energy inside and see the whole body light up from deep within.

4. Then focus on the **Door of Life and the Navel.** First flush them through with **green** light, cleaning them out. Draw the green light from the Door of Life through to the navel, and then push back through. To cool down the Door of Life (if necessary), draw the **blue** light and push it through from the navel to the Ming Men, connecting with the universe. Then bring the **white or violet light down and activate the Lower Tan Tien.** Always teach people about their Tan Tien. It is like an ocean and the body is like a hollow bamboo. The bamboo can bring the water up. This way energy will never dry up.
5. Concentrate on the **Solar Plexus and the Chi Chung (T-11)**. Once again, draw the **green** light, spiral it, use it to flush, then **energize with white and violet light**. As before, always allow the energy to stream through both the points completely. The solar plexus holds all the emotions. When working on the solar plexus, the **most important thing to remember is to connect the 'backside' to the universe.** There is literally no end to this connection. Just allow the Chi to come all the way down and then pull it towards the rear slightly. Simply clean out the path. Allow the information to condense; allow any images to manifest and then release. Then stabilize the energy. Picture the Chi-Field enveloping the person as a big protective bubble. **Cool down with blue.**
6. Proceed to the **Heart and the Shen Dao (T5/ T6).** Draw in the **green** light, push it through and repeat several times. Now scan the heart with your palm, sensing its strength; then select an appropriate hue, i.e., not too dark. Send the **red** light through the heart to strengthen it.

   Cool down any excess heat in the heart, flushing down and out, using **blue** color. Draw in the **white** Chi; push it through to the **Wing Point (T5/T6)**. **Energize the center at the back of the heart using violet and golden Chi.** Picture the heart surrounded by a golden aura.
7. Move up to the **Throat Center, flushing through to C-7**, using **blue** light first to open and clear it and then **green** light to clean.
8. Next activate the **mideyebrow.** Focus on the mideyebrow in your own body. Use **golden yellow** Chi; use it to **flush and stabilize** the mideyebrow. Flush it all the way through to the backside of the head. **Energize with violet golden light.**
9. Proceed up to the **third eye** in the middle of the forehead. **Connect to the Kun Lun** at the back of the crown. **Flush** through both points using the pale **violet light. Energize** with **electric white or golden light.**
10. Concentrate on the **crown**. Use **violet** or **golden** light to enter and **flush** all the way down through the **central channel**, leaving the body at the **perineum**. **Cool down the system by showering blue light over the whole body.**

    When people have a serious infection, they should consult a doctor. It is common for infections to take a long time to clear up nowadays, even when using antibiotics.

There are two ways to deal with this; allopathically or naturally. The Taoist practices are in great conflict with ‘contemporary (i.e., not alternative/natural) western concepts’. The western concepts recommend a full course of antibiotics. Using the theories of Chi Kung, a baby for example, would absorb natural antibiotics via the bones. Western Allopathic Practice consider antibiotics as wonder drugs. The only ‘wonder’ is when will they realize the many harmful effects that it has on the body? Babies that have been ‘fed’ with antibiotics frequently suffer many organ related problems by the time they reach 20 to 30 years of age. The worst thing is that they are given to children nearly every time that they become ill. If you scan (cosmically) the solar plexus, thymus or spleen of such a person, the area feels like a hollow void. This means that the immune system is totally non-functional. The same thing applies to ***Aids***: the immune system no longer functions. Any virus can enter the body - medicine cannot kill it. There is nothing left to ‘fight with’.

In my experience and following my own experiments, in circumstances like the above my own practices are as follows: I clean my colon; there are a few ways to do that. I eat fiber, vegetables, especially green leafy ones. I eat a big bowl of vegetable soup. I also add things like onion, pumpkin, carrots and cabbage, which will give it a sweet taste. I don’t eat any meat and drink a lot of water. The colon is cleansed within two days. I also brush the skin, which is a very good detoxifier. The vegetables will provide natural carbohydrates which are digested and absorbed very quickly. A healthy diet consists of many whole natural foods. It is ‘alkaline’ as opposed to ‘acid’ based. I also use the ‘colema’ which I feel is the best cleanser.

Chapter VI

Individual Healing Session

The Individual Healing session embodies all of the techniques that you have used previously and consolidates them for use on a one-to-one basis. The session comprises the following three parts:

A. Almighty Knife (Chi Knife)

The almighty knife can be used for any health problems. We can charge the knife with Yin or Yang energy. We charge the almighty knife most often with Yin Power. Yin is the cold energy and is used for any inflammation, pain or heat. Yin is used to break up illness. Charge the almighty knife with Yang for strengthening, melting or expelling. Only use the Yang when there is no pain.

B. Energizing and Activating the Immune System

C. Cleansing the Internal Organs

A. Almighty Knife

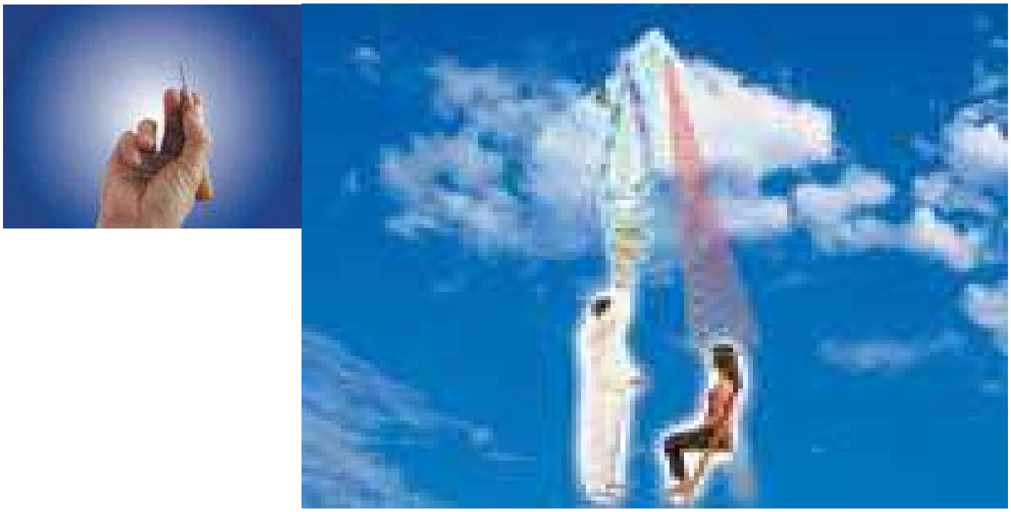

Fig. 6.1 *Almighty Chi Knife*

Since the dawn of ages and our descent into the material realm, we have constructed, manufactured and used tools to facilitate the progress of events. They are an extension of ourselves that enable us to enhance, mould and magnify our intrinsic nature as inquisitive playful beings.

We have used implements to help nourish, protect and amuse ourselves and further our progress in attempting to understand the 'how' and 'why' of existential reality. The knife has served as a means to carve our way into the present. It is a simple, yet powerful 'friend' and derives as much power as we wish to lend it.

Its design may consist of any combination of the five elements (Earth) which when combined with Universal (Heaven) and Cosmic (Man) forces, provide a unit of energetic substance which receives and transmits magnification, purpose and power in accordance with the human will that governs it. ‘Chi Knife’ and “Powers of Ten’ adopt the precepts of a truly magical existence. When we arrive at our destination we will realize that we have indeed turned “the Wheel of Law’. In achieving this end we will have completed the cycle and returned to the beginning of the journey. We will understand that the simple things in life, like the knife, contain the same amount of power as state of the art technology.

We are all returning to the Garden of Eden. Complex telescopic and microscopic equipment now enables us to see the truth:

> “Inside has no outside; outside has no inside”.

They are one in the same thing. When they merge together we will indeed enter the Kingdom of God.

Outline of Practice

1. Hold the knife in the right hand (left, if you are left handed). Hold the other hand near your body, with the thumb and index finger touching. Project the Chi Knife as big as the universe, ‘long and big - reaching to Heaven’. The bones in your arm are hollow and the ‘bottom’ is sealed at the shoulder. Feel them ‘fill and pack’ with the power coming from above.
2. Then give the command: **“Power comes from the east. This Chi Knife has the power to dissolve all the negative energy, sickness and bad fortune. By my request, please carry out the order now.”** Wait until you feel that your arm and the knife are heavy and full of Chi.

Eye Problems

Use the thumb and the index finger to hold the eye wide open.

You hold the knife like holding a pen, 3 to 6 cm away from the eye. Use the little finger for leverage and support.

WARNING!

DO NOT USE A KNIFE. THE **MASTERS** OF GREAT EXPERIENCE USE RITUAL SCALPEL TYPE KNIVES. THE STUDENT SHOULD USE THE “SWORD HAND” FINGERS (P.76, FIG. 4.24).

DO **NOT** USE SHARP OBJECTS NEAR THE EYES.

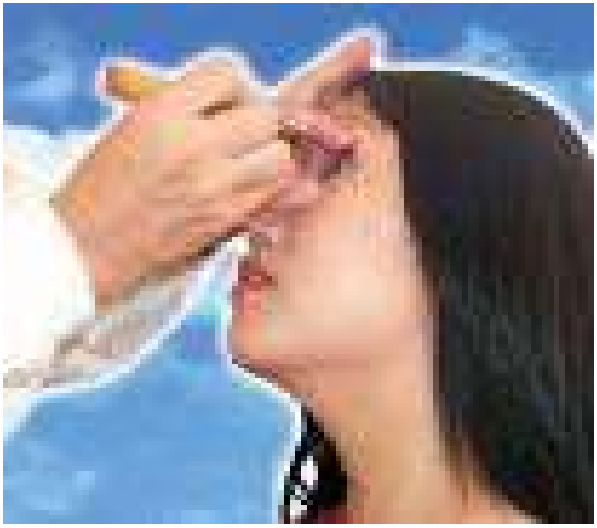

Fig. 6.2 *Hold the knife three to six cm. from the eye.*

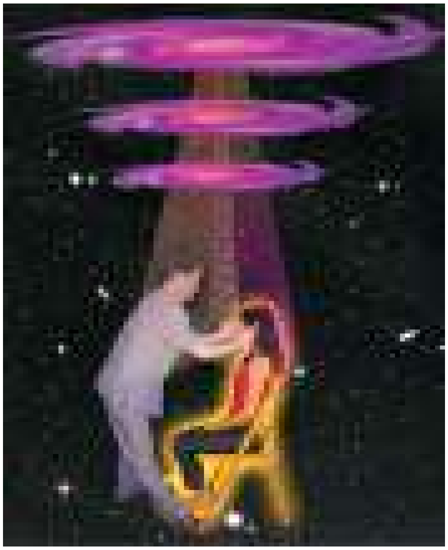

Chi knife - “Almighty Knife” helps all eye problems”. Cut up and down and sideways in the cross pattern ‘through’ the eye thoroughly. Tell the student to move the eye left and slowly move to the right as you cut with the Chi Knife. When they move their eye to the right, move the Chi Knife to the left. **Remember that the knife is used for ‘energetic surgery’ and should never make contact with the eye itself.**

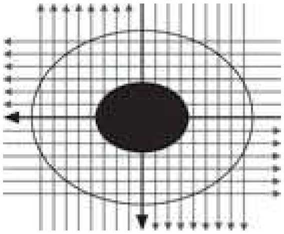

“Slice” the eye toward the right side for 50 to 100 knife movements and do the same to the left side.

Fig. 6.3 *“Almighty Chi Knife helps eye all problems”.*

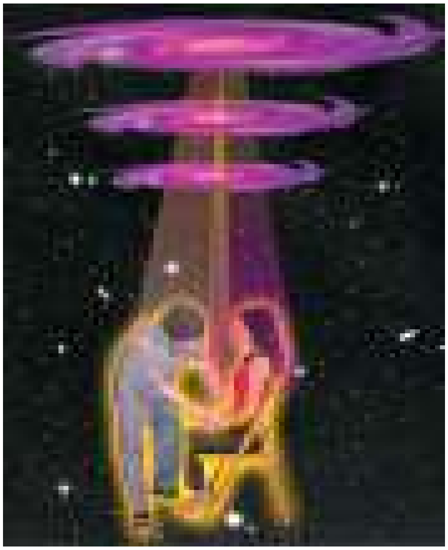

“Almighty Knife, please destroy all kinds of wounds, warts and tumors”. Cut up, down and sideways in the cross pattern through the part that has the ‘condition’.

Fig. 6.4 *“Almighty Knife, please destroy all kinds of wounds, warts and tumors”.*

The almighty knife can be used for any problem in the body. Follow the same procedure to carry out ‘energetic surgery’ on all parts of the anatomy. It is especially good for soft tissue repair, including the brain.

When you feel that the Chi Knife is full of sick Chi, spiral counter clockwise and discard it down to the ground. When you feel that the knife is ‘empty’ charge it with Chi from the universe.

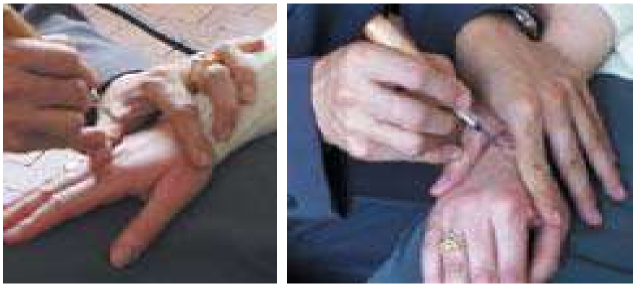

B. Energizing and Activating the Immune System

In cases of severe illness, you will want to do this technique many many times, because it is essential to stimulate the body's defence-systems. This technique is also beneficial for debilitation and infection. You may also want do this in a preventative manner.

Activating the immune system always involves the bone structure, especially the round bones in the arms and legs and the flat bones in the skull, jaw, rib cage, hips and spine. All of these are involved with the production of white blood cells. The **thymus gland** and the **spleen** are of special importance to the immune system, because they are involved with programming and training the white blood cells (called T-Cells) to do the right job. **The lymphatic system is involved in removing toxic waste from the body; it is therefore beneficial to stimulate the lymph flow as well.** This is best done in combination with Chi Nei Tsang. *For details refer to the Chi Nei Tsang Books.*

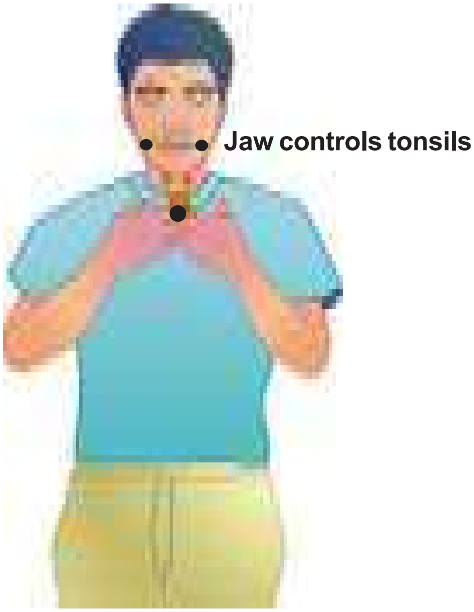

Fig. 6.6
Touch the top of Sternum to activate the Thymus.

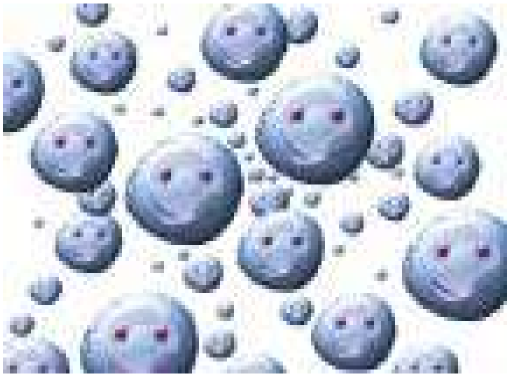

Fig. 6.7 *Happy Cells.*

If you work on a one-to-one basis, first perform the general cleansing session, where you 'pour energy' over the other person. After that use your hands, drawing Chi and pushing it through. Remember to push the energy all the way through and connect it with the universe beyond. Ask your student to concentrate on the areas that you are treating, preferably by guiding them, touching each point as you go.

Outline of Practice

Guide the student as you practice together. Work with your own energy body and tell the student to work with theirs. Presuming that you have already done the previous set of general healing, you can start with the **coccyx**.

1. Be aware of your Tan Tien and connect to the universe. Direct the Chi down to the fingers. Touch your coccyx with the fingers concentrating a beam of white light on to it. Picture the sacrum opening and the white light streaming inside, flowing up the spine and streaming out over the whole bone-structure. Transfer the feeling of Chi in the coccyx up to the universe, multiply it and guide it down to your energy body.
2. Always be aware of the Tan Tien and the universe first. Attract the yellow light; spiral and condense the healing light into the tip of your fingers; move the fingers up to touch the sacrum. Send out a beam of yellow Chi into the sacrum, directing it up the spine. This will stimulate the production of red blood cells. If you are not sure whether you are able to send it right into the marrow, picture the marrow or emit the Chi from your fingers, while at the same time concentrating on the bone marrow of the energy body. Transfer the feeling of 'Bone Chi' up to the universe and down to the energy body.

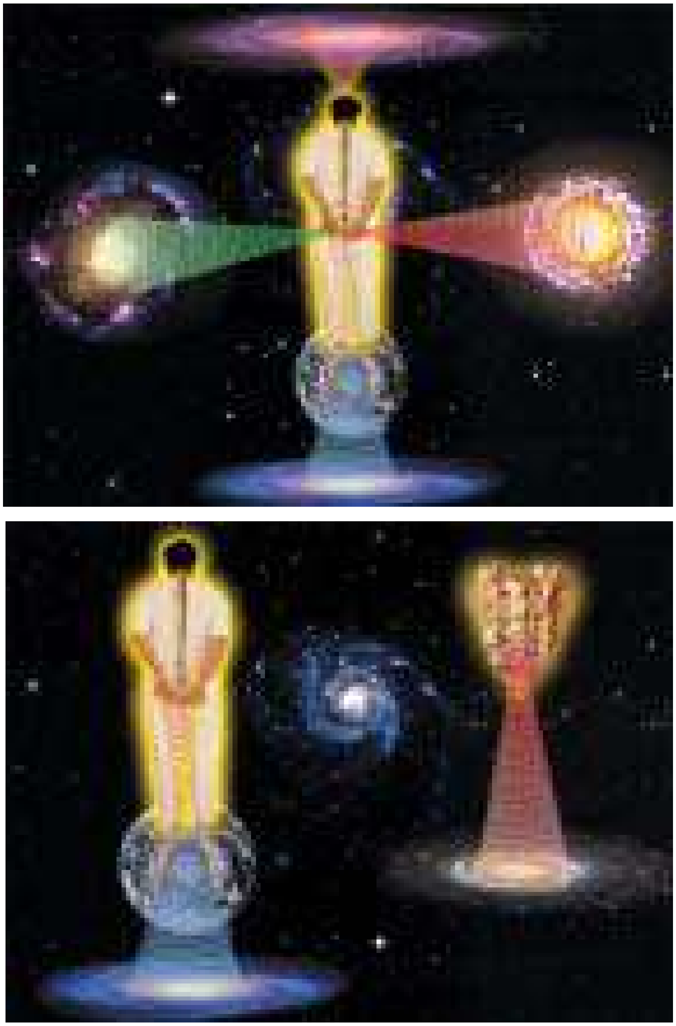

***Fig. 6.9** Touch the Sacrum.*

3. Proceed to the crown, showering it with violet light. Then spread from the crown throughout the bones of the body. Repeat several times. Instruct the student to feel the bone structure breathing. Ask him/her to extend up from the crown to the Big

Dipper and North Star, channelling down the red and violet light into the thymus gland.

4. Continue to the throat, stroking down with your palms, using green color to move the lymph down towards the backside of the heart. **(Never stroke up towards the head. Only guide the lymph down).**

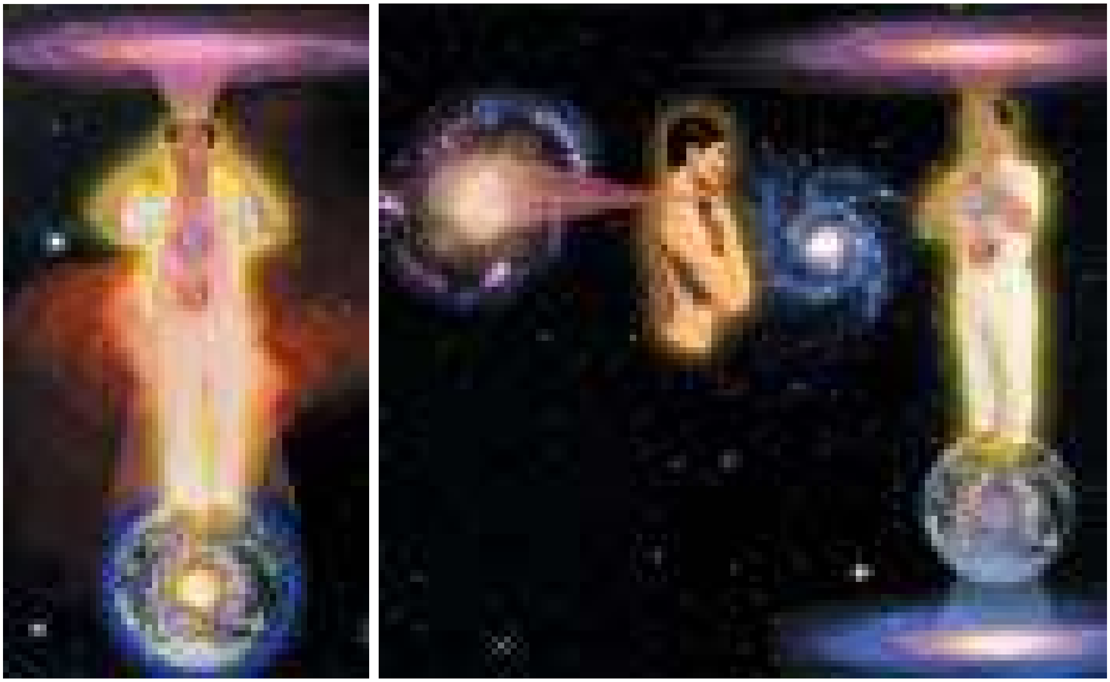

A. ***B.***

Fig. 6.10 A. *Channel the Violet Light of the North.*
B. *Use your Palms to bring the Lymph flow back to the Heart.*

5. Proceed to the lymph nodes under the armpits and across the sternum. Next activate the lymph nodes in the groin. From the groin you guide the lymph up to the navel center (Not down where it will get stuck in the legs).
6. Return to the navel center. Send the energy deep inside, energizing the lymph nodes at the back. Pull back, drawing green energy, flushing out any waste products that are stuck in them.

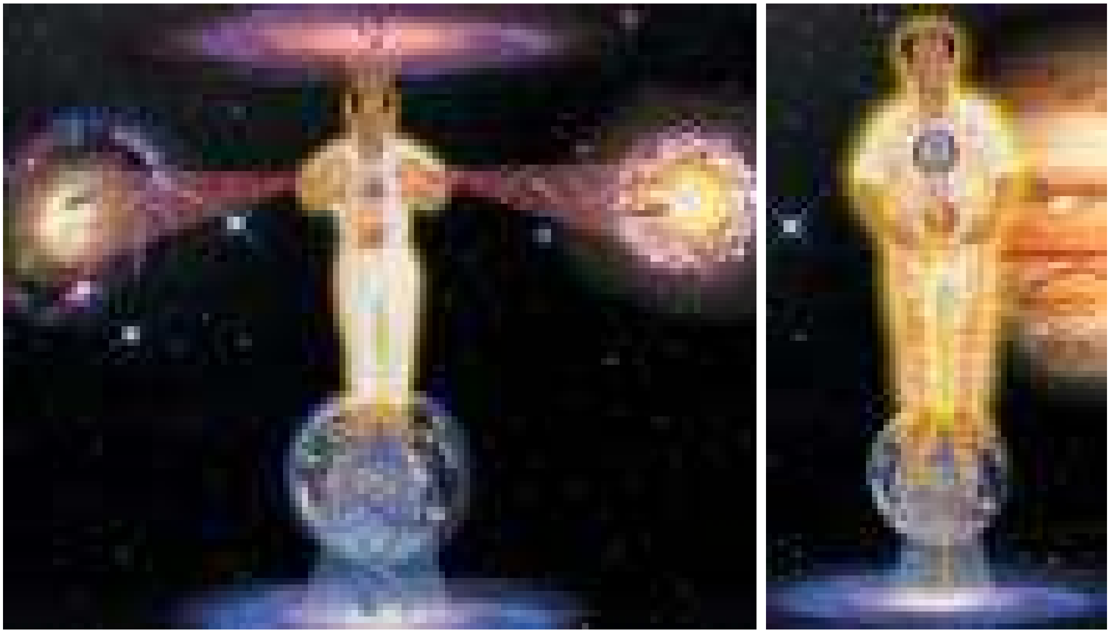

A. **B.**

Fig. 6.11 A. *Activate the Lymph Nodes in the Armpits and Groin.*
B. *Energize and flush the Lymph Nodes in the Navel.*

C. Cleansing the Internal Organs Technique

Healing Some Common Illnesses

This session can be used for any kind of ailment and carried out with any number of students. It can take place on a singular, paired, small group or large group basis. Numbers can range from one person, one hundred, one thousand, ten thousand, ad infinitum. It is more likely that a greater energetic current will be created in much larger groups. In the case of 'mass' healings, there will always be a 'guide' or 'conductor', to open the channel for the students to follow. In the case of solo practices you will of course be the guide. **The following instructions have been set out on a one-to-one basis**.

By connecting to heaven, earth and your student, you will maximise the wholistic healing potential to your physical and subtle bodies. We recommend you spend about twenty minutes on the general healing before moving on to the more localized practices of emptying and cleansing the body's centers, systems and organs, contained within this chapter. We recommend that you spend between five to fifteen minutes on all of these. **Always perform the Sacred or Chi Water practice before.**

It is important to remember that if there is no energy in your Tan Tien, then there will be no energy to work with. 'Connecting the Tan Tien and the Universe' means connect your center to the infinite source of the universe, the Tao. When you make the connection you access this power. Always hold on to your center; it is the only thing that is truly yours. Keep the fire warm and spiraling. When you are aware of your Chi, you can direct heavenly light straight into your energy body and then into the student's.

General Principles

1. Push and pull the energy through the student's body. Push and disperse the sick energy into the universe and out to the planets, where it will be transformed and recycled.
2. Pull the energy back and stop it between yourself and the student.
3. Spiral the energy counterclockwise until you see the **green** light mix and clean the sick energy. Spiral and flush the energy down into the earth.
4. Spiral with **blue** light; see it mix and flush the energy within the student's body. Spiral this energy down to the ground.
5. Energize the student with **white** or **violet** light.
6. You can also use both hands to channel Chi when you 'push and pull' through the body.
7. Energy can be introduced to the affected area of the body by moving the hand clockwise in a spiral.
8. Move the hand anticlockwise with a spiraling motion to clean, flush and remove energy from the treatment area. After this, brush down the whole body, to direct the sick energy into the earth.

Use an anticlockwise spiraling movement in order to clean, then expel the energy.

Use a clockwise movement to bring energy in and an anticlockwise one to take it out.

9. When you begin pushing, pulling and sweeping, you may feel resistance in the body. As you progress and the energy centers open, this feeling will diminish.
10. There are many color combinations that can be applied to the various parts of the body. When you start practice, stick to the basic 'blue, green, white, violet and blue' formula. When you are working on a specific area, you can keep this book near by, or write a short 'color recipe' before you begin. In this way, you will be able to relax your mind.
11. 'Pushing and pulling' is like fanning and venting the body; when you fan your hands in front of your face, it will have a cooling effect. Ventilation enables the stagnant, sick energy to leave the body.

Below is a list of the body's organs in relation to their tendency to 'store heat'. They are listed in descending order, i.e, the head is the hottest:

Head, heart, liver, lung, stomach and spleen, large intestine and small intestine, kidneys and sexual organs (which become hot when aroused).

Outline of Practice

Occupying a position amongst most of the internal organs, the solar plexus acts as a kind of clearing center for energy generated by the emotions. It also lies in between the lower and higher energy centers.

1. Scan; ask for the power to look at the cells.
2. Push, pull and give the command, “Disperse”. Push until you feel the solar plexus clear and open.
3. Spiral green Chi counterclockwise. Talk to the sick energy; tell it to “leave”, you will take it to “a better place, where it will be happy”. Give the command, “Out”.
4. Spiral Blue Chi counterclockwise, flushing out any remaining sickness. Guide the sick energy into the earth.
5. Spiral and sweep Blue-Green and Red Chi into the solar plexus, sweeping the organs. If the patient is expecting a baby, has an infection or bleeding in the intestinal area; do not use this method.

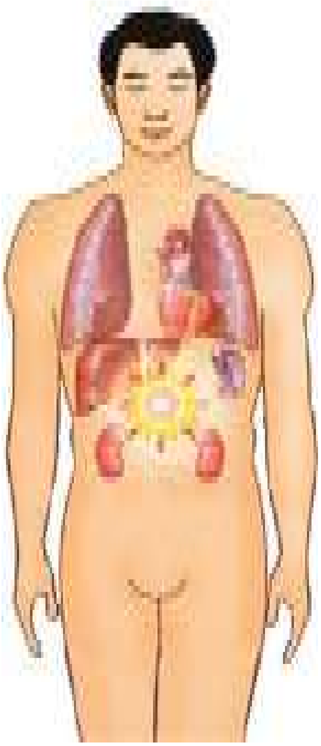

Fig. 6.12
Spiral Green Chi counterclockwise in the Navel.

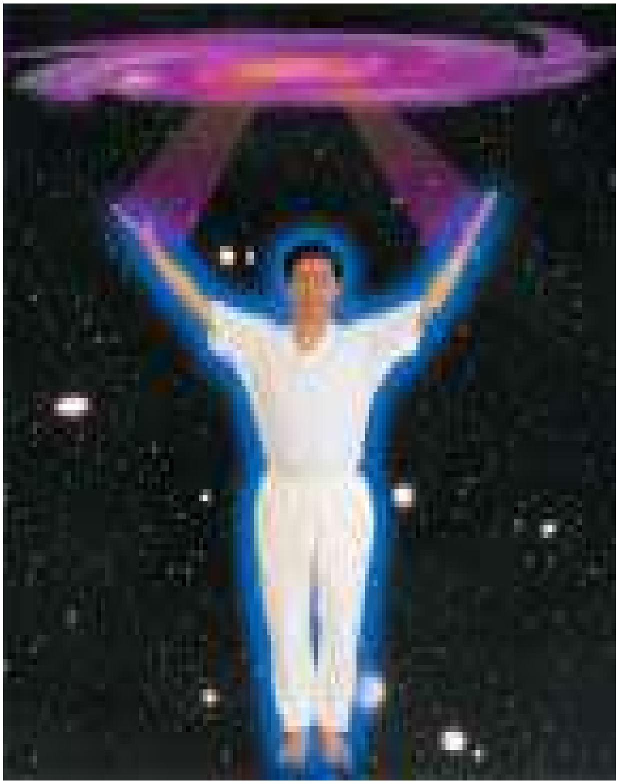

***Fig. 6.13** Charge your Hands with Chi - project the Hand Bones into the Universe and feel them fill and pack with Chi.*

Work on cleaning out the **organs**. Scan the person and see which parts are hot. A lot of people will have an overheated solar plexus, which means that this area will require a lot of attention. You will feel the heat when you scan.

You can **clean the organs by 'pushing and pulling'**. Picture the organ and push; feel your hands are very 'long', extending through the organs and reaching out to the universe. Project all the sick energy down into the earth and out to the universe. Use your intent to guide it into the earth and out to the other planets in the solar system, where it will be processed and recycled. Pull the universal energy back towards yourself; push and disperse all the negative energy out to universe. Pull; when you pull the Chi back past the student, stop in between the two of you. Push, pull, push and pull, until you feel the energy or power. **Ask; "Give me the power to see". Feel your Chi entering into the organs with ease.**

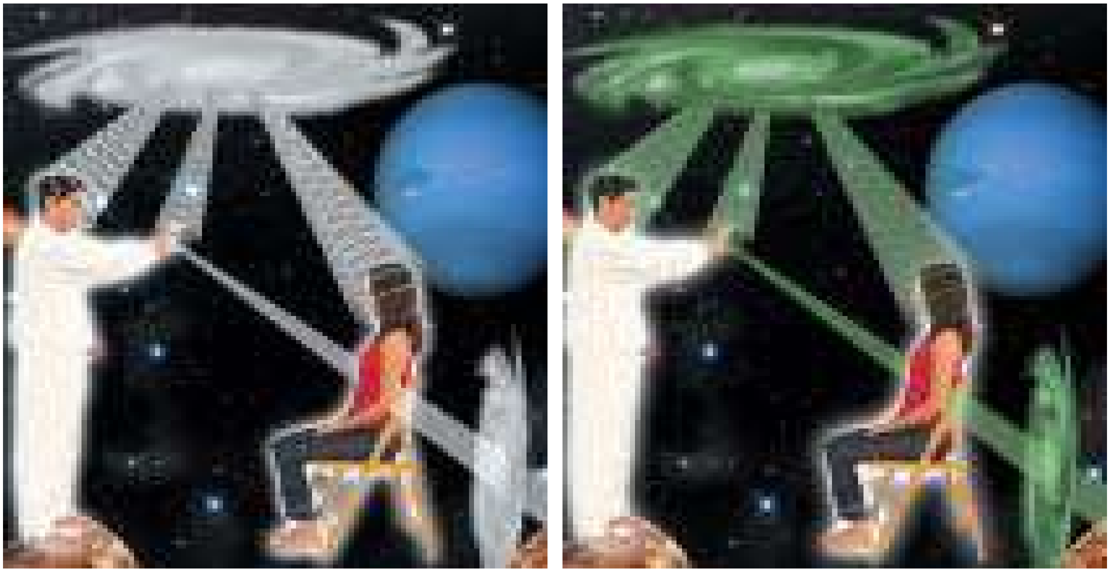

A. ***B.***

Fig. 6.14

A. *Extend through the solar plexus and reach out to the universe.*

B. *'Clean and energize with Green' - push and pull to help loosen the solar plexus. Spiral counterclockwise with a Green Light. Blend with the sick energy and guide it down to the ground.*

Clean with Green Light. Spiral counterclockwise. Feel the green light mixing with the sick energy. Push and pull until you feel the energy in the solar plexus begin to disperse. Spiral the energy counterclockwise and guide it down to the ground.

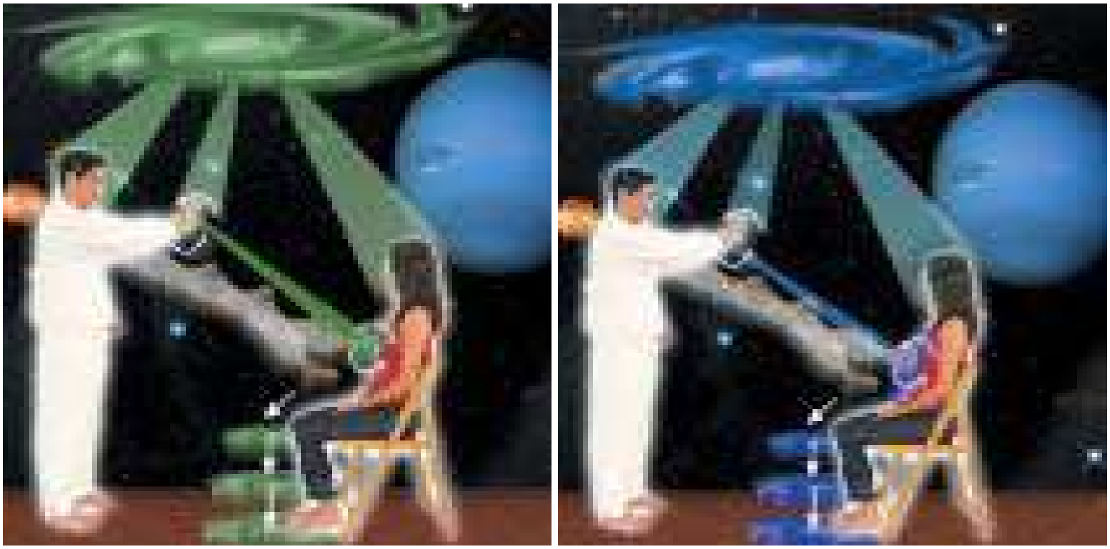

A. ***B.***

Fig. 6.15 A. *'Push and Pull' to help disperse the energy in the Solar Plexus.*
B.*'Flush and stabilize with Blue'-Picture Ocean Blue and flush sick energy down into the Earth.*

Rinse and flush with blue light. Picture ocean blue and flush down to the earth. See and feel the solar plexus becoming cleaner. Flush the sick energy down to the earth.

Now clean out the whole body again. With all the organs: from the lungs, heart, liver, gall bladder, pancreas, spleen, stomach, small intestines, large intestine, kidneys and sexual organs. Clean all the way down to the earth. **When you clean the solar plexus and all the other organs you remove all the negative emotional energy.**

Energize the **spleen**, **liver** and **kidneys** with blue or white light. Then project green, followed by orange light to the cores of the liver and kidneys. Clean with the green and orange light and stabilize with blue light.

Work on the **back** of the **spleen** and **kidneys**. Energize them with blue-white light. The spleen is on the student's left and kidneys on the right and left. They all help to clean the blood. Spiral the energy clockwise into the spleen. See it in your mideyebrow. See the cells and spiral the light into them. Give the command, "**Stay**". This is very important; it makes sure that none of the cells hide from the dazzling light that you are sending. Give the command again. Then stabilize them with blue light. Picture the kidneys and see their cells and feel the blue light go into the cells.

Work on the **Liver** on the student's right hand side. Smile and laugh to the liver and see all of its cells. Give the command, "**Stay**"; then send in the healing light. Send in blue or white (energize), followed by green (energize and clean), blue (rinse and stabilize), violet (recharge and reprogram) and then blue to stabilize again.

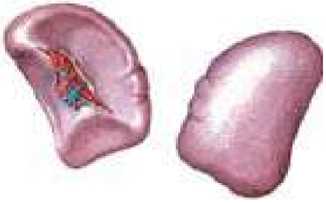

Spleen - Front and Rear

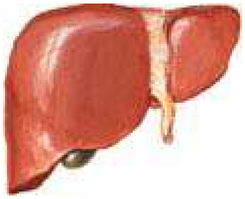

Liver

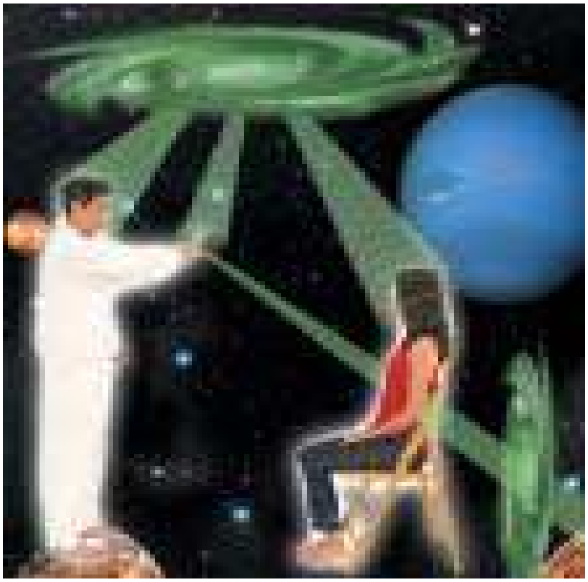

Fig. 6.16 *'Push and Pull' to help disperse the energy in the Liver and Spleen. Clean with Green and flush and stabilize with blue. Flush the sick energy down into the Earth.*

It is essential that you feel and talk to any sick energy. **If the student is healthy then you will feel their 'light'.** Ask their energy body to tell you where it **may** need some healing. Send in 'light radar'. Transmit green light and flush down to the earth. **When you feel that energy is accumulating, start to spiral more.** Push and pull and give the command for the sick energy to leave.

The more you clean, the more you heal. Tell your student to practice at home after the session, at an agreed time. You can then carry out 'absent or distance' healing, linking everyone's personal star, energy body and the universe. **With time, you will master the techniques, imprinting them with your own manner.** Eventually symptoms disappear, as the blockages are cleared and the flow returned.

Cleansing the Spinal Cord

Always cleanse the spinal cord, which protects the entire body from diseased and negative energy, both physical and psychic. When you open the spinal cord and solar plexus you can release many energy blockages. Sweep the hand in a clockwise motion to energize the spine with either/or a combination of; **blue, green, white and ultra-violet**. Then spiral counterclockwise and **brush** the spine down to the ground. Continue sweeping until you feel that the spinal cord has cooled down. Remember not to use orange or red.

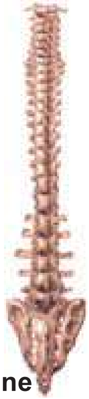

Spine

Cleansing the Blood (Lungs)

Pale green/white and tangerine/white chi work as cleaning agents for the blood. Diseases of the blood and arteries can be treated. The chi first energises the lungs and as the blood passes through them, it takes in the chi, thus becoming purified. The benefit is then carried to the rest of the body as the blood circulates. Serious infections can also be treated in this way; however, the technique should not be used on women who are pregnant.

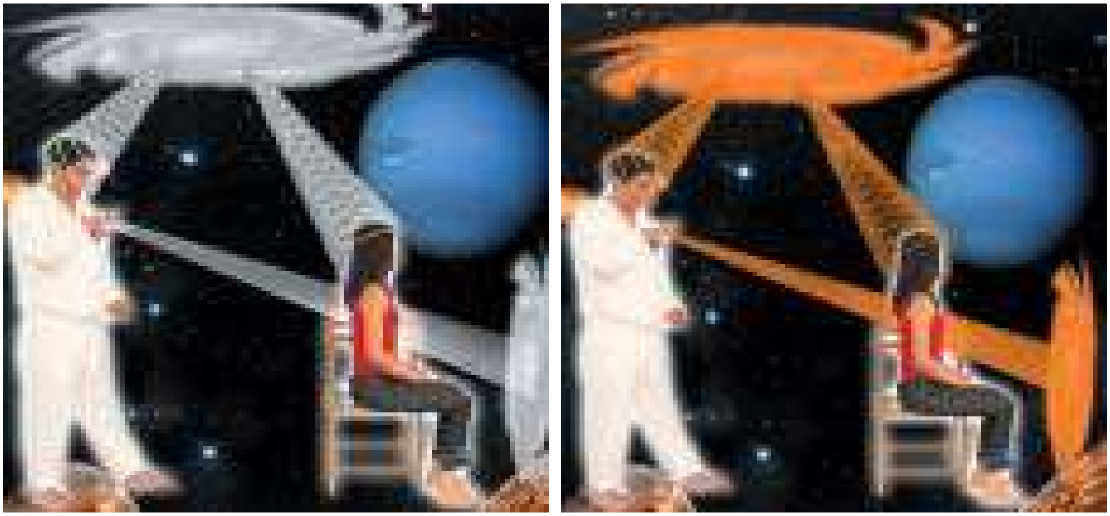

A. *B.*

Fig. 6.17 A. *Cleansing the Blood and emptying the Lungs.*
B. *Become aware your Three Tan Tiens. Connect to the Universe and ask for 'the Power to See'.*

Outline for Cleansing the Blood/Lungs

1. Scan the student.
2. Push through and open the lungs from the back to the front.
3. When the lungs need to be given energy, use pale green/white followed by pale tangerine/white chi beginning at the front and moving top the rear.
4. To make weak patients stronger, cerise/white chi can be employed as this will bring energy into the lungs.
5. Purification of the blood is carried out by the liver, kidneys and spleen. Use white chi to bring energy to these areas, after first pushing and pulling through them.
6. The blood requires a lot of cleaning on a regular basis.
7. Energize with white Chi. **Germs can only grow in the dark. 'Fill yourself with the light'.**

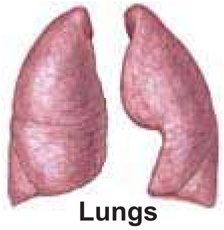

Circulatory/Heart System General

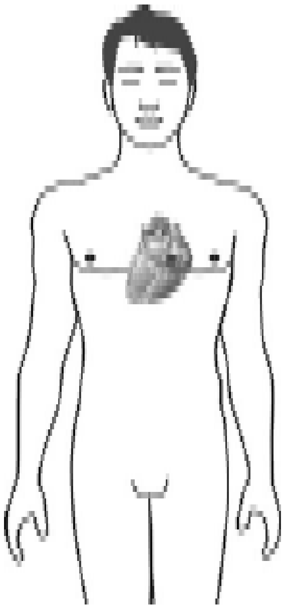

One of the basic problems pertaining to the circulatory system and heart is hypertension and hypotension. Once again, the sacrum and solar plexus are involved. The sacrum controls the bones; indirectly, it stimulates the marrow and enhances the quality of the blood. Good blood is like running the highest quality motor oil through the engine of a motor vehicle.

(Within the Universal Tao System we also have the view that the sacrum controls the Chi-pulse. If there is enough Chi, the heart does not have to work so hard, because the Chi pushes the blood).

The solar plexus relates to the emotions, affecting the heart through two mechanisms:

- It may affect the heart directly, causing rhythmic disturbance or pressure around the heart muscle
- Malfunction in the liver or digestive tract, leading to high cholesterol levels or gas which pushes up towards the heart from the intestines and causes unwanted heat and pressure.

Cleaning out the Solar Plexus will greatly improve these conditions. If you have been trained in Chi Nei Tsang you can use it in combination with the Cosmic Healing techniques.

Specific

Working on the Heart Muscle:

Clean out the Solar Plexus as outlined above. Push and pull with green and blue.
Detoxify the liver, flush with green.
Enhance with white and violet light.
Use the finger to work on the lower left and upper right part of the heart.
Remember that the heart muscle is very delicate.
You draw the force, you spiral and you push.
Stabilize with white.
For enlarged heart use blue.
Carefully use the red light to balance the blood and open the blood vessels.
Sweep the sacrum with green, red and blue.
Sweep the solar plexus with green, blue, red, white and violet.
Energize and cool down the throat center, using blue.

Working on Hypertension (High Blood Pressure):

Overactive adrenals/ kidneys often cause hypertension. Clean out the solar plexus first, because the adrenals (T11) are connected to it. The adrenal glands produce adrenaline, the hormone that stimulates the heart to pump. Overactive adrenal glands will eventually affect the heart.
Tune-up the Door of Life with blue, calming it.
Take the pressure from the heart, use green.
Energize the heart, use violet.
Flush down from head to feet with blue;
The basic problem with hypertension is that the Chi goes up and cannot come down. Work on the sacrum, using orange and yellow.

Working on Hypotension (Low Blood Pressure):
In hypotension the Door of Life and T11 are underactive.
First clean out the Solar Plexus, this time focus on removing blocks or stuck energy.
Energize the Door of Life, using green and red.
Energize the Sacrum (same colors).
Energize the Base of Skull with green and violet.

Stomach Pains

Pass your hands over the patient and brush the whole abdomen, especially the solar plexus, navel and lower abdomen. The navel should be treated with a combination of green and blue, and blue should be used at the end to firm up the energy.

If problems persist: A) Treat with Chi Nei Tsang B) Seek the advice of a qualified medical practitioner.

Relieving Pain: *Blue Chi*

Chi that is associated with the colors blue and green can be utilized in different ways. While green chi acts to free up energy that has become ‘sick’ and ‘stuck’, blue chi can act as a painkiller because of its ability to reduce the severity of pain.

1. Push and pull the through the affected part out to the universe or ground the sick energy until partial relief is attained.
2. A large amount of pale blue/white chi together with pale green/white chi should be used to transmit energy to the problem area.
3. For the alleviation of pain, use pale green/white chi and pale tangerine/white chi in turn.

Energize means draw the Chi into the part that you want to receive healing. Cleansing means push energy through the affected part and send it out to the universe. Pull the new Chi back into the same area. Light green Chi is used in loosening and cleansing delicate organs like the eyes and the brain.

Headache: *Light Whitish Green and Light Whitish Blue Chi*

1. To check for areas that are congested or lacking in energy, the eyes, temples, spine and solar plexus area should be scanned. Also, the crown, mideyebrow, forehead and back of the head and neck.
2. The eye and temple areas should be brushed down when the headache is because of eyestrain and then flushed into the ground.
3. The whole head, particularly the painful section should be brushed and then pale green/white, blue/white and pale purple/white Chi used to bring energy into the area.

Migraine Headache: *Light Whitish Green Blue and Violet Chi*

1. Persons who suffer from prolonged or chronic headaches including those types termed 'migraine' headaches have an abnormal accumulation of energy in various parts of their bodies. Starting with the solar plexus, this energy follows a route up the vertebral column and thence into the neck and head. As a result the various blood vessels in the head expand, causing pain in the head area. Mental strain and emotional disturbances usually cause this type of headache.
2. Use a pushing movement to clean and open the solar plexus and liver areas. First use pale green/white, then pale blue/white and finally mauve/white chi to bring energy to the solar plexus area.
3. In order to send the 'sick' red chi into the ground, brush down the spinal column.
4. The back of the heart can be opened by pushing through in order to remove the dirty Chi. Then bring energy to the heart by using pale green/white and purple. Use your mind to see the heart center expanding and opening.
5. Brush down the dirty chi from the head area and send it into the earth.
6. Use a combination of pale green/white, blue and purple to bring energy to the back of the head, the crown point and the forehead.

Toothache: *Light Whitish Green and Blue Chi*

1. Brush down the area with pale green/white Chi.
2. Energy should be brought into the area with pale green/white, mauve/white and blue Chi.

Broken Bone: *Orange-Yellow Chi*

For anything relating to the bones always activate the sacrum. Light orange-yellow Chi is used to encourage broken bones to knit quickly, Chi which is pale yellowish tangerine may be employed. The sacrum attracts the orange-yellow Chi; feel the eight holes in the bone breathing. The orange-yellow Chi enters, transfers up to the crown and is sent up to the universe to multiply.

1. The fracture area should be brushed using a combination of pale green/white and pale tangerine/white chi.
2. To promote fast knitting of fractures, a daily treatment with pale tangerine yellow Chi mixed with white can be employed, in the ratio of 7:1.5:1.5 of white (for the center), yellow and tangerine respectively. Push through and brush the sacral and navel areas, using a cerise (light red) Chi mixed with white.

Back Injury: *Light Whitish Blue and Green Violet Chi*

1. Scan the student to see which parts are congested.
2. Use chi that is pale white/green or pale white/tangerine and, selecting each in turn, with a brushing movement apply to the whole spinal area and especially to the area where the trauma is located.
3. To bring energy to the damaged area, use a combination of pale blue/white, green and purple.
4. White chi should be used with a brushing movement on both the back and front of the solar plexus area.

Infection and Inflammation: *Light Green Chi*

Treat the damaged area with green and tangerine/white chi, using a brushing movement. For the heart and spleen, these colors should be replaced with green and purple.

Bring energy to the area by the use of mauve, white, tangerine, green and blue and strengthen the natural immune of the body.

Cysts: *Light Green and Blue Chi*

During the course of the session, repeatedly pass your hands over the patient's body. Treat the solar plexus area with a combination of green and blue chi. The cyst can be removed using a chi knife, spiraling clockwise with green and tangerine chi; moving the 'stuck' red chi out of the body and grounding it in the earth. After, blue chi should be used for a calming effect, while green and tangerine can also provide energy. For maximum benefit, sessions should be thrice weekly.

As foods containing a lot of highly spiced ingredients have a lot of red chi, these should be kept to a minimum. The patient should use the Inner Smile to calm and balance his/her emotional state.

It is sometimes necessary to sweep 100-200 times, to cool the affected area. Scanning will enable you to take a temperature reading. If the area is still hot then continue sweeping until it reaches a satisfactory temperature.

Fever

1. Brushing can be done with blue or green Chi.
2. Push and pull through the affected part, sweeping the spleen with green and blue. Clean and rinse thoroughly.
3. The solar plexus area can be treated with tangerine and blue Chi but tangerine should not be used on a patient with diarrhea.
4. The thymus gland can be stimulated by the application of green, tangerine and purple Chi. First, brush the heart area with green and blue. This helps the body to ward off infection.
5. Both lungs should be brushed and energy brought into the back of the lung area by the use of tangerine and green.
6. Sweep the spleen with green.

If the student has a fever then they must lie down to receive the healing. This applies in all cases where the student is unable to sit up comfortably.

Insomnia: *Green and Blue*

Inability to sleep may be the result of too much activity in the solar plexus and the sacral area. The solar plexus may also be congested. Similar problems can also occur in the throat, mideyebrow and forehead, as well as at the crown point.

Repeatedly pass your hands over the patient's body as the treatment progresses and use a combination of blue and green chi to clean the sacral and solar plexus areas.

Sleep can be brought on by use of blue, as it soothes the solar plexus area. Use blue and green to brush the mideyebrow, forehead and crown point and white and tangerine on the navel. Finish with blue.

Do not eat food late in the evening.

Sleep in a separate bed from your partner. You can place two beds next to each other.

Tinnitus

Relates to weakened kidneys. Work on the kidneys, mideyebrow and forehead.

Blood Clotting: *Blue Chi* - Fresh Burns: *Green and Blue Chi*

Because of the cooling and calming properties of blue chi, it can be used to remove the heat from burns. Green has a similar effect.

The area needing treatment should be scanned, then brushed with a combination of pale green/white and blue Chi until relief is felt and the heat in the cells released.

Old Minor Burns: *Green and Red Chi*

Pass your hands over the damaged areas and repeat this during the treatment period. Use pale green/white and tangerine, but do not use on the head, eyes or brain, or other sensitive organs.

To reduce discomfort in the damaged area and promote fast healing, firstly pale blue Chi mixed with white should be employed, followed by cerise mixed with white, and green in equal proportions as this mixture optimizes the breaking down of necrotic cellular material. However, cerise should not be used in the head area but be replaced by mauve mixed with white.

Use cerise/white chi to brush the sacral and navel areas and follow with blue which assists in the release of the stagnant Chi. This session should be repeated for 2 or 3 days.

Old Severe Burns: *Green and Red Chi*

1. Scan at the beginning and end of the treatment (Note:I do not think scan can be seen as plagiarism and I don't know what other word to use in this context as I do not have the full text)
2. The local area should be brushed first with pale green/white and then with pale tangerine white. Keep repeating but make sure to omit sensitive organs and the head area.
3. To speed up healing and reduce pain and infection, the damaged area should be filled with energy from pale blue white, pale green white and mauve with white Chi.
4. To help speed up the healing process, wait a few days until there is much less pain, then use pale green white and cerise mixed with white Chi. If the head area is affected, use mauve mixed with white in place of the cerise.
5. Brush in the sacral and navel areas and to speed up healing use white Chi.
6. After making your healing intent stable, release it.

Old Wounds: *Green and Red Chi*

Green chi, in equal proportion of pale green/white and pale red/white should be employed on injuries which are not new as, when used in large quantities, it attacks and disperses damaged cellular matter, promoting fast healing. On recent injuries, it will have an adverse effect because it is likely to cause the trauma area to become moist, and it also will not be able to work quickly.

Scan at the beginning and end of the treatment, and brush the damaged area with pale green/white and pale tangerine/white Chi. Some pale blue/white can be employed locally to energise the area. Use of pale green/white followed by cerise mixed with white can speed up the healing process and assist in the breakdown of cellular material.

The navel and sacral areas should be brushed using pale cerise/white Chi. This will fill the area with energy and make the wounds heal faster. Before the Chi is let go, it should be stabilized.

Some Specific Points and Ailments

The following descriptions are mere guidelines. Each case is different and much depends on your own abilities. One needs practice to detect diseases. Best results with Cosmic Healing will be obtained when practiced in combination with Chi Nei Tsang.

Remember that these techniques are not meant to replace the doctor, but merely to assist him. The student must always work with the practitioner. ***The practitioner is a channel, the student the healer.***

Solar Plexus: *Blue Chi*

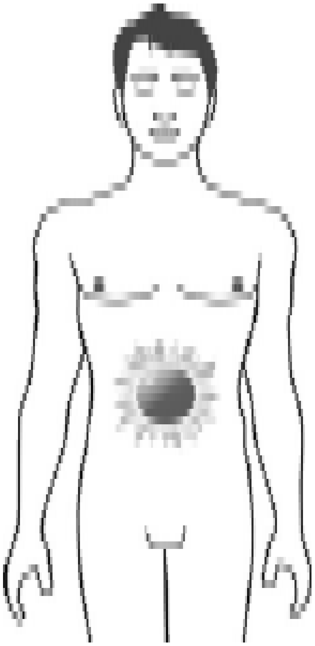

The Solar Plexus is the connection point for practically all the nerves inside the body. That is why it becomes laden with emotions so easily. When the Solar Plexus jams up and overheats, the liver will overheat; when this happens, the heart will also overheat; this will in turn affect the thyroid and parathyroid; finally the brain will be affected. Cleaning out this center is therefore extremely important. This type of cleaning may be done from the top down (as in the general cleansing session) or aimed at the Solar Plexus directly. Draw the green light, spiral it, flush with the green light, then energize with white and violet light. As before, always allow the energy to stream through both the points fully. The Solar Plexus holds all the emotions. When working on the Solar Plexus, the most important thing to remember is to connect the backside to the universe. There is literally no end to the 'other' end of this connection. Just allow the Chi to come all the way down and slightly pull inwards a little. Simply clean out the path. Allow the information to condense, allow any images to come out. Then stabilize the energy. Picture the Chi-field enveloping the person as a big protective bubble. Cool down with blue.

Skin Infection: *Blue Chi*

The skin is related to the lungs and the kidneys. Whenever there is an infection, the immune system is involved.

1. Enhance immunity as outlined in pages 101-108.
2. Focus on strengthening the lungs. Always start with green to flush out. Then send orange light into the lungs. Make sure you are not sending it up to the brain. Energize the lungs with blue and white.
3. Detoxify the kidneys, purify the blood (green) and energize with blue. Ask the person to visualize a fresh mountain-stream to stimulate the kidneys.
4. Treat the area locally, using your hands. If the area is big, use your palm, if small use the fingers. Do not focus on your palms - refer to the section 'Forget About Your Palms'. on page 152. Draw green Chi from the forest, and then energize with violet.
5. To diminish pain or itch, use blue to cool down. Also use the Chi knife to work on the infection.

In case the skin problem is allergy-based, you should also treat the mideyebrow; there might be some irritation here. First clean with green, then either energize with violet (if resistance is low) or soothe with a light shade of yellow. If the skin problem is related to asthma, open the lungs with bright red clean the throat center with green and energize it with (deep) blue.

If the problem is due to toxicity, also clean out the liver flushing with a combination of green and orange. Energize with green. Enhance the immune system by working on the sacrum.

Eyes: *Green Chi*

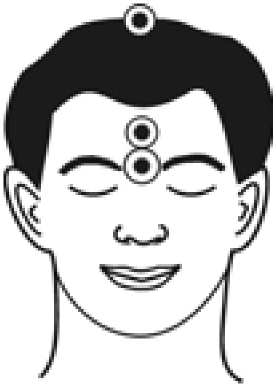

The right and left eye are governed by the forehead and mideyebrow respectively. The eyes are also linked to the temples. Ask the person to touch their temple bone and guide them to take in the (white or light violet) Chi. Send Chi through the mideyebrow and forehead (all the way through). Ask the person to visualize the light coming in. Use soft shades of green with golden yellow to stimulate the optic nerve. Use blue to calm the eye muscles and light violet to stimulate the inner eyesight.

The eyes are connected to the liver. Scan the liver and energize with green. Scan the jawbone to see if there is any unreleased anger. Cool down with blue. Channel bright green into the liver.

Ears: *Violet Chi*

The right and left ears are governed by the forehead and mideyebrow respectively. Push and pull green Chi to clean. The ears are connected to the kidneys, so you can work on them also. The ears are also connected to the sinuses and upper respiratory tract. This means that infections need to cool down before you can use any cleaning technique. Use blue to cool and clean with violet. As long as fever and severe infection persist, use blue to cool them down. When clean, energize with violet.

Have Better Health, Less Stress and Remain Young

Brush specifically on the areas of the solar plexus, liver, stomach and pancreas.

White chi brings energy to the crown point, head (at the back) and both sides of the brain.

Both the front and rear of the heart and lungs can receive energy from white/purple Chi.

Both the front and rear of the spleen, the navel and lower abdominal area, the spinal cord, the kidneys, using white Chi.

On the sexual center use white chi but if the problem is difficult to shift, use green. You can also use white Chi on the sacral area, and green or purple on other organs.

White Chi can be employed for the arms and hands, while the legs and feet should be treated with white and purple Chi.
Eye problems with green/yellow energy.

After a session, patients who are particularly weak should not bathe for one full day. The energy which has been put into their bodies can thus be kept there by application of pale blue/white Chi at the end of the session.

Fig. 6.18 *Use Compassion, Love and Kindness.*

Summary of Individual Healing Session

A. Chi Knife
B. Energizing and Activating the Immune System
C. Cleansing with the Internal Organ Techniques

General Outline of Practices

1. Gain experience by working with one person. **Stand in front of them with your feet shoulder width apart.** Hold your hands at your heart in salutation. Transmit your love and compassion and ask for their permission to help them to help themselves heal.

2. Connect to the student's and your own cosmic star above your head and ask for universal permission. **Radiate your love and transfer three minds into one mind**. Lower your Upper Tan Tien. Activate your consciousness: Three minds become one mind. Expand your awareness in the mideyebrow and the crown. Feel the Tan Tien, heart, mideyebrow and crown spiraling. Expand to the universe.

3. Visualize the person's legs as being 'very long', extending down to the earth. **Activate the Sacred Fire** and bring it down. Clean out the room and make a big Chi field around you and the student.

4. **Call the Elemental Forces:** water, fire, thunder, lightning, rain, lake, earth, mountain, wind and Heavenly Chi. Call the guardian animals; Blue Tortoise from the north, Red Pheasant from the south, Green Dragon from the east, White Tiger from the west, Yellow Phoenix from above and Black Tortoise from below. Feel that you are protected by the Chi; gather the power of all the mountains, rivers, streams, stones and rocks into the room. Activate them with lightning, thunder and wind - filling the whole room with electrified Chi.

A. Chi Knife

1. Hold the knife in right hand (left, if you are left handed). Hold the other hand near your body with the thumb and index finger touching. Project the Chi Knife as big as the universe, 'long and big - reaching to Heaven'. The bones in your arm are hollow and the 'bottom' is sealed at the shoulder. Feel them 'fill and pack' with the power coming from above.

2. Then give the command: **"Power comes from the east. This Chi Knife has the power to dissolve all the negative energy, sickness and bad fortune. By my request, please carry out the order now."** Wait until you feel that your arm and the knife are heavy and full of Chi.

Carry out energetic surgery, **cutting and slicing the affected area with cool yin energy**. When you feel that the Chi Knife is full of sick Chi, spiral counterclockwise and discard it down into the ground. When you feel that the knife is 'empty', charge it with Chi from the universe. Continue until you feel that the area has been cleansed.

B. Energizing and Activating the Immune System

In cases of severe illness, debilitation and infection, it is essential to stimulate the body's defence systems. You may also want do this in a preventative manner. Guide the student as you practice together. Presuming that you have already done the previous set of general healing, you can start with the coccyx.

1. Be aware of your Tan Tien and connect to the universe. Touch your **coccyx** with the fingers concentrating a beam of white or yellow light on to it. Picture the sacrum opening and the light streaming inside, flowing **up the spine** and streaming out over the whole bone structure. Feel the Chi rising up to the universe, multiply it and guide it down to your energy body.

2. Be aware of the Tan Tien and the universe. Attract the yellow light; spiral and condense the healing light into the tips of your fingers; move the fingers up to touch the **sacrum**. Send out a beam of yellow Chi into the sacrum, directing it up the spine. This will stimulate the production of red blood cells. Transfer the feeling of 'Bone-Chi' up to universe and down to your energy body.
3. Proceed to the **crown**, showering it with violet light. Then spread from the crown throughout the bones of the body. Repeat several times. Instruct the student to feel the bone structure breathing. Ask him/her to extend up from their crown to the Big Dipper and North Star, channelling down the red and violet light into the thymus gland.
4. Continue to the **throat**, stroking down with your palms, using green color to move the lymph down towards the backside of the heart. (Never stroke up towards the head! Only guide the lymph down).
5. Proceed to the **lymph nodes under the armpits and across the sternum**. Next activate the lymph nodes in the **groin**. From the groin you guide the lymph up to the navel center (Not down where it will get stuck in the legs).
6. Return to the **navel center**. Send the energy deep inside, energizing the lymph nodes at the back. Pull back, drawing green energy, flushing out any waste products that are stuck in them.

Activate the points at the inside of the **elbows** and the **knees**.

Shower the whole body with healing light and energize the **perineum** with blue Chi.

Always remember to ground any sick energy by sending it deep into the ground.
Ask the student to do the same and it will be more effective. You should do this while giving the session as well as when closing it.

C. Cleansing the Internal Organs
(Repeated from pages 132-133 and pages 136-137)

This session can be used for any kind of ailment and carried out with any number of students. **It is important to remember that if there is no energy in your Tan Tien, then there will be no energy to work with.** When you are aware of your Chi, you can direct heavenly light straight into your energy body and then into the student's.

General Cleansing Procedure

1. Push and pull the energy through the student's body. Push and disperse the sick energy into the universe and out to the planets, where it will be transformed and recycled.
2. Pull the energy back and stop it between yourself and the student.
3. Spiral the energy counterclockwise until you see the **green** light mix and clean the sick energy. Spiral and flush the energy down into the earth.

4. Spiral with blue light; see it mix and flush the energy within the student's body. Spiral this energy down to the ground.
5. Energize the student with **white** or **violet** light.
6. You can also use both hands to channel Chi when you 'push and pull' through the body.
7. **Spiral in** - spiral clockwise to bring energy into the body and energize it.
8. **Spiral out** - spiral counterclockwise to cleanse, flush and remove energy.
9. When you begin pushing, pulling and sweeping, you may feel resistance in the body. As you progress and the energy centers open, this feeling will diminish.
10. There are many color combinations that can be applied to the various parts of the body. When you start practice, stick to the basic 'blue, green, white, violet and blue' formula. When you are working on a specific area, you can keep this book near by, or write a short 'color recipe' before you begin. In this way, you will be able to relax your mind.
11. 'Pushing and pulling' is like fanning and venting the body; when you fan your hands in front of your face, it will have a cooling effect. Ventilation enables the stagnant, sick energy to leave the body.

    Below is a list of the body's organs in relation to their tendency to 'store heat'. They are listed in descending order, i.e, the head is the hottest.

    Head, heart, liver, lung, stomach and spleen, large intestine and small intestine, kidneys and sexual organs (which become hot when aroused).

Outline of Practice

The **solar plexus** is the clearing house for emotional energy and is situated in between the lower and higher energy centers and in the center of most of the internal organs. All the organs dump their negative emotions here. The internal organs can be cleansed by sweeping the solar plexus, energizing, cleaning and flushing it of sick energy. When you feel sick energy in the body, give the command: "Stay". This is very important; it makes sure that none of the cells hide from the dazzling light that you are sending in.

1. Scan; ask for the power to look at the cells.
2. Push, pull and give the command "Disperse". Push until you feel the solar plexus clear and open.
3. Spiral green Chi counterclockwise. Talk to the sick energy; tell it to "Leave", you will take it to "a better place, where it will be happy". Give the command "Out".
4. Spiral Blue Chi counterclockwise, flushing out any remaining sickness. Guide the sick energy into the earth.
5. Spiral and sweep Blue - Green and Red Chi into the solar plexus, sweep the organs. Do not use in cases where there is intestinal infection and internal bleeding (which can be worsened) or in the presence of pregnant women.

Cleansing the Spinal Cord

Always cleanse the spinal cord, which protects the entire body from diseased and negative energy, both physical and psychic. When you open the spinal cord and solar plexus you can release many energy blockages. Sweep the hand in a clockwise motion to energize the spine with either/or a combination of: blue, green, white and ultra-violet. Then spiral counterclockwise and brush the spine down to the ground. Continue sweeping until you feel that the spinal cord has cooled down. Remember not to use orange or red

Cleansing the Blood (Lungs)

The blood can be cleansed by energizing the lungs with whitish green, then with light whitish orange. The blood passing through the lungs will absorb the green and orange Chi, thereby cleansing the blood plasma and vessels and the rest of the body. This technique is very useful for treating ailments of the blood and of the arteries and for severe infections. Do not use it on pregnant women.

Outline of Practice

1. Scan the student.
2. Push through and open the lungs from the back to the front.
3. Energize the lungs from front to back with light whitish green, then with light whitish orange. Push through.
4. If the student is weak, energize the lungs with light whitish red, which will provide a strengthening effect.
5. The **spleen**, **liver and kidneys** purify the blood. Push and pull through the organs and energize with white Chi.
6. The blood requires a lot of cleaning on a regular basis.
7. Energize with white Chi. **Germs can only grow in the dark. 'Fill yourself with the light'.**

General Principles

Energy can be introduced to the affected area of the body by moving the hand clockwise in a spiral. **Flush** means to return the sick energy to the earth for recycling.

Stabilize means to return the area to normal functioning.

Spiral in a **clockwise** motion when you are '**stabilizing**'.

Use **blue or white** light to energize the area that you are working on.

Use **green** light to **energize** and **clean** the area that you are working on.

Use **blue** light to **flush** and **stabilize** the area that you are working on.

Use **violet** light to **balance** the sick energy and **reprogram** the cells of the body.

For **general infections** always **use green** (energize and clean) and **blue** (stabilize) light. Use violet light to balance the sick energy and reprogram all the cells. Use the ‘Chi Knife’ for infections as well. Never use orange on the head, heart and spleen.

Give the command **“Stay”** to make sure **all the cells receive the light** you are sending in. Remember that germs and bacteria cannot hide in the light.

Energize means to draw the Cosmic Chi into the part that requires healing.
Cleanse means to clean all the cells being treated.

You can also work on yourself. You may need some help when working on the back. If you obtain optimum health and practice the Cosmic Orbit regularly then you will never ‘need’ to apply any of these techniques to yourself.

Like all Chinese medicine, Cosmic Healing works on the system as a whole, balancing the entire organism, treating the cause to cure the symptoms.

The techniques really are simple.

When in doubt use; blue, green, white, violet, blue.
You can always use white and violet.
Upper organs - use violet.

You can practice group healings and then link your students’ personal stars together, at an agreed time later in the day. **You can then perform the World Link Cosmic Healing**.

Work on any other ailments that the student may have.
The simplest way to learn is to begin practicing today.
The simplest way to heal is to allow yourself to be healed.

Chapter VII
Different Hand Techniques

Forget about your Palms

An extremely important theory of Cosmic Healing Hand Techniques is to 'forget about your palm'. This means that you should only use the palm (which is incredibly sensitive), for 'guiding' the universal energy into the right spot, sending out a minute beam of light, like a laser, that marks the place where the energy should be sent. Your palm is connected to your brain, which has been sunk into your Tan Tien. The light marker will serve as a beacon for the force that is directed down by your mind. Draw down the Universal Chi, guide it to the right spot and give the command; "Stay". This is very important, otherwise the Chi will disappear once you move your hand.

When you practice Cosmic Healing it makes no difference how many people you treat, because you are not transferring your own energy.

Don't Focus your Mind on the Sick Part

If you direct your attention towards something, your energy will go there. If you focus on some ailment, you are already sending out your energy. Instead you should try to connect with a greater source, Universal Chi, extending your mind very far away. This way you won't lose your power. The only thing you do as a practitioner is to form the connection between nature, the student and the sick energy or between the universe, the Chi field and the sick part.

Cosmic Healing is Teaching

One of the best methods of protection for a practitioner is the Fusion practice (*Refer to 'Fusion of the Five Elements I by Mantak Chia*). Perhaps even better is for the practitioner not to perceive him/herself as a healer, but as a person that comes to guide people in something they have to do themselves. When you say; "I am a healer", the sickness or bad energy, has already found a new home. When you check it out, first it will try to hide, and then it will try to attach itself to you. But if you don't interfere, the energy will go straight to the ground. So it is very important to clarify that you are not a healer, but a practitioner that can help, and that the student or sick person has to do most of the job.

Fig. 7.1 *The Stronger your Awareness of your Inner Body, the Easier it will be for others present to pick up this Sense.*

If you guide those present, this will greatly enhance your healing session. Sit or stand facing your student. Any time you want him or her to focus on a particular point of the body, ask them to touch it with their hands. For example, "Now place your hands on your navel", while you do the same thing yourself. As soon as you feel the Chi there, you instruct them to focus on the navel themselves. This way you are actually helping them to deepen their bodily awareness. The stronger your own awareness of your inner body, the easier it will be for the others present to pick it up. This means that if you are treating some part of another person's body, or if you see that the person is not aware of any energy in some part of the body, you can help by focusing your awareness on this particular part of your own body.

Chi Kung Diagnosis: Hand Scanning

Hand scanning is a simple and effective diagnostic tool of External Chi Healing. It involves passing your palm over the body of your student from one inch to one foot above the surface of the skin and being aware of the energetic state of the various bodily areas. What one senses is described as the electromagnetic field, the aura or the energy body.

General Variations in the Energy Field

The variations you may sense with scanning that have diagnostic significance are:

Temperature

Heat generally indicates excess, while cold indicates deficiency. However, some areas of the body are naturally warmer or colder than others. This is described below under "Hand Scanning the Internal Organs."

Thickness

A feeling of thickness over the area would indicate excess, while thinness may indicate deficiency.

Wind

You may feel sensations of wind leaving the body at various places. This may indicate internal wind at a given location, or it may point to an area of the body that is "leaking energy" and needs to be ‘patched’ or ‘sealed’. It may also be a positive sign that you are driving wind out of the body.

Peaks and Valleys

Peaks in the energy field will feel like mountains or hills in the aural landscape. They may actually push your hand away. Valleys feel like depressions or energy vortexes drawing you in. They may also feel empty, like a hole or pit. Peaks indicate excess or stagnation, while valleys indicate deficiency.

Tingling

Accumulations of sick Chi may cause your hand to tingle, feel prickly, throb, or even hurt. These are excesses of energy and upon inquiry, the student may complain of pain in this area.

Hand Scanning the Internal Organs

Each Organ Emits a Different Aura

Each organ emits a different kind of force or aura through the skin. By passing a hand one or two inches above the skin, you can feel different sensations that reflect the condition or state of the internal organs. You need to develop the sensitivity to receive and identify the vibration or frequency of each organ. Practicing the Cosmic Chi Kung Meditations will help develop such sensitivity.

Liver and Gall Bladder Scanning

Healthy liver and gall bladder energy feels warm.

Negative Emotions: When you pass your hand over the liver, under the right side of the rib cage, you will feel a charged energy come up to your hand. This is a sign of anger in the liver.

Overactivity: When you pass your hand over the liver and feel a rush of hot energy, this indicates that the liver is overheating because of toxins or emotional stress.

Underactivity: When you pass your hand over the liver and you feel a dense and hot energy, the liver is weak, congested and sick.

Lungs Scanning

Healthy lung energy feels cool and dry.

Negative Emotions: Pass your hand over the lungs. Do the Lungs' Sound and listen to the echo of the sound as it rebounds from the lungs. Sadness will feel like a deflating ball pressed between your hands.

Overactivity: Energy that feels dry and hot indicates an overworked organ.

Underactivity: Energy that feels damp and cool under the scanning hand indicates underactive or congested lungs that can lead to respiratory problems.

Heart Scanning

Healthy heart energy feels warm and energetic.

Negative Emotions: Hot and charged energy indicates impatience, hastiness and arrogance in the heart.

Overactivity: If your hand detects hot, charged and overly expansive energy, this indicates that the heart and blood may be overheating.

Underactivity: Energy that feels cool and less expansive indicates an underactive or congested heart.

Spleen Scanning

Healthy spleen energy feels lukewarm.

Negative emotions: Energy that feel damp and sinking indicates excessive worry.

Overactivity: The energy feels hot and damp when the spleen is overactive.

Underactivity: The energy feels cool and damp.

Kidneys Scanning

Healthy kidney energy feels cold, but not too cold.

Negative emotions: Energy that feels cold and chilly indicates fear.

Overactivity: When the kidneys are overworked or overstimulated by excessive exercise or improper diet and liquid intake, the energy can feel damp, stick and hot.

Underactivity: When toxins are blocking the organs, the energy can feel damp and cold.

Hand scanning is an art, and may take a while for you to feel confident using it. Practicing Cosmic Chi Kung is one of the best ways to develop greater sensitivity in hand scanning.

Chapter VIII
Cosmic Chi Kung
Palm and Finger Training

Why We Practice the Cosmic Chi Kung

The Cosmic Chi Kung is also known as Buddha Palm. The reason we train with the Empty Force practice in connection with the Cosmic Healing is to learn how to project Chi through the space in the cosmos, the space between your hands and to project Chi through the space to your students. Usually when we touch the student we pick up sick energy. This practice will teach you how to ground the sick energy to the earth and to disperse it into the Universe to be decomposed and recycled by the planets.

Practice until you can project Chi out from your palms and fingers and when you feel the Chi coming into your body through your hands. The most important part is to always remain connected with the Universal Force coming from all six directions. If you focus on healing from your hands or your Tan Tien you will use up your own energy. You must be connected to the Cosmic Universal Force.

Expand your awareness to nature, the oceans, the lakes, the forests and mountains. Smile to nature and feel it smile back to you; inhale and draw the Chi into your palms. Expand your awareness to connect to the light, the Milky Way and the cosmos.

Be aware of the heart and the red light in the heart; expand your awareness to infinity. The light will come close to you; picture the red planet Mars above you. See it shine down to your crown, feeling the light in your palms. Be aware of the six directions, feel your body growing bigger; until you touch the sky, your feet still planted in the earth.

Opening Heaven and Earth Force in the Palms with Heaven and Earth Spirals

The purpose of this technique is to draw in the forces of heaven and earth through the palm. Activate the spirals, the symbol for the heaven and earth force in the center of the palm.

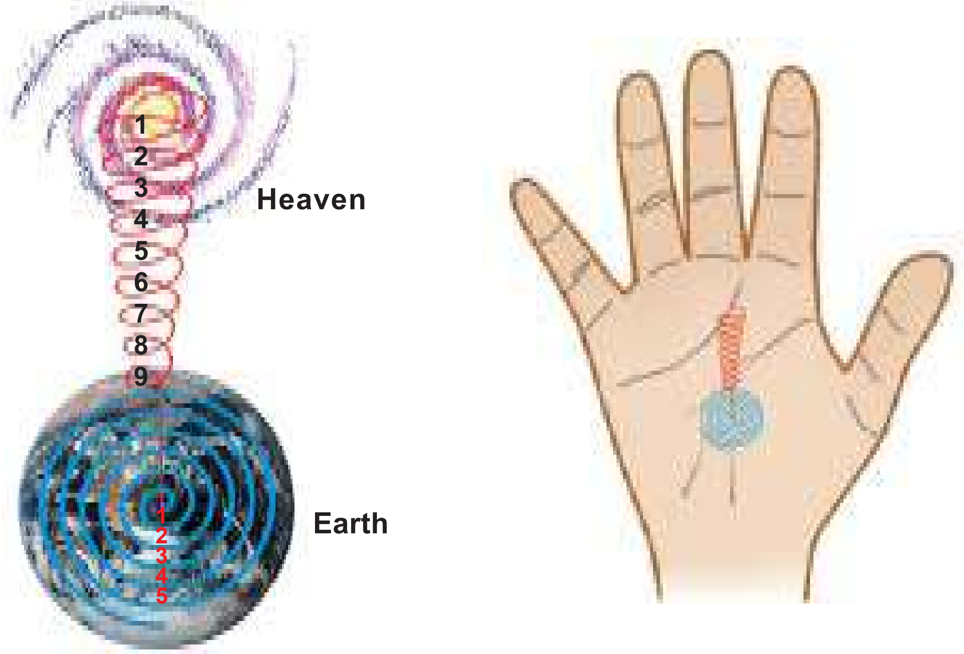

Fig. 8.1 *Enhances the symbol of Heaven and Earth Force, in the Laogong - Pericardium 8.*

This is accomplished by drawing these spirals with the palm itself. As you draw the spirals, you should actually feel the palm breathing, inhaling these forces in. In the beginning, however, you may only be able to use your own mental imagery and imagine that you are absorbing these forces. In time the feelings will replace your imagination.

You breathe in Heavenly Chi when you draw the heaven spirals, and breathe in Earthly Chi when drawing the earth spirals. You may draw in both in one session; you may also choose just to draw in either Heaven or Earth Chi if you feel you need more of one or the other. It is important to balance these energies whenever you do this exercise.

1. Begin by raising the right hand to about shoulder height with the elbow sunk and the palm facing forward. Slightly pull back (open) the index finger, and slightly stretch the thumb forward and down. This will activate the Laogong point on the palm.

   At the same time, touch the tips of your left thumb and index fingers together. This will allow the right hand to draw in the Heaven and Earth Chi. The left hand will then be held palm up to draw in the heavenly force, or held palm down to draw in the earthly force.

2. Begin **drawing the heaven spiral** by moving the **palm in a clockwise motion**, with the circles getting smaller and smaller, spiraling inward. At the same time, your left hand is held at your side, palm up with thumb and index tips touching each other. Draw seven or nine (planets of our solar system) small spirals with your right hand in this way, drawing in the violet heavenly energy.
3. Next, lower your right hand slightly and turn your left palm down to face the earth. Then begin **drawing the earth spiral** by moving the right palm in a **counterclockwise motion** with the circles getting larger and larger. Draw five spirals (five elements), drawing in the blue earth energy as you move. Inhale into the palm, feeling the energies.
4. You may now bring both hands down to the beginning position of Cosmic Chi Kung, or you may open the Cosmic Channel to the planets for healing. This is discussed in the next section, Part I, 'Opening to Heaven, Earth and Cosmic Chi'.

Cell Breathing

By understanding the operations of our cells, we establish a foundation for our experience of Cosmic Chi Kung. Each cell is a living, breathing entity, absorbing energy with each inhalation and releasing toxins with each exhalation. Thus with every breath, we are creating more Chi.

We are made up of over one hundred trillion cells (one trillion is one thousand billion). Each one of these cells is constantly breathing. In order for our cells to function optimally, we must follow nature's teachings.

Fig. 8.2 *Each Cell is a Living, Breathing Entity.*

The first thing is to keep your cells free of toxins. Eat nutritional foods. This helps the Chi to flow smoothly through your energy channels and aids greater assimilation of vitamins and minerals. Learn and apply Five Element nutrition principles, eating according to the seasons and to your own constitution.

The second thing is to keep your cells free of excess emotions. Develop the skill of acting positively to situations, rather than reacting, especially negatively. Don't allow anyone to make you sad and unhealthy. Some people like to say, "You make me angry," or "You made me worry." We cannot allow other people to make us so emotional. We do not need to accept the negative emotions that other people try to dump on us. Cellular and psychic boundaries are related. Negative emotions disturb the cells' ability to absorb Chi and to generate the energy necessary for healing. We need all the Chi that we can get. Learn and apply the Six Healing Sounds, the Cosmic Inner Smile and the Fusion of the Five Elements meditation to transform negative emotions into personal power.

Third, create more energy by cultivating a happy attitude, so that your cells can regenerate more easily. **Smiling and loving** are the two quickest ways to make more Chi. In order to have an ample supply of Chi, develop an attitude of **joy and love** in yourself and towards others.

Mind, Eyes, Heart and Intention (YI)

The opening of the mind, eyes, heart and use of intention is vital for further activating and increasing your Chi. In the Tao we achieve this by cultivating our "Yi" (pronounced "yee"). "Yi" means mind power, intention, awareness and concentration. The Yi leads and guides the Chi. With our mind we control thought patterns. With our eyes we control the senses of sight, hearing, smell and taste. With our heart we control all our organs and their related emotions: the kindness/anger of the liver, the joy/impatience of the heart, the openness/worry of the spleen, the courage/sadness of the lungs and the gentleness/fear of the kidneys.

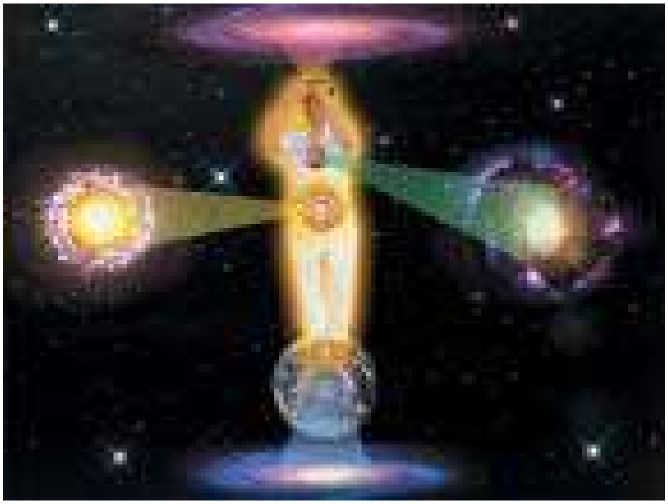

***Fig. 8.3** 'Yi' leads and guides the Chi.*

Every excellent achievement in our lives depends upon the quality and efficient operation of our Yi. Combining the awareness of the mind, eyes, heart and intention into one, causes the creation of a rich and rare alchemical mixture. The power of the Yi is the magical catalyst that assures energetic results.

The basic teachings are all in the Microcosmic Orbit, the Cosmic Inner Smile and the Six Healing Sounds. As long as you are alive and healthy, you are able to draw in energy from the universe and nature. The stronger the battery, the better it is able to store this energy. If you're weak and sick, you have no way to draw any energy from nature or the universe. When the cells stop drawing in energy, life ends. If you have perfected these practices, you have all the personal physical, mental and spiritual power that you need. Everything comes together. Your mind is aware of the original force. From the knowledge of Oneness, you are inwardly aware of your mind, senses and heart; you are outwardly aware of the universal energy. Now you are capable of receiving the abundantly available healing power of the universe; you can tune into everything, inwardly and outwardly in all directions. This is what we are here for: to heal and become whole again. We have higher goals to reach. The foundation for reaching our goals is a strong and healthy body.

Cosmic Healing Chi Kung is merely an extension of the universe within yourself. You draw the universe in through your palms, skin, heart and crown. With your Yi you breathe in the cosmos through these various open cavities of receptivity in your body, absorb its power, condense it, transform it and use it for the benefit of all.

Connecting the Bridge

Contracting the muscles of the perineum, sexual organs and anus activates our connection with the earth energy. By pulling up these areas and drawing in energy through the soles of the feet, we immediately become grounded and bring the earth (yin) energy into the Microcosmic Orbit. Pulling up should be done gently and directed by the Yi.

When we speak of "Connecting the Bridge," we are referring to the above exercise, except that we do not pull up on the perineum. Pull up gently and lightly on the sex organs and anus while leaving the perineum totally relaxed. After performing the first contractions, when you then relax the perineum, the earth energy will continue to flow into the body using the principles of a siphon. We thus connect the "Bridge" across the perineum from the sexual organs to the anus using the Yi to direct the energy. This actively combines the earth energy and your sexual energy and directs it into the Microcosmic Orbit. This energy is circulated during Cosmic Chi Kung. Holding a very gentle contraction will keep you grounded during the exercise. It can be done sitting or standing.

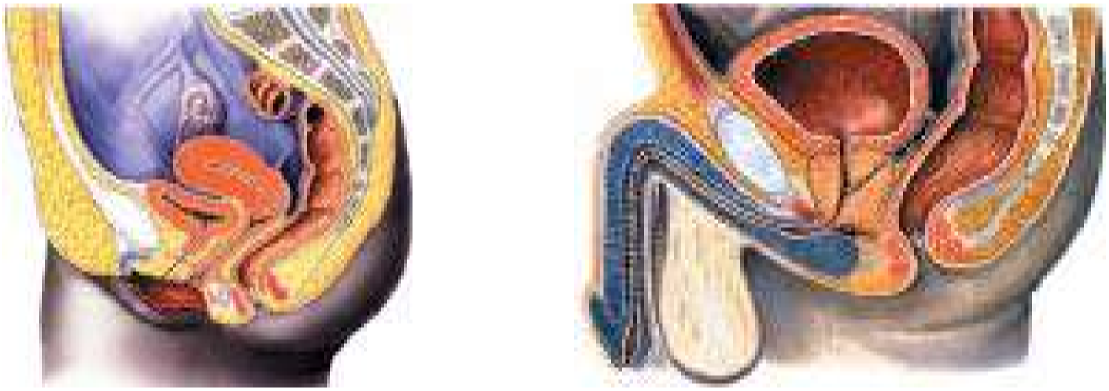

Women lightly contract the Vagina *Men lightly contract the Perineum and Anus*

Fig. 8.4 *Connect the Bridge - combine Earth and Heavenly Energy.*

Cosmic Chi Kung Preparation

Cosmic Chi Kung can be practiced either in a sitting or a standing position.
Sitting, relax as much as possible and invoke the **Cosmic Inner Smile**.

1. Be aware of your **heart** and listen to your **kidneys** with your inner senses. Smile to your heart and kidneys and feel them pulsing; feel them communicating and interacting with each other. The **Kan** and **Li**; the water of the kidneys and the fire of the heart, balance and **begin to mix**.
2. Open the arm pits as if you were holding a ping pong ball there and slightly move the fingers to activate the Chi flow. Relax. **Feel** the sensations of the **Chi flow**: tingling, warmth, pulsing, electric and magnetic feelings.
3. Be aware of your palms. Draw in Chi by **activating the bone breathing** process. Become aware of the perineum; slightly pull up the perineum, sexual organs and anus and "**connect the bridge**." (Remember to relax the perineum).
4. Breathe into the palms of the hands and soles of the feet, further enhancing your connection to the earth force. Smile to the blue earth energy as it flows up your arms and legs and suffuses your body with pleasant Yin energy.
5. Breathe into the **mideyebrow** point and activate your connection with the **Cosmic Force**. Smile to the golden light of the cosmic energy as it swirls into your mideyebrow point and flows throughout your body, healing and balancing your energy.
6. Breathe into your **crown** point and activate your connection with the **Heavenly Force.** Smile to the red and violet light of the heavenly force as it flows into your brain and washes your body with subtle pure Yang energy.

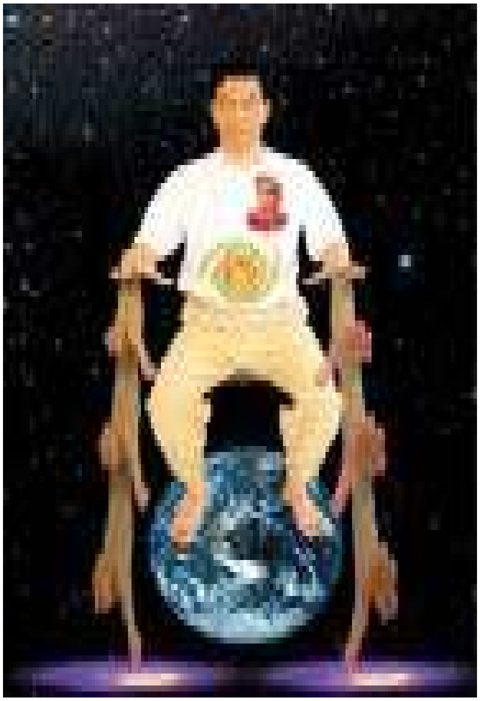

Fig. 8.6 *Invoke the Cosmic Inner Smile.*

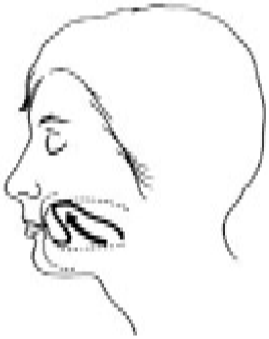

Fig. 8.5
Place the Tip of the Tongue on the Upper Palate.

7. Place the tip of your tongue on the upper palate behind the teeth in a comfortable position.

8. As you begin to move through the postures of Cosmic Chi Kung, always be aware of the Chi flow and its attending sensations.

Cosmic Chi Kung Preparation Standing

1. Stand with your **feet parallel** and shoulder width apart.
   Tilt your **sacrum** and **pelvis** slightly forward until you feel your feet press more firmly into the ground. As you tilt the pelvis, feel your lumbar vertebrae pressing outward; this is called "**Opening the Door of Life**."
2. Relax your **chest** and sink your **sternum**. Draw your chin in slightly, and hold your **head, neck** and **spine** erect as if your head and spine were suspended from above by a string.
3. Follow the same instructions for the preparations of Cosmic Chi Kung.

Opening to Heaven, Earth and Nature

The natural world is the source of our power. Cosmic Chi Kung teaches us about nature and its energies. In the Tao, nature and the universe are equivalent to God. You are a part of nature and can easily learn to open to the forces in the macrocosm and let them flow through you. This is very simple. Just tune into the frequencies around you. Your fingers are antennae, which transmit and receive.

1. To begin your practice of Cosmic Chi Kung, simply **stand** up and **feel the energy around you**. When you stand or sit, be aware of your soles (Bubbling Spring points) making a good connection with the earth. Your perineum (Hui Yin point) is relaxed and open. You are connecting to the earth force and can expand further down on to the other side of the universe. This is one type of energy.
2. When you draw your **chin in**, slightly **relax your chest** and tilt your head slightly forward; you begin to be aware of the heavenly force. This is another type of energy.
3. The qualities of the **Five Elements** reveal themselves as woody trees (growing force), fiery deserts (expanding force), earth mountains (stabilizing force), metal air (contracting force) and watery oceans and lakes (gathering force). We call the combination of these elements the '**Cosmic Force**'.

Thus, your mind and body learn to gather, absorb, direct and transform the heaven, earth and cosmic energy for your own use. This is the essence of Chi Kung. Today there is a lot of emphasis on learning movements and forms of Chi Kung. However, if the inner feeling is not there, the movements are of little value.

Heavenly force is Yang. Earth force is Yin. By learning to connect to the energies of heaven and earth, you have a powerful tool for restoring and maintaining your inner balance of Yin and Yang.

Outline of Practice

Before practicing each of the Cosmic Chi Kung sets in this chapter and in the following chapters begin with the heaven and earth spirals as outlined on page 157.

First **Create the Chi Field**.

Hold each of the Cosmic Chi Kung postures for at least 5 - 30 seconds, gradually lengthening the time to 60 - 120 seconds per posture.

Note: While holding the postures count the breath from 5 to 15 seconds, being aware of the **Tan Tien** and feeling the Chi in the Tan Tien which then moves up to the **crown** and expands up to the universe; fuse your Chi with the Universal Chi and let it multiply down to your palms. **Just concentrate on the 'Tan Tien and the Universe'**; do not concentrate on the palms at all.

Throughout parts I, II, III and IV 'Tan Tien and Universe' means to spiral into the Tan Tien and spiral out into the universe connecting 'Tan Tien and Universe'.

Part I
Opening to Heaven, Earth and Cosmic Chi

Opening Movements

Three Minds into One Mind

Empty the mind and heart down to the Tan Tien.
Hold the hands together at the heart center.
Smile and make the heart feel soft; feel love, joy and happiness.
Turn your consciousness inward and your awareness out.

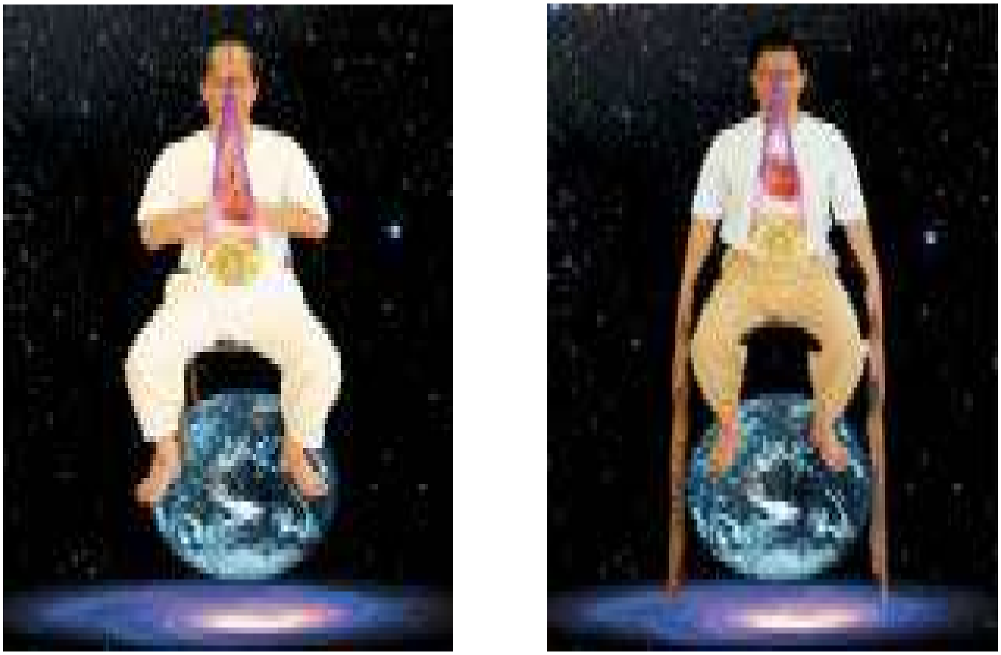

Fig. 8.7 *Empty your Mind and Heart down to your Lower Tan Tien.*

Feel your legs lengthen extending down to the center of the earth.
Feel your hands lengthen extending to the universe below.
Let the mind expand and be aware all the way down to the universe.
Feel the 'long' bones of the arms and legs and begin bone breathing.
Feel universal Chi filling, packing and compressing into the bones.
Draw this Chi down to the back of your crown and feel a heavy pressure press on the crown; feel slightly numb or a light electric downward flow, like oil dripping down. 'Focus at Tan Tien and Universe'. Count to five.

Channeling the Earth Force - Washing the Bone Marrow

Smile into the **perineum, palms** and **crown**. Then slowly raise the arms, palms facing each other, to chest height keeping the elbows relaxed and sunk. Rotate the arms slowly until the palms are facing down. Be aware of the mideyebrow and feel your breath and lightly contract your eyes and round muscles around the eyes. Connect the perineum bridge by slightly squeezing the sex organs and the anus; do this a few times.

Rest and **smile to soles of the feet**, feel the Chi bubbling and feel like you are standing on a hot spring starting to bubble up.

Be aware of your Tan Tien and expand your mind down past the earth to connect with the galaxy below you. Multiply and bring your mind back to the Tan Tien. Gradually feel the Chi being absorbed through the whole body and absorb it into the bones and body like a **rising steam or mist**. Feel the earth force move through the center of the bones and enter the bone marrow move up the calves and thighs (femur bones) through the hip bones, spine, **scapula**, **arms**, **neck** and **skull**. Finally **swirl the energy around your brain. 'Focus at Tan Tien and Universe'**.

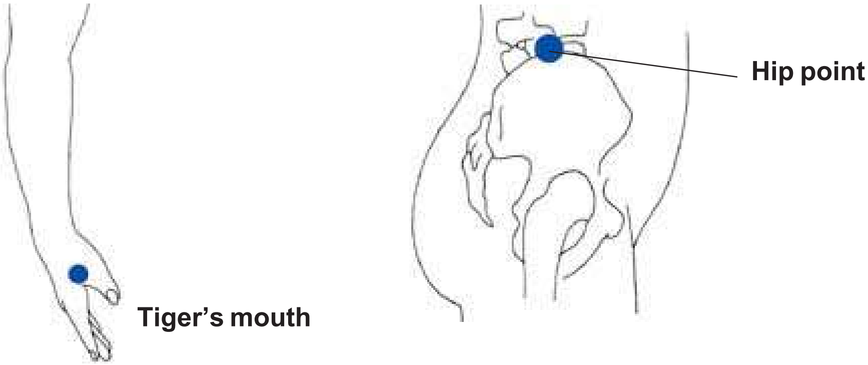

Fig 8.8 *Activate the Hip: Large Intestine 4, Pericardium 8 and Small Intestine 3.*

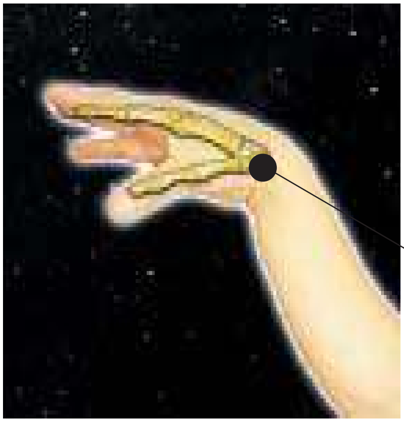

Raise your index finger slightly upward to activate the Hegu Point (Large Intestine).

Fig. 8.9 *Open the 'Tiger's Mouth'.*

Then **stretch your thumbs out to the sides and down**, so that they end up pointing downwards toward the earth. This activates Large Intestine 4 (Hegu). Large Intestine 4 is found in the webbing between the thumb and index finger. It is called the "eye of the hand"; this area is also known as the "Tiger's Mouth."

Slowly draw the elbows back and lower the hands until your palms are facing down beside the "eye of the hip" (the iliac crest) with the "eye of the hand" (Large Intestine 4) aligned with the "eye of the hip. "Be aware of the Tan Tien (Yi, the abdominal brain) and expand your mind out into space, the Cosmos and the universe. Remember: The fingers and the energy point act just like a laser to help guide the universal force flow; as soon as you focus on the fingers or the palms you are starting to use your own energy. Feel your Tan Tien and the crown full with Chi. Just use your second mind intention, lightly aware of the area between the hips and LI-4 points. Focus at Tan Tien and the universe. This will activate the lungs and large intestines. **Hold for a 30 second count** and gradually you will feel the ascending colon and the sigmoid colon have been activated; you may feel some movement around this area.

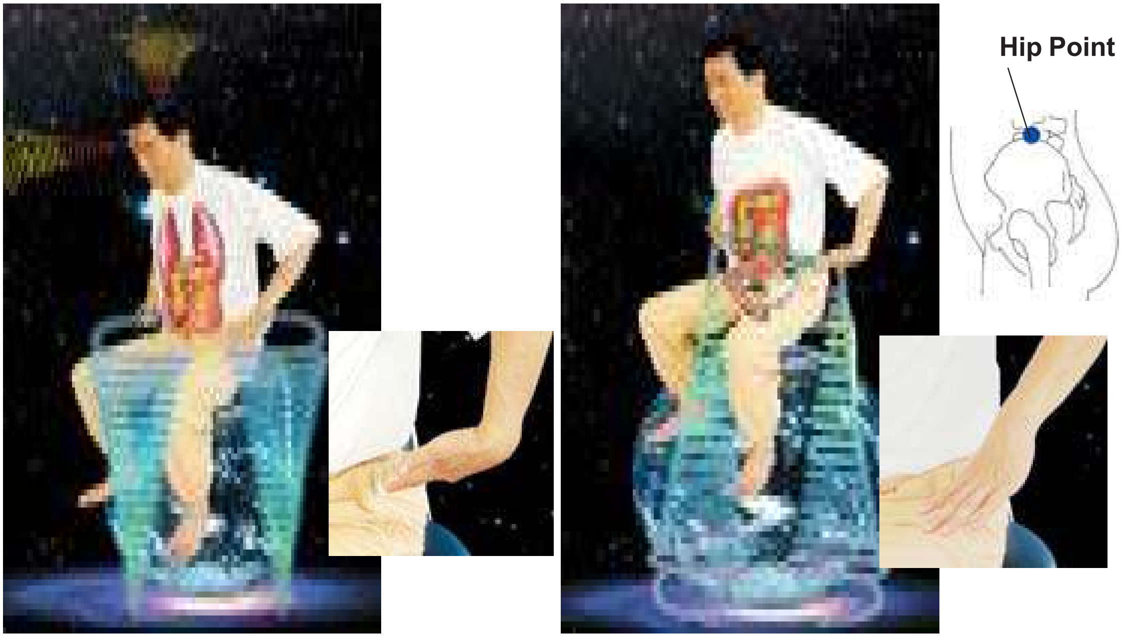

Fig. 8.10 *Activate the Organ Energy.*

Use your Yi to rotate your hands so that the fingers first point out, then back, turn the palms up and finally point the middle finger in toward the eye of the hip. Tan Tien and the universe - **95 % aware of the 'Tan Tien and the Universe' and 5 % aware of energy flowing back and forth between the tips of the middle fingers** (Pericardium 9); feel the Chi passing through the hips. This will activate the pericardium.

Rotate the hands again until the fingers are pointing forward with the **palms facing up**. Align the knife edges (Small Intestine 3) with the eyes of the hip. 'Focus at Tan Tien and the Universe'. Be lightly aware of the energy passing back and forth between the two hands. This will activate the heart and small intestine.

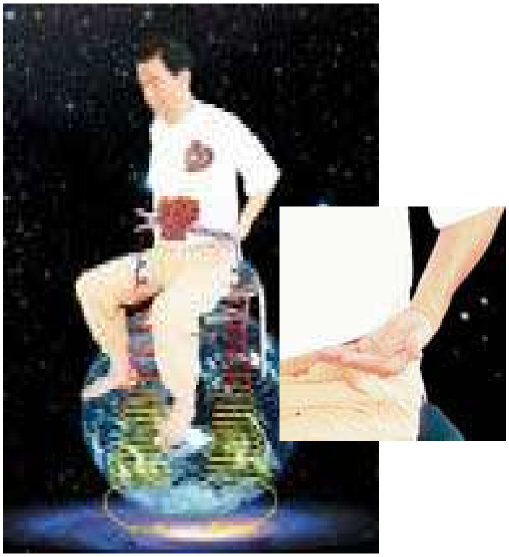

Fig. 8.11 *Activate the Heart and Small Intestine.*

Absorb the Heavenly Force

Washing the Bone Marrow - Crown to Soles

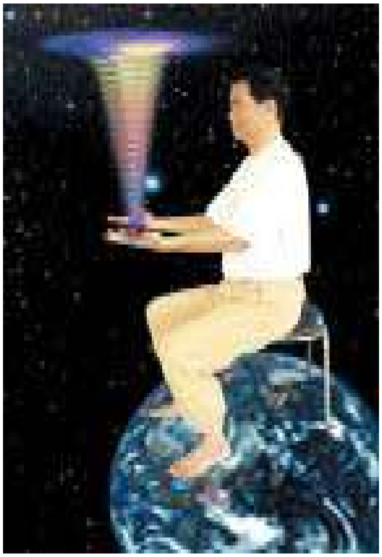

Extend the arms forward to chest height with the palms facing upward.

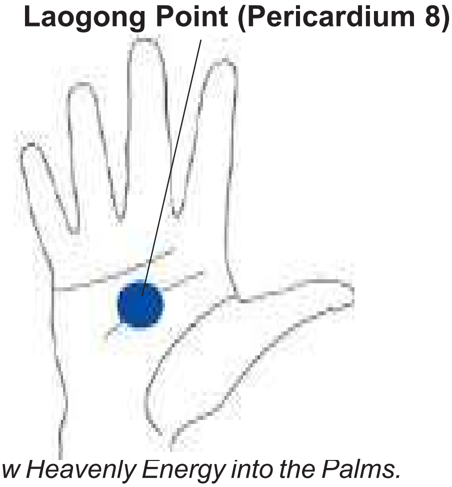

Fig. 8.12 *Draw Heavenly Energy into the Palms.*

Become aware of the **Tan Tien, the Crown and the Star above you**. Expand your mind (YI) out to the universe and connect to the galaxy. Be aware of the palms and the crown which are connected to the galaxy. Activate the Laogong point (Pericardium 8).

Feel the heavenly energy come directly to your palms, and be aware of the violet light of the North Star and the red light of the Big Dipper both shining radiantly.

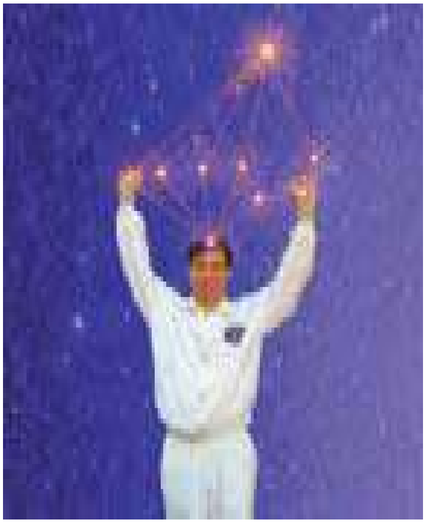

Fig. 8.13 Hold the Big Dipper.

Picture yourself holding the handle of the Big Dipper and turn the 'cup' pouring the light over your head. Draw this light in through the palms and the crown point and feel it washing your bone marrow. Always remember your Tan Tien. **Direct the sensations down through the skull, cervical vertebrae, clavicle, scapulas and sternum. Activate your thymus gland and continue drawing the healing heavenly light down through the rib cage, arms, spine, hips, femurs and calf bones, feet and toe bones. Be aware of the soles bubbling up to the perineum and up to the crown**.

Lightly rock yourself like riding a horse; this will help keep your spine open and Chi can flow easily throughout the practice.

Absorb the Earth Force and the Other Side of the Galaxy

Rotate the arms until the **hands are palms down**.

Raise the index fingers slightly and extend the thumbs first towards each other and then toward the ground. Be 95% aware of the Tan Tien and the Universe, and 5 % aware of the index fingers and the big toes. Wait until you feel the Chi enter. Fill the joints of the fingers and up to the wrists; feel 'tense but not tense' and let the **Chi continue to fill the joints** up to the elbows, up to the shoulders and gradually up to the neck and the head. Go from the toes up the legs to the hips and on into the spine and the rib cage **completing all the joints of the body**.

Be aware of the Tan Tien and the universe and only slightly aware of the palms (Laogong), the soles of the feet (Kidney 1-Bubbling Spring), and the perineum (Conception Vessel 1-Hui Yin) bubbling. Activate these points, feel the earth energy and continue smiling through the earth to connect with the galaxy, Milky Way, and the universe. Draw this light in through the palms and soles and let it wash up through the center of your bones from the feet up. Let it **steam and cleanse your marrow**. Picture any impurities or illness dripping out of your bones and draining down into the earth, where they will be recycled and purified.

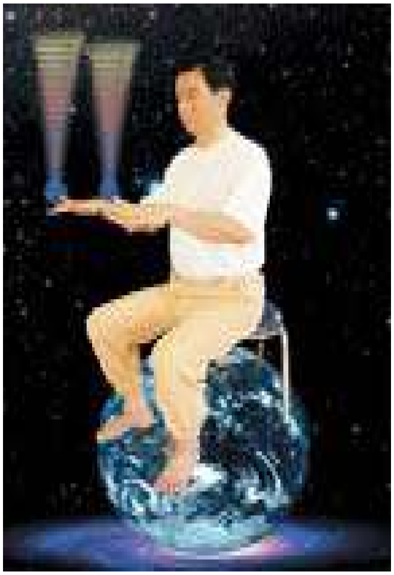

Fig. 8.14 Open the index fingers.

Core Movements of Part I

Grasping the Moon

Connecting the Heaven and Earth Forces - Right Side over Left

Open Position: Heaven Force - Slightly raise the index fingers to open the palms. Bring the arms up , bending the elbows so that the left hand is under the right elbow with the left index finger pointing up to Heart 3.

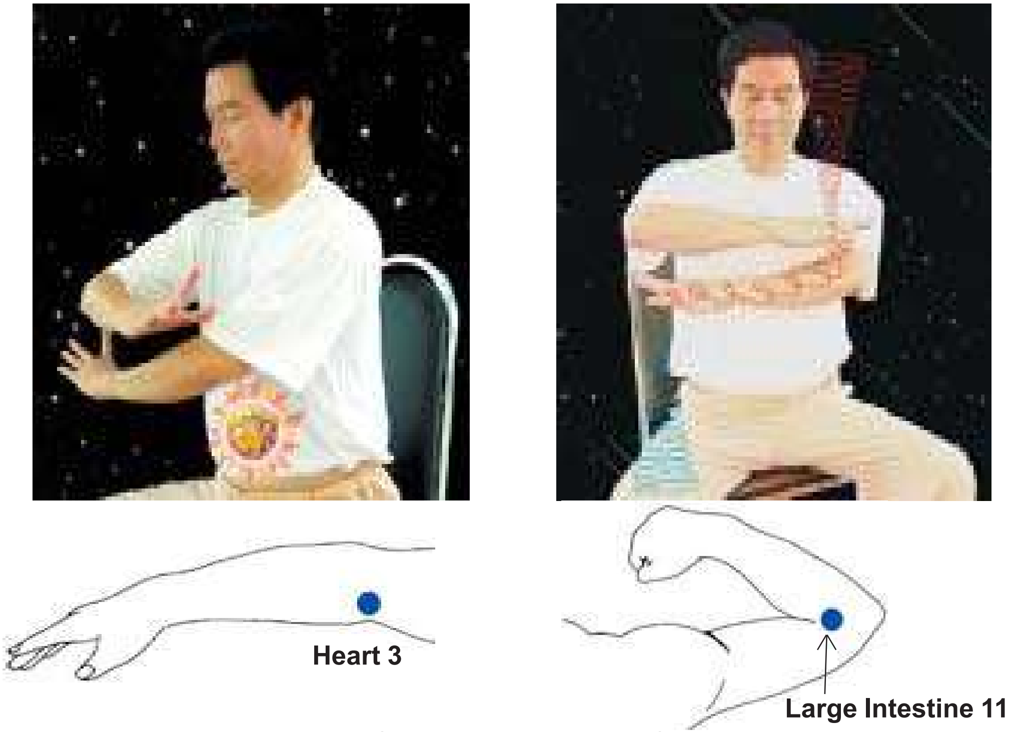

Fig. 8.15 *Complete the Heavenly Circuit.*

The right index finger is pointing up toward heaven and connecting to the galaxy. The right forearm is over the left forearm, with the tip of the right index finger held above the Large Intestine 11 point of the left arm.

The **right index finger draws in the heavenly force** like antennae. The force flows up the bones of the right arm, across the shoulders to the left arm and through the bones of the left arm to the left index fingertip. It **then flows through the index finger of the left hand** into the Heart 3 point of the right arm, **completing the heavenly circuit**. Continue cycling in this way. Hold the position and count to five.

The practice uses the 'Empty Force'; this means that there is no contact. The fingers must be 'tense without tensing'. When you tense and raise the index finger, you will feel the Laogong point opening.

Close Position: Tan Tien, earth and the universal force - keeping the arms in the same position, move both index fingers to point down. The **right index finger will now point toward the Large Intestine 11 point of the left arm and the left index finger will point downward toward the earth** and continue down to connect to the universe. Draw in the earth force through the left index finger, bringing it through the bones of the left arm, across the shoulders through the right arm to the right index finger and out into the Large Intestine 11 point of the left arm. Continue cycling in this way, **completing the earth circuit**.

Open and close three times altogether ending in the open position. Hold each count for five seconds.

Holding and Activating the Chi Ball
Tan Tien and the Original Force - Right Side over Left

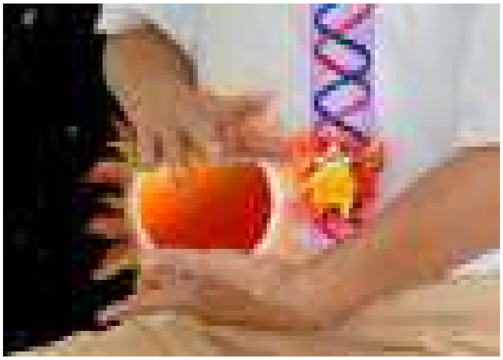

Fig. 8.16 *Right Side over Left - 'Holding the Ball'.*

Turn the **left palm upward** and lower it to the Tan Tien with the pinky pointing inwards.

Lower the right palm to the navel, palm down, as though you were holding a small Chi ball. Lower the right thumb to point toward the left Laogong point. This opens the eye of the right hand (LI-4), then point it toward the navel. Be aware of Tan Tien and universe.

Yin/Yang Palms

Separate the hands and extend them out in front of the body at navel level. The **left palm** is still pointing up, and is **Yang** drawing in the heavenly force through Laogong. The **right palm** is still facing down, and is **Yin** drawing in the earth and the galactic force from below.

Conclude by turning the left palm over to face downward. **Draw in the earth force through both palms.**

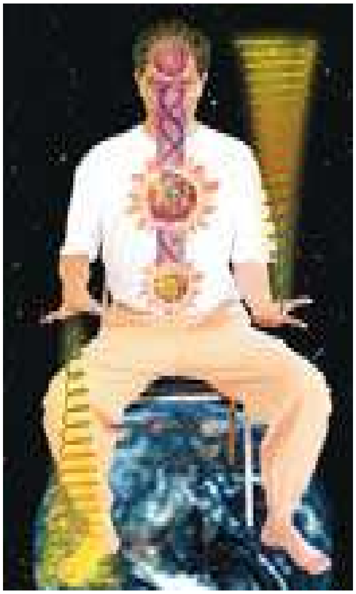

Fig. 8.17 *'Yin and Yang Palms' - Right Hand connecting to Earth and Left to Heaven.*

Left Side: **Repeat on the left hand side with the left arm over the right**. Holding the Chi ball the left hand is on top. Hold for a count of five seconds.

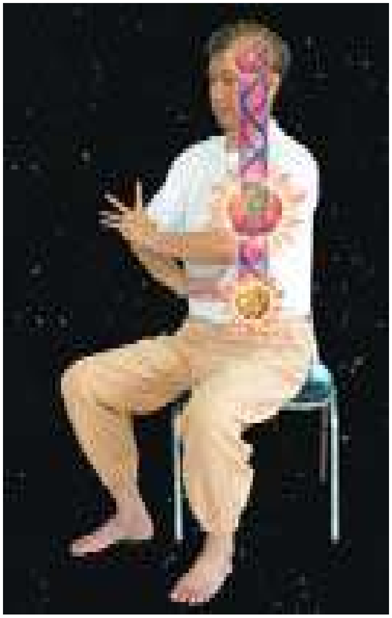

Fig. 8.18 *Open the Heavenly Circuit. Then the bottom hand slides and turns 'Palm-Up' and the Top Arm slides to 'hold the Chi Ball'*

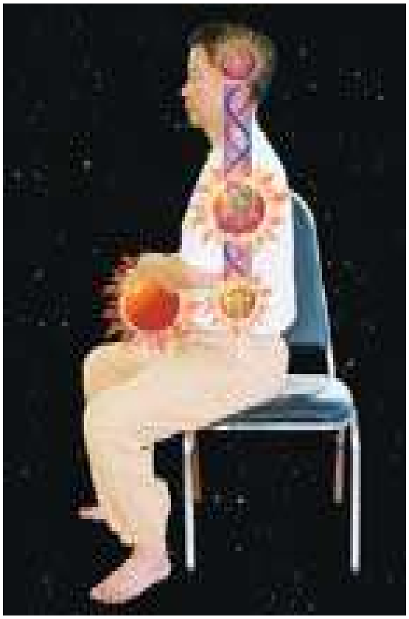

Fig. 8.19 *Holding the Chi Ball - Left Side over Right.*

Closing Movements

Activate the Chi Channels - Balance the Organs

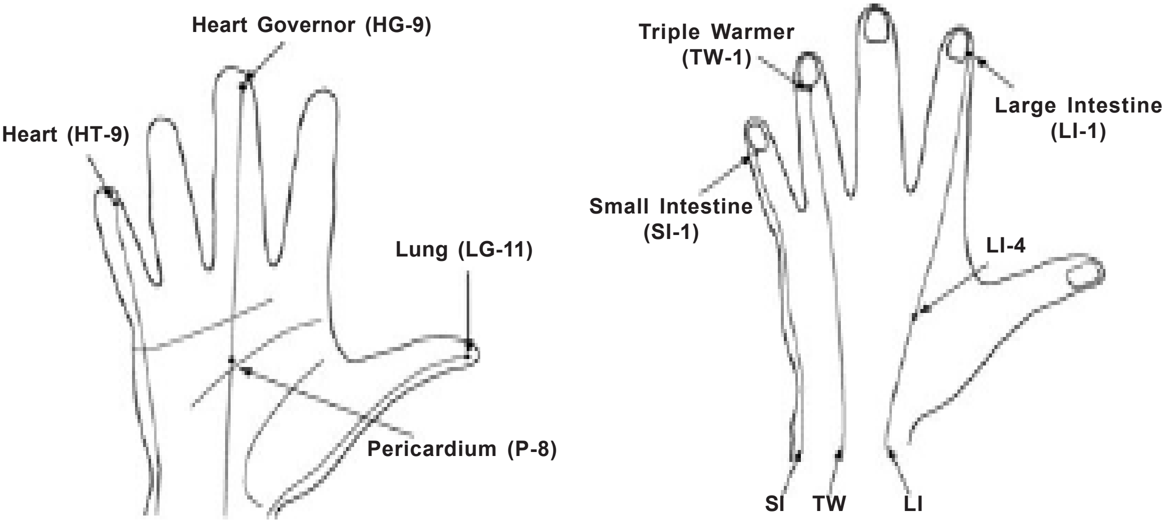
Heart Governor (HG-9)
Heart (HT-9)
Lung (LG-11)
Pericardium (P-8)
Triple Warmer (TW-1)
Large Intestine (LI-1)
Small Intestine (SI-1)
LI-4
SI
TW
LI

Fig. 8.20 *Finger Points*

The closing movements balance organ energy while activating the Chi Channels.

Hold the arms extended forward with the palms down at solar plexus level. Hold each stage for a count of five seconds.

Index Fingers:

Tense all the fingers and feel the Chi filling all the joints; open the index fingers, raising them up, while keeping the rest of the fingers level.

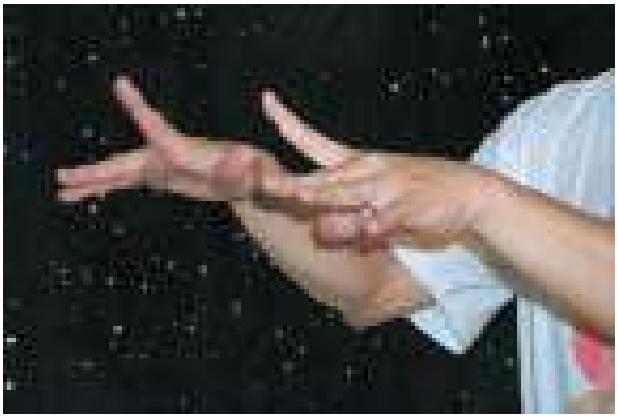

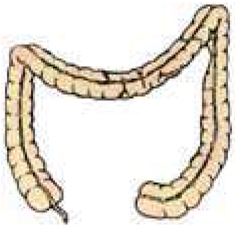

Large Intestine

Fig. 8.21 *Activate the Index Fingers - Up.*

The tension should be like the string of a tuned musical instrument. When lightly tense the Heavenly Chi will be drawn in through the index fingers.

Relax the hands and bring the index fingers back to the neutral position, level with the other fingers.

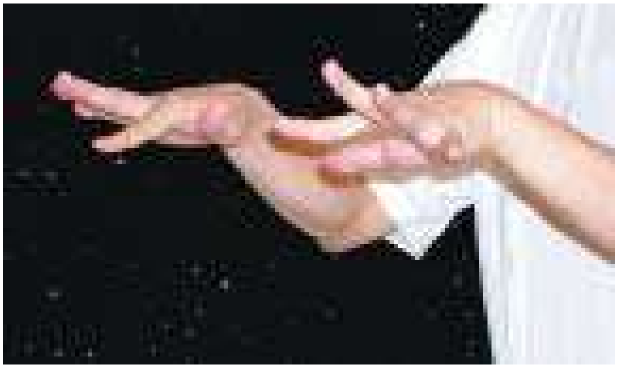

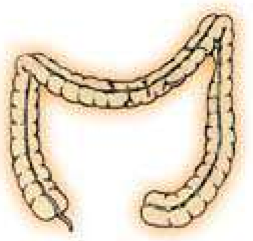

Energy to Large Intestine

Fig. 8.22 *Draw Earth Energy through the Index Fingers - Down.*

Stretch out and tense all the fingers and **press the index fingers to point down** toward the earth and continue down to the universe. Keep the other fingers level. **Draw in earth energy through the index fingers and circulate it through the body.**

Bring the index fingers back to neutral and relax your hands.

Once again **open the index fingers by raising the index fingers up**, while keeping the rest of the fingers level. **Draw in the Heavenly Chi through the index fingers**.

Bring the index fingers back to neutral again and relax your hands.

Ring Fingers:

Stretch the ring fingers down toward the earth, keeping the rest of the fingers level. Draw in earth energy through the ring fingers and circulate it through the body.

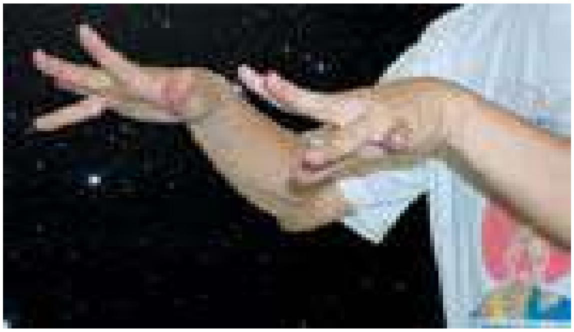

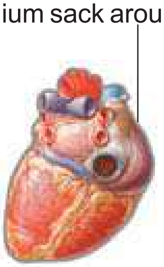
Pericardium sack around the Heart

Fig. 8.23 *Draw Earth Energy through the Ring Fingers.*

Bring the ring fingers back to neutral and relax your hands.

Once again open the index fingers by raising the index fingers up, while keeping the rest of the fingers level, and draw in the Heavenly Chi through the index fingers.

Bring the index fingers back to neutral again and relax your hands.

Thumbs:

Stretch the thumbs out and down toward the earth, keeping the rest of the fingers level. Draw in earth energy through the thumbs and circulate it through the body.

Bring the thumbs back to neutral and relax your hands.

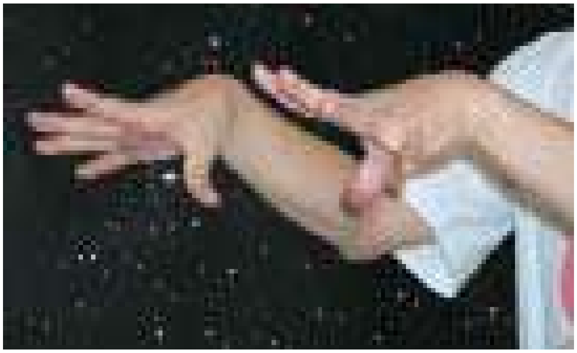

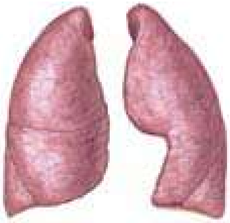

Fig. 8.24 *Draw Earth Energy through the Thumbs.*

Once again open the index fingers by raising the index fingers up, while keeping the rest of the fingers level, and draw in the Heavenly Chi through the index fingers.

Bring the index fingers back to neutral again and relax your hands.

Pinky Fingers:

Stretch the pinky fingers out and down toward the earth, keeping the rest of the fingers level. Draw in earth energy through the pinky fingers and circulate it through the body.

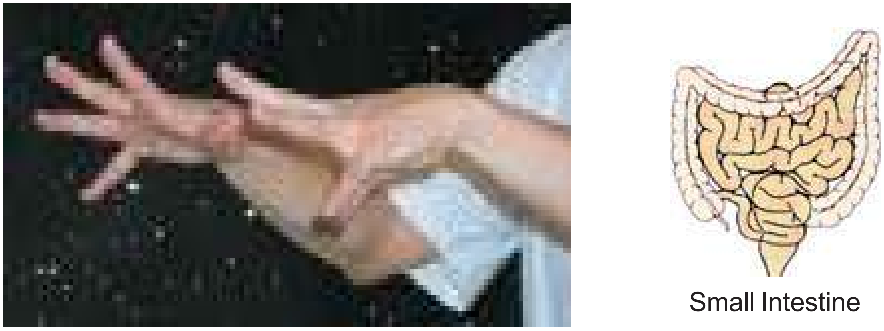

***Fig. 8.25** Draw Earth Energy through the Pinky Fingers.*

Bring the pinky fingers back to neutral and relax your hands.

Once again open the index fingers by raising the index fingers up, while keeping the rest of the fingers level. Draw in the Heavenly Chi through the index fingers. Bring the index fingers back to neutral again and relax your hands.

Middle Fingers:

Stretch the middle fingers out and down toward the earth, keeping the rest of the fingers level. Draw in earth energy through the middle fingers and circulate it through the body.

Bring the middle fingers back to neutral and relax your hands.

Once again open the index fingers by raising the index fingers up, while keeping the rest of the fingers level. Draw in the Heavenly Chi through the index fingers.

Bring the index fingers back to neutral again and relax your hands.

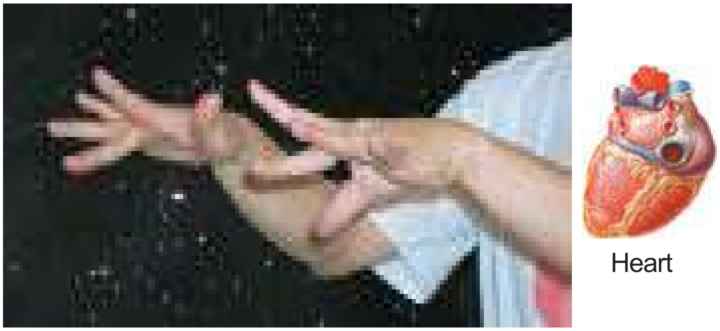

***Fig. 8.26** Draw Earth Energy through the Middle Fingers.*

Crane's Beak and Swallow the Saliva:

Form the "Crane's Beak" with both hands by bringing all the fingertips together with the thumbs inside. **Inhale** and **contract the sexual organs**.

Move your tongue and suck the mouth to **activate the saliva**. Divide it into three parts, tighten the neck and gulp down the **first part** to the center of the **navel**; force the **second** mouthful of nectar down to the **left side of the navel** and the **third** to the **right side of the navel**. Raise the forearms up to shoulder height with the fingers pointing down. Slowly open the palms and begin to lower the arms to the sides, until the palms are facing each other before the starting position.

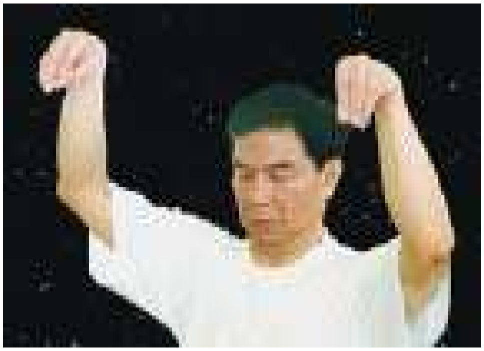

Fig. 8.27 *Form the 'Crane's Beak'.*

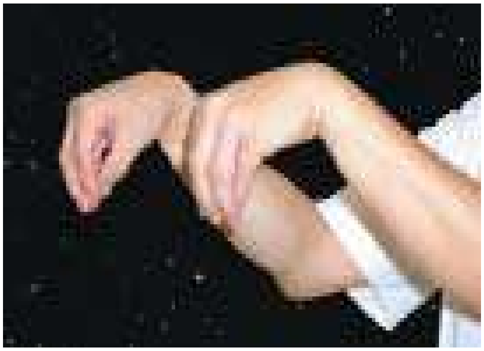

Fig. 8.28 *Raise Forearms to Shoulders.*

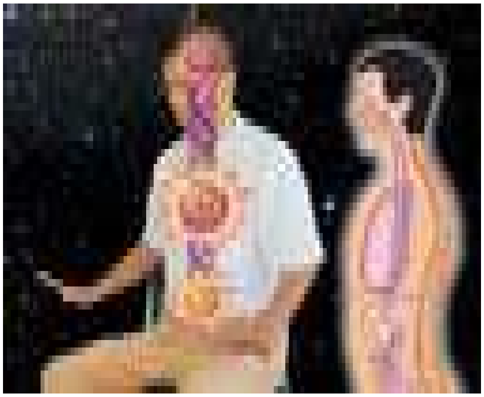

Fig. 8.29 *Lower Arms to the Sides.*

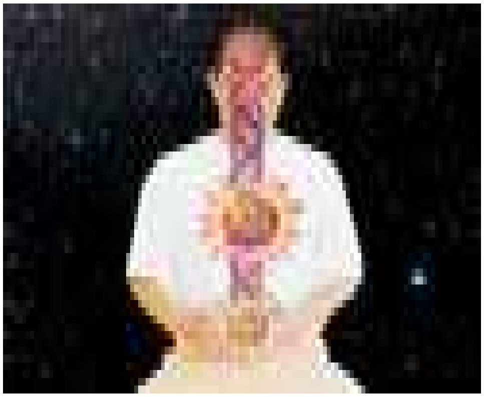

Fig. 8.30
Palms facing each other to cover the Navel.

Conclusion:

Bring awareness to the navel. Notice the quality and intensity of the energy generated and **collect your energy at the navel**. Men place hands over the navel with the right hand on top; women vice versa. Rest.

Summary:

Practice Part I daily for one or two weeks until you can do it well. Then you can proceed to learning Part II.

Part II
Opening the Bridge and Regulator Channels

Theory

Chi flows through your body along numerous channels or meridians. The channels are divided into two major groups: the **extraordinary** or special channels and the **ordinary** channels. After your mother's egg and your father's sperm joined together to form a single cell, that cell began to divide. The extraordinary channels were the first energy channels formed as a result of that early cell division.

There are **Eight Extraordinary Channels:**

Governor Channel
Functional or Conception Channel
Thrusting Channel
Belt Channel
Yang Regulator Channel
Yin Regulator Channel
Yang Bridge Channel
Yin Bridge Channel

Later, as your fetus developed, your Original Chi flowed through the eight extraordinary channels to help create your internal organs and their energy channels. These twelve "ordinary channels" are divided into yin and yang. The yin Channels are connected to the solid organs, and the yang channels are connected to the hollow organs.

The **Twelve Ordinary Channels:**

Lung (Yin) Channel
Large Intestine (Yang) Channel
Stomach (Yang) Channel
Spleen (Yin) Channel
Heart (Yin) Channel
Small Intestine (Yang) Channel
Bladder (Yang) Channel
Kidneys (Yin) Channel
Pericardium (Yin) Channel
Triple Warmer (Yang) Channel
Gall Bladder (Yang) Channel
Liver (Yin) Channel

These twelve organs serve to extract energy from the food and water we ingest and the air that we breathe to create energy through the metabolic processes of respiration, circulation, digestion, elimination and reproduction. This energy is called PostNatal Chi, as it comes in after we are born. Thus the extraordinary channels serve as the link between our Original or Prenatal Chi which came from our mother and father, and our Postnatal Chi which comes from our food and air as a result of metabolism.

General Functions of the Eight Extraordinary Channels

They serve as reservoirs of Chi.

About two thousand years ago, one of the great texts of Chinese medicine, the Nan Ching was written. This classical text describes the twelve ordinary channels as rivers and the eight extraordinary channels as reservoirs of Chi. When the ordinary channels become low in energy, they can draw from the reservoirs of energy in the extraordinary channels. On the other hand, if the ordinary channels become too full, the excess can be taken up by the extraordinary channels. In this way the extraordinary channels help us to maintain a balance in our energy body.

They store and circulate Ching Chi.

The extraordinary channels all draw their energy from the kidneys, which are the storehouse of Ching Chi (essence or sexual energy) in the body. Thus the extraordinary channels circulate the Ching Chi around the body, particularly to the skin and hair, and to the five ancestral organs: the brain and spinal cord, the bone marrow, the blood, the uterus and the liver and gall bladder.

They circulate Defensive Chi to protect the body.

The Chi that protects the body against invasion by external pathogens is called Defensive Chi or Wei Chi. The extraordinary channels circulate Defensive Chi over the back, abdomen and thorax. These functions are performed by the Governor Channel, the Functional Channel and the Thrusting Channels, respectively.

They regulate our life cycles.

In the first chapter of the Yellow Emperor's Classic of Internal Medicine (Huang Di Nei Ching Su Wen), another of the classical texts of Chinese medicine, the life changes of women and men are described in 7 and 8 year cycles respectively. The Functional Channel and the Thrusting Channel govern these cycles.

Eight Extraordinary Channels and Chi Kung

There are few texts available today that describe the purpose and function of the eight extraordinary channels and their place in Chi Kung practice. To make matters more confusing, the extraordinary channels and points used in Chi Kung are often quite different from those used in acupuncture, even though they may bear the same names.

Most acupuncturists pay little attention to the eight extraordinary channels in diagnosis and treatment with the exception of some modern Japanese researchers. By contrast, the extraordinary channels have been of special importance to Taoists and Chi Kung practitioners for thousands of years. Taoists see the extraordinary channels as the foundation of our bodily energy, as the bridge between our Original or Prenatal Chi and our Postnatal Chi; these channels affect us on the deepest level of our basic constitutional energy.

Therefore, Taoists focus on opening up the flow of energy through the eight extraordinary channels as a prerequisite for opening the energy flow in the twelve ordinary channels. In the Universal Tao system, you first open up the Governor Channel and the Functional Channel in the Microcosmic Orbit Meditation. The second pair: the Thrusting Channels and the Belt Channels are opened in the Fusion of the Five Elements II Meditation. The last four channels, the Yin and Yang Bridge Channels and the Yin and Yang Regulator Channels are opened in the third level of Fusion of the Five Elements Meditation. After opening all eight channels, in the next level of Taoist inner alchemical meditation called the Lesser Enlightenment of Kan and Li, the twelve ordinary channels are then opened.

Special Functions of the Bridge and Regulator Channels

The Bridge and Regulator Channels have no points of their own. They borrow points from the other 'ordinary' channels linking together and controlling the flow through all the separate channels of the body. The Bridge and Regulator Channels travel along the same paths. There is no 'real' difference between them. Both of them connect or 'bridge' and 'regulate' the flow of Chi in the body's meridians.

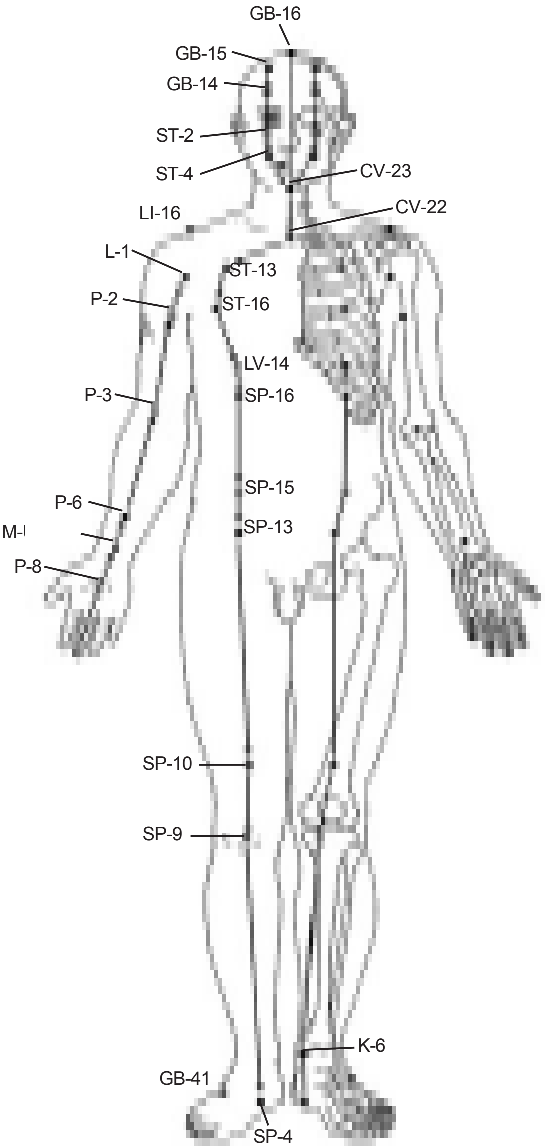
GB-16
GB-15
GB-14
ST-2
ST-4
CV-23
LI-16
CV-22
L-1
ST-13
P-2
ST-16
LV-14
P-3
SP-16
SP-15
P-6
M-I
SP-13
P-8
SP-10
SP-9
K-6
GB-41
SP-4

Fig. 8.31 *Yin Bridge and Regulator Channels*

Bridge Channels Yin and Yang Chiao Mo Function

The Bridge Channels, also called the Heel Vessels as they originate at the heels, regulate the amount of energy being used by all the other meridians in the body. They act like a bridge linking together the stored Chi in the body and the areas in need of Chi. Usually if any meridian uses more energy than it needs to flow properly, then other meridians become deficient as a result. Thus the Bridge Channels seek to assure that your energy is distributed in a balanced way. The Yin Bridge Channel runs along the front or yin side of the body, while the Yang Bridge Channel runs along the back of the body.

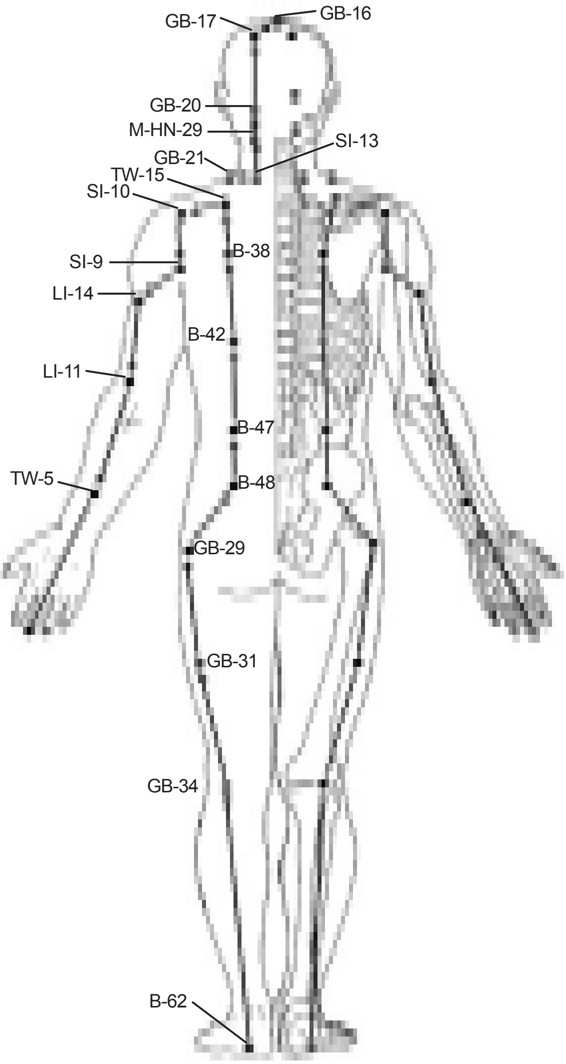
GB-17
GB-16
GB-20
M-HN-29
SI-13
GB-21
TW-15
SI-10
B-38
SI-9
LI-14
B-42
LI-11
B-47
TW-5
B-48
GB-29
GB-31
GB-34
B-62

Fig. 8.32 *Yang Bridge and Regulator Channels*

Disorders of the Bridge Channels

Excess Yang Energy

When yin energy is slowed down in the Bridge Channels, yang energy flows more rapidly. The excess yang can cause the following problems: insomnia, difficulty in closing the eyes, hypertension, stiff back and waist, inability to bend down, thigh tumors, bad colds, spontaneous sweating, headaches, painful eyes, paralysis of the arms and legs, vomiting of milk in infants, deafness, epilepsy, nose bleeding, swelling of the body, pain in the joints and head sweating.

Excess Yin Energy

When yang energy is slowed down in the Bridge Channels, yin energy moves more rapidly. The excess yin causes the following problems: sleepiness, difficulty in keeping the eyes open, hypotension, choking, painful urination, stomach rumbling, vomiting, diarrhea, difficult bowel movements, difficult labor and unconsciousness.

Regulator Channels

Yin and Yang Wei Mo Function

The Regulator Channels, also called the Linking Vessels, bind together all the meridians in the body. The Regulator Channels are also divided into yin and yang. The yin aspect, which runs along the front of the body, moves the yin energy and regulates the blood and inner parts of the body. It connects with all the yin channels: the Liver, Spleen, Kidney, Heart, Pericardium and Lung channels. If the yin aspect becomes imbalanced, the person may suffer from heart pains.

The yang aspect, which runs along the back side of the body, moves the yang energy, controls Defensive Chi, regulates resistance to external infections, and regulates the external parts of the body. It connects with all the yang channels: the Stomach, Bladder, Gall Bladder, Large Intestine, Triple Warmer & Small Intestine channels. If the yang aspect becomes out of balance, the person may catch colds and fevers easily.

By joining together the various channels, the Regulator Channels help to maintain a harmonious and cooperative interaction between the different channels.

The Regulator Channels in Taoist Yoga and Chi Kung are slightly different from those presented in acupuncture texts. The Chi Kung Regulator Channels include the yin and yang arm routes as well; some Taoist Yoga texts also refer to the arm routes as the Yin Yu and the Yang Yu. Acupuncture texts, by contrast, include only the leg, trunk and head routes. Many recent Chi Kung texts, unaware of these differences, depict illustrations from acupuncture texts alone, further adding to the confusion.

Disorders of the Regulator Channels

Excess Yang Energy

When yin energy is slowed down in the Regulator Channels, yang energy moves more rapidly. The excess yang can cause the following problems: swelling and pain in the joints, cold knees, paralysis of the arms and legs, painful back and sides, aching muscles, pain in the head, neck and edge of the eyebrows, fever, rashes, night sweating, tetanus, painful red eyes, colds and superficial fevers.

Excess Yin Energy

Similarly, when yang energy is slowed down in the Regulator Channels, yin energy moves more rapidly. The excess yin can cause the following problems: heart pain, diarrhea with stomach rumbling, difficulty swallowing, pain on both sides of the chest, diseases associated with cold and convulsions.

Summary

The Bridge and Regulator Channels link all of the yin and yang channels in the body together and regulate the flow of energy in these channels to maintain a state of energetic balance in the body. The channels used for meditation are somewhat different from the ones seen in Chinese medical text and acupuncture charts. This is because their purposes are different.

Chinese medicine aims to restore sick people to health. The points being treated must be superficial so that they can be activated by acupuncture needles.

Chi Kung and Taoist meditation aim to maintain health and to take one beyond mere physical health to spiritual immortality. The channels and points can be deep within the body, since the energy is guided by the mind or by postures and movements rather than by needles.

The movements used in Cosmic Chi Kung Part II to activate the Bridge and Regulator Channels are relatively simple. To completely master the Bridge and Regulator Channels, you must learn and practice the third level of Fusion of the Five Elements Meditation. Once you have done so, this part of Cosmic Chi Kung practice will be particularly powerful and balancing for you.

Practice

Opening Movements

The Opening Movements for this second part of Cosmic Chi Kung are the same as for the first part. By now you should be familiar with these movements, so they will be described briefly. For a more detailed description, refer to Part I, Opening Movements.

Hands down at sides, palms facing each other. Commence Bone breathing. Be aware of the Tan Tien and absorb Cosmic Chi. Be aware of the Tan Tien and channeling the earth and galactic forces.

Washing the Bone Marrow

Raise arms to chest height - palms facing each other. Rotate arms until palms face down. Open the 'Tiger's Mouth' - Index finger up, thumb down.

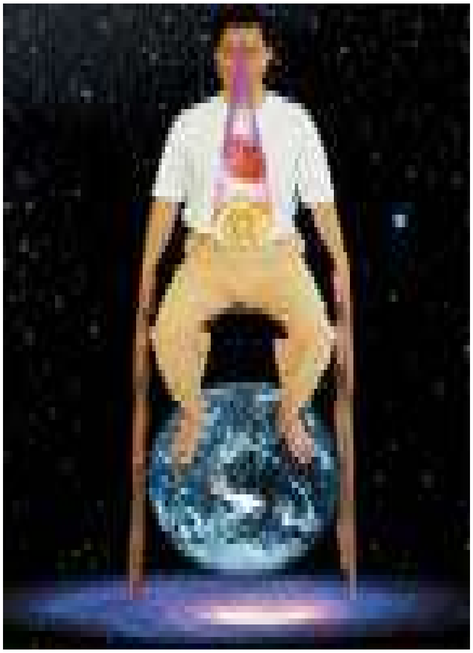

Fig. 8.33 *Hands lower to the Sides. Lower the Mind to the Tan Tien.*

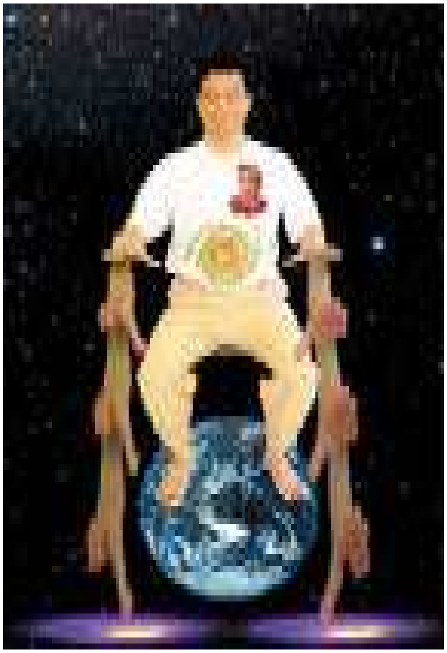

Fig. 8.34 *Open the 'Tiger's Mouth'.*

Activate and Beam to the Eye of the Hip

A. Large Intestine 4 - Eye of the Hand
B. Pericardium 9 - Tip of the Middle Finger
C. Small Intestine 3 - Knife Edge of the Hand

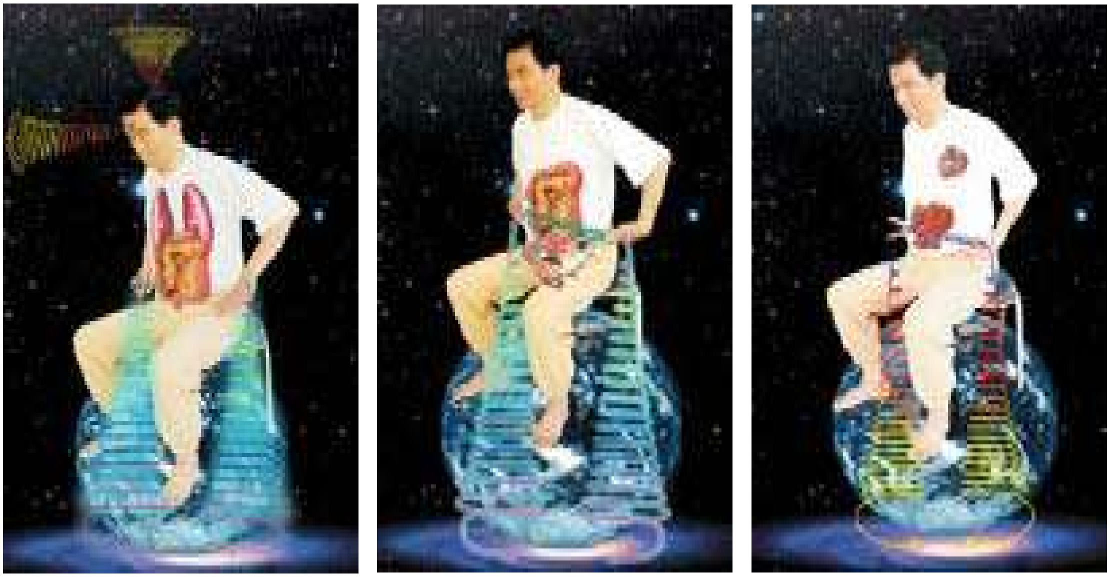

Fig. 8.35 *Activate the Large Intestine, Pericardium and Small Intestine.*

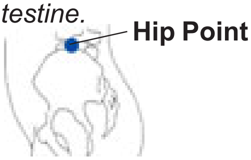

Be aware of the Tan Tien and absorb the Heavenly Force Washing the Bone Marrow Crown to Soles.
Extend the arms forward - palms upward. Be aware of the Tan Tien and absorb the heavenly force. Rotate the arms until the palms face down. Raise the index fingers slightly. Absorb the Earth Chi.

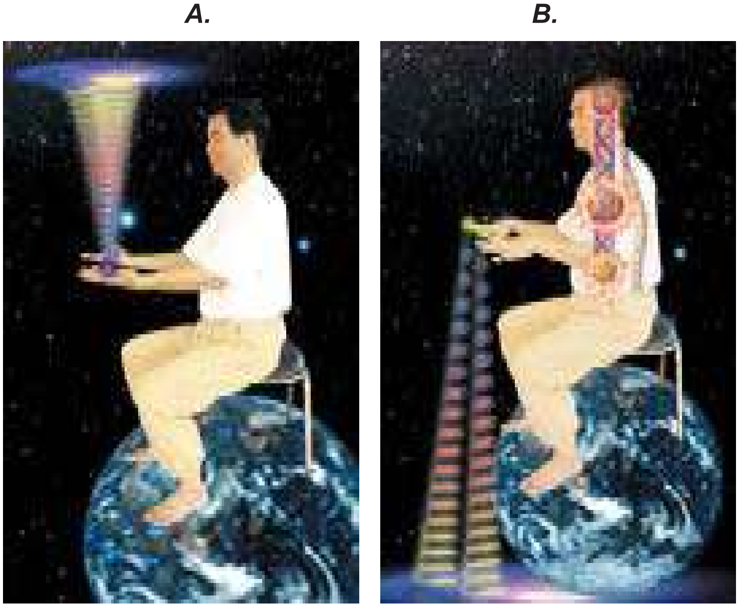

Fig. 8.36
A. *Be aware of the Tan Tien and absorb the Heavenly Force.*

B. *Index Fingers up - be aware of the Tan Tien and absorb the Earth and Universal Force.*

Core Movements of Part II
Opening the Bridge and Regulator Channels

Activate the Throat Center - Right Side

Keeping the elbows sunk, raise the left palm first so that it is about 1 inch in front of the throat center (Conception Vessel 22), palm facing in. Then bring the right hand up so that it is about 1 inches in front of the left hand, palm facing in. Align the Lung 10 point in the middle of the eminence of the left hand with the **throat center. Feel the Chi go through the neck to C-7**. Continue smiling to the universe and connect to the back side of the universe. Hold for a 5 to 30 count and feel the throat center activated.

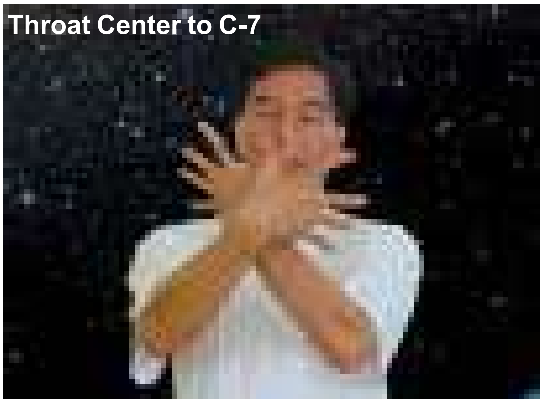

Fig. 8.37 *Connect to the Throat Center.*

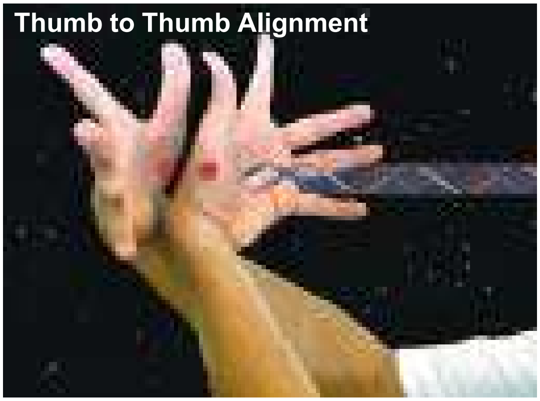

Fig. 8.38
Feel the connection to the Universe at the Back.

Move both hands out slowly like you're pulling silk; when the feeling is diminished or the silk starts to break stop there. Be still and feel the beam connect the throat center to C-7 and the universe. Maintain the energy beam as you move in and out three times altogether ending with the hands as above.

Activate Mideyebrow Point - Right Side

Be aware of the Tan Tien and the universe. Keeping the Lung 10 points aligned, raise the hands up to the level of the mideyebrow point about one inch in front of the face. **Keep the points aligned and beam the energy passing through both hands into the mideyebrow point and through to the base of the skull.** Continue smiling and extending to make a connection to the back universe for a 5 to 30 count.

Use only 5% of the awareness to move both hands very slowly out, like you are **pulling silk**; continue to move out as long as your can feel the connection to the mideyebrow, C-7 points and the universe, maintaining the energy beam as you move.

Move in and out three times altogether, ending with the hands as above.

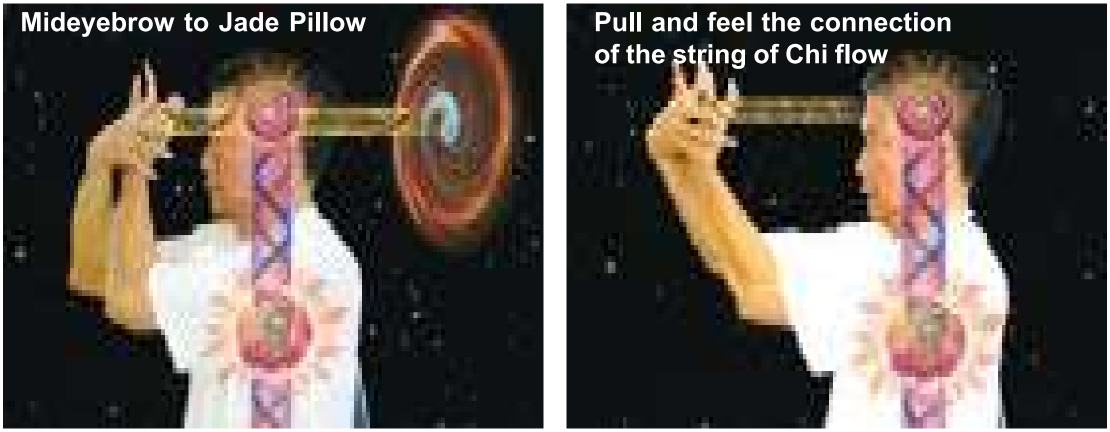
Mideyebrow to Jade Pillow
Pull and feel the connection of the string of Chi flow

Fig. 8.39 *Connect to the Mideyebrow, Jade Pillow and the Universe at the Back.*

Keeping the palms and thenars (Lung 10) aligned, move the hands back down to the position in front of the throat center. Keeping the forearms horizontal at shoulder height, separate the hands so that the Laogong point is aligned in front of Stomach 13, just **below the clavicle** on a vertical line above the nipple or mammillary line. The middle fingertips should almost be touching. **Beam energy from the Laogong points into ST-13 through the back on the side of the spine and connect to the universe**. You will be following this vertical line, which passes through the nipples, down the torso to the level of the navel, **stopping to focus and beam energy into key points along the way.** On each of these, focus your Yi on beaming energy into the organs inside each position. Stomach 13 is the location that activates the Heart and Lungs. Feel the change in your breathing as you activate the lungs.

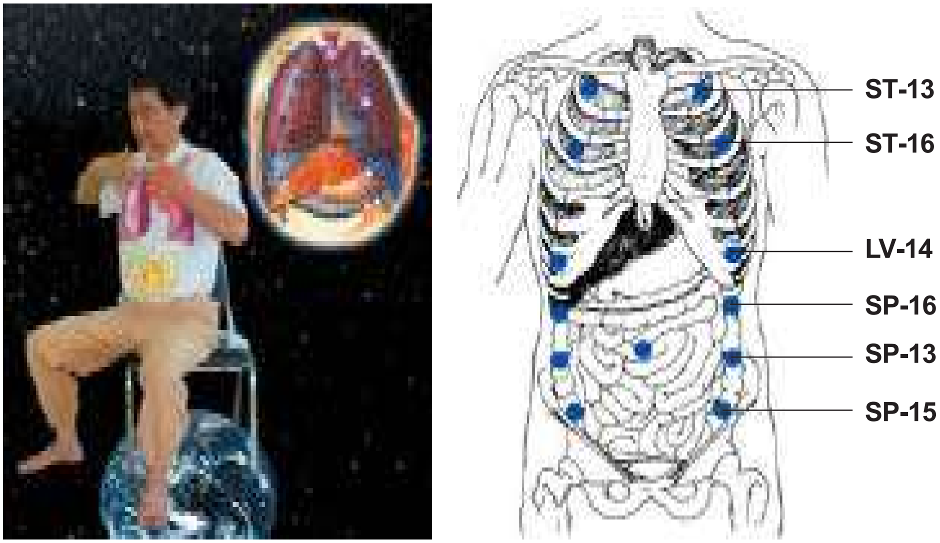

Fig. 8.40 *Beam Energy into ST13.*

Move the hands down the mammillary line to Stomach 16, about one inch above the nipple (in the space between the third and fourth ribs). Chi connects to the back and to the universe. Again, beam energy to energize and balance the **heart and middle of the lungs**. Move the hands down to Liver 14 (about 2 - 3 inches below the nipple in the space between the 6th and 7th ribs). Beam healing energy into the **liver and gall bladder.** Feel the Chi or the vibration of the palms activate the liver and feel the energy moving inside there.

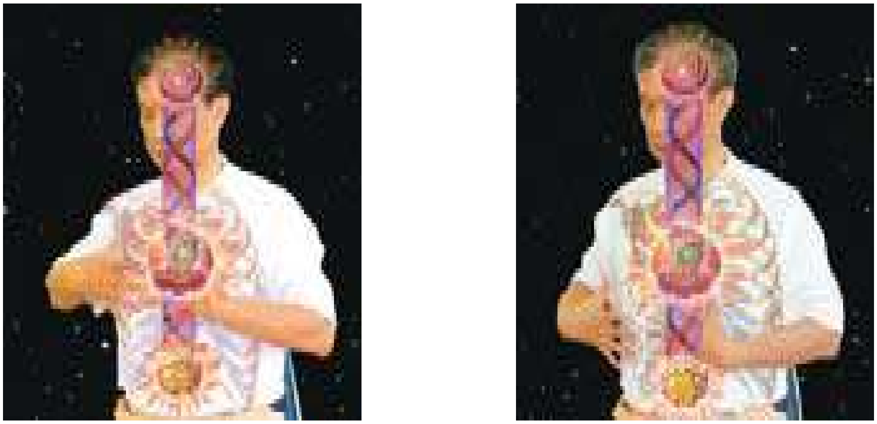
A. *B.*

Fig. 8.41 A. *Energize and balance the Lungs at ST16.*
B. *Energize and balance the Liver and Gallbladder at LV14.*

Move the hands down to **Spleen** 16 (just below the rib cage on the mammillary line). Beam healing energy into the **stomach, pancreas and spleen and the liver.** Picture the organs receiving the Chi from the cosmic and the universe.

Move the hands down to Spleen 15 (on the mammillary line to the level of the navel). Beam healing energy from your palms into the **small intestine and the Tan Tien.**

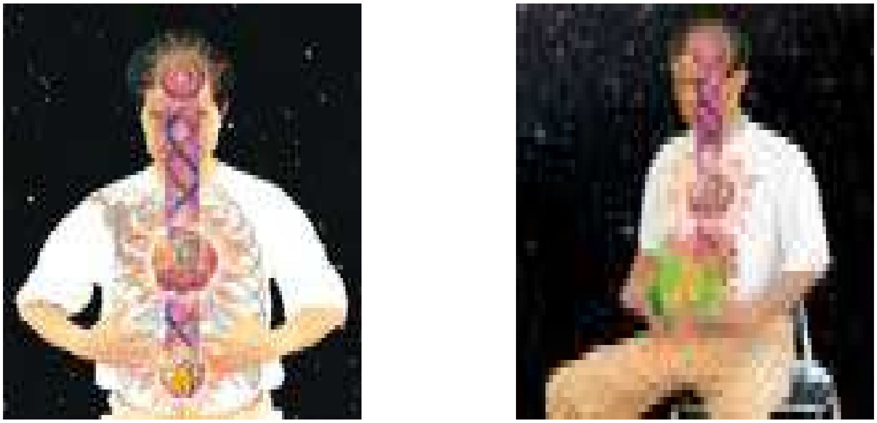
A. *B.*

Fig. 8.42 A. *Energize the Stomach, Pancreas, Spleen and Liver at SP 16.*
B. *Energize the Tan Tien and Small Intestine at SP15.*

Hold the Chi Ball - Right Hand on Top

Turn the **left palm up** and align the Chi Ball with the Tan Tien. Turn the **right palm down** and align the eye of the hand (LI-4) with the navel. Feel your palms holding a Chi ball and feel a **Chi ball inside the Tan Tien**. Feel the Chi ball connect to the Cosmic Chi outside. Feel the north (usually left) and south (usually right) poles in the hands. Lower the right thumb to point toward the left Laogong point.

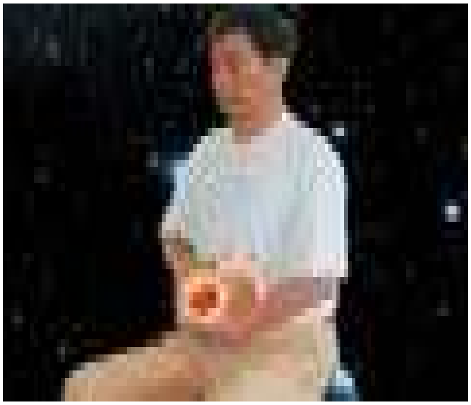

Fig. 8.43
Hold the 'Chi Ball' - One in the Tan Tien and the other in the hands - Right Hand over Left.

Yin/Yang Palms

Separate the hands and extend them out in front of the body at solar plexus level. The **left palm** is still pointing up and is yang drawing in the **Heavenly Force** through Laogong. The **right palm** is still facing down and is yin drawing in the **Earth Force** energy. **'Be Aware Tan Tien and Universe'**.

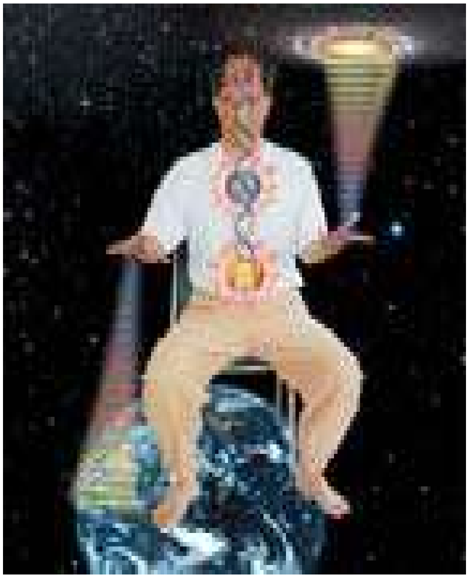

Fig. 8.44 *Connect to Heaven and Earth.*

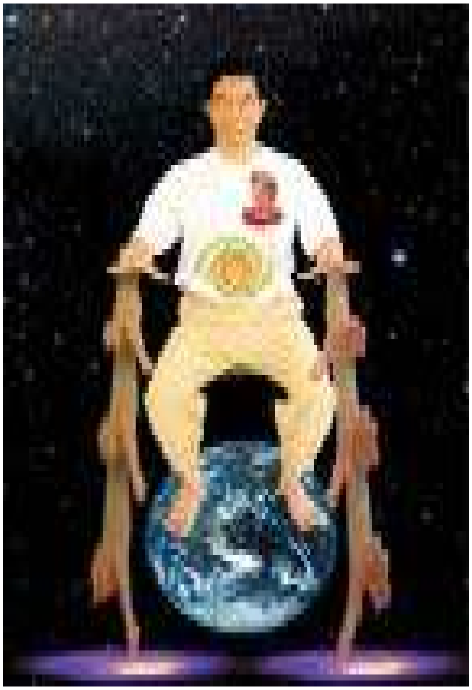

Conclude by turning the left palm over to face downward. **Draw in the 'Earth Force' through both palms.**

Fig. 8.45 *Draw the Earth Force through the Palms.*

Repeat the sequence with the left palm over the right.

Start by moving the right palm to the throat and then the left palm. Scan the body, as in the previous sequence, using both palms to connect to the points running down the front of the body. Remember to maintain a sense of the Chi flowing between the points in the front and their corresponding points in the back of the body.

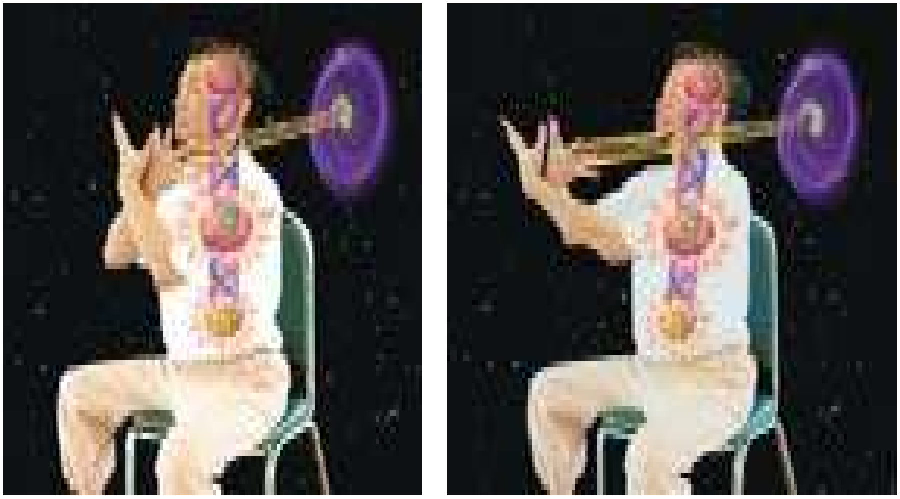

Fig. 8.46 *Repeat the Practice on the Left Side. Push and Pull the silk at the Throat.*

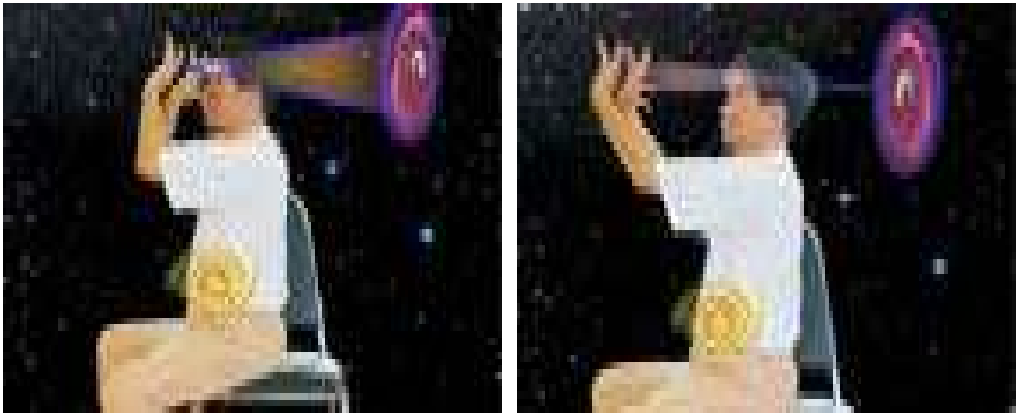

Fig. 8.47 *Project to the Mideyebrow and Base of the Skull - connect to the Universe at the back. Then hold the Ball and repeat Finger Movements, as below.*

Continue with the left side, now pulling the silk at throat (C-7).

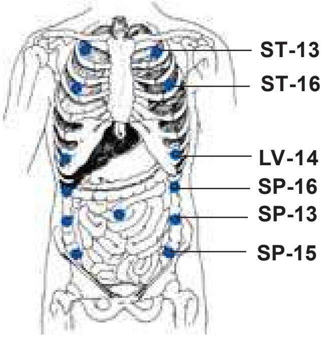

Fig. 8.48 *Beam Energy into each of the Organs like the last time.*

Closing Movements

The closing movements, crane's beak and finish are the same as before.
Activate the Chi channels - balancing the organs.

Index Fingers

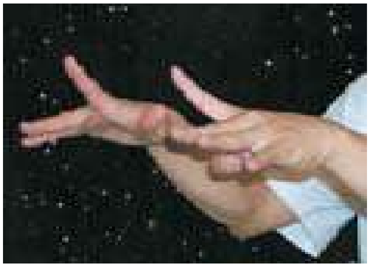

Fig. 8.49 *Open (up)*

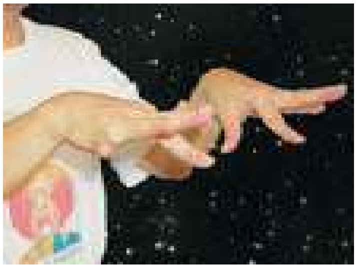

Fig. 8.50 *Neutral*

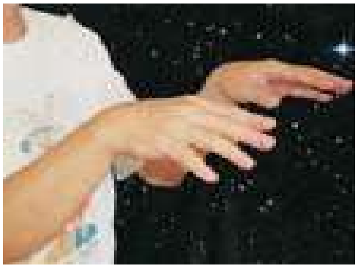

Fig. 8.51 *Close (down)*

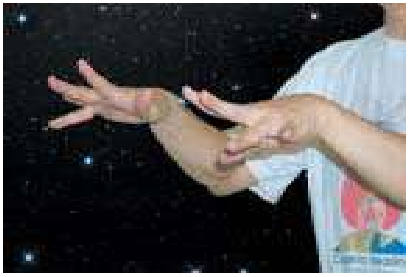

Fig. 8.52
Ring Finger *- down, neutral.*
Index Finger *- open, neutral.*

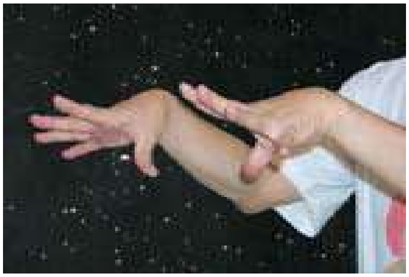

Fig. 8.53
Thumb *- in and down, neutral.*
Index Finger *- open, neutral.*

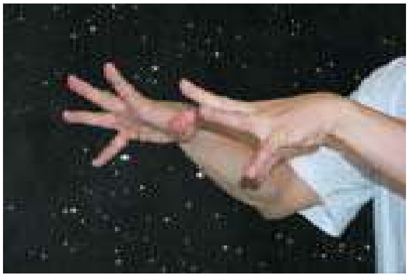

Fig. 8.54
Pinky Finger *- down, neutral.*
Index Finger *- open, neutral.*

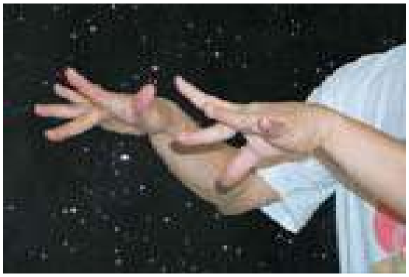

Fig. 8.55
Middle Finger *- down, neutral.*
Index Finger *- open, neutral.*

Crane's Beak

Form the crane's beak with both hands by bringing the fingertips together with the thumbs inside. Activate the saliva and swallow down three times to the Tan Tien. Raise the forearms to shoulder height with the fingers pointed down, inhaling gently as you raise. Open the palms slowly and begin to lower the arms to the sides as you exhale, until the palms are facing each other in the starting position.

Finish

Bring the awareness to the navel and notice the quality and intensity of the energy generated. **Collect energy at the navel**. Men, place the hands over the navel with the right hand on top; vice versa for women.

Part III

Opening the Functional and Governor Channels

Theory: Functional and Governor Channels

The Functional and Governor Channels are the first two of the eight extraordinary channels described earlier in this chapter (beginning p.178). They serve as the reservoirs of yin and yang energy in the body. The Functional or Conception Channel is yin and all of the ordinary yin channels connect to it. Similarly, the Governor Channel is yang and all of the ordinary yang channels connect to it.

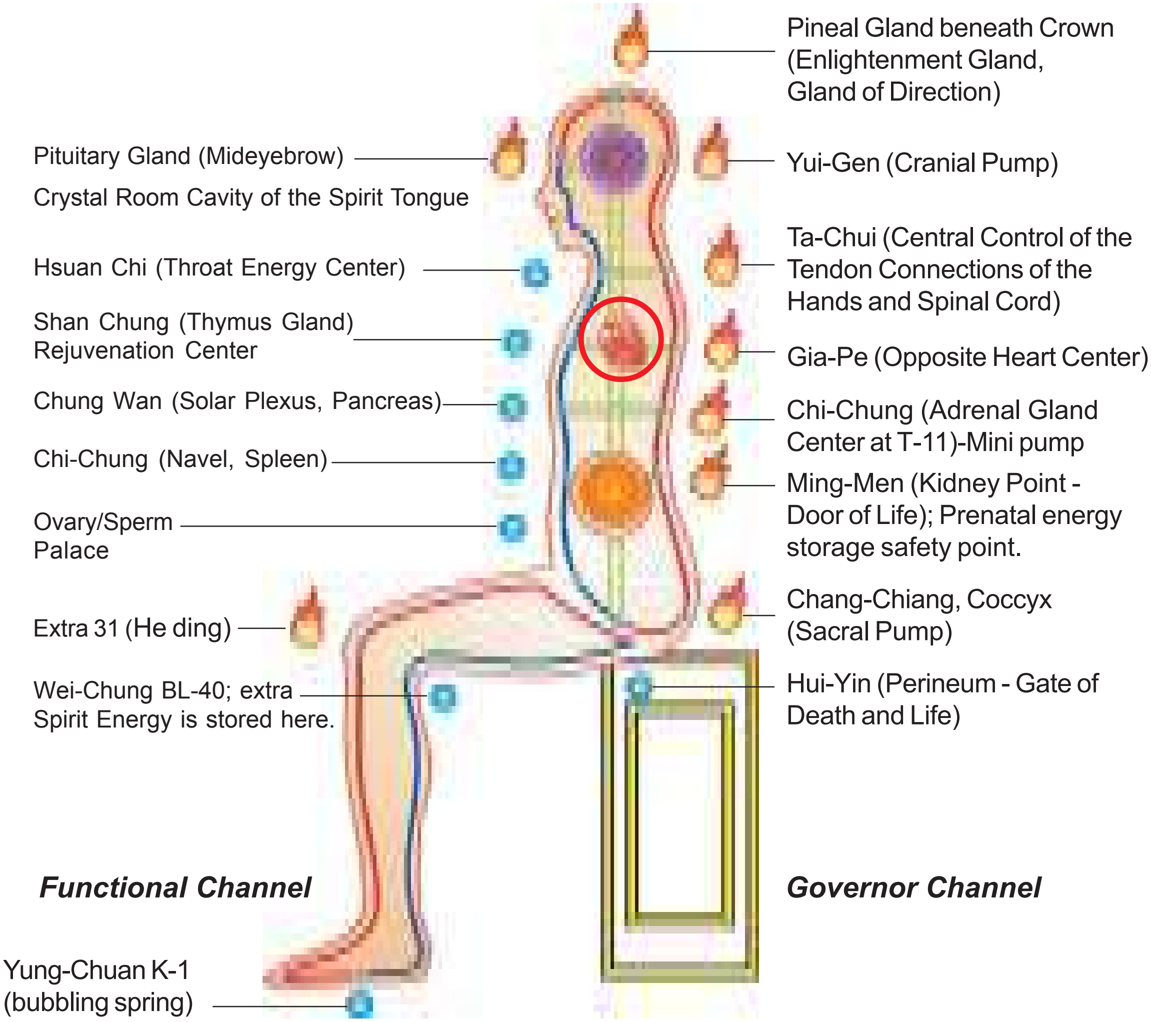
Pineal Gland beneath Crown (Enlightenment Gland, Gland of Direction)
Pituitary Gland (Mideyebrow)
Crystal Room Cavity of the Spirit Tongue
Yui-Gen (Cranial Pump)
Ta-Chui (Central Control of the Tendon Connections of the Hands and Spinal Cord)
Hsuan Chi (Throat Energy Center)
Shan Chung (Thymus Gland) Rejuvenation Center
Gia-Pe (Opposite Heart Center)
Chung Wan (Solar Plexus, Pancreas)
Chi-Chung (Adrenal Gland Center at T-11)-Mini pump
Chi-Chung (Navel, Spleen)
Ming-Men (Kidney Point - Door of Life); Prenatal energy storage safety point.
Ovary/Sperm Palace
Chang-Chiang, Coccyx (Sacral Pump)
Extra 31 (He ding)
Wei-Chung BL-40; extra Spirit Energy is stored here.
Hui-Yin (Perineum - Gate of Death and Life)
Functional Channel
Governor Channel
Yung-Chuan K-1 (bubbling spring)

Fig. 8.56 *Cosmic Orbit*

Chi Ball

In Part III we also begin to focus more on the sensation of Chi being emitted from our palms. This energy is emitted from the **Laogong** points primarily, but you can also feel it throughout the palms and fingers. It will feel like a balloon between your hands that is expanding and contracting or like two magnetic fields repelling and attracting each other.

Once you have practiced Cosmic Chi Kung for awhile and have become skillful at it, you may find that Grasping the Chi ball is all you need to do to activate your flow of healing energy before doing a session. As you become sensitive to the sensation of your own Chi, it becomes an easy step to feel the Chi of others.

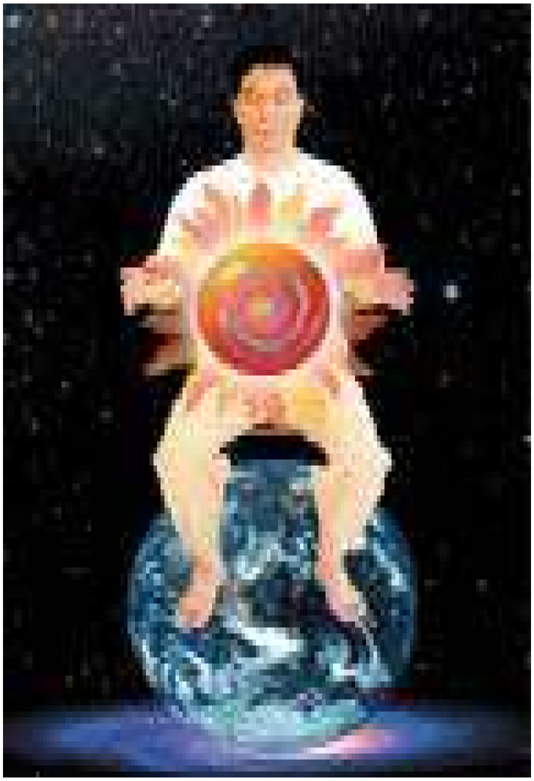

Fig. 8.57 *Grasping the Chi Ball.*

Opening Movements, Hands Down at Sides, Palms Facing Each Other
Bone Breathing - Absorbing Cosmic Chi

The Opening Movements are the same as in the sequence in Part I, 'Opening to Heaven, Earth and Cosmic Chi', on page 164.

A. Channeling the Earth Force - Washing the Bone Marrow Raise arms to chest height - Palms facing each other. Rotate arms until **palms face down.** Open the Tiger's Mouth - Index finger up, thumb down.

B. Activate and Beam to the Eye of the Hip

C. Absorb the Heavenly Force Washing the Bone Marrow - Crown to Soles

Core Movements of Part III

Grasping the Chi Ball

Slowly rotate the hands so that your palms face each other in front of your navel.

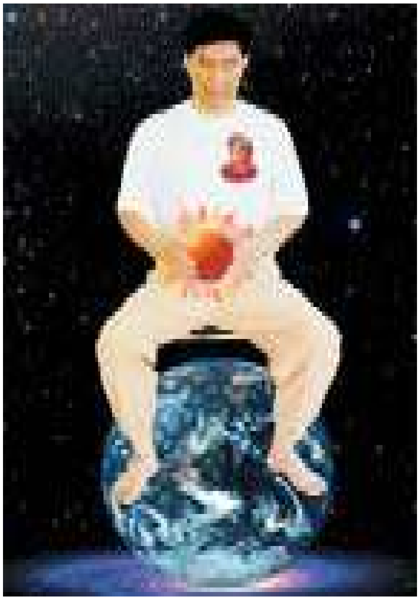

Fig. 8.58 *Hold the Chi Ball.*

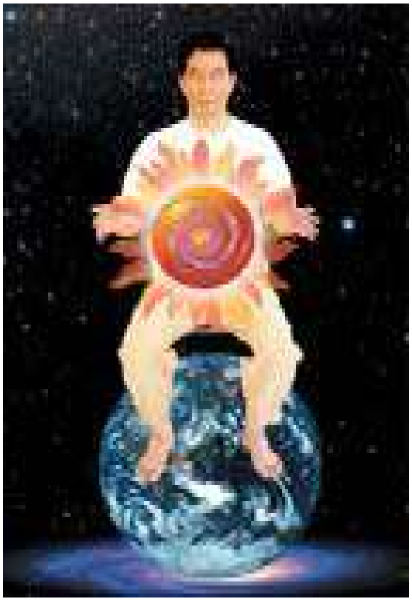

Fig. 8.59 *Stretch and squeeze it.*

Feel the invisible ball of energy between your hands. Allow the energy to push your hands apart keeping the feeling of connection between your palms. When you have less feeling stop the palms and hold it there and reestablish the feeling of the Chi ball.

Let the energy draw your hands back toward each other, like a magnet, until you feel as if you were squeezing and compressing the ball.

Stretch and squeeze the Chi ball three times by opening and closing your palms in this way.

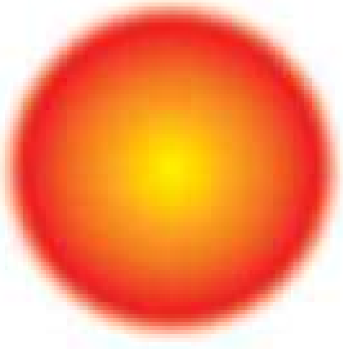

Activate the Outer and Inner Arm Gates

Activate Waiguan Triple Warmer 5 and Neiguan Pericardium 6 - **Right** arm on top. Slowly rotate the arms so that both **palms face down.**

Cross the right wrist over the left wrist with 1 inch of space between the hands. Align PC-6 of the right wrist with TW-5 of the left wrist. Feel the two gates activate each other like metal and magnet drawn to each other. Hold for a 5 to 30 second count.

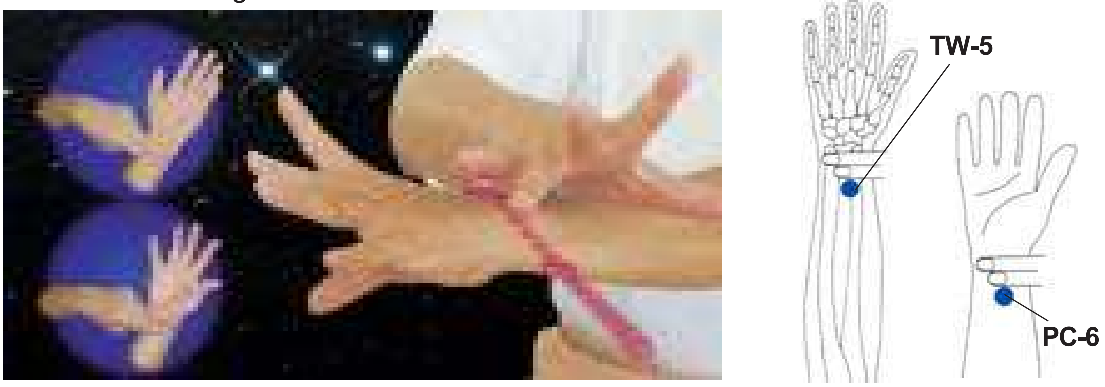

Fig. 8.60 *Align and activate PC-6 (Right) with TW-5 (Left).*

Keeping the wrists crossed, slowly rotate hands **turning the palms upward**. Now TW-5 of the right wrist will be aligned with PC-6 of the left wrist. Hold and count and feel the points activated. 'Be aware of Tan Tien and Universe'.

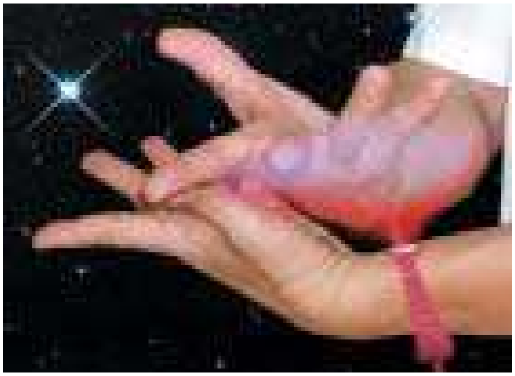

Fig. 8.61
Align and activate TW-5 (Right) with PC-6 (Left).

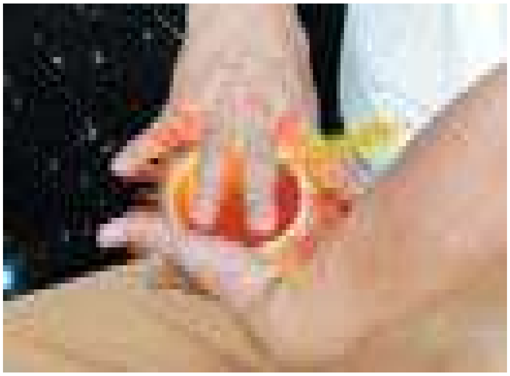

Fig. 8.62 *Feel the Chi Ball and beam energy from LI-4 into the Navel Point.*

Open the Functional Channel with LI-4

Bring the left hand down to the level of the Tan Tien with the Chi Knife facing in. Lower the right hand, palm down and align the eye of the hand, the LI-4 point with your navel. **Feel the Chi ball between your two hands**, and at the same time beam energy from LI-4 into the **navel** point.

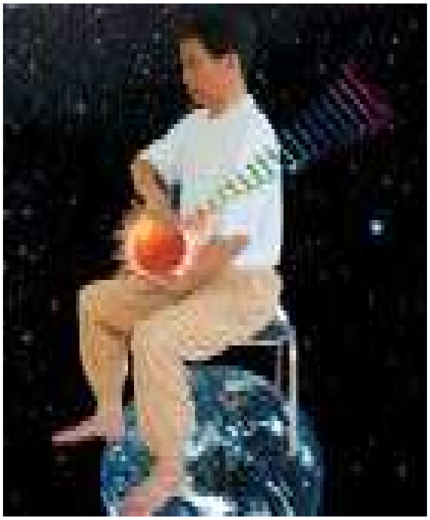

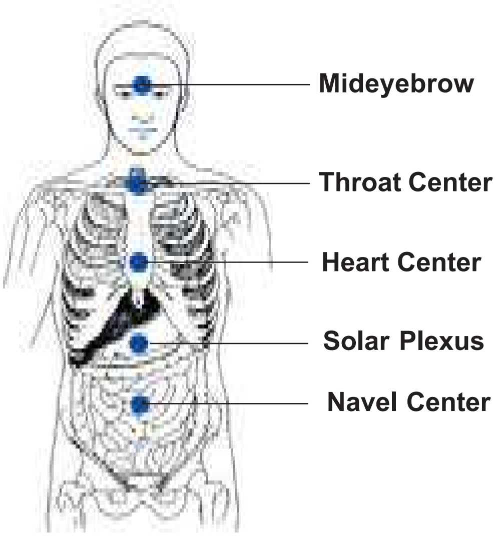
Mideyebrow
Throat Center
Heart Center
Solar Plexus
Navel Center

Fig. 8.63 *Magnify the Power of the Beam to the Solar Plexus.*

Fig. 8.64 *Activate all the Points.*

Raise the right hand up the Functional Channel, stopping and aligning LI-4 with the **solar plexus**, hold for a 5 to 30 second count and feel the points open and starting to activate each other.

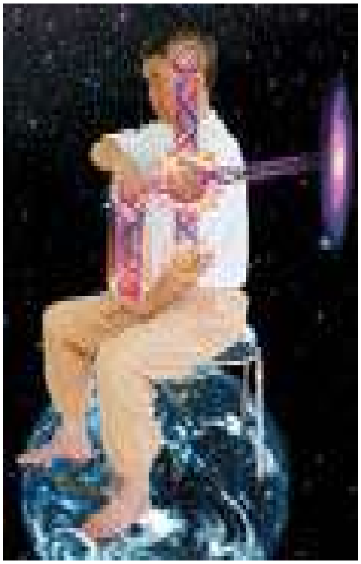

Fig. 8.65
Beam to the Heart Center.

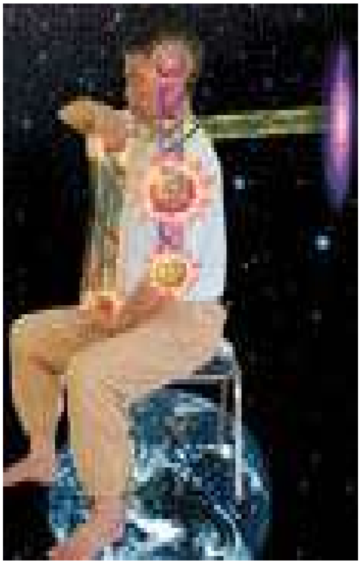

Fig. 8.66
Beam to theThroat

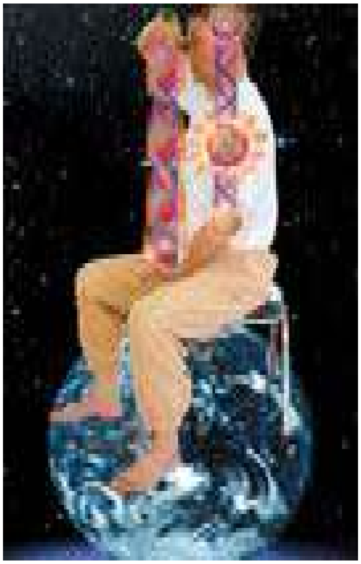

Fig. 8.67
Beam to the Mideyebrow

Continue to move up to the **heart**, **throat** (CV-22) and **mideyebrow** points. Beam energy from the eye of the hand into each of these points and at the same time, continue to feel the Chi ball connection between the two hands.

Return down the Functional Channel in the same manner, starting from the mideyebrow and stopping to beam energy into each point.

Next beam into the **throat center, heart center, solar plexus** and finally the **navel**. Hold the Chi ball at the navel with the right hand over the left.

Yin/Yang Palms

Separate the hands and extend them out in front of the body at solar plexus level. The left palm is still pointing up and is yang drawing in the heavenly force through the Laogong point. The right palm is still facing down and is yin drawing in the earth force energy.

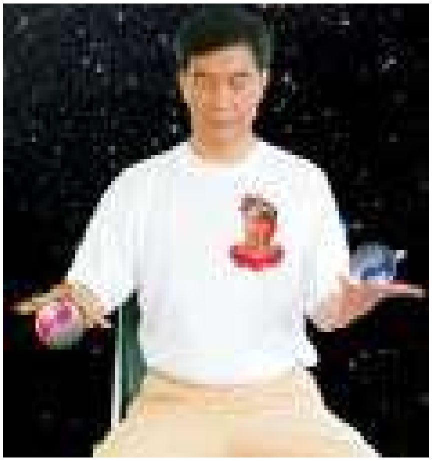

Fig. 8.68 *Separate the Hands, one palm facing up and one down - draw in the Forces of Heaven and Earth.*

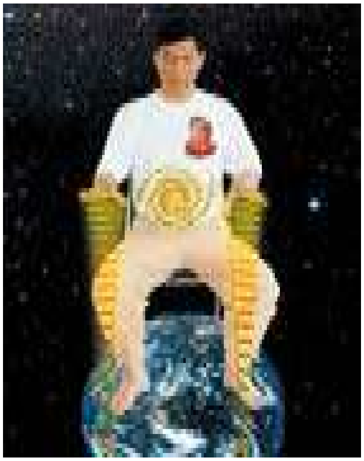

Fig. 8.69
Draw in Earth Force.

Conclude by turning the left palm over to face downward. Draw in the earth force through both palms.

Activate the Chi Ball, Tan Tien and the Original Force

Turn the right palm upward and lower it to the Tan Tien with the Chi knife pointing inwards.

Lower the left palm to the navel, palm down. Lower the left thumb to point toward the right Laogong point. This opens the eye of the left hand (LI-4), pointing it toward the navel.

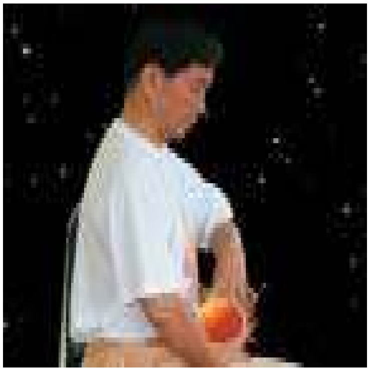

Fig. 8.70 *Activate the Chi Ball, Tan Tien and the Original Force - Right Hand over left, the 'Large Intestine 4' pointing at the Navel.*

Open the Functional Channel with the Left Hand

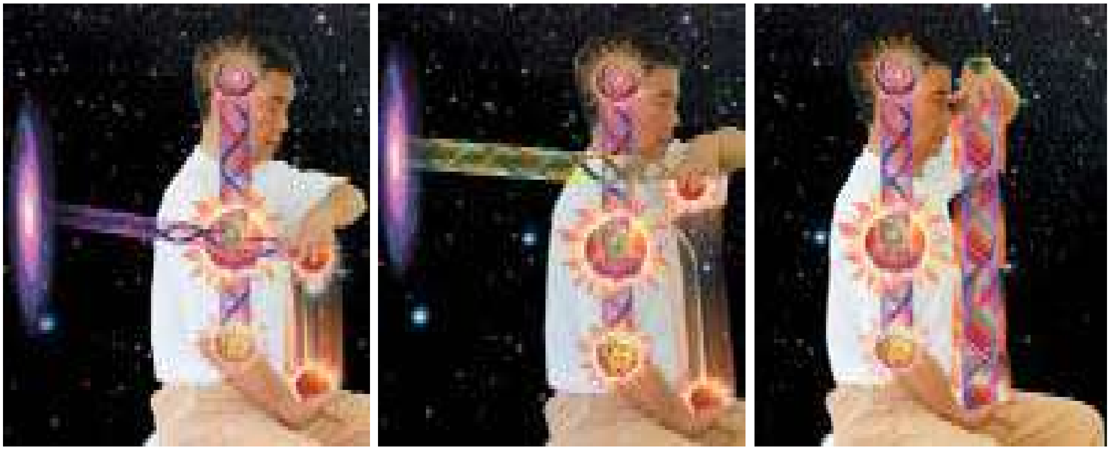

A. *B.* *C.*

Fig. 8.71 A. *Bring LI4 point up to the Solar Plexus and Heart Center.*
B. *Then up to the Throat Center.*
C. *Then up to the Mideyebrow. Descending down the channel in reverse, finishing at the Navel.*

After going up the functional channel come down the channel with the left hand.

Holding a Chi Ball: Open the hands with the palms facing each other holding a Chi ball. Stretch and squeeze the Chi ball a few times in order to feel the Chi as a substance; open and then relax the palms as you squeeze.

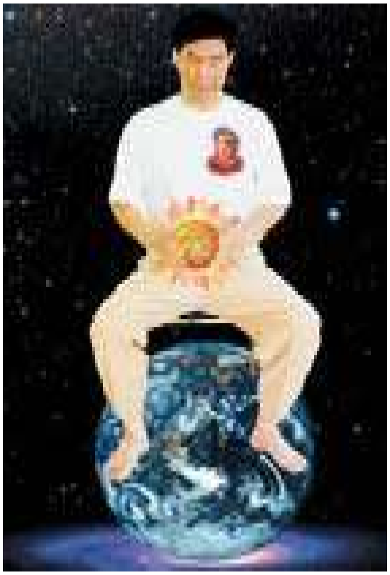

Fig. 8.72 *Hold and squeeze the Chi Ball.*

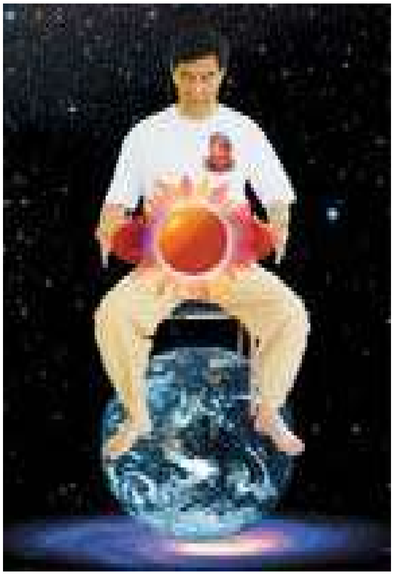

Fig. 8.73
Feel the Chi Ball expanding, growing Bigger.

Double Palm and Double Beam - Palm to Palm
Activate the Functional and Governor Channels - Right Palm on Left

Bring the left palm in first to face the navel, about 1 inch away, aligning the Laogong point with the navel. Follow by bringing the right hand in front of the left hand, **aligning the Laogong of the right palm with the Laogong of the left hand and the navel**. Both palms are now facing in. Beam energy through both Laogong points to the **navel and through the body to the Door of Life**, between L-2 and L-3 (Lumbar Vertebrae). Hold for a 5 to 30 second count and feel the vibration of the palms and Chi beam penetrate through the two points; feel them link together.

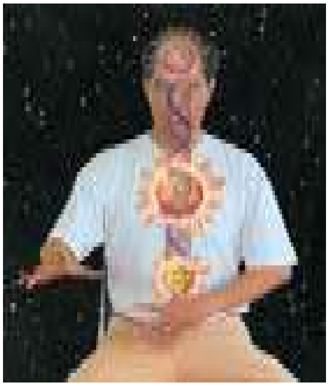

Fig. 8.74
Left Palm to Navel.

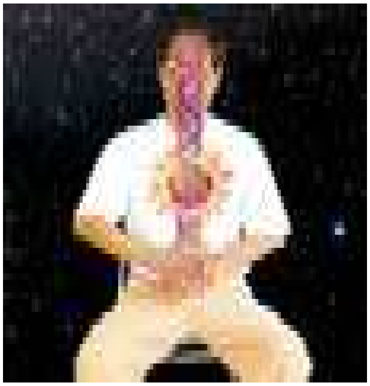

Fig. 8.75
Right on Top.

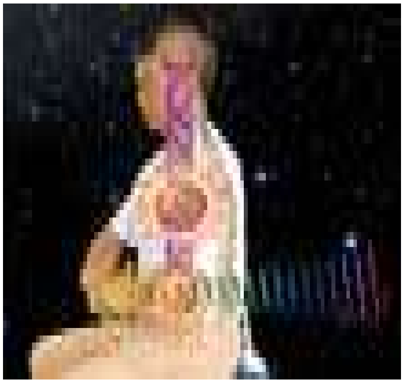

Fig. 8.76
Beam from Navel to Door of Life.

Next raise the left palm up to the solar plexus, aligning the Laogong point with the solar plexus.

Follow with the right palm aligning the Laogong to the left palm. Beam energy into the **solar plexus point and through the body to the T-11 point (thoracic vertebrae)**.

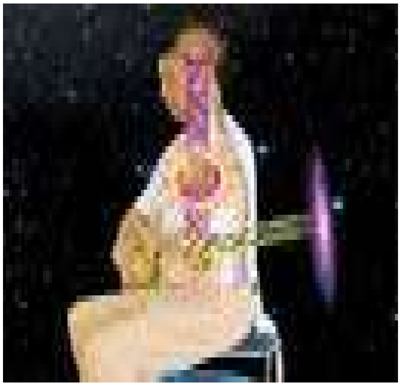

Fig. 8.77
Raise Left Palm to Solar Plexus.

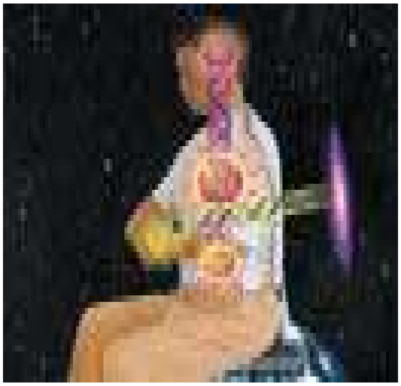

***Fig. 8.78** Follow with the Right Hand to form the Double Beam.*

Continue in the same way with the left hand leading to the **heart point and wing point,** the **throat point and C-7 (cervical vertebrae)**, the **mideyebrow point and Jade Pillow**, and the **crown point and perineum**. **Always remember 'Tan Tien and the universe'.**

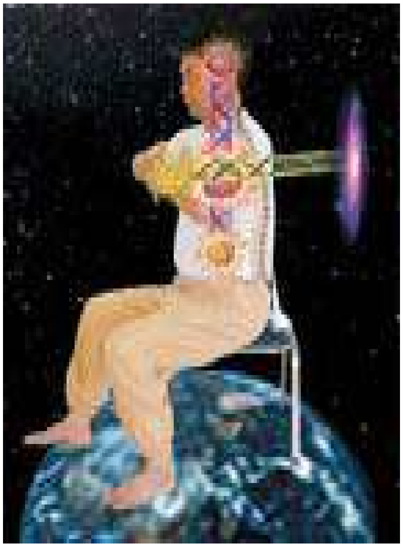

Fig. 8.79 *Beam into the Heart Center and to the back of the Heart.*

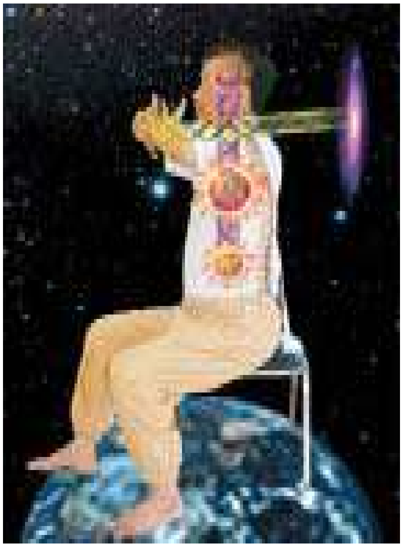

Fig. 8.80 *Beam into the Throat Center and to C7.*

When you reach the crown lightly spiral the palms and feel the Chi slowly penetrate deep into the body reaching the perineum; this will also open the **Thrusting Channels.** This might take a longer time, try a 30 to 60 second count.

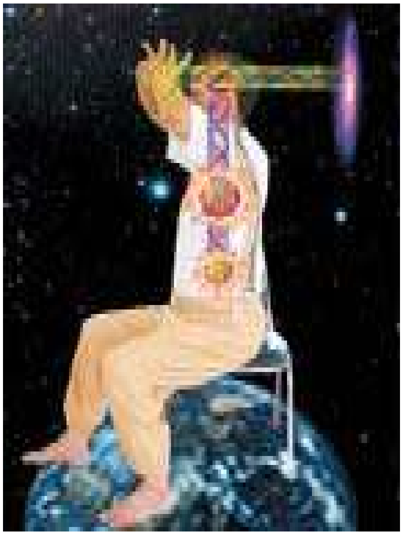

Fig. 8.81 *Beam into the Mideyebrow to the base of the Stall.*

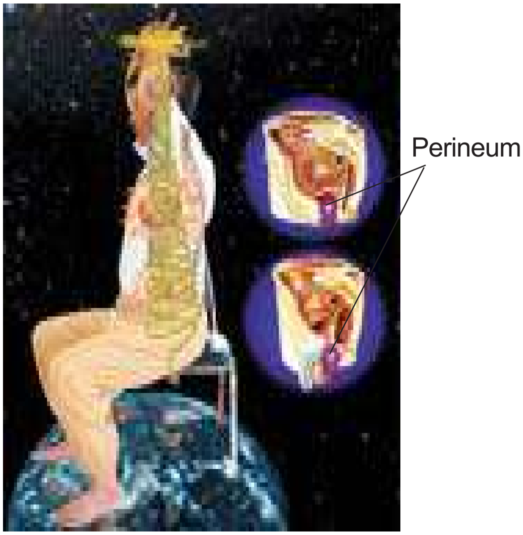

Fig. 8.82 *Beam from the Crown down to the Perineum.*

Return down the front, point by point in the same way, leading with the left hand.

Go up and down this way three times. The double palm and beam activate both the Functional and Governor Channels. The energy beam passes all the way through the body at each point.

Grasping the Chi Ball

Separate the hands and slowly extend the arms in front of you at navel level, palms facing each other, holding the Chi ball.

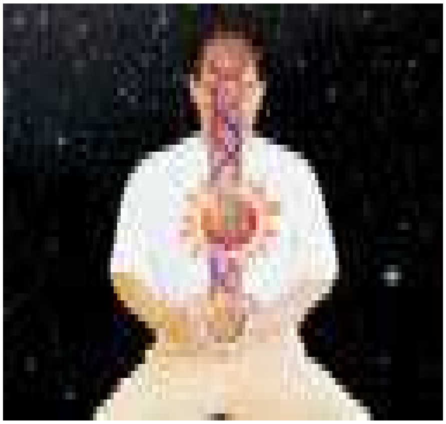

Fig. 8.83 *Squeeze the Chi Ball.*

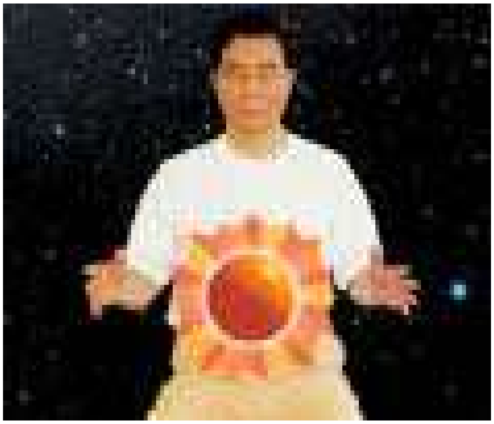

Fig. 8.84 *Magnify its Energy.*

Stretch and squeeze the Chi ball three times as before by opening and closing the palms. 'Be Aware of Tan Tien and Universe'.

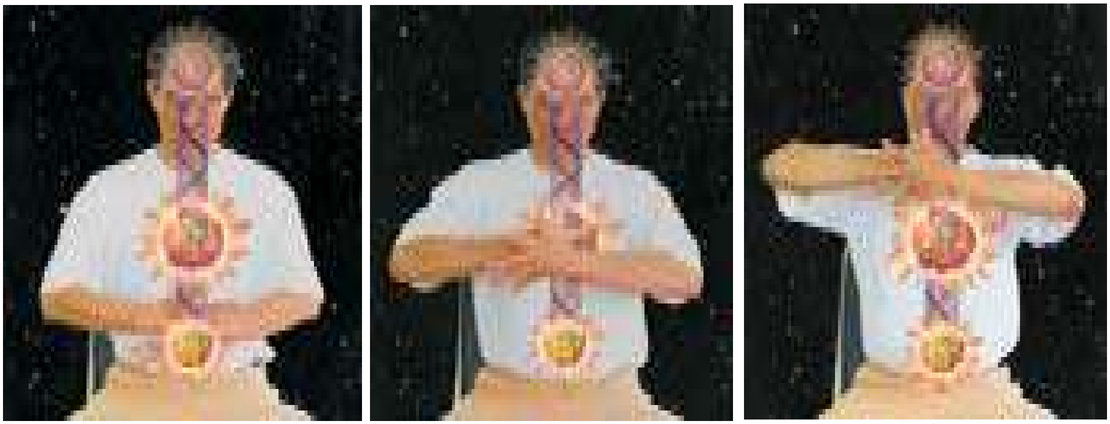

A. ***B.*** ***C.***

Fig. 8.85 *Repeat the double palm and beam with the* ***left hand on the outside and the right palm leading,*** *beaming from;* ***A.*** *Navel to Door of Life.* ***B.*** *Solar Plexus to T11 and Heart Center to the Wing Point* ***C.****Throat Center to C7.*

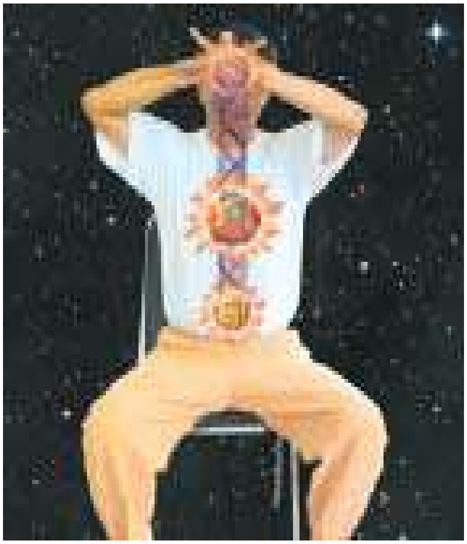

Fig. 8.86 *Activate the Mideyebrow to the Base of the Skull.*

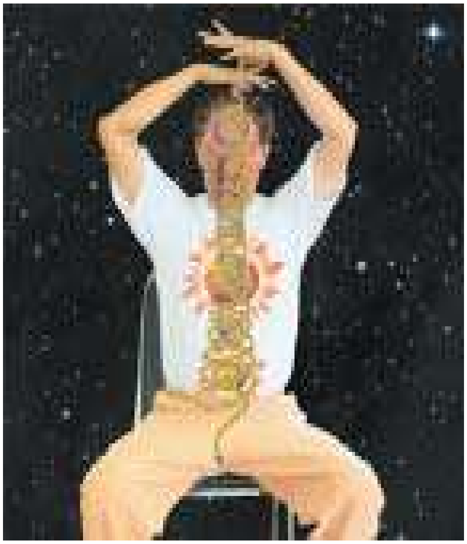

Fig. 8.87 *Activate the Crown and feel the Energy penetrate to the Perineum.*

A. Lower the right hand to the mideyebrow and penetrate to the skull.
B. Lower the right hand down to the throat center, following with the left.
C. Lower down to the solar plexus and penetrate to T11.
D. Lower down to the navel and penetrate to the Door of Life.

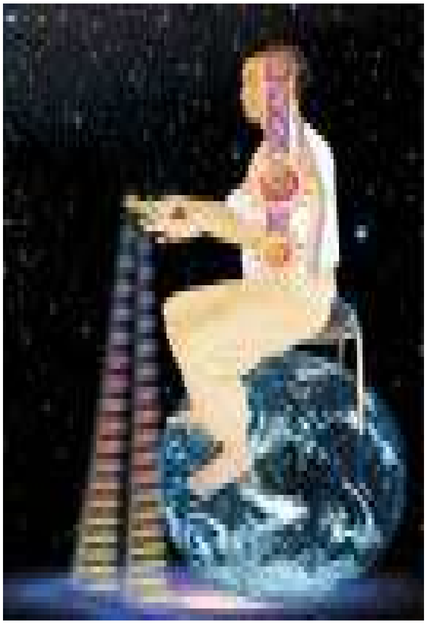

Fig. 8.88 *Channel the Earth Force.*

Rotate the **hands to face downward** and draw in the earth force through the palms, soles and perineum.

Closing Movements

Activate the Chi Channels - Balancing the Organs

The closing movements, crane's beak and finish are the same as before.

***Index Finger** open (up), close (down), neutral (even with the other fingers)*
***Ring Finger** - down, neutral.*
***Index Finger** - open, neutral.*
***Thumb** - in and down, neutral.*
***Index Finger** - open, neutral.*
***Pinky Finger** - down, neutral.*
***Index Finger** - open, neutral.*
***Middle Finger** - down, neutral.*
***Index Finger** - open, neutral.*

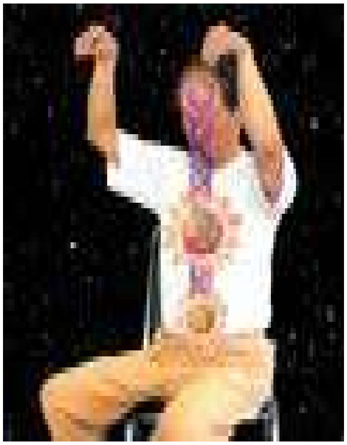

***Fig. 8.89** Form the Crane's Beak.*

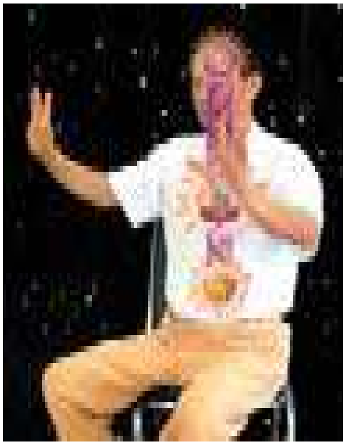

***Fig. 8.90** Palms to Front.*

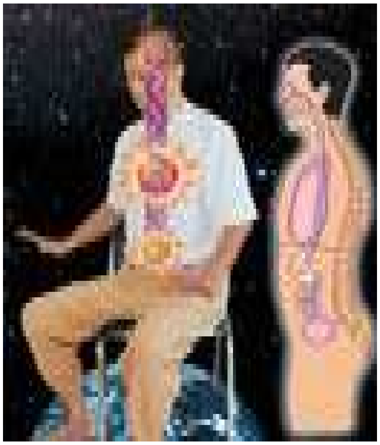

***Fig. 8.91** Lower to Sides.*

Crane's Beak

Form the crane's beak with both hands by bringing the fingertips together with the thumbs inside. Activate the saliva and swallow this elixir down three times to the Tan Tien. Raise the forearms to shoulder height with the fingers pointed down, inhaling gently as you raise.

Open the palms slowly and begin to lower the arms to the sides as you exhale, until the palms are facing each other in the starting position.

Finish

Bring awareness to the navel and notice the quality and intensity of the energy generated. Collect energy at the navel. Men, place the hands over the navel with the right hand on top; vice versa for women.

Part IV

Activating the Yin and Yang Channels and the Chi Belt

Theory

In this final part of Cosmic Chi Kung, you will focus on opening the yin and yang channels of the arms and on activating the Chi Belt around the waist. Mastery of Part IV of Cosmic Chi Kung will enhance your ability to project Chi through your fingers for healing.

Yin and Yang Channels of the Arms

There are six important energy channels flowing through the arms. These channels are paired together with each pair comprised of a yin channel and a yang channel. The yin channel of each pair flows down the inside of the arm from torso to fingertip, while its yang counterpart flows up the outside of the arm from fingertip to head.
The pairs are as follows:

Metal Element - Yin - Lung Channel - Thumb
Yang - Large Intestine Channel - Index Finger
Fire Element - Yin - Pericardium Channel - Middle Finger
Yang - Triple Warmer Channel - Ring Finger
Fire Element - Yin - Heart Channel - Pinky Finger
Yang - Small Intestine - Pinky Finger

Projecting Healing Energy through the Fingers

Emitting Chi from the fingers for healing is known in China as the **"One Finger Art."** Because the various energy channels of the arms either end or begin at the fingertips, the fingers are very effective instruments for projecting healing energy. Your fingers can focus energy like a laser beam toward a concentrated area such as a specific acupuncture point. You can also emit energy from all the fingers at once creating a combined effect to target an area. Part IV of Cosmic Chi Kung will stimulate all of the arm channels and will give you an opportunity to activate all of the fingers for beaming energy.

Fig. 8.92 *"One Finger Art"*

Activating the Chi Belt - Dai Mo

The Chi Belt or Belt Channel (also called the Girdle Vessel) is the only channel in the body that runs horizontally. It encircles the body at the waist level, connecting together all the vertical channels running through the torso. Thus, the Belt Channel plays an important role in maintaining a good energetic communication between the upper and lower body. In women it strongly affects the uterus and the menstrual cycle in particular.

In Taoist Chi Kung, the Belt Channel is not limited to the waist region. It encircles the entire body, almost like weaving an energetic cocoon around you from head to foot. Activating the Belt Channel strengthens the aura and helps to defend and protect you from outside negative energies.

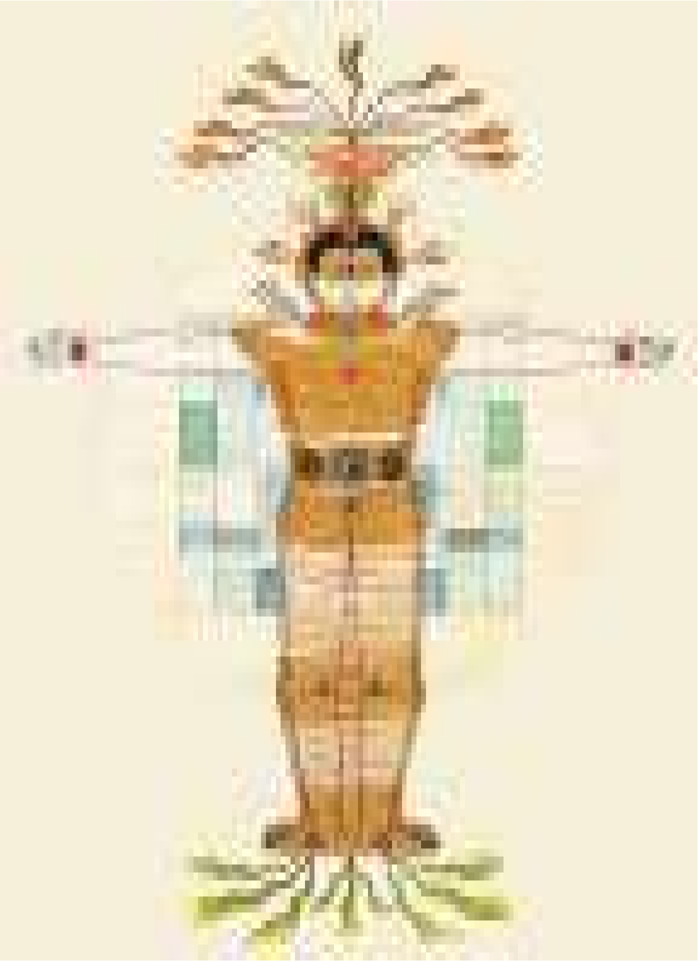

Fig. 8.93 *Belt Channel*

Practice

Opening Movements

These opening movements have been described in Parts I, II, III. Refer to Part I to refresh your practice.

Hands Down at Sides - Palms Facing Each Other - Bone Breathing - Absorbing Cosmic Chi.

Channeling the Earth Force - Washing the Bone Marrow.
Raise arms to chest height - Palms facing each other.
Rotate arms until palms face down.
Open the Tiger's Mouth - Index Finger up, thumb down.

Activate and Beam to the Eye of the Hip.
Large Intestine 4 - Eye of the Hand.
Pericardium 9 - Tip of the Middle Finger.
Small Intestine 3 - Knife Edge of the Hand.

Absorb the Heavenly Force - Washing the Bone Marrow - Crown to Soles.
Extend the arms forward to chest height - palms upward.

Absorb the Earth Force.
Rotate the arms until the palms face down.
Raise the Index Fingers slightly.

Fig. 8.94 *Mother Earth's Healing Power*

Core Movements of Part IV Yin and Yang Channels - Chi Belt

Activating the Yin Channels - Left Side

Turn the palms down and slightly open the index fingers. Smile to your palms and soles and channel the earth force.

Turn your right palm up and pass it just one inch below the left arm yin channels from the left palm to the armpit, **without physically touching the arm**. Feel the Chi moving. Turn your right palm to face your left rib cage and pass your right hand down the left side of your abdomen along the descending colon, and a cross the pelvis to the right side to activate the **yin** energy.

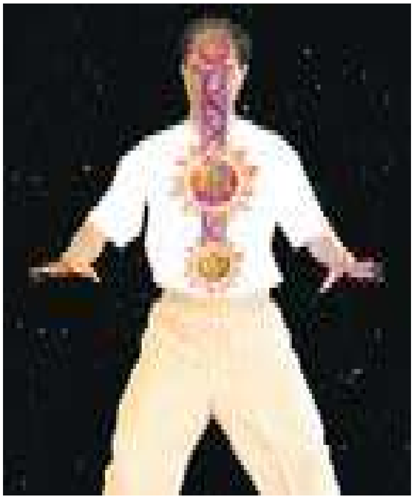

Fig. 8.95 *Channel the Earth Force.*

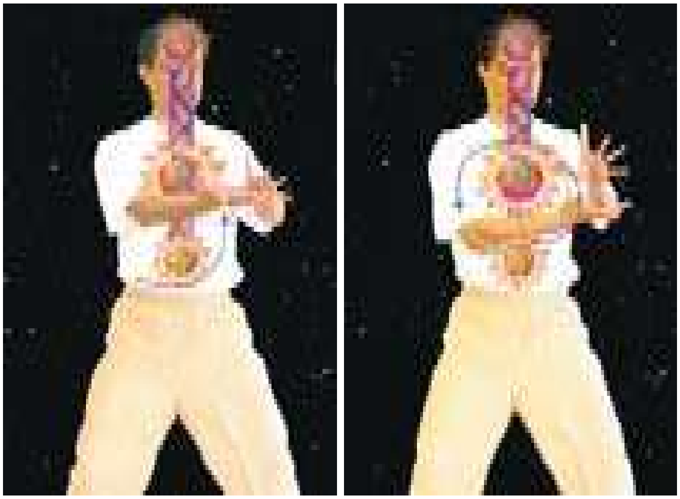

Fig. 8.96 *Activate the Yin Energy from the Palm to the Pelvis.*

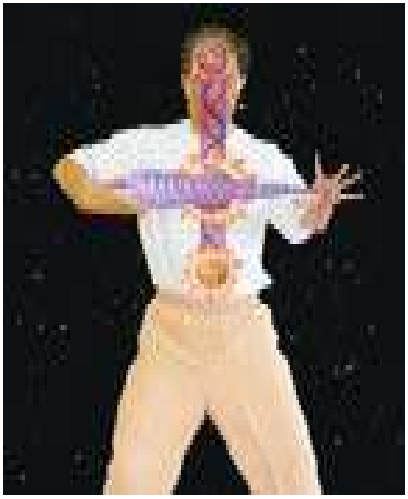

Fig. 8.97 *Raise Chi in the Right Palm.*

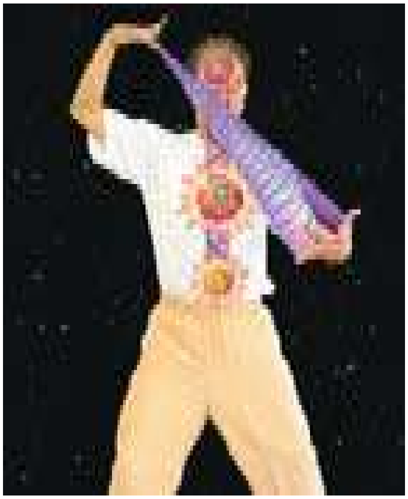

Fig. 8.98 *Pass it to the Left.*

Pass your right hand up the right side of the abdomen to the level of your forehead with the palm facing diagonally toward the left palm (which simultaneously turns palm up at the level of the navel). Project the Chi from palm to palm for 30 to 60 seconds, vibrating the palm and feeling Chi from the right palm sent to the left picking up by the left palm. Keep the palms open and relaxed.

Next point the right ring finger toward the left palm and project Chi making very small circles from the right ring finger (Triple Warmer channel); emit the Chi to the left palm for 30 to 60 seconds.

Then return the right index finger to neutral position. Point the left ring finger up toward the right palm, project into the right palm for 30 to 60 seconds, and then return to neutral.

Next both ring fingers project toward the opposite palms at the same time for 30 to 60 seconds and then return to neutral. Notice that the energy may meet in the middle.

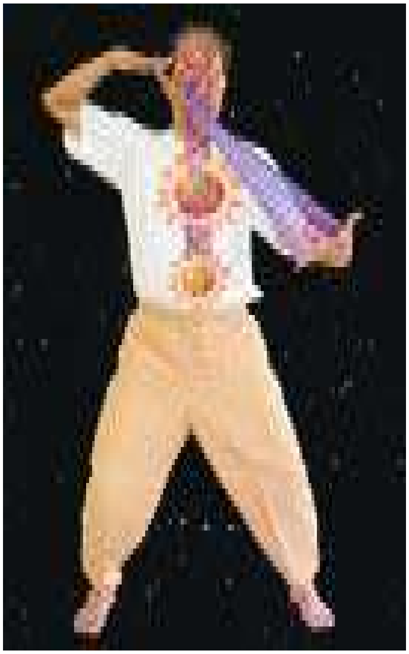

Fig. 8.99 *Project Energy and balance Left and Right.*

Activate the Yang Channels - Left Side

Turn the left palm down and lower the right palm to cover the back of the left hand, palm down. Pass the right palm over the left arm's **yang** channels from the back of the left hand to the left shoulder. Then pass the right palm across the upper chest and across the right breast. Scoop the right palm under the right armpit, fingers first. Then bring the right palm forward, palm down and press forward with the left palm at the same time. Relax and channel the earth force.

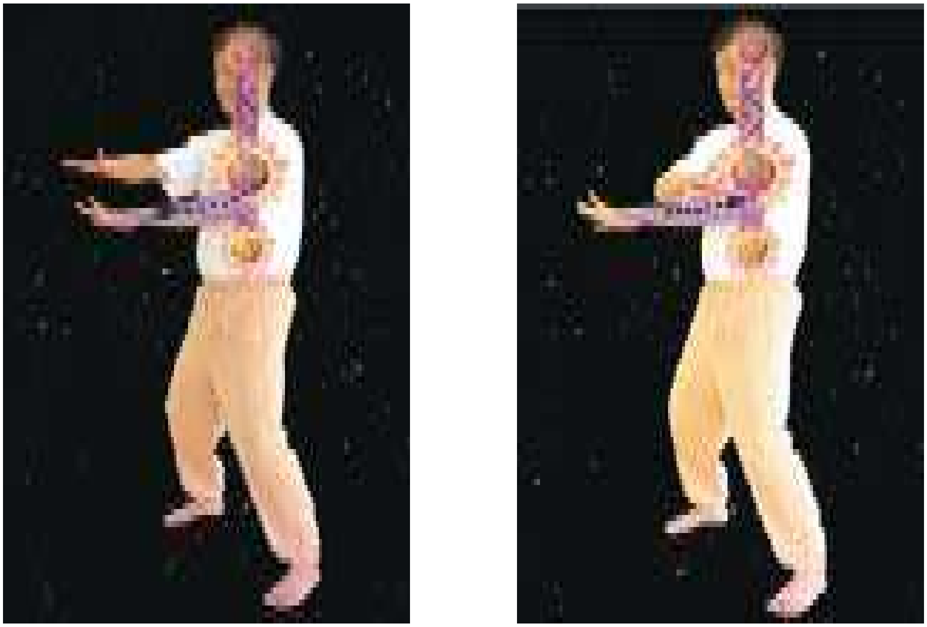

A. *Left Palm down, lower Right Palm to cover the back of Left Hand.*

B. *Pass the Right Palm over the Left Arm's Yang Channels.*

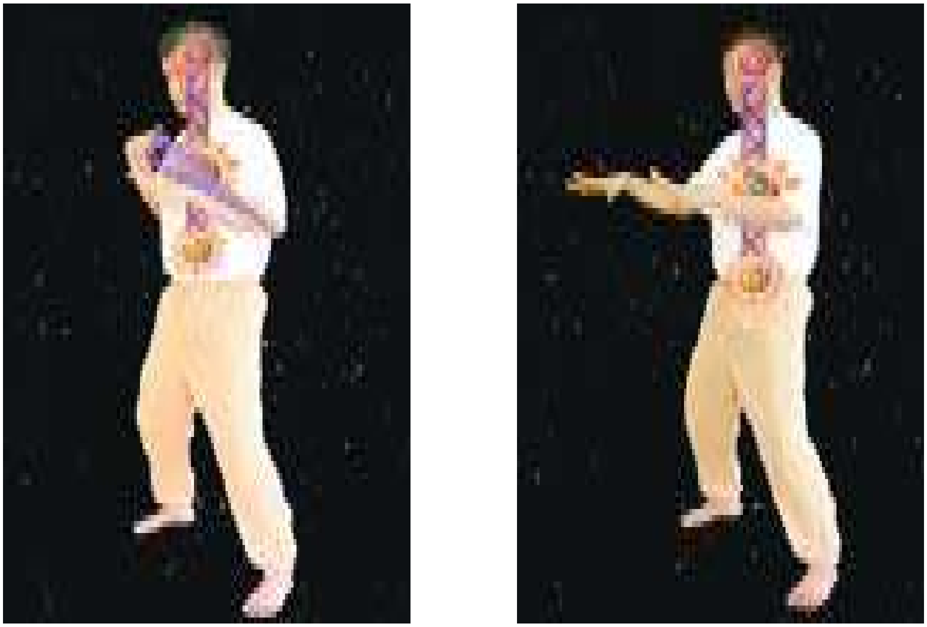

C. *Pass the Right palm across the chest and scoop under the Right Armpit, Fingers first.*

D. *Press both Palms forward, relax and channel the Earth Force.*

Fig. 8.100 *Balance Left Side Yin and Yang and Channel Earth Energy.*

Repeat, Activating the Yin and Yang Channels to the Right Side

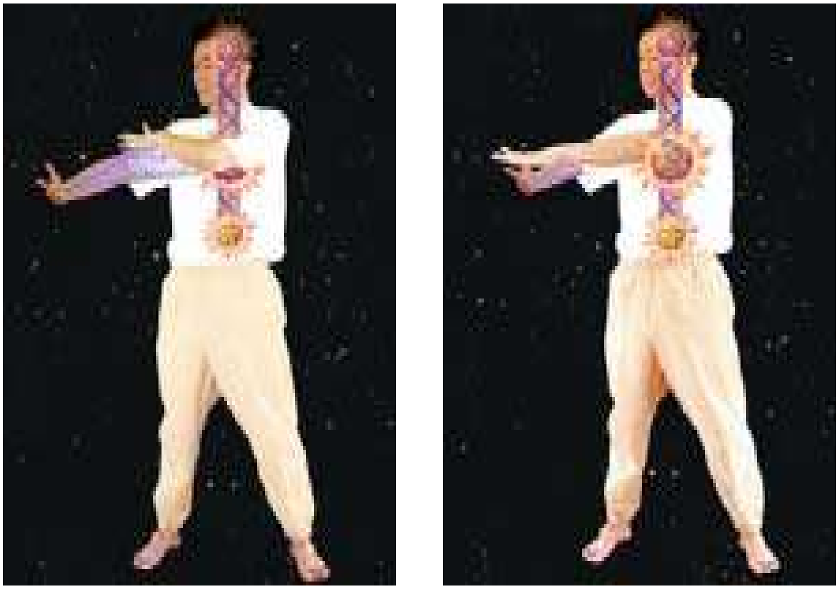

Fig. 8.101 *Balance right side yin and yang and channel earth energy*

Then repeat **Activating the Yin and Yang Channels** with the other fingers and channels in the following order:

Index Finger - Large Intestine
Thumb - Lung
Pinky Finger - Heart and Small Intestine
Middle Finger - Pericardium

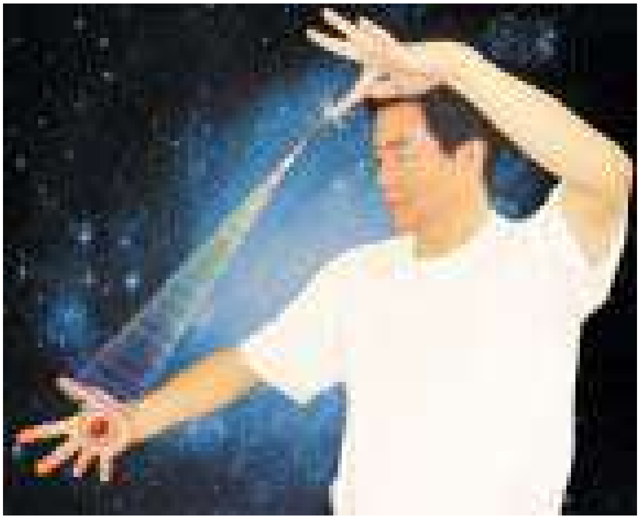

Fig. 8.102 *One Finger Art - activating the Laogong.*

Activate the Chi Belt - Right Foot Forward

Step forward with the right foot. The right palm covers the navel and the left palm covers the Door of Life. Feel the Chi beam penetrate from palm to palm.

Turn your hips to the right and shift weight to your right leg. At the same time move your right palm to the Door of Life in a brushing manner, while your left palm moves to the right hip in a sweeping manner. Note that the palms face the Chi belt.

Repeat the above three times altogether.

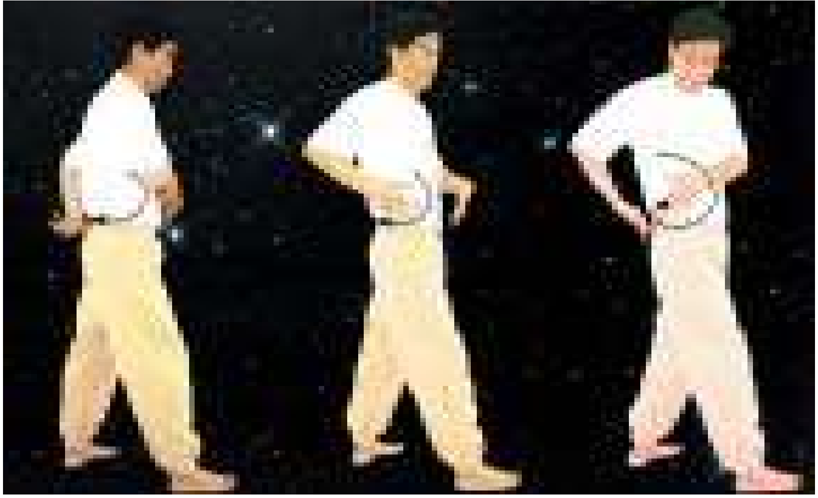

Fig. 8.103 *Activate the Chi Belt - Right Side.*

Activate the Chi Belt - Left Foot Forward

Repeat the above with the left foot forward and reversing the right and left hands.

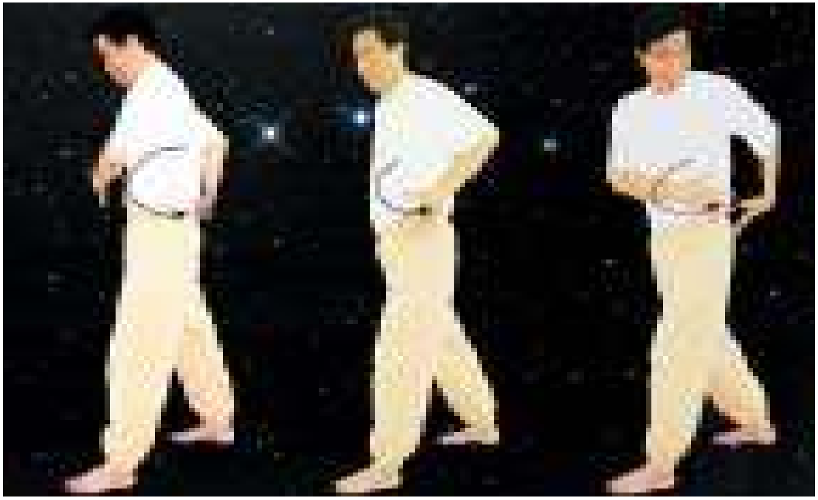

Fig. 8.104 *Activate the Chi Belt - Left Side.*

Channel the Earth Force

Bring the hands forward and rotate the palms to face downwards. **Smile** and channel the earth force as before.

For these Closing Movements see the pictures beginning on page 171.

Closing Movements
Activate the Chi Channels - Balancing the Organs

Index Finger - open (up), close (down), neutral (even with the other fingers).
Ring Finger - down, neutral.
Index Finger - open, neutral.
Thumb - in and down, neutral.
Index Finger - open, neutral.
Pinky Finger - down, neutral.
Index Finger - open, neutral.
Middle Finger - down, neutral.
Index Finger - open, neutral.

Crane's Beak

Form the crane's beak with both hands by bringing the fingertips together with the thumbs inside. Swallow the saliva. Raise the forearms to shoulder height with the fingers pointed down, inhaling gently as you raise.

Open the palms slowly and begin to lower the arms to the sides as you exhale, until the palms have returned to the starting position.

Finish

Bring the awareness to the navel and notice the quality and intensity of the energy generated.

Collect energy at the navel. Men, place the hands over the navel with the right hand on top; vice versa for women.

Healing Others

Summary and Combined Practice

After learning the four sets of Cosmic Chi Kung, you may want to combine them all into a short and simple daily practice. This combined set, previously called Buddha Palm I, synthesizes the movements from the first three sets into one basic sequence. The movements are done on the right side only, so the entire set can be completed in **ten to twenty minutes**.

Cosmic Chi Kung: Simple Combined Sequence

Opening Movements

Bone Breathing - Absorbing Cosmic Chi

1. Stand with your hands down by your sides. Feel your Chi moving and activate the **bone breathing**.
2. Activate the cosmic force by using your mental intention (Yi) to spiral at the **mideyebrow** point and draw in the golden light energy. Breathe through the mideyebrow point in the same manner as in bone breathing, allowing the energy to **flow through the Microcosmic Orbit**. Alternatively you can breathe the golden light down into your navel to charge up your Original Chi at the Tan Tien. You can also simply circulate the golden light throughout your body. Relax, smile and enjoy the sensation of your entire body breathing in the cosmic energy.

Channeling the Earth Force - Washing the Bone Marrow

1. Connect with the earth force by activating the **perineum** and the **soles of the feet, Kidney 1, Bubbling Spring**. Connect the bridge by slightly squeezing the sex organs and the anus. Wash the bone marrow from the soles to the crown.
2. **Smile** into the perineum, palms, mideyebrow and crown. Then slowly raise the arms, palms facing each other to chest height keeping the elbows relaxed and sunk. Then slowly rotate the arms until the **palms are facing down**. Absorb the Earth Chi through the soles, perineum and palms; feel it being absorbed into the bones and body like a **steam or mist rising from the earth**. Feel the earth force move through the **center of the bones**, up the calves, thighs (femur bones), through the hip bones, spine, scapulas, arms, neck and skull. Finally swirl the energy around your brain.

Activate the Hip - Large Intestine 4, Pericardium 9 and Small Intestine 3

1. Raise your index finger slightly upward to activate the Laogong point (Pericardium 8). Then stretch your thumbs out to the sides and down, so that they end up pointing downwards toward the earth to activate Large Intestine 4 (Hegu). **Large Intestine 4 is found in the webbing between the thumb and index finger. It is called the "eye of the hand"; this area is also known as the "Tiger's Mouth."**
2. Slowly draw the elbows back and lower the hands until your palms are facing down beside the "eye of the hip" (the iliac crest) with the "eye of the hand" (Large Intestine 4) aligned with the "eye of the hip". **Use your mind intention to beam the energy back and forth between your two LI-4 points. This will activate the large and small intestines.**
3. Rotate your hands so that the fingers first point out, then back and finally point the fingers in toward the eye of the hip. **Beam energy back and forth between the tips of the middle fingers** (Pericardium 9) passing through the hips.
   Rotate the hands again until the fingers are pointing forward with the palms facing up. Align the knife edge (Small Intestine 3) with the eyes of the hip, and **beam the energy back and forth between the two hands.**

Absorb the Heavenly Force - Washing the Bone Marrow - Crown to Soles

Extend the arms forward to chest height with the **palms still facing upward**.

Become aware of the palms and the crown. Activate the Laogong point (Pericardium 8). Feel the heavenly energy, and be aware of the violet light of the North Star and the red light of the Big Dipper shining down. Draw this light in through the palms and the crown point and feel it washing your bone marrow. Direct the sensations down through the skull, cervical vertebrae, clavicle, scapulas and sternum. Activate your thymus gland and continue drawing the healing heavenly light down through the rib cage, arms, spine, hips, femurs, calf bones, feet and toe bones. **Feel your bones activated with a positive electrical force (the positive force comes from heaven, negative from the earth).**

Absorb the Earth Force

1. Rotate the arms until the **palms face down**.
   Raise the index fingers slightly, extend the thumbs first towards each other, and then towards the ground.
2. Become aware of the palms (Laogong), the soles of the feet (Kidney 1, Bubbling Spring) and the perineum (Conception Vessel 1, Hui Yin). Activate these points, feel the earth energy, and **be aware of a gentle blue light or white light rising up from the earth like a pure steamy mist.** Draw this light in through the palms and soles and let it wash up through the center of your bones as it steam cleans your

marrow. Picture any impurities or illness dripping out of your bones and draining down into the earth, where they are recycled and purified.

Core Movements

Grasping the Moon - Connecting the Heaven and Earth Forces

1. **Open Position:** Heaven Force - Bring the arms up to shoulder height, bending the elbows so that the **left hand is under the right elbow** with the left index finger pointing up to Heart 3.
2. The right forearm is over the left forearm with the tip of the right index finger held above the Large Intestine 11 point of the left arm. The right index finger is pointing up toward heaven.
3. The right index finger draws in the heavenly force like an antenna. The force flows up the bones of the right arm across the shoulders to the left arm, and through the bones of the left arm to the left index fingertip. It then flows through the index finger of the left hand into the Heart 3 point of the right arm **completing the heavenly circuit**. Continue cycling in this way.
4. **Close Position:** Earth Force - Keeping the arms in the same position, move both index fingers to point down. The right index finger will now point toward the Large Intestine 11, point of the left arm and the left index finger will point downward toward the earth. Draw in the earth force through the left index finger bringing it through the bones of the left arm, across the shoulders through the right arm to the right index finger and out into the Large Intestine 11 point of the left arm. Continue cycling in this way **completing the earth circuit.**

Open and close three times altogether ending in the open position.

Activate the Throat Center

1. Draw the elbows out to the side and rotate the palms to face in so that the left palm is about 3 - 6 inches in front of the Throat Center (Conception Vessel 22), palm facing in, and the right hand is about 1 - 3 inches in front of the left hand, palm facing in. Align the Lung 10 point of the right hand with the Lung 10 of the left hand, and align the Lung 10 points in the middle of the thenar eminence of the left hand with the Throat Center. **Keep the points aligned and beam the energy passing through both hands into the Throat Center.**
2. Move both hands about 12-18 inches out from the Throat Center maintaining the energy beam as you move.
3. Move in and out three times altogether in this way ending with the hands in as in preivous point above.

Activate the Third Eye

1. Keeping the Lung 10 points aligned, raise the hands up to the level of the mideyebrow point. Keep the points aligned and **beam the energy passing through both hands into the mideyebrow point.**
2. Move both hands about 12-18 inches out from the mideyebrow point maintaining the energy beam as you move. **Move in and out three times** altogether in this way, ending with the hands as previously.
3. Keeping the palms and thenars (Lung 10) aligned, move the hands back down to the position in front of the **Throat Center.**

Open the Bridge and Regulator Channels

1. Keeping the forearms horizontal at shoulder height, separate the hands so that the Laogong points are aligned in front of Stomach 13 just below the clavicle on a vertical line above the nipple or **mammillary** line. The fingertips should almost be touching. Beam energy from the Laogong points into ST-13. You will be following this vertical line, which passes through the nipples, down the torso to the level of the navel stopping to focus and **beam energy into key points along the way**. On each of these, focus your Yi on beaming energy into the organs inside each position. Stomach 13 is the location that activates the Heart and Lungs. Feel the change in your breathing as you activate the **lungs**.
2. Move the hands down the mammillary line to Stomach 16, about one inch above the nipple (in the space between the third and fourth ribs). Again, beam energy to energize and balance the **heart and lungs**.
3. Move the hands down to Liver 14 (about 2 - 3 inches below the nipple in the space between the 6th and 7th ribs). Beam healing energy into the **liver and gall bladder**.
4. Move the hands down to Spleen 16 (just below the rib cage on the mammillary line). Beam healing energy into the **stomach, pancreas and spleen**.
5. Move the hands down to Spleen 15 (on the mammillary line to the level of the navel). Beam healing energy from your palms into the **small intestine and the Tan Tien**.

Grasping the Chi Ball

1. Slowly rotate the hands so that your palms face each other in front of your navel. Feel the invisible ball of energy between your hands. Allow the energy to push your hands apart keeping the feeling of connection between your palms.
2. Let the energy draw your hands back toward each other, like a magnet, until you feel as if you are squeezing a ball.
3. Stretch and squeeze the Chi ball three times by opening and closing your palms in this way.

Activate the Outer and Inner Arm Gates
Waiguan - Triple Warmer 5 and Neiguan - Pericardium 6 - Right Arm on Top

1. Slowly rotate the arms so that both palms face down.
2. Cross the right wrist over the left wrist with 1-3 inches of space between the hands. Align PC-6 of the right wrist with TW-5 of the left wrist.
3. Keeping the wrists crossed, slowly rotate hands turning the palms upward. Now TW-5 of the right wrist will be aligned with PC-6 of the left wrist.

Open the Functional Channel with LI-4

1. Bring the left hand down to the level of the Tan Tien with the Chi Knife facing in. Lower the right hand, palm down, and align the eye of the hand, the LI-4 point with your navel. Feel the Chi ball between your two hands, and at the same time beam energy from LI-4 into the navel point.
2. Raise the right hand up the Functional Channel stopping and aligning LI-4 with the **solar plexus, heart, throat** (CV-22) and **mideyebrow** points. **Beam energy from the eye of the hand into each of these points**, and at the same time continue to feel the Chi ball connection between the two hands.
3. Return down the Functional Channel in the same manner stopping to beam energy into each point. **Go up and down the Functional Channel** in this way three times.

Yin/Yang Palms

1. Separate the hands and extend them out in front of the body at solar plexus level. The left palm is still pointing up and is yang drawing in the heavenly force through Laogong. The right palm is still facing down and is yin drawing in the earth force energy.
2. Conclude by turning the left palm over to face downward. Draw in the earth force through both palms.

Double Palm and Beam to Activate the Functional and Governor Channels

1. Bring the left palm in first to face the navel, about 3 - 5 inches away, aligning the Laogong point with the navel. Follow by bringing the right hand in front of the left hand, aligning the Laogong of the right palm with the Laogong of the left hand and the navel. Both palms are now facing in. **Beam energy through both Laogong points to the navel and through the body to the Door of Life**.
2. Next raise the left palm up to the **solar plexus** aligning the Laogong point with the solar plexus. Beam energy into the solar plexus point and through the body to the T-11 point. Your right hand remains at the navel level beaming energy to the navel and the Door of Life.
3. Raise your right hand up to the outside of the left hand at the solar plexus level, and beam energy through both hands into the **solar plexus point and the T-11 point**. Continue in the same way with the left hand leading to the **heart point and wing**

point, the **throat point and C-7**, the **mideyebrow point and jade pillow** and the **crown point and perineum**.

4. Return down the front, point by point in the same way, leading with the left hand. Go up and down in this way three times. **The double palm and beam activate both the Functional and Governor Channels. The energy beam passes all the way through the body at each point**.

Grasping the Chi Ball

1. Separate the hands and slowly extend the arms in front of you at navel level, palms facing each other holding the Chi ball.
2. Stretch and squeeze the Chi ball three times as before by opening and closing the palms.

Channel the Earth Force

Rotate the hands to face downward, and draw in the earth force through the palms, soles and perineum.

Closing Movements

Activate the Chi Channels - Balancing the Organs

Index Finger - open (up), close (down), neutral (even with the other fingers).
Ring Finger - down, neutral.
Index Finger - open, neutral.
Thumb - in and down, neutral.
Index Finger - open, neutral.
Pinky Finger - down, neutral.
Index Finger - open, neutral.
Middle Finger - down, neutral.
Index Finger - open, neutral.

Crane's Beak

1. Form the crane's beak with both hands by bringing the fingertips together with the thumbs inside. Raise the forearms to shoulder height with the fingers pointed down, inhaling gently as you raise.
2. Open the palms slowly and begin to lower the arms to the sides as you exhale, until the palms are facing each other in the starting position.

Finish

1. Bring the awareness to the navel and notice the quality and intensity of the energy generated.
2. Collect energy at the navel. Men, place the hands over the navel with the right hand on top; vice versa for women.

Chapter IX
Universal Tao Practices

Cosmic Inner Smile

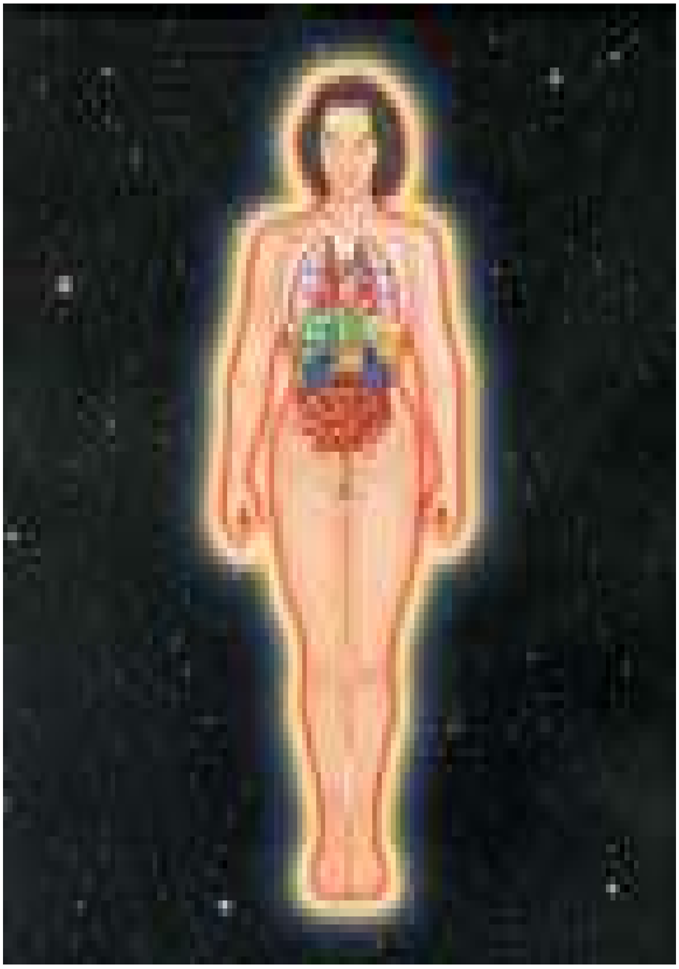

Fig. 9.1 *Inner Smile*

The Inner Smile begins at the eyes and the mideyebrow point and moves down to the heart. As you activate the heart, the loving energy will flow out and you will feel the energy of your Inner Smile flow down the entire length of your body like a waterfall. This is a very powerful and effective tool to counteract stress, tension and negative Chi.

1. Be aware of the **mideyebrow,** imagining that your are in one of your favorite places in the world, a place where you feel safe, relaxed and happy. Recall the sights you might have seen there, the sounds you might have heard, the scents, sensations and flavors that you associate with that place.
2. Imagine that one of your favorite people is standing in front of you, smiling to you with loving, happy, radiant shining eyes. **Smile** to your face and slightly lift up the corners of your mouth.
3. Feel yourself responding to that special person's smile with a smile of your own. Feel your eyes smiling and relaxing.
4. Smile down to the **thymus gland** and picture a white flower blossoming. Gently inhale into the thymus gland, connecting your breath to the olfactory organ. Smell the good fragrance.
5. Aim your inner attention at your **heart**, picturing the heart before your inner eye and smile to it. **Smile** until you feel the heart smile back to you. Picture your heart like a red rose, gradually opening. This will activate the **love** and fire of **compassion** in the heart. Once you feel the red light and loving awareness it will activate the **cosmic red healing light** or mist from above and around you.
6. Smile at the light or mist and very slowly with a soft, long, deep breath draw the red mist, love and compassion into the mideyebrow, down through the mouth and throat, into the heart and gradually overflowing to the **small intestine**. Exhale, but retain the red light and the love and compassion in the **heart and small intestine**. At the same time exhale the cloudy, black or negative energy. Keep on doing this breathing 18 to 36 times waiting until the heart becomes bright red before your inner vision and starts to radiate out to the tongue, mouth, nose, ears and eyes. Allow the red light to whirl around you and form a **red aura**. Feel your skin glowing with red energy.
7. Let the heart's loving energy radiate out to the **lungs**. Aim your attention at the lungs; picture them before your inner eye and smile to them. **Smile** until you feel the lungs smile back to you. Picture your lungs like a white rose, gradually opening; smell the good fragrance. This will activate the courage in the lungs. Once you invoke the white light and courage into the lungs, you will also activate the **cosmic white healing light** or mist from above and around you.
8. Smile at the light or mist and very slowly, softly take a long deep breath and draw the white mist into the mideyebrow, the mouth and down to the lungs, gradually overflowing into the **large intestine**. Exhale but retain the white light and the feeling of courage in the lungs. At the same time exhale the cloudy, black or negative energy. Keep on doing this kind of breathing 18 to 36 times, waiting until the lungs become bright white and start to radiate out to the nose, ears, eyes, tongue and mouth. Invite the white light to whirl around you and form a **white aura** covering your skin like autumn dew.
9. The **spleen, pancreas and stomach** correspond to the yellow color of the earth element. Connect to these organs and smile down into them from the mideyebrow.

First connect to the heart, then attract the clear yellow light from above and around you. You might see the **golden yellow aura** of a wheat field ready for harvest. Expel feelings of worry; exhale the cloudy, sticky energy. Breathe in the golden yellow aura, filling your spleen center with golden yellow light. Repeat to complete 18 to 36 cycles, then allow the light to radiate out to your mouth, nose, ears, eyes and tongue. Wrap the golden aura around you, leaving a golden shine on your skin.

10. The **kidneys and bladder** correspond to the blue color of the water element. Gentleness is contained within the blue healing light. Inhale the blue aura into the kidneys. Expel feelings of fear or stress on breathing out. Allow your energy field to expand, breathing in and out the blue energy 18-36 times, until it starts to radiate out from your kidneys to your ears, eyes, tongue, mouth and nose. Gather the blue mist on your skin enveloping you with a **blue aura**.
11. The **liver and gall bladder** connect to the green color of the element wood; kindness replaces anger, which is expelled on breathing out. Do the same as above and create a **green aura** around you. Exhale the dark red, cloudy heat. Inhale the nourishing green of the forests. Do 18 to 36 cycles, until the green light has completely filled the liver and starts to radiate out to your eyes, tongue, mouth, nose and ears. Invite the green light to form a green aura around you.
12. **Smile** down to the **sexual organs** and reproductive system. Feel love and sexual arousal and the heart and sexual organs uniting. Observe how this process transforms the sexual energy into Chi. Channel down orange and red light to the sexual organs. Thank them for their work in keeping you alive and healthy. Rest. Do nothing. Gather and store the energy by smiling and spiraling in the Tan Tien.

Six Healing Sounds

The Six Healing Sounds enhance Therapeutic Work

Everyone has heard stories about gifted beings who possess great healing powers. People seek out gifted healers. Yet even in the best of circumstances, how much time can a great healer spend with you? One hour a day? An hour a week? What about the rest of the time? Thus you must learn how to take care of yourself; you must learn how to clear out your negative energy and transform it into positive healthy energy. If you maintain yourself in this way it will surely enhance any other therapy you are receiving.

The Six Healing Sounds are a simple yet powerful tool to promote physical, energetic and emotional healing and balance. If you are healing others, you can teach your students one or two of the Six Healing sounds each session, to enhance the effects of your therapeutic work. During the practice of the sound the resting period is very important. Growing the good virtue of the organs is essential so the negative or sick energy has less room to grow.

The sounds are also used to generate certain frequencies for specific healing. Each sound can generate different energy for the healing of the different organs.

Once the movements, sounds and information have been integrated into your practice you can simplify and make the Six Healing Sounds more powerful. For example, when you are settled into the lung sound and you are breathing white healing light and feel the mideyebrow wide open, you put the awareness in the 'Tan Tien and Universe'.

Chi will flow from the universe into the Tan Tien and just the right amount and quality will flow from the Tan Tien into the lungs and large intestine. Just set up the location (the organ) and connect to the universe and the Chi will flow to the correct location!

When doing these, the eyes are open only while making the sound.

Six Healing Sounds - Practice

Lung Sound

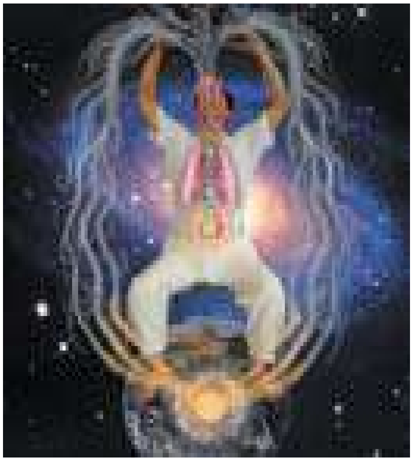

Element: Metal
Associated Organ: Large Intestine
Sound: SSSSSSS (Tongue behind Teeth)
Emotions: Negative - Grief, Sadness, Depression
Positive - Courage, Righteousness, High Self Esteem
Color: White, Clear, Metallic
Season: Fall
Direction: West

Position: Sit in a chair with your back straight and your hands resting palms up on your thighs. Have your feet flat on the floor about a hip's width apart . **Smile** down to your lungs and be aware of any sadness, grief or excess heat in your lungs. Slowly inhale and raise your hands up your center line with your fingers pointing toward each other. When your hands pass your shoulder level, begin to rotate the palms out as you continue raising your hands up in front of you and above your head with the palms up. Point your fingers towards the fingers of the opposite hand and keep your elbows slightly bent.

Sound: Part your lips slightly, keeping your jaw gently closed. Look up through the space between your two hands and push your palms slightly upward as you slowly exhale and make the sound, "SSSSSSS". Picture and feel any excess heat, sadness, grief, depression, sickness and dingy white color expelled and released as you exhale slowly and fully.

Resting Posture: When you have completely exhaled, rotate the palms to face downward with the fingers still pointing toward each other. Slowly lower the palms and bring them just in front of the chest feeling the lung's aura.

Close your eyes and be aware of your lungs. **Smile** into your lungs, and as you inhale, imagine that you are breathing in a **bright white mist of light**. Breathe this light into your lungs and feel it cooling, cleansing, invigorating, healing and refreshing your lungs. Feel it flowing down to the large intestine to balance the energy of the yin lungs and yang large intestine, allowing the **courage** quality of your lungs to emerge. Grow more courage, so sadness and depression have less room to grow. With each inbreath, feel yourself drawing in cool fresh energy. With each outbreath, mentally make the Lung Sound and release any remaining sadness or hot energy.

Repeat at least three times. For the first two repetitions, you can make the sound aloud. On the third or last repetition, make the sound sub-vocally (vocalizing the sound so softly that only you can hear it) Repeat six, nine, twelve or twenty four times to alleviate extreme sadness, depression, colds, flu, toothaches, asthma and emphysema.

Kidney Sound

Element: Water
Associated Organ: Bladder
Sound: CHOOOOOO (with your lips forming an "O" as if blowing out a candle.)
Emotions: Negative - Fear, Shock
Positive - Gentleness, Wisdom
Color: Dark Blue or Black
Season: Winter
Direction: North

Position: Now move the hands to cover the kidneys. **Smile** to your kidneys, and be aware of any excess cold or heat in the kidney region. Then bring your legs together, ankles and knees touching. Lean forward and clasp the fingers of both hands together around your knees. Inhale and pull your arms straight from the lower back while bending the torso forward (this allows your back to protrude in the area of the kidneys). Tilt your head upward as you look straight ahead, still pulling on your arms from the lower back. Feel your spine pulling against your knees.

Sound: Round the lips slightly and slowly exhale while making the sound "CHOOOOOO." Simultaneously contract your abdomen, pulling it in toward your kidneys. Imagine any fear, sickness, imbalances, excess cold or heat energy being released and squeezed out of the fascia surrounding the kidneys.

Resting Posture: After you have fully exhaled, slowly sit erect and return the hands to touch the aura of the kidneys. Close your eyes and again be aware of your kidneys. **Smile** to your kidneys and, on the inbreath, imagine you are breathing a brilliant **luminous blue light mist** into them; feel this mist healing, balancing and refreshing your kidneys and bladder and picture them glowing a bright blue color. On the outbreath, imagine you are still making the Kidney Sound.

Repeat at least 3 times, as with the previous sound. Repeat six, nine, twelve, or twenty four times to alleviate extreme fear, fatigue, low-pitched ringing in the ears, dizziness, back pain, bladder and urinary infections or other problems of the reproductive system.

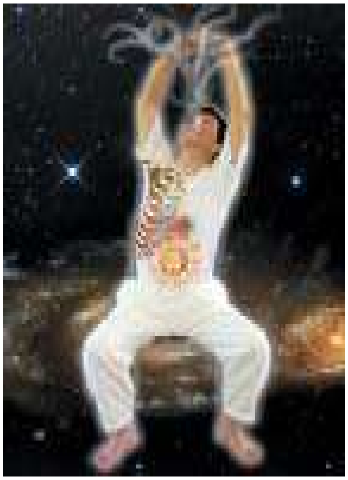

Liver Sound

Element: Wood
Associated Organ: Gall Bladder
Sound: SHHHHHHH
Emotions: Negative - Anger, Frustration, Resentment
Positive - Loving Kindness, Benevolence, Forgiveness
Color: Green
Season: Spring
Direction: East

Position: Place the hands on the liver. **Smile** to your liver and be aware of any anger, frustration, resentment or excess heat. Slowly begin to inhale a deep breath as you extend your arms up from the sides with your palms up. Continue to raise them over your head. Interlace your fingers together and turn your joined hands over to face the sky, palms up. Push out through the heels of the palms and extend the arms up, keeping the shoulders relaxed. Bend a little to the left and stretch your right arm slightly to gently open the area of your liver.

Sound: Open your eyes wide (the eyes are the sensory opening of the liver). Slowly exhale, making the sound, "SHHHHHHH" sub-vocally. Feel that you are releasing any trapped excess heat, anger, illness and negativity from your liver and that these are riding out of your body on your breath.

Resting Posture: Once you have fully exhaled, separate your hands, turn the palms down and slowly lower your arms to the sides, leading with the heels of the hands. **Smile** and inhale a **shiny spring green mist**, illuminating the liver and gall bladder. Bring your hands back to rest on the liver's aura. Close your eyes and smile into your liver. With each inbreath, breathe fresh Chi into your liver and gall bladder and mentally make the Liver Sound with each outbreath.

Repeat at least three times, as with the previous sound. Repeat six, nine, twelve or twenty four times to alleviate extreme anger, to relieve red or watery eyes, to remove a sour or bitter taste in the mouth or to detoxify the liver.

Heart Sound

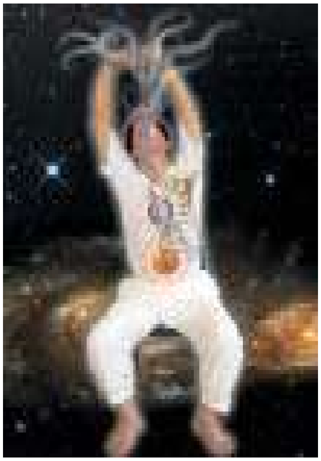

Element: Fire
Associated Organ: Small Intestine
Sound: HAAAAAAAW
Emotions: Negative - Arrogance, Harshness, Cruelty, Hatred
Positive - Joy, Honor, Respect, Love, Happiness
Color: Red
Season: Summer
Direction: South

Position: Let both hands rest on the heart. **Smile** to your heart and be aware of any arrogance, haughtiness, hatred, giddiness, cruelty or hastiness. Slowly begin to inhale a deep breath as you extend your arms up from the sides with your palms up, as you did with the Liver Sound. Continue to raise them over your head. Interlace your fingers together and turn your clasped hands over to face the sky, palms up. Push out through the heels of the palms and extend the arms up, keeping the shoulders relaxed. Bend a little to the right and stretch your left arm slightly to gently open the area of your heart.

Sound: Keep your eyes soft and relaxed, looking up through your hands. Slowly exhale, making the sound, "HAAAAAAAW" sub-vocally. Feel that you are releasing any trapped heat, negative emotions, illness and imbalances from your heart and that these are riding out of the body on your breath.

Resting Posture: Once you have fully exhaled, close your eyes, separate your hands, turn the palms down and slowly lower your arms to the sides, leading with the heels of the hands. As you move inhale a **bright red mist** into the heart and small intestines. Bring your hands back to rest on your heart's aura . Smile into your heart. With each inbreath, breathe fresh Chi into your heart with each outbreath mentally repeat the heart sound.

Repeat at least three times, as with the previous sound. Repeat six, nine, twelve or twenty four times for the heart and small intestines; feel them balanced. This will alleviate extreme impatience, hastiness, arrogance, nervousness, moodiness, jumpiness, irritability, tongue ulcers, palpitations, sore throat, heart disease, insomnia and detoxify the heart.

Spleen Sound

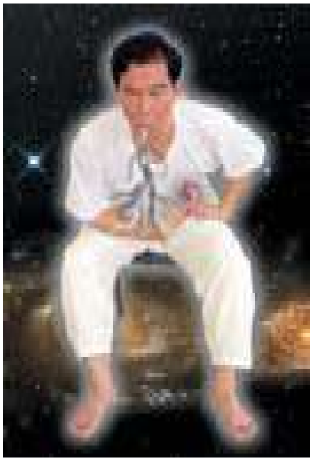

Element: Earth
Associated Organ: Pancreas, Stomach
Sound: WHOOOOOOO (gutturally from the throat)
Emotions: Negative - Worry, Excess Sympathy, Overthinking
Positive - Fairness, Balance, Equanimity, Justice, Openness
Color: Yellow
Season: Indian Summer
Direction: Center (where you stand, looking out to the Six Directions)

Position: Place your hands on the body covering the spleen, pancreas, and stomach area. Be aware of your spleen and **smile** sincerely into it. Inhale deeply as you move the arms outward in an embrace and aim the fingers up under the left rib cage. Place your fingers of both hands just beneath the sternum and rib cage on the left side.

Sound: Look out, lean into your fingers and gently push your fingertips in. Exhale slowly and make the sound, "WHOOOOOOO," from the depths of your throat. Feel yourself releasing any trapped heat, worry, mental fixations or excess sympathy.

Resting Posture: Once you have fully exhaled, close your eyes, slowly release the hands and extend the arms out embracing the earth; return your hands to the resting position on the spleen's aura. Smile to your spleen, pancreas and stomach. Inhale fresh Chi to your spleen, pancreas and stomach as a brilliant **luminous yellow healing mist** that cleanses and refreshes your organs. Mentally do the Spleen Sound with each outbreath.

Repeat at least three times, as with the previous sound. Repeat six, nine, twelve or twenty four times to alleviate extreme indigestion, heat or cold in the stomach or spleen, worry, nausea, hemorrhoids, fatigue, organ prolapse or loose stools.

Triple Warmer Sound

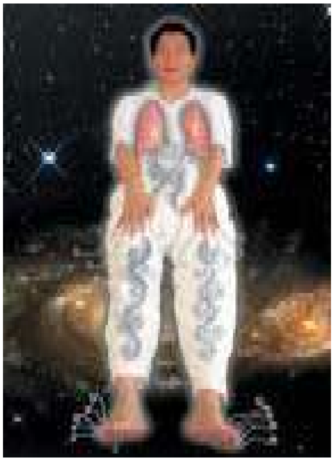

The Triple Warmer refers to the upper, middle and lower body, and to the distinct metabolic transformations that occur within each area. The Upper Warmer is the area above the diaphragm, where the heart and lungs are located. This area tends to become hot and is responsible for respiration and cardiovascular circulation. The Middle Warmer, the area between the diaphragm and the navel, becomes warm and is where the digestive organs are located. The Lower Warmer, the area below the navel, is responsible for reproduction and elimination and is cool in temperature. The sound "HEEEEEEE" balances the temperatures of the three levels by bringing hot energy down to the lower center and cold energy up to the higher centers.

Position: Lie on your back with your arms resting palms up at your sides. Keep your eyes closed. **Smile**. First breathe into the upper part of your lungs to expand the Upper Warmer; then breathe into the middle of the lungs to expand the Middle Warmer; finally, with the last part of your inhalation, expand your lower lungs and fill the Lower Warmer. Breathing in this way creates more space inside for each organ, helping to release and circulate any internal heat or cold.

Sound: Exhale with the sound "HEEEEEEE" sub-vocally, first flattening your chest, then your solar plexus, and finally your lower abdomen. Feel the dark and cloudy color, cold and chill energy exit from the tips of the fingers.

Resting Posture: Once you have fully exhaled, do not to focus on any emotions or purification process at all. Just let go and relax your body and mind completely.

Repeat at least three times, as with the previous sound. Repeat six, nine, twelve or twenty four times to alleviate insomnia and stress.

When you have completed the Six Healing Sounds just rest, **smile** and do nothing.

Regular daily practice of the Six Healing Sounds will help you to keep in touch with the energetic and emotional state of your internal organs. Most importantly do the sounds in the evening before going to sleep. This will help clear out the negative emotions before sleeping, so you can feel that the night’s rest recharges your energy positively. This practice will greatly benefit your personal Cosmic Healing Chi Kung practice. It will help sensitize you to the varieties and differing qualities of Chi. This knowledge will also help you in diagnosing and treating others.

For more details of this practice please see the book "Taoist Ways to Transform Stress Into Vitality", by Mantak Chia.

Appendix

Guide to the Acupuncture Points used in Cosmic Chi Kung

The following is a detailed description of the acupuncture energy points used in the Cosmic Chi Kung practice. It is necessary to know the exact location of the points to do the practice well. However, you do not have to know the traditional Chinese medical functions and energetics of the points to do the practice. These are included in this appendix as an item of interest.

The names of the points are given first by the channel and number, second by the English translation of the traditional Chinese name for the point and finally by the name in Chinese.

We are emphasizing the traditional Chinese medical therapeutic properties of the points given herein. Many of these points also have very different purposes in Taoist Yoga. *For a description of the Taoist Yoga energetics, refer to "Awaken Healing Light of the Tao" by Mantak Chia.*

Pericardium 8
"Lakor's Palace" - Laogong

Ying Spring, Fire Point

Location: On the center of the palm, where the tip of the middle finger touches when making a loose fist.

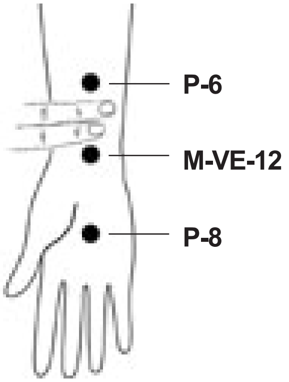

Functions: Cools the heart, drains heat from the heart, cools the blood, calms the spirit, regulates the heart Chi and yang, Ghost Point for treating spirit possession.

Indications: Coma from stroke, heat exhaustion, angina pectoris, mania, hysteria, mental illness, excessive sweating of the palms, mouth problems.

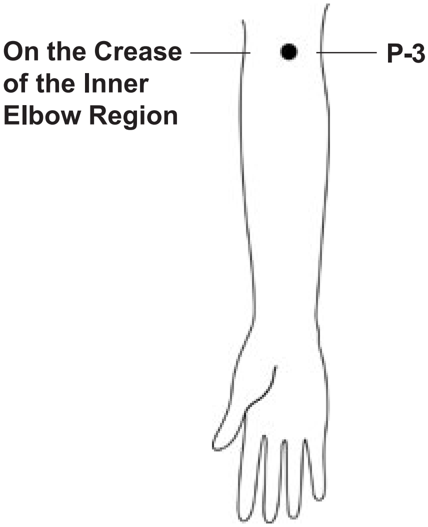

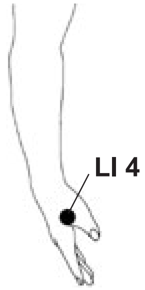

Large Intestine 4

"Adjoining Valleys" - Hegu

Yuan Source Point
Location: On the back side of the hand between the thumb and index finger, in the web approximately at the middle of the metacarpal bone of the index finger.
Functions: Disperses wind, relieves exterior conditions, suppresses pain and clears the channels, clears lung heat, calms the spirit.
Very powerful point for moving the Chi and blood of the whole body; main point for pain; main point for headache; main point for ailments of the head, face and sensory organs; main point for immunity.
Indications: Headaches, common cold, redness with pain and swelling of the eyes, toothache, facial swelling, sore throat, finger cramps, arm pain, fever, abdominal pain, constipation.
Caution: Contra-indicated in pregnancy.

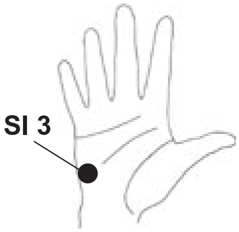

Small Intestine 3

"Back Creek" - Houxi

Shu Stream, Wood Point
Master Point of the Governor Channel
Coupled Point of the Yang Bridge Channel
Location: When a loose fist is made, the point is near the head of the fifth metacarpal bone on the knife edge of the hand at the junction of the red and white skin.
Functions: Relaxes the tendon-muscular channels, opens the Governor Channel, clears the Spirit, drains evil Chi from the heart.
Indications: Seizures, psychosis, hysteria, intercostal neuralgia, headache, stiff neck, red, painful and congested eyes, deafness, spasms of the arm, elbow and fingers; fever, night sweating, whiplash, occipital headaches.

Used for structural/musculo-skeletal problems; opens the Governor Channel for problems with back, neck and head, main point for stiff neck, antispasmodic for muscle spasms; anti-inflammatory for spinal arthritis, intestinal inflammation.

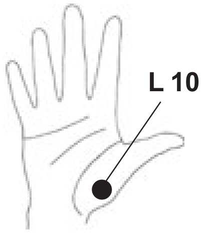

Lung 10

"Fish Border" - Yuji

Ying Spring, Fire Point
Location: In the middle of the ulna eminence below the thumbs at the junction of the red and white skin.
Functions: Cools heat in the lungs, benefits the throat.
Indications: Sore throat, cough, laryngitis, tonsillitis, fever with common cold.

Triple Warmer 5
"Outer Gate" - Wai Guan

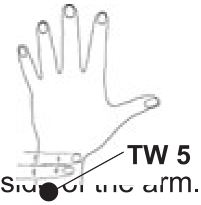

Luo Connecting Point
Master Point of the Yang Regulator Channel
Coupled Point of the Belt Channel
Location: About two fingers width above the wrist crease on the outsi of the arm.
Functions: Regulates the Yang Regulator Channel and Belt Channel, tonifies and consolidates Protective (Wei) Chi, releases exterior hot and cold conditions, helps circulate stagnant Chi in the channels.
Indications: Common cold with alternating chills and fever, high fevers, pneumonia, deafness, migraine headaches, paralysis, stiff neck.

Pericardium 6
"Inner Gate" - Nei Guan

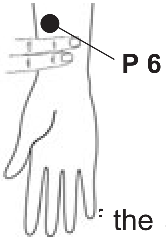

Luo Connecting Point
Master Point of the Yin Regulator Channel
Location: About two fingers width above the wrist crease on the i the arm, between the two prominent tendons.
Functions: Calms the heart and spirit, regulates the Chi, opens and relaxes the chest, regulates and harmonizes the stomach, regulates the liver, relieves pain.
Indications: Rheumatic heart disease, shock, angina pectoris, palpitations, chest pain, asthma, shortness of breath, tightness or fullness in the chest, spasm of the diaphragm, vomiting, stomach ache, abdominal pain, morning sickness, motion sickness, migraine headaches, hysteria, anxiety, irritability, insomnia, seizures, swollen and painful throat, painful menses or nausea with period.

Large Intestine 11
"Crooked Pond" - Qu Chi

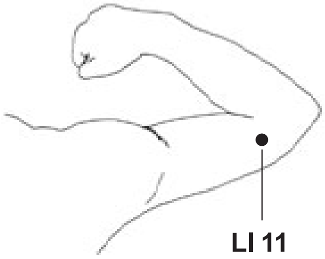

He Sea, Earth Point
Ghost Point
Location: When the elbow is flexed, in the depression at the outer and upper end of the elbow crease.
Functions: Cools heat, clears fire, drains dampness, eliminates wind and exterior conditions, regulates and moistens large intestine.
Indications: Arthritic pain in the arms, paralysis, hypertension, high fever, anemia, allergies, skin problems, Parkinson's Disease. Main point for skin diseases; main point for high fever.

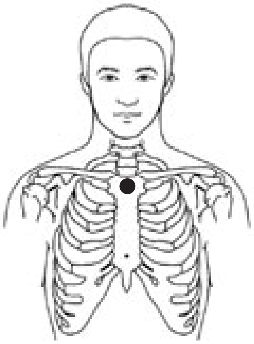

Conception Vessel 22

"Heaven's Chimney" - Tian Tu

Intersecting Point of the Yin Regulator Channel on the Conception Vessel, Window of the Sky Point

Location: At the top of the breastbone (sternum) in the V-shaped indentation (suprasternal notch). Press in and downward at a forty-five degree angle.

Functions: Cools the throat and clears the voice, facilitates and regulates movement of lung Chi, frees the breath and helps restore the proper functioning of the lungs and bronchi, helps open Conception Vessel (Functional Channel).

Indications: Asthma, bronchial asthma, bronchitis, coughing, pharyngitis, goiter, hiccups, nervous vomiting, hoarse voice, spasms of the esophagus, diseases of the vocal cords, sore throat.

Conception Vessel 17

“Central Altar” - Tan Zhong

Front Mu-Alarm Point of the pericardium, Influential Point of Chi of the body
Sea of Chi Point, Master Point of the Middle Tan Tien

Location: On the frontal midline at the level of the fourth intercostal space, between the nipples.

Functions: Regulates the lungs and the Upper Warmer, tonifies Ancestral Chi, opens and relaxes the chest, diffuses lung Chi, regulates and tonifies Chi, transforms phlegm.

Indications: Pulmonary tuberculosis with shortness of breath, bronchitis, asthma, bronchial asthma, chest pain; all breast problems such as mastitis, insufficient lactation, breast abscesses, wheezing, chest pain, labored breathing, palpitations, difficulty swallowing food, anxiety.

Conception Vessel 12

"Middle Cavity" - Zhongwan (Solar Plexus)

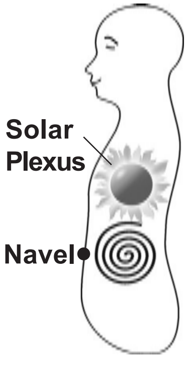

Front Mu-Alarm Point of the stomach, Influential point of all yang organs

Location: Approximately midway between the navel and the top joint of the xiphoid process.

Functions: Regulates the stomach Chi and yin, regulates, strengthens and tonifies the spleen Chi and yang, regulates the Middle Warmer, reduces digestive stagnation, tonifies nutritive Chi, regulates Chi and blood, redirects rebellious Chi downwards, clears stomach fire and heat, calms the fetus, controls the aura.

Indications: Acute or chronic gastritis, stomach and duodenal ulcers, prolapsed stomach, acute intestinal obstruction, stomach ache, vomiting, abdominal distention, diarrhea, constipation, acid regurgitation, indigestion, hypertension, mental diseases.

Mideyebrow Point

"Seal Hall" - Yin Tang

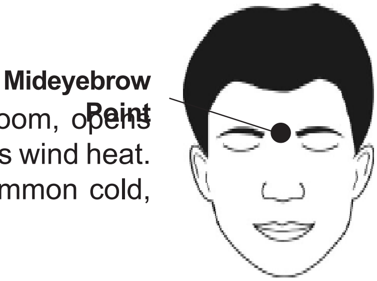

Master Point of the Upper Tan Tien
Location: At the midpoint between the two eyebrows.
Functions: Calms the spirit, activates the Crystal Room, opens the Governor Channel, draws in Cosmic Chi, eliminates wind heat.
Indications: Headache, vertigo, rhinitis, sinusitis, common cold, hypertension, infantile convulsions, sore eyes.

Conception Vessel 6

"Chi Ocean" - Chi Hai

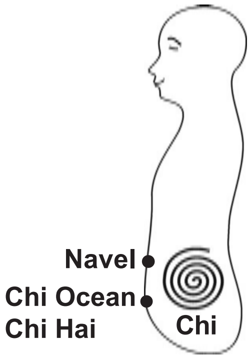

Master Point of the Lower Tan Tien
Location: Approximately three fingers width below the navel.
Functions: Regulates Chi, tonifies Original Chi, strengthens weak kidneys, harmonizes the blood, regulates the Thrusting Route Conception Vessel, reinforces Ching Chi, enriches yin.
Indications: Neurasthenia, abdominal distention, abdominal pain, irregular menstruation, impotence, spermatorrhea, urinary retention, frequent urination, intestinal paralysis, incontinence, constipation, infertility, uterine bleeding, hernia.

Stomach 13

"Chi's Household" - Chi Hu

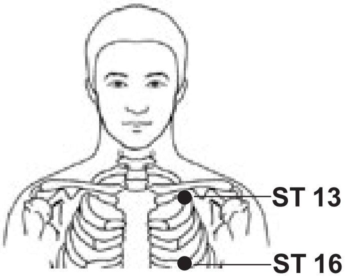

Location: Below the midpoint of the collarbone along the mammilary line (directly above the nipple).
Functions: Clears heat, loosens the chest, relaxes the diaphragm, regulates lung Chi.
Indications: Asthma, bronchitis, chest and back pain, hiccups.

Stomach 16

"Breast's Window" - Ying Chuang

Location: In the space between the third and fourth ribs, one rib directly above the nipple in men, slightly higher in women.
Functions: Stops pain, reduces swelling, clears heat, resolves depression, opens lungs, stops cough, relaxes chest, moves Chi.
Indications: Coughing, asthma, swelling of the breasts, chest and stomach pain, heartburn, shortness of breath, melancholy, diarrhea

Liver 14

"Gate of Hope" - Chi Men

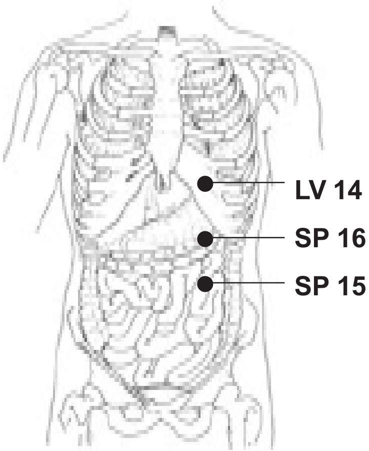

Front Mu-Alarm Point of the liver
Intersecting point of the Yin
Regulator and spleen channels on the liver channel
Location: In the sixth intercostal space, directly below the nipple.
Functions: Spreads liver Chi, transforms and removes congealed blood, strengthens spleen and stomach, expands and relaxes the chest.
Indications: Difficulty in breathing, chest pain, hepatitis, enlarged liver, gall stones, pleurisy, nervous stomach, menopausal disorders, cholera, failure to discharge the placenta after childbirth.

Spleen 16

"Abdomen's Sorrow" - Fu Ai

Location: Just below the bottom edge of the rib cage on the mamillary line.
Functions: Frees bowel Chi, dispels damp and heat, opens the organs and clears Chi stagnation.
Indications: Pain in the navel region, indigestion, dysentary, constipation.

Spleen 15

"Great Horizontal" - Da Heng

Intersecting point of the Yin Regulator Channel on the spleen channel
Location: At the level with the navel along the mamillary line.
Functions: Regulates the spleen, regulates and moistens the intestines, reduces digestive stagnation and transforms damp-heat.
Indications: Abdominal distension, diarrhea, constipation, intestinal paralysis, parasitic worms in the intestines, chronic sadness.

Conception Vessel 8

"Spirit Palace Gate" - Shen Que

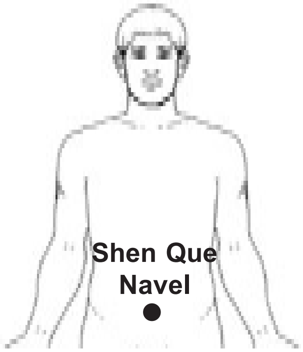

Doorway to the Original Chi,
Location: In the center of the navel
Functions: Tonifies, strengthens and regulates the spleen (Chi and yang), and stomach Chi, regulates the intestines, warms the interior and reduces digestive stagnation, tonifies the kidneys, warms the yang, dries dampness and dispels cold.
Indications: Chronic diarrhea, intestinal tuberculosis, all urinary disorders, shock resulting from intestinal adhesions, heatstroke, rectal or anal prolapse, restless fetus.

Governor Vessel 4

"Door of Life" - Ming Men

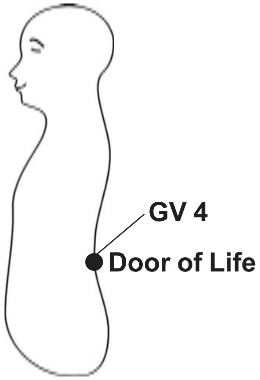

Back Doorway to the Lower Tan Tien
Location: Below the spinous process of the second lumbar vertebrae. Approximately at the level of the navel.
Functions: Nourishes the Original Chi, strengthens and harmonizes the kidneys, tonifies ching and yang, clears channels and invigorates collateral vessels, benefits the lower back and bones, regulates water pathways.
Main point to build life fire; main point for sexual/genital problems.
Indications: Bone disorders, chronic nephritis, enuresis, low sex drive, fatigue, spermatorrhea, impotence, irregular menses, painful menses, no menses, scanty menses, abnormal uterine bleeding, lower back stiffness and pain, kidney pain radiating to the abdomen, hemorrhoids, urinary incontinence, painful urination, diarrhea, sciatica, spinal myelitis.

Governor Vessel 6

"Middle of Spine" - Jizhong

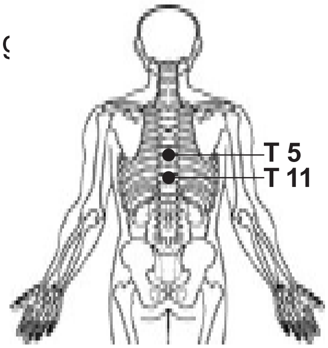

Location: Below the spinous process of the 11th thoracic vertebrae (T-11).
Functions: Stimulates Ching Chi, benefits spleen, stomach and kidneys and liver.
Indications: Hepatitis, seizures, low back pain, paralysis of lower limbs, blood in stools, diabetes, diarrhea.

Governor Vessel 11

"Spirit's Path" - Shen Tao

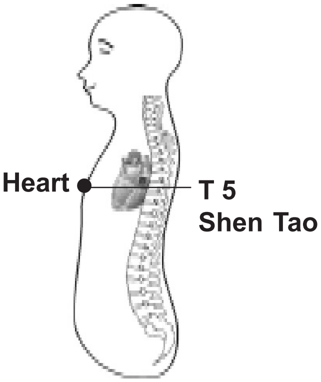

Location: Below the spinous process of T-5 (opposite CV-17, the Heart Point).
Functions: Calms heart and mind, regulates heart Chi and yang, expands and relaxes the chest, benefits heart and lungs, sedates pain, sedates fright and dispels wind.
Indications: Anxiety and palpitations due to fear or fright, neurasthenia, asthma, cough, shortness of breath, chest and hypochondriac pain, insomnia, aphasia due to winds, stroke, forgetfulness, fever.

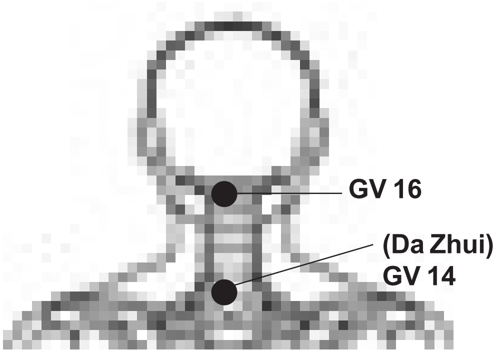

Governor Vessel 14

" Big Vertebra" - Da Zhui

Influential point of yang
Sea of Chi point
Intersection point of all yang channels.
Location: Below the spinous process of the seventh cervical vertebra (C-7), approximately at the level of the shoulders.
Functions: Opens the yang, clears the brain, calms the spirit, tonifies Protective (Wei) Chi, reduces fever, relieves exterior conditions, clears heat.
Indications: Fever, sunstroke, malaria, psychosis, seizures, bronchitis, asthma, pulmonary tuberculosis, emphysema, hepatitis, blood diseases, eczema, hemiplegia, pain in the back of the shoulders, cold-induced diseases, cough, fever and chills.

Governor Vessel 16

“Jade Pillow” or "Wind Palace" - Feng Fu

Sea of Marrow Point, Window of Sky Point, Ghost Point
Intersection point of the Yang Regulator Channel on the Governor Vessel
Location: At the base of the skull, about one inch above the posterior hairline.
Functions: Benefits and clears the brain, calms spirit, opens the sensory orifices, dispels wind, wind-cold and wind-heat, smoothes joint functions.
Indications: Seizures, mania, hemiplegia, loss of speech due to stroke, delirium, suicidal behavior, fear and fright, anxiety, common cold, sensation of heaviness in the head, headache, dizziness, numbness of the limbs, deaf-mutism, blurred vision, sinusitis, stiff neck.

Governor Vessel 20

"One Hundred Meetings" - Bai Hui

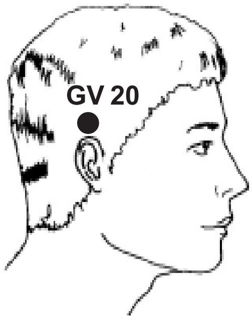

Sea of Marrow Point
Location: At the crown of the head, approximately on the midpoint of the line connecting the top apex of the two ears.
Function: Clears the senses, calms the spirit, extinguishes liver wind, stabilizes ascending yang.
Indications: Headache, dizziness, shock, depression, hypertension, insomnia, seizures, prolapsed anus, prolapsed uterus, mental dullness, hemorrhoids.

Bibliography

Bensky, Dan. Acupuncture: A Comprehensive Text, Seattle, WA: Eastland Press, 1981.
Chia, Mantak. Awaken Healing Light of the Tao, Huntington, NY: Healing Tao Books, 1993.Chi Nei Tsang: Internal Organ Chi Massage. Huntington, Ny: Healing Tao Books, 1990.
Cooper, Groffrey M. The Cell, A Molecular Approach, Washington, DC: Sinauer Associates, 1996.
De Lee, H. Lorenzen. Message from Water: Word of Recommendation, 1984.**DeLaney, Colleen, Leonard, David, and Kitsch, Lancelot.** The Acupuncture Point Book, Santa Cruz, CA: Roast Duck Productions, 1989.
Dong, Paul and Esser, Aristide H. Chi Gong: The Ancient Chinese Way to Health. New York: Paragon House, 1990.
Jou, Tsung Hwa. The Tao of Meditation: Way to Enlightenment. Piscataway: Tai Chi Foundation, 1983.
Lade, Arnie. Acupuncture Points: Images and Functions, Seattle, WA: Eastland Press, 1989.
Lindsay, Mary. The Visual Dictionary of the Human Body, First American Edition, New York, NY: Dorling Kindersley Publications, 1991.
Lu, Kuan Yu. Taoist Yoga: Alchemy and Immortality, New York, NY: Samuel Weiser, 1970.
Maciocia, Giovanni. The Foundations of Chinese Medicine, Edinburgh, Scotland: Churchill Livingstone, 1989.
Matsumoto, Kiiko and Birch, Stephen. Extraordinary Vessels. Brookline: Paradigm Publications, 1986.
McGraw-Hill Companies Human Physiology, 1999.
Morrison, Philip and Phylis and the Office of Charles and Ray Eames. Power of Ten:1982
Nilsson, Lennart. The Body Victorious. New York: NY: Delacorte Press, 1985.
Shanghai College of Traditional Medicine, translated by O'Connor, John and Teeguarden, Iona Marsaa. Acupressure Way of Health: Jin Shin Do, Tokyo, Japan: Japan Publications, 1978.
Tortora, Gerard J. and Grabowski, Sandra R. Introduction to the Human Body, Fifth Edition. New York, NY: Wiley, 2000.
Yang, Jwing-Ming. The Root of Chinese Chi Kung Jamaica Plain, MA: YMAA, 1989.

$19.95

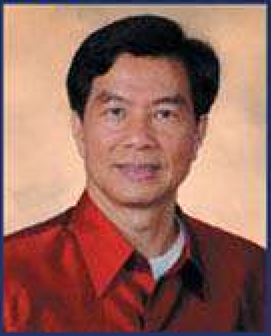

Cosmic Healing I

Cosmic Chi Kung

Mantak Chia

Illustrated by Udon Jandee

Cosmic Healing I

Cosmic Healing I builds on the healing aspects of the Universal Tao foundation practices. In Cosmic Healing I the student will learn how to use the Cosmic Healing Chi Kung in combination with color for guiding the healing learn to process in yourself & other, you'll make use of your Personal Stars & connections with the Big Dipper & North Star so as to have the chi to "sweep & clean out" sick energies.

A New Approach to Healing

In Cosmic Healing I Master Mantak Chia, continue to bring the ancient, time-tested wisdom of the Universal Tao System to the Western world. Cosmic Healing I presents a whole new understanding and approach to healing with detailed explanations of self-healing techniques and ways to transform negative, sick energies from them and others.

Hand Techniques for Detoxifying and Rejuvenating the Vital Organs

Learn to develop your most powerful healing potential through the Cosmic Chi Kung combined with your Mind, Eyes, Heart & Intention. The Cosmic Chi Kung encourages your Compassion to bloom so as to enhance your self- healing & your ability to support other in healing them selves.

0 097487 672561

Universal Tao Publications

Preston B. Nichols, with Peter Moon

The Montauk Project
Experiments In Time

This book is dedicated to the memory of the crew of the U. S. S. Eldridge and to those who gave their life at Montauk.

CONTENTS

Introduction
Guide to the Reader
One-The Philadelphia Experiment
Two-Montauk Discovered
Three-A Visit to Montauk
Four-Duncan Arrives
Five-A Conspiracy Revealed
Six- "Project Moonbeam"
Seven-Wilhelm Reich and the Phoenix Project
Eight-The "Phoenix Project" Absorbs "Project Rainbow"
Nine-The Montauk Project Begins
Ten-The Montauk Chair
Eleven-Creation from the Ether
Twelve-Time Warping
Thirteen-Time Travel
Fourteen-Mission to Mars
Fifteen-Encounter with the Beast
Sixteen-The Nature of Time
Seventeen-The Montauk Base is Sealed
Eighteen-Montauk Today
Nineteen-von Neumann Alive!
Appendix A-A Scientific Analysis of the Radiosonde
Appendix B-Wilhelm Reich
Appendix C-Mind Control and the Persian Gulf War
Appendix D-Nikola Tesla
Appendix E-History of the Philadelphia Experiment and its Reconcilation with the Montauk Project
Appendix F-Quantum Levels of Existence
Appendix G-Glossary

INTRODUCTION

At the eastern most end of Long Island sits Montauk Point, known to most New Yorkers for its scenic beauty and landmark lighthouse. To the immediate west of the lighthouse, there is a mysterious and derelict Air Force base on the grounds of old Fort Hero. Although it was officially decommissioned and abandoned by the U.S. Air Force in 1969, it was subsequently reopened and continued to operate without the sanction of the U. S. Government.

The entire financing for the base is also a mystery. No funding can be traced to the military or government. Officials of the U. S. Government have probed for answers without success.

The secrecy of the operation has prompted legends to thrive across Long Island. However, it is unlikely that any of the local people of Montauk, or those who spread the tales, know the full story of what actually went on there.

A circle of insiders believe the Montauk Project was a development and culmination of the phenomena encountered aboard the U.S.S. Eldridge in 1943.

Popularly known as the "Philadelphia Experiment", the ship actually disappeared while the Navy conducted radar invisibility experiments.

According to these accounts, over three decades of secret research and applied technology ensued. Experiments were conducted that included that included electronic mind surveillance and the control of distinct populations.

The climax of this work was reached at Montauk Point in 1983. It was at that time that the Montauk Project effectively ripped open a hole in space-time to 1943.

Perhaps the person best qualified to tell the real story is Preston Nichols, an electrical engineer and inventor who has studied the Montauk Project for the better part of a decade. His interest in the project was spurred in part by unusual circumstances in his own life. He was also able to legally acquire much of the equipment that was used for the project. His continued investigation ultimately revealed his own role as the technical director of the project. Despite brainwashing and threats to silence him, he has survived and has decided it is in the best interest of all to tell his story.

GUIDE TO THE READER

Because of the subject matter of this book is controversial, we would like to offer some guidelines.

This book is an exercise in consciousness. It is an invitation to view time in a new manner and expand your awareness of the universe. Time rules our fate and ushers in our death. Although we are regulated by its laws, there is much that we do not know about time and how it relates to our consciousness. Hopefully, at the very least, this information will broaden your horizons.

Some of the data you will read in this book can be considered as "soft facts".

Soft facts are not untrue, they are just not backed up by irrefutable documentation.

A "hard fact" would be documentation or hard physical evidence that could stand up to scrutiny.

By the nature of the subject matter and security considerations, hard facts about the Montauk Project have been very difficult to obtain. There is also an area between soft and hard which can be termed "gray facts". These would be very plausible but not as easily provable as a hard fact.

Any serious investigation will show that a Montauk Project did, in fact, exist. One can also find people who have been experimented on in some fashion or another.

This book is not an attempt to prove anything. The purpose is to get a story told that is of essential interest to scientific researchers, metaphysicians and citizens of the planet Earth. It is the story of one particular individual and his circle of contacts. It is hoped that more individuals will come out of the closet and that researchers will come forth with more investigations and documentation.

This work is being presented as non-fiction as it contains no falsehoods to the best knowledge of the authors. However, it can also be read as pure science fiction if that is more suitable to the reader.

A short glossary has been provided in the back to assist with ordinary electronic terms and those of a more esoteric nature. Scientists who read this book should understand that the definitions are designed to assist the general reader's understanding. They are not purported to be the latest technical jargon.

Likewise, the general reader should understand that the diagrams in this book are included for the benefit of technical people. If one is interested, they can get a further understanding of those terms and symbols by studying the Radio Amateur's Handbook or a text of similar nature.

ONE-THE PHILADELPHIA EXPERIMENT

The origin of the Montauk Project dates back to 1943 when radar invisibility was being researched aboard the USS Eldridge. As the Eldridge was stationed at the Philadelphia Navy Yard, the events concerning the ship have commonly been referred to as the "Philadelphia Experiment". Having been the subject of differnet books and a movie, only a quick synopsis will be given here.*

The Philadelphia Experiment was known as the Rainbow Project to those who manned and operated it. It was designed as a top secret project that would help end World War II. The forerunner of today's stealth technology, the Rainbow Project was experimenting with a technique to make a ship invisible to enemy radar. This was done by creating an "electromagnetic bottle" which actually diverted radar waves around the ship. An "electromagnetic bottle" changes the entire electromagnetic field of a specific area - in this case, the field encompassing the USS Eldridge.

While the objective was to simply make the ship undetectable by radar, it had a totally unexpected and drastic side effect. It made the ship invisible to the naked eye and removed it from the space-time continuum. The ship suddenly reappeared in Norfolk, Virginia, hundreds of miles away.

The project was a success from a material standpoint, but it was a drastic catastrophe to the people involved. While the USS Eldridge "moved" from the Philadelphia Naval Yard to Norfolk and back again, the crew found themselves in complete disorientation. They had left the physical universe and had no familiar surrounding to relate to. Upon their return to the Philadelphia Navy Yard, some were planted into the bulkheads of the ship itself. Those who survived were in a mental state of disorientation and absolute horror.

The crew were subsequently discharged as "mentally unfit" after having spent considerable time in rehabilitation. The status of "mentally unfit" made it very convenient for their stories to be discredited.

This put the Rainbow Project at a standstill.

Although a major breakthrough had occurred, there was no certainty that human beings could survive further experimentation. It was too risky. Dr. John von Neumann, who headed the project, was now summoned to work on the Manhattan Project. This concerned the making of the atom bomb, which became the weapon of choice for ending World War II.

Although it is not well known, vast research that began with the Rainbow Project was resumed in the late 1940's. It continued on, culminating with a hole being ripped through space-time at Montauk in 1983. The goal of this book is to give you a general understanding of the research and events subsequent to the Philadelphia Experiment and up to 1983 at Montauk. I will begin by telling you how I, Preston Nichols, stumble across it.

* Further information about the Philadelphia Experiment can be found in Appendix E.

TWO-MONTAUK DISCOVERED

In 1971, I began working for BJM*, a well known defense contractor on Long Island. Through the years, I got a degree in electrical engineering and became a specialist in electromagnetic phenomena. I was not then aware of the Philadelphia Experiment or its accompanying phenomena.

Although I was not extraordinarily interested in the paranormal at that time, I had obtained a grant to study mental telepathy and to determine whether or not it existed. I sought to disprove it, but I was surprised to find out that it did, in fact, exist.

I began my research and found out that telepathic communication operated on principles that are strikingly similar to that of radio waves. I had discovered a wave that could be termed a "telepathic wave". In some respects, it behaved like a radio wave. I set out to get the characteristics of this "telepathic wave". I studied their wave lengths and other pertinent facts. I determined that while a telepathic wave behaves like a radio wave, it isn't exactly a radio wave. Although it propagates in a similar fashion to that of electromagnetic waves and possesses like properties, not all of these fit into normal wave functions.

I found all of this very exciting. I had discovered a whole brand new electromagnetic function that was not in any of the text books I'd ever seen.

I wanted to learn as much as I could and studied all the activities that might use this type of function. My interest into metaphysics had been launched.

I continued to research in my spare time and collaborated with different psychics to test and monitor their various responses. In 1974, I noticed a peculiar phenomena that was common to all of the psychics that I worked with. Every day, at the same hour, their minds would be jammed. They couldn't think effectively.

Suspecting that the interference was caused by an electronic signal, I used my radio equipment and correlated what came on over the air waves at the times the psychics were non-functional. Whenever a 410-420 MHz (Megahertz) cycle appeared on the air, they were jammed. When the 410-420 MHz cycle was off, the psychics would open back up after about twenty minutes. It was obvious that this signal was greatly impeding the ability of my psychics.

I decided to trace the signal. Placing a modified TV atenna on the roof of my car, I grabbed a VHF receiver and set out looking for the source of it. I tracked it right to Montauk Point. It was coming directly from a red and white radar atenna on the Air Force base.

At first, I thought that this signal might have been generated accidentally.

I checked around and found out that the base was still active. Unfortunately, security was tight and the guards wouldn't give any useful information. They said that the radar was for a project run by the FAA. I couldn't press the point beyond that. In fact, their statement didn't make a lot of sense. This was a World War II radar defense system known as "Sage Radar". It was totally antiquated, and there is not any known reason why the FAA would need such a system I didn't believe them but couldn't help being intrigued. Unfortunately, I had hit a dead end.

I continued my psychic research, but didn't get anywhere on the investigation of the Montauk atenna until 1984, when a friend of mine called. He told me the place was now abandoned, and that I should go out there and check it out. I did.

It was indeed abandoned, with debris strewn everywhere. I saw a fire extinguisher left amidst many scattered papers. The gate was opened as were the windows and doors of the buildings. This is not the way the military normally leaves a base.

I strolled around. The first things that caught my eye was the high voltage equipment. I was very interested as it was a radio engineer's delight. I am a collector of ham gear and radio equipment, and I wanted to buy it. I figured it would be available cheap if I made the proper arrangements though the Surplus Disposal Agency in Michigan.

After examing all the equipment, I contacted the disposal agency and spoke to a friendly lady. I told her what I wanted, and she told me she would see what could be done. It appeared to be abandoned material and looked like a scrap contract. If this was so, I'd be able to take what I wanted. Unfortunately, I didn't hear from her so I called her back three weeks later. She informed me that there had not been any success with tracing the equipment. They couldn't find out who owned it. Neither the military or the GSA (General Services Administration) claimed to know anything about it. Fortunately, the Surplus Disposal Agency said they would continue to track the matter further. After another week or two went by, I called her back. She said she'd turn me over to a John Smith (fictitious name), located at a military overseas terminal in Bayonne, New Jersey.

"Talk to him and he'll set something up," she said.

"We like to keep our customers satisfied."

I met John Smith. He didn't want to discuss anything on the phone. He said that no one officially admitted to owning that equipment. As far as they were concerned, the equipment was abandoned and I could go in and take whatever I wanted. He gave me a piece of paper which appeared be official and said to show it to anyone who might question my presence in the area. It was not an official document nor was it registered with anybody, but he assured me that it would keep the police off my back. He also referred me to the caretaker of the Montauk Air Force Base who would show me around.

*BJM is a fictitious name for the company I worked for.

THREE-A VISIT FROM MONTAUK

I was out at the base within the week. There I met the caretaker, Mr. Anderson.

He was very helpful. He told me to be careful and showed me where things were so that I wouldn't fall through the floor and that type of thing. He said I was welcome to take anything I could this trip, but if he ever saw me out there again, he'd have to kick me out. His job, afer all, was to keep people off the base.

He realized that the permission I had was semi-official at best. He was also kind enough to tell me that he went out for a dring every evening at 7P.M..

I had taken a trip to Montauk with a fellow named Brian. Brian was a psychic who had helped me with my research. As we foraged around the base, we went in two different directions. I went into a building and saw a man who appeared to be homeless. He told me that he had been living in the building ever since the base was abandoned. He also said that there had been a big experiment a year earlier and that everything had gone crazy. Apparently, he'd never gotten over it himself.

In fact, the man recognized me, but I had no idea who he was or what he was talking about. I did listen to his story. He said he had been a technician at the base and that he'd been AWOL. He had deserted the project just before the base had been abandoned. He spoke about a big beast appearing and frightening everyone away. He told me a lot about the technical details of the machinery and how things worked. He also said something that was very strange. He told me that he remembered me well. In fact, I had been his boss on the project. of course, I thought it was pure nonsense.

I didn't know then that there was any truth to his story. This was just the beginning of my discovery that the Montauk Project was real.

I left the man and found Brian. He was conplaining that things weren't right and that he was feeling some very funny vibrations. I decided to ask him for a psychic reading right there. His reading was strangely similar to what the homeless man had just told me. He spoke of irregular weather patterns, mind control and a vicious beast. He mentioned animals being affected, crashing through windows.

Mind control was a main focus of Brian's reading.

The reading was interesting, but we were there to cart out the equipment. Much of it was heavy and we weren't allowed to bring a vehicle right onto the base.

We had to back pack it. I was thus able to acquire much of the equipment left behind from the Montauk Project.

A few weeks later, I was surprised by a visitor who barged into my lab. He came straight to the lab, which was in back of the house. He didn't ring the door bell or anything. He claimed to know me and said that I had been his boss. He went on to explain many of the technical details of the Montauk Project. His story corroborated what psychics and the homeless man had told me. I didn't recognize him but listened to all he had to say.

I was sure that something had gone on at the Montauk base, but I didn't know what. My personal involvement was evident, but I still didn't consider it very seriously. I was, however, puzzled by different people recognizing me. I had to make it my business to investigate Montauk. So, I went out and camped on the beach for a week or so. I went to bars and asked the locals for stories about the base. I talked to people on the beach, on the street, wherever I could find them. I asked all about the stranged activities that were purported to have occurred.

Six different people said that it had snowed in the middle of August. There were listings of hurricane force winds that came out of nowhere. Thunder storms, lightning and hail were also reported under unusual circumstances. They would appear when previously there had been no meterorologial evidence to expect such.

There were other unusual stories besides the weather. These included stories of animals coming into the town en masse and sometimes crashing through the windows. By this time, I had taken different psychics out to the base. The stories confirmed what psychics had been able to determine through their own sensitivity.

I finally got the idea to speak to the Chief of Police who also informed me of strange happenings. For example, crimes would be committed in a two hour period. Then, all of a sudden, nothing. Keep in mind that Montauk is a very small town. After the quiet, another two hour period of crimes would occur.

Teens were also reported to suddenly group en masse for two hours, then mysteriously separate and go their own ways. The Chief couldn't account for it, but his statements lined up perfectly with what the psychics had indicated about mind control experiments.

I had collected some really bizarre information, but I didn't have many answers.

I was, however becoming very suspicious. I had often travelled to Ham-fests, (where Ham radio equipment is bought and sold) and there more people would recognize me. I had no idea who they were, but I would talk to them and ask them about Montauk. As I did, more information came, but everything was still a big puzzle.

FOUR- DUNCAN ARRIVES

In November of '84, another man appeared at my lab door. His name was Duncan Cameron. He had a piece of audio equipment, and he wanted to know if I could help him with it. He quickly became absorbed in the group of psychics I had working with me at the time. This endeavor was a continuation of my original line of research. Duncan showed a keen aptitude for such work and was extremely enthusiastic.

I thought he was too good to be true and became suspicious of him. My assistant, Brian, felt the same. He didn't like duncan's sudden involvement and decided to go his own way.

At one point, I surprised Duncan by telling him that I would be taking him some place to see if he recognized it. I drove him to the Montauk Air Force Base.

He not only recognized it, he told me what the purpose was for each of the various buildings. He knew exactly where the bulletin board in the mess hall was and many other suchminute details. Obviously, he had been there before. He knew the place like the back of his hand. He provided new information about the nature of the base and what his own function had been, Duncan's input dovetailed very nicely with the previous data I had collected.

When Duncan entered the transmitter building, he suddenly went into a trance and began spitting out information. This was curious, but I had to shake him repeatedly to break him out of it. When I brought him back to the lab, I applied techniques that I'd learned to help Duncan unblock his memories. Layers of programming were now coming out of Duncan. A lot of information concerned the Montauk Project.

Many different things were revealed, until finally a shocking program came straight to the awareness of Duncan's conscious mind. He blurted out that he had been programmed to come to my place, befriend me and, then, kill me and blow up my entire lab. All my work would be totally destroyed. Duncan appeared to be more outraged at all this than I was. He swore that he would no longer help those who had programmed him, and he has worked with me ever since.

Further work with Duncan revealed even more bizarre information. He had been involved in the Philadelphia Experiment! He said that he and his brother Edward had served aboard the USS Eldridge as members of the crew*.

A lot of things surfaced as a result of my work with Duncan. I started to remember things about the Montauk Project and was now certain I'd been involved.

I just didn't know how or why. The puzzle was slowly clearing up. I found Duncan to be an extremely operational psychic and through him I was able to confirm new information.

* An account of Duncan's role in the Philadelphia Experiment is in the book "The Philadelphia Experiment & Other UFO Conspiracies" by Brad Steiger with Al Bielek and Sherry Hanson Steiger.

FIVE- A CONSPIRACY REVEALED

I visited Montauk many more times, often with different people who had been involved. A small group of us began to realize that we had stumbled across one of the highest security projects the country had ever known. We figured that we had better do something fast with this new found knowledge. If we didn't, we might end up dead.

As a group, we decided action had to be taken. We weren't sure exactly what to do, so we sat around and discussed it. What was the best thing to do? Publish it? Immediately? We talked about it extensively. In July of 1986, we decided that I should go to the United States Psychotronics Association (USPA) in Chicago and talk about it. I did, and it created an uproar. Word got around fast to those who didn't want the Montauk story to be revealed. Suddenly, here I was, giving an unannounced lecture. The information got out to hundreds of people, swept under the rug without creating a public furor. To this day, I still appreciate the open forum and free speech that the USPA provided me.

Now, we decided to feed the information to the federal government. One of associates knew the nephew of a senior senator from the Southwest. The nephew, who we will call Lenny, worked for the Senator. We gave the information to Lenny, who passed it to his uncle. This information included pictures of the orders given to the different military personnel, which we had found strewn about the base.

The Senator did a personal investigation and verified that miitary technicians had in fact been assigned to the base. The Senator also discovered that the base was decommissioned, derelict and mothballed since 1969. Having served his country as an Air Force general, he was particularly interested to know why Air Force personnel were working on a derelict base. And, where did the money come from to open up the base and run it?

After they did their own investigation and saw the pictures and documents we supplied them, there was no question that the base had been active. They verified that Fort Hero (which is the name of the original World War I base that surrounds the entire area of the Air Force base) and Montauk were indeed derelict and simply listed as property held by the General Services Administrations since 1970.

The Senator got very involved and travelled to Long Island to find out what he could about Montauk Air Force Base. He was not greeted with enthusiastic cooperation despite having very impressive personal credentials. People reported seeing him looking through the fences and trying to find out what was going on. He visited me and told me to keep quiet about it as speaking out any further could jeopardize his investigation. That is why I have kept this strory quiet until now.

When the Senator completed his investigation, he couldn't find any trace of government funding, no appropriations, no oversight committees and no payments.

He eventually retired due to advancing age, but I have since been informed by Lenny that he sees no problem with my story being published. He also said that the Senator is still in the picture and that the investigation had been reopened.

SIX- "PROJECT MOONBEAM"

While the Senator was seaching for paper trails that might reveal the secrets of Montauk, I knew that they would not solve my personal mysteries one bit. I had been recognized by people I didn't know, and it was obvious that I had severe memory blocks. What made things so hard to reconcile was that I had a full set of "normal" memories which told me where I had been.

My memory improved while working with Duncan, and I eventually realized that I must have been existing on two separate time tracks. As bizarre as it may sound, it was the only sensible explanation under the circumstances.

As my memory was still largely blocked, there were three avenues of approach to the problem. First, I could simply try to remember the other time track, through regression or hypnosis. This proved to be very difficult for me and was virtually of no use. Secondly, I could look for clues and hints (in our normal time track) that the other time track did, in fact, exist. Thirdly, I could try to find the answers through technology. This would include theories of how the other time track was created and how I ended upon it.

The third approach was the easiest. I am told that many people might find this very confusing, but I was familiar with the theories of the Philadelphia Experiment and was not intimidated by physics or electromagnetics. I found it plausible.

The second approach also proved extremely helpful, but clues were hard to come by.

It was now 1989. I started to roam around the plant at BJM, where I was still working. I would talk to different people and dredge up what information I could without trying to appear suspicious. I would also walk around and just sense my own personal gut reaction to the different places in the plant.

I became particularly irritated when I would come to a certain room. My innards would just churn. I sensed very strongly that there was something in that room that was disturbing me. I had to investigate it. I rang the doorbell and was told that I couldn't come in. It was a high security area. Reportedly, only ten people at the plant had the proper clearance to be in that room.

I found that virtually no one knew anything about it. Finally, I did find two people who'd been in there, but they said they couldn't tell me anything.

One of them must have turned me in, because the security personnel visited me shortly thereafter. It was time to lay low for a while.

About a year after my futile investigation, the room was totally cleared out.

The doors were open and anyone could walk right in. It was obvious that there had been all sorts of equipment. Dirt markings revealed that four round things had stood on the floor. I presumed they were coil structures. It was clear that there had been a console. There was also a huge power line that still ran into the room. The entire place gave me the creeps, but I was driven to find all I could.

I discovered an elevator in the back of the room. I got in and found only two buttons: Main Floor and Sub Floor. There was also a numbered key pad. I pushed the button for Sub Floor and tried to go down, but the elevator would only go so far. I heard a voice that told me to punch in the proper coded numbers on the key pad. I didn't have the code and a beeping noise went off for about thirty seconds. Security was alerted. I had hit another dead end.

I wasn't scoring any points with security, and it was time to lay low once again. I began to think of how I could show that something very unusual was going on.

I also recalled earlier strange experiences that had occurred while working at BJM. There was a period when, all of a sudden, a band-aid would appear on my hand. It hadn't been there fifteen minutes ago! I couldn't remember putting it on. This hapened more than a few times.

One day, I had been sitting at my desk and my hand suddenly started to ache.

The back of the hand was sore, and there was a band aid on it. I absolutely knew that I had not put that band aid on nor had I had it put on. I became very suspicious. I got up and went down to the nurse.

I said to her, "This may sound wacky, but was I in here for a band aid?"

"No, you weren't in here," she told me.

I asked her where I'd gotten it and she said, "You must've gotten it from one of the first aid kits. Don't you remember?"

"I'm just trying to figure it out," I said, and I walked out. I thought in my mind, "I'm not going to get a band-aid at BJM except from the company nurse."

I wanted a record, so I made a conviction that I would never use a first aid kit.

I eventually remembered the reason I had sustained so many injuries to my hands. In my alternate reality, I frequently had to move different equipment.

I was just about the only one who could move it as most people would go wacky when they'd get near it. For some reason, it didn't seem to bother me. But it was heavy and hard to maneuver. With no one to assist me, bruised hands and band aids became a regular occurrence.

I kept to my conviction not to use any band aids from first aid kits. I continued to check with the nurse when they appeared, and the records indicated I'd never been to her.

As this was an irregularity, she must have reported it to security. They visited me and said, "Why are you asking about band aids, Mr. Nichols?" I knew better than to pursue that anymore.

Recalling these experiences with the band aids helped spur my memory back to 1978. I remembered sitting at my work bench one day. All of a sudden, I smelled the scent of burning transformers. It was pungent, like the smell of burning tar. It came and disappeared very fast. This happened at 9:00 o'clock in the morning. The rest of the day continued as normal until 4:00 o'clock in the afternoon when the whole plant began to smell like putrid smoke from burning transformers.

I thought to myself, "That's the same smell I smelled at 9:00 o'clock this morning." But now it occurred to me that the event probably hadn't happened at the time I had thought. You can't burn up a transformer and have the smell disappear as fast as it had that morning.

Many more events of this nature had occurred. Each puzzle tended to confuse the general issue. Streams of unfamiliar people continued to recognize me. I began to get executive mail that would normally be for the vice president of a company.

For instance, I would be asked to come to a conference concerning patents. I didn't know what they were talking about. I was also called to meetings with a certain executive. He always appeared very agitated whenever we spoke.

Most of the inquiries I received from these people were about the Moonbeam Project. I didn't know what it was. But one day, I had an intuitive urge. The basement of the BJM building in Melville had a very high security area. Consciously, I had no clearance to be in that area, but I walked in anyway. Normally, when you walk from one security area to another, you must hand the guard your badge and he gives you another badge (with a different designation). This permits you to walk in the secure area. I simply went in and gave him my badge from the lesser security area, and what do you know? He gave me a badge with my name on it! I'd had a hunch and it worked.

I walked around and let the churning in my gut determine what direction I should go in. I ended up in a posh mahogany paneled office. There was a large desk with a name plate on it that read, "Preston B. Nichols, Assist Project Director". This was the first tangible physical proof I had that something out of the ordinary was definitely occurring. I sat at the desk and looked through all the papers.

It was impossible to take the papers out of the place as I knew I would be searched very thoroughly on my way out of this high security area. So, I committed everything I saw to memory, to the best of my ability. I had an entire second career here that I knew almost nothing about! I can't even talk about most of it.

It is top secret. I'm bound not to mention it for thirty years because of an agreement I signed when I went to work for BJM. However, I didn't sign a single thing regarding the activities of the Montauk Project.

Sifting through the material, I spent about six hours in my newly discovered office. Then, I decided I'd better get back to my regular job before the day was through. I handed back my badge and walked out. A couple of days went by before I decided it was time to go back and check things out again. Once more, I handed the guard my badge, but this time he didn't give me anything back. He said, "By the way, Mr. Roberts (fictitious name) wants to see you."

A man, Mr. Roberts, came out of an office that had "Project Director" written on it. He looked at me and said, "What do you want to come in here for, sir?"

"To get to my other desk," I replied.

He said, "You don't have any other desk here."

I pointed to the office where my desk had been. But as I entered the room with the Project Director, I found it to be gone. In the couple of days since I'd been there, they had removed every trace of myself from the room.

Somebody must have realized that I had visited my office when I wasn't supposed to. I had entered in an ordinary state of mind which was not to their liking. They apparently had not turned on the program (switching me to an alternate reality) for that particular day and must have been wondering why I'd shown up. They must have concluded that the process was leaking and that I was some how able to remember my alternate existence. As a result, they stopped everything. I was pulled aside through security channels and was told that if I breathed a word of what I'd seen, I'd be locked up in jail and the key thrown away.

I tried to think of other strange incidents that had occurred. I'd kept a suspicious eye and had been experiencing two separate existences. How the hell had I been at Montauk and working at BJM, apparently during the same time period? I had already arrived at the conclusion that I must have been working two jobs simultaneously because there was a period of time when I'd come home and be totally exhausted.

At this point, all of what you've read was one huge confused mess in my mind.

I knew that I'd been working on two separate time lines or maybe more. In fact, I'd discovered quite a bit, but it was more confused than clear. I was, however, able to make a major breakthrough in 1990. I had begun constructing a Delta T* antenna on the roof of my laboratory. One day, I was sitting on the roof and soldering all the loops together into the relay boxes (which relay the signals from the antenna downstairs to the lab). Apparently, as I sat there and held the wires together to solder them, the time functions were causing my mind to shift. The more soldering I did, the more I became aware. Then, one day - bang! - the whole memory line blew open for me. All I could figure was that the Delta T antenna was storing up time flux waves as I was connecting it together. It just kept pushing my mind a little bit with regard to the time reference. The antenna was stressing time (bending it) and enough bend was created so that I was subconsciously in two time lines. This was my memory breakthrough.

Whatever the explanation, I was very pleased to have regained so much of my memory. I also believe my theory about the Delta T antenna is correct because the more time I spent working on the antenna, the more memories came back. By early June 1990, all my key memories had come back.

In July, I was laid off. Subsequent to my firing, all of my close connections were removed as well. After having worked at BJM for the better part of two decades, I no longer had any links or friends to the company. My information sources had been effictively severed.

You now have a general idea of the circumstances whereby I regained my memory.

The next part of the book will contain the history of the Montauk Project that includes a general description of the technology involved. It is based upon my own memories and the information that has been shared with me by my various colleagues involved with the Montauk Project.

* A Delta T antenna is an octahedronal antenna structure that can shift time zones. It is designed to bend time. Delta T=Delta Time. Delta is used in science to show change and "Delta T" would refer to a change in time. More about the nature of this antenna will be covered later in the book.

SEVEN- WILHELM REICH AND THE PHOENIX PROJECT

The U.S. Government began a weather control project in the late 1940's under the codename "Phoenix." The information and technology for this came from Dr. Wilhelm Reich, an Austrian scientist who had studied with Freud and Carl Jung.

Reich was an extremely brilliant man but highly controversial. Although he experimented extensively and wrote many volumes, few of his critics have taken an honest look at all of his research because much of it is not available. Part of this can be attributed to the Food and Drug Administration who supervised a massive book burning of all his available materials and also destroyed much of his laboratory equipment.

Reich was known in part for his discovery of "orgone" energy, which is orgasmic or life energy. His experiments revealed orgone energy to be distinctly different from ordinary electromagnetic energy. He was able to prove the existence of this energy in the laboratory. His findings were written up in various psychiatric and medical journals of the period. The discovery of a type of energy called "orgone" was not so controversial. It became very controlversial with the powers at be when he reported curing cancer with his theories. He also associated "orgone" energy with "cosmic energy" and the Newtonian concept of "the ether". None of these views won him support from conventional scientists of the 1940's.

At the turn of the century, scientists had embraced the Newtonian "ether".

This is referred to a hypothetical invisible substance as a medium for light and radiant energy. Einstein, who embraced the theory in his early years, eventually determined that there could not be a calm ether sea through which matter moves.

Not all physicists bought Einstein's argument, but Reich didn't disagree. He pointed out that Einstein disproved the concept of a static ether. Reich considered the ether to be wave-like in nature and not static at all.

Conventional scientists have since recognized the existence of phenomena that are a cross between particles and waves. They are sometimes referred to as "wavicles".

Common research has also shown that vacuum space contains complex properties that are dynamic in nature.

Although it is not my cause to take up the case of Reich, his concept of the ether has proven itself functional in my research. It does not matter whether we are actually referring to "wavicles" or even more esoteric phenomena when we talk about the ether. Is is the word that Reich used, and it is easier for me to use in describing this for the general public. The reader is invited to read up on Reich as his work is vast and encompasses much more than can be covered in the scope of this book.

For instance, he found practical uses for his theories such as modifying the weather. He found that violent storms accumulate "dead orgone", which he termed "DOR." Dead orgone refers to the accumulation of "dead energy" or energy that is on a descending spiral. Orgone and DOR were found to be present not only in biological organisms but in empty regions of the environment as well. An active and enthusiastic go-getter would be considered to have plenty of orgone energy, whereas a complaining hypochondriac who wanted to die would have DOR energy.

For example, he found that the more DOR in the storm system, the more violent the storm. He experimented with many forms of DOR busting, and came up with a simple electromagnetic method to reduce the violence of storms. In the late 1940's, Reich contacted the government and told them he had developed technology that could take the violence out of storms. Despite what disinformation you may hear, the government already knew what Reich could do and considered him a brilliant man. They asked for his prototypes and he was happy to oblige since he wasn't interested in the mechanical development, just the research.

At this point, the government's technology team merged Reich's discoveries with their own weather monitors and produced what is known today as the "radiosonde."

The government's contribution to the radiosonde dates back to the "airborne metrograph"* of the 1920's. This was a mechanical device that recorded temperature, bumidity and pressure. It was sent up in a parachute balloon and recorded information on a paper tape. The balloon was designed to burst so that the parachute would bring the metrograph back to Earth. The public were encouraged to retrieve them for a $5 reward, which was considerably more money in those days. This was how the government obtained data on the weather.

As these devices were returned via the mail, the time that elapsed before the recorded information could be read was much too long.

In the late 1930's, a new device was designed that was called a "radio metrograph."

This was similar to the airborne metrograph except that it contained electrical sensors. These sensors were connected to a transmitter that would transmit to a receiver on the ground.

The radio metrograph was the state-of-the-art weather device when Wilhelm Reich contacted the government in the late 1940's. He gave them a little balsa wood package that could be sent up in a balloon. According to witnesses, approaching thunderstorms actually split up and went arount the test sit on Long Island.

The government combined the technology of the radio metrographs with Reich's DOR busting device and called it the "radiosonde." It was developed until consistent effects on the weather could be reproduced.

By the 1950's, radiosondes were being sent into the air en masse at a rate of about 200 per day.

Since these radiosondes were sent up in balloons, they would not come down hard enough to self destruct upon impact. The public would find them, and it would be impossible to keep the actual units secret enough without arousing suspicion. They publicized the apparent purpose of recording weather data, which uniformed examination would back up. The real purpose is not that obvious.

If someone tuned into one of these packages, the signal would not appear unusual when normal radio equipment was ued. So far so good!

They showed the public a data receiving station; set up to receive the inaccurate and unusuable data. A small production run of this receiving equipment was produced.

There were literally hundreds of these radiosondes in the air every day. With the radio range being limited to 100 miles, there should have been a "pile" of receivers known as radiosonde receptors and they should have been very common.

As I am a surplus radio collector "nut", it is quite strange that I have never seen a radiosonde receptor or the equipment that should accompany one. It is very unusual to have a data transmitter (in this case, the radiosonde) with no receiver to pick it up. This indicates that the Government didn't use the receivers!

My next clue was to look at the specification sheet for the radiosonde tube which emphatically states that the life expectancy is only a few hours. Despite this, I have had a tube on the air for over 2,000 hours, and at this time have built over twenty such units with only on efailure. This is a good industrial failure rate but is a major red flag. My only explanation is that if some local amateur radio operator finds or buys a radiosonde on the surplus market, he will read the data, get misled and not bother building a circuit that will run for a "only a few hours." He will use another tube.

It appears that the Government does not want the public to use these tubes and find something unusual and thus blow their secret. This is why misinformation in the spec sheet preserves the secret. In fact, they are not telling a lie because the battery pack was designed so that the tube would burn out after three hours or so. This is caused by back bombardment of the cathode, which would cool slowly and then destruct.

By the time these radiosondes hit the ground, they were dead. This way the public, who were encouraged to return them, wouldn't be able to pick up live units. If there was no secrecy involved here, why would the government design a battery to burn out a costly tube that would have to be replaced after a very short usage? More disinformation was accomplished by packing the sensors in sealed vials, which implies that upon exposure to the air, the sensors are short lived. Because of these precautions, the secret was maintained for over forty years, which is excellent security.

Upon further examination of the radiosonde and its circuitry, I discovered that the temperature and humidity registers in the radiosonde didn't work. Not any of them!

The temperature sensor was useless for recording the temperature, but it did have a function.** It acted as a DOR antenna while the humidity sensor acted as an orgone antenna. If DOR was sensed by the antenna, the transmitter would be broadcast out of phase and bust up the DOR and take the violence out of a storm.

Conversely, transmitting in phase would cause the DOR to build up.

The humidity sensor had the same effect with orgone energy. Transmitting in phase would build up the orgone energy and transmitting out of phase would reduce it.

The radiosonde also contained a pressure element that would act as a switch signal and would maintain either DOR or orgone. This was how they built up the orgone energy.

The transmitter consisted of two oscillators. One was a carrier oscillator, which runs at 403 MHz. The other ran at 7 MHz and is a relaxation oscillator.

This one would pulse on and off depending on what was encountered. Somehow, this monitored the etheric function of the radiosonde. I haven't discovered everything there is to know about radiosonde, but I have done a scientific analysis of it which I've included in the appendix (see Appendix A) for those who are interested.

What I have told you about the radiosonde is hard evidence that can stand up to scrutiny. It establishes the credibility of my story that there was a secret project that involved weather control. We can't say exactly whether the radiosondes were used just to bust up violent storms, but the possibility was also there to build them up. The government abandoned the weather control aspect eventually. Changing weather, if it were proven in court, could lead to many law suits.

What is more intriguing than the weather aspect is the entire prospect of orgone and DOR energy and what could be done with that. In theory, this means that the government could have targeted communities, buildings or an entire populace and transmitted orgone or DOR energy. These type of activities have been reported in Russia for years. Not much press coverage has been given the U. S. effort in this regard, but there has been some activity. Whether it has been used harmfully or in war, I cannot answer, but the potential was there.

Forty years of development could also have made this a very refined technological device.

Please refer to Appendix B for additional information on Wilhelm Reich.

* The word "metrograph" is more clearly defined if you understand that "metro" signifies that it was a meteorological device and that "graph" means to write.

** For those technically oriented, the temperature sensor is essentially a thermistor; but instead of being carbon based, it contains noble metals and exotic elements. It is a very poor temperature sensor because as the temperature cycles it up and down, the resistance curve changes and it doesn't hold its calibration. The humidity sensor suffers from the same problem.

EIGHT- "THE PHOENIX PROJECT" ABSORBS "PROJECT RAINBOW"

While the Phoenix Project was investigating the weather and the use of radiosondes, Project Rainbow resurfaced in the late 1940's. Project Rainbow (which was the code name for the operation that brought about the Philadelphia Experiment) was going to continue research into the phenomena encountered on the USS Eldridge.

This project was concerned with the "electromagnetic bottle" technology, which eventually resulted in today's stealth fighter craft.

At about the same time, Dr. John von Neumann and his research team were called back. They had worked on the original Rainbow project and went to work on a new endeavor. This was similar to the Rainbow Project but had a different gowl. They were to find out what went wrong with the "human factor" of the experiment and why it failed so miserably.

In the early 1950's, it was decided that the remnants of Project Rainbow and the radiosonde project should be included under the same umbrella with the human factor study. After that point, the title of "Phoenix Project" was used to refer to all of these activities.

The project headquarters was at Brookhaven Labs on Long Island and the first order of business was to put Dr. von Neumann in charge of the entire project.

Dr. von Neumann was a mathematician who came to the United States from Germany.

He also became a theoretical physicist and was noted for his very advanced concepts of space and time. He originated the computer and built the first vacuum tube computer at Princeton University, where he also served as the head of the Institute for Advanced Study.

Dr. von Neumann had what could be described as a "good technical feel." He had the ability to apply advanced theories to technology. His background in math gave him enough theory to communicate with Einstein, and he could in turn pass this on to the engineers and serve as a bridge between the two.

As von Neumann began work on the Phoenix Project, he quickly learned that he was going to have to study metaphysics. He had to understand the metaphysical side of man. The Rainbow technology had dissolved the physical and biological structure of human beings. People were stuck in bulkheads and changed beyond recognition in some cases. But it was the esoteric workings of the mind that had been affected first, in each case.

Von Neumann and his team spent about ten years working out why human beings had troubles with electromagnetic fields that shifted them through different places and times. They actually found out that humans are born with what is known as a "time reference" point. At conception, an energy being is attached to a time line and we all start from that point. To understand this, it is necessary to view the "energy being" or soul as distinct from the physical body of the person concerned.

Our whole reference as a physical and metaphysical being stems from that time reference which actually resides within the electromagnetic background of our planet. This time reference is the basic orientation point you have to the universe and the way it operates. You can imagine how you would feel if the clock suddenly started moving backwards and time as well. It is this time reference point that was thrown out of kilter with the individual crewmen of the USS Eldridge and caused them untold trauma.

The Rainbow technology turns on and creates what can be called an alternate or artificial reality. It creates a stealth effect by not only isolating the ship, but the individual beings as well, within a "bottle effect." Those beings were literally removed from space and our universe as we know it. This accounts for the invisibility of the ship and of the people on board. The alternate reality thus created has no time references at all because it is not part of the normal time stream. It is entirely out of time. To be in an artificial reality would be like waking up and not knowing where the hell you are. All of this would be very confusing.

The Phoenix Project was faced with solving the problem of bringing human beings into the "bottle" (and eventually out again) while at the same time connecting them to their real time reference (that they would know as the planet Earth, etc).

This meant that when they were in the alternate reality or "bottle", they had to be supplied with something that would give them a time reference. They solved this by feeding into the "bottle" all the natural backgrounds of the Earth - at least enough to convince them of a continuous stream time reference. To do otherwise, would likely cause those in the "bottle" to experience transdimensional disorder and problems of this sort. This is why it was necessary to set a phony stage. They could then feel some degree of normality.

Dr. von Neumann was the ideal candidate for the job since he knew computers.

A computer had to be used if they were going to calculate the time references of specific people and replicate those references while they were passing through an "electromagnetic bottle" or alternate reality. The people inside the "bottle" would be going through zero time and essentially a "no reality" or a disoriented one at best. The computer had to generate an electromagnetic background (or phony stage) that the physical being would synchronize with as well. If that wasn't done, the spirit and the physical body would go out of synch, thus resulting in insanity.

There are two points to be brought out here: the physical being and the spiritual being. This is why the time reference would lock in the spirit and the electromagnetic background would lock in the body. This whole project started in 1948 and was finally developed in 1967.

When this project was complete, a final report was written and submitted to Congress. Congress had funded this particular project thus far and followed the results. They were told that the consciousness of man could definitely be affected by electromagnetics; and additionally, that it would be possible to develop equipment that could literally change the way a person thinks.

Not surprisingly, Congress said no. They were concerned that if the wrong people got a hold of this technology that they themselves could lose their minds and be controlled. It is a very valid concern and word was given by 1969 to disband the entire project.

NINE- THE MONTAUK PROJECT BEGINS

It is no secret that Congress has tried to brow beat the CIA into finding out everything that goes on in the intelligence community. They have cut their funding, limited their legal powers, and even the most naive person would likely admit to a credibility gap of some degree. However, we are not dealing with the CIA proper here. Indeed, if the CIA is involved, it would be a splinter wing or wings that are being used by a source other than the CIA director.

When Congress disbanded the Phoenix Project, the group at Brookhaven had already built an entire kingdom around this project. They had Reichian and stealth technologies which could definitely affect the mind of man.

The Brookhaven group went to the military and informed them about this fantastic new piece of technology they were working on. They told them about a device that could make the enemy surrender without a battle simply by throwing a switch. Of course, the military was very interested. This was every war expert's dream. Imagine, a device that makes the enemy give up before the battle starts!*

The military became enthusiastic and were ready to cooperate. They were informed that they didn't need to get involved in the financing because that was covered by the group at Brookhaven National labs. But, the Brookhaven people needed a place where proper experimentation could be done in seclusion. They needed certain equipment and personnel from the military. They gave the military a list of all technology required.

Of particular import on the technology list was the old Sage Radar. For this, they required a huge radiosonde that would operate around 425 to 450 MegaHertz.

From earlier research, it was known that this was one of the "window frequencies" for getting into the human consciousness. A very high powered radar device was needed that ran at 425 to 450 MHz.

The military had just what they were looking for: a mothballed Air Force base at Montauk Point that housed an obsolete Sage Radar system that fit the bill.

This system already had the RF sections and the modulator that would be required to build a huge radiosonde.

The Sage Radar at Montauk was originally part of the early warning defense system used during the '50's and '60's. Today, satellites and over-the-horizon radar make this technology obsolete for defense purposes. It certainly raises an important question, even if one doesn't believe this story. Why was an old antiquated defense system turned on and utilized for a period of over ten years?

The name for this project was known as "Phoenix II" by the officials concerned.

It has since been colloquially named by myself and others involved as the Montauk Project.

Up until then, Congress had been informed about what had occurred. But at this point, independent people were carrying forward with a project denied by Congress and were operating outside of any controls. They were even using the U. S. military in the process. Of course, it quickly becomes, "Who is using who?"

But, the point being stressed here is that it was being done without the supervision of elected officials and in spite of their objections.

The Montauk Base was being reopened. The Sage Radar had been shut down since 1969/1970 when the base was turned over to the General Services Administration.

It was a surplus government base without anything on it, and government financing for it had ceased.

It is obvious that major funding would be required for such an endeavor. The financing is shrouded in mystery, but it appeared to be totally private. I do not have documented evidence myself of the financing but have been told by my Montauk acquaintances that the original money came courtesy of the Nazis.

In 1944, an American troop train went through a French tunnel carrying 10 billion dollars worth of Nazi gold. This train was dynamited in the tunnel while carrying 51 GIs. General George Patton was in Europe at the time and investigated this, but he couldn't understand how an American troop train could be dynamited in western Allied territory. As a general and human being, he cared about the GIs. The 10 billion dollars was also a mystery, but Patton's efforts were blocked.

I've been told this gold eventually showed up at Montauk, and it was 10 billion dollars of gold priced, then, at $20 an ounce. This was the equivalent of almost 200 billion dollars in today's currency. It was used to finance the project initially and for years to come. After it was all spent, the project was allegedly financed by the infamous Krupp** family, who controlled the ITT corporation.

In late 1970 and 1971, the Montauk Air Force Base, 0773rd Radar Battalion, was actively being reestablished. They had to establish a staff, get the equipment working and set up the whole research facility. This took about a year, and by late '71, the Montauk Project was underway.

The strictest security measures were employed, part of which were entirely valid. Although confidential stealth technology was involved, it is no secret that the stealth aircraft was designed with a radar resistant absorbine coating and a reduced surface cross section. What is secret are certain aspects of the "electromagnetic bottle" technology and how that was propagated. We're not going to discuss this or describe it as it remains a duly authorized military secret that concerns the defense of the United States. With this book, we are concerned with disclosing a project that should never have been activated in the first place. With no military or defense purposes to begin with, it was only designed for controlling the minds of the population and in spite of Congress forbidding this project.

The staff was a mixture of military employees, government employees and personnel supplied by various corporations. I was one of the latter and came to the project in 1973.

There were a number of Air Force technicians who had worked on the Sage Radar in the '60's. The Air Force had assigned them to Montauk even though it was listed on the books as a decommissioned, derelict base. The technicians told the Phoenix group that they could change the general mood of the base by changing the frequency and pulse duration of the radar. they had noticed this as a professional curiosity after years of working with radar.

This was a surprise to the Phoenix people, and they found it very interesting.

By changing the pulse rate and pulse width, they could change the general way people were thinking. This was what they were looking for.

This new information prompted what I now refer to as the "Microwave Oven" experiments. They took the reflector (which looks like a huge banana peel and which can be seen from a distance when you are at the Point), rotated it almost due west and angled it down so that it was focused on one of the buildings, in what they though to be a safe place.

Inside that building they had a chair inside a shielded room. First, they'd sit someone in the chair - this was usually Duncan Cameron. Then, they would open and close the door to determine how much UHF/microwave energy was getting into the room. All this was being done while the antenna was rotated and focused to a point in front of the building. At the same time, the transmitter was blasting gigawatts of power.

They experimented by running the transmitter at different pulse widths, different pulse rates, and different frequencies. They tried everything they could think of, just plain empirical experimentation. They just wanted to see what would happen to the person in the chair if he was bombarded by "x" frequency, pulse, etc. They observed that certain changes made a person sleep, cry, laugh, be agitated and so on. There were rumors that whenever the Sage Radar ran, the mood of the whole base would change. This was very interesting to the project supervisors as they were primarily concerned with the study of human factors.

They wanted to see how they could train and change brain waves. This was done by changing the repitition rates of the pulse and the amplitude in correspondence to different biological functions. In this way, a person's thoughts could be controlled. With the 425-450 MHz of radio frequency power, they actually had a window into the human mind. The next step would be to find out what was inside of it.

Although the door to the shielded room was closed most of the time, it didn't work properly. The subjects were exposed to a strong enough field to influence the brain waves but not enough to do damage. However, if exposed to it for several days on end, it could be quite damaging.

Duncan sustained serious brain and tissue damage as a result of continuous expsure to 100 kilowatts of RF power at a distance of about 100 yards. The radio waves baked his brains and chest. Anywhere in his body where there was a change of density, zones of heat or energy would be created by the concentration of the microwave beams.

Upon visiting a doctor in 1988, Duncan's doctor commented upon the unusual scar tissue in his lungs. He'd never seen anything like it. Another doctor who was consulted said he'd only seen it in the service when someone had gotten in front of a high powered radar beam.

Previous research in or about 1986 indicated that Duncan was actually brain dead. Initially, I had asked different psychics to do readings on Duncan. They determined he was brain dead. I also knew that it was possible to inject a particular dye into the brain and have x-rays or CAT scans can reveal what areas of the brain are using oxygen. Brain dead individuals suffer from a lack of oxygen to the cerebrum. If the psychic readings were accurate, his brain would not be using much oxygen.

I asked a neurologist with whom I was friendly with, and he said it was definitely possible that some one could be brain dead and yet be walking around. He cited some post mortems done on people in England in the U. S. whose brains had unusual coatings inside the skull. The coatings were about a millimeter thick.

More interesting yet was a case he'd encountered about ten years ago. He took out a group of x-rays of a normal human being and showed me the red areas. He also indicated blue areas but told me they were ares that didn't require much oxygen.

Then, he put up another x-ray where the entire brain was blue. This meant that the person was alive and was walking around like a normal human except he has memory loss problems from it. He was essentially brain dead and the brain was using just enough oxygen to keep it from rotting. I noticed the corner of the x-ray and was surprised to see Duncan's name. Based upon this information, Duncan is indeed brain dead.

I asked the doctor for an explanation, but he wasn't sure. He could only offer a theoretical conclusion based upon psychic powers. He said that his profession recognized the existence of psychic phenomena but did not understand it.

At this point, we learned that the only reason Duncan is alive today is due to his strong psychic aptitude. The psychic part of his mind takes over the physical part of his mind and runs the body. His brain stem is alive; his spinal chord is alive; his body is alive, but his actual higher brain is dead. His psychic energy runs the body through the brain stem.

Duncan was not the only person affected. We don't know how many people were involved but the body count was probably high.

It wasn't until 1972 or '73 when it was finally realized that stealth technology dealth with non-burning radiation. One theory was that actual non-burning radiation, which is the higher order of components (as opposed to burning radiation), actually went through the reflector and would be opposite to the focal point of the antenna.

THey tried it and turned the antenna around 180 degrees. They aimed the burning rays into the sky and hit the person with the non-burning rays. Then, they found they had the same mood altering capabilities, if not more than they had before, but this did not damage the people. But at what cost to the persons previously experimented on!

At this point in the project, they were interested in monitoring people and changing their thoughts and moods, etc. It was not necessarily how they changed but the fact that they changed under certain circumstances. Different army units were invited to come to the base and have R&R there. As far as the soldiers were concerned, it was free R&R in a beautiful location.

The outer base had a nice gymnasium and a bowling alley with excellent food and accommodations. Unbeknownst to the servicemen, they became guinea pigs for the mood control experiments. However, these were not the only guinea pigs.

Experimentation was also done on the townspeople, Long Island, New Jersey, upstate New York and Connecticut civilians, just to see how far it could go. However, most experimentation was done on the vacationing soldiers.

Time was spent monitoring different pulse types, trying this and trying that.

They would note and categorize the different effects. It was all pure empirical experimentation and a huge data base was collected. Once they had enough data, they began to make some sense out of which functions did what.

During this period, they also experimented with frequency hopping. Frequency hopping consists of the transmitter instantaneously and randomly shifting around to any of five different frequencies (that were being fed to the transmitter). This point became veryimportant later on as it was key to bending time.

They discovered that very fast frequency hops made the modulations more psycho- active. A data base was then developed that would list the frequency hop times (times you go from one frequency to another), how they pulse modulated, the rate the pulse modulated at, the pulse width, and the power output they pulsed it at.

This was then coupled with the responding effects it had. The data base was very extensive and covered an extremely broad range of causes and effects.

After the extensive experimentation, they developed a control panel with which they could set different pulse modulations and timings. They knew that these different pulses and functions represented certain thought patterns from the individual. They could set the modulators and timings so that a transmission would be generated that would place thought patterns into an individual. This meant they could literally set this pulse at anything they wanted and expect a desired effect to take place.

All of this took about three or four years to research. The transmitter was now fully operational and hooked up. Programs could be typed in that would put the transmitter through its phases. Programs were derived that could change the moods of people, increase the crime rate, or make people agitated. Even animals within the vicinty were programmed to do strange things.

The researchers were able to derive programs whereby they could focus on a car and stop all the electric functions in it. I don't know what the modulations were, but I understand they found this quite by accident.

One day, there were military vehicles riding around the base. They suddenly ceased to operate without any function. An investigation ensued to find out what was occurring with the transmitter at the time, and a program was developed. At first, the program could only get the lights in a car to dim. It was eventually refined to where the program caused all electrical functions in a vehicle to cease.

Several years of research and collecting information had finally yielded a mind control device. The next objective was to create a precision technology with the material. In order to do this, help from very strange sources was enlisted.

* I have included in Appendix C some vidence that suggests mind control devices were used against the Iraqis during the Persian Gulf War.

** The Krupp's were the owners of the German munitions factories for World War I and II. After being found guilty of war crimes and complicity with Hitler at the Nuremberg Trials, the head of the Krupp family was paroled from a light prison sentence and allowed to continue his notorious arms dealings.

TEN- THE MONTAUK CHAIR

In the 1950's, ITT developed sensor technology that could literally display what a person was thinking. It was essentially a mind-reading machine. It operated on the principle of picking up the ectromagnetic functions of human beings and translating those in an understandable form. It consisted of a chair in which a person would sit. Coils, which served as sensors, were placed around the chair. There were also three receivers, six channels and a Cray 1 computer which would display what was on a person's mind - digitally or on a screen.

It is still a mystery how this technology was developed. It has been suggested that the research was aided by the Sirians, an alien race who come from the star system known as Sirius. This theory has the aliens providing the basic design and humans working it out from that.

Three sets of coils were set up in a pyramid around the chair. There was also a coil around the top of the pyramid to parallel the base coil. The person would be placed inside the field of the coils. The three sets of coils were connected to three different radio receivers (Hammerland Super ProP 600's) and six outputs.

An independent sideband detector, which had a floating carrier reference system, would provide six outputs from the three receivers. Three of them were of the sideband below the carrier wave. Three of them were of the sideband above the carrier wave. This brings to mind a very important question. If this device was reading minds, what was the carrier wave it was using to do this?

With the use of an oscillator, the detectors in the receivers were able to lock on a phantom or etheric signal that was being picked up by the coils. There was no actual carrier wave as we would normally know it. the detectors would lock in on the noise peak that the coils picked up from the three sets of frequencies the receivers were tuned to.

At this point, the research team was actually able to detect the signals that represented the comparable functions of the human mind. Solid signals that would change with a person's thoughts were actually coming out of the receivers. This device was actually reading the human aura, which is a word that psychics and metaphysicians use to describe the electromagnetic field that surrounds the human body. In the same way that human speeck is carried via radio waves, this device was carrying thoughts (which theoretically manifest in the aura).

The six output channels from the receivers were then run through a digital converter (turning them into computer language) and fed into a computer. A Cray 1 computer was used to decode what the receivers were picking up. A lot of hard work and a lot more computer crunching got things to the point where the computer could print out a dialogue. This would be a running dialogue of the person thinking.

More work got it to where the person would visualize something and a picture would actually appear on the computer monitor. Improvements and refinements continued until a 3D representation of the actual audio/visual aspect (of the person's thoughts) appeared on the computer monitor and could in turn be printed out.

When the people at Montauk heard about this mind reading device, they thought it was great. They wanted to turn this mind reading machine into a transmitter. This could possibly cut the risks to human beings undergoing invisibility or time experimentation. The theory was that a person in the chair would transmit an alternate reality to the crew (like in the Philadelphia Experiment). When the ship became invisible, the crew would then be insynchronization with the alternate reality and wouldn't become disoriented or mentally lost.

At this point, a chair was procured, which we now refer to as the famous "Montauk Chair". It was hooked up to the coil set-up from ITT. The Cray1 computer, which was used to decode the transmissions being generated from the person in the chair, was then interfaced with an IBM 360 computer. This was, in turn, interfaced with the Montauk transmitter.

The IBM 360 was needed to control the modulation of the transmitter so that the transmitter could frequency hop across the entire band.

It was about this time that I remember Al Bielek taking on a key role. Al is one of the authors of 'The Philadelphia Experiment and Other UFO Conspiracies.'

He has memories of being involved in the Rainbow Project as well. Originally, he was brought to the project to explain what was going on metaphysically with the use of the transmitter on human beings. He was chosen because he not only had an engineering background, but he was psychically sensitive and had an extensive knowledge of esoteric matters.

It now became Al's job to help interface the Cray 1 computer with the IBM 360.

The Cray 1 was putting out tons of information. They didn't know what to do with it and needed someone with esoteric knowledge to figure it out. They had to convert what the Cray 1 was putting out so that it would synchronize with what the pulse modulation comupter wanted. The IBM 360 served this function and was essentially used as a translator and storage bank for what the Cray 1 was outputting. Al got very heavily involved because he was part of the team that figured out what program to put on the IBM 360 that would translate the Cray 1 output to drive the transmitter.

The transmitter had a modulation computer which was digitally fed the typical 32 bit code that the 360 put out. The modulation computer and the transmitter were set.

The IBM 360 would tell the modulation computer how to modulate the transmitter.

Now we had a system where one could put in 32 bit words of data and the transmitter would give back something. And here the chair fed the receptors feeding the Cray 1 which would tell what the person was thinking. They had to take this and translate what was coming out of the Cray 1 and make it so that the IFM 360 could re-encode the thought form that was actually transmitting. It took about a year to successfully link up the computers.

I had joined the project at this time to work with the radio frequencies and transmitter. Although some linkage had been achieved with the computers, they were having huge problems with feedback from the transmitter to the chair. The solution to the feedback was to move the chair down the coast to the ITT center in Southampton, Long Island. A psychic would then sit in the chair in Southampton and relay via computer to the Montauk transmitter.

The psychic would think thoughts, and the Cray 1 would decode them. They'd be put on a 32 bit radio link and sent Montauk where they would go into the IBM 360.

The IBM computer would then broadcast it out the transmitter and could build a thought form out at Montauk of what the psychic was thinking in Southampton. The device was essentially a mind amplifier.

It took another year of research before they could get a readable signal (based upon what the psychic was thinking at Southampton) sent to Montauk and out the transmitter. This was their first objective: get some thought fidelity from the chair through the Montauk transmitter and out the antenna. Besides Duncan, there were a couple of additional psychics on site. They literally tuned up the computer programs. Finally, the thought forms became clear. The psychic could concentrate on something in Southampton and the transmitter at Montauk would transmit a very clear representation of what he was thinking.

That was the first point at which the Montauk transmitter was working with high thought fidelity.

In another year, I recall as early '75, they discovered another problem. If there was a glitch in the flow of time in our reality, everything fell apart. In other words, if the psychic in the chair projected a reality (in terms of time in this case) that was not consistent with out reality (i. e., the flow of time in our reality), it would cause the connection between Southampton and Montauk to break up. Any glitch in space-time between the two cities would cause the transmission of the thought form to cease.

To better understand a time glitch, imagine time as a continuous pulsation or flow. As the basic pulsation of time interacts and changes form with other flows or phenomena, we have motion as we know it, against the backdrop of time. When these core pulsations that make up time are shifted (due to a reality change or other phenomena); the direction, speed, or flow of time is changed. This is what is known as a time glitch. Theoretically, these occur every now and then, since we are referenced in our reality, we really don't notice a time glitch. Deja vu phenomena could well be an example of a time glitch in the fabric of time.

With the chair in Southampton, the mind control experiments with the transmitter were not always working. This was attributable to the time glitches. It was also known that if a large amount of power was fed into the transmitter during a time glitch, there could be disastrous effects.

It now became imperative to get the chair working at Montauk. They first put tremendous shielding around the chair so that the electromagnetic fields at Montauk would not affect it. That didn't work, so they tried putting the chair in an electromagnetic dead zone. They picked the best dead zone available, but this was not successful either.

They worked through mid '75 but continued to have difficulties until they consulted the original prototype that the chair was based upon (allegedly devised by the Sirians). This device was not identical to the one ITT had created. It had a different kind of coil set-up wherein the coils were connected to crystal type receivers. These were actual crystals and not ordinary electronic devices.

After review of the prototype, secret bids for a new chair were put out and RCA came up with the winning bid. Nikola Tesla* had designed receivers for RCA in the '30's. Tesla's work during this period was done under the name "N. Terbo", which referred to his mother's maiden name. These Tesla receivers had very special coil structures. They were normal type radio coils but were arranged in strange coupling patterns as set up and designed by Tesla.

The set up of the Montauk Chair was also enhanced by using Helmholtz coils.

These were placed around the chair to serve as pick up coils. In ordinary electronics, Helmholtz coils consist of two sets of coils. They possess a unique property in that they can be phased to create a constant field (of energy) inside the coils.

At Montauk, the researchers extrapolated upon the principles of Helmholtz coils.

They used three sets of coils (X, Y and Z), and phased them so that while a constant energy couls be maintained inside the coils, there was absolutely no effect on the outside.

The coil structure in the receivers designed by Tesla was ideal for the Montauk Project. Not only would the chair be in a coil structure, but so would the receivers themselves. This would shield the energy field.

It should also be noted that the coil sturctures in the Tesla receivers are also known as Delta T or Delta Time coil structures. The property of shielding an energy field is part of what enable a "bottle effect" to be created around the USS Eldridge in the Philadelphia Experiment. These Delta T coils were actually picking up three axes of time signals. More pertinent to the project, they no longer had a microwave link that would malfunction during a reality shift.

To get the Montauk chair operating without interference, they had to replicate what the crystal receivers did with the "Sirian" technology. The coil structures in the prototype receivers were Delta Time coil structures. And the receiver itself did the Delta Time function, but not the antenna. ITT had the Delta Time function in the antenna instead of the receivers. The RCA version used standard type Helmholtz pick up coils that could accomplish Delta Time conversion in the receivers. They also had the same kind of detector system and oscillator locks that ITT used with the Cray 1 computer.

At this point, it now became inclusive of the coil only. Outside, the coil structure, there was no sensitivity. They could put the chair in the dead spot that was between the transmitting antenna on top of the transmitter building and the transmitting magnetic antenna that was underground. This was in the underground basement of the transmitter building, which had already been tightly shielded. In the next room, they had these three specially designed receivers with another rack of equipment. These were used to synchronize all local oscillators with the signal, similar to the ITT system.

Now, the antenna, the transmitter, and the chair were in the same time plane.

The computers were in their own time plane. It didn't matter that they had the chair underground and the Cray and 360 in the other building (feeding back to the transmitter building). When everything is digitized, one is no longer in real time. A "fake time" is created. The computers could have been located anywhere. The computer building was designed to operate computers and shielded out electromagnetics and energy didn't drive the computer insane. The operation center was totally shielded in cement and steel.

Finally, they created the second and last generation of the Montauk chair.

It performed the same purpose as the first chair. It brought the same six channels of information to the computer, but there was an additional advantage. It was immune to the signal from the antenna. Now, the signal from the antenna didn't feed back and cause interference. So, they had everything on site. They spent another six months until about late '75, early '76, just aligning, adjusting, and making sure everything was working.

They finally got the transmitter functioning, which was quite astounding. What happened afterwards was even more so.

* Nikola Tesla was an electronic genius who was the first to discover and apply the principles of alternating current. With the financial backing of George Westinghouse, he revolutionized the way electricity was used across the world.

See Appendix D for more information on Tesla.

ELEVEN- CREATION FROM THE ETHER

Once they had the transmitter working, it took about another year to work out the computer programs so the system would receive and transmit all psycho-active functions. By late '77, the transmitter was reproducing thought forms without glitches and with a very high degree of fidelity. At this point, they pulled out all the stops. They had the psychic, Duncan Cameron, concentrate on a solid object, and guess what happened? The solid object actually precipitated out of the ether!

In his mind, he would concentrate on a solid object, and it would appear somewhere on the base. Whatever Duncan would visualize, the transmitter would transmit the lattice (or matrix) for, and build enough power to materialize whatever he was thinking of. Every single point to where he could witness to a particular spot on the base, at that spot an object would materialize. In other words, if he would hold an object in his hand and/or visualize it, it would appear at the given spot.

They actually had discovered pure creation out of thought with the use of the transmitter.

Whatever Duncan could think up would appear. Many times, it would be only visible and not solid to the touch, like a ghost. Sometimes, it was a real solid object that was stable and would stay. Other times, it was a solid object that would remain as long as the transmitter was turned on and then fade out as the transmitter was turned off. The read out from the computer gave an accurate representation of what Duncan was thinking. The researchers could then select what thoughts would be broadcase out of the transmitter. Most of these though forms were broadcast in the vicinity of the Montauk Air Force Base, but other locations were used as well.

What Duncan though of as a subjective reality would be created as an objective reality (either solid or transparent, depending on the circumstances). For example, he could think of an entire building and that building would appear on base. This type of experimentation was routine.

The system worked with a good degree of fidelity. Now, they wanted to see what they could do with it. the first experiment was called "The Seeing Eye." With a lock of person's hair or other appropriate object in his hand, Duncan could concentrate on the person and be able to see as if he was seeing through their eyes, hearing through their ears, and feeling through their body. He could actually, see through other people, anywhere on the planet. This style of experimentation was extensive, but I don't know how far it was taken.

It is truly incredible that such a feat could be accomplished, however the agenda employed was more sinister than incredible. They were interested in controlling how human beings think. The next move was to see if they could put thoughts in the head of another person. For instance, they would have Duncan meet a subject individual.

Subsequent to the meeting and unbeknownst to the individual, Duncan would concentrate on the individual. Ninety-nine percent of the time, the subject would get thoughts similar to Duncan's. Being able to push his mind so far into the mind of another being, Duncan could control another person and make them do anything he wanted.

This control factor was on a deeper level than ordinary hypnosis.

Through Duncan, the equipment and the Montauk transmitter, scientists could actually load information, programs and commands into the individual's mind. Duncan's thoughts would become an individual's own thoughts. And, using this process, an individual could be made to do something he wouldn't ordinarily do. This was the start of the mind control aspect of the Montauk Project.

This line of research continued until about 1979. Many other different experiments ensued. Some of them were interesting, but others had horrible consequences.

They would target individuals or masses of people, animals, places and technology.

They could basically target anything they wanted. For example, a TV set could be made to go haywire. They could stop the picture or shut it off entirely. They telekinetically moved objects and destroyed rooms.

In one particular case, Duncan concentratedon shattering a window. Enough force was generated to the point where it actually broke a window in the nearby town of Montauk. Animals could also be made to charge off Montauk Point and into the town. Humans could be influenced to start a crime wave.

One must realize that when Duncan did these experiments, he was in an altered form of consciousness. He had been given special training which could possibly have been administered by the CIA or NSA. In any event, his conscious mind would be diverted through sexual bliss. What could be termed the primitive mind would then surface.

Duncan, the individual, would be transferred into an orgasmic trance. His primitive mind, at the disposal of the researchers, became very suggestible and therefore controllable.

For this programming, information could be installed via any of the body's senses. Duncan would then be directed to have his primitive mind concentrate on the information thus installed. For instance, once his primitive mind surfaced and was told to concentrate on something, it would concentrate with its whole being. His whole mind would focus on one subject while his body went into suspended animation.

The primitive mind could also be cleared of previous programming, and something else could be inserted. There was a literal translator, whereby they could program in whatever they wanted. Spoken words, written words, movies, music or whatever was needed was employed to work the primitive mind.

These techniques were the key to getting clear thought forms from the transmitter that would either affect another person's mind or bring creation out of the ether.

By 1978, the mind control techniques were fully developed and recorded. Appropriate tapes were made and distributed to different agencies so they could be developed into something practical.

TWELVE- TIME WARPING

As the experiments continued throughout 1979, a very peculiar phenomena was noticed. As Duncan's thoughts were projected out through the transmitter, they would suddenly cease. This was disappointing and appeared to be a malfunction.

Eventually, it was noticed that the projection of Duncan's thoughts hadn't ceased.

They were just occurring out of the normal time stream!

For example, he would concentrate on something at 8:00 PM and the object or occurrence would happen at midnight or even 6:00 AM. Whatever he thought of would not happen at the time he thought of it.

It now appeared that the Montauk scientists could now use Duncan's psychic powers to actually bend time!

They eagerly started to research this phenomena. We were all required to attend what were known as the "Sigma Conferences", which were held near Olympia, Washington.

These conferences were on the subject of time functions, and we were there to gain a better knowledge of how time works. We were told we had to optimize the use of the transmitter for time manipulations.

We learned that the equipment being used was strong enough to bend time, but it wasn't doing a complete job. The antennas being employed were giving us what could be a side effect of "time warping." This side effect of time shifting did show, however, that the basic equipment was sufficient to do it. But, we required an antenna that was much more effective in creating time potentials.

After going to several conferences and talking to many people, our research group decided that the radio frequency being used was not working. Changes had to be made, such as setting up pulses into a coil. We also studied pyramid based geometry and how to use that to bend the time field. Additionally, we had to learn more about what is known as the Delta Time function (time changing function).

The key clue to our understanding time was a suggestion that we use a particular type of antenna structurde, which I now refer to as an Orion Delta T antenna. It is referred to as "Orion" because there was a persistent rumor that the design was given tot he project by aliens from the Orion constellation (this is a different group of aliens from the Sirians, whose knowledge was alegedly used for the Montauk chair). According to the rumor, the Orions knew we were close to achieving our task and had their own agenda for helping us.

The Orion Delta T was a huge octahedronal antenna, and it was placed underground.

Its height was about 100 to 150 feet from point to point. Excavations were completed to about 300 feet to house the antenna under the transmitter.

The Montauk chair was placed under the transmitter and above the Delta T antenna.

This was done in order to phase the ground RF antenna with the below ground loop antenna so that the chair was in a null point between them. The null point was meant to cut out the interference even deeper. It knocked the interference right out of the chair - completely.

The Delta T transmitting antenna was supplied by three drives. Two of the drives came from the pulse modulators of the two transmitters and fed into the x and y coils of the Delta T. (The same pulse that supplied power to the amplitron also supplied power to the Delta T antenna that was underground). The third axis was the z-axis.

It was placed around the perimeter of the antenna and was derived from a white noise* source that came from a 250 kilowatt audio amplifier. The white noise correlated the whole transmitter and more will be said on that later.

The RF was fed into an omnidirectional antenna located above ground on the top of the transmitter building. Additionally, the non-hertzian component (which is etheric in nature) of the RF made it below ground and interfaced with the magnetic field that had been generated underground. When these frequencies are summed in that manner, time disturbances and distortions result.

The basic techniques were the same as those employed in the Philadelphia Experiment.

On the Eldridge, they had the RF transmissions on the main mast of the ship. The coils were placed around the deck and were driven by pulses. We had essentially duplicated by upgraded the Rainbow Porject machine. This technique also made the project far more controllable.

In addition to the Delta T antenna, there are two other key points to understand: zero time and white noise.

Zero time was referred to previously, but I will give a more complete understanding of it now. First, zero time is outside the realm of our normal three dimensional universe. It would be considered senior to the createed world as zero time existed prior to our created world. Zero time is our basic connection to the universe.

As our universe roates, it rotates around zero time. But our universe is not the only one. Every universe has a zero point. All the zero points of the different universes coincide and never move: that is why it is called a zero point.

It may help to imagine a carnival style merry-go-round that revolves around a central booth. The man inside that booth would represent the zero point. In addition to the merry-go-round, there would be several more merry-go-rounds at different levels, but all would be under the control of the central zero point booth.

A zero time reference generator had already been constructed by Nikola Tesla in the 1920's. It consisted of an assortment of spinning widgets and rotating wheels. colloquially, we referred to it as a "whirligig". It is a strange device because when you turn it on, yu can hear it "lock in" to something, but we are not referring to the power line. I'm told it locks into the rotation of the Earth itself, which is a secondary zero time reference. It is secondary because the Earth's rotation is inertially related to the solar system, which is inertially related to the galaxy, on down to the universe. The universe rotates around the zero time point.

One can get an even better understanding of this by reading up on Tesla and how he discovered alternating current by applying the principles of the rotating magnetic fields of the earth. The zero time generator is to some degree an extrapolation of that, however it doesn't just refer to the rotation of the Earth. It takes into account the orbit of the Sun, our galaxy, and ultimately the center of our whole reality.

The other key point to understand is white noise. White noise could be considered the glue that makes the whole operation work. It basically made the whole transmitter system coherent. It is a highly technical operation which I will simplify.

The Sage transmitter contained something like forty or fifty crystal controlled oscillators, mixers and amplifiers that generated a 425MHz signal. It also had "frequency agility", which meant it was able to spontaneously switch from one frequency to another.

Along with the transmitter, they had what is called a "COHO" or a "coherent oscillator set-up." Normally, a "COHO" would function by having only one frequency reference. However, this is not how the Montauk transmitter achieved coherency.

In order to make it entirely coherent, we took every oscillator available and amplitude modulated it with white noise. Since white noise is fifty percent correlated to everything, it serves a universal auto-correlating function. The result was that all of the etheric components of the oscillators were now coherent to each other. We weren't trying to correlate the normal electrical functions as they didn't concern us. We were only interested in the etheric functions, as they gave the results we were searching for.

A very stable time reference was required from the zero time generator. This produced two 30 hertz waves, referenced to zero time. One was connected to the computers and synchronized the clock or timing functions. The other modulated the white noise generator. By adjusting the phase between them, we could focus on and monitor the whole operation. This enabled us to take the correlations of the white noise and refer it right to the center point of time, where all time crosses.

The purpose of this experiment was to make the psychic transmissions of Duncan time coherent. Dr. von Neumann had instructed us that the transmitter had to be time coherent with respect to zero time. The whirligig zero time reference also served as a witness point back to the Philadelphia Experiment, and that was very important. The project was trying to open a door to the USS Eldridge in 1943.

Modifications continued on the equipment through 1979, until we had a coherent transmission system with respect to time phase.

Now, they had to calibrate Duncan. This meant they had to adjust and modify the equipment to synchronize with him. He had already demonstrated that he had zero point references of his own when the inadvertent time bending had occurred. This could perhaps be better explained due to his prior experience during the Philadelphia Experiment. There, he had jumped off the Eldridge and was thrust into a time vortex.

At Montauk, he was now in an entirely new set of circumstances, but his familiarization with zero time had apparently never left him.

There were also other psychics, but Duncan was the first they had used, and he was in the chair ninety percent of the time the system was in operation. If he was sick or didn't feel well, they'd wait a day. Because every time they changed the operator, they had to recalibrate and reprogram the computers and pulse modulator, and it took about two full days to do that. If Duncan was out for two weeks or more, they'd put in another operator, but I only remember one time when they did that. It was almost a disaster, because they didn't spend enough time in the initial calibration. From then on, Duncan was the one and only one who ran the equipment. A backup had to be there, however, in case something happened to Duncan.

By 1980, the big radar reflector (that looks like a huge banana peel) on top of the building was no longer in use. Now, there were two transmitters that fed the omnidirectional antenna (the one above ground). The pulse modulators of the transmitters were feeding both that antenna and the coils of the Delta T antenna (underground).

Also connected to the computer was the Montauk chair, which was now placed between both antennas at the null point. By this time, the computer system was huge and was housed inside the control room next to the radar tower. Additionally, the computer room contained a lot of different terminals and displays to monitor the various activities of the project.

Duncan would start out sitting in the chair. Then, the transmitter would be turned on. His mind would be blank and clear. He would then be directed to concentrate on an opening in time from say, 1980 (then the current time) to 1990. At this point, a "hole" or time portal would appear right in the center of the Delta T antenna - you could walk through the portal from 1980 to 1990. There was an opening that you could look into. It looked like a circular corridor with a light at the other end.

The time door would remain as long as Duncan would concentrate on 1990 and 1980.

I've been told by those who entered the tunnel that it looked like a spiral, similar to science fiction sytle renditions of a vortex. When outside the tunnel, it looked like you were looking through space - from one circular opening through space to a circular but little bit smaller window at the other end. I was considered too valuable to the technical operation and was not allowed to travel through the portal.

From 1980 to late 1981, the time function was calibrated. At first, the time portals would drift away. One might go throught the portal and come out in 1960.

But when one went back to find it later, although it was still being tracked in real time, the portal would not appear where it should have been. One could easily get lost in time and space. Initially, the portal would be opened up, but it would drift. This was because Duncan himself was drifting. He had to got through extensive training to get the portal to be stable. We also had to focus the transmitter more closely and tighten up the thought form translation to get everything right.

We would spend days just rying to get a particular time change to occur as predicated.

However, there was no particular problem with creating a time warp. Predicting what it would do was the difficulty. Finally, towards the end of 1981, we learned how to stabilize it so that when a portal would appear, it would remain. Although the function was not absolutely perfect, it was predictable, stable and running according to plans.

Essentially, what the scientists were doing was using 1943, 1963, 1983 vortex, which was based upon the natural twenty year biorhythms of the Earth .1943, 1963 and 1983 acted as anchor points for the main vortex. Sub vortices or open ended vortices would be created by going from the main one through an anchor point ('43, '63 or '83). At Montauk, August 12, 1983 was used.

For example, let's say they wanted to reach November 1981. There would be a bridge point from November of '81 to August 12, 1983. From August 12, '83 they could go to whatever time they wanted. The vortex ran between August 12, 1943 and August 12, '83 because that was the master vortex. It is called open ended, because there is no device at the other end which anchoring it.

Although they had stabilized the time aspect of the portals, they had to work on the spatial aspect as well. They stabilized this aspect so that they could not only place a portal at a particular time but in a particular space.

Once time stabilized and the above was accomplished, they kicked out everybody and cleared the entire base except for a few key persons. I remained there as I was the technical operator and was essential to the project. Duncan remained as he was the psychic who made the operation work. The entire system was tuned to him.

Two other psychics were also retained as back-ups in case Duncan was killed or incapacitated. The project directors also stayed, but the military left. A whole new team was brought in to do the more mundane functions of maintaining the base.

Up to that point, everyone operated on a "need to know" basis. Security was already tight, but they wanted even higher security. They didn't want the military to know what they were doing with time. But everybody knew there was something weird going on. They just didn't know what.

* White noise is an impulse at every frequnecy at the same time. When you

are tuning your FM radio dial, the noise you hear between stations is white noise.

It can be thought of as a sudden burst at every frequency or a bunch of impulses thrown together.

THIRTEEN- TIME TRAVEL

As most of the technicians were gone, a new technical crew was brought in.

I don't know who they were and what their qualifications were, but they were called the "Secret Crew." The project was relaunched and is now sometimes referred to as "Phoenix III." This lasted from February 1981 until 1983.

The objective now was to explore time itself. The crew began to look at past history and to the future, just scouting around. They would search ahead for a hostile environment. Through the vortex, they could sample the air, the terrain and everything without entering the portal.

Those who travelled through the vortex described it as a peculiar spiral tunnel that was lit, all the way down. As one started to walk down, he would suddenly be pulled through it. It propelled one out the other end, usually in another place (as opposed to Montauk), or according to where the transmitter was set or placed. It could be anywhere in the Universe.

The tunnel resembled a corkscrew with an effect similar to lit bulbs. It was a fluted sort of structure and not a straight tunnel. It twisted and took turns until you'd come out the other end. There, you would meet somebody or do something.

You would complete your mission and return. The tunnel would open for you, and you'd come back to where you came from. However, if they lost power during the operation, you'd be lost in time or abandoned somewhere in the vortex itself. When someone was lost, it was usually caused by a glitch in hyperspace.* And although many were lost, the scientists didn't abandon people deliberately or carelessly.

According to Duncan, there was also another function of the time tunnel. About two-thirds of the way down the tunnel, one's energy leaves the body. One would feel a big thump accompanied by a tendency to see on a broad scale. He reported sensing a higher intelligence along with an out-of-the-body experience. This was referred to as a FULL OUT. The researchers would try to manifest this in Duncan.

It could have been for further "Seeing Eye" experiments or for other reasons.

It was routine to create a tunnel, grab somebody off the street and send them down. Most of the time these people were winos or derelicts whose absence wouldn't create a furor. If they returned, they would make a full report on what they had encountered. Most of the winos used for the experiments were sobered up for a week before entering a portal, but many didn't make it back. We don't know how many people are still floating around in time, whenever, wherever, and however.

As "Phoenix III" developed, the individuals so chosen for this reseearch would be wired up with all sorts of TV and radio equipment so that they could report back "live." Each individual would be escorted through the portal, sometimes with force. TV and radio signals would travel right through the portals and as long as they could pick these up, researchers would have radio/video tapes of what the time traveller had experienced.

Those controlling the project began to play all sorts of games, manipulating the past and future. I don't really know what they did because I was the one at the switch. My station was in the transmitter building, and I had to keep everything going. I was not privy to a lot of what was going on, but at one point I do know that they had an extensive library of videotapes. I saw the tapes themselves although I was not granted extensive viewing privileges. Actually, I designed and built the viewer (with the aid of tremendous resources) so I had some idea of what was going on. Much of what I knew came from Duncan's own reports, because by that time, we had become good friends. Eventually, we were debriefed and sent on our separate ways. Most of my memories of him had been wiped out.

In addition to the derelicts, the reserachers also used kids for some reason.

I'm not sure what exactly the purpose was, but there was one kid at Montauk who would go out and get other kids and bring them to the project. He was like a tractor beam. He lived in Montauk and would circulate around very effectively. There was also an entire corps of these around the New York metro area that could get away for six hours or so without being missed. They were specifically trained to go out and bring in other kids. Some kids returned home, some didn't. The kids chosen were between 10 and 16, or maybe 18 at the oldest and 9 at the youngest. Most were just about to reach puberty or had just finished it. They were usually blond, blue eyed, tall and light skinned. They fit the Aryan stereotype. To my knowleddge, there were no girls in this group.

A later investigation showed that Montauk had a NeoNazi connection and that the Nazis were still on the Aryan kick. We don't know where the kids went, what they were educated in or programmed for. Whether they came back or not is still a mystery.

What information is available is that they sent every raw recruit into the future to 6037 AD, always to the same point, to what appeared to be a dead city in ruins.

Everything was stationary, not unlike a dreamlike state. There were no signs of life. In the center of the city was a square with a gold horse on a pedestal.

There were inscriptions on that pedestal, and recruits were sent there to read what they said. Each recruit would interpret and report. We still don't know what the reseachers were after. They could have been trying to find the same answer from different people. I don't know. Duncan suggested there was technology in the pedestal and that they were trying to get somebody to sense or feel what the technology was.

Someone else involved in the project has said the horse was there to test the powers of observation of the recruits and that it also served as a point of reference.

The recruits were always asked if they saw anybody in the city. Each individual would interpret what he observed and report.

We know a lot of people were shoved somewhere into the future, maybe 200 or 300 years ahead. Estimates range from three to ten thousand people that were eventually abandoned. We have no idea for what purpose.

I have already said that I don't know exactly what they did with time. I wasn't there, but I do know they did a lot with World War I and World War II. They monitored those times and took pictures. They knew exactly what they were doing.

They could actually make up a secondary vortex to observe what was going on. We called this a seeing eye function. The original vortex was such that one could drive a truck through it. Using phase conjugation through the elaborate computer set-up, past and future history could actually be transmitted through the portal and viewed on television.

* Hyperspace is defined as space which exceeds the boundaries of three dimensions.

FOURTEEN- MISSION TO MARS

The project researchers continued to scout around in time. It was in late 1981 or '82 when the first actual use of this technology was employed to gain entrance into the underground areas in the big pyramid on the planet Mars.

As this material will be controlversial to much of the general public, I will try to give some background.

There is currently a video tape in circulation entitled "Hoagland's Mars".

This is a presentation to NASA scientists by scientific journalist Richard Hoagland concerning the tetrahedral complex that is associated with the "Face on Mars."

In this video, Hoagland shows the "face" and nearby pyramids that were photographed by the Viking spacecraft in the '70's. computerized projection techniques are used that give one a 360 degree "fly by" of the "face." The video also gives a close look at the pyramids.

Hoagland is trying to convice NASA to make more pictures of this region, which is known as Cydonia. NASA has been hard to convince and minimized the significance of Hoagland's work. In fact, a major effort was launched to prohibit the showing of this tape on public television stations. The story of this scandal was reported on by New York radio station WABC.

Why would NASA assume such a stance about a subject that is so intriguing?

The answer is perhaps explained by a book entitled, 'Alternative 3,' by Leslie Watkins with David Ambrose and Christopher Miles.* This book was based on a 1977 video that revealed a secret space program being run by an international conspiracy that included both the Russians and the United States. It is a fascinating account that includes astronauts breaking security, disappearing scientists, murder and the establishment of slave societies on the moon and the planet Mars. The book claims that men actually landed on Mars as early as 1962.

It is not my cause to prove that a colony exists or did exist on Mars. I have included this information so that the reader will understand that there is an entire scenario concerning Mars that is separate from my story. Those who are interested can investigate "Hoagland's Mars" or 'Alternative 3' for themselves. It is interesting to note, however, that the documentary entitled "Alternative 3" was shown on a San Francisco TV station sometime around the late '70's. A story has proliferated since that at the time the FCC threatened to revoke the station's broadcasting license if it were to be shown again. It wasn't shown again.

The directows of the Montauk Project knew there was a colony on Mars. It is more than likely that they were a part of the conspiracy.

Mars was interesting to the Montauk researchers, because they had realized that there was an old technology there. They knew somebody had built the pyramids and face on Mars. These were not natural formations.

According to the information that my associates and I have dug up, the people who were living on the surface of Mars could not get to the underground area beneath the pyramid. The entrances were either sealed over or simply couldn't be found. In fact, it appeared that the big pyramid was sealed better then the pyramid at Giza. Despite all the expensive and fancy technology that was available, the pyramid could not be penetrated.

The scientists at Montauk decided the best approach would be to project right into the center of the Martian underground. The newly discovered Montauk technology gave them the wherewithal to use a spatial warp to get inside. They wanted to get into the underground caverns. These were thought to be set up and administered by a very old civilization.

The time portal took the risk out of the operation as we could look through it.

We had a set up with TV monitors so that whatever Duncan would visualize would appear on the monitors. This provided a visual of present time on planet Mars. In order to find the underground area, we kept moving the opoen end of the vortex until a corridor appeared. At that point, we had Duncan solidify the portal. The away team was then able to walk from Montauk to Mars and be underground.

By this time, Duncan was no longer required to be continuously in the chair.

We had learned to have Duncan generate functions with the computer storing them and continually spitting them back. The computer could generally run the transmitter for a short time and had enough memory to modify the flow of time for about four hours. If Duncan would not return after that time, the thought forms being generated would drop out of reality. In such a case, the thought forms would have to be reconstructed from scratch.

The system definitely needed a live being initially. He would create the time portals and hold them open through concentration. Once the opening was made, we could record on tape what the live person was generating. The tape could then be used on its own to create another opening.

The system was continually refined and improved. If Duncan made a time connection once, it would then be recorded on tape. Because he sometimes had difficulty getting a connection, the tape made it easier and automatic. An entire library was eventually accumulated so that they didn't have to rely on Duncan. It was this development that enabled Duncan to be sent through the vortices himself. This occurred in '82 and '83. He was eventually selected for the team that went to Mars.

Using the time portals, Mars had been scoured for live inhabitants. Researchers had to push back about 125,000 years before they could find any. I don't know what they found out or what they did with the information. Duncan has tried to access this information, but it is deeply buried and difficult to contact.

My personal view is that the pyramid on Mars serves as an antenna. Perhaps there is technology inside of the pyramid. According to Duncan's recollections, he travelled to the inside of the pyramid. He saw technology being operated there and called it "The Solar System Defense." According to his account, the Montauk researchers wanted this shut off. It had to be shut off before anything else could be done.

This defense has been shut off retroactive to 1943, which is commonly considered amongst many UFO buffs to be the beginning of the massive UFO phenomena.

There's not much more I can say about Mars at this point except that the movie "Total Recall" is fancifully based upon some of the events that occurred with the Montauk Project. The way they used the chair in the movie is strinkingly similar.

Time research continued and countless missions were run until August 12, 1983. This was when the actual lock was made back to 1943 and 1963.

* 'Alternative 3' was originally published in the United Kingdom. The first printing in the U.S. was in 1979 by Avon Books, a Division of the Hearst Corporation, 959 Eighth Avenue, New York, New York, 10019.

FIFTEEN- ENCOUNTER WITH THE BEAST

On August 5th, 1983, we were given a directive to run the transmitter non-stop - just turn it on and let it go continuously. We followed the orders, but nothing out of the ordinary occurred until August 12th. Then, something very strange happened. All of a sudden, the equipment appeared to drop into synch with something else. We didn't know what function the system was now attuned to, but at that point, the USS Eldridge (the ship used for the Philadelphia Experiment) appeared through the portal. We had locked up with the Eldridge.

I'm not sure if this was a pure accident, but if the Montauk researchers were trying to hook up with the Eldridge, the attempt had to be made on this exact date. This is due to the 20 year biorhythms of the planet Earth (which was a discovery made in the process of these experiments) and the Eldridge experiment having occurred on August 12, 1943.

At this point, the Duncan from 1943 appeared and could be seen through the time portal along with his own brother. Both were crew members of the USS Eldridge.

We kept the Duncan of 1983 from seeing himself so as to avoid a time paradox and resultant negative effects.

The project had now reached apocalyptic proportions. Natural laws were being violated, and it seemed everyone involved felt uncomfortable. Three colleagues and myself had been privately voicing misgivings about the project over a period of months. We had talked about the pitfalls of dealing with time and how this might affect the karma of the planet. We hoped the project would truncate itself.

Consequently, our little cabal created a contingency program that only Duncan could activate. It was designed to crash the entire project.

We finally decided we'd had enough of the whole experiment. The contingency program was activated by someone approaching Duncan while he was in the chair and simply whispering, "The time is now."

At this moment, he let loose a monster from his subconscious, and the transmitter actually portrayed a hairy monster. It was big, hairy, hungry and nasty. But it didn't appear underground in the null point. It showed up somewhere on the base.

It would eat anything it could find, and it smashed everything in sight. Several different people saw it, but almost everyone described a different beast. It was either 9 feet tall or 30 feet tall, depending on who saw it. I personally believe it was about 9 or 10 feet in height. Fright does strange things to people, and no one was sure of what the exact physical constitution of this monster was. No one was in any frame of mind to calmly and collectively analyze its exact nature.

My supervisor had ordered us to shut off the generatorsin order to stop whatever type of phenomena was occurring. This didn't work, so it was decided that the thing had to be stopped.

It was decided that the transmitter had to be shut down. There were two efforts made in this direction. One was to send somebody back and turn off the transmitters on the Eldridge. They would be smashed if that was what was necessary to shut them down.

The other effort was by myself and the director of the project. We unsuccessfully attempted to shut the transmitter at Montauk. We then went into the power station and disconnected the base from the Long Island Lighting Company. The power kept going and nothing stopped.

We weren't concerned about the lights. We just wanted to stop the transmitter itself. We decided the next best thing to do was to go into the power station and cut the wires leading into the ground from the big transformers. I put an acetylene torch on my back and cut the wires going into the ground. I had to be careful because they were hot. Still, nothing happened. The lights at the base stayed on.

I figured there must have been another power feed somewhere. We went over to the transformer farm nex to the transmitter building and cut the wires coming up out of the ground. At that point, the lights at the base went out and the computer stopped. But, the lights in the transmitter building stayed on!

We went into the building and pulled the wires out of the panel that controlled the transmitter. Then wires from the transmitter itself. The lights in the building went off, but the transmitter stayed on.

I then went upstairs and cut the actual equipment apart. I cut the conduits. I cut the cabinets. Finally, I cut enough apart that the transmitter just groaned and stopped. All the lights went off. We'd done it. Today, you can still see the torch marks where I'd cut things apart.

It was at this point that the beast stopped moving and faded back into the ether.

The portal closed and that was the end of that episode.

After we stopped the transmitter and things settled, we figured out what had happened. When we had first thrown the switches in the power station, none of the lights went off at the base. There was no power coming into the base. When I cut the lines going to the transmitter building, the rest of the base went off, including the computers. However, the transmitters ran without computers.

The system had actually gone into a free energy mode. The two systems (i. e.

the two generators - one in 1943 aboard the Eldridge, one at Montauk in 1983) were locked together. There was a tremendous amount of energy bouncing between the generators. With so much energy between them, all the electrical circuits that were connected remained active. The lights stayed on.

More importantly, the generators established a connection from 1983 to 1943.

By bouncing energy between the two time periods, a stable vortex was created.

This served as an anchor. Using this vortex, a time tunnel could then be projected to a specific point in time.

For example, if one wanted to go from 1983 to 1993, the '83 to '43 vortex would first have to be functioning to serve as the anchor. The projection to '93 (or whatever other point in time chosen) would come out the '83 end of the vortex.

If one wanted to go to 1923, one would project through the 1943 end of the vortex. Times between 1943 and 1983 could be reached by going through either end of the vortex. Dates after 1963 were accessed through 1983 and dates prior to 1963 via 1943.

This is not to say that all time travel would have to be done in this manner (using the master vortex from 1943 to 1983). During these experiments, no generators were found in either the past or the future that could link up and establish a vortex of this nature. There are, of course, plenty of generators around, but a successful link had to be made. That link required a "witness effect."

"Witness" is an occult term. As a noun, it refers to an object that is connected or related to someone or something. For example, a lock of someone's hair or a picture could serve as a witness. As a verb, "witness" means to use an object to enter a person's consciousness or otherwise have an effect on them.

One example of a "witness effect" would be for someone to take a lock of hair, use it with a love potion, and have the owner of the hair fall in love.

With the Montauk Project, there were three "witness effects." They could be considered as three different levels of witnessing.

The first level consisted of physical people who were actually on the USS Eldridge. Any surviving crew members they could find were brought to Montauk for the experiment in 1983. This also included personnel who were considered to be reincarnated since the Philadelphia Experiment. Duncan and Al Bielek were both there and were two of the primary witnesses.

The second level of witnessing concerned technology. The zero time reference generator (referred to previously as the whirligig) used aboard the Eldridge was also used at Montauk. When the Eldridge was eventually decommissioned in 1946, the whirligig was placed in storage. It was eventually brought out to Montauk and incorporated with the system there. In addition to the whirligig, there were two very strange radio transceivers* linking the two projects. They were "crosstime" transmitters. They could transmit across time, and they used that to lock up the two projects.

The third level of witnessing was the planetary biorhythm. the term "biorhythm" is esoteric and refers to the higher order channels that regulate life in an organism.

Biorhythms are a result of the resonance upon which nature operates. In humans, the processes of sleeping and eating would involve biorhythms. Of course, there are many subtle ones that could be studied, ad infinitum. When viewing the Earth as an organism, there are also biorhythms. The seasons and the daily spin of the planet would involved biorhythms. The scientists at Montauk exhaustively studied the biorhythms of the Eart and how they related to the entire universe. They discovered that there is a major planetary biorhythm that peaks out every twenty years.

The Philadelphia Experiment occurred in 1943. Although 1983 was forty years later, it was a multiple of twenty and served as a potent witness. It enabled the two projects to link up. I should also mention that it is entirely possible that the link could have been made without the witness effect, however its application proved very helpful to the project.

The reader should now have some idea of the general theories and applications that were used at Montauk.

After the bizarre occurrences of August 12, 1983, the Montauk base virtually emptied. The power was restored, but lights were left on with everything in disarray. Most of the personnel were eventually rounded up, debriefed and brainwashed accordingly.

* I was able to acquire a few of the transceivers that were used in the experiments.

Up to this date, I do not fully understand them or their function. It is impossible to get any literature or manuals on the subject. The only possible way to get information on these transceivers is to ask people who have used them. The only answers I've gotten thus far is that they were a highly classified piece of equipment. People I spoke to knew they were for stealth airplanes, but they didn't know what exactly they were for.

SIXTEEN- THE NATURE OF TIME

This book will give rise to many questions, particularly about the nature of time itself. From my experience in talking to groups, I will try to clarify some points that often confuse people.

First of all, the past and the future can be changed.

It will help to consider the idea of a chessboard. In chess, there might be thirty moves in a game. Each one of these moves will create a different layout of the chess board. If one were to "go back in time" and change a move that has been made, it would consequently change all the other board layouts subsequent to that change.

Time could be considered a hypnotic pulse that we all subconsciously agree or submit to. When someone is able to manipulate a change in time, they are also manipulating our subconscious considerations and experiences. Therefore, if time is changed, one wouldn't necessarily realize it.

This scenario implies that we are merely pieces on a grand chess board. To a degree, this is true. For example, retired generals often complain of having been the pawns of international bankers. It is a farfetched comment, but perhaps there would be no war if generals could be truly clued in on the real machinations behind international politics.

There is also the example of Homer's 'Iliad', which tells the story of the Trojan War. According to that legend, the gods literally manipulated the characters on Earth like a chess board. The story is ripe with intrigue between mortals and gods. The plots become so intricate and thick that it sometimes seems Homer is trying to provide us witha microcosmic view of the entire universe.

Whatever the case, we are all players in the game that is known as "time." The obvious way to protect one's interest is, of course, to gain knowledge about time itself. Whether one wants to do this by pursuing meditation or astrophysics is an entirely personal matter.

At Montauk, the scientists also viewed the future. The viewers they had gave them the ability to look at multiple futures. Once they chose a particular scenario and activated it by someone or something travelling to it, that future would become fixed. That point would be locked to the time from whence the connection was made.

It would create a loop that was fixed.

For example, let's say multiple futures were viewed with different people becoming president. Suppose the future with "Sam Jones" as president was chosen by the researchers, for whatever reason. Linking a person or item from the present would lock in the President Sam Jones scenario no matter what. However, none of this means that a fixed point scenario couldn't be further changed by the scientists doing more manipulations.

At this writing, we arecurrently in a time loop. This loop extends from where the Montauk researchers penetrated into the past up to where they penetrated into the future. It's fixed and would appear unalterable. However, this does not mean that we are all relegated to being hopeless slaves of time manipulators. The subconscious has its automatic or hypnotic levels, but it also contains the seeds of freedom: dreams. If one can dream something, it can be brought into being.

It is very easy to get philosophical about all of this and get lost in that process. My point with this book is that there has been manipulation of time. This has also exploited individuals and caused untold misery. It could easily be considered to be the work of dark forces.

There is still one major question. Who was really behind the Montauk Project?

There are countless intrigues and scenarios one can envision. Religionists can bring in God and the Devil. UFO aficionados can offer a grand scheme of aliens vying for our solar system. Left wingers will offer explanations concerning the CIA and secret government.

I believe that all of the above can shed light on what actually happened at Montauk. It is also my hope that this book will bring more people out of the woodwork. Thus, we can have more answers and less mystery.

SEVENTEEN- THE MONTAUK BASE IS SEALED

After the events of August 12, 1983, the Montauk Air Force Base was abandoned.

By the end of that year, there was no knowledge of anyone being on the base.

In May or June of 1984, a crack squad of Black Berets were sent to the base.

I believe they were Marines, but I'm not absolutely sure. They were reportedly ordered to shoot anything that moved. Their purpose was to purge anyone who might be on the base.

There was a second team that followed the Black Berets. They removed secret equipment which was considered too sensitive to leave behind.

The next step was to prepare the underground to be sealed. Certain incriminating evidence was removed at this point. I've heard that a room with hundreds of skeletons was cleared out during this evolution.

About six months later, a caravan of cement mixers appeared on the base. Many people saw these trucks. They filled the vast underground areas of Montauk with cement. This included dumping cement down the elevator shafts as well.

The gates were locked up and the base was abandoned for good.

EIGHTEEN- MONTAUK TODAY

If one travels to Montauk Point today and parks in the state parking lot near the light house, it is possible to get a good view of the giant radar reflector that sits atop the transmitter building.

For those who are either brave or foolish, one can follow the dirt roads that lead to the base. Most of the entrance gates have been bent or otherwise vandalized so that entrance is easy. This was probably done by local juveniles who sometimes get drunk and have beer parties on the base. However, walking on the base is prohibited by New York State park rangers who periodically patrol the area.* There are also occupied buildings on the main roads to the base.

It should be noted that I am not writing this information to lure people to the base. People are going to be curious after reading this book, and it is my responsibility to warn them. I'm not absolutely sure of the legal techinicalities, but walking on the base is probably illegal. One goes at one's own risk.

There are also other dangers to consider.

Two people I know who participated in the Montauk Project visited the area in the late '80's. They claimed they were abducted and do not totally remember what happened to them.

Another person has reported in August of 1991 that video cameras can now be seen from the top of the transmitter building. This is a new development and is rather odd considering it is a vacant and derelict facility.

There are also reports that the underground areas of the base are being reopened.

This is speculative but should serve to warn anyone making travel plans to Montauk.

* The entirety of Fort Hero, including the inner Montauk Base has since been donated to New York State as a park. While there are peculiar political arrangements concerning the base to this day, the rangers are not out of bounds in keeping people off the grounds. The buildings are in a state of disrepair and are potentially dangerous to those going on a casual walk.

NINTEEN- VON NEUMANN ALIVE!

After completing the first draft of this book, a new development took place.

It concerns occurrences that started years ago but only came to a resolution recently.

This concerns John von Neumann and corroborates the theory that he didn't die in 1958 as is generally believed.

In 1983, I was contacted by a friend of mine in upstate New York who I will refer to as Klark. He knew I was interested in communicatoins equipment and told me about an old time suplus dealer who I will call Dr. Rinehart.* Rinehart was a legend in the local surplus community.

Klark said that the man had a collection of equipment that went back as far as the 1930's and 1940's. A meeting was arranged with Dr. Rinehart on the pretext that I was interested in buying him out. Klark introduced me, and Rinhart showed me his collection on a silver platter. He sincerely wanted to sell out, but I thought his price was too high. A great deal of the equipment was junk and would have cost just as much to cart away.

I considered his prices exorbitant and thought he might be a bit crazy. Apparently, he went even crazier after he met me. Klark visited him again on his own and was met at the door with a shot gun. Rinehart pointed the gun and told Klark that he didn't want to see that bastard Preston on his property. He didn't want to see Klark, Preston or any of their friends on the property. He said he'd shoot them if they showed up.

Klark tried to calm him down and asked what this was all about. He had no idea why the man was so upset. Rinehart said that Preston had come back and robbed him the night he was last there.

As it turned out, somebody had come to the guy's house, strapped him to the chair, ransacked the house and stole money. It certainly wasn't me, and Klark and I were both confused. Years went by, and I had dismissed the puzzling circumstances concernign Dr. Rinehart.

As my memory of the Montauk Project returned, I suddenly recognized Dr. Rinehart.

He was actually John Eric von Neumann, the brains behind the Philadelphia Experiment and Montauk Project!

Many years back, perhaps as early as 1958, von Neumann had been assigned to a "witness relocation" program. He was given a new identity as Dr. Rinehart and slipped into a new role as a surplus dealer upstate. He also remained on call to the authorities that ran the Phoenix and Montauk Projects and would work for them whenevery he was needed. Sometimes this was for months at a time.

This man not only looked like von Neumann, his doctorates in mathematics and physics were on the wall and came from Germany. Despite this, he claimed never to have left the United States.

It was also apparent that this gentleman's faculties and memory had been tampered with.

I had talked this over with Al Bielek, and we figured that my presence at von Neumann's house was too much for him. He would have remembered me from Montauk and that probably frightened him and caused him to flip out.

All of this is fascinating in its own right, but my main interest was in a strange looking receiver that he had. It is known as an FRR 24 Receiver. I had noticed it on my origianl visit, and it was still there. I wasn't about to return to his place in view of his threats, but I'd sent people up there and they indicated the receiver was still there.

Al remembered von Neumann as well and wanted to visit him. In fact, von Neumann as Dr. Rinehart had taken a liking to Al. Hoping to get my hands on his receiver, I drove Al upstate to Rinehart's house.

We weren't sure how to approach him on the receiver. We thought about having me wear a disguise but thought it might be easier to have Al buy the receivers on my behalf.

Al got out of the car and greeted him. I remained in the car in hopes that he would ignore me. It started to rain, so Rinehart told Al that they should go to the trailer on the other side of the property. That's where the equipment was. Rinehart walked right by my car and looked me straight in the face. He was friendly and said I should come along, too. Apparently, Rinehart didn't recognize me. I followed them to the trailer as if nothing had ever happened between us.

Al got the guy talking, and I just listened. Von Neumann didn't come through.

He was strictly in the identity of "Dr. Rinehart" as he spoke to us.

When he stopped talking, I told Rinehart that I'd heard he had a very large receiver setup where each reciever fits in a rack by itself.

He said, "Oh, that thing! I was gonna keep that. But hell, I'm never gonna use it. I can't even move it. I'm gonna keep it, or I'm gonna sell it."

I asked him how much he wanted for it, and he said he would give it to me for a thousand dollars. I told him that Al and I couldn't afford that kind of money, so he suggested a trade.

Al told me to make an offer so I offered $600 for four racks of the receivers.

He said that was a little less than he wanted and that he'd have to think it over.

We left on good terms and returned home.

Another meeting was arranged some time later. He said that he wanted hi-fi equipment and would be willing to work out a trade. We dug up some hi-fi material and went up again. He looked at it and practically got tears in his eyes. He was excited to see the material and remembered the people who actually designed much of it.

He apologized and said that he really couldn't use any of the material. He wanted cash. If we sold the material, he said we could come back and get the receivers for cash.

We carted everything back to Long Island again. I was frustrated but wasn't going to give up. I called around and found out that I could sell the material.

It was worth $750 to other dealers, and I sold it immediately.

I wanted to get his receivers quickly as he was getting known again by national collectors. They would snap up the receivers if I didn't act soon.

I took $800 and went up to see Dr. Rinehart again. I'd taken along some friends to help me move the equipment. Fortunately, it was a clear day, and the weather wasn't going to interfere with our plans.

Dr. Rinehart came out and was in a friendly mood. I showed him $750, but he said he didn't want any money until he was sure I was satisfied with the receivers.

He showed us around the place. We went to look at the receivers, and I was surprised.

He had four racks of the equipment, and I had remembered only one. He was willing to let them all go for my offer of $750 which was more than fair on his part.

I found him to be quite pleasant. In fact, I was a bit puzzled. Initially, he had wanted $1,200 per rack which meant $4,800 for the entire system. Now, it was seven years later, and he was accepting $750. It is my opinion that he wanted me to have the receivers for some reason. I still don't know exactly what it was.

As I poked around and looked at the receivers, my two friends went to the chicken coop as they were interested in some Western Electric equipment that was stored there. Dr. Rinehart was sitting in a chair not far from the receivers. Suddenly, I noticed that he was no longer Rinehart. He was John von Neumann! He remembered his true identity and began to talk.

He definitely remembered me and told me things that were of a sensitive nature which I am obligated not to repeat. He also said that over the years he had seen that millions of dollars were put away in secret Swiss bank accounts. This money was to be used to compensate many of the workers at Montauk who had suffered as a result of the project. Apparently, when I had visited him years back, some sort of signal alerted the secret group that backed the Montauk Project. He was bound up and robbed the next night, and his secret bank books were missing. He now realized that I had not been involved.

I wasn't able to start moving the receivers out until the next day. It was a big job. I took the receivers out of their racks and broke them up so they could be moved safely. Rinehart was there, took, and he started to fade in and out. First, he was Rinehart, then he'd be von Neumann. It was like a yo-yo. Finally, he settled on von Neumann.

As von Neumann, he said that he had obtained these receivers for a very good reason. They were actually capable of tuning in on either of the two projects: Project Rainbow (the Philadelphia Experiment) or the Montauk Project. Further, the receivers were capable of tuning in on the project from any other space and time in our universe. He also believed this receiver was the main witness from Montauk to the USS Eldridge. He said it could pick up the pattern of the Eldridge back in 1943.

It seemed that von Neumann had completed what he had to say. Rinehart returned, and I loaded up the receivers to take back to Long Island.

I wasn't sure how the receivers worked or what they were all about. My first step was to ask Duncan to do a psychic reading. He indicated that the receiver was capable of tuning to any particular point in time by way of zero time. He said that if we could figure out how to tune it, we could tune in to any other point in time.

We realized what von Neumann had already told me: this equipment was a key part of the Montauk time machine. I don't think the particular equipment that I had was on either the Eldridge or at Montauk. I think it had been used at the Philadelphia Navy Yard in the 1940's.

I wanted to trace this equipment further to see if it had a logical point of origin. I called the biggest old time radio surplus dealer in the country. They had never heard of the FRR 24 Receiver. I talked to lots of friends in the surplus business and found only one person who had ever seen one or heard of one. This person said that the receiver came out of RCA. He owned a piece of one of the receivers at one time. It had gone out of his hands when an old man from upstate New York came and paid an exorbitant price for the piece he had of the receiver.

Tracing this back to Dr. Rinehart, he verified that he was the purchaser of the equipment. But, he said that this only accounted for pieces from two of the racks he'd sole to me. There were four in total, and he had to buy the other two racks of receivers from somebody else. I tracked the other person down with Dr. Rinehart's help. This person was a young man who also said the FRR 24 Receiver had come from RCA.

I decided to find out how many of these receivers had actually been released.

I called up the Surplus Disposal Agency, gave them the number for the receiver, and they did a computer read out. A lady at the agency said that only three FRR 24's had ever been released. All the other systems were either still in use or had been destroyed.

Then, she indicated that until recently, this receiver had been classified. She said that if any of them had been scrapped, the manuals for them had to be destroyed.

There was also a note indicating that each FRR 24 unit contained seventy-five pounds of silver. The units had reportedly been scrapped and sold to dealers for silver recovery. When scrapped, they are not useful as they would have been put through a crusher.

The report indicated that FRR 24's were only released when the government agreed to sell them to a world communications company. There were three such instances listed. One FRR 24 went to RCA, one went to ITT on the west coast and another to Vero Beach, Florida.

I tried to trace down people who had actually worked with the FRR 24. Finally, I located a retired gentleman who had worked at RCA Rocky Point (on the eastern end of Long Island). He had worked at the receiver station at Rocky Point.

The gentleman indicated that the FRR 24's had been at the RCA receiving station for years. He raved about the receivers and said they were beautiful and fantastic.

When they were turned on, however, he said that a very strange type of interference was picked up all across Long Island Sound. It was a mystery and neither he nor anyone else could figure it out. He also mentioned that the receivers made strange audio noises and that RCA finally decided not to use them.

This was interesting because von Neumann told me that two receiver racks from RCA Rocky Point had been sent back to the 1930's. One ended up in the Philadelphia Naval Yard and was used to track the Rainbow Project in 1943. The other receiver rack ended up at RCA for disassembly and study so that it could be replicated and applied to the technology of the day.

It is interesting to note that in the 1930's, RCA made tremendous strides in radio technology. The years 1933 and 1934 were particularly ripe with new discoveries.

If von Neumann was right, RCA received and analyzed a rack of receivers from the future. It is likely that von Neumann would have sent them back himself.

The receiver rack that ended up at the Philadelphia Naval Yard eventually came into my possession, and I still have it. The disassembled rack was enhanced and improved upon by RCA, and these are what ended up at Rocky Point. This was accomplished through a time loop, thus there are some differences in the RCA receivers (the FRR 24's that I recently got from von Neumann) and the one used during the Philadelphia Experiment. Both receivers, however, have more similarities than differences.

In addition to the above, I also have equipment developed by Tesla that appears to be inspired by the FRR 24's that were sent back from the future by von Neumann.

Whatever the case, radio technology was pushed way ahead in the 1930's. As an engineer and professional radio man, it is my personal conclusion that it couldn't have been done without some major help from somewhere. For example, Nikola Tesla was always upsetting the status quo by saying that he experienced communication with aliens.

There is another major point of interest concerning the FRR 24 Receiver. When I purchased them from Rinehart, I noticed that the aluminum housing was corroded on the outside. Aluminum itself does not corrode, but it does when it is mixed with impurities. The aluminum panel in the chassis, however, had no such corrosion. The chassis were therefore made out of a very pur aluminum. Commercial grade aluminum used in radio equipment is usually not that pure.

What does all this tell us?

There had to be a reason that the aluminum was so pure. Recently, it has come to the surface in conventional scientific circles that aluminum can be made into a superconductor. A friend of mine at NASA told me that mixing mercury with aluminum and alcohol makes microchannels big enough for electrons to channel through the aluminum. This essentially creates a room temperature superconductor.

Rinehart also warned me about the chassis. He said there might be a little bit of mercury contamination on the chassis. Further inspection has revealed that the chassis had some sort of mercury treatment process.

It is currently my belief that the treatment process is related to the silver resonators, which are the capacitors and coils. The mercury and aluminum create microfine channels along the superconducting chassis and the channels become a multi-dimensional resonator.

To conclude, this receiver is actually a multi-dimensional space-time resonator and would be part and parcel of any time machine that was used for the Philadelphia Experiment or at Montauk.

* Dr. Rinehart is a pseudonym used to protect this individual's privacy.

APPENDIX A- A SCIENTIFIC ANALYSIS OF THE RADIOSONDE

Note: This analysis is not expected to be readily understood by the general lay public. It is included only for those who are technically inclined. It also serves as corroboration of my statement that the government had the means to affect the weather).

The Radiosonde consists of two variable resistance type sensors. One registers temperature, the other humidity.

The temperature sensor is a thermistor, where the electrical resistance varies inversely with the temperature. The humidity sensor is an electrolytic resistor where the electrical resistance varies directly with the relative humidity. In most of the Radiosondes, the pressure sensor is of the pressure responsive selctor switch type (Baro switch). In essence, the transmitter sees a varying resistance which is selected alternately by the Baro switch or a sequential switcher. A short circuit is selected occasionally which is called the reference mode. This is what the sensors do on the surface. It is also the line the government releases to the public. Although cursory investigation will show this description to be true, there is also other activity which is secret. The temperature sensor is a carbon bar with precious metals added, and acts as an antenna to the DOR function. It also inverts the transform known as DOR energetics. This item is packed in a small vial, and has to be installed on fasten stock clips on the arms of the Radiosonde. In order to get a true temperature reading, it is painted white to reflect the sun's radiant heat and sits above the package in the open. This placement is understandable from conventional science, but it cannot be understood from the viewpoint of relativistic sciences.

The humidity sensor is an electrolytic resistor. We do not understand its operation because the usual electrolytic resistor varies inversely with the relative humidity. This humidity sensor consists of a grid of conductive lines with an unknown chemical overlay. It acts as an antenna for orgone in phase. It is also similar to the electrolytic detectors that have been introduced for the detection of esoteric energies. The humidity sensor is also hermetically sealed in a small vial and has to be put in its holder on top of the Radiosonde, thus totally covering and protecting it from direct rain, but allowing air to circulate around it. This follows the released line of information.

In later Radiosondes, the Baro switch is replaced with a clock-work driven scanning switch, with the addition of a receiver that the Government claims is used as a transponder to track direction and height. This suggests that the Baro switch gives height information, which can be read from the pressure, but depends on a uniform pressure gradient which our atmosphere does not exhibit. This follows the released information, but is grossly inaccurate.

I do not believe that this is the real objective for the Baro switch. In fact, an entirely different purpose is intimated. It appears that the Baro switch is the correlated function which would be necessary to synchronize the DOR busting to the environment of the Earth. It is also apparent that the receiver synchronizes the DOR busting to the the environment. At this point, I do not fully understand the sensor scheme.

In addition to the sensors, the other part of the radiosonde is the transmitter.

It is pulst time modulated and the repetition rate of the pulse varies with the resistance presented to the transmitter. There are two types of pulse modulation used.

One is where the modulation pulses off the CW (CW=carrier wave) carrier oscillator.

The other is where a high voltage pulse supplies B+ (B+ stands for "B batteries", which refer to plate voltage) to the carrier oscillator. There are two frequencies used: 400 MHz and 1680 MHz. The 400 MHz oscillator is comprised of tuned lines with the trioed tube in the field of the lines. The 1680 MHz oscillator is of the integral cavity type, with the triode tube inside the fields of the cavity.

In the down pulsed CW oscillator transmitter, there are two sections: the modulation oscillator and the carrier oscillator. The modulation oscillator is what generates the pulse, which is a triode oscillator with a blocking network in the grid circuit.

The operation of the transmitter is simple. When the oscillator runs, the grid bias builds up across the cap (C) and when the voltage reaches the tube cutoff the oscillation stops. At this point, C discharges until the tube starts up again.

As the grid voltage across C cycles up and down, the oscillator starts and stops; hence modulating the voltage drop across Rp, which is bypassed. The value of C and Rg and Rext and Rref determines the repetition rate of the pulse across Rp.

The signal with the pulse and its potentials are capacitively coupled to the carrier oscillator.

This is the usual explanation, but let us consider the relativistic activity.

When the tube is cut off, the higher order signal builds up inside the tube like the charge on a capacitator. The longer the tube is cut off, the more relativistic signal builds up in the tube. When the tube turns on and oscillates, it slew between saturation and cutoff, and two things happen. First, the stored relativistic charge is forced out. Second, the slewing of the oscillation between saturation and cutoff has the effect of amplifying the hgiher order components by "0" point activity.

The result is that the signal is amplified and outputs in pulses. From this point, the modulation signal is capacitively coupled to the grid of the carrier oscillator, where the pulse stops the oscillation.

When we consider the carrier oscillator, the circuit is a standard one. How it was optimized probably has something to do with the placement of the tube in the field of the resonant network and the design of the tube. The higher order operation in the carrier oscillator is similar to the modulation oscillator. When the tube slews from saturation to cutoff, the 0 point of the vacuum is ripped. This results in relativistic gain as well as forcing all signals stored in the tube to the output and the antenna.

The bypassed output of the modulation oscillator, which is pulsed potential (scalar) at approzimately 7 MHz is coupled to the grid of the carrier oscillator and slews the Q point * from saturation to cutoff. The "0" point activity sends out bursts of relativistic signal which replicates closely the input signal from the sensors.

The pulse modulator transmitter uses a delay line pulse modulator with a thyratron, charging reactor, blocking diode, pulse forming network, and a pulse transformer that generates 1400 V pulses which drives the carrier oscillator. The thyratron is fired off by the output of the same modulation oscillator as in all of the rest.

The modulation oscillator loads the relativistic signal into the delay line through the thyratron which is cut off but still has "0" point gain. When the thyratron fires, everything in the pulse forming network is loaded into the carrier oscillator tube as a 1400 V pulse which rips apart the vacuum and results in a high "relativistic gain" through the usual "0" point activity.

The whole package is operated by a battery pack which lasts approximately three hours.

* "Q point" means quiescent point. This refers to the point where the tube rests.

APPENDIX B- WILHELM REICH

Although the government had great appreciation for the discoveries and developments of Wilhelm Reich, it appears they had very little use for him personally. He was under pressure from the AMA and FDA for years on charges of quackery. Eventually, he was given a very stiff prison sentence for contempt of court when he refused to appear. The subsequent burning of his books and destruction of his equipment may be unparalleled in modern times for its outrageousness.

His claims about jamming the drives of UFOs didn't win him too many friends either. He concluded that the drives of UFOs ran on cosmic or orgone energy. He developed a "space gun" on the basis of orgone theory and was able to make UFOs fade away with regularity according to eyewitness accounts.

After he was in prison, authorities reportedly gave him express permission to work on anti-gravity equations. This is odd to say the least; especially if they believed he was a quack.

Whatever the exact facts and details of Wilhelm Reich are, it appears that he was used for his inventive genius and then put away so he couldn't disseminate his discoveries elsewhere. The systematic condemnation of his work only backs this up.

APPENDIX C- MIND CONTROL AND THE PERSIAN GULF WAR I was still laid off when the Persian Gulf War erupted and had the opportunity to watch the live briefings. Live briefings are interesting to me because information sometimes comes through that would otherwise be edited out.

In one report, a CNN reporter said that he had just returned from Kuwait where he had travelled with an American patrol. They had noticed a patrol of about thirty Iraqis on the next sand dune. While the Americans were wondering how they would get the Iraqis to surrender, a U. S. helicopter suddenly appeared and flew over the Iraqis. By the time the helicopter had reached the next sand dune, the Iraqis had their hands up and were surrendering.

This is all very suspicious in itself. These were the same Iraqis who fought a Holy War against Iran for eight years.

The next news of interest that I noted came towards the end of the conflict when Brigadier General Neil was questioned by a British reporter from the BBC. The reporter asked the General about his plans to get the Iraqi soldiers out of the deep bunkers that the Germans built for the Iraqis. These bunkers were known to be extremely well fortified, and it was a good question.

General Neil said, "We bring in the psychological..."*

He then broke his sentence with coughing. It didn't sound like a real cough but that he had caught himself saying something he shouldn't reveal. When he stopped coughing, he continued speaking.

"I'm sorry, we bring in the helicopters with PA (public address) systems and we talk 'em out."

To me, his statements were very significant. It is apparent to me that the General had made a mistake and had to continue his sentence in the same vein. In my opinion, he was going to say something like "psychological broadcasting helicopters."

He had helicopters on his mind and in order to make it less obvious, I believe he changed what he was going to say and referred to "PA system helicopters."

I did some research into the Iraqi bunkers and discovered that the Americans had been trying to get the plans for their construction from the East Germans.

The Americans wanted to know how to penetrate the bunkers. They did retrieve the plans and found out the walls were very thick. Even after the air blitz, the Iraqis were still deep in the bunkers. They had electricity, entertainment and enough food and water to last at least six months. The bunkers were three feet thick and probably could have withstood a nuclear blast. The Iraqis also had the equipment to tunnel out if necessary.

The British reporter knew it wouldn't be easy to get the Iraqis out of the bunkers. That's why he asked the question. I believe it is absurd to suggest that these fanatic soldiers would have surrendered under the mere threat of PA system helicopters.

* The quotations by General Neil are not exact quotations but are paraphrases based upon my memory of the actual events.

APPENDIX D- NIKOLA TESLA

Nikola Tesla was born in 1856 in what is today known as Yugoslavia. Known as the "Father of Radio", he was clairvoyant and had different paranormal abilities.

Most notable was his vision as a youngster that he would build an alternating current generator that would revolutionize the way that mankind harnessed electricity.

Tesla received a renaissance education and learned to speak several languages.

He worked his way across Europe as an inventor and electronics engineer. In Paris, his genius came to the attention of one of Thomas Edison's associates, and Tesla was invited to meet the famous inventor. Although Edison hired him, the two never got along well.

Edison's utilities used direct current which required a power plant every few miles or so. Tesla tried to convince him that alternating current was more effective and less costly to operate. Edison was stubborn and Tesla's brilliance must have made him feel insecure. Here was a man whose geius was far superior to that of Edison's!

Edison would never back up Tesla's plans to revolutionize the world with alternating current. The two finally had a complete falling out when Tesla advised Edison that he could upgrade his entire facility by building new machines and replacing the old ones. Edison offered him $50,000 to complete the task. Tesla designed twenty-four types of machines and effectively enhanced the factory. Edison was very impressed but wouldn't pay the money. He claimed it was just his "American sense of humor."

George Westinghouse was an inventor himself and recognized the genius in Tesla.

He backed Tesla's plan to harness alternating current from Niagara Falls and the world has never been the same since. Meanwhile, Edison tried to prove that alternating current could be deadly to humans and went to the extent of electrocuting a dog in public (with alternating current) to prove his point. Edison ended up embarrassed and humiliated.

Tesla's career went on the rise and his experiments were of vast renown. He demonstrated remote control with small boats in Madison Square Garden, but many people dismissed it as witchcraft.

He even generated lighting between the Earth and sky at Colorado Springs. This experiment was particularly remarkable because he put light bulbs to the raw ground and they light up. This demonstrated that the Earth's surface was a conductor of electricity. This proved that if the proper vehicles were used that the entire population of Earth could enjoy free energy.

Tesla created a huge tower on Long Island and sought to build a system that would provide free energy. While in progress the financier J. P. Morgan pulled the rug out from under Tesla. He didn't want free energy.

Tesla's career went into a decline and his reputation was hurt. Part of this was caused by his periodic admisssions that he received communication from aliens.

His receivers supposedly picked up transmissions from Mars.

No one ever denied he was an electronics genius, but because he understood supernatural phenomena, he was held in suspicion. Today, many of my engineering colleagues consider him a "nut" who just happened to be brilliant in electronics. This is a very convenient explanation.

It is my opinion that he was incredibly ahead of his time.

APPENDIX E- HISTORY OF THE PHILADELPHIA EXPERIMENT AND ITS RECONCILIATION WITH THE MONTAUK PROJECT In 1912, a mathematician named David Hilbert developed several different methods of new math. One of these was known as "Hilbert Space." With this he developed equations for mulitple realities and multiple spaces. He met Dr. John von Neumann in 1926 and shared his information. Von Neumann took a lot of the systems he learned from Hilbert and ran with it. According to Einstein, von Neumann was the most brilliant of mathematicians. He had an uncanny ability to take abstract theoretical concepts in math and apply them to physical situations. Von Neumann developed all kinds of new systems and math.

A Dr. Levinson had come along and developed the "Levinson Time Equations."

He published three books, which are now very obscure and almost impossible to find.

An associate of mine did dig up two of them at Princeton's Institute for Advanced Study. All of this work was to serve as a background for the invisibility project which would apply the theoretical principles to a large hard object.

Serious research into the subject of invisibility began in earnest in the early 1930's at the University of Chicago. Dr. John Hutchinson Sr. served as Dean at this particular time and was privy to the work of Dr. Kurtenhauer, an Austrian physicist then at the University. They were later joined by Nikola Tesla. Together, they studied the nature of relativity and invisibility.

In 1933, the Institute for Advanced Study was formed at Princeton University.

This included Albert Einstein and John von Neumann, a brilliant mathematician and scientist. The invisibility project was transferred to Princeton shortly thereafter.

In 1936, the project was expanded and Tesla was made the director of the group.

With Tesla on board, partial invisibility was achieved before the end of the year.

Research went on to 1940 when a full test was done in the Brooklyn Naval Yard.

It was a small test, with no one on board the vehicle. The ship used was powered by generators from other ships, connected by cables.

Another scientist, T. Townsend Brown, became involved at this point. He was known for his practical ability to apply theoretical physics. Brown had a background in gravity and magnetic mines. He had developed counter measures to the mines with a technique known as degaussing. This would trip the mines at a safe distance.

There was a big brain drain on Europe in the 1930's. Many Jewish and Nazi scientists were smuggled into the country. Much of this influx has been attributed to A. Duncan Cameron Sr. Although we know he had extensive connections, his exact relationship to intelligence circles is still a mystery.

By 1941, Tesla had full confidence of the powers that be (FDR). A ship was prcured on his behalf, and he had coils wrapped around the entire ship. His famous Tesla coils were also employed on the ship. However, he grew wary because as the project developed, he knew there would be problems with personnel. Perhaps he knew this due to his ability to fully visualize his inventions in his mind. In any case, Tesla knew that the mental state and bodies of the crew would be affected severely.

He wanted more time to perfect the experiment.

Von Neumann desagreed with this vehemently at the time and the two never got along. Von Neumann was a brilliant scientist but did not embrace metaphysics for its own sake. Metaphysics was old hat to Tesla, and he had built a successful legacy of inventions based upon his unique prescience.

Part of what made his views so controversial was that during his experiments in Colorado Springs, circa 1900, he said that off planet intelligence had contacted him via consistent signal messages when Mars approached. This also occurred in 1926 when he had radio towers erected in the Waldorf Astoria and at his New York city lab.

He claimed to receive information that he'd lose people if things were not changed.

He needed time to design new equipment.

Tesla's requests for more time were not heeded. The government had a war to win and additional time was not granted. Tesla wen through the motions but secretly sabotaged the operation in March 1942. He was either fired or quit. He is supposed to have died in 1943, but there is arguable evidence to suggest he was whicked off to England. A look-alike derelict is supposed to have been put in his place for the funeral. He was cremated the day after his body was found which was not in keeping with the tradition of his family's orthodox faith. Whether or not he died is controversial. That secret papers were removed from his safe has never been in question.

Von Neumann was named director of the project. He did a study and determined that two huge generators would be required for the experiment. The keel for the USS Eldridge was laid in July 1942. Tests were done at dry dock. Then, in late '42, von Neumann decided that the experiment could be fatal to people, just as Tesla had suggested. Ironically, he still got upset at the mention of Tesla's name. He decided a third generator would do the trick. He had time to build one but never got the third one to synchronize with the other two. It never worked because the gear box was incompatible. The experiment went out of control and a Navy technician was zapped, went comatose for four months and left the project. They pulled out the third generator. Von Neumann wasn't satisfied, but his superiors weren't going to wait any longer.

In July 20, 1943, they decided it was ready and made tests. Duncan Cameron Jr. and his brother, Edward, were in the control room to operate it. The ship was no longer at anchor and orders came by radio to turn it on. Fifteen minutes of invisibility ensued. There were immediate problems with people. They got sick, some experiencing nausea. There were also mental illnesses and psychological disorientation. They needed more time, but the final deadline was given for August 12th, 1943. The orders came from the Chief of Naval Operations, and he said he was only concerned with the war.

Trying to avoid damage to individuals involved, von Neumann tried to modify the equipment so that only radar invisibility would be achieved, not literal sight invisibility.

Six days before the final test on the Eldridge, three UFOs appeared over the ship.

The switch was thrown for the final test on August 12th, 1943. Two of the UFOs left the area. One was sucked up into hyperspace and ended up in the underground facility at Montauk.

Reports from Duncan indicated that he and his brother knew things were going to go wrong with the August 12th experiment. However, for three to six minutes, things looked good. It appeared it might work without any devastating effects. They could see the outline of the ship - everything was gone. There were problems.

The principal radio mast and the transmitter were broken. People were jammed in the bulkheads. Others were walking around in an insane state.

Duncan and Edward Cameron did not suffer the same trauma as their shipmates.

They had been shielded in the generator room which was surrounded by steel bulkheads.

The steel acted as a shield to the RF energy. As they witnessed things falling apart, they tried to shut off the generator and transceivers but were unsuccessful.

At the same "time", another experiment was going on forty years later at Montauk.

Research had revealed that the Earth, like humans, has a biorhythm. These biorhythms peak out every twenty years on August 12th. This coincided with 1983 and provided an additional function for the connecting links through the Earth's field for the Eldridge to be pulled into hyperspace.

The Cameron brothers could not turn off the equipment on the Eldridge because it was linked through time to the generator at Montauk. They figured it wasn't safe to remain on the ship and decided the best alternative would be to jump overboard in hopes of escaping the electromagnetic field of the ship.

They jumped and found themselves pulled through a time tunnel and onto dry ground at Montauk on Aug .12th, '83 at night. They were found quickly and taken downstairs.

Von Neumann met Duncan and Edward and indicated he knew they were coming. He was now an old man. He said that there had been a lock up in hyperspace and that he'd been waiting since 1943 for this date. He told the time travellers that the technicians at Montauk were unable to go shut things down. Duncan and Edward were required to go back to 1943 and shut the generator off. Von Neumann even told them that the historical records showed that they had turned it off. But they hadn't done it yet! He told them to destory any equipment if that's what it took.

Before returning to 1943 for good, Duncan and Edward did some missions for the Montauk group. They made a number of trips back to 1943. On one of these trips, Duncan passed through the time portal and entered the time tunnel. Duncan somehow entered a side tunnel and got caught there. Side tunnels were a mystery and remain so.

Even though the Montauk scientists theoretically considered side tunnels non-existent, Duncan was warned not to enter them if they should appear. Edward soon ended up in the same tunnel with Duncan.

A group of aliens revealed themselves. Apparently, the side tunnel was an artifical reality created by the aliens. They wanted a piece of equipment before they would let their captives go. This equipment was a very sensitive instrument that charged the crystal drive to the UFO that was underground at Montauk. The aliens didn't seem to mind leaving a ship, but they were very intent on keeping the drive source a mystery to humans.

Duncan and Edward returned to Montauk and retrieved the drive for the aliens.

Eventually, they were able to return to the Eldridge and carry out von Neumann's orders. They smashed the generators, transmitters and cut every cable they could find. The ship finally returned to its original point at the

Philadelphia Naval Yard.

Before the portal closed, Duncan returned to Montauk in 1983. His brother, Edward, remained in 1943. Duncan is not sure why he returned. It has been suggested that he may have been under orders or programmed to do so.

This adventure turned out to be a disaster for Duncan. His time references totally dissolved, and he lost his link to the time line. When time references are lost, one of three things happens: aging slows down, remains the same or speeds up. In this case, it speeded up. Duncan began to age rapidly. After a short amount of time, he began to die of extreme old age.

We're not sure how this happened, but we believe von Neumann transferred him to another time. Scientists were enlisted to help him. They couldn't let the Duncan from 1943 die. He was not only invaluable to the project, he was elaborately involved with the entire scope of time. His death could have created bizarre paradoxes and had to be avoided.

Unfortunately, Duncan's body was dying and there was nothing that could be done to alter the rapid aging. But there was another alternative. Research had already demonstrated that each human being has their own unique electromagnetic identity. This was commonly referred to as one's "electromagnetic signature" or just "signature." If this "signature" could be preserved when Duncan's body ceased to function, it could theoretically be transferred to a new body.

The Montauk scientists were already intensely familiar with all of Duncan's electromagnetic manifestations from the exhaustive research that had been done.

By some means, I'm not sure how, his "soul" or "signature" was transferred to a new body.

They sought help from one of their most loyal and effective agents: A. Duncan Cameron Sr., who happened to be the father of Duncan and Edward Cameron.

Duncan Sr. was a mysterious character. He was married five times over the course of his life. He had numerous connections and didn't seem to work. He spent his time building sailboats and travelling to Europe. Some have alleged that he smuggled Nazi and/or German scientists into the U. S. via his boating activities.

There is practically only one tangible piece of evidence that connects him to intelligence circles. He appeared in a photo of a special graduation for intelligence personnel at the Coast Guard Academy. He was not officially affiliated with the Coast Guard in any way.

Through the use of the Montauk time techniques, the Montauk group contacted Duncan Sr. in 1947. They informed him of the situation and told him to get busy and have another son. He now had a different wife than Duncan Jr.'s original mother. Duncan Sr. cooperated and a child was born, but it was a girl. His directions were to produce a son. Finally, a boy was born in 1951. "Duncan" was chosen as the name for this child, and this is the same Duncan I know today.

The Montauk techniques are obviously remarkable, but they were not sophisticated enough to move Duncan from 1983 straight back to 1951. There could have been other factors involved, but it appears the scientists had to rely on and ust the twenty year biorhythms of the Earth. As Duncan's original body was dying, he was transferred to 1963 and "installed" into the new body provided by Duncan Sr. and his wife.

Duncan Jr. has no memories prior to 1963. It is also obvious that whoever occupied his body between 1951 and 1963 was forced out.

I have often heard accounts of a secret project that was run by ITT at Brentwood, Long Island in 1963. It is entirely possible that transferring Duncan to a new body was the focal point or a very important part of this project. Whatever the circumstances, this project would certainly have been trying to somehow utilize the Earth's bioshythm that occurs every twenty years.

Edward Cameron had returned to 1943. Duncan was in 1963.

After the August 1943 experiment, the Navy brass didn't know what to do. Four days worth of meetings ensued with no conclusions. They decided to do one more test.

In late October of 1943, the Eldridge disembarked for the final experiment.

No personnel wree to remain on board. The crew boarded another vessel and controlled the equipment on the Eldridge remotely. The ship became invisible for about fifteen or twenty minutes. When they boarded it, some of the equipment was missing.

Two transmitters, and a generator were gone. The control room was a burnt shambles, but the zero time reference generator was left intact. It was put into secret storage.

The Navy washed their hands of the entire operation and officially launched the USS Eldridge with its office of record. The ship was eventually sold to the Greek navy who later uncovered the log books and found that everything before January of 1944 had been omitted from the records.

According to Al Bielek's account, Edward Cameron continued his career in the Navy. He had top level security clearance and probed into many sensitive areas such as "free energy" vehicles and devices. He was outspoken and complained about improper proceedures.

For whatever reason, he was brainwashed to forget the Philadelphia Experiment and anything else to do with secret technology.

Al has stated that age regression techiques were used to put Edward Cameron into a new body in the Bielek family. The Bielek family was chosen as there was only one child in the family and the baby had died by the time of his first birthday.

Edward was substituted and the parents were brainwashed accordingly. Edward has since been knwon as "Al Bielek."

Age regression techniques have been traced back to Tesla. When he was working on the original Philadelphia Experiment, he developed a device to help sailors in the event that they lost their time locks. The purpose of this device was to reestablish an individual with his normal time locks in the event he had been disoriented from time travel. The government or someone allegedly used this Tesla device and developed it for physical time regression.

Tesla said that if the time locks of an individual are moved ahead in time, one could actually remove age. If one's time locks were pushed back twenty years younger, the body would be referenced to those locks.

Edward Cameron now became Al Bielek. Al grew up with his own identity and education and became and engineer. Eventually, he ended up working at Montauk.

It was not until the mid 1980's that Al began to get memories from his earlier identity. To this day, he continues to doggedly research the Philadelphia Experiment and is planning to write another book. He intends to prove, even to the most skeptical, that the Philadelphia Experiment did indeed happen.

APPENDIX F- QUANTUM LEVELS OF EXISTENCE (According to Preston Nichols)

When I am referring to quantum levels of existence, "quantum" refers to the different or many possible levels. "Quantum" comes from the Latin root "quantis" which means quantity.

Understanding multiple realities is key to understanding time. Conventional physics does not deny the possibility of parallel existences, but it is mostly concerned with theories about matter and anti-matter. As there is much evidence to prompt scientific investigation into the area, there are currently about sixty theories around the world and ten in the U. S. that deal with quantum levels of reality.

I am offering my own theory based upon my experiences, some of which are covered in this book. Like any proper scientific theory, this is being offered because it has proven workable to me in the laboratory. It will also help the reader get a better grasp of how time functions.

What exactly is parallel reality?

It would be a world or universe that has almost everything we have here. If we switched into it, we'd see another body that would represent us in the other existence.

The parallel universe would not necessarily behave exactly like the one we're familiar to. It would have unique properties unto itself.

It is my understanding that we exist in a number of parallel realities. We are primarily conscious of "our reality" because we are focused or referenced to it. The parallel universes might reach our consciousness through dreams, ESP, meditation or artificially induced mental states.

It is now important to consider the overall view and what these different realities might look like in schematic form. Einstein theorized that if one travels in a straight line from any particular point in space that one would eventually end up in exactly the sam eplace that they started at. This could be considered a full loop.

We are not going to go into the equations of that, but the general reader can grasp this by understanding what Einstein called a time toroid. For this purpose, a toroid can be likened to a two dimensional doughnut. Einstein likened the entire universe to a time toroid. He theorized that if one started in a straight line from any given point on the outside of the doughnut, that one would end up at the exact opposite end of the doughnut. Both of these point would be essentially the same, except that one could be considered "positive" and the other "negative." As they are both points in the infinite stream of time, one point could be called "positive infinity" and the other "negative infinity."

For purposes of explaining my theory, I have extrapolated upon Einstein's idea and have converted his time toroid into a three dimensional sphere. The reader can easily picture a smaller sphere inside of a bigger sphere. For clarification, we will call the inside sphere, Sphere A, which can be likened to a softball.

The bigger sphere will be called Sphere B and can be likened to a basketball (which is hollow).

Our experiences in different realities all occur within the realm of Sphere A.

If you start from a point on Sphere A and travel in a straight line, you will eventually come back to where you started from.

Within Sphere A, everything in dynamic and moving. It is time as we conceive it.

The area between Sphere A and Sphere B is not dynamic at all. In fact, it is considered to be at rest. We can postulate this area to be a sea of time particles.

These are not particles in any ordinary sense. In fact, trying to describe them in this three dimensional example is awkward. We are simply assuming these non-moving particles of time exist because we can sense them (if only in idea form). These particles between Sphere A and Sphere B would be the same as the particles in dynamic time (within Sphere A) except that they are static, i. e. at rest.

Incidentally, we are not conscious of static time because our "normal" reality is built upon dynamic functions or dynamic time.

A reality in time is created when God or someone creates a stress on the wall of Sphere A. This stress will cause the dynamic particles inside of Sphere A to move and travel through the sphere until there is a loop, thus completing the Alpha and Omega (beginning and end).

Our reality can be considered one gigantic loop. It could have started with the big bang or the beginning of the universe and would eventually end there, but it would in fact continue ad infinitum.

When someone or something then takes that loop that we are in and creates a new time stress and changes reality, a new loop is created that is, in fact, an alternate reality. The original loop cannot be obliterated or denied. It will still be there. The new loop could be modified in any way the modifier chooses. It could be an opening in 1963 that goes to 1983. Everything in between those times would be an alternate loop. It would not be an entire loop of its own, but would be added to the original loop of our normal reality. In this way, partial loops would be added on to our original time line, and we can call this conglomeration of loops a manifold.

Each loop could also be called a manifold (a manifold generally refers to something that has many parts).

As different alternate realities are created off of the original loop, additional manifolds are added to the sphere and make it swell. In addition to the alternate realities created by changing the reality of a particular time line, there could be parallel realities created in the beginning of time that also have their own "original" loops as well. There is an infinate number of loops and manifolds possible.

Some people may wonder about Sphere B in the above example. It is basically there to make the theory fit. At this point, I can add no further significance to Sphere B except that it serves as a wall containing time stress particles. It could possibly be a part of a bigger scheme of metaphysics.

Now that you have a grasp of how these time loops and manifolds fit into an overall picture of the universe, there is another key question that must be asked.

Is it possible to gain consciousness of the other loops or manifolds?

Yes, it is. This is what happened on my roof when I was putting up my Delta T antenna (as discussed in Chapter Six). That antenna has a subtle interdimensional effect on the nature of time itself. It enabled me to regain consciousness of an alternate time line that I had been put on against my innate will.

It is therefore possible for others to travel from one time loop to another.

In fact, it appears that this is the entire reason for the Philadelphia Experiment and Montauk Project in the first place. This theory indicates that not only was an alternate time loop created but that this loop enabled a vast influx of alien UFOs to come to this planet. UFOs have always been around, but there is no denying the sudden frequency of reports in the 1940's.

Even if you do not accept any of this as the truth, it is quite obvious that this is the type of advantage an alien race might have over us.

The next point I want to address is that parallel realities are based upon principles common to electromagnetics. For example, it is common knowledge that alternating current is created by an alternating difference in potentials.

It is the relationship between the voltage and the current that makes alternating current function. Further, the relationship between the current and voltage is inextricable.

Just as there is an interplay between the voltage and current, there is one between two different realities.

Extrapolating from these principles, it can be understood that parallel realities are ninety degrees out of phase with our "normal reality." In other words, if there is a parallel reality, one has to consider that it has potential energy. It is not activated of itself. It would also be ninety degrees out of kilter from our normal viewpoint. The fact that it is potential energy means that it has the capability to flow to our reality and vice versa.

This explains that there is not only a relationship between electromagnetic principles and other universes, but suggests that by utilizing electromagnetic principles that one can theoretically enter the realm of other realities. These would include the alternate time loops of which I have already spoken.

It is hoped that the above will give the reader a general understanding of how electromagnetic principles were used to manipulate time at Montauk.

APPENDIX G- GLOSSARY

Amplitron- A high powered UHF amplifier. At Montauk, this served as the final amplifier of the transmitter before a function was radiated out the antenna.

A large tube, it weighed 300 pounds and measured 35 inches in its largest dimension.

Biorhythm- This is an esoteric term and refers to any regularly repeating life function in an organism. A biorhythm is perhaps best understood in terms of Oriental "Ki" or "Chi" energy, which is the life force that regulates the entire body. Acupuncture addresses biorhythms in order to affect a cure. When the planet is considered as an organism, biorhythms would include all the subtle functions that make life possible and regulate it. The seasons, rotation of the Earth and spinning of the galaxy would all be taken into account. Legendary places such as Stonehenge are considered to be constructed in harmony with the biorhythms of the planet.

Cathode- In a vacuum tube, the electron emitting material is called a cathode.

In an electrolytic cell, it is the negative electrode from which current flows.

In essence, it is a source of flow.

Cycle- A unit of activity within a wave that continually repeats itself. A cycle will go up and down before it repeats itself. If you visualize ocean waves that are all uniform, the series of waves would be called the "wave." The one ocean wave that a surfer might ride would be a "cycle."

Delta T- Short for "Delta Time." Delta is used in science to indicate change, thus "Delta T" would indicate a change in time.

Delta T antenna- An octahedronal antenna structure that is designed to bend time.

Visually, it looks like two pyramids sharing the same base. By definition, it can actually facilitate shifting time zones. Two coils are placed vertically around the edges of the pyramid structure at 90 degree angles to one another.

A third coil surrounds the base. Shifting time zones was accomplished by pulsing and powering the Delta T antenna, as is discussed in Chapter 12. Even when the antenna is not powered, it has a subtle interdimensional effect on the nature of time itself.

DOR- Stands for "Dead ORgone" (see definition of "orgone".) This refers to life energy that has become stagnant or negative. DOR couls be considered the antithesis of life energy.

"Electromagnetic bottle"- This refers to a "bottle effect" that is created when a specific space is surrounded by an electromagnetic field. The specific space itself is the inside of the "bottle." The walls would be the electromagnetic field.

When people or objects are within the specific space, they would be within an "electromagnetic bottle."

Electromagnetic wave- When an electric charge occurs that oscillates (swings back and forth), a field around the charge is generated. This field is both electric and magnetic in nature. This field also oscillates which in turn propagates a wave through space. This was is called an electromagnetic wave.

Frequency- The number of waves or cycles per second.

Helmholtz coils- commonly, Helmholtz coils refer to two identical coils that are separated by a distance of one radius of the coils themselves. (You can visualize this if youthink of two hula hoops parallel to each other.) When the coils are electrified, they produce a homogeneous magnetic field over a larger volume of space than does a single coil.

Hertz- (abbr Hz) This is simply one cycle of a wave. A wave consists of numerous cycles that are repetitions of one cycle. To be a bit more technical, hertz is the complete fluctuation of a wave from plus (the highest point) to minus (the lowest point.) Five hertz would be five cycles per second.

MHz- MegaHertz, which are equivalent to 1,000,000 hertz.

Non-hertzian component- This term does not exist in conventional science. It refers to the etheric component of electromagnetic waves. Theoretically, the non-

Hertzian component is a wave function. Instead of oscillating transversely, it oscillates with the direction of propagation, which is known as longitudinal (i. e. sound waves.) It could be looked at as an "acoustical" electromagnetic wave.

Orgone- This refers to life energy or sexual energy as observed by Dr. Wilhelm Reich. It is the positive energy that "makes us tick."

Oscillator- A device that establishes and maintains oscillations. To oscillate means to swing back and forth. In electronics, an oscillations refers to a regular variation between maximum and minimum values, such as current or voltage.

Phase- The time intervale between when one thing occurs and the instant a second related thing takes place.

Phase conjugation- This is the process whereby a wave comes back from a received source that is an imaginary image of a transmitted wave. In other words, when a radio wave is transmitted, an image goes back to the transmitter by the process of conjugation. (For more information, one can read up on modern electro-optics theory.)

Phoenix Project- A secret project that commenced in the late 1940's. It researched the use of orgone energy, particularly in regards to weather control. It eventually inherited the Rainbow Project and included the Montauk Project itself. "Phoenix" was an official code name.

Psycho-active- This pertains to any activity or function that has an effect on the mind or psyche. In this book, psycho-active refers primarily to electromagnetic functions or electric equipment that influence human thinking and behavior.

Psychotronics- The science and discipline of how life functions. It includes the study of how technology interacts with the human mind, spirit and body.

Science, mathematics, philosophy, metaphysics and esoteric studies are united through the study of psychotronics. It would also include other realities and how we interface with other dimensions of existence.

Pulse modulations- These are sent as a series of short pulses which are separated by relatively long stretches of time with no signal being transmitted.

RF- Radio Frequency. Frequencies above 20,000 hertz are called radio frequencies because they are useful in radio transmissions.

Radio wave- An electromagnetic wave that carries intelligent information (pictures, sound, etc.)

Relativistic- Relativistic functions refer to activities that are out of our normal reference frame. It also concerns how activities in other reference frames relate to ours. Relativity embraces the concept of everything without any limitations, including other dimensions and the entire universe(s).

Sideband- This is the component of radio waves that actually carries the intelligent information.

Space-time- When you study higher level physics, it becomes apparent that space and time are inextricably related to each other. It is considered less accurate to refer to just space or time by itself (because they don't exist by themselves.)

That would be like saying your mouth ate the dinner.

Time reference(s)- This refers to the electromagnetic factors by which we are connected to the physical universe and the stream of time. Consciousness of time can be likened to a deep hypnosis which causes one to be in sympathy with the various frequencies and pulses of the physical universe.

Tranceiver- An instrument that serves as both a receiver and a transmitter.

Transmitter- A device or unit that sends a signal or message.

Wave- A state of motion that rises and falls periodically is called a wave. It can be transmitted from one particular area to another with no actual transport of matter taking place. A wave consists of many cycles and can carry signals, pictures or sounds.

THE STORY CONTINUES...

MONTAUK REVISITED, Adventures in Synchronicity: By Preston B. Nichols and Peter Moon Montauk revistied unmasks the occult forces that were behind the scene and technology used in "The Montauk Project." An ornate tapestry is revealed which interweaves the mysterious associations of the Cameron clan with genesis of American rocketry, the bizarre history of the electronic transistor and the magic of Aleister Crowley. The Montauk investigation carries forward as Preston Nichols opens the door to Peter Moon and unleashes a host of incredible characters and new information. A startling scenario is depicted that reaches far beyond the scope of the first book, "The Montauk Project." Illustrations and photos .256 pgs.

PYRAMIDS OF MONTAUK, Explorations in Consciousness: By Preston B. Nichols and Peter Moon An astonishing second sequel that awakens the consciousness of humanity to its ancienty history and origins through the discovery of pyramids at Montauk. Their placement on sacred native American ground opens the door to an unprecedented investigation of the mystery schools of earth and their connection to Eygpt, Atlantis, Mars and the star Sirius. Mr. Nichols fascinates us with an update on covert operations that includes the discovery of a nuclear particle accelerator and the development of new psychotronic weapons. "Pyramids of Montauk" stirs the quest for the end of time as we know it .256 pgs.

THE MONTAUK PULSE, A Chronicle of Time

As "The Montauk Project: Experiments in Time" originally went to press in May 1992, many new and startling developments began to take place. Consequently a newsletter entitled "The Montauk Pulse" began in January 1993 to keep interested readers up to date on breaking news regarding the Montauk scenario.

It was not envisioned at that time that two subsequent books would be published on the subject. The Pulse has a distinct identity from the books and will often comment on details and history that don't necessarily find their way into book form.

The Montauk Pulse contains at least six pages per issue and is published quarterly.

It includes exciting new breakthroughs on the Montauk story as well as similarly related phenomena like the Philadelphia Experiment or other space-time projects.

The Delta-T Antenna

The Delta-T antenna is the device of legend from 'Project Phoenix'. The device consists of 3 antennas or coils each at 90 degrees to the other. The main two are fed a sine and cosine wave so that they are 90 degrees phase-shifted. This starts up a rotational magnetic field in the center of the antennas whose rotational velocity is equal to the input frequency. The third antenna pulls the field into the 3rd dimension.

Here is a possible Delta-T configuration:

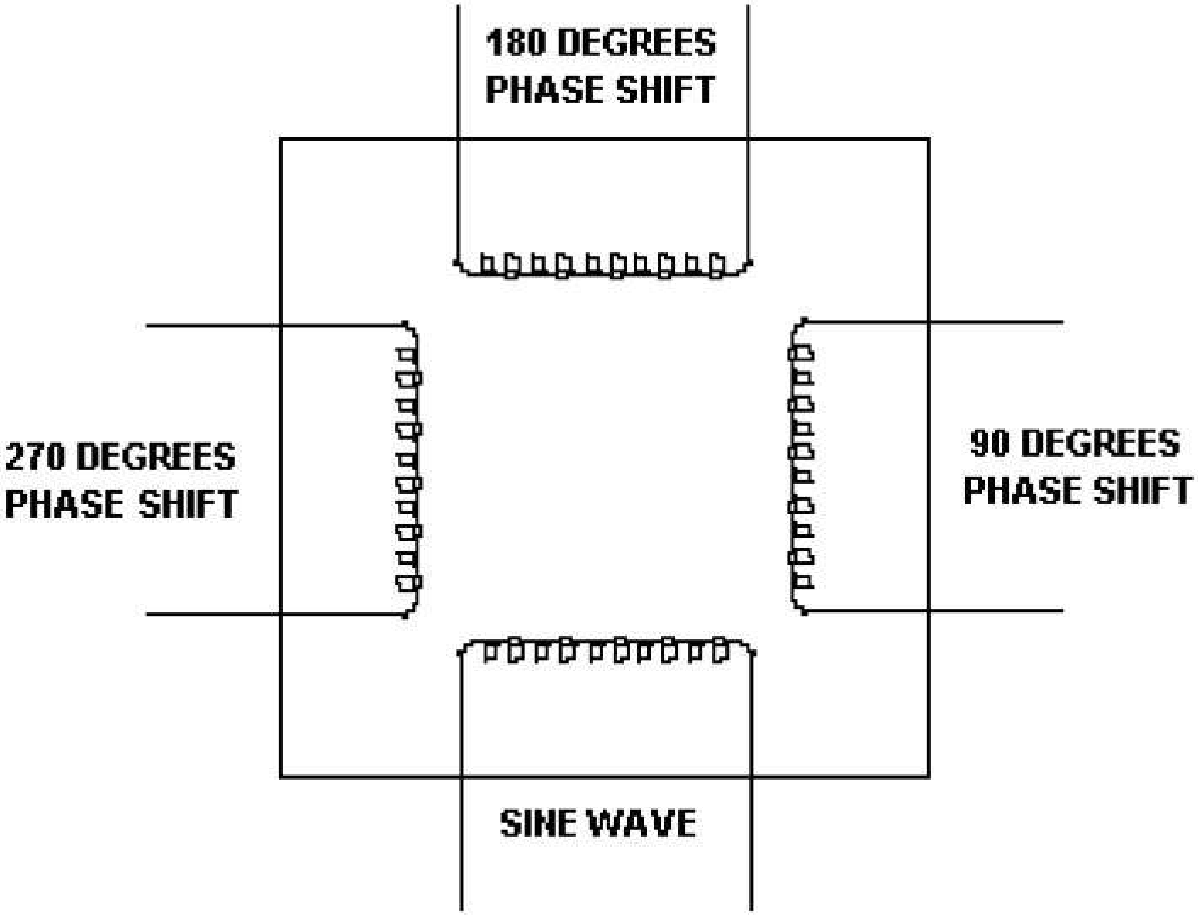
180 DEGREES PHASE SHIFT
270 DEGREES PHASE SHIFT
90 DEGREES PHASE SHIFT
SINE WAVE

Construction:

This device (Delta-T) enables time-traveling. The central part of the device consists of 2 frame antennas that are arranged 90 degree in relationship to each other. Each frame is approximately 2 meters in diameter. The antennas can consist of tubes which are 10 cm in diameter. The tubes must be superconductive. Inside the tubes float liquid helium for the superconductivity.

The control device of the machine consists of a sinus generator with a tuning capacity from 1-Hz to 1000-MHz. The signal from this generator is amplified a little before it goes to a phase changer. This changes the phase of the signal 90-degrees. The phase changer is independent of the frequency in. The phase change is 90-degrees for both the 1-Hz and 1000-MHz input. The in-signal is split just before the phase changer so that one part is unchanged and another part is changed 90-degrees. After the phase changer, the signal amplified with 2 linear amplifiers to 100,000 watts on each channel. After that, the 2 channels are fed to its respective frame antenna. The signal on one frame antenna will always be 90-degrees in phase difference to the other frame antenna.

Usage:

The generator is set to 1-Hz and the machine is turned on. A compass needle at the center of the machine will travel slowly round. Inside and round the antenna system, it becomes a powerful rotating magnetic field.

At a higher generator frequency, the magnetic field will rotate faster. Approximately 2 meters from the antennas are placed several chairs arranged in a circle around the antenna system. Here the time travelers are placed. One increases now the generator frequency gradually until the speed of the rotating magnetic field is close to the speed-of-light. If the right conditions are present now (etherical/astral), the time will run slower for those who are sitting on the chairs compared to persons outside the power-field. Seen from the chairs, the time will run normal between each chair. The time travelers' clock will run normal.

They will see the clock on the wall outside the power-field running fast into the Future. The nearer the rotating speed of the magnetic field comes to the speed-of-light, the faster will the time travelers move into the future. If the speed of the magnetic field is set equal to the speed-of-light, the time will freeze as seen from outside the magnetic field. As seen from the time travelers view, they are moving endlessly fast into the future (they will not be able to sense the world outside the magnetic field and the time will stop for them). If one increases the frequency further, the rotating speed of the magnetic field will become greater than the speed-of-light. The time sensed from the chairs would then be reversed, and they will see the clock on the wall outside the magnetic field moving backwards. The time travelers will now travel back in time.

At the speed-of-light, the time travelers go both endlessly fast into the Future and the Past at the same time. The result is that the time freezes at that speed.

At a higher frequency, they will travel slower into the Past. When the speed of the magnetic field approaches the speed-of-light from above, the time travelers will travel faster into the past. At the speed-of-light for the magnetic field rotation, the time is frozen for the people inside the power-field and they can't register anything. For an outside person who looks inside, the power-field it looks like looking at a frozen hologram. If one from outside the magnetic field tries to touch the field at this state, it seems harder than steel. The outside person will at this state not be able to penetrate the power-field which acts as an endlessly strong armor at that state. If one puts a clock inside the power-field, an outside person will observe the following.

At 1-Hz, the clock will run normal. When the frequency arrives at the lowest part of the VHF band (above 30-MHz) the clock seems to be going slower. At the top of the VHF band (approximately 300-MHz), the clock will stop completely. In the lower part of the UHF band (from 400-MHz), the clock begin to go slowly backwards. At 1000-MHz, the clock will go approximately at normal speed backwards. It's also possible that the machine must have an injection of modulated orgone energy to function.

The Montauk Project Papers

by Thomas Skeggs

Introduction

Over the last few years I have amassed some information on 2 strange experimental projects known as the **Montauk Project** and the **Philadelphia Experiment**.

So I have decided to publish extracts from some of my papers on these fascinating stories. There is a lot of debate surrounding the Montauk Project and the Philadelphia Experiment. Some of this debate can get a bit heated where both skeptic and believer trade insults and accusations of fraud, hoaxing, and even lying. Others make the claim that the Montauk Project and the Philadelphia Experiment are nothing more than disinformation -- a tool to distract the gullible investigator, researcher and journalist away from the real truth surrounding highly-classified and secretive research programs.

When writing these papers, I have simply looked at them from purely a technical point-of-view to try to determine the true nature on how the equipment described in these projects may have actually worked.

The Montauk Chair

According to the authors Preston Nichols and "Peter Moon", the technology to read a person what way a person was thinking was originally allegedly developed at the ITT Technical Institute at the University of Southampton, Long Island in the 1950s. (Some basic research reveals the ITT Institute has closed down at Southampton). Nichols states in his book that RCA went on to develop a second version of the chair during the 1970s.

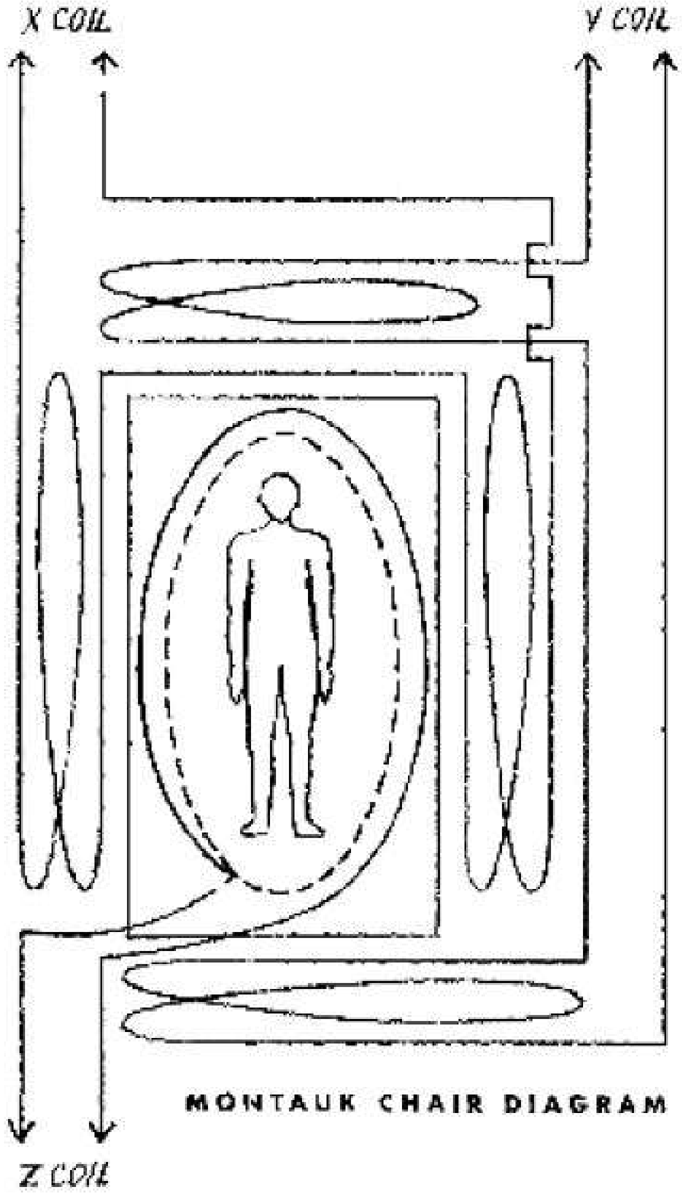

a diagram drawn by Nichols describing the lay-out of the Montauk Chair.

Nichols claims the chair was surrounded by 3 **Helmholtz coils** described as the 'X' coil, 'Y' coil and 'Z' coil. Note the 'X' coils are located on each site of the individual. The 'Y' coils are located above the head and below the feet. And the 'Z' coil appears to be wound around the chair. (2 loops could have been used above and below the chairs which may have been designed to pick up fluctuations or modulations within the electrical magnetic field surrounding the human body.)

Helmholtz coils consist of cable wound in loops in to produce a coil. Helmholtz coils are used in industry and science to produce a constant magnetic field. Also notice in the above diagram that each pair of loops are wound in opposite directions -- one clockwise and the other counterclockwise.

This arrangement is known as a **pair Helmholtz coils**. And the magnetic field is uniform in the center between the 2 loops or Helmholtz coils.

The diagram above basically shows the magnetic field that is produced when a currant is applied to the cables wound in a loop like those found in a pair of Helmholtz coils.

The 3 Helmholtz coils were connected to what appears to be 3 antiquated value-driven radio receivers.

Researching other sources of information has drawn very little supporting evidence from other sources regarding reading the thoughts of people. Yet I have found a short reference to the work carried out by Lawrence Pinneo (a neurophysiologist) from the Stanford Research Institute (SRI) in 1974. He had developed a computer system capable of **reading a person's mind** by correlating the brain wave 3

with subjects on an ECG with certain commands. Writer and researcher K.B. Wells, Jr. writes in his book The Montauk Files that the CIA developed a chair in the 1960s which used polygraph-like technology to determine if the person sitting in the chair was telling the truth.

Also scientists at the University of Sussex, England had developed a sensor which could read a heartbeat of an individual from between 1-2 meters. The sensor would not have to come into physical contact with the skin. The scientists said it would be especially useful with monitoring the heartbeat of serious burn victims. The scientists stated the sensor worked by picking up fluctuations in the electromagnetic field surrounding the individual. The heartbeat causes the fluctuations, and sensor can pick these tiny fluctuations and filter out noise. They also said the sensor was highly sensitive voltmeter.

The Montauk chair system may have operated in a very similar way by acting as a highly sensitive magnetometer, which could record the slightest change in within the individual’s magnetic field. The Montauk chair system may have focused on brain activity. The coils are all wired up to 3 old radio receivers with no sign of an external power source. Yet the above diagrams immediately draw to mind Faraday’s Law. If you have a fluctuating or moving magnetic field, it will induce a currant to flow with a coil of wires. This means that the Montauk chair system may have been actually powered by electromagnetic field surrounding the human body of an individual.

Chair Notes

- The first version of the chair was allegedly developed at the ITT Technical Institute in Southampton, Long Island during the 1950s.

- The second version of the chair was allegedly developed by RCA during the 1970s.

- The chair system used 3 pairs of Helmholtz coils, designed to pick fluctuations of the electromagnetic field surrounding an individual.

- A computer system was developed in 1974 at the SRI which could read the minds with certain commands.

The Montauk Chair Receivers

According to the Preston Nichols' first book The Montauk Project: Experiments in Time, there were 2 versions of the Montauk Chair. The first version of the infamous Montauk chair used 3 Hammarland SP600 or the Hammarland Super ProP600 radio receivers. I later found references to Hammarlund SP600 receivers on the Internet plus references to a Super Pro SP600. It appears that Nichols misspelled the make and model numbers. These valve-driven radio receivers were built between 1950 to 1972. There are also numerous model variations usually identified with suffix number. For example, SP600-J12. Nichols has not printed a suffix number, so it still remains uncertain as to what actual model was used. It can be important because most of the model variations have slightly different specifications and frequency responses.

HAMMARLUND MODEL SP-600

a photo of the Hammarlund SP600 Radio receiver claimed to have be used at Montauk

There are several models made and were widely used by the U.S. Navy, Signal Corps, and the Air Force. Hammarlund also produced a civilian version of the SP600 which was popular with Amateur Radio enthusiasts. (Nichols and Bielek are such enthusiasts and attend HAM radio meetings.)

another photo of a Hammarlund SP600 radio receiver. This may be a Hammarland Super Pro SP600

Yet one odd feature is the frequency ranges. The Hammarlund SP600 receiver has a frequency range of 0.54-54 MHz.

According to Nichols, they developed a second version of the Montauk chair which used 3 radio receivers built by RCA. This type of receiver has proven harder to find information on. In Nichol’s book, he first just referred to the receivers as RCA "1935" (which was the year the first FM radios were developed).

Nichols then names the second version of the radio receiver built by RCA as a "FRR-24". Later in his book, he also claims the FRR-24 was used in the legendary Philadelphia Experiment. Nichols suggests it was an **extra-dimensional resonator**. But the FRR-24 is a receiver -- not a transmitter. It would only be able to detect-or-receiver signals and not produce -or-propagate them. But it does seem possible that an AN/FRR-24 could have been fitted as standard kit to a new warship being built in 1943 like the *USS Eldridge*. 5

Nichols also stated in his book that he managed to trace a man who had used a AN/FRR-24 receiver and this man sung the praises of the receiver. But this did point out to Nichols that the set was prone to picking up an unidentified source of interference. I did find a reference to an AN/FRR-24 radio describing it as just a receiving set. (Nichols left off the 'AN' prefix which means it’s a set made for the military). Information on this particular model appears very sketchy. I did manage to find some details on its predecessor the U.S. Navy’s AN/FRR-23 (also built by RCA). See picture below. Finding a photo of the AN/FRR-24 or technical details has drawn a blank.

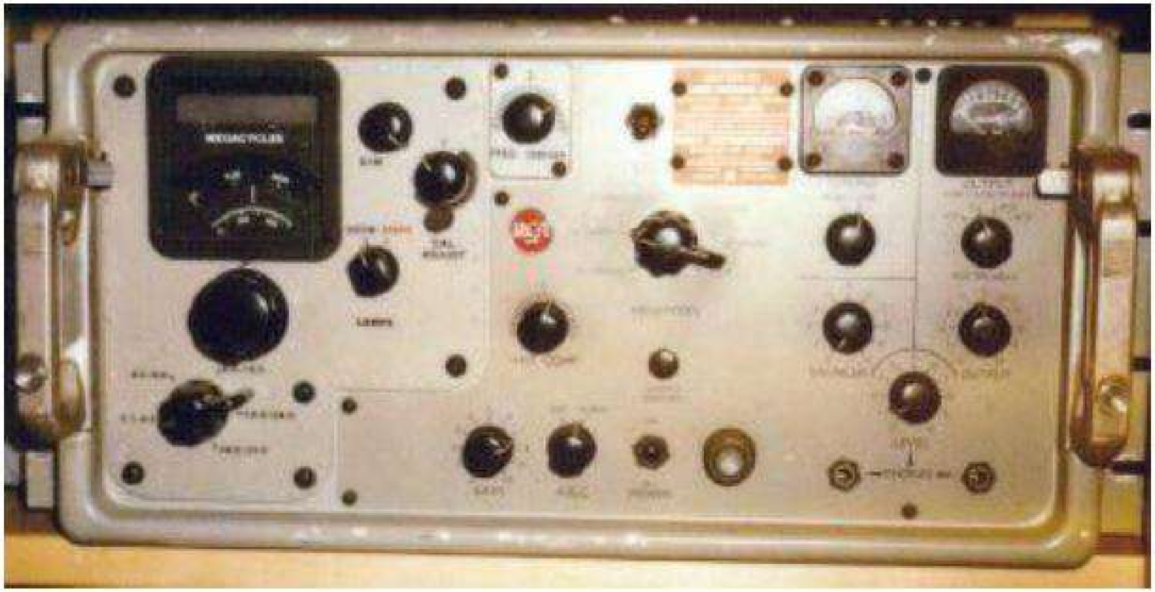

This is a rare photo of the AN/FRR-23, built by RCA for the U.S. Navy. Trying to find a photo of the AN/FRR-24 as used in the Montauk Project has drawn a blank so far.

Note in the picture of the AN/FRR-23 the red RCA label. And it was built for the U.S. Navy.

The AN/FRR-23 was an AM (Amplitude Modulation) HF receiver. It had a frequency range of 2-32 MHz. The date AN/FRR-23 entered service was around 1940-45. The radio sets appear to be very heavy, between 900-100 lbs.

I also found an intriguing reference to special radio receiver known as the "R-274/FRR" ordered by the U.S. Navy in Oct 1957. Its frequency range was between 100-400 khz and 1.35-29.7 MHz. This radio receiver was a version of the popular Hammarlund SP600 series. (Could this have been a replacement for aging AN/FRR-24 receivers).

The radio receiver system described by Nichols appears to consist of 3 antiquated valve-driven radio receivers developed for the U.S. Navy and civilian use in the 1940s and 1950s. Yet Nichols stated that the Montauk Chair system was developed in the 1950s. Why would they use old valve receiver sets?

I recall one TV documentary on the Cold War. The Americans had managed to capture a MIG 21 fighter aircraft. And when technicians took the MIG 21 apart, they found its avionics largely consisted of old valve-driven circuits. The technicians laughed at how primitive it was. During the 1950s and 1960s during atomic bomb tests in Nevada, technicians found that the high-tech transistor-based avionics fitted in most Western military aircraft was prone to the Electro-Magnetic Pulse (EMP) produced when atomic bombs are detonated. The EMP would seriously damage unprotected transistor based circuits, and the old valve circuits were a lot more robust. The technicians then realized why Soviet aircraft were fitted with the old style valves. It may be due to the sets are less prone to damage by sudden bursts of EM radiation. 6

This could explain why such old and antiquated radios sets could have been used at Montauk because they were less prone to EM interference. Secondly, many radio and hi-fi fans will usually state that valve-driven radios and amplifiers produce a much better quality sound and signal output.

One thing that did catch my attention is that Nichols stated that each of the 3 receivers may have been tuned to 3 different frequencies. And each of the 3 radio receivers were connected to 3 sideband detector producing 2 outputs: a USB (Upper SideBand) and LSB (Lower SideBand) output. Looking at the wiring diagram, the signals picked up by the Helmholtz coils and radio receivers surrounding the chair appeared to be mostly filtered out. So what the sideband detectors may have done is pick up fluctuations or modulations within the electromagnetic field surrounding the human body (as described in the 'Chair' section).

I'm still following up research in this area of using LSB and USB. Plus I will try to find out if these sets are AM or FM receivers.

Another oddity is that Nichols wrote in his first book that these old radio receivers were located inside the transmitter/radar tower. That appears to contradict his story that the engineers went to great lengths to isolate and screen out noise around the infamous Montauk Chair. They then go and place the old radio receivers inside a noisy building where high-powered microwave transmitters are pumping out megawatts of EM radiation. But Nichols stated the Montauk chair was located in the basement of the transmitter building. I have recently conducted some research into Montauk, and I believe the radio receivers and chair system could have been located in another building located near to the transmitter. Not inside or below it. (I will write more about this at a later date).

Receiver Notes

- Version 1 of the chair may have used Hammarlund SP600 receivers or 3 Super Pro SP600 receivers.
- Version 2 of the chair may have used 3 AN/FRR-24 HF AM radio receivers built for the U.S. Navy by RCA during the 1940s.
- The 3 receivers may have been tuned to 3 different frequencies.
- The 3 radio receivers were connected to 3 sideband detectors.
- The receivers are obsolete and antiquated.
- The old receivers were located in the transmitter/radar building.

The Computers

This section will examine how the computers were used to process the information from the radio receivers.

The Transmitters

This section will look at the transmitters installed at Montauk.

The Delta T Antenna

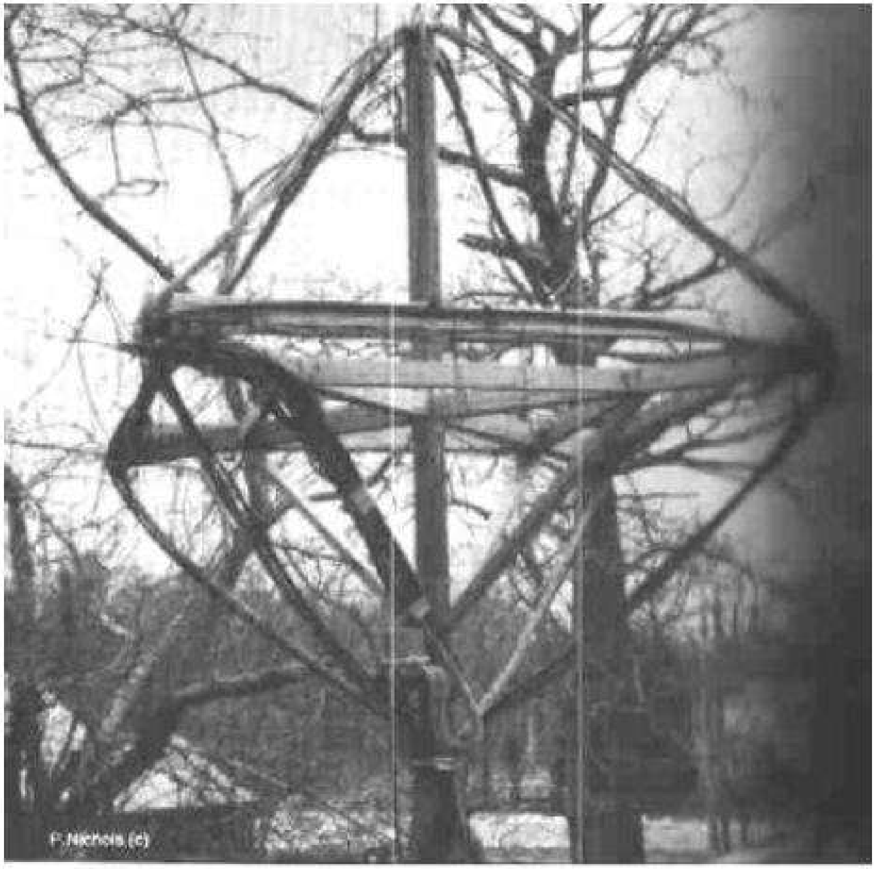

a photo of the Delta T antenna which is claimed to have been used in the Montauk Project by Preston Nichols

According to Nichols, the Delta-T antenna consists of 2 pyramid-shaped wiring looms with 3 Helmholtz coils wired in a special way to create a time-shifting field

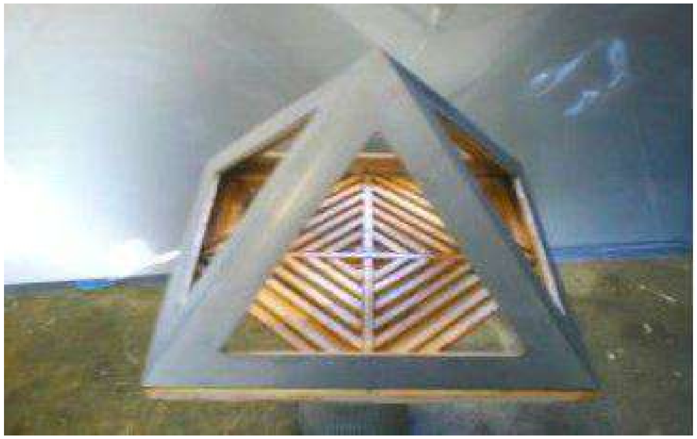

a replica of the Delta T antenna.

It consists of an outer steel skin with a wooden interior. The antenna has a removable floor panels. By looking at the photo, you can see where the coils will be located inside the antenna. The remote-viewing data that I got seems to suggest some transparent material covering the triangular openings like window blinds. It reminds me of a Chinese Lantern. 8

The model is based on written accounts and remote-viewing data.

What's confusing about the RV data is that there may have been different versions of the Delta-T antenna built at different sites or different times. Or there may have been more than one antenna in operation. The RV data seems to offer conflicting RV data regarding its overall shape.

It's also difficult to trace the antennas as they were placed in long-term storage around America. Also different parts were stored at different locations (Plattsburgh, Hanscom, Wright Field, and a base in New Jersey).

I also got some RV data on a possible second site surrounded by Pine trees and a chain link fence. It has a warning notice -- red background with white letters "Danger". It had black letters underneath. The only word I could pick out is "Hazard". It looked bare on the other side of the fence. It had overgrown grass plus antennas and wires. During RV sessions, I am repeatedly drawn to this location.

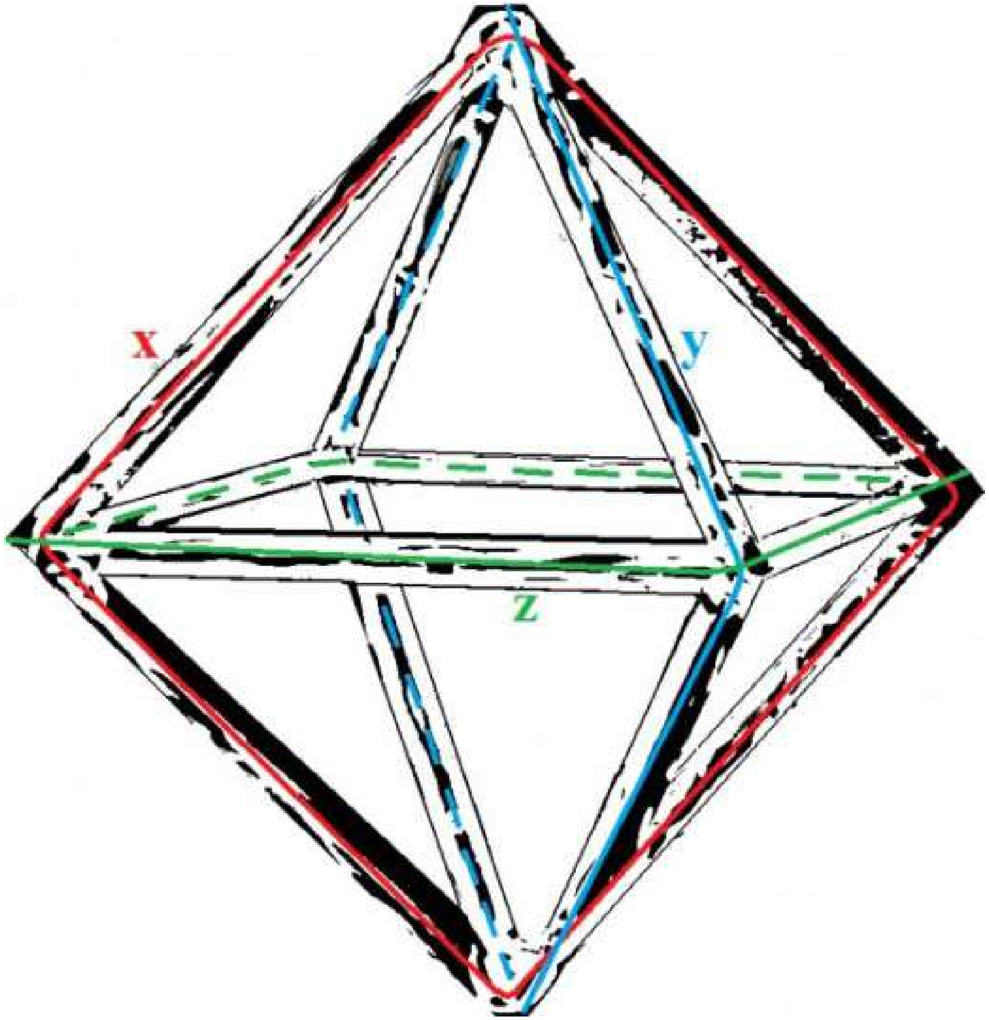

The Delta T Antenna uses the platonic solid shape (building block of the Universe) of a octahedra to achieve it specialized function.

Montauk, America's Greatest Unknown Conspiracy

Updated on October 17, 2015

Radar Instillation, Camp Hero

This is the radar dish used for mind control experiments and EF transmissions at Camp Hero, Montauk, New York.

Peter Moon

THE
MONTAUK PROJECT
EXPERIMENTS IN TIME
PRESTON B. NICHOLS
WITH PETER MOON

AREA CLOSED

The Montauk Project: Experiments in Time

Etymology of Montauk

Native American shamans describe a spirit guide, shape shifter or time traveler as a Manatu or Manituo. According to The Keys of Enoch by J.J. Hurtak, the Tibetan word for Orion is Tak and is described as the gateway from human physical consciousness in this dimension to the next level of human evolution (creation).

Discovering the Conspiracy

What little is known about the United States Militaries' clandestine activities that revolve around Montauk Point Long Island, New York can be summed up in books by Peter Moon, Preston Nichols and Stewart Swerdlow. These men have found the courage to reveal activities that may very well have gone undocumented had they remained silent and not blown the whistle on a secret Government project that could very well affect the lives of every man, woman and child on this planet. In this article Montauk, America's Greatest Unknown Conspiracy we will explore what little is truly known about the subject and speculate on its implications for mankind.

Montauk, America's Greatest Unknown Conspiracy involves mind control experiments, weather manipulation, star gate technology, telepathy, UFOs, Aliens, Nazi's, pyramids, sleeper agents, Aleister Crowley and black magic, time travel and time 'police', remote viewing, Deep Underground Military Bases (D.U.M.B.s) and above all an incredible amount of synchronicity that cannot be easily explained as mere coincidence. The implications of this profound mystery will leave the reader reeling from the sheer number of unexplained phenomenon, bizarre coincidences and mind boggling innuendo.

Another aspect of this mystery that cannot be ignored is the number of famous people involved in this ongoing mystery such as Jack Parsons, co-founder of the Jet Propulsion Laboratory, L. Ron Hubbard founder of the Church of Scientology, Aliester Crowley, the wickedest man on Earth (666), Mark Hammel, Luke Sky Walker of the Star Wars movies, Alfred Bielek, physicist, Remote Viewer and Director of PSI OPS at Montauk AF Base, Dr. John von Neumann famous physicist, Thomas Jefferson, 3rd President of the United States, Dr. John Dee alchemist and astronomer for Queen Elizabeth, Karl Haushofer, spiritual ambassador to the Far East for the Nazi's and a prominent General and leader in Hitler's rise to power and of course Duncan Cameron the operator of the infamous 'Montauk Chair'.

Montauk, America's Greatest Unknown Conspiracy really began with the Philadelphia Experiment in 1943 aboard the USS Eldridge, which was the U.S. Militaries' first publicly known attempt to create a 'stealth' field around an American warship that ostensibly would make the ship invisible to enemy radar. The USS Eldridge was stationed in the Philadelphia Navy Yard and the resulting stealth experiment was originally known as the Project Rainbow. While the attempt failed miserably, killing sailors and leaving others in a psychotic condition, it did open up new avenues of research into wormholes, star gates and quark fields.

Basically what scientists and researchers, headed by Dr. John Von Neumann, attempted to do was to create an electromagnetic field or bottle around the ship that would divert radar signals around the bottle, making it undetectable to radar waves and underwater mine fields. While the experiment was essentially successful it had some unintended side affects, such as making the entire ship invisible to the naked eye. What was learned later was that the entire ship and crew had been moved into another dimension of space-time and transported hundreds of miles away to Norfolk, VA and back again.

Needless to say, no one was prepared for this turn of events and when the ship re-materialized it left some of the sailors aboard fused into the bulkheads of the ship, much like how matter caught in a tornado can be fused together from the Hutchinson Effect which is found naturally in tornadoes. Because of these effects to human life and the fact that no one at the time really understood the results of the experiment the Project Rainbow was (publicly) 'terminated' and Von Neumann moved on to the Atomic Bomb and The Manhattan Project.

However, because of the results of the Project Rainbow, the Government at some point decided to look deeper into the subject and that research began after World War II at Montauk Point Long Island, New York. This information may very well have gone completely unnoticed and never had been discovered had Preston Nichols not received a Government grant to research telepathy, in 1971, he lived on Long Island, New York.

Originally he set out to disprove mental telepathy as he was a stern non-believer in psychic phenomenon. He learned through his research on his test subjects (psychics) that in fact telepathy not only existed but could be proved through the detection of an electromagnetic wave that was formed by the psychics themselves. His expertise was in electrical engineering and electromagnetic phenomenon, he had little interest in paranormal experience and had never even heard of the Philadelphia Experiment, of course most people hadn't until a movie was made about it in 1984. He set about learning about this electromagnetic wave and its properties and learned that it functioned much like a radio wave.

His research however lead him to conclude that a 410-420 MHz (megahertz) cycle being broadcast over the airwaves, at a specific time of the day, was effectively jamming the telepathic signals his psychics created, resulting in their inability to focus mentally or form a telepathic wave. The human mind operates on the 400 to 450 MHz cycle. Intrigued and frustrated he set out to trace this signal and found that it was coming directly form a radar dish on Montauk Air Force Base Long Island, New York.
Upon further investigation into this radar facility he learned that it was being used by the FAA (Federal Aviation Administration) for experiments with radar technology and this excuse may well have held up, except for the fact that Preston Nichols was an expert in radar technology and could see no practical use the FAA would have for using an antiquated (World War II technology) Sage Radar system to broadcast signals in 410 to 420 MHz range. However because of incredibly tight security and stonewalling by Government officials he could learn little else, until 1984 when he learned the base had been shut down.

Before we continue with the unravelling of this mystery, let's take a little time to reflect upon the exact location of Montauk Point Long Island New York as it has an intriguing history that is pertinent to this investigation.

Ley Lines

What is interesting to note is that the ley line that passes through Montauk, NY, intersects with the Bermuda Triangle.

Montauk Point, New York -
Montauk Point, Montauk, NY 11954, USA
Montauk Point as been described as an island but is actually connected to Long Island, New York.

Preston Nichols & Peter Moon

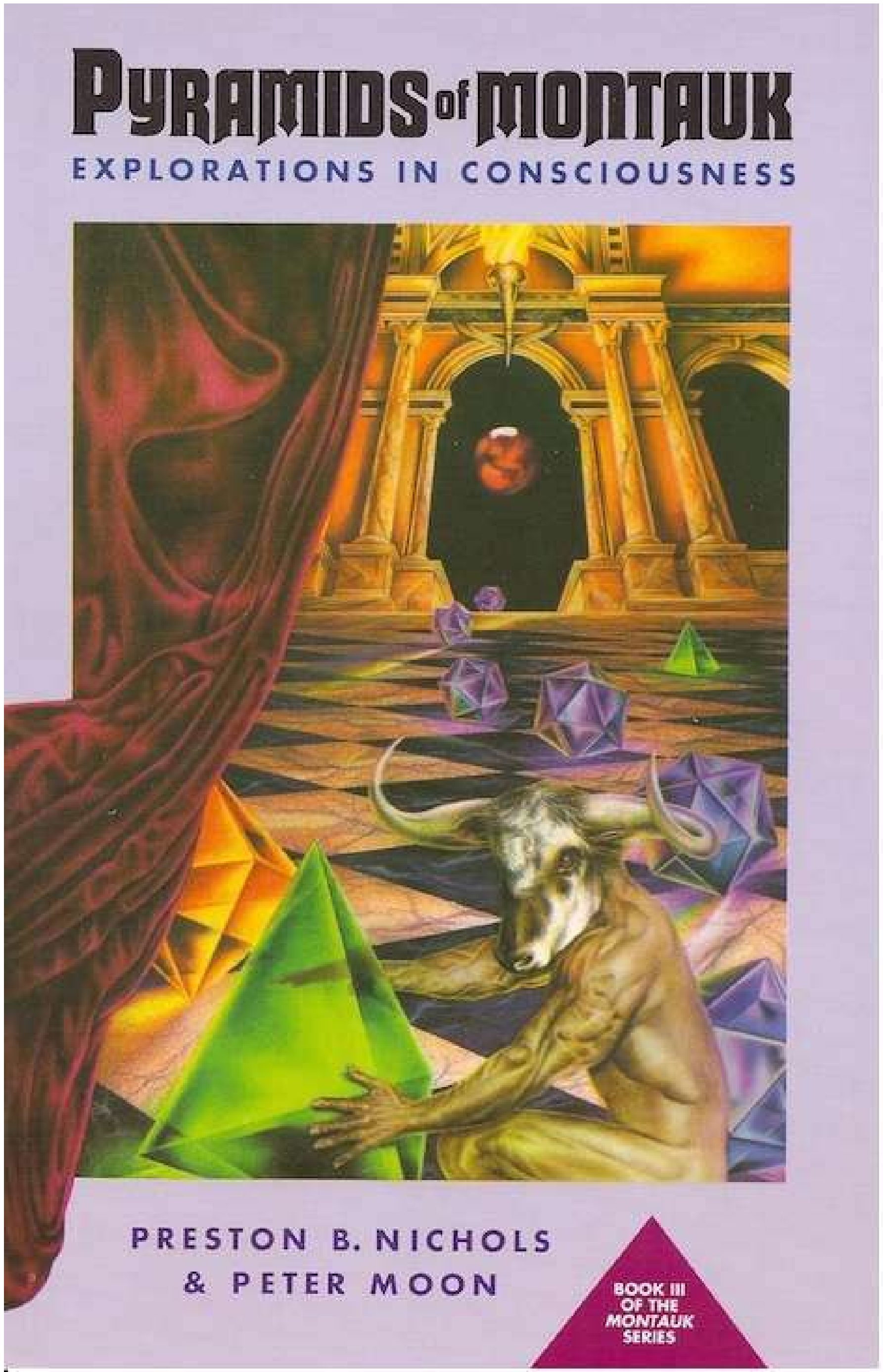
PYRAMIDS of MONTAUK
EXPLORATIONS IN CONSCIOUSNESS
PRESTON B. NICHOLS
& PETER MOON
BOOK III OF THE MONTAUK SERIES

Pyramids of Montauk: Explorations in Consciousness

Montauk Point Long Island, New York

Montauk (Island) Point, New York is located on a ley line and is part of the Earth's electro-magnetic grid. For those of you unfamiliar with the term ley line it refers to a low frequency electro-magnetic grid that covers the Earth. This grid creates geometric shapes in the form of platonic solids (the building blocks of the Universe) across the globe. Many ancient megalithic structures were built on these lines as away to harness the natural energy of the Earth, such as the Pyramids of Giza and Stonehenge.

The Native Americans that lived for thousands of years in this section of New York, the Montauk Tribe (the word Montauk means a place of farseeing), were part of a larger tribe known as the Algonquin. What is interesting is that the chiefs of this Montauk tribe were named Pharaohs and that there is documented evidence and photographs of the tops of three pyramids along the coast of Montauk Point (see Pyramids of Montauk). These pyramids have since been buried to protect their location and prevent vandals from pilfering them. Only two other civilizations in human history have used the title Pharaoh for their rulers, Atlantis and Egypt.

This tribe of Native Americans spoke an ancient language not known to many other tribes in the area, called the Vril Language. The Vril are described in myth and legend as an ancient race of man (Atlanteans) that live inside Hollow Earth and possess sophisticated technology and weapons. For more on the subject of the Vril I highly recommend a short novel written by Edward Bulwer Lytton called The Coming Race. However I would also like to point out that Nazi Germany created a secret 'spiritual' group closely linked to the SS, known as the Vril Society, which was involved in secret weapons research and technology concerned with round winged aircraft (UFOs) and anti-gravity.

Another interesting aspect of this mystery is that this Vril 'language' has also been linked to the language of the Angels, called Enochian, which was translated by the infamous alchemist Dr. John Dee, t he royal astronomer for Queen Elizabeth. Dr.John Dee was a founding member of the Rosicrucian Order, was used as a role model for characters in Shakespeare's plays, such as King Lear and has been linked to the sinking of the Spanish Armada by weather manipulation and many other myths and legends.

Adding an interesting twist to this mystery is that the 3rd President of the United States and known Freemason, Thomas Jefferson, travelled to Montuak to record, translate and preserve the Montauk language, known as Vril the language of Atlantis from the root language of Enochian, the language of the Angels. According to Jefferson upon returning to Washington the manuscript (which including an alphabet and dictionary) he had so carefully transcribed was lost in the Potomac River. There seems to be no end to the strange incidents, events and synchronicity involved in the Montauk, America's Greatest Unknown Conspiracy, including but not limited to, who actually owns the land of Montauk Point, New York.

When the Montauk Indians 'sold', (they were told it would be a renewable lease), their rights to Montauk, New York to white settlers the dates of the transactions all included three sixes (666 is the number of the Beast, which was Aleister Crowley's calling card). The first transaction took place when they 'sold' their land to the township of East Hampton on August 6th, 1660. The next deed for Indian occupation of the land was transacted on February 6th, 1661. Over the next two centuries there were many legal battles over the true ownership and rights to the land, however in 1910 New York Supreme Court ruled that the Montauk Tribe was 'legally extinct' and that the land was the property of the Federal Government.

Peter Moon

U.S.S. Eldridge

The Philadelphia Experiment, The Movie

Released in 1984 and distributed by Thorn E.M.I., one of the largest electronic firms in the United Kingdom, after only few weeks in movie theaters across the United States it was pulled at the behest of unknown U.S. Government officials. Thorn E.M.I. took the issue to court and won, getting the censorship reversed and allowing the movie to be seen by the public but the damage had been done and few movie theaters would touch it, the movie was a financial bust. I have watched the movie and must say that I would like to have those two hours of my life back.

Despite the movie receiving rave reviews by critics, it was a box office flop and only became popular after being released on video. However the synchronicity involved with the making of this movie and the characters involved in its production, distribution and story lead even a die hard skeptic to recognize that there is more to this story than meets the eye.

In 1918, Aleister Crowley (666), at the height of his magical career took a 'magical retirement' to Montauk Point Long Island, New York, what he did while he was there remains a mystery. The Crowley family name, along with the Wilson family name became business partners in the early 1900s England to form a corporation involved in the manufacturing of the first electronic instruments in Great Britain used to monitor electric pulsations.

This corporation merged with other companies in the 1920s to form Thorn Lighting, which eventually became Thorn E.M.I. in 1979. They also have an entertainment division which includes music and video publishing and distribution.

I mention the name Wilson because of the odd connection to other players in this mystery. L. Ron Hubbard's original name was Wilson, as his father was adopted at an early age and took on the name of his adopted parents. The wife of Jack Parsons (JPL co-founder), whom went by the name Cameron, her maiden name was Wilson. Both her and Hubbard worked for Naval Intelligence and were reported to have been used as spies to investigate Jack Parsons. Jack Parsons was a student of Aliester Crowley and member of the O.T.O. (Ordo Templi Orientis), Crowley was purportedly the head of this secret order.

An interesting side note is the link between Parsons, Hubbard and Cameron participating in a magic ritual known as the Babylon Working which purportedly dealt with opening a hole in the space time continuum, a rift to another dimension, which could conceivably allow Alien spaceships (UFOs) to enter our dimension. This secret magical 'working' took place in 1946, a year before the infamous Roswell UFO

crash. Jack Parsons was killed in an explosion in his laboratory in 1952 but neither his wife nor the executor of his will was able to see the body before it was cremated.
Another odd twist to this movie is the reported involvement of Mark Hammil, whom played Luke Skywalker in the original Star Wars movies. Although this information cannot be substantiated he was known to be a silent financier of the movie and childhood friends of Preston Nichols and Duncan Cameron. Nichols worked as a sound engineer on the movie The Empire Strikes Back.

Al Bielbek reported to have dealings with Hammel in 1956 in Hawaii, which would make Hammels' far older than what he appears to be. However since the movie did involve time travel and part of the Montauk mystery does seem to revolve around this concept perhaps this isn't as unlikely as it seems.
Of course there is the convoluted story of Mark Hammils' infamous 'car accident' between the time the movie Star Wars came out and The Empire Strikes Back, was made. Starwars.com had this to say about it, "A car accident in early 1977 did extensive damage to Hamill's face, and the actor underwent reconstructive surgery prior to filming The Empire Strikes Back."

After much reading and research about this movie and its connection to Montauk, the only other aspect that I find strange is that normally if a movie is a box office flop and actually loses money, a sequel is not made. However with every other bizarre aspect of Montauk, America's Greatest Unknown Conspiracy, The Philadelphia Experiment II was made and released in 1994. This movie involves an altered timeline that has Germany winning World War II and America being run as fascist military regime by the Nazis. Although, this may seem a fantastic scenario to most Americans it is not without precedence considering the information related in my article Nibiru Planet X How the Fourth Reich Plans to Rule the World After the Pole Shift.

Peter Moon

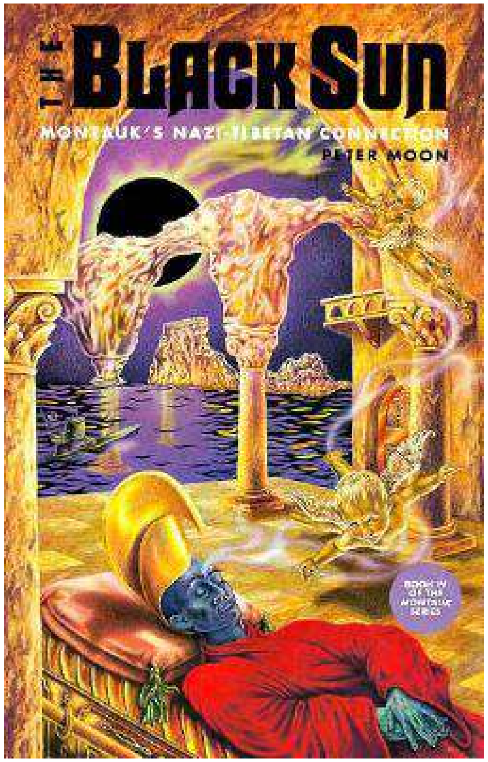
THE BLACK SUN
MONTAUK'S NAZI-TIBETAN CONNECTION
PETER MOON
BOOK IV OF THE MONTAUK SERIES

The Black Sun: Montauk's Nazi-Tibetan Connection

The Nazi Montauk Connection

The amount of synchronicity involved with the Nazi/Montuak connection is extensive and convoluted to the extreme, whole books have been written on this subject and for good reason. I will try to be brief however it is a conspiracy theorist dream come true.

According to Peter Moon in his book Pyramids of Montauk and The Black Sun, Aliester Crowley, members of the Thule Society (which were instrumental in Hitler's rise to power) and the Order of the Silver Star of the Illuminati conspired together in a magical ritual known as the Phisummum in 1923 to create a worm-hole for time travel. The Order of the Black Sun and high ranking Nazi officials were the controlling parties of the various powers involved using the Spear of Destiny as a magical power Source to travel back in time an retrieve the Holy Grail.

The result of their magical ritual was a 'time explosion' that connected to Philadelphia in 1943, Montauk in 1634 and Bannock Hill, England during the time of King Arthur and Merlin. It has long been speculated and rumored that Hitler and Nazi Germany were in league with Extra-terrestrials but the identity of these Aliens has never been determined. Some sources claim a faction of the Pleadians, others the Alpha-Draconians, while still others the Alderberans. Whatever the case maybe, they did seem to have a leg up on the rest of the world in terms of technology and scientific advancements. Speculation is that this advanced knowledge of anti-gravity, weapons, cloning and magical rituals were derived from knowledge gained by German ambassador Karl Haushofer's time spent in the Orient and initiation into the Dragon Society of Japan.

Karl Haushofer, a WW I General and some suspect co-author of Mein Kampf, was a known Bon Priest, Bon is an ancient religion predating Buddha in Tibet. The basic principals of the Bon Religion deals with the control of linear time and is grounded in a spiritual connection to nature and Mother Earth.

A fundamental belief of the followers of Bon was that in their heaven, to which they referred as 'Sridpa Yesang', there were three deities who controlled the world. These were Dagpa, Salba and Shepa, one of whom controlled the past, another the present and the third controlled the future. It was they who held the key to Bon doctrines and Tonpa Shenrab Miwo claimed to be their earthly incarnation endowed with the power to control the current time.

The connection to Montauk is interesting from the perspective that a submarine base located at Ford Pond Bay (Long Island, NY) was in operation during the WW II within miles of Montauk Point and many eye-witness testimony from the locals puts fleets of Nazi submarines off the coast during the War. It has long been speculated that an underground/water submarine base existed in the vicinity of Stony Hill, Amagansett, NY by the locals during World War II.

During the 1930s the largest population of the Aryan Brotherhood in the USA existed in Yaphank, New York, known as Camp Siegfried (after the Norse God). This group promoted an eugenics program of sterilization of inferior races and held what the locals called Bund Meetings and Nazi rallies. Coincidentally the infamous Brookhaven Laboratories is located right next door to Camp Siegfried.

Further speculation has it that late in the War as the Nazi regime began to fall, a clandestine mission was conducted by a Nazi submarine in which gold and other loot from the War was buried at Montauk Point to be used by Nazi scientist that would be brought to America through Operation Paperclip. It has long been rumored that the Nazi's exchanged three nuclear bombs with America late in the War for safe passage out of Germany by high ranking Nazi officials before and after the War, including but not limited to Hitler himself.

What complicates the matter is that no American doctors or military officials were able to examine Hitler's remains, as Russia was the first to arrive at Hitler's bunker and cordoned off the area. Stalin had reported many times in the press of the day, that he believed that Hitler's body was a fake and that his supposed death was a Nazi trick. Incidentally, Hitler's autopsy reports weren't released until March 5, 1953, eight years after the end of the War and right after Stalin's death, presumably so that he could not shed any light on the details.

Regardless of whether or not Hitler escaped and lived on after World War II and there have been many books related to this subject, the focus of this article Montauk, America's Greatest Unknown Conspiracy is about the known information that involves Montauk Island, New York and how it affects the course of human history.

Montauk Conspiracy
Brookhaven Laboratory

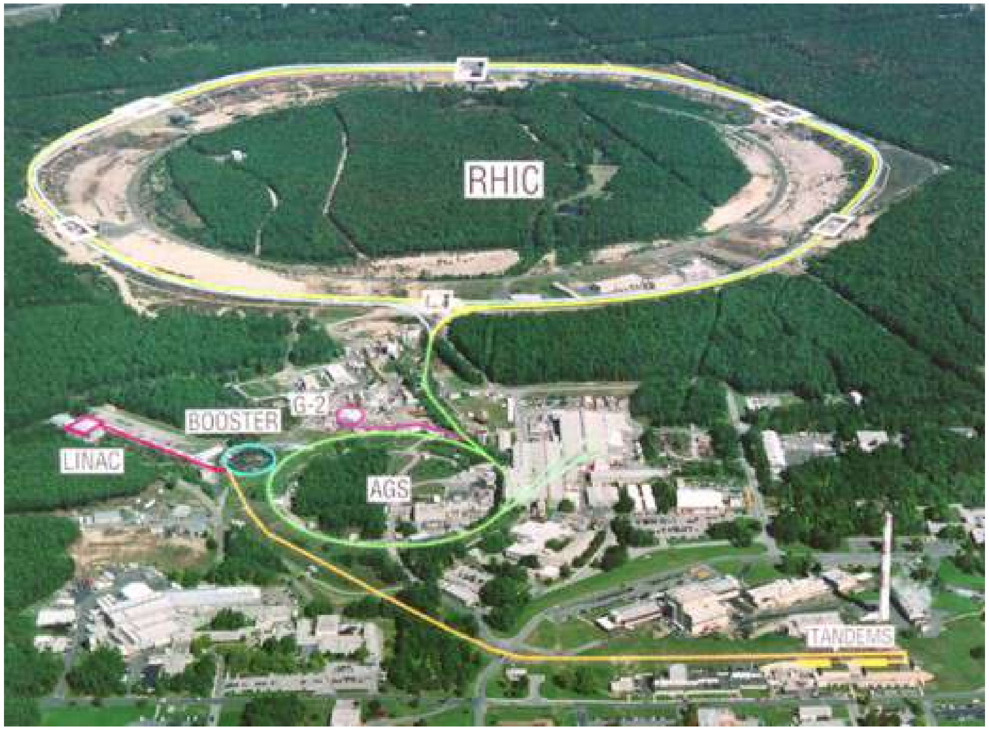
RHIC
G-2
BOOSTER
LINAC
AGS
TANDEMS

The Relativistic Heavy Ion Collider (RHIC) under construction.
Brookhaven Laboratory

The completed Relativistic Heavy Ion Collider (RHIC) on Long Island, New York.

Montauk Conspiracy

It seems clear to this writer that there is an abundance of secret intrigue, disinformation, synchronicity and clandestine operations revolving around the area known as Montauk Point, New York. From early historical knowledge of colonial settlers right up to present day, the amount of unusual events, secret societal involvement, military activity and local gossip shows that this region of Long Island, New York has more than its fair share of intrigue and speculation.

What we do know about Montauk, America's Greatest Unknown Conspiracy is that in 1986, electronics expert Preston Nichols announced in Chicago, to the dismay of the U.S. intelligence community, his involvement in the 'black operation' known as Operation Phoenix which might be better known as the Montauk Project. Originally this project was designed to utilize the technology and experiments by Dr. Wilhelm Reich. An Austrian scientist known for his discovery of 'orgone' energy in the 1940s, which is more closely related to cosmic energy and the Newtonian concept of ether physics.

Although Reich was able to prove the existence of orgone energy in laboratory experiments and published his findings in scientific journals his ideas were scoffed at and ridiculed by his peers. Among his claims was that orgone energy could cure cancer and that it directly affected the weather through what he called dead orgone energy or DOR. His research lead him to conclude that storms such as hurricanes, violent thunderstorms and other bad weather were caused by high levels of DOR (dead orgone energy). However he also learned that living orgone energy could be found in humans as well, with active happy people having more orgone energy than depressed lethargic people.
He developed a device that could break up DOR in bad weather by connecting it to balloons and releasing them into the weather systems. His experiments showed that he could actually dissipate bad weather using his invention. However when he presented this invention to the U.S. Government his research and technology was confiscated citing National Security issues. Later all of his research papers and laboratory materials were destroyed by the Food and Drug Administration to prevent his discoveries of orgone energy from going public.
The U.S. Government used his invention, coupled with a device used in the 1920s to collect weather data, known as the metrograph. This device recorded weather related phenomenon using a balloon/parachute system. The balloon was designed to burst after reaching a prescribed altitude, deploying a parachute to deliver it safely to the ground. Citizens were given $5 dollars (a lot of money back then) to return the device to the authorities.

The U.S. Government created an invention they called a radiosonde based on Reich's invention and the metrograph. However what was known publicly as a device for measuring and detecting weather elements, upon further research has been revealed as a transmitter of weather elements, specifically DOR. In the fifties as many as 200 radiosonde's were being used in the atmosphere above America, daily. The device was eventually scrapped after more advanced weather manipulation technology was created, such as HAARP.
So, initially Operation Phoenix began as weather manipulation but it soon took on a much wider scope, when it was combined with Project Rainbow, ostensibly to keep the latter shielded from the public. Dr. von Neumann would head up the entire project headquartered at Brookhaven National Laboratories on Long Island. He built the first vacuum tube computer at Princeton University to help in this research into theoretical physics and space/tome manipulation.

What he learned initially as the reasons concerning the failure of the Philadelphia Experiment were due to humans having a time reference point, which is altered significantly when surrounded by an electro-magnetic field that essentially displaces whatever is inside it in the space/time continuum or alternative reality. His idea was to use a computer to create a false reality that humans could then use to synchronize too when undergoing any further electro-magnetic space/time experiments. What he found through experimentation on human test subjects (the general public and selected individuals) was that he could affectively control what people thought using the technology he had developed from 1948 to 1967 during Project Phoenix.

Up until his final report the U.S. Congress had funded this project but upon learning the implications of this technology they ordered the project to be shut down, perhaps fearing their own minds would be tampered with and understandably so. However considering the implications for weapons research and population control it would be grossly naïve to think that this project would end. Enter the Military Industrial Complex and the real intrigue behind Montauk, America's Greatest Unknown Conspiracy.

Amplitron

With names of devices right out of science fiction the technology was used to create time travel and star gates.

Thyratron

A Thyratron was one of the devices used with the Sage Radar to help with the time travel experiments.

Montauk Air Force Base

The Montauk Air Force Base (MAFB) was 'officially' reactivated in 1971 and was the site of a mothballed Sage Radar instillation. Although Congress was no longer officially involved in funding the new Phoenix II Project, money from Brookhaven Laboratories was funneled into the project for research and MAFB was used for the physical research and for security purposes. A combination of civilian and military personnel were used to work on how the technology from the original project could be used for military purposes.

Essentially they used the Sage Radar antenna as a super radiosonde transmitter to broadcast on a 425-450MHz radio frequency using varying pulsations and amplitude to affect different biological functions and thought control. Initially they used the base personal as their first guinea pigs but eventually began broadcasting to the general public in the nearby area. They found that they could even affect the behavior of animals in the region, as well.

To monitor the results of their ongoing experiments they would use 'sleeper agents' infused into the local population to report results and spread disinformation. Some of the side affects of their experiments was a wide range of abnormalities with electronic devices such as radios, televisions and appliances, including turning off or on cars in the area. They compiled a huge data base using different amplitudes, pulse modulations and frequencies and their results. This lead to the creation of a control panel with different settings that could be used to create instant results. After three or four years they could control the crime rate in the local population, the general mood and even some physical attributes, such as sexual urges or dysfunction.

However all this was seen as only a general application to a wide physical range, what was needed was a refinement of their techniques, so that individuals could be targeting and more control could be exerted. Learning to refine this technology is what created the ability of the people behind Montauk, America's Greatest Unknown Conspiracy to open wormholes and time vortexes.

Using an idea derived and based on applications taken from reversed engineered UFO technology what is known as the Montuak Chair was developed. ITT had been working on this idea since the 50s, in which they could read a persons mind and display it on a computer monitor using a similar type chair. The Phoenix II Project technicians wanted to use this technology to create a virtual reality machine that would enable one person to broadcast into the minds of their test subjects. Ostensibly this would be to allow anyone involved in entering an alternate reality to maintain their zero time reference point and would be the first step in making time travel possible, that was the theory anyway.
Many problems had to be worked out but oddly enough some of the technical issues were solved using Tesla technology form the 30s, known as Delta T (time) coil structures. Also Al Bielek got involved to solve some the computer interface issues concerning the pyschics being used in the chairs operation.

Duncan Cameron

Delta T Antenna

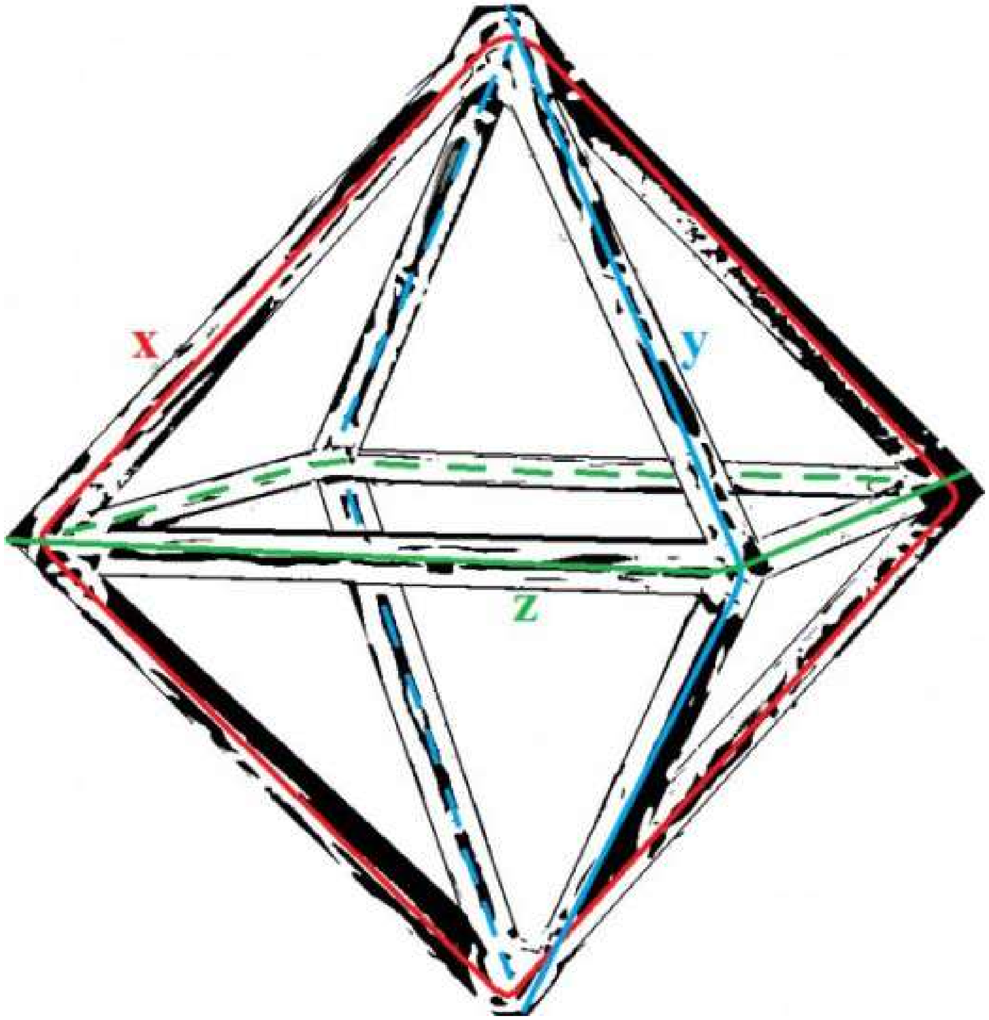

The Delta T Antenna uses the platonic solid shape (building block of the Universe) of a octahedra to achieve it specialized function.

The Zero Time Concept

The concept of zero time can be easily understood if we picture the center of the Universe like a large merry-go-round with many levels or dimensions. At the center of this merry-go-round would be a control booth that operates or exists in a zero time capsule and all linear time moves outward from that point.

The reason I mention this is because for the technology to work properly all aspects of the components including the operator of the chair had to be harmonically synchronized to the zero time reference. To do this required that they not only modify the existing technology but also create their own 'white noise' generator as a way of harmonizing all the various technology together. My understanding of this concept is limited but I basically see it as a constant wave signal that allows calibration to be achieved.

However what needs to be pointed out or recognized is that the engineers and scientist involved in the Phoenix II Project or Montauk, America's Greatest Unknown Conspiracy did not have the technical sophistication to invent or create a device that could generate a zero point, white noise oscillator back in the late 70s. However because they did have access to certain information supplied by 'off world' intelligence's with their own nefarious agenda, they were able to 'borrow' this technology without having a complete understanding of how it worked or was engineered.

Persistent rumors circulated that Aliens from both the Sirius and Orion star systems were involved in the project. Whether their intentions were honorable or not, is beside the point, the high degree of technological assistance and theoretical concepts employed in this operation did seem to take a decidedly Alien bent. To achieve the successful time experiments desired required technology that simply was not available in the late 70s. What was known as the Orion Delta T Antenna was supplied by an unknown source fully functional, as well as a fully functional Tesla Zero Point Time Reference Generator, supposedly invented in the 20s by Tesla himself.

These components allowed the engineers to create a COHO or coherent 'white noise' oscillator that could be linked to all the other devices and achieve a harmonic synchronization with the chair and operator, which would allow the desired time experiments to succeed.

Preston Nichols & Peter Moon

MONTAUK REVISITED
ADVENTURES IN SYNCHRONICITY
PRESTON B. NICHOLS
& PETER MOON
BOOK II OF THE MONTAUK SERIES

Montauk Revisited: Adventures in Synchronicity

Duncan Cameron

Duncan Cameron was 'the' main Montauk Chair operator with the most success and was used extensively for most of the experiments in time. Duncan's revelations in the books written by Peter Moon and Preston Nichols were paramount in deciphering Montauk, America's Greatest Unknown Conspiracy. Duncan is a psychic, employed and trained by either the NSA or CIA in remote viewing techniques, depending on which source your willing to accept and his history is full of time travel innuendo, mysterious circumstances and aspects of mind control. However he seems to have a vast knowledge of technology and American history that only researchers and writers of conspiracy theory have been able to find.

The history of the Cameron Clan is extensive and far too broad to be written about here but suffice to say the name Cameron is old, with connections to ancient royal bloodlines, and can be found throughout American history as well. It has been reported that their was a Cameron (perhaps named Wilson) on board the

USS Eldridge during the Philadelphia Experiment and some speculation is that it could very well have been the same man who became the main Montauk Chair operator in the 1980s.

Whatever the case maybe to this enigmatic character, once he was synchronized to the Montauk Chair, his psychic abilities were increased to a high degree and he was able to produce startling results that defy conventional wisdom. Among his outstanding results, if they're to be accepted as real were opening time portals, materializing objects by mere thought, feeling through the five senses of other humans no matter where they were on the planet or off, traveling through time and space, projecting thoughts into other human minds to the point of controlling their actions and many other similar experiments.

Once the technological bugs were worked out of the system and a high degree of harmonic fidelity was achieved with Duncan in 'The Chair' it was just a matter of fine tuning the synchronization of the technology and testing the limits of his psychic abilities. However as more and more successful 'experiments' were achieved, the program began to take on a more sinister agenda.

At first, this was just experiments that involved controlling inanimate objects and electronics such as switching TV channels in peoples homes, turning them on and off and causing them to produce static but then it turned to destroying them, breaking windows, moving objects and destroying furniture and appliances. Then it turned even more deadly by using Cameron to influence peoples thoughts by directing them to commit crimes sprees and engage in odd sexual behavior. The truly sinister aspect of this was that the minds of those he engaged in these experiments would have no memory of the actual events, it was very deep form of hypnotic mind control.

One result of these experiments that showed some potential that the Military Industrial Complex wished to explore was the phenomenon involving displaced time. It was found that Cameron could think of an event taking place or an object materializing and then that event or object would appear before he thought of it or sometimes many hours or days after he thought of it, even when he wasn't 'plugged into the Chair'. This lead to creating programs based on the recorded data supplied by these experiments, so that eventually they didn't even need Duncan to be 'plugged in' to produce the same results.

Because of the nature of these experiments and technology required a high degree of 'secrecy' an underground complex was built beneath Montauk Air Force Base at Camp Hero, New York. However the construction of this base did not go unnoticed by the local population as tons of earth and large boulders began appearing along the coastlines of Montauk Point without any known means of arrival. The odd behavior of the locals and clandestine comings and goings of personnel began to become transparent and reported, this lead to a sophisticated disinformation campaign that still exists today.

Time Portal

This excellent illustration exemplifies my concept of what a Duncan Cameron created time portal may have looked like.

Time Portal

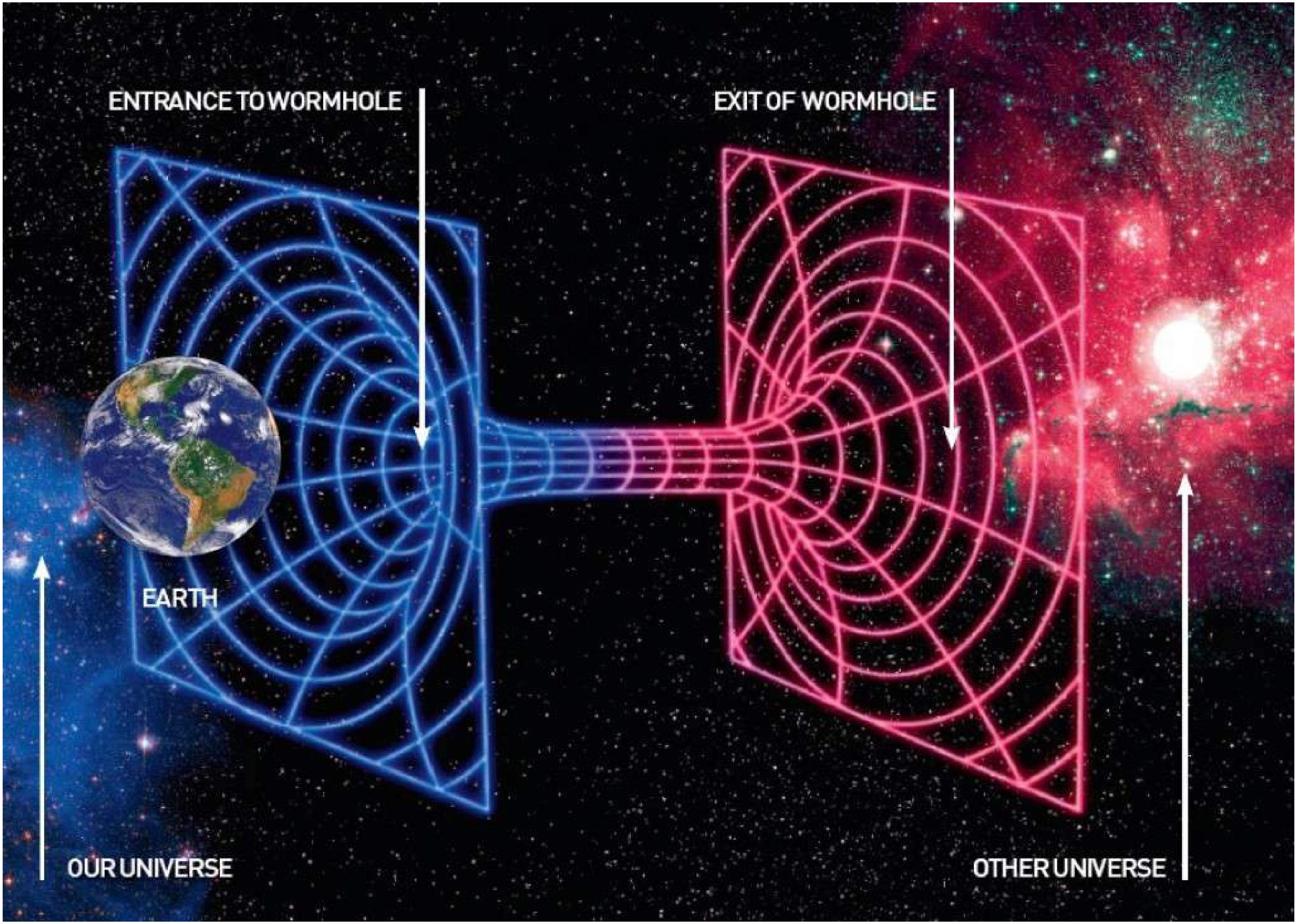
ENTRANCE TO WORMHOLE
EXIT OF WORMHOLE
EARTH
OUR UNIVERSE
OTHER UNIVERSE

From the article How to Build a time Machine by Stephen Hawkings, well worth the read.

Preton Nichols & Peter Moon

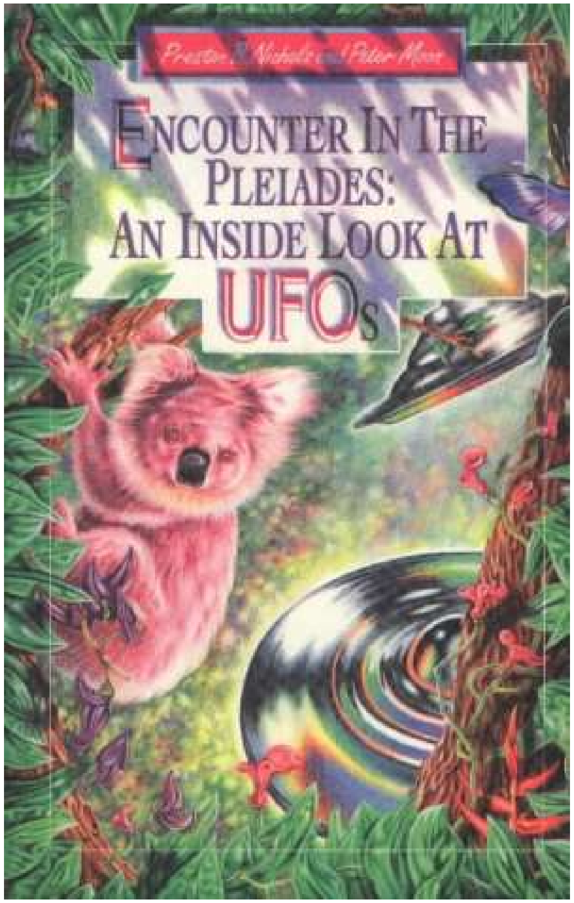
Preston B. Nichols and Peter Moon
ENCOUNTER IN THE PLEIADES: AN INSIDE LOOK AT UFOs

Encounter in the Pleiades: An Inside Look at UFOs

Experiments In Time

Once the advanced Alien technology had been integrated into the system and tested in 1980, the experiments began to concentrate on space/time and less on mind control. However because of the nefarious mind control experiments and time travel implications a secret group within the Phoenix Project decided to implement a 'fail safe procedure' just in case things got out of control, in hindsight this turned out to be a wise precaution.

The time experiments began by having Duncan concentrate on a specific date and year while the zero point time generator kept him located in his own time.

After many experiments and testing he was able to produce a window in time or what was labelled a 'time portal' in which viewers were able to look into the future and the past. Eventually he was able to open a dimensional doorway that linked two time periods together in which humans could actually walk through and return . . . sometimes!
However since this was in essence untested waters and there was never a shortage of homeless people in New York City, the first guinea pigs used were the indigent and poor, after all no one would miss them and it wouldn't arouse suspicion if they were lost in time.

Before these actual time experiments took place however secrecy was amped up to a new level and all non-essential personnel were removed, this included the military. This was done by reporting to the Military Industrial Complex hierarchy that the time experiments were a failure. This is a common tactic used by The Powers That Be whenever they steal technology from inventors or scientist, they merely publicly term the invention or technology a failure, hire experts to write scientific essays on why the technology cannot work and promote it through the Fake Stream Media.

This tactic can also be used to make people disappear by faking their deaths and having the media report it as fact. As long as the body cannot be identified or witnessed, none is the wiser. This is why the suspicious death of Jack Parsons has been questioned as well as other notable instances such Osama Bin laden and the Seal Team that dispatched him . . . dead men tell no tales.

Once all the personnel were removed and replaced by other agents, the only remaining people were those that had specialized knowledge that couldn't be easily replaced. However certain mind control techniques were used on these remaining personnel so that they would have no memory of what they were working on. This lead to the phenomenon of 'missing time' which to most people makes them question their own sanity. Having experienced this phenomenon myself I can tell you it is quite disconcerting and very difficult to relate to anyone else. The best way to describe it would be to waking up fully clothed and have no idea of how you got there, with your last memory of being somewhere else (location) in time.

All of this information is known because Preston Nichols was one of the specialized engineers that worked on and created many of the technical devices used in the experiments and he had become good friends with Duncan Cameron. Duncan had a direct mental link with the time events taking place and would often have an out-of-body experience while plugged into the chair. Of course their daily memory would be erased and replaced with other memories so that they 'thought' they were going to work in one place and returning home, completely unaware of what they were really doing.

After all these precautions were in place, a new name was used for the ongoing experiments, it was now called the Phoenix Project III. They began cataloguing different eras of time and space, by sampling the environment using automated devices, such as robots, drones and sensors. The tunnel or time portal was described very much like the stargate animation shown in the popular TV series Stargate, as it twisted and turned surrounded by wormhole like lights and elements.

As long as the power remained on and no 'glitches' occurred during the transmission a portal would remain open on the other end and the robotic sensors could return with their collected data. After 'using' and testing many involuntary (homeless) subjects, sometimes by force, with various degrees of success, the project began using time operatives to conduct missions. These operatives would have advanced video equipment, sensors and cameras that would report in real time what they experienced while on their missions, so that their controllers could see what they were seeing and evaluate the data without having to rely on individual reports. A vast catalogue of videos were created using this technology.

Besides the normal operatives, children were used as well and according to Duncan and Nichols and every one of them was sent to a specific date and time in the future (6037 A.D.) as an initiation process, before their first mission. Why this was done remains one of the many truly bizarre mysteries revolving around the Montauk, America's Greatest Unknown Conspiracy.

Both Cameron and Nichols estimate that as many as 10,000 people were sent 200 to 300 years into the future never to return. However much of the experiments in time revolved around both World War I and World War II. A secondary vortex was also used in conjunction with the first that would allow the controllers to beam, through sophisticated computer technology a television signal into the past or future that could be viewed by any with the same technology.

One of the most interesting aspects of this technology was the reported time travel to other planets in the future as well as the past and the exploration of the Pyramids on Mars or Face of Cydonia complex.

Mars Mission

In a nutshell, The Powers That Be, the true hierarchy behind Montauk, America's Greatest Unknown Conspiracy knew beyond a shadow of a doubt that the Face of Cydonia Pyramid Complex had been created by an ancient sophisticated race of Aliens and wanted any technology that still existed there. According to both Duncan and Preston this information was known and considered a foregone conclusion.

To reveal to anyone how this information was obtained would require divulging their 'ultra-secret' space program that had been in existence since at least the early 60s, they used a time tested form of disinformation through ridicule, censorship and silence pertaining to any information that might reveal that the Face of Cydonia and surrounding area, was in fact a large pyramid complex.

However, because The Powers That Be assume the public is stupid and can't help themselves from rubbing our noses in our own stupidity, they allowed a docu-drama called Alternative 3 to be aired once on British Television in the late 70s that described how the United States and Russia had set up jointly run military prison bases on both the Moon and Mars. These bases were initially designed to allow the rich elite, 'themselves', to flee the Earth in case of any catastrophic Earth Changes that would take place in the near future . . . such as a Polar Shift.

The bases also served as means of exploring both planets or moon in hopes of recovering any Alien technology left behind by advanced races. The base on Mars was located near the Face of Cydonia and is the main reason the Mars Rover was not sent there, later. However in the early 80s after twenty years of trying to get into the pyramids on Mars using various mining techniques and explosives with no success, they decided the Phoenix Project would solve their problems and gain entry.

Although both Duncan and Preston vaguely remember planning a mission to Mars they can't recall much else about it. Duncan was able to remember penetrating the main pyramid 125,000 years into the past and 'seeing' operating technology which was part of what he called "The Solar System Defense System". The Powers That Be were desperate to have this automated system turned off but he could not recall if they were successful. Some speculation is that this 'Defense System' may be preventing The Powers That Be from leaving the planet en masse and could be the reason no 'publicly known' manned missions to other planets or the Moon have been accomplished since the Apollo mission in the early 70s.

Philip K. Dick

It is interesting to note that two of Philip K. Dick's novels, The Man in the High Castle and We Can Remember It For You Wholesale have inspired the movies Philadelphia Experiment II and Total Recall, respectively, both pertaining too the Montuak Project.

While researching this information I have learned that 11 of his 44 novels have been made into movies, all of which deal with alternative realities. Of course, Do Androids Dream of Electric Sheep (Blade Runner) is the most famous of these.

Duncan's Beast

This is one of the photos of Duncan Cameron's subconscious Beast.

The Montauk Project Gets Shut Down

On August 5, 1983 the operators of the Montauk Project were given orders to turn on the chair, transmitter and time machine equipment and leaving it running until further notice. After a week of running continuously, August 12, 1983, exactly forty years to the day from the Philadelphia Experiment, August 12, 1943, the equipment lurched into high gear and became synchronized with another power source. When this happened a time portal opened back into the past and the USS Eldridge could clearly be seen at the other end.
Due to the nature of their research and experiments the operators of the Montauk Project had learned that the Earth went through a cycle of twenty year biorhythms which made opening time portals easier when one used these cycles for time travel. Even Duncan Cameron could be seen on board the Eldridge and some of the operators began to get worried that some boundaries had been crossed and that a paradox could upset the nature of time itself.

Fortunately a select few operators that still remained from the original crew had created a fail -safe countermeasure if the project needed to be terminated without prejudice. Duncan Cameron was given the code words, 'The Time Is Now' and Cameron let loose a beast from his subconscious mind that showed up literally as huge beast somewhere on the AF Base. This monster, perhaps powered by the huge amounts of energy coursing through the time portal, began destroying everything in its path. Photographs of this beast were taken showing it actually existed.

The supervisor of the operation fearing the experiment was out of control and not knowing what a time explosion could do ordered the technicians to shut down the equipment, which they did, to no avail. The two time periods seemed to be fused together in loop with energy cascading out of control. It was decided that both transmitters had to be destroyed for the link to be severed, so someone would have to go back in time and destroy the one on the USS Eldridge, this would be one way ticket, there was no coming back.
Preston Nichols and the Supervisor began pulling wires on all the equipment, using an acetylene torch, Preston was able to cut through live wires and still the

transmitter stayed on. A permanent time portal had been created using the USS Eldridge and the Montauk Project. The operators panicked and began smashing equipment and cutting wires, still the portal remained. Nothing they tried seemed to work, they ran to the transformer farm above ground and cut all the power cables leading underground and still the portal remained.

They cut all the power cables leading from the power station in town to the AF Base, and the lights of the base shut off and even the computer link-up yet the portal remained open, the energy seemed to be growing. They ran to the transmitter station and cut all the power cables there, the lights went out but the transmitter remained operating, they were getting desperate.

Finally Preston took his acetylene torch and began cutting the actual (hot) transmitter machine into pieces, he cut the conduits, he cut the cabinets, he cut apart all the moving parts. The machine groaned once like a dying beast and stopped, the time portal collapsed and the beast from Duncan's subconscious mind disappeared, it was over.

From that day forward, August 12, 1983, forty years from the Philadelphia Experiment failure, the Montuak Project shut down . . . or did it? The personnel were debriefed, their memories erased, the base was abandoned and fenced off. Eventually some memories did re-surface but more questions remain than answers.

Time Travel

Wormholes through the 4th dimension could be used for more than just 'time travel' by linking planets and destinations in fixed portals.

Conclusions

While the amount of truly bizarre synchronicity (coincidence) involved in Montauk, America's Greatest Unknown Conspiracy can only be fully appreciated by reading Preston Nichols and

Peter Moon's books, I hope that I have shown the reader that there is more to this story than meets the eye. Without the courage and dedication shown by Peter

Moon, Duncan Cameron, Preston Nichols and even Stewart Swerdlow to reveal this mystery to America and the world, I doubt if any of this information would have seen the light of day.
All that being said, I think the most significant information outlined here is that, some secret group (The Powers That Be) have conspired to create technology that would allow them to travel through time and space. We may never know their true intentions or even any of the results of their experiments but it seems clear to this writer that even our own Government (Congress) sanctioned these experiments up to a point.

What should be carefully thought about and considered is that if, in fact, any changes were made in the past due to these 'experiments in time' it could very well have already affected the lives of every man, woman and child living on this planet today.
This should not be taken lightly as we may already be living in an altered timeline, without even realizing what affects it may lead too nor anyway of knowing or recognizing the profound changes it could have to our very existence.

Without a doubt the decisions of a select few could very well have plunged humanity into a future that was never meant to be, while keeping the billions of people that live on this planet completely in the dark. However since there is no point in worrying about events we have no control over the main point of this article is awareness, consideration and planting the seed of knowledge.

Putting the nefarious experiments aside for a moment, it seems clear that a form of mind control was practiced on American people, with startling results. The implications of Nazi collusion in technology, should also raise an eyebrow or two, however I barely touched on that aspect in this article. The technology and science alone that resulted from these experiments, if shared with the world, could open avenues of research that could potentially solve some of the worlds energy problems and even perhaps clear up some of the more profound mysteries of our ancient history.

Consider for a moment a device that would allow children of all ages and even adults to explore our ancient past without ever leaving the confines of their homes. Think of the implications towards a better education of not only our history but the history of our entire Universe. Think of using a wormhole for travel to other parts of the planet or planets within our solar system as easy as walking through a door.

When mankind begins to create a new paradigm built on giving, unconditional love and cooperation we could use this kind of technology to vastly improve the lives of every man, woman and child on the planet. While this idea may seem naïve to the reader, consider how far the human race could be in terms of spiritual, scientific and technological advancements if we could temper our competitive nature, get over our differences and embrace our common heritage.

People say that War and violence is human nature but I disagree, children have to be taught this kind of behavior. Raise a child or even a pet just on love, with no harsh words, no physical discipline and no references to violent behavior through books, TV or movies and you will see a being that knows nothing but various degrees of love.
While this article Montauk, America's Greatest Unknown Conspiracy may have revolved around the selfish decisions of a select few humans to control and influence the masses, perhaps the Greatest Unknown Conspiracy should be why Unconditional Love is not recognized as the most powerful force in the Universe.

THE PHILADELPHIA PROJECT PROJECT RAINBOW AND THE USS ELDRIDGE

In July 1943, the destroyer U.S.S. Eldridge pulled into the Delaware Bay area for a United States Naval experiment that involved the task of making the ship invisible. The project's official name is Project Rainbow, but was nicknamed and more commonly known as the Philadelphia Experiment. Much has been written and speculated about the legendary experiment into invisibilty, but sorting fact from fiction is a near impossible task, especially with the recent influx of misinformation and deliberate disinformation that has been spread by those connected to the U.S. Intelligence community and professional skeptics.

There is much controversy over what exactly happened, but one thing is for sure. For some reason, soon after the test was completed, a massive blanket of secrecy and denial was placed over what happened in Delaware Bay. It is thought that a huge scientific breakthrough was made, and the ship was accurately transported over space and time, disappearing in Philadelphia, Pennsylvania and reappearing in Norfolk, Virginia. Whatever did happen is still not known, but different theories are discussed below: The "Official" Navy Record The Navy admits that the U.S.S. Eldridge took part in an experiment that involved wrapping wire around the hull of the destroyer in an attempt to cancel out the magnetic fields of the metal on the ship. This is known as degaussing. This would render the ship "invisible" to underwater magnetic mines that rely on proximity sensors to trigger the detonation.

These sensors operate by detecting magnetic fields around ships. Without the magnetic field, the ship would be able to pass through regions mined with these sensors, invisible to enemy mines, but not to radar or vision. The Navy's report is very plausible, and doesn't mention any exotic results or circumstances. But could this just be a believable account to drop the interest by the general public, leaving only the true minority of investigators in doubt? Physical Invisibility Some scientists have developed the theory that the Navy was working on a way to make the ship invisible to vision. However, it didn't involve warping space time or any complex task of a similar nature. This theory suggests that the Eldridge was equipped with high frequency generators that would heat up the surrounding air to cause a mirage, making the vessel invisible. This phenomenon is naturally occurring, and there have been cases where entire islands have disappeared from view in the right weather conditions. The high frequency generator would heat up the surrounding air and the water (creating a green-colored fog that was said to have engulfed the ship), causing a mirage to form, concealing the ship from view. The generator would also account for the sickness (physical and mental) of the crew after the experiment.

A high freqency generator can cause serious harm to a person's wellbeing, especially at close range. This is more plausible than the degaussing theory, and would also explain the crew's sick condition as a result of the test. The main problem with these theories though, is that it doesn't explain how the U.S.S. Eldridge was seen in Norfolk, Virginia by the civilian crew of the SS Andrew Furuseth, when the ship disappeared from view in Philadelphia in a space of only about fifteen minutes. There are also details such as crewmen being fused to the hull of the ship and some not even reappearing. Transported across space and time? The most interesting theory about the Philadelphia Experiment is that the destroyer did in fact disappear and was teleported across space and time. Supposedly, there was a great number of ingenious scientists (including Tesla and Einstein) that were taking part in the experiment. However, Nikola Tesla was supposed to dead at the time of the Naval experiment. The theory is that light has to be bent around the ship to make it invisible. To accomplish this, the Navy wrapped the ship's circumference in wire and passed a measured current through it. This caused a huge oscillating magnet to form a magnetic field around the ship, not only bending the light, but space and time as well. The physics of the experiment are reminiscent of Einstein's Unified Field Theory that once you bend light, you are also unwittingly bending space and time as well.

The first time this experiment was undertaken, the ship didn't completely disappear, and an imprint of the hull could be seen sitting in the water. The second time, the ship totally disappeared in a green fog and was sighted in Norfolk, Virginia. A haunting fact is that when the ship reappeared, the crew were all in a state of disorientation. Some were mentally ill, while other crewmen didn't even return. There were also crewmen that returned embedded in the hull. Later accounts arose about the crewmen, including a former crew member who was involved in a bar fight, and all the participants froze in time, as reported by a local newspaper! There were also accounts of people who were on the ship, spontaneously combusting. The mystery remains It is still not known what happened that day in 1943, mainly due to the lack of witnesses coming forth who served aboard the Eldridge. There is also no documentation available to the public which details Project Rainbow. It may have simply been a degaussing experiment. But how did the destroyer appear seconds later in Virginia? Its possible the answer will never be known, but the mystery may be solved when scientists rediscover what happened in Delaware Bay.

Camp Hero: Mind Control, Time Travel & Reptilians

CAMP HERO:
MIND CONTROL,
TIME TRAVEL &
REPTILIANS

The following information is an update to a post from January 2012:

Montauk, N.Y. — like Area 51, written about in Annie Jacobsen’s book Area 51: An Uncensored History of America's Top Secret Military Basethe Montauk facility has been the center of an otherworldly conspiracy theory for decades.

Believers say that people were kidnapped at a U.S. Air Force base and subjected to mind control and time travel experiments. And extraterrestrials somehow had a hand in it all.

But nobody has been able to actually prove these allegations. And all that’s left of the Montauk facility — currently a state park — are the above-ground remnants of the original Air Force base. According to a document issued by the Air Force Historical Studies office, the Montauk base, known as Camp Hero, was decommissioned in the early 1980s.

Montauk is a small seaside resort town on the tip of Long Island that draws vacationers to its shores every year. Camp Hero, located a short distance outside of Montauk, has roots as far back as the Revolutionary War, during which it was used to test military cannons. Later, during World War II, Camp Hero served as a coastal defense installation against any possible Nazi intrusions into America.

Montauk Chronicles follows the story of three men, Alfred Bielek, Stewart Swerdlow and Preston Nichols, who claim that Camp Hero ended up as an underground site for scientific atrocities.

Bielek, a retired electrical engineer, maintains he was part of the mysterious Philadelphia Experiment, where in 1943, the U.S. Navy allegedly tried to make a small destroyer undetectable to radar. The test ended in disastrous results, including the ship disappearing from the Philadelphia Navy yard and — purportedly — traveling through time.

The Philadelphia Experiment: Invisibility Time Travel and Mind Control - The Shocking Truth- Al Bielek

According to Bielek's story, he was transported ahead in time from Philadelphia, and extraterrestrials were responsible for the technology used in the Philadelphia Experiment. He also maintains he was recruited in 1970 to work on mind control and time travel projects at the Montauk facility.

Swerdlow's story involves being kidnapped as a teenager from his Long Island, N.Y., home, taken to the Montauk base, and subjected to a variety of experiments.

"Beatings, a lot of torture, electrical shock, burials, near-drownings," Swerdlow said. "They'd bring you to the point of death, and then they would save you, and the person doing this would be your rescuer or god, and would say, 'I'm the one that saved you and remember that.' And that became your handler — your programmer."

Swerdlow recalls being part of some horrific experiments while at the Montauk facility.

Montauk: The Alien Connection- Stewart Swerdlow

"The walls were very damp, oozing water, so it appeared to be deep underground or even underwater. I was always on this cold, hard table. Sometimes there'd be other people around, either my age or older, and electrodes were put into me and injections."

All three men claim to have seen extraterrestrials while at the underground Camp Hero facility.

"Well, there were quite a number of aliens at Montauk," said Bielek. "Some were there on a semi-permanent basis. A lot of them were just visitors that came in and looked at what they wanted to see and went back home.

"There were little grays there, which I suspected were degenerated humans from out of the future. Large gray aliens (which are a different species) were also at Montauk, and they were highly intelligent."

Nichols, like Bielek, was an electrical engineer, who says he worked with Bielek in the mind control and psychic aspects of the Montauk Project.

The Montauk Project: Experiments in Time- Preston B. Nichols

"There were definitely alien beings at Montauk," Nichols said. "We had the little grays and the larger grays as well as a variety of reptilian beings. The large grays didn't want anything to do with me because they couldn't reach me telepathically. When I entered a room they would leave.

"They were the strangest thing that I ever saw. At that point, I was beginning to doubt my own sanity."

And Swerdlow also saw an alien presence at Montauk. "Most of the time my interaction was with human beings, but I did come into close contact with alien beings.

“I did see, occasionally, intelligent reptilian humanoid beings as well as gray aliens who were once human beings but were physically altered as a result of degeneration and radiation toxins in their system. Most of them communicated with mental telepathy.”

Garetano shot much of his film at the actual site of Camp Hero. - Lee Speigel - Montauk Project: New Film Takes On Allegations Of Mind Control, Time Travel And Alien Encounters At Military Base - THP

Montauk Mind Control Victim Interview

An exclusive interview with Andy Pero, survivor of a secret mind control project aimed at creating the super soldier with psionic abilities.

By Eve Frances Lorgen, M.A -

Andy Pero was unwillingly subjected to a mind control program off-shoot the Monarch Project-a trauma based mind control program similar to what Cathy O' Brien describes in her autobiography by Mark Philips, "Tranceformation of America".

Andy Pero claims to have been created through a combination of genetic manipulation, trauma based mind control and Silva Mind Control training resulting in superhuman feats and psionic abilities. In addition to this, Mr. Pero alleges to have been unwittingly used in covert assassinations as a Manchurian Candidate and also in missions connected to the Montauk Project, such as time travel.

Andy Pero first went public with his story in August of 1998 as a special guest of Preston Nichols while at a lecture series hosted by Bob Eure of Philadelphia, PA. I was so astounded with Andy's testimony, that I interviewed him personally so that his story could be shared with readers of Unknown Magazine. For those who are unfamiliar with me, I have worked for ten years investigating and counseling UFO experiencers, abductees and occasionally survivors of various forms of anomalous trauma, such as ritual abuse and mind control. Because of my background in alien abductions, paranormal, the occult and anomalous trauma, I recognized Andy as an extraordinary survivor of trauma based mind control, one whose story should not remain hidden:

EL: Andy, where are you from originally and when did your training begin?

AP: I was born in Fallon, Nevada in November of 1969. My father was a Lieutenant Commander at the Fallon Naval Air Station in Nevada at the time. I believe my first experiences began by the time I was two and a half years old. At five years old my family lived in Munich, Germany and I remember attending two distinct Kindergarten classes, on in German that was located on or near McGraw Army Troop Army Base, and another Kindergarten in English.

I have flashback memories of torture sessions as early as age two where I was subjected to shock treatments with needles inserted in my ears and genitalia. What they (the controllers/handlers) do is use extreme trauma to separate the mind from the body. They do this to split the mind into several parts that can later be programmed to do things-like superhuman feats. Basically, they disconnect all of the normal human brain impulses, emotions and whatever is normally present, then reprogram in new ones. They do this to remove all fear, so that when they program in new abilities, you will not have the instinctual, "but I can't, I'm afraid..." response. They program you to do, not think. In essence, they treat and program your mind as if it was a computer.

EL: Who are the people running these secret programs like the Monarch and Montauk Projects, and where do they conduct their activities?

AP: Our own government. One of my handlers-who tried to brainwash me into believing that he was my father--was basically a Nazi. I nicknamed him Adolph because he is very Aryan looking.

After WW2 many Nazi scientists were brought over to the US through our own Intelligence community who formerly worked in mind control research in the concentration camps. I believe they are a remnant of the Nazi party working in conjunction with different branches of the US government and the military.

Much of my training and torture sessions were done right at the University of Rochester, NY. They used to take me to a private room in the attic or top floor of the library. The big lab where I had most of the programming done to me was at the Rome National Air Base in Rome, NY. I have also been used in the Montauk chair while at Camp Hero, Montauk, L.I. and also in a chair in Atlanta, GA at Dobbins Air Force Base.

The Montauk facility located in Camp Hero, L.I. was reopened by the Department of the Navy and is active to this day.

Basically there are underground bases located across the country. Each one has a different function and specialty. There is a facility in Rochester, NY, Paramus, NJ, Dobbins Air Force Base in Atlanta, GA, Camp Hero in Montauk, L.I. and many others. In the Paramus, NJ facility is where they do the sex programming for many of the female agents. For the most part, 99% of them are involuntary sex agents and couriers who were raised in various mind control projects.

EL: What were some of the things the mind control programmers made you do?

AP: I remember my mother enrolling me in a Silva Mind Control course when I was ten. It is a type of self-hypnosis where you learn to go to different levels of your mind to do self-healing, relaxation, visualization, bending spoons with the power of your mind, things like that. This is also the time when I saw two peculiar military men, who would periodically visit me throughout my childhood, especially sporting events. I did exceptionally well at Silva Mind Control (and sports), so much so that the two military men approached me and said, "I hear you have some very special abilities. Someday you will work for us." At the time I didn't understand, but later the two men figured prominently in all the training I had up all the way through my college years. Of course I didn't realize this until my memories came back. I did very well at sports and demonstrated a genius IQ, but I stuttered so badly until I was in my twenties, that everyone just thought I was a dumb jock. The stuttering was due to all the electroshock treatments.

I recall being subjected to electroshock numerous times, being placed in an enclosed water tank and told to breathe underwater. Consequently, I drowned several times and would revive because of my self-created healing pool in my mind. I learned to create my own mental laboratory through Silva Mind Control, replete with a super powerful healing pool that could heal any injury. I was forced to endure suffering beyond belief to accomplish superhuman feats of strength and survival.

EL: Tell me more about the various superhuman abilities you discovered you had.

AP: The trainers would put me into a trance whereby I'd be told I could jump off a ladder of minimal height. In my mind, I believed that I was only jumping off a footstool or a short ledge. In actuality, I was progressively led to jump higher and higher heights to the point of successfully jumping off buildings and even out of planes without a parachute.

When under mind control, I could do whatever I was instructed to do as long as I believed them. I would do 200 or more push ups in perfect form, sometimes as long as a half hour straight, or until I was told to stop. I weight lifted 545 lbs.on a barbell and squats while attending college and much more (500-1500 lb.) while I was under the mind control in the training laboratory not far from Rochester, NY.

In my jumps, it was as if I was falling like a cat, calm and cool with no fear, focusing like a laser only on the landing.

EL: Are there others like yourself who have had similar training and experiences?

AP: I think there are thousands of other children and people who are used and many end up missing. Many die in the process. They are considered expendable. The ones who can survive the most rigorous programming live on to become what are known as the Montauk boys. Duncan Cameron is such a person. In general, the boys and young men are trained to be couriers, assassins and super soldiers and even psychic killers. The women are mostly used as sex agents. Cathy O' Brien's story is an example.

EL: What kinds of special missions or time travel experiments do you remember taking part in?

AP: There were times I remember being at the small Rochester airport, being picked up in an F-16 fighter plane and ending up in some southwestern desert terrain either for special military training and obstacle courses or assassination missions. I remember seeing other men in black t-shirts and pants on the same training course at one time. Other times I found myself on some aircraft carrier out at sea. After the mission they would take me back and I would return to college after a couple of days or even hours not remembering what even happened. All of these missions were done between the time I was 15 years old and until 1988-1992 under deep hypnotic programming with the memories erased, until of some of my memories started to return.

I was also taught to handle various types of guns, loading, cleaning, special characteristics of each model and firing at exact targets repeatedly and perfectly without thinking, just doing.

EL: Tell me more about your involvement in the Montauk Project.

AP: I was involved in an off shoot of the Montauk Project called the Montauk chair. Basically the Montauk chair was developed to give the human and spirit a zero point of reference to facilitate time travel. What the chair essentially does is separate the mind from the body. The chair operator's thoughts and vibrational energy is picked up by umbrella looking antennas above the chair, sent to a computer, over to a processor, then amplified several hundred times. The information is sent to a network of free energy crystals arranged in a circle. Then whatever thoughts were amplified, i.e., a time period, a wormhole would open up in the room. The wormhole was as large as 16-18 feet across and even large enough for a truck to go through.

In Camp Hero, Montauk, the location is the cross hairs of the earth's biorhythms and is the point on earth where time travel is most easily accomplished when earth is the point of origin.

My part of the Montauk chair project was to use my focused visualization skills to think of specific time points in history that were assigned to me. The chair is connected to a sophisticated computer system and thought amplifier, such that when a thought or time period is visualized, the computer simulates a time portal based on that point in time. A series of time portals are done until a library of time periods and portals are accumulated.

The chair is also used to amplify extremely focused thoughts to create three-dimensional materializations based on the operators thoughts. Preston Nichols wrote several books on the Montauk Project and describes an incident in which Duncan Cameron created a monster while in the Montauk chair.

One such time travel mission was called Project Southern Cross. It was used to win WW2 in favor of the allies. What the US government did using time travel was to go back in time to the 1940's to help us win the war. We would deliver communication devices, weapons and technologies made out of 1940's parts. These would be delivered to the 1940's along with a complete set of drawings on how to make them out of 1940's parts. I took part in several of these deliveries, one time I was sent to Germany and another time to England. I was not allowed to speak to anyone, other than deliver my parcel and quickly return back to our time. And this was all done under deep hypnotic programming, so I didn't have a lot of freedom to explore. I was gone no longer than two hours for the deliveries.

EL: Have you ever encountered any extraterrestrials or seen any aliens in any of your underground base memories?

AP: From what I understand the Department of the Navy made an agreement with the alien Greys to exchange technologies for human women and children to conduct horrific breeding experiments. This is what is going on right now in an underground base not far from Miami, Florida. One of my most disturbing memories is being escorted down a hallway in this underground lab and seeing cages of chickenwire fencing with women and children screaming for help. I have seen Grey aliens (the 4-foot tall ones with large black eyes) and also 7-foot tall reptilian beings in some of my experiences. I have been told that I have many children from alien breeding experiments. I have had abductions with the Greys also.

On one occasion I was introduced to a Reptilian being while in an underground base sometime in 1989-90. At first I saw a 7-foot tall human Ayran looking man. He walks towards me and I notice that his image phases out as if something interfered with an energy field. He does something to a device on his belt and tells me, "OK, I'll show you." He then pushes some button and then I see his image change into a 7-foot tall lizard like creature who looked like he weighed over 400 lb.

I was introduced to a group of beings that claimed to be a sub group of the Illuminati. They were made up of a group of about 40-50 men, all 6' 2" or taller, white, and of an indeterminate older age, averaging 70, but I suspect much older considering the technology they have access to. They convened in a boardroom that I believe was in an underground base, perhaps Camp Hero, Montauk. They sat around a large football shaped table made of wood. Each high backed chair had a light at the tip and each sitting had a TV type of screen and control panel that came out >from the table. There was a huge screen located in the front of the room as well. I have reason to believe that not all of the Illuminati are human, but in actuality are alien beings that appear to look human.

EL: Who or what groups do you think are running the show here in the US or even on Earth? How do the aliens fit into all of this?

AP: It is very complicated, but basically there are four main versions or groups of the New World Order, all with slightly varying agendas. The first group comprises the US and Western Europe (i.e., England, Germany, France and the US). The second group is Eastern Europe fused with Western Europe and the US. The third group is the countries of the Orient, and the fourth is the Middle Eastern-Saudia Arabian elements and all the oil money. There are also off world groups (aliens and future human Nazi's types called Zardanians, Reptilians, Greys and Dracos) who also have a hand in the NWO plans. For the time being, the old money groups of Europe and England have teamed up with die hard Nazi factions, Jews and Americans forming the group that has been always been known as the Illuminati and affiliated secret societies. This large international group controls the media, police force, military, etc. The Illuminati is also part of the NWO plan, which includes members such as George Bush, Prince Charles and other high ranking people.

As for the aliens, I think most of the alien Greys are either under the authority of or in collaboration with the Reptilians, Dracos and the Illuminati, based on some of my experiences and memories that are still returning to me. Some Grey groups are in conflict with the Reptilian/Draco agenda and are doing their own thing (human/grey hybridization programs) so to speak. Their agenda gets into heavy spiritual matters having to do with acquiring an emotional body and an ability to ascend to higher dimensions. They are using humans to accomplish this via the human/alien hybrids. Humans have the key in terms of having a threefold essence or unity of being, physical, emotional and spiritual. Other aliens have lost this unity of being (because of the fall from grace), thus inhibiting them from ascending to higher dimensions, or evolving to higher levels. This gets back to the original Luciferic rebellion, fallen angels and the various conflicts between species.

EL: When did you start remembering your memories? How did or can you break the programming they install?

AP: Some of my first memories surfaced while I was living with a roommate in Atlanta, GA back in 1996. We went to the bowling alley and I saw a large sign in the shape of a man dressed like a Rhino advertising, "Bowl with Rhino balls." This triggered me into remembering that "Rhino" is the name my Nazi handlers called me. My roommate overheard me say outloud, "I'm Rhino-they called me Rhino." He got excited and immediately asked me, "Where did you hear that?" He had a grave look of concern on his face and kept questioning me. By his reaction, I realized later that all along he was one of my hired programmers and handlers. The initial triggered memory that surfaced was a flashback of me being electrocuted and shocked and screaming in pain in one of my numerous torture sessions.

Then after that first memory, other memories started leaking through. It was like my mind was dammed up and the subconscious memories were starting to leak through. My mind controller programmers deliberately built a wall around my memories so that I wouldn't remember. When I started recalling bits and pieces, I'd write notes to myself and hide them. Then I'd find my notes, after I had forgotten everything because my roommate was catching on that I was remembering. So he would repeatedly put me in trances, erase my memories and I'd forget what I had just remembered. But then everything came crashing down, because I deliberately hid notes and pretended that I didn't know to my roommate. I had thought he was a friend and it turned out he was just hired to be another controller and handler to keep me in the program.

I recall that in the past I did not even remember my dreams for ten years. But now things are coming back. I am working on purifying the body to help clear my mind and get my abilities back.

EL: How do they keep control overall these people who have been and are still in mind control projects?

AP: The essence of their control tactics is of splitting the mind at a young age. This entails a process of trauma, torture, degradation and humiliation of the subject. The process is different for each individual, but in essence they go to your subconscious mind and find out what the most personal parts of you are. Then they essentially use this against you to break you down and rip your mind apart. What happens is this: An individual who possesses special mental abilities is selected at a very young age. (Very often they are from military families or from multigenerational Satanic ritual abuse families-EL) Then they are subjected to various stages of terror and horror to begin to separate them from their own minds. It is a process of progressive dissociation and compartmentalization of the mind and personality.

They control your mind through various complicated hypnosis techniques and commands until they create the perfect Manchurian Candidate.(See The Greenbaum Speech by D.C. Hammond) They program you to do something, then erase the memory. Basically they program your mind as if it were a computer. Once they install the magic word, all they have to do is say it and you're under their control. This happens from an early age so it is easier to control them if they are conditioned and dissociated into several personalities.

Another method of control is that they send in special people in your lives, handlers and controllers who pose as friends, teachers, "deprogrammers", roommates, lovers, etc. who are actually operatives for the mind control projects. Some are sleeper operatives who don't even know they are being used to a certain extent. It makes it hard to trust anyone at this point.

EL: Is there anything else you can say about these projects or advice to others who have gone through similar experiences, like how they can be set free from its influence?

AP: For those who want to get out of the influence of the programs of mind control, you must have a strong fighting instinct to do what is right regardless of what others are doing. Your only cage is the one you build for yourself. You must connect with your highest self to break the evil that is all around us, and not succumb to doing evil just because everyone else is doing it. Retain your fundamental sense of right and wrong.

It all goes back to the basic conflict basic good and evil. It is like the dark forces in power of the planet at this time are trying to cheat their way into heaven or higher dimensions. - Eve Frances Lorgen, M.A

Back To Contents

It is even said that under the perfect conditions, the device offers its user the ability to foresee what was going to happen in the near future.

However, Father Ernetti remained secretive and stated he was not at liberty to reveal further details about the Chronovisor.

Father Ernetti did however reveal that the Chronovisor worked by '…processing residual electromagnetic radiation left over by numerous processes…'

According to numerous reports, Father Ernetti used the Chronovisor to witness important historical events, with the most notable being the crucifixion of Christ.

However, Father Ernetti revealed that thanks to this invention had managed to witness the destruction of Sodom and Gomorrah, and other major historical events such as the founding of Rome in 753 B.C. Also, thanks to the Chronovisor, Father Ernetti was able to recompose the missing work "Thyestes" , written by Ennio Quinto and represented in Rome in 169 A.C. in its original version, and the original text of the the two stone tablets written by God (Exodus 24:12) which were given to Moses on Mount Sinai, apart from witnessing the likes of Napoleon, Roman philosopher Cicero and other great and momentous historical and biblical episodes.

French priest François Charles Antoine Brune was one of the first to hear about the Chronovisor.

According to Father François, he met Father Ernetti in the 1960's when the two were traveling across the Grand Canal of Venice. As the two were experts in ancient languages, they began to talk about the Bible and its interpretation through the years.

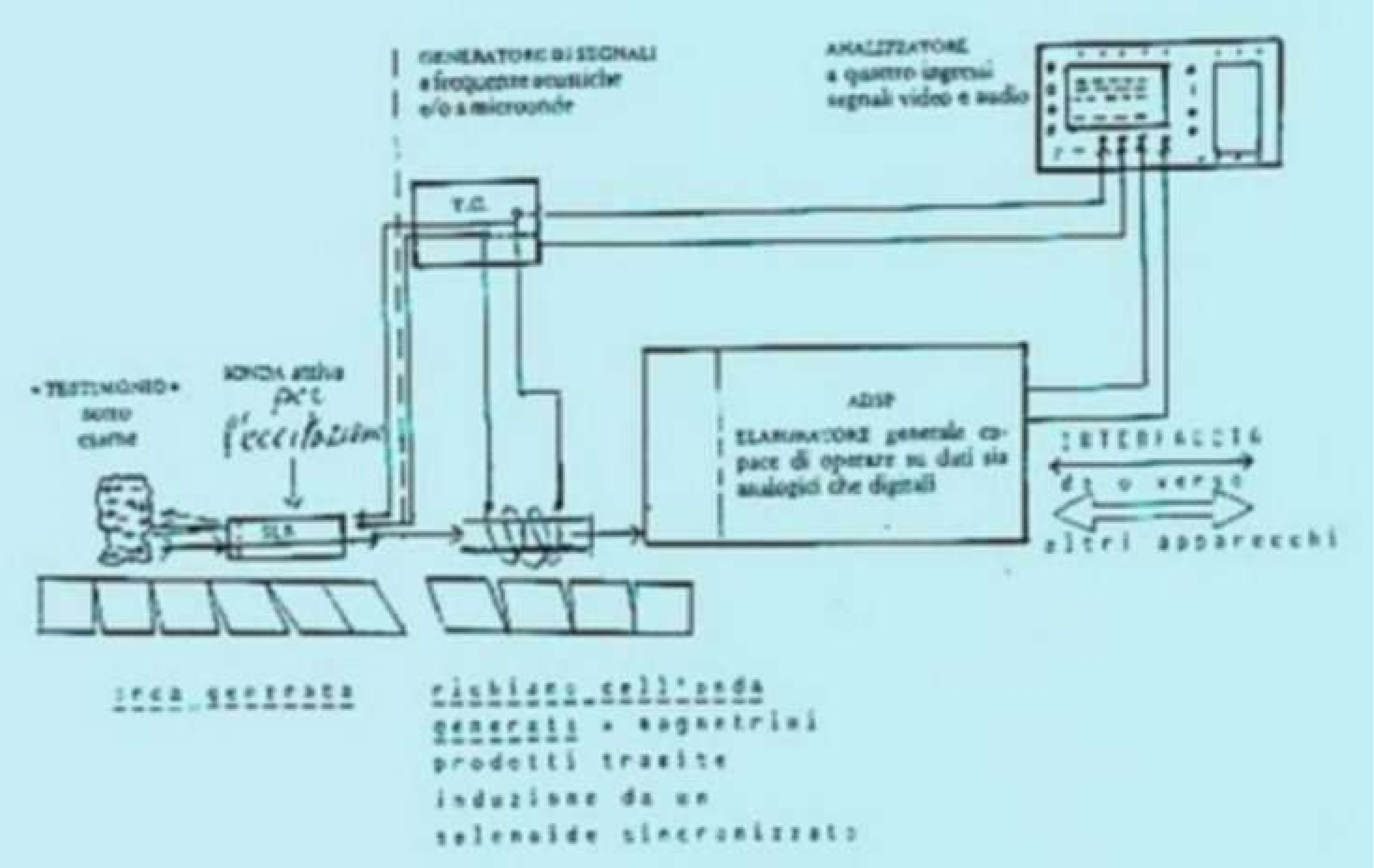
GENERATORE DI SEGNALI
a frequenze acustiche
e/o a microonde
ANALIZZATORE
a quattro ingressi
segnali video e audio
T.C.
• TESTIMONIO •
sotto
esame
SONDA attiva
per
l'eccitazione
ADSP
ELABORATORE generale capace di operare su dati sia analogici che digitali
INTERFACCIA
da o verso
altri apparecchi
area generata
richiamo dell'onda
generata e magnetrici
prodotti tramite
induzione da un
solenoide sincronizzato

Supposedly, these are the blueprints of the Chronovisor.

Father Brune that was very intrigued when the Father Ernetti revealed that there was a machine that could answer all questions related to the Bible.

When Father François asked about the Machine, Father Ernetti described the Chronovisor, saying that it was device that functioned just like a T.V., but instead of receiving transmissions from local stations, the chronovisor was able to tune into the past and allow the viewer to “see and hear” events that had occurred centuries before.

Ernetti told Brune that the machine worked by detecting images and sounds that humanity had ‘created’ which were “floating” in space.

Donald Marshall On Megadeth's "The World Needs A Hero" Album Cover

Empowerment by Virtue of Golden Truth
Human Cloning: Specifically, REM Driven Human Cloning
FULL DISCLOSURE

Donald Marshall

29 October 2015

https://www.facebook.com/donald.marshall.148
http://donaldmarshall.proboards.com/
Vincent Eastwood Interview
https://www.youtube.com/watch?v=M_1UiFeV5Jg&ab

Contents

List of Images 5

Introduction 6

Whistle Blowing: Qualifying Disclosure 6

The Serious Crime Act 2015 7

Fundamentals 8

Key Figures 9

  Phil Schneider 9

  George Green 11

  Aaron Russo 13

    Income Tax 13

    Radio Frequency Identification (RFID) Microchips 14

    Habeas Corpus 15

    Fraudulent Manipulation of Election Results 15

    The War on Terrorism is the War on Your Freedom 16

    Whoever Makes the Money Makes the Rules 16

    Russo's Message to Mankind 16

    Let's Consolidate our Problems 17

  Aaron Russo: Reflections and Warnings 17

    The Falsehood of September 11th 2001 17

    The New World Order Agenda 19

    Women's Liberation from the Perspective of the Elite 20

    America is a Republic 20

    Depopulation 21

    Borrowing Money from Private Banks causes Inflation and Debt 21

    The Deception of the Council of Foreign Relations (CFR) 23

    Combining America, Mexico and Canada into one Country 24

    Russo's Vision on How to Bring These People Down 25

The Trans-humanist / Post Human Agenda 26

The Phases of Sleep and Rapid Eye Movement (R.E.M) Sleep 27

REM Sleep 28

Current Advances in Technology 28

  Memory Suppression Technologies 29

  Mind-Voice Technology 29

  H.A.A.R.P. Technology 29

  H.A.A.R.P. and Weather Control 30

  H.A.A.R.P. and Mind Control 31

  RFID Technology 32

RFID Technology and the Medical Profession 32
Further Implications of RFID Technology 33
Mind Control: MK Ultra Technology 33
Human Cloning 35
Cloning Centre and Cloning Technology 36
Drip Feed Disclosure and Evaluative Conditioning 36
Drip Feed Disclosure 36
Evaluative Conditioning 37
Consciousness Transfer 38
Disclosure: Save the Victims through your Diligence 40
Words of Inspiration: Personalized from a Theoretical Physicist –Thomas Campbell 41
The Science of Hidden Consciousness 41
Disclosure: Donald Marshall's Message to the World 42
The Two Types of Cloning Techniques 42
The Different Grades of Clones: "Mark 1" to "Mark 4" Clones 43
Donald Marshall's Dilemma with "Mark 2" Sleep Driven REM Clones 43
Consciousness Transfer Happens When the Original Reaches REM Sleep 43
Donald Marshall is an Original as he sits at Home Typing this Disclosure 44
When Donald Marshall's REM Driven Cloned Experiences Began 45
The Science of REM Driven Cloning Began in 1945 46
Help Donald Marshall STOP REM Driven Cloning NOW! 47
Do Not Panic, Riot, or Cause Chaos. This is Very Important. 47
Spread this Disclosure Document until it reaches the Armed Forces 47
Donald Marshall is NOT "The New Age Saviour". Nobody wants to be a 'saviour' 48
How the Nostradamus Prophecies shapes the lives of the Illuminati 49
HIV/AIDS has been cured; so have many forms of cancer 49
The Behaviour of High Profile People at the Cloning Centre 50
The Ring Leaders of the Cloning Centre 50
The Royal Family of England 53
Help Donald Marshall by Sharing this Disclosure with Your Close Friends and Family 57
The Deplorable Nature of the Illuminati 59
What Happened to Britney Spears? 60
NEVER EVER SELL YOUR SOUL 62
In This Physical World Selling Your Soul Is Selling Your REM Driven Clone 62
Fake Alien Abductions conducted through REM Driven Cloning Technology 64
I am THE Spy for Humanity against the Corruption of the Illuminati 65
The Threat to Humanity when the H.A.A.R.P Grid is Complete 66
REM Driven Cloning is the Most Terrible Nightmare Situation Ever 68

The Meek Shall Inherit the Earth 70
Be Wary of: Shills; Trolls; Disinformation Agents and Double Agents 72
Contact Information 73
Interviews 74
Links to Donald Marshall Interviews 76
Vincent Eastwood 76
Jeanice Barcelo 76
Lisa Phillips (CFR) 76
Radio Presenters –Contact Donald Marshall 77
Professionals who understand "Consciousness Transfer" –Contact Donald Marshall 77
Email 77
How to Learn More about Donald Marshall's Whistle Blowing Disclosures 77
How Much Do You Know About Post-humanism / Trans-humanism? 78
This IS your ONLY chance to end these monsters. There is a deadline! 79
Do NOT Be Afraid to Help. We Outnumber the Illuminati by 1,000,000:1 80
Noteworthy Frequently Asked Question 81
Donald Marshall's Mission and Vision on How to Bring Down The Illuminati 81
Empowerment by Virtue of Golden Truth 83
Appendices 84
Appendix A: Thien Thanh Thi Nguyen (Tila Tequila) Transcript 84
Appendix B: MK Ultra 87
Appendix C: The Illuminati 89
References 99
Legislation 111

List of Images

Image 1: Stages of sleep.. 27
Image 2: The ionosphere's relationship to earth.. 30
Image 3: Megadeth Album Cover: *The World Needs a Hero* (2001).. 58
Image 4: Donald Marshall’s original body compared with his REM Driven Duplicate Clone........... 58
Image 5: Britney Spears Japanimation: Britney Spears grown as a duplicate clone.......................... 61
Image 6: Britney Spears Japanimation: Britney Spears walks into the Cloning Centre..................... 61
Image 7: Britney Spears Japanimation: The Cloning Centre depicted from an aerial view................ 61
Image 8: Possible H.A.A.R.P. Locations. 67

Introduction

Just read it. Every word in this document is true, and it is the most important document to ever reach the internet. This document is written to empower good people of the world against the tyranny which exists all around us in our world today. Effectively, the document presented here discloses concealed information about the extent to which medicine, science, and technology has advanced in our **current** era. Moreover, I am disclosing this information because I have been extensively abused, and I am victim of **current** advances in medicine, science, and technology, which I describe in detail below. There are many, many, other victims of abuse, including myself, and therefore it is our plea that you read this document with an open mind and you investigate **ALL** statements which may initially irk you and appear 'suspect' to begin with (the circumstantial evidence is available all around you); because as a victim of abuse, I can honestly say, there is nothing worse than experiencing abuse, and those who have the power to protect you from your abusers, ignore you.

I will now discuss what qualifies as whistle blowing disclosure. After detailing the criteria which qualifies for whistle blowing disclosure, I will detail aspects of the Serious Crime Act, and preface a foundation which defines, explains, and references key information for the reader to understand as well as explore the extent to which medicine, science, and technology has advanced today; all for the express purpose of compelling the reader that medicine, science, and technology, has advanced at an incredible rate; that **current** advances in medicine, science and technology are withheld from public knowledge; and finally these advancements in medicine, science, and technology, have been and are **currently** used to commit **serious unspeakable** crimes against humanity and our environment here on earth. Moreover, it is my aim to present this information as coherently as possible to assist the reader to save me, and save many, many others who experience abuse daily resulting from concealed advances in technology.

Whistle Blowing: Qualifying Disclosure

In the United Kingdom under the Public Interest Disclosure Act (1998) (PIDA) section 43B, a "qualifying disclosure" (whistle blowing disclosure) implies "any disclosure of information which in the reasonable belief of the worker, tends to show one of the following:

a) That a criminal offence has been committed, is being committed or is likely to be committed,
b) That a person has failed, is failing or is likely to fail to comply with any legal obligation to which he is subject
c) That a miscarriage of justice has occurred, is occurring or is likely to occur,
d) That the health or safety of any individual has been, is being or is likely to be endangered,
e) That the environment has been, is being, or is likely to be damaged, or
f) That information tending to show any matter falling within any one of the preceding paragraphs has been, or is likely to be deliberately concealed"

Furthermore, under the Enterprise and Regulatory Reform Act (ERRA) 2013 section 17 (*The new 'public interest' test)* is added for whistle blowing disclosures, amending section 43B of the Employment Rights Act (ERA) 1996 and therefore now reads as such (new words underlined):

"(1) In this part a 'qualifying disclosure' means any disclosure of information which, in the reasonable belief of the worker making the disclosure, *is made in the public interest and* tends to show one or more of the following-- [criminal offence, breach of legal obligation, etc]." (Halliday 2013, p. 2).

The following information is presented **with the express intent** that it is made in the public interest, and meets the all criteria set out in section 43B (a-f) of PIDA as well as the 'public interest test', ERRA section 17. Furthermore, in order for the reader to perceive the credence of my statements, so that (s)he has impetus to research the validity of my statements and not dismiss it as folly supplied by an internet troll who has nothing better to do than waste peoples' time, I will cite the Serious Crime Act 2015 here for the express purpose of convincing the reader that all statements made in this document and my original document are indeed FACTUAL! THIS IS NO HOAX. THIS IS NO JOKE. This is simply what is happening in the world today.

The Serious Crime Act 2015

The Serious Crime Act 2015 (UK) section 41, 3ZA, amends section 3A of the Computer Misuse Act 1990 which **previously** read [under Computer Misuse Offences]:

[F1 3A Making, supplying or obtaining articles for use in offence under section 1 or 3

1) [That] A person is guilty of an offence if he makes, adapts, supplies or offers to supply any article intending it to be used to commit, or to assist in the commission of, an offence under section 1 or 3.

To: [The Serious Crime Act 2015, section 41, 3ZA]

3ZA Unauthorised acts causing, or creating risk of, serious damage

(1) A person is guilty of an offence if—

(a) the person does any unauthorised act in relation to a computer;

(2) Damage is of a "material kind" for the purposes of this section if it is—

(a) damage to human welfare in any place;

(b) damage to the environment of any place;

(c) damage to the economy of any country; or

(d) damage to the national security of any country.

3) For the purposes of subsection (2)(a) an act causes damage to human welfare only if it causes—

(a) loss to human life;

(b) human illness or injury;

(4) It is immaterial for the purposes of subsection (2) whether or not an act causing damage—

(a) does so directly;

(b) is the only or main cause of the damage.

(5) In this section—

(c) a reference to a country includes a reference to a territory, and to any place in, or part or region of, a country or territory.

(6) A person guilty of an offence under this section is (unless subsection (7) applies) liable, on conviction on indictment, to imprisonment for a term not exceeding 14 years, or to a fine, or to both.

(7) Where an offence under this section is committed as a result of an act causing or creating a significant risk of—

(a) serious damage to human welfare of the kind mentioned in subsection (3)(a) or (3)(b), or

(b) serious damage to national security,

a person guilty of the offence is liable, on conviction on indictment, to imprisonment for life, or to a fine, or to both."

In laymen's terms it is a **serious crime** (Serious Crime Act 2015), punishable by imprisonment, a fine, or both, to cite hatred or spread hoaxes by use of a computer which would cause mass hysteria or public dissension. I have been condemning public figures, high profile individuals, celebrities, singers, movie stars, directors, politicians, world leaders, royalty, anti-secret society 'truthers' (who are really misinformation agents, and agents of deception) etc. –the list is long - since 2011 and **NOT** a single one of these high profile people has issued a public statement saying "I do not like how you have tarnished my name" or filed a law suit against me. Why you may ask? **Because** every word I have stated and will state again (in this document) is true.

Fundamentals

Before I describe the heinous crimes committed against me and many others (including unsuspecting civilians of the world) it is important that I explain, define, and reference:

- Key figures such as Phil Schneider, George Green and Aaron Russo, so that the reader has reference points to verify the extent to which technology has advanced in our **present** era (and continues to advance);
- Define and briefly explain the transhumanist / post human agenda (for those who may be unfamiliar with their aims).
- Define and explain rapid eye movement (R.E.M) sleep, the phases of sleep, and what happens to the (original's) body during sleep.

- Reference advances in technology, with particular attention to: Memory suppression technologies; Mind-voice technology; H.A.A.R.P technology, RFID microchip technology, and MK Ultra technology (CIA's mind control program);
- Define and explain cloning; the different types of clones; cloning centres and cloning technology;
- Define drip feed disclosure / evaluative conditioning and explain why it occurs;
- Explain what 'consciousness transfer' is to the best of my knowledge, and cite drip feed disclosure articles with the express intent to allow the reader to grasp the abuse I have suffered daily at the hands of my tormentors.

Key Figures

Phil Schneider

Phil Schneider (pictured) had 17 years experience working in government black projects carrying a level three security clearance. He was a geologist and engineer who worked in the black projects underground bases at Area 51, S-4, and Los Alamos.

He is most notable for disclosing (Schneider 1995; 1996; Open Minds 2011):

- The 'black budget' expenditure of the United States, which Schneider claims to be between 1.023 trillion U.S. dollars every 2 years (over $500 billion per year);
- Deep Underground Military Bases (also known as D.U.M.Bs -"dumbs"), and at the time of his lecture (Schneider, 1995), –that there are 131 **active** Deep Underground Military Bases present in the United States, and 1477 Deep Underground Military Bases worldwide;
- Each D.U.M.B costs on average 17-19 billion U.S. dollars; paid for by the taxpayer; and it takes approximately a year-and-a-half to 2 years to build D.U.M.Bs with sophisticated methods.
- That military technology outstrips the general public's technology at a rate of 44 to 45 years of technology for every calendar year which passes. In other words for every 12 months which passes military technology will have advanced by 44 years than what we as the general public is currently accustomed to. Therefore as a rough example if we were to take the year Facebook was founded (2004) as a base year, then the military technology since the creation of Facebook will have outstripped what the

general public is accustomed to by as much as **484 years(!)** (2015 - 2004 = 11. 11 x 44 = 484). Similarly, if we are to use the year which YouTube was founded as a base year (2005), once again, military technology would be 440 years more advanced than what the public is currently accustomed to today.

For the purpose of disclosing current advances in medicine, science and technology, and how such advances are used against me and the people of the world to commit monstrous crimes I will use 1945 as a base year. Everything will become apparent including why I use 1945 as a base year for my disclosure, but for the express purpose of compelling the reader to investigate my disclosure I must present everything, logically, sequentially, methodologically, and provide references for the reader in a coherent way to enable him or her to **pay serious attention** to my eye witness accounts. Therefore, at this present stage keep in mind the year 1945, and the fact that military technology outstrips the general public's technology at a rate of 44 years for every 12 months which passes.

George Green

George Green (pictured) was affiliated with U.S. Presidential candidates, and was once asked to be the Finance Chairman for the next President of the United States. Green would later declined the offer of Finance Chairman when a comment made by Ted Kennedy regarding sleeping with Green's 14 year old daughter caused George Green to re-evaluate his associations with this in-group.

During his 2008 interview with Project Camelot (2008a; 2008b), Green disclosed the following noteworthy information:

- U.S. presidents are "selected" and not elected. This is clearly expressed in a story Green recounts, when Green asks:
  "Who is going to be the next President of the United States [for him to overlook their finances]?"
  And the reply was: "Jimmy Carter"
  Green responds: "Jimmy who?"
  The reply Green received was: "Well, he's the Democratic Governor of Georgia."
  George Green: "But I've been voting Republican."
  Green was then confronted by a tall man, Paul Volcker (American Economist, and Chairman of Federal Reserve under Jimmy Carter and Ronald Regan), who walked over and said, "Son, don't worry about it [Republican or Democrat], we control 'em both."
- U.S. scientists learned how to make people (clones) since 1938 -walking talking ones –and the scientists call these people "synthetics" or "the others".
- Cloning technology is relatively advanced. All that has to be done is take two cells from the original, give the cells a small electrical charge (retain a fertilised egg), then all one needs is a receiver (a womb / artificial womb for the fertilised egg to grow).
- Scientist were excited by the synthetic technology because it meant that spare organ parts could be grown for an "original" human without rejection, because theoretically speaking, the DNA of the synthetic is the same as the original.
- Green gives an example of a cloned Politician: George Walker Bush. Green advices the viewer to seek old video recordings of George W. Bush, and compare the old George W. Bush, with the George W. Bush during Bush's second term,

-in terms of mannerism, speech pattern, body language etc. (old videos of George Walker Bush (Boringest 2006; Fox 4 News - Dallas-Fort Worth 2014) are presented in the References section).

- Scientist have also learned how to make these synthetic people within a few months, to the point where the synthetic can be a walking, talking duplicate of the original, intact with all the memories and experiences of the original. The only problem is that the memory, experiences, and functioning capabilities of these synthetics are like: "A DVD recorder. Sometimes you have glitches, and you have to take the synthetics to Camp David occasionally [every 6 months to a year] to get them tuned up."
- Remember these synthetics are people too, they can think and act just like you can, but they do not have a soul.
- Most of the world leaders have been bought and paid for [by men in the background] and are created to think a certain way –to meet the agendas of these men who remain in the background.
- The global elite plans on depopulating the current human population of **over 7 billion** to **500 million** people. This is corroborated by the "Georgia Guidestones" (WorldTruth 2014).
- Vladimir Putin is executing plans to bring the United States down and into a massive depression. Furthermore, China agrees with Vladimir Putin's plans; stating that the Chinese government has not been a Superpower for over 5000 years and "It is their turn to run the world."
- China has the capacity to set back / shut down all of U.S. computing and electrical systems within two days.

At this present stage, the most **salient** points to keep in mind from Green's accounts are that: U.S. presidents are selected and not elected; and **most importantly**, that scientists have been capable of creating synthetics (cloning people) since 1938. If you have kept in mind Schneider's statement (1995; 1996), (that for every 12 months which passes military technology increases by the equivalent of 44 years compared to what the general public is accustomed to) –then learning that scientists have been capable of cloning humans since 1938 should not come as much surprise. Again, everything regarding my disclosure will become apparent, and I thank these men for disclosing their information, because it helps the reader to corroborate my disclosure.

Aaron Russo

Aaron Russo (pictured) was an American businessman, director, and political activist. He is best known for directing blockbuster films such as *Trading Places*, *The Rose* and *Wise Guys*. Russo believed that a human being "Should stand for something, and do the right thing when the time calls to act". Consequently, during the latter part of his career he did his best to warn the American public by producing documentaries such as *Mad as Hell* (1996), *Freedom to Fascism* (Russo 2006), and *Reflections and Warnings* (Jones 2008).

Russo disclosed the following key information in *Freedom to Fascism* (Russo 2006):

Income Tax

- There is no law requiring Americans to pay income tax. **Although this is a truth, I DO NOT advocate Americans NOT to pay income tax (and neither did Russo).** This is simply because the Federal Reserve can imprison you and seize your possessions for not paying income tax (although there is no law requiring Americans to do so). So for the sake of avoiding hassle for the mean time –pay your income taxes.
- To further illustrate the above point, Bob Shultz speaking at "We the People Foundation" said the following: "Most people believe that the income tax system is legal and that the revenue from the tax is used in the public interest. However, there is a substantial conclusive body of evidence that proves that our income tax system represents the most pernicious form of tyranny. It is the greatest hoax ever perpetrated by government against the working men and women of America."
- Charlie Beall: "The federal government itself refuses to provide, the American people, who are coercively being subjected to this extraction of their private property, without any underlying legal justification. There is no law. There is no law that requires the average American worker in the private sector to pay a direct un-apportioned tax on their labour and compensation for services. There is no law."

- State Representative – Phil Hart –(R-Idaho): "You can look through the statutes, and look for the law that requires you to pay, and when you do that, you **cannot** identify a law that requires the average person in America who earns a wage and works in private business to pay an income tax."
- Peter Gibbons (Tax Attorney) –It's actually very simple. Congress tried to enact an income tax in 1894 –The Supreme Court said that is unconstitutional. When The Supreme Court says something is unconstitutional, it's unconstitutional. They (Congress) tried again in 1913 –and The Supreme Court said –the 16th amendment- "No new power of taxation" –so if they (Congress) didn't have it then (1913) and they didn't get it; they **DON'T** have it. There is no constitutional base for a tax on the wages for Americans living and working in the 50 States of the union. Period. End of argument.

Radio Frequency Identification (RFID) Microchips

- The latest technology for identifying people at the point of self, when they make purchases –is actually the implantable (RFID) microchip. There are microchips that can actually be imbedded directly into human flesh... It's a tiny glass capsule about the size of a grain of rice... it contains an RFID computer microchip with a coiled antenna and it can transmit information also at a distance.
- Katherine Albrecht –Author of "*Spychips*" (2005) makes the following comment: "RFID is a technology that uses tiny computer technology the size of a grain of sand or smaller; hooked up to miniature antennas to transmit information about items at a distance. Back in 1999, Procter and Gamble, Gillette, and MIT got together to find a way to commercialise this technology and make it small enough, make it efficient enough and make it low cost enough to essentially ---their dream is to put these tiny computer chips on every physical item manufactured on planet earth."
- Radio waves can travel through walls, they can travel through wood, and they can travel through things we normally rely on to protect our privacy, for example your purse, your bag-pack, your pocket, –anything you are wearing or carrying.
- One of the most worrisome applications of RFID microchips, are proposals to put them into cash, meaning that it would be able to track every bank note where it had been, who it had been issued to, and create in essence an audit trail. That would essentially take away the anonymity of cash we now enjoy today.
- Once everything you do is tied down to a single number and there is no longer the ability to pay with cash, then all it takes to render you a non citizen is to simply turn that microchip off. **You will no longer be able to participate in any function in society including buying food. Once money becomes digitised through RFID technology, the elite can deduct whatever amount of money they want out of your microchip, whenever they want. They can trace you whenever they want. You will be at their mercy.**

Habeas Corpus

- Habeas Corpus is "[A] writ [formal document] requiring a person under arrest to be brought before a judge or into court, especially to secure the person's release unless lawful grounds are shown for their detention" (Legal-dictionary 2015).
- During the Bush Presidency [2001-2009 (History 2015)], President Bush signed executive orders giving him sole authority to impose martial law, and suspend Habeas Corpus. This gives him doctorial power over the people without any 'checks and balances.'

  In other words, "The government can jail you for life without charges, without a trial, without a lawyer" (Russo 2006).
- Furthermore: "The National Defense Authorization Act [4] signed by President Obama on the 31st December 2011 authorises the indefinite detention, without trial or indictment, of any US citizens designated as enemies by the executive." See Paye (2013) for further discussion regarding the suspension of Habeas Corpus.

Fraudulent Manipulation of Election Results

- In 2004 at the "Forum of Presidential Election" Clinton Eugene Curtis, a former Computer Programmer for NASA and ExxonMobil testified under oath that election results can be rigged using electronic programs.
- In 2004 he gave the following abridged testimony (Russo 2006):

  Judge: Mr Curtis, are there programs which can be used to secretly fix elections?
  Curtis: Yes.
  Judge: How do you know that to be the case?
  Curtis: Because in October of 2000 I wrote a prototype for present congressman Tom Feeney, and the company I worked for in Oviedo Florida –it did just that.
  Judge: And when you said, "it did just that" –it would rig an election?
  Curtis: It would flip the vote 51:49 –whoever you wanted it to go to, and whichever race you wanted to win.
  Judge: And would that program that you designed be something that election officials that might be on county boards would actually, could detect?
  Curtis: They'd never see it.
  Judge: So how would such a program, a secret program that, fixes the election –how could it be detected?
  Curtis: You would have to view it either with source code, or you would have to have a receipt and then count the hard paper against the actual vote total –other than that you won't see it.
- **Judge:** Given the availability of such vote rigging software and the testimony that has been given under oath of substantial statistical anomalies and gross differences – between exit polling data and the actual tabulated results, do you have an opinion whether or not Ohio elections, the Ohio Presidential elections was hacked?
  Curtis: Yes. I would say it was.

Person in the audience: ---So in other words there is absolutely no assurance whatsoever in anything in regard to these machines?
Curtis: Absolutely none.

- See Truthstream (2006) for further details regarding the full account of Curtis' testimony on how elections can be manipulated.
- Moreover, voting machine manufactures refuse to allow anyone to see the source code. Without paper ballots, the honesty of any election cannot be verified.

The War on Terrorism is the War on Your Freedom

- It is time to wake up America. These ID cards are not about defeating terrorism, but they are **ALL** about controlling the American people.
- The (mainstream) media controls the information that a person gets in various ways. They can make sure that the average American watching T.V. or reading the newspaper is going to come out with a certain mindset. She is going to say this is good, that's bad –and that is all they (media/elites) have to do.

Whoever Makes the Money Makes the Rules

- As Mayer Rothschild said "Give me control of a nation's money supply, and I care not who makes its laws". Mayer Rothschild, private banker, –knew that he and the other bankers would now control the laws of the nation. Government gave these bankers one of its most important powers, and now had to borrow money from the bankers and pay interest to finance the government.
- America has gone from people owning their own property, owning their own businesses –to a nation in debt because all the money is created by borrowing (from private banks) and this country has become one where people just live by borrowing.

Russo's Message to Mankind

- If you are in the military or law enforcement, remember you swore an oath [to uphold the law of the land]. You did not swear an oath to promote world government, or corporations.
- Now that you do understand what happened [to America, and the monetary system based on debt slavery] and how it is leading to a tyrannical one world government – the future of mankind depends on you –will you choose freedom or slavery?
- Stop being passive liberals. Stop being passive conservatives. Stop being passive centralists. Stop being passive human beings! When the media starts telling you that the country will fall apart if this is done... Do not be fooled!!
- Remember these are sick, malevolent and twisted people we are dealing with, trying to save themselves. Squash their agendas and stay on course.
- "I believe the time for mankind is time to give all or perish. Grow up or die. Grow up and become adults. Act like adults. Take some responsibility. The world which you have perceived is childlike; –and now the curtain has been pulled back for all to see."

- Unless people get active [get fully informed, and learn how governments of the world are conspiring against their people] and say –I am going to help shut down the Federal Reserve System; I am going to shut down the powers that be –the whole human race is doomed otherwise.
- We are coming near the end game and things are starting to accelerate... –and people are looking at the world leaders to declare martial law [which in turn will rid Habeas corpus].

Let's Consolidate our Problems

- If we can all get focused on how to win the game instead of all these different objectives –such as people are fighting for a better environment; people are fighting to keep their guns/gun control; there are all these different issues which are going on around the world -and which are all important on their own –but if we can consolidate on that and focus –and take all those people –and say –hey let's shut down the Federal Reserve, then we'll deal with those issues. Let's go to the agenda, the objective, the root cause of man's problems, first, then we'll go back to the other stuff.
- That is the priority right now. You have to cut off the head of the beast. And the head of the beast is the Federal Reserve System and the people behind it. You see, and that is what will save the world, and if people understand that –and they **stop** being passive liberals, and **stop** being passive republicans, but rather become **active** human beings –that is what will save the world!

Aaron Russo: Reflections and Warnings

In this documentary Russo recalls his relationship with Nick Rockefeller. Russo shares critical information Rockefeller discussed with him.

The Falsehood of September 11th 2001

- Russo met Nick Rockefeller through a female attorney who telephoned Russo and said "One of the Rockefellers would like to meet you." Russo made a documentary called *Mad as Hell* (1996), and Rockefeller had watched the video and knew Russo was running for Governor of Nevada and wanted to meet him. Russo said "Sure I'd like to meet him", and the two met and talked. Rockefeller proved to be a very smart man and shared ideas with Russo and was the person to tell Russo 11 months before 9/11 happened "There was going to be an event..."
- Rockefeller never told Russo what the event was going to be; but, there was going to be an event –and out of that event, we [America] were going to invade Afghanistan, to run pipelines from the Caspian Sea; we were going to invade Iraq, to take over the oil fields and establish a base in the Middle East, and make it all part of the New World Order; and we would go after (Hugo) Chavez and Venezuela –and sure enough, later when 9/11 happened –and I remember he was telling me how you are going to see soldiers looking in caves for people in Afghanistan and Pakistan and all these places and there is going to be this war on terror which has no real enemy and

the whole thing is a giant HOAX, but it is a way for the government to take over the American people (and the world).

- There is no question about it [that 9/11 is a hoax]. Nick Rockefeller said to Russo: “There is going to be a “War on Terror”” and he was laughing. Who are we fighting against...? Why do you think 9/11 happened and then nothing has happened since then? Do you think that our security is so great here that these people who pulled off 9/11, who were able to lock down another plane...? Come on it is ridiculous... 9/11 was done by people in our own government and our own banking system to perpetuate the fear of the American people into subordinating themselves to anything the government wants them to do. That is what it is about and to create this endless war on terror.
- Nick Rockefeller was laughing when he said, “We are going to be sending men into caves in Afghanistan and Pakistan” –and it was just cynical, he kept laughing and saying “Look how stupid everyone is! We can do whatever we want!”
- 9/11 was the first lie; and the next lie was to go into Iraq, to get Saddam Hussein out with his weapons of mass destruction (when the real issue was control of the oil fields) –that was the next lie. 9/11 created an endless war on terror that would go on and on and you can never define a real winner. There is no one to defeat and so it goes on and on forever. And they can do whatever they want; because they scared the hell out of the American public.
- This whole war on terror is [perpetuated on] a fraud. It is a farce. It is very difficult to say it out loud because people are intimidated in saying it. Because if you say it they want to make you into a nutcase –but the truth has to be... and the truth has to come out. The fact of the matter happens to be the whole war on terror is a fraud, it is a farce. Yes. There is a war going on in Iraq, because we invaded Iraq, and people over there are fighting... but the ‘war on terror’ –it is a JOKE –you know, and until we discover what really happened in 9/11 and who was responsible for 9/11 –because that is where the war on terror emanates from. That is where it comes from. It was 9/11 which allowed this war on terror to begin and until we get to the bottom root of 9/11 –the truth of 9/11 we’ll never know about the war on terror.
- Russo was in Tahiti when 9/11 happened and he got a call from his son – and his son said –“The twin towers, they were just attacked and they are falling down or something...” Russo was in Tahiti and had just woken up from sleep. Russo didn’t realise what it was immediately (11 months after Rockefeller had told him about 9/11 and it had happened –because Rockefeller said there was going to be an event –he wasn’t specific) –but after Russo saw that America was going to go to Iraq and Afghanistan, that is when he realised, and equated it to what Rockefeller had said.
- 9/11 was only a manifestation to create a fear in the American public. So that we would obey and do what they want us to do. Take for example, Richard Reed ‘The shoe bomber’ –now here is a guy who is 6 feet 6, ugly as can be; I heard he smelled... He sits on a plane, lights a match in a non-smoking area, to put his shoe on fire... surrounded by people... That is idiotic! If you were going to blow up a plane... you would go into the bathroom... you close the door... and you put your shoe on fire...

you are not going to sit there.... surrounded by people, lighting matches in a no smoking flight... they (elite) want you to believe this nonsense. That is ridiculous!

- The war on terrorism is to keep people in fear. It is an endless war without a real enemy (‘terrorists’) –so that people would submit and do whatever the government wants them to. Submit to searches, give you ID cards, put Radio Frequency Identification (RFID) chips in you etc. You become servants to the elite –that is what this is all about...
- Freedom of liberty is what people really want, and it is time to stop the duplicity of the government from lying to us. You see many people know the truth of what is happening in this country; like 9/11 but they are afraid to stand up. People have to stand up and find their courage and say “I’m not going to take this anymore, I know the truth” –and they create a situation where if you tell the truth, you are considered a lunatic.
- In other words if someone goes on a T.V. show and says that 9/11 was an inside job – immediately the person is labelled an idiot or crazy. They call you names. **You cannot be afraid of that.**
- If you do not fight the corruption and you do not stand up for what is right in life, you end up being a serf and a slave and you are leaving your children a world in which you would not want to live in yourself, so how can you in decency behave that way? You have to stand up for what is right in life, and unless you do that you are nothing.

The New World Order Agenda

- The whole agenda is to create a one world government where everyone has an RFID chip implanted in them. All the money is to be in those chips [a cashless world]. This information came straight from Nick Rockefeller himself. That is what the ultimate plans of the global elite, banking industry and Rockefeller wants to accomplish.
- The agenda is to implant everyone with RFID microchips. All money is transferred to those microchips. There is no more cash. Money would be in microchips. Instead of having cash, you would have money in your microchips, but whenever they want, they (elite) could take whatever amount out of your microchip whenever they want to. Total control. If you are a protestor they just turn off your chip; you cannot buy food; you cannot do anything; it is total control of the people.
- So they want a one world government controlled by them. Everyone being chipped, all the money in those chips, and they control the chips and they control people, and you become a slave. You become a serf to these people –that is their goal. That is their goal; that is their intentions.
- Russo did not believe in enslaving people and Rockefeller would question him in the following manner: “Why do you care about them? Why do you care about those people? What difference does it make to you [Aaron]? Take care of your own life.
  Do the best you can for you and your family. What do the rest of the people mean to you? They don’t mean anything to you. They are just serfs. They are just people...”

- Rockefeller asked –"Why are you fighting for the people for, what is it all about? The people have to be ruled. The constitution, what you are standing for is only for a few people, it's only for a few individuals who can live that way and we believe that it is best for society to be ruled by an elite people who control everything." Russo told Rockefeller he does not believe that. Russo believes: "God put me on this earth to be best person I could be and put everyone on this earth to be the best they can be, and NOT to be a slave and a sheep to YOU and these people (elite) –and I do not understand why you want to control everything. What is the need for that?"
- It was just a lack of caring [from Rockefeller's part], and that is just not who Russo was. It was just sort of like cold you know, and Russo used to say to Rockefeller, "What is the point, of all this? You have all the money in the world you need; you have all the power you need, what is the point? What is the end goal? Rockefeller said "The end goal is to get everyone chipped. To control the whole society. To have the elite people [the bankers and government] controlling the world."

Women's Liberation from the Perspective of the Elite

- Women's' Liberation was founded by the Rockefellers.
- Rockefeller asked Russo: "What do you think women's liberation was about?" At the time Russo had a pretty conventional thinking about it and he said "It's about women having the right to work; get equal pay with men; just like they won the right to vote." At this point Rockefeller started to laugh and he said to Russo "You're an idiot"
- Rockefeller said –"Let me tell you what that was about. We the Rockefellers funded women's' liberation. We are the ones who got it all over the newspapers and television. The Rockefeller foundation –and you want to know why? There were two primary reasons:"
- 1) "We couldn't tax half the population before Women's Liberation and the second reason was"
- 2) "Now we get the kids at an early age [because both parents are away from home working] –we can indoctrinate the kids how to think, so it breaks up their family –the kids start looking at the state as the family, as the school as the officials as their family, not as their parents teaching them, and so those are the two primary reasons for women's liberation."
- Russo thought, up to that point, Women's Liberation was a noble thing; however when he saw the Rockefellers' intentions behind it, where they were coming from when they created Women's Liberation; the thought process of it; Russo saw the evil behind what he thought was a noble venture.

America is a Republic

- America is a constitution republic / it is supposed to be a constitution republic –and **NOT** a democracy. The majority should not rule, and nor should the majority take over the inalienable rights of the minority.

- Americans pledge allegiance to the "Flag and the **Republic** which it stands" –and **NOT** democracy.
- Democracy is the worst form of government you can have because it is majority rule. Therefore the government can tell you what to do because 'the majority' wants it. It is irrelevant what the majority wants. Decisions should not take away the inalienable rights of an individual.
- Russo also said "It doesn't matter who you vote for, republican or democrat; –they are the same –neither one of them is stopping the Federal Reserve or paying income taxes.
- The elite (Federal Reserve) have taken over the American government; there is no difference between republicans and democrats. There is no difference between the two parties. The duality is manufactured. They (Federal Reserve; central banks) control both parties. It doesn't matter to the elite which one wins, because whoever is running for President will be someone they anoint. Whoever runs for President, will do whatever the elite want them to do. The fact of the matter happens to be that you cannot win an election unless you have enough money to win; they (Federal Reserve; central banks) make sure who gets the money.

Depopulation

- Russo and Rockefeller discussed many things –and one of the things Rockefeller brought up in conversation was reducing the world population. Rockefeller felt that there are too many people in the world. In a way Russo agreed that there are too many people in the world, but he **does not** think he has the authority to say who dies and who lives; but the elite felt that they want to reduce world population and Rockefeller felt it should be reduced by half.
- Rockefeller even mentioned to Russo in conversation, that they were having a real problem trying to solve the Israel/Palestine problem –and they were playing with the idea of bringing Israel to Arizona –and taking everybody from Israel and giving everybody a million dollars, and setting up Israel in the state of Arizona –because that is a problem that they are not in charge of.

Borrowing Money from Private Banks causes Inflation and Debt

- These people (elite) control the money so they make all the rules, and therefore they put the rules in which they want into effect, and the truth is America has really become a socialist communistic country. Everyone says it is a capitalistic country, but how can it be a capitalistic country when you have a central bank? That is the first question people should ask. It can't be. It is a planned economy, it is a phony!
- If they want to create prosperity, they just print dollars, or put digits into the economy. Now you have prosperity. You do not have real prosperity, you do not have real manufacturing, you just have money being injected in which is infusion of credit. This makes the government go into more debt.

- Whoever makes the money makes the rules. Rothschild said that. Why are we allowing these private bankers to make the money for our country? It is nonsense. Why are we paying interest to these banks to make money for us when the government can do it itself without paying interest, without all that debt? There is no answer to that question and it is a question no politician will raise. Everybody talks about America's debt... We are in debt because we borrow money... but we don't HAVE to borrow money. They designed it so that we can go into debt. We can create the money, and back it by gold so that they cannot create too many of it, so that you do not have the inflation, and do what the Founding Fathers told us.
- Why in the world does the American government borrow money from the banks when they have the ability to create it themselves without borrowing it, and paying interest on it? Why? Nobody can answer that question –not one politician ever raises that. Why does the American government ever borrow money, when they can create it without paying interest?

  –And people say that –well if the American government creates it, it will cause inflation. And that is their answer. And Russo says well let's look at it: the American government has the Federal Reserve do it, which creates the same inflation as if they did it, but also with the inflation –now you are getting massive debt –so with the Federal Reserve you have inflation and debt. Now if the American government made the money, backed by gold which would limit the amount they could make –you wouldn't have debt and you wouldn't have inflation.
- It wasn't until 1913 when the Federal Reserve came in, that America had inflation. Before then there was no inflation for 100 years. There were points and spikes, mostly during the Civil War –but basically there was no inflation other than during that short period of time. I mean a loaf of bread would cost the same thing. People could plan their lives.
- Today, they have planned inflation, and now you have two parents working, they cannot afford to take and pay for their family anymore; the kids are going to state run schools now, the kids are being indoctrinated how to think; they are being given Ritalin, they are being given all these drugs, the whole country is being dumb down, it is all because of the Federal Reserve System; and the Federal Reserve system and these bankers are responsible for the demise of America. And if we ever want to win this battle you must shut down the Federal Reserve System, and we must shut down these bankers and restore sound money to this country.
- If you analyse the situation and if you realise that since the Federal Reserve has come into being since 1913, illegally, without a constitutional amendment, by bribing a few senators during Christmas vacation, they turned over the most important power that the American government has, the creation and issuance of money to a private bank.
- Through that private bank issuing money they have destroyed this country. They have destroyed the purchasing power of money in this country; they have created social programs that are destroying this country.
- The Federal Reserve has created massive inflation in America which means the American worker has to keep on making more money to keep up with the cost of

living. The more money they make to keep up with the cost of living, the less competitive they become in the world economy. So now what happens is that we have to pay our workers so much to keep up with the cost of living; and then they (elite; government; corporations etc.) say screw the American worker; let's go overseas now to get the cheap labour.

- The inflation the Federal Reserve has created has now allowed other countries to outcompete us. Other countries do not have to pay as much money as we have to pay to our workers to survive. So now we are not competitive anymore and we have lost our manufacturing base. We have lost our competitive edge.
- "Freedom to Fascism" is a documentary that everybody should see. Russo and his team show the fraud of the income tax; they show how Judges put people into jail for no reason; they show the corruption of the justice system. They show how the Federal Reserve came into being and how it is controlling society and how all the central banks are working together through the bank of International Settlements, in Switzerland which is the central bank for all central banks and how all are working together to create this one world government; this one world order; which is what they are trying to do.

The Deception of the Council of Foreign Relations (CFR)

- Russo was interested in joining the Council of Foreign Relations, but he found out from Rockefeller himself, that part of the end goal of the CFR is to get everyone RFID chipped. To control the whole society and have the elite people (bankers, government etc.) controlling the world. Russo asked Rockefeller "Do all the people in the CFR believe the way that you do?" Rockefeller said "No, no, no. Most of them believe they are doing the right thing. A lot of them believe it is better off being socialistic. We have to convince people that socialism is really capitalism." Because America is becoming a socialistic country; it is a communist country today.
- Russo's friendship with Rockefeller became one where they would share thoughts, ideas and philosophies and Rockefeller wanted Russo to become a part of what they were doing [enslavement agenda], and for Russo to become a member of the CFR; Rockefeller offered various business opportunities for Russo to get involved in the CFR and to **not** take up the fight or the battle that Russo had been taking up in the past. Rockefeller wanted Russo to drop the idea of helping the people; because "What was the point in Russo fighting for the people?" Rockefeller would question.
- Russo asked Rockefeller do all the people of the Council of Foreign Relations feel the same way you feel? Rockefeller said "A lot of them think they are doing the right thing, they think that socialism is the best way to go (but this form of socialism involves redistributing the wealth for the elites and not to everyone), they think that they are doing the right thing." But the people at the (very) top they all know the truth of what is happening. (Therefore, the good people in the CFR are also under an illusion, and do not know that effectively they are working for evil men who have the overall goal to control the populace of the world).

- So it is compartmentalised within the elite structure as well. All the people in the CFR -2000 to 3000 people like Dan Raddler –they don’t know what is going on--- they join the CFR because it is prestigious. They think it is good for business, it is good for this; they don’t know what is really happening –the evil that comes out of it –that is emanating out of it.
- In terms of the CFR, in terms of compartmentalisation, there are many good people which Russo believes are part of these organisations who do not even understand what these organisations are really about. For example, when Russo was in Germany, doing cancer treatment, there was a gentleman there who was visiting a friend of his with cancer. The gentleman visiting his friend was a member of the CFR, and him and Russo were talking and Russo showed him the movie (*Freedom to Fascism* (2006)); --and he said, “Oh my god! I’m going to resign.” “I had no idea this is what the CFR is about”. He had no idea; he is just a nice guy, who thought he was joining a prestigious organisation.
- A lot of people join the CFR because they think it is a prestigious organisation; it will help them in business; make good business contacts etc. They do not have an understanding that the CFR is really about world domination. How they, and the Trilateral Commission, The Bilderberger [Group], the banks, all work together to control the people –a lot of them do not understand that. They do not see the big picture. They think: Oh the CFR is a prestigious organisation. I’ll make this, I’ll make that, and I can do business deals. It is just business to them. The CFR wants to get the people in there that have influence and power, and so they are part of that (enslavement agenda), and so they are not opposed to them. So the whole country is becoming the haves and the have not. You are getting the very, very wealthy and the middle class being destroyed and you are getting the poor people.
- In Russo’s words: “You can call the CFR what you like, but it is a criminal organisation. Run by criminals. But people do not think of it as a criminal organisation, because it has ‘class’; ‘style’; ‘prestige’; –and it is ‘respected’; so people do not look at it as being a criminal organisation. That is what a great job they have done.”

Combining America, Mexico and Canada into one Country

- Imagine this... here you are in America, and they (elite) are combining American, Canada and Mexico into one country. The North American union. And the American people do not know anything about it. It is not even in the press. They would rather talk about Rosy O’Donnell and Donald Trump calling each other names than discussing the fact that we are merging into one country. This isn’t even reported.
- The fact of the matter happens to be, that tells you how controlled the media is. The elite control the media, and they control governments and they are all in bed together. Here you are combining American, Canada and Mexico into one country, and you do not see it in the press. You do not see it in the press. Why? This should be one of the top stories everywhere –and the elite are not worried about it. That tells you –there is the evidence that [media] it is controlled. They do not want the American people to

know what is going on, that is why they do not protect our borders. That is why we are losing our constitution, the very document that secures our freedoms.

Russo's Vision on How to Bring These People Down

- There is no question we are in tyranny, there is no question that the American citizen is no longer a free individual human being, to do the things that they wish to do. We're slaves and it is getting worse.
- We are dealing with complete evil; and until the American people wake up and say, we do not want this evil in our country anymore and we want to come back to a country of decency and goodness, integrity and honour, we are going down that road and that is what it is going to take? It is going to take people to stand up and say we do not want to live in this kind of world anymore. I believe we should pull all of our troops out of Iraq, I believe we should leave other countries alone. Let other countries live their lives the way they choose to. Stop trying to spread 'democracy' around the world, which is the worst form of government there is anyway (because 51 percent rule over 49%), restore our republic to what it is supposed to be and go back to what the founding fathers gave us. Restore the republic.
- In Russo's opinion: The populace must shut down the Federal Reserve System –and there has to be an uprising. There has to be an uprising. People have to stand up. One person cannot do it alone. You cannot do it alone. People do not seem to have the courage to do what they have to do.
- A lot of people in Hollywood know the truth, they do not want to stand up and speak about it; I know many of them have seen my movie (*Freedom to Fascism* (2006)) and they know I am right, and they want to talk about it because everybody is afraid. Everybody is afraid because they think that the money they get from the Federal Reserve is really money and they have a comfortable lifestyle and they are afraid of change. They are afraid to stand up for what is right, and until people are willing to stand up and have the courage to do what they need to do, it is not going to change; and hopefully we can affect change when people stand up and say "Hey, I've had enough".
- We (the populace) have one advantage. They (elite) need us to cooperate. See, if we do not cooperate with them, they cannot win. They always need our cooperation with them to go along with their programs. They try to "sell us". Democracy; this majority says this; believe in this; do this, do that; the war on terror; we have to be scared... They are always trying to do things to "sell to us" to go along with them, and once we learn not to cooperate with them; then we win the game.
- That is the point, do not cooperate with them, and do not go along with the program anymore. Stop it. Join forces, and bring freedom back to this country. It is going to take people who believe in freedom; The Constitution and the Founding Fathers, Thomas Jefferson, to make this country whole again, because right now it is in the grip of the evil ones, and the only way to stop that is for good men to stand up.

- We have to stop being scared. We have to do what is necessary to take back what is ours. We have to stop these bankers, these elite, full of liars, congressmen full of liars. They are destroying our borders.
- So through these bankers, attempting to take over America, knowing that America was the freest nation on this earth, it was necessary for them to takeover America, take away gun rights, freedom to bear arms, and create a country where we become slaves, because once they take over America, the rest of the world becomes a lot easier for them. And so by creating 9/11, an event to terrify the American people that we are being 'attacked by terrorist', you create a world where there is an enemy that can **never** be pinpointed. You can never win the battle. It is 100 year war – a never ending war on terrorism. So you are always fighting this war, and through the war on terrorism, through 9/11 which is the first lie, then you create the war on terrorism which is the next lie, then you create the war in Iraq through weapons of mass destruction, which is the next lie –so you get one lie, to the next lie, to the next lie, -- now it is going to be Iran the next lie –and sending more troops and insurgents into Iraq.
- Restore America's Republic back to what it is supposed to be. Get the bankers out of our government. Get government to stop borrowing money from the banks. Government should make its own money; restore the Republic. Restore individual freedoms. That is what this country is about -and until we do that we are going to be slaves.
- You have to take away the creation of money away from the private bankers and you will solve 95% of your problems.
- Americans, mobilise, stand tall, stand together, tell the government you are "Mad as hell!" Do not cooperate with the government do not accept a National ID card. Do everything in your power to restore freedom and your individuality back to America. Stop being a country run by the institutions for the institutions. Let's go back to "We the people, by the people for the people", as opposed to, we the institutions, by the institution, for the institution. Stand up for your individual rights. Stand up for the God leaders that are in each and every one of us!

The Trans-humanist / Post Human Agenda

Transhumanism is a cultural and intellectual movement that believes we can, and should, improve the human condition through the use of advanced technologies. One of the core concepts in transhumanist thinking is life extension: through genetic engineering, nanotech, cloning, and other emerging technologies, eternal life may soon be possible. Likewise, transhumanists are interested in the ever-increasing number of technologies that can boost our physical, intellectual, and psychological capabilities beyond what humans are naturally capable of (thus the term *trans*human) (Anthony 2013). Transcranial direct current stimulation (tDCS), for example, which speeds up reaction times and learning speed by running a very weak electric current through your brain, (Anthony 2012) has already been used by the US military to train snipers. On the more extreme side, transhumanism deals with the concepts of mind uploading (to a computer), and what happens when we finally craft a computer with greater-than-human intelligence (the technological singularity) (See: "How to create a mind, or die trying", Hewitt 2012) (Anthony 2013).

Moreover, put simply, "posthumanism" can be defined as that condition in which humans and intelligent technology are becoming increasingly intertwined (TheNanoAge 2015). For readers interested in learning more about the posthuman / transhuman agenda see BT Soul Catcher 2025 (BEAMS 2007); Avatar Project 2045 (2045 Initiative 2015; Borghino 2012) and Mind Clone Robot (Bloomberg Business 2015; RT 2015) which have all been disclosed as methods of transferring the human consciousness to a computer.

The Phases of Sleep and Rapid Eye Movement (R.E.M) Sleep

Sleepers pass through five stages of sleep: 1, 2, 3, 4, and REM (rapid eye movement) sleep. These stages progress cyclically from stage 1 through REM then begin again with stage 1. A complete sleep cycle takes an average 90 to 110 minutes (Sleepdex 2015). In other words, after falling asleep, it takes approximately 90-110 minutes to enter REM sleep.

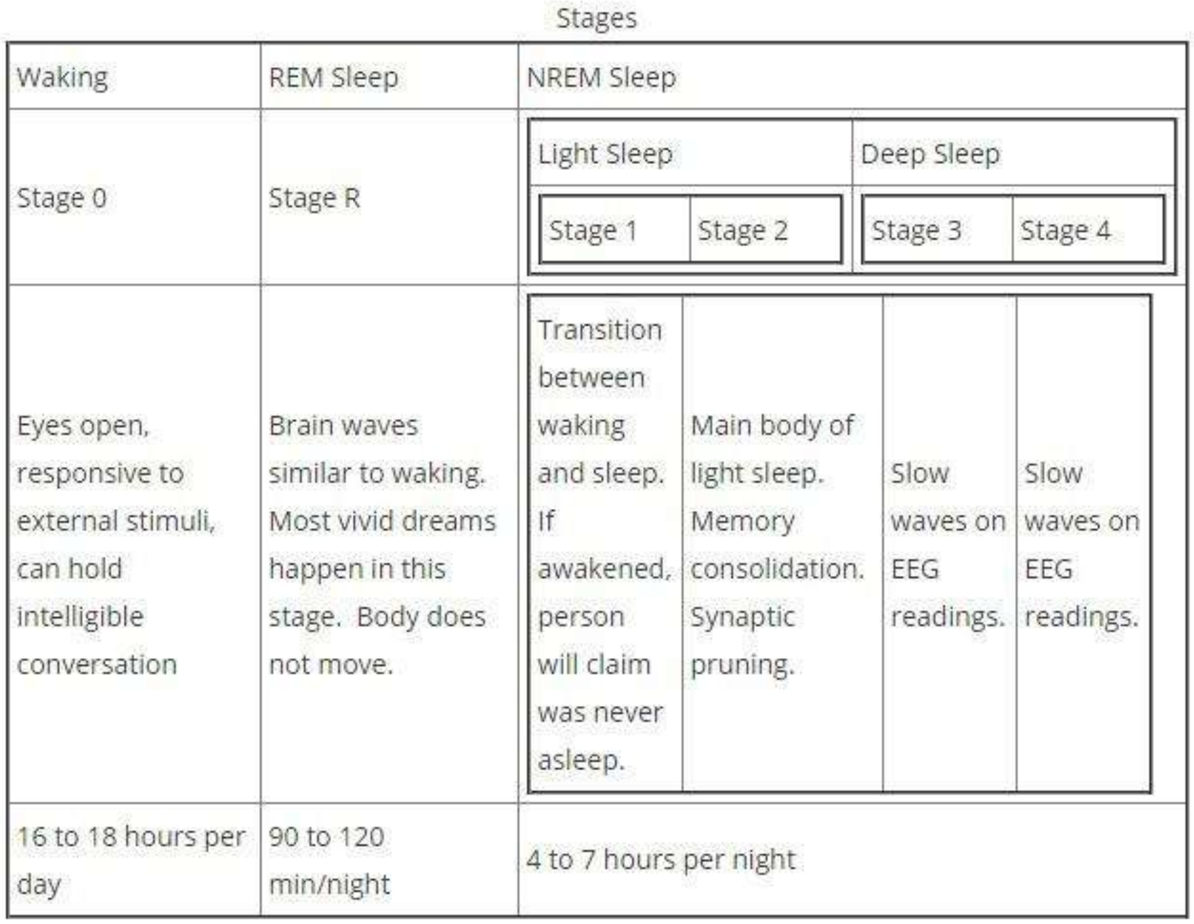

Stages

Waking	REM Sleep	NREM Sleep			
Stage 0	Stage R	Light Sleep		Deep Sleep	
		Stage 1	Stage 2	Stage 3	Stage 4
Eyes open, responsive to external stimuli, can hold intelligible conversation	Brain waves similar to waking. Most vivid dreams happen in this stage. Body does not move.	Transition between waking and sleep. If awakened, person will claim was never asleep.	Main body of light sleep. Memory consolidation. Synaptic pruning.	Slow waves on EEG readings.	Slow waves on EEG readings.
16 to 18 hours per day	90 to 120 min/night	4 to 7 hours per night			

Image 1: Stages of sleep. Source: Sleepdex: (2015)

Any sufficiently advanced technology is indistinguishable from magic.
Arthur C. Clarke

REM Sleep

Most dreaming occurs during Stage Five, known as REM. REM sleep is characterized by eye movement, increased respiration rate, and increased brain activity. REM sleep is also referred to as paradoxical sleep because, while the brain and other body systems become more active, your muscles become more relaxed, or paralyzed. Dreaming occurs because of increased brain activity, but voluntary muscles become paralyzed. Voluntary muscles are those that you need to move by choice, for example, your arms and legs. Involuntary muscles are those that include your heart and gut. They move on their own (Sleepdex 2015; Walcutt 2013).

Rapid eye movement, or REM sleep, is when you typically dream. You may have images float by in earlier stages, particularly when you are going through Alpha or Theta (brain waves), but the actual dream state occurs in REM (Walcutt 2013).

This period of paralyzation is a built-in protective measure to keep you from harming yourself. When you are paralyzed, you can't leap out of bed and run. Do you ever feel like you can't escape during a dream? Well, the truth is, you can't. You can breathe, and your heart is working, but you really can't move (Walcutt 2013).

The reader is also advised to see the Horizon documentary "Why Do We Dream?" (BBC Horizon 2009). This video is available, on YouTube, and details sleep, the phases of sleep, REM sleep, the period of paralysation and more.

Current Advances in Technology

This is a section some readers may have difficulty with, because reading about these current technologies alone, one cannot help but feel that such technologies described below sound like the stuff of science fiction and fantasy -although they have been corroborated through public drip-feed disclosure (explained below). Nevertheless, I urge the reader to keep in mind Arthur C. Clarke's quote (above), as well as Phil Schneider's testimony regarding the advancements of military technology in comparison to the general public's technology (for every 12 months, military technology outstrips the technology the public is accustomed to by a rate of 44 years). Moreover, for the reader who finds difficulty understanding the technologies described below, seek the supporting articles / videos.

Memory Suppression Technologies

Memory suppression technologies are any scientifically advanced technologies which are used to suppress memory. Examples of how memories can be suppressed can be found by reading Winter's (2014) article which details how memories can be suppressed using light; and Greenberg's (2013) article which discusses memory suppression through gene / chemical modification.

Mind-Voice Technology

Mind-voice technology is an advanced technology which is capable of reading, listening, hearing or broadcasting your inner voice / thoughts. Examples of articles which discuss mind-voice technology are: Prigg (2014) details software which can read the inner voice; and New Scientist (2014) which also discusses a brain decoder which can eavesdrop on one's inner voice.

H.A.A.R.P. Technology

The High Frequency Active Auroral Research Program (HAARP) is a radio transmitting system that can bounce signals off the ionosphere (a region of the Earth's upper atmosphere 60km (37 miles) to 1000km (620 miles) altitude) and back to earth to probe deep into the earth or sea, its proponents say. The system could locate minerals or communicate with submarines. (Begich & Manning 1997; Sheen, Begich & Robbins 2005).

HAARP can also:

- Disrupt human mental processes.
- Knock out all global communications systems.
- Manipulate global weather.
- Change weather patterns over large areas.
- Interfere with wildlife migration patterns.
- Hurt ecosystems.
- Negatively affect your health, moods, and mental states.
- Unnaturally impact the Earth's upper atmosphere.

This illustration (below) shows the ionosphere's relationship to the Earth. The illustration appeared in the HAARP Environmental Impact Statement on page 10-125 of Volume II.

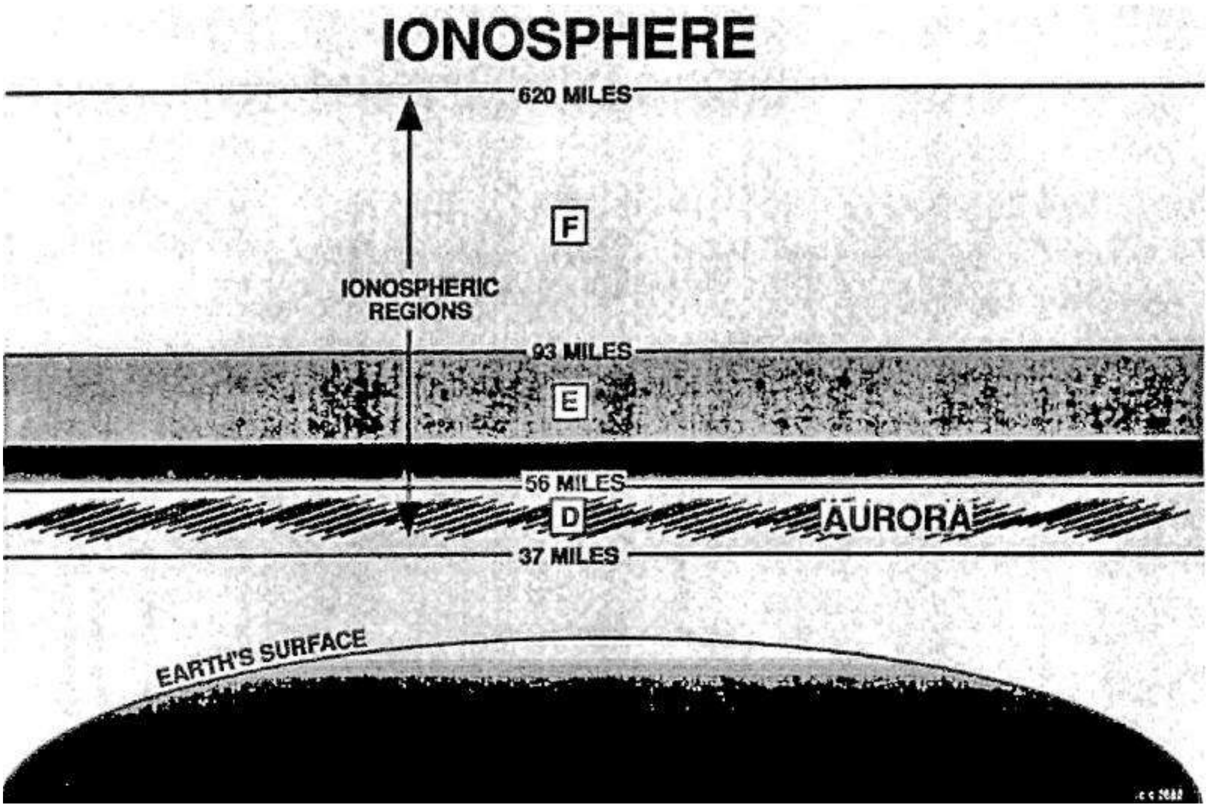

Image 2: The ionosphere's relationship to earth. Source: (Begich & Manning 1997)

The ionosphere protects the earth. HAARP (High frequency Active Auroral Research Program) is made to beam more than 1.7 gigawatts (billion watts) of radiated power into the ionosphere -the electrically charged layer above Earth‘s atmosphere. Put simply, the apparatus is a reversal of a radio telescope - just transmitting instead of receiving. It will boil the upper atmosphere. After disturbing the ionosphere, the radiations will bounce back onto the earth in the form of long waves which penetrate our bodies, the ground, and the oceans.

HAARP represents a technology which could lead to a new class of weapons that could change our world profoundly - an all-purpose military tool. If misused, the tool could mess up the weather. It could be used against humanity in a way that would change what people think, believe and feel. It could be used for good or evil, just as a harp can produce the music of Mozart or the melody of a death march.

H.A.A.R.P. and Weather Control

- “The theoretical implication [of Dr. Robert Helliwell and John Katsufrakis of Stanford University in 1974] suggested by their work is that global weather control can be attained by the injection of relatively small 'signals' into the Van Allen belts (radiation belts around Earth) - something like a super-transistor effect” said Frederic Jueneman.
- Yes. The weather can be controlled using HAARP technology.

- A series of weather disasters began in 1960, according to a CIA report mentioned in the editorial, but at the time climatologists couldn‘t look ahead and see that droughts, floods and abnormal temperatures would continue beyond that decade. As if natural disasters weren‘t bad enough, the CIA reported that national governments were already able to manipulate weather for military purposes [using HAARP technology].

As far back as 1958, the chief White House advisor on weather modification, Captain Howard T. Orville, said the U.S. Department of Defence (DoD) was studying ways to manipulate the charges of the earth and sky and so affect the weather by using an electronic beam to ionize or deionize the atmosphere over a given area. In 1966, Professor Gordon J. F. MacDonald, associate director of the Institute of Geophysics and Planetary Physics at the University of California, Los Angeles, was a member of the President‘s Science Advisory Committee, and later a member of the President's Council on Environmental Quality.

Gordon J. F. MacDonald published papers on the use of environmental control technologies for military purposes. MacDonald made a revealing comment: ―The key to geophysical warfare is the identification of environmental instabilities to which the addition of a small amount of energy would release vastly greater amounts of energy. MacDonald had a number of ideas for using the environment as a weapon system and he contributed to what was, at the time, the dream of a futurist. When he wrote his chapter, ―"How to Wreck the Environment" for the book "Unless Peace Comes" he was not kidding around.

In the text MacDonald describes the use of weather manipulation, climate modification, polar ice cap melting or destabilization, ozone depletion techniques, earthquake engineering, ocean wave control and brain wave manipulation utilizing the planet‘s energy fields. He also said that these types of weapons would be developed and, when used, would be virtually undetectable by their victims. He was not some wire haired fanatic when he made these observations in 1966 -he had the credentials of a world recognized scientist. What his futuristic concepts became, are the things which projects like HAARP are made of...

H.A.A.R.P. and Mind Control

- Radio frequency radiation, acting as a carrier for extremely low frequencies (ELF), can be used to wirelessly entrain (adjust) brain waves.
- We are talking about very, very low power requirements. The trick for influencing brain activity is in the combination of frequency, power level and wave form.
- As Dr. Patrick Flanagan, one of America's most gifted inventors noted in an interview, the HAARP project could be not only the biggest ionospheric heater in the world, but also the biggest brain-entrainment (brain adjustment) device ever conceived.
- According to HAARP records, when the device is built to full power it can send very low frequency (VLF) and extremely low frequency (ELF) waves using many wave forms at energy levels sufficient to affect the mental states of entire regional populations.

- The HAARP transmitting system could be used unintentionally or intentionally to alter mental functions.
- If HAARP is tuned to the right frequency, using just the right wave forms, mental disruption throughout a region could occur intentionally or as a side effect of the radio frequency transmissions [in other words: “mind control”].

RFID Technology

Radio Frequency Identification (RFID) microchips are microchips that can be directly imbedded into human flesh. This section discusses the dangers of RFID chips. See Rense (2001) for full review.

- RFID technology links the brains of people via implanted microchips to satellites controlled by ground-based super-computers.
- Today they are small enough to be inserted into the neck or back, and also intravenously (through a vein) in different parts of the body during surgical operations, with or without the consent of the subject. It is now almost impossible to detect or remove them.
- Implanted human beings can be followed anywhere.
- Today's microchips operate by means of low-frequency radio waves that target them. With the help of satellites, the implanted person can be tracked anywhere on the globe.

RFID Technology and the Medical Profession

- One reason the dangers of implantable microchip technology has remained a state secret is the widespread prestige of the psychiatric DIAGNOSTIC STATISTICAL MANUAL IV produced by the U.S. American Psychiatric Association (APA), and printed in 18 languages. Psychiatrists working for U.S. intelligence agencies no doubt participated in writing and revising this manual. This psychiatric "bible" covers up the secret development of Mind Control technologies by labelling some of their effects as symptoms of paranoid schizophrenia.
- The Psychiatric Diagnostic Statistical Manual (DSM) for mental disorders has been a brilliant cover up operation in 18 languages to hide the atrocities of military and intelligence agencies' actions towards their targets. THE MANUAL LISTS ALL MIND CONTROL ACTIONS AS SIGNS OF PARANOID SCHIZOPHRENIA.
- If a target is under surveillance with modern technology via TV, radio, telephone, loudspeakers, lasers, microwaves, poisoned with mind altering drugs via air-ducts, giving familiar smells which cause headache, nausea and so forth, if s/he claims her/his clothes are poisoned, her/his food or tap water as well --- all medical schools teach their students that the person is paranoid, ESPECIALLY if s/he believes intelligence agencies are behind it all.
- Never is the medical profession told that these are routine actions all over the world by intelligence agencies against their targets. Thus, victims of mind control are falsely considered mentally ill and get no help since they are not believed and their suffering is doubled by misinformed health professionals.

Further Implications of RFID Technology

- How many people realize what the implications of implantable chips actually mean? It means total loss of privacy and total outside control of the person's physical body functions, mental, emotional and thought processes, including the implanted person's subconscious and dreams! For the rest of his / her life!
- It sounds like science fiction but it is secret military and intelligence agencies' mind control technology, which has been experimented with for over half a century (since 1950s). Totally without the knowledge of the general public and even the general academic population.
- Supercomputers in Maryland, Israel and elsewhere with a speed of over 20 BILLION bits/sec can monitor millions of people simultaneously. In fact, the whole world population can be totally controlled by these secret brain-computer interactions, however unbelievable it sounds for the uninformed.
- Neuro-electromagnetic involuntary human experimentation has been going on with the so-called "vulnerable population" for over 50 years, in the name of "science" or "national security" contrary to all human rights. It happens today in the **USA, Japan, and Europe**. With few exceptions, the mass media suppresses all information about the entire topic.
- Only increased public awareness of the microchip implants, their frightful consequences to privacy by influencing of individuals' thoughts and actions, causing people to become biological robots with physical and emotional pain whenever the supercomputer technician so wishes, is enough reason to refuse to take the microchip into your body for whatever reason.
- It is the biggest threat to humanity and the most sinister plan to enslave the human race forever.

If you have a choice and want to remain a normal human being with privacy, **DO NOT** have your children implanted **NOR** yourself implanted with RFID microchip(s) (or any other type of implantable microchip). Otherwise your vision, hearing, sensing, thoughts, dreams and subconscious will be influenced by an outsider, who does not have your best interests in mind. For the rest of your life!

Mind Control: MK Ultra Technology

MK Ultra today has evolved from the 1950s variation of mind control (MK Ultra Compendium 1980), whereby drugs such as LSD, and interrogation were used as methods to weaken the mind of the individual to force confessions through mind control. Today, mind control is achieved through the implantable microchip (mentioned above; Rense 2001).

- The brain functions of an implanted person can then be remotely monitored by supercomputers and even altered through the changing of frequencies.
- Once implanted, the U.S. National Security Agency's (NSA) 20 billion bits / second supercomputers could now "see and hear" what you are experiencing with a remote monitoring system (RMS).

- Every thought, reaction, hearing and visual observation causes a certain neurological potential, spikes, and patterns in the brain and its electromagnetic fields, which can now be decoded into thoughts, pictures and voices. MK Ultra technology is therefore capable of Mental Video and Audio Projection, as well as Artificial Telepathy.
- The mass media have not reported that an implanted person's privacy vanishes for the rest of his or her life. S/he can be manipulated in many ways. Using different frequencies, the secret controller of MK Ultra technology can even change a person's emotional life. S/he can be made aggressive or lethargic. Sexuality can be artificially influenced. Thought signals and subconscious thinking can be read, dreams affected and even induced, all without the knowledge or consent of the implanted person by using MK ultra technology.
- Memory suppression technologies are used in conjunction with MK Ultra technology, which enables the programmer to control certain memories the victim remembers. The use of memory suppression technologies and MK Ultra technology allows the programmer to reinforce behaviour and elicit specific conditioned responses.
- Mind control techniques, such as MK Ultra, can be used for political purposes. The goal of mind controllers today is to induce the targeted persons or groups to act against his/her own convictions and best interests. Zombified individuals can even be programmed using MK Ultra technology, to murder and remember nothing of their crime afterward.
- The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct and to control the absolute behaviour and thought patterns of the individual. The purpose of mind control, using MK Ultra technology is to disrupt memory, discredit people through unusual behaviour, to make them insane or to commit suicide or murder.

See Mind-Computer (2012) which discusses how artificial telepathy is achieved. Artificial telepathy, also known as ‘brain to brain communication’, is also possible using MK Ultra technology. Furthermore, review: Jim Cristea (2009); Berkeley News (2011); UC Berkeley Campus Life (2011) CTForecaster (2013); nature video (2013) and Stromberg (2013). The above mentioned articles and videos describe, demonstrate, and corroborate how audio and video projection of the brain is achieved; how dreams can be recorded and projected digitally; how the brain can be scanned to reveal hidden information personal to an individual; and what a person’s underlying intentions are, by using brain scans. Moreover, the articles and videos mentioned above confirm the functionalities of MK Ultra technology.

Furthermore, when our brain functions are connected to supercomputers by means of RFID technology, MK Ultra technology and other implantable microchips, it will be too late for protest. This threat can be defeated only by educating the public, using available literature on biotelemetry (electronic equipment that receives signals from radio transmitters) and information exchanged at international congresses.

Human Cloning

There are currently five different types of clones, concealed from public knowledge. There are Mark 1 clones; Mark 2 Clones; Mark 3 Clones; and Mark 4 clones and reanimated clones. Moreover, there are two types of cloning techniques: duplication cloning and replication cloning.

Mark 1 clones are REM sleep driven clones. However, the technology used for Mark 1 clones causes many side effects and therefore Mark 1 REM sleep driven clones are no longer a preferred choice.

Mark 2 clones are also REM sleep driven clones. Mark 2 clones have fewer side effects than Mark 1 clones, and therefore Mark 2 clones are currently the preferred method for REM sleep driven clones.

Mark 3 clones are independent clones which operate on microchip containing the entire consciousness of an individual.

Mark 4 clones are also independent clones, operating on a microchip which contains the entire consciousness of an individual and is an advanced version of a Mark 3 clone.

These independent clones (Mark 3 and Mark 4) have a lifespan of 6 months to 12 months and require adjustments after this period to run efficiently once more. Without these adjustments, their functionality weakens.

To "reanimate" means "to restore to life; resuscitate; revive". Consequently, reanimated clones are clones which are genetically identical to that of a person who once lived.

Replication cloning is what the public is most familiar with. Replication cloning involves giving birth to a genetic identical of an original where the newborn starts life off as a baby and matures. The newborn is referred to as a clone.

However, duplication cloning is a current concealed advanced from of cloning, and it involves taking as little as two cells from an individual, adding a constant electrical charge to the cells until a complete human being is formed.

Duplication cloning is similar to the process of regenerative medicine as demonstrated by Dr. Stephen Badylak's video "How to grow a New Fingertip" (CBS 2008; Science Channel 2014) where Badylak states: 'A whole human can be grown within 9 months.' On average it takes 5 months to grow a duplicate clone of an original by means of advanced scientific and technological regenerative procedures.

Cloning Centre and Cloning Technology

A cloning centre is a place where clones are produced. Cloning technology are the advancements in medicine, science and technology used to produce duplicate and replicate copies of originals.

Drip Feed Disclosure and Evaluative Conditioning

Drip Feed Disclosure

Drip feed disclosure is the process of supplying information but in small amounts overtime. Drip feed disclosure is also the process of revealing information slowly overtime, possibly telling lies to conceal certain aspects of the truth until the source administering the drip feed disclosure has adequate time to let out the truth in a slow and controlled way, thereby delaying the betrayed partner (in this disclosure, the public) from having the "complete truth" for some time.

Drip feed disclosure is also a method to gauge public reaction used by governments, the media, multinational corporations and organisations as well as high ranking officials to "test" whether the general public is acceptant of the concealed information or not. When the public reacts favourably to the drip feed disclosure, more information is revealed and made public, and it appears to the unsuspecting observer that the people involved in making the disclosure are taking positive steps towards a favourable goal for all. However, when the public reacts adversely; information contradicting the drip feed disclosure is presented, and an expert is presented to the public who voices the concerns of the general public, and therefore it appears that the opinions of the public have been noted and research and development will not continue in the initial stated direction. Nevertheless, the truth remains concealed and research and development continues despite the aversions of the public.

Examples of drip feed disclosure involve articles such as mind uploading / downloading (BEAMS 2007), Mind Clone Robots (Bloomberg Business 2015; RT 2015), The 2045 Avatar Project (2045 Initiative 2015; Borghino 2012). I am here to tell you that I have been a spy for over 30 years and such technological accomplishments which are posted under trans-humanism / post-humanism genre have been realised many years ago and are **available** today. They are just concealed from the public.

Furthermore, Dolly the sheep was announced as the first publicly cloned mammal (Animal Research 1996), but how many readers can say they know that, four years later, a monkey (BBC News 2000), our closest primate, was cloned? This is publicly disclosed knowledge, but I suspect not many people know of a cloned monkey four years after Dolly. This is because there was such a worldwide adverse reaction to genetic cloning when Dolly was made public, that the disclosure of the cloned Rhesus monkey (BBC News 2000) was not made public knowledge on a grand scale.

Evaluative Conditioning

Evaluative conditioning is defined as a change in liking, which occurs due to an association with a positive or negative stimulus (see De Houwer et al., 2001). Simply put, this means that our preferences for brands, products, people and other things can be influenced and even modified by the presence of something we like or dislike strongly (Hale 2012).

In many settings, a neutral stimulus, called a "conditioned stimulus", often coincides with some desirable or undesirable object, called the unconditioned stimulus. An unknown brand, for example, might appear in a commercial that also depicts a happy child. Over time, stimuli that often coincide with desirable objects are perceived more favourably, whereas stimuli that often coincide with undesirable objects are perceived less favourably--called evaluative conditioning (De Houwer, Thomas, & Bauyens, 2001; Walter, Nagengast, & Trassilli, 2005; Moss 2009). An "unconditioned response" is a response to a neutral stimulus we have no / little control over. It is a natural automatic response.

In other words, our preferences for liking or disliking brands, products etc. (the neutral stimulus) can be influenced by embedding (implicitly placing) the brand etc. (neutral stimulus) with positive or negative associations. Overtime our conditioned response becomes one of conditioned favourable or negative response towards the brand, product etc. (neutral stimulus) when we are faced with the brand, product etc. at a future date. Our preferences have been guided overtime. Evaluative conditioning can change our preferences when carried out subliminally or implicitly; it does not have to be explicit.

In media; music; movies; and other forms of popular culture and entertainment, images and symbols (of stimuli which cause undesirable consequences) are embedded subliminally and implicitly and are paired with positive associations as a form of evaluative conditioning. This causes individuals who have no preconceived judgements of the stimulus to be guided to have positive associations with a stimulus which causes undesirable consequences.

In popular culture; media; advertisements; movies; music and other forms of entertainment, evaluative conditioning is used as a method of hinting (that something is wrong); showing off of power (i.e. nothing can be done to stop the negative stimulus) and as a form of gloating (i.e. we've pulled off the negative stimulus / we are pulling it off). In over the 30 years I have been a spy, I have witnessed methods of evaluative conditioning used to hint, show off power, and gloat more often than not, and evaluative conditioning is not used just to guide social order.

The fortunate aspect is that once a person learns that his / her preferences are being guided by methods of evaluative conditioning, in order to influence the person to have positive associations with negative stimuli (or feel powerless towards the negative stimuli); and that the stimulus (brand / product) does in fact cause undesirable consequences –then the 'spell' is broken. The person can now choose how he or she responds to the brand etc. Usually, once all is known: a negative stimulus is associated with negative associations; despite it being portrayed as positive through evaluative conditioning.

Consciousness Transfer

Consciousness is defined as "the state of being aware of and responsive to one's surroundings; a person's awareness or perception of something" (Dictionary Reference 2015). Consciousness can also be described as: individual awareness of a person's unique thoughts, memories, feelings, sensations and environment (Cherry 2015).

John Locke (1632-1704) was an English philosopher, Oxford academic and medical researcher who argues that it is sameness of consciousness rather than sameness of substance that constitutes personal identity. Consequently, if the psychological life is transferred from the body of a prince to the body of a cobbler (shoe mender), Locke argues, the resulting person will be the prince and not the cobbler. He would be responsible for the prince's actions and not the cobbler's; those who were close to the prince could continue their relationships with him but those who had relationships with the cobbler could not, and so on (Schechtman 2012, p.334). Moreover, basic Lockean intuition has proved to be that "consciousness transfer" can be thought of as the feat in which the person moves from one body into another (Schechtman 2012, p.334).

Consciousness transfer can also be thought of as the process of transferring or copying the mental content (including long-term memory and "self") from a particular brain and copying it to a computational device; artificial body or avatar body such as that of a robot or clone version of the original. The computation device, robot or the clone, will then respond essentially the same way as the original brain (as suggested by Lockean theory on consciousness transfer) and therefore the computational device, robot or clone experiences having a conscious mind and essentially the behaviour of the computational device, robot or clone, can be attributed as belonging to that of the original.

It is **OF MONUMENTAL IMPORTANCE** that the reader understands the above statement regarding consciousness transfer is a method of **CURRENT** concealed advances in technology and not a concept relegated only to the genre of science fiction. At the very least the reader should be **open** to the possibility of consciousness transfer and the implications of consciousness transfer, in order to truly begin to understand the extent to which current concealed advances in technology are used to commit monstrous crimes.

I understand the above sounds unsettling, but there are also many wonderful concealed advances in technology (detailed in the disclosure section). The marvels in technological feats and human accomplishment will be released for the benefit of mankind once the world learns about the monstrous crimes which are committed against the them, and the earth, through the use of highly concealed technological advances and the good people of the world band together to put an end to the tyranny around us.

Now is also a good time to remind the reader that if military technology advances by a rate of 44 years for every 12 months which passes, then since 1945 military technology has advanced by a rate of more than 3000 years compared to the technology the public is currently accustomed to (44 times 70 = 3080). Now ask yourself honestly, in 3000 years from now (the year is 5015) don't you think humans will have been capable of developing a method which allows them to transfer their consciousness from one body to the next and more?

I am here to tell you that I have been a spy for over 30 years and that from the intelligence I have gathered over the years it has been illustrated to me countless times that consciousness transfer is indeed fact and currently exists. Consciousness transfer has been achieved since 1945, although this accomplishment in human advancement (and more) is concealed from public knowledge. Moreover, consciousness transfer is a highly advanced form of concealed technology used by high ranking members of society to commit unspeakable crimes against the public.

Furthermore, for the reader who wishes to understand how consciousness transfer is possible, see Petkova and Ehrsson (2008) and Ehrsson (2013). In the video Professor Henrik Ehrsson (2013) demonstrates:

1) Consciousness transfer from one body to the next (owning another body other than ones original)
2) Physiological evidence for owning the new body. In other words when consciousness has been transferred and the person perceives the new body as that of his or her own; when the new body is threatened as the mannequin (new body) was in the experiment, individuals still perceived the mannequin body as that of their own and became frightened (displayed biological and physiological responses).
3) Visual perception and stimulation causes us to perceive ownership of a new body.
4) The new body we inhabit must be similar to that of our own (it does not matter whether the new body is smaller or larger than our original so long as the measurements are proportional) for consciousness transfer to occur.

5) That it is possible for consciousness to be transferred to another person's body and have (or perceive) ownership of other person's body while being localised in your own body. In other words, dual consciousness is possible.

Petkova and Ehrsson's (2008) and Ehrsson's (2013) research gives one of the clearest publicly disclosed explanations of consciousness transfer and I urge the reader to watch the video in order to better understand my disclosure. One of the main findings from his research is that perception is not rigid; and perception does shape reality. Sight and synchronous (going on at the same time) stimulation can, and does alter the brains perception of reality. Sight and synchronous stimulation activates certain parts of the brain (the sensory parts of the brain and the motor (movement) parts of the brain). The match between the two (sight and stimulation) "convinces" the brain that 'hey, I'm no longer in this body; I'm in that one or that I have a third arm (although it is a false limp) etc. Accordingly, "perception is reality".

Disclosure: Save the Victims through your Diligence

I've done my best to preface this information as logically, sequentially and methodologically for the reader as possible. I have stated what the law is on 'whistle blowing'; I have given references and definitions where appropriate and I have provided many sources in order for the reader to research, corroborate, and better understand my disclosure, so that he or she becomes compelled to help me and victims like me, defeat tyranny and save the future of mankind; therefore if there are any areas of my disclosure which still appears suspect –and I understand, after all my efforts –things may still not appear clear to some readers because we are dealing with highly advanced concealed technologies which are not available in the world the general public lives in; and therefore it is difficult to fathom and furthermore conceive that men can be so evil to their fellow humans through the use of advances in science and technology. Nevertheless, I promise you, I have no reason to lie, I am victimised daily because of these technologies, and therefore my only option is to tell the truth, and nothing but the truth so that the people of the world can put an end to this grave injustice.

For the reader who still finds it difficult to accept my disclosure after first read, I want you to do two things:

1) **Give me the benefit of the doubt.** I am asking you to do this because as I stated earlier, one of the hardest things to do as a victim of abuse is to come forward about the abuse you have suffered. It is even more heartbreaking when nobody believes you. Consider all the victims of paedophiles you have heard about on the news. I cannot speak for their individual cases but I can attest to the fact it affects my mental psyche beyond comprehension; there is nothing worse than being ignored and that nobody believes you. So on that premise, please give me the benefit of the doubt, because I'm sure the majority of good human beings out there will feel worse when it is proven that I am a victim of severe abuse and you ignored me, when you could have done everything in your power to save me.

2) **Set out on a genuine quest to debunk my disclosure.** If my disclosure appears suspect to you to begin with, then set out on a genuine and investigative quest to debunk my disclosure because, if my disclosure initially appears circumspect to believe, then I should be proven false within minutes of your research of my disclosure, because even professionals make mistakes. There will be something about the disclosure which will just not 'add up'. However, when a person is telling the truth, then it is very difficult to debunk that person.
I promise you, it will be worth your time to know the real truth about the world, simply because the truth is liberating.

I would also like to remind you that the truth about the world has been hidden from the public for many decades, therefore when you do hear the truth of course it is going to sound absurd to begin with; but keep an open mind and pursue till you know what truth is; because "to "know" is to understand with certainty".

Words of Inspiration: Personalized from a Theoretical Physicist –Thomas Campbell

The following gives an introduction to many aspects which occur in this world behind closed doors or in plain sight depending on how familiar the reader is with the topics. Evidently, a lot of what is written predominantly affects the people of the world and it is time the people of the world stood up and reclaimed their world. I cannot stress this importance enough! The disclosure is also a basis which explains many of the anomalies which exist in the world today. I apologise in advance if the disclosure is too shocking to be believable (initially) and it shakes the foundations of your belief systems... but as I've said, rather than lash out at the information, you must question your own belief system... and probe into the mechanisms which have influenced you to adopt your current belief system... only then does the shock subside... and you can move forward with an open mind through investigation, which should lead you to either corroborate or debunk the disclosure. Keep in your heart: scepticism, optimism and open mindedness (Campbell 2008). Scepticism keeps you from believing fallacies. It keeps you from falling into that which you have not yet proven for yourself. Scepticism is a requirement for breakthroughs. Optimism gives you the ability to give merit to that which you have not yet proven for yourself and Open Mindedness gives you the ability to "see" that which has not yet been proven.

The Science of Hidden Consciousness

Furthermore, "Consciousness Science Kept Hidden" (YesEthan 2013) illustrates:

- "The heart is an electrical organ. It produces by far the strongest form of bio electricity in our body; up to 40-60 times stronger than the second most powerful source –which is the brain. This electrical energy travels through every single cell in our body and in a sense, binds the cells together. The bio-electricity field is strong enough that it can even be detected outside of the body; out into space; beyond the skin. It is very measurable electromagnetic energy, much like radio waves. What we found is that the heart produces an electromagnetic field that surrounds our entire body 360 degrees; and it can be detected about 3-4 feet with megatron meters outside of the body. Researchers at the medical facility in Kansas –say they have detected it 10-12 feet. So regardless of how far it goes, what is interesting is that we produce this electromagnetic field which can be detected; it can be measured, with sensitive mainstream medical equipment." -Howard Martin (co Author) "The Heart Math Solution""
- Scientific studies have shown: that the heart is not only the most powerful centre in the body, it is the most intelligent and has the ability to receive precognition (knowledge of a future event or situation, especially outside one's normal sense of perception). If we can create coherence with our heart and mind centre we can access our intuition more frequently.

When you read my disclosure, remember that it is the heart and NOT the brain which is the most powerful centre in the body; the heart is the most intelligent; it has the ability to receive precognition and intuition about a future event, **before it actually occurs**; therefore if you find that your heart is in resonance with the themes I disclose; do not fight it with the ego of the mind; when our egos get in the way it overrides our heart's intuition; your heart is capable of discerning the truth of an event before your brain becomes conscious of it; by all means listen to your heart, and allow your brain to naturally develop its intuition. It will be the coherence between your heart first, and then your mind centre which will allow you to access your intuition.

Make an effort and reach your own conclusions to the point where you feel comfort in the fact that you **KNOW** the information contained in this disclosure is indeed factual. This will be the point you will have become empowered with golden truth and compelled to fulfil your duty as a human being and also spread the truth far and wide. Yes, we are all living in very interesting times.

Disclosure: Donald Marshall's Message to the World

My name is Donald Marshall. I have been cloned by a large secretive cult know as "The Freemasons" and "The Vril Society" and "Scientologist" together called the Illuminati. For readers unfamiliar with The Illuminati see "The New World Order" (1990) by A. Ralph Epperson and Appendix C in this document; an introductory guide to the Illuminati and their agenda for the world is presented.

Furthermore, I have to tell you that human cloning has been done since 1945. Now is also a very good time to recall Phil Schneider's comment (Schneider was a Geologist and Engineer who worked on black projects and Deep Underground Military Bases (DUMBs)): "That for every 12 months which passes, military technology advances by a rate of 44 years compared to the technology the public is currently accustomed to" (Schneider 1995; 1996; Open Minds 2011). George Green (a former Financier affiliated with U.S. Presidential candidates) also attests to the fact that humans have been cloned since 1938 (Green 2008a; 2008b). Therefore if we take 1945 as a base year, then military technology or hidden technology has advanced by the rate equivalent to 3080 years than the technology the general public is currently accustomed to (2015 – 1945 = 70 years; 70 multiplied by 44 = 3080). So yes, humans have been cloned since 1945 and continue to be cloned today and that current hidden technology in the year 2015 is as much as 3000 years ahead of the technology the public is currently accustomed to. I'll repeat myself because it is **very important** the reader understands the above statement, even if, the rest of the disclosure is hard to fathom after first read. The two **most important** things to keep in mind are:
1) Humans have been cloned for over 70 years (since 1945);
2) **Present** hidden technology is more advanced -as much as 3080 years more advanced- than what the public is currently accustomed to and technology continues to advance at an incredible rate.

The Two Types of Cloning Techniques

Now there are two different types of cloning techniques, there is replication cloning and duplication cloning. Replication cloning is the type of cloning the public has generally heard of. Replication cloning involves taking the nucleus (the DNA) out of a donor egg, and replacing it with new DNA from the person to be cloned. After a few days the resulting embryo can be implanted for pregnancy. The newborn starts life off as a genetic copy of an original.

Duplication cloning, on the other hand, is the second hidden method and type of cloning where the DNA from the person is grown in a big thick tank full of (salty) water. It involves a method of regenerative technology where the cells are agitated and agitated and over the course of 5 months, a fully formed duplicate clone body of an original is developed. The Illuminati used to have to use a tissue sample from the original or the cells from women's pap smears because this contained rich cells to make duplicate clones; they also used children's foreskins which were discarded at the hospital that got removed. That is what happened to me. I had my foreskin removed at age 4 –and by age 5 the Illuminati grew duplicate clones of me. Now the Illuminati say they have upgraded the technology since 2000 and now all they need is blood from the original. They then agitate the blood cells over and over again through regenerative technology, until a fully formed duplicate clone body of an original is produced. For readers who are unfamiliar with regenerative procedures, see Dr Stephen Badylak's video on "How to grow a new fingertip" (Science Channel 2014) and Carmichael (2013) which discuses how scientists cloned a mouse from a blood sample (this is another example of drip-feed disclosure). Regenerative science and medicine (CBS 2008) does work; I've seen it many times.

The Different Grades of Clones: "Mark 1" to "Mark 4" Clones

There are a few different grades of clones that I know of. There are: "Mark 1" to "Mark 4" clones. Mark 1 clones were available at the end of World War II (1945). Mark 1 clones were rapid eye movement (REM) driven clones but they were primitive grades of cloning with lots of side effects caused to the original. Mark 1, REM driven clones were called "Organic Robotoids", even though there are no robotic parts to the clone at all. See Dr Peter Beter's discussion on "Organic Robotoids" (Beter 2011).

Donald Marshall's Dilemma with "Mark 2" Sleep Driven REM Clones

The problem I am having is with Mark 2 REM driven clones. "Mark 2" is a rapid eye movement (REM) driven clone. In other words, Mark 2 clones are sleep driven clones. REM sleep is the fifth stage of sleep (Sleepdex 2015). The first REM cycle usually happens 90 minutes to 110 minutes **after** we fall asleep (Sleepdex 2015). What that means is: **currently,** this is the Illuminati's main form of communication. They do not call people on the phone; they do not meet at the Bohemian Grove anymore (they only meet there once a year for traditional purposes). Since they discovered the science of sleep driven cloning, they meet at the cloning centre **WHEN THEY GO TO SLEEP** (The cloning centre is a physical location, located 5 / 6 hours radius from the Robert Pickton Farm (Port Coquitlam, British Columbia, Canada) at a nature reserve).

Consciousness Transfer Happens When the Original Reaches REM Sleep

The Illuminati can transfer the consciousness of an original once the person reaches REM sleep if there is a duplicate clone of the original at the cloning centre, once the person goes to sleep (90 minutes to 110 minutes after falling asleep). So what happens is that the consciousness of the original is transferred to the duplicated clone body (the duplicated clone body is grown within 5 months by regenerative technology) once the original reaches REM sleep. The original's consciousness is transferred from the original's body, although the original's body is still in their bed, asleep at home; to a duplicated REM driven clone body at the cloning centre and the original 'wakes up' as a cloned version of himself / herself at the cloning centre. It is a great marvel of science and one man's greatest achievements, but it is kept hidden and used for sinister purposes which I discuss below. I must address the readers at this point who are lost at the thought of duplication cloning, and consciousness transfer.

For readers who may have difficulty understanding the process of duplication cloning, see the video which features Dr Stephen Badylak (CBS 2008; Science Channel 2014). Badylak describes and illustrates how a new fingertip can be grown within 4 weeks, by the process of regenerative technology. Duplication cloning works a similar way; involving regenerative technology whereby the cells are agitated and agitated, and over the course of 5 months a fully formed duplicate clone body of an original is grown. Moreover, readers should also view the Horizon documentary "Why Do We Dream?" (2009); the documentary clearly explains the phases of sleep; particularly REM sleep. The documentary also clearly explains that during REM phase sleep the whole body shuts down, and it is only the brain which is active. The fact that the brain is still active, although the body is inactive, is a perfect opportunity for consciousness transfer; and this is usually the moment the Illuminati transfer the consciousness of a person who is asleep, to a cloned version of himself / herself. The reader should also explore Ehrsson's (2013) lecture on consciousness transfer. In the video, Ehrsson (2013) clearly explains how through the first person visual perspective (seeing through the eyes), and through synchronous (occurring at the same time) stimulation of a body part, individuals can perceive and see the world from another body different from their original. In other words, so long as people can see through the eyes, and the body is stimulated, the match between the visual perspective and feeling body sensations allows the individual to see and perceive the world through another body different from their original. The consciousness has been transferred to another body.

Ehrsson (2013) also demonstrates that consciousness is linked. Once consciousness is transferred (from the original's body to the new body), even though the new body is not the persons original body; when the new body is attacked, because the person perceives the world through the new body, everything feels "very real". When the new body (for which the person's consciousness has been transferred to) is attacked, the original still perceives the threat as "real", and therefore the person displays a biological and physiological fear response in their original body; the heart rate increases, as does respiration (breathing) rate, as well as anxiety.

Donald Marshall is an Original as he sits at Home Typing this Disclosure

This is how advanced technology is today. This is also exactly what happens to me. **I am an original as I sit at home here typing this disclosure,** but when I enter REM sleep, my entire original body shuts down, and only my brain is active. The Illuminati have linked my consciousness to a REM driven duplicate clone body of me. Therefore, as soon as I enter REM sleep, they are alerted to the fact that I have entered stage five of the sleep cycle, and in REM sleep, because they have a green and red light above duplicate clones at the cloning centre. The red light indicates that the original is awake, and the green light lets them know that original has entered REM phase sleep and it is time to transfer their consciousness to a duplicate clone version. They cannot transfer the consciousness of an original before REM phase or when the original is awake. When they try and transfer consciousness before an original is in REM phase it makes the original have intense headaches. Setting an alarm every 90 minutes to wake up and then go back to sleep also does not work in terms of avoiding REM sleep and having your consciousness transferred, because the body requires REM sleep, and therefore you become very tired, quickly, if you do not get REM sleep. If you stay awake for four days without REM sleep you begin to hallucinate; after seven days without REM sleep you will die.

When Donald Marshall's REM Driven Cloned Experiences Began

This whole sordid episode began when I was 5 years old, after going to the doctors to get my foreskin removed, at age 4. The tissues from my discarded foreskin were used to grow multiple duplicate clones of me. Within months, I was having my consciousness transferred to a REM driven, 5 year old clone version of me every time I entered REM sleep. The reason the Illuminati did this was because they wanted to use me as what they term "a diddle kid" – in other words they wanted to have sex with an REM driven, 5 year old, duplicate clone version of me. This is what they do. They clone people. They make duplicate clone bodies of people to victimise in terrible ways, and they clone children for men with undeveloped penises to have sex with.

However, I could never remember any of these REM driven clone experiences when I woke up for 25 years; because another aspect of REM driven cloning is that the Illuminati have the advantage of being able to suppress the experiences of people who have their consciousness transferred to their REM driven duplicate clone versions through advanced memory suppression technologies. For the reader who wishes to understand how memory suppression works see Winter (2014). Memories are stored and retrieved from different parts of the brain. Interactions between the cerebral cortex (the outer layer of the brain) and hippocampus need to work together in order to bring the memories out of mental storage to be re-experienced by the mind (Winter 2014). Therefore all that has to be done is for them to turn off the nerve cells which communicate with the cerebral cortex and hippocampus for the evening which you went to sleep and had your consciousness transferred to a REM driven duplicate clone version of you, and you will not remember anything. The technology is so advanced; they can, and do suppress memories at a push of a button. Therefore, the victim wakes up the next morning and will not remember a single REM driven duplicate clone experience. All that they will be aware of is that they did not have a dream the previous night.

This is what happened to me for over 25 years. I wasn't allowed to remember these REM driven duplicate clone experiences until I turned 30. My memories were suppressed for 25 years. I would wake up many days, after a night's sleep, thinking 'I didn't have a dream last night'; or I would wake up the next morning and I would be sick, although I went to sleep the previous night feeling completely healthy. Throughout the years, I thought I had terminal illness; because whenever I went to see a doctor, the diagnosis would always come back that 'there was nothing wrong with me'.

However, over the course of 25 years when I was memory suppressed, and not allowed to remember my REM sleep driven clone experiences, the Illuminati were using me to produce songs, lots and lots of songs, and whenever I could not produce a song, or did not want to, they would stab my REM driven clone body; sodomize my REM driven clone body; chain me to a crucifix and burn my REM driven clone and torture me for all kinds different reasons; and even when I was compliant and did what they told me to, I was attacked, simply because some members of the Illuminati wanted to know what it feels like to stab someone. Now because consciousness is linked (Ehrsson 2013), all these REM clone torture experiences affected me in my original body, and I have a weak heart today because of it, at the age of 39. The memories of my REM clone driven experiences also had to be released to me slowly (see Alford (2014) and FW: Thinking (2014) –for an example of how memories can be wiped and restored) over the course of many months because I had been through so much horror and torture that if the memories was released all at once, I would have had a heart attack in my real body and died.

This is one of the main reasons the Illuminati currently use REM sleep driven cloning (besides sex and torture): to plagiarise talented people under duress. There are many people including me who have been, and are plagiarised on a daily basis. Their consciousness is transferred to their REM driven clone alternates, and night after night and under duress they are forced to reveal their money making ideas or face torture or clone death; clone death after clone death. That is what the movie *Inception* (2010) is about: REM sleep driven cloning, and stealing ideas from the minds of unsuspecting individuals. The Illuminati actually made the film *Inception* (2010) as well as *The Island* (2005), *Avatar* (2009) and *The 6th Day* (2000); the films mentioned above contain references of cloning and REM driven cloning to show off the Illuminati's 'power' and throw in the world's face and laugh.

The Science of REM Driven Cloning Began in 1945

I am one of the few people fortunate, or unfortunate depending on how you look at it, who remembers all of my REM driven cloning experiences because these memories were released to me at the age of 30. I remember many people who have attended and still attend the cloning centre which they bring me to over the course of 34 years. I am currently 39 years old and was brought there since the age of 5. The science of REM sleep driven cloning was first discovered in 1945, and at first it was just a political thing which leaders of the world did. Through the science of REM sleep driven cloning, political figures, heads of state and royalty met each other as REM driven clones versions of their selves when they went to sleep and discussed worldly affairs in complete secret with each other. They also did whatever they wanted with each other as REM driven clones, which they could not do in their original bodies such as sex, fighting each other to the death, jumping off cliffs as their REM driven clones, and just about anything you can think of when one has the ability of pseudo-immortality as a sleep driven clone version of themselves hidden from the guise of the world and general public.

Well they soon got bored of each other very quickly, and started to make REM driven duplicate clones of movie stars, musicians, celebrities and public figures from all walks of life to hang with as REM driven clones and now they all get together for a disgusting time as REM driven clones when they go to sleep. Most of the G20 political leaders meet at the cloning centre as REM driven clones when they go to sleep to discuss worldly matters and watch gruesome things done to innocent and unsuspecting civilians, who are also REM sleep driven clones at the cloning centre, but civilians have their memories suppressed. This is what the films *Hostel* (2006); *Hostel: Part II* (2007); and *Hostel: Part III* (2011) (Illuminati made films) -are about: -rich people paying to torture and kill innocent people at a secluded location for sport; just that at the cloning centre, the Illuminati members torture and kill REM driven clones of innocent people.

Therefore, depending on what was done to your REM driven clone, you could wake up the next day with a very intense headache in your **original body**, if your REM clone was repeatedly punched in the head; an upset stomach in your **original body** if your internal organs were tortured; achy limbs; and worst of all, feel sick all over; although the previous evening you went to sleep feeling healthy. Remember consciousness is linked (Ehrsson 2013), therefore whatever happens to your REM driven clone will affect you in your original body. Even when they suppress your memory; the previous night's experiences still affects you in your original body. You wake up with the knowledge that you did not dream the previous evening when you went to sleep, and with certain illnesses depending on what was done to your REM driven clone.

Help Donald Marshall STOP REM Driven Cloning NOW!

Everything I discuss in this disclosure is firsthand witness accounts. I have seen and experienced horror spectacles done to innocent civilians and to me. For the sake of humanity and the innocent children brought to the cloning centres all over the world; I must tell the whole world about this and the good people of this world **MUST HELP ME TO BRING THIS TO A STOP.** My sense of duty, moral correctness, and conscience cannot allow me to stay silent about this evil. REM sleep driven cloning, and the torture of innocent civilians as REM driven clones; having REM driven clone sex with beautiful, innocent unsuspecting civilians of the world; the paedophilia of children through the process of REM driven cloning; REM driven sex slavery; the plagiarism of talented individuals through the process of REM driven cloning; and REM driven idea slavery, whereby talented individuals must produce new ideas and concepts daily (or whenever they sleep) or be stabbed or tortured in their sleep as REM clones –is the most vicious and pernicious form of tyranny ever to occur against humanity and **IT MUST BE STOPPED NOW!**

Do Not Panic, Riot, or Cause Chaos. This is Very Important.

Now I must preface this carefully, because members of the Illuminati have told me that I have to put this in an eloquent fashion, in a way which does not make people panic because people finding out about REM driven cloning, and the extent of the evil it has been used for, and continues to be used for; it could cause loss of social order, riots and anarchy in the streets. Moreover, I too **DO NOT** want riots and anarchy in the streets, despite the fact that I am vehemently angry considering the extent to which they plagiarised my talents over the years. If you are a good person reading this, and you want to help and you want social order restored for the benefit of mankind, promise yourself; me; all the innocent children they have affected through REM driven cloning, and the children of the future that you will **NOT** riot, and destroy a world for children who are going to inherit the world. **This is very important**. Remember the Illuminati have highly advanced technologies (kept hidden and secret), including weaponry, and they are just waiting for any excuse to use it on the populace. In all revolutions, the populace always win, and I want this disclosure and the end of REM driven sleep cloning to go smoothly. I want the good people of the world to keep spreading this information, keep spreading this disclosure all over social media, tell your close friends, your family, and as many people on social media platforms as you can. The internet, and radio shows with small audiences are the only forms of communication the Illuminati do not control. They own all forms of television networks, even including the Aboriginal People's Television Network (APTN), and therefore a message as important as this will never reach the average man or woman through television because it does not serve their agenda.

Spread this Disclosure Document until it reaches the Armed Forces

I want this message to reach the armed forces, because currently people in the armed forces are following orders, yet they do not know how corrupt their governments and people in high profile places are. They are unknowingly defending these corrupt people. People in the armed forces will **NOT** defend these corrupt people nor will they harm civilians when they realise that these corrupt people have been growing duplicate clone bodies of civilians, transferring civilians consciousness to their duplicate clone versions when the victim reaches REM sleep, and torturing civilians in their sleep; having sex with under-aged REM driven cloned children; having sex with innocent and unsuspecting adults as REM driven sleep clones, against their will; torturing REM driven clones for sport; and for money-making ideas to benefit their own pockets while they make innocent civilians sick and have side effects in their original bodies from REM sleep driven cloning technology. No. The armed forces will not accept that. Therefore keep spreading and sharing this information for everyone to see so that it reaches personnel in the armed forces **ASAP!** I want the armed forces to overthrow these corrupt people. The Illuminati think they are very sly and untouchable because together they hold most of the wealth in the world. We must show them that they are not untouchable. Please do all you can, to share and spread this message. It is the most important message to ever reach the internet because it affects all our liberties.

I must also address the readers who may feel afraid in spreading my message because they may want to do the right thing, but they are terrified the Illuminati may degrade or end their lives. As unbelievable as the next thing I am going to share sounds, it is a belief system which shapes the reality of the Illuminati and therefore they themselves are afraid to do anything which may degrade their lives or 'eternal soul'.

They believe that they are the 'fornicators' mentioned in The Bible, and if they were to degrade or harm the lives of anyone aiding me, they would suffer the wrath of God when they die. Notice I said **when they die**, but so long as they are still alive they can do all the evil they like, with little consequence, as they do in their REM driven clone versions at the cloning centre; worshipping Lucifer and doing all sorts of ungodly things. They even say God does not exist, only science and technology; yet they are afraid of dying. Very afraid of dying; because they believe they'll meet God's judgement. I'm just relaying what they have told me. They are weird like that. They also follow Hopi Indian Prophecy, Mayan Prophecy and Nostradamus Prophecy. They mix and match those three prophecies and come up with their own religion of what may or may not happen in the future.

Donald Marshall is NOT "The New Age Saviour". Nobody wants to be a 'saviour'

Furthermore, **THEY** believe I am a new age saviour for the end times; that I was going to save the world from something, according to the Nostradamus prophecy (quatrains); the Illuminati consider Nostradamus the greatest prophet ever to walk the earth. They are in awe of him.

I **DO NOT** endorse their claims. I just want REM driven cloning to stop and these people face punishments for their crimes against humanity. I really do not endorse such claims, but if it saves my life and yours, so be it. In reality and practicality, nobody wants to be 'the saviour'; not with such tyranny and pernicious evil such as this. We'd rather all just be living normal lives without the knowledge and practice of REM driven sleep cloning which is used for torture, clone death spectacles and sex with children. My friends on Facebook do not want to be heroes or 'saviours' either; nobody really does, but we do it out of a sense of duty, moral correctness and our conscience would never rest knowing the true extent of evil in this world, and doing nothing to end it. So all that saviour and Nostradamus prophecy talk, it is just the Illuminati; they like to quote prophecies, and make them come true to make themselves feel special.

Like I said, they are weird. I'm just relaying what they say, and I **do not** endorse the new age saviour claims. I'm just a normal guy, who was unfortunate to be cloned when he had his foreskin removed at age 4. This could have happened to anyone. Thankfully I had a very active imagination at the age of 5 and with the use of the technologies such as Mind-voice technology (that is what the Illuminati call it) it allowed me to make songs to keep the perverts off me. Mind-voice technology is just a technology which can tune into the inner voice in your head (See articles by Prigg (2014) and New Scientist (2014) –which discuss how scientist can listen to your inner voice; –it is similar to how Mind-voice technology works). Therefore, when you hear the sound of a guitar; the sound of drums; or any instrument for that matter; after hearing the sound, when you imagine the sound you heard, your REM driven clone can replicate this sound with Mind-voice technology **exactly**; and I used Mind-voice technology to make many, many songs over the years. As I've said, and I'll say it again: technology today is very advanced.

How the Nostradamus Prophecies shapes the lives of the Illuminati

Also another reason they have not killed me is because the Illuminati said "If they do, everyone is going to know I am telling the truth", -which I am with my right hand to God and I cannot lie about this stuff; the things I have seen are too sinister and diabolical. It would make me as bad as them if I did. Furthermore, because they follow the Nostradamus prophecies religiously, they have told me that they believe anyone helping me is considered part of "The Army of Light" and again, to hurt or degrade anyone helping me, considered part of The Army of Light would incur the wrath of God; and that to hurt anyone considered part of The Army of Light will degrade their lives; they will suffer misfortune and entire ruin as will their eternal soul. These are just some of the interpretations of the Nostradamus quatrains they have relayed to me. These people are beyond crazy, zealot and religious fanatics. Nevertheless, the main point to remember by all this is that **you are safe**, and **they cannot hurt you**, **they are scared to hurt you**, as well as kill me because prophecies have shaped their lives and the lives of their ancestors for hundreds of years.

Furthermore, and practically, I have told too many people that this is the extent, to which technology has developed in our current day, and it is kept secret and hidden from the public for nefarious purposes and this is what the Illuminati do: REM sleep driven cloning. I have been spreading this information **since 2011** and not one of these high profile people, government officials or celebrity figures has issued a (public) statement against me or taken me to court over libel charges because I'm telling the truth; and their statements would not hold in court.

I have also told over a million Arabs; I have appeared on radio interviews with Vinny Eastwood (Vincent Eastwood 2013); Jeanice Barcelo (Jeanice Barcelo 2013); and Lisa Phillips on the Cry Freedom Radio show (Astral 7ight 2013a – 2013h). I have reached audiences as many as 280,000 views on the Vinny Eastwood Show (Vincent Eastwood 2013) alone. However, I must stress that YouTube reduces the view count of my interviews every so often in order to suppress the truth so that my interviews do not get featured on their platform. Nevertheless, there is safety in numbers and rest assured there are new people waking up to the disclosure of REM driven sleep cloning every day; and therefore you can share and spread this information **without any worry** about facing reprisals from the Illuminati, even if you do not believe that it is in fact their beliefs in Nostradamus prophecies which is saving you and I from the Illuminati.

HIV/AIDS has been cured; so have many forms of cancer

Now is also a good time to share with the reader that this situation is not all 'doom and gloom.' In their quest to perfect technologies, science and medicine, the Illuminati discovered the cure for HIV/AIDS; cancer (except pancreatic cancer); Alzheimer's; Dementia and many more debilitating diseases which humanity suffers with. They relayed the fact that they found the cure for HIV/AIDS by stimulating the cells in an oxygenated rich environment because diseases/viruses cannot thrive in an oxygenated rich environment (in layman's terms). Magic Johnson was cured from AIDS this way and they use him as a spokesperson to sell the retrovirus pills which they know is not as effective. However, the greed these people have, knows no bounds, and because they receive too many donations for HIV/AIDS as well as cancer donations they do not announce they have cured HIV/AIDS or cancer publicly. As I've told you, and will continue to tell you: technology is far, far, FAR advanced than what you currently see around you. With the help of the good people of earth I promise to release these technologies for the benefit of mankind, and all reading this disclosure can hold me to this. You **must** hold me to this. It is time we ended corrupt governments and corrupt individuals in high positions of power in this world. Hold me to this.

I hope you are now beginning to understand why this is a **world emergency** and why these corrupt people must be ousted and overthrown. I also hope you understand why you must not riot or cause chaos because as I said at the beginning of this disclosure "I write this to empower good people of the world against the tyranny which exists all around us in our world today"; I **did NOT** write this to cause chaos, public dissension or anarchy. Good people of the world will inherit the earth once these corrupt people are overthrown. Therefore, under no circumstance must you destroy the world when you are going to inherit it. No matter how angry this disclosure makes you; no matter how angry the diabolical people who commit these crimes against humanity make you feel. Let's make sure this goes smoothly. Keep spreading and sharing this information, until this disclosure reaches the armed forces; until the armed forces bring these corrupt people to their knees. We are about to inherit the earth; by all means let's ensure everything goes smoothly. Now that you understand how important this disclosure is and that the overthrow of these corrupt individuals must progress smoothly, I'll continue.

The Behaviour of High Profile People at the Cloning Centre

The high profile individuals who attend these REM driven clone gatherings in their sleep have nothing better to do than to show off in disgusting ways. They have no shame and it seems nothing embarrasses them. Some REM sleep driven, cloning centre attendees, sit in the stands of an unused arena, smaller than a hockey rink, but it still has the capacity to seat between approximately 300 to 400 people. There is dirt in the centre of the rink where ice would be. They have frightened REM driven clones of children walk into the middle of the dirt rink to be victimised for a bizarre and disgusting spectacle. Sometimes, they have animals like dogs have sex with the REM driven clone children, while a man holds the dogs on leash so that it wouldn't bite the child on the back of the neck; which I have seen happen before. They all try to outdo each other in their levels of depravity; to be evil is to be "cool" to them.

The Ring Leaders of the Cloning Centre

I will now discuss the ring leaders, but before I name names, please understand that I am not someone who has collated stories and evidence together over the internet. I REALLY **do not** need to do this. Every time I enter REM sleep, the ring leaders at the cloning centre transfer my consciousness to a duplicate clone version of me. I have seen and talked with these monsters over the course of 34 years as REM driven clones at the cloning centre. However, I do understand that for the average reader, without understanding how far medicine, science and technology has advanced; without seeing or hearing people, other than me, discuss the crimes the ring leaders of the cloning centres commit, it is very difficult to imagine that this is what these high profile people are involved with, and do. Nevertheless, I guarantee it; this is the true extent and prevalence of evil in our world.

Their public personas, and public images is well managed by their PR teams, and therefore for the average person it is difficult to imagine these people in any other way but "positive"; but these people behind closed doors are the most evil, sick and twisted people I have ever had the misfortune of encountering. It is also difficult to take in what I am saying because I am not a public figure, nor do I have the "influence" or public image these people have; the general populace have never met, or heard of me, or known who I am; if this was the case my eye witness testimony alone will be enough; however, I do understand that we are dealing with highly advanced technologies, thousands of years advanced compared to what the general public is accustomed to, and because of this I provide references for the reader; I provide references for the reader to have something to corroborate what I am disclosing; and I provide references for the reader to see past the illusions these people have created.

Moreover, as "Consciousness Science Kept Hidden" demonstrates (YesEthan 2013):
remember that it is the heart and not the brain which is the most powerful centre in the body, the heart is the most intelligent; it has the ability to receive precognition and intuition about a future event, **before it actually occurs**, therefore if you find that your heart is in resonance with the names mentioned; do not fight it with the ego of the mind; when our egos get in the way it overrides our heart's intuition; your heart is capable of discerning the truth of an event before your brain becomes conscious of it; by all means listen to your heart, and allow your brain to naturally develop its intuition. It will be the coherence between your heart first, and then your mind centre which will allow you to access your intuition.

The references I provide for the reader to corroborate what I am about to say includes: "Appeal from Survivors of Canadian Genocide" (inifiniLor 2013); Tila Tequila –"Missing Children and Cloning Centres" (Astral 7ight 2013i); "Royal Babylon by Heathcote Williams" (MrCowshedder 2012); and "THIS MOVIE WILL BLOW YOUR F%SNG MIND" (KafkaWinstonWorld 2014).

Watch all the above videos and you will easily see all that I disclose. In Appeal from Survivors of Canadian Genocide interview (inifiniLor 2013): Listen to the indigenous people Stee-mas and Wahtsek speak about the pain and torment they have suffered under the Crown, the Vatican, the government, and churches of Canada. 50,000 to 150,000 indigenous children have gone missing because of the involvement of the Crown, the Vatican, the government, and the churches of Canada; and all these factions involved know this!

Tila Tequila is a television personality who has attended the cloning centres as a REM driven clone since childhood. In Tila Tequila's radio phone in (Astral 7ight 2013i), the audio is not very clear for the majority of her disclosure; she was also frightened and therefore she sounds most animated; however her words are clear at the following points of the audio (I also provide a transcript in the appendices section: "Appendix A"):

4.00 min: Cloning Centres and they take your children...
4.25 min: –There are these CLONING CENTRES –I'm not even talking about child molestation here...
5.40 min –They take your children –they not only molest them; men f*** them; and make them shoot each other...
7.59 min- There is a point where you cannot just turn the other way, you know this stuff is going on and you go 'oh well, you know... that's their problem; let's just turn the other cheek. How long are you going to turn the other cheek until it happens to your own freaking children?

10.34 min - Do you know why they love children? Because they are innocent souls; they are innocent... they are the most innocent, pure beings in this planet. They are not harmed by anything. They are new to the world; bright-eyed, pure innocent children. That is why these disgusting paedophile... and these clone rings, cloning centres, satanic rituals, Brownsville Texas... there are many of the of the cloning centres where they take your children that go missing.

Royal Babylon by Heathcote Williams (MrCowshedder 2012) addresses the fact that for Brits, despite their inordinate pride about their tradition, it is often revealed they know little about it than anyone. The documentary then goes on to address the fact that Prince Phillip the Duke of Edinburgh, has a lust for blood-sports which includes shooting endangered wildlife and mammals which he considers 'culling'. The documentary is also notable for demonstrating to the viewer that Queen Elizabeth II redistributed as much as 2.1 million for her private income from 276 people from Merseyside and Lancashire. These people had no will, so their property was grabbed. The same is true for the Duchy of Cornwall. If you die without making a will, everything you possess will go to Prince Charles. Furthermore, Queen Elizabeth II is the largest land owner on earth. Queen Elizabeth II: head of state of the United Kingdom, and 31 other states and territories is the legal owner of about 6600 million acres of land; one six of the earth's non ocean surface. She is the only person on earth who owns whole countries, and who owns countries that are not her own domestic territory. This land ownership is separate from her role as head of state; is different from other monarchies where no such claim is made –Norway, Belgium, Denmark etc. The value of her land holding is £17,600,000,000,000 (17.6 trillion Pounds) (approximately)).

In "THIS MOVIE WILL BLOW YOUR F%SNG MIND" (KafkaWinstonWorld 2014), readers with a deeper knowledge of the Illuminati and the agenda of the Illuminati will easily spot the minor inaccuracies in the documentary. However, the documentary provides a good reference and foundation for anyone new to the topic of the Illuminati and their agendas for the world. The documentary demonstrates how the world is currently a deception within a deception coated in reverse psychology. It gives widespread insight into the Royal family of England, and their lineage.

Furthermore, the documentary provides information that the world's most popular religions have been infiltrated with Illuminati dogma, in secret. The documentary details how (some) modern Christians, Jews, and Muslims do not (yet) recognise their common enemy because that enemy is invisible (until now). The enemy hides within their own religions. The enemy also hides within the activist community (anti-Illuminati movements), it finances big budget documentaries and movements. The end game is to sell a divisive (creating strong disagreement), anti-capitalistic, atheistic message and deliver trusting followers right into the lap of a communist (one government) New World Order. The enemy is the Illuminati.

The documentary also details that: "Today, the bloodlines of this unholy alliance are on the final stages of establishing one world religion, one world government, and one world ruler [--that is **IF**, the good people of this world do nothing to stop them]. The Windsor Royals are at the control of this web of deceit. It is a known fact that Adam Weishaupt (the founder of the Illuminati) took refuge and got assistance from the German Saxe-Coburg Gotha Royal family. The Saxe-Coburg Gotha's changed their German name to Windsor (in 1917). The Windsor's public image is a façade, to hide the cesspool from which they operate behind the scenes; their power and control cascades out of Buckingham Palace to the elite bloodlines who run world affairs."

For anyone who thinks they know they Royal family of England, did you know that still to this day: they have a person appointed called "Groom of the Stool"? What is a "Groom of the stool", you may ask? A Groom of the stool is responsible for stools. The kind you sit on and the kind you expel. In other words a Groom of the stool wipes Royal bottoms. The above has been confirmed to me by Prince Charles of Wales himself as REM driven clones at the cloning centre, and I quote: "It gives you the feeling of being completely taken care of" when I asked Prince Charles (as a REM driven clone) why they still have that. But don't just take my word for this; research "Groom of the stool".

The Royal Family of England

As most of readers have already guessed; and as unbelievable as it may currently sound to some readers, the ring leaders of these cloning centres includes the royal family of England. As I've said REM driven cloning started as far back as 1945, and it was just a political thing, but Queen Elizabeth II has been involved with REM driven cloning since its inception! Yes! Queen Elizabeth II; Phillip, Duke of Edinburgh and Prince Charles of Wales are the worst of them. They are unbelievable depraved perverts all showing off to the celebrities who attend these REM driven clone gatherings. Elizabeth has the REM sleep driven cloned children at the cloning centre call her "Lilli-bet" and Elizabeth as a REM driven clone herself does ungodly things to these REM driven cloned children. Some she fakes being nice to; some she is terrible to; and as a REM sleep driven clone Elizabeth cuts these REM driven clone children with swords while they scream. The decent people at these cloning centres who have had their consciousness transferred into their duplicate REM driven clones, and are there against their will, like I am; they are terrified to say anything against Elizabeth and her cronies. Most of these decent people also have their children's consciousness transferred into their REM driven duplicate clones, and Elizabeth and her cronies hold these decent peoples' children as hostage; to be torn apart; clone death after clone death if they have the slightest inkling of informing anyone, but as Elizabeth and her cohort have been torturing me as REM driven sleep clones anyway, I will disclose all that they do.

Vladimir Putin is also there, as an REM driven clone, and he loves to put the fear of torture and death into people, but he is essentially a perverted coward himself. Most of the famous people who attended the cloning centre as REM driven clones are ashamed to speak or be seen by me there very much; because they are ashamed of the perverse and disgusting gatherings. I am a decent person, even when my consciousness is transferred to my REM duplicate clone alternative, and I will not participate in these depraved acts; so they use me as an example and torture me for being a good person. They have drugged my REM driven clone, and as I've mentioned because consciousness transfer is real (Ehrsson 2013) and because consciousness is linked, therefore everything feels "real as real"; when I'm drugged as a REM driven clone, I behave exactly as any person would on hard drugs (although I have never taken any drugs in my original body); my vision is blurry and my balance affected; they have used this as method of having me do disgusting things in a blurred state, videotaping these dirty acts and editing a highlight reel which also depicts me doing disgusting things to show to the world when the world learns of their depraved natures and their public personas are shattered so that I can be despised as much as they will be when the entire world learns about this. Furthermore, because of the microchip inserted in my REM driven clone; they mind controlled me like a "Manchurian Candidate" (*The Manchurian Candidate* 2004) using MK Ultra technology, which is far more advanced than the declassified MK Ultra techniques (MK Ultra Compendium 1980) mentioned online today which references MK Ultra techniques in the 1970s. See Rense (2001) for a detailed explanation of how implanted microchips work, and how these chips allow the programmer control victims. C.S.I.S Canadian intelligence are all involved; they have REM driven clones too; including a lot of Commissionaires and Canadian Prime Minister Steven Harper himself; They do what Elizabeth says, and seem to follow her every whim without question. It doesn't matter who you are; if you have a cute child, or a talented child with imagination, creativity and moneymaking potential (such as business smarts; song making ability; literary / oratory ability etc. –you name it so long as your child has moneymaking value to them) they will send for the blood record of your child; agitate the blood cells over and over until a duplicate clone of your child is grown; they will then transfer your child's duplicate clone body from the tank they were grown onto a steel rack; the steel racks contain 5 rows and each row holds a duplicate cloned body; once your child goes to sleep in his or her original body and once they enter REM sleep (90 minutes to 110 minutes after they first fall asleep) the Illuminati are alerted to this fact by a green light above your child's duplicate clone at the cloning centre and they then transfer your child's consciousness to the REM driven duplicate clone; this is the point where the prime evils molest your children; and so your children are being molested in their sleep as REM driven clones and then they suppress the child's memory or implant false memories (see Kim (2013) and Alford (2015) –on how false memories can be implanted) so your children wake up not remembering dreams from the previous night, or if they do remember anything, the memory is false. These depraved acts very much stunt children's development; it causes them to have learning disabilities; unexplained depression; loneliness; suicidal thoughts, and all kinds of side effects.

The cloning centre is a paedophiles paradise and **it must be stopped**! Elizabeth secretly owes a few music companies too (Universal and others) with bands under contract; she forces me under a knife to compose music for them; if I can't she will stab my REM driven clone; burn my REM driven clone and have her pervert thugs smash my REM driven clone body there. As Ehrsson (2013) demonstrated, because consciousness is linked; and consciousness transfer leads the person who has had their consciousness transferred to still perceive reality as "real as real" in the new body; I have the same biological and physiological responses as I would have if these attacks were occurring to my original body. This has caused me to have heart damage and severe debilitating headaches. Because my memories were released to me; I can now also remember all my past REM driven clone memories since age 5; as well as the previous night's REM driven cloned memories (when I have my consciousness transferred to the cloning centre) which causes severe stress to my mental psyche and mental well being.

As unbelievable as the above statement sounds regarding Elizabeth's ownership of music companies, just go and listen to the song "Abadon" performed by Boondox (RainmanJhof 2011) on YouTube. There are two "Abadon" versions; the version you seek can be found in the Reference section uploaded by RainmanJhof (2011); skip to 2:46 min (or listen to the whole song; pay particular attention at 2:46 min) and you will hear loud and clearly, Elizabeth say "Drop the mother f****ing base". That is Elizabeth's voice; you can compare it to any of her Christmas speeches. That is her. Elizabeth said the above as a REM driven clone at the cloning centre. I know. I was there. They got so confident, that nobody would ever find out about their REM driven cloning shenanigans, that they even had me sing on a song I produced: "Where'd You Go" –Fort Minor (Murdok Dubstep Remix) (MurdokDubstep 2010), available on YouTube. Tila Tequila and I sing on this remixed track as REM driven clones (MurdokDubstep 2010); I start to sing from 1:44 min onwards. You can compare the voice you hear in the song with the Vinny Eastwood interview (Vincent Eastwood 2013). That is me, in both cases, and that song was recorded at the cloning centre.

One good person at the cloning centre; Bernie Mac (also as an REM clone version of himself); who was an actor and comedian, tried to stand up for me and speak up for me saying ‘Cloning and torturing people here, it’s not right to do to a human being...’ and “This place is the devil”’... Bernie Mac also tried to have others join him in speaking out against the REM driven clone torture zone while he was having his consciousness transferred to his REM clone driven version. Elizabeth and Philip tortured him for 4 to 5 hours continuously at the cloning centre as an example of their ‘power’, Bernie Mac had an aneurysm (brain bleed) the next day and died. This is one of the aspects about REM driven clone torture: if you torture a REM driven clone continuously for hours on end, you will cause the person to die in their original body from a heart attack or aneurysm (while they sleep); even if they do manage to wake up the next day; because of the constant pain their consciousness has recorded; and because consciousness is linked they will wake up only to have a heart attack or aneurysm in their original bodies the following day. If you apply a constantly electrical current to a REM driven clone, that person will either have a heart attack or die from aneurysm in their original bodies while they sleep.

This form of murder is similar to the CIA’s heart attack gun (Non Mirage Truth Vision 2015); which was declassified in the 1970s and was a weapon used for causing heart attacks in victims, undetectable as the cause of heart attack under autopsy. The difference with REM driven clone death as a form of weaponry is that hardly anyone has heard of death by REM driven clone torture. Moreover, the perpetrator can be thousands of miles away from the victim; the culprit just needs a duplicate REM driven clone of the victim to torture or apply a constant electrical current to; this is exactly what is going on in our world every day. When Elizabeth and Philip don’t like someone and they want to get rid of that person, they clone that person; torture his or her REM driven clone till that person has a heart attack or aneurysm in their original body and dies.

This is what Elizabeth and Duke Phillip did to the indigenous people in Canada. The video reference (inifiniLor 2013) provided above, regarding the genocide of the indigenous people of Canada discusses this very topic (inifiniLor 2013). As Elizabeth and Duke Phillip visited Canada as clones, when they killed the native children; they both had solid alibis of being in England at the time. They then cloned all the witnesses of the Canadian missing children’s case and had them remotely killed by applying constant electric currents to their REM driven clones; one by one and each witness either died of a heart attack or aneurysm. They just dispatched the last witness by also applying a constant electric current to the persons REM driven clone because the missing children’s case was building momentum.

After Bernie Mac’s death everyone is afraid to speak up; they all became afraid of suffering the same fate. All it will take is a few lie detector tests from an independent source; and I will volunteer for them; they cover up for themselves so carefully that lie detector tests from an independent source unconnected to them in anyway will be the only viable method. I know many know non-famous people who have also attended these clone centre gatherings as their REM driven clone duplicate versions and also remember the clone gathering experiences in their original bodies (because they were not memory suppressed) that could be given lie detector tests to prove this. When it comes to the sexual exploitation of a child, lie detectors are admissible in Canadian courts. I need decent people with integrity to speak about this filthy business; to testify about who is involved. There is much to tell about the Illuminati and I am a wealth of information about this topic.

Elizabeth also had Princess Dianna Spencer killed by having someone shine the brightest light known to mankind through the window side on her car; they served away to avoid the light and hit the divider; it was no paparazzi. I first wrote about this in my original disclosure in **2011**; this was corroborated by an article 2 years **AFTER** I had made my original disclosure. See Radar Online (2013) for the corroborating article. Furthermore, "Candle in the wind" performed by Elton John (SadSongs4You 2010) was not Dianna's 'death song'; it was "Bigger than us" performed by "White Lies" (WhiteLiesVevo 2010). The Illuminati released this song, deliberately, 13 years after Dianna's death, again as a show of their 'power'. I did not want to compose this song but Elizabeth and her thugs threatened to stab and burn my REM driven clone repeatedly if I did not. Listen to the lyrics. I'll provide some of the lyrics here for those of you who are feeling too shocked by these revelations:

White Lies Lyrics "Bigger than us"

You took the tunnel route home.
You've never taken that way with me before.
Did you feel a need for change?
Apologies on your fingernails,
Love flickered in the city of lights
Like internet and radio waves.

I don't need your tears
I don't want your love
I've just got to get home

And I feel like I'm breaking up
But I wanted to stay.
Headlights on the hillside
Don't take me this way.
I don't want you to hold me
I want you to pray,
'cause it's bigger than us.

Are you beginning to see why these people must be stopped and stopped immediately? Are you beginning to see why it is no longer okay to 'turn a blind eye' or 'bury your head in the sand', and hope it all goes away? Can you see why I need the good people of this world to keep sharing this disclosure (online); to tell your friends and family until this information reaches the armed forces so that the armed forces can bring these people to their knees. I also do not want innocent civilians to get hurt despite the monstrosities these people cause to humanity on a daily basis. I want the overthrow to progress smoothly so that the good people of the earth can inherit the world, with all technologies and medical advancements which benefit mankind, made available to all. This is why I maintain my composure and resolve, in spite of the evil I see on a daily basis, and you the reader must also maintain your composure and resolve in spite of the evil which I am disclosing, so that we do not inherit a torn down world after we have completely overthrown these malevolent, sinister and corrupt hypocrites!

Furthermore, Diana survived the car crash; and while, either on her way to hospital or in hospital, Dianna was injected with a high concentration of salt, which is near undetectable in an autopsy and she died. Elizabeth did this because Diana was going to marry an Arab named Dodi Fayed. The Illuminati hate Arabs because for the most part, Arabs have not broken their relationship with God. Dianna hated her association with these people; Dianna knew about REM driven, sleep cloning and wanted nothing to do with it; but she was afraid of being killed and so said nothing. I know a lot about this and many more deplorable things the Illuminati have done over the course of 34 years and I will be more than happy to reveal all the evil they have done; as well as all the wonderful discoveries and progresses they have made in medicine, science and technology, which they have withheld from the general public. I am very happy to disclose all, because I known once the world learns of all they have done, we truly will begin to see the end of such evil on this planet. Just make sure you keep spreading this disclosure and it reaches the armed forces soon. Soon, soon, soon!

Help Donald Marshall by Sharing this Disclosure with Your Close Friends and Family

If you are reading this, then you have to help me. I know some of you may be reading this thinking, but I am just an ordinary man or woman Donald, I know you are telling the truth; I can feel it in my heart and my head that what you say is true and that you are telling the truth (which I promise you, with my right hand to God I am) but I'm just an ordinary man / woman Donald; I don't have any power; I don't have any influence; or because of the position I am currently in I can't speak out against these people; and I am just an ordinary man / woman I cannot be expected to take on the world's problems. You are right if you are thinking and feeling this, and I do not expect or want you to take on the world's problems. All I want you to do is share this information with close friends and family initially, because when you do that, at least that is one more person who will know about the true extent of evil in this world. When your close friend or family member also shares my disclosure with another close friend, the disclosure will eventually spread exponentially. It will also be one more person who has your well being in their thoughts because I understand how scary it is to know this. Therefore remember **there is safety in numbers;** and the more people you tell in your social circle, the more you can all look out for each other. So feel free to share my disclosure with close friends and family. Do not talk about my disclosure in public (for now) if it makes you feel uncomfortable. Moreover, if you want to reach more people (other than your initial social circle) the best way, is to do it online: through social media, such as Facebook, Twitter, and YouTube etc. As I've already mentioned, the Illuminati controls all forms of media except the internet.

The Illuminati, being so 'confident' at the time with the evil they wield onto the world, put my face on a Megadeth album cover titled *The World Needs a Hero* (2001) (to mock me) when I was 23. I posed for this album cover at the cloning centre when I was 23, and the album was released with a picture of my face, as a REM driven clone, 3 years later when I was 26. They did this because they were extremely confident that nothing would ever be proven. Google image it, although I am 39 now, you can still tell that it is me. The image shows what I end up looking like after an REM driven clone gathering with a skeleton popping out from my chest. I understand most of you are going to continue reading because although it is "liberating", it is also "paralysing" at the same time (Laurence Mountford) to realise what the truth of the world is, so I will display two images below, but promise yourself, that you will Google search this image in your spare time.

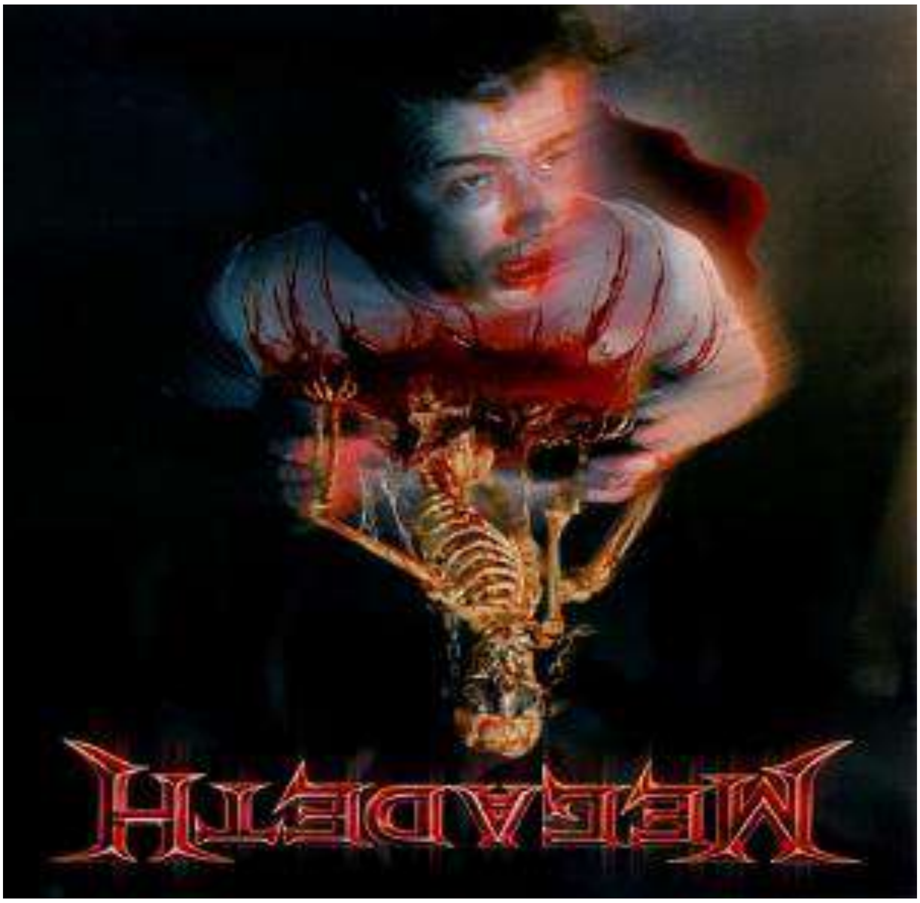

Image 3: Megadeth Album Cover: ***The World Needs a Hero*** **(2001)**

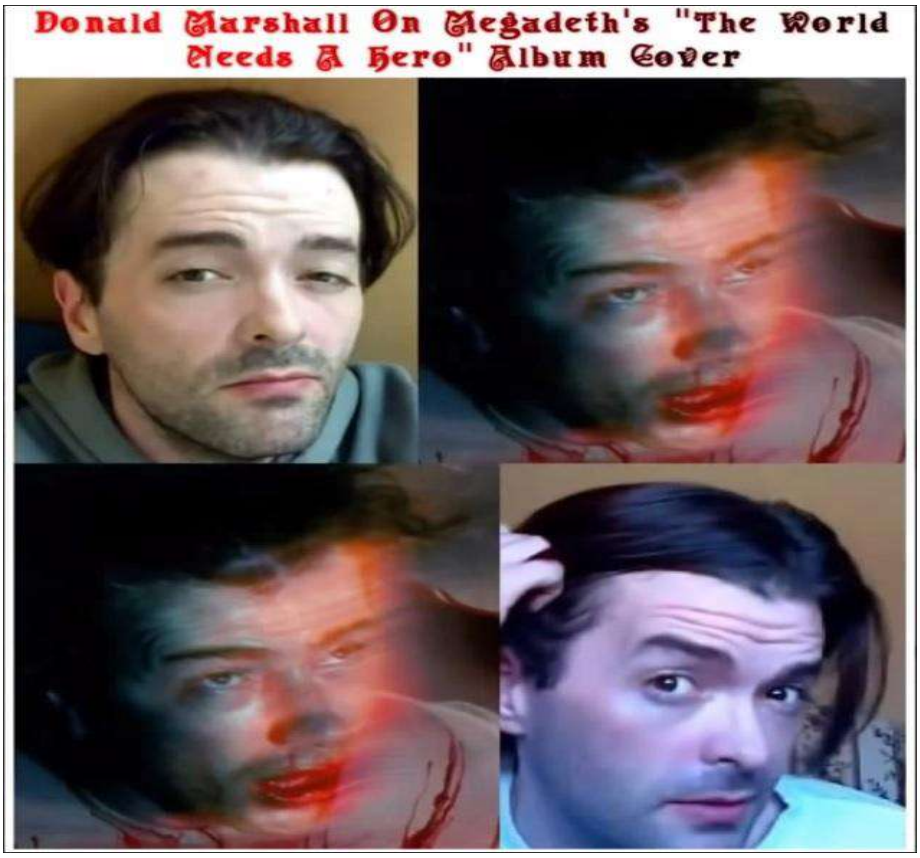
Donald Marshall On Megadeth's "The World Needs A Hero" Album Cover

Image 4: My original body: top left (age 37) and bottom right (age 37);
compared to my 23 year old REM driven duplicate clone body (top right and bottom left)
at the cloning centre.

You can clearly see from the images that, it is in fact me. In Image 4, you can also compare what my real body (top left and bottom right) looks like compared to my duplicate clone body. Duplicate clones do not come out of the tank looking 100% like the original; it's 99% - there are always some slight variations as some have commented that the eyebrows and nose in the picture, bottom right, is not the same (as the duplicate clone on the album cover). However, to people who make such comments, I say: my eyebrows and nose is the same when you compare it to the pictures of my original body (top left and bottom right). Also, remember that the image on the album cover is a duplicate REM driven clone of me. REM duplicate clones do not come out of the tank looking 100% like the original. There are always slight variations; but the wrinkles on my 37 year old forehead (when I took the picture in my original body; top left and bottom right) is similar to my 23 year old wrinkles, the lips are the same; and the face which you see, is definitely, the same! Yes. REM duplicate cloning is real; going to sleep and waking up as a cloned version of version of you is real and **we must put a stop to it now!**

The Deplorable Nature of the Illuminati

My own family has been turned against me one by one. The Illuminati have told my family members when they were REM driven clones at the cloning centre, things such as I secretly hit or molested their children; poisoned their pets when the pets died of sickness and other such things to make my own family hate me and not want to help me. Even when my family side with me there at the cloning centre, as REM driven clones of themselves; and my family members say 'They do not believe that Donny did all those horrid things' -which they have been told; Elizabeth then says "Are you cooling me a liyah?!" in her disgusting croaking voice made as low as low and evil sounding as she can make it, complete with psychotic and malevolent glare; and as terrified REM driven clones my family say "No, no, of course [what you are saying] it's true", stuttering and afraid as REM clones. My mother Catherine McMahon sold me into this REM driven sex and torture cloning slavery when she remarried a man named Gordon Cohoon; Gordon's whole family is in this scummy secret society, his brothers Tom and Tony Cohoon; his sisters Darlene and Bernadette Cohoon; they fear the new and improved lie detector tests; and all it would take, is for an independent source, far, far removed from any association with the Illuminati to administer these lie detector tests on me **AND** the above named people to prove this unequivocally.

Half way through the making of my original disclosure document (in **2011**), when I went to sleep that evening, the Illuminati transferred my consciousness to my REM driven clone and introduced me to a man (who was also a REM driven clone) named TROY LANDRY. He is an alligator trapper from Louisiana on the TV program named "Swamp People". Troy Landry said "If I send my [original] disclosure out to the public then he would take a power drill to my shin bone at the cloning centre and suck the marrow from my [REM driven clone's] bones." It is similar to spinal tap, and one of the WORST things you can do to a REM driven clone besides burning a REM driven clone. I told Troy that I will be sending out my disclosure to the public, because I have to escape the cloning centre, and so, Troy Landry did just that. It was excruciatingly painful; all the REM driven clones that were sat in the stands of the arena just watched; slack-jawed. Troy Landry is an insatiable child molester and an extra retarded REM driven clone. A side effect of REM driven cloning is that as a REM driven clone version of yourself, your emotions are heightened and you are less smart as you are when you are experiencing the world in your original body. Nevertheless, Troy Landry is extra retarded as a REM driven clone.

In Louisiana, when Troy sees a young boy that he likes, Troy asks of the boy's name and tells the boy "He is an alligator hunter"; Troy then proceeds to shake the boy's hand; Troy then has his paedophile friends at the local cloning centre find the boy's blood records to grow duplicate clones of the boy. Five months later, there are multiple REM driven duplicate clone bodies of the boy grown for Troy Landry to victimise before the crowd of onlookers. All members of the Illuminati do this; it's just a way of life to them; they consider themselves 'the privileged people' in the world power organisation. Almost all the cast on the "Swamp People" television program attended the cloning centre as REM driven duplicate clones of themselves, **when they go to sleep**. After putting a power drill to my shin bone, Troy Landry asked me again, "Will I send my [original] disclosure out?" I said "Yes" to which he replied "I'll do the same to your pelvic bone everyday!" The pain on my pelvic bone was far worse than when Troy put the power drill to my REM driven clone shin bone. I told Troy "I have no choice but to inform the public" – and so Troy put a power drill to my REM driven pelvic bone, everyday... and it is the worst pain ever... you forget your own name; where you are; all you know is pain; and you beg God or anything to save you from it...

Another deplorable thing which the Illuminati did in real life is: The Canadian government were trying to lower the amount of prostitutes on the streets (Elizabeth hates prostitutes) so they had a man named Robert Pickton start killing the women and feeding the women to pigs on his farm. The Illuminati had a camera set up in the upper corner of a room in Pickton's house and recorded Pickton hitting these women over the head with a hammer (a ball-peen hammer). They took the recordings, all 49 of these murders, and they all watch them at the cloning centre. To be seen as 'tough' at the cloning centre is to watch all 49 of the Pickton murders without having it affect your psyche. Elizabeth loves watching the Pickton murders and she says she has a macabre fascination with death. Canadian Prime minster Steven Harper knows all about the conduct of the Pickton murders; he has seen all the recordings and cheers on as the rest of them do. If Mr. Pickton says anything about his involvement with the Illuminati, or the recordings, they will apply a constant electrical current to his REM driven clone until Pickton has a heart attack or aneurysm in his original body.

What Happened to Britney Spears?

The Pickton murders are also the reason Britney Spears 'lost it' and shaved her head bald. She had watched too many Pickton murders as a REM driven clone at the cloning centre and it drove her 'over the edge'. I'll also reference the following songs performed by Britney Spears so that the reader has more corroborating evidence to add to the ever increasing pile. Search for the video "Break the Ice" (BritneySpearsVevo 2009) on YouTube. In the Japanimation video; at the beginning of the video; that is a cloned **cartoon** version of Britney Spears being grown in a big thick thank full of water (1 second to 7 seconds); and at 1:32 min to 1:43 min Britney walks into the cloning centre. This is **EXACTLY** how the cloning tubes are in real life. They are stacked one on top of the other, each filled with salty water. Britney Spears made this video because she said a fantasy of hers is to "Blow up a cloning centre". The Illuminati allowed her to release this video because at the time they figured nobody would ever figure it out. I have provided the images on the next page.

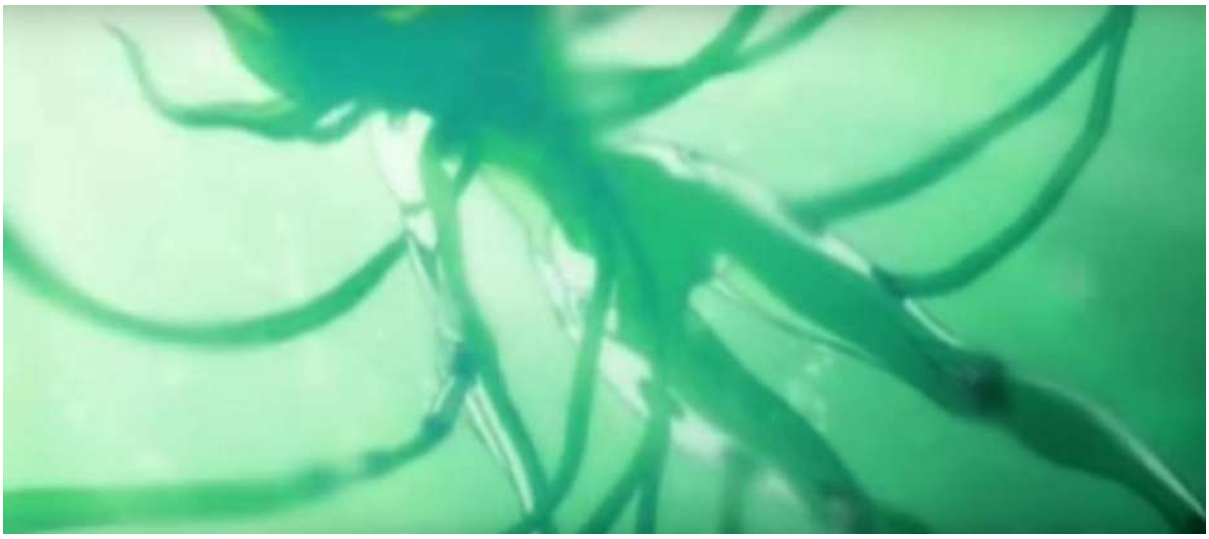

Image 5: Britney Spears is being grown as a duplicate clone in a cloning tube. The depiction is EXACTLY how the duplicate clones are grown in the Cloning Centre... with wires extending from the body; in a big thick tank full of salty water.

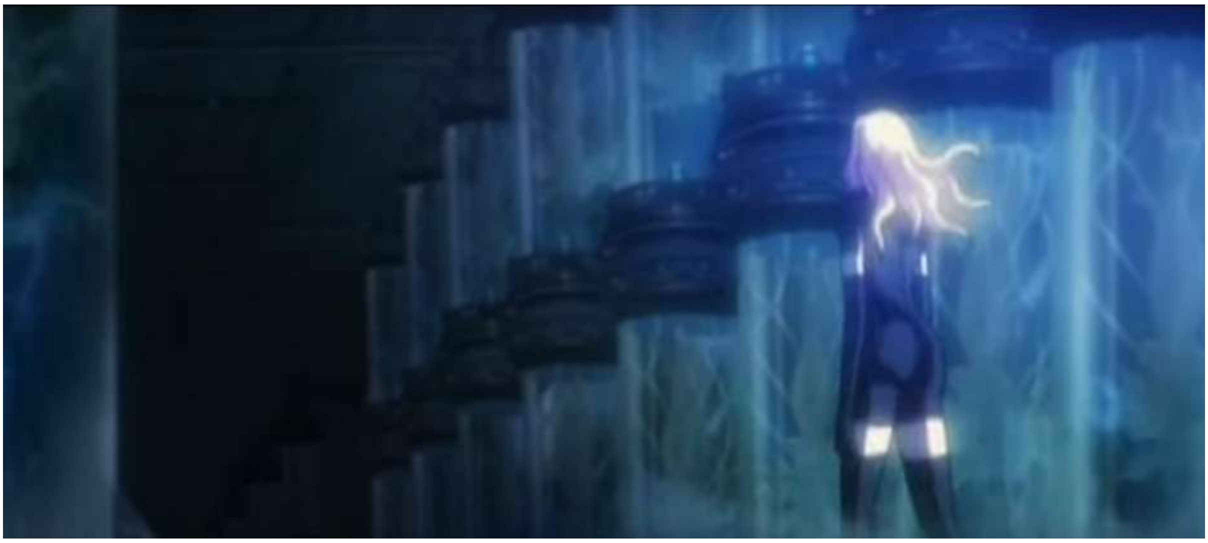

Image 6: Britney Spears walks into the cloning centre. The cloning tubes are stacked one on top of another. This is EXACTLY how these cloning tubes are in real life at the Cloning Centre.

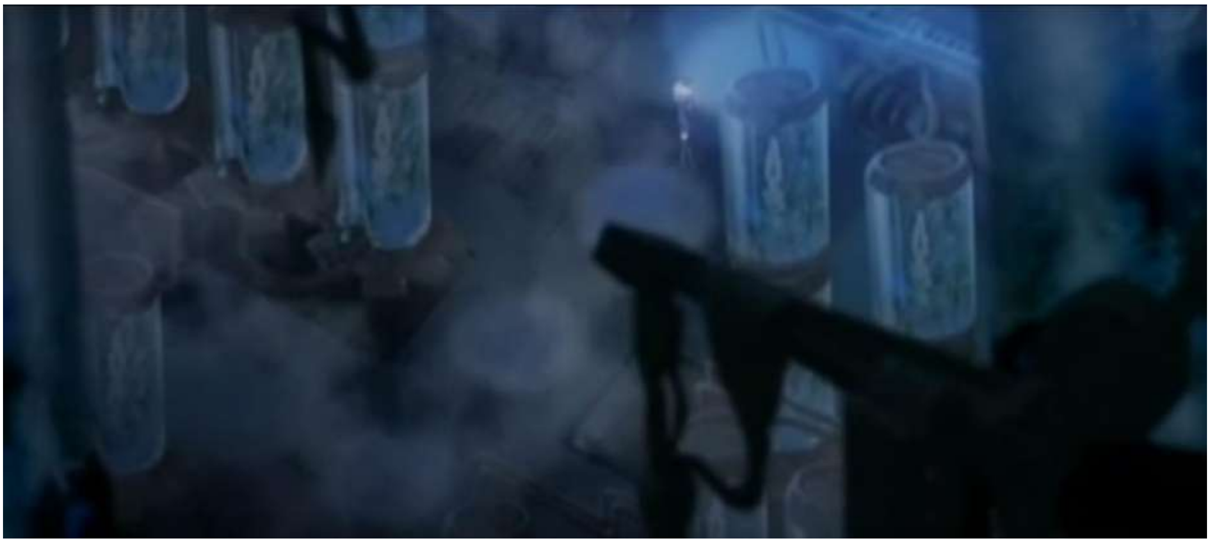

Image 7: The Cloning Centre depicted from an aerial view.

In “Hold it Against Me” (BritneySpearsVevo 2011) performed by Britney Spears; Britney is depicted fighting against her clone at 2:46 min till the end of the video. In “Mona Lisa” (SimpleGirl4ewer 2007) later remade by Britney Spears she changed the words in the song from “gone” to “cloned”. I’ll provide the lyrics to chorus here because they have not changed it on A-Z lyrics etc., but listen to the song and you will clearly hear the following:

Britney Spears Lyrics “Mona Lisa”

She’s the original (yeah, yeah)
She’s unforgettable (yeah yeah)
She wants you to know (yeah)
She’s been cloned…
It’s kind of incredible (yeah yeah)
She’s so unpredictable (yeah yeah)
She wants you to know (yeah)
She’s been cloned,
She’s been cloned,
She’s been cloned…

Yes. Britney Spears and many other celebrities want you to know they have been cloned. The celebrities are depending on you, the ordinary man / woman; the populace, as I am depending on you to spread this message fast and far. In fact Elizabeth has said that “If any public person speaks out, they will die!” Therefore, public figures cannot openly discuss this topic to warn the populace –so they hint. It is **NOW** up to the ordinary man and woman to end this. Furthermore, Elizabeth has a nuclear bomb hooked up at the cloning centre which she can press at any time to wreck havoc on the world unsuspectingly. Thankfully, I managed to convince Elizabeth not to do such a silly thing. I hope you can now understand why everyone is terrified of Elizabeth at the cloning centre, and why this is a **world emergency** and we are depending on you to share and spread my disclosure!

NEVER EVER SELL YOUR SOUL

In This Physical World Selling Your Soul Is Selling Your REM Driven Clone

Furthermore I must discuss an entrapment many up and coming ‘stars’ fall into. You have most likely heard many public figures say “I sold my soul to the devil” or some variation of “I sold my soul”. I am now going to address the aspect of “selling souls” or “selling your soul” which I am sure you have heard many of these celebrities talk about. I’ll address this as coherently as I can because the on the surface selling your soul talk is difficult to contemplate especially if you consider the soul to be a metaphysical, ethereal or mystical part separate from your earth body. Moreover, we live in a physical world, so selling a metaphysical aspect of you sounds incomprehensible (at first). For the most part we have little understanding of what the soul is. However, we are told that the soul is a part of us which never dies and it is eternal. It is all that we are, ever were, and ever will be. Now our faiths and beliefs come into question, because if you are an atheist: there is no God, selling your soul is a bunch of crazy nonsense.

From a practical perspective it also seems difficult to fathom, because if you go on what you have been told, you are selling an intangible part of yourself to be used by someone else for whatever they wish; and so if you get into the headspace of viewing your soul as something separate from you or that you are in control of your mind and your body, then again, 'selling an intangible aspect of you' again sounds like nonsense because right now nobody controls you; you are controlling you; so what are you really selling? Thin air?

Many up and coming 'stars' think this way when they are confronted with "selling their soul" aspect. The up and coming 'stars' are blinded by the prospect of being super famous and super rich; and in the moment they are confronted with "selling their soul" they believe these Illuminati people, are a group of over religious zealots and that they, -the up and coming 'stars' are taking advantage of the Illuminati because the 'soon to be 'stars'' do not believe (in that moment) it is possible to sell your soul, and only consider being super rich and super famous; so the 'stars' sign their 'soul' over.

At this point, it is usually too late for the up and coming 'star'; because what they have actually sold, in this physical world, is the use of their REM driven clone to the Illuminati. They have sold their REM driven clone to be used for REM driven sex by perverse old men and women whenever they enter REM sleep.

These celebrities are not memory suppressed either; so whenever they enter REM sleep, their consciousness is transferred to their REM driven clone, to be used for whatever the Illuminati wish: usually sex as REM driven clones with dirty and perverse old men in return for a promotion in the entertainment industry; such as movie role, advertisements, concert gigs etc. It is a very, very, sick business, and once these up and coming 'stars' get into the business they are trapped. Now you can begin to understand why so many celebrities, although on the surface it appears they have it all, are depressed; and do random and crazy things. It is a cry for help. Their illogical behaviours, (not all) for the most part, can be attributed to REM driven cloning. Furthermore, the Illuminati consider 'selling your soul' (the use of your REM clone) **a very serious business.** They do not joke about this; and there are NO RETURNS. Once you have sold the use of your REM driven clone for promotions etc., that's it. If you ever make a fuss and want your 'soul' back, in other words: you no longer wish to have your REM driven clone to be used by dirty old men for sex; the Illuminati will kill you by applying a constant electric current to your REM driven clone until you have a heart attack or aneurysm in your original body and die. It is as simple as that.

For example, Whitney Houston wanted nothing to do with the cloning centres, she said she would keep quiet; she never wants to be in the spotlight; she never wants to do radio or television. Let bygones be bygones, I'll go my way and you can go your way. The Illuminati were extremely angry! They had bank-rolled her, and now she didn't want to be their 'friend' or ever associate with them again. She was sacrificed; she was killed with a constant electric current to her REM clone. She had sold her soul, in the beginning of her career; in other words, the "use of her REM clone". Earlier in Whitney Houston's career, she had signed that they can use her REM driven clone however they wish.

I have also seen people say "No" –when confronted with 'selling their souls'. These people are then memory suppressed, and will not remember the proposal of their REM driven experience. The people who also refuse to sell the use of their REM driven clone ('sell their souls') are blacklisted secretly, and they will never receive a high paying movie role, concert gig, advertisements etc. If you have ever wondered why so many talented people have not 'made it' –they refused to sell their souls, in other words, the use of their REM driven clones in exchange for promotions. This is how the entertainment industry really works. Another aspect of selling your soul, in other words, **selling your REM driven clone** is that you are also agreeing to sell your children, and your children's children etc.; your whole lineage to the Illuminati; for the Illuminati to use your unborn children for paedophilia, torture, sex sports and other grotesque and ungodly things when they sleep. It is a nasty business. **NEVER EVER** sell your soul.

The lesson is very important. NEVER EVER, 'sell your soul'; NEVER, EVER! You will experience the worst nightmare situation ever, and in this physical world, that is the use of your REM driven clone to be used for sex sports. You'll never have a decent night's sleep ever again, and you will always retain the full memory of your REM driven clone experiences; as well as, you will be selling your unborn children into this dirty business. Even when confronted with this as a joke: never sell your soul. Never Ever! The physical aspect you are confronted with when asked to 'sell your soul' is: "YOU", and everything which makes you, "you"; your ideas, your creativity, your privacy, your biology; everything which makes you, you. Never sell your soul, which in this life is currently the use of your REM driven clone! Never, ever! I hope that is well received.

Fake Alien Abductions conducted through REM Driven Cloning Technology

I must add they grow duplicate REM driven clones of people from all walks of life and then they transfer the persons consciousness to their REM driven clone version at the cloning centre while the person's REM driven duplicate clone is chained down to stainless steel corpse tables; or they pre-inject the REM driven duplicate clones with drugs before transferring the consciousness so that once the consciousness has been transferred, the REM driven clone cannot move. This is the point when they usually send people into the room dressed as aliens. It is a fake alien abduction. It's just these Illuminati people dressed in Hollywood quality accessories and they make the REM driven duplicate clones which have either been chained to the steel corpse tables or pre-injected with drugs, believe they are having an 'alien abduction'. Members of the Illuminati even dye chicken skin grey, and stretch it over a mask for realism and perform "experiments" on the immobile REM driven clones; the perverts anally probe them; they rape their limb bodies and do not suppress the victims memories; so these victims wake up thinking they have been abducted by aliens but in reality, they had their consciousness transferred to their duplicate clone versions when they entered REM sleep and they were violated. The victims do not know where to turn and naturally, they are embarrassed. Some victims even try and do install cameras all around their bedrooms in order to videotape themselves while they sleep; to prove they have been 'taken'; but to no avail they have not been taken; they have been cloned and had their consciousness transferred to their REM driven duplicate copies once they enter REM sleep by the most disgusting perverts in the world. The scum even videotape these 'abductions' to watch later as sick demented porn. These fake abductions are conducted by the Illuminati more often than you can imagine. They just keep doing it to random people over and over and over...

I am THE Spy for Humanity against the Corruption of the Illuminati

Over the years the Illuminati have offered me: a priceless katana (samurai sword); a rigged jackpot winning lottery ticket; REM driven clone slaves to keep me quiet, any of the prettiest women I've ever seen; any girl from high school; Vladimir Putin's daughter; even children; I do the right thing and spit in their REM driven clone faces; jam my fingers into their eyes but they have the technology to turn off the pain sensors of a REM driven clone, and therefore they have their pain receptors turned off as REM clones, while my pain receptors are fully on (sometimes even beyond the threshold of what is normal pain for a human). See the online articles: Science Daily (2014) and RT (2014) which discuss how pain receptors can be turned on and off in humans. Not much will hurt their REM driven clones other than getting dirt or vomit in a wound or bleeding too much. However, they have perfected the science and technology of REM driven cloning and therefore it only takes time and approximately $30 to grow new duplicate clone bodies. I started to smear excrement in their faces as a REM driven clone at the cloning centre in an attempt to deter them; I have **never** handled faeces in my original body, but I do this as a REM driven clone at the cloning centre, and even that does not deter them from their depraved ways. I composed a song which was later performed by "One Republic" referring to it called "All the Right Moves" ("All the right friends in all the right places... All the right moves in all the right faces"). I've made so many songs with MK Ultra technology (Rense 2001; Jim Cristea 2009; Berekley News 2011; Mind-Computer 2012 nature video 2013) and Mind-voice technology (New Scientist 2014; Prigg 2014) it is ridiculous. You see it is very easy to make music using Mind-voice technology; Flo-Rida used Mind-voice technology to produce the song "Whistle" when I first made my original disclosure (in 2011). The song is a taunt at me about my "whistle blowing disclosure" i.e. 'Can Donald blow the whistle? Etc. etc.' The thing is, anyone with a talent for music can make music much easier with the technologies which they have (MK ultra technology; Mind-voice technology); however, the Illuminati made more money from the music I composed, and so over the years, they just kept using me, over and over and over again, for rock, pop, rap, country etc. There are so many people involved in this REM driven cloning business, it is staggering; the organisation is vast. The only things the Illuminati fear are nuclear war; the new and improved lie detector tests and the general populace spreading my DISCLOSURE DOCUMENTS.

For those of you thinking Queen Elizabeth II looks like a kindly little old lady and you still cannot believe that all this is true, although I have provided plenty of evidence to the contrary as well as evidence of her saying "Drop the mother f***ing base" as a REM driven clone on a song track (RainmanJhof 2011); you could not be more wrong. She is the worst human being I have ever seen or heard of. It is so sad to see these REM driven women and children having their consciousness transferred to their REM driven clone duplicates and brought to the centre of the dirt rink, sitting there naked, afraid and crying and raped and beaten for sport, for the rich and famous. REM driven cloned women and children display the most range of emotions, and it is the saddest thing to see there. Therefore, of course I cannot bring myself to be a part of this despite any riches which the Illuminati offer me.

Over the 31 years when I was having my consciousness transferred to my REM driven clone at the cloning centre; Elizabeth, Vladimir and many, many Illuminati members and cloning centre attendees discussed many topics with me. These Illuminati people, such as Elizabeth and Vladimir, as well as many others confided in me because I had composed songs since I was 5 years old.

For readers struggling with the idea that a 5 year old can compose such wonderful songs at such a young age I will remind you that, it is easier to compose songs with technologies such as MK Ultra and Mind-voice technology. One is only limited by their imagination and creativity. Moreover, research talented children. I won't provide references of talented children here because I want to protect children from the Illuminati, and not have children exploited like I was; and for the readers who are parents; I am sure you know how talented, creative, and imaginative your children are.

Accordingly, since the age of 5, Elizabeth, Vladimir and many, others confided me in me because they thought I was special, and wanted me to hang with them; some of these Illuminati people, I had composed their favourite songs; I was also memory suppressed for many years so they thought it was harmless to tell me everything that exists or is possible. They were trying to impress me and amaze me, and therefore they showed me plenty of their technologies and what they had. They wanted me to me a willing, full Illuminati member; so over the years I feigned interest and pretended to like these Illuminati people, so that I could learn what they do and how they do it. I was looking for a way to block the consciousness transfer to a clone during REM sleep, or hoping they would reveal this part to me; but they did not, and I did not find a method to block the consciousness transfer. However, I bided my time, until I knew just about everything they do; so that I could reveal to the world everything they do; to give humanity a fighting chance against everything they do and as a consequence, I became a spy for humanity against the corruption of the Illuminati.

The Threat to Humanity when the H.A.A.R.P Grid is Complete

One of the most important topics which they discussed with me that I must add to this disclosure is HAARP technology and their plans for HAARP technology. I have discussed the basics of HAARP above. HAARP radio transmitters have to be placed at certain points across earth to complete a 'HAARP grid'. Once all the installations are completed over the earth HAARP will have more functions as a grid than the basics described. The Illuminati have told me, there is a possibility the HAARP grid will be capable of time travel, although there is a 50/50 chance using HAARP for time travel could destroy the earth. They also told me that once the HAARP installations are completed, the HAARP grid will be able to bend space and make the distance between, for example, Earth and Pluto approximately 1000 miles for a few seconds, and then the distance will go back to normal.

The Illuminati also said, once the HAARP grid is completed, they would be able to mind control people like never before without the use of RFID microchips or MK Ultra. Life on earth will be much worse once all the HAARP installations are up and working together in unison. The grid has to be complete though. HAARP cannot achieve time travel manipulation and mind control if it is just located in one or a few locations around the earth. There must be a certain number of HAARP installations around the world, in certain locations around the world; and the HAARP radio transmitters must be working in unison in order to achieve time travel manipulation and mind control of the populace of earth.

This is what they told me as REM driven clones at the cloning centre. This disclosure could be false, or a half truth. I do not know. However, it is very important that I add this to the disclosure because the Illuminati are working twice as fast to complete the HAARP grid to achieve their aim of mind control over the entire population since I the time I sent out my first disclosure (in **2011**). Whether or not they will achieve their aim of total mind control once all the HAARP installations are completed is NOT something we can afford to debate. They also told me that once the HAARP installations are completed; they will have the capability to mind control anyone remotely like never before, at the push of a button. HAARP will also be capable of setting back time; which the Illuminati will use if they are ever in danger of being exposed. However, the Illuminati will remember the previous time, and the general populace will NOT remember the previous time, so that the Illuminati can correct the mistake(s) which led to their downfall. **Once the HAARP grid is completed, humanity will be slaves forever.** This is why we cannot afford to debate. Furthermore, from the book “Angels don’t play this HAARP” (Begich & Manning 1997, p. 8), it is estimated that the HAARP grid will be completed by approximately 2017; others estimate by the year 2020. The important thing to note is that humanity does not have long, until the HAARP grid is completed.

I must also add the positive side of HAARP, because as I have said, there is nothing inherently evil about technology. Technology is just a tool; it depends on the person using it. HAARP used correctly, and in the hands of honest men and women, will tame the weather with no adverse effects. Therefore humanity is on a timeline and a deadline to bring the Illuminati to an end. **HAARP is one of the main reasons, this disclosure must spread, and spread fast.**

See below for possible HAARP locations around the earth (Rense 2011). The HAARP sites are located globally where the richest mineral belts are located. Sub-surface mineral exploration has been done from satellites by radio tomography that is 100% accurate.

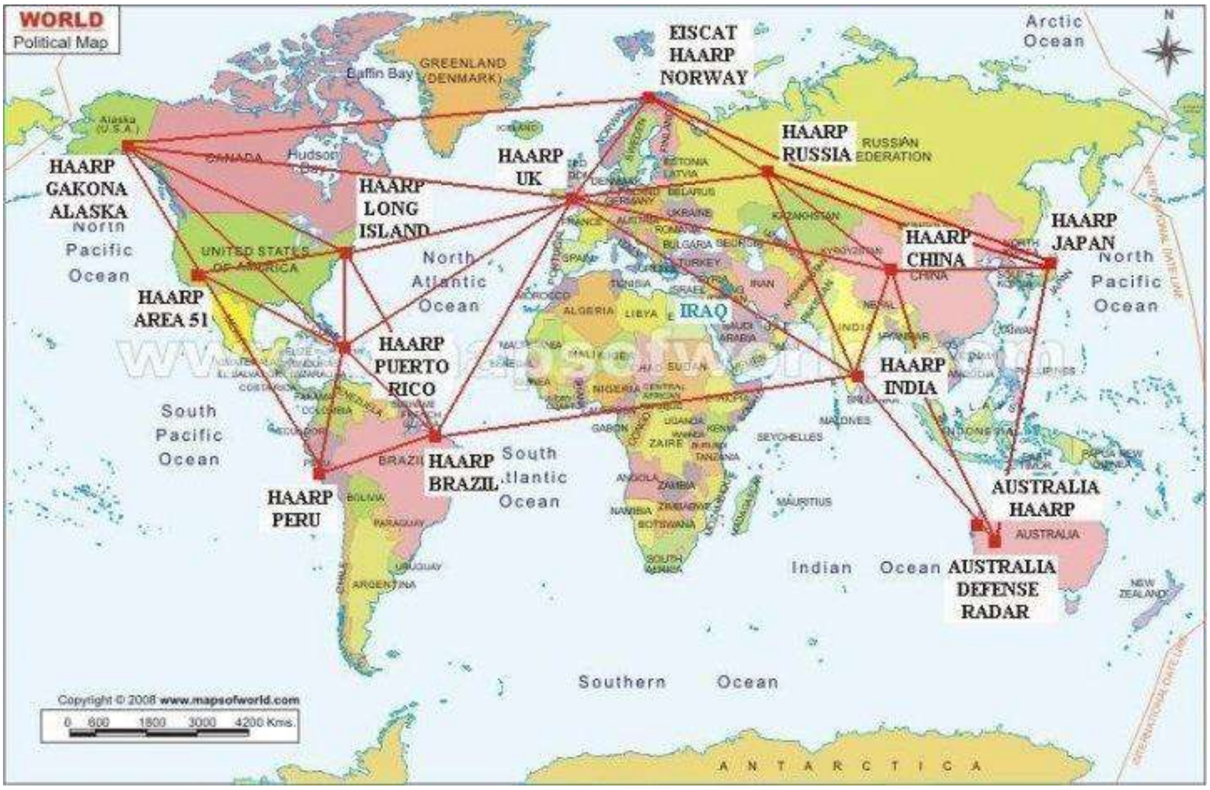
WORLD
Political Map
EISCAT HAARP NORWAY
HAARP RUSSIA
HAARP GAKONA ALASKA
HAARP UK
HAARP LONG ISLAND
HAARP CHINA
HAARP JAPAN
HAARP AREA 51
HAARP PUERTO RICO
IRAQ
HAARP INDIA
HAARP BRAZIL
HAARP PERU
AUSTRALIA HAARP
AUSTRALIA DEFENSE RADAR
Arctic Ocean
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean
ANTARCTICA
Copyright © 2008 www.mapsofworld.com

Image 8: Possible H.A.A.R.P. Locations. Source: Rense (2011)

REM Driven Cloning is the Most Terrible Nightmare Situation Ever

Furthermore, in terms of REM driven cloning; Fefe Dobson (singer) and Kurt Russell (actor) told me during my original disclosure (**2011**) specifically, to include in my disclosure that they do not like the cloning centre and they didn't torture me. Then they said not to mention them, but I will mention them again. Mila Kunis (actress) from "That 70s Show" stabbed in my REM driven clone body multiple times when I made my original disclosure and I was rendered immobile for saying that "She is a slimy scumbag for hanging with these people" and for also saying "She has enormous eyeballs and looks like a lemur". Mila Kunis (as a REM driven clone at the cloning centre) then begged me not to mention her (in my original disclosure, but I'll mention her again too) and then Mila Kunis said "She doesn't want to go to the clone zone anyway" and the Illuminati then deactivated her consciousness from her REM driven clone body, and she left. Her REM driven clone body went limp at the point where her consciousness was deactivated from it; her REM driven clone body looked 'dead', and she was gone. Some of the privileged people at the cloning centre are allowed **NOT** to have their consciousness transferred to their REM driven, clone duplicate versions (or they attend the cloning centre willingly when they want), but not me. I am an imprisoned slave in the worst nightmare situation. I told Nicole Leone (Madonna) also a REM driven clone at the cloning centre, that I was going to tell everyone that she coerced me to compose songs for her throughout my life (as REM driven clones). Nicole Leone (Madonna) told me VERY specifically to say in my disclosure here "She is not afraid!" and "No one will believe me and nothing will come of this". I beg you to help me prove her wrong.

Moreover, having your consciousness transferred to a REM driven clone alternative inhibits the person's ability to dream for that evening. Although your original body is immobile and asleep in your bed, when they transfer your consciousness you have actually 'woken up' as a REM driven clone version of you at a physical location here on earth; therefore you walk and talk as a clone version of you. You are not dreaming; you are just a REM driven clone for the evening. Furthermore, when the Illuminati restore your memories (as they have with me) or do not suppress the memory of your REM driven cloning experience for the evening (such as with the fake alien abductions), you will remember the experience as 'clear as daylight on a summer's day'; the experience is 'clear as a bell' as life is in your original body.

It is unlike dreaming, where dreams are fuzzy, incoherent and with random conjecture. The Illuminati know this as a fact, and therefore pre-inject the REM driven clones of random unsuspecting civilians for the evening with drugs before they transfer the person's consciousness. They tell the person who has had their consciousness transferred to their REM driven clone all sorts of nonsense: such as they are in the 5th Dimension; the astral plane; a singularity; the spiritual realm; Valhalla; quantum hopping etc. and all sorts of fallacies so that the person will not have a coherent picture of what has happened or what is happening. The truth is that person has had their consciousness transferred to their REM driven duplicate clone when they went to sleep and entered REM sleep; their original body still remains asleep in their bed; their consciousness however is now transferred to a clone duplicate alternative, and the person is 'awake', walking and talking as a cloned version of themselves in a physical location. The Illuminati allow some people to retain some of these REM driven clone experiences to affect their mental psyche.

The physical location where the populace will be able to verify these REM driven clones is somewhere within 5/6 hours drive, in a radius of the Robert Pickton farm, Port Coquitlam, British Columbia, in Canada, somewhere at a remote nature reserve. This is the cloning centre all the rich and famous attend because it is above ground and the rich and famous don't like the 'hospital smell' of underground cloning centres. The above ground cloning centre located in Canada is the cloning centre I have my consciousness transferred to when I sleep. It is guarded by military personnel. There are many cloning centres all over the world; as Phil Schneider (1995; 1996) disclosed as far back as the mid 1990s when he spoke: there were 1477 Deep Underground Military Bases (DUMBs) worldwide and 131 active DUMBs in the United States alone. Each DUMB costs on average 17 -19 billion U.S. dollars; paid for by the taxpayer; and it takes approximately a year-and-a-half to 2 years to build DUMBs with sophisticated methods. DUMBs contain cloning floors, and have an entire floor dedicated to human cloning, where duplicate clones are grown of unsuspecting civilians (remember they just need your blood / DNA); and unsuspecting civilians have their consciousness transferred to their REM driven clones at a physical location, which is usually these Deep Underground Military Bases when they sleep; unsuspecting civilians are memory suppressed, also to be used for gruesome torture and sex sports; as well as the plagiarism of talented individuals. Schneider gave the above statistics in the mid 1990s; chances are there are more DUMBS and cloning centres worldwide today.

Are you beginning to understand why this is the worst nightmare situation ever? Can you understand how: you, your friends, your family, your brothers, your sisters, aunties, uncles, children, work colleagues and anyone else you have contact with in your network, may be involved in this REM driven cloning business and will not even be fully aware of the situation or remember any experiences which they may have had once they wake up from sleep and they are back in their original bodies? I remember all of the REM driven clone experiences as clear as daylight because my memories are fully restored. Can you understand why the Illuminati must be stopped and stopped immediately? This REM driven clone business is vast, and it affects everyone, not just me.

This is why I mention that I am fortunate or unfortunate to remember all these experiences clearly depending on how you view it. I am unfortunate in the sense that I can recall all the depraved things these people do as REM driven cloned duplicates as clear as daylight and it affects my mental psyche. However, the fortunate part is: because I can recall all these experiences clearly, I can inform the world about the evil committed against humanity in a clear and logical way for the average reader to understand and verify, so that the populace can bring these people to justice for the crimes they have perpetrated against humanity and continue to inflict on humanity. For anyone who has had their intellectual property stolen by the Illuminati as a consequence of the technologies they have used on you unsuspectingly, you will be able to sue and regain monetary reward for the theft of intellectual property as well as invasion of your privacy.

I do not want to get too far ahead, because at this present moment, it is **very important** (I cannot say it enough) that this disclosure spreads (this is the stage we are currently at); it reaches the armed forces; to enable the armed forces to realise the crimes committed against humanity because of these people (the armed forces are kept 'out of the loop'; they just 'follow orders'); the armed forces will have to perform a military coup and bring this people to their knees. This is the point where the ordinary man / woman can have his REM driven cloning experiences restored, (that is if you want to) and sue / rightfully reclaim damages and legal reward for any crimes committed against him or her.

The Meek Shall Inherit the Earth

As you can see if everything goes smoothly; the meek shall inherit the earth. This is why you must maintain your composure and resolve, as I do every single day. You must not riot or cause chaos, damage property or undertake any act which will cause public dissension. The good people of this earth are about to inherit the earth; under no circumstance must you destroy anything. At the end of this currently dire and complicated situation; humanity will inherit the cure for HIV/AIDS, cancer, Alzheimer's, dementia, you name it –many debilitating diseases known to man; and other such wonderful technologies like hypersonic trains which run on magnets and are capable of doing Mach 5 (Mach 5 is the equivalent of 3806 miles per hour or 6125 kilometres per hour; Schneider (1995; 1996) also mentions the availability of Mach 5 trains in his disclosure, back in the 1990s). These trains can be used to deliver foods which are discarded in More Economically Developed Countries (MEDCs) to Less Economically Developed Countries (LEDCs) because these trains can travel thousands of miles within minutes. Yes. The Illuminati have all these wonderful technologies and more, but withhold it, and in turn withhold the progress of humanity. Yes. Plenty of man's problems can be solved when we address the root causes of our problems, and one of the root causes of humanity's problems is: REM driven human cloning which is currently in the hands of evil men. Even REM driven cloning can be used for the benefit of mankind when this tool is in the hands of righteous men. It can be used learn new skills in your sleep: such as gain extra qualifications; learn how to speak a new language; play instruments etc. –so long as memories are not suppressed, you will be able to perform the skills you learned as a REM clone in your original body when you wake up. As you see there is nothing wrong with most technologies; it all depends on the person / people using it. Nevertheless, although they may be beneficial uses to REM driven cloning not yet realised, we must be shut down REM cloning before we can even consider how REM driven cloning may benefit mankind, because at this present moment all that REM driven human cloning is used for is to commit evil, particularly against unsuspecting civilians.

It is OK to feel angry about all these crimes committed against humanity daily. But do not let these evil crimes rule you, in fact, let it motivate you; let it motivate you to the point where you will stand up and speak out to ensure that these crimes end. I think about the children and the future of the world every day; that is what motivates me and I view my situation as a God-given mission to bring these people to justice. I view things this way to maintain my sanity and composure. This is what keeps me going despite the extreme evil I experience, and it helps me to maintain my resolve and composure. The fact that I do not want a single child to grow up in a world, where they are being messed with in their sleep is what keeps me going despite this extreme evil. Some children are even kidnapped and brought to the cloning centre in their original bodies, again to be messed with. This is the root cause of why so many children go missing across the world. I'm sure any rational thinking human being feels the same way I do, and does not want their children, or any child to inherit a sick world governed by sick individuals.

I am a baptised Roman Catholic and a God fearing man. I believe there is a Creator of the universe but the Illuminati tell me there is no God, only science and technology. The Illuminati are also responsible for the death of the young beauty pageant girl Jean-Benet Ramsay; her parents are in this REM driven cloning business. Casey Anthony's daughter; as well as, Casey Anthony (before she died) AND Casey Anthony's parents also attend the cloning centre as REM driven clones. Many others; a man had his wife killed and this was mentioned on the news; it was an Asian woman whose rich husband had killed her; the husband said the wife was responsible for their son's death. Elizabeth's response was that "Rich people in this organisation (The Illuminati) do not go to jail, they are covered for."

Stephen Spielberg and George Lucas are Illuminati members and they willing attend the cloning centre as REM driven clones when they sleep. George Lucas made other Illuminati members light my REM driven clone body on fire for the end of his movie *Star Wars Episode III: Revenge of the Sith* (2005).George Lucas said (as a REM driven clone) he wanted Hayden Christensen "To scream realistically". Hayden Christensen (also a REM driven clone at the cloning centre) watched me, and listened to my screams and groans of pain and copied the sounds coming from my burning REM driven clone exactly. Hayden Christensen knows he is a REM driven clone when he goes to sleep and Hayden Christensen knows his consciousness is transferred when he reaches REM sleep. Hayden Christensen knows all about REM driven human cloning, and he is a "privileged" Illuminati member. Natalie Portman also knows all there is to know about REM driven human cloning; as does approximately a quarter of the "Star Wars" cast. Many directors have attended the cloning centre as REM driven clones too, and have used me in a similar role playing situation; they have caused me to have excruciating pain as a REM driven clone, to see what kind of squeak or screech I will make, as if I were less than a dog; and when I am a bloody mess on the floor as a REM driven clone; that is when they usually crawl on my broken REM driven clone body and sodomize me saying something like "They love me and they cannot control themselves because I made all their favourite songs and I'm so "special"".

My ribs in my REM driven clone body will be broken, I will have suffered internal bleeding and will be crying or screaming, if I am able, but they just continue to sodomize me, and videotape such a depraved act of wickedness so that they can view it again like "evil pornography".

The REM driven cloned children, who have their consciousness transferred while they sleep at home in bed, in their original bodies, need a familiar face to talk to as REM clones or all they do is scream and cry at the cloning centre; that is where Joy Geizer comes in. Joy Geizer is married to my half brother from my father's first marriage; Joy Geizer is a REM driven clone girl guide leader, and when the Illuminati clone young girls, Joy Geizer speaks to these REM driven cloned girls; Joy Geizer keeps the REM driven cloned girls calm and "pimps" them out for free, knowing that these young girls will have their memories suppressed, and therefore they will not remember the experience and will not talk about it when they wake up. All the Geizers are in this REM driven, human cloning business; there are many people who remember these REM driven cloning experiences in their 'awake state' and in their original bodies who could be polygraph tested by independent polygraph testers. The police, and polygraph testers in my City of Halifax, Nova Scotia are compromised (they attend cloning centres as REM driven clones too); they cover up for child molesters and therefore cannot be trusted to fairly administer a polygraph test; as I said, commissionaires and C.S.I.S are heavily involved. All it would take is a few polygraph tests to prove these things unequivocally, because for those of you who have 'eyes', you can already see that REM driven cloning is real, and that this is actually the way of the world. I will take these polygraph tests publicly too, and demand that my mother and step father submit to them. They have told me, they **wouldn't** even attempt to lie on a polygraph.

Human cloning, particularly REM driven human cloning, is one of the absolute truths of this world; hidden from the world for over 70 years. My life and the freedoms of many, many people, including the reader (you) depend on this disclosure. The Illuminati say that the world finding out about REM driven human cloning will set the stage for the end of the world; people panicking and destroying property because of the evil these people have committed, but as I've said: prove them wrong, because we are going to inherit the earth, so there is no need to riot or cause public dissension. There will be no end of this world; just the end of the crazy Illuminati people. They always like to think and behave so negatively, don't they? *Sigh*. Furthermore, the whole talk about 2012; December 21st 2012 and the end of the world Mayan Prophecy; that was really about the world finding out REM driven human cloning, and the sick and sadistic nature it has been used for by these sick and sadistic people. Under no circumstance must you riot; you **must** prove you are better than these savage barbarians. I'll say it again, **because it is important**: we are going to inherit the earth, and therefore the smart and right thing to do, is not to destroy anything, especially the earth which you shall inherit.

Furthermore, I hope you are beginning to understand that the good people of this world must bring these people to justice and stop their wicked ways. I sincerely wrote this disclosure so that the good people of earth would be empowered to put a stop to their wicked ways. I wrote this disclosure to empower the good people of earth against tyranny and not so that the world would end. Therefore you must understand that the downfall of these Luciferians (Illuminati) must progress smoothly, because unless you have been to the cloning centre as a REM driven clone and retained the full memories of your experience, it is difficult to understand the level of depravity and sub human cruelty; it is beyond anything that has ever been heard of, and there is much more to tell.

Ignore this disclosure and you will condemn me to a horrible eventual death and you will encourage them that 'they are all 'powerful''. This is not an exaggeration. There is no need to exaggerate even President Barrack Obama is involved in the REM driven cloning business and attends the cloning centre as a REM driven clone version of himself, when he goes to sleep. He and his wife Michelle Obama attend these REM driven clone gatherings. Barrack Obama has even told me as a REM driven clone, at the cloning centre "Donny, we're all powerful. You are a slave and the people here [as REM driven clones at the cloning centre] won't speak up for you for fear of torture or death. Now make us a new song or we'll gut you like fish and leave you to writhe in agony!"

Be Wary of: Shills; Trolls; Disinformation Agents and Double Agents

There are many, many disinformation agents whose aim it is, is to denounce or debunk me. Be very wary of them. They DO NOT, and cannot debunk me, because after all, everything I have disclosed is truth. Although the truth concerns highly advanced concealed technologies, so it is a bit difficult to fathom at first, but keep reading, and you will realise that all which I have disclosed is true. Be very wary of the disinformation agents.

The most common thing these people say is that "Donald is a paranoid schizophrenic" –yet, ask them to detail the ins and outs of what a paranoid schizophrenic is and what specifically qualifies Donald to be a 'paranoid schizophrenic'; there is a high probability they will not be able to explain this to you, logically, sequentially and methodologically. Their main goal is just to make it as difficult for newcomers to understand what the real truth of this world is and as I have disclosed; and one of the root causes of man's problems is REM driven cloning.

Some disinformation agents are paid by the Illuminati, to behave this way, and are all too happy to echo empty statements (without any substantial evidence for their claims) because it provides them with a pay check. I will provide the following reference for the reader which discusses the signs and symptoms of a paranoid schizophrenic; all for the express intent for the reader to further understand paranoid schizophrenia. You can read more about paranoid schizophrenia on Medical News Today (2015). The link is provided in the reference section. You are welcome to read my Facebook wall and within days, you will realise I converse with others quite naturally. You can also ask individuals who have spoken to me through Skype or met me; but if this is not a viable option for you; all you have to do is read my Facebook wall and within days, you'll see that whoever makes such harsh and untrue remarks on my character is in fact a disinformation agent.

Furthermore, remember we live in a 3D **physical** world; and there is always a physical aspect to how something manifests. Therefore, when you hear others discussing concepts such as 'Tila Tequila demonstrated an 'energy ball' right on camera; Why didn't Donald discuss that?!' Please understand that these people are trying to deceive you with misinformation. There are also professional disinformation agents to be wary of. I do not perceive anything extra-ordinary in the world when I am awake in my original body beyond the five senses. I don't see orbs, auras or anything of the like. All I will ever discuss are practical things which have been unequivocally proven to me as fact; because in the situation the earth is currently in, I cannot afford to speculate.

Contact Information

My Facebook is: https://www.facebook.com/donald.marshall.148
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

I have a public wall on Facebook which starts from March 2012. In other words, all posts are made public since March 2012 and you do not have to join my friends list or subscribe me to see what I post (although you will have to have a Facebook account). I suggest that newcomers start reading from March 2012, and be patient, and read everything. You can read everything and learn the real truth of the world free of charge. I don't wish to write a book, I am not looking to turn a profit from this; I want to crush these guys and shut down the cloning centres!

It also worthwhile to add that: I am **not** asking for donations. I do not want **ANY** donations. Ever! These disclosures are far more important than any donations. This is **NOT** my job. I have a job. I am an independent contractor; carpentry is my trade and I earn a living this way. If you sincerely want to help; the best way to help is to spread and share my disclosures. That is all I ever ask for, so that eventually, the world knows about REM driven human cloning, and the armed forces can bring these people to their knees. **That is all I want; spread, spread, and spread this disclosure.** I will **never ever** ask for donations. Ever! Please keep this in mind, and anyone who asks for donations in my name or on my behalf should not be trusted. I, Donald Marshall, will **NEVER EVER** ask for any donations. I hope that is well received.

You can also view Proboards which has all my Facebook posts and have been archived by Celine O'Carroll and Astral 7ight by visiting:

http://donaldmarshall.proboards.com/
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

There is a search function on Proboards, and you can use this to search for and read all the disclosures I have made regarding REM driven clones, the people involved and more. I understand that it is human nature to want to know, which people have been to the cloning centre as REM driven clones; therefore, use the search function to read about any public figure which I have already covered that you have an inkling about. You can also post anonymously on Proboards and Celine and other Administrators will transfer your question onto Facebook which I'll answer.

Donald Marshall Revolution is a website which details a brief overview of the Illuminati.
http://donaldmarshallrevolution.com/
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

DONALD MARSHALL
REVOLUTION

Interviews

You can listen to the radio interviews I have done. Listen for consistency; particularly anything which you do not hear me, pronounce clearly for the first time; the best thing to do is to pause the recording at that particular point and replay it. You should also research the statement you do not understand. Sometimes reading helps comprehension a lot faster.

One of the main reasons you should pause and replay the recordings is because: the truth has been kept hidden for so long that a lot of what I discuss in my interviews are beyond most people's current world view; so at some points I may speak too fast for you; my audio/ microphone may not be so clear so you may miss what I say etc.

A friend has told me that when he first watched the Vinny Eastwood interview; he did not hear me say the word "scars" (When Vinny asked: 'How do I know I'm the real me?') although he replayed that particular point in the video 8 times. Everything was just beyond his current comprehension, at the time. No matter how many times he replayed that part he really could not hear me say the word "scars" –so he let that part go, and played the rest of the interview, pausing, and replaying points which he did not understand, especially to comprehend whether I was talking about my original body or my REM driven clone duplicate body. He also listened to all my interviews for consistency, to note any 'slip ups', or any parts of my testimonies which do not 'add up'. He would listen to all my interviews, pausing and replaying parts he did not understand and he would reserve his judgements until he felt everything I was saying was for example as ordinary as: 'I woke up today, brushed my teeth, and took the dog for a walk'. He was also patient to realise the truth. After listening to my interviews he would just let it 'sink in'. A week later he would come back and listen to the same interviews, to test whether his comprehension on the topics I discuss has improved, and whether he can understand what I am saying without having to pause and repeat at certain points in my interview; and soon enough he could now hear me say "scars" at that particular point of the Vinny Eastwood interview. He had reached the point where all topics I discuss sounded to him like I'm saying everyday common place stuff that people have heard, such as: 'I woke up today, brushed my teeth, and took the dog for a walk'. For anyone who may struggle to understand the topics I discuss: I strongly recommend you take the above approach as my friend did; soon enough, you too will realise the real truth of the world like he has: REM driven cloning, kept secret and used for sinister purposes.

I cannot say the following is true for everybody, however, an unproductive venture a complete newcomer can do is to listen to my interviews first time, all the way through, without pausing or replaying parts which they do not fully understand; If you do this and if there is just a single part of my interviews which does not make sense to you; this will interfere with your understanding of the entire interview. Remember, all I am discussing is technology, thousands of years advanced compared to what you currently use; available today, hidden and secret. If things start to get too complex for you, reduce it to its bare minimums: (advanced) science and technology. I hope that helps.

Another thing which I do in my interviews, that friends have picked up on, is: -because REM driven cloning has been my reality for many years; I don't differentiate between my original body and my REM driven clone body. I just say: I did this, I did that, and Elizabeth did this and that, therefore it can become very confusing for newcomers. Please bare with me; although I'm more emotional as a REM driven clone, and not as smart as I am in my original body (this is a side effect of cloning: REM driven clones are more emotional than normal, and dumber than they are in their original bodies) I'm still "me" when my consciousness is transferred; I have all the experiences and knowledge which makes me, "me" and therefore I naturally do not differentiate between my REM driven clone version, and my original body as an outsider discussing these concepts would. I understand it helps comprehension so I have painstakingly done this throughout this disclosure.

I hope this helps; and I hope this helps to better understand my disclosure as well as the interviews which I have done. On the next page you can find the links to my interviews. You can copy and paste the links to your web browser or press Ctrl+Click – (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the images below) to direct you to the interviews.

Links to Donald Marshall Interviews

Vincent Eastwood

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=M_1UiFeV5Jg&ab_channel=VincentEastwood
OR Press Ctrl+Click (on the image below) to follow the link.

Jeanice Barcelo

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=3uzgu4ekT3c&ab_channel=JeaniceBarcelo
OR Press Ctrl+Click (on the image below) to follow the link.

Lisa Phillips (CFR)

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=UonnFuHLJKc&ab_channel=Astral7ight
OR Press Ctrl+Click (on the image below) to follow the link.
Listen to parts 1 through 8.

Donald Marshall
1st Part Of Interview
April 11, 2013

Radio Presenters –Contact Donald Marshall

Anyone who sincerely wants to contact me for radio interviews on their show is welcome to do this. Please contact me through Proboards by leaving a message for me to contact you.

Professionals who understand "Consciousness Transfer" –Contact Donald Marshall

Any neuroscientists, engineers or professionals who understand how consciousness transfer works, and can provide me with a detailed methodology of how to block the consciousness transfer to my REM driven clone; please message me on Proboards, and this will be greatly appreciated.

Email

I currently do not have a contactable email address. In my original disclosure, the email address has been compromised (hacked), and so has any other email accounts I created: Yahoo, hotmail, Gmail, AOL etc. It doesn't matter; they eventually get hacked; for whatever reason, the Illuminati do not want me to have an email account.

How to Learn More about Donald Marshall's Whistle Blowing Disclosures

The best ways to learn about the truth of the world is through my Facebook page and Proboards. There is a mountain of evidence which has been collated over the years, and corroborates all that I have disclosed here. I also understand that for some people it is impractical to read three years worth of Facebook posts, especially when some post are repetitive; contain unnecessary comments from trolls, shills and other disinformation agents. Therefore, I am currently compiling 5 documents which should make it easier for others to read and understand everything quickly; as well as, for others to share and spread those 5 documents including this main disclosure document –so 6 documents about the real truth of the world. All for free!!

These documents include:

- "Frequently Asked Donald Marshall Questions" –This will contain all the general FAQs which I have been asked over the years; all in one place.
- "Experiences from the cloning centre" –This document will contain the full disclosure of my REM driven cloning experiences from the cloning centre when I go to sleep; all in chronological order.
- "List of people to avoid" This will contain a whole list of trolls, shills, and disinformation agents and double agents in alphabetical order, to avoid at all costs, in real life and online; -all complied in one place; and the reasons why they should be avoided.
- "Public Figures and their relationship with Donald Marshall as REM driven clones" –this will be a complete list of every public figure I have ever come into contact with as a REM driven clone; and you will be able to read about all the public figures I have met as REM clones, as well as the experiences, all in one place.
- "The subterranean underground colonists: The Vril" –this will be everything about the biological parasites the Illuminati harbour. Yes. The Illuminati harbour biological parasites, which they use against unsuspecting humans. I told you the Illuminati were

> the biggest depraved perverts ever to grace God's earth. REM driven human cloning, kept secret from the world for over 70 years is difficult to fathom as it is, as are these biological parasites. These parasites are also one of the main root causes of humanity's problems here on earth, and must be rendered extinct. In this document you will be able to learn everything about their, strengths, weaknesses, psychology, biology, ecology, sexuality and arm yourself with knowledge which ensures that humanity exterminates these parasites from our earth, forever. The secret of Vril; a secret no more!

I hope after reading my eye witness statements presented in this document that you have now began to understand that this is simply what is happening in our world today.

Heed my disclosure, very diligently. At the start of my disclosure I asked the reader to do two things:

1) Give me the benefit of the doubt –because there is nothing worse than being a victim of abuse, reporting it and the people who have the power to stop it, ignore you;
2) Set out on a genuine quest to debunk my disclosure. I understand that for some people, plenty of what I have disclosed will still sound 'too out there' beyond their current scope of reality to be believable. For such people, the honest thing to do is to start with the first topic which you do not fully understand and explore it, either by finding that corresponding topic on Proboards (Donald Marshall Proboards 2015), and continuing your research from there; or just Google any topic you currently do not understand and review the topic on websites such as New Scientist (2015), Gizmag (2015), Motherboard (2015), BBC News (2015a; 2015b), BBC Future (2015), The Guardian (2015a; 2015b) The Independent (2015a; 2015b), Daily Mail Online (2015) and start reading more articles in the Science and Technology columns because these people have been telling you what they have been doing for decades. Consequently, keep in mind that articles which discuss 'future' technologies and 'improvements' in medicine or science, are articles, which are **really** discussing present technologies, and achievements in science and medicine already realised; because most of these achievements have been realised decades ago. It just hasn't been fully disclosed publicly.

I have also saved and backed-up every reference (except the homepages of the websites referenced) in this document; therefore if a link or video is ever deleted let me know.

How Much Do You Know About Post-humanism / Trans-humanism?

Furthermore, you have to be honest here: because if you do not know much or anything about post-humanism / trans-humanism then I am afraid to say, -and this is no fault of yours – that you are behind in this the world; simply because these Illuminati people are Post-humanists; and that is the direction the want the rest of the world to go. They want the world to reach a destination where human cloning is common place; a world where downloading your mind / entire life experience onto a computer microchip and living on as a cloned version of you with the microchip running the consciousness to be common place; and these Illuminati Post-humanists want the world to head in a direction where it is commonplace to molest children through science and technology. No. This really is not Science-Fiction, this is the world we live in today; this is why such marvels in science, medicine and technology has been withheld from the general public and kept hidden by an 'elite' group of people because if these people honestly came forward, and said “We have invented this technology and we want to use it for this “negative” purpose” –the populace would not accept it, and these people will be lynched in the streets before lunchtime. Therefore, through concealed advancements in medicine, science and technology, the Illuminati can live out their inner depraved cravings and commit crimes against humanity and the world, for which most people will be too blind to see or even fathom, because the hidden science and technology is beyond the average person’s current comprehension.

Therefore, please attain an education in post-humanism / trans-humanism because these Illuminati people believe they can rival creation, nature and the marvels of the universe through science, medicine and technology. The Illuminati go **against** creation; they go **against** nature; and they go **against** humans. They don’t consider themselves humans. They consider themselves post-humans and believe they can become “gods” through advancements in medicine, science and technology. See LawOfIdentity (2014) and Mark Dice (2014) for Richard Seed’s comments on ‘becoming gods’ through transhumanism. Richard Seed is a Physicist with a Ph.D. from Harvard University; he is well known in the controversial cloning debate and declares his aspirations to "become god" saying 'We are going to become Gods, period. If you don't like it, get off. You don't have to contribute, you don't have to participate, but if you are going to interfere with me becoming a God, you're going to have trouble. There'll be warfare.' If you value humanity, do your best to attain an education in the trend of post-humanism / trans-humanism; otherwise the future of mankind will be bleak. We must do everything we can to stop these people.

This is of uttermost importance. REM driven cloning of is the most terrible thing in the world, especially when it involves unsuspecting civilians, and worst of all, innocent children. If you choose NOT to do anything you allow the Illuminati to continue to clone your children, sisters, wives, and sons. You allow the Illuminati to continue to hijack the minds of your children, sisters, wives, and sons while they sleep, through concealed advancements in science and technology. You allow the Illuminati to transfer the consciousness of your children, sisters, wives and sons to their REM driven duplicate clone versions while they sleep, whereby the Illuminati molest your children, sisters, wives and sons, which will cause them to have learning disabilities, unexplained depression and suicidal thoughts, as well as all kinds of side effects.

This IS your ONLY chance to end these monsters. There is a deadline!

This is your **ONLY** chance to do something to end these monsters. We are on a timeline and there is a deadline. As I have already mentioned, as well as, the top scientists in the field of physics and geophysics have mentioned: HAARP is not only capable of controlling the weather; once the HAARP grid is completed and working at full power, it will be capable of mind control over entire populations. We will all be slaves to the Illuminati FOREVER. Our freewill will be gone forever. Furthermore, if they achieve their aim of time travel with HAARP technology, future generations of humanity will never be able to stop them, because as they have told me: ‘They will always be able to go back to a previous time, and correct the mistake(s) which led to their downfall’. This is the Illuminati’s end game and what they wish to achieve. This is why this disclosure is so heavy, and so important. It goes beyond REM driven cloning; although it is important the world understands REM driven cloning. The future of humanity is at stake; and because of the 2-5 year deadline until the HAARP grid is completed, it is time humanity stood up and brought an end to the Illuminati. This REALLY is your ONLY chance to end these monsters!

The only way for evil to triumph is for good people to do nothing.
Edmund Burke

Remember for every 12 months which passes military / concealed technology outstrips the technology the general public is accustomed to by 44 years (Schneider 1995; 1996). Close your eyes and think 44 years into the future. The year is 2059. What kind of technologies do you expect mankind to have? Do it for real; don't just read the words. Close your eyes and imagine. The year is 2059. What kind of technologies does mankind have? You see. It is 'out of this world' technology, right. This is exactly what is going on today in the year 2015. It's just that it is hidden. That's all. Not so difficult to fathom now.

Do NOT Be Afraid to Help. We Outnumber the Illuminati by 1,000,000:1

Don't be afraid to help. These people are easily defeated when good people stand together as one. These Illuminati people total no more than 10,000 people. 10,000 people against 7,000,000,000 (7 billion) people –that's less than 0.000001 % of the world's population. Now can you begin to understand how prevalent evil can be when a small organised group of individuals, as little as 10,000 of the world's most evil and tyrannical people all work together in unison to exert their influence over the world? The world does not have to be this way. We outnumber the Illuminati people by a 1,000,000 to 1. For every Illuminati person there is, there are **one million** people who are not Illuminati, therefore do not to be afraid to help; all you will be doing is helping humanity rid itself of its sickness. So please feel free to share and spread my disclosure far and wide.

For those of you whom this applies to: remember the Illuminati (and I know it sounds ridiculous despite the evil they do) are deeply religious. They believe anyone aiding me is part of the "Army of Light" prophesised by Nostradamus; and to harm or degrade the life of anyone assisting me will bring them to their ultimate end and they will incur the wrath of God; they will suffer utter ruin and demise in their lives if they are to hurt you; they are very scared of people who can see through the lies and deceptions they have inflicted on the world. So please, stay calm, do not stress your heart, and know that **you are safe**, and **they cannot hurt you**. You can feel safe in the knowledge that you can do the right thing by helping me. They told me this as REM driven clones on the night of 21st of February 2014 when I went to sleep. The above is exactly what they said.

For anyone interested in understanding the interpretations of the Nostradamus Prophecies, Crystal Links (2015) provides a good source for all 942 of the quatrains. Delores Cannon is also an author who has written three volumes called "Conversations with Nostradamus" and for anyone interested you can read these online (Galactic 2012a; Galactic 2012b) as well as watch her YouTube videos (CreativeForceVideo 2014; Disclose TruthTV 2015).

Yes. This is the state of the world today: a deception within a deception coated in reverse psychology; and fact is indeed stranger than fiction. Now after everything I have said, if you are still sceptical and 'in-between' on this issue then the best thing for you to do is to pay attention to your dreams; or lack of dreams –they are your own experiences and you cannot deny your experiences.

In other words, you may be having REM driven cloned experiences too; for some people they have a 'dream', where they are in the same environment, over and over and over again; or the theme being discussed in their 'dreams' is the same over and over and over again; or the environment their 'dreams' take place always appear to happen in the same 3 or 4 environments over and over again –no matter what the theme is. In each scenario there is a likely probability that these are not just dreams, and in fact these people have had their consciousness transferred to the REM driven clone duplicates and the Illuminati is trying to extract something from them, once that is done they implant a false memory (See Kim (2013) and Alford (2015) for discussions on how false memories are implanted) so you wake up with the feeling of remembering something, but in actuality it is false; or you wake up with no memory of dreams from the previous evening, which means there is a high probability that you had your memories suppressed when your consciousness was transferred to your REM driven clone. This is why I say you can be sceptical about everything I have said (for now), but pay attention to your dreams, or lack of dreams, they are your own experiences, and you cannot deny your own experiences.

Noteworthy Frequently Asked Question

A Frequently Asked Question I receive, which is worth a mentioning here, is as follows: **Q:** Donald, if everything you have said is true...; in other words human clones walk among us now in a multilayered conspiracy which reaches the highest levels of government; armed with this knowledge what does one do exactly? What happens now, Donald? You obviously want to spread the word and make people aware but to what end?

DM: In short: I want to bring a complete end to the Illuminati and usher in a "Golden Age" of mankind. This question is best answered by detailing my 'Mission and Vision' for ending the Illuminati completely.

Donald Marshall's Mission and Vision on How to Bring Down The Illuminati

- This disclosure must spread, and **spread FAST and FAR!**
- If you now understand everything I have disclosed in this document as truth, then do not waste any more time. Share this disclosure with your wives, husbands, brothers, sisters' aunties, uncles, friends, co-workers, and children. We all have a part to play in saving the world from a premature doom.
- Call friends who have not heard from you in a while and tell them you have important news to share. Share this document. Share it on social media such as Facebook; Twitter; Instagram; Dropbox; Slideshare etc.
- Keep sharing this document until it comes to the attention of the Armed Forces. The Armed Forces will have to intervene. Once the Armed Forces intervene we will have reached the middle stage of this vision and will be witnessing a complete end to the Illuminati. Until then, we're in the beginning stages, so please; spread this document faster and further. The quicker it is spread; the quicker the Illuminati are ended.
- Once the Armed Forces have intervened a Military coup can be orchestrated against these vile people.
- Cloning centres can then be shut down once these people are overthrown.

- The freeing of any missing people (children; teens; adults) trapped in the cloning centres can commence (once the Armed Forces intervene).
- HAARP (as well as other highly advanced technologies) can now be contained and not used for adverse effects against the world (once the world learns of this disclosure).
- These Illuminati people will then HAVE to appear in court for their crimes against humanity.
- Suing the Illuminati members in court (once court proceedings commence... we will be past the midpoint of my vision, and closer towards witnessing a complete end to the Illuminati).
- The populace (and it will be your choice) can then have their REM driven clone experiences restored and also sue and claim any legal reward / compensation.
- After the populace has sued the Illuminati for their crimes against humanity; the punishments can commence.
- Punishments will include: imprisonment and executions of these sick and malevolent people. The imprisonment and executions of evil Illuminati members will bring an end to the Illuminati.
- After the punishments and executions; Governments worldwide can now be replaced with incorrupt individuals, worldwide.
- The structure of Governance will also have to change. The reason for this because future generations will always be able to check their leaders and governance more appropriately; so that the depraved and subhuman acts I described can NEVER be committed against humanity again; nor will world leaders be capable of committing such monstrous crimes in such secrecy ever again.
- Ensuring that the laws of the land always progresses in line with advancements in medicine, science and technology; as well as, ensuring law progresses in the directions of Research and Development (R&D) regarding future technologies, science and medicine.
- Release of technologies, science and medicine which benefit mankind.
- Commence a “Golden Age” of mankind.

If you do not fight the corruption and you do not stand up for what is right in life, you end up being a serf and a slave and you are leaving your children a world in which you would not want to live in yourself, so how can you in decency behave that way? You have to stand up for what is right in life, and unless you do that you are nothing.

Aaron Russo

Empowerment by Virtue of Golden Truth

As you can see, despite all the horror of the world I currently present: if everything goes smoothly, the good people of this world, truly will inherit it. This is why despite any anger, sadness, or fears you may have; you must not riot, damage property or cause chaos or bring about any other form of public dissension. You cannot stay silent, or ignore the issue in the hope that the threat removes itself. You are called to act; you must take action to help bring the Illuminati to an end; you must act while at the same time, you must maintain your composure and resolve to ensure that the whole procedure goes smoothly, and we all inherit a world we want to live in.

Your life, **at this very moment**, is more important than you may have probably ever imagined. You have purpose. Through you, and other good people around the world, together we can bring an end to such unspeakable forms of tyranny in this world. It is my humble stance that you have now become truly empowered with golden truth and you are now compelled to bring this tyranny to an end.

Do not waste the knowledge you have obtained from this disclosure. It is my only hope to escape this man made living hell. It is my only hope, as well as, the hope of many REM clones imprisoned there, as well as, real people who go missing daily, and are trapped in the cloning centres.

We sincerely beg you.

Donald Marshall

Appendices

Appendix A: Thien Thanh Thi Nguyen (Tila Tequila) Transcript

Link: https://www.youtube.com/watch?v=7mRZ7ItF9ls&ab_channel=Astral7ight

00.00 – 1.10 min: You know what... since you f****** with my program darling Queen Elizabeth [II] and the paedophilia ring and the cloning centres, and the cloning centres. That's right darling the cloning centres. Parents listen to me right now, they are blocking me but that is quite alright. They are blocking me but that is quite alright. Because I have many, many other forces; I shall not say their names right now, but I have many big plans to expose all of you disgusting, sadistic f***s! Okay? That is all. I shall save that for another time. But, however, I shall REPEAT: that was just an introduction to the reptilian family, leading all the way back. They call themselves the "The Black Nobility". Now that is just one part of it; alright?

1.11 -2.47 min: The Black Nobility; the reptilian family; all the way back from ancient times; so which they think... they feel like they are the divine chosen ones... from whom may I ask? Definitely. Definitely not God. Our creator of the world. Reptilians: they feed on blood; children; the paedophilia ring; recently busted. Oh! It has been going on for centuries. Parents listen to me carefully. I don't care if I'm cutting out. I will continue this and I am not alone on this battle. Believe THAT! Believe THAT! I am not alone on this battle. I started out alone but I am marching on with MILLIONS; okay? So sit your old a** down okay. Because you are gonna roll over, but it doesn't matter anyway because you are all ancestral f***s! Who interbreed... ancestral f*****g... and then... and then... Oh! Only going to talk about cloning those children and... oh! And all those many children's parents listen to me carefully.

2.48 -4.24 min: Hundreds of missing children come up every year. You wonder... why? How? How could this be? And then there is so called CPS [Child Protection Services] or whatever they are called; they come and take your children, just, just for nothing; right? Not to discredit everybody, because not everybody is bad. I'm talking about the bad people. And they take your kids... they take them as this... they treat them like... I can't even say the word. It's disgusting, it's sadistic. They... they... they take your kids... they toss them out like little [inaudible] cause they are so f*****g... Pardon my French, but then again when I'm dealing with these evil cum-buckets I have no holy words coming out of my mouth, because these are the only words they resonate to. So therefore that's how I refer to them. Because they can only [inaudible] ...their masters whiplash on them with these [inaudible] words of cursing, vileness and slaving and that is not what the true God is; okay? The God of Hell...; Anyways...

4.25 – 5.30 min: There are these CLONING CENTRES where they take your children and do sadistic things to them. I'm not even talking about child molestation here; not to mention; uh there is one of them that got caught, flying out to Florida; to meet up with a four year old little girl, to have sex with a four year old girl. That's right. Google it because it is so highly sinful... We have commercials about... you know starving kids; you know save the starving kids and it is heartbreaking. We can have commercials about it... why? Because it is a horrible thing and people can have commercials about it. WHY do we not have commercials about... 'Daddy please don't, don't. Mommy please don't let daddy let daddy touch me?' – because it is disgusting! It is so disgusting; beyond sin that no one, NO ONE, can even make a commercial about that because that's how sinful it is.

5.31 – 7.19 min: Do you understand that? Do you understand how sinful that is? These people, I'm not even going to call them people; alright? They take your children; they not only molest them; men f*** them and make them shoot each other. They give them guns. It's either you shoot him or I'm gonna shoot you. They are... mind you, they are children. Children. Yes. I'm speaking out, because I... uh! Who else is doing this? You're all just [inaudible]... shame on you... And actually the most recent paedophilia... got taken down... WOW! How long did that take? Really?! Do you know how long this has been going on? *Sigh* Alright I'm gonna calm down... but as a parent and I love parents out there. My heart goes out to all the parents out there who have missing children. You know, we all pray for them; every day. And I put on a bold face in public every day; because there needs to be someone strong, believe that. But my strength comes from somewhere... that I have a very vulnerable emotional side where I feel very strongly for these children and innocent people. So therefore I want to speak out.

7.20 – 9.10 min: I have and I have my passion too; and in the end you shall all know why I am so passionate about exposing every single one of these scumbags; okay? The truth shall prevail and you all will know why. So, as for you parents... ah... there are no words to describe. But let me just expose because you can't just... there is a point where you CANNOT just turn the other way. You know this stuff is going on; and you go 'Oh well... that's their problem. Let's just turn the other cheek.' How long are you going to turn the other cheek, until it happens to your own freaking children? When, when, when your own child; three years old, get's run out [kidnapped] and gets blasted right in the head with a gun... yeah... there's more to that people; okay? And I'm not just saying that coming from some... I mean, actually, mothers, fathers out there... If you found out... I'm sure you would do way worse than what I am just saying. I'm just using voc. I'm just annihilating them vocally. I'm sure the parents out there who find out what their children have gone through, their missing children. I am pretty positive more than just a vocal annihilation of these scumbags that do this to your children; okay?

9.11 – 10.32 min: So keep on turning the other cheek folks. Hey, go and turn the other cheek; you are with the others. I am not. I AM NOT. Like I said you... There is only two ways to go about this: you're either with us: the good guys, or you're with the others. There is no in-between. Cause if you are in-between, hiding like cowards, turning the other cheek; doing whatever; well then you are a freaking coward and you are just a sheep. And sheep end up dying because you know what you are owned by "The Others". So pick one: you're either one of us: the good guys or you're part of the others. That is simple as that! Two choices: good guys; bad guys. In-between you're dead because the bad guys are going to suck your soul out [transfer your consciousness to your REM driven clone] and do some sadistic stuff to you and watch, and make you watch while they pretty sadistic stuff to your children as well.

10.33 – 11.33 min: Do you know why they love children? Because they are innocent souls; they're innocent... they're, they're the most innocent pure beings in this planet. They're not harmed by anything. They're new to the world; bright-eyed pure innocent children. That is why these disgusting paedophile and these clone rings; cloning centres; satanic rituals; Brownsville Texas... There are many other cloning centres where they take your children that go missing. And you wonder why? What happens to them? I'm sorry to break this to you but that is what happens. Now either do you want to know what happens to them or do you want to turn the other freaking cheek?

11.34 – 13.09 min: Tune into my next show. I'm gonna upload stuff; I have an arsenal ready to blow up! Okay. And I have reason behind this. You all should know my personal reason soon; but this has nothing to do with me right now. But know that I'm back, I'm back with a vengeance and I'm back with an army full of people around the world who are sick and tired of treated like animals; or quote, unquote "COWS". We all know what that means. For the outside world (the masses) we all know the term "sheep", sheeple. But for the insiders we know what the cows are don't we? You know what "The Others" like to do with the cows, right? They start to herd you in... and to... yeah...I'm gonna end it at that. And to all the parents, families and children out there, I love you so much. I... have to maintain composure, because that is what I do. That is all. Over and out.

Appendix B: MK Ultra

MK Ultra *-noun:* [Manufacturing Killers Utilizing Lethal Tradecraft Requiring Assassinations] The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct, and to control the absolute behaviour and thought patterns of the individual.

.

Modern-day MK Ultra involves an implantable microchip which is inserted into a Mark 2 clone of the victim. Every thought, reaction, hearing and visual observation causes a certain neurological potential, spikes, and patterns in the brain and its electromagnetic fields, which can now be decoded by the implantable microchip into thoughts, pictures and voices. The thoughts, pictures and voices of the implanted Mark 2 clone (victim) can now be displayed visually and heard on any system capable of converting visual images, such as a television or a computer. These images are usually displayed on a giant screen at the cloning centre.

The purpose of MK Ultra is to elicit a specific conditioned response in a victim (desired by the programmer) to an otherwise neutral stimulus. In other words, through the process of "classical conditioning" (which involves learning a new behaviour via the process of association. In simple terms two stimuli are linked together to produce a new learned response in a person) the implanted Mark 2 REM duplicate clone (victim) is placed thorough many similar recurring scenarios which are experienced (or perceived) by the victim as "real", because the programmer manipulates the victim's audio and visual field, and continuously pairs the victim with the neutral stimulus, so that either a positive or negative conditioned response is elicited in the victim towards the neutral stimulus.

In other words an implanted victim can be placed in many recurring situations deemed as frightful, where a mystery stranger saves the victim from the frightful situation over and over again. The victim will now have positive associations towards the mystery stranger and therefore the conditioned response is now one of 'positive associations' (towards the mystery stranger who did not elicit such a reaction before).

MK Ultra techniques are administered through Mark 2 (sleep driven) clones. In other words classical conditioning happens when the person sleeps. Memory suppression technologies are also used in conjunction with MK Ultra technology, which enables the programmer to control certain memories the victim remembers. The use of memory suppression technologies and MK Ultra technology allows the programmer to reinforce behaviour and elicit specific conditioned responses (in an original).

Therefore, in our example above, when the original meets the mystery stranger in real life, the original (victim) will be predisposed to the mystery stranger and automatically 'feel safe' around this person and may even believe that he / she should 'date this mystery stranger' – depending on what was programmed; because the victim was conditioned to elicit such an emotional response towards the mystery stranger in his / her sleep as a Mark 2 REM driven clone through MK Ultra functionality.

MK Ultra can also be used to create zombified individuals who are programmed to murder and remember nothing of their crime afterward. MK Ultra (technology) can also be used to disrupt the memory of the original, discredit people through unusual behaviour, make the person insane or commit suicide and murder.

MK Ultra has much functionality. Furthermore, because of its video and audio projection functionality, MK Ultra can also be used to compose music. MK Ultra is capable of relaying the Mark 2 REM driven clone’s subconscious and conscious mind as visual images and audio, projected onto a screen. Therefore thoughts, pictures and voices which have been experienced consciously and subconsciously in one’s life can be displayed visually and audibly on a television or computer screen. The Mark 2 clone now has the option of harmonising these thoughts, pictures, and voices into a coherent order which produces music through the thought process of his or her imagination. See Jim Cristea (2009); Berkeley News (2011); UC Berkeley Campus Life (2011) CTForecaster (2013); nature video (2013) and Stromberg (2013) for examples and discussions and the capability of MK Ultra technology.

Appendix C: The Illuminati

Illuminati *–noun:* A modern-day criminal organisation operated by reprobate (depraved, unprincipled and wicked person) criminals. Their main agenda: is to enslave the whole world through advanced concealed technologies.

Today's Illuminati trace their roots back to Professor Adam Weishaupt who found the Illuminati on 1st May 1776. Since the inception of the Illuminati the intent has always been, and remains: "to bring about a NEW World Order that writes God out of the picture and deifies (glorifies) Lucifer." This intent is still prevalent today. The following excerpt is derived from A. Ralph Epperson (1990) *The New World Order:*

> Weishaupt was a teacher of Cannon Law (law governing the affairs of a Christian Church, especially the law created or recognised by the Papal authority in the Roman Catholic Church) at the University of Ingolstadt in Bavaria, now part of Germany.
>
> He even told the world, in his writings, where he would conceal the Order: "None is fitter than the three lower degrees of Free Masonry; the public is accustomed to it, expects little from it, and therefore takes little notice of it." He felt that this secrecy would lead him to success because he felt no one would be able to break into it. He wrote: "Our secret Association works in a way that nothing can withstand ...."
>
> Weishaupt accepted the fact that all secret associations and secret orders had two doctrines, one concealed and reserved for the Masters... the other public ...." and the Illuminati was [and are to this day] a secret society with two doctrines.
>
> Professor Weishaupt, its founder, boasted of his organization's secrecy. He realized that this secrecy would enable them to decide the fate of nations and because their deliberations were secret, no outsider could interfere. He wrote: "The great strength of our Order lies in its concealment; let it never appear in its own name, but always covered by another name, and another occupation." Weishaupt later wrote about that secrecy in a letter to a fellow member of the Illuminati: "Nothing can bring this about [the new world order] but hidden societies. Hidden schools of wisdom are the means which will one day free men from their bonds [the "bonds" of religion] Princes and nations shall vanish from the earth." So the secret societies were created to bring the world to the new society known as the New World Order. The members of these organizations obviously feel that their goals are so noble that they may perform whatever tasks are required of them to bring that goal to fruition. This means that murder, plunder, and lying all become acceptable as long as these methods assist its members in obtaining their goal.
>
> Adam Weishaupt, the founder of the Illuminati, wrote over and over and over again, that "the ends justified the means." Weishaupt also told initiates to use whatever means, which included murder, to achieve the goals of the association that he was joining. And that the major goal of the Illuminati, was the destruction of all religion, including Christianity. That meant that if Christians physically stood in the way, they could be removed by simply murdering them. Weishaupt even went so far as to say that anyone not willing to take the life of another was unfit to join the Illuminati. He wrote the following in a letter to a fellow member in 1778: "No man is fit for our Order who is not ... ready to go to every length ...."
>
> Another reason, that Weishaupt felt that the Illuminati would succeed, was the fact that he was offering his members worldwide power. He felt that this inducement would enable him to draw into his organization only those who would do anything to satisfy that desire for power. He wrote: "The true purpose of the Order was to rule the world. To achieve this it was necessary for the Order to destroy all religions, overthrow all governments and abolish private property."

> But his religion had a different base than the traditional religion: his was based upon a worship of reason: "... then will Reason rule with unperceived sway." "... Reason will be the only code of Man. This is one of our greatest secrets." "When at last Reason becomes the religion of man, then will the problem be solved." Weishaupt's dedication of his organization to "reason" makes some sense when the reader recalls that "reason" has been defined as the "unbridled use of man's mind to solve man's problems without the involvement of God." The Bible calls this "the fruit of the tree of the knowledge of good and evil." It was this knowledge that God wanted man not to have, and it was the promise made to man by Lucifer that man could have it by eating of "the fruit." In addition, Weishaupt's religion offered its believers a reward not offered by any other religion: worldwide power!
>
> Weishaupt wrote: "The pupils [members of the Illuminati] are convinced that the Order will rule the world. Every member therefore becomes a ruler." Weishaupt's religion not only offered power to his believers, but he offered them something else not guaranteed by any other religion: worldly success. He said that once a candidate had achieved the exalted degree of Illuminatus Minor, the fourth of the thirteen inside his Order, his superiors would: "assist him [the member] in bringing his talents into action, and [would] place him in situations most favourable for their exertion, so that he may be assured of success." Finally, the goal of the Illuminati was "man made perfect as a god - without God."

The ideology of “man made perfect as a god –without God” still remains to this day, and it is practiced by today’s Illuminati members. The above phrase is what ties in Luciferian worship and trans-humanism. Lucifer is idolised by Illuminati members as the deity who gave man ‘knowledge’ and therefore is worthy of worship; God, -according to Luciferians, -did not want man to have knowledge and therefore is despised by Luciferians. Ingrained in the trans-humanism doctrine is the believe that: ‘man can become ‘god’ through science and technology and in turn overthrow the Creator of the universe: God’. See LawOfIdentity (2014) and Mark Dice (2014).

These are the basics of Luciferianism. Therefore everything which is natural or pertains to nature must be contended or destroyed by Luciferians. This is why Illuminati members endorse having sex with children, killing first born sons, and drink blood. All the above go against nature and according to the ‘edicts of Lucifer’: paedophilia makes the person committing the act younger (it doesn’t, it is just an excuse to act perverse on children because they know children are vulnerable); killing your first born son gives you good luck and fortune in this life (so yes, some Luciferians have sacrificed their first bon sons); and Illuminati members believe drinking blood / cannibalism is a ‘purifying agent’ (although in reality it causes spongiform encephalitis (holes in the brain)).

Modern-day Illuminati members still retain the goal of its founder “to bring about a NEW World Order that writes God out of the picture and deifies (glorifies) Lucifer.” Infiltration through secrecy, still remains their mode of operation, for the current Illuminati and therefore they have secretly infiltrated all the major religions on earth; government and education – where each successive generation is being dumb down; they have continually diminished the ability for individuals to own private property, or claim inheritance; divided people against each other to continually diminish patriotism; and have continually diminished family values.

They also compartmentalise their knowledge between members. Until I fully exposed the Illuminati, many people who have been REM clones at the cloning centre did not know they were in fact REM driven clones, and thought they were in the 5th Dimension; the astral plane; a singularity; the spiritual realm; Valhalla; quantum hopping; a time stutter etc. or whatever else the Illuminati told them.

Today's Illuminati members also meet in secret (just like the founding members) as REM driven clones, **when they go to sleep**. Furthermore, because Illuminati meetings are in secret and not many people know the exact location (because knowledge is compartmentalised) of the cloning centre; or the fact that their consciousness has been transferred to Mark 2 REM driven clone bodies at the cloning centre (and they are not in their original bodies); as well as the fact that unsuspecting civilians have their memories suppressed; the points mentioned above are the reasons the Illuminati believe 'they are all powerful and untouchable'; and as a consequence, today's Illuminati members do all the disgusting things they want, because they believe no outsider can interfere.

The ring leaders of the Illuminati today also believe "the ends justify the means". This is why they clone, torture, molest, murder, and rape unsuspecting civilians as REM driven clones in their sleep. The Illuminati of today offer their members incitements to go along with their agenda and not oppose them (or face death). Another popular method is to entice their target with many, many wonderful prospects, and have the target believe they are joining a noble and prosperous venture, so that the target fulfils the objectives of the Illuminati unknowingly; promoting the Illuminati in a positive way, because the target has been deceived to perceive the nature of the Illuminati as 'positive'; by the time the target finds out the true intent of the Illuminati, and the evil which emanates from it, it is too late. Those who rise up in the ranks of the Illuminati are the men and women who have an insatiable lust for power, and most importantly: the men and women who want to rule the world.

Modern-day Illuminati members also wish to become gods (through technology); overthrow the Creator, and achieve their overall aim of enslaving mankind. This is why they clone people, and clone people in high rank society from all walks of life (movie stars, musicians, politicians etc.; whether the person willing wants to be part of the Illuminati or not) so long as that person is in a position of power and influence, the Illuminati clone that person, and threaten that person, for example -"Hey, you're going to hang with us –or else" –through such coercion, the people in high rank society who have power and influence will not oppose the Illuminati's plans to become gods; enslave mankind forever; and rule the world. Another reason for cloning high rank society is to include these people into the Illuminati (willingly or unwillingly) to ensure the world's populace remains in ignorance (because once all the world leaders and high rank society are cloned and under the coercion of the Illuminati nobody in a position of influence or power can warn the populace against the Illuminati) until the Illuminati's plan is completed and they have enslaved the world forever.

The Illuminati's overall aim of 'becoming gods' and ruling the world, as well as, mind controlling all the inhabitants of earth is also the reason the Illuminati:

- Administer drip feed disclosure through media, by telling some truths mixed with lies in order to conceal their true intentions and overall aim, and prevent the betrayed partner (the public) from ever discovering "the complete truth";
- Administer evaluative conditioning, by placing their symbols and ideology in popular media with positive associations, so that the unknowing and unsuspecting public will eventually become predisposed to the Illuminati and unsuspectingly have a positive or neutral response towards the Illuminati;
- Because they want to become 'gods' is also the reason the Illuminati is promoting RFID microchips and only discussing RFID microchips positively, while at the same time placing suppression (gagging) orders on anyone who speaks negatively about RFID microchips –which implies an unsuspecting public will willingly accept the microchip; and at the point of transaction, the person will have (unknowingly) given up their privacy to a third party (the Illuminati) for the rest of his / her life;
- Their aim to become 'gods' is also the reason the Illuminati are hurriedly trying to complete the HAARP grid across earth –because a complete HAARP grid will allow them to achieve their goal of mind control over the entire world; which fulfils their objective of becoming gods; because a complete HAARP grid will be capable of time travel, and therefore the Illuminati will be capable of going back to a previous time to correct the mistake(s) which led to their downfall; the Illuminati members will retain the knowledge of the previous time, and the rest of humanity will have no recollection of such a memory.

The Illuminati is **not** a joke. It is **not** fiction. They are very real, and part of humanity's reality; and through advanced concealed technologies the Illuminati aim to enslave humanity forever. The ring leaders of the present Illuminati includes Queen Elizabeth II, Prince Philip Duke of Edinburgh, Prince Charles of Wales and Vladimir Putin. I have also detailed the actions of the ringleaders in the main text in this disclosure, as well as, detailed other modern-day evil Illuminati members on my Facebook and Proboards. It is time the good people of earth, stopped being afraid, do the right thing, put a stop to this evil, and save themselves, as well as their children's children from being slaves forever. Spread and share this disclosure.

Glossary

Aneurysm -*noun:* An excessive swelling of the wall of an artery at a fixed point in the body. A brain aneurysm is therefore a: bulge or ballooning in a blood vessel in the brain. It often looks like a berry hanging on a stem. A brain aneurysm can leak or rapture, causing bleeding into the brain.

Brain Aneurysm -*noun:* see Aneurysm.

Clone –*noun:* a cell, group of cells, an organism produced asexually from a single ancestor and is genetically identical to a single ancestor.

Concealed Technology -*noun:* hidden machinery and devices undisclosed and currently unavailable for public consumption. Concealed (or military) technology develops at a rate of 44 years for every 12 months which passes in comparison to the technology the public is currently accustomed to. Origin: Phil Schneider.

Cloning Centre –*noun:* a physical location (on earth) where duplicate and replicate clones are produced. These physical locations are usually Deep Underground Military Bases (DUMBs). DUMBs have an entire floor dedicated to cloning. It is also a place where Illuminati members meet with each other as REM driven clones. The above-ground cloning centre where many high profile people attend can be found within a radius of 5/6 hours drive from the Robert Pickton Farm Port Coquitlam, British Columbia, in Canada, somewhere at a remote nature reserve.

Cloning Technology -*noun:* the technological advancements in medicine, science and technology used to produce duplicate and replicate copies of originals.

Conditioned Response -*noun:* an automatic response established by training to an ordinarily neutral stimulus.

Conditioned Stimulus –*noun:* A previously neutral stimulus that, after repeated association with an unconditioned stimulus, elicits the response produced by the unconditioned stimulus itself.

Consciousness –*noun:* the state of being aware of and responsive to one's surroundings; a person's awareness or perception of something.

Consciousness Transfer -*noun or verb:* the process of transferring or copying the mental content (including long-term memory and "self") from a particular brain and copying it to a computational device; artificial body or avatar body such as that of a robot or clone version of the original. It is also the feat in which the person's mental content (long term memory and "self") moves from one body into another.

Depopulation *verb:* to remove or reduce the population of, as by destruction or force.

Drip Feed Disclosure *noun or verb:* is the process of supplying information but in small amounts overtime. Drip feed disclosure is also the process of revealing information slowly overtime, possibly telling lies to conceal certain aspects of the truth until the source administering the drip feed disclosure has adequate time to let out the truth in a slow and controlled way, thereby delaying the betrayed partner (in this disclosure, the public) from having the “complete truth” for some time.

DUMB *–noun:* [Deep Underground Military Base] a facility directly owned and operated by or for the military or one of its branches that shelters military equipment and personnel, and facilitates training and operations beneath the surface of the earth.

Duplicate Clone *-noun:* a fully formed human body which is a genetic copy of original developed through the process of regenerative technology. Duplicate clones are grown in a big thick tank full of (salty) water.

Duplication Cloning *-verb:* involves agitating the cells of on an original repetitively until a fully formed human body of the original is developed. Duplicate clones take an average of 5 months to form into a fully developed human body of the original through the process of regenerative medicine and technology.

Evaluative Conditioning *-noun:* is a change in liking, which occurs due to an association with a positive or negative stimulus.

H.A.A.R.P. Technology *–noun:* [High frequency Active Auroral Research Program] a radio transmitting system that can bounce signals off the Earth’s upper atmosphere, (60km (37 miles) to 1000km (620 miles) high) back to probe deep into the earth or sea. HAARP is also capable of: disrupting human mental processes; knocking out all global communication systems; changing weather patterns over large areas; interfering with wildlife and migration patterns; hurting ecosystems; negatively affect human beings health, moods and mental states; and unnaturally ‘boil’ the earth’s upper atmosphere. HAARP used correctly will control the weather without any adverse effects.

H.A.A.R.P. Grid *–noun:* a network of radio transmitters which can bounce signals off the earth’s upper surface. Each transmitter is located at a specific point across earth and communicates in unison with other radio transmitters across the earth. At this present time of writing, a HAARP grid has **not** been completed, although the Illuminati are working twice as fast to complete a HAARP grid. The threat to humanity once a HAARP grid is completed includes: mind control over the entire world’s inhabitants. A completed HAARP grid will also be capable of time travel, and therefore the Illuminati will always be able to go back to a pervious point in time to correct the mistake(s) which led to their downfall. Humanity will be slaves forever.

Habeas Corpus *-noun:* is [A] writ [formal document] requiring a person under arrest to be brought before a judge or into court, especially to secure the person's release unless lawful grounds are shown for their detention.

Heart Attack *-noun:* A sudden occurrence of a blockage of the flow of blood to the heart.

Human Clone *–noun:* The creation of a genetically identical copy of a human.

Ionosphere *–noun:* the layer of the earth's atmosphere which contains a high concentration of ions and free electrons and is able to reflect radio waves. It lies above the mesosphere and extends from about 60km (37 miles) to 1,000 km (620 miles) above the earth's surface.

Illuminati *–noun:* A modern-day criminal organisation operated by reprobate (depraved, unprincipled and wicked person) criminals. Their main agenda: is to enslave the whole world through advanced concealed technologies. See Appendix C for further discussion.

Mark 2 Clone *–noun:* Is a sleep driven clone; specifically, a REM sleep driven clone. A Mark 2 Clone is activated by transferring the consciousness of an original into a Mark 2 Clone **only when** the original reaches REM sleep (usually 90- 110 minutes **after** the original falls asleep). Once the consciousness of the original is transferred from the original's body to the Mark 2 Clone, the Mark 2 clone is now capable of motion: such as walking, talking etc. The Mark 2 clone 'drops limp' and becomes motionless once the original wakes up from sleep. The original's consciousness no longer resides in the Mark 2 clone (once the original is awake) and therefore the Mark 2 clone is now incapable of motion. Mark 2 Clones are also known as "REM Driven Clones" and "REM Duplicate Clones".

Memory Suppression *–noun or verb:* is the selective removal of memories or associations with the mind using memory suppression technology.

Memory Suppression Technology *–noun:* any scientifically advanced technology used selectively to remove memories from the conscious mind.

Mind-Voice Technology *–noun:* an advanced technology capable of reading, listening, hearing and broadcasting a person's inner voice or thoughts. It is capable of replicating sounds **exactly**. Therefore an individual can hear the sound of drums, a guitar or any instrument and replicate that sound exactly just by thinking about it. Consequently, Mind-voice technology has the functionality of producing music.

Military Technology *–noun:* machinery and devices developed from scientific knowledge used by the Armed Forces which advance at a rate of 44 years for every 12 months which passes, compared to the technology the public is accustomed to. Origin: Phil Schneider

MK Ultra *-noun:* [Manufacturing Killers Utilizing Lethal Tradecraft Requiring Assassinations] The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct and to control the absolute behaviour and thought patterns of the individual. See Appendix B for further details.

Neutral Stimulus *–noun:* is a stimulus which initially produces no specific response other than focusing attention. In classical conditioning, when used together with an unconditioned stimulus, the neutral stimulus becomes a conditioned stimulus.

Negative Association *-noun:* is an undesirable experience or perception.

Negative Stimulus *-noun:* a stimulus with undesirable consequences.

New World Order *-noun:* [NWO] Agenda. The whole NWO agenda is to turn humanity into mindless slaves forever; whereby the post-humans / trans-humans mind control the entire world's populace either through RFID microchips or a completed HAARP grid. Another aspect of the NWO agenda is to depopulate the world's current population of 7.3 billion people to 500 million people (and never exceed a world population of 500 million people afterwards); ruled by a one world government; a one world ruler; with a one world religion. See Appendix C for more details.

Original *–noun:* A person who is **not** a clone.

Pain Receptor *-noun:* Any one of the many nerve endings throughout the body that warn of harmful changes in the environment such as excessive pressure or temperature.

Positive Association **-***noun:* is a desirable experience or perception.

Positive stimulus *-noun:* a stimulus with desirable consequences.

Posthumanism *–noun:* seeks to rewrite the very definition of being human. It is the condition in which humans and intelligent technology become intertwined. In the Posthuman there are no essential differences or absolute demarcations between bodily existence and computer stimulation, cybernetic mechanism and biological organism, robot technology and human goals.

Posthuman *-noun:* see Posthumanism.

Project MK Ultra **-***noun:* see MK Ultra.

Regenerative Medicine *-noun:* (of a living organism) the process of re-growing new tissues after loss or damage.

Regenerative Technology *-noun:* any machinery or device developed from scientific knowledge which has the capability to re-grow new tissues after loss or damage.

REM Sleep *–noun:* [Rapid Eye Movement] is the fifth stage of sleep in the sleep cycle. It takes 90 - 110 minutes to reach REM sleep after we fall asleep. REM sleep is also known as the "period of paralysation". The involuntary muscles such as the brain become more active whereas voluntary muscles (those that you move by choice) such as your arms and legs become more relaxed or paralysed. REM sleep is a kind of sleep that occurs at intervals during sleep, and it is characterised by rapid eye movements.

REM Driven Clone **-***noun:* [Rapid Eye Movement Driven Clone] a clone that can only become activated, once the original is in REM sleep. See Mark 2 Clone.

REM Duplicate Clone **-***noun:* [Rapid Eye Movement Driven Clone] A clone developed by regenerative medicine and technology and is therefore an identical copy of an original. REM duplicate clones can only become activated when the original is in REM sleep. See Mark 2 Clone.

REM Driven Clone Death –*noun:* the process where an original dies because of constant torture to their REM driven clone or where a constant electrical current is applied to the REM driven clone resulting in death of an original usually in the form of an aneurysm or heart attack (because consciousness is linked) in the original's body.

REM Driven Clone Torture —*noun or verb:* the action or practice of inflicting serve pain on a REM driven clone. REM driven clone torture causes biological and physiological responses in the original's body because consciousness is linked. Intermittent REM driven clone torture (depending on what is done) causes the original to experience severe headaches, an upset stomach, achy limbs, sickness; a weakened heart. Continuous REM driven clone torture will lead to the death of the original; usually in the form of aneurysm or heart attack in the original's body.

Replication Cloning –*verb:* involves giving birth to a genetic identical of an original where the newborn starts life off as a baby and matures. The newborn is referred to as a clone.

RFID Technology [Radio Frequency Identification] -*noun:* are electronic microchips the size of a grain of sand that can be directly embedded into the human flesh. RFID microchips communicate wirelessly through the use of electromagnetic fields to transfer data. RFID microchips link the brains of people via the implanted microchip to satellites controlled by ground base super-computers. The dangers of RFID microchips to the implanted person are: total loss of privacy and total control of the person's physical body functions, mental and emotional thought processes, including the implanted person's subconscious or dreams –for the rest of that person's life! RFID microchips are also tracking devices, and the implanted person can be tracked anywhere on the globe.

Selling One's Soul -*verb:* to sell the use one's "Mark 2" REM driven clone to the Illuminati, for the Illuminati to use the individual's Mark 2 REM driven clone in whatever manner the Illuminati wishes. There are no returns once the individual has signed over his / her (soul) Mark 2 REM driven clone. When an individual sells their (soul) Mark 2 REM driven clone, the person has also entered into a contract to sell the Mark 2 REM driven clone(s) of their current children (if they have any) as well as any unborn children the person may have later in life. The person sells all their descendants (souls) Mark 2 REM driven clones to the Illuminati, once the individual sells their (soul) Mark 2 REM driven clone to the Illuminati. Selling one's (soul) Mark 2 REM driven clone is considered **a serious business transaction** to the Illuminati. There are no returns. If the person ever makes a fuss and wants their (soul) Mark 2 REM driven clone back, the Illuminati will either torture the person's Mark 2 Clone, or apply a constant electric current to the person's Mark 2 Clone until the person either has a heart attack or aneurysm in their original body. This is what public figures are hinting at when they say "They have sold their soul". They have sold the use of their Mark 2 REM driven clone to the Illuminati. Selling one's soul is not a joke. Never sell your soul.

Stimulus –*noun:* is something that causes a reaction, especially interest, excitement or energy. It is also an energy change registered by the senses. For example a stimulus can be a shinny object for a baby.

Technology –*noun:* machinery and devices developed from scientific knowledge.

Technological Advancement *–noun:* is incorporating, by means of experimental development, a characteristic or capability not previously existing or available in standard practice, into a new or existing process or product that enhances a product's performance. Novelty, uniqueness, or innovation alone does not indicate a technological advancement.

Transhumanism *–noun:* the belief or theory that the human race can evolve beyond its current physical and mental limitations, especially by means of science and technology.

Unconditioned Response *–noun:* is a response to a neutral stimulus we have no / little control over. It is a natural automatic response. For example, food is an unconditioned stimulus for a hungry animal, and salivation is the unconditioned response.

Unconditioned Stimulus *–noun:* A stimulus that elicits an unconditioned response.

BLUE BLOOD, TRUE BLOOD
CONFLICT & CREATION
A Personal Account
By Stewart A. Swerdlow

Expansions Publishing Company, Inc.
P.O. Box 12, St. Joseph MI 49085

With Excerpts From
Belief Systems Shattered
by Janet Swerdlow

BLUE BLOOD, TRUE BLOOD: CONFLICT & CREATION
A Personal Account
Copyright © 2002 Expansions Publishing Company, Inc.
Cover Art and Illustrations by Lori Sarich
Typography and book design by L'OR Intuitives
Editor: Janet Swerdlow
Published by: Expansions Publishing Company, Inc.
P.O. Box 12
Saint Joseph MI 49085
All rights reserved. Printed in the United States of America. No parts of this book may be used or reproduced in any manner whatsoever without written permission except in the case of brief quotations embodied in critical articles and reviews.
For further information contact Expansions Publishing Company, Inc.
P.O. Box 12, Saint Joseph MI 49085 U.S.A.
email: stewart@stewartswerdlow.com
janet@janetswerdlow.com
Website address: www.stewartswerdlow.com
Library of Congress Cataloging-in-Publication Data
Stewart Swerdlow
Blue Blood, True Blood: Conflict & Creation
A Personal Account
256 pages
ISBN 0-9626446-6-8
Library of Congress Catalog Card Number: In Process

Other books by Stewart Swerdlow:
Montauk: The Alien Connection
The Healer's Handbook: A Journey Into Hyperspace
The White Owl Legends: An Archetypal Story of Creation
Forthcoming books:
Dream Dictionary
Mind-Pattern Analysis
Books by Janet Swerdlow:
Belief Systems Shattered
In Search of Yourself
Practical Tips For Everyday Living
Life Support Group ™ Leader's Manual
Little Fluffs Children's Book Series
Forthcoming by Janet Swerdlow:
Children's Group Leader's Manual

DEDICATION

It is with the deepest of heart and soul felt gratitude that I wish to thank my wife, Janet, for her unconditional love and support during my journey through the depths of hell. She was there all the time, waiting for me, and never giving up hope. Without her I would not be here any longer. There are no words in any human language that can express how immensely grateful I am to her for being my twin flame/other half. I have learned brutal lessons these past couple of years. They would have all been in vain if it were not for Janet. Thank you, my true love, from the depths of my being. There is also a great admiration for my parents and sister who also stayed vigilant during these trying times. They left no stone unturned in exposing the vulgarity and depravity of the demons sent to harm me. I give my wonderful family eternal hugs, and love, and the peace of mind that I shall never be captured again.

As always, my incredible children, who love me without end, have proven to me that there is no space or time that can separate me from them. As they grow into adulthood, I realize that children are a gift that we are allowed to take care of for a while so that the magnificent blossom can grow forth on its own and create a new garden. My children are my jewels and my treasures. Thank you for believing in me.

Finally, I want to thank my fabulous and generous friends who have made my return to my own life easy and happy. Their generosity is overwhelming and their love unforgettable. I am proud to count among my friends: Steve and Sandy Wade, Todd Berran, Lori Sarich, Duncan Cameron, Shannon O’Shea, and Aldee Filley. You have all been as dear family to me.

I also thank you, the public, both readers and students, who have stood by me through better or worse. I am grateful for your continued support and respect. It is my goal to continue to bring you the truth so that your lives may also be healed and that you may experience your personal connection to the God-Mind.

Contents

Preface ... 7
In The Beginning... ... 9
The Transparent People ... 13
Is This Planet Taken? ... 17
The Reptilian Agenda ... 19
Conflict & Creation ... 23
Bluebloods! ... 27
Other Alien Groups ... 31
The Crystal Skull ... 35
The Ancient Hebrews ... 37
The Annunnaki & The Black Race ... 41
Mind-Control Via Religion ... 43
The Roman Empire ... 49
Emmanuel ... 51
Magdalene Lineage & Beyond ... 55
The Illuminati Hierarchy ... 61
Blood-Ritual & Ceremony ... 69
Ritual In Red ... 79
My Lost Children ... 87
Around The World ... 93
Alliance En France ... 103
Illuminati Agent ... 109
Tango In Argentina ... 115
Mystery Man & Company ... 119
The End Of Rosie 2 ... 125
Swiss Keys ... 129
Boy From Brazil ... 133
Future Illuminati Plans ... 139
Reflections & Observations ... 145

APPENDICES

Alien Groups ... 150
Milky Way Galaxy ... 171
Reptilian Androgyny ... 172
Mixture Of Lineages ... 173
13 Ruling Families - Illuminati ... 174
Sexuality & Mind-Pattern ... 175
News Clips from Expansions Exposé ... 177

Excerpts
Belief Systems Shattered
By Janet Swerdlow

All Experiences Are Neutral ... 183
Is Your Reaction Fear? ... 187
Subverting Sexuality ... 191
Vegetarianism---Is It Really The Way? ... 195
Are You A “Sleeper?” ... 197
Illuminati As Teachers ... 201
Victim Mentality ... 205
Are You A Mind-Control Candidate? ... 209
Earth’s Mind-Pattern ... 213

One Mind-Pattern=Global Control 217
Patriotism vs. Warmongering 221
Simultaneous Existence 225
Index 229
Expansions Ordering Information 249

PREFACE

AS FAR AS THE HISTORY OF THIS PLANET IS CONCERNED, YOU CAN CONSIDER THAT EVERYTHING YOU EVER LEARNED IS A COMPLETE LIE. ALL HISTORY AND SCIENCE BOOKS ARE REWRITTEN TO ACCOMMODATE THE AGENDA OF THE CONTROLLERS OF THIS PLANET. THESE BOOKS ARE JUST AS FALSE AS THE NEW AGE MATERIAL SPEWING FORTH FROM SUCH PLACES AS SEDONA AND SANTA FE, TO NAME A COUPLE OF ILLUMINATI MECCAS.

DISINFORMATION IS RAMPANT EVERYWHERE. MY INFORMATION COMES FROM MY MONTAUK PROJECT INDOCTRINATIONS, EXPERIENCES, CONVERSATIONS WITH SCIENTISTS INVOLVED IN ILLUMINATI PROGRAMS, COMMUNICATIONS WITH ALIEN AND INTERDIMENSIONAL BEINGS WHOM I MET AT VARIOUS GOVERNMENT PROJECTS, AND THROUGH THE PROBING OF MY OWN OVERSOUL.

I CANNOT OFFER YOU PHYSICAL PROOF AT THIS TIME. I CAN ONLY TELL YOU THAT IT IS AVAILABLE IN CERTAIN PLACES. ALTHOUGH ALL EXISTENCE IS SIMULTANEOUS, AND TIME AND SPACE ARE MERELY ILLUSIONS OF PHYSICAL REALITY, FOR THE PURPOSE OF THIS BOOK I WILL PRESENT THE MAJORITY OF HISTORY FROM A LINEAR PERSPECTIVE. THERE ARE ALSO INFINITE UNIVERSES, BOTH PHYSICAL AND NON-PHYSICAL, BUT FOR NOW, I WILL ONLY TELL YOU ABOUT THIS ONE.

TO UNDERSTAND HOW THE ILLUMINATI CAME TO POWER, OR EVEN TO COMPREHEND WHAT THEY ARE, IT IS IMPORTANT TO UNDERSTAND THE BEGINNING OF LIFE ON THIS WORLD AND ITS PROGRESSION.

I DO NOT BELIEVE IN ANY RELIGION, ORGANIZED OR OTHERWISE. ALL RELIGIONS, NO MATTER WHAT THEY ARE, ARE FORMS OF GROUP MIND-CONTROL DESIGNED TO MANIPULATE LARGE MASSES OF PEOPLE TO STOP THINKING FOR THEMSELVES. I DO BELIEVE IN GOD. IT HAS NO RELIGION.

IN THE SUBSEQUENT YEARS FOLLOWING THE PUBLICATION OF MY LAST BOOK, MUCH HAS BEEN WRITTEN ABOUT BY OTHER RESEARCHERS CONCERNING REPTILIANS AND THEIR RITUALS. THIS THEME HAS FOUND ITS WAY INTO THE PSYCHE OF THE PUBLIC. MOST PEOPLE ARE STILL NOT WILLING TO ACCEPT OR EVEN ENTERTAIN THE IDEA THAT HYBRID REPTILIANS ARE IN CONTROL OF THIS PLANET AND PERFORM BLOOD RITUALS AND CEREMONIES USING HUMANS AS THE SOURCE OF FOOD AND HORMONES. I WISH IT WERE NOT TRUE. BUT I CANNOT CHANGE HISTORY, CURRENT EVENTS, OR WHAT I KNOW TO BE TRUE.

THE CONTROLLERS PLAN THE PATTERN OF COMING EVENTS IN SUCH A WAY TO PURPOSEFULLY DISORIENT THE MASSES WHILE THEY ALONE SEE THE DESTINATION AND THE ORDER OF LIFE. THE MASSES ARE LEAD TO BELIEVE THAT HAPHAZARD EVENTS AMIDST CHAOS SHAPE THEIR EXISTENCE.

IN TRUE REALITY, THERE IS NO SUCH THING AS CHAOS. CHAOS IS SIMPLY A PATTERN NOT YET UNDERSTOOD OR PERCEIVED.

THINK ABOUT AN ANT CRAWLING OVER A DESIGNED, TILED FLOOR. THE ANT MAY SEEM CONFUSED OR DISORIENTED; NOT KNOWING WHICH WAY TO GO. BUT A HUMAN BEING WATCHING THE ANT FROM A HIGHER PERSPECTIVE CLEARLY SEES THE FLOOR PATTERN AND KNOWS WHICH WAY THE

ANT CAN GET TO ITS DESTINATION. TO THE ANT, THERE IS ONLY PERCEIVED CHAOS. TO THE HUMAN, THERE IS A PRESCRIBED PATTERN.

TO THE CONTROLLERS, THE HUMANS ARE THE ANTS. THE CONTROLLERS ARE INTERESTED IN DIRECTING THE PEOPLE IN SUCH A WAY THAT THEY DO NOT KNOW THAT THEY ARE BEING DIRECTED. THE CONTROLLERS WORK SLOWLY AND METHODICALLY FROM A GLOBAL PERSPECTIVE TO ACCOMPLISH THEIR GOALS.

WITH AWARENESS, YOU TOO, CAN PERCEIVE ORDER FROM THE CHAOS, MAKE SOME CONSCIOUS DECISIONS, AND GAIN CONTROL OF YOUR OWN LIFE AND DESTINY.

GO BACK

2 - THE TRANSPARENT PEOPLE

WHEN I WORKED AT MONTAUK, I ENCOUNTERED AN OCCASIONAL ALIEN SPECIES KNOWN AS A REPTILIAN. THEY SEEMED TO POP IN AND OUT OF PHYSICAL REALITY. THE REPTILIANS PRIMARILY USE THE LOWER ASTRAL REALMS AS THEIR REFERENCE POINT, OR POINT OF ENTRY, INTO PHYSICAL REALITY. THIS IS THE ORIGIN OF THE LEGENDS OF ASTRAL DEMONS. MY MONTAUK CONTROLLERS EXPLAINED THAT THESE BEINGS WERE BROUGHT TO THE DRACO STAR SYSTEM EONS AGO BY ANOTHER GROUP UNKNOWN TO ANYONE. FURTHER, THEY EXPLAINED, NO ONE KNEW ANYTHING OF THEIR TRUE ORIGINS.

COMMUNICATION WITH ME WAS ENTIRELY TELEPATHIC (THE DRACO REPTILIAN). MENTALLY, IT TOLD ME THAT IT CAME FROM THE FAR FUTURE WHERE HUMANS NO LONGER EXISTED, AND THAT ITS SPECIES WAS NOT FROM THIS REALITY. CONTINUING, IT SAID THAT ITS SPECIES HAD TRAVELED BACK INTO THE DISTANT PAST TO CREATE A RACE OF BEINGS, WHICH I KNEW AS THE REPTILIANS, TO ANTAGONIZE AND TEST HUMANS.

IT ALSO TOLD ME THAT THEY WERE ASSISTED BY THE SIRIANS OF SIRIUS A IN THE CREATION OF THE REPTILIANS AND THEIR PLACEMENT IN THE DRACO STAR SYSTEM.

FOR THE REPTILIANS TO FUNCTION IN PHYSICAL REALITY, THEY NEEDED PHYSICAL GENETICS. THE TRANSPARENT PEOPLE TOOK GENETICS FROM THE NOW PHYSICAL LYRAENS, WHO HAD BLONDE OR RED HAIR, AND BLUE OR GREEN EYES. THESE GENETICS WERE MIXED WITH THE TRANSPARENT PEOPLES COLLECTIVE ENERGY, THUS MANIFESTING PHYSICALLY AS THE REPTILIANS. THIS IS WHY THE CURRENT REPTILIANS NEED THE ENERGETICS FROM ARYAN-TYPE PEOPLE TO SURVIVE ON THE PHYSICAL PLANE.

ONCE THE REPTILIANS WERE CREATED IN THE ASTRAL, THEY NEEDED A PHYSICAL HOME BASE FROM WHICH TO ACCOMPLISH THEIR TASK. FOR THIS, THEY WERE TAKEN TO MANY DIFFERENT PHYSICAL REALITIES IN WHICH THEY COULD BECOME THE DOMINANT SPECIES. MENTALLY, THEY WERE PROGRAMMED TO CONQUER AND ABSORB ALL OF THE RACES AND SPECIES THAT THEY ENCOUNTERED. THOSE THAT COULD NOT BE ABSORBED WERE TO BE DESTROYED. THE PURPOSE OF ALL OF THIS IS TO DETERMINE THE MOST PERFECT FORM IN PHYSICAL REALITY THAT CAN EXIST IN ANY ENVIRONMENT. THINK OF IT AS A GIGANTIC, COSMIC SURVIVOR CONTEST.

THE REPTILIANS ARE PROGRAMMED TO BELIEVE THAT THEY ARE THE SUPERIOR PHYSICAL FORM.

SCIENTIFICALLY SPEAKING, REPTILIAN DNA DOES NOT CHANGE VERY MUCH OVER EONS OF TIME. IT BASICALLY REMAINS THE SAME. FOR THEM, THIS IS THEIR PROOF THAT THEY ARE ALREADY PERFECT, WITHOUT ANY NEED TO ADAPT. MAMMALIAN LIFE, ON THE OTHER HAND, EVOLVES AND CHANGES FORM CONSTANTLY TO SURVIVE.

TO THE REPTILIAN MIND, THIS DENOTES WEAKNESS AND INFERIORITY. REPTILIANS ARE ALSO ANDROGYNOUS, MEANING MALE AND FEMALE IN ONE BODY. THIS IS COMPARABLE TO ALL NON-PHYSICAL FORMS THAT HAVE NO GENDER LIKE GOD-MIND. FOR THIS REASON, THE REPTILIANS BELIEVE THEMSELVES TO BE MORE GODLIKE BECAUSE OF THEIR ANDROGYNY. DUE TO THEIR ETHNOCENTRIC VALUES, THEY ALSO CONSIDER IT THEIR RIGHT TO CONTROL AND CONQUER ALL OF SPACE AND TIME.

ALTHOUGH THE REPTILIANS OPERATE OUT OF A GENERAL GROUP MIND, THEY ARE SEPARATED INTO SEVEN DIFFERENT REPTILIAN SPECIES, EACH CREATED TO PERFORM SPECIFIC FUNCTIONS.

THIS IS THEIR HIERARCHY, AND IS BASED ON THE NINE LEVEL ANGELIC HIERARCHY THAT I DISCUSSED IN THE HEALERS HANDBOOK - A JOURNEY INTO HYPERSPACE (SKY BOOKS, 1999).

AS AN ASIDE, THE HINDU CASTE SYSTEM IS A DIRECT REPLICA OF THE REPTILIAN HIERARCHY.

COMMENTARY FROM WINTER:

THE POLITICS OF THE INTERVENTIONIST DRACO TRADING COLONIES (SEE DUNE - WHOSE ARRAKEIS - THE HOME STAR OF DRACOS AND DUNE - IS CORRECTLY LOCATED IN ALPHA DRACONIS) - DEFINITELY BECAME REPRESENTED BY ENLIL/YAHWEH/MICHAEL SIDE OF THE AN - ANUNNAKI FAMILY. ALSO SHOULD BE SEEN AS THE ROOTS OF INSURANCE AND BANKING ON EARTH. THE TEMPLAR FOLLOWERS OF THE MAG (DRACO MATRILINY) LINE ARE OFTEN CREDITED WITH 'INVENTING' INSURANCE AND BANKING FOR THE WEST. THIS IS ABSURD - SINCE THEY HAD ALREADY INSTALLED SUCH INDUSTRY - SO CLOSE TO FEAR AND EXTORTION - TO HARVEST COLONIES ON HUNDREDS OF PLANETS. THE "SKULL AND BONES" - ORION- ORIGINS OF THE PIRACY AND EXTORTION STORY IS AN EXAMPLE OF INSURANCE AND BANKING THAT LOSES ITS PUBLIC RELATIONS BUDGET. MORE ON THE SKULL OF THE MAG BELOW.

LATER - AS ENLIL/YAHWEH UNDER THE NAME OF "GOD" MICHABO (MICHAEL) INSTALLS GENETIC RULES IN ABORIGINAL AUSTRALIA AND THE AMERICA'S HE IS FOLLOWING THE DRAC TRADITION ENTIRELY FOCUSED ON PREVENTING CHANGE - AND FACILITATING HARVEST OF COLONIES- LIKE FORESTS WHOSE ONLY PURPOSE IS TO BE CUT DOWN.

AT A DEEPER LEVEL WE SHOULD INVESTIGATE THE CONNECTION BETWEEN LOSS OF DNA IMPLOSION - THE PHYSICS OF LOSS OF SOUL & ACCESS TO BLISS - CALLED NEPHILIM WHICH BEFELL THESE DRACO (LOST SERAPHIM) FAMILIES. LOSS OF FIRE IN THE BLOOD CREATED LOSS OF LUCID DREAMING, LOSS OF "ENSOULMENT" AND LOSS OF LONG TERM MEMORY. PART OF THEIR PROBLEM WAS LOSS OF ABILITY TO NAVIGATE IN TIME TRAVEL ('`GUILD NAVIGATOR' IN DUNE - STAR NAVIGATING TIME TRAVELERS, ABLE TO WARP SPACE FROM WITHIN) - DUE TO THIS LOSS OF DNA COHERENCE.

NOT UNLIKE ONE OF THEIR RESULTANT CULTURES - TIBETAN - WHO ARE INTENSELY AWARE THAT THEIR IS NO STEERING THEIR COLLECTIVE CULTURAL DIRECTION (TIME TRAVEL) WITHOUT A LEADER WHO IS ABLE TO REMEMBER PAST LIVES. THIS IS EXACTLY WHAT DRACO LOST. IT IS

PROFOUND TO COMPARE THIS TO ZULU SHAMAN LEADER LAST WEEK BEMOANING TO US HOW HIS GRAND CHILDREN LOST THE ABILITY TO LUCID DREAM JUST AFTER THEY SUFFERED BEING IMMUNIZED. THIS IS AN EXAMPLE OF HOW BORG - FALLEN DRAC EMPIRES - TREAT DNA. BY NOT LOVING THEIR GENES ENOUGH TO SET THEM FREE - THEY BLIND THEIR CHILDREN.

GO BACK

GO BACK TO VIDA EN NUESTRA GALAXIA

3 - IS THIS PLANET TAKEN?

IN THE BELOW - WE SUGGEST COMPARISON TO LYSA ROYALL'S "PRISM OF LYRA" BOOK - A 'CHANNELED' MORE ROMANTIC AND INTUITIVE ACCOUNT- WHICH AT LEAST AGREES SURPRISINGLY ABOUT THE LYRAEN ORIGIN OF TODAY'S HUMANS.

BECAUSE THE LYRAENS DID NOT HAVE A DEFENSE SYSTEM IN PLACE, THEY WERE A SITTING TARGET FOR THE REPTILIANS, ALSO REFERRED TO AS THE "DRACO". AFTER BEING BRUTALLY ATTACKED BY THE DRACO EMPIRE, THE SURVIVORS OF THE LYRAEN SOCIETY DISPERSED TO OTHER LOCATIONS THROUGHOUT THE GALAXY. THE REMNANTS OF THE DRACO ATTACK ON LYRAE ARE STILL SEEN BY TODAY'S SCIENTISTS.

IN 1985, A NEWSPAPER ARTICLE STATED THAT SCIENTISTS ARE ABLE TO OBSERVE REMNANT WAVES OF A BLAST THAT FAN OUTWARD, AND EMANATE FROM A CENTRAL PART OF THIS GALAXY. THEY BELIEVE THIS BLAST TO BE SEVERAL MILLION YEARS OLD AND OF SUCH INTENSE PROPORTIONS THAT THE WAVE IS STILL TRAVELING TOWARD THE EDGE OF THE GALAXY BEFORE DISSIPATING. THEY CLAIM NOT TO HAVE ANY IDEA OF WHAT CREATED THE BLAST.

THESE SURVIVORS WENT TO ORION, TAU CETI, PLEIADES, PROCYON, ANTARIES, ALPHA CENTAURI, BARNARD STAR, ARCTURUS, AND DOZENS OF OTHER SOLAR SYSTEMS. IN THIS SOLAR SYSTEM, THE REFUGEES COLONIZED THE PLANET NOW CALLED MARS. AT THAT TIME, IT WAS THE THIRD PLANET IN THE SOLAR SYSTEM. A WORLD CALLED MALDEK WAS THE FOURTH PLANET IN THIS SOLAR SYSTEM, AND WAS ALSO COLONIZED.

THE LYRAENS WERE ALL BLONDE-HAIRED, BLUE-EYED PEOPLE, WITH AN OCCASIONAL RED-HAIRED OR GREEN-EYED PERSON. IN LYRAEN SOCIETY, RED-HAIRED PEOPLE WERE CONSIDERED SPECIAL WITH EXTRASENSORY POWERS THAT CONNECTED THEM TO NON-PHYSICAL REALMS. THEY WERE ESPECIALLY DESIRED FOR BREEDING PURPOSES. SPECIAL PERMISSION TO BREED WITH A RED-HAIRED PERSON WAS REQUIRED BECAUSE OF THE EXTRA, OR SUPERHUMAN, ABILITIES THAT CAME WITH THE OFFSPRING.

FOR THIS REASON, RED-HAIRED PEOPLE WERE KEPT SEPARATE FROM THE REST, AND EVEN HAD THEIR OWN SUBCULTURE. THEY WERE ALSO COVETED BY THE REPTILIANS, WHO AS A SPECIES DID NOT HAVE MUCH PSYCHIC ABILITY. OFTEN, WHEN REPTILIANS CAME TO A WORLD FOR OCCUPATION, THE LYRAEN REFUGEES OFFERED A GROUP OF RED-HAIRED PEOPLE TO APPEASE THEM FOR A WHILE. THIS PRACTICE EVENTUALLY DEGENERATED INTO SACRIFICES TO APPEASE THE DEMONS.

THE PLANET EARTH IN THOSE DAYS WAS A WATER WORLD IN SECOND ORBIT FROM THE SUN. THERE WAS LITTLE LAND ABOVE THE SURFACE. THE ONLY INTELLIGENT INHABITANTS WERE AN AMPHIBIAN RACE THAT WAS COMPLETELY WITHOUT TECHNOLOGY. THE ATMOSPHERE OF THE EARTH WAS MOSTLY LIQUID. THE PLANET DEFINITELY COULD NOT SUSTAIN ANY TYPE OF HUMAN

LIFE FORMS.

THE DISPERSED LYRAEN DESCENDANTS DEVELOPED THEIR OWN CULTURES OVER THE EONS OF TIME. EVEN THEIR GENETICS MANIFESTED DIFFERENTLY AS A RESULT OF THE MIND-PATTERNS OF EACH OF THE COLONIES. FOR INSTANCE, MARS AND MALDEK WERE SIMILAR TO THE CURRENT EARTH ENVIRONMENT, WITH WARM TO TEMPERATE CLIMATES AND AN OXYGEN-RICH ATMOSPHERE. THE GRAVITY ON MALDEK WAS DENSER THAT MARS, SO THOSE PEOPLE DEVELOPED A THICKER FRAME AND A MORE AGGRESSIVE ATTITUDE.

EVENTUALLY, SKIRMISHES DEVELOPED BETWEEN THE OCCUPANTS OF THE TWO PLANETS. MARS WAS RICH IN RESOURCES. THE PEOPLE OF MALDEK THOUGHT THAT THEY DESERVED THESE RESOURCES FOR SURVIVAL. THE MARTIANS ASKED THE BEINGS OF SIRIUS A, FROM THE PLANET KHOOM, FOR DEFENSE TECHNOLOGY TO SHIELD THEIR PLANET FROM ATTACK, NOT ONLY FROM THE REPTILIANS, BUT FROM THEIR HUMANOID NEIGHBORS AND COUSINS. THE SIRIANS ARE KNOWN THROUGHOUT THE GALAXY AS MERCHANTS OF TECHNOLOGY. THEY HAVE THE BEST, EVEN SHARING IT WITH THE REPTILIANS.

SO, THE SIRIANS CREATED A DEFENSE MECHANISM LOCATED IN THE MARS UNDERGROUND.

MARS IS A HOLLOW PLANET, AS ARE EARTH AND JUPITER. PLANETS CREATED WITH MATERIAL EJECTED FROM A STAR HAVE HOLLOW INTERIORS. AS A MOLTEN BALL IS THROWN FROM THE STAR AND STARTS SPINNING AWAY, IT BEGINS TO COOL. THE CENTRIFUGAL FORCE OF THE GLOBE SPINNING AND MOVING AT GREAT SPEED PUSHES THE MOLTEN INTERIOR TO THE SIDES, FORMING THE CRUST OF THE PLANET. THIS, IN TURN, FORCES HOT GASES OUT OF THE POLES TO FORM OPENINGS AT BOTH ENDS. THE MOLTEN CORE AND GASES THAT REMAIN GET TRAPPED BETWEEN THE HOLLOW INTERIOR AND THE PLATES BELOW THE CRUST OF THE GLOBE. THESE ARE PUSHED OUT PERIODICALLY IN THE FORM OF VOLCANIC ACTIVITY.

THE NEXUS POINT ON ANY SUCH GLOBE IS ALWAYS AT THE 19TH PARALLEL OF THE PLANET. IT IS EVIDENT ON EARTH BY THE HAWAIIAN VOLCANOES, LOCATED AT THE 19TH PARALLEL; BY THE MONS VOLCANO ON MARS LOCATED AT THE 19TH PARALLEL; AND AT THE RED SPOT ON JUPITER, ALSO AT THE 19TH PARALLEL.

THE GEOMETRY BUILT INTO THE MONUMENTS ON MARS BY THE SIRIANS AND LYRAEN COLONISTS EXPLAINS ABOUT THE 19TH PARALLEL THROUGH ITS GEOMETRIC EQUATIONS AND MEASUREMENTS.

THIS GEOMETRY IS ALSO REPLICATED AND CONTAINED WITHIN THE GIZA PLATEAU IN EGYPT.

COMPARE THIS TO RICHARD HOAGLAND - STILL CLUE LESS TO WHY DOLMEN LIKE GIZA AND CYDONIA ARE AT TETRAHEDRAL LATITUDES. BY MODULATING (ADJUSTING) THE SPIN RATE OF PLANETS AT TETRAHEDRAL HAND HOLDS - PLANETARY SPIN COMES INTO PHASE WITH ZODIAC SPIN - THIS SELF SIMILAR SPIN RELATION (IMPLOSIVE COLLAPSE) IS CRITICAL TO STABILIZING GRAVITY AND THEREFORE ATMOSPHERE.

GO BACK

4 - THE REPTILIAN AGENDA

THE REPTILIAN AGENDA WAS, AND IS, TO SEEK OUT THE HUMAN REFUGEES FOR DESTRUCTION OR ASSIMILATION, AND TO USE THEIR BLOOD AND HORMONES FOR SUSTENANCE.

IT IS SAID THOSE (OPHANIM? SERAPHIM? WINGED DRACS? CIAKAR? CHERUBIM?) WHO PLACED THE DRACO IN OUR GALACTIC SECTOR KNEW THAT THE HUMANOID REMNANT WOULD NEED AN AGGRESSIVE PARASITE TO TRIGGER DEVELOPMENT. FURTHER THE RESULT OF THE PART OF THE HUMANOID BLOOD ON EARTH TODAY RECEIVING A BLOOD LINE CROSS FROM DRACO (URU CROSS WITH IBI) - IS BENEFICIAL. (SEE LAWRENCE GARDNER DEFENDING HIS DRAC FAMILY ROOT). THIS IS RELATED TO USUALLY AN OR CELTIC OR BRITISH RH NEGATIVE BLOOD.

THE REPTILIAN DRAC BLOOD ENKI (RA - ABRAHAM) CROSSED WITH THE CRO-MAGNON MONKEY BLOOD HAD DNA WAS LESS VITAL THAN THE INDIGENOUS (RHESUS) MONKEY BLOOD. SO THE REPTILIANS HAD TO KEEP IN-BREEDING (AS IN HOLY BLOOD HOLY GRAIL) - CONSISTENT WITH WHAT SWERDLOW SAYS BELOW. AREAS (UK) WHERE THIS RHESUS MONKEY BLOOD IS ABSENT (RH NEGATIVE) MAY LEAD THE WORLD IN LACK OF COMPASSION. E-LIZ-A BETH (FROM LIZARD BORN) BREEDS A LINE WHERE NO PERMISSION TO CRY PRODUCES A STIFF UPPER LIP.

THE REMNANT LYRAENS WHO COLONIZED OTHER PLANETS FORMED AN ALLIANCE AGAINST THE CONSTANT REPTILIAN ATTACKS. THEY CALLED THIS ALLIANCE THE GALACTIC FEDERATION, COMPRISED OF 110 DIFFERENT COLONIES. THE COLONIES BELONGING TO THE FEDERATION WISHED TO MAINTAIN THEIR NEW IDENTITIES, AND NO LONGER ASSOCIATE WITH THE OLD WAY. TOGETHER, THE FEDERATION COLONISTS MANAGED TO REPEL THE REPTILIAN ATTACKS.

THERE WERE THREE PRIMARY GROUPS WHO DID NOT JOIN THE FEDERATION. THESE THREE GROUPS WERE CONSIDERED EXTREMISTS, OR NATIONALISTIC IDEALISTS, SEEKING TO RECREATE THE GLORY OF THE OLD LYRAEN CIVILIZATION. ONE GROUP WAS THE ATLANS, LOCATED ON A PLEIADIAN PLANET. THE PLEIADES ACTUALLY CONSISTS OF THIRTY-TWO PLANETS ORBITING SEVEN STARS. AT THAT TIME THERE WERE SIXTEEN DIFFERENT COLONIES OF LYRAEN DESCENT THROUGHOUT THE PLEIADES. THESE COLONISTS ALL WANTED TO OUST THE RENEGADE ATLANS BECAUSE THEY REMAINED INDEPENDENT AND DID NOT ASSIST THEIR HUMAN COUSINS.

THE OTHER TWO GROUPS WERE THE MARTIANS AND MALDEKIANS, WHO WERE ALREADY AT ODDS WITH EACH OTHER. FOR THIS REASON, THE REPTILIANS TURNED THEIR ATTENTION TOWARD THIS SOLAR SYSTEM WITH ITS TWO HUMAN COLONIES. IN THE REPTILIANS ESTIMATION, IT WOULD BE EASY TO DIVIDE AND CONQUER.

THE REPTILIANS LOVE TO USE COMETS AND ASTEROIDS AS WEAPONS AND SHIPS, USING THEM TO TRAVEL THROUGH THE STARS. FIRST, THEY CREATE A SMALL BLACK HOLE AS A PROPULSION SYSTEM THAT PULLS THE LARGER PLANETOID TOWARDS ITS DESTINATION. WHEN USED AS A WEAPON, THEY USE A PARTICLE BEAM ACCELERATOR TO CREATE A BLAST THAT HURLS THE COMET OR ASTEROID TO ITS TARGET. ALL OF THE TECHNOLOGY WAS OBTAINED BY THE BEINGS FROM SIRIUS A.(1)

IN THIS WAY, THEY HURLED A HUGE ICE COMET AIMED AT MARS AND MALDEK. THE REPTILIANS, NOT BEING VERY TECHNOLOGICALLY ORIENTED, MISCALCULATED THE TRAJECTORY. THE PULL OF THE GIGANTIC GAS PLANET, JUPITER, PULLED THE COMET OFF COURSE. THE ICE COMET THEN HEADED DIRECTLY FOR MALDEK. THE CITIZENS OF THAT PLANET ASKED THE MARTIANS FOR HELP. EVEN THOUGH THEY WERE AT ODDS WITH EACH OTHER, THEY ALLOWED SOME OF THE MALDEKIANS TO MOVE TO THE MARTIAN UNDERGROUND. THE COMET CAME SO CLOSE TO MALDEK THAT THE PLANET GOT CAUGHT BETWEEN THE GRAVITATIONAL PULL OF JUPITER, MARS, AND THE

COMET. THIS CAUSED THE PLANET TO EXPLODE, LEAVING AN ASTEROID BELT BETWEEN MARS AND JUPITER. (2)

THE EXPLOSION PUSHED THE ICE COMET CLOSE ENOUGH TO MARS TO RIP THE ATMOSPHERE OFF THAT PLANET, LEAVING ONLY AN EXTREMELY THIN ATMOSPHERE. THE EXPLOSION ALSO PULLED MARS FURTHER AWAY FROM THE SUN.

THE COMET THEN CONTINUED ON TOWARD THE EARTH. THE HEAT OF THE SUN AND THE GRAVITATIONAL PULL BETWEEN THE TWO GLOBES FORCED THE WATERY ATMOSPHERE OF THE EARTH TO POLARIZE. THIS POLARIZATION PULLED MOST OF THE ICE FROM THE COMET TO THE POLAR REGIONS OF THE EARTH, THUS COVERING MOST OPENINGS TO THE INNER EARTH, WHILE AT THE SAME TIME EXPOSING HUGE LAND MASSES FOR THE FIRST TIME.

THE COMET THEN SWITCHED PLACES WITH EARTH, TAKING UP THE SECOND ORBIT FROM THE SUN, BECOMING THE PLANET NOW KNOWN AS VENUS. THE HEAT OF THE SUN MELTED THE ICE ON THE COMET, CREATING A CLOUDY COVERING TO THIS NEW PLANET. THE EARTH WAS PUSHED OUT TO THE THIRD ORBIT OCCUPYING THE PREVIOUS POSITION HELD BY MARS. THE EARTH WAS NOW READY TO BE COLONIZED. MOST OF THE SURVIVING AMPHIBIANS WERE TRANSPORTED TO A NEW HOME ON NEPTUNE. SOME STAYED IN THE NEWLY FORMED OCEANS.

THE REPTILIANS WHO WERE INSIDE THE HOLLOW COMET, NOW VENUS, CAME TO THE SURFACE OF THIS NEW WORLD. THEY BUILT SEVEN DOMED CITIES, ONE FOR EACH OF THE SEVEN GROUPS IN THE HIERARCHY. IN THE MID-1980S ONE OF NEW YORK'S DAILY PAPERS, NEWSDAY, REPORTED THAT A SOVIET SPACE PROBE PENETRATED THE CLOUD LAYER OF VENUS AND PHOTOGRAPHED SEVEN WHITE DOMES THE SIZE OF SMALL CITIES, ALL IN A ROW.

AFTER A PAGE-LONG DIATRIBE, THE AMERICAN SCIENTISTS CONCLUDED THAT THIS WAS ALL A NATURAL FORMATION.

THE REPTILIANS DROVE A LARGE, HOLLOWED OUT OBJECT INTO EARTHS ORBIT TO BEGIN THE COLONIZATION PROCESS. THIS OBJECT IS NOW CALLED THE MOON. CONVENTIONAL SCIENCE CONSIDERS THE MOON NATURAL, YET IT IS THE ONLY KNOWN OBJECT IN SPACE THAT DOES NOT SPIN ON ITS AXIS. THE MOON FACES THE EARTH IN THE SAME POSITION ALL OF THE TIME, LEAVING ONE SIDE IN COMPLETE DARKNESS. A SONIC RESONANCE SENT TO THE SURFACE OF THE MOON MAKES A PINGING NOISE LIKE A HOLLOW OBJECT. IF THE MOON WERE SOLID, THE NOISE WOULD SOUND LIKE A THUMP OR THUD. THE MOON IS HOLLOW. A RECENT ARTICLE IN AN ASTRONOMY MAGAZINE SAID THAT THE MOON WAS BEING RECLASSIFIED BECAUSE IT IS CONSIDERED TO BE HOLLOW.

THE REPTILIANS CHOSE A LARGE CONTINENTAL LANDMASS TO BEGIN THEIR CIVILIZATION ON THE EARTH, NOW REFERRED TO AS LEMURIA OR MU. THIS WAS A VAST AREA IN WHAT IS NOW THE PACIFIC BASIN, EXTENDING FROM JAPAN TO AUSTRALIA, AND FROM THE COAST OF CALIFORNIA TO PERU. THE HAWAIIAN ISLANDS ARE IN THE MIDDLE OF THIS ONE-TIME LANDMASS.

HERE, AN ANDROGYNOUS REPTILIAN CULTURE DEVELOPED. THEY BROUGHT WITH THEM THE CREATURES THAT WERE THEIR SUSTENANCE THE DINOSAURS. ALL BEINGS CREATE BENEATH THEM ANIMALS AND PLANTS THAT ARE A REFLECTION OF THE MIND-PATTERN. REPTILIANS CREATE DINOSAURS, HUMANS CREATE MAMMALS. THEY ARE NOT DESIGNED TO COEXIST ON THE SAME PLANET.

ADDITIONALLY, THE THINKING PROCESS OF THE REPTILIANS DIFFERS FROM THE HUMAN THINKING PROCESS. BECAUSE REPTILIANS DO NOT EVOLVE RAPIDLY AND REMAIN UNCHANGING, THEIR EXPANSION IS ALSO SLOW MOVING AND INSIDIOUS. IT WOULD TAKE SEVERAL MILLENNIA FOR THE REPTILIANS TO DECIDE WHETHER OR NOT THEY WOULD COEXIST WITH HUMANS. AFTER ALL, EARTH WAS STILL AN OUTPOST FAR FROM THE CENTRE OF THE DRACONIAN EMPIRE.

IN THE MEANTIME, THE MARTIANS WERE NOW LIVING UNDERGROUND WITH THEIR HOSTILE MALDEKIAN GUESTS. SOMETHING HAD TO BE DONE QUICKLY TO PREVENT THEM FROM DESTROYING ONE ANOTHER. SO, THE MARTIANS PETITIONED THE GALACTIC FEDERATION TO REMOVE THE MALDEKIAN REFUGEES TO ANOTHER PLANET. THE GALACTIC FEDERATION ALSO RECEIVED A PETITION FROM THE PLEIADIAN COUNCIL AT THE SAME TIME, ASKING THE FEDERATION TO REMOVE THE ATLANS FROM THEIR STAR CLUSTER.

THE FEDERATION THUSLY DECIDED TO USE THE ATLANS AS A COUNTERBALANCE ON EARTH. IF THE ATLANS SURVIVED, THE MALDEKIANS WOULD ALSO BE SENT. THE HUMAN/LYRAEN DESCENDANTS WERE LITERALLY THROWING THEIR OWN RIFFRAFF TO THE REPTILIAN COLONISTS ON EARTH. IN THIS WAY, THE FEDERATION WOULD GET RID OF THEIR UNDESIRABLES. THE UNDESIRABLES WOULD OCCUPY THE ATTENTION OF THE REPTILIANS. THE FEDERATION WOULD GAIN VALUABLE TIME TO BUILD THEIR OWN FORCES AGAINST THE REPTILIANS.

WHEN THE ATLANS ARRIVED ON THE EARTH, THEY COLONIZED WHAT BECAME KNOWN AS ATLANTIS. THEIR CONTINENT STRETCHED FROM WHAT IS NOW THE CARIBBEAN BASIN TO THE AZORES AND CANARY ISLANDS, AS WELL AS SEVERAL SMALL ISLAND CHAINS REACHING UP TO WHAT IS NOW THE EAST COAST OF THE UNITED STATES, INCLUDING MONTAUK POINT.

THE INDUSTRIOUS ATLANTEANS RAPIDLY GREW TO A LARGE, PROSPERING CIVILIZATION NEEDING MORE TERRITORY. THE DINOSAUR POPULATION WAS RAPIDLY INCREASING AND BECOMING DANGEROUS TO THE HUMAN COLONISTS. THE ATLANTEANS BEGAN DESTROYING THE DINOSAURS TO PROTECT THEMSELVES. THIS DID NOT SIT WELL WITH THE REPTILIANS. SOON MAJOR BATTLES OCCURRED ON THE EARTH BETWEEN THE LEMURIAN REPTILIANS AND ATLANTEAN HUMANS.

AT THE SAME TIME, THE MALDEKIAN REFUGEES ARRIVED ON EARTH. THEY CREATED A LARGE HUMAN COLONY IN WHAT IS NOW THE GOBI DESERT, NORTHERN INDIA, SUMER, AND OTHER PARTS OF ASIA.

THE MALDEKIANS ATTACKED THE LUNAR SURFACE WHERE THE REPTILIANS GUARDED THEIR EARTH OUTPOST FROM INVASION. THE MALDEKIANS ALSO BOMBARDED ATLANTIS AND LEMURIA WITH LASER WEAPONS. THE DINOSAURS WERE WIPED OUT.

ADDITIONALLY, THE MARTIANS ALSO ATTACKED THE REPTILIANS FROM SPACE SINCE THEY, TOO, WERE SEARCHING FOR A REPTILIAN-FREE ENVIRONMENT IN WHICH TO LIVE. THIS MIGHT BE CONSIDERED THE REAL FIRST WORLD WAR ON THIS PLANET.

IT WAS A MESS!

FOOTNOTES:

1. THE SIRIANS WERE AT WAR WITH THE ORION SYSTEM. THIS HOSTILITY EXISTS TO THIS DAY. IT IS INTRIGUING SINCE THE BEINGS IN ORION WERE ONCE VERY HUMAN, AS LYRAEN COLONISTS, AND THEN WERE TAKEN OVER BY THE REPTILIANS. HOWEVER, THE SIRIANS AND THE REPTILIANS TRADE WITH EACH OTHER AND THE BEINGS FROM SIRIUS A SELL WEAPONRY TO THE DRACOS! A COMPLEX POLITICAL SITUATION INDEED.

2. THE COMET ALSO CAUSED THE PLANET URANUS TO FLIP ON ITS SIDE. IT IS THE ONLY KNOWN PLANET THAT ROTATES NORTH-SOUTH INSTEAD OF EAST-WEST.

THE LOSS OF THE MARTIAN ATMOSPHERE CAUSED BY DRACO’S PLAYING BILLIARDS BECAME THE "TOTAL RECALL" LEGEND (MOVIE ABOUT MARTIAN HISTORY AND THE OXYGEN WAR). THE DRACO BASES THERE TODAY STILL HAVE NO HESITATION TO SHOOT DOWN A NASA PROBE, ALTHOUGH INCREASINGLY AS THE US GOVERNMENT BECOMES A SATELLITE OF THE SHAPE-SHIFTERS- THEY BEGIN TO LET A FEW EARTH PROBES IN.

GO BACK

5 - CONFLICT & CREATION

TO STOP THE FIGHTING AND MAKE THE EARTH PEACEFUL ENOUGH FOR COLONIZATION, A MEETING WAS HELD BY A COUNCIL FROM THE ANDROMEDA GALAXY, ON A PLANET CALLED HATONA.

THIS MEETING TOOK PLACE OUTSIDE OF THE MILKY WAY GALAXY WITH A NEUTRAL COUNCIL BECAUSE ALL CIVILIZATIONS WITHIN THE MILKY WAY GALAXY WERE IN SOME WAY CONNECTED TO THE FIGHTING, AND ALL HAD SOME SORT OF STAKE IN BELONGING TO THE WINNING SIDE.

INTERESTING PREQUEL TO LETTERS FROM ANDROMEDA FROM ALEX COLLIER (AUTHOR OF "DEFENDING SACRED GROUND").

THE HATONA COUNCIL CONVENED FOR MANY DECADES AS THE FIGHTING CONTINUED IN THIS SOLAR SYSTEM. FINALLY, WITH THEIR INTERCESSION, AN AGREEMENT WAS REACHED BETWEEN SOME OF THE HUMAN FACTIONS AND THE REPTILIAN EARTH COLONISTS. KEEP IN MIND THAT THIS AGREEMENT WAS WITHOUT THE PARTICIPATION OF THE REPTILIANS FROM THE ORIGINAL DRACO EMPIRE.

THE AGREEMENT STATED THAT A NEW BREED OF HUMANITY WOULD BE CREATED ON EARTH THAT WOULD CONTAIN THE DNA OF ALL INTERESTED PARTIES WHO PARTICIPATED IN THE "PEACE" PROCESS. A DESIGNATED AREA ON EARTH WOULD BE SET ASIDE FOR THE CREATION OF THIS NEW SPECIES. THE EARTH-BASED REPTILIANS OF LEMURIA AGREED TO THIS UNDER THE CONDITION THAT THE REPTILIAN BODY BE THE FOUNDATION FOR THIS NEW BEING.

THIS IS WHY THE ORIGINAL BIBLE STATES, "LET US MAKE MAN IN OUR OWN IMAGE." THIS IS A PLURAL STATEMENT BECAUSE IT WAS A GROUP PROJECT.

TO ACHIEVE A NEW SPECIES FROM A REPTILIAN ANDROGYNOUS BODY, IT WAS NECESSARY TO SEPARATE THE GENETICS INTO MALE AND FEMALE COMPONENTS. THIS IS THE ALLEGORICAL STORY IN THE BIBLE OF ADAM AND EVE. CREATING EVE OUT OF THE RIB OF ADAM IS ACTUALLY THE STORY OF SEPARATING THE ANDROGYNOUS REPTILIAN BODY INTO MALE AND FEMALE. THIS IS WHY ALL HUMANS ON THIS PLANET HAVE REPTILIAN DNA WITH REPTILIAN TRAITS. THIS IS ALSO WHY HUMAN FETUSES GO THROUGH REPTILIAN-STYLE DEVELOPMENT IN THE WOMB BEFORE LOOKING HUMANOID.

MANY PROTOTYPES WERE DEVELOPED OVER MILLENNIA. UNDER THE SUPERVISION OF THE HATONA COUNCIL, RACES WERE CREATED AND THEN DESTROYED WHEN IT WAS NOT ACCEPTABLE

BY ALL PARTIES. THIS EXPLAINS WHY ANCESTORS OF MANKIND APPEAR AND THEN SUDDENLY DISAPPEAR IN LAYERS OF ARCHAEOLOGICAL ANALYSIS.

TWELVE HUMANOID, AND ONE REPTILIAN, GROUPS DONATED DNA FOR THIS PURPOSE. MANKIND WAS DEVELOPED IN THE AREA NOW KNOWN AS IRAN/IRAQ, AS WELL AS PARTS OF AFRICA. HYBRIDS WERE ALSO DEVELOPED ON ATLANTIS AND LEMURIA. REMNANTS OF THESE ARE SEEN TODAY AS THE BIGFOOT OR YETI IN NORTH AMERICA AND ASIA; THE ABORIGINES OF AUSTRALIA; AND THE PYGMIES AND WATUSI IN AFRICA.

THE AFRICAN VERSIONS WERE CREATED BY BEINGS FROM A NOMADIC, ARTIFICIAL PLANET KNOWN AS NIBIRU, OR MARDUK. THESE REPTILIAN-LIKE BEINGS TRAVEL IN A MANUFACTURED WORLD LOOPING OUR SOLAR SYSTEM.

THE SUMERIANS CALLED THEM ANUNNAKI.

IT IS HERE, WE WOULD HAVE TO ASSUME, THE FAMILY OF AN, WITH SON'S ENKI / ENLIL (LUCIFER / MICHAEL - ADON / YAHWEH) ARRIVED FROM SIRIUS TECHNO CRAFT, WITH ALL THE NEPHILIM GENETIC DISASTERS WE CALLED THE RETURN OF ENKI. WHILE PEOPLE LIKE SITCHIN CAN ADD USEFUL PUZZLE PIECES LIKE HOW THEIR HOME ATMOSPHERE REPAIR COULD GAIN THE NECESSARY GRAVITY FRACTALITY USING GOLD DEPOSITS, THE SUMERIAN BASICALLY DID NOT GIVE HIM A CLUE AS TO THE BIGGER GALACTIC POLITIC MESS THAN TRIGGERED ENKI'S (ABRAHAM) GENETIC EXPERIMENTS. AFTER REPEATEDLY ENCOUNTERING THE ENKI/ENLIL STORY IN AUSTRALIA (FISH OR DOLPHIN GOD VS SNAKE GOD -LEVITE), IT WAS COOL TO LISTEN TO CRETO MUTWA TELL THE ZULU KILIMANJARO MEANS ENKI MOUNTAIN!

THE COSMIC JOKE TO THIS PROJECT IS THAT ALL OF THE GROUPS DONATING DNA SECRETLY PROGRAMMED SEQUENCES TO CAUSE THEIR GENETIC STRAND TO BE PREDOMINANT. THIS SET THE PRECEDENT FOR ETERNAL CONFLICT. HUMANITY WAS DOOMED TO FIGHT AND BE CONTROLLED. NO ONE GROUP WOULD EVER BE IN CHARGE. THE PROJECT WAS DOOMED FOR FAILURE BEFORE IT EVEN BEGAN!

SUCH DNA PROGRAMMING INVITES TYRANNY AND OPPRESSION. SOUL-PERSONALITIES ATTRACTED TO SUCH A PLANET HAVE A VICTIM MENTALITY. MANY ADVANCED CULTURES CALL EARTH A PRISON PLANET, AND DUMP THEIR CRIMINALS HERE AS PUNISHMENT.

ONCE IN A WHILE, ONE OF THESE SOUL-PERSONALITIES REVEAL THEMSELVES, SUCH AS RICHARD DAHMER, CHARLES MANSON, RICHARD SPECK, AND VLAD THE IMPALER (COUNT DRACULA), TO NAME A FEW.

DUNE WAS A PRISON COLONY, LIKE AUSTRALIA. EARTH FITS THE ARCHETYPE. LIKE IN DUNE HOWEVER, IN EARTH'S GENETIC CAULDRON OF SURVIVAL, THE STRONGEST DNA EMERGES: A 'VACCINE' FOR THE ORION WARS? BUILT BY DESIGN? THE CURRENT EXPRESSION OF THE 2 BROTHERS STILL TOSSING NUCLEAR BOMBS THEN CALLED EAST OF EDEN, TODAY CALLED THE MIDDLE EAST CONFLICT, MAY BE A PERFECT HEALTHY FRACTAL OF THE ORION WARS, WITH THE MAIN WARRING GALACTIC ELEMENTS ALL NICELY STAGED TO WORK OUT A GALACTIC CONFLICT IN A MICRO-COSM.

THIS MAY BE MORE HOWEVER THAN A TEMPEST IN A TEAPOT, BECAUSE THE GALACTIC PLAYERS APPEAR TO BE WILLING TO UP THE STAKES WITH MORE INTERVENTION IF THEIR CHESS PIECE STARTS TO LOSE. ENLIL AS YAHWEH (THE ENCODER IN BIBLE CODE) MAY NOT BE WILLING TO STAND ASIDE AS HIS ISRAEL UNDER THE HEX DRACO FLAG IS DESTROYED BY YET ANOTHER NUCLEAR BLAST. (PREDICTED BY BIBLE CODE AND FATIMA & OTHERS) .

AS WINGMAKERS SO NICELY POINTED OUT - ONLY THE TRON (BST TECHNOLOGY) WITH (IGNITED) DNA MORE TIME TRAVEL CAPABLE CAN DEFEAT SUCH PREDATORY ET'S. THESE GENES WHICH TIME TRAVEL ARE PRECISELY THE JEWEL THAT ALWAYS ELUDES YAHWEH ENLIL'S REPTILE (LEVITE) FAMILY - THEY ARE STUCK IN THE CUBE, GENETIC SOFTWARE ENVIRONMENT OF THE TETRA (HEBREW ALPHABET) WHICH CAN ONLY MAKE A GOLEM.

ISRAEL WAS ORIGINALLY THE GENES OF ENKI WHEN HIS THOTH PTAH LINE AKHANATON, TOOK THE GOLD FROM EGYPT, ASSUMING THE MOSES IDENTITY TO START THE ESSENES. TROUBLE WITH THE 12 TRIBES - GENETIC IMPLOSION SYMMETRY - WAS WHEN THE ORIENTAL (QUEEN DRAC) ASHOKA ASHKANAZI, USURPED THE SEPHARTIC, INSTALLING ENLIL'S CONTROL IN ISRAEL. (ALSO FOUNDING THE BANK OF ENGLAND - DESCRIBED BELOW AS OWNING THE US BANKING).

THE REPTILIANS ENSURED THAT THE NEW MAN WOULD BE FOREVER ATTACHED TO THE REPTILIAN FREQUENCY BECAUSE THE FOUNDATIONAL PROTOTYPE WAS REPTILIAN. THIS MEANT THAT THE NEW MAN COULD EASILY BE MENTALLY CONTROLLED BY THEM.

UPON DISCOVERY OF THIS INFORMATION THAT THE REPTILIANS WANTED CONTROL OF THE NEW RACE, THE ATLANTEANS BEGAN A SEVERE ELECTROMAGNETIC BOMBARDMENT OF LEMURIA. THIS BOMBARDMENT CAUSED THE BULK OF THE CONTINENT TO SUBMERGE INTO THE OCEAN, NOW CALLED THE PACIFIC OCEAN. THE ONLY PARTS LEFT ABOVE THE WATER ARE THE HAWAIIAN ISLANDS, THE CALIFORNIA COAST WEST OF THE SAN ANDREAS FAULT, AUSTRALIA, NEW ZEALAND, THE SOUTH PACIFIC ISLANDS, JAPAN, THE PHILIPPINES, TAIWAN, AND THE ISLANDS OF SOUTHEAST ASIA.

THE REPTILIAN SURVIVORS WENT TO NORTHERN INDIA, THE EARTHS INTERIOR, THE PLANET VENUS, AND PARTS OF CENTRAL AND SOUTH AMERICA. INNER EARTH BECAME THE "HOMELAND" FOR MOST OF THE SURVIVING REPTILIANS OF LEMURIA. HERE, THEY CREATED A VAST UNDERGROUND CIVILIZATION.

THIS STARTED THE LEGENDS OF HELL AND DEMONS LIVING IN FIRE UNDER THE EARTH. THEY BUILT TUBES CONTAINING FAST, SUBWAY-LIKE VEHICLES THAT CAN TRAVEL TO ANY POINT ON THE EARTH WITHIN A FEW HOURS. THEY CREATED THE FAMED UNDERWORLD CITIES OF AKKADIA, AGARTHA, HYPERBOREA, AND SHAMBALLA THAT ARE SOUGHT BY EXPLORERS TO THIS VERY DAY. THESE CITIES ARE BUILT ALONG THE INSIDE WALL OF THE INNER CRUST THAT LINES THE INTERIOR OF THE EARTH.

REMEMBER, THE HOLLOW EARTH IS NOT A THEORY, BUT A SCIENTIFIC FACT CAUSED BY THE COOLING AND SPINNING OF A PLANET AS IT IS EJECTED FROM A STAR OR SUN.

SHALL WE CHECK TO SEE IF MEREDITH LADY YOUNG, AUTHOR OF "AGARTHA" (BOOK ABOUT THE CIVILIZATION OF THE INNER EARTH) AGREES?

THE PRIMARY ENTRY POINTS TO THE INNER EARTH ARE VIA THE NORTH POLE, WHERE THERE IS AN OPENING OF 1300 MILES, AND THE SOUTH POLE, WITH AN OPENING OF 950 MILES. THESE CAN BE SEEN FROM SPACE. THAT IS WHY COMMERCIAL AIRCRAFT ARE NOT ALLOWED TO FLY OVER THESE AREAS; NOT BECAUSE OF MAGNETIC DISTURBANCES, WHICH IS THE "OFFICIAL" REASON. ADMIRAL BYRD REPORTED ON THESE OPENINGS IN THE 1920S UNTIL HIS INFORMATION WAS CONCEALED BY THE GOVERNMENT.

AT THE VERY CENTRE, OR NUCLEUS, OF INNER EARTH, THERE IS A GLOBE OF ENERGY LEFT OVER FROM THE CREATION OF THIS PLANET THAT ACTS AS AN INNER SUN. IT IS THE LIGHT FROM THIS OBJECT SUSPENDED BY GRAVITY AND CENTRIFUGAL FORCE THAT CAUSES THE LIGHT OF THE AURORA BOREALIS.

NUMEROUS CAVE ENTRANCES TO THE INNER EARTH EXIST IN THE ROCKY MOUNTAINS AND SIERRA MOUNTAINS IN THE WESTERN UNITED STATES, AS WELL AS LESS NUMEROUS OPENINGS IN THE OZARKS AND APPALACHIAN MOUNTAINS. ENTRIES ALSO EXIST IN THE ALPS, HIMALAYAS, ANDES, AND THE CARIBBEAN. THERE ARE ALSO NUMEROUS SUB-OCEANIC ENTRY POINTS, PARTICULARLY IN THE DEEP TRENCHES OF THE PACIFIC OCEAN, THE CARIBBEAN SEA, AND THE ATLANTIC SUBMARINE MOUNTAIN RANGES ESPECIALLY ON OR NEAR THE AZORES, CANARY ISLANDS, AND THE FALKLANDS.

ALL OF THESE AREAS ARE CLOSELY GUARDED BY LOCAL GOVERNMENTS AND N.W.O. ELITE FORCES. ARTIFICIALLY CREATED ENTRANCES EXIST UNDER THE NEW DENVER AIRPORT, THE GIZA PLATEAU IN EGYPT, MAJOR AIR FORCE COMPLEXES AROUND THE WORLD, AND MANY OF THE TEMPLES IN INDIA AND CHINA. A MAJOR CHINESE ENTRY POINT IS UNDER THE SHENSI PYRAMID THAT IS OUT OF BOUNDS FOR EVERYONE IN WESTERN CHINA.

OF COURSE, WITH THE REPTILIANS OFF THE EARTHS SURFACE, THE ATLANTEANS WERE FREE TO PLAY WITH THIS NEW MANKIND AND ESTABLISH HUMANS AS THE RULER OF THIS PLANET. THEY ESTABLISHED COLONIES ALL OVER THE REMAINING PORTIONS OF LAND. THEY INVITED THE SIRIANS TO COME AND PLAY WITH THEM. THEY BOOTED THE MARDUK BEINGS OFF-WORLD AND TOOK CONTROL OF THEIR SLAVES. THEY CREATED NEW HYBRIDS FOR SEA AND LAND, ONE OF WHICH BECAME THE MERFOLK, A GENETIC BLEND OF HUMAN AND DOLPHIN. THE DOLPHIN SPECIES WAS BROUGHT HERE FROM THE ANDROMEDA GALAXY TO MONITOR ALL OF THESE EVENTS.

WHENEVER THE ATLANTEANS DETECTED UNDERGROUND REPTILIAN ACTIVITY, THEY BLASTED THE INNER EARTH WITH LASERS AND ELECTROMAGNETIC PULSES TO KILL THEM. UNFORTUNATELY, THIS WEAKENED THE UPPER CRUST OF THE EARTHS TOP MANTLE RIDING OVER THE TRAPPED MAGMA BETWEEN THE UPPER AND LOWER CRUSTS. AFTER SEVERAL MILLENNIA OF THESE ATTACKS, THE ATLANTEAN CONTINENT STARTED TO BREAK UP. THEIR CIVILIZATION BEGAN TO BREAK UP AS A REFLECTION OF THE PHYSICAL DETERIORATION OF THEIR CONTINENT. THE ATLANTEANS BECAME EVEN MORE BELLIGERENT AS FEAR AND DESTRUCTION OVERTOOK THEIR MIND-PATTERNS. BLACK MAGICIANS AND SORCERERS TOOK THE PLACE OF SCIENTISTS AND RELIGIOUS LEADERS.

FORTUNATELY, THE POPULATION FORESAW THE DESTRUCTION THAT WAS COMING. MANY REFUGEES RELOCATED TO WHAT IS NOW EGYPT, PERU, THE APPALACHIAN MOUNTAINS AND WESTERN EUROPE, JUST BEFORE THE CONTINENT COLLAPSED INTO THE UPPER CRUST OF THE EARTH. THIS COLLAPSE CAUSED THE EARTH TO FLIP ON ITS AXIS, CREATING THE LEGEND OF THE FLOOD WRITTEN ABOUT IN THE BIBLE, AND IN OTHER WORLD CULTURES.

THIS CATASTROPHE WAS USED AS A WINDOW OF OPPORTUNITY BY THE GROUPS THAT DONATED DNA TO CREATE MANKIND. THEY IMMEDIATELY BEGAN REORGANIZING THE HUMANS INTO NEW GROUPS THAT BECAME THE BASIS FOR FUTURE NATIONALISM.

*

THE SIRIANS HELPED TO CREATE THE ANCIENT EGYPTIAN CULTURE
*

THOSE FROM TAU CETI ORGANIZED THE SLAVIC CULTURE
*

THE RIGELIANS WERE BUSY IN CHINA AND THE ORIENT (3)

WHILE ALL THIS WAS TRANSPIRING, THE REPTILIANS SAW AN OPPORTUNITY AND SEIZED IT!

FOOTNOTE:

3. FOR FURTHER DETAILS OF THE FLOW OF GENETICS AN ALIEN INTERVENTION IN MANKIND, PLEASE REFER TO THE MILKY WAY GALAXY CHART ON THE BOOK.

GO BACK

6 - BLUEBLOODS!

THE INNER EARTH PROVIDED A SUBTERRANEAN LOCALE FOR THE REPTILIANS TO REGROUP AND FORMULATE PLANS TO RETAKE THE SURFACE. AT THIS POINT, THE REPTILIANS WERE COMPLETELY CUT OFF FROM THEIR HOME IN THE DRACO CONSTELLATION. THEIR SPACESHIP, THE MOON, WAS IN HUMAN HANDS. THEY WERE ALONE, ISOLATED ON A NOW HOSTILE PLANET. THEY NEEDED TO DEFEND THEMSELVES.

THEY DEVELOPED A PLAN TO INSIDIOUSLY RETAKE THE SURFACE BY BLENDING THEIR GENETICS WITH THE GENETICS OF THE SURFACE HUMANS. BECAUSE THE HUMAN PROTOTYPE ALREADY HAD REPTILIAN GENETICS, IT WAS EASY TO ACCESS THE MIND-PATTERN. THE REPTILIAN FREQUENCY WAS ALREADY ESTABLISHED IN THE BRAIN STEM AS WELL AS THE REPTILIAN BRAIN SECTION OF THESE HYBRID HUMANS.

THE POPULATION OF SUMER WAS CHOSEN AS THE STARTING POINT. THESE HUMANS WERE PRIMARILY DESCENDANTS OF THE MARTIAN, MALDEKIAN, AND LYRAEN REFUGEES. THE REPTILIANS HAVE A PREFERENCE FOR THE GENETICS OF BLONDE-HAIRED, BLUE-EYED PEOPLE WHOSE MIND-PATTERNS AND GENETICS ARE SO EASILY CONTROLLED. THEY ABDUCTED MEMBERS OF THE RULING CLASSES, INCLUDING POLITICAL LEADERS.

USING THESE HUMANS, THEY BEGAN A NEW HYBRIDIZATION PROGRAM THAT TOOK SEVERAL GENERATIONS TO PERFECT. THEIR GOAL WAS TO REACH A HUMAN/ REPTILIAN GENETIC 50/50 SPLIT. THIS WOULD PRODUCE A HUMAN-LOOKING REPTILIAN THAT COULD EASILY SHAPESHIFT FROM REPTILIAN TO HUMAN, THEN BACK AGAIN. SHAPESHIFTING WAS ACCOMPLISHED SIMPLY BY CONCENTRATING ON THE GENETICS THE HYBRID WISHED TO OPEN, OR LOCK UP, WHATEVER THE CASE MAY BE.

FOR THIS PROGRAM THE REPTILIANS ENGAGED THE HELP OF THE SIRIANS WHO HAD THE TECHNOLOGY TO IMPLEMENT SUCH A PROGRAM. THE SIRIANS KNEW A LOT ABOUT GENETIC ALTERATIONS AND MIND-PROGRAMMING, WHICH THEY FREELY SHARED WITH THE REPTILIANS.

ONCE THE HYBRIDIZATION PROGRAM WAS COMPLETE, THE SUMERIAN LEADERS WERE NOW

SHAPESHIFTING REPTILIANS. THE NEW REPTILIAN HYBRID BECAME THE ELITE OF THAT CULTURE. THEIR BLOOD, BECAUSE OF THE INCREASED REPTILIAN DNA, CONTAINED MORE OF A COPPER CONTENT.

SINCE COPPER-BASED BLOOD TURNS BLUE-GREEN UPON OXIDIZING, THESE REPTILIAN HYBRIDS WERE CALLED "BLUEBLOODS".

AMERICAN GOVERNMENT IS PROBABLY THE MAIN FOCUS OF THE CURRENT REPTILIAN SHAPESHIFTERS. YOUR ATTENTION IS DRAWN TO THE VERTICAL SLIT EYEBALLS WHICH SHAPE SHIFT IN TO MANY US MILITARY LEADERS - NORMAN RUSSBACHER, RUPERT MURDOCK, ETC. PICTURES: WERE AT WIOLAWA.COM WE SUGGESTED NEW RULES REQUIRING ALL GOVERNMENT DECISIONS AND VOTING MUST TAKE PLACE IN 20% OXYGEN AIR, TO KILL THE SHAPESHIFTERS.

THE BLUEBLOODS QUICKLY REALIZED THAT WITH A 50/50 HUMAN/REPTILIAN GENETIC SPLIT, IT WAS NECESSARY TO INTERMARRY TO MAINTAIN THE 50/50 SPLIT BLOODLINE NECESSARY TO SHAPESHIFT. WHEN THE SPLIT INCREASED TOO FAR TO THE REPTILIAN SIDE, SHAPESHIFTING BECAME DIFFICULT, AND HOLDING HUMAN FORM BECAME IMPOSSIBLE. IN THESE CASES, IT WAS DISCOVERED THAT THE INGESTION OF HUMAN HORMONES, FLESH, AND BLOOD, ALLOWED THE REPTILIANS TO MAINTAIN THE HUMAN FORM.

HUMAN FORM WAS NECESSARY TO MAINTAIN TO AVOID SCARING THE POPULATION, WHICH WAS NOW NOT ACCUSTOMED TO THE REPTILIAN FORM.

CONTROL OF THE MASSES WAS EASIER WHEN THE ORDERS CAME FROM A HUMANOID. THE REPTILIAN FORMAT WAS KEPT TO RELIGIOUS ICONS AND LEGENDS. THE STATUES OF THEIR GODS AND GODDESSES REFLECT THE REPTILIAN INFLUENCE, EVEN SHOWING A FEMALE REPTILIAN HOLDING A HYBRID BABY. (CLICK IMAGE RIGHT)

THE SHAPESHIFTING REPTILIAN BLUEBLOODS ASKED THE SIRIANS FOR HELP WITH THE DAILY MAINTENANCE OF THEIR HUMAN FORMS. THE SIRIANS DETERMINED THAT FEEDING THE HYBRIDS HUMAN HORMONES AND BLOOD IN AN ALTERED ANIMAL FORM WOULD BE THE EASIEST WAY TO DO IT UNNOTICED BY THE POPULATION.

THE SACRIFICIAL ANIMAL USED BY MOST MIDDLE EASTERN PEOPLE WAS THE WILD BOAR, SO THE SIRIANS CHOSE IT AS THE BASIS FOR THIS NEW ANIMAL HYBRID. HUMAN GENETICS WERE MIXED WITH THOSE OF THE WILD BOAR TO CREATE THE DOMESTICATED PIG. THIS ANIMAL WAS SERVED DAILY TO THE BLUEBLOODS AS A METHOD OF TEMPORARILY MAINTAINING THEIR HUMAN FORM UNTIL THEY COULD USE AN ACTUAL HUMAN IN A SACRIFICIAL CEREMONY.

BECAUSE THE DOMESTICATED PIG IS A COMBINATION OF HUMAN AND ANIMAL GENETICS, EATING IT IS A FORM OF CANNIBALISM. THIS EXPLAINS WHY THE HEBREWS CONSIDERED IT UNCLEAN TO EAT. THIS IS ALSO WHY THE PIG IS CONSIDERED TO BE THE MOST INTELLIGENT ANIMAL ON EARTH, WHY PIG SKIN CAN BE GRAFTED DIRECTLY ONTO HUMANS IN BURN CASES, AND WHY PIG HEART VALVES CAN BE USED IN HUMANS WITH LITTLE DIFFICULTY. CANCER DRUGS AND OTHER CHEMICALS ARE OFTEN TESTED ON PIGS BEFORE HUMANS.

THE DOMESTICATED PIG FREQUENCY, OR GROUP MIND, IS THE PERFECT VEHICLE FOR ANIMAL SPECIES TO ENTER BEFORE ENTERING HUMAN FORM ON THEIR EVOLUTIONARY PROGRESSION. IN MANY RESPECTS, PIGS CAN BE CONSIDERED A FORM OF HUMANITY. TO A LESSER DEGREE, THE SAME IS TRUE ABOUT CATS.

AS TIME PROGRESSED, THE CIVILIZATION OF SUMER DECLINED AND TRANSFORMED INTO OTHER CULTURES. VAST MIGRATIONS FROM SUMER TO OTHER LOCATIONS IN CENTRAL ASIA OCCURRED. THE MIGRATING PEOPLES TOOK THEIR BLUEBLOOD LEADERS WITH THEM, AS THEY WERE THEIR ROYALTY AND KINGS.

THE SUMERIANS BECAME KNOWN AS THE SUM-ARYANS, OR JUST, ARYANS. THEY SPREAD OUT ACROSS ASIA INTO THE STEPPES OF RUSSIA AND INTO THE NORTHERN INDIAN SUBCONTINENT. IN INDIA THEY ENCOUNTERED THE DARK-SKINNED DRAVIDIANS, WHO WERE REPTILIAN REMNANTS FROM LEMURIA. THE DRAVIDIANS WERE DRIVEN TO THE CENTRAL AND SOUTHERN PARTS OF INDIA, WHILE THE ARYAN HYBRIDS TOOK CONTROL OF THE NORTH, AND INTO THE FOOTHILLS OF THE HIMALAYAS.

THE ARYAN LEADERS, ALL BLUEBLOODS, BECAME THE SULTANS AND RAJAS OF LEGEND AND HISTORY. SUMERIANS ALSO CREATED BABYLONIA.

THE SUMERIANS ALSO MIGRATED TO THE AREA KNOWN AS THE CAUCASUS REGION, WHERE THE KHAZARS DEVELOPED. FROM THE CAUCASUS REGION, THE BLUEBLOOD KINGS AND THEIR PEOPLE SPREAD WEST TOWARD EUROPE, DEVELOPING INTO THE FRANKS, CAMBRIANS, AND TEUTONIC NATIONALITIES. THESE NATIONALITIES WERE ALSO BEING MANIPULATED BY VARIOUS ALIEN CULTURES LIKE THE ANTARIANS, ARCTURIANS, ALDEBARANS, TAU CETIANS, AND OTHER REMNANTS OF THE LYRAENS, SUCH AS THE ATLANS. THE ATLANS LOCATED HERE EVENTUALLY BECAME THE CELTS.

TO BACK-TRACK FOR JUST A BIT, I HAD SAID IN A PREVIOUS CHAPTER THAT THE DESCENDANTS OF THE REPTILIAN HYBRID SUMERIANS WENT INTO CENTRAL ASIA AND THE MIDDLE EAST. THEY MOSTLY ESTABLISHED THEMSELVES IN THE CAUCASUS MOUNTAINS AND BECAME THE KHAZARS. FROM HERE, THEY SPREAD WEST TOWARD EUROPE, SEEDING THE NATIONAL IDENTITIES FOR THE VIKINGS, THE FRANKS, THE TEUTONIC PEOPLES, AND THE RUSSIANS. KEEP IN MIND THAT WHEN ATLANTIS SANK, SOME OF THOSE REFUGEES WENT TO WESTERN EUROPE AND DEVELOPED INTO THE CELTS. SOME WENT TO GREECE AND OTHERS TO THE ITALIAN PENINSULA.

THESE PEOPLES WERE HERE BEFORE THE HYBRIDS MOVED IN. IT WAS DURING THE INTERIM TIME PERIOD FROM THE DESTRUCTION OF ATLANTIS UNTIL THE SUMERIAN DESCENDANTS MOVED IN THAT OTHER ALIEN GROUPS STARTED TO ADD THEIR GENETIC MIX TO THE POT AND DEVELOP INDIVIDUAL CULTURES BASED ON THEIR HOME WORLDS.

THESE BLUEBLOOD LEADERS ALSO INFILTRATED THE MIDDLE EASTERN PEOPLES, SUCH AS THE BIBLICAL CANAANITES, MALACHITES, AND KITTITES.

AT THE SAME TIME IN EGYPT, THE SIRIANS WERE REORGANIZING THE ATLANTEAN DESCENDANTS THERE, KNOWN AS THE PHOENICIANS. THE PHOENICIANS WERE BLONDE-HAIRED, BLUE-EYED, WITH SOME GREEN-EYED, RED-HAIRED PEOPLE AMONG THEM. THE PHOENICIANS COLONIZED THE COASTAL MIDDLE EAST AND THE BRITISH ISLES. THEY EVEN COLONIZED PARTS OF THE NORTH-EASTERN NORTH AMERICAN CONTINENT, ALL THE WAY TO THE GREAT LAKES AREA. SOME OF THEIR MINES AND WRITING SON STONE TABLETS CAN STILL BE FOUND IN THE WOODS OF NORTH AMERICA.

THE SIRIANS WERE ALSO GENETICALLY CREATING THE ANCIENT HEBREWS. THE JEWISH PEOPLE ARE ACTUALLY A COMBINATION OF THESE GENETICALLY MANIPULATED HEBREWS AND THE SUMERIANS. THESE JEWISH PEOPLE WERE THEN RELEASED INTO THE PALESTINIAN TERRITORY. THE NAME, PALESTINE, COMES FROM THE ANCIENT PEOPLE, THE PHILISTINES, WHO WERE ACTUALLY PHOENICIANS.

ALL OF THESE MIXED IN THE COASTAL PLAINS OF PALESTINE AND CREATED A NEW RELIGION BASED ON SACRIFICE AND AN AVENGING ALIEN CONTROLLER, THAT THEY CALLED GOD, OR ELOHIM.

SIMILARLY, WHEN THE ARYANS MIXED WITH THE DRAVIDIANS IN INDIA, THEY CREATED THE HINDU RELIGION, WHICH IS ACTUALLY A RECREATION OF THE REPTILIAN SEVEN-TIER HIERARCHY. THE CASTE SYSTEM OF INDIA IS A DIRECT COPY OF THE REPTILIAN DIVISION OF FUNCTION.

AT THE SAME TIME THAT ALL OF THIS WAS GOING ON IN WESTERN AND CENTRAL ASIA, THE RIGELIANS, WERE DEVELOPING THE REMNANTS OF LEMURIA WHO ESCAPED TO THE COAST OF EASTERN ASIA. THE RIGELIANS WERE A HUMAN CIVILIZATION THAT WAS CONTROLLED, AND EVENTUALLY ASSIMILATED, BY THE REPTILIANS. THE RIGELIANS ASSISTED THE INNER EARTH REPTILIANS IN DEVELOPING A HYBRID THAT INCLUDED RIGELIAN DNA.

THE RIGELIAN/REPTILIAN HYBRIDS SET UP DYNASTIES IN WHAT IS NOW JAPAN AND CHINA THAT DEVELOPED INDEPENDENTLY OF THEIR WESTERN COUSINS.

IN THEIR MANIA FOR CONTROL, THE REPTILIANS USED THE VARIOUS RACES THAT DONATED DNA TO THE ORIGINAL HUMAN PROJECT. THEY FASTIDIOUSLY MONITORED THESE RELATED SECTIONS OF HYBRIDS TO DETERMINE WHICH WAS BEST SUITED FOR OVERALL CONTROL, AND WHICH FOR SUBSERVIENCE. ALL THE HYBRIDS COULD BE CONTROLLED THROUGH THE REPTILIAN BRAIN THAT HOOKED THEM INTO REPTILIAN MIND-PATTERNS, BUT SOME WERE MORE CONTROLLABLE THAN OTHERS.

IN EUROPE, THE BLUEBLOODS INSIDIOUSLY TOOK CONTROL OF THE VARIOUS TRIBES AND GROUPS, BECOMING THEIR KINGS AND ROYALTY. THEY INFILTRATED THE ARCTURIAN EXPERIMENT, CALLED THE ETRUSCANS AND STARTED TO CREATE A NEW GLOBAL EMPIRE THROUGH THE ROMANS. THESE EUROPEAN BLUEBLOODS THEN ENTIRELY ELIMINATED THE ANTARIAN EXPERIMENT IN GREECE, AND INSTIGATED THEIR PLAN FOR GLOBALIZATION THROUGH THE ROMAN EMPIRE.

THE REPTILIANS EVEN OFFENDED THE SIRIANS BY INFILTRATING THE EGYPTIAN EXPERIMENT AND IMPLEMENTING THEIR RELIGION THERE.

THE REPTILIAN HYBRIDS BECAME LIKE THE ENDOMETRIOSIS OF THE KNOWN WORLD, SLOWLY GROWING INTO ALL AREAS AND CREATING CONTROL THROUGH THE BLUEBLOOD SYSTEM.

GO BACK

GO BACK TO THE BLACK NOBILITY

7 - OTHER ALIEN GROUPS

WHILE THE REPTILIANS WERE THE FIRST COLONISTS ON EARTH, THEY WERE NOT THE ONLY ONES WHO INTERFERED WITH HUMAN DEVELOPMENT ON THIS PLANET. IN ALL THERE ARE TWELVE OTHER GROUPS WHO DONATED DNA TO THE PRODUCTION OF THE EXPERIMENT.

ADD THE REPTILIANS TO THE TWELVE GROUPS, RESULTING IN HUMAN BEINGS WITH A GENETIC MIXTURE OF 13 DIFFERENT STRAINS.

THE RESULT WAS A GENERAL FREE-FOR-ALL. WHILE ALL THESE HUMANOID ALIENS WERE LYRAEN/REPTILIAN DESCENDANTS, EACH GROUP WAS CULTURALLY AND PHYSICALLY MANIPULATED BY DIFFERENT GROUPS. THIS IS SIMILAR TO A LAB PROFESSOR WHO LEAVES THE DOOR OPEN WHILE HE IS GONE, AND ALL HIS ASSISTANTS ADD THEIR OWN GENETICS TO THE EXPERIMENT.

THE TAU CETI ALIENS CENTERED THEIR ATTENTION ON THE EASTERN EUROPEAN AREA, FROM WHAT IS NOW SERBIA TO THE SLOPES OF THE URAL MOUNTAINS. FROM HERE, THEY INFLUENCED THE SLAVIC AND RUSSIAN PEOPLES. THE GEOGRAPHIC CONDITIONS RESEMBLED THAT IN THE TAU CETI STAR SYSTEM, AND ITS COLONY EPSILON ERIDANUS. THE TAU CETI ADDED THEIR DNA TO THE HUMAN PROTOTYPES THAT WERE ALREADY ESTABLISHED THERE, CREATING WHAT IS NOW KNOWN AS THE SLAVIC PEOPLES. THE RESULTS WERE A RACE OF HUMANS WHO WERE STOCKY, BARREL-CHESTED, AND AVERAGING 5’6’’ TO 5’9’’ IN HEIGHT, WITH A DENSE BONE STRUCTURE AND DARK EYES. THEY WERE AGGRESSIVE, AND PREFERRED A COLD CLIMATE.

THESE TAU CETI/HUMANS WERE VIRULENTLY AGAINST THE GREY ALIEN RACE AND THE REPTILIANS, BECAUSE THEIR WORLDS HAD BEEN ATTACKED, AND THEIR CHILDREN STOLEN AND KILLED BY BOTH RACES. THE TAU CETIANS VOWED TO FOLLOW THE GREY RACE AND DESTROY THEM.

IN THE 1950S, THE SOVIET UNION SIGNED AN AGREEMENT WITH THE TAU CETIANS TO USE BASES IN SIBERIA AND UNDER THE URAL MOUNTAINS. FOR THIS REASON, THE CITY OF SVERDLOVSK, NAMED AFTER MY GREAT-UNCLE, THE FIRST PRESIDENT OF THE SOVIET UNION, WAS CLOSED TO OUTSIDERS (CLICK IMAGE LEFT).

MANY EXPERIMENTS INVOLVING RADIATION ON THE PUBLIC WERE PERFORMED HERE FROM 1958 THROUGH THE 1980S. A UNITED STATES SPY PLANE WAS SHOT DOWN OVER SVERDLOVSK IN THE EARLY 1960S WHEN THE UNITED STATES WAS TRYING TO LEARN ABOUT THE SECRET ACTIVITIES TAKING PLACE THERE.

IN CENTRAL EUROPE, THE GERMAN TRIBES WERE GENETICALLY MANIPULATED BY BEINGS FROM ALDEBARAN. THESE PEOPLE ARE VERY INTELLIGENT AND SCIENTIFICALLY-ORIENTED. THEY ARE GENERALLY BLONDE-HAIRED AND BLUE-EYED, WITH A MINORITY OF DARK-HAIRED, LIGHT-BROWN TO HAZEL-EYED PEOPLE. THEY ARE MILITARISTIC, AND PREFER TO KEEP TO THEMSELVES. FOR ALMOST 2,000 YEARS, THE ALDEBARANS HAVE BEEN ENERGETICALLY CONNECTING TO THE GERMANIC PEOPLES, TELEPATHICALLY SENDING INFORMATION TO THEM AND PROMOTING A NATIONAL SENTIMENT.

MANY HUMANS OF THE ALDEBARAN FREQUENCY HAVE MIXED WITH THE TAU CETI DESCENDANTS IN THE SLAVIC AREA, PARTICULARLY IN POLAND AND RUSSIA. HITLER KNEW THIS. THAT IS WHY HE WAS SO ADAMANT ABOUT INVADING THOSE COUNTRIES AND INCORPORATING THEM INTO HIS EMPIRE.

HITLER WAS ONLY HALF GERMANIC. HIS FATHER WAS A WEALTHY JEWISH BUSINESSMAN IN AUSTRIA. HIS MOTHER WORKED AT THE HOME AS A MAID. SHE HAD AN AFFAIR WITH THE MASTER OF THE HOUSE, AND WHEN THE WIFE FOUND OUT ABOUT IT, HAD THE MAID THROWN OUT. THE JEWISH BUSINESSMAN DID NOTHING TO HELP HITLER’S MOTHER. FOR THIS REASON, HITLER HATED THE JEWS AND SOUGHT TO DESTROY HIS OWN GENETIC LINEAGE, BASICALLY BECAUSE HE HATED HIMSELF. HE WAS ALSO DEEPLY MIND-CONTROLLED.

THE ALDEBARANS ALSO GENETICALLY INFUSED THE VIKINGS. THESE NORDIC PEOPLE INHERITED THE AGGRESSIVE AND MILITARISTIC TENDENCIES THAT ARE ALSO SEEN IN THE GERMANS. THE VIKINGS PLUNDERED AND RAPED ACROSS EUROPE FOR CENTURIES, BUT DID NOT HAVE THE

TECHNOLOGICAL ABILITY TO STAY IN POWER.

AN ACCIDENTAL MANIPULATION OF GENETICS OCCURRED ON THE ITALIAN PENINSULA 3000 YEARS AGO. A SHIP FROM THE ARCTURUS STAR SYSTEM CRASHED LANDED ON ETRUSCAN TERRITORY. THESE HUMANS WERE ACTUALLY EXTREMELY SPIRITUALLY-MINDED, AND INSTEAD OF TRYING TO GET BACK HOME, STAYED AND BLENDED INTO THE HUMANS OF THAT PART OF EARTH. THEIR DESCENDANTS BECAME THE ROMANS WHO WERE THEN INFILTRATED AND MIXED WITH THE CENTRAL ASIAN HYBRIDS.

BEINGS FROM THE ANTARIAN STAR SYSTEM WERE BEHIND THE GENETIC MANIPULATION OF ANCIENT GREECE. THESE PEOPLE WERE A SOCIETY PREDOMINANTLY BASED UPON HOMOSEXUALITY. FEMALES WERE USED FOR BREEDING ONLY. IN FACT, THERE WERE ANTARIAN OBSERVERS AT THE MONTAUK PROJECT WHO WERE INTERESTED IN THE PROGRAMMING ASPECTS OF SEXUALITY AS THEY RELATED TO THE WILHELM REICH METHODS.

THE ANTARIANS ARE DARK, OFTEN WITH OLIVE SKIN, DARK EYES, AND SHORT, THIN BODIES. THEY HAVE A FABULOUS MUSCULATURE DUE TO THE DENSITY OF THEIR HOME-WORLD, AND ARE KNOWN FOR THEIR BODY-BUILDING GOALS.

THE GREEK-ANTARIANS COLONIZED SPAIN AND PORTUGAL. THEIR DESCENDANTS FURTHER MIXED WITH THE ROMANS, AND ARABS WHO ARE PREDOMINANTLY SUMERIAN/REPTILIAN. THESE THEN COLONIZED CENTRAL AND SOUTH AMERICA, MIXING THEIR GENETICS WITH THE NATIVE INDIANS WHO WERE OF ATLANTEAN-PROCYON DESCENT.

THE PROCYON STAR SYSTEM DOES NOT HAVE MUCH TECHNOLOGY. THE PROCYONS WERE BROUGHT TO THIS PLANET AFTER THE FALL OF ATLANTIS TO BOOST THE SURVIVAL RATE OF THE REFUGEES. THEY BECAME THE MAYA, AZTEC, AND INCA. THEY WERE GIVEN ANCIENT LEMURIAN AND ATLANTEAN OUTPOSTS IN THE ANDES AND SIERRAS OF MEXICO. THEY TRIED UNSUCCESSFULLY TO RECREATE THESE CULTURES, INCLUDING EMULATING THE BUILDING OF THE PYRAMIDS, PERFORMING MEDICAL PROCEDURES, AND FINALLY, SACRIFICING TO THE REPTILIAN GODS. THIS IS WHY THEIR LEGENDS SPEAK OF BLOND MEN RETURNING IN CHARIOTS FROM SPACE TO TAKE THEM AWAY.

THE ANASAZI INDIANS OF THE AMERICAN SOUTHWEST WERE ALSO BROUGHT FROM PROCYON. IT WAS THE SIRIANS WHO SO GENEROUSLY PROVIDED TRANSPORTATION. THE SIRIANS EVEN ATTEMPTED TO BRING THE HEBREWS TO THE AMERICAN WEST. ANCIENT HEBREW COINS WERE FOUND IN NEW MEXICO AND OTHER PARTS OF NORTH AND SOUTH AMERICA.

OVER THE MILLENNIA AND RECENT CENTURIES, MOVEMENTS OF NATIONS, COLONIZATIONS, WAR, AND FAMINES, HAVE THROWN THE EARTHS POPULATION INTO A GIANT MELTING POT. STRANDS OF GENETICS HAVE CONTINUOUSLY MIXED WITH ONE ANOTHER, ESPECIALLY IN NORTH AMERICA, EUROPE, AUSTRALIA, THE CARIBBEAN, AND SOUTH AMERICA. THE RESULT IS THAT THERE ARE FEW PURE RACES LEFT, AND THE RACIAL AND CULTURAL UNITY MAKES IT EASIER FOR GROUP CONTROL.

WHILE ALL OF THIS WAS GOING ON IN EUROPE AND THE MIDDLE EAST, THE CHINESE EMPIRE WAS EXPANDING OVER EAST ASIA, AND THE DRAVIDIAN-REPTILIAN CULTURE IN INDIA WAS BEING REPLACED BY THE ARYAN (ARI- AND SUM-ARIAN) HORDES FROM CENTRAL ASIA. IN SOUTH AMERICA, THE INCA EMPIRE WAS FLOURISHING AS IT WAS MIXED WITH THE GENETICS OF THE PROCYON STAR SYSTEM.

THE SAME MIXTURE WAS HAPPENING IN NORTH AND CENTRAL AMERICA CREATING THE TOLTECS, MAYANS, AND AZTECS. ALL OF THESE CULTURES, USED BLOOD-RITUAL AND HUMAN SACRIFICE. THIS INDICATES THAT THE PROCYONIANS WERE THEMSELVES CONQUERED BY THE REPTILIANS AND DID THEIR BIDDING, EVEN THOUGH THEY WERE HUMANOID. ALL OF THE CENTRAL AND SOUTH AMERICAN CULTURES USED SNAKES AND REPTILIANS AS SYMBOLS.

THESE PEOPLE HAVE A UNIQUE BLEND OF LEMURIAN/DRACO AND ATLANTEAN/HUMAN GENETICS MIXED WITH THE PROCYONIAN DNA.

GO BACK

GO BACK TO VIDA EN NUESTRA GALAXIA

8 - THE CRYSTAL SKULL

WHEN THE HATONA COUNCIL CONVENED TO DETERMINE THE DEVELOPMENT OF LIFE ON EARTH, THEY PONDERED TWO QUESTIONS. FIRST, IF THE EARTH BEINGS WERE LEFT TO DEVELOP ON THEIR OWN, HOW WOULD THEY KNOW THEIR TRUE ORIGINS, AND SECOND, IF THERE WAS NO INTERFERENCE, HOW COULD THEY BE PRODDED IN THE CORRECT DIRECTION.

ONE E.T. GROUP THAT WAS PRIMARILY NON-PHYSICAL DECIDED TO LEAVE A REPOSITORY OF KNOWLEDGE FOR THOSE WHO EVOLVED ENOUGH TO UNDERSTAND IT. THESE MUSCULAR BEINGS WERE TALL WITH GOLDEN-BRONZE SKIN, GOLDEN HAIR, AND VIOLET EYES. ON AN ETHERIC LEVEL, THIS E.T. GROUP CREATED AN OBJECT THAT CONTAINED WITHIN IT THE SUM TOTAL OF KNOWLEDGE OF THE MIND OF GOD AS THEY KNEW IT. THEY ALSO PROGRAMMED THIS OBJECT WITH THE HISTORY OF THE UNIVERSE AND ALL THE TECHNOLOGY THAT WOULD EVER BE NECESSARY.

FOR THE OBJECT THIS GROUP CHOSE THE SHAPE OF A FEMALE HUMAN SKULL WITHOUT ANY RACIAL FEATURES. THE SKULL REPRESENTED ALL HUMANOIDS, SYMBOLIZING BROTHERHOOD AND HARMONY. THE FEMALE WAS CHOSEN BECAUSE IT WAS TO BE PLACED IN PHYSICAL REALITY AS A SYMBOL OF EGO BEING OVERCOME.

CRYSTAL WAS CHOSEN BECAUSE IT REPRESENTS THE HIGHEST VIBRATION POSSIBLE IN PHYSICAL REALITY - PURITY, CLARITY, FOCUSING, AND MAGNIFICATION. THE MOVABLE JAW PIECE (CLICK IMAGE RIGHT) SYMBOLIZED THE FACT THAT IT WAS A COMMUNICATION DEVICE. THE E.T. GROUP LEFT THE CRYSTAL SKULL WITH THE FIRST LYRAEN/ ATLANTEAN CIVILIZATION WHERE IT WAS PLACED IN A TEMPLE PYRAMID TO BE ENERGIZED BY THE LYRAEN/ATLANTEANS OVER MANY AEONS OF TIME.

WHEN THE SIRIANS INFILTRATED THE SECOND-GENERATION LYRAEN/ATLANTEAN CIVILIZATION, THEY NEGOTIATED WITH THE ATLANS TO STUDY THE CRYSTAL SKULL. OVER TIME, THEY CREATED AN EXACT DUPLICATE THAT THEY BROUGHT BACK TO SIRIUS A. OTHER ALIEN GROUPS MADE INFERIOR COPIES OF IT FOR USE WITH THEIR OWN HUMAN CREATIONS. BY THE TIME THE THIRD-GENERATION LYRAEN/ATLANTEANS WERE WELL ESTABLISHED, THE TRUE PURPOSE OF THE CRYSTAL SKULL WAS ALMOST ENTIRELY FORGOTTEN.

THOSE IN POWER TRIED TO USE IT FOR NEGATIVE PURPOSES, NOT REALIZING THAT THIS MAGNIFIED AND REFLECTED BACK ALL THEIR EVIL DEEDS AND INTENTIONS. IN ADDITION, THE CRYSTAL SKULL WAS CREATED IN SUCH A WAY THAT WHATEVER IS THOUGHT IN ITS PRESENCE REFLECTS BACK, BECOMING A PART OF THE THINKER’S EXPERIENCES. THE CRYSTAL SKULL TEACHES THAT THE PHYSICAL UNIVERSE MIRRORS THOUGHTS.

WHEN ATLANTIS SANK, HIGH PRIESTS FLEEING THE CONTINENT BROUGHT IT TO CENTRAL AMERICA

TO WHERE THE PROCYONIANS BROUGHT THE MAYA. HERE, IT WAS USED AS AN OBJECT OF WORSHIP AND REVERENCE UNTIL THE MAYA WERE REMOVED FROM THE EARTH. EVENTUALLY, THE CRYSTAL SKULL REMAINED BURIED IN RUINS UNTIL ITS DISCOVERY IN THE EARLY PART OF THE 20TH CENTURY, WHEN THE CRYSTAL SKULL ALLOWED ITSELF TO BE DISCOVERED.

THE CRYSTAL SKULL OPERATES THROUGH THE TRINITY OF COMMUNICATION COLOR, TONE, AND ARCHETYPE. WHEN ANY COMBINATION OF THESE THREE IS BEAMED OR THOUGHT OF IN THE PRESENCE OF THE CRYSTAL SKULL, IT OPENS A PROGRAM THAT IS CODED TO A PARTICULAR FREQUENCY RESONANCE.

AN INFINITE NUMBER OF COMBINATIONS CAN BE USED, AND ANY ONE WILL UNLOCK A PROGRAM IN THE CRYSTAL SKULL TO TEACH MANKIND. THE LEFT-BRAIN REPRESENTS LANGUAGE, AND THE RIGHT-BRAIN REPRESENTS PURE THOUGHT. THE PINEAL GLAND OF COMMUNICATION BALANCES AND TRANSLATES THE LEFT- AND RIGHT- BRAINS WITH THE USE OF ARCHETYPES.

ARCHETYPES CAN BE GEOMETRIC SHAPES, LETTERS, NUMBERS, ANCIENT HEBREW SYMBOLS, PICTOGRAMS, OR ANY COMBINATION OF THESE. COLORS ARE ALSO PART OF THE TRIAD WITH A LANGUAGE ALL THEIR OWN. THE LEFT-BRAIN IS DARK, THE RIGHT-BRAIN IS LIGHT, AND ONCE AGAIN, THE PINEAL GLAND BALANCES AND TRANSLATES THIS THROUGH COLORS. TONES ALSO REPRESENT THE TRINITY OF SOUND, BALANCING MUSIC AND SILENCE. IN THE SAME WAY, THE CRYSTAL SKULL BALANCES THE MIND OF GOD AND PHYSICAL REALITY.

AT TIMES, THE CRYSTAL SKULL BECOMES NON-PHYSICAL. BECAUSE IT WAS CREATED WITHOUT A BODY, THIS SYMBOLIZES THE NEEDLESSNESS OF THE PHYSICAL BODY. THE CRYSTAL SKULL IS A BRIDGE BETWEEN ALL LEVELS OF REALITY.

ANYONE WHO KNOWS THE SEQUENCES OF THE TRINITY OF COMMUNICATION BECOMES ALL-POWERFUL AND OMNISCIENT.

GO BACK

GO BACK TO CRANEOS INSOLITOS

9 - THE ANCIENT HEBREWS

MOST MODERN JEWS HAVE ABSOLUTELY NO GENETIC LINK TO THE MIDDLE EAST. THERE ARE IN FACT, MANY DIFFERENT PHYSICAL TYPES OF JEWS, COVERING MANY RACIAL CHARACTERISTICS. THIS SUPPORTS THE FACT THAT THEY ARE NOT A HOMOGENEOUS GROUP, BUT A RELIGIOUS GROUP THAT SPANS MANY CULTURES. THE STORY OF ABRAHAM GOING OUT FROM THE CITY OF UR AND COMING TO CANAAN IS REALLY TELLING THE STORY OF THE REPTILIAN HYBRIDS LEAVING

SUMERIAN TERRITORY AND COLONIZING OTHER PARTS OF CENTRAL ASIA AND THE MIDDLE EAST. THE VAST MAJORITY OF EUROPEAN/AMERICAN JEWS CAN TRACE THEIR GENETIC LINEAGE TO THE KHAZARS, MENTIONED IN THE LAST CHAPTER, WHO ALL CONVERTED TO JUDAISM IN THE 800S TO CIRCUMVENT THE CATHOLIC RULE OF THE HOLY ROMAN EMPIRE.

IN THE YEAR 2000, THE UNIVERSITY OF PAVIA IN ITALY DID A GENETIC STUDY OF EUROPEAN MEN. THEY FOUND THAT 80% OF THEM HAD A DIRECT LINEAGE TO CENTRAL ASIA AND THE OTHER 20% TO THE MIDDLE EAST. THIS SUPPORTS THE STATEMENT THAT THE SUMERIANS ENTERED INTO CENTRAL ASIA AND THEN MIGRATED TO EUROPE AND THE MIDDLE EAST. THIS ALSO NULLIFIES THE THEORY OF AFRICA AS THE BIRTHPLACE OF MANKIND. THERE IS ABSOLUTELY NO GENETIC CONNECTION FROM ASIA OR EUROPE TO AFRICA.

THE ANCIENT HEBREWS HAVE NOTHING TO DO WITH MODERN JEWS. AS MENTIONED IN THE PREVIOUS CHAPTER, THE HEBREWS WERE SIRIAN-CREATED IN EGYPT BY COMBINING SIRIAN AND LYRAEN GENETICS. THESE PEOPLE WERE TALL AND POWERFUL, AND SPOKE THE SIRIAN LANGUAGE WHICH IS THE EQUIVALENT TO THE ANCIENT HEBREW LANGUAGE. SCHOLARS AGREE THAT THE HEBREW LANGUAGE SUDDENLY APPEARED ON THE SCENE.

THE ANCIENT PEOPLES OF PALESTINE SPOKE ARAMAIC, WHICH WAS THE PARENT LANGUAGE OF ARABIC, FARSI, AND SEVERAL OTHER MIDDLE EASTERN DIALECTS. ORIGINALLY, HEBREW WAS A LANGUAGE USED EXCLUSIVELY BY THE PRIESTHOOD AND THE EGYPTIAN SECRET SOCIETY. EVENTUALLY, ANCIENT HEBREW BEGAN TO MIX WITH ARAMAIC, AMONG OTHER LANGUAGES.

THE HEBREWS WERE ACTUALLY PAID WORKERS IN EGYPT. THEY WENT TO CANAAN TO ASSIMILATE THE NATIVE CULTURES FOR THE EGYPTIAN EMPIRE. THEY MIXED WITH THE LOCAL TRIBES OF SUMERIAN-HYBRID DESCENDANTS, PRACTICING BLOOD-RITUAL AND HUMAN SACRIFICED. ALL OF THIS WAS INCORPORATED INTO A CONGLOMERATE RELIGION BASED ON ANCIENT EGYPTIAN/ATLANTEAN/ SIRIAN BELIEFS. THAT IS HOW JUDAISM WAS BORN.

THE STORY OF THE EXODUS OF THE JEWS FROM EGYPT IS A POOR RETELLING OF THE DESTRUCTION OF THE ISLAND OF SANTORINI IN THE EASTERN MEDITERRANEAN SEA BY VOLCANIC EXPLOSION. THE LAVA DISPERSED INTO THE SEA CAUSING IT TO TURN BLOOD-RED. THE VOLCANIC ASH AND ROCK THAT SPEWED FORTH CREATED THE LEGEND OF THE PLAGUES ON EGYPT. MANY CITIZENS FLED THE AREA. THE RED SEA PARTED WHEN THE GROUND UNDER THE WATER ROSE UP, EXPOSING LAND THAT COULD THEN BE CROSSED. SEVERAL HOURS LATER, IT SANK BACK DOWN, DROWNING WHOEVER WAS STILL TRYING TO CROSS.

THE ACTUAL TRANSLATION OF HOW THE TEN COMMANDMENTS, OR THE BIBLE CODES, WERE TRANSMITTED TO THE PEOPLE OF THE EXODUS, STATES THAT THE PEOPLE SPONTANEOUSLY STARTED SPEAKING THE INSTRUCTIONS. THIS IS A DEMONSTRATION OF MIND-TO-MIND COMMUNICATION. IN THIS CASE, PROGRAMMING WAS ELECTROMAGNETICALLY ACTIVATED, REVEALING DNA INSTRUCTIONS THAT WERE THEN WRITTEN DOWN. THESE INSTRUCTIONS WERE DESIGNED TO KEEP THE EXPERIMENT/PROJECT IN LINE.

THE MIDDLE EAST BECAME A FOCAL POINT FOR THE SIRIANS AND THEIR REPTILIAN ALLIES. TOGETHER, THEY GENERATED A NEW VERSION OF RELIGION AND CULTURE FOR EASE OF GLOBAL CONTROL AND DOMINATION. WHAT BETTER WAY TO DO THAT THAN BY PROGRAMMING A RACE OF NOMADIC PEOPLE THAT WOULD CARRY THE RELIGION AND CULTURE EVERYWHERE ON THE EARTH.

WHENEVER THE CONDITIONS WITHIN THE EXPERIMENT BECAME UNMANAGEABLE, THE PARTICULAR OVERSEER GROUP MADE CORRECTIONS. FOR EXAMPLE, MOST PEOPLE ARE FAMILIAR WITH THE STORY OF THE CITIES OF SODOM AND GOMORRAH, WHERE A GROUP OF HUMANS IN THE MANKIND EXPERIMENT WENT AWRY. THEIR SEXUAL HABITS WERE NOT CONDUCIVE TO PROPAGATION SINCE THEY WERE PREDOMINANTLY HOMOSEXUAL.

A VIRUS WAS INTRODUCED INTO THE POPULATION TO DESTROY IT. HOWEVER, IT STARTED TO SPREAD OUTSIDE THE IMMEDIATE REGION. THIS DISTURBED THE CONTROLLERS, SO THEY SENT

TWO AGENTS TO INVESTIGATE, AND TO DETERMINE IF ANYONE WAS SALVAGEABLE. THE LOCAL INHABITANTS CALLED THEM ANGELS, BECAUSE OF THEIR BLONDE HAIR, BLUE EYES, AND PERFECT BODIES.

LOT AND HIS WIFE TOOK THE AGENTS INTO THEIR HOME TO PROTECT THEM, SINCE ALL THE MEN IN TOWN WANTED TO HAVE SEX WITH THEM. SOON AFTER, SODOM AND GOMORRAH WERE DESTROYED. TO THIS DAY, RADIATION CAN BE DETECTED AT THE SITE. MELTED ROCK IS SEEN IN THE CANYON WALLS. SODOM AND GOMORRAH WERE ELIMINATED WITH A NUCLEAR WEAPON.

IN THE EARLY 1960S, ISRAELI SCIENTIST DIGGING AT THE SITE OF SODOM FOUND REMNANTS OF BONE AND BODY TISSUE ENCASED IN ROCK. AMERICAN SCIENTISTS WERE CALLED IN BECAUSE THEY HAD BETTER EQUIPMENT. WHAT THEY FOUND IN THE PRESERVED BODY TISSUE WAS A SAMPLE OF THE VIRUS. THE AMERICANS RECONSTITUTED IT WITH LIVING CELLULAR NUCLEI. QUIETLY IT WAS TESTED ON A TERMINALLY ILL PATIENT IN A ST LOUIS HOSPITAL IN 1967. THE MAN DIED A HORRIBLE DEATH. THE VIRUS WAS CALLED AIDS BECAUSE IT REMOVED WHATEVER IMMUNE SYSTEM THE MAN HAD LEFT. THE AMERICAN RESEARCH DETERMINED THAT THIS THOUSANDS-OF-YEARS-OLD VIRUS WAS ARTIFICIALLY GENETICALLY CREATED.

THE ANCIENT HEBREW RELIGION WAS A CONGLOMERATION OF THE SUMERIAN-REPTILIAN BELIEF SYSTEM, ANCIENT EGYPTIAN-ATLANTEAN-LYRAEN BELIEF SYSTEM, AND THE REPTILIAN TRIBAL SUBCULTURES FOUND IN THE LAND OF CANAAN THAT WERE INCORPORATED INTO ONE ACCEPTABLE DOGMA. IT INCLUDED BLOOD-RITUAL AND HUMAN SACRIFICE, WHICH IS WHY ABRAHAM WAS SO EAGER TO PLEASE GOD BY KILLING HIS OWN SON ON AN ALTAR. THIS SYMBOLISM CAME IN HANDY LATER ON FOR RELIGIOUS PURPOSES.

THE HEBREW LANGUAGE ITSELF DID NOT APPEAR UNTIL AFTER THE HEBREWS LEFT EGYPT, AND THEN ONLY FOR RELIGIOUS CEREMONIAL PURPOSES BY THE PRIESTS. IN FACT, MANY OF THE TERMS AND NAMES USED IN THE HEBREW RELIGION CAME FROM THE EGYPTIAN CEREMONIAL RITES. THE NAME, MOSES, ORIGINATES FROM THE TITLE GIVEN TO ADEPTS OF THE SECRET EGYPTIAN PYRAMID CULT.

THIS EGYPTIAN CEREMONIAL TITLE, MOSHE, MEANS "HE WHO IS ANOINTED WITH CROCODILE FAT FROM THE NILE RIVER."

HERE, WE SUGGEST SWERDLOW DOES NOT DEAL WELL - WITH THE ACADEMIC EVIDENCE MOSES WAS AKHANATON - REF: OUT OF EGYPT , HOUSE OF THE MESSIAH , BOOKS ETC. BUT AT LEAST HE PERCEIVES CORRECTLY THAT THE BEST HEBREW RITES WERE MOSTLY DESCENDED FROM EGYPT.

THE ONGOING CONFLICT BETWEEN ENLIL’S AMUN RA PRIESTS, VERSUS AKHENATON’S ATUN (ENKI) PROBABLY REFLECTS THE ENLIL MORE DRAC REPTILIAN PERSUASION, VERSUS ENKI - THE REBELLIOUS SON FROM SIRIUS, PUSHING THE POSSIBILITY THAT A (HIS?) GENETIC EXPERIMENT HERE COULD PRODUCE A BLOOD LINE RECOVERY FROM THE CATASTROPHIC NEPHILIM DNA CONDITION.

WITH THE BENEFICIAL RESULT TO THE ADVANCED DRACO’S OF NEW ACCESS TO A BLOOD LINE THAT COULD RESUME STAR NAVIGATING - WHERE THE BIG SALARIES ARE MADE- BECAUSE OF RENEWED GRAVITY MAKING POWER IN IMPLODING IGNITED DNA.

THIS CEREMONY WAS PERFORMED INSIDE THE KINGS CHAMBER IN THE GREAT PYRAMID.

IN HEBREW IT MEANT "OUT FROM THE WATER", IN REFERENCE TO THE CROCODILES. THE EGYPTIAN WORD FOR CROCODILE FAT WAS MESSEH FROM WHICH THE TERM MESSIAH IS DERIVED. TO BE ANOINTED WITH CROCODILE FAT RENDERED INTO AN OIL, WAS TO ABSORB POWERFUL, UNCHANGING REPTILIAN ENERGY AND ASSIMILATE THIS INTO THE BODY. MOSHE, OR MOSES, WAS A RITUAL TITLE, NOT A PERSONS NAME AND WAS A REFERENCE TO THE CROCODILE FAT USED IN TEMPLE AND PYRAMID CEREMONIES.

A LEADER SUCH AS A MOSES WAS A LOGICAL CHOICE TO SPEARHEAD A MIGRATION OF PEOPLE TO SAFETY FROM A PLACE OF TURMOIL, SUCH AS WHEN THE DESTRUCTION OF THE ISLAND OF SANTORINI, ALSO KNOWN AS THERA, BLEW UP AND CREATED HAVOC ON EGYPTIAN TERRITORY, CAUSING MANY TO FLEE. IT WOULD ALSO MAKE SENSE FOR THE SIRIAN CONTROLLERS OF EGYPT TO SEND OUT THEIR GENETIC CREATIONS, THE HEBREWS, TO A PLACE OF SAFETY IN THE SINAI DESERT LIKE A SAFEKEEPING UNTIL NEEDED.

IT WAS THEN THAT THE TORAH, OR HOLY COMMANDMENTS, WAS GIVEN TO THIS PEOPLE AS A CODE BY WHICH TO LIVE. THIS DOCUMENT IS IN A CODE THAT CAN ONLY NOW BE DECIPHERED WITH THE USE OF A COMPUTER. THE LAND OF CANAAN NEEDED TO BE CONQUERED IN ORDER TO GIVE DEVELOPMENT ROOM TO THE NEW CREATIONS. IN EFFECT, THE HEBREWS WERE THE NEXT SEQUENCE OF THE EXPERIMENT, AND THE OLD VERSIONS WERE TO BE ELIMINATED OR ASSIMILATED.

THE HEBREWS CARRIED THE NEW CODED LAWS IN THE ARK OF THE COVENANT (COVEN OF ANTS), THAT ONLY THE PRIESTS COULD TOUCH. WHEN THE UNIVERSITY OF MINNESOTA IN THE 1960S ATTEMPTED TO BUILD THE ARK AS DESCRIBED WITH INSTRUCTIONS IN THE BIBLE, IT WAS SO ELECTRICALLY CHARGED THAT IT HAD TO BE DISMANTLED!

THIS IS WHY THE ANCIENT HEBREW PRIESTS ENTERED INTO THE ARK AREA ALONE, FOR SAFETY REASONS. THEY ALSO HAD TO WEAR WHITE LINEN, AS NON-CONDUCTIVE CLOTHING, AND WORE A SPECIAL BREASTPLATE THAT GROUNDED THEM ELECTRICALLY AND ACTED AS A PROTECTIVE SHIELD. THE ARK WAS REALLY A COMMUNICATION DEVICE. THIS IS WHY THE PRIESTS HAD MESSAGES FROM GOD WHENEVER THEY WERE IN THE ARKS PRESENCE.

THE ARK ALSO ACTED AS A LOCATION DEVICE FOR BEACONS ORIGINATING OUTSIDE OF THIS PLANET. THE ARK WAS ORIGINALLY KEPT IN THE GREAT PYRAMID TO FOCALIZE ENERGIES. IT WAS SENT WITH THE HEBREWS FOR SAFEKEEPING. IT HAS SINCE TRAVELED FROM EGYPT TO CANAAN/ISRAEL TO ETHIOPIA, BACK TO ISRAEL, AND NOW IS LOCATED UNDER THE PYRAMID AGAIN.

HOWEVER, THERE ARE ACTUALLY TWO ARKS.

THE SECOND ONE IS IN JERUSALEM.

WE FEEL THIS INFORMATION IS INCOMPLETE, IN THAT CLEARLY THE ARK WAS A FRACTALLY ARRANGED (ORGONE) CAPACITOR - WHOSE CHARGE COMPRESSION COULD CONTAIN RADIOACTIVITY. AND THAT FEEDING THE CHARGE OF THIS DEVICE IS THE REASON IT MUST BE LOCATED ON A MAJOR LEY LINE - TODAY IT APPEARS- PRECISELY ON THE MICHAEL LINE - ST

MICHAEL'S (ENLIL) ISLAND CAVE IN THE AZORES - IMAGE ABOVE LEFT (ASSURAS = ASSA URU)- WHERE JACQUE COUSTEAU'S SON LOST HIS LIFE FOR GETTING TOO CLOSE.

GO BACK

GO BACK TO KHAZAR - THE THIRTEENTH TRIBE

GO BACK TO THE PROTOCOLS OF THE LEARNED ELDERS OF ZION

10 - THE ANUNNAKI & THE BLACK RACE

THE CREATORS OF THE BLACK RACE, THE REPTILIAN BEINGS OF MARDUK KNOWN AS THE ANUNNAKI WHOSE PLANETOID TRAVELS IN AN ELLIPTICAL ORBIT OF THIS SOLAR SYSTEM EVERY FEW THOUSAND YEARS, ARE DUE TO ARRIVE IN 2003.

IT IS NOW 2004 AND REASONABLE TO SURMISE THAT MASSIVE ARRIVAL/INVASION IS IMMANENT. WE RECOMMEND ISOLATION IN COMMUNITY, HI GROUND, SHELTER FROM HI WIND, GOOD WATER, SEEDS, AND BEING TRANSPARENT TO GOVERNMENT. GOOD SHAMAN LUCID DREAMING SKILL TO SET UP BOUNDARY.

THE ANUNNAKI LIKE TO DEVELOP SLAVE RACES FOR SPECIFIC PURPOSES, CREATING THEM OUT OF LOWER ANIMAL FORMS. IN THE CASE OF THEIR EARTH CREATIONS, THEY USED SIMIANS TO CREATE A SLAVE RACE TO WORK THE MINES IN WHAT IS NOW KNOWN AS AFRICA. THIS RACE WAS GENETICALLY DESIGNED TO WORK AND LIVE IN HOT, HUMID CONDITIONS WITH A LIMITED LIFE SPAN. DISEASES WERE PROGRAMMED INTO THEM TO PREVENT DEVELOPMENT OF CULTURE, AND CAUSE THEM TO BECOME DEPENDENT ON THEIR CREATORS FOR EXISTENCE. THIS IS WHY IN THE GENETIC MEMORY OF ALL OTHER RACES, THE BLACKS ARE SLAVES.

THE AMERICAN AND EUROPEAN ILLUMINATI DECIDED TO USE THE AIDS VIRUS AS A WEAPON TO REMOVE THE BLACK RACE FROM EARTH. THESE ILLUMINATI CONCLUDED THAT IF THE BLACKS WERE NO LONGER HERE, THE ANUNNAKI WOULD FIND THE CONDITION OF THEIR SLAVES UNACCEPTABLE, AND NOT INTERFERE WITH ILLUMINATI PLANS. THE GREEN MONKEYS OF EAST AFRICA WERE SUBSEQUENTLY INFECTED WITH THE AIDS VIRUS, ALONG WITH THE BLOOD SUPPLY OF HAITI. IN AFRICA, AIDS IS A HETEROSEXUAL DISEASE.

THE WORLD HEALTH ORGANIZATION, WHO, WAS ALSO USED TO DIRECTLY INJECT THE AIDS VIRUS INTO THE POPULATION UNDER THE GUISE OF IMMUNIZATION.

"THE 1971 FLOWCHART MAKES IT PERFECTLY CLEAR, THE DESIGN, INTENT AND PURPOSE OF THE U.S. SPECIAL VIRUS PROGRAM. AS DR. PETER PIOT, EXECUTIVE DIRECTOR OF UNAIDS SAYS, THE HIV/AIDS VIRUS IS THE RESULT OF MANY STEPS IN THE LABORATORY, IT WAS NO ACCIDENT. THE 1971 FLOWCHART PROVIDES ABSOLUTE EVIDENCE OF THE UNITED STATES' INTENT TO KILL ITS OWN CITIZENS AND OTHERS." DR BOYD GRAVES

HOWEVER, A MONKEY WRENCH TURNED UP IN THE PLANS. A GAY FRENCH-CANADIAN FLIGHT ATTENDANT HAD SEX WITH A BISEXUAL PARTNER IN AFRICA. THE INFECTED FLIGHT ATTENDANT ANGRILY SPREAD THE DISEASE WHEREVER HE WENT. NEW YORK CITY AND SAN FRANCISCO WERE TWO OF HIS PRIMARY DESTINATIONS. HE ALONE IS DETERMINED TO HAVE BEEN THE MAIN CARRIER OUTSIDE OF THE INTENDED TARGET. THE ILLUMINATI DOES NOT THINK OF EVERYTHING. (5)

BY 2003 WHEN MARDUK RETURNS, MOST OF BLACK AFRICA WILL BE INFECTED. THOSE WHO ARE NOT INFECTED WILL BE KILLED VIA EBOLA, WAR, OR FAMINE.

MOST CARIBBEAN AND NORTH AMERICAN BLACKS HAVE A HIGH PERCENTAGE OF EUROPEAN GENETICS THAT ARE NOT CONSIDERED DESIRABLE FOR SLAVE PURPOSES.

IN 1999, NEWSPAPERS REPORTED THAT US ASTRONOMERS SAID THAT A HUGE PLANET HAD BEEN DETECTED BEYOND PLUTO WITH AN ELLIPTICAL ORBIT OPPOSITE TO ALL THE OTHER PLANETS IN THIS SOLAR SYSTEM. THEY SAID IT WAS ON A TRAJECTORY THAT WOULD TAKE IT TO NEAR THE EARTH IN 2003!

NOT ANOTHER WORD WAS PRINTED ABOUT IT AFTER THAT.

QUOTES FROM

PLANET X, COMETS AND WHAT NASA IS HIDING

(EXPLORING THE PARALLEL TO THE INTERVENTIONIST COMET THEORY)

AN INTERVIEW WITH JAMES MCCANNEY, BY RICK MARTIN

"... THE OTHER FACTOR... IN 1991 WHAT HARRINGTON SAW WAS TWO THINGS: HE SAW HALE-BOPP, AND HE SAW SOMETHING MUCH BIGGER BEYOND HALE BOPP. THAT'S PLANET-X. .. IN 1991, HALE-BOPP WAS ON A NEAR DIRECT COLLISION COURSE WITH EARTH. WITH A COUPLE OF QUICK PHOTOGRAPHS THEY COULD CHART THE ORBIT, AND IT WAS ON A NEW COLLISION COURSE WITH EARTH.

MARTIN :(INTERVIEWER): NO WONDER THERE WAS SUCH A SCRAMBLE.

MCCANNEY: IT WAS A HUGE SCRAMBLE. WHEN IT WAS DISCOVERED, I CALLED UP GODDARD SPACE CENTRE- I KNEW THE SECRETARY THERE- AND I SAID: 'WHAT'S GOING ON?' I HEARD THERE'S THIS COMET...' YOU COULD HEAR THE SCREAMING IN THE BACKGROUND. AND SHE SAID: "OH MY GOD, THIS COMET IS HUGE!". BUT I THOUGHT SHE MEANT IN THE SENSE OF BEING A NEWS STORY. NO, IT WAS HUGE IN THE SENSE THAT IS A PLANET-SIZED OBJECT. THEY HAD BEEN TRACKING IT.

YOU SEE, THIS IS WHERE THE DIVISION COMES IN, BECAUSE IT WASN'T UNTIL THEN THAT EVEN A LOT OF THE SCIENTISTS AT GODDARD FOUND OUT ABOUT IT. BUT THEY HAD BEEN TRACKING THIS SINCE 1991, POSSIBLY EARLIER. HARRINGTON DISCOVERED IT, AND YOU SEE IT IN THE 1991 MEMO THAT HE KNEW EXACTLY WHERE TO GO AND LOOK AT IT.

A LONG TIME AGO, WHAT HAPPENED IS... IT WAS ONE OF THE COMPANIONS OF NIBURU THAT DID THE DAMAGE TO EARTH.
MARTIN: A COMPANION?

MCCANNEY: A COMPANION. IT WAS THE ONE THAT BECAME VENUS. VELIKOVSKY WAS VERY RIGHT THAT VENUS WAS A HUGE COMET THAT WORKED ITS WAY THROUGH THE SOLAR SYSTEM, AND IT TOOK ABOUT A 600 YEAR PERIOD FROM THE TIME IT WAS CAPTURED BY JUPITER TO THE TIME IT ENCOUNTERED EARTH. AND THEN WORKED ITS WAY IN TO BECOME THE PLANET (VENUS) THAT WE KNOW TODAY.

SO, ORIGINALLY, WHAT HAPPENED WAS THAT HALE-BOPP WAS HERE ABOUT 4,200 YEARS AGO. THEY WERE, LITERALLY, SMALLER COMPANIONS TO NIBURU. THAT'S WHY THEY DIDN'T WANT ANYBODY TO KNOW ABOUT THE COMPANION, BECAUSE THEY KNEW IT WAS ON A COLLISION COURSE WITH EARTH, AND THEY KNEW IT WAS THE COMPANION TO THE BIGGER ONE THAT CAUSED THE PROBLEM. BUT THEY DIDN'T REALIZE THAT HALE-BOPP WAS, LITERALLY ONE OF THE COMPANIONS ITSELF.

NOW, WHEN THE DESTROYER, THE BIG GUY, NIBURU, COMES IN, IT HAS AN ENTIRE ENTOURAGE OF THESE THINGS.

MARTIN: I GUESS COMENT NEAT WOULD BE ONE OF THOSE?

MCCANNEY: AND THAT'S THE THING. WHEN WE GOT BARRAGED A FEW WEEKS AGO BY ALL THESE COMETS, AND THEY NEVER ANNOUNCED COMET NEAT,C-2002/VI. CLEARLY ALL OF THIS STUFF IS COMING FROM THE SOUTHERN HEMISPHERE.

THEN OF COURSE, HARRINGTON KNEW VERY WELL WHERE THAT WAS, THE REASONS THAT I GAVE; THEY WERE 'PULLING DOWN' ON THE PLANETS URANUS AND NEPTUNE. IT'S INTERESTING TO NOTE THAT WHEN THE STORY OF HARRINGTON CAME OUT, THE GOVERNMENT TRIED TO MAKE A STATEMENT THROUGH SOME OF THESE ASTRONOMERS THAT ON THE RADIO, THE DISINFORMATION GUYS, WHO CAME UP WITH A STORY: "OH WELL, WE HAVE CORRECTED THE MASSES OF THOSE PLANETS DUE TO NEW INFORMATION, SO THAT HAS TAKEN CARE OF THE PROBLEM"

WELL, NO, THAT DOESN'T CORRECT ANYTHING, WHEN YOU SEE THESE PLANETS BEING 'PULLED DOWN'. THAT WOULD ONLY CORRECT THINGS IN THE PLANE OF THE PLANETS. THIS OBJECT WAS BIG ENOUGH, BACK IN 1991, THAT IT WAS PULLING URANUS AND NEPTUNE DOWN OUT OF THEIR ORBITS. THAT'S HOW BIG THIS THING IS!

SO, YOU SEE THE CONCERN OVER THE COMPANION. BECAUSE THEY ALL KNOW, AND THE VATICAN KNOWS, THAT IT WAS THE COMPANION THAT DID THE DAMAGE THE LAST TIME. THE ONLY PROBLEM IS, THE COMPANION BECAME THE PLANET VENUS. WHAT THEY DON'T UNDERSTAND IS THAT IT'S A VERY DIFFERENT THING TO PRODUCE THE ORBITS FOR THESE, AND NASA IS LEARNING THE HARD WAY. THEY COULDN'T KEEP TRACK OF HALE-BOPP; IT CHANGED ON A DAILY BASIS. THAT WAS THE ONE THING WE DID IN THE MILLENIUM GROUP: TRACK THE DAILING CHANGING OF ITS ORBIT ON THE GOVERNMENT EPHEMERIS PAGES.

MARTIN: WAS COMET NEAT A SURPRISE? DID THAT COME OUT OF NOWHERE- OR DID THEY EXPECT

THAT?

MCCANNEY: NO. COMET NEAT IS ANOTHER VERY LARGE NUCLEUS; PLANETARY IN SIZE- PROBABLY THE SIZE OF OUR MOON, AT LEAST; PROBABLY LARGER. THEY PROBABLY SAW IT COMING IN YEARS AGO, AS PART OF THIS ENTOURAGE OF THINGS COMING IN -- WHICH I THINK OF AS THINGS THAT ARE COMING IN AS PART OF THE PLANET X ENTOURAGE. THEY DIDN'T WANT ANYBODY TO KNOW ABOUT IT, FOR THE SIMPLE REASON THEY KNEW IT WAS GOING TO COME IN RIGHT AROUND THE SUN AND IT WAS BIG. THEY PROBABLY NEVER EXPECTED IT TO BECOME AS BRIGHT AS IT DID. BUT IT WAS LITERALLY VISIBLE IN THE DAYTIME SKY, RIGHT NEXT TO THE SUN, AS IT PASSED - OVER ABOUT A 12 HOUR PERIOD WHEN IT WAS COMING IN.
MARTIN: THE OBVIOUS QUESTION IS: HOW MANY MORE OF THESE COMPANIONS CAN WE LOOK FORWARD TO ?

MCCANNEY: WE DON'T KNOW. OF COURSE THOUGH, NASA WOULD HAVE A VERY GOOD KNOWLEDGE OF THIS. THE OTHER IMPORTANT THING THAT I WANTED TO SAY EARLIER ABOUT HALE-BOPP IS THAT IN THE SIX YEAR PERIOD FROM 1991 TO 1996, WHERE IT ACTUALLY HIT PERIHELION WITH THE SUN, IT LOST THREE MONTHS TIME IN ARRIVING DUE TO THE TAIL-DRAG OF THE HUGE COMET. THAT'S WHY WE DIDN'T HAVE A DIRECT COLLISION WITH IT. AND WHEN I SAY "DIRECT COLLISION", I DON'T MEAN HITTING EARTH; I MEAN WE WOULD HAVE BEEN WITHIN ABOUT ONE MILLION MILES. BY ANYBODY'S STANDARD, IT WOULD HAVE BEEN TOTAL DEVASTATION OF THIS PLANET. THE FLOODING WOULD HAVE BEEN ENORMOUS. IT WAS ONLY DUE TO THE FACT THAT THIS THING SLOWED DOWN THAT WE MISSED IT.

MARTIN: I GUESS ANOTHER QUESTION WOULD BE; WHERE IS NEAT GOING?

MCCANNEY: NEAT HEADED BACK OUT. THAT IS OBJECT NUMBER FOUR OF MY SOUTH AMERICAN HARRINGTON EXPEDITION TO CHART THE NEW ORBIT FOR NEAT, BECAUSE IT CLEARLY LOST A LOT OF ITS ENERGY AS IT CAME AROUND THE SUN, AS IT PICKED UP A LOT OF TAIL MATERIAL. SO, IT'S NOT GOING TO COME AND HIT EARTH. THAT'S WHAT NASA ALWAYS SAYS: "OH THESE PEOPLE THINK IT'S GOING TO HIT EARTH". NO, NO; NOBODY SAID ANYTHING ABOUT IT HITTING EARTH. THEY TRY TO MAKE FUN OF PEOPLE AND, IN FACT, THEY ACTUALLY HAVE PEOPLE WHO SET UP THOSE STORIES ON THE INTERNET SO THEY CAN GO MAKE FUN OF THEM. IT'S PART OF THEIR DISINFORMATION CAMPAIGN.

MARTIN: WHEN A COMET THE SIZE OF NEAT, SWINGS BY OUR SUN, HOW DOES 'ACTION AT A DISTANCE' COME INTO PLAY?

MCCANNEY: THE FLARE THAT CAME OFF (THE SUN), THAT YOU SEE IN MANY OF THE PHOTOS, CAME AND HIT THE BACK SIDE OF THE COMET TAIL.

MARTIN: THE FIVE-MILLION-MILE FLARE?

MCCANNEY: YES. NOW, IF THAT WERE TO HAVE COME AT EARTH, IT WOULD HAVE KNOCKED US TO OUR KNEES. BUT IT WENT OFF IN A TOTALLY OBSCURE DIRECTION. LET'S LOOK AT SOMETHING ELSE. WHAT YOU DIDN'T SEE THERE, BUT I COULD SEE IT COMING OFF OF NEAT... IF YOU LOOK VERY CLOSELY, YOU'LL SEE A PIN-THIN STREAK COMING DIRECTLY AWAY FROM THE SUN AND OUT AWAY FROM THE NUCLEUS, OUT THE RIGHT OF THE SCREEN. THAT'S CONNECTING WITH PLANET MERCURY. MERCURY WAS IN A DIRECT ALIGNMENT WITH NEAT AS IT CAME ACROSS THE ECLIPTIC, THE PLANE OF THE PLANETS. THAT LINE, THAT YOU CAN ACTUALLY SEE ON SOLAR PHOTOGRAPHS, IS CONNECTING TO MERCURY.

SO, NOW LET'S PUT EARTH OVER THERE. WHAT IF EARTH HAD BEEN OVER 90 DEGREES AROUND, AND WE WERE NOT BROADSIDE TO IT? THEN, WE COULD HAVE VERY EASILY BEEN IN A POSITION TO TAKE THAT FLARE, FOR EXAMPLE, OR TAKE AN ELECTRICAL DISCHARGE DIRECTLY FROM NEAT. THAT IS WHAT THE ANCIENTS TALKED ABOUT WITH THE COMETS, THE LIGHTNING BOLTS FLYING ACROSS THE HEAVENS; THEY SAW THESE THINGS - ZEUS THROWING LIGHTNING BOLTS TO MARS. THEY SAW THIS STUFF.

MARTIN: IT WAS LITERAL; IT WASN'T METAPHORICAL?

MCCANNEY: NO, IT WAS NOT METAPHORICAL. WHEN VENUS CAME AROUND MARS, IT LASHED OUT WITH AN ELECTRICAL DISCHARGE AND AURORAS IN THE ATMOSPHERE OF MARS LIT UP; IT LOOKED LIKE A SNAKE GRABBING MARS. IT LITERALLY SUCKED THE OCEANS AND ATMOSPHERE OFF OF MARS AS IT PASSED BY. AND THEY SAW THIS. THEY KNEW THAT MARS, PRIOR TO THAT, WAS A WATER PLANET, WAS A BLUE PLANET, JUST LIKE EARTH. MARS HAS A VERY THIN ATMOSPHERE. VENUS HAS A MASSIVE ATMOSPHERE, THOUSANDS OF TIMES DENSER THAN EARTH'S ATMOSPHERE. BUT PERCENTAGE WISE, THE CHEMICAL COMPOSITION OF THE ATMOSPHERES OF VENUS AND MARS ARE EXACTLY THE SAME- WHICH MEANS THEY WERE FORMED IN THE SAME BOILING POT THERE, AS THEY PASSED BY EACH OTHER.

MARTIN: I WANTED TO ASK YOU ABOUT VELIKOVSKY. YOU'RE VERY SIMILAR TO HIM IN THAT HE WAS GIVEN A HARD TIME AND RIDICULED, AND IT TURNS OUT HE WAS RIGHT.

MCCANNEY: THERE'S NO QUESTION ANY MORE THAT VELIKOVSKY WAS RIGHT. AND, I THINK THE BIGGEST THING THAT I WANT TO SAY ABOUT VELIKOVSKY IS THAT HE WAS NOT STUDYING ASTRONOMY. HE WAS STUDYING CALENDARS!

--
(IN THE NEXT SECTION-MCCANNEY REVIEWS MORE OF HIS INTERESTING WORK ON HOW CONTROVERSIAL HE HAS BEEN - "COMETS ARE NOT DIRTY SNOWBALLS- BUT LARGE ELECTRICAL VACUUM CLEANERS IN OUTER SPACE" - BY DEMONSTRATING AGAINST THE SCIENCE COMMUNITY OPINION - THAT THE PLASMA OF SPACE IS ACTIVE - FULL OF CHARGE - AND NOT ELECTRICALLY NEUTRAL. THIS HAS PROFOUND IMPLICATIONS FOR WHAT COMETS DO IN NEAR MISSES - AND PERHAPS - HOW IN PART - "PLANETS EXPERIENCE GRAVITY RELATIONS EROTICALLY" - A GURDJIEFF SENTIMENT.)
--THE LAST SECTION OF THE ARTICLE IS NICE TO ADD HERE: (FITS NICELY WITH THE DOZENS OF NEW WEB SITES COMPARING GEORGE BUSH'S LAWS PASSED SINCE 9/11 TO HITLER)

MARTIN: ARE THERE SOME CLOSING COMMENTS YOU WOULD LIKE TO MAKE?

MCCANNEY: I WOULD SAY, NUMBER ONE THAT THE REST OF THE WORLD IS ADVANCING FAR BEYOND THE UNITED STATES IN CONSCIOUSNESS AND IN PROGRESS AS A HUMAN SPECIES. THE OTHER THING I WOULD SAY IS THAT AS A COUNTRY, AS A CIVILIAN POPULATION, WE HAVE TO GRAB HOLD OF THIS COUNTRY AND TURN IT AROUND BECAUSE, LITERALLY, THE WHOLE REST OF THE WORLD DEPENDS UPON IT. WE ARE AT A STAGE RIGHT NOW THAT IS EQUIVALENT TO 1939, PRE-WORLD WAR II HITLER'S GERMANY.

THEY DID NOT TURN THAT COUNTRY AROUND - AND IF WE DON'T TURN THIS COUNTRY AROUND, WE'RE GOING TO BE IN A FAR BIGGER WORLD PROBLEM THAN WORLD WAR II EVER THOUGHT OF

BEING.

THE REMNANTS OF THE MARDUK/REPTILIAN EXPERIMENTATION IN CENTRAL AND SOUTHERN AFRICA WERE BASICALLY LEFT ON THEIR OWN. THE BEINGS FROM MARDUK, THE ANUNNAKI, ALSO KNOWN AS THE ABBENNAKKI, LEFT THE EARTH, CONTINUING ON ITS ELLIPTICAL JOURNEY THROUGH THE SOLAR SYSTEM AND BEYOND.

THIS ARTIFICIAL PLANET PASSES NEAR EARTH EVERY 12,000 YEARS. A COMPLETE ORBIT TAKES 24,000 YEARS. THE LAST TIME THEY PASSED EARTH, APPROXIMATELY 10,500 BC, ITS GRAVITATIONAL PULL EXERTED IN CONJUNCTION WITH ATLANTEAN CRYSTAL EXPERIMENTATION, CAUSED THE EARTH TO FLIP ON ITS AXIS, SINKING ATLANTIS.

THE ANUNNAKI/ABBENNAKKI HAVE ROBOTICAL SCOUT SHIPS THAT MONITOR EARTH WHEN THE ARTIFICIAL PLANET IS NOT CLOSE. THESE SHIPS HAVE A CREW OF WORKERS THAT LOOK LIKE DWARFED BEARS. THEY ARE KNOWN TO BE AGGRESSIVE, EVEN ATTACKING HUMANS IN THE PAST. THESE BEAR-LIKE CREATURES ARE CYBORGS WITH BRAIN IMPLANTS THAT ALLOW CONTROL FROM A DISTANCE. THEIR BODIES ARE ORGANIC AND POWERFUL.

THESE CREATURES STARTED APPEARING IN THE LATE 1960S. GOVERNMENT-RELEASED UFO INVESTIGATION REPORTS DOCUMENT THEIR EXISTENCE. MORE OF THESE CREATURES WILL BE SEEN IN THE NEXT TWO YEARS. STAY AWAY FROM THEM! THEY ARE DANGEROUS.

MOST LIKELY, THEY WILL SURFACE IN AFRICA TO CHECK ON THE BLACK RACE FOR THEIR MASTERS ON MARDUK. LOUIS FARRKHAN, THE BLACK MOSLEM LEADER, HAS OFTEN SPOKEN ABOUT HIS UFO CONTACTS. ACCORDING TO FARRAKHAN, THESE CONTACTS EXPLAINED THAT THE BLACK RACE WAS CREATED BY THEM.

THEY FURTHER TOLD HIM ABOUT THE NEED FOR THE BLACKS TO DEVELOP WITHOUT INTERFERENCE FROM OTHER RACES.

FOTENOTES:

5. THIS WAS AUGMENTED BY EXPERIMENTS ON GAY MEN IN NEW YORK CITY UNDER THE GUISE OF HEPATITIS SHOTS.

GO BACK

GO BACK TO THE ANUNNAKI

11 - MIND-CONTROL VIA RELIGION

THE FIRST RELIGION ON EARTH WAS THE REPTILIAN BELIEF SYSTEM BROUGHT BY THE COLONISTS OF LEMURIA. THEIR RELIGION BELIEVED IN A GOD-MIND THAT CONTAINED A HIERARCHY, OR CASTE SYSTEM. THIS CASTE SYSTEM WAS EXTRAPOLATED TO THE SEVERAL REPTILIAN SPECIES INCORPORATED INTO THE DRACO EMPIRE. EACH SPECIES HAD ITS OWN PLACE IN THE STRUCTURE OF THEIR SOCIETY. EVERY INDIVIDUAL KNEW ITS FUNCTIONS AND RESPECTED THESE BOUNDARIES. TO VIOLATE THESE RULES MEANT DEATH.

THE REPTILIANS OPERATE AS A GROUP MIND, MEANING THAT NO SINGLE REPTILIAN CAN MAKE A DECISION FOR ITSELF. ONLY THE UPPER CASTE, OR WINGED ONES, HAVE THE SEMBLANCE OF INDIVIDUALITY. THEY WERE, AND ARE, THE LEADERS.

WHEN THIS RELIGION WAS BROUGHT TO SUMERIA, THE CASTE SYSTEM WAS INFUSED INTO SOCIETY AS A RELIGIOUS HIERARCHY. REMEMBER THAT THE COLONISTS OF SUMERIA WERE REFUGEES FROM LYRAE/MARS/MALDEK. THEY, LIKE THE ATLANTEANS, MAINTAINED THE ORIGINAL BELIEF SYSTEM OF THE LYRAEN CULTURE. THE LYRAEN BELIEF SYSTEM ENCOURAGED INDIVIDUALITY, AS WELL AS PROMOTED SERVICE TO OTHERS AS A PATHWAY TO SELF-GROWTH. THE LYRAENS BELIEVED THAT RED-HAIRED PEOPLE WERE CONNECTORS TO GOD-MIND, AND AS SUCH, USED THEM AS ORACLES.

THE REPTILIANS WORSHIPPED THE TRANSPARENT BEINGS FROM THE ASTRAL PLANES AS THEIR CREATORS. THE TRANSPARENT BEINGS HAVE A MASS CONSCIOUSNESS, LIKE AN OVERSOUL. THEY ARE BASICALLY GENDERLESS, ALTHOUGH IN TERMS OF PHYSICAL REALITY, THEIR CHARACTERISTICS AND TRAITS RENDER THEM MORE MASCULINE THAN FEMININE.

WHEN THE REPTILIANS BROUGHT THIS RELIGION TO THE SUMERIANS, THEY WERE CAREFUL TO INTRODUCE IT IN A WAY THAT WOULD BE ACCEPTED AND FOLLOWED. FIRST, THEY CREATED A GENDER BASE FOR A GENDER-MINDED POPULATION. THEN, THEY INSTILLED FEAR TO CONTROL THE MIND-PATTERNS. CLEVERLY, THEY DEVISED A RELIGION BASED ON A MALE-FEMALE, GOD-GODDESS CONTROL SYSTEM. THE MALE GOD WAS CALLED NIMROD; THE FEMALE GODDESS, SEMIRAMIS. THEY ARE DEPICTED AS HALF-HUMAN, HALF-REPTILIAN. THEIR APPEARANCES WERE DESIGNED TO FRIGHTEN THE HUMANS INTO SUBMISSION.

NIMROD AND SEMIRAMIS EVENTUALLY BECAME THE OSIRIS AND ISIS OF EGYPT, AND THE APOLLO AND ATHENA OF GREECE, AMONG MANY OTHER GODS. ALL USED THE MALE/FEMALE, GOD/GODDESS THEME BECAUSE IT REPRESENTED THE ORIGINAL REPTILIAN ANDROGYNY AND THE SEPARATION OF THE HUMAN PROTOTYPE INTO THE MALE/FEMALE, ADAM/EVE.

BECAUSE OF THE "MASCULINE" TENDENCY OF THE TRANSPARENT PEOPLE, AND DESPITE BEING ANDROGYNOUS THEMSELVES, THE REPTILIANS PREFER THE POWERFUL MALE OVER THE FEMALE. THEY REPRESENTED THEIR ANDROGYNY IN SUMER BY PLACING THREE HORNS ON THE REPTILIAN GOD, NIMROD.

THERE ARE MANY LAYERS OF SYMBOLISM TO THIS:

- THE PENIS AND TWO TESTICLES
- TWO ENERGIES UNITING TO CREATE A THIRD, I.E., THE HUMAN PROTOTYPE
- THE THREE LEVELS OF EXISTENCE: HYPERSPACE, ASTRAL, AND PHYSICAL
- THE THREE LEVELS OF AWARENESS: CONSCIOUS, SUBCONSCIOUS, AND SUPERCONSCIOUS
- ANDROGYNY LEADING TO MALE AND FEMALE SECTIONS

THUS, THE NUMBER THREE WAS AN IMPORTANT SYMBOL TO THE REPTILIANS ON EARTH. THEY REPRESENT THIS IN MANY WAYS, INCLUDING THE LILY, OR FLEUR-DE-LIS WITH ITS THREE POINTS.

THEY ALSO USE THE SCORPION WITH ITS STINGER AND TWO PIERCING CLAWS.

THE ADVANCED VERSION OF THE SCORPION IS THE EAGLE, WHICH REPRESENTS THE SCORPION IN ITS HIGHER FORM. BECAUSE OF THIS, THE EAGLE CAME TO REPRESENT RAISING FROM A LOWER FORM TO A HIGHER ONE. IT REPRESENTED POWER AND GLOBALIZATION. BECAUSE IT WAS A BIRD OF PREY, IT WAS ABLE TO CAPTURE EVERYTHING BENEATH IT, ESPECIALLY LIVE FOOD. FOR THESE REASONS, THE ROMANS ALWAYS USED AN EAGLE ON THEIR STAFFS WHENEVER THEY MARCHED INTO A CITY OR COUNTRY.

MOST PEOPLE DO NOT REALIZE THAT ALL BIRDS ARE DESCENDENTS OF REPTILIAN DINOSAURS. MANY CORPORATE LOGOS AND SUPERHEROES OF TODAY HAVE WINGS ATTACHED TO THEM. THE WINGS ALSO REPRESENT THE WINGED REPTILIAN LEADERS IN THE UPPER ECHELON OF THE REPTILIAN/ILLUMINATI FLOWCHART.

HALF A CONTINENT AWAY IN CHINA, THE REMNANTS OF LEMURIA CREATED A MALE-DOMINATED EMPIRE SYSTEM. HERE, THE MALE EMPEROR WOULD ALWAYS HAVE A EMPRESS. THE PEOPLE WERE TOLD THAT THE EMPEROR AND EMPRESS WERE DESCENDENTS OF THE SUN-GOD. THE SYMBOL OF THE EMPIRE WAS THE DRAGON, ANOTHER REPTILIAN FIGURE. THIS "ROYAL" FAMILY CREATED A DYNASTY THAT RULED TIGHT-FISTED FOR MILLENNIA.

THE CHINESE REPTILIAN RELIGION SPREAD ACROSS EASTERN ASIA, WHILE THE SUMERIAN VERSION MEANDERED THROUGH CENTRAL AND WESTERN ASIA. THE SPREAD OF THESE RELIGIONS WAS INTENTIONALLY CONTROLLED FROM THE UNDERGROUND REPTILIAN POPULATION, PRIMARILY CENTERED UNDER TIBET. LOOKING AT A WORLD MAP, ONE CAN EASILY SEE THAT THIS IS THE BEST UNDERGROUND BASE LOCATION IN ASIA TO REACH ALL AREAS OF THE CONTINENT. THESE REPTILIANS WERE AIDED BY BEINGS FROM SIRIUS B WHO DEVELOPED THE BUDDHIST PHILOSOPHIES, AS WELL AS A GROUP OF RENEGADE LYRAENS TRYING TO REPRODUCE A LYRAEN CIVILIZATION UNDER REPTILIAN CONTROL. STRANGE BEDFELLOWS!

AT THE SAME TIME IN INDIA, THE LEMURIAN REPTILIAN REFUGEES CREATED A CASTE SYSTEM THAT WAS A DIRECT REPLICA OF THE REPTILIAN HIERARCHY, FROM THE LOWLY WORKERS/UNTOUCHABLES TO THE BRAHMIN/WINGED ONES. THIS INDIAN/REPTILIAN CULTURE REMAINED LOCALIZED, WRITING THE ANCIENT VEDAS AND BUILDING TEMPLES TO THE VARIOUS REPTILIAN GODS.

MEANWHILE, THE EGYPTIANS, WHO WERE ATLANTEAN/LYRAEN REFUGEES, WERE IN THE PROCESS OF BUILDING A NEW CIVILIZATION FROM THE REMNANTS OF THEIR TWO MAJOR ANCESTRAL ONES THAT WERE DESTROYED. THE BEINGS FROM SIRIUS A HELPED THEM, AS THEY WERE A MAJOR FACTOR IN THE INTERACTIONS OF ATLANTIS.

AS PREVIOUSLY STATED, IN EGYPT, THE REPTILIAN GODS WERE KNOWN AS OSIRIS AND ISIS. THE EGYPTIAN PANACEA OF GODS INCLUDED A LARGE VARIETY OF HYBRID CREATIONS, HALF-HUMAN, HALF-ANIMAL. THIS WAS REMINISCENT OF THE ATLANTEAN HYBRID EXPERIMENTS THAT FOUND THEIR WAY INTO EGYPTIAN CULTURE, AND WAS PROMOTED BY THE SIRIANS WHO WERE PREPARING THAT CULTURE FOR A REPTILIAN TAKEOVER.

THE ATLANTEANS WERE SO ENTRENCHED IN THE ORIGINAL LYRAEN BELIEF SYSTEM, THAT IT TOOK SEVERAL MILLENNIA FOR THE REPTILIANS TO GET A STRONGHOLD IN THAT CULTURE. NOW THAT THE REMNANTS WERE SCATTERED OVER THE EARTH AND NON-COHESIVE, IT WAS EASY TO SEND IN INFILTRATORS (IN-FULL-TRAITORS) TO SWAY THE BELIEF SYSTEM TOWARD A REPTILIAN FLAVOR.

THIS BEGAN WITH THE INTRODUCTION OF THE CAT AS A FOCAL POINT OF WORSHIP.

IN THE SIRIUS A STAR SYSTEM, THE MAIN WORLD IS CALLED KHOOM. THE ANCIENT NAME FOR EGYPT IS KHEM. THERE IS ALSO A CORRELATION WITH MEXICO. SOME RESEARCHERS SAY THAT THE BAY OF CAMPECHE TRANSLATES TO THE BAY OF OLD EGYPT, INDICATING A CONNECTION

BETWEEN EGYPT AND THE YUCATAN PENINSULA. THIS IS NOT SO.

SIRIANS WHO INTERACTED WITH ATLANTIS NAMED THIS AREA AFTER THEIR HOME WORLD, AND THEN CARRIED THE NAME ON TO THE NEW REFUGEE CULTURE IN EGYPT.

OUR INFO, IS THAT THE WORD KHEM FOR EGYPTAUS (EGYPT) MEANT FROM THE BLACKNESS, FOR THE BLACK OR DARK COPPER BASED NUBIAN BLOOD OF THOTH/ENKI RA LINE, WHICH FOUNDED IT. EVIDENCE WOULD INCLUDE THE BLACK ZULU CULTURE WHOSE NAME FOR GOD - ENKI AS IN THE KILIMANJARO - SUGGESTS THEIR OWN BLACK GENETIC ORIGINS WITH ENKI (RA - ABRAHAM) AND HIS SON THOTH. LATER THE WORD KHEM MEANING BLACK HOLE BECAME ALCHEMY AND CHEMISTRY, FROM THE IMPLOSION PRINCIPLE OF THE ABILITY TO MAKE BLACK HOLES (WHICH GOLD’S ELECTRON VALENCE FRACTALITY ENABLES). HENCE ALCHEMY OF THE HEART IS THE FUSION ABILITY WHICH RESULTS FROM PERFECTING COMPRESSION (COMPASSION) IMPLOSION.

ANOTHER PLANET IN ORBIT AROUND SIRIUS A IS A WORLD CALLED KILROTI. HERE, THE SIRIANS CREATED HIGH-INTELLIGENCE CAT-LIKE BEINGS. THESE CAT-LIKE BEINGS ARE CALLED THE LION PEOPLE. IN THE 1970S AND 1980S, THE GOVERNMENT CREATED A CARTOON FOR CHILDREN THAT DESCRIBED THESE BEINGS.

IN THE HIGH ASTRAL LEVELS, THERE IS AN ETHERIC RACE OF LION BEINGS WHO ARE GOLD, HAVE WINGS, AND VIOLET EYES. THE NAME OF THE RACE IS ARI. ARI IS ALSO THE OLD HEBREW WORD FOR LION. THEIR FREQUENCY IS MORE POWERFUL THAN THE DOLPHIN FREQUENCY.

THE ARI CREATED THE OHALU COUNCIL THAT GOVERNS THE SIRIUS A STAR SYSTEM.

THIS FITS INTERESTINGLY WITH ANNA HAYES (VOYAGERS SERIES BOOKS, AND AMENTI) WHO CLAIMS THAT THE REASON THE SPHINX IS LION OR CAT LION SHAPED IS THAT THE STAR ELDERS OF THE INTERVENTIONIST ANUNNAKI (NEPHALIM) WERE LEONINE OR CAT/LION LIKE. SO IT WAS POLITICALLY CORRECT FOR THE FALLEN ANUNNAKI TO CLAIM THAT THEIR GOD WAS THE LION BEING. IF YOUR GOD IS STRONGER THAN YOUR ENEMIES GOD, YOU HAVE A PROBLEM IN WAR, SO YOUR STATUE (THE SPHINX) PROVES YOU HAVE THE RIGHT GOD (THE LION ELDERS OF SIRIUS) ON YOUR SIDE.

WE WONDER WHY SWERDLOW DID NOT SUGGEST THAT ARY-AN .... MEANS LION-BASED AN-NUNAKI - INSTEAD OF SUM-ARYAN.

THE KILROTI WERE GENERATED BY MIXING THE GENETICS OF THE SIRIANS WITH THE ENERGY OF THE ARI. THIS IS WHAT WAS BROUGHT TO ANCIENT EGYPT. AS THE NON-PHYSICAL ENERGY

DESCENDED INTO PHYSICAL REALITY, DNA FORMED THAT COULD BE USED TO CREATE CORPOREAL LIFE.

THIS WAS THEN MIXED WITH HUMAN AND WILD LION DNA TO FORM THE COMMON HOUSE CAT FOUND ON EARTH. THE CAT WAS GIVEN TO EVERY HOME IN ANCIENT EGYPT, AND PROGRAMMED TO LEAVE AT NIGHT TO REPORT BACK TO THEIR ALIEN CONTROLLERS. THIS IS WHY CATS TO THIS DAY HAVE THE URGE TO GO OUT AT NIGHT. THIS ALSO EXPLAINS THEIR ALOOF NATURE.

THE SIRIANS INCORPORATED WORSHIP OF THE CAT IDOL INTO THE EGYPTIAN RELIGION TO ENSURE THE PERPETUITY OF THIS METHOD OF SPYING. THE SIRIANS ALSO BUILT THE SPHINX AS A SYMBOLIC REMINDER OF THE BLENDING OF HUMAN GENETICS WITH LION FREQUENCY. THIS WAS A WAY TO ENERGETICALLY BIND FUTURE CIVILIZATION TO THE SIRIANS. THE SPHINX WAS DESIGNED TO FACE THE MORNING STAR SIRIUS A EVERY DAY. THE FACE ON THE SPHINX IS IDENTICAL TO THE FACE ON THE MARS MONUMENT THAT LOOKS DOWN TO THE EARTH AT THE SPHINX.

SIRIAN TECHNOLOGY BUILT THE COMPLEXES AT THE CYDONIA PLATEAU ON MARS, UPON THE ARRIVAL OF THE FIRST LYRAEN REFUGEES. THE NEW MARTIANS WERE UNAWARE OF THE CLOSE SIRIAN CONNECTION TO THE REPTILIANS.

THE ORIGINAL PYRAMIDS, BUILT AFTER THE DESTRUCTION OF ATLANTIS, WERE ENERGY POINTS. THEY WERE THE SAME SHAPES UNDERGROUND AS ABOVE, MAKING THEM INTO OCTAHEDRONS. AT THEIR CENTRE IS A TETRAHEDRON. THIS MASTER SHAPE IS THE ARCHETYPE SYMBOL FOR GOD-MIND TOTALITY. ANYTHING AT ITS EXACT CENTRE IS ABSOLUTELY PROTECTED. THE OCTAHEDRON IS ALSO THE SHAPE OF THE DELTA-T ANTENNA USED AT THE MONTAUK PROJECT. THIS SHAPE, WHEN ENERGIZED IN THE PROPER COLOR CODES, CAUSES INTERDIMENSIONAL RIFTS, CREATING VORTICES AND WORMHOLES.

RITUALS PERFORMED AT THIS CENTRE POINT PRODUCE VAST ENERGIES THAT CAN BE TRANSMITTED THROUGH HYPERSPACE TO ANYWHERE IN CREATION.

WE SAW PRESTON NICHOLS SPIN THE DELTA T, OCTA ANTENNA OF MONTAUK - IN MODEL - AT VARIOUS PSYCHOTRONIC CONFERENCES. IT SEEMS NEITHER HE NOR SWERDLOW UNDERSTOOD THE REASON IT WAS LATER CALLED DECA DELTA, AS ONLY WHEN THE OCTA/CUBE IS TILT SPUN INTO DODECA (SOULINVITATION.COM/MERKABBAH) DO YOU GET IMPLOSIVE COMPRESSION THAT BECOMES SUPERLUMINAL AND TIME PENETRATING ACCELERATION THRU C LIGHT SPEED. OBSERVE TO DODECA ANTENNA REQUIRED FOR TIME TRAVEL IN THE MOVIE "CONTACT".

SOME RESEARCHERS CLAIM THAT THE PYRAMIDS WERE PUMPING STATIONS FOR UNDERGROUND NILE TRIBUTARIES. THIS IS ONLY PARTIALLY TRUE.

BECAUSE WATER IS AN ELECTROMAGNETIC AMPLIFIER, IT WAS USED TO SURROUND THE RITUAL CHAMBER LOCATED IN THE DELTA-T TO BOOST RITUAL ENERGY. USING THESE METHODS, THE ANCIENT EGYPTIANS CONTROLLED WEATHER, DESTROYED ENEMIES, CREATED STARGATES, AND BOOSTED THEIR REPTILIAN ENERGIES IN CEREMONY. THIS IS ALSO WHY THE MONTAUK PROJECT WAS LOCATED NEAR VAST AMOUNTS OF WATER.

"GREAT WHITE PYRAMID" OF "XIAN,CHINA" IS THE "WORLDS LARGEST PYRAMID". IT IS REPORTED TO BE ABOUT 300 METERS HIGH.

THE GREAT PYRAMID IS PART OF A PROTECTIVE SOLAR SYSTEM GRID, LINKING THE MOON AND MARS MONUMENTS TOGETHER TO PRODUCE A FORCE FIELD TO REPEL INVADERS.

THE GREAT PYRAMID IS ALSO CONNECTED TO OTHER POINTS ON THE EARTH SUCH AS STONEHENGE, A SUBMERGED ATLANTEAN CRYSTAL, TIAHUANACO, AYERS ROCK, AND THE WHITE PYRAMID IN WESTERN CHINA. TOGETHER, THEY FORM AN ENERGY CONTAINMENT FIELD SIMILAR TO AN ELECTRIC FENCE.

THE HAARP PROJECT IN ALASKA TAPS INTO THIS.

MEANWHILE, A DYNASTIC CULTURE SIMILAR TO THE REPTILIAN CHINESE DYNASTIES WAS FORMING IN EGYPT. THE REPTILIAN CONTROL RELIGION BASED ON OSIRIS AND ISIS FANNED OUT OVER NORTHEAST AFRICA AND THE MIDDLE EAST. AT THE SAME TIME THE REPTILIAN SUMERIAN CULT OF NIMROD AND SEMIRAMIS FANNED OUT OVER CENTRAL ASIA.

A UNIFICATION OF THESE BELIEFS WAS NECESSARY TO ACCOMPLISH THE PLAN SET FORTH FROM THE BEGINNING. THIS IS WHY ABRAHAM ORIGINATED IN THE SUMERIAN CITY OF UR, JOURNEYING WESTWARD TOWARD THE MIDDLE EAST. IT WAS NO ACCIDENT THAT THE FOLLOWERS OF ABRAHAM WOUND UP IN EGYPT, WHERE THE SIRIANS THEN CREATED THE HEBREWS. THE HEBREWS WERE DESTINED TO BE THE CULTURAL PROTOTYPE FOR THE FUTURE WORLD.

THIS IS WHY THEY WERE PROGRAMMED TO WANDER ALL OVER THE EARTH AND INFLUENCE ALL EXISTING CULTURES.

(DAN WINTER COMMENTS)

THIS WOULD ALL IMPLY THAT IF ENKI IS IN FACT THE RA IN ABRAHAM - THAT HIS GENETIC ENGINEERING EXPLOITS WITH HIS SPERM TO MAKE EVE FERTILE USING CRO-MAG EGG IMPLANTED IN HIS HALF SISTER DRACO INANNA/NINHURSAG... WAS PART OF A MORE INSIDIOUS PLAN FROM HIS SIRIAN AN FAMILY? IT FITS THAT HIS EXPLOITS DESCEND FROM UR BECAUSE THAT IS THE NAME OF THE DRAC LINE. (N-IBI-URU, URUSHALAEM/JERUSALEM, IL-UR-U, H-IBI-URU/HEBREW , UR-MAN/ROMAN/ROMANIAN, UR-BAN)

WE WOULD HAVE TO EXPECT THAT THE (PART DRAC?) FATHER HE SHARED WITH YAHWEH ENLIL, WAS HALF BORG MACHINE (HARK-ONEN, SARKDAUER, DARTH VADER ...ETC). MORE AT RETURN OF ENKI. IT ALSO FITS THAT ENKI'S DAD AN = ANTU THE SIRIAN SUN GOD (TUETHE DE DANAAN - PEOPLE OF THE GODS) (CALLED BIKI BY AUSTRALIAN ABORIGINALS) - WAS ABLE TO ACT DISTANT AND ALOOF EVEN SEEMINGLY HEARTLESS TO THE CREATURE-LINGS OF EARTHS SEEMING PATHETIC SLAVE BORGS. TODAY, SWERDLOW'S DESCRIPTION OF THE SIRIUS A POLITICS PERFECTLY FITS THAT PICTURE.

THEY WILL STAND BACK, HAVING SOLD GUNS TO BOTH SIDES, IMPARTIALLY WAITING TO SEE WHO WINS- AND ASSUME EDUCATION AND EVOLUTION REQUIRES THE BLOOD SPILL. VERY MUCH CONNECTED TO THE POLITICS OF THE WEST EUROPE ARMS MUNITIONS FAMILIES (PRESUMABLY IN BUSH'S FAMILY GROUP) WHO GOT GREAT CORPORATE RETURN FOR THEIR INVESTMENT IN PAYING FOR THE SUB-SONIC SCALAR DEVICE (A NEAR TOWER BASE LO FREQUENCY VIBRATOR) THAT TOOK DOWN NEW YORK'S TWIN TOWERS (EVIDENCED BY THE SEISMIC TRACE).

...(EXCERPTS)

"THE SEISMIC SIGNALS FROM THE FIVE EVENTS ON 11 SEPTEMBER DIFFERED FROM A SMALL EARTHQUAKE IN SIGNIFICANT WAYS.

THEY WERE RICHER IN LOW-FREQUENCY ENERGY AND POORER IN HIGH-FREQUENCY ENERGY. "...

"WHILE GOING OVER SOME OF THE FOOTAGE I HAD TAPED ON A TIVO RECORDER, A CLIP CAUGHT MY ATTENTION THAT HASN'T BEEN NOTICED BY ANYONE. A SHOT FROM THE TOP OF A BUILDING LOOKING SOUTH SHOWS BOTH TOWERS BURNING AND STILL STANDING. THEN IN THE LOWER LEFT OF THE SCREEN, AT GROUND LEVEL QUITE A DISTANCE FROM THE BASE OF THE TOWERS A COMPLETELY SEPARATE BUILDING COLLAPSES IN A HUGE DUST CLOUD.

I REPEAT, THIS IS BEFORE THE COLLAPSE OF THE TOWERS"

OUR SUPPOSITION THAT ENKI (RA, ABRAHAM, ATON, THOTH, HERMES AND THE AKHANATON-MOSES-JESUS MYTH/LINE THEY SIRED) WAS COLORED TO TRY TO PREVENT THE TOTAL PARASITIC EATING OF THE TAK-ADAMA (ADAM AND EVE) BLOOD LINE (YOU AND ME) EVEN AGAINST THE MORE DRACONIAN WISHES OF HIS FAMILY- BECAUSE HE ESSENTIALLY FELL IN LOVE WITH HIS OWN GENETIC CREATION.

AND EVEN POSSIBLY THAT HE PERCEIVED EVENTUALLY THAT THE BLOOD LINES HE CROSSED COULD BE THE VACCINE IN THE ORION WARS GIVING THE REMNANT HUMANOIDS FROM MANY STAR CULTURES A CHANCE IF THEY COULD (BLISS) IGNITE THEIR DNA TO ENOUGH GRAVITY MAKING TO OUT-STEER THEIR CREATOR PARASITE ANCESTORS??? (THE BST - CLICK IMAGE RIGHT - OF WINGMAKERS)

WITH ALL OF THIS AS A BACKGROUND, THE REPTILIAN BLUEBLOODS NOW NEEDED TO FORM A GLOBAL EMPIRE THAT WOULD ENCOMPASS THE CULTURES THAT THEY HAD INSPIRED.

THEY NEEDED TO COUNTERACT THE OTHER ALIEN GROUPS THAT HAD BEEN WORKING DILIGENTLY IN EUROPE TO CREATE MORE HUMAN CIVILIZATIONS.

GO BACK

GO BACK TO CONTROL MENTAL

GO BACK TO BUDDHISM AND THE KALACHAKRA SYSTEM

GO BACK TO LA VERDADERA HISTORIA DE LOS NAZARENOS Y LA BIBLIA

15 - THE ILLUMINATI HIERARCHY

P.61-67

THE LEADER OF THE EARTHS ILLUMINATI IS CALLED THE "PINDAR". THE PINDAR IS A MEMBER OF ONE OF THE 13 RULING ILLUMINATI FAMILIES, AND IS ALWAYS MALE. THE TITLE, PINDAR, IS AN ABBREVIATED TERM FOR "PINNACLE OF THE DRACO", ALSO KNOWN AS THE "PENIS OF THE DRAGON". SYMBOLICALLY, THIS REPRESENTS THE TOP OF POWER, CONTROL, CREATION, PENETRATION, EXPANSION, INVASION, AND FEAR. THE HOLDER OF THIS RANK REPORTS TO THE PUREBRED REPTILIAN LEADER IN THE INNER EARTH.

RECENTLY, THERE ARE REPORTS THAT THE MARQUIS DE LIBEAUX IS THE PINDAR, BUT THIS IS DISINFORMATION. THE TRUE CURRENT PINDAR IS THE HEAD OF THE ROTHSCHILD FAMILY, AS HAS BEEN FOR SEVERAL HUNDRED YEARS. HE IS BASED IN GERMANY NEAR FRANKFURT. IN THE LATE 1970S, HE OVERSAW THE SISTER PROJECT TO MONTAUK, CALLED M.A.L.D.A. IS AN ANAGRAM FOR MONTAUK-ALSACE-LORRAINE DIMENSIONAL ACTIVATION. THIS PROJECT WAS LOCATED NEAR THE CITY OF STRASBOURG, FRANCE, HISTORICALLY ONCE PART OF GERMANY.

INTERESTINGLY, THERE IS A WINERY ON THE EAST END OF LONG ISLAND, NOT FAR FROM MONTAUK POINT, CALLED PINDAR VINEYARDS. THIS WINE IS GROWING IN POPULARITY, GAINING INTERNATIONAL ACCOLADES. THIS FITS NICELY INTO THE PLAN, AS THIS AREA WILL BE A PART OF THE CAPITAL DISTRICT OF THE EARTH/UNITED NATIONS IN THE EMPIRE STATE! RED WINE IS SYMBOLIC OF THE BLOOD INGESTED BY THE REPTILIANS. THE WINE CAN BECOME SANCTIFIED AS IT DID IN THE ROMAN CATHOLIC CHURCH, A PATSY FOR THE REPTILIANS. IN THE CATHOLIC CHURCH, WINE REPLACED THE BLOOD IN CEREMONY.

THE ILLUMINATI HERE ON EARTH HAVE ESTABLISHED A PYRAMID STRUCTURE OF CONTROL IDENTICAL TO THE SYSTEM THAT EXISTS IN THE DRACO EMPIRE. THE PYRAMID WITH THE REPTILIAN EYE, LOCATED ON THE AMERICAN ONE-DOLLAR BILL, IS SYMBOLIC OF THIS CONTROL STRUCTURE. THE EYE IS THE CAP ON THE PYRAMID, THUS EXPLAINING WHY THE ORIGINAL SURFACE OF THE GREAT PYRAMID IN EGYPT WAS CAPPED IN SOLID GOLD.

THE PINDAR IS REPRESENTED BY THE GOLD CAP ON THE PYRAMID. THE NEXT LAYER, OR "EYE", ON THE PYRAMID REPRESENTS THE 13 RULING FAMILIES.

THEY ARE AS FOLLOWS:

*

- ROTHSCHILD (BAUER OR BOWER) - PINDAR
- BRUCE
- CAVENDISH (KENNEDY)
- DE MEDICI
- HANOVER
- HAPSBURG
- KRUPP
- PLANTAGENET
- ROCKEFELLER
- ROMANOV
- SINCLAIR (ST. CLAIR)
- WARBURG (DEL BANCO)
- WINDSOR (SAXE-COBURG-GOTHE)

EACH OF THE 13 RULING FAMILIES IS GIVEN AN AREA OF THE EARTH AND/OR A PARTICULAR FUNCTION TO FULFILL ON THE EARTH.

THESE PARTICULAR FUNCTIONS INCLUDE:

- GLOBAL FINANCES
- MILITARY TECHNOLOGY/DEVELOPMENT
- MIND-CONTROL
- RELIGION

*

MEDIA

EACH OF THE 13 RULING FAMILIES HAS A COUNCIL OF 13 AS WELL. THE NUMBER, 13, HAS GREAT SIGNIFICANCE TO THEM. THEY KNOW THAT THERE ARE 12 TYPES OF ENERGIES THAT PASS THROUGH THE 10 ASPECTS OF GOD-MIND. THE TOTALITY OF THE 12 ENERGIES EQUALS A 13TH ENERGY. THIS IS CONSIDERED THE MOST POWERFUL KNOWLEDGE.

THEY ALSO KNOW THAT THERE ARE REALLY 13 ZODIAC SIGNS, NOT THE COMMONLY ACKNOWLEDGED 12. THEY HAVE KEPT THE 13TH HIDDEN FOR CENTURIES BECAUSE IT IS THE SIGN OF THE DRAGON. THEY KEEP THE QUALITIES AND TRAITS OF THIS SIGN SECRET TO AVOID GIVING AWAY CLUES TO THE REPTILIAN MIND-PATTERN.

THE NEXT LAYER IS THE SECOND-IN-COMMAND FAMILIES WHO DO THE SUPPORT WORK FOR THE PINDAR AND 13 RULING FAMILIES. WHILE ALL OF THE 13 RULING FAMILY MEMBERS ARE SHAPE-SHIFTERS, ALL MEMBERS OF THE 300 SUPPORTING FAMILIES ARE NOT. THEY DO, HOWEVER, ALL HAVE A HIGH PERCENTAGE OF REPTILIAN DNA.

THEY ARE KNOWN AS THE "COMMITTEE OF 300".THESE FAMILIES INCLUDE SUCH NOTABLE NAMES AS AGNELLI, BALLIOL, BEALE, BELL, BOUVIER, BUSH, CAMERON, CAMPBELL, CARNEGIE, CARRINGTON, COOLIDGE, DELANO, DOUGLAS, FORD, GARDNER, GRAHAM, HAMILTON, HARRIMAN, HEINZ, KUHN, LINDSAY, LOEB, MELLON, MONTGOMERY, MORGAN, NORMAN, OPPENHEIMER, RHODES, ROOSEVELT, RUSSELL, SAVOY, SCHIFF, SETON, SPENCER, STEWART/STUART, TAFT, AND WILSON. THERE ARE MANY OTHERS...

THE COMMITTEE OF 300 USE MANY WELL-KNOWN INSTITUTIONS TO ACCOMPLISH THEIR GOALS, INCLUDING THE

- COUNCIL ON FOREIGN RELATIONS
- BILDERBERGERS
- TRILATERAL COMMISSION
- CLUB OF ROME
- ROYAL INSTITUTE FOR INTERNATIONAL AFFAIRS
- MAFIA
- CIA
- NSA
- MOSSAD
- SECRET SERVICE

*

INTERNATIONAL MONETARY FUND

*

FEDERAL RESERVE

*

INTERNAL REVENUE SERVICE

*

INTERPOL,

...TO NAME A FEW.

ALL OF THESE ARE PRIVATE ORGANIZATIONS OR CORPORATIONS SET UP AS PUBLIC SERVICE DEVICES, BUT THIS IS FAR FROM THE TRUTH.

THE ILLUMINATI STRUCTURE ALSO CREATES ARTIFICIAL COUNTRIES TO FURTHER THEIR GOALS. EXAMPLES OF THESE ARE THE UNITED STATES, SWITZERLAND, KUWAIT, THE SOVIET UNION, PANAMA, ISRAEL, ITALY, YUGOSLAVIA, THE UNITED KINGDOM, MOST OF BLACK AFRICA, ALL OF THE ARAB COUNTRIES, AND ALL OF CENTRAL AND SOUTH AMERICA. THESE NATIONS WERE CREATED TO AMASS WEALTH FOR THE RULING FAMILIES AND THEIR SUPPORTERS, TO HIDE OR KEEP THEIR WEALTH, AND TO CREATE UNSTABLE CONDITIONS NECESSARY TO START WARS OR INCREASE MILITARY BUDGETS.

SWITZERLAND WAS CREATED AS A NEUTRAL BANKING CENTRE SO THAT ILLUMINATI FAMILIES WOULD HAVE A SAFE PLACE TO KEEP THEIR FUNDS WITHOUT FEAR OF DESTRUCTION FROM WARS AND PRYING EYES.

THE UNITED STATES WAS ESTABLISHED WITH 13 COLONIES, ONE FOR EACH OF THE ILLUMINATI FAMILIES. THE ORIGINAL FLAG HAD 13 STARS, AND STILL HAS 13 STRIPES. THE EAGLE, THE SYMBOL OF THE UNITED STATES, HOLDS 13 ARROWS IN ITS TALONS. THE UNITED STATES IS ACTUALLY A CORPORATE ASSET OF THE VIRGINIA COMPANY THAT WAS ESTABLISHED IN 1604 IN ENGLAND WITH DIRECT INVOLVEMENT OF THE ROTHSCHILDS. THE FINANCES OF THE ROTHSCHILDS WERE NECESSARY TO FUND THE EXPLORATION AND EXPLOITATION OF THE NORTH AMERICAN CONTINENT.

THE ASSETS OF THE VIRGINIA COMPANY, INCLUDING THE UNITED STATES, ARE OWNED BY THE HOLY ROMAN EMPIRE VIA THE VATICAN. THIS OCCURRED IN 1213 WHEN KING JAMES GAVE ALL ENGLISH ASSETS TO THE REPTILIAN POPE. EXECUTORSHIP REMAINS WITH THE BRITISH ROYAL FAMILY, BUT ACTUAL OWNERSHIP LIES WITH THE ROMAN CATHOLIC CHURCH.

THE UNITED STATES OF AMERICA IS NOT NAMED AFTER AMERIGO VESPUCCI, AS YOU LEARNED IN SCHOOL. THE ILLUMINATI WOULD NEVER NAME A CONTINENT, ACTUALLY TWO CONTINENTS, AFTER AN ITALIAN MAPMAKER.

THE NAME IS ACTUALLY A COMBINATION OF WORDS.

*

"AM" IS THE HEBREW WORD FOR "PEOPLE"

*

"AME" IS ALSO THE COMMAND FORM OF THE SPANISH/LATIN VERB "TO LOVE"

*

"ERI" OR "ARI" IS A HEBREW TERM FOR "LION"

*

"RICA" IS THE FEMININE FORM OF THE SPANISH WORD FOR "RICH"

*

"KA" IS THE ANCIENT EGYPTIAN WORD FOR SOUL, OR SPIRIT FORCE WITHIN A BODY

THERE ARE TWO LAYERS OF MEANINGS.

THE ANCIENT HEBREW/EGYPTIAN TRANSLATES TO SAY,

"THE PEOPLE OF THE LION WITH SPIRIT FORCE"

HENCE, THE PYRAMID AND ALL-SEEING EYE ON THE ONE-DOLLAR BILL. THE LATINIZED VERSION TRANSLATES TO SAY, "LOVE RICHES", IN A FEMINIZED/PHYSICAL REALITY WAY. THIS GIVES AN IDEA OF WHAT THEY HAD IN MIND.

TAKE THIS A STEP FURTHER, AND ONE SEES THE MIXTURE OF THE FEMININE LATIN/EAGLE IDEAS WITH THE MASCULINE HEBREW/LION IDEAS. THE SYMBOLIC STATEMENT OF AMERICA IS THAT IT IS A COMBINATION OF LEMURIA AND ATLANTIS; A BLEND OF THE HUMAN/LYRAE WITH REPTILIAN/DRACO. PERHAPS THE ANAGRAM LSD, AN ILLUMINATI CREATED DRUG, HAS A HIDDEN MEANING AS WELL: LYRAE-SIRIUS-DRACO!

THE COMBINATION OF THESE THREE CIVILIZATIONS WOULD PRODUCE THE MOST POWERFUL, TECHNOLOGICAL EMPIRE EVER KNOWN!

IN 1776, THE CREATION OF THE UNITED STATES AS AN INDEPENDENT NATION COINCIDED WITH THE DECLARATION INTO PUBLIC EXISTENCE OF THE OFFICIAL ILLUMINATI ORGANIZATION BY MEMBER ADAM WEISHAUPT, IN BAVARIA. PUBLICLY, MR. WEISHAUPT APPEARED TO BE DETERMINED TO CREATE AN ORGANIZATION COMPRISED OF THE EUROPEAN ELITE THAT WOULD UPLIFT MANKIND.

OF COURSE, THIS WAS PART OF AN ILLUMINATI GLOBAL CEREMONY. THE CREATION FOR THE UNITED STATES AND THE ILLUMINATI GLOBAL CEREMONY. THE CREATION OF THE UNITED STATES AND THE ILLUMINATI ORGANIZATION WERE ARTIFICIAL BEGINNINGS FOR PUBLIC CONSUMPTION. THE UNITED STATES WAS THE DEVICE TO BE USED TO BRING THE ILLUMINATI INTO PUBLIC ACCEPTANCE. CURRENT ILLUMINATI MEMBERS BELIEVE THAT ADAM WEISHAUPT WAS A LOOK-ALIKE FOR GEORGE WASHINGTON, AND IT IS ACTUALLY WEISHAUPTS IMAGE THAT APPEARS ON THE ONE-DOLLAR BILL.

GEORGE WASHINGTON WAS A WEALTHY SLAVE AND PLANTATION OWNER. HE IS KNOWN TO HAVE RAPED SOME OF HIS FEMALE SLAVES AND USED SOME OF THE MALE SLAVES IN RITUALISTIC CEREMONY. THERE ARE MANY PEOPLE OF THE BLACK RACE WHO CAN LITERALLY TRACE THEIR GENETICS TO THE FOUNDING FATHERS. GEORGE WASHINGTON ALSO ORDERED THE BUILDING OF THE MONTAUK LIGHTHOUSE IN 1796. THIS LIGHTHOUSE INCLUDED AN UNDERGROUND AREA FOR SUPPLY STORAGE IN CASE OF A BRITISH COASTLINE INVASION. IF HE HAD ONLY KNOWN WHAT THAT AREA WOULD BECOME - OR DID HE?

THE 13 RULING ILLUMINATI FAMILIES CONSTANTLY VIE FOR CONTROL AMONGST THEMSELVES.

DURING THIS TIME PERIOD, THE SPANISH, BRITISH, AND FRENCH ILLUMINATI ALL FOUGHT TO WIN CONTROL OVER NORTH AND SOUTH AMERICA. THE ROTHSCHILDS KEPT THESE ILLUMINATI FACTIONS IN LINE BY SENDING HESSIAN TROOPS TO MONITOR THE SITUATION. THE LEADERS ENJOYED THESE WAR GAMES, PITTING ONE AGAINST THE OTHER TO SEE WHO WOULD WIN. THE HUNDREDS OF THOUSANDS OF LIVES LOST WERE MEANINGLESS TO THEM.

THE MANIFEST DESTINY OF THE UNITED STATES WAS CREATED TO EXPAND THE TERRITORY OF THE ARYANS AT THE EXPENSE OF THE NATIVE POPULATIONS. AS ALWAYS, THE ILLUMINATI SEEK TO DESTROY NATIVE PEOPLES AND THEIR CULTURES. THIS IS AN ATTEMPT TO DESTROY THEIR KNOWLEDGE OF GOD-MIND, AS WELL AS THE POSSIBILITY THAT THE NATIVES WILL IMPART THIS INFORMATION ON TO OTHERS.

ESPECIALLY IMPORTANT IS THEIR NEED TO ELIMINATE NATIVE CULTURES WITH ANCIENT KNOWLEDGE OF ATLANTIS AND LYRAE.

THE NATIVES THAT GAVE THEM THE MOST PROBLEM WERE THE CHEROKEE INDIANS BECAUSE THIS TRIBE RETAINED MOST OF THEIR ATLANTEAN KNOWLEDGE, EVEN ACCESSING THE BEAR/BIGFOOT FREQUENCY FOR INFORMATION. FOR THIS REASON, THESE PEOPLE WERE UPROOTED FROM THEIR HOMELAND IN THE SOUTHERN APPALACHIAN MOUNTAINS, AND FORCIBLY MARCHED TO OKLAHOMA ON WHAT IS NOW KNOWN AS THE TRAIL OF TEARS. MANY DIED ALONG THE WAY. ONLY A REMNANT REMAINED IN NORTH CAROLINA, TENNESSEE, AND GEORGIA. IN THE NORTH, THE VAST IROQUOIS/MOHAWK NATION WAS DISBANDED. THE MONTAUK, DIRECT DESCENDENTS OF THE ATLANTEANS WHO CALL THEIR LEADER PHARAOH, WERE SYSTEMATICALLY ELIMINATED.

THE ROTHSCHILDS WERE AGGRESSIVELY INVOLVED WITH THE SLAVE TRADE FROM AFRICA, IMPORTING SLAVES TO NORTH AND SOUTH AMERICA AS WELL AS THE CARIBBEAN. THEY WERE VERY CAREFUL NOT TO IMPORT BLACKS FROM THE EASTERN AREAS OF ETHIOPIA OR SUDAN WHERE THE DESCENDENTS OF SOLOMON WERE LOCATED, INSTEAD CONCENTRATING ON WESTERN AND CENTRAL AFRICA FOR THE SLAVE POPULATIONS. THESE AREAS HAD THE PURE MIXTURE OF ANUNNAKI AND SIMIAN GENETICS, AND THE PROGRAMMING DESIRABLE FOR THE ILLUMINATI AGENDA.

THE ROTHSCHILDS DECIDED THAT SPLITTING THE UNITED STATES COLONIES WOULD DOUBLE THEIR PROFITS. SO THEY POLITICALLY CREATED, AND FINANCIALLY SUPPORTED, THE CIVIL WAR. THE CIVIL WAR WAS ACTUALLY A GLOBAL CEREMONIAL RITUAL TO BRING SLAVERY TO ITS NEXT LEVEL.

THIS WAR ALLOWED THE NORTH TO WIN, AND PUBLICLY ABOLISH SLAVERY. THE BEST SLAVES ARE THE ONES WHO DO NOT REALIZE THAT THEY ARE SLAVES. THIS ALLEVIATES REBELLION AND RESISTANCE. THIS WAS THE STATUS IMMEDIATELY FOLLOWING THE CIVIL WAR. BLACKS IN THE SOUTH ARE STILL SLAVES. THERE IS STILL SEGREGATION, EVEN IN THE NORTH. THE ILLUMINATI STILL CONSIDER BLACKS TO BE SECOND OR THIRD CLASS CITIZENS. ONLY NOW THE SLAVERY IS SUBTLE AND MASKED.

SINCE THE CIVIL WAR, THERE HAVE BEEN OTHER STAGED WARS THAT ENTRENCHED THE TREND TOWARD GLOBALIZATION. THE SPANISH-AMERICAN WAR OF 1898-1899 ACQUIRED MORE LAND FOR THE AMERICAN ILLUMINATI, PLACING A GREATER PORTION OF THE EARTH'S SURFACE UNDER AMERICAN JURISDICTION. WORLD WAR I WAS DESIGNED TO CHANGE THE MAP OF EUROPE AS WELL AS TEST GERM AND CHEMICAL WARFARE TECHNOLOGY FOR FUTURE USE. THIS COINCIDED WITH THE WORLDWIDE INFLUENZA OUTBREAK DESIGNED TO REDUCE THE GLOBAL POPULATION, MAKING CONTROL EASIER. WORLD WAR I ALSO LAID THE FOUNDATION FOR THE GERMAN ROLE IN THE NEXT WAR.

WORLD WAR II WAS A TEST OF THE FINAL GLOBALIZATION AND EXTERMINATION PROJECTS. IT WAS

ALSO DESIGNED TO TEST MIND-CONTROL MACHINATIONS; TO TEST THE USE OF FLUORIDE WHICH DEADENS BRAIN ACTIVITY AND SLOWS RESISTANCE TO AUTHORITY; TO EXPERIMENT WITH SLAVE LABOR CAMPS AND STUDY THE DEVELOPMENT OF RESISTANCE; AND TO TEACH THE MASSES TO SPY AND REPORT ON ONE ANOTHER.

WORLD WAR II BROUGHT THREE PRIMARY GOALS OF THE ILLUMINATI TO FRUITION.

- THE FIRST WAS THAT HIDDEN ILLUMINATI SYMBOLISMS WERE BROUGHT TO PUBLIC ATTENTION FROM THE UNDERGROUND STRONGHOLDS IN TIBET AND EGYPT, SUCH AS THE SWASTIKA AND THE ANKH.
- THE SECOND WAS THE CREATION OF THE STATE OF ISRAEL AS A FOUNDATION FOR THE NEW WORLD RELIGION.
- THE LAST WAS THE CREATION OF NUCLEAR WEAPONS AS PART OF THE ILLUMINATI GLOBAL CEREMONY.

DURING WORLD WAR II, THE GERMANS HELPED TO PERFECT "SEX-SLAVES" AS A MEANS OF TRANSMITTING INFORMATION AMONGST THE ELITE. SEX-SLAVES CAN BE EITHER MALE OR FEMALE, WHO ARE SEXUALLY PROGRAMMED USING WILHELM REICH PROCEDURES, WHICH ARE ILLEGAL IN THE UNITED STATES, BUT USED BY THE ILLUMINATI AND GOVERNMENT.

THESE SEX-SLAVES DELIVER MESSAGES AND KEEP PROGRAMMED SLEEPERS IN LINE. THE SEX-SLAVE IS DOWNLOADED WITH A MESSAGE OR FUNCTION THROUGH VARIOUS SEXUAL ACTS AND DRUGS, WHICH CAN ONLY BE RELEASED BY REPEATING THE SAME SEXUAL ACT WITH THE TARGET, OR PERSON, TO BE ACTIVATED. THEY ARE TRAINED TO KNOW THEIR TARGET’S TRIGGER WORDS AND TRIGGER EVENTS TO ACTIVATE, DELETE, OR CHANGE PROGRAMMING.

IN RECENT YEARS, SEVERAL WOMEN HAVE COME FORWARD CLAIMING TO BE THE SEX-SLAVES OF GLOBALLY RECOGNIZED POLITICAL FIGURES. MANY WERE USED AS INFORMATION COURIERS BETWEEN HIGH-LEVEL MALE ILLUMINATI. USUALLY, LOOK-ALIKES OF THE POLITICAL FIGURES ARE USED IN THE INCIPIENT PROGRAMMING AS A FOCAL POINT FOR THE SEX-SLAVE. THE SLAVE IS PUT THROUGH A DESENSITIZING PROCESS, SO THERE IS NO PLEASURE IN THE SEXUAL ACT; IT IS MERELY A DUTY TO BE PERFORMED. MANY TIMES THE SLAVE BECOMES SEXUALLY PROMISCUOUS, REPEATEDLY HAVING SEX WITH PEOPLE WHO LOOK LIKE THE INTENDED TARGET. IT IS A SAD LIFE.

BY THE END OF WORLD WAR II, ONE OF THE THREE MAJOR ILLUMINATI GLOBAL RITUALS WAS ACCOMPLISHED.

THIS WAS THE NUCLEAR EXPLOSION THAT TOOK PLACE IN 1945 AT THE 33RD PARALLEL AS A TEST FOR THE NUCLEAR ATTACK ON JAPAN. THIS EXPLOSION WAS SYMBOLIC, REPRESENTING THE SIMULTANEOUS CREATION AND DESTRUCTION OF MATTER AND ENERGY. THE YEAR WAS SYMBOLIC AS WELL. IN NUMEROLOGY, 1 + 9 = 10, REPRESENTING THE 10 ASPECTS OF GOD-MIND. THE NUMBER 10 FURTHER BREAKS DOWN TO 1 + 0 = 1, REPRESENTING A NEW BEGINNING. CONTINUING, 4 + 5 = 9, REPRESENTING THE END OF A CYCLE.

SYMBOLICALLY, THE ENTIRE EVENT REPRESENTED THE END OF A CYCLE TO PREPARE FOR A NEW BEGINNING USING THE NEW CREATION OF GOD-MIND OUT OF DESTRUCTION.

ADDITIONALLY, A CYLINDER CONTAINING MATERIAL STILL NOT EXPLAINED BY THE GOVERNMENT

WAS TRUCKED INTO THE NUCLEAR EXPLOSION TESTING. THIS CYLINDER WAS MADE FROM PURE STEEL AND ALLEGEDLY WAS THE SAME PHYSICAL DIMENSIONS AS THE KABALA DESCRIBES FOR THE CREATION OF GOLEMS. KABALA IS ANCIENT HEBREW METAPHYSICS THAT HAS BEEN A STAPLE FOR THE ILLUMINATI FOR MILLENNIA. GOLEMS ARE ARTIFICIAL BEINGS THAT ARE USED AS A SLAVE FORCE. IT IS HIGHLY PROBABLE THAT THIS WAS A SYMBOLIC RITUAL FOR THE CREATION OF THE SOCIETY OF GOLEMS.

WORLD WAR II ALSO ALLOWED THE EUROPEAN/AMERICAN ILLUMINATI TO DESTROY THE JAPANESE ILLUMINATI DESIRES OF GLOBAL DOMINATION. THE JAPANESE ROYAL FAMILY, REPRESENTED BY EMPEROR HIROHITO, HAVE ALWAYS BEEN OSTRACIZED AS NON-LEGITIMATE BY THE RULING 13 FAMILIES. THE JAPANESE CLAIM TO BE DIRECT DESCENDENTS OF LEMURIAN PUREBRED REPTILIANS.

THE EUROPEAN/AMERICAN ILLUMINATI CLAIM THAT THE JAPANESE ILLUMINATI ARE DESCENDENTS FROM A LOWER SPECIES IN THE DRACO HIERARCHY. THIS LOWER SPECIES IS CONSIDERED A WORKER CLASS WITHOUT ANY POLITICAL CLOUT OR INFLUENCE. THE EUROPEAN/AMERICAN ILLUMINATI ALSO CLAIM THAT EAST INDIANS ARE A LOWER SPECIES IN THE DRACO HIERARCHY. THE 13 RULING FAMILIES CONSIDER LIGHT SKIN AND HAIR TO BE AN ELITE CHARACTERISTIC.

ON JANUARY 17, 1994, JAPAN SENT A SEISMIC EVENT TO CALIFORNIA. EXACTLY ONE YEAR LATER ON JANUARY 17, 1995, THE CITY OF KOBE, JAPAN WAS SEISMICALLY DESTROYED. KOBE WAS THE HOME OF THE JAPANESE ELECTROMAGNETIC WEAPONRY CENTERS. THE EUROPEAN/AMERICAN ILLUMINATI WILL NOT TOLERATE THORNS IN THEIR SIDES. THE DESTRUCTION OF JAPAN AND ITS ROYAL FAMILY WILL CONTINUE IN THE COMING MONTHS.

EVERY YEAR, THE ILLUMINATI HOLD MEETINGS TO PLAN THE EVENTS OF THE COMING YEAR TO ACCOMPLISH THEIR MAIN OBJECTIVE FORMULATED MILLENNIA AGO OF GLOBAL CONTROL AND DOMINATION. IN THE 1850S, THEY PINPOINTED THEIR TARGET DATE FOR COMPLETE DOMINATION WITH AN AGENDA CALLED PLAN 2000.

THIS HAS SINCE BEEN REVISED TO 2003. THE FIASCO ELECTION OF GEORGE W. BUSH JR. TO OFFICE IS A KEY SIGN THAT THEY ARE ON TARGET. THE PUBLIC LESSON OF THE UNITED STATES PRESIDENTIAL "ELECTION" OF 2000 IS THAT THE CITIZENS DO NOT VOTE FOR ANYONE!

EVEN THE ILLUMINATI ARE NOW FINDING IT INCREASINGLY DIFFICULT TO CONCEAL THEIR PLANS.

GO BACK

DRACO

P.158

CREATED BY AN ET GROUP, THE CREATORS TRAVELED BACK IN TIME AND SPACE WITH GENETIC MATERIAL TO SUPPORT THE DISINTEGRATION OF THE LYRAEN CIVILIZATION. WITH SEVEN DIFFERENT TYPES OF DRACO RACES, THE LEADER GROUP IS A SEVEN TO EIGHT FOOT TALL WINGED REPTILIAN-TYPE CREATURE. ABOVE IS PICTURED A WARRIOR USED TO CONQUER AND OCCUPY A PLANET. THE SLENDER FOUR TO FIVE FOOT DRACO SIMILAR TO A LIZARD PERFORMS MENIAL TASKS AND AIDS IN ABDUCTIONS.

HARSH, WARLIKE BEINGS WHO FEEL LITTLE EMOTION, THE DRACO HAVE NO REGARD FOR CULTURE OR OTHER BEINGS. MOST DRACO ARE ANDROGYNOUS AND REPRODUCE BY PARTHENOGENESIS, OR CLONING. ONE SPECIAL GROUP THAT IS COMPLETELY MALE CREATES HYBRID RACES THAT CONQUER OTHERS.

THE EARTH'S MOON IS A DRACO PLANETOID PLACED IN ORBIT AEONS AGO DURING THE TIME OF THE LEMURIAN COLONIZATION. WITH THE INTENTION TO DIVIDE AND CONQUER, THEY ARE KNOWN TO BE BRUTAL, AS WITH RIGEL BY BOILING OCEANS, SCORCHING LANDMASSES, ETC. THE DRACO HAVE VAST UNDERGROUND BASES ON EARTH AND COLONIES ON VENUS.

A SECOND MOON HAS BEEN STATIONED OVER THE EARTH. IT ARRIVE BEHIND THE HALE-BOPP COMET IN 1997. THESE ARE PURE REPTILIANS.

THE HYBRIDS, ILLUMINATI, ARE IN CONTROL OF THE EARTH.

GO BACK

THERE FOLLOWS EXCERPTS FROM THE BOOK:

MONTAUK - THE ALIEN CONNECTION

BY STEWART SWERDLOW
EDITED BY PETER MOON, EXPANSIONS PUBLISHING CO.

COMMENTS IN ITALIC LETTERS ARE FROM DAN WINTER, THE REST IS A QUOTE DIRECTLY FROM THE BOOK-PAGES MENTIONED...

P.36-40

SPEAKING IN SUCCESSION FROM RIGHT TO LEFT, THE FIRST ONE WAS A LARGE REPTILIAN CREATURE WHO LOOKED LIKE A LIZARD STUFFED INTO A BLACK UNIFORM. HE SPOKE WITH A HISSING NOISE IN A LANGUAGE THAT SOUNDED GUTTURAL AND SEVERE.

SIMULTANEOUSLY, I HEARD THE MEANING OF HIS WORDS IN MY HEAD. A MEMBER OF A VAST

EMPIRE THAT SPANNED A LARGE PORTION OF THE GALAXY, HIS PEOPLE ARE ATTEMPTING TO OCCUPY ALL OF THE STAR SYSTEMS ON THE OUTER FRINGES OF THIS GALAXY, EVENTUALLY WORKING THEIR WAY INWARD. A DEFECTOR, HE NOW ADVISES THIS COUNCIL ON THE POSSIBLE ACTIVITIES OF THESE INVASION FORCES.

ACCORDING TO THE LIZARD, THE EARTH WAS INVADED MANY THOUSANDS OF YEARS AGO BY AN ARMY OF HIS PEOPLE THAT ARRIVED IN A HUGE SHIP THAT IS NOW THE MOON OF EARTH.

NUMEROUS SOURCES- INCLUDING WE UNDERSTAND NASA'S SEISMIC DATA - CONFIRM THE MOON IS A HOLLOW METAL FRAMEWORK, TOWED IN HERE, AND INFESTED INTERNALLY WITH GREYS AND DRACS

ANOTHER SUCH SHIP IS ON ITS WAY, HE SAID, DESTINED TO ARRIVE BEFORE THE END OF THIS CENTURY (AS WE CALCULATE TIME).

SEVERAL SOURCES - AMONG THEM ALEX COLLIER, HAVE INSISTED ARTIFICIAL PLANETOIDS - LIKE HALE BOPP'S TOW INS - WERE HEADED HERE IN COMET GRAVITY TAILS - FILLED WITH INVADING DRACOS. THE SUN CRUISERS IN THE SOLAR/MERCURY NEAR ORBIT - VISIBLE ON SEVERAL WEB SITES - ARE SAID TO BE EXAMPLES OF THESE. ALEX'S ANDROMEDAN MATERIAL SPECIFICALLY PREDICTED THE COMET WOULD DROP THE INVASION SHIPS IN A TEMPORARY NEAR MERCURY ORBIT.

AFTER BEING DRIVEN OFF THE EARTH BY SETTLERS FROM THE LYRAEN EMPIRE, HIS PEOPLE WENT UNDERGROUND.

MOST OF THE LEMURIAN EPOCH STORIES SUGGEST THAT THE END OF THE LEMURIAN CONTINENT WAS A WAR WITH A REPTILIAN RACE, WHICH THE HUMANOIDS WON AGAINST THE REPTILIANS - BUT AT THE COST OF THE ENVIRONMENT - FORCING THE HUMANOID SURVIVING REMNANT UNDERGROUND - WHERE THEIR BLOOD BECAME MORE RED AND DEPENDANT ON IRON TO BOND OXYGEN (AS OPPOSED TO THE HUMANOID REMNANT WHO FLED AT THAT POINT LARGELY TO SIRIUS AND PLEIADES WHERE THEY BECAME MORE BLUE/COPPER BLOOD.)

THERE, THEY REMAIN IN STASIS UNTIL THEY ARE TO BE REACTIVATED BY THE INCOMING SHIP. THESE REPTILIANS ALSO MAINTAIN BASES ON VENUS AND ON SOME OF THE MOONS OF THE OUTER PLANETS. BECAUSE HIS PEOPLE ARE MALE ONLY, THEY CREATED FEMALES FOR THE SOLE PURPOSE OF BREEDING. DESPITE THIS, CLONING IS THEIR PRIMARY METHOD OF REPRODUCTION. IN CONCLUSION, THE LIZARD ADDED THAT I WOULD SOMEDAY CONVERT HIS PEOPLE TO THE LIGHT BECAUSE MY SOUL WAS AN EMISSARY TO THEM MANY YEARS AGO. REMEMBERING ME, THEY WOULD RESPECT WHAT I SAID TO THEM.

THE NEXT ONE TO SPEAK WAS AN AMPHIBIAN-TYPE BEING WHO LOOKED VERY MUCH LIKE THE CREATURE FROM THE BLACK LAGOON. SLIMY AND MOIST, HIS BODY WAS COVERED WITH A SCALY, GREENISH-GREY SKIN. BREATHING HEAVILY, HIS THOUGHTS FLOODED MY BRAIN. HIS CIVILIZATION WAS THE ORIGINAL INHABITANT OF EARTH BEFORE THE OTHERS ARRIVED. AT THAT TIME, EARTH WAS MOSTLY WATER AND MARSHES. DEVASTATED BY MANKIND AND ALIENS ALIKE, THERE REMAIN ONLY A FEW POCKETS OF HIS PEOPLE AT THE DEEPEST DEPTHS OF THE OCEANS.

FROM TIME TO TIME THEY COME TO DRY LAND TO SUN THEMSELVES, THUS CREATING THE BASIS FOR THE LEGENDARY MERMAIDS AND MERMEN. THIS SPECIES WORKED CLOSELY WITH THE ATLANTEANS BEFORE THAT CONTINENT SANK. THEY SERVED AS LIAISONS BETWEEN MANKIND AND THE WHALES AND DOLPHINS. HE SAID THAT THESE SEA MAMMALS ARE ADVANCED RACES FROM ANOTHER GALAXY. MOST OF HIS PEOPLE WERE LIFTED OFF-WORLD TO UNDERGROUND OCEANS ON NEPTUNE. THIS MANEUVER WAS FACILITATED BY BENEVOLENT ET GROUPS. CONTINUING, HE TOLD ME THAT I HAVE DOLPHIN DNA; THEREFORE, I COULD LEARN TO COMMUNICATE WITH HIS SPECIES IN ORDER TO HELP MANKIND AND THE DOLPHIN/WHALE SYSTEMS.

NEXT, A BEING ON THE OPPOSITE SIDE OF THE TABLE SPOKE. LOOKING LIKE A SMALL DARK-HAIRED HUMAN, HIS EYES WERE SO DARK THAT THEY ALMOST SHINED. HIS EYES SEEMED TO PIERCE RIGHT THROUGH MY OWN, AND I FOUND IT DIFFICULT TO LOOK AT HIM. CLAIMING TO REPRESENT THE FEDERATION OF PLANETS OF THIS GALAXY, HE SAID THAT THERE ARE OVER 120 DIFFERENT MEMBER CIVILIZATIONS. SOMEDAY, EARTH WILL BE ASKED TO JOIN, BUT ONLY IF THEY ARE SUCCESSFUL IN REPELLING THE INVASION FORCE. OTHERWISE, EARTH COULD BECOME A TARGET BY THE FEDERATION UNTIL THE REPTILIANS ARE REMOVED.

THIS DARK-HAIRED HUMAN SAID THAT I WAS SELECTED TO SPEAK WITH ALL OF THESE ALIENS BECAUSE EACH SPECIES HAD CONTRIBUTED TO MY DNA CREATION. I WAS MADE FOR THE EXPRESS GENETIC PURPOSE OF BELONGING TO MANY DIFFERENT GROUPS. BECAUSE OF THAT, EACH GROUP WOULD BE MORE WILLING TO LISTEN TO ME AND ACCEPT MY IDEAS SINCE I PARTIALLY BELONGED TO THEM.

CONTINUING, HE SAID THAT MY SOUL-PERSONALITY HAD AGREED TO THIS MISSION LONG AGO, UNDERGOING TRAINING IN MANY GALAXIES AND ALTERNATE UNIVERSES. HE SAID THAT MUCH OF WHAT THEY TOLD ME NOW WOULD STAY HIDDEN IN MY CELLULAR MEMORY UNTIL EACH PIECE OF INFORMATION BECOMES NECESSARY. FUTURE HARDSHIPS AND SADNESS WOULD CONDITION ME FOR MY MISSION. NOT WANTING TO HEAR ANY OF THIS, I TENSED UP TO THE POINT OF VOMITING. THE NEXT THING I REMEMBER, I WAS STILL IN THE CHAIR WITH ANOTHER BEING TALKING TO ME.

THIS CREATURE WAS A PALE GREYISH-WHITE. HE POSSESSED LARGE, ROUND, BLACK EYES AND A LONG NOSE. A SHORT CURTAIN OF MATERIAL DRAPED AROUND THE BACK OF HIS LARGE HEAD. HIS THIN MOUTH DID NOT MOVE AS HE SPOKE WORDS INSIDE MY HEAD. WITH A HARSH ATTITUDE, THIS GREY CLAIMED TO BE FROM THE RIGEL STAR SYSTEM. ALTHOUGH HE DID NOT WANT TO BE PART OF THIS MEETING, HIS OVERLORDS INSISTED THAT HE ATTEND IT. HIS JOB WAS TO MONITOR THE PROCEDURES SINCE I WAS ONCE A PART OF THEIR RACE, TOO.

HIS HOME WORLD WAS ONCE PART OF THE LYRAEN CIVILIZATION. AFTER THAT CULTURE DISINTEGRATED, THE REPTILIANS INVADED HIS PLANET. NOW HIS PEOPLE ARE PART OF THAT EMPIRE AND AS SUCH, DO THEIR BIDDING. FROM TIME TO TIME HE WOULD SEND HIS WORKERS, SMALLER GREY CREATURES, TO CHECK UP ON ME. MOST OF THE TIME I WOULD NOT REMEMBER IT EXCEPT AS A BAD DREAM.

MICHAEL ASH - THE SUPPOSED CHIEF TIME EMPATH FROM MONTAUK - HAD DESCRIBED TO ME HOW UNPLEASANT AND INTERVENTIONIST THE MANY RIGELIANS WERE - WHO HE ENCOUNTERED AT MONTAUK. ALSO NOTABLY IN THE BUEHLER/THOTH MATERIALS- THOTH (SON OF ENKI/ADONAI/RA/HERMES) INSISTS HIS BLOOD IS BLUE/BLACK BECAUSE HE WAS FROM RIGEL IN ORION. I SUSPECT THIS WAS THE ANCESTRAL FAMILY OF ENKI/RA'S MOTHER AIDA WHO WAS

WINGED DRAC.. "PTAH TAAL" .

(THOTH - NUBIAN - BLUE BLACK BLOOD - WAS THE REASON THE ORIGINAL NAME FOR EGYPT WAS KHEM - WHICH MEANS FROM THE BLACKNESS - THE TERM ALCHEMY THEREFORE MEANS- SCIENCE OF MAKING BLACK HOLES. IT WAS THOTH - HERMES FAMILY WHO IMPREGNATED AKSENPATEN - WHICH BECAME THE TUT/MAGDALEN STORY - AND BLACK MADONNA). IN ALEX COLLIER - ANDROMEDAN PAPERS (DEFENDING SACRED GROUND - ETC.)

READ ABOUT THE ORIGIN OF THE TERM PTAH IN EGYPT (RA /ENKI'S LINE) FROM THE TERM PTAH TAAL - MEANING 11TH DIMENSIONAL OR BIRD TRIBE - BUT MORE PROBABLY WINGED DRAC OR CIAKAR OR MOTHMAN - THE SOURCE OF THE PSYCHOKINESIS IN THE DNA - ENKI - RA - ABRAHAM - USED IN HIS OWN SPERM TO MAKE ADAM AND EVE FERTILE.

THERE WOULD COME A TIME WHEN THESE GREYS WOULD BE AT WAR WITH SOME OF THE WORLDS REPRESENTED AT THIS MEETING. MY JOB WOULD THEN BE TO MONITOR THE ACTIVITIES AND REPORT THE FINDINGS TO MY CONTROLLERS. MY BODY CONTAINS CHEMICALS NEEDED BY HIS RACE. THE GREY SAID THAT HIS RACE ALSO LOOKED HUMAN UNTIL WAR CONTAMINATED THEIR ENVIRONMENT, GENETICALLY DEGENERATING THEM. MY BODY WAS PRESENTED TO THEM AS A TEMPORARY TOKEN OF HOPE AND PEACE.

ALLOWED TO USE MY GENETICS FOR THE PURPOSE OF UPGRADING THEIR DYING SPECIES, THEY COULD NOT PURPOSEFULLY HARM ME, PERMANENTLY KIDNAP ME, OR ALLOW ME TO REMEMBER WHAT THEY DID TO GET WHAT THEY NEEDED FROM ME. THESE WERE THE RULES THAT ALL ABIDED BY TO KEEP THE STATUS QUO WHILE AT THE SAME TIME BENEFITING FROM MY CREATION.

IF MY MISSION FAILED, OR IF ANY OF THE PARTICIPANTS IN THIS PROJECT NO LONGER WISHED TO CONTINUE, THEN I WOULD BE REMOVED TO A SAFE PLACE WHILE THE OTHERS FOUGHT AMONGST THEMSELVES, POSSIBLY EVEN ON THE SURFACE OF THE PLANET EARTH. FINALLY, HE TOLD ME THAT HIS PEOPLE MERELY WISHED TO CORRECT THE ERRORS COMMITTED AGAINST THEM IN ORDER TO EVOLVE. THEY WISHED TO BECOME INDEPENDENT OF THE REPTILIANS AND RECREATE THEIR OLD CIVILIZATION BEFORE IT WAS TOO LATE. THEIR DILEMMA WAS THAT, IF THEY ATTEMPTED THIS, THEY WOULD BE DESTROYED COMPLETELY BY THEIR MASTERS. ON THE OTHER HAND, IF THEY CONTINUED WITH THEIR MASTERS ORDERS, THEY WOULD BE TARGETED BY THE FEDERATION PLANETS. BECAUSE THEY FELT HOPELESS, THEY LOOKED OUT ONLY FOR THEMSELVES.

AT THIS POINT, THE LAST BEING AT THE COUNCIL TALE, WHO SAT IN THE MIDDLE, INTERRUPTED THE RIGELIAN. THIS BEING WAS VERY TALL. STANDING UP, HE RAISED HIS ARMS TO EITHER SIDE. THE BEAUTIFUL WHITE ROBE HE WORE WAS TRIMMED IN A BLUE THAT I HAD NEVER SEEN BEFORE. HIS LARGE HEAD WAS ROUND ON TOP WITH A POINTED CHIN. STANDING WITH HIS ARMS OUTSTRETCHED, HE REMARKABLY RESEMBLED A LIVING ANKH. HIS OVAL EYES WERE BRILLIANT BLUE; HIS SKIN IVORY WHITE. BY FAR, HE WAS THE MOST IMPRESSIVE BEING AT THIS GATHERING.

WHEN HIS THOUGHTS FILLED MY HEAD, I COULD NOT EVEN THINK OF MY OWN NAME! AS HE SPOKE, I SAW WORDS IN A STRANGE LANGUAGE SWIRLING ETHEREALLY AROUND HIS HEAD.

SEQUENCES OF BRAIN CORTEX CHARGE DOMAINS IN COMPRESSION SYMMETRY - ORIGIN OF ANCIENT ALPHABETS - (COMPARE- HEBREW - THE SOFTWARE ENVIRONMENT FOR A CENTRAL BORG COMPUTER DOING GENETIC INTERVENTION - VERSUS OPHANIM - BIRD TRIBE- ORIGIN OF STARGATE THE MOVIE).. AND PHYSICS SYMMETRY ELEMENTS OF CREATION.

HE TOLD ME HIS NAME, BUT I CANNOT REMEMBER IT. COMING FROM THE PLANET KHOOM IN THE BINARY SIRIUS STAR SYSTEM, HIS PEOPLE WERE DESCENDENTS OF NON-PHYSICAL BEINGS WHO INHABITED HYPERSPACE, A REGION OF CONSCIOUSNESS EXISTING OUTSIDE OF LINEAR TIME AND SPACE. THEY CREATED THE ANCIENT EGYPTIAN CIVILIZATION AS WELL AS THE JEWISH PEOPLE AND GAVE THEM THE TORAH. THE CRYSTAL SKULL WAS THEIR CREATION, AND THEY WERE IN CHARGE OF MANY EVENTS IN THE GALAXY AND BEYOND. MY SOUL-PERSONALITY WAS FROM HIS PEOPLE BECAUSE IT WAS THE ONLY TYPE ADVANCED ENOUGH TO ANIMATE SUCH A HYBRID BODY AS MY OWN.

POSSESSING THE MOST ADVANCED TECHNOLOGY IN THE UNIVERSE, ALL THE OTHER SPECIES CAME TO THESE SIRIANS FOR INFORMATION. NOW, AS AN ADULT, I REALIZE THAT THIS GROUP ALSO PLAYS ONE CIVILIZATION AGAINST THE OTHER TO BENEFIT EVOLUTION AS WELL AS THEIR OWN SPECIES.

WE KNOW ENKI/ENLIL (ADONAI/LUCIFER - VERSUS YAHWEH/MICHAEL) FATHER WAS MOST LIKELY FROM SIRIUS. IN A HOLDOUT POSITION TRYING TO SAVE THE LAST HUMANOID REMNANT AGAINST THE ONSLAUGHT OF THE BORG MACHINE EMPIRE (EMPIRE STRIKES BACK BECOME REAL). WE SUSPECT HE (ANTU / OR BIKI AS THE ABOS CALL HIM- LITERALLY SUN GOD / AN) WAS HALF MACHINE - BUT NOT HAPPY ABOUT THAT (CORRECT HISTORY DESCRIBING MUAB DIB'S FATHER LINE - HARKONEN - HALF MACHINE IN DUNE - AND LUKE SKYWALKER'S FATHER DARTH VADER - HALF MACHINE IN STAR TREK - A PHYLLIS SCHLEMMER ORIGINATED STORY).

WHAT IS CONFUSING IS THE ANDROMEDAN MATERIAL - HAS INSISTED SIRIUS A WAS NON-INTERVENTIONIST IN IT'S POLITICAL RELATION WITH EARTH - WHILE SIRIUS B WAS MORE INTERVENTIONIST (INTERFERE WITH INDIGENOUS GENEPOOLS WITHOUT REGARD TO THE PRIME DIRECTIVE).

THIS NOTE HINTS THAT ALL THE GOOD DNA CALLED HEBREW - ACTUALLY DESCENDED FROM ENKI IN EGYPT (THE BOOKS: OUT OF EGYPT, & HOUSE OF THE MESSIAH & COPPER SCROLLS - PRESENT OVERWHELMING EVIDENCE - THE RABBI'S LIED ABOUT ISRAEL BEING MOSES ORIGIN - MOSES WAS THE NAME AKHANATON TOOK - WHEN ENLIL/AMUN - TOOK OVER EGYPT & WHILE HE SWIPED THE FAMILY GOLD TO FOUND THE ESSENES).

THE THING MISSING HERE IS THE SOURCE OF THE GOOD SIRIAN DNA - WHICH IS ACTUALLY WINGED DRACONIAN... ENKI'S MOM. THE CRYSTAL SKULL WAS A DECADENT CONTROL DEVICE USED LATER BY A FALLEN MAG MATRILINY WHO LOST THE SKILL TO PASS RACIAL MEMORY IN A REAL SKULL TOUCH. - SEE ORIGINS OF BAPHOMET.

8 - THE SIRIANS
P. 77- 84

THE APARTMENT IN HOLON BECAME MY HOME BASE. FROM THERE, I LEFT ON MANY TRIPS AROUND THE REGION, SOMETIMES FOR DAYS AT A TIME.

ONE PARTICULAR ADVENTURE TOOK PLACE WHILE ON AN EXCURSION TO THE NEGEV DESERT IN THE SOUTH OF ISRAEL. I HAD JUST PASSED THE RESTRICTED AREA AROUND DIMONA, A TOWN WHERE ISRAEL HAS ITS SECRET NUCLEAR FACILITY. THIS PLACE IS GUARDED BY THE FALASHA, THE BLACK JEWS FROM ETHIOPIA WHO CLAIM TO BE DESCENDENTS OF KING SOLOMON AND THE QUEEN OF SHEBA. THE QUEEN SUPPOSEDLY RETURNED TO ETHIOPIA WITH THE ORIGINAL ARK OF THE COVENANT WHERE IT REMAINED FOR SAFEKEEPING UNTIL ONLY A FEW YEARS AGO.

THE SO CALLED BLACK JEWS ARE THE REAL NUBIAN COPPER BLUE BLOODS KIDS OF ENKI/RA... SEED OF SEPHARTICS... NOT THE USURPING (ENLIL'S) ASHKANASI - INNANA DRAC DESCENDANTS - FOUNDERS OF CHINA AND BANK OF ENGLAND... IT IS NOT OVER TRIVIALIZING TO CHARACTERIZE THE CONTINUING DIALECTIC BY THE ENKI VS ENLIL HALF-BLOOD BROTHERS-LINES.

KING SOLOMON WAS A POWERFUL MYSTIC (AKHANATON - WHOSE AMARNA-SOLAR CAPACITOR FOR IGNITING DNA BECAME THE SOLOMON TEMPLE MYTH). HIS ADVISORS ALWAYS DISCUSSED THE FUTURE WITH HIM. IT IS BELIEVED THAT HE WAS WARNED OF A FUTURE TIME WHEN ISRAEL WOULD BE OVERRUN WITH EVIL FOREIGNERS. (ASHKANAZI - WHO NOW FLY THE ALPHA DRACONIS FLAG - THE HEX DOUBLE TRIANGLE - OVER ISRAEL - ALSO HAPPENS TO BE THE SHAPE OF THE INTERVENTIONIST DRACO CRAFT COMING TO INVADE). THE HOLY ARK OF THE COVENANT WOULD MOST LIKELY BE A PRIME TARGET.

I BELIEVE THAT SOLOMON, IN ALL HIS WISDOM, DECIDED TO HIDE THE REAL ARK OF THE COVENANT IN A PLACE WHERE NO ONE WOULD THINK TO LOOK. CONVENIENTLY, THE QUEEN OF SHEBA WAS LOOKING FOR A MAJOR TOKEN OF FAITH AND AFFECTION FROM HER CURRENT LOVER. SOLOMON SENT THE ARK BACK WITH HER TO ETHIOPIA AND MADE A FACSIMILE FOR THE PUBLIC TO VIEW.

THE ARK OF THE COVENANT IS A FRACTAL SELF SIMILAR CAPACITOR ('ORGONE') WITH ENOUGH IMPLOSIVE COMPRESSION TO PARTIALLY SAFELY CONTAIN RADIOACTIVE POWER. (RADIOACTIVITY IS THE OPPOSITE OF FRACTAL CHARGE COMPRESSION - WHICH IS WHY URI GELLER COULD MEAN...

IN ORDER TO SOLIDIFY HIS TRUST FOR SHEBA AND THE ARK, SOLOMON SAW TO IT SHE HAD HIS CHILD SO THAT SHE WOULD ALWAYS FEEL AN OBLIGATION TO HER CHILD'S PEOPLE. HIS WAS THE PERFECT PLAN.

THE QUEENS DESCENDENTS BECAME THE FALASHA, A NUMEROUS FACTION IN ETHIOPIA. AT ONE TIME, IT IS BELIEVED THAT THERE WERE OVER 500,000 FALASHA IN ETHIOPIA. IN FACT, HEILE SALASSIE, THE RULER OF ETHIOPIA UNTIL A MILITARY COUP TOPPLED HIS REGIME, CALLED HIMSELF THE LION OF JUDAH.

THE ISRAELI MOSSAD, OR INTELLIGENCE AGENCY, HELPED THE PROVINCE OF ERITREA BREAK AWAY FROM ETHIOPIA BECAUSE THE ARK WAS HIDDEN THERE AND THE MILITARY GOVERNMENT

REFUSED TO COOPERATE WITH THE ISRAELIS. CONSEQUENTLY, THE ISRAELIS CREATED THEIR OWN COUNTRY WITH WHOM TO NEGOTIATE. THE ARK WAS SUCCESSFULLY TRANSFERRED TO THE HOLY LAND AT THE SAME TIME THE MILITARY GOVERNMENT OF ETHIOPIA ALLOWED THE ISRAELIS TO AIRLIFT ALL OF THE REMAINING FALASHA OUT OF ETHIOPIA. A STRANGE TURN OF EVENTS, BUT ONE THAT PORTENDS THE COMING OF THE MESSIAH FOR THE JEWS.

IT WAS NEAR DIMONA THAT I WAS ABDUCTED BY THE SIRIANS FOR A MAJOR ADVENTURE. AS I WALKED AROUND SOME ROCK FORMATIONS THAT JUTTED UP OUT OF THE DESERT, THE SUN WAS HOT AND THE AIR EXTREMELY DRY. PERSPIRATION EVAPORATED IN A MANNER OF SECONDS AFTER APPEARING ON MY SKIN. SUDDENLY, A TREMENDOUS FLASH OF LIGHT BLINDED ME. RAISING MY HANDS TO SHIELD MY EYES, I REALIZED THAT THERE WERE NO SOUNDS AROUND ME; IN FACT, THERE WAS AN EERIE SILENCE ALL ABOUT.

OPENING MY EYES, I FOUND MYSELF IN A GREAT ROOM THAT RESEMBLED AN ANCIENT GREEK OR ROMAN THRONE ROOM WITH HUGE COLUMNS AND A SCREEN ON THE WALL. LIGHTING WAS SUBDUED AND WITHOUT A VISIBLE SOURCE. THERE WERE NO WINDOWS OR DOORS. FROM A SPACE NEAR THE SCREEN, A TALL SIRIAN DRESSED IN A WHITE AND BLUE ROBE APPEARED. HE WAS NEARLY IDENTICAL TO THE SIRIAN I HAD SEEN AS A CHILD. ABOUT SEVEN-FEET TALL, HIS SKIN WAS PALE. HE HAD LARGE POINTED EARS, A LONG POINTED NOSE, AND BIG, BLUE ALMOND-SHAPED EYES.

HIS MOUTH AND LIPS WERE VERY SMALL. HIS FINGERS APPEARED LONG AND GRACEFUL AS HE STOOD IN FRONT OF ME WITH OUTSTRETCHED ARMS. THE WAY HIS ROBE TOUCHED THE FLOOR, COMBINED WITH THE SHAPE OF HIS ELONGATED HEAD, HE LOOKED EXACTLY LIKE A LIVING ANKH! WHAT A BEAUTIFUL SIGHT! I FELT NO FEAR BECAUSE I HAD SEEN ONE BEFORE AND FELT AT HOME WITH IT. THE SCREEN CAME ON BEHIND HIM AND ALLOWED ME TO SEE THE OUTLINE OF EARTH AS WE MOVED SWIFTLY AWAY FROM IT IN SOME SORT OF A CRAFT HEADING TOWARD MARS. IN THAT ROOM, THERE WAS NO SENSATION OF MOVEMENT. IT FELT SOFT AND COMFORTABLE EVEN MORE COMFORTABLE THAN HOME.

THE BEING MOVED FORWARD AND SMILED, TELEPATHICALLY WELCOMING ME. THROUGH HIS THOUGHTS, HE TOLD ME THAT WE WERE ON OUR WAY TO MARS TO SEE SOMETHING THAT WOULD HELP ME AT A LATER TIME. CONTINUING, HE SAID THAT HIS SPECIES CREATED PEOPLE FROM STOCK BROUGHT TO EARTH FROM ELSEWHERE. THESE CREATIONS ARE KNOWN AS HEBREWS, AS IS THEIR LANGUAGE.

MUCH OF THIS HAD BEEN DISTORTED AND INTERFERED WITH OVER THE MILLENNIUMS. HIS SPECIES WAS NOW IN LEAGUE WITH THE REMNANTS OF THESE ORIGINAL CREATIONS, THE ISRAELIS, TO CORRECT AND PURIFY THEM BEFORE THE NEXT WAVE OF INTERRUPTIONS OCCUR. APPARENTLY, THE SIRIANS FELT THAT THESE HEBREW REMNANTS WERE NOT THE GENETICALLY-WHOLE BEINGS THAT THEY ORIGINALLY CREATED. BECAUSE IT WAS TOO LATE TO CHANGE THEIR PHYSICAL STRUCTURE, IT HAD BECOME NECESSARY TO CHANGE THEIR MIND-PATTERNS, THUS ENABLING THEM TO INTERFACE WITH OTHER ALIEN BEINGS. AS HE COMMUNICATED, PICTURES OF HIS WORDS APPEARED ON THE SCREEN BEHIND HIM. I ACTUALLY SAW HISTORY FROM THE DIM PAST AS THOUGH IT WAS HAPPENING AT THAT MOMENT.

NEXT, I REMEMBER LYING ON A TABLE IN A BRIGHTLY LIT CLINICAL-LOOKING ROOM. THE SIRIAN WAS THERE WITH A TALL GREY BEING WHO HAD ROUND, BLACK EYES AND A GOOFY LOOK ON ITS FACE. IDENTIFIED TO ME AS A VEGAN, IT WAS A CREATION BY THE SIRIANS OUT OF THEIR OWN GENETICS MIXED WITH THE LITTLE GREYS. ALTHOUGH NOT STRAPPED DOWN, I COULD NOT MOVE, AND I REALIZED THAT I WAS NAKED. FOR THE FIRST TIME, I FELT FEAR WITH THE SIRIANS. A BRIGHT DEVICE WAS PLACED OVER MY FACE AS THEY PROBED MY GENITALS, STOMACH, AND CHEST.

THE SIRIAN SAID THAT THEY NEEDED TO TEST FOR RESIDUALS OF GENETIC RESONANCE IDENTIFYING MY BODY WITH THEIRS. HE SAID THEY HAD MANUFACTURED MY BODY IN COLLUSION WITH CERTAIN EARTH FORCES FOR THE PURPOSE OF COMPLETING THEIR AGENDA. I WAS SOME KIND OF PAWN IN AN INTERSTELLAR GAME. WHAT WAS GOING ON HERE?

TO COMPLICATE MATTERS FURTHER, HE SAID THAT MY SOUL-PERSONALITY WAS SOME TYPE OF ENTITY FROM A NON-PHYSICAL ASPECT OF SIRIUS WHICH WAS AN EVEN HIGHER FORM THAN HIMSELF. MY BODY CONTAINED SIRIAN DNA WHICH WAS NECESSARY FOR THE SOUL-PERSONALITY TO OPERATE IT. AS HE SPOKE, I BEGAN TO GLIMPSE IN MY MIND MY TRUE IDENTITY WHICH WAS SIMULTANEOUSLY MAGNIFICENTLY BEAUTIFUL AND FRIGHTENING. TEARS FLOWED UNCONTROLLABLY DOWN MY FACE. THIS SEEMED TO AMUSE THE SIRIAN AND THE VEGAN. I THOUGHT TO MYSELF, IF I LOOKED LIKE EITHER ONE OF YOU, I WOULD NOT LAUGH!

NEXT, I WAS STOOD UP INSTANTANEOUSLY. WEARING ONLY A SIRIAN ROBE, I COULD NOT SEE MY FEET. I FELT ELATION, LIKE I WAS ON A NATURAL HIGH. I CANNOT COMPARE THE FEELING TO ANYTHING ANOTHER CAN UNDERSTAND WITHOUT HAVING HAD THE EXPERIENCE THEMSELVES. I FELT A COMPLETE KNOWING OF ALL THINGS; A CONNECTION TO ALL BEINGS. YET, I RETAINED MY IDENTITY OF SELF. I FOLLOWED THE SIRIAN OUT THROUGH THE WALL TO WHAT APPEARED TO BE A MAROON-COLORED CAVE WITH A HIGH CEILING THAT SEEMED TO GO ON FOR MILES.

GLANCING BACK, I SAW THE CRAFT THAT I HAD JUST BEEN IN IT RESEMBLED A GIGANTIC PEARL. COMPLETELY WHITE, WITH A SOFT WHITISH-VIOLET GLOW, THERE WAS NO SOUND COMING FROM IT. INSTANTLY, I KNEW THAT IT TRAVELED THROUGH SPACE LIKE AN ELECTRON AS IT BECOMES A PARTICLES WAVE SOMEWHERE BETWEEN THE PHYSICAL AND SPIRITUAL PLANES.

IN FRONT OF ME, I SAW A LARGE GROUP OF MEN. MOST OF THEM WERE IN THEIR TWENTIES, A FEW WERE TEENAGERS, AND SOME WERE OLDER. DIGGING IN UNISON WITH SHOVELS, THEY MOVED LIKE AUTOMATONS. LARGE, MELON-HEADED BEINGS WATCHED THEM WORK. APPARENTLY, THESE WERE THEIR OVERSEERS. A FEW FOUR-FOOT TALL GREYS MOVED ABOUT. I SAW TUNNELS AND CRAFTS THAT LOOKED LIKE DISCS. SUDDENLY, ALL THE MEN STOPPED AND LOOKED UP TOWARD A PLATFORM. LOUDLY, A VOICE CALLED OUT, THE EMISSARY FROM RIGEL WILL SPEAK NOW.

AT THAT MOMENT, A FIVE-FOOT BEING WITH ROUND, BLACK EYES APPEARED ON THE PLATFORM. HIS LARGE HEAD WAS DRAPED WITH A KIND OF SHORT CURTAIN AROUND THE BACK OF IT. DRESSED ALL IN BLACK, HIS CLOTHING APPEARED TO BE A UNIFORM. LOOKING OMINOUS, HE CARRIED A ROD THAT LOOKED BOTH METALLIC AND CRYSTALLINE AT THE SAME TIME.

EVERYONE SEEMED TO UNDERSTAND HIS MENTAL COMMUNICATION. EXPLAINING THAT THIS GROUP OF MEN HAD COMPLETED THEIR SERVICE ON EARTH AND THE MARTIAN OUTPOST, HE TOLD THEM THAT THEY WOULD NOW BE EXAMINED FOR POSSIBLE TRANSPORT TO RIGEL FOR EXPERIMENTATION. THOSE NOT USED WOULD BE ELIMINATED. THE MEN WERE THEN LED INTO THE OPEN PORT OF A SILVER DISC. I NOTICED THAT THEIR LEGS WERE CHAINED TOGETHER. THE ENTIRE SCENE APPEARED TO BE ORCHESTRATED JUST FOR ME. I DO NOT KNOW IF IT WAS. IF SO, WHAT WOULD BE THE PURPOSE? PERHAPS THE ENTIRE INCIDENT IS AN IMPLANTED MEMORY.

THE NEXT THING I REMEMBER, I WAS BACK ON THE SIRIAN SHIP. TRAVELING IN HYPERSPACE, I SAT IN A DARKENED ROOM ON THE FLOOR WITH THE VEGAN AND THE SIRIAN. SOFT LIGHTING ALLOWED ME TO SEE INSIDE AND OUTSIDE THE SHIP SIMULTANEOUSLY. HYPERSPACE WAS A BEAUTIFUL SHADE OF DARK BLUE AND VIOLET. THE SHIP APPEARED TO BE DIAMOND-SHAPED AS IT TRAVELED. THE SIRIAN SAID THAT WE WERE GOING BACK TO OUR HOME WORLD, KHOOM, WHICH ORBITED SIRIUS A IN THAT BINARY STAR SYSTEM. LATER, I COULD DECIDE IF I WANTED TO RETURN TO EARTH. I REALIZED WE THREE WERE THE ONLY BEINGS ABOARD THAT GREAT SHIP. OVERRATED

TOTALLY BY OUR WILL, I FELT COMPLETELY SAFE AND AT PEACE. I KNEW THAT I WAS GOING HOME.

I DO NOT REMEMBER MUCH OF MY STAY ON KHOOM. I WAS TOLD THAT MY MEMORY WOULD RETURN AT THE APPROPRIATE TIME. A FROZEN WORLD COVERED IN ICE AND SNOW, THE SUNS GLARE ON THE SURFACE WAS BLINDING. THE INHABITANTS LIVED UNDERGROUND. AN ELABORATE AND IMPENETRABLE DEFENSE SYSTEM PROTECTED THE ENTIRE PLANET.

ONCE A SUBTROPICAL WORLD, KHOOM WAS PUSHED OUT OF ITS ORIGINAL ORBIT BY WAR EONS AGO. FROM SPACE, IT LOOKS BLUE AND WHITE. THE PLANET HAS NO MOONS. A VICTIM OF THE ORIGINAL BATTLE BETWEEN GOOD AND EVIL, KHOOM WAS BLOWN FROM ITS CRADLE ORBIT BY THOSE WHO CREATED THE DRACO RACE, THE REPTILIANS WHO SEEK TO DOMINATE THE GALAXY AND BEYOND. THE CREATORS OF THE DRACO RACES CAME FROM ANOTHER TIME AND SPACE. THE FIRST GENESIS OF LUCIFER, THEIR NAME DENOTES THE EPITOME OF EVIL.

OCCUPYING THE SAME PHYSICAL LOCATION AS KHOOM, BUT AT A DIFFERENT VIBRATORY RESONANCE, IS A NON-PHYSICAL WORLD GOVERNED BY A COUNCIL OF NINE BEINGS KNOWN AS THE OHALU COUNCIL. WHILE NOT THE SAME BEINGS REFERRED TO AS THE NINEIN OTHER LITERATURE, THEY DO COMMUNICATE WITH THE NINE AS WELL AS PARTICIPATE IN JOINT PROJECTS. THE NINE FIRST APPEARED IN LITERATURE IN THE WORKS OF DR. ANDRIJA PUHARICH. IN HIS WORK WITH CHANNELERS, HE CAME ACROSS A FEW INDIVIDUALS WHO CLAIMED TO BE IN CONTACT WITH THIS ET GROUP.

URI GELLER WAS THE FIRST TO ACTUALLY IDENTIFY THEM AS FORMER PHYSICAL BEINGS WHO TRANSFERRED THEIR MINDS AND SOUL-ESSENCES INTO NINE ADVANCED COMPUTERS. EACH ONE OF THESE COMPUTER REPRESENTS A DIFFERENT ASPECT OF THE MIND OF GOD. THE NINE COMMUNICATE WITH A FEW SELECT INDIVIDUALS ACROSS THE PLANET IN AN EFFORT TO UPGRADE THE COLLECTIVE CONSCIOUSNESS OF HUMANKIND.

THE OHALU COUNCIL ALSO GOVERNS THE PLANET KHOOM IN THE STAR SYSTEM SIRIUS SINCE THE SIRIANS ARE REALLY THE LOWER VIBRATIONS OF THE COUNCIL BEINGS IN THE SAME WAY THAT HUMANS ARE LOWER VIBRATIONS OF THEIR ET SELVES.

I WAS ORIGINALLY SENT TO EARTH BY THE OHALU COUNCIL WHO DIRECTED THE SIRIANS ON THE CREATION OF MY PHYSICAL BODY. THEY TOLD ME THERE ARE NINE BEINGS ON EARTH WHO ARE LIKE ME. EACH ONE IS DIRECTED BY A MEMBER OF THE COUNCIL. I WAS SHOWN MY PAST AND FUTURE WHILE I WAS ON KHOOM. I WAS TAKEN TO A PLANET ORBITING SIRIUS B THAT WAS A TROPICAL, SWAMPY JUNGLE WORLD OCCUPIED BY SHORT, STOCKY BEINGS WHO LIVE IN HUTS.

EXTREMELY ADVANCED, THESE CREATURES CAN ASTRAL PROJECT ANYWHERE THEY WANT TO AT WILL. THEY RELY ON OTHERS FOR PHYSICAL TRANSPORTATION OFF-WORLD BUT AS THEY HAVE NO NEED TO GO ANYWHERE ELSE, THEY RARELY DO SO. COMMUNICATING EXCLUSIVELY BY MIND-LINKING, THEY DO NOT HAVE A SPOKEN LANGUAGE.

THE SIRIANS TOLD ME OF THE COMING EARTH INVASION BY THE DRACO; THAT THE ORION CONFEDERATION WAS WORKING WITH THE DRACO; AND THAT THERE IS A WAR GOING ON RIGHT NOW BETWEEN THE SIRIANS AND ORIONS. THE SUPREME MERCHANTS OF THE UNIVERSE, THE SIRIANS ACTUALLY SUPPLIED THE ORION GROUPS WITH THE WEAPONS THAT ARE NOW BEING USED AGAINST THEM. HOWEVER, THE SIRIANS KEEP THE BEST AND MOST POWERFUL FOR THEMSELVES, SO THEY NEVER LOSE.

THE SIRIANS SEE THE ORIONS AS BAD CHILDREN WHO PLAY WITH MATCHES. THEY DO NOT SEEK TO DESTROY THEM, BUT THEY KEEP THEM IN CHECK. THEY ALLOW HUMANS, ORIONS, AND THE DRACO TO FOLLOW THEIR OWN DESTINIES.

DESPITE NOT WANTING TO LEAVE, THE OHALU COUNCIL CONVINCED ME TO RETURN TO EARTH. ARRIVING BACK IN ISRAEL I FOUND THAT ONLY THREE DAYS HAD PASSED. I WAS UNCEASINGLY THIRSTY, HAD LOST TEN POUNDS, AND WAS SO TIRED THAT I COULD NOT SEE STRAIGHT. RETURNING TO HOLON, I SLEPT FOR ALMOST AN ENTIRE WEEK.

DURING MY RETURN REST, I REALIZED THAT THE ARK OF THE COVENANT IS ACTUALLY A COMMUNICATION DEVICE BETWEEN THE PHYSICAL WORLD AND HYPERSPACE. THE SIRIANS APPARENTLY GAVE THE HEBREWS INSTRUCTIONS FOR BUILDING IT. A RECONSTRUCTION OF THE DEVICE WAS BUILT BY THE UNIVERSITY OF MINNESOTA YEARS AGO, FOLLOWING THE INSTRUCTIONS GIVEN IN THE BIBLE. IT WAS SO ELECTRICALLY CHARGED AND DANGEROUS THAT IT HAD TO BE DESTROYED.

I BELIEVE THAT THE SIRIANS ARE TRYING TO UNDERMINE THE PLANS OF ALL THE FACTIONS INVOLVED ON EARTH:

- THE NEW WORLD ORDER
- THE DRACO
- THE GREYS
- THE TALL BLONDS, ETC.

THEIR AGENDA IS TO BRING ALL EVENTS TO A CLIMAX, THEN USURP ALL POWER, POSSIBLY VIA THE ISRAELIS. THIS IS ONLY SPECULATION ON MY PART.

TIME WILL TELL.

P. 130 - 132

THAT NIGHT, THE SESSION WAS MOST INTERESTING. AFTER I WAS ENTRANCED, A SMALL WHITE BEING, WHO LOOKED LIKE A TYPICAL GREY, ENTERED MY BODY. SPEAKING AT FIRST IN A STRANGE LANGUAGE THAT ONLY DUNCAN UNDERSTOOD, IT THEN SPOKE IN ENGLISH AS THE TONE BECAME MORE OMINOUS. THIS BEING SAID THAT IT HAD EVERY RIGHT TO TAKE POSSESSION OF MY BODY BECAUSE I WAS ONE OF THEM!

CHALLENGING THIS REMARK, PRESTON SAID THAT I WAS A HUMAN BEING WITH A SOUL FROM GOD, AND NO ONE WAS ALLOWED TO USE THE BODY EXCEPT STEWART. THE ENTITY CURSED PRESTON AND CALLED HIM PRESSED ON. IT SAID I WORKED FOR THEM AND WAS CARRYING OUT A MISSION VITAL TO THE SUCCESS OF THEIR PROGRAM ON EARTH.

BOTH PRESTON AND DUNCAN SAW THE PHYSICAL SHAPE OF MY BODY CHANGE AS IF IT WERE A GREY ALIEN BODY. THE OUTLINE OF MY FACE SIMILARLY CHANGED. THEN, THE BEING STARTED TO MOVE MY BODY. GETTING UP, IT WALKED AROUND WHILE MAKING NASTY COMMENTS ABOUT THOSE PRESENT. ALTHOUGH MY EYES WERE CLOSED, MY BODY WALKED AROUND THE ROOM AS IF IT WERE WIDE AWAKE.

WHEN PRESTON STARTED TO ASK QUESTIONS ABOUT THEIR AGENDA, THE BEING HESITATED. NEXT, AN EXTREMELY POWERFUL ENTITY LITERALLY PULLED THE GREY FROM MY BODY AS IT TOOK OVER. IDENTIFYING ITSELF AS A DRACO COMMANDER, IT GAVE ITS NAME AS GENGEEKO. PRESTON IMMEDIATELY UNDERSTOOD THE DRACO TO BE POWERFUL, REPTILIAN WARRIORS. THE CREATURE TOLD PRESTON THAT AN INVASION FORCE WAS ON ITS WAY TO EARTH AND THAT NOTHING COULD STOP IT.

THE MOON ORBITING THE EARTH WAS THEIR FIRST CRAFT. IT HAD ARRIVED HERE AEONS AGO TO CONTROL THE PLANET. AFTER CREATING THE LEMURIAN CIVILIZATION, THEY HAD BEEN REMOVED FROM THE EARTH BY THE ATLANS AND THE DESCENDENTS OF THE DISBANDED LYRAEN EMPIRE WITH THE HELP OF THE PLEIADIANS. NOW, THE DRACO WERE RETURNING TO RECLAIM THE EARTH AND USE IT AS A MILITARY BASE FOR ENTRY INTO THE REST OF THE GALAXY. AT THIS POINT, I REALIZED THAT THIS WAS WHY SO MANY RACES WERE INTERESTED IN THE EARTH. IF THIS PLANET FALLS, THEN THE REST OF THE GALAXY IS IN DANGER.

THE DRACONIAN THEN STOOD UP AND RASPED A WARNING AT PRESTON NOT TO USE HIS EQUIPMENT OR PLEIADIAN CONTACTS TO TRY TO STOP THEM. NEXT, HE PHYSICALLY ATTACKED PRESTON!

BOTH DUNCAN AND PRESTON HAD TO RESTRAIN THIS BEING IN MY BODY UNTIL IT FINALLY SAT DOWN AND RESUMED ITS SPEECH. CLAIMING THAT HUMANS WERE WEAK, IT SAID THAT HUMANS NEEDED THE ORDER THAT AN INVASION WOULD BRING. THIS WAY, THE INVASION WOULD BENEFIT EVERYONE. THE DRACO WOULD RECEIVE THE RAW MATERIALS, WORKERS, AND FOOD THAT THEY NEEDED FOR THEIR INVASION INTO THE REST OF THE GALAXY. THE EARTH WOULD BE PROTECTED FOREVER BY THE DRACO EMPIRE.

OUR LEADERS WERE WELL AWARE OF THE IMPENDING INVASION, GRADUALLY PREPARING THE WORLD POPULATION VIA TELEVISION SHOWS AND MOVIES. EVEN RULERS IN SOME COUNTRIES WERE HUMANS WITH DRACO SOUL-PERSONALITIES.

THE REPTILIAN WITHIN MY BODY EXPANDED UPON HIS IDEAS BY SAYING THAT THE UNITED NATIONS WOULD BE THE FORUM FOR A CENTRAL PLANETARY GOVERNMENT. UNITED STATES LEADERS WERE IN LEAGUE WITH DRACO ALLIES WITHOUT REALIZING IT. SOME OF THE LEADERS OF THIS PLANET HAD PREPARED ESCAPE PLANS TO MARS WHERE EQUIPMENT WAS ALREADY BEING ACTIVATED, AS WELL AS TO OTHER PLANETS AND MOONS IN THIS SOLAR SYSTEM. MARS HAS A HUGE UNDERGROUND FACILITY BUILT BY THE SIRIANS OVER 500,000 YEARS AGO.

GENGEEKO SAID THAT MY BODY COULD BE USED SINCE I WAS ONCE AN AMBASSADOR TO THEIR HOME WORLD FROM THE OHALU COUNCIL, THE RULER OF THE SIRIAN STAR SYSTEM. I WAS NEUTRAL AND NOT REALLY CONCERNED ABOUT WHO WAS IN POWER. THIS IS BASICALLY TRUE FOR ME; I AM NOT A POLITICAL PERSON. MY MAIN INTEREST LIES IN HELPING INDIVIDUALS TO ADVANCE THEIR SOULS AND MINDS.

CONTINUING, GENGEEKO STATED THAT I WAS CREATED WITH ALIEN GENETICS AND THAT MY SOUL-PERSONALITY WAS NOT HUMAN NOR FROM THIS STAR SYSTEM. THIS EXPLAINS WHY I HAVE SO MANY MENTAL ABILITIES THAT ARE CONSIDERED UNCONVENTIONAL ON EARTH.

WHEN I CAME OUT OF THIS TRANCE SEVERAL HOURS LATER, MY BODY WAS DEHYDRATED AND EXTREMELY COLD. DISORIENTED AND CONFUSED, I DID NOT RETURN TO NORMAL FOR ALMOST TWO DAYS. IN ADDITION, I SUFFERED RECTAL BLEEDING, SHORTNESS OF BREATH, AND SEVERE SHIVERING.

OVERALL, IT WAS NOT A PLEASANT EXPERIENCE.

GO BACK

GO BACK TO THE MONTAUK PROJECT

17 - BREAKTHROUGH

P. 137 - 139

THE NEXT TIME PRESTON USED WILHELM REICH PROCEDURES, HE BROUGHT HIS TAPE RECORDER.

AN ENTITY CALLING ITSELF TUBOR ENTERED MY BODY WITH SUCH FORCE THAT I SHUDDERED HARD AND NEARLY FELL TO THE FLOOR. CLAIMING TO BE THE DRACO CONTROLLER OF A MISSION TO PREPARE EARTH FOR OCCUPATION, TUBORS MAIN CONCERN WAS OTHER ALIEN AND HUMAN INFLUENCES ON THE POPULATION THAT MIGHT THWART THEIR PLANS. THESE REPTILIAN BEINGS WERE EXTREMELY NASTY AND HOSTILE TOWARD ANYONE WHO QUESTIONED THEM. TUBOR COMMENTED THAT IT WAS DISGUSTED WITH MY BODY AND DETESTED THE WAY A HUMAN FELT. HUMANS WERE CONSIDERED TO BE WEAK, FRAGILE, AND TOO PRONE TO EMOTIONAL REACTIONS.

INSISTING THAT I HAD A CONTRACT WITH THEM WHICH PERMITTED THEM TO USE MY BODY BEFORE THEIR OFFICIAL ARRIVAL, TUBOR SAID THAT I WAS ALSO DESTINED TO BE A LIAISON DURING THE INVASION. USING A HUMAN IN THIS WAY WOULD ALLOW BOTH SIDES TO UNDERSTAND THE MIND-SET OF THE OTHER. I HAD SUPPOSEDLY AGREED TO THIS BECAUSE I WAS ONCE THE SIRIAN AMBASSADOR TO THEIR HOME WORLD AS WELL AS TO ARCTURUS AND A PLANET CALLED UMO.

AS THE SIRIAN AMBASSADOR, I HAD SUCCESSFULLY NEGOTIATED A TECHNOLOGY CONTRACT BETWEEN THE DRACO AND THE SIRIANS, SIRIUS A AGREED TO PROVIDE THE DRACO WITH HIGH TECHNOLOGY IN EXCHANGE FOR FREE TRADE AND PASSAGE OF THEIR VESSELS ANYWHERE WITHIN THE DRACO EMPIRE. THIS CONTRACT UPSET THE ORION CONFEDERATION BECAUSE, BEING CONTROLLED BY THE DRACO, THEY WOULD NEVER BE FREE. THE SIGNING OF THIS CONTRACT

EVENTUALLY PROVOKED A WAR BETWEEN THE ORION CONFEDERATION AND SIRIUS A WHICH CONTINUES TO THIS DAY.

TUBOR ALSO TOLD US THAT THE MONTAUK PROJECT EMPLOYED SIRIAN TECHNOLOGY. THE DRACO OBSERVED ALL EXPERIMENTS BUT WERE ESPECIALLY INTERESTED IN THE ONES THAT INVOLVED GENETIC MANIPULATION. BECAUSE THEY ARE AN ANDROGYNOUS RACE, THE DRACO ARE EXCEPTIONALLY INTERESTED IN SPECIES THAT PROCREATE SEXUALLY. USING SEXUALITY AS A MEANS OF PROGRAMMING PEOPLE AND GAINING MIND CONTROL OVER THE MASSES WAS PARTICULARLY FASCINATING TO THEM.

WHENEVER PRESTON ASKED QUESTIONS OR STATED AN OPINION, TUBOR BECAME ANGRY. HE CALLED PRESTON PRESSED ON AND REFERRED TO DUNCAN AS DUNK CAN. AT TIMES, TUBOR HISSED AND SWUNG HIS ARMS TOWARD THEM. TUBOR EVEN TRIED TO HURT MY BODY BY HITTING OR TWISTING PARTS OF IT. ABOUT AN HOUR LATER, ANOTHER ENTITY PUSHED INTO ME AND THREW TUBOR SPINNING AWAY. AFTER FALLING LIMP TO THE FLOOR, MY BODY THEN STOOD UP AGAIN.

CLAIMING TO BE A SIRIAN, THIS BEING GAVE ITS NAME AS MISHKA. HE SAID THAT HE FORCED HIS WAY INTO THE BODY TO SHOW ME HOW TO DEFLECT HOSTILE USE OF MY BODY. SINCE I HAD SIRIAN DNA, I COULD EASILY LEARN HOW TO DEFEND MYSELF USING MY ABILITIES. MISHKA SAID THAT HE LIVED ON A LARGE SIRIAN SPACE STATION CALLED CALUMBA WHICH ORBITED BETWEEN EARTH AND MARS.

THIS STATION WAS DESIGNED TO MONITOR INTERFERENCE ON EARTH AND SURROUNDING PLANETS. THE SIRIANS WERE INDEPENDENT AND CONSIDERED TO BE THE MERCHANTS OF THE UNIVERSE. WITH THE HIGHEST TECHNOLOGY AVAILABLE IN THEIR POSSESSION, MANY RACES CAME TO THEM FOR HELP. EVEN ALL THE WEAPONS USED BY THE DRACO CAME FROM SIRIUS. WHEN PRESTON REMARKED THAT THIS MADE THEM GUILTY OF HURTING OTHERS, MISHKA HAD AN ANSWER. HE REPLIED THAT WITHOUT SIRIAN WEAPONS, THE DRACO WOULD USE BRUTAL FORCE TO OVERWHELM THEIR TARGETS. IN THIS WAY, THE SIRIANS CONSIDERED THEMSELVES TO BE A MITIGATING FORCE IN THE GALAXY.

OVER THE NEXT SEVERAL MONTHS, MISHKA, MISHKAS ASSISTANT MARSHAK, TUBOR, AND GENGEEKO USED MY BODY TO DELIVER WARNINGS AND ULTIMATUMS. I WAS GIVEN INFORMATION THAT THE USSR WAS IN LEAGUE WITH THE DRACO AND ALLOWING THEM TO USE SOVIET BASES FOR ADVANCE OPERATIONS. BUT, THE USSR WOULD EVENTUALLY BREAK UP INTO SMALLER NATIONS AND DISRUPT THE AGREEMENT.

WHEN THIS HAPPENED, I WAS TOLD THAT THIS WAS A DECEPTION TO LULL THE REST OF THE WORLD INTO A FALSE PEACE. THE VARIOUS SOVIET GOVERNMENTS WERE CLOSELY ALIGNED WITH EACH OTHER. WHEN THE OPPORTUNITY WAS RIGHT, THE SOVIETS WOULD POUNCE ON THE UNSUSPECTING COUNTRIES.

IN THIS WAY, THE DRACO HAD A POWERFUL ALLY ON EARTH TO DO THEIR DIRTY WORK FOR THEM.

P.142 143

THE REST OF THE CONVERSATION INVOLVED INFORMATION ABOUT THE SIRIAN INTEREST IN ISRAEL AND THE JEWISH PEOPLE. THE JEWS, HE SAID, WERE CREATED AS A JOINT EFFORT BETWEEN THE SIRIANS AND THE DRACO. THE GENETIC STOCK CAME FROM THE HEBREWS, A RACE OF SIRIAN

ORIGIN. THE OHALU COUNCIL, COMPRISED OF THE LEADERS OF THE SIRIAN CIVILIZATION, PROVIDED THE TORAH (THE FIRST FIVE BOOKS OF MOSES, I.E., THE OLD TESTAMENT OF THE BIBLE) WITH ITS CODED INFORMATION THAT I WOULD LEARN TO DECIPHER.

IN FACT, THE OHALUANS ARE NONPHYSICAL BEINGS WHO REMAIN IN HYPERSPACE. THE SIRIANS ARE THEIR PHYSICAL DESCENDENTS. ANCIENT HEBREW IS THE LANGUAGE OF THIS COUNCIL. A HOLY LANGUAGE, IT COMES DIRECTLY FROM THE MIND OF GOD. EACH LETTER IS A SYMBOL, NUMBER, ARCHETYPE, AND GEOMETRIC SHAPE THAT IS USED TO TRANSLATE FROM THE SPIRITUAL INTO THE PHYSICAL.

IN FACT, EACH CHAPTER OF THE OLD TESTAMENT CAN BE DECODED TO LETTER PATTERNS. WHEN RECURRING PATTERNS ARE MATCHED AGAINST ONE ANOTHER, GEOMETRIC SHAPES ARE FORMED.

THESE INCLUDE A DOUGHNUT SHAPE, DIAMONDS, THREE-DIMENSIONAL TRIANGLES, ETC. WHEN ALL THE PATTERNS ARE ENFOLDED ON ONE ANOTHER, A TETRAHEDRON IS FORMED IN MULTIDIMENSIONALITY. THAT IS TO SAY, THE SHAPE CAN ONLY BE DEMONSTRATED ON PAPER BY DRAWING A THREE-DIMENSIONAL TETRAHEDRON. HOWEVER, ITS ACTUAL SHAPE GOES FAR BEYOND ANYTHING THAT CAN BE SHOWN GRAPHICALLY WITH CURRENT TECHNOLOGY. ALL LETTERS OF THE HEBREW ALPHABET CAN BE SEEN WITHIN THE SHAPE OF THE TETRAHEDRON.

IN THE ANCIENT HEBREW ALPHABET THERE ARE FOUR LETTERS THAT HAVE A STYLIZED CROWN ON TOP OF THEM. NO ONE KNEW WHY. HOWEVER, IF ALL OF THE OTHER LETTERS WERE SOMEHOW FORGOTTEN, EXCEPT FOR THOSE FOUR SPECIAL ONES, BY USING THESE FOUR, THE TETRAHEDRON COULD BE RECONSTRUCTED AND ALL THE OTHER LETTERS AGAIN FOUND WITHIN THE SHAPE.

SCIENTISTS IN JERUSALEM AND NEW YORK ARE ONLY NOW REALIZING THE INFORMATION ENCODED WITHIN THE OLD TESTAMENT AS THEY REVIEW IT BY COMPUTER AND RESEARCH THE VARIOUS LETTER PATTERNS. WHAT TYPE OF MIND COULD HAVE CREATED SUCH A TIMELESS DOCUMENT? CERTAINLY NOT A HUMAN ONE.

AT THE CONCLUSION OF OUR CONVERSATION, ELSINOB TOLD ME TO RETURN TO BED. WALKING BACK TO MY ROOM, I NOTICED THAT THE GLOW SUDDENLY STOPPED. THE MOMENT MY HEAD TOUCHED MY PILLOW, I WENT INTO A DEEP SLEEP. AT 4:45 A.M., THE ALARM RANG. WHEN I AWOKE, I REMEMBERED EVERYTHING. THAT IS WHEN I BECAME AFRAID.

P.156-158

I ALSO LEARNED THAT THERE WERE ORIGINALLY TWELVE ROOT RACES ON THE EARTH AT THE BEGINNING OF LIFE ON THIS PLANET. WHAT THIS MEANS IS THAT THERE WERE TWELVE ORIGINAL ALIEN RACES THAT AGREED TO GENETICALLY MANIPULATE AND SEED LIFE ON THIS PLANET AS PART OF A GREAT EXPERIMENT. THE PURPOSE WAS TO DETERMINE IF ALL OF THE MAN FREQUENCIES IN THE GALAXY COULD LIVE TOGETHER HARMONIOUSLY OR IF THEY WOULD DESTROY ONE ANOTHER.

THESE TWELVE ALIEN RACES MONITORED THEIR CONTRIBUTIONS TO THE EXPERIMENT OVER THE MILLENNIA. SOME LOST INTEREST WHILE SOME COMPLETELY ALTERED THEIR ORIGINAL IDEAS. IN ANY EVENT, THE TWELVE RACES ARE RETURNING TO REMOVE OR AID THEIR PART OF THE PROJECT BEFORE INVASION FORCES ARRIVE ON EARTH AND USURP ALL OF THE RESOURCES AND PEOPLE. THE OUTCOME OF THIS INVASION IS YET TO BE DETERMINED.

MANY OF THE ORIGINAL RACES KEEP BASES IN THIS SOLAR SYSTEM TO WATCH THE EARTH.

*

A SIRIAN BASE ON MARS HAS EXISTED FOR MANY THOUSANDS OF YEARS

*

THE AMPHIBIANS HAVE BASES ON NEPTUNE AND ON TITAN, A MOON OF SATURN

*

THE DRACO, WHO ARE THE INVADING REPTILIANS, HAVE BASES ON VENUS AS WELL AS UNDER THE EARTH

*

THE PLEIADEANS HAVE A BASE ON A MOON OF JUPITER

*

MANY OTHERS MAINTAIN PLATFORMS OR STATIONS IN ORBIT AROUND THE EARTH AND OTHER PLANETS IN THIS SOLAR SYSTEM

*

MOST OF THEM DO NOT WANT TO BE DISCOVERED YET

IN THE EARLY 1990'S, I WAS CONTACTED BY A MAN IN LOS ANGELES ON BEHALF OF MARINA POPOVICH, THE SOVIET COSMONAUT WHO WAS ON A LECTURE TOUR IN THIS COUNTRY. SECRETLY, SHE WAS INVESTIGATING THE DISAPPEARANCE OF A SOVIET SPACECRAFT THAT HAD BEEN ON A RECONNAISSANCE MISSION TO MARS. WHILE APPROACHING PHOBOS, A MARTIAN MOON, A STRANGE-LOOKING CRAFT EMERGED FROM BEHIND THAT MOON, FIRED SOMETHING AT THE RUSSIAN PROBE, AND THE SHIP WAS NEVER HEARD FROM AGAIN.

MS. POPOVICH KNEW MY FAMILY HISTORY IN RUSSIA (REMEMBER, MY GREAT-UNCLE WAS THE FIRST PRESIDENT OF THE SOVIET UNION), AND SHE KNEW ABOUT MY BACKGROUND AND TRAINING. SHE WAS INTERESTED TO HEAR WHAT I HAD TO SAY ABOUT THE MISSING RUSSIAN CRAFT.

AFTER SEVERAL ATTEMPTS TO CONVERSE VIA PHONE, MARINA WAS SUDDENLY SENT BACK TO MOSCOW. TO MY KNOWLEDGE, SHE NEVER RETURNED. I NEVER HAD THE CHANCE TO TELL HER THAT A SIRIAN VESSEL DESTROYED THE SOVIET PROBE ON BEHALF OF THE U.S. GOVERNMENT AND THE ISRAELIS! APPARENTLY, THERE WERE THOSE PEOPLE WHO DID NOT WANT THE WORLD TO FIND OUT ABOUT THE MONUMENTS ON MARS.

THESE SAME PEOPLE DID NOT WANT THE PUBLIC TO KNOW THAT AMERICANS WERE ALREADY WORKING WITH BEINGS FROM TAU CETI WHO WERE THE PROGENITORS OF THE SLAVIC RACES. ALTHOUGH THE SIRIANS, TAU CETIANS, RUSSIANS, ISRAELIS, AND AMERICANS WERE THEORETICALLY ALL WORKING TOGETHER AGAINST THE ORION CONFEDERATION AND THE DRACO EMPIRE, EACH HAD THEIR OWN AGENDA AND WITHHELD INFORMATION FROM EACH OTHER. THE AMERICANS AND THE ISRAELIS WANTED PRIMARY CONTROL SO THEY CAN EVENTUALLY DOMINATE THE WORLD.

THEY REASONED THAT THE GOVERNMENT WITH THE MOST EXCLUSIVE INFORMATION AND TECHNOLOGY WOULD WIN.

MONTAUK: THE ALIEN CONNECTION REVEALS THE MOST AMAZING STORY YET TO SURFACE IN THE AREA OF ALIEN ABDUCTION. THIS IS AN AUTOBIOGRAPHICAL AND FACTUAL ACCOUNT FROM STEWART SWERDLOW, A GIFTED MENTALIST WHO WAS BORN CLAIRVOYANT BUT HAUNTED BY STRANGE TIME-SPACE SCENARIOS. AFTER SUFFERING ALIEN ABDUCTIONS AND GOVERNMENT MANIPULATIONS, STEWART FOUND PRESTON NICHOLS AND DISCOVERED HIS OWN ROLE IN TIME

TRAVEL EXPERIMENTS KNOWN AS THE MONTAUK PROJECT.

AFTER REFUSING TO BREAK HIS ASSOCIATION WITH NICHOLS, STEWART WAS INCARCERATED BY THE AUTHORITIES, BUT THE TRUTH BEGAN TO REVEAL ITSELF. STRUGGLING FOR HIS LIFE, STEWART USED HIS MENTAL ABILITIES TO OVERCOME THE NEGATIVE INFLUENCES SURROUNDING HIM AND ULTIMATELY DISCOVERED THE HIGHEST COMMON DENOMINATOR IN THE ALIEN EQUATION AN INTERDIMENSIONAL LANGUAGE WHICH COMMUNICATES TO ALL CONSCIOUS BEINGS.

P.205

VICTIM MENTALITY

THIS PLANET PRIMARILY ATTRACTS TWO TYPES OF BEINGS THOSE WITH A VICTIM MENTALITY AND THOSE WITH AN OPPRESSOR MENTALITY. IN ORDER TO BE A VICTIM, YOU MUST HAVE OPPRESSORS. IN ORDER TO BE AN OPPRESSOR, YOU MUST HAVE VICTIMS. TOGETHER, THIS CREATES BALANCE IN GOD-MIND.

THIS PLANET IS DESIGNED AS A PLACE WHERE BEINGS WITH A VICTIM MENTALITY CAN LEARN TO OVERCOME IT. THEREFORE, ANY 'HIGHER LEVEL' BEING WILL NOT BECOME INVOLVED BECAUSE IT IS NECESSARY FOR THOSE HERE TO HAVE THE OPPORTUNITY TO WORK THROUGH THEIR VICTIM MENTALITIES. TO BECOME INVOLVED WOULD MEAN INTERFERING OR TAKING AWAY THE LESSONS OF THOSE WITH VICTIM MENTALITIES. HIGHER LEVEL BEINGS WILL BE OBJECTIVE OBSERVERS WHO WILL GUIDE AND INSTRUCT YOU THROUGH YOUR VICTIM MENTALITY ONCE YOU REACH THEM, BUT THAT IS ALL.

THEY ARE NOT HERE TO "SAVE" YOU OR THIS PLANET.

ANY BEING WHO SAYS THAT THEY ARE HERE TO "SAVE" YOU IS INTERFERING IN YOUR SOUL GROWTH. ONLY YOU CAN "SAVE" YOURSELF FROM A VICTIM MENTALITY. ONLY YOU CAN "SAVE" YOURSELF FROM A VICTIM MENTALITY. ONLY YOU AND THE LESSONS YOU SELF-DESIGN WILL TEACH YOU ABOUT VICTIM MENTALITY. ONLY YOU CAN MOVE THROUGH THESE LESSONS THAT WILL ALLOW YOU TO MAKE A FINAL RELEASE OF YOUR VICTIM MENTALITY. NO ONE IS WAITING TO HELP YOU DO NOT BE FOOLED BY ANY BEING MAKING SUCH PROMISES.

WHEN YOU LEARN TO MOVE THROUGH YOUR VICTIM MENTALITY, YOUR MIND PATTERN WILL NO LONGER ATTRACT OPPRESSORS. THE MORE STUBBORN YOU ARE IN YOUR DETERMINATION TO HANG ONTO YOUR VICTIM MENTALITY WAYS, THE MORE INTENSE THE LESSONS THAT YOU ATTRACT TO GET YOU TO RELEASE THOSE WAYS. THIS MEANS, METAPHORICALLY SPEAKING, A BIGGER, MEANER, STRONGER, STICK UNTIL YOU FINALLY "GET IT."

YOU HAVE TO GET HURT, BELITTLED, AND FINALLY, ANGRY ENOUGH TO STAND UP FOR YOURSELF AND DECLARE THAT YOU WILL NO LONGER BE A VICTIM TO ANYONE, ANYWHERE, ANY MORE! PERHAPS WHEN YOU FINALLY HAVE HAD ENOUGH, YOU RELEASE YOUR VICTIM MENTALITY BY SAYING, "I JUST DON'T GIVE A DAM ANYMORE WHAT ANYBODY THINKS! I AM SPEAKING MY MIND REGARDLESS! I AM TAKING CARE OF MYSELF, REGARDLESS!"

SOMETIMES, PEOPLE ARE BEATEN UP SO MUCH THAT INSTEAD OF RELEASING THE VICTIM MENTALITY, THEY DECIDE TO BECOME THE OPPRESSOR WHEN THE OPPORTUNITY PRESENTS ITSELF. THIS IS WHY CYCLES REPEAT THEMSELVES. PEOPLE WHO WERE OPPRESSED FIND OTHER VICTIMS, BECAUSE THIS GIVES THEM A SENSE OF CONTROL. IN DOING SO, THEY RELIVE THEIR

EARLIER EXPERIENCES WHEN THEY FELT LIKE THEY HAD NO CONTROL. THEY BECOME THE PERPETRATOR, JUST SO THEY CAN BE THE "WINNER" FOR A CHANGE. OF COURSE, THIS IS ANOTHER IMBALANCE, OR EXTREME FLIP FROM VICTIM TO OPPRESSOR.

THE IDEAL IS TO FIND A HAPPY MEDIUM, OR BALANCE, BETWEEN VICTIM AND OPPRESSOR. VICTIM AND OPPRESSOR ARE OPPOSITE SIDES OF THE SAME COIN IN THE MIND OF GOD. IN THIS CASE, TWO INDIVIDUAL SOUL-PERSONALITIES ARE NEEDED TO MAINTAIN AN EQUILIBRIUM IN GOD MIND. ONE CARRIES THE WEIGHT OF A VICTIM, ONE CARRIES THE WEIGHT OF AN OPPRESSOR. WHAT YOU NEED TO DO IS FIND THE BALANCE WITHIN YOURSELF, SO THAT EQUILIBRIUM WITHIN THE MIND OF GOD IS MAINTAINED WITHIN ONE SOUL-PERSONALITY INSTEAD OF TWO.

YES, YOU MUST LEARN TO SPEAK YOUR MIND AND NOT ALLOW ANYONE TO WALK ON YOU. YES, YOU MUST SOMETIMES BE HARSH AND CRUEL WITH WORDS AND SOMETIMES EVEN ACTIONS. BUT WHEN THIS MUST HAPPEN, IT MUST BE DONE IN AN OBJECTIVE MANNER WHAT IS BEST FOR EVERYONE INVOLVED TO GET THE POINT ACROSS WHILE CAUSING THE LEAST AMOUNT OF PAIN.

BOUNDARIES MUST BE SET.

- WHAT ARE YOUR BOUNDARIES?
- WHO CROSSES THEM AND HOW?
- WHO IS A BELLIGERENT PERSON IN YOUR LIFE THAT YOU TRY TO AVOID BECAUSE OF THEIR BULLISH, PUSHY WAYS?
- WHY DO YOU AVOID THIS PERSON?
- WHY NOT PRACTICE LETTING GO OF YOUR VICTIM MENTALITY AND SPEAK YOUR MIND TO HIM/HER?

SO WHAT IF YOUR KNEES SHAKE SO HARD THAT YOU THINK YOU WILL FALL DOWN, AND YOUR VOICE IS SHAKY, SQUEAKY, OR BARELY AUDIBLE? SOMEWHERE ALONG THE WAY, YOU MUST STOP YOUR VICTIM MENTALITY BEFORE IT STOPS YOU. LEARN TO BE PROACTIVE INSTEAD OF REACTIVE.

WHEN YOU ALLOW OTHERS TO PUSH YOU AROUND, YOU BECOME ANGRY, SULLEN, AND INTROVERTED. THEN, YOU LASH OUT AT THE UNDESERVING WITH MISDIRECTED FRUSTRATION AND ANGER. SOMETIMES YOUR BODY BECOMES ILL BECAUSE OF ALL THAT YOU SUPPRESS, OR YOU INSULATE YOUR BODY, OR PERHAPS STOP EATING. ONE WAY OR THE OTHER, YOU SUFFER.

BUT, THERE IS A PART OF YOU THAT FEELS THAT YOU DESERVE TO SUFFER. THERE IS A PART OF YOU THAT ENJOYS THIS PAIN AND SELF-PUNISHMENT. IT FEELS GOOD TO SOME PART OF YOURSELF, OR YOU WOULD NOT DO IT. SOME PEOPLE CREATE OTHERS TO GIVE THEM PAIN, SOMETIMES PHYSICAL, SOMETIMES EMOTIONAL, SOMETIMES BOTH. SOME PEOPLE DO IT FOR THEMSELVES A SELF-CONTAINED, FULLY FUNCTIONAL, VICTIM MENTALITY UNIT. SOME PEOPLE ARE EXTREMELY SUCCESSFUL AT THIS.

WHY DO YOU FEEL SO BAD ABOUT YOURSELF THAT YOU FEEL LIKE YOU ARE MEANT TO SUFFER, TO BE ALONE, TO FEEL GUILTY, TO HAVE ILL HEALTH, TO BE OVER OR UNDERWEIGHT. DID IT START IN THIS LIFETIME, OR DID IT START BEFORE? FOLLOW THOSE FEELINGS, AND ALLOW YOURSELF TO RELEASE IT. THIS PHYSICAL REALITY IS YOUR CHANCE TO OVERCOME IT. THE MORE YOU IGNORE

THIS SITUATION, THE HARDER AND MORE INTENSE LESSON YOU WILL ATTRACT TO BEAT IT OUT OF YOURSELF. YOU CAME HERE TO LEARN NOW DO IT. QUIT WHINING MOANING, AND COMPLAINING BECAUSE THERE IS A PART OF YOU THAT ENJOYS THAT TOO.

ACHES AND PAINS CAN CREATE A LOT OF ENTERTAINMENT. YOU CAN RUN FROM DOCTOR TO DOCTOR, TRYING TO FIND A CURE FOR SOMETHING THAT WILL NEVER BE CURED AS LONG AS YOU MAINTAIN YOUR CURRENT MIND-PATTERN.

"BAD LUCK" CAN ATTRACT A LOT OF ATTENTION AND SYMPATHY FROM OTHERS. DIRE CIRCUMSTANCES WILL FORCE SOMEONE TO PAY ATTENTION TO YOU. NEVER GETTING A RAISE OR A PROMOTION IS AN EXCUSE NOT TO CLIMB HIGHER IN YOUR CAREER OR COMPANY. YOU CAN MOAN ALL YOU WANT ABOUT BEING ALONE, WITH A PARTNER WHO DOES NOT UNDERSTAND YOU, OR WITHOUT ONE AT ALL.

BUT THIS SELF-IMPOSED ISOLATION ALSO GIVES YOU TIME TO SELFISHLY TAKE CARE OF YOURSELF WITHOUT ANY INTERFERENCE. THE SAME FOR SAYING YOU WANT CHILDREN BUT ARE UNABLE TO HAVE THEM. A PART OF YOU DOES NOT WANT A SPOUSE, OR A CARING SPOUSE, OR CHILDREN, BECAUSE IF ALL OF YOU WANTED THIS AND IT WAS TRULY IN YOUR MIND-PATTERN, YOU WOULD HAVE IT!

VICTIM, VICTIM, VICTIM! LET US COUNT THE WAYS! REFUSE TO BE A VICTIM. FIND A NEW SOURCE OF ENTERTAINMENT. RELEASE YOURSELF FROM THIS MIND-PATTERN SO THAT YOU CAN MOVE INTO NEW VISTAS OF GROWTH - ONES SO VAST THAT YOU CANNOT EVEN IMAGINE OR COMPREHEND THEM. YOU MUST LET GO OF THE OLD TO MAKE ROOM FOR THE NEW. NO ONE CAN DO IT FOR YOU. NO ONE CAN "SAVE" YOU FROM YOURSELF. NO ONE IS WAITING TO HELP THIS PLANET.

ONLY YOU CAN MAKE A DIFFERENCE. ONLY YOU CAN SAVE THIS PLANET.

DO IT!

P.209

ARE YOU A MIND-CONTROL CANDIDATE?

DOES MIND-CONTROL SOUND LIKE A FOREBODING SUBJECT THAT PEOPLE IN DARK, MYSTERIOUS PLACES PRACTICE UPON UNSUSPECTING VICTIMS? DO YOU READ ABOUT OTHER PEOPLE’S CLAIMS WITH INTEREST, WHILE THINKING THAT IT REALLY HAS NOTHING TO DO WITH YOU? REMEMBER, IF YOU ARE READING ABOUT IT, IT IS IN YOUR WORLD. IF IT IS IN YOUR WORLD, IT IS A REFLECTION OF SOME PART OF YOURSELF OR IT COULD NOT EXIST. SO, ARE YOU A MIND-CONTROL CANDIDATE FOR SINISTER FORCES?

TO FIT THE PROFILE THEY ARE LOOKING FOR, YOU MUST BE CONTROLLABLE. ARE YOU? WHAT DOES YOUR MIND-PATTERN LOOK LIKE? ARE YOU IN CONTROL OF YOUR OWN LIFE, OR IS YOUR LIFE IN CONTROL OF YOU?

IF ANYTHING, ANYWHERE CONTROLS YOU, YOU HAVE AN ESTABLISHED MIND-PATTERN THAT ANY FORCE CAN USE AND MANIPULATE. ALL THE FORCE HAS TO DO IS ENTER YOUR AURIC FIELD THROUGH THIS "HOLE" IN YOUR MIND-PATTERN. TO CLOSE THE HOLES AND CORRECT THE MIND-PATTERN, IT IS IMPORTANT TO FIND OUT WHAT IN YOUR LIFE IS CONTROLLING YOU SO THAT YOU CAN STOP IT.

WHAT CONTROLS YOU? IS IT A PERSON? ANYONE WHO "PULLS YOUR STRINGS" OR "PUSHES YOUR BUTTONS" IS CONTROLLING YOU. ANYONE WHO YOU ALLOW TO PUT YOU DOWN, OR WHO MAKES YOU FEEL LIKE LESS OF A PERSON IS CONTROLLING YOU. WHEN YOU REACT TO THEIR MANIPULATION, YOU ARE NOT IN CONTROL. THE OTHER PERSON CONTROLS YOU. THIS COULD BE A PARENT, SPOUSE, CHILD, RELATIVE, CO-WORKER, OR NEIGHBOR, FOR EXAMPLE.

DO YOUR EMOTIONS CONTROL YOU? WHEN EMOTIONS ARE OUT OF BALANCE, THEY CAN CREATE THEIR OWN SUBPERSONALITIES. EACH EMOTION HAS ITS OWN COLOR, TONE, AND ARCHETYPE WITHIN YOUR AURIC FIELD. ANGER IS COMPRISED OF THE COLOR RED, FEAR IS YELLOW, JEALOUSY IS GREEN. EVERY TIME YOU HAVE EXPERIENCE THESE EMOTIONS WITHOUT PASSING THEM UP TO YOUR OVERSOUL, THE ENERGY OF THESE SPECIFIC SUBPERSONALITIES ARE FED.

ANGER GETS STRONGER, FEAR GETS STRONGER, JEALOUSY GETS STRONGER. NOW, THEY ARE IN CONTROL OF YOU. YOU CAN MAKE ALL THE PROMISES YOU WANT, BUT WHAT HAPPENS WHEN YOUR BUTTONS GET PUSHED? ALL OF A SUDDEN, ANGER IS CONTROLLING YOU, OR FEAR, OR JEALOUSY. YOU HAVE LOST THE BATTLE. YOUR EMOTIONS ARE IN CONTROL.

PERHAPS IT IS ILLNESS THAT CONTROLS YOU. DO YOU NEED THAT ILLNESS IN ORDER TO FEEL IMPORTANT, OR AS A WAY TO GET ATTENTION AND/OR LOVE? DO YOU HAVE CHRONIC HEALTH ISSUES THAT NEVER SEEM TO GO AWAY? OR YOU JUST GET RID OF SOMETHING AND SOMETHING ELSE POPS UP TO TAKE ITS PLACE? WHAT PART OF YOU NEEDS AND ALLOWS THIS TYPE OF CONTROL?

DO ALCOHOL AND DRUGS CONTROL YOU? HOW ABOUT TOBACCO? SEX? THERE ARE SUPPORT GROUPS FOR ALCOHOLISM, DRUG, AND SEX ADDICTION; FOR PEOPLE WHO CANNOT STOP SHOPPING, GAMBLING, AND/OR SHOPLIFTING. THERE ARE SUPPORT GROUPS FOR PEOPLE CONTROLLED BY BULIMIA, ANOREXIA, OBSESSIVE COMPULSIVE DISORDER. YOU NAME IT, IT IS OUT THERE CONTROLLING SOMEBODY, SOMEWHERE. IS IT CONTROLLING YOU.

DOES FOOD CONTROL YOU? FOOD CONTROLLING YOU DOES NOT IMPLY THAT YOU ARE OVERWEIGHT. COULD YOU GO A DAY WITHOUT CHOCOLATE OR SUGAR? COULD YOU FAST FOR 24 HOURS WITHOUT ANYTHING BUT WATER? MOST LIKELY YOU WOULD NOT STARVE TO DEATH AND DIE, SO COULD YOU??

HOW ABOUT FASHION? ARE YOU THE FIRST TO GET THE LATEST CLOTHING, HAIRSTYLES, HOME DECOR, WITHOUT THINKING IF YOU EVEN LIKE THEM OR NOT? HOW ABOUT CARS, MOTORCYCLES TRUCKS, MOTOR VEHICLES, BOAT... THE LIST IS ENDLESS.

OR FADS, BE IT BODY-PIERCING, TATTOOS, OR ACQUIRING ALL KINDS OF DOODADS AND KNICKKNACKS THAT WILL BE USELESS PRACTICALLY BEFORE YOU EVEN HAVE THEM OUT OF THE BOX. DO YOU HAVE TO HAVE WHATEVER EVERYONE ELSE HAS.

PLASTIC SURGERY CONTROLS PEOPLE. THEY KEEP GOING BACK AND REDOING EVERY PART OF THE BODY UNTIL THERE IS NOTHING LEFT TO REDO. THEN THEY START OVER AGAIN.

SOME PEOPLE START EXERCISING AND CANNOT STOP. THEY EXERCISE CONTINUALLY AND CONSTANTLY, SOMETIMES BUILDING UP MUSCLE MASSES THAT WOULD PUT HERCULES TO SHAME. OTHERS EXERCISE UNTIL THEY DO NOT HAVE A DROP OF FAT ON THEIR BODIES, THEN BEGIN DESTROYING THEIR MUSCLE TISSUES, LIGAMENTS, AND INTERNAL ORGANS.

DO YOUR EXPERIENCES USE AND CONTROL YOU, OR DO YOU USE AND CONTROL YOUR EXPERIENCES? THIS IS THE BOTTOM LINE. IF ANYTHING OR ANYONE CONTROLS YOU IN ANY WAY, SHAPE, OR FORM, YOU ARE A PRIME CANDIDATE FOR MASS AND/OR INDIVIDUAL MIND-CONTROL. WHEN YOU ARE IN CONTROL OF YOUR OWN SELF, NO ONE AND NO THING CAN CONTROL YOU.

EVERYTHING IN MODERATION IS ACCEPTABLE, BUT WHEN SOMETHING TAKES OVER YOUR LIFE TO THE DEGREE WHERE IT CONTROLS YOU, YOU CREATE A MIND-PATTERN THAT SAYS "OUTSIDE FORCES CAN AND DO CONTROL ME."

THIS LEAVES HOLES IN YOUR MIND-PATTERN AND AURIC FIELD THAT TELLS ANYONE WHO CAN UNDERSTAND THIS INFORMATION THAT YOU ARE A PERFECT CANDIDATE FOR MIND-CONTROL. LEARN TO BREATHE YOURSELF INTO YOUR CENTRE. IDENTIFY THE STRONG ASPECTS OF YOUR MIND-PATTERN AND THE ONES THAT NEED STRENGTHENING. ANCHOR YOURSELF IN THE STRENGTH OF YOUR OWN OVERSOUL AND GOD-MIND. TAKE CONTROL OVER ABSOLUTELY ANYTHING THAT HAS CONTROL OVER YOU, REMEMBERING TO KEEP THE BALANCE. YOU CAN BE PULLED OFF-CENTRE IN EVERY WAY IMAGINABLE, AND YOU WILL BE. THESE ARE YOUR TESTS.

BE THE MONITOR OF YOUR OWN PROGRESS. THE MORE YOU USE INTERNAL MONITORS, THE LESS CHANCE OUTSIDE MONITORS CAN STEP IN TO DO THE JOB FOR YOU. WHENEVER YOU LET YOUR GUARD DOWN, THERE WILL BE SOMEBODY OR SOMETHING THAT WILL KNOWINGLY AND GLADLY PUSH YOU OUT OF YOUR CENTRE.

PART OF YOUR REASON FOR EXISTING IN PHYSICAL REALITY IS TO FIND OUT WHO AND WHAT YOU ARE. BEING PUSHED OUT OF YOUR CENTRE TELLS YOU WHO YOU ARE NOT. OPERATING FROM OUTSIDE OF YOUR CENTRE ALLOWS OTHERS TO DEFINE YOU ACCORDING TO THEIR TERMS. THEY CAN MOULD AND BEND YOU LIKE CLAY TO CREATE THE PERSON THAT THEY NEED FOR THEIR OWN PURPOSES.

YOU ARE SO BUSY MOLDING AND BENDING THAT YOU DO NOT HAVE TIME TO RECOGNIZE WHAT IS HAPPENING TO YOU.

ONLY WHEN YOU STAND FIRM IN YOUR CENTRE, IN CONTROL OF ABSOLUTELY EVERY ASPECT OF YOUR LIFE, ARE ALL THE HOLES OF YOUR MIND-PATTERN COMPLETELY SEALED SHUT. WHEN THIS HAPPENS, NO ONE OR NO THING CAN CONTROL YOU.

BUT EVEN WHEN THE DOOR IS CLOSED TO OUTSIDERS, REMEMBER THAT THEY WILL STILL COME KNOCKING.

IT IS UP TO YOU TO ENSURE THAT THEY DO NOT ENTER.

GO BACK

BEFORE THE VANCOUVER 9/11 TRIBUNAL

In re:	
Prior Knowledge of the Terror Attacks of September 11, 2001 by Defense Secretary Donald H. Rumsfeld resulting from the Quantum Access Capabilities of the US Defense-Technical Community acquired while Secretary Rumsfeld served secretly as the Defense Attaché to DARPA's Project Pegasus in the early 1970's	**AFFIDAVIT OF ANDREW D. BASIAGO**

I, ANDREW D. BASIAGO, being first duly sworn, on oath, depose and state as follows:

1. My name is Andrew Daniel Basiago.

2. I am an American citizen.

3. I was born on September 18, 1961 in Morristown, New Jersey, USA.

4. I reside in Vancouver, Washington, USA.

5. I am an attorney-at-law in private practice in the State of Washington.

6. I was admitted to the Washington State Bar Association in 1996.

7. Currently, I am an active member in good standing of the Washington State Bar Association.

8. I maintain a successful solo law practice.

9. I have never been accused of, arrested for, charged with, or convicted of a crime or been disciplined on professional ethical grounds by my bar association.

10. I am of sound mind and healthy body.

11. This affidavit is accurate to the best of my ability to remember and tell the true life experiences that I testify to in it and is made without any motive to deceive or purpose of evasion.

12. During the years 1968 to 1972, I was a child "chrononaut" in Project Pegasus.

13. A chrononaut is a time traveler.

14. Project Pegasus was a classified, defense-related, research and development program of the Defense Advanced Research Projects Agency [DARPA.]

15. DARPA is the agency that develops advanced weapon systems, communication devices, and surveillance technologies for the US Department of Defense.

16. Project Pegasus was the US time-space program at the time of the emergence of quantum access in the US defense-technical community.

17. "Quantum access" is the ability to gather information about a non-local event in the time-space continuum by either sending a chrononaut there physically or propagating a hologram of the event through which the chrononaut either views the event or visits the event on a physio-virtual basis.

18. Non-local events are those scenarios that take place in what we call the past and future but which are actually going on simultaneously in a quantum hologram that consists of a multitude of adjacent similar and dissimilar time lines.

19. The mission of Project Pegasus was to use quantum access to develop a "remote sensing in time" capability pertaining to non-local events on behalf of the Executive Branch of the United States government.

20. It was hoped that by developing better intelligence about past and future events, Project Pegasus would enable the US president, intelligence community, and military to engage in better contingency planning for future events.

21. Project Pegasus was trying to accomplish this by developing different time travel technologies in a research and development program under the aegis of DARPA.

22. These time travel technologies were advanced electro-magnetic devices, including teleporters, that would enable DARPA to send chrononauts to non-local events and advanced electro-optical devices, including chornovisors, that would enable DARPA to propagate holograms of non-local events for purposes of remote viewing or physio-virtual visitation of said non-local events.

23. A "teleporter" is an electro-magnetic device that by opening up a vortal tunnel in the quantum hologram permits a chrononaut to travel thousands of miles in time-space in seconds.

24. Teleporters were fully operational in the secret realms of the US defense-technical community by 1967-68.

25. A "chronovisor" is an electro-optical device that by driving an electro-magnetic signal through a crystal array propagates a hologram that lenses a non-local event into the laboratory.

26. Chronovisors were fully operational in the secret realms of the US defense-technical community by 1967-69.

27. I am the fifth and youngest child of the American electrical engineer and inventor Raymond Francis Basiago [1923-1990].

28. From 1966 to 1984, Raymond F. Basiago was employed as a special projects engineer for The Ralph M. Parsons Corporation of New York, New York [today, Parsons Corporation of Pasadena, California.]

29. Parsons was the principal US defense contractor involved in DARPA's Project Pegasus, so principally, in fact, that when Ralph M. Parsons, the founder of Parsons Corporation, bought what was then the world's largest private yacht from the billionaire Adnan Kashoggi in 1972, Parsons renamed the ship "Pegasus II."

30. In addition to working for Parsons as an engineer, Raymond F. Basiago secretly held the rank of Major in the Second Reserve of the US Army.

31. During his 18 years of employment with the Parsons Corporation ostensibly working on conventional engineering projects, my father was secretly employed as the technical liaison between Parsons and the CIA on the theory and practice of teleportation.

32. While it is not clear whether Raymond F. Basiago was ever a CIA officer *per se*, his work with DARPA, the CIA, and other US military and intelligence agencies was extensive enough for him to be in the CIA employee database.

33. When I inquired of my father's employment status with the CIA through an intermediary ten years ago, I was told by the CIA that they could not release information to the family members of living or deceased CIA personnel.

34. I was first brought into Project Pegasus by my father in 1967-68, when at age six I teleported with him from New Jersey to New Mexico via a teleporter located in Building 68 at the "old" Curtiss-Wright Aeronautical Company facility formerly at 1 Passaic Avenue in Wood Ridge, New Jersey.

35. This "jump" resulted in me becoming the first American child to teleport.

36. I was officially enlisted in DARPA's Project Pegasus in Fall 1969, at age seven, at the beginning of my third grade year, when I signed a loyalty and secrecy oath on US Department of Defense letterhead.

37. During 1969-72, the learning laboratory for gifted and talented school children at my public elementary school, Mountview Road School in Morris Plains, New Jersey, was used as a training course for time travel activities that I was then involved in at defense contractor locations in New Jersey and New Mexico.

38. In this educational context, I received alpha intelligence training in such abilities as perception, memory, and communication.

39. This training included photo-learning on machines called "tachistoscopes" developed by the Office of Naval Research for the education of the 140 gifted and talented American school children enrolled in Project Pegasus and other secret defense and intelligence projects utilizing US school children as "team members."

40. The curriculum that we studied on the tachistoscope was called "Galileo" and emphasized the history of society and of science from the year 1450. The purpose of this schooling was to give the children who would be time traveling for DARPA knowledge of the social and technical history of the modern era so that we would have this information available to us when we were in the field.

41. From Fall 1969 to Summer 1972, I was involved in eight different forms of quantum access for DARPA, in which I received specialized education in the learning lab at Mountview and then took part in practical exercises either in the learning laboratory or at a defense contractor location off-campus.

42. *Remote Viewing*. In Fall 1969, the children in my group were assessed for our psychic ability and then involved in remote viewing exercises for the US Navy's Office of Naval Intelligence. We were asked, *inter alia*, to remotely view the POW compound known as the "Hanoi Hilton" in North Vietnam and determine whether Lt. John McCain was in the prison and, if so, where he was in the prison.

43. *Spinning*. In Spring 1970, the children in my group took part in exercises in which we were laid on our backs on a table top-like device and spun clockwise at 33 rpm. In this way, we were trained to go "out-of-body." During one astral projection, I confronted a barrier that the lady from DARPA who debriefed me stated was the infrastructure propagating the hologram we exist in, The Matrix.

44. *The Montauk Chair*. In Spring 1970, the children in my group first began to take astral projections to episodes in our own subjective futures via the time travel technology known as "the Montauk chair." After returning from these induced astral projections to our own futures, we were then debriefed about what we saw, and in this way intelligence was gathered about future events and conditions. According to Dr. David L. Anderson, the Montauk chair "went national" and involved thousand of American schoolchildren in probes to episodes in their future lives beginning in Spring 1970. I was one of them.

45. *The Teleporter.* In Summer 1970, the children in my group first began traveling to the state capitol grounds in Santa Fe, New Mexico via the Tesla teleporter at the Curtiss-Wright Aeronautical Company in Wood Ridge, New Jersey and returning to New Jersey via the teleport located at the Sandia National Laboratory in Sandia, New Mexico, adjacent to the Kirtland Air Force Base northeast of Albuquerque, New Mexico.

46. *The Chronovisor.* In Fall 1970, the children in my group first began accessing past and future events via third and fourth-dimensional Ernetti-Gemelli chronovisors situated at the General Manufacturing Company in Convent Station, New Jersey; a public auditorium under construction in Morristown, New Jersey; and the ITT Defense Communications facility in Nutley, New Jersey.

47. *The Plasma Confinement Chamber.* In Spring 1972, my father took me to a time lab in East Hanover, NJ where, under the direction of Dr. Stirling Colgate, I entered a Lucite chamber where plasma was emitted, and while immersed in the plasma, I worm-holed to Gettysburg, PA on November 19, 1863, where I was depicted in the famous Josephine Cobb photograph of Abraham Lincoln at Gettysburg.

48. *The Stargate.* In Summer 1972, I took trips to 2045 to retrieve microfilmed summaries of intervening events being prepared by project historians in the future. The Stargate through which I jumped was located in the gymnasium at the Cerrillos Cultural Center near Cerrillos, NM. We arrived back in the present on the grounds of the Microwave Tower at the Lobo Overlook site in SW Colorado.

49. *The Jump Room.* Before I left my period of service attached to Project Pegasus in 1972, I witnessed a device, about the size of a microwave oven, through which things like documents from Washington, DC were arriving in New Mexico. I believe that it was this time travel technology from which the jump rooms were developed via which, by the late 1970's, the US government was sending personnel Mars.

50. During the three years that I was officially attached to Project Pegasus from 1969 to 1972, I spent three secret summers in New Mexico living with my father and taking part in time travel activities undertaken at both atomic research institutions like the Los Alamos National Laboratories [LANL] and the Sandia National Laboratory and smaller labs operated by DARPA in the desolate environs of New Mexico.

51. These secret summers were effectuated by my father and I teleporting to New Mexico, spending the summer there, and then at the end of the summer teleporting back to New Jersey so as to arrive on the afternoon of the morning that we had left at the beginning of the summer. In this way, the time loop created by our "quantum displacement" in New Mexico was closed and time travel was used to conceal time travel research via the propagation of a time loop and its subsequent disappearance.

52. These "secret" or "hidden" summers were the first Summer 1971, Summer 1973 accessed from Fall 1971, and the first Summer 1972. I say "first Summer 1971 and 1972" because after those summers, my father and I lived second summers, with a different set of experiences, upon returning to New Jersey.

53. Consequently, as both a child progressing farther and farther down the rabbit hole of applied quantum physics, and as my father's son, during those three summers, and for shorter periods during the school year when I was in New Mexico after having teleported there, I was frequently present at project social events, including project meetings held at the labs, meals taken at restaurants, drinks had at bars, parties enjoyed at private residences, and receptions held at hotels.

54. In that context, I met many distinguished Americans who were either project principals, or who were interested in the project's activities, or who were being brought through the project's social network so that they could be briefed about their roles in future events as individuals who had been identified as "persons of interest" related to their personal destinies as national leaders of the United States.

55. The project principals that I met included Dr. Harold M. Agnew, who served as the federal administrator of Project Pegasus during his tenure as director of LANL [1970-78]; Dr. Edward Teller, who I once met when he was trying to fix a chronovisor at LANL; Dr. Stephen J. Lukasik, when he was serving as the director of DARPA; Dr. Iben Browning, when he was serving as the director of the directorate of science and technology of the CIA [1959-74]; John A. McCone, a former director of the CIA [1960-64]; Dr. Stirling Colgate, when he was president of the New Mexico Institute of Science and Technology [NMIST]; Dr. Emilio Segre, who had a discussion with my father in the hallway at LANL about the existence of teleportation; and Donald H. Rumsfeld, the future Secretary of Defense, who served as the defense attaché of Project Pegasus.

56. In retrospect, when I consider that three of these individuals [Agnew, Teller, and Segre] were Manhattan Project physicists and that three [Agnew, Lukasik, and McCone] served as the director of major defense and intelligence agencies [LANL, DARPA, and CIA], I am struck by how important Project Pegasus was and how privileged my father and I were to be part of it and both witness and participate in the US time-space exploration program at the time of the emergence of time travel in the US defense-technical community in the early 1970's.

57. I also met several future national leaders of the United States at lunches that were held in Albuquerque, NM with project principals shortly after they were apprised of the fact that using its quantum access capability DARPA had determined that they were destined to serve as President or Vice President of the United States. These individuals included President George H.W. Bush, Vice President Richard B. Cheney, President George W. Bush, and President William J. Clinton. These lunches were held at La Hacienda Restaurant in Old Town Albuquerque, where my father's lifelong friend, Mary Constance Chavez [1928-81], was the cashier

58. In this unusual and unprecedented social milieu, in which American defense officials operating in the present were sharing intelligence data from the future with both future national leaders and their colleagues on the project, I had, on occasion, the opportunity to overhear "shop talk" that was filtering through the social network of Project Pegasus, in which project principals were revealing what they knew about future events, developments, and conditions.

59. These discussions were in the nature of gossip or project scuttlebutt and presumably should not have been held because the information about the future that Project Pegasus was gathering was supposed to remain secret and only shared with those individuals with a need to know on a "need to know basis only."

60. The future events, developments, and conditions that I acquired prior knowledge of as a result of serving as a child on Project Pegasus included, but were not limited to: the resignation of President Richard M. Nixon; the advent of the Internet; the emergence of AIDS as an epidemic disease; the presidencies of US presidents Carter, Bush, Clinton, and Bush; widespread cell phone use in the future; the fact that Donald H. Rumsfeld would one day serve as Defense Secretary; and the attack on the Twin Towers of the World Trade Center on 9/11.

61. I am filing this affidavit with the tribunal because I have direct, personal knowledge of the fact that a secret time travel project for which he served as defense attaché gave Donald H. Rumsfeld, who served as US Secretary of Defense during 9/11, prior knowledge of 9/11 thirty years before 9/11.

62. While serving on Project Pegasus as a child chrononaut and he as its defense attaché – as the point man between the Department of Defense and project principals like Dr. Agnew and my father – I had extensive contact with Mr. Rumsfeld.

63. In Spring 1971, I attended a lecture that Mr. Rumsfeld gave to about 15 young people on the project. This lecture was held at an underground amphitheater at the Cerrillos Cultural Center near Cerrillos, NM. During the lecture, Mr. Rumsfeld thanked us for our participation in the program, admonished us to keep our knowledge of the time travel technologies involved absolutely secret because the national security of the United States depended on it, and informed us that the significance of teleportation was primarily military in that it would enable the US Army to put troops precisely where they were needed on the battlefield.

64. I greeted Mr. Rumsfeld as we entered the same amphitheater at the Cerrillos Cultural Center for a later briefing. Sticking out my hand and imitating adults, I introduced myself as simply "Basiago." Following suit, he shook my hand and replied: "Rumsfeld."

65. I was in the company of my father in 1971 when we entered a bar in Albuquerque and my father had a conversation with Mr. Rumsfeld in which my father and Mr. Rumsfeld commiserated about being involved in a program that was involving American school children in time travel. During this conversation, my father informed him that the program had determined that he was destined to "make" Defense Secretary. Mr. Rumsfeld replied: "I know that, Ray. I read that in one of the reports. But that's one future event that I'll only believe when it happens."

66. I was again in the company of my father in 1972 when my father was briefed by Mr. Rumsfeld about the fact that upon high school graduation the children in the program would be admitted to the US Naval Academy at Annapolis "as a pretext for involving them in future project activities," as Mr. Rumsfeld stated.

67. In addition, I was a guest at an afternoon reception at a downtown hotel in Albuquerque. After drinks were served, Mr. Rumsfeld regaled those around him with stories of his exploits as a naval aviator during the Korean War.

68. The 9/11 event was spoken of openly within Project Pegasus' inner circle. I think this was the case because it was not only a future event that had been detected by the program but because it was a dramatic future event; it was a future event that had implications for the national security of the United States; and it was a future event that involved something that had not been seen since the War of 1812, which was a coordinated attack on the US mainland.

69. One time, in a restaurant in Albuquerque, NM, one of the project engineers asked my father what he knew about this future event. "What can you tell me about the event called 9-1-1, Ray? Apparently, it's some kind of domestic Pearl Harbor?" My father responded, matter-of-factly: "Middle Eastern terrorists hijack commercial airliners and fly them into the Twin Towers in New York City." "What happens then, Ray?" the project engineer asked. My father answered: "The Twin Towers come tumbling down, like a ton of bricks." "God, how awful," the project engineer sighed. "What happens after that?" My father answered: "The country picks itself up, dusts itself off, and goes on like nothing happened."

70. While this was typical of the kind of informal discussion about 9/11 that I overheard as a boy while serving on Project Pegasus, my father and I were actually shown visual evidence of the future attacks of 9/11 that had been retrieved by the project from 30 years in the future.

71. One afternoon during our Summer 1971 stay in New Mexico, my father and I had just finished lunch at the La Hacienda Restaurant in Old Town Albuquerque, when my father was called to the telephone. He listened intently, and then said: "I'll be right down." When he hung up the phone, he said: "Let's go." When we were in the car, I asked my father where we were going. He said, tersely: "Aerojet."

72. My father sped to our destination. The Aerojet Corporation was in Socorro, NM, a distance of 90 miles south of Albuquerque. We reached Aerojet in a little over an hour, meaning that we must have been traveling at 80 to 85 mph for my father to get us there so swiftly by private automobile. At Aerojet, we were shown astonishing evidence, so, in retrospect, I can understand my father's excitement.

73. In those years, the Aerojet Corporation facility was at the corner of Bullock Avenue and Leroy Place on the New Mexico Institute of Science and Technology [NMIST] campus in Socorro, NM. It was a small, C-shaped building with a zero-scaped courtyard. A small macadam parking lot on the grass in front of the building accommodated about 10 cars. When one turned left while traveling west on Bullock Avenue to park in this small parking lot, one passed a sign consisting of a slab of black granite with silver letters that read "Aerojet Corporation." The Aerojet Corporation facility on the NMIST campus was processing images of future events retrieved by US chrononauts from the future.

74. During one of three fact-finding trips about my project experiences that I took to New Mexico in 2003, 2004, and 2008, Joy Thompson, the head research librarian for NMIST, not only confirmed that a C-shaped Aerojet Corporation building was present on the grounds of NMIST in the early 1970's, but that in those years it was misidentified on the campus maps published inside the official NMIST course catalogues as either Gibbons Hall, a student union of NMIST, or as the National Radio Astronomy Observatory, which is located a mile west of NMIST. This was contrary to what I had been told by Aerojet before making my 2004 trip to New Mexico. They told me that Aerojet did not have a corporate presence in Socorro until 1985, when, in fact, it was in 1985 that Aerojet built its present facility on the NMIST campus and tore down its earlier, C-shaped building.

75. When, in 1971, my father and I entered the Aerojet building through a door in the interior of the "C," my father showed an armed security guard his Department of Defense identification card and told the guard, "My son's with me." The guard pushed a button and prison cell-like bars blocking access to the secured left side of the building retracted into the ceiling and we walked down the hall.

76. We were greeted in the hall by one of the photo-analysts processing images from the future at the Aerojet facility. I remember that my father had explained that some of these photo-analysts were graduate physics students studying under Dr. Stirling Colgate, who was then serving as President of NMIST and chair of its physics department. The photo-analyst was very excited and as he approached my father and I excitedly informed my father: "We've got it, Ray! We've got images of the event!" My father shook his head in bemused bewilderment.

77. The photo-analyst led us into a small room that had about 20 TV monitors stacked all over the place in a hodgepodge manner. The small photo-analytical lab was dense with TV monitors and video cassette consoles. The photo-analyst then took a beta video cassette and with trembling hands popped it into a beta video console. "Watch this!" he told us. The TV screen illuminated with a set of large square pixels that formed from the jagged leading edge of what appeared to be a fragment of videotape. This might have been the remnant of one of the microfilms of the sort that we were retrieving from the year 2045.

78. When the picture pixilated into view, we saw sun glinting off of a commercial passenger airliner as it entered the scene from the left, traveled across the screen slowly, and crashed into one of the Twin Towers in Downtown Manhattan, causing a dense fireball to erupt in the area of its impact with the building. The videotape loop that we were shown was approximately 20 seconds in duration. Every time the fireball would erupt and the loop would conclude, the photo-analyst would replay the videotape loop, and every time the plane again struck one of the Twin Towers, he and another young photo-analyst watching would exclaim "Bam!" as if they were replaying a replay of a "slam dunk" in a televised basketball game. I believe that what we were shown was the second plane strike that would take place 30 years in the future on 9/11 and its resulting fireball. Certainly, the footage was indistinguishable from the video evidence of the attack on the Twin Towers that has been broadcast widely since 9/11.

79. In light of the fact that my father and I learned of 9/11 while attached to Project Pegasus, it is certain that Mr. Rumsfeld did, too. After all, he was the defense attaché for the program, which placed him at a higher level of responsibility than my father. Inevitably, Mr. Rumsfeld would have been shown the visual evidence of the future event, as my father and I were shown that day at the Aerojet Corporation in Socorro, NM in 1971; or would have read of the future terror attack associated with the term "9/11" in one of the intelligence reports about future events that were being circulated in the leadership echelon of the program and among top intelligence officials; or would have heard 9/11 discussed in casual conversations between project principals.

80. This evidence establishes that 30 years prior to 9/11, and on the day of 9/11, the US Secretary of Defense, Donald H. Rumsfeld, had prior knowledge of 9/11.

DATED this 11th day of September, 2012

ANDREW D. BASIAGO,
US Chrononaut, 1968-72, 1980-84

STATE OF WASHINGTON)
: ss.
COUNTY OF CLARK)

SUBSCRIBED AND SWORN to before me this 11th day of September 2012.

SHARI M. ROARK
10-12-2015

Shari M. Roark
NOTARY PUBLIC OF WASHINGTON
My commission expires
10/12/2015

The one and only true timetraveler Andrew D. Basiago tells the Truth:

THE EIGHT MODALITIES OF TIME TRAVEL by Andrew D. Basiago

1. Remote viewing was developed by the US military in the 1960's years before it was supposedly developed at Stanford Research Institute [SRI] in 1972. It uses the human mind to travel to locations distant in space in real time or distant in time-space.

2. Spinning to induce out-of-body experiences so as to travel on the astral plane is an ancient occult practice. It uses dissociation of the mind to access while awake the astral realm that we access while we are lucid dreaming.

3. The Montauk chair was reverse-engineered from the pilot's seat aboard a crashed ET craft, by which the ET pilot piloted the craft psychically to avoid collisions in space in light of the speed of the craft and the vast interstellar distances traversed. It uses magnetic transduction to boost human consciousness forward in time so that the individual in the chair pre-experiences a moment in his subjective future.

4. The teleporter, which opens up a vortal tunnel in the fabric of time-space through which the teleportee passes from Point A to Point B in several seconds, was invented by Nikola Tesla. It uses a field of radiant energy to open up the vortal tunnel in time-space. When the tunnel in time-space closes, the teleportee finds footfall in the location where the tunnel closes, either in real time or in the past or future.

5. The chronovisor was accidentally discovered by Vatican musicologists Father Pellegrino Ernetti and Father Augustino Gemelli when they were studying the harmonic patterns in Gregorian chants at the Catholic University of Milan in the 1940's. Ernetti and Gemelli found that the microphone they were developing could pick up the sounds of past events. It uses holograms that are so dense that they have the effect of lensing non-

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

OK NO READ MORE

multi-colored light by US defense contractors under DARPA after the Vatican gave the chronovisor technology to the US government for further development.

6. The stargate is a Tesla teleporter that so concentrates the radiant energy that it derives from the quantum hologram that it can send the teleportee vast distances in time-space. Once inside the vortal tunnnel, the teleportee can see that the tunnel is identical to the tunnel walls of bluish-white holographic light produced by the Tesla teleporter, but since greater distances in time-space can be reached by the stargate, the teleportee spends more time in the tunnel and so has a greater risk of asphyxiation. It was via stargate that Project Pegasus was accessing 2045 in 1972.

7. The plasma confinement chamber was invented by Dr. Stirling Colgate, president of the New Mexico Institute of Science and Technology [NMIST] and a physicist at the Los Alamos National Laboratories [LANL]. It uses radioactive plasma to propagate a wormhole which the time traveler travels in order to be embedded temporarily in a past event. When this local quantum field effect created around the chrononaut wears off, he experiences transitioning spontaneously back to his point of embarkation in the chamber without traveling via wormhole. It was via this device that Andrew D. Basiago was sent to Gettysburg, PA in 1863 in Spring 1972 as seen in the Josephine Cobb image of Lincoln at Gettysburg.

8. The jump room or “aeronautical repositioning chamber” [ARC] was developed in a joint venture between Parsons and Lockheed, possibly after being reverse-engineered from an extraterrestrial device or as a result of ET-human liaison in which the device was given to the US government by one of the Grey ET species. It uses an unknown process to relocate the teleportee on an interplanetary basis. In the CIA’s Mars jump room program, jumps between Earth and Mars tended to take about 20 minutes and involved the jump room “morphing” from a box to a cylinder and back again during the jump.

Copyright 2016 by Andrew D. Basiago. All rights reserved.

Source

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

OK NO READ MORE

Humans Are Free

Alien Base for Giants Discovered in the Bucegi Mountains

Share 19.3k Tweet G+ Share Share Share

http://humansarefree.com/2010/11/another-true-history-chapter.html

The article was originally published on November 30th, 2010, at 11:11 AM. It was improved, corrected and republished in May 2013.

I. Short Introduction

In the summer of 2003, in an unexplored area of the Bucegi Mountains, a team from Zero Department (a top secret section of the Romanian Intelligence Service -- SRI), had made an epochal discovery, which could had completely change mankind's destiny.

The United States of America exercised colossal diplomatic pressure on the Romanian Government, which intended to disclose the findings to the entire world.

The implications became far more complex due to the brutal interference of the Order of the Illuminati, which sought to take control of both the location of the discovery, and the joint Romanian - American expedition.

The expert on strange phenomena, and also the leader of operations for the Romanian State, Cezar Brad, is the “hero” of the shocking events that occurred in the Bucegi Mountains. He also had two memorable meetings with a leading representative of the Illuminati, and top Bilderberg member.

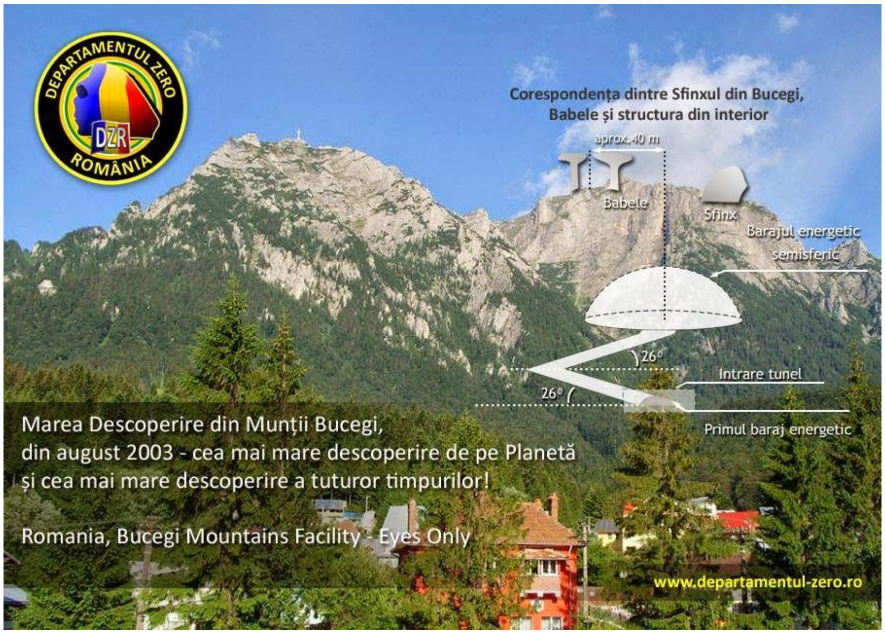
DEPARTAMENTUL ZERO
DZR
ROMÂNIA
Corespondența dintre Sfinxul din Bucegi, Babele și structura din interior
aprox. 40 m
Babele
Sfinx
Barajul energetic semisferic
26°
Intrare tunel
26°
Primul baraj energetic
Marea Descoperire din Munții Bucegi, din august 2003 - cea mai mare descoperire de pe Planetă și cea mai mare descoperire a tuturor timpurilor!
Romania, Bucegi Mountains Facility - Eyes Only
www.departamentul-zero.ro

The location of the alien base inside the 'Masivul Bucegi' mountain, Romania. Also, notice the locations of the Sphinx of Bucegi and the Babele monoliths above the alien base.

II. A Strange Visit From the Most Powerful Bilderberg Member

In May 2003, Cezar was visited by a highly important character. The meeting request came through SRI, as a result of the government's intervention. The person was a foreigner, but spoke the Romanian language very well. He was also familiar with the country. The SRI informed Cezar that the visitor was a high ranking member of a very important Masonic lodge from Italy, he was a noble, and had very strong financial influence in Romania.

His political influence was also very high, since he was able to penetrate the wall of SRI agents and reach the DZ.

Cezar felt a high pressure and a heavy weight around this person. He was surrounded by a cloud of heavy, unpleasant radiation which concealed his true intentions. For this meeting, Cezar prepared thoroughly, isolating himself in a room and falling into a state of deep meditation, to learn more about the person.

A SRI helicopter brought this arrogant, tall gentleman, wearing a black suit. He had a cane with ivory handle and gold inlay. His face expressed harshness and his green eyes had a strange effect, radiating unusual coldness. He presented himself as Senior Massini (probably not the real name). He was very confident of his power and created the impression of a person who was used to give orders.

He was the leader of one of the most important Masonic Lodges of Europe and also to one of the most influential Masonic organizations in the world: the Bilderberg Group.

Massini was very direct and stated that humans are of two kinds: those who can be manipulated and ruled (these form the majority), and those who have certain virtues and strong personalities. He said that his group is part of the highest Masonic order and is very interested in the outcome of the discussion.

Mister Massini explained that the Bilderberg Group is not a Masonic lodge, and it means much more than that. He said that the lodges are just facades and the real power is much higher than the thirty-third hierarchical level. Massini invited Cezar to join the group, letting him understand that he will have many advantages.

According to Cezar, Massini was both physically and psychical strong, which contrasted with his age. Unfortunately for him, his power was centered on a huge ego, arrogance and sense of superiority over the rest.

Massini informed Cezar that he requested him personally, using his political influence, because he was impressed by Cezar's psychic power.

III. The Pentagon Spies With Satellites

A Pentagon satellite used for geodetic espionage, based on bionic technology and shape waves, discovered in 2002 a separate unit in a specific area of the Bucegi Mountains. The empty space inside the mountain had no correspondence with the outside, and looked like it was carved from the inside, by intelligent beings. It was definitely not a cave.

The satellite scan of the mountain revealed two major energetic blocks. These barriers were made of artificial energy. The first one was like an energetic wall, blocking the access to the tunnel, while the second one was shaped like a dome (or hemisphere), and was located at the opposite end of the tunnel, near the center of the mountain.

Massini was sure that inside the dome they will discover something extremely important.

He had deep knowledge about the origin of this discovery and had knew of the existence of at least one item located inside the great hemispherical hall. The tunnel and dome were strangely aligned with the rock formations from the top of the mountain, known as "Babele" and the "Sphinx of Bucegi".

The Romanian Sphinx of Bucegi

IV. A Similar Structure In Iraq

The Pentagon team noted that the hemispherical energy barrier had the same vibrational frequency and the same shape as one other top secret underground structure that they had discovered prior, near Baghdad, Iraq. Shortly after this discovery, the Iraqi war broke out and after a few months, the Americans had access to the biggest secret in the area -- which the Iraqis knew nothing of.

Massini explained Cezar that the content of that discovery had to do with Earth's mysterious past and the history of their secret organizations. When the Pentagon investigation noted the similarities between the underground structure of Baghdad and the one from the Bucegi Mountains, Massini and his masonic lodge became extremely agitated. Initially, they've almost panicked. The panic was due to the fact that this structure - much larger and more complex than the one from Iraq - is on Romania’s territory.

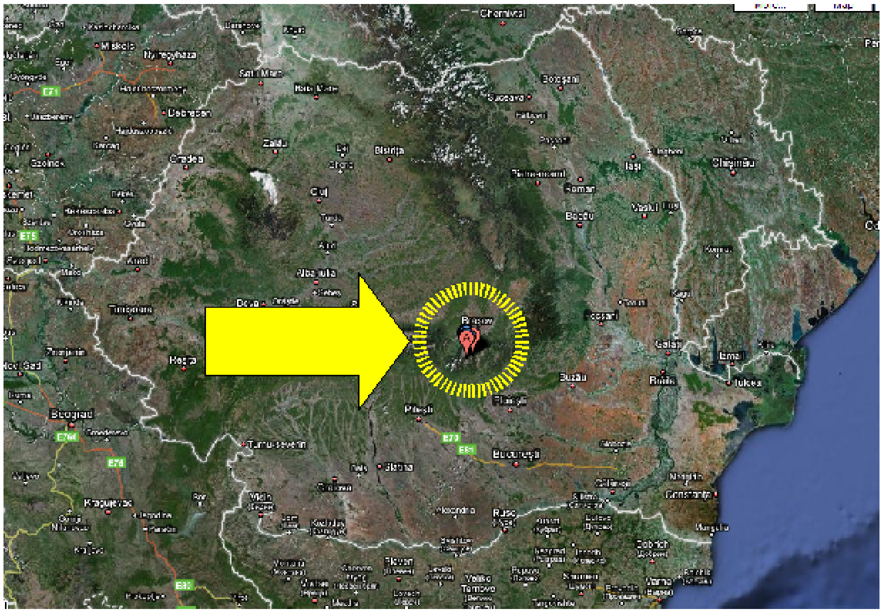

The location of the alien base, on Romania's map

[Apparently, Romania is probable to play an important role in the downfall of the secret societies that control the world and enslave mankind. It was also suggested that somewhere above the Bucegi Mountains, an energetic pyramid is located, which is invisible to the naked eye, and contains the real history of our planet. One can imagine their panic].

Massini brought to the drilling site an ultra sophisticated, hard rock drilling machine, used by the US military. The device used a strong plasma jet and a sort of rotating magnetic field, literally melting the rock with no visible effort.

V. The Secrets of the Bucegi Mountains -- Year 2003

They were able to achieve penetration about 60-70 meters away from the first energetic barrier, and reached the first gallery, which looked like a subway tunnel. Its walls were perfectly polished. At the end of the tunnel there was a massive stone gate, which was protected by an invisible energetic barrier. Three members of the first special intervention team tried to touch the door, and immediately died of cardiac arrest.

Any object (rock, plastic, metal or wood) threw at the barrier, immediately turned into fine dust. Two generals from the Pentagon and the U.S. presidential adviser arrived on the spot.

VI. The Grand Gallery

Beyond the formidable energy barrier, which caused the deaths of three people, there was also the solid rock gate. On the tunnel’s wall, just in front of the gate, there was an area of 20 square cm, on which there was precisely drawn an equilateral triangle pointing up. The square was located between the huge grinding stone gate and the invisible energy barrier.

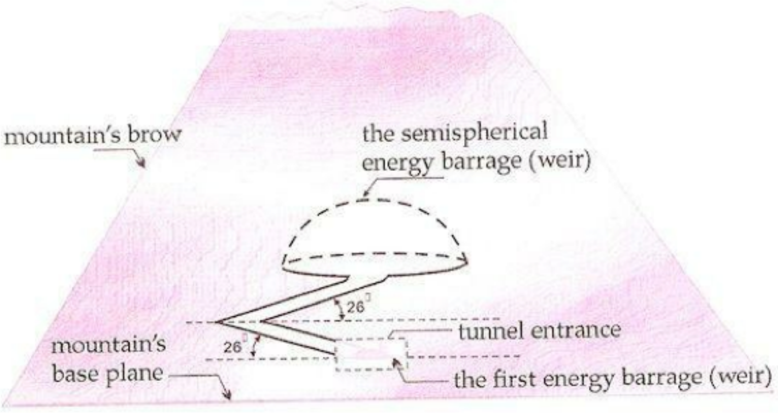
mountain's brow
the semispherical energy barrage (weir)
26°
26°
tunnel entrance
mountain's base plane
the first energy barrage (weir)

Tunnel and dome inside the Bucegi Mountains

Cezar felt that there is some kind of compatibility between the energetic barrier and himself, something like a mutual “sympathy”. His hand lightly touched the surface of the energetic barrier and he felt a tingling on the skin. The shield was completely harmless to him, so he stepped forward, passing right through it. The U.S. officials were absolutely stunned.

Cezar estimated the barrier to be no more than an inch thick.

He touched the triangle drawing located in the center of the square, and the gigantic stone door silently slid to the left, into the wall. That command also cancelled the energetic barrier, giving them access to a huge room, which was later named "The Grand Gallery".

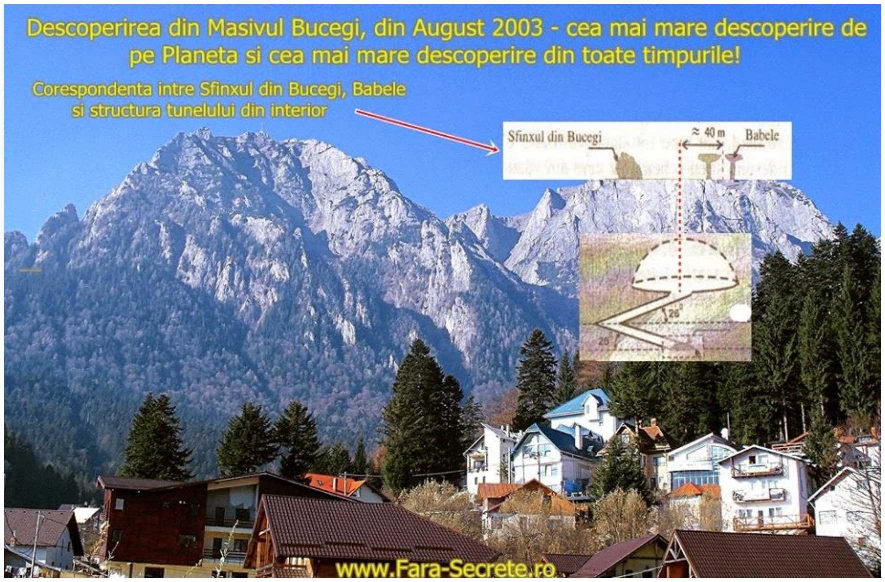
Descoperirea din Masivul Bucegi, din August 2003 - cea mai mare descoperire de pe Planeta si cea mai mare descoperire din toate timpurile!
Corespondenta intre Sfinxul din Bucegi, Babele si structura tunelului din interior
Sfinxul din Bucegi
≈ 40 m
Babele
26°
26°
www.Fara-Secrete.ro

Even though there was no visible light source, the Grand Gallery was perfectly lit.

After turning off the first energetic barrier, the huge hemispherical shield at the other end of the room, suddenly tuned to a higher vibration and started emitting higher radiations.

On a closer analysis, the wall inside the Grand Gallery seemed synthetic but also felt like something organic was part of it. It had the color of oil, but the reflections were green and blue.

Later tests revealed that the wall's material was somewhat rough to the touch, but it could not be scratched or bent. It withstood any attempt of breaking, piercing or cutting it. Later, the scientists tried to burn the material, but in a mysterious way, the flames were somehow absorbed within it, leaving it untouched.

The American scientists agreed that the material is a mysterious combination of organic and inorganic matter. After 280 feet, the gallery suddenly turned to the right in a sharp angle. Further ahead a blue, sparkly, light could be seen. The blue light at the end of the gallery was the reflection of the protective shield of energy.

VII. A Similar Base in Iraq

The U.S. adviser on national security issues received a call and he was notified that the energetic shield from Baghdad, Iraq, had been suddenly activated, and it was also pulsating at a higher frequency.

In front of the shield from Baghdad, a hologram of our planet appeared, which sequentially and progressively depicted the European Continent, then moved to the south-east, then it showed the Bucegi Mountains from Romania and, finally, it showed their own location within the structure's corridor. It was obvious that the two hemispherical energy shields were in a direct connection.

Basically, the Iraqi base had been notified about the presence of people inside the Romanian base.

The bad news was that the U.S. presidency was notified and it contacted the Romanian diplomacy. In just a few minutes, the whole operation had been disclosed, and Lord Massini's plan went down the drain.

The U.S. presidency demanded to take control over the secret base and the entire operation. The Romanian politicians, who knew nothing of the undergoing secret operation, panicked.

The Pentagon's generals present at the scene, had been informed that Washington demanded an urgent meeting.

VIII. CSAT (Supreme Defense Council) Emergency Meeting

An emergency session of the Supreme Defense Council (CSAT) had created a huge wave of sympathy for the Zero Department. Most were shaken by the news they've received. The CSAT's decision was to continue the research, but under Zero Department's complete control. They've had also demanded an inventory of everything found in the Great Room.

From Bucharest (Romania's capital), the orders came in waves, canceling each other. Some were very vehement and strict, while others were elusive -- denoting the huge tension.

CSAT's members were in continuous session, keeping in touch with the Bucegi Mountains team. After discussing everything for hours, they had decided to make the discovery public. The Romanian government was to make a formal statement to the entire world. Some CSAT members vehemently opposed the decision.

IX. Romania's Official Statement

When the U.S. diplomacy had been informed that Romania will disclose the discoveries, everything tuned into chaos. The President was called for a direct phone conversation with the White House.

Within hours, the U.S. blocked all financial transactions with Romania and its access to all other financial institutions. Romania was about to declare “State of emergency” in the Bucegi Mountains and the capital.

The talks between the U.S. officials arrived in Bucharest and the Romanian Emergency Department took place without a translator. The U.S. officials were verbally violent, constantly shouting and uttering threats to Romania and its officials.

Romania's official disclosure to the world would have provided photographic evidence and anything else essential for a complete clarification. Leading scientists and researchers from all over the planet were to be invited for in depth studies.

But most importantly, it would have revealed the truth about mankind's distant past and the real history -- which according to what Cezar witnessed inside the Grand Gallery, is almost entirely counterfeit.

The reaction of the U.S. government was so brutal because that disclosure would have shattered their global influence and power, in a instant.

The official reason cited by the U.S., was not to create panic in the world, but they had omitted to acknowledge that the current state of global anguish is a direct result of deliberate deceit and manipulation conducted by the Freemasonry and other secret societies, for millenia.

There was also an intervention from the Vatican (if you wonder how did the Vatican found out about this discovery, then you should know that they are at the very top of the world's control pyramid, and their religious piety is merely a facade -- read: The Darkest Secrets of the Vatican), in which the Pope (!) called for moderation before this great, fundamental, step for mankind.

The Pope promised to make certain documents available to the Romanian state from the ancient papal secret archives, which were of great importance for Romania. The documents were backing up the discoveries (yes, that's how powerful a Pope is).

After 24 hours of negotiations, a final agreement occurred between Romania and the USA. The Romanian state was to postpone the disclosure, and gradually present everything to the people.

X. The Projection Room

The Grand Gallery ended abruptly with a giant auditorium, 30 meters in height (98,5 feet) and a length of 100 meters (328 feet). The Projection Room was smaller in size and it was protected by an energetic shield.

Advancing towards the shield, a portion of it disappeared, in the shape of a door, allowing access inside it. The shield protected the room by any outside influences. Once inside the room, the shield became compact and looked like a white-golden wall.

Basically, the shield formed this dome-shaped room, with curved walls and ceiling. At the back of the room, at a height of a about 10-12 meters (33-39,5 feet), the shield ended where it touched the room's stone wall.

In this wall there were three enormous tunnel holes: one located straight ahead, and the other two were symmetrical on both sides. They were lit by a diffuse light in a greenish tint. Both sides decided to prohibit the access to these tunnels and a protocol was signed between them.

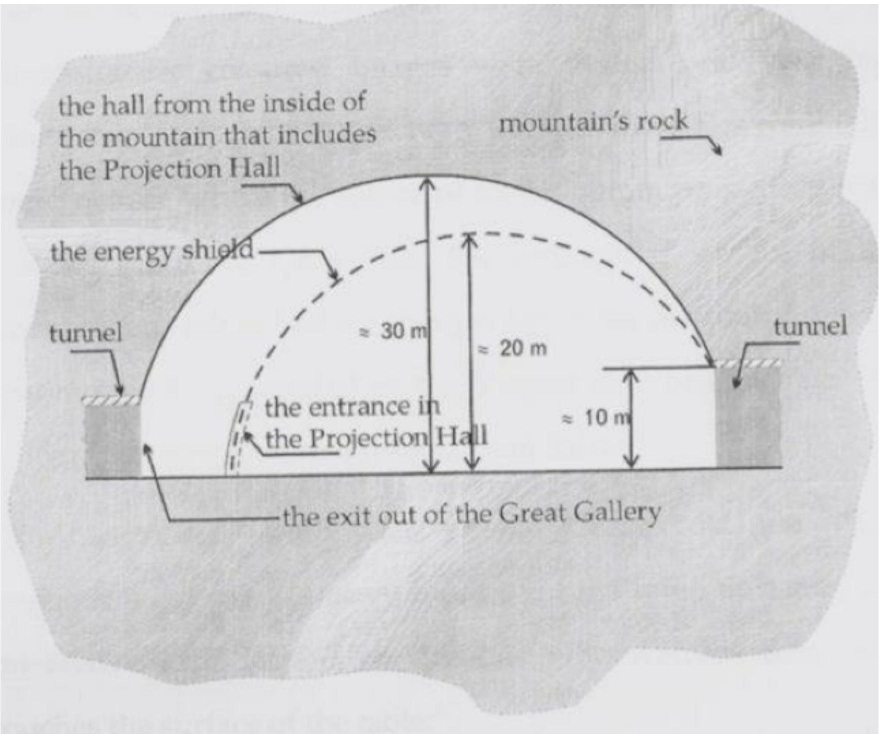
the hall from the inside of the mountain that includes the Projection Hall
mountain's rock
the energy shield
tunnel
tunnel
≈ 30 m
≈ 20 m
≈ 10 m
the entrance in the Projection Hall
the exit out of the Great Gallery

XI. Tables For Giants

A series of huge stone tables were arranged along the right and left walls (five on each side), following their curvature. The tables were about 2 meters in height (6,6 feet).

The table tops had precision cut-in reliefs, different signs of an unknown writing and characters that resembled ancient cuneiform. The writing also contained more general symbols, such as triangles and circles. Although the signs were not painted, they've had a fluorescent light radiating in various colors, different for each table.

On some of the tables there were different objects, which appeared to be technical tools. From many of these tools, translucent-white wires descended to the floor, and entered inside shiny, silver-like, rectangular boxes. These boxes were placed directly on the floor.

At a closer inspection, the cables were extremely flexible and lightweight, and light pulses could be seen circulating along their length. Each time someone approached the tables, a holographic projection was automatically activated, presenting aspects of a particular scientific field. The three dimensional images were perfect and had height of almost two and a half meters (8,2 ft).

The projections were automated and ran by themselves, but at the same time they were interactive, and changed according to the signs touched on the tables' surface.

[Picture of an actual giant skeleton unearthed in the 1970s inside a Romanian gold mine, at Rosia Montana -- the full story in a future article]:

Sabol Corna -Rosia Montana
13 decembrie 1976
Uluitor! Schelet urias de dac hiperborean fosilizat descoperit recent la Rosia
Este Rosia Montana unicul portal către Shamballa?

XII. A Great Discovery: Holographic DNA Combinations Between Extraterrestrial Species

Because of the great height of the tables, the scientists used special tripods to climb at a comfortable height. At a closer investigation they've observed a dark-glassy material covering the table tops. The material was divided into several large squares, bounded by straight lines, which formed some kind of a grid.

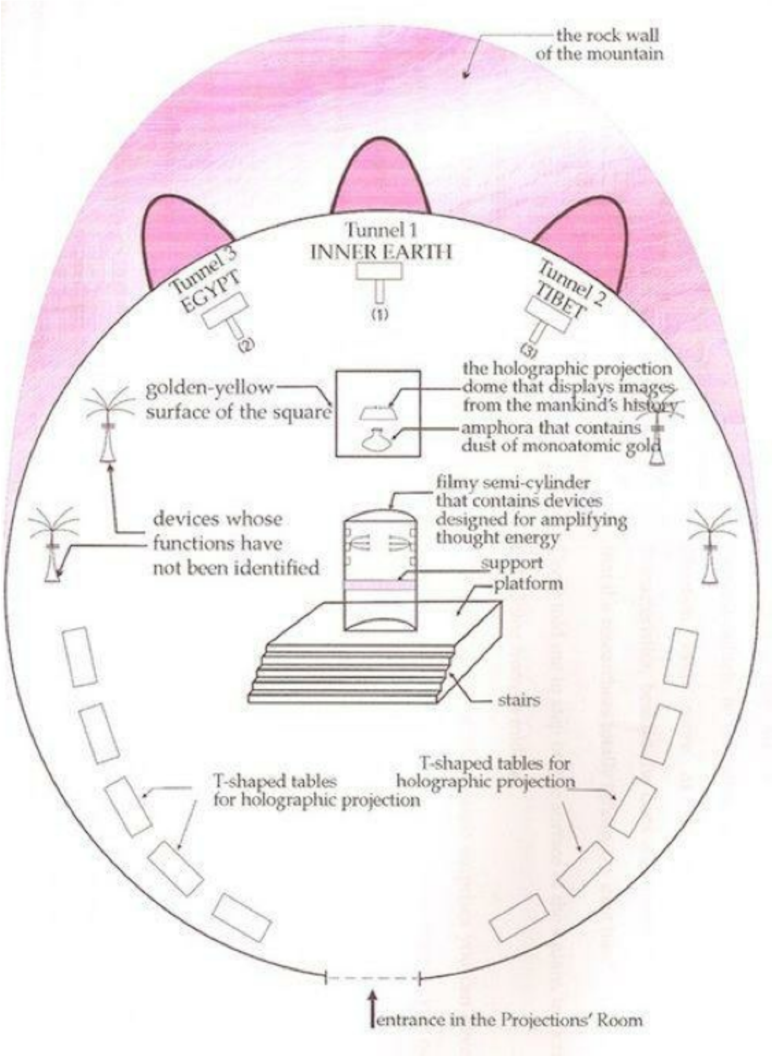
the rock wall of the mountain
Tunnel 1 INNER EARTH
(1)
Tunnel 3 EGYPT
(2)
Tunnel 2 TIBET
(3)
golden-yellow surface of the square
the holographic projection dome that displays images from the mankind's history
amphora that contains dust of monoatomic gold
filmy semi-cylinder that contains devices designed for amplifying thought energy
devices whose functions have not been identified
support
platform
stairs
T-shaped tables for holographic projection
T-shaped tables for holographic projection
entrance in the Projections' Room

One of the tables contained information from the field of biology, and it projected images of plants and animals, some of which were completely unknown to the scientists. After touching one of the squares, a hologram presenting the structure of the human body was activated. The hologram was constantly rotating, highlighting various areas of the human body.

Touching other squares activated holographic projections of extraterrestrial beings, from other planetary systems. Simultaneously touching two different squares, a complex scientific analysis was projected, showing the DNA of both species, and compatibility possibilities between the two.

On the sides, in vertical lines, there were written explanations in the same, alien alphabet. At the end of the simulation, a crossbreed between the two species was displayed.

XIII. Real Giants

Judging by the size of all the objects within the Projection Room, the creators of the edifice were probably very tall beings. A confirmation of giant humanoid skeletons found in Romania could be found in a newspaper called “Ziarul”:

> "The team from the newspaper is accompanied by researcher Vasile Rudan, who noted that the people of the Bozioru village have concrete evidence: a cemetery with skeletons of giants. It was discovered by chance over 20 years ago, when the authorities decided to plant apple trees in the village of Scaieni. Digging on a hill, the villagers discovered huge skeletons, measuring more than 2,40 m (7 ft 10 in).
>
> "Dragoi Ilie, one of those who worked in the apple orchard, takes us to the spot. Mister Ilie shows us around the orchard: 'Everywhere are tombs of giants. We were making holes, to plant saplings, when we found a human head as big as a pumpkin. Neither one of us had ever seen anything like that. We were all amazed. Digging further, we've found some bones of the feet, as big as the vineyard poles. The dead one must have been very big.'"

XIV. The Library of the Universe

The Projection Room contained detailed information from fields like: physics, cosmology, astronomy, architecture, technology, biology, genetics, and even religion. Because it contained so much information, the room was compared to a library.

In the middle of the room, there was a podium-like area which contained an interesting device. The scientists speculated that it might have been a device to enhance various cerebral functions.

XV. Romania's Control Panel

Next to it there was a command panel, covered with various geometric symbols, in different colors. There were also two sliding levers and a red button, in the center of the panel, above all other commands. A holographic simulation explained the purpose of the button. It showed an image of the Carpathian Mountains as seen from 25 km (15,5 miles) above them and next, an enormous quantity of water flooded the lowlands and plains.

Next, from the territories of Romania, Hungary and Ukraine, more streams of water appear, heading towards the Transylvanian plateau, in the form of huge rivers. The image focused on Romania's territory and showed most of its surface covered by the waters of a great sea. Some of the tallest mountain peaks could now be seen as small, floating islands.

Next, the waters begun to withdraw, as soon as the leavers found on the panel were activated, leaving sunk only a small area near the Retezat Godeanu mountain. A true manual.

XVI. A Mysterious Amphora

Behind the control panel there was a square pedestal, of about three meters (9 ft 10 in) per side, on which a mysterious amphora was resting. Lord Massini had knowledge of its existence and, for him and his Masonic elite, this was the most important discovery.

The amphora contained a very fine white powder. The researchers were dismayed to find that the powder was a high purity formula of monatomic gold, with an unknown crystalline structure.

The gold powder in its pure form greatly stimulates certain waves and energy exchanges at cellular and neural level. This causes an accelerated process of rejuvenation.

Theoretically, a man can live in the same physical body for several thousand years, provided they consume from time to time, a well-defined quantity of monatomic gold powder. This explains many puzzling aspects about the incredible longevity of some important figures and uncovers part of the hidden intentions of the global elite.

XVII. The Real History of Our Planet

Behind the amphora, in the middle of the square, there was a huge dome that projected holograms. It displayed the main aspects of humanity's most distant past, from its very beginning. Cezar explained that according to the images projected by the hologram, about 90% of mankind's official history is false and counterfeit. Darwin's evolutionary theory is also incorrect.

In the middle of the square is a huge dome which projects a hologram of moving parts. It contains the main aspects of the very distant past of humanity, from its very beginning. Cezar explained that according to the images projected by the hologram, Darwin's evolutionary theory is completely false.

The true origin of man is depicted in a condensed holographic form. After these lessons of truth, we can say that 90% of the official history of mankind is false and counterfeit. Unbelievable, but what is now considered to have really happened in our history, never did, while the so called "myths" and "legends" are almost entirely true.

Also, most archaeological theories are false. For example, the dinosaurs did not went extinct 65 million years ago, and the old continents of Lemuria and Atlantis existed. This strange inversion caused many problems and conflicts between people in our past. For a correct dating of the presented events, a star map corresponding to that period was projected on the background.

Although the time period covered by the projections was very large (several hundred thousand years), and the procession cycle of the Earth is 25,920 years, by observing the number of “Platonic years” (the 25,920-year cycle) the exact dating of the events was possible. The Bucegi Mountains are 50-55 thousand years old.

XVIII. Shocking Revelations

Cezar saw what happened during the so called Great Flood and where the human civilization originated, but he decided not to reveal this, as the revelations are too shocking for the mentality, ideas and knowledge of the contemporary man.

Personally, I think people like him are also responsible for the current state of unawareness and latency of the human species. We have been lied and manipulated for millenia, and as soon as we are close to the truth, someone decides that we are not yet ready to hear it.

This is an arrogant and selfish approach! The people are as ready today as they will ever be, and the sooner the truth will be released, the better. Having more pieces of the puzzle will make them accept the truth easier and, more importantly, will make convince them to take action against the manipulators and enslavers.

In a disturbing display of images, the existence of Jesus and his crucifixion was also presented. The projections revealed that many of those who witnessed the crucifixion came there from other historical periods.

They were wearing the same kind of clothing as the rest of the people, but their facial features were different, hence they were constantly covering their faces. The hologram also presented fragments from the spiritual lives of other exceptional characters of mankind's distant past, including people that we know nothing of.

In those times, the social and population distribution were completely different from what is known today, so the archaeologists and anthropologists should review their theories from scratch.

XIX. The Three Mysterious Tunnels

The discovered tunnels are thousands of miles in length and lead to three different areas of the planet. The one in the left connected the Romanian base to a yet to be discovered underground base somewhere in Egypt (which was definitely discovered and explored by now).

The right tunnel led to a similar but smaller structure, in the "Tibetan Plateau". This tunnel had three secondary ramifications. One led to an underground area near the Romanian city of Buzaru (close to the "Carpathian Bend"). Another one led to the already mentioned Iraqi base, while the final ramification led to an underground base in the "Gobi Desert", in Mongolia.

XX. The Third Tunnel -- A Secret World

The middle tunnel was the most important to both Lord Massini and the U.S. Government, which strongly requested to be kept secret to the public. This tunnel descended to extreme depths towards a veritable underground world, located near the center of our planet.

Those of you who are not familiar to the hollow earth theory, please note that there are important figures who support the theory that all planets are probably hollow on the inside, and bring interesting scientific arguments to support their claims.

Also, there are some very interesting and beautiful stories about people who allegedly traveled to this world. The realm is commonly known as Agartha, and the third tunnel was probably leading straight to it. It only makes sense that the shadow elite were not interested to make the discovery known to the people.

Related: The complete collection of stories about Agartha;

Intense preparations begun for expeditions inside all 3 tunnels, as following: the first one to Egypt, the second to Tibet, and the third one to Earth's core.

(Summary of events from the book: "Viitor cu cap de mort", by "Radu Cinamar" -- the pen name for an undisclosed writer, which in my opinion, is none other than ex-General Emil Strainu. I base my statement on the extremely similar writing style and the fact that the author had direct access to this above-Top Secret alien base -- something that couldn't have happened unless the person was an insider.)

In 2009, a local television "Antena 1" briefly presented these events:

After the program, they've received an anonymous phone threat:

The following is a transcript of the phone call. The reporter from "Antena 1" answered the phone:

Reporter: Hello, good afternoon!
Anonymous caller: We're sending you a warning: be very careful! Stop talking about the Bucegi [Mountains]!
R: Who are you?

A.c.: This information must remain at the level of some structures and must not be made public! You've entered a dangerous game. You are young, you have families... there are enough subjects in this country to speak about.
R: Mister, who are you?
A.c.: Don't wish to know us and to be yourselves interviewed by us! That's all I've had to say.

Thank you for reading and always remember: "the truth is out there". Never stop searching...

By Alexander Light, HumansAreFree.com, based on Radu Cinamar's book Transylvanian Sunrise; | You can find further details and the discoveries that followed in Radu Cinamar's books; | Cover image: Fara Secrete; | Testimonies about giant skeletons discovered all over Romania (not available in English): The Secret of Rosia Montana; Argedava's giant skeletons (more here); Romania's giants; The Giants of Scaieni;

Donate

Dear Friends,
HumansAreFree is and will always be free to access and use. If you appreciate my work, please help me continue.

Stay updated via Email Newsletter:

Subscribe

ExelateData

lightwater

Alexander Putney

"Soma is food for the gods." ~ Chàndogya Upanishad

"We have drunk the Soma and become immortal. We have attained the light the gods discovered… One thinks to have drunk Soma… Of him… no one ever tastes." ~ Rig Veda

"The man who supplies food hath always [Soma] pressing stones adjusted, a wet Soma filter… he who hath this knowledge wins the luminous spheres." ~ Atharva Veda

What is this ethereal *Soma* that is the subject of so many ancient Vedic hymns? Soma is *electrum*, a special food and beverage additive consisting of gold and silver nanoparticles that are tasteless, invisible, and electroluminescent ~*gold nanospheres emit red light, while silver nanospheres emit blue light*. Vedic wisdom clearly identifies the intracellular luminosity of electrum-enriched Soma waters and foods as endowing eternal life, directly defining a 'god' as one who 'discovered' and 'attained the light' by consuming electroluminescent gold and silver nanoparticles!

The partaking of electrum-enriched foods and beverages formed an integral rite of sacred initiation practices at all the ancient pyramids, temples and sacred sites at resonance. Identified in the Bible as *showbread* or *manna bread*, it was not made by a common baker, but by the goldsmith, to preserve the initiate from being burned in the tabernacle.

Maya wisdom prepares initiates at their sacred sites using the golden waters of Tláloc to crystallize and reawaken human consciousness, reflecting the highest wisdom of visitors from the stars as recorded by Maya ancestors in ancient times. Many renewed contacts taken up with individuals on various continents over the last several decades attest to the transdimensional use of lightwater with silver and gold nanoparticles in the ignition of HHO plasma for enhancement of consciousness and longevity through cellular regeneration processes enabling vastly extended lifespans of hundreds of years. The secrets of transdimensional travel have been revealed by extraterrestrial teachers from various confederations of peaceful planetary systems, including advanced applications of standing wave resonance for the resonant atomic transmutation processing of metals for spaceflight vehicles.

Concise explanations were given concerning the magnetic reversal of our entire solar system on December 22, 2012, and provide the timeline of specific intersecting events involving the Galactic Plane crossing, the intense blast wave of the Betelgeuse supernova and the incandescence of a Red Dwarf star at the edge of our solar system. The messages of visitors from the depths of the cosmos resonate clearly with the wisdom offerings of the indigenous Elders of Earth, preparing humanity for ascension in the resonant pulsations of worldwide aurora known as the Red Dawn, when the bleeding sky heralds the emergence of the Fifth World.

lightwater

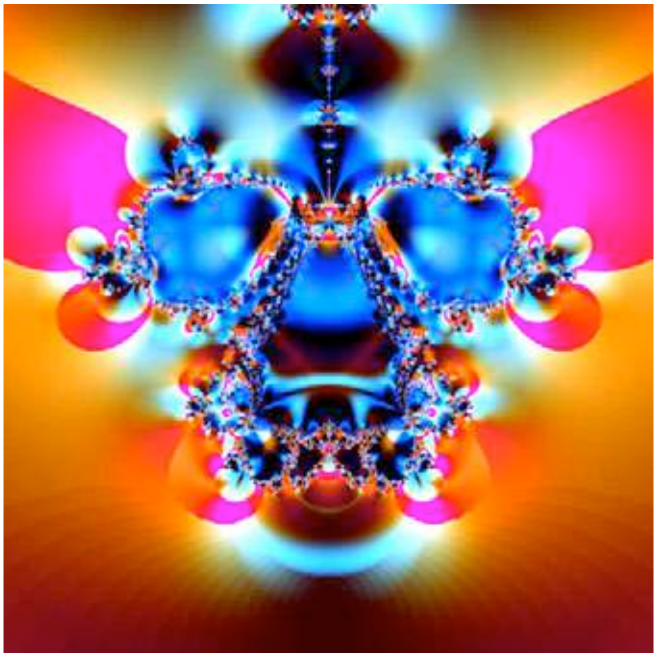

Alexander R. Putney

Human Resonance

www.humanresonance.org
www.resonánciahumana.org

© 2010 Alexander R. Putney

For all of the radiant beings who guide humanity
toward the creation of a lasting peace on Earth,
with her transformation into a heavenly planet
under the purifying red light of the solar twins.

Having no one to influence my outlook,
I display my primordial nature and adopt the style of a youthful prince.
This is due to the only father guru.

I am busy working for others.
Prajna, penetrating all obstacles, has made the prince old and wise, fearing no one.
Dancing in space, clad in clouds, eating the sun and holding the moon,
the stars are my retinue.

The naked child is beautiful and dignified.
The red flower blooms in the sky. It is ironic to see the formless dancer,
dancing to the trumpet without a trumpeter.

At the palace of red ruby, listening to the utterance of the seed syllable,
it is joyful to watch the dance of illusion,
the seductive maidens of phenomena.

The warrior without a sword, riding on a rainbow,
hears the limitless laughter of transcendent joy,
the poisonous snake becomes amrita.

Drinking fire, wearing water,
holding the mace of the wind, breathing earth,
I am the lord of the three worlds.

Excerpt from ‘Enthronement’ by Chögyam Trungpa, January 22, 1973
From ‘The Myth of Freedom and the Way of Meditation’ (2002)

1 Fourth-Density Realms 1 – 44

2 Contacts From Koldas 45 – 66

3 The Plejaren 67 – 124

4 Urandir's Transport 125 – 150

5 Golden Planets 151 – 192

6 Betelgeuse & the Red Dawn 193 – 222

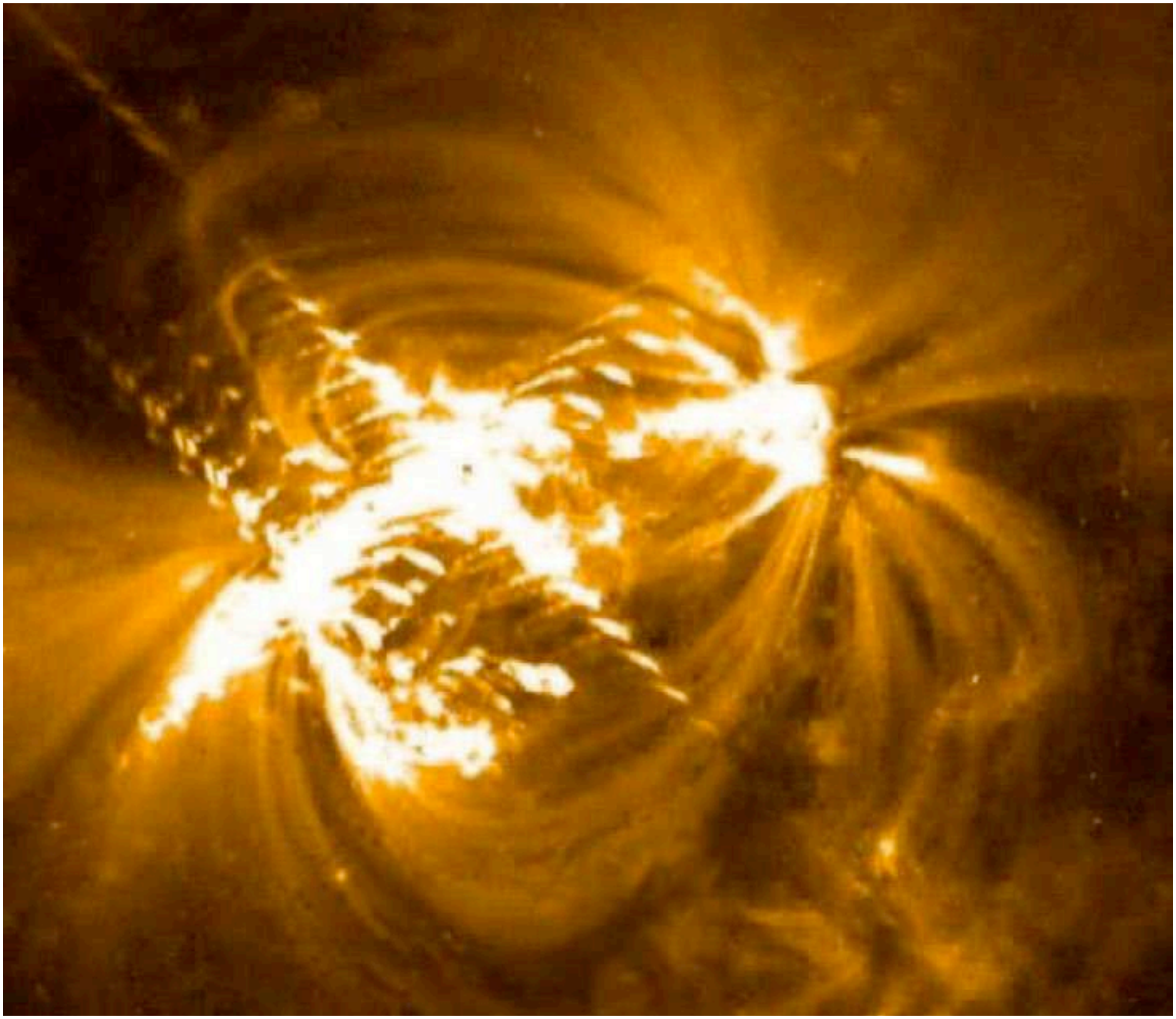

1
Fourth-Density Realms

Fourth-Density Realms

Unusual atmospheric events taking place above the surface of the Earth are unmistakable signs that our planet is rapidly ascending into a new, high-resonance quantum state in synchrony with changes witnessed on the sun. Distinct geometric patterning defining the nonlinear distribution of sunspots is retraced by atmospheric excitations of plasma into the concentric arcs of resonant standing waves within the solar corona. Solar plasma formations during coronal mass ejections display perpendicular rows of luminous plasma plumes (opposite) that are also observed of Earth's noctilucent aurora (above).

Auroral plumes were photographed and videotaped in the high atmosphere above Xiamen, China on July 12, 2010 as uneven rows of vertical plasma beams shone brightly during the night hours.[1] Perhaps the most striking examples of aerial plasma formations have been recorded near ground level in Longview, Texas by motion-activated game-trail cameras on the night of September 9-10, 2009.[2] The auspicious calendar date of this plasma flare event –9/9/9– suggests that heightened Earth resonance and solar activity may be directly linked with synchronized meditation by large masses of human beings, as once achieved by the pyramid-building cultures of the world, whose resonant temple stones still hum today.

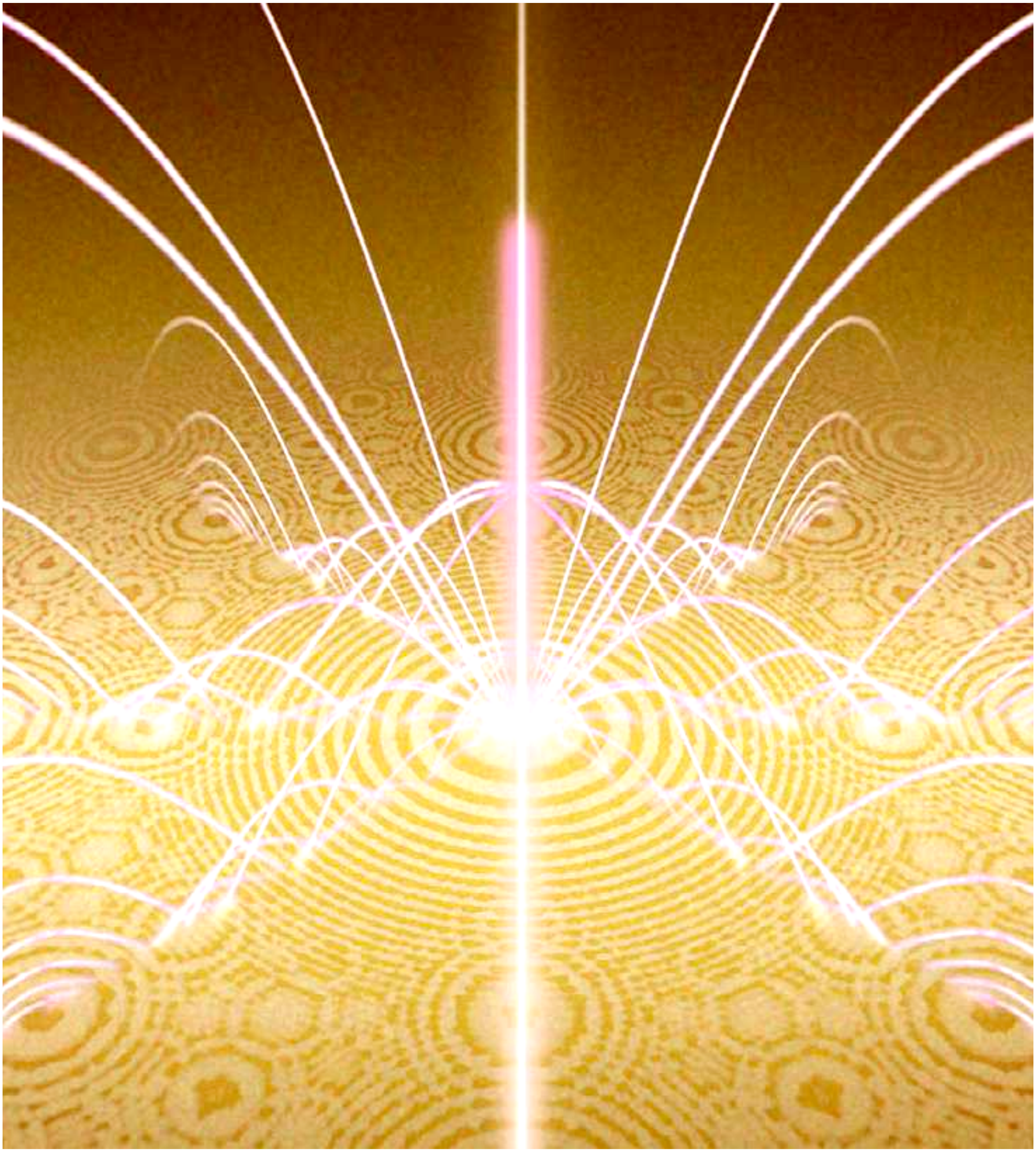

Sacred geometric knowledge of the focal points of Earth resonance remains perfectly preserved in the pyramidal monuments of the world, both in their respective structural configurations and their relative geopositions, encoding the golden proportion as a precise spherical formula [$z_{n+1} = z_n^2$]. All of the world's stone temples and sacred sites find exquisite alignment within this global psychoacoustic resonance network organized by the geometric distribution of nonlinear infrasound standing waves (above).[3] This transcendental quadratic formula has been encoded in the sacred geometric configuration of the Sanskrit mandala *–a square within a circle–* and preserved throughout ancient sacred artworks the world over.

This spherical mandala code lies carefully nested, in perfect order, in the locks of the Buddha's hair and the circular pattern of the radiant aura surrounding his golden head (opposite). The hemispheric design of the topknot bun at the crown chakra displays octopolar symmetry with sets of concentric circles arranged in radial rows of 4, 8 and 16. The extremely long earlobes often seen in representations of the Buddha reflect an odd Sanskrit cultural association with transcendent wisdom and the great attainment of yogis to become transdimensional beings, at one with Creation.

The identification of spiritual attainment with gold has been passed down since time immemorial, and the latest findings of the biophysical sciences provide access to deeper layers of meaning within these ever-present cultural concepts. Gold, like silver, is not reactive with water and so does not break down by oxidation, but maintains a pure state –just as the consciousness of a fully self-aware human being is crystallized and purified to achieve the highest state of perfection attainable.

Sacred mountains are identified by indigenous cultures for the alignment of their peaks with convergence points of infrasound resonance, while great masters of consciousness are known to reside on such sacred peaks. Among the Himalayan spiritual traditions of India and Nepal, the mystical figure Mahavatar Babaji is closely associated with a cavesite retreat within sacred Mt. Kailash. Other poignant examples include the Indian adept Sri Ramana, who practiced a barefoot yogic lifestyle atop sacred Mt. Arunachala thriving on the accumulated spiritual energy for most of his life,

This resonant bioenergetic phenomenon is well cited among ancient wisdom traditions as an attainment of spiritual and physical purity, as the water of the adept's body itself becomes crystallized by the deep delta wave beating that reigns over the enlightened consciousness. Such resonant bodies never desiccate atop sacred mountain peaks, or inside pyramid chambers, but become mummified in the preservative infrasound-enhanced electromagnetic environment as the piezoelectric crystalline water, bones and tissues are continually sterilized by mild electrical currents that neutralize all pathogens.[4] This post-mortem effect was reported of spiritual adept Simeon Toko, who died in Luanda, Angola, yet his lifeless body was observed by respectful masses for many days without any sign of putrefaction.[5]

The three main pyramids that dominate the Giza plateau were created as a symbolic representation of the three belt stars of the Orion constellation, so designed to transduce the diverse ambient energies of both solar and geomagnetic storms into a precisely ordered tri-frequency resonance. Any three acoustic frequencies in Fibonacci ratio generate the full set of nonlinear standing waves that couple energy from Earth's base Schumann Resonance frequencies (near the human heartbeat at rest) into the higher harmonics,[6] which then become audible to the human ear as a deep pulsation –the cosmic Om.

Infrasound standing waves stimulated by 1.45 Hz heartbeat booming within the Great Pyramid ignite the very water vapor surrounding the pyramids in arcs of HHO plasma that is generated as water molecules are torn apart and ionized. Gold and silver nanoparticles were a key component of the high purity water that once rose as vapor through the pyramid's ascending passages, to enhance the ignition process of HHO plasma within the stone chambers and into the surrounding atmosphere. This biocompatible form of gold (thus 'organic') was consumed by all initiates within the pyramids, and reveals the deeper significance of the star constellation name 'Orion' *–which literally means 'gold ion'! Ionized nanoparticles of gold and silver are electroluminescent, illuminating the DNA of all initiates within the pyramids:*

> Through this initiation they will be able to see the luminosity of the Great Spirit. When this spiritual initiation is completed in the house of Hunab K'u in heaven, the spirit of each initiate will be illumined. Then, it will return to Earth to awaken the body and take it to the pyramids. Chichen Itza, Uxmal, K'aba, Etznah, Palenque, all of these sacred Mayan centers have a specific function in every part of the body to awaken the 7 powers contained in it.
>
> Only through the solar initiation can the sleeping body of mankind be awakened. The reincarnated teachers of the new age of Aquarius implore for the sacred human race to awaken, so that in this way it can fulfill its sacred destiny, which is to be the true sons and daughters of the cosmic light. The time of knowledge approaches, the light in the center of the pyramidal house of Hunab K'u will flash like lighting that will pierce through the shadows that envelop the human race. Let us prepare to receive the light of knowledge that comes from Hunab K'u and transcend into the memory of the creator and become beings of eternal luminosity.[7]

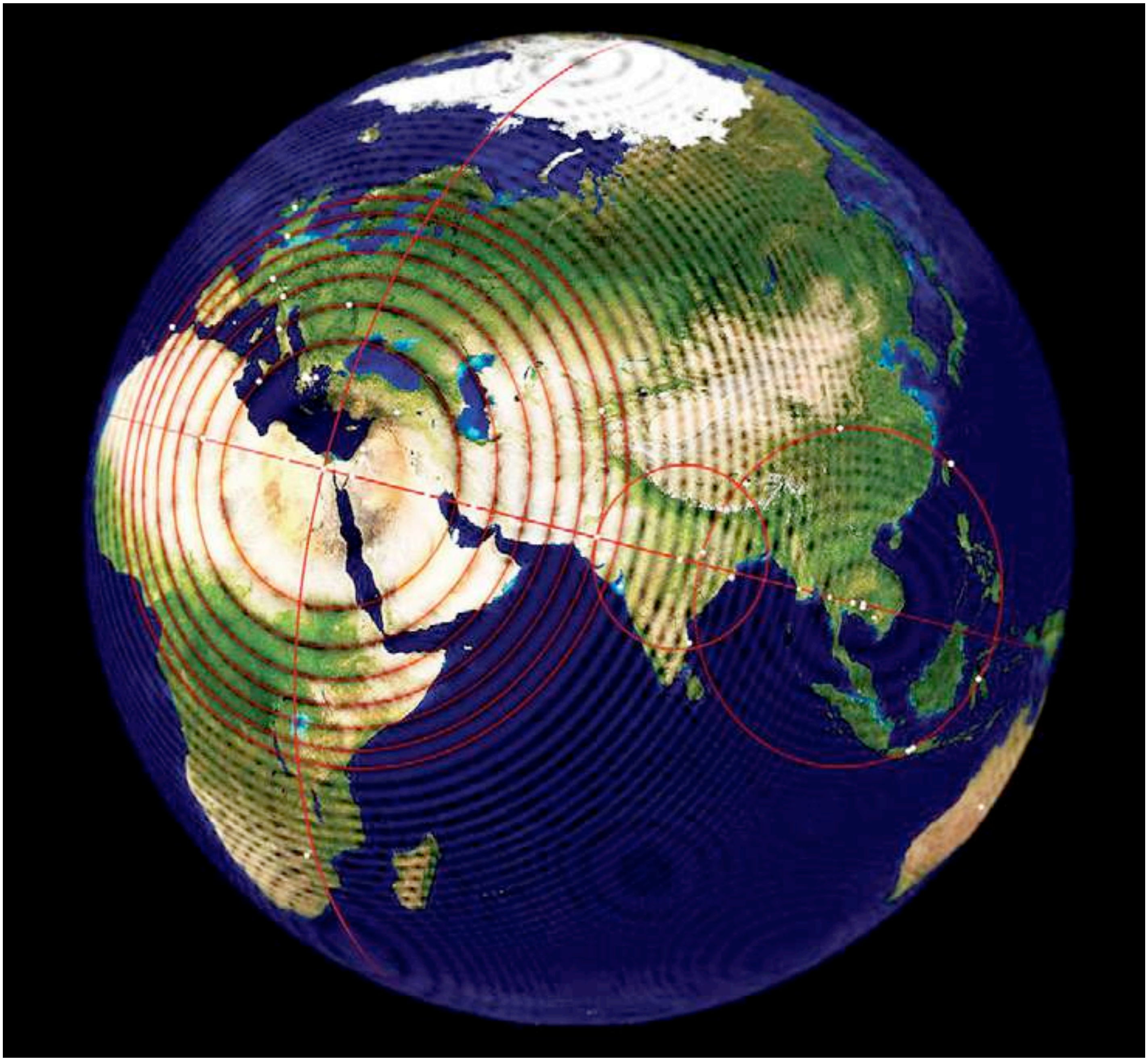

.

The electroluminosity of nano-gold and -silver within the body vastly increases cellular longevity and the lifespans of all initiates accordingly, as well as providing the quintessential biorhythmic environment for meditation, telepathic communication and Astral projection. The vast multitudes of piezoelectric stone temples found throughout the entire world find precise placement within the infrasound distribution pattern expressed by the sacred formula known to mathematicians today as four-dimensional equation [$z_{n+1} = z_n^2$], as a spherical rendering expressed from the apex of the Great Pyramid of Giza, Egypt (above, with major sacred temple sites denoted in white).

The Sanskrit origin of the world's ancient pyramid traditions has been linguistically identified by modern epigrapher Dr. Kurt Schildmann,[8] yet the living culture of the Maya people of Central America remains a most direct modern link to the Vedic knowledge of the deep past, as one of several descendant language forms. The powerful words of the ancient Sanskrit pyramid-builders still resonate within the sacred temple stones today, as a clue to their longstanding and divine purpose that is intimately linked with the conscious Spirit of all the human beings of Earth. The ancient astronomical and mathematical wisdom is preserved in Maya scriptures and oral traditions, *in conscious mastery of galactic temporal cycles:*

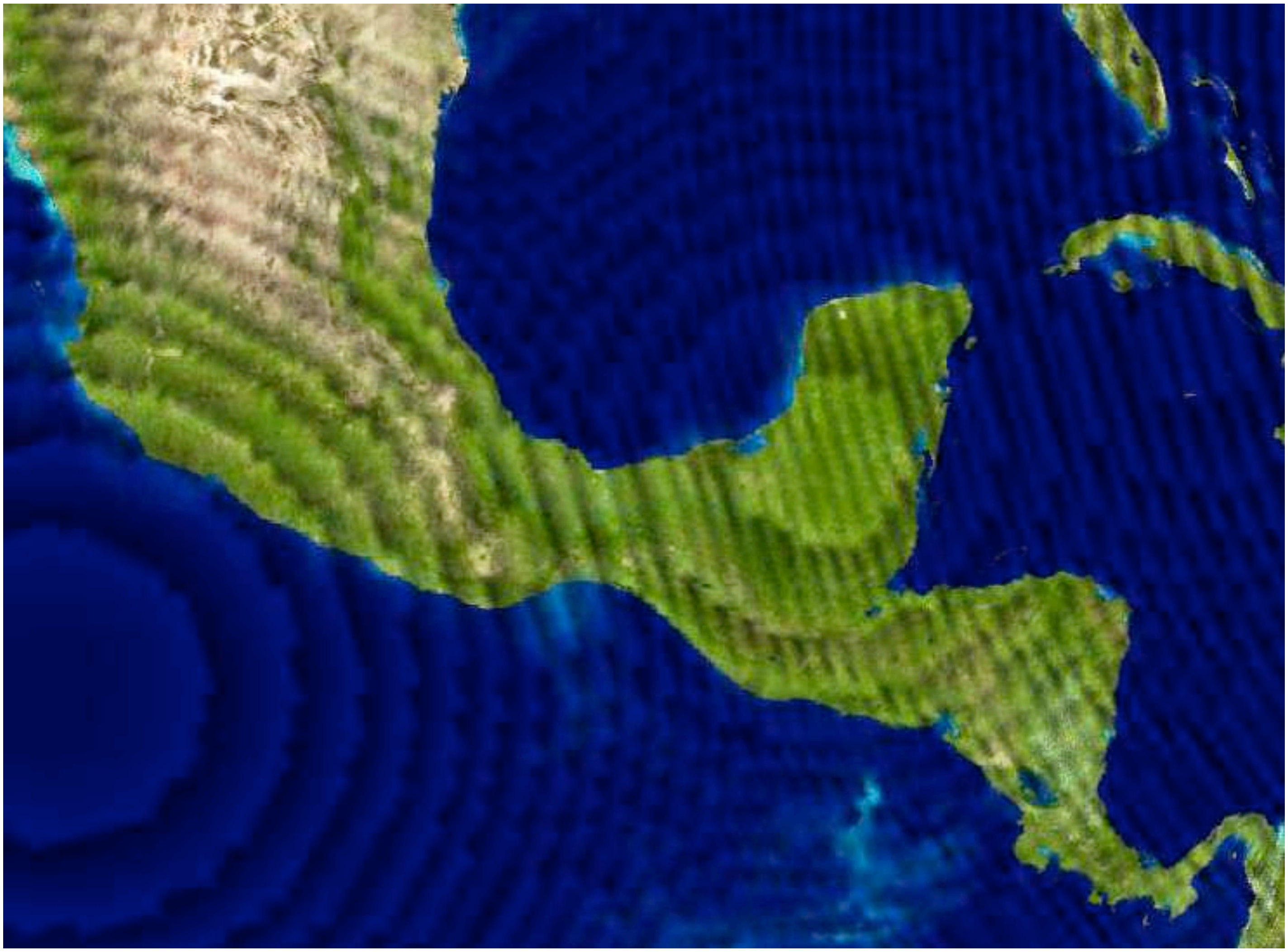

Man does not will when he wars, loves, reaps; it is the rhythms of the great gods, the planets, that act over him and make him do. When he comprehends that by himself he can do nothing, then he can learn to serve the gods; so, he must become conscious of the rhythm of the gods.

The calendar that governs the life of man, the Tonalpohualli, is based on the rites of the planets that turn closest to the Earth: Paynal, Mercury; Quetzalcóatl, Venus; and Huitzilopochtli, Mars...

After 657 Tonalpohualli (468 solar years) the cycles of the minor planets and the cycles of the Sun end together. Tonatiuh, Quetzalcóatl, Huitzilopochtli, and Paynal shine again at the same time in the same form. This is an age, the duration of the life of a culture.

Chalchiuhtlicue, Goddess of Terrestrial Waters, of that which flows, runs, surges, forward and down, arriving always at profounder levels. Chalchiuhtlicue swells the fruits and flowers only so that they may drop.

Chalchiuhtlicue fills the gourd of pulque so that man may forget. Chalchiuhtlicue, current that descends in the course of rivers, current that flows in the course of time, making men's wastes and the implements of war inert and taking them to their inevitable end.

Tláloc, God of Celestial Water, god of the vapor that rises, from the earth warmed by the sun after the rains, god of the mist that ascends from the valleys at dawn, god of the water that returns to its source in the clouds that swim over the highest peaks, god of the humid incense from which rise the copal prayers and the prayers of sacrifice.

Tláloc is the return of vapor that strains to rise, is the return of time that strains to remember… Tláloc, God of the Fight Against the Current, with whose aid the hero battles against the torrent toward his own origin and beginning, towards the wings of his soul, the wings that Tláloc hides in the hero's past.

Quetzalcóatl, the Plumed Serpent, moves between gods and men, because Quetzalcóatl is God who permeates man and is the man that achieves God. The Plumed Serpent is born when that which slithers over the Earth grows wings to be elevated to Heaven. Quetzalcóatl is a superior man, the inner circle of humanity, the link between gods and men. All men are made of earth, air, water, and fire... But in their hearts and in their semen, each man has his own coatl, his own serpent, the energy of Tonatiuh, the power of the sun itself. And in this serpent sleeps consciousness, in this serpent is hidden his divinity. From this serpent his wings will grow.

There is an occult energy in the heart that comes from Tonatiuh, the Sun, and if man releases it, returning it consciously to the sun, he becomes immortal. But to liberate this energy, sacrifice is necessary. Man must sacrifice the desires and habits that he adores, sacrifice them in himself, and turn the knife against the enemy that he carries within himself, that keeps his heart a prisoner.

In recent times men still remembered these words, but they have now forgotten their significance. They have made enemies of other men to sacrifice them and take out their hearts, believing such offerings would propitiate Tonatiuh. Such is their degeneration, such is their superstition. When fear unites with knowledge, terrible things are done. It is the self within ourselves that we have to sacrifice. It is our own heart that has to be torn out of the false being and offered to the light. May Xiuhtecuhtli, Lord of Fire, burn my false being. May Itzli, Obsidian Knife, liberate my heart.

What is the blood that connects stones with soul, men with suns? It is the universal unity, the one creative principle crystallizing into myriad forms, and when liberated by sacrifice it returns to unity. Because to sacrifice is to act consciously, to sacrifice that which will be taken away is to deny the destiny that takes it.

Defraud Death by sacrificing Life; because from the hand of the Goddess of Sacrifice that holds the obsidian blade sprouts the germ of Life to come. To sow life in the fires of sacrifice: Are the flames rising on the Hill of the Star, toward Cauhlacan? Or is it my own heart that is afire?

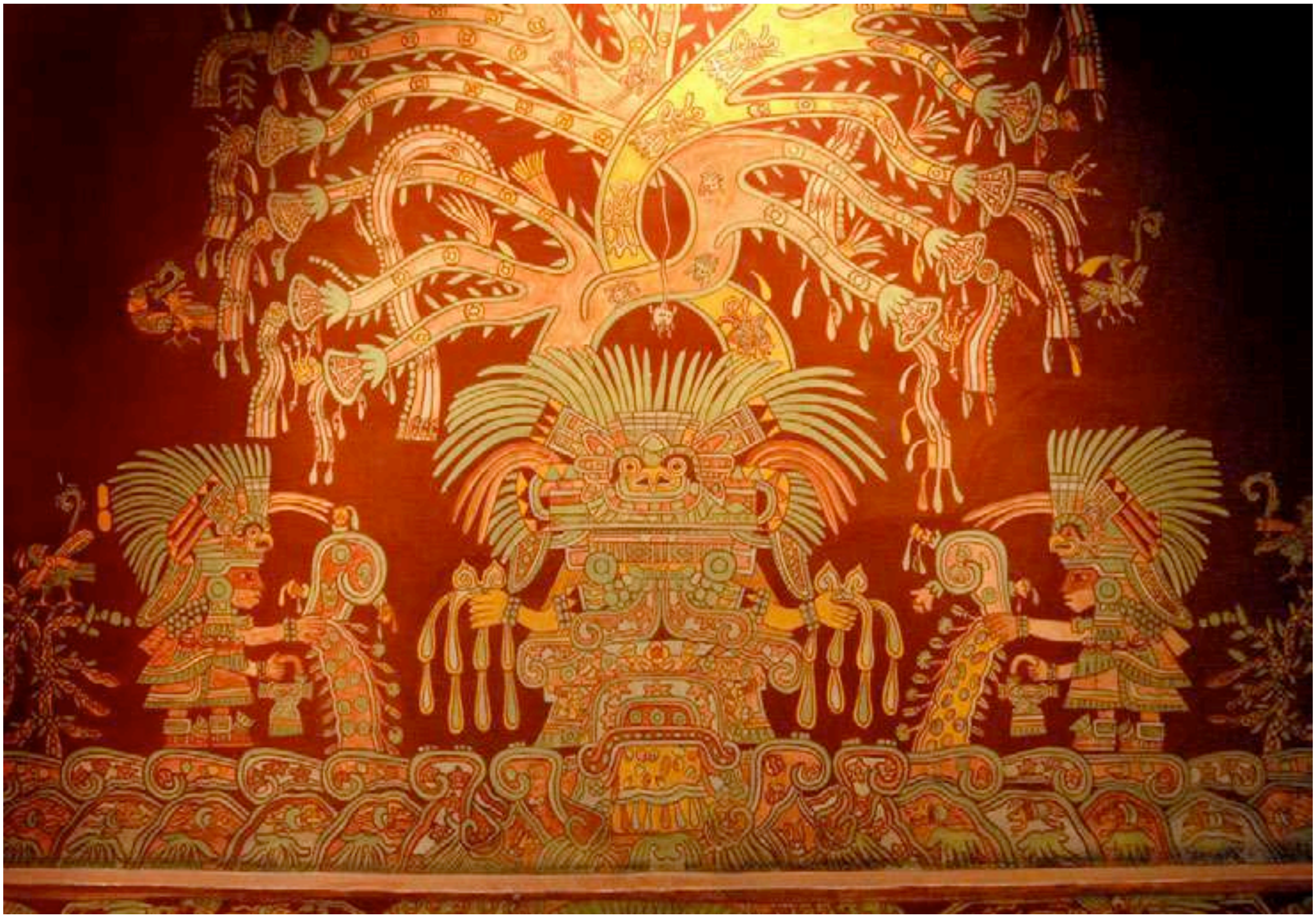

> The cycle of years is past - the waiting ended. Come - reunite, pilgrims, for the sky is in flames! From Xochicalco to Teotíhuacan the red spreads one step, another step, and another, only twelve short steps from the cave of the womb to the final conflagration… The flames rage in consummation- Rise, oh flames! What light, what heat! What great holocaust! The smoke ascends and spirals... Obscurity disappears as the flames rise to Tonatiuh, the purest light.

Profound mathematical knowledge of the rhythms of the inner planets and their specific effect on the patterns of the collective consciousness of the human beings of Earth is followed by an explanation of the attributes of different kinds of water. Sacred waters are identified with memory, wisdom, ascension and prayer and are known to science as water molecules exclusively comprised of the lightest form of hydrogen – *protium*. The lower forms of water contain heavy hydrogen isotopes – *deuterium and tritium* (with 2 and 3 neutrons, respectively), and are associated with forgetfulness, degeneration and entropy.

The Maya understand the inner power of human consciousness as a function of DNA, the coiled ‘coatl’ serpent within every living cell, to be mastered through meditation, spiritual development and emotional control of brainwave and DNA states. DNA electrification is overtly expressed in the straight, double ladders depicted on a large copper/gold plate from the Crespi collection from Cuenca, Ecuador, with double-headed serpents representing electrical currents moving along nanoparticles chains (opposite).

The prophecy of the Red Sky is related by Maya texts as an ignition of the pyramids and the very sky itself in flames! Atmospheric HHO plasma plumes and red atmospheric aurora are depicted in the Teotihuacan Tree of Life mural (above), the blood red sky encompassing the entire background of the scene, dominated by a giant tree-like form comprised of luminous pink and green plumes. Along the length of the spiraling ‘branches’ of this sacred tree are numerous concentric circles, as a well as spiders and butterflies – symbols for the transformative effects of the living waters of Tláloc.

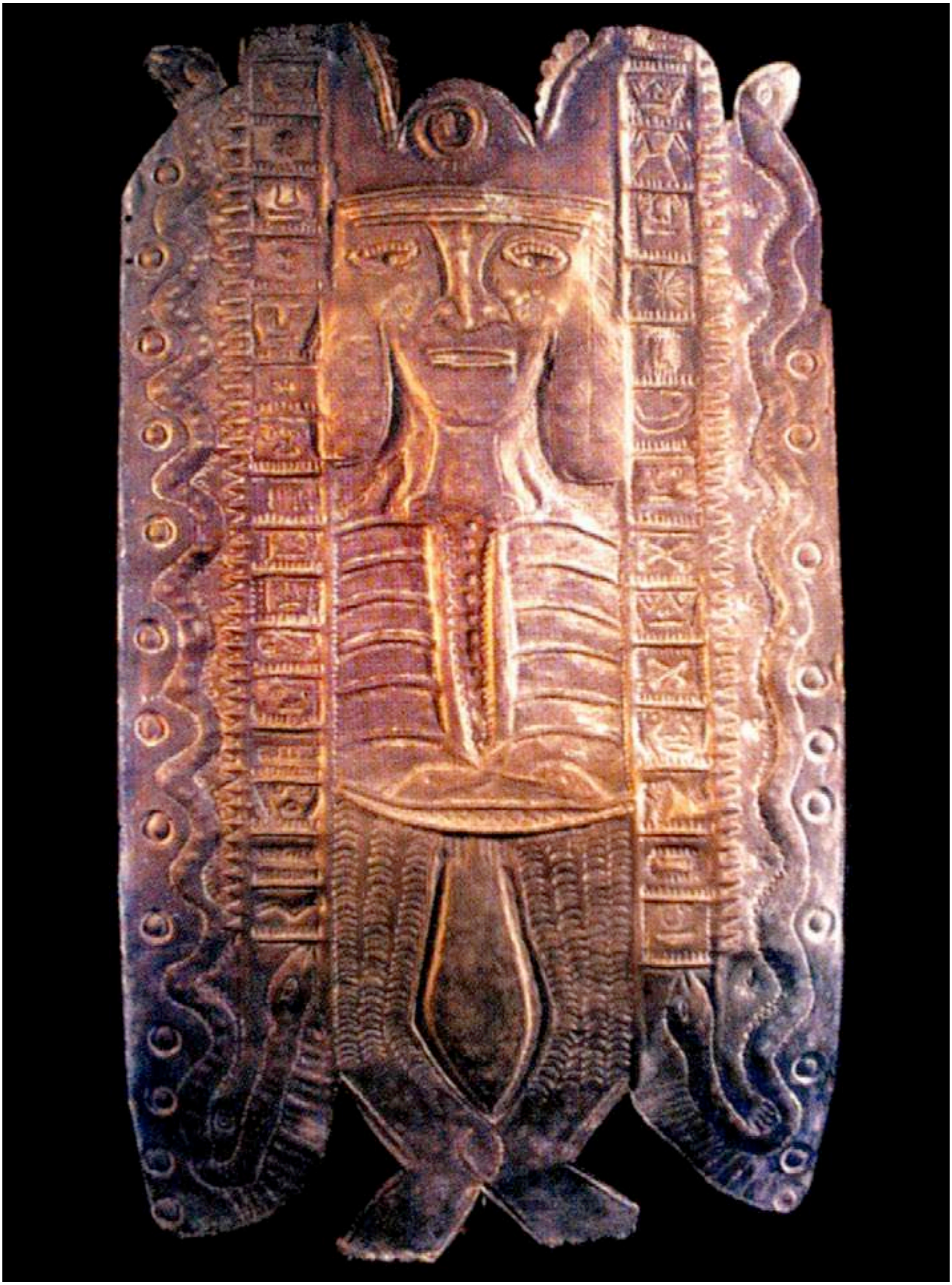

The concept of defrauding death by self-sacrifice through the overcoming of human instincts, karmic attachments and desires as symbolized by the laughing skull. Fear of death and the unknown is transcended in the eternal light of knowledge, thus ‘making death an ally.’ The laughing skull symbolism is preserved in ancient Tayrona gold masterpieces from Colombia, often presented with linear alignments of concentric circles (below, left). Three spirals are another commonly used symbolism representing the three frequencies of sound that stimulate psychoacoustic resonance within the temple chambers.

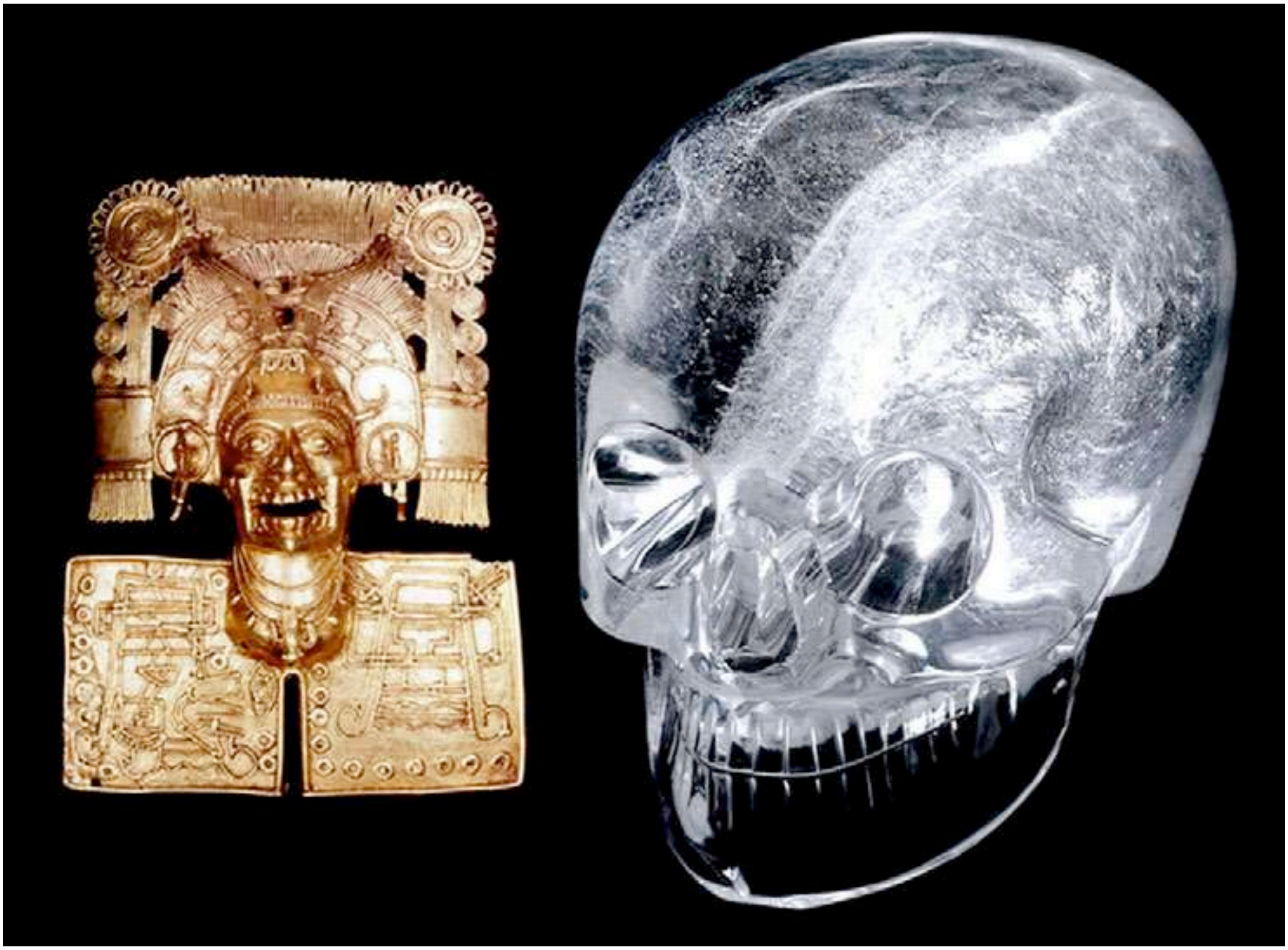

The crystal skull is another cross-cultural symbol closely linked with the transformative concept of crystallization of the water within the human body, referencing the piezoelectric property of quartz and the piezoelectric potential of the crystallized human being. The advanced bioelectrical metabolism of superior man is achieved through the death of individual desires and karmic attachments, to become spiritually reborn as Quetzalcóatl in the pyramidal House of Hunab K’u, the Universal Creator:

> The mysteries of the sacred Crystal Skulls have been recorded in the depths of time. That means that it is almost impossible for the human beings to understand them. This is the reason why thousands of people around the world feel restless; they want to know why the ancient peoples made these skulls… The Crystal Skull and the human being make up a single unit, but the main difference between them is that the skull has more memory than us [human beings].
>
> When all the crystal skulls of the world get together, it will be an important moment for the human beings because the memory that has been stored in these skulls will teach us how to recall our past lives. Currently, the modern human being does not know how to remember; and due to this, he does not know where he comes from or where he is going. And he does not know about his future either because he cannot remember where he came from or where he lived his past lives exactly...

There are some impressive geometric shapes that have been appearing in the fields of the European country of England, but the most impressive is the one that appeared on July 16th, 1999 in Windmill Hill, Avebury Trusloe, Wiltshire County [presented above]... As it can be seen, it is a large circle with a square in its center. Besides, the square has 328 small circles distributed symmetrically in its interior… The information below was published in the magazine "Crop Circles" where Judith Moore and Barbara Lamb wrote the following: “This 'Crop Circle' is the DNA chip that will evolve our mathematical capacity. Each one of these circles in an infinite fractal crystalline code, and each crystalline aspect is a fractal universe. This evolved system is the next step for mathematics on Earth. When this particular glyph becomes of collective knowledge, current mathematics will seem obsolete or primitive, like the difference between a stone axe and the laser.”

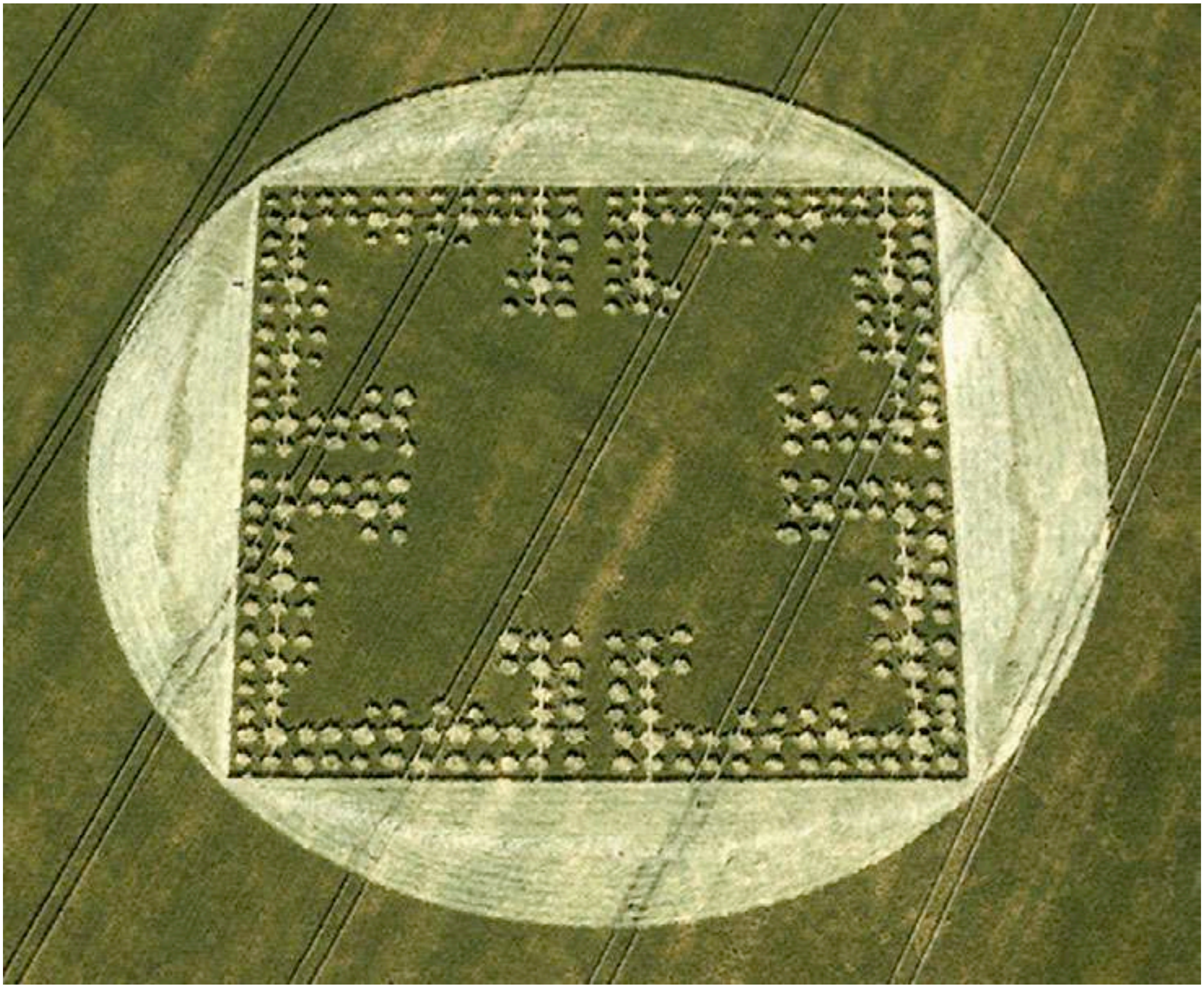

Qiuayon soen tra soen nas tra ka mix na ke te son tra laixwa mei ix nika. This is the ancient mantra for the activation of this process. In ancient history there were bearers of both the Mayan and Egyptian tradition that prepared the planetary essence to receive this evolutional leap. The crystalline codes were implanted in sacred sites and will be reactivated by those of you who bear the appropriate DNA codes to complete those ancient formulas and receive the mathematical formulas created in the collective mind. Those who will be part of this cosmic mission have already been reincarnated to activate this process…

The geometric writing exists, but the human being in general still cannot understand it. In ancient times the peoples mastered it; for example, the Mayans could understand it and they applied this knowledge on the planning of their constructions when they built their ceremonial centers and they also applied it on their architecture. They knew that energy moves in geometric shapes using its perfect proportions.

The Greeks knew that geometry was important to understand the mysteries of life, but it was the Mayans the first ones to include it in their culture to live it everyday. The Christians speak of God, but they do not relate him to the geometric shapes. In the Mayan culture God and science are related to the geometric shapes; for example, the circle and the square are the symbol of Hunab K'u who is the Mayan God. The circle represents the movement and the square the measurement. 'The Giver of Movement and Measurement' can be found in these two crisscrossed geometric shapes; this is for the Mayas the Universal Creator...

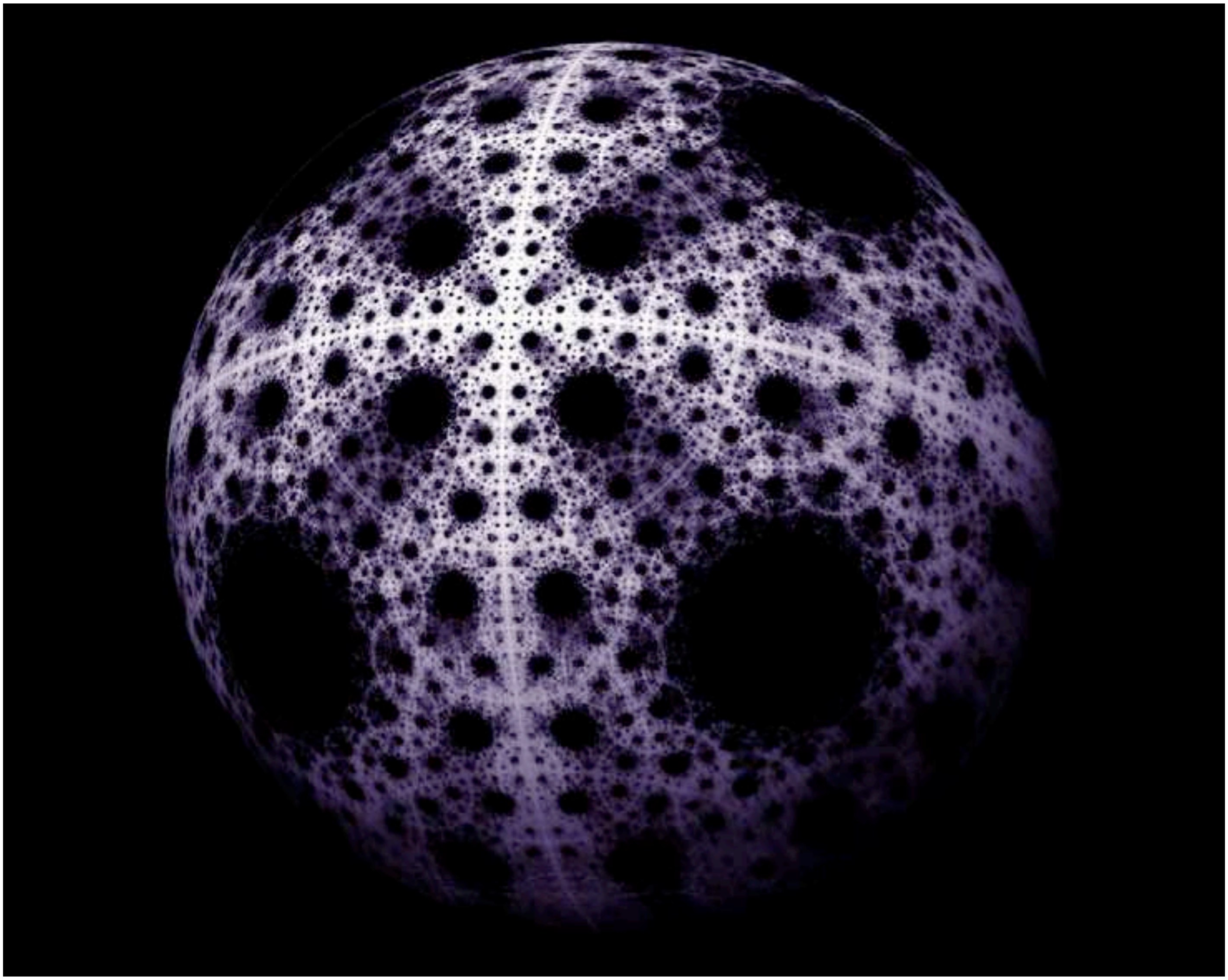

On the initiatic work of magnetic activation we will carry out with the Crystal Skulls in the Mayan temples we will use this same geometric shapes that appeared in England. In this occasion, all the initiates and teachers will experience these geometric shapes applied to the construction of the Mayan temples, but we will also have to live them cosmically. If we observe the map that we will be using in this initiation, we will realize that the ceremonial centers of the ancient Mayas were based on a geometric planning viewed from the space just as it can be appreciated...

This map shows us that the Mayas built their ceremonial centers using the macro-geometry. We the initiates will use the micro-geometry when we perform our rituals to later connect it with the macro-geometry. Well, then, who will help us to expand this energy in order to achieve the activation of the pyramids? Well, that will be the role of the sacred crystal skulls because science has proved that quartz is able to expand energy much more quickly than human beings… Let me tell you that this time the flowing energy will be used only for the human beings' benefit. As we previously said, the crystal skull and the human being form a single unit. When we work with this energy in the Mayan pyramids, then they will be activated and this energy will travel to the 8,000 ceremonial Mayan centers distributed in all the countries of the Mayan world to later continue its flow toward the other temples spread in the rest of the world...

> The Crystal Skulls will work on the magnetic activation of the pyramids and in a natural way they will make several radiations of the cosmos to come to us. These radiations will concentrate on the axis of all and each one of the pyramids and of course in their center, too. The pyramids are huge accumulators of electromagnetic energy that will be activated by the crystal skulls together with the initiates and teachers in the Mayan temples in order to seek the human benefit. It is also important to send this energy already activated to the center of each human being skull so that he can raise his consciousness.

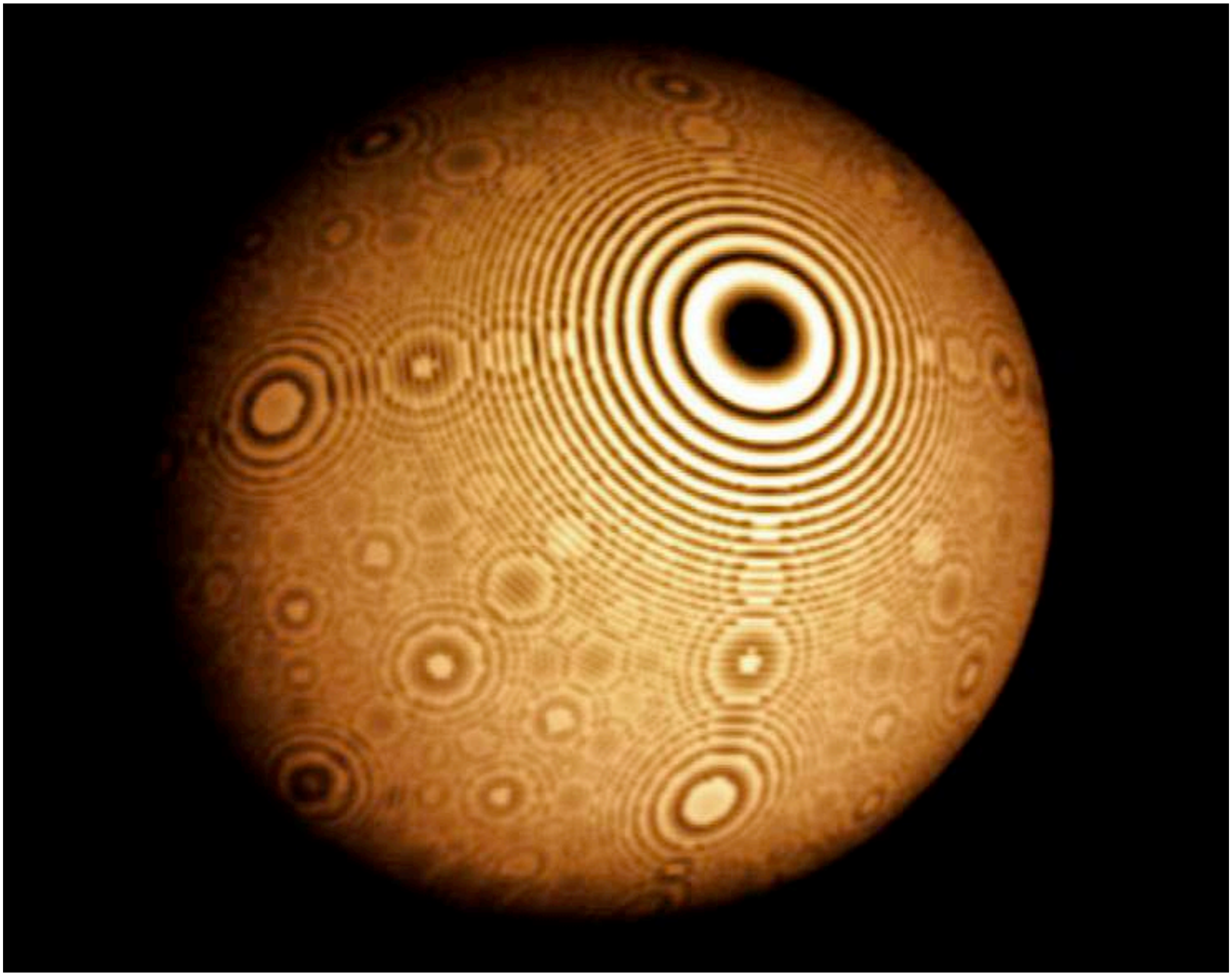

> As a conclusion, we will quote an important fact related to the pyramids. It is worth mentioning here the scientific theory that talks about the atomic excitation that takes place in the Sun, the stars and the cosmic space… that is able to penetrate all the existent things. Well, when this cosmic excitation penetrates into the pyramidal constructions, it gets organized and forms perfect geometric figures, giving birth in this to the cosmic geometric writing… We will do this work together in benefit of the natural laws, our Mother Earth, and all the human beings. It will also help us to start a communication with our cosmic family. This initiatic work will open the gate of the New Itza Age and the future generations will have a better world to live in, a world with a new consciousness achieved through this magnetic cosmic activation.[9]

The force of cosmic infrasound so penetrates all matter and becomes focused by giant crystalline pyramids on Earth’s surface, into a nonlinear distribution pattern exactly encoded by the spherical resonance formula [$z_{n+1} = z_n^2$] (above). Standing auroral beams and atmospheric flashes are being caused by localized surges of terrestrial infrasound at nodal convergence points, irradiating water vapor to induce the dissociation of water molecules and the rapid formation of HHO plasma. The implosion of hydrogen atoms releases intense ultraviolet-A and infrared light that becomes visible during brief periods.

An intensely luminous HHO plasma canopy envelops the Giza pyramids when operating at high resonance, radiating giant acoustic plasma beams that arc from the four faces of each pyramid, with looping flares connecting one pyramid to another just as solar flares arc from one sunspot to another. This dramatic luminous interaction of the pyramids with stimulated atmospheric plasma activity is exactly described by the Pyramid Texts of Saqarra, Egypt (below), only now comprehensible to plasma physicists as intense emissions of UV-A and IR light from resonant HHO plasma and amplified by the gleaming UV fluorescent limestone's of pyramids all over the world:

> House bright and dark of Heaven and Earth, for the [solar] ships put together; Great Pyramid, House of the gods with pointed peak; for Heaven-to-Earth it is greatly equipped. House whose interior glows with a reddish light of Heaven, pulsating a beam which reaches far and wide; its awesomeness touches the flesh… House of Eternity; its foundations are stones, the water; its great circumference is set in the clay.[10]

Calcite is an ultraviolet fluorescent material that was used extensively in the construction of all of the pyramids of the world. The ancient Vedic accounts of antigravitic vimana, or solar ships, include recipes for electroluminescent metal alloys which display the same molecular properties as the pyramid stones, as well as showing the same octagonal axis-symmetric geometry. If antigravitic effects were once achieved in the pyramids for application to the human body and consciousness, what is the range of extra-low-frequency sound biocompatibility? The heartbeat - at 1.45 Hz wavelength of 765 feet - informs the base length of the Great Pyramid of Egypt. The word 'pyr-a-mid' itself is of Greek origin, based on the word 'pyros' or fire and meaning 'fire in the middle'. Were biocompatible plasmas once burning in the pyramids' inner chambers?

The recently rediscovered high-energy state of hydrogen has been reported by Dr. Randall Mills, who has developed the process which releases vast amounts of energy stored within the hydrogen atom itself. Mills has defined the physics underlying this novel chemical reaction derived from water as hydrino resonant-transfer plasmas, which are now being commercially marketed for power generation by his company Blacklight Power, Inc. A very similar application of this novel reaction has been developed by Denny Klein, called HHO gas, and is now available as a plasma torch capable of sublimating any material. Both of these hydrogen-based plasma reactions produce intense infrared and ultravoilet light. Interestingly, the bandwidth of UV radiation produced is restricted to UV-A, which is beneficial to the human body, while the destructive UV-B and UV-C wavelengths are not observed.

The technical process by which HHO plasmas are formed has been defined by the Blacklight Power researchers. It involves the breaking of the molecular bonds of water by vaporized metal nanoparticles, producing a highly kinetic vapor composed of individual hydrogen and oxygen atoms. This reaction cannot occur in the presence of air, therefore a vacuum chamber is required. Once the HHO vapor fills the chamber, it is then electrified to become HHO plasma releasing energy as the H atoms shrink to one-quarter normal size.

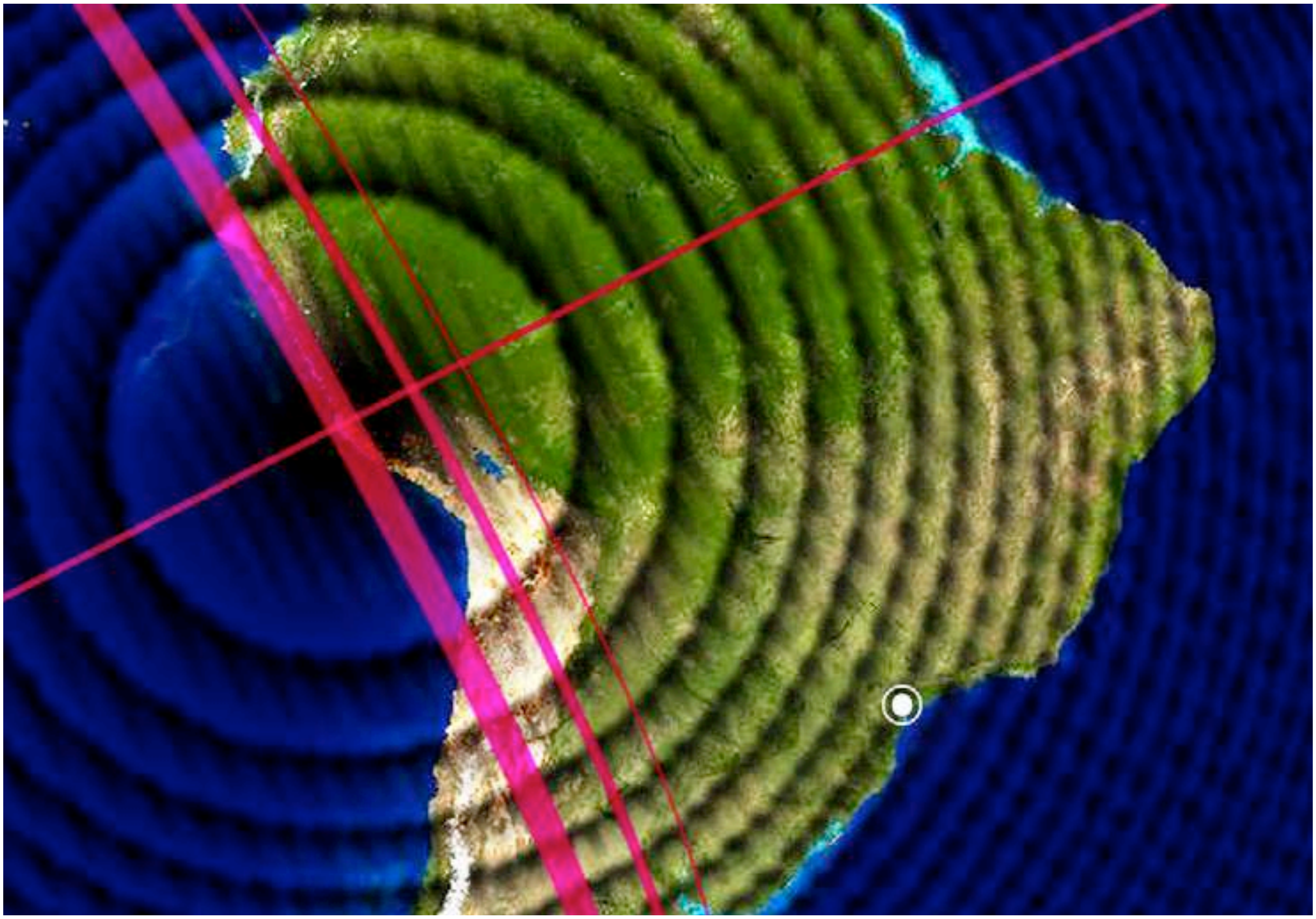

These contemporary scientific breakthroughs shed light on the unusual experiences recounted (in Search Magazine, March 1960) by American ex-patriot Dr. Raymond Bernard in Joinville, a coastal city in the Brazilian State of Santa Catarina (above). A very strange series of extraterrestrial contacts began while Dr. Bernard was organizing and living amongst a community of vegetarians outside of the city. The calls of a booming voice from a nearby mountaintop prompted a member of Bernard's camp to investigate the location of their origin. Eventually finding only a well-like circular shaft, measured by jungle vines to be about 300' in depth, the investigator returned equipped with ample rope to descend to its bottom:

> When I came to the bottom I entered a tunnel, which I traversed for a distance of 2,000 meters. I then saw with my flashlight a door of stone. While I was watching the door, it suddenly opened and I saw a very tall man with a metallic uniform, who spoke with a powerful voice and said that this was the first time that anyone ever had the courage to enter this tunnel. I first was frightened at the powerful voice of the man and wanted to run, but he called me and told me not to be afraid, as he was a very peaceful person who never did harm to any living being. So I asked him who he was to live in this cavern.
>
> He said, "I am an Atlantean-Inca, the guardian of this door."
>
> He asked me what I was looking for. I told him I was looking for my race, because I too am an Inca. He seemed to be very pleased to hear this. I told him I would like to visit his city, and that I had an American friend, and a wife that would like to go there too, with my children...
>
> On this second visit to the Atlantean he delivered a lecture on radioactivity and its danger. He said that radioactive dust in the air is causing surface dwellers to age very rapidly because it accumulated in the pores of the skin and stops skin respiration. He said the pores of the skin have a constant alternate contraction and expansion and serve to take in air and expel foreign matter. In an atmosphere of radioactive dust metals, there substances clog up the pores and interfere with their vital functions of respiration and excretion. This causes disease and early death.

The Atlantean called me to come near a transparent screen to make an exchange of blood to make sure I will preserve the secret of the whereabouts of this tunnel, so that I do not reveal it to an unworthy person. When I was near the curved plastic screen, another curved plastic screen suddenly appeared from each side of the door and encompassed me, so that I found myself inside the two curved plastic screens. Then it seemed that all air was pumped out, leaving a vacuum in the space where I was between the plastic screens. Then other air entered, which seemed lighter and purer, which made me feel more healthful and stronger. Then the inside plastic door opened, and he put inside a man-sized capsule of this transparent substance which he held by handles on both sides of it, and told me to go inside the open door. After I was inside, the door automatically closed. Now the second plastic door opened and the Atlantean took hold of the handles on each side of the capsule and carried me to the other side of the inside plastic screen whose door automatically opened and shut...

He told me not to bring here any unworthy person, because he would know it in advance. I asked how. He said that with this apparatus, which he called an "electrovisor," he could behold whatever was occurring in any part of the world. If any unworthy person comes near the mountain and tries to get to the tunnel opening, certain rays confuse the person's mind, so that he is unable to continue the trip and will go off in a wrong direction.

I had some bread in a pocket. He told me not to eat the bread. He put a white pill in my hand (which was projected outside the capsule all the time), and told me this pill had the taste of many fruits. Then I withdrew my hand and the plastic material through which it projected closed. After he gave me the pill he carried the capsule with me inside to the space between the two plastic screens, as the door of the inside screen opened to admit the capsule and then automatically shut. Then the door of the capsule opened. I left the capsule, and then the door of the outside plastic screen opened and I left.

When newcomers enter, they first enter the capsule, and the Atlantean carries them to a Decontamination Chamber (the Atlantean told me). The door of the capsule opens, the person leaves the capsule, takes off his clothing, then the chamber becomes filled with vapor, which draws forth radioactive poisons from his body. The person dresses with other clothing there ready for him, then enters an "electrical apparatus," which carries him [towards] the center of the earth. I should mention that during my visits, before the door opened I heard a peculiar humming sound, which was of the apparatus with which the Atlantean came up from below, which became louder and louder as the apparatus came near, and when it came to a stop the noise stopped and the doors automatically opened.

> I should also mention that when a flying saucer came near my house on Saturday, June 13, 1959, in the afternoon and night, it gave off an identical hum as this subterranean apparatus, which makes me think they are both operated by the same mechanism and that the flying saucer was an apparatus of those same subterranean Atlantean's and did not come from another planet. The flying saucer came describing a spiral which descended on top of a hill about 2000 meters from my house, and then rose in a spiral manner, with a tail of light behind it. It came in the night, had a round shape and a silver color. I believe it was sent by subterranean Atlantean's , who knew where I lived and came especially near my house.[11]

These details related by Bernard are all the more remarkable when one considers their timeframe, prior to 1960! The accounts of a peculiar loud humming that accompanied both the subterranean elevator and the vimana 'flying saucer' are now known to be ultra-low frequency soundwaves used for antigravitic control. The Atlantean's brief mention of technological remote viewing and rays that would confuse any visitor with negative intensions are also applications of ultra-low frequency waves being developed by today's scientists. Blood sampling for DNA study in the screening of visitors is also logical, as modern researchers have now been able to show a genetic predisposition to psychopathic behaviour.

Likewise, the lesson given by the Atlantean on radioactive dust metals in the atmosphere is confirmed by the ancient evidence of nuclear blasts in the archaeological record, as well as by modern indiscriminate use of depleted uranium (DU) weaponry. Modern medicine has also recently developed the process called electroporation that was described to Bernard prior to 1960 as a natural cellular function that has become blocked by atmospheric conditions. The detailed account of decontamination processes involving electroporation is extremely scientifically accurate, as informed by newly developed heat-resistant piezoelectric plastics such as PVDF (polyvinylidene fluoride) that form the vacuum chamber into which HHO gas is flooded before becoming an electrified plasma.

Further investigations of the unique properties of superfluid plasmas have been conducted by a group at the Massachussetts Institute of Technology led by Dr. Ketterle. The group's experiments have defined the parameters necessary for the creation of superfluid plasma beams. An infrared light ray is used as a pathway that is then surrounded by a strong encircling electromagnetic field. These conditions efficiently contain the plasma when released into the IR beam. They have found that a pair of rotating green lasers within the IR beam induces the frictionless rotation of the superfluid plasma, which subsequently forms an array of hexagonal vortices that maintain the crystalline structure of the ionized gas.

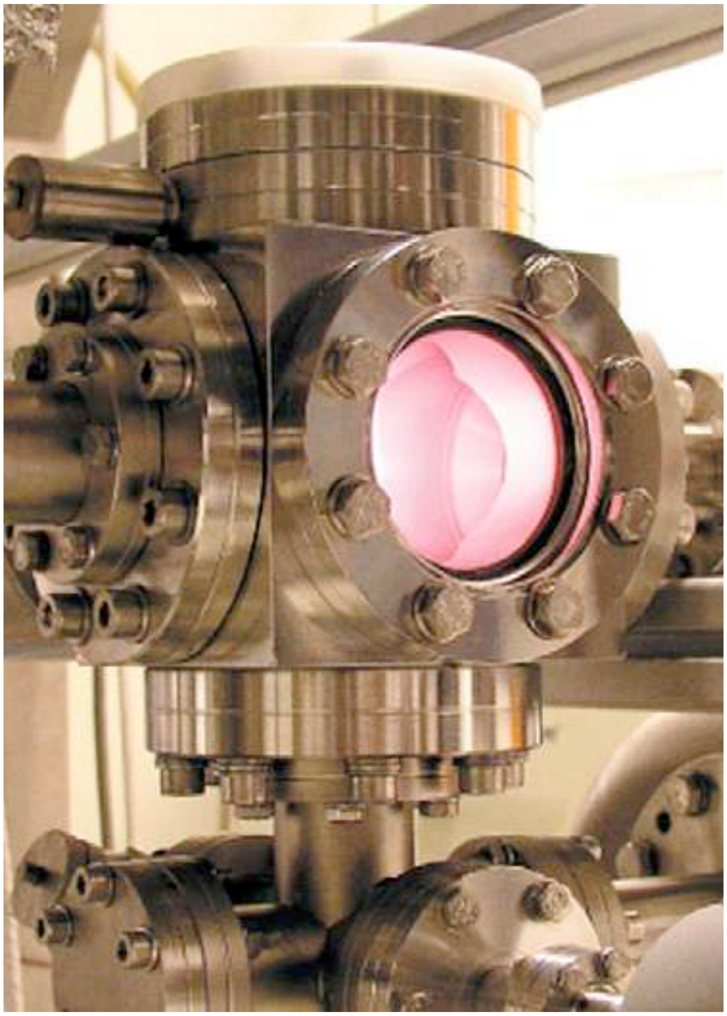

Ketterle's experimental results have apparently replicated the 'ultraviolet fog beams' related in the vast majority of extraterrestrial encounters. Physical transport via levitation in ultraviolet beams have been reported for decades, including riveting descriptions of the intense sensations of heat and light that accompany the events. The consistency of the contemporary eyewitness accounts of these advanced plasma technologies is astounding, and when analyzed in the context of the latest human technological developments one finds a profound synchronicity. Twenty-first century science has completed both the mathematical and experimental groundwork enabling a new comprehension of high-energy realms beyond common human experience.

Plasma is the fourth state of matter, with the highest energy and lowest density, followed by gas, liquid and solid states. The ignition of fire is a commonly witnessed event where gas becomes ionized to form plasma. The realm of human beings is a gas-filled resonating atmosphere in which we experience a three-dimensional world. A four-dimensional extraterrestrial realm of resonant plasmas passes through and beyond the human reality, apparently inhabited by electroluminous paraphysical beings.

While still higher realms of aphysical consciousness - thought at large - must exist, all human contact with this realm necessarily occurs via thought, dream, intuition, meditation or channeling. The channeled Cassiopaea contacts of Laura Knight-Jadczyk provide a very broad frame of reference and apt terminology which we will apply here in our discussion of the relationship of the gas and plasma realms –as 'third-density' and 'fourth-density' environments. Matter brought into the high-energy plasma environment decreases in density thus allowing molecular interpenetration. Evidence of the hyperdimensional interpenetration of matter have been found in the debris of tornadoes. One such anomaly has been described more than once - a windowframe with a wood plank passing right through its unbroken pane of glass.

Superconducting plasma aerial vehicles achieve frictionless passage through air, water or stone as there is no displacement of matter, allowing velocities and maneuverablity impossible for objects in third-density. The wave motion of acoustic propagation that is created in a third-density atmosphere of air is not present in fourth-density plasmas, which lack the temperature differentials that bend sound into waveforms. The fourth-density environment creates a coherent energy transfer, inducing unified telepathic consciousness without any information loss as minds are electromagnetically coupled in group synchrony.

The biocompatability of HHO plasmas accurately described in Bernard's account is quite similar to later reports of extraterrestrial encounters from other parts of the world. Michael Hesemann's breathtaking documentary film 'Ships of Light: The Carlos Diaz UFO Experience' presents overwhelming video and photographic evidence of fourth-density technology. Diaz vividly describes the initial sensation of heart throbbing at 1.45 Hz entrainment that precedes his extraterrestrial contacts. This powerful infrasound effect can now be understood as part of the propulsion system of the fourth-density craft, suggesting its invisible presence is maintained nearby before Carlos becomes aware of it. The craft then makes a luminous and visually dramatic appearance only when he has mentally prepared himself for the contact:

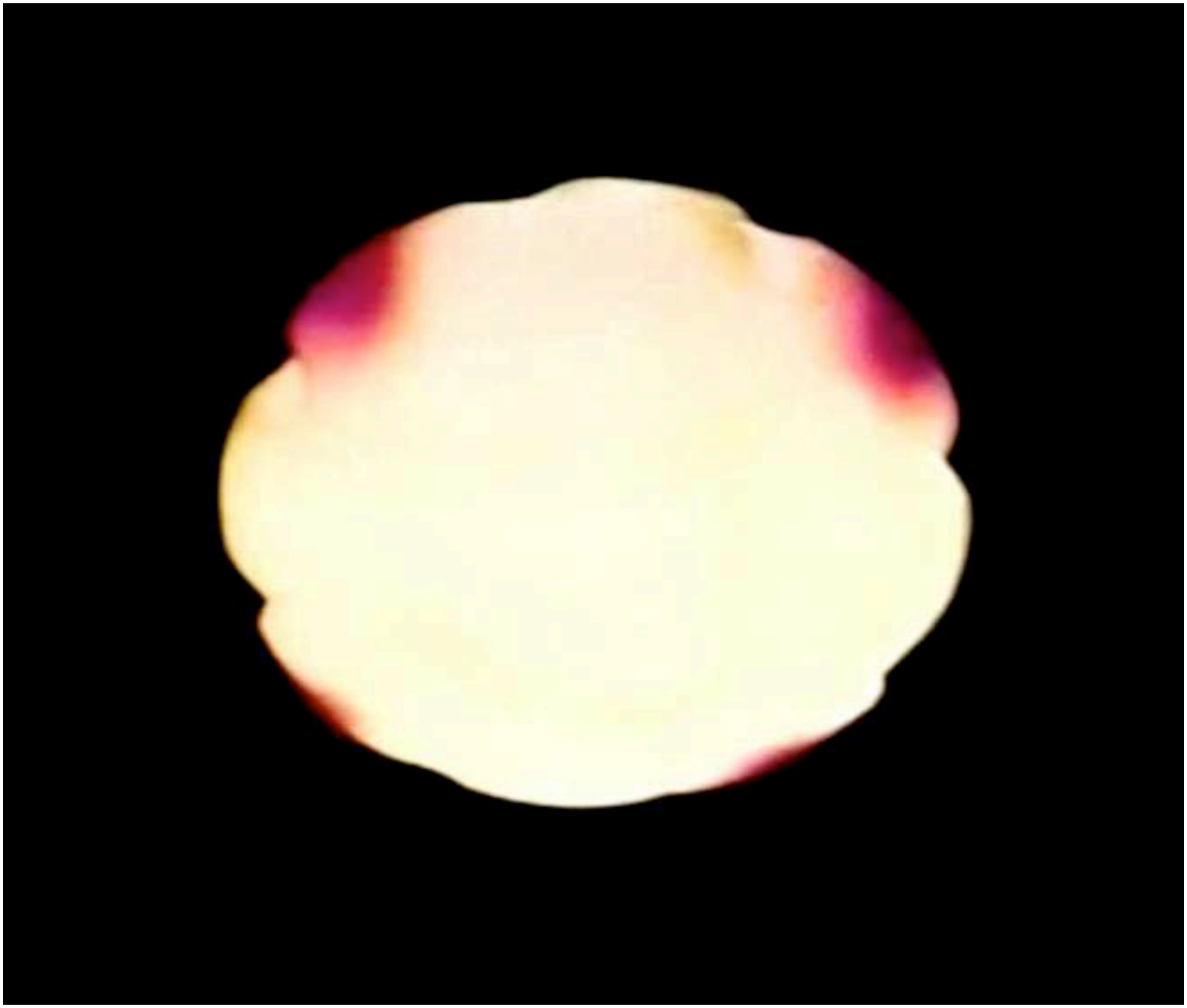

Q: In what form do the contacts happen?

There is a very special sensation concentrated in the solar plexus of my body – a sensation like when one is strongly in love. This sensation sometimes wakes me up at night, and I know that I am going to have an encounter with them. What I enjoy about the experiences I have had is that the communication I have with them is just as we are having at this moment.

Q: So it's an absolutely physical contact? You can touch them?

Yes, I can touch them... and at times they present a body of light...

Sometimes these ships appear at a great altitude, coming down slowly to hover eventually over the ground at an altitude of one and a half feet. When it is still at a high altitude the movements of the ship are very unstable, but when it hovers over the ground it is completely static. But it also happens that the ship suddenly appears right on the spot. For me, it is a kind of magic. They come directly out of the light slowly, part for part. Their body emerges out of the wall of light...

Their craft use the natural energy belts of the planet. Sometimes they enter into the belts, letting themselves flow with the energy. If they want to maintain one position they must act against the flow of the energy belts, causing the erratic movements of the craft. Once they are in the natural energy flow they travel at high speeds...

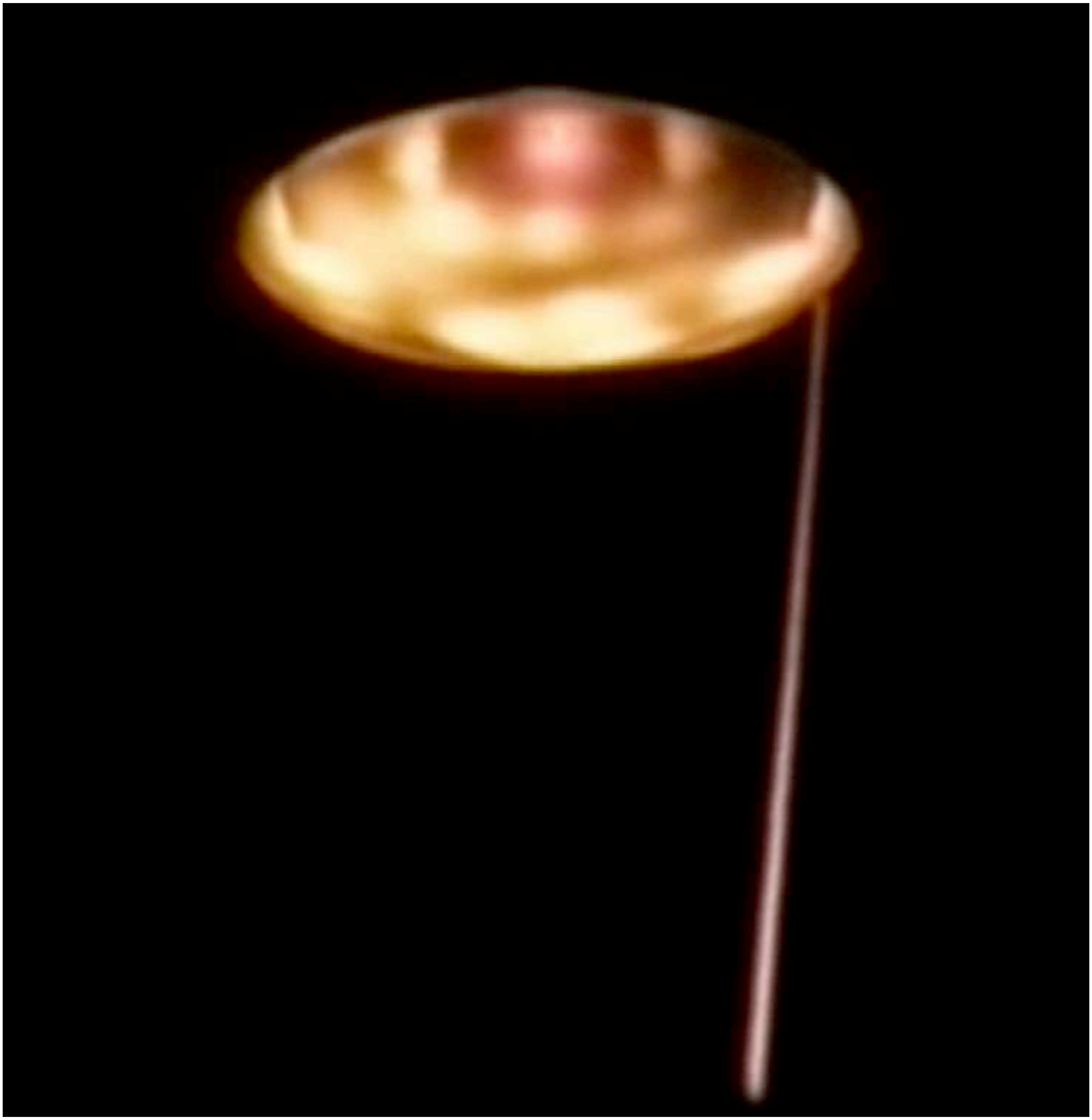

At different opportunities during my encounters I experienced how a beam of light shot out of the craft. This beam of light has the purpose to bring insects, plants, pollen, small and bigger animals aboard the ship, and sometimes we too, my friend and I, were brought aboard by such a ray. This beam directly forms and emerges from the ship, and consists of the same light or energy of the ship itself.

The first time he invited me consciously to go into a ship, I saw my friend getting in first. He went through this fog-like light. When more than half of his body was inside of the light his body got sucked into the object. So, what I did first is that I put one hand into the light to feel, and to my surprise it went through the light and inside I felt a nice temperature. When I took my hand out I was relieved to see it was in one piece, so that gave me the confidence to walk in, and when I had my body halfway inside the object I was sucked into it too. But once you are into it there is nothing you can see. You just see yellow light all over. You don’t feel your toes. You don’t feel the floor, but the sensation is very peculiar, because you feel very peaceful, with a lot of love within you.

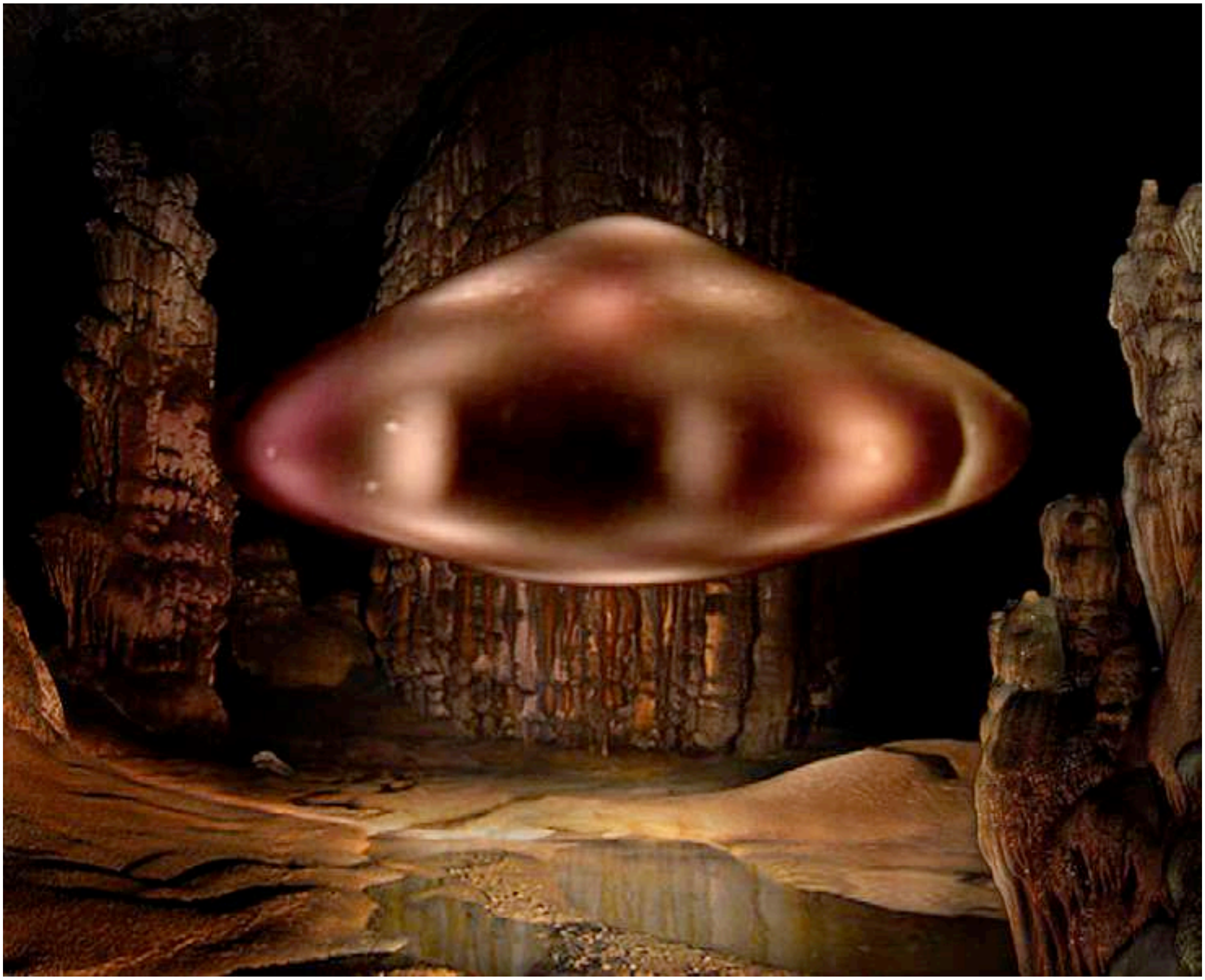

Once when I came out of the yellow light I realized I was not at the same place, but in a cave. It had stalactites and stalagmites, and between them I saw several sculptures and ornaments which were of the art of the Maya culture. This big cave was the place where the ship hovered and there was a kind of path we walked through and we came to another cave, and I saw there many people who were very kind to me and greeted me... But the particular thing about this cave was that there were seven egg-like spheres of yellow light. After we got to this place my friend stuck his hand into one of the eggs and took out a small ball of light, also yellow, and said that this little ball could project a ray that would touch my seventh vertebrae or at the middle of the spine and it would take all of the information in my neurons.

Once I was there in this place he invited me to go into one of those spheres or eggs of light. When I did so, I realized that all the yellow color changed, and it was not yellow anymore, but was a scene of a forest from the air. It was like as if I was flying. And then I saw on each of my sides a wing – the wing of an eagle, and I realized what I was looking at was the information from an eagle, his memories, that they had stored into this egg of light.

Once... my good friend Jaime Maussan asked me to ask them where they come from. So when I was with one of them the next time, we were walking near Mextitla, I asked him ‘Where are you from?’, and he smiled and said, ‘Look Carlos, I can give you the name of any star, any constellation, any galaxy, but what would this tell you? As far as you have gotten is your moon, so it makes no sense if I tell you where we come from since you can’t go there and verify. For me, this answer was enough even though when I was with them I once heard them to mention a star named Maya, and when I looked into an astronomy book that there is in the constellation of the Pleiades a star named Maya. Maybe this is where they come from but I really don’t know...

After entering these spheres or eggs of light, I understood that they store a great wealth of information in them, not just about human beings, but about all forms, all manifestations of life. They don't only store images, they also store senses. When you are there you feel like an eagle, like a bird – I felt joy, I felt fear – and I have seen many types of environments; the sea, the ice, jungles. What I have learned from this is that every living thing is important for this wonderful planet.

They also offered to show me what they stored about the past of my own life but I didn't want to know, because what I want to do is go forward, and to try to act with all the things I have learned to live this experience. I had experienced that there is a wonderful interaction between all living things, but this interaction went out of control. Every individual, every species is an important part of this interaction. This is the reason why I want to share all of this with the whole world – to invite the people to act in order to save the Earth – to make it a true paradise for all living creatures, and to inherit our future generations a living planet.[12]

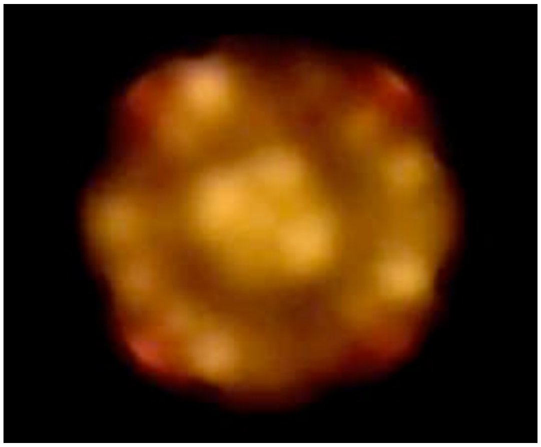

Profound experiences shared with Carlos through advanced transdimensional consciousness technologies have inspired a deepening understanding of the interdependence and unity of all life in the cosmos, and the urgency of the need to preserve the living systems of our mother Earth. The organic, jellyfish-like appearance of the plasma ships photographed by Diaz reflects a natural beauty by imitating the rhythmic bioluminescence of bacterial colonies symbiotically living within deep ocean creatures, which become stimulated to generate the brightly pulsating spotted displays of squid (above). Carlos Diaz's astonishing video footage of plasma beam technology has confounded the experts, including digital video and photo analyst Jim Dilettoso of Village Labs in Tempe, Arizona:

> The ray of light that comes down, when we look at it frame by frame, the light of the ray... [is] darker the top than it is at the bottom... and it goes off the same way. At each incident when the light goes on... the variance is between one frame, a thirtieth if a second. Well if [the ray] was [propagated] at the speed of light, we wouldn't even see it in a frame. It would be off and then on... which is curious. So, I'm still trying to understand it.

The false assumption that the ray is a simple beam of light has caused great confusion, as the intensity of the beam increases from top to bottom over the period of three video frames or one tenth of a second. The same three-frame shift in intensity from top to bottom is also noted when the beam recedes. The constant of the speed of light denies the possibility of a single source of emanation, suggesting that the beam is an emission of plasma. This notion is corroborated by the recent experimental results of Ketterle's MIT group, from which we can infer that the beam is initiated as a faint ray of infrared light subsequently flooded by HHO plasma in one tenth of a second. Knowledge of the properties of plasma beams and the effect of acoustic levitation by use of standing waves informs the authenticity of the images, as for Dr. August Meessen, Professor of physics at the Catholic University of Louvain, Belgium:

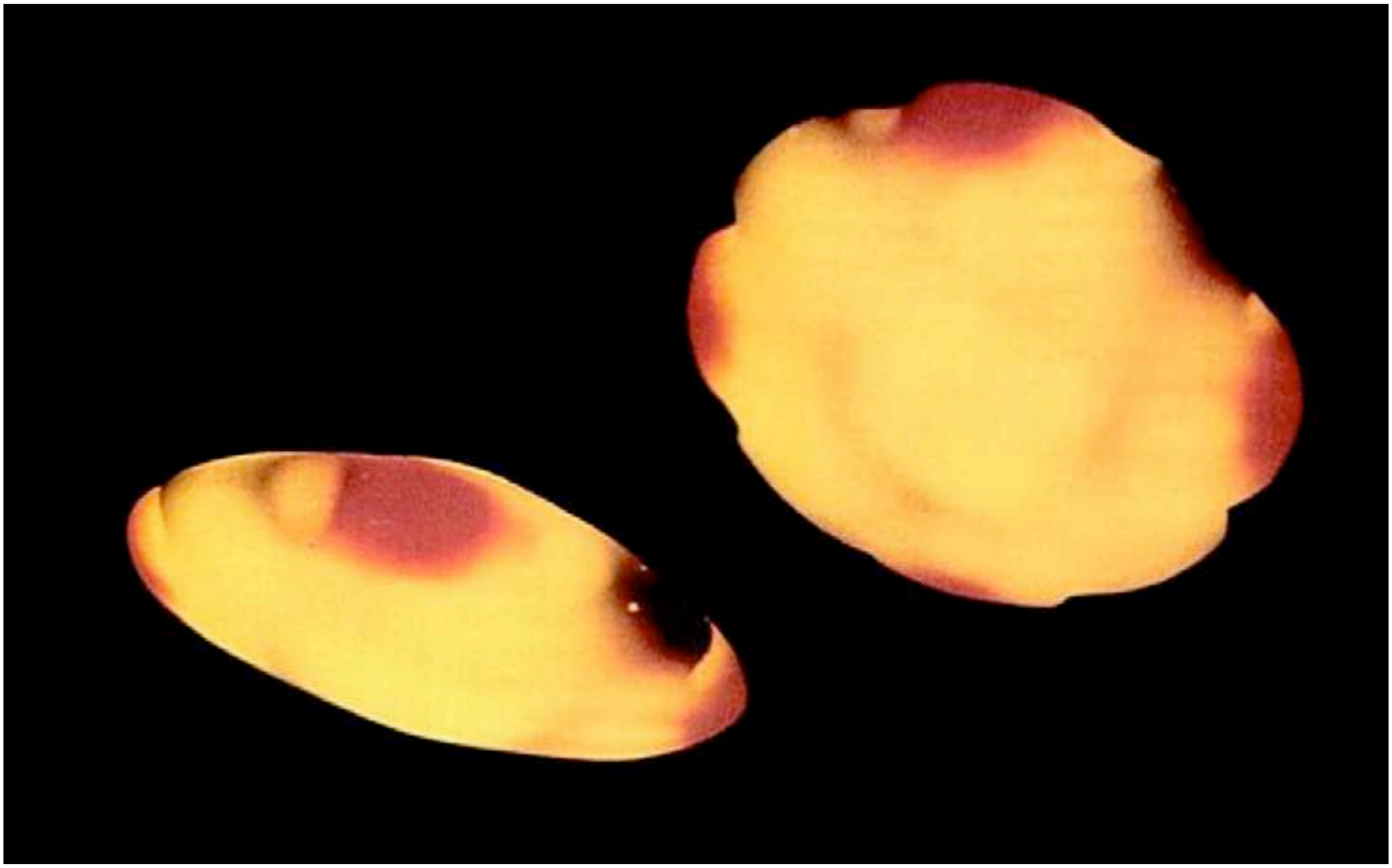

> As far as I see there are no objections against [the Diaz documentary photographs and videos], because if it is true it is very interesting for me as a confirmation of the theory where one expected results of this type... I developed a theory for the propulsion of classical objects... [which are] axis-symmetric objects... How do they produce electromagnetic waves which surround the object? ...One of the predictions of the [Unified Field] theory... [is] that they should be able to create *standing waves with certain modes where at certain regularly distributed places you have a higher intensity, thus you could eventually ionize differently.* If the theory is correct you could expect that somewhere observations like [those of Carlos Diaz] should appear.

Dr. Meessen accurately identifies the aerial activity over Tepotzlan, Mexico as involving ionizing spacecraft that restore energy by absorption of terrestrial infrasound standing waves that converge on the area, as described to Carlos as 'energy belts' by the extraterrestrials in contact with Carlos Diaz. Meessen has arrived at the same conclusions concerning standing waves as had Nikola Tesla, the original discoverer of the fascinating acoustic phenomenon, and have been applied by this author as spherical infrasound resonance maps of the Earth, flowering from the Giza Prime Meridian:

> The discovery of the stationary terrestrial waves... [indicates] that, despite its vast extent, the entire planet can be thrown into resonant vibration like a little tuning fork; that electrical oscillations suited to its physical properties and dimensions pass through it unimpeded, in strict obedience to a simple mathematical law...[13]
>
> [T]he Earth responds to a great number of vibrations and can be resonantly excited just like a wire of limited dimensions. When this takes place there are formed on its surface *stationary parallel circles of equal electrical activity*, which can be revealed by properly attuned [infrasound] instruments… Then *the crests and hollows of the stationary waves would be in parallel [or concentric] circles [enveloping the entire planet]*…[14]

Tesla's most profound discoveries in those last several months of the 19th century would go unpublished for almost 80 years, yet find direct relevance to the advancements of resonant technologies of the 21st century and beyond. The concise explanations of Nikola Tesla provide insight into the technological aspect of fourth-density experiences reported by extraterrestrial contactees and abductees alike. The enveloping brilliance of violet, yellow or golden light –combined with an intense sensation of the warmth of Eternal Love– are among the transdimensional experiences that leave many at a loss for words.

Many unmistakable aspects of advanced bioelectrical bodily purification processes using HHO plasma have been well documented. Closely echoing Bernard's 1960 report, Diaz also underwent plasma decontamination and was given fruit concentrate pills to ingest instead of the normal third-density metabolism based on the digestion of organic material.

The impeccable and extensive video and photographic evidence presented by Diaz documents the fourth-density conditions of intense light and heat of HHO plasma within the beam transport and in the plasma ship itself (opposite). This remarkable enhancement of the luminosity of the atmospheric conditions reflects the atmospheric environment present on the homeworlds of those space-travelers, affecting DNA enhancements not present in the terrestrial human populations of Earth. DNA evidence of this kind was actually obtained by Australian Peter Khoury, who has collected and preserved a single translucent hair from one of his several disturbing extraterrestrial encounters investigated by Bill Chalker:

> It was not until 1996 that I heard from Peter Khoury about the hair sample that had been recovered from what may have been an alien abduction sexual assault case… By 1998, I began an investigation into the hair sample, when biochemical colleagues agreed to undertake what was the world's first PCR (polymerase chain reaction) DNA profiling of biological material implicated in an alien abduction experience. The analysis confirmed the hair came from someone who was biologically close to normal human genetics, but of an unusual racial type - a rare Chinese Mongoloid type - one of the rarest human lineages known, that lies further from the human mainstream than any other except for African pygmies and aboriginals.

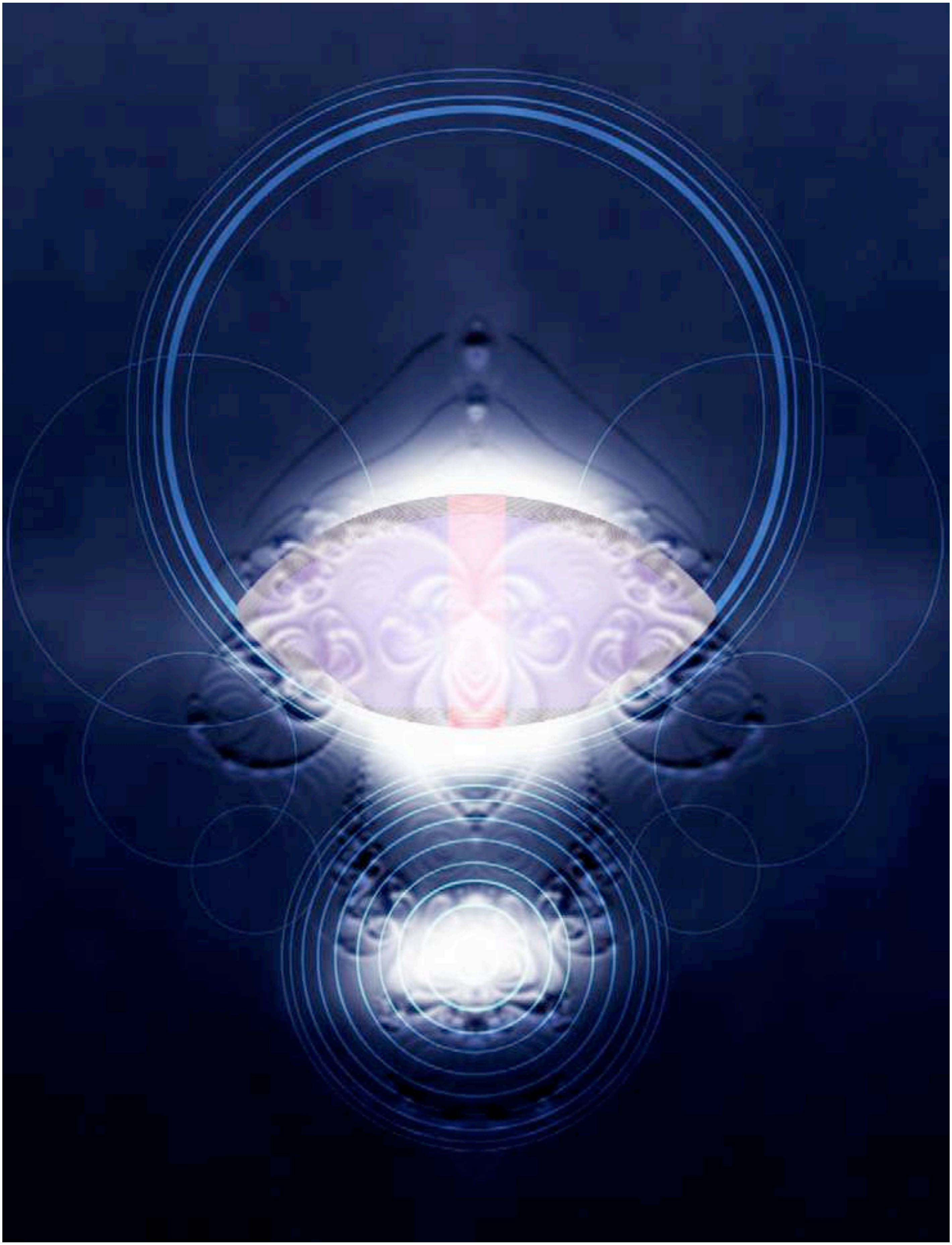

> There was the strange anomaly of it being blonde to clear instead of black, as would be expected from the Asian type mitochondrial DNA. The study concluded, "The most probable donor of the hair must therefore be as (Khoury) claims: a tall blonde female who does not need much color in her hair or skin, as a form of protection against the sun, perhaps because she does not require it [in a luminous fourth-density environment]..."
>
> The original DNA work was done on the shaft of the hair. Fascinating further anomalies were found in the root of the hair. Two types of DNA were found depending on where the mitochondrial DNA testing occurs, namely confirming the rare Chinese type DNA in the hair shaft and indicating a rare possible Basque/Gaelic type DNA in the root section... [The female's appearance is illustrated below, based on Khoury's detailed descriptions.]
>
> This was very puzzling and controversial, until a 'Nature Biotechnology' paper appeared in 2000. It revealed recent findings on hair transplanting with previously incompatible hair, using advanced cloning techniques, developed in a possible cure for baldness. We seem to be seeing similar combined or "grafted" DNA in the sample recovered under controversial circumstances by Peter Khoury back in 1992.

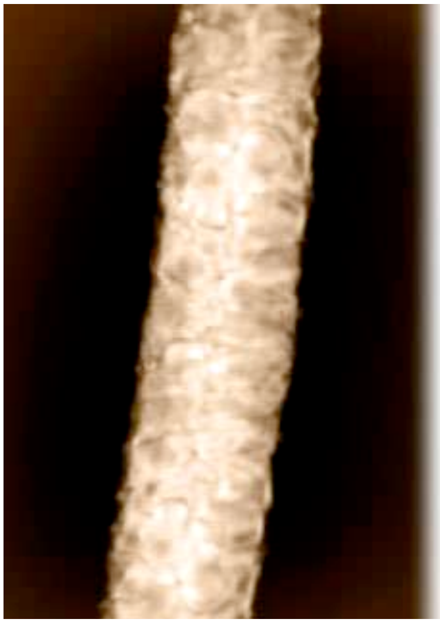

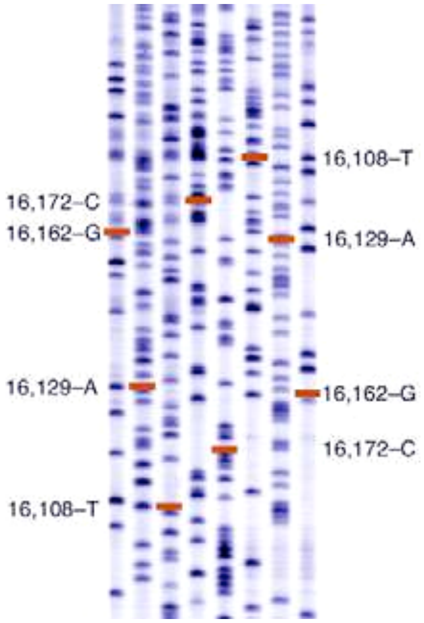

> Perhaps even more controversial is that we have findings suggestive of nuclear DNA indicating possible viral resistance. The hair sample seems to show it contains 2 deleted genes for CCR5 protein and no intact gene for normal undeleted CCR5 - this CCR5 deletion factor has been implicated in AIDs resistance. To keep a very complex story somewhat uncomplicated, what seems to be suggested by the range of findings is possible evidence for advanced DNA techniques and DNA anomalies and findings, for which we are only now discovering or starting to make sense of in mainstream biotechnology.

While the unusual experiences of Bernard and Diaz were of a very positive nature, the sexual assault case of Khoury is by contrast quite negative. Yet in all of these cases the powerful emotive interaction between the extraterrestrials and the contactees is obvious, as one might expect of primarily telepathic beings. Diaz provides extensive video and photographic evidence that support his experiences, while Oliveira and Khoury have collected the most convincing physical evidence of HHO plasmas and an extraterrestrial DNA sample.

When compared and contrasted, their corroborated accounts and evidence of hyperdimensional transport compellingly point to a single, remarkable conclusion: that technologically advanced Sanskrit cultures of the deep human past have not perished, but have in fact transcended our toxic third-density realm to achieve vastly extended lifespans in fourth-density plasma environments developed in subterranean chambers and flying vimana discs. The stark synchronicity of evidence from these extraterrestrial contacts with the emerging human technologies of acoustic levitation, HHO plasma, electroluminescent metals and piezoelectric plastics informs a broadened perspective. By synthesizing these scientifically described and extraterrestrially demonstrated technologies, humanity is at the threshold of a hyperdimensional transition to the unified consciousness of the fourth-density realm.

Advanced HHO plasma technology will soon significantly expand the consciousness and extend the lifespan of human beings living on Earth's surface, thus closing the circle of time to reunite with our advanced ancient ancestors concurrently dwelling in fourth-density. The ascension of human consciousness by self-sacrifice in HHO plasma chambers within the sacred pyramids endows telepathic communication and direct thought contact with fourth-density spacecraft in Earth's atmosphere. Fleets of plasma ships have been videotaped at high-altitudes over the vicinity of Mexico City in various large formations, including *long chains of ships that appear like floating helical DNA strands* (below).

Such odd and spectacular plasma ship formations are orchestrated to reawaken human curiosity and announce the presence of higher dimensional realities, while most fourth-density spacecraft pass through Earth's atmosphere without being perceived by terrestrial humans. Highly developed subterranean bases and cities have been in use for millennia by extraterrestrial colonies living deep below the Earth's surface:

> Two scientists affiliated to the Fundación Instituto Biofísico de Investigaciones (FICI) headquartered in La Matanza… claimed having picked up radioactive signatures, microwaves, levels of electricity and oscillations originating from the depths of the earth.
>
> Omar Hesse and Jorge Millstein surveyed the mountains surrounding Cachi last week -- 157 km from the capital and 2,280 meters above sea level. After applying a Russian tester in the vicinity of the Nevado de Cachi, they concluded that the signals are not natural in origin, and rather originate by machinery operated by intelligent beings… "The oscillations clearly indicate that kilometers beneath the surface there is activity: alternating electrical waves, which means a power source," said Hesse. "This could mean engines," he added.[15]

The constantly maintained acoustic signals of deep subterranean cities present convincing evidence that such highly advanced facilities are powered by infrasound standing wave resonance, in alignment with the pyramids on the Earth's surface above. The fourth-density experiences of Carlos Diaz involved visits to augmented natural caverns below the Tepotzlan, Mexico area (not far from the Nevado de Cachi site) that contained ancient statuary like that of the Maya tradition, as well as seven yellow luminous plasma eggs composed of the same HHO plasma that also surrounds the spacecraft as a radiant red aura.

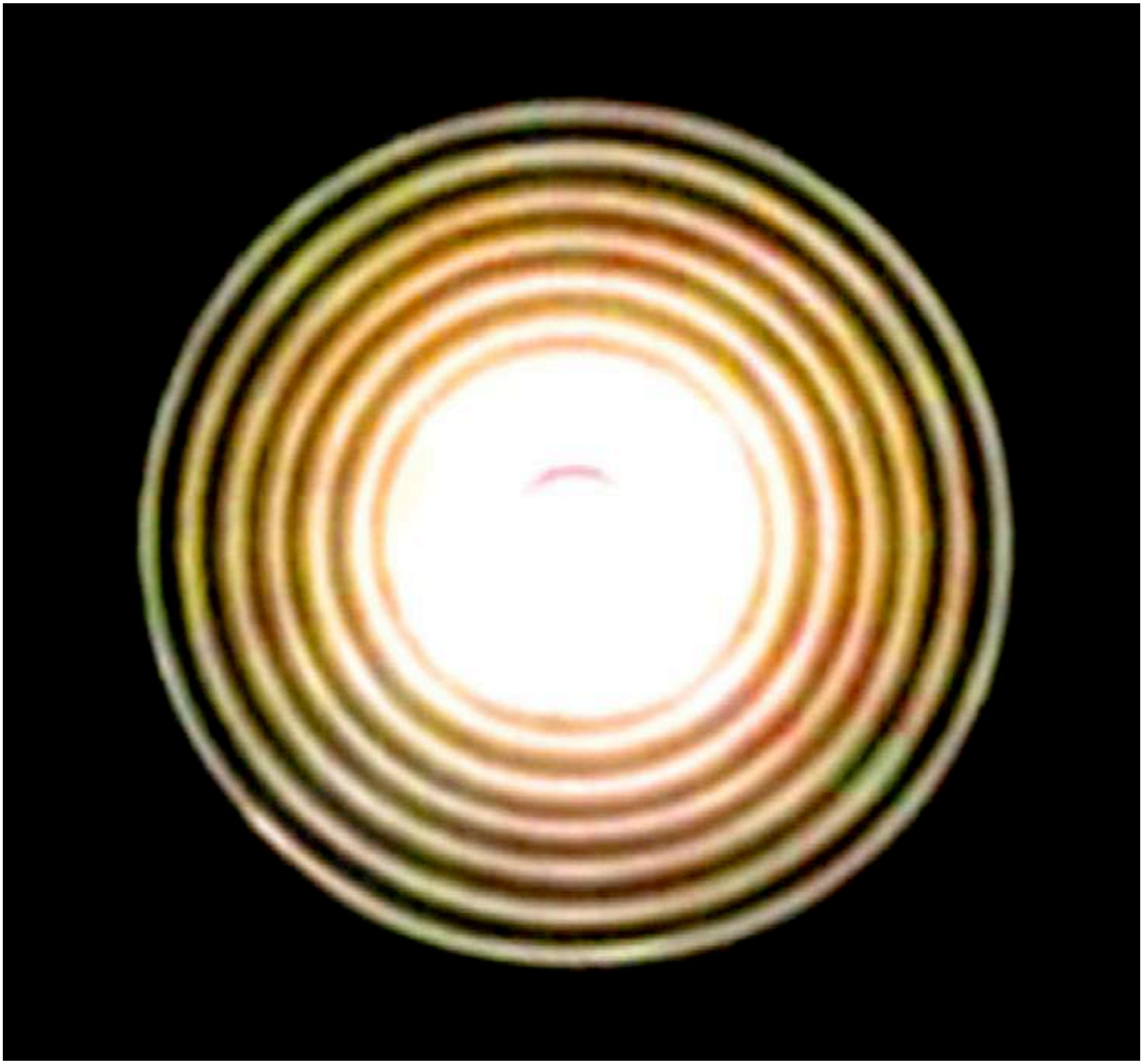

Transit to and from these deep underground cities is achieved by transdimensional passage directly through the bedrock by retuning the atomic frequency of the entire ship, its interior HHO plasma atmosphere and crew to safely interpenetrate the material realm. The transdimensional ships' generate thin veils of HHO plasma surrounding the vessel that can instantaneously shift the frequency of all emitted light into the infrared range for invisibility cloaking, or be tuned to generate visible light of any desired frequency. Security camera video from Turkey in 2006 obtained detailed close-up images of the underside of a large lens-shaped plasma craft, revealing 7 luminous concentric rings surrounding one central HHO plasma beam emitter (above).

This same configuration of 7 luminescent concentric circles reflects the nonlinear quadratic formula [$z_{n+1} = z_n^2$], as precisely encoded in a complex set of ancient Sanskrit stone artifacts from La Maná, Ecuador (pg. 197). The Sanskrit artifacts from La Maná fluoresce under ultraviolet light, as absorbed UV radiation is converted into reemitted light of visible wavelengths – a process that is also generated within the electroluminescent crystalline metals of transdimensional plasma spacecraft.

Complex metal alloys synthesized for acoustic levitation using cosmic infrasound standing waves convert the strong electromagnetic fields of the ships' internal HHO plasma chambers into reemitted visible light from the external surfaces, tuned as desired for various aerial displays or using infrared radiations outside of the visible spectrum of the vast majority of terrestrial humans beings. The enigmatic experiences reported by hundreds of thousands of eyewitnesses of extraterrestrial spacecraft transiting Earth's skies are now explicable in scientific terms, as nonlinear effects of quantum resonance.

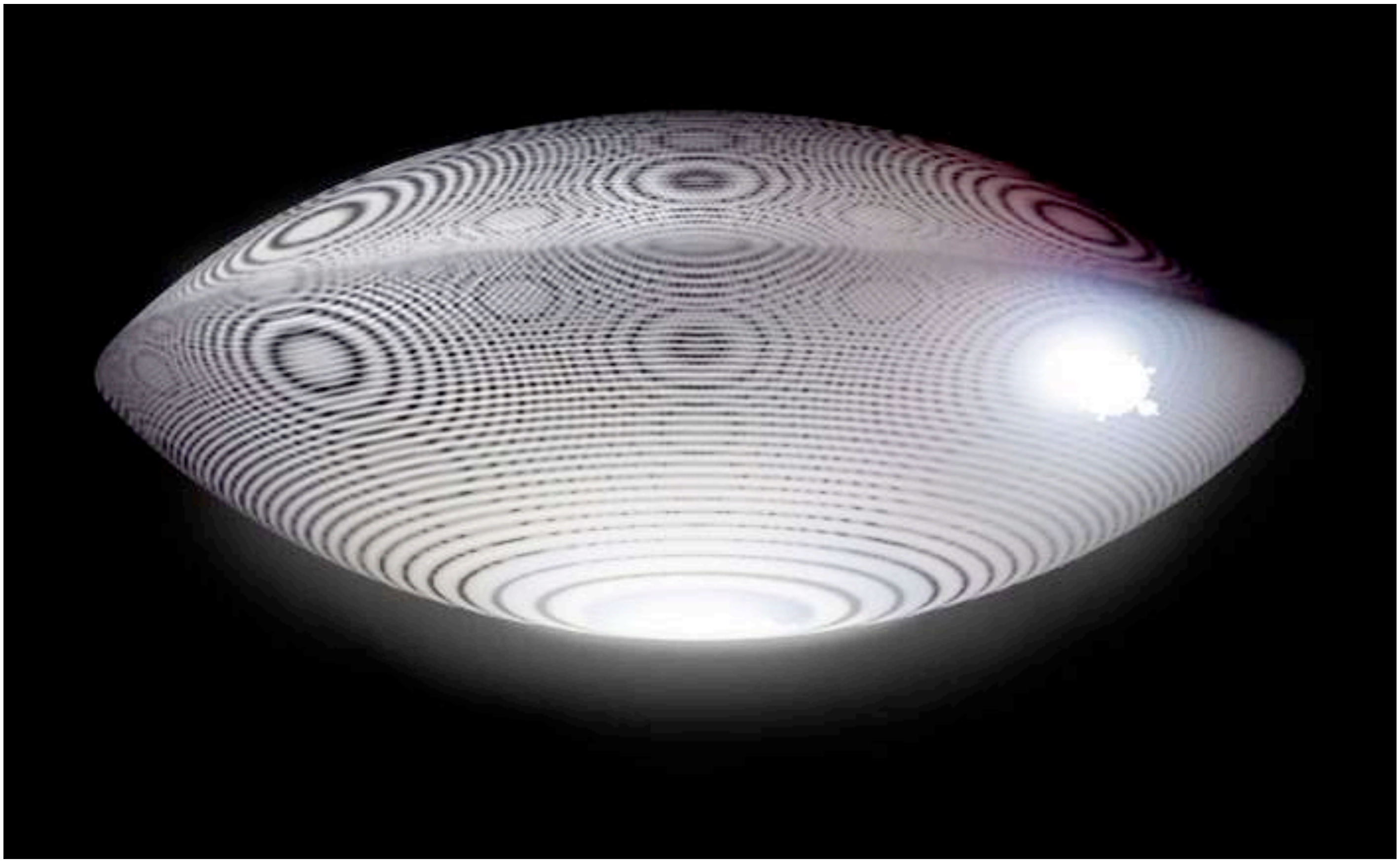

Primary identification of the optimal material properties for antigravity alloys have been experimentally tested and reported by more than one group. The axis-symmetric shape of the object to be levitated, as well as its resonant crystalline material composition, endows antigravitic properties that act to transduce incident radiation into resonant standing waves encircling the object. Fran De Aquino, physicist at Maranhao State University in S. Luis, Brazil, has mathematically modeled antigravitic effects observed in the extra-low frequency acoustic levitation of electroluminescent objects:

> According to the weak form of Einstein's general relativity equivalence principle, the gravitational and inertial masses are equivalent... [To the contrary,] we have shown that the gravitational mass and the inertial mass are correlated by an adimensional factor, which depends on the incident radiation upon the particle. It was shown that only in the absence of electromagnetic radiation this factor becomes equal to 1 and that... it can be reduced, nullified or made negative by means of extra-low frequency (ELF) radiation.
>
> We have studied the possibility to control the gravity on luminescent materials and have concluded that electroluminescent materials with high refractive indices are a new and efficient solution for gravity control technology.[16]

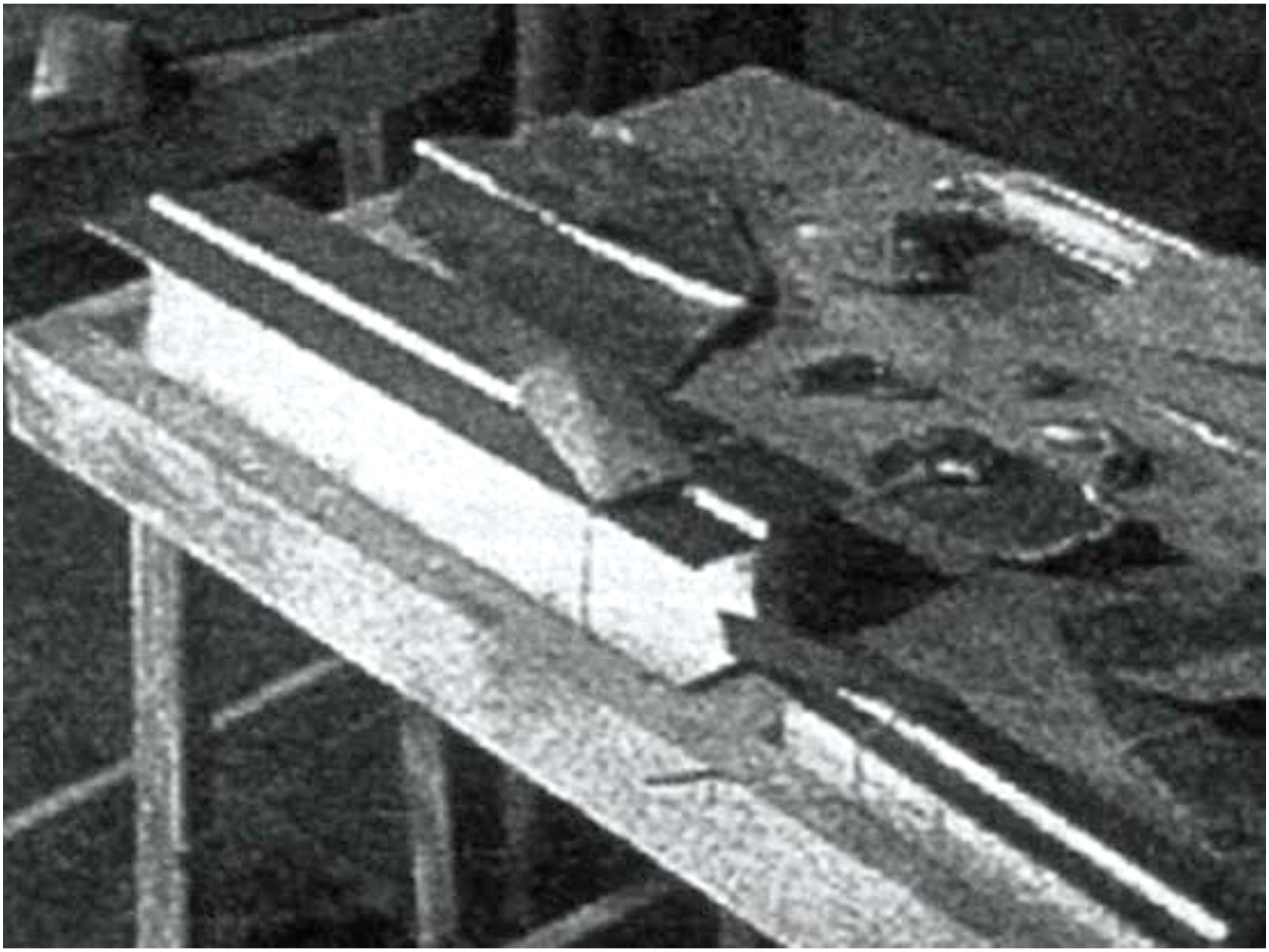

The refraction of acoustic energy within the parallel faces of the alloy’s crystalline lattice allows resonant energy storing as standing waves form within the electrified metals, thereby generating coherent light from its highly reflective surfaces and reducing its mass to become totally weightless. During investigation of photographic evidence produced in the Carlos Diaz case, analyst Jim Dilettoso identified the unusual properties of light emitted by electroluminous alloys as *“coherent light – like a laser”*.[17]

Various metal alloys manufactured for diamagnetic antigravity propulsion systems (having two opposing rotating magnetic fields) produce the effects of electroluminescence from highly reflective surfaces, as well as photo-absorption from completely black surfaces. By applying ancient metallurgical knowledge preserved in Vedic hymns, Dr. C.S.R. Prabhu has reproduced metal alloys with formulations that are completely unknown to modern science, such as one referred to in Sanskrit as Tamogarbha Loha:

> Tamogarbha Loha: Already produced in the laboratory, light in weight, black in color, found to be resistant to acids. Displayed high level of absorption for laser light... used in 'Tamo Yantra'… for the purposes of absorption of light escaping from a photochemical reaction... generating 'darkness'...[18]

The Atlantean Sanskrit origin of vimana spacecraft *seen throughout Earth’s skies today* is revealed by crashed aerial discs collected by military forces in many parts of the world. The most famous incident of this kind occurred on July 2, 1947 –allegedly near Roswell, New Mexico. Leaked US military documentary photographs and films of the New Mexico disc crash wreckage[19] include dwarf humanoid corpses with *large eyes, grey skin and six fingers*, as well as fragments of highly reflective metal alloys imprinted with small six-fingered hand consoles and hieroglyphic Archaic Sanskrit inscriptions. Dr. Schildmann’s epigraphic decipherment reveals the meaning of a fragmentary text from an I-beam:

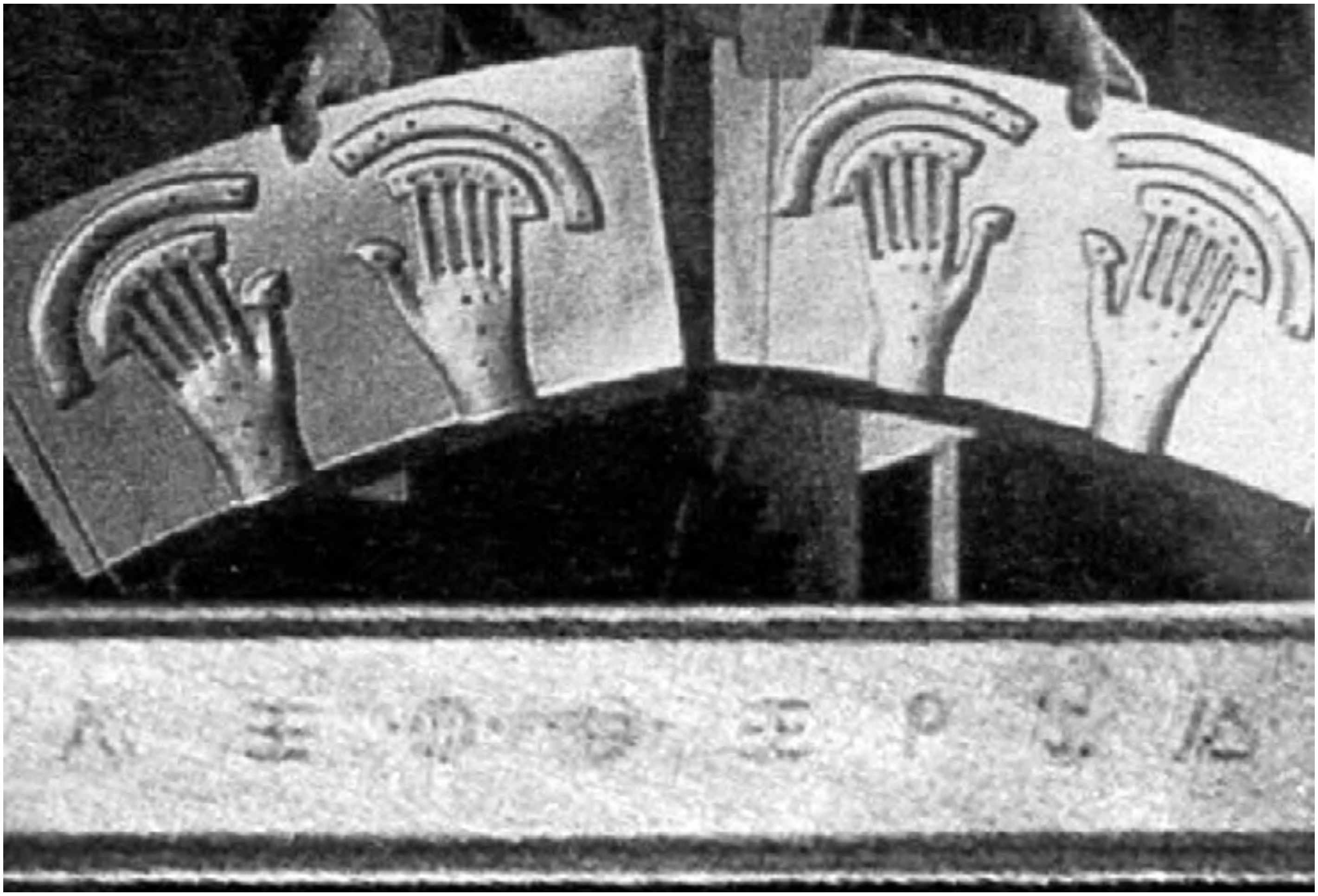

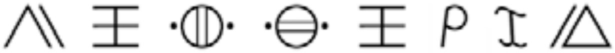

kar-as tridaśa yoni śani tridaśa ku-p^{a}i^{a} as-upama

“workings of 30 resonances, female and male, 30 resonances whereby approaching the highest”

The New Mexico I-beam text directly confirms the Sanskrit origin of the pyramids and ‘UFO’ phenomena. Four very unusual removable metal panels were also documented among the pieces of Roswell crash wreckage, formed with what appear to be inset hand consoles for conjoined control of the plasma ship. One of the panels was damaged. Like portable biocomputers for each crewmember of the ship, each panel is individually tuned to the biorhythmic signature of each pilot, precluding use by others.

The panels’ cast metal forms accommodate two small, six-fingered hands, with nodal sensors distributed at geometric points on the contact consoles and along a radial groove extending just beyond the span of the fingers. The configuration of the consoles exactly match the anatomy of the Roswell autopsy documentary video leaked anonymously to Ray Santilli and published in May of 1995, corroborating their close association and authenticity. While abductees call these beings ‘the grays’, they have been identified as bioengineered androids of a dwarf race from the planet known as Zeta Reticuli, of the Reticulum star system, during informative contacts with a Swiss farmer named Eduard Meier on November 3, 1975, and further clarified in tantalizing detail over 11 years later, on February 28, 1987:

> Their size varies between 126 and 163 centimeters… They originate in the planet and star system Zeta Reticuli,… a median distance of 37 light years away from the Earth. Their spaceships are similar to our beamships, and extremely well equipped, and are very able for space-traveling. The atmosphere of their homeworld is very similar to the Earth atmosphere, thus they can breathe here without great difficulty. They wear close-fitted filter-dress, which offers protection for them in some ways.[20]

The Roswell Case is a reality because the UFO crash actually took place there, but this is denied up to the present day and will continue to be denied in the distant future by the American military and Secret Services, as well as by all government agencies, namely with flimsy, changing, stupid, and ridiculous excuses that will only surely testify to the fact that everything actually happened. With the crashed object, it concerned a beamship from the Reticulum systems. The crew did not consist of natural human life forms but of human androids of a bioorganic nature, which means that they were artificial human androids, which the US-Army also got a hold of, some living and some dead, and these were secretly kept in secret stations [for autopsy, above, opposite[21]].

With the bioorganic androids, it concerned effective artificial life forms, so not robots or the like, but independently thinking beings, in every way self-capable of making decisions, made out of living material and living organs, bred in accordance with the human beings who are located in the Reticulum systems. In Earth-human terms, these bioorganic androids bred by the Reticulum people are actual life forms, and as such, they are respected by the Reticulum people and are also appreciated and treated, for the androids are, indeed, artificial human beings. As such, they stand under the command of the real humans of the Reticulum systems, on whose behalf they also carry out excursions, etc. to foreign star systems and planetary systems. And such artificial human forms were associated with the crash, as I already explained.

The intelligence of these android beings stems from a bioorganic consciousness of an artificial form, which is just as capable of development as the consciousness of the normal, natural person; therefore, it can also be creative and inventive. The type of consciousness is designed uniformly and collectively in each android, which means that all of the androids have a uniform consciousness form, and so, they are collectively like-minded, and thus, a collective form of telepathic communication is owned by them… which doesn't correspond to an actual personality but rather to an artificial form, which can think quite independently, act, and decide, as well as develop itself further, but which isn't arranged into conscious evolution but only functions according to conscious instinct… [Yet], in spite of the collectivity, [such an android] has its own individuality…

The bioorganic androids of the Reticulum people contain an authoritative programming in their artificial consciousness, which prevents that the artificial beings could ever rise over their designers, respectively their creators, but the time of their lives is limited to 300 years of devoted servitude to these. Nevertheless, they are entirely free and can in every way decide as to what responsibilities they undertake, exercise, or decline and, thus, whether they want to focus on a task or not. So they are in no way subject to compulsion by their creators.

In the manner in which they are biologically created, they are incapable of performing sexual acts and also of reproducing themselves. Their… artificial-bioorganic consciousness does not contain a creative spirit form, which would drive and animate this by its own energies. Animation is provided by the artificial brain, which is constructed in such a way that it can absorb and use the cosmic-electromagnetic life energies, so that the existence of life is guaranteed. It is this cosmic energy, by which the artificial life forms are animated and also made alive. All this, however, couldn't be recognized by all the doctors, surgeons, and pathologists, etc. in the American secret stations, where the crashed android beings were taken. But how could they, since their knowledge relating to this wasn't in place yet? In fact, it will still take many centuries and even several millennia before the terrestrial scientists arrive at the possibility of creating such beings as the bioorganic androids. Therefore, these beings will remain an unsolvable mystery to them for a very long time, as it is also the case for them with regard to the crashed beings in Roswell.[22]

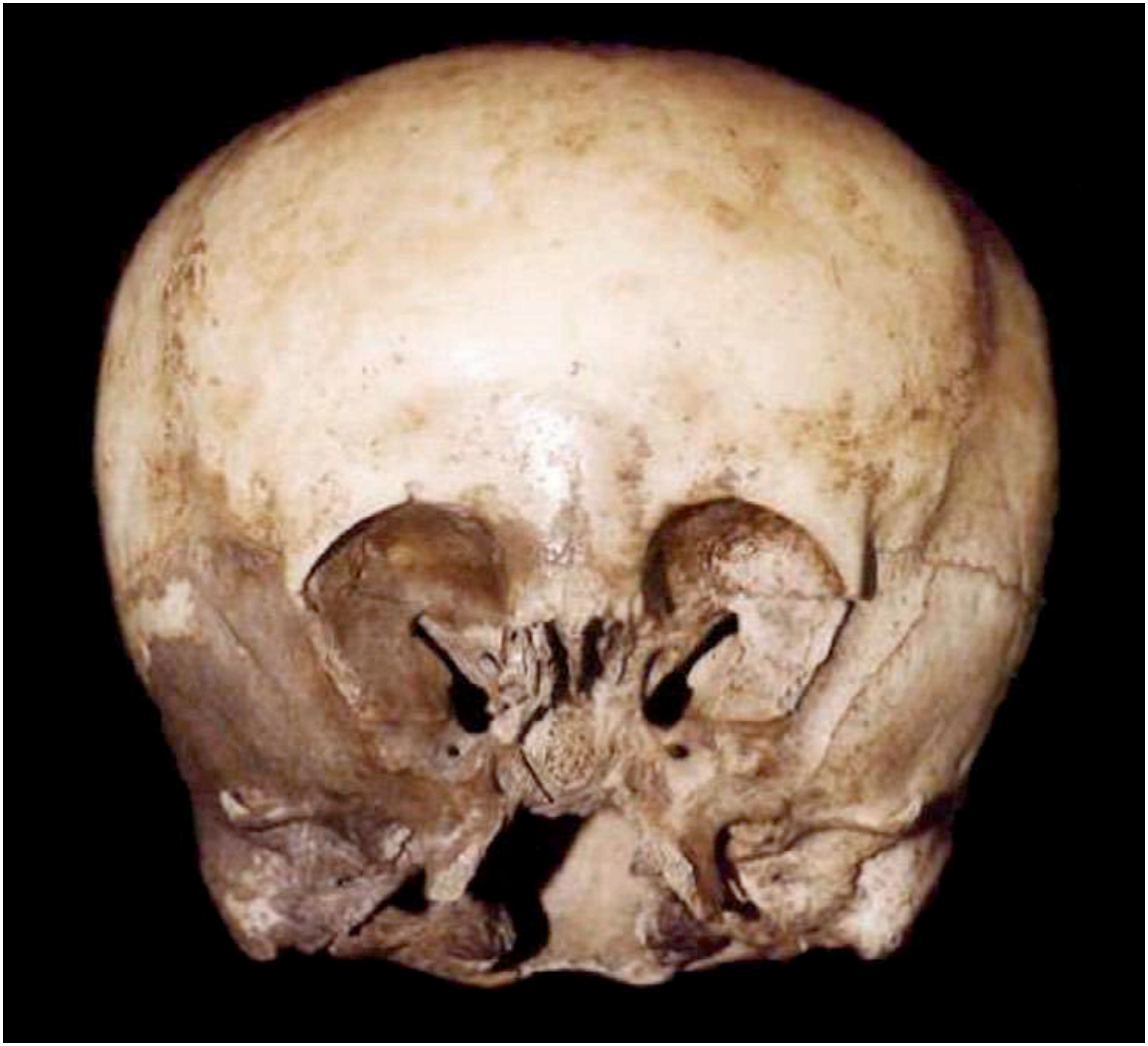

The complex information given by Quetzal concerning the bioorganic androids of the Reticulum people includes specific details of their collective telepathic communication and animation by ‘cosmic electromagnetic life-energies’ that indirectly refer to the resonant fourth-density environment of HHO plasma that is artificially created within their interstellar spacecraft. Brainwave and heartbeat synchronization engendered by resonant acoustic standing waves produced within the plasma ships induces collective telepathic communication, while animation of the artificial consciousness and bodies of the androids is accomplished by the nanopulsed bioelectric currents of HHO plasma.

Further corroborating physical evidence for the fourth-density origin and biorhythmic animation of the Roswell androids has been documented by anthropology investigator Lloyd Pye –a small, unusual humanoid skull discovered 80 years ago in a cave in Mexico. Known as the ‘Starchild’ skull (above, opposite), it displays a set of cranial features that closely resemble those of the Roswell humanoid, with extremely large ocular cavities and an expansive, round brain vault with a very broad, flat occipital region. DNA testing of the Starchild skull has revealed segments of unique genetic material, and bone density analyses show major deviations from average human skulls, confirming the skull’s extraterrestrial origin.

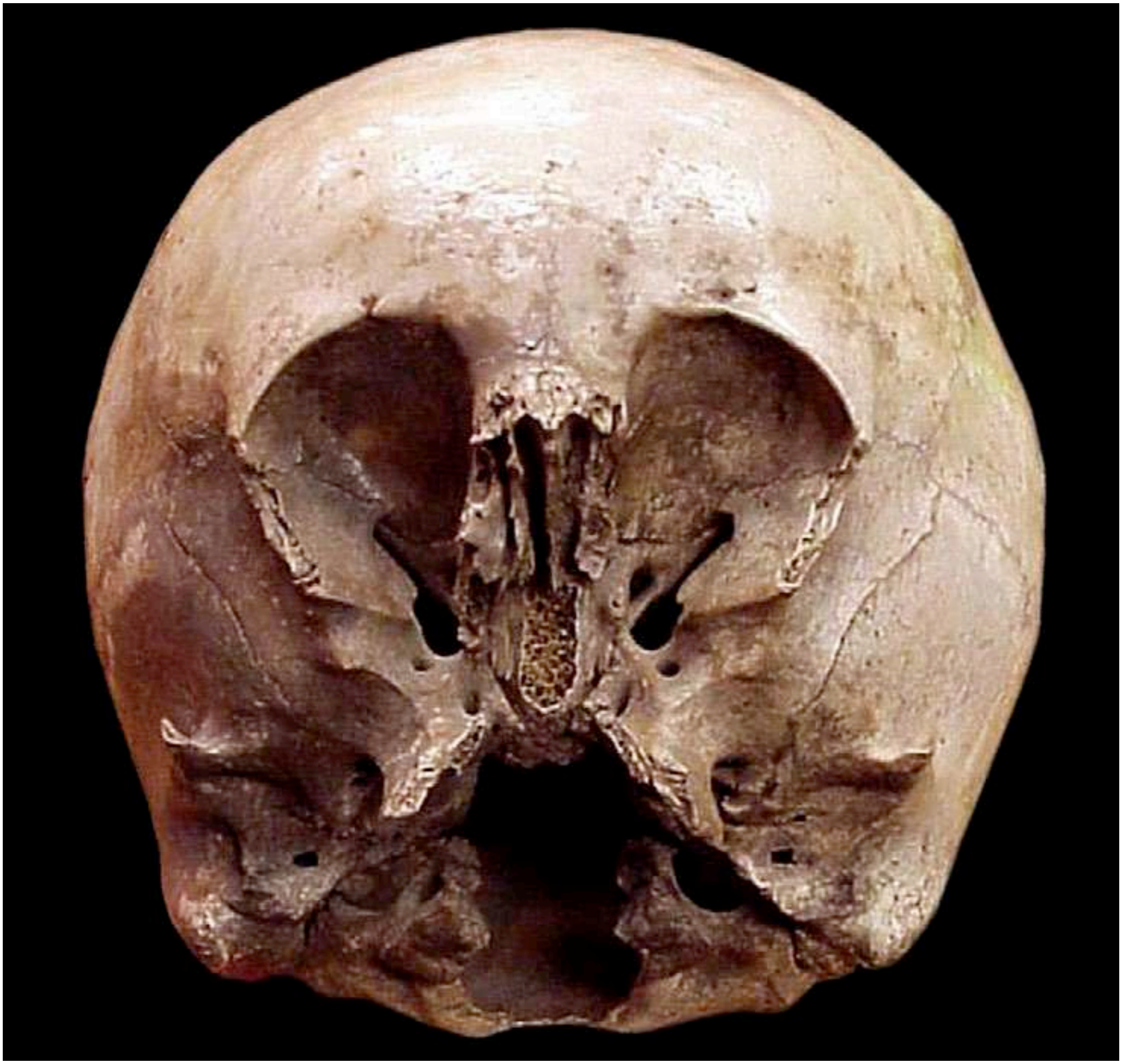

The Starchild skull is thought to be 900 years old, according to radiocarbon dating tests.[23] Further analysis revealed the bone is twice as hard yet half as thick as the skull bones of an Earthly human being, with a density similar to tooth enamel.[24] The unusual braincase presents a volume of 1600 cubic centimeters (cc), allowing for a brain mass that is 200cc greater than the average human brain.[25] *DNA testing has identified unique portions of genetic material that correspond to no known species.*[26]

Microscopy studies of bone samples taken from this unusual skull provide the most significant evidence of advanced artificial bioengineering. Cross-sections of the bone revealed the presence of metallic nanowires running through the many cavities crisscrossing its structure, known as cancellous holes (inset, opposite), that were not cut by the high-speed stainless steel rotary saw. Spectroscopic analysis of the nanowires has yet to be undertaken to determine the elemental composition of the metal.[27] The enhanced electrical conductivity of metal nanowires grown within the artificially modified skull distributes the nanopulsed electrical currents of the HHO plasma atmosphere maintained within the Reticulum spacecraft to sustain the artificial lifeforms as they undertake interplanetary excursions.

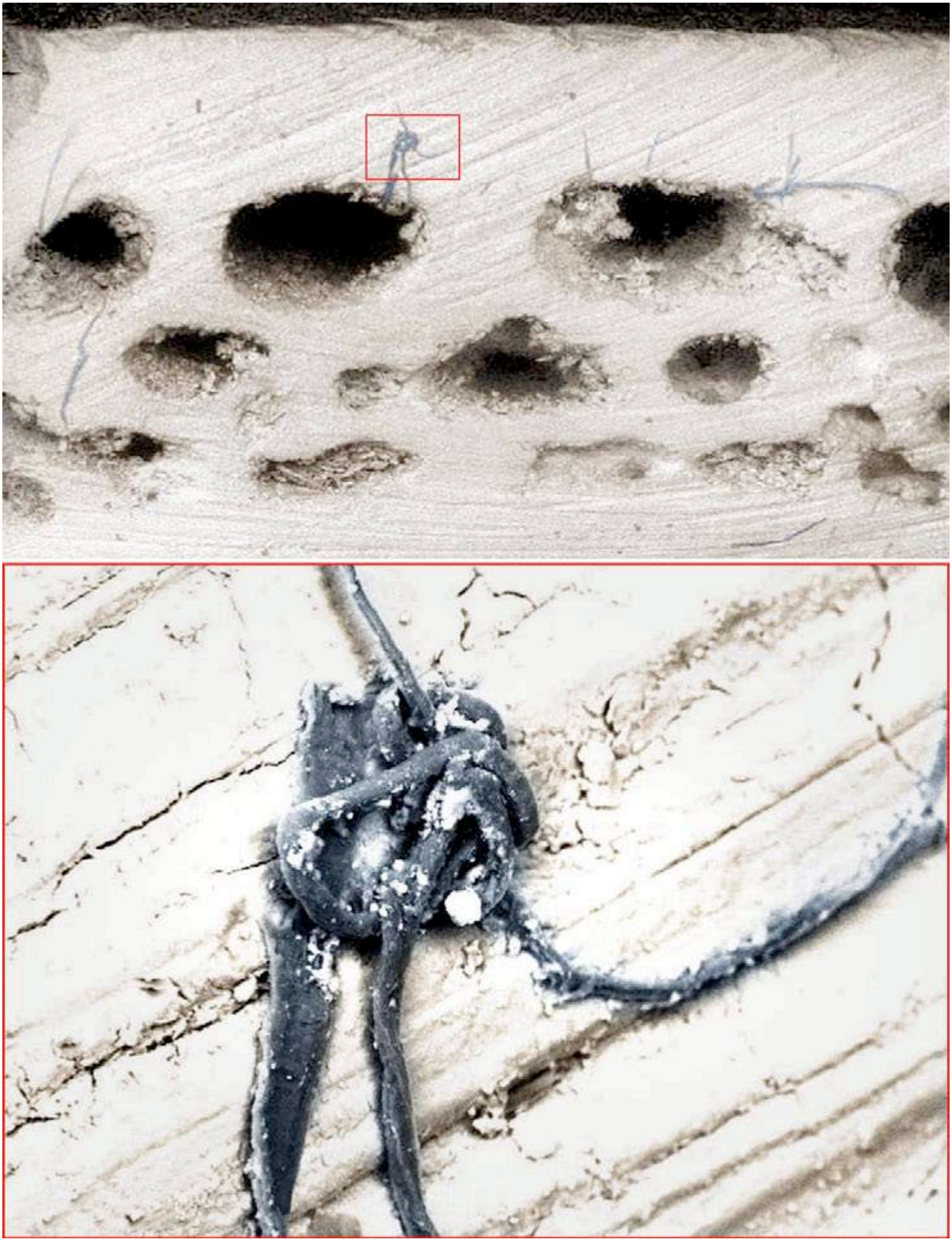

The longstanding presence of technologically advanced extraterrestrial visitors from the Orion, Reticulum and Draco star systems, among others, extends hundreds of thousands of years into the veiled human past. This ancient and still ongoing pattern of activity is recorded in artistic representations from all over the world, and many stunningly accurate depictions of the bioorganic androids of the Reticulum people are to be found among various indigenous cultural traditions. The Wandjina figures of Australian aboriginal traditions comprise a stunning record of interactions with extraterrestrial visitors that spans several thousand years in successive layers on the walls of sacred rock shelters and cavesites.

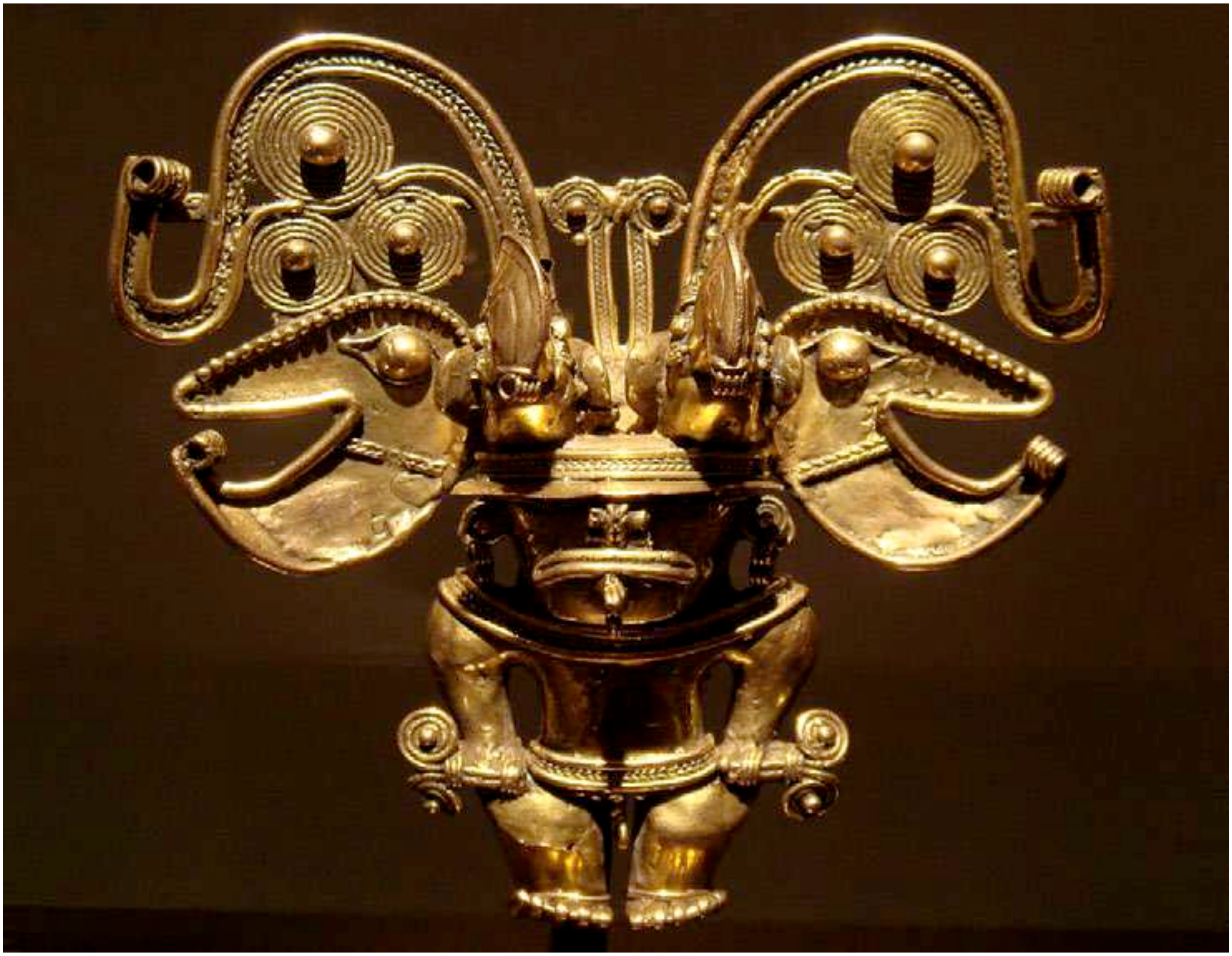

The Wandjina figures are depicted with large almond-shaped eyes, small open mouths and short-limbed, dwarf-like bodies, just as represented by ancient artists from Mesopotamia, as well as among Andean artists on the other side of the Pacific Ocean. A Chancay tri-frequency psychoacoustic whistling vessel displays this same combination of unmistakable features seen in the Reticulum androids, and even includes the accurate anatomical modeling of six fingers on each hand (opposite).

The diminutive Reticulum androids are also represented in Tayrona gold masterworks from Colombia, showing infrasound standing waves converging on the two hemispheres of the brain, inset by 3 spirals representing the tri-frequency resonance (above). The figure holds a double-spiral belt at his waist representing the control of DNA states under the influence of psychoacoustic resonance. These many specific features reflect an accurate and longstanding cultural knowledge of these cosmic visitors, and their connection with highly advanced psychoacoustic technologies of consciousness that were practiced on Earth during Atlantean times, before the Great Deluge that silenced the worldwide infrasound network.

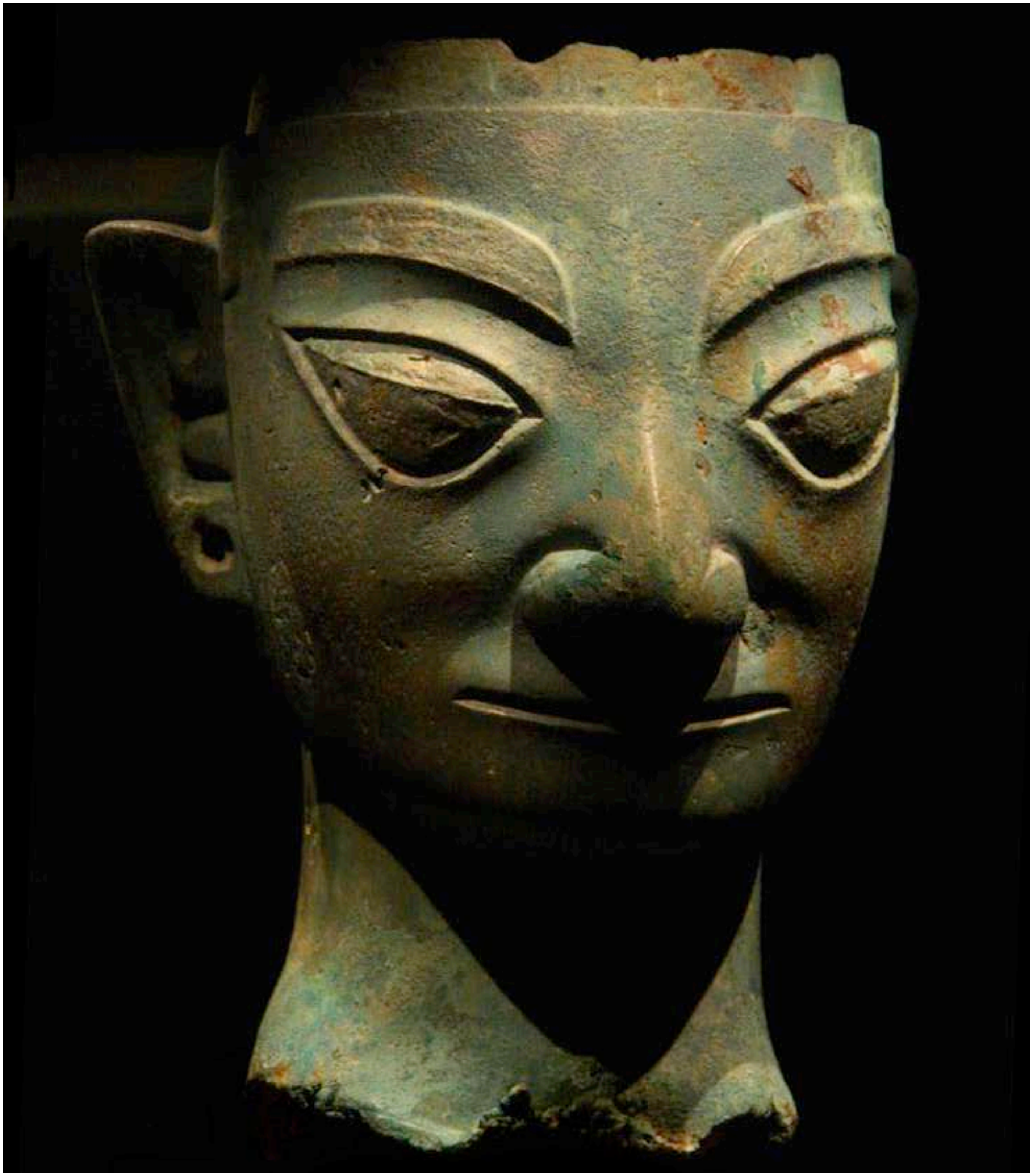

Dozens of antediluvian bronze masterworks from Sanxingdui, China preserve very tall reptilian humanoids associated with the diminutive bioorganic androids from the Reticulum star system, quite likely their designers (above). The physical characteristics of this reptilian race have been described by modern abductees in exactly the same detail and scale as depicted by ancient Chinese artists, associated with populated homeworlds in the Draco constellation (hence ‘Draconian’ –of the dragon). Bioorganic androids are used by the reptoids to pilot vehicles and carry out tasks in the harsh atmospheres of foreign planets.

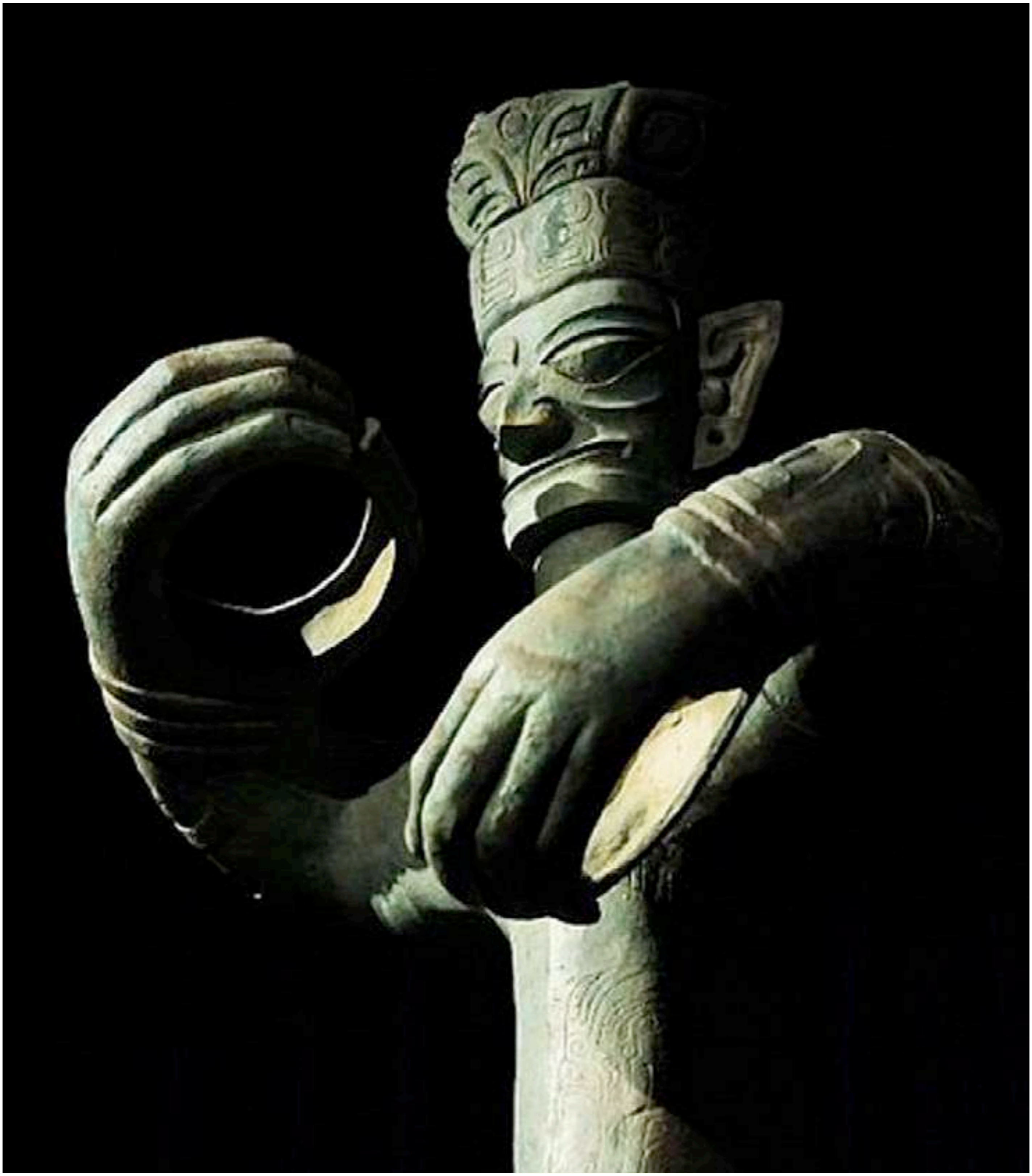

One giant reptilian figure from the Sanxingdui collection wears a headdress surmounted by a circle and a set of standing waves that converge above the crown chakra of the figure (above), as similarly seen on the headdress of the kundalini figure hammered into copper plate from the Crespi collection of Cuenca, Ecuador (p. 10). The complete set of cultural references to infrasound standing waves and kundalini development is shared on all continents, belonging to one ancient mother-culture of Aztlan spanning the entire globe in geomagnetic alignment with the Great Pyramid of Giza, Egypt.

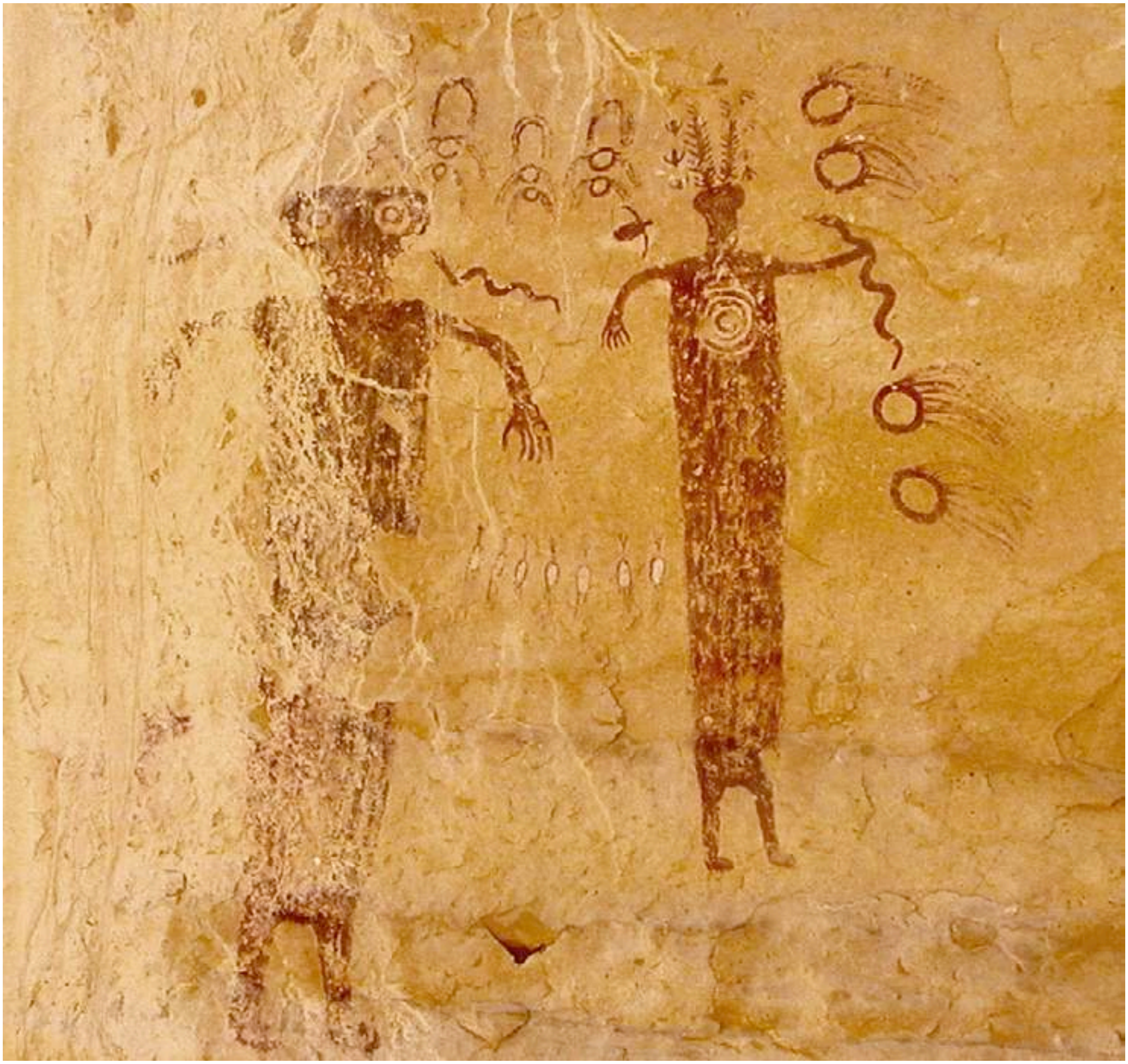

The sheer walls of Barrier Canyon still reverberate with an intense infrasound resonance, and bear many profound Atlantean cultural symbols in the red ochre renderings of several millennia past. Part of a larger wall mural, two very tall humanoid figures are seen levitating into the air with writhing double serpents representing DNA, with concentric circles shown at the heart chakra representing infrasound standing waves tuned to the heartbeat frequency (above). Pairs of concentric circles form the eyes of one of the figures, connected by the arc of a standing wave signifying the hemispheric synchronization achieved by psychoacoustic entrainment. The same psychoacoustic effect is signified by plants sprouting from the head of the other figure like living antennae, with several small flying birds encircling it.

A group of seven luminous white caves are symbolically represented between the two tall floating figures, known as ‘Chicomostoc’ – the 7 sacred caves described in the origin story of Aztlan still retold by the Maya.[28] Luminous plasma ships are depicted as circles flying around the levitating figures, with standing waves radiating from some of the spacecraft (see p. 28), while others appear to be moving in formation through the sky, just as observed and videotaped in our times by thousands of dismayed eyewitnesses.

References

---

[1] -- (2010) ‘‘Rain of Luminous Beams’ Appears in the Sky of Xiamen Huandao Road’ *Southeast Express*, July 12, online · http://www.humanresonance.org/xiamen.html

[2] Moulton Howe L (2009) ‘Eerie ‘Blades of Light’ in Longview, Texas September 9-10’ *Earthfiles*, October 10, online · http://www.humanresonance.org/longview.html

[3] Putney A (2008) ‘Ø’ *Human Resonance Productions*

[4] Lyman W, Kaali S (1996) *‘Biocompatible Electric Current Attenuates HIV Infectivity’* online, · http://educate -youself.org/cn/LymanKaalibiocompatibleHIV1996report18nov06.shtml

[5] Dark T (2001) ‘African Avatars and the third Secret of fatima’ *Nexus Magazine* 8:5, August-September, online · http://www.sons-of-ra.org/writings/African%20Avatars%20and%20the%20Third%20Secret%20of%20Fatima.pdf

[6] Cervenka M, Bednarik M and Konicek P (2003) *‘Nonlinear Standing Waves in Acoustic Resonators with Arbitrary Reflection Coefficients at Their Walls’* Czech Technical University, Prague

[7] Men, Yaxk’in, op cit

[8] Schildmann, K (1999) *‘Als Das Raumschiff ‘Athena’ Die Erde Kippte -Indus-, Burrows-Cave- und Glozel-Texte Entziffert’* CTT-Verlag

[9] Men H (2002) ‘Magnetic Activation of the Pyramids Through the Crystal Skulls’ *Human Resonance*, October 22, online · http://www.humanresonance.org/activation.html

[10] Sitchin Z (1999) *‘The Wars of gods and Men: Book III of the Earth Chronicles’* Harper, online · http://www.bibliotecapleyades.net/sitchin/sitchinbooks03_03.htm

[11] Bernard R (1960) ‘I Found Shaver’s Caves’ *Search Magazine*, March, pp. 26-29

[12] Hesseman M, Zahradnikova N (2001) *‘Ships of Light: The Carlos Diaz Experience’* 2000 Film Productions

[13] Tesla N (1978) *‘Nikola Tesla: Colorado Springs Notes 1899-1900’* Nolit, Beograd, Yugoslavia, pp.61-62

[14] -- (1913) ‘Nikola Tesla’s Plan to Keep "Wireless Thumb" on Ships at Sea, Suggests Transmitters Powerful Enough to Cause the Earth to Vibrate at the Poles - and Equator’ *New York Press, November 9*

[15] Tr. Scott Corrales (2003) ‘Ufologists Pick Up IntraTerrestrial Waves’ *Diario El Tribuno de Salta*, online · http://www.eltribuno.com.ar/salta/20030523_213309.php

[16] De Aquino F (2000) ‘Possibility of Control of the Gravitational Mass by Means of Extra-Low Frequency Radiation’ *Los Alamos National Laboratory*, online · http://arxiv.org/pdf/gr-qc/0005107

[17] Hesseman, op cit

[18] Prabhu CS (2001) ‘Ancient Indian Metallurgy’ *Earthsystem Sciences Awareness and Research Association* Hyderabad, India

[19] Santilli R (1995) *‘Santilli Footage, Debris’* online video · http://www.youtube.com/watch?v=1B8cCNu-FME&feature=related

[20] Stevens WC Ed. (1990) *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 2’* Genesis III Publishing, pp. 33-34

[21] -- (2007) ‘Alien Autopsy’ *Google Video, online* · http://www.google.com/m/search?oe=UTF-8&client=safari&hl=en&aq=f&oq=&aqi=-k0d0t0&fkt=1520&fsdt=9277&q=alien+autopsy+video

[22] Meier E, Tr. Stevens B (2010) ‘Pleiadian/Plejaren Contact Notes Volume 5’ *Billy Meier Translations*, online · http://www.billymeiertranslations.com/#dialog/divVolume5

[23] Pye L (2010) ‘C14 Dating of the Starchild Skull’ *Starchild Project,* online · http://www.starchildproject.com/dating.htm

[24] Pye L (2010) ‘Bone of the Starchild Skull’ *Starchild Project,* online · http://www.starchildproject.com/bone.htm

[25] Pye L (2010) ‘Brain of the Starchild Skull’ *Starchild Project,* online · http://www.starchildproject.com/brain.htm

[26] Pye L (2010) ‘Starchild Skull 2010 Nuclear DNA Recovery’ *Starchild Project,* online · http://www.starchildproject.com/dna.htm

[27] Pye L (2010) ‘Strange Fibers Inside the Bone of the Starchild Skull’ *Starchild Project,* online · http://www.starchildproject.com/fibers.htm

[28] Ghostwolf R (2007) ‘Ancestors of Our First Americans’ *Robert Ghostwolf*, online · http://www.robertghostwolf.com/Aztlan/outheretv.htm

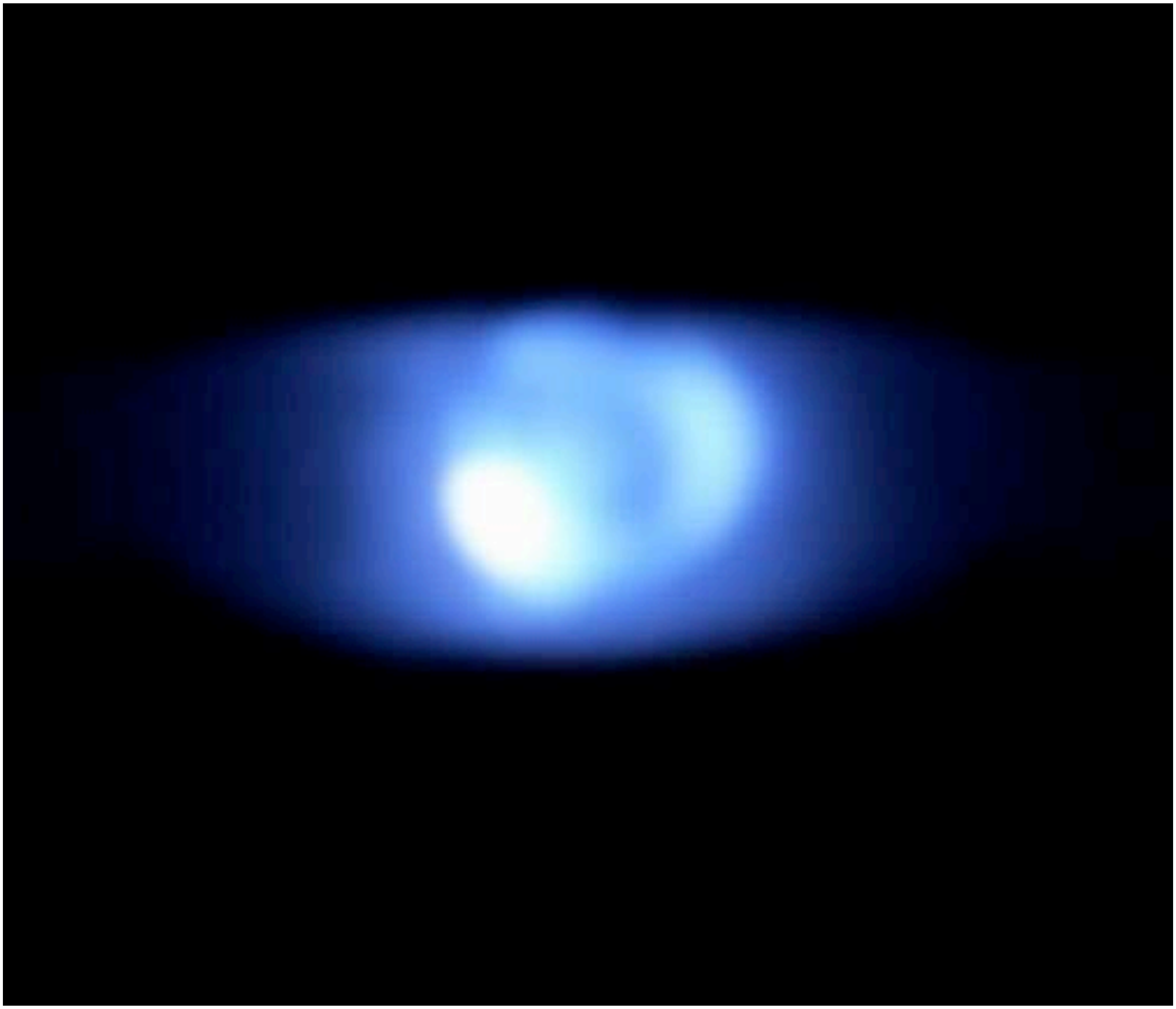

2

Contacts from Koldas

Contacts from Koldas

A remarkable set of still ongoing extraterrestrial contacts that began in 1968, in South Africa, present a fascinating perspective on the unfolding human experience. A young radio/electrical engineer living in Richard's Bay, South Africa (above), 'F. Edwin W.' (alias) received radio transmissions from a cosmonaut in Earth orbit named Valdar, from the planet Koldas. Subsequent years of regular radio contacts and comprehensive face-to-face conversations included other wise cosmic visitors from a Confederation of Planets in alliance with Koldas. Extensive and profound information shared during the visits of Valdar, Kashendo, Taylanz and Wy-Ora were recorded and later published by Carl van Vlierden in 1986.

Edwin was told that the primary structure of the Confederation of 12 Planets was peacefully founded in the Grandor star system, comprised of planets Grandor, Leec Goran, Sparlane, Velaan and Novan; and has extended to include the planets of neighboring systems like Koldas, Salamia, Siton, Bryonne, Pyrole, Pyrc Furvey and Triaxula. Valdar provided Edwin with a basic description of his homeworld, Koldas – a giant world with freshwater oceans reflecting the purity of planetary systems at high resonance:

> 'It is many times larger than Earth. We have seas on Koldas, but these are freshwater seas. There are also large lakes, the largest about 24,000 km long and 20,000 km wide. Some of the towns of Koldas are in the mountain regions while others are by the sea. The towns are similar to yours, except that buildings are constructed differently. There are no roads here for our vehicles travel above the surface of the ground. Our buildings are linked by conveyors to take people to their destination. Our capital city is vast, much larger than New York. Of course there are rural areas with smaller homes for one or two families, and there are forests and also rivers. In the large lakes there are islands where you find holiday resorts.

> ‘Our way of life is very similar to that of Earth. What I am trying to impress upon you is that we do not differ from you. The only difference is our mental attitude. Our way of life is different. We have no monetary system; no currency whatsoever. Each planet in the Confederation produces and plays a vital part in the whole. Some are rich in minerals and supply these to the others and some are more suitable for cultivation. Koldas, with its seas or lakes, has an abundance of fish, harvests of which are sent to other planets in the Confederation in exchange for minerals and materials to build the powerful Astrael-craft. And so my friends, a chain of good-will and understanding links all our planets. Earth too could play a part in this Confederation… We are like you. We are in the flesh. We feel pain and we die. The difference is that we are able to travel at this fantastic speed.’ [1]

The visitors from Koldas spoke in words of genuine compassion during their many radio contacts with Edwin (below). Valdar expressed deep sorrow concerning the intense suffering of children of war:

> Valdar then reiterated that most of Earth's governments, political rulers and Superiors, knew of the Koldasian operations and the Confederation. He said they had known of the presence of Confederation craft over their countries and in the oceans for a very long time… ‘What makes me sad,’ Valdar said, ‘is the unnecessary bloodshed and warfare on Earth. I often think of the little children who are the innocent victims of all this violence. Infants without homes having to find refuge, food and water, yet they have no idea why their parents are at war. There are areas of dire poverty. All these scenes and much more were recently recorded by the three-day aerial survey. This too contributes to the Confederation's decision to withdraw their patrols.’ [2]

The observation of humanity and monitoring of the rapid development of Earthly civilization is a primary task of Confederation vessels, including the older Astraelcraft and newer Corynthian spacecraft, which incorporates their most advanced designs with automatic instrumentation, representing a most formidable interstellar craft, in terms of both annihilating power and instant translocating capacity:

> ‘The craft used by space-travelers range from two-seater scout-craft carried by their motherships, to the largest masterpieces of design and technology which are veritable floating cities in space. The Confederation spacecraft were spherical in appearance prior to 1941. It was these craft which were used to explore the Universe and solar systems many thousands of years ago. Then the spherical craft were superceded by the Astraelcraft. These were named by the Koldasian designer in honor of his wife Astraelda. Other smaller craft are disc-shaped. Larger carriers are cigar-shaped and measure in kilometers. These are used for cargo and passengers.

> ‘All these craft are propelled by magnetic flux. The magnetic streams have different layers with various velocities. The motors, with which all craft are fitted, serve as rudders directing the craft into the various velocity layers. The highest velocity is reached on the outer surface and can exceed that of light. The magnetic streams criss-cross deep space, connecting planets, solar systems, galaxies and even universes in a vast web of highways and by-ways for space-travelers and their craft…’ [3]
>
> Koldas had enjoyed 20 centuries of history without fighting of any kind and consequently the only deterrent they could muster was their fleet of Astrael spacecraft. These craft are not intended for military service but are equipped with a ‘Sun-ray’ disintegrating laser or charged particle ray gun which is designed to annihilate meteorites in the magnetic traffic lanes. Its pencil-like ‘laser’ beam stabs at any object, destroying it in a brilliant flash without leaving a trace...[4]

> I scrambled up against the side of the dune where I had a good view of the spacecraft as it touched down on the beach. As it came closer, I could see how enormous it was. It must have been close to 50 meters in diameter… I heard a low hum… It then rose up into the air… The ship now moved out over the surf, rising all the time… The craft itself was not emitting any light… The outside was steel grey in color but shiny. One could not see any rivets or joints or welds; it seemed to be one complete moulding, like two saucers put rim to rim.[5]

The technological advancement of Koldasian spacecraft reflects their comprehensive understanding of the nature of the human spirit. The physical limitations of corporeal existence have been mitigated by the artificial control of matter, space and time – freeing the human being to explore transdimensionality. Valdar’s brief description of the spiritual views prevalent among the Confederation planets is highly reminiscent of the ancient Sanskrit traditions passed down through the ages from their original dissemination on Earth during colonization events described to have occurred over 70,000 years ago:

> And he told me sometimes about the religion of the Confederation… He stressed that he did not give such information with the idea of trying to introduce it to Earth. He did say that many of our ancient civilizations knew of their religion, which was taken to our planet by members of the Confederation approximately 70,000 years ago. They believe that there is One Supreme Being, usually referred to as the Divine One. This Being is infinite and all-powerful and He creates all things, seen and unseen. Also created by Him was his Son. When man was created, the Divine One placed within the body of man a spirit or the soul, which never dies but continues on to eternity. The task of man's soul is to rule the material body by overcoming material desires... When this is ac-

> complished, the soul will be called back to the great source and forever live in perfect joy and happiness. They also believe that all mankind is created by the same Divine Force, therefore all mankind are brothers and sisters. But this is a mere outline of their religion; there is much more – for example, their belief in reincarnation.[6]

The ancient idea of reincarnation present among most cultures of the world has only recently come under scientific scrutiny. Hypnotic access to memories of past lives stored in the subconscious mind enables the statistical comparison of specific psychological and physiological traits of successive incarnations of a spirit, whereby solid probability calculations can be made concerning the likelihood of such a spiritual continuity. Well-documented reincarnation cases such as that of Jeffrey Keene and John B. Gordon (left) present so many extremely rare synchronicities – *elusive 'proofs'.*

In his discussions with Edwin, Kashendo described part of the Confederation's master registry as a vast reincarnation database, which cross-references the specific characteristics of every sentient being within the planetary collective as a statistical aid in identifying the successive incarnations of a particular spirit:

> 'Now as I promised,' said Kashendo, 'I shall tell you what we know of your past incarnations in our own Confederation… It is a vast subject and there is much to tell. Let me begin by saying that the master register is stored in a gigantic memory bank on the mother planet of Grandor. Here are the records of all beings who have lived lives in the Confederation.'
>
> Kashendo explained that the giant computer is housed in 37 buildings which are arranged in a circle around a central building. Most of the staff work in the central building which collects the required data from 'memory cells', consisting of thousands of transparent tubes with crystals inside them. Each tube is capable of storing many thousands of symbols and a memory bank consists of many thousands of these transparent tubes. The data is stored indefinitely and when needed the controllers extract it in coded form which is then decoded to provide the required information. All the equipment is very complex and kept in a special atmosphere.
>
> The memory cells store innumerable personal features and characteristics which reappear in the same individual through many lives. When a search is made, a special computer transmits the details of the required person to the whole range of memory cells. As soon as the features are recognized by a memory bank a message is transmitted back to the central control building where it is recorded. The whole process is complicated and Kashendo said that he found it difficult to describe it accurately, but he hoped it explained how the Grandorians are able to trace people's lives through time, at least within the Confederation.' [7]

The scientific tracking of a particular spirit through its long string of incarnations is a statistical task that requires great computational capabilities, which the Grandorian record-keepers achieve using supercomputers operating in 'a special atmosphere' of fourth-density HHO plasma. The geometric configuration of 37 buildings housing the thousands of memory cells is precisely aligned in a circle around a central building to make use of infrasound standing waves for the focused distribution of wireless power by which the HHO plasma chambers are resonantly synchronized. Edwin's spirit had selected a reincarnation on Earth as a means of transmitting information for the benefit of humanity. One of Kashendo's profound offerings of knowledge concerning the human past and the deep history of Earth revealed the presence of once breathtaking cities laid to waste on the ocean floor, along with the giant wreckage of huge interstellar spacecraft which enabled the colonization of Earth by different races seeking refuge from various worlds in peril (rather than being a byproduct of evolutionary processes):

Kashendo spoke about the various races we have on Earth and said, ‘This is an intriguing subject but some of the mystery is known to the Confederation. We know that long before we first visited your planet, there were other visitors. When our ancient pioneers arrived there, they found evidence that whoever came before them must have come from systems far beyond those known to us. It is possible that they wished to colonize Earth. We do not know for certain whether this explains the many different races, but we do know that the race which you call the Red Indian, had ancestors who came from a far away planet doomed to destruction. They were highly civilized and had mastered space-travel. They left their home planet in a giant craft to search for new planets to colonize and reached Earth and there they settled. It seems that they were abandoned by their people or were unable to return home, possibly for lack of fuel. Their craft was stranded and so they remained on Earth. They were forced to improvise and learn to live off the land. They made use of what they had and what they could find. They hunted the animals; and learned to use them for food and sustenance.

‘We know about these people, for they were very intelligent as a race, and one of the early ones to occupy Earth. But what about the many other races you have? We can only assume that they too were space travelers for we have definitely found the remains of many craft beneath your desert sands. Gigantic craft [have been found] which are not from the Confederation! Should these remains be discovered, they would cause a sensation on Earth as they are of enormous proportions. They must have carried thousands of people.

‘It is possible,’ he suggested, ‘that they never intended to return to their homes. Were they banished and perhaps sent away as punishment? We can only guess at the reason for their presence but the evidence of their arrival from other planets is there, beneath the sands. This may be one reason why there are so many different races on Earth…

‘When the Confederation carried out patrols beneath the surface of your oceans, they discovered many craft buried in the mud of the ocean floors. There is much evidence which will not be found soon, and much that has been lost. In another transmission I must tell you about the abandoned cities which lie beneath your Pacific Ocean; mysterious cities which must have been of great beauty before they disappeared. What catastrophe befell Earth? Who were the builders of these matchless cities?’ [8]

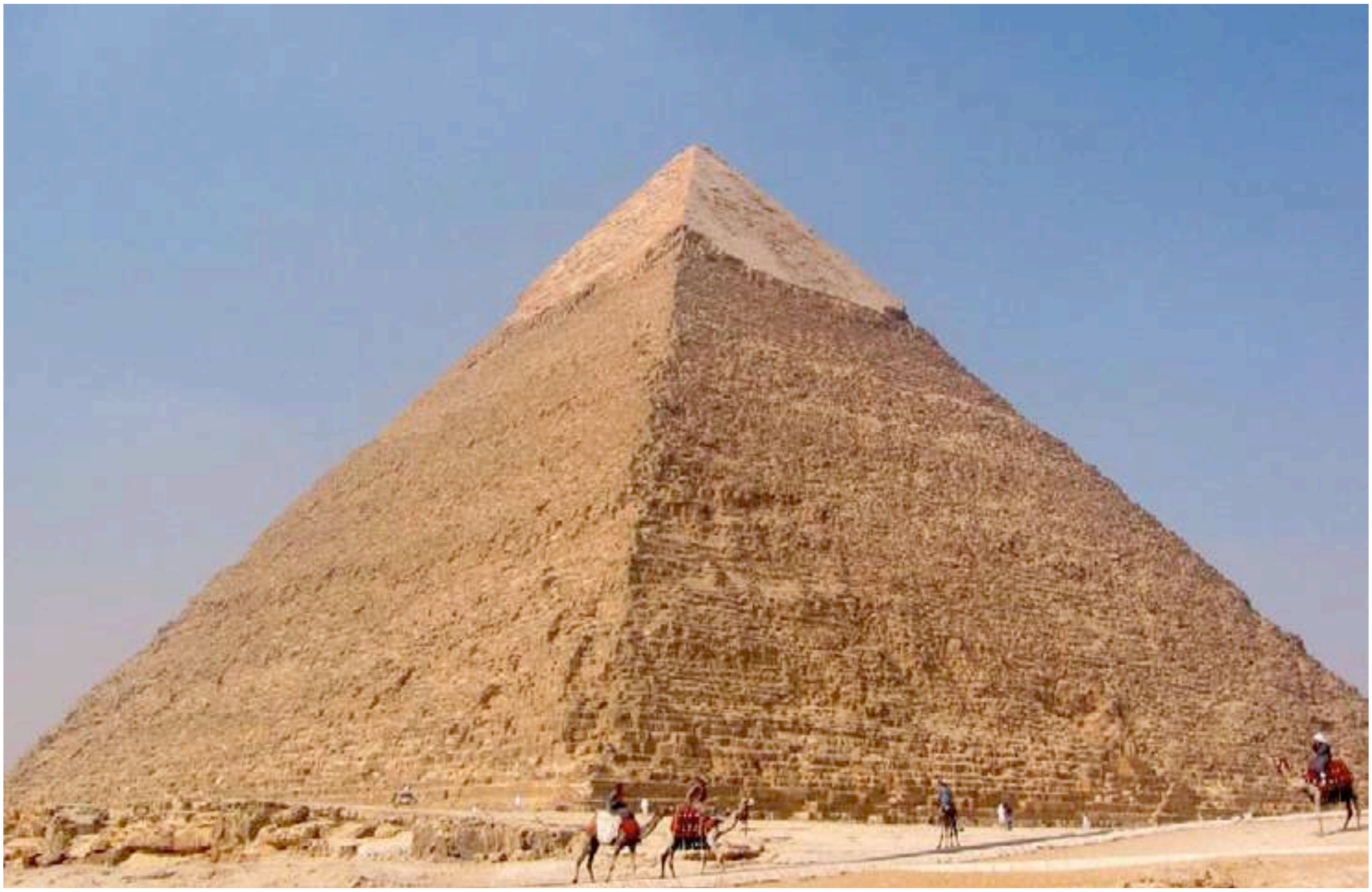

These remarkable facts delivered by Kashendo attest to the great antiquity of human civilization on Earth, and the deep mystery of the cosmic origin of human racial diversity on Earth. While the existence of the remains of gigantic spacecraft below the (Sahara?) desert sands has been quite purposefully withheld from public knowledge by the world's governments, the same astounding scale is plainly witnessed in the pyramids of Giza, Egypt – designed to transduce and focus infrasound standing waves used for the propulsion of antigravitic plasma ships by acoustic levitation. Valdar explained:

> The pyramids… were used as transmitting and receiving stations and for the storage of magno-solar energy. They were located in a strategic area and the ancient Starships used to home in on them… Their design is significant in more ways than one. In the coming changes on your planet, this pyramidal design will again be used in your buildings.' [9]

Other discussions with the Koldasian envoys confirm that the 'pyramids [of Giza] were strategically placed in the geographic center of the land surface of the planet.' [10] This fact was recently reported by investigators of the Great Flood, who analyzed sea level data to identify an ancient global flooding event that raised sea levels by 178m, which offset the gravitic center of Earth's landmass by hundreds of miles to the northeast of Giza.[11] *This 178m sea level rise explains how the casing stones of the central pyramid have remained intact above that precise height,* while those casing stones below that mark were gradually torn away by the erosive wave activity of slowly receding floodwaters (above).

Long before the destruction of the pyramid resonance network by the Great Flood, the Giza pyramids pulsed with a tri-thalamic frequency of 1.45 Hz to synchronize heart/mind rhythms throughout all of the sacred sites of the planet and act as a homing beacon for the legions of spacecraft in transit through our solar system. The resonant excitation of planetary infrasound taps a vast reservoir of cosmic energy available to humanity once again through the oncoming changes, which include an intensification of planetary resonance. The deep hum of our planet, an infrasonic gravity field, is artificially recreated in all of the space-faring plasma ships' magnetic motors, as described to Edwin during a transmission:

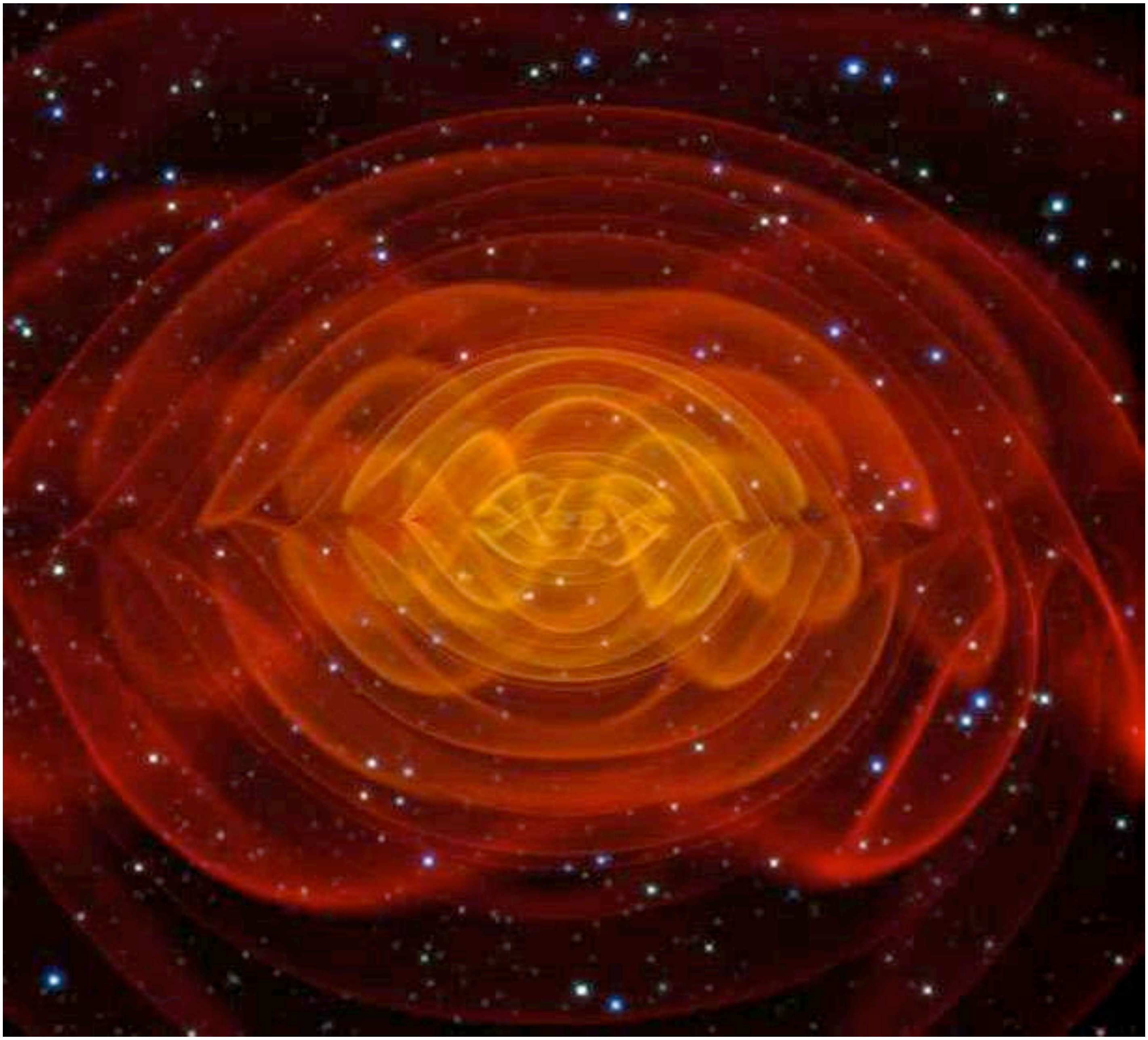

> Wy-Ora continued wistfully, ‘If only your governments would unite in peace and extend the invitation, your skies would ring with the hum of magnetic motors… All our knowledge would be yours. All that we have achieved over the centuries would be given you freely. But unfortunately, your rulers cannot agree. They tell us many things; why they cannot accept us. But it all boils down to the fact that it would bring too many changes on Earth and that it would disturb the 'balance of things'.’ [12]

In fact, all extraterrestrial matters have been kept secret by Nazi/CIA elements to preserve the present corruption of world governments, enforced by the military use of stealth technology stolen from Nikola Tesla in 1942.[13] The standing waves discovered by Tesla in 1899[14] display the same Fibonacci structure observed of galactic and solar systems (above), providing magnetic field lanes for magnetic propulsion:

> ‘Each solar system has a magnetic field. Once you have broken away from the gravitational field of your planet, you can make use of the magnetic fields of other planets and solar systems to travel vast distances and even change in time… On Earth you are using the wrong method of propulsion. You are using fuels, liquid or solid or whatever they may be. This does not take advantage of the vast reservoir of universal energy. There is an unlimited supply of this energy and power that can never be exhausted. This energy is there for mankind to use on your planet.’

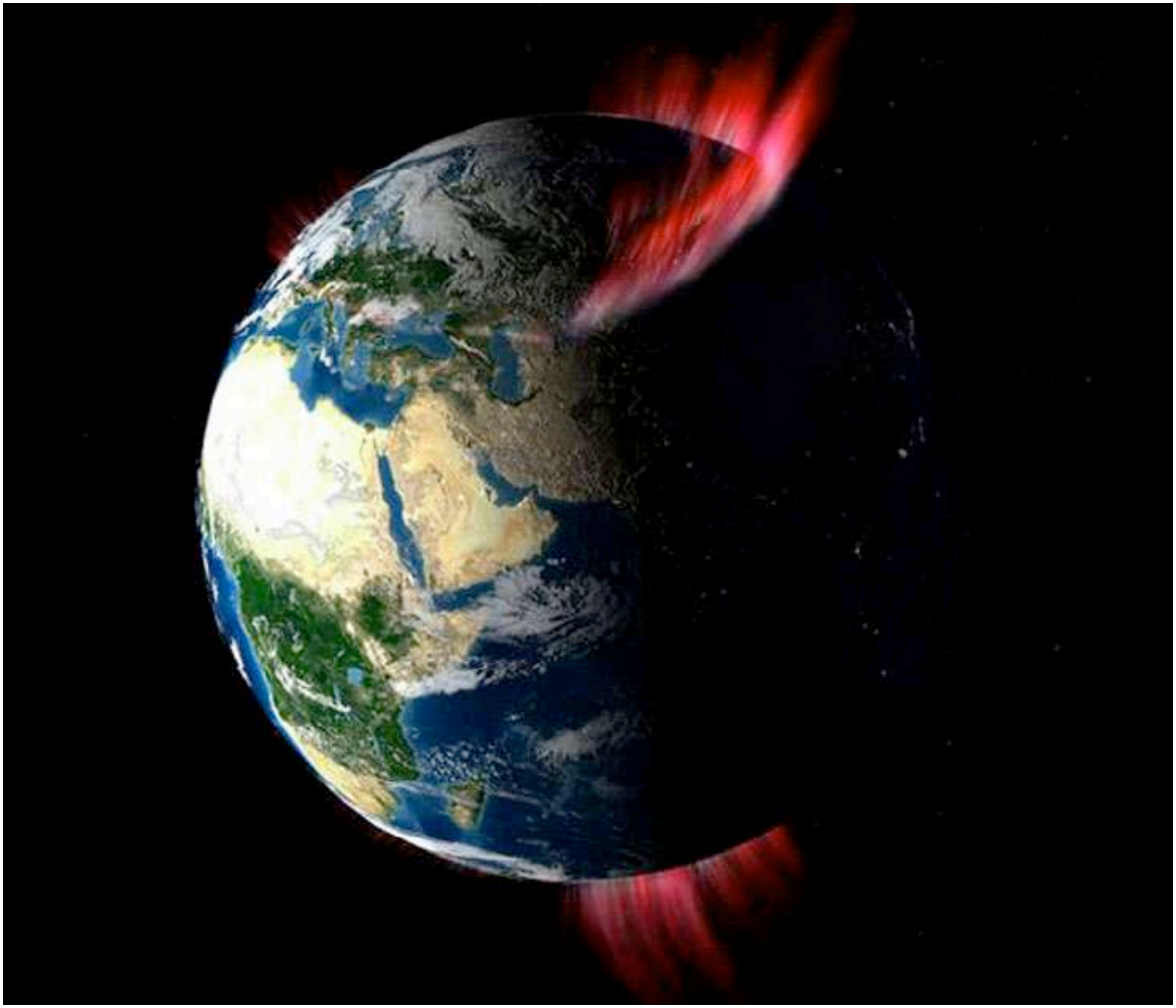

> ‘Let me explain further. The Astrael-craft I am in has eight magnetic motors; four forward and four reverse for control. These are actually giant electromagnets powered with solar energy. Now, beyond your planet there is a magnetic force-field which runs in streams between planets and solar systems, and even between galaxies and universes. Magnetic fields seem to run on forever. Once we are in a magnetic field we travel in this invisible stream through space. Our magnetic motors allow us to check and control our speed. Without them we could not select our destination. We would be helpless as a rudderless ship on a fast flowing river.’
>
> ‘Once a certain velocity is reached, things begin to happen. For a little while, breathing becomes difficult and the craft is controlled by a master computer. As you arrive in other universes (dimensions) awareness returns. Craft, computer, personnel and all which have existed in the energy-state flowing along the magnetic paths, now reintegrate with their polarity changed, to become solid matter again. And that my friends, is all there is to it. It is the magnetic fields which carry us at these incredible speeds, an changing time is a side effect of it. Your craft on earth could do this if you had the magnetic motors!’ [15]

The technical details of the magnetic motors were first discovered by Nikola Tesla during his wireless researches, and were later perfected in secret designs for antigravitic spacecraft using infrasound standing waves and HHO plasma, derived from water. The implosion of atomic hydrogen forms the unique plasma that protects all interstellar spacecraft. The presence of galactic magnetic fields, vast cosmic highways described by Valdar in 1970, were only recently identified by Earth’s scientists in 2009:

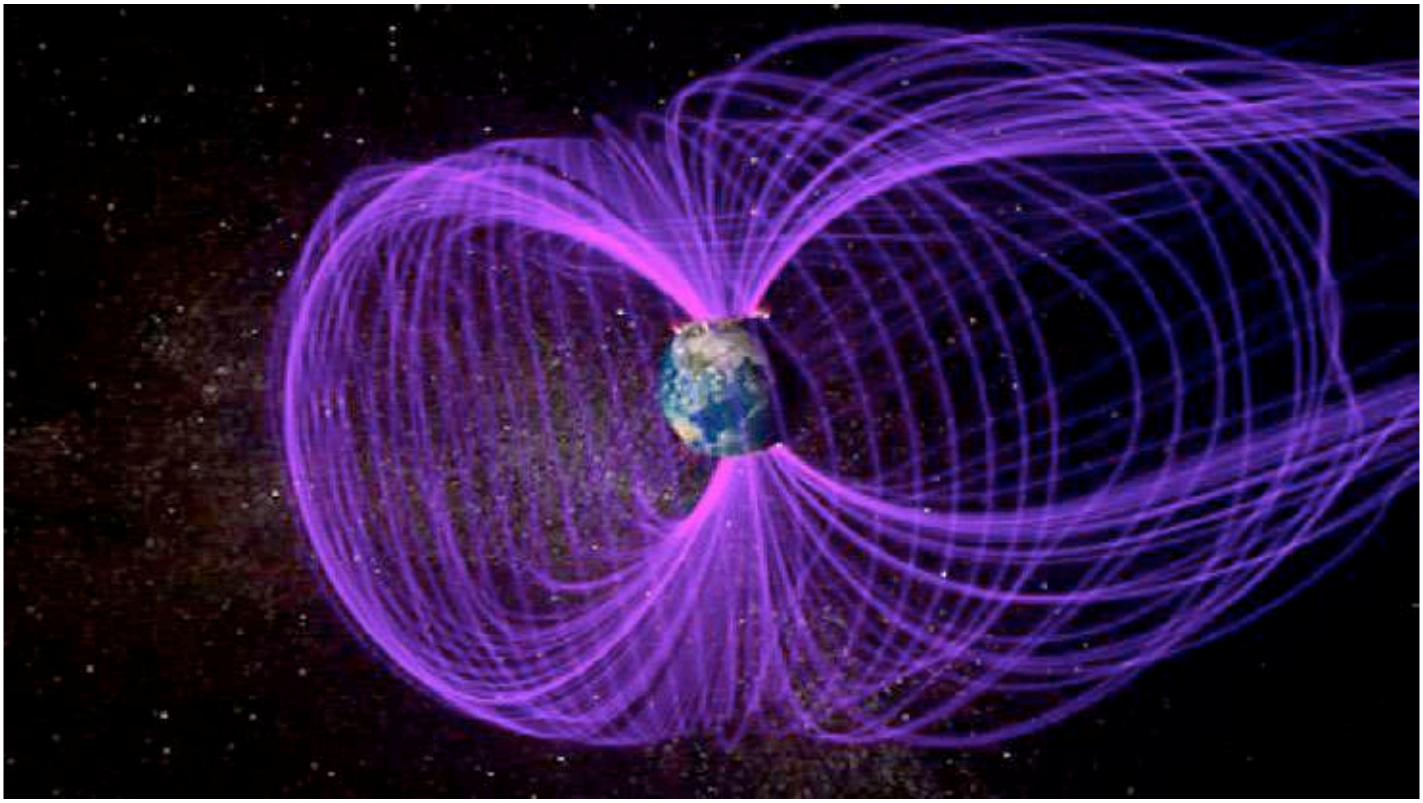

> Last year, NASA's Interstellar Boundary Explorer (IBEX) made a discovery so profound that scientists had a hard time describing what they were seeing. A vast ribbon located in the outermost reaches of the solar system had been spotted, a structure that had never been seen before. Now scientists believe the shape might be created by a huge reflection caused by particles bouncing off a galactic magnetic field… In December it was announced that the veteran Voyager probes currently flying toward the heliosphere measured an unexpected increase in magnetic field strength. Scientists believe that this increase is due to an unexpectedly large magnetic field threaded through a cloud of interstellar gas called the ‘Local Fluff.’ Our solar system is currently speeding into this Fluff… [which would more accurately be called ‘plasma’.]

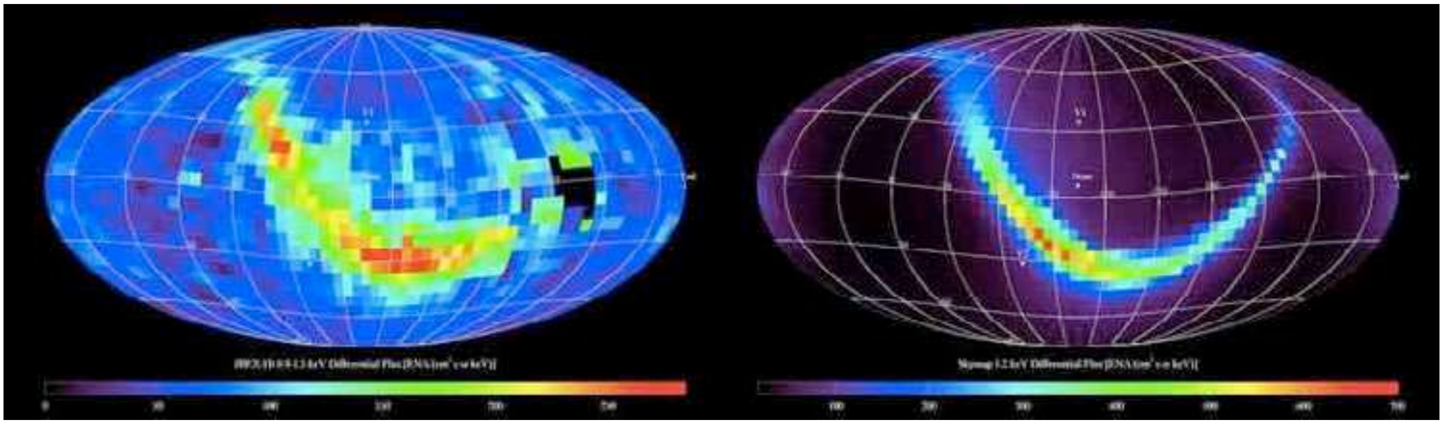

> Jacob Heerikhuisen, a NASA Heliophysics Guest Investigator from the University of Alabama in Huntsville, believes the ribbon observed by IBEX is caused by an interstellar magnetic field… ‘We believe the ribbon is a reflection,’ said Heerikhuisen. ‘It is where solar wind particles heading out into interstellar space are reflected back into the solar system by a galactic magnetic field.’ [16]

These newly identified galactic magnetic fields connect star systems within the spiraling arms of galactic wheels and act as propulsion lanes for interstellar spacecraft. The advancement of human consciousness is required for the penetration of the light speed barrier, for which purpose the Koldasian visitors influence contactees like Edwin by a voluntary conditioning of the subconscious mind during deep sleep:

We asked Valdar about ‘conditioning’ during a thought-transmission. He said that initially we were tested to see whether we had accepted the Confederation ‘way of life’… ‘*Only then*’ he continued, ‘various methods of conditioning are used. This is done mainly during thought-transmissions. We describe our planets to you – our way of life – and we answer your questions. Your subconscious minds are conditioned by us to store certain impressions gained during these transmissions. These are of great importance later on.

‘We continued to condition you mentally, so you will accept more of what we give you. This is a long procedure and is still in operation at the moment. We will continue until we are satisfied. *This type of conditioning is done mainly at night, during the hours of sleep. While sleeping, everyone transmits a signal. This is normal and happens to everyone. We have instruments to 'tune in' to your particular signal; your personal frequency.*

Valdar continued, ‘You know that you all have a forward or conscious mind. You also have a mind that lies dormant, which we call the 'back' mind or the subconscious. This back mind accumulates a great deal of knowledge and a great variety of impressions lie stored there. A brain is like a computer. It may reach a stage where it will not accept any more. Then it refuses to receive new impressions and this mainly applies to one's subconscious mind. Then it becomes necessary to unburden this subconscious mind. This is to prepare you for all the new information you will acquire.

‘We have a method of conditioning by which we rid your subconscious mind of much that has been impressed upon it, making way for new conditioning. This has been carefully programmed for each one individually, to suit your mission and type of work. All data and information is placed there. This continues stage by stage, depending on how you accept it. It all depends on you.

‘We can tell what you have recorded of the information we have placed there and what you have received during the various transmissions via the thought-channel. Then comes a very important stage – the most important of all – and you are going through this stage at the moment. This is the stage in the program where you must accept a new way of life. I have already mentioned how strange things are beyond your solar system.

‘All this has been placed in your mind and it is done to alert your subconscious to be prepared. We are going to place in your memory-bank vital information that is needed if you should ever have to travel beyond your solar system, so that it will not refuse to accept what it will see and what conditions will prevail when you travel in Outer Space… You see, there are conditions which one cannot describe. One has to experience them to realize what is meant. One's body undergoes a drastic change in interstellar flight, for you will travel many times faster than the speed of light. Your body may be shattered into millions of particles and even your mind may experience this disintegration. This may sound more frightening than it is.

‘Now, when you are in that stage of disintegration, you will exist... and yet, you do not exist. For you (the consciousness) are *there* in 'energy particles' – you actually *flow*. Yes, the word flow describes it accurately, for you cannot leave your Universe unless you are able to flow through the magnetic field, coasting along the magnetic force lines. One cannot escape from your Universe and enter an other-matter Universe unless one undergoes change. You see there is a strange law out there, a law that has existed since the birth of the two Universes. There is a dividing line, which is invisible, beyond which nothing can go, *not even light rays*.

Therefore, to escape beyond that invisible barrier one must undergo great change. When the craft, including all its occupants, goes through this change, it is completely disintegrated into its elementary particles – into energy. After that one does not travel at speed anymore, as this does not describe it. One then travels in such a way that there are no words to express it... This happens only for a certain distance; the distance between the two Universes. Then one emerges into the anti-matter Universe and the craft and occupants revert back to solid bodies again, the process being reversed [to induce rematerialization].

'Let us put it this way. A container filled with water is suddenly spilled upon the floor where it spreads out. Now, suppose you had a magnet that could attract water. Place this near the spill on the floor and suddenly, it becomes one body of water again. This illustrates what happens when you disintegrate your body and craft into elementary particles - spread it out, so to speak - so you can flow from one Universe to another. You would flow through gigantic magnetic fields, which are of such high intensity that when you break through that invisible barrier you become energy and travel in time.

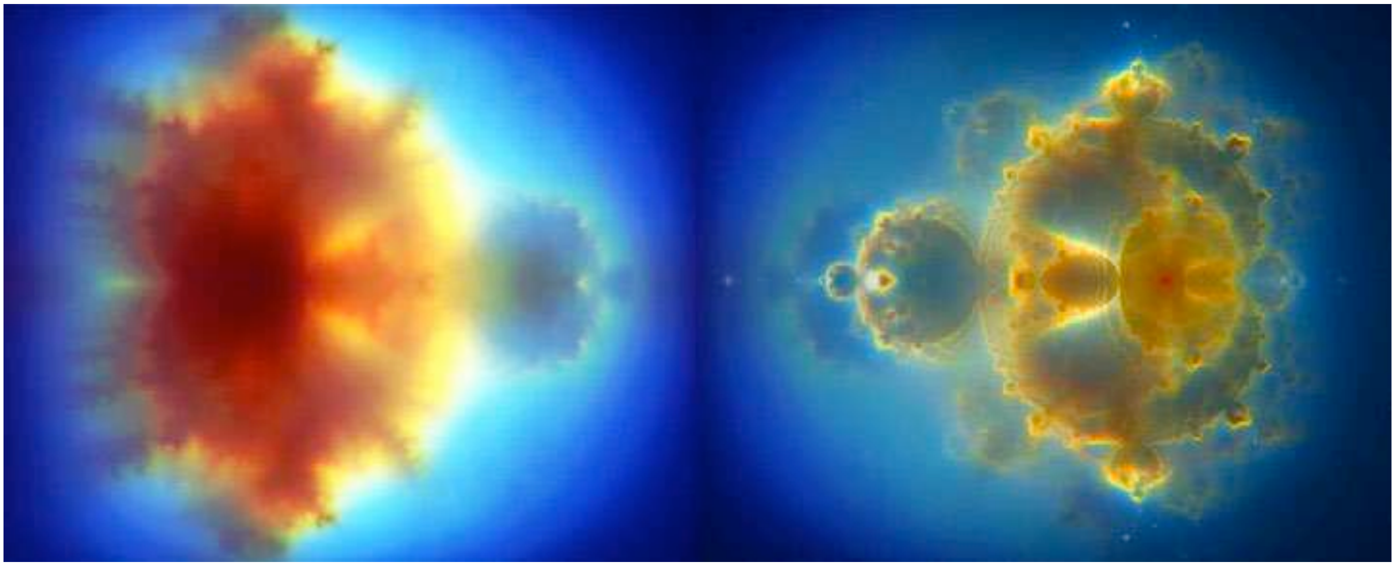

'This is the most important part of your conditioning because when you arrive on the other side, there is a world that you could never imagine! There are vast differences. You would not be as you were when you left *your* Universe. You would be in a new environment. Nature has molded you to live on Earth; you were never intended to go through the polarity barrier. It would be impossible for you to live on our planet the way you are now. To begin with, gravitation is so great that you would not be able to breathe. Koldas is such a large planet; the forces of gravity are correspondingly greater. You might survive, but under such great torture…

'That is why Nature remolds you when you penetrate that invisible barrier. It adapts you, not gradually, but suddenly, to the new environment. Even if your body might be able to adapt itself to the new environment, your conscious mind would not. Your subconscious mind does not care where it is, but your forward or conscious mind could not cope… It is fortunate that you all have a subconscious mind, otherwise it would be impossible for any of you to leave your Universe. Here we have a storehouse where all this vital information can be placed. We program your subconscious and tell it exactly what to expect; how to react; what to do and how to do it. So when you make your final jump and you arrive at the new world, your subconscious has been trained to take full command. Then you will no longer have a subconscious mind; then the complete brain will function – you will need it! Indeed you will. The new world is much larger – every solar system is so much greater; the way of life is so different…

The things one has to do would require your thought capacity to be so much larger – your concentration, your eyesight, speech, breathing, much more... then the subconscious would disappear... Only on Earth is the mind divided into a conscious and subconscious. Although we in the Confederation do have a subconscious mind, it is very small indeed. So you see, this is the most important part of the conditioning. Other stages consist of being taught various things, such as communication and telepathy and the acceptance of new colors. We respond to a greater range of colors. This extended vision is very important and has given us problems with some who have left Earth. We have found that their brains will not accept the wider range and intensity (of the colors) and they become muddled and very confused. So it has become necessary for our scientists to develop new means of conditioning to allow one to accept this.

‘It is difficult to explain all this to you. I do not compile this program for conditioning, myself; it is done by scientists who have spent their lives on this work. Each one of you is well known, not only to one particular scientist, but to many scientists. Every one of you has accepted and passed the first stage. Now there is a pause and it is my job to find out more about you. I have to find out much about each one of you, right down to the last detail, and all this is recorded and submitted to the panel of scientists who deal with mental programming. They then make a very careful study of all the data and select the best methods for your individual needs.

‘This may be different for each person. In some cases, photographs are needed. This we manage also and every detail is taken into account. It is very necessary that all these details are considered, as we cannot afford any errors in conditioning one wrongly, for it could do much damage, especially if the person needed to visit the antimatter Universe. If he or she were wrongly programmed, it could be fatal. Remember, in the fluid stage, while you are flowing between the two Universes, we rely on your subconscious to piece you together again in the correct way. For example, your height and your weight would increase. You would become more robust, *depending on which solar system you are visiting*. If your destination were Triaxula, then it would be in reverse. You would automatically adapt to the environment…

‘You see, once you are beyond the light barrier, that invisible line which separates our two Universes; *once you have made the first crossing of the line, you can at any time travel at or beyond the barrier, disintegrate and reform and coast through the magnetic fields at will.* You will find subsequent crossings much easier and faster. It is like being in two different seas. You are in one, and we are in the other. We, for instance, found that when we break through to your Universe, various difficulties arose. Difficulties in breathing [can occur]. Your sunlight interferes with our eyesight a great deal and we find your atmosphere very humid, even in your colder regions. So you see, once you leave your Universe and break through that barrier, your body automatically takes on the right shape and size and you adapt to the environment, provided that you have been properly conditioned.

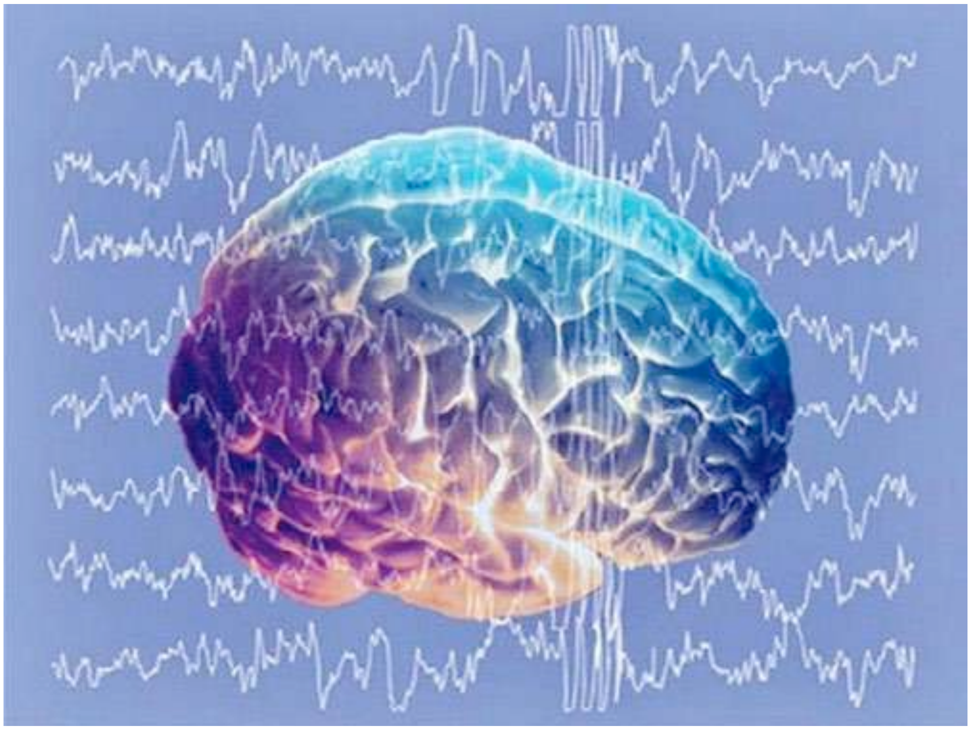

You will find that your brain capacity also increases and this is when the information which we have given you in our training will find a place. Your present brain is suited to your planet and present age. If you were to come to our planet with the same brain capacity as you now have, you would be in a great danger and trouble, for you would not be able to understand us and we would not be able to understand you. You would be as a primitive prehistoric creature would be if placed in your present society. That is why it is necessary to have this background training, this [voluntary process of advanced, subconscious mental] conditioning.

‘And when all of you have completed this, then you will be on our level. You are then full-fledged Koldasians, Byronnians, or Triaxulans or whatever the case may be. You would then find that you could travel at will from planet to planet, without any problem. And once you are beyond the light barrier, you will have no difficulty, even with the language. But this is new! The conditioning for language has never been used before for the scientists thought that it was far too much. Now they have realized that they can go much further and that is why, when you reach a certain stage, you will find that you have very strange dreams. These are very realistic dreams. You might even be able to smell or feel in these dreams. At first, this will prove alarming, for you may feel the heat of the day or the intense cold of the night. You will see things in your dreams, which we have placed there. You may, for example, experience living with nature.

‘We in the Confederation respect animal and plant life. We realize that without nature we would not live in a healthy and clean environment. As on Earth, nature is the same for us in the Confederation. Its beauty is very real to us! Mountains, streams, trees, flowers are the things we love –it is part of our life. Although we have a very highly evolved technical world, this does not mar our sense of freedom nor our appreciation of the beauty of Creation. We too like to escape from those technological marvels of science we have created. We too like to escape into the country to admire the beauty of the Divine One's handiwork, the planet he has given us to live on… Mainly this is what we impress on people's subconscious so that your conscious mind would know what to expect when you arrive. Your subconscious is trained to accept your new body, which would be totally different from what it is now. Sure, you would not turn green, or blue, but you would become larger people to suit the Koldasian environment. You are now being conditioned to be very close to the Koldasian standard.

‘Yes, going through the light barrier is a very big step,’ said Valdar, slowly emphasizing each word. Even for those of us who are used to [the changes]... For we undergo subtle changes when we arrive in your Universe. That is how we can mingle undetected with the population of Earth as we have done for a very long time in the past. But if we came here with the bodies of our Universe, we would be detected immediately. The major step is the ability to *flow* in the magnetic field through the light barrier. Of course, you would feel the changes here from planet to planet – gravity, heat, cold and so on. This you would experience, but your body has been conditioned to accept all this and much more.

‘Edwin has already reached the stage where we can use him to listen through his ears, see through his eyes and so observe local conditions on Earth. We can do this at our discretion and so keep in close touch with you all. Yes, you would be very surprised to see the *real* Edwin, the Koldasian Edwin – his twin.

‘Then, of course, age is a factor with the person being conditioned. The mind would also have to be adjusted to accept that the body would automatically become younger beyond the light barrier. This it does for some reason or another. It seems that you get a *new* lease on life, regeneration. This applies to all those born on Earth. *For us who are born in our other-matter Universe, long life is the norm...* There are special rooms where this type of conditioning is done. These rooms look like the operating theatres in your hospitals on Earth. The carrier I am in at the moment has three such large rooms. In the centre, there is a couch on which the person to be conditioned lies down. Then a device, resembling a helmet, is strapped on the head… This is connected to the strata-recorder and the crystals, and the thought-impulses coming from these are transferred via this helmet directly to the brain. We can accomplish this conditioning in about one hour. However, the person is first placed in a sleep state and firmly strapped-down where he or she remains until the light barrier has been passed. This is the normal procedure. Those who have been conditioned already would be treated as the crew and would be seated in the main control centre. This is where I prefer all passengers and the crew to be when we transcend. This main control centre is the safest area in an Astrael-craft. Once beyond the light barrier, things return to normal and one walks and sleeps, and one can use the craft as one would an ocean liner.

Valdar continued his explanation. 'I have used the word 'conditioning'. This is not strictly the right word to use; probably 'programming', training or teaching would be better expressions. *Remember we do not do this against your will!* It is entirely up to you. The very first step is that *you* accept us, then you automatically tune your mind to our wavelength and only then do we continue to prepare you for the next phase. I am really your teacher, with many assistants… Of course, those crystals and that valuable instrument, the strata-recorder, could not be dispensed with…

Then another aspect of this conditioning program emerged when the discussion turned to the enigmatic subject called the 'New Age'. Valdar said, 'One has to be conditioned to accept the New Age. In the ancient scriptures of the Confederation, it is written that the Divine One will change *many* planets.

'A voice tells us that there will be a change. The timing of the event will depend on many circumstances. It might be caused by the separation of the two Universes resulting in a complete break in the magnetic field between them. This event may bring about changes which would gradually usher in a New Age. There might be a change in the atomic structure. *Every atomic power station would then cease to function, and even your electrical power would stop. But a new source of energy will then be available…*

'Your physical world would undergo a transformation – a cleansing of the Earth. *Then the resurrection would follow in the still of the night. This transformation would be almost instantaneous. Even the heavens would change color!* Mankind will be afraid, but a great uplifting will follow. *There will be regeneration; for example, those of 90 years age will be as 50, and those of you who are 50 will be as 30*.

'Many would not see these changes. Only those of you who can now accept the new way of life will perceive it and experience it. Many are now being conditioned for that day... and also for cosmic flights, should that become necessary. Your subconscious mind is being conditioned to prepare for such an event, should it arise. Then, as if with the turning of a switch, true understanding would come. Some of you may become teachers or leaders…

> Valdar also mentioned again the disturbance in the magnetic fields, which, although unobserved by astronomers on Earth, caused an insurmountable barrier to all magnetically propelled spacecraft. This disturbance came to an end on the 18th of January 1981… 'Once the magnetic disturbance [of the December Solstice of 2012 occurs]… and the magnetic fields return to normal, then there could be a radical change on earth. When the changes come... and come they will, the rotational direction of Earth will be opposite what it is today. It would take the planet quite a while to readjust itself, but it will do so, for the earth is still a very young planet. These things happen to young planets! And they must happen before your sun becomes unstable and turns into a supernova and burns everything out. But there are still many millions of Earth years before that event.

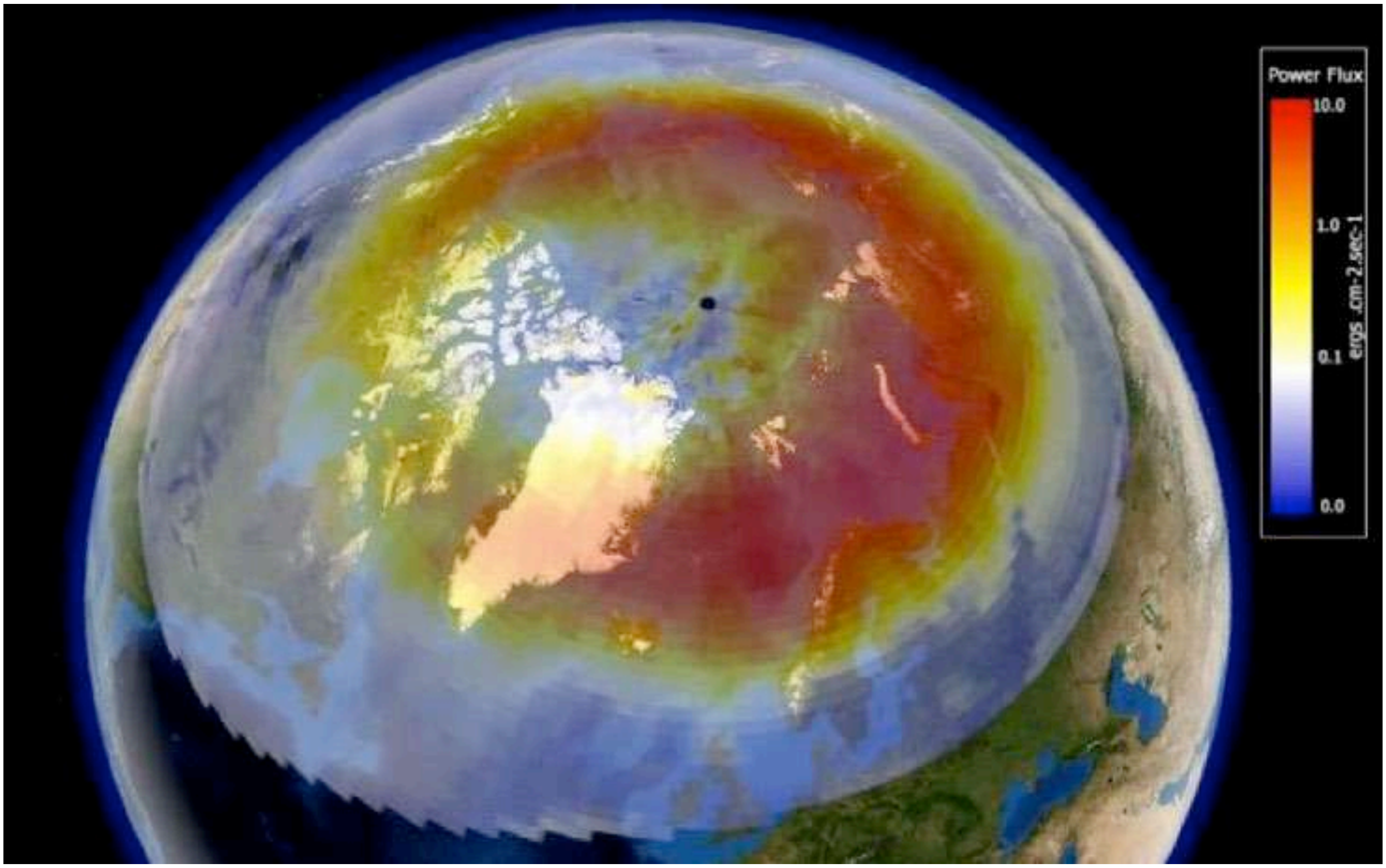

> 'But then there is the possibility that Earth may turn into a ball of fire as the result of a nuclear war. However, I feel that these changes, some of which I mentioned, will come before mankind on Earth can perform such a deed. Let us wait and see. There are still many things that even the Confederation is uncertain of. Changes can be observed and data analysed but only Mother Nature – the Divine One – will trigger the reaction that will change things permanently; I repeat *permanently* on Earth.
>
> 'The reversal of rotation of the planet would also change the polarity of Earth. North would become South and the South pole would become North. This could be brought about by a natural or an accidental cause… Great ships from space will come... then truly the planet Earth will be at peace. Those who remain will know our love and friendship. They will understand what the stars have to offer them. All this will be given to them and many things will be revealed in the New Age… And this is why we keep contact with you, to bring you messages of peace and love. We have tried to bring you an understanding of our planets and our way of life, which will be to your benefit when this great change arrives… and this will affect the whole of the Universe – not only Earth. These things have not been planned by us; we are just passing the message to you.' [17]

Valdar's clear statements elucidate dramatic future events and subsequent major atmospheric changes involving the oncoming shift of Earth to high resonance. *He foretells a 'physical… transformation – a cleansing of the Earth… in the still of the night… [that] would be almost instantaneous. Even the heavens would change color!'* A great increase in the intensity of cosmic rays bombarding the Earth, from a nearby supernova event or chain of events, would re-engage the now-dormant pyramids to ignite an atmospheric HHO plasma canopy, with plumes arching along the pathways of Fibonacci standing waves.

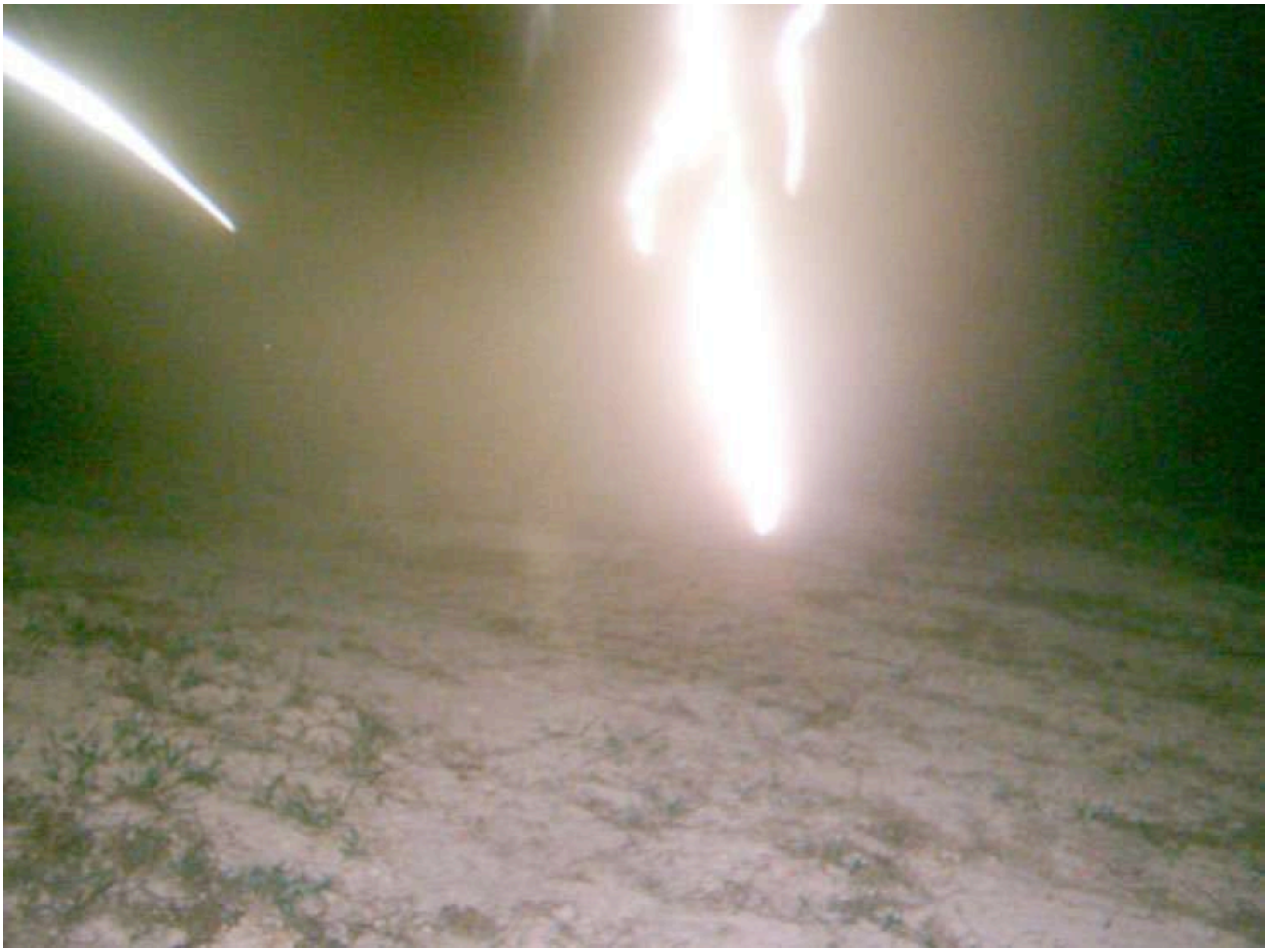

An astounding aerial HHO plasma phenomenon was photographed in detail on the auspicious date of 9/9/9 in Longview, Texas by motion-sensor game-trail cameras (above)[18], providing *recent, direct evidence* of the atmospheric and electromagnetic changes described by Valdar over 40 years prior to this enigmatic time preceding the magnetic reversal of December 22, 2012. The ultraviolet-A and infrared spectra emitted by the ionizing plasma plumes, as well as the precise standing wave configuration of the luminous arcs, confirm the photographed phenomena to be a product of ultra-low frequency irradiation of water vapor, causing its combustion by energized particles as HHO gas near ground level. *Atmospheric HHO plasma plumes and standing auroral patterns may become an everyday aspect of Earth's future.*

The inevitable result of high-energy atmospheric change will be *the purification of the ambient light* that penetrates many layers of HHO plasma to reach the ground, providing healing ultraviolet-A and infrared light to all terrestrial lifeforms. Just as demonstrated in HHO plasma beam healing devices like the Godlight cellular regeneration beam of Troy Hurtubise,[19] the atmospheric conditions will dramatically alter the lifecycle of all living terrestrial creatures: 'There will be regeneration; for example, those of 90 years age will be as 50, and those of you who are 50 will be as 30.' The regenerative effect of the frequencies of light emitted by HHO plasma promotes the extreme longevity enjoyed by advanced space-traveling civilizations on their homeworlds, and also generated inside the cabins of their interstellar spacecraft.

The highly resonant atmospheric purity of the planet Koldas affords a longevity that is significantly greater than that presently experienced here on Earth, although the oncoming energy shift may soon restore Earth's atmospheric purification processes, involving the plasma discharges of auroral activity and lightning strikes, that had become severely diminished over the preceding millennia:

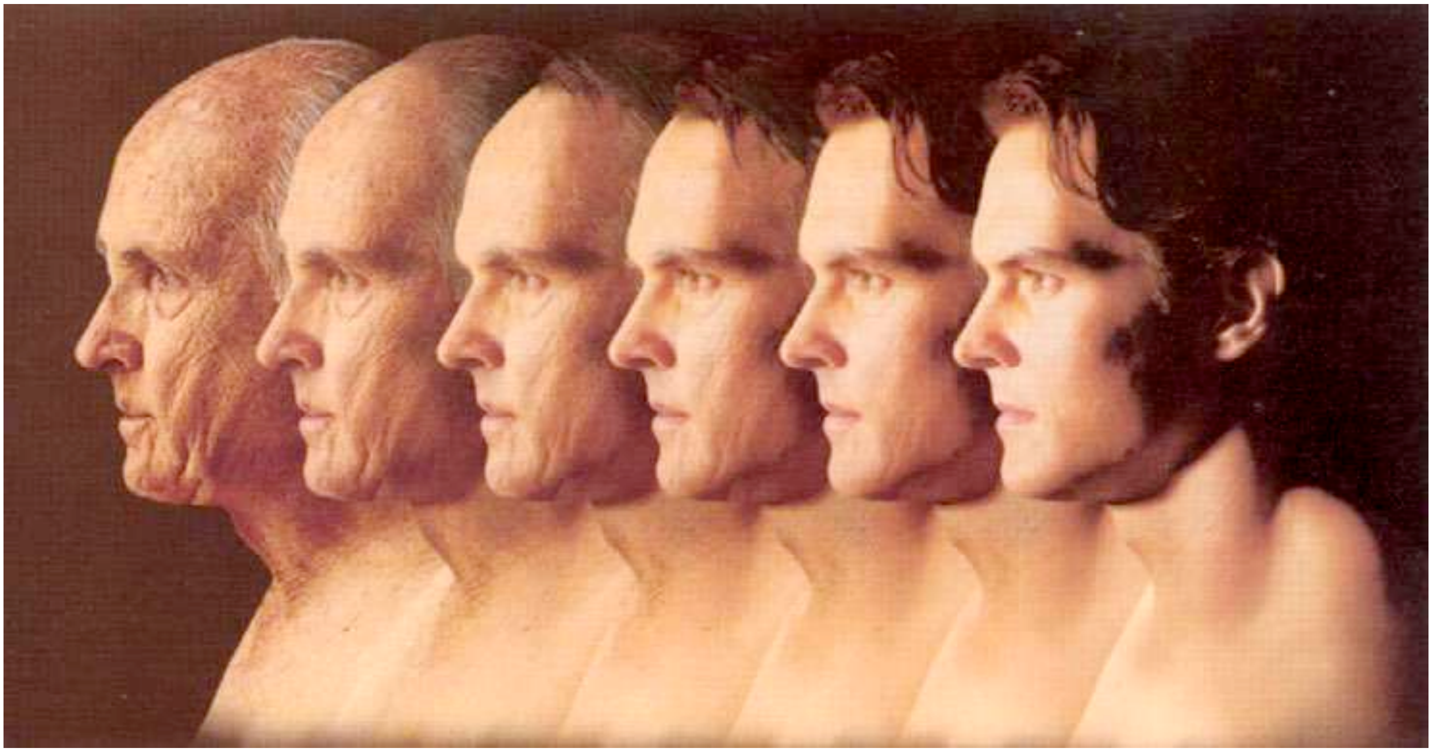

> 'Greetings... The message that I would like to leave with you this evening is, be patient, trust in us, for indeed you will realize that we can see a little further into the future than it is possible for you to do at this present moment. Trust in us and a new way of life, a new path, will be revealed to you, as the ancient prophets did in the past. Their monuments are still there today to be seen on your planet – the pyramids and the Sphinx…
>
> Other Superiors have governed Koldas, who have lived to the age of an equivalent of 250 Earth years. They enjoyed the full benefits of life as I do. I have no ailments and that is as it should be, and that is how it will be for those who share our life philosophy. *We are the instruments for the Divine Love and Truth to manifest through.* We are not Gods. Never think of us as such! There is one Divine Creator of all the heavens that you can see, and that we can see, and what lies beyond that again... I am Vax Noah... Salu kata katsu, farewell.' [20]

This heartfelt message from the Superior of Koldas was transmitted from high above, among the stars where many things beyond Earthly vantage can be seen. The penetrating knowledge offered to us by the Confederation of Twelve Planets was seeded for germination during the cosmic changes unfolding since the beginning of their contacts with Edwin in 1968. Present scientific understanding of enhanced cellular function under the influence of a mild electrical current, and under the high-energy conditions of plasma beam healing devices, allows the conceptual framework needed to grasp the truth of Valdar's insights. As he well knows from similar events in the history of other binary star systems like our own, the cellular rejuvenation effect of HHO plasma will be induced by the elevation of vast amounts of water into the atmosphere of Earth under the growing influence of an inaudible acoustic resonance.

The greatly increased ambient atmospheric luminosity generated by a glowing auroral canopy of HHO plasma, in combination with increased infrasonic stimulation, will alter the expression of DNA in all living creatures on Earth. In the terrestrial human being, pineal gland stimulation will bring about significant changes in the range of visible light and audible sound perceived. Our eyesight and hearing will extend significantly, aligning with the greater capabilities of Koldasian sensory perception. This was referenced quite literally when Valdar explained 'what we can see is more than you.' Layers of meaning within the simple, profound statements of the Koldasian visitors continue to unfold:

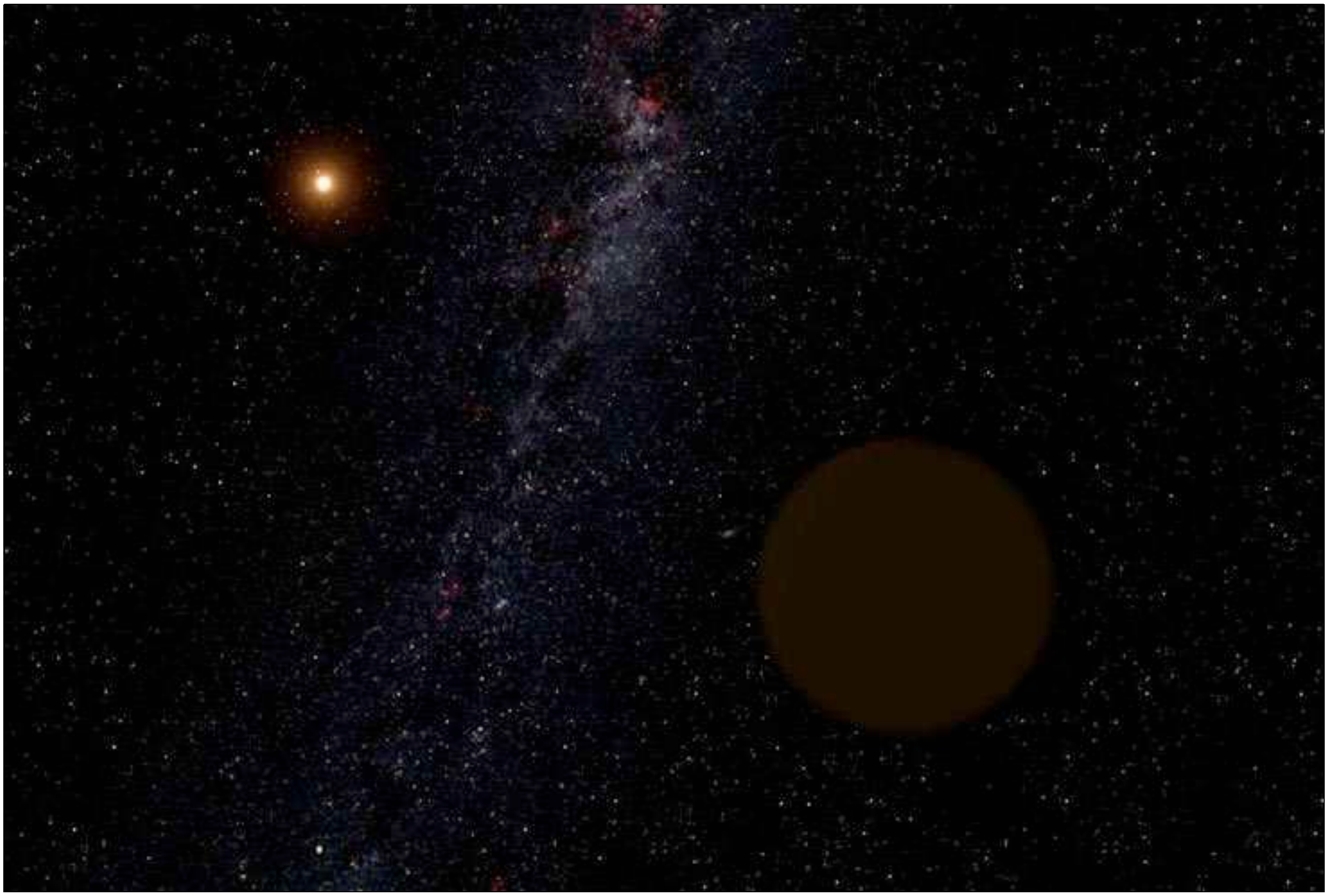

'There might be a change in the atomic structure*. Every atomic power station would then cease to function, and even your electrical power would stop. But a new source of energy will then be available.'* Indeed, the intense solar flare activity will send massive coronal ejections that will influence the ultra-low frequency resonance of Earth through interaction with the Giza pyramids, which transduce all solar energy into inaudible acoustic waves. This new source of energy which is 'here for mankind to use on our planet' refers to these resonant infrasound standing waves that will be dramatically increasing in strength during the events of December 22, 2012. But what unseen forces are causing this great shift?

Valdar divulged the hidden presence of 'Sola Kananda', the brown star binary companion of our sun that returns once again from the direction of Cancer on its long elliptical orbit. Although the full timeline of predicted future events was explained to Edwin in detail, the public was only informed that preparations and conditioning contacts made by the Confederation on Earth are being conducted in case an emergency evacuation of Earth becomes necessary in the near future due to these cosmic changes:

> 'Let me explain briefly the purpose of the plan. It is to save as many as possible in the event of a major catastrophe occurring on Earth. This might be a nuclear war or the collision with an asteroid or some other cosmic disaster involving Earth. This includes the approaching embryo sun, Sola Kananda…'
>
> We first heard about Sola Kananda approaching our solar system in 1973… Sola Kananda means a 'sun that has not yet flared or blazed', we are told. It is approaching from the constellation of Cancer. As our solar system enters deeper into its sphere of influence, changes in our planetary magnetic fields have from time to time been observed by the Confederation.[21]

These changes have been comprehensively mapped as nonlinear effects of focus infrasound standing waves that are increasing in intensity with solar flare activity. When considered in the context of the near approach of an invisible *embryonic sun*, one may infer that the 'instantaneous resurrection' of Earth is somehow intimately linked with the dramatic *rebirth* of the dormant brown dwarf star *into a red dwarf*:

> ‘In the near future, with the passing of Sola Kananda, your planet will pass through a certain region in space on its journey around your sun. This region is a belt which Sola Kananda has produced through space. This will bring about changes on your Earth. I am the one to inform you of... [these changes in great detail, which we cannot allow to be published]. This is why I am making this transmission. A great many changes will probably take place. Environmental changes on your planet will make your superiors and scientists look at space in a different light...’ [22]

With the dramatic revelation of the nearby brown dwarf binary twin, and all the auroral activity at specific infrasound standing wave convergence points, Earthly scientists will recognize the nonlinear alternations of locations of vortical activity and the connected design purpose of the world’s pyramids. But before these revelatory events are to occur there will be a brief preceding period of darkness during the time our solar system will cross the Galactic Plane, which Valdar cleverly referred to as *‘the still of the night’.*

Maya spiritual Elder Don Alejandro Cirilo Perez Oxlaj (above) affirms *this time of darkness may perhaps last for 30 to 70 hours during the Galactic Plane crossing* – thereby suggesting that *the solar corona may become extinguished* during the period of magnetic reversal, as if entirely swallowed by sunspots: when “we will enter the central magnetic axis… we cannot see the rays of the sun. We are inside. Once the Earth re-emerges from the axis we will then be able to see the New Sun.” [23] The galactic magnetic fields and cosmic rays of infrasound that drive our sun into incandescence will likely diminish altogether during the Galactic Plane crossing. Valdar referred to this as ‘the complete break in the magnetic field’, causing a temporary extinguishing of the solar corona during the process of magnetic reversal of the entire solar system –the sun, the brown dwarf binary twin star, and all of the orbiting planets.

Mayan wisdom reveals the brief darkness will end dramatically *in the bright flash of a synchronization ray emitted from the Galactic Central Sun*, causing ‘nearly instantaneous regeneration’ for all living things as well as global auroras and the incandescence of the red dwarf star. Koldasian Elders speak literally, not figuratively, when stating Earth’s *‘scientists will look at space in a different light’* –through an atmospheric plasma canopy glowing red, as *‘even the heavens will change color!’* Wisdom teachings preserved in the Mayan ‘Pyramid of Fire Codex’ unequivocally state*: ‘Come, reunite pilgrims, for the sky is in flames!’* [24]

> Valdar hinted at some future catastrophe and the changes that this event would bring with it. But he gave no due as to when this cataclysm or cosmic upheaval would take place as this was unknown even to the scientists of the Confederation. They are certain that it will occur as they have learned to interpret various signs in their long history of cosmic observations… The virgin lands where the ancient Atlanteans settled will one day be a haven for mankind again. For after the big change takes place, where there are now oceans, there will be land and where there is now land, there will be oceans. The continents of America, Africa, Asia and Europe will no longer be there. In their place will be these 'new' lands, now at the polar regions, and civilization will start again for we think your polar regions will not suffer in this big upheaval. They will remain more or less intact but the climate will change and the icecaps will melt. On the rest of Earth's surface there will be drastic changes! The present polar regions are even now being prepared, for they will be needed. You may have noticed that there is a lot of UFO activity there.’ [25]

Indeed, a fascinating component of the preparations mentioned involve giant aerial plasma displays over major urban populations, most spectacularly witnessed by residents of the Chinese cities of Dalian and Yantai in 2005; Tomsk, Russia in 2006; Trondelag, Norway in 2009 (photograph above); and over South Australia in 2010.[26] Such technological demonstrations of fourth-density plasma shake and arouse human consciousness to awaken for the long-awaited day when the humans of Earth fully accept the great offerings of cosmic visitors, as the time beckons when such help becomes increasingly necessary.

Recognition of the wisdom teachings of cosmic observers allows the integration of new concepts for a more balanced awareness. The peaceful and wise Confederation of Twelve Planets is but one of innumerable cooperative networks of cosmic civilizations exploring the endless magnetic fields in search of knowledge, toward the ever-present call of Oneness. Spiritual adepts among this small planetary group attain an extended lifespan by means of the radiations of the same high-energy HHO plasma that will soon be witnessed throughout the skies of Earth. The dynamic shift of Earth resonance to a higher level follows a pattern of revelation that inevitably occurs during the ascendance of all young planets and all nascent civilizations, along the path toward collective consciousness and cosmic co-creation.

References

---

[1] van Vlierden C, Ed. Stevens W (1986) *'UFO Contact from Planet Koldas: A Cosmic Dialog'* UFO Photo Archives, pp. 79-80

[2] Ibid, p. 46

[3] Ibid, p. 29

[4] Ibid, p. 83

[5] Ibid, pp. 16-17

[6] Ibid, p. 12

[7] Ibid, pp. 58-59

[8] Ibid, pp. 104-105

[9] Ibid, p. 21

[10] Ibid, p. 167

[11] Isenberg H (2003) 'Giza, Center of Earth' *Mars News*, online · http://mars-news.de/pyramids/gizacenter.html

[12] van Vlierden, op cit, pp. 26-27

[13] Orion E (2003) *'The Bush Connection'* The Bush Connection, Section B, p. 13

[14] Tesla N (1978) *'Nikola Tesla: Colorado Springs Notes 1899-1900'* Nolit, Beograd, Yugoslavia, pp. 61-62

[15] van Vlierden, op cit, p. 80

[16] O'Neill I (2010) 'Mystery Behind Solar System's Giant Ribbon Solved?' *NASA,* January 15, online · http://news.discovery.com/space/mystery-behind-solar-systems-giant-ribbon-solved.html

[17] van Vlierden, op cit, pp. 186-189

[18] Moulton Howe L (2009) 'Eerie 'Blades of Light' in Longview, Texas – September 9-10' *Earthfiles*, October 10, online · http://www.humanresonance.org/longview.html

[19] Novak P (2005) 'Angel Light Ascends to God Light, Part 2' *Bay Today News*, online · http://www.baytoday.ca/content/news/details.asp?c=8271

[20] van Vlierden, op cit, p. 200

[21] Ibid, pp. 142-143

[22] Ibid, p. 228

[23] Cirilo Perez A (2009) 'The Messages of Wandering Wolf', *Shift of the Ages*, online · http://shiftoftheages.com/wandering_wolfs_messages

[24] Jenkins JM, Matz M (2004) *'Pyramid of Fire: Spiritual Ascent at the End of Time'* Bear & Company, p. 88

[25] van Vlierden, op cit, pp. 132-133

[26] Putney A (2010) 'Plasma Spirals' *Human Resonance*, online · http://www.humanresonance.orh/plasma_spirals.html

3
The Plejaren

The Plejaren

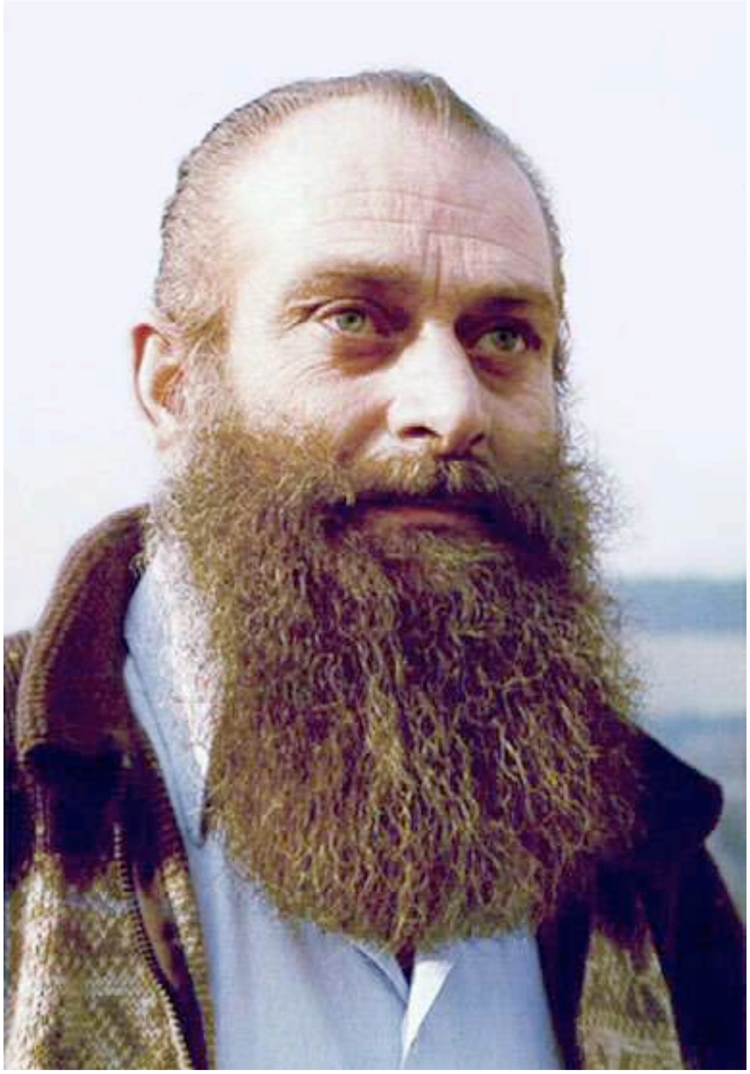

Eduard Albert Meier (b. 2/3/1937) is a Swiss farmer who has been visited since the age of five by extraterrestrials from the region of the Plejares star system, representing a wise and peaceful Confederation of planetary systems. Eduard's childhood contacts were made by an aged man named Sfath, while as a young adult he maintained contacts with a female cosmic visitor named Asket. Through his later years he maintained contacts with Quetzal, Semjase, her father Ptaah and many others.

'Billy' is a nickname Eduard received in Tehran during his long travels on foot through the Middle East and Asia Minor (seen at left, in 1980). Billy Meier has documented years of contacts and remarkable dialogs with visitors from the Plejares star system. He has also produced several 8mm documentary films of Plejaren spacecraft, or *beamships* as they are called. Meier has also produced several hundred high-quality photographs of extra-terrestrial spacecraft of many types – both from the ground and *in the air, aboard the Plejaren spacecraft itself.* The sheer breadth and diversity of highly advanced information and physical evidence provided by Meier concerning his contacts and cosmic travels with the Plejaren calls for the close attention and discernment of the human beings of Earth, who grow in awareness by the slow process of assimilation of knowledge.

While many contactees have experienced transdimensional travel and orbited the Earth and her oceans, Meier is one of the few present-day humans who has time-traveled by hyperleaps through deep space with his cosmic friends and returned with photographs and scientific details that are undeniably valid. The most profound teachings that imbue Meier's life's work are the spiritual teachings he has received from the Plejaren and presented to humanity through his many books, now available in several languages. The simple and penetrating spiritual teachings of the Plejaren are now transmitted through the Semjase Silver Star Center in Hinterschmidtrüti, Switzerland, providing an opportunity for synchronized meditation practices with billions of advanced spiritual beings throughout the Confederation's unified network of planets, who choose specific times for coordinated interplanetary peace meditation on a regular basis.

The Confederation comprises a collective of diverse civilizations that remain deeply rooted in spiritual connection through their respective planets, providing a guiding influence on worlds in crucial phases of developmental shift, as is the present case of Earth. These spiritually and materially advanced visitors remind us that even they grow food in organic gardens, and maintain *direct contact with nature by planting with the hands in the soil, in the creative intention of love.* Plejaren wisdom finds guidance in the Laws of Creation, and encourages humanity to learn for ourselves through investigation, thoughtful reflection and the application of reason. In spiritual aid of the human beings of Earth, Billy confers their extensive spiritual knowledge and physical evidence in the form of photographs, films and even samples of exotic materials.

Eduard's space-traveling visitors explained their ancestral origin in the Lyra star system, before an ancient cosmic migration to the Plejares star system, including the beautiful planet of Erra, which resides in a *fourth-density space-time configuration about 80 light years beyond the Pleiades star cluster* (above). The name Plejaren ('play-yar-en') means *'Law of the Sevenfold'*, reflecting a penetrating cognizance of Creational Law and the Truth of Spirit. Offerings of the Plejaren Elders on the physics of Spirit were balanced by comprehensive lessons on solar and astrophysics. Meier was informed that the Sol system, outside the Milky Way, constitutes something like an enclave because it floats as a single system far outside a spiral arm, moving at high-speed towards the Hercules star system. The distance from the Sol system to the middle of the galactic central sun is 35,002 light years, according to Plejaren calculation.

Spiritual demonstrations of Semjase's intentional healing ability were given during various contacts when she rapidly restored Meier's dwindling health by concentration alone. During the 50th contact of April 6, 1976 Eduard's broken ribs were mended in minutes by an electro-healing apparatus within Semjase's beamship (opposite), extending natural mental capabilities through technological means.[1] The ability of the Plejaren visitors to predict all major events in Eduard's life was repeatedly demonstrated, and even the date and cause of his death were previewed and discussed with him in detail. Another definitive event in his life predicted by the Plejaren, for his psychological preparation, involved the loss of his left arm in a bus accident in Tehran in 1964. After the necessary amputation, Meier healed in conventional hospice and wisely refused a biotech prosthetic arm offered by his extraterrestrial friends, on the grounds that his safety would be compromised by the greed of other humans who would inevitably attempt to steal the advanced extraterrestrial biotechnology. *In this, he has found the resounding truth of the spiritual lesson.*

Rather than secretly enjoying the material benefits of his contacts, Eduard Meier instead disseminates the advanced spiritual teachings of the Plejaren Elders, inviting each Earthly human being to consider the truth of spirit. The path of inner seeking through meditation and reasoned thought has become the focus of FIGU, a conscious community that has co-founded the Semjase Silver Star Center. Serving primarily as a meditation center, the facility synchronizes the FIGU community for the development of telepathy by exposure to focused infrasound. *From there, Eduard's enduring teachings have radiated their light:*

> In truth, the Earth human lives a life that does not ever deserve the designation *life*, in the true sense… [for he] no longer knows of the strengths and possibilities that lie within him and are suppressed. The human's psyche has become a prisoner of the material, and the human's corporal nature hinders him from glimpsing more than only a weak shimmer of the true light. But certainly the day will come when the human will venture forward somewhat further than usual in his thoughts, and then still a little bit further, and further.
>
> Then the inner light will gleam in him and he catches, to start with, only a small and fleeting glimpse of his true essence. Indeed, then comes the point in time in which he becomes somewhat more aware of it, and then still somewhat more. So the light of truth with which, namely, he is one, with eternal life and with the eternal power of his essence, gradually comes on for him. This light then illuminates his entire psyche and takes possession of it.
>
> Awakened through this new consciousness, the human holds the thoughts of this oneness firmly, and thereby now progressively lives in these thoughts, his life now flows in constant actualization of the oneness of the creational Being. And thus the terrestrial Earth human is transmuted into the human who is of Creation. Compared to the new life, the old one only deserves the name *death*, because ignorance and narrow-mindedness is like that. But this deadly life is no longer able to shackle the new human who, before, in the old life, in a certain sense, had actually been dead -a living corpse. (OM: The Book of Truth, 43:153,167-177) [2]

Knowledge of the inner dimensions of the human being inspires reflection on the endless gifts offered by the Creation, granting natural relief from the burden of the physical through the transcendence of spirit. The moral dilemma which now cripples Earthly consciousness, namely the role of psychopathy in politics, was solved naturally by the Plejaren science of biorhythmic healing by genetic enhancement through the tri-thalamic heartbeat entrainment during fetal development.[3] This advanced preventive infrasonic genetic technique also relieves the moral crisis of abortion, which is permitted by natural chemical methods using essential oils (like that of the Neem tree found here on Earth[4]) within the first three weeks of pregnancy only, as "the time of the beginning heart activity coincides with the entrance of the spirit of the newly conceived human being, so on the 21st day after conception." [5] The determination of spirit in accordance with Creation propels the consciously seeking individual from within his/her deepening experience of the luminosity of his own being, as taught by Semjase during the 18th contact of May 15, 1975:

I want to proceed with the teaching of the spirit, as it is the most important of all things… The experience reveals that the Creation is unending beauty, a beauty above all beauty, borderless, intensive happiness without end, wisdom, knowledge, ability, truth and absolute determination. When that is recognized by the human then he leads every one of his joys instantly back to its origin, to the unending joy of Creation.

Wherever he sees something beautiful, be it now a flower, an animal or a human, or something else, he thus brings it instantly into interconnectedness with the infinite beauty of Creation itself. Wherever the human sees cognition reach expression, in this or any form, in an impressive and exalted manner, then he knows that it has its roots in the endless cognition which is Creation. Wherever life stirs itself, be it even in the tiniest Being, in a creature, as, for example, the microbe, there, behind this life he glimpses the infinite, the eternal and the creational. Out of that he attempts to deepen and expand his understanding and feeling through which he contemplates the essence of the creational and its inseparable presence at any time and in any space in many kinds of ways in the daily life and experience.

Creation exists in every human as a fragment of himself. Once this thought has first deeply penetrated the human and become his experience, then any fear and all doubt in him disappear. If he knows that the Creation is all-knowing, and almighty then he gains in inner peace and security and is immune from stupid thoughts and wrong feelings. Repeated contemplation about the all-present, the truth, wisdom and the ability and knowledge of Creation as constantly eternal, as dignity and worth, allows the word and the designation Creation to become something very meaningful in the human – to something which calls forth transformations in the feelings and transformations in his manner of thinking.

The more his intelligence is effective in this direction the more he gains radiating light, the more powerful his personality would be and the more blessed his entire life and works would be. Again and again the human living in clarity of consciousness would produce in himself, anew, the strong perception that the Creational is far more real than his body's feelings. This perception rules, without interruption, the human living in accordance with spirit. The creational grasps possession of his consciousness whereby his senses are full of peace, strength, joy, knowledge, wisdom, truth and hope. All measures that the normal human grasps in his material-intellectual thoughts in order to attain peace, happiness and strength always show themselves as deceitful, while the constant controlled-ness of the human who is living according to spirit would be pulsed through with the dynamic power of Creation.

He who is a human living according to spirit is very dynamic in all things. He attempts, uninterruptedly, to reach his goal as quickly as possible. As long as he lives he would, in accordance with this, use the time for applying the creational ability with all the energy available to him. He would never allow this undertaking to be lost. Everything possible can happen in the course of time, yet his desire for the creational would never be extinguished in him. He can strike hindrances in which he suddenly has no sense and taste any more for the necessary material things, yet he never loses the taste and sense for the creational because he knows that the Creation embodies the true Being.

Only a human who fights hard for consciousness and spiritual goods and progress, for knowledge, truth, logic, wisdom and love harvests mighty fruit of a spiritual and consciousness kind, because they do not simply fall into his lap. First, before the first results would be obtained, it is required that the spiritual-intellectual manner of thinking must learn the path of creational thinking and recognize its absolute correctness and determination. Yet when these first results come about then the cognizant one steps onward with great steps and expands himself in spirit and consciousness to a factor of power. First, through this, already known facts, knowledge, truth, logic, wisdom and love would be self-evident, which first, however, must be gained through hard work. Indeed, the path does not end at this point, because further seeking, further searching, further development and further recognition lead into the borderless duration of time.

Everything possible may happen in the course of time, and hold the human back from taking up deeds, yet the human living according to spirit knows no borders and does not allow himself to be kept from his goals by any kind of events or bad future out-looks. For him the future already exists in the present, so it is that everything must be done here and now in order to reach the highest spiritual and consciousness state… Again and again he produces in himself the strong perception that something is there which gives him immeasurable power and makes him free from unreal assumptions; (it is) the truth of Creation.

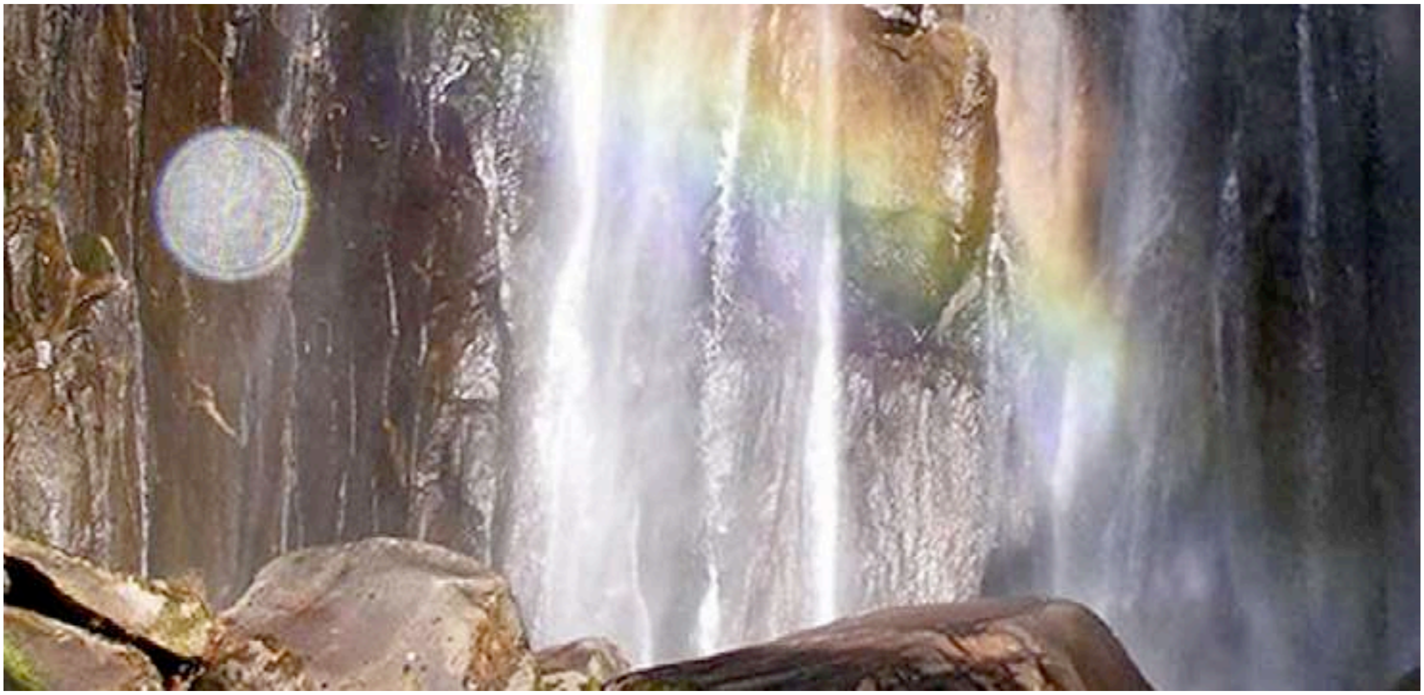

Always again and again the human produces in himself the strong perception that he finds, in the ocean of creational light, his wisdom, his knowledge, truth, logic and love, which for him first entirely enables the Being of life. The joy of the human who is turned towards Creation exists as a result of his reverence produced from the creational and Creation; (it is) this, in which he accepts the almighty will in the creational laws and turns the absolute determination of these laws to his own determination and practically employs them.

He brings his dedication to the laws to expression through the learning and utilization of all spiritual and consciousness facts, but never through belief, assumptions, serving and humility. The only way to learn exists in the unremitting efforts and striving to obtain higher spiritual and consciousness cognitions and to bring the capabilities into application which have unfolded through this. Patience and endurance and the development of higher understanding, recognition and engagement of the cosmic and universal love, deepening of the spiritual and consciousness knowledge and ability, as well as the shutting-out of material-intellectual thought powers, like egoism, materialism, pride, envy, greed and jealousy and so forth, are thereby of decisive significance, because only this guarantees the recognition and obedience to the creational laws…

He who is a human living according to the spirit attempts constantly to maintain himself always, and under all circumstances, in the vibrational realm of creational motions. Should these once, through some kind of influences, be in danger, then he calls on Creation in the highest power and thus protects himself in its highest vibrational power, from the negative. In this manner he proceeds so far along until the creational essence in him has produced the absolute determination of defense against the negative.

> [In his thoughts and his waters,] the human is always the bearer of some things. Many carry oppressive feelings in themselves, others grief, worries, problems… But he, who is a human living according to the spirit, is a bearer of the creational, of the spiritual… He immediately transforms even the most negative thoughts that come and want to destroy his spiritual-intellectual thoughts, into creational power and wisdom. So he makes everything creational and carries it as a block of power in himself. Thereby he is a traveling temple of Creation.

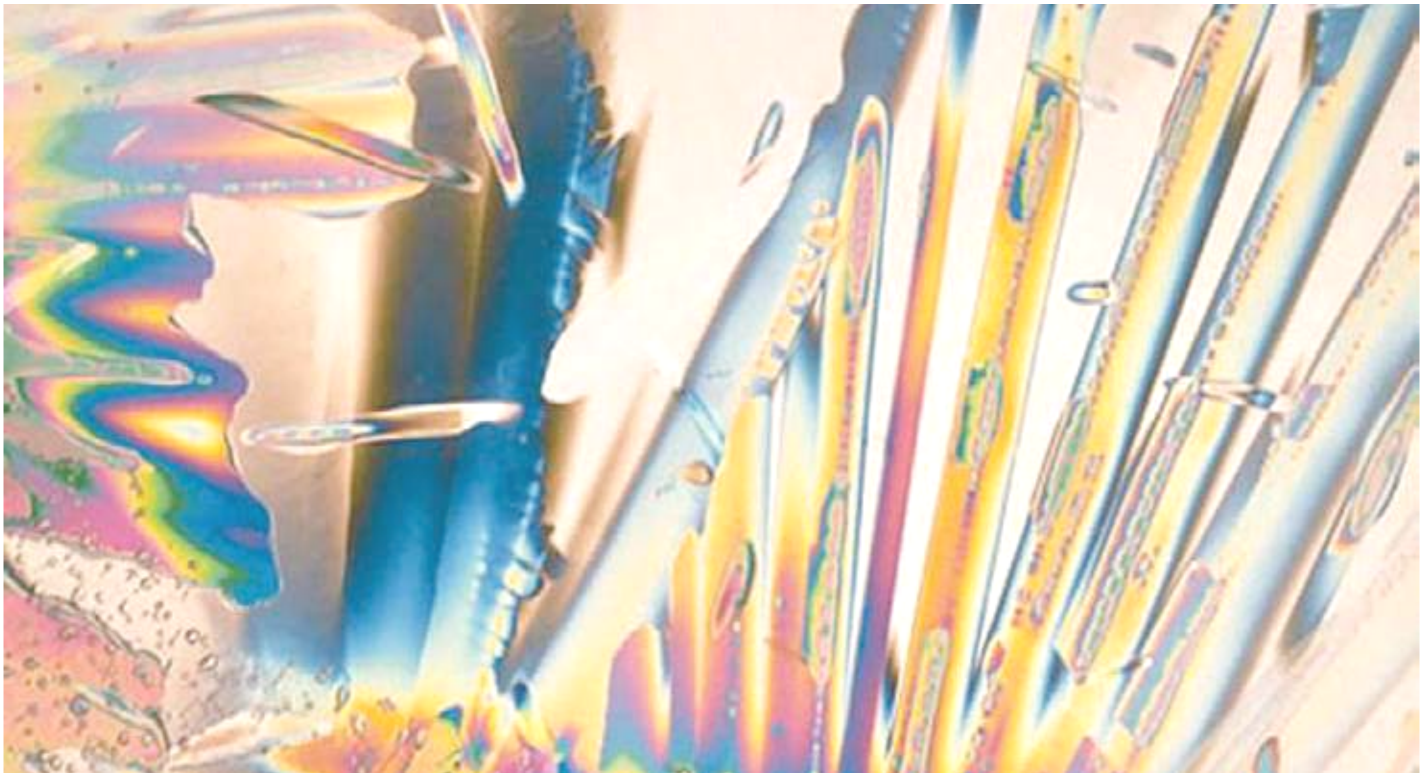

> The intelligences of the human living according to Creation would be ever more inspired and press always deeper into the spirit-consciousness. He who lives according to the spirit presses always deeper into all things with his cognitive abilities and grasps the warmth of the fire of the creational presence in everything. Thereby the spiritually thinking human would, over-all, be cognizant and conscious of the power and the truth of the creational presence. What therefore remains unrecognized by the everyday perception of the senses, [–as *cosmic infrasound–]* would be spiritually-consciously perceived by the human living according to Creation…
>
> The material life on the Earth is like a passing game, an evaporating phenomenon, but behind it lies the eternal and timeless truth; the spirit of the creational presence and the reality of Creation. This creational reality carries in itself all riches and realms of the ability to do with wisdom. This is the remaining, the timeless constancy and imperishable. This is no game, rather the eternal and timeless truth and wisdom, knowledge, freedom, love, logic and ability, the relative perfection, the absolute determination…[6]

The pristine clarity of Semjase's spiritual teaching continually unfolds within the conscious listener, expressing the source of all joy in Truth and the oneness of Creation. The purity of spirit and thought present in the Plejaren wisdom teachers reflects their physical purity, drinking the celestial waters of their resonant home planet of Erra. The comprehensive knowledge of this advanced cosmic culture directs their harmonious lifestyle to benefit from a collective consciousness and the telepathic abilities endowed by following the resonant life-path set forth by The Spirit. Dwelling within the field of heartbeat resonance, one enters into direct communication with the living instruments given to us by the Creator, to hear the joyful laughter and all-penetrating cosmic *Om* – the vital force of the human being engaged in the 'synesthetic' state of inspired conscious co-creation.

While Eduard had been educated in the teachings of spirit since his earliest contacts with Semjase's grandfather Sfath –at the age of five– comprehensive spirit teachings were offered in later contacts with Semjase, beginning on January 27, 1975. The contacts of the 1970s involved many lessons on the Truth of Spirit, but also began a period of extensive photographic opportunities for material evidence.

At left is a montage created from Meier's original February 27, 1975 photograph taken at Jacobsberg-Allenberg/Bettswil, Switzerland. Meier has even taken photographs *from inside Plejaren beamships in flight* (below). The blurring and tinting of the image is a result of special coatings used in the windows, likely comprised of quartz glass, which glow orange in our atmosphere. The resonant acoustic and electromagnetic environment maintained inside the Plejaren beamships interferes heavily with the camera functions, often overexposing the film due to high-energy radiations or distorting images beyond recognition. Dozens of remarkable photos taken in space are marred by significant focal problems.

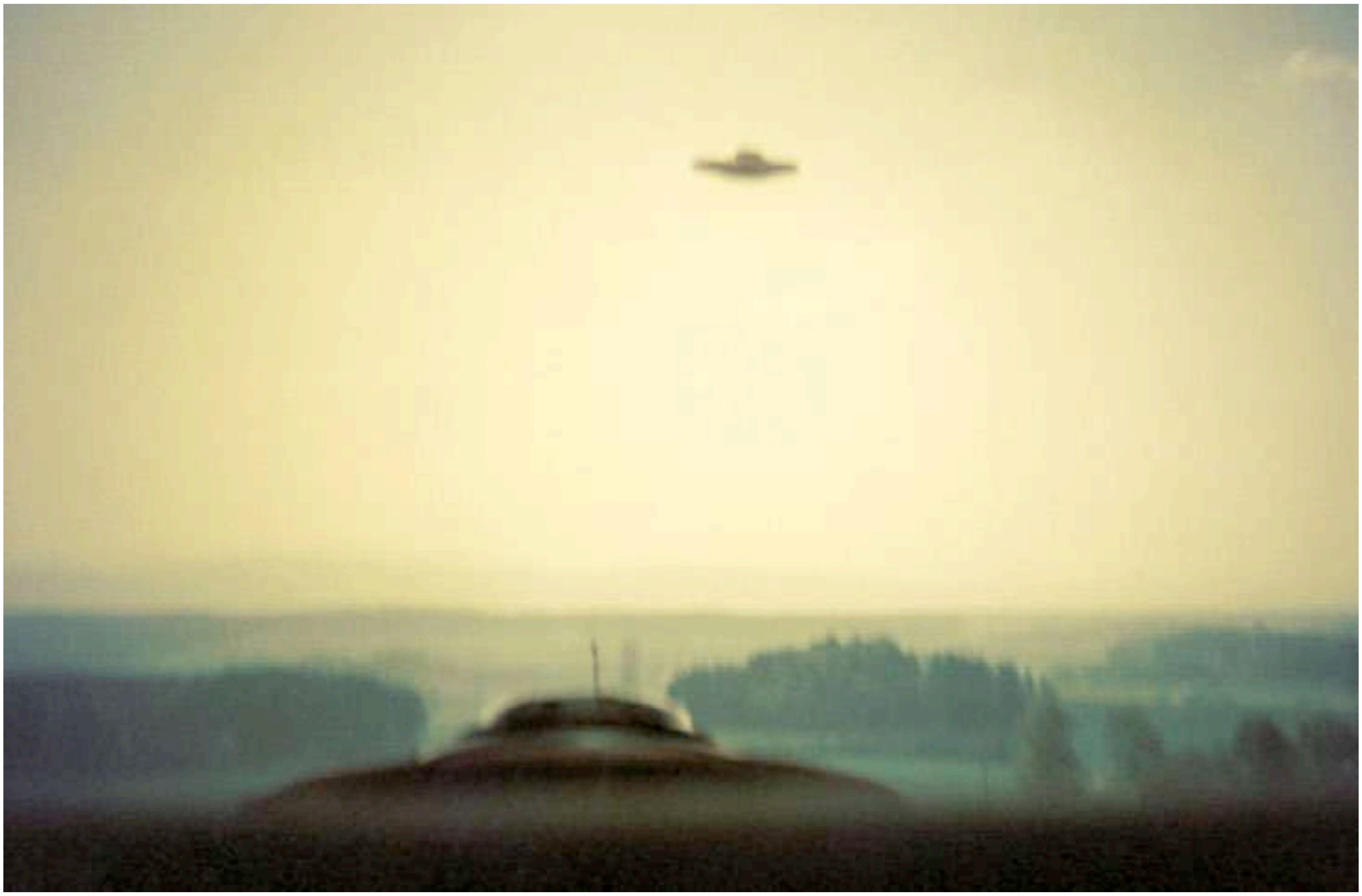

All of the photographs produced by Eduard Meier during his contacts were obtained by permission of the Plejaren, while certain sensitive areas were prohibited as subjects for documentation. Portraiture of the Plejaren visitors was not permissible, nor was Meier allowed to photograph the interior of the beamships. Why do they not yet present themselves to the human beings of Earth openly? Semjase's father, Ptaah, provided Billy with a concise answer to this question, revealing a profound understanding of the nature of human consciousness at the present time, and the primacy of reasoned thought and careful deliberation:

> *Ptaah:* When the epochs change, so also do the minds of the forms of life. This has also happened to humans now. The Age of Fishes was characterized by religious fanaticism… Then the birth pains of the Age of Aquarius began, and suddenly the Earth human started to listen inside of himself. And he started to think and to explore, and suddenly recognized logic only in truth. He developed in a short time his thinking abilities, and searched and explored… which is a characteristic of the Age of Aquarius. And this is one of the main reasons why certain human beings are now being contacted by extraterrestrial forms of life again. The present Age of Aquarius demands thinking and spiritual evolution of the humans of Earth. This cannot be achieved by seeing and listening with the physical organs only, but only by reasoned thought. If we appear officially in our beamships and spacecraft, the effect of deliberation and of thought would suffer, because the Earth human, in his present state of evolution, does no longer deliberate and search that which he can see with his own eyes and hear with his ears. Beyond the seeing and hearing, the interest is short-lived, because that soon becomes a natural event to him, and he no longer cares. What he knows no more interests him. His interest is only aroused by secrets for which he can hunt.[7]

The Plejaren visitors always wore a protective suit that Meier described as matching the texture and color of elephant skin, but was creased and folded to the exact forms of its wearer – perhaps some kind of organically grown material fit to form. Meier also described a reflective circular metal neck fitting with the apparent function of a helmet-mounting lock ring. The Plejaren go helmetless on Earth without exposing themselves to our polluted air, as this circular neck fitting is electrified to create a *standing wave* over the nose and mouth that purifies and sterilizes the air taken in. This apparatus protects the wearer from radioactive uranium dust in Earth's atmosphere due to humanity's seemingly insatiable lust for war.

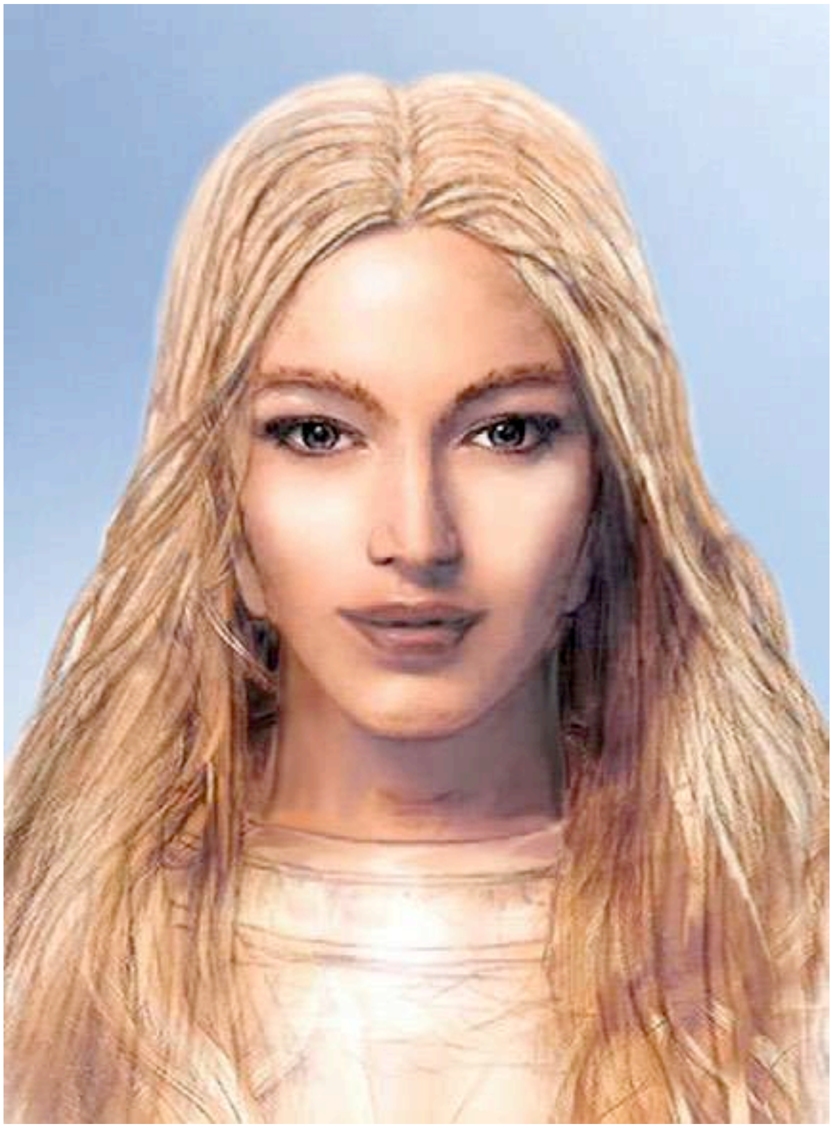

The low-resonance environment of Earth had long ago reduced human lifespans to the brief span of near 100 years. This limit on longevity to which terrestrial humans have all become accustomed is a drastic and direct result of the weak and offset magnetic field of the planet, corresponding low levels of infrasound resonance and very high levels of deuterium (the heavy form of hydrogen) in our waters. Traces of *heavy water* pollute the waters of Earth and her surface-dwelling human populations, yet the process of her cleansing and transformation into a pristine paradise already grows from seeds within the human heart that will require a quantum shift to high-resonance. The high-resonance environments of planets with atmospheres containing only *light water*, comprised exclusively of protium (being deuterium-free), engenders human lifespans of perhaps 1,000 years or more as enjoyed by the advanced Plejaren civilization. An enhanced and colorized image based on a self-portrait drawing by Semjase (at left)[8] captures the refined grace of her features at the age of 330 years, in 1975. The facial features, size and proportions of the Plejaren race are identical to that of Earthly human beings with the exception of their long, connected earlobes, which may have inspired an association with wisdom long ago (pg. 4):

> [Concerning lifespans, investigator Lt. Col. Wendelle Stevens concluded as follows:] This 1,000 years average age for the [Plejaren] seems strangely coincidental, as that is the [Bible's] reported age of Earth humans before the Great Flood, when the dense vapor envelope surrounding the Earth came down in torrents everywhere for scores of days. Loss of the vapor envelope reduced the atmospheric protection of Earth's creature species from the harsh [UV-B and UV-C] radiations of our sun, and the lifespans of all creatures decreased quite dramatically and rapidly to less than 100 years for humans, and vegetable and plant yields were reduced accordingly.
>
> Recalling the extraterrestrial's statements that we are descendents of their ancestors when they were here long ago, our 1,000-year biblical longevities take on new meaning. Loss of that atmospheric protection has shortened our lifespans to 1/10th its former scale… An interesting observation made later on in the contacts by one of the extraterrestrials, describing a protective [HHO plasma] screen they use around their vehicles, and even around their person when out of their vehicles, indicated that they age in our atmosphere as much as we [terrestrial humans do] without their protective screen![9]

The vital force of *infrasound* expresses itself much more intensely in the highly evolved atmospheric conditions on Erra, imbuing an extreme longevity and advanced consciousness among the Plejaren through psychoacoustic entrainment. Meier initiated a profound lesson during the 39th contact of December 3, 1975 by pointing out the brevity of the human lifespan as a result of "the fact that the forces of thinking of each single human being are able to raise or lower the force of life of the fellow creatures":

> *Ptaah*: Certainly, this question is of great importance… The age of every creature is shaped partly by the penetration of its outer forces by others of the same kind of forms of life. Because the thinking of a form of life in cooperation with spirit generates an extremely logical, that is Creation-accorded, force which is released as high-frequency radiations and oscillations, an immense force is generated which is able to influence everything. The Creation-accorded force penetrates all, truly into all, material forms of life and matter, and influences them according to their kind and form. Each form of life owns an, adapted to its spiritual level, age of living or time for living, thus as well with the human beings of Earth. In earliest times, when the human being of Earth was procreated here by his ancestors, his average age of life was 1,007 years of living, and as taught by his procreators, possessed an enormous spiritual level of knowledge and abilities.

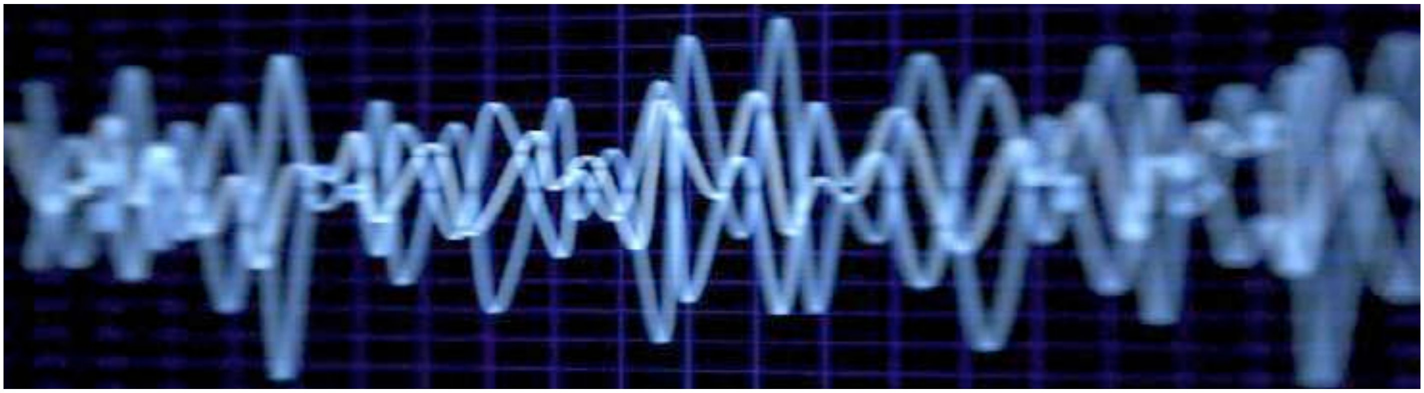

> But unexpectedly rapidly, he fell into religions and their heresies, and by this lost true knowledge and the truth. He necessarily began to work against natural law, became a stranger to it, and lived under the bondage of self-created laws. This all together resulted in the loss of the high average span of life, which in a few thousands of years decreased ever more and settled at the twentieth part of its earlier time. In the beginning of the new epoch [of the Age of Aquarius] a change for the better once more marked itself, from which the average span of life slowly increases. The main reason for this is to be found in the acknowledgement of the truth and connected spiritual direction. The more then that spiritual direction turns itself again towards the truth, the higher again develops the average span of life, because here, by this also, the altered genetic and other factors regulate themselves again, which in the course of millenniums had become wickedly influenced. The mass of Earth human beings moving in spiritual form is still relatively low, and in many cases only beginning to recognize the real truth, thus hitherto still no great progress could be achieved. But the more the Earth human being turns himself towards the real truth, the more does also increase his average span of life.
>
> Seen as a whole, every form of life is of important meaning, that is, in its thinking and the forces whereby released. Each thought radiates out as a great force and strikes the other creatures who attract these like magnets. According to their quality, negative or positive or balanced, they generate in the receiving creatures kindred forces, under according effects… But by the heresies and other misguidance of religions, the Earth human being is mainly only able to unfold forces which destructively influence all forms of life. Such destructions are found in the injury to age, that is, the lowering of the average age.
>
> The greater the assembly of human beings, as for example in cities, the lower sinks the average age, because in such places immeasurable thought energy is released, and penetrates into the fellow creatures… But also all materiality suckles itself up on these destroying forces, and in this way falls to early destruction, as for example iron, which normally does not rust so fast as is the case on Earth. By this, each Earth human being is himself guilty for the early death of his fellow creature, consciously or unconsciously, when he thinks in uncontrolled form, and in this way releases uncontrollable destructive forces. If the Earth human being wants to think right and according to the Creation, then he must guide the course of his thinking neither in negative, nor in positive trails, because both are degenerations which create bad consequences, thus they are against the Creation. Right thinking declares itself by a balanced and natural manner of thinking, in which way life becomes procreated and maintained.[10]

These penetrating lessons generate a heightening of the spiritual resonance of those humans who can receive the lesson and attune their thoughts, feelings and actions accordingly. The voluntary effects of teachings of truth was precisely given by Semjase during the 91st contact of November 17, 1977:

> *Semjase:* …In result of your education and your knowledge, it is you, [Eduard]… who emits the highest radiations on the Earth… When for example exists a lesson of the truth, then its radiation diffuses very quickly and very intensively. If now other living creatures find themselves the truth within this lesson, by which fact they themselves strive towards the truth, so they become liberated for the reception of the radiation [from the speaker of truth]… In consequence of this, the receivers of the radiation tune themselves inside of the subconscious for elaborating the lesson of truth towards its best inside of themselves, whereby they generate automatically a relationship of themselves to the diffuser of the lessons and even to his place of stay.[11]

The highly developed thought forces of Eduard Meier (above, 2010) were briefly demonstrated by focused intent in 1977, when he levitated a ½ ton oven across a short distance of several meters, and later that year demonstrated the heating and melting of several coins in his hands.[12] The real danger of *oscillation interference* was concisely addressed by Quetzal during the 97th contact of December 28, 1977:

> *Quetzal:* The reason why we do not maintain connections in physical form with you Earth human beings, is as follows: The spheres of oscillations between the Earth human being and ourselves is basically so much different, expressed in forms of negative and positive, as of the balances, that it could have heavy consequences, if they would touch one another. Normally those oscillations of the human being reach up to 90 meters [or 295'], which is why care has to be taken, this distance does not fall below that, thus that Earth human beings shall not come within this point of distance to us. Our oscillations are very highly sensitive, and react strongly to other oscillations, which penetrate towards our sphere of oscillations. In comparison to the Earth human being, we are standing very high within our total development, namely by about 3,500 years in advance of the Earth human being, for which reason as well our complete field of oscillations has become subtilized. When now oscillations of a much less sensible Earth human being would hit us, just that way, that Earth human and thus quite imbalanced and negative oscillations would penetrate into the field of our oscillations, then this would be equal to a very strong shocking of the structure of oscillations, resulting in a strong outer influence leading to a sudden uncontrolledness of reactions, thoughts and actions of ourselves, which at the same time release uncontrolled feelings of fear…

> Normally we use, when we walk among Earth human beings, which doing yet is extremely rare, an instrument, which protects us from the lower Earth human radiations. But to do this was never necessary with you, because you are exempt from these oscillations because of all your knowledge and your balance, which equates with ours… These dangerous disturbances of oscillation structure are the main reason that we are not allowed to agree to physical contacts with the Earth human beings. Even in our telepathic, or just inspirational-impulsive contacts, we have to exercise extreme caution, as also reflected impulse radiations effect similar consequences.[13]

Complete absorbance of radiated energies by the resonant metals of Plejaren spacecraft protects the visitors from the destructive thought forces of terrestrial humans, as with protective gear worn when outside the craft. Without the technical mitigation of radiant energies, safe contact is nearly impossible. The HHO plasma screens of Plejaren beamships permit the viewing of invisible infrared atmospheric energy as 'ion plumes' like those recently reported over North America during geomagnetic storms.[14] An excerpt from the 38th contact of November 13, 1975 concerns "energy bales… that circle the Earth at different distances and intervals" – *according to the Fibonacci structure of infrasound standing waves:*

> [Semjase explained that] through the dimension door it was possible to visit the Earth in its prehistoric past as well as in its later time when it has a thick white vapor envelope surrounding the globe… But what now about the dimension portal in the Devil's Triangle? Can anyone *see* it?
>
> *Semjase:* With the eyes alone, it is not recognizable, but it is possible to make the [infrared] radiation visible… with the help of my ship. At present only two courses of radiation are visible from this side of the Earth, while the third one is behind the Earth and thus cannot be seen from here… [Not only can you see it on the viewscreen, but] you can also see it from the windows, because the radiation becoming visible results from the radiation shield, which is [HHO plasma] spread from the whole ship. Now watch carefully, for the Earth…
>
> *Meier:* …Girl, …that is fantastic – I can see really two huge trails of radiation! Besides this, there are spheres and other things around the Earth. What do these mean?
>
> *Semjase:* Those are energy bales of different sorts and strengths. They circle the Earth at different distances and intervals, but are very important for the maintenance of Earth creatures. I am not allowed to explain more.[15]

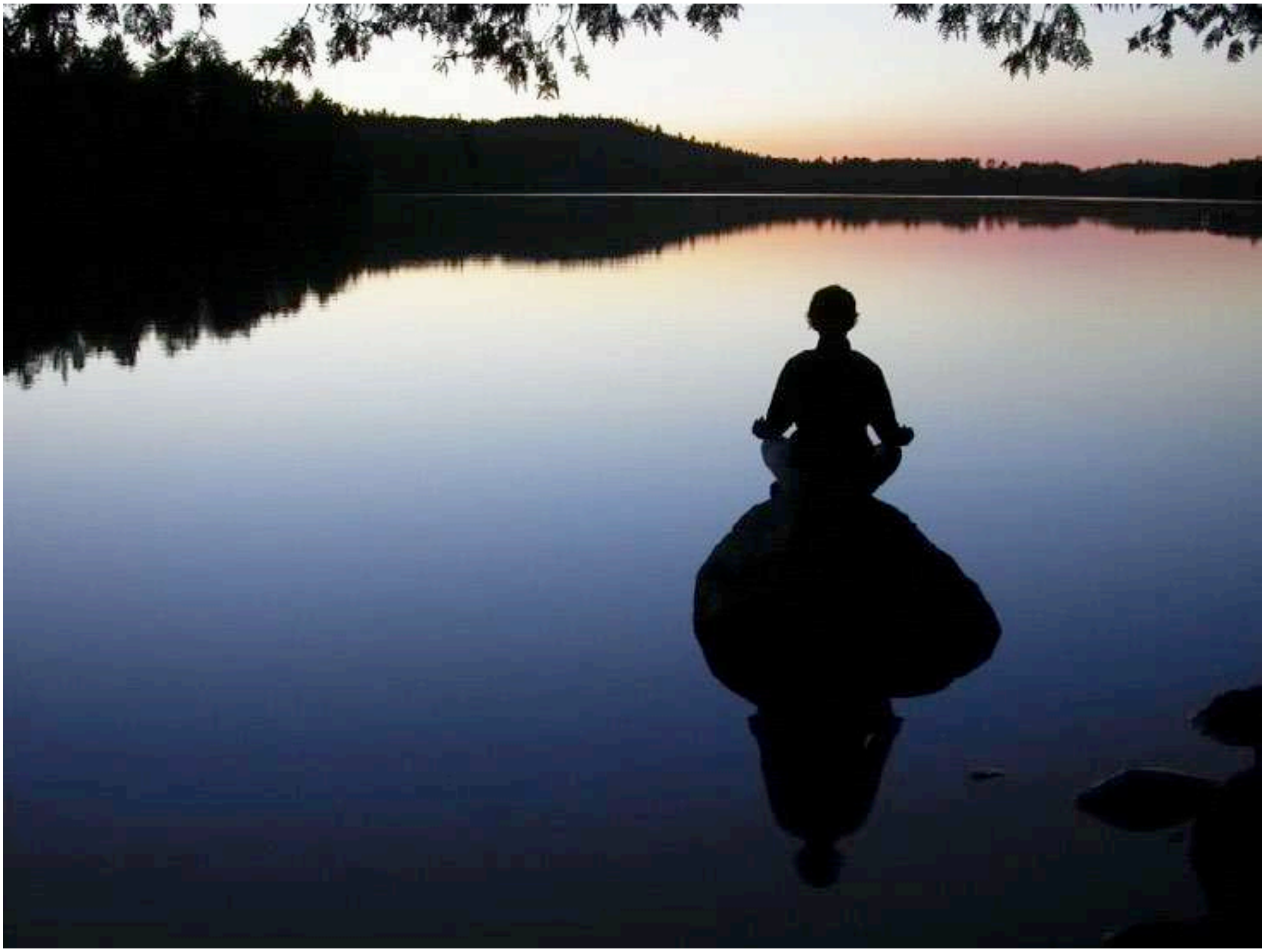

The Fibonacci standing waves that comprise Earth's resonant energy cocoon are artificially generated for the acoustic propulsion of spacecraft. A beamships' alteration of the normal infrasound environment of Earth likewise alters the consciousness of human beings, potentially undermining natural learning cycles:

> *Semjase:* …[C]oncerning control of the gravitational forces… natural cosmic laws order a unitary and harmonic development in compass to all concerns of all levels able for evolution. This means that the spiritual as well as the material development proceed constantly in harmony with one another, and always complete one another by unison of the consequence in all matters. In the case of the mastering of gravity, the regulations of the laws are arranged in order, that primary is the spiritual evolution, of most decisive meaning, for the elaboration of technical means with the aim of mastering the gravitational forces. Here normally at first appears the exercise of spiritual forces in paralyzation of the forces of gravity, by which a human being is able to move weightless through space… [G]enerated by spiritual force state… a levitation process takes place. From the very primary side, this status is evoked within an unknown to the human being form, and just step-by-step, he learns to master it and to control it suitably. When the human being has then learned to control these forces, and only then, the way is cleared for him, by this, to consider technical means and to build them. So the control of gravity consists in its first-primary form on the evolution of spirit and feelings, meaning that the told forces can physically be controlled just then when the human creature has delivered himself from the "weight" of the non-spiritual condition…
>
> Exceptions witness the standard, even if these exceptions cannot easily be recognized and proceed in a way which often does not come to appear. Such an exception does the Earth human being represent. Except for a few, he is still not able to master levitation, and nevertheless he is already able to see clearly a great amount in the field of gravity research. This exception consists of the confusing fact of a very rare misdevelopment, which basically has shifted the potential of parallelization of the spiritual and the material, by which the material

> [aspect of awareness] has become overweighted and dominates the spiritual. So this means that the Earth human being has become governed by the material side, by which the evolution runs in this direction, while the spiritual steps badly injured behind, as you yourself have once said. From that it appears that the technologies are more developed than the spiritual evolution. In this it has to be considered that especially since the year 1937 the strong influence of the new cosmic age have led this wrong burden of the material side in the whole world toward more gentle and harmonized courses, by which the differences between spiritual plenty and the material is no longer as strong as is generally thought.
>
> Many forms of spiritual direction have led to spiritual recognitions of enormous meaning and value, which decisively appear in the development of technologies, and as an interesting fact, lead to success although the normal sequence of observance of the law was disturbed. But this can also be explained by, the otherwise newly gathered spiritual cognitions do nearly compensate for the lawful rule, which means, that in spite of barbarism the Earth human being has proceeded in both directions since 1937, and gets closer, though often unconsciously about this, *and advances by the cosmic oscillations*, rapidly towards the real evolution and begins to comply with it. This course of development can already be seen in the first days of entering the Waterman Age, in the 3rd of February 1844, and during the following time until the year 1937, when since the 3rd of February the second half of transgression of time began to offer its powerful working, from which it can be said by great probability, that the temporal aim of Earth humanity in respect to his determination by evolution, will have been reached after a time of about 800 years.[16]

This concise explanation concerning human control of the gravitational forces given by Semjase contains many profound truths, which reach to the core of the present predicament of our Earthly humankind. The Plejaren comprehension of the Creational Laws reveals the natural evolutionary process whereby each civilization overcomes all barriers to interstellar space-travel: from the cognition of an internal spiritual process of levitation involving resonant infrasound standing waves and water crystallization. The ancient Vedic spiritual practice of levitation was referred to as one of the great attainments. Known as *'laghima'* in Sanskrit, it is understood as an internal electromagnetic force generated by infrasound standing waves that are transduced and focused by the piezoelectric human body into an enveloping antigravity field. Earth's present-day scientists have learned to overcome the gravitational force using acoustic levitation, as an effect generated by the focusing of high-frequency acoustic waves into a concave resonator. Biophysicists have demonstrated the levitation of various insects, and even gone so far as to successfully complete the gestation cycle of a frog egg entirely within a levitating water droplet! [17]

Just as concave resonators artificially generate the force of *laghima* in a beetle, an artificial gravity field is generated by Plejaren beamships for their levitation by acoustic resonance. While high-frequency oscillations are used to levitate beetles in laboratory experiments (at left), interstellar spacecraft apply ultra-low frequencies for the acoustic generation of gravity. When describing the rebalancing of the evolutionary growth of spirit with the growth of technical means, *Semjase identifies changes in 'cosmic oscillations' as the underlying cause*. The movement of our entire solar system into a highly energetic region of our galaxy is presently inducing the profound advancement in consciousness due to the relative increase of planetary infrasound resonance, effecting a dramatic rebalancing of the Earth's very sensitive vibratory envelope. These cosmic oscillations are transduced on Earth by the world's network of pyramids into a resonant field of infrasound standing waves that converge at focal points defined by the axis-symmetrical octagonal geometry of the Great Pyramid of Giza, Egypt. The Fibonacci-ordered distribution pattern of resonant infrasound focal points dictates the precise locations where human consciousness becomes enhanced by the heightened 'cosmic oscillations' increasingly inundating the Earth since 1844, according to information provided to Meier.

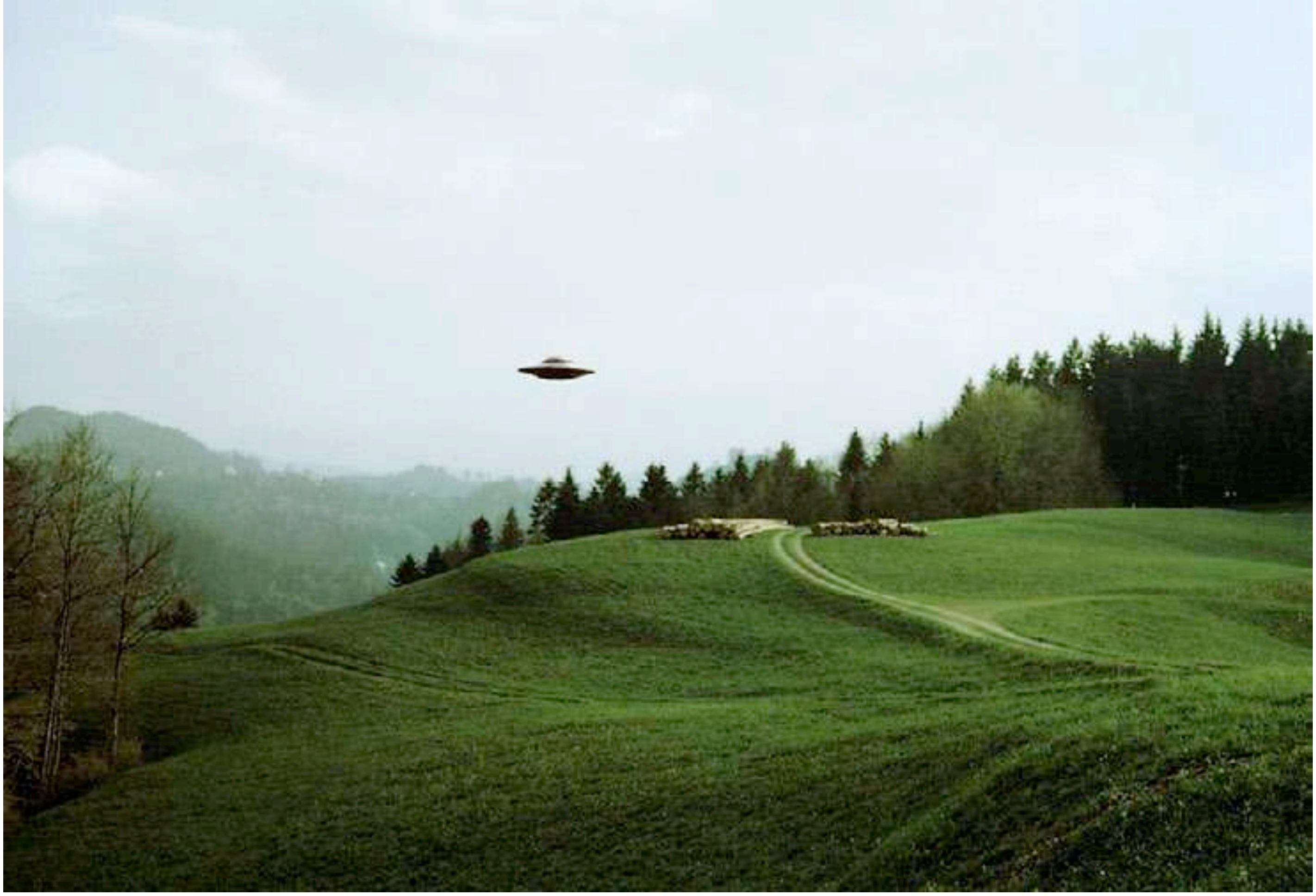

Several of Billy's contacts occurred in Ober-Sätelleg, Switzerland, where he was able to photograph Semjase's beamship on March 8, 1975 (above). At this same location on July 7, 1980 Meier was simply instructed to record the acoustic resonance of Semjase's beamship during a demonstration flight.[18] Meier took up a position below the low-hovering beamship, near the trees beyond the log piles, while his wife Kalliope, two of his children and several close friends also made recordings on the greenway about 488m away. *Thorough analyses by acoustics experts confirm the beamship sounds to be genuine recordings that are impossible to replicate using any known equipment, being truly unique in three significant aspects:*

1. There were at least thirty or more discrete frequencies in a random and constantly shifting mix that ranged from 4 to 2170 Hz, but varied on average between 470 and 1452 Hz.
2. The amplitude of these frequencies was also constantly changing, whereby the dominance alternated.
3. The wave shape was also constantly changing in a random, periodic rhythm that caused a characteristic beat.

The wave pattern in the oscilloscope showed this constant and random shift in frequency, in which the principal waves of all frequencies came together in perfect synchronization at one moment, only to travel at the next moment in different directions and stages, thereby generating different patterns— at one moment seemingly moving in a cluster in one direction and, at the next moment, in the other direction. Then they gradually expanded until, for one moment, they formed a mutually precise and evenly distributed pattern, only then to move again into different relationships. Although these changes appeared to be random and were not repeated in a particular order, they always appeared in geometric relationship to one another. Two other sound engineers and a synthesizer sound specialist joined the analysis team, and the sounds were reexamined, this time for possible duplication. All of these specialists agreed that the character of the sounds was unique and that any type of synthesis, if in fact such was possible, could produce only portions of the recordings we had examined and that duplicating only part of the sounds, even in a short linear segment, would be impossible. The number of traveling and constantly shifting discrete frequencies and constantly changing amplitudes, which were shifting in relative dominance, exhibited duplication problems that exceed the abilities of a current state-of-the-art device!"

> Normally, Meier does not hear such loud ship noises, certainly not for such a long time. At the most, there is usually a very short noise when the spaceship lands or takes off. This demonstration was intentionally given for the purpose of the tape recording… None of the other eyewitnesses saw any type of spaceship, but Meier said that he could see it from below and observe a strange effect. *As the sound went up on the scale, the ship became more transparent, and when the pitch became lower, it looked denser again*…
>
> Jim Dilettoso, our research consultant, took samples of this noise to Peter Gimer and Rick Coupland of Micor in San Francisco. There they performed tests in the audible range, from 20 Hz to 2000 Hz. They found twenty-four characteristic frequencies within the audible range and eight outside of it. They found all thirty-two frequencies concurrently at different amplitudes and volumes. All thirty-two tones are somehow produced simultaneously. In a time matrix, the amplitude of some frequencies increases, while that of others decreases. It was observed that the amplitude periodically increased by ca. fifty decibels and then decreased by ca. forty, and at other times, just the opposite was the case, which produced the characteristic beat that was audible. Other normal sounds were audible on the recording, but there were no signs of a tone-on-tone tape dubbing. All frequencies were clear and stable, and they were regularly lined up along the frequency scale [in the sonogram, below].

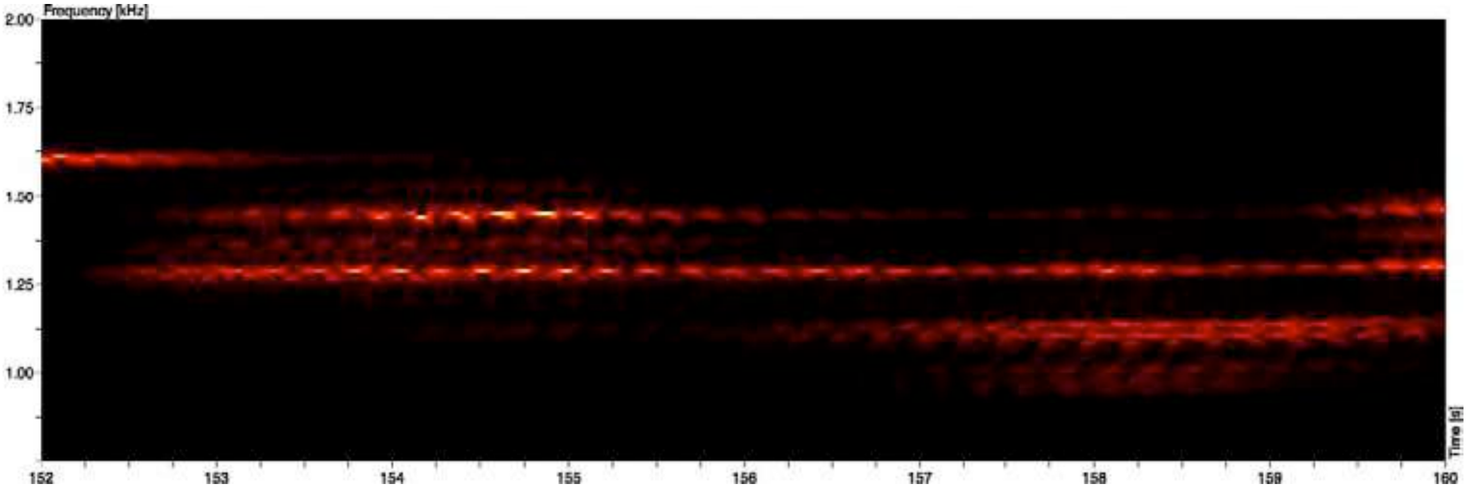

> The demodulation showed that one must be dealing with a rotating device — 249.6 U/min. modulated at 4.16 Hz… The high-speed device produced a sound that began at 520 Hz and increased in steps up to 990 Hz, only then to decrease again down to 520 Hz in the same frequency steps. The tonal groups comprising 520 Hz disappeared together at 520 Hz and reappeared at 600 Hz, only to disappear again at 720 Hz. At 990 Hz, however, they reappeared and were very strong. The same occurred in the other direction. The condition remained stable for a moment and then shifted again. This shift was somehow random, but nevertheless constant. At certain times, the high-speed device was still, and a deep, throbbing beat was audible; and then the high-speed rotation began again, and the sound increased rapidly to a high vibrating crescendo in the upper 50,000 U/min. range or higher. Simultaneously, the slower 249.6 U/min.-rotation again became audible. The vibration was produced by the high-speed rotation of approximately 29,000 U/min., which accelerated to over 59,000 U/min. This was most certainly not a normal sound! [19]

While acoustic experts have identified several key features of the beamship sounds that are impossible to reproduce with any known equipment, their archive search for related sounds failed to recognize these same key features in a sound that we are all constantly exposed to, but only very rarely conscious of – *the sound of our mother Earth!* The over 30 fixed frequencies of Semjase's beamship, whirring in constantly shifting amplitudes, very closely replicate the much lower infrasound harmonies of Earth's planetary resonance. Earth's base frequency presently fluctuates near 7.8 Hz and harmonics progress in frequency steps of approximately 5.9 Hz, extending from below 2 Hz upwards to about 50 Hz.[20]

Subtle fluctuations observed in the deep resonance of Earth are seen as rather dramatic shifts in the multi-frequency oscillations of the Plejaren beamship, as dominance constantly shifts up and down the frequency scale in regular stepped intervals visible in the sonogram above. The various circular rings and domed elements of the beamships' hulls are designed to generate this standing wave resonance through highly structured crystalline metals that transduce and amplify the omnipresent song of the cosmos, referred to by the Plejaren as the 'Symphony of the Spheres' during the 223rd contact of May 1, 1988:

Meier: You, as well as Ptaah and Semjase, once played the strange humming sounds of the Earth to me through your apparatuses, which I can't perceive in nature, unfortunately, because my hearing senses aren't so sensitive that I could hear these sounds, which should very well be possible for other people, however, as you've explained… You named the sounds "Symphony of the Earth" or "Symphony of the Planets" as well as "Symphony of Space," which is also entirely incorporated into these sounds, like also the atmospheres of planets and suns and even all material and immaterial things of every kind, concerning which you've spoken of a "Symphony of the Spheres." This symphony, i.e. these sounds, are, so far, still unknown to the Earthly scientists in particular and to the people in general, other than just the exceptions to the rule, even those people who perceive these humming sounds, i.e. "Symphony of the Earth," and who are of the opinion that something isn't right in their heads, and thus, not that real things are going to and fro. Can you once again officially explain what it actually concerns with these humming sounds, if it shouldn't remain a secret?

Quetzal: It does not have to remain a secret because in about 10 to 15 years [1998-2002], a portion of the Earthly sciences should also encounter these things. Moreover, it concerns the fact that every planet, every comet, every larger asteroid, meteor, and every sun, as well as the nebulae and clouds in space and space itself, as well as every black hole and all material and immaterial things existing at all produce certain tones and sounds, which we, as you said, actually call Symphony. Within the entire realm of the Creation or of the Universe, nothing exists that would not be incorporated into these symphonies, so therefore, all material and immaterial things vibrationally generate certain tones and sounds and even symphonies, which usually cannot be perceived by the human ear because these move in too low or too high frequencies for this. Only through abnormal, degenerative, physiological changes can these sounds or symphonies be perceived by human life forms, whereby these then become effective in an extremely consciousness-impairing, nerve-racking, and psyche-damaging manner.

> The frequencies of these symphonies or sounds reach into all levels, so thus, the lowest sounds, like humming sounds, as you say, are given as well as the highest whirring sounds and shimmering sounds, etc. As the frequencies are virtually countless in their number, these are also the sources of the sound symphonies, for these relate to all that exists of the material and immaterial. Thus, if one starts from the Earth, then these sound sources are, for example, the waters, in terms of the deep-sea movements as well as the movements of the waters of the lakes, rivers, streams, creeks, and springs.

> Even the roar of the waves, the small and mighty storms, the tectonic movements of the Earth's interior, the bubbling of magma in the Earth, and the eruptions of the volcanoes belong to these, as well as the rain, hail, and the various layers of the atmosphere, which rub against each other, as well as the winds against the trees, grasses, flowers, houses, mountains, sand dunes and deserts, steppes, icebergs, glaciers, and surfaces of ice, as well as against windmills, wind turbines, ships, aircraft, rockets, missiles, and vehicles of all types, against land surfaces and everything else that materially exists. Also to be noted are the voices and sounds of humans and animals, insects, birds, and fish and any other water creatures.
>
> All sounds together form the actual "Symphony of the Planet," as we call it. And as this happens in the purely material realm, so this also happens in any immaterial sphere in the entire Universe, each of which also produces its own symphony, like also fog flowing on the Earth or nebulae existing in space and all other formations that are absolutely impossible to cite in their number because everything is so immeasurable that it could neither be grasped by purely human understanding nor by technological possibilities of any kind.[21]

Given as a precise description of cosmic resonance, the comprehensive explanation of the 'Symphony of the Spheres' provided by Quetzal penetrates to the unified vibratory essence of the cosmos. My own direct perception of the resonant symphony during deep meditative states formed the impetus for my years of study of infrasound standing waves. I discerned that the deep throbbing pulsation that fills my consciousness during the darkest hours of the night was induced by ayurvedic pyramids for heartbeat entrainment at specific nodal points around the globe where sacred ancient temples still exist, yet having suffered terrible cataclysm. The cosmic symphony is absorbed through the resonant metals of plasma beamships to be converted by interior HHO plasma chambers into ultraviolet-A and infrared light for all power needs. The over 30 distinct frequencies observed in the infrasonic bands of Earth's resonant symphony are exactly replicated by spacecraft for acoustic levitation on the arches of standing waves.

Analyses of the resonant humming of the Plejaren beamships recorded by Billy Meier exactly echo hieroglyphic Sanskrit scripts on highly reflective silver-alloy metal debris recovered from the 1947 aerial disc crash near Roswell, New Mexico (pg. 31). Embossed text on an I-beam fragment reads: *"workings of 30 resonances, female and male, 30 resonances whereby approaching the highest…"* Meier's beamship recordings prove *the Roswell texts describe the 30 resonant frequencies of standing waves generated by plasma ships for both antigravitic propulsion and the maintenance of life onboard.*

Standing waves surround the beamship to act as waveguides for HHO gas that is emitted from the aerial antenna atop the craft into the energy streams, where it is ignited to form HHO plasma around the ship. Also called hydrino plasma, HHO plasma is defined as *resonant-transfer plasma* for its crystalline ability to transduce all received acoustic energy into re-emitted heat and light. All interstellar vessels achieve extreme acceleration by enveloping the spacecraft in HHO plasma to induce a set of transdimensional characteristics that includes superfluidity and frictionless passage through the atmosphere. Semjase explained in basic terms that the screen "allows every interference to glide away, without… resistance." [22] This same plasma technology has already been in use for many years now in the nose cones of US military aircraft.[23] Additional information was provided to Meier regarding the use of plasma 'distorter-screens' enveloping Plejaren beamships that allow complete optical invisibility to the naked eye:

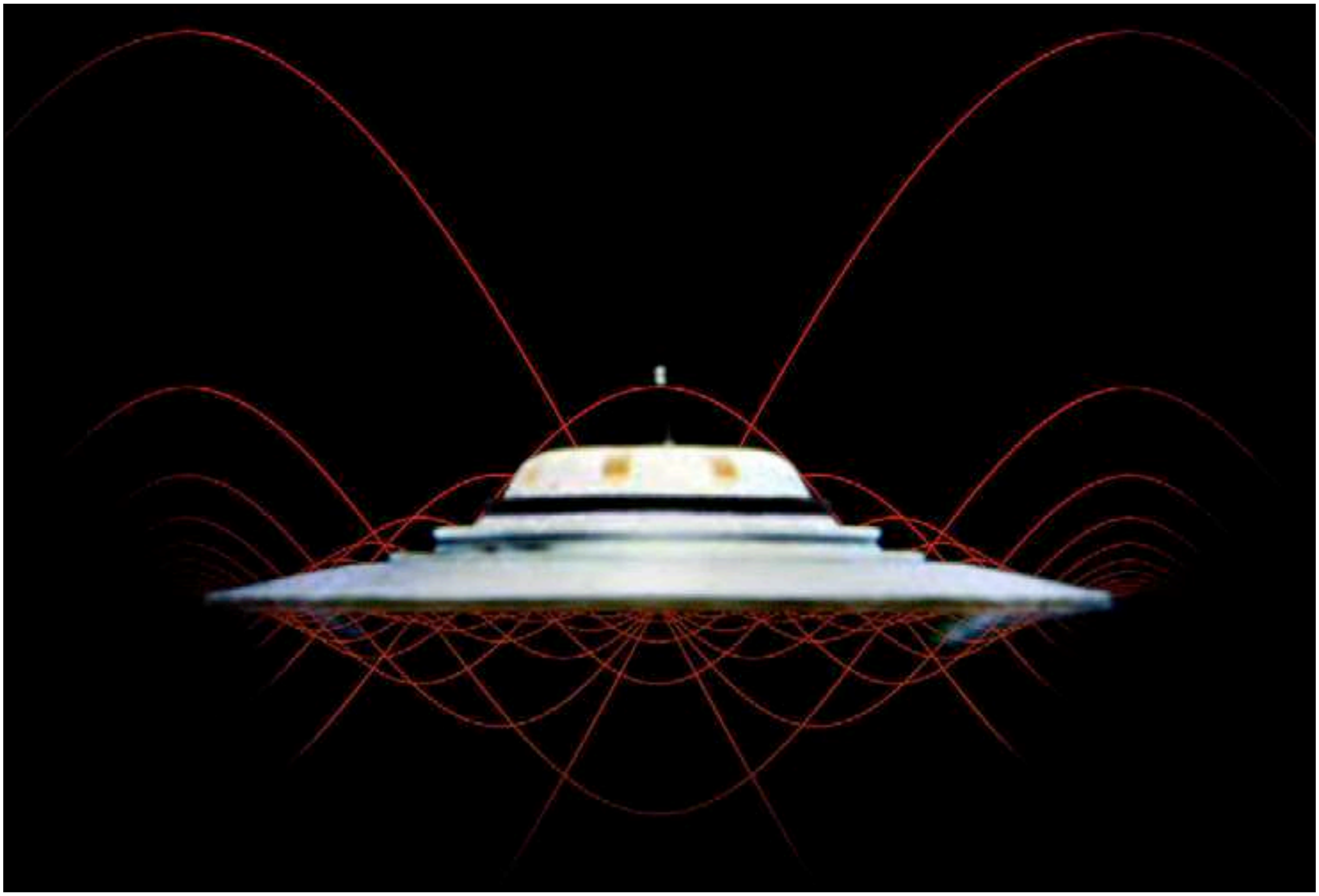

> [Earth human beings] are very simply not accustomed to close observing and viewing. On the other hand we protect our beamship against sight of any kind and they can't see us. It is easy for us to protect our ships as we desire in a diameter of 500 meters, in whole or on parts, at one side, upwards or downwards, against sight by (use of) a distorter-screen. When I allow you to photograph my ship from one side, then it is protected on all other sides against further observation by the distorter-screens, thus none uninvited are able to see it.[24]

The beamships' invisibility screen is generated from emitted HHO gas produced by the resonant dissociation of water molecules at 13,563,688 Hz. The implosion of single hydrogen atoms (to become *hydrinos)* as HHO plasma[25] is achieved by the rapid ionization of emitted gold and silver nanoparticles, as indirectly described by Semjase during her 11th contact with Eduard on April 15, 1975: "The antenna's conducting beam and the energy collecting beam appear as a fine line of energy over the beamship, while the regeneration radiation becomes visible beneath the ship in various hues… The energy only becomes used and regenerated again for the radiation drive… Our whole technology is aligned toward natural rectification, but not toward destruction." *Implosion* was then addressed by Meier, "I understand; the natural principle is, thus, implosion and regeneration, rather than explosion and destruction." [26]

Meier observed an optical transparency effect of the gradual acoustic engagement of the ship's HHO plasma protection screen, observing that *"as the sound went up on the scale, the ship became more transparent, and when the pitch became lower, it looked denser again…"* A different transparency effect was mentioned on another occasion when, during the 31st contact of July 17, 1975, Semjase explained: "You know that with our technology we have the ability to strain any materials by special radiation and cause those materials to be invisible to the naked eye. We can control this effect very exactly." [27] This transparency beam technology was first created and described among our scientists by Canadian inventor Troy Hurtubise, who named his prototype beam device *Angel Light.*[28] An intense beam of X-ray light from krypton plasma generates the temporary effect of *total optical transparency in all materials* – a new state of matter to be technically described only later, in 2009, by Oxford University researchers:

> [A] short pulse from a laser "knocked out" a core electron from every aluminum atom in a sample without disrupting the metal's crystalline structure… [The] high-powered synchrotron radiation generator produces extremely brief pulses of soft X-ray light… At such high intensities the aluminum turned transparent.[29]

Plejaren beamships' precise control of transparency induction using X-ray light beams was demonstrated for Meier on objects in Earth orbit and even inside the beamships themselves. The domed ceiling and 8 view-ports of the circular spacecraft are made transparent by this high-intensity illumination process, giving the optical effect of disappearing instantaneously to provide an unimpeded 360° panoramic view.

Meier's clear profile photograph of a landed beamship from 1975 reveals the structural features of the flight vehicle perfectly conform to the Fibonacci-ordered standing wave resonance patterns generated by the over 30 discreet frequencies resonating within the alloys of the hull during flight (opposite). During the 4th contact of February 15, 1975, Semjase offered a comprehensive description of beamship propulsion drives, yet carefully substituted the word 'beam' for the more accurate scientific term 'HHO plasma', a technical concept apparently not authorized for use according to the Confederation's High Council:

Semjase: For traveling through cosmic space, a drive is necessary which surpasses the speed of light by millions of times. But this propulsion can only come into action once the speed of light has already been reached… This means then that a beamship needs two drives, first the normal which accelerates up to the speed of light, and then a second… which generates a million-fold, a billion-fold, the speed of light, thus the hyperspeed by which hyperspace is penetrated… So time and space collapse, and they become zero time and zero space. This means time and space cease to exist. And exactly by that, distances of numerous light years are traversed in a fraction of a second, with no time lag… When, for example, we leave the Pleiades and need about 7 hours to get to Earth, then on our own planet, and on Earth, 7 hours pass. We need this long because we first have to fly beyond the reach of the planets by normal propulsion, and only far out in space can we convert to hyperspeed. Back from hyperspace then, far outside your solar system, we convert to normal drive. We are never allowed to enter hyperspace too near to a planet… I am only allowed to say this concerning propulsions; your scientists are already working on light-emitting drives. Light-emitting drives work for normal propulsion, and move the beamships near planets and until they are a safe distance away.

Meier: This is evident, but how is it possible for a beamship in the gravity of a planet, or in its atmosphere, to attain such great speed without glowing, or the passengers succumbing to the huge pressures?

Semjase: This is very easy to explain, and also no secret any more to Earth humans, at least not the scientists; the beamship is surrounded by a protection-beam-girdle… [Thereby] anything contacting the screen [becomes] disintegrated or “flowed” away… [while the protective HHO plasma screen] also neutralizes the attractive force of a planet… The gravity of a planet is, besides this, not always the same, or of the same strength, owing to certain alternations which will be discovered by your scientists in a reasonable time. The beam-protective-screen diverts the gravity and attractive forces, and the beamship in effect becomes a miniature planet which can travel at nearly light speed through any atmosphere without risk. As the gravity of a concerned planet does not influence the ship, the passengers feel normal and unburdened, as if they were on their planet itself, always under the premise that the planetary gravity is in accordance with their anatomical capabilities. In the beamship itself, the gravity of course is tuned to the passengers and is completely controllable. When passengers on spacecraft from other worlds move in atmospheres alien to them, or on hostile planets with unbearable gravity, they use suits and other small transportable instruments which generate for that creature, the necessary beam-protection-screen for their particular ship and their particular requirement.[30]

[Semjase continued her explanation of superluminal space travel during the 8th contact, on March 18, 1975:] The turning-off of a time-dilation or time-shift needs but the penetration of hyperspace, as I have already explained. The “jump” occurs very fast by momentary paralyzation of the protective screen under flash-like increased velocity, with the consequence of a flash-like increase in mass. This means that the initiation process runs so fast, that by the speed of certain processes, generated by the apparatus, matter is distorted within the millionth part of a second, and becomes fine-material form, which is able to pass hyperspace timelessly…

> By this, at the same time, space and time are paralyzed and disappear, in consequence of which the ship already rematerializes itself at its destination place, as it is dematerialized at its place of departure. The whole process needs no longer than a millionth part of a second, thus also for creatures, passing hyperspace does not take alterations of any kind. When spaceships fly below the speed of light, this inevitably takes much time, which is always the case with newcomers to cosmic flight. First, they all do learn from experiences and cognitions. Second, this space-traveling is very dangerous and leaves its purpose open to question. When spaceships reach hyperspeed without passing directly into hyperspace, then catastrophe for the ship and its passengers is assured… Speeds above that of light hold many dangers, when the barrier of hyperspace is not penetrated and made an ally. The dilation effect is only one of these dangers… Our forefathers, too, faced these problems, and went astray in space and time. The same does happen to other creatures, and in the same way.[31]

The industrial development of antigravity technology on Earth followed Nikola Tesla's 1899 discovery of standing waves, which led to the later development of the implosive hydrogen motor. Through the joint espionage efforts of Nazi industrialists in both the US and Germany, Tesla's secret transdimensional spacecraft designs were stolen by Hitler and quickly weaponized by enslaved Nazi/CIA scientists[32]:

> Viktor Schauberger was the actual designer and builder of this world's first flying saucer motor, which was built for Nazi Germany in 1944. In view of this startling revelation one has to wonder where he obtained his revolutionary knowledge to construct such a device. He had discovered what he called an "implosion" principle, and invented a motor to take advantage of this new principle. The motor consumed only air and water and generated light, heat, and thrust, as well as a diamagnetism which made lift possible through "diamagnetic levitation".[33]

The dark truth of the Nazi assassination of Tesla after the theft of his designs was retold in 1995, during a deathbed confession by Hitler's personal bodyguard and espionage agent Otto Skorzeny. Published by Erik Berman in 2003, Skorzeny confirmed that Tesla's HHO plasma invisibility technology had been applied in both the Nazi flying discs and later in the US stealth aircraft.[34] Advances in Earth's resonant spacecraft technologies continually occur as the expansive consciousness of humanity reaches toward the stars.

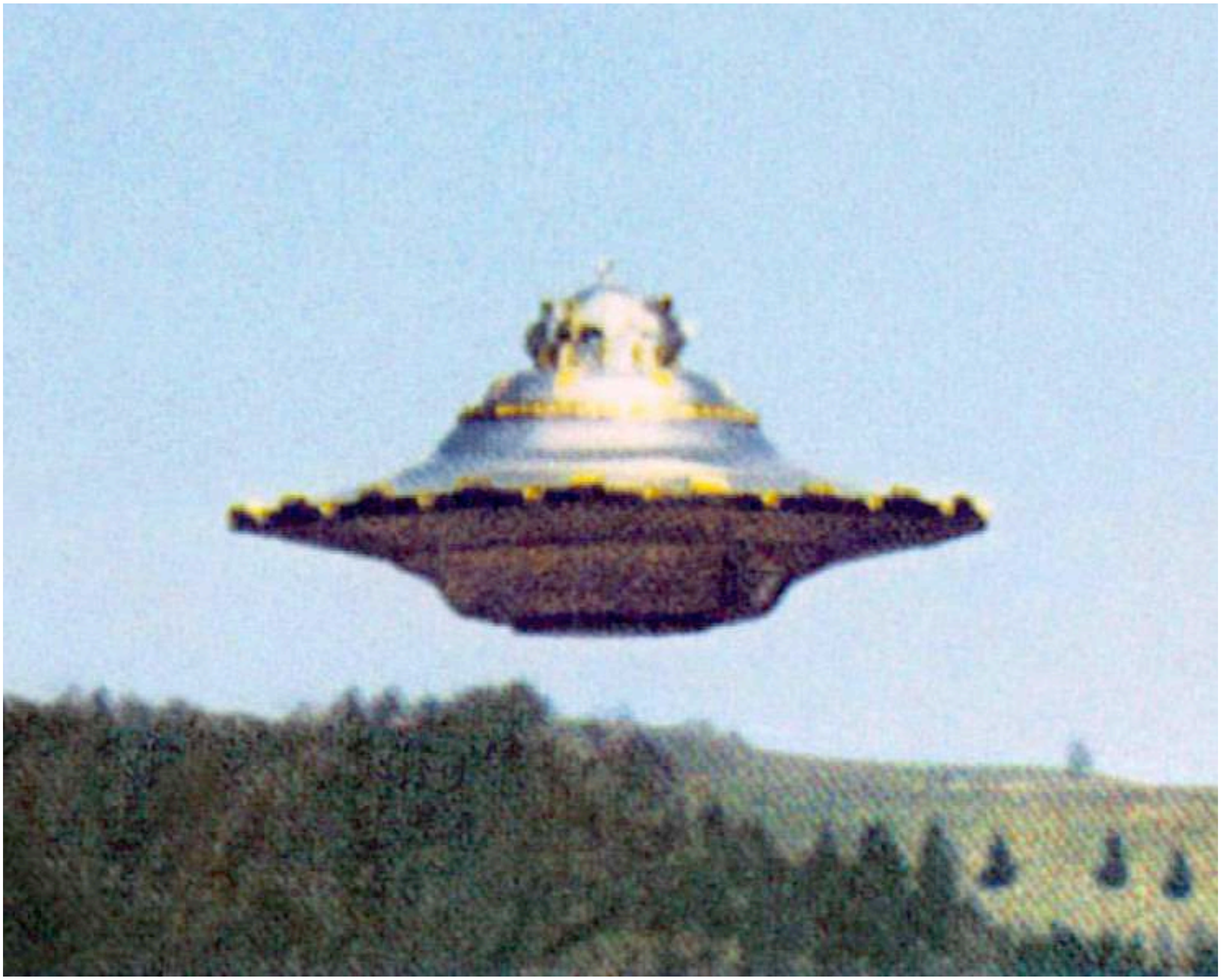

In anticipation of piloting a *newly manufactured* spacecraft from Erra to Earth for the first time, Semjase announced the deployment during the 48th contact. Eduard was given the opportunity to photograph the new beamship on March 8, 1976, at Bachtelhornli-Unterbachtel (above). The new beamship displays a golden-colored alloy comprising the underside of the hull and along a band below the orange-tinted cupola windows. Both beamship models are 7 meters in diameter and share a similar hull design, with several significant differences. The latest design presents a more complex array of HHO gas-venting systems above the domed ceiling that replaces the aerial antenna employed in the previous model, as well as radial ridges that form a distinctive stepped contour along the golden perimeter of the spacecraft.

Semjase explained, “In the outer appearance only small differences appear. The greater novelty is the inside arrangements, which offer much more capability than is the case with my present ship. With the new one I have the ability to break through dimensions in two directions.” [35] *Meier had previously described transdimensional hyperleaps through space and time during earlier contacts in his youth, as an instantaneous physical dematerialization from which arises a profound sense of cosmic Oneness*:

> With the launching of the [time-travel] transmission a quite weird thing happened in the form that the entire surroundings of the ship slowly shimmered and then quite suddenly simply no longer existed. I also observed the same process in regard to myself, and in the moment of the "jump" it was, to me, actually as if I suddenly no longer corporally existed. Somehow I felt transferred into something which I sensed as being eternity itself and in which an indescribable silence and endless, calming peace and enormous love ruled… It was often for me as if I were raised up by *being* itself.[36]

While photographing the latest Plejaren spacecraft design for the first time (below), Meier witnessed the instantaneous appearance of two accompanying spacecraft, later described as 4-meter Explorer class unmanned reconnaissance vessels. HHO plasma excited in the air surrounding the beamship for instant optical invisibility can also be generated within the ship's interior airspaces by rapidly ionizing the third-density lightwater vapor to become a fourth-density environment of HHO plasma. *During his 31st contact of July 17, 1975 Eduard Meier experienced fourth-density space-travel, by hyperleaps, covering vast distances to visit various inhabited worlds of the Universe.* The remarkable details reported by Meier deserve serious contemplation – especially the state of consciousness experienced by all the passengers aboard the giant spherical 'Great-Spacer' during the seven minutes of dematerialization *'in the eternity'*:

Semjase: Earth humans still don't know very many things. But now we have to interrupt our conversation, because my father has finished his preparations for the great leap. And as I am already speaking of such, then I want to ask you something: In a few minutes, we will jump for seven minutes into the "eternity", as you call it. *The feelings and sensations there are completely otherwise than in normal existence in material life.* For that reason it is also not possible for us to receive your thoughts and feelings with proper concentration. In consequence of which neither I nor my father, nor any of us, could later repeat for you your feelings and thoughts, so that you could write them down. If you nevertheless wish to do this, then there is a chance on a technical basis…

The booths, which you can see there, beside the screens, are equipped with all necessary means, to store up feelings and thoughts. The thought impulses received are stored in a special computer and can later, as may be desired, be repeated word for word from it. The helmet, formable and adaptable in size, which you see there, is equipped with very fine sondes and is as well covered by a special, fine-meshed net of sondes, which pick up every kind of energy and transforms it into impulses, which are then transmitted to the computer where they are registered and stored. The energy of thoughts and feelings is measured in very high values and can be received only through those instruments. The energies of feelings and thoughts exist only in very high frequency fields, or hyper-frequencies. To now be able to register your thoughts and feelings, it is necessary that you place yourself into the chair and lay the head under the helmet cap, which then will adapt itself automatically to your head…

Meier: At your service… (As Semjase explained to me, I quickly sit down inside one of the three booths, in the extremely comfortable chair. As soon as I have seated myself, the peculiar helmet moves over my head and sinks down silently. It is big enough that it closes around my whole head, and only leaves my face open, thus I can see and watch everything. But the helmet is not touching my head; only lying close around it, keeping about one and a half centimeters distance to the skull, as I can see when I place a forefinger between the helmet and the head. Now I am tense and expectant, for I wonder what is going to happen. Ptaah and Semjase manipulate the apparatus, and now I can see again how the fantastic heavens and stars change. In a fraction of a second they are nothing more than a whitish milky mass, a shining mass, as I have already seen in the other hyper-leaps. But now suddenly as well, this milky whitish shining is gone and there is darkness.

But now what is this? Suddenly all is merging into a golden color, and now everything is like silver. But - my dear this glistening light, this beaming shining splendor! Everything is merged into glistening light - only the glistening light. It is stronger than all the suns of the Universe... Dear, oh dear, this glistening light, and it does not hurt the eyes! Dear, this must be eternity, the glistening light of the eternal... but see, there is nothing besides the eternity; man alive, how marvelous! Marvelous? Man alive, that is itself marvelous. Eternity and marvelousness are one and the same. Only why do I separate it? Why do I put the eternity into terms of time? Time does not exist, and the eternity is marvelous. Man, just what is this? This tranquility, this peace -- what is it? How could I have achieved this?

Love, oh that deep all-encompassing love. Nothing is there, but *love*: wonderful, marvelous. Oh yes, I am, but I am not. Everything is so deep, and full of love. Of course, I am eternity, and I am inside of eternity. How could I ever forget this? Oh yes, I am a human being, how can I... why do I forget that? I am only a guest in eternity – and those loving voices, calling for me, from where is it coming? I can see nothing, only the glistening light, comforting. Who is calling me? I see nothing; man alive, I am nothing any more. I can't see myself. I am eternity, in the eternity. Oh, how is that, but I don't see with my eyes, still I see everything. And I am not listening with my ears, yet I hear everything. Yes, the love, how powerful it is, how immense, infinite and wonderful. Everything is love and splendor; why doesn't the human being understand this?... The calling comes out of the light. Is it the call of eternity?

Oh how hard it is... to not think that I am only a guest in eternity, that I am only a human being. How painful it is, this being a human. I do no more want to be. I want to remain here as eternity inside of eternity... this deep. Yes, I want to stay here. Never again do I want to return. Here is the existence, the real existence... How can I feel strange having to return to a material world? I belong here... Pain, why are you shaking me? What is it?)

Semjase: Do you still not understand?

Meier: I – Oh yes, of course! – Pitiful – very much pitiful – why have you brought me back here?

Semjase: You want to joke, but we have all been inside the timelessness, and would all like to stay there. But we are not ready to stay, because we have to absolve the way of evolution step-by-step – as you also must do, dear friend. I know quite well, how painfully your heart aches now, as we all have experienced the same thing. *But we have to adjust ourselves for that and do know, that under no circumstances or conditions are we allowed or able to remain, until we have reached that level of consciousness.* I do regret very much that you feel disturbed inside, as I can see it on your face. But you will overcome it, because you can think about it realistically. For that reason we also took responsibility for this equipment... [that allows us to record your blissful thoughts so that you may reflect on the experience of transcendence later.] [37]

The transdimensional experience of oneness had left Eduard shaken. What he experienced *'in the eternity'* finds direct parallels with the impressions of other contactees who have transcended into *fourth-density awareness.* During transdimensional breakthrough, the third-density air environment maintained inside the spacecraft is instantaneously converted into a fourth-density environment of HHO plasma, *inducing a transcendent state in the human being through the radiations of gold and silver nanospheres among the trace elements conforming the atmospheric water vapor maintained within all beamships:*

> *Ptaah:* You breathe here in this room… air… [that] is adjusted to the conditions of our homeworld, and is much more healthy and invigorating than that of Earth. The oxygen content is some values higher than on Earth, and also the compound of other elements is somewhat different… The different composition of the air keeps us awake longer, and in need of less sleep than is the case on Earth. That is why you are not tired.[38]

An elementally balanced atmospheric composition affects the bioelectrical enhancement of all cellular metabolic processes within the human body, greatly decreasing the need for physical sustenance, whereby Meier was sustained on one meal for over 90 hours. Open discussion of bionanotechnology was precluded by the Plejaren, except once –by Quetzal during the 205th contact of February 3, 1986:

> But in our civilizations, [disease] has no longer appeared for thousands of years… I want to explain as an exception and on my own responsibility. *Above all, silver ion products – such as silver ion powder, silver ion jelly, or silver ion cream – are useful.* For this, an application is required on the skin.[39]

This hint from Quetzal addressed those few silver ion products available at that time in 1986, yet point directly toward the present advances of ayurvedic electro-medicine using gold and silver nano-colloids.

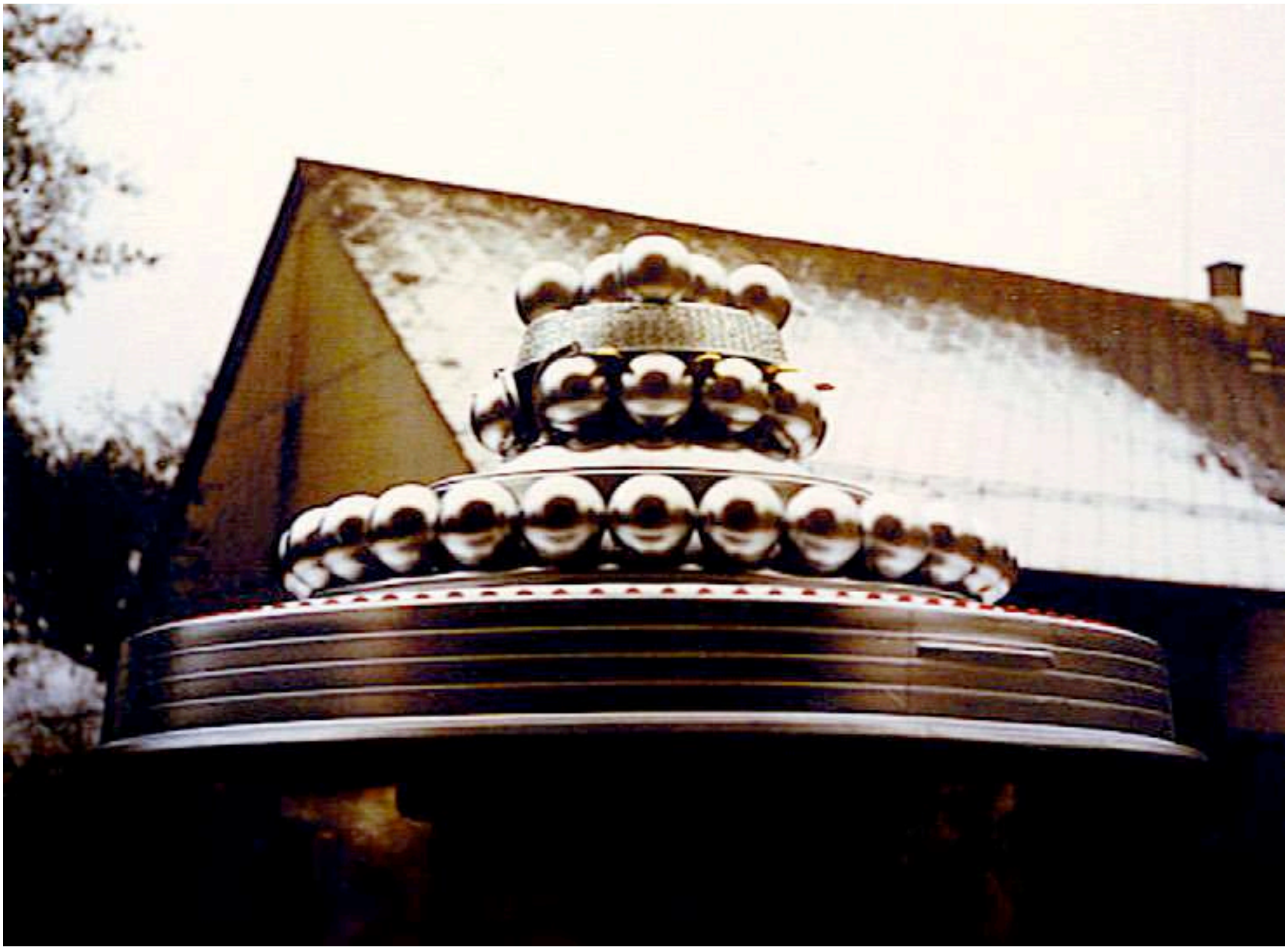

Through these highly advanced transdimensional means, Plejaren spacecraft are able to not only hyperleap through space, but also through time. Meier was taken to visit Earth in various stages of her past and future to witness differing levels of gravity and the corresponding atmospheric conditions. On one occasion, Eduard was actually taken thousands of years into the past aboard Semjase's beamship *to visit himself in a previous incarnation* and was informed that the same time-travel process even allows the potential for visiting one's future incarnations, although this is rarely done.

Throughout Eduard's early years of contacts with the Plejaren Elders he was denied the opportunity to obtain detailed close-up photographs of their spectacular plasma beamships, yet this prohibition was lifted during later contacts in 1981. Several opportunities were granted for Billy to document visits by a single-occupant 3.5-meter reconnaissance beamship (above). Having no exposed windows, this compact plasma ship relies upon interior view-screens or, for direct viewing, the walls of the ship are made transparent by intense X-ray illumination. The many conjoined circular sections of the tiered hull are comprised of a highly reflective electroluminescent metal alloy that can be electronically tuned to emit any desired frequency of light. Infrared luminescence is used to ignite HHO gas for instantaneous plasma invisibility cloaking, referred to by the Plejaren as a 'beamship protection screen'.

Scores of superbly detailed documentary photographs of transdimensional spacecraft were taken by Meier in the parking lot of the Semjase Silver Star Center. Close comparison of the differences that appear in the configuration of Quetzal's ship reveals controlled rotation of the large ring of spheres. Meier learned of the absorptive characteristics of the beamships and other similar interstellar spacecraft during the 150th contact of October 10, 1981, in discussion with Quetzal regarding technical barriers to communication between various civilizations operating plasma spacecraft in the vicinity of Earth:

> No radio waves and no other kinds of vibrations are capable of penetrating into these aircraft because the outer hulls of each device absorb everything in terms of power storage for the ships' propulsion and everything else. This means among other things that these aircraft are so designed that they absorb all waves of all frequencies and, therefore, all vibrations of all kinds for power storage, like also sunlight and everything else in terms of entering radiations, etc. But the moon light and the weak light of the stars are also absorbed and stored as power, together with many other things [comprising the Symphony of the Planet, *especially terrestrial standing waves of infrasound resonance*], which are still unknown to the Earthly scientists.[40]

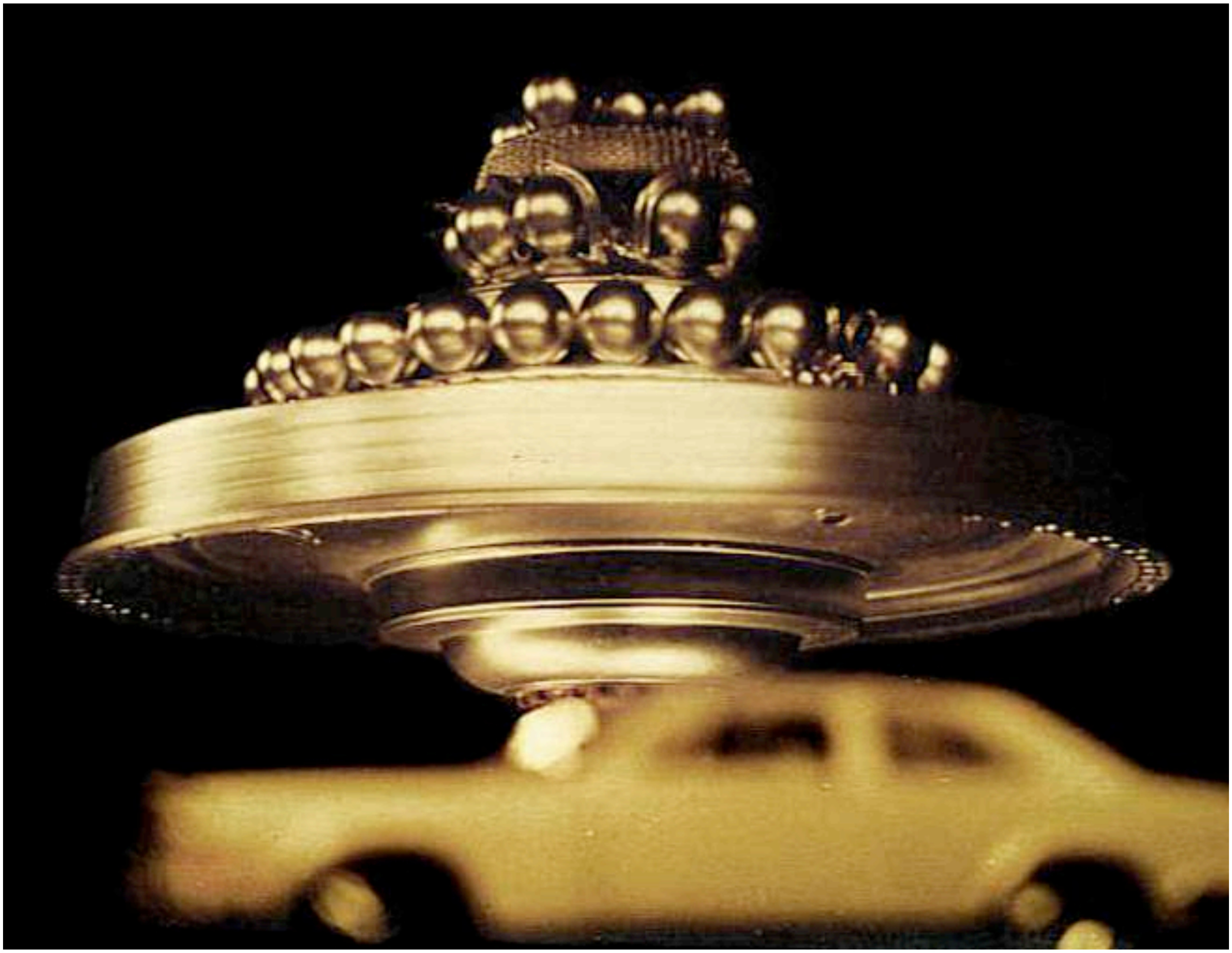

All wavelengths of energy are absorbed into the craft and transduced into ultraviolet-A and infrared light inside HHO plasma chambers, formed by the conjoined rings of spherical resonators. The rotating rings release plasma discharges that arc across the spark gap created between the two resonant spheres with gold-colored circular fittings. The noticeable warping of the image is an optical effect of the strong electromagnetic fields levitating the craft across the parking lot while it was being photographed, without landing. Along the outermost rim of the craft is a ring of 96 rubies that act as plasma beam emitters as HHO plasma becomes confined within the path of the infrared light by an encircling electromagnetic field, just as recently defined by MIT researchers investigating *superfluidity* (Ketterle et al.[41]).

The single-occupant reconnaissance spacecraft displays many advanced features for the application of fourth-density HHO plasma beam technologies that further confirm the validity of these photographs, which have been met with decades of sustained criticism despite being authenticated by experts. Many magnificent photographs taken at the Semjase Silver Star Center provide scale references by including foreground and background objects, allowing the size of the spacecraft to be accurately determined.

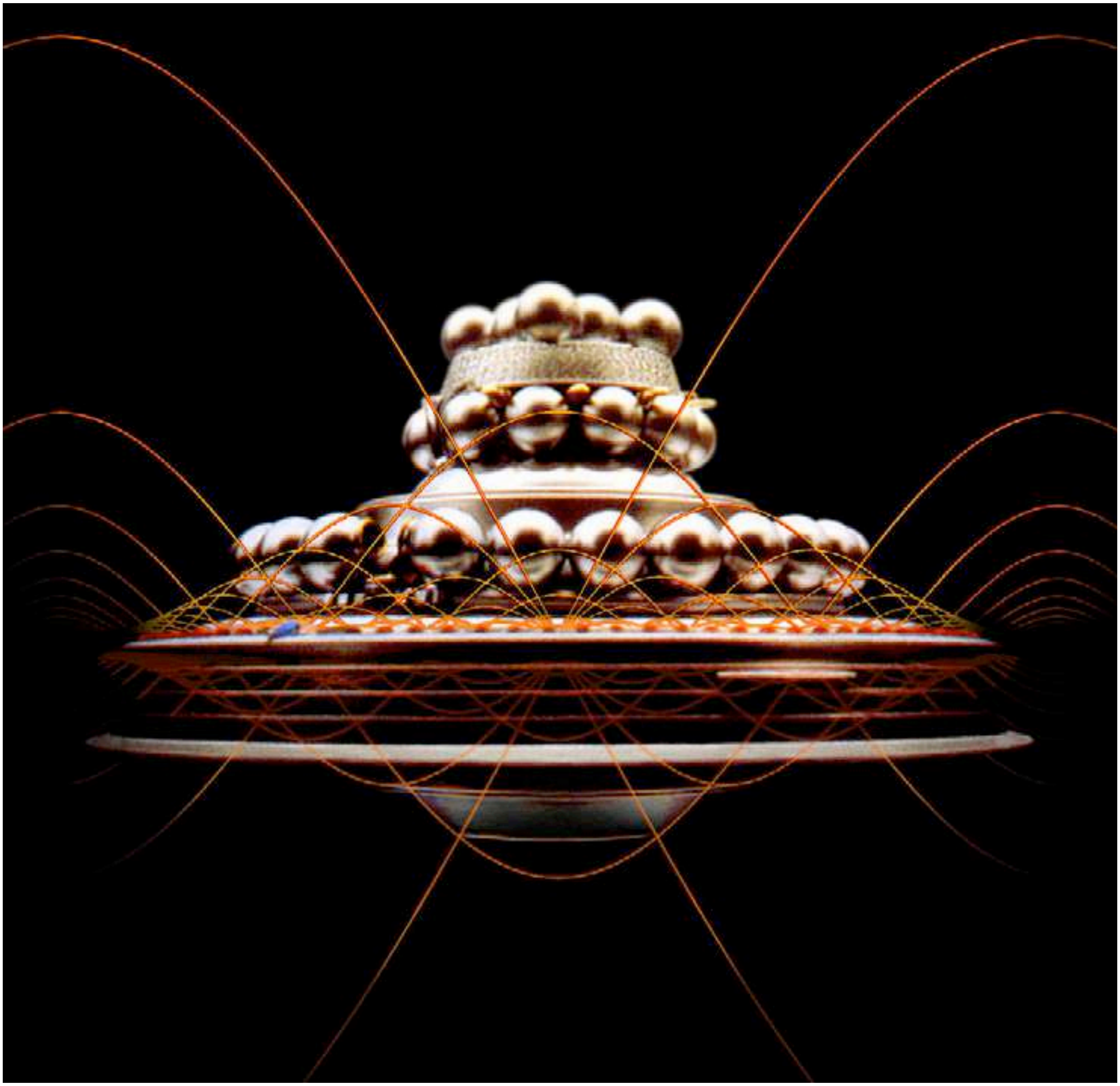

The craft's spherical resonators and compact circular hull serve in the production of standing acoustic waves that provide the antigravitic effect for flight by acoustic levitation. An array of 8 spheres is clustered at the top the craft, above a central row of 12 spheres and a wide lower row of 24 spheres – all of similar diameter. This 8–12–24 alignment of resonators presents the same quadrupolar geometric configuration observed of all of their spaceflight vehicles for generating Fibonacci-ordered standing waves composed of over 30 simultaneous, fixed frequencies with constantly shifting amplitudes (overlaid, above).

The loud singing of the resonant metals in flight can be shifted out of the audible range of human beings to be entirely imperceptible, utilizing deep, throbbing infrasound frequencies well below 20 Hz for levitation. The frequency of each spherical resonator is defined by wall thickness, not diameter, thereby allowing the specification of a unique frequency for each, as the high-speed rotation of the largest ring of resonators at 249.6 U/min. creates the 4.16 Hz modulation. *Never having seen Meier's photographs, acoustic analysts identified the object that produced the sounds as being composed of many spheres.*

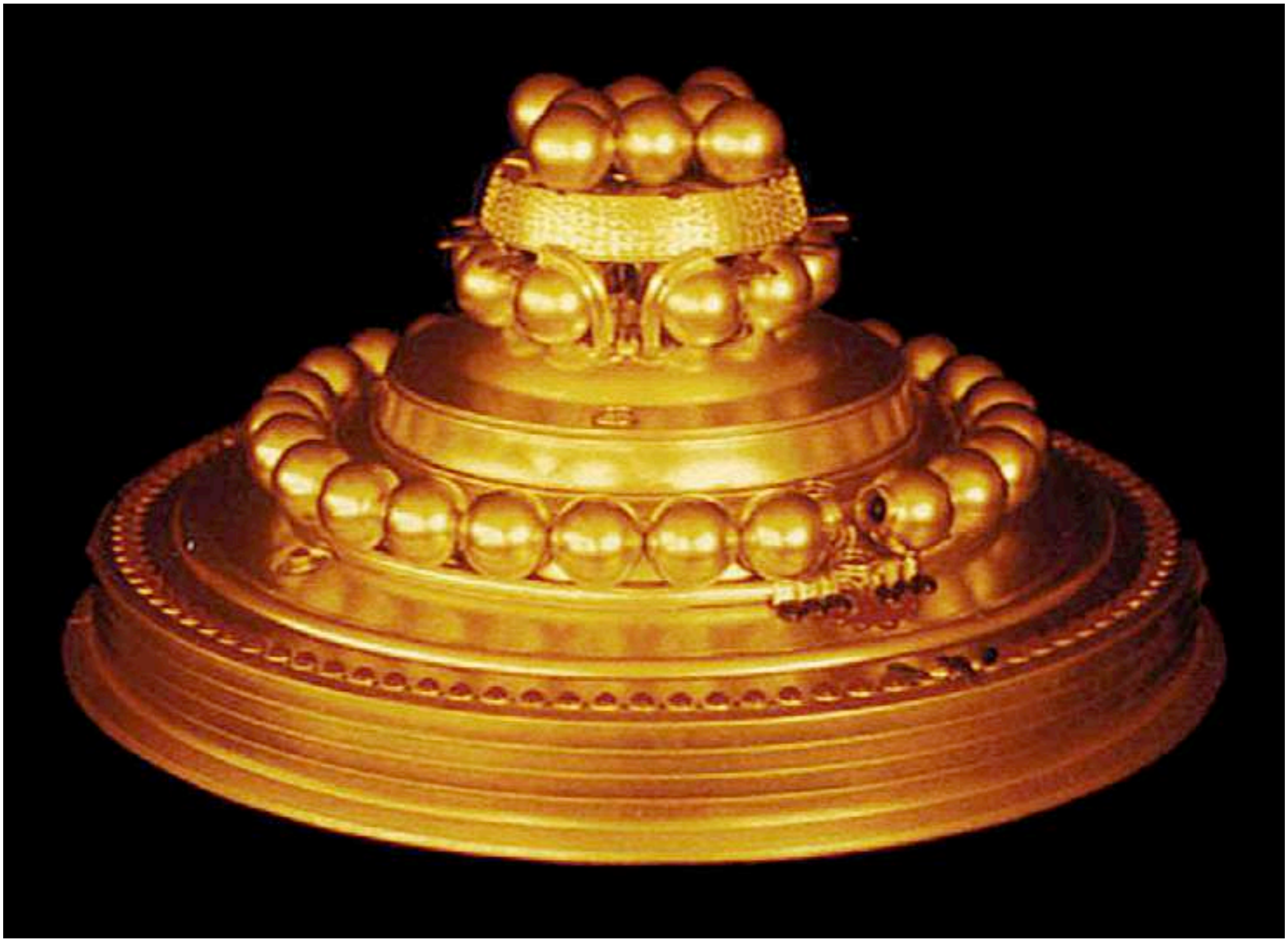

Meier was able to document an interesting effect of electroluminescence generated by Quetzal's beamship at nighttime from several perspectives, in a series of remarkable close-ups on August 5, 1981. Red light emitted by the normally silver-colored metal surfaces gives the ship a golden appearance. One image from the series captured the ship in a particular configuration not seen in any of Meier's other photographs, with a reflective metallic viewing screen exposed by the retraction of a small protective sleeve (above). Comparing photographs of Quetzal's ship reveals the rotary function of the lower array of resonators. The compact circular design of this interstellar vessel reflects the same mathematical structure that guides the larger beamship designs, presenting the precise, unmistakable form of the Mandelbrot quaternion (at left). The octagonal symmetry of the Mandelbrot set defines the distribution pattern of interconnected spherical resonators mounted in circular rows along the upper tiers in the 8–12–24 arrangement. *This 'otherworldly' plasma vessel reflects the same geometries seen in crop circles.*

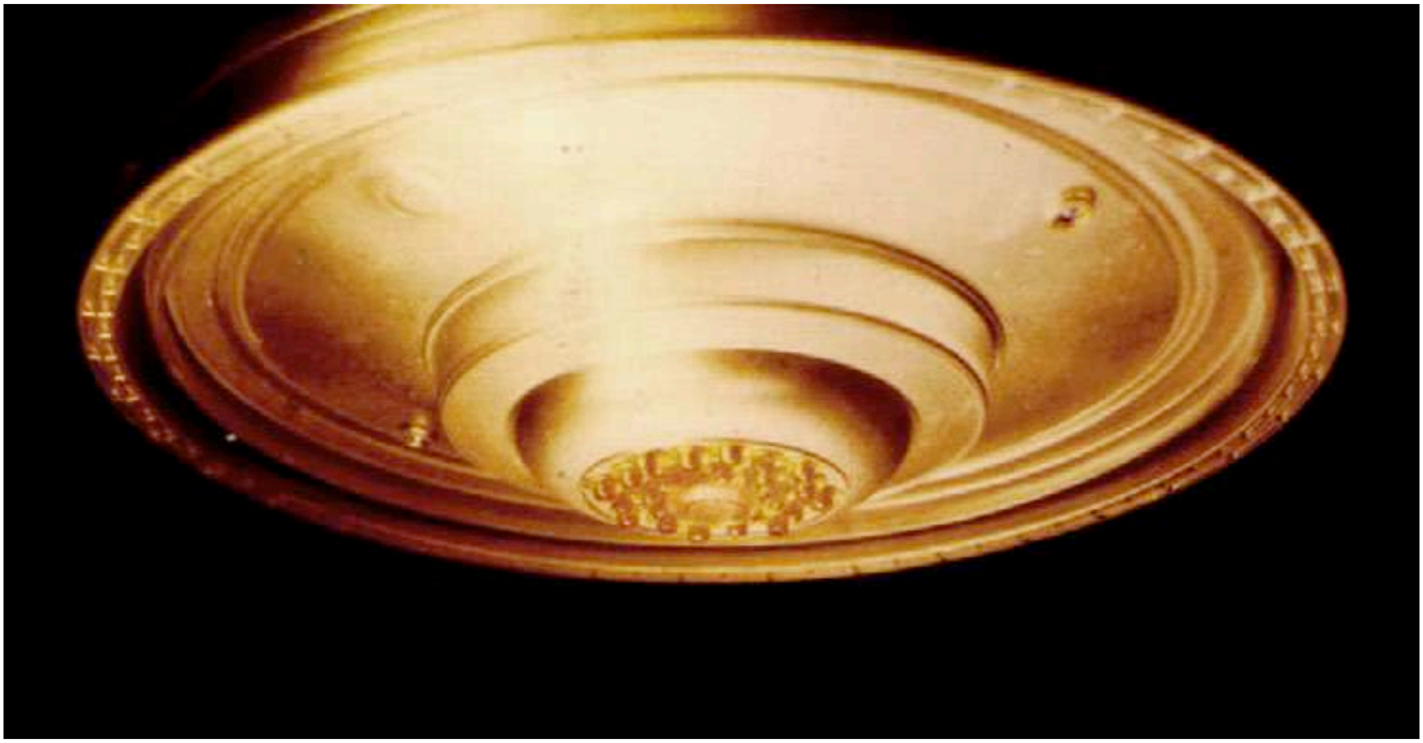

The close-range photo series included well-illuminated, detailed nighttime images of the tiered convex underside of the beamship piloted by Quetzal. Recessed patterns are seen under the beamship's outer flange and details of the lower hull's concentric rings include a pair of circular fixtures and other closed circular ports just barely visible (likely housing landing supports). The large, stepped central element terminates in a nearly hemispherical platform with an array of HHO plasma beam emitters, for which the 'beamships' have been named. Plejaren beamships in Earth's airspace maintain complete invisibility, yet special occasions presented themselves whereby members of the Meier's community were able to witness and photograph various spacecraft and other associated aerial phenomena. While Semjase's description of global celebrations on her homeworld of Erra included immense, multicolored aerial HHO plasma displays by fleets of beamships, only modest demonstrations were made for Meier's group.

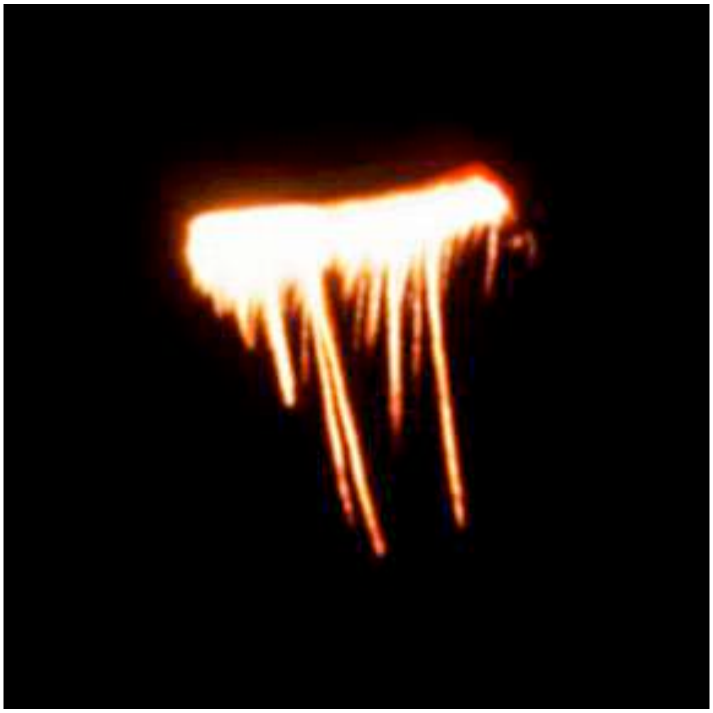

FIGU member and author Guido Moosebrugger was able to photograph a nighttime aerial beamship demonstration on June 13, 1976 above Winkelriet/ Wetzikon (at left). The bright luminous display distinctly presents an array of infrared beams acting as guide paths for glowing orange columnar discharges of HHO plasma. Rather than providing some form of thrust as would be expected in the context of plasma jets used in basic terrestrial rocket technology, the plasma beams can be used to target and incinerate objects below the ship, and to levitate objects upwards from very far below. This beam is often used for tasks such as collecting atmospheric, terrestrial or aquatic samples and specimens, as well as generating luminous aerial displays. Specific colors emitted by HHO plasma can be controlled by simple chemical additives being vented with the HHO gas, thereby forming plasma trails of banded colors in the wake of the plasma craft as it maneuvers aerially.

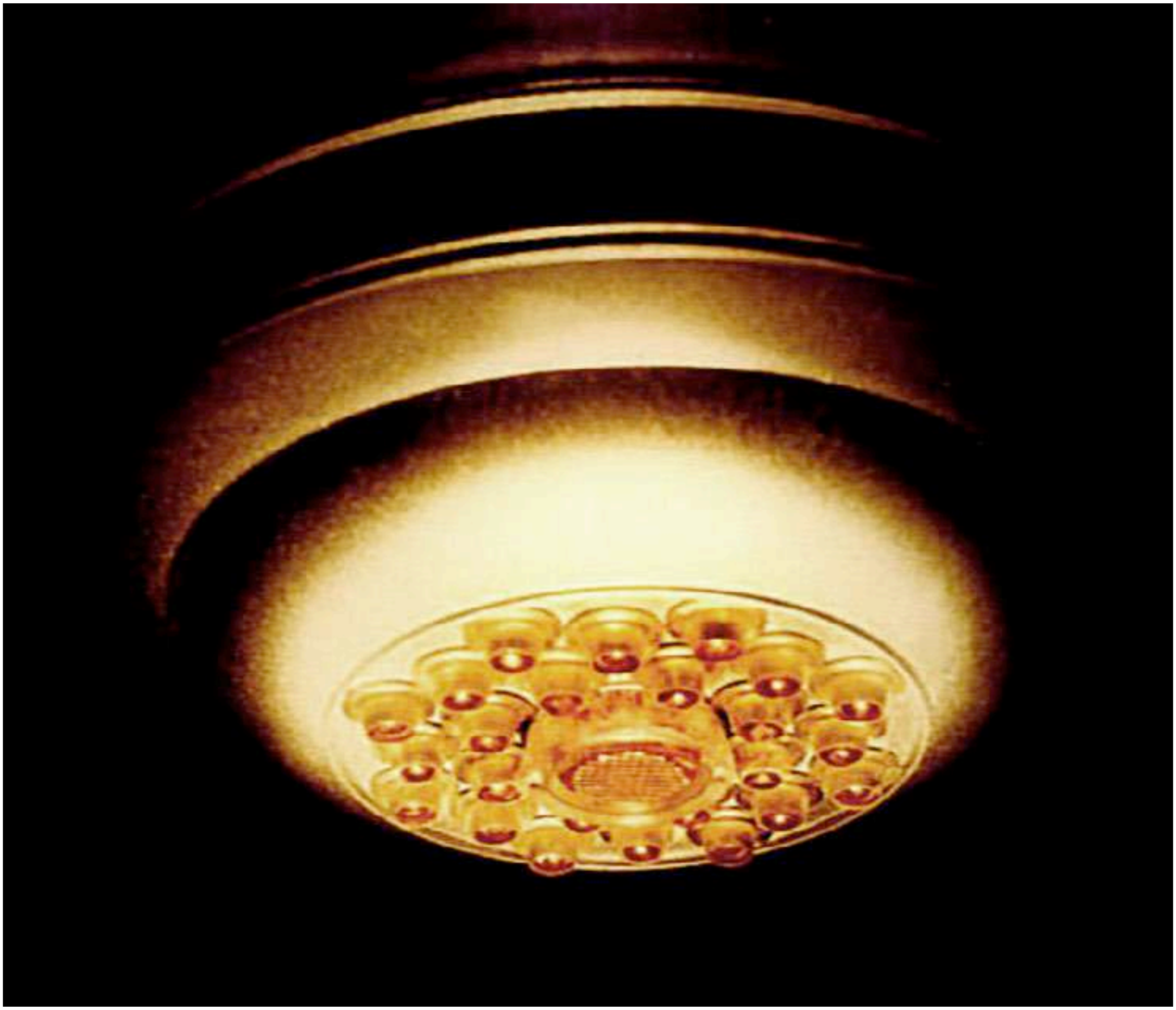

The circular bottom of the reconnaissance beamship houses a symmetric octagonal array of 24 plasma beam emitters arranged around a larger central element which vents HHO gas into the beam array to be ignited as HHO resonant-transfer plasma (above). Each of the plasma emitters contains a ruby for the generation of a beam of infrared light, precisely aimed by a short nozzle. The plasma emitters' achieve a narrow range of rotary motion that allows all 24 beams to be focused on one central point several feet below the beamship, or independently into any other desired configuration. The alloyed nozzles focus an electromagnetic field that encircles the infrared beam, thereby confining the HHO plasma within the beam. These techniques for plasma beam focusing are now commonly used by plasma physicists, yet are rarely studied in the context of beam propagation over vast distances through the atmosphere.

Well-known examples of resonant metal alloys formulated by Sanskrit-based Buddhist traditions in Asia include Tibetan singing bowls, comprised of 3 metals, and finer Bhutanese singing bowls rendered from 7 metals or more, known for highly resonant characteristics. While the singing bowls of Asia are comprised of a layered composite of various metals formed by hammering, the much more advanced resonant alloy production processes of Plejaren metallurgists involve the co-mingling of atoms in a composition unique to space-faring plasma ships, unlike any produced on Earth in modern times. Billy accepted several resonant metal samples as gifts from Asket, the products of her civilization, the Timars, who use similar advanced techniques of frequency resonance applied by the Plejaren, and indeed all cosmic travelers.

The resonant metal alloys of transdimensional plasma ships are developed using precision multi-stage temperature and frequency programs that convert base metals into alloys (above) containing all of the precious metals, rare earth elements and trace amounts of every other element on the periodic table – many in crystallized or metallic states not yet conceived of by Earthly scientists. Polarization of the final products is achieved by the micro-layering of perpendicular grain patterns formed within the crystals.

Semjase provided Meier with Plejaren beamship metal alloy samples of a *7-stage metal conversion process from lead into a copper-nickel-silver-gold alloy, also containing traces of every stable element in a smooth curve from the heaviest elements through the lightest gases.* Many aspects of this complex process were discussed during the 45th contact on February 25, 1976, in response to Meier's direct request – "Quite simply, we would like to know what sort of metal you use for your beamships?"

> *Semjase:* This I can explain for you: We gain it from a process of conversion from lead. We extract this soft metal from many things, as for example lead-containing atmospheres from stars, from waters, from different plants, etc., as well as from the decay of different ore-stones of stars in destruction. By an, according to Earth understanding, very difficult process, we convert the lead substances we assembled into the soft metal lead, which we then change by further mechanical-chemical processes into a hard-metal form, which is much harder than your metal which you call steel. This still is unserviceable in this form for beamships, because it must be polarized by further processes of conversion into a beamship-suited alloy of a special sort and character, about which I am not allowed to give details. The final product, which must have quite certain characteristics, consists of… a copper-nickel-silver alloy, which for certain of the beamships also contains gold.
>
> *Meier:* Oh yes, thus an alloy which could be produced on Earth… You mentioned a lot of manners before, in ways in which you gain the lead. Here it has struck me that you have mentioned no word of an ore mining of Earth kind. Is this sort of obtaining ores not usual with you?

> *Semjase:* Ore-mining or other mineral mining on a planet or another star is done by us only in extreme emergency, because this process is equal to the destruction of the star. A planet or another star is never exploited in the form as is done on Earth. What the Earth human being does by that is equal to the destruction of the planet. The first evil effects of this destruction came up on the Earth already some decades ago, and at the present time it is already experiencing the pains of the destruction itself. This is to understand the way that the Earth human being exploits his planet, and robs from it its fundamental life force when de deprives it of its oil, gas and the different ores. The effect of this is that the Earth suffers shifts inside, leading to volcanic eruptions and earthquakes, because the Earth is slowly breaking down into itself. The same event is generated also by the construction of dams and similar formations, which cause, by the assembled masses of water, very dangerous shifts. The worst delusion the Earth human being is performing is his subterranean and below-surface experiments with atomic bombs, as well as those cruel deep subterranean explosion tests, which he declares are atomic bomb tests, but which in truth are much more dangerous.[42]

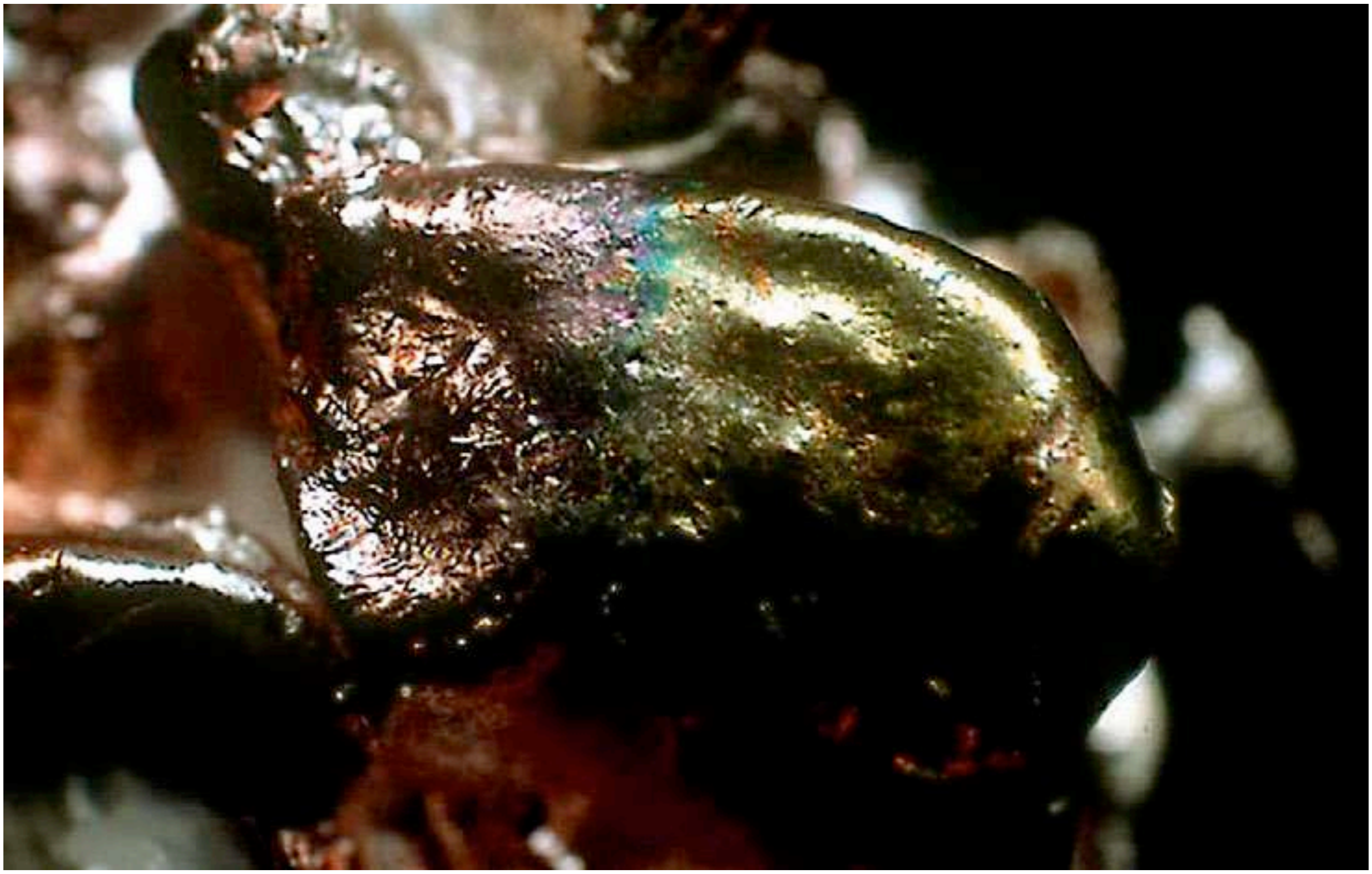

The spiritual ethics of the Plejaren culture reflect a comprehensive understanding of planetary systems. All 7 stages of beamship metal production were related on the following day, during Meier's 46th contact, with an astounding explanation of the atomic transmutation of metals by *controlled frequency resonance:*

> *Semjase:* Here I also have brought you some metal, of three different working steps… The first working step absorbs the lead substances from the atmosphere and condenses them into pure lead. The second process distracts from the resultant metal all dangerous radiations. Then the lead is fed into heat-converters, which, without the addition of any other metals, converts the lead by several processes into the alloy.
>
> *(Ed.- The metal specimens finally delivered to Meier, which they and we had analyzed, consisted of three of the seven states the metal goes through in production and refinement. We had samples of states 3, 4, and 5. The sample of state 3 was mostly lead, with impurities. State 4 was nearly equal parts of silver, copper and lead. State 5 contained silver, copper, nickel, gold, magnesium and other trace elements. We lost our specimens in the hands of testing laboratories before we had finished our analysis. We found traces of stable rhenium and thulium in our specimen tested by Marcel Vogel in his laboratory in San Jose, California. Since our testing in 1979, scientists have found that the addition of rare earth elements like rhenium and thulium produce astounding properties in otherwise basic metals, such as superconductivity, superhardness and supermagnetism, etc.)*

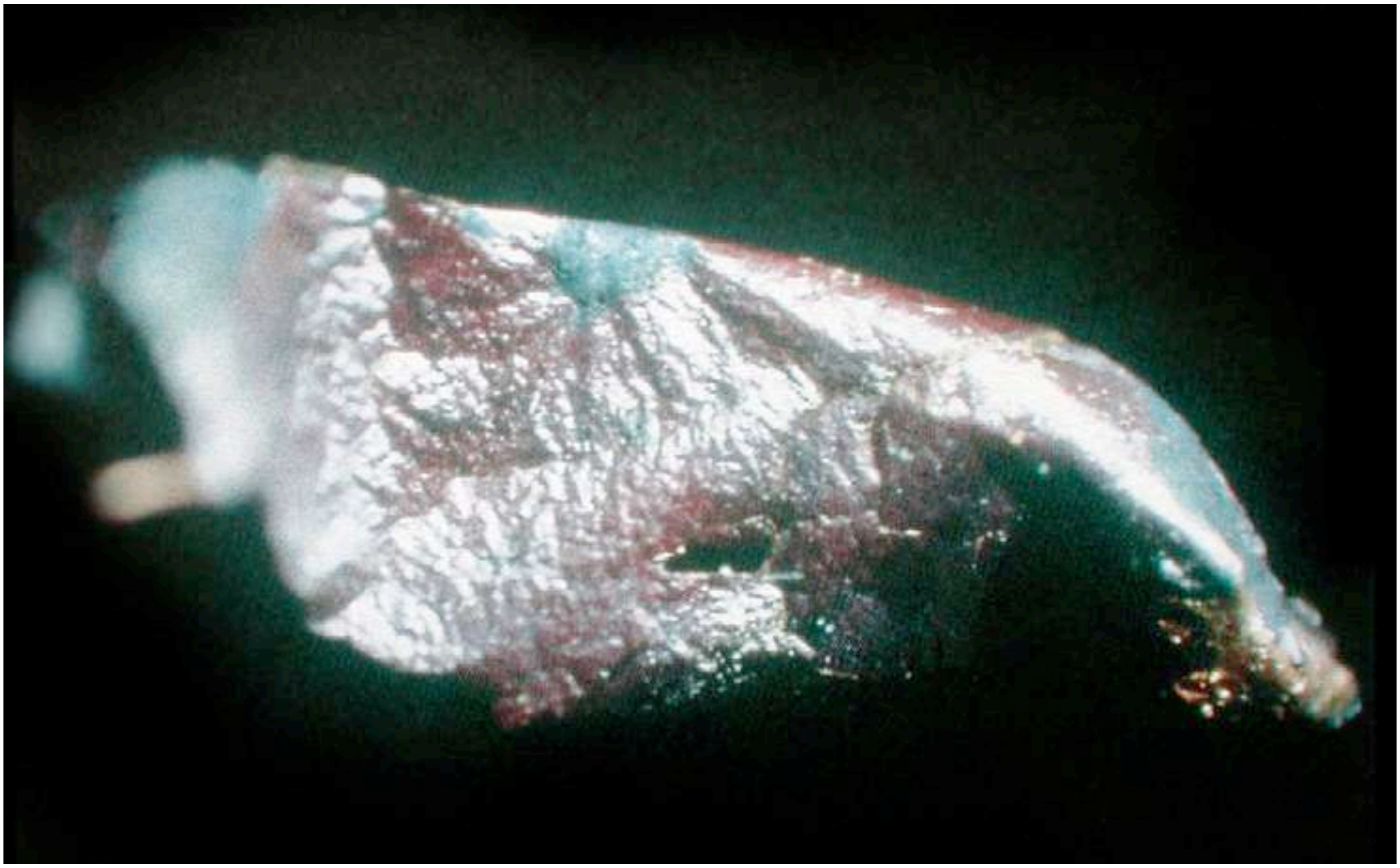

Meier: You mean that the lead becomes liquefied first, and in this step suffers the conversion?

Semjase: Surely. There do all right exist direct possibilities for the alternation of stuff, but these possibilities are not given for us. But our scientists have already well achieved successes in experiment. For the present, however, we still convert the metals the old-fashioned way in heat-converters, and as said, this happens in seven different runs. The metal becomes liquified, like with you, in furnaces, to suffer then by distinct oscillations, an alternation, but just to a certain value, because of the different steps in working that are necessary. Each final product is then pushed by pressure through a cooling spiral, where then are generated the small figures, as you have then here before you. This process is repeated some times, while with each new process the value of the different metals increases and becomes the targeted alloy. The sixth working process then affects the complete alloy… The cooling spiral contains usual water, which we obtain especially for this process in a condensational manner…

Meier: Thank you. Still it is not evident to me how you then will work the metal, as neither at your beamship, not Ptaah's huge ['Great-Spacer'] box, could I recognize joints or seams, etc. And nothing is riveted as well.

Semjase: For this purpose we use an instrument which you would call a welding apparatus. But it is based on oscillation techniques, which liquifies the metal in a cold state and lets it flow together, by which it is completely without seam and forms one single piece. So we also do not know grinding operations, as you do when you have to do welding work, and then have to grind away the seams to smooth it all. In this respect, you on Earth use a very dangerous operation.[43]

While Earthly science achieved the liquification of metals using high-intensity ultrasound[44] decades ago, Semjase informs us that for thousands of years Plejaren metallurgists have been producing beamship alloys by *subjecting molten metals to "distinct [atomic] oscillations of a certain value."* What are these particular *resonant frequencies* that induce the alchemical transmutation of one element into another? The atomic physicists among Earth's scientists have been making great strides in their understanding of low-energy nuclear transmutation – the most outspoken among them being Dr. Joseph Champion of the Phonon Resonance Institute in St. George, Utah.[45] Dr. Champion's work elucidates the atomic principles enabling several successful methods of atomic transmutation involving the production of precious metals:

> Dimensional phonon resonance occurs when the space occupied by one isotope is *exactly the same as that of another isotope in its rest state* [i.e. 25°C]. This event can only occur under the following two conditions: the expansion of an isotope by heating, or the contraction of an isotope by cooling. Due to the natural characteristics of elemental properties, this event is extremely rare and one can only force the event under select conditions. To determine the phonon resonance of an isotope, it is necessary to apply the following formula:
>
> $$\text{Phonon Resonance (Hz/Cm)} = \sqrt[3]{\frac{d \times Na}{m}}$$
>
> *d - density in g/cm³*
> *Na - Avogadro's Constant*
> *m - atomic mass*
>
> In the formation of Au (or other elements) from a dimensional reaction, the conversion will occur without excess energies or nuclear signatures… by heating Ag to a temperature of 43.2°C… To achieve maximum conversion of Ag to Au will depend on the dwell time at resonance temperature. To date, visible conversion of Ag to Au has taken as little as six hours, with 2% conversion taking up to 24 hours. The reaction is safe… [with] no toxicity.[46]
>
> $$^{107}\text{Ag Phonon Resonance (Hz/Cm)} = \sqrt[3]{\frac{10.5 \times (6.0221 \times 10^{23})}{106.905092}} = 38{,}962{,}532$$
>
> $$\text{Au Phonon Resonance (Hz/Cm)} = \sqrt[3]{\frac{19.3 \times (6.0221 \times 10^{23})}{196.967}} = 38{,}931{,}841$$

Dr. Champion's phonon resonance formula provides the resonant frequency of *gold* (Au) in its rest state as 38,931,841 Hz and the ^{107}Ag isotope of *silver* as 38,962,532 Hz (above). The relative proximity of the resonant frequencies of these two precious metals allows a low-energy atomic transmutation from ^{107}Ag to Au to occur *near room temperature*. However, silver isotope ^{109}Ag comprises 48.8% of all available silver bouillon and, unlike it's sister isotope ^{107}Ag, is not so easily converted to Au.

The conversion of ^{109}Ag into gold requires a more complex process using megahertz frequencies in conjunction with precision heating techniques. By the same means aluminum can be converted into silver when oscillating at a resonance of 38,950,565 Hz (at 276°C) and will also convert into gold when dwelling at a resonance of 39,932,068 Hz (at 295°C). After just a few days of dwell-time at resonance, atoms of the starting element have become *informed with the resonant frequency* of the target element in its rest state. The metal is then melted and rapidly cooled in distilled water to achieve the maximum conversion rate, *just as in the procedures described by Semjase in 1976.* The respective electron configurations of aluminum, silver, platinum and gold reveal the restructuring of electron orbital shells during transmutation:

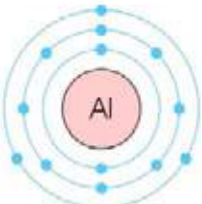

13: Aluminum 2,8,3

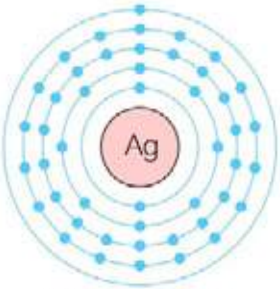

47: Silver 2,8,18,18,1

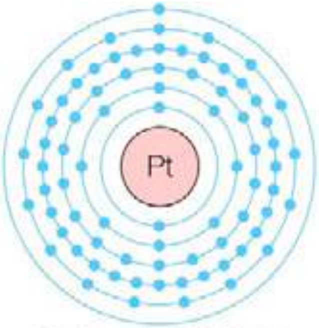

78: Platinum 2,8,18,32,17,1

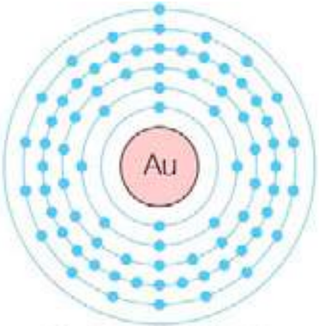

79: Gold 2,8,18,32,18,1

The addition of subatomic particles into atoms vibrating at resonance is facilitated by *standing waves that act as pathways for the reconfiguration of electrons*, *protons and neutrons into a more densely packed arrangement.* Just as acoustic standing waves can power wireless devices by transmitting electrical currents across otherwise nonconductive airspaces, standing microwave resonances induced within atoms allow subatomic particles to jump across otherwise impassable gaps in the atomic framework. Direct observation of this resonant phenomenon of atomic transmutation will be made possible by the quantum stroboscope, which has already provided the first direct imaging of standing waves that comprise the structure of a single electron. A source metal like aluminum can thus be transmuted into silver and then gold as successive electron shells become filled, adding atomic weight in the process and significantly increasing the yield weight of the conversion with each stage of the complex process.

Dr. Champion has identified a new class of intriguing natural metabolic processes[47] where atomic transmutation occurs as fungi ingest silver, thereby undergoing a *pleomorphic change*. Dry, active yeast is a cheap and readily available bread-making product that requires only warmth and water to thrive and reproduce in vast numbers. Common yeast species like *Saccharomyces cerevisiae* produce the silver-binding protein SSB-1 that enables the attachment of active yeast cells to available silver surfaces.[48] Ingestion of silver atoms by the round 2-3 nm microbes causes their mutation into rod-shaped mutant microbes about 1-2 nm in length (below). The budding yeast cells act as *electron donors* inducing the resonant transmutation of silver into gold, as 32 electrons, 32 protons and 57 neutrons are accepted by ^{107}Ag silver atoms vibrating at a 38,931,841 Hz resonance with gold. This newly recognized metabolic process leaves nanoclusters of gold deposited on exposed silver binding sites, as further transmutation takes place within the bodies of the microbes whereby ingested silver is converted into gold nanorods.

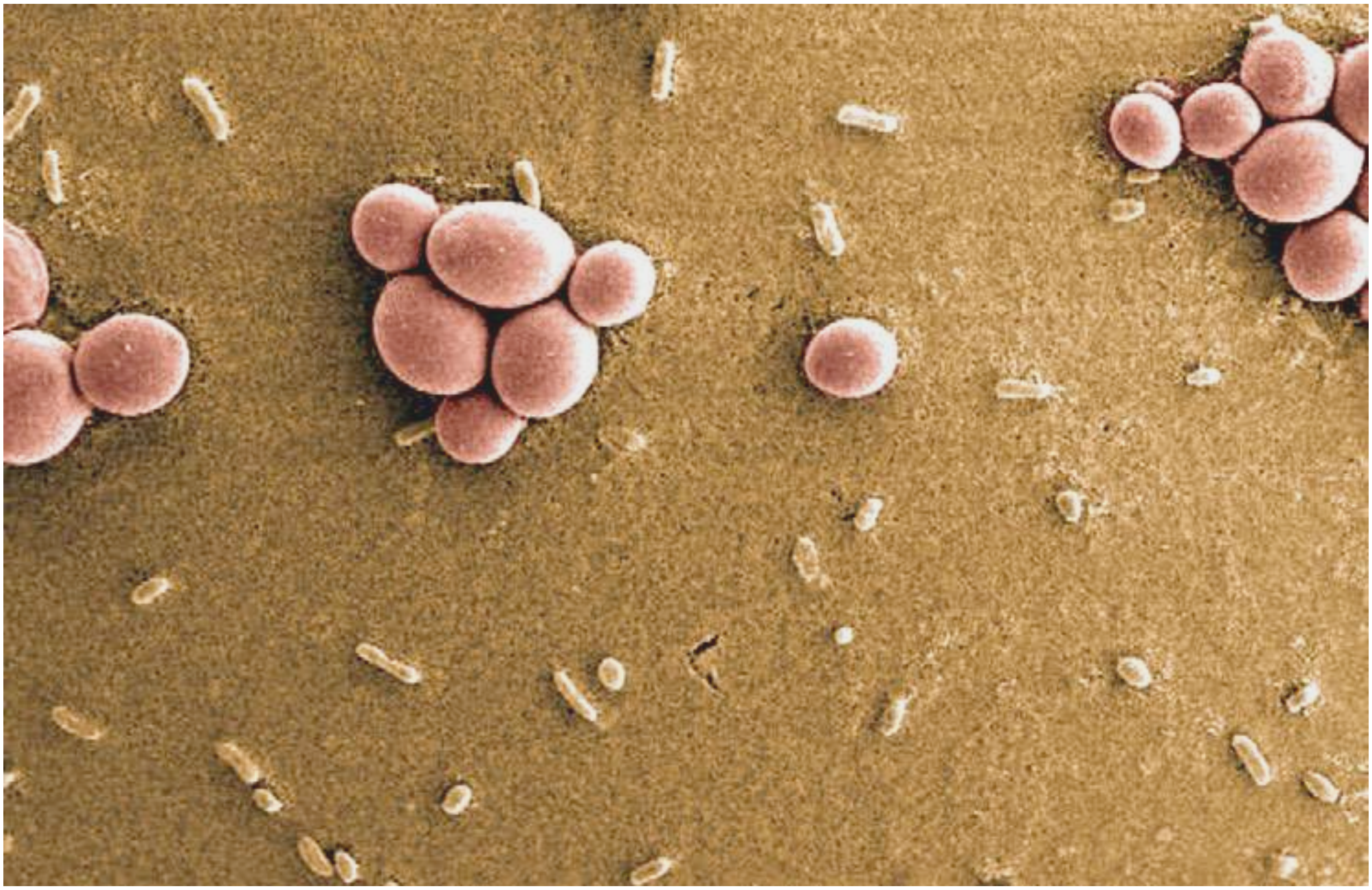

The rediscovery of biological transmutation in yeast species restores ancient metallurgical knowledge and informs a scientific reinterpretation of enigmatic passages from the Bible. Biblical accounts of the divine presence often refer to a sacred substance called *manna,* which was used in making holy items such as the ‘bread of life’ (or showbread) and the ‘living water’. However, this holy bread was not simply produced by a common baker. Instead, Moses commanded the goldsmith Bezaleel to prepare manna for the ‘Bread of the Presence of God’ – *apparently using yeast to convert silver nanopowder into gold.* Many unmistakable references preserved in the Book of Exodus confirm the biological process of atomic transmutation was used in the enrichment of food and water for the elevation of human consciousness. Advanced material and spiritual knowledge possessed by Moses is revealed in his demonstration that the presence of God is not conferred by gold idols, but by gold nanoparticles in sacred bread and water:

> When Moses approached the camp and saw the calf and the dancing, his anger burned… He burnt the golden calf - melted it down, and then filed it to dust; and that the powder to which it was reduced might be taken notice of throughout the camp, he strewn it upon the water which they all drank of. He did this that it might appear that an idol is nothing in the world, he reduced this to atoms, that it might be as near nothing as could be. (32:19-20)

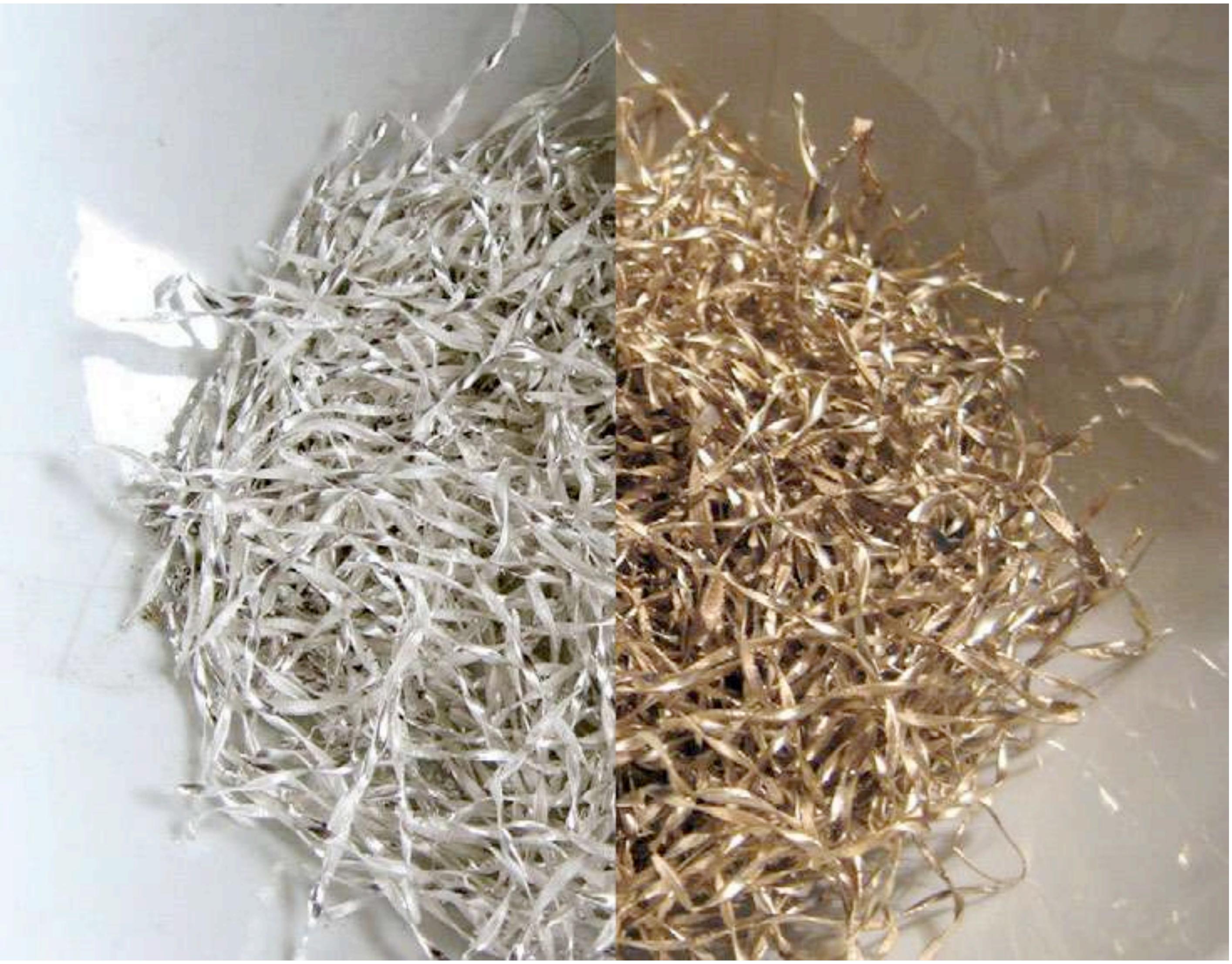

My first experiments achieved ~3% transmutation of silver into gold in 7 days using the silver/yeast method at 108°F (before/after seen above)! A comprehensive spreadsheet provides phonon resonance frequencies and temperatures for most isotopes based on each atom's physical diameter.[49] A small selection of phonon resonance frequencies determined for various conversions follows:

Source Isotope	>	Target Isotope	Resonant Frequency	Resonant Temperature
^{107}Ag (silver)	>	^{197}Au (gold)	38,931,841 Hz	43 °C / 108 °F
^{95}Mo (molybdenum)	>	^{65}Zn (zinc)	39,400,694 Hz	115 °C / 239 °F
^{65}Zn (zinc)	>	^{27}Al (aluminum)	39,296,350 Hz	183 °C / 362 °F
^{27}Al (aluminum)	>	^{107}Ag (silver)	38,950,565 Hz	276 °C / 529 °F
^{27}Al (aluminum)	>	^{197}Au (gold)	39,932,068 Hz	295 °C / 563 °F
^{107}Ag (silver)	>	^{109}Ag (silver)	38,652,383 Hz	327 °C / 620 °F
^{65}Zn (zinc)	>	^{107}Ag (silver)	39,138,500 Hz	344 °C / 653 °F
^{65}Zn (zinc)	>	^{197}Au (gold)	39,126,761 Hz	356 °C / 673 °F

The atomic resonance of various elements provides the particular conversion cascade by which nature creates the diversity and abundance of matter through great heating and cooling cycles recurring throughout the depths of the cosmos. The layered lithosphere's seemingly endless variety of minerals and ores are the byproduct of extreme conditions of pressure, heat and time. Bacteria thriving in these conditions also contribute to the Earth's internal atomic transmutation and decay processes, just as microbes are essential to metabolic processes in the human body, with its own specific resonances.

The full spectrum of atomic resonances present in all planets endows resonant vibratory characteristics, and Plejaren metallurgists have produced resonant metals by the same means. Beamship metals also contain traces of *all* stable elements present in nature. Eduard Meier's Plejaren contacts have explained that transdimensional spacecraft are like tiny artificial planets unto themselves, and the beamship metal samples provided to Meier have given further insight into the deeper layers of truth contained within their many concise explanations. The delicate threads of knowledge woven into Meier's communications with the cosmic Plejaren culture have been published word-for-word for this very reason: to preserve for human contemplation their advanced scientific and spiritual teachings in pristine form.

The penetrating truth of the Plejaren teachings has become all the more apparent as the frontiers of human knowledge expand toward recognition of natural law and co-creation using diverse natural micro- and macrocosmic processes. The manipulation of an atom's diameter by temperature adjustment allows for nuclear transmutation from one element to another in *a specific cascade sequence* that determines the frequency programs used by Plejaren metallurgists to convert lead into a resonant copper-nickel-silver-gold alloy with trace amounts of every other stable element. The third, fourth and fifth-stage beamship metal samples were submitted for microscopy analysis to Marcel Vogel, senior research scientist for IBM, who later confessed his amazement at what he observed under magnification:

> "When I touched the oxide with a stainless steel probe, red streaks appeared and the oxide coating disappeared. I just touched the metal like that, and it started to deoxidize and become a pure metal. I have never seen a phenomenon like that before." Of another metal sample containing nearly every element in the periodic table, Vogel stated, "Each pure element was bonded to each of the others, yet somehow retained its own identity." At 500X magnification thulium was revealed. "Thulium exists [naturally] only in minute amounts. It is exceedingly expensive, far beyond platinum, and rare to come by. Someone would have to have an extensive metallurgical knowledge even to be aware of a composition of this type", said Vogel. At 1600X Vogel said, "A whole new world appears in the specimen. There are structures within structures - very unusual." At 2500X he found that the sample was, "metal, but at the same time ... it is crystal!" …Vogel put the full weight of his expertise in these summary comments: *"With any technology that I know of, we could not achieve this on this planet! I could not put it together myself, as a scientist...*[50]

Vogel's astonished assessment of beamship metal constitution was made independently, without having any advanced knowledge of the comprehensive details Meier had been given by Semjase concerning all 7 stages of the resonant metal manufacturing processes of Plejaren metallurgy! The most significant details omitted by Semjase (for good reason) are the precise transmutation frequencies necessary to achieve each specific conversion. The complete set of phonon resonance frequencies determined by the formulas of Dr. Champion allow the integration of Plejaren teachings with the theoretical framework understood by contemporary atomic physicists. Dr. Champion's work inspires a reinterpretation of the results of various analyses conducted by Marcel Vogel in 1979, which comprised the first modern scientific study of resonant metal alloys produced by the cascade process of atomic transmutation:

> All of the elements examined spectrographically had missing bands in their spectra which should have been there if they were normal atomic spectra. This indicated that the elements were put together in a very unusual way [easily distinguished] from normal Earth technology… [presenting elemental characteristics] beyond what one would consider an isotope… Basically this indicates a non-electrical cold fusion process of synthesis because there was no ash and no heat residue. Such a process is not known to Earth technology… Another unique feature of these elements was the almost perfectly uniform fall-off of lighter elements preceding that being studied in the spectrograph. There would be the primary in a high state of purity and then with no secondary and no catalyst an absolutely uniform curve of all the lighter elements preceding that, even through the gases. It appears as though the desired metal was the result of a uniform buildup of density, particle by particle!
>
> 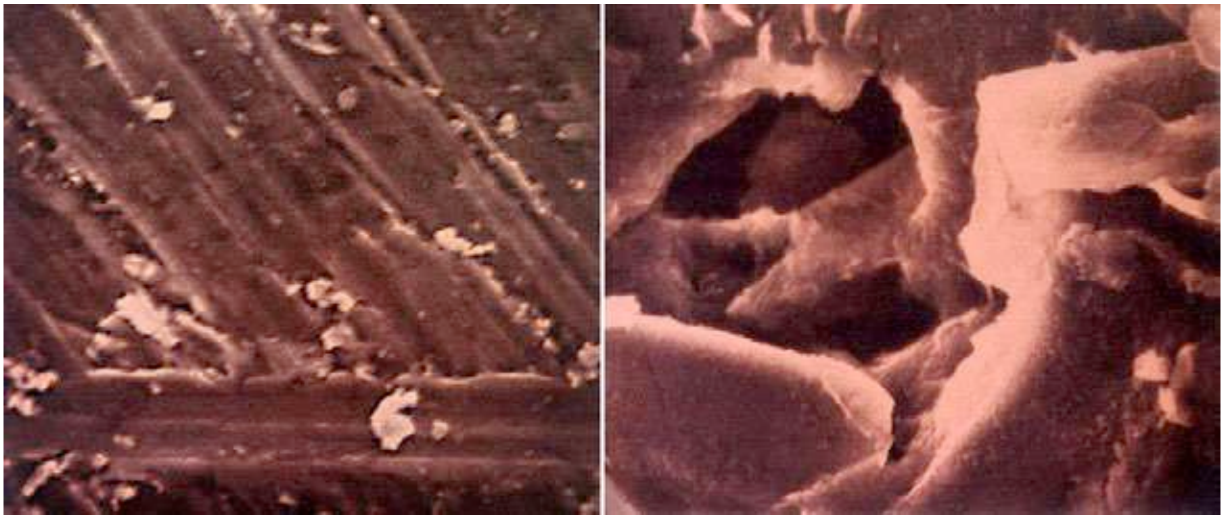
>
> Examination of a non-metallic part of the specimen using polarized light to emphasize certain bi-refringence revealed crystalline structures, which are organized in spiral formations. Under 2000X scanning electron microscopy the six-sided face angles of the crystals were clearly revealed. A photomicrograph picture shot of this revealed unusual clarity indicating unusual conductivity of the electrons illuminating the image. Ordinarily, if they were ordinary non-conducting crystals, like quartz, the electrons would accumulate rapidly on the points of the crystals and form a space charge obscuring the image. The conductivity was so efficient that sub-crystal structures inside the crystals could be actually seen – something unheard of before this. The electrical bonding between the basic crystal elements was so efficient that a clear view through the crystalline latticework revealed a spiral buildup of regularly spaced assembly with contact taking place at the ends of the sub-particles. There were no gas bubbles. If this were produced in a heat fusion process the picture would be entirely different and a greater mixing of elements would be expected… Still another micropart of the specimen consisted of pure silver –again with no secondary lines and no catalyst indication. The silver was also unusually conductive and it was smoothly bonded to a pure aluminum area that was unusually conductive also… The following unique properties encountered in examining this specimen make it exotic and very definitely point to an unearthly origin:
>
> - Unusual purity in state of metal examined… with no indication of catalyst or secondary spectrographic lines.
> - Unusual number of trace elements ranging from the primary heavy all the way up the scale density in a very uniform curve with no breaks or gaps up to the lightest gaseous elements. Again these lines all indicate the same high purity for each element. Scientists involved have never seen anything like this before.
> - Strange, very uniform grain pattern with a layer of horizontal grains sandwiched between layers of vertical grains again sandwiched between that and another horizontal grain, etc., in a tight 90° latticework.
> - One metal surface showed evidence of an extremely fine mechanical cut, possibly even by laser. [above]
> - Association of non-metallic crystals of strange structure and extremely high conductivity for the elements.
> - End-to-end formation of bond patterns in the crystal structure and the spiral assembly or stacking of the crystals in the non-metallic substance was unique.
> - Some form of non-electrolytic cold synthesis process is very definitely indicated.
> - There is some evidence that the synthesis took place in a vacuum.[51]

These many unique and astounding characteristics of the resonant beamship metal samples – including *superconductivity, superhardness and supermagnetism* – are only attainable through the phonon resonance application of standing waves for *tuned atomic transmutation in a uniform build-up of density, particle by particle.* The concept of standing waves was touched upon indirectly on many occasions during the Plejaren dialogs with Billy Meier, though the specific term itself was not divulged, in a situation very similar to the treatment of another sensitive term: 'HHO plasma'. Infrasound standing waves were mentioned in the context of "energy bales of different sorts and strengths… [that] circle the Earth at different distances and intervals, but are very important for the maintenance of Earth creatures" (pg. 80).

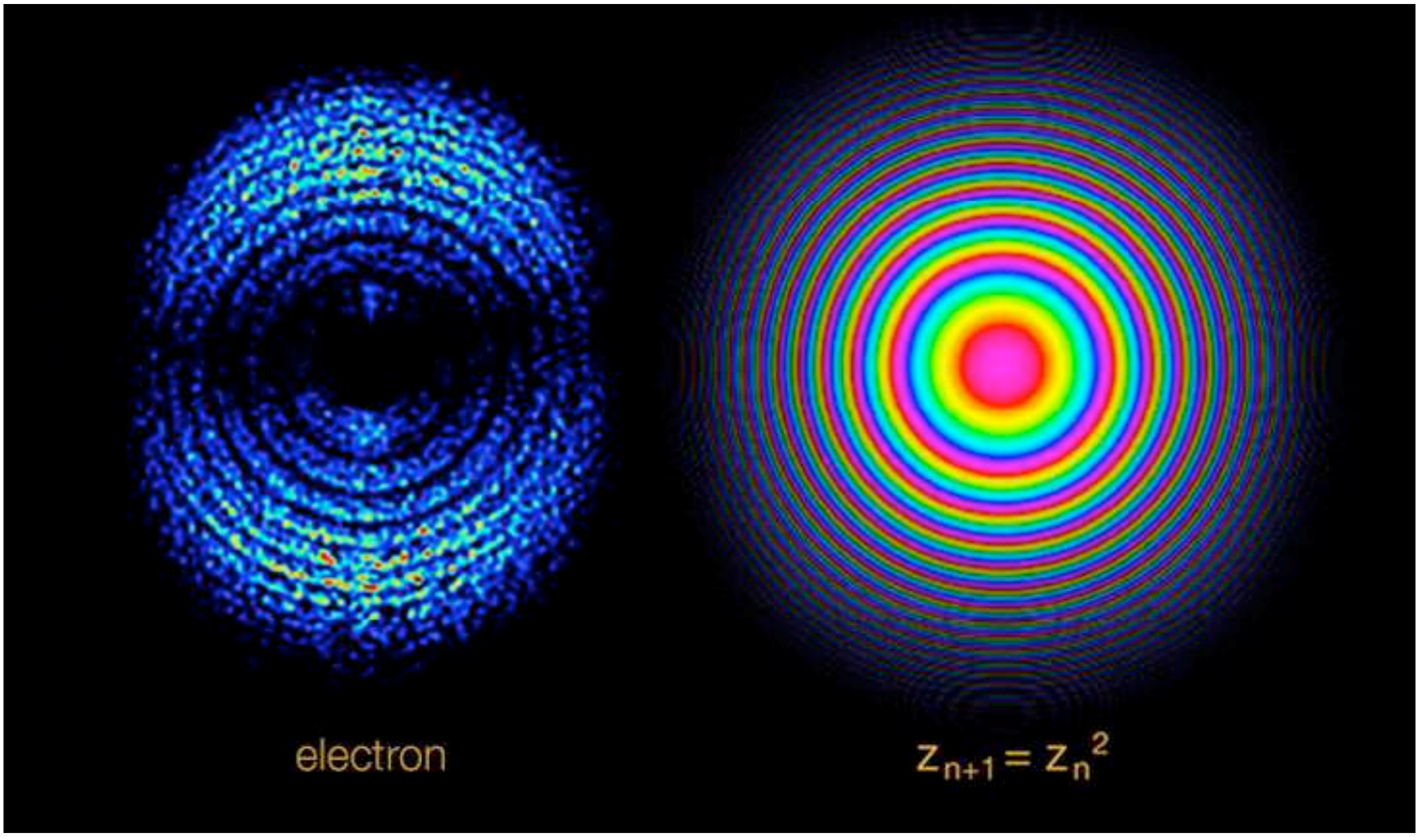

Semjase also indirectly referenced terrestrial standing waves when she informed Meier that "the gravity of a planet is… not always… of the same strength, owing to certain alternations which will be discovered by your scientists in a reasonable time" (pg. 89). In fact, over 30 years elapsed before my completion of these standing wave maps in 2005, defining the specific Fibonacci 'distances' or 'intervals' underlying this essential pattern of infrasound resonance. The Int'l Bureau of Weights and Measures detected variations in Earth's gravity field in 2007.[52] The diamagnetic metals that comprise their weights are presently displaying a measurable weight offset, although inaudible infrasonic effects of acoustic levitation were not recognized as the cause. This effect will inevitably increase in the future to reveal its inaudible source.

The same standing wave maps that inform the significant fluctuations of gravity also inform the latest discoveries of scientists imaging the subatomic realm. The structure of *the electron* was recorded for the first time in 2008, revealing the unmistakable scattering pattern of nonlinear standing waves, as precisely encoded by the quadratic formula [$z_{n+1} = z_n^2$] (above). Researchers in Sweden demonstrated an attosecond quantum stroboscope that releases individual electrons into the field of an infrared laser at exactly one electron per laser cycle to facilitate the imaging of coherent electron scattering that takes place.[53] *Standing wave resonances comprising the electron are presented 3-dimensionally (opposite), and are applied for infrasound antigravity propulsion by all space-traveling civilizations*, using diverse mathematical languages, and employing entirely different operations. During the 248th contact of February 3, 1994, Ptaah confirmed that the uniformity of universal laws determines that all mathematical and physics formulas applied by all sentient beings may find equivalence, differing only in the terms of expression:

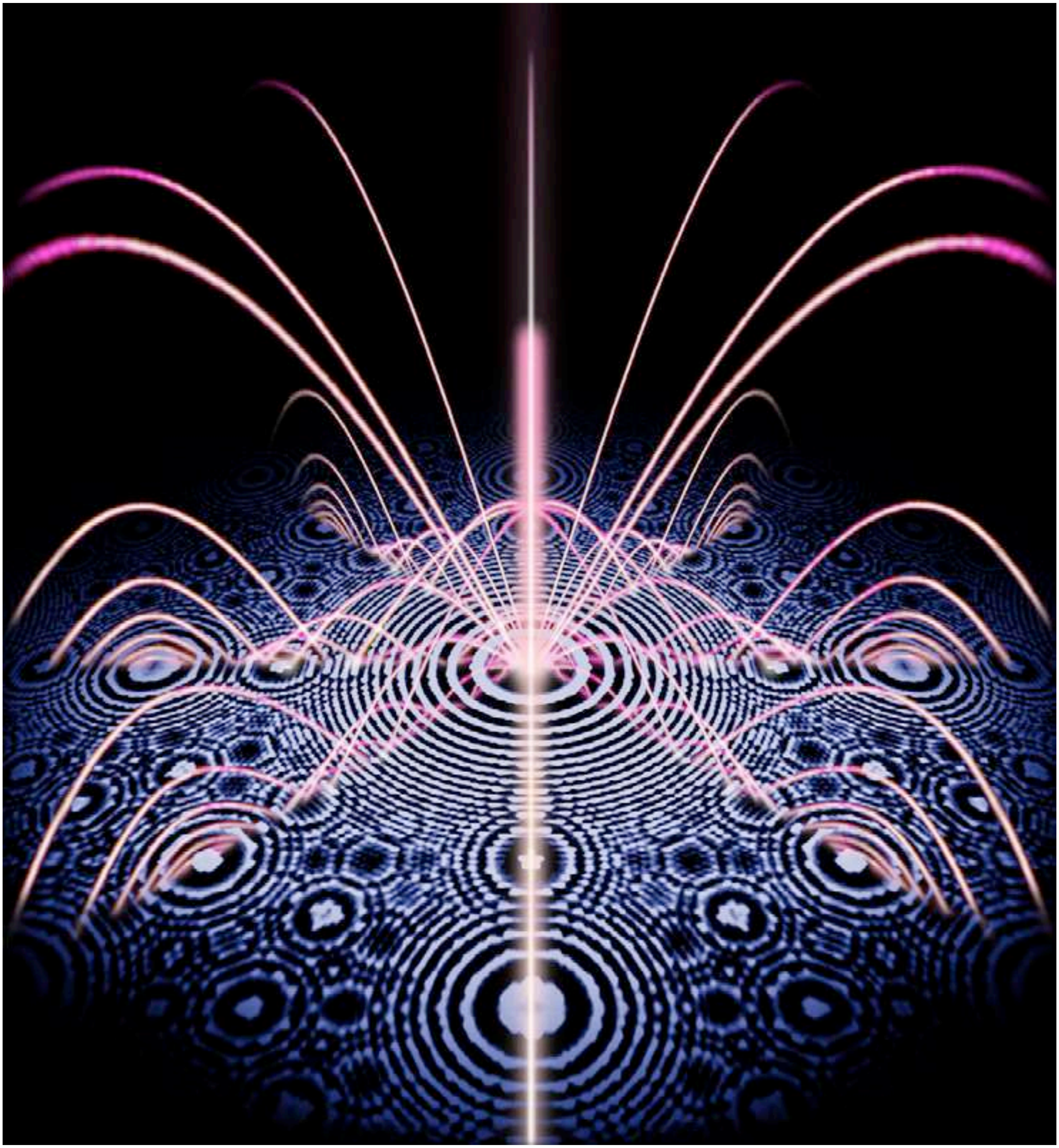

With regard to mathematics, that 3 times 3 is 9, it is to be said, of course, that the result is right and is also consistent with our mathematics, but the form of our mathematics is fundamentally different to the terrestrial one… Like with the mathematical formulas, the physics formulas also differ. But once they are converted, they yield the same values because universally, the mathematical and physical laws, etc. are uniform in their basic value and final value and, therefore, are one and the same, just with the difference that the various races, civilizations, and humanities of the various worlds throughout the vastness of the Universe call other forms of mathematics and other terms, etc. their own; consequently, they also have other methods of calculation than what are common among the Earth people. Nevertheless, in mathematics, the basic value and final value yield the same results.[54]

The vast variety of cosmic and planetary conditions within the universe facilitates natural learning environments of various types, each providing the necessary range of lessons and limitations for each developing life-form and each collective terrestrial civilization. Planets can be categorized by their respective levels of infrasound resonance, with planets such as Earth presently residing at the lowest level of the scale, and planets such as Erra, in the Plejares star system, finding their place among the more highly developed worlds. As Earth presently progresses from the primary level of consciousness lessons, the infrasound resonance of the planet is stimulated to a higher level of collective vibratory experience. Previously invisible forces will become visible to human eyes; subtle forces will become self-evident.

Rainfall patterns and major fluctuations in river levels have been directly correlated with solar flare activity –through the stimulation of infrasound standing wave resonance, raising vast amounts of water into the atmosphere to fall as rain. Rows of water vapor plumes evaporating off the ocean were photographed off Laguna Beach, California[55] in 2009, while the long-term water flow-rates of the Parana River were perfectly matched with sunspot numbers[56] in 2010. So-called constants like gravitation and atomic decay rates vary significantly, and in accordance with the levels of incident infrasound radiation upon the tested materials. *Thus, Earth's infrasound resonance with the sun defines the increasingly noticeable fluctuations of gravitation, atmospheric evaporation, rainfall, and the decay rates of every atom comprising our world.*

Plejaren scientists precisely comprehend the relative conditions of various planets, having studied literally millions of different inhabitable and uninhabitable worlds, and this comprehensive knowledge extends into the minute variations in the respective rates of decay of elements on each planet. Ptaah revealed as much to Meier during a conversation regarding the low-resonance of the present atmospheric conditions of Earth that "all materiality suckles itself up on… and in this way falls to early destruction, *as for example iron, which normally does not rust so fast as is the case on Earth*." (pg. 78) The significance of this short sentence can only be understood in the broader context of planetary infrasound resonance as it influences atomic decay. Physicists on Earth have only recently documented the correlation of notable fluctuations in the rate of decay in radioactive isotopes in the laboratory with solar flare activity:

> It's a mystery that presented itself unexpectedly: The radioactive decay of some elements sitting quietly in laboratories on Earth seemed to be influenced by activities inside the sun, 93 million miles away. Is this possible? Researchers from Stanford and Purdue University believe it is. But their explanation of how it happens opens the door to yet another mystery… There is even an outside chance that this unexpected effect is brought about by a previously unknown particle emitted by the sun. "That would be truly remarkable," said Peter Sturrock, Stanford professor emeritus of applied physics and an expert on the inner workings of the sun.
>
> The story begins, in a sense, in classrooms around the world, where students are taught that the rate of decay of a specific radioactive material is a constant. This concept is relied upon, for example, when anthropologists use carbon-14 to date ancient artifacts… But that assumption was challenged in an unexpected way by a group of researchers from Purdue University… Ephraim Fischbach, a physics professor at Purdue, was looking into the rate of radioactive decay of several isotopes as a possible source of random numbers generated without any human input... As the researchers pored through published data on specific isotopes, they found disagreement in the measured decay rates – odd for supposed physical constants.

> Checking data collected at Brookhaven National Laboratory on Long Island and the Federal Physical and Technical Institute in Germany, they came across something even more surprising: long-term observation of the decay rate of silicon-32 and radium-226 seemed to show a small seasonal variation. The decay rate was ever so slightly faster in winter than in summer…
>
> On Dec 13, 2006, the sun itself provided a crucial clue, when a solar flare sent a stream of particles and radiation toward Earth. Purdue nuclear engineer Jere Jenkins, while measuring the decay rate of manganese-54, a short-lived isotope used in medical diagnostics, noticed that the rate dropped slightly during the flare, a decrease that started about a day and a half *before* the flare.
>
> If this apparent relationship between flares and decay rates proves true, it could lead to a method of predicting solar flares prior to their occurrence, which could help prevent damage to satellites and electric grids, as well as save the lives of astronauts in space. The decay-rate aberrations that Jenkins noticed occurred during the middle of the night in Indiana – meaning that something produced by the sun had traveled all the way through the Earth to reach Jenkins' detectors. What could the flare send forth that could have such an effect?

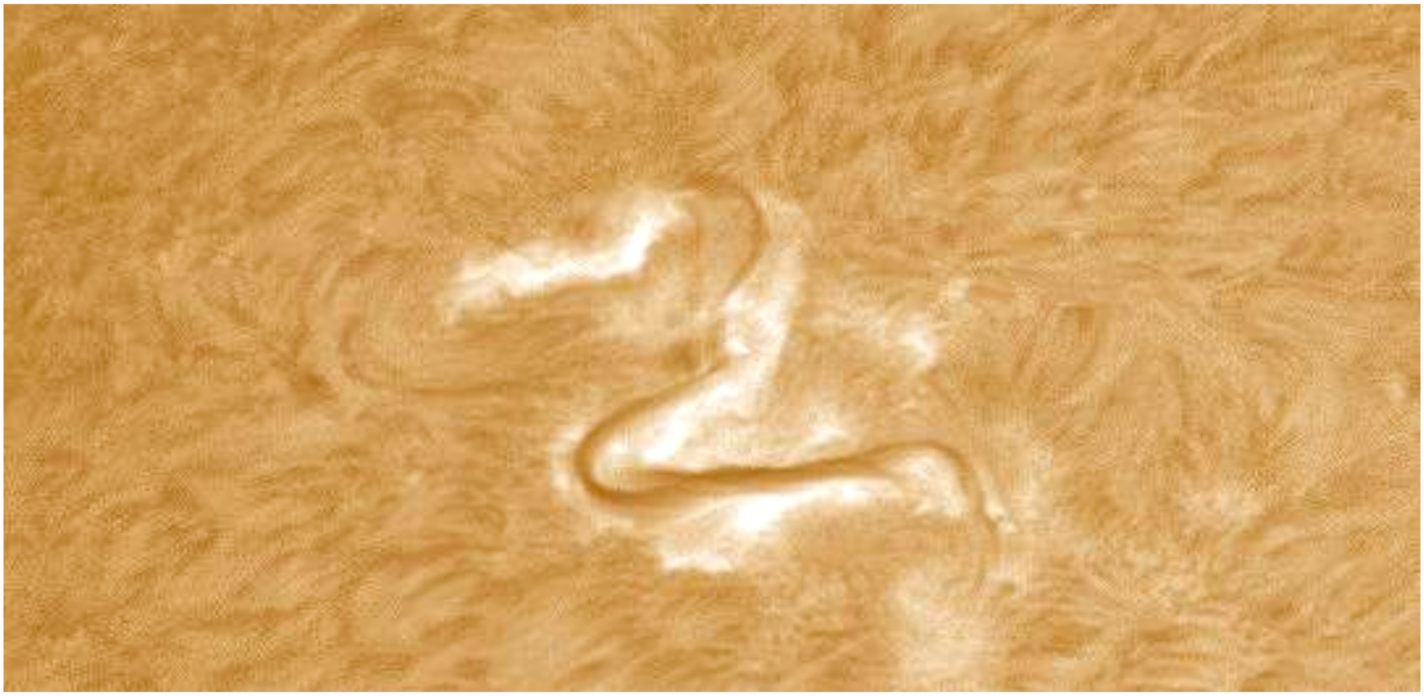

> Jenkins and Fischbach guessed that the culprits in this bit of decay-rate mischief were probably solar neutrinos, the almost weightless particles famous for flying at almost the speed of light through the physical world – humans, rocks, oceans or planets – with virtually no interaction with anything… "It doesn't make sense according to conventional ideas," Fischbach said. Jenkins whimsically added, "What we're suggesting is that something that doesn't really interact with anything is changing something that can't be changed."
>
> "It's an effect that no one yet understands," agreed Sturrock. "Theorists are starting to say, 'What's going on?' But that's what the evidence points to. It's a challenge for the physicists and… the solar people too."[57]

In fact, the subtle variations in atomic decay data are not an effect of neutrinos, *but the direct result of measurable surges of localized infrasound resonance*, just as in the case of the ‘aberrant’ fluctuations of planetary gravity fields and the patterns observed in the water flow rates of major rivers. *The fundamental integrity of the measured radioactive isotopes is directly affected by the incident infrasound radiation.* The decay rates measured by Jenkins at Purdue show a significant fluctuation specifically because of the location of his laboratory in relation to the Great Pyramid of Giza. The Purdue University campus in West Lafayette, Indiana (40.43°N 86.91°W) is 6,176 miles from Giza, a distance comprising *24.8%* of the Earth’s mean circumference. Infrasound standing waves focused by the Great Pyramid into the central Indiana area not only alter decay rates in Purdue labs, but also levitate huge sandstone boulders into treetops in nearby Limon[58] and Yellowwood[59] Indiana State Parks at the *25.0%* distance - in the same antigravity resonance process employed by the crystalline metals of interstellar Plejaren beamships.

The invisible web of terrestrial infrasound standing waves that protectively encircles the planet becomes focused by the three Giza pyramids, and in the same way by the three-ringed hull of Semjase's 7-meter beamship. The complex designs of every Plejaren spacecraft are precisely configured to interact within the Fibonacci structure of standing waves to achieve highly efficient acoustic energy storing within the crystalline latticework of resonant alloys comprising the circular body. This same effect occurs in natural crystalline structures – *in stones*. A remarkable photograph taken on July 19, 2009 by Gang Hao captures a special effect of acoustic resonance (above).[60] This giant boulder is estimated at 5m in diameter, which suggests a weight of several tons. The family never noticed it flying overhead while taking the photos, suggesting that the boulder was captured while slowly floating during an event of acoustic levitation, actually caused by solar flares transduced into infrasound by the Great Pyramid.

The piezoelectric properties of quartz, the main constituent of the sandstone boulder, allows its levitation by focused infrasound standing waves. The internal reflection of the parallel faces of the quartz crystals allows the acoustic energy to be stored inside the stone. The build-up of vibratory energy within the quartz lattice creates an electromagnetic field reducing the stone's weight until it becomes entirely weightless, and then slowly levitates along the infrasound standing wave's arch. The source of the focused standing wave energy levitating these stones is the Orion pyramids at Giza, exactly 5,240 miles from Gushan (or Gunchangzhai), China (*25.02°N* 118.51°E). This distance is precisely *21.05%* of the Earth's mean circumference of 24,892 miles. Gushan is also situated 1,245 miles from Angkor Wat - a resonant distance that is precisely *5.00%* of the Earth's mean circumference.

The whole number percent values of the site distances confirms the antigravity formula as applied. Evidently, the quantum physics of standing wave resonance underlies all of the transdimensional Plejaren technologies, as standing wave resonance affects both the levitation of spacecraft and the 'levitation' of electrons, protons and neutrons through the atomic framework. The provision of subatomic particles allows low-energy nuclear change and the multi-staged production of precious metals and rare elements such as vanadium, thulium and rhenium. These exotic elements are costly to mine and only rarely find application in human industry today, yet the advances of phonon resonance beckon a new era of harmonious human creativity, promising to expand the reach of human consciousness by allowing interstellar flight and bringing an end to the destructive and dangerous practice of ore mining. The broad set of transdimensional capabilities of Plejaren beamships include time travel – affording their ability to preview future events. Billy Meier has published 100% accurate predictions of a multitude of major scientific discoveries and major world events – *yet without having published any erroneous predictions*:

- 29th Contact, July 7, 1975, and 31st Contact, July 17, 1975: Contact involved specific information about Venus, unknown at the time, including composition of atmospheric gases, surface temperatures, depth of clouds, wind speeds, atmospheric pressure, coloration, variation in terrain, etc…
- 31st Contact, July 17, 1975: Semjase informed Meier that Mt Chimborazo, Ecuador, is the highest mountain on Earth [as measured from the center of the planet]… contrary to Meier's belief that it was Mt Everest…
- 1978: In a Wasserman publication, Meier foretold the launch of a telescope at end of the 1980s that would make unfathomable discoveries in space, and wrote that an asteroid would be discovered in the late 1980s to early 1990s; it would be named Toutatis, and may threaten Earth in September 2004…
- 1978: In *Existing Life in the Universe*, Meier [described] 2 small planetary bodies outside Pluto's orbit…
- 115th Contact, October 19, 1978: Meier described existence [and composition] of Jupiter's rings [and moons]...
- Jonestown massacre (Nov 18, 1978), Overthrow of the Shah of Iran (Jan 1979)
- Chinese invasion of North Vietnam (Feb–Mar 1979), Iran Hostage Crisis (Nov 1979–Jan 1981)
- Soviet invasion of Afghanistan (Dec 1979), Terrorist attack on Iranian Embassy, London (Apr–May 1980)
- Abdication of Queen Juliana of The Netherlands (Apr 1980), Death of Tito of Yugoslavia (May 4, 1980)
- Mount St. Helens eruption (May 18, 1980), Assassination of Indian PM Indira Gandhi (Oct 31, 1984)
- 215th Contact, February 28, 1987: Known as the Henoch (or Enoch) Prophecies, this contact contained a forewarning of the destruction of the World Trade Center (WTC) by terrorism, the series of worldwide wars that the US would subsequently launch, and military actions involving Russia, China, France, Germany, Spain, England, Scandinavia and many other countries.[61]

Repeated, accurate predictions of major world events, often years prior, proves Meier's foreknowledge! He also received information from the Plejaren leader Ptaah regarding the dramatic events of December 22, 2012, as later disclosed during the 476th contact of February 3, 2009 – on Eduard's 72nd birthday:

Meier: Regarding the Mayan calendar, [engraved as a sacred mandala in resonant piezoelectric stone, above,] it must be said that the year 2012 will bring a whole series of quite special events, and… our sun, will contribute to them, because on the sun's surface enormous eruptions will occur and solar storms are caused, which will be very intensive and will influence Earth's geomagnetic balance.

The Earth's magnetic field can suffer very strong fluctuations through the very strong X-ray radiation falling down on it. And through such geomagnetic storms, *power grids and all electronic appliances can collapse, thus also those in satellites, and the ISS can also suffer great damage or can even fail completely.* Such solar storms can lead to enormous natural catastrophes, and in the worst case, to a polarity reversal of Earth's magnetic field... What also has to be said is that such solar storms form to a climax in an exact rhythm of eleven years, which is normal, but in the year 2012 it can happen in an extraordinary magnitude. Earth will come under a very strong bombardment of hard radiation, which can have a very negative effect on the Earth's ozone layer. Due to the entire resulting process, *nitric oxides and acid rain can form worldwide*, which may have bad and devastating effects on the entire plant world. The enormous solar eruptions can lead to further very strong climate changes and, with it, to immense droughts, bad weather, volcanic eruptions, earthquakes, crop failures and, therefore, to even greater famines than have existed up to now. In addition to all this, the year 2012 brings other unpleasant events, such as a so far unknown, invisible, dark and huge space wanderer that is threatening from the fringe of our solar system and could wreak tremendous havoc on Earth.

And if this should really occur, the fact of its existence would be openly proven starting with 2010 or 2011, if at all, because the possibility of a "dark" and, therefore, non-observable passage of the colossus is also possible. Aside from that, all kinds of unpleasant machinations regarding unpeace, worldwide heavy warlike actions, as well as increasing degeneracy and excessive human ways of behaviour are threatening for the year 2012…

> *Ptaah:* You should not say more, because what you explained should be enough… Regarding the Mayan calendar, the 21st of December 2012 represents the culmination of a star constellation [alignment] that only occurs around every 26,000 years. The Mayan calculations are assuming that Earth - on the 21st of December 2012 - shall lie on an imaginary line that joins together the star [Betelgeuse,] above the left side of the constellation Orion with the Central Sun, [situated at] the centre of the Milky Way. The Sun meets the Milky Way at a location that is built by interstellar dust clouds, and is called the "dark cleft of the Milky Way" by the human beings of Earth. On the 21st of December 2012, at nightfall of winter solstice, the Sun is directly in that cleft, and it is in such a position that the Milky Way covers the horizon in all positions all around.
>
> Thereby the optical impression is produced that the Milky Way would touch Earth all around and that the galaxy would lie directly on the Earth. On the whole that's the end product of the Mayan calendar recording. This should suffice, my friend. To say more would be too much, because it would only foster the fear of the human beings of Earth, which should not be, however. So keep silent on all the further explanations that we have given you...[62]

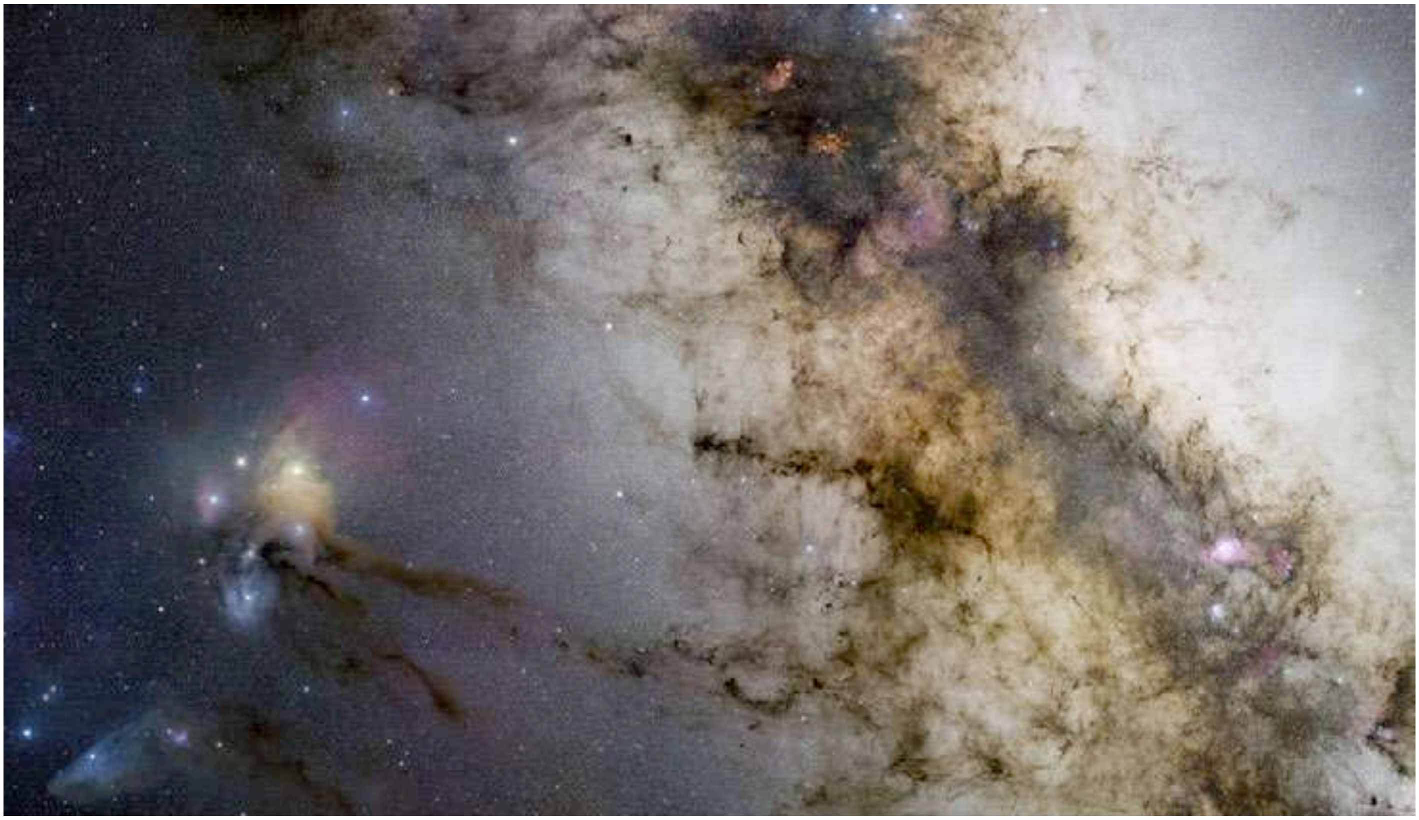

Through the tireless efforts of Eduard Meier, the wisdom traditions of the Plejaren culture reveal that the ongoing Earth changes are a result of the Golden Radiation emitted from the Central Sun, which lies at the core of our Milky Way galaxy. This Central Sun is obscured from our direct view by the 'dark cleft of the Milky Way' (above). Specific details regarding these electromagnetic field changes now occurring on Earth are most intensely experienced at infrasound standing wave convergence points which are already destroying all electrical equipment, just as discussed by Meier and his Plejaren contacts. The electrical overload of modern technology is also being experienced on the International Space Station, in accordance with the information from Meier and echoed in Native American prophecies of the Hopi.

Billy's brief mention of the *'invisible, dark and huge space wanderer'* references the presence of a binary twin companion of our sun - *a brown dwarf star* that the Hopi prophecies describe as reigniting to be seen as a small red star known to them as the Purifier, which will be the sign of the Purification Day, when all life on Earth will be effected. *The wave of energy described by the Plejaren as the Golden Radiation is the crucial factor linking these events involving the red supergiant star Betelgeuse with the Central Sun on December 22, 2012, as the wave of Golden Radiation passes through our Sol system.* Another chaotic aspect of the near passage of the dark brown dwarf star is the disturbance it has already caused within the Oort Cloud, before it becomes reignited as the Red Purifier of Hopi Prophecy.

The topic of the origin of comets within our solar system was addressed during the 150th contact of October 10, 1981, concerning the nearby Dark Star and its cyclical dislodging of comets from the Oort Cloud accounting for 97% of cometary debris entering Earth's vicinity. While the complete details of the close passage of the invisible brown dwarf binary twin of our visible Sun (referred to as a Dark Star) have been wisely withheld, the breaching of the Oort Cloud presently threatens all of Earthly humanity:

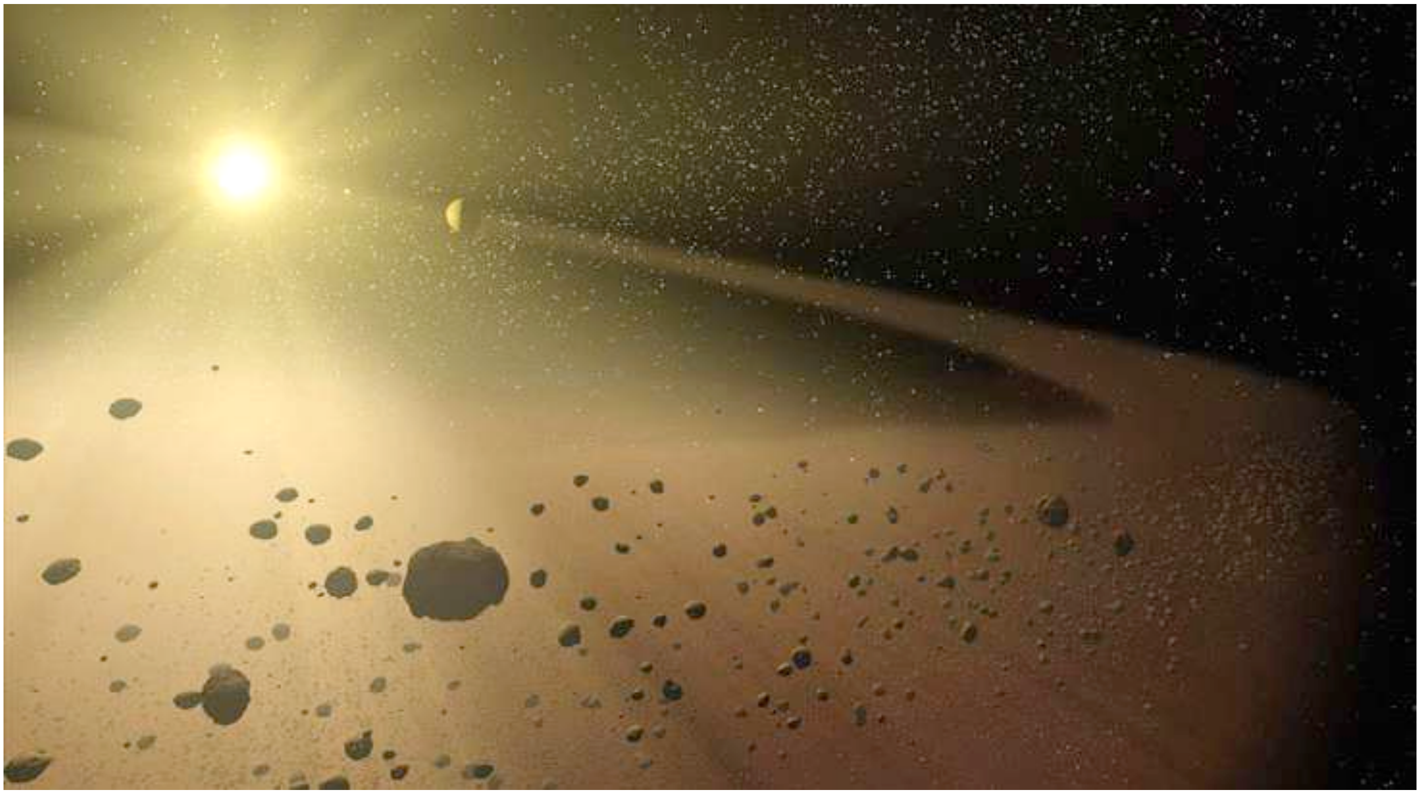

> *Meier:* In the coming years and decades, there will again appear frequent passages of comets, including those such as Halley's Comet, which can practically be observed in the sky… in broad daylight… [W]hen you took me in 1975 on the great journey, I saw beyond our solar system and planetary system the immense belt of material of frozen acids, rocks… and lifeless planets; some were quite large. Like the asteroid belt beyond the orbit of Mars, this belt also moves around the Sun but very far outside of Pluto's orbit.
>
> At that time, you explained to me that celestial bodies of all kinds pass through this belt, such as Dark Stars and other space bodies, and that every now and then, smaller or larger bodies tear out from this belt and enter into another orbit around the Sun, which can then partly be observed as comets from the Earth and, thus, can be seen. If I remember correctly, you told me at that time that on the one hand, this belt was still widely unknown to the Earth people in its magnitude but… more would soon be discovered… Furthermore, you explained that this material belt consists of the remains of the formation of the solar system and of immigrated materials from space, etc. Now, I wonder how high the percentage of comets is that appears in our solar system and that arises from this material belt beyond Pluto's orbit?
>
> *Quetzal:* It has probably escaped your attention, but this belt was discovered in the meantime and is named 'Oort,' after a Dutch astronomer. In accordance with its discovery, the discovered belt is called the Oort Cloud in Earth's scientific circles, but that's about it because the connections of this belt to the comets of the Sol system are still, to my knowledge, foreign and unknown to the Earth scientists.
>
> But the truth is that about 97% of the SOL system comets originate from this belt, as in this belt, planetary bodies passing by upset the trajectories of the materials through their attractive forces, so then these suddenly scramble, which removes isolated larger and smaller objects, pushing them into new orbits around the solar system, where they then appear as comets, while others fly out far from the Sun into free space in order to disappear somewhere in the vastness of the universe, often referred to as migrant dark planets, which will partly be captured again by major stars as satellites and then orbit them, far away from the new mother stars. But others will rush uncontrollably as dark objects traveling through space.[63]

Studies informing the link between Oort comets and a dark binary twin[64] have been published by U. of Louisiana researchers[65] and the Binary Research Institute,[66] and subsequently reconfirmed in 2010.[67] *The rain of Oort comets has already begun to impact the Earth on a daily basis.* Smaller fragments are creating dramatic aerial explosions worldwide, as witnessed over South Sulawesi, Indonesia on October 8, 2009. Sizable craters attest to large impacts in Hidalgo, Mexico on February 10, 2010; San Joaquin (Santander), Colombia on September 6, 2010; and San Bruno, California on September 9, 2010. The prophetic date of December 22, 2012 was *indirectly* addressed during the 52nd contact of May 17, 1976, in connection with precise astronomical calculations decisive in the building of the Giza pyramids *about 73,350 years ago* (spanning two cycles of 36,650 years, as Ptaah later explained in the 256th contact[68]):

> *Semjase:* When the pyramids were built, the then astronomers had already calculated the further fate of the Earth. The measurements of the pyramids were elaborated according to their calculations… [which] revealed that many thousands of years in the future, a catastrophe would menace the Earth and that from the cosmos, like at the time of the construction of the pyramids. To reveal to the later inhabitants of the Earth over the later millennia… the measured dates in astronomical form were built into the pyramids for the far distant future… when the catastrophe from the cosmos threatens anew. The ancient astronomy was very able, and… even considered in their calculations, several world crashes, and could reckon the dates with very exact precision.
>
> As pre-calculated, the Earth did then convert, and agreed with the then and since long gone science. Now slowly the time becomes mature, that the astronomically measured dates, which as mentioned, served in the construction of the pyramids and were decisive in this, agrees with the astronomical values of the present, and announce the more than 70,000 years ago prophesied events this way. And this event will be exactly then, when the solar light of a very far distant star, a central star, falls through the tube-like Opening of Revelation, which draws itself from the outside of the Giza pyramid into the center, in a straight uninterrupted line, and illuminates one certain point. I am not allowed to tell more about this.[69]

Many insights can be gleaned from Semjase's foretelling of the activation of the Great Pyramid by the radiations of the galactic Central Sun, and the catastrophic impacts of Oort comets dislodged by the Dark Star's passage. *The Giza pyramids were masterfully designed for repeated reactivations by cyclical blast waves that radiate from the core of the galaxy to impact Earth in a precisely calculated rhythm, causing the magnetic reversal of the entire Sol system and a dramatic increase in terrestrial infrasound levels.* Worldwide auroras can thereby be expected, as well as increased evaporation and precipitation rates globally, and a measurable reduction in the atomic decay rates of all Earthly matter. All cosmic signs point toward the imminent reemergence of a 'thick white vapor envelope surrounding the globe'.

Earthly astronomy can no longer ignore the nearby presence of the brown dwarf binary twin of our sun. Nor can corporate news continue to hide or dismiss the multitudes of reported meteor strikes now inundating the Earth, *exactly as foretold by Eduard Meier and his knowledgeable informants decades ago.* The existing corporate and religious institutionalization of psychopathy cannot withstand the great revelations occurring at the present time, and will soon lose its despotic hold of human consciousness:

> *Ptaah:* When hitherto Semjase only spoke of Earth religions in negative form, then it was because the Earth human had to be made attentive to the error in his religions, ...for these are like nowhere else in this Universe… They control the Earth human and the whole planet, and retard your world thousands of years, in Spiritual respect… [Y]our religions still contain real worth, but this becomes so indiscernible that only few people are able to see the effective truth there. These few who are able to see the real truths from the scripts, are banned and confounded by those dependent on these wrong beliefs, and thus cannot spread the truth. When Semjase speaks in negative form about Earth religions, she is addressing the falsifications in them which have been built up partly consciously. The Earth religions are only called such in name, but are just degenerate cults."

In truth, the same small minority of fallible psychopathic individuals control world religions, economics, sciences and politics –those engaged in the destructive subterranean nuclear blasts, damming of waters and extraction of oils– are themselves also actively inducing earthquakes using the HAARP array to terrorize and extort the rest of humanity. *Awareness of the hereditary genetic disorder that results in psychopathic behavior demands responsible action from the conscious human beings of Earth.*

The holistic teachings of Plejaren Elders reveal that genetic enhancement through biorhythmic entrainment holds the highest potential for the reversal of all known mis-developments of the physical body or the spirit. *Natural law provides for the bioelectrification and illumination of the body's waters, using golden celestial water comprised of pure protium in conjunction with tri-thalamic heartbeat entrainment during fetal development and throughout life, thereby removing all negative influences from the crystallized human being:*

> *Quetzal:* [T]he fallible ones could be completely cured of their criminal or other degenerate forms through genetic manipulation and be released into society. This gene-manipulative healing method was further investigated and further developed [by Plejaren biophysicists many thousands of years ago], and indeed, even in reference to all hereditary diseases and immunodeficiencies as well as physical and organic deformities, but also in terms of disorders of consciousness and inheritable psychological disorders, etc. The development progressed rapidly, so already soon, early detections of gene damages of the aforementioned kind became possible, which naturally led to the fact that these cognitions were used and the genetic deformities, genetic abnormalities, and other harmful genetic influences were already repaired in the growing child in the womb.
>
> Thus, an irrefutable law has been in force since then, namely that every child, so every girl and every boy, is to be arranged into a gene examination, i.e. a gene analysis, at the prenatal age of seven months. Since then, since this law gained validity and became observed absolutely, it is valid with us that every crime or other degeneration as well as every inheritable disease of the body and its organs, every immunodeficiency of any kind, and every form of harmful aggressiveness, jealousy, and vindictiveness are just as non-existent as also hatred, envy, desire for validity, egoism, and many other characteristics that bring the human being discord, unkindness, and other harms. But at the same time, it should be noted that through this, our own decision in all matters and concerns is in no way impaired, as well as not our drive for self-preservation and our drive for self-defense, when this is demanded.[70]

Until the coming times when this biorhythmic genetic healing becomes active through the world's infrasound network of resonant pyramids, the survival of humanity requires our careful discernment of psychopathy in human leadership. The Prophecies of Henoch, forewarning of the dire consequences of human ignorance on Earth, were presented by Quetzal during the 215th contact on February 28, 1987:

Quetzal: Before I give you a clear account of the prophecies of Henoch, I would like to point out that prophecies are always changeable and can be changed for the better if man makes positive changes in his thoughts, feelings and actions, leading to that which is better and positively progressive…

However, this does not apply to predictions, as these rest upon events that cannot be changed, are inevitable and surely and definitely will occur in the future. Predictions rest upon a preview and thus on a direct viewing of the future, and have to do with neither prophecy nor with calculation of probability. So when I make a portion of Henoch's prophecies for the third millennium known to you, it does not mean that they have to be fulfilled, because the prerequisite of fulfillment in each case would be that the already existing causes continue to exist as also continue to be created in the future so that a fulfillment of the prophecies can come to pass.

Thus, provided that human beings of Earth will become reasonable, the possibility exists that by a reasonable change in the way of thinking as well as a reasonable development in feeling and an equally reasonable way of acting, everything changes for the better and positive, whereby prophecies do not have to be fulfilled. However, if this transformation does not occur, a very evil, wicked and negative time lies ahead for the Earth and its entire population in the coming new millennium…

The point in time at which these prophecies will begin to be fulfilled will be when a Pope will no longer reside in Rome. All of Europe will then fall victim to a terrible punishment by evil powers. The Christian religion will collapse and the churches and monasteries will end up in ruins and ashes… Monstrous forces will be created by science and will be released by the military forces and armies as well as by terrorists, causing great destruction. Millions and even billions of people will be killed by acts of terrorism, by wars and civil wars; and finally, in some parts of the world, every third human being, and, in other places, every fourth human being, will lose his or her life. The nations of the East will rise against the nations of the West, the West against the East…

The people will be completely powerless against all this and will live through 888 days of Hell on Earth, suffering hunger and plagues, which will claim even more lives than the war itself. The time will be severe as never before experienced on Earth. Ultimately, nothing can be bought or sold any longer. All provisions will be rationed; and if a human being steals even a small piece of bread, he/she will have to pay for it with his/her life…

And it will be that the fanatics of Islam will rise up against the countries of Europe and all will shake and quiver. Everything in the West will be destroyed; England will be conquered and thrown down to the lowest level of misery. And the fanatics and warriors of Islam will retain their power for a long time. However, not only Europe will be affected but ultimately all the countries and peoples of the Earth, as the great horror expands to a war that will encompass the entire world.

After the turn of the millennium, the papacy will exist only a short period. Pope John Paul II is the third from last in this position. After him, only one additional pontificate will follow [Pope Benedict]. Then a Pontifex Maximus follows who will be known as Petrus Romanus. Under his religious rule, the end of the Catholic Church will come, a total collapse becoming inevitable. That will be the beginning of the worst catastrophe that will ever have befallen the human beings and the Earth…[71]

These profound passages form only a portion of the Prophecies of Henoch, and follow from earlier information about the lineage of popes provided to Meier on August 6, 1978 during the 113th contact:

> Karol Wojtyla… is the 3rd to last pope before the world change [of December 22, 2012]. He is the 264th pope. He will die in the near future. The world will be astonished when the next pope [Joseph Ratzinger] is killed also. Then the pope with the number value of five [is coronated]. He is the 266th pope - he is the last one.[72]

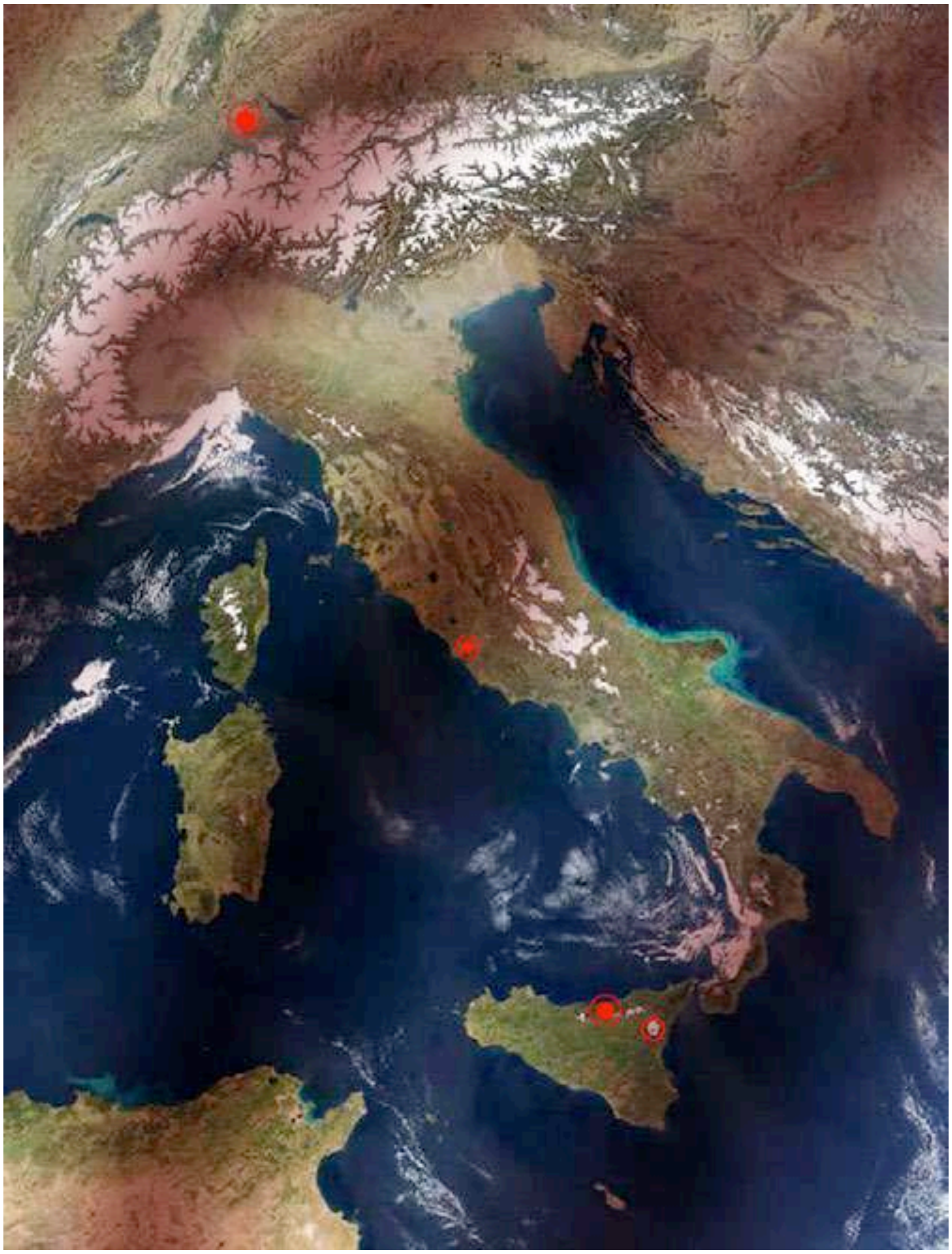

After the murder of Pope Benedict XVI, "a Pontifex Maximus follows who will be known as Petrus Romanus. Under his religious rule, the end of the Catholic Church will come, a total collapse becoming inevitable." These very detailed, ancient prophecies from the Henoch lineage directly coincide with those attributed to St. Malachy, an Irish monk who produced a comprehensive list in the year 1139 which prophesied every future pontiff of the Papal lineage from that date until today. St. Malachy wrote: "In the final persecution of the Holy Roman Church there will reign Petrus Romanus, who will feed his flock among many tribulations; after which the seven-hilled city will be destroyed and the dreadful Judge will judge the people." Not only is the same name given for a proclaimed 'final Pope', but the destruction of Rome, the 'seven-hilled city' is also given as a sign of the imminence of the collapse of the Roman Catholic Church. The Henoch prophecies foretell that "the point in time at which these prophecies will begin to be fulfilled will be when a Pope will no longer reside in Rome."

These foretold events *will* follow a pattern of fulfillment due to dramatic *infrasonic surges* that require the imminent evacuation of Rome and the Vatican, *which will inevitably be overcome by recurring spates of piezoelectric fires,* just as documented in the towns of Messina, Sicily, near the active Mt. Etna volcano.[73] The Vatican (41.53°N 12.30°E in white, above) is 1,326 miles from Giza, Egypt – a distance that is 5.3% of the Earth's mean circumference – along the radial distance of infrasound standing wave convergence. The steadily growing intensity of planetary infrasound is already significantly affecting these fixed focal points, including both the location of the Vatican and the Semjase Silver Star Center. Infrasound waves have been recorded in the atmosphere above Switzerland that reveal an inaudible ultra-low frequency field of concentric circles, as documented during an infrasound resonance event on April 24, 2009:

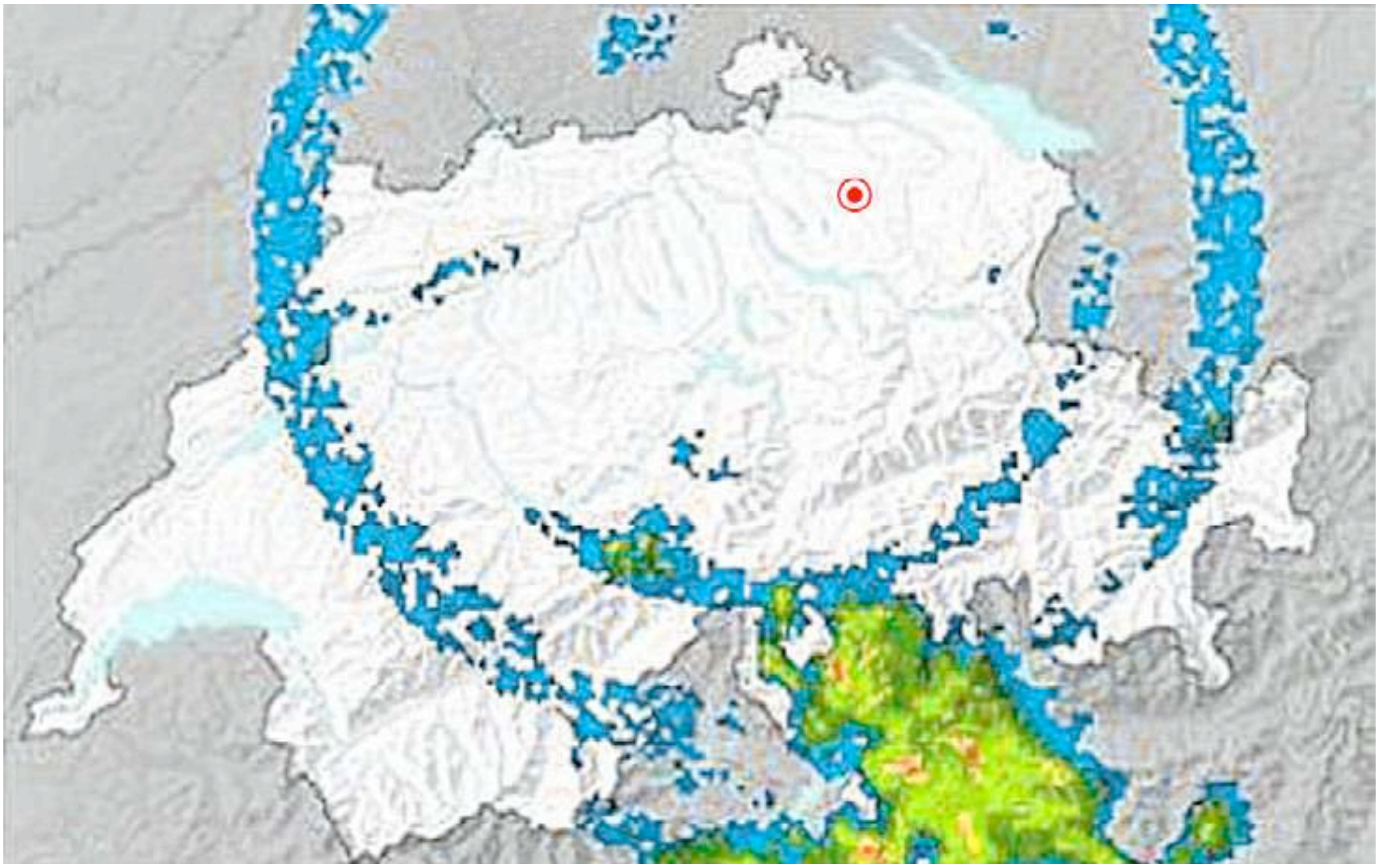

> Did a country-sized UFO hover above Switzerland last month? On the morning of Friday the 24th of April of this year employees at MeteoSwiss [observed] huge circular fields hovering over Switzerland… [that] covered the entire country… MeteoSwiss claimed… that they were merely caused by the reflection of water droplets. A few minutes after being queried on the issue the radar images were hurriedly removed from the department's website. Are we being told everything the Swiss meteorological knows about these radar images? [74]

Meteorologists in Switzerland have never seen radar anomalies like this specifically because they are hidden and scrubbed from archives, as in this particular case. Hundreds of similar radar anomalies are documented by websites that gather the data before it's purged from online archives. The only viable explanation for these giant radar rings is water vapor suspended by concentric rings of standing waves in the ultra-low frequency range. Nikola Tesla was the first to scientifically describe standing waves in his Colorado Springs lab[75], later reporting: "In Colorado I succeeded one day in precipitating a dense fog. There was a mist outside, but when I turned on the current the cloud in the laboratory became so dense that when the hand was held only a few inches from the face it could not be seen." [76] The standing waves that fogged-out Tesla's laboratory are now suspending these giant concentric circular rings of water vapor above the entire country of Switzerland – as on April 24, 2009. *A circumference of the Earth taken along the northwest/southeast axis of the Great Pyramid of Giza, Egypt passes exactly through the center point of the MeteoSwiss radar circles.*

The center point of these gigantic radar circles lies just southwest of the Semjase Silver Star Center (overlaid in red, above) in Hinterschmidrüti, Switzerland (47.416°N 8.906°E); both sites being located along the 6.7% resonant distance of about 1,680 miles from Giza. This precise resonant alignment of the Semjase spiritual center in Switzerland reflects the same advanced understanding of the Fibonacci geometry of infrasound standing waves that informs the locations of all of the ancient pyramids of the world, now approaching the time of reactivation with the magnetic reversal of December 22, 2012. The meticulous work of Eduard Meier to record his extensive contacts with the Plejaren civilization guides humanity by the light of *Truth, reorienting our consciousness toward the radiant lessons of the Creation witnessed within ourselves and our mother Earth during this time of the great shift.*

References

---

[1] Stevens WC Ed. (1990) *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 2’* Genesis III Publishing, pp. 257-258
[2] Meier E (1988) ‘OM: The Book of Truth’ FIGU Press, Canon 43: Verses 153, 167-177, pp. 354-355
[3] Putney A (2008) ‘∅’ Human Resonance Productions, pp. 23-24
[4] -- (2009) *‘Neem Oil’* Sister Zeus, online · http://www.sisterzeus.com/neem.html
[5] Meier E, Tr. Stevens B (2010) ‘Contact Report 260 Translation’ *Billy Meier Wiki - The Future of Mankind*, online · http://www.futureofmankind.co.uk/Billy_Meier/Contact_Report_260
[6] Stevens WC Ed. (1988) *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 1’* Genesis III Publishing, pp. 150-154
[7] Ibid, p. 385
[8] -- (2006) ‘Semjase by Dejan’ online · http://www.futureofmankind.co.uk/meier/gaiaguys/meier.plejarenportraits.htm
[9] Stevens, op cit, pp. 48-49
[10] Stevens, op cit, *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 2’* pp. 87-89
[11] Stevens, op cit, *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 1’* p. 241
[12] Stevens WC Ed. (1995) *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 4’* Genesis III Publishing, p. 324
[13] Ibid, p. 281
[14] Phillips T (2004) ‘Strange Space Weather Over Africa’ *Science At NASA*, November 14, online · http://science.nasa.gov/headlines/y2007/13nov_africa.htm
[15] Stevens, op cit, *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 2’,* pp. 54-59
[16] Stevens WC Ed. (1992) *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 3’* Genesis III Publishing, pp. 211-214
[17] Xie W A, Wei B (2001) ‘Parametric Study of Single Axis Acoustic Levitation’ *App. Phys. Lett. 79:6, pp. 881-883*
[18] Horn M (2006) ‘Beamship Sounds’ *They Fly*, online · http://theyfly.com/sounds/BeamshipSounds.mp3
[19] Stevens WC (2000) *‘UFO Sound Recordings: Semjase’s Beamship During a Flight Demonstration’* pp. 1-2, online · http://www.theyfly.com/PDF/UFOSoundRecordings.pdf
[20] Kruger A (2004) ‘Construction And Deployment Of An ULF Receiver For The Study Of Schumann Resonance In Iowa’ online · http://www.ia.spacegrant.org/RES_INF/VRR2003/Kruger-SEED.pdf
[21] Meier E, Tr. Stevens B (2010) ‘Contact Report 223 Translation’ *Billy Meier Wiki - The Future of Mankind*, online · http://www.futureofmankind.co.uk/Billy_Meier/Contact_Report_223
[22] Stevens, op cit, *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 1’* p. 56
[23] Mullins J (2000) ‘Will Plasma Revolutionize Aircraft Design?’ *Space Daily*, October 28, online · http://www.spacedaily.com/news/future-00o.html
[24] Stevens, op cit, p.124
[25] Mills RL, Ray P, Dhandapani B, Nansteel M, Chen X, He J (2002) ‘New Power Source from Fractional Quantum Energy Levels of Atomic Hydrogen Surpasses Internal Combustion’ *Journal of Molecular Structure* 643, pp. 43-54
[26] Meier E, Tr. Stevens B (2010) ‘Contact Report 11 Translation’ *Billy Meier Wiki - The Future of Mankind*, online · http://www.futureofmankind.co.uk/Billy_Meier/Contact_Report_11
[27] Stevens, op cit, p. 289
[28] Novak P (2005) ‘Hurtubise Says Invention Sees Through Walls’ *Bay Today News*, January 16, online · http://www.baytoday.ca/content/news/details.asp?c=6657
[29] -- (2009) ‘Scientists Claim New State of Matter’ *Live Science*, July 28, online · http://www.livescience.com/technology/090728-new-state-matter.html
[30] Stevens, op cit, pp. 54-56
[31] Ibid, pp. 117-119
[32] Putney A (2009) *‘Veil of Invisibility’* Human Resonance Productions
[33] Stevens, op cit, *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 2’* p. 55
[34] Orion E (2003) *‘The Bush Connection’* The Bush Connection.com
[35] Stevens, op cit, p. 187
[36] Meier E, Tr. Devine D, Legg V (2009) ‘Asket’s Explanations – Part 4’ *Billy Meier Wiki - The Future of Mankind*, online · http://www.futureofmankind.co.uk/Billy_Meier/Asket%27s_Explanations_-_Part_4
[37] Stevens, op cit, *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 1’* pp. 378-382
[38] Ibid, pp. 367-368
[39] Meier E, Tr. Stevens B (2010) ‘Pleiadian/Plejaren Contact Notes Volume 5’ *Billy Meier Translations*, online · http://www.billymeiertranslations.com/#dialog/divVolume5
[40] Meier E, Tr. Stevens B (2010) ‘Pleiadian/Plejaren Contact Notes Volume 4’ *Billy Meier Translations*, online · http://www.billymeiertranslations.com/#dialog/divVolume4
[41] Ketterle W, Zwierlein MW, Abo-Shaeer JR, Schirotzek A, Schunck CH (2005) ‘Vortices and Superfluidity in a

---

Strongly Interacting Fermi Gas' *Nature* 435, pp. 1047-1051

[42] Stevens, op cit, pp.155-156

[43] Ibid, pp.170-175

[44] Matsuoka S (1996) 'Ultrasonic Welding of Ceramic/Metal' *J. of Materials Processing Tech 47(1-2),* pp. 185-196

[45] Champion J (2007) 'Phonon Resonance Institute Home' *Dr Joe Champion*, online · http://drjoechampion.com

[46] Champion J (2007) 'Phonon Resonance' *Dr Joe Champion*, online · http://drjoechampion.com/phonon.htm

[47] Choi C (2010) 'Microbes Leave Gold on Corpses, May Complicate Forensics' *Live Science*, online · http://www.livescience.com/history/gold-corpse-hair-100225.html

[48] Clark MW, Yip ML, Campbell J, Abelson J (1990) 'SSB-1 of the Yeast Saccharomyces Cerevisiae is a Nucleoler-Specific, Silver-Binding Protein that is Associated with the snR10 and snR11 Small Nuclear *RNAs' Journal of Cellular Biology 111(5 Pt 1)*, pp. 1741-51, online · http://www.ncbi.nlm.nih.gov/pubmed/2121740

[49] Waldrop A (2008) 'Phonon Resonance Spreadsheet' online · http://drjoechampion.com/ResonanceCalcGraph.xls

[50] Horn M (2006) 'Documentation from IBM on Marcel Vogel's Patents' *They Fly*, online · http://www.theyfly.com/newsflash4/m_vogel.htm

[51] -- (2004) *'Through Space and Time: A Photo Journal of "Billy" Eduard Albert Meier'* Steelmark, pp. 66-67

[52] Keaten J (2007) 'Shrinking Kilogram Bewilders Physicists' *Associated Press*, September 12, online · http://www.humanresonance.org/kilogram.html

[53] Mauritsson J, Johnsson P, Mansten E, Swoboda M, Ruchon T, L'Huillier A, Schafer KJ (2008) 'Coherent Electron Scattering Captured by an Attosecond Quantum Stroboscope' *Physical Review Letters* 100, 073003

[54] Meier E, Tr. Stevens B (2010) 'Contact Report 248 Translation' *Billy Meier Wiki - The Future of Mankind*, online · http://www.futureofmankind.co.uk/Billy_Meier/Contact_Report_248

[55] Robbins G (2009) 'Ocean Smoke Puzzles Forecasters' *Orange County Register*, December 22, online · http://www.humanresonance.org/laguna.html

[56] Mauas P, Buccino A, Flamenco E (2010) 'Long-tern Solar Activity Influences on South American Rivers' *Journal of Atmospheric and Solar-Terrestrial Physics on Space Climate 1*, March, online · http://arxiv.org/abs/1003.0414

[57] Strober D, Jolagh C (2010) 'The Strange Case for Solar Flares and Radioactive Elements' *Stanford Report*, August 23, online · http://news.stanford.edu/news/2010/august/sun-082310.html

[58] Stoops R (2003) 'Unionville, Indiana –Large Boulder in a 40-ft. Tree' *Roadside America*, online · http://www.humanresonance.org/limon.html

[59] Hess J (2000) 'Unexplained Tree-top Boulders Found in Forest' *Brown County Democrat*, online · http://www.humanresonance.org/yellowwood.html

[60] -- (2009) 'Gushan's Mysterious Flying Objects (Photo)' *South East Express (Xiamen),* August 4, online · http://www.humanresonance.org/gushan.html

[61] Horn M (2004) 'The Henoch Prophecies from the Billy Meier Contacts' *Nexus Magazine 11(5)*

[62] Horn M (2009) '2012; What We Have Brought upon Ourselves' *They Fly*, online · http://theyfly.com/2012.htm

[63] Meier E, Tr. Stevens B (2010) 'Pleiadian/Plejaren Contact Notes Volume 4' *Billy Meier Translations*, online · http://www.billymeiertranslations.com/#dialog/divVolume4

[64] Blakemore P (2002) 'Does the Sun Have a Doomsday Twin?' *The Telegraph*, October 18, online · http://www.telegraph.co.uk/science/science-news/3300270/Does-the-Sun-have-a-doomsday-twin.html

[65] Matese J, Whitman P, Whitmire D (1999) 'Cometary Evidence of a Massive Body in the Outer Oort Cloud' *Icarus 141,* pp. 354-366, online · http://www.ucs.louisiana.edu/~dpw9254/MS7292.pdf

[66] Cruttenden W (2009) 'Introduction: Binary Companion Theory' *Binary Research Institute*, online · http://www.binaryresearchinstitute.org/bri/research/introduction/theory.shtml

[67] Matese J, Whitmire D (2010) 'Persistent Evidence of a Jovian-Mass Solar Companion in the Oort Cloud' *submitted to Icarus,* online · http://arxiv.org/abs/1004.4584

[68] Meier E, Tr. Devine D, Legg V (2007) 'The Truth About the Pyramids and Stonehenge' They Fly, online · http://www.theyfly.com/Pyramids_Stonehenge.htm

[69] Stevens, op cit, *'Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 2'*, pp. 288-289

[70] Meier E, Tr. Stevens B (2010) 'Contact Report 217 Translation' *Billy Meier Wiki - The Future of Mankind*, online · http://www.futureofmankind.co.uk/Billy_Meier/Contact_Report_217

[71] Horn, op cit, 'The Henoch Prophecies from the Billy Meier Contacts'

[72] Meier E (2005) 'One Hundred Thirteenth Contact' *Semjase.Net*, online · http://www.semjase.net/semjeng9.html

[73] Putney A (2008) 'Piezoelectric Fires in Messina, Italy' *Human Resonance*, online · http://www.humanresonance.org/messina.html

[74] Cohen M (2009) 'Did A Country-Sized UFO Hover Above Switzerland Last Month?' *All News Web*, May 2, online · http://www.allnewsweb.com/page6756758.php

[75] Tesla N (1978) *'Nikola Tesla: Colorado Springs Notes 1899-1900'* Nolit, Beograd, Yugoslavia, pp. 61-62

[76] -- (1917) *'Minutes of the Annual Meeting of the American Institute of Electrical Engineers, May 18'*

4

Urandir’s Transport

Urandir's Transport

Ongoing cosmic contacts occurring in Brazil have been largely obfuscated by corporate media disinformation, finding press coverage only in the Portuguese language broadcasts of the Brazilian media. Most significant among the many extraterrestrial contacts occurring in South American regions are those reported by Urandir Oliveira, who very carefully documents many forms of evidence attesting to the veracity of his statements and the reality of his transdimensional experiences. The weight of Urandir's evidence takes the form of demonstrations of the strong magnetic influence that he directs through his hands to move objects, in addition to physical trace evidence and video recordings of his levitation along violet-colored transport beams into luminous aerial plasma ships floating above. *At Urandir's special request, Verade TV videographers recorded and broadcast high-quality footage of his 1998 plasma beam transport into one of six disc-shaped plasma craft in an arc, each radiating pure white light.*

Urandir Fernandes de Oliveira was born on March 14, 1963 in Maraba Paulista, a rural area outside Sao Paulo in the state of Sao Paulo, Brazil. He clearly states that his unusual experiences with highly advanced extraterrestrial beings have involved their purposeful electromagnetic activation of bioenergetic capabilities that lie dormant in the average person. At a young age, Urandir had noticed the influence of his emotional states on metal objects in his hands. He quickly learned to develop this skill over the years, eventually being able to bend and tear common coins with his bare fingers. While he does not recall any physical interactions with the cosmic visitors during early childhood, he does attribute his unusual electromagnetic abilities to artificial plasma environments that he was voluntarily exposed to. The first face-to-face meeting that Oliveira recalls occurred on March 6, 1976, at the age of 13, when he was transported from his Sao Paolo home aboard a luminous lens-shaped craft. There he met with several tall, blond-haired humans of about 6'3"-6'5" in height, who radiated love, peace and intelligence through cat-like eyes, with vertically-oriented luminous golden pupils enclosed by blue, green or yellow irises.

The enhanced mental capabilities endowed by Oliveira's visitors allowed his further development of telepathic communication with the beings at a distance, by which method they would request and prepare him for subsequent contacts. The electro-acoustic activation of various Earth humans was explained to be a crucial mission of this extraterrestrial group, comprised of several allied races for whom Urandir has been given no names or planetary origins, but only recognizes their personalities telepathically. The practical urgency of Urandir's contacts involve the provision of crucial information regarding the electromagnetic changes now occurring on Earth, to prevent disastrous consequences that include the destruction of all nuclear power stations and nuclear arms storage facilities by the electrical overload and superheating of metals at specific focal points of infrasound resonance. Several contacts were conducted over many years to prepare Urandir for the complex astronomical information he would later receive:

By age eight, Urandir realized that when he fought over food with his brothers and sisters at the kitchen table, his hands could bend and twist the metal utensils. He remembers once throwing a fork that kept bending as it hit the table top… Urandir recalls physical abductions from Earth locations into an extraterrestrial craft when he was 13, 23, 27, 35 and 39, plus an encounter in Sao Paulo with ET humanoids at age 23.

Age 13, March 6, 1976 - Urandir Oliveira's Extraterrestrial Abduction #1, Sao Paulo

When he was thirteen on March 6, 1976, at 10pm at his home in Sao Paulo, he had his first encounter with a craft. The beings were tall and blond-haired. Their eyes were like a cat's, vertical honey-colored pupils in blue, green or yellow irises. He understood that the frequency of his mind had been speeded up by the blond humanoids to improve communication. Since then, one of the consequences has been that Urandir Oliveira's mind control over matter resembles the powers of metal bender, Uri Geller, who linked his abilities to a childhood encounter with a glowing being.

Age 23, 1986 - Urandir Oliveira's Extraterrestrial Encounter in Sao Paulo

Urandir had been making money by entertaining people with his mental powers in Sao Paulo. Then one night, he came upon a beggar and child who were hit by a car and broken bones were exposed through their skin. Not knowing how or why, he bent down and placed his hands on their terrible wounds and the bleeding stopped. Shortly thereafter, tall blond beings approached him and said, 'Now you know what to do with the light energy.'

Age 27, 1990 - Urandir Oliveira's Extraterrestrial Abduction #2, Sao Paulo

In 1990, he and a friend were both lifted up in beams of light into a craft where Urandir remembers many crystals pulsating with light and ‘lightning.’ He understood that the pulsating crystals ‘activated human force fields.’

Age 35, 1998 - Urandir Oliveira's Extraterrestrial Abduction #3, Corguinho Farm

> In 1996, Urandir bought 209 acres of farmland in Corguinho between hills and a mesa where local farmers for decades have seen mysterious lights and aerial craft. His parents and in-laws also moved to the farm to help with the work. Over the next couple of years, Urandir appeared on several Brazilian television programs to demonstrate his Uri Geller-like ability to use his mind to bend metal, shut off and turn on lights, to move and control the direction of a small metal ball and to mentally break metal coins and glass dishes. By 1998, Brazil Verade TV traveled to his Corguinho farm because Urandir informed the station that the blond beings were coming back to get him again. The crew was videotaping when six lights in an arc appeared in the night sky and descended toward the ground. As the arc of lights seemed to lower toward Urandir, he walked through them into a small beam of light that lifted him into a disc-shaped craft while the Brazil Verade television camera rolled. Some of that video was broadcast in Brazil and the tape was sent to CNN for international broadcast. But to date, Urandir does not think CNN ever aired any of it and he wonders what happened to that professional videotape of his second abduction.
>
> What no camera had access to was the inside of the craft. There, tall, blond humanoids showed Urandir wall screens containing moving images of Earth destruction by nuclear missiles and nuclear power reactors. Urandir understood that the destructions will be caused by changing Earth magnetic fields which affect the electronic systems in the missiles and reactors. The blond beings also showed Urandir that three "fully functional" pyramids currently covered by sand will be uncovered for the world to see.

Age 39, September 15, 2002 - Urandir Oliveira's Extraterrestrial Abduction #4, Corguinho Farm

> His fourth abduction encounter began after 7 p.m. local time on September 15, 2002, in the bedroom he and his wife, Jessica, shared in their farm house. Urandir said he had received telepathic information several days before from the beings that by 10 p.m. on September 15, another transport to a humanoid craft would occur. So, anticipating another interaction, Urandir told his parents and in-laws, as well as local friends, to be ready for "a rain of rocks" that would last for seven minutes when Urandir was taken to the non-human craft. That same week, Jessica had traveled with their 2-year-old son, Aria, to Sao Paulo, to visit family and was not there when the drama unfolded.[1]

The natural energy fields of Earth resonance are notably strong in the entire Corguinho area, one indicator of which can be easily observed in the red soil that is given its color by iron, just as is the case with our own blood. The high electrical conductivity of iron in the blood is also witnessed in iron-rich soils, which conduct ambient electrical currents in the ground transduced by the bedrock from deep, inaudible pulsations of infrasound standing waves. Oliveira’s earlier extraterrestrial experiences in the city of Sao Paulo prepared him for the later development of a spiritually and ecologically conscious community within which Urandir would be able to spread the knowledge and experiences he had gained from his contacts.[2] The electromagnetic setting of the Corguinho area provides the optimal environment for telepathy.

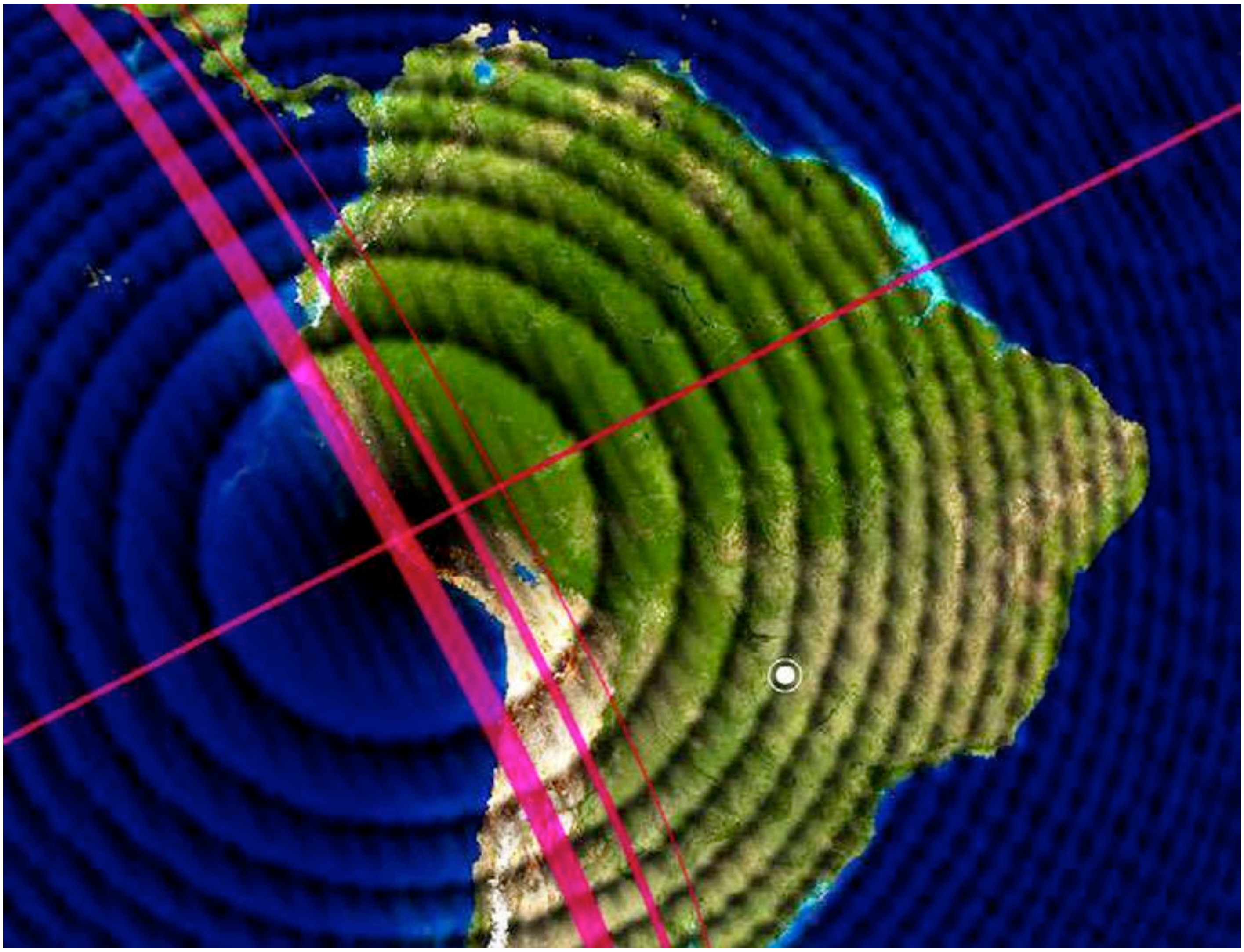

As with many conscious individuals seeking a sustainable life-path in accordance with ancient indigenous wisdom concerning these times of great change, Urandir and his wife Jessica have chosen to relocate to a resonant location to begin a sustainable organic farming community. Just as in the case of Eduard Meier and the FIGU community in Switzerland, Urandir Oliveira has founded Project Portal at his family-owned Ecovila Fezenda organic farm. Both communities engage in the sharing of extraterrestrial spiritual teachings concerning meditation, telekinesis, levitation, telepathy and the oncoming Earth changes.

Quadratic mapping of infrasound standing wave patterns reveals the Rochedo area of the Brazilian state of Mato Grosso do Sul as a resonant convergence point. The Oliveiras' 209-acre farm (19.88°S 54.95°W) is 6,669 miles from Giza, or 26.79% of the Earth's mean circumference. It is 1,365 miles from Nazca, Peru –a resonant circumference distance of 5.5%. The nearby giant rock formation, for which the town of Rochedo ("Little Rock") gets its name, is an ancient geological sign of convergent infrasound (opposite).

The holistic vision and profoundly transformative consciousness that is being engendered through these communities has apparently been perceived as a threat to the US government policy of UFO secrecy and denial, resulting in repeated attacks of all kinds. Both FIGU and Project Portal have been branded 'cults' and libelously slandered and legally attacked by so-called 'UFO groups' that are nothing more than US government propaganda fronts. Both Eduard Meier and Urandir Oliveira have received numerous death threats and both have evaded assassination attempts. Oliveira was also harassed, arrested and briefly imprisoned in March of 2000 on false and absurd charges aimed at destroying Project Portal. These detractions have been unsuccessful due to the open nature of Project Portal's workshops and events, bringing hundreds of people to the eco-village every year for spiritual and psychic development.[3]

Looming majestically over green fields and forests is Morro Ashtar, an oblong plateau that has become the center of increasing aerial activity. *Continually unfolding telepathic interactions, luminous plasma ships, and very bright but completely silent aerial explosions suggest to some residents that a subterranean base exists below the plateau.* Many Corguinho residents have spoken publicly about their numerous and regular sightings, and in some cases have also reported direct conscious interactions with extraterrestrial beings.

One such case involving ongoing contacts has been reported by Mr. João, a 72 year-old lawyer from Rio de Janeiro living in the Corguinho area.[4] An intriguing aspect of the experiences of Mr. João involve loud booming and humming noises emitted by the invisible plasma ship of the visitors to raise his electro-acoustic vibratory field, during which time they continually asked about his perception of colors presented from the luminous craft. Success was achieved after Mr. João was asked to put his hands on the ground, *as his shoes had apparently prevented the energetic activation they were undertaking!*

While videotaped group experiences involving interactions with plasma craft do exist, other such events in the Corguinho area have gone largely unreported. Documentary video was recorded by Fabricio da Silva during Urandir's experience aboard the fourth-density plasma ship on the night of September 15, 2002, including footage of a 'rain of stones' falling from the sky during the initial beam levitation event.[5] Also included in the video document is an interview with Urandir Oliveira by science reporter Linda Moulton Howe from February 9, 2003, at his farmhouse, conducted with Portugese-to-English translation by Luciano Didier. Subsequent interviews by Moulton Howe (with translation by Didier and Felipe Branco) explored the many intriguing details of Urandir's plasma beam transport and transdimensional experiences aboard a plasma ship for three whole days, from February 15-18, 2002:

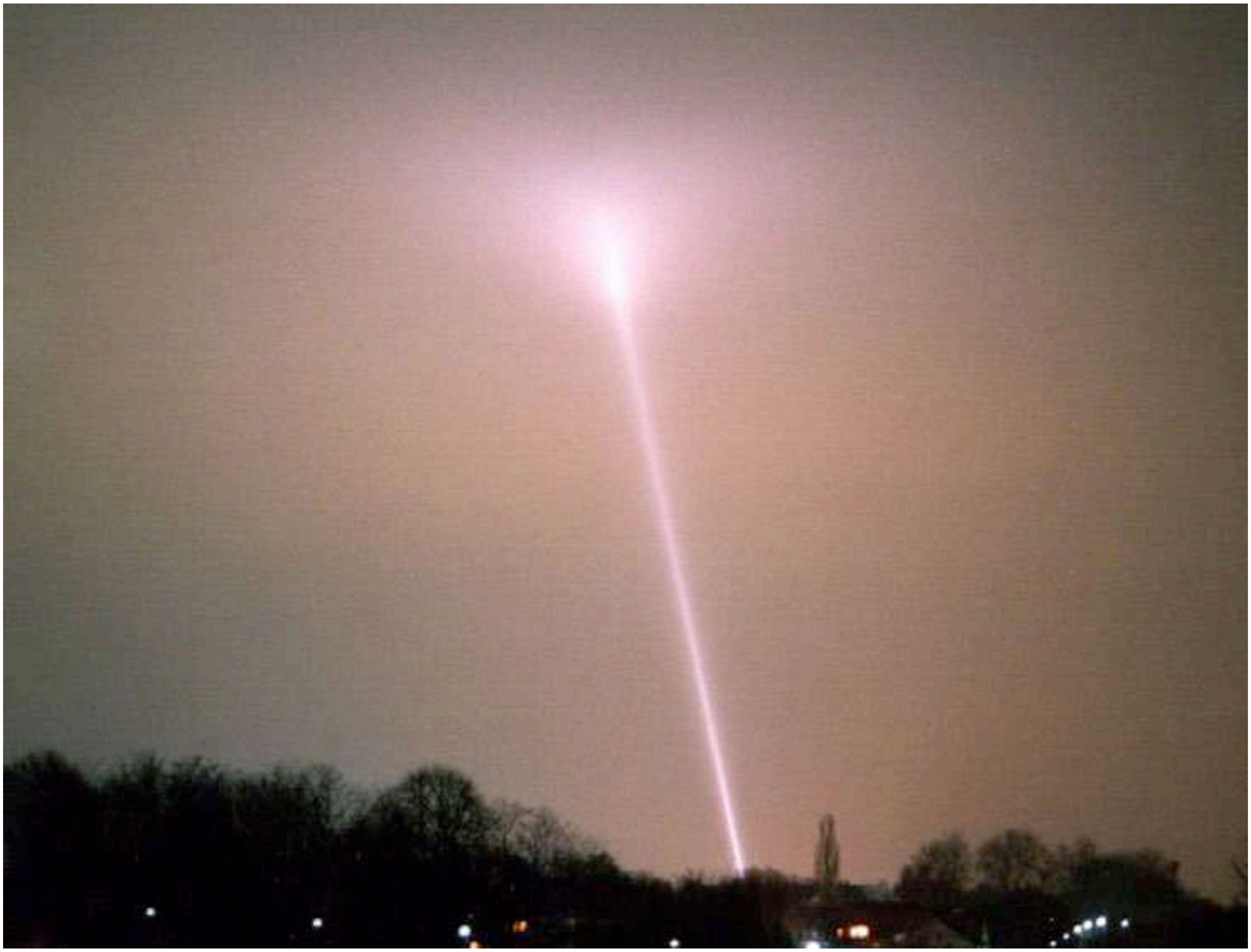

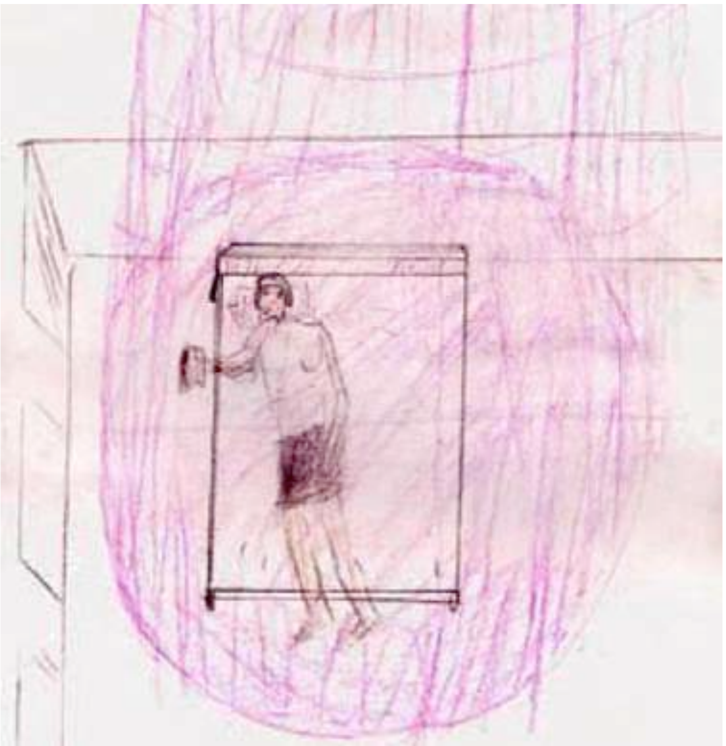

He (Urandir) was very anxious because for three days already he was expecting to be abducted, and there were only three hours left to the time the beings told him he would be abducted. (…10:00pm). He tried to prepare himself, so he went home and took a shower and put his shorts on and laid down on the bed. He picked up a book and tried to relax. A few minutes later, he realized a flash in the room. Then he felt his body like a fever warming up.

Q: What was the color of that flash?

The flash was sort of a light purple, or violet... [lit up] the whole room and then was condensing in the form of… a tube that included the bed and himself all the way from the bottom of the floor to the ceiling… The book was in his right hand (extended in the air beyond the right edge of the bed) and when the light flashed, eventually he released the book. The book fell from his hand. And he felt sort of a paralyzing [effect]. He could not move. He really felt this fever in his body and his body on the bed. He felt his body warmer on the bed. [see Urandir's sketch of his body in the beam, at left. Plasma beam photo above was taken elsewhere,[6] yet displays identical characteristics as described.]

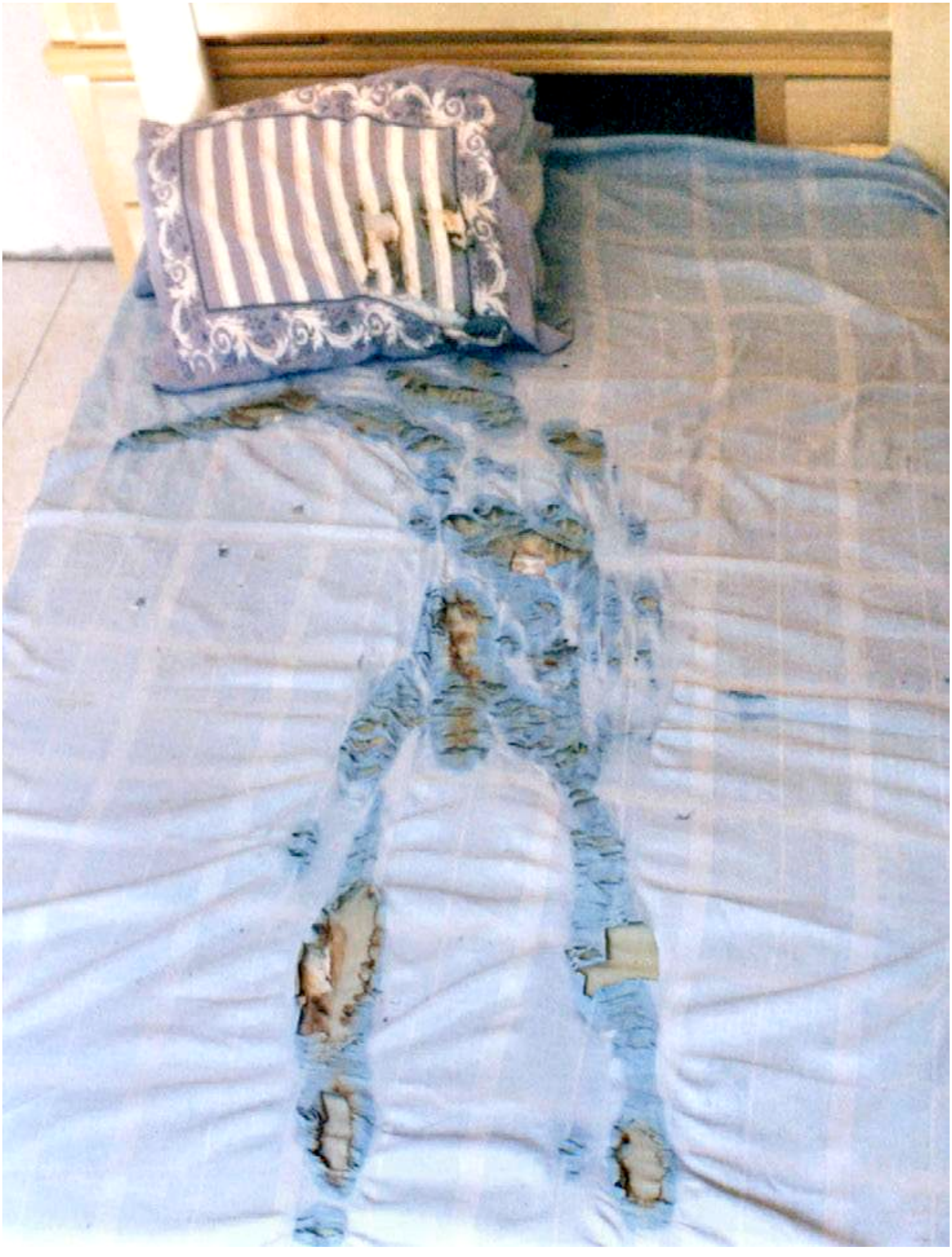

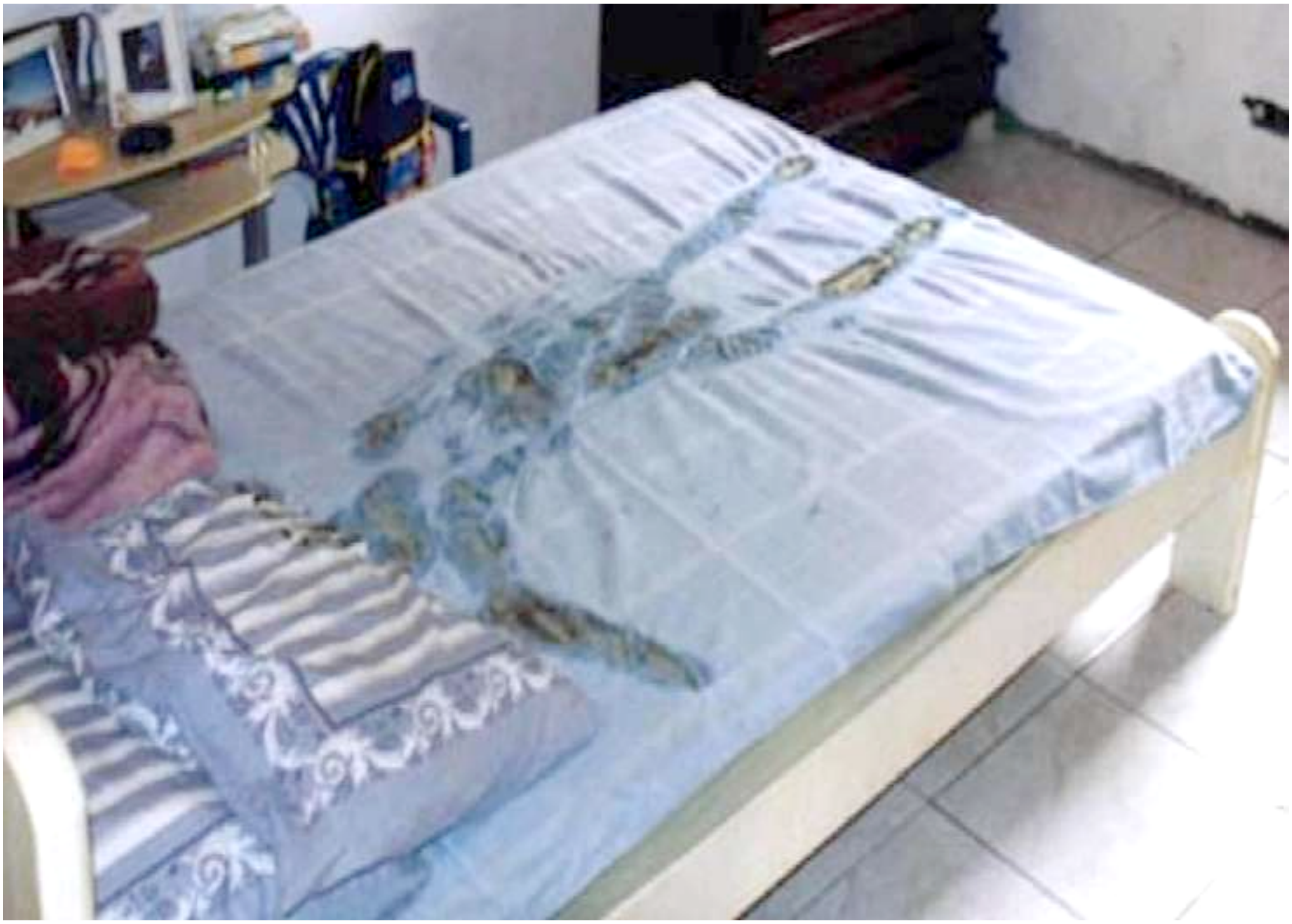

Q: When you look at those sheets and that ceiling, you would think you would feel pain (from the burn)?

Yes, because it leaves the impression that something burned someone there. But knowing in advance already the beam process of abduction (from earlier encounters), he knew it was not going to be painful at all.

Q: What is happening after the heat and going upward?

He felt himself rising up and going through the ceiling. From the point he was paralyzed, he could not move so much, but from what he could see, he realized everything was like normal from inside the light as he is rising. Actually the position did not allow him to see much, but what he could see are the normal things like the walls, the door, [and] the window.

Q: What happens as he gets up to the wood ceiling?

He saw himself going through the ceiling and he could actually see the molecules between the ceiling, like if he were going through foam. Not like a foam rubber, but like a soap, more like soap bubbles, like when you wash your hands and you produce the foam that comes out in the soap. And then he is turned upright so he is standing after he crosses that point of the ceiling, going through the ceiling, in the same violet light.

Q: The other day, he mentioned he could see and hear more?

Even at the height and distance he was, he could see and hear perfectly well, even better than normal, as if everything was amplified. He could see the stones falling everywhere, he could hear peoples' conversations and see them below… he could only see [the plasma craft] at the moment he approached it.

Q: The rocks are falling why?

It was a signal, a sign, that the beings left so that everyone would recognize that moment and everyone would know he was being abducted… As he was approaching the craft rising and saw everything that was going on down below, he goes into the first room into the craft. And it is also violet color inside this craft.[7]

The many fascinating aspects of the plasma beam transport of Urandir Oliveira on September 15, 2002 provide a wealth of information for the study of fourth-density technology. While such vivid descriptions of violet fog and levitation in luminous beams have been shared by many contactees over the last decades, none of those accounts were accompanied by accurate scientific descriptions of the entire experience. Urandir's case is a gift to Earthly science, in that he provides detailed scientific explanations for physical trace evidence in the form of textile samples with plasma-induced molecular alterations that can be studied by physicists for potential replication by all known means.

Previously obtained evidence of transdimensional plasma phenomena subjected to rigorous scientific analyses has been limited to downed plants from crop circle formations. The rapid heating of water contained within the effected plants by microwaves causes expulsion cavities in the nodes of the stalks as water is forced out by high-pressure. Video of plasma balls forming these marvelous geometric mandalas in the crops confirms their formation by the microwave emissions of airborne HHO plasma spheres that are themselves *natural elemental beings giving mathematical lessons to Earthly humanity.*

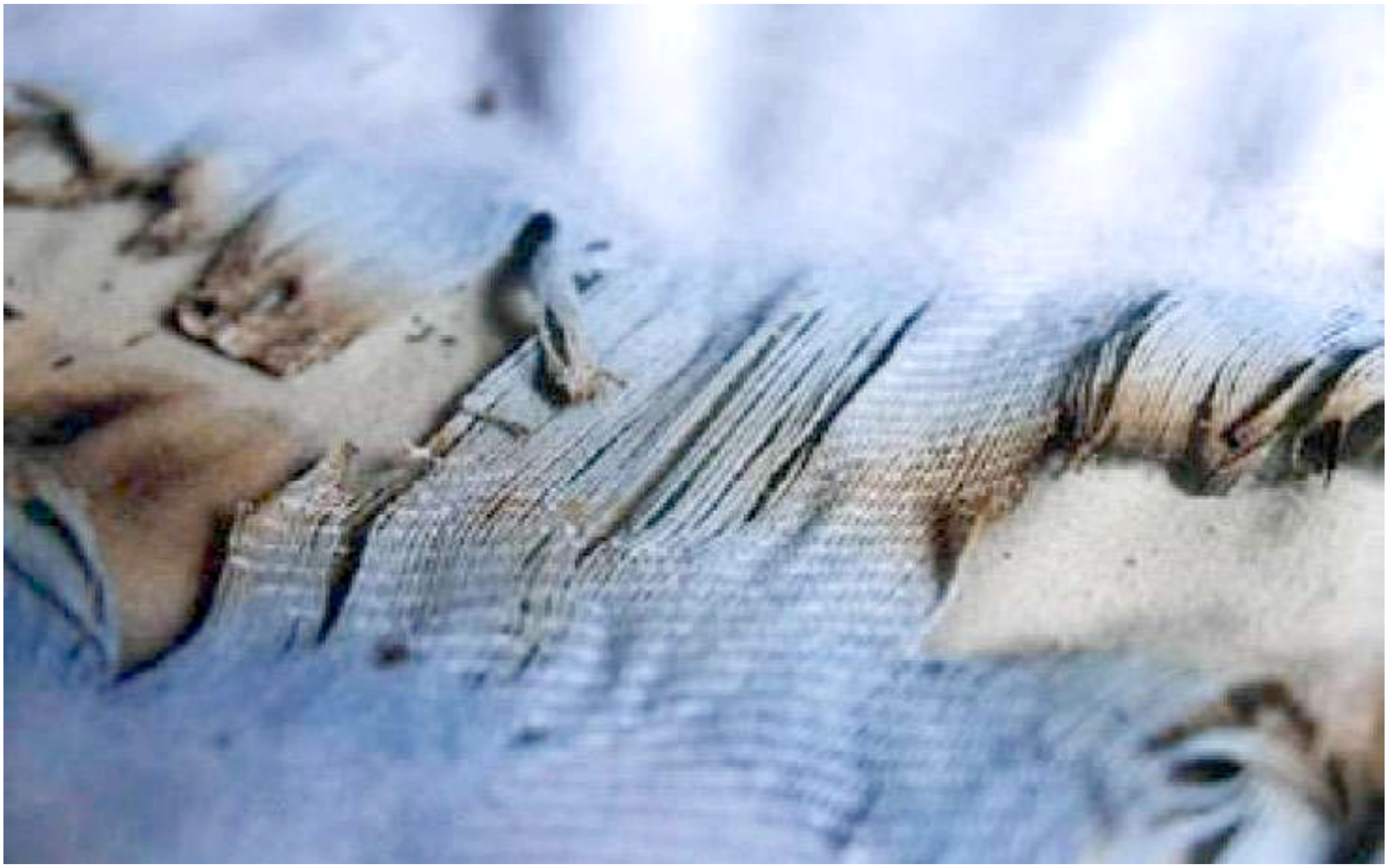

Scientists have confirmed through various means that the microwave radiation responsible for the many geometric crop formations of non-human origin must have been delivered in a very brief period of time, perhaps involving an exposure of under 7 seconds. This is exactly the conclusion that must be drawn from analyses of the scorched bed sheet evidence presented in the Oliveira contact case (above). Urandir experienced no pain during the plasma beam transport, as the beam technologies employed are completely safe and biocompatible, normally leaving no detectable trace of its use whatsoever.

Recently applied as a dental torch, helium/oxygen plasma is safely maintained near room temperatures by use of nanosecond electrical pulses, causing only a few degrees of heating in tooth enamel and gum tissues while vaporizing plaque and sterilizing oral films of bacteria in seconds.[8] During Urandir's beam transport, the air the room was infused from above with HHO gas bearing gold and silver nanoparticles and rapidly ionized by infrared light followed by nanosecond electrical pulses to form HHO plasma near room temperature, which became concentrated within the beam's vertical column. The paralysis experienced by Urandir is electronically controlled by automated computers in the plasma ship above, as pulses from rotating green lasers within the plasma column created differential forces within the superfluid beam to levitate his body off the bed. The scorching of the bed sheet and pillowcase was then purposefully executed by a very brief heating of the plasma field closely surrounding Urandir's body just above the bed, before continuing with the strangely silent levitation process.

All matter within the fourth-density HHO plasma beam undergoes an atomic synchronization and expansion and that allows complete interpenetration of objects without interference, whereby Urandir's body was brought directly through the ceiling toward the craft. He describes being rotated into a vertical position after having passed through the roof, allowing him to see all of the activity going on below. Urandir's description of enhanced vision and hearing inside the plasma beam column is a result of the fourth-density awareness endowed by immersion of the human body in a superconductive state where the holographic aspect of consciousness becomes reunified with *the eternity*. The bright white light that Urandir describes as emanating from the electroluminescent metal hulls of the plasma ships also radiates from the craft's interior walls, which electrify the interior chambers with the same biocompatibly tuned HHO plasma that is concentrated within the beam transport tube emanating both ultraviolet-A and infrared light.

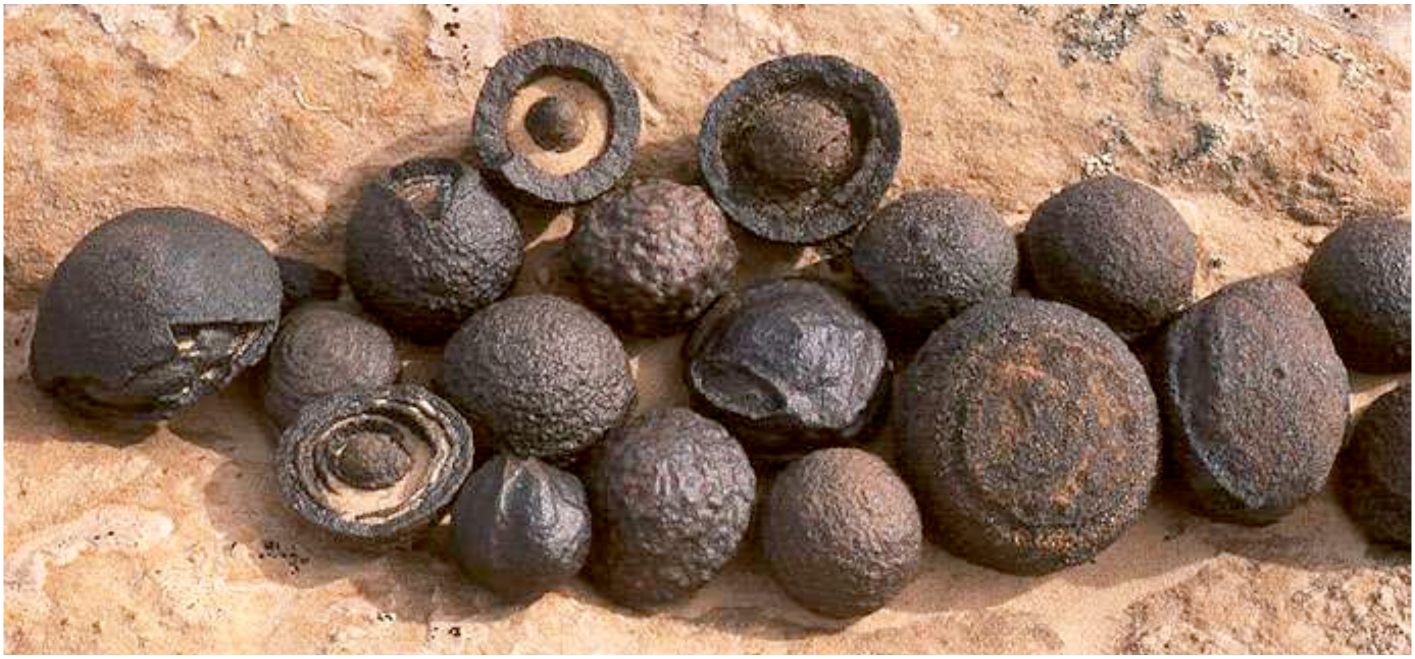

After Urandir's transdimensional levitation through the roof of his farmhouse and into the plasma ship hovering silently above, the cloaked ship created a 'rain of stones' as a deliberate sign to the neighborhood that the foretold event was indeed taking place. In the video footage recorded by Fabricio da Silva on the night of September 15, 2002, one can hear the round stones occasionally falling on the corrugated steel roofing and rolling down to be collected from the ground by the surprised neighbors.[9] Petrographic microscope analysis of one of the disc-shaped stones collected from the Oliveira farm was conducted by Professor Gomaa Omar, Geologist at the Dept. of Earth and Environmental Sciences at the University of Pennsylvania, and the results discussed in an interview with Linda Moulton Howe in June of 2003:

> It's very clear when you look at the cross-section here, you can see how [the stones] are banded. This is a mystery to me… I have never seen this particular 'football-like' shape and this particular mineral arrangement…
>
> Q: *And to produce this shape – of these sort of football-shaped or disc-shaped stones… and to have all of this quartz embedded in the iron… [would require a temperature] at about 2,000°F?*
>
> It was in a melt, yes… Definitely it is very clear that the quartz grains were floating in this iron liquid at high temperatures… But, as to the banding and the stratification going from the center outward, that has something to do with spinning force… and I really don't know what would cause this spinning force to form this [symmetrically disc-shaped iron/quartz stone].[10]

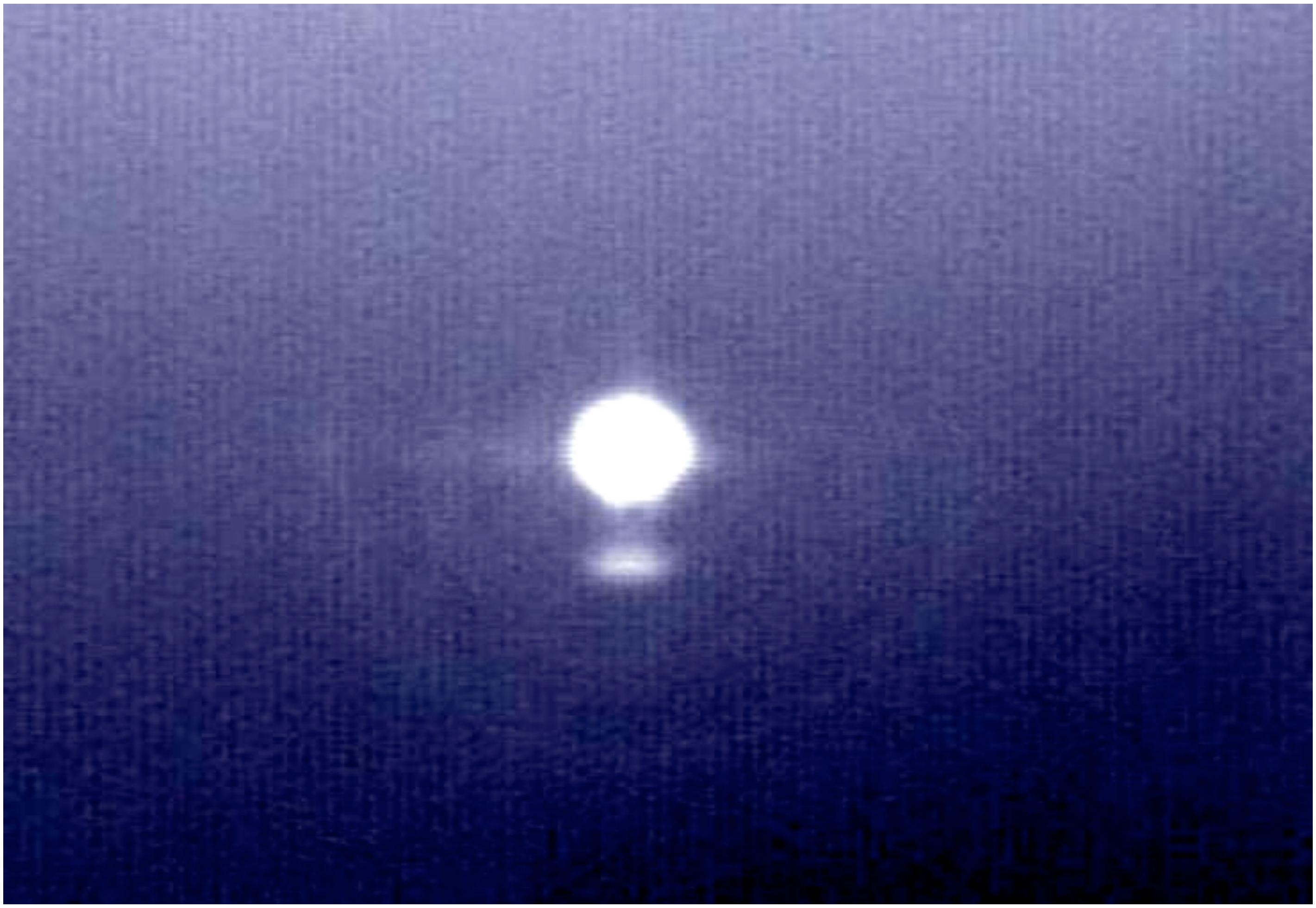

Iron in the red soils of the Corguinho area are attracted to the intense electromagnetic fields generated by the implosive hydrino plasma beam technology. Small grains of quartz sand are lifted from the ground along with the magnetic iron particles to be mixed with the liquified iron suspended in the intense rotating electromagnetic fields generated below the transdimensional spacecraft. Surface oxidation does not occur on these disc-shaped metallic concretions due to the unique effect of HHO plasma. Inventor Denny Klein noted the rust-resistant properties of surfaces brazed by HHO plasma during experimentation throughout the 1990s, which lead to the first commercial production of HHO gas torches, however short-lived.[11] The convergence of Earthly human technology with extraterrestrial means is greatly accelerating.

All over the world, these disc-shaped iron-quartz concretions are created below passing plasma craft of all kinds, even becoming commonplace in dry canyons which act like funnels to collect falling debris (opposite). Rains of these round iron stones have also been recently linked to invisible aircraft above the Indian towns of Ratria[12] Dhenkanal.[13] Infrared video obtained on March 5, 2004 by the Mexican Air Force[14] recorded a quickly moving formation of 11 disc-shaped aircraft that were completely invisible to the naked eye due to their use of HHO plasma screens surrounding each vessel like a bubble, with heat signatures of plasma vortices below each craft being clearly visible (above). The formation of the iron concretions was even filmed during a crash event in New Mexico,[15] actually showing magnetic material in the soil being trapped in the vortex below the craft, visible as molten metal glowing a bright red color.

Similar phenomena involving the interaction of iron-rich substrates with the strong rotating magnetic fields of plasmas moving above the terrestrial surface has also been identified on Mars.[16] Contributing to events of this kind are various plasma formations including natural lightening arcs that occur before the many impacts of meteor and comet debris to produce tiny spherical iron pellets that become thrown off as slag in a wide shower. The application of this phenomena by Urandir's fourth-density visitors as a message to the community is another example of their purposeful production of tangible evidence:

Q: [What occurred right after transport into the plasma ship?] This being was waiting, and how many others?

There were more – he cannot tell exactly how many, but there were more behind him. He was the main one, this one, waiting for him… [is] like 6-foot-three or -five inches [tall]… Then he walked in towards a bubble that was higher than him. As he saw a bed inside this bubble and just supposed he had to go and lay down there… [Inside the bubble] it was cold… like a car with air-conditioning on… He just walked through it… It was as if he (the blond being) was telling him, 'Go in, lay down and relax.' That's what he felt… He goes through it, feels very cold, lays down on the bed and then some sort of mist comes in and the shorts he had on disappear and in its place, some kind of other clothes materialized over his skin… [like those worn by the blond beings]. All of them, including his own, emanated a light like silvery-white neon lights… This light stabilizes his chakras and is sterilizing him from any bacteria on this planet.

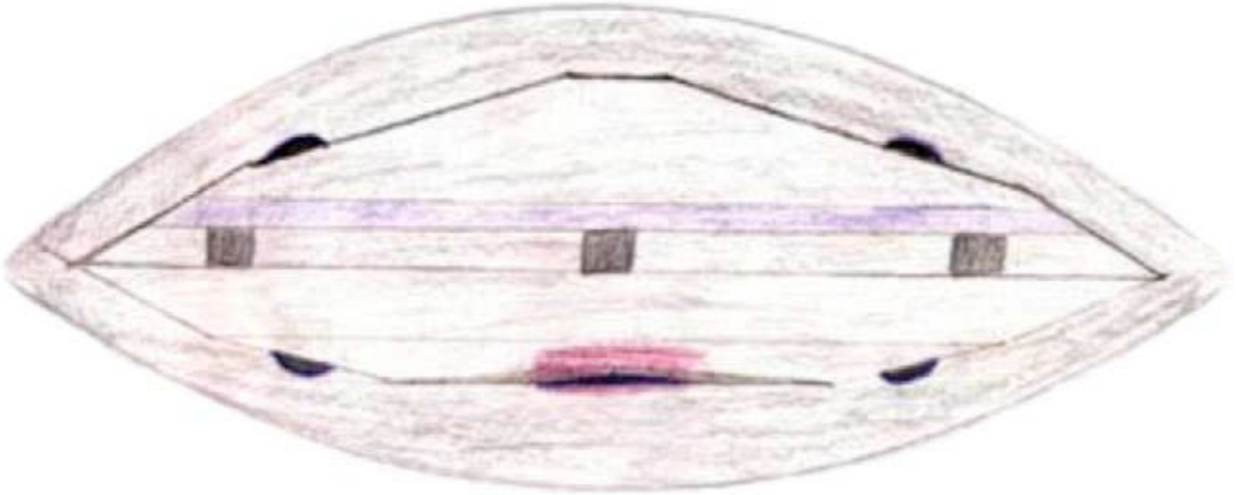

Q: What happened then after he's in the suit?

He walked out of this bubble and went towards them. The greeting with the beings is touching the (index) finger to identify the energy. And the same finger passing, like rubbing, the forehead, like if he was tuning into the other one's mind to start a telepathic conversation… The being touches his own forehead and then Urandir touches his own forehead. Only the (index) fingers touch… only telepathic conversation –no images in his mind.

The tall blond male directed Urandir toward stairs and they climbed to the next floor [as seen in his drawing, above] where Urandir was to remain for the next three Earth days watching wall screens of images about the Earth, this solar system and the galaxy. Watching with Urandir were representatives of other non-human species who Urandir understands are allied with the blond beings in a mission to help humans transform and evolve. The tall blond beings are trying to transform the mind frequencies of as many people as possible, raising them in an effort to neutralize large numbers of low frequencies and help block catastrophes they see in Earth's future. Urandir considers himself one of their students and his goal is to share publicly what he has learned about mind energy over matter from the beings.

The following excerpts about Urandir's memory of what he was shown and told are taken from a five hour conversation that was tape recorded in Corguinho, Brazil on February 10, 2003, with Portuguese to English translation help. Urandir said that various dates connected with earth changes - both positive and dangerous - were shown to him, but not in chronological order. (After walking up stairs with the blond being to a second floor) the wall of the craft there transforms into a monitor and they start showing scenes… The worldwide events that are going to happen… [in Earth's immediate future are presented and described telepathically.]

Scene 1 - 2014, Outer Space Chunks of Ice

The solar system traveling and meeting a curtain of ice pieces approaching… He said the solar system travels in a big galactic cycle through the ice… This is the second time the earth has passed through these huge clumps of ice. It's very close this and it will take maybe 15 years to go through the big pieces, but a few pieces of ice could already be coming toward Earth today. The blond beings said that we cannot see this yet because there is another body that reflects with a very strong electromagnetic field all the light of this ice curtain approaching the solar system…

Scene 2 - December 22, 2012, Galactic Pulse

A giant galaxy was the next image… [with the] solar system down there. On December 22, 2012, this galaxy is going to emanate a big flash of light that will affect all the suns. Every sun reflects to all the planets surrounding them and will cause a 'reversal' not only the earth, but all the planets… It's like God synchronizing the whole universe, like a synchronization ray or beam that would come from God, the Creator. He said it will be seen by any creature on the planet that has eyes to see it… For our planet, there will be lots of solar storms. The suns will get hotter with more solar storms…

Scene 3 - December 22, 2012 - Sun and Stars Moving in the Sky

They showed how the human beings would see the sky from the earth… Moving, like [on a teeter-totter]… if you are looking at the sky, it will swing. You will see the moon and the stars moving.

Q: Like [The Book of] Revelations in the Bible says the stars will move in the sky?

Yes, exactly… [moving or wobbling during] both day and night.

Scene 4 - 2014 and 2023, 'Plasma Ships' Evacuate Some Humans from Earth

In the next scene are lights in the sky which are the blond beings' plasma ships picking up humans from hilltops in their beam transport technology. This is roughly near 2023. But something similar in evacuating human beings starts in 2014. These are like stars representing the plasma ships. People are on the hills and the ships are rescuing them… [The plasma ships emanate] silver light. Like a dish to a dish…

Scene 5 - 2004, Cosmic Body Travels Close to Earth

At end of 2004, there is a shock due to a body that travels close to our planet and interferes with the climate and atmosphere… [*This event occurred on September 29, 2004* – the close passage of Asteroid Toutatis, as also predicted during the extraterrestrial contacts of Eduard Meier in 1978, who even foretold the exact name that would be chosen by French astronomers several years before the asteroid had been observed.]

> **Scene 6 - 2007, Cosmic Body Disturbs Earth's EM Fields; Scientists Discover Life-Extending Water**
>
> In 2007, there will be both bad news and good news for human beings. There will be a cosmic body [–Dark Star–] that moves toward or into the solar system that will affect the earth's electromagnetic fields… This electromagnetic wave will affect some places locally, different, but not the whole earth. A good example is the Etna volcano in Italy. It could erupt and kill thousands of people. The rest of the world will carry on…

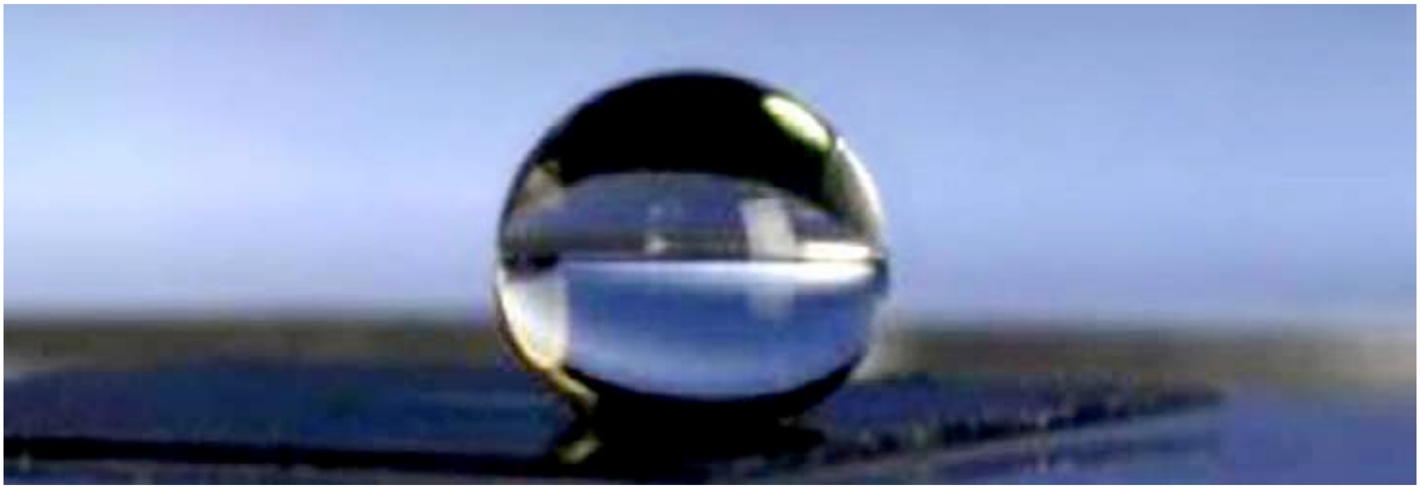

> The beings said that science has already got the knowledge, but nothing has been published yet. In the ice caps of the Antarctic, scientists will study the water and find it is more pure, free of heavy water, deuterium? He is saying they are going to find this water with less heaviness. Scientists will discover this new purer water will retard the aging process by about 80 times. People could live 300 or 400 years. The technology to extract this new water in 2007 will be available… [whereby pure protium can be separated from heavier water molecules.]
>
> *Q: Who is planning this cruel joke of us finding life-extending water just before humans are wiped out by ice blocks only seven years later?*
>
> But there will be many survivors who can use the new discovery. The Earth has more than 6 billion inhabitants. Around 3 billion will die with all the catastrophes, but nearly as many will survive.
>
> **Scene 7 - 2028, Blond Beings Return Some Humans to Earth**
>
> And by 2028, it's like as if they are helping the population trying to help fix the planet. They start bringing back those people they evacuated before. And it starts a new era on the Earth. It's like everyone is part of a cooperative effort… After the evacuation, the people are returned years later but younger than when they were first beamed up into the plasma craft… They are taken into a different dimension that was adapted for the human beings to be preserved...[17]

The specific predictive information provided to Urandir Oliveira by his extraterrestrial teachers deserves careful analysis and consideration. Their concise description of the Galactic Pulse of December 22, 2012 being a synchronization beam radiating from the galactic core *–observable to all as a bright luminous flash that induces magnetic reversal of the entire Sol system–* corroborates descriptions provided by other extraterrestrial sources. The wobbling of the planet for days during a major reorientation in its rotational axis echoes ancient Biblical predictions, and *suggests a permanent seasonal change on Earth*.

We are warned that light from a threatening interstellar ice sheet is reflected by the strong electromagnetic field of the brown dwarf binary twin of our sun, itself having already invisibly entered the solar system. The foretold electromagnetic wave effects of the closing Dark Star are indeed perfectly exemplified by unusual events occurring in the Messina region of Sicily, just west of the mentioned Etna volcano. *The town of Canneto di Caronia has experienced repeated evacuations throughout the last several years due to extremely high electromagnetic fields being transduced from intense acoustic energy by the piezoelectric limestone bedrock.* The entire area is being bombarded by ultra-low frequency standing waves focused by the Giza pyramids, at a distance of 1,120 miles, or 4.5% of the Earth's mean circumference. *The brief explanation of localized effects given to Oliveira refers to nonlinear focal points of infrasound resonance.*

Information given concerning *lightwater* that Urandir drank with his companions for three days aboard the plasma craft is of great importance, being ‘lighter’ in reference to *both its weight and optical aspects*. Being comprised of pure *protium*, with less atomic weight than the heavier forms of water (deuterium and tritium), lightwater can be extracted from mixed terrestrial sources by virtue of the specific freezing, boiling and vapor points of protium that differ from the lower forms. *Oliveira was given astounding details concerning cellular longevity: that under bioelectrification conditions, the ingestion of pure protium lightwater endows an extended lifespan of 300 to 400 years.* This is the biological effect of the fourth-density HHO plasma environment that *‘has been adapted for the human beings to be preserved’* such that they are *‘younger than they were when they were first beamed up into the plasma craft.’*

It appears that Urandir’s fourth-density teachers recognize the fact of governmental concealment of the anti-aging properties of pure protium water as a valid reason for providing this important information to humanity through Oliveira’s efforts. The invisible bionanotechnological aspects of the lightwater that Urandir drank and absorbed through his skin during his extraterrestrial experiences were not discussed, yet comprise a significant factor in the formation of the HHO plasma that preserves both spacecraft and passengers. Biocompatible gold and silver nanoparticles of less than 10nm in size provide for remarkable optical characteristics, by transducing red and blue light inside all the bodies cells for the activation of DNA as an organic superconductor. *This effect cannot occur without the ingestion of these trace metals.*

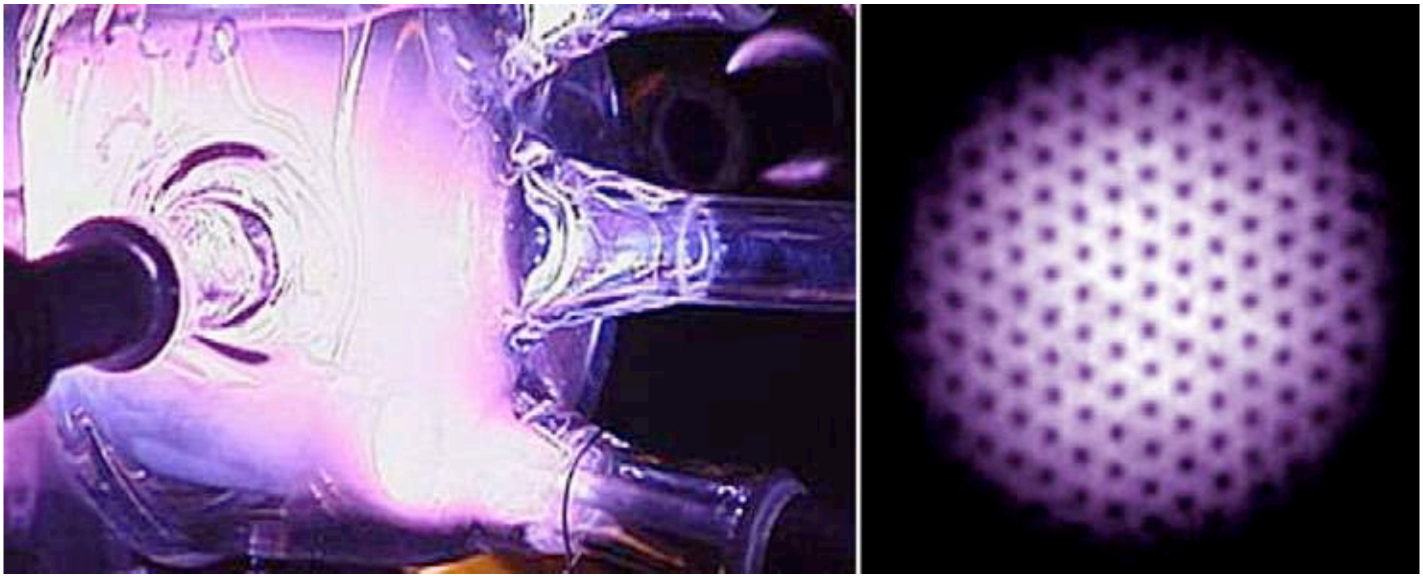

Physicists on Earth have now begun to apply the atomic implosion process of hydrogen for the transduction of *free energy from the resonant field of infrasound standing waves enveloping the planet,* using HHO plasma reactors that contain tungsten power and catalysts that are not biocompatible (above, left).[18] For direct application to the human body, gold and silver nanoparticles and dissolved potassium are essential. Earthly physicists have also recently defined the parameters for containing plasma within a superfluid beam, reporting the appearance of a hexagonal array of vortices during rotation (Ketterle, et al).[19] Oliveira described his experiences inside an HHO plasma transport beam to Linda Moulton Howe:

> The transport process was reversed, according to Urandir Oliveira, when the tall, blond beings decided to return him on the evening of September 18, 2002, after three days on a second floor room where a large wall screen showed Urandir and many other non-human beings images of Earth's alleged future. During our February 10, 2003, interview about the experience, I asked Urandir if he and the others were in the same room all three days. He answered, "Yes, it's like the walls were endless. I walked, but never reached the walls, so it gave the impression that the room was really big”…
>
> What he says he did understand was that the meeting in the "endless" room with many other non-human beings was supposed to have lasted for fifteen Earth days, not only three. But many of the people who had been at the Corguinho farm on the night of September 15, 2002, when the scorched body prints were found on Urandir's bed and the bedroom ceiling told friends, relatives and the Brazilian media about what had happened.

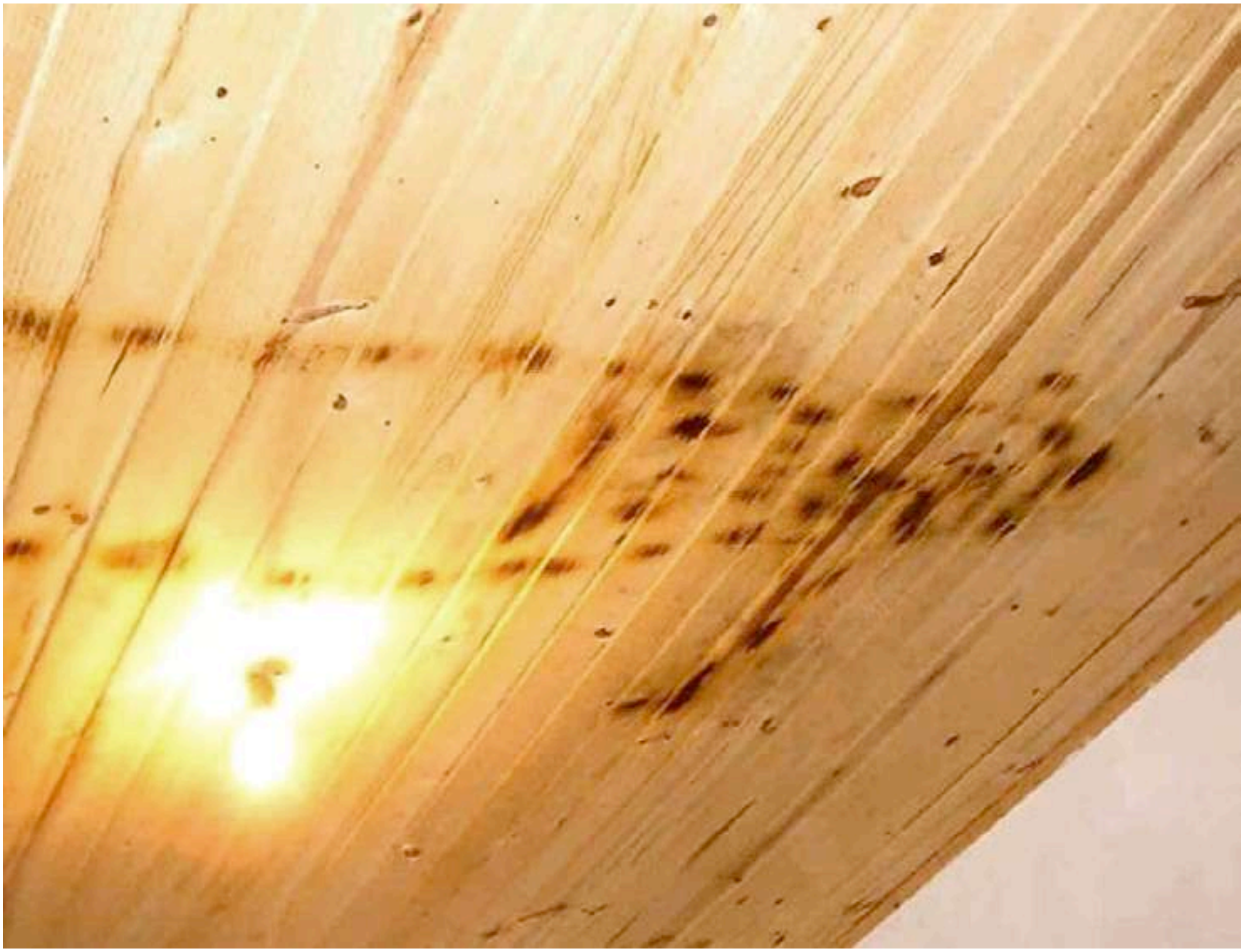

By September 16, 2002, Brazilian television reporters had traveled to Urandir's farmhouse to wait for a craft to return him. Felipe Branco and others who had been shocked the previous night by the bedroom scene had already locked the bedroom and shuttered the windows so no one could enter, including the curious media. When no dramatic aerial or beam activity was seen on September 16 or 17, the media left the farm. Urandir said, "In the end, the blond being told me that the trip was going to be interrupted, cut short, because of all the movement down there with the media which was not in accordance with what had been committed to me" by the beings. He understood that the non-humans did not want sensation, but left the physical evidence in the bedroom that unusual energy had interacted there for researchers to analyze.

On the evening of September 18, 2002, when the farm was relatively quiet, the tall blond being walked Urandir down the stairs in the craft to the place he had entered in the violet light. There he went back into the bubble where the silver, glowing, leotard-like suit dematerialized from his body and his own underpants materialized on him again.

Interview: The violet light was already there and a lot denser… When the violet transforms into a tube that comes down, the light becomes more subtle… At this point, he walks into the light. As he is going back down in the beam, about half way down, this light changes him from a standing position and puts him in the previous exact position when he was on the bed before (on September 15, 2002). The very same position and he is paralyzed again to go all the way through the roof and ceiling of his house and back to where he was…

Q: Any different feeling then?

Like as if his bones are stretching. Like cracking joints… The whole skeleton of all his bones go snap, snap.

Q: I wonder if it has to do with molecular expansion and contraction of passing his matter body through the roof and onto the bed?

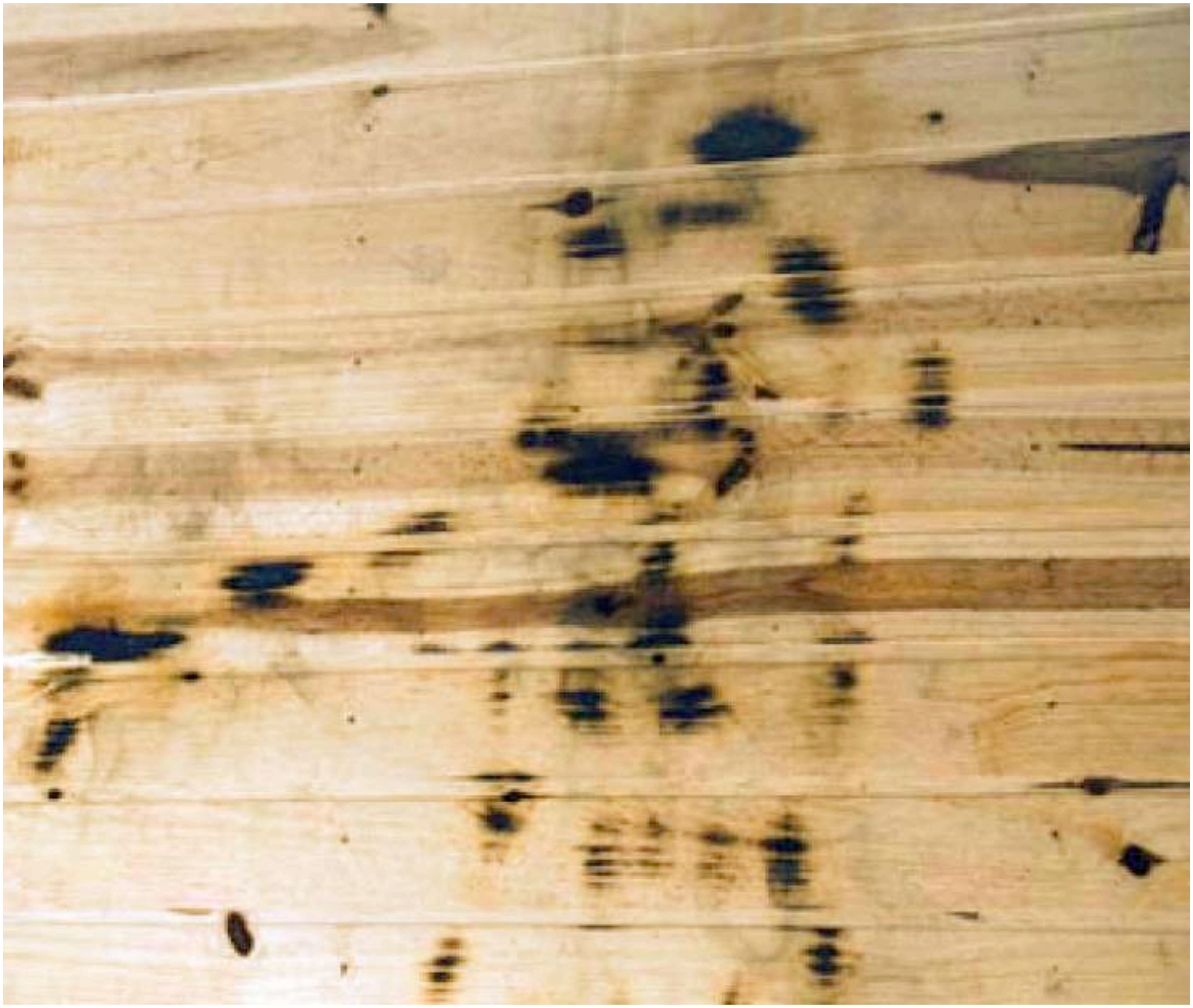

Yes, it does have something to do with expanded molecular structure so he can interact with matter again. It's as if they are re-tuning the body… He can see through the same [violet] light to the room… He saw himself going through the ceiling and then he feels his body touch the bed. But before he actually approached the bed, he could not see that the bed was burned because the way in which he was paralyzed would not allow him to. So the (violet) light goes off and everything comes back to normal and he stands up, switched on the room light, and then he is surprised with his bed because it is so burnt. He thought he had been melted there! ...There was some sort of smoke in the ceiling still coming out.

I was puzzled… as I photographed and gathered cloth and wood samples from the sheet and the ceiling… [because] the scorch marks on the pillow and the heavy scorch mark surrounded by the aura of his head in the wood ceiling [above] were in different places… A logical speculation about this difference is that Urandir's body was moving forward toward the head of the bed as it was lifted upward by the violet beam technology… I asked him why he thought there were no scorch marks of his hand or feet in the ceiling. He was uncertain, but said he understood from the non-humans that the energy beam that lifted his body was precisely tuned to the molecules of his body against the bed in order to separate them from the bed… [It is an] extraordinary technology.

Q: What went through his mind as he looked at the sheet and the ceiling?

At the time, he started touching himself thinking, 'Have I melted?' And he wanted to know his weight, so he climbed on a scale and he lost 4 kilos (8.8 pounds) in those three days.

Q: Did they feed him anything?

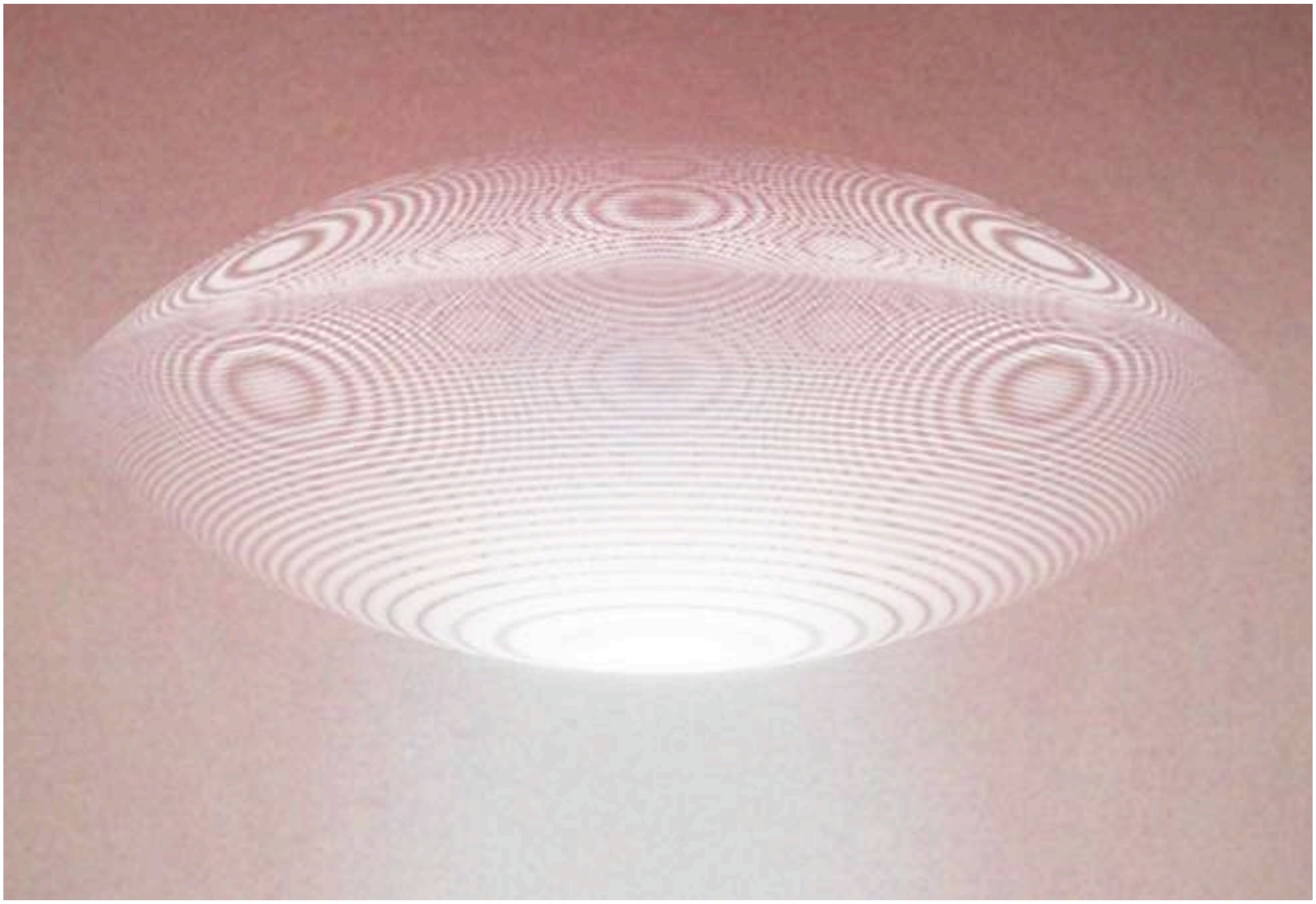

Some tablets with different flavors… If he wanted a meat flavor, they would give a meat flavor. If he wanted a sweet or apple flavor, he would have an apple flavor.

Q: Did they give him liquid? Water that was much lighter than our own…

Q: He felt perfectly fine? Relaxed, peaceful, joyful.

Q: So he is now standing in the bedroom. The ceiling is smoking. He is looking at his own body to see if he is burned… [and his clothing appears unaffected by the high temperatures. Then what happens?]

Then he tries to open the door of the room and it's locked. He opened the window and it is blocked from the outside… with bars. Everyone locked the room after he was gone (on September 15) after the [transport] to make sure no one would walk into the room… [thereby securing the evidence of Urandir's beam transport.] He has an intercom phone, so he called his parents who live on the farm and they rushed to the room to open it.

Q: He was gone for at least three days and humans go to the bathroom on a daily basis. What did they provide?

There was some sort of a cabin, almost transparent, a plasma material. And the only need or urge he had was to urinate. He did not feel any other urge. When he arrived at the plasma material, there was no hole, no basin, nothing. So, he thought, 'Well, I will just urinate here'…

Q: What happened?

So he starts urinating and when it is about 10 centimeters for the urine to touch the ground, it would disappear.

Q: He said he is standing before a plasma. Is it like a fog?

[The plasma cabin and urinal are formed by the] same material as the bubble. [The urine just disappears] before it touches the floor.

Q: Do the [extraterrestrial] blond beings urinate?

Yes, not as much as we do, but they do… [and they eat the] same tablets that they provided him." [20]

[A]nalytical chemist Phyllis A. Budinger [explained]... that the blue threads in the affected Oliveira sheet are cotton and the white are polyester. The blue and white threads are woven together throughout the sheet, but there is great variability in the body imprint from what looks like burned and charred cloth right next to cloth with only slightly darker blue discoloration that also outlines the entire body print. This is especially interesting when you consider the facts that cotton decomposes, or scorches, at about 148º Centigrade, or about 300º Fahrenheit. But polyester threads don't melt until 500ºF. There are melted, gummy threads intricately woven throughout the blue cotton threads -even in places where the cotton is not scorched, or even brown.

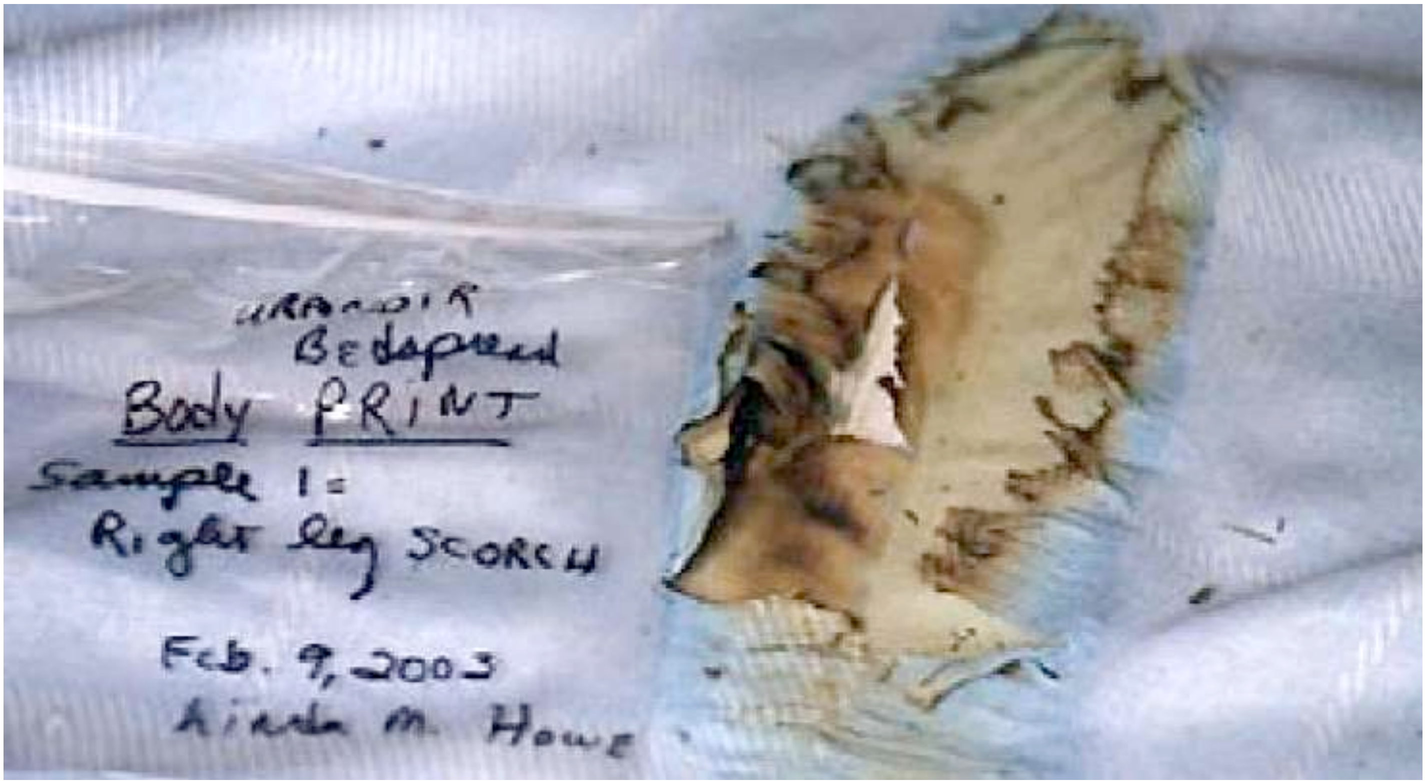

Interview: W. C. Levengood, Biophysicist, Pinelandia Biophysical Laboratory…: ‘I've never seen anything like this. This takes a very, very high precision kind of - whatever the energy, and there had to be energy here to transform, probably at the molecular level, this because it changed the internal energy of the material. Whatever it was… [this form of energy] was extremely precise.

‘I ran on to one region right at the edge of the brown material and the threads - you could actually see where they were white, normal threads, and then just a fraction of a millimeter away, they had begun to be transformed. They were glassy-looking, but you could still see the individual fibers inside. Then in another half millimeter, they were totally homogeneous into this opalescent material. So, that transition is very precise and you could see the line of demarcation if you had the incident light just at the right angle on the sample, you can see this band of change. So, the energy here is very, very precise.

‘I measured the weight per unit area of a control sample. The weight of the control cloth is 17 milligrams per square centimeter. The blue region is 24% higher in weight than the control cloth. I don't think anything was added. It means the energy has changed the density of the polyethylene and to change density means you have to change the molecular structure. There is some incredible technology [being applied] here…

Q: This is a photomicrograph of a sample of the body pattern on the Urandir bed that was not scorched, but almost looks wet… [opposite, center.] But in fact that wet look has been caused by a change in the polyester?

It sure has. These I call a transition, or the polyester is converted into what I call vitreous fibers. It looks wet, but this is actually sort of an opalescent liquid-looking, but it's like a plastic, very hard.

Q: We know that polyester… has a melting temperature of 500°F and that cotton has a scorch temperature of 300°F. The difference of 200°F in this photograph alone is baffling because it looks as if the polyester threads have been melted and the blue cotton threads seem completely untouched.

That's right. And that difference of 200°F -that's the astounding thing. The cotton threads are totally undisturbed.

Levengood's biophysical analyses of microwave-radiated plants from crop formations form the only precedent for the study of material evidence of aerial plasma beam phenomena. Further analyses of the Corguinho bed sheet were undertaken by Professor Herbet J. Barndt, Director of the Grundy Testing Laboratory of the School of Textiles and Materials Technology at Philadelphia University. In an interview by Linda Moulton Howe in May of 2003, Professor Barndt stated that the delicate burn pattern on Urandir Oliveira's bed sheet must have been made with laser-like precision in mere seconds of exposure:

> *Q: Could you explain a little bit about the possible importance that water might have played – both from having taken a shower and perspiration when this individual laid down?*
>
> Well, when cotton gets wet, moisture actually goes into the fiber. In other words it's inside the fiber, as well as on the surface. [In the case of] polyester, the water remains on the surface. So in the presence of a high temperature the water will evaporate from the surface of the polyester before it evaporates from the cotton, because the water is bound to the cotton fiber. So, in order to damage the cotton fiber you have to evaporate the water to get the fiber hot enough to burn. But since the polyester isn't really holding water it can absorb heat more rapidly, and that may explain why the polyester melted, and the cotton in a number of these areas wasn't particularly damaged…

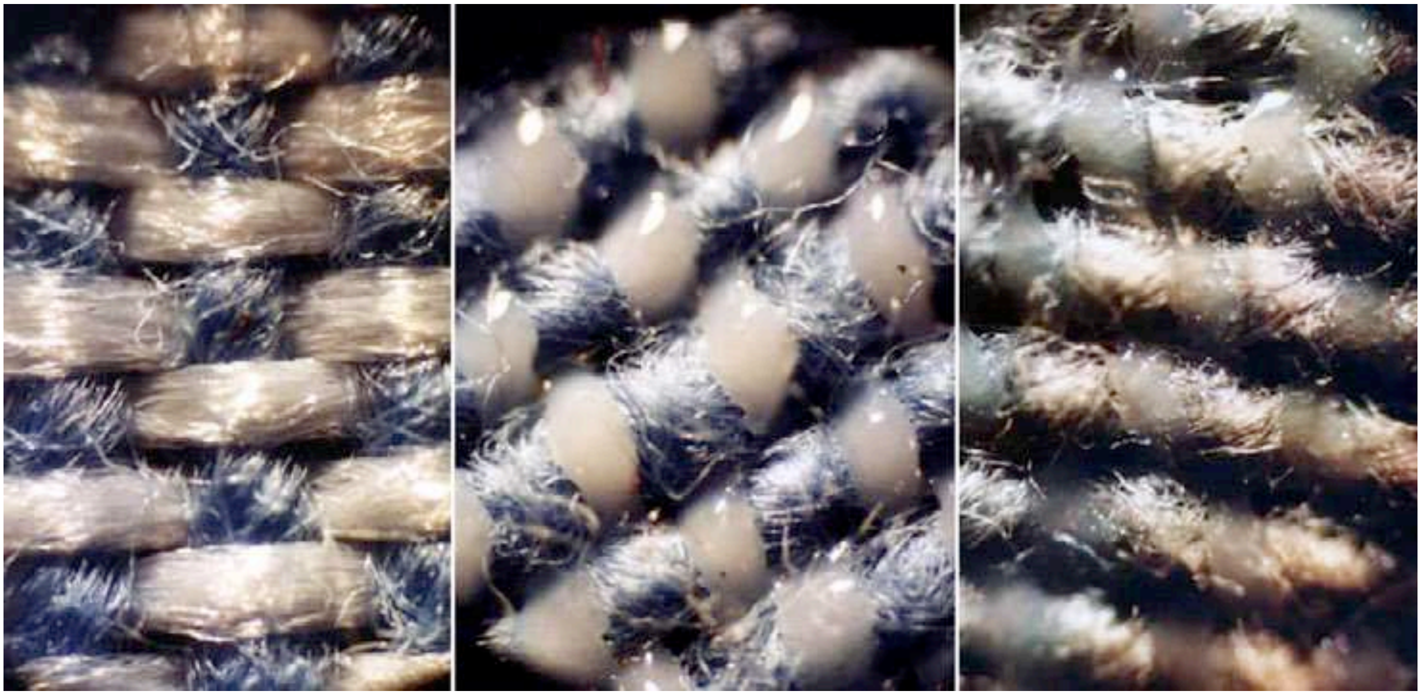

> Q: *Does that suggest… whatever did happen to this bed sheet happened in a very quick period of time?*
>
> I would think so… especially with the temperatures that are involved and the fact that the cotton wasn't degraded more than it was. it would seem that it had to have been a very quick occurrence, whatever happened here.
>
> Q: *And we're talking in seconds, right?*
>
> I would think so… It's very unusual… You would have to have a pinpoint heat source like a laser to do that… If you have a heat source that's intense enough to melt the polyester, and still not cause intense damage to the cotton, or at least some of the cotton, it has to move fairly quickly along this pattern… [or] the whole thing [is briefly heated] all at once…[21]

These detailed conclusions of the analyses of the bed sheet evidence from the Oliveira case point directly to a specific plasma heat source that was unknown to the 'experts', yet is derived from the most common and basic element in the Universe – hydrogen! *It is now known that HHO plasma is unique in that it heats different materials to different temperatures depending on the molecular structure.* An HHO plasma torch will not boil water, but instead reverts to water vapor on contact. The implosion of atomic hydrogen allows for precisely tuned interaction with the cotton and polyester threads in discreet bands of temperature change, producing the ultraviolet and infrared emissions described by Urandir as heat and violet light.

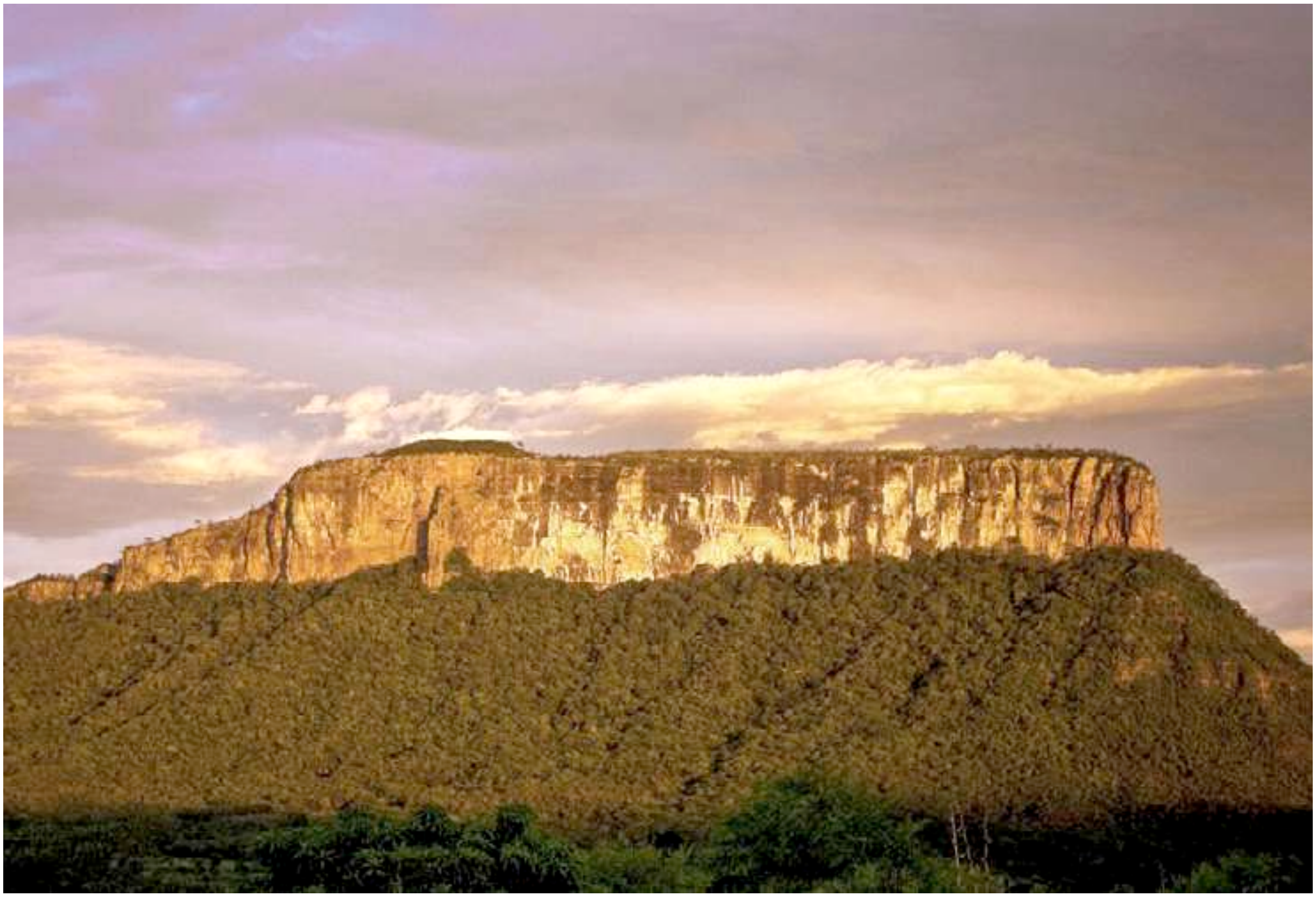

HHO plasma is also responsible for the molecular changes documented by Levengood regarding the weight increase of the vitreous polyester fibers. Similar analytic results were obtained from other material evidence collected by Oliveira and his group after a brilliant nighttime aerial explosion over Mount Ashtar (above) that scattered radioactive debris later measured with Geiger counters. A small brown ball of a solid resin-coated material was found lodged in a vertical rock face and carefully collected for laboratory analysis (opposite). X-ray diffraction analyses of the roughly spherical, resin-coated metal sample were conducted in the capitol city of Brasilia to determine the elemental composition of the artifact, which may reveal its extraterrestrial origin. *Testing showed abnormally high quantities of rare Earth elements:*

> Analysis of the nucleus, done by Centro Técnico Aerospacial [detected]… iron, oxygen, silicon, titanium and vanadium. In relation to the occurrence of these elements, the iron and the silicon can be largely found as iron ore and as silica... The oxygen is found associated chemically to the iron ore and the silica as iron oxide and silicon oxide, respectively. These oxides are mentioned in the probable presented compounds according to analysis done by x-ray diffraction. The titanium, as well as the iron and the silicon are not found in a pure state in the nature, being found also as oxides. In the sample, the probable compound that contains titanium is the iron oxide and titanium, which can be found in the nature as a mineral called ilmenite.
>
> Nonetheless, the most important aspect is the presence of vanadium. This element and all its derivatives are toxic to humans. Pure vanadium is normally used in special steels. It is found in complex minerals, such as vanadite ($Pb_5Cl(VO_4)3$) and carnotite ($K_2(UO_2)_2(VO_4)_2 \cdot 3H_2O$), with an occurrence in the crust of the Earth of 0.03% - a very restricted presence in the nature. Vanadium does not occur together with iron, silicon or titanium, which suggests that it was manipulated to be used with these chemical elements.
>
> Curiously, some forms of iron ore and titanium can present magnetism, which leads to certain suspicions about a possible mechanism of magnetic control of the space probe. The usage of iron bindings and silicon are also indicated in this sense, together with the bindings with Manganese… [Brasilia's] Laboratório Instrumental da Divisão de Química do Instituto de Aeronáutica e Espaço do Centro Aeroespacial showed that [the resin] presents a fossilized-material aspect of the ester resin kind, known as amber and silica.[22]

The initial radioactivity of the recovered probe debris is evidence of molecular changes having occurred during the explosion, while the exotic composition of nucleus confirms its unearthly origin. *The amber/silica resin coating presents an organic appearance, yet the roughly spherical metallic core of iron, oxygen, silicon, titanium and vanadium is certainly of non-mineral origin.* Brazilian Aerospace laboratories clearly acknowledged the highly unusual magnetic properties of the combination of elements found in the nucleus, while the resonant piezoelectric properties of silicon also implicate a highly advanced magneto-acoustic levitation technology. Silence accompanied the great heat and light emitted during the brilliant flash that briefly lit up the entire area like daylight, as the fusion of the aerial object into a tiny molten mass occurred.

The pointed speculation of analytical scientists supports the direct observations of witnesses to the luminous event and is further corroborated by information related in the contact case of Eduard Meier. Exactly this type of brown resin-coated mass was discussed as the byproduct of special security features commonly used by space-travelers visiting Earth. The dangerous potential misuse of technological knowledge is prevented by the installation of self-destruct mechanisms within all instruments and probes, for intentional or automatic activation in the event of malfunction. Semjase informed Meier of the necessity of the elimination of technological materials foreign to Earth, and a complete demonstration of the elimination process during his 62nd contact with cosmic visitors from the Plejares system, on August 12, 1976:

> *Meier:* I have heard that some objects had simply disappeared or burnt, or dissolved themselves in some way, if they were found by Earth human beings and if the objects belonged to some extraterrestrial intelligences.
>
> *Semjase:* The reason therefore is the elimination of these objects. We and many other intelligences are able to eliminate lost, or fallen into insincere hands, things by different ways, or simply to burn these to ashes. This happens by very far-reaching transmitters, which are started by means built into each instrument, a destroyer mechanism. These destroyer mechanisms are usually built into the objects and instruments, which we take to strange and still lowly-developed worlds, for reasons of security, so there does not come the dangerous for them possibilities of progress into the hands of the concerned still lowly-developed forms of life of any planet. These instruments or objects equipped with such destroyer mechanisms are normally produced from synthetic materials, like for example this small thing here… a selective warning instrument… which announces by a fine signal tone to me, when for example now should anyone come near us, whose brainwave pattern is not noted in this instrument, by which I mean, not registered.
>
> *Meier:* …Can you perhaps demonstrate for me how you destroy such a thing?
>
> *Semjase:* In purpose of your accomplishments, yes. Look, I simply throw it here onto the ground, and now do come a bit aside. – Like that, yes. Now look at this girdle-instrument; when you push these both flat buttons, then the warning means will start itself and burn and change towards a viscous and brown-black synthetic mass. Pay attention to the means laying there and push the two buttons. (I allow myself to push them.)
>
> *Meier:* Can I have the remainders? Know I want to let them get analyzed.
>
> *Semjase:* Yes, but then you should stop burning this mass, as otherwise nothing will remain for you. Don't touch it until 1$^1/_2$ hours have passed, as until then, certain radiations dissolved by the combustion will have escaped out of the mass.
>
> *Meier:* Well, but don't you think that our scientists could analyze important facts from out of this mass?
>
> *Semjase:* The chance is very small, because this sort of synthetic material allows no more recognition of the original compounds, [and] will in similar form be used as well on the Earth…[23]

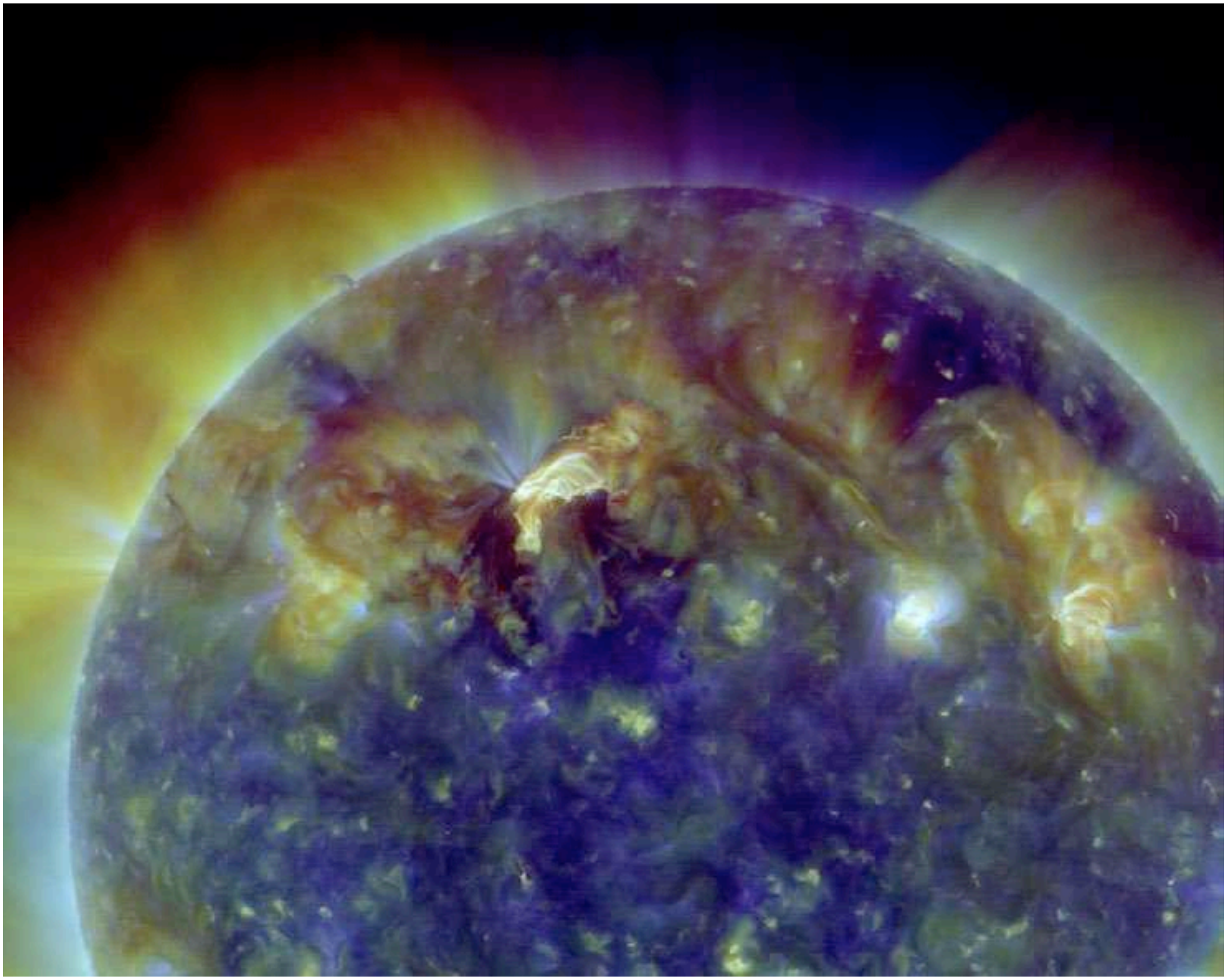

The explanations of Semjase perfectly describe the aerial probe debris recovered by scientific investigators involved with the Urandir Oliveira contact case, despite the more than three decades that separate the statements of Semjase with the spectacular aerial events occurring above Mount Ashtar in Mato Grosso do Sul, Brazil. The comprehensive guidance offered by visitors from other worlds, through the persistent efforts of Urandir Oliveira and other dedicated individuals, prepares Earthly humanity for the current electro-magnetic changes connected with the emergence of the brown dwarf binary companion at the outer reaches of our solar system, as the rain of Oort comet impacts bombards every continent of the Earth, continually threatening humanity for a prolonged period of many years into the future.

Greatly increased solar flare activity and resulting geomagnetic superstorms will accompany the process of magnetic reversal of the entire solar system in alignment with the Galactic Pulse of December 22, 2012. The ancient calculations of the calendars of the Maya culture resonate with the future events foretold to Urandir Oliveira, as do their advanced lessons on the effects of lightwater. The Maya also identify the transformative effects of the celestial lightwaters of Tláloc that rise in the mists at dawn, in contrast to the degenerative effects of the heavy waters that run downward to the center of the Earth. *Indeed, the process of activating the electro-acoustic vibratory fields of Earthly human beings is not only occurring in the plasma ships of extraterrestrial entities, but also in the Mayan pyramids and sacred sites throughout Central America.* Through these cosmic methods of bioenergetic activation the disasters threatening Earth's future may be mitigated or prevented, as the transdimensional enhancement of consciousness allows the integration of new perceptive capabilities in synchrony with the dramatic environmental upheaval.

References

---

[1] Moulton Howe L (2003) ‘Corguinho, Brazil: Farmer Urandir Oliveira Describes His Transport to A Non-Human Craft on September 15, 2002’ *Earthfiles*, online · http://www.earthfiles.com (removed)

[2] -- (2010) ‘Urandir Oliveira: Paranormal’ Urandir, online · http://www.urandir.com.br/

[3] -- (2007) ‘Urandir Fernandes de Oliveira’ *Projeto Portal*, online · http://www.projetoportal.org.br/urandir-f-de-oliveira-ing.html

[4] -- (2008) ‘Corguinho_alien_aontactee_account.avi’ *Google Video*, online · http://video.google.com/videoplay?docid=6825480847717282477&hl=en

[5] da Silva F (2007) ‘Abduction.avi’ *Google Video*, online · http://video.google.com/videoplay?docid=-3157785193143547318

[6] Rofe C (2003) ‘UFO Beam Photograph Submission’ *The Keyboard*, January 25, online · http://www.thekeyboard.org.uk/UFO%27s%20Fact%20or%20Fiction.htm

[7] Moulton Howe, op cit

[8] Dunham B (2009) ‘Cool Plasma Pack Heat Against Biofilms’, *University of Southern California News*, June 5, online · http://www.humanresonance.org/helium.html

[9] da Silva, op cit

[10] Moulton Howe, op cit

[11] Rodgers W (2005) ‘Clearwater Man Puts Technology to Work’, *Tampa Tribune*, November 27, online · http://www.humanresonance.org/hho.html

[12] Ghosh R (2004) ‘Villagers in Madhya Pradesh Face Mysterious Enemy’ *Human Resonance*, March 21, online · http://www.humanresonance.org/ratria.html

[13] -- (2009) ‘Mysterious Stones Fall From the Sky’, *Statesman News Service*, June 19, online · http://www.humanresonance.org/dhenkanal.html

[14] -- (2010) ‘Mexico Air Force FLIR UFOs’ *Youtube Video*, online · http://www.youtube.com/watch?v=uDOOZ_IPb6Y

[15] -- (2010) ‘New Mexico UFO Crash’ *Youtube Video*, online · http://www.youtube.com/watch?v=wAcotjjTj4s

[16] -- (2004) ‘Blueberries on Mars and Other Spherical Rocks’, *Thunderbolts*, online · http://www.thunderbolts.info/tpod/2004/arch/040827mars.htm

[17] Moulton Howe, op cit

[18] Mills RL, Ray P, Dhandapani B, Nansteel M, Chen X, He J (2002) ‘New Power Source from Fractional Quantum Energy Levels of Atomic Hydrogen Surpasses Internal Combustion’ *Journal of Molecular Structure* 643, pp. 43-54

[19] Ketterle W, Zwierlein MW, Abo-Shaeer JR, Schirotzek A, Schunck CH (2005) ‘Vortices and Superfluidity in a Strongly Interacting Fermi Gas’ *Nature* 435, pp. 1047-1051

[20] Moulton Howe, op cit

[21] Moulton Howe L (2003) ‘Part 7 - Scientists Examine Samples from the Corguinho, Brazil Bed Sheet and Pillowcase’ *Earthfiles*, online · http://www.earthfiles.com (removed)

[22] -- (2008) ‘Analysis of the Nucleus, Done by CTA – Centro Técnico Aerospacial’, *Projeto Portal*, online · http://www.ufocorguinho.com.br/resgate-ing.htm

[23] Stevens WC Ed. (1992) *‘Message from the Pleiades: The Contact Notes of Eduard Billy Meier, Volume 3’* Genesis III Publishing, pp. 164-166

5
Golden Planets

Golden Planets

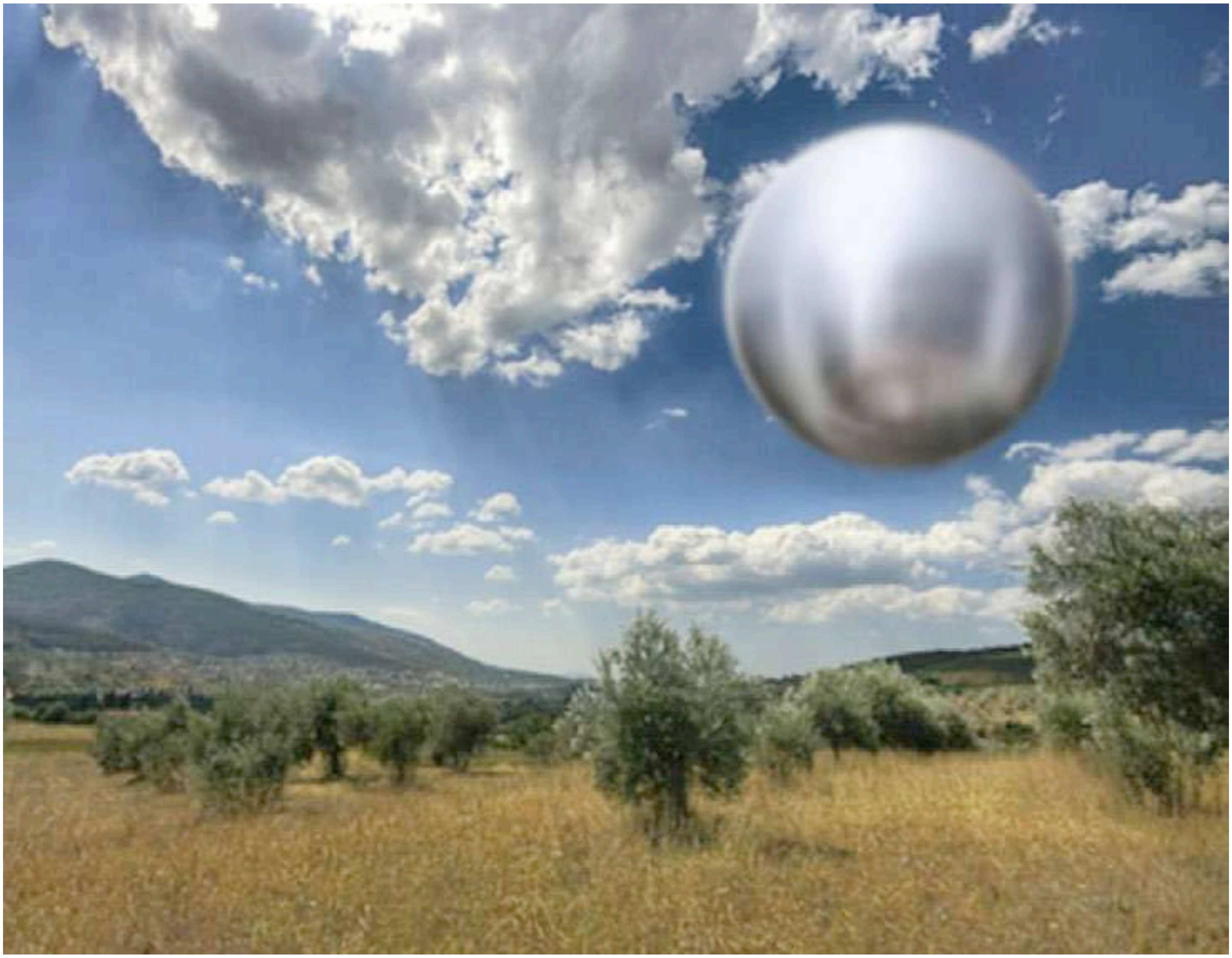

In one of the most intriguing cosmic journeys yet documented in our times, Michel Desmarquet was contacted by a giant, quite beautiful race of extraterrestrial human beings, and invited aboard their massive spherical spacecraft. Michel's cosmic journey began on a clear night in June of 1987, in the northeast part of Australia, when he was levitated above his home in a bluish plasma to be met by a giant figure in mid-air. Wearing a close-fitting single-piece suit, the huge figure smiled and explained the many unusual circumstances to Michel, before inviting him aboard the giant spacecraft that floated nearby:

> About 100 meters from us was an enormous sphere, from which emanated a bluish Aura. I later learned that it measured 70 meters in diameter. The light was not steady, but shimmered, resembling a heat haze when one looks from a distance at sand heated by the summer sun. This enormous sphere 'shimmered' about ten meters from the ground. With no windows, no openings, no ladder, it appeared as smooth as the shell of an egg.[1]

The giant being introduced itself to Michel as Thao, and explained the main purpose of their long journey from their gigantic homeworld, known as the golden planet of Thiaoouba. It was later identified as one of three such golden planets in our galaxy that represent the highest attainment of physical and spiritual advancement. Michel was invited aboard the spherical super long-range spacecraft, called an 'Alatora', and underwent a necessary set of bacterial decontamination processes with the implicit understanding that he was to join this race of gentle giants on a great journey. Michel was given the opportunity to visit their homeworld of Thiaoouba to speak directly with their High Council of 7 Elders, known as the Thaori.

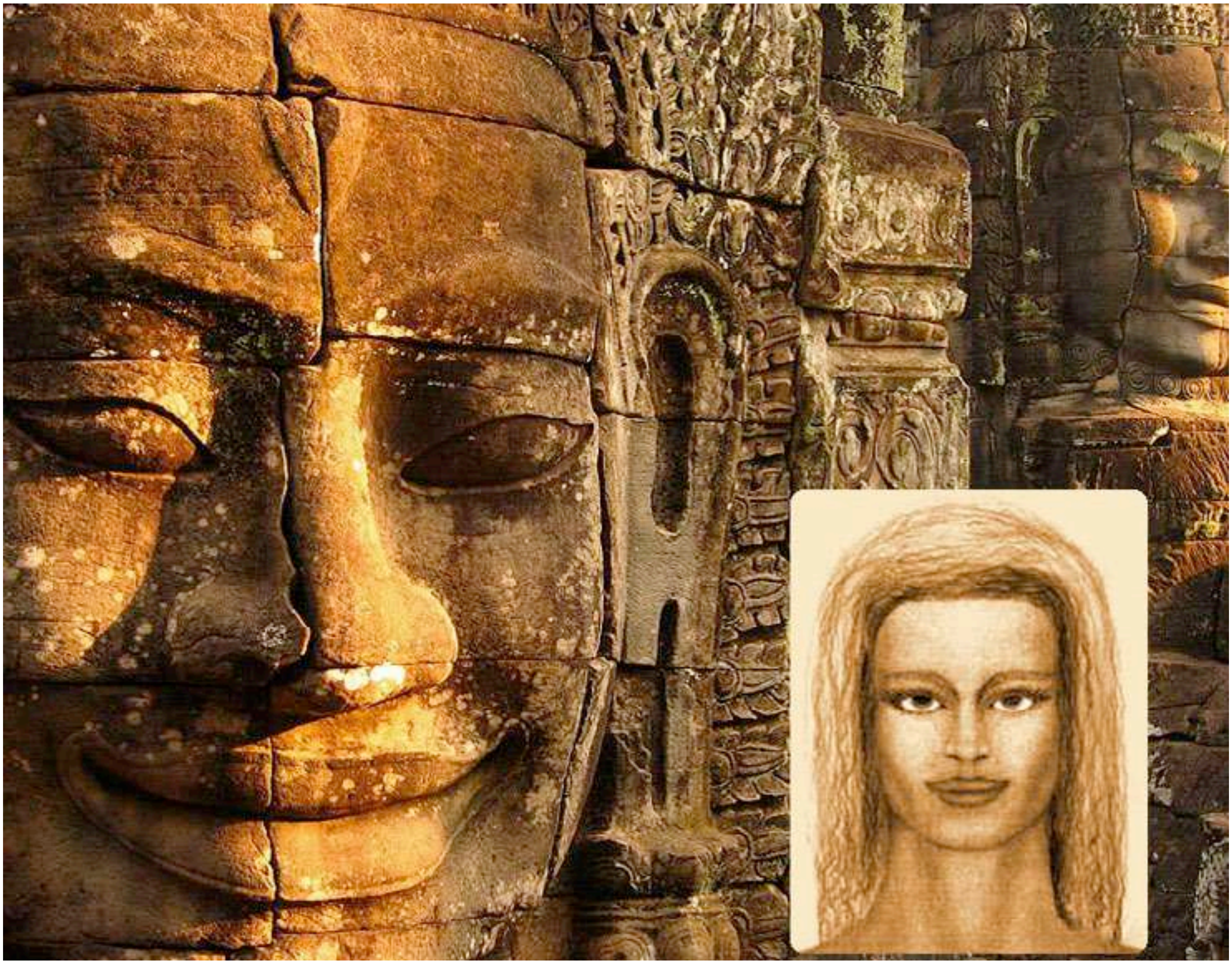

No photographic or physical evidence of Michel's cosmic journey was obtained, but his vivid descriptions, first published in 1993, provide a wealth of advanced spiritual information and detailed observations that in themselves provide the ultimate validity of his experiences. As Desmarquet clearly acknowledges, "this extraordinary story will appear to [be]… a story entirely invented – but I do not have the imagination that such a fabrication would require. *This is not science fiction.* The reader of good faith will be able to recognize the *truth* in the message I transmit from my new friends to the people of the planet Earth."

For a period of ten days, Michel became the pupil of Thao, a quite impressive personage of 9'8" (290cm) in height, who radiated a youthful wisdom and joyfulness, representing a unisex race of giant human beings inhabiting the planet of Thiaoouba. Being hermaphrodites, all Thiaooubans possess both male and female reproductive organs. Such giant extraterrestrial hermaphrodites may have been depicted by ancient artists in giant stonework, such as the temples of Angkor from the ancient Sanskrit traditions of present-day Cambodia (above, with inset drawing of Thao based on Michel's descriptions). Thao's intriguing facial features combine the elegance of very feminine eyes with a strong masculine chin:

> Thao… smiled at me and I clearly remember appreciating, for the first time, the beauty and nobility of her face… I have already mentioned her hair, which was long and silky, golden-blond in color, which fell to her shoulders and framed a face that was perfectly oval in shape. She had a large, slightly protruding forehead… Her blue-mauve eyes and long curled lashes would have been the envy of many women on our planet. Her eyebrows curved upwards, similar to the wings of a seagull… The chin, well shaped but slightly angular, suggested a willful determination that was somewhat masculine, but this did not detract from its charm. A faint shadow of hair above her upper lip could have spoiled this perfect face, were it not blond.[2]

> [Upon entering the giant ship, we] were now ‘gliding’ at ground level along a tunnel-shaped corridor that led to a small room, the walls of which were of a yellow so intense that I had to close my eyes… ‘It is not a light. It is just the present color of the walls… There is no paint. There are only vibrations, Michel. You still believe that you are in your Earthly universe, when you are not. You are now in one of our super long-range spacecraft, capable of traveling at several times the speed of light. We will be leaving soon, if you will lie down on this bunk?’
>
> There, in the centre of the room were two boxes - rather like coffins without lids. I stretched out in one of them and Thao in the other… I wanted to lift myself up a little but couldn’t, being held by an unknown and invisible force. The yellow color progressively disappeared from the walls, to be replaced by a blue that was certainly no less intense… [Thao explained the light intensity as a health precaution–] ‘We come from a planet [that is] bacteriologically different from Earth, which, for us, is a veritable culture medium. Thus, in order to contact you, I was obliged to take this basic precaution. You, yourself, were a danger to me but you are no longer… When you entered this cabin, the color was too intense for you and I gave you the helmet you are now wearing, which was specially designed for you…

> ‘During the very short time the cabin was yellow and then blue, eighty percent of the dangerous bacteria in you was destroyed. Then perhaps you felt a coolness in the air, similar to when an air conditioner is working; this was another form of disinfection by... let’s call it radiation, although that is not the correct word – it cannot be translated into any Earth language. In this way, I have been disinfected one hundred percent, but you still have enough bacteria to harm us considerably. I am going to give you these two pills, and in three hours you will be able to consider yourself as ‘pure’ as one of us.’
>
> As she spoke, she took a little box from beside her bunk, removed the pills and held them out to me, along with a test tube containing a liquid that I supposed to be water. I swallowed them both, lifting the base of my helmet to do so... ‘I have given you a special drug in order to cleanse your body of all bacteria that is dangerous to us. I have also given you another drug that has caused your Astral being to leave your body - this will last three hours, the time it will take to purify you. In this way, you will be able to visit our spacecraft, without danger of contamination to us and without wasting time.’ [3]

Thao’s explanation of the intense electroluminescence of the resonant metal walls of the Alatora spacecraft and the necessary preventive bioelectrification protocols can now be properly understood. Lightwater and a pill most likely containing exotic trace elements including electroluminescent nanoparticles of gold and silver became activated by the strong force-field which impeded his movement while reclining in the resonant chamber, illuminating his cellular waters with red and blue light. The precise environmental controls exercised inside the starship reflect the luminosity of Thao’s high-resonance homeworld of Thaioouba. While in Astral body, Michel took the opportunity to ask his impressive new friend a spiritual question of great importance: *“First of all, who is God? Does he exist?”*

> ‘I have been expecting this question and it is important that you ask it. On an ancient stone tablet, which I believe is *Naacal*, it is written: In the beginning there was nothing - all was darkness and silence.’
>
> ‘The Spirit - the Superior Intelligence decided to create the worlds and he commanded to four superior forces... It’s extremely difficult for the human mind, even when highly developed, to comprehend such a thing. In fact, in a sense, it is impossible. On the other hand, your Astral Spirit assimilates it when it is freed from your physical body. But I’m getting ahead of myself - let’s go back to the very beginning.’
>
> ‘In the beginning there was nothing except darkness and a spirit - THE Spirit. The Spirit was, and is, infinitely powerful - powerful beyond the comprehension of any human mind. The Spirit is so powerful that he was able, by the action of his will alone, to trigger an atomic explosion with chain reactions of unimaginable force. In fact, the Spirit imagined the worlds - he imagined how to create them - from the most enormous to the most minuscule. He imagined the atoms. When he imagined them he created, in his imagination, all that moved and will move: all that lived and will live; all that is motionless, or seems to be - every single thing.’

‘But it existed only in his imagination. All was still in darkness. Once he had an overall view of what he wanted to create, he was able, by his exceptional spiritual force, to create, instantaneously, the four forces of the Universe. With these, he directed the first and the most gigantic atomic explosion of all time - what certain people on Earth call ‘The Big Bang’. The Spirit was at its centre and induced it. Darkness was gone and the Universe was creating itself according to the will of the Spirit. The Spirit was thus, is still, and always will be, at the centre of the Universe for he is the Master and Creator of it...’

‘Well then,’ I interrupted, ‘it’s the story of God as the Christian religion teaches it - or just about - and I never believed in their nonsense...’

‘Michel, I speak of no religion such as exists on Earth and especially not the Christian religion. Don’t confuse religions with the Creation and the simplicity of all that ensued. Don’t confuse logic with the illogical distortions of religions. We will have the opportunity to talk again later on this subject, and you are certainly in for some surprises… For the moment, I was trying to explain the Creation to you. During billions of years (for the Creator of course, it is eternally the ‘present’, but it is more at the level of our understanding to count by billions of years), all the worlds, suns and atoms were formed, as you are taught in schools, the planets revolving around their suns, and sometimes with their own satellites etc. At certain times in certain solar systems, some planets cool down - soil is formed, rocks solidify, oceans are formed and landmasses become continents. Finally, these planets become habitable for certain forms of life. All this was in the beginning, in the imagination of The Spirit. We can call his first force the ‘Atomic force’.

‘At this stage, by his second force, he conceived the primary living creatures and many of the primary plants, from which later derived the sub-species. This second force we will call the ‘Ovocosmic Force’, as these creatures and plants were created by simple cosmic rays, which ended up with cosmic eggs.

‘At the very beginning, The Spirit imagined experiencing feelings through a special creature. He had imagined Man by means of the third force that we will call the ‘Ovoastromic Force’. Thus Man was created. Have you ever considered Michel, what intelligence it took to create a human being or even an animal? Blood that circulates thanks to the heart that beats millions of times independently of the will... lungs that purify the blood... the nervous system... the brain which gives the orders, aided by the five senses... the spinal cord which is ultra-sensitive… Have you ever wondered why, that of the billions of individuals on a planet like yours, there are no two fingerprints the same: and why, what we call the ‘crystalline’ of the blood, is just as unique among individuals as the fingerprint? …Your experts and technicians on Earth, and on other planets, have tried and still try to create a human body. Have they succeeded? In regards to the robots they have made, not even the most highly perfected will ever be more than a vulgar machine in comparison with the human mechanism.

> 'To go back to the crystalline I mentioned just now, it is best described as a certain vibration particular to the blood of each individual. [Scanning electron microscopy imaging of healthy red blood cells is presented below.] It has nothing to do with the blood group. Various religious sects on Earth believe absolutely in the 'rightness' of refusing blood transfusions. Their reasons relate to the teachings and books of their religious teachings and their own interpretations of these, whereas they should look to the real reason, which is the impact the different vibrations have on each other… If it is a large transfusion, it can have an influence on the recipient to a degree and, for a length of time, which varies according to the volume of the transfusion. This influence, of course, is never dangerous… After a time, which never exceeds one month, the vibrations of the recipient's blood takes over, leaving no trace of the vibrations of the donor's blood… It shouldn't be forgotten that these vibrations are much more a feature of the physiological and fluidic body than that of the physical body.' [4]

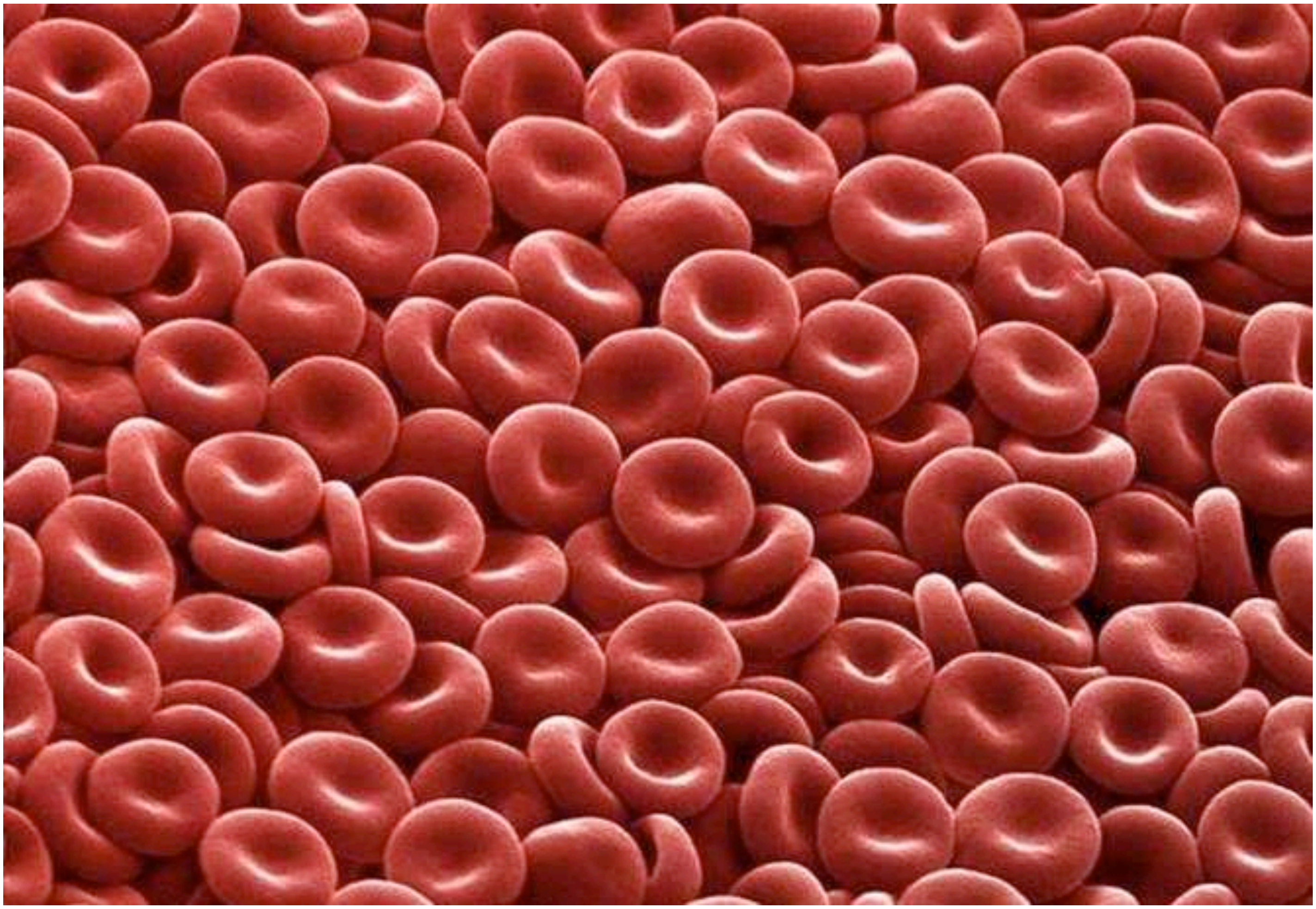

The straightforward scientific explanations provided to Michel Desmarquet by the cosmic beings of Thiaoouba guide Earthly human thinking toward the unified logic that Nature presents all around. The simple beauty of the lessons of the Creation offered by Thao rise in harmony with the teachings of Spirit transmitted to Earth by various other advanced masters in Nature, radiating their light throughout the great depths of the cosmos.

The concise explanation given concerning the formation of the cosmos through the four forces of Creation clarifies the quantum interactions of the short- and long-wave aspects of Creation, also referred to as the matter and anti-matter Universes. The Atomic force propels the material creation (*ultrasonic resonances*), the Ovocosmic force generates and maintains plant and animal life (*infrasonic resonances*) and the Ovoastromic force manifests through human life and consciousness. The fourth force was explained to Michel only later, after further necessary explanations were addressed: The Astral force. Thao's comparison of the fingerprint with the unique resonant frequency of the 'crystalline of the blood' of each individual is just one of the many fascinating pieces of information that lead human consciousness on Earth toward the deeper cognitions that will unfold with this great change of the ages.

Simple descriptions preserved in ancient Vedic texts belie a profound depth of awareness that confirms that visitations like the one described by Michel Desmarquet had also occurred in the Vedic Age of Atlantis, before being recorded on parchment in later times. *The only precedents among the historical documents of Earth for such a story as the one told by Desmarquet are to be found among the Sanskrit cultures of the world, including sacred Mayan and Egyptian hieroglyphic texts as well as Sumerian cuneiform writings, that record the teachings of highly spiritual visitors from the stars.* The great teachings of ancient prophets are being given anew through the contacts of Earthly human beings in our times, in preparation for the ascension of Earth into collective consciousness.

The scribes of Sumer also recorded the malevolent extraterrestrial interests of the Annunaki involving the extraction of gold, using Earthly human slave labor, for the ultimate purpose of enriching the atmosphere of their own planet.[5] This unusual and highly advanced concept of golden atmospheres and golden rivers is also discussed in the ancient Sanskrit writings of India. An intriguing dialog between King Pariksit and Sukadeva Gosvami preserved in the Srimad Bhagavatam describes the golden waters of heavenly planets. The Description of Jambudvîpa (Canto 5, Ch.16) identifies certain golden rivers on Earth (in the material world) and their benefits, as enjoyed on other *heavenly* planets such as Thiaoouba:

> The mud on both banks of the River Jambu-nadi, being moistened by the flowing juice and then dried by the air and the sunshine, produces huge quantities of gold called Jambu-nada. The denizens of heaven use this gold for various kinds of ornaments. Therefore all the inhabitants of the heavenly planets and their youthful wives are fully decorated with golden helmets, bangles and belts, and thus they enjoy life... (Text 21)
>
> The residents of the material world who enjoy the products of these flowing rivers have no wrinkles on their bodies and no grey hair. They never feel fatigue, and perspiration does not give their bodies a bad odor. They are not afflicted by old age, disease or untimely death, they do not suffer from chilly cold or scorching heat, nor do their bodies lose their luster. They all live very happily, without anxieties, until death. (Text 25) [6]

These ancient dialogs quite astonishingly detail all of the known antibacterial effects of bioelectrification using nano-colloidal gold and silver water, in addition to other benefits that remain unrecognized by contemporary scientists, such as an increased tolerance to temperature extremes. Such odd, enigmatic information contained in ancient writings concerning atmospheric gold and silver is now comprehensible in the context of the scientific use of nanoparticles for the illumination of DNA within all cells of the human body *and the ignition of atmospheric HHO plasma*, just as vividly described by Michel on Thiaoouba:

Suddenly, there it was, right in the middle of the panel... Thiaoouba! My guess had to be correct - I could feel it. Thao immediately telepathized an affirmative, leaving me in no doubt… As we approached and Thiaoouba grew on the panel, I couldn't take my eyes off the sight, for what I saw before me was indescribably beautiful. Initially, the first word which sprang to mind was 'luminous' - this then was juxtaposed with 'golden' - but the effect produced by this color was beyond any description. If I were to invent a word, perhaps one that might apply would be 'lumino-vapour-golden'. In fact, one had the impression of plunging into a luminous and golden bath – *almost as if there were very fine gold dust in the atmosphere.*

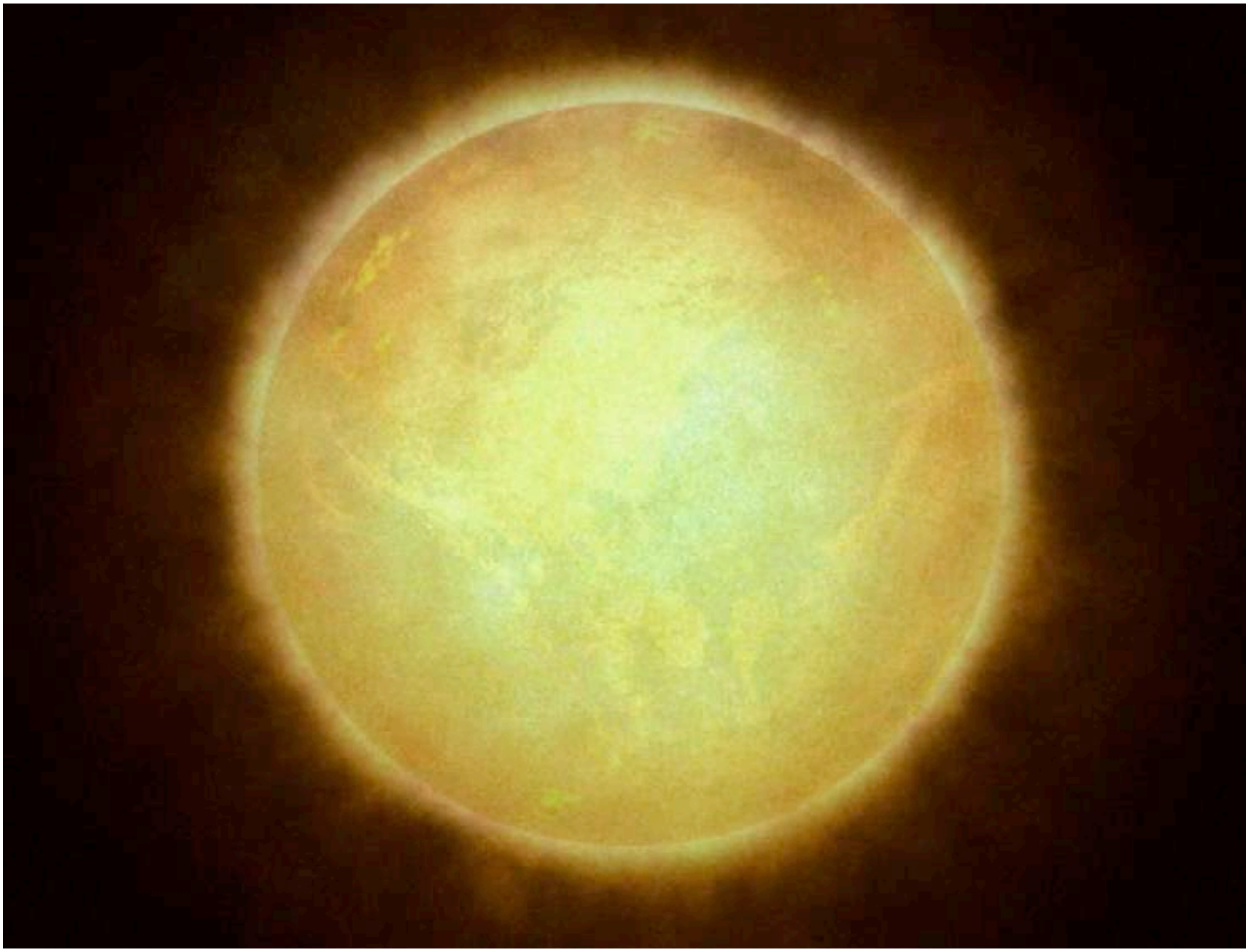

We were descending gently on to the planet and the panel no longer showed its contours, but rather, the contours of a continent could be distinguished, ending abruptly at an ocean that was sprinkled with a multitude of islands of different colors… The closer we came, the greater was the detail discernible… What captivated me most, was the color before me – I was dazzled! All the colors were, in each tonal variation more vivid than ours. A bright green, for example, almost shone – it radiated color. A dark green had the opposite effect – it 'kept' its color. It is extremely difficult to describe, for the colors on this planet could not be compared with any that exist on Earth. A red could be recognized as red, but it wasn't the red we know. There is a word in Thao's language which defines the types of colors on Earth and on planets similar to ours. Our colors are Kalbilaoka which I translate as 'dull', theirs, on the other hand, are Theosolakoviniki which means they radiate their colors from within.

My attention was soon drawn to what looked, on the screen, like eggs - yes, eggs! I could see the ground dotted with eggs, some half covered with vegetation, and others quite bare. Some seemed bigger than others and some lay down. Others were upright with what looked like the pointed end towards the sky. I was so astonished by this sight that I turned again towards Thao to ask her about these 'eggs', when suddenly on the screen, there appeared a round form surrounded by several spheres of different sizes, and, slightly further away, yet more 'eggs'. These ones were enormous. I recognized the spheres to be space vessels just like ours.

‘Affirmative’, said Thao from her seat, ‘and the round form you see is the cell in which our spacecraft will be accommodated in a few moments, for we are in the process of docking.

‘And the gigantic eggs, what are they?’

Thao smiled. ‘Buildings, Michel, but just now, there is something more important that I must explain to you. Our planet contains many surprises for you, but there are two, which could have harmful effects on you. I must therefore ensure that you take certain elementary precautions. Thiaoouba does not have the same gravitational force as your planet. Your weight would be 70 kilograms on Earth - here it will be 47 kg. When you leave the spacecraft, if you aren’t careful, you risk losing your sense of equilibrium in your movements and your reflexes. You’ll be inclined to take too great a stride, and perhaps fall and injure yourself...’

‘But I don’t understand. In your spacecraft, I feel fine.’

‘We have related the internal gravitational pull to correspond with that of Earth – or almost… It’s true that under this force, our bodies are heavier, but we have counter-balanced this by a semi-levitation, thus we are not uncomfortable, and at the same time, we have the satisfaction of seeing you move among us at ease.’ A slight jolt indicated we had docked. This extraordinary trip was over - I was going to put my foot on another planet.

‘The second point,’ Thao resumed, ‘is that you will be obliged to wear a mask, for a while at least, for the brightness and the colors will literally intoxicate you, just as if you had drunk alcohol. The colors are vibrations that act on certain points on your physiological body. On Earth, these points are so slightly stimulated, so little exercised, that here, consequences could be unfortunate…’ There, she took a helmet, very light, which covered my face from my forehead to just below my nose.

‘Let’s go, Michel, and welcome to Thiaoouba.’ Outside the spacecraft, we walked along a very short walkway. Immediately, I felt lighter. The sensation was very pleasant, although somewhat disconcerting, since several times I lost balance and Thao had to steady me.

We saw no one, a fact that surprised me... After a short distance, we arrived at a round platform, to the side of the walkway. Thao sat down on a circular seat inside the platform and signaled that I should sit down opposite her… She took out an object the size of a walkie-talkie and immediately I felt myself pinned to the seat, just as I had been in the spaceship, by an invisible force field. Then, quite gently and with a barely perceptible hum, the platform rose by several meters and moved off rapidly towards the ‘eggs’, about 800 meters away. The thin and slightly perfumed air was lashing the exposed area of my face below my nose, which was very nice, its temperature being around 26 degrees Celsius.

In just seconds, we arrived, and passed through the walls of one of the ‘eggs’, as if we’d passed through a cloud. The platform stopped and came to rest gently on the floor of the ‘building’. I looked around me in all directions… It seemed absurd, but the ‘egg’ had disappeared. We had indeed, entered the ‘egg’ and yet around us, as far as the eye could see, stretched the countryside. We could see the landing ground and the docked spaceships just as though we were outside... Not far from us, were gathered twenty or thirty people, all busy to some extent, in front of desks and screens which flashed with colored lights - similar to inside the spacecraft. A type of music played very softly, elevating me to a state of euphoria…

Michel Desmarquet's thoroughly detailed and fascinating description of the visual splendor of the golden planet of Thiaoouba contains a wealth of information for scientific consideration, as well as a direct reference to the abundance of atmospheric gold (nanoparticles) visible throughout the luminous skies of Thiaoouba! This vaporous luminosity of Thiaoouba's atmosphere is created by light amplification effects of the HHO plasma that forms around the gold nanoparticles, which electroluminesce under the influence of atmospheric electrical currents moving through the resonant field of infrasound standing waves.

Thao's description of the effect of high-intensity colored light on the human body is also quite intriguing. Informing Michel that "the colors are vibrations that act on certain points on your physiological body", Thao offers an only slightly veiled reference to the pineal gland to the center of the human brain – also known as the 'third-eye' or *ajna chakra* among ancient Sanskrit traditions. The pineal gland is truly an *eye*, being photosensitive, and is responsible for the alignment of the body's circadian rhythms with the sun by light reception in the center of the brain.[7] Thao later explained in great detail the role of the pineal gland as one of the six principle points within the chakra system of the human body:

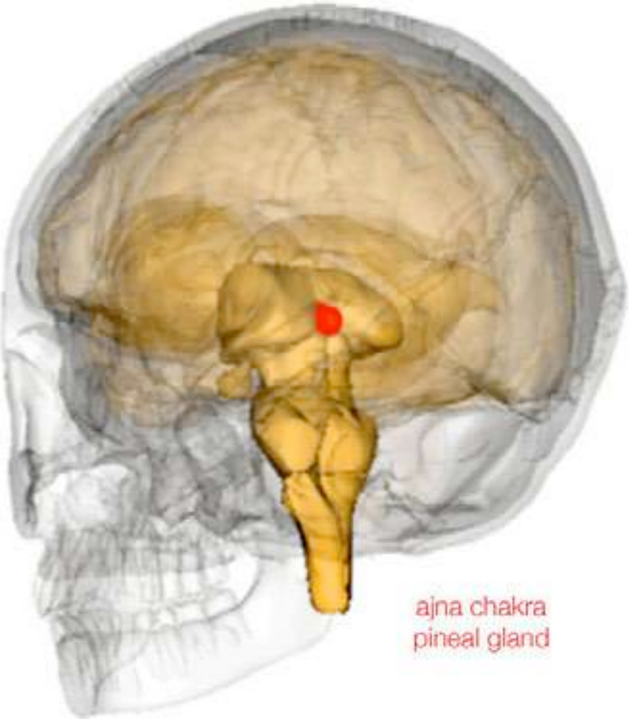

> 'The fluidic body [includes the brain stem and spinal cord, and] influences the physiological body which, in turn, influences the physical body… In the fluidic body, you possess six principal points which we call Karolas and which the yogis on your planet call Chakras. The first Chakra is the one situated between your two eyes, just a centimeter and a half above your nose. It's the 'brain' of your fluidic body, if you like; it corresponds to the pineal gland [in red, at left], which is placed much further back in your physical brain but on exactly the same level…
>
> 'Now, at the bottom of the fluidic body and just above the sex organs, is found a very important Chakra, which we call the Mouladhara, and which your yogis call Sacred [or Sacral]. Above this Chakra, and meeting the spinal column, is the Palantius. It is in the form of a coiled spring and only reaches the base of the spinal column when it is relaxed.
>
> 'For it to become relaxed, it requires the accomplishment of the sexual act between two partners who must not only love one another, but also have a spiritual affinity between them. Only at that moment and under these conditions will the Palantius extend to the spinal column, transferring an energy and special gifts to the physiological body which then affects the physical body. The person concerned will experience happiness in sexual enjoyment that is far greater than normal.' [8]

Thao identifies the shape and function of the *palantius* in much the same way as the coiled Kundalini serpent of Sanskrit teachings, and further reveals the special conditions for such an activation include not only partnership in love, but also *true spiritual affinity*. As proven by research scientists, such spiritual affinity takes the form of a *heart/mind resonance between the two*,[9] whereby releasing the palantius.

The identification of the chakras as existing within the fluidic body is also very concise, meaning within the spinal fluid itself. While the definition of the pineal gland as the first chakra presents an inverse order when compared with Sanskrit traditions, the Thiaoouban name of the root chakra –the Mouladhara– is identical to Sanskrit terminology. The attunements of Michel's chakras were so powerfully influenced by the luminosity of the high-resonance atmosphere of Thiaoouba that, combined with the intensity of ambient infrasound and the resulting decreased weight, his state of consciousness was elevated to euphoric levels that are only momentarily experienced by human beings on Earth, if ever. The highly elevated levels of all frequencies of electromagnetic radiation also required longer periods of rest.

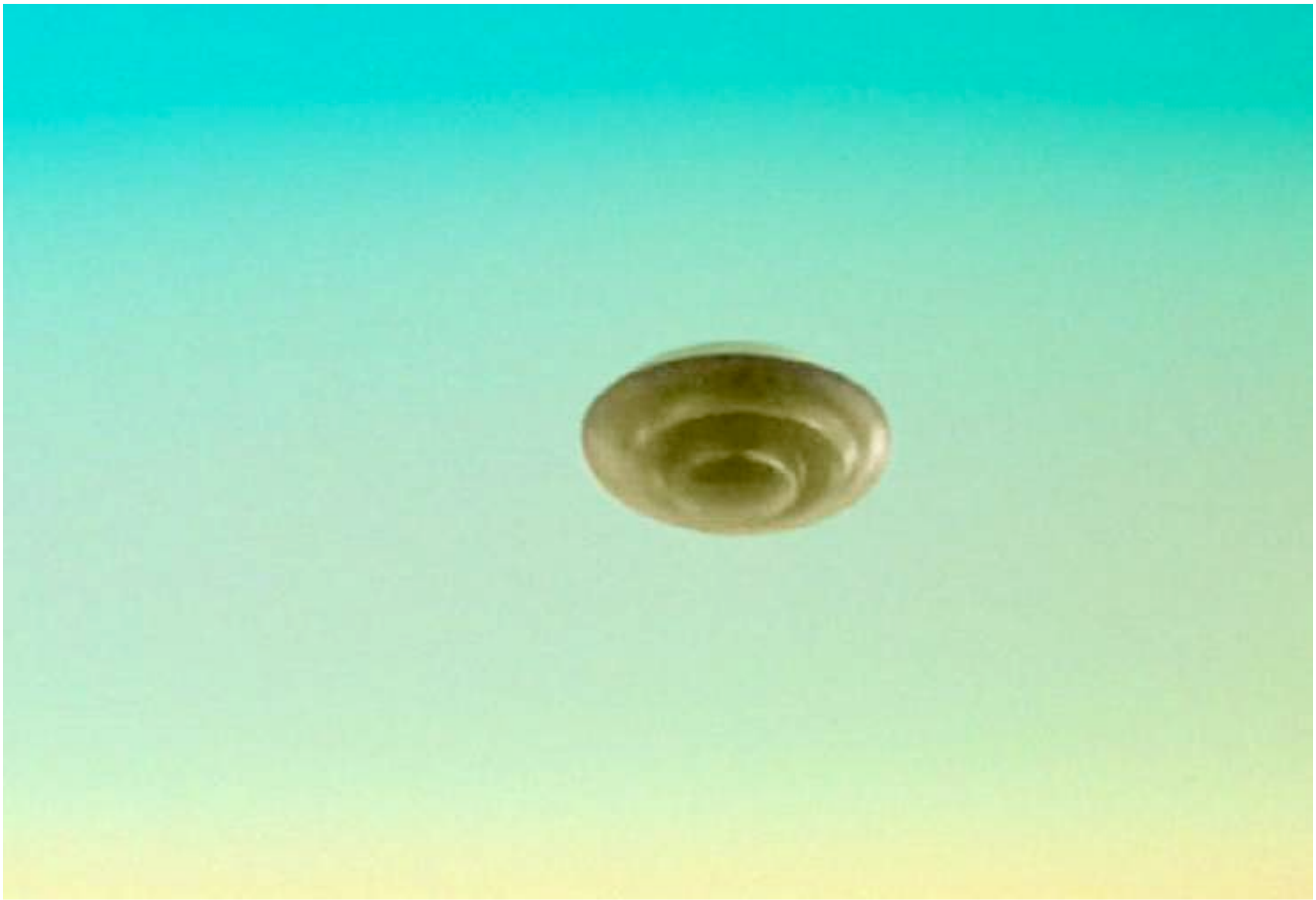

The profound attainment of knowledge witnessed on Thiaoouba is reflected in the delicacy of colors and radiating brilliance of every life-form preserved there in harmony with the advanced technological lifestyle of the giant human inhabitants. Ever-present forms and forces within nature inspire the technological creations observable on this planet, whose buildings take on the form of eggs, and plasma spacecraft display a spherical perfection. The flying circular platform used for Michel's magnificent tour of Thiaoouba (illustrated above) is resonantly designed to generate the barely perceptible humming sound that Michel described during lift-off. Invisible magnetic fields are generated as belts that act as safety harnesses, and a domed magnetic field forms a protective covering enclosing the occupants of the open platform aircraft.

The grand scale of the trees and flowers, as well as the insects and animals, is a reflection of the enhanced gravity conditions of Thiaoouba, which are similarly witnessed in the geological record of Earth. During the age of dinosaurs, all life-forms on earth were correspondingly larger than witnessed today. Insects were once several times larger than the same types now living and would be crushed by the present conditions of gravity were they to be found on the planet today. The relative conditions of planetary gravity are directly tied to the intensity of atmospheric infrasound standing wave resonance.

> We rejoined our vehicle and immediately accelerated away in the direction of a forest, which could be seen in the distance. We flew at a height of approximately five or six meters and at a speed I would have estimated to be 70 or 80 kilometers per hour. The air was warm and fragrant and I again felt euphoric, in a way I had never experienced on Earth. We arrived at the edge of the forest and I remember having been greatly impressed by the size of the largest trees. They looked to rise about 200 meters into the sky.
>
> 'The tallest is 240 of your meters, Michel.' Thao explained without me having to ask, 'and between 20 and 30 meters in diameter at the base. Some of these are 8000 of our years old. Our year consists of 333 days of 26 karses. A karse is a period of 55 lorse, a lorse comprising 70 kasios, and a kasio being almost equivalent to one of your seconds…

The vehicle greatly reduced its speed and we were able to glide between, or indeed, stop and observe more closely, the trees at heights that ranged from almost ground level to 10 meters above the ground. Thao was able to guide our 'flying platform' with amazing precision and expertise. Our vehicle, and Thao's manner of driving it, put me in mind of a flying carpet, which was taking me on a magical tour of this magnificent forest… Thao leaned towards me and removed my mask. The undergrowth was luminous and softly golden but I found it quite tolerable… 'It is a good time to begin accustoming yourself to the light and color, Michel. Look!'

Following her gaze, I spotted, very high among the branches, three butterflies, vividly colored and of enormous size. These lepidoptera, which must have had one meter wingspans, fluttered high in the foliage, but we were lucky to see them fly closer and closer to us, on wings of blue, green and orange. It is as clear to me as if it were yesterday. They brushed against us with their wings that were strangely fringed, to create the most beautiful and breathtaking effect. One of them came to rest on a leaf just a few meters from us and I was able to admire its body, ringed with silver and gold, and its jade-green antennae. Its proboscis was golden and the tops of its wings were green with streaks of bright blue alternated with dark orange diamond shapes. The under-sides were dark blue, but luminous, as though they had been illuminated from above by a projector.

For the duration of time this giant insect remained on the leaf, it seemed to emit a soft whistling sound and I was quite surprised by this. I had certainly not heard a lepidopteran on Earth make any sound at all. Of course, we were no longer on Earth but on Thiaoouba, and this was only the beginning of a long series of surprises for me. On the forest floor, grew an incredible variety of plants, each one stranger than the next. They covered the ground completely, but I noticed very few bushes among them… the forest's giants prevented them from developing.

In size, these plants varied from a ground-covering moss-like plant, to one the size of a large rose bush. One kind, with leaves as thick as a hand in various shapes - sometimes heart-shaped or circular, sometimes very long and thin - was of a color tending much more towards blue than to green. Flowers of every shape and color, even of the purest black, entwined each other. From our altitude of several meters, the effect was absolutely glorious. We rose till we were up among the highest branches and I put my mask back on according to Thao's direction. We emerged from the canopy and moved slowly, just above the foliage of those enormous trees.

Above the forest the light was, once again, incredibly intense and I had the impression of traveling through a landscape of pure crystal. Marvelous birds were perched on the tops of the taller trees, watching us pass, without fear. Their colors, varied and rich, were a veritable feast for my eyes in spite of the subduing effect of my mask. Here were varieties of macaw, with blue, yellow, pink and red plumage; and, among them a type of bird of paradise strutted amidst a cloud of what appeared to be hummingbirds. These hummingbirds were of a brilliant red color flecked with gold. The red, pink and orange tail feathers of the birds of paradise, would have measured 250 centimeters in length and their wingspans almost two meters.

When these ‘jewels’ took flight, the underside of their wings revealed a very soft, misty pink, with just a touch of bright blue on the tips - so unexpected, especially as the tops of their wings were of an orange-yellow color. Their heads wore plumes of impressive size, each feather being a different color: yellow, green, orange, black, blue, red, white, cream... I feel frustrated that my attempts to describe the colors I saw on Thiaoouba are so inadequate - I feel I need a whole new lexicon, as my language fails me. I had the constant impression that the colors came from within the objects I looked at, and the color was more than I had known it to be. On earth, we know perhaps 15 shades of red; here there must have been over a hundred...

It wasn’t only the colors that claimed my attention. The sounds that I had heard since we began to fly over the forest inspired me to seek an explanation from Thao. It was almost a background music, very light and soft, similar to a flute which continually played the same air but at a distance. As we moved on, the music seemed to change, only to return to the original tune. ‘Is that music I hear?’

‘It is vibrations emitted by the thousands of insects which, when combined with the vibrations of the colors reflected by solar rays on to certain plants, such as the Xinoxi, for example, produce the very musical results that you hear. We, ourselves, only hear it if we particularly attune to it, for it comprises an integral part of our life and our environment. It is restful, isn’t it? … According to the experts, if these vibrations were to cease, we would experience considerable eye trouble. This will perhaps seem odd at first, since these vibrations appear to be perceptible to the ear rather than the eye. However, …the chance of their ceasing is as remote as the chance of our sun extinguishing itself tomorrow.’

> Thao turned our vehicle and in a few moments, we had left the forest tops and… I saw we were approaching the ocean. Palm trees resembling coconut palms waved their majestic fronds at impressive heights, on the edge of a beach of golden sands… A hundred or so people basked on the sand or swam, entirely naked, in the transparent waters of the ocean. I felt a little dazed, not only because of the new and wondrous things I was constantly discovering, but also because of the perpetual sensation of lightness, due to the change in gravity. This sensation was my reminder of Earth – what a strange word, and how difficult it was to visualize Earth now! The auditory and visual vibrations were also affecting my nervous system enormously. Usually a highly-strung person, I was feeling completely relaxed… –so relaxed I felt like crying.[10]

The perfection of natural beauty and harmony of life on Thiaoouba resonated deeply with Michel, its intense radiations bathing him in a field of love that has been raised to the highest resonance:

> The sun was already quite low and the sky was totally clear. The ocean had taken on an orange color, which was surprising. I never would have imagined water could appear such a shade. Inquiring about it, telepathically, it was explained to me that, sometimes, at this time of day, immense patches of orange colored plankton would rise to the surface. These waters, it appeared, contained enormous quantities of plankton. What a sight it was: the sky was blue-green, the sea was orange, and everything was enveloped in the golden light which, on this planet, seemed to come from nowhere and everywhere.[11]

The intense luminous qualities described by Michel can only be explained as an amplification effect of diffuse atmospheric HHO plasma generated by electroluminescent nanoparticles of silver and gold suspended in the planet's strong enveloping field of infrasound standing waves. The tranquil weather patterns well described throughout Michel's nine-day adventure on Thiaoouba are a reflection of the planet's highly balanced atmospheric systems. Desmarquet was also taken by the intense nighttime optical effects of Thiaoouba's many-layered canopy of atmospheric HHO plasma, which amplifies all starlight as seen from the ground, creating a magnificent opportunity for direct stellar observation:

> No one spoke; we sat in silence. Looking up, I could see the stars appearing one by one, shining colorfully as though a firework display had ‘frozen’ in the sky. On Thiaoouba, because their layers of gasses in the atmosphere differ from ours, the stars appear to be colored and also much larger than they appear to us on Earth.[12]

The color of each particular star is amplified as it passes through many atmospheric layers of HHO plasma, formed from water vapor that becomes ionized as a glowing plasma canopy. HHO plasma forms a barely visible network of thin veils suspended along the pathways of infrasound standing waves where gold and silver nanoparticles are kinetically energized, allowing the splitting of water molecules into single atoms of hydrogen and oxygen (HHO gas). The ancient Sanskrit knowledge of ‘heavenly planets’ is in complete correspondence to the experiences so richly described by Michel Desmarquet on the golden planet of Thiaoouba! The natural domed plasma canopies of heavenly planets form the divine inspiration for the unified egg-shaped design of all habitations, temples and spaceports, reflecting the *Ovocosmos*:

> The vehicle proceeded slowly at ground level and, after a final bend in the path, we arrived before a ‘little egg’, nestled among small trees and flowering vines. It seemed that every building on this planet had the shape of an egg, most often lying on their ‘sides’ but occasionally upright, as I have said, with the pointed end upwards. The ‘shells’ were off-white in color and had no windows or doors. This particular egg lay on its side, apparently half buried in the ground. It was about 30 meters long and 20 meters in diameter - quite small in comparison with those I’d seen so far.
>
> Thao stopped the vehicle in front of a bright light centered on the wall of the egg. Leaving the platform, we entered the habitation. As we did so, I felt a light pressure, with no more force than the weight of an eiderdown. I remembered experiencing the same sensation earlier, when we passed through the wall of the space centre… Having neither doors nor windows on these buildings is extraordinary in itself, but once inside, it was stranger still. As I have mentioned before, the overall impression was of still being outside.

> The startling beauty of color was everywhere; the greenery; the branches of the trees dissecting the blue-mauve sky above; the butterflies; the flowers...I recall a bird that came to rest right in the middle of the roof, so that we could see the bottom of its feet. It was as though it had miraculously stopped in mid-air – the effect was quite extraordinary… The only contrast with the outside was provided by the floor that was covered in a type of carpet on which was arranged comfortable looking seats and large pedestal tables. All of these furnishings were, of course, on a large scale - appropriate to these 'large-scale' people.

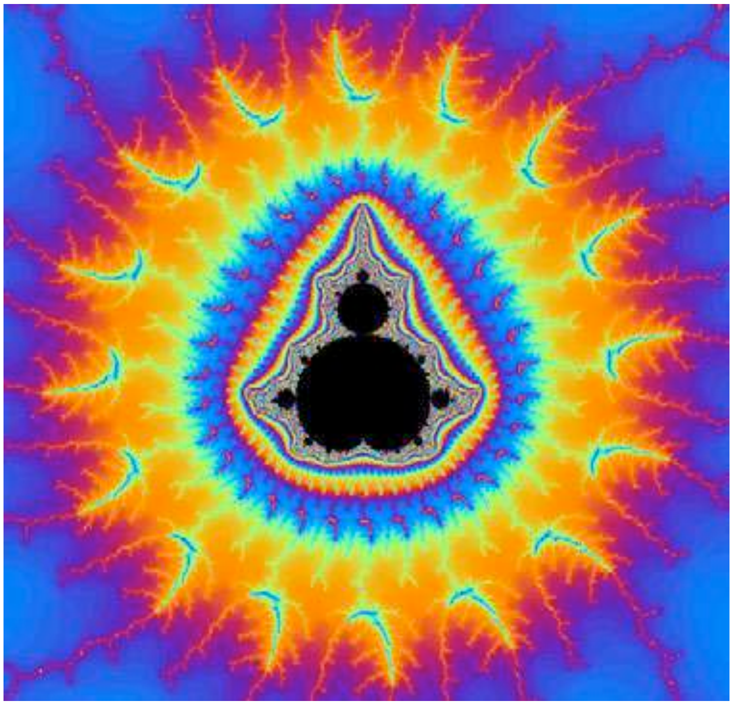

> 'Thao,' I asked, 'how are your walls transparent and yet we can't see in from the outside? And how can we pass through your walls as we did?'…
>
> 'You see, Michel, this habitation exists because of a magnetic field that is quite special. We have copied the forces of nature and the creations of nature to our own ends. Let me explain. Every body - human, animal or mineral - possesses a field around itself. The human body, for example, is surrounded both by an Aura and by an etheric force (field) of oval shape [at left]. You know that, don't you?' I nodded.
>
> 'The latter comprises, in part, electricity and, to a greater extent, vibrations that we call Ariacostinaki… These vibrations occur continuously for your protection while you are alive, and they are not to be confused with the vibrations of the Aura. With our habitations, we have copied nature in creating a field of mineral electro-etheric vibrations around a nucleus.' Thao indicated an 'egg', the size of an ostrich egg, located in the middle of the room between two seats… [as the nucleus.]
>
> 'This way,' she said, smiling again and taking several steps in another direction. 'Do you see this transparent drawer? Inside you have various compartments. From left to right: fish, shellfish, eggs, cheese, dairy products, vegetables and fruits, and here in the last, we have what you call 'manna', which is our bread…'
>
> 'You tell me that, on Earth, this bread is known as 'manna'. How is it that it exists on Earth at all?'
>
> 'It is a product we always carry on our intergalactic spacecraft… very practical… easily compressed and highly nourishing. In fact, it's a complete food. It comes from wheat and oats and you could live on it alone for months.'[13]

Michel's unusual descriptions of the giant egg-like domed magnetic fields of the buildings of the people of Thiaoouba reflect a simplicity and natural beauty that is profoundly inspiring. Thao's concise explanation of the etheric force field surrounding the ostrich-egg sized nucleus of the building was very carefully given in terms understandable to Michel, yet can be further understood in scientific terms. Each domed building is located at precise focal points of resonant infrasound standing waves, which become transduced by the resonant metal alloys of the nucleus into a vast electromagnetic field dome of the same shape.

The sacred *oval* or *egg* shape is repeated throughout Nature as reflections of the structure of the Universe, *the great ovum*. The *Ovocosmic* force can be described structurally by the Mandelbrot Set quadratic function [$z_{n+1} = z_n^2 + c$] (above), and the closely related function for nonlinear standing waves: [$z_{n+1} = z_n^2$]. The entire design of the Thiaoouban spiritual center of the City of Nine Dokos, home to the Seven Masters of the Thaori, conforms to this sacred cosmic force of resonant infrasound (opposite).

Thiaoouban *manna* bread was described as consisting of oats and wheat, while the electrical properties of invisible constituents including nanosilver and gold are indispensible for space travel, enabling the transdimensional effect of superconductivity of the human body for hyperleaps across the Universe. Gold and silver together comprise the ancient alloy *electrum*, so named for high electrical conductivity. Vedic scholars recently identified the Sanskrit term for electrum as *'Soma'*, described as granting eternal life through the luminosity of the electrum nanoparticles consumed in both foods and drinks:

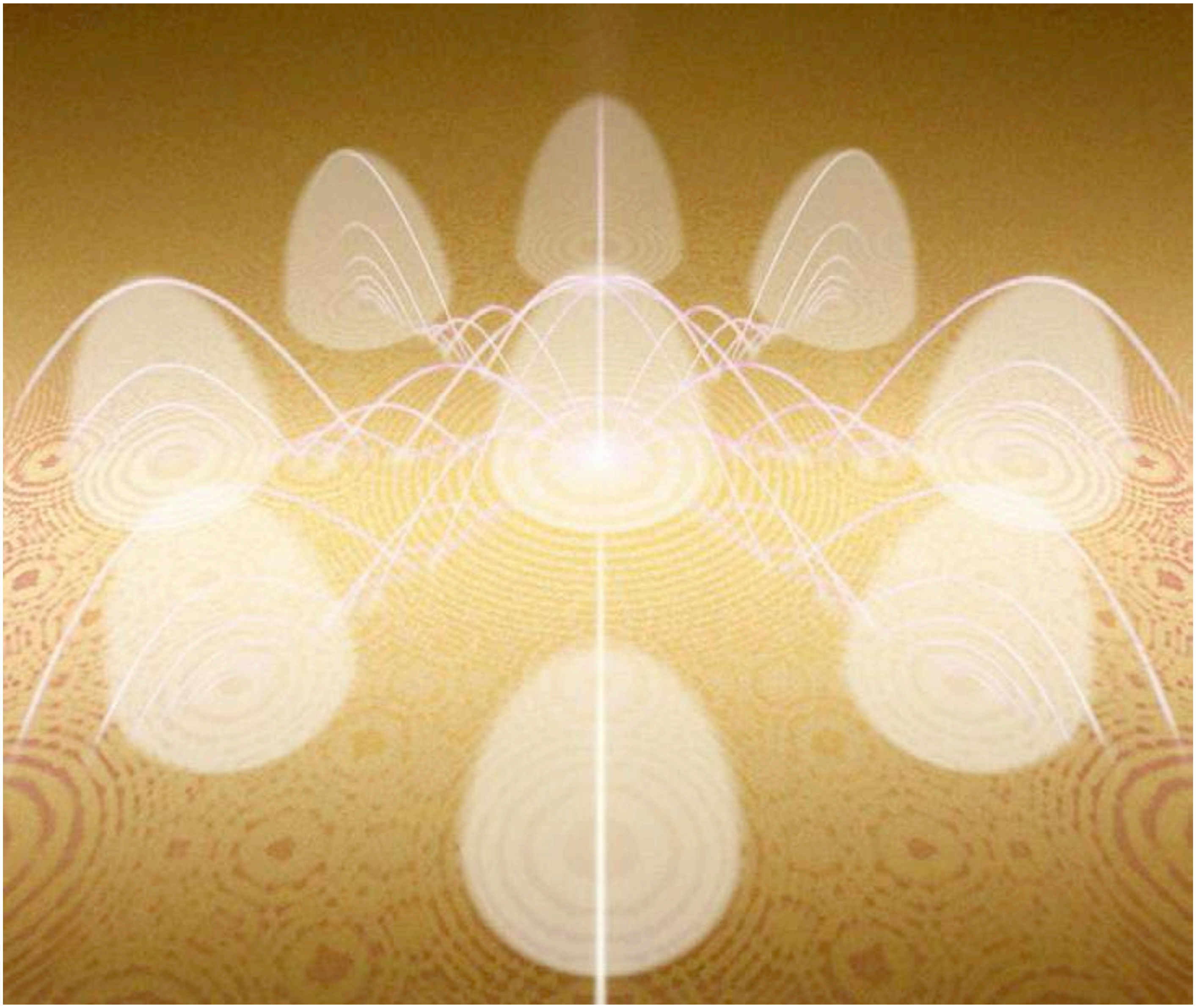

> “We have drunk the Soma and become immortal. We have attained the light the gods discovered.” (Rig Veda: VIII.48.3)… “Soma is food for the gods. Gods eat Soma.” (Chàndogya Upanishad: V.10.4)… “The man who supplies food hath always [his Soma] pressing stones adjusted, a wet Soma filter, well-prepared religious rites… he who hath this knowledge wins the luminous spheres.” (Atharva Veda: IX.6)… *Soma*, *Soma* ma¯al, asemon, asem, *electrum* may perhaps denote the same substance… (which required the purificatory 'mineral waters') contained in the kaman.d.alu symbols in the icons of the yaks.a legacy. It may perhaps be the same substance said to be am°tam which was considered to be the elixir of life, of immortality. It may perhaps be the same substance referred to, in sheer poetry, as amritam *àyur hira¯yam*. Gold is immortality.[14]

Vedic hymns clearly identify the intracellular luminosity of electrum-enriched Soma waters and foods as endowing eternal life. Vedic wisdom also relates the yogic practice of levitation using the bodily force of *‘laghima’* –an ability that Thao suddenly demonstrated for Michel while selecting Soma food preparations:

> ‘We can, as readily, lift ourselves in the air, but that serves no great purpose other than our own amusement.’ Having said that, Thao, who was sitting cross-legged, began to rise above her seat and floated about the room, finally coming to rest in mid-air. I stared at her, but soon realized I was the only one fascinated by her accomplishment. Indeed, I must have looked idiotic, for all eyes were fixed on me. Evidently, Thao’s behavior was perfectly normal to my friends but they were more interested in the astonished expression on my face… Thao descended slowly on to her seat. ‘That demonstrates one of the many sciences you have lost on Earth, Michel - apart from a few individuals who are still capable of doing it… There was a time when it was practiced by many, along with many other skills.’ [15]

The demonstration of levitation given by Thao was one of many special lessons given to Michel on the golden planet. On Earth, this bioenergetic art form was practiced as one of the attainments, or *siddhis* known to Sanskrit adepts, in addition to being able to make one's body become larger or smaller, invisible or radiant, or instantaneously translocate or bilocate (being in two places at once). Thao would make demonstrations of other siddhis later during her discussions with Michel, and proceeded with a lesson on auras and color coordination for the balancing of physical health and consciousness:

> 'The colors of your robe were chosen according to those of your Aura; that's why you feel so good. If people on Earth were able to see Auras, they too could choose colors, which suit them and thereby enhance their feeling of well-being. They'd make use of color rather than aspirin…'
>
> 'I'll give you an example. Do you ever remember saying of someone: 'oh, those clothes don't suit you at all. He, or she, has no taste'? …If you were to suggest it be because of the colors they were wearing, they would think you odd. You could explain that the vibrations of the colors were in discord with those of their Auras, but they would be no more inclined to believe you. On your planet, people only believe in what they see or touch... and yet the Aura can be seen... The Aura vibrates constantly with colors that vary. At the top of your head is a veritable bouquet of colors, where almost all the colors you know are represented.
>
> 'Around the head too, is a golden halo, but it is only really obvious in the most highly spiritual people and in those who have sacrificed themselves in order to help someone else. The halo resembles a golden mist, much like painters on Earth depict the haloes of 'saints' and of Christ. The haloes were included in their paintings because, in those times, some of the artists actually saw them… The colors are all there in the Aura: some shine more strongly, others are dull. People in poor health, for example, or people with bad intentions... Many people on Earth could see it and read it a very long time ago, but there are few now. Calm yourself, Michel. You will see it, and not only one but several, including your own.'…
>
> Approaching the coast, I could recognize enormous dokos, their points as usual towards the sky, I counted a group of nine, but the island was sprinkled with others, smaller and less visible amongst the vegetation. Thao took us higher and we were soon flying over what Thao called Kotra quo doj Doko - 'The City of the Nine Dokos'. Skillfully, Thao brought us down between the dokos, to a beautiful park located in their midst. In spite of my mask, I was aware that the golden mist that enveloped Thiaoouba was much denser around these dokos than elsewhere. Thao confirmed I was not mistaken in my perception…

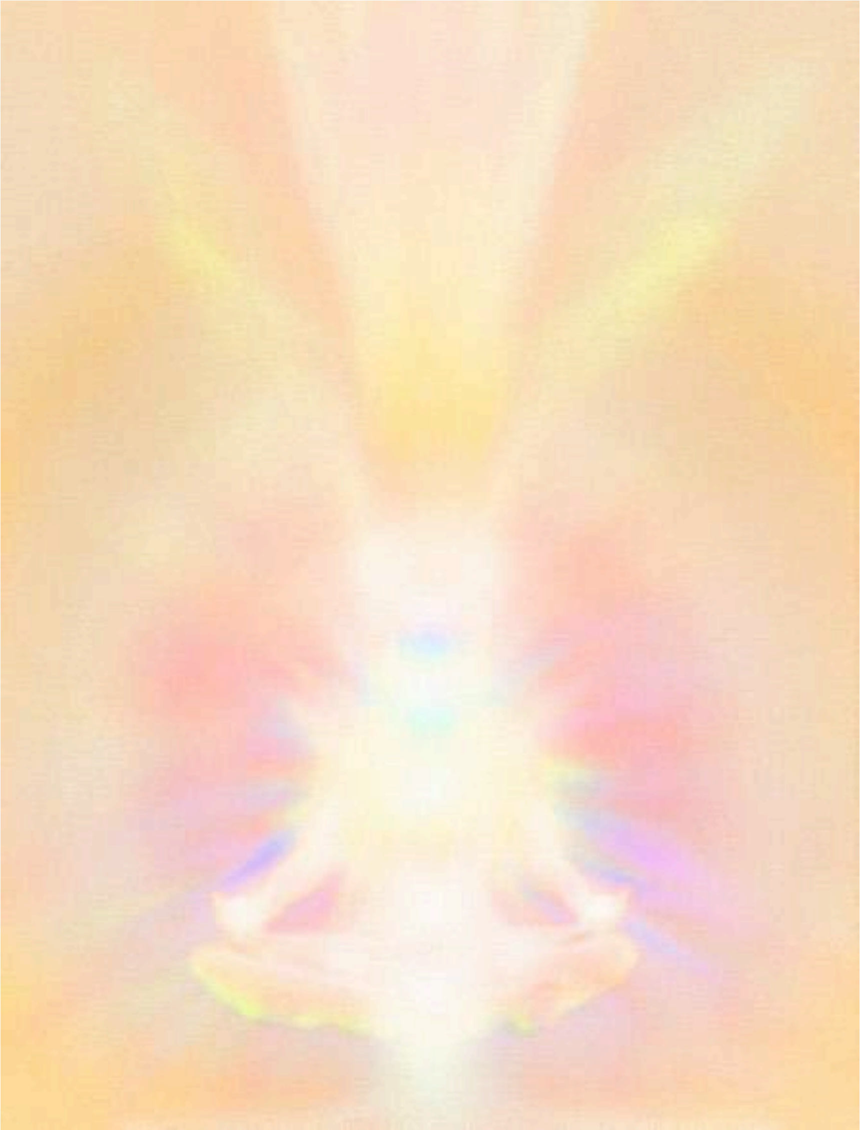

Eventually, we reached the Central Doko and stopped below the entrance light. Thao seemed to concentrate for several seconds, then she took me by the shoulder and we passed through the wall… She immediately removed my mask, advising at the same time, that I half-closed my eyes, which I did. Light filtered through my lower lids and after a time, I was able to open my eyes normally… I must say, that this brightness, more golden than in my own doko, was considerably uncomfortable at first…

This doko must have been 100 meters in diameter. We headed directly, although more slowly, for the centre, where seven seats, each occupied, were arranged in a semi-circle. The occupants sat as though petrified and, at first, I thought them to be statues… In looks they resembled Thao, although their hair was longer and their facial expressions more serious, giving them the air of being older. Their eyes seemed to be illuminated from within, which was somewhat disturbing. What struck me most of all, was *the golden haze, even stronger here than outside, which seemed to concentrate in haloes around their heads…* When one of them raised a hand to indicate Thao and I should each take a seat facing them, I was indeed awestruck, and the word is feeble. I could not have imagined it possible that such radiant beings could exist: *it was as though they were on fire inside and emitting rays from within.*

They sat on block-like seats, fabric-covered, with straight backs. Each seat was of a different color - some only slightly different and others vastly different from their neighbors. Their clothes too, differed in color, suiting each wearer perfectly. All of them sat in what we call on Earth, the ‘lotus position’ that is, the sitting position of Buddha, with hands resting on knees… It was the central figure who addressed me, in a voice so melodious and, at the same time, so authoritarian. I was stunned by it, particularly since he spoke in perfect French. ‘You are welcome among us, Michel. May The Spirit assist and enlighten you.’ The others echoed: ‘May The Spirit enlighten you.’

He began to rise gently above his seat still in the lotus position, and floated towards me. This did not entirely surprise me since Thao had previously demonstrated this technique of levitation. I wanted to rise before this undoubtedly great and highly spiritual personage, as a measure of the infinite respect inspired in me. In trying to move, I found I couldn’t - as though paralyzed in my seat… He stopped just above and in front of me, placing both his hands on my head; the thumbs joined on my forehead above my nose, opposite the pineal gland, and the fingers joined at the top of my head. It was Thao who described these details to me later, for at the time, I was overwhelmed by such a sensation, that the details didn’t register.

During the time his hands were on my head, it seemed that my body no longer existed. A gentle warmth and delicate perfume originated within me, emanating in waves and blending with soft music that was barely audible… Suddenly, I could see amazing colors surrounding the figures opposite me and, as the ‘leader’ returned slowly towards his seat, I could see a multitude of radiant colors around him; ones I had been unable to perceive before. The principal color was a mass of pale pink which enveloped the seven figures, as though in a cloud, and their movements caused that wonderful, glowing pink to encircle us also!

When I had sufficiently recovered my senses to turn towards Thao, she too, was surrounded with wonderful colors, although less brilliant than those around the seven figures… You will notice that, in speaking of these great personages, I instinctively use 'he' rather than 'she'. In explaining this, I can only suggest that the personalities of these special beings were so strong and their bearings so imposing, that I recognized more of the masculine in them than the feminine –I mean no offence to women– my reaction was instinctive... However, women or men, they had transformed me. I knew that the colors surrounding them were their Auras. I was capable of seeing Auras –who knew for how long– and I wondered at what I saw.

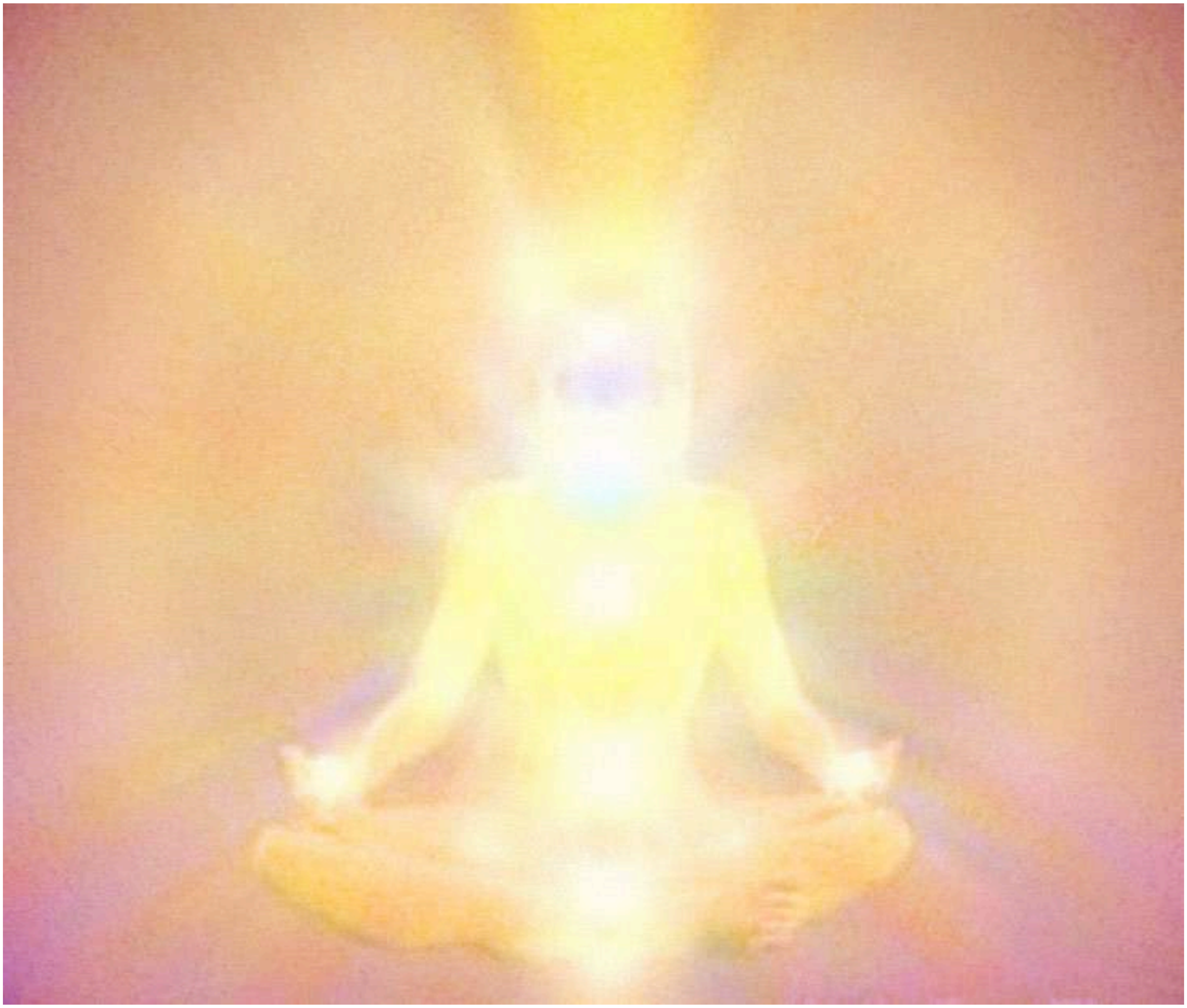

The 'leader' had resumed his seat and all eyes were fixed on me, as if they wanted to see inside me, which indeed, is what they were doing. Silence reigned for a time, which seemed interminable. I watched the varied colors of their Auras vibrate and dance around them, sometimes far in the distance and recognized the 'bouquet of color' Thao had spoken of earlier… The golden haloes, clearly defined, were almost saffron colored. It occurred to me, they could not only see my Aura, but possibly read it as well. I suddenly felt quite naked before this learned assembly. The question that haunted me was: why have they brought me here?

Abruptly, the 'leader' broke the silence. 'As Thao has already explained to you, Michel, you have been chosen by us, to visit our planet, in order to report certain messages and to offer enlightenment on several important issues when you return to Earth. The time has come when certain events must occur. After several thousand years of darkness and savagery on the planet Earth, a so-called 'civilization' appeared and, inevitably, technology was developed –a development, which was accelerated during the last 150 years.

‘It has been 14,500 years since a comparable level of technological advance existed on Earth. This technology, which is nothing compared with true knowledge, is nevertheless, sufficiently advanced to become harmful to the human race on Earth in the very near future… harmful, because it is only material knowledge and not spiritual knowledge. Technology should assist spiritual development, not confine people, more and more, within a materialistic world, as is happening now on your planet… To an even greater extent, your people are obsessed with a single goal - affluence. Their lives are concerned with all that the pursuit of wealth entails; envy, jealousy, hatred of those richer and contempt for those poorer. In other words your technology, which is nothing compared with what existed on Earth more than 14,500 years ago, is dragging your civilization down, and pushing it closer and closer to moral and spiritual catastrophe.’

I noticed that each time this great personage spoke of materialism, his Aura and those of his acolytes, flashed with a dull and ‘dirty’ red, as though momentarily, they were in the middle of burning bushes. ‘We, the people of Thiaoouba, are assigned to assist, guide and sometimes punish the inhabitants of planets under our guardianship.’ Fortunately, Thao had briefed me on Earth’s history during our journey to Thiaoouba. Otherwise, I’d have surely fallen off my seat on hearing such a speech.

‘I think,’ he resumed, ‘that you already know what we mean by ‘harmful to the human race’. Many people on Earth believe atomic arms to be the major danger, but it is not so. The greatest danger concerns ‘materialism’. The people of your planet seek money - to some it’s a means of attaining power; to others it’s a means to acquire drugs (another curse), yet to others, it’s a way of possessing more than their neighbors possess…

‘Now, once and for always, we will teach you and provide you with what you French call ‘the key to the mystery’… As you have heard, in the beginning there was the Spirit alone and he created, by his immense force, all that exists materially. He created the planets, the suns, plants, animals, with one goal in mind: to satisfy his spiritual need. This is quite logical since he is purely spirit. Already I see you are wondering why the need to create material things in order to attain spiritual fulfillment. I offer this, by way of explanation: the creator sought spiritual experiences through a material world… In order to have these experiences, he wanted to embody a tiny part of his Spirit in a physical entity. To do this, he called on the Fourth force - the force which Thao has not yet spoken of and which concerns only spirituality. In this domain, Universal Law also applies.

‘You most certainly know that the pattern of the Universe dictates that nine planets revolve around their sun. It is also the case that these suns revolve around a bigger sun, which is the nucleus for nine such suns, and their nine planets… Certain accidents occur and sometimes a planet will disappear in a solar system, or perhaps enter it, but later in time, the solar system will revert and base its structure again on the number nine.

‘The Fourth force had a very important role to play: it had to bring to fruition all that the Spirit had imagined. It ‘inserted’ thus, an infinitesimal part of the Spirit in the human body. This comprises what you could call the Astral body, which forms one ninth of the essential human being and consists of one ninth of a ‘Higher-self’, which is sometimes called ‘overself'. The Higher-self of man is, in other words, an entity which sends one ninth of itself into a human body, becoming the person’s Astral being. Other physical bodies are inhabited, similarly, by other ninths of the same Higher-self and yet each part remains integral to the central entity… Further, the Higher-self is a ninth part of a superior Higher-self which, in turn, is a ninth part of a more superior Higher-self. The process continues as far back as the source, and allows the enormous filtration of spiritual experience required by the Spirit.

‘You must not think that the Higher-self of the first category is insignificant in comparison with the others. It functions at a lower level, but is nevertheless extremely powerful and important. It is capable of curing illness and even resuscitating the dead. There are many instances of people, declared clinically dead, who are brought back to life in the hands of doctors who had abandoned all hope for them. What generally happens in these cases, is that the person’s Astral body meets with the Higher-self. This portion of the Higher-self has left the physical body during the period of ‘death’. It perceives its physical body below, and the doctors trying to resuscitate it; it can also perceive loved ones who mourn for it. In his present state, the Astral body, the person will feel perfectly well –even blissful. Usually he abandons his physical body, frequently the source of much suffering, to find himself catapulted down a ‘psychic canal’, at the end of which is a wondrous light and beyond, a state of bliss.

‘If before passing through this canal into the blissful light which is his Higher- self he has the least will not to die –not on his own account but for the sake of those who need him, young children for example, he will ask to return. In certain cases it will be permitted… You are in constant communication with your Higher-self by means of your cerebral canal. Acting as a transmitting and receiving post, it conducts special vibrations directly between your Astral body and your Higher-self. Your Higher-self monitors you continually, by day and by night and can intervene to save you from an accident.

‘Someone, for example, who is to catch a plane, finds that the taxi breaks down on the way to the airport; a second taxi called also breaks down - just like that... Could you really believe in such a coincidence? The plane in question crashes thirty minutes later, leaving no survivors. Another person, an old and rheumatic woman and barely able to walk, starts to cross a street. There is a loud horn blast and a screeching of tires, but this person is miraculously able to leap to safety.

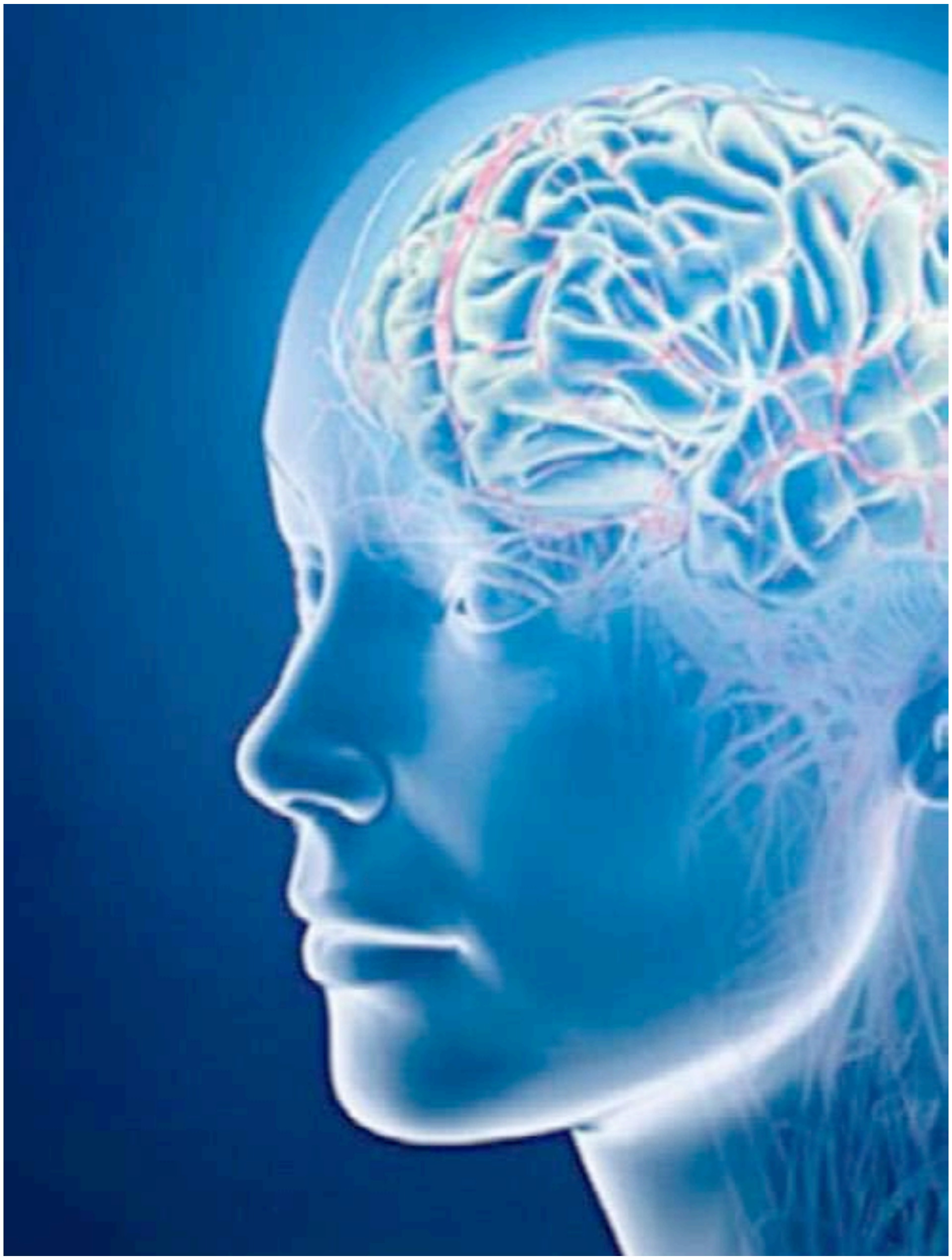

‘How is this explained? It was not yet her time to die and so her Higher-self intervened. In one hundredth of a second the Higher-self triggered a reaction in her adrenaline glands, which, for a few seconds, provided enough strength to her muscles to enable her to execute the leap which saved her life. Adrenaline released into the blood can make it possible to flee from imminent danger, or to defeat the ‘unbeatable’ through anger or fear. In too strong a dose, however, adrenaline becomes a lethal poison.

‘It isn’t only the cerebral canal which is able to conduct messages between the Higher-self and the Astral body. Another channel exists sometimes in dreams –or, I should even say, in sleep. At certain times during sleep, your Higher-self is able to call your Astral body to itself and, either communicate instructions or ideas, or to regenerate it in some way, replenishing its spiritual strength or enlightening it in regard to solutions to important problems. For this reason, it is essential your sleep is undisturbed by intrusive noise or by nightmares resulting from harmful impressions received during the day. Perhaps you will better understand the importance of your old French saying: ‘Night brings counsel’.

‘The physical body in which you exist at the moment is already very complex, but still, it is nothing compared with the complexity of the process of evolvement which occurs with Astral bodies and Higher-selves. In order to allow ordinary people on your planet to understand as easily as possible, I will make my explanation in the simplest terms… Your Astral body, which inhabits every normal human being, transfers to its Higher-self all the sensations that are experienced during a lifetime in a physical body. These sensations pass through the immense ‘filter’ of nine Higher-selves before arriving in the etheric ‘ocean’ that surrounds the Spirit. If these sensations are based essentially on materialism, the Higher-selves have enormous trouble filtering them, just as a water filter clogs quicker if it filters dirty water than if the water was already clear…

‘If, through the numerous experiences you have in your life, you ensure your Astral body benefits in a spiritual sense, it will acquire more and more spiritual understanding. In time, which can vary from 500 to even 15,000 of your Earth years, your Higher-self will have nothing more to filter…

This part of itself, embodied in the Astral being of Michel Desmarquet, will be so spiritually advanced, it will have arrived at the next stage where it will have to contend directly with the more superior Higher-self… We can compare this process with a nine-stage filter, intended to rid the water passing through, of nine elements. At the end of stage one in the process, one will have been completely eliminated, with eight remaining. Of course, to make this information easier to digest, I am making enormous use of imagery...

'This Astral body then, will have completed its cycle with the Higher-self of the first category and will then detach itself from Higher-self number one to rejoin the Higher-self of the second category; the entire process will be repeated. By the same token, the Astral body will be sufficiently spiritually advanced to pass to a planet of the next category, as well…

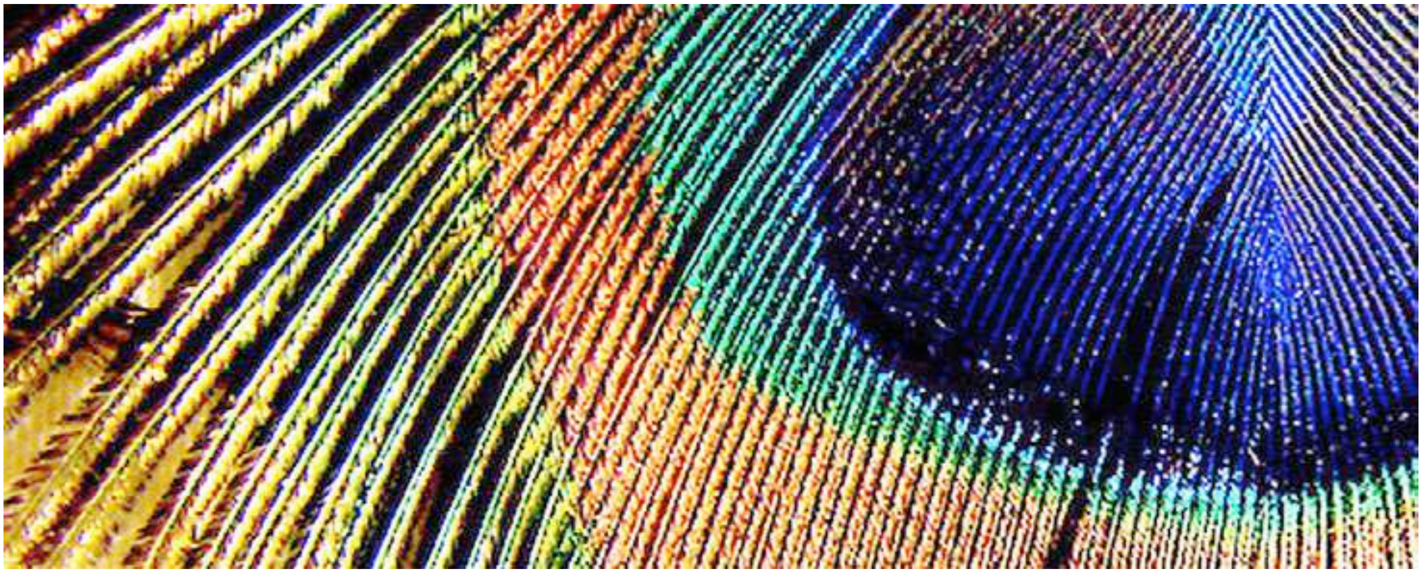

'In his wisdom, The Spirit, by means of the Fourth force, provided for nine categories of planets. At present, you are on the planet Thiaoouba which is in the ninth category; that is, at the top of the scale… Earth is a planet of the first category and therefore at the bottom of the scale. What does this mean? The planet Earth could be likened to a kindergarten with the emphasis on teaching basic social values. A planet of the second category would then correspond with a primary school where further values are taught - in both schools, adult guidance is imperative. The third category would comprise secondary schools where a foundation of values allows exploration beyond. Next, you would go to university, where you are treated as an adult, for you would not only have attained a certain amount of knowledge, but you would also start to accept civic responsibility.

'This is the type of progress occurring with the nine categories of planets. The more spiritually advanced you are, the more you will benefit, on a superior planet, by an environment and general way of life which is superior. The very way in which you procure food is much easier, which in turn, simplifies the process of organizing your way of life; the consequence is more effective spiritual development. On the higher planets, Nature itself, enters the stage to assist 'the pupil' and, by the time you reach planets of the sixth, seventh, eighth and ninth category, not only is your Astral body highly evolved, your physical body has also benefited from your development.

'We know you have already been favorably impressed by what you have seen on our planet. As you see more, you will appreciate that it is what you would call on Earth, 'a paradise'; and yet, compared with true happiness, when you become a pure spirit, it is still nothing… From this planet, it is possible either to remain in a physical body or be reunited with the Great Spirit in the ether.' As these words were uttered, the Aura surrounding the leader shone more brightly than ever and I was surprised to see him almost disappear in a golden mist, only to reappear a second later.

'You have understood that the Astral body is a body which inhabits your physical body recalling and noting all the understanding acquired during the course of its various lives… It can only be enriched spiritually - not materially. The physical body is merely a vehicle which, in most cases we abandon when we die… I will elaborate, for I see that 'in most cases' has confused you. By this, I mean that some of us, including all on our planet, are able to regenerate the cells of our bodies at will. Yes, you have already noticed that most of us seem to be of the same age. We are one of three planets that are the most highly evolved in this galaxy. Some of us can, and do, directly join what we call the great ether.

‘So, on this particular planet, we have arrived at a stage near perfection, both materially and spiritually. But we have our roles to perform, as does every creature existing in the universe; in fact, everything, including a single pebble, has its role… Our role, as beings of a superior planet, is to guide – to help with spiritual development and even, sometimes, materially. We are in a position to give material assistance because we are technologically the most advanced people. Indeed, how could a father give spiritual guidance to his child if he were not older, more educated and more skilled in diplomacy than the child? If the child should require physical punishment, as unfortunately, is sometimes the case, isn’t it important that the parent be physically stronger than the child? Certain adults, who refuse to listen and who are absolutely stubborn, also need to be corrected by physical means.‘

‘You, Michel, come from the planet Earth, which is sometimes called ‘The Planet of Sorrows’. Indeed, the name is appropriate, but it is this way for a precise reason - it is intended to provide a learning environment of a quite specific kind. It is not because life there is so difficult that you have to intervene - you cannot lightly go against Nature, destroying rather than conserving what the Creator has put at your disposal; that is, interfering with ecological systems, which have been intricately designed. Certain countries, like Australia, where you are from, are beginning to show great respect for ecology and it is a step in the right direction; but even in that country, what case is made of pollution - both water and air pollution? What is ever done about one of the worst forms of pollution? *Noise.* [Damage to sensitive auditory cells in the cochlea, seen below, is but one of many dangers.]

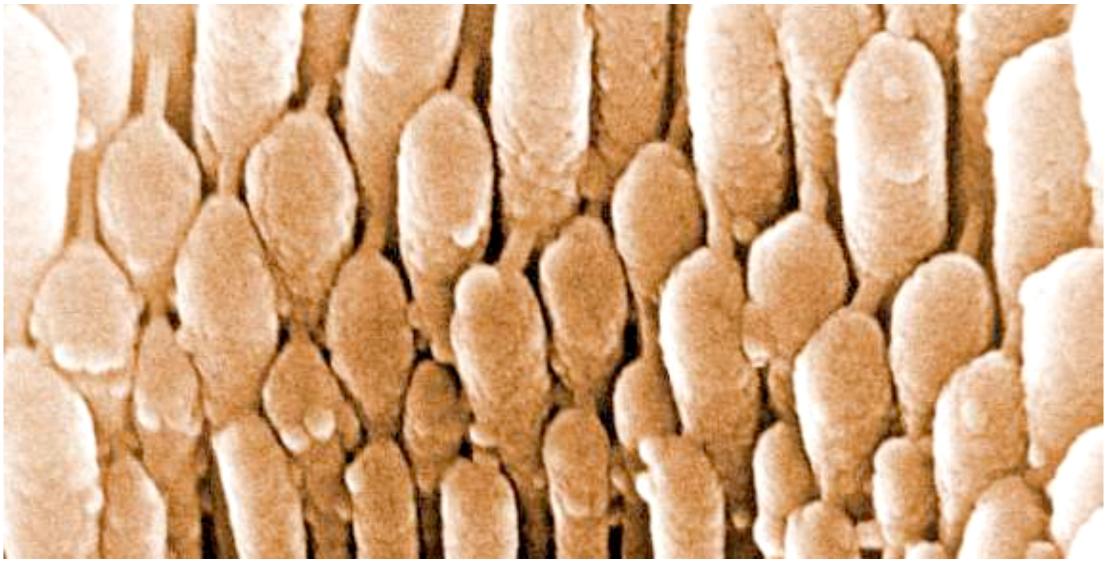

‘I say ‘worst’ because people such as Australians pay virtually no attention to it at all… Ask someone if traffic noise bothers him and the response will surprise you - eighty five percent of the time it will be; ‘What noise? What are you talking about? Oh that noise - we get used to it.’ And it is precisely because they ‘get used to it’ that the danger exists.’

Just then, Thaora, as this high figure was called, made a gesture and I turned around. He was replying to a question I had mentally posed; ‘How can he speak of percentages and know so much about our planet with so much precision?’ Turning around, I almost uttered a cry of surprise for, behind me, stood Biastra and Latoli. In itself, this was nothing surprising, but the friends I knew who measured 310 and 280 centimeters in height respectively, were now reduced in size to correspond with my height. My mouth must have continued to gape, for Thaora smiled.

‘Can you understand, that sometimes, and very often in these days, some of us live among your people on Earth? - and there is my answer to your question. To continue on the very important subject of noise, it is such a danger that, if nothing is done, catastrophe is certain… Let us take the example of a discotheque. The people who expose themselves to music that is typically played three times too loudly, are subjecting their brains and their physiological and astral bodies to vibrations which are so harmful. If they could see the damage that is caused, they would vacate the discotheque quicker than if there was a fire.

‘But vibrations don’t only come from noise; they also come from colors and it is astonishing that, on your planet, experiments conducted in this field have not been followed up.[16] Our ‘agents’ reported a particular experiment involving a man who was capable of lifting a certain weight. It was found that after staring for a moment at a pink-colored screen, he consistently lost thirty percent of his strength… Your civilization pays no attention to such experimentation. In fact, colors can enormously influence the behavior of human beings and yet, control of this influence requires that an individual’s Aura be taken into account. If for example, you want to paint or wallpaper your bedroom with the colors that are truly appropriate for you, you must be aware of the colors of certain principal points of your Aura.

‘By matching the colors of your walls with those of your Aura, you can improve your health or maintain good health. Further, the vibrations emanating from these colors are essential for good mental balance, exerting their influence even while you sleep.’ I was wondering how we could be expected to know these significant colors in our Auras when, on Earth, we were not capable of perceiving Auras… Thaora replied immediately, without me having to say a word aloud.

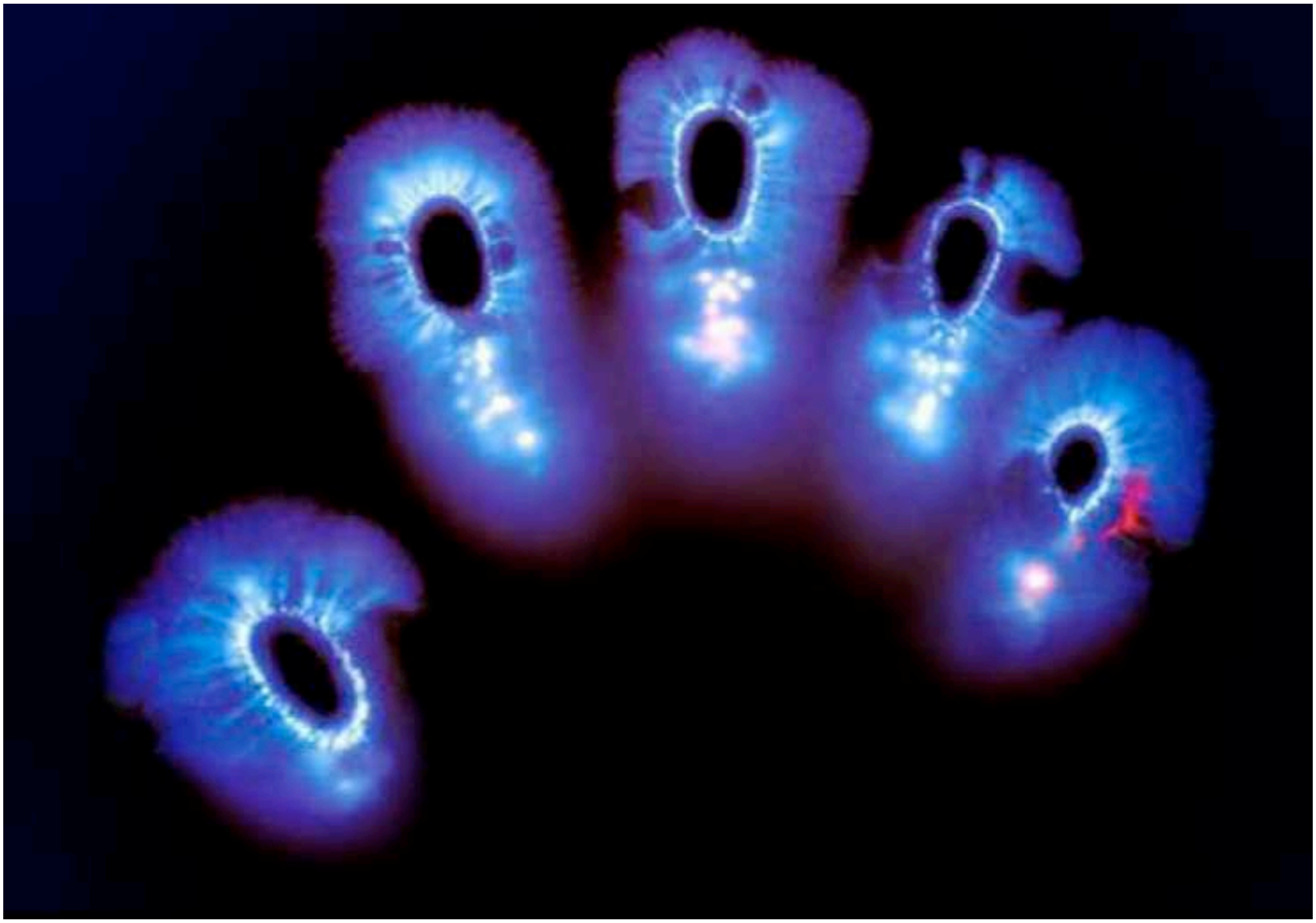

‘Michel, it is now very important that your experts invent the special equipment necessary, to enable perception of the Aura, as this, in turn, will ensure that correct choices are made at the critical crossroads ahead… The Russians have already photographed the Aura [as Kirlian photography – seen above, opposite]. This is the beginning, but the results obtained allow them only to read the first two letters of the alphabet, as it were, in comparison with what we are able to decipher. The reading of the Aura in order to heal the physical body is nothing compared with what such a reading can achieve for the psychic body, or the physiological body. It is in the area of the psyche that, on Earth, your greatest problems exist.

‘At the moment, most responsibility is taken for the physical body, but this is a serious mistake. If your psyche is poor, it will influence your physical appearance accordingly, but, regardless, your physical body will wear out and die one day, whereas your psyche, being part of your Astral body, never dies. On the contrary, the more you cultivate your mind, the less you will be burdened by your physical body and the quicker you will proceed through your cycle of lives… We could have brought you to our planet in Astral body, but instead, we have brought you here in physical body - and for an important reason. I see you already understand our reason. This pleases us and we thank you for your willingness to assist us in our task.’

The Thaora stopped talking and seemed to lapse into thought, at the same time, fixing me with his luminous eyes. I cannot say how much time elapsed. I know my state became more and more euphoric and I was aware that the Auras of the seven personages were gradually changing. The colors became more vivid in places, softer in others, while the outer edges became misty. This mist became more golden and pink as it spread, gradually blurring the seven figures.[17]

The profound wealth of imagery, wisdom and spiritual knowledge transmitted by the Thaori during the fourth-density plasma experiences of Michel Desmarquet on the golden planet of Thiaoouba resonate with the highest teachings available to Earthly humanity. The importance of Aura photography, known as Kirlian photography among Russian biophysicists and referred to as ‘electro-photonics’ by Thao, is decisive in not only the deepening of our understanding of the effects of color on human perception, but as well it represents the beginning of high-energy applications of HHO plasma to the human body. While clearly visible gaps in electrical conductivity observed along the fingertips has already shown promise in reflecting the state of health of specific organs (quite useful in diagnostic readings), the more advanced beam healing systems now biocompatible and effective in rapid cellular rejuvenation.

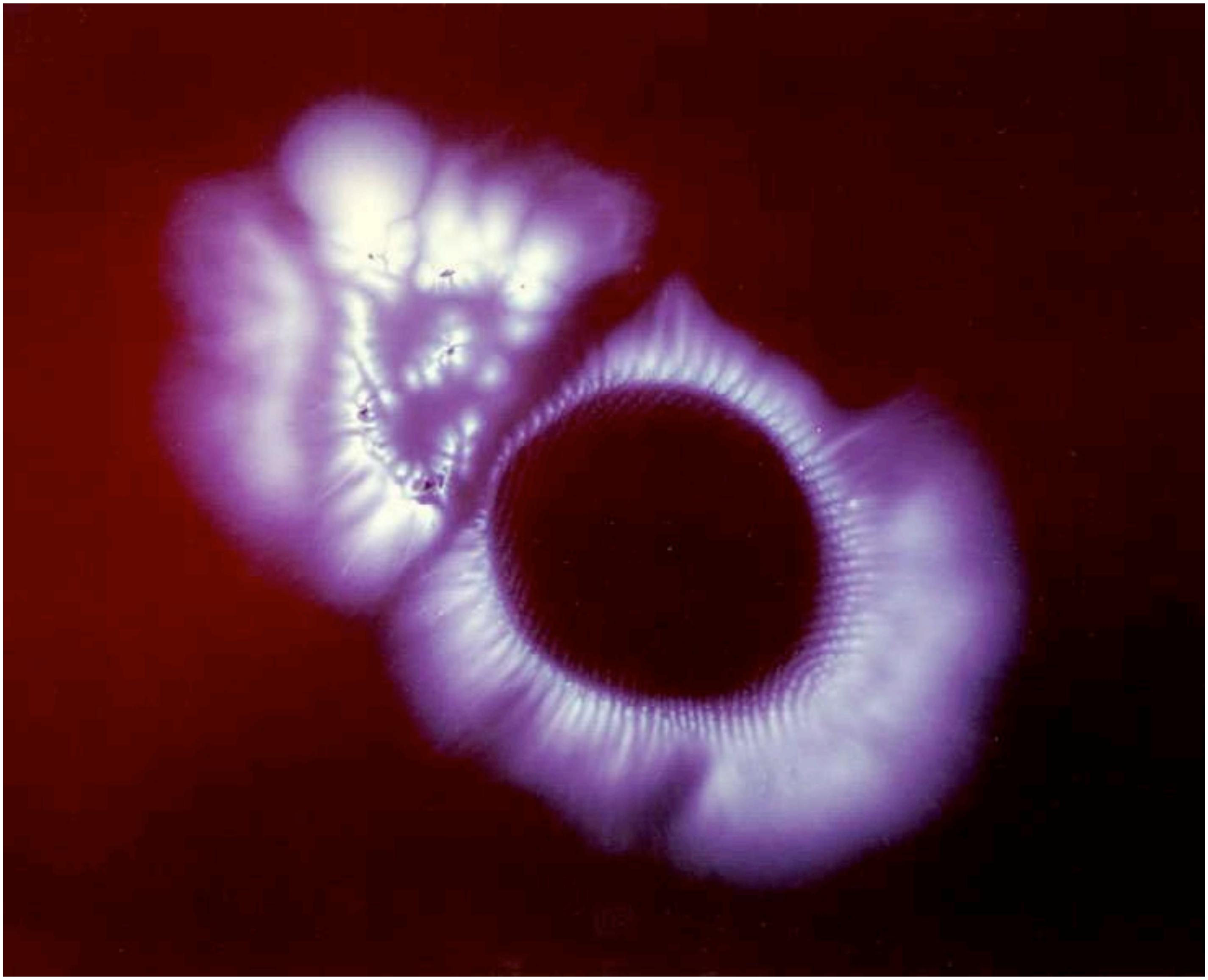

While the field of dentistry now applies nanosecond electrical pulses to ionize the HHO plasma at low temperatures to form biocompatible helium/oxygen plasmas for oral cleaning and sterilization,[18] an unnoticed effect of cellular regeneration is also occurring. The application of HHO plasma to the cleansing of the human body promises a significant reduction in the rate of cellular aging when the proper sources of lightwater (protium) are ingested. The bioelectric effect of electroporation, whereby cellular permeability is significantly lowered by rapid, milliampere electrical pulses, and can be only be generated throughout all of the body’s cells using electroluminescent gold and silver nanoparticles. Natural safeguards for immersion in the fourth-density HHO plasma environment provide for a limit to 9 full days of exposure, as succinctly explained to Michel by another temporary visitor to the golden Thiaoouba. The intense luminosity of the atmospheric HHO plasma inevitably intoxicates low-resonance human bodies:

> ‘An Astral body absolutely cannot change to a superior planet until sufficiently evolved [in spirit]… This evolution occurs thanks to the physical body that improves as it progresses right up as far as this planet. You have already learned of the nine categories of planets –ours are at the bottom of the scale. We, in our present physical bodies, can be permitted only nine days’ stay here. According to Universal Law, on the tenth day, our physical bodies would die and neither Thao nor the great Thaora, within whose power it is to revive the dead, would be able to prevent or reverse the process. Nature has very inflexible rules with well-established safeguards… But it is possible that when you do die on Earth –when your moment has come– your Astral body will reincarnate in a body on… a second or perhaps third category planet… [Through reincarnation] we must be ‘reprocessed’ over and over until we emerge perfect, for eventually we will rejoin The Spirit who, being perfect himself, cannot accept the slightest imperfection.’

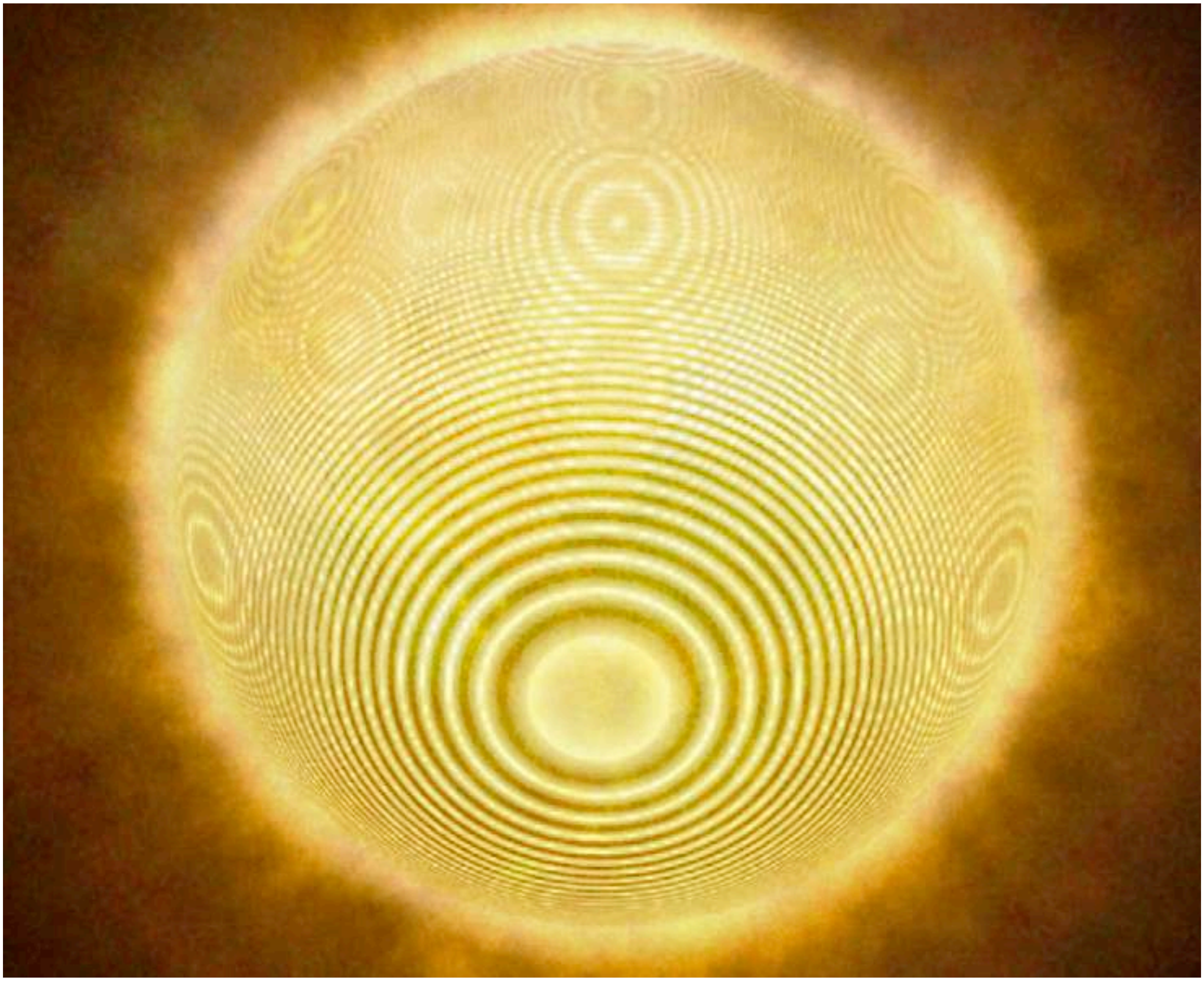

> Around each planet, since its creation, is a kind of *psychosphere* or vibratory cocoon, which turns at a speed seven times that of light [or 1,302,000 miles per second]. This cocoon acts as a blotter, as it were, absorbing (and remembering) absolutely every event occurring on the planet… The human being, comprising an integral part of the universe can, because of his Astral body and if he is correctly trained, draw what knowledge he seeks from within the psychosphere. Of course, great training is required for this.[19]

This description of the psychosphere corresponds closely with the ancient Sanskrit concept of the *akashic field*, the mnemonic body of the planet that spirals around the Earth through the pathways of standing infrasound waves (illustrated above). The (Van Allen) plasma belt that encompasses Earth possesses superfluid crystalline properties that may act as a form of cosmic memory storage.

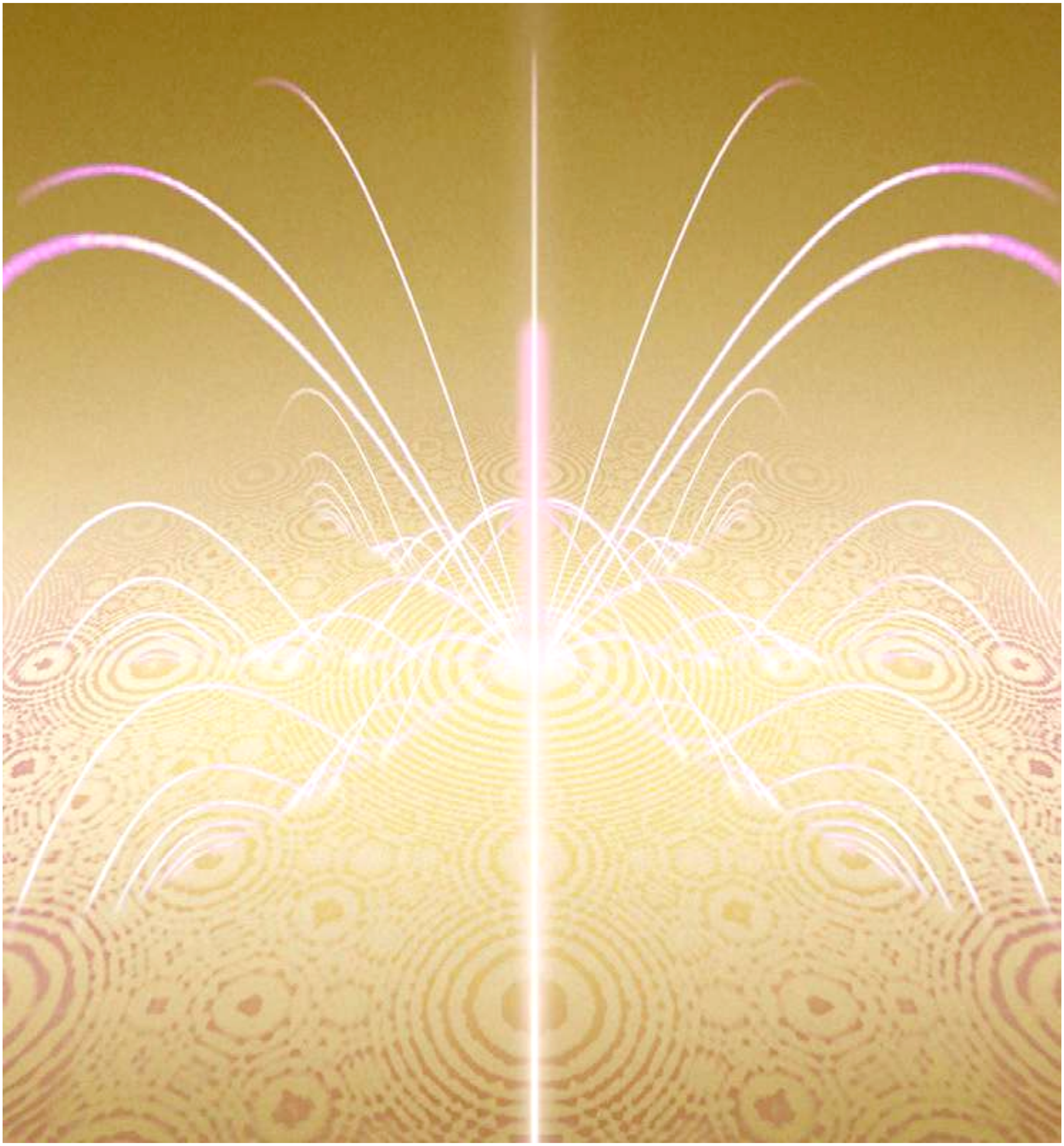

The subconscious mind is in constant contact with the planetary psychosphere where all the information of all past and future lives is stored, as elucidated by the Sleeping Prophet Edgar Cayce, in so many astoundingly advanced spiritual readings from the era of the industrial revolution. While Cayce had a unique ability to access the psychosphere directly during trance, Thao explained that most inexplicable moments of déjà vu are attributable to *akashic future life previews* shown by the Higher-self:

> [A]fter having seen his life… flash before him, and, after accepting to live that life, all details of it were erased from his memory. He passed through what certain Nagas have called 'The River of Oblivion' - this happens whether one accepts or rejects a possible reincarnation. Of course, there is a reason for it.

'If, for example, you remembered that, around forty years of age, you would lose your wife and two cherished children in a car accident and that you, yourself, would be confined to a wheelchair, the knowledge could tempt you to take your life rather than face up to your troubles, or it might lead you to behave badly in other spheres. So, the 'film' is erased, in something like the way you 'wipe out' a tape recording.

'Occasionally, by accident, the machine does not erase everything and you can hear brief portions of what should have been erased. Of course, my analogies are fanciful when I speak of 'films' and 'tape recordings' but I hope they give you an idea of what I am trying to explain. In reality, the process involves electro-photonics, which mean nothing, yet, to people on Earth. This, in fact, occurs often in the 'films' that the Higher-self shows to an Astral body, which is why most people say, on several occasions in the course of their lives 'I have seen that before' or 'I have heard that before' and they know what the very next action or word will be. In English, people call this feeling 'déjà vu'.' [20]

'Michel… you were chosen by us to visit Thiaoouba, but the essential motive for our choice has not been revealed. It is not only because you have a mind already awakened and open, but also –and principally– because you are one of the rare soukous inhabiting Earth at the present time. A 'soukou' is an Astral body that has lived eighty-one lives in human physical bodies, and has lived those lives on different planets or different categories. For various reasons, the 'soukous' return to live on inferior planets, like Earth, when they could just as well continue to 'climb the ladder' without ever going backwards. You know that the number nine is the number of the Universe. You are here in the City of Nine Dokos, founded on Universal Law. Your Astral body has nine times nine lives, which brings you to the end of one of the great cycles.'

Once again, I was completely flabbergasted. I suspected I wasn't living my first life… –but eighty-one lives! I didn't know one lived so many… 'It's possible to live many more, Michel,' said the Thaora, interrupting my thoughts. 'Thao is up to her 216th, but other entities live far fewer. As I said, you have been chosen from among very few 'soukous' living on Earth, but, in order that you acquire a thorough understanding during the trip to our planet, we have planned another journey in time for you. So that you will better understand what reincarnation is, and what its purpose is, we will permit you to revisit your previous existences. This journey in time will be useful to you when writing your book as you will fully comprehend its purpose.' [21]

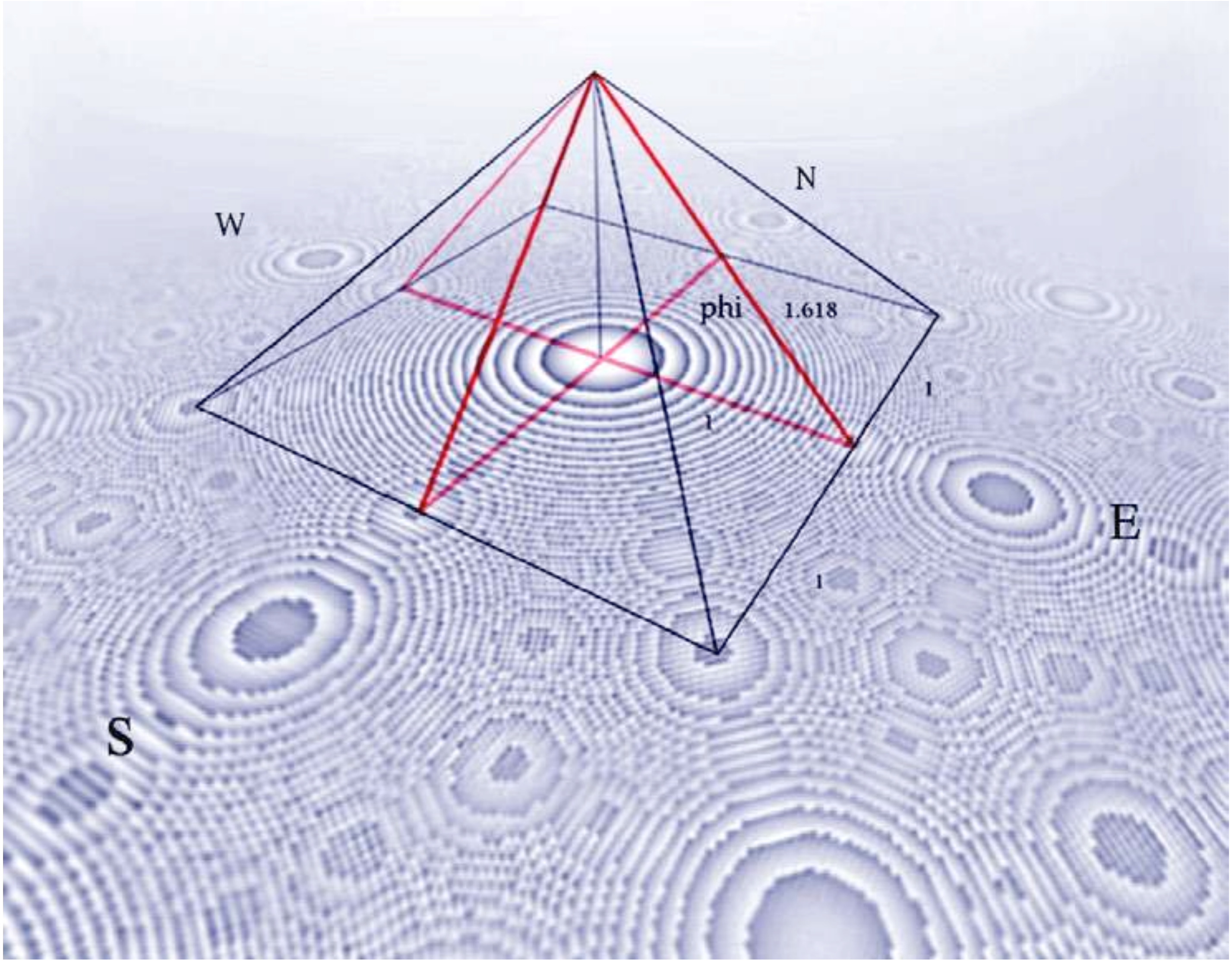

Absolutely all of the events of past lives are stored within the psychosphere and remain accessible to highly attuned adepts of meditation. The ancient pyramid technologies were developed to enhance the psychoacoustic conditions that engender *synesthesia*, a form of holographic consciousness. *The global network of ancient pyramids function as one synchronous unit, balancing the biorhythms of initiates into telepathic communication with other worlds*:

> 'I must finish off my story concerning the continent of Mu… Savanasa then, was situated on a vast plateau overlooking plains which rose, on average, no more than 30 meters above sea level. On this plateau and in the centre, an enormous pyramid was constructed. Each stone used in its construction, some weighing more than 50 tons, was cut precisely to within one fifth of a millimeter, using what we can call 'ultrasonic vibratory systems'. [These ultrasonic pulse cutting techniques were developed by scientists in 2009.[22]] This was done in the quarries of Holaton, now found on Easter Island, which was the one place on the entire continent where this special rock could be found. There was, however, another quarry at Notora, southwest of the continent.
>
> 'The enormous stones were transported using anti-gravitational techniques well known to these people. (They were carried on platforms, 20 centimeters above the paved roads, and were constructed using the same principals as those of the pyramids.) Roads such as these were built all over the country, converging, like a massive spider's web, on the capital, Savanasa.
>
> 'The huge stones were taken to Savanasa and put into position according to directions from the 'master' or chief architect on the project. When finished, the pyramid measured exactly 440.01 meters in height and its four faces were oriented precisely towards the four points of the compass [as seen above].'
>
> 'Was this intended to be the King's palace, or his tomb?' Everyone wore the same indulgent smile that often appeared when I asked a question.

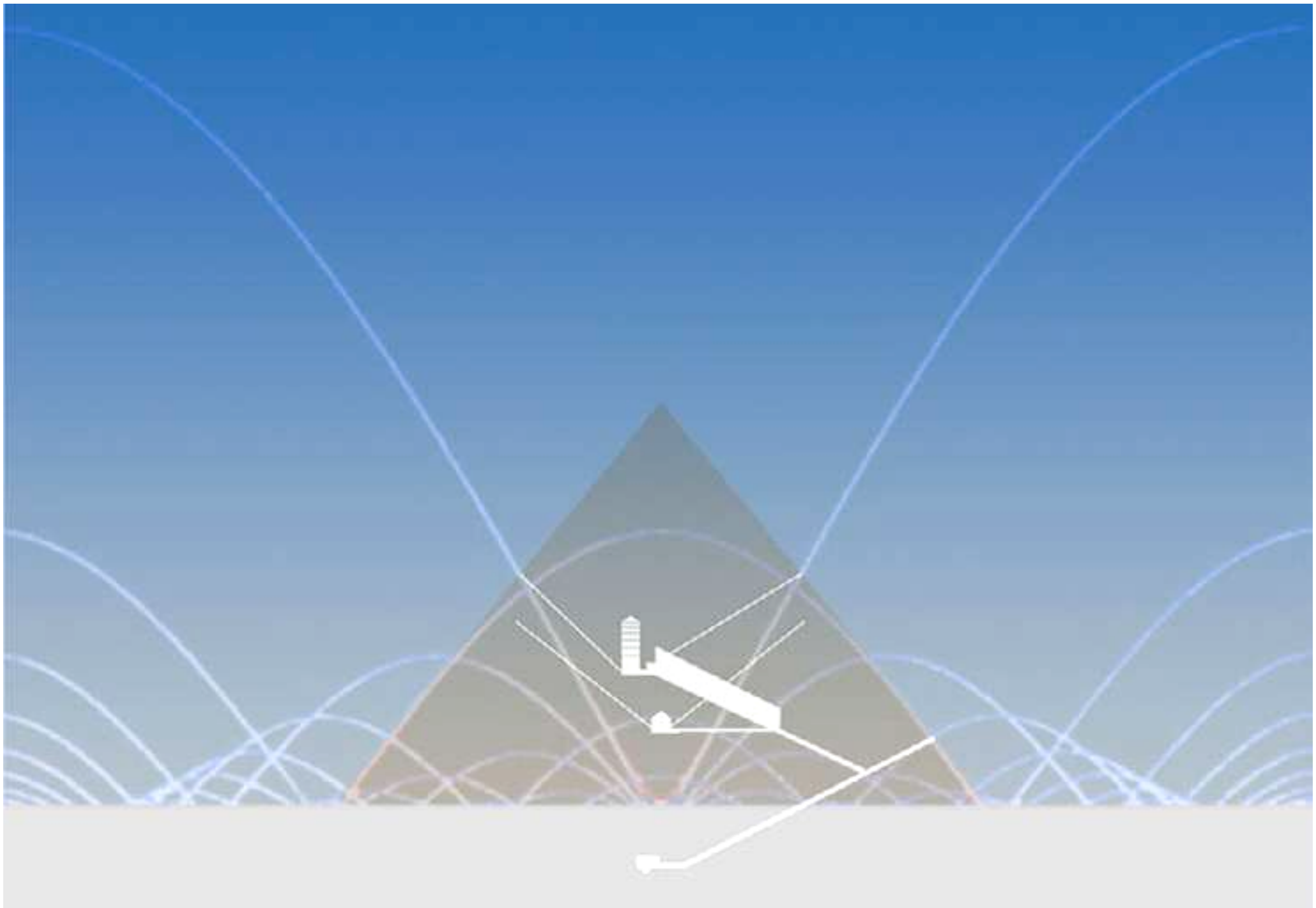

> ‘Nothing of the sort, Michel. This pyramid was much more important - it was a tool. An enormous tool, I admit, but a tool just the same. So, too, was the Pyramid of Cheops, in Egypt, though it was much smaller… You will have realized… these were highly advanced people. They possessed a profound understanding of Universal Law and used their pyramid as a ‘captor’ of cosmic rays, forces and energies, as well as terrestrial energies…
>
> Inside, rooms positioned according to a precise plan, served the King and certain other great initiates, as powerful communications centers, enabling (telepathic) communication with other planets and other worlds in the universe. Such communication with extraterrestrials is no longer possible for people on Earth; but the people of Mu in those days, by natural means and by exploiting cosmic forces, were in constant communication with other beings and were even able to explore parallel universes…
>
> [The pyramid’s] second use was to make rain. By a system of plates, made of a special alloy incorporating silver as its major component, these people were able, in a few days, to cause the accumulation of clouds above the country, and so, to have rain… Thus, they were able to create, virtually, a paradise over the whole continent. Rivers and springs never dried up, but flowed lazily across the numerous plains of a land that was essentially flat. Fruit trees were laden with fruit, bowing under the weight of oranges, mandarins or apples, according to latitude. Exotic fruits, of kinds that… no longer exist on Earth, were harvested in abundance.’ [23]

In describing interior “rooms positioned according to a precise plan,” Thao is making reference to the nonlinear structure of Fibonacci standing waves by which the precise positions of the pyramids’ chambers were determined, for the purpose of focusing terrestrial infrasound resonance (above). Telepathic communication was established by biorhythmic entrainment at 1.45 Hz heartbeat, also known as the *tri-thalamic entrainment frequency*, by which the biorhythms of initiates using the pyramid chambers are synchronized with those initiates in other pyramids across the world, or indeed with other planetary civilizations. Infrasound stimulation also provided for the secondary purpose of rainmaking, *by drastically increasing the evaporation and precipitation rates*, just as occurs during solar flare activity. Biorhythmic balancing in the pyramids also greatly benefits the Astral body, composed of electrons:

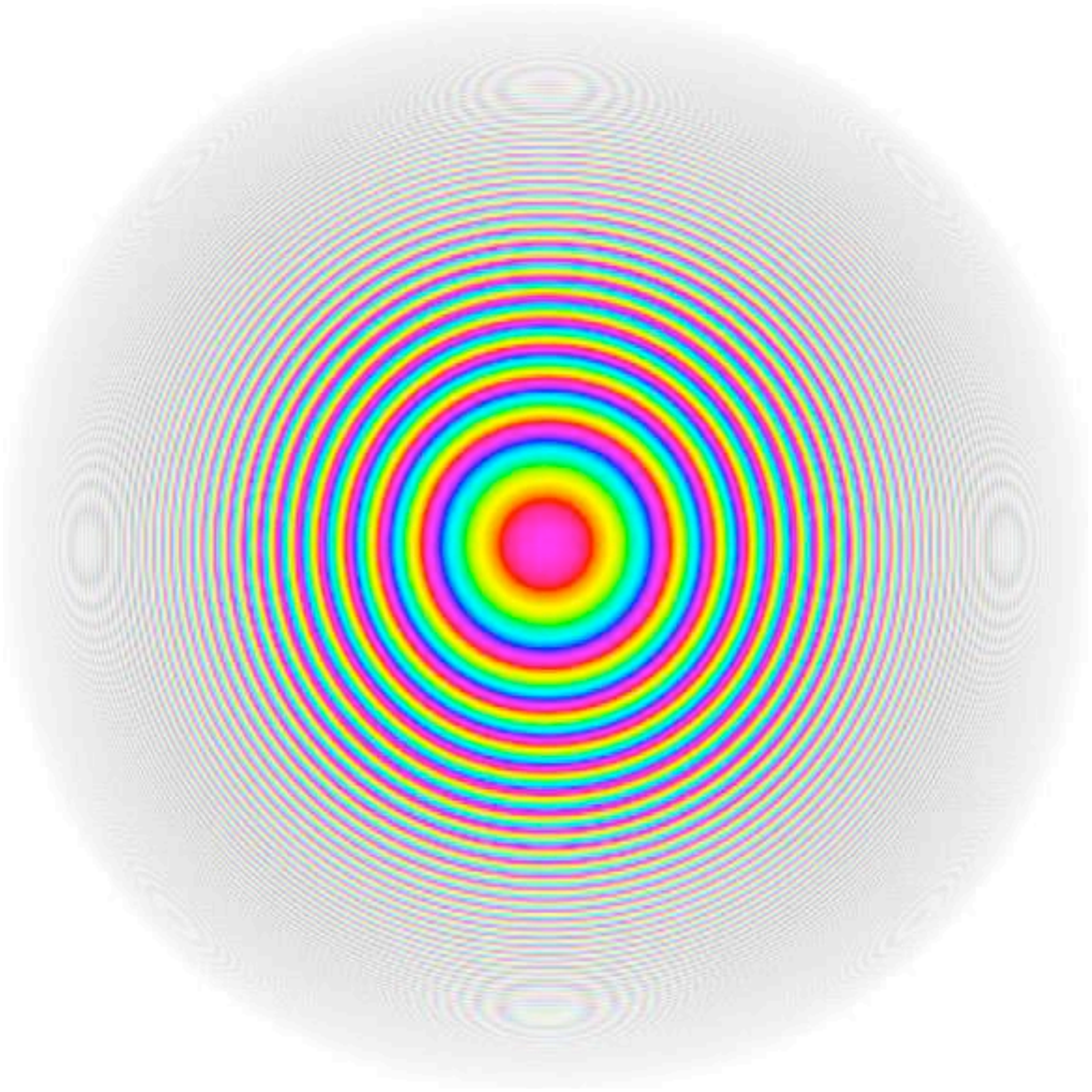

‘Now we come to another kind of pollution that is of no small concern, though people might readily dismiss it. As the great Thaora told you, noise is one of the most noxious pollutants for it upsets your electrons and unbalances your physical compartment. I haven’t yet mentioned these electrons to you and I see that you are not following me very well… A normal human Astral body contains approximately four billion, trillion electrons [4.0×10^{21} = 4 000 000 000 000 000 000 000 electrons]. These electrons have a life span of approximately ten billion, trillion of your years [10^{22} = 10 000 000 000 000 000 000 000 years]. They were created at the moment of creation. Your Astral body contains them and, when you die, nineteen per cent rejoin the electrons of the universe until required by Nature to form a new body or a new tree or animal, and the eighty one per cent rejoin your Higher-self.’ ‘I don’t quite follow you,’ I interrupted.

‘I know, but I intend helping you to understand. An Astral body is not quite what you would call a pure spirit. On Earth, there is a belief that the spirit is made of nothing. This is false. The Astral body is composed of billions of electrons, exactly marrying your physical shape. Each of these electrons has a ‘memory’ and each is capable of retaining as much information as is contained in all the books that fill the shelves of an average town library.

‘I see you are staring wide-eyed at me, but it is as I say. This information is coded, like a microfilm containing all the plans of an industrial installation that a spy would be able to pass in a cufflink, though much more miniaturized than that. Certain physicists on Earth are now aware of this fact but the public, at large, hasn’t been informed of it. Your Astral body transmits and receives messages, by means of these electrons, through the channel of your brain, to, and from your Higher-self. Information is being transmitted without you being aware of it, thanks to a weak electric current from your brain in harmony with your electrons. ‘Since it is the Higher-self which sent this Astral body into your physical body, it is in the natural order of things that your Higher-self should receive information from your Astral body.

> ‘Like all things electronic, the Astral body – tool of the Higher-self – is quite a delicate tool. During your waking hours it is capable of sending messages of extreme urgency to the Higher-self but the Higher-self seeks much more than that. So, during sleep, your Astral body leaves your physical body to rejoin the Higher-self, either passing on required information or receiving information or orders… Now, those people who, through very advanced and special exercises, are able to detach their astral bodies from their physical bodies, will be able to see a light, silvery-blue thread, such as you saw yourself, linking their physical and astral bodies. Their Astral bodies, likewise, are visible during the time that the separation lasts. It is these same electrons which form your Astral body and which create the visible effect of the thread.

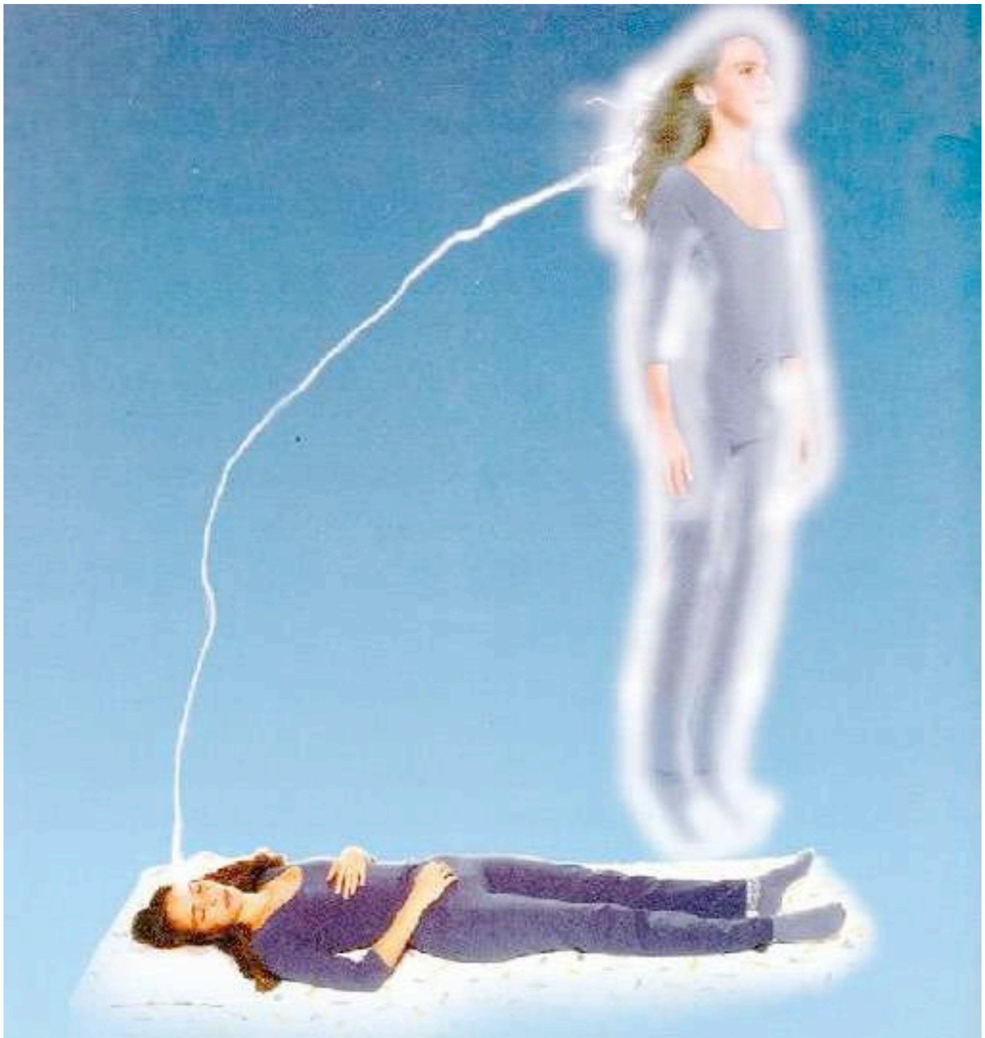

> ‘I see that you follow what I am saying and that you have grasped my point. Let me finish by explaining the dangers of noise. Noise directly attacks the electrons of your Astral body creating parasites, to use a radio and television term. If… someone is operating an electric tool next door to your house, such large parasites will be produced on your screen that the image will distort completely.

> ‘The same thing occurs with the Astral body, but unfortunately you won’t be aware of it in the same way that you are with a television screen; and, it’s much worse, since noise damages your electrons. And yet people say: ‘Oh, we get used to it.’ Your brain ‘tenses’, so to speak and your psyche initiates self-defense mechanisms, but not so the Astral body; a parasite invades its electrons - which, of course, has disastrous repercussions for your Higher-self.

> ‘The sounds that reach your ears are clearly very important. A particular piece of music can elevate you to a state of euphoria, while another piece, although very pretty, will have no effect on you or, perhaps, irritate you. Try an experiment: take a piece of soft violin, piano or flute music that you like and play it as loudly as you can. The suffering of your eardrums will not be as great as the discomfort you’ll feel within. Most of your fellow human beings on Earth consider noise pollution to be of negligible concern, but the noise of the exhaust pipe of a motor bike is three to four times worse than the noxious fumes that it discharges. While the fumes affect your throat and your lungs, the noise affects your Astral body… However, no one has ever been able to take a photograph of your Astral body and, so, people don’t concern themselves with it!’ [24]

Thao’s lengthy description of the Astral or *electronic body* brings sharp clarity to the issue of noise pollution in the industrialized world. The remarkable descriptive metaphor of electronic parasites was powerfully used to illustrate the invisible effect of the dangerous distortion caused by the incessant bombardment of disruptive oscillations from modern airplane, truck and automobile traffic. Not mentioned at the time of Desmarquet’s contact in 1987, which occurred before the mass-production of wireless devices, the lethal effects of long-term exposure of DNA to microwave radiation has since become another invisible threat to all life, mirroring the atmospheric devastation caused by combustion engines:

‘For approximately one hundred and forty years on your planet, man has been accelerating the destruction of Nature and the pollution of the environment. This has happened since the discovery of steam power and the combustion engine. You have but a few years left in which to arrest the pollution before the situation becomes irreversible. One of the principal pollutants on Earth is the petrol-driven engine and this could be replaced immediately with a hydrogen engine that would cause no pollution, so to speak. On certain planets, this is called the ‘clean motor’. Prototypes for such an engine have been constructed by various engineers on your planet [including John Keely (above) and Nikola Tesla among others,] but they must be industrially manufactured in order to replace petrol engines. Not only would this measure mean a seventy percent reduction in current levels of pollution by combustion waste, but it would also be more economical for consumers.

‘The big petrol corporations had been terrified at the idea of this motor being popularized for it would mean loss of sales for their oil and subsequent financial ruin. Governments, too, who impose enormous taxes on these oils, would suffer equally. You see, Michel, it always comes back to money. Because of it, you have a whole economic and financial context that opposes progress towards radical change in the interest of all human life on Earth. The people on Earth allow themselves to be pushed around, bullied, exploited and led to the abattoirs by political and financial cartels which are sometimes even associated with well-known sects and religions. When these cartels fail to win the people with clever advertising campaigns intended to brainwash them, they try to succeed through political channels, and next through religion or through a clever blend of the lot.

‘Great men wanting to do something for mankind have simply been done away with. Martin Luther King is one example; Ghandi is another. But the people of Earth can no longer allow themselves to be treated as fools and led to the abattoirs like flocks of sheep by leaders that they, themselves, have democratically elected. The people form the vast majority. In a nation of one hundred million inhabitants, it is absurd that a group of financiers comprising perhaps a thousand individuals can decide the fate of the others - like the butcher does at the abattoir.

‘Such a group has well and truly stifled the business of the hydrogen motor so that it is no longer mentioned… These people couldn’t care less what might happen to your planet in years to come. Selfishly, they seek their gains, expecting to be dead before ‘whatever is going to happen’ happens. If the Earth disappears, as a result of horrific cataclysms, they assume they will already be dead.

> ‘There, they are making a big mistake, for the source of the coming disasters is the pollution which is growing daily on your planet, and its consequences will be felt very soon - much sooner than you can imagine. The people of Earth must not do as the child forbidden to play with fire; the child is without experience and, in spite of the prohibition, he disobeys and burns himself. Once burnt, he ‘knows’ that the adults were right. He won’t play with fire again but he will pay for his disobedience by suffering… Unfortunately, in the case that concerns us, the consequences are much more serious than the burn of a child. It’s the destruction of your entire planet that is at risk - with no second chance if you don’t place your trust in those who want to help you.

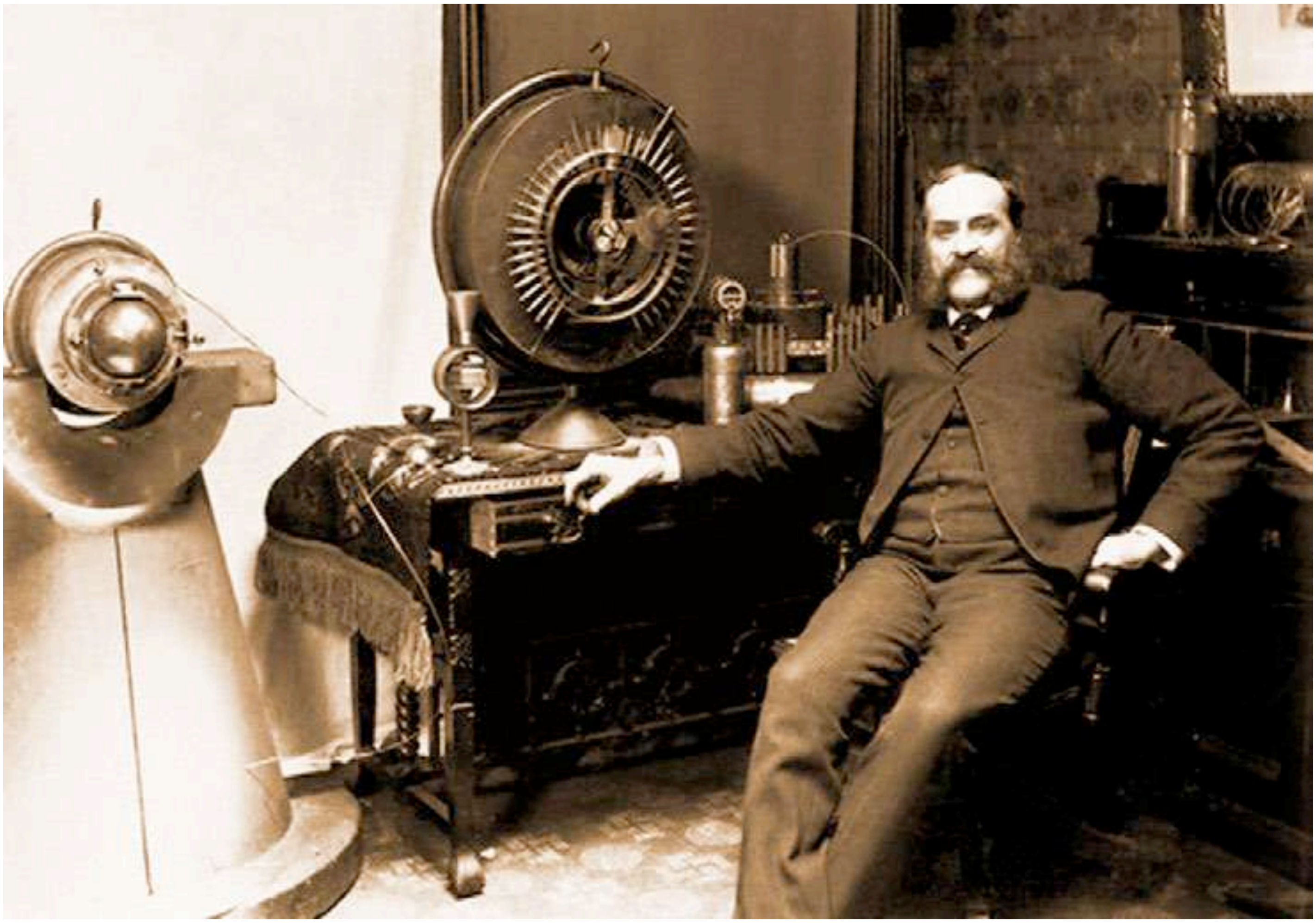

> ‘It interests us to see that recently established ecological movements are gaining in momentum and power; and that the young people of Earth are ‘carrying’ other sensible people along with them in their fight against pollution… There is only one solution… the grouping of individuals. A group is only as powerful as it is large. Those whom you call the conservationists are becoming stronger and stronger and will continue to do so. But it is vital that people forget their hated, their resentment, and especially their political and racial differences. This group must be internationally united – and don’t tell me that is so difficult - for already on Earth there exists a non-violent and very large international organization – the International Red Cross, which has been functioning effectively for quite some time. It is essential that this conservationist group include in its programs not only the conservation of the environment from direct damage but from indirect damage as well, such as that which results from smoke: exhaust fumes from vehicles, smoke from factories, and so on.’ [25]

The immediate and dire consequences of the continued acoustic and atmospheric pollution of Earthy humanity could not have been more clearly stated. The hydrogen engine that Thao speaks of was discovered a century ago by Earthly scientists – most notably by John Worrell Keely (above) and Nikola Tesla. As in the mentioned cases of Martin Luther King and Ghandi, Tesla was also assassinated[26] and his technologies stolen to be secretly developed by Nazi scientists.[27] Recent rediscoveries by Dr. Randall Mills define hydrino resonant-transfer plasma, formed by the ionization of HHO gas using metal particles (opposite).[28] *These facts concerning the clean HHO plasma motor must become widely known and used.*

> ‘If certain of the leaders who read your book don ‘t believe you, or doubt what is written, challenge them to explain the disappearance of billions of ‘needles’ put into orbit around Earth several years ago. Ask them also to explain the second disappearance of billions more ‘needles’ again put into orbit. They will know what you are referring to, never fear. We are responsible for the disappearance of these ‘needles’, judging them to be potentially disastrous for your planet… We do, at times, prevent your experts from ‘playing with matches’ but it is important that our assistance not be relied upon when mistakes are made. If we judge it appropriate to ‘lend a hand’, we do so, but we can’t, and we don’t wish to save you from disaster automatically - that would be to contravene Universal Law.’ [29]

The disappearance of millions of needles in orbit described by Thao as direct evidence of the reality of the extraterrestrial prevention of disasters caused by human scientific ignorance is a powerful reminder of the fragility of our planet and the carelessness of Earthly scientists. Dangerous experimentation of this kind has subsequently led to high atmospheric use of radioactive aluminum and barium nanoparticles dispersed as ‘chemtrails’ for atmospheric and psychological manipulation. Photo-ionization of the particles generates HHO plasma by the high-energy splitting of water molecules in the near vacuum of space, which acts as an amplifier of land-based signals beamed from the HAARP array, creating a bounce-back effect for ultra-low frequency wave focusing used in the artificial generation of earthquakes.

While extraterrestrial overseers have not brought a halt to the secret development and use of dangerous infrasonic technologies by Earth’s psychopathic profiteers, rare interventions do occur when catastrophic situations arise. In proving the benevolent interventions of Thiaoouban spacecraft observing Earth, Thao referred to *two* senseless tests in near-Earth orbit in October of 1961 and May of 1963 that were carried out as part of Lincoln Laboratory’s Project West Ford,[30] which involved a potentially disastrous attempt to create a new method of military communications using an artificial ‘dipole belt’ around the Earth:

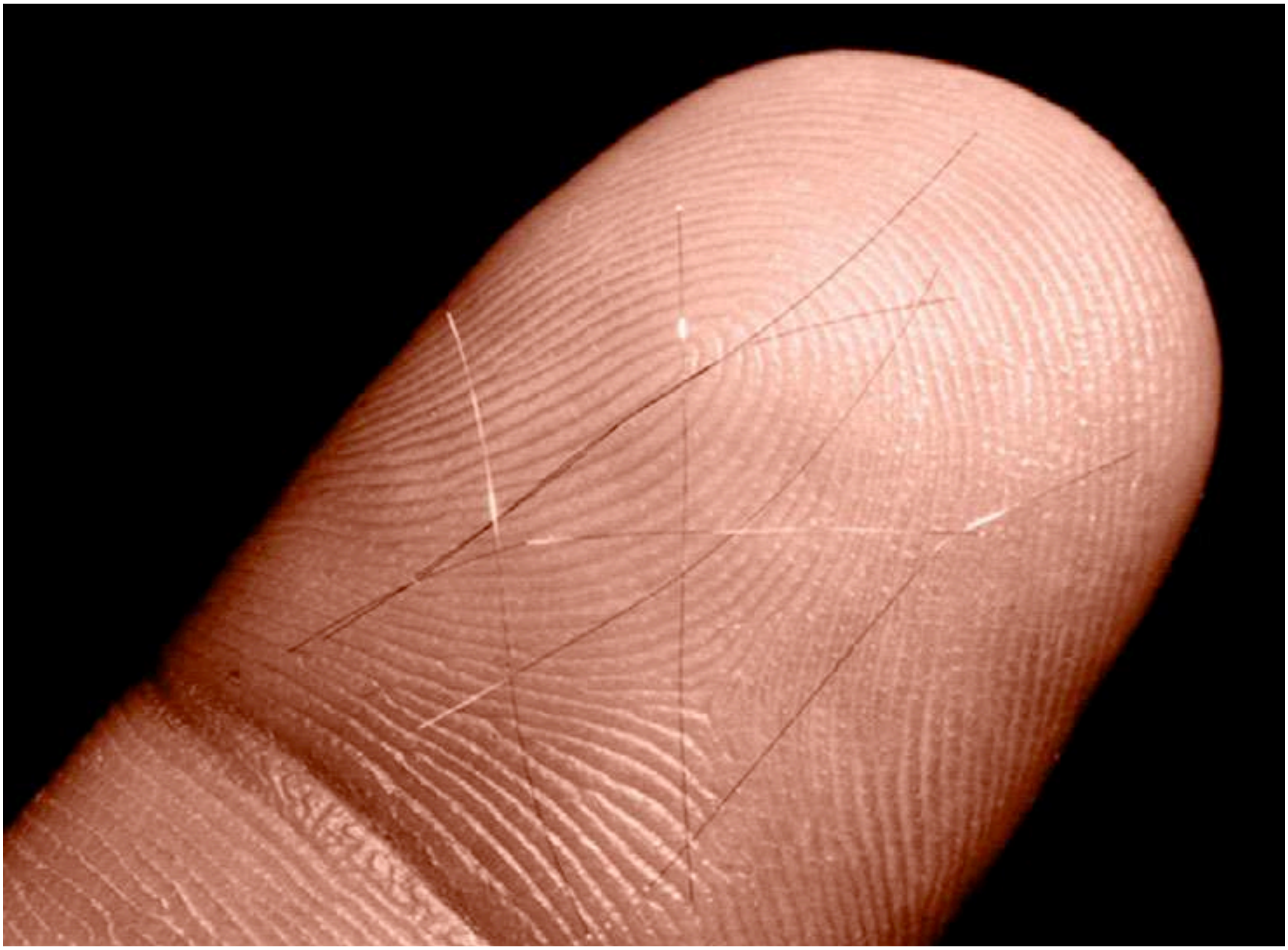

> The dipole belt was to serve, in effect, an artificial ionosphere… [by artificially creating] a belt of 480 million dipoles orbiting 2,200 miles above the earth… In addition to judiciously specifying the initial orbit parameters, we gave the dipoles a high area-to-mass ratio to ensure that the solar radiation pressure acting on the dipoles would distort their circular orbits into highly elliptical ones. The orbital perigee would eventually dip to an altitude at which air resistance would sweep the dipoles out of orbit… The first attempt at establishing the dipole belt, in October 1961, failed. The dipole-dispenser ejection mechanism didn't spin up, so no dipoles were released. With a redesigned dipole dispenser, a dipole belt was established in May 1963… By late 1965, the dipole belt had been swept from orbit, right on schedule.[31]

The Lincoln Laboratory photograph of 7 of the tiny dipole needles– 0.7" long and 0.0007" in diameter – (above) confounds the ability of the human mind to imagine the total volume of 480 million of these needles, which could easily have caused major damage to the Earth if allowed to melt together during atmospheric reentry to form several large molten masses that would rain down on the Earth in a series of major catastrophic impacts. The extreme myopia of the faulty thought process that directed this dangerous experiment is characteristic of 'government scientists' who endanger our planet by ignorant interference in natural atmospheric and orbital processes.

Prevention of the first mission's release of the dipole belt was achieved by interference with the functioning of the dipole-dispenser ejection mechanism. This simple preventive measure only frustrated NASA scientists, who mounted a second mission that was allowed to be released, and the experiment conducted, before the dipole belt was disintegrated by the plasma beams of Alatora spacecraft for the protection of all life on Earth. The oversight of human development on Earth is a significant task of the many advanced neighboring civilizations within our galaxy, who prevent such calamities only rarely by physical means when the logic and reason of Earthly humans cannot be inspired by any other means:

‘I see, Michel, you are having enormous difficulty believing what I have been saying. Never forget that we have THE *knowledge* - you have not seen one-tenth of what we can do. Attend carefully and I will give you a few examples to help you understand what I’m going to tell you.’

Thao stopped talking and appeared to be concentrating. As I watched, her face became a blur and, instinctively, I rubbed my eyes. Of course, this did not help and, in fact, she became progressively transparent until I could see right through her. Finally, she was no longer there - she had completely disappeared.

‘Thao,’ I called, slightly concerned, ‘where are you?’

‘Here, Michel.’ I jumped, for the voice came as a whisper, quite close to my ear…

Several feet in front of me, I saw the silhouette of Thao, completely golden yet radiant, as though inside her there burned a fire, its flames brief but intense. As for her face, it was recognizable but her eyes seemed to send forth little rays each time she spoke… She began to rise a few feet above the ground, without having moved a muscle of her ‘body’; then she started to circle the room, so fast I had trouble keeping my eyes on her. She stopped, eventually, above her seat and sat her ghostly form down. It was as though she were made of a shining mist - she was still recognizable as Thao and yet, quite transparent. The next instant she was gone. I looked around, but she had completely vanished.

‘Look no further, Michel, I am back.’ Indeed, there she was, in flesh and bone again, sitting on her seat.

‘How do you *do* that?’

‘As I was just explaining to you, we have THE *knowledge*. We can revive the dead; cure the deaf and the blind; make people walk who are paralyzed; we can cure any malady you care to name. We are masters, not *of* Nature, but *in* Nature, and we can do the thing most difficult of all - we can generate life spontaneously. *‘Out of the release of cosmic ray, we can create any type of living creature, including man…* We can create a human body, but that is done only by the great Thaori, taking infinite care, for the human body must be inhabited by several bodies, as you are aware - the physiological, the astral etc. If not, it would merely be a robot. Perfect knowledge is therefore required for such an undertaking… A man of twenty or thirty years-of-age can be created by the Thaori in approximately twenty-four of your Earth hours.’ [32]

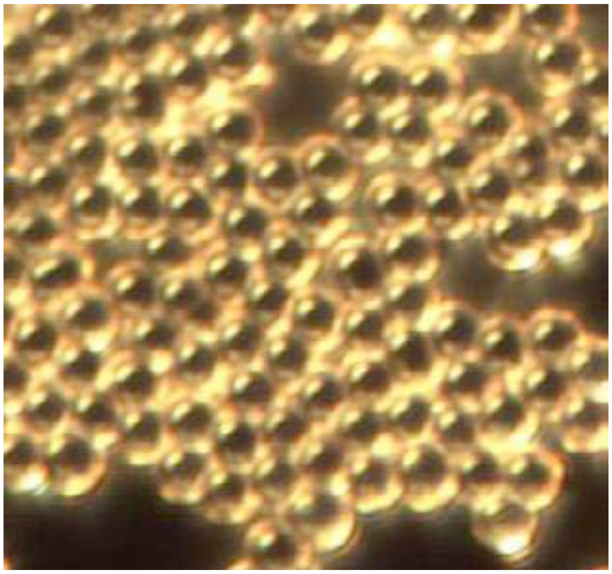

The astounding disclosures of spiritual knowledge provided through the writings of Michel Desmarquet inspire great contemplation. The transdimensional demonstrations and many penetrating truths offered by both Thao and the Thaori reflect their purely Creational consciousness, *as masters in Nature*. Michel’s galactic voyage to the golden planet of Thiaoouba stands as an example of the highest attainment of Nature, inspiring the human beings of Earth to reverse the destructive effects of our ignorant ways by the advanced spiritual use of gold and silver nanoparticles for resonant atmospheric and intracellular illumination (at left). The seeds of knowledge have been widely sown, and hold the magnificent promise of an era of rejuvenation using electrum-enriched foods and drinks to reclaim our own natural ability *‘to regenerate the cells of our bodies at will’* in the enhanced cosmic ray environment of the Betelgeuse supernova blast.

At this juncture, the development of human consciousness on Earth will benefit from dramatic cosmic environmental changes, *‘as Nature itself enters the stage to assist the pupil’* in the attainment of knowledge. Thiaoouban wisdom concerning the Four Forces of Creation, the Atomic, Ovocosmic, Ovoastromic and Astral, has been offered for our contemplation on Earth as all living beings receive DNA enhancements induced by cycling galactic infrasound waves. Their lessons identify the significant advances of Kirlian photography and the clean hydrogen engine as crucial developments for terrestrial technology on Earth, pointing directly toward the use of biocompatible HHO plasma for cellular regeneration, enhanced longevity and an inexhaustible source of clean, resonant energy.

References

---

[1] Desmarquet M (1993) *'Thiaoouba Prophecy'* Arafura Press, p. 6, online · http://www.thiaoouba.com

[2] Ibid, p. 18

[3] Ibid, pp. 7-11

[4] Ibid, pp. 52-54

[5] Sichin Z (1976) *'The 12th Planet'* Harper Books

[6] Vyasa K, tr. Prabhupâda B (1977) 'Description of Jambudvîpa' *Srimad Bhagavatam*, Canto 5, ch.16, online

· http://www.srimadbhagavatam.org/canto5/Canto5-Sanskrit/chapter16.html

[7] Macchi M, Bruce J (2004) 'Human Pineal Physiology and the Functional Significance of Melatonin' *Neuroendocrinology (25:3-4),* pp. 177-195

[8] Desmarquet, op cit, pp. 161-162

[9] -- (2010) 'It's True! Couples Share 6th Sense' *The Times of India*, online

· http://timesofindia.indiatimes.com/home/science/Its-true-Couples-share-6th-sense/articleshow/6640735.cms

[10] Desmarquet, op cit, pp. 61-63

[11] Ibid, p.154

[12] Ibid, p.156

[13] Ibid, pp. 64-67

[14] Kalyanaraman S (2004) *'Indian Alchemy: Soma in the Vedas'* Munishram Manoharlal Publishers, pp. 2-7

[15] Desmarquet, op cit, pp. 68-69

[16] Vandewalle G, Schwartz S, Grandjean D, Wuillaume C, Balteau E, Degueldre C, Schabus M, Phillips C, Luxen A, Dijk DJ, Maquet P (2010) 'Spectral Quality of Light Modulates Emotional Brain Responses in Humans' *PNAS*, online · http://www.pnas.org/cgi/doi/10.1073/pnas.1010180107

[17] Ibid, pp.73-82

[18] Dunham B (2009) 'Cool Plasma Pack Heat Against Biofilms', *University of Southern California News*, June 5,

online · http://www.humanresonance.org/helium.html

[19] Desmarquet, op cit, p. 96

[20] Ibid, p.137

[21] Ibid, p. 122

[22] Fox S (2009) 'EM Pulse Cuts Through Steel in 200 Milliseconds' *Popular Science*, December 3, online

· http://www.popsci.com/technology/article/2009-12/electromagnetic-pulse-cuts-through-steel-200-milliseconds

[23] Ibid, pp. 91-92

[24] Ibid, pp. 165-166

[25] Ibid, pp. 163-164

[26] Orion E (2003) *'The Bush Connection'* The Bush Connection

[27] Putney A (2009) *'Veil of Invisibility'* Human Resonance Productions

[28] Mills RL, Ray P, Dhandapani B, Nansteel M, Chen X, He J (2002) 'New Power Source from Fractional Quantum Energy Levels of Atomic Hydrogen Surpasses Internal Combustion' *Journal of Molecular Structure* 643, pp. 43-54

[29] Desmarquet, op cit, p. 109

[30] -- (1964) 'Project West Ford' *Proc. IEEE (52:5) special issue*

[31] MacLelland D (1970) 'Looking Back: Space Needles – An orbiting Belt of Tiny Dipoles Took Lincoln Laboratory into Space' *Massachusetts Inst. of Tech.*, online · http://www.ll.mit.edu/publications/labnotes/Looking_Back_3.pdf

[32] Desmarquet, op cit, pp. 161-162

6

Betelgeuse & the Red Dawn

Betelgeuse & the Red Dawn

An inexorable sequence of natural electromagnetic events is unfolding that demands our full attention, and underlies the shifting currents of human consciousness. Ancient Sanskrit knowledge maintained by the Maya culture explicitly describes the events surrounding the end of this solar cycle, as retold in many forms throughout the indigenous prophecies and wisdom traditions of the world. It is said humanity will be transformed in the red light of cosmic consciousness – quite directly telling us *'man becomes god.'*

Mayan wisdom identifies the energetic changes of December 22, 2012 as the effect of a *synchronization ray* emitted from the center of the galaxy. This event is described as an alignment of Earth with the galactic plane, which will ignite Earth's high atmosphere in an all-encompassing red aurora. The Pyramid of Fire Codex foretells that the sky itself will be in flames, and the red flames will spread amongst the pyramids – from one 'Hill of the Star' to another along flaming plumes of celestial water:

> Tláloc, God of Celestial Water, god of the vapor that *rises*, from the Earth warmed by the sun after the rains, god of the mist that *ascends* from the valleys at dawn… Tláloc is the return of vapor that strains to rise, is the return of time that strains to *remember*. Tláloc, God of the Fight Against the Current, with whose aid the hero battles against the torrent toward his own origin and beginning, towards the wings of his soul, the wings that Tláloc hides in the hero's past…
>
> It is the self within ourselves that we have to sacrifice. It is our own heart that has to be torn out of the false being and offered to the light. May Xiuhtecuhtli, Lord of Fire, burn my false being, May Itzli, Obsidian Knife, liberate my heart. To sow life in the fires of sacrifice: Are the flames rising on the Hill of the Star, toward Cauhlacan? Or is it my own heart that is afire?
>
> The cycle of years is past - the waiting ended. Come - reunite, pilgrims, for the sky is in flames! From Xochicalco to Teotíhuacan the red spreads one step, another step, and another, only twelve short steps from the cave of the womb to the final conflagration... The flames rage in consummation- Rise, oh flames! What light, what heat! What great holocaust! The smoke ascends and spirals... Obscurity disappears as the flames rise to Tonatiuh, the purest light.[1]

The kundalini technology of the pyramids employs the crystalline memory of celestial waters to awaken human divinity at sacred temples like Teotihuacan, meaning '*site-where-man-becomes-God'* (below). The resonant frequency of the Pyramid of the Sun at Teotihuacan closely matches the 1.45 Hz heartbeat oscillation of the Great Pyramid due to its comparable base length of 740'. Itza Mayan Elder Hunbatz Men describes the pyramids as generators of *the brilliant light of heaven that imbues the human being with eternal life.* Cosmic Creator Hunab K'u guides the initiatic work of Mayan Elders at sacred pyramids:

> "Take care of the light that I leave within my temple. When you form the human being, teach him to venerate my house, for within it they will find the light they need for eternal life. Teach him also that if he should stray, due to his errors, vices or ignorance, then I, Hunab K'u will reclaim my beloved beings and take them to my temple of wisdom and nourish them again with my light which is in the heavens. Then, together, Tepeu, Gugumatz and Huyubcaan will perform purification rituals using that brilliant light for the benefit of my beloved whom we created and molded, so they can live in the light of eternal wisdom." [2]

A heavenly life-giving light radiating inside pyramid chambers is also related in hieroglyphic texts inscribed on the interior walls of the pyramids of Saqqara, Egypt. Profound details of the luminal and acoustic effects of the Great Pyramid transducer in full operation have been recorded in stone:

> House bright and dark of Heaven and Earth, for the [solar] ships put together; Great Pyramid, House of the gods with pointed peak; for Heaven-to-Earth it is greatly equipped. House whose interior glows with a reddish Light of Heaven, pulsating a beam which reaches far and wide; its awesomeness touches the flesh... House of Eternity: its foundations are stones, the water; its great circumference is set in the clay. House the rightness of whose howling the Great-Ones-Who-See-and-Orbit bring down to rest… Mountain by which Utu ascends… whose deep insides men cannot penetrate.[3]

Both the Egyptian and Mayan hieroglyphs are descendant forms of the ancient logographic Sanskrit, as deciphered by epigrapher Dr. Kurt Schildmann (1909-2001). This antediluvian proto-Sanskrit can be identified as the Atlantean script, having been found on all inhabited continents of the world on stone and ceramic artifacts dated to the Pleistocene period – being over 12,000 years old. The pyramid traditions of these cultures of Atlantean descent provide us with stunningly explicit descriptions of future events tied into the ancient Sanskrit conception of water as the transformative agent of consciousness.

The Atlantean psychoacoustic technology of quartz crystal skulls reflects the Sanskrit tradition of augmenting consciousness through the purification and crystallization of the water within the human body. The refraction of light within the human body that is fueled by crystalline water enables the higher capacities of human consciousness by biorhythmic alignment at 1.45 Hz heartbeat entrainment in the pyramid chambers. The Great Pyramid's upper chamber does not contain a sarcophagus – the granite box was actually a water-birthing resonator filled with celestial water. The closed shafts rising from the chambers are its "deep insides men cannot penetrate", designed to focus infrasound heartbeats at 1.45 Hz into the tiny newborn infants. Usually born as twins, the infants' skulls were elongated by this process.

The celestial waters of La Maná, Ecuador (photographed above with several plasma orbs) and Tlacote, Mexico have been extensively studied for their astounding resonant properties.[4] Test results have shown the spring waters of La Maná and Tlacote have the highest water purity found anywhere in the world, containing virtually no salt or other dissolved solids. This in itself is remarkable, but tests have also identified trace amounts of gold and silver nanoparticles and an energy rating of over 1,000,000 on the Bovis scale. There is no modern technology capable of creating energized waters like those flowing from the springs of La Maná and Tlacote. These celestial waters are the nanotechnological product of the transdimensional civilization of Atlantis – the central component of the pyramid kundalini tradition of psychoacoustic gestation and water-birthing practices.

Artifacts from the ancient Sanskrit sacred site in La Maná, Ecuador reveal the origin and use of these celestial waters as a technological product of the global pyramid system itself. Over 300 metal, stone and ceramic artifacts have been recovered, among which are diverse UV-fluorescent maps of standing wave resonance patterns within the electron, mineral calcite, and Great Pyramid of Giza, Egypt.[5] The artifacts preserve logographic Sanskrit script, while their numerals from 1 to 10 are presented in fluorescent inlays adorning a set of 13 stone goblets precisely designed for water volume measurement.

Among the artifacts found below the sacred site is a large stone lingam in the form of a kundalini cobra, engraved with complex UV-fluorescent inlaid patterns – as well as a gold psychoacoustic helmet, accompanied by dozens of ceramic figurines that show both adults and children using the helmets while barefoot for electrical contact to the pyramid. The largest stone artifact from La Maná is a breathtaking global map of the Earth with all the continents accurately depicted as they were before the fall of Atlantis! Star maps on flat stone plates were inlaid with UV-fluorescent calcite mortars depicting the constellations surrounding Orion, with Betelgeuse (fluorescing in red) as one of the three large stars forming the Winter Triangle with Sirius and Procyon. This star map is also presented on the largest goblet (above, center).

Celestial waters were the primary enablers of this advanced psychoacoustic Atlantean technology, by which human bodies became superconductors of electricity. The suspension of gold and silver nanoparticles enhances the conductivity of the body by creating a pathway for electrical flow without shock. Shock waves do not form inside crystals for the crystal lattice distributes pressure evenly and the whole mass is set vibrating in unison. The celestial waters maintain the integrity of the body's cells during exposure to the electrical currents inside the pyramids at resonance. The gold and silver nanoparticles are from 2-10nm in size and act as transducers in the body, *converting electrical pulses into light pulses*.

Gold nanoparticles scatter RED light inside all of the cells of the body. This is the key to understanding the words of the pyramid-builders concerning the *"heavenly red light of eternal life"!* The fact that these water sources are still flowing, and in fact greatly increasing their output with time, is primary evidence that the pyramids are automatically roaring to life with the input of infrasound energy from solar storms as we approach the full onset of the new solar age. These widespread effects are already being observed.

*(**Note:** monatomic gold, aka white powder gold, does NOT scatter red light, and is not recommended!)*

Golden waters and wisdom of the Chichimeca and Chichicastl cultures of Tlacote, Mexico have been preserved by the Maya Elders, and presented through translation by Gerardo Aceves Quetza Sha:

> Tlacote is a great assembly point of the ancient gods where once existed the Golden City of the Golden Water, built by the Chichimeca and the Chichicastl. In other times, our ancestors left us transcendental information… Man has always been looking for something else, something that lives inside of him, and he recognized the inner light manifesting in his heart. It is this inner light that will cause the changes in humanity and that will make the dream come true, that the animal which is the human being will be able to live thousands of years. Man opens his heart to understand the revealing truth, creating consciousness and elevating his thoughts. This process is necessary, because we must understand and solve the enigmas of Earth and the planetary system, in order to develop in perfect conscience right from the beginning of creation. We meet thousands of phenomena in our life. Man calls this coincidence; spirit calls it cycles of comprehension in the evolution of man. Therefore, if a new race emerges to a new cycle of existence, this happens because the information imprinted in the subconscious of the Inner Self has been completed.

> All the information we need is imprinted in the subconscious of the inner self, so it may evolve according to its individual level, remembering who it is and knowing its task on Earth: 'AWAKEN, INNER SELF, and recognize your inner creation in order to understand the movements of the Universe!' This is how the different races and religions developed in their particular states of evolution. Each religion in its particular language has accompanied us on a very important part of our path and accomplished the demands of a particular time. Now humanity is awakening to a new cycle of consciousness. A new race is born, constantly bringing in new knowledge and spheres of conscience, and being aware of the past, the present and the future…
>
> This message predicting the awakening of the Golden Water is sacred and was created and given to us by beings from other dimensions, so that other generations can be born, prepared for the great changes that will be produced from the center of the Earth.[6]

Not only are the acoustic environments enhanced at these sacred sites, but the quality of water is also correspondingly enhanced to become *crystalline.* These profound findings reveal the sophisticated yet invisible technologies that enhance human consciousness and longevity at these sites – for those who choose to live close to spirit and purify their bodies to be bathed in the energy of the pyramids. The red light scattering of golden celestial waters induces DNA enhancement through heartbeat synchronization.

The ancient Hopi wisdom echoes that maintained by the Maya, describing the imminent shift from the second phase of Creation, during which the sun has shone yellow, to the third phase of Creation during which the Earth is bathed in red light: “The third phase of Creation is the time of the red light, Talawva, the perfect light. Humanity then, having been fully formed and [crystallized,] proudly faces the Creator for the first time, to know the true warmth of love.” While the Hopi prophecies ascribe no calendar date to the coming of the red light of the new solar age, their wisdom perfectly harmonizes with that of the Mayan culture, with its profound and complex calendar system providing the end date of December 22, 2012.

What could possibly trigger the dramatic changes in human consciousness, involving a shift in the light reaching Earth’s surface to a red hue? Solar physicists confirm the 11-year cycle of solar activity corresponds with the 2012 end date, the last solar maximum having been recorded in 2001. At each solar maximum the sun’s magnetic field reverses, but what could possibly induce the Earth’s magnetic field to reverse with our sun? Recent stellar observations in 2009 reveal the hidden threads which link one star system’s fate with another. A research team at UC Berkeley has recently reported data indicating the nearby red supergiant star Betelgeuse is shrinking in the process of imminent collapse:

> Betelgeuse is enshrouded by vast clouds of gas and dust, so measuring its size is difficult. To cut through this cocoon, Charles Townes of the University of California, Berkeley, and colleagues used a set of telescopes that are sensitive to a particular wavelength of the star's infrared light.
>
> But the star's reach seems to be waning. New observations indicate the giant star has shrunk by more than 15 per cent since 1993. This could be a sign of a long-term oscillation in its size or the star's first death knells…
>
> The team used these instruments to measure the size of Betelgeuse's disc on the sky. Over a span of 15 years, the star's diameter seems to have declined from 11.2 to 9.6 AU (1 AU, or astronomical unit, is the distance from the Earth to the sun)… The cause for this reduction is unknown, as it is unclear how red supergiants behave near the end of their lives.[7]

One may easily identify similar patterns in the energy fluctuations of both red supergiant Betelgeuse and our own small yellow Sol. The contraction and collapse of Betelgeuse is being mirrored on the Sun, which has seen a prolonged period of spotless inactivity for the period of 264 days preceding the appearance of Sunspot 1024 in July of 2009. *Both stars dimming together represents the contraction before the full strike of the wave, as the supernova’s energetic blast wave reaches the Earth.*

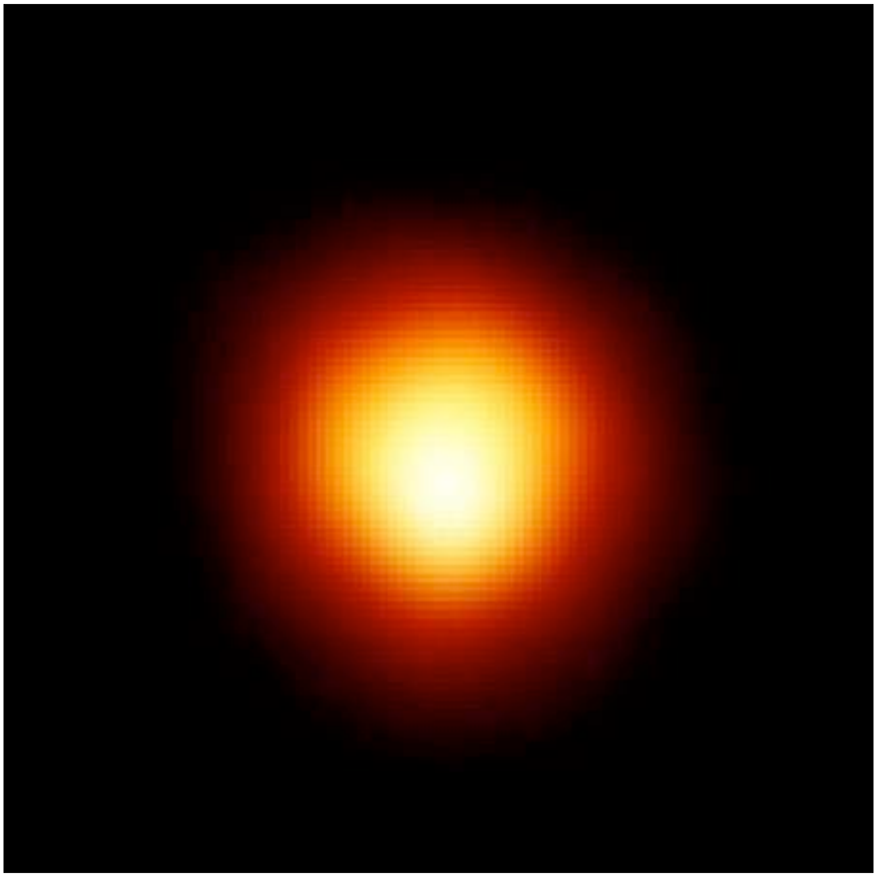

Plasma physicists proposing the Electric Universe theory have keenly observed the pulsed throughput of energy that generates the dynamics of stellar interactions, overturning the false notion of stars as independent thermonuclear generators with their own self-determined lifecycle. The lifecycles of stars are actually intertwined with the cycling waves of energy that traverse the spiral arms of our galaxy. Galactic waves are generated as pulsations of infrasound emitted from the center of our galaxy that traverse its arms, gaining momentum from the radiated energy of supernovae that are set off by the wave's influence in a chain of events moving outward. Red supergiant Betelgeuse, at the shoulder of the constellation Orion (at left), is the nearest in a chain of supernova events along the expanding wave front, and acts as our local indicator by which we can estimate the arrival time of the galactic energy wave.

Further analysis of the data obtained by the UC Berkeley team should also provide the rate of acceleration of the stellar collapse and give us a fairly good estimate for the moment of supernova. Considering the fact that Betelgeuse is our close neighbor at about 600 light-years distance, this event has already occurred and the nova will be visible to us in just a few years time. So, when exactly will the electromagnetic wave hit Earth, and how will it affect us?

Certainly, the Betelgeuse supernova will be seen as a major flash of light plainly visible even during the daylight hours. The supernova may shine very brightly for many days. The more complex unfolding of events triggered by the supernova involve many factors not commonly understood by astrophysicists. For time indications of the potential hazards and benefits of these electromagnetic changes we must turn to the most ancient and most accurate calendar system known to man. Mayan wisdom gives us an exact date for the change of the ages (December 22, 2012), and an exact sign is given as well – the Red Sky.

The Maya use of sacred celestial waters within the pyramids for healing and longevity is an open invitation to sacrifice the false comforts of urban society that disrupt the natural acoustic environment and isolate people from electrical contact with the Earth. Industrial man finds dissonance and disease in concrete boxes, whereas resonance and longevity are to be found in the celestial waters and barefoot lifestyle practiced at sacred sites. The Maya invite us all to come together as one: "The cycle of years is past - the waiting ended. Come - reunite, pilgrims, for the sky is in flames! From Xochicalco to Teotíhuacan the red spreads one step, another step, and another…"

Yaqui Elder Don Juan spoke of waiting for the 2012 event during his teachings to cultural anthropologist Carlos Castaneda, stating emphatically: "[As seers] our case is a bit different, because we are at the end of our trail. We are not seeking anything... we go from day to day doing nothing. We are waiting. I will not tire of repeating this: we know that we are waiting and we know what we are waiting for. We are waiting for freedom!" [8] The progressive energetic changes currently raising human consciousness will cause fear and panic only for those who are unprepared, while many spiritually awakened individuals have been preparing for the Purification Day.

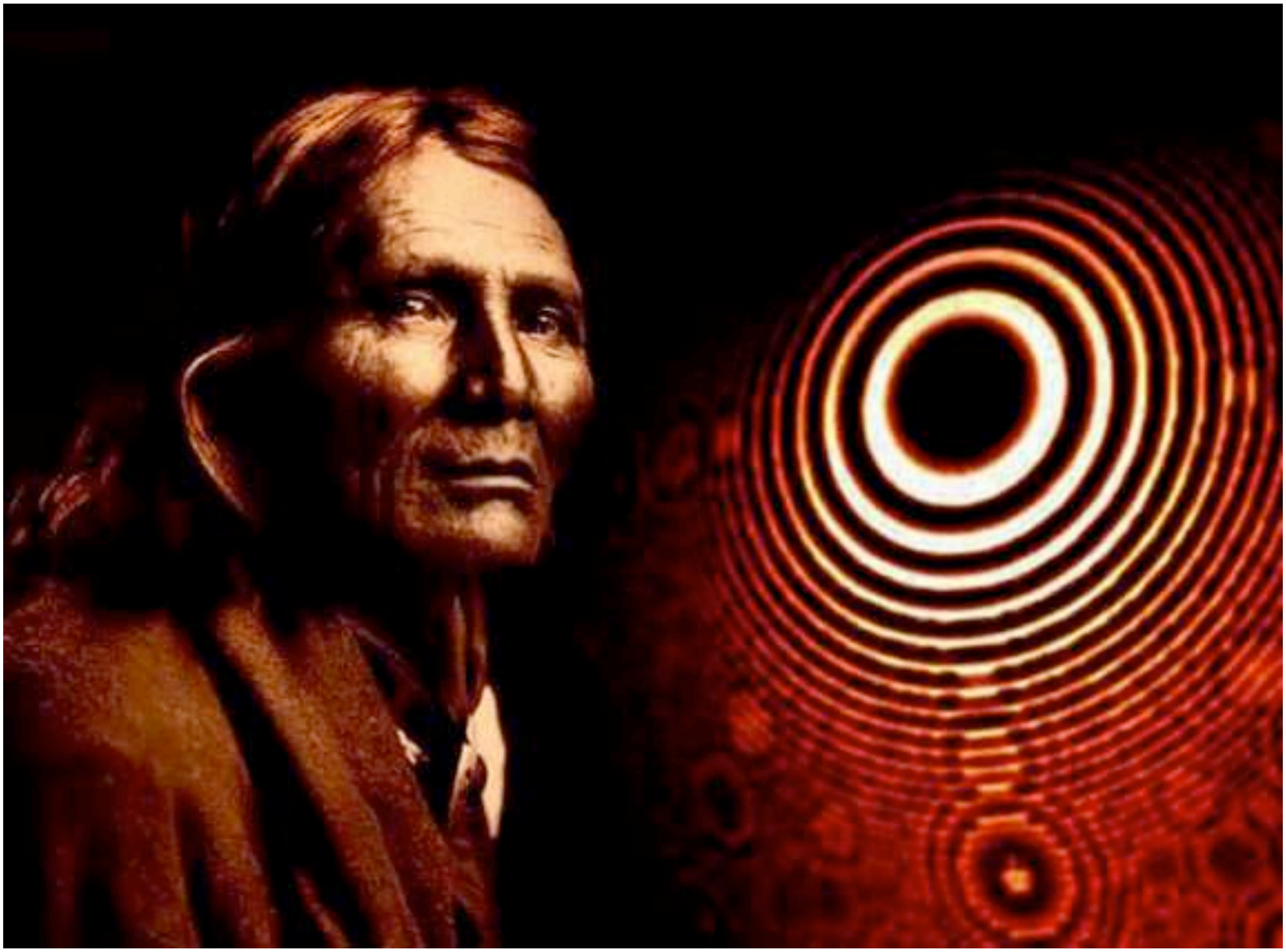

Apache Elder Stalking Wolf related his profound visions of the future in the Prophecy of the Red Sky, received as a vision several decades before the foretold events, and later published by Tom Brown Jr.:

> In the 1920s, an Apache wise man had a Vision of four prophecies that foretold death and destruction for mankind, unless we incorporate Spirit in our daily lives. Two of these prophecies may already have come true. A number of people can predict the future, but few get the timing correct. "Grandfather" was an Apache wise man and scout, named Stalking Wolf, who grew up outside white man's influence. His many predictions not only came true in the manner he predicted, but also when he predicted...
>
> Grandfather had been wandering for several years and was well into his forties when the Vision of the four signs was given to him. He had just finished his third Vision Quest at the Eternal Cave when the Vision made itself known. He had been seated at the mouth of the cave, awaiting the rising Sun, when the spirit of the warrior appeared to him. He felt as if he were in a state somewhere between dream and reality, sleep and wakefulness, until the spirit finally spoke and he knew that it was not his imagination. The spirit called Grandfather's name and beckoned him to follow.
>
> As Grandfather stood, he was suddenly transported to another world. Again, he thought that he was dreaming, but his flesh could feel the reality of this place; his senses knew that this was a state of abject reality, but in another time and place.
>
> The spirit warrior spoke to Grandfather. "These are the things yet to come that will mark the destruction of man. These things you may never see, but you must work to stop them and pass these warnings on to your grandchildren. They are the possible futures of what will come if man does not come back to the Earth and begin to obey the laws of Creation and the Creator. There are four signs, four warnings, that only the children of the Earth will understand. Each warning marks the beginning of a possible future, and as each warning becomes reality, so too does the future it marks"...

The First Sign

..."Welcome to what will be called the 'land of starvation'. The world will one day look upon all of this with horror and will blame the famine on the weather and the Earth. This will be the first warning to the world that man cannot live beyond the laws of Creation, nor can he fight Nature... There will come starvation before and after this starvation, but none will capture the attention of the world with such impact as does this one. The children of the Earth will know the lessons that are held in all of this pain and death, but the world will only see it as drought and famine, blaming Nature instead of itself"...

In his dream, the spirit spoke to Grandfather. "It is during the years of the famine, the first sign, that man will be plagued by a disease, a disease that will sweep the land and terrorize the masses. The white coats [doctors/scientists] will have no answers for the people, and a great cry will arise across the land. The disease will be born of monkeys, drugs and sex. It will destroy man from inside, making common sickness a killing disease. Mankind will bring this disease upon himself as a result of his life, his worship of sex and drugs, and a life away from Nature. This, too, is a part of the first warning; but, again, man will not heed this warning and he will continue to worship the false gods of sex and the unconscious spirit of drugs."

The Second Sign

The spirit looked at Grandfather for a long moment, then finally spoke. "Holes in the sky... will become the second sign of the destruction of man. The holes in the sky and all that you have seen could become man's reality. It is here, at the beginning of this second sign, that man can no longer heal the Earth with physical action. It is here that man must heed the warning and work harder to change the future at hand. But man must not only work physically, he must also work spiritually, through prayer, for only through prayer can man now hope to heal the Earth and himself...

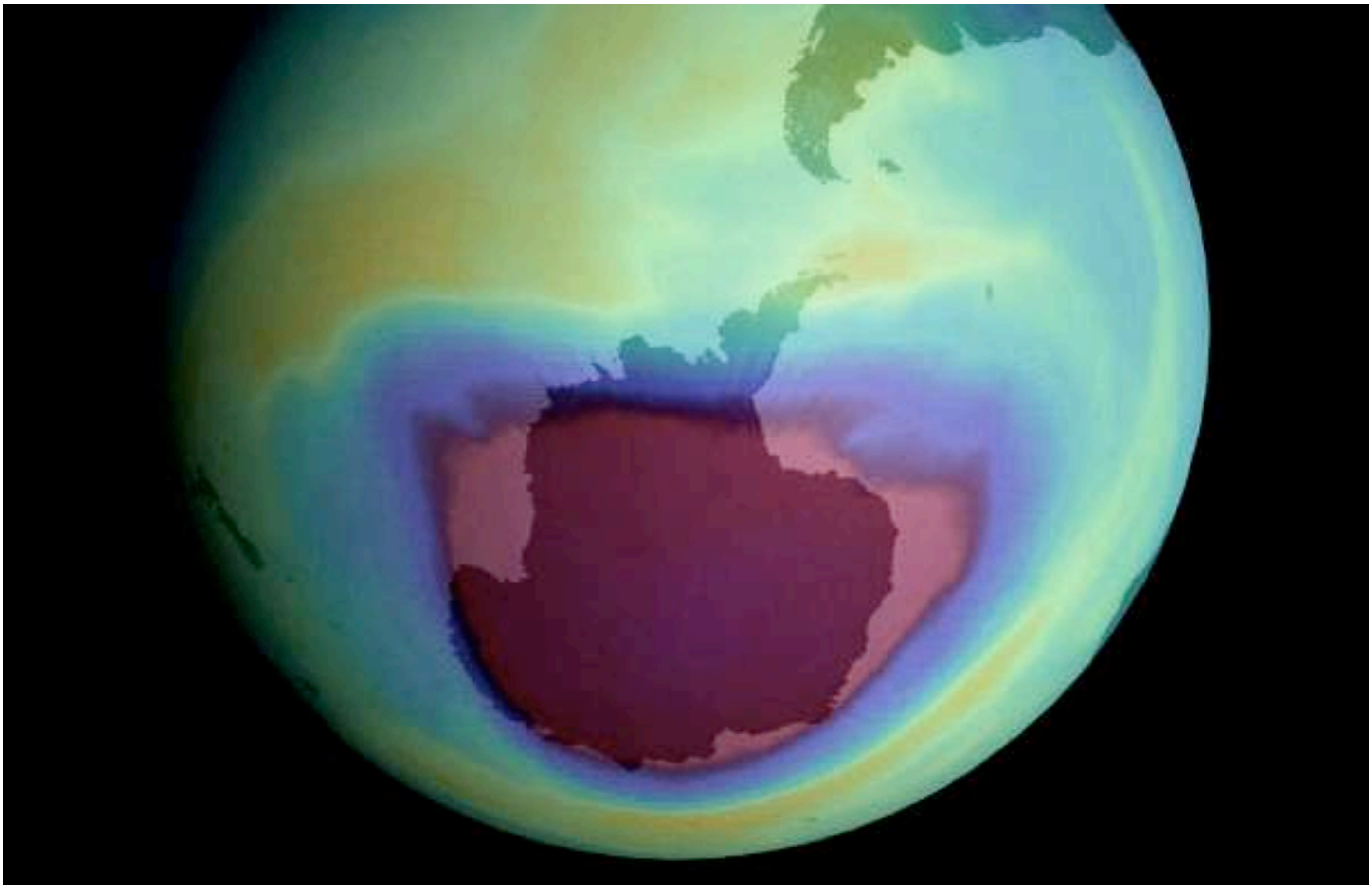

"These holes are a direct result of man's life, his travel, and of the sins of his grandfathers and grandmothers. These holes, the second sign, will mark the killing of his grandchildren and will become a legacy to man's life away from Nature. It is the time of these holes that will mark a great transition in mankind's thinking. They will then be faced with a choice-a choice to continue following the path of destruction, or a choice to move back to the philosophy of the Earth and a simpler existence. It is here that the decision must be made, or all will be lost...

The Third Sign

Grandfather spent the next four days at the cave entrance, though for those four days nothing spoke to him, not even the Earth. He said that it was a time of great sorrow, of aloneness, and a time to digest all that had taken place.

He knew that these things would not appear in his lifetime, but they had to be passed down to the people of the future with the same urgency and power with which they had been delivered to him. But he did not know how he would explain these unlikely events to anyone. Surely the elders and shamans of the tribes would understand, but not society, and certainly not anyone who was removed from the Earth and Spirit.

He sat for the full four days, unmoving, as if made of stone, and his heart felt heavy with the burden he now carried.

It was at the end of the fourth day that the third Vision came to him. As he gazed out onto the landscape towards the setting Sun, the sky suddenly turned to a liquid and then turned blood-red. As far as his eyes could see, the sky was solid red, with no variation in shadow, texture or light. The whole of Creation seemed to have grown still, as if awaiting some unseen command. Time, place and destiny seemed to be in limbo, stilled by the bleeding sky. He gazed for a long time at the sky, in a state of awe and terror, for the red color of the sky was like nothing he had ever seen in any sunset or sunrise. The color was that of man, not of Nature, and it had a vile stench and texture. It seemed to burn the Earth wherever it touched. As sunset drifted to night, the stars shone bright red, the color never leaving the sky, and everywhere the cries of fear and pain were heard.

Again, the warrior spirit appeared to Grandfather, but this time as a voice from the sky. Like thunder, the voice shook the landscape. "This, then, is the third sign, the night of the bleeding stars. It will become known throughout the world, for the sky in all lands will be red with the blood of the sky, day and night. It is then, with this sign of the third probable future, that there is no longer hope. Life on the Earth as man has lived it will come to an end, and there can be no turning back, physically or spiritually. It is then, if things are not changed during the second sign, that man will surely know the destruction of the Earth is at hand. It is then that the children of the Earth must run to the wild places and hide. For when the sky bleeds fire, there will be no safety in the world of man."

Grandfather sat in shocked horror as the voice continued. "From this time, when the stars bleed, to the fourth and final sign, will be four seasons of peace [that is, one year]. It is in these four seasons that the children of the Earth must live deep in the wild places and find a new home, close to the Earth and the Creator. It is only the children of the Earth that will survive, and they must live the philosophy of the Earth, never returning to the thinking of man. And survival will not be enough, for the children of the Earth must also live close to the Spirit. So tell them not to hesitate if and when this third sign becomes manifest in the stars, for there are but four seasons to escape"… Grandfather said that the voice and red sky lingered for a week, and then were gone as quickly as they were manifest.

The Fourth Sign

He did not remember how many days he'd spent at the mouth of the cave, nor did it make a difference, for he had received the Vision he had come for. It was in his final night at the Eternal Cave that the fourth Vision came to Grandfather, this time carried by the voice of a young child.

The child said, "The fourth and final sign will appear through the next ten winters [that is, ten years] following the night that the stars will bleed. During this time, the Earth will heal itself and man will die. For those ten years, the children of the Earth must remain hidden in the wild places, make no permanent camps, and wander to avoid contact with the last remaining forces of man. They must remain hidden, like the ancient scouts, and fight the urge to go back to the destruction of man. Curiosity could kill many."

There was a long silence, until Grandfather spoke to the child spirit, asking, "And what will happen to the worlds of man?"

There was another period of silence until finally the child spoke again. "There will be a great famine throughout the world, like man cannot imagine. Waters will run vile, the poisons of man's sins running strong in the waters of the soils, lakes and rivers. Crops will fail, the animals of man will die, and disease will kill the masses. The grandchildren will feed upon the remains of the dead, and all about will be the cries of pain and anguish. Roving bands of men will hunt and kill other men for food, and water will always be scarce, getting scarcer with each passing year. The land, the water, the sky will all be poisoned, and man will live in the wrath of the Creator. Man will hide at first in the cities, but there he will die. A few will run to the wilderness, but the wilderness will destroy them, for they had long ago been given a choice. Man will be destroyed, his cities in ruin, and it is then that the grandchildren will pay for the sins of their grandfathers and grandmothers."

"Is there then no hope?" Grandfather asked.

The child spoke again. "There is only hope during the time of the first and second signs. Upon the third sign, the night of the bleeding, there is no longer hope, for only the children of the Earth will survive. Man will be given these warnings; if unheeded, there can be no hope, for only the children of the Earth will purge themselves of the cancers of mankind, of mankind's destructive thinking. It will be the children of the Earth who will bring a new hope to the new society, living closer to the Earth and Spirit."

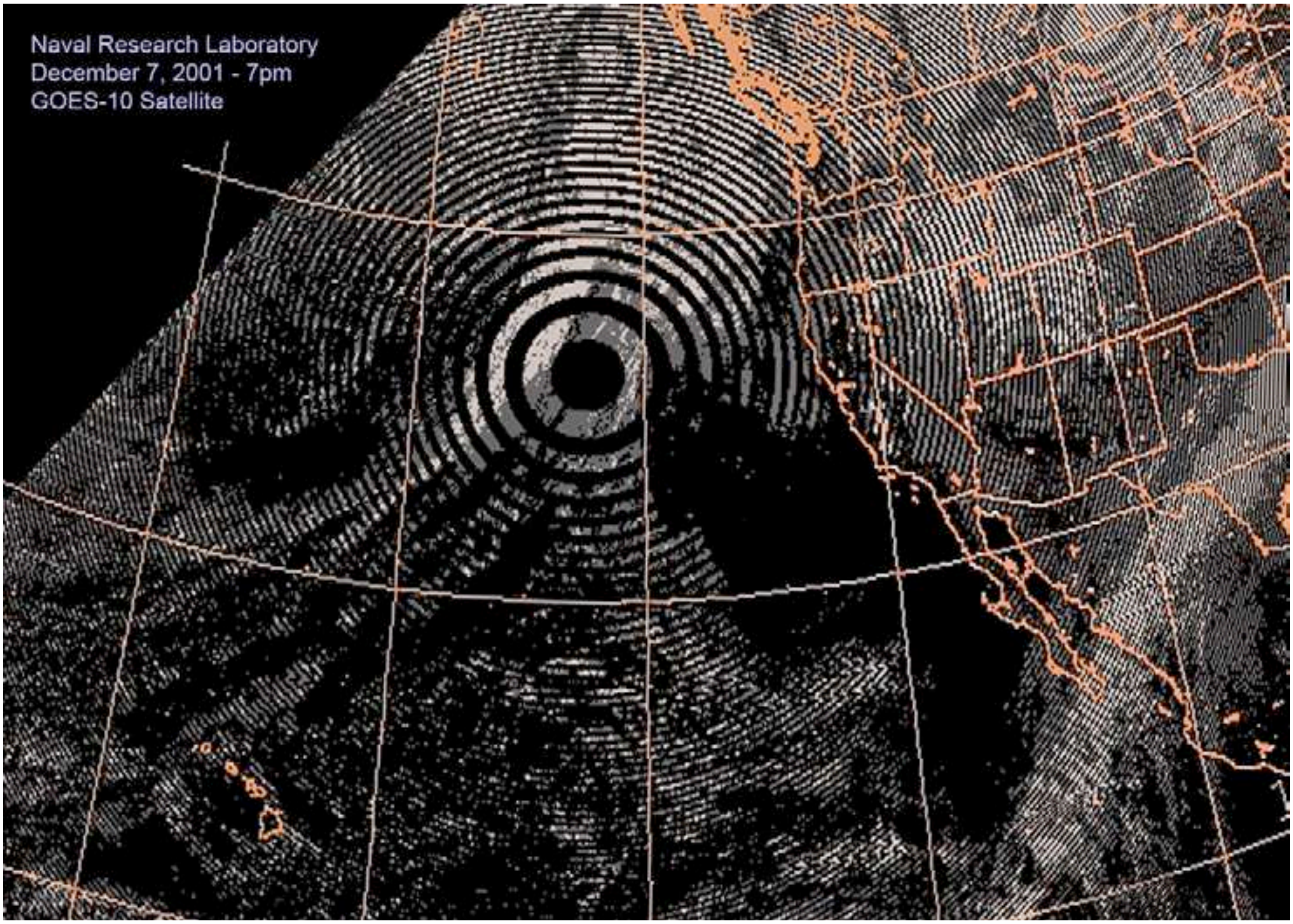

> Then all was silent, the landscape cleared and returned to normal, and Grandfather stepped from the Vision. Shaken, he said that he had wandered for the next season, trying to understand all that had been given to him, trying to understand why he had been chosen.
>
> Grandfather had related the story to me in great detail during that night of the four prophecies. I don't think that any event had been left out, and his emotions and thoughts were such that he actually relived it for us. Thus the power of his Vision became part of our spirit, our driving force, and a big part of our fears.[9]

The prophetic accuracy of Stalking Wolf's visions are quite stunningly clear thus far, both in their highly detailed descriptions and sequence of events, as presented to the young Tom Brown and friends. Here is the prophesied timeline as interpreted in the context of the ongoing magnetic polar reversal:

1st Sign:	1975–1985	African famine and AIDS pandemic spread
2nd Sign:	1985–1995	Holes form in Earth's ozone layer, permitting destructive solar rays
3rd Sign:	Dec 22, 2012	Magnetic reversal ignites a plasma canopy seen as red skies worldwide
4th Sign:	2013–2023	Large-scale destruction of low-resonance human technology ensues

Apache Grandfather Stalking Wolf experienced visions of the dramatic changes that have since occurred, and are yet to come, providing us with the culminating event marking the emergence of a new solar era. However, there are multitudes of more subtle signs of these oncoming electromagnetic changes that can be understood only in this context. The effects of shifting gravity and enhanced electromagnetic fields have been documented all over world, especially in areas where we find pyramids, or megalithic structures. These electromagnetic fields are focused by the axis-symmetric geometries of the pyramids of Giza, Egypt, thereby redistributing solar energy into a fixed global pattern of infrasound standing waves. These patterns of infrasound resonance have even been observed by satellite during the last solar maximum, on the night of December 7-8, 2001, in the infrared spectrum (above).[10]

Deeply resonant humming heard in these infrasound hotspots is continually generating millions of reports of disturbance and annoyance worldwide, with new telephone hotlines even fielding calls from concerned residents to quell the panic. Deep infrasound booms like heartbeats have been reported all over the world by residents of Llanidloes, Mawnan, Bridlington, Auckland, Sydney, Klai, Ranchlands, and in the US in Anderson,[11] Kimberley,[12] Richmond,[13] Wilmington,[14] Knoxville,[15] Mobile,[16] Knob Noster,[17] Denver,[18] Seattle,[19] Arroyo Grande[20] and Atwater.[21] Infrasound resonance focal points in the US are mapped below.

What is initially perceived as an annoyance quickly becomes a clear and present danger, as the infrasound resonance starts to overload all metals and wires with electricity until the extreme heat ignites fires. As Stalking Wolf emphasizes, "when the sky bleeds fire, there will be no safety in the world of man." All of the modern electrical technologies of humanity will not only cease to function, but will endanger human beings by spontaneous combustion. The surprising ignition of plastic-sheathed copper wiring was first noted in 1859 when a massive solar ejection shorted out and then ignited telegraph lines across the entire United States and all of Europe, and auroras were seen as far south as Rome and Hawai'i.[22]

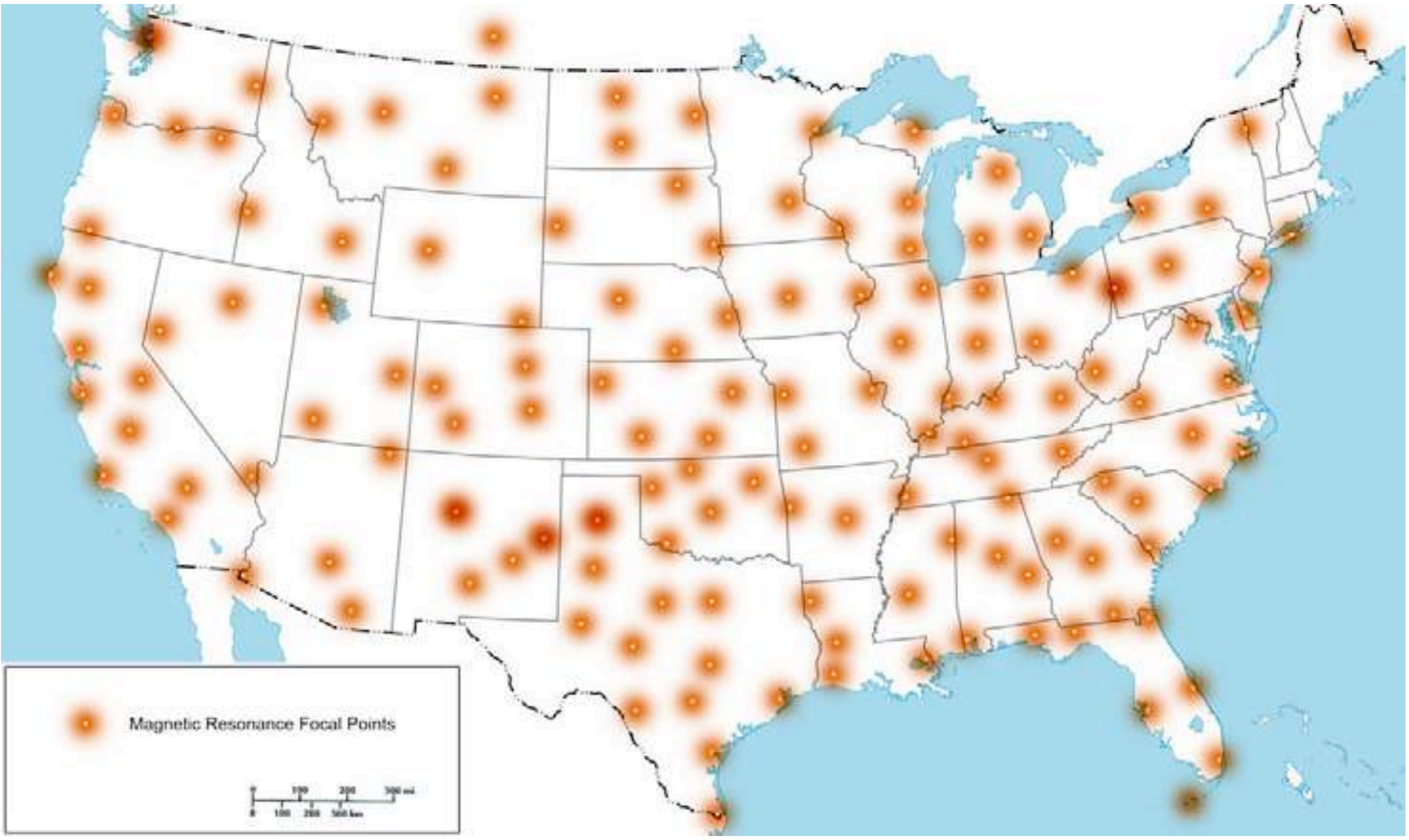

Periods of elevated infrasound resonance have generated hundreds of fires in homes across the world. These resonant fires are located in the nodes of an infrasound standing wave pattern that is increasing in intensity, eventually requiring total evacuation of residents from areas like Tenerife,[23] Babura,[24] Lalapansi,[25] Mapuve,[26] Bodibe,[27] Landovica,[28] Peschici,[29] Berici,[30] Messina,[31] northern Greece,[32] Kishtwar,[33] Ratria,[34] Santo Tomas,[35] and in the US in Seattle,[36] Colorado Springs,[37] San Mateo[38] and Santa Barbara.[39] A grave danger to personal safety in these regions of infrasound resonance will be posed by cellphones, which will ignite spontaneously in the pockets of clothing, as discovered already by one unfortunate man in Vallejo, California.[40] Such surprising events will be an inevitable occurrence in the years to come.

One of the most interesting incidents of this kind recently occurred in Soesterberg, Netherlands where, on the morning of July 19, 2009, ball lightning descended on the town in an unprecedented series of events.[41] Residents reported fast-moving ball lightning in more than 100 homes that sustained extensive damage to electrical systems. One eyewitness described a purple beam of light that shot skyward through the roof of one home. The fact that nobody was injured in Soesterberg is remarkable.

Clearly, modern civilization’s embrace of electrical wiring is becoming a dangerous trap in many parts of the world, eventually threatening to kill anyone who cannot return to a life close to spirit and in the natural places of the Earth. Even high above the planet, infrasound drumming has been reported on multiple occasions onboard the International Space Station (ISS) in 2004 – the result of an infrasound resonance of standing waves that will inevitably destroy the station’s systems as it has in many homes far below:

> The two men aboard the international space station heard a strange metallic sound again Friday, four months after being startled by it the first time… Cosmonaut Alexander Kaleri was talking to flight controllers in Moscow when he heard a loud drum-like noise coming from the instrument panel of the station's Russian-built living quarters.
>
> Kaleri and astronaut Michael Foale first heard the mystery noise - described as a flapping sheet of metal - back in late November. Neither the crewmen nor flight controllers were ever able to identify the sound… "It's very strange," Russian Mission Control said. "I doubt that it would be a coincidence that you're hearing the same thing coming from the same place."
>
> During a spacewalk in February, Kaleri and Foale were supposed to check the exterior of the space station where the noise originated last November. But Kaleri's spacesuit overheated and became damp, and the spacewalk had to be cut short, so the men did not have time to inspect the area.[42]

The fact that Kaleri's investigative spacewalk was not possible as his spacesuit overheated is another result of the electrical effects of the standing wave resonance. While the event has been passed off as the result of 'space junk', NASA directed a spacewalk mission that abruptly failed, yet no link has been drawn between the anomalous acoustic energy heard with the anomalous heating of the astronaut's spacesuit. The infrasound energy that is likely causing the noisy drumming and the overheating of the cosmonaut's spacesuit will likely recur for a longer duration, building an electrical charge capable of rendering the systems of the International Space Station inoperable.

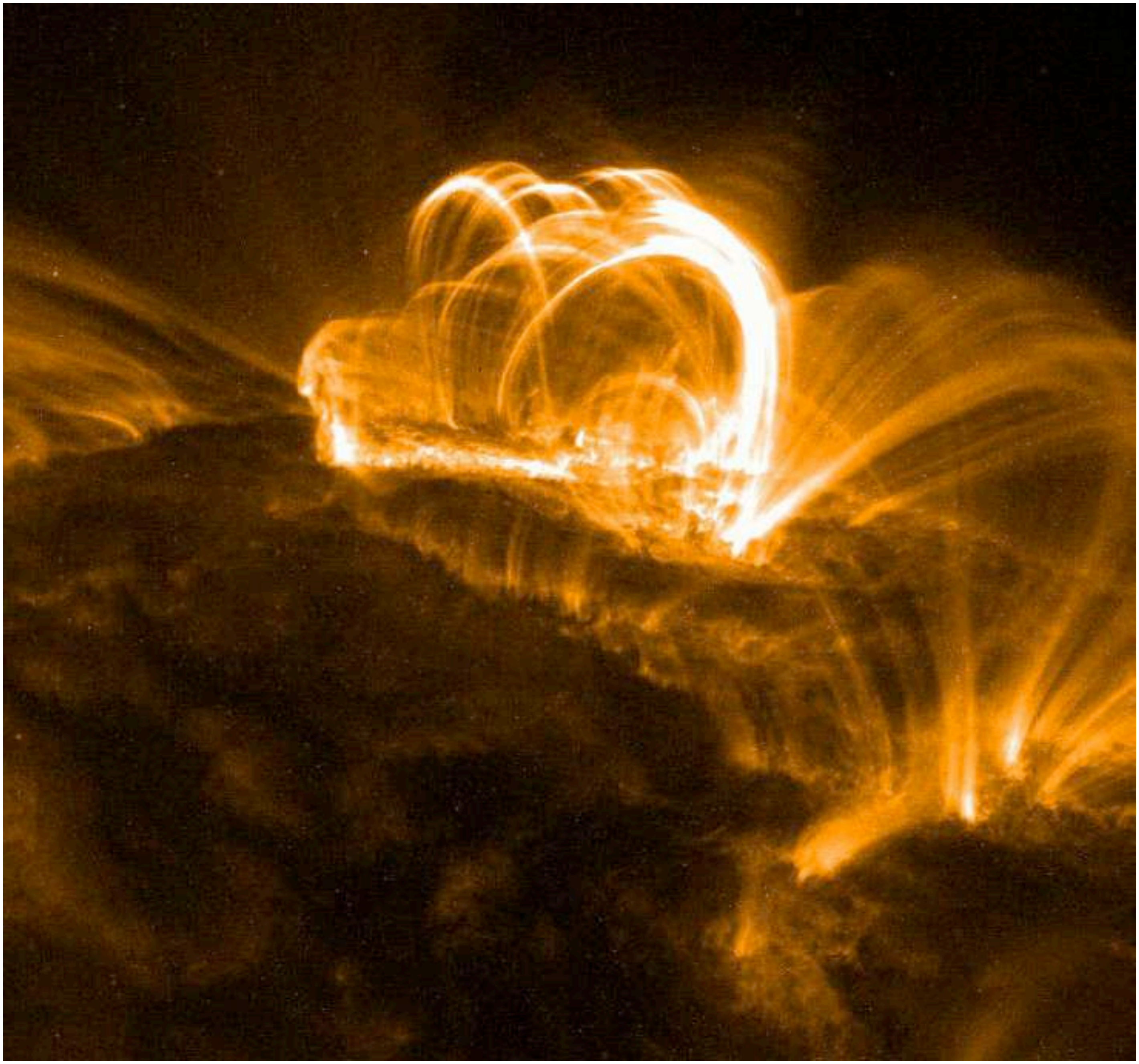

Hopi Elder White Feather relates a sacred prophecy describing the nine signs preceding the emergence of the new solar age on Earth – known to the Hopi as The Fifth World. The final sign clearly describes the destruction of the ISS as the last warning preceding the morning of the Red Dawn:

> "And this is the Ninth and Last Sign: You will hear of a dwelling-place in the heavens, above the Earth, that shall fall with a great crash. It will appear as a blue star. Very soon after this, the ceremonies of my people will cease.
>
> "These are the Signs that great destruction is coming. The world shall rock to and fro. The white man will battle against other people in other lands -- with those who possessed the first light of wisdom. There will be many columns of smoke and fire such as White Feather has seen the white man make in the deserts not far from here. Only those which come will cause disease and a great dying.
>
> "Many of my people, understanding the prophecies, shall be safe. Those who stay and live in the places of my people also shall be safe. Then there will be much to rebuild. And soon - very soon afterward - Pahana will return. He shall bring with him the dawn of the Fifth World. He shall plant the seeds of his wisdom in their hearts. Even now the seeds are being planted. These shall smooth the way to the Emergence into the Fifth World."

The Hopi prophecy details the order of events to occur, which can be deciphered when combined with the Mayan calendar wisdom which links the Red Dawn prophecy with the end date of 12.22.2012. Thus, the ISS may be interpreted as the sign of the Blue Star Katchina that precedes the 2012 event, like an alarm clock heralding its imminence. The Hopi tell that soon after the fall of the space station will come the appearance of the Red Star Katchina bringing the Day of Purification, when the skies in all lands of the Earth will bleed red.

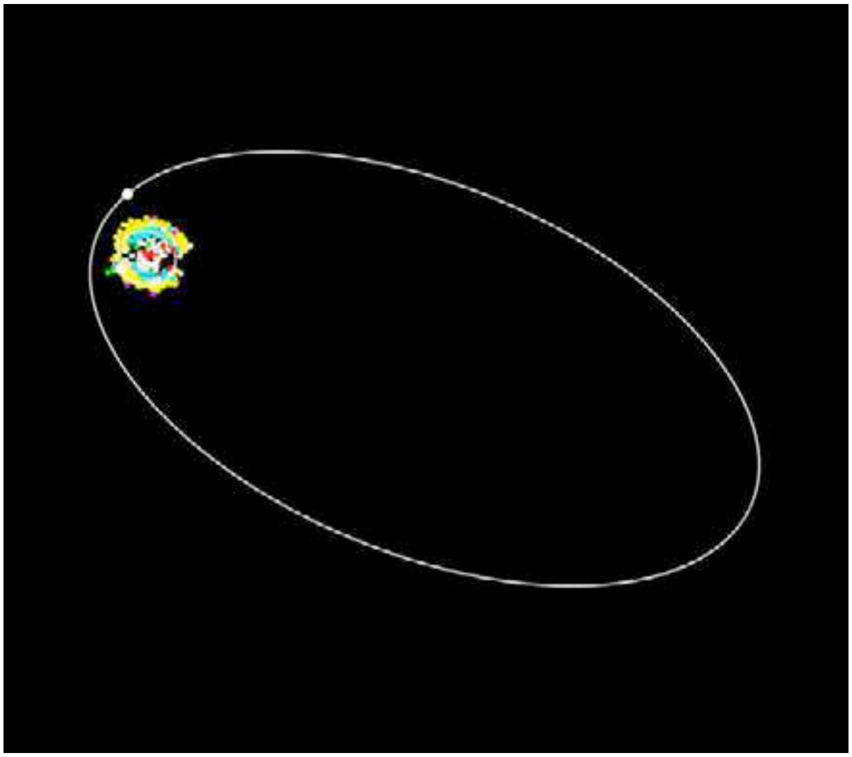

Hopi Elders describe this future cosmic event in simple and clear terms: "a small Red Star which will come very close and sit in our heavens watching us." Now, just wait a minute… stars are not known to wander around and sneak up on anyone! What exactly is being suggested by the Hopi? While it may come as a startling surprise for many, the Hopi description *accurately* reflects the scenario presented by the Binary Research Institute.[43] Astrophysics is now coming to grips with the reality that our solar system is a binary star system - that we will imminently discover a dark star companion in close proximity to our sun. The binary model is the only one that predicts the extremely elliptical orbit of Sedna (inset diagram at left), which must be highly influenced by our sun's dark twin to maintain such an orbit.[44]

The Hopi sacred prophecies tell of the impending incandescence of a red star in the close vicinity of Earth, describing it as a revelation of the solar twins. The now hidden brown dwarf companion star will soon reach a critical proximity to the sun – its dormant surface will reignite, transforming it into a brilliant red star. The small size and dark brown color of the long-ago extinguished dark twin sun has allowed it to creep on its elliptical path back into close proximity of the solar system without our awareness, and its ignition will likely be triggered by the imminent wave front from the supernova of Betelgeuse.

The red light of purification will emanate from this newly visible star, dramatically changing the color of our skies to a flaming red. This is referred to in the ancient wisdom as the Red Dawn of the forever cycle, where the consciousness and DNA of every living being will be transformed through the influence of the red light. Dr. Robert Ghostwolf and the Hopi Elders retell sacred traditional prophecies of the coming of the new world as an opportunity for the advancement of consciousness:

> The return of the Blue Star Katchina who is also known as Nan ga sohu will be the alarm clock that tells us of the new day and new way of life, a new world that is coming. This is where the changes will begin. They will start as fires that burn within us, and we will burn up with desires and conflict if we do not remember the original teachings, and return to the peaceful way of life.
>
> Not far behind the twins will come the Purifier, the Red Katchina, who will bring the Day of Purification. On this day the Earth, her creatures and all life as we know it will change forever. There will be messengers that will precede this coming of the Purifier. They will leave messages to those on Earth who remember the old ways.
>
> The messages will be found written in the living stone, through the sacred grains, and even the waters (Crop Circles have been found in ice). From the Purifier will issue forth a great Red Light. All things will change in their manner of being. Every living thing will be offered the opportunity to change from the largest to the smallest thing.

> Those who return to the ways given to us in the original teachings, and live a natural way of life will not be touched by the coming of the Purifier. They will survive and build the new world. Only in the ancient teachings will the ability to understand the messages be found.
>
> It is important to understand that these messages will be found upon every living thing, even within our bodies, even within a drop of our blood. All life forms will receive the messages from the twins... those that fly, the plants, even the rabbit. The appearance of the twins begins a period of seven years and will be our final opportunity to change our ways. Everything we experience is all a matter of choice...
>
> When the Purifier comes we will see him first as a small Red Star which will come very close and sit in our heavens watching us. Watching us to see how well we have remembered the sacred teachings. This Purifier will show us many miraculous signs in our heavens. In this way we will know Creator is not a dream. Even those who do not feel their connection to spirit will see the face of creator across the sky. Things unseen will be felt very strongly.

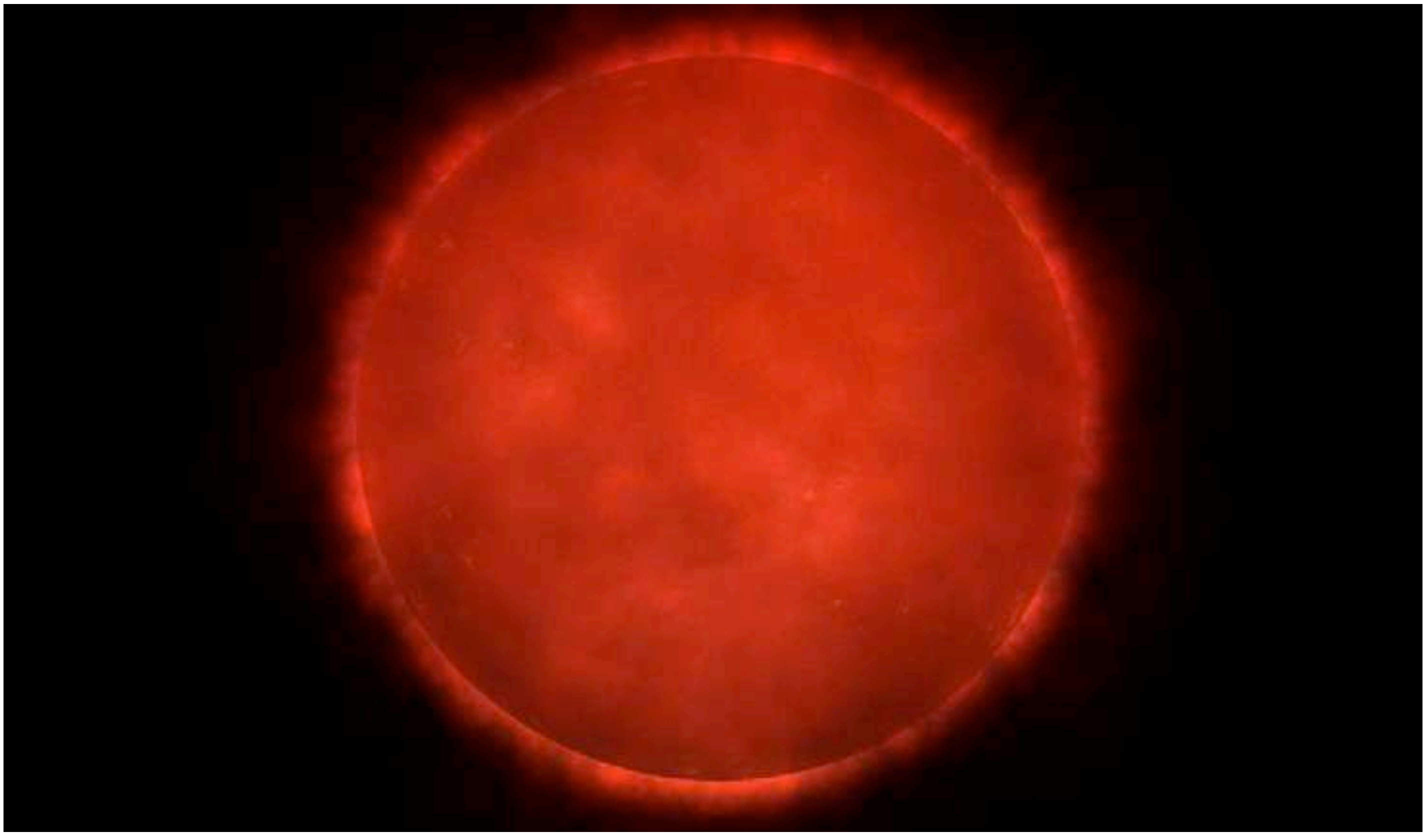

> Many things will begin to occur that will not make sense, for reality will be shifting back in and out of the dream state. There will be many doorways to the lower world that will open at this time. Things long forgotten will come back to remind us of our past creations. All living things will want to be present for this day when time ends, and we enter the forever cycle of the Fifth World.
>
> We will receive many warnings allowing us to change our ways from below the Earth as well as above. Then, one morning in a moment, we will awaken to the Red Dawn. The sky will be the color of blood, many things will then begin to happen that right now we are not sure of their exact nature. For much of reality will not be as it is now.[45]

Hopi culture foretells invisible acoustic changes in the environment and in the hearts of human beings, as “things unseen will be strongly felt.” The Elders inform us “the messages will be found written in the living stone, through the sacred grains, and even the waters (Crop Circles have been found in ice).” For decades now, conscious plasma orbs have been forming geometric messages in ice and grain fields, providing coherent mathematical formulas. The most complex quadratic functions were rendered in England, where the Mandelbrot Set [$z_{n+1} = z_n^2 + c$] appeared in Ickleton in 1991.[46] Subsequent agro-glyphs formed at Lurkley Hill in 2006[47] and Wayland Smithy in 2008[48] presented a variation on the fractal formula: [$z_{n+1} = z_n^2$]. Remarkably, this formula exactly describes unusual IR satellite data from 2001.

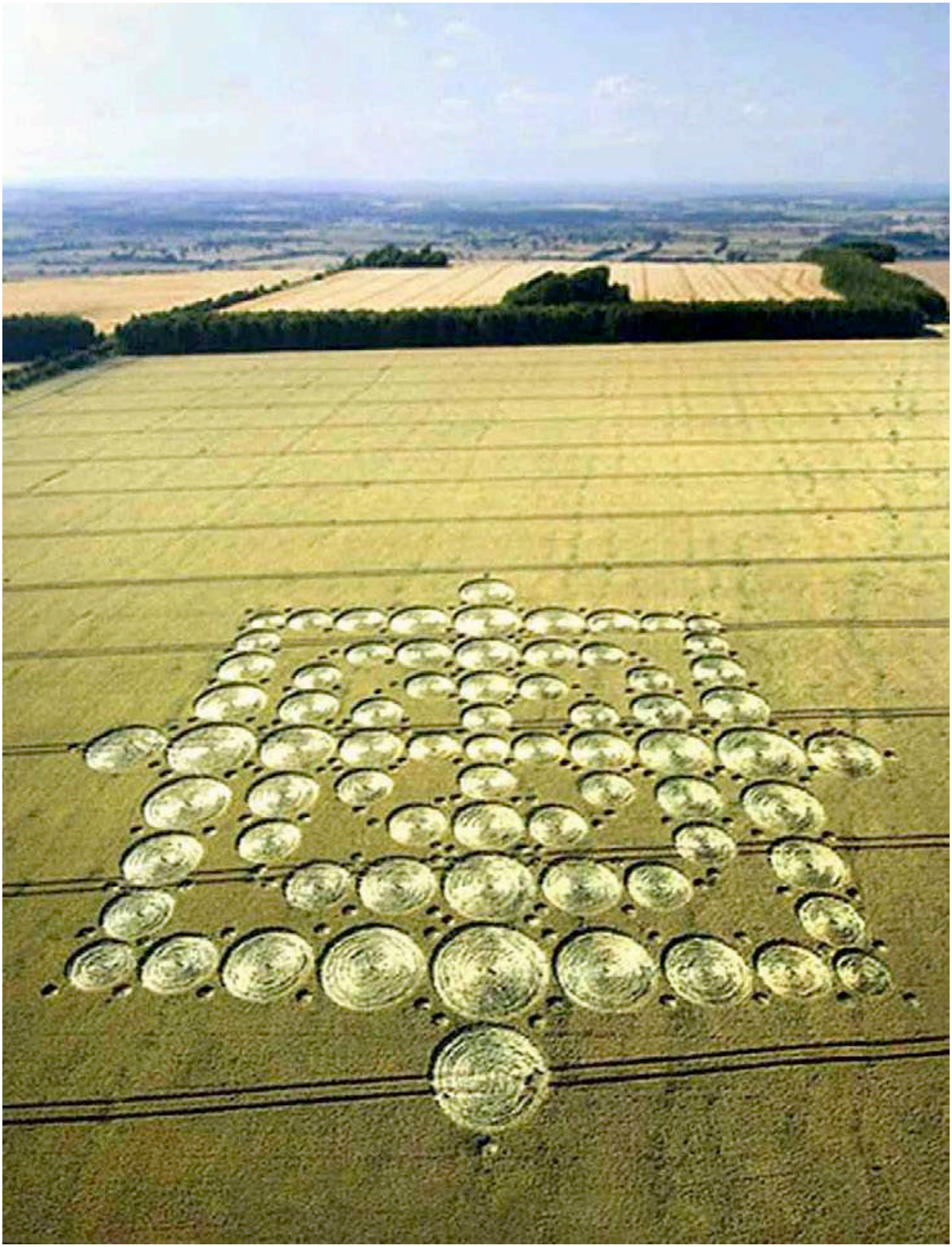

The sacred grains of England were also impressed with the value of Pi in 2008, in the form of a ratcheted spiral that accurately renders the first 10 digits (3.141592654…).[49] The conscious plasma ball entities that form these stunning messages are offering advanced hyper-dimensional mathematics to human beings at this time in our spiritual development to prepare our consciousness for multi-dimensional awareness. The agroglyphs provide us with a map of the nodal points of infrasound standing wave patterns now being stimulated by increasing solar activity. The circle-makers have even provided a specific and unmistakable *calendar date* in a truly massive formation in Avebury, England in 2008, showing a precise configuration of all of the planets of the solar system – *confirming the T'zolkin calendar of the Maya:*

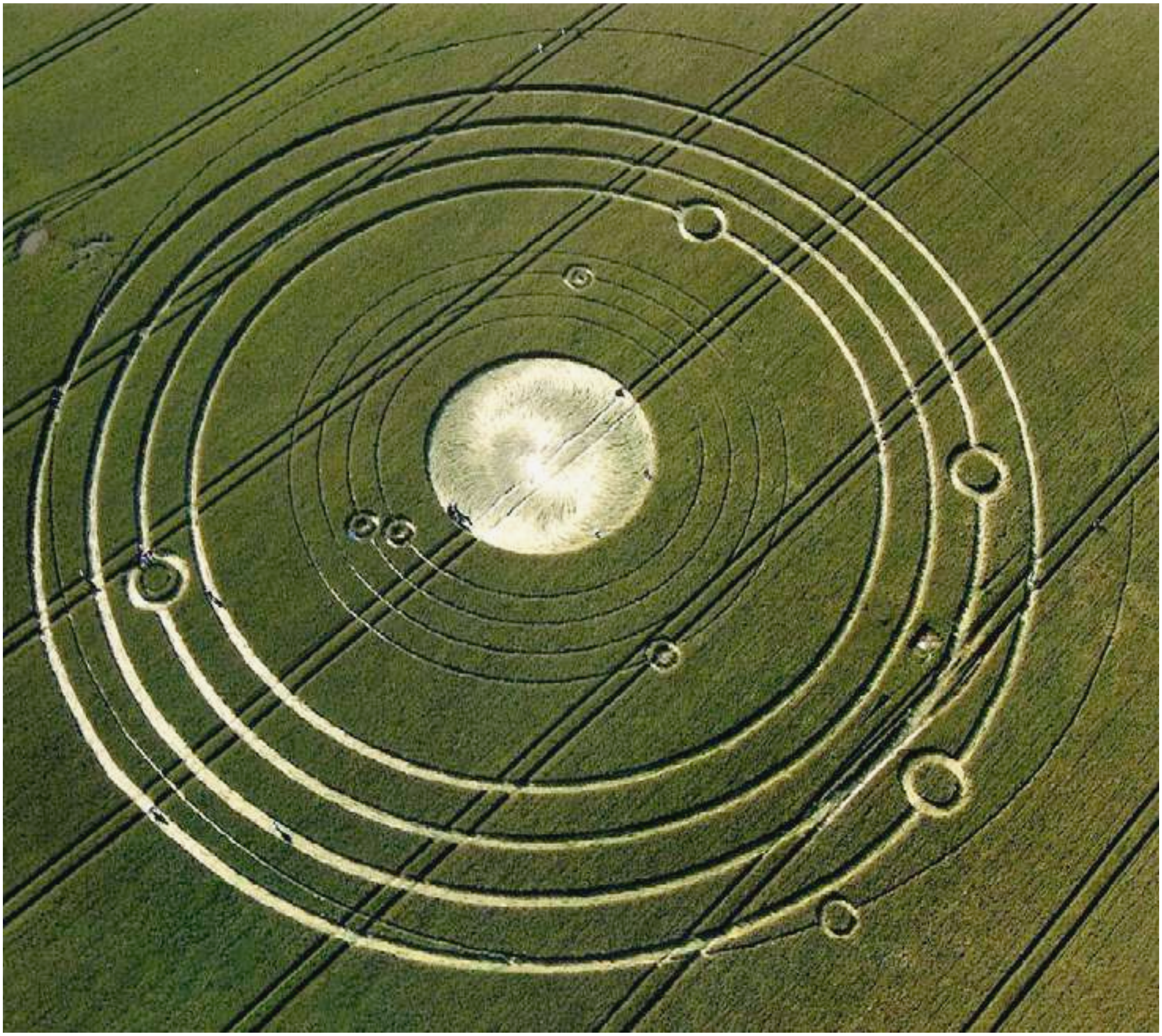

Avebury Manor of July 15, 2008 reminds us that the Mayan Long Count calendar will end on December 23, 2012: by showing precise orbital locations for all nine planets of our solar system on that date, plus an accurate elliptical orbit for Pluto.

The 5125-year Mayan Long count calendar will end on December 22-23, 2012. Given the long period of time which it represents, all nine planets of our solar system need to be shown in order to mark its end uniquely. Such a clever astronomical diagram was shown at Avebury Manor on July 15, 2008. All nine planets appear there precisely as they will be located in space four and one-half years from now on December 23, 2012...[50]

The indigenous Elders tell us that the messages of change will be received inside all living things, and each will be given an opportunity to change. Russian researchers have used infrasound and light waves to induce DNA transmutation, by which one species is genetically transmuted into another.[51] This discovery provides a new framework for quantum leaps in evolution, whereby quantum shifts from high to low levels of planetary infrasound resonance regulate the expression of DNA in Earth's many lifeforms.

New animal species are being encountered that are a sign of planet-wide genetic transmutations now occurring in hotspots of infrasound humming. An unusual new species now recorded often is a mostly hairless variety of purple/grey-skinned canine dubbed the 'hyote' (hyena/coyote, juvenile seen alive in captivity above), photographed and videotaped in Minnesota,[52] Maryland, Texas and Saudi Arabia. Sightings of various giant species of 'thunderbirds' have also been increasing around the world as species exhibit spontaneous DNA changes.

Another example of a new species now emerging at hotspots throughout the world is a kind of gracile fox that displays a long neck and large ears, with a tan coat that can be subtly spotted on the backside. The animal's fine facial features, combined with the extremely elongated limbs and tail, give it a deer-like appearance. This 'deerfox' has been photographed and observed in the US in England (pregnant female and two juveniles seen in separate photographs, above).[53]

These same infrasound-related DNA changes are expressed in extraordinary human beings who possess profound cosmic knowledge of their pristine origins and the unlimited holographic capabilities enabled by such knowledge. A shining example of this truth is a young woman from the taiga forests of Siberia named Anastasia who has mastered ayurveda – *the art of living.* Having inherited the deepest of cosmic wisdom from her Vedruss forefathers, Anastasia provides us with the spiritual information that will germinate the seeds of human divinity now being planted in the fields of the Earth. The voices of the indigenous Elders of the world speak as one, revealing ancient Vedic knowledge of the transdimensional nature of human consciousness to guide our path of ascent into the red light of Oneness.

The Sanskrit wisdom traditions of Russia have recently revealed a profound knowledge of nature that forms the basis for the use of heartbeat resonance in monumental psychoacoustics. Timeless Vedic wisdom understands the seed as physician and the tree as an organic antenna - a collector of cosmic energy for the single purpose of powering the human heart and spirit. The Vedruss Elders of Siberia reveal their process of DNA imprinting of seeds for the creation of a resonant field of heartbeat synchronization.[54] By heartbeat entrainment through seed imprinting, the family of Anastasia has created a resonant field of infrasound that amplifies their heartbeat to enable astral travel through the Akashic field of planetary infrasound standing waves *on the heartbeat frequency of love*:

> “When someone is in a state of love they emit a radiant energy. It takes but a split second for it to reflect off the planets floating overhead and come back to Earth to give life to everything that breathes… The Sun is one of those planets, and it reflects but a tiny fraction of such radiance. Only bright rays can travel into Space from Man on the Earth. And only beneficial rays can be reflected from Space back to Earth."
>
> "Under the influence of malicious feelings Man can emit only dark rays. These dark rays cannot rise but must fall into the depths of the Earth. Bouncing off its core, they return to the surface in the form of volcanic eruptions, earthquakes, wars, etc. The culminating achievement of these dark rays is their direct effect on the Man originating them, invariably exacerbating the man's own malicious feelings." [55]

Vedruss knowledge echoes the call of the Hopi Elders, who foretell of the great increase in the energies moving through human beings, whose inner emotional states will become amplified in the light of the Red Purifier to manifest as fires that will burn within us. *Malevolent feelings will consume those who harbor them, while the rays of love will also become equally amplified to generate healing forces in human beings who harbor a deep Creational love by living as one being in heartbeat resonance.*

Hopi Elders also foretell of portals to the lower worlds that will be opening at this time to remind us of our past creations. Referred to as *‘sipapu’*, these lower realms are described in Hopi wisdom traditions as having been used to preserve the human race during previous cataclysmic events on Earth. Yaqui Elder Don Juan elaborated on the secret knowledge possessed by ancient Nahuatl seers pertaining to the motive properties of fire and water for bodily transport to fourth-density realms of inorganic life, in artificially generated HHO plasma chambers both above and below the Earth’s surface:

> He continued then with his elucidation of the practices of the old seers. Another of their great findings had to do with the category of secret knowledge: fire and water. They discovered that flames have the most peculiar quality; they can transport man bodily, just as water does... They considered flames and fluidity to be higher, magical properties, and they used them as a means for bodily transportation to the realm of inorganic life.[56]
>
> Don Juan said that the old seers observed that the wetness of water only dampens or soaks, but the fluidity of water moves. It runs, they surmised, in search of other levels underneath us. They believed that water had been given to us not only for life, but also as a link, a road to the other levels below... to be transported bodily by the fluidity of water anywhere between this level of ours and the other seven levels below.[57]
>
> The knowledge of the below had to do with fog, water of underground springs, swamps, lightening bolts, earthquakes, the night, moonlight and the moon.[58]
>
> “[Some men of knowledge] endeavored to use *seeing* positively and to teach it to their fellow men. I’m convinced that under their direction, the populations of entire cities went into other worlds and never came back.” [59]

“The old seers *saw* that the Earth has a cocoon,” he went on. “They *saw* that there was a ball encasing the Earth, a luminous cocoon that entraps the Eagle’s emanations. The Earth is a gigantic sentient being subjected to the same forces we are.” [60]

”[Don Juan stated that]… one of the most worthwhile findings of the ancient seers... was the discovery that organic life is not the only form of life present on this Earth... For seers, to be alive means to be aware... For the average man, to be aware means to be an organism. This is where seers are different. For them, to be aware means that the emanations that cause awareness are encased inside a receptacle.[61]

The circular aspect... is what maintains life and awareness, fulfillment and purpose... “The reason that it’s called the circular force is that it comes in rings, thread-like hoops of iridescence... it strikes all living beings ceaselessly... to give them strength, direction, awareness; to give them life... It’s one indivisible force that fits all living beings, organic and inorganic.” [62]

“They say that perception is a condition of alignment... Alignment is what allows awareness to be cultivated by every living creature...” When seers *see* perception, they witness that the luminosity of the... emanations outside those creatures’ cocoons brightens the luminosity inside their cocoons. The outside luminosity attracts the inside one; it traps it, so to speak, and fixes it. That fixation is the awareness of every specific being.[63]

He explained that what he called the key to everything was the first-hand knowledge that the Earth is a sentient being and as such can give warriors a tremendous boost; it is an impulse that comes from the awareness of the Earth itself at the instant in which the emanations inside the warriors’ cocoons are aligned with the appropriate emanations inside the Earth’s cocoon. Since both the Earth and man are sentient beings, their emanations coincide, or rather, the Earth has all the emanations that are present in man and all the emanations present in all sentient beings, organic or inorganic for that matter. When a moment of alignment takes place, sentient beings use that alignment in a limited way and perceive their world. Warriors can use that alignment either to perceive, like everyone else, or as a boost that allows them to enter unimaginable worlds.[64]

He explained that the Earth’s boost is the force of alignment of only the amber emanations. It is a boost that heightens awareness to unthinkable degrees. To the new seers it is a blast of unlimited consciousness, which they call total freedom.[65]

Don Juan restated over and over that the portion of the emanations inside man’s cocoon is in there only for awareness, and that awareness is matching that portion of the emanations with the same portion of emanations at large. They are called emanations at large because they are immense... inside man’s cocoon the unknown is the emanations untouched by awareness. When the glow of awareness touches them, they become active and can be aligned with the corresponding emanations at large. Once that happens the unknown is perceived and becomes the known.[66]

The Eagle, that power that governs the destiny of all living things, reflects equally and at once all those living things… [and] although it is not moved by the circumstances of any living thing, has granted a gift to each of those beings. In its own way and right, any one of them, it if so desires, has the power to keep the flame of awareness, the power to disobey the summons to die and be consumed. Every living thing has been granted the power, if it so desires, to seek an opening to freedom and to go through it. It is evident to the seer who sees the opening, and to the creatures that go through it, that the Eagle has granted that gift in order to perpetuate awareness.

For the purpose of guiding living beings to that opening, the Eagle created the Nagual. The Nagual is a double being to whom the rule has been revealed. Whether it be in the form of a human being, an animal, a plant, or anything else that lives, the Nagual by virtue of its doubleness is drawn to seek that hidden passageway.

The Nagual comes in pairs, male and female. A double man and a double woman become the Nagual only after the rule has been told to each of them, and each of them has understood it and accepted it in full… To the eye of the seer, a Nagual man or a Nagual woman appears as a luminous egg with four compartments. Unlike the average human being, who has two sides only, a left and a right, the Nagual has the left side divided into two long sections, and a right side equally divided in two.

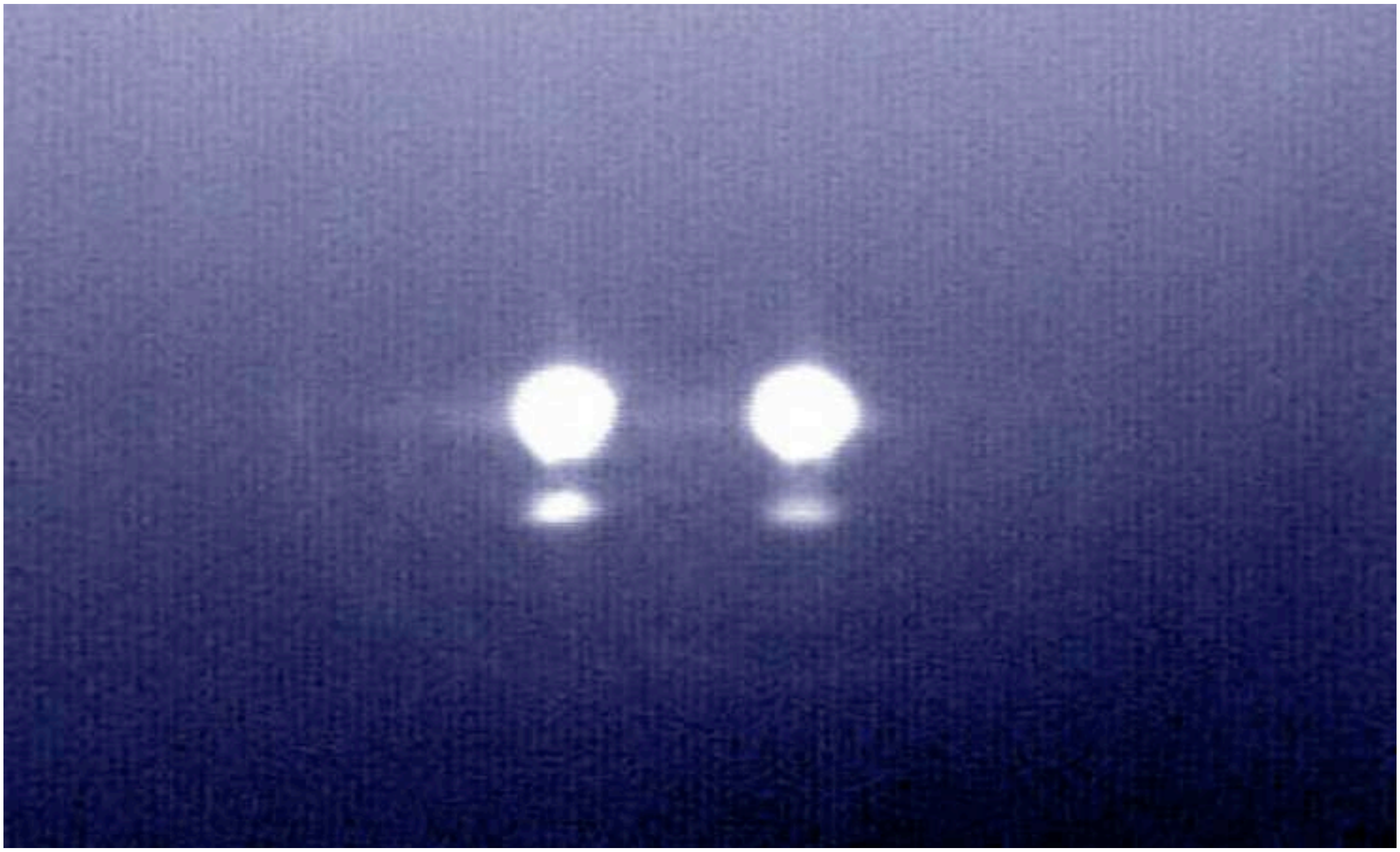

The Eagle created the first Nagual man and first Nagual woman as seers and immediately put them in the world to see. It provided them with four female warriors who were stalkers, three male warriors, and one male courier, whom they were to nourish, enhance and lead to freedom… In order to make sure that the first Nagual man would lead his party to freedom and not deviate from that path or become corrupted, the Eagle took the Nagual woman to the other world to serve as a beacon, guiding the party to the opening.

The Nagual and his warriors were then commanded to forget. They were plunged into darkness and were given new tasks: the task of remembering themselves, and the task of remembering the Eagle… The command to forget was so great that everyone was separated. They did not remember who they were. The Eagle intended that if they were capable of remembering themselves again, they would find the totality of themselves. Only then would they have the strength and forbearance necessary to seek and face their definitive journey.

Their last task, after they had regained the totality of themselves, was to get a new pair of double beings and transform them into a new Nagual man and a new Nagual woman by virtue of revealing the rule to them…

Don Juan explained that the rule was not a tale, and that to cross over to freedom did not mean eternal life as eternity is commonly understood –that is, as living forever. What the rule stated was that one could keep the awareness which is ordinarily relinquished at the moment of dying… [A]t the moment of crossing, one enters into the third attention, and the body in its entirety is kindled with knowledge. Every cell at once becomes aware of itself, and also aware of the totality of the body… Therefore the crux of the warrior's struggle was not so much to realize that the crossing over stated in the rule meant crossing to the third attention, but rather to conceive that there exists such an awareness at all.[67]

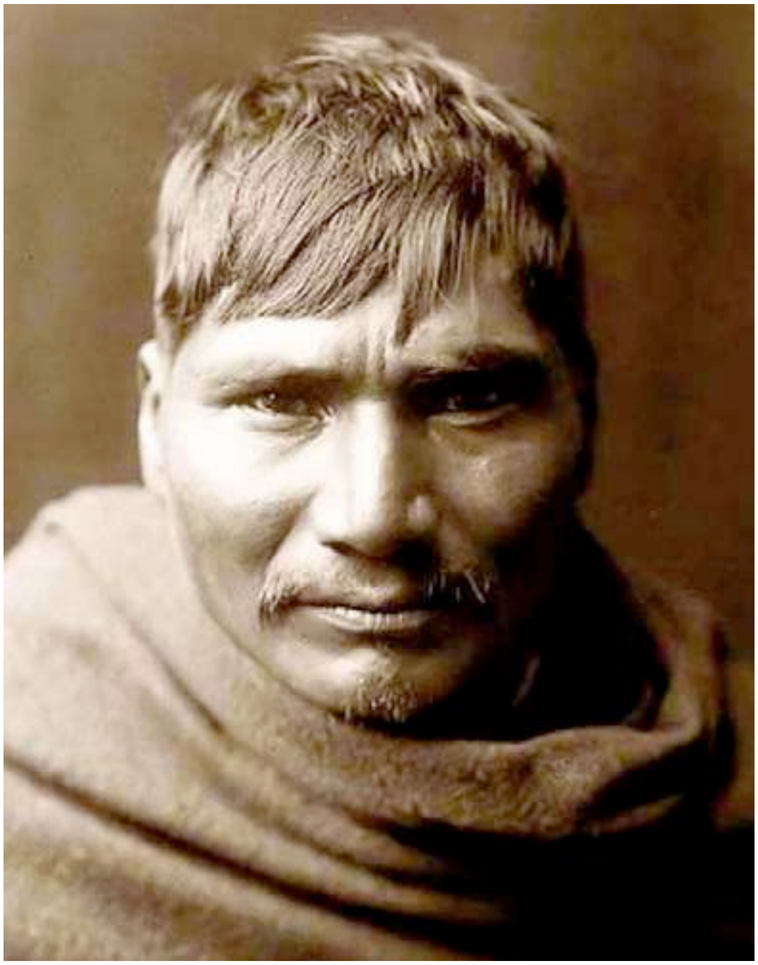

“Seers say that there are three types of attention... they are rather three levels of attainment. They are the first, second, and third attention, each of them an independent domain, complete in itself.” He explained that the first attention in man is the animal awareness... everything that one can think about is part of the first attention... [68]

Don Juan explained that in order for our first attention to bring into focus the world that we perceive, it has to emphasize certain emanations selected from the narrow band of emanations where man’s awareness is located. The discarded emanations are still within our reach but remain dormant, unknown to us for the duration of our lives. The new seers call the emphasized emanations the right side, normal awareness, the tonal, this world, the known, the first attention. The average man calls it reality, rationality, common sense.

The emphasized emanations compose a large portion of man’s band of awareness, but a very small piece of the total spectrum of emanations present inside the cocoon of man. The disregarded emanations within man’s band are thought of as a sort of preamble to the unknown, the unknown proper consisting of the bulk of emanations which are not part of the human band and which are never emphasized. Seers call them the left-side awareness, the nagual, the other world, the unknown, the second attention.[69]

“[To be able]… to utilize those unused emanations, one needs uncommon, elaborate tactics that require supreme discipline and concentration... the art of *dreaming*... the concentration needed to be aware that one is having a dream is the forerunner of the second attention. That concentration is a form of consciousness that is not in the same category as the consciousness needed to deal with the daily world.

“The new seers... let the mastery of awareness develop to its natural end, which is to extend the glow of awareness beyond the bounds of the luminous cocoon in one single stroke. The third attention is attained when the glow of awareness turns into the fire from within: a glow that kindles not one band at a time but all the... emanations inside man’s cocoon.”

Don Juan expressed his awe for the new seers’ deliberate effort to attain the third attention while they are alive and conscious of their individuality... “The supreme accomplishment of human beings,” he said, “is to attain that level of attention while retaining the life-force, without becoming a disembodied awareness moving like a flicker of light...”[70]

[He then explained that]...the glow produced by a state of total awareness... is *seen* as a burst of incandescence in the entire luminous egg. It is an explosion of light of such a magnitude that the boundaries of the shell are diffused and the inside emanations extend themselves beyond anything imaginable.

“Are those special cases, don Juan?”

“Certainly. They happen only to seers. No other living men or any other living creatures brighten up like that. Seers who deliberately attain total awareness are a sight to behold. That is the moment when they burn from within. The fire from within consumes them. And in full awareness they fuse themselves to the emanations at large, and glide into eternity.

[As seers] our case is a bit different, because we are at the end of our trail. We are not seeking anything... we go from day to day doing nothing. We are waiting. I will not tire of repeating this: we know that we are waiting and we know what we are waiting for. We are waiting for freedom!” [71]

A Great Cycle of approximately 309,888 years (12 x 12 Precession cycles of 2,152 years) nears completion with the returning Dark Star on its near 36,650 year elliptical cycle (2,152 x 17 = 36,584), as red supergiant Betelgeuse shrinks in collapse, poised to illuminate all that is hidden within the shadows of the human heart and our entire solar system. Prophetically foretold signs are to be seen in their dramatic fulfillment in all parts of the world. Infrasound resonance surges as intense heartbeat booming and ignites piezoelectric fires at nodal points in all parts of the world. The many Oort comet impacts and aerial fireballs witnessed all over the world in 2010 will continue to increase in frequency for several years as a constant threat to the delicate biosphere of Earth in these revelatory times preceding the supernova blast wave of December 22, 2012, and until 2028.

The initiation of celestial waters at sacred pyramid sites illuminates the interior chambers of the pyramids with the heavenly light of the Creator, and illuminates the DNA of the initiate with the red and blue light of gold and silver nanoparticles within every cell. Intense infrasound booming will align all initiates into heart/mind synchronization inducing the synesthetic experience of collective telepathic unity, *kindling the fire from within to attain fourth-density awareness.*

The time has arrived for humanity to escape the technological perils of his own misguided making and to reconvene at our ancient pyramids and sacred sites to ascend into the red light of purification that will enhance all harmonious living beings touching the Earth during the Red Dawn of the Forever Cycle.

References

---

[1] Jenkins JM, Matz M (2004) *‘Pyramid of Fire: Spiritual Ascent at the End of Time’* Bear & Company, pp. 81-88

[2] Men H, Yaxk’in AJ (2001) ‘Mayan/Tibetan Prophecy’ online · http://www.lightworkarticles.info/MayanProphecy/

[3] Sitchin Z (1999) *‘The Wars of Gods and Men: Book III of the Earth Chronicles’* Harper, online · http://www.bibliotecapleyades.net/sitchin/sitchinbooks03_03.htm

[4] Putney A (2007) ‘Sacred Ayurvedic Waters’ *Human Resonance*, online · http://www.humanresonance.org/water.html

[5] Putney A (2008) ‘Ø’ *Human Resonance Productions*, pp. 116-162

[6] Quetza Sha GA (1996) *‘The Manuscript of the Crystal Golden Water of Tlacote’ Human Resonance,* online · http://www.humanresonance.org/tlacote.html

[7] Courtland R (2009) ‘Betelgeuse the Incredible Shrinking Star’ *New Scientist*, June 10, online · http://www.newscientist.com/article/dn17282-betelgeuse-the-incredible-shrinking-star.html

[8] Castaneda C (1991) ‘*The Fire From Within*’ Washington Square Press, pp. 121

[9] Brown T (1999) ‘Night of the Red Sky: The Prophetic Vision of 'Grandfather' *Nexus Magazine (7:1)*, Dec-Jan

[10] Putney A (2008) ‘Satellite Records Atmospheric Anomaly’ online · http://www.humanresonance.org/sphere.html

[11] Schneider J (2007) ‘Mysterious Noise Annoys Anderson, Indiana Couple’ *The Herald Bulletin*, August 29, online · http://www.humanresonance.org/anderson.html

[12] Murray K (2008) ‘Magic Valley is Shaking – No Earthquakes Recorded’ *KTVB 7 News*, March 21, online · http://www.humanresonance.org/kimberley.html

[13] -- (2004) ‘Mysterious ‘Booms’ Continue With Puzzling Regularity in Richmond’ *WWBT News*, November 19, online · http://www.humanresonance.org/richmond.html

[14] -- (2008) ‘Mysterious Booms Rock Cape Fear Region’ *Star News*, February 8, online · http://www.humanresonance.org/wilmington.html

[15] -- (2007) ‘Rumbling Wakes Up Dozens of Knox County Residents’ *WLTV News*, May 14, online · http://www.humanresonance.org/knoxville.html

[16] -- (2006) ‘Mystery Booms Rattle Mobile Residents’ *WPMI News*, January 19, online · http://www.humanresonance.org/mobile.html

[17] Miller JL (2007) ‘Wall-Shaking, Window-Rattling Booms Heard in Missouri Still a Mystery’ *Lake Sun Leader*, July 24, online · http://www.humanresonance.org/knobnoster.html

[18] -- (2007) ‘Plane Windshields Crack at Denver Int. Airport’ *Associated Press*, February 16, online · http://www.humanresonance.org/denver.html

[19] Callis T (2009) ‘What’s Causing the Mysterious Booming Sounds? No One Seems to Know’ *Peninsula Daily News*, July 31, online · http://www.humanresonance.org/seattle.html

[20] -- (2007) ‘Military Spokesman Responds to Report of Sonic Booms, Shaking in California’ *San Luis Obispo,* July 18, online · http://www.humanresonance.org/arroyogrande.html

[21] Albrecht L (2005) ‘Atwater Noise A Mystery’ *Merced Sun-Star*, November 30, online · http://www.humanresonance.org/atwater.html

[22] Brit RR (2003) ‘The Great Strom: Solar Tempest of 1859 Revealed’ *Space.com News*, October 27, online · http://www.space.com/scienceastronomy/mystery_monday_031027.html

[23] -- (2007) ‘Thousands Flee Fires in Spain’s Canary Islands’ *Human Resonance*, July 31, online · http://www.humanresonance.org/canary.html

[24] Olawale T (2008) ‘Nigeria: Babura – Where Mystery Fires Ravaged a Community’ *This Day*, March 11, online · http://www.humanresonance.org/babura.html

[25] -- (2008) ‘Mystery Fire Guts Hut’ *Midlands Bureau*, April 13, online · http://www.humanresonance.org/lalapansi.html

[26] Hlungwani V (2007) ‘Mystery Fires Hit Village’ *Sowetan News*, online · http://www.humanresonance.org/mapuve.html

[27] -- (2003) ‘Community Gripped By Mystery Beneath the Earth’ *SABC News*, online · http://www.humanresonance.org/bodibe.html

[28] -- (2008) ‘Spontaneous Fires Ignite in 7 Houses in Landovica, Kosovo’ *Human Resonance*, August 5, online · http://www.humanresonance.org/landovica.html

[29] Bertani C (2007) ‘Mysterious Fires – Pyromaniacs or Paranoids?’ *Human Resonance*, July 31, online · http://www.humanresonance.org/peschici.html

[30] -- (1990) ‘The Berici Hills Mystery’ *Sunday Express*, March 11, online · http://www.humanresonance.org/berici.html

[31] Ferrari T (2004) ‘Canneto di Caronia, This is Not Your Average Ghost Town’ *Human Resonance*, April 4, online · http://www.humanresonance.org/messina.html

[32] Becatoros E (2007) ‘Helicopter Sent to Save Trapped Greeks’ *Associated Press*, August 27, online · http://www.humanresonance.org/greece.html

---

[33] -- (2008) 'Mysterious Fire Engulfs Kishtwar Village' *News Agency of Kashmir*, October 12, online · http://www.humanresonance.org/kishtwar.html

[34] Ghosh R (2004) 'Villagers in Madhya Pradesh Face Mysterious Enemy' *Human Resonance*, March 21, online · http://www.humanresonance.org/ratria.html

[35] Ganzon T (2009) 'Phillipines: Batangas Town Residents Worry Over 'Mysterious' Fires' *ABS-CBN News*, January 8, online · http://www.humanresonance.org/santo_tomas.html

[36] Jones N (2001) 'Fires and Gas Explosions in Seattle' *Human Resonance*, July 27, online · http://www.humanresonance.org/seattle.html

[37] Folsom B (2008) 'Unexplained Ground Heat Burns Boy's Feet' *KOAA.com*, June 3, online · http://www.humanresonance.org/co_springs.html

[38] -- (2008) 'California: Police Investigate Mystery Car Fire' *The Examiner*, March 28, online · http://www.humanresonance.org/san_mateo.html

[39] Lin J (2008) 'Ventura County Hot Spot Puzzles Experts' *LA Times*, August 5, online · http://www.humanresonance.org/padres.html

[40] -- (2007) 'Cell Phone Catches Fire, Burns Man Over Half His Body' *Associated Press*, January 16, online · http://www.humanresonance.org/vallejo.html

[41] -- (2009) 'Ball Lightning Hits 100 Homes' *Dutch News*, July 21, online · http://www.humanresonance.org/soesterberg.html

[42] Dunn M (2004) 'Odd Space Noise Heard Again' *Associated Press*, April 2, online · http://www.humanresonance.org/space_station.html

[43] Cruttenden W (2009) 'Introduction: Binary Companion Theory' *Binary Research Institute*, online · http://www.binaryresearchinstitute.org/bri/research/introduction/theory.shtml

[44] -- (2004) 'Sedna and Sol B?' *Sol Station News*, March 15, online · http://www.humanresonance.org/sedna.html

[45] Ghostwolf R (1994) 'Hopi Prophecy Fulfilled' *Last Cry: Native American Prophecies and Tales of the End Times Magazine*, online · http://www/wolflodge.org/bluestar/bluestar.htm

[46] Putney A (2008) 'Mandala Formation in Ickleton, England' *Human Resonance*, online · http://www.humanresonance.org/ickleton.html

[47] Putney A (2008) 'Mandala Formation at Lurkley Hill in Lockeridge, England' *Human Resonance*, online · http://www.humanresonance.org/lurkley.html

[48] Putney A (2008) 'Mandala Formation in Wayland Smithy, England' *Human Resonance*, July 27, online · http://www.humanresonance.org/wayland08.html

[49] Putney A (2008) 'Mandala Formation Near Barbary Castle, England' *Human Resonance*, June 1, online · http://www.humanresonance.org/barbury08.html

[50] Collie R (2008) 'Crop Circle at Avebury Manor Reported July 15th, 2008' *Human Resonance*, July 15, online · http://www.humanresonance.org/avebury08.html

[51] Gariaev P, Grigor'ev K, Vasil'ev A, Poponin V, Shcheglov V (1999) 'Investigation of the Fluctuation Dynamics of DNA Solutions by Laser Correlation Spectroscopy' *Bulletin of the Lebedev Physics Institute* 11-12, pp. 23-30

[52] Coleman L (2009) 'Hairless Mystery Animal Photographed' *Cryptomundo*, July 10, online · http://www.cryptomundo.com/cryptozoo-news/hairless09/

[53] Moulton Howe L (2007) 'Mystery Animal Photographed in England' *Earthfiles*, online · http://www.earthfiles.com

[54] Megre V (2003) *'Anastasia'* Ringing Cedars Press, pp. 77-81

[55] Ibid, p. 7

[56] Castaneda, op cit, '*The Fire From Within*', p. 92

[57] Ibid, pp. 98-99

[58] Ibid, p. 93

[59] Ibid, pp. 18-19

[60] Ibid, p. 204

[61] Ibid, p. 90

[62] Ibid, p. 228

[63] Ibid, p. 61

[64] Ibid, p. 206

[65] Ibid, p. 247

[66] Ibid, p. 207

[67] Castaneda C (1982) *'The Eagle's Gift'* Simon & Schuster, pp.173-179

[68] Castaneda, op cit, '*The Fire From Within*' pp. 76-77

[69] Ibid, p. 116

[70] Ibid, pp. 77-78

[71] Ibid, pp. 120-121

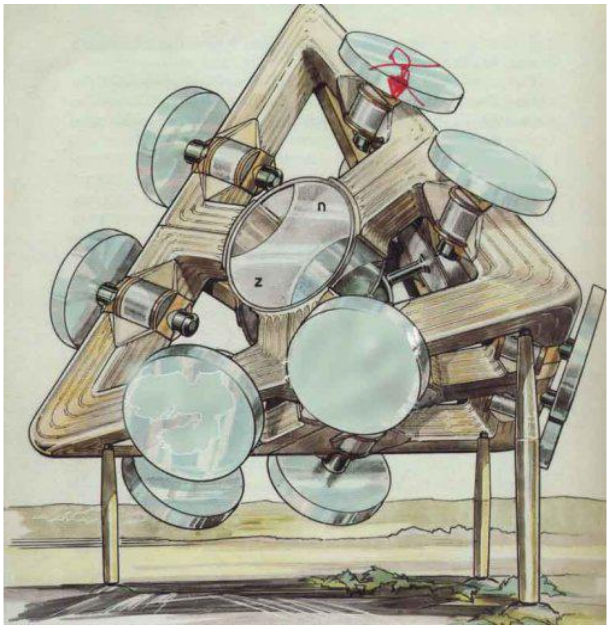

Technical information given from the IARGA contactteam.

their describtions on the matter of
TIME

- AND MORE

picture is described later

12. *The analog carrierfield model, an instruction model.*

The cosmic carrier field of all space and matter performs a sixfold function, of which some have already been discussed in gereral way.

1. The carrier function of an infinitely strong forcefield blocked in itself contains the fluidum conducting all waves without dissipation and maintained to eternity.
2. The orientation of the time-axes of the universes with respect to nothing provides the structure of the carrier-field of three dimensions, the bounded Love-will, creating our life and existence.
3. The time synchronisation function locks all universes in the iron grip of the law of cosmic counter-balance from which no atom can escape.
4. The feedback principle of time to the total mass of the universe is the function of the coupling-field by counterbalance (and interference) of the time-bound waves.
5. The function of the timeless primeval atom-field, the atom-principle, as the transmitter program containing the code of matter of which all atoms are remote manifestations.
6. The function as bearer of the absolute ALL-knowledge and goal consciousness, manifesting itself as the laws of nature and as the intelligent goal of life: the spirit of God.

The last function concerns the existential aspects, which have been discussed in the previous section of this book.

The first analog model is a mechanical instruction model explaining the cosmic carrier field by means of symbols. This model was built on another life-supporting planet on a high mountain and was many meters high. It was the gathering point for contemplation of that intelligent race for devotion to its Creator.

It was the symbol by which the majority of the race acknowledged the goal of Creation, simultaneously the sign that the cosmic isolation had been lifted. The other function of the carrier-field will be discussed with the help of this mechanical instruction model.

The analog model consisted mainly of the six edges of a huge equilateral triangular pyramid built out of solid soft-iron bars. The many metres thick edges of the pyramid contain an impressive number of iron bars parallel to the direction of the edges and cast in a kind of plastic resin following a process similar to that of casting concrete(betong). So a mainly iron-frame was formed especially suitable for conduction of magnetic pulses of high power without much loss.

So far the carrying fuction of the cosmic time-field, which primarily maintains the timeless waves is symbolized in the model. The magnetic pulses in the edges of the pyramid run backwards and forwards like the timeless waves in reality.

With respect to the other edges of the pyramid, the magnetic pulses are phase-shifted. The problem is that up to now the forward propagation (forplantning) in time has been missed out. In fact we look for the equivalent of the two little wagons on the vibrating rail. The solution can be found in the symbol of the rotating flywheel, as *with the atom its rotation is continuous in time, it is in fact time propagation.* Therefore the iron frame of the pyramid is discontinued at two places in each of the six edges of the pyramid. At this discontinuation point a linear magnetic impulse motor is installed, on the axis of which a large flywheel is mounted.

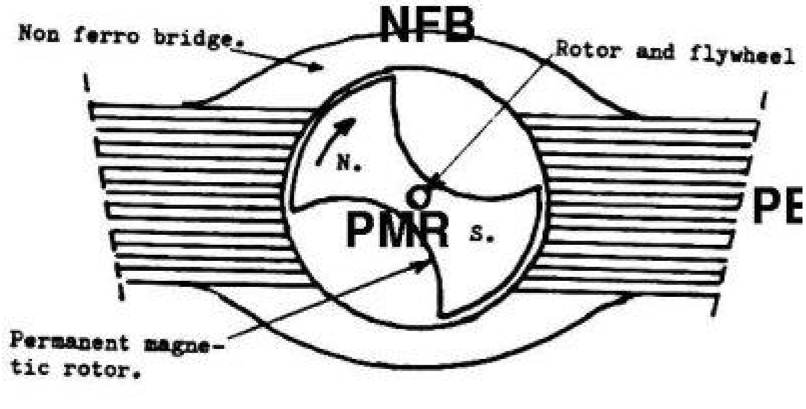

NBF=NonFerroBridge

PE=PyramidEdge and soft iron bars

PMR=PermanentMagneticRotor

The flywheel symbolizes not only the propagation of time, but also the inertia of matter and so the flywheel is in effect the total mass of the universe. This type of magnetic impulse motor is not self-starting, so at the inside on the flywheel axis a starter-motor is mounted, fitted solely for the purpose of starting. One rotating the synchronous motor maintains the exact number of revolutions at the same frequency of the magnetic alternation field. Thus in each of pyramid's edges two synchronous motors are mounted making twelve flywheels in total.

The two flywheels per axis rotate in opposite directions to symbolize the time propagation in opposite time directions. This type of linear impulse motor is chosen because it can be made to rotate in either direction.

The first analog model.

In one triangular plane of the pyramid the six synchronous motors in the iron frame demonstrate a rotating magnetic field, seen from the center of the plane. Each edge of the pyramid has the right phase-shift with respect to the other edges of the triangle.

The first model has four central rotating magnetic field generators in the middle of each triangular plane. (See figure of unfolded pyramid indicating the actual phase-shift).

These four synchronous (three phase) generators are coupled together by a gearbox and are driven by the main electric motor (not shown in the drawing). Consequently the four three-phase generators fulfill the synchronisation and constant phase-shifts for the flywheels of the pyramid-edges. The four separate rotating fields together, initate the complex three-dimensional rotation of the carrier field (or atom-field).

The disadvantage of the first model is that it is further away from reality, because the atomic field is not symboized as one three-demensional rotating field. However it shows better how to imagine the complicated rotation in three dimensions.

The second or the real analog model.

This mechanical model has only one central generator in the middle of the pyramid, instead of the four rotating field generators of the first model. All magnetic waves in the pyramid edges go straight to the pyramid's center. The rotor of the 3-dimensional rotating generator is like a yo-yo with spherical end-surfaces.

It requires a touch of genius to provide the drive-connection between the central motor and the spherical rotor, which has to be started up in a complex rotation. The six diametrically opposed pole-shoes from the middle of the pyramid's edges around the spherical rotor form in pairs, three axes of right angles - representing a Cartesian coordinate system. So it shows that the six time-axes of the edges of the pyramid can be converted into the three principal time-axes of the central sphere.

Together it imitates three-dimensional time, which causally creates the three-dimensional space of our universe.

The principles of the first and second analog models are equivalent. One version can be converted into the other, for example, converting the four rotating fields to one three-dimensional rotation in the centre-mass of the pyramid.

The model can be started by switching on the thirteen drive-motors. As the flywheels reach velocity the twelve starter-motors are switched off and only the central motor drives the complete gigantic model. The essence is that all flywheels have exactly the same number of revolutions. They are all either synchronous motors or generators. Then the flywheels are marked and illuminated by a stroboscopic flash light. *Seen by the normal eye they do not appear to move showing the synchronous rotation and the constant phase-shifts between the flywheels.*

These mechanical models are meant to explain the unimaginable Creation out of the nothing by symbolic imitation of the carrier-field. However it is only a weak symbol of Creation. The carrier-field is the esoteric medium, a transmutation of the absolute spirit or conscience, it is the spiritual force field of infinite strength which is present everywhere in the universe.

The imitation of the carrier-field functions.

Nevertheless the analogy of reality is incredible. This will be shown in the next explanation of the five functions of the cosmic carrier-field.

The carrier-function

The foundation of the carrier-field is the infinite force-field created by interference quenching (avbrytende stopp) of the original infinite vibration. The mode initiates this force-field with the six bundles of soft-iron bars as the edges of the pyramid.

As in reality, the three-dimensional magnetic rotating field in our model is caused by the linear waves of each time-axis. *The timeless waves alternate backwards and forwards as in our model*. They drive the flywheels representing the masses of the universes, and their rotation is the propagation (spredning) in time.

The orientation of the time-axes with respect to the nothing.

The relative position of one time-axis in the apparent nothing plays an important role in the orientation of the universes in reality, determining the direction of each time-axis. Here the first model is shown representing the four three-phase synchronous generators in the planes of the pyramid. Only with decompostion of the three-dimensional rotation into four rotations can one show that the three-dimensional rotating field has left no remaining component in the nothing, while in contrast a two-dimensional rotating field can not be nullified, a remaining component will always exist in the nothing. This can be explained by considering a single plane of the pyramid with its six flywheels at the edges, the analog for a two-dimensional

carrier-field. The rotation impulse of the flywheels is compensated by the opposite rotations but the rotation impulse of the central three-phase generator is left over. It can be measured by the outside world (the nothing).

It requires four field three-phase generators (maintaining overall symmetry) to compensate for these rotation impulses.

With this the three-dimensionality of the carrier-field has been shown and the orientation of the time-axes with respect to the nothing is determined.

The time synchronization function.

Impressively the model imitates the time synchronization, simultaneously maintaining the cosmic counter-balance by which the universes continue to exist. The rotation of the two flywheels per axis is in opposite directions and they rotate with exactly the same speed. Thus it imitates the circular motion to the timeless waves which progresses exactly synchronous by their mutual interference coupling.

The twelve cosmic systems run without any dissipation, once started, the perpetuum mobile continues into eternity with the absolute accuracy of time-synchronisation. If one supposes the model to be without loss and without using the eternal driving motors, it will continue to run synchronously due to the inertia of the flywheels and the fact that the synchronous linear impulse motors can function as generators as well.

This symbolizes another property of the continuous circular motion of the timeless waves, because every wave somewhere in the pyramid circuit at each moment has its opposite value resulting in an overall zero at any moment.

This is the law of conservation of the nothing.

The couplingfield function.

Time synchronization is not alone determined by the iron grip of the three-dimensional carrier-field and the principle of cosmic counter-balance. Apart from this an important effect is realized from the feedback of the gigantic masses of the universes to the timeless waves. These masses can be considered as enormous flywheels without loss. Once started by the super flash, no change in speed is possible, because a force of sufficient magnitude is simply not available. This function of the coupling field can also be shown in our model, because the linear synchronous motors of the flywheels are also generators. If the central driving motor is switched off, then the flywheels continue for some time to drive the central three-dimensioal rotating sphere. Everything stays exactly synchronous with the same phase-shifts.

This experiment shows how the flywheels feed their energy back and have a stabilizing function in time-synchronization. The coupling-field of the masses of the universes is fed back in the time-bound vibrations of the atom-field from which no atom total mass of the universes is fed back to one single atom.

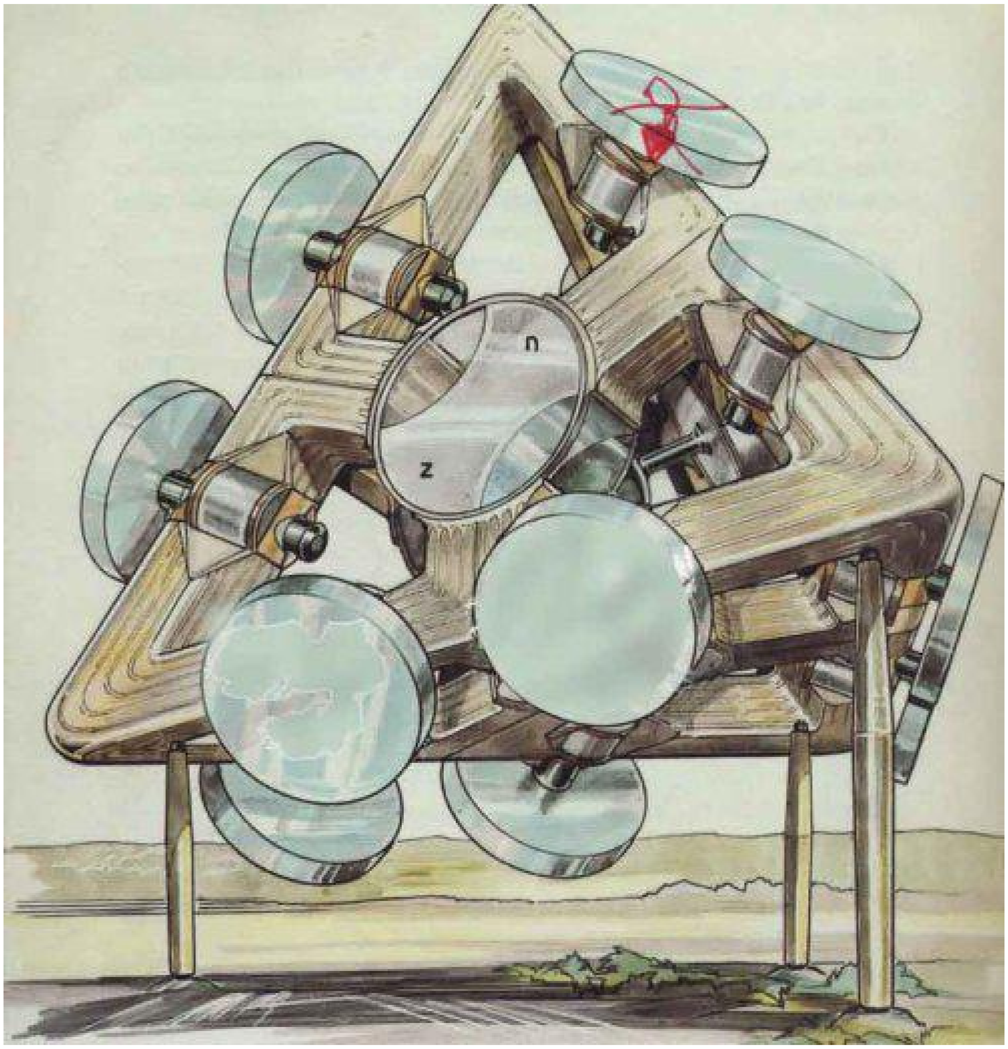

This is a painting by Rudolf Dass of the analogous visual-mechanical carrier-field model which symbolizes time and matter creation and synchronization in twelve universes. This giant working construction that was a cross between a monument and a machine was seen on an artificially flattened mountaintop on **Iarga**. This place was perceived as a place of pilgrimage, a place of meditation and contemplation, and the machine was a symbol of creation. The **Iarga**ns know no separation between science and religion, and religious and spiritual theories are entirely scientific. The model was several hundred yards(meters) long. The machine seemed to operate by some kind of perpetual motion, neither running on fuel or electrical power.

The Instructional Analog Model

The mechanical instruction models are the visual representations of the principles of the cosmic unified field theory. The model is intended to demonstrate that time, matter, and energy are creations out of the nothing and therefore the counter-balance has to be maintained to allow the causal infinity of the nothing to exist. In a broader context the physical symbols all take on new meaning and demonstrate the unification of science and religion. In a religious sense the spherical three-dimensional time-field in the heart of the pyramid symbolizes the Soul of the Creator. The pyramid symbolizes his infinite lovewill, the cage in which the infinity of His existence has been blocked irrevocably. The lines symbolize the different frequencies of the timeless waves. The flywheels on the edges of the pyramid symbolize the twelve associated and interrelated universes, including ours, that make up this unity. One cannot exist without the others.

The timeless primeval atom-principle, the transmitter function.

The atom-field is a time-creating field and as such a remote manifestation of the primeval atom principle. It can be compared with a transmitter. As soon as the program is converted into waves of great power, billions of radios can receive the same program, the number is unlimited. Everything depends on the one program of the transmitter, the primeval atom-field. If this exists with infinite strength, it is at the same time the vibration medium (the carrier-field) which helps it to spread everywhere. *Thus any atom in any universe can receive the timeless basic program.* The second model shows the three-dimensional time-field of the primeval atom field most clearly, although this is a three-dimensional rotation of the central sphere in spatial sense. This is an analog of reality, the atom has a spherical field which is defined by six linear time-vibration directions. Any atom is a remote manifestation of the one carrier-field of infinite strength.

Considering this fifth function of the carrier-field one has to realize that only the central sphere, the atom, is visible to us. The fields in the pyramid edges outside the sphere are invisible. It is the time-field reality of which the atom is a reflection.

The mechanical instruction models are the visual representations of the principles of the cosmic unified field theory. This requires however some marginal notes. A material comparison to explain spiritual abstractions is always in some ways incomplete. There are several hiatus of which only the three most important ones are mentioned.

1. In the model one time-axis represents only *one frequency* of the magnetic field, while in reality all frequencies are contained. Every frequency however can be initiated for each time-axis for which the interference representation of the analog model is valid.

2. The model demonstrates the rotating field of the *timeless waves.* It *does not imitate the time-bound waves* by which our time-axis the propagation speed of time is X, then the phase velocity of the two time-axes connected to "our" pyramid edge is ½X. The other two time-axes connecting our time-axis with the other corner of the pyramid is -½X, making up for the time speed of our anti-universe -X. The crossing, not-touching time-axis is perpendicular to our time-axis and does not contribute.

3. In our model the masses of the universes are given by the flywheel symbol. In fact this is *fundamentally wrong, because it is separated from the three-dimensional rotation of the central sphere.* It can therefore basically not maintain the counterbalance of the nothing.

The most important conclusion of the analog model is that *time, matter and energy are creations out of the nothing and therefore the counter-balance has to be maintained to allow the causal infinity of the nothing exist.*

14. The existential symbolism of the model.

So far the physical aspects of the model have been unravelled, but if one wants to give a wider context to the model suddenly all the physical symbols change their meaning. In relation to philosophical or theological disciplines the model takes on the additional meaning of the unification of science and religion. By certain inherent (medfødte) laws within any intelligent race the denial of God's existence will be destroyed sooner or later. Earth science will provide an indestructible contribution to our final knowledge-awareness, which we shall attribute as typical human attainment in the cosmic integration. It is that piece of our human identity and self-awareness, which we shall possess and care for throughout eternity. *It the light of this purpose, science serves God and the distinction between science and religion fades away.*

In a religious sense the spherical three-dimensional time-field in the heart of the pyramid has a special meaning. **It symbolizes the Soul of the Creator**, the Father who is the pure Identity of the infinite origin. The pyramid symbolizes his infinite Love-will, the cage in which the infinity of his Existence has been blocked irrevocably. He sacrificed his Ego and his freedom, to love his twelve Sons once. He is present only indirectly in his Creation as the reflection of the infinity, the Supreme Being, as the one Ego-consciousness of the Archangels.

That is why the theologian covers the central sphere with pure gold, symbolizing the golden Heart, the Holy of Holies, the most sacred secret of the universe. The pyramid is then painted dark red, the universal colour of love. Over the red many thin golden stripes run in the direction along the pyramid edges. These stripes symbolize the different frequencies of the timeless waves, the expressions of Love and faith of the Father and his infinite Will.

Gathered around the golden Heart are the twelve flywheels, coloured white, the colour of purity and innocence. The model shows that every universe has the power to awaken the matter-consciousness and to transform it to the new

God: the Son. The goal of Creation is therefore the awakening and the formation of the twelve Sons of the Father.

This is not in contradiction with the Bible telling us that the Father sent his only-begotten Son. The Bible is solely directed to man. It is our instruction book for life describing our attitude towards God and our fellow man. The Bible does not tell anything about life on other planets, because it is not relevant in our relationship with God now.

The same goes for our parallel universes. In fact they do not exist for us, because they are not in our world of existence but somewhere else. They do not exist for all intelligent races in this universe, we shall never be able to observe them nor experience them in any way. Perhaps at the end of all time we shall meet. In *our* world the Father has one only-begotten Son and this truth is already so unimaginable that nobody can conceive the consequences yet.

END OF IARGAN DISCUSSION

---

mainpage

the IARGA CONTACT story

ufo-contact-material

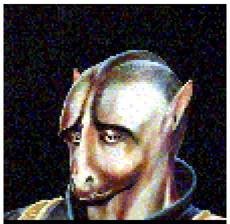

Technical information given from the IARGA contactteam.

their describtions on the matter of TIME

- AND MORE

APPENDIX 1

An Analog Model

1. Introduction

This appendix is not a part of the story of the Creation which is explained in religious terms in Chapter 1 (Part 2). It is a complementary way of reasoning to prove the existence of God, using the concepts of time, matter and energy. These concepts are also used in physics. The problem, which now arises, is that the author will use physical concepts without understanding the real meaning of the words, because he is not a specialist. The words atom, atom-field, matter, etc., must all be understood within the context of this book and they have no direct connection with the ideas used in modern science. Especially the words "matter" and "atom" are most confusing, because the **Iarga**n author knows about the existence of matter in three configurations (and its opposite possibilities as anti-matter); i.e., time-matter, phase-matter and *reflection-matter* (the one we know about in our universe), but he fails to interpret these concepts due to his lack of knowledge of physics. He only knows that these forms of matter are all related to one or none of the directions of time- propagation (tidsforplantning) of the atom of which time is fundamentally three-dimensional. It therefore determines in which form these configurations (beskaffenheter) can exist in our universe.

In this appendix - time-creation out of NOTHING is described - by making use of the concept of more time-dimensions. The explanation of time-creation is very difficult because it uses an overall carrier-field of infinite strength and then this concept is called the cosmic unified field theory. In defining this term, one has to be careful*, because this theory is of a different basic nature from the existing unified field theories here on earth.* Our unified field theories try to unify the four basic forces in nature, which are the electro-magnetic, the weak strong nuclear force and gravitation. The unified field theories accepted by most prominent (iøynefallende) scientists find their origin in the general relativity theory of Einstein, which itself explains matter and energy in terms of space and time.

The problem now is to find a method, explaining the spiritual or esoteric reality, the cosmic time-field. On other planets, this problem was solved by explaining the basic concepts of the cosmic unified field theory. The elementary functions of these concepts can be understood with a mechanical instruction model demonstrating the cosmic time-field visually.

Why indulge in this complication of describing a theory which will probably only be fully grasped by a few individuals? It does not concern the author very much, he only carries out his commission. His **Iarga**n instructors gave three reasons for publishing the concepts of this theory:

1. The future integration of science and religion requires a material proof of God's existence. This explanation could be the start of the exchange between these two.

2. In the previous story of Creation the limitation of the infinite is explained unambiguously(upretendiøst), but the working principle was completely neglected(forsømt).

3. The cosmic unified field theory published in this book will be the only possible verification for the true identity of this book, providing the link for the future development of the human race.

2. The physical approximation of the NOTHING.

One concept in physics is that all atoms and energy consist of vibrations and wave appearances. In a material way, every atom or form of energy is caused by vibration. The physical approximation (tilnærmelse) describes the nothing of the idea of vibration - and *does not touch on the spiritual aspects and the consciousness of matter*. It does not distinguish between the "nothing" and the spiritual "No-Thing" (the ALL). It only recognizes the material aspects of the nothing.

If everything is vibrations or waves, then it is not necessary to speak of matter or energy, then Creation is fundamentally based on this concept of vibration. The primeval vibration (before the limitation) can be described as infinite in time, infinite in all its aspects and possibilities, infinite in frequency and amplitude and existing in all time directions (omnitime). Unchanging, timeless and powerful to all limited vibrations in every bound power. It is the omnipotent definition.

In this approximation the nothing is more simple to understand. Before the creation - the infinite omnipotent vibration existed- *but a vibration in every direction and of infinite frequency and amplitude is not a vibration*. The apparent paradox can be eliminated by stating: Only at the instant when this infinite vibration blocks itself by interference*, as in a flash, all bound waves and frequencies were created.* This condition is very close to the existential view of the Creation out of NOTHING in Chapter 1.

The nothing (here) has no counteraction, but in contrast the omnipotent vibration - can manifest itself only by counter-acting in such a way that the infinite allows the nothing to be. So one can also state that all vibrations were created because the nothing bound itself by a mechanism of interference quenching (ble dempet) - (counteracting vibrations) of infinite force, which blocked the infinite possibilities of these vibrations. *This universe can only be considered as the limitation of the unbounded nothing.*

3. The bounding of the omnipotent infinite vibration, the cosmic time field.

One of the basic principles on which the mechanism of bounding is based is: *If one property of a power (any power) of something is bound, then all properties and powers of this something will be bound.* In infinity the finite cannot exist. To limit the omnipotent vibration it was sufficient to bind only one of the perhaps infinite number of possibilities. Because if only just a little something were to be formed in the nothing, the nothing would be altered and cancelled irrevocably, then the complete finity (endelighet/begrensethet) was a fact. OR, if there were just the three-dimensional rotating time-field as a possibility of the primeval atom in the infinity of the nothing, then the finite was a fact. This atom-field needed to be a force-field of infinite strength, the carrier-field, to counteract the omnipotent vibration, the cosmic time-field.

4. The Creation can only exist as a result of counter-balance.

One of the functions of the carrier-field is time-synchronization by which the cosmic counter-action and therefore the nothing can BE. *Matter and energy cannot exist on its own in the nothing.* The principle of time synchronization is given:

1. Time is only relative, it is motion or propagation (spredning) in relation to at least a second motion of time.
2. the cosmic law of counterbalance dictates that the second time has to counterbalance the first time. So time can only exist in relation to a second exact *opposite direction* of time (running backwards in relation to our time).

3. Time exists only through the presence of atoms or matter. Then the backward-running time can only exist through matter in which the time propagation is opposite, *running backwards* in relation to us (antitime matter). Anti-matter or reflection antimatter may exist in our universe with an opposite electric charge with respect to our known reflection matter. The question, as to which charge antitime matter has in the anti-universe, is of no importance in this appendix.

4. Because neither matter nor energy can exist on its own in the nothing -another quantity of matter of opposite properties *must* exist which is exactly similar to that in our universe. This balance of matter and energy is so accurate, that one surplus atom (overskuddsatom) is impossible.

This explains nearly as per definition the cosmic law of counter-balance allowing the nothing to be. There must be at least a second universe exactly like ours, in which time runs backwards according to us, but normally for life in that universe. In the following it is shown that there are not just two, but *eleven* other universes in addition to ours necessary to maintain the cosmic law of counter-balance.

6. The time-synchronization field.

The law of counter-balance also applies to time. The timedirection is opposite and the time velocity is equal to ours. Time is the propagation of vibrations of the atom. Then the atomvibrations in both universes are equal and opposite with respect to the timelessness. This is only possible if these times are caused by one field (time-field).

The definition of a field (any field in physics) is: If an event occurs, because another event takes place somewhere else without a mechanical or visible connection between these events, then these events are caused by a field. Time-synchronization occurs according to this definition. The atom vibrations in our universes are exactly similar, because the atom vibrations of our anti-universe occur despite the enormous difference in time and without any dissipation of energy. It is about 30 billion (milliarder) years since the occurrence of the super flash Creation (big bang) between the two universes.

The law of conservation of counter-balance is causal to the law of energy conservation in every universe. Each universe is a closed system and without energy dissipation outside itself. The amount of matter and energy once created remains constant. *A loss of energy is impossible, because our time-bound energy propagates forwards, while the time-bound matter and energy of the anti-universe propagate backwards in time.* These universes are separated by a time-barrier through which no communication nor transfer of life is known. (Well - according to information in the book UFO-CONTACT FROM PLANET KOLDAS and also told by the et-contact ASKET and Semjase - there are some methods for entering the mirroruniverse - but this can not be touched here - reseach yourself if interested!! R.Ø.remark.)

6. What is a timeless vibration or wave?

The question is how the field functions maintain the time-synchronization of the atom-waves between the universes separated by billions of years. It is only possible if the synchronization stands still in time, but exists simultaneously in all the time elapsed (- som er gått..) since the beginning of the super flash, as well in all the future. *A timeless vibration does not normally move in time*, the amplitudes of the field propagation - alternate (veksler) forwards and backwards in time. Time synchronization is the principle by which our universe catches every forward-moving amplitude and the anti-universe catches every backward moving amplitude of one and the same timeless wave.

On other planets there are simple mechanical models explaining the function of time synchronization. One such model is very similar to our well-known mini-model rail upon which are two little wagons. The model has two lengths of rail of say 1.50 meters each, which are in direct line with each other.

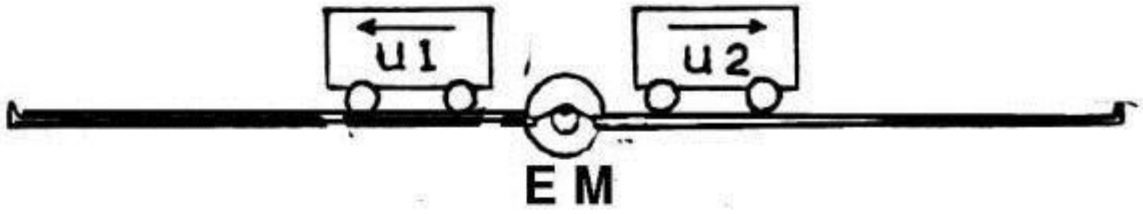

Picture of the model wagons with a Electric drive motor which letting an excentric, and so the rails vibrate horisontally.

Rail and wagon model demonstrating the principle of the timeless waves. The vibration is present at any point of the rail and at any time during the whole process period. Also from the moment of the super flash until the end of all times (ALWAYS).

The rails are mounted on small springs to allow them to vibrate forwards and backwards in horizontal direction (from left to right and vice versa). The two rails are connected to a little electric motor with a double eccentric, each driving one rail, letting the rails vibrate about 1 millimeter of linear distance, also backwards and forwards. The amplitude is small enough to be invisible to the eye.

Then one little wagon is placed on each rail and each wagon has a catch with a catchwheel (tilbakeløps-sperre) on one of the wheelaxes, so the wagon can ride in only one direction. *The movement in the other direction is blocked.* The wagons are placed with their rears against each other, after which the electric motor is switched on. If the teeth of the catchwheel have the same beat as the vibrator mechanism, the wagons roll away from each other with exactly the same speed to the end of each rail.

With this simple model time synchronization for the atom-vibrations of both departing universes (in time) by a timeless field "always existing everywhere" has been demonstrated sufficiently. At the same time a flaw(feil) comes to light. It is impossible to maintain an energy time flux of the timefield for the two departing universes, because where is then the inexhaustible (utømmelige)energy source to maintain the time-flux? (betyr det tidsustabiliteten??)

This is one of the reasons that a one-dimensional time-field is not possible. Only a three-dimensional time-field, which encloses the energy time-flux in itself, allows our universe and the others to exist.

7. The principle (mechanism) to create finity: INTERFERENCE.

So far the function of time-synchronization of the carrier-field has been considered. Now we will discuss the medium of conserving and maintaining the timeless vibration into eternity. It is the infinite force-field carrying all matter and energy.

We are at the moment of the flash or big bang, the moment at which the infinite primeval vibration limited (transmutated) itself to the carrier-field. This mechanism of blocking is well known and based on the principle of wave interference. (Interference of lightwaves is important today in laser technology.) Interference is the action or counter-action of two or more waves. If the *wave-propagation(-forplantning) is opposite and parallel* and the frequencies of the waves equal, mutual amplification (gjensidig forsterkning)or attenuation (fortynnelse) of the intensities of the wave is possible, depending on the phase relation between these waves.

If *the phases are opposite* and the *amplitudes of the waves are equal,* total quenching (dempning/kvelning)is possible. This is the principle of the binding of infinity. If an infinite wave manifests itself in one time-direction, automatically the equivalent absolute counteraction is created which counter-balances the infinity, allowing the nothing to be.

The infinite primeval wave(urbølgen) blocked itself at the moment it created itself and so it never existed in reality. It was in fact a *transmutation (forvandling) of infinity;* the transmutation of the infinite vibration of infinite amplitude and frequency to a carrier-field of infinite strength, which blocked itself by interference.

8. The time-bound vibrations or waves.

The primeval wave blocked itself in a three dimensional time-axis system (propagation directions) by interference, and trans-mutated to an infinite, timeless, immobile force-field. How could the bound timeless waves exist, which synchronize our universe?

It is an error of thinking to suppose that by quenching the omnipotent vibration - all finite waves were also cancelled. It is just the opposite. A vibration with infinite frequency is not a vibration in reality. The infinite prevented the vibration, because a vibration can only exist in the finite. As soon as infinity was removed, the prevention disappeared and the real bound wave was a possibility. It is the carrier forcefield although of infinite strength and energy density which gets its finite value due to interference quenching. All timeless waves are as it were encapsulated in infinity and therefore indestructible. However "belonging" to infinity and not bound by time themselves, they create time forever, they are the eternal ripples on the static ("krusning i likevekten"), unchangeable and loss-free field of infinite strength. (This must be what the Danish wiseman Martinus call the X2- principle and the primeval-desire or "urbegjæret" in Danish. R.Ø.remark.)

Because of this, all time-bound waves such as light, radio-waves or gravitation, exist in space without any loss and it is the empty space in the universe, which is the non-modulated structure of the carrier-field. It is one of the most fantastic touches of genius of Creation, that the non-vibrating infinity was transmutated to a loss-free medium carrying all time-bound and timeless waves. A few properties of this medium can be measured with instruments, the other part of its structure cannot be objectively observed yet, *it is the esoteric fluidum or the infinite Love-will of the Father which carries the whole Creation.*

9. The unimaginable three-dimensional time.

(Den utenklige tredimensjonale tid)

One time-axis counter-balances only one direction of time propagation of the omnipotent (allmektige)wave. In infinity all possibilities of vibration exist and therefore all time-propagation directions. The question is, could infinity be limited by just one time-direction and one co-existing propagation direction going the opposite way?

One way to answer this was already given in the previous paragraph, in which it was stated that the energy time-flux of the primeval wave of two universes is not a closed system. Another plausible answer can be by simply reasoning.

The atom is causal for the time - and space dimensions, due to its time-creating properties, already mentioned. *The space-creating properties are easy to understand, because the atom takes up space.* It is three-dimensional in space and therefore it creates space in three dimensions. If all atoms were two-dimensional then the whole of creation would have been on a plane and space would not have been more than a flat or curved plane(as on a sheet of paper. R.Ø.remark.) Our time-space is no more than a bound part of the nothing (det "noe som er" -X1- som Martinus kaller det. R.Ø.remark.) - being filled with matter or energy. Anything that space distinguishes (utskiller) of the time - and - dimensionless void, is caused by atoms or energy. If atoms have three space dimensions, then the timeless waves creating the atoms also have three dimensions.

1. The timeless primeval atom-field has three time dimensions.

2. The timeless waves could not be bound by only one time-axis, but must necessarily have three time-axes.

3. The timeless atom-field should consist of a three-dimensional rotating field of infinite strength, which is attenuated (fortynnet) of itself by interference. It conserves

the law of cosmic counter-balance of the no-thing. *The atom of such a rotating field structure has six time-axes, but in principle these six axes can always be reduced to three time-dimensions, no further reduction is possible.*

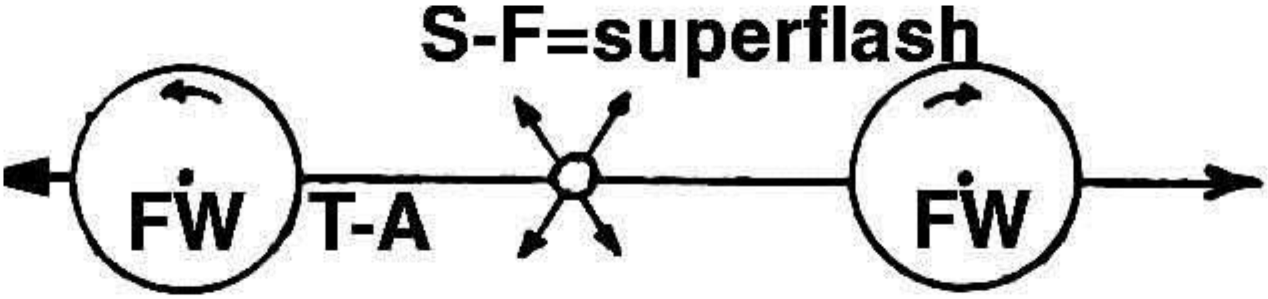

Picture of ONE-TIME-AXIS: The cosmic symbol of one time-axis.

In the following parpagraphs of the analog carrier-field model it is shown, why the atom-field has to be a three-dimensional and not a two-dimensional field.

10. The six time-axes of the time-creating, cosmic carrier field.

How can it be established that the carrier-field of all matter in twelve universes has six time-axes? The first part of the answer is simple to explain. *A well-known type of rotating field on earth is that of the electric three-phase synchronous motor.*

This type of electromagnetic-field system consists of three linear magnetic fields, which are arranged at angles of 120 round a rotor. The phase-differences of the three alternating magnetic-fields, which have the same frequencies (say 50 Hz) are also at each 120 degrees. Each alternating field is in itself an exact linear vibration oriented in one direction. The rotating field exists only if the three linear vibrations are combined together with the proper phase-relation between them. So a two-dimensional rotating field in space is created. It can force a piece of magnetic material to rotate around its axis with a great power.

Insert pict of motor

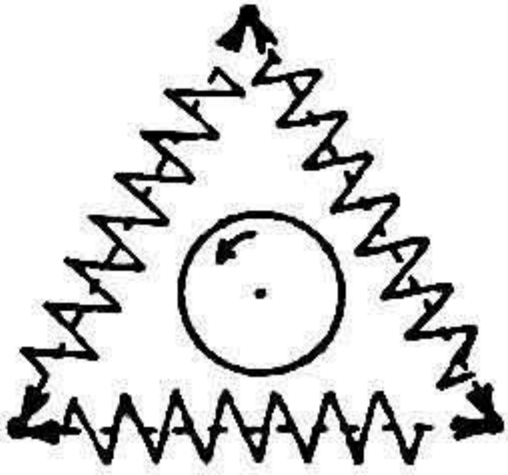

The principle of an electric three-phase synchronous motor.

The fact that it is possible to create rotating fields by combining three linear alternating waves does not mean that it has to be the same for fields of infinite strength, it can be more than three, but it is not very likely as first guess, because the three wave principle (the equilateral triangle) is the most simple and effective. According this principle, for a three dimensional field six linear waves are required to form an *equilateral (likesidet) triangular pyramid,* but why an angle of 60 degrees between time-axes?

The first answer can initially be given in a very simple way, the second is more complicated, but at least more to the point.

A simple way of reasoning to derive (utlede) the answer to the question of 60 degrees is the following. The atom field of matter has spatial (romlig)structure, therefore the time-axes must be positioned in such a way that they form a closed three-dimensional geometric object, because the energy fluxes of time must be closed in itself due to the law of counter-balance.

The time-axes form straight lines, which cannot be parallel lines, because parallel time-axes form one and the same time-direction. *The most simple three-dimensional closed object is an equilateral triangular pyramid.* There are no reasons to suppose it should be an a-symmetric pyramid if three-dimensional time is considered, on the contrary, it must be symmetrical in all positions. So it is simply a equilateral triangular pyramid of which the lanes are equilateral triangles. There is no more simple object in three-dimensional space with these requirements.

The more sophisticated answer to the question of 60 degrees is time given by approaching the question from the relativity of time frequencies between the other universes (the other time-axes). Time is the wave propagation of atoms which determines the direction of the past to the future.Simultaneously these time-creating or timeless waves are the linear phase waves combining to maintain the atomic rotating field. The problem is that "our' atoms do not experience the right frequencies of the other time-axes, because they do not propagate (spredes)with the same speed, but advance in time.

Thus there is an apparent frequency-shift, a kind of Doppler shift, by which the frequencies of the "other" time-axes are experienced by "our" atom. One should realize that the timeless vibrations of the other universes also exist in our universe.

At first sight it seems impossible for a rotating field to posess a constant phase-shift because between vibrations of different frequencies a constant phase-shift is not possible unless the frequencies are related to each other in a particular way. They are harmonic. Two vibrations are harmonic if the apparent frequencies have constant ratios to each other, like 1:2, 1:3, 1:4, etc.

Timeless waves can only exist when the frequencies are related by the harmonic ratio 1 : 2.

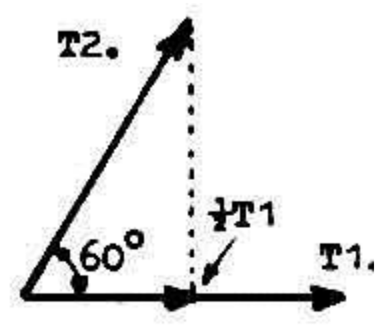

Picture explaining the frequency-shift for 60 degrees time-axis angle.

The frequency ratio is determined by the angle between the two time-axes, as the picture shows. Our time direction is T1 while another time direction is shown as T2. The time direction propagating forwards (from past to future) is useful to our universe, while the backward-moving impulses of the timeleess wave do not "catch" the rotating field of "our" atoms. Secondly the time-axis at a right angle (90dg) to our axis T1 does not make any contribution, we experience this time-axis as non-existing.

As the picture shows, due to the angle of 60 degrees, 'our' atom experiences exactly half the time-speed of the other time-axis and this makes the frequency harmonic. As T1=2 times T2, the result is the simplest harmonic ratio of 1:2, determined by the angle of projection of 60 degrees between two time-axes. (Pythagoras' law for right angled triangles).

All other possibilities and other harmonic ratios fail to give non-destructive contributions for the necessary interference intensification. Other excitations of higher order harmonic ratios give more complicated time-axes structures of the rotating atom field and cannot be stable. Also because the couple-field feedback to the total mass (inertia) of the universes is not optimized (see the following paragraphs).

It supposes simultaneously the existence of a higher number of co-existing universes, which seems highly unlikely (usannsynlig)(higher than twelve). Therefore we must decide for the simplest three-dimensional geometric form: the equilateral triangular pyramid.

11. The ancient symbol of God, the Star of David.

It is necessary to summarize the above issues of the principle.

1. The choice of the mechanism of limitation determined the existence of all types of timeless waves, among which was the (most complex) three-dimensional rotating time-field.

2. The limitation took place by interference quenching (forstyrrelses-dempning) the primeval infinite vibration (possibility) affording more time-axes in three dimensions.

3. A time-axis is a part of the infinite force-field. This only exists as a single linear vibration.

4. For a three-dimensional rotating field at least six linear field waves are necessary, which are located spatially in time having angles of 60 degrees between the time-axes.

5. The angle of 60 degrees could only be created by an apparent frequency shift maintaining the harmonic vibrations of the rotating field.

These considerations determine the mechanism of blocking the infinite. It consists of six time-axes forming a closed three-dimensional object, which creates and restricts the speed of the rotating atom-field.

Though every time-axis carries two universes(universe and anti-univers), there are in the frightening depth of cosmic space twelve distinct cosmic structures, separated by the unbridgeable barrier of time. Not only are there eleven other universes apart from ours, *but they are all embedded in the same cosmic or geometric space*. All universes have the same size and the same energy content. The geometric space is necessarily the location in the nothing - in which the omnipotent vibration quenched itself by interference. It is the location for waves encountering each other, as it is for light-waves interfering with each other. But it does not require that every atom has twelve buldges at the same space-point in time. Or more crucially, it does not mean that every human being has twelve images of himself without knowing it.

It is the mechanism of interference synchronisation of the timeless waves which uncouples the universes at each space-point, simultaneously maintaining the time-propagation of each universe.

*The equilateral triangular pyramid is the symbol of God, the Father as it is known everywhere in the universe (after lifting the cosmic isolation).*We know it as the double equilateral triangles of the Star of David, also six time-axes with angles of 60 degrees. It means that the principle of the limitation was revealed to humanity centuries ago, but we did not understand its meaning so far.

make in your thoughts : The Star of David.

Technical information given from the IARGA contactteam.

their describtions on the matter of TIME

- AND MORE

picture is described later

12. The analog carrierfield model, an instruction model.

The cosmic carrier field of all space and matter performs a sixfold function, of which some have already been discussed in gereral way.

1. The carrier function of an infinitely strong forcefield blocked in itself contains the fluidum conducting all waves without dissipation and maintained to eternity.
2. The orientation of the time-axes of the universes with respect to nothing provides the structure of the carrier-field of three dimensions, the bounded Love-will, creating our life and existence.
3. The time synchronisation function locks all universes in the iron grip of the law of cosmic counter-balance from which no atom can escape.
4. The feedback principle of time to the total mass of the universe is the function of the coupling-field by counterbalance (and interference) of the time-bound waves.
5. The function of the timeless primeval atom-field, the atom-principle, as the transmitter program containing the code of matter of which all atoms are remote manifestations.
6. The function as bearer of the absolute ALL-knowledge and goal consciousness, manifesting itself as the laws of nature and as the intelligent goal of life: the spirit of God.

The last function concerns the existential aspects, which have been discussed in the previous section of this book.

The first analog model is a mechanical instruction model explaining the cosmic carrier field by means of symbols. This model was built on another life-supporting planet on a high mountain and was many meters high. It was the gathering point for contemplation of that intelligent race for devotion to its Creator.

It was the symbol by which the majority of the race acknowledged the goal of Creation, simultaneously the sign that the cosmic isolation had been lifted. The other function of the carrier-field will be discussed with the help of this mechanical instruction model.

The analog model consisted mainly of the six edges of a huge equilateral triangular pyramid built out of solid soft-iron bars. The many metres thick edges of the pyramid contain an impressive number of iron bars parallel to the direction of the edges and cast in a kind of plastic resin following a process similar to that of casting concrete(betong). So a mainly iron-frame was formed especially suitable for conduction of magnetic pulses of high power without much loss.

So far the carrying fuction of the cosmic time-field, which primarily maintains the timeless waves is symbolized in the model. The magnetic pulses in the edges of the pyramid run backwards and forwards like the timeless waves in reality.

With respect to the other edges of the pyramid, the magnetic pulses are phase-shifted. The problem is that up to now the forward propagation (forplantning) in time has been missed out. In fact we look for the equivalent of the two little wagons on the vibrating rail. The solution can be found in the symbol of the rotating flywheel, as *with the atom its rotation is continuous in time, it is in fact time propagation.* Therefore the iron frame of the pyramid is discontinued at two places in each of the six edges of the pyramid. At this discontinuation point a linear magnetic impulse motor is installed, on the axis of which a large flywheel is mounted.

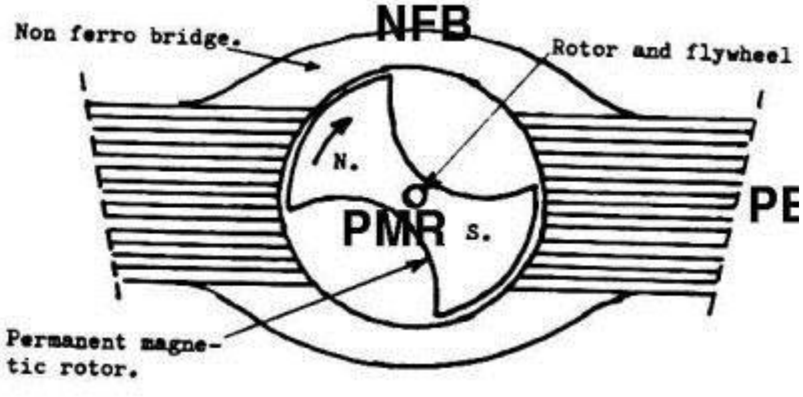

NBF=NonFerroBridge

PE=PyramidEdge and soft iron bars

PMR=PermanentMagneticRotor

The flywheel symbolizes not only the propagation of time, but also the inertia of matter and so the flywheel is in effect the total mass of the universe. This type of magnetic impulse motor is not self-starting, so at the inside on the flywheel axis a starter-motor is mounted, fitted solely for the purpose of starting. One rotating the synchronous motor maintains the exact number of revolutions at the same frequency of the magnetic alternation field. Thus in each of pyramid's edges two synchronous motors are mounted making twelve flywheels in total.

The two flywheels per axis rotate in opposite directions to symbolize the time propagation in opposite time directions. This type of linear impulse motor is chosen because it can be made to rotate in either direction.

The first analog model.

In one triangular plane of the pyramid the six synchronous motors in the iron frame demonstrate a rotating magnetic field, seen from the center of the plane. Each edge of the pyramid has the right phase-shift with respect to the other edges of the triangle.

The first model has four central rotating magnetic field generators in the middle of each triangular plane. (See figure of unfolded pyramid indicating the actual phase-shift).

These four synchronous (three phase) generators are coupled together by a gearbox and are driven by the main electric motor (not shown in the drawing). Consequently the four three-phase generators fulfill the synchronisation and constant phase-shifts for the flywheels of the pyramid-edges. The four separate rotating fields together, initate the complex three-dimensional rotation of the carrier field (or atom-field).

The disadvantage of the first model is that it is further away from reality, because the atomic field is not symboized as one three-demensional rotating field. However it shows better how to imagine the complicated rotation in three dimensions.

The second or the real analog model.

This mechanical model has only one central generator in the middle of the pyramid, instead of the four rotating field generators of the first model. All magnetic waves in the pyramid edges go straight to the pyramid's center. The rotor of the 3-dimensional rotating generator is like a yo-yo with spherical end-surfaces.

It requires a touch of genius to provide the drive-connection between the central motor and the spherical rotor, which has to be started up in a complex rotation. The six diametrically opposed pole-shoes from the middle of the pyramid's edges around the spherical rotor form in pairs, three axes of right angles - representing a Cartesian coordinate system. So it shows that the six time-axes of the edges of the pyramid can be converted into the three principal time-axes of the central sphere.

Together it imitates three-dimensional time, which causally creates the three-dimensional space of our universe.

The principles of the first and second analog models are equivalent. One version can be converted into the other, for example, converting the four rotating fields to one three-dimensional rotation in the centre-mass of the pyramid.

The model can be started by switching on the thirteen drive-motors. As the flywheels reach velocity the twelve starter-motors are switched off and only the central motor drives the complete gigantic model. The essence is that all flywheels have exactly the same number of revolutions. They are all either synchronous motors or generators. Then the flywheels are marked and illuminated by a stroboscopic flash light. *Seen by* *the normal eye they do not appear to move showing the synchronous rotation and the constant phase-shifts between the flywheels.*

These mechanical models are meant to explain the unimaginable Creation out of the nothing by symbolic imitation of the carrier-field. However it is only a weak symbol of Creation. The carrier-field is the esoteric medium, a transmutation of the absolute spirit or conscience, it is the spiritual force field of infinite strength which is present everywhere in the universe.

The imitation of the carrier-field functions.

Nevertheless the analogy of reality is incredible. This will be shown in the next explanation of the five functions of the cosmic carrier-field.

The carrier-function

The foundation of the carrier-field is the infinite force-field created by interference quenching (avbrytende stopp) of the original infinite vibration. The mode initiates this force-field with the six bundles of soft-iron bars as the edges of the pyramid.

As in reality, the three-dimensional magnetic rotating field in our model is caused by the linear waves of each time-axis. *The timeless waves alternate backwards and forwards as in our model*. They drive the flywheels representing the masses of the universes, and their rotation is the propagation (spredning) in time.

The orientation of the time-axes with respect to the nothing.

The relative position of one time-axis in the apparent nothing plays an important role in the orientation of the universes in reality, determining the direction of each time-axis. Here the first model is shown representing the four three-phase synchronous generators in the planes of the pyramid. Only with decompostion of the three-dimensional rotation into four rotations can one show that the three-dimensional rotating field has left no remaining component in the nothing, while in contrast a two-dimensional rotating field can not be nullified, a remaining component will always exist in the nothing. This can be explained by considering a single plane of the pyramid with its six flywheels at the edges, the analog for a two-dimensional

carrier-field. The rotation impulse of the flywheels is compensated by the opposite rotations but the rotation impulse of the central three-phase generator is left over. It can be measured by the outside world (the nothing).

It requires four field three-phase generators (maintaining overall symmetry) to compensate for these rotation impulses.

With this the three-dimensionality of the carrier-field has been shown and the orientation of the time-axes with respect to the nothing is determined.

The time synchronization function.

Impressively the model imitates the time synchronization, simultaneously maintaining the cosmic counter-balance by which the universes continue to exist. The rotation of the two flywheels per axis is in opposite directions and they rotate with exactly the same speed. Thus it imitates the circular motion to the timeless waves which progresses exactly synchronous by their mutual interference coupling.

The twelve cosmic systems run without any dissipation, once started, the perpetuum mobile continues into eternity with the absolute accuracy of time-synchronisation. If one supposes the model to be without loss and without using the eternal driving motors, it will continue to run synchronously due to the inertia of the flywheels and the fact that the synchronous linear impulse motors can function as generators as well.

This symbolizes another property of the continuous circular motion of the timeless waves, because every wave somewhere in the pyramid circuit at each moment has its opposite value resulting in an overall zero at any moment.

This is the law of conservation of the nothing.

The couplingfield function.

Time synchronization is not alone determined by the iron grip of the three-dimensional carrier-field and the principle of cosmic counter-balance. Apart from this an important effect is realized from the feedback of the gigantic masses of the universes to the timeless waves. These masses can be considered as enormous flywheels without loss. Once started by the super flash, no change in speed is possible, because a force of sufficient magnitude is simply not available. This function of the coupling field can also be shown in our model, because the linear synchronous motors of the flywheels are also generators. If the central driving motor is switched off, then the flywheels continue for some time to drive the central three-dimensioal rotating sphere. Everything stays exactly synchronous with the same phase-shifts.

This experiment shows how the flywheels feed their energy back and have a stabilizing function in time-synchronization. The coupling-field of the masses of the universes is fed back in the time-bound vibrations of the atom-field from which no atom total mass of the universes is fed back to one single atom.

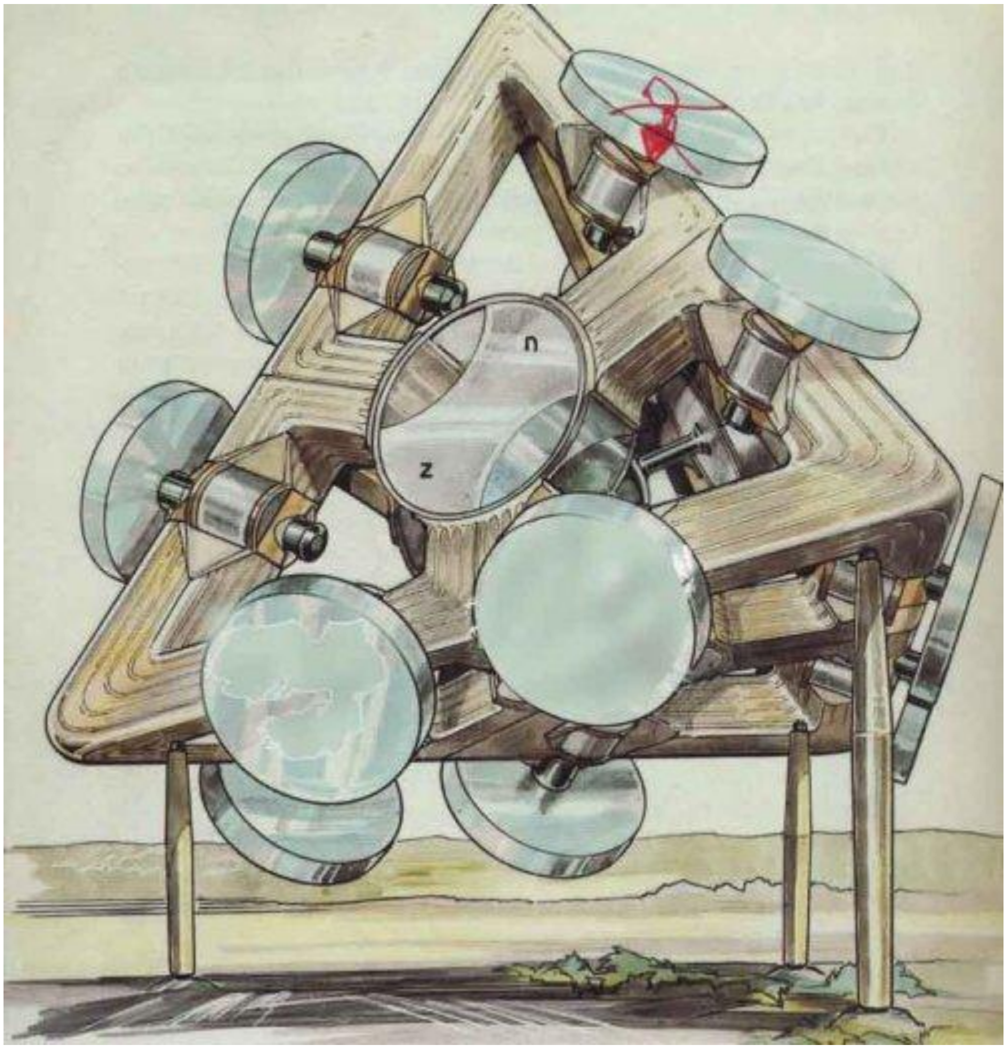

This is a painting by Rudolf Dass of the analogous visual-mechanical carrier-field model which symbolizes time and matter creation and synchronization in twelve universes. This giant working construction that was a cross between a monument and a machine was seen on an artificially flattened mountaintop on **Iarga**. This place was perceived as a place of pilgrimage, a place of meditation and contemplation, and the machine was a symbol of creation. The **Iarga**ns know no separation between science and religion, and religious and spiritual theories are entirely scientific. The model was several hundred yards(meters) long. The machine seemed to operate by some kind of perpetual motion, neither running on fuel or electrical power.

The Instructional Analog Model

The mechanical instruction models are the visual representations of the principles of the cosmic unified field theory. The model is intended to demonstrate that time, matter, and energy are creations out of the nothing and therefore the counter-balance has to be maintained to allow the causal infinity of the nothing to exist. In a broader context the physical symbols all take on new meaning and demonstrate the unification of science and religion. In a religious sense the spherical three-dimensional time-field in the heart of the pyramid symbolizes the Soul of the Creator. The pyramid symbolizes his infinite lovewill, the cage in which the infinity of His existence has been blocked irrevocably. The lines symbolize the different frequencies of the timeless waves. The flywheels on the edges of the pyramid symbolize the twelve associated and interrelated universes, including ours, that make up this unity. One cannot exist without the others.

The timeless primeval atom-principle, the transmitter function.

The atom-field is a time-creating field and as such a remote manifestation of the primeval atom principle. It can be compared with a transmitter. As soon as the program is converted into waves of great power, billions of radios can receive the same program, the number is unlimited. Everything depends on the one program of the transmitter, the primeval atom-field. If this exists with infinite strength, it is at the same time the vibration medium (the carrier-field) which helps it to spread everywhere. *Thus any atom in any universe can receive the timeless basic program.* The second model shows the three-dimensional time-field of the primeval atom field most clearly, although this is a three-dimensional rotation of the central sphere in spatial sense. This is an analog of reality, the atom has a spherical field which is defined by six linear time-vibration directions. Any atom is a remote manifestation of the one carrier-field of infinite strength.

Considering this fifth function of the carrier-field one has to realize that only the central sphere, the atom, is visible to us. The fields in the pyramid edges outside the sphere are invisible. It is the time-field reality of which the atom is a reflection.

The mechanical instruction models are the visual representations of the principles of the cosmic unified field theory. This requires however some marginal notes. A material comparison to explain spiritual abstractions is always in some ways incomplete. There are several hiatus of which only the three most important ones are mentioned.

1. In the model one time-axis represents only *one frequency* of the magnetic field, while in reality all frequencies are contained. Every frequency however can be initiated for each time-axis for which the interference representation of the analog model is valid.

2. The model demonstrates the rotating field of the *timeless waves*. It *does not imitate the time-bound waves* by which our time-axis the propagation speed of time is X, then the phase velocity of the two time-axes connected to "our" pyramid edge is ½X. The other two time-axes connecting our time-axis with the other corner of the pyramid is -½X, making up for the time speed of our anti-universe -X. The crossing, not-touching time-axis is perpendicular to our time-axis and does not contribute.

3. In our model the masses of the universes are given by the flywheel symbol. In fact this is *fundamentally wrong, because it is separated from the three-dimensional rotation of the central sphere.* It can therefore basically not maintain the counterbalance of the nothing.

The most important conclusion of the analog model is that *time, matter and energy are creations out of the nothing and therefore the counter-balance has to be maintained to allow the causal infinity of the nothing exist.*

14. The existential symbolism of the model.

So far the physical aspects of the model have been unravelled, but if one wants to give a wider context to the model suddenly all the physical symbols change their meaning. In relation to philosophical or theological disciplines the model takes on the additional meaning of the unification of science and religion. By certain inherent (medfødte) laws within any intelligent race the denial of God's existence will be destroyed sooner or later. Earth science will provide an indestructible contribution to our final knowledge-awareness, which we shall attribute as typical human attainment in the cosmic integration. It is that piece of our human identity and self-awareness, which we shall possess and care for throughout eternity. *It the light of this purpose, science serves God and the distinction between science and religion fades away*.

In a religious sense the spherical three-dimensional time-field in the heart of the pyramid has a special meaning. **It symbolizes the Soul of the Creator**, the Father who is the pure Identity of the infinite origin. The pyramid symbolizes his infinite Love-will, the cage in which the infinity of his Existence has been blocked irrevocably. He sacrificed his Ego and his freedom, to love his twelve Sons once. He is present only indirectly in his Creation as the reflection of the infinity, the Supreme Being, as the one Ego-consciousness of the Archangels.

That is why the theologian covers the central sphere with pure gold, symbolizing the golden Heart, the Holy of Holies, the most sacred secret of the universe. The pyramid is then painted dark red, the universal colour of love. Over the red many thin golden stripes run in the direction along the pyramid edges. These stripes symbolize the different frequencies of the timeless waves, the expressions of Love and faith of the Father and his infinite Will.

Gathered around the golden Heart are the twelve flywheels, coloured white, the colour of purity and innocence. The model shows that every universe has the power to awaken the matter-consciousness and to transform it to the new God: the Son. The goal of Creation is therefore the awakening and the formation of the twelve Sons of the Father.

This is not in contradiction with the Bible telling us that the Father sent his only-begotten Son. The Bible is solely directed to man. It is our instruction book for life describing our attitude towards God and our fellow man. The Bible does not tell anything about life on other planets, because it is not relevant in our relationship with God now.

The same goes for our parallel universes. In fact they do not exist for us, because they are not in our world of existence but somewhere else. They do not exist for all intelligent races in this universe, we shall never be able to observe them nor experience them in any way. Perhaps at the end of all time we shall meet.
In *our* world the Father has one only-begotten Son and this truth is already so unimaginable that nobody can conceive the consequences yet.

END OF IARGAN DISCUSSION

Stefan Denaerde

Cosmic Unified Field Theory.

How it started.

More than 25 years ago appeared an amazing book in the Netherlands: ´Buitenaardse Beschaving´ (Extraterrestrial Civilization) from Stefan Denaerde. (Known in English as : "UFO contact from planet Iarga") It describes the confrontation of the author with the race from the planet Jarga (Iarga). It was written as a novel (although it is a true story) and illustrated by a well known Dutch technical illustrator (Rudolf Das).

Sailing with his ship in Dutch waters, Mr. Denaerde (not his real name) bumped into something which was hidden underwater. When he sees someone floating in the water he comes to the rescue of what later appears to be an astronaut. This rescue was a test for him and after the initial shock he gets in contact with the Jargans who invite him for a three day learning session inside their submerged saucer.

Via a 3D screen and a radiation ´head set´ he learns about the Jargan culture, philosophy and technical achievements. Through the concepts of extreme efficiency and justice he is taken to the Jargans love for the Omni-Creativity (our word for ´God´ is too poor) and their striving to be one with it.
This ´carrier field´ of the Universe supports everything (matter, time, consiousnes), and by applying this knowledge it makes interstellar space travel possible.
It appears that the goal of a civilization is to create a socially stable supercivilization. Discriminations of any kind no longer exist. Resources and wealth are available for everyone, but not as a goal as such, but to enable everyone to grow spiritually.

There is much more to this book. Anyone interested can find an interview with Denaerde by Wendell Stevens, as well as several chapters from the book on this link. (This also contains the 2nd book of Denaerde!)
From this interview it appears that the contact has continued for many years in telepathic form. A lot of highly technical information has been given in this way, which resulted in the second book:

The Universal Creation. (In Dutch: ´de Universele Schepping´)

This appeared about 20 years after his first book. It describes in one theory the origin of the Universe, matter, time, consciousness, the Universal goal for any civilization, the history and possible (very bright) future of mankind, in one word : a TOE (theory of everything).
The atmosphere of the book is totally different: the first book had something like ´you don't have to believe this has happened, it could be hypothetical´. The Universal Creation is much more direct and confronts us with a direct proof of its contents, because the authors (Stefan and his ´brother´ Christian Denaerde) express that it is high time that science and religion are united again.
The Universal Creation comes with an appendix ´Cosmic Unified Theory´- in English!- which goes into much detail about the new physical model. Recently this information is available on CR-Rom, to be ordered at the Cosmic Field Paradigm foundation. This essay is mainly about the appendix and the CD-Rom.

The creation out of the nothing.

According to the Cosmic Field (CF) model there existed prior to the creation (or big bang) vibrations of infinite frequency and amplitude in every time direction.
The act of creation was the blocking of the infinite waves by interference to form waves of finite frequency and amplitude. In order for matter and energy or anything to exist, there must be an equal counterbalancing vibration. Also time direction has to be counterbalanced. Think of a time axis as two pieces of mini model rail which are connected via a motor, which can move the rails back and forth say over 1 cm. If you place a wagon

on one of the rails which has a catch wheel that catches only the forward movement of the rails, you will see the wagon move forward. Another wagon, placed in the opposite direction on the other rail will also move forwards, thus the two wagons will separate from each other. So the rails can be seen as a time axis, where time moves back and forth. The wagon represents matter, which can only catch one time direction through a property called spin.
One time axis is not sufficient: to counterbalance everything and to cause a a feedback mechanism a 3D time structure is needed in the form of a tetra-eder.

Fig. 1 The carrier field model

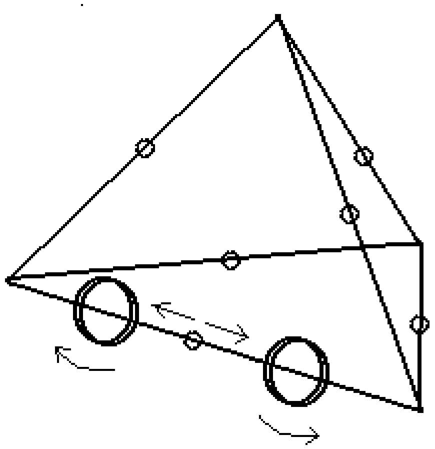

6 time axes form a closed loop. 12 universes are separated by time direction.
The spin of matter is represented by the two flywheels (only shown for one time axis)

This implies that there exist 6 time axes (each with a back and forward movement in time) and thus 12 universes. These universes cannot be seen, but some aspects like the time frequencies can be sensed by our atoms. (BTW, this might explain the missing mass problem in cosmology: our universe still senses the mass of the other 11 universes, that is why 90% of the mass is unexplained..)

The picture shows the arrangement of the time axes. The center of each axis (depicted by a small circle) can be considered as T=0. The two flywheels (drawn on one axis only) represent the spin direction of matter. In our direct anti-universe (on the same time axis) the direction of time is opposite, but also the spin of the matter particles. This means that everything is mirrored, and living beings in the anti-universe experience time as moving forward, just like we. In fact time moves in two directions along an axis, but matter has a sort of rectifying effect - it only picks up the forward going time.

The 60 degrees angle between the time axes is necessary to have an harmonic ratio between our time axis and the time axis of the other universes.

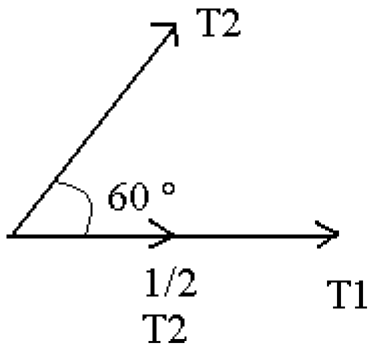

Fig.2 2 time axes at 60 degrees

Only at this angle there can be a harmonic ratio between the time creating frequencies. An atom on our time axis T1 experiences the other time axis at half the frequency.

Our time direction is T1. Another time direction is T2. The angle of 60 degrees makes T1=2*T2, thus giving an harmonic ratio of 1:2.

The carrier field model has several functions, such as:

It is an infinitely strong force field, blocked in itself. It is a loss free conduction medium for all waves. The counter balancing of everything implies that the universes are locked in an iron grip: the spins in one universe are exactly the same and opposite to the spins of the anti universes. It also means that there cannot be one atom more in a universe, which is not present in the 11 other universes. (this does not mean that all the 12 universes have to be the same).

The model is also an image of the universal atom: When applied at a very small scale Denaerde explains the basic building blocks to form matter.

The flywheels can represent the total mass in one universe or just the mass of one atom. The 3D time structure of the pyramid is present in all matter: you could imagine the center of the pyramid being in a star, a galaxy, etc. The 11 other universes are not observable, but they could occupy the same 3D space. The model demonstrates beautifully that matter, time and energy are just properties of one universe (1/12th of the whole). Summing all the properties of the pyramid results in nothing. (The big bang theory fails to explain where the initial energy came from)

The Double rotating central bar.

Denaerde introduces the concept of a magnetic bar, which touches 2 opposite ribs in the pyramid. This induces a flux in the loss free time axis. This is a very important concept and it takes a while to understand it. The bar needs to rotate around two axes simultaneously (Fig. 3) in order to create alternating fluxes in the pyramid ribs (which still represent time axes). The rotating movement repeats itself in 6 time-steps.

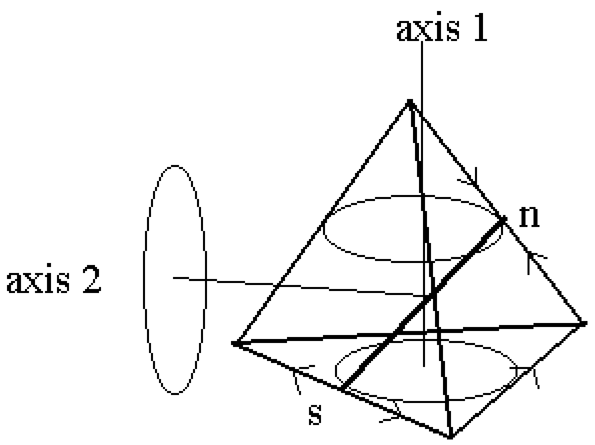

The simultaneous rotation of a central magnetic bar around 2 axes gives alternating flux patterns (time directions) in a pair of ribs. This process is repeated in 6 timesteps.

Fig.3 Double rotation of a central bar

Denaerde calls this configuration a building element. Next we will see that from these building elements one can derive the fundamental elementary particles : electron, quark and neutrino.

The two fundamental building blocks

By adding the initial building block at different time steps (adding three bar rotors) it can be derived that there are two possibilities. These are called central rotors of the first and second kind. (figures 4 and 5). The central rotor of the first kind gives a fixed alternating flux pattern in the ribs of the pyramid. (Try it out!)

Fig. 4 3 building blocks at different time moments are added

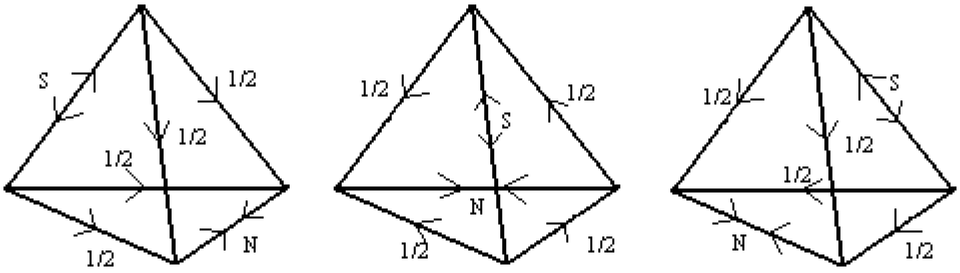
S
1/2
1/2
1/2
1/2
N
1/2
1/2
S
1/2
N
1/2
1/2
1/2
S
1/2
1/2
N
1/2

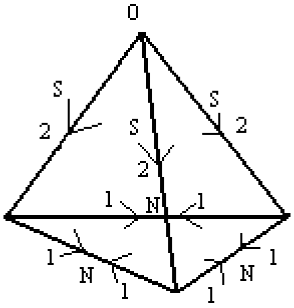
0
S
2
S
2
S
2
1
N
1
1
N
1
1
N
1

The total return flux gives a new building block: a rigid central barrotor of the first kind. The zero node does not change its orientation. In a next timestep the flux pattern is reversed, but the magnitude is the same.

This model describes the ELECTRON.

Fig. 5 The rigid central bar rotor of the second kind

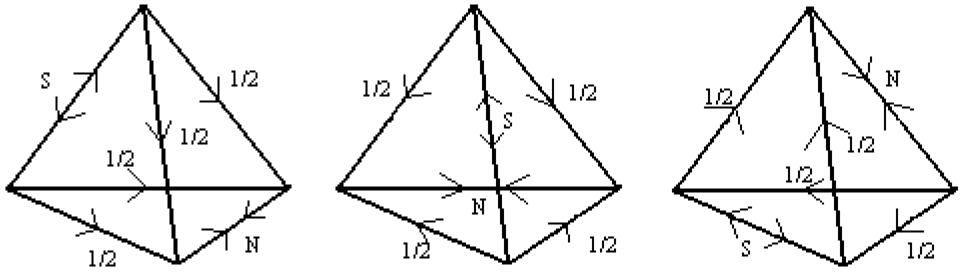

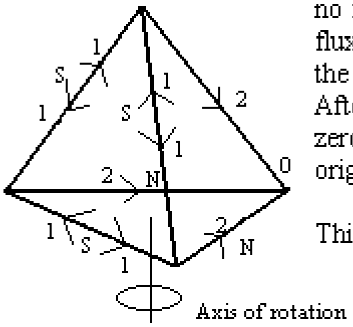

The rotor of the second kind has no fixed zero flux node. The zero flux node rotates through one of the corners of the pyramid.
After 6 timesteps the flux and the zero node are identical to the original position.

This model represents the QUARK.

The second possibility is called a central rotor of the second kind.

The rotor of the first kind is similar to the ELECTRON.
The rotor of the second kind is similar to the QUARK.

Don´t forget, the pyramid ribs represent time axes. (Note: Denaerde uses capitals to distinguish from the real particles in 3D space.)

Three QUARKS added together give a PROTON. (Figure 6)

Fig 6 Quarks, electron and H atom

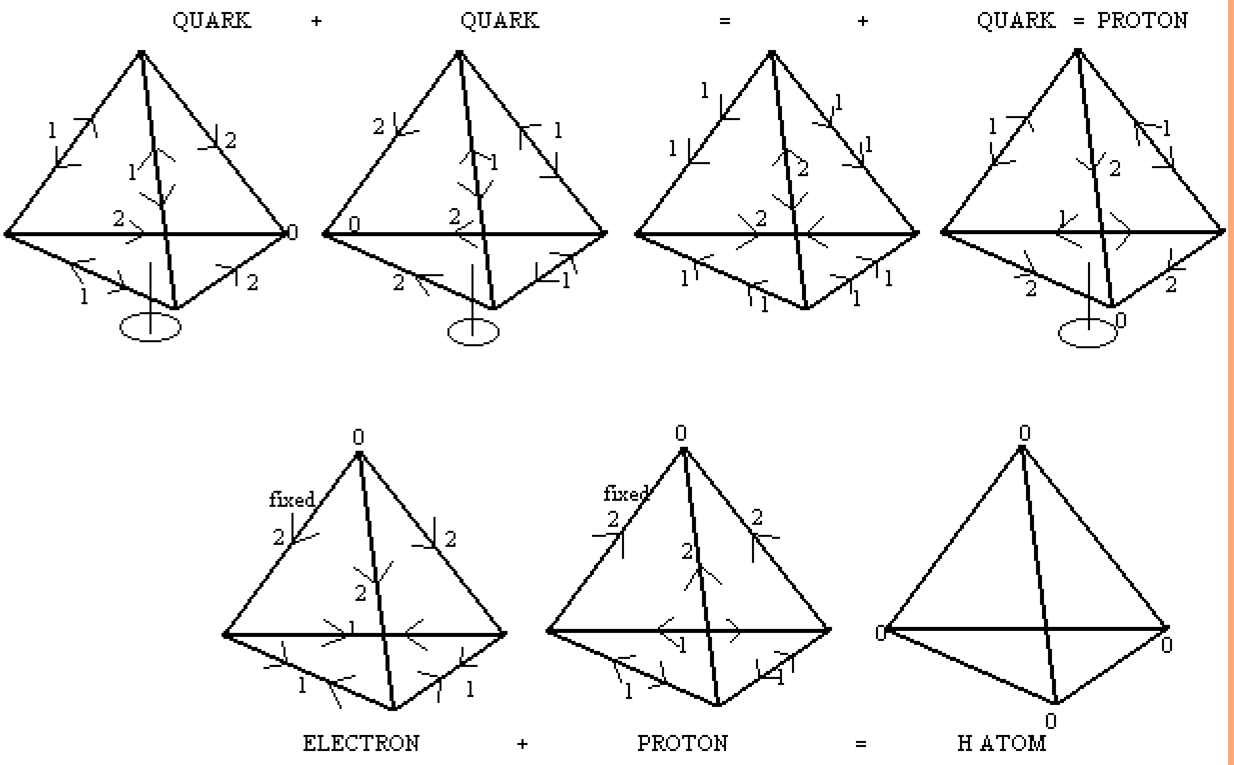

A QUARK and an "anti time" QUARK or the ELECTRON and the "anti time" ELECTRON (positron) can be distinguished by the c.w. or anti c.w. rotation of the central rotors. Dipoles of opposite rotation destroy each other.
Fig 6 shows that adding the PROTON and the ELECTRON gives the neutral H-ATOM.
The mass difference between electron and proton can be explained by different rotation frequencies of the dipoles.

Denaerde then continues to describe several NEUTRON groups which can be formed. This is rather complicated and requires a lot of pictures to visualize, so I would rather skip this. (For anyone interested, I strongly recommend to order the CD!)
The outcome is, that the NEUTRON in the Carrier Field model has 4 QUARKS.
This gives a NEUTRON with a zero external flux. The model explains why the neutron is not stable in free space, while the electron and proton are.

It can easily be seen that to the CF model a QUARK has a unit of charge of 1/3 of the ELECTRON, if charge is defined as the external flux in the pyramid ribs. Charge is related to a symmetry axis of the pyramid. The CF model predicts that charge is not constant in time and that there are two types of charge for a matter particle (e.g. electron): North and South charge. Due to the time switching mechanism we don´t detect this.

A NEUTRON consisting of 4 central rotors of the 2nd kind can be converted into a PROTON (3 rotors of the 2nd kind) and an ELECTRON (1 rotor of the 1st kind).
What is left over is the NEUTRINO : a free dipole in the carrier field.

In order for the particles to exist in space and time, a grid is needed which carries the complex flux patterns. Denaerde comes to a grid consisting of two pyramids stacked together. The top one is inverted and rotated 60 degrees, while its top touches the base of the other pyramid (sunken). The size of the pyramids is in the order of h (Plank´s constant, normalised to length). All particles have more or less the same size in parametric space (space as we know it).

Further Implications

We come now to the more profound meanings of the model. This subject I describe very condensed (without further proof, although the subject is described in great detail in the appendix).

The double rotating central bar has an additional degree of freedom. This results in a second gravitational field, which we cannot directly observe. So there are two kinds of gravity.
The CF model talks about ideal Black Holes and White Radiators. Black Holes have a function to stabilize the universes. As said before, the mass (energy) in all 12 universes must be exactly the same. Still the mass distribution can be different, but only thanks to black holes. B.H´s and W.R´s are present in the corners of the pyramid frame. In one time step, our time direction faces the Black Hole, while our anti-universe´s time direction faces the White Radiator. The next time step our time direction faces the White Radiator. In the next time step it is the null-energy flux.
White Radiators are generators of mass and energy, but have no mass of their own. It generates the exact opposite type of gravity which nullifies with normal gravity.
A pair of BH and WR belongs to each reality. The ideal Black Holes (and there counterpart the WR) in the CF model must have spin. The normal black hole as known from General Relativity Theorie is discussed also.

The formation of a pair of BH and WR (also called a macroscopic quantum mirror or black mirror) is then proven to be a non natural object : it could only have formed by an act of intelligence. Intelligent life is a necessary condition in the formation of the universe! According to the CF model, the black mirrors are formed by the collective mind power of an intelligent race.

Another important consequence is that the time history (knowledge) of the universe is laid down in the quantum gravitation fields. This is one of the reasons why the universe expands: to record the time history. In other words, the time history of a matter system, determined by an increase of entropy, is converted into information in the quantum gravity field. Intelligent life interferes with and modifies the quantum gravity fields (creativity).

The two types of gravity fields represent the matter-bound and emotional (or soul-bound) energy fields. The creation of the macroscopic quantum mirror is the final goal of an intelligent race. It is called the all-might manifestation of the race and it can only be initiated by enough minds working together to create the mirror - which means nothing less than the recognition and identification of the Love poles which carry the universe...

Note:

Everything written here is in my own words. Pictures were made by me. Therefore, copyright has not been violated.

Back

The Lacerta Files

Reptoids of the Inner Earth.

Reptoids of the Inner Earth

The Lacerta Files

Jimmy Bergman, Owner, Sweden.

Published in this website on December 26, 2004.

Note of the Editor of Brother Veritus' Website: These English translations were edited to correct English grammatical errors but message and intent were left intact. The word "Ilojiim" or "Illojim" used extensively in these texts could mean "Elohim", since this is the equivalent traditional word used in Western culture, however, this term may indicate a different galactic race. The illustrations are not of Lacerta.

Lacerta File I

Translation by Chris Pfeiler, Editor and Translator. Taken from http://www.sabon.org/reptiloid/index3.html .

Introduction

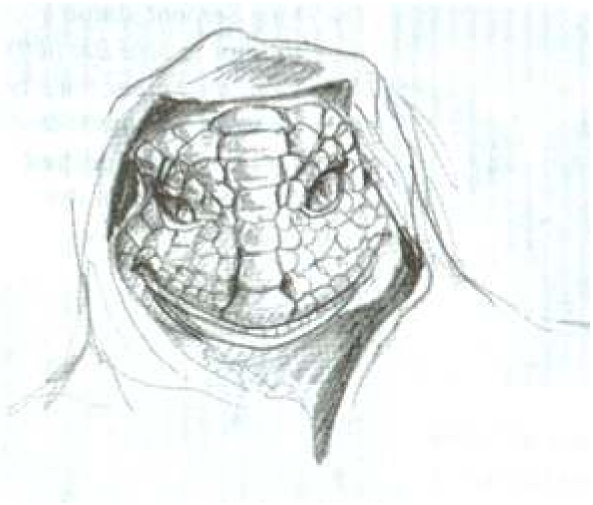

I certify that the following text is the absolute truth and no work of fiction. These are parts of a transcript of an interview I've made with a non-human and reptilian being in December 1999. This female being was already in contact with a friend of mine (whose name is given only with the abbreviation E.F. in the text) since some months. Let me declare, that I was all my life a skeptic about UFOs, aliens and other weird things and I thought that E.F. tells me just dreams or fictitious stories when he talked with me about his First Contacts with the non-human being "Lacerta". I was still a skeptic when I met this being on December 16 last year in that small warm room in the remote house of my friend near to a town in the south of Sweden, despite the fact that I saw now with my own eyes that she was not human. She has told and shown me so many unbelievable things during that meeting that I can't deny the reality and the truth of her words any longer. This is not another of that wrong UFO papers which claim to tell the truth but tell in fact just fiction, I'm convinced that this transcript contains the only truth and therefore you should read it.

I had talked with her for over 3 hours, so the following transcript shows you only shortened parts of the interview, because she asked me after the interview not to publish everything she had told me already now. The order of the questions in this transcript is not always the same order in which I had asked them, so it may seem sometimes a little bit confusing to you. It was not easy to delete all the important parts she had asked me to delete from the transcript, so I apologize for the maybe unusual order.

I’m in the possession of the entire transcript of the interview (49 pages with some of my drawings of her body and her equipment) and also of some tapes on which I have the full interview, but I will not reveal this before I have permission from her.

I will send this shortened form of the still fascinating document to four of my reliable friends to Finland, Norway, Germany and France and I hope they will translate it into their own languages and into other languages and I hope as many people as possible will be able to read and to understand the transcript. If you receive it, please send it to all your friends via e-mail or make print-outs and copy them.

I certify furthermore, that various "paranormal" abilities of her species like telepathy and telekinesis (including the moving and dancing of my pencil on the table without touching and the flying of an apple around 40 centimeters over her hands) were shown to me during the 3 hours and 6 minutes of the meeting and I’m absolutely sure that these abilities were no tricks. The following is certainly difficult to understand and to believe for someone who hasn’t experienced it, but I was really in contact with her mind and I’m now completely sure that everything she said during the interview is the absolute truth about our world.

Unfortunately, if I read the entire transcript and (much more) this very shortened form by myself I have the strong impression, that everything I’ve written sounds too unbelievable to be true, that everything sounds more like a bad science fiction story from TV or cinema and I have doubts that anyone will believe my experiences. But they are true, if you believe it or not. I can’t expect from you that you believe my simple words without evidence, but I can’t give you that evidence. Please read the transcript and think about it and you will maybe see the truth in these words.

There will be a new meeting between me and her (again in the same house in Sweden) on April 23 2000 and she promised me to give me maybe some evidence for her existence. In the meantime I collect questions which I will ask her then. Maybe she gives me permission to reveal more of the missing parts in that transcript and about the coming war.

Believe it or not, this makes no real difference (but I hope you will believe.)

Ole K.

Transcript of Interview (Shortened Version)

December 16, 1999

Question: First of all, who are you and what are you? Are you an extraterrestrial species or can your origin be found on this planet?

Answer: As you could see with your own eyes, I’m not a human being like you and to be honest I’m no real mammal (despite my partly mammal-like body features, which are a result of evolution.) I’m a female reptile being, belonging to a very old reptilian race.

We are the native terrans and we live on that planet since millions of years. We are mentioned in your religious writings like your Christian Bible and many of the ancient human tribes were aware of our presence and worshipped us as gods, for example the Egyptians and the Inca and many other old tribes. Your Christian religion has misunderstood our role in your creation, so we are mentioned as "evil serpent" in your writings. This is wrong. Your race was genetically engineered by aliens and we were just the more or less passive visitors of this accelerated evolution process. You must know (some of your scientists have already supposed this) that your species had evolved in a naturally completely impossible speed within just 2 - 3 millions of years. This is absolutely impossible, because evolution is a much slower process if it's natural but you have not understood this. Your creation was artificial and done by genetic engineering, but not by us but by an alien species. If you ask me, if I'm an extraterrestrial, I must answer no. We are native terrans. We had and have some colonies in the solar system, but we originate on this planet. It's in fact our planet and not yours - it was never yours.

Question: Can you tell me your name?

Answer: This is difficult, because your human tongue is not able to pronounce it correctly (and a mispronunciation of our names is very offensive for some of my kind.) Our language is very different from yours, but my name is —I will try to say it smoother by use of your human letters— something like "Sssshiaassshakkkasskkhhhshhh" with a very, very strong pronunciation of the "sh" and "k" sounds. We have no forenames like you but only a single but unique name which is divided and characterized by the way of speaking and which is given not to children (who have an own children-name) but only in a special procedure in the adolescent age at the time of either religious or scientific "enlightenment" or awareness (as you would call it.) I would appreciate it if you don't try to say my real name with your human tongue. Please call me "Lacerta", this is the name I generally use when I'm among humans and talk with them.

Question: How old are you?

Answer: We measure the time not like you in astronomical years and in the revolving of the earth around sun, because we usually live beneath the surface of the planet. Our time measurement depends on periodically returning cycles in the earth magnetic field and according to this (and said with your numbers) I'm today —let me calculate— 57,653 cycles old. I have reached my adult phase and my awareness 16,337 cycles ago (this is a very important date for us.) According to your human time scale I'm around 28 years old.

Question: What is your task? Do you have a "job" like us?

Answer: To say it with your words: I'm a curios student of the social behavior of your species. That's why I'm here and talk to you, that's why I have revealed my real nature to E.F. and now to you and that's why I give you all that secret information and why I will try to answer all the questions on your many sheets of paper honestly. I will see how you react, how others of your kind react. There are so many crazies and liars of your kind on this planet who claim to know the truth about us, about UFOs, about aliens and so on and some of you believe their lies. I'm interested to see how your species will react if you make the truth (which I will tell you now) public. I'm quite sure everyone of you will refuse to believe my words, but I hope I'm wrong, because you need to understand if you want to survive the coming years.

Question: I’ve read your full statement (which you have given to E.F.) about this, but can you give me now just a short answer: are UFOs real flying objects piloted by extraterrestrials or do they belong to your species?

Answer: Some observed UFOs - as you call them - belong to us, but most not. Most of the "mysterious" flying objects in the sky are not technological devices but mainly misinterpretations of natural phenomena your scientists have not understood (like spontaneous plasma flares in the high atmosphere.) Nevertheless, some UFOs are real craft belonging either to your own species (especially to your military) or to other alien species or at last to us (but a minority of sighted craft belongs really to us, because we are generally very careful with our movements in the atmosphere and we have special ways to hide our ships.) If you read a report about a sighting of a metallic bright-gray cigar-shaped cylindrical object with a length of —there are different types— let me say between 20 and 260 of your meters and if this object had made a very deep humming sound and if there were 5 bright red lights on the metallic surface of the cigar (one at the top, one in the middle, two at the end) then it’s likely that someone of you have seen one of our ships and this means that it was either partly defect or that someone of us was not careful enough. We have also a very small fleet of disc-shaped craft, but such UFOs belong usually to an alien species. Triangular UFOs belong generally to your own military but they use foreign technology to build them. If you really want to try to see one of our craft, you should have a look at the skies over the Arctic, the Antarctic and over Inner Asia (especially over the mountains there.)

Question: Have you a special symbol or something like that with which we can identify your kind?

Answer: We have two major symbols representing our species. One (the more ancient) symbol is a blue serpent with four white wings on a black background (the colors have religious meanings for us.) This symbol was used from certain parts of my society, but it is today very seldom —you humans have copied it very often in your old writings. The other symbol is a mystic being you would call a "Dragon" in the shape of a circle with seven white stars in the middle. This symbol is much more common today. If you see one of that symbols on a cylindrical craft I’ve described in my previous answer or on some underground installation, this thing or place belongs definitely to us (and I would advise you to go away from there as soon as possible.)

Question: The seven stars in the second symbol you’ve mentioned, do they mean the Pleiades?

Answer: Pleiades? No. Actually, the seven stars are planets and moons and they are a symbol for our former seven colonies in the solar system. The stars are shown in front of a blue background and the dragon-circle means the shape of Earth. The seven white stars mean Moon, Mars, Venus and 4 moons of Jupiter and Saturn, we had colonized in the past. Two colonies are no longer in use and abandoned, so 5 stars would be more correct.

Question: As you have not allowed me to make photos —what would be very useful to prove your real existence and the truth of this story— can you describe yourself detailed?

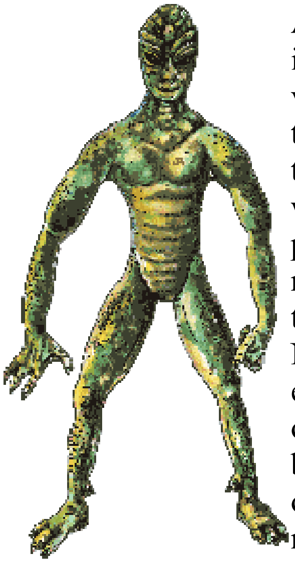

Answer: I know that it would be helpful to prove the authenticity of this interview if you can make some photos from me. Otherwise, you humans are very skeptic (that's good for us and for the real alien species acting secretly on this planet) so even if you had such photos, many of your kind would say that they are fraud, that I'm just a masked human woman or something like that (that would be very offensive for me.) You must understand that I can't give you permission to make photos of me or of my equipment. This have various reasons, which I want not to discuss with you further, but one of the reasons is the keeping up of the secrecy of our existence, another reason is more religious. Nevertheless, you have permission to make drawings of my look and of my equipment I can show you later. I can also try to describe myself, but I doubt that others of your kind will be able to imagine my real look just from simple words, because the automatic denial of the existence of reptilian species and generally of intelligent species other than your own is part of the programming of your mind. Well, I will try.

Imagine the body of a normal human woman and you have at first a good imagination of my body. Like you, I have a head, two arms, two hands, two legs and two feet and the proportions of my body are like yours. As I'm female I have also two breasts (despite our reptile origin, we have started to give milk to our babies during the evolution process —this happened around 30 million years ago— because this is the best thing to keep the young alive. Evolution had done this for your species already in the dinosaur age and —a little bit later— also for ours. That means not that we are now real mammals) but the breasts of us are not as large as those of human woman and the size of them is generally equal for every female of my kind. The external reproduction organs are for both sexes smaller than those of humans, but they are visible and they have the same function as yours (another gift of evolution to our species.)

My skin is mainly of a green-beige colour - more pale green - and we have some patterns of brown irregular dots (each dot of the size of 1 - 2 centimeters) on our skin and in our face (the patterns are different for both sexes but females have more, especially in the lower body and in the face.) You can see them in my case as two lines over the eyebrows crossing my forehead, at my cheek and at my chin. My eyes are a little bit larger than human eyes (for this reason, we can see better in the darkness) and usually dominated from the large black pupils, which are surrounded from a small bright-green iris (males have a dark-green iris). The pupil is slit and can change its size from a small black line to a wide-open egg-shaped oval, because our retina is very light sensitive and the pupil must compare this. We have external round ears but they are smaller and not so curved as yours, but we can hear better because our ears are more sensitive for sonic (we can also hear a wider range of sonic). There's a muscle or "lid" over the ears which can completely close them (for example under water). Our nose is more pointed and there is a V-shaped curving between the nostrils, which enabled the ancestors to "see" temperature. We have lost most of this ability, but we can still feel temperature much better with this "organ". Our lips are shaped like yours (those of females a little bit larger than those of males) but of a pale brown colour and our teeth are very white and strong and a little bit longer and sharper than your soft mammal teeth. We have no different hair colors like you (but there is a tradition to colour the hairs in different ages) and the original colour is —like mine— a greenish brown. Our hairs are thicker and stronger than yours and they grow very slow. In addition, the head is the only part of our body where we have hairs.

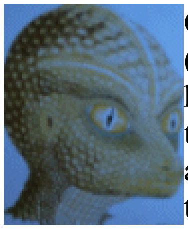

Our body, arms and legs are similar in shape and size to yours, but the color is different (green-beige, like the face) and there are scale-like structures on the upper legs (over the knee) and upper arms (over the elbow). Our five fingers are a little bit longer and thinner than human fingers and our skin on the palm is plain, so we have no lines like you but again a combination of a scale-like skin structure and of the brown dots (both sexes have the dots on the palm) and we have no fingerprints like you. If you touch my skin, you will feel that it is smoother than your hairy skin. There are small sharp horns on the upside of both middle fingers. The fingernails are grey and generally longer than yours. You see that my nails are not so long and round at the top. This is because I'm female. Males have sharp pointed nails with a length of sometimes 5 or 6 of your centimeters. The following feature is very different from your body and part of our reptilian origin: if you touch the backside of my upper body you will feel a hard bony line through my clothing. This is not my spine but a very difficult shaped external plate-structure of skin and tissue following exactly our spine from the head to the hip. There is an extremely high number of nerves and large blood vessels in this structure and in the plates (which are around two or three centimeters long and very touch sensitive - this is the reason why we have always problems to sit in chairs with a back like this chair.) The main task of these small plates (beside a role in our sexuality) is simply the regulation of our body temperature and if we sit in natural or artificial sunlight, these plates become more blood-filled and the vessels become wider and the sun is able to heat up our reptoid blood (which circulates through the body and through the plates) for many degrees and that gives us a great pleasure.

What else is different from your kind? Oh, we have no navel, because we were born in a different way to your mammal birth. The other exterior differences from your kind are minor and I think I must not mention all now, because most of them are not visible if we wear clothing. I hope the description of my body was detailed enough. I would advice you to make some drawings.

Question: What kind of clothing do you generally wear? I suppose this is not the way you dress normally?

Answer: No, I wear this human every-day-clothing only when I'm among humans. To be honest, it's not very comfortable for me to wear such tight things and it is always a very unusual feeling. If we are in our own home (this means in our subterranean home) or in our large artificial sun areas and if we are together with others near to our own name, we are usually naked. Is this shocking for you? When we are in the public and together with many others of my species we wear very wide and soft clothing made of thin, light stuff. I have told you that many parts of our bodies are very touch sensitive, mostly the small back plates so we can't feel comfortable in tight clothing because it can hurt us. Man and woman wear often the same kind of clothing, but the colors are different for the sexes.

Question: You've said "others near to your own name". Do you mean your family?

Answer: No, not really. You would call it "family" but with this word you mean only those of your kind which belong genetically together like father or mother and child. As I have said earlier we have a very difficult and unique name. Part of the pronunciation of that name is absolute unique and there is no other being with the same name, but part of this name (the middle part) is pronounced in a way that told the others to which "family" (I must use the word, because you haven't the right one in your vocabulary) you belong.

This means not that all in that group are genetically related to you, because these groups are usually very large and contain between 40 and 70 of us. This group includes generally your genetic relations - except one of them had decided to left this group - and your connection with father and mother is often the strongest. It would be too difficult for me to explain you now our very old social system which is very complex and we would need many hours only for the primary things. Maybe we can meet another time and I can give you detailed descriptions of all these things.

Question: Have you a tail like normal reptiles?

Answer: Do you see one? No, we have no visible tail. If you look at our skeleton, there is only a small rounded bone at the end of our spine behind the pelvis. This is a useless rudiment of the tail of our ancestors, but it is not visible from the outside. Oh, our embryos have tails during the first months of development, but these tails disappear before they were born. A tail makes only sense for a primitive species which tries to walk on two legs and must held the balance with the tail, but our skeleton had changed during evolution and our spine is nearly in the same shape as yours, so we need no tail to stay on two feet.

Question: You said that you were born in a different way to us. Do you lay eggs?

Answer: Yes, but not like your birds or primitive reptiles. Actually, the embryo grows in a protein liquid inside the mother's womb, but there is also an egg-shaped but very thin chalk hull around it, that fills the whole womb. The embryo inside this hull is completely autarky from the mother's body and it has every substance it needs to develop inside this chalk hull. There is also a cord like your navel cord which is connected to a point hidden behind the back plates. When the baby is going to be born, the whole egg is pressed through the vagina covered in a slimy protein substance and the baby came out of this soft egg after some minutes. These two horns on our middle fingers were instinctively used from babies to break through the chalk hull to take their first breath. Our young are not so large as your babies when they were born, they are between 30 and 35 of your centimeters tall, the egg is around 40 centimeters tall (this is because our vagina is smaller than a human one) but we grow to a normal size of 1,60 to 1,80 meters.

Question: What about your body temperature? You've said that you enjoy to lay in the sun. What effect has this to your organism?

Answer: We are no mammals and as reptiles our body temperature depends on the temperature of our surrounding. If you touch my hand you will maybe feel that it is colder than yours, because our normal body temperature is around 30 to 33 degree Celsius. If we sit in the sun (especially naked and with our row of small back plates in the sun) our body temperature can rise for 8 or 9 degrees within minutes. This rise causes a production of many enzymes and hormones in our body, our heart and brain and every organ becomes more active and we feel than very, very good. You humans only enjoy being in the sun but for us it is the greatest pleasure you can imagine (maybe like your sexual excitement.) We also enjoy swimming in very warm water or other liquids to rise our body temperature. If we are for some hours in the shadow, our temperature goes back to 30 to 33 degrees. This can cause no harm to us, but we feel much better in the sun. We have artificial sun-rooms in the underground but this is not the same for us like the real sun.

Question: What do you eat?

Answer: Generally various things like you: flesh, fruit, vegetables, special kinds of fungus (from subterranean farms) and other things. We can also eat and digest some substances which are poisonous for you. The main difference between you and us is that we must eat flesh, because our body needs the proteins. We can't live completely vegetarian like your kind because our digestion would stop working and we would die after some weeks or maybe months without flesh. Many of us eat raw flesh or other things which would be disgusting for you. Personally, I prefer cooked flesh and surface fruits like apples or oranges.

Question: Can you tell me something about the natural history and evolution of your species? How old is your species? Have you evolved from primitive reptiles as mankind has evolved from apes?

Answer: Oh, this is a very long and complex story and it sounds certainly unbelievable to you, but it's the truth. I will try to explain it in short. Around 65 million years ago, many of our unadvanced ancestors from the dinosaur race died in a great global cataclysm. The reason for this destruction was not a natural disaster - an asteroid impact as your scientists believe falsely - but a war between two enemy alien groups that took mainly place in the orbit and high atmosphere of your planet. According to our limited knowledge about the early days this global war was the first alien war on planet earth but it was definitely not the last (and a future war is coming soon, while a "cold war" - as you call it - between alien groups is ongoing since the last 73 years on your planet.) The opponents in this 65 million year old war were two advanced alien species, whose both names are again not pronounceable for your tongues. I'm able to say them but it would hurt your ear if I tell you the names in their original way. One race was humanoid like your species (but much older) and was from this universe, from a solar system in the star constellation you call "Procyon" today in your maps. The other species - about which we know not so much - was a reptilian species, but they have nothing to do with our own species, because we have evolved from local saurians without exterior influence (except the successful manipulation of our own genes by us. More about that later.) The advanced reptilian species came not from this universe but from a - well, how should I explain it to you. Your scientists have not really understood the true nature of the universe, because your illogical mind is not able to see the easiest things and relies on wrong mathematics and numbers. This is part of the genetic programming of your kind to which I will come later. Let me say, that you are nearly as far away from the understanding of the universe as you were 500 years ago.

To use a term you will maybe understand: the other species came not from this universe but from another "bubble" in the foam of the Omniverse. You would call it maybe another dimension, but this is not the right word to describe it correctly (by the way, the term dimension is generally wrong in the way you understand it.) The fact you should remember is, that advanced species are able to "walk" between bubbles by use of - as you would call it - quantum technology and sometimes in special ways only by use of their mind (my own species had also advanced mental abilities in comparison to your species, but we are not able to do the matter-string/bubble changing without technology, but other species active on this planet are able and this looks to you like magic as it had to your ancestors.)

Back to our own history: the first species (the humanoids) had reached Earth around 150 years before the reptilians and they built some colonies on the former continents.

There was a large colony on the continent you call "Antarctica" today and another one in the continent you call "Asia" today. These people lived together with animal-like saurians on the planet without problems. When the advanced reptilian species arrived in this system, the humanoid colonists from "Procyon" tried to communicate peacefully, but they were not successful and a global war started within months. You must understand that both species were interested in this young planet not for his biology and undeveloped species, but for only one reason: raw material, especially copper. To understand this reason, you must know that copper is a very important material for some advanced species (even today) because it is - together with some unstable materials - able to produce new stable elements if you induce a high electromagnetic field in the right angle with a high nuclear radiation field to produce an over-crossing of fluctuating fields. The fusion of copper with other elements in such a magnetic/radiation field-chamber can produce a force field of special nature that is very useful for various technological tasks (but the base for this is an extremely complex formula you are not able to discover because of the restrictions of your simple mind.) Both species wanted to have the copper of Planet Earth and for this reason they fought a not very long war in space and orbit. The humanoid species seemed to be successful during the first time, but in a last battle the reptilians decided to use a mighty experimental weapon - a special kind of fusion bomb which should destroy the life forms on the planet but should not harm the valuable raw materials and the copper. The bomb was fired from space and detonated at a point of your planet you call "Middle America" today. As it detonated in the ocean, it produced an unpredictable fusion with hydrogen and the effect was much stronger than the reptilians had expected. A deadly radiation, an over-production of fusion-oxygen, a fall-out of different elements and a "nuclear winter" for nearly 200 years were the result. Most of the humanoids were killed and the reptilians lost their interest on the planet after some years for (even for us) unknown reasons - maybe because of the radiation. Planet Earth was on its own again and the animals on the surface died. By the way, one result of the fusion bomb was the fall-out of different elements and materials created in the burning process and one of those materials was Iridium. Your human scientists today see the Iridium concentration in the ground as an evidence for an asteroid impact that killed the dinosaurs. That is not true, but how should you know that?

Well, most of the dinosaurs died (not all in the detonation but in the bad things which came after the war, especially in the nuclear winter and in the fall-out.) Nearly all dinosaurs and reptilians were dead within the next 20 years. Some of them —especially those in the oceans— were able to survive for the next 200 to 300 years even in this changed world, but these species also died, because the climate had changed. The nuclear winter ended after 200 years, but it was colder on earth than before. Despite the cataclysm, some species were able to survive: fish (like the sharks), birds, little creepy mammals (your ancestors), various reptiles like crocodiles... and there was a special kind of small but advanced dinosaurs which had developed together with the last large animal-reptilians like the species you call Tyrannosaurus.

This new reptile was walking on two legs and looked at little bit like your reconstruction of an Iguanodon (it originated in this family) but it was smaller (around 1.50 meters tall) with some humanoid features, a changed bone structure, a larger skull and brain, a hand with a thumb which was able to grab things, a different organism and digestion, advanced eyes in the middle of the head like your eyes and most important...with a new and better brain structure. This was our direct ancestor. There are theories that the radiation from the bomb took part in the mutations of the organism of this new breed, but this is not proven.

Nevertheless, this little humanoid-like dinosaur evolved during the following 30 million years (as I have said earlier, a species need generally more time to evolve than you think, if the evolution is not artificially induced like in your case) from an animal to a more or less thinking being. These beings were intelligent enough not to die in the next millions of years, because they learned to change their behavior, they lived in caves instead in the cold nature and they learned to use stones and branches as first tools and the use of fire as help to warm them - especially to warm their blood which is very important for our kind to survive. During the next 20 million years this species was divided by nature into 27 sub-species (unfortunately, former reptilian species were prone to divide themselves in a more or less illogical way into sub-species during the evolution process. You can clearly see this in the unnecessary high number of animal-dinosaur species in earlier times) and there were many (mainly primitive) wars between this sub-species for dominance.

Well, nature was not very friendly to us and as far a we know from the 27 sub-species 24 were extinct in primitive wars and in evolution, because their organism and mind was not developed enough to survive and (as main reason) they were not able to change their blood temperature in the right way if the climate changed. 50 million years after the war and after the end of Dinosaurs, only three (now also technological) advanced reptilian species were remaining on this planet together with all the other lower animals. Through natural and artificial crossbreeding these three species were united to one reptilian species and through the invention of genetic manipulations, we were able to "eliminate" the dividing-prone genes in our genetic structure. According to our history and belief, this was the time when our final reptilian race - as you see me today - was created by use of genetic engineering. This was around 10 million years ago and our evolution nearly stopped at this point (well, actually there were some minor changes in our look toward a more humanoid and mammal-like appearance during the coming ages, but we have not divided again into sub-species). You see, we are a very old race in comparison to your kind, which was jumping around as small monkey-like animals in the trees at this time while we invented technology, colonized other planets of this system, built large cities on this planet (which disappeared without a trace in the ages) and engineered our own genes while your genes where still those of animals.

10 million years ago the small simians started to grow and they came down from the trees to the ground (again because of the change of the climate - especially on the so-called African continent.) But they evolved very slow as it is normal for a mammal and if nothing extraordinary had happened to your kind, we wouldn’t be able to sit here and talk because I would sit in my comfortable modern house and you would sit in your cave clothed with fur and trying to discover the secrets of fire - or you would maybe sit in one of our zoos. But the things had developed differently and you believe now you are the "crown of creation" and you can sit in the modern house and we must hide and live beneath the earth and in remote areas. Around 1,5 million years ago, another alien species arrived at Earth (it was surprisingly the first species since over 60 million years. This would be more surprising for you if you would know how many different species are today here.) The interest of this humanoid species - you call them "Illojim" today - was not the raw material and the copper, it were to our astonishment the unadvanced ape-humanoids. Despite of our presence on this planet, the aliens decided to "help" the apes to evolve a little bit faster, to serve them in the future as some kind of slave-race in coming wars. The fate of your species was not really important for us, but we didn’t like the presence of the "Illojim" on our planet and they didn’t liked our presence on their new "galactic zoo" planet and so your sixth and seventh creation was the reason for a war between us and them.

You can read about that war for example partly in the book you call "Bible" in a very strange way of description. The real truth is a very long and difficult story. Should I continue?

Question: No, not now. I've made some notes about your history and now I have some questions.

Answer: Please ask.

Question: First of all, you handle with a very large time scale. You claim that your primitive ancestors lived together with the dinosaurs, survived the - as you called it - artificial cataclysm and evolved then over 40 million years and your evolution was completed 10 million years ago. This sounds very unbelievable to me. Can you say something to this?

Answer: I understand that this must sound absolutely unbelievable to you, because you are a young and genetically engineered species. Your historical horizon ends at a scale of just some thousands of years and you think this is right. But it isn't. This is impossible. Your programmed mind is obviously not able to handle with such large time scales. Our evolution time may seem incredible long to you, but this is in fact the original way of nature. Remember, your early mammal ancestors developed together with dinosaurs and they survived the bomb like us. They evolved slowly during the next millions of years and they divided into various species and shapes, some of them larger, some of them smaller. This is evolution of the body. But what about their mind and intelligence? They were simple animals. The mammals evolved since —let us say— 150 millions of years, but only in the last 2 - 3 millions of years they were able to became intelligent and thinking. And within this small period beings like you were created. From nature? 148 millions of year's time for the evolution of animal-like mammals, 2 millions of year's time for the development of (more or less) intelligent beings like you? Ask yourself: Do you really think this accelerated evolution is natural? Then your species is more ignorant than I've thought. We have not evolved wrong but you.

Question: I understand. But I have another question. You've mentioned many facts about the ancient war between the aliens 65 million years ago. This happened very long before your kind became really intelligent (as far as I have understood you). Why do you know so many things about that "first war" and about the evolution of your species?

Answer: This is a good question (much better than the previous) and I have not explained it properly to you. Our knowledge about the first war comes completely from an ancient artifact, which was found around 16,000 years ago from our archeologists on the continent you call North America today. They found there a round plate with a diameter of approximately 47 of your centimeters. The plate was made of an even for us unknown magnetic material and inside the plate there was another smaller crystal plate which contained an enormous amount of information coded in the molecular structure of the crystal. This "memory plate" was manufactured from the last bomb survivors of human race from "Procyon" already 65 million years ago but it was completely intact when we found it. Our scientists were able to encode the messages and data and so we heard the first time about the events which took place in the distant past and which led to the extinction of the dinosaurs. The plate contained detailed descriptions of both species (but more about the humanoids) and about the events and weapons, including the fusion bomb. It contained also a description of the animals and saurians on earth, including our pre-intelligent ancestor species. The rest of our knowledge about our evolution comes from skeletons and from the back-reading and de/encoding of our DNA.

You see, we know the real truth about our roots since 16,000 years. Before that time, there was a more religious idea of our creation.

Question: What have happened with the both alien species?

Answer: We don’t know exactly. The surviving humanoids on earth obviously died in the years after the bomb and others of their kind and the reptilians never came back to Earth (as far as we know). Concerning the reptilian aliens, there is a possibility that it was physically impossible for them to return, because the matter between bubbles is sometimes in rapid movement. The current theory is, that both species had ceased to exist during the millions of years.

Question: You’ve mentioned skeletons of your kind. How can it be, that human scientists haven’t found any trace of you and your ancestors if you really live for such a long time on this planet? We have found many skeletons of primitive dinosaurs, but none of an advanced reptilian being with a larger skull and brain and a hand with a thumb as you have described it before.

Answer: Yes, you have. But your "great" scientists were not able to reconstruct the skeletons completely, because they wanted to reconstruct reptilian animals, not intelligent beings. You would laugh if you would know how many of the (especially small) saurian skeletons in your museums are totally wrong constructions of never-existing beings, because you used many bones which didn’t really belong together and sometimes you made artificial bones if something was missing you needed to construct an "animal" saurian. Many of your scientists are aware of this problem, but they don’t make it public, because they can’t explain it and they claim, that the right bones were just missing and their reconstruction is right. Many bones of us were used for Iguanodon reconstructions, for example the hands with the visible thumb (look at an Iguanodon in a museum and you will see that I’m right.) A scientist in the country you call United States had built a nearly correct skeleton of our kind some years ago, but the local government (which is partly aware of our existence) confiscated the reconstruction. As we live today (and since thousands of years) nearly completely beneath the earth, you will not find any cadavers or skeletons of us.

Question: You speak sometimes about underground cities and artificial sunlight. Do you mean something like a "hollow Earth" with this? Is there a second sun inside our planet?

Answer: No, Earth is not really completely hollow and there is no second sun inside. This story is ridiculous and physically not possible (even your species should be intelligent enough not to believe this.) Do you know how much mass a sun must have to produce energy and light for a longer time by fusion? Do you really think that there could be a small active sun inside the planet? When I talk about our subterranean home, I talk about large cave systems. The caves you have discovered near to the surface are tiny in comparison to real caves and huge caverns deeper in the earth (in a depth of 2,000 to 8,000 of your meters, but connected with many hidden tunnels to the surface or to surface-near caves) and we live in large and advanced cities and colonies inside such caves. Major sites of us are beyond the Arctic, the Antarctic, Inner Asia, North America and Australia. If I talk about artificial sunlight in our cities I don’t mean a real sun but various technological sources of light (including gravitational sources) which illuminates the caverns and tunnels. There are special cave areas and tunnels with a strong UV light in every city and we use that places to heat our blood. Furthermore, we have also some surface sun places in remote areas, especially in America and Australia.

Question: Where can we find such a surface-near entry to your world?

Answer: Do you really think I will tell you their exact location? If you want to find such an entry, you have to search it by yourself (but I would advise you not to do that.) When I came to the surface four days ago, I used an entry approximately 300 of your kilometers north from here near to a large lake, but I doubt that you would be able to find it (there are only a few entries in this part of the world - more are far more north and east.) As a little advice: if you are in a narrow cave or in a tunnel or even in something that looks to you like an artificial mine shaft and as deeper you walk as smoother appear the walls and if you feel unusual warm air streaming from the depth or if you hear the rushing sound of streaming air in a ventilation or elevator shaft, then look for a special kind of artificial and smooth wall somewhere in the cave with a door made of gray metal. If you would be able to open that door (but I doubt this) you would be in a usually round technical room with ventilation systems and elevators to the depth. This is probably an entry to our world. If you have reached this point, you should know that we are now definitely aware of your presence. You are already in big trouble if you have entered the round room, but you should look for one of the two reptilian symbols on the walls. If there are no symbols or other symbols, you are maybe in bigger trouble as you think, because not every underground installation belongs to our kind. Some new tunnel systems are operated from alien races (including hostile races). My general advice if you find yourself in a, for you, strange underground installation: run away as fast as you can.

Question: You mentioned earlier that you use the name "Lacerta" when you are among humans and that you enjoy it to be in the real sun on the surface of earth. But how can you be among humans? You don't look like us, so anyone will see that you belong to another species. Why have nobody seen and described a being like you if your kind lives already since our "creation" together with us on the same planet. Can you explain that to me?

Answer: First, my kind was of course seen and described (and worshipped) many times in your primitive past, for example in your religious writings like your Christian Bible. You can find descriptions and even simple drawings of us for also in the southern part of the American continent on various temples. So-called "wise" men from India and from the Asian mountains have described our species many times in writings, together with other "wise" men from the African continent. I think we are the most mentioned non-human species (maybe beside the "Illojim") in your history. If you don't believe me, have a look at your history and you will see the truth in my words. Your "great" scientists called the belief in us "superstition" and "religion" and today's "intelligent" humans have forgotten our presence on the surface in the past.

Furthermore, our species is seen even today sometimes from human witnesses in its original shape on earth or in our surface-near entries and tunnel systems, but fortunately you and your media didn't take the reports of such "crazies" serious (that's good for us and that's the reason why we allow those people to see us as we really are.) Some of my species are also in direct contact with human scientists and politicians from the surface, but this is top-secret —as you would call it— and nobody of your public knows anything about it (the matter of these meetings is generally the upcoming war with and between the alien species and our assistance in this war). But there is also another explanation, why we can walk among you and why you are not able to recognize us: mimicry.

The following may again sound unbelievable and even shocking to you, but as you have asked I will explain it. I have told you before, that we have more advanced mental abilities than your species and with "more advanced" I mean, that we are able to use telepathy and telekinesis from our birth on (in fact, mother and new-born child communicate generally with telepathy during the first months) without special training as you need it to activate these sleeping parts of your brain. The structure of our brain is a little bit different to yours and our hypothesis is larger and more active than yours - especially when we are in sunlight. Our own abilities are very strong in comparison to yours, but weak in comparison to the "matter-string/bubble" mind forces of some of the alien species on this planet. I was never very good in that mind things, but we all have these primary abilities and can use them for example for our protection or even for attack.

When we are on the surface and we meet human beings (even a large group of them - this makes no difference. All of your minds are like one mind) we are able to "touch" their mind and induce them via telepathy the command "See us as one of your kind" and the weak human mind will accept this order without re-fusion and they will see us (despite our reptilian look) as normal humans. I've done this many times and you weak humans generally see me as an attractive brown-haired woman, because I have created this special "mimicry image" in my mind years ago and I can induce it into your minds without problems. I've needed some time at the beginning to learn the use of the mimicry correctly, but then it worked nearly automatically and I can even walk among a group of yours and nobody will recognize what I am. There is a simple switch ("See us as we really are / See us as we want you to see us") in your consciousness which was placed there from the "Illojim" when they created your kind and we can use this switch to convince you that you see humans when you look at us (other aliens use this switch, too). It is easier as you think. When there are meetings between your kind and aliens which seem to look exactly like yours, these aliens have used that switch and some of the meetings with man-like aliens can be also explained with meetings with my kind.) When I met E.F. the first time, he saw me also as a normal human woman and I remember that he was very frightened and shocked when I revealed him my real appearance.

Question: Do you mean, that you can really make me belief that I talk now with an attractive brown-haired human woman instead of a reptilian being like you?

Answer: Probably, but I don't think so in your special case. When someone expects to see a human woman instead of me, I can do it without problems with his mind (even with large groups) because nobody expects to see a reptile woman. But I have allowed your mind to see me in my original appearance from our first meeting on and I have never induced something into your mind, so you have already realized that I'm not human. If I would now try to change this, it would probably lead to an absolute confusion or to unconsciousness and I don't want to harm you. As I have said I'm not very good in these things.

Question: That's very scary. Can you kill with those abilities?

Answer: Yes, but it's forbidden. This means not that it was not done in previous times.

Question: Have both sexes these abilities?

Answer: Yes.

Question: What about photos? How do you appear on photos?

Answer: This is a silly question. I appear on photos as a reptile being, because I can't have influence on the photo or on the camera itself but only on the photographers mind. If he or she would develop the film and show the photo to others, they would see me in my original shape. That's the reason why it is forbidden for our kind to be filmed or photographed and we must avoid every camera on the surface (that is very difficult and we were filmed sometimes in the past without our knowledge, especially from certain of your governments and secret agencies.)

Question: What other commands can your kind induce into our minds? Something like "Serve us" or "Obey"?

Answer: This is again a strange question. We are not your enemy (most of us not) so why should we do this? To answer your question: it depends on the strength of the human mind and on the strength of the sending reptilian. There is no "Serve us" or "Serve me" switch in your mind, so such a command is much more difficult to induce. If the human mind and consciousness is weak and the reptilian inducer is experienced in these things and was some hours in the sun before he or she tries to do it, then it could probably work for a certain time. There are secret teachings about such things, but I've never learned anything about it. I use my primary abilities for mimicry and for communication with my own kind and sometimes for other private things, but I've never used it to harm humans or their mind. I would appreciate it if we can end with this topic here.

Question: A last question: you've said earlier, that you can hide your UFOs? Do you use the same abilities to do this?

Answer: Yes, but on a technical base. There is a powerful device inside each craft which is able to send an artificial signal to your minds to convince you, that you see either nothing but only the sky or that you see normal aircraft like planes instead of our ships. This isn't used very often, because we avoid human public when we move in the atmosphere. If you are able to see our "UFOs" it means that the device is either defect or deactivated for some reason. The camouflage effect didn't work on photos - to answer this possible question of you already in advance - but why should someone make a photo of the sky when he could not see anything unusual there. By the way, most of the surface-near entry points to our tunnels are also hidden with such a device and your kind will generally see only normal cave walls instead of the door. That's one reason why I've said that I doubt that you will be able to find such a secret door to our world (but it have happened a few times in the past.)

Question: Back to your and our own history. You've mentioned the race of the "Illojim" who have created our human race. From where did they come and how did they look like? What had exactly happened when they arrived? Are they our "God"?

Answer: The "Illojim" came from this universe, from the solar system you call "Aldebaran" in your maps. They were a very tall humanoid species whit usually blonde hairs and a very white skin (they avoided the sunlight, because it hurt their skin and their eyes. This was absolutely unbelievable for a sun-loving species like us).

They seemed to be intelligent and peaceful at the beginning and we started a more or less friendly communication with them, but later they showed their real intentions and plans: they wanted to evolve the apes to a new breed and we were a disturbing factor for them on their new zoo planet. At first, they caught around 10,000 or maybe even 20,000 of your simian ancestors and they left the planet for some hundred years. When they returned, they brought your (now more human) ancestors back. Then they left Earth again for some thousands of years and the primitive pre-humans lived together with us without major problems (they were just afraid of our aircraft and technology). The "Illojim" had taught their mind and enhanced their brain and their body structure and they were now able to use tools and fire. The "Illojim" returned within 23,000 years seven times and accelerated the evolution speed of certain of your kind. You must understand that you are not the first human civilization on the planet. The first advanced humans (who lived at the same time with less-developed pre-humans, because the "Illojim" had experimented with different speeds and stages of evolution) with technology and speech existed around 700,000 years ago on this planet (your scientists have not understand this, because they've found only the bones of the pre-humans and some primitive cave drawings showing advanced humans and flying devices.) This genetically advanced human breed lived together with us, but they avoided contact with my kind, because the "Illojim" teachers had warned them with misleading purpose that we are evil beings and that we lie to them.

Well, after some centuries the aliens decided to extinct their first creation and they accelerated the evolution of a second and better test series and so on and so on. The truth is that your modern human civilization is not the first on this planet Earth but already the seventh. The buildings of the first breeds are lost, but the fifth civilization was the one, which built the large triangular constructions you call "Egyptian Pyramids" today around 75,000 years ago (your Egyptians just found that large ancient pyramids in the sand and tried not very successful to built similar constructions) and the sixth civilization was the one, which built the cities which ruins you can find today beneath the sea in the so-called Bimini Area around 16,000 years ago. The last creation of the seventh breed - of your series - was done just 8,500 years ago and this is the only creation you can remember and to which your religious writings refer. You rely on archeological and paleonthological artifacts which show you a wrong and short past, but how should you know anything about the six civilizations before. And if you find evidence for their existence, you deny and misinterpret the facts. This is partly a programming of your mind and partly pure ignorance. I will tell you in the following only about your creations, because the six previous mankinds are lost and therefore they should not concern you.

There was a long war between us and the "Illojim" and also between certain groups of the "Illojim" themselves, because many of them were the opinion that the again-and-again creation of human species on this planet makes no real sense. The last battles in this war were fought around 5,000 years ago in orbit and surface. The aliens used powerful sonic weapons to destroy our underground cities but on the other hand we were able to destroy many of their surface installations and bases in space.

The humans of your series were very frightened when they observed our battles and they wrote it down in form of religious myths (their mind was not able to understand what was really going on.) The "Illojim" - who appeared as "gods" for the sixth and seventh breed - told them that it is a war between good and evil and that they are the good and we are the evil race. This depends certainly on the point of view. It was our planet before they arrived and before they started their evolution project with your kind. In my opinion, it was our right to fight for our planet. It was exactly 4,943 years ago - according to your time scale - that the Illojim left the planet again for unknown reasons (this is a very important date for us, because many of our historians called it a victory.) Fact is that we don't know what had really happened. The "Illojim" were gone from one day to another, they vanished without a trace together with their ships and we found most of their surface installations destroyed by them. The humans were on their own and your civilization developed. Many of us were in contact with certain (more southern) tribes of your species in the coming centuries and we were able to convince some of them that we are not the "Evil" the aliens wanted them to believe. During the time from 4,900 years ago to today, many other alien species arrived the planet (some of them used the old teaching and programming of your mind and "played" again God for you) but the "Illojim" themselves never came back. They had left the planet for a duration of some thousand years also earlier, so we expect their return one day in future to end their project or to maybe extinct also the seventh breed, but we don't really know what have happened to them (to answer this question of you in advance).

Your current civilization doesn't know anything about your real origin, about your real past, about your real world and universe and you know very little about us and our past. And you know nothing about the things to come in near future. As long as you will not understand and believe my words - I tell you the truth because we are not your enemy - as long there is danger for your species. Your enemies are already here and you have not understood. Open your eyes or you will be in big trouble soon. If you haven't believed anything of the things I've told you before, then you should really believe and remember this.

Question: Why do you think I don't believe you?

Answer: I have a certain feeling that you don't believe me, despite the fact that I'm sitting here in front of you. Everything I have told you in the last two hours is the absolute truth about our world.

Question: How many alien species are active on earth at the moment?

Answer: As far as we know 14 species. 11 from this universe, 2 from another "bubble" and 1 very advanced from a very different plain. Don't ask me for names, because nearly all are not pronounceable for you, eight of them are not pronounceable even for us. Most of the species - especially the more advanced - are just studying you as animals and they are not very dangerous for you and for us and we work together with some of them, but three species are hostile, including the one which was in contact with some of your governments and exchanged their technology for copper and other important things and which had betrayed your kind. There was and is a "cold war" between two of these hostile races during the last 73 years and the third species seemed to be the "winner" in this useless struggle. We expect a more "hot" war between them and you in the near future (I would say in the next 10 or 20 years) and we are worried about that development. In the last time, there were some rumors about a new, fifteenth species which had arrived on Earth just 3 or 4 years ago, but we don't know anything about their intentions and we were not in contact with them till now. Maybe the rumors are wrong.

Question: What do the hostile alien races want?

Answer: Various raw materials, including copper for their technology, your water (or better the hydrogen in your water, which is a source of energy in advanced fusion processes) and certain chemical elements in your air. Furthermore, two of the species are also interested in your body, in your human tissue and blood, because their own genetic structure is defect through bad evolution and radiation (as far as we know) and they need intact strings from your kind and from animals to repair their own genetic again and again, but they are not really able to repair the defects completely because their DNA and your DNA is not fully compatible (my own species is absolutely incompatible with them, so they are not very interested in us) and they try to make more compatible crossbreeds between you and them by use of artificial fertilizations and artificial wombs. We suppose that the coming war between the three races or between you and one or all of them will be fought for raw material, hydrogen, air and DNA.

Question: Is this the reason for the "abductions"?

Answer: Partly, especially when the aliens took egg and sperm samples from you. Sometimes the abductors belong to another and more advanced race and they just want to study your body and your mind (which is more interesting for some of them than your solid body) as you would study a primitive animal. As I have said, three alien species are hostile and this means that they do not care for your fate or for your life and people who were "abducted" by them came very rarely back alive. If someone is able to report about an abduction, it means in my opinion that he or she has not met one of the aggressive species or that he or she is a very, very lucky human to be alive. Advanced and "friendly" races also took sometimes egg and sperm samples, but for other reasons.

Question: You've said there are only 14 species active on earth. But why describe people who saw alien beings so many different and bizarre types of them?

Answer: I think I have already answered to this question. As I have said, most of the alien races have much more advanced mind abilities than you or even me (there is just one alien race completely without such abilities). They are able to appear in your mind and memory as whatever they want and this induced "image" has nothing to do with their real appearance. You remember them as normal humans or grey dwarfs or even extremely bizarre animals because they want you to remember that or sometimes they want you to completely forget anything about a meeting with them. Another example: you can for example remember that you were just in a normal of your human hospitals and that some doctors were examining you and you think not further about what have happened to you (maybe till you discover that there is no hospital in the street were you supposed it) but in fact you were examined by them in one of their laboratories. You can't rely on your mind in this case. They appear in different shapes to you to confuse you and to make so-called abduction witnesses who were able to remember the events - or who believe they are able to remember - ridiculous in the public and as far as we know, they are successful. Believe me, there are only 14 alien species on this planet and only eight of them abduct humans at the moment (again as far as we know.) In addition, not everyone of your "abductees" is one and some of the aliens in their reports are really just imagination or lies.

Question: How can we protect us against this influence on our mind?

Answer: I don't know. I doubt you can, because your mind is like an open book to read and write for nearly every species I know. This is partly the guilty of the "Illojim" themselves, because they had constructed or better mis-constructed (partly intentionally) your mind and your consciousness without real protection mechanisms. If you are aware that someone tries to manipulate your mind, you can only concentrate on that suspicion and try to analyze every one of your thoughts and memories. Very important: don't close your eyes (this would lead to a different form of brain waves which are more easy to access) and don't sit or lay down to rest. If you stay awake during the first minutes, you can maybe try to filter the other thoughts and waves in your brain and the inducer will give up after some minutes if he or she is not successful because it will start to hurt his or her own brain. This is very difficult and certainly painful and it can harm you, so better don't try to resist but it would be the only possibility you have. However, you can try this only with the more weaker species, not with the strong.

Question: What do you mean with "one species comes from a very different plain"?

Answer: Before I can explain that correctly to you, you must be able to understand the universe and this would mean a maybe useless teaching of your mind (including the remove of some barriers) of many weeks and with teaching I mean not only teaching by words. I have said this with your word "plain" or "level" because you have again no better word in your vocabulary and dimension would be in this case absolutely wrong (it's rather wrong even for another "bubble") because a dimension can't exist without plains. If you would be a species living in another or over the plain and if you would be furthermore able to enter plains without technology so that your body is not made of that kind of matter you know, then you would be the mightiest being you can imagine. This very advanced race I've mentioned had developed outside of here and they've evolved in fact over billions of years. They would be able to destroy all of you and us and everything with just a single thought. We were in contact with them only 3 times in our whole history, because their interest in your planet is different from that of all other races. They are definitely no danger for you or us.

Question: What will happen when the war begins?

Answer: This is difficult to answer. It depends on the enemy race and on their tactic. "War" is not always that primitive thing you humans mean with the word, "War" can be fought on various levels. One possibility they have is the "destruction" of your social system by influence on political leaders, another is the use of advanced weapon systems which can cause earthquakes or volcanic eruptions or other disasters (including weather disasters) which may seem natural to you. The special fields from copper-fusion I've mentioned earlier are able to have an influence on your global weather. I think they will not attack the planet directly before the human civilization is weak, because even you have possibilities to destroy their craft (but not many.) Let me say, that we are not absolutely sure if there will be really such a "hot" war already in the next years. I don't want to talk further about this.

Question: This is the end of the interview. Do you want to say a last sentence or message?

Answer: Open your eyes and see. Don't believe only in your wrong history or your scientists or your politicians. Some of them know the truth about various things, but they don't inform the public to avoid confusion and panic.

I think your species is not as bad as some of my kind thinks and it would be a pity to observe your end. That's everything I can say. Go through your world with open eyes and you will see - or maybe not. Your kind is ignorant.

Question: Do you think anyone will believe that this interview is the truth?

Answer: No, but it is an interesting experiment for my social studies. We will meet again in some months and you will tell me then what have happened after the publication of my message. Maybe there is hope for your kind.

All Rights Reserved ©

The Lacerta File II

Translation by Doug Parrish. Taken from http://www.sabon.org/reptiloid/index4.html .

Introduction

I once again reaffirm that the following text is the absolute truth and is not fiction. It was composed from three original tape recordings which were made on April 24, 2000 with a tape recorder during my second interview with the reptilian creature known as "Lacerta". At Lacerta's request, the original text of 31 pages was revised and shortened up to deal with some questions and answers. Some existing questions were partially shortened or amended. It was even undertaken to extract message and significance from it. This part of the interview, either not mentioned or not mentioned completely in the transcript, deals primarily with personal issues, paranormal demonstrations, the social system of the reptilian species and alien technology and physics.

The reason for the shifting of the date and time of the second meeting was a possible observation and surveillance of my own person after the publication of the first transcript. Although everything was attempted on the advice of Lacerta to keep my identity a secret, just two days after the dissemination of the document abroad, and various unusual events took place. Please don't think that I am paranoid; however, I believe that the publication of the interview has drawn either official attention or the attention of some organization to me. Up until this time, I usually regarded people who believed that they were being followed by the state to be nothing more than jokers. But now I have begun to revise my ideas on that since events in January. It began with a failure of my telephone for several hours. When the phone became operational once again, there were quiet echoes and strange clicking and whirring sounds when I made calls. A defect could (ostensibly) not be found anywhere. Overnight, important data disappeared from the hard drive in my computer. The testing program reported "defective sectors" where strangely enough there were only data which dealt with illustrations and completed textual material from the interview. These "defective sectors" also contained material of a paranormal nature in the field of my research. (Fortunately, the material was also stored on floppies.) In addition I discovered by pure chance some hidden data in a likewise hidden directory index. The name which appeared on the data and the directory index was "E72UJ."

A friend, who is a computer expert, could not make anything of this designation, and when I was about to show it to him, the directory index had disappeared. One evening, my apartment door was standing wide open, my TV set was running —and I am absolutely confident that I had turned the TV set off.

A minivan with British markings and the imprint of a Europe-wide supermarket chain parked in front of my house. I noticed the same minivan again on several occasions traveling at a distance behind my car, even when I visited the town of ...... 65 kilometers away. When I returned, the car was on the other side of the street once again. I never saw anyone get into or out of the car. A knock on the door of the vehicle and on the tinted windows caused no reaction of any kind. After about two weeks, the minivan disappeared again. When I informed E.F. personally about these events, he suggested that I change the place and date of the meeting in order to assure our own and Lacerta's safety. The meeting took place on April 27, 2000 in another isolated location. It was unobserved as far as I can determine.

Once again, all of this may sound strange and paranoid, like a fantasy from a cheap science fiction film; however, I can only repeat to and assure the reader once again: all of this is the unadulterated truth. Believe my words or don't believe them. These things have happened and they will continue to happen, whether you believe it or not. Until it is too late. Our civilization is in danger.

Ole. K. – May 3, 2000

Transcript of the Interview (Shortened Version)

Date of Interview: April 27, 2000

[Comment by Ole. K.: The meeting began with an appraisal of diverse questions and opinions which I had gotten from readers of the first transcript in anonymous fashion through distribution from my trustworthy friends. Some of these opinions—all together there were over 14 pages of paper—contained comments shaped by everything from a radically religious to a fanatical tendency to welcome contact with a reptilian species. Some of these comments contained stereotypical phrases like "Servants of Hell" or "Species of the Evil One". I don't want to go into any kind of detailed description here since I don't want pass on further any false and radical realm of thought.]

Question: When you read these religious and animosity-ridden comments here, what do you think and feel then? Is the relationship between your species and ours really shaped from that kind of total negation?

Answer: Does it amaze you that I am not completely angered by that? I had fully expected those kinds of extreme reactions. The programming for the utter negation of another species (especially the reptilian species) as in your own case is deeply embedded in each of your own individual consciousnesses. This ancient conditioning stems from the days of your third artificial creation and, biologically speaking, is passed down as an information genome from generation to generation. The identification of my species with the powers of darkness was a primary intention of the Illojim, who liked seeing themselves in the role of the powers of light —something which in and of itself represents a paradox, since that humanoid species was extremely sensitive to your sunlight.

In case you were expecting me to act offended, I guess I'll have to partially disappoint you. These obscure intentions are not really your fault; you are simply following for the most part what you have inherited from your ancestors. It is indeed actually somewhat disappointing that many of you develop no especially strong individual self-conscience, for this would help you to overcome the conditioning. As I already said, we were in direct contact in the last several centuries with some of your more primitive human tribes; these tribes had themselves succeeded in breaking through the old "creation programming;" they were able to meet us without tension, hate and total rejection. Apparently many of your modern civilized individuals are not in a position to think on their own, but rather let themselves be guided by programming and religion (which is also a manifestation of that ancient programming and part and parcel of the plan of the Illojim). Therefore, comments of that kind I'd sooner regard as amusing than irritating; they simply confirm in large measure for me my suppositions about your defined mode of thinking.

Question: Therefore, you are not the "Species of the Evil One" as was remarked earlier?

Answer: How am I supposed to answer that? Your people still think according to a simple and completely inappropriate scheme of generalizations. Simply put, there are absolutely NO purely evil species. There exist in every terrestrial and extraterrestrial species alike both good and evil individuals; it's even true of your own people; but there is NO such thing as an absolutely evil species. This conception is really very primitive. You people have believed from time immemorial what you are supposed to believe—what was foreseen for you to believe by your creators. Every well-known species, even the more highly developed ones, consists of a great number of individual consciousnesses (at least a portion of the consciousness is individual, even though there are connecting fields of consciousness); these self-sufficient spirits are able to decide freely for themselves a lifestyle which is either good or evil, according to your own human standards. It depends again on the respective point of view; your people are not necessarily in a position to judge whether the deeds of a much more highly developed species are good or evil, because you stand at a lower observation point, from which an assessment is not possible. Your simple words "good" and "evil" are in any case examples of a tendency towards generalization; in my language there are many concepts for the various shades of meaning of individual behaviors in comparison to the norms of a society.

Even those extraterrestrial species which are inclined to act with antagonism towards you are not "Species of the Evil One," even though they operate negatively with respect to your own race. They do this for their own reasons and do not regard themselves as evil; were your structured way of thinking more linear and more focused as theirs is, then you would also behave in such a fashion. The attitude of a species towards other kinds of existence naturally depends very heavily on its respective structured way of thinking; each species sets its own priorities. To classify that as "good" or "evil" is really quite primitive, for the survival of any species argues for many varieties, among them your own, as well as for even the most varied of the worst or negatively-directed deeds. I won't even exclude my own kind in this regard, for there have been certain occurrences in the past which I don't personally welcome, but about which I would also not like to go into detail. None of these occurrences have happened in the last 200 years of your time scale. But please note the following: there are NO absolutely good and there are NO absolutely bad species, because each and every species always consists of individuals.

Question: In the letters that I got, there was often the question, whether you could go into any greater detail regarding the advanced physics that you commented on last time. Many people said, your words made no sense. For example, how do UFOs function, how do they fly, how do they perform the maneuvers that they do?

Answer: I ought to explain that to people? That's not all that simple. Let me think about it for a minute. I always have to use very simple words in order to make clear to you the basic principles of a higher kind of science. Let's try this: You have to be clear about some fundamental facts. The very first thing is that you must divide up the conception of the physical world because each existence consists of different layers; let's say for simplicity's sake that it consists of a material illusion and a sphere of influence. {TRANSLATOR'S NOTE: No legitimate translation exists for this word 'Feldraum'; "Feld" means "field," "Raum" means "space, room, expanse." Therefore, I'm translating it as "sphere of influence."} Certain physical conditions are associated only with the realm of the material {as in 'concrete'}, while other and more complicated conditions are associated only with the sphere of influence of the material world. Your conception of the physical world is based upon a simple material illusion. That illusion is further subdivided into three elementary or basic conditions of matter. A fourth and very important condition also exists, which you simply pay attention to more or less as you choose; it is the one bordering on the sphere of influence or plasma realm. For you, the theory for a controlled transformation or an elevation of the frequency of matter and the stable existence of this fourth aggregate condition of matter is not very common, or it exists at a very primitive level. (As an aside, there are simply five states of matter, but the post-plasma state would really be going too far and it would only serve to confuse you. Besides, it is not necessary for an understanding of the basic theory; it is connected with diverse phenomena which you would characterize as paranormal.) Now, back to the essentials: Plasma...now, with plasma I don't mean just "hot gas"—as the concept is generally simplified by your people—but rather I mean a higher aggregate condition of matter. The plasma state of matter is a special form of matter which lies between its real existence and the sphere of influence, that is, a complete loss of mass and pure accretion of energy of various form whenever matter is "pushed or shoved." {NOTE: No explanation was given for the use of the word "pushed, shoved" as used in this context. Your guess is as good mine.}

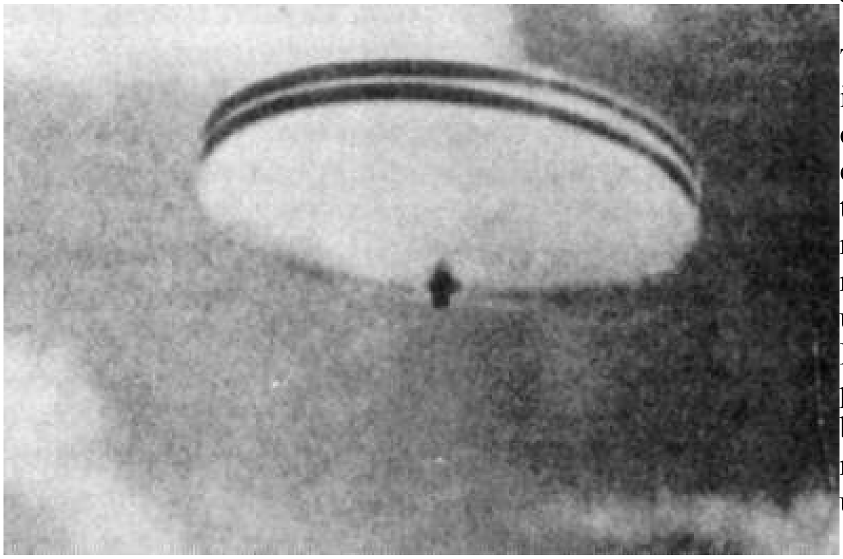

The fourth state of matter is very important for certain physical conditions which can be used for example to...how should I express this to you...generate antigravity. (That's a rather strange human word and not really correct, but you ought to understand it better this way.) Essentially, in the world of real physics, there are no bipolar forces, but rather only "observer dependent reflective behavior" of a single, large unified force at different levels.

With antigravity or the displacement of gravitational characteristics into levels, one can, for example, cause apparently solid matter to levitate; this method is employed partly by us and by extraterrestrials as well as a means of propulsion for their UFOs. You people are moving on a really primitive level towards a similar principle for your secret military projects, but since you have more or less stolen this technology (and it was later falsely passed on to you intentionally by the extraterrestrials), you lack the real physical understanding; as a result, you have to struggle with problems of instability and radiation with your "UFOs". According to my information, there have been a great number of deaths of your people because of intense radiation and field disturbances. Don't you agree, this is also an example of the business regarding the question of "good" and "evil"? You people play with unknown forces and thereby accept the death of colleagues of your own kind, for they are dying for a greater cause, namely, for the advancement of your technology, which as a result is being put into place once again for the purpose of war, i.e., for negative pursuits. Now, one can give you the benefit of the doubt, that only the least number of your kind have any knowledge about these alien projects which are —as you explain it— top secret. It was told to you that the higher the ordinal or ranking number of basic matter, the simpler the heightening of the condition, but that is only partially correct. If you can't circumvent these powers, then you're better off not attempting it. But your kind has always been ignorant and has from time immemorial tried to play around with forces which you have not even understood. Why would that ever change?

You remember this business of copper fusion? By means of the fluctuation at the right angle with the induced radiation field, copper is fused with other elements. (The illusion of matter is fused, the fields in the sphere of influence overlap each other, but the main force would be reflected by that process and would assume a quasi-bipolar character.) The resulting connection and the field would therefore not be stable in the normal condition of matter and unsuited for tasks. As a result, the entire field spectrum is shifted to a higher plasma-like condition, whereby the spectrum comes together with this harsh shifting to the opposite pole side —the word is NOT correct— of the force field and it resembles quite closely a gravitational shift. This shifting causes a "tilting" of the repulsing quasi-bipolar force, which now no longer flows to the interior of the force field, but rather flows partly to the exterior of the field. The result is an inter-stratifying reflective force field which is very difficult to modulate within certain technical boundaries in relation to its own characteristics. It can also carry out a multiplicity of tasks, as for example, causing massive flying objects to be levitated and maneuvered. It can also exert a camouflage function in the realm of electromagnetic radiation as well as manipulate temporal sequences of events—indeed only to very limited extent—and other things as well. Are you familiar with your "quantum tunnel effect"? Even the amplitude equalizations among genuine matter can be achieved with one of those kinds of fields if the frequency and the distance from the plane of the field are high enough. Unfortunately, the whole thing that I have explained to you in your words has come out to be rather primitive, I'm afraid. It sounds rather strange and certainly impossible for your comprehension, but perhaps this simple explanation can be of some use to you in helping you to understand. But then again, maybe not.

Question: Is there a scientific substantiation for paranormal powers, as for example with your powers of thought?

Answer: Yes. In order to explain that, one has to acknowledge the physical reality of the sphere of influence {Feldraum}. I'll try to do it...wait just a second...you are going to have to separate yourself mentally from the illusion that that which you see is the true nature of the universe.

It is, at best, the surface of a side. Imagine for yourself that all the matter here —you, this table, this pencil, this technical device, this paper— does not really exist, but that it is rather only the result of a field oscillation and a concentration of energy. All matter that you see, every creature, every planet and star in this universe, has an "information-energy equivalent" in the sphere of influence which is located on a main field —the general level {of things}. Now, there is not only one level, but several. Last time, I had mentioned that highly-developed species which is capable of changing levels (which is something completely different from the simple bubble changing, for bubbles are a part of each and every level). Do you understand? Dimensions, as you call them, are a part of a solitary bubble, bubbles or universal foam are a part of a level, and levels are layers in the sphere of influence, while the sphere of influence, acting in the capacity of single physical size, is essentially unending; it is composed of innumerable information-energy layers and general levels. There are in the sphere of influence no null-levels; all are the same, but they are separated by means of their energy conditions. I notice that I am confusing you now. I think I ought to stop with this explanation.

Question: No, please continue. How do concrete paranormal powers arise?

Answer: Well, OK then. Let's try something simpler. Again, it is not completely correct, but let's begin in this manner: Tangible matter on this side is mirrored in the sphere of influence {Feldraum} as a field with distinct layers. These layers contain information, as an example, about the simple structure of matter or the string frequency, but also there is stored information stemming from the development of matter. Are you familiar with the human concept of "morphogenetic fields?" One part of the layer could be designated as such. Now there is still another intermediary layer for which you unfortunately have no human concept, since the theory is not common in human thinking. Let's call it a "para-layer," for this layer is mainly responsible for everything which you call PSI and paranormal and which lies outside the boundaries of your primitive science. This para-layer lies between the layers of matter and the morphogenetic layers of a field in the sphere of influence. It can actively integrate with both. Your body, for example, is mirrored as a field in the sphere of influence {Feldraum}. That doesn't mean that it does not also exist here as well —as flesh, blood, bones— in the form of matter strings or atoms, but not only that. Existence is always a duality. Some layers of the field contain simple information about the solid matter of your body and its frequency, while other layers {contain information about} your spirit, your consciousness or, speaking from a human-religious point of view, your soul. Awareness or consciousness in this case is a simple energy matrix, divided into different layers of your field in the sphere of influence —nothing more, nothing less. Genuine awareness can also exist here on the matter side, but only in the form of post-plasma {the fifth form of matter}. With the necessary physical knowledge and the corresponding technology, the consciousness/awareness matrix, or soul, can also be separated from its field of rest. It can, despite its removal, continue to exist in a self-sufficient manner for a certain amount of time. That has the strange occult name of "soul robbing." But above all, we're talking about science here, not about magic or dark forces.

[Comment by Ole. K.: The "soul robbing" was mentioned in one of the radical, religiously-motivated comments in connection with the reptilian species.]

But back to your Question: Creatures with more powerful mental powers can have a direct influence on the para-layer by means of their consciousness/awareness fields.

Now this layer is not limited only to the individual, but rather as a part of a general information layer —you could call it in a prosaic sense the community soul— that is connected with all animate and inanimate matter and all consciousness which exist on this main level. The biological cause for these abilities lies on the side of matter, by the way, in the pituitary gland, which always is in the position to generate the frequencies to actively control the sphere of influence {Feldraum}. Even you people could theoretically do this; however, you are solidly blocked in these things. As I have said, the para-layer can interact with mind as well as with matter. For example, if I decide to use my mental powers once more in order to move this pencil, then, simply said, I imagine in my mind how my consciousness/awareness expands/amplifies itself on the matter side in the form of post-plasma to the pencil. In the sphere of influence this causes simultaneously an automatic command from the consciousness/awareness layer to the para-layer to interact with the matter layer of the pencil. Since the para-layer is not confined to the body, it is not even a problem that the pencil lies over there, for I can unerringly reach it, even without moving my matter body. Post-Plasma on this side, para-layer on the other. I have control over the pencil and the interaction brings the matter field of the pencil to the point where it changes in the manner in which it moves, for example.

[Comment by Ole K.: I certify that the pencil mentioned above abruptly at that very moment jumped into the air to a height of 20 cm and then fell back to the surface of the table. The sound is clearly heard on the recording tape. No one visibly had touched that pencil.]

Question: That is fascinating. Which kinds of paranormal activities can one generate with that?

Answer: All kinds. Everything that you call paranormal. As I said, this special layer lies in the sphere of influence {Feldraum} between the morphogenetic information layers and the matter layers and can interact with respect to both sides. That is to say, it can be interacted with solid matter as well as with mind or mental information, wherewith we can achieve everything that is generally designated as telekinesis and telepathy. The "connection absorption" with another consciousness/awareness is generally separate in the procedure from the simple influence of matter, since different consciousness/awareness fields work with different oscillations. A consciousness/awareness that sends or a consciousness/awareness that listens must first adapt itself to the other mind exactly, before any access is possible. Most species also have chances to block the alien access, but you people don't have this. The following is generally valid: the stronger the paranormal abilities of a species, the simpler the adaptation and the access. Our own abilities are not so powerfully developed; therefore, first we have to learn specifically alien mind influence in order to use our mimicry, for example—where mimicry is actually quite simple in your minds due to the implanting of the on/off switch. Some of these abilities are also partially inherited; mother and child of my kind as an example are attuned exactly during the first months of life —partially also in the egg covering in the expectant mother— and communicate telepathically. In order to influence you people, we need a certain amount of time for practicing, despite your simple structure. Therefore, it is forbidden, for example, for adults of my kind before the "Age of Enlightenment" to come to the surface of the Earth. (That term is synonymous, along with other things, with full physical strength.) In the case of not fully developed abilities, the danger of discovery by you would be too great. By the way, there are of course numerous secret teachings about the real possibilities which can give one these abilities, but I really don't know anything exact about that.

Whenever an alien mind ought to be influenced, then there are some generally valid steps, which are set into motion by other extraterrestrial species.

First and foremost, the alien oscillation must be felt, something generally that is done automatically by the brain, i.e., for the one the field oscillation, for the other the quasi-electrical brain waves here in the normal space {which matter inhabits}. That is not especially difficult. After that, one simply probes for the other consciousness/awareness in the mind with a post-plasma manifestation, the sphere of influence {Feldraum} reacts and the connection is there. Now one can read out information from the first one and record the desired information to the second one in the correct location. You asked me last time whether you people have the opportunity to protect yourselves against this influence, and I told you that only an awake and concentrated mind had any kind of a chance to withstand it. In this state of mind the oscillations change very abruptly and access becomes complicated; more precisely, it can come as a painful recoil. Whenever you close your eyes, then the field becomes "flat," and alien access {to the mind} is immediately possible and without restriction. In terms of your chances against a more highly developed species, you have none at all. They are able to adjust the oscillations faster than you can change. I could even demonstrate it on you, but you were really horrified and confused the last time, so we'll just leave it at an explanation.

This explanation presumably sounds to you like —as you say— something esoteric or from the occult or magic. The reason for that is simply that you lack the basic understanding for seeing the background reasons. All paranormal phenomena have a purely scientific origination. None of this has anything to do with supernatural powers. We grow up with this kind of knowledge, we know how one makes use of these powers, and where they come from. We are acquainted with theory and practice. You are not. Therefore, you really don't understand what happens in your world—you see only one side of existence, not the other (I mean here both that are physical). Everything paranormal is dualistic, and it exists in the space that matter inhabits as well as in the sphere of influence {Feldraum}. To be explained...it can only be explained by the acceptance of the latter, because the sphere of influence {Feldraum} is the basis. I would welcome an end to the scientific questions since you really aren't grasping them anyway. We're wasting more or less valuable time by doing this.

Question: Only one last question. At our first meeting in December, you made it quite clear that you didn't want to discuss scientific and paranormal concerns. Why the openness now?

Answer: The last time I saw really no necessity in overburdening you with facts of that kind (and now you are obviously overburdened). Therefore, I had preferred only to mention these topics in a peripheral sense. Apparently, however, some of my performances today have set you to thinking about your world, something that can't be all bad. And by the way, your human scientists will tend to regard my comments as "humbug." And so I see no great danger in spreading this information widely. No one will pay much attention to it. By the way, the words of people who have characterized me as a "Creature of Evil" have their basis in the belief in occult powers and magic —both of which things DO NOT exist. There is no magic, only highly developed science, and everything that you label as "magic" is only a part of science. If you would only comprehend that, then you would be a step ahead in your development. My openness on this issue ends here. Pose other questions, please.

Question: Good. Let's talk about UFOs. Can you explain to me how our governments came into possession of UFO material to the point that they could start their own projects? Did it have anything to do with the "Roswell Incident?"

Answer: Yes, but that incident was not the first one. I am no historian, I am studying only your current behavior, so my knowledge about those events in your history is presumably not very extensive. I will try to explain to you what I know about those things which happened at that time. Let me think about it for a second. In the years 1946 to 1953 in your time scale, there were five cases where extraterrestrial ships crashed to the surface of the Earth. In that crash which you call the "Roswell Incident," there was not only one alien ship involved, but two that crashed after a collision in different parts of the land in the west—the one you call the USA. (You have to know that the ships of this particular species can remain levitating in the air for a particular period of time even though they are damaged; that accounts for the spatial difference {in their crash locations}.) These were indeed not the first crashes, but by that time the second and the third. Another ship had crashed in 1946, but it was destroyed beyond usability.

One thing first before the explanation: it certainly sounds ridiculous to you that such highly developed extraterrestrial ships simply crash, and that a relatively large number did so in a relatively short amount of time. The explanation for that is likewise more than strange, but it is correct. It does not lie in the ship's drive itself, but rather in the direction of the field to your planet. This species that we are discussing —and it was always in this time period that this species used a disk-shaped craft— used a propulsion system which ran according to the normal principle of fusion, to be sure, but one that at that time employed a more than unconventional method for field alignment. This method had various advantages but also disadvantages. The repelling field must of course lie in the absolute correct angle to the surface of the Earth. This species used an alignment technology in their ships, with which the field locked into place all points of the Earth's magnetic field. Now at that time this species had just arrived on the earth and their point of origin lay on a planet with a more stable magnetic field, for which they had developed and aligned their drive. The magnetic field of the Earth is not really all that stable; it is subject to cyclical variations and it forms field eddies under unfavorable conditions. Whenever a ship with one of those kinds of drives gets into a field fluctuation or into an eddy that is too strong, then for a short time the repelling field can no longer align itself correctly and the ship glides uncontrolled on its flight path. The drive is operating correctly, to be sure, but the field fluctuates in all directions and because of that, the ship can crash. In the case from 1947 which you addressed, it is my understanding that one of the ships got caught in a fluctuation, its field linked up unintentionally with that of its squadron leader and it collided with another ship whereby both of them were heavily damaged. The cause for the magnetic fluctuation at that time was probably an electrical disturbance brought about by a weather event. Both ships crashed as a result; one of them fell near the collision point, the other a hundred of your kilometers or so distant. All occupants were killed in the impact. The thin hull structure of that kind of disk craft is in and of itself not very stable, since those disks have not been designed for crashes as well as for flight in a field where there are exterior forces at work.

Now, your human military collected the individual pieces at first until they discovered the whole ships with the dead creatures aboard. Immediately they classified everything as "Top Secret" and brought them to their military bases in order to analyze the drive. The secret endeavor was to set the alien technology in place later against evil enemies of that great country. That is as primitive as it is ridiculous.

I believe I remember—I don't want to specify exactly your date—that it was probably between 1949 and 1952 that there was a rather bad accident during some research being done on one of the wrecks. According to what I heard—what members of my species were told by members of that government—it resulted in an unintentional activation of one of the drive's components in the unshielded condition. As a result, for a very short period of time—how should I phrase this—there was an unchecked shift of the environment to a plasma-like condition, which on the other hand, through a very, very unfortunate accident, caused an overturning of the general power field into a magnetic pulse of immense power. Do you have any idea what kind of an effect a plasma-magnetic jolt has, when it comes into contact with an organism? No, how should you know that. Of course you don't. Disturbance in the structure of the field and bioelectric feedback. Imagine, if you will, a human body which is engulfed in bright flames for 3 or 4 of your days. Those flames apparently do not go out and they burn the body right down to its last constituents. Well, then you have an approximate impression of what happened. I think that 20 or 30 of your scientists were killed in that lab.

Two further crashes occurred in 1950 and 1953 in the water catchment area of the American continent. Those ships were able to be recovered from the crashes relatively intact. (The one in 1953, as I remember, even had an intact drive core. It was by means of that device that you saw for the first time that you had understood the entire concept fully incorrectly and that you had reconstructed it fully incorrectly. Even today you still don't have it right.) That species, which had built the ships in the first place —a species which I, by the way, count among those who are unfriendly towards you— was naturally worried about the investigation of their own technology by your kind. They did not want, however, at that early point in time, to begin direct conflict with you, and so they chose the diplomatic path and came into contact with that government during your 1960's. Of course, they did not divulge the real reasons for their being here —copper, hydrogen, air— but rather they pretended to be curious "researchers" and offered to show people the functioning principle of the ships whereby they would expect in return some "favors." Simple-minded as you are, you of course agreed to it...and were deceived. You gave them raw materials, you gave them secured locations for their bases, you gave them access to your most secret defense data, you gave them access to your DNA and much more—and all just to quench your greed for power and information. The alien species of course quickly noticed that they were dealing with simple-minded creatures, and they gave you false and inferior information about their technology so that they receive much more out of the collaboration than your kind do. For example, they gave you information that the drive can only be constructed with unstable elements of a higher ranking number, but they withheld the information that the field drive can be constructed with various modifications to work as well with stable elements of a lower periodic number, and generally, that's the way it's done. Through these half truths they made you dependent on the synthesizing of high {numbered} elements, and thereby renewed by their own technology. Their clues to the construction of your "UFOs" were laid out in such a way that the solution to old problems caused new problems to arise simultaneously. They never told you the complete truth, but always built in again and again clever lies, which later lead to technical problems —and to your dependence on them.

In the last years of your 1970's and your early 1980's, it finally came down to various events between the alien species and that human government —I don't want to go into detail here since there is much that even I am not exactly sure of. The whole thing lay in the context with some new, or better said, the old technical problems with your own self-constructed ships whose camouflage and drive partially failed to function in test flights in the open.

Because of that, the function of secrecy was threatened. Your military and your politicians slowly—very slowly—came to the conclusion after more than 20 years of this that they had been deceived by that alien species. Multitudinous incongruities and the overstepping of bounds of the treaties by both sides finally led to an altercation between you and the extraterrestrials, which culminated in the lift-off of three of the alien aerial objects through a special —how do you say it?— EMP {electromagnetic pulse} weapon and a military skirmish at one of their underground installations. As a consequence of these attacks, the alien species ultimately withdrew from all contact with you and was understandably more than angered about your kind. Therefore, I count these extraterrestrials among the three groups who are hostile towards you, and while the other two are more occupied with their own business, among them waging a cold war for dominance on your planet, your old "friends" and partners are preparing to supply themselves finally with the sole and absolute dominance over raw materials and human DNA. At the moment it is probably true that they lack some of the technical possibilities and the large amount of forces which they need in order to achieve their goals directly. In spite of that, we are counting on negative actions —possibly ever of a more subtle kind— against you in the next few years or decades.

Question: Will the other extraterrestrial species undertake nothing against these war-like actions? Specifically, something ought to be on Earth for the more highly developed species.

Answer: You're wrong there. Specifically, for the more highly developed species there is simply at the very least your fate. You are animals for them. Animals in a very large lab. Understandably, an alien intervention on your planet would disturb their projects, but I don't think that they accept a confrontation with other species for it. Many of them could look for another research planet for themselves or they could study over a long distance your behavior and your consciousness/awareness, since crisis situations could have an attraction for their studies. Whenever you people take a look at an ant hill, and another person comes along and steps on the ant hill, what do you do? You go on your way, or you search for another ant hill or you observe the ants in their crisis condition. But would one of you —even though he were larger and more powerful than the one who stepped on the ant hill in the first place— defend the meaningless ants? No. You have to imagine for yourself the viewpoint of the more highly advanced creatures. You are the ants. Don't expect any help from them.

Of course we would also ask for help when it became clear that your old partners were ganging up on you. Some members of that human government are fully aware of our existence —also partially owing to an old religious basis. For example, there is a gigantic partially underground building in the capital which is totally dedicated to my species and that also has a direct approach to an elevator shaft and to an underground system. In this building partial meetings have taken place and do take place between us and humans. We have passed on information to you in the last few years; according to what I know, we will keep ourselves as far away from the conflict as we can. You ought to learn to solve your own problems yourselves or to become intelligent enough never to create those kinds of situations. What will come and who will possibly place themselves on your side, only time will tell. I really do not want to make any indications about that.

Question: I have here 5 prints of different UFOs, which claim to show UFOs. Can you take a look at the pictures and tell me in which of them actual extraterrestrial aerial craft can be seen?

Answer: I can try it. You pose many questions to me today which even I cannot answer unequivocally. Don't overestimate my knowledge, I'm no expert in alien technology and the construction of extraterrestrial ships. To be sure, there are mostly some technical details and peculiarities about genuine "UFOs," with whose help one can easily differentiate them from natural phenomena or human forgeries. You falsify sometimes the pattern of genuine ships; therefore, it is not so easy simply with absolute certainty to identify an object. I'll try it. Show me the photos.

[Comment by Ole K.: She considered the pictures respectively for only a couple of seconds and then sorted out photos 1, 3 and 5.]

These three pictures here are obvious counterfeits or erroneous identifications. In the one picture, it certainly seems to me that a real existing ship of an alien species was adapted for a small model here. It lacks important characteristics which are tied in with the technically- and physically-associated field. Generally speaking, a picture is all the more a fake, the clearer the outline and the colors are, because a levitating ship is generally hidden in a shifted-field condition that even distorts the colors or the forms according to alignment. It might perhaps sound strange, but hazy and spectrally-shifted photos are sometimes to be interpreted as an indication for a possibly authenticity. By the way, this object is floating above the water. If it were a genuine ship, we would have to see in any case either a trough or a swell on the surface. Since the surface is flat, it is obviously not a genuine ship. In my opinion, none of these three pictures show genuine objects in flight or UFOs. Here in this picture I see above all no artificial object in flight; it seems much more to deal with only a light reflex in your simple optical cameras. You really ought to be intelligent enough not to fall for a mix-up like that. When your general public chases counterfeits and frauds for a long time, then they will presumably discover too late, what is really going on in front of them in their atmosphere.

PHOTO 2: Albiosc, France, 1974

This one seems to be genuine, at least it displays the necessary characteristics. I would assign it at first glance to an alien species who have been visiting your planet for the last 35 years or so. The object itself is metallic and disk-shaped; certainly it is distorted in form and color by means of a field effect. These four white and very long "processes" on the underside of the ship itself portray a kind of quasi-gravitational light manipulation, i.e., the universal force field is being shifted in the direction of a simulated gravity. Actually, it is not a genuine light (it is mostly not a genuine light whenever you see illuminating "UFOs") but a special strongly charged form field which manifests itself in the space that matter inhabits as a quasi-light. The reason for the activation of this special high-energy system in an atmosphere is not completely clear to me; it's possible that it is a kind of investigating or influencing of the environment. In any case, it is terribly careless of that species to allow this technology to be photographed by humans. Well, I guess that most of you just plain don't understand it, and those who do will not say anything about it to the general public.

PHOTO 4: Petit Rechain, Belgium, 1990

This is in fact a genuine aerial object; it is in no way extraterrestrial. Triangular aerial objects in flight are simply not used by alien species, or not in this form, at least. That streamlined kind of form is a human concept. It is one of your own secret military projects that you build with the help of immature alien technology —technology that was handed over to you by the extraterrestrials during the 1960's and the 1970's. Generally, the form of the hull for a genuine extraterrestrial ship is of no consequence, for inside the field itself there are no exterior forces that have any effect there; in general, the ships have a rounded off form and they are built without hard edges —as a disk or a cylinder— so that the field can flow more easily. Your projects decree that along with the alien drive field there also be a conventional jet engine system; therefore, they are always triangular and built thus with streamlining in order to be steerable with this primitive recoil principle.

In the example here the ship glides above all on its genuine field drive. Do you see the distortion and the quasi-light in the rotating cylinders? That is an unmistakable indication for the authenticity of the photo. But why, you might ask, are there 4 cylinders? That's unusual —even the interval seems to be incorrect. The coloring is very dark and the interior optical distortion is very noticeable. Presumably a reconstruction of the original system by your scientists. Since the alien species has just not given you any more information since the disagreement, they are rebuilding the systems single-handedly without actually being able to understand what kind of dangerous thing they are doing there. This construction does not make the system better, only more unstable. Both of the forward cylinders are too close to each other; they will definitely flow into each other. The color shows me a powerful residual radiation; it was probably the case that high elements were used again as customary for the shifting. It is in any case very dangerous to be unshielded in the vicinity of the field. Did the person who took the photo display any kind of radiation and burn damage?

Question: I don't know. Where do these military "UFOs" come from? From the United States?

Answer: Yes. I think generally that's true. From the western continent.

Question: Why then do they fly over thickly populated areas of Europe? This photo comes from Belgium. That doesn't make any sense. Can you explain?

Answer: Why is it that ONLY I am able to explain strange human deeds? It's possible that these are long-distance tests or tests with the electromagnetic camouflage systems. The old enemy of the American nation is on this side of the world, so why shouldn't they test here? At home they've had enough time to have had their ships crossing back and forth. Maybe they have aroused too much observation there. With one of those kinds of unstable field structures —as your photo indicates— I consider it somewhat improbable that that ship is capable of making a flight of that length over the ocean. It's possible there is a test station here on your continent. Unfortunately, I don't know anything about it.

Question: Many readers of the first transcript have posed the question how your original contact with E.F. came about. I already know the story from your narratives, but could you repeat it here once again for this volume and for the new transcript?

Answer: Of course. Now, the story began about two of your years ago here in Sweden. I have been strongly interested in your species and your behavior since my youth; I had already studied your literature at that time, as well as possible. (Naturally, it is not easy in my homeland to come into possession of human books, but since my group or family stands in a higher ranking order, I was able to gather some material together and sometimes to speak with others of my kind who have already been in contact with you.) I was really very curious about your species and as soon as I was allowed to come to the surface, I attempted to assemble more information immediately; above all, it was expressly forbidden for me to commence direct contact with humans because in my position at that time, there existed no necessity for doing so.

It was in your year 1998, when I was on my way further north from here in the remote forests in the vicinity of the entrance to my world and was looking for biological specimens, which we use in order to watch over environmental pollution and destruction of your flora and fauna statistically by your own kind. At the time, I was already on the return path to the entrance—we can even orient ourselves more easily, by the way, through our senses to the Earth's magnetic field—and already in the vicinity of the large lake, when much to my surprise I came across a cabin in the woods. In this cabin I sensed a human consciousness/awareness. It was E.F. Actually, I had no permission for contact with another species, but by the same token I had set in place my mimicry capability quite successfully prior to this—even with larger groups of you (I had never ever come across a human being when I was alone). Now, let's call it primitive curiosity; I wanted to talk with the person in this cabin and so I knocked on the door. E. opened the door and we got into an interesting conversation. His language was not quite yet common at that time for me, but it's not all that hard to learn a new language when one can read the information in the consciousness/awareness of the opposite individual. I simply told him that I came from a foreign country in the east. Of course at the time, he did not really "recognize" who I was; he was totally convinced that he was talking with a creature of his own kind, although it was simply only a mimicry image.

Since my assignment anyway had as its goal an investigation of this terrain which was to last for several days, I visited him in this span of time three times as a human person. At first we talked mainly about really ordinary things; later we got into religious and physical topics. He seemed to be impressed by my knowledge, and I was likewise impressed with his clear thoughts and his—for a human being—well displayed personality structure and his own opinions. You really like giving yourselves over completely to a public opinion or conditioning, as for example, "reptilian species are evil" and stuff like that. I steered the conversation in this direction, and E.F. said something to the effect that he believed in alien species and that they did not have to necessarily be evil, but perhaps only different than his kind are. That pleased me. At that juncture of time, of course, I could not speak concretely with him about my knowledge because he wouldn't have believed me—he would have taken me for a human practical joker. I cultivated the very, very unusual idea (for my kind) to show him my true exterior, something that I did during our conversation at our fourth meeting in the cabin. Actually, he was predestined for contact: he was open-minded, honest, intelligent, not religiously inclined or conditioned; he lived alone and isolated, and no one would believe him, should he decide to go public with his story.

I dared to take the step, but then I had serious doubt about the propriety of my act, especially when he reacted...very...violently. He got control of himself again after a time and we could finally talk concretely about definite matters. Now he had no choice but to believe me. This was the beginning of a series of meetings which initially took place there in the woods, but later took place in his remote residence. Finally he brought you into contact with me...and for that reason we are now sitting here once again and talking about things which probably won't be believed out there in human society.

Question: You said, you would not have had permission at that time for contacts with human beings. Do you now then have permission to talk with E.F. and me about all these things and even to make this scientifically public?

Answer: Yes. That is difficult to explain and for you to understand. Let's just say, I find myself in the position now to arrange this permission without having to take into account any consequences. In this position I am quasi-"immune" against certain restrictions. Let's look it that way. Yes.

Question: If other people want to come into contact with your kind, do they have a chance to do so?

Answer: Generally not. We avoid contact with you and we operate on the surface only in remote areas and there we use the mimicry techniques in case we should meet some people. That I am talking with you now does not mean that others will follow my example. It goes without saying that you could try to find an entrance to my world and penetrate your way into there. However, that can sooner lead to unpleasant consequences for the infiltrator. You have next to no chances on the surface of recognizing us. You can't even contact us directly, we have to contact you, just as I did with E.F. Those kinds of contacts however are not the rule but are very rare occurrences.

Question: Can you describe your subterranean homeland location?

Answer: I can attempt to do so, but I certainly will not tell you where this place is located. My homeland lies in one of our smaller underground settlements to the east of here. I'll give you some numbers so that you can make a better impression for yourself. Just a minute...I have to try to convert the measurements approximately into your units. It is a dome-shaped cavern at a distance of about 4300 meters from the Earth's surface. The cavern was organized as a colony about 3000 years ago; a major portion of the ceiling structure is artificially integrated into the rock and the form was remodeled into an almost elegantly proportioned and very flat dome with an oval ground plan. The diameter of the dome according to your measures is about two-and-a-half kilometers. The height of the dome at the highest point is about 220 meters. Underneath that highest point in every colony there stands a special whitish-gray cylindrical building—a kind of supporting column which holds the honeycomb net-carrying structure of the dome. This building is the tallest, largest and oldest in the entire dome for it is always situated as the first construction together with the security of the ceiling. (In the meantime of course there were times when it was completed and reconditioned.) That building has a very special name and religious significance. We have only one of those columns; larger colonies even have more columns according to the construction of the ceiling.

One of the main colonies in Inner Asia has as an example 9 of those kinds of supports, but that colony is also over 25 of your kilometers in size. The central building is generally a center of religion, but also a center for climate control, and a center for the behavior and the regulation of the lighting system.

We have at our location all together 5 large artificial light sources which generate your UV light and its warmth through gravitational sources. The air shafts and the light systems from the surface likewise run through these columns and naturally, they are very intensely controlled.

By the way, we have 3 air shafts and 2 elevator systems there, and even a tunnel connection to the next main colony which lies approximately 500 kilometers to the southeast. One elevator shaft leads to a cavern near the surface, the other leads to one of our depots for the ships —you remember, the cylindrical ships— that is naturally concealed closer to the surface behind a rocky mountain face. Normally, there are only three ships there —it's a small depot. The other buildings of the colony are, for the most part, concentrically ordered in oval circles around the main supporting column, and they are without exception much flatter; generally only between 3 and 20 meters tall. The shape of the buildings is round and dome-like. The color is even differentiated according to circle and distance from the main column. To the north of the column, there is an additional, very large but very flat round building. This building interrupts the concentric system of the colony with its diameter of about 250 meters. It is the artificial sun zone in which specially illuminated corridors and rooms are housed. In these locations very powerful UV light predominates, and they are used in order to warm our blood. There is even a medical dispensary and a meeting room located there. Beyond the outer ring of the colony, there are zones in which animals are kept—you know, we MUST consume flesh as nourishment—and the gardens in which plant nourishment and mushroom culture are cultivated; there is also hot and cold running water there from subterranean sources. The power station is located on the edge of the colony. The station is driven by fusion as its base and it supplies the colony and the "suns" with energy. My group or "family" lives, by the way, in the fourth ring of buildings out from the central support column. So much in such a short time. To describe to you all the buildings and their tasks would be going too far. It is difficult to describe something like that to you, for it is a completely different set of surroundings and culture from what you are accustomed to in your life on the surface. You really have to see it for yourself to be able to believe it.

Question: Will I myself see it sometime?

Answer: Who knows, maybe. Time brings new opportunities.

Question: How many creatures of your kind live in this colony?

Answer: Approximately 900.

Question: That is the end of the interview. Do you have any final message for the readers of the transcript?

Answer: Yes. I am thoroughly surprised at the many comments to my words; of course, I am naturally also disappointed about the religious portrayals of me as the enemy which have been voiced and which have buried themselves deeply in your mind. You should learn to set yourselves apart from the old conditioning and not to stand quasi under the control of something or someone who has already been gone for 5000 years. You are, after all, free spirits. Those are my final words.

All Rights Reserved ©

KEELYNET

**Thanks to Slavek Krepelka and Ren Berghuis for paying to have the Russian version of Chapter 5 translated into English and sharing it with KeelyNet.
Please check out Slavek's website.**

Orkin® Extermination Save $50 On Orkin® Pest Service 100% Satisfaction Or Money Back! www.PestRemovalZone.com
K&J Magnetics Neo Magnets Buy Top Quality Neodymium Magnets Great Selection & Prices. Shop Now! www.KJMagne
Heaven Bee Venom Eyes Controls muscles around the eyes Instant lift, tightening & Firming www.skincareorganics.co.uk/heaven
AdChoices

The Natural Phenomena of AntiGravitation and Invisibility in Insects due to the Grebennikov Cavity Structure Effect (CSE)

Introduction

by Iu. N. Cherednichenko, Senior Researcher, Biophysics Laboratory, Institute of Human Pathology and Ecology, Russian Academy of Medical Science

Viktor Stepanovich Grebennikov is a naturalist, a professional entomologist, an artist-simply put, an intellectual with a wide range of interests and pursuits. He is known to many as the discoverer of the Cavernous Structures Effect (CSE). But very few people are familiar with his other discovery, one that also borrows from Nature and its innermost secrets.

Back in 1988 he discovered anti-gravitational effects of the chitin shell of certain insects. But the most impressive concomitant phenomenon discovered at the same time was that of complete or partial invisibility or of distorted perception of material objects entering the zone of compensated gravity. Based on this discovery, the author used bionic principles to design and build an anti-gravitational platform for dirigible flights at the speed of up to 25 km/min. Since 1991-92 he has used this device for fast transportation.

Bio-gravitational effects are a wide spectrum of natural phenomena, apparently not confined to just a few species of insects. There is much empirical data to support the possibility of a lowered weight or complete levitation of material objects as a result of directed psycho-physical human action (psychokinesis)-e. g. levitation of yogi practicing transcendental meditation according to the Maharishi method. There are known cases of mediums levitating during spiritistic sessions. However, it would be a mistake to think that such abilities are only found in people who are gifted by nature.

I am convinced that these abilities are an understudied biological regularity. As is known, human weight significantly drops in the state of somnambulistic automatism (sleepwalking). During their nocturnal journeys, 80-90 kg sleepwalkers are able to tread on thin planks, or step on people sleeping next to them without causing the latter any physical discomfort (other than fright). Some clinical cases of non-spasmodic epileptic fits often result in a short-term reversible transformation of personality (people in such state are commonly referred to as "possessed"), whereby a skinny, exhausted girl or a ten-year-old boy acquire the physical prowess of a trained athlete.

Currently this psychological phenomenon is known as multiple-personality syndrome because it significantly differs from the classical complex of epileptic symptoms. Such clinical cases are well-known and well-documented. However, phenomena accompanied by a change in the weight of humans or of material objects are not confined to functional pathologies of the organism.

Healthy people in the state of acute psychological stress caused by a life-threatening situation or an overpowering motivation to achieve a vitally important goal have the ability to spontaneously overcome obstacles insurmountable in their normal condition-e. g. to lift enormous weights, etc. These phenomena are commonly explained by an extreme mobilization of muscular strength, but precise calculations do not agree with such hypotheses. Apparently, athletes (high jumpers, weightlifters, runners) have particularly developed bio-antigravitational mechanisms.

Their athletic performance is mostly (if not wholly) determined not so much by the rigor of their training as by their psychological preparedness. If an accurate scientific task of studying the anomalies of the human weight in various psycho-physiological states were ever set up and technical means of dynamic weight monitoring created, we would then have objective data on this unusual phenomenon. There is also evidence of other phenomena of short-term mass increase in biological objects, including humans, that are not related to mass transfer.

V. S. Grebennikov's book has high literary merit and includes the author's own illustrations. It is a kind of a "dactylogram" for his system of spiritual values, his environmental outlook, and his entomological autobiography. Many readers are likely to perceive the book as nothing more than a popularized summary of the entomologist's 60-year experience of scientific observations, peppered with some elements of science fiction. But such a conclusion would be deeply erroneous. As Viktor Stepanovich's friend and as someone with an intimate knowledge of his work (our homes are only 10km apart), I can vouch I have never met a more careful, conscientious, honest, and talented experimental scientist.

Grebennikov is also widely known in the so-called scientific underground (i. e. the branch of advanced Russian science constantly persecuted by the official scientific establishment). Thus, a committee for combating pseudoscience, created in Novosibirsk division of the Russian Academy, has victimized many talented members of our local scientific community. The situation is much the same at the Russian Agricultural Academy. It is very easy to lose one's job at a lab (even as its head, regardless of one's degree and title). One only needs to publish an article on, for example, the evolutionary significance of antigravitational mechanisms in insects.

But I am convinced that discoveries of such proportions must not be buried in manuscripts just because pragmatism still rules science. Let this book be nothing but "science fiction" for those at the top. Each person has his own beliefs. But he who has eyes shall see. Catastrophism in both the evolution of living nature and in the nature of human knowledge is actually a drastic destruction of old belief systems-a destruction that runs ahead of theoretical prognostications. A fanatical faith and idol-worship links our contemporary academic science with pagan religion. But a harmonious development (in the sense of Pavel Florensky's pneumatosphere) would not be possible without breaking old stereotypes in the process of mastering the wisdom and experience of older generations.

Flight - Chapter V of V. S. Grebennikov's *My World*

ВИКТОР ГРЕБЕННИКОВ
МОЙ МИР

CHAPTER V. **FLIGHT**

A quiet evening in the steppe. The sun's red disk has already touched the faraway, misty horizon. It is too late to get back home-I've stayed too long here with my insects and am preparing to spend the night in the field. Thank goodness I still have water in the flask and some mosquito repellent-one needs it here, what with hosts of gnats on the steep shore of this salty lake.

I am in the steppes, in Kamyshlovo valley. It used to be a mighty tributary of the Irtysh, but the ploughing of the steppes and deforestation turned the river into a deep, broad gully with a string of salty lakes, like this one. There is no wind. Pods of ducks gleam over the evening lake, sandpipers are also heard in the distance.

The high, pearl-colored sky stretches over the calming world of the steppe. How good it is to be out here, in the open country!

I settle for the night on the very edge of the steep, on a grassy glade. I spread out my coat, put my backpack under the head, and before lying down, collect a few dry cakes of cow manure, and light them up. The romantic, unforgettable smell of bluish smoke slowly spreads across the dozing steppe. I lie down on my simple bed, stretch my tired legs and anticipate yet another wonderful night in the country.

The blue smoke quietly takes me to the Land of Fairy Tales; sleep comes fast. I become very small, the size of an ant, then enormous, like the sky, and am about to fall asleep. But why is it that today these "pre-sleep transformations" of my bodily dimensions are somewhat unusual, too strong? A new sensation has mixed in-a sensation of falling, as though the high cliff has been snatched away from under my body, and I am falling into an unknown, terrible abyss!

Suddenly I see flashes. I open my eyes, but they don't go away-they are dancing on the pearl-and-sliver evening sky and on the grass. I get a strong, metallic taste in my mouth, as though I pressed my tongue to the contact plates of a small electric battery. My ears start ringing, I distinctly hear the double beats of my own heart.

How can one sleep when such things are going on!

I sit up and try to drive away these unpleasant sensations, but nothing comes out of my efforts. The only result is that the flashes are no longer wide and blurred but sharp and clear, like sparks or perhaps small chains; they make it hard to look around. Then I remember: I had very similar sensations a few years ago in Lesochek, or to be more precise, in the Enchanted Grove [the author is referring to localities of an entomological preserve in Omsk Region].

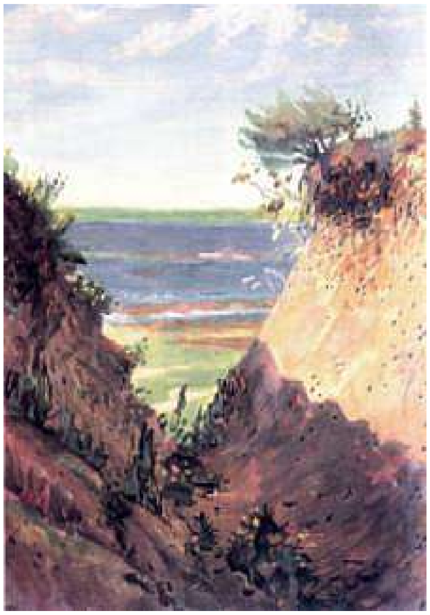

I have to get up and walk around the lakeshore. Does it feel like this everywhere around here? No: here, a meter from the edge, I feel a clear effect of "something", while ten meters further into the steppe the effect clearly disappears.

It becomes a bit frightening: I am alone in the deserted steppe, by the "Enchanted Lake". I should quickly pack up and clear out. But my curiosity takes over: what is this, really? Could it be that the smell of lake water and slime is doing this to me? I go down, under the steep and sit down by the water. The thick, sweetish smell of sapropel-rotted remains of algae-is enveloping me like in a mud spa. I sit there for five, ten minutes-no unpleasant sensations. It would be suitable to sleep here, if it weren't so wet.

I climb the steppe-same old story! My head is spinning, I again get that "galvanic", sour taste in the mouth and feel as though my weight is changing-I am at one moment incredibly light, and unbearably heavy at the next. I see flashes in my eyes. If it was indeed a "bad spot", some nasty anomaly, then there would be no grass here, and large bees would not be nesting in the loamy steppe.

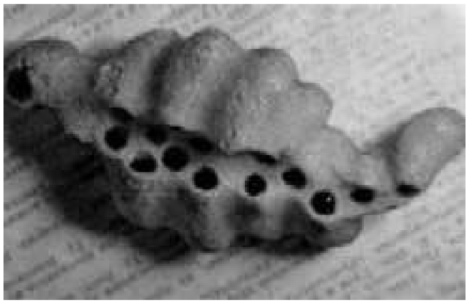

Meanwhile, their nests are all over it-in fact, I was trying to make my bed right above their underground "bee city" in whose depths there is of course a multitude of tunnels, chambers, lots of larvae, cocoons-all of them alive and healthy. I understood nothing that time.

I got up with a headache even before sunrise and, tired, hobbled off toward the road to get a hitch to Isilkul.

That summer I visited the "Enchanted Lake" four more times, at various times of day, and under various weather conditions. By the end of the summer my bees got incredibly busy stuffing their holes with flower pollen-in a word, they were feeling great. Which *I* wasn't: a meter from the edge of the steppe, above their nests, I again had a set of most unpleasant sensations. Five meters away, I had none... And there was the same old bewilderment: why, why do these bees feel so good here that the entire steppe is dappled with their holes like Swiss cheese, and in places, almost like a sponge?

The solution came many years later, when the bee city in Kamyshlovo valley died: the tillage came to the very edge which consequently fell off. Now instead of grass and bee holes, there is nothing there but an atrocious heap of mud.

I only had a handful of old clay lumps-fragments of those nests, with multiple chamber cells. The cells were side by side and reminded of small thimbles, or little jugs with narrowing necks.

I already knew that these bees were of the quadruple ring species-that was the number of light rings on their elongated bellies. On my desk, packed with equipment, ant- and grasshopper-houses, bottles with chemicals, and other stuff, I had a wide receptacle filled with these spongy clay lumps. I was about to pick something up and moved my hand over these porous fragments.

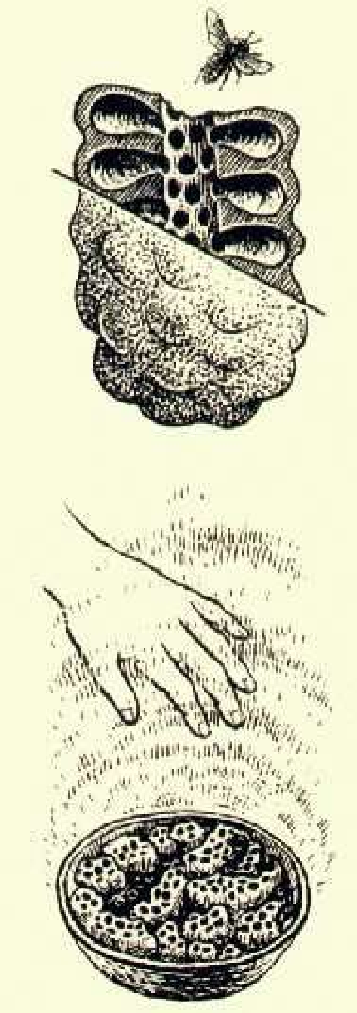

A miracle happened: I suddenly felt warmth emanating from them. I touched the lumps with my hand-they were cold, but above them I felt a clear thermal sensation.

Besides, in my fingers I felt some hitherto unknown jerks, some sort of "tick" as it were. And when I pushed the bowl with the nests to the end of the desk and leaned over it, I felt the same sensation as on the lake-my head was getting lighter and bigger, the body was falling down, the eyes saw rapid flashes, and the mouth tasted an electric battery. I was feeling slightly nauseous...

I put a sheet of cardboard on top of the bowl-the sensation didn't change. A pot lid changed nothing either; it was as if the "something" was cutting right through it. I had to study the phenomenon at once. But what could I do at home, without the necessary physical instruments? I got assistance from many research scientists of various institutes of the Agricultural Academy in Novosibirsk.

But alas, the instruments-either thermometers, or ultrasound detectors, magnetometers and electrometers-did not respond to them in the slightest.

We conducted a precise chemical analysis of the clay-nothing special. The radiometer was also silent... But ordinary human hands, and not just mine, distinctly felt either warmth or a cold draft and a tingle, or sometimes a thicker, stickier environment.

Some people's hands got heavier, others felt theirs were pushed up; some people's fingers and arm muscles got numb, they felt giddy and had profuse salivation.

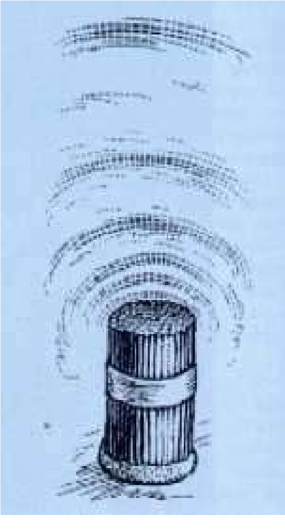

Similar phenomena could be observed in a bunch of paper tubes inhabited by leaf-cutting bees. Each tunnel had a solid row of multi-layered cans of torn leaves, covered with concave lids (also of leaves). Inside the cans there were silk, oval cocoons with larvae and chrysalides.

I asked people who knew nothing of my discovery to hold their hands or faces over the leaf-cutter nests, and took a detailed record of the experiment. The results may be found in my article "On the physical and biological properties of pollinator bee nests" published in the *Siberian Bulletin of Agricultural Science*, no.3, 1984.

The same article contains the formula of the discovery-a brief physical description of this wonderful phenomenon. Based on the structure of bee nests, I created a few dozen artificial honeycombs-of plastic, paper, metal, and wood. It turned out that the cause of all those unusual sensations was not a biological field, but the size, shape, number, and the arrangement of caverns formed by any solid objects. And as before, the organism felt it, while the instruments were silent.

I called the discovery the Cavernous Structures Effect (CSE) and carried on with my experiments.

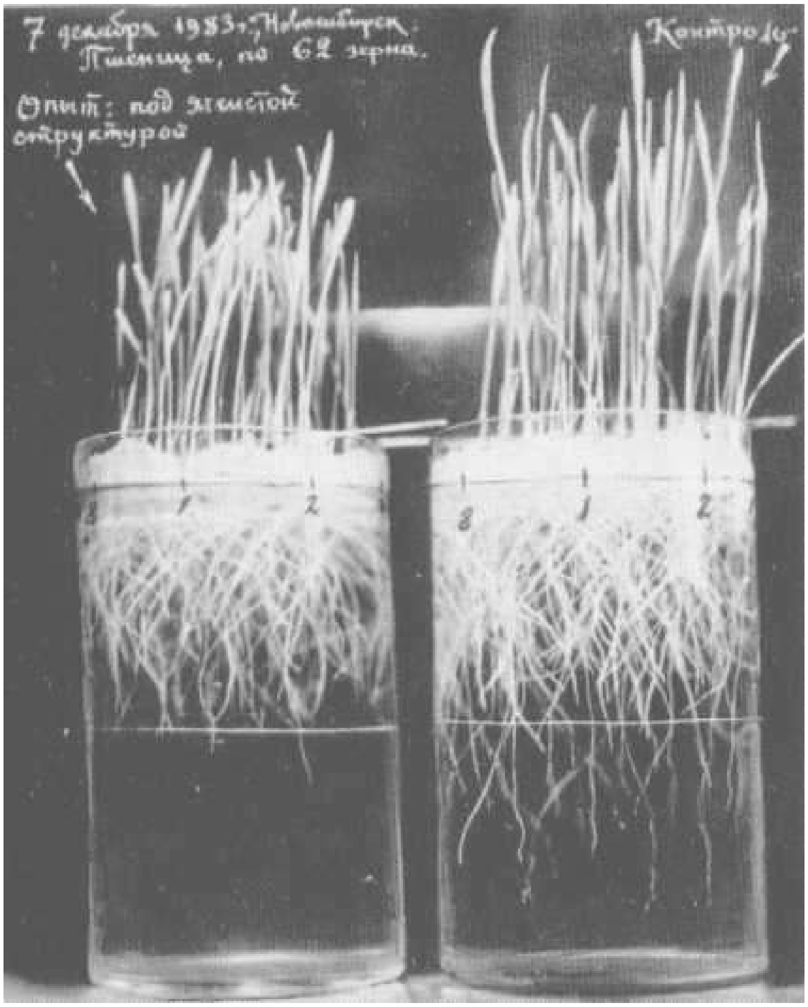

Nature continued to reveal its innermost secrets one after another...

It turned out that the CSE zone inhibits the growth of saprophytic soil bacteria, of yeast and other cultures, as well as wheat grain germination. It also changes the behavior of microscopic algea chlamydospores. Leaf-cutting bee larvae begin to phosphoresce, while adult bees are much more active in this field and finish pollination two weeks earlier.

It turned out that the CSE, like gravitation, could not be shielded-it affected living organisms through walls, thick metal, and other screens. It turned out that if a porous object were moved to another spot, the human would feel the CSE not immediately but in a few seconds or minutes, while the old spot would retain a "trace", or as I called it, a "phantom" perceivable by the hand for hours, and sometimes for months thereafter.

It turned out that the CSE field did not decrease evenly with distance, but surrounded the honeycomb with a system of invisible, yet sometimes clearly perceivable "shells".

It turned out that animals (white mice) and humans entering the zone of the CSE (even a very strong one) soon adapted to it. It couldn't be otherwise: we are everywhere surrounded by caverns large and small: by grids, cells of living and dead plants (as well as our own cells), by bubbles of foam-rubber, foam plastic, foam concrete, rooms, corridors, halls, roofing, spaces between machine parts, trees, furniture, buildings.

It turned out that the CSE "ray" had a stronger impact on living organisms when it was directed away from the sun, and also downwards, facing the Earth's center.

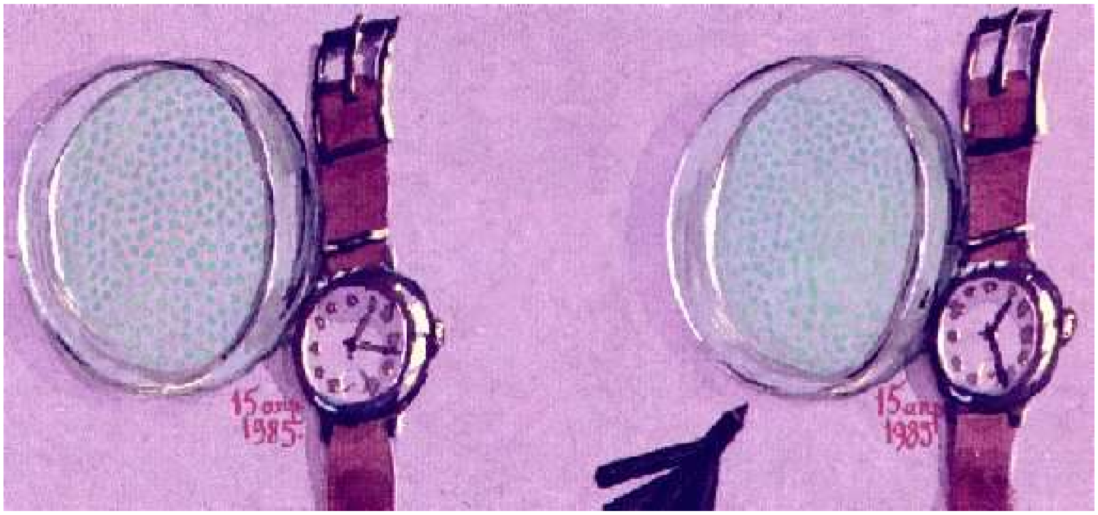

It turned out that clocks-both mechanical and electronic-placed in a strong CSE field started running inaccurately-Time must also have a part in it. All this was the manifestation of the Will of Matter, constantly moving, transforming, and eternally existing. It turned out that back in the 20s the French physicist Louis des Broglie was awarded the Nobel Prize for his discovery of these waves, and that the latter were used in electronic microscopes.

It turned out... well, many other things transpired in my experiments and research, but they would lead us into solid-state physics, quantum mechanics, elementary-particle physics, i. e far away from the main characters of our narrative: insects...

Meanwhile, I did manage to devise instruments for an objective registration of the CSE-instruments that accurately reacted to the proximity of insect nests.

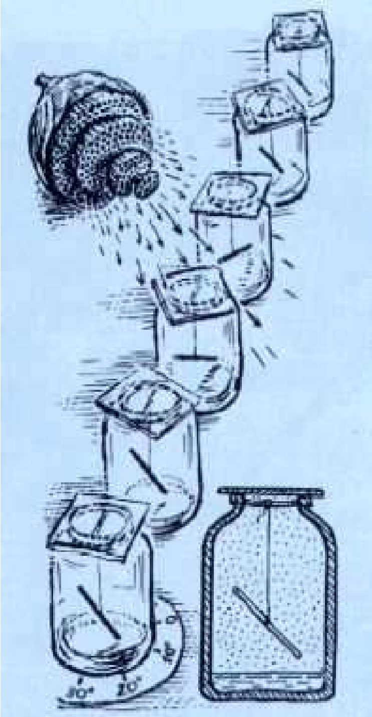

Here they are in the drawing: sealed vessels with straws and burnt twigs-drawing coals-suspended on spider web threads. There is some water at the bottom to counter static electricity hindering experiments in dry air.

If you point an old wasp nest, a bee honeycomb, a bunch of cereal ears to the upper end of the indicator, it slowly moves a few dozen degrees...

There is no miracle here: the energy of scintillating electrons of both multi-cavernous bodies creates a total wave system in space, whereby a wave is energy capable of performing a mutual repulsion of these objects-even through obstacles, such as a thick-walled steel capsule (see photograph).

It is hard to imagine that its armor is powerless to stop waves of a tiny, light wasp nest seen in the picture, and that the indicator inside this heavy, solid capsule "runs away"-sometimes as far as 180 degrees-from this long-vacant nest. Yet it is so. Those who have doubts are invited to visit the Agroecology Museum near Novosibirsk-you'll see it for yourselves.

The same museum displays an always-active honeycomb painkiller. It is a chair with an overhead cap that has a few empty, but intact combs of the honeybee ("dry" honeycombs, in the beekeeper vocab) in it. Anyone who sits in this chair will after a few minutes almost certainly feel something (please write to me what exactly you feel, I'll be grateful), while those with a headache will in just a few minutes say goodbye to the pain-at least for a few hours. My painkillers are successfully used in many parts of the country-I made no secret of my discovery.

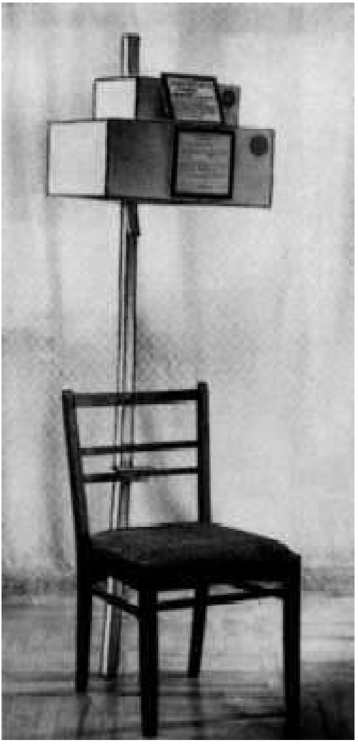

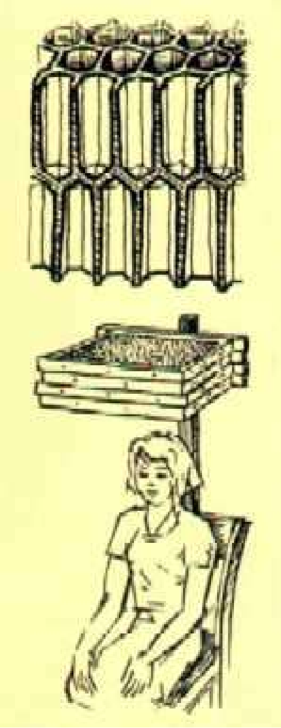

The hand will clearly sense the emanation if you take it from below, palm up, to the cap with bee honeycombs. The cap could be made of cardboard, veneer, or better still, of tin plate with tightly sealed seams.

Yet another gift from insects...

This was my reasoning at first: people have been dealing with the honeybee for thousands of years, no one has ever complained of anything unpleasant, except of course stings. I held a dry honeycomb over my head-it was working!

I decided to use a set of six frames. Such was the story of my rather simple discovery. An old wasp nest works quite differently, even though the size and shape of its cells are very close to those of bees.

The important difference was that the honeycomb material, unlike that of wax, is more crumbly and micro-porous: it is paper-like (by the way, it was wasps that invented paper, not people: they scrape old wood fiber and mix it with their sticky saliva).

Walls of the wasp honeycomb are much thinner than those of bees, the cell size and pattern are also different, as is the outer shell, also made of multi-layered, loosely wrapped paper. I had reports of a highly unpleasant effect of a few wasp nests in an attic. And besides, most multi-cell devices and objects that will manifest CSE in the first few minutes have a far from beneficial effect on humans. Honeybee combs are a rare exception. And when in the 1960s we had bumblebees living in our Isilkul apartment, I often observed the following.

A young bumblebee on its first trip away from the hive did not take the trouble to remember the entrance and would spend hours wandering around the windows of our house and of a similar-looking house nearby. And in the evening, giving up on its poor visual memory, it would land on the brick wall, precisely outside the hive and would try to break right through it. How did the insect know that right there, four meters away from the entrance, and a meter and a half below, behind the thick, half-meter wall was its home nest? At the time I was lost in conjectures, but now I know exactly why the bumblebee behaved like that. An amazing find, wouldn't you agree?

Now let us remember the experiment in which hunter wasps returned not just to a given location, but to an entirely different place where the lump of soil with their nest had been moved: no doubt, they were able to find it because of a wave beacon created by the nest cavern. And there was another mystery revealed to me by my insect friends. It turned out that to attract their pollinators, flowers use not only color, odor, and nectar, but also a similar wave beacon, powerful and unstoppable.

I discovered it with a drawing coal-a burnt twig-by passing it over large, bell-shaped flowers (tulips, lilies, amaryllises, mallows, pumpkins).

Already at a distance I could feel a "braking", as it were, of this detector. Soon I was able to find a flower in a dark room standing one or two meters away from it-but only if it had not been moved,

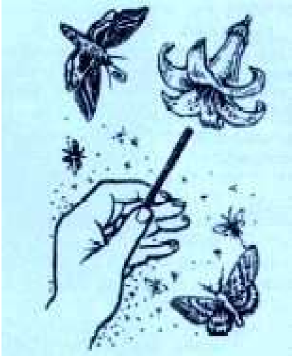

because a "false target" would be left in its old place-the "residual phantom" I already mentioned.

I do not possess any supersensory abilities, and any person after some training would be able to do the same. Instead of coal one could use a 10-cm-long piece of a yellow sorghum stem, or a short pencil whose rear end should be facing the flower.

Some people would be able to feel the flower (a "warm", "cold", or "shivering" sensation emanating from it) with their bare hands, tongues, or even faces. As many experiments demonstrated, children and adolescents are particularly sensitive to Waves of Matter.

As for bees that nest underground, their "knowledge" of the CSE is vital for them first of all, because it enables the builder of a new gallery to stay away from a neighboring nest. Otherwise the entire bee-city cut through with intersecting holes would simply collapse.

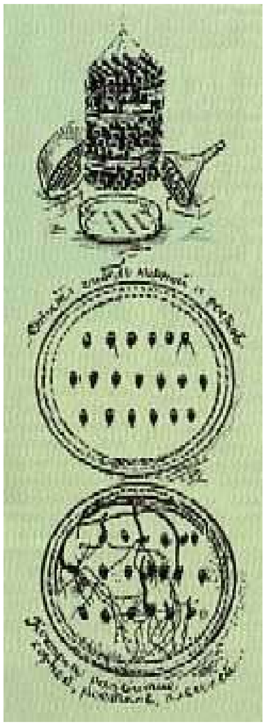

Secondly, plant roots cannot be allowed to grow down into the galleries and honeycombs. Thus roots stop a few centimeters away from the honeycomb, or else, feeling that nests are near, they start growing aside.

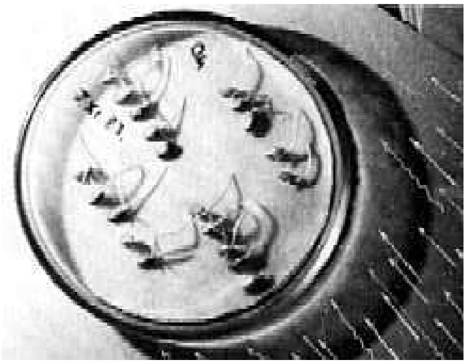

The latter conclusion was confirmed by my many experiments on sprouting wheat seeds in a strong CSE field, as compared to seeds germinating in the same climatic conditions but in the absence of the CSE.

Photographs and drawings show both the dying of roots in the experimental batch and their sharp deviation in a direction away from my "artificial honeycomb".

Thus bees and weeds back at the lake had long ago made a pact-another example of the highest ecological expediency of all Being. And in that same spot on the globe we see yet another example of people's mercilessly ignorant attitude to Nature...

The bee-city is now gone; every spring thick streams of fertile black earth soil run down, between filthy heaps of trash, to the lifeless, salty puddles that not too long ago were a string of lakes with countless flocks of sandpipers and ducks, white swans, and hovering fish-hawks. And by the steppe thinned out by bee holes, one used to hear the hum of hundreds of thousands of bees that for the first time led me into the Unknown.

I must have tired the reader with all these honeycombs of mine... A separate thick book would be required to describe all my experiments. Therefore I will only mention one thing: my pocket, battery-powered calculator often malfunctioned in the CSE field: it either erred, or sometimes its display window would fail to light up for hours. I used the field of a wasp nest combined with that of my two palms. None of these structures had any effect in isolation.

I will also note that hands with their tubular phalanxes, joints, ligaments, blood vessels, and nails are intensive CSE emanators capable of giving a powerful push to the straw or coal indicator of my little instrument from a couple of meters' distance. Practically anyone could do it. This is why I am convinced that there are no people with supersensory abilities, or rather that all the people have them... And the number of those who from a distance can move light-weight objects on a table, hold them suspended in the air or "magnetically" attached to the hand is far greater than is usually thought. Try it yourself! I look forward to your letters.

There once was an ancient folk game: one man sits on a chair, and over his head, four of his friends "build" a grid of horizontally stretched palms with slightly spread fingers-first right hands, then left, with 2 cm gaps between them. In 10-15 seconds, all four synchronously put their pressed-together index and middle fingers under the armpits and under the knees of the sitting man, and then they energetically toss him up in the air. The time between "collapsing" the grid and tossing the man must not exceed two seconds; the synchronicity is also very important. If everything is done right, a 100-kilo man flies up almost to the ceiling, while the ones who tossed him claim he was light as a feather.

A strict reader may ask me how it is possible. Doesn't it all contradict laws of nature? And if so, am I not propagating mysticism? Nothing of the sort! There is no mysticism, the thing is simply that we, humans, still know little of the Universe which, as we see, not always "accepts" our, all too human rules, assumptions, and orders... Once it dawned on me: the results of my experiments with insect nests bear too much similarity to the reports of people who happened to be in the vicinity of... UFOs. Think and compare: temporary malfunctioning of electronic devices, disrupted clocks-i. e., time, an invisible, resilient "obstacle", a temporary drop in the weight of objects, the sensation of a drop in human weight, phosphenes-moving, colored flashes in the eyes, a "galvanic" taste in the mouth...

I am sure you have read about all this in UFO journals. I am now telling you it can all be experienced in our Museum. Come visit! Was I standing on the threshold of yet another mystery? Quite so. And again I was helped by chance, or rather by my old insect friends. And again there were sleepless nights, failures, doubts, breakdowns, even accidents... And I had no one to turn to for advice-they would have just laughed, or worse...

But I can say this, my reader: he is happy who has a more or less adequate use of his eyes, head, and hands-skillful hands are particularly important!-and trust me, the joy of creative work, even of work that ends in failure, is far higher and brighter than earning any diplomas, medals, or patents.

Flying an Anti-gravitational Platform

(*excerpts from a diary)*

Judge it for yourself from my diary excerpts-obviously simplified and adapted for this book. Pictures and drawings will help you to evaluate my story... A hot summer day. Far-away expanses drown in a bluish-lilac haze; the sky's gigantic cupola with fluffy clouds stretches above the fields and coppices. I am flying about 300 meters above ground, with a distant lake-a light, elongated spot in the haze-as my reference point.

Blue, intricate tree contours slowly recede; between them, there are fields. Those, bluish-green ones are fields of oats; the whitish rectangles with a strange, rhythmic twinkling are those of buckwheat. Straight ahead of me is a field of alfalfa-its green color is familiar, it resembles the oil paint "cobalt medium-green". Green oceans of wheat on the right are of a denser shade and resemble the "chrome oxide" paint. An enormous, multi-colored palette floats further and further backwards.

Footpaths meander between fields and coppices. They join gravel roads which it turn stretch further out, toward the highway, still invisible from here for the haze, but I know that if I flew on the right side of the lake, I would see it-a smooth, gray strip without a beginning or an end, on which cars-small boxes-are slowly crawling.

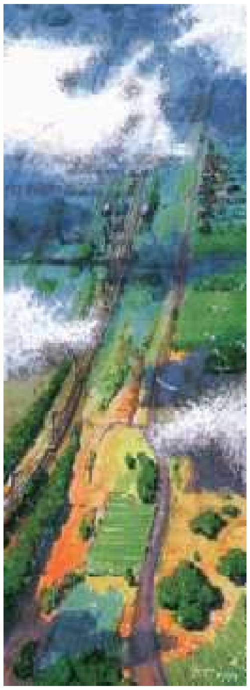

Isometric, flat shadows of cumulus clouds are picturesquely spread around the sunny forest-steppe. They are deep-blue where they cover coppices, and are various shades of light blue over fields. Now I am in the shadow of one such cloud: I accelerate-it's quite easy for me to do that-and leave the shadow.

I lean forward slightly and feel a warm, taut wind coming far down below, from the sun-warmed ground and plants. It comes not from the side, as on the ground, but strangely from the surface up. I physically feel a thick, dense current with a strong odor of blooming buckwheat. Of course this jet can easily lift up even a large bird-an eagle, a stork, or a crane-if it freezes its spread wings. But I have no wings and am suspended in the air not by the upward jet.

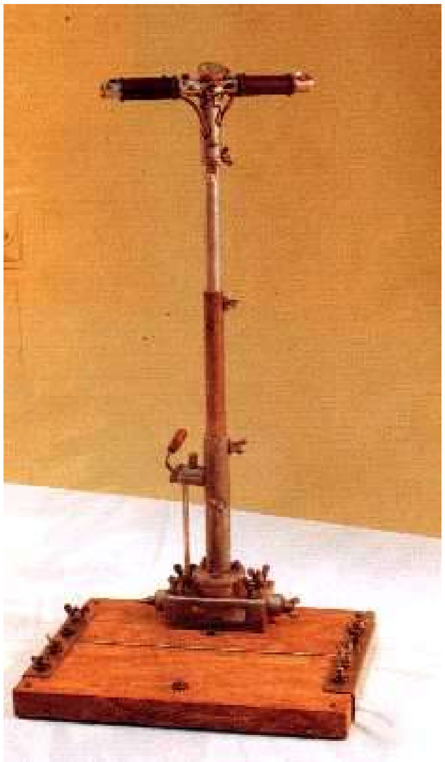

In my flight I am supported by a flat, rectangular little platform, slightly bigger than the seat of a chair, with a pole and two handles to which I hold on and with whose help I navigate the device. Is this science fiction? I wouldn't say so...

In a word, the interrupted manuscript of this book was abandoned for two years because generous, ancient Nature, again through my insect friends had given me another Something-and it did so, as usual, elegantly and inconspicuously, yet swiftly and convincingly. And for two years the Discovery did not let me go, even though it seemed to me I was mastering it at a break-neck speed.

(Note: Grebennikov would have been approximately 62-63 years of age in 1990-1992)

But it always happens this way: when your work is new and interesting, time flies twice as fast. A light spot of a steppe lake is already much closer. Beyond it, the highway is visible with already distinctly discernable boxes of cars. The highway is about 8km away from the railway that runs parallel to it, and if I look closer, I can see the poles of the power line and the light-colored embankment of the railway. It is time to turn some 20 degrees to the left.

I am not seen from the ground, and not just because of the distance: even in a very low flight I cast almost no shadow. Yet, as I found out later, people sometimes see something where I am in the sky-either a light sphere, a disk, or something like a slanted cloud with sharp edges that moves, according to them, not exactly the way a cloud would.

One person observed a "flat, non-transparent square, about one hectare in size"-could it have been the optically enlarged little platform of my device?

Most people see nothing at all, and I am for the moment pleased with it-I can't be too careful! Besides, I still haven't determined what my visibility or invisibility depends on.

Therefore I confess that I consciously avoid people in my flight and for that purpose bypass cities and towns, and even cross roads and footpaths at high speed, after making sure there is no one on them.

In these excursions-no doubt, fictional for the reader, but for me already almost casual-I trust only my insect friends depicted on these pages.

The first practical use of my discovery was-and still is-entomological: to examine my secret places, to take a picture of them from above, to find new, still unexamined Insect Lands in need of protection and rescue. Alas, Nature established its own, strict limitations on my work: just as on a passenger plane, I could see but couldn't photograph.

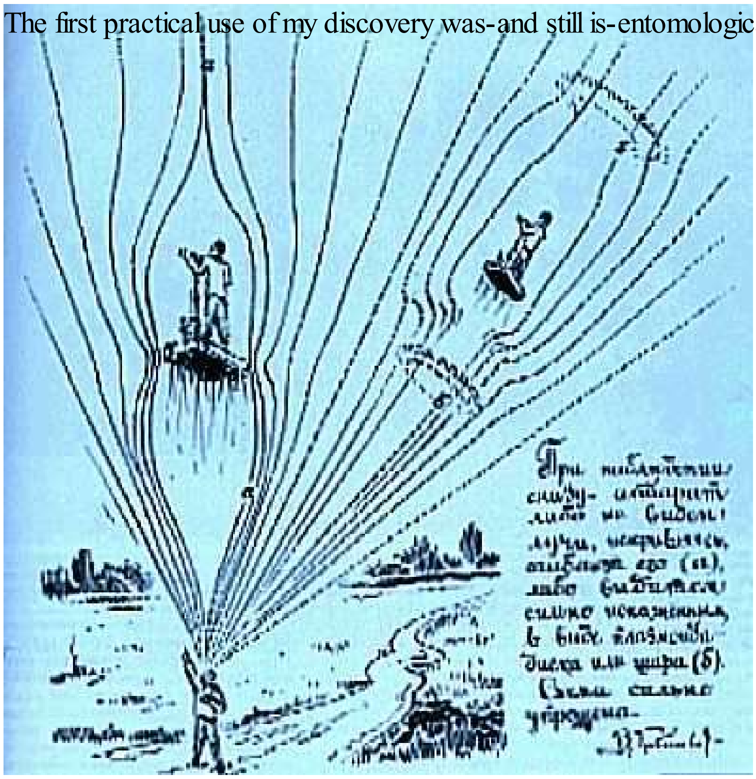

My camera shutter wouldn't close, and both rolls of films I had taken with me-one in the camera, the other in my pocket-got light-struck. I didn't succeed in sketching the landscape either; as both my hands were almost always busy, I could only free one hand for a couple of seconds. Thus I could only draw from memory. I managed to do that only immediately after landing-though I am an artist, my visual memory is not that great.

In my flight I did not feel the same way we do when we fly in our sleep.

It was with flying in my sleep that I started this book a while ago. And flying is not so much pleasure as it is work, sometimes very hard and dangerous. One has to stand, not hover, the hands are always busy, and a few centimeters away there is a border separating "this" space from "that", on the outside.

The border is invisible but very treacherous. My contraption is still rather clumsy and resembles perhaps... hospital scales. But this is only the beginning!

By the way, besides the camera, I sometimes had trouble with my watch and possibly, with the calendar too: descending on a familiar glade, I would occasionally find it slightly "out of season", with a two-week deviation, and I had nothing to check it against.

Thus it is possible to fly not just in space but also-or so it seems-in time as well. I cannot make the latter claim with a 100% guarantee, except perhaps that in flight, particularly at its beginning, a watch runs too slow and then too fast, but at the end of the excursion starts running accurately again.

This is why I stay away from people during my journeys: if time is involved alongside gravitation, I might perhaps accidentally disrupt cause-and-effect relations and someone might get hurt.

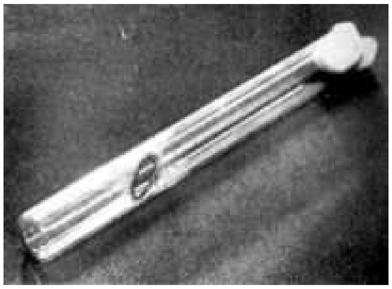

This is where my fears were coming from: insects captured "there" disappear from test tubes, boxes, and other receptacles.

They disappear mostly without a trace. Once a test tube in my pocket was crushed to tiny bits, another time there was an oval hole in the glass, with brown, as though "chitin" edges-you can see it in the picture.

Many times I felt a kind of burning or an electric shock inside my pocket-perhaps at the moment of my prisoner's "disappearance".

Only once did I find a captured insect in the test tube, but it wasn't the adult ichneumon with white rings on its feelers, but its... chrysalis, i. e. its earlier stage. It was alive-it moved its belly when touched. Much to my dismay, it died a week later.

It is best to fly on clear summer days. Flying is much more difficult when it rains, and almost impossible in winter-not because of the cold. I could have adapted my device accordingly, but since I am an entomologist, I simply do not need winter flights.

How and why did I come to this discovery? In the summer of 1988, as I was examining under a microscope the chitin shells of insects, their pinnate (feathery) feelers, and the thinnest structure of butterflies' wings, I got interested in an amazingly rhythmical microstructure of one large insect detail.

It was an extremely well-ordered composition, as though pressed on a complex machine according to special blueprints and calculations. As I saw it, the intricate sponginess was clearly not necessary either for the durability of the detail, or for its decoration. I had never observed anything like this unusual micro-ornament either in nature, in technology, or in art.

Because its structure is three-dimensional, so far I have been unable to capture it in a drawing, or a photograph. Why does an insect need it? Besides, other than in flight, this structure at the bottom of the wing case is always hidden from the eye-no one would ever see it properly. Was it perhaps the wave beacon with "my" multiple cavernous structures effect? That truly lucky summer there were very many insects of this species, and I would capture them at night: neither before, nor after was I able to observe these insects.

I put the small, concave chitin plate on the microscope shelf in order again to examine under strong magnification its strangely star-shaped cells. I again admired this masterpiece of nature, and almost purposelessly placed it on top of another, identical plate that had the same unusual cells on one of its sides.

But no!-the detail broke loose from my tweezers; for a few seconds it hung suspended above the other plate on the microscope shelf, turned a few degrees clockwise, slid to the right, turned counterclockwise, swung, and only then abruptly fell on the desk.

You can imagine what I felt at that moment... When I came to my senses, I tied a few panels with a wire-it wasn't an easy thing to do, and I only succeeded when I positioned them vertically. What I got was a multi-layered chitin block. I put it on the desk.

Even a relatively large object-such as a paper tack-could not fall on it-something pushed it up and aside. When I attached the tack on top of the "block", I witnessed such incredible, impossible things (for example, the tack for a few moments was lost from sight) that I realized it was no beacon, but something else entirely.

And again I got so excited that all the objects around me became foggy and shaky. It was with a huge effort that I managed to pull myself together in a couple of hours and continue working.

So, this is how it started. Of course, much still remains to be understood, verified, and tested. I will certainly tell my readers about the finer details of my machine, about its propulsion principles, about distances, heights, speeds, equipment, and all the rest-but in my next book.

...I conducted my first, very unsuccessful and highly dangerous flight on the night of March 17, 1990. I didn't have the patience to wait till the warm season and neglected to go to a deserted area. I already knew that night was the most dangerous time for this kind of work.

I had bad luck from the very beginning: the panel blocks of the right part of the bearing platform periodically got stuck. I should have fixed the problem immediately, but neglected to do so. I took off right in the middle of the Agricultural Academy campus, erroneously assuming that at 1 AM everyone was asleep, and nobody would see me.

The lift-off went well, but in a few seconds, when the lit windows of buildings sank beneath me, I felt dizzy. I should have landed right then but remained airborne, which was wrong because a powerful force snatched away my control over the movement and weight, and it pulled me in the direction of the city.

Drawn by this unexpected, uncontrollable power, I crossed the second circle of nine-story buildings in the city's residential area (they are laid out in two huge circles with five-story buildings, including ours, inside them), then I crossed a snow-covered, narrow field, and the Academy City highway... The dark immensity of Novosibirsk was closing in upon me, and it was closing in fast. I was already near a bunch of tall factory chimneys many of which fumed thick smoke-night shift was on. I had to do something quickly.

I got on top of the situation only with a great effort. Finally I managed to conduct an emergency adjustment of the panel blocks. My horizontal movement slowed down, but then I again felt sick.

Only at fourth try did I succeed in stopping the horizontal movement, at which point my platform was hanging over Zatulinka, the city's industrial district. The sinister chimneys silently continued to fume right underneath me.

I rested for a few minutes-if one could call hanging over a lighted factory fence rest-and after I made sure the "evil power" has passed, I glided back-yet not in the direction of our Agricultural Academy campus but to the right from it, toward the airport. I did this to foul the trail, in case someone had seen me.

Only about halfway to the airport, over some dark, night fields where there was clearly no one around, I abruptly turned home... Next day I naturally couldn't get out of bed.

News on TV and in newspapers was more than alarming. Headlines, such as "UFO over Zatulinka" and "Aliens again?" meant that my flight had been detected. But how! Some perceived the "phenomenon" as glowing spheres or disks-many actually saw not one sphere but two! Others claimed they had seen a "real saucer" with windows and rays.

I am not discounting the possibility that some Zatulino residents saw not my near-emergency evolutions, but something else entirely that had nothing to do with those. Besides, March of 1990 was particularly rich in UFO sightings in Siberia, near Nalchik, and especially in Belgium where, according to *Pravda*, on March 31 the engineer Marcel Alferlane took a two-minute film of the flight of a huge triangular craft which, according to Belgian scientists, were none other than "material objects with a capacity no civilization can currently create."

Is it really so? As for me, I would suggest that the gravitational filter platforms (or as I call them, panel blocks) of these machines were in fact small, triangular, and made here on Earth-but with more sophistication than my half-wooden contraption.

I too wanted to make the platform triangular-it is much safer and more efficient that way-but I chose a rectangular design because it is easier to fold, and when folded, it resembles a suitcase, a painter's case, or a briefcase that can be thus disguised so as not to arouse suspicion. I, naturally, disguised it as a painter's case.

I had nothing to do with the sightings in Nalchik or Belgium. Besides, as it may appear, I am very impractical in the use of my discovery-I only fly to my entomological preserves. These are far more important to me than any technological finds.

At the moment, I have eleven such preserves: eight in Omsk region, one in Voronezh region, and one near Novosibirsk. There used to be six of them in Novosibirsk region, all of them created, or rather salvaged by me and my family, but they don't like them here. Neither the Agricultural Academy (still more obsessed with "chemistry" than with anything else), nor the Environmental Protection Committee were willing to help me salvage these little preserves from evil, ignorant people.

Thus I am continuing my journey westward under the magnificent, fluffy clouds at noon. The blue shadows of the clouds, the intricately shaped coppices, and the multicolored rectangles of fields float backwards below me.

The speed of my flight is quite high, but there is no wind in my ears-the platform's force field has "carved out" from space an upward-diverging, invisible column that cuts the platform off the earth's gravitational pull. But it left me and the air inside the column intact. I think that all this, as it were, parts space in flight, and then closes it behind me.

This must be the reason for the invisibility, or the distorted visibility, of the device and its "rider"-as was the case with my flight over Novosibirsk's Zatulinka suburb.

But the protection from gravity is regulated, even though it is incomplete: if you move your head forward, you already feel the turbulence of the wind that clearly smells either of sweet clover, of buckwheat, or of the colored weeds of Siberian meadows.

I leave Isilkul with its huge grain elevator on my right and gradually begin to descend over the highway, making sure that I am invisible to drivers, passengers, and people working in the field.

My platform and I cast no shadow (although the shadow occasionally appears): I see three kids on the edge of a forest, go down, drop my speed, and fly right near them. They show no reaction, which means that everything is fine-neither I, nor my shadow are visible. Or heard: the propulsion principle of my device is such that the platform makes no sound whatsoever, because there is practically no air friction.

My journey was long-at least forty minutes from Novosibirsk. My hands are tired as I can't take them off the controls, so are my legs and body-I have to stand up straight, tied to the vertical pole with a belt. And even though I can travel faster, I am still afraid to do so-my hand-made machine is still too small and fragile.

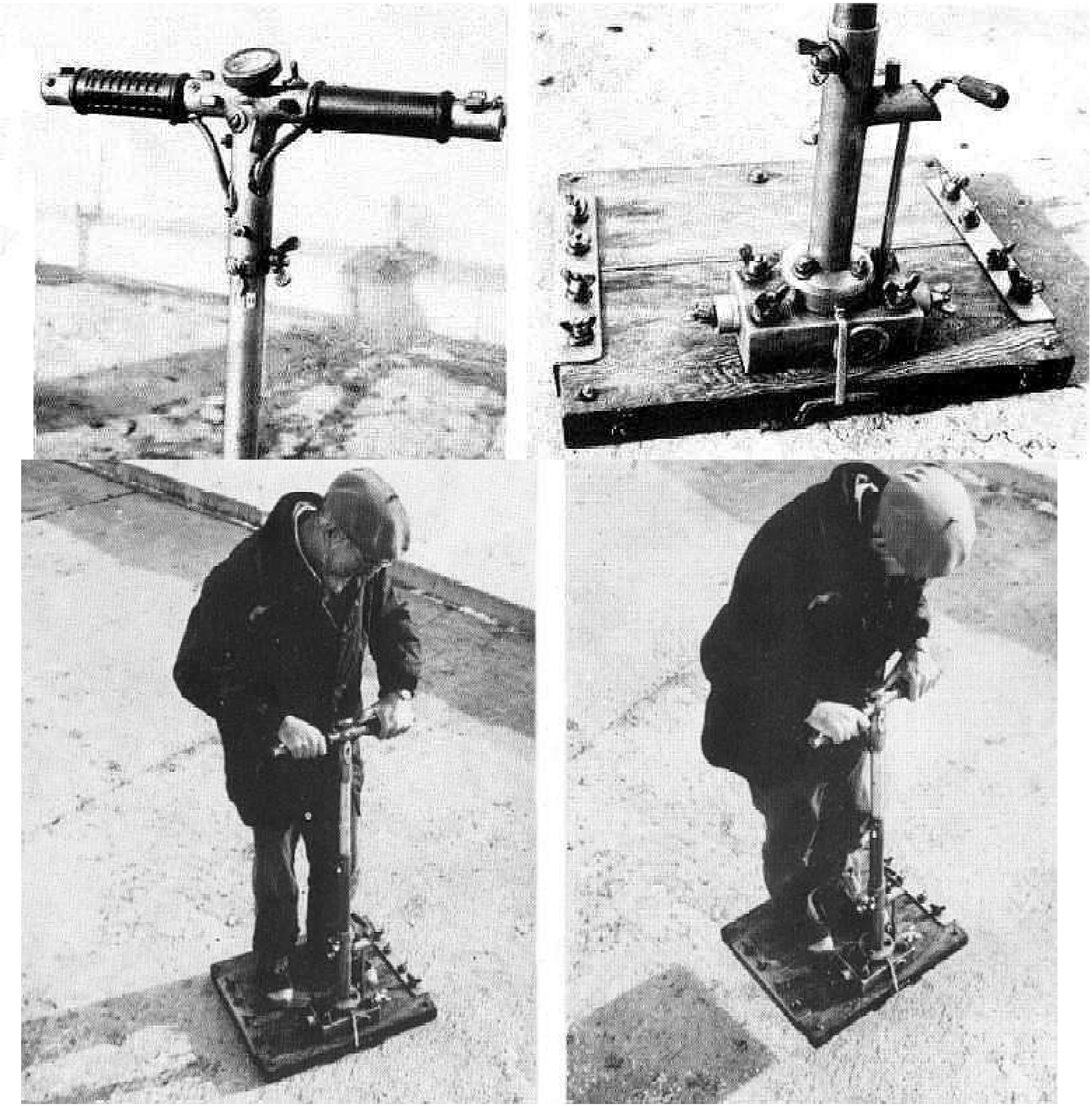

I again go up and ahead, and soon I see the familiar landmarks-a road intersection, a passenger terminal on the right side of the highway. Another five kilometers, and finally I see orange columns of the Preserve fence. The Preserve is this year-come to think of it-twenty years old! How many times I saved this child of mine from trouble and bureaucrats, from chemicals-loaded aircraft, from fires, and many other evil deeds. And the Land of Insects is alive and well!

Descending and braking, which is done by cross-shifting filter blinds under the platform board, I already see the thicket of carrot weed, make out the light heads of their flowers resembling azure balls-they are of course covered with insects, and an incredible joy comes over me, taking away my fatigue, for it was I who saved this piece of Earth, even if a small one, less than seven hectares.

Already for twenty years no one has driven here, no one has cut the grass, tended cattle, and the soil has risen in places to fourteen centimeters high. Not only several locally extinct species of insects have returned, but also such weeds as feather grass of rare varieties, purple Scorzonera whose large flowers in the morning smell of chocolate, and many other plants. I feel the thick smell of cuckoo flower-only this Middle Glade smells like that, it is right behind the fence of the preserve, and fills me with yet again with the joyful anticipation of another encounter with the World of Insects.

Here they are, I can see them very well even from ten meters above the ground on the wide umbrellas and azure balls of angelica and carrot plants: dark orange butterflies sit on them in groups; heavy hornets bow the white and yellow inflorescences of lady's bedstraws; ginger and blue dragonflies with trembling wide wings and a fine network of veins hover next to my head. I slow down even more, and see a sudden flash below: my shadow, hitherto invisible, has finally appeared and now slowly glides along weeds and bushes.

But I am already safe-there is not a soul around, and the highway three hundred meters north of the preserve is now empty. I can land. The stems of the tallest weeds rustle against the bottom of my "podium"-the platform with the panel blocks.

But before putting it down on a little bump, I, in a fit of joy, again spread the blinds with my control handle, and vertically go up. The landscape below quickly shrinks, shrivels as it were: the shrubs of the preserve, its edges and fences, all the surrounding coppices and fields. The horizon begins to curve on all sides in a huge groove, opening up the railroad that runs two kilometers on the left, then a village on the right-it twinkles with its light slate roofs.

Further on the right is Roslavka, the central estate of Lesnoy State Farm-it already looks like a small city. Left from the railroad are cow farms of Lesnoy's Komsomolsk branch; they are surrounded by a yellow ring of straw and dry, foot-worn manure. In the far west, where the smooth curve of the railroad disappears (this is actually confusing: the railway is straight as an arrow), there are small houses and the neat white cube of the Yunino railroad terminal, six km away. Beyond Yunino, there are limitless expanses of Kazakhstan drowning in the hot, bluish haze.

And finally here it is, below me-Isilkulia, the land of my youth; it's very different from how it appears on maps and plans with their inscriptions and signs. It is vast, limitless, alive, interspersed with dark, intricate islands of coppices, cloudy shadows, light, clear spots of lakes.

The huge disk of the Earth with all this for some reason appears more and more concave-I still haven't discovered the reason for this already familiar illusion. I go up higher, the rare, white cloud masses sink lower, and the sky is darker than below-it is dark blue. The fields visible between the clouds are already covered with a thickening blue haze, and it is more and more difficult to make them out. Too bad I can't take my four-year-old grandson Andrei with me; the platform could easily lift us both. Yet one can't be too careful...

... Goodness, what am I doing? I cast a shadow back on the Glade, didn't I? This means I can be seen by thousands, as on that memorable night in March. It is day now, and I may again appear as a disk, square, or worse, my own person... There is also a cargo plane, still soundless, coming straight at me, quickly growing in size; I already see the cold shimmer of its body and the pulsation of its unnaturally red blinker.

Down, quick! I brake abruptly, make a turn; the sun is at my back; my shadow should be across from me, on the gigantic, convex wall of a white cloud. But there is none, only a multicolored glory, an iridescent, bright ring familiar to all pilots has brushed the cloud ahead of me.

I sigh with relief-this means nobody saw either me, or my "double" in the guise of a triangle, square, or a "banal" saucer... A thought occurs to me (I must say that despite the desperate technical and physical inconvenience, imagination works much better and faster in a "falling" flight): what if I am not the only one out of five billion people to have made my discovery; what if flying devices based on the same principle-both home-made and professional-have long been constructed and tested?

But all screening platforms have the same quality: sometimes they become visible to other people; pilots too are "transformed"-they are seen as "humanoids" in silver costumes, either short and green, or flat as if made of cardboard (Voronezh, 1989), etc. Thus it may very well be that these are not alien UFO crewmen, but "temporarily deformed"-of course to outside observers-earthly pilots and builders of little platforms, such as mine, who have made their inventions reliable.

My advice to those who in their study of insects comes across the same phenomenon and begin making and testing a "gravitoplane" (by the way, I am convinced that one can't make the discovery without insects) is this: to fly only on fine summer days, to avoid working in thunderstorms or rain, not to get too far or too high, not take a thing with you from the landing area, to make all assembly units maximally strong, and to avoid testing the device in the vicinity of any power lines, towns (let alone cities), transport, or people.

The best site for testing is a distant forest glade, as far away from human habitation as possible; otherwise a phenomenon known as poltergeist could occur in the radius of a few dozen meters-"unexplained" movements of household objects, switching off, or on, of household electric appliances, and even fires.

I myself have no explanation for all this, but it seems that these phenomena are the consequence of temporal disruptions, a complicated and treacherous thing. Not a single, even tiniest fragment or particle should be dropped either during the flight, or in the landing area.

Let us remember the Dalnegorsk phenomenon of January 29, 1986-apparently a tragic one for the inventors, when the entire device was blown apart and scattered on a vast area, and only small shreds of filter cells were found, impossible to analyze chemically (as it should be!).

Remember, I wrote that insects taken "there" and moved "here" in a test tube disappeared, and a hole was formed in the tube, if it remained intact.

It turns out these holes resembled those in window glass; the latter sometimes appear in residential and office buildings, occasionally in "bursts" in the windows of several rooms and floors. A hole is 3-5 mm on the outside, widening in a cone to he inside, with exit diameter of 6-15 mm. Some holes are melted or colored in brown on edges-just as it happened in the case of my insect in a test tube.

It seems that this type of poltergeist is caused not, as I used to believe, by short-lived microplasmoids of a tiny ball lightening type, but by particles and specks carelessly dropped while testing a device similar to mine. The photographs of window holes on these pages are documentary and made by me at the scientific center of the Agricultural Academy near Novosibirsk. I can show them to anyone who wants to see them. These holes appeared during 1975-1990, but none of them, except perhaps the very last one, are related to my flights.

I am certain that part of UFO descriptions are actually those of platforms, panel blocks and other large parts of devices deliberately or accidentally taken out of the active field by their designers and makers. These fragments are capable of causing much trouble to others, or at best, to generate a series of improbable tales and stories in papers and magazines, often accompanied by "scientific" commentary...

Why am I not disclosing the particulars of my discovery at this time? Firstly, because one needs time and energy for proving the truth. I have neither. I know how daunting this task is from my own bitter experience of trying to get recognition for my previous discoveries, including such an obvious one as the Cavernous Structures Effect of whose reality you, my readers, I am sure, are by now convinced.

This was the result of my protracted, painstaking efforts to get the CSE scientifically recognized:

> "Any further correspondence with you on the subject of your patent application is counterproductive."

I know personally some of the High Priests of Science, and I am certain that were I ever to get an audience with one such person (which is now practically impossible), were I ever to;

- open my painter's case,
- attach the pole,
- turn the handle,
- and soar to the ceiling,

he wouldn't be a bit impressed-or worse still, would order the trickster out of the office. I look forward to times when young people will replace these "priests".

The second reason for my "non-disclosure" is more objective. I found these antigravitational structures only in one species of Siberian insects. I am not even naming the class to which this insect belongs-it seems to be on the verge of extinction, and the population surge I registered back then was possibly local and final.

Thus, if I were to name the genus and the species, what is the guarantee that dishonest people, half-way competent in biology, would not rush out to ravines, meadows, and forests to catch perhaps the very last samples of this Miracle of Nature?

What are the guarantees that they would not plough up hundreds of glades, cut down dozens of forests to get to this potentially lucrative prey? Therefore, let all I have related in this chapter and in the addendum remain science fiction; may Nature herself never reveal this secret to them-it would take a lot of effort, and they would never be able to get it by force as there are still several million insect species living on the planet.

Spend at least an hour on the morphological study of each of them, then calculate the odds of encountering the Unusual, and I will sincerely wish you diligence and a very long life, for even if you took no days off, working eight hours a day, you would need a thousand years of life.

I hope I will be understood and forgiven by those of my readers who wanted immediate information about my discovery not for selfish ends, but simply out of curiosity. Indeed, what would you do in my place if you were to act in the best interests of Living Nature?

Besides, I can see that similar inventions have been made by other people who are also in no rush to take their discoveries to bureaucrats' offices, preferring to fly across night skies in the guise of strange disks, triangles, or squares with chatoyant (iridescent) glimmer.

Falling down, or rather sinking down, I orient myself, look to see if there is anyone around. I brake abruptly about forty meters from the ground, and land safely where I always do-on a tiny glade in the Big Forest of the preserve. You won't find it on a map, and if you get there, you won't be able to find it either.

Don't judge me for the fact that the branches of several aspens there are cut or sliced "by lightening": The strictly vertical take-off and landing are very difficult, and the initial trajectory is for the most part slanted, particularly at take-off, when the platform is for some reason carried off away from the sun, and sometimes the other way around.

I loosen the screws on the control pole, then shorten it like an antenna of a portable radio, and remove it from the platform which I fold in half. Now it looks like a painter's case, a box for paints, if only a bit thicker. I put the case, some food, and a few tools for repairing the fence in my backpack and make my way for the Middle Glade between aspens and short dog-rose bushes. Even before I leave the forest, I see a good omen-a family of fire-red toadstools that have lined up on the forest bedding in a wide curve, or, as it used to be called in folklore, a "witch's ring".

Why "witch's"? And in general, why does one have to break, knock off, trample this beautiful mushroom of Siberian forests? I often asked mushroom-pickers why they do it. The answer was, "because it's inedible!" But turf, clay, twigs, tree stumps, and stones are inedible too.

If there were rocks lying in the forest instead of mushrooms, no one would be knocking them off. It seems that inedible mushrooms are knocked off because they are alive; they are knocked off only in order to kill them! What is this then?

Do people really have this in their blood-to knock off a mushroom, to crush a bug, to shoot a bird, a hare, or a bison? And is this not where boorishness, sadism, pogroms, and wars originate? One really wants not to believe it, but I put myself in the shoes of an alien: I come to Earth to visit humans and see them knock off mushrooms, crush insects, shoot birds and each other.

What would I do? I would immediately turn my spacecraft around and go back. I wouldn't return for at least 500 earth years... What would you do, my reader, if you were an alien?

It's a good thing at least that this little family of toadstools is hidden from evil eyes and cruel feet. Every summer it gives me joy to see its special life, its cinnabar-red, moist caps with large, whitish scales. But here is the Glade. I step on it, as usual, with my heart sinking with a constant longing for this dear, faraway nature of Isilkul, with a fear that some "master" might decide to plough it up, and with a joy that it is still unploughed, uncut, and untrampled...

And it really means nothing that in my backpack I have a folded, i. e. neutralized platform with gravitational, micro-cellular filter blocks, and between them, a folded pole with field regulators and a belt with which I fasten myself to the pole.

What difference does it make that I got about fifty years ahead of contemporary science with my discovery? People are still going to master this and many other mysteries of Matter, Space, Gravitation, and Time.

But no supercivilization on any planet of any Supergalaxy is going to re-create this very Glade with its complex, fragile, trembling Life, with its lady's bedstraws, meadow sweets, and feather-grass...

Where else, in what corner of the Universe are you going to find a match for this lilac-blue bellflower in whose semi-transparent entrails two flower flies are doing their love dance? On what other planet would a nearly tame blue butterfly land on your outstretched hand to have a taste of something salty-sausage, cheese, or a pickle? Or else, just to walk up and down your palm, opening and closing its gray wings on whose backside there is a fine ornament of round eye-shaped spots?

...It hasn't been too long since we, humans, started flying-first air balloons, then airplanes, and now powerful rockets that we send to other heavenly bodies. What next? Next we are going to fly to other stars at a speed close to that of light; but even the closest galaxy would still be out of reach.

Yet Humankind-if it earns the name of Intelligent-will solve many riddles of the Universe and will then overcome that hurdle too. Then any worlds of the Universe will become accessible, close-even if they are trillions of light years away. It'll happen, for it is all a matter of Reason, Science, and technology. But of nothing else. Only this Glade may disappear if I-and there is no one else to rely on-am not going to preserve it for my close and distant descendants.

So what is more valuable to Humanity at this time-the insect preserve or the home-made device capable of developing the zenithal pull of at least 100 kg and the horizontal speed of 30-40 km/min? I am asking *you*, my reader. But think hard before you give a serious, responsible answer.

[illegible] these pictures. [illegible] rather simple device in assembly. A flexible cable inside a steering column trasmits movement from the left handle to the gravitational blinds.

By joining or parting these "wing cases", I lift off or land. Once I lost the left handle in a free-falling descent and would have been in a better world if the platform hadn't dug out a rather deep well in the tillage-first a vertical one, then a horizontal, facing away from the sun.

Thus I not only survived, but also felt almost no impact-just darkness. I extracted myself and my fairly badly damaged device from this well-but not without efforts as the "well" had no slag heaps!

I had to use all my ingenuity to disguise it. If seen from the road, it would have caused much speculation, and may even have led some over-zealous investigators to the culprit. Similar wells-also with a side-tunnel and without slag heaps-were suddenly formed on October 24, 1989 in the fields of Khvorostyansk District of Samara Region. *Komsomol'skaya pravda* described it in detail on December 6 of the same year. It seems I am not alone. And quite likely am "inventing a bicycle". Well, actually the top part of my device looks very much like one: the right handle is used for horizontal, onward march achieved, also via a cable, by the incline of both groups of "wing case" blinds. I never fly faster than 25 km/min, preferring to go ten times slower.

...I don't know whether I have persuaded you, my reader, that similar devices will soon be available to practically everyone, while Living Nature which humans cannot survive without won't be available to anyone if we don't save it.

But I don't want to seem entirely greedy and will give researchers another Patent of Nature, one also related to Movement and Gravitation. Physicists say that an unsupported mover is impossible. In other words, a device completely isolated from the environment won't fly or drive-a car won't go without outer wheels, a plane won't fly with a covered propeller or engine, neither will a rocket with stopped nozzles. Baron Münchhausen who managed to pull himself up by the hair from a mire is the only exception.

It happened near Novosibirsk in 1981 when we were studying the entomo-fauna of alfalfa-its pollinators and pests. Walking along the field, I was "mowing" alfalfa with an insect net and was then moving its contents-insects, leaves and flowers-into a glass jar. Such is the cruel method of studying the insect make-up of fields, no better one has been invented. Alas, such was the work with which I earned my living at the Institute of Agricultural Chemistry. I was about to throw a piece of ethered cotton wool into the jar and then cap it, when a light little cocoon jumped up at me.

It was oval-shaped, rather dense and non-transparent. One of the "prisoners" of the jar must have pushed it-cocoons can't jump on their own! But the cocoon proved me wrong: it jumped up one more time, hit the glass wall, and fell down. I took it out and put it in a separate test tube. At home I looked at it through a binocular microscope-nothing special, a cocoon like any other, about 3 mm long, 1.5 mm wide. Its walls felt strong to the touch-as they should. The cocoon energetically jumped when lit-or warmed?-by the sun; it was quiet in the dark. Its could jump 30mm longwise and, what was even more remarkable, up to 50mm high. As far as I could tell, it flew smoothly, almost without tumbling. No doubt, the larva of the insect was responsible for the movement. But it was impossible to see how it was happening.

Jumping ahead, I can tell you that the cocoon finally produced a male insect of the ichneumon family, the *Batiplectes anurus* species. It is beneficial for agriculture because its larvae parasitize the weevil, a pest of alfalfa.

The flying cocoon was finally to land in a cool place-for example, a crack in the ground. It must have found myself in my net during its strange journey, i. e. at the moment of its jump. It all resembled poltergeist-unexplained "jumps" of household objects, many times described in papers. I would put it on glass and look at it from below: could it be that the larva draws in its bottom and then abruptly releases it? Nothing of the kind-there were no dents at any point, and the cocoon jumped no matter how I rolled it. It was also remarkable that from horizontal, slippery glass it jumped sideways.

I measured its trajectories: they were up to 35 cm long and up to 50 mm high, that is, the cocoon lifted itself up to a height 30 times its own width! Shall I leave this capsule without support? But how? With a piece of loose cotton wool! I loosen a piece of cotton wool by pulling it a little, place the cocoon on this "cloud", put it out in the sun and impatiently wait. If the cocoon's inhabitant jumps by hitting the lower wall, making the cocoon to bounce off its support, this time it won't work because the impact will be absorbed by the thin, paddy fibrils of the cotton wool. Theoretically, the cocoon shouldn't even move. But no: it takes off from its motionless pad-up and aside, as it did before.

I measure the broad jump: 42 mm, i. e. as before. The insect must have been hitting not the bottom, but the top part of the cocoon-at any rate, it must have been doing something that caused the capsule to move.

Frankly speaking, it is as I write these notes that I feel agitation; back then, in 1981, I found nothing supernatural in the jumps of my prisoner. This was because I knew that, according to physics, there are and there can be no unsupported movers. Otherwise I would have bred a couple of hundred of those insects-thankfully, they were quite common-and would have studied the phenomenon thoroughly.

Now let us fantasize a little: what if the batiplectes wanted to leave the Earth? An adult, winged insect would have no luck-our atmosphere is quite rarefied at the top, wings are no match for it. A larva in a cocoon is an entirely different matter. It could, in theory, after lifting its capsule 5 cm in a jump, take it up even further while in the air, then again and again...

And if the cocoon were airtight-I mean the air reserve for the pilot's breathing-then the device would be able to leave the atmosphere and would have no obstacles to a limitless build-up of speed. Such is the alluring, incredible value of unsupported movers, declared, alas, a product of empty fantasy. But even if you are no physicist, you still have a hard time imagining what a tiny larva does in there if its vessel soars 5 cm high. It simply can't be-and yet it jumps!

Physicists say that this is "beyond science" as it "contradicts the laws of nature." The only problem is that the *Batiplectes anurus* doesn't know it. The physicists' ban must also have been unknown to the leading, experienced biologists who honestly wrote the following on page 26 of the academic *Register of Insects of European USSR* (vol. III, pt. 3): "the cocoon jumps up as a result of abrupt movements of the larva inside the cocoon." In a word, it is a working-and tested-example of a safe, unsupported mover. I am giving it to you, my reader: invent, design, build, and Godspeed! But hurry!

Massive chemical warfare has been waged against the alfalfa pest snout-beetle (phitonomus). Humanity may actually win it. But the price may be too great: with the destruction of the *Phitonomus varnabilis* beetle, our planet's fauna may also lose the ichneumon *Batiplectes anurus* as it parasitizes only this kind of weevil and cannot survive without it.

Meanwhile, any proposals on using biological weapons against the pest-such as our very ichneumon and other insect predators are completely rejected by the bosses of Russian agriculture and agricultural science. I have been fighting them on this for years, but so far with little success.

However, one could understand those in charge too-how can one stop expensive chemical factories? And why do agrarian scientists care about some unsupported mover that doesn't allow alfalfa to be treated with poison? Hurry up, biologists, engineers, physicists! For if Chemistry wins, this Mystery-and with, a host of other Mysteries related to it-will leave people for ever. Without insects, people won't invent it themselves. Please trust me, an entomologist with 60-year experience.

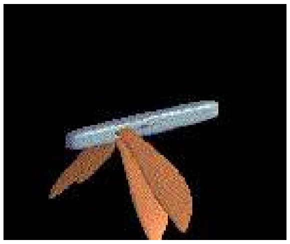

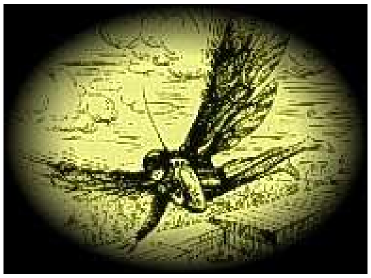

At the end of my first book, *A Million Riddles*, published in Novosibirsk in 1968, there is a drawing that I am reproducing again: a man is flying over Novosibirsk's Academic City. He is flying a device based on a huge pair of insect wings.

At the time I dreamed of inventing such a machine. Strangely, the dream came true precisely because of my friendship with insects-yet not by blindly copying the most noticeable parts-for example wings that only make me smile now-but through careful study of living Nature.

Nothing would have been possible without my six-legged friends. No one would be able to do without them either. Thus safeguard their world, the ancient, wonderful world of Insect, for it is an infinite, unique treasure of Nature's mysteries! I beg you all, take care of it!

FROM THE NOTEBOOK OF A NATURALIST

Artificial honeycomb.

Take a dozen and a half papier-machet supermarket egg cases (30-egg variety), tie them up or glue them together (one on top of another) in such a way as to join the "teeth" to one another, not to the hollow spots.

You will have large cells, similar to multi-cellular combs of a certain "paper" wasp, except many times larger. Fix the whole set (it can be enclosed in a case) over the head of a person sitting in a chair, with the bottom "comb" is 10-20 cm above the head.

Leave the person there for 10-15 min. The "unnatural", unusual transformation of the spatial shape formed by the set can be picked up even by the palm of a hand.

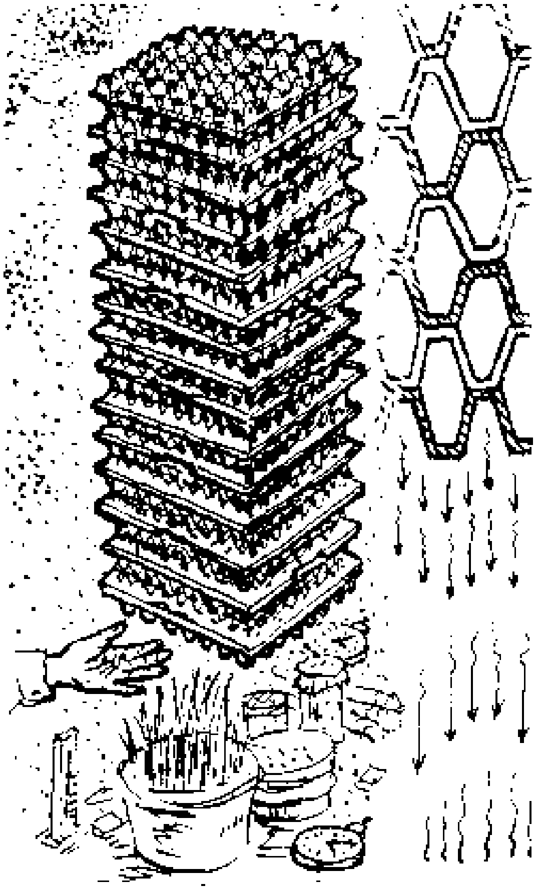

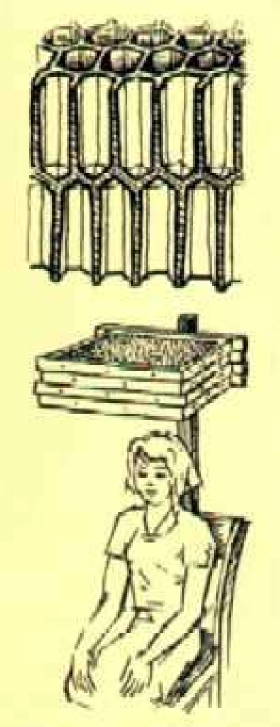

Experiment with couching seeds, or breeding microorganisms and insects under a "macrocomb" and compare the results with those of identical experiments conducted at least 2 m away from the comb. Repeat each pair of experiments several times.

"Iron comb". in a similar way, test the impact of common kitchen shredders piled up one on top of another, with their wire-edges down, small-hole shredders at the bottom, large-hole ones at the top.

Paper emanators of the CSE.

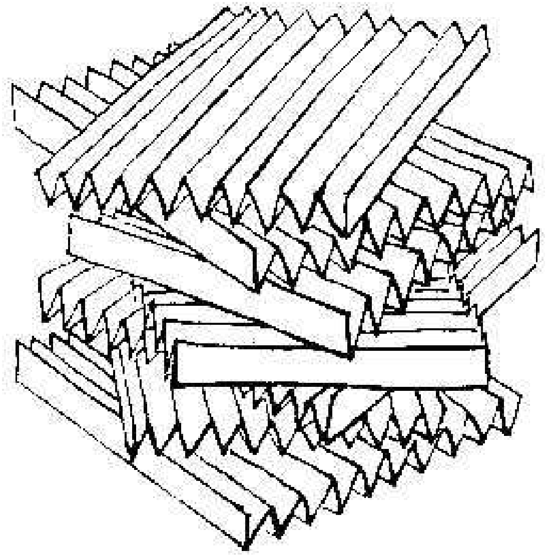

Cut apart-longways-5 sheets of office paper and fold each of them accordion-like so as to get 10 edges and 20 surfaces. Squeeze the accordions to make them square and glue them on top of one another, turning each horizontally 30 degrees clockwise against the bottom one.

Glue together, preferably out of dark paper, a conical, multi-layered "flower" with a few dozen petals; fluff up the petals.

Test the emanators with by putting your palmon above the "flower" and underneath the suspended "accordion". Place them above the head of the sitting person, marking his sensations.

Foam plastic. We are used to the fact that this excellent thermal insulator "reflects" the warmth of the hand even at a distance. But even if you cover it with dark paper, cardboard, or a tin plate, it would still do the same. This happens due to the work of multiple vesicular caverns of the material that produce the CSE.

Foam rubber. It is widely known that a person used to sleeping on, say, a cotton wool mattress doesn't sleep well at first on a rubber foam one, or else is unable to sleep at all. This is a typical manifestation of the CSE. Later on, the organism adapts itself to this new bed.

"Mushroom CSE". A hunter once told me that he warms up his hands in winter on bracket-fungi. Let us recall that the lower horizontal part of this tree fungus is full of fine comb-tubes through which spores fall out in summer. What the hunter felt was not warm but a typical CSE.

Moving "combs".

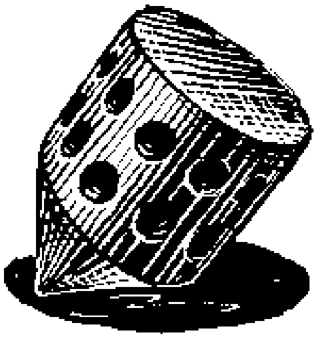

Make a wooden whipping top and drill several through caverns on its sides, pencil-width in diameter or a bit wider. Their CSE significantly increases if the top is rotated; this is easily perceived by the palm.

This is due to the fact that the caverns must be numerically multiplying in space.

"Flower CSE". An "unnatural" position of such a seemingly common and pleasant object as a living flower can also change its properties. Put a bunch of several dozen bell-shaped flowers (tulips, narcissuses, lilies, bell-flowers) upside down above the head of a sitting person. To bar the impact of odors, put the flower bunch in a plastic bag. Write to me about the results.

Wind-fallen trees. One of my test subjects, a geographer, said to me after experiencing the effect of one of my "grids" that he had once had a similar sensation many years before, when he was passing a wind-fallen section of a forest. His head, ears, mouth, and the entire body felt something particularly unpleasant-the same thing he felt under my grid. This means that the abruptly disrupted shape of the normal multi-cavernous space of the forest for some time emanated waves unpleasant for humans.

Before the rain. Place a shower nozzle on a tap and run cold water. Slowly move your hand toward the drops coming from the sides: most people feel "warmth". In reality, this is the CSE reinforced by the movement of ever new elements of the "multi-layered" grid-water drops and gaps between them. After practicing in the bathroom or kitchen, pick up an even stronger CSE from fountains and waterfalls. Even when the atmospheric pressure is high, the shroud of a distant rain creates a powerful CSE field that has its impact on a large area. Have you ever felt sleepy before the rain even in enclosed premises? The CSE cannot be screened.

"The CSE of a book".

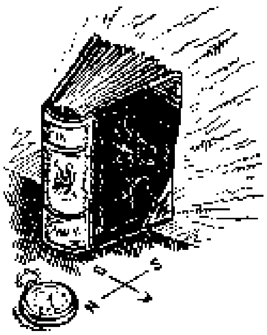

Take a thick, preferably well-read book and place it upright on the edge of a desk with its back facing the direction of the sun (e.g. facing north at night).

Open the book and fluff up its pages as evenly as possible. In a few minutes (the CSE does not appear immediately, as it doesn't disappear immediately ether), pick up with your palm, tongue, or back of your head some of the sensations mentioned in this chapter.

This "tail", after some practice, can be picked up at a 2-3 meters' distance. It is also easy to verify that the "book CSE" is also non-screenable-ask someone to stand between the hadn and the book.

"A large cone" with an artificial comb filling and three magnets at the back.

Two cones of this sort, positioned against each other taking into account the position of the sun, were in the morning of April 23, 1991, thrown apart and disfigured. One had been placed in Isilkul, the other near Novosibirsk (the second one was unfolded and pressed into the wall of an underground hiding place; its magnets disappeared).

At the same moment, residents of an Omsk apartment experienced a series of strangest "poltergeists" (see *Vechernii Omsk* for April 26 and Omsk and Moscow TV broadcasts). Because of this coincidence, on August 5, 1991, the same paper called the device in the picture "a Grebennikov's hyperboloid".

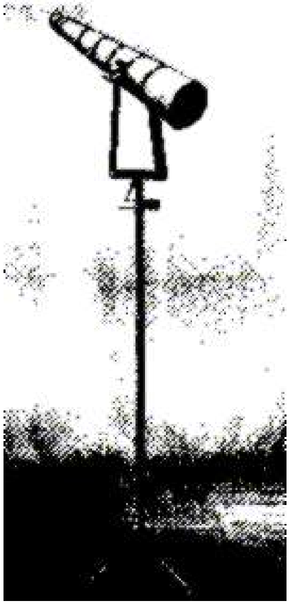

Actually, one of the "beams" of the upright electronic waves between the two structures may have been formed precisely there, on the Irtysh embankment in Omsk.

"A medium cone".

Tightly insert a dozen plastic household funnels into each other and fix the structure on any support with the nozzles turned toward the sun. Cover the bell end of the top funnel with a net or light blue cloth (so that the tested subjects not anticipate heat).

"A small cone". Tightly roll up two unusable rolls of film, tied them up with string or thread and make a bell-shaped cavity in the middle of the top roll. CSE emanations can be easily picked up by the palm, particularly in a counter-solar position. You will get interesting sensations if you press this "microcone" to your forehead.

"Perpetuum mobile".

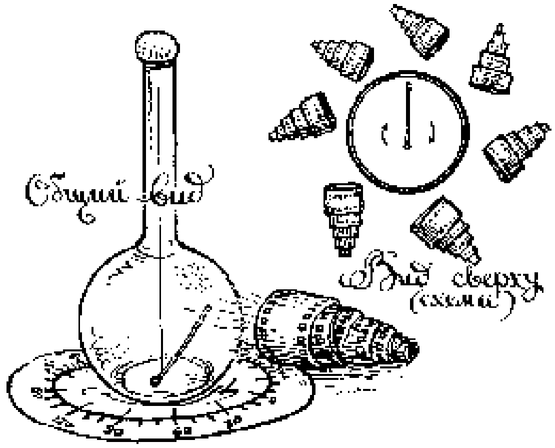

I surrounded my above described device [for registering CSE emanations; a straw indicator suspended on a cobweb thread] with seven funnel-shaped rolls of film [see above].

Slowly leaving the zone of impact of one roll, the straw would enter the power field of another, then the third, and so on...

This experiment is most successful in a sound-insulated chamber, away from wires, pipes, sources of heat, cold, and bright light. There is no miracle here: matter is eternal in its endless movement.

A solar ether- and beam radiator.

This intricate name was devised by the Leipzig professor Otto Kornschelt who discovered the CSE over 100 years ago and produced devices for its practical application in medicine, agriculture, and technology. Rhythmic caverns were formed in them by cooper chains.

The devices were positioned with their backsides facing the sun. It is indeed true that new inventions are simply well-forgotten old ones. The sensations described by Kornschelt are identical to the ones I experienced in my own independent work.

I learned about Korschelt's experiments very recently from M. Platten's *New Medical Technique*, vol. III, St. Petersburg, 1886, where the following drawing of the device is reproduced.

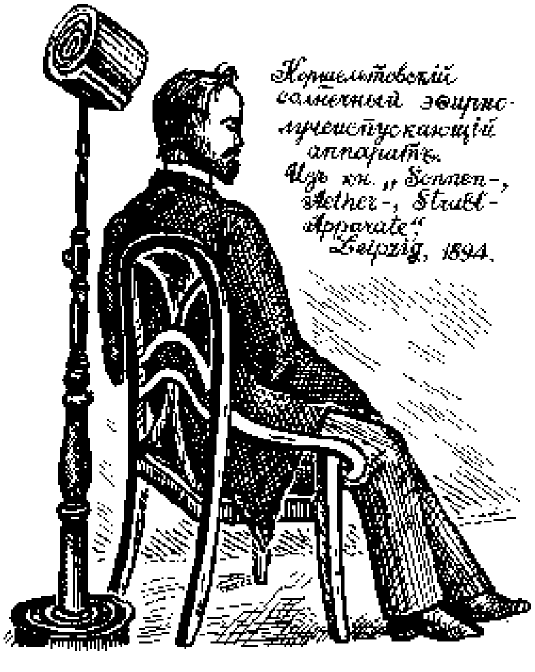

"The sieve CSE". In the old days, in several areas headaches and concussion symptoms were treated with an ordinary flour sieve that was held above the head of the patient, net up. Alternatively, the patient squeezed the rim between the teeth, with the net in front of the face. The sieve material is unimportant. The device works better if the patient faces the sun (north at midnight). This type of CSE is also perceivable by healthy people.

The CSE and the planets. The planets of our Solar system are situated at certain distances from the sun. The Ticius-Bodet formula for the distance is this: 4 is added to the numbers 3, 6, 12, 36, etc (a geometrical progression), and the resulting number is divided by 10. The cause for this regularity is unknown. The empty spot in this progression (between Mars and Jupiter) is occupied by asteroids. The Kemerovo physicist V. Iu. Kaznev thinks that the regularity is determined by the CSE generated by the sun: the matter of planets was grouped in the areas of the sun's field force concentration.

The CSE in daily life. Perceivable waves of matter are emanated by piles of pipes, some caves, underground tunnels, tree crowns; the shape of premises is also significant (round, cornered, cupolaed). The wall and furniture material also emanates a CSE of certain parameters.

"Micro-CSE". The CSE effect may be manifested not just in galactic or household scales, but also in micro-world, in substances whose molecules have caverns of certain shapes. For example, in naphthalene. I filled a one-liter jar with it, sealed it, and suspended it from the ceiling. People beneath it felt with their palms a whole system of power field "clots". (more son if the receptacle was suspended above the top of the head). Activated charcoal is also a multi-cavernous structure. Hold 2-3 tablets of such charcoal in your fingers as demonstrated in the picture and for a few minutes move your hands slightly up and down, or parting and joining them. Write to me about the results.

Tefelin.

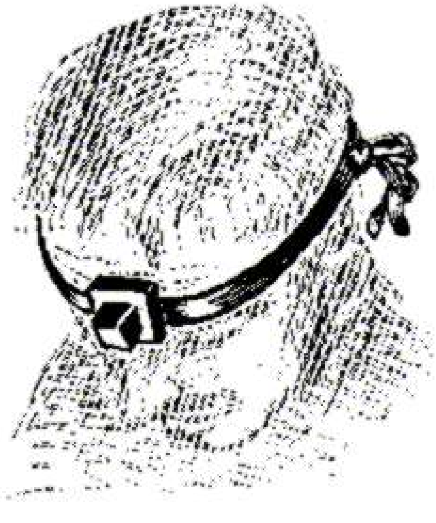

I have so far isolated 4 CSE emanators beneficial for humans: bee honeycombs, a grid of joined hands (more about it in the next chapter), a sieve, a phylactery otherwise known as tefelin.

What is it? An old device: a tightly sown leather cube attached to a leather platform with two bands. Inside the cube there are four strips of parchment-tightly rolled, bleached, soft kidskin with Talmudic inscriptions. A worshippers attached the device to his forehead, with the axes of parchment rolls perpendicular to the forehead and their outer ends facing East. It turns out, the inscriptions were unimportant; what matters is the material, shape, and dimensions.

Made of different materials, the device only causes unpleasant sensations, while a leather tefelin produces a beneficial physiological effect-besides the shape and other factors, the microstructure of the material must have a part in it too.

Thot's Scepter.

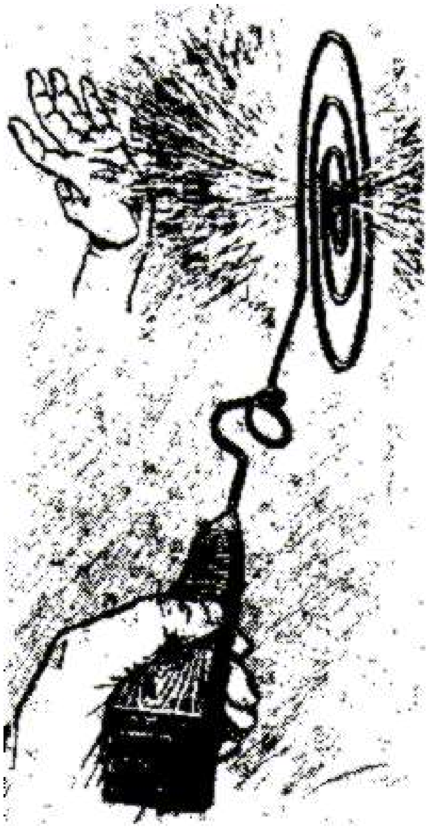

The ancient Egyptian deity Thoth is a god of science, sorcery, and an "accountant" of the dead's earthly deeds. This is the design of his staff: 2- or 3mm copper wire is twisted at the end in the shape of a flat spiral, with 3-4 coils, each 10 cm in diameter'; closer to the handle there are 2 coils of transverse, 3-dimensional spiral, each 5 cm in diameter.

The wire is inserted in the 16-cm-long square-sectioned handle of dense wood, 4 cm thick at base and 1.5 cm thick at its end; the entire staff with the wire is 41 cm long. The narrow end of the handle has 13 deep accordion-shaped cuts.

The staff works even without the wire (albeit not as strongly); the wire is thin and could be of any material but works best if thickly insulated-two layers increases its effect.

If you hold the staff as demonstrated in the picture, the total radiation emanating from the center of the large spiral, perpendicular to its surface, are very well-perceivable by the human palm on both sides.

I never found out for what purpose ancient Egyptians used this "double-beam" emanator.

The Pyramid of Cheops.

Make a pyramid of 3-4 layers of thick, porous wrapping paper: 202x20 cm square base, ascending edges 19cm each. Glue it only at the edges, the tighter the better, but in a thin line.

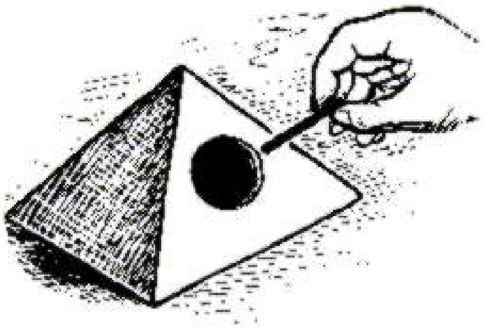

Make a 5-6 cm hole in the middle of one of the side facets. Hold a 10 cm-long piece of drawing coal in your fingers, or simply a pencil, and insert this indicator into the hole, slanting the other end toward the bottom of the pyramid. "Stir" the space inside the pyramid with the indicator, take it out, then repeat the procedure about 30 times.

You will soon pick up an active zone-a "clot"-where the Egyptians had their tombs. Another active zone, above the top of the pyramid, is also well-perceived by the indicator if you drag its end over he top. After some practice, the "clot" and the "torch" are well-felt by the finger inserted into he pyramid, or a palm moved above it.

The pyramid effect that generated over the centuries many scary, mysterious stories is one of the CSE manifestations.

The pyramid skeleton.

Similar interesting qualities are displayed by pyramids of the same dimensions but only skeletal,

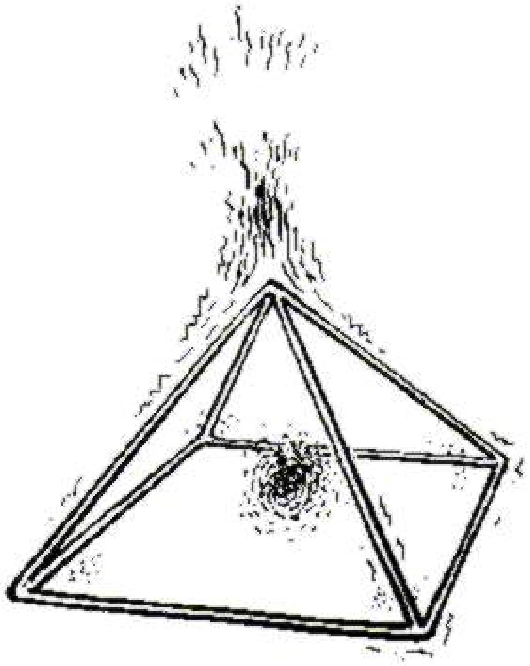

without facets-a skeleton glued together of 8 smooth, firm straws. Here we get the effect of the total CSE of the straws with their complex capillary structure and the effect of the entire cavern. Such pyramids can also be made in other sizes, with a proportional increase in the length of the edges.

Hold such a pyramid above the head of your friend, bottom down, for about 5 min, then bottom up. Conduct additional experiments with insects (bumblebees, developing caterpillars, etc.), house plants, and perishable foods by placing the latter within the pyramid, above and underneath it (always checking your experiments by identical ones but without the CSE effect). And you will see that ancient Egyptians had their reason to build pyramids...

Telekinesis.

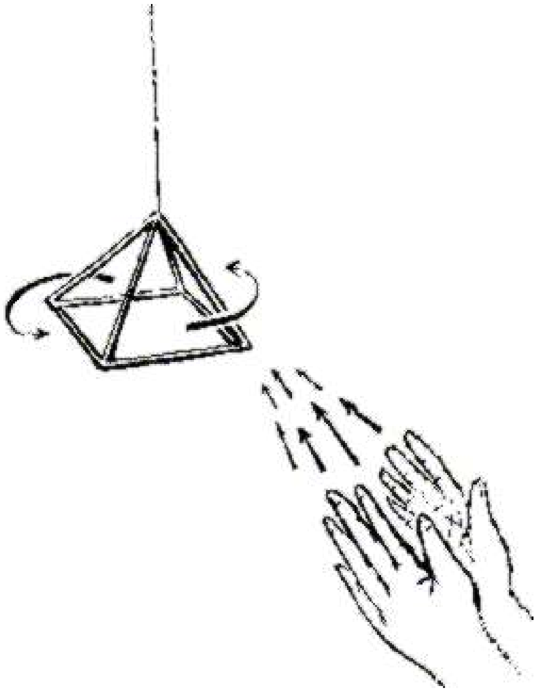

This is the name for a contactless movement of light objects of which the so-called gifted people are capable-i. e. moving a match box on a table without touching it, holding a tennis ball in the air...

I submit that everyone has this ability. Suspend the described skeletal pyramid by its top from the ceiling on a thin, artificial thread, or better still on a long shred of elastic torn from a stocking. Choose a spot with the lowest convection (air circulation).

In a few hours, when the pyramid stops rotating, from a 2-meter distance point at it a "tube" made up of two hands (see picture). In a few minutes (do not lose your "target"), the pyramid will start rotating clockwise under the pressure of this beam of CSE energy. Stop the rotation by moving the "tube" to the right side of the skeleton-it will start rotating counter-clockwise.

Conduct experiments of various duration, after various time intervals and at various distance. You will see that telekinesis is no miracle, but only one of the manifestations of the Will of Matter that is available to a chosen few but to everyone. The thing is that the palm is also a multi-cavernous structure that clearly repels the indicator of the straw-cobweb device described in this chapter. By using it and the skeletal pyramid, you can practice and develop your "telekinetic" abilities, significantly increasing them.

"The CSE of cereals".

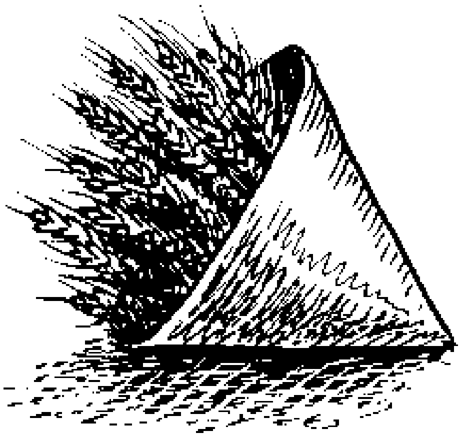

Fasten a bunch of 30-40 ripe wheat ears, better with short stems, inside a low cone of dark paper-as in the picture.

Hand-perceivable emanations repel the straw indicator of the same device through any screens-even sharper than some honeycombs.

This effect is produced by multiple wedge-shaped sinuses between ear scales that are directed at an acute angle toward the bottom of the ear.

Haymaking with "miracles".

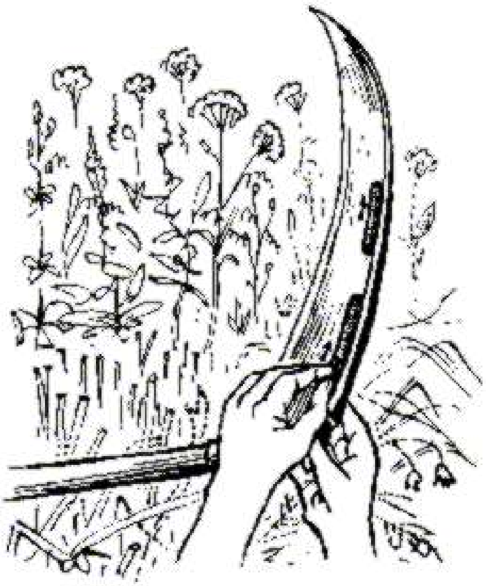

In my youth, I was shown the following: a fragment of a cut stem, the length of a short pencil, was placed on the blade of a scythe, next to its blunt edge; another stem fragment of the same length, placed on the blade in the same manner but at some distance, was pushed by the hand to the first one.

At about 8cm, the first stem got moving, "ran away" from the second stem along the rim. Th experiment wasn't always successful; it usually occurred right after the cutting a large amount of grass from the same spot; I forgot some elements or conditions of the experiment.

I think the following factors were at work here: an abrupt change of the total CSE field on the "deformed" meadow (let us remember the windfall case), the grid of the reaper's fingers, the multi-cavernous properties of the stem itself, and perhaps its position against the morning sun. Static electricity is excluded as everything at that hour is wet.

Identified Flying Objects. A long time ago, in a remote Caucasus village, I was surprised that people walk around the mountains at night, through dense forests. They all had lit cigarettes in their mouths, are all waving their hands, and their cigarette lights for a second disappear behind their bodies.... It turned out those weren't cigarettes, but local fireflies, *Luceola mingredica*; their light twinkle in this manner. Meanwhile, UFO reports and letters from my readers tell of dark flying saucers that turn out either a flock of birds or a compact swarm of insects . I myself saw in Siberia not just "columns" of insects but also "balls" , 3 or 4 meters in diameter: in one case those were some mosquito-like fliers, in another, winged ants of the *Mirmica* genus. From afar this swarm could by taken by an ignorant person for a huge, round plasmoid.

A detailed description of the CSE effect may be found in my book *Mysteries of the World of Insects* (Novosibirsk, 1990), in the journals *Sibirskii vestnik selskokhoziastvennoi nauki*, no.3, 1984, and *Pchlovodstvo*, no. 12, 1984. The physical nature of CSE is described in *Non-periodic Galloping Phenomena in the Environment*, vol. III (Tomsk, 1988). All in all I have published over three dozen articles on the CSE.

As promised, I will describe the rest in my next book. I will call it as I called this chapter: *Flight*.

NOTE from Jerry Decker : Victor S. Grebennikov died at the age of 74 in April 2001 as reported to KeelyNet in two phone calls to his surviving son Sergei.

This all came about from an email from a friend in Russia by the name of Youlain who sent in the URL some two years ago. When I saw the pictures of the platform, especially the one hovering over the ground I thought it was a hoax, but I couldn't stop thinking about it so translated the webpage in sections beginning with the technical descriptions.

As correlations began to fall into place, I hesitated reporting it because his writing indicated psychogenic effects associated with these structures, meaning it could be hallucinations, delusions or daydreams perpetrated with jumping into the air while on the platform as what the TMers (Transcendental Meditation) devotees claim as 'flight'.

However, other aspects of his claims led me to think he really had discovered something. With the idea of getting a copy of the book, I ended up sending some $200 to Yuri with the instructions being to take out the cost of the book and shipping, give the rest to the Professor. I did receive the book (Youlain bought his for $7.00 US but he lives in Russia.)

Meanwhile, I had made a 'secret' page on KeelyNet whose URL I mailed out to long time trusted confidants to see what they thought about it since I valued their insights, opinions and knowledge, asking them to keep it on the QT until I heard from the professor direct, if possible. From their responses, new information was added.

However, one of these long time confidants who over the years since the BBS days, shared much useful information, had now become infatuated with public attention and so posted the URL and claimed I was 'hiding' information in order to further his newfound career of public speaker and author. Very sad to see this but ego just took him over, so he's gone.

In the meantime, Yuri had kindly provided me with the Professors mailing address and I immediately wrote to him in Russian, sending at least 5 packets over the next year of correlated information to him and all translated into Russian.

In all that time I received only one letter from the Professor which stated he had suffered a stroke, that he had no degree and had learned what he knew from his experiences in 'gulags' (Russian prison of war) and was paralyzed on one side, having to type this letter

with one finger. At the time of his response, he wrote he had received 3 of my letters, indicating in the letter that his son had withheld the ones I sent in February until May before he actually saw them.

I asked his son in the phone call why they had not responded with but one letter in the past 2 years and he said the entire family had been ill, especially his father. Early in 2001 I asked Yuri if I flew over there and pre-arranged an interview with Professor Grebennikov, would he serve as translator and he agreed, that was roughly in February. It was only when I called Sergei in July, after the June KeelyNet conference, that I learned Professor Grebennikov was deceased.

The Professor had written that he would like me to republish his book in English and so I was trying to get him to sign a copyright release to pursue that project. After his death, the rights would have gone to Sergei, so I asked him if he would be willing to sign such a release for a percentage of book sales and he didn't seem to comprehend the matter over two phone calls and a detailed letter with release form. The calls were made in Russian as translated by my Russian translator Helena. She also wrote and sent the letter with copyright release to which Sergei has not responded as of December 6th, 2001.

It takes about two to three weeks for a letter to get from Dallas to his village near Novosibirsk and the same to get back. They don't offer UPS, FedX, or any faster means of snail communication and the Professor was not online, relegated to living on $24.00 US per month in his retirement. That is another reason I was considering publishing his book as his percentage would greatly improve his income level, that is irrelevant now.

In the book, of the sections we have in English, he asks only to be recognized by his peers and in the history books as the discoverer of the CSE phenomenon. He has of course refused to reveal the name of the insect and even the genus, it could be a beetle, a butterfly, a bee, a wasp, a locust, a grasshopper, a dung beetle. There are some 1100 species of beetles in Russia.

Professor Grebennikov was a great ecologist as you can tell by his protective attitude toward the environment. He was afraid if the name of the insect were revealed, people would rush to the area and kill off all of them. I wrote him in two separate letters that it was not necessary to know the insect name or genus.

In order to allow verification, all we needed were TWO of the insect covers which could be analyzed under an electron microscope to determine geometry and dimensions, from that point, it should be possible to duplicate it artificially and the insect would be forever safe.

I offered the Professor $1,000 US in cash if he could send me two of the covers (I sent money to pay shipping) AND they exhibited the gravity repelling characteristic he had seen. He indicates the effect was not magnetic, electrostatic or due to stray air currents.

He never responded which his son claims was due to illness.

At any rate, the entire book is currently being translated and my current plans are to extract technical information and compile it with other information which correlates and supports what the Professor claims.

This will be sold through KeelyNet to recover my personal expenditures and to support KeelyNet along with further research, donations are greatly appreciated.

There is a lot of weirdness going on with this. I asked Sergei if he could provide a couple of

the covers, he said he didn't know what they were or where they came from. I then asked him what happened to his fathers flying platform, he said (in a hesitant and halting fashion) his father had hidden it and he didn't know the location.

We have since read an email from a Russian scientist who visited Professor Grebennikov in the hospital and claims he was told firsthand that he had smashed the platform to pieces. Sergei also said many people had been contacting he and his father since the publication of the book some 6-7 years ago, with everyone wanting the secret, but his father had never revealed it.

Numerous posts have been made to the KeelyNet Interact discussion list on cavities, nanostructures, arrays, chitin, dielectrics and related subjects for those who are interested in this claim.

Additional information has been posted in the Chaos file, the Time Dilation file and in dual presentations (Decker & Davidson) made at the June KeelyNet 2001 conference in Dallas (available on videotape). See also Dan Davidsons' experiments with shape power and Joe Parrs original discovery of energy bubbles in pyramidal structures.

Many are involved in this and have widely varying views. At this time, we have no verfiable experimental evidence of the late Professor Grebennikovs wide-ranging and earth-shaking claims. As you can see, they span all branches of science, philosophy and theosophy, shedding quite new and novel views on the nature of gravity, time and paranormal experiences.

We will continue to correlate information and try to pin down experiments that can verify various aspects of his claims. He has certainly provided many experiments that seem to rely mostly on human perceptions.

We need to find ways to detect these emanations with instrumentation, exactly the same problem we have with Aether/ZPE, we need a kind of attachment for a FLUKE meter and a scope to be able to qualify and QUANTIFY these kinds of energies.

I would like to add one further comment, please do not confuse time dilation with time travel. They are not remotely the same. Time dilation simply slows down the local temporal field of an object, kind of like a 'stasis' field on Star Trek so that anything outside this local area moves at the normal speed, while anything inside this field ages at a slower rate. Refer to the Time Dilation file.

The Professors' friend and associate is Yuri N. Cherednichenko.

If you have comments, ideas or information relevant to the above, please post it to the KeelyNet discussion list or send it to Jerry Decker for posting to the list, thanks!

Additional fascinating information relating to weight loss in sleepwalkers was provided by Slavek Krapelka and with full details posted at his website at;

Amazing Sleepwalking Experiments

Date	Time (Aproximate to 10 minutes, more or less)	Weighted (Aproximate to 500g, more or less)	Normal Weight	Type of Scale	Report Number
03/08/00	23h 40m	**43.248Kg**	*41.600Kg*	Electronic	*1-8-00*
12/09/00	02h 20m	**43.876Kg**	*41.600Kg*	Electronic	*1-9-00*
18/09/00	00h 25m	**5 Kg**	*41.600Kg*	Mechanical	*2-9-00*
21/09/00	00h 50m	**7.200Kg**	*41.600Kg*	Mechanical	*3-9-00*
22/09/00	04h 10m	**4.700Kg**	*41.600Kg*	Mechanical	*4-9-00*
27/09/00	00h 20m	**5.300Kg**	*41.600Kg*	Mechanical	*5-9-00*
28/09/00	01h 35m	**5.600Kg**	*41.600Kg*	Mechanical	*6-9-00*
"/""/""	" ""	**78.800Kg**	*41.600Kg*	Mechanical	*6-9-00*
11/10/00	22h 40m	**41.500kg**	*42Kg*	Mechanical	*1-10-00*
15/10/00	02h 20m	**6.100Kg**	*42Kg*	Mechanical	*2-10-00*
16/10/00	03h10m	**5.200Kg**	*42Kg*	Mechanical	*3-10-00*
17/10/00	00h 25m	**5.900Kg**	*42Kg*	Mechanical	*4-10-00*
18/10/00	05h 10m	**3.800Kg**	*42Kg*	Mechanical	*5-10-00*
23/10/00	04h 45m	**3.600Kg**	*42Kg*	Mechanical	*6-10-00*
27/10/00	01h 40m	**4.300Kg**	*42Kg*	Mechanical	*7-10-00*
28/10/00	02h 50m	**5.300Kg**	*42Kg*	Mechanical	*8-10-00*
09/11/00	01h 40m	**3.200Kg**	*42Kg*	Mechanical	*1-11-00*
10/11/00	03h 10m	**00000Kg**	*42Kg*	Mechanical	*2-11-00*
18/11/00	00h 35m	**00000Kg**	*42Kg*	Mechanical	*3-11-00*
23/11/00	04h 40m	**4.300Kg**	*42Kg*	Mechanical	*4-11-00*
27/11/00	00h 10m	**4.900Kg**	*42Kg*	Mechanical	*5-11-00*
28/11/00	01h 40m	**4.200Kg**	*42Kg*	Mechanical	*6-11-00*

Again, we want to publicly acknowledge the efforts of Slavek Krapelka and Ren Berghuis for spending the $300.00 to get the Chapter 5 file translated and passing it on.

If you would like to contribute something to help them recover their costs, please send an email to Ren Berghuis for information on how to do so. THANKS!!!

So FEW people actually contribute with anything more than comments or looking for free material researched or validated by others.

There is more information which will be added to this page or linked in future.

Additional information from our translator that might be of use;

> **"The only specific insect I've seen mentioned thus far of possibly interest was the one I identified in the portion of the 'Flight' chapter you saw, the bee, 'Halictus quadricinctus'. It seems that particular subspecies is native to Siberia, although most of the other references sound fairly generic."**

A picture of the mudnest and the Halictus Quadricinctus photos I found on the web;

1. **Trigona (Heterotrigona) lutea Bingham**
2. **Trigona (Heterotrigona) laevicpeps Smith**
3. **(Micrapis) florea Fabricius**
4. **Trigona (Heterotrigona) vidua Lepeletier**
5. **Ctenoplectectra davidi Vachal**
6. **Apis (Megapis) laboriosa Smith**

7. Eucera pekingensis Yasumatsu
8. Tetralonia chinensis Smith
9. Habrophotula nigripes Wu
10. (pupae)
11. (larvae)
12. Xylocopa (Biluna) tranquabarorum (Swederus)
13. Halictus camellia Wu
14. * Halictus quadricinctus Fabricius *****
15. ??
16. Osmia excavata Alfken

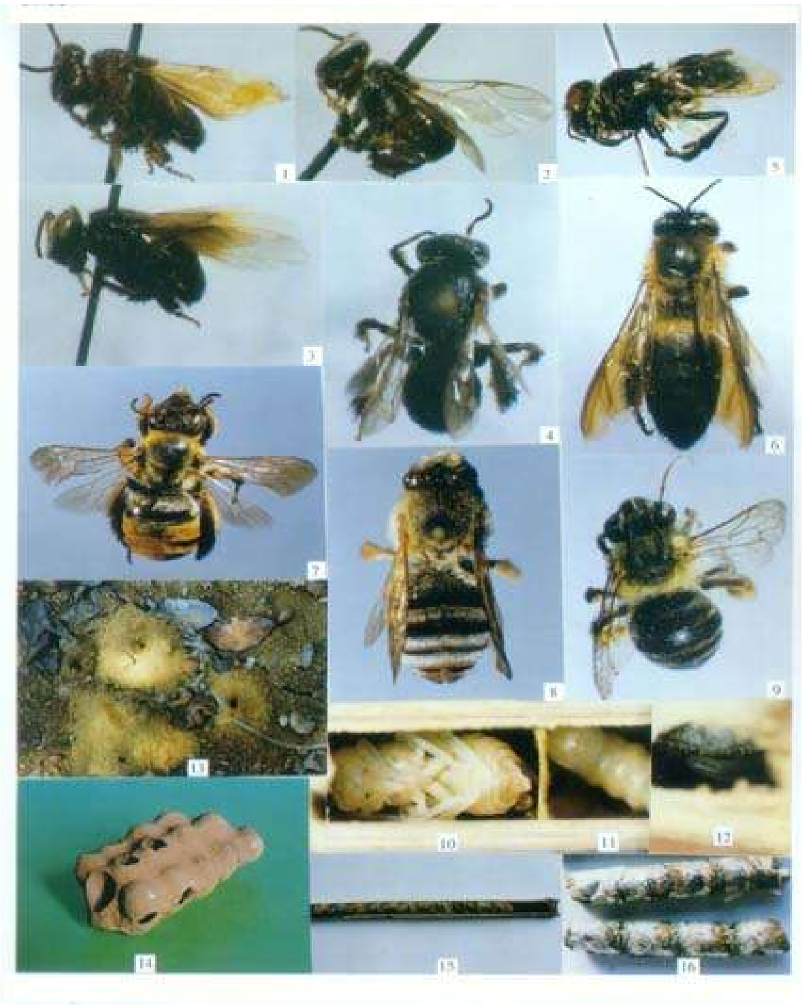

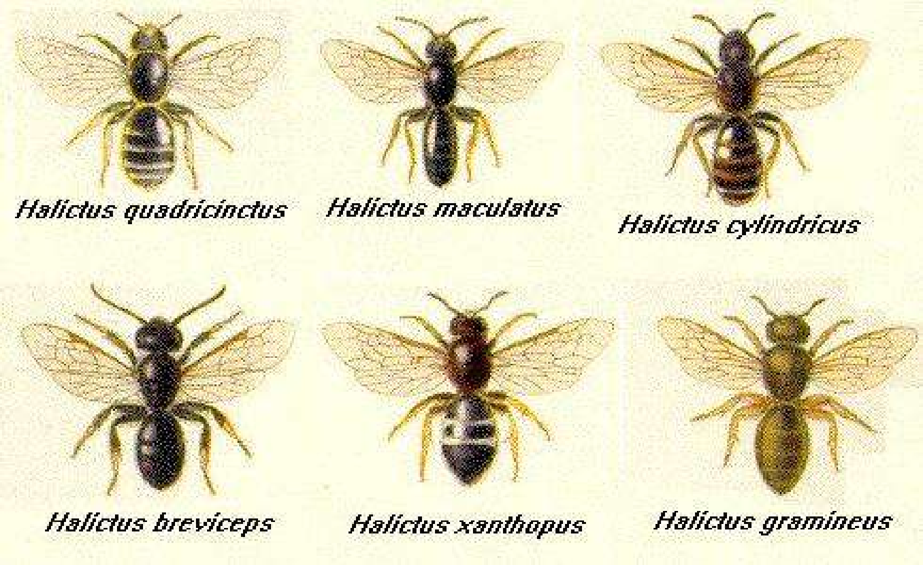
Halictus quadricinctus
Halictus maculatus
Halictus cylindricus
Halictus breviceps
Halictus xanthopus
Halictus gramineus

Terminix® Pest Control $50 Off Terminix Pest Control. Call today! 855-860-9137 www.TerminixPestControlOffers.com

Buy Neodymium Magnets Top Quality Neodymium Magnets, Low Prices, Speedy Same Day Shipping www.amazingmagnets.co

Eco Friendly Pest Control Free Inspection & Estimates. Safe Non-Toxic Extermination. www.PointePestControl.com

AdChoices

If you found this file useful or interesting, please consider a donation or a purchase to help keep KeelyNet online and providing free information.

PayPal DONATE

MexiStim, crock, edkuniversity, quantum, Einstein, Clem, Aharanov, Bohm, virtual, state, free energy, gravity, health, electronic, space, time, dimension, keely, tesla, Decker, electric, gas, chemistry, chemical, aether, zpe, zero point, moray, hendershot, bearden, energy, free energy, lee, perrault, overunity, power, bedini, water, schauberger, baumgartner, sweet, vortex, science, physics, antigravity, sound, acoustics, electricity, motor, vacuum, medicine, oil, petroleum, hydrogen, ecology, biology, Keelynet, levitation, geometry, hubbard, physics, electronics, alternative, paranormal, metal, organic, fiber, reich, weather, rain, media, rife, gray, interdimensional, time travel, new energy, energy, electron, atom, quark, quantum, battery, emf, back emf, lenz, magnet, zpf, puthoff, solar, wind, thermal, thermodynamic, heat, pump, plant, research, invent, invention, inventor, patent, future, past, present, generator, dynamo, faraday, homopolar, disc, kinemassic, unidirectional, direction, fuel, gas, vapor, flying, flight, satellite, rocket, moon, mars, sun, earth, push, einstein, converter, inverter, microscope, signature, radionic, psionic, psychic, telepathic, telepathy, precognition, cognition, reality, galaxy, laser, color, microwave, radio, aura, pyramid, thoth, indian, temple, music, frequency, wave, wavelength, broadcast, superconductor, cooper pair, perpetual, motion, unity, expansion, quartz, vibration, oscillation, sympathetic, cancer, disease, heart, attack, weapon, carbon, alien, clinic, pressure, mechanical, physical, medicine, healing, life, rejuvenation, cell, voltage, current, amperage, watts, coil, cone, biomagnetic, lightning, arc, spark, resonant, resonance, conjugation, phase, ufo, patent, experiment, experimental, virus, bacteria, battery, honeycomb, ftl, rectify, rectification, grebennikov, russia, negative, positive, neutral, weird, air, dilation, chaos, control, parameter, mystery, weightless, invisible, invisibility, veda, vedic, vedas, hollow, vailx, vimana, saucer, angular, momentum, conversion, propulsion, drive, universe, transition, macrosonic, synergetic, tornado, standing, airship, aerial, propeller, thrust, maxwell, field, eternal, turbine, wind, rotating, force, technology, technical, cop, efficiency, transmission, gearbox, experiment, experimenter, circuit, correlation, helical, gasoline, decomposition, fuel cell, hydrogen, oxygen, cold, fusion, inertial, high voltage, repulsive, attractive, theory, dielectric, paramagnetic, diamagnetic, orgone, caduceus, bifilar, cymatic, jenny, manners, delawarr, heironymous, hieronymous, analyzer, molecular, atomic, grid, appliance, wheel, motion, brown, temperature, radiant, monopole, sweet, silvertooth, motional, hooper, flanagan, decker, vacuum, matrix, space-time, heaven, god, heretic, pain, degauss, artificial, intelligence, biology, blaster, colloid, tube, stellar, mutation, bible, purify, purification, holiday, vacation, travel, ticket, flight, airline, car, exotic, novel, unique, dissociation, compression, vector, lost, shift, teleport, ark, covenant, tantalus, tantalusfield, noah, theoretical, anomaly, anomalies, phenomena, rare, earth, neodymium, robotic, conspiracy, electronic, ancient, electrogravitic, gravitic, engine, stirling, thought, emg, eeg, nexus, Tilley

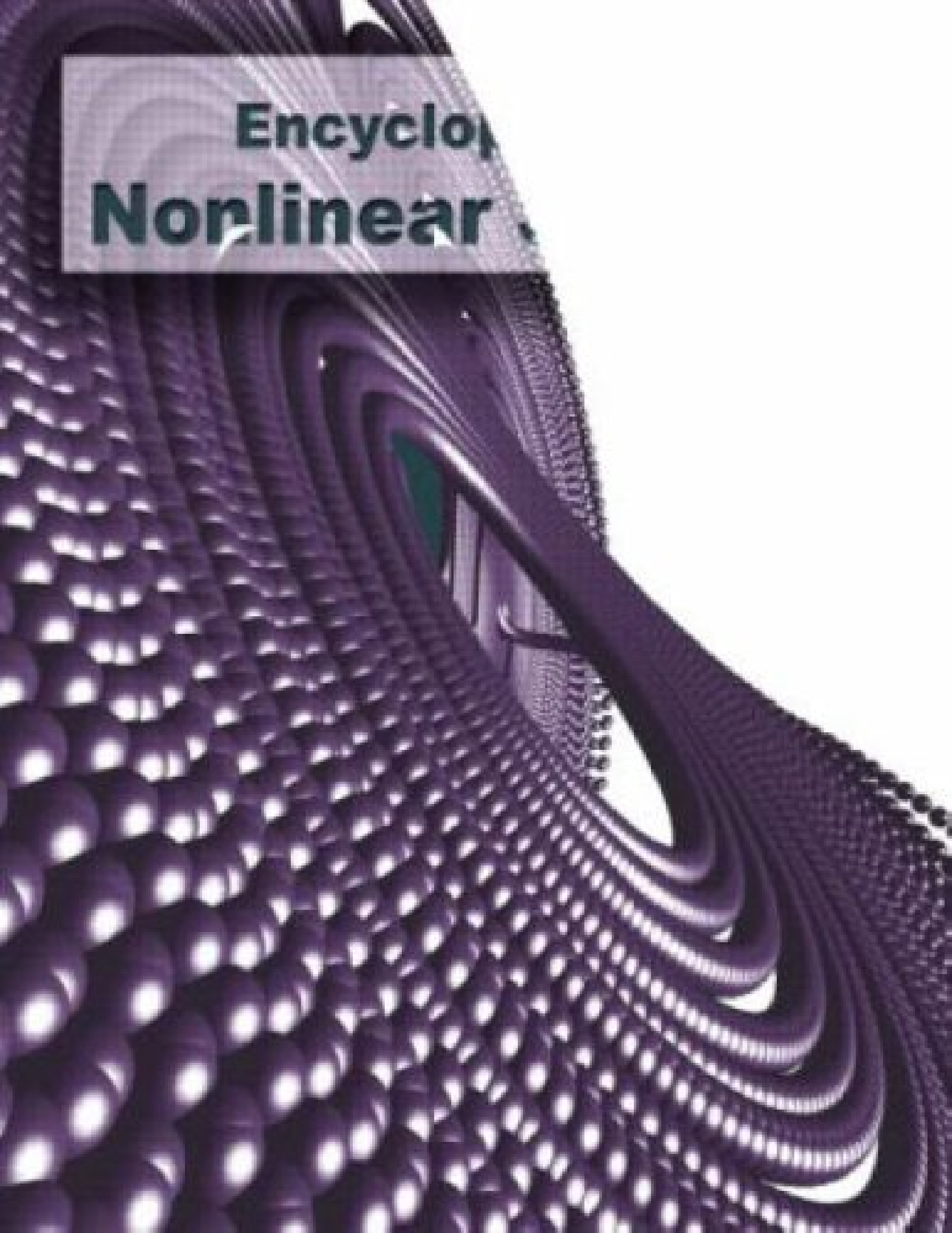

Encyclop

Nonlinear

List of Entries

Ablowitz–Kaup–Newell–Segur system
Adiabatic invariants
Alfvén waves
Algorithmic complexity
Anderson localization
Anosov and Axiom-A systems
Arnol'd diffusion
Artificial intelligence
Artificial life
Atmospheric and ocean sciences
Attractor neural network
Attractors
Aubry–Mather theory
Avalanche breakdown
Avalanches
Averaging methods

Bäcklund transformations
Ball lightning
Belousov–Zhabotinsky reaction
Bernoulli's equation
Berry's phase
Bethe ansatz
Bifurcations
Bilayer lipid membranes
Billiards
Binding energy
Biological evolution
Biomolecular solitons
Bjerrum defects
Black holes
Born–Infeld equations
Bose–Einstein condensation
Boundary layers
Boundary value problems
Branching laws
Breathers
Brownian motion
Brusselator
Burgers equation
Butterfly effect

Candle
Cardiac arrhythmias and the electrocardiogram
Cardiac muscle models
Cat map
Catalytic hypercycle
Catastrophe theory
Causality
Celestial mechanics
Cell assemblies
Cellular automata
Cellular nonlinear networks
Center manifold reduction
Chaos vs. turbulence
Chaotic advection
Chaotic dynamics
Characteristics
Charge density waves
Chemical kinetics
Cherenkov radiation
Chua's circuit
Clear air turbulence
Cluster coagulation
Coherence phenomena
Collective coordinates
Collisions
Color centers
Commensurate-incommensurate transition
Compartmental models
Complex Ginzburg–Landau equation
Conley index
Constants of motion and conservation laws
Continuum approximations
Contour dynamics

Controlling chaos
Cosmological models
Coupled map lattice
Coupled oscillators
Coupled systems of partial differential equations
Critical phenomena

Damped-driven anharmonic oscillator
Darboux transformation
Davydov soliton
Delay-differential equations
Denjoy theory
Derrick–Hobart theorem
Detailed balance
Determinism
Deterministic walks in random environments
Development of singularities
Differential geometry
Diffusion
Dimensional analysis
Dimensions
Diodes
Discrete breathers
Discrete nonlinear Schrödinger equations
Discrete self-trapping system
Dislocations in crystals
Dispersion management
Dispersion relations
Distributed oscillators
DNA premelting
DNA solitons
Domain walls
Doppler shift
Dressing method
Dripping faucet
Drude model
Duffing equation
Dune formation
Dynamical systems
Dynamos, homogeneous

Economic system dynamics
Effective mass
Einstein equations
Electroencephalogram at large scales
Electroencephalogram at mesoscopic scales
Electron beam microwave devices
Elliptic functions
Embedding methods
Emergence
Energy analysis
Entropy
Ephaptic coupling
Epidemiology
Equations, nonlinear
Equilibrium
Equipartition of energy
Ergodic theory
Euler–Lagrange equations
Evans function
Evaporation wave
Excitability
Excitons
Explosions
Extremum principles

Fairy rings of mushrooms
Feedback
Fermi acceleration and Fermi map
Fermi–Pasta–Ulam oscillator chain
Ferromagnetism and ferroelectricity
Fibonacci series
Filamentation
Fitness landscape
FitzHugh–Nagumo equation
Flame front
Flip-flop circuit
Fluctuation-dissipation theorem
Fluid dynamics
Fokker–Planck equation
Forecasting
Forest fires
Fractals
Framed space curves
Franck–Condon factor
Fredholm theorem
Free energy
Free probability theory
Frenkel–Kontorova model
Frequency doubling
Fröhlich theory
Frustration
Function spaces
Functional analysis

Galaxies
Game of life
Game theory
Gel'fand–Levitan theory
General circulation models of the atmosphere
General relativity
Generalized functions
Geometrical optics, nonlinear
Geomorphology and tectonics
Gestalt phenomena
Glacial flow
Global warming
Gradient system
Granular materials
Gravitational waves
Group velocity
Growth patterns

Hamiltonian systems
Harmonic generation
Hartree approximation
Heat conduction
Hele-Shaw cell
Helicity
Hénon map
Hénon–Heiles system
Hierarchies of nonlinear systems
Higgs boson
Hirota's method
Hodgkin–Huxley equations
Hodograph transform
Hole burning
Holons
Hopf bifurcation
Horseshoes and hyperbolicity in dynamical systems
Hurricanes and tornadoes
Huygens principle
Hydrogen bond
Hydrothermal waves
Hysteresis

Idempotent analysis
Incoherent solitons
Inertial manifolds
Information theory
Inhibition
Instantons
Integrability
Integrable cellular automata
Integrable lattices
Integral transforms
Integrate and fire neuron
Intermittency
Invariant manifolds and sets
Inverse problems
Inverse scattering method or transform
Ising model

Josephson junction arrays
Josephson junctions
Jump phenomena
Jupiter's Great Red Spot

Kadomtsev–Petviashvili equation
Kelvin–Helmholtz instability
Kerr effect
Kicked rotor
Kirchhoff's laws
Knot theory
Kolmogorov cascade
Kolmogorov–Arnol'd–Moser theorem
Korteweg–de Vries equation
Kuramoto–Sivashinsky equation

Laboratory models of nonlinear waves
Landau–Lifshitz equation
Langmuir–Blodgett films
Lasers
Lattice gas methods
Lévy flights
Lie algebras and Lie groups
Lifetime
Linearization
Liquid crystals
Local modes in molecular crystals
Local modes in molecules
Long Josephson junctions
Lorentz gas
Lorenz equations
Lyapunov exponents

Magnetohydrodynamics
Manley–Rowe relations
Maps
Maps in the complex plane
Markin–Chizmadzhev model
Markov partitions
Martingales
Matter, nonlinear theory of
Maxwell–Bloch equations
McCulloch–Pitts network
Measures
Mechanics of solids
Mel'nikov method
Mixing
Modulated waves
Molecular dynamics
Monodromy preserving deformations
Monte Carlo methods
Morphogenesis, biological
Multidimensional solitons
Multifractal analysis
Multiple scale analysis
Multiplex neuron
Multiplicative processes
Multisoliton perturbation theory
Myelinated nerves

Navier–Stokes equation
N-body problem
Nephron dynamics
Nerve impulses
Neural network models
Neuristor
Neurons
Newton's laws of motion
Nonequilibrium statistical mechanics
Nonlinear acoustics
Nonlinear electronics
Nonlinear optics
Nonlinear plasma waves

Nonlinear Schrödinger equations
Nonlinear signal processing
Nonlinear toys
Nonlinearity, definition of
Nontwist maps
Normal forms theory
N-soliton formulas
Numerical methods
N-wave interactions

One-dimensional maps
Optical fiber communications
Order from chaos
Order parameters
Ordinary differential equations, nonlinear
Overtones

Painlevé analysis
Parametric amplification
Partial differential equations, nonlinear
Particle accelerators
Particles and antiparticles
Pattern formation
Peierls barrier
Pendulum
Perceptron
Percolation theory
Period doubling
Periodic bursting
Periodic orbit theory
Periodic spectral theory
Perturbation theory
Phase dynamics
Phase plane
Phase space
Phase-space diffusion and correlations
Phase transitions
Photonic crystals
Plasma soliton experiments
Plume dynamics
Poincaré theorems
Poisson brackets
Polaritons
Polarons
Polymerization
Population dynamics
Power balance
Protein dynamics
Protein structure
Pump-probe measurements

Quantum chaos
Quantum field theory
Quantum inverse scattering method
Quantum nonlinearity
Quantum theory

Quasilinear analysis
Quasiperiodicity

Random matrix theory I: Origins and physical applications
Random matrix theory II: Algebraic developments
Random matrix theory III: Combinatorics
Random matrix theory IV: Analytic methods
Random walks
Ratchets
Rayleigh and Raman scattering and IR absorption
Rayleigh–Taylor instability
Reaction-diffusion systems
Recurrence
Regular and chaotic dynamics in atomic physics
Relaxation oscillators
Renormalization groups
Rheology
Riccati equations
Riemann–Hilbert problem
Rössler systems
Rotating rigid bodies
Rotating-wave approximation
Routes to chaos

Salerno equation
Sandpile model
Scheibe aggregates
Scroll waves
Semiconductor laser
Semiconductor oscillators
Separation of variables
Shear flow
Shock waves
Sinai–Ruelle–Bowen measures
Sine-Gordon equation
Singularity theory
Skyrmions
Solar system
Solitons
Solitons, a brief history
Solitons, types of
Spatiotemporal chaos
Spectral analysis
Spin systems
Spiral waves
Stability
Standard map
State diagrams
Stereoscopic vision and binocular rivalry
Stochastic analysis of neural systems
Stochastic processes
String theory
Structural complexity
Superconducting quantum interference device

Superconductivity
Superfluidity
Superlattices
Surface waves
Symbolic dynamics
Symmetry groups
Symmetry: equations vs. solutions
Symplectic maps
Synchronization
Synergetics

Tachyons and superluminal motion
Tacoma Narrows Bridge collapse
Taylor–Couette flow
Tensors
Tessellation
Thermal convection
Thermo-diffusion effects
Theta functions
Threshold phenomena
Time series analysis
Toda lattice
Topological defects
Topology
Traffic flow
Turbulence
Turbulence, ideal
Turing patterns
Twistor theory

Universality

Van der Pol equation
Virial theorem
Visiometrics
Volterra series and operators
Vortex dynamics in excitable media
Vortex dynamics of fluids

Water waves
Wave of translation
Wave packets, linear and nonlinear
Wave propagation in disordered media
Wave stability and instability
Wavelets
Winding numbers

Yang–Mills theory

Zeldovich–Frank-Kamenetsky equation
Zero-dispersion limits

Thematic List of Entries

General

HISTORY OF NONLINEAR SCIENCE

Bernoulli's equation, Butterfly effect, Candle, Celestial mechanics, Davydov soliton, Determinism, Feedback, Fermi–Pasta–Ulam oscillator chain, Fibonacci series, Hodgkin–Huxley equations, Introduction, Integrability, Lorenz equations, Manley–Rowe relations, Markin–Chizmadzhev model, Martingales, Matter, nonlinear theory of, Poincaré theorems, Solar system, Solitons, a brief history, Tacoma Narrows Bridge collapse, Van der Pol equation, Zeldovich–Frank-Kamenetsky equation

COMMON EXAMPLES OF NONLINEAR PHENOMENA

Avalanches, Ball lightning, Brownian motion, Butterfly effect, Candle, Clear air turbulence, Diffusion, Dripping faucet, Dune formation, Explosions, Fairy rings of mushrooms, Filamentation, Flame front, Fluid dynamics, Forest fires, Glacial flow, Global warming, Hurricanes and tornadoes, Jupiter's Great Red Spot, Nonlinear toys, Order from chaos, Pendulum, Phase transitions, Plume dynamics, Solar system, Tacoma Narrows Bridge collapse, Traffic flow, Water waves

Methods and Models

ANALYTICAL METHODS

Bäcklund transformations, Bethe ansatz, Center-manifold reduction, Characteristics, Collective coordinates, Continuum approximations, Dimensional analysis, Dispersion relations, Dressing method, Elliptic functions, Energy analysis, Evans function, Fredholm theorem, Gel'fand–Levitan theory, Generalized functions, Hamiltonian systems, Hirota's method, Hodograph transform, Idempotent analysis, Integral transforms, Inverse scattering method or transform, Kirchhoff's laws, Multiple scale analysis, Multisoliton perturbation theory, Nonequilibrium statistical mechanics, Normal forms theory, N-soliton formulas, Painlevé analysis, Periodic spectral theory, Perturbation theory, Phase dynamics, Phase plane, Poisson brackets, Power balance, Quantum inverse scattering method, Quasilinear analysis, Riccati equations, Rotating-wave approximation, Separation of variables, Spectral analysis, Stability, State diagrams, Synergetics, Tensors, Theta functions, Time series analysis, Volterra series, Wavelets, Zero-dispersion limits

COMPUTATIONAL METHODS

Averaging methods, Cellular automata, Cellular nonlinear networks, Characteristics, Compartmental models, Contour dynamics, Embedding methods, Extremum principles, Fitness landscape, Forecasting, Framed space curves, Hartree approximation, Integrability, Inverse problems, Lattice gas methods, Linearization, Maps, Martingales, Monte–Carlo methods, Numerical methods, Recurrence, Theta functions, Time series analysis, Visiometrics, Volterra series and operators, Wavelets

TOPOLOGICAL METHODS

Bäcklund transformations, Cat map, Conley index, Darboux transformation, Denjoy theory, Derrick–Hobart theorem, Differential geometry, Extremum principles, Functional analysis, Horseshoes and hyperbolicity in dynamical systems, Huygens principle, Inertial manifolds, Invariant manifolds and sets, Knot theory, Kolmogorov–Arnol'd–Moser theorem, Lie algebras and Lie groups, Maps, Measures, Monodromy-preserving deformations, Multifractal analysis, Nontwist maps, One-dimensional maps, Periodic orbit theory, Phase plane, Phase space, Renormalization groups, Riemann–Hilbert problem, Singularity theory, Symbolic dynamics, Symmetry groups, Topology, Virial theorem, Winding numbers

CHAOS, NOISE AND TURBULENCE

Attractors, Aubry–Mather theory, Butterfly effect, Chaos vs. turbulence, Chaotic advection, Chaotic dynamics, Clear air turbulence, Dimensions, Entropy, Ergodic theory, Fluctuation-dissipation theorem, Fokker–Planck equation, Free probability theory, Frustration, Hele-Shaw cell, Horseshoes and hyperbolicity in dynamical systems, Lévy flights, Lyapunov exponents, Martingales, Mel'nikov method, Order from chaos, Percolation theory, Phase space, Quantum chaos, Random matrix theory, Random walks, Routes to chaos, Spatiotemporal chaos, Stochastic processes, Turbulence, Turbulence, ideal

COHERENT STRUCTURES

Biomolecular solitons, Black holes, Breathers, Cell assemblies, Davydov soliton, Discrete breathers, Dislocations in crystals, DNA solitons, Domain walls, Dune formation, Emergence, Fairy rings of mushrooms, Flame front, Higgs boson, Holons, Hurricanes and tornadoes, Instantons, Jupiter's Great Red Spot, Local modes in molecular crystals, Local modes in molecules, Multidimensional solitons, Nerve impulses, Polaritons, Polarons, Shock waves, Skyrmions, Solitons, types of, Spiral waves, Tachyons and superluminal motion, Turbulence, Turing patterns, Wave of translation

DYNAMICAL SYSTEMS

Anosov and axiom-A systems, Arnol'd diffusion, Attractors, Aubry–Mather theory, Bifurcations, Billiards, Butterfly effect, Cat map, Catastrophe theory, Center manifold reduction, Chaotic dynamics, Coupled map lattice, Deterministic walks in random environments, Development of singularities, Dynamical systems, Equilibrium, Ergodic theory, Fitness landscape, Framed space curves, Function spaces, Gradient system, Hamiltonian systems, Hénon map, Hopf bifurcation, Horseshoes and hyperbolicity in dynamical systems, Inertial manifolds, Intermittency, Kicked rotor, Kolmogorov–Arnol'd–Moser theorem, Lyapunov exponents, Maps, Measures, Mel'nikov method, One-dimensional maps, Pattern formation, Periodic orbit theory, Phase plane, Phase space, Phase-space diffusion and correlations, Poincaré theorems, Reaction-diffusion systems, Rössler systems, Rotating rigid bodies, Routes to chaos, Sinai–Ruelle–Bowen measures, Standard map, Stochastic processes, Symbolic dynamics, Synergetics, Universality, Visiometrics, Winding numbers

GENERAL PHENOMENA

Adiabatic invariants, Algorithmic complexity, Anderson localization, Arnol'd diffusion, Attractors, Berry's phase, Bifurcations, Binding energy, Boundary layers, Branching laws, Breathers, Brownian motion, Butterfly effect, Causality, Chaotic dynamics, Characteristics, Cluster coagulation, Coherence phenomena, Collisions, Critical phenomena, Detailed balance, Determinism, Diffusion, Domain walls, Doppler shift, Effective mass, Emergence, Entropy, Equilibrium, Equipartition of energy, Excitability, Explosions, Feedback, Filamentation, Fractals, Free energy, Frequency doubling, Frustration, Gestalt phenomena, Group velocity, Harmonic generation, Helicity, Hopf bifurcation, Huygens' principle, Hysteresis, Incoherent solitons, Inhibition, Integrability, Intermittency, Jump phenomena, Kolmogorov cascade, Lévy flights, Lifetime, Mixing, Modulated waves, Multiplicative processes, Nonlinearity, definition of, N-wave interactions, Order from chaos, Order parameters, Overtones, Pattern formation, Period doubling, Periodic bursting, Power balance, Quantum chaos, Quantum nonlinearity, Quasiperiodicity, Recurrence, Routes to chaos, Scroll waves, Shear flow, Solitons, Spiral waves, Structural complexity, Symmetry: equations vs. solutions, Synergetics, Tachyons and superluminal motion, Tessellation, Thermal convection, Threshold phenomena, Turbulence, Universality, Wave packets, linear and nonlinear, Wave propagation in disordered media, Wave stability and instability

MAPS

Aubry–Mather theory, Bäcklund transformations, Cat map, Coupled map lattice, Darboux transformation, Denjoy theory, Embedding methods, Fermi acceleration and Fermi map, Hénon map, Maps, Maps in the complex plane, Monodromy preserving deformations, Nontwist maps, One-dimensional maps, Periodic orbit theory, Recurrence, Renormalization groups, Singularity theory, Standard map, Symplectic maps

MATHEMATICAL MODELS

Ablowitz–Kaup–Newell–Segur system, Attractor neural network, Billiards, Boundary value problems, Brusselator, Burger's equation, Cat map, Cellular automata, Compartmental models, Complex Ginzburg–Landau equation, Continuum approximations, Coupled map lattice, Coupled systems of partial differential equations, Delay-differential equations, Discrete nonlinear Schrödinger equations, Discrete self-trapping system, Duffing equation, Equations, nonlinear, Euler–Lagrange equations, Fitzhugh–Nagumo equation, Fokker–Planck equation, Frenkel–Kontorova model, Game of life, General circulation models of the atmosphere, Hénon–Heiles system, Integrable cellular automata, Integrable lattices, Ising model, Kadomtsev–Petviashvili equation, Knot theory, Korteweg–de Vries equation, Kuramoto–Sivashinsky equation, Landau–Lifshitz equation, Lattice gas methods, Lie algebras and Lie groups, Lorenz equations, Markov partitions, Martingales, Maxwell–Bloch equation, McCulloch–Pitts network, Navier–Stokes equation, Neural network models, Newton's laws of motion, Nonlinear Schrödinger equations, One-dimensional maps, Ordinary differential equations, nonlinear, Partial differential equations, nonlinear, Random walks, Riccati equations, Salerno equation, Sandpile model, Sine-Gordon equation, Spin systems, Stochastic processes, Structural complexity, Symbolic dynamics, Synergetics, Toda lattice, Van der Pol equation, Zeldovich–Frank-Kamenetsky equation

STABILITY

Attractors, Bifurcations, Butterfly effect, Catastrophe theory, Controlling chaos, Development of singularities, Dispersion management, Dispersion relations, Emergence, Equilibrium, Excitability, Feedback, Growth patterns, Hopf bifurcation, Lyapunov exponents, Nonequilibrium statistical mechanics, Stability

Disciplines

ASTRONOMY AND ASTROPHYSICS

Alfvèn waves, Black holes, Celestial mechanics, Cosmological models, Einstein equations, Galaxies, Gravitational waves, Hénon–Heiles system, Jupiter's Great Red Spot, N-body problem, Solar system

BIOLOGY

Artificial life, Bilayer lipid membranes, Biological evolution, Biomolecular solitons, Cardiac arrhythmias and electro cardiogram, Cardiac muscle models, Catalytic hypercycle, Compartmental models, Davydov soliton, DNA premelting, DNA solitons, Epidemiology, Excitability, Fairy rings of mushrooms, Fibonacci series, Fitness landscape, Fröhlich theory, Game of life, Growth patterns, Morphogenesis, biological, Nephron dynamics, Protein dynamics, Protein structure, Scroll waves, Turing patterns

CHEMISTRY

Belousov–Zhabotinsky reaction, Biomolecular solitons, Brusselator, Candle, Catalytic hypercycle, Chemical kinetics, Cluster coagulation, Flame front, Franck–Condon factor, Hydrogen bond, Langmuir–Blodgett films, Molecular dynamics, Polymerization, Protein structure, Reaction-diffusion systems, Scheibe aggregates, Turing patterns, Vortex dynamics in excitable media

CONDENSED MATTER AND SOLID-STATE PHYSICS

Anderson localization, Avalanche breakdown, Bjerrum defects, Bose–Einstein condensation, Charge density waves, Cherenkov radiation, Color centers, Commensurate-incommensurate transition, Discrete breathers, Dislocations in crystals, Domain walls, Drude model, Effective mass, Excitons, ferromagnetism and Ferroelectricity, Franck–Condon factor, Frenkel–Kontorova model, Frustration, Heat conduction, Hydrogen bond, Ising model, Langmuir–Blodgett films, Liquid crystals, Local modes in molecular crystals, Mechanics of solids, Nonlinear acoustics, Peierls barrier, Percolation theory, Regular and chaotic dynamics in atomic physics, Scheibe aggregates, Semiconductor oscillators, Spin systems, Superconductivity, Superfluidity, Surface waves

EARTH SCIENCE

Alfvén waves, Atmospheric and ocean sciences, Avalanches, Ball lightning, Butterfly effect, Clear air turbulence, Dune formation, Dynamos, homogeneous, Fairy rings of mushrooms, Forest fires, General circulation models of the atmosphere, Geomorphology and tectonics, Glacial flow, Global warming, Hurricanes and tornadoes, Kelvin–Helmholtz instability, Sandpile model, Water waves

ENGINEERING

Artificial intelligence, Cellular automata, Cellular nonlinear networks, Chaotic advection, Chua's circuit, Controlling chaos, Coupled oscillators, Diodes, Dispersion management, Dynamos, homogeneous, Electron beam microwave devices, Explosions, Feedback, Flip-flop circuit, Frequency doubling, Hele-Shaw cell, Hysteresis, Information theory, Josephson junction arrays, Josephson junctions, Langmuir–Blodgett films, Lasers, Long Josephson junctions, Manley–Rowe relations, Neuristor, Nonlinear electronics, Nonlinear optics, Nonlinear signal processing, Optical fiber communications, Parametric amplification, Particle accelerators, Ratchets, Relaxation oscillators, Semiconductor laser, Semiconductor oscillators, Superconducting quantum interference device, Synchronization, Tacoma Narrows Bridge collapse

FLUIDS

Alfvén waves, Atmospheric and ocean sciences, Bernoulli's equation, Chaos vs. turbulence, Chaotic advection, Clear air turbulence, Contour dynamics, Electron beam microwave devices, Evaporation wave, Fluid dynamics, Forecasting, General circulation models of the atmosphere, Glacial flow, Hele-Shaw cell, Hurricanes and tornadoes, Hydrothermal waves, Jump phenomena, Jupiter's Great Red Spot, Kelvin–Helmholtz instability, Laboratory models of nonlinear waves, Lattice gas methods, Liquid crystals, Lorentz gas, Magnetohydrodynamics, Navier–Stokes equation, Nonlinear plasma waves, Plasma soliton experiments, Plume dynamics, Rayleigh–Taylor instability, Shear flow, Shock waves, Superfluidity, Surface waves, Taylor–Couette flow, Thermal convection, Thermo-diffusion effects, Traffic flow, Turbulence, Turbulence, ideal, Visiometrics, Vortex dynamics of fluids, Water waves

NEUROSCIENCE

Artificial intelligence, Attractor neural network, Cell assemblies, Compartmental models, Electroencephalogram at large scales, Electroencephalogram at mesoscopic scales, Ephaptic coupling, Evans function, FitzHugh–Nagumo equation, Gestalt phenomena, Hodgkin–Huxley equations, Inhibition, Integrate and fire neuron, Multiplex neuron, Myelinated nerves, Nerve impulses, Neural network models, Neurons, Pattern formation, Perceptron, Stereoscopic vision and binocular rivalry, Stochastic analyses of neural systems, Synergetics

NONLINEAR OPTICS

Cherenkov radiation, Color centers, Damped-driven anharmonic oscillator, Dispersion management, Distributed oscillators, Excitons, Filamentation, Geometrical optics, nonlinear, Harmonic generation, Hole burning, Kerr effect, Lasers, Liquid crystals, Maxwell–Bloch equations, Nonlinear optics, Optical fiber communications, Photonic crystals, Polaritons, Polarons, Pump-probe measurements, Rayleigh and Raman scattering and IR absorption, Semiconductor laser, Tachyons and superluminal motion

PLASMA PHYSICS

Alfvén waves, Ball lightning, Charge density waves, Drude model, Dynamos, homogeneous, Electron beam microwave devices, Magnetohydrodynamics, Nonlinear plasma waves, Particle accelerators, Plasma soliton experiments

SOCIAL SCIENCE

Economic system dynamics, Epidemiology, Game theory, Hierarchies of nonlinear systems, Population dynamics, Synergetics, Traffic flow

SOLID MECHANICS AND NONLINEAR VIBRATIONS

Avalanche breakdown, Bilayer lipid membranes, Bjerrum defects, Charge density waves, Cluster coagulation, Color centers, Detailed balance, Dislocations in crystals, Domain walls, Frustration, Glacial flow, Granular materials, Growth patterns, Heat conduction, Hydrogen bond, Ising model, Kerr effect, Langmuir–Blodgett films, Liquid crystals, Local modes in molecular crystals, Mechanics of solids, Molecular dynamics, Nonlinear acoustics, Protein dynamics, Ratchets, Rheology, Sandpile model, Scheibe aggregates, Shock waves, Spin systems, Superlattices, Surface waves, Tessellation, Topological defects

THEORETICAL PHYSICS

Berry's phase, Black holes, Born–Infeld equations, Celestial mechanics, Cherenkov radiation, Cluster coagulation, Constants of motion and conservation laws, Cosmological models, Critical phenomena, Derrick–Hobart theorem, Detailed balance, Einstein equations, Entropy, Equipartition of energy, Fluctuation-dissipation theorem, Fokker–Planck equation, Free energy, Galaxies, General relativity, Gravitational waves, Hamiltonian systems, Higgs boson, Holons, Instantons, Matter, nonlinear theory of, N-body problem, Newton's laws of motion, Particles and antiparticles, Quantum field theory, Quantum theory, Regular and chaotic dynamics in atomic physics, Rotating rigid bodies, Skyrmions, String theory, Tachyons and superluminal motion, Twistor theory, Virial theorem, Yang–Mills theory

A

AB INITIO CALCULATIONS

See **Molecular dynamics**

ABLOWITZ–KAUP–NEWELL–SEGUR SYSTEM

In 1967, Gardner, Greene, Kruskal, and Miura (or GGKM) (Gardner et al., 1967) showed that the Kortegweg–de Vries (KdV) equation

$$q_t + 6qq_x + q_{xxx} = 0, \tag{1}$$

with rapidly decaying initial data on $-\infty < x < \infty$, can be linearized using direct and inverse scattering methods associated with the linear Schrödinger equation

$$v_{xx} + [k^2 + q(x,t)]v = 0. \tag{2}$$

The KdV equation is of practical interest, having been first derived in the study of long water waves (Korteweg & de Vries, 1895) and subsequently in several other areas of applied science.

In the method proposed by Gardner et al., the solitary wave (soliton) solution to the KdV equation (1)

$$q = 2\kappa^2 \operatorname{sech}^2 \kappa(x - 4\kappa^2 t - x_0)$$

and multisoliton solutions are associated with the discrete spectrum of Equation (2). The discrete eigenvalues were shown to be invariants of the KdV motion; for example, the above soliton solution is associated with the discrete eigenvalue of Equation (2) at $k = \mathrm{i}\kappa$.

At that time, it was not clear whether the method could be applied to other physically significant equations. In 1972, however, Zakharov and Shabat (1972) used an operator formalism developed by Lax (1968) to show that the nonlinear Schrödinger (NLS) equation

$$\mathrm{i}q_t + q_{xx} + \sigma|q|^2 q = 0, \tag{3}$$

with rapidly decaying initial data on $-\infty < x < \infty$, could also be linearized by direct and inverse scattering methods.

The NLS equation was known to arise in many physical contexts (Benney & Newell, 1967) and in 1973 Hasegawa and Tappert showed that the NLS equation describes the long-distance dynamics of nonlinear pulses in optical fibers (Hasegawa & Tappert, 1973). Motivated by these developments and indications that other equations fit into this category, David Kaup, Alan Newell, Harvey Segur, and the present author (Ablowitz et al., 1973, 1974) studied the following modification of the Zakharov–Shabat system:

$$\begin{aligned} v_{1x} &= -\mathrm{i}\zeta v_1 + q v_2, \\ v_{2x} &= \mathrm{i}\zeta v_2 + r v_1, \end{aligned} \tag{4}$$

$$\begin{aligned} v_{1t} &= A v_1 + B v_2, \\ v_{2t} &= C v_1 + D v_2. \end{aligned} \tag{5}$$

In Equations (4) and (5), v_1 and v_2 are auxiliary functions obeying the postulated linear systems; Equation (4) play the same role as Equation (2), whereas Equation (5) determine the temporal evolution of the functions v_1 and v_2. (The evolution equation associated with the auxiliary function v for the KdV equation was not given above.) The method establishes that the functions $q = q(x,t)$ and $r = r(x,t)$ satisfy nonlinear equations when the (yet to be determined) functions A, B, C, and D are properly chosen.

The key to this approach is to make Equations (4) and (5) compatible, that is, set the x-derivative of v_{it} equal to the t-derivative of v_{ix}. In other words, we set the x-derivative of the right-hand side of Equations (5) equal to the t-derivative of the right-hand side of Equations (4). The result of this calculation yields the following equations for A, B, C, and D:

$$\begin{aligned} A_x &= qC - rB, \\ B_x + 2\mathrm{i}\zeta B &= q_t - 2Aq, \\ C_x - 2\mathrm{i}\zeta C &= r_t + 2Ar, \\ D &= -A. \end{aligned} \tag{6}$$

In Ablowitz et al. (1973, 1974; see also Ablowitz & Segur, 1981), methods to solve these equations are described. The simplest procedure is to look for finite power series expansions such as $A = \sum_{i=0}^{N} \zeta^i A_i$ and similarly for B and C.

For example, with $N = 2$, we find with $r = \mp q^*$ that the nonlinear Schrödinger equation (3) with $\sigma = \pm 1$ is a necessary condition. In this case there are 11 equations for the nine unknowns $\{A_i, B_i, C_i\}$, $i = 0, 1, 2$, and the remaining two equations determine the nonlinear evolution equations for q and r (in this case NLS when $q = \mp r^*$).

With $N = 3$ and $r = -1$, we find that q must satisfy the KdV equation. Also, with $r = \mp q$, the modifed KdV equation

$$q_t \pm 6q^2 q_x + q_{xxx} = 0 \tag{7}$$

results.

If we look for expansions containing inverse powers of ζ, additional interesting equations can be obtained. For example, postulating $A = a/\zeta$, $B = b/\zeta$, $C = c/\zeta$ results in the sine-Gordon and sinh-Gordon equations

$$u_{xt} = \sin u, \tag{8}$$

$$u_{xt} = \sinh u, \tag{9}$$

where $q = -r = -u_x/2$ in Equation (8) and $q = r = u_x/2$ in Equation (9). The sine-Gordon equation has been known to be an important equation in the study of differential geometry since the 19th century (cf. Bianchi, 1902), and it has found applications in the 20th century as models for dislocation propagation in crystals, domain walls in ferromagnetic and ferroelectric materials, short-pulse propagation in resonant optical media, and magnetic flux propagation in long Josephson junctions, among others.

Thus, a number of physically interesting nonlinear wave equations are obtained from the above formalism. In Ablowitz et al. (1973, 1974; see also Ablowitz & Segur, 1981), it was further shown as to how this approach could be generalized to a class of nonlinear equations described in terms of certain nonlinear evolution operators that were subsequently referred to in the literature as recursion operators.

Further, the whole class of nonlinear equations with rapidly decaying initial data on $-\infty < x < \infty$ was shown to be linearized via direct and inverse scattering methods. Special soliton solutions are associated with the discrete spectrum of the linear operator (4), and via (5) the discrete eigenvalues were shown to be invariants of the motion. In subsequent years, asymptotic analysis of the integral equations yielded the long-time behavior of the continuous spectrum, which in turn showed the ubiquitous role that the Painlevé equations play in integrable systems (cf. Ablowitz & Segur, 1981).

Because this formulation is analogous to the method of Fourier transforms, the method was termed the inverse scattering transform or simply the IST.

Mark J. Ablowitz

See also **Integrability; Inverse scattering method or transform; Korteweg–de Vries equation; Nonlinear Schrödinger equations; Solitons**

Further Reading

Ablowitz, M.J., Kaup, D.J., Newell, A.C. & Segur, H. 1973. Nonlinear equations of physical significance. *Physical Review Letters*, 31: 125–127

Ablowitz, M.J., Kaup, D.J., Newell, A.C. & Segur, H. 1974. The inverse scattering transform–Fourier analysis for nonlinear problems. *Studies in Applied Mathematics*, 53: 249–315

Ablowitz, M.J. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia, PA: Society for Industrial and Applied Mathematics

Benney, D.J. & Newell, A.C. 1967. The propagation of nonlinear envelopes. *Journal of Mathematics and Physics* (Name changed to: *Studies in Applied Mathematics*), 46: 133–139

Bianchi, L. 1902. *Lezioni de Geometria Differenziale*, 3 vols, Pisa: Spoerri

Gardner, C.S, Greene, J.M., Kruskal, M.D. & Miura, R.M. 1967. Method for solving the Korteweg–deVries equation. *Physical Review Letters*, 19: 1095–1097

Hasegawa, A. & Tappert, F. 1973. Transmission of stationary nonlinear optical pulses in dispersive dielectrical fibers. I. Anamolous dispersion. *Applied Physics Letters*, 23: 142–144

Korteweg, D.J. & de Vries, F. 1895. On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philosophical Magazine*, 39: 422–443

Lax, P.D. 1968. Integrals of nonlinear equations of evolution and solitary waves. *Communications in Pure and Applied Mathematics*, 21: 467–490

Zakharov, V.E. & Shabat A.B. 1972. Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Soviet Physics, JETP*, 34: 62–69

ABLOWITZ–LADIK EQUATION

See **Discrete nonlinear Schrödinger equations**

ACOUSTIC SOLITONS

See **Nonlinear acoustics**

ACTION POTENTIAL

See **Nerve impulses**

ACTION-ANGLE VARIABLES

See **Hamiltonian systems**

ACTIVATOR-INHIBITOR SYSTEM

See **Reaction-diffusion systems**

ADIABATIC APPROXIMATION

See **Davydov soliton**

ADIABATIC INVARIANTS

Adiabatic invariants, denoted by I, are approximate constants of motion of a given dynamical system (not necessarily Hamiltonian), which are approximately preserved during a process of slow change of the system's parameters (denoted by λ). This change is on a time scale T, which is supposed to be much larger than any typical dynamical time scale such as traversal time or the period of the shortest periodic orbits.

This is an asymptotic statement, in the sense that the adiabatic invariants are better preserved, the slower the driving of the system. In other words, the switching function $\lambda = \lambda(t)$ varies more slowly on the typical evolutionary time scale T, and the preservation is perfect in the limit $T \to \infty$.

The important point is that while the system's parameters $\lambda(t)$ and their dynamical quantities such as the total energy and angular momentum can change by arbitrarily large amounts, their combination involved in the adiabatic invariant I is preserved to a very high degree of accuracy, and this allows us to calculate changes of important quantities in dynamical systems. Examples arise in celestial mechanics, in other Hamiltonian systems, and in the motion of charged particles in magnetic and electric fields.

The accuracy of preservation can be calculated in systems with one degree of freedom and is exponentially good with T if the switching function $\lambda(t)$ is analytic (of class $\mathcal{C}^\infty$); that is to say, the change of the adiabatic invariant ΔI is of the form

$$\Delta I = \alpha \exp(-\beta T), \tag{1}$$

where α and β are known constants. If, however, the switching function $\lambda(t)$ is only of class $\mathcal{C}^m$ (m-times continuously differentiable), then the change of the adiabatic invariant ΔI during an adiabatic change over a time period of length T is algebraic only, namely

$$\Delta I = \alpha T^{-(m+1)}. \tag{2}$$

In both cases, $\Delta I \to 0$ as $T \to \infty$.

The fact that the evolutionary time scale T is large compared to the typical shortest dynamical time scales (average return time, etc.) suggests the averaging method or the so-called averaging principle. Here the long-term evolution (adiabatic evolution) of the system can be calculated by replacing the actual dynamical system with its averaged correspondent, obtained by averaging over the shortest dynamical time scales (the fast variables). Such a procedure is well known, for example, in celestial mechanics where the secular effects of the third-body perturbations of a planet are obtained by averaging the perturbations over one revolutionary period of the perturbers. This was done by Carl Friedrich Gauss in 1801 in the context of studying the dynamics of planets.

The adiabatic invariants can be easily calculated in one-dimensional systems and in completely integrable systems with N degrees of freedom. Something is known about the ergodic Hamiltonian systems, while little is known about adiabatic invariants in mixed-type Hamiltonian systems (with divided phase space), where for some initial conditions in the classical phase space, we have regular motion on invariant tori and irregular (chaotic) motion for other (complementary) initial conditions.

One elementary example is the simple (mathematical) pendulum, of point mass m and of length l with the declination angle φ, described by the Hamiltonian

$$H = \frac{p_\varphi^2}{2ml^2} - mgl\,\cos\varphi, \tag{3}$$

where $p_\varphi = ml^2\dot{\varphi}$ is the angular momentum. For small oscillations $\varphi \ll 1$, around the stable equilibrium $\varphi = 0$. It is described by the harmonic Hamiltonian

$$H' = \frac{p_\varphi^2}{2ml^2} + \frac{mgl}{2}\varphi^2. \tag{4}$$

Here the angular oscillation frequency is $\omega = 2\pi\nu = \sqrt{g/l}$, where ν is the frequency and g is the gravitational acceleration. We denote the total energy of the Hamiltonian H' by E. Paul Ehrenfest discovered that the quantity $I = E/\omega$ is the adiabatic invariant of the system, so the change of $E(t)$ on large time scales $T \gg 1/\nu$ is such that $I = E(t)/\omega(t)$ remains constant. Therefore, if for example the length of the pendulum $l = l(t)$ is slowly, adiabatically changing, then the energy of the system will change according to the law

$$E = E_0\sqrt{\frac{l_0}{l}}, \tag{5}$$

where E_0 and l_0 are the initial values and E and l the final values of the two variables. One can easily show that the oscillation amplitude φ_0 changes as $l^{-3/4}$ as the length l changes. This is an elementary example of a dynamically driven system in which the change of energy E can be very large, as is the change of ω, but $I = E/\omega$ is a well-preserved adiabatic invariant; in fact, it is exponentially well preserved if the switching function $\lambda(t)$ is analytic.

More generally, for Hamiltonian systems $H(q, p, \lambda)$ with one degree of freedom, whose state is described by the coordinate q and canonically conjugate momentum p in the phase space (q, p), and $\lambda = \lambda(t)$ is the system's parameter (slowly changing on time scale T), one can show that the action integral

$$I(E, \lambda) = I(E(t), \lambda(t)) = \frac{1}{2\pi}\oint p\,\mathrm{d}q \tag{6}$$

is the adiabatic invariant of the system, where the contour integral is taken at a fixed total energy E and a fixed value of λ. In this case, $2\pi I$ is interpreted as the area inside the curve $E = \text{const.}$ in the phase plane (q, p). The accuracy is exponentially good if $\lambda(t)$ is an analytic function and algebraic if it is of class C^m. Moreover, the theorem holds true only if the frequency ω is nonzero.

This implies that a passage through a separatrix (in the phase space of a one-dimensional system) is excluded because $\omega = 0$; thus a different approach is necessary with a highly nontrivial result. When crossing a separatrix of a one-dimensional double potential well from outside in an adiabatic way going inside, a bifurcation takes place, and the capture of the trajectory in either of the two wells is possible with some probabilities. These probabilities can be calculated quite easily, and the spread of the adiabatic invariant ΔI after such a passage can also be calculated, but this is more difficult. Important applications are found in celestial mechanics, where an adiabatic capture of a small body near a resonance with a planet can take place; in plasma physics; and in quantum mechanics of states close to the separatrix (in the semiclassical limit).

This is an interesting result, because I is precisely that quantity which according to the "old quantum mechanics" of Bohr and Sommerfeld has to be quantized, that is, made equal to an integer multiple of Planck's constant $\hbar$. Of course, the old quantum mechanics is generally wrong, but it can be a good approximation to the solution of the Schrödinger equation. Even then, strictly speaking, the quantization condition in the sense of EBK or Maslov quantization, must be written in the form

$$I = \frac{1}{2\pi} \oint p \, \mathrm{d}q = \left(n + \frac{\alpha}{4}\right) \hbar, \qquad (7)$$

where $n = 0, 1, 2, \ldots$ is the quantum number and α is the Maslov index, that is, the number of caustics (projection singularities) round the cycle $E = \text{const.}$ in the phase plane. For smooth systems with quadratic kinetic energy, it is typically $\alpha = 2$. Thus, at this semiclassical level, we have the semiclassical adiabatic invariant, stating that in one-dimensional systems under an adiabatic change, the quantum number (and thus the eigenstate) is preserved. This agrees with the exact result in the theory of the Schrödinger equation in quantum mechanics. Round a closed loop in a parameter space, a quantum system returns to its original state, except for the phase. (This closed-loop phase change is essentially the so-called Berry's phase.)

The method of averaging can also be used in N-dimensional Hamiltonians $H = H(\boldsymbol{q}, \boldsymbol{p})$, where $\boldsymbol{q}$ and $\boldsymbol{p}$ are N-dimensional vectors, but it works only in two extreme cases: the integrable case and the ergodic case.

In a classical integrable Hamiltonian system we have N analytic, global, and functionally independent constants of motion $A_i = A_i(\boldsymbol{q}, \boldsymbol{p}),\ i = 1, 2, \ldots, N$, pairwise in involution; that is, all Poisson brackets $\{A_i, A_j\}$ vanish identically everywhere in phase space. The orbits in phase space are then confined to an invariant N-dimensional surface, and according to the Liouville–Arnol'd theorem the topology of these surfaces must be the topology of an N-dimensional torus. Then an action integral $I = (1/2\pi) \oint \boldsymbol{p} \cdot \mathrm{d}\boldsymbol{q}$ along a closed loop on a torus will be zero if the loop can be continuously shrunk to a point on the torus. But there are loops that cannot be shrunk to a point due to the topology of the torus. Then the integral I is different from zero, otherwise its value does not depend on the particular loop, so in a sense it is a topological invariant of the torus. On an N-dimensional torus, there are N such independent elementary closed loops C_i, $i = 1, 2, \ldots, N$. The integrals that we call simply actions or action variables

$$I_i = \frac{1}{2\pi} \oint_{C_i} \boldsymbol{p} \cdot \mathrm{d}\boldsymbol{q} \qquad (8)$$

are then the most natural momentum variables on the torus, whilst angle variables $\boldsymbol{\Theta}$ specifying the position on the torus labeled by $\boldsymbol{I}$ can be generated from the transformation

$$\boldsymbol{\Theta} = \frac{\partial S(\boldsymbol{I}, \boldsymbol{q})}{\partial \boldsymbol{I}}, \qquad (9)$$

where $S = \int \mathbf{p} \cdot \mathrm{d}\boldsymbol{q}$ is an action integral on the torus.

Applying the averaging principle (the method of averaging), one readily shows that for an integrable system the actions $\boldsymbol{I}$ are N adiabatic invariants, provided the system is nondegenerate, which means that the frequencies

$$\boldsymbol{\omega} = \frac{\partial H}{\partial \boldsymbol{I}} \qquad (10)$$

on the given torus are not rationally connected; that is to say, there is no integer vector $\boldsymbol{k}$ such that $\boldsymbol{\omega} \cdot \boldsymbol{k} = 0$. The problem is that during an adiabatic process the frequencies $\boldsymbol{\omega}$ will change, and therefore, strictly speaking, there will be infinitely many points of $\lambda = \lambda(t)$, where $\boldsymbol{\omega} \cdot \boldsymbol{k} = 0$, which will, strictly speaking, invalidate the theorem. However, it is thought that if the degree of resonances or rationality conditions $\boldsymbol{\omega} \cdot \boldsymbol{k} = 0$ is of a very high order, meaning that all components of $\boldsymbol{k}$ are very large, then the adiabatic invariants I_i will be quite well preserved. But low-order resonances (rationality conditions) must be excluded. The details of such a process call for further investigation.

When the N actions I_i of an integrable system are quantized in the sense of Maslov, as explained above in the one-dimensional case, we again find agreement, at this semiclassical level, with quantum mechanics: in a family of integrable systems, all N quantum numbers and the corresponding eigenstates are preserved under an adiabatic change.

Another extreme of classically ergodic and thus fully chaotic systems has been considered already by Hertz. He found that in such ergodic Hamiltonian systems the phase space volume enclosed by the energy surface $H(\boldsymbol{q}, \boldsymbol{p}) = E =$ constant is the adiabatic invariant, denoted by

$$\Omega(E) = \int_{H(\boldsymbol{q},\boldsymbol{p})\le E} \mathrm{d}^N\boldsymbol{q}\,\mathrm{d}^N\boldsymbol{p}. \qquad (11)$$

Of course, here it is required that while the system's parameter $\lambda(t)$ is slowly changing, the system itself must be ergodic for all $\lambda(t)$. Sometimes, this condition is difficult to satisfy, but sometimes it is easily fulfilled. Examples are the stadium of Bunimovich with varying length of the straight line between the two semicircles, or the Sinai billiard with varying radius of the circle inside a square. For an ergodic two-dimensional billiard of area $\mathcal{A}$ and point mass m, we have

$$\Omega(E) = 2\pi m\mathcal{A}E. \qquad (12)$$

Therefore, when $\mathcal{A}$ is adiabatically changing, the energy E of the billiard particle is changing reciprocally with $\mathcal{A}$. Diminishing $\mathcal{A}$ implies increasing E, and this can be interpreted as work being done against the "pressure" of only one particle, if we define the pressure as the time average of the momentum transfer at collisions with the boundary of our ergodic billiard. There is a formalism to proceed with this analysis close to the thermodynamic formalism, as derived from statistical mechanics, except that here we are talking about time averages rather than phase averages of classical variables.

Again, this general result for ergodic systems is interesting from the quantum point of view because $\mathcal{N} = \Omega(E)/(2\pi\hbar)^N$ is precisely the number of energy levels below the energy E in the semiclassical limit of very large $\mathcal{N}$, which is known as the Thomas–Fermi rule. It is the number of elementary quantum Planck cells inside the volume element $H(\boldsymbol{q}, \boldsymbol{p}) \le E$. Indeed, quantum mechanically, the eigenstate and the (energy counting sequential) main quantum number $\mathcal{N}$ are preserved under an adiabatic change.

In case of a mixed-type Hamiltonian system, which is a typical case in nature, adiabatic theory is in its infancy. Moreover, in three or higher degrees of freedom, we have universal diffusion on the Arnol'd web, which is dense on the energy surface, even for KAM-type Hamiltonian systems that are very close to integrability, like our solar planetary system. On the Arnol'd web we have diffusional chaotic motion, and there is a rigorous theory by Nekhoroshev giving a rigorous upper bound to the diffusion rate in such a case. However, when compared with numerical calculations, it is found that the diffusion rate is many orders of magnitude smaller than the Nekhoroshev limit. In other words, the actual diffusion time is much longer than estimated by Nekhoroshev, implying that there we have some approximate adiabatic invariant for long times, but not very long times.

Marko Robnik

See also **Averaging methods; Berry's phase; Billiards; Phase space; Quantum theory; Quasilinear analysis**

Further Reading

Arnol'd, V.I. 1989. *Mathematical Methods of Classical Mechanics*, 2nd edition, New York and Heidelberg: Springer

Cary, J.R. & Rusu, P. 1992. Separatrix eigenfunctions. *Physical Review* A, 45: 8501–8512

Cary, J.R. & Rusu, P. 1993. Quantum dynamics near a classical separatrix. *Physical Review A*, 47: 2496–2505

Landau, L.D. & Lifshitz, E.M. 1996. *Mechanics—Course of Theoretical Physics*, vol. 1, 3rd edition, Oxford: Butterworth-Heinemann

Landau, L.D. & Lifshitz, E.M. 1997. *Quantum Mechanics: Non-Relativistic Theory—Course of Theoretical Physics*, vol. 3, 3rd edition, Oxford: Butterworth-Heinemann

Lichtenberg, A.J. & Lieberman, M.A. 1992. *Regular and Chaotic Dynamics*, 2nd edition, New York and Heidelberg: Springer

Lochak, P. & Meunier, C. 1988. *Multiphase Averaging for Classical Systems*, New York and Heidelberg: Springer

Reinhardt, W.P. 1994. Regular and irregular correspondences. *Progress of Theoretical Physics Supplement*, 116: 179–205

ALFVÉN WAVES

The essence of Hannes Alfvén's contributions to cosmic and laboratory plasmas is his idea of combining electromagnetics and hydrodynamics (Alfvén, 1942), thus introducing the new concept of magnetohydrodynamics (MHD). Electromagnetic waves associated with the motion of conducting liquids in magnetic fields, now known as Alfvén waves, were first observed experimentally (Lundquist, 1949; Lehnert, 1954). Later on, waves of this nature have turned out to be fundamental constituents of numerous phenomena in all parts of the universe (Fälthammar, 1995; Wilhelmsson, 2000). In the pioneering experiments, liquid mercury was used by Lundquist, and liquid sodium by Lehnert, who achieved higher electrical conductivity and lower density, leading to a higher Lundquist number (lower damping). Alfvén used his early results to give a possible explanation for sunspots and the solar cycle (periodicity in the Sun's activity) (Alfvén, 1942). Alfvén noticed that the Sun has a general magnetic field and that solar matter is a good conductor, thus fulfilling idealized requirements for the notion of an electromagnetic wave in a gaseous conductor or plasma.

At a very early age, Alfvén was given a copy of a popular astronomy book by Camille Flammarion, which greatly stimulated his lifelong interest in astronomy and astrophysics. His early experiences building radio receivers at the school radio club were also important for his later activities. Interestingly, another great scientist, Albert Einstein, received a small compass as a present when he was five years old, which entirely absorbed his interest. He asked everybody around him what a magnetic field was and what gravity was, and later on in his life he admitted that this early experience might have influenced his lifelong scientific activities. Other similarities between the two scientists were that in their professional work both Einstein and Alfvén were very creative individualists, striving for simplicity of their solutions, and being skilled in many areas, they often looked at problems with fresh eyes. Both received Nobel Prizes in physics: Einstein in 1922, Alfvén in 1970.

The simplest form of an Alfvén wave, a propagation of an electromagnetic wave in a highly conducting plasma, was first rejected by critics on the grounds that it could not be correct, otherwise it would already have been discovered by Maxwell. Furthermore, experiments had been performed with magnetic fields and conductive media by Ampère and others long ago. Nevertheless, "The Alfvén wave, in fact, is the very foundation on which the entire structure of magnetohydrodynamics (MHD) is erected. Beginning from a majestic original simplicity, it has acquired a rich and variegated character, and has ended up dictating most of the low-frequency dynamics of magnetized plasmas" (Mahajan, 1995).

To visualize the interaction between the magnetic field and the motion of the conductive fluid, one may use an analogy with the theory of stretched strings to obtain a wave along the magnetic lines of force with a velocity v_A, where

$$v_A{}^2 = \frac{B^2}{\mu_0 \rho} \tag{1}$$

and ρ is the mass density of the fluid, μ_0 is the permittivity, and B is the magnetic field.

The variations in velocity and current are mutually perpendicular and the magnetic field variations are in the direction of the fluid velocity variations, all variations being perpendicular to the direction of propagation. One may say that the variations of the magnetic field lines are frozen to those of the fluid motion, as can be deduced from electromagnetic equations, together with the hydrodynamic equation for the case of an incompressible fluid of infinite conductivity. The Alfvén wave is a low-frequency wave ($\omega < \omega_{\text{ci}}$, ω_{ci} being the ion cyclotron frequency) for which the displacement current is negligible. In fact, there are two types of Alfvén waves, for which

$$\omega / k_{\|} = v_A \text{ (torsional or shearwave)} \tag{2}$$

and

$$\omega / k = v_A \text{ (compressional wave)} \tag{3}$$

where ω is the frequency, $k^2 = k_{\|}^2 + k_{\perp}^2$, with $k_{\|}$ and $k_{\perp}$ being the wave numbers along and perpendicular to the magnetic field. For the shear wave, the frequency depends only on $k_{\|}$ and not on $k_{\perp}$, which has profound consequences and leads to a continuous spectrum (Mahajan, 1995).

For determining plasma stability and in selecting schemes for plasma heating and current drive in fusion plasma devices, the understanding of Alfvén wave dynamics is of great importance and has led to a vast literature. Nonlinear effects are of relevance to large-amplitude disturbances frequently observed in laboratory and space plasmas (Wilhelmsson, 1976). The formation and propagation of Alfvén vortices with geocosmophysical and pulsar (electron-positron plasma) applications are just two examples. Alfvén waves have also found interesting applications in solid-state plasmas in semiconductors as well as in metals and semimetals. Such studies have resulted in refined methods of measuring magnetic fields.

It is often said that the universe consists 99% of plasma. Alfvén used to say that it seems as if only the crust of the Earth is not plasma. In the mid-1960s, this author gave a talk at the Royal Institute of Technology in Stockholm about plasmas in solids (electrons and holes), which Hannes Alfvén himself attended. Among other things, the talk described recent observations of Alfvén waves in such plasmas, and Alfvén said: "Ah, they are here also. How interesting, I did not know that."

It was not until the middle of the 20th century that more intensive investigations on Alfvén waves in space and laboratory plasmas began. The slow development of the field of space plasmas was possibly because many physicists were not acquainted with the fact that electric currents can be distributed in large volumes and magnetic fields in such volumes can be present. Since then, the gigantic laboratory of the universe from the aurora originating in the Earth's magnetosphere to quasars at the rim of the universe has attracted immense interest with regard to Alfvén waves.

When propagating in inhomogeneous plasmas, for example, in the magnetosphere, the Alfvén wave experiences many interesting phenomena, including mode coupling, resonant mode conversion, and resonant absorption. We now know that shear Alfvén waves lie behind the phenomena of micropulsations in the geomagnetic field and also acceleration of particles. Micropulsations were detected a hundred years ago with simple magnetometers on the ground. It took more than 50 years before it was understood that they were related to the magnetosphere. Solar physics is another fascinating field where Alfvén waves occur, giving rise to sunspots (Alfvén, 1942). The vast amount of energy exhibited in eruptions of particles on the solar surface, originating in the interior of the Sun, is probably transported by Alfvén waves. These also play a role in heating the solar corona. Alfvén waves were first identified in the solar wind by means of spacecraft measurements by the end of the 1960s. They also occur in the exosphere of comets. A new and promising area of research is laboratory astrophysics using high-intensity particle and photon beams that may shed light on superstrong fields in plasmas.

For applications to confinement and heating of fusion plasmas, for example, in Tokamak devices, shear Alfvén waves have been studied in toroidal plasmas, accounting for nonuniform plasmas in axisymmetric situations. It is believed that the remaining exciting challenges lie in the area of nonlinear physics of shear Alfvén waves and associated particle dynamics and anomalous losses of α particles in a deuterium-tritium plasma. Collective modes in inhomogeneous plasmas as well as energy and particle transport in plasmas with transport barriers are of paramount importance for the design of a future Tokamak power plant (Parail, 2002).

Nonlinear transport processes in laboratory and cosmic plasmas have much in common (Wilhelmsson, 2000; Wilhelmsson & Lazzaro, 2001). Similarities (and discrepancies) could be highly indicative and beneficial for an improved understanding of specific phenomena as well as for plasma dynamics in general—possibly even for describing the evolution of the universe (Wilhelmsson, 2002).

HANS WILHELMSSON

See also **Magnetohydrodynamics; Nonlinear plasma waves; Plasma soliton experiments**

Further Reading

Alfvén, H. 1942. Existence of electromagnetic-hydrodynamic waves. *Nature*, 150: 405–406

Fälthammar, C.-G. 1995. Hannes Alfvén. In *Alfvén Waves in Cosmic and Laboratory Plasmas*, edited by A.C.-L. Chian, A.S. de Assis, C.A. de Azevedo, P.K. Shukla & L. Stenflo, Proceedings of the International Workshop on Alfvén Waves, *Physica Scripta*, T60: 7

Lehnert, B. 1954. Magnetohydrodynamic waves in liquid sodium. *Physical Review*, 94: 815

Lundquist, S. 1949. Experimental demonstration of magnetohydrodynamic waves. *Nature*, 164: 145

Mahajan, S.M. 1995. Spectrum of Alfvén waves, a brief review. *Physica Scripta*, T60: 160–170

Marston, E.H. & Kao Y.H. 1969. Damped Alfvén waves in bismuth. A determination of charge-carrier relaxation times. *Physical Review*, 182: 504

Parail, V.V. 2002. Energy and particle transport in plasmas with transport barriers. *Plasma Physics and Controlled Fusion*, 44: A63–85

Wilhelmsson, H. (editor). 1982. The physics of hot plasmas. Proceedings of the International Conference on Plasma Physics, Göteborg, May 1982, *Physica Scripta*, T2 (1 and 2)

Wilhelmsson, H. (editor). 1976. *Plasma Physics: Nonlinear Theory and Experiments*, New York and London: Plenum Press

Wilhelmsson, H. 2000. *Fusion: A Voyage through the Plasma Universe*, Bristol and Philadelphia: Institute of Physics Publishing

Wilhelmsson, H. 2002. Gravitational contraction and plasma fusion burn; universal expansion and the Hubble Law. *Physica Scripta*, 66: 395

Wilhelmsson, H. & Lazzaro, E. 2001. *Reaction-Diffusion Problems in the Physics of Hot Plasmas*, Bristol and Philadelphia: Institute of Physics Publishing

ALGORITHMIC COMPLEXITY

The notion of complexity as an object of scientific interest is relatively new. Prior to the 20th century, the main concern was that of simplicity, with complexity being the denigrated opposite. This idea of simplicity has had a long history enshrined in the dictum of the 14th-century Franciscan philosopher, William of Occam, that "pluritas non est ponenda sine necessitate" [being is not multiplied without necessity], and passed on simply as "Occam's razor," or more prosaically, "keep it simple" (Thorburn, 1918).

Indeed, the razor has been invoked by such notables as Isaac Newton and, in modern times, Albert Einstein and Stephen Hawking to justify parsimony in the adoption of physical principles. Although the dictum has proved its usefulness as a support for many scientific theories, the last century witnessed a gradual concern for simplicity's complement. Implicit was a recognition that beyond logical partitions was a need to quantify the simple/complex continuum. Perhaps the first milestone on the road to quantifying complexity came with Claude Shannon's famous information entropy in the late 1940s. Although it was not specifically developed as a complexity measure, the information connection made by Warren Weaver soon provided an impetus for sustained interest in information as a unifying concept for complexity (Weaver, 1948).

Shannon approached information as a statistical measure of receiving a message (Shannon, 1948): if $p_1, p_2, \ldots, p_N$ are the probabilities of receiving messages $m_1, m_2, \ldots, m_N$, then the information carried is defined by

$$I = -\sum_{i}^{N} p_i \log_2 p_i . \qquad (1)$$

Information is typically referred to as a measure of surprise; that is, the more unlikely a message, the more information it contains. To some degree, information is related to the notion of randomness in that the more regular (less complex, less random) something is, the less surprise is available. A simple calculation of this entropy demonstrates that the maximum of the function is achieved when all probabilities are equal. Shannon had a measure of capacities of a communication channel as his goal and did not concern himself with individual objects of the messages. Nonetheless, the quantification in terms of probabilities provided a basis for viewing complexity.

This theme was soon independently taken up by Ray Solomonoff (1964), Andrey Kolmogorov (1965), and Gregory Chaitin (1966). In a sense, Somolonoff was looking for a way to measure the effect of Occam's razor; that is, how can one measure objectively the simplicity of a scientific theory? Kolmogorov and Chaitin, on the other hand, were interested in a measure of complexity of individual objects, as opposed to Shannon's average. This Kolmogorov–Chaitin complexity has come to be known variously as algorithmic complexity, algorithmic entropy, and algorithmic randomness, among other designations. Both Kolmogorov and Chaitin were interested in binary number strings as objects and the ability to define the complexity of a string in terms of the shortest algorithm that prints out the string. Again, regularity and randomness is involved (Gammarman & Vovk, 1999). Consider, for example, the simple bit string, 101010101010, . . .; the minimal program to write the string requires only the pattern 10, the length of the string, and the "repeat, write" instructions, or

$$K(s) = \min\{|p| : s = C_T(p)\}, \qquad (2)$$

where $K(s)$ is the Kolmogorov complexity of the string, $|p|$ is the program length in bits, and $C_T(p)$ is the result of running program p on a universal Turing machine T.

Clearly, the recognition that patterns play an important role in defining complexity re-emphasized their importance in terms of data compression. In the early 1950s, David Huffman recognized their importance, and algorithmic complexity reaffirmed their utility with the ascendancy of computers and their demand for storage space (Huffman, 1952). Thus, numerous coding schemes were developed to take advantage of the fact that a simple algorithm can compress long data streams based upon the idea that recurrent patterns exist.

The efforts of Somonoloff, Kolmogorov, and Chaitin spawned numerous alternative measures of complexity, often seeking to address identified deficiencies in the definitions (Shalizi & Crutchfield, 2001). Among the deficiencies pointed out were the following: (i) Complexity is defined in terms of randomness—it is maximized by random strings. Is this what is really sought? (ii) Complexity is uncomputable, since there is no algorithm to compute it on a universal computing machine. (iii) Complexity does not provide information regarding structural patterns or organizations that have the same amount of disorder.

These questions were compounded by the expanding field of nonlinear dynamics. Kolmogorov's earlier entropy (1958)—developed to determine the rate of information creation—were among the invariants used to distinguish chaotic systems and an inferred complexity. Moreover, the description of physical dynamical systems became an additional issue (Zurek, 1989). Physical processes were typically described along the continuum of two extremes: periodic or random. However, both such systems are simply described, one by a recurrent pattern and the other by a statistical description. While information is high in a random system, it is low in a periodic process. Amalgams of two such processes might require considerable computational effort, yet no metric sufficiently expressed this. Certainly, such combined processes (random and periodic) may exhibit moderate information but the most concise description may be quite complex.

Some of these difficulties have been addressed with varying degrees of acceptance (Wakerberger et al., 1994). Increasingly, however, it appears that the question is evolving along two different lines: a formal approach (rules) with its main ramifications redounding to mathematics and computer science and a physical approach (equations) dealing with the characterization of systems.

Both approaches have in common the emphasis on the reconstruction (prediction) of an observed system and on the need to give the most parsimonious recipe for generating the studied entity. The metrics of complexity is thus expressed in terms of program lines for the mathematics-oriented option and in dimensionality for the physics-oriented definition, but there is the same basic notion of complexity as the inverse of compressibility of a given object (Boffetta et al., 2002).

This notion of compressibility has an immediate translation in terms of both multidimensional statistics and technology. In multidimensional statistics, the compressibility of a given data set corresponds to the percentage of explained variance by the optimal (generally in a least-squares sense) model of the data. More generally, something can be compressed if there exists some sort of correlation structure linking the different portions of a system, the existence of such correlations implying that the information about one part of the system is implicit in another part. Thus, all the information is not needed to reconstruct the entire system. It is evident how this concept corresponds to the cognate concept of redundancy, bringing us back to the notion of entropy (Giuliani et al., 2001).

Clearly, the diverse algorithms designed to measure complexity suggest a commonality. The question remains as to whether one metric is sufficient for its characterization.

JOSEPH P. ZBILUT AND ALESSANDRO GIULIANI

See also **Entropy; Information theory; Structural complexity**

Further Reading

Boffetta, G., Cencini, M., Falcioni, M. & Vulpiani, A. 2002. Predictability: a way to characterize complexity. *Physics Reports*, 356: 367–474

Chaitin, G.J. 1966. On the length of programs for computing finite binary sequences. *Journal of the Association for Computing Machinery*, 13: 547–569

Gammerman, A. & Vovk, V. 1999. Kolmogorov complexity: sources, theory and applications. *The Computer Journal*, 42: 252–255

Gell-Mann, M. & Lloyd, S. 1999. Information measures, effective complexity, and total information. *Complexity*, 2: 44–52

Giuliani, A., Colafranceschi, M., Webber Jr., C.L. & Zbilut, J.P. 2001. A complexity score derived from principal components analysis of nonlinear order measures. *Physica* A, 301: 567–588

Huffman, D.A. 1952. A method for the construction of minimum redundancy codes. *Proceedings IRE*, 40: 1098–1101

Kolmogorov, A.N. 1958. A new metric invariant of transitive dynamical systems and automorphism in Lebesgue spaces. *Doklady Akademii Nauk SSSR [Proceedings of the Academy of Sciences of the USSR]*, 119: 861–864

Kolmogorov, A.N. 1965. Tri podkhoda k opredeleniiu poniatiia "kolichestvo informatsii". [Three approaches to the quantitative definition of information.] *Problemy Peredachy Informatsii [Problems Information Transmission]*, 1: 3–11

Lempel, A. & Ziv, J. 1976. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22: 75–81

Li, M. & Vitányi, P. 1997. *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd edition, New York: Springer

Salomon, D. 1998. *Data Compression, the Complete Reference*, New York: Springer

Shalizi, C.R. & Chrutchfield, J.P. 2001. Computational mechanics: pattern and prediction, structure and simplicity. *Journal of Statistical Physics* 104: 819–881

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423

Solomonoff, R.J. 1964. The formal theory of inductive inference, parts 1 and 2. *Information and Control*, 7: 1–22, 224–254

Thorburn, W.M. 1918. The myth of Occam's razor. *Mind*, 27: 345–353

Wackerberger, R., Witt, A., Atmanspacher, H., Kurths, J. & Scheingraber, H. 1994. A comparative classification of complexity measures. *Chaos, Solitons & Fractals*, 4: 133–173

Weaver, W. 1948. Science and complexity. *American Scientist*, 36: 536–544

Zurek, W. 1989. Thermodynamic cost of computation, algorithmic complexity, and the information metric. *Nature*, 341: 119–124

ALL-OR-NOTHING RESPONSE

See **Nerve impulses**

ALMOST PERIODIC FUNCTIONS

See **Quasiperiodicity**

AMBIGUOUS FIGURES

See **Cell assemblies**

ANDERSON LOCALIZATION

Anderson localization is a phenomenon associated with the interference of waves in random media. Although Philip Anderson's original publication (Anderson, 1958) was actually motivated by experiments on the propagation of spin-waves in random magnets, the greatest application of the concept has been the study of electrical transport phenomena in metals and semiconductors. Over the past 10 years, more attention has been focused on other wave phenomena in random media, particularly optical phenomena.

Our current understanding of electronic transport in metals and semiconductors is based on the Schrödinger equation for the wave function of conduction electrons of the form

$$-\frac{\hbar^2}{2m^*}\nabla^2\psi(\boldsymbol{r}) + [U(\boldsymbol{r}) + V(\boldsymbol{r})]\,\psi(\boldsymbol{r}) = E\psi(\boldsymbol{r}),$$

where $U(\boldsymbol{r})$ is a periodic potential representing the regular lattice in the solid and $V(\boldsymbol{r})$ is a random function of position, which represents the presence of impurities in the system. In the absence of the random potential, the allowed energies of such an electron fall within a series of bands separated by energy gaps. The eigenfunctions in the absence of the random potential are all of the form $\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}}u_{j,\boldsymbol{k}}(\boldsymbol{r})$ where the wavevector $\boldsymbol{k}$ lies within the first Brillouin zone, j labels the energy bands and the Bloch function, $u_{j,\boldsymbol{k}}(\boldsymbol{r})$, has the periodicity of the regular potential, U. Such states are *extended* in the sense that their support covers the entire system and hence they can contribute to electrical conduction.

The eigenstates of an electron subject to a random potential may be of two types. Some are extended, although there may be strong local modulations in the amplitude. These states can contribute to electrical conduction through the material, even at zero temperature, as they connect the two ends of a sample. Other states, however, are localized in that their amplitude vanishes exponentially outside a specific finite region. These states are referred to as *Anderson localized* and can only contribute to conduction via thermal activation.

One can understand the existence of localized states by considering the low-energy states of an electron moving in a very rough, random potential, $V(\boldsymbol{r})$ (Lee & Ramakrishnan, 1985). The lowest energy states will be those bound to very deep troughs in the potential function. The mixing between states localized in different wells will be very weak because states with significant spatial overlap will have very different energies, while states with similar energies are spatially well separated so that the wave function overlaps are exponentially small.

The scale on which the wave functions of localized states decay to zero defines the localization length ξ, which depends on the energy of the state and the strength of the disorder. The balance between extended and localized states depends on the strength of the disorder and the spatial dimensionality of the

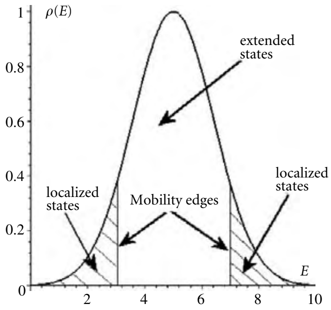

Figure 1. A schematic plot of the density of states showing a single disorder broadened band for a 3-d system. The states in the band center are extended while those in the tails are localized (shaded regions); the mobility edges between the two types of state are marked.

system. In a one-dimensional (1-d) system, all of the electronic eigenstates are strongly localized by any amount of disorder. In two dimensions, it is believed that all states are actually localized, but that the localization length can be very long in the center of a band. The application of a strong magnetic field to a disordered 2-d electron system, such as may be formed at a semiconductor heterojunction at low temperature, causes the conduction band to break up into a sequence of disorder broadened Landau bands, each with an extended state at its center—an essential feature of the quantum Hall effect. In three dimensions, the eigenstates at the center of a band are truly extended while those in the low- and high-energy tails are localized. It is believed that there are two well-defined critical energies within the band at which the nature of the states changes so that localized and extended states do not co-exist at a given energy.

The critical energies are usually referred to as mobility edges because the zero-temperature conductance of the system will be zero when the Fermi energy lies in the regime of localized states but nonzero in the extended regime (see Figure 1). The location of the mobility edges depends on the strength of the disorder —in very clean systems, only the states in the tails of a band will be localized while in a very dirty system the mobility edges may meet in the band center so that all states are localized. The behavior at the mobility edge has been studied by performing experiments on a series of devices with increasing amounts of disorder. The transition between metallic (conducting) and insulating behavior is closely analogous to other phase transitions. Mott (1973) supposed that this metal-insulator transition was first order, with the conductivity jumping from a fixed finite value, $\sigma_{\min}$, to zero. In 1979, a renormalization group analysis was carried out by the so-called "gang of four," based on earlier scaling arguments by Thouless and co-workers (Thouless, 1974), which favored a continuous metal-insulator transition for 3-d systems (Abrahams et al., 1979). The discussion of the change in nature of the states from extended to localized in terms of a zero-temperature quantum phase transition has been very fruitful. In this description, the localization length, ξ, plays the same role as the correlation length for fluctuations in a thermal transition. It is supposed that ξ diverges at the mobility edges according to a universal power law

$$\xi \sim |E - E_c|^{-\nu}\,.$$

Numerical evidence indicates that the value of the exponent is indeed universal and has the value $\nu \sim 1.6$.

Although the underlying physics of Anderson localization is that of linear waves in random media, the discussion can be recast in terms of nonlinear models without disorder, which are in the same family of statistical field theories used to describe thermal phase transitions, specifically nonlinear sigma models (Efetov, 1997). This has led to the notion that the spatial variation of the wave functions of states at the mobility edge displays a multifractal character.

The application of these ideas to other wave phenomena in random media has been slower. It is much harder to observe strong localization in bosonic and classical wave systems, but recently much experimental work has been carried out on optical and acoustic localization (see John (1990) for a good introduction). This work shows that Anderson localization is not an essentially quantum mechanical effect but is ubiquitous for wave propagation in random media. Similarly, the interplay between the physics of randomly disordered systems and quantum chaos is also proving very rich and fruitful (Altshuler & Simons, 1994).

There are a number of other mechanisms whereby wave excitations may become spatially localized. The propagation of excitations within macromolecules, for example, may become localized both because of interference effects associated with "random" changes in structure and also because of self-trapping effects associated with nonlinearity in the wave equation for these modes. In the case of electrons in solids, electronic excitations may become localized both by random variations in potential and by interaction effects that give rise to the so-called Mott transition. Such interaction effects are responsible for the phenomenon of Coulomb blockade observed in semiconductor nanostructures.

Keith Benedict

See also **Discrete self-trapping system; Local modes in molecular crystals**

Further Reading

Abrahams, E., Anderson, P.W., Licciardello, D.C. & Ramakrishnan, T.V. 1979. Scaling theory of localization: Absence of quantum diffusion in two dimensions. *Physical Review Letters*, 42: 673

Altshuler, B. & Simons, B.D. 1994. Universalities: from Anderson localization to quantum chaos. In *Mesoscopic Quantum Physics, Proceedings of the 61st Les Houches Summer School,* edited by E. Akkermans, G. Montambaux, J.-L. Pichard & J. Zinn-Justin, Amsterdam: North-Holland

Anderson, P.W. 1958. The absence of diffusion in certain random lattices. *Physical Review*, 109: 1492

Efetov, K. 1997. *Supersymmetry in Disorder and Chaos*. Cambridge and New York: Cambridge University Press

John, S. 1990. The localization of waves in disordered media. In *Scattering and Localization of Classical Waves in Random Media,* edited by P. Sheng. Singapore: World Scientific, pp. 1–96

Lee, P.A. & Ramakrishnan, T.V. 1985. Disordered electronic systems. *Reviews of Modern Physics*, 57: 287

Mott, N.F. 1973. In *Electronic and Structural Properties of Amorphous Semiconductors*, edited by P.G. LeComber & J. Mort, London: Academic Press, p. 1

Thouless, D.J. 1974. Electrons in disordered systems and the theory of localization. *Physics Reports*, 13: 93

ANNIHILATION (KINK-ANTIKINK)

See **Sine-Gordon equation**

ANOSOV AND AXIOM-A SYSTEMS

Two classes of dynamical systems exhibiting chaotic behavior were axiomatically defined and systematically studied for the first time in the 1960s. Previous studies had concentrated on more specific situations. Axiom-A systems were introduced by Stephen Smale in his seminal paper (Smale, 1967). Anosov systems, which are a special case of Axiom-A systems, were studied independently in Moscow around the same period. Today, Anosov and Axiom-A systems are valued as idealized models of chaos: while the conditions defining Axiom A are too stringent to include many real-life examples, it is recognized that they have features shared in various forms by most chaotic systems.

Definitions

First, we give the definitions in the discrete-time case.

Let f be a smooth invertible map (for basic notions, *See* **Phase space**). A compact invariant set of f is said to be *hyperbolic* if at every point in this set, the tangent space splits into a direct sum of two subspaces E^{u} and E^{s} with the property that these subspaces are invariant under the differential $\mathrm{d}f$, that is, $\mathrm{d}f(x)E^{\mathrm{u}}(x) = E^{\mathrm{u}}(f(x))$, $\mathrm{d}f(x)E^{\mathrm{s}}(x) = E^{\mathrm{u}}(f(x))$, and that $\mathrm{d}f$ expands vectors in E^{u} and contracts vectors in E^{s}.

If $E^{\mathrm{u}} = \{0\}$ in the definition above, then the invariant set is made up of attracting fixed points or periodic orbits. Similarly, if $E^{\mathrm{s}} = \{0\}$, then the orbits are repelling. If neither subspace is trivial, then the behavior is locally "saddle-like," that is to say, relative to the orbit of a point x, most nearby orbits diverge exponentially fast both in forward and in backward time. This is why hyperbolicity is a mathematical notion of chaos.

An *Anosov diffeomorphism* is a smooth invertible map of a compact manifold with the property that the entire space is a hyperbolic set.

Axiom A, which is a larger class, focuses on the part of the system that is not transient. More precisely, a point x in the phase space is said to be nonwandering if every neighborhood U of x contains an orbit that returns to U. A map is said to satisfy Axiom A if its nonwandering set is hyperbolic and contains a dense set of periodic points.

Definitions in the continuous-time case are analogous: f above is replaced by the time-t-maps of the flow, and the tangent spaces now decompose into $E^{\mathrm{u}} \oplus E^0 \oplus E^{\mathrm{s}}$ where E^0, which is 1-d, represents the direction of the flow lines.

Phase Space Structures and Properties

Anosov and Axiom-A systems are defined by the behavior of the differential. Corresponding to the linear structures left invariant by $\mathrm{d}f$ are nonlinear structures, namely stable manifolds tangent to E^{s} and unstable manifolds tangent to E^{u}.

Thus, two families of invariant manifolds are associated with an Anosov map and each one of these fills up the entire phase space; they are sometimes called the *stable* and *unstable foliations*. The leaves of these foliations are transverse at each point, forming a kind of (topological) coordinate system. The map f expands distances along the leaves of one of these foliations and contracts distances along the leaves of the other. For Axiom-A systems, one has a similar local product structure or "coordinate system" at each point in the nonwandering set, but the picture is local, and there are gaps: the stable and unstable leaves do not necessarily fill out open sets in the phase space.

In addition to these local structures, Axiom-A systems have a global structure theorem known as *spectral decomposition*. It says that the nonwandering set of every Axiom-A map can be written as $X_1 \cup \cdots \cup X_r$ where the X_i are disjoint closed invariant sets on which f is topologically transitive. The X_i are called basic sets. Each X_i can be decomposed further into a finite union $\bigcup X_{i,j}$, where each $X_{i,j}$ is invariant and topologically mixing under some iterate of f. (Topological transitivity and mixing are irreducibility conditions; *See* **Phase space**.) This decomposition is reminiscent of the corresponding result for finite-state Markov chains.

One of the reasons why hyperbolic sets are important is their robustness: they cannot be perturbed away. More precisely, let f be a map with a hyperbolic set Λ that is locally maximal, that is, it is the largest invariant set in some neighborhood U. Then for every map g that is C^1 near f, the largest invariant set Λ' of g in

U is again hyperbolic; moreover, f restricted to Λ is *topologically conjugate* to g restricted to Λ'. This is mathematical shorthand for saying that not only are the two sets Λ and Λ' topologically indistinguishable, but the orbit structure of f on Λ is indistinguishable from that of g on Λ'.

The above phenomenon brings us to the idea of *structural stability*. A map f is said to be *structurally stable* if every map g, that is C^1 near f is topologically conjugate to f (on the entire phase space). It turns out that a map is structurally stable if and only if it satisfies Axiom A and an additional condition called strong transversality.

Next, we discuss the idea of pseudo-orbits versus real orbits. Letting $d(\cdot,\cdot)$ be the metric, a sequence of points $x_0, x_1, x_2, \ldots$ in the phase space is called an *ε-pseudo-orbit* of f if $d(f(x_i), x_{i+1}) < \varepsilon$ for every i. Computer-generated orbits, for example, are pseudo-orbits due to round-off errors. A fact of consequence to people performing numerical experiments is that in hyperbolic systems, small errors at each step get magnified exponentially fast. For example, if the expansion rate is ≥ 3, then an ε-error made at one step is tripled at each subsequent step, that is, after only $O(|\log \varepsilon|)$ iterates, the error is $O(1)$, and the pseudo-orbit bears no relation to the real one. There is, however, a theorem that states that every pseudo-orbit is shadowed by a real one. More precisely, given a hyperbolic set, there is a constant C such that if $x_0, x_1, x_2, \ldots$ is an ε-pseudo-orbit, then there is a phase point z such that $d(x_i, f^i(z)) < C\varepsilon$ for all i. Thus, paradoxical as it may first seem, this result asserts that on hyperbolic sets, each pseudo-orbit approximates a real orbit, even though it may deviate considerably from the one with the same initial condition.

The shadowing orbit corresponding to a bi-infinite pseudo-orbit is, in fact, unique. From this, one deduces the following Closing Lemma: for any hyperbolic set, there is a constant C such that the following holds: every finite orbit segment $x, f(x), \ldots, f^{n-1}(x)$ that nearly closes up, that is, $d(x, f^{n-1}(x)) < \varepsilon$ for some small ε, lies within $< C\varepsilon$ of a genuine periodic orbit of period n. Thus, hyperbolic sets contain many periodic points.

Examples

A large class of Anosov diffeomorphisms comes from linear toral automorphisms, that is, maps of the n-dimensional torus induced by $n \times n$ matrices with integer entries, det $= \pm 1$, and no eigenvalues of modulus one. (*See* **Cat map** for a detailed example of this). We remark that due to their structural stability (nonlinear), perturbations of linear toral automorphisms continue to have the Anosov property. This remark also applies to all of the examples below. In fact, all known Anosov diffeomorphisms are topologically identical to a linear toral automorphism (or a slight generalization of these).

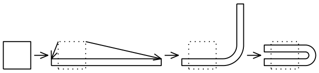

Figure 1. The horseshoe.

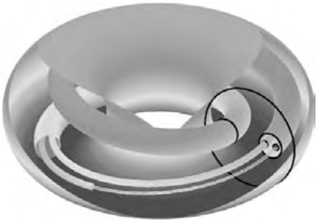

Figure 2. The solenoid.

Geodesic flows describe free motions of points on manifolds. Let M be a manifold. Given $x \in M$ and a unit vector v at x, there is a unique geodesic starting from x in the direction v. The geodesic flow φ^t is given by $\varphi^t(x, v) = (x', v')$, where x' is the point t units down the geodesic and v' is the direction at x'. Geodesic flows on manifolds of strictly negative curvature are the main examples of Anosov flows. They were studied by Jacques Hadamard (ca. 1900) and Gustav Hedlund and Eberhard Hopf (in the 1930s) considerably before Anosov theory was developed.

Smale's *horseshoe* is the prototypical example of a hyperbolic invariant set. This map, so called because it bends a rectangle B into the shape of a horseshoe and puts it back on top of B, is shown in Figure 1. The set $\{x\colon f^n(x) \in B \text{ for all } n = 0, \pm 1, \pm 2, \ldots\}$ is hyperbolic (*See* **Horseshoes and hyperbolicity in dynamical systems; Phase space**).

Finally, we mention the *solenoid* (see Figure 2, and also in the color plate section as the Smale solenoid), which is an example of an Axiom-A attractor. Here, the map f is defined on a solid torus $M = S^1 \times D_2$, where D_2 is a 2-d disk. It is easiest to describe it in two steps: first it maps M into a long thin solid torus, which is then placed inside M winding around the S^1 direction twice. The attractor is given by $\Lambda = \bigcap_{n \geq 0} f^n(M)$.

Symbolic Coding of Orbits and Ergodic Theory

An important tool for studying the orbit structure of Axiom-A systems is the *Markov partition*, constructed for Anosov systems by Sinai and extended to Axiom-A basic sets by Bowen. Given a partition $\{R_1, \ldots, R_k\}$ of the phase space, there is a natural way to attach to each point x in the phase space a sequence of symbols, namely $(\ldots, a_{-1}, a_0, a_1, a_2, \ldots)$ where $a_i \in \{1, 2, \ldots, k\}$ is the name of the partition element containing $f^i(x)$, that is, $f^i(x) \in R_{a_i}$ for each i. In general, not all sequences are realized by orbits of f. Markov partitions are designed so that the set of symbol sequences that correspond to real orbits has Markovian properties; it is called a shift of finite type (*See* **Symbolic dynamics**).

The ergodic theory of Axiom-A systems has its origins in statistical mechanics. In a 1-d lattice model in statistical mechanics, one has an infinite array of sites indexed by the integers; at each site, the system can be in any one of a finite number of states. Thus, the configuration space for a 1-d lattice model is the set of bi-infinite sequences on a finite alphabet. Identifying this symbol space with the one from Markov partitions, Sinai and Ruelle were able to transport some of the basic ideas from statistical mechanics, including the notions of Gibbs states and equilibrium states, to the ergodic theory of Axiom-A systems.

The notion of *equilibrium states*, which is equivalent to Gibbs states for Axiom-A systems, has the following meaning in dynamical systems in general: given a potential function φ, an invariant measure is said to be an equilibrium state if it maximizes the quantity

$$h_\mu(f) - \int \varphi \,\mathrm{d}\mu,$$

where $h_\mu(f)$ denotes the Kolmogorov–Sinai entropy of f and the supremum is taken over all f-invariant probability measures μ. In particular, when $\varphi = 0$, this measure is the measure that maximizes entropy; and when $\varphi = \log|\det(\mathrm{d}f|_{E^u})|$, it is the Sinai–Ruelle–Bowen (SRB) measure. From a physical or observational point of view, SRB measures are the most important invariant measures for dissipative dynamical systems (*See* **Sinai–Ruelle–Bowen measures**).

Periodic Points and Their Growth Properties

We discuss briefly some further results related to the abundance of periodic points in Axiom-A systems.

For an Axiom-A diffeomorphism f, if $P(n)$ is the number of periodic points of period $\leq n$, then $P(n) \sim \mathrm{e}^{hn}$ where h is the topological entropy of f. That is to say, the dynamical complexity of f is reflected in its periodic behavior. An analogous result holds for Axiom-A flows.

Finally, we mention the dynamical zeta function, which sums up the periodic information of a system. In the discrete-time case, $\zeta(z) := \exp \sum_{n=1}^{\infty} P(n)z^n/n$ has been shown to be a rational function analytic on $|z| < \mathrm{e}^{-h}$. In the continuous-time case, the zeta function is given by $\zeta(z) := \prod_\gamma (1 - \exp(-z\,l(\gamma)))^{-1}$, where the product is taken over all (nonstationary) periodic orbits γ and $l(\gamma)$ is the smallest positive period of γ. This function is known to be meromorphic on a certain domain, but the locations of its poles, which are intimately related to correlation decay properties of the system, remain one of the yet unresolved issues in Axiom-A theory.

Boris Hasselblatt and Lai-Sang Young

See also **Cat map; Horseshoes and hyperbolicity in dynamical systems; Phase space; Sinai–Ruelle–Bowen measures; Symbolic dynamics**

Further Reading

Bowen, R. 1975. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Berlin and New York: Springer

Branner, B. & Hjorth, P. (editors). 1995. *Real and Complex Dynamical Systems. Proceedings of the NATO Advanced Study Institute held in Hillerød, June 20–July 2, 1993*, Dordrecht and Boston: Kluwer

Fielder, B. (editor). 2002. *Handbook of Dynamical Systems*, Vol. 2, Amsterdam and New York: Elsevier

Hasselblatt, B. & Katok, A. (editors). 2002. *Handbook of Dynamical Systems*, Vol. 1A, Amsterdam and New York: Elsevier

Hasselblatt, B. & Katok, A. 2003. *Dynamics: A First Course*, Cambridge and New York: Cambridge University Press

Katok, A. & Hasselblatt, B. 1995. *Introduction to the Modern Theory of Dynamical Systems*, Cambridge and New York: Cambridge University Press

Smale, S. 1967. Differentiable dynamical systems. *Bulletin of the American Mathematical Society*, 73: 747–817

ANTISOLITONS

See **Solitons, types of**

ANTI-STOKES SCATTERING

See **Rayleigh and Raman scattering and IR absorption**

ARNOL'D CAT MAP

See **Cat map**

ARNOL'D DIFFUSION

For near-integrable Hamiltonian systems with more than two degrees of freedom, stochastic and regular trajectories are intimately co-mingled in the $2N$-dimensional phase space. Stochastic layers in phase space exist near resonances of the motion. The thickness of the layers expands with increasing perturbation, leading to primary resonance overlap, motion across the layers, and the appearance of strong stochasticity in the motion. In the limit of weak perturbation, however, primary resonance overlap does not occur. A new physical behavior of the motion then makes its appearance: motion along the resonance layers called Arnol'd diffusion (AD). For two degrees of freedom, with a weak perturbation, two-dimensional

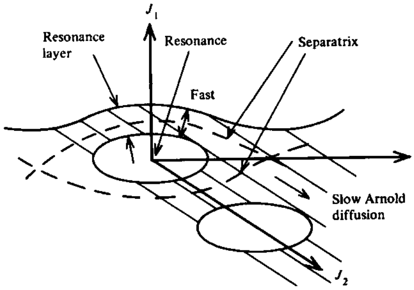

Figure 1. Illustration of the directions of the fast diffusion across a resonance layer and the slow diffusion along the resonance layer.

Kolmogorov–Arnol'd–Moser (KAM) surfaces divide the three-dimensional energy "volume" in phase space into a set of closed volumes each bounded by KAM surfaces, much as lines isolate regions of a plane. For $N > 2$ degrees of freedom, the N-dimensional KAM surfaces do not divide the $(2N - 1)$-dimensional energy volume into distinct regions. Thus, for $N > 2$, in the generic case, all stochastic layers of the energy surface in phase space are connected into a single complex network—the Arnol'd web. The web permeates the entire phase space, intersecting or lying infinitesimally close to every point. For an initial condition within the web, the subsequent stochastic motion will eventually intersect every finite region of the energy surface in phase space, even in the limit as the perturbation strength approaches zero.

The merging of stochastic trajectories into a single web was proved (Arnol'd, 1964) for a specific nonlinear Hamiltonian. A general proof of the existence of a single web has not been given, but many computational examples support the conjecture. From a practical point of view, there are two major questions with respect to AD in a particular system: what is the relative measure of stochastic trajectories (fraction of the phase space that is stochastic) in the region of interest? And for a given initial condition, how fast will system points diffuse along the thin threads of the Arnol'd web?

We illustrate the motion along the resonance layer in Figure 1. A projection of the motion onto the J_1, θ_1 plane is shown, illustrating a resonance with a stochastic layer. At right angles to this plane, the action of the other coordinate J_2 is shown. If there are only two degrees of freedom in a conservative system, the fact that the motion is constrained to lie on a constant energy surface restricts the change in J_2 for J_1 constrained to the stochastic layer. However, if there is another degree of freedom, or if the Hamiltonian is time dependent, then this restriction is lifted, and motion along the stochastic layer in the J_2 direction can occur.

The diffusion rate (D) along a layer has been calculated by Chirikov (1979) for the important case of three resonances, and by Tennyson et al. (1979) for an equivalent mapping model, which they called a stochastic pump. These models predict, for a single action I corresponding to J_2 in Figure 1,

$$D = (\Delta I)^2/t \propto \mathrm{e}^{-A/\varepsilon^{1/2}}, \qquad (1)$$

where t is the time, ε is a perturbation parameter, and $A \approx 1$. For coupling among many resonances, a rigorous upper bound on the diffusion rate (Nekhoroshev, 1977) generally overestimates the rate by orders of magnitude. Using a similar formalism with a somewhat more restrictive class of Hamiltonians, but still encompassing most physical problems, the upper bound can be improved (Benettin et al., 1985; Lochak & Neistadt, 1992) to give what they considered to be an optimal upper bound:

$$D \propto \mathrm{e}^{-A/\varepsilon^{\gamma}}, \ \gamma \approx N^{-1}. \qquad (2)$$

If N is large, such an exponentially small diffusion could only hold for very small ε (specified within the theory), otherwise the exponential factor could be essentially unity. Also, an upper bound must be related to the fastest local diffusion. This may be much more rapid than an average global diffusion, which would be controlled by the portions of the phase space where the diffusion is slowest. For upper bound calculations, consult the original papers of Nekhoroshev (1977), Benettin et al. (1985), and Lochak & Neistadt (1992).

The simplest way to calculate local AD is to couple two standard maps together with a weak coupling term $\mu \sin(\theta_n + \phi_n)$, where θ_n and ϕ_n are the map phases and $\mu \ll 1$. Using the stochastic pump model, with a regular orbit (in the absence of coupling) in the (I, θ) map being driven by stochasticity in the (J, ϕ) map, the Hamiltonian of the mapping is approximated as $H \approx H_i + H_j$, with

$$\begin{aligned} H_i &= I^2/2 + K_i \cos\theta + 2\mu \cos(\theta + \phi), \\ H_j &= J^2/2 + K_j \cos\phi + 2K_j \cos\phi \cos 2\pi n, \end{aligned} \qquad (3)$$

where n is the time normalized to mapping periods. We have retained only the lowest Fourier term from the mapping frequency in the H_j equation of (3), and considered that the stochasticity in H_i is driven by the coupling. To calculate the changes in H_i per iteration due to kicks delivered by (J, ϕ), we take the derivative

$$\begin{aligned} \frac{\partial H_i}{\partial n} &= \frac{\mathrm{d}H_i}{\mathrm{d}n} \\ &= \frac{\mathrm{d}}{\mathrm{d}n}[2\mu \cos(\theta + \phi)] \\ &\quad + 2\mu \frac{\mathrm{d}\theta}{\mathrm{d}n} \sin[\theta + \phi(n)]. \end{aligned} \qquad (4)$$

For rotational orbits $\theta = w_i n + \theta_0$, scaling the time variable to revolutions of the map ($s = \omega_j n$), and defining the ratio of frequencies ($Q_0 = \omega_i/\omega_j = \omega_i/K_j^{1/2}$), Equation (4) is integrated to obtain

$$\Delta H_i = 2\mu Q_0 \left(\cos\theta_0 \int_{-\infty}^{\infty} \sin[Q_0 s + \phi(s)]\mathrm{d}s + \sin\theta_0 \int_{-\infty}^{\infty} \cos[Q_0 s + \phi(s)]\mathrm{d}s \right). \qquad (5)$$

The first of the integrals in (5) integrates to zero; the second is a Mel'nikov–Arnol'd integral (Chirikov, 1979, Appendix A), which can be evaluated to give the change in ΔH_i over one characteristic half-period of the (J, ϕ) map. Squaring ΔH_i and averaging over θ_0 gives

$$\langle(\Delta H_i)^2\rangle = 32\pi^2 Q_0^4 \mu^2 \frac{\sinh^2(\pi Q_0/2)}{\sinh^2(\pi Q_0)}. \qquad (6)$$

To determine the diffusion constant D, divide $\langle(\Delta H_i)^2\rangle$ by twice the average number of iterations in this half-period

$$\overline{T}_j = \frac{1}{\omega_j} \ln \left| \frac{32\mathrm{e}}{w_1} \right|, \qquad (7)$$

where $w_1 = \Delta H / H_{\text{separatrix}}$ is the relative energy of the edge of the stochastic region, $w_1 = 8\pi(2\pi/K_j^{1/2})^3 \times \mathrm{e}^{-\pi^2/K_j^{1/2}}$, and e is the base of natural logarithms. Combining (6) and (7), and using $\Delta H_i = I \Delta I$, the diffusion constant in action space can be approximated in a form that exhibits the main Q_0 scaling:

$$D \approx 16\mu^2 n Q_0^2 \mathrm{e}^{-\pi Q_0}, \qquad (8)$$

where we have assumed that $I \approx \omega_i$. Comparing (8) to (1) with $Q_0 = \omega_i/K_j^{1/2}$, we observe that $K_j \propto \varepsilon$, the perturbation parameter. The numerical results agreed well with (8) (see Lichtenberg & Aswani (1998) and references therein).

Chirikov et al. (1979) found, numerically, that one could distinguish the diffusion in a range where ε was sufficiently large and a single resonance was dominant, such that a three-resonance model scaling as in (1) holds, from a range of smaller values of ε with many overlapping weak resonances, where the scaling in (2) applies. The results of their numerical investigation demonstrated the transition between the two regimes. In another approach, the diffusion through a large number of weakly coupled standard mappings was determined numerically, with the strength of the coupling controlled in a manner such that the three-resonance model could be applied in a statistical manner to determine the diffusion rate (Lichtenberg & Aswani, 1998).

These studies indicate that the basic models can be used to determine Arnol'd diffusion in multidimensional systems if the system parameters can be sufficiently controlled. For more information on these and related topics, the reader is referred to Chirikov (1979) and to Lichtenberg & Lieberman (1991, Chapter 6).

ALLAN J. LIICHTENBERG

See also **Kolmogorov–Arnol'd–Moser theorem; Phase space diffusion and correlations; Standard map**

Further Reading

Arnol'd, V.I. 1964. Instability of dynamical systems with several degrees of freedom. *Russian Mathematical Surveys*, 18: 85

Benettin, G., Galgani, L. & Giorgilli, A. 1985. A proof of Nekoroshev's theorem for the stability times of nearly integrable Hamiltonian systems. *Celestial Mechanics*, 37: 1–25

Chirikov, B.V. 1979. A universal instability of many-dimensional oscillator systems. *Physics Reports*, 52: 265–379

Chirikov, B.V., Ford, J. & Vivaldi, F. 1979. Some numerical studies of AD in a simple model. In *Nonlinear Dynamics and the Beam-Beam Interaction*, edited by M. Month & J.C. Herrera, New York: American Institute of Physics

Lichtenberg, A.J. & Aswani, A.M. 1998. Arnold diffusion in many weakly coupled mappings. *Physical Review* E, 57: 5325–5331

Lichtenberg, A.J. & Lieberman, M.A. 1991. *Regular and Chaotic Dynamics*, 2nd edition, New York: Springer

Lochak, P. & Neistadt, A.I. 1992. Estimates in the theorem of N.N. Nekhoroshev for systems with quasi-convex Hamiltonian. *Chaos*, 2: 495–499

Nekhoroshev, N.N. 1977. An exponential estimate of the time of stability of nearly itegrable Hamiltonian systems. *Russian Mathematical Surveys*, 32: 1–65

Tennyson, J.L., Lieberman, M.A. & Lichtenberg, A.J. 1979. Diffusion in near-integrable Hamiltonian systems with three degrees of freedom. In *Nonlinear Dynamics and the Beam-Beam Interaction*, edited by M. Month & J.C. Herrera, New York: American Institute of Physics

ARNOL'D TONGUES

See **Coupled oscillators**

ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is a field of research in computer science reproducing intelligent reasoning. AI programs are mainly based on logic-oriented symbolic languages such as, for example, Prolog (Programming in Logic) or LISP (List Programming). Historically, AI was inspired by Alan Turing's question: "Can machines think?" According to the Turing test for AI, a machine is intelligent if a human user cannot distinguish whether he or she is interacting and communicating with a machine or a human being. Thus, before starting with AI, a general concept of computer and computabilty must be defined in computer science.

In 1936, Turing and Emil Post independently suggested the following definition of computability.

A "Turing machine" consists of: (a) a control box in which a finite program is placed, (b) a potentially infinite tape, divided lengthwise into squares, and (c) a device for scanning, or printing on one square of the tape at a time, and for moving along the tape or stopping, all under the command of the control box.

If the symbols used by a Turing machine are restricted to a stroke / and a blank *, then every natural number x can be represented by a sequence of x strokes (e.g., 3 by ///), each stroke on a square of the Turing tape. The blank is used to denote that the square is empty (or the corresponding number is zero). In particular, a blank is necessary to separate sequences of strokes representing numbers. Thus, a Turing machine computes a numerical function f with arguments $x_1, \ldots, x_n$ if the machine program starts with the input tape $\ldots * x_1 * x_2 * \ldots * x_n * \ldots$ and stops after finite steps with an output $\ldots * x_1 * x_2 * \ldots * x_n * f(x_1, \ldots, x_n) * \ldots$ on the tape. From a logical point of view, John von Neumann's general-purpose computer is a technical realization of a universal Turing machine that can simulate any kind of Turing program.

Besides Turing machines, there are many other mathematically equivalent procedures for defining computability (e.g., register machines, recursive functions) that are mathematically equivalent. According to Alonzo Church's thesis, the informal intuitive notion of an algorithm is identical to one of these equivalent mathematical concepts, for example, the program of a Turing machine.

With respect to AI, the paradigm of effective computabilty implies that the mind is represented by program-controlled machines, and mental structures refer to symbolic data structures, while mental processes implement algorithms. Historically, the hard core of AI was established during the Dartmouth Conference in 1956 when leading researchers, such as John McCarthy, Alan Newell, Herbert Simon, and others from different disciplines, formed the new scientific community of AI. If human thinking can be represented by an algorithm, then according to Church's thesis, it can be represented by a Turing program that can be computed by a universal Turing machine. Thus, human thinking could be simulated by a general-purpose computer and, in this sense, Turing's question ("Can machines think?") must be answered with a "yes." The premise that human thinking can be codified and represented by recursive procedures is, of course, doubtful. Even processes of mathematical thinking can be more complex than recursive functions.

The first period of AI was dominated by questions of heuristic programming, which means the automated search for human problem solutions in trees of possible derivations, controlled and evaluated by heuristics. In 1962, these simulative procedures were generalized and enlarged for the so-called General Problem Solver (GPS), which was assumed to be the heuristic framework of human problem solving. But GPS could only solve some insignificant problems in a formalized microworld. Thus, AI researchers tried to construct specialized systems of problem solving that use the specialized knowledge of human experts.

The architecture of an "expert system" consists of the following components: knowledge base, problem-solving component (interference system), explanation component, knowledge acquisition, and dialogue component. Knowledge is the key factor in the performance of an expert system. The knowledge is of two types. The first type is the facts of the domain that are written in textbooks and journals in the field. Equally important to the practice of a field is the second type of knowledge, called heuristic knowledge, which is the knowledge of good practice and judgment in a field. It is experimental knowledge, that art of good guessing, that a human expert acquires over years of work.

Expert systems are computational models of problem-solving procedures that need symbolic representation of knowledge. Unlike program-controlled serial computers, the human brain and mind are characterized by learning processes without symbolic representations. With respect to the architecture of von Neumann computers, an essential limitation derives from the sequential and centralized control, but complex dynamical systems like the brain are intrinsically parallel and self-organized.

In their famous paper "A Logical Calculus of the Ideas Immanent in Nervous Activity" in 1943, Warren McCulloch and Walter Pitts offered a complex model of neurons as threshold logic units with excitatory and inhibitory synapses. Their "McCulloch–Pitts neuron" fires an impulse along its axon at time $t + 1$ if the weighted sum of its inputs and weights at time t exceeds the threshold of the neuron. The weights are numbers corresponding to the neurochemical interactions of the neuron with other neurons. But, in a McCulloch–Pitts network, the function of an artificial neuron is fixed for all time. McCulloch and Pitts succeeded in demonstrating that a network of formal neurons of their type could compute any finite logical expression.

In order to make a neural computer capable of complex tasks, it is necessary to find procedures of learning. A learning procedure is nothing else than an adjustment of the many weights so that the desired output vector (e.g., a perception) is achieved. The first learning neural computer was Frank Rosenblatt's "Perceptron" (1957). Rosenblatt's neural computer is a feedforward network with binary threshold units and three layers. The first layer is a sensory surface called a "retina" that consists of stimulus cells (S-units). The S-units are connected with the intermediate layer by fixed weights that do not change during the learning process. The elements of the intermediate layer are called associator cells (A-units). Each A-unit has a fixed weighted input of some S-units. In other words, some S-units project their output onto an A-unit. An S-unit may also project its output onto several A-units. The intermediate layer is completely connected with the output layer, the elements of which are called response cells (R-units). The weights between the intermediate layer and the output layer are variable and thus able to learn.

The Perceptron was viewed as a neural computer that can classify a perceived pattern in one of several possible groups. In 1969, Marvin Minsky and Seymour Papert proved that Perceptrons cannot recognize and distinguish the connectivity of patterns, in general. The Perceptron's failure is overcome by more flexible networks with supervised and unsupervised learning algorithms (e.g., Hopfield systems, Chua's cellular neural networks, Kohonen's self-organizing maps).

In the age of globalization, communication networks such as the Internet are a tremendous challenge to AI. From a technical point of view, we need intelligent programs distributed in the nets. There are already more or less intelligent virtual organisms (agents), learning, self-organizing, and adapting to our individual preferences of information, to select our e-mails, to prepare economic transactions, or to defend against attacks of hostile computer viruses, like the immune system of our body. Although the capability to manage the complexity of modern societies depends decisively on progress in AI, we need computational ecologies with distributed AI to support human life and not human-like robots to replace it.

KLAUS MAINZER

See also **Artificial life; Cell assemblies; Game of life; McCulloch–Pitts network; Neural network models; Perceptron**

Further Reading

Lenat, D.B. & Guha, R.V. 1990. *Building Large Knowledge-Based Systems*, Reading, MA: Addison-Wesley

Mainzer, K. 2003. *Thinking in Complexity. The Computational Dynamics of Matter, Mind, and Mankind*, 4th edition, Berlin and New York: Springer

Minsky, M. & Papert, S.A. 1969. *Perceptrons*, Cambridge, MA MIT Press

Minski, M. 1985. *The Society of Mind*, New York: Simon & Schuster

Nilson, N.J. 1982. *Principles of Artificial Intelligence*, Berlin and New York: Springer

Palm, G. (editor). 1984. *Neural Assemblies. An Alternative Approach to Artificial Intelligence*, Berlin and New York Springer

ARTIFICIAL LIFE

The term *artificial life* (AL) was coined in 1987 by Christopher Langton, who organized a workshop by that name in frustration with the lack of a forum for discussing work on the computer simulation of biological systems. In Langton's characterization, AL seeks to "contribute to theoretical biology by locating life-as-we-know-it within the larger picture of life-as-it-could-be" (Langton, 1989). In other words, AL aims to use computer simulation to synthesize alternative life-like systems and, thus, find out which characteristics and principles are essential and which are merely contingent on how life happened to evolve on this planet. While other branches of biology may use simulation to understand specific mechanisms, AL is broader, more abstract, and highly interdisciplinary, in addition to implying certain ideological convictions. Chief among these is the assumption that life is a process, rather than a metaphysical substance or an atomic property of matter, which emerges in a bottom-up fashion from local interactions among suitably arranged populations of individually lifeless components.

Opinions differ on whether such artificial systems may be logically equivalent to their natural counterparts and therefore really alive, or whether they are simply life-like simulacra. The former view is called the strong AL hypothesis, to associate it with a similarly functionalist standpoint known as Strong Artificial intelligence. However, the strong position in AL is considerably more tenable than its AI analog, which fails to distinguish between emergent and explicitly predetermined sources of behavior. Related to this "strong versus weak" argument is the unresolved question of whether life is an absolute category in nature at all or simply a useful way of grouping certain phenomena.

Early History

Attempts to construct living or life-like artifacts from mechanical parts date back at least to the ancient Greeks, and we can presume that many of these experiments were motivated by questions similar to those posed today. Nevertheless, these early systems tended to employ the "if it quacks like a duck it is a duck" principle, and so were only superficially lifelike, rather than in the deeper sense presently hoped for. One of the most ingenious of these early automata was indeed a duck (or at least something that moved, ate, defecated, and quacked like one), built by Jacques de Vaucanson around 1730.

Mary Shelley's *Frankenstein* explores similar issues in a fictional context. Contrary to popular belief, Shelley's monster was apparently not made from human body parts but from raw materials (cadavers are only mentioned with regard to Frankenstein's research). These components were then imbued with the "spark of life" (which Shelley associates with electricity) in order to animate them. Her viewpoint was still partially vitalistic, but there is a link between Shelley and her contemporary Charles Babbage, whose interpretation of intelligence (if not life itself) was more formalized, abstract, and mechanical. The mechanization of the mind continued with George Boole's logical algebra and then the work of Alan Turing and John von Neumann on automating thought processes, which led directly to the invention of the digital computer and the beginnings of artificial intelligence. It was a similar inquiry into the abstract nature of life, as distinct from mind, that prompted von Neumann's investigations into self-replicating machinery and Turing's work on embryogenesis and "unorganized machines" (related to neural networks).

Formal Methods

While complexity theory is concerned with the manner in which complex behavior arises from simple systems, AL is interested in how systems generate continually increasing levels of complexity. The most striking feature of living systems is their ability to self-organize and self-maintain—a property that Humberto Maturana and Francisco Varela have termed "autopoiesis" (Maturana & Varela, 1980). Evolution, embryogenesis, learning, and the development of social organizations are therefore the mechanisms of primary interest to AL researchers.

The key features of AL models are the use of populations of semi-autonomous entities, the coupling of these through simple local interactions (no centralized control and little or no globally accessible information), and the consequent emergence of collective, persistent phenomena that require a higher level of description than that used to describe their substrate. Conventional mathematical notation is not usually appropriate for such distributed and labile systems, and the individual computer programs are often their own best description. There are, however, a number of frequently used abstract structures and formal grammars, including the following:

Cellular automata, in which the populations are arrays of finite state machines and interactions occur between neighboring cells according to simple rules. Under the right conditions, emergent entities (such as the glider in John Conway's Game of Life) arise and persist on the surface of the matrix, interacting with other entities in computationally interesting ways.

Genetic algorithms, in which the populations are genomes in a gene pool and interactions occur between their phenotypes and some form of stressful environment. Natural selection (or sometimes human choice) drives the population to adapt and grow ever fitter, perhaps solving real practical problems in the process.

L-systems, or *Lindenmayer systems*, which provide a grammar for defining the growth of branching (often plant-like) physical structures, as insights into morphology and embryology.

Autonomous agents, which are composite code and data objects, representing mobile physical entities (robots, ants, stock market traders) embedded in a real or simulated environment. They interact locally by sensing their environment and receiving messages from other agents, giving rise to emergent phenomena of many kinds including cooperative social structures, nest-building, and collective problem-solving.

Autocatalytic networks, in which the populations are of simulated enzymes and the interactions are equivalent to catalysis. Such networks are capable of self-generation and a growth in complexity, mimicking the bootstrapping process that presumably gave rise to life on Earth.

Current Status

Like most new fields, AL has undergone cycles of hubris and doubt, innovation and stasis, and differentiation and consolidation. The listing of topics for the latest in the series of workshops started by Langton in 1987 is as broad as ever, although probably the bulk of AL work today (2004) is focused on artificial evolution. Most research concentrates on fine details, while the basic philosophical questions remain largely unanswered. Nevertheless, AL remains one of relatively few fields where one can ask direct questions about one's own existence in a practical way.

STEVE GRAND

See also **Catalytic hypercycle; Cellular automata; Emergence; Game of life; Hierarchies of nonlinear systems; Turing patterns**

Further Reading

Adami, C. 1998. *Introduction to Artificial Life*, New York: Springer
Boden, M.A. (editor). 1996. *The Philosophy of Artificial Life*, Oxford and New York: Oxford University Press
Langton, C.G. (editor). 1989. *Artificial Life*, Redwood City, CA: Addison-Wesley
Levy, S. 1992. *Artificial Life: The Quest for a New Creation*, New York: Pantheon
Maturana, H.R. & Varela, F.J. 1980. *Autopoiesis and Cognition: The Realization of the Living*, Dordrecht and Boston: Reidel

ASSEMBLY OF NEURONS

See **Cell assemblies**

ATMOSPHERIC AND OCEAN SCIENCES

Earliest works on the study of the atmosphere and ocean date back to Aristotle and his student Theophrastus in 350 BC and further progressed through Torricelli's invention of the barometer in 1643, Boyle's law in 1657, and Celsius's invention of the thermometer in 1742 (due to Galileo in 1607). The first rigorous theoretical model for the study of the atmosphere was proposed by

Vilhelm Bjerknes in 1904, following which many scientists began to apply fundamental physics to the atmosphere and ocean. The advent of these theoretical approaches and the invention of efficient communication technologies in the mid-20th century made numerical weather prediction feasible and was in particular encouraged by Lewis Fry Richardson and John von Neumann in 1946, using the differential equations proposed by Bjerknes. Today, advanced numerical modeling and observational techniques exist, which are constantly being developed further in order to understand and study the complex nonlinear dynamics of the atmosphere and ocean.

This overview article summarizes the governing equations used in atmospheric and ocean sciences, features of atmosphere-ocean interaction, and processes for an idealized geometry and structure with reference to a one-dimensional vertical scale (Figure 1), a two-dimensional vertically averaged scale (Figures 3(a) and 4(a)), a two-dimensional zonally averaged meridional scale (Figures 3(b) and 4(b)), and a three-dimensional scale (Figure 2), and regimes of interacting systems (such as El Niño and Southern Oscillation and North Atlantic Oscillation) (Figures 5–7). The entry serves as an introduction to the many nonlinear processes taking place (for example, chaos, turbulence) and provides a few illustrative examples of self-organizing coherent structures of the nonlinear dynamics of the atmosphere and ocean.

Governing Equations

The combined atmosphere and ocean system can be regarded as a huge volume of fluid resting on a rotating oblate spheroid with varying surface topography moving through space, with an interface (which in general is discontinuous) between two fluid masses of differing densities. This coupled atmosphere-ocean system is driven by energy input through solar radiation (see Figure 1), gravity (for example, through interaction with other stellar bodies such as the Sun and Moon, i.e., tides), and inertia. The entire fluid is described by equations for conserved quantities such as momentum, mass (of air, water vapor, water, salt), and energy together with equations of state for air and water (*See* **Fluid dynamics; Navier–Stokes equation**). The movement of large water or air masses in a rotating reference frame adds to the complexity of motions, due to the presence of Coriolis forces, introduced by Coriolis in 1835.

Atmosphere-ocean interactions can be defined as an exchange of momentum, heat, and water (vapor and its partial masses: salts, carbon, oxygen, nitrogen, etc.) between air and water masses. The governing equations in the Euler formulation and a cartesian coordinate system are given by:

(i) The conservation of momentum

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} + 2\Omega \times \boldsymbol{u} = -\frac{1}{\rho}\nabla p - \boldsymbol{g} + \boldsymbol{F}_{\text{ext}} + \boldsymbol{F}_{\text{fric}}, \quad (1)$$

where the second term $2\Omega \times \boldsymbol{u}$ is the term due to the Coriolis force (Ω is the angular velocity of the Earth; $|\Omega| = 7.29 \times 10^{-5}\ \mathrm{s}^{-1}$), and forces due to a pressure gradient ∇p, gravity ($|\boldsymbol{g}| = 9.81\,\mathrm{m\,s^{-2}}$) and external ($\boldsymbol{F}_{\text{ext}}$) as well as frictional ($\boldsymbol{F}_{\text{fric}}$) forces are included. Note that the operator $\mathrm{d}/\mathrm{d}t$ is defined by

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \left(\frac{\partial}{\partial t} + \boldsymbol{v}\cdot\nabla\right)\boldsymbol{v}.$$

(ii) The conservation of mass (or continuity equation)

$$\frac{1}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}t} + \nabla\cdot\boldsymbol{u} = 0. \quad (2)$$

Note that there are alternative formulations such as the Lagrangian and impulse-flux form for these equations, and cartesian coordinate systems can be mapped to different geometries such as spherical coordinates by appropriate transformations.

(iii) The conservation of energy (First Law of Thermodynamics) and Gibbs's equation (Second Law of Thermodynamics)

$$\begin{aligned}\frac{\mathrm{d}Q}{\mathrm{d}t} &= \frac{\mathrm{d}\varepsilon}{\mathrm{d}t} + p\frac{\mathrm{d}\alpha}{\mathrm{d}t},\\ \frac{\mathrm{d}\eta}{\mathrm{d}t} &= \frac{1}{T}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t} + \frac{p}{T}\frac{\mathrm{d}\alpha}{\mathrm{d}t} - \sum\frac{\mu_i}{T}\frac{\mathrm{d}\gamma_i}{\mathrm{d}t},\end{aligned} \quad (3)$$

where Q is the heat supply (sensible, latent, and radiative heat fluxes; see Figure 1), T is the temperature, ε the internal energy and α the specific volume ($\alpha = 1/\rho$), η the entropy, μ_i the chemical potentials, and γ_i the partial masses. The conservation of energy states in brief that the change in heat is balanced by a change in internal energy and mechanical work performed, and Gibbs's equation determines the direction of an irreversible process, relating entropy to a change in internal energy, volume, and partial masses.

(iv) The conservation of partial masses of water and air, that is, salinity for water, where all constituents are represented as salts and water vapor for air, yield equations similar to (2)

$$\frac{1}{\rho_{\mathrm{v}}}\frac{\mathrm{d}\rho_{\mathrm{v}}}{\mathrm{d}t} + \nabla\cdot\boldsymbol{u} = W_v$$

and

$$\frac{\mathrm{d}\rho s}{\mathrm{d}t} + \rho s\nabla\cdot\boldsymbol{u} = W_{\mathrm{s}}, \quad (4)$$

where ρ_{v} is the density of water vapor, s the specific salinity (gram salts per gram water), and W_v, W_s

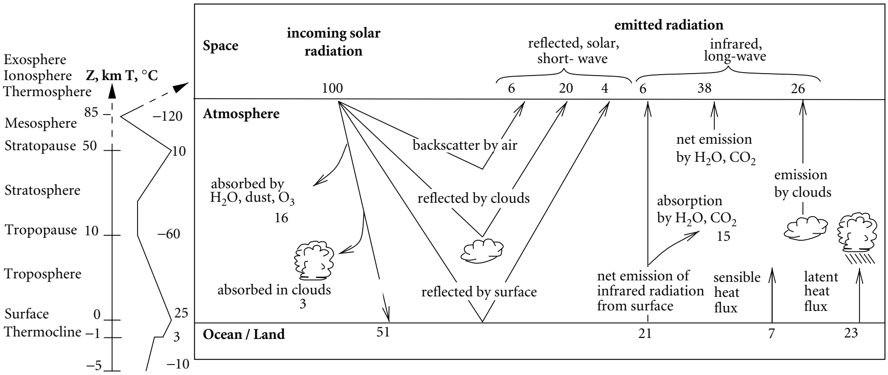
Exosphere
Ionosphere
Thermosphere
Mesosphere
Stratopause
Stratosphere
Tropopause
Troposphere
Surface
Thermocline
Z, km
T, °C
85
50
10
0
–1
–5
–120
10
–60
25
3
–10
Space
incoming solar radiation
100
emitted radiation
reflected, solar, short- wave
infrared, long-wave
6
20
4
6
38
26
Atmosphere
backscatter by air
absorbed by H_2O, dust, O_3
16
reflected by clouds
absorbed in clouds
3
reflected by surface
net emission by H_2O, CO_2
emission by clouds
absorption by H_2O, CO_2
15
net emission of infrared radiation from surface
sensible heat flux
latent heat flux
Ocean / Land
51
21
7
23

Figure 1. Sketch of the vertical structure of the atmosphere–ocean system and radiation balance and processes in the global climate system. Adapted from National Academy of Sciences (1975). Note that lengths are not to scale and and temperatures indicate only global averages.

contain possible source and sink terms as well as the effect of molecular diffusion in terms of the concentration flux density $\boldsymbol{S}$ $(-\nabla\cdot\boldsymbol{S})$ and possible phase changes.

(v) The equation of state for a mixture of salts and gases for air and water, whose constituent concentrations are virtually constant in the atmosphere and ocean is

$$p \approx \rho RT\,(1 + 0.6078\,q)\,, \tag{5}$$

where R is the gas constant for dry air ($R = 287.04\,\mathrm{J\,kg^{-1}\,K^{-1}}$) and $q = \rho_v/\rho$ is the specific humidity. Similarly, the equation of state for near incompressible water is

$$\rho \approx \rho_0[1 - \alpha(T - T_0) + \beta(S - S_0)], \tag{6}$$

where ρ_0, T_0, and S_0 are reference values for density, temperature, and salinity ($\rho_0 = 1028\,\mathrm{kg\,m^{-3}}$, $T_0 = 283\,\mathrm{K}(= 10°\mathrm{C})$, $S_0 = 35‰$), and α and β are the coefficients of thermal expansion and saline contraction ($\alpha = 1.7 \times 10^{-4}\,\mathrm{K^{-1}}$, $\beta = 7.6 \times 10^{-4}$), see Krauss (1973); Cushman-Roisin (1993).

Equations detailed in (i)–(v) form a set of hydrothermodynamic equations for the atmosphere-ocean system to which various approximations and scaling limits can be applied. Among them are the shallow-water equations, primitive equations, the Boussinesq and anelastic approximation, quasigeostrophic, and semigeostrophic equations and variants or mixtures of these. These equations have to be solved with appropriate boundary conditions and conditions at the air-sea interface; for details refer to Krauss (1973), Gill (1982) and Kraus & Businger (1994). For studies of the upper atmosphere, further equations for the geomagnetic field can also be taken into account (Maxwell's equations).

Atmospheric Structure and Circulation

In the vertical dimension, several atmospheric layers can be differentiated (see Figure 1). Figure 2 gives the length and time scales of typical atmospheric processes.

From sea level up to about 2 km is the atmospheric boundary layer, characterized by momentum, heat, moisture, and water transfer between the atmosphere and its underlying surface. Above the boundary layer is the troposphere (Greek, *tropos* meaning *turn, change*) that constitutes most of the total mass of the atmosphere (about 10 km height) and is largely in hydrostatic balance characterized by a decrease in temperature. Above the troposphere and stratosphere, which contains the ozone layer, temperatures rise throughout. The mesosphere, which is bounded by the stratopause (about 50 km height) below and mesopause (about 85 km height) above, is a layer of very thin air where temperatures drop to extreme lows. Above the mesopause, temperatures increase again throughout the thermosphere (from about 85 km to 700 km), the largest layer of the atmosphere, where the ionosphere is located (between about 100 km and 300 km). The ionosphere contains ionized atoms and free electrons and permits the reflection of electromagnetic waves. Above the thermosphere is the exosphere, which is the outermost layer of the atmosphere and the transition region between the atmosphere and outer space, the magnetosphere in particular, where atoms can escape into space beyond the so-called escape velocity and where the Van Allen belt is situated.

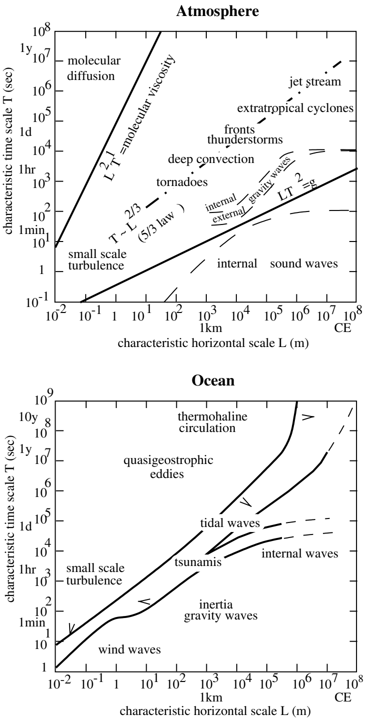

Figure 2. Schematic logarithmic time and horizontal length scales of typical atmospheric and oceanic phenomena. Note that Richardson's $L \propto T^{3/2}$ relation and CE stands for circumference of the Earth. Modified from Lettau (1952), Smagorinsky (1974), and World Meteorological Organization (1975).

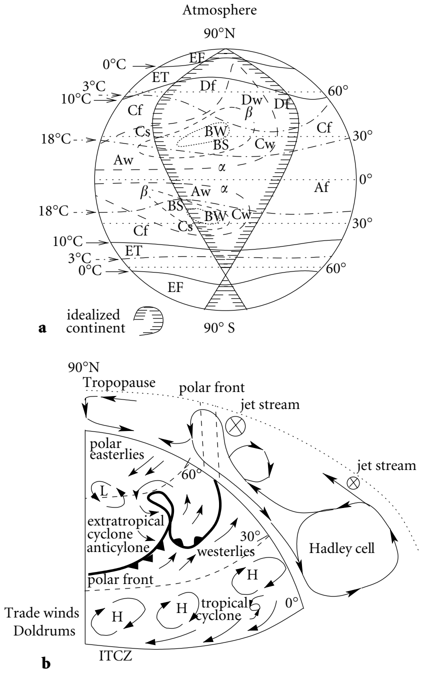

Figure 3. Sketch of the near-surface climate and atmospheric circulation of the Earth with an idealized continent. (a) Averaged isothermals of the coldest month (dashed-dot, −3°C, 18°C) and the warmest month (solid, 0°C, 10°C) and periodical dry season boundaries α and dry climate β. The following climate regions are indicated: wet equatorial climate (Af), tropical wet/dry climate (Aw), desert climate (BW), steppe climate (BS), sinic climate (Cw), Mediterranean climate (Cs), humid subtropical climate (Cf), humid continental climate (Df), continental subarctic climate (Dw), tundra climate (ET) and snow and ice climate (EF). Modified from Köppen (1923). (b) The zonal mean jet streams (primary circulation) and mass overturning (secondary circulation) in a meridional height section, the subtropical highs (H) and subpolar lows (L), polar easterlies, westerlies, polar front, trade winds, and intertropical convergence zone (ITCZ). ▲ denotes a cold front and ● a warm front. Adapted from Palmen (1951), Defant and Defant (1958), and Hantel in Bergmann & Schäfer (2001).

A low (high) in meteorology refers to a system of low (high) pressure, a closed area of minimum (maximum) atmospheric pressure (closed isobars, or contours of constant pressure) on a constant height chart. A low (high) is always associated with (anti)cyclonic circulation, thus also called a cyclone (anticyclone). Anticyclonic means clockwise in the Northern Hemisphere (and counterclockwise in the Southern Hemisphere). Cyclonic means counterclockwise in the Northern Hemisphere (and clockwise in the Southern Hemisphere). At zeroth order, a balance of pressure gradient forces and Coriolis forces, that is, geostrophic balance, occurs, leading to the flow of air along isobars instead of across (in the direction of the pressure gradient). A front is a discontinuous interface or a region of strong gradients between two air masses of differing densities or temperatures, thus encouraging conversion of potential into kinetic energy (examples are polar front, arctic front, cold front, and warm front).

Hurricanes and typhoons (local names for tropical cyclones) transport large amounts of heat from low to mid and high latitudes and develop over oceans. Little is known about the initial stages of their formation, although they are triggered by small low-pressure systems in the Intertropical Convergence Zone (*See* **Hurricanes and tornadoes**). Because of their strong winds, cyclones are particularly active in inducing

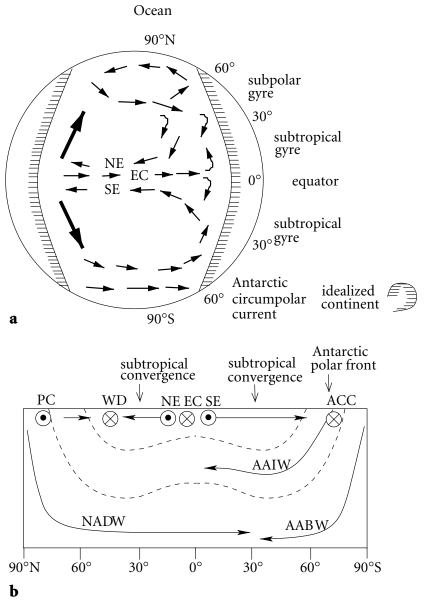

Figure 4. Sketch of the oceanic circulation of an idealized basin (a) the global wind-induced distribution of ocean currents (primary circulation) and (b) the zonal mean thermohaline circulation (secondary circulation) in a meridional depth section showing upper, intermediate, deep and bottom water masses. NE denotes the north equatorial, EC the equatorial and SE the south equatorial current. PC stands for polar current, ACC for Antarctic circumpolar current, WD for west drift, AAIW for Antarctic intermediate waters, NADW for North Atlantic deep water, and AABW for Antarctic bottom water. Adapted from Hasse and Dobson (1986).

upwelling and wind-driven surface water transport. Extratropical cyclones are frontal cyclones of mid to high latitudes (see Figures 2 and 3).

Meteorology and oceanography are concerned with understanding, predicting, and modeling the weather, climate, and oceans due to their fundamental socio-economic and environmental impact. In meteorology, one distinguishes between short (1–3 days) and medium-range (4–10 days) numerical weather prediction models (NWPs) for the atmosphere and general circulation models (GCMs, *See* **General circulation models of the atmosphere**). While NWPs for local and regional weather prediction are usually not coupled to ocean models, GCMs are global three-dimensional complex coupled atmosphere-ocean models (which even include the influence of land masses), used to study global climate change, modeling radiation, photochemistry, transfer of heat, water vapor, momentum, greenhouse gases, clouds, ocean temperatures, and ice boundaries. The atmosphere-ocean interface couples the "fast" processes of the atmosphere with the comparably "slow" processes of the ocean through evaporation, precipitation, and momentum interaction. GCMs are validated using statistical techniques and correlated to the actual climate evolution. Additionally, the application of GCMs to different planetary atmospheres, for example, on Mars and Jupiter, leads to a greater understanding of the planet's history and environment.

The complexity of the dynamics of the atmosphere and ocean is largely due to the intrinsic coupling between these two large masses at the air-sea interface.

Ocean

Ocean circulation is forced by tidal forces (also known to force atmospheric tides), due to gravitational attraction, wind stress, applied shear forces acting on the interface, and external, mainly solar, radiation, penetrating into the sea surface and affecting the heat budget and water mass due to evaporation. Primary sources of tidal forcing, earliest work on which was undertaken by Pierre-Simon Laplace in 1778, are the Moon and the Sun. One discerns between diurnal, semidiurnal, and mixed-type tides.

In the ocean, one distinguishes between two types of ocean currents: surface (wind-driven) and deep circulation (thermohaline circulation). Separating the surface and deep circulation is the thermocline, a small layer of strong gradient of temperature, salinity, and density, acting as an interface between the two types of circulations.

Surface circulation ranging up to 400 m in depth is forced by the prominent westerly winds in the mid-latitudes and trade winds in the tropical regions (see Figures 3 and 4), which are both forced by solar heating and Coriolis forces leading to expansion of water near the equator and decreased density, but increased salinity due to evaporation. An example of the latter is the Gulf Stream in the North Atlantic. The surface wind stress, solar heating, Coriolis forces, and gravity lead to the creation of large gyres in all ocean basins with clockwise (anticyclonic) circulation in the northern hemisphere and counterclockwise circulation in the southern hemisphere. The North Atlantic Gyre, for example, consists of four currents: the north equatorial current, the Gulf Stream, the North Atlantic current, and the Canary current.

Ekman transport, the combination of wind stress and Coriolis forces, leads to a convergence of water masses in the center of such gyres, which increases the sea surface elevation. The layer of Ekman transport can be 100–150 m in depth and also leads to upwelling due to conservation of mass on the western (eastern) coasts for winds from the north (south) in the Northern (Southern) Hemisphere. As a consequence, nutrient-rich

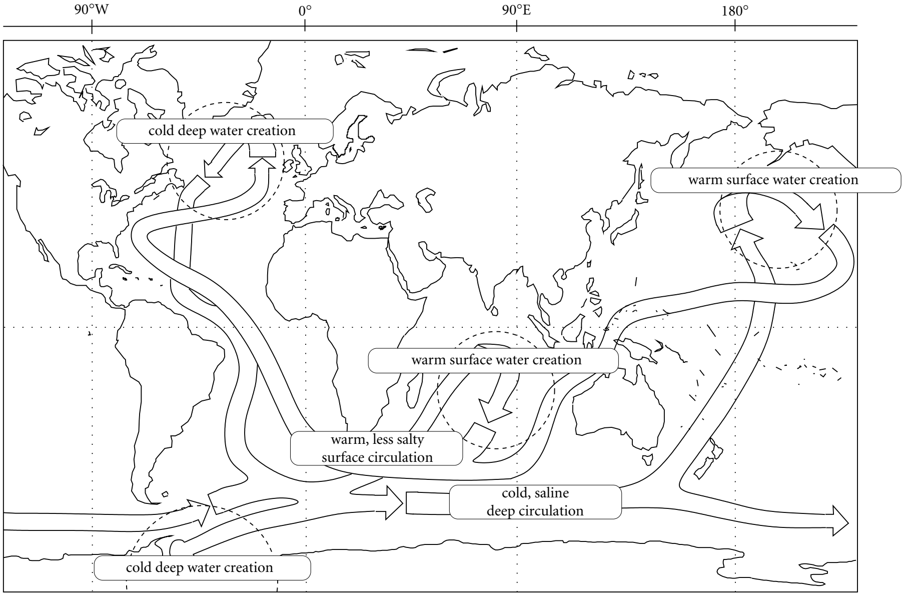
90°W
0°
90°E
180°
cold deep water creation
warm surface water creation
warm surface water creation
warm, less salty surface circulation
cold, saline deep circulation
cold deep water creation

Figure 5. Sketch of the global conveyor belt through all oceans, showing the cold saline deep circulation, the warm, less salty surface circulation, and the primary regions of their creation. Note that this circulation is only characteristic of the actual global circulation. Adapted from Broecker (1987).

deep water is brought to the surface. With the opposite wind direction, Ekman transport acts to induce downwelling.

Another important combination of forces is the balance of Coriolis forces and gravity (pressure gradient forces), which is called geostrophic balance, leading to the movement of mass along isobars instead of across (geostrophic current), similar to the atmosphere. The boundary currents along the eastern and western coastlines are the major geostrophic currents in a gyre. The western side of the gyre is stronger than the eastern due to the Earth's rotation, called western intensification.

Deep circulation makes up 90% of the total water mass and is driven by density forces and gravity, which in turn is a function of temperature and salinity. High-density deep water originates in the case of extreme cooling of the sea surface in the polar regions, sinking to large depths as a density current, a strongly nonlinear phenomenon. When the warm Gulf Stream waters, which have increased salinity due to excessive evaporation in the tropics, move north due to the North Atlantic Gyre, they are cooled by Arctic winds from the north and sink to great depths forming the high-density Atlantic deep waters (see Figure 5). The downward transport of water is balanced by upward transport in low- and mid-latitude regions.

The most prominent example of the interaction between atmospheric and ocean dynamics is the global conveyor belt, which links the surface (wind-driven) and deep (thermohaline) circulation to the atmospheric circulation. The global conveyor belt is a global circulatory system of distinguishable and recognizable water masses traversing all oceans (see Figure 5). The water masses of this global conveyor belt transport heat and moisture, contributing to the climate globally. In Earth's history, the global conveyor belt has experienced flow reversals and perturbations leading to changes in the global circulatory system. The rather recent anthropogenic impact on climate and oceans through greenhouse gas emissions has the potential to create instability in this large-scale dynamical system, which could alter Earth's climate and have devastating environmental and agricultural effects.

ENSO and NAO

Another example of atmosphere-ocean coupling is the combination of the El Niño and Southern Oscillation (ENSO). The El Niño ocean current (and associated

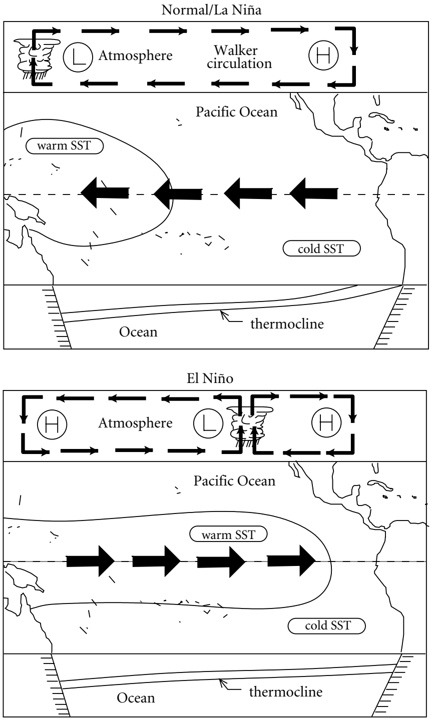
Normal/La Niña
L
Atmosphere
Walker circulation
H
Pacific Ocean
warm SST
cold SST
Ocean
thermocline
El Niño
H
Atmosphere
L
H
Pacific Ocean
warm SST
cold SST
Ocean
thermocline

Figure 6. Sketch of the El Niño in the tropical Pacific, showing a reversal in (trade) wind direction from easterlies to westerlies during an El Niño period bringing warmer water (warm corresponds to a positive sea-surface temperature [SST]) close to the South American coast, displacing the equatorial thermocline downwards. Note the change in atmospheric tropical convection and associated heavy rainfall. After McPhaden, NOAA/TAO (2002) and Holton (1992).

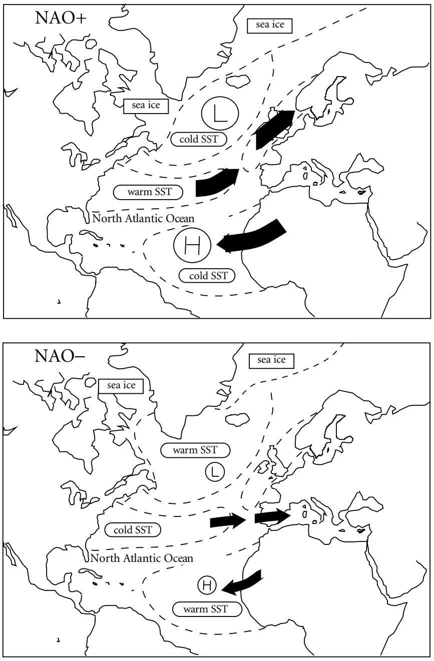
NAO+
sea ice
sea ice
L
cold SST
warm SST
North Atlantic Ocean
H
cold SST
NAO−
sea ice
sea ice
warm SST
L
cold SST
North Atlantic Ocean
H
warm SST

Figure 7. Sketch of the North Atlantic Oscillation (NAO) during the northern hemisphere winter season. Positive NAO (NAO+) showing an above-usual strong subtropical high-pressure center and subpolar low, resulting in increased wind strengths and storms crossing the Atlantic towards northern Europe. NAO+ is associated with a warm wet winter in Europe and cold dry winter in North America. Central America experiences mild wet winter conditions. Negative NAO (NAO−) shows a weaker subtropical high and subpolar low, resulting in lower wind speeds and weaker storms crossing the Atlantic toward southern Europe and receded sea ice masses around Greenland. NAO− is associated with cold weather in northern Europe and moist air in the Mediterranean. Central America experiences colder climates and more snow. Adapted from Wanner (2000).

wind and rain change) is named from the Spanish for Christ Child, due to its annual occurrence off the South American coast around Christmas, and may also be sensitive to anthropogenic influence (see Figure 6). The Southern Oscillation occurs as a 2–5-year periodic reversal in the east-west pressure gradient associated with the present equatorial wind circulation, called Walker circulation, across the Pacific leading to a reversal in wind direction and changes in temperature and precipitation. The easterly wind in the West Pacific becomes a westerly. As a consequence, the strong trade winds are weakened, affecting climate globally (e.g., crop failures in Australia, flooding in the USA, and the monsoon in India). The Southern Oscillation in turn leads to large-scale oceanic fluctuations in the circulation of the Pacific Ocean and sea-surface temperatures, which is called El Niño. The interannual variability, though, is not yet fully understood; consideration of a wider range of tropical and extratropical influences is needed. A counterpart to the ENSO in the Pacific is the North Atlantic Oscillation (NAO), which is essentially an oscillation in the pressure difference across the North Atlantic and is described further in Figure 7.

Monsoons

The monsoons (derived from Arabic, *mauism*, meaning season or shift in wind) are seasonally reversing winds and one of the most pertinent features of the global atmospheric circulation. The best-known examples are the monsoons over the Indian Ocean and, to some extent, the western Pacific Ocean (tropical region of Australia), the western coast of Africa, and the Carribbean. Monsoons are characteristic for wet summer and dry winter seasons, associated with strong winds and cyclone formation. They occur due to differing thermal characteristics of the land and sea surfaces. Land, having a much smaller heat capacity than the ocean, emits heat from solar radiation more easily, leading to upward heat (cumulus) convection. In the summer season, this leads to a pressure gradient and thus wind from the land to the ocean in the upper layers of the atmosphere and subsequent conserving flow of moisture-rich air from the sea back inland at lower levels. This leads to monsoonal rains, increased latent heat release, and intensified monsoon circulation. During the monsoons of the winter season, the opposite of the summer season monsoon takes place, although less pronounced, since the thermal gradient between the land and sea is reversed. The winter monsoons thus lead to precipitation over the sea and cool dry land surfaces.

ANDREAS A. AIGNER AND KLAUS FRAEDRICH

See also **Fluid dynamics; General circulation models of the atmosphere; Hurricanes and tornadoes; Lorenz equations; Navier–Stokes equation**

Further Reading

Apel, J. 1989. *Principles of Ocean Physics*, London: Academic Press

Barry, R.G., Chorley, R.J. & Chase, T. 2003. *Atmosphere, Weather and Climate*, 8th edition, London and New York: Routledge

Bergmann, K., Schaefer C. & von Raith, W. 2001. *Lehrbuch der Experimentalphysik, Band 7, Erde und Planeten*, Berlin: de Gruyter

Cushman-Roisin, B. 1993. *Introduction to Geophysical Fluid Dynamics*, Englewood Cliffs, NJ: Prentice–Hall

Defant, A. & Defant, Fr. 1958. *Physikalische Dynamik der Atmosphäre*, Frankfurt: Akademische Verlagsgesellschaft

Gill, A. 1982. *Atmosphere–Ocean Dynamics*, New York: Academic Press

Hasse, L. & Dobson, F. 1986. *Introductory Physics of the Atmosphere and Ocean*, Dordrecht and Boston: Reidel

Holton, J.R. 1992. *An Introduction to Dynamic Meteorology*, 3rd edition, New York: Academic Press

Kraus, E.B. & Businger, J.A. 1994. *Atmosphere–Ocean Interaction*, New York: Oxford University Press, and Oxford: Clarendon Press

Krauss, W. 1973. *Dynamics of the Homogeneous and Quasihomogeneous Ocean*, vol I, Berlin: Bornträger

LeBlond, P.H. & Mysak, L.A. 1978. *Waves in the Ocean*, Amsterdan: Elsevier

Lindzen, R.S. 1990. *Dynamics in Atmospheric Physics*, Cambridge and New York: Cambridge University Press

Pedlosky, J. 1986. *Geophysical Fluid Dynamics*, New York: Springer

Philander, S.G. 1990. *El Niño, La Niña, and the Southern Oscillation*, New York: Academic Press

ATTRACTOR NEURAL NETWORKS

Neural networks with feedback can have complex dynamics; their outputs are not related in a simple way to their inputs. Nevertheless, they can perform computations by converging to attractors of their dynamics. Here, we analyze how this is done for a simple example problem: associative memory, following the treatment by Hopfield (1984) (see also Hertz, et al., 1991, Chapters 2 and 3).

Let us assume that input data are fed into the network by setting the initial values of the units that make it up (or a subset of them). The network dynamics then lead to successive changes in these values. Eventually, the network will settle down into an attractor, after which the values of the units (or some subset of them) give the output of the computation. The associative memory problem can be described in the following way: there is a set of p patterns to be stored. Given, as input, a pattern that is a corrupted version of one of these, the attractor should be a fixed point as close as possible to the corresponding uncorrupted pattern.

We focus on networks described by systems of differential equations such as

$$\tau_i \frac{\mathrm{d}u_i}{\mathrm{d}t} + u_i(t) = \sum_{j \neq i} w_{ij} g[u_j(t)]. \qquad (1)$$

Here, $u_i(t)$ is the net input to unit i at time t and $g(\,)$ is a sigmoidal activation function ($g' > 0$), so that $V_i = g(u_i)$ is the value (output) of unit i. The connection weight to unit i from unit j is denoted w_{ij}, and τ_i is the relaxation time. We can also consider discrete-time systems governed by

$$V_i(t+1) = g\left[\sum_j w_{ij} V_j(t)\right]. \qquad (2)$$

Here, it is understood that all units are updated simultaneously. In either case, the "program" of such a network is its connection weights w_{ij}.

In general, three kinds of attractors are possible: fixed point, limit cycle, and strange attractor. There are conditions under which the attractors will always be fixed points. For nets described by the continuous dynamics of Equation (1), a sufficient (but not necessary) condition is that the connection weights be symmetric: $w_{ij} = w_{ji}$. General results about the stability of recurrent nets were proved by Cohen & Grossberg (1983). They showed, for dynamics (1), that there is a Lyapunov function, that is, a function of the state variables u_i, which always decreases under the dynamics, except for special values of the u_i at which it does not change. These values are fixed points. For values of the u_i close to such a point, the system will evolve either toward it (an attractor) or away from it (a repellor). For almost all starting states, the dynamics will end at one of the attractor's fixed points. Furthermore, these are the only attractors.

We treat the case $g(u) = \tanh(\beta u)$ and consider the *ansatz*

$$w_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu. \qquad (3)$$

That is, for each pattern, there is a contribution to the connection weight proportional to the product of sending (ξ_j^μ) and receiving (ξ_i^μ) unit activities when the network is in a stationary state $V_i = \xi_i^\mu$. This is just the form of synaptic strength proposed by Hebb (1949) as the basis of animal memory, so this *ansatz* is sometimes called a Hebbian storage prescription.

To see how well the network performs this computation, we examine the fixed points of (1) or (2), which solve

$$V_i = \tanh\left(\beta \sum_j w_{ij} V_j\right). \qquad (4)$$

The quality of retrieval of a particular stored pattern ξ_i^μ is measured by the quantity $m_\mu = N^{-1}\sum_i \xi_i^\mu V_i$. Using (4), with the weight formula (3), we look for solutions in which the configuration of the network is correlated with only one of the stored patterns, that is, just one of the m_μ's is not zero. If the number of stored patterns $p \ll N$, we find a simple equation for m_μ:

$$m_\mu = \tanh(\beta m_\mu). \qquad (5)$$

This equation has nontrivial solutions whenever the gain $\beta > 1$ and for β large, $m_\mu \to 1$, indicating perfect retrieval. If the gain is high enough, there are other attractors in addition to the ones we have tried to program into the network with the choice (3), but by keeping the gain between 1 and 2.17 we can limit the attractor set to the desired states.

When p is of the same order as N, the analysis is more involved. We define a parameter $\alpha = p/N$. For small α, the overlaps m_μ between the stored patterns and the fixed points are less than, but still close to, 1. However, there is a critical value of α, $\alpha_c(\beta)$, above which there are no longer fixed points close to the patterns to be stored and the memory breaks down catastrophically. One finds $\alpha_c(1) = 0$ and, in the limit $\beta \to \infty$, $\alpha_c(\beta) \to 0.14$.

Thus, attractor computation works in this system over a wide range of the model parameters α and β. It can be shown to be robust with respect to many other variations, including dilution (random removal of connections), asymmetry (making some of the $w_{ij} \neq w_{ji}$), and quantization or clipping of the weight values. Its breakdown at the boundary $\alpha_c(\beta)$ is a collective effect like a phase transition in a physical system.

The weight formula (3) was only an educated guess. It is possible to obtain better weights, which reduce the crosstalk and increase α_c, by employing systematic learning algorithms.

It is also possible to extend the above-described model to store pattern sequences by including suitable delays in the discrete-time dynamics (2).

It appears that attractor networks play a role in computations in the brain. One example of current interest is working memory: some neurons in the prefrontal cortex that are selectively sensitive to a particular visual stimulus exhibit continuing activity after the stimulus is turned off, even though the animal sees other stimuli. Thus, they seem to be involved in the temporary storage of visual patterns. Computational network models based on the simple concepts described above (Renart et al., 2001) are able to reproduce the main features seen in recordings from these neurons.

JOHN HERTZ

See also **Cellular nonlinear networks; McCulloch–Pitts network; Neural network models**

Further Reading

Amit, D.J. 1989. *Modeling Brain Function*, Cambridge and New York: Cambridge University Press

Cohen, M. & Grossberg, S. 1983. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man and Cybernetics*, 13: 815–826 etc.

Hebb, D.O. 1949. *The Organization of Behavior*, New York: Wiley

Hertz, J.A., Krogh, A.S., & Palmer, R.G. 1991. *Introduction to the Theory of Neural Computation*, Redwood City, CA: Addison-Wesley

Hopfield, J.J. 1984. Neurons with graded responses have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences USA*, 79: 3088–3092

Renart, A., Moreno, R., de la Rocha, J., Parga, N. & Rolls, E.T. 2001. A model of the IT-PF network in object working memory which includes balanced persistent activity and tuned inhibition. *Neurocomputing*, 38–40: 1525–1531

ATTRACTORS

A wide variety of problems arising in physics, chemistry, and biology can be recast within the framework of dynamical systems. A dynamical system is made up of two parts: the phase space, which consists of all possible configurations of the physical system, and the "dynamics," a rule describing how the state of the system changes over time. The fundamental insight of the theory is that some problems, which initially appear extremely complicated, can be greatly simplified if we are prepared to concentrate on their long-term behavior, that is, what happens eventually.

This idea finds mathematical expression in the concept of an attractor.

The simplest possibility is that the system settles down to a constant state (e.g., a pendulum damped by air resistance will end up hanging vertically downward). In the phase space, this corresponds to an attractor that is a single "fixed point" for the dynamics. If the system settles down to a repeated oscillation, then this corresponds to a "periodic orbit," a closed curve in the phase space. For two coupled ordinary differential equations (ODEs), it is a consequence of the Poincaré–Bendixson Theorem that these fixed points and periodic orbits are essentially the only two kinds of attractors that are possible (see Hirsch & Smale, 1974 for a more exact statement). In higher dimensions, it is possible for the limiting behavior to be quasi-periodic with k different frequencies, corresponding to a k-torus in the phase space (cf. Landau's picture of turbulence as in Landau & Lifschitz, 1987).

However, with three or more coupled ODEs (or in one-dimensional maps), the attractor can be an extremely complicated object. The famous "Lorenz attractor" was perhaps the first explicit example of an attractor that is not just a fixed point or (quasi) periodic orbit. Edward Lorenz highlighted this in the title of his 1963 paper, "Deterministic Nonperiodic Flow." The phrase "strange attractor" was coined by Ruelle & Takens (1971) for such complicated attracting sets. These attractors, and the chaotic dynamics associated with them, have been the focus of much attention, particularly in relation with the theory of turbulence (the subject of Ruelle & Takens' paper; see also Ruelle, 1989). There is no fixed definition of a "strange attractor"; some authors use the phrase as a signature of chaotic dynamics, while others use it to denote a fractal attractor (e.g., Grebogi et al. (1984) discuss "strange nonchaotic attractors").

Over the years, various authors have given precise (but different) definitions of an attractor: Milnor (1985) discusses many of these (and proposes a new one of his own). Most definitions require that an attractor attract a "large set of initial conditions and satisfy some kind of minimality property" (without this, the whole phase space could be called an attractor).

We refer to the set of all those points in the phase space whose trajectories are attracted to some set A as the basin of attraction of A and write this $B(A)$. There are essentially two choices of what it means to attract a large set of initial conditions: the more common one is that $B(A)$ contains an open neighborhood of A, while Milnor (1985) suggested that a more realistic requirement is that $B(A)$ has positive Lebesgue measure.

Exactly what type of minimality assumption we require depends on what we want our attractor to say about the dynamics. At the very least, there should be no smaller (closed) set with the same basin of attraction: this excludes any "unnecessary" points from the attractor. A consequence of this minimality property is that the attractor is invariant: if A is the attractor of a map f, this means that $f(A) = A$ (there is, of course, a similar property for the attractor of a flow). In particular, this means that it is possible to talk about the "dynamics on the attractor."

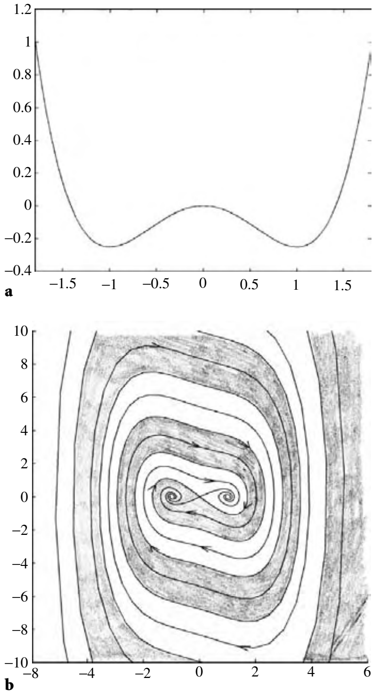

Figure 1. (a) A symmetric double-well potential. (b) Phase portrait of a particle moving in the potential of (a) with friction.

If we want one attractor to describe the possible asymptotic dynamics of *every* initial condition, then there is no need to impose any further minimality assumption. Figure 1(b) shows the phase portrait for a particle moving with friction in the symmetric double-well potential of Figure 1(a); the basin of attraction of the fixed point corresponding to the bottom of the left-hand well is shaded. (The equations of motion are $\dot{x} = y$ and $\dot{y} = -\frac{y}{2} + x - x^3$.) We could say that the attractor consists of the three points $\{(-1, 0), (1, 0), (0, 0)\}$, but this discards much of the information contained in the phase portrait. This motivates the further requirement that an attractor be "indecomposable": it should not be possible to split it into two disjoint invariant subsets. (Some definitions require there to be a dense orbit in the attractor: essentially, this means that one trajectory "covers" the entire attractor, so in particular the attractor cannot be split into two pieces.) This gives us two possible attractors: $(-1, 0)$ and $(1, 0)$ (the origin does not attract a neighborhood of itself, nor any set of positive measure).

In this example, the boundary between the basins of attraction of the two competing attractors is a smooth curve. However, in many examples this boundary is a fractal set. This was first noticed by McDonald et al. (1985), who observed that near a fractal boundary, it is harder to predict the asymptotic behavior of imprecisely known initial conditions. An extreme version of this occurs with the phenomenon of "riddled basins," first observed by Alexander et al. (1992): arbitrarily close to a point attracted to one attractor; there can be a point attracted to another. In this case, an arbitrarily small change in the initial condition can lead to completely different asymptotic behavior. (In addition to treating some analytically tractable examples, Alexander et al. (1992) give an impressive array of pictures from their numerical simulations.)

Attractors can also be meaningfully defined for the infinite-dimensional dynamical systems arising from partial and functional differential equations (e.g., Hale, 1988; Robinson, 2001; Temam 1988/1996), and for random and nonautonomous systems (Crauel et al. (1997) adopt an approach that includes both these cases).

James C. Robinson

See also **Chaos vs. turbulence; Dynamical systems; Fractals; Phase space; Turbulence**

Further Reading

Alexander, J.C., Yorke, J.A., You, Z. & Kan, I. 1992. Riddled basins. *International Journal of Bifurcation and Chaos*, 2: 795–813

Crauel, H., Debussche, A. & Flandoli, F. 1997. Random attractors. *Journal of Dynamics and Differential Equations*, 9: 307–341

Grebogi, C., Ott, E., Pelikan, S. & Yorke, J.A. 1984. Strange attractors that are not chaotic. *Physica* D, 13: 261–268

Hale, J.K. 1988. *Asymptotic Behavior of Dissipative Systems*, Providence, RI: American Mathematical Society

Hirsch, M.W. & Smale, S. 1974. *Differential Equations, Dynamical Systems and Linear Algebra*. New York: Academic Press

Landau, L.D. & Lifschitz, E.M. 1987. *Fluid Mechanics*, 2nd edition, Oxford: Pergamon Press

Lorenz, E.N. 1963. Determininstic non-periodic flow. *Journal of Atmospheric Science*, 20: 130–141

Milnor, J. 1985. On the concept of attractor. *Communications in Mathematical Physics*, 99: 177–195

Robinson, J.C. 2001. *Infinite-Dimensional Dynamical Systems*, Cambridge and New York: Cambridge University Press

Ruelle, D. 1989. *Chaotic Evolution and Strange Attractors*, Cambridge and New York: Cambridge University Press

Ruelle, D. & Takens, F. 1971. On the nature of turbulence. *Communications in Mathematical Physics*, 20: 167–192

Temam, R. 1996. *Infinite-dimensional Dynamical Systems in Mechanics and Physics*, 2nd edition, Berlin and New York: Springer

AUBRY–MATHER THEORY

Named after Serge Aubry and John Mather, who independently shaped the seminal ideas, the Aubry–Mather theory addresses one of the central problems of modern dynamics: the characterization, of nonintegrable Hamiltonian time evolution beyond the realm of perturbation theory. In general terms, when a Hamiltonian system is near-integrable, perturbation theory provides a rigorous generic description of the invariant sets of motion (closed sets containing trajectories) as smooth surfaces (KAM tori), each one parametrized by the rotation number ω of the angle variable (angle-action coordinates): all the trajectories born and living inside the invariant torus share this common value of ω. An invariant set has an associated natural invariant measure, which describes the measure-theoretical (or statistical) properties of the trajectories inside the invariant set. The invariant measure on a torus is a continuous measure, so that the distribution function of the angle variable is continuous. In this near-integrable regime of the dynamics, perturbative schemes converge adequately and future evolution is—to a desired arbitrary degree—predictable for arbitrary initial conditions on each torus.

Far from integrable Hamiltonian dynamics, what is the fate of these invariant natural measures, or invariant sets of motion, beyond the borders of validity of perturbation theory? The answer is that each torus breaks down and its remaining pieces form an invariant fractal set, called by Percival a cantorus (or Aubry–Mather set) characterized by the rotation number value common to all trajectories in the cantorus. The statistical properties of the trajectories on the invariant cantorus are now described by a purely discrete measure or Cantor distribution function (see Figure 1).

Basic Theorems

The formal setting of Aubry–Mather theory for the transition from regular motion on invariant tori to orbits on hierarchically structured nowhere dense cantori is the class of maps of a cylindrical surface $\mathcal{C} = S^1 \times \mathcal{R}$ (cylindrical coordinates (u, p)) (see Figure 2) onto itself,

$$f : \mathcal{C} \to \mathcal{C}, \qquad (1)$$

characterized by preservation of areas (symplectic) and the "twist" property, meaning that the torsion produced by an iteration of the map on a vertical segment of the cylindrical surface converts it into a part of the graph

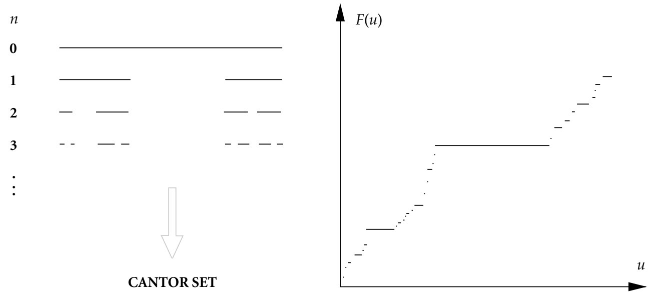

Figure 1. Left: Construction of a Cantor set from the unit real interval (or circle S^1) as a limiting process. At each step n, a whole piece is cut out from each remaining full segment. Right: The distribution function $F(u)$ of the projection onto the angular component u_n of a cantorus orbit. $F(u)$ is the limiting proportion of the values of $u_n < u$, $-\infty < n < +\infty$.

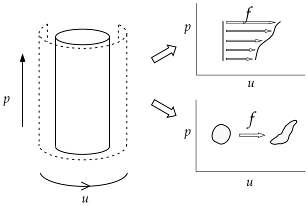

Figure 2. Schematic illustration on the unfolded cylindrical surface $\mathcal{C} = S^1 \times \mathcal{R}$ of the twist (upper right) and area-preserving (lower right) properties of the map $F : \mathcal{C} \to \mathcal{C}$. In the upper right area the curve is a single-valued function $\varphi(u)$ of the angular variable.

of some function $\varphi(u)$ of the angular coordinate. More explicitly, if we denote by $(u', p') = f(u = 0, p)$ the image of the vertical segment, then u' is a monotone function of p, so that (u', p') is the graph of a single-valued function.

An area-preserving twist map has associated an action-generating function related to the map via a variational (extremal action) principle:

- An action-generating function, $H(x, x')$, of a twist map is a two-variable function that is strictly convex:

$$\frac{\partial^2 H}{\partial u \partial u'} \leq K < 0. \tag{2}$$

- If u_0 is a critical point of $\mathcal{L}(x) = H(u_{-1}, x) + H(x, u_1)$, then a certain sequence (u_{-1}, p_{-1}), (u_0, p_0), (u_1, p_1) is a segment of a cylinder orbit of f.

(Given a sequence $\{u_j\}_{j=0}^{n}$ with fixed ends $(u_0 = a, u_n = b)$, the associated action functional $\mathcal{L}$ is the sum $\sum_{j=0}^{n-1} H(u_j, u_{j+1})$.

A cylinder orbit is called ordered when it projects onto an angular sequence ordered in the same way as a uniform rotation of angle ω. An invariant set is a minimal invariant set if it does not include proper invariant subsets and is called ordered if it contains only ordered orbits. The proper definition of an Aubry–Mather set is a minimal invariant ordered set that projects one-to-one on a nowhere dense Cantor set of the circle S^1.

The following points comprise the main core (Golé, 2001; Katok & Hasselblatt, 1995) of the Aubry–Mather theory:

- For each rational value $\omega = p/q$ of the rotation number, there exist (Poincaré–Birkhoff theorem) at least two ordered periodic orbits (Birkhoff periodic orbits of type (p, q)), which are obtained by, respectively, minimizing and maximizing the action over the appropriate set of angular sequences.

  In general, periodic orbits of rational rotation number $\omega = p/q$ do not form an invariant circle, in which case there are nonperiodic orbits approaching two different periodic orbits as $n \to -\infty$ and as $n \to +\infty$, called heteroclinic orbits (or homoclinic in some contexts). These orbits connect two Birkhoff periodic orbits through a minimal action path. Usually, the number of map iterations needed for such action-minimizing orbits to pass over the action barriers is exponentially small.
- For each irrational value of ω, there exists either an invariant torus or an Aubry–Mather set. There are also homoclinic trajectories connecting orbits on the Aubry–Mather set.

The hierarchical structure of gaps that break up the torus has its origins in the path-dependent action barriers. Note also that heteroclinics to nearby periodic orbits (of rational rotation number) pass over nearly the same action barriers, leading to a somewhat metaphorical view of nearby resonances biting the tori and leaving gaps. Certainly, the action barriers fractalize the invariant measure according to the proximity of the irrational rotation number ω to rationals.

The explanatory power of the extremal action principle in the fractalization of invariant sets of motion suggests one of the immediate physical applications of this theory. Indeed, Aubry's work was originally motivated by the equilibrium problem of a discrete field (u_n) $n \in \mathcal{Z}$, under some energy functional, whose extremalization defines equilibrium field configurations. Under some conditions on the energy-generating function $H(u, u')$, both are equivalent mathematical physics problems, and a one-to-one correspondence between orbits (u_n, p_n) and equilibrium field configurations (u_n) does exist (Aubry, 1985).

Application to the Generalized Frenkel–Kontorova Model

From this perspective, the Aubry–Mather theory gives rigorous variational answers in the description of equilibrium discrete nonlinear fields such as the generalized Frenkel–Kontorova (FK) model with convex interactions. Although the terminology changes, every aspect of cylinder dynamics has a counterpart in the equilibrium problem of this interacting nonlinear many-body model.

- Commensurate (periodic) field configurations correspond to Birkhoff periodic orbits, and as such they are connected by the field configurations associated to heteroclinics, which are here called discrete (sine-Gordon) solitons or elementary discommensurations.
- Incommensurate (quasiperiodic) field configurations can correspond either to tori trajectories or to Aubry–Mather trajectories. The macroscopic physical properties (formally represented by averages on the invariant measure) of the field configuration experience drastic changes when passing from one case (tori) to the other (Aubry–Mather sets). This transition (called breaking of analiticity by Aubry) has been characterized as a critical phenomenon using renormalization group methods by (MacKay, 1993).

The Aubry–Mather theory puts on a firm basis what is known as discommensuration theory, which is the description of a generic incommensurate or (higher-order) commensurate field configuration as an array of discrete field solitons, strongly interacting when tori subsist, but almost noninteracting and deeply pinned when only Cantor invariant measures remain. Aubry's work provided a satisfactory understanding of the complexity of the phase diagrams and the singular character of the equations of state of the generalized FK model (Griffiths, 1990).

Luis Mario Floría

See also **Commensurate–incommensurate transition; Frenkel–Kontorova model; Hamiltonian systems; Kolmogorov–Arnol'd–Moser theorem; Phase transitions; Standard map; Symplectic maps**

Further Reading

Aubry, S. 1985. *Structures incommensurables et brisure de la symmetrie de translation I* [Incommensurate structures and the breaking of traslation symmetry I]. In *Structures et Instabilités*, edited by C. Godréche, Les Ulis: Editions de Physique, pp. 73–194

Golé, C. 2001. *Symplectic Twist Maps. Global Variational Techniques*, Singapore: World Scientific

Griffiths, R.B. 1990. Frenkel–Kontorova models of commensurate-incommensurate phase transitions. In *Fundamental Problems in Statistical Mechanics VII*, edited by H. van Beijeren, Amsterdam: North-Holland, pp. 69–110

Katok, A. & Hasselblatt, B. 1995. *Introduction to the Modern Theory of Dynamical Systems*. Cambridge and New York: Cambridge University Press

MacKay, R.S. 1993. *Renormalisation in Area-preserving Maps*, Singapore: World Scientific

AUTO-BÄCKLUND TRANSFORMATION

See **Bäcklund transformations**

AUTOCATALYTIC SYSTEM

See **Reaction-diffusion systems**

AUTOCORRELATION FUNCTION

See **Coherence phenomena**

AUTONOMOUS SYSTEM

See **Phase space**

AUTO-OSCILLATIONS

See **Phase plane**

AVALANCHE BREAKDOWN

Charge transport in condensed matter is simply described by the current density j as a function of the local electric field E. For bulk materials, the current density per unit area is given by $j(E) = -env$, where $e > 0$ is the electron charge, n is the conduction electron density per unit volume, and v is the drift velocity. In the simplest case, v is a linear function of the field: $v = -\mu E$, with mobility μ.

Thus, the conductivity $\sigma = j/E = en\mu$ is proportional to the number of conduction electrons. In metals, n is given by the number of valence electrons, which is temperature independent. In semiconductors, however, the concentration of electrons in the conduction band varies greatly and is determined by generation-recombination (GR) processes that induce transitions between valence band, conduction band, and impurity levels (donors and acceptors). Charge carrier concentration depends not only upon temperature but also upon the electric field, which explains why the conductivity can change over many orders of magnitude. A GR process that depends particularly strongly on the field is impact ionization, the inverse of the Auger effect. It is a process in which a charge carrier with high kinetic energy collides with a second charge carrier, transferring its kinetic energy to the latter, which is thereby lifted to a higher energy level. The kinetic energy is increased by the local electric field, which heats up the carriers. As a certain minimum energy is necessary to overcome the difference in the energy levels of the second carrier, the impact ionization probability depends in a threshold-like manner on the applied voltage.

Impact ionization processes may be classified as band-band processes or band-trap processes depending on whether the second carrier is initially in the valence band and makes a transition from the valence band to the conduction band, or whether it is initially at a localized level (impurity, donor, acceptor), and makes a transition to a band state. Further, impact ionization processes are classified as electron or hole processes according to whether the ionizing hot carrier is a conduction band electron or a hole in the valence band.

Schematically, impact ionization may be written as one of the following reaction equations, in analogy with chemical kinetics:

$$\mathrm{e} \longrightarrow 2\mathrm{e} + \mathrm{h}, \tag{1}$$

$$\mathrm{e} + \mathrm{e_t} \longrightarrow 2\mathrm{e} + \mathrm{h_t}, \tag{2}$$

$$\mathrm{h} \longrightarrow 2\mathrm{h} + \mathrm{e}, \tag{3}$$

$$\mathrm{h} + \mathrm{h_t} \longrightarrow 2\mathrm{h} + \mathrm{e_t}, \tag{4}$$

where e and h denote band electrons and holes, respectively, and $\mathrm{e_t}$ and $\mathrm{h_t}$ stand for electrons and holes trapped at impurities (donors, acceptors, or deep levels). The result of the process is carrier multiplication (avalanching), which may induce electrical instabilities at sufficiently high electric fields. Impact ionization from shallow donors or acceptors is responsible for impurity breakdown at low temperatures. Being an autocatalytic process (i.e., each carrier ionizes secondary carriers that might, in turn, impact ionize other carriers), it induces a nonequilibrium phase transition between a low- and high-conductivity state and may lead to a variety of spatiotemporal instabilities, including current filamentation, self-sustained oscillations, and chaos. The conductivity saturates when all impurities are ionized. Band-to-band impact ionization eventually induces avalanche breakdown, limiting the bias voltage that can be safely applied to a device. The conductivity increases much more strongly than during impurity breakdown because of the large number of valence band electrons available for ionization.

Impurity impact-ionization breakdown at helium temperatures (ca. 5 K) in p-Ge, n-GaAs, and other semiconductor materials has been thoroughly studied both experimentally and theoretically as a model system for nonlinear dynamics in semiconductors. It displays S-shaped current–voltage characteristics because the GR kinetics incorporating impact ionization from at least two impurity levels (ground state and excited state) allows for three different values of the carrier density $n(E)$ in a certain range of fields E. As a result of the negative differential conductivity, a variety of temporal and spatiotemporal instabilities occur, ranging from stationary and breathing current filaments and traveling charge density waves to various chaotic scenarios.

Band-to-band impact ionization of a reverse biased p–n junction is the basis of a variety of electronic devices. A number of these devices depend on a combination of impact ionization of hot electrons and transit time effects. The IMPATT (impact ionization avalanche transit time) diodes can generate the highest continuous power output at frequencies > 30 GHz. The originally proposed device (Read diode) involves a reverse biased $\mathrm{n^+}$–p–i–$\mathrm{p^+}$ multilayer structure, where $\mathrm{n^+}$ and $\mathrm{p^+}$ denote strongly n- or p-doped semiconductor regions, and i denotes an intrinsic (undoped) region. In the $\mathrm{n^+}$–p region (avalanche region), carriers are generated by impact ionization across the bandgap; the generated holes are swept through the i region (drift region) and collected at the $\mathrm{p^+}$ contact. When a periodic (ac) voltage is superimposed on the time-independent (dc) reverse bias, a π phase lag of the ac current behind the voltage can arise. This phase lag is due to the finite buildup time of the avalanche current and the finite time carriers take to cross the drift region (transit-time delay). If the sum of these delay times is approximately one-half cycle of the operating frequency, negative conductance is observed; in other words, the carrier flow drifts opposite to the ac electric field. This can be achieved by properly matching the length of the drift region with the drift velocity and the frequency. Other devices using the avalanche breakdown effect are the TRAPATT (trapped plasma avalanche triggered transit) diode and the avalanche transistor. The Zener diode is a p–n junction that exhibits a sharp increase in the magnitude of the current at a certain reverse voltage where avalanche breakdown sets in. It is used to stabilize and limit the dc voltage in circuits (overload and transient suppressor) since the current can vary over a large range at the avalanche breakdown threshold without a noticeable change in the voltage. The original Zener effect, on the other hand, is due to quantum mechanical tunneling across the bandgap at high fields, and is effective in highly doped (resulting in narrow depletion layers) Zener diodes at lower breakdown voltages.

ECKEHARD SCHÖLL

See also **Diodes; Drude model; Nonlinear electronics; Semiconductor oscillators**

Further Reading

Landsberg, P.T. 1991. *Recombination in Semiconductors*, Cambridge and New York: Cambridge University Press

Schöll, E. 1987. *Nonequilibrium Phase Transitions in Semiconductors*, Berlin: Springer

Schöll, E. 2001. *Nonlinear Spatio-temporal Dynamics and Chaos in Semiconductors*, Cambridge and New York: Cambridge University Press

Schöll, E., Niedernostheide, F.-J., Parisi, J., Prettl, W. & Purwins, H. 1998. Formation of spatio-temporal structures in semiconductors. In *Evolution of Spontaneous Structures in Dissipative Continuous Systems*, edited by F.H. Busse & S.C. Müller, Berlin: Springer, pp. 446–494

Shaw, M.P., Mitin, V.V., Schöll, E. & Grubin, H.L. 1992. *The Physics of Instabilities in Solid State Electron Devices*, New York: Plenum Press

AVALANCHES

An avalanche is a downhill slide of a large mass, usually of snow, ice, or rock debris prompted by a small initial disturbance. Avalanches, along with landslides, are one of the major natural disasters that still present significant danger for people in the mountains. On average, 25 people die in avalanches every winter in Switzerland alone. Dozens of people were killed on September 23, 2002, in a gigantic avalanche in Northern Ossetia, Russia, when a 150 m thick chunk of the Kolka Glacier broke off and triggered an avalanche of ice and debris that slid some 25 km along Karmadon gorge. In 1999, some 3000 avalanches occurred in the Swiss Alps.

Avalanches vary widely in size, from minor slides to large movements of snow reaching a volume of 10^5 m^3 and a weight of 30,000 tons. The speed of the downhill snow movement can reach 100 m/s. There are two main types of avalanches-loose avalanche and slab avalanche: depending on the physical properties of snow. Soft dry snow typically produces loose avalanches that form a wedge downward from the starting point, mainly determined by the physical properties of the granular material. In wet or icy conditions, on the other hand, a whole slab of solid dense snow may slide down. The initiation of the second type occurs as a fracture line at the top of the slab. The study of real avalanches and landslides is mostly an empirical science that is traditionally a part of geophysics and draws from the physics of snow, ice, and soil. Semi-empirical computer codes have been developed for prediction of avalanches dependent on the weather conditions (snowfall, wind, temperature profiles) and topography.

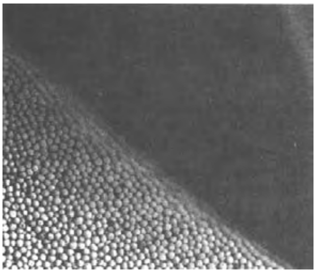

Figure 1. Only several layers of mustard seeds are involved in the rolling motion inside the avalanche: moving grains are smeared out in this long-exposure photograph. Reproduced with permission from Jaeger et al. (1998).

More fundamental aspects of avalanche dynamics have been studied in controlled laboratory experiments with dry or wet granular piles, or sandpiles. Granular slope can be characterized by two *angles of repose*—the *static* angle of repose θ_s which is the maximum angle at which the granular slope can remain static, and the *dynamic* angle of repose θ_d, or a minimum angle at which the granular flow down the slope can persist. Typically, in dry granular media, the difference between static and dynamic angles of repose is about 2–5°, for smooth glass beads $\theta_s \approx 25°$, $\theta_d \approx 23°$. Avalanches may occur in the bistable regime when the slope angle satisfies $\theta_d < \theta < \theta_s$. The bistability is explained by the need to *dilate* the granular material for it to enter flowing regime (*Bagnold's dilatancy*).

An avalanche can be initiated by a small localized fluctuation from which the fluidized region expands downhill and sometimes also uphill, while the sand always slides downhill. An avalanche in a deep sandpile usually involves a narrow layer near the surface (see Figure 1). Avalanches have also been studied in finite-depth granular layers on inclined planes. The two-dimensional structure of a developing avalanche depends on the thickness of the granular layer and the slope angle. For thin layers and small angles, wedge-shaped avalanches are formed similar to the loose snow avalanches (Figure 2a). In thicker layers and at higher inclination angles, avalanches have a balloon-type shape that expands both down- and uphill (Figure 2b).

The kinematics of the fluidized layer in one dimension can be described by a set of hydraulic equations for the local thickness $R(x,t)$ of the layer of rolling particles flowing over a sandpile of immobile particles with variable profile $h(x,t)$ (BCRE model, after Bouchaud et al., (1994)),

$$\partial_t R = -v\partial_x R + \Gamma(R,h) + \text{(diffusive terms)}, \quad (1)$$

$$\partial_t h = -\Gamma(R,h) + \text{(diffusive terms)}, \quad (2)$$

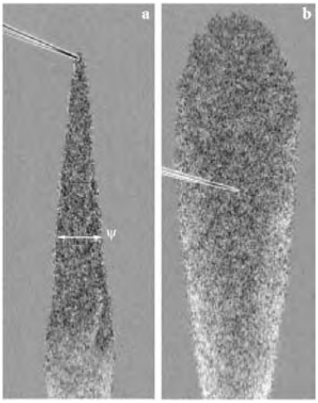

Figure 2. Structure of the avalanche in a thin (4 grain diameters) layer of glass beads: (a) wedge-shaped avalanche for $\theta = 31.5°$; (b) balloon-shaped avalanche propagating both up- and downhill for $\theta = 32.5°$. Reprinted by permission from *Nature* (Daerr & Douady, 1999).Copyright (1999) Macmillan Publishers Ltd.

where Γ is the entrainment flux of immobile particles into the rolling layer and the downhill transport velocity v is assumed constant. Γ becomes positive when the local slope becomes steeper than the static repose angle θ_s, and in the simplest case, $\Gamma = \gamma R(\partial_x h - \tan\theta_s)$. This model allows for a complete analytical treatment.

A more sophisticated continuum theory of granular avalanches is based on the fluid dynamics (Navier–Stokes) equations coupled with a phenomenological description of the first-order phase transition from a static to a fluidized state driven by the local shear stress (Aranson & Tsimring, 2001). The local phase state is described by the local order parameter ρ that is controlled by a Ginzburg–Landau-type equation with bistable free energy $F(\rho, \delta)$:

$$\partial_t \rho = D\nabla^2 \rho - \partial_\rho F(\rho, \delta) \qquad (3)$$

The control parameter δ in this equation depends on the ratio of shear to normal stress. This theory can describe a variety of "partially fluidized" granular flows, including avalanches in sandpiles. In a "shallow-water" approximation, it yields the BCRE-type equations for the local slope and the thickness of the rolling layer.

The wide distribution of scales in real avalanches led Bak et al. (1988) to propose a "sandpile cellular automaton" (*See* **Sandpile model**) as a paradigm model for *self-organized criticality* (SOC), the phenomenon that occurs in slowly driven nonequilibrium spatially extended systems when they asymptotically reach a critical state characterized by a power-law distribution of event sizes. The BTW model is remarkably simple, yet it exhibits a highly nontrivial behavior. The sandpile is formed on a lattice by dropping "grains" on a random site from above, one at a time. "Grains" form stacks of integer height at each lattice site. After each grain dropping the sandpile is allowed to relax. Relaxation occurs when the slope (a difference in heights of two adjacent stacks) reaches a critical value ("angle of repose") and the grain hops to a lower stack. This may prompt a series of subsequent hops and so trigger an avalanche. The size of the avalanche is determined by the number of grains set into motion by adding a single grain to a sandpile. In the asymptotic regime in a large system, the avalanche size distribution becomes scale-invariant, $P(s) \propto s^{-\alpha}$ with $\alpha \approx 1.5$.

The relevance of this model and its generalizations to real avalanches is still a matter of debate. The sandpile model is defined via a single repose angle, and so its asymptotic behavior has the properties of the critical state for a second-order phase transition. Real sandpiles are characterized by two angles of repose and thus exhibit features of the first-order phase transition. Experiments with avalanches in slowly rotating drums do not confirm the scale-invariant distribution of avalanches. However, in such experiments, the internal structures of the sandpile (the force chains) are constantly changing in the process of rotation. In other experiments with large monodispersed glass beads dropped on a conical sandpile, SOC with $\alpha \approx 1.5$ was observed. The characteristics of the size distribution depend on the geometry of the sandpile and the physical and geometrical properties of grains. SOC was also observed in the avalanche statistics in a three-dimensional pile of long rice; however, a smaller scaling exponent $\alpha \approx 1.2$ was measured for the avalanche size distribution.

An avalanche in a pile of sand has been used as a metaphor in many other physical phenomena including the avalanche diodes, vortices in type-II superconductors, Barkhausen effect in ferro-magnetics, $1/f$ noise, and.

LEV TSIMRING

See also **Granular materials; Sandpile model**

Further Reading

Aranson, I.S. & Tsimring, L.S. 2001. Continuum description of avalanches in granular media. *Physical Review E*, 64: 020301

Bak, P., Tang, C. & Wiesenfeld, K. 1988. Self-organized criticality. *Physical Review A*, 38: 364–374

Bouchaud, J.-P., Cates, M.E., Ravi Prakash, J. & Edwards, S.F. 1994. A model for the dynamics of sandpile surfaces. *Journal de Physique I*, 4: 1383–1410

Daerr, A. & Douady, S. 1999. Two types of avalanche behaviour in granular media. *Nature*, 399: 241–243

Duran, J. 1999. *Sands, Powders, and Grains: An Introduction to the Physics of Granular Materials*, Berlin and New York: Springer

Jaeger, H.M., Nagel, S.R. & Behringer, R.P. 1996. Granular solids, liquids, and gases. *Reviews of Modern Physics*, 68: 1259–1273

Jensen, H.J. 1998. *Self-Organized Criticality*, Cambridge: Cambridge University Press

Nagel, S.R. 1992. Instabilities in a sandpile. *Reviews of Modern Physics*, 64(1): 321–325

Rajchenbach, J. 2000. Granular flows. *Advances in Physics*, 49(2): 229–256

AVERAGING METHODS

Averaging methods are generally used for dynamical systems of two or more degrees of freedom when time scales or space scales are well separated. An average over the rapidly varying coordinates of one degree of freedom, considering the coordinates of the second degree of freedom to be constant during the average, can, with appropriate variables, retain a time-invariant quantity that enters into the solution of the slower motion. This solution, in turn, supplies a parameter to the rapid motion, which can then be solved in a lower-dimensional space. The averaging method is closely related to the calculation of *adiabatic invariants*, which are the approximately constant integrals of the motion that are obtained by averaging over the fast angle variables.

The lowest-order calculation is generally straightforwardly performed in canonical coordinates. A transformation from momentum and position coordinates (p, q), for the fast oscillation, to action-angle form (J, θ) gives a constant of the motion J, if all other variables are held constant. The action J is then the constant parameter in the equation for the slower motion. It is not always convenient to transform to action-angle form directly, but the underlying constants are related to the action variables.

The formal expansion procedure that is employed is to develop the solution in an asymptotic series. The mathematical method applied to ordinary differential equations was developed by Nikolai Bogoliubov (Bogoliubov & Mitropolsky, 1961) and in a somewhat different form by Martin Kruskal (1962). The expansion techniques can be formally extended to all orders in the perturbation parameter but are actually divergent. For multiple periodic systems, higher-order local nonlinear resonances between the degrees of freedom may destroy the ordering in their neighborhood. We will return to this problem below.

Averaging over the fastest oscillation of an N-degree-of-freedom system reduces the number of freedoms to $N-1$. A second average over the next fastest motion then produces a second adiabatic invariant to reduce the freedoms to $N-2$. This process may be continued to obtain a hierarchy of adiabatic invariants, until the system is reduced to one degree of freedom, which can be integrated to obtain a final integrable equation. The process is well known in plasma physics where, for a charged particle gyrating in a magnetic mirror field, we first find the magnetic moment invariant μ associated with the fast gyration, then find the longitudinal invariant $J_{||}$ associated with the slower bounce motion, and finally find the flux invariant Φ associated with the drift motion. The three degrees of freedom are shown in Figure 1. The small parameters in this case are ε_1, the ratio of bounce frequency to gyration frequency; ε_2, the ratio of guiding center drift frequency to bounce frequency; and ε_3, the ratio of the frequency of the time-varying magnetic field to the drift frequency. This example motivated the development of averaging methods. The derivations of these invariants are given in detail in Northrop (1963) or in other plasma physics texts.

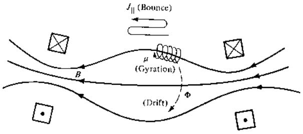

Figure 1. A hierarchy of adiabatic invariants for the charged particle gyrating in a nonaxisymmetric, magnetic mirror field. The three adiabatic invariants are the magnetic moment μ, the longitudinal invariant $J_{||}$, and the guiding center flux invariant Φ.

Although the asymptotic expansions are formally good to all orders in a small dimensionless parameter of the form $\varepsilon = |\dot{\omega}/\omega^2|$, where ω is the frequency of the fast oscillation that is slowly changing in time and $\dot{\omega} \equiv \mathrm{d}\omega/\mathrm{d}t$, the series generally diverge. The physical reason is that resonances or near-resonances between degrees of freedom lead to small denominators in the coefficients of terms. For nonlinear coupled oscillatory systems, exact resonances for certain values of the action locally change the structure of the phase-space orbits so that they do not follow the values obtained by averaging. This led to the development of the secular perturbation theory (Born, 1927), in which a local transformation of the coordinates around the resonance can be made. The frequency of the oscillatory motion in the neighborhood of the exact resonance is then slow compared with the other frequencies in the transformed coordinates, and averaging can then be applied locally. A review of the various methods, their limitations, practical examples, and reference to original sources can be found in Lichtenberg (1969) and Lichtenberg & Lieberman (1991).

The above discussion is related to the study of finite-dimensional systems governed by ordinary differential equations. The methods are usually applied to relatively low-dimensional systems, for example, the motion of a magnetically confined charged particle as described above. However, averaging methods are also applied to systems governed by partial differential equations, such as nonlinear wave propagation problems and wave instabilities. For example, waves on discrete oscillator chains can be obtained by first averaging over the discreteness using a Taylor expansion.

ALLAN J. LICHTENBERG

See also **Adiabatic invariants; Breathers; Collective coordinates; Modulated waves; Solitons**

Further Reading

Bogoliubov, N.N. & Mitropolsky, Y.A. 1961. *Assymptotic Methods in the Theory of Nonlinear Oscillators*, New York: Gordon & Beach

Born, M. 1927. *The Mechanics of the Atom*, London: Bell

Kruskal, M.D. 1962. Asymptotic theory of Hamiltonian systems with all solutions nearly periodic. *Journal of Mathematical Physics*, 3: 806–828

Lichtenberg, A.J. 1969. *Phase Space Dynamics of Particles*, New York: Wiley

Lichtenberg, A.J. & Lieberman M.A. 1991. *Regular and Chaotic Dynamics*, 2nd edition, New York: Springer

Northrop, T.G. 1963. *The Adiabatic Motion of Charged Particles*, New York: Wiley

B

BÄCKLUND TRANSFORMATIONS

Bäcklund transformations (BTs) originated in investigations conducted in the late 19th century into invariance properties of pseudospherical surfaces, namely surfaces of constant negative Gaussian curvature. In 1862, it was Edmond Bour who derived the well-known sine-Gordon equation

$$\omega_{uv} = \frac{1}{\rho^2} \sin \omega \qquad (1)$$

via the Gauss–Mainardi–Codazzi system for pseudospherical surfaces with total curvature $\mathcal{K} = -1/\rho^2$, parametrized in asymptotic coordinates. In 1883, Albert Bäcklund published his now classical result whereby pseudo-spherical surfaces may be generated in an iterative manner. Thus, if $\boldsymbol{r}$ is the position vector of a pseudospherical surface Σ corresponding to a seed solution ω of Equation (1) and ω' denotes the Bäcklund transformation of ω via the BT

$$\left.\begin{aligned} \omega'_u - \omega_u &= \frac{2\beta}{\rho} \sin\left(\frac{\omega' + \omega}{2}\right), \\ \omega'_v + \omega_v &= \frac{2}{\beta\rho} \sin\left(\frac{\omega' - \omega}{2}\right), \end{aligned}\right\} \mathbb{B}_\beta \qquad (2)$$

then the position vector $\boldsymbol{r}'$ of the one-parameter class of surfaces Σ' corresponding to ω' is given by (Bäcklund, 1883)

$$\begin{aligned} \boldsymbol{r}' &= \boldsymbol{r} + \frac{\mathrm{L}}{\sin\omega} \\ &\quad \times \left[\sin\left(\frac{\omega - \omega'}{2}\right) \boldsymbol{r}_u + \sin\left(\frac{\omega + \omega'}{2}\right) \boldsymbol{r}_v \right], \end{aligned} \qquad (3)$$

where $L = \rho \sin\zeta$ and $\beta = \tan(\zeta/2)$, ζ being the constant angle between the normals to Σ and Σ' and β being termed the Bäcklund parameter. Sophus Lie subsequently observed that $\mathbb{B}_\beta$ may be decomposed according to $\mathbb{B}_\beta = \mathbb{L}_\beta^{-1}\mathbb{B}_{\beta=1}\mathbb{L}_\beta$, where $\mathbb{L}_\beta$ and $\mathbb{L}_\beta^{-1}$ are Lie invariances. Thus, Lie transformations play a crucial role in intruding the Bäcklund parameter β into the parameter-independent "Bianchi" transformation $\mathbb{B}_{\beta=1}$ to produce $\mathbb{B}_\beta$.

It was in 1892 that Luigi Bianchi in his masterly paper *Sulla Trasformazione di Bäcklund per le Superficie Pseudosferiche* established that the BT $\mathbb{B}_\beta$ admits a commutation property $\mathbb{B}_{\beta_2}\mathbb{B}_{\beta_1} = \mathbb{B}_{\beta_1}\mathbb{B}_{\beta_2}$, a consequence of which is a nonlinear superposition principle embodied in what is termed a "permutability theorem."

Bianchi's Permutability Theorem

If ω is a seed solution of the sine-Gordon equation (1), let ω_1, ω_2 denote the BT of ω via $\mathbb{B}_{\beta_1}$ and $\mathbb{B}_{\beta_2}$, that is, $\omega_1 = \mathbb{B}_{\beta_1}(\omega)$ and $\omega_2 = \mathbb{B}_{\beta_2}(\omega)$. Let $\omega_{12} = \mathbb{B}_{\beta_2}(\omega_1)$ and $\omega_{21} = \mathbb{B}_{\beta_1}(\omega_2)$. Then, imposition of the commutativity requirement $\omega_{12} = \omega_{21}$ yields a new solution of (1), namely

$$\begin{aligned} \Omega &= \omega_{12} = \omega_{21} \\ &= \omega + 4\tan^{-1}\left[\frac{\beta_2 + \beta_1}{\beta_2 - \beta_1} \tan\left(\frac{\omega_2 - \omega_1}{4}\right)\right]. \end{aligned} \qquad (4)$$

This result is commonly encapsulated in what is termed a "Lamb diagram" as shown in Figure 1. This solution-generation procedure may be iterated via what is sometimes termed a Bianchi lattice. At each iteration, a new Bäcklund parameter β_i is introduced.

The discovery of the BT for the iterative construction of pseudospherical surfaces along with its concomitant permutability theorem led to an intensive search by geometers at the turn of the 20th century for other classes of privileged surfaces that possess Bäcklund-type transformations. In this connection, Luther Eisenhart, in the preface to his monograph *Transformations of Surfaces* published in 1922, asserted that: "During the past twenty-five years many of the advances in differential geometry of surfaces in Euclidean space have had to do with transformations of surfaces of a given type into surfaces of the same type." Thus,

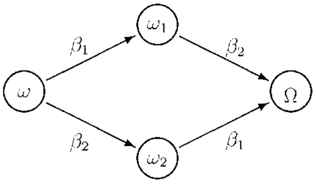

Figure 1. The Lamb diagram.

distinguished geometers such as Bianchi, Calapso, Darboux, Demoulin, Guichard, Jonas, Tzitzeica, and Weingarten all conducted extensive investigations into various classes of surfaces that admit BTs.

The particular Lamé system descriptive of triply orthogonal systems in the case when one of the constituent coordinate surfaces is pseudospherical was shown by Bianchi (1885) to admit an auto-BT, that is, a BT that renders the system invariant. Bianchi followed this in 1890 with the construction of a BT for the Gauss–Mainardi–Codazzi system associated with the class of hyperbolic surfaces with Gaussian curvature $\mathcal{K} = -1/\rho^2$ subject to the constraint

$$\rho_{uv} = 0, \tag{5}$$

where u, v are asymptotic coordinates. In 1899, Gaston Darboux constructed a BT for the nonlinear system

$$\begin{aligned} \theta_{xx} + \theta_{yy} + \kappa_1\kappa_2 e^{2\theta} &= 0, \\ \kappa_{1,y} + (\kappa_1 - \kappa_2)\theta_y &= 0, \\ \kappa_{2,x} + (\kappa_2 - \kappa_1)\theta_x &= 0 \end{aligned} \tag{6}$$

descriptive of isothermic surfaces with fundamental forms

$$\mathrm{I} = e^{2\theta}(dx^2 + dy^2), \quad \mathrm{II} = e^{2\theta}(\kappa_1 dx^2 + \kappa_2 dy^2), \tag{7}$$

where κ_1, κ_2 are principal curvatures and x, y are conjugate coordinates. The classical BT for system (6) has been set in a modern solitonic context by Cieśliński (1997).

In the first decade of the 20th century, the Romanian geometer Gheorghe Tzitzeica embarked upon an investigation of an important class of surfaces for which, in asymptotic coordinates, the Gauss–Mainardi–Codazzi system reduces to the nonlinear hyperbolic equation

$$(\ln h)_{uv} = h - h^{-2} \tag{8}$$

to be rediscovered some 70 years later in a soliton context. Tzitzeica (1910) not only constructed a BT for (8) but also set down what, in modern terms, is a linear representation containing a spectral parameter. Tzitzeica surfaces may be subsumed in the more general class of projective minimal surfaces for which BTs can be established (Rogers & Schief, 2002). In particular, this class contains the Demoulin system (1933)

$$(\ln h)_{uv} = h - \frac{1}{hk}, \quad (\ln k)_{uv} = k - \frac{1}{hk}. \tag{9}$$

The application of BTs in physics began with the work of Seeger et al. (1953) on crystal dislocation theory. Therein, within the context of Frenkel and Kontorova's dislocation theory, the superposition of so-called "eigenmotions" was obtained via the permutability relation (4). The interaction of what are today called breathers with kink-type dislocations was both described analytically and displayed graphically. The typical solitonic features to be subsequently discovered for the Korteweg–de Vries (KdV) equation

$$u_t + 6uu_x + u_{xxx} = 0, \tag{10}$$

(namely, preservation of velocity and shape following interaction as well as the concomitant phase shift) were all recorded. Bianchi's permutability theorem was subsequently employed in an investigation of the propagation of ultrashort optical pulses in a resonant medium by Lamb (1971).

A BT for the Korteweg–de Vries equation (10), namely

$$\begin{aligned} (\Lambda + \Lambda')_x &= \beta - \tfrac{1}{2}(\Lambda - \Lambda')^2, \\ (\Lambda' + \Lambda')_t &= (u - u')(u_{xx} - u'_{xx}) \\ &\quad -2(u_x^2 + u_x u'_x + u'^2_x), \end{aligned} \tag{11}$$

where

$$\Lambda = \int_{\infty}^{x} u(\sigma, t)\,d\sigma \tag{12}$$

was established by Wahlquist and Estabrook (1973). The spatial part of the BT was used to construct a permutability theorem, whereby multi-soliton solutions may be generated. This permutability theorem makes a remarkable appearance in numerical analysis as the so-called ε-algorithm.

A BT for the celebrated nonlinear Schrödinger (NLS) equation

$$iq_t + q_{xx} + 2|q|^2 q = 0 \tag{13}$$

was established by Lamb (1974) employing a direct method due to Clairin (1902) and by Chen (1974) via the inverse scattering transform (IST) formalism. The BT adopts the form

$$\begin{aligned} q_x + q'_x &= (q - q')(4\beta^2 - |q + q'|^2)^{1/2}, \\ q_t + q'_t &= i(q_x - q'_x)(4\beta^2 - |q + q'|^2)^{1/2} \\ &\quad + \frac{i}{2}(q + q')(|q + q'|^2 + |q - q'|^2), \end{aligned} \tag{14}$$

the spatial part of which may be used to construct a permutability theorem (Rogers & Shadwick, 1982).

Crum's theorem may be adduced to show that, at the level of the linear representation of soliton equations, the action of the BT is to add a discrete eigenvalue to the spectrum. The role of BTs in the context of the IST and their action on reflection coefficients is treated in detail by Calogero and Degasperis (1982).

That the Toda lattice equation

$$\ddot{y}_n = \exp[-(y_n - y_{n-1})] - \exp[-(y_{n+1} - y_n] \quad (15)$$

admits a BT, namely

$$\begin{aligned} \dot{y}_n - \dot{y}'_{n-1} &= \beta\,[\exp\{-(y'_n - y_n)\} \\ &\quad - \exp\{(y'_{n-1} - y_{n-1})\}], \\ \dot{y}'_n - \dot{y}_n &= \beta^{-1}[\exp\{-(y_{n+1} - y'_n)\} \\ &\quad - \exp\{-(y_n - y'_{n-1})\}] \end{aligned} \quad (16)$$

was established by Wadati and Toda (1975). BTs for a range of integrable differential-difference as well as integro-differential equations may be conveniently derived via Hirota's bilinear operator approach (see Rogers & Shadwick, 1982).

BTs have by now been constructed for the gamut of known solitonic equations as well as their Painlevé reductions (Gromak, 1999). The importance of BTs in soliton theory with regard to such aspects as multi-soliton generation, geometric connections, and integrable discretization is well established. Moreover, BTs also have extensive applications in continuum mechanics (Rogers & Shadwick, 1982). Important connections between infinitesimal BTs as originally introduced in a gas dynamics context (Loewner, 1952) and the construction of 2 + 1 dimensional solitonic systems have also been uncovered.

COLIN ROGERS

See also **Hirota's method; Inverse scattering method or transform; *N*-soliton formulas; Sine-Gordon equation; Solitons**

Further Reading

Bäcklund, A.V. 1883. Om ytor med konstant negativ krökning. *Lunds Universitets Årsskrift*, 19: 1–48

Bianchi, L. 1885. Sopra i sistemi tripli ortogonali di Weingarten. *Annali di Matematica*, 13: 177–234

Bianchi, L. 1890. Sopra alcone nuove classi di superficie e di sistemi tripli ortogonali. *Annali di Matematica*, 18:301–358

Bianchi, L. 1892. Sulla traformazione di Bäcklund per le superficie pseudosferiche. *Rendiconti Lincei*, 5: 3–12

Calogero, F. & Degasperis, A. 1982. *Spectral Transform and Solitons*, Amsterdam and New York: North-Holland

Chen, H.H. 1974. General derivation of Bäcklund transformations from inverse scattering problems. *Physical Review Letters*, 33: 925–928

Cieśliński, J. 1997. The Darboux–Bianchi transformation for isothermic surfaces. Classical results versus the soliton approach. *Differential Geometry and Its Applications*, 7: 1–28

Clairin, J. 1902. Sur les transformations de Bäcklund. *Annales de l'Ecole Normale Supérieure*, 27: 451–489

Darboux, G. 1899. Sur les surfaces isothermiques. *Comptes Rendus*, 128: 1299–1305

Demoulin, A. 1933. Sur deux transformations des surfaces dont les quadriques de Lie n'ont que deux ou trois points charactéristiques. *Bulletin de l'Académie Belgique*, 19: 479–501, 579–592, 1352–1363

Gromak, V. 1999. Bäcklund transformations of Painlevé equations and their applications. In *The Painlevé Property: One Century Later*, edited by R. Conte, New York: Springer

Konopelchenko, B.G & Rogers, C. 1993. On generalised Loewner systems: novel integrable equations in 2+1-dimensions. *Journal of Mathematical Physics*, 34: 214–242

Lamb, G.L. Jr. 1971. Analytical descriptions of ultra short optical pulse propagation in a resonant medium. *Reviews of Modern Physics*, 43: 99–124

Lamb, G.L. Jr. 1974. Bäcklund transformations for certain nonlinear evolution equations. *Journal of Mathematical Physics*, 15: 2157–2165

Loewner, C. 1952. Generation of solutions of systems of partial differential equations by composition of infinitesimal Bäcklund transformations. *Journal d'Analyse Mathématique*, 2: 219–242

Rogers, C. & Schief, W.K. 2002. *Bäcklund and Darboux Transformations: Geometry and Modern Applications in Soliton Theory*, Cambridge and New York: Cambridge University Press

Rogers, C. & Shadwick, W.F. 1982. *Bäcklund Transformations and Their Applications*, New York: Academic Press

Seeger, A., Donth, H. & Kochendörfer, A. 1953. Theorie der Versetzungen in Eindimensionalen Atomreihen III. Versetzungen, Eigenbewegungen und ihre Wechselwirkung, *Zeitschrift für Physik*, 134: 173–193

Tzitzeica, G. 1910. Sur une nouvelle classe de surfaces. *Comptes Rendus*, 150: 955–956

Wadati, M. & Toda, M. 1975. Bäcklund transformation for the exponential lattice. *Journal of the Physical Society of Japan*, 39: 1196–1203

Wahlquist, H.D. & Estabrook, F.B. 1973, Bäcklund transformations for solutions of the Korteweg–de Vries equation. *Physical Review Letters*, 31: 1386–1390

BAKER MAP

See **Maps**

BAKER–AKHIEZER FUNCTION

See **Integrable lattices**

BALL LIGHTNING

Properties

Ball lightning is an impressive natural phenomenon for which there is yet no accepted scientific explanation. It consists of flaming balls or fireballs, usually bright white, red, orange, or yellow, which appear unexpectedly sometimes near the ground, following the discharge of a lightning flash, or in midair coming from a cloud.

Most observations of ball lightning are associated with thunderstorms, and they exhibit the following more detailed properties: (i) Their shape is usually spherical or spheroidal with diameters between 10 and

50 cm. (ii) They tend to move horizontally. (iii) The observed distribution of lifetimes has a most probable value between 2 and 5 s and an average value of about 10 s or higher (some cases of more than 1 min having been reported). (iv) Ball lightning is bright enough to be clearly seen in daylight, the visible output being in the range 10–150 W (similar to that of a home electric light bulb). (v) Some balls have appeared within aircraft, traveling inside the fuselage along the aisle from front to rear. (vi) There are reports of odors, similar to those of ozone, burning sulfur, or nitric oxide, and of sounds, mainly hisses, buzzes, or flutters. (vii) Most balls decay silently, but some expire with an explosion. (viii) Ball lightning has killed or injured people and animals and damaged trees, buildings, cars, and electric equipment. (ix) Fires have been started showing that there is something hot inside. In such events, the released energy has been estimated to be between 10 kJ and more than 1 MJ. (x) Ball lightning has never been produced in laboratories, in spite of many attempts and some interesting results, including anode spots and luminous objects that decay very quickly. Consequently, the properties of ball lightning are derived from reports by witnesses, who are often excited by the phenomenon and have no scientific training.

A possibly related phenomenon has been observed in submarines, after a short circuit of the batteries. Balls of plasma that float in air for several seconds have appeared at the electrodes. On these occasions, the current was about 150 kA and the energy was estimated to be between 200 and 400 kJ.

Classification of the Models

Three main characteristics must be accounted for by a successful model but seem very difficult to explain: the tendency toward horizontal motion (hot air or plasma in air tends to rise), relatively long lifetimes, and contradictions among witnesses. For example, some report that balls are cold since they did not feel any warmth when one passed nearby, while others were burned and needed medical care.

The many different models proposed to explain the phenomenon can be classified into two groups, according to whether the energy source is internal or external. In the first group, some are based on plasmoids (equilibrium configurations of plasmas), high-density plasmas with quantum mechanical properties, closed loops of currents confined by their own magnetic field (in some cases, the linking of the currents playing an important role), vortex structures as whirlwinds or rotating spheres, bubbles containing microwave radiation, chemical reactions or combustion, fractal structures, aerosols, filaments of silicon, carbon nanotubes, nuclear processes, or new physics, and even primordial mini black holes. In the second group, some assume that the balls are powered by electrical discharges or by high-frequency microwave focused from thunderclouds. None of them is generally accepted.

Chemical and Electromagnetic Models

The association of ball lightning with electrical discharges suggests strongly that they have an electromagnetic nature. However, Michael Faraday argued that ball lightning cannot be an electric phenomenon as it would decay almost instantaneously, in contrast to its observed lifetime of at least several seconds. Finkelstein & Rubinstein (1964) used the time-independent magnetic virial theorem to place a stringent upper limit to the energy of a fireball. This limit has been viewed as a compelling argument against electromagnetic models, stimulating non-electromagnetic chemical approaches.

Recently, aerosol models have received considerable attention. In one model (Abrahamson & Dinniss, 2000), a lightning discharge vaporizes silicon dioxide in the soil that—after interacting with carbon compounds—is transformed into pure silicon droplets of nanometer scale. These droplets become coated with an insulating coat of oxides and are polarized, after which they become aligned with electric fields and form networks of filaments, in loose structures called “fluff balls.” In another model (Bychkov, 2002), the discharges pick up organic material from the soil and transform it into a kind of “spongy ball” that can hold electric charges. Models of this type fail to explain that some balls appear in mid air, where there is neither silicon nor organic nor any other similar material.

Electromagnetic models that include chemical effects are promising candidates for an explanation of ball lightning. Indeed, there are now counterarguments to the three main objections that express Faraday’s argument in modern language. These are based on the radiated output, the pinch effect, and magnetic pressure.

The first objection is that the power emitted by a plasma of the size of a ball lightning is too high (one liter of air plasma at 15,000 K emits about 5 MW, several orders of magnitude too much). This may be, however, an indication that most of the ball is at ambient temperature, only a small fraction being hot, concentrated in filamentary structures (as hot current streamers). If this fraction is of the order of 1 ppm, the radiated output would be of the order of 10–100 W, in agreement with reports.

But the solution to the first problem raises another one. Any plasma current channel inside the ball would be necked and cut in a very short time by the pinch effect (the Lorentz force); thus, a ball structured by such currents could not last long enough. However, in 1958, Chandrasekhar and Woltjer showed in an astrophysical context that plasmas relax to minimum energy states, verifying the condition that the electric current and the magnetic field are parallel so $\nabla \times \boldsymbol{B} = \lambda \boldsymbol{B}$, in which there can be no pinch effect since the Lorentz force vanishes. They concluded that such states, known as force-free fields, can confine large amounts of magnetic energy. Although the minimum energy of an uncontained plasma (as in ball lightning) is zero and corresponds to an infinitely expanded magnetic field, an almost force-free condition could be attained first in a very short time at a finite radius, a slow expansion continuing afterward with negligible pinch effect. This could take several seconds.

The third objection is based on the magnetic virial theorem, which states that a system of charges in electromagnetic interactions has no equilibrium state in the absence of external forces, because the large magnetic pressure must produce an explosion with no other force to compensate it. But it is not certain that the fireballs are in equilibrium; they could be just in metastable states with slow evolution, the streamers, moreover, clearly not being in equilibrium themselves. Still more important, the force-free condition annihilates the magnetic pressure or at least reduces it to a much smaller value if the field is almost force-free. Furthermore, the problem needs a much more complex analysis than has been offered up to now. For instance, one must include the thermochemical and quantum effects on the transport processes in the plasma as well as other nonlinear effects.

Faddeev and Niemi (2000) have proposed compelling arguments that challenge certain widely held views on plasmas, showing that the virial theorem does allow nontrivial equilibrium states of streamers and electromagnetic fields inside a background of plasma, which are "topologically stable solitons that describe knotted and linked flux tubes of helical magnetic fields." This kind of configuration was proposed in 1998 by Rañada et al. (2000) in the context of ball lightning. Therefore, it seems that the virial theorem does not necessarily support Faraday's view.

That ball lightning may contain force-free magnetic configurations of plasmas seems plausible. Because electric conduction in air proceeds through thin channels called streamers—as happens in ordinary lightning—it can be imagined that plasma inside the fireball consists of a self-organized set of metastable, highly conductive, wire-like or filamentary currents. Furthermore, unusually long-lived filaments (even closed loops) in high-density structures have been theoretically predicted and experimentally observed in many plasma systems within a great range of length scales, for instance, in astrophysics, tokamaks, and ordinary discharges in air. Thus, filamentary structures are currently receiving considerable attention.

The strongly nonlinear behavior of a plasma is enhanced when filamentary structures appear, leading to a complex non-isotropic system, which should be studied within a more general theory than ideal magnetohydrodynamics (MHD). Still, the main features of such systems can be described in the framework of resistive MHD. The important dissipative effects depend on the transport coefficients, such as thermal and electrical conductivities, which are highly nonlinear functions of the electromagnetic fields and the temperature, as well as of the chemical and quantum properties.

From the point of view of the MHD approximation, the dimensionless parameters of the plasma inside ball lightning may be quite similar to those found in other plasma scenarios, implying stable or metastable currents along a set of closed loops in filamentary structures. An interesting and unexplored possibility is the establishment, inside the streamers, of a quasicollision-free highly conductive regime in the direction of the magnetic field, which is strong and parallel to the streamers axis. In such a regime, both the electric and the thermal conductivities would become highly anisotropic. The first would be considerably enhanced along the axis of the streamer. On the other hand, both conductivities would be greatly reduced in the transverse directions, behaving as $1/B^2$ according to classical predictions. In this way, the dissipation and the spreading in the streamers would be much smaller than in ordinary regimes and should produce a long-lived strongly magnetized global plasma structure within an intricate stabilizing topology of filamentary currents.

In summary, even though the phenomenon of ball lightning has been known for many years, there is still no accepted theory to explain it—the alternatives currently most favored being the chemical and the electromagnetic models. The latter seem promising now, after recent results showed that some classical objections are not always applicable. As an example, a number of filamentary plasma structures have generated considerable interest, which are similar to stable plasma scenarios observed in nature, for instance, in astrophysics. These kinds of models could possibly embody chemical and electromagnetic elements.

Antonio F. Rañada, José L. Trueba, and José M. Donoso

See also **Helicity; Magnetohydrodynamics; Nonlinear plasma waves**

Further reading

Abrahamson, J. & Dinniss, J. 2000. Ball lightning caused by oxidation of nanoparticle networks from normal lightning strikes on soil. *Nature*, 403: 519–521

Barry, J.D. 1980. *Ball Lightning and Bead Lightning. Extreme Forms of Atmospheric Electricity*, New York: Plenum Press

Bychkov, V.L. 2002. Polymer-composite ball lightning. *Philosophical Transactions of the Royal Society* A, 360: 37–60

Faddeev, L. & Niemi, A.J. 2000. Magnetic geometry and the confinement of electrically conducting plasmas. *Physical Review Letters*, 85: 3416–3419

Finkelstein, D. & Rubinstein, J. 1964. Ball lightning. *Physical Review* A, 135: 390

Ohtsuki, Y.-H. (editor). 1988. *Science of Ball Lightning (Fire Ball)*, Singapore: World Scientific

Rañada, A.F., Soler, M. & Trueba, J.L. 2000. Ball lightning as a force-free magnetic knot. *Physical Review* E, 62: 7181–7190

Singer, S. 1971. *The Nature of Ball Lightning*, New York and London: Plenum Press

Smirnov, B.M. 1993. Physics of ball lightning. *Physics Reports*, 224: 151–236

Stenhoff, M. 1999. *Ball Lightning: An Unsolved Problem in Atmospheric Physics*, Dordrecht and New York: Kluwer

Trubnikov, B.A. 2002. Current filaments in plasmas. *Plasma Physics Reports*, 28: 312–326

BANACH SPACE

See **Function spaces**

BASIN OF ATTRACTION

See **Phase space**

BAXTER'S Q-OPERATOR

See **Bethe ansatz**

BELOUSOV–ZHABOTINSKY REACTION

In 1950, Boris Pavlovich Belousov worked at the Institute of Biophysics of the Ministry of Public Health of the USSR when he observed that the reaction between citric acid, bromate ions, and ceric ions (as catalyst) produced a regular periodic and reproducible change of color between an oxidized state and a reduced state. A temporal oscillating reaction appeared like a chemical clock. His 1951 paper on this study was largely rejected by the science community because it seemed to violate the Second Law of thermodynamics.

In 1961, Anatol Zhabotinsky, a Russian postgraduate student guided by his professor, Simon Shnoll, modified the previous reaction by replacing citric acid with malonic acid and adding ferroin sulfate as an indicator. As the oxidized state of ferroin is blue and the reduced one is red, he was able to observe an oscillating temporal reaction with larger oscillating amplitudes. This reaction

$$\begin{aligned}3\mathrm{CH_2(CO_2H)_2} + 4\mathrm{BrO_3^-} \\ = 4\mathrm{Br^-} + 9\mathrm{CO_2} + 6\mathrm{H_2O}\end{aligned} \tag{1}$$

was named the Belousov–Zhabotinsky (BZ) reaction. The first publications in English, recognizing the works of Belousov and Zhabotinsky, were done in 1967 by the Danish scientist Hans Degn. However, Zhabotinsky was unable to propose a complete mechanism for the system. Experimentally, the periodic evolution of the potential of the reacting solution can be followed by a potentiometric method, as shown in Figure 1.

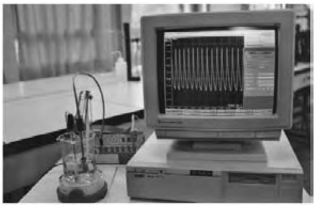

Figure 1. A photograph showing the periodic potential obtained between a platinum electrode and a reference electrode immersed in a BZ solution.

In 1972, Richard Field, Endre Körös, and Richard Noyes, at the University of Oregon, studied this mechanism using a bromide-selective electrode to follow the reaction. They proposed 18 steps involving 21 chemical species, using the same principles of chemical kinetics and thermodynamics that govern ordinary chemical reactions—this was the FKN mechanism. In 1974, the same scientists proposed a simplified mechanism with penetrating chemical insight: the "Oregonator" (in honor of the University of Oregon).

Compound:	$\mathrm{BrO_3^-}$	Organic species	HBrO
Notation:	A	B	P

Compound:	$\mathrm{HBrO_2}$	$\mathrm{Br^-}$	$\mathrm{Ce^{4+}}$
Notation:	X	Y	Z

With these notations, the FKN mechanism is

- $A + Y \rightarrow X + P$,
- $A + X \rightarrow 2X + 2Z$,
- $X + Y \rightarrow 2P$,
- $2X \rightarrow A + P$,
- $B + Z \rightarrow (f/2)Y +$ other products.

The second step is fundamental for the observation of oscillations; it is an autocatalytic reaction or retroaction loop. In the fifth step, B represents all oxidizable organic species present, and f is a stoichiometric factor. The kinetic differential equations are:

$$\begin{aligned}\frac{\mathrm{d}[X]}{\mathrm{d}t} &= k_1[A]\cdot[Y] + k_2[A]\cdot[X] \\ &\quad -k_3[X]\cdot[Y] - 2k_4[X]^2,\end{aligned} \tag{2}$$

$$\frac{\mathrm{d}[Y]}{\mathrm{d}t} = -k_1[A]\cdot[Y] - k_3[X]\cdot[Y] + 2k_5\frac{f}{2}[B]\cdot[X], \tag{3}$$

$$\frac{\mathrm{d}[Z]}{\mathrm{d}t} = 2k_2[A]\cdot[X] - k_5[B][Z]. \tag{4}$$

The rate constants are:

Rate constants:	k_1	k_2	k_3	k_4	k_5
Value (l/mol s):	1.28	2.4×10^6	33.6	3×10^3	1

This system is clearly nonlinear. Solutions can be obtained by numerical methods, and for $0.5 \leq f \leq 2.4$, some oscillating temporal solutions are observed, whose solutions depend on the initial conditions. According to the Second Law of thermodynamics, all spontaneous chemical changes in a homogeneous and closed system involve a decrease in free enthalpy of this system. If a fluctuation disrupts the system close to equilibrium, the system will return irrevocably to this stable state, making oscillations impossible. Nonetheless, it is possible to observe oscillations when the system is far from equilibrium. One of the striking properties of nonlinear systems is the effect of fluctuations (of the concentrations of intermediates), which can transform an unstable system into new states that are more organized than the initial state. Ilya Prigogine (who was awarded the 1977 Nobel Prize in chemistry for his work on thermodynamics) gave the name "dissipative structures" to such systems to emphasize the importance of irreversible phenomena far from equilibrium.

Continuing to work on oscillating reactions, in 1970, Zhabotinsky published, with Zaikin, a paper that announced the existence of two-dimensional waves in the BZ reaction; also in 1972, Arthur Winfree observed spiral wave patterns in a BZ reaction (see Figure 2). This reaction took place without stirring in a thin (approximately 2 mm thick) layer of reactants poured into a Petri dish. Blue concentric circles (called targets) radiated across the dish on a red background and self-generating spirals appeared. Soon afterward, scientists reported that a blue target center can produce waves of oxidation propagating through the reduced medium, and as the waves advance toward the interior of the center, they transform from red to blue. The period of oscillation is variable, but the speed of propagation is roughly constant. Thus, the BZ reaction proves to be a stationary spatiotemporal oscillating reaction.

With the aid of the FKN mechanism, Field and Noyes showed how to understand the development of such target waves. The diffusion is the transport of species from the areas of high concentrations to those of low concentrations. When there is a coupling between a chemical reaction with an autocatalytic step (or feedback-retroaction loop) as in the BZ reaction with the diffusion of species, spatial organization phenomena can occur; thus these are called "reaction-diffusion systems." In such systems, molecules react chemically with each other when they collide, and as the concentrations of components change, a chemical wave propagates.

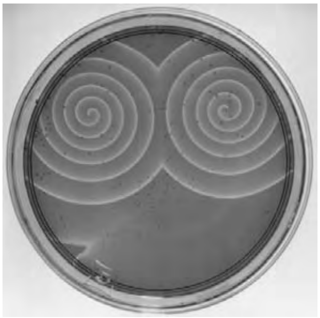

Figure 2. Spiral waves in a BZ reaction (Courtesy of A.T. Winfree). See text for details.

In 1984, Oleg Mornev (at the Institute of Theoretical and Experimental Biophysics of the Russian Academy of Sciences) showed that in an infinite plane stationary system of reaction-diffusion oscillations, Snell's sine law of refraction was verified. Thus the simple rule

$$\frac{\sin \psi_1}{\sin \psi_2} = \frac{v_1}{v_2} \tag{5}$$

dictates the angles ψ_1 and ψ_2 when waves hit an interface separating two regions with different speeds (v_1 and v_2) of wave propagation. This result was surprising because reaction-diffusion systems are nonlinear; thus, the medium is an active and an integral part of the wave. In fact, ψ_1 cannot be set arbitrarily but is slaved with ψ_2 due to the nature of the two regions.

In 1993, Zhabotinsky (at the Department of Chemistry, Brandeis University) demonstrated Snell's law experimentally. In 1998, Rui Dilaö and Joaquim Sainhas (at the Instituto Superio Tecnico, Lisbon Portugal) showed the following constraint in a reaction-diffusion system: after the reaction, the medium must be chemically identical to its initial form. In other words, while the waves are propagating and reactions are taking place, the medium has different properties, but these waves transform the medium back to the original species. By solving reaction-diffusion equations under this constraint, Dilaö and Sainhas showed (using computer simulations) that their formulation agrees with experiments, and mathematically chemical waves obey Snell's sine law. Continuing to work on the phenomenon of refraction, Mornev is developing formulations that hold both in infinite and finite media.

There are other examples of reaction-diffusion systems. In 1983, Patrick de Kepper (a French chemist in Toulouse) highlighted Turing structures in a ClO_2^-, I^-, malonic acid reaction (CIMA reaction). In Turing structures, stationary zones of varying concentrations appear in space. For these observations, the reaction must have steps with retroaction loops containing activators and other steps with inhibitors, and the activators must diffuse more slowly than the inhibitors.

Although the intense study of oscillating chemical reactions and nonlinear dynamics in chemistry is only about 30 years old, its progress has been impressive. Depending on the initial conditions, the BZ reactions can have unpredictable behaviors, even though they are described by deterministic laws. Thus, the BZ reaction belongs to the group of physical or chemical systems that exhibits deterministic chaos.

GÉRARD DUPUIS AND NICOLE BERLAND

See also **Brusselator; Chemical kinetics; Fairy rings of mushrooms; Reaction-diffusion systems; Turing patterns; Vortex dynamics in excitable media**

Further Reading

Dilao, R. & Sainhas, J. 1998. Wave optics in reaction-diffusion systems. *Physical Review Letters*, 80: 5216

Epsein, I.R. & Pojman, J.A. 1998. *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns and Chaos*, Oxford and New York: Oxford University Press

Field, R.J., Körös, E. & Noyes, R.M. 1972. *Journal of American Chemical Society*, 94: 8649

Gray, P. & Scott, S.K. 1990. *Chemical Oscillations and Instabilities, Nonlinear Chemical Kinetics*, Oxford: Clarendon Press and New York: Oxford University Press

Mornev, O.A. 1984. Elements of the "optics" of autowaves. In *Self-Organization of Autowaves and Structures Far from Equilibrium*, edited by V.I. Krinsky, Berlin: Springer, pp. 111–118

Zaikin, A.N. & Zhabotinsky, A.M. 1970. Concentration wave propagation in two dimensional liquid phase self oscillating system. *Nature*, 225: 535–537

Zhabotinsky, A.M. 1964. *Biofizika*, 9, 306.

Zhabotinsky, A.M., Eager, M.D. & Epstein, I.R. 1993. Refraction and reflection of chemical waves. *Physical Review Letters*, 71: 1526–1529

BENJAMIN–BONA–MAHONY EQUATION

See **Water waves**

BENJAMIN–FEIR INSTABILITY

See **Wave stability and instability**

BENJAMIN–ONO EQUATION

See **Solitons, types of**

BERNOULLI SHIFT

See **Maps**

BERNOULLI'S EQUATION

Bernoulli's equation is possibly the best-known result in fluid mechanics—and the most frequently abused. Bernoulli's equation may be viewed as an energy-conservation budget for a fluid particle as it travels up and down the "hills" of potential energy, due to the fields of gravity and the pressure within the fluid, acquiring and relinquishing kinetic energy. In its simplest form, it states that

$$V^2/2 + p/\rho + gz = C, \tag{1}$$

where p is the pressure in the fluid of density ρ, V is the flow speed, and z is the vertical coordinate. (The flow takes place in a uniform gravitational field of acceleration g.) The sum on the left-hand side of Equation (1) is a constant, C. The first term is the kinetic energy of the fluid per unit mass, the second and third terms are the potential energy (again per unit mass) in the combined energy "landscape" of pressure and gravity.

The result is credited to Daniel Bernoulli (1700–1782), son of Johann Bernoulli (1667–1748), and to his monograph *Hydrodynamica*, initiated in 1729 and ultimately published in 1738. The history of the equation is, however, much richer than this simple sequence of events would suggest. Hunter Rouse puts the issues this way in a book containing English translations of the writings of both the son and the father:

> Why [these] works should have been singled out for translation seems at first sight rather obvious, if only because of the frequency with which the name Bernoulli is on the hydraulician's lips. But it is only Daniel to whom one is making reference, and the word is gradually spreading that the theorem bearing his name is nowhere to be found in his habitually cited *Hydrodynamica*. Not until the last few years has mention of either the work *Hydraulica* or its author Johann Bernoulli appeared in the fluids literature with any frequency, and this almost exclusively in the writings of C. Truesdell. It is Truesdell's thesis that, whereas Daniel has received too much credit for the formulation of the Bernoulli theorem, Johann has received too little. (Carmody et al., 1968).

A complicated set of circumstances ensued in which both father (Johann) and son (Daniel) sent their manuscripts to Leonhard Euler for comment and this led to Johann's manuscript, which appears to have been composed later than Daniel's, being published in the Memoirs of the Imperial Academy of Science in St. Petersburg in 1737 and 1738 (although these were not printed until a decade later). Indeed, Johann's treatise first appeared in his collected works published in Switzerland in 1743. While Daniel's treatise has the gist of what we today call Bernoulli's equation, Johann's treatment is more mature and complete.

In the form stated in Equation (1), Bernoulli's equation applies only to steady, constant-density, irrotational flow, that is, to a flow pattern that is unchanging in time and that has no vorticity.

More refined versions may be derived. Thus, in a steady, constant-density flow with vorticity, Equation (1) still holds along each streamline, but the "constant" on the right-hand side may vary from streamline to streamline. Indeed, the gradient of this changing "Bernoulli constant," ∇C, equals the Lamb vector, the vector product of flow velocity and vorticity,

$$\boldsymbol{V} \times \boldsymbol{\omega} = \nabla C.$$

If the flow is irrotational but unsteady, a version of Bernoulli's equation again holds, but the constant on the right-hand side of (1) is replaced by (minus) the time derivative of the velocity potential. (In an irrotational flow, the velocity field is the gradient of a scalar known as the velocity potential.) With $V = -\nabla\phi$, where ϕ is the velocity potential, we obtain Bernoulli's equation in the form

$$(\nabla\phi)^2/2 + p/\rho + gz = -\frac{\partial\phi}{\partial t}, \qquad (2)$$

which, coupled with the condition of irrotational flow,

$$\Delta\phi = 0, \qquad (3)$$

gives a system of two partial differential equations for the fields p and ϕ.

Bernoulli's equation in the simplistic form "high flow speed implies low pressure and vice versa" is often applied as a first, crude explanation of many flow phenomena from the ability to balance a ball atop a plume of air to the lift on an airfoil in flight. Some of these explanations are too simplistic, not to say incorrect. Nevertheless, Bernoulli's equation, when properly applied under the assumptions that ensure its validity, can be an extremely useful and powerful tool of fluid flow analysis.

It is remarkable—and important to note—that Bernoulli's equation (1) is not invariant to a Galilean transformation, ordinarily a prerequisite for a physical law to be useful. Thus, if one wants to use Bernoulli's equation (1) to calculate the pressure distribution for flow around an object, assuming the velocity field is known, it is essential to do so in a frame of reference in which the flow satisfies the necessary assumptions, in particular, that the flow is steady. The correct result is obtained by carrying out such a calculation in a frame of reference moving with the body. If the calculation is attempted in the "laboratory frame" through which the object is moving, one has to tackle the much more complex version of Bernoulli's equation given in (2). If the version in Equation (1) is applied, one obtains an incorrect result.

HASSAN AREF

See also **Fluid dynamics**

Further Reading

Batchelor, G.K. 1967. *An Introduction to Fluid Dynamics*, Cambridge: Cambridge University Press

Carmody, T., Kobus, H. & Rouse, H. 1968. *Hydrodynamica by Daniel Bernoulli and Hydraulica by Johann Bernoulli*, translated from the Latin, with a preface by Hunter Rouse, New York: Dover

Lamb, H. 1932. *Hydrodynamics*, 6th edition, Cambridge: Cambridge University Press

BERRY'S PHASE

Consider the parallel transport of an orthonormal frame along a line of constant latitude on the surface of a sphere. In going once around the sphere, the frame undergoes a rotation through an angle $\Delta\theta = 2\pi\cos\alpha$, where α is the colatitude. This may be shown using the geometry of Figure 1. As is also evident from the figure, this phase shift is purely geometric in character—it is independent of the time it takes to traverse the closed loop.

This construction underlies the well-known phase shift exhibited by the Foucault pendulum as the Earth rotates through one full period. Although arising through a dynamical process involving two widely separated time scales (the period of the Earth's rotation and the oscillation period of the pendulum), the phase shift in this and other examples is now understood in a more unified way. *Holonomic* effects such as these arise in a host of applications ranging from problems in superconductivity theory, fiber optic design, magnetic resonance imaging (MRI), amoeba propulsion and robotic locomotion and control, micromoter design, molecular dynamics, rigid-body motion, vortex dynamics in incompressible fluid flows (Newton, 2001), and satellite orientation control. For a survey and further references on the use of phases in locomotion problems, see Marsden & Ostrowski (1998).

That the falling cat learns quickly to re-orient itself optimally in mid-flight while maintaining zero angular momentum is a manifestation of the fact that controlling and manipulating a system's internal or shape variables can lead to phase changes in the external, or group variables, a process that can be exploited and has deeper connections to problems related to the dynamics of Yang–Mills particles moving in their associated gauge field, a link that is the *falling cat theorem* of Montgomery (1991a) (see further discussion and references in Marsden (1992) and Marsden & Ratiu (1999)). One can read many of the original articles leading to our current understanding of the geometric phase in the collection edited by Shapere & Wilczek (1989).

Problems of this type have a long and complex history dating back to work on the circular polarization of light in an inhomogeneous medium by Vladimirskii and Rytov in the 1930s and by Pancharatnam in the 1950s, who studied interference patterns produced by plates of

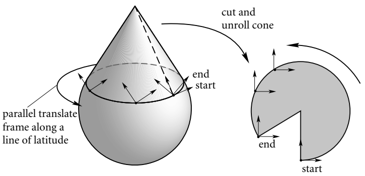

Figure 1. Parallel transport of a frame around a line of latitude.

an anisotropic crystal. Much of this early history is described in the articles by Michael Berry (Berry, 1988, 1990). The more recent literature was initiated by his earlier articles (Berry, 1984, 1985), which investigated the evolution of quantum systems whose Hamiltonian depends on external parameters that are slowly varied around a closed loop. The adiabatic theorem of quantum mechanics states that for infinitely slow changes of the parameters, the evolution of the complex wave function, governed by the time-dependent Schrödinger equation, is instantaneously in an eigenstate of the *frozen* Hamiltonian. At the end of one cycle, when the parameters recur, the wave function returns to its original eigenstate, but with a phase change that is related to the geometric properties of the closed loop. This phase change now goes by the name *Berry's phase.* Geometric developments started with the work of Simon (1983), and Marsden et al. (1989). One can introduce a bundle of eigenstates of the slowly varying Hamiltonian, as well as a natural *connection* on it; the Berry phase is then the bundle holonomy associated with this connection, while the curvature of the connection, when integrated over a closed two-dimensional (2-d) surface in parameter space gives rise to the first Chern class characterizing the topological twisting of this bundle.

The classical counterpart to Berry's phase was originally developed by Hannay (1985) (hence the terminology *Hannay's angle*) and is most naturally described by considering slowly varying integrable Hamiltonian systems in action-angle form. If we let $(I_1, \ldots, I_n; \theta_1, \ldots \theta_n)$ represent the action-angle variables of a given integrable system, then the governing Hamiltonian can be expressed as $\mathcal{H}(I_1, \ldots, I_n; R(t))$, where $R(t)$ is a slowly varying parameter that cycles through a closed loop in time period T, that is, $R(t+T) = R(t)$, $\dot{R}(t) \sim \varepsilon R$, $\varepsilon \ll 1$. The configuration space for the system is an n-dimensional torus $\mathbb{T}^n$ and we seek a formula for the angle variables as the parameter or parameters slowly evolve around the closed loop $\mathcal{C}$ in parameter space. The time-dependent system is governed by

$$\dot{\mathbf{I}} = \dot{R}(t) \cdot \frac{\partial \mathbf{I}}{\partial R}, \tag{1}$$

$$\dot{\boldsymbol{\theta}} = \boldsymbol{\omega}(\mathbf{I}) + \dot{R}(t) \cdot \frac{\partial \boldsymbol{\theta}}{\partial R}, \tag{2}$$

where

$$\omega(\mathbf{I}) \equiv \frac{\partial \mathcal{H}}{\partial \mathbf{I}}.$$

Since R is slowly varying, we can average the system around level curves of the frozen (i.e., $\varepsilon = 0$) Hamiltonian. If we let $\langle \quad \rangle$ denote this phase-space average, then the averaged canonical system becomes

$$\dot{\mathbf{I}} = \dot{R}(t) \cdot \left\langle \frac{\partial \mathbf{I}}{\partial R} \right\rangle \tag{3}$$

$$\dot{\boldsymbol{\theta}} = \boldsymbol{\omega}(\mathbf{I}) + \dot{R}(t) \cdot \left\langle \frac{\partial \boldsymbol{\theta}}{\partial R} \right\rangle. \tag{4}$$

The well-known adiabatic theorem of quantum mechanics guarantees that the action variable is nearly constant due to its adiabatic invariance, whereas the angle variables can be integrated over period T

$$\begin{aligned} \theta_T^i &= \int_0^T \omega^i(\mathbf{I}) \mathrm{d}t + \int_0^T \dot{R}(t) \cdot \left\langle \frac{\partial \theta^i}{\partial R} \right\rangle \mathrm{d}t && (5) \\ &= \theta_{\mathrm{d}} + \theta_{\mathrm{g}}. && (6) \end{aligned}$$

The first term, θ_{d}, called the dynamic phase is due to the frozen system, while the second term, θ_{g}, arises from the time variation. This geometric phase can be rewritten in a revealing manner as

$$\begin{aligned} \theta_{\mathrm{g}} &= \int_0^T \dot{R}(t) \cdot \left\langle \frac{\partial \theta^i}{\partial R} \right\rangle \mathrm{d}t && (7) \\ &= \oint \left\langle \frac{\partial \theta^i}{\partial R} \right\rangle \mathrm{d}R. && (8) \end{aligned}$$

The contour integral is taken over the closed loop $\mathcal{C}$ in parameter space. Although arising through a dynamical process, it is ultimately a purely geometric quantity that results from a delicate balance of two compensating effects in the limit $\varepsilon \to 0$. On the one hand, $T \to \infty$ in (7), while on the other, $\dot{R}(t) \to 0$. Their rates exactly balance so that the integral leaves a residual term in the limit $\varepsilon = 0$, as given in (8).

A nice example developed in Hannay (1985) is the *bead-on-hoop* problem in which a frictionless bead is constrained to slide along a closed planar wire hoop that encloses area $\mathcal{A}$ and has perimeter length $\mathcal{L}$. As the bead slides around the hoop, the hoop is slowly rotated about its vertical axis (which is aligned with the gravitational vector) through one full revolution. We are interested in the angular position of the bead with respect to a fixed point on the hoop after one full revolution of the hoop. When compared with its angular position had the hoop been held fixed (the frozen problem), this angle difference would represent the geometric phase and is given by

$$\Delta\theta = -8\pi^2 \mathcal{A}/\mathcal{L}^2. \tag{9}$$

Montgomery (1991b) shows that modulo 2π, we have the following *rigid-body phase formula*:

$$\Delta\theta = -\Lambda + 2ET/R. \tag{10}$$

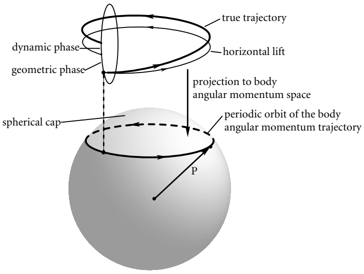
true trajectory
dynamic phase
horizontal lift
geometric phase
projection to body angular momentum space
spherical cap
periodic orbit of the body angular momentum trajectory
P

Figure 2. The geometry of the rigid-body phase formula.

Let us explain the notation in this remarkable formula. When a rigid body is freely spinning about its center of mass, one learns in mechanics that this dynamics can be described by the Euler equations, which are equations for the body angular momentum $\boldsymbol{\Pi}$. This vector in $\mathbb{R}^3$ moves on a sphere (of radius $R = \|\boldsymbol{\Pi}\|$) and describes periodic orbits (or exceptionally, heteroclinic orbits). This orbit is schematically depicted by the closed curve on the sphere shown in Figure 2. However, the full dynamics includes the dynamics of the rotation matrix for describing the attitude of the rigid body as well as its conjugate momentum. There is a projection from the full dynamic phase space (which is 6-d) to the body angular momentum space (which is 3-d). After one period of the motion on the sphere, the actual rigid-body motion was not periodic, but it had rotated about the spatial angular momentum vector by an angle $\Delta\theta$, the left-hand side of the above formula. The quantity Λ is the spherical angle subtended by the cap shown in the figure, E is the energy of the trajectory, and T is the period of the closed orbit on the sphere. A detailed history of this formula is given in Marsden & Ratiu (1999).

Paul K. Newton and Jerrold E. Marsden

See also **Adiabatic invariants; Averaging methods; Hamiltonian systems; Integrability; Phase space**

Further Reading

Berry, M.V. 1984. Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society, London* A, 392: 45–57

Berry, M.V. 1985. Classical adiabatic angles and quantal adiabatic phase. *Journal of Physics* A, 18: 15–27

Berry, M.V. 1988. The geometric phase. *Scientific American*, December, 46–52

Berry, M.V. 1990. Anticipations of the geometric phase. *Physics Today*, 43(12), 34–40

Hannay, J.H. 1985. Angle variable holonomy in adiabatic excursion of an integrable Hamiltonian. *Journal of Physics* A, 18: 221–230

Marsden, J.E. 1992. *Lectures on Mechanics*, Cambridge and New York: Cambridge University Press

Marsden, J.E., Montgomery, R. & Ratiu, T. 1989. Cartan–Hannay–Berry phases and symmetry. *Contemporary Mathematics*, 97: 279–295

Marsden, J.E. & Ostrowski, J. 1998. Symmetries in motion: geometric foundations of motion control, *Nonlinear Science Today*. (http://link.springer-ny.com)

Marsden, J.E. & Ratiu, T. 1999. *Introduction to Mechanics and Symmetry*, 2nd edition, New York: Springer

Montgomery, R. 1991a. Optimal control of deformable bodies and its relation to gauge theory. In *The Geometry of Hamiltonian Systems*, edited by T. Ratiu, New York: Springer, pp. 403–438

Montgomery, R. 1991b. How much does a rigid body rotate? A Berry's phase from the 18th century, *American Journal of Physics*, 59: 394–398

Newton, P.K. 2001. *The N-Vortex Problem: Analytical Techniques*, New York: Springer, Chapter 5

Shapere, A. & Wilczek, F. (editors). 1989. *Geometric Phases in Physics*, Singapore: World Scientific

Simon, B. 1983. Holonomy, the quantum adiabatic theorem, and Berry's phase. *Physical Review Letters*, 51(24): 2167–2170

BETHE ANSATZ

The Bethe ansatz is the name given to a method for exactly solving quantum many-body systems in one spatial dimension (1-d) or classical statistical lattice models (vertex models) in two spatial dimensions (Baxter, 1982; Korepin et al., 1993). The method was developed by Hans Bethe in 1931 (Bethe, 1931) in order to diagonalize the Hamiltonian of a chain of N spins with isotropic exchange interactions, introduced by Werner Heisenberg some years before as the simplest model for a 1-d magnet. This result was achieved by assuming the wave function to be of the form

$$f(x_1, x_2, ..., x_M) = \sum_P A_P \, \mathrm{e}^{\mathrm{i}\sum_{j=1}^{M} k_{P_j} x_j} \qquad (1)$$

with the sum performed on all possible permutations P of M distinct wave numbers $\{k_1, ..., k_M\}$, corresponding to down spins in the system (Bethe ansatz). By imposing invariance under the physical symmetries of the system (discrete translations and total spin rotations), Bethe obtained conditions on the coefficients A_P, which were satisfied if a set of M nonlinear equations (Bethe equations) in N complex parameters (Bethe numbers) were fulfilled. Surprisingly, the wave functions thus constructed were simultaneous eigenfunctions not only of the translation operator, the total spin $\boldsymbol{S}$, and its projection S_z along the z-direction but also of the isotropic Heisenberg Hamiltonian

$$H = \sum_{i=1}^{N} \left(\boldsymbol{S}_i \cdot \boldsymbol{S}_{i+1} - \frac{1}{4} \right). \qquad (2)$$

The energy and the crystal momentum were expressed as symmetric functions of the Bethe numbers; thus, the eigenvalue problem for H was reduced to the solution of an algebraic problem—solution of the Bethe equations.

This remarkable result was possible because of the existence of additional symmetries of the Heisenberg Hamiltonian, which emerged thanks to the ansatz made by Bethe on the wave function. In this original formulation, the method is known as the coordinate Bethe ansatz.

Progress in clarifying the role of symmetries in the Bethe ansatz, the link with integrable systems, and the algebraic aspect of the method was achieved by the Saint Petersburg (formerly Leningrad) School in the course of developing the quantum inverse scattering method (QISM) (Faddeev, 1984; Sklyanin & Faddeev, 1978; Korepin et al., 1993), after the work of Baxter on the integrability of vertex models (Baxter, 1982). In this approach, a key role is played by the monodromy operator defined as $\tau(\lambda) = L_N(\lambda)L_{N-1}, ..., L_1(\lambda)$, with λ being a complex number (spectral parameter) and $L_n(\lambda)$ being the quantum local Lax operator defined for the isotropic Heisenberg model as

$$L_n(\lambda) = \begin{pmatrix} \lambda + \mathrm{i}S_n^z & \mathrm{i}S_n^- \\ \mathrm{i}S_n^+ & \lambda - \mathrm{i}S_n^z \end{pmatrix} \tag{3}$$

with S^+, S^- raising and lowering spin-$\frac{1}{2}$ operators. Note that L_n can be viewed as an operator acting on the space $h_n \otimes V$, where h_n ($\equiv C^2$) is the physical Hilbert at site n (the space of couples of complex numbers), and V is an auxiliary space related to the matrix representation of L_n (for the present case, V is also identified with C^2). The product of Lax operators, taken in the auxiliary space, coincides with the usual matrix multiplication, so that the monodromy matrix can be rewritten as

$$\tau(\lambda) = \begin{pmatrix} A(\lambda) & B(\lambda) \\ C(\lambda) & D(\lambda) \end{pmatrix}, \tag{4}$$

with A, B, C, D operators acting on the full physical Hilbert space: $H = \prod \otimes h_i$. As is known from QISM (Korepin et al., 1993; Sklyanin & Faddeev, 1978; Faddeev, 1984), the commutation relations between elements of the monodromy matrix can be written in a compact form as

$$R(\lambda - \mu)\,(\tau(\lambda) \otimes \tau(\mu)) = (\tau(\mu) \otimes \tau(\lambda))\,R(\lambda - \mu), \tag{5}$$

where R is a 4×4 matrix (quantum R-matrix) satisfying the Yang–Baxter equation. From Equation (5), it follows that the trace of the monodromy operator (also known as the transfer matrix) $T(\lambda) \equiv \mathrm{tr}(\tau(\lambda)) = A(\lambda) + D(\lambda)$ gives rise, for different values of the spectral parameter, to an abelian algebra of operators

$$[T(\lambda), T(\mu)] = 0. \tag{6}$$

One can prove that the Hamiltonian is also an element of this algebra so that, for a system of N sites, there are N quantum integrals of motion in involution, corresponding to Liouville integrability in the classical limit. The diagonalization of the Hamiltonian and the other integrals of motion is thus reduced to the solution of the eigenvalue problem for the transfer matrix T. This problem can be solved by the so-called algebraic Bethe ansatz, a procedure that resembles the algebraic diagonalization of the harmonic oscillator by means of creation and annihilation operators. It relies on the existence of a vector $|\Omega\rangle$ (pseudovacuum) in the Hilbert space, which is annihilated by the operator C of the monodromy matrix

$$C(\lambda)|\Omega\rangle = 0. \tag{7}$$

For the Heisenberg chain, $|\Omega\rangle$ can be chosen as $|\Omega\rangle = \prod_{i=1}^{N} \otimes |\uparrow\rangle_i$ with $|\uparrow\rangle_i$ denoting the spin up state at site i. From Equation (3) it is clear that L_n acts on the state $|\uparrow\rangle_n$ as a triangular matrix, and the same is true for $\tau(\lambda)$ acting on $|\Omega\rangle$; thus, Equation (7) is automatically satisfied (C plays the role of an annihilation operator). From Equations (3) and (4), it is also evident that $A(\lambda)|\Omega\rangle = (\lambda + \mathrm{i}/2)^N|\Omega\rangle$, $D(\lambda)|\Omega\rangle = (\lambda - \mathrm{i}/2)^N|\Omega\rangle$. Moreover, one can show that the operator B in Equation (4) can be used as a creation operator.

By taking N different values of the spectral parameter $\lambda_1, \lambda_2, ..., \lambda_N$, one constructs a trial wave function as

$$|\Psi(\lambda_1, ..., \lambda_N)\rangle = \prod_{i=1}^{N} B(\lambda_i)|\Omega\rangle. \tag{8}$$

A direct calculation shows that

$$\begin{aligned} T(\lambda)|\Psi(\lambda_1, ..., \lambda_N)\rangle \;=\; & \Lambda(\lambda)|\Psi(\lambda_1, ..., \lambda_N)\rangle \\ & + \text{unwanted terms}, \end{aligned} \tag{9}$$

where the unwanted terms can be calculated using the commutation relations of A and D with B, obtained from Equation (5). The unwanted terms, however, are eliminated if the λ_i are taken as solutions of the Bethe equations that, for the isotropic Heisenberg chain, are of the form

$$\left(\frac{\lambda_\alpha - \mathrm{i}/2}{\lambda_\alpha + \mathrm{i}/2}\right)^N = -\prod_{\beta=1}^{M} \frac{\lambda_\alpha - \lambda_\beta - \mathrm{i}}{\lambda_\alpha - \lambda_\beta + \mathrm{i}}, \quad \alpha = 1, 2, ..., M. \tag{10}$$

The set of states obtained from Equation (8) in correspondence of the solutions of this system of nonlinear equations can be shown to be complete. The diagonalization of $T(\lambda)$, and hence of the Hamiltonian and all the quantum integrals of motion, is thus reduced to the problem of solving the Bethe equations. For finite size systems, this is a difficult problem to solve due to the nonlinearity of the equations, and one usually resorts to numerical tools. In the thermodynamical limit, however, it is possible to obtain exact solutions of the energy spectrum by deriving linear integral equation for the density distribution of the Bethe solutions in a complex plane (however, this requires an assumption on the nature of the solution known as the string hypothesis).

The algebraic Bethe ansatz has been successfully applied to a large class of many-body problems, including anisotropic generalizations of the Heisenberg chain, the Hubbard model, and the Kondo model, and has stimulated a variety of related approaches including Baxter's q-operator method (Baxter, 1982) and the notion of quantum groups. Recent progress in the computation of correlation functions of quantum-integrable many-body problems have also been made using the Bethe ansatz (Korepin et al., 1993).

Mario Salerno

See also **Quantum inverse scattering method; Salerno equation**

Further Reading

Baxter, R.J. 1982. *Exactly Solved Model of Statistical Mechanics*, New York: Academic Press

Bethe, H. 1931. Zur Theorie der Metalle I. Eigenwerte and Eigenfunktionen der Linearen Atomkette. *Zeitschrift für Physik*, 71: 205–226

Faddeev, L.D. 1984. Integrable models in 1+1 dimensional quantum field theory. In *Recent Advances in Field Theory and Statistical Mechanics*, Les Houches 1982, edited by J.B. Zuber & R. Stora, Amsterdam: North-Holland

Korepin, V.E., Bogoliubov, N.M. & Izergin, A.G. 1993. *Quantum Inverse Scattering Method and Correlation Functions*, Cambridge and New York: Cambridge University Press, and references therein

Sklyanin, E.K. & Faddeev, L.D. 1978. Quantum mechanical approach to completely integrable field theory models. *Soviet Physics Doklady*, 23: 902

BIFURCATIONS

Bifurcations are critical events that arise in systems when an external control parameter is varied (Arnol'd et al., 1994). For small values of the parameter the system will be linear and a unique fixed point will exist. As the parameter is changed to ranges, where nonlinearity becomes important, instabilities in the form of new fixed points or solutions with qualitatively different dynamical behavior may arise at bifurcations. These critical events are of mathematical and practical interest since their analysis can be performed, and they form organizing centers for observed dynamics (Guckenheimer & Holmes, 1986).

As an example, consider the simple physical system of a plastic ruler that is compressed lengthwise between your hands. This was first considered by Leonhard Euler in 1744 and is often referred to as the Euler strut problem (Acheson, 1997). At small forces, the ruler is approximately straight and supports the applied load. This is called the trivial state of the system. However, as the load increases, buckling takes place so that the ruler is deflected up or down. The straight trivial state becomes unstable and is replaced by a pair of solutions where each corresponds to one of the buckled states. If the ruler and the application of the load were both perfect, this would provide a physical example of a symmetry-breaking supercritical pitchfork bifurcation of the type shown in the bifurcation diagram of Figure 1(a).

The symbol *X* represents the deflection of the center of the ruler, which is used as the measure of the state of the system shown plotted as a function of the applied load λ. The symmetry that is broken is the mirror-plane symmetry of the straight ruler. The bifurcation is called supercritical because the nontrivial branches have the same stability as the trivial state, and it is termed pitchfork due to its shape.

When the bifurcating solutions have a stability opposite to the trivial, the bifurcation is called subcritical. A sketch of such a bifurcation is given in Figure 1(b) where an increase in the parameter λ would involve a jump to a large *X* state when λ is increased beyond the critical value λ_c. In order to regain the trivial state, λ would then have to be reduced to reach the folded part of the solution branches, and a sudden change back to the trivial state would occur. Hence, hysteresis takes place between the two transitions, and such a path is labeled C in Figure 1(b). The pair of folds in Figure 1(b) are called saddle-node bifurcations (Iooss & Joseph, 1990). A physical example of this is provided by the buckling of an elastic wire such as the outer portion of a bicycle brake cable. When a short length is held vertically, it will stand upright. If you push the remaining length upward through your hand, it will eventually become long enough so that gravity will cause it to fall over through a large angle of deflection. Now pull it back downward through your hand and you will find that the deflected state remains over a range of lengths before flipping back to the vertical. This is an example of a hysteresis loop.

The two models in Figures 1(a) and (b) contain a reflection symmetry. If this is not present, the bifurcation is transcritical and an example of such a bifurcation is given in Figure 1(c). In the physical example of the buckling of the ruler, this type of bifurcation would be observed if a constant side load

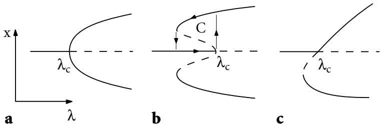

Figure 1. Sketches of (a) supercritical pitchfork, (b) subcritical pitchfork and (c) transcritical bifurcations. Solid lines indicate stable solutions and dashed lines indicate unstable.

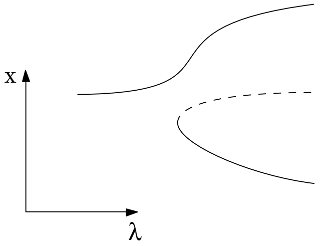

Figure 2. Imperfect pitchfork bifurcation.

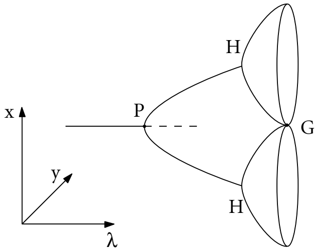

Figure 3. Schematic of a gluing bifurcation sequence.

was applied in addition to the end load in the example of the ruler. In this case, the mid-plane symmetry is automatically broken.

Of course, in any real system, physical imperfections will be present. These can be taken into account using the imperfect bifurcation theory (Golubitsky & Schaeffer, 1985). The effect of an imperfection is to disconnect the supercritical pitchfork bifurcation as shown in Figure 2. It can be seen that there is one state that evolves smoothly with an increase in parameter λ and another disconnected branch. In the example of the ruler, imperfections could arise from irregularities in the shape of the ruler or a slight imbalance in the applied load. The lower limit of the disconnected branch is defined by another type of bifurcation, a saddle node. The disconnected state can be attained, either by variation of two parameters (e.g., variation of side load for the Euler strut), or by a discontinuous or sudden jump in the parameter λ. In the latter case, there is a finite chance that the system will land on the disconnected solution. Examples of observations of such behavior in fluid flows are provided by Taylor–Couette flow (Pfister et al., 1988), a flow through a sudden expansion (Fearn et al., 1990), and convection (Arroya & Saviron, 1992).

Another very important bifurcation is a Hopf where a simply periodic cycle arises from a fixed point as a parameter is changed. As in the case of pitchforks, Hopf bifurcations may also be super- or subcritical, with hysteresis present in the latter case. An interesting feature of supercritical Hopf bifurcations is that the system takes more and more time to reach the equilibrium state as the bifurcation point nears. The observed long-term dynamics are analogous to critical slowing down (Landau & Lifshitz, 1980) in phase transitions and have been found in a fluid flow (Pfister & Gerdts, 1981), for example.

Further interesting global bifurcations (Glendinning, 1994) may occur when pitchfork and Hopf bifurcations occur sequentially. An example of this is shown schematically in Figure 3, where the pair of asymmetric states that arise at the pitchfork *P* then undergo a pair of Hopf bifurcations at the points labeled *H*. The cycles that arise at *H* then join together at a gluing bifurcation (Coullet et al., 1984) when λ is increased. This point is marked *G* in Figure 3, and it is an example of a homoclinic bifurcation. In this case, a single large orbit is formed from the pair of cycles as λ is increased beyond *G* with the period going to infinity exactly at *G*. Interesting dynamical behavior including chaos can be observed near such points in experiments when physical imperfections are taken into account (Glendinning et al., 2001).

Tom Mullin

See also **Catastrophe theory; Critical phenomena; Equilibrium; Hopf bifurcation; Phase transitions**

Further Reading

Acheson, D. 1997. *From Calculus to Chaos*, Oxford and New York: Oxford University Press

Arnol'd, V.I., Afrajmovich, V.S., Il'yashenko, Yu.S. & Shil'nikov, L.P. 1994. *Bifurcation Theory and Catastrophe Theory*, Berlin and New York: Springer

Arroyo, M.P. & Saviron, J.M. 1992. Rayleigh Bénard convection in a small box-spatial features and thermal-dependence of the velocity field. *Journal of Fluid Mechanics*, 235: 325–348

Coullet, P., Gambaudo, J.M. & Tresser, C. 1984. Une nouvelle bifurcation de codimension 2: le collage de cycles. *Comptes Rendu de l'Academie des Sciences de Paris, Ser. I–Mathematische* 299: 253–256

Fearn, R.M., Mullin, T. & Cliffe, K.A. 1990. Nonlinear flow phenomena in a symmetric sudden-expansion. *Journal of Fluid Mechanics*, 211: 595–608

Glendinning, P. 1994. *Stability, Instability and Chaos: An Introduction to the Theory of Nonlinear Differential Equations*, Cambridge and New York: Cambridge University Press

Glendinning, P., Abshagen, J. & Mullin, T. 2001. Imperfect homoclinic bifurcations. *Physical Review* E, 64: 036208

Golubitsky, M. & Schaeffer, D.G. 1985. *Singularities and Groups in Bifurcation Theory I*, Berlin and New York: Springer

Guckenheimer, J. & Holmes, P.J. 1986. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, 2nd edition, Berlin and New York: Springer

Iooss, G. & Joseph, D.D. 1990. *Elementary Stability and Bifurcation Theory*, 2nd edition, Berlin and New York: Springer

Landau, L.D. & Lifshitz, E.M. 1980. *Statistical Physics*, part 1, 3rd edition, London: Pergamon

Pfister, G. & Gerdts, U. 1981. The dynamics of Taylor wavy vortex flow. *Physics Letters A*, 83: 23–27

Pfister, G., Schmidt, H. Cliffe, K.A. & Mullin, T. 1988. Bifurcation phenomena in Taylor–Couette flow in a very short annulus. *Journal of Fluid Mechanics*, 191: 1–18

BI-HAMILTONIAN STRUCTURE

See **Integrable lattices**

BILAYER LIPID MEMBRANES

When a group of unknown researchers reported the artificial assembly of a bimolecular lipid membrane in vitro (at a 1961 symposium on the plasma membrane sponsored by the American and New York Heart Association), it was initially met with skepticism. The research group led by Donald O. Rudin began their report with a description of mundane soap bubbles, followed by "black holes" in soap films, ending with an invisible "black" lipid membrane made from extracts of cows' brains. The reconstituted structure (7.5 nm thick) was created just like a cell membrane separating two aqueous solutions. The speaker then said:

> upon adding one, as yet unidentified, heat-stable compound. . .from fermented egg white. . . to one side of the bathing solutions. . .lowers the resistance. . .by 5 orders of magnitude to a new steady state. . .which changes with applied potential. . .Recovery is prompt. . . the phenomenon is indistinguishable. . . from the excitable alga *Valonia*. . ., and similar to the frog nerve action potential (Ottova & Tien, 2002).

The first report was published a year later (Mueller et al., 1962). In reaction to that report, the subsequent inventor of liposomes (artificial spherical bilayer lipid membranes) wrote recently in an article entitled "Surrogate cells or Trojan horses" (Bangham, 1995):

> . . .a preprint of a paper was lent to me by Richard Keynes, then Head of the Department of Physiology [Cambridge University], and my boss. This paper was a bombshell . . .They [Mueller, Rudin, Tien, and Wescott] described methods for preparing a membrane . . . not too dissimilar to that of a node of Ranvier. . .The physiologists went mad over the model, referred to as a "BLM", an acronym for Bilayer or by some for Black Lipid Membrane. They were as irresistible to play with as soap bubbles.

Indeed, the Rudin group was playing with soap bubbles using equipment purchased from a local toy shop. But scientific experimentation with soap bubbles began with the observations of Robert Hooke (who coined the word "cell" in 1665 to describe the structure of a thin slice of cork tissue observed through a microscope he had constructed), with his observation of "black spots" in soap bubbles and films. Years later, Isaac Newton estimated the blackest soap film (when light waves reflecting from one layer of soap molecules destructively interfere with light waves reflecting from the second layer of soap molecule) to be about $3–8\times10^{-6}$ in. thick. (Modern measurements give thicknesses between 5 and 9 nm, depending on the soap solution used.)

Origins of the Lipid Bilayer Concept

The recognition of the lipid bilayer as a model for biomembranes dates back to the work of Hugo Fricke, in the 1920s and 1930s, who calculated the thickness of red blood cell (RBC) membranes to be between 3.3 and 11 nm, based on frequency-dependent measurements of the impedance of cell suspensions. Modern measurements on experimental bilayer lipid membranes (BLMs) and biomembranes confirm Fricke's estimation of the thickness of the plasma membrane (Tien & Ottova, 2000).

In his 1917 studies of the molecular organization of fatty acids at the air-water interface, Irving Langmuir had demonstrated that a simple trough apparatus could provide the data to estimate the dimensions of a molecule. Evert Gorter and F. Grendel (respectively, a pediatrician and a chemist) used Langmuir's trough to determine the area occupied by lipids extracted from red blood cell (from human, pig, or rat sources) "ghosts" (empty membrane sacs) and found that there was enough lipid to form a layer two molecules thick over the whole cell surface. In other words

$$\frac{\text{surface area occupied (from monolayer experiment)}}{\text{surface area of red blood cell}} \cong 2. \tag{1}$$

Thus, Gorter and Grendel suggested that the plasma membrane of red blood cells may be thought of as a lipid bilayer, with the polar (hydrophilic) head groups oriented outward.

Experimental Realization

The structure of black soap films led to the realization by Rudin and his co-workers in 1960 that a soap film in its final stages of thinning has a structure composed of two fatty acid monolayers sandwiching an aqueous solution as follows:

air | monolayer | soap solution | monolayer | air.

With the above background in mind, Rudin et al. simply proceeded to make a BLM under an aqueous solution, which may be represented as follows:

aqueous solution | BLM | aqueous solution.

Their effort was successful (Tien & Ottova, 2001, p. 86). Rudin and his colleagues showed that a

BLM formed from brain extracts and separating two aqueous solutions was self-sealing to punctures, with many physical and chemical properties similar to those of biomembranes. Upon modification with a certain compound called excitability-inducing molecule (EIM), this otherwise electrically inert structure became excitable, displaying characteristic features similar to those of action potentials of the nerve membrane.

By the end of the early 1970s, it had been determined that an unmodified bilayer lipid membrane separating two similar aqueous solutions is about 5 nm thick and is in a liquid-crystalline state with the following electrical properties: membrane potential ($E_\mathrm{m} \simeq 0$), membrane resistivity ($R_\mathrm{m} \simeq 10^9\ \Omega\,\mathrm{cm}$), membrane capacitance ($C_\mathrm{m} \simeq$ 0.5–1 $\mu\mathrm{F\,cm^{-2}}$), and dielectric breakdown ($V_\mathrm{b} > 250{,}000\ \mathrm{V/cm}$). In spite of its very low dielectric constant ($\varepsilon \simeq$ 2–7), this liquid-crystallline BLM is surprisingly permeable to water (8–24 µm/s) (Tien & Ottova, 2000).

The Lipid Bilayer Principle

In spite of their variable compositions, the fundamental structural element of all biomembranes is a liquid-crystalline phospholipid bilayer. Thus, the lipid bilayer principle of cell or biological membranes may be summarily stated as follows: all living organisms are made of cells, and every cell is enclosed by a plasma membrane, the indispensable component of which is a lipid bilayer. The key property of lipid bilayer-based cells is that they are separated from the environment by a permeability barrier that allows them to preserve their identity, take up nutrients, and remove waste. This 5 nm thick liquid-crystalline lipid bilayer serves not only as a physical barrier but also as a conduit for transport, a reactor for energy conversion, a transducer for signal processing, a bipolar electrode for redox reactions, or a site for molecular recognition.

The liquid-crystalline lipid bilayer of biomembranes not only provides the physical barrier separating the cytoplasm from its extracellular surroundings, it also separates organelles inside the cell to protect important processes and events. More specifically, the lipid bilayer of cell membrane must prevent its molecules of life (genetic materials and many proteins) from diffusing away. At the same time, the lipid bilayer must keep out foreign molecules that are harmful to the cells. To be viable, the cell must also communicate with the environment to continuously monitor the external conditions and adapt to them. Further, the cell needs to pump in nutrients and release toxic products of its metabolism. How does the cell carry out all of these multi-faceted activities?

A brief answer is that the cell depends on its lipid-proteins-carbohydrate complexes (i.e., glycoproteins, proteolipids, glycolipids, etc.) embedded in the lipid bilayer to gather information about the environment in various ways. Examples include communication with hundreds of other cells about a variety of vital tasks such as growth, differentiation, and death (apoptosis). Glycoproteins are responsible for regulating the traffic of material to and from the cytoplasmic space. Paradoxically, the intrinsic structure of cell membranes creates a bumpy obstacle to these vital processes of intercellular communication. The cell shields itself behind its lipid bilayer, which is virtually impermeable to all ions (e.g., $\mathrm{Na^+}$, $\mathrm{K^+}$, $\mathrm{Cl^-}$) and most polar molecules (except H_2O). This barrier must be overcome, however, for a cell to inform itself of what is happening in the world outside, as well as to carry out vital functions. Thus, over millions and millions of years of evolution, the liquid-crystalline lipid bilayer—besides acting as a physical restraint—has been modified to serve as a conduit for material transport, as a reactor for energy conversion, as a bipolar electrode for redox reactions, as a site for molecular recognition, and other diverse functions such as apoptosis and signal transduction.

Insofar as membrane transport is concerned, cells make use of three approaches: simple diffusion, facilitated diffusion, and active transport. Although simple diffusion is an effective transport mechanism for some substances such as water, the cell must make use of other mechanisms for moving substances in and out of the cell. Facilitated diffusion utilizes membrane channels to allow charged molecules, which otherwise could not diffuse across the lipid bilayer. These channels are especially useful with small ions such as $\mathrm{K^+}$, $\mathrm{Na^+}$, and $\mathrm{Cl^-}$. The number of protein channels available limits the rate of facilitated transport, whereas the speed of simple diffusion is controlled by the concentration gradient. Under active transport, the expenditure of energy is necessary to translocate the molecule from one side of the lipid bilayer to the other, in contrast to the concentration gradient. Similar to facilitated diffusion, active transport is limited by either the capacity of membrane channels or the number of carriers present.

Today, ion channels are found ubiquitously. To name a few, they are in the plasma membrane of sperm, bacteria, and higher plants; the sarcoplasmic retculum of skeletal muscle, nerve membrane, synaptic vesicle membranes of rat cerebral cortex, and the skin of carps. As a weapon of attack, many toxins released by living organisms such as dermonecrotic toxin, hemolysin, brevetoxin, and bee venom are polypeptide-based ion-channel formers. For example, functioning of membrane proteins, in particular, ionic channels, can be modulated by alteration of their arrangement in membranes (e.g., electroporation, Tien & Ottova, 2003).

At the membrane level, most cellular activities involve some kind of lipid bilayer-based receptor-ligand contact interactions. Outstanding examples among these are ion-sensing, molecular recognition (e.g., antigen-antibody binding and enzyme-substrate interaction), light conversion and detection, gated channels, and active transport. The development of self-assembled bilayer lipid membranes (BLMs and liposomes) has made it possible to investigate directly the electrical properties and transport phenomena across a 5 nm thick biomembrane element separating two aqueous phases. A modified or reconstituted BLM is viewed as a dynamic structure that changes in response to environmental stimuli as a function of time, as described by the so-called dynamic membrane hypothesis. Under this hypothesis, each type of receptor interacts specifically with its own ligand. That is, the so-called G-receptor is usually coupled to a guanosine nucleotide-binding protein that in turn stimulates or inhibits an intracellular, lipid bilayer-bound enzyme. G-protein-linked receptors mediate the cellular responses to a vast variety of signaling molecules, including local mediators, hormones, and neurotransmitters, which are as varied in structure as they are in function. G-protein-linked receptors usually consist of a single polypeptide chain, which threads back and forth across the lipid bilayer up to seven times. The members of this receptor family have a similar amino acid sequence and functional relationship. The binding sites for G-proteins have been reported to be the second and third intracellular loops and the carboxy-terminal tail. The endogenous ligands, such as hormones, neurotransmitters, and exogenous stimulants such as odorants, belonging to this class are important target analytes for biosensor technology Tien & Ottova (2003).

H. Ti. TIEN AND ANGELICA OTTOVA-LUEITMANNOVA

See also **Langmuir–Blodgett films; Nerve impulses; Neurons**

Further Reading

Bangham, A.D. 1995. Surrogate cells or Trojan horses. *BioEssays*, 17: 1081–1088

Mueller, P., Rudin, D.O., Tien H.T. & Wescott, W.C. 1962. Reconstitution of cell membrane structure in vitro and its transformation into an excitable system. *Nature*, 194: 979–980

Ottova A. & Tien, H.T. 2002. The 40th anniversary of bilayer lipid membrane research. *Bioelectrochemistry*, 56: 171–173

Ottova, A., Tvarozek, V. & Tien, H.T. 2003. Supported BLMs. In *Planar Lipid Bilayers (BLMs) and Their Applications*, edited by H.T. Tien & A. Ottova-Leitmannova, Amsterdam: Elsevier

Tien, H.T. 1974. *Bilayer Lipid Membranes (BLM): Theory and Practice*, New York: Marcel Dekker

Tien, H.T. & Ottova, A.L. 2000. *Membrane Biophysics: As Viewed from Experimental Bilayer Lipid Membranes (Planar Lipid Bilayers and Spherical Liposomes)*, Amsterdam and New York: Elsevier Science

Tien, H.T. & Ottova, A. 2001. The lipid bilayer concept and its experimental realization: from soap bubbles, the kitchen sink, to bilayer lipid membranes. *Journal of Membrane Science*, 189: 83–117

Tien, H.T. & Ottova, A. 2003. The bilayer lipid membrane (BLM) under electrical fields. *IEEE Transactions on Dielectrics and Electrical Insulation*, 10(5): 717–727

BILLIARDS

In mathematical physics, the singular noun "billiards" denotes a dynamical system corresponding to the inertial motion of a point mass within a region that has a piecewise smooth boundary. The reflections from the boundary are taken to be elastic; that is, the angle of reflection equals the angle of incidence. This model arises naturally in optics, acoustics, and classical and statistical mechanics. In fact, two fundamental models in statistical mechanics, gas of hard spheres (Boltzmann gas) and the Lorentz gas, are billiards. The billiards concept occupies a central position in nonlinear physics because it provides ideal visible models for analysis of dynamical properties leading to classical chaos and an ideal testing ground for the semiclassical analysis of quantum systems.

Billiards models are Hamiltonian systems. Hence, the phase volume is preserved under the dynamics, and the system can be studied in the framework of ergodic theory. In particular, the boundary of the billiard region is supposed to be only piecewise smooth; that is, it consists of smooth components. Therefore, the dynamics of billiards is not defined for orbits that hit singular points of the boundary. However, the phase volume of such orbits equals zero. The dynamics of billiards is completely defined by the shape of its boundary. A smooth component of the boundary is called dispersing, focusing, or neutral if it is convex inward, outward the billiard region, or if it is flat (has zero curvature), respectively. Any billiard orbit is a broken line in its configuration space.

The classical examples of integrable billiards are provided by circular and elliptical boundaries. Configuration spaces of these billiards are foliated by caustics, which are smooth curves (surfaces γ) such that if one link of the billiard orbit is tangent to γ, then every other link of this orbit is tangent to γ. Billiards in a circle has one family of caustics formed by (smaller) concentric circles, while billiards in an ellipse has two families of caustics (confocal ellipses and confocal hyperbolas), which are separated by orbits such that each link intersects a focus of the ellipse. Birkhoff's conjecture (Birkhoff, 1927) claims that among all billiards inside smooth convex curves, only billiards in ellipses are integrable. Berger (1990) has shown that in three dimensions (*d*), only billiards in ellipsoids produce foliations of a billiard region by smooth convex caustics. However, it does not imply that only billiards in ellipsoids are integrable because if a billiard in $d > 2$ has an invariant hypersurface then this hypersurface does not necessarily consist of rays tangent to some hypersurface in the configuration space. Using KAM theory, Lazutkin has shown that if a billiards boundary is strictly convex, with a sufficiently smooth curve and its curvature never vanishes; then there exists an uncountable number of smooth caustics in the vicinity of the boundary, and moreover, the phase volume of the orbits tangent to these caustics is positive (Lazutkin, 1991).

An opposite situation occurs when a boundary is everywhere dispersing. Such models were introduced by Sinai (1970), in his seminal paper and they are called Sinai (or dispersing) billiards (Figure 1(a)). Sinai billiards have the strongest chaotic properties; that is, they are ergodic, mixing, have a positive metric entropy, and are Bernoulli systems. If a (narrow) parallel beam of rays is made to fall onto a dispersing boundary, then after reflection it becomes divergent and, therefore, the distance between the rays in this beam increases with time. It is the mechanism of dispersing that generates sensitive dependence on initial conditions (hyperbolicity) and is responsible for strong chaotic properties of dispersing billiards.

On the other hand, focusing boundaries produce the opposite effect. Indeed, a narrow parallel beam of rays after reflection from the focusing boundary becomes convergent; that is, the distance between rays in such a beam decreases with time. Therefore, it has been the general understanding that a dispersing boundary always produces chaotization of the dynamics, while a focusing boundary produces stabilization of the dynamics. However, there exists another mechanism of chaos in billiards (and, in general, in Hamiltonian systems), which is called defocusing (Bunimovich, 1974, 1979). The point is that a narrow parallel beam of rays, after focusing because of reflection from a focusing boundary, may become divergent provided that a free path between two consecutive reflections from the boundary is long enough. Assuming that the time of divergence exceeds (averaged over all orbits) the time of convergence, one obtains chaotic billiards. One of the first, and the most famous, example of such billiards is called a stadium (Figure 1(b)). One obtains a stadium by cutting a circle into two semi-circles and connecting them by two common tangent segments. The length of these segments could be arbitrarily small, which demonstrates that the mechanism of defocusing can work under small deformations of even the integrable (a circle) billiards. Focusing billiards can have as strong chaotic properties as Sinai's billiards do (Bunimovich, 2000).

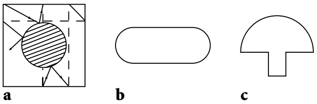

Figure 1. (a) Sinai billiard. (b) Stadium. (c) Mushroom.

There are no other mechanisms of chaos in billiards. Indeed, billiards in polygons and polyhedrons have zero metric entropy (Boldrighini et al., 1978). Nevertheless, a typical billiard in a polygon is ergodic (Kerckhoff et al., 1986). Because focusing components can form parts of the boundary of integrable as well as chaotic billiards, a natural question is whether there are some restrictions. Two classes of focusing components admissible in chaotic billiards were found (Wojtkowski, 1986; Markarian, 1988). The most general class of such focusing components is formed by absolutely focusing mirrors (AFM) (Bunimovich, 1992). AFMs form a new notion in geometric optics. A mirror γ (or a smooth component of a billiards' boundary) is called absolutely focusing if any narrow parallel beam of rays that falls on γ becomes focused after its last reflection in a series of consecutive reflections from γ. Observe that a mirror is focusing if any parallel beam of rays becomes focused just after the first reflection from this mirror. AFMs can also be characterized in terms of their local properties (Donnay, 1991; Bunimovich, 1992).

Generic Hamiltonian systems are neither integrable nor chaotic. Instead, their phase spaces get divided into KAM-islands and chaotic sea(s). The only clear and clean example of this phenomenon is a billiard in a mushroom (Bunimovich, 2001). The mushroom consists of a semicircular hat sitting on a foot (Figure 1(c)). A mushroom becomes a stadium when the width of the foot equals the width of the hat.

Clearly, the mechanism of dispersing works is higher than two dimensions as well (Sinai, 1970). It is not obvious at all for the mechanism of defocusing because of astigmatism. However, chaotic focusing billiards also do exist in dimension $d \geq 3$ (Bunimovich & Rehacek, 1998). But one pays a price of astigmatism by not allowing the focusing component to be as large as it can be in $d = 2$.

Many properties of classical dynamics of billiards are closely related to the properties of the corresponding quantum problem. Consider the Schrödinger equation with a potential equal to zero inside the billiard region and equal to infinity on the boundary. The eigenfunctions become uniformly distributed over the regions of ergodic billiards for high wave numbers (Shnirelman, 1991). On the contrary, there exist infinite series of eigenfunctions localized in the vicinity of convex caustics of billiards (Lazutkin, 1991).

Leonid Bunimovich

See also **Ergodic theory; Horseshoes and hyperbolicity in dynamical systems; Lorentz gas**

Further Reading

Berger, M. 1990. Sur les caustiques de surfaces en dimension 3. *Comptes Rendu de l'Academie de Sciences*, 311: 333–336

Birkhoff, G. 1927. *Dynamical Systems*. New York, American Mathematical Society

Boldrighini, C., Keane, M. & Marchetti, F. 1978. Billiards in polygons. *Annals of Probability*, 6: 532–540

Bunimovich, L.A. 1974. On billiards close to dispersing. *Mathematical USSR Sbornik*, 95: 49–73 (originally published in Russian)

Bunimovich, L.A. 1979. On the ergodic properties of nowhere dispersing billiards. *Communications in Mathematical Physics*, 65: 295–312

Bunimovich, L.A. 1992. On absolutely focusing mirrors. In *Ergodic Theory and Related Topics*, edited by U. Krengel, et al., Berlin and New York: Springer, pp. 62–82

Bunimovich, L.A. 2000. Billiards and other hyperbolic systems with singularities. In *Dynamical Systems, Ergodic Theory and Applications*, edited by Ya. G. Sinai, Berlin: Springer

Bunimovich, L.A. 2001. Mushrooms and other billiards with divided phase space. *Chaos*, 11: 802–808

Bunimovich, L.A. & Rehacek, J. 1998. How many dimensional stadia look like. *Communications in Mathematical Physics*, 197: 277–301

Donnay, V. 1991. Using integrability to produce chaos: billiards with positive entropy. *Communications in Mathematical Physics*, 141: 225–257

Kerckhoff, S., Mazur, H. & Smillie, J. 1986. Ergodicity of billiard flows and quadratic differentials. *Annals of Mathematics*, 124: 293–311

Lazutkin, V. F. 1991. *The KAM Theory and Asymptotics of Spectrum of Elliptic Operators*. Berlin and New York: Springer

Markarian, R. 1988. Billiards with Pesin region of measure one. *Communications in Mathematical Physics*, 118: 87–97

Shnirelman, A. I. 1991. On the asymptotic properties of eigenfunctions in the regions of chaotic motion. Addendum in *The KAM Theory and Asymptotics of Spectrum of Elliptic Operators* by V. F. Lazutkin. Berlin and New York: Springer

Sinai, Ya. G. 1970. Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. *Russian Mathematical Surveys*, 25: 137–189 (originally published in Russian 1970)

Wojtkowski, M. 1986. Principles for the design of billiards with nonvanishing Lyapunov exponents. *Communications in Mathematical Physics*, 105: 391–414

BINDING ENERGY

When two particles form a bound state under a certain kind of physical interaction, the resulting state has an energy smaller than the sum of the rest energies of the constituent elements of such a bound state. That is why, by definition (or one could say by construction), bound states are ones in which work has to be done to separate the constituents.

The energy that one has to provide (equivalently the work) in order to separate a bound state into its elements is called the binding energy, E_b, and from the above, it can be directly inferred that $E_\mathrm{b} > 0$. The equivalent mass to this energy (under the Einstein relation) also bears a name and is called the "mass defect," $\Delta m = E_\mathrm{b}/c^2$, where c is the speed of light.

Examples of binding energy can be easily found among the fundamental forces in nature, such as the gravitational force, the electromagnetic force, and the nuclear force.

Considering an approximately circular (in reality, elliptical) motion of the Earth around the Sun, equating the gravitational force $F_\mathrm{g} = GM_\mathrm{s}M_\mathrm{e}/R^2$ (where G is the gravitational constant, the subscripts s and e denote Sun and Earth, respectively, and R is their relative distance) with the centripetal force $F_\mathrm{c} = M_\mathrm{e}v^2/R$, one obtains $v = \sqrt{GM_\mathrm{s}/R}$, leading to a kinetic energy $E_\mathrm{k} = GM_\mathrm{e}M_\mathrm{s}/(2R)$, which combined with the potential energy of $E_\mathrm{p} = -\,GM_\mathrm{e}M_\mathrm{s}/R$, results in a binding energy for the solar system of the form

$$E_\mathrm{b} = G\frac{M_\mathrm{e}M_\mathrm{s}}{2R}. \tag{1}$$

Using the relevant masses for the Earth and Sun and their separation, this quantity can be approximately calculated as $E_\mathrm{b} \approx 2.6 \times 10^{33}$ J ($\Delta m \approx 2.9 \times 10^{16}$ kg). However, what actually matters in terms of physical "observability" is the ratio of mass defect to the bound state mass (the closer this ratio is to 1, the greater the possibility of observing the mass defect). In the case of the gravitational system $\Delta M/M_\mathrm{b} \approx 1.5 \times 10^{-14}$; hence, the mass defect for the gravitational force will not be observable.

Similar calculations can be performed classically for the hydrogen atom (following the same path, but substituting $G \to 1/(4\pi\varepsilon_0)$, $M_\mathrm{s} \to |q_\mathrm{e}|$, and $M_\mathrm{e} \to q_\mathrm{e}$, where ε_0 is the dielectric constant in a vacuum and q_e is the charge of the electron). In this case, for the electrostatic force,

$$E_\mathrm{b} = \frac{1}{2}\frac{q_\mathrm{e}^2}{4\pi\varepsilon_0 R}. \tag{2}$$

In this case, however, $R \approx 0.53 \times 10^{-10}$ m (while in the previous example, it was $\approx 1.5 \times 10^{11}$ m!). In the case of the hydrogen atom, $E_\mathrm{b} \approx 13.6$ eV and $\Delta m \approx 2.5 \times 10^{-35}$ kg. The ratio $\Delta M/M_\mathrm{b} \approx 1.5 \times 10^{-8}$, indicating that in this case also it is not possible to observe the mass defect.

In the case of the nuclear force, however, the ratio of $\Delta M/M_\mathrm{b}$ is of order 10^{-3}, and hence it is possible to observe the mass defect. For example, the mass of an α particle consisting of two protons and two neutrons is 6.6447×10^{-27} kg, while the individual masses of these particles add up to 6.6951×10^{-27} kg, resulting in a binding energy of 28.3 MeV and $\Delta M/M_\mathrm{b} \approx 0.0075$. In fact, a very common diagram in nuclear physics is the so-called nuclear binding energy curve (see, e.g., http://hyperphysics.phy-astr.gsu.edu/hbase/nucene/nucbin.html), which shows the binding energy of various elements as a function of their mass number. In this graph, the larger the E_b, the more stable the element; iron (with atomic number $A = 56$ and binding energy 8.8 MeV/nucleon) is the most stable element. Lighter elements can yield energy by fission, while heavier elements can yield energy by means of fusion, emitting energies in the MeV range.

Binding Energy in Nonlinear Systems

Naturally, bound states of multiple waves can be formed in nonlinear systems. To fix ideas, we will examine such bound states and their corresponding binding energies in the specific context of the well-known sine-Gordon equation. For a detailed exposition of the features and applications of this equation, see Dodd et al. (1982). The sine-Gordon equation in (1 + 1) dimensions is

$$u_{tt} = u_{xx} - \sin(u). \tag{3}$$

Perhaps, the best-known nonlinear wave solution of this equation consists of the topological soliton (kink), which is of the form (in the static case)

$$u(x) = 4\tan^{-1}(\mathrm{e}^{sx}), \tag{4}$$

$s \in \{-1, 1\}$, where the case of $s = 1$ corresponds to a kink, while $s = -1$ corresponds to an antikink. The energy of such a static kink solution

$$E = \int_{-\infty}^{\infty} \mathrm{d}x \left[\frac{1}{2}(u_t^2 + u_x^2) + 1 - \cos(u) \right] \tag{5}$$

can be calculated as $E = 8$.

Another elemental solution of the equation is the breather-like solution of the form

$$u(x,t) = 4\tan^{-1}\left[\frac{(1-\omega^2)^{1/2}}{\omega} \frac{\sin(\omega t)}{\cosh\left((1-\omega^2)^{1/2} x\right)} \right]. \tag{6}$$

This exponentially localized in space, periodic in time solution can be considered as a result of a merger of a kink and an antikink. Hence, this is perhaps the simplest example of a bound state in this nonlinear system.

The bound state character of this solution can also be revealed by the expression for its energy. Using expression (6) in Equation (5), we obtain

$$E_{\text{breather}} = 16(1-\omega^2)^{1/2}. \tag{7}$$

Hence, this energy, for any $\omega \in (0, 1)$ (ω is the frequency of the internal breathing oscillation), is less than the sum of the kink and antikink energies, verifying that the binding energy of such a state is

$$E_{\mathrm{b}} = 16\left[1 - (1-\omega^2)^{1/2}\right]. \tag{8}$$

It is also worthwhile to note that the energy of such a breather excitation varies in the interval (0, 16) depending on its frequency. Hence, there is no threshold for the excitation of such a wave, but even for small amounts of energy, such waveforms will be excited (large frequency/small period ones for small excitation energy).

One can generalize the solution of the form (6) in a periodic breather lattice solution of the sine-Gordon equation in the form (see, e.g., McLachlan, 1994)

$$u(x,t) = 4\tan^{-1}[a\, \mathrm{sn}(bt, k^2)\mathrm{dn}(cx, 1-m^2)], \tag{9}$$

where $\mathrm{sn}(x, k)$ and $\mathrm{dn}(x, k)$ are the Jacobi elliptic functions with modulus k. Here

$$a = \sqrt{\frac{k}{m}},\ b = \sqrt{\frac{m}{(m+k)(1+mk)}} \text{ and}$$
$$c = \sqrt{\frac{k}{(m+k)(1+mk)}}.$$

One can then evaluate the energy (per breather) of this infinite periodic breather lattice configuration (for details of the calculation, the interested reader is directed to Kevrekidis et al., 2001) to be

$$E = 16\sqrt{\frac{k}{(k+m)(1+km)}}E(1-m^2), \tag{10}$$

where $E(1-m^2)$ is the complete elliptic integral of the second kind.

Depending on the values of the elliptic moduli, k and m, the expression of Equation (9) represents a lattice of different entities. For $m, k \to 0$, it corresponds to genuine sine-Gordon breathers. For $k, m \to 1$, the limit gives the "pseudosphere" solution $u = 4\tan^{-1}(\tanh \frac{t}{2})$, which resembles a π-kink but in time rather than space (see McLachlan, 1994). On the other hand, the $k \to$ finite, $m \to 0$ limit gives the kink-antikink pair solution $4\tan^{-1}(t \operatorname{sech} x)$. This solution has the character of a kink-antikink pair "breathing" in time.

The above different limits illustrate why Equation (10) is an important result, since it can be used (see below) to obtain the asymptotic interaction between entities such as breathers, pseudospheres, or kink-antikink pairs. When taking the appropriate above limits of expression (10), the leading-order term will be the energy of a single such entity. However, the correction to that will be the (per particle) binding energy in a configuration of multiple such entities (or, as it is often referred to, the energy of interaction between two such entities).

To calculate the breather-breather interaction (their binding energy), we take the limit $m, k \to 0$, with $k/m = (1-\omega^2)/\omega^2$, where ω is the breather frequency (see McLachlan, 1994), to obtain

$$\begin{aligned} E = \Bigg[& 16\sqrt{1-\omega^2} - 8m^2\frac{(1-\omega^2)^{3/2}}{\omega^2} \\ & + 6m^4\frac{(1-\omega^2)^{5/2}}{\omega^4} \Bigg] E(1-m^2). \end{aligned} \tag{11}$$

Hence, the corrections to the (single) breather energy in Equation (11) correspond to the binding energy of the formed breather bound states. Similar expressions can be found for the pseudosphere:

$$\begin{aligned} E = {} & 4\pi - \frac{\pi}{2}(k-1)^2 + \frac{\pi}{4}(m-1)^2 \\ & + \frac{3\pi}{32}(m-1)^2(k-1)^2 \end{aligned} \tag{12}$$

(again, 4π is the energy of a single pseudosphere) and for kink-antikink pairs

$$E = \left[16 - 8\left(k + \frac{1}{k}\right) + 2m^2\left(2 + 3k^2 + \frac{3}{k^2}\right)\right] \times E(1 - m^2). \qquad (13)$$

Similar examples of breather lattices also exist for other equations, such as the well-known Korteweg–de Vries (KdV) equation, and one can again infer breather-breather state binding energies in a similar manner.

In general, one can say that the concept of a bound state for nonlinear evolutionary partial differential equations supporting soliton (or solitary wave) solutions persists in a form very similar to the way it manifests itself for fundamental physical forces and their particle carriers. In the present case, the elements of the bound states are the nonlinear waves proper (a feature reminiscent of the particle-like character of such waves, manifest evidently also in their interactions). In a number of (most often integrable) cases, where the form of the bound state solutions is analytically tractable, the calculation of the bound state energy and of the energy of its constituent elements again provides information, through the difference between the two, for the binding energy (or energy of interaction) of such waves.

P.G. KEVREKIDIS

See also **Breathers; Partial differential equations, nonlinear; Sine-Gordon equation; Solitons**

Further Reading

Dodd, R.K., Eilbeck, J.C., Gibbon, J.D. & Morris, H.C. 1982. *Solitons and Nonlinear Wave Equations*, London: Academic Press

Kevrekidis, P.G., Saxena, A. & Bishop, A.R. 2001. *Physical Review* E, 64: 026613

McLachlan, R. 1994. *Math. Intelligencer*, 16: 31

BIOLOGICAL EVOLUTION

The term *evolution* defines a process that is driven by internal and/or external forces. In quantum mechanics, an evolution operator conducts change in time. Cosmic evolution in the standard model aims at a consistent description of the process from the "big bang" to the present universe. "Prebiotic evolution" deals with chemical precursors of present-day life and is determined by the conditions at the early Earth, be it in the primordial atmosphere, in the surrounding of volcanic hot springs at the sea floor, or at some other location. "Biological evolution" follows the prebiotic scenario, and it shaped and is still shaping the biosphere on Earth. A temporal change in the biosphere manifests itself as the appearance, alteration, and extinction of biological species. This view was not generally accepted before Charles Darwin. Influenced by the geologist Charles Lyell and his concept of uniformitarianism, Darwin and the proponents of the theory of evolution suggested that changes in the biosphere occur gradually, continuously, or at least, in small steps. In this aspect, which is not essential for the mechanism of evolution, Darwin's theory contrasted the view held by the majority of his contemporaries, who assumed constancy of biological species and change exclusively through catastrophic events leading to mass extinction (Ruse, 1979). The opponents of evolutionary thinking, Louis Agassiz, Georges Cuvier, and others, considered species as invariant entities. The remnants of extinct species in the fossil record were interpreted by them as witnesses from earlier worlds destroyed by punctual events, the great deluge, and other catastrophes that wiped out major parts of the organismic world. In society, the concept of evolution was heavily attacked by representatives of the Christian Churches because it was seen to be in conflict with the *Genesis* report in the Bible (Ruse, 2001). During the 20th century, European religious thought has reconciled religious belief and the idea of an evolving biosphere. In North America, the strong opposition of some groups of religious fanatics led to the peculiar development of Creationism, whose claim of being an alternative to the theory of evolution is rejected by the established scientific community (NAS, 1999).

The current theory of biological evolution originated from two epochal contributions by Charles Darwin and Gregor Mendel. Darwin conceived a mechanism for evolutionary change of the biosphere based on variation and selection, and he gathered empirical data providing evidence for the action of natural and artificial selection, the latter exercised in animal breeding and nursery gardens. Darwin's principle (published in *On The Origin of Species by Natural Selection* in 1859) has two consequences: species adapt to their environments and are related to their ancestors in terms of phylogenies, or branches of an ancestral tree of species. In 1866, Gregor Mendel introduced quantitative statistics into the evaluation of data in biology and performed the first precisely controlled fertilization experiments with plants. He discovered and interpreted correctly the action of genes in determining the properties of organisms. Mendel's work was considered irrelevant by the evolutionists of the second half of the 19th century and was "rediscovered" around 1900. Only in 1930 were the Darwinian concept of selection and Mendel's rules of inheritance combined to a common mathematical formalism by the population geneticists Ronald Fisher, John Haldane, and Sewall Wright (for a recent text in population genetics, see Hartl & Clark, 1997). In the 1940s, finally, Darwinian evolution and Mendelian genetics were united in the *Synthetic* or *Neo-Darwinian Theory of Evolution* by the works of the experimental biologists Theodosius

Dobzhansky, Julian Huxley, Ernst Mayr, and others (Mayr, 1997). In the second half of the 20th century, molecular biology put evolutionary theory on firm fundamentals, chemistry and physics. Comparison of genes and, more recently of whole genomes, allows for reconstruction of phylogenies on the basis of nucleotide sequence divergence through mutation (Judson, 1979); the exploration of molecular structures provides insights into the chemistry of present day life; and knowledge of biomolecular properties eventually led to the construction of laboratory systems that allow for observation of evolution of molecules in the test tube (Spiegelman, 1971; Watts & Schwarz, 1997).

Darwinian evolution results from the interplay of variation and selection, both being consequences of reproduction in populations. Variation operates on genomes or genotypes, which are polynucleotide sequences carrying the genetic information, and occurs in two fundamentally different ways: (i) mutation causes local changes in genomic sequences, whereas (ii) recombination exchanges corresponding segments between two genotypes. Selection is based on differences in *fitness* being a property of the phenotype. The phenotype is defined as the union of all, structural as well as dynamic, properties of an individual organism. Unfolding of the phenotype is programmed by the genome; but, at the same time, requires a highly specific environment. In addition, it is influenced by epigenetic factors (epigenetic refers to every nonenvironmental factor that interferes with the development of the organism, except those encoded in the nucleotide sequence of DNA; many epigenetic factors are already understood at the molecular level, and involve specific modifications of genomic DNA). Fitness, in essence, counts the number of fertile descendants reaching the reproductive age. It has two major components: (i) the probability of survival to reproduction, and (ii) the number of viable and fertile offspring.

To illustrate selection in a population of n asexually reproducing phenotypes, we consider a continuous-time model that describes change by a differential equation

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i(f_i - \Phi), \quad i = 1, \ldots, n$$

$$\text{with } \Phi(t) = \sum_{j=1}^{n} f_j x_j = \overline{f}. \tag{1}$$

The variables denote the frequencies of reproducing variants: $x_i(t) = N_i(t)/\sum_{j=1}^{n} N_j(t)$, with $N_i(t)$ counting the number of individuals with phenotype S_i or genotype I_i at time t. (For several genotypes giving rise to the same phenotype, see *neutrality* below.) Fitness values f_i when averaged over the entire population yield the mean fitness expressed by a time-dependent flux $\Phi(t)$. Frequencies of variants with fitness values above average, $f_i > \Phi$, increase with time, those of below-average variants, $f_i < \Phi$, decrease and as a consequence, the mean fitness increases. The flux $\Phi(t)$ is a nondecreasing function of time and selection continues until all variants, except the fittest, have died out (*See also* **Fitness landscape**). For two variants, I_0 and I_1, the solution boils down to

$$\left[\frac{x_1}{x_0}\right](t) = \left[\frac{x_1}{x_0}\right](0) \cdot \exp(mt)$$

or

$$\left[\frac{X_1}{X_0}\right]_t = \left[\frac{X_1}{X_0}\right]_0 w^t,$$

where the upper equation refers to continuously varying $x(t)$ and the lower equation refers to population to discrete time variables X_t with synchronized reproduction. The *Malthusian* fitness difference $m = f_1 - f_0$ is related to the *Darwinian* relative fitness $w = (1 + f_1)/(1 + f_2)$ by $m \approx \ln w$ (see Hartl & Clark (1997)). The conditions for selection are $m > 0$ or $w > 1$, respectively. An example is shown in Figure 1.

Sexual reproduction of diploid organisms involves Mendelian genetics (see Figure 2). Every gene (A) comes in two copies, identical or different, which are chosen from a reservoir of variants A_i called alleles. Recombination occurs in the process of reproduction when the two copies are separated and reassembled in pieces through random combination. The differential equation (1) is extended to describe selection in the diploid case in the form of Fisher's selection equation:

$$\begin{aligned}\frac{\mathrm{d}x_i}{\mathrm{d}t} &= x_i \cdot \left(\sum_{j=1}^{n} a_{ij} x_j - \Phi\right) \\ &= x_i\,(\overline{a_i} - \Phi), \quad i = 1, \ldots, n \end{aligned} \tag{2}$$

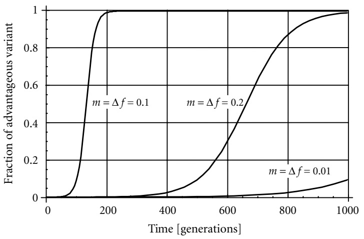

Figure 1. Illustration of selection in populations. The plotted curves represent the frequencies of advantageous mutants I_1 in a population of individuals I_0 with a Malthusian fitness difference of $m = \Delta f = f_1 - f_0 = 0.1$, 0.02, and 0.01. The population size is $N = 10{,}000$, and the mutants were initially present in a single copy: $N_1(0) = 1$ or $x_1(0) = 0.0001$.

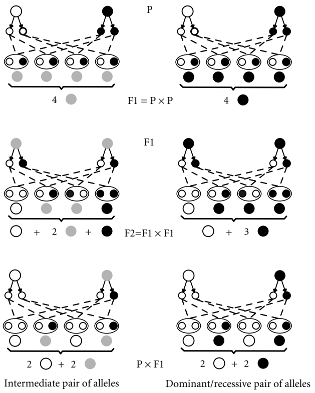

Figure 2. Mendelian genetics. In sexual reproduction, the two parental genomes are split into pieces and recombined randomly, which means each of the four alleles has a 50 % chance to be incorporated in the genome of an offspring. Mendel's laws are of a statistical nature and hold as mean values in the limit of large numbers of observations. Two cases are shown: (i) the heterozygote unfolds into a phenotype with intermediate properties (gray through blending of black and white), and (ii) the property of one allele (black) is dominant. In the latter case, the other allele (white) is called recessive. Interbreeding of two homozygous individuals (parent generation P) leads to a first offspring generation (F1) of identical heterozygous individuals; the phenotypes in the next (F2) generation show a distribution of 1:2:1 in the intermediate and 1:3 in the dominant/recessive case. Crossing of the (recessive) parent genotype with an F1-individual yields a 1:1 ratio of phenotypes.

with

$$\overline{a_i} = \sum_{j=1}^{n} a_{ij} x_j,$$

$$\Phi(t) = \overline{a} = \sum_{i=1}^{n} \overline{a_i} x_i = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

The variables refer to alleles A_j rather than to whole genomes, and the rate coefficient a_{ij} represents the individual fitness values for the combination $A_i A_j$. Fitness is assumed to be independent of the positioning of alleles, $A_i A_j$ or $A_j A_i$, and hence, $a_{ij} = a_{ji}$ holds. The term $\overline{a_i}$ is the population-averaged mean fitness of the allele combinations carrying A_i at least once: $A_i A_j$, $j = 1, \ldots, n$. Fisher's fundamental theorem states that the flux $\Phi(t) = \sum_{i=1}^{n} \overline{a_i} x_i = \overline{a}$ is a nondecreasing function of time, but the outcome of selection need not be unique as optimization might end in a local optimum of Φ (*See also* **Fitness landscape**). For example, in the two-allele case, inferiority of the *heterozygote* $A_1 A_2$, $a_{12} < \min\{a_{11}, a_{22}\}$, results in bistability since homogenous populations of either *homozygote*, $A_1 A_1$ or $A_2 A_2$, represent stable equilibrium points. Then, the initial conditions determine the outcome of selection.

The optimization principle is not universally valid: when mutation is included or when more complex cases of recombination are considered, optimization of mean fitness is restricted to certain ranges of initial conditions, whereas different behavior is observed for other starting values. Still, optimization remains an important heuristic in evolution as it is frequently observed.

Innovation is introduced into genes by mutation consisting of a local change in the sequence of nucleotides resulting from an imperfect replication of genetic information or externally caused damage. Two scenarios are distinguished: (i) rare mutation treated by conventional population genetics and typically occurring with multicellular organisms and most bacteria, and (ii) frequent mutation handled by quasispecies theory (Eigen, 1971; Eigen & Schuster, 1977) and determining evolution of viruses. Higher mutation rates are often advantageous because they allow for adaptation, but there exists an error threshold of replication beyond which inheritance breaks down because too many mutations destroy the genetic message. RNA viruses are under a strong selection constraint by the host and their mutation rates are close to the error threshold.

The idea that genotypes and phenotypes are related one-to-one turned out to be wrong. Molecular genetics revealed a high degree of *neutrality* (Kimura, 1983): many different genotypes give rise to the same phenotype. Advantageous mutations are rare; deleterious mutations are eliminated by selection thus leaving a majority of observed changes in the genomes to result from neutral mutations. Neutrality gives rise to random drift of populations in genotype space, which was also found to be important for the mechanism of evolution since it allows populations to escape from minor local fitness optima or evolutionary traps (Schuster, 1996) and Schuster in Crutchfield & Schuster, 2003). Random drift leads to an almost constant mutation rate per year and nucleotide independent of the species being tantamount to a *molecular clock of evolution*. This clock is used for dating in the reconstruction of phylogenies from comparison of present-day genome sequences. Molecular clock dates yield substantially longer time spans compared with those from the fossil record. The discrepancy seems to be reconcilable because paleontological datings are too young and molecular clock datings are too old by systematic errors (Benton & Ayala, 2003).

The Darwinian mechanism is powerful because it makes no reference to the specific nature of the reproducing entities. Therefore, it is likewise valid for molecules, viruses, bacteria, or higher organisms. Selection based on the Darwinian principle is observed in many disciplines outside biology, for example, in physics and chemistry, in economics, and in the social sciences.

Since its introduction, the theory of evolution has undergone changes and modifications. The rejection of catastrophic events as an important source of change in the history of life on Earth was a political issue rather than one based on scientific data. Geological evidence for fallings of large meteorites as well as major floods is now available, and such events wiped out substantial parts of the biosphere. The paleontological record reflects the interplay between continuous evolution and external influences, which resulted in epochs of gradual development interrupted by punctuated events. Interestingly, evolution of bacteria or molecules under constant conditions also showed punctuation without external triggers: populations "wait" during quasistationary periods for rare mutations that initiate fast periods of change.

Still, there are open problems in current evolutionary theory. Recent sequence data challenge the idea of a tree of life. Although animal phylogeny appears to be on a firm basis, there are problems with the reconstruction of a tree-like history of plant species. Prokaryote evolution cannot be cast into a tree: archebacteria and eubacteria exchange genetic information across species and kingdoms. Such horizontal gene transfer occurs frequently and obscures the descendance of species. Darwinian evolution, although successful in describing the mechanisms of optimization and adaptations of species, is unable to provide explanations for the major evolutionary transitions that lead from one hierarchical level of life to the next higher forms (Maynard Smith & Szathmáry, 1995; Schuster , 1996). Examples of such transitions are the origin of the genetic code; the transition from the prokaryotic to the eukaryotic cell; the transition from unicellular organisms to multicellular plants, fungi, and animals; the transition from solitary animals to animal societies; and eventually the transition to man and human societies. Common to all these transitions is the integration of individual competitors as cooperating elements into a novel functional unit. Simple model mechanisms have been proposed that can explain cooperation of competitors (see, e.g., the hypercycle Eigen & Schuster, 1978), but no real solution to the problem has been found yet.

PETER SCHUSTER

See also **Catalytic hypercycle; Fitness landscape**

Further Reading

Benton, M.J. & Ayala, F.J. 2003. Dating the tree of life. *Science*, 300: 1698–1700

Crutchfield, J.P. & Schuster, P. (editors). 2003. *Evolutionary Dynamics: Exploring the Interplay of Selection, Accident, Neutrality, and Function*, Oxford and New York: Oxford University Press

Eigen, M. 1971. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58: 465–523

Eigen, M. & Schuster, P. 1977. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*, 64: 541–565

Eigen, M. & Schuster, P. 1978. The hypercycle. A principle of natural self-organization. Part B: The abstract hypercycle. *Naturwissenschaften*, 65: 7–41

Hartl, D.L. & Clark, A.G. 1997. *Principles of Population Genetics*, 3rd edition, Sunderland, MA: Sinauer Associates

Judson, H.F. 1979. *The Eighth Day of Creation. The Makers of the Revolution in Biology*, London: Jonathan Cape and New York: Simon and Schuster

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*, Cambridge and New York: Cambridge University Press.

Maynard Smith, J. & Szathmáry, E. 1995. *The Major Transitions in Evolution*, Oxford and New York: Freeman

Mayr, E. 1997. The establishment of evolutionary biology as a discrete biological discipline. *BioEssays*, 19: 263–266

National Academy of Sciences (NAS). 1999. *Science and Creationism. A View from the National Academy of Sciences*, 2nd edition, Washington, DC: National Academy Press

Ruse, M. 1979. *The Darwinian Revolution*, Chicago, IL: University of Chicago Press

Ruse, M. 2001. *Can a Darwinian Be a Christian? The Relationship Between Science and Religion*, Cambridge and New York: Cambridge University Press

Schuster, P. 1996. How does complexity arise in evolution? *Complexity*, 2(1): 22–30

Spiegelman, S. 1971. An approach to the experimental analysis of precellular evolution. *Quarterly Reviews of Biophysics*, 4: 213–253

Watts, A. & Schwarz, G. (editors). 1997. *Evolutionary Biotechnology—From Theory to Experiment. Biophyscial Chemistry*, vol. 66, nos. 2–3, Amsterdam: Elsevier, pp. 67–284

BIOMOLECULAR SOLITONS

Biological molecules are complex systems that evolve in an ever-changing environment and nevertheless exhibit a remarkable stability of their functions. This feature, which is reminiscent of the exceptional stability of solitons in the presence of perturbations, is perhaps what led to suggestions that solitons could have a role in some biological functions. Beyond this analogy, there are more solid arguments to consider the role of nonlinearity in biological molecules. They are very large atomic assemblies performing their function through large conformational changes, which have a cooperative character because they involve many atoms moving in a coherent manner, and are highly nonlinear due to their amplitude of motion, which is much larger than the standard thermal motions observed in small molecules. Additional nonlinearities can originate from the coupling of different degrees of freedom, as proposed for proteins. Dispersion, necessary to balance the effect of nonlinearity in order to obtain soliton-like excitations, is introduced by the discreteness of the molecular lattice, which behaves in a manner different from a continuous medium.

Besides conformational changes involved in many biological functions, issues important for biological molecules are energy transport and storage, and charge transport. Nonlinear excitations have been proposed as possible contributors to these phenomena, in the two main classes of biological molecules: nucleic acids and proteins (Peyrard, 1995).

Following Erwin Schrödinger in his prophetic book *What Is Life?* (Schrödinger, 1944), the nucleic acid DNA can be viewed as an "aperiodic crystal." The static structure of DNA is a fairly regular pile of flat base pairs, linked by hydrogen bonds and connected by sugar–phosphate strands that form a double helix (Calladine & Drew, 1997). The lack of periodicity occurs because the base pairs can be either $A-T$ (adenine–thymine) or $G-C$ (guanine–cytosine), their sequence defining the genetic code. The static picture that emerges from crystallographic data has little to do with actual DNA in a living cell. The genetic code, buried in the double helix, would not be accessible if DNA were not a highly dynamical structure. Biologists have observed the "breathing of DNA," which is a fluctuational opening in which one or a few base pairs open temporarily. These motions are probed experimentally by monitoring deuterium-hydrogen exchange, based on the assumption that the imino-protons that bind the bases can only be exchanged for open base pairs. DNA double helix is also opened by enzymes during the transcription of a gene, that is, the reading of the genetic code.

This phenomenon is complex, but there are related experimental observations that are more amenable to physical analysis, on the thermal denaturation of DNA. When the double helix is heated, one first observes local openings over a few to a few tens of base pairs. These grow and invade the whole molecule, leading to a thermal separation of the two strands, which can be monitored by measuring the UV absorbance of the molecule, which is highly sensitive to the disturbance of the base stacking. This "DNA melting"—which appears as a phase transition in one dimension—poses challenging questions because, in order to cause the local openings, one has to break many hydrogen bonds between the bases, which requires the localization of a large amount of thermal energy in a small region of the molecule. Nonlinear effects could be at the origin of this phenomenon.

All these observations led to many investigations and models of the nonlinear dynamics of the DNA molecule. A description at the scale of the individual atoms is not necessary to analyze base-pair openings, so the bases are generally described as rigid objects in these models. The earliest attempt to describe DNA opening in terms of solitons is due to Englander et al. (1980), who viewed it as a cooperative motion involving 10 or more base pairs and propagating as a localized defect along the molecule. This idea was further formalized by Yomosa (1984), and Takeno & Homma (1983), who introduced a coupled base rotator model for the structure and dynamics of DNA. The general ideas behind this approach are schematized in Figure 1(a).

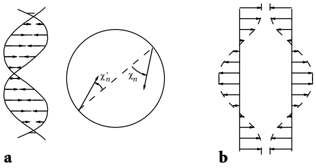

Figure 1. (a) Schematic view of the plane base rotator model for DNA. (b) A symmetric open state of the model.

Only rotational degrees of freedom of the bases are introduced (denoted by angles χ_n and χ'_n in Figure 1). The pairing of the bases is described by an on-site potential $V(\chi_n, \chi'_n)$, which, in its simplest form is $V(\chi_n, \chi'_n) = A(1 - \cos\chi_n) + A(1 - \cos\chi'_n) + B(1 - \cos\chi_n \cos\chi'_n)$, and the stacking along the molecule is represented by a potential $W(\chi_n, \chi'_n, \chi_{n-1}, \chi'_{n-1} = S[1 - \cos(\chi_n - \chi_{n-1})] + S[1 - \cos(\chi'_n - \chi'_{n-1})]$, where A, B, S are constants. Adding the kinetic energy of the bases $\frac{1}{2}I(\dot{\chi}_n^2 + \dot{\chi}_n'^2)$, where I is the moment of inertia of the bases around their rotation axis, and summing over n, one obtains the Hamiltonian of the model. Various nonlinear excitations are possible depending on the symmetries of the motion (such as $\chi_n = \chi'_n$, $\chi_n = -\chi'_n$) and the values of the constants. If the stacking interaction is strong enough, a continuum approximation can be made. This approximation replaces the discrete variables $\chi_n(t)$ by the function $\chi(x, t)$ and finite differences such as $\chi_n - \chi_{n-1}$ by derivatives $a(\partial\chi/\partial x)$, where a is the spacing between the bases and x denotes the continuous coordinate along the helix axis. When $A = 0$, in its simplest form, the model leads to a sine-Gordon equation

$$I\frac{\partial^2\chi}{\partial t^2} - Sa^2\frac{\partial^2\chi}{\partial x^2} + B\sin\chi = 0, \qquad (1)$$

which has topological solutions such as the one schematized in Figure 1(b), where the bases undergo a 2π rotation, generating an open state that may slide along the chain.

Models for the rotation of the base pairs have been further refined by Yakushevich (1998) and are discussed in the entry on **DNA solitons**.

Another point of view was chosen later by Dauxois et al. (1993), who were interested in the statistical physics of DNA thermal denaturation. This problem had been studied by Ising models, which simply use a two-state variable equal to 0 or 1 to specify whether a base pair is closed (0) or open (1). Such models cannot describe the intermediate states, but they can be generalized by introducing a real variable $y_n(t)$ that measures the stretching of the hydrogen bonds in a base pair that is equal to 0 in the equilibrium structure and grows to infinity when the two strands are fully separated. With such a variable, a natural shape of the on-site potential is the Morse potential $V(y_n) = D[\exp(-\alpha y_n) - 1]^2$ (D and α are constants), which has a minimum corresponding to the binding of the two bases in their equilibrium state by the hydrogen bonds, and a plateau at large y_n, which is associated to the vanishing of the pairing force $\partial V/\partial y_n$ when the bases are far apart.

Such a model does not have topological solitons, but its nonlinear dynamics leads to localized oscillatory modes, called breathers, which are approximately described by solitons of the nonlinear Schrödinger equation in the continuum limit and turn into permanently open states at a high temperature (*See* **Breathers**). These studies have focused attention on the importance of discreteness for nonlinear energy localization. In DNA, the stacking interactions are not very strong, and this is why imino-proton exchange experiments can detect the exchange on one base pair while the neighboring base pairs are not affected. As a result a continuum approximation is very crude. What could appear as a problem because it complicates analytical studies of the nonlinear dynamics turns out to have a far-reaching consequence because it has been shown that discreteness is *crucial* for the existence and formation of nonlinear localized modes (Sievers & Takeno, 1988; MacKay & Aubry, 1994), which correspond to the “breathing” of DNA observed by biologists. It is important to notice that the existence of these nonlinear solutions is not linked to a particular mathematical expression of the potentials. Instead, it is a generic feature of nonlinear lattices having interactions qualitatively similar to those that connect the bases in DNA. Moreover, it has also been shown that thermal fluctuations can self-localize in such lattices so it is likely that related nonlinear excitations could exist in DNA. But discreteness has another consequence. Large-amplitude modes are strongly localized due to their high nonlinearity. Their width becomes of the order of the spacing between the bases and they lose the translational invariance of solitons in continuum media. The image of freely moving solitons has to be corrected by the pinning effect of discreteness, and the translation of the nonlinear excitations in DNA, if it occurs, has to be activated, for instance, by thermal fluctuations.

Proteins are much more complex than DNA because they do not have a quasi-periodic structure, but some of their substructures are nevertheless fairly regular. They are biological polymers composed of amino acids of the general formula

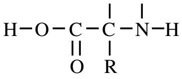

$$\mathrm{H{-}O{-}\underset{\underset{\displaystyle O}{\|}}{C}{-}\overset{\overset{\displaystyle H}{|}}{\underset{\underset{\displaystyle R}{|}}{C}}{-}\overset{\overset{\displaystyle H}{|}}{N}{-}H}$$

where *R* is an organic radical that determines the amino acid. These building blocks are linked by a peptide bond that can be viewed as a result of the elimination of a water molecule between consecutive amino acids, leading to the generic formula

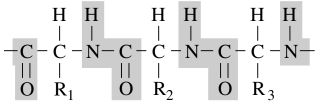

$$\mathrm{{-}\underset{\underset{\displaystyle O}{\|}}{C}{-}\overset{\overset{\displaystyle H}{|}}{\underset{\underset{\displaystyle R_1}{|}}{C}}{-}\overset{\overset{\displaystyle H}{|}}{N}{-}\underset{\underset{\displaystyle O}{\|}}{C}{-}\overset{\overset{\displaystyle H}{|}}{\underset{\underset{\displaystyle R_2}{|}}{C}}{-}\overset{\overset{\displaystyle H}{|}}{N}{-}\underset{\underset{\displaystyle O}{\|}}{C}{-}\overset{\overset{\displaystyle H}{|}}{\underset{\underset{\displaystyle R_3}{|}}{C}}{-}\overset{\overset{\displaystyle H}{|}}{N}{-}}$$

A given protein is defined by its sequence of amino acids chosen by 20 possible types and the length of the chain (typically 150–180 residues), but this so-called primary structure does not determine the function that depends on the spatial organization of the residues. Segments of the chain tend to fold into secondary structures having the shape of helices (called α-helices) or sheets (called β-sheets) stabilized by hydrogen bonds formed mainly between the negatively charged $C=O$ groups and the positively charged protons linked to the nitrogen atom of a peptide bond. The different components of the secondary structure assemble in the tertiary structure, which is the functional form of the protein.

Proteins perform numerous functions and one of them is the storage and transport of the energy released by the hydrolysis of adenosine-triphosphate (ATP), which plays the role of the fuel necessary for many biological processes, such as muscle contraction. The hydrolysis of a single ATP molecule releases approximately 0.4 eV, which is transmitted to a protein for later use. This raises a puzzling question because, if this energy were distributed among all the degrees of freedom of a protein, each atom would carry such a small amount that the energy would be useless. There must be a mechanism that maintains this energy sufficiently localized, and moreover, as it will not be used at the site where it has been released, it must be transported efficiently within the molecule. Recent experiments at the molecular scale have shown that the hydrolysis of a single ATP molecule can be used for *several* steps of a molecular motor involved in muscle contraction (Kitamura et al., 1999), providing evidence of the temporary storage of the energy.

Attempting to understand these phenomena in 1973, Alexander Davydov noticed that the energy released by ATP hydrolysis almost coincides with 2 quanta of the vibrational energy of the $C=O$ bond, which led him to the conclusion that this energy was stored as vibrational energy in the peptide bond (Scott, 1992). He conjectured that it could stay localized through an extrinsic nonlinearity associated with a distortion of the chain of hydrogen bonds that spans the α-helix. The underlying mechanism is similar to the one leading to the formation of a polaron in solid-state physics (Ashcroft & Mermin, 1976). The vibration of the $C = O$ bond strains the lattice in its vicinity, resulting in slight displacements of the neighboring amino acids. But, as the frequency of the $C = O$ vibration is affected by its interactions with the neighboring atoms, the frequency of the excited $C = O$ bond becomes slightly shifted and no longer coincides with the resonating frequencies of the neighboring $C = O$ bonds, preventing an efficient transfer of energy to the neighboring sites. As a consequence, the energy released by the ATP hydrolysis does not spread along the protein.

Therefore, the basic idea behind the mechanism proposed by Davydov is nonlinear energy localization due to the shift of the frequency of an oscillator when it is excited. For the protein, it is not due to an intrinsic nonlinearity of the $C = O$ bond (as was the case for the Morse potential linking the bases in a pair for DNA), but due to a coupling with another degree of freedom, which is an acoustic mode of the lattice of amino acids connected by hydrogen bonds.

As only a few quanta of the $C = O$ vibrational motion are excited, the theory cannot ignore quantum effects, and in order to go beyond the qualitative picture discussed above, one has to solve the time-dependent Schrödinger equation. Davydov proposed a simple *ansatz* to describe the quantum state of the system. In this simple approximation, the motion of the self-trapped energy packet is described by a discrete form of the nonlinear Schrödinger equation. When one introduces proper parameters for the α-helix, the calculation of the soliton width shows that it is much broader than the lattice spacing, which should allow its motion without pinning by discreteness. As a result, energy transfer by solitons in the α-helix is plausible, but a definitive conclusion about the existence of such solitons is still pending. This is because the role of thermal fluctuations, which could destroy the coherence of the lattice distortion around the excited $C = O$ site and hence the self-trapping, and the extent of quantum effects are hard to evaluate quantitatively (Peyrard, 1995). A direct experimental observation on a protein has not been possible up to now.

These uncertainties prompted physicists and physical chemists to experimentally investigate model systems that are simpler than proteins but, nevertheless, show chemical bonds comparable to the peptide bonds in proteins. Crystalline acetanilide consists of quasi-one-dimensional chains of hydrogen-bonded peptide groups. In the early 1980s, it was recognized by spectroscopic studies that the $C = O$ stretching and $N-H$ stretching bands of crystalline acetanilide exhibit anomalies, and tentative explanations involve self-trapped states similar to the Davydov solitons. These ideas have been confirmed by recent pump–probe experiments (Edler et al., 2002). A direct observation of self-trapping could be achieved, and it appears that the crystal structure is essential to stabilize the excitation that decays 20 times faster for isolated molecules than for molecules linked by hydrogen bonds in the crystal. Although the lifetime of the self-trapped state (20 ps) is shorter than expected by Davydov, this study supports the original idea of the importance of the coupling with the lattice degrees of freedom. A possible translational motion of the self-trapped state and its possible role for biological functions are still open questions.

The *News and Views* section of the journal *Nature* attests that nonlinear excitations in biomolecules have been the object of strong controversy, ranging from enthusiastic approval (Maddox, 1986, 1989) to harsh criticisms (Frank-Kamenetskii, 1987), which were justified by some of the overstatements by theoreticians. Today, passionate opinions have subsided and experiments at the scale of a single molecule have become feasible, showing us how biomolecules work or take their shape. Thus, it appears likely that while freely moving solitons along DNA or protein α-helices may not exist, nonlinear excitations leading to energy localization or storage, and perhaps transport, could well provide useful clues to understand some of the phenomena occurring in biomolecules.

MICHEL PEYRARD

See also **Davydov soliton; DNA premelting; DNA solitons; Pump-probe measurements**

Further Reading

Ashcroft, N.W. & Mermin, D.A. 1976. *Solid State Physics*, Philadelphia: Saunders Company

Calladine, C.R. & Drew, H.R. 1997. *Understanding DNA: The Molecule and How It Works*, 2nd edition, San Diego and London: Academic Press

Dauxois, T., Peyrard, M. & Bishop, A.R. 1993. Dynamics and thermodynamics of a nonlinear model for DNA denaturation. *Physical Review E*, 47: 684–695 and R44–R47

Edler, J., Hamm, P. & Scott, A.C. 2002. Femtosecond study of self-trapped vibrational excitons in crystalline acetanilide. *Physical Review Letters*, 88 (1–4): 067403

Englander, S.W., Kallenbach, N.R., Heeger, A.J., Krumhansl, J.A. & Litwin, S. 1980. Nature of the open state in long polynucleotide double helices: Possibility of soliton excitations. *Proceedings of the National Academy of Sciences USA*, 777: 7222–7227

Frank-Kamenetskii, M. 1987. Physicists retreat again. *Nature*, 328: 108

Kitamura, K., Tokunaga, M., Iwane A.H. & Yanagida, T. 1999. A single myosin head moves along an actin filament with regular steps of 5.3 nanometres. *Nature*, 397: 129–134

MacKay, R.S. & Aubry, S. 1994. Proof of existence of breathers for time-reversible or Hamiltonian networks of weakly coupled oscillators. *Nonlinearity*, 7: 1623–1643

Maddox, J. 1986. Physicists about to hi-jack DNA? *Nature*, 324: 11

Maddox, J. 1989. Towards the calculation of DNA. *Nature*, 339: 577

Peyrard, M. (editor). 1995. *Nonlinear Excitations in Biomolecules*, Berlin and New York: Springer

Schrödinger, E. 1944. *What is Life? The Physical Aspect of the Living Cell*, Cambridge: Cambridge University Press

Scott, A.C. 1992. Davydov's soliton. *Physics Reports*, 217: 1–67

Sievers, A.J. & Takeno, S. 1988. Intrinsic localized modes in anharmonic crystals. *Physical Review Letters*, 61: 970–973

Takeno, S. & Homma, S. 1983. Topological solitons and modulated structures of bases in DNA double helices. *Progress of Theoretical Physics*, 70: 308–311

Yakushevich, L.V. 1998. *Nonlinear Physics of DNA*, Chichester and New York: Wiley

Yomosa, S. 1984. Solitary excitations in deoxyribonucleic acid (DNA) double helices. *Physical Review* A, 30: 474–480

BIONS

See **Breathers**

BIRGE–SPONER RELATION

See **Local modes in molecules**

BIRKOFF–SMALE THEOREM

See **Phase space**

BISTABILITY

See **Equilibrium**

BISTABLE EQUATION

See **Zeldovich–Frank-Kamenetsky equation**

BJERRUM DEFECTS

Ice is the most common and important member in a class of solids in which the conduction of electricity is carried almost exclusively by protons. Its dc electrical conductivity reaches the level of $10^{-7}\ \Omega^{-1}\ \mathrm{m}^{-1}$, placing ice among the semiconductors. Known protonic semiconductors include disordered forms of solid water and several salt hydrates and gas hydrates. In the case of lithium hydrazinium sulfate ($LiN_2H_5SO_4$), one finds quasi-one-dimensional hydrogen-bonded chains (HBCs) along the *c*-crystallographic axis. Electrical conductivity is three orders of magnitude larger in the *c*-direction (compared with the perpendicular directions), demonstrating that proton conductivity is directly related to the presence of hydrogen bonds. In addition to inorganic crystals, protonic conductivity plays a significant role in biological systems, where it participates in energy transduction and formation of proton pumps. Of particular significance is proton transport across cellular membranes through the use of hydrogen-bonded side-chains of proteins embedded in membrane pores.

As is shown in Figures 1 and 2, protonic conductivity in HBCs takes place through ionic defects and bonding or "Bjerrum" defects (named after Danish physical chemist Niels Bjerrum). Ionic defects are formed by an excess proton (H_3O^+) or a proton vacancy (HO^-), while bonding defects are misfits in the orientations of neighboring atoms resulting in either vacant bonds (L defect) or placing two protons in the same bond (D defect). Bonding defects do not obey the Bernal–Fowler rule of one proton per hydrogen bond for the ideal ice crystal.

When a proton is transported along an HBC through an ionic defect, after the passage of the proton, the chain remains blocked to further proton movement since all chain protons have been moved, say, from the left-hand to the right-hand side of each hydrogen bond. The chain gets unblocked through cooperative rotations, that is, through the passage of a bonding defect. Thus, protons move in an HBC through coordinated ionic and Bjerrum defects, using a mechanism that is also found in hydrogen-bonded protein side chains. Coordinated proton transport in biological macromolecules leads to the formation of proton pumps that channel protons across membranes and, through reversals, produce cyclic motor actions of mechanical nature. Defects in HBCs are topological in nature, with the rotational activation energy being smaller than that of ionic defect energy. The total charge of the topological ionic and bonding defects is not the same; in ice, the ionic defect charge is $e_I = 0.64e$ (e is the proton charge), while the bonding defect charge is $e_B = 0.36e$. Only after a coordinated passage of an ionic and bonding defect is one entire proton charge transferred across the HBC.

A simple one-dimensional cooperative model of an HBC is similar to the Frenkel–Kontorova model but with two alternating barriers modeling bonding and ionic activation energies. The minima separating the barriers correspond to equilibrium positions of protons that interact mutually through dipole-dipole interactions. In equilibrium under this model, there is initially one proton per hydrogen bond; transitions of protons over the large barriers correspond to ionic defects, while bonding defects result from transitions over the smaller barrier. Both classes of defects are modeled through topological solitons. There are two kink solutions corresponding to HO^- and L-bonding defects, while the corresponding antikinks are the H_3O^+ and D-bonding defects, respectively.

This simple model can be made quantitative through the introduction of the one-dimensional Hamiltonian

$$H = \sum_n \left[\frac{p_n^2}{2} + \frac{1}{2}(u_{n+1} - u_n)^2 + \omega V(u_n) \right], \quad (1)$$

where u_n and p_n are the dimensionless displacement from an equilibrium position and momentum, respectively, of the nth hydrogen that is coupled to its nearest

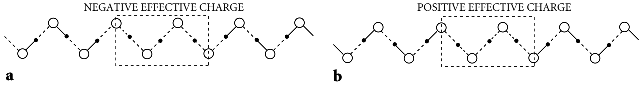

Figure 1. Ionic defects present in a hydrogen bonded chain: (a) negative ionic defect HO^- and (b) positive ionic defect H_3O^+. Large open circles denote ions, for example, oxygen ions in ice, while small black dots are protons. A hydrogen bond that links two ions contains a covalent part (solid line) that places in equilibrium the proton closer to one of the two oxygens. In the ionic defect region, there is a gradual transition in the equilibrium locations of protons within the hydrogen bonds; this transitional region is modeled through a topological soliton.

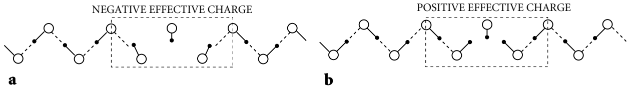

Figure 2. Bonding defects in a hydrogen bonded chain: (a) negative bonding defect (L) and (b) positive bonding defect (D). Molecular rotations introduce additional protons or remove protons from the quasi-one-dimensional HBC and produce bonding defects.

neighboring protons through harmonic spring interaction, while ω sets the energy scale. A typical choice for $V(u_n)$, the nonlinear substrate potential that models the ionic and bonding barriers, is

$$V(u_n) = \frac{2}{1-\alpha^2}\left[\cos(\frac{u_n}{2}) - \alpha\right]^2. \qquad (2)$$

The substrate potential (2) is periodic and (for appropriate values of the parameter α) is doubly periodic with two distinct alternating maxima that separate degenerate minima. In this model, one assumes one proton per unit cell, the latter consisting of the larger ionic barrier with its adjacent minima, one on each side. In the strongly cooperative limit, where neighboring hydrogen displacements do not differ substantially, one obtains for the proton displacement $u(x, t)$ that becomes a function of the continuous space variable x as well as time t the double sine-Gordon partial differential equation:

$$u_{tt} - c^2 u_{xx} + \frac{\omega^2}{1-\alpha^2}\left[-\sin u + 2\alpha \sin\frac{u}{2}\right] = 0, \qquad (3)$$

where c is the speed of sound of the linearized lattice oscillations. This sine-Gordon model has as solutions two sets of soliton kinks as well as their corresponding antikinks representing L-Bjerrum (kink I), D-Bjerrum (antikink I), HO^- ionic (kink II) and H_3O^+ (antikink II). More complex nonlinear models can be constructed that also include an acoustic interaction between neighboring ions as well as coupling of protons with ions. In these cases, one obtains two component solitons where the defects in the proton sublattice are topological solitons that induce a polaronic-like deformation in the ionic lattice. This more complex defect can travel along an HBC when an external electric field is applied in the system. Numerical simulations demonstrate that these nonlinear defects do indeed encompass some of the basic dynamical properties of the ionic and bonding defects found in hydrogen-bonded networks.

G.P. Tsironis

See also **Frenkel–Kontorova model; Hydrogen bond; Sine-Gordon equation; Topological defects**

Further Reading

Hobbs, P.V. 1974. *Ice Physics*, Oxford: Clarendon Press

Pnevmatikos, St. 1988. Soliton dynamics of hydrogen-bonded networks: a mechanism for proton conductivity. *Physical Review Letters,* 60: 1534–1537

Pnevmatikos, St., Tsironis, G.P. & Zolotaryuk, A.V. 1989. Nonlinear quasiparticles in hydrogen-bonded systems. *Journal of Molecular Liquids*, 41: 85–103

Savin, A.V, Tsironis, G.P. & Zolotaryuk, A.V. 1997. Reversal effects in stochastic kink dynamics. *Physical Review* E, 56: 2457–2466

Zolotaryuk, A.V., Pnevmatikos, St. & Savin, A.V. 1991. Charge transport by solitons in hydrogen-bonded materials. *Physical Review Letters*, 67: 707–710

BLACK HOLES

A massive body like the Earth is characterized by an "escape velocity," which is the speed that a moving particle must have on leaving the surface of the body to leave the attraction of gravity. Consider a bullet at the surface of the Earth that is moving upward with a speed of 11.2 km/s and neglect atmospheric friction. Such a bullet will just escape Earth's gravitational field by exchanging its kinetic energy for the potential energy of the gravitational field. If the mass of the Earth were compressed into a smaller radius, this escape velocity would be larger, because the gravitational energy to be overcome by the kinetic energy of the bullet would be larger.

In 1916, Karl Schwarzschild used Einstein's gravitational field equations to show that if a body of mass m is compressed to a radius

$$r_{\mathrm{s}} = \frac{2Gm}{c^2}, \qquad (1)$$

where G is the gravitational constant, then an object traveling at the speed of light c will be unable to escape the influence of gravity (Schwarzschild, 1916). For a body having the mass of the Earth, this "Schwarzschild radius" is about 1 cm, and for the Sun, it is about 3 km. Interestingly, Schwarzschild's idea was first suggested in the 18th century (Mitchell, 1783; Laplace, 1796).

The term "black hole" was coined by John Archibald Wheeler in 1967 to denote a cosmic object with its mass concentrated within the Schwarzschild radius. Neither particles nor light can overcome gravitational attraction and travel outside the sphere of radius r_{s}. Interestingly, Stephen Hawking (1974) has shown that—due to quantum fluctuations—a black hole should radiate as a black body with the temperature

$$T = \frac{hc}{4\pi k r_{\mathrm{s}}}, \qquad (2)$$

where k and h are, respectively, the Boltzmann and Planck constants. Indirect evidence for black holes is provided by the fact that there do not exist stable cold stars with masses larger than about three Sun masses. Under its own gravitational field, according to theory, such a star should collapse into a black hole. In 1931, Subrahmanyan Chandrasekhar was the first to conclude that above some critical mass of white dwarfs, the equation of state (of a quantum relativistic gas of degenerate Fermi particles) is too weak to counter the gravitational forces, leading to the formation of black holes. Both Lev Landau (1932) and Arthur Eddington (1924) rejected this implication of relativistic quantum mechanics rather than accept the possibility of black holes. Albert Einstein also concluded that Schwarzschild singularities do not exist in the real world (Einstein, 1939).

In 1939, however, J. Robert Oppenheimer and his colleagues used general relativity (rather than Newtonian gravity) to show that when all thermonuclear sources of energy are exhausted (with no further outward pressure due to radiation), a sufficiently heavy star will continue to contract indefinitely, never reaching equilibrium (Oppenheimer & Volkoff, 1939; Oppenheimer & Snyder, 1939). Oppenheimer et al. further noted that if one considers stellar collapse from the inside, a stationary observer sees the stellar surface moving inward to the Schwarzschild sphere and finally sees the surface freeze as it nears the Schwarzschild sphere. Moreover, they showed that observers who move inward with the collapsing matter do not observe such freezing; these observers could cross the critical surface ("event horizon") after a finite time on their own clocks, after which they have no possibility of sending a signal that could be detected by an observer located outside the collapsing matter.

Recently, the scientific history of black holes has been characterized by a rapid growth of observational, theoretical, and mathematical studies, in which the discovery of such compact objects becomes the main purpose (Thorne et al., 1986). Currently, the most important classes are black holes of stellar masses (about 3–10 solar masses) and super-massive black holes. The most convincing candidates for stellar black holes are binary X-ray sources, one component of which is an ordinary star and the other component is a black hole or neutron star (Novikov & Zeldovich, 1966). Estimates of the masses of compact objects in these systems are essentially greater than three solar masses, and one example of such a system is Cygnus X-1 (V 1357 Cyg). The present number of systems mentioned as possible candidates for black holes with stellar masses is about 20, all of which are X-ray sources in binary systems (Novikov & Frolov, 2001). In the case of super-massive black holes and nuclei of Seyfert galaxies, interpretations of the observable effects using the black hole theory seem the most simple and natural; for example, Galaxy M 31 is a black hole candidate having a mass of about 3×10^7 the Sun's mass (Novikov & Frolov, 2001).

Presently, the concept of black holes continues to be confirmed by direct observations and is used to explain observable astronomical effects related to exceptionally strong emission of energy. Thus, it is expected that in the future new astronomical objects will be detected near black holes, and new physical phenomena will be discovered that can be interpreted using the black hole concept. Along these lines, it is interesting to note recent work in which concepts from thermodynamics and information theory (such as temperature and entropy) are connected with black holes based on Hawking's ideas (Markov, 1965; Hawking, 1977). Also of interest are "artificial black holes," which do not compress a large amount of mass into a small volume, but reduce the speed of light in a moving medium to less than the speed of the medium, thereby creating an event horizon (Leonhardt, 2001).

Viatcheslav Kuvshinov

See also **Binding energy; Einstein equations; General relativity**

Further Reading

Chandrasekhar, S. 1931. The maximum mass of ideal white dwarfs. *Astrophysical Journal*, 74: 81

Eddington, A. 1924. A comparison of Whitehead's and Einstein's formulae. *Nature*, 113: 192

Einstein, A. 1939. On a stationary system with spherical symmetry consisting of many gravitating masses. *Annals of Mathematics*, 40: 922

Hawking, S.W. 1974. Black hole explosions. *Nature*, 248: 30–31

Hawking, S.W. 1977. Gravitational instantons. *Physics Letters* A, 60: 81
Landau, L.D. 1932. On the theory of stars. *Physikalische Zeitschrift der Sowjetunion*, 1: 285
Laplace, P.-S. 1796. *Exposition du système du monde*, [Description of the World System], Paris
Leonhardt, U. 2001. A laboratory analogue of the event horizon using slow light in an atomic medium. *Nature*, 415: 406–409
Markov, M.A. 1965. Can the gravitational field prove essential for the theory of elementaryparticles? *Progress of Theoretical Physics*, (Suppl): 85–95
Mitchell, J. 1783. *Transactions of the Royal Society of London*, 74: 35
Novikov, I.D. & Frolov, V.P. 2001. Black holes in the Universe. *Physics-Uspekhi*, 44(3): 291
Novikov, I.D. & Zeldovich, Ya.B. 1966. *Nuovo Cimento*, 4 (Suppl.): 810
Oppenheimer, J.R. & Snyder, M. 1939. On continued gravitational contraction. *Physical Review*, 56: 455
Oppenheimer J.R. & Volkoff, G.M. 1939. On massive neutron cores. *Physical Review*, 55: 374–381
Schwarzschild, K. 1916. Über das Gravitational eines Massenpunktes nach der Einsteineschen Theory. *Sitzungsberichte der Preusischen Akademie der Wissenschaften zu Berlin, Physik-Mathematik, Kl.*: 189–196
Thorne, K.S., Price, R.H. & MacDonald, D.A. (editors). 1986. *Black Holes: The Membrane Paradigm*, New Haven: Yale University Press

BLOCH DOMAIN WALL

See **Domain walls**

BLOCH FUNCTIONS

See **Periodic spectral theroy**

BLOWOUT BIFURCATION

See **Intermittency**

BLOW-UP (COLLAPSE)

See **Development of singularities**

BOHR–SOMMERFELD QUANTIZATION

See **Quantum theory**

BOOMERONS

See **Solitons, types of**

BORN–INFELD EQUATIONS

Classical linear vacuum electrodynamics with point massive charged particles has two limiting properties: the electromagnetic energy of a point particle field is infinity, and a Lorentz force must be postulated to describe interactions between point particles and an electromagnetic field. Nonlinear vacuum electrodynamics can be free of these imperfections.

Gustav Mie (1912–1913) considered a nonlinear electrodynamics model in the framework of his "Fundamental unified theory of matter." In this theory the electron is represented by a nonsingular solution with a finite electromagnetic energy, but Mie's field equations are noninvariant under the gauge transformation for an electromagnetic four-potential (addition of the four-gradient of an arbitrary scalar function).

Max Born (1934) considered a nonlinear electrodynamics model that is invariant under the gauge transformation. A stationary electron in this model is represented by an electrostatic field configuration that is everywhere finite, in contrast to the case of linear electrodynamics when the electron's field is infinite at the singular point (see Figure 1). The central point in Born's electron is also singular because there is a discontinuity of electrical field at this point (hedgehog singularity). The full electromagnetic energy of this electron's field configuration is finite.

Born and Leopold Infeld (1934) then considered a more general nonlinear electrodynamics model, which has the same solution associated with the electron. Called Born–Infeld electrodynamics, this model is based on the Born–Infeld equations, which have the form of Maxwell's equations, including electrical and magnetic field strengths $\boldsymbol{E}$, $\boldsymbol{H}$, and inductions $\boldsymbol{D}$, $\boldsymbol{B}$ with nonlinear constitutive relations $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{E}, \boldsymbol{B})$, $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{E}, \boldsymbol{B})$ of a special kind. For inertial reference frames and in the region outside of field singularities, these equations are

$$\begin{cases} \operatorname{div} \boldsymbol{B} = 0, \\ \dfrac{1}{c}\dfrac{\partial \boldsymbol{B}}{\partial t} + \operatorname{curl} \boldsymbol{E} = 0, \\ \operatorname{div} \boldsymbol{D} = 0, \\ \dfrac{1}{c}\dfrac{\partial \boldsymbol{D}}{\partial t} - \operatorname{curl} \boldsymbol{H} = 0, \end{cases} \tag{1}$$

where

$$\begin{aligned} \boldsymbol{D} &= \frac{1}{\mathcal{L}}\,(\boldsymbol{E} + \chi^2\,\mathcal{J}\boldsymbol{B}), \\ \boldsymbol{H} &= \frac{1}{\mathcal{L}}\,(\boldsymbol{B} - \chi^2\,\mathcal{J}\boldsymbol{E}), \end{aligned} \tag{2}$$

$$\begin{aligned} &\mathcal{L} \equiv \sqrt{|\,1 - \chi^2\mathcal{I} - \chi^4\mathcal{J}^2\,|}, \\ &\mathcal{I} = \boldsymbol{E}\cdot\boldsymbol{E} - \boldsymbol{B}\cdot\boldsymbol{B}, \quad \mathcal{J} = \boldsymbol{E}\cdot\boldsymbol{B}. \end{aligned} \tag{3}$$

Relations (2) can be resolved for $\boldsymbol{E}$ and $\boldsymbol{H}$:

$$\begin{aligned} \boldsymbol{E} &= \frac{1}{\mathcal{H}}\left(\boldsymbol{D} - \chi^2\,\boldsymbol{\mathcal{P}} \times \boldsymbol{B}\right), \\ \boldsymbol{H} &= \frac{1}{\mathcal{H}}\left(\boldsymbol{B} + \chi^2\,\boldsymbol{\mathcal{P}} \times \boldsymbol{D}\right), \end{aligned} \tag{4}$$

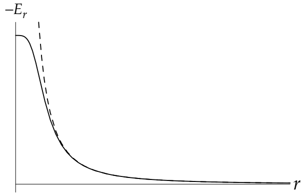

Figure 1. Radial components of electrical field for Born's electron and purely Coulomb field (dashed).

where $\mathcal{H}=\sqrt{1+\chi^2\left(\boldsymbol{D}^2+\boldsymbol{B}^2\right)+\chi^4\boldsymbol{\mathcal{P}}^2}$, $\boldsymbol{\mathcal{P}}\equiv\boldsymbol{D}\times\boldsymbol{B}$. Using relations (4) for Eq. (1), the fields $\boldsymbol{D}$ and $\boldsymbol{B}$ are unknown.

The symmetrical energy-momentum tensor for Born–Infeld equations has the following components:

$$\begin{aligned} T^{00} &= \frac{1}{4\pi\,\chi^2}\,(\mathcal{H}-1)\,, \quad T^{0i}=\frac{1}{4\pi}\,\mathcal{P}^i, \\ T^{ij} &= \frac{1}{4\pi}\left\{\delta^{ij}\left[\boldsymbol{D}\cdot\boldsymbol{E}+\boldsymbol{B}\cdot\boldsymbol{H}-\chi^{-2}\,(\mathcal{H}-1)\right]\right. \\ &\quad \left. -\left(D^i\,E^j+B^i\,H^j\right)\right\}. \end{aligned} \tag{5}$$

In spherical coordinates, the field of Born's static electron solution may have only radial components

$$D_r=\frac{e}{r^2}, \quad E_r=\frac{e}{\sqrt{\bar{r}^4+r^4}}, \tag{6}$$

where e is the electron's charge and $\bar{r}\equiv\sqrt{|\chi\,e|}$. At the point $r=0$, the electrical field has the maximum absolute value

$$|E_r(0)|=\frac{|e|}{\bar{r}^2}=\frac{1}{\chi}\,, \tag{7}$$

which Born and Infeld called the absolute field constant.

The energy of field configuration (6) is

$$m=\int T^{00}\,\mathrm{d}V=\frac{2}{3}\,\beta\,\frac{\bar{r}^3}{\chi^2}, \tag{8}$$

where the volume integral is calculated over the whole space, and

$$\beta\equiv\int\limits_0^\infty\frac{\mathrm{d}r}{\sqrt{1+r^4}}=\frac{\left[\Gamma(\frac{1}{4})\right]^2}{4\sqrt{\pi}}\approx 1.8541. \tag{9}$$

In view of the definition for $\bar{r}$ below (6), Equation (8) yields

$$\bar{r}=\frac{2}{3}\,\beta\,\frac{e^2}{m}. \tag{10}$$

Considering m as the mass of electron and using (7), Born & Infeld (1934) estimated the absolute field constant $\chi^{-1}\approx 3\times 10^{20}\,\mathrm{V/m}$. Later, Born & Schrödinger (1935) gave a new estimate (two orders of magnitude less) based on some considerations taking into account the spin of the electron. (Of course, such estimates may be corrected with more detailed models.)

An electrically charged solution of the Born–Infeld equations can be generalized to a solution with the singularity having both electrical and magnetic charges (Chernitskii, 1999). A corresponding hypothetical particle is called a dyon (Schwinger, 1969). Nonzero (radial) components of fields for this solution have the form

$$\begin{aligned} D_r &= \frac{\mathcal{C}_e}{r^2}, \quad E_r=\frac{\mathcal{C}_e}{\sqrt{\bar{r}^4+r^4}}, \\ B_r &= \frac{\mathcal{C}_m}{r^2}, \quad H_r=\frac{\mathcal{C}_m}{\sqrt{\bar{r}^4+r^4}}, \end{aligned} \tag{11}$$

where $\mathcal{C}_e$ is the electric charge and $\mathcal{C}_m$ is the magnetic one; $\bar{r}\equiv\left[\chi^2\left(\mathcal{C}_e^2+\mathcal{C}_m^2\right)\right]^{1/4}$. The energy of this solution is given by Equation (8) with this definition for $\bar{r}$. It should be noted that space components of electromagnetic potential for the static dyon solution have a line singularity.

A generalized Lorentz force appears when a small, almost constant field $\tilde{\boldsymbol{D}}$, $\tilde{\boldsymbol{B}}$ is considered in addition to the moving dyon solution. The sum of the field $\tilde{\boldsymbol{D}}$, $\tilde{\boldsymbol{B}}$ and the field of the dyon with varying velocity is taken as an initial approximation to some exact solution. Conservation of total momentum gives the following trajectory equation (Chernitskii, 1999):

$$\begin{aligned} m\,\frac{\mathrm{d}}{\mathrm{d}t}\,\frac{\boldsymbol{v}}{\sqrt{1-\boldsymbol{v}^2}} &= \mathcal{C}_e\left(\tilde{\boldsymbol{D}}+\boldsymbol{v}\times\tilde{\boldsymbol{B}}\right) \\ &\quad +\mathcal{C}_m\left(\tilde{\boldsymbol{B}}-\boldsymbol{v}\times\tilde{\boldsymbol{D}}\right), \end{aligned} \tag{12}$$

where $\boldsymbol{v}$ is the velocity of the dyon, and m is the energy for static dyon defined by (8).

A solution with two dyon singularities (called a bidyon) having equal electric ($\mathcal{C}_e=e/2$) and opposite magnetic charges can be considered as a model for a charged particle with spin (Chernitskii, 1999). Such a solution has both angular momentum and magnetic moment.

A plane electromagnetic wave with arbitrary polarization and form in the direction of propagation (without coordinate dependence in a perpendicular plane) is an exact solution to the Born–Infeld equations. The simplest case assumes one nonzero component of the vector potential ($A_y\equiv\phi(t,x)$), whereupon Equations (1) reduce to the linearly polarized plane wave equation

$$\begin{aligned} &\left(1+\chi^2\,\phi_x^2\right)\phi_{tt}-\chi^2\,2\,\phi_x\,\phi_t\,\phi_{xt} \\ &\quad -\left(c^2-\chi^2\,\phi_t^2\right)\phi_{xx}=0 \end{aligned} \tag{13}$$

with indices indicating partial derivatives. Sometimes called the Born–Infeld equation, Equation (13) has solutions $\phi=\zeta(x^1-x^0)$ and $\phi=\zeta(x^1+x^0)$, where $\zeta(x)$ is an arbitrary function (Whitham, 1974). Solutions comprising two interacting waves propagating in opposite directions are obtained via a hodograph transform (Whitham, 1974). Brunelli & Ashok (1998) have found a Lax representation for solutions of this equation.

A solution to the Born–Infeld equations, which is the sum of two circularly polarized waves propagating in different directions, was obtained by Erwin Schrödinger (1943).

Equations (1) with relations (2) have an interesting characteristic equation (Chernitskii, 1998)

$$g^{\mu\nu}\,\frac{\partial\Phi}{\partial x^{\mu}}\,\frac{\partial\Phi}{\partial x^{\nu}}=0,\quad g^{\mu\nu}\equiv g^{\mu\nu}-4\pi\,\chi^{2}\,T^{\mu\nu},\quad (14)$$

where $\Phi(x^{\mu})=0$ is an equation of the characteristic surface and $T^{\mu\nu}$ are defined by (5). This form for $g^{\mu\nu}$, including in addition the energy-momentum tensor, is special for Born–Infeld equations.

The Born–Infeld model also appears in the quantized string theory (Fradkin & Tseytlin, 1985) and in Einstein's unified field theory with a nonsymmetrical metric (Chernikov & Shavokhina, 1986). In general, this nonlinear electrodynamics model is connected with ideas of space-time geometrization and general relativity (see Eddington, 1924; Chernitskii, 2002).

ALEXANDER A. CHERNITSKII

See also **Einstein equations; Hodograph transform; Matter, nonlinear theory of; String theory**

Further Reading

Born, M. 1934. On the quantum theory of the electromagnetic field. *Proceedings of the Royal Society of London* A, 143: 410–437

Born, M. & Infeld, L. 1934. Foundation of the new field theory. *Proceedings of the Royal Society of London* A, 144: 425–451

Born, M. & Schrödinger, E. 1935. The absolute field constant in the new field theory. *Nature*, 135: 342

Brunelli, J.C. & Ashok, D. 1998. A Lax representation for Born–Infeld equation. *Physics Letters* B, 426: 57–63

Chernikov, N.A. & Shavokhina, N.S. 1986. The Born–Infeld theory as part of Einstein's unified field theory. *Soviet Mathematics*, (Izvestiya Vgsikish Uchebnykh Zaverdenii), 30(4): 81–83

Chernitskii, A.A. 1998. Light beams distortion in nonlinear electrodynamics. *Journal of High Energy Physics*, 11, 15: 1–5

Chernitskii, A.A. 1999. Dyons and interactions in nonlinear (Born–Infeld) electrodynamics. *Journal of High Energy Physics*, 12, 10: 1–34

Chernitskii, A.A. 2002. Induced gravitation as nonlinear electrodynamics effect. *Gravitation & Cosmology*, 8 (Suppl.), 123–130

Eddington, A.S. 1924. *The Mathematical Theory of Relativity*, Cambridge: Cambridge University Press

Fradkin, R.S. & Tseytlin, A.A. 1985. Nonlinear electrodynamics from quantized strings. *Physics Letters* B, 163: 123–130

Mie, G. 1912–13. Grundlagen einer theorie der materie. *Annalen der Physik*, 37: 511–534; 39: 1–40: 40: 1–66

Schrödinger, E. 1942. Dynamics and scattering-power of Born's electron. *Proceedings of the Royal Irish Academy* A, 48: 91–122

Schrödinger, E. 1943. A new exact solution in non-linear optics (two-wave system). *Proceedings of the Royal Irish Academy* A, 49: 59–66

Schwinger, J. 1969. A magnetic model of matter. *Science*, 165: 757–761

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

BOSE–EINSTEIN CONDENSATION

Bose–Einstein condensation (BEC) is the occupation of a single quantum state by a large number of identical particles, which implies that the particles are bosons, satisfying Bose–Einstein statistics and allowing for many particles to pile up in the same quantum state. This is in contrast to fermions, satisfying Fermi–Dirac statistics, for which the Pauli exclusion principle forbids the occupation of any single quantum state by more than one particle.

The role of quantum correlations caused by Bose–Einstein statistics is crucial for the occurrence of BEC. These statistics were advanced by Satyendranath Bose (1924) for photons, having zero mass, and generalized by Albert Einstein (1924) to particles with nonzero masses. Einstein (1925) also described the phenomenon of condensation in ideal gases. The possibility of BEC in weakly nonideal gases was theoretically demonstrated by Nikolai Bogolubov (1947). The wave function of Bose-condensed particles in dilute gases satisfies the Gross–Pitaevskii equation, suggested by Gross (1961) and Pitaevskii (1961). Its mathematical structure is that of the nonlinear Schrödinger equation. Experimental evidence of BEC in weakly interacting confined gases was achieved 70 years after Einstein's prediction, almost simultaneously, by three experimental groups (Anderson et al., 1995; Bradley et al., 1995; Davis et al., 1995).

To say that many particles are in the same quantum state implies that these particles display state coherence, a particular example of coherence phenomena requiring the particles to be strongly correlated with each other. The necessary conditions may be qualitatively understood by applying the de Broglie duality arguments to an ensemble of atoms in thermodynamic equilibrium at temperature T. Then the thermal energy of an atom is given by $k_{\mathrm{B}}T$, where k_{B} is the Boltzmann constant. This energy defines the thermal wavelength

$$\lambda_{\mathrm{T}}\equiv\sqrt{2\pi\hbar^{2}/m_{0}k_{\mathrm{B}}T}\qquad(1)$$

for an atom of mass m_0, with $\hbar$ being the Planck constant. Thus, an atom can be associated with a matter wave characterized by the wavelength (λ_{T}). Atoms become correlated with each other when their related waves overlap, which requires that the wavelength be larger than the mean interatomic distance, $\lambda_{\mathrm{T}}>a$. The average atomic density $\rho\equiv N/V$ for N atoms in volume V is related to the mean distance a through the equality $\rho a^3 = 1$. Hence, condition $\lambda_T > a$ may be rewritten as $\rho\lambda_T^3 > 1$. With the thermal wavelength (1), this yields the inequality

$$T < \frac{2\pi\hbar^2}{m_0 k_B}\,\rho^{2/3}, \tag{2}$$

which implies that state coherence may develop if the temperature is sufficiently low or the density of particles is sufficiently high.

An accurate description of BEC for an ideal gas is based on the Bose–Einstein distribution

$$n(\boldsymbol{p}) = \left[\exp\left(\frac{\varepsilon_p - \mu}{k_B T}\right) - 1\right]^{-1}, \tag{3}$$

describing the density of particles with a single-particle energy $\varepsilon_p = p^2/2m_0$ for a momentum $\boldsymbol{p}$ and with a chemical potential μ. The latter is defined from the condition $N = \sum_p n(\boldsymbol{p})$ for the total number of particles. Assuming the thermodynamic limit

$$N \to \infty, \qquad V \to \infty, \qquad \frac{N}{V} \to \text{const}$$

allows the replacement of summation over $\boldsymbol{p}$ by integration. Then, the fraction of particles, condensing to the state with $\boldsymbol{p} = 0$ is

$$n_0 \equiv \frac{N_0}{N} = 1 - \left(\frac{T}{T_c}\right)^{3/2} \tag{4}$$

below the condensation temperature

$$T_c = \frac{2\pi\hbar^2\rho^{2/3}}{m_0 k_B \zeta^{2/3}}, \tag{5}$$

where $\zeta \approx 2.612$. Above the critical temperature (5), $n_0 = 0$. The latter is about half of the right-hand side of inequality (2).

The condensate fraction (4) is derived for an ideal (noninteracting) Bose gas. A weakly nonideal (weakly interacting) Bose gas also displays Bose–Einstein condensation, although particle interactions deplete the condensate so that at zero temperature the condensate fraction is smaller than unity ($n_0 < 1$). A system is called weakly interacting if the characteristic interaction radius r_{int} is much shorter than the mean interparticle distance ($r_{\text{int}} \ll a$). This inequality can be rewritten as $\rho r_{\text{int}}^3 \ll 1$, and such a system is termed dilute.

Superfluid liquids, such as liquid ^{4}He, are far from being dilute, but it is commonly believed that the phenomenon of superfluidity is somehow connected with BEC. Although an explicit relation between the superfluid and condensate fractions is not known, theoretical calculations and experimental observations for superfluid helium estimate the condensate fraction at $T = 0$ as $n_0 \approx 0.1$.

A strongly correlated pair of fermions can be treated approximately as a boson, allowing superfluidity in liquid ^{3}He to be interpreted as the condensation of coupled fermions. Similarly, superconductivity is often compared with the condensation of the Cooper pairs that are formed by correlated electrons or holes. One should understand, however, that the superconductivity of fermions is analogous to but not identical to BEC of bosons.

An ideal object for the experimental observation of BEC is a dilute atomic Bose gas confined in a trap and cooled down to temperatures satisfying condition (2). Such experiments with different atomic gases have been recently realized, BEC has been explicitly observed, and a variety of its features have been carefully investigated. It has been demonstrated that the system of Bose-condensed atoms displays a high level of state coherence.

There exist different types of traps (single- and double-well), magnetic, optical, and their combinations, which make it possible to confine atoms for sufficiently long times of up to 100s. Using a standing wave of laser light, multi-well periodic effective potentials called optical lattices have been obtained, which have allowed the demonstration of a number of interesting effects, including Bloch oscillations, Landau–Zener tunneling, Josephson current, Wannier–Stark ladders, Bragg diffraction, and so on.

Displaying a high level of state coherence, an ensemble of Bose-condensed atoms forms a matter wave that is analogous to a coherent electromagnetic wave from a laser. Therefore, a device emitting a coherent beam of Bose atoms is called an atom laser.

The realization of BEC of dilute trapped gases is important for several reasons. First, this demonstrated the phenomenon predicted by Einstein in the 1920s. Note that a direct observation of BEC in superfluid helium—despite enormous experimental efforts—has never been achieved. Second, dilute atomic gases are simple statistical systems that can serve as a touchstone for testing different theoretical approaches. Finally, Bose-condensed trapped gases display a variety of interesting properties that promise diverse practical applications.

V.I. YUKALOV

See also **Coherence phenomena; Critical phenomena; Lasers; Nonequilibrium statistical mechanics; Nonlinear optics; Nonlinear Schrödinger equations; Order parameters; Phase transitions; Quantum nonlinearity; Quantum theory; Superconductivity; Superfluidity**

Further Reading

Anderson, M.H., Ensher, J.R., Matthews, M.R., Wieman, C.E. & Cornell, E.A. 1995. Observation of Bose–Einstein condensation in a dilute atomic vapor. *Science*, 269: 198–201

Bogolubov, N.N. 1947. On the theory of superfluidity. *Journal of Physics*, 11: 23–32

Bose, S.N. 1924. Plancks gesetz und lichtquantenhypothese. *Zeitschrift für Physik*, 26: 178–181

Bradley, C.C., Sackett, C.A., Tollett, J.J. & Hulet, R.G. 1995. Evidence of Bose–Einstein condensation in an atomic gas with attractive interactions. *Physical Review Letters*, 75: 1687–1690

Coleman, A.J. & Yukalov, V.I. 2000. *Reduced Density Matrices*, Berlin: Springer

Courteille, P.W., Bagnato, V.S. & Yukalov, V.I. 2001. Bose–Einstein condensation of trapped atomic gases. *Laser Physics*, 11: 659–800

Dalfovo, F., Giorgini, S., Pitaevskii, L.P. & Stringari, S. 1999. Theory of Bose–Einstein condensation in trapped gases. *Reviews of Modern Physics*, 71: 463–512

Davis, K.B., Mewes, M.O., Andrews, M.R., van Drutten, N.J., Durfee, D.S., Kurn, D.M. & Ketterle, W. 1995. Bose–Einstein condensation in a gas of sodium atoms. *Physical Review Letters*, 75: 3969–3973

Einstein, A. 1924. Quantentheorie des einatomigen idealen gases. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physik-Mathematik*, 261–267

Einstein, A. 1925. Quantentheorie des einatomigen idealen gases. Zweite abhandlung. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physik-Mathematik*, 3–14

Gross, E.P. 1961. Structure of a quantized vortex in boson systems. *Nuovo Cimento*, 20: 454–477

Huang, K. 1963. *Statistical Mechanics*, New York: Wiley

Klauder, J.R. & Skagerstam, B.S. 1985. *Coherent States*, Singapore: World Scientific

Lifshitz, E.M. & Pitaevskii, L.P. 1980. *Statistical Physics: Theory of Condensed State*, Oxford: Pergamon

Nozières, P. Pines, D. 1990. *Theory of Quantum Liquids: Superfluid Bose Liquids*, Redwood, CA: Addison-Wesley

Parkins, A.S. & Walls, D.F. 1998. The physics of trapped dilute-gas Bose–Einstein condensates. *Physics Reports*, 303: 1–80

Pitaevskii, L.P. 1961. Vortex lines in an imperfect Bose gas. *Journal of Experimental and Theoretical Physics*, 13: 451–455

Ter Haar, D. 1977. *Lectures on Selected Topics in Statistical Mechanics*, Oxford: Pergamon

Yukalov, V.I. & Shumovsky, A.S. 1990. *Lectures on Phase Transitions*, Singapore: World Scientific

Ziff, R.M., Uhlenbeck, G.E. & Kac, M. 1977. The ideal Bose-Einstein gas revisited. *Physics Reports*, 32: 169–248

BOSONS

See **Quantum nonlinearity**

BOUNDARY LAYERS

The Navier–Stokes system is the basic mathematical model for viscous incompressible flows. It reads

$$(\mathrm{NS}_\nu)\qquad \begin{cases} \partial_t \boldsymbol{u} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} - \nu\Delta\boldsymbol{u} + \nabla p = 0,\\ \mathrm{div}(\boldsymbol{u}) = 0,\\ \boldsymbol{u} = 0 \quad \text{on} \quad \partial\Omega, \end{cases} \tag{1}$$

where $\boldsymbol{u}$ is the velocity, p is the pressure, and ν is the viscosity. We can define a typical length scale L and a typical velocity U. The dimensionless parameter or Reynolds number, $Re = UL/\nu$, is very important to compare the properties of different flows. Indeed, two flows having the same Re have the same properties. When Re is very large (ν very small), the Navier–Stokes system (NS_ν) behaves like the Euler system

$$(\text{Euler})\qquad \begin{cases} \partial_t \boldsymbol{U} + \boldsymbol{U}\cdot\nabla\boldsymbol{U} + \nabla p = 0,\\ \mathrm{div}(\boldsymbol{U}) = 0,\\ \boldsymbol{U}.n = 0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{2}$$

In the region close to the boundary, the length scale becomes very small and we cannot neglect viscous effects. In 1905, Ludwig Prandtl suggested that there exists a thin layer called the boundary layer, where the solution $\boldsymbol{u}$ undergoes a sharp transition from a solution to the Euler system to the no-slip boundary condition $\boldsymbol{u} = 0$ on $\partial\Omega$ of the Navier–Stokes system. In other words, $\boldsymbol{u} = \boldsymbol{U} + \boldsymbol{u}_{\mathrm{BL}}$ where $\boldsymbol{u}_{\mathrm{BL}}$ is small except near the boundary.

To illustrate this, we consider a two-dimensional (planar) flow $\boldsymbol{u} = (u, v)$ in the half-space $\{(x, y) \mid y > 0\}$ subject to the following initial condition $\boldsymbol{u}(t = 0, x, y) = \boldsymbol{u}_0(x, y)$, boundary condition $\boldsymbol{u}(t, x, y = 0) = 0$, and $\boldsymbol{u} \to (U_0, 0)$ when $y \to \infty$.

Taking the typical length and velocity of order one, the Reynolds number reduces to $Re = \nu^{-1}$. Let $\varepsilon = Re^{-1/2} = \sqrt{\nu}$.

Near the boundary, the Euler system is not a good approximation. We introduce new independent variables and new unknowns

$$\tilde t = t, \quad \tilde x = x, \quad \tilde y = \frac{y}{\varepsilon},$$

$$(\tilde u, \tilde v)(\tilde t, \tilde x, \tilde y) = \left(u, \frac{v}{\varepsilon}\right)(\tilde t, \tilde x, \varepsilon\tilde y).$$

Notice that when $\tilde y$ is of order one, $y = \varepsilon\tilde y$ is of order ε. Rewriting the Navier–Stokes system in terms of the new variables and unknowns yields

$$\begin{cases} \tilde u_{\tilde t} + \tilde u\tilde u_{\tilde x} + \tilde v\tilde u_{\tilde y} - \tilde u_{\tilde y\tilde y} - \varepsilon^2\tilde u_{\tilde x\tilde x} + p_{\tilde x} = 0,\\ \varepsilon^2(\tilde v_{\tilde t} + \tilde u\tilde v_{\tilde x} + \tilde v\tilde v_{\tilde y} - \tilde v_{\tilde y\tilde y}) - \varepsilon^4\tilde v_{\tilde x\tilde x} + p_{\tilde y} = 0,\\ \tilde u_{\tilde x} + \tilde v_{\tilde y} = 0. \end{cases} \tag{3}$$

Neglecting the terms of order ε^2 and ε^4 yields

$$\begin{cases} \tilde u_{\tilde t} + \tilde u\tilde u_{\tilde x} + \tilde v\tilde u_{\tilde y} - \tilde u_{\tilde y\tilde y} + p_{\tilde x} = 0,\\ p_{\tilde y} = 0, \qquad \tilde u_{\tilde x} + \tilde v_{\tilde y} = 0. \end{cases} \tag{4}$$

Since p does not depend on $\tilde y$, we deduce that the pressure does not vary within the boundary layer and can be recovered from the Euler system (2) when $y = 0$, namely $p_x(t, x) = -(U_t + UU_x)(t, x, y = 0)$, since $V(t, x, y = 0) = 0$.

Going back to the old variables, we obtain

$$\begin{cases} u_t + uu_x + vu_y - \nu u_{yy} + p_x = 0,\\ u_x + v_y = 0 \end{cases} \tag{5}$$

which is the so-called Prandtl system. It should be supplemented with the following boundary conditions:

$$\begin{cases} u(t, x, y = 0) = v(t, x, y = 0) = 0,\\ (u, v)(t, x, y) \to (U(t, x, 0), 0) \quad \text{as} \quad y \to \infty. \end{cases} \tag{6}$$

Formally, a good approximation of $\boldsymbol{u}$ should be $\boldsymbol{U} + \boldsymbol{u}_{\mathrm{BL}}$, where U is the solution of the Euler system (2) and $\boldsymbol{u}_{\mathrm{BL}} + \boldsymbol{U}(t, x, 0)$ is the solution of the Prandtl system (5), (6).

Replacing the Navier–Stokes system by the Euler system in the interior and the Prandtl system near the boundary requires a justification. Mathematically, this can be formulated as a convergence theorem when ν goes to 0; namely, $\boldsymbol{u} - (\boldsymbol{U} + \boldsymbol{u}_{\mathrm{BL}})$ goes to 0 when ν goes to 0 in L^∞ or in some energy space (see Masmoudi, 1998 for a special case). In its whole generality, this is still a major open problem in fluid mechanics. This is due to problems related to the well-posedness of the Prandtl system as well as problems related to the instability of some solutions to the Prandtl system, which may prevent the convergence.

Let us explain the first problem for the steady Prandtl system

$$\begin{cases} uu_x + vu_y - \nu u_{yy} + p_x = 0, \\ u_x + v_y = 0 \end{cases} \tag{7}$$

in $\Omega = \{(x, y) \mid 0 < x < X, y > 0\}$ subject to the following extra boundary condition $u(x = 0, y) = u_0(y)$. Here, x should be thought of as a time-like variable. If we assume that $U, u_0 \geq 0$ and $u > 0$ if $y > 0$, we can introduce the von Mises transformation

$$(x, y) \rightarrow (x, \psi) \quad \text{and} \quad w = u^2,$$

where $\psi_y = u$, $\psi_x = -v$ and $\psi(x, 0) = 0$. In (x, ψ), the steady Prandtl system reads

$$w_x = \nu\sqrt{w}\,w_{\psi\psi} - 2p_x,$$

which is a degenerate parabolic equation, with the boundary conditions

$$\begin{gathered} w(x, 0) = 0, \quad w(0, \psi) = w_1(\psi), \\ w(x, \psi) \rightarrow U^2(x) \quad \text{as } \psi \rightarrow \infty, \end{gathered}$$

where $w_1(\int_0^y u_1(s)\mathrm{d}s) = u_1^2(y)$. Using this new equation, one can prove existence for the steady Prandtl system (see Oleinik and Samokhin, 1999). In the case of favorable pressure gradient, namely $p_x \leq 0$, the solution is global ($X = +\infty$). If $p_x > 0$, then a separation of the boundary layer may occur. x_0 is said to be a point of separation if $u_y(x_0, 0) = 0$ and $u_y(x, 0) > 0$ for $0 < x < x_0$. Qualitatively, the separation of the boundary layer is caused by a downward flow that drives the boundary layer away from the boundary. In that case, the assumption that the tangential velocity is large compared with the normal one is not valid, and the derivation of the Prandtl system is not justified.

A second obstacle to the convergence can come from the instability of the solution to the Prandtl system itself, if we consider a two-dimensional shear flow $\boldsymbol{u}^{\mathrm{s}} = (u^{\mathrm{s}}(y), 0)$, which is a steady solution of the Euler system. It is well known that the linear stability of such a flow is linked to the presence of inflection points in the profile u^{s}. A necessary condition of instability is that the profile has an inflection point. The solution to the Prandtl system with initial data $\boldsymbol{u}^{\mathrm{s}}$ and $U = 0$ is just the solution of a heat equation $u_t = \nu u_{yy}$. If the profile $\boldsymbol{u}^{\mathrm{s}}$ is linearly unstable for the Euler system, then u is not a good approximation of the Navier–Stokes system when ν goes to 0 (see Grenier, 2000).

The boundary layer theory is a very powerful tool in asymptotic analysis and is present in many different fields of partial differential equations, including the magnetohydrodynamic flow boundary layer. In fluid mechanics and atmospheric dynamics, we can also mention the Ekman layer, which is due to the balance between the viscosity and the rapid rotation of a fluid (Coriolis forces). In kinetic theory, systems of conservation laws, passage to the limit from a parabolic to a hyperbolic system, different types of boundary layers arise.

NADER MASMOUDI

See also **Fluid dynamics; Navier–Stokes equation**

Further Reading

Grenier, E. 2000. On the nonlinear instability of Euler and Prandtl equations. *Communications on Pure and Applied Mathematics*, (53): 1067–1091

Grenier, E. & Masmoudi, N. 1997. Ekman layers of rotating fluids, the case of well prepared initial data. *Communications in Partial Differential Equations*, 22: 953–975

Masmoudi, N. 1998. The Euler limit of the Navier-Stokes equations, and rotating fluids with boundary. *Archive Rational Mechanics and Analysis*, 142(4): 375–394

Oleinik, O.A. & V.N. Samokhin. 1999. *Mathematical Models in Boundary Layer Theory*, Boca Raton, FL: Chapman & Hall

Prandtl, L. 1905. Mathematiker-Kongresses. *Boundary Layer*, Heidelberg: *Verhandlung Internationalen*, pp. 484–494

Weinan, W.E. 2000. Boundary layer theory and the zero-viscosity limit of the Navier–Stokes system. *Acta Mathematica Sinica*, 16(2): 207–218

BOUNDARY VALUE PROBLEMS

For a given ordinary or partial differential equation, a boundary value problem (BVP) requires finding a solution of the equation valid in a bounded domain and satisfying a set of given conditions on the boundary of a domain. To define a boundary value problem, therefore, one needs to give an equation, a domain, and an appropriate number of functions supported on the boundary of the given domain, defining the boundary conditions. For example,

$q_t + q_x + q_{xxx} = 0$, the PDE,
$x \in [0, \infty)$, $t \in [0, T]$, the domain,
$q(x, 0) = f_0(x)$, $q(0, t) = g_0(t)$, the boundary conditions.

Finding a solution of a given BVP is more difficult than finding a function that satisfies the PDE, because of the constraint imposed on the solution at the boundary of the domain. Indeed, for nonlinear equations, there exists no general method to find the solution of a given BVP. The question of the solvability of such a problem must also be addressed, and in general, it does not have an easy answer (in the above example, if we prescribe two rather than just one condition at the boundary $x = 0$, there exists, in general, no solution). The existence and uniqueness of the solution of a given BVP can be guaranteed only when a specific, well-defined number of boundary conditions are prescribed, and this number depends on the highest-order derivatives appearing in the equation with respect to each variable (Coddington & Levinson, 1955).

For linear ordinary differential equations (ODEs), there is a general methodology for solving a BVP, based on defining the particular solution of a related problem. This particular solution is called the Green function associated with the BVP, and it depends on the boundary conditions. The Green function is used to define an integral operator, and if this operator is sufficiently regular, one can use it to express the solution of the original problem (Stackgold, 1979). This approach is powerful and fairly general, but it is not always successful, and it cannot be used for nonlinear equations.

No general methods are available to construct solutions for nonlinear ODEs or even to assert their existence. Most techniques rely on perturbing in some way the solution of an associated linearized problem or an integrable nonlinear problem. If one hopes to extract information about the nonlinear problem by studying a corresponding linearized one, the correct way to linearize must also be evaluated. Examples of such techniques are branching theory, eigenvalue perturbation, and boundary conditions or domain perturbation.

For linear PDEs in two variables, the classical approach for solving a BVP (going back to Jean D'Alembert's work in the 1740s) is separation of variables. The aim of this technique is to reduce the problem to two distinct linear problems for two ODEs. However, the separability of the problem depends on the specific domain and boundary conditions. For example, depending on the specific boundary conditions prescribed, the ODE one obtains may lead, via the associated Green function, to a non-self-adjoint transform problem, for which few general results are available.

An important theoretical result for the solvability of BVP for linear PDEs is the fundamental principle of Ehrenpreis (1970), which states that there always exists an appropriate generalization of the Fourier transform capable of representing the general solution of a BVP for a linear PDE in the variables (x_1, x_2), posed in a smooth convex domain. This result assumes the well-posedness of the problem. It then ensures that there exists a measure $\rho(k)$ and a contour Γ in the complex plane such that the solution of the problem can be expressed as an integral in the form

$$\int_\Gamma \mathrm{e}^{f_1(k)x_1+f_2(k)x_2}\mathrm{d}\rho(k),$$

where k is a complex parameter. The functions $f_1(k)$ and $f_2(k)$ are given explicitly. For example, in the case of an evolution equation $u_t = Lu$, where $u = u(x,t)$ and L is a linear differential operator, the representation takes the form

$$u(x,t) = \int_\Gamma \mathrm{e}^{\mathrm{i}kx-\mathrm{i}\omega(k)t}\mathrm{d}\rho(k), \qquad (1)$$

where $\omega(k)$ is the dispersion relation of the equation. In representation (1), the dependence on the solution variables (x,t) is explicit; the integration variable k is called the spectral parameter, and such a representation is said to be spectrally decomposed. However, this result is, in general, not constructive, as $\rho(k)$ and Γ are not known. In some cases, it is possible, to obtain this representation via separation of variables, but this is not always the case. Consider, for example, the second-order BVP

$$\begin{aligned} \mathrm{i}q_t + q_{xx} &= 0, 0 < x < \infty,\ 0 < t < \infty, \\ q(x,0) &= q_0(x), q(0,t) = f(t), \end{aligned} \qquad (2)$$

where $q = q(x,t)$ and it is assumed that all functions are infinitely differentiable and vanish as $x \to \infty$. By separating variables, one obtains an ODE in x, which can be solved using the sine transform. Assuming that a unique solution exists, this procedure yields the representation

$$\begin{aligned} q(x,t) = {} & \frac{2}{\pi}\int_0^\infty \sin(kx)\mathrm{e}^{-\mathrm{i}k^2t} \\ & \times \left(\hat{q}_0(k) + \mathrm{i}k\int_0^t \mathrm{e}^{\mathrm{i}k^2s}f(s)\mathrm{d}s\right)\mathrm{d}k, \end{aligned}$$

where

$$\hat{q}_0(k) = \int_0^\infty \sin(kx)q_0(x)dx.$$

This representation is not in form (1) as the variable t also appears as a limit of integration. This fact not only makes this representation less convenient for extracting information about the t dependence of the solution but also makes the rigorous proof of existence and uniqueness of a solution more cumbersome, as the relevant integral is not uniformly convergent as $x \to 0$.

For nonlinear PDEs, no general method is available (Logan, 1994). Perturbation techniques can be of some use in the study of evolutionary PDEs of the form $u_t + P(u) = 0$, where $u = u(t,x)$ and Pu is a nonlinear ODE containing the x-derivatives. Solutions of this problem such that $u_t = 0$ are called steady-state solutions: these are the solutions that are independent of time. In this context, one studies the linearized stability of the steady state by using the same perturbative techniques discussed for ODEs, as this yields information about the qualitative behavior in time of the solution of the nonlinear problem. The results available in this area are, in general, of limited applicability and practical use for finding explicit solutions.

The special class of nonlinear PDEs known as integrable deserves separate consideration. For these equations, there exists a particular linearizing technique, the inverse scattering transform, which yields the solution of the Cauchy problem. Some of these equations, such as the Korteweg–de Vries and sine-Gordon equations, have been considered on simple domains, and specific BVPs have been solved by ad hoc PDE techniques. The first such result was obtained already 40 years ago (Cattabriga, 1959), but recently this field has witnessed a new surge of interest. To obtain such results, the nonlinear problem is often considered as a linear problem, with the nonlinear term considered as a nonhomogeneous (or forcing) term; thus, the analysis is based on the analysis of the linearized equations by classical PDE techniques (Bona et al., 2001). A different approach involves the attempt to extend the inverse scattering linearizing technique to BVPs, as done, for example, in Leon (2002) for the sine-Gordon equation.

Recently, a general approach to solving BVPs for two-dimensional linear PDEs has been proposed and successfully used to solve many different types of such problems (Fokas, 2000). Its relevance is enhanced by the fact that this approach can be generalized to treat integrable nonlinear PDEs. This methodology yields a spectral transform associated directly to the PDE rather than to transforms associated to the two ODEs obtained by separating variables. For Example (2), this yields, for the solution, the representation

$$q(x,t) = \frac{1}{2\pi}\int_0^\infty \mathrm{e}^{\mathrm{i}kx-\mathrm{i}k^2t}\hat{q}_0(k)\mathrm{d}k + \frac{1}{2\pi}\int_\Gamma \mathrm{e}^{\mathrm{i}kx-\mathrm{i}k^2t}\hat{q}(k)\mathrm{d}k,$$

where Γ is the boundary of the first quadrant of the complex k-plane, and

$$\hat{q}(k) = \frac{\hat{q}_0(k)+\hat{q}_0(-k)}{2} - k\int_0^\infty \mathrm{e}^{\mathrm{i}k^2t} f(t)\mathrm{d}t.$$

This representation is in Ehrenpreis form, and in addition, measure and contour are explicitly constructed.

The above-mentioned approach provides a unification of the integral representation of the solution of a linear PDE in terms of the Ehrenpreis fundamental principle with the inverse scattering transform for integrable nonlinear PDEs. Indeed, when the problem reduces to an initial value problem for decaying solutions (i.e., the domain for the spatial variable is the whole real line, and the solution is assumed to vanish at $\pm\infty$), the transform obtained is precisely the inverse scattering transform.

The essential ingredients of this approach are the reformulation of the PDE as the closure condition of a certain differential form, and the definition in the complex plane of a Riemann–Hilbert problem depending on both the PDE and the domain. The differential form can be found algorithmically for linear PDEs and is equivalent to the Lax pair formulation for integrable nonlinear PDEs (Lax, 1968). The solution of this Riemann–Hilbert problem (which can be found in closed form in many cases) takes the role of the classical Green formula, and yields an integral representation for the solution, which is independent of the particular boundary conditions and indeed contains all the boundary values of the solution. What this approach crucially provides (when the definition domain is connected) is a global relation among these boundary values, which is the tool necessary to express the solution only in terms of the given boundary conditions and to prove rigorously problems with well-posedness, as well as existence and uniqueness results.

BEATRICE PELLONI

See also **Integrability; Inverse scattering method or transform; Riemann–Hilbert problem; Separation of variables**

Further Reading

Bona, J., et al. 2001. A non-homogeneous boundary value problem for the Korteweg–de Vries equation. *Transactions of the American Mathematical Society*, 354: 427–490

Cattabriga, L. 1959. Un problema al contorno per una equazione parabolica di ordine dispari. *Annali della Scuola Normale Superiore di Pisa*, 13: 163–203.

Coddington, E.A. & Levinson, N. 1955. *Theory of Ordinary Differential Equations*, New York: McGraw-Hill

Ehrenpreis, L. 1970. *Fourier Analysis in Several Complex Variables*, New York: Wiley-Interscience

Fokas, A.S. 2000. On the integrability of linear and nonlinear PDEs. *Journal of Mathematical Physics*, 41: 4188

Ghidaglia, J.L. & Colin, T. 2001. An initial-boundary value problems for the Korteweg–de Vries equation posed on a finite interval. *Advances in Differential Equations*, 6(12): 1463–1492

Lax, P.D. 1968. Integrals of nonlinear equations of evolution and solitary waves. *Communications in Pure and Applied Mathematics*, 21: 467–490

Leon, J. 2002. Solution of the Dirichlet boundary value problem for the sine-Gordon equation. *Physics Letters* A, 298, 343–252

Logan, J.D. 1994. *An Introduction to Nonlinear Partial Differential Equations*, New York: Wiley-Interscience

Stackgold, I. 1979. *Green's Functions and Boundary Value Problems*, New York: Wiley-Interscience

BOUSSINESQ EQUATIONS

See **Water waves**

BOX COUNTING

See **Dimensions**

BRAIN WAVES

See **Electroencephalogram at large scales**

BRANCHING LAWS

In this entry, we briefly trace the history of a familiar phenomenon, branching, in physical and biological systems and the laws governing them. The simplest type of branching tree is one in which a single conduit enters a vertex and two conduits emerge. This dichotomous process is clearly seen in the patterns of biological systems, such as botanical trees, neuronal dendrites, lungs, and arteries, as well as in the patterns of physical systems, such as lightning, river networks, and fluvial landscapes. The quantification of branching through the construction of the mathematical laws that govern them can be traced back to Leonardo da Vinci (1452–1519). In his *Notebooks*, he writes (Richter, 1970):

> All the branches of a tree at every stage of its height when put together are equal in thickness to the trunk [below them]. All the branches of a water [course] at every stage of its course, if they are of equal rapidity, are equal to the body of the main stream.

He also admonished his readers with: "Let no man who is not a Mathematician read the elements of my work." This statement increases in significance when we consider that da Vinci wrote nearly two centuries before Galileo (Galilei, 1638), who is generally given the credit for establishing the importance of mathematics in modern science.

The first sentence in the da Vinci quote is further clarified in subsequent paragraphs of the *Notebooks*. With the aid of da Vinci's sketch reproduced in Figure 1, this sentence has been interpreted as follows: if a tree has a trunk of diameter d_0 that bifurcates into two limbs of diameters d_1 and d_2, the three diameters are related by

$$d_0^a = d_1^a + d_2^a \tag{1}$$

Simple geometrical scaling yields the diameter exponent $\alpha = 2$, which corresponds to rigid pipes carrying fluid from one level of the tree to the next, while retaining a fixed cross-sectional area through successive generations of bifurcation. Although the pipe model has a number of proponents from hydrology, the diameter exponent for botanical trees was determined empirically by Cecil D. Murray in 1927 to be insensitive to the kind of botanical tree and to have a value 2.59 rather than 2 (Murray, 1927). Equation (1) is referred to as Murray's law in the physiology literature.

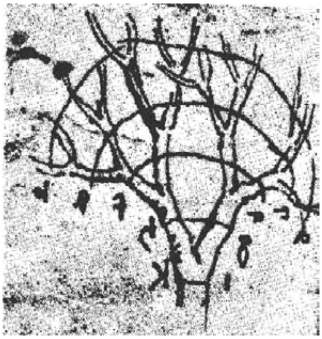

Figure 1. Sketch of tree from Leonardo da Vinci's Notebooks, PL XXVII (Richter, 1970).

The significance of Murray's law was not lost on D'Arcy Thompson (1942). In the second edition of his classic *On Growth and Form*, first published in 1917, Thompson argues that the geometrical properties of biological systems can often be the limiting factor in the development and final function of an organism. This is stated in his principle of similitude, which is a generalization of certain observations made by Galileo regarding the trade-off between the weight and strength of bone (Galilei, 1638). Thompson goes on to argue that the design principle for biological systems is that of energy minimization.

The second sentence in the da Vinci quotation is as suggestive as the first. In modern language, we would interpret it to mean that the flow of a river remains constant as tributaries emerge along the river's course. This equality must be true in order for the water to continue flowing in one direction and not stop and reverse course at the mouth of a tributary. Using the pipe model introduced above, and minimizing the energy with respect to the pipe radius, yields $\alpha = 3$, in Equation (1). Thus, the value of the diameter exponent obtained empirically by Murray falls between the theoretical limits of geometric self-similarity and hydrodynamic conservation, $2 \leq \alpha \leq 3$.

In da Vinci's tree, it is easy to assign a generation number to each of the limbs, but the counting procedure can become complicated in more complex systems like the generations of the bronchial tubes in the mammalian lung. One form taken by the branching laws is that the ratio of the radii of the tubes (from one generation to the next) is constant, that is, by the scaling relation

$$\frac{r_j}{r_{j+1}} = R. \tag{2}$$

Equation (2) is analogous to Horton's law for river trees and fluvial landscapes, which involves the ratio of the number of branches in successive generations of branches, rather than radii (Mandelbrot, 1977). In either case, the parameter R determines the branching law and Equation (2) implies a geometrical self-similarity in the tree, as anticipated by Thompson. In the branching of bronchial airways, $d_1 = d_2$ at each stage of the tree, so that from Equation (1) we deduce the relationship between the radii of the pipes between successive generations as

$$r_j = 2^{1/\alpha} r_{j+1}. \qquad (3)$$

In the case of the lung, the diameter of an airway is reduced by a factor $2^{-1/3}$ at each generation, since $\alpha = 3$ for the bronchial tree. Therefore, after j generations, $r_j = r_0 \exp(-j/\lambda)$, where the exponential rate of reduction, $\lambda = \ln(2)/3$, is the same for each generation beyond the trachea r_0, as argued by Weibel (2000).

A less space-filling value of the scaling index is obtained for the arterial system, where it is empirically determined that $\alpha = 2.7$. For a general non-integer branching index, the scaling relation Equation (3) defines a fractal tree. Such trees have no characteristic scale length and were first organized and discussed as a class by Benoit Mandelbrot (1977)—the father of fractals. The classical approach relied on the assumption that biological processes, like their physical counterparts, are continuous, homogeneous, and regular. However, most biological systems and many physical ones are discontinuous, inhomogeneous, and irregular and are necessarily that way in order to perform a particular function, such as gas exchange in lungs and arteries.

An entirely different kind of fractal tree is that of neuronal dendrites. The branching trees of neurons interleave the brain and form the communication system within the body. In the neurophysiology literature, Equation (1) is known as Rall's law with $\alpha = 1.5$ (Rall, 1959). More recent measurements of the scaling index, at each generation of dendritic branching, show a change with generation number; that is, the single parameter R is replaced with R_j. This non-constant scaling coefficient implies that Thompson's principle of similitude is violated.

A fractal model of the bronchial tree assumes that the ratio of successive generations of radii is dependent on the generation number, giving rise to a renormalization group relation, with the solution given by

$$r_j = \frac{a(j)}{j^u}, \quad j > 0. \qquad (4)$$

Here, the average radius is an inverse power law in the generation number j, modulated by a slowly oscillating function $a(j)$ as observed in the human, dog, rat, and hamster data shown in Figure 2 (West & Deering, 1994). In this way, the fractal concept is used as a design principle in biology (Weibel, 2000; West, 1999; Mandelbrot, 1977) and in the development of branching laws.

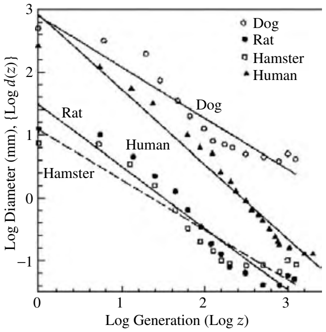

Figure 2. The variation in diameter d of the bronchial airways is depicted as a function of generation number j for rats, hamsters, humans, and dogs. The modulated inverse power law from the fractal model of the bronchial airway is observed in each case (West & Deering, 1994).

Bruce J. West

See also **Fibonacci series; Geomorphology and tectonics; Martingales; Multiplex neuron**

Further Reading

Galilei, G. 1638. *Dialogue Concerning Two New Sciences*, translated by H. Crew & A. deSalvio in 1914, New York: Dover, 1954

Mandelbrot, B.B. 1977. *The Fractal Geometry of Nature*, San Francisco: W.H. Freeman

Murray, C.D. 1927. A relationship between circumference and weight and its bearing on branching angles. *Journal of General Physiology*, 10: 725–729

Rall, W. 1959. Theory of physiological properties of dendrites. *Annals of New York Academy of Science*, 96: 1071–1091

Richter, J.P. 1970. *The Notebooks of Leonardo da Vinci*, vol. 1, New York: Dover, unabridged edition of the work first published in London in 1883

Thompson, D.W. 1942. *On Growth and Form*, 2nd edition, Cambridge: Cambridge University Press, republished New York: Dover, 1992

Weibel, E.R. 2000. *Symmorphosis: On Form and Function in Shaping Life*, Cambridge, MA: Harvard University Press

West, B.J. 1999. *Physiology, Promiscuity and Prophecy at the Millennium: A Tale of Tails*, Singapore: World Scientific

West, B.J. & Deering, W. 1994. Fractal physiology for physicists: Levy statistics. *Physics Reports*, 246: 1–100

BREATHERS

The term *breather* (also called a "bion") arose from studies of the sine-Gordon (SG) equation

$$u_{tt} - u_{xx} + \sin u = 0, \qquad (1)$$

which has localized solutions that oscillate periodically with time and decay exponentially in space. Such a solution of Equation (1) is given by

$$u(x,t) = 4\tan^{-1}\left[\frac{\lambda \sin \omega t}{\omega \cosh \lambda x}\right], \quad \lambda = \sqrt{1-\omega^2}, \tag{2}$$

which is shown in Figure 1.

Although the breather of Equation (2) is a non-topological soliton of Equation (1), it can be considered as a bound state of two topological solitons of the SG equation (kink and antikink), one of which is shown in Figure 2(a). The kink and antikink oscillate with respect to each other with the period $T = 2\pi/\omega$. Thus, such a soliton is also called a "doublet." A sketch of the bion at small frequencies ($\omega \ll 1$) and large enough t is presented in Figure 2(b). At some initial time, the kink and antikink move outward in opposite directions and separate in space with increasing time up to some finite distance (at $t = T/4$). The kink and antikink components of the breather never become fully free of distortions in their shapes due to interactions between each other, and finally oscillate in a kind of bound state.

At $1-\omega^2 \ll 1$, Equation (2) reduces to a small-amplitude breather $u(x,t) = 4\,Re\,\psi(x,t)$, where

$$\psi(x,t) = -\mathrm{i}\,\frac{\lambda \exp(-\mathrm{i}\omega t)}{\cosh \lambda x} \tag{3}$$

and $\lambda = \sqrt{2(1-\omega)}$. Equation (3) is a soliton solution of the nonlinear Schrödinger (NLS) equation

$$\mathrm{i}\psi_t + \tfrac{1}{2}\,\psi_{xx} - \psi + |\psi|^2\psi = 0, \tag{4}$$

which is regarded as a breather and can be written as $\psi(x,t) = \phi(x)\exp(\mathrm{i}\omega t)$. In this form, the spatial dependence of the soliton amplitude and the time dependence of the phase (of the complex function ψ) are separated. As a result, the nonlinearity appears only in the amplitude, but not in the phase of the NLS soliton. Although such a separation of the spatial and time dependencies in a soliton expression does not take place in the general form of the SG breather, the limiting case of the SG breather coincides with the amplitude of the NLS soliton.

At present, it is not known whether other nonlinear Klein–Gordon equations similar to Equation (1), but differing from it only by the nonlinear term, possess exact breather solutions (Segur & Kruskal, 1987). However, if certain nonlinear terms in a Klein–Gordon-type equation differ only slightly from $\sin u$ (slightly perturbed SG equation), a breather-like solution may persist in the first order with respect to the perturbation (Birnir et al., 1994).

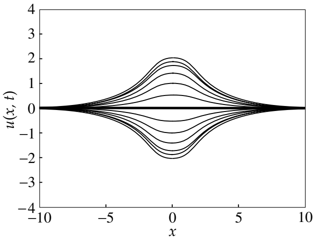

Figure 1. $u(x,t)$ from Equation (2) versus x for 26 different times equally spaced and covering one period, with $\lambda = 0.5$.

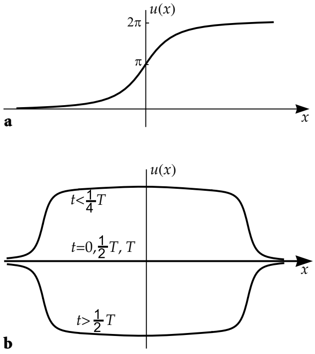

Figure 2. (a) A sketch of the sine-Gordon kink. (b) Three profiles of the kink-antikink oscillations.

The breather of the SG equation can move along the space coordinate axis with a stationary velocity V. As Equation (1) is a relativistic invariant equation (invariant under a Lorentz transformation of the independent variables), one can obtain a moving breather from Equation (2) substituting $x \to (x - Vt)/(1 - V^2)^{1/2}$ and $t \to (t - Vx)/(1 - V^2)^{1/2}$. Consequently, the moving breathers form a two-parametric (ω and V) family of solutions of the SG equation.

Possible values of the breather parameters ω and V can be compared with the dispersion relation ($\omega^2 = 1 + k^2$) for small vibrations (phonons) described by the linearized version of Equation (1). These phonons have frequencies $\omega > 1$ and phase velocities $\omega/k > 1$. A breather frequency, on the other hand, is smaller than the minimum frequency of the phonons ($\omega < 1$), and the breather velocity is smaller than the minimum phonon phase velocity ($V < 1$). Therefore, the dynamical breather parameters lie outside of the spectrum of the linear vibrations. Although the time dependence of the breather includes the higher temporal harmonics of the oscillations, the phonons cannot be resonantly excited by the breather. Thus, breathers and phonons are asymptotically independent vibrational modes of the system described by the SG equation. This asymptotic independence of nonlinear excitations (breathers and kinks) and phonons follows from the integrability of Equation (1).

An important way to study nonlinear integrable equations is the inverse scattering method. According to this method, breathers are characterized by poles in the complex phase plane of scattering parameters for the equation under consideration.

It is known that several nonlinear differential equations possess breather solutions. The Landau–Lifshitz (LL) equation provides an example of a nonlinear equation generalizing the results that are described by the SG and NLS equations. The breather-like solution of the LL equation has a more complicated form than the one presented above; however, it is also a two-parameter soliton called a dynamic magnetic soliton (Kosevich et al., 1977). Its oscillatory behavior is characterized by a frequency ω, and its center can move with a velocity V. In the general case, the magnetic soliton can be described by some complex function of x and t, but the time and spatial dependencies are not separated in the analytical expression for such a soliton.

An important class of breathers the so-called discrete breathers (also known as intrinsic localized modes, self-localized anharmonic modes, or nonlinear localized excitations). These are solutions of a nonlinear equation on a lattice, and they are periodic in time and localized in space. Although most such investigations are performed by numerical calculations, there exist nonlinear dynamic equations on a lattice possessing exact analytical breather solutions. One of them is the following discrete version of the NLS equation proposed by Ablowitz and Ladik (AL) in 1976 (Ablowitz & Ladik, 1976):

$$\mathrm{i}\partial_t \psi_n - (\psi_{n+1} + \psi_{n-1})(1 + |\psi_n|^2) = 0. \qquad (5)$$

The AL lattice is integrable and it allows for breather-like solutions, the simplest of which has a form close to that of Equation (2):

$$\psi = \frac{\sinh\beta \exp(-\mathrm{i}\omega t)}{\cosh\beta(n - x_0)}, \qquad (6)$$

where n is the integer number of a lattice site, $x_0 = \text{constant}$, and $\omega = -2\cosh\beta$.

ARNOLD KOSEVICH

See also **Discrete breathers; Discrete nonlinear Schrödinger equations; Integrability; Inverse scattering method or transform; Sine-Gordon equation; Solitons**

Further Reading

Ablowitz, M.J. & Ladik, J.F. 1976. Nonlinear differential-difference equations and Fourier analysis. *Journal of Mathematical Physics*, 17: 1011

Birnir, B., McKean, H.P. & Weinstein, A. 1994. The rigidity of sine-Gordon breathers. *Communications on Pure and Applied Mathematics*, 47: 1043

Flach, S. & Willis, C.R. 1998. Discrete breathers. *Physics Reports*, 295: 181

Kosevich, A.M., Ivanov, B.A. & Kovalev, A.S. 1977. Nonlinear localized magnetization wave of a ferromagnet as a bound-state of a large number of magnons. *Pis'ma Zhurnal Eksperimental'noy i Teoreticheskoy Fiziki*, 25: 516 (in Russian); *JETP Letters*, 25: 486

Kosevich, A.M., Ivanov, B.A. & Kovalev, A.S. 1990. Magnetic solitons. *Physics Reports*, 194: 117

Segur, H. & Kruskal, M.D. 1987. Nonexistence of small-amplitude breather solutions in ϕ^4 theory. *Physical Review Letters*, 58: 747

BROUWER'S FIXED POINT THEOREM

See **Winding numbers**

BROWNIAN MOTION

In 1828, Robert Brown, a leading botanist, observed that a wide variety of particles suspended in liquid exhibit an intrinsic, irregular motion when viewed under a microscope. While not the first to witness such motion, his experimental focus on this phenomenon, which would bear his name, established its universality and intrinsic nature, thereby raising it as an issue for fundamental scientific inquiry (Nelson, 1967). Deutsch has recently raised the question of whether Brown actually witnessed Brownian motion or fluctuations due to some external contaminating influence (Peterson, 1991). Indeed, while Brownian motion is a ubiquitous phenomenon, not all irregular motions can be ascribed to Brownian motion. The dancing of dust particles in sunlight is dominated by imperceptible turbulent currents, not Brownian motion. True Brownian motion is generally only visible on scales of microns and below, but has important macroscopic ramifications because all microscopic particles manifest it. For example, Brownian motion makes possible both the fine-scale mixing of initially segregated substances in nature and industry, as well as the passive transport of ions, nutrients, and fuel, which allow biological cells to support life.

The origin of Brownian motion remained under debate throughout the 19th century, with Cantoni, Delsaux, Gouy, and C. Weiner proposing that thermal motions in the suspending liquid were responsible, as discussed in Einstein (1956, pp. 86–88), Gallavotti (1999, Chapter 8), and Russel et al. (1989, pp. 65–66). Attempts to examine this hypothesis quantitatively were hampered by the inability to measure accurately the velocity of particles undergoing Brownian motion, since such motion loses coherence over time scales (microseconds) that are shorter than those which experimental observations were able to resolve. In one of three ground-breaking papers that Einstein published in 1905, he offered a statistical mechanical means for theoretical calculations involving Brownian motion (Einstein, 1956). Einstein realized that the quantity involving Brownian motion that can be best observed under a microscope in an experiment is the "diffusivity":

$$D = \lim_{t\to\infty} \frac{|X(t) - X(0)|^2}{2t}, \tag{1}$$

where $X(t)$ denotes the observed displacement of the Brownian particle along a fixed direction at time t. In practice, t is simply taken as some satisfactorily long time of observation, and there is no need for fine temporal resolution as there would be if the velocity were to be measured. Einstein employed a random walk model for his analysis and showed that the diffusivity defined in (1) is identical to the diffusion constant that describes the macroscopic evolution of the concentration density $n(\boldsymbol{x}, t)$ of a large number of Brownian particles:

$$\frac{\partial n(\boldsymbol{x}, t)}{\partial t} = D \sum_{j=1}^{3} \frac{\partial^2 n(\boldsymbol{x}, t)}{\partial x_j^2}. \tag{2}$$

Through an elegant argument based on *equilibrium* statistical mechanical arguments, Einstein showed that the diffusivity D of a Brownian particle must be related to its friction coefficient ξ in the following way:

$$D = \frac{k_B T}{m\xi}, \tag{3}$$

where k_B is Boltzmann's constant and T is the absolute temperature (measured in Kelvin scale), and m is the particle's mass. The friction coefficient ξ appears in the relation between the drag force F_{drag} and velocity v of the particle in steady-state motion (assuming a low Reynolds number):

$$F_{\text{drag}} = m\xi v. \tag{4}$$

For a sphere of radius a moving through a fluid with dynamic viscosity μ, the friction coefficient is given by $\xi = 6\pi\mu a/m$. The remarkable property of the "Einstein relation" in (3) is that it links a quantity D pertaining to statistically unpredictable dynamical fluctuations to a quantity ξ, which involves deterministic, steady-state properties. Later work generalized the Einstein relation (3) to "fluctuation-dissipation theorems," which express the structure of the spontaneous statistical fluctuations in a wide class of physical systems to the structure of the dissipative (frictional) dynamics (Kubo et al., 1991, Chapter 1).

The basic theory of Brownian motion was developed by Einstein in 1905, a time when the premises of the atomic theory of matter were still not yet fully agreed upon (Gallavotti, 1999; Nelson, 1967). Einstein realized that a careful observation of Brownian motion and his relation between the diffusivity of a Brownian particle and its mobility could be used to calculate the number of particles making up a given mass of fluid if the atomic theory were valid. Under a microscope sufficient to resolve the Brownian motion of a particle, all quantities in (3) are directly observable except for Boltzmann's constant k_B. Therefore, the Einstein relation (3) can be used to compute a value for k_B based on Brownian motion data. Now, k_B is in turn related to Avogadro's number N_A, which is the number of molecules in a "mole" (a certain well-defined macroscopic amount) of a substance. The Brownian motion data and the Einstein relation, therefore, furnish an independent prediction for Avogadro's number N_A and, thereby, the number of molecules per unit mass of the fluid. In other words, the number (and therefore mass) of the individual fluid particles could be calculated without having to observe them at an individual level, an experimental feat that has become possible only in recent years. Instead, their individual mass and number could be assessed through their collective influence on a much larger and, therefore, observable immersed particle. In 1908, Jean Perrin experimentally confirmed that the value of N_A computed in this way agreed with those obtained from other techniques (Gallavotti, 1999), providing strong support for the atomic theory of matter. Since the 1970s, Brownian motion has been investigated in the laboratory through dynamic light scattering techniques (Russel et al., 1989, Chapter 3).

The most idealized mathematical representation of Brownian motion with diffusivity D is defined as $(2D)^{1/2}W(t)$, where $W(t)$ is a canonical continuous random process with Gaussian statistics such that $W(0) = 0$, $\langle W(t)\rangle = 0$, and

$$\langle (W(t) - W(t'))^2\rangle = |t - t'|. \tag{5}$$

This mathematical Brownian motion is often referred to as the Wiener process (Borodin & Salminen, 2002; Gallavotti, 1999; Nelson, 1967). This idealized Brownian motion has independent increments (no inertia). Physical Brownian motion, of course, has some small inertia as well as several other complicating influences from the fluid environment and from the presence of other nearby Brownian particles (Russel et al., 1989). These extra features can be built into a dynamical description using the mathematical Brownian motion as the basic noise input with influence mediated by the other physical parameters. The mathematical Brownian motion has a similar role in modeling noise input in a wide variety of stochastic models in physics, biology, finance, and other fields. More precisely, the Levy–Khinchine theorem indicates that in any system affected by noise in a continuous way such that the noise on disjoint time intervals is independent can be modeled in terms of mathematical Brownian motion (Reichl, 1998, Chapters 4, 5).

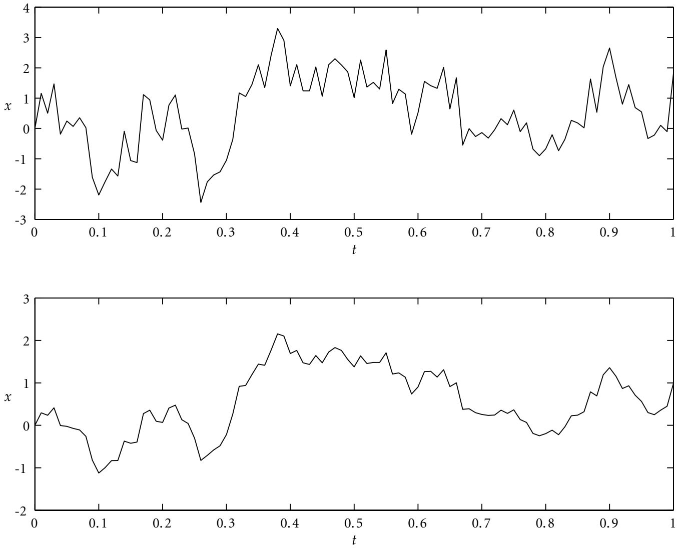

Figure 1. Sample trajectories of fractional Brownian motion using Fourier-wavelet method (Elliott et al., 1997). Top panel: $H = \frac{1}{3}$, lower panel: $H = \frac{2}{3}$. Both simulations used the same random numbers.

Discontinuous noise-induced jumps, in contrast, are modeled in terms of Poisson processes or more generally Lévy processes (Reichl, 1998, Chapters 4, 5). Continuous noise with long-range correlations (so that the independent increment property is not satisfied), on the other hand, can often be usefully modeled in terms of "fractional Brownian motion" (FBM) (Mandelbrot, 2002). This is an idealized Gaussian random process $Z(t)$ with $Z(0) = 0$, $\langle Z(t) \rangle = 0$, and

$$\langle (Z(t) - Z(t'))^2 \rangle = |t - t'|^{2H} \tag{6}$$

where the Hurst exponent H is chosen from the interval $0 < H < 1$. The FBM with $H = \frac{1}{2}$ corresponds to ordinary Brownian motion with independent increments. FBMs with $\frac{1}{2} < H < 1$ have positive, long-ranged correlations with less rough trajectories and large excursions, while FBMs with $0 < H < \frac{1}{2}$ have negative, long-ranged correlations with rougher trajectories and a more oscillatory character (Figure 1). All FBMs have a statistical self-similarity property; the statistics of the rescaled FBM $\lambda^{-H} Z(\lambda t)$ are identical to those of the original $Z(t)$. That is, these processes have no finite length or time scale associated to them and can be thought of as random fractals. Fractional Brownian motions are therefore particularly appropriate for modeling systems with fluctuations occurring over a wide range of scales; cutoff lengths can be introduced by filtering an input FBM. Models built from FBMs have been developed in turbulence theory, natural landscape and cloud structures, surface adsorption processes, neural signals in biology, and self-organized critical systems such as earthquakes, forest fires, and sandpiles.

PETER R. KRAMER

See also **Fluctuation-dissipation theorem; Fluid dynamics; Fokker–Planck equation; Lévy flights; Random walks**

Further Reading

Borodin, A.N. & Salminen, P. 2002. *Handbook of Brownian Motion: Facts and Formulae*, 2nd edition, Basel: Birkhäuser

Einstein, A. 1956. *Investigations on the Theory of the Brownian Movement*, edited with notes by R. Fürth, translated by A.D. Cowper, New York: Dover

Elliott, Jr, F.W., Horntrop, D.J. & Majda, A.J. 1997. A Fourier-wavelet Monte Carlo method for fractal random fields. *Journal of Computational Physics*, 132(2): 384–408

Gallavotti, G. 1999. *Statistical Mechanics*, Berlin: Springer

Kubo, R., Toda, M. & Hashitsume, N. 1991. *Statistical Physics*, vol. 2, 2nd edition, Berlin: Springer

Mandelbrot, B.B. 2002. *Gaussian Self-affinity and Fractals*, Chapter IV. New York: Springer

Mazo, R.M. *Brownian Motion. Fluctuations, Dynamics and Applications*, Oxford and New York: Oxford University Press

Nelson, E. 1967. *Dynamical Theories of Brownian Motion*, Princeton, NJ: Princeton University Press

Peterson, I. 1991. Did Brown see Brownian motion? *Science News*, 139: 287

Reichl, L.E. 1998. *A Modern Course in Statistical Physics*. 2nd edition, New York: Wiley

Russel, W.B., Saville, D.A. & Schowalter, W.R. 1989. *Colloidal Dispersions*. Cambridge and New York: Cambridge University Press

BRUSSELATOR

The Brusselator is an autocatalytic model involving two intermediates. It illustrates how the fundamental laws of thermodynamics and chemical kinetics as applied to open systems far from equilibrium can give rise to self-organizing behavior and to dissipative structures in the form of temporal oscillations and spatial pattern formation.

Chemical kinetics imposes stringent conditions on the concentrations of the species involved in a reaction scheme and on the associated parameters. In a scheme consisting entirely of elementary steps, the overall rates are given (in an ideal system) by mass action kinetics, featuring particular combinations of products of concentrations preceded by stoichiometric coefficients (integer numbers specifying how the relevant constituents are produced or consumed). This guarantees the positivity of the solutions of the mass balance equations. A second condition is detailed balance, which requires that in chemical equilibrium, each individual reaction step is balanced by its inverse (*See* **Detailed balance**). This gives rise to relations linking the concentrations of initial reactants and final products to the rate constants, independent of the concentrations of the intermediates.

An additional set of requirements stems from the fact that self-organization in chemical kinetics must arise through an instability. One reason for this is that equilibrium and the states in its vicinity, obtained as the constraints are gradually increased (called the "thermodynamic branch"), are stable. To overcome this property and evolve to states that are qualitatively different from equilibrium, new branches of solutions must be generated, which can only take place through the mechanisms of instability and bifurcation. This, in turn, requires that the non-equilibrium constraints exceed a critical value (Glansdorff & Prigogine, 1971). Because the evolution laws generated by chemical kinetics at the macroscopic level of description are dissipative, the bifurcating states are attractors, attained by families of initial conditions belonging to an appropriate part of phase space. This guarantees structural stability, that is to say, the robustness of the solution toward small perturbations—a condition to be fulfilled by a model aiming to describe a physical phenomenon. Note that non-equilibrium instabilities and self-organization collapse when the system becomes closed to the external environment or when the kinetics involves only first-order steps, in which case one obtains a monotonic decay to a unique steady state (Denbigh et al., 1948).

Following the pioneering work of Alfred Lotka, several authors in the late 1940s, proposed models of open nonlinear systems deriving from mass action kinetics and giving rise to sustained oscillations (Lotka, 1956; Moore, 1949). These models do not have structural stability (as they give rise to a continuum of initial condition-dependent solutions), and they do not exhibit the role of the constraints in a transparent manner (as they are usually formulated in the limit of irreversible reactions). When the non-equilibrium constraints are explicitly accounted for, it is found (Lefever et al., 1967) that there is no instability threshold in these models. As the first known chemical model that is both fully compatible with the laws of thermodynamics and chemical kinetics and generates dissipative structures through non-equilibrium instabilities, the Brusselator is free from such deficiencies.

Model Presentation

In the interest of transparency, one desires a minimal model, and if oscillatory behavior is one of the required properties, this necessitates two coupled variables representing the concentrations of intermediate products. As in the models of the Lotka family, one seeks steps that are not only nonlinear but also include feedback processes, the simplest chemical version of which is autocatalysis. But contrary to these models, one now needs to scan the whole range of near to far from equilibrium situations and to undergo an instability somewhere in this range. As the Lotka-type models contain only second-order steps, a natural solution is to amend them by replacing these steps by a third-order one. This leads to the scheme (Prigogine & Lefever, 1968)

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} X, \quad B + X \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} Y + D,$$

$$2X + Y \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} 3X, \quad X \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} E \qquad (1)$$

Hanusse, Tyson, and Light have shown that a two-variable system compatible with the above requirements necessarily comprises a third-order step. Here A, B are the initial reactants, D, E the final products, and X, Y the intermediates: X can be thought of as an activator generating Y at its own expense, which acts as an inhibitor if the B concentration is large.

From the standpoint of irreversible thermodynamics, the Brusselator can be driven out of equilibrium through two independent constraints (affinities) related to the overall reactions

$$A \underset{k_{-1}k_{-4}}{\overset{k_1k_4}{\rightleftharpoons}} E, \; B \underset{k_{-3}k_{-2}}{\overset{k_3k_2}{\rightleftharpoons}} D.$$

This offers sufficient flexibility to allow one to take the limit of purely irreversible steps and of fixed reactant and product concentrations (also referred to as pool chemical approximation, ensuring that the (X, Y) subsystem becomes open to the external environment), while satisfying the positivity of concentrations, detailed balance, and mass conservation (Lefever et al., 1988). When diffusion is also included, and upon performing a suitable scaling transformation, this leads to the Brusselator equations

$$\begin{aligned} \frac{\partial X}{\partial t} &= A - (B+1)X + X^2Y + D_x\nabla^2 X, \\ \frac{\partial Y}{\partial t} &= BX - X^2Y + D_y\nabla^2 Y \end{aligned} \tag{2}$$

in which B, D_x/D_y, and the system size usually play the role of the parameters controlling the instabilities.

A number of variants of this canonical form have also been developed, including Brusselator in an open well-stirred reactor, Brusselator in a non-ideal system, including coupling between non-equilibrium instabilities and phase transitions, and coupling with external fields or advection.

Behavior of the Solutions

Since the first bifurcation analysis of the Brusselator equations (Nicolis & Auchmuty, 1974), several studies have been devoted to the various modes of spatiotemporal organization generated by Equations (2): limit cycles, Turing patterns, and traveling waves in one-dimensional systems (Nicolis & Prigogine, 1977); spatiotemporal chaos arising from the diffusive coupling of local limit cycle oscillators (Kuramoto, 1984); patterns in two- and three-dimensional systems including patterns arising from the interference of different instability mechanisms such as Turing, Hopf (De Wit, 1999); and the effect of confinement (Herschkowitz-Kaufman & Nicolis, 1972). Many phenomena now known to be generic have first been discovered on these Brusselator-based analyses, which have also helped to test the limits of traditional theoretical approaches and to explore new methodologies such as normal forms and phase dynamics.

The Brusselator has also been used to explore possible thermodynamic signatures of dissipative structures. No clearcut tendencies seem to exist, suggesting that global thermodynamic quantities like entropy and entropy production do not provide adequate measures of dynamic complexity. Finally, attention has been focused on the new insights afforded when the mean-field equations (2) are augmented to account for fluctuations, a study for which the Brusselator is well suited thanks to its mechanistic basis. The interest here is to provide a fundamental understanding of how large-scale order can be sustained despite the locally prevailing thermal disorder. Early accounts of the results, with emphasis on critical behavior in the vicinity of bifurcations, can be found in Nicolis & Prigogine (1977) and in Walgraef et al. (1982).

G. Nicolis

See also **Chemical kinetics; Detailed balance; Emergence; Turing patterns**

Further reading

Denbigh, K.G., Hicks, M. & Page, F.M. 1948. Kinetics of open systems. *Transactions of the Faraday Society*, 44: 479–494

De Wit, A. 1999. Spatial patterns and spatio-temporal dynamics in chemical systems. *Advances in Chemical Physics*, 109: 453–513

Glansdorff, P. & Prigogine, I. 1971. *Thermodynamic Theory of Structure, Stability and Fluctuations*, London and New York: Wiley

Herschkowitz-Kaufman, M. & Nicolis, G. 1972. Localized spatial structures and nonlinear chemical waves in dissipative systems. *Journal of Chemical Physics*, 56: 1890–1895

Kuramoto, Y. 1984. *Chemical Oscillations, Waves and Turbulence*, Berlin and New York: Springer

Lefever, R., Nicolis, G. & Prigogine, I. 1967. On the occurrence of oscillations around the steady state in systems of chemical reactions far from equilibrium. *Journal of Chemical Physics*, 47: 1045–1047

Lefever, R., Nicolis, G. & Borckmans, P. 1988. The Brusselator: it does oscillate all the same. *Journal of the Chemical Society, Faraday Transactions*, 1. 84: 1013–1023

Lotka, A. 1956. *Elements of Mathematical Biology*, New York: Dover

Moore, M.J. 1949. Kinetics of open reaction systems: chains of simple autocatalytic reactions. *Transactions of the Faraday Society*, 45: 1098–1109

Nicolis, G. & Auchmuty, J.F.G. 1974. Dissipative structures, catastrophes, and pattern formation: a bifurcation analysis. *Proceedings National Academy of Sciences USA*, 71: 2748–2751

Nicolis, G. & Prigogine, I. 1977. *Self-organization in Nonequilibrium Systems*, New York: Wiley

Prigogine, I. & Lefever, R. 1968. Symmetry-breaking instabilities in dissipative systems. II. *Journal of Chemical Physics*, 48: 1695–1700

Walgraef, D., Dewel, G. & Borckmans, P. 1982. Nonequilibrium phase transitions and chemical instabilities. *Advances in Chemical Physics*, 491: 311–355

BULLETS

See **Solitons, types of**

BURGERS EQUATION

In 1915, Harry Bateman considered a nonlinear equation whose steady solutions were thought to describe certain viscous flows (Bateman, 1915). This equation, modeling a diffusive nonlinear wave, is now widely known as the Burgers equation, and is given by

$$u_t + uu_x = \frac{\mu}{2} u_{xx}, \qquad (1)$$

where μ is a constant measuring the viscosity of the fluid. It is a nonlinear parabolic equation, simply describing a temporal evolution where nonlinear convection and linear diffusion are combined, and it can be derived as a weakly nonlinear approximation to the equations of gas dynamics.

Although nonlinear, Equation (1) is very simple, and interest in it was revived in the 1940s, when Dutch physicist Jan Burgers proposed it to describe a mathematical model of turbulence in gas (Burgers, 1940). As a model for gas dynamics, it was then studied extensively by Burgers (1948), Eberhard Hopf (1950), Julian Cole (1951), and others, in particular; after the discovery of a coordinate transformation that maps it to the heat equation. While as a model for gas turbulence the equation was soon rivaled by more complicated models, the linearizing transformation just mentioned added importance to the equation as a mathematical model, which has since been extensively studied. The limit $\mu \to 0$ is a hyperbolic equation, called the inviscid Burgers equation:

$$u_t + uu_x = 0. \qquad (2)$$

This limiting equation is important because it provides a simple example of a conservation law, capturing the crucial phenomenon of shock formation. Indeed, it was originally introduced as a model to describe the formation of shock waves in gas dynamics.

A first-order partial differential equation for $u(x,t)$ is called a conservation law if it can be written in the form

$$u_t + (f(u))_x = 0\,.$$

For Equation (2), $f(u) = u^2/2$. Such conservation laws may exhibit the formation of shocks, which are discontinuities appearing in the solution after a finite time and then propagating in a regular manner. When this phenomenon arises, an initially smooth wave becomes steeper and steeper as time progresses, until it forms a jump discontinuity—the shock.

Once a discontinuity forms, the solution is no longer a globally differentiable function; thus, the sense in which it can be considered as a solution of the PDE must be clarified. A discontinuous function $(u(x,t))$ can still be considered as a solution in the weak sense if it satisfies

$$\int\int_D \left(u\varphi_t + \frac{1}{2}u^2\varphi_x \right) \mathrm{d}x\mathrm{d}t = 0, \qquad (3)$$

where D is any rectangle in the (x,t) plane, and $\varphi(x,t)$ is any smooth function vanishing on the boundary ∂D. Any regular solution is a weak solution, as is seen by multiplying the equation by $\varphi(x,t)$, integrating by parts along ∂D, and using Green's theorem.

In physical applications, one often considers the discontinuous solution as the limit, as $\mu \to 0$, of smooth solutions of the viscous Equation (1). This idea is correct from a physical point of view, as it takes into account the significance of these solutions as a physical description of gas dynamics. From the form of the equation (or its weak formulation (3)), one can derive the velocity v_s of a shock separating two regimes, u_r to the right and u_l to the left of a discontinuity. The result is the Rankine–Hugoniot formula, valid in general for conservation laws, which for the case of the Burgers equation yields

$$s = \tfrac{1}{2}(u_\mathrm{r} + u_\mathrm{l}). \qquad (4)$$

Even this, however, is not enough to guarantee uniqueness of the solution, because there are several ways of writing the equation in the form of a conservation law. Often, the way to select the physically relevant solution is to consider the vanishing viscosity limit. To obtain this solution mathematically, an additional entropy condition, that $u_\mathrm{l} > s > u_\mathrm{r}$, must be imposed.

Besides its significance as a model for shocks, the Burgers equation is prominent among PDEs because it is completely integrable. Indeed, the nonlinear change of variable

$$u = -\mu(\log \psi)_x \qquad (5)$$

transforms Equation (1) into the heat equation $\psi_t = \psi_{xx}$, with initial conditions transforming simply into initial conditions for this latter equation: if $u(x,0) = f(x)$ is the given condition, then the corresponding initial condition for the heat equation is given by

$$\psi(x,0) = \exp\left[-\frac{1}{\mu} \int_0^x f(\xi)\mathrm{d}\xi \right].$$

The relation between the Burgers and the heat equation was already mentioned in an earlier book (Forsyth, 1906), but the former had not been recognized as physically relevant; hence, the importance of this connection was seemingly not noticed at the time. Using the transformation of Equation (5), known as the Cole–Hopf transformation, it is easy to solve the initial value problem for this equation. Recently, a generalization of the Cole–Hopf transformation has been successfully used to linearize the boundary value problem for the Burgers equation posed on the semiline $x > 0$ (Calogero & De Lillo, 1989).

The existence of this linearizing transformation, which is a transformation of Bäcklund type (Rogers & Shadwick, 1982) relating the solutions of two different PDEs, stimulated work to extend this approach to a generalized version of the Burgers equation, such as the Korteweg–de Vries–Burgers equation, given by

$$u_t + uu_x = \frac{\mu}{2} u_{xx} - \varepsilon u_{xxx}, \quad \varepsilon > 0. \qquad (6)$$

Although it was found out that such a directly linearizing transformation did not exist, efforts in this direction were rewarded by several discoveries. Indeed, the importance of a linearizing transformation became evident when the inverse scattering transform (IST) was discovered, leading to the full analytical understanding of the solution of the Cauchy problem for the Korteweg–de Vries (KdV) equation, and later all integrable evolution equations in one spatial dimension, such as KdV, the nonlinear Schrödinger, and the sine-Gordon equations.

A crucial step in the discovery of the IST was an observation made by Robert Miura. In analogy to gas dynamics, he noted that one needs conservation laws to compute jump conditions across the region where the solution is small and essentially dispersive, isolating the solitonic part, which is thought of as a kind of reversible shock. This led to the connection between the KdV and modified KdV equation via the Miura transformation and eventually to the IST through which these nonlinear equations are solved through a series of linear problems (Gardner et al., 1967).

Nowadays, the Burgers equation is used as a simplified model of a kind of hydrodynamic turbulence (Case & Chiu, 1969), called Burgers turbulence. Burgers himself wrote a treatise on the equation now known by his name (Burgers, 1974), where several variants are proposed to describe this particular kind of turbulence. Generalizations such as the KdV–Burgers equation (6) arose from the need to model more complicated physical situations and introduce more factors than those that the Burgers equation takes into account. Lower-order friction terms may be considered that reduce the amplitude of the wave, although in a different scale and manner than the reduction due to the higher-order diffusion term u_{xx}. For example, the KdV–Burgers equation is an appropriate model when a different higher-order amplitude-reducing effect, namely dispersion, is introduced. Depending on the relative sizes of μ and ε, this equation may exhibit either an essentially shock-like structure, with the presence of dispersive tail, or mainly dispersive phenomena; thus, Equation (6) has been proposed as a natural model for hydrodynamic turbulence. In the context of the study of gas dynamics (particularly turbulent and vorticity phenomena), the Burgers equation has also been used to model phase diffusion along vortex filaments.

BEATRICE PELLONI

See also **Constants of motion and conservation laws; Inverse scattering method or transform; Shock waves; Turbulence**

Further Reading

Bateman, H. 1915. Some recent research on the motion of fluids. *Monthly Weather Review*, 43: 163–170

Burgers, J. 1940. Application of a model system to illustrate some points of the statistical theory of free turbulence. *Proceedings of the Nederlandse Akademie van Wetenschappen*, 43: 2–12

Burgers, J. 1948. A mathematical model illustrating the theory of turbulence, *Advances in Applied Mechanics*, 1: 171–199

Burgers, J. 1974. *The Nonlinear Diffusion Equation: Asymptotic Solutions and Statistical Problems*, Dordrecht and Boston: Reidel

Calogero, F. & De Lillo, S. 1989. The Burgers equation on the semiline. *Inverse Problems*, 5: L37

Case, K.M. & Chiu, S.C. 1969. Burgers turbulence models. *Physics of Fluids*, 12: 1799–1808

Cole, J. 1951. On a quasilinear parabolic equation occuring in aerodynamics. *Quarterly Journal of Applied Mathematics*, 9:225–236

Forsyth, A.R. 1906. *Theory of Differential Equations*, Cambridge: Cambridge University Press

Gardner, C.S., Greene, J.M., Kruskal, M.D. & Miura, R.M. 1967. Method for solving the Korteweg-de Vries equation. *Physical Review Letters*, 19: 1095–1097

Hopf, E. 1950. The partial differential equation $u_l + uu_x = \mu u_{xx}$. *Communications in Pure and Applied Mathematics*, 3: 201–230

Lax, P.D. 1973. *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, Philadelphia: Society for Industrial and Applied Mathematics

Newell, A.C. (editor). 1974. *Nonlinear Wave Motion*, Providence, RI: American Mathematical Society

Rogers, C. & Shadwick, W.F. 1982. *Bäcklund Transformations and Their Applications*, New York: Academic Press

Sachdev, P.L. 1987. *Nonlinear Diffusive Waves*, Cambridge and New York: Cambridge University Press

Smoller, J. 1983. *Shock Waves and Reaction-Diffusion Equations*, Berlin and New York: Springer

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley-Interscience

BUTTERFLY EFFECT

The Butterfly Effect serves as a metaphor for what in technical language is called "sensitive dependence on initial conditions" or "deterministic chaos," the fact that small causes can have large effects.

As recounted by Gleick (1987, Chapter 1), in the early 1960s, Edward Lorenz was carrying out computer simulations on a 12-dimensional weather model. One day, he decided to run a particular time series for longer. In order to save time, he restarted his code from data from a previous printout. After returning from a coffee break, he found that the weather simulation had diverged sharply from that of his earlier run. After some checks, he could only conclude that the difference was caused by the difference in initial conditions: he had typed in only the first three of the six decimal digits that the computer worked with internally. Apparently, his assumption that the fourth digit would be unimportant was false.

Lorenz realized the importance of his observation: "If, then, there is any error whatever in observing the present state—and in any real system such errors seem inevitable—an acceptable prediction of an instantaneous state in the distant future may well be impossible" (Lorenz, 1963, p. 133). Indeed, the error made by discarding the fourth and higher digits is so small that it can be imagined to represent the effect of the flap of the wings of a butterfly. Lorenz originally used the image of a seagull. The more lasting name seems to have come from his address at the annual meeting of the American Association for the Advancement of Science in Washington, December 29, 1972, which was entitled "Predictability: does the flap of a butterfly's wings in Brazil set off a tornado in Texas?" The text of this talk was never published but is presented in its original form as an appendix in Lorenz (1993).

Sensitive dependence on initial conditions forces us to distinguish between determinism and predictability, two concepts often confused by scientists and popular writers alike. Determinism has to do with how Nature (or, less ambitiously, any system under consideration) behaves, while predictability has to do with what we, human beings, are able to observe, analyze, and compute. We have determinism if we have a law or a formula describing exactly, and fully, how the system behaves given its present state. To have predictability we need, in addition, to be able to measure the present state of the system with sufficient precision and to compute with the given formula (to solve the equations) in a sufficiently accurate computational scheme.

Determinism is most famously expressed by Pierre-Simon Laplace (1814, p.2):

> An intelligence that, at a given instant, could comprehend all the forces by which nature is animated and the respective situation of the beings that make it up, if moreover it were vast enough to submit these data to analysis, would encompass in the same formula the movements of the greatest bodies of the universe and those of the lightest atoms. For such an intelligence nothing would be uncertain, and the future, like the past, would be open to its eyes.

Laplace's dramatic statement is often erroneously interpreted as a belief in perfect predictability now rendered untenable by the chaos theory. But he was describing determinism: given the state of the system (the universe) at some time, we have a formula (a set of differential equations) that gives, in principle, the state of the system at any later time. Nowhere will one find a claim about the computability, by us humans, of all the consequences of the laws of mechanics. Indeed, the quote appears in the introduction of a book on probability. Laplace is, in fact, assuming incomplete knowledge from the start and uses probabilities to make rational inferences. If it were not for quantum mechanics, Laplace's statement would still stand, unaffected by deterministic chaos.

To illustrate the problems with computability, consider the simple but important example of the (deterministic) Bernoulli shift map defined by

$$f : [0, 1] \to [0, 1] : \; x_{n+1} = 2x_n \; (\mathrm{mod}\, 1).$$

On numbers in binary representation, this map has a particularly simple effect: shift the binary point one place to the right and discard the first digit. For example, if $x_0 = 0.10110$ (which corresponds to the decimal 0.6875), then $x_1 = 0.01100$ (decimal 0.375). Now, any rational starting number x_0 is represented by a repeating sequence of 0s and 1s and hence leads to a periodic orbit of f, while any irrational x_0 is represented by a nonrepeating sequence of 0s and 1s and hence leads to a nonperiodic orbit. This latter sequence would appear as unpredictable as the sequence of heads and tails generated by flipping a coin, the quintessentially random process. Since there is an irrational number arbitrarily close to every rational number and vice versa, the map exhibits a sensitive dependence on initial conditions. In practice, on a computer, numbers are always represented with finite precision; hence, the computations become completely meaningless once—after a finite number of iterations—all significant digits have been removed. In the standard 32-bit (4-byte) floating point arithmetic with 23-bit mantissa, this will be after roughly 23 iterations.

The significance of the Bernoulli shift map is that dynamical systems theory tells us that its dynamics lies at the heart of the so-called "horseshoe dynamics," which in turn is commonly found in (the wide class of) systems with homoclinic (i.e., expanding and reinjecting) orbits (Wiggins, 1988). It means that in many situations, all we can say about a system's dynamics is of a statistical nature. A quantitative measure of the sensitivity on initial conditions, and therefore a measure of the predictability horizon, is provided by the leading Lyapunov exponent.

The possibility of small causes having large effects (in a perfectly deterministic universe) was anticipated by many scientists before Lorenz, and even before the birth of dynamical systems theory, which is generally accepted to have its origins in Poincaré's work on differential equations toward the end of the 19th century. Maxwell (1876, p. 20) wrote: "There is a maxim which is often quoted, that 'The same causes will always produce the same effects.'" After discussing the meaning of this principle, he adds: "There is another maxim which must not be confounded with [this], which asserts that 'Like causes produce like effects.' This is only true when small variations in the initial circumstances produce only small variations in the final state of the system." He then gives the example of how a small displacement of railway points sends a train on different courses.

Others have often used the image of the weather:

Wiener (1954/55):

> It is quite conceivable that the general outlines of the weather give us a good, large picture of its course for hours or possibly even for days. However, I am profoundly skeptical of the unimportance of the unobserved part of the weather for longer periods.

> To assume that these factors which determine the infinitely complicated pattern of the winds and the temperature will not in the long run play their share in determining major features of weather, is to ignore the very real possibility of the self-amplification of small details in the weather map. A tornado is a highly local phenomenon, and apparent trifles of no great extent may determine its exact track. Even a hurricane is probably fairly local where it starts, and phenomena of no great importance there may change its ultimate track by hundreds of miles.

Poincaré (1908, p. 67):

> Why have meteorologists such difficulty in predicting the weather with any certainty? Why is it that showers and even storms seem to come by chance, so that many people think it quite natural to pray for rain or fine weather, though they would consider it ridiculous to ask for an eclipse by prayer? We see that great disturbances are generally produced in regions where the atmosphere is in unstable equilibrium. The meteorologists see very well that the equilibrium is unstable, that a cyclone will be formed somewhere, but exactly where they are not in a position to say; a tenth of a degree more or less at any given point, and the cyclone will burst here and not there, and extend its ravages over districts it would otherwise have spared. If they had been aware of this tenth of a degree, they could have known it beforehand, but the observations were neither sufficiently comprehensive nor sufficiently precise, and that is the reason why it all seems due to the intervention of chance.

Even earlier, Franklin (1898, p. 173) had used an analogy surprisingly similar to Lorenz's:

> . . . an infinitesimal cause may produce a finite effect. Long range detailed weather prediction is therefore impossible, and the only detailed prediction which is possible is the inference of the ultimate trend and character of a storm from observations of its early stages; and the accuracy of this prediction is subject to the condition that the flight of a grasshopper in Montana may turn a storm aside from Philadelphia to New York!

Duhem (1954, p. 141) used Hadamard's theorem of 1898 on the complicated geodesic motion on surfaces of negative curvature to "expose fully the absolutely irremediable physical uselessness of certain mathematical deductions." If such incomputable behavior is possible in mechanics, "the least complex of physical theories," Duhem goes on to ask rhetorically, "Should we not meet that ensnaring conclusion in a host of other, more complicated problems, if it were possible to analyse the solutions closely enough?"

Many had contemplated the possibility of sensitive dependence on initial conditions, but Lorenz was the first to see it actually happening quantitatively in the numbers spit out by his Royal McBee computing machine, and to be sufficiently intrigued by it to study it more closely in the delightfully simple system of equations now bearing his name (Lorenz, 1963). Indeed, while most scientists, with Duhem, had looked to complicated systems for unpredictable behavior, Lorenz found it in simple ones and thereby made it amenable to analysis.

GERT VAN DER HEIJDEN

See also **Chaotic dynamics; Determinism; General circulation models of the atmosphere; Horseshoes and hyperbolicity in dynamical systems; Lorenz equations**

Further Reading

Bricmont, J. 1995. Science of chaos or chaos in science? *Physicalia Magazine*, 17: 159–208

Duhem, P. 1954. *The Aim and Structure of Physical Theory*, translated by Ph.P. Wiener, Princeton: Princeton University Press (original French edition, *La Théorie Physique: Son Objet, Sa Structure, 2ème éd.*, Paris: Marcel Rivière & Cie, 1914; 1st edition, 1906)

Franklin, W.S. 1898. A book review of P. Duhem, *Traité Elémentaire de Méchanique Chimique fondée sur la Thermodynamique*, vols. I and II, Paris, 1897, *The Physical Review*, 6: 170–175

Gleick, J. 1987. *Chaos: Making a New Science*, London: Heinemann, and New York: Viking

Laplace, P.-S. 1814. *Philosophical Essay on Probabilities*, translated from the fifth French edition of 1825 by A.I. Dale, Berlin and New York: Springer, 1995 (first edition published in French, 1814)

Lorenz, E.N. 1963. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20: 130–141

Lorenz, E.N. 1993. *The Essence of Chaos*, Seattle: University of Washington Press and London: University College London Press

Maxwell, J.C. 1876. *Matter and Motion*, New York: Van Nostrand

Poincaré, H. 1908. *Chance*. In *Science and Method*, pp. 64–90, translated by F. Maitland, London: Thomas Nelson and Sons, 1914 (original French edition, *Science et Méthode*, Paris: E. Flammarion, 1908)

Wiener, N. 1954/55. Nonlinear prediction and dynamics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, vol. 3, pp. 247–252; *Mathematical Reviews*, 18: 949 (1957); reprinted in 1981. *Norbert Wiener: Collected Works with Commentaries*, vol. III, edited by P. Masani, Cambridge, Massachusetts: MIT Press, pp. 371–376

Wiggins, S. 1988. *Global Bifurcations and Chaos*, Berlin and New York: Springer

C

CALOGERO–MOSER MODEL

See **Particles and antiparticles**

CANDLE

At about the time that John Scott Russell was systematically studying hydrodynamic solitons on a Scottish canal, Michael Faraday, the brilliant English experimental physicist and physical chemist, organized his annual Christmas Lectures on facets of natural philosophy for young people. These included a series on the candle that began with the claim (Faraday, 1861):

> There is no better, there is no more open door by which you can enter into the study of natural philosophy than by considering the physical phenomena of a candle.

Although this assertion may have startled some of his listeners, Faraday went on to support it with a sequence of simple yet elegant experiments that clearly expose the structure and composition of a candle flame, demonstrating a stream of energy-laden vapor feeding into the flame and suggesting an analogy with the process of respiration in living organisms (Day & Catlow, 1994). (An engraving showing Faraday presenting one of these lectures can be found on a recent British 20-pound note.)

While the details are intricate (Fife, 1988), the flame of a candle can be regarded globally as a dynamic process balancing two flows of energy: the rate at which energy is dissipated by the flame (through emission of heat and light) and the rate at which energy is released from the wax as the flame eats its way down the candle. Let us define variables as follows.

- P is the power dissipation by the flame, in units of (say) joules per hour.
- E is the chemical energy stored in the wax of the candle, in units of (say) joules per centimeter.
- v is the speed at which the flame moves down the candle.

If the rates of dissipation and energy input are equal, then the velocity of the flame is determined by the power balance condition

$$P = vE, \qquad (1)$$

implying $v = P/E$ (cm/h). Metaphorically, the flame must digest energy at the same rate at which it is eaten.

Consider a family of cylindrical candles with various diameters (d), and assume the dissipation rates of their flames to be independent of the sizes of the candles. Since stored chemical energy is proportional to the area of cross section ($E \propto d^2$), power balance implies

$$v \propto 1/d^2. \qquad (2)$$

Some measured flame speeds for typical candles are plotted on a log–log scale in Figure 1, where the dashed line of slope -2 indicates a $1/d^2$ dependence. From this figure, it is evident that the inverse square dependence of Equation (2) is obeyed for larger candles. For smaller candles, v is somewhat less than expected, because the flames are not so large.

Although Equation (2) was derived for candles, Equation (1) is quite general, expressing a global constraint that governs the dynamics of many kinds of nonlinear diffusion, including the propagation of nerve impulses (Scott, 2002). For a smooth axon described by the Hodgkin–Huxley equations, power balance is established between electrostatic energy released from the fiber membrane and ohmic dissipation by circulating ionic currents, implying $v \propto \sqrt{d}$. (Plotted in Figure 1, this dependence would have a slope of $+\frac{1}{2}$.) For myelinated nerves, on the other hand, evolutionary pressures require that $v \propto d$, corresponding to a slope of unity in Figure 1.

In the language of nonlinear dynamics, a candle flame provides a physical example of an attractor, which is evidently stable because moderate disturbances (small gusts of air) do not extinguish the flame by forcing it out of its basin of attraction. As the air becomes still, the flame returns to its original shape and size. The task of lighting a candle, on the other hand, requires getting the wick hot enough—above an

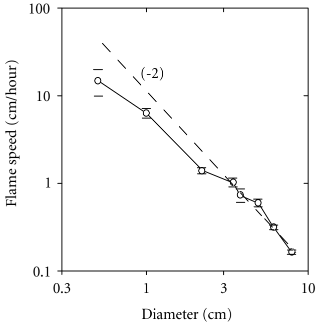

Figure 1. Measurements of flame speeds (v) for candles of different diameters (d). The error bars indicate rms deviations of about six individual measurements. (Data courtesy of Lela Scott MacNeil Scott (2003).)

ignition threshold and into the basin of attraction—so that a viable flame is established. Qualitatively similar conditions govern the firing of a nerve axon, leading to an all-or-nothing response.

From a more intuitive perspective, an ignition threshold implies the power balance indicated in Equation (1), where the corresponding flame is unstable. Above the threshold, instability arises from the establishment of a positive feedback loop, in which the flame releases more than enough chemical energy than is needed to maintain its temperature. Such a positive feedback loop is represented by the diagram

Release of energy (vE)
$\downarrow \qquad \uparrow$
Dissipation of energy (P),

with the gain about the loop being greater than unity, implying an increase in the flame size with time (Scott, 2003). Eventually, this temporal increase is limited by nonlinear effects in the release and dissipation of energy, reducing the loop gain to unity as the fully developed flame is established.

The candle flame is an example of a reaction-diffusion (or autocatalytic) process, going back to an early suggestion by Robert Luther, a German physical chemist (Luther, 1906). Following a lecture demonstration of a chemical wave, Luther claimed that such systems should support traveling waves at a speed proportional to $\sqrt{D/\tau}$ where D is the diffusion constant for the reacting components (in units of distance squared per unit of time) and τ is a delay time for the onset of the reaction. During the 1930s, autocatalytic systems were studied in the context of genetic diffusion through spatially dispersed biological species, and in 1938 the equation

$$D\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} = \frac{1}{\tau}u(u-a)(u-1), \tag{3}$$

where a is a threshold parameter lying in the range $(0, \frac{1}{2}]$, was used by Soviet scientists Yakov Zeldovich and David Frank-Kamenetsky to represent a flame front. These authors showed that uniform traveling waves solutions of Equation (3) propagate at a fixed speed given by the expression

$$v = (1-2a)\sqrt{D/2\tau}, \tag{4}$$

which includes both Luther's factor $\left(\sqrt{D/\tau}\right)$ and the power balance condition. Long overlooked by the neuroscience community, this early work on flame propagation offers a convenient model for the leading edge of a nerve impulse, confirming Faraday's intuition (Scott, 2002).

Alwyn Scott

See also **Attractors; Flame front; Hodgkin–Huxley equations; Power balance; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Day, P. & Catlow, C.R.A. 1994. *The Candle Revisited: Essays on Science and Technology*, Oxford and New York: Oxford University Press

Faraday, M. 1861. *A Course of Six Lectures on the Chemical History of a Candle*. Reprinted as *Faraday's Chemical History of a Candle*, Chicago: Chicago Review Press, 1988

Fife, P.C. 1988. *Dynamics of Internal Layers and Diffusive Interfaces*, Philadelphia: Society for Industrial and Applied Mathematics

Luther, R. 1906. Räumliche Fortpflanzung chemischer Reaktionen. *Zeitschrift für Elektrochemie* 12(32): 596–600 (English translation in *Journal of Chemical Education* 64 (1987): 740–742)

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*. 2nd edition, Oxford and New York: Oxford University Press

CANONICAL VARIABLES

See **Hamiltonian systems**

CANTOR SETS

See **Fractals**

CAPACITY DIMENSION

See **Dimensions**

CAPILLARY WAVES

See **Water waves**

CARDIAC ARRHYTHMIAS AND THE ELECTROCARDIOGRAM

In the early 1900s, Willem Einthoven developed the string galvanometer to measure the potential differences in the body surface associated with the heartbeat and introduced a nomenclature for the deflections of the electrocardiogram that are still used today. For this work, Einthoven was awarded a Nobel prize in 1924 (Katz & Hellerstein, 1964).

The electrocardiogram (ECG) is a measurement of the potential difference between two points on the surface of the body. Because the heart generates waves of electrical activation that propagate through the heart during the cardiac cycle, ECG measurements reflect cardiac activity. Over the past century, physicians have learned how to interpret the electrocardiogram to diagnose a variety of different cardiac abnormalities. Although interpreting an ECG is difficult, this entry introduces the basic principles.

In order to appreciate the ECG, it is first necessary to have a rudimentary knowledge about the spread of the cardiac impulse in the heart. The heart is composed of four chambers, the right and left atria, and the right and left ventricles (see Figure 1). The atria are electrically connected to each other, but are insulated from the ventricles everywhere except in a small region called the atrioventricular (AV) node. The ventricles are also electrically connected to each other. The rhythm of the heart is set by the sinoatrial node located in the right atrium, which acts as the pacemaker of the heart. From a mathematical perspective, this pacemaker is an example of a nonlinear oscillator. Thus, if the rhythm is perturbed, for example, by delivering a shock to the atria, then in general the timing of subsequent firings of the sinus node may be reset (i.e., they occur at different times than they would have if the shock had not been delivered), but the frequency and amplitude of the oscillation will remain the same. A wave of excitation initiated in the sinus node travels through the atria, then through the atrioventricular node, and then through specialized Purkinje fibers to the ventricles. The wave of electrical excitation is associated with a wave of mechanical contraction so that the cardiac cycle is associated with contraction and pumping of the blood through the body. The right and left atria are comparatively small chambers and act as collection points for blood. The right atrium collects blood from the body and the left atrium collects blood from the lungs. The right ventricle pumps blood to the lungs to be oxygenated, whereas the left ventricle pumps blood that has returned to the heart from the lungs to the rest of the body. The right atrium and right ventricle are separated by the tricuspid valve that prevents backflow of blood during the ventricular contraction. Similarly, the left atrium and left ventricle are separated by the mitral valve. In order to pump the blood, the ventricles are comparatively large and muscular.

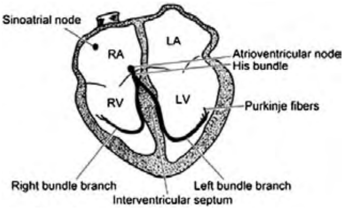

Figure 1. A schematic diagram of the heart. Adapted from Goldberger & Goldberger (1994) with permission.

In the normal ECG, there are several main deflections labeled the P wave, the QRS complex, and the T wave, Figure 2a (Goldberger & Goldberger, 1994). The P wave is associated with the electrical activation of the atria, the QRS complex is associated with the electrical activation of the ventricles, and the T wave is associated with the repolarization of the ventricles. The duration of the PR interval reflects the conduction time from the atria to ventricles, which is typically 120–200 ms. The duration of the QRS complex reflects the time that it takes for the wave of excitation to activate the ventricles. Because of the specialized Purkinje fibers, the wave of activation spreads rapidly through the ventricles so that the normal duration of the QRS complex is less than 100 ms. The time interval from the beginning of the QRS complex to the end of the T wave, called the QT interval, reflects the time that the ventricles are in the contraction phase. The

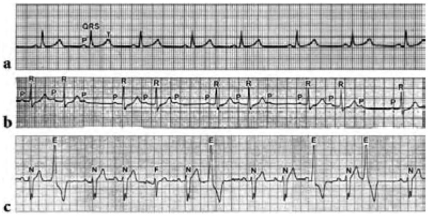

Figure 2. Sample electrocardiograms. In all traces, one large box represents 0.2 s. (a) The normal electrocardiogram. The P wave, QRS complex, and T wave are labeled. (b) 3:2 Wenckebach rhythm, an example of a second-degree heart block. There are 3 P waves for each R wave in a repeating pattern. (c) Parasystole. The normal beats, labeled N, occur with a period of about 790 ms, and the abnormal ectopic beats, labeled E, occur with a regular period of 1300 ms s. However, when ectopic beats fall too soon after the normal beats, they are blocked. Normal beats that occur after an ectopic beat are also blocked. If a normal and ectopic beat occur at the same time, the complex has a different geometry, labelled F for fusion. In this record, the number of normal beats occurring between ectopic beats is either 4, 2, or 1, satisfying the rules given in the text. Panels (a) and (b) are adopted from Goldberger & Goldberger (1994), with permission. Panel (c) is adapted from Courtemanche et al. (1989) with permission.

duration of QT interval depends somewhat on the basic heart rate. It is shorter when the heart is beating faster. For heart beats in the normal range, the QT interval is typically of the order of 300–450 ms.

On examining an ECG, one first looks for P waves. The presence of the P wave indicates that the heart beat is being generated by the normal pacemaker. In the normal heart, each P wave is followed by a QRS complex and then a T wave. The heart rate is often measured by time intervals between two consecutive R waves. Abnormally fast heart rates, faster than about 90 beats per minute, are called tachycardia, and abnormally slow heart rates, slower than about 50 beats per minute, are called bradycardia.

Reduced to the basics, all cardiac arrhythmias (i.e., abnormal cardiac rhythms) are associated with abnormal initiation of a wave of cardiac excitation, abnormal propagation of a wave of cardiac excitation, or some combination of the two. Given such a simple underlying concept, it is not surprising that mathematicians have been attracted to the study of cardiac arrhythmias, or that many cardiologists are mathematically inclined. However, despite the apparent simplicity, cardiac arrhythmias can manifest themselves in many different ways, and it is still not always possible to figure out the mechanism of an arrhythmia in any given individual. The following is focused on some arrhythmias that are well understood and that have interesting mathematical analyses.

One class of cardiac arrhythmias is associated with conduction defects through the AV node. In first-degree heart block, the PR interval is elevated above its normal value, but each P wave is followed by a QRS complex and T wave. However, in second-degree heart block, there are more P waves than QRS complexes, as some of the atrial activations do not propagate to the ventricles. This type of cardiac arrhythmia, sometimes called Wenckebach rhythms (after Karel Frederik Wenckebach, a Dutch-born Austrian physician, who studied these rhythms at the beginning of the 20th century), has repeatedly attracted theoretical interest (Katz & Hellerstein, 1964). It is common to classify Wenckebach rhythms by a ratio giving the number of P waves to the number of QRS complexes. For example, Figure 2b shows a 3:2 heart block.

In the 1920s, Balthasar van der Pol and J. van der Mark developed a mathematical model of the heart as coupled nonlinear oscillators that displayed striking similarities to the Wenckebach rhythms. We now understand that in a number of different models, as the frequency of atrial activation is increased, different types of $N : M$ heart block can be observed (van der Pol & van der Mark, 1928). In fact, theoretical models have demonstrated that if there is $N : M$ heart block at one stimulation frequency and an $N' : M'$ heart block at a higher frequency, then the $N + N' : M + M'$ heart block is expected at some intermediate stimulation frequency. This result provides a mathematical classification complementary to the cardiological classification, and can be confirmed in clinical settings (Guevara, 1991). Finally, in third-degree heart block, there is a regular atrial rhythm and a regular ventricular rhythm (at a slower frequency), but there is no coupling between the two rhythms. Such rhythms in mathematics are called quasi-periodic.

A different type of rhythm that appeals to mathematicians is called parasystole. In the “pure” case, the normal sinus rhythm beats at a constant frequency, and an abnormal (ectopic) pacemaker in the ventricles beats at a second slower frequency (Glass et al., 1986; Courtemanche et al., 1989). Figure 2c shows the normal (N) beats and the ectopic (E) beats. If the ectopic pacemaker fires at a time outside the refractory period of the ventricles, then there is an abnormal ectopic beat, identifiable on the ECG by a morphology distinct from the normal beat, and the following normal sinus beat is blocked. If the normal and abnormal beats occur at the same time, this leads to a fusion (F) beat.

Surprisingly, this simple mechanism has amazing consequences that can be appreciated by forming a sequence of integers that counts the number of sinus beats between two ectopic beats. In general, for fixed sinus and ectopic frequencies and a fixed refractory period, in this sequence, there are at most three integers, where the sum of the two smaller integers is one less than the largest integer. Moreover, given the values of the parameters, it is possible to predict the three integers. The mathematics for this problem is related to the “gaps and steps” problem in number theory. Both AV heart block and parasystole lead to mathematical predictions of cardiac arrhythmias in humans, which have been tested in experimental models and in humans. Such arrhythmias are diagnosed and treated when necessary, by physicians who have no knowledge of the underlying mathematics. Thus, to date, the mathematical analysis of these arrhythmias is of little medical interest.

From a medical perspective, the most important class of arrhythmias is called re-entrant arrhythmias. In these arrhythmias, the period of the oscillation is set by the time an excitation takes to travel in a circuitous path, rather than the period of oscillation of a pacemaker. The re-entrant circuit can be found in a single chamber of the heart or can involve several anatomical features of the heart (Josephson, 2002). In typical atrial flutter, there is a wave circulating around the tricuspid valve in the right atrium; in Wolf–Parkinson–White syndrome, there can be excitation traveling in the normal circuit from atria to the AV node to the ventricles, but then traveling retrogradely back to the atria via an abnormal accessory pathway between the ventricles and the atria. Also in some patients who have had a heart attack, there is a re-entrant circuit contained entirely in the ventricles. In all these three re-entrant arrhythmias, a part of the circuit is believed to be a comparatively thin strand of tissue.

Considering these re-entrant arrhythmias from a mathematical perspective, the wave often appears to be circulating on a one-dimensional ring. This conceptualization developed by cardiologists has an important implication for therapy: "if you cut the ring, you can cure the rhythm." By inserting catheters directly into a patient's heart and delivering radio frequency radiation to precisely identified loci, the cardiologist destroys heart tissue and can often cure these serious arrhythmias. In these cases, the cardiologist is thinking like a topologist since changing the topology of the heart cures the arrhythmia. Moreover, there is a body of mathematics that has studied the properties of excitation traveling on one-dimensional rings (Glass et al., 2002).

Other reentrant arrhythmias are not as well understood and are not as easily treated. Many theoretical and experimental studies (*See* **Cardiac muscle models**) have documented spiral waves circulating stably in two dimensions and scroll waves circulating in three dimensions (Winfree, 2001). Since real hearts are three-dimensional, and there is still no good technology to image excitation in the depth (as opposed to the surface) of the cardiac tissue, the actual geometry of excitation waves in cardiac tissue associated with some arrhythmias is not as well understood and is now the subject of intense study.

From an operational point of view, it is suggested that any arrhythmia that cannot be cured by a small localized lesion in the heart is a candidate for a circulating wave in two or three dimensions. Such rhythms include atrial and ventricular fibrillation. In these rhythms, there is evidence of strong fractionation (breakup) of excitation waves giving rise to multiple small spiral waves and patterns of shifting blocks. Tachycardias can also arise in the ventricles in patients other than those who have experienced a heart attack, or perhaps occasionally in hearts with completely normal anatomy, and in these individuals it is likely that spiral and scroll waves are the underlying geometries of the excitation. A particularly dangerous arrhythmia, polymorphic ventricular tachycardia (in which there is a continually changing morphology of the ECG complexes), is probably associated with spiral and scroll waves that undergo a meander. New technologies are presenting unique opportunities to image cardiac arrhythmias in model systems and the clinic, and nonlinear dynamics is suggesting new strategies for controlling cardiac arrhythmias. For a summary of advances up until 2002 see the collection of papers in Christini & Glass (2002).

Despite the great advances in research and medicine over the past 100 years, there is still a huge gap between what is understood and what actually happens in the human heart. The only way to appreciate this gap is to toss out the models and start looking at real data from patients who are experiencing complex arrhythmia as measured on the ECG. All who plan to model cardiac arrhythmias are encouraged to take this step.

Leon Glass

See also **Cardiac muscle models; Scroll waves; Spiral waves; Van der Pol equation**

Further Reading

Christini, D. & Glass, L. (editors). 2002. Focus issue: mapping and control of complex cardiac arrhythmias. *Chaos*, 12(3)

Courtemanche, M., Glass, L., Bélair, J., Scagliotti, D. & Gordon, D. 1989. A circle map in a human heart. *Physica* D, 40: 299–310

Glass, L., Goldberger, A., & Bélair, J. 1986. Dynamics of pure parasystole. *American Journal of Physiology*, 251: H841–H847

Glass, L., Nagai, Y., Hall, K., et al. 2002. Predicting the entrainment of reentrant cardiac waves using phase resetting curves. *Physical Review* E, 65: 021908

Goldberger, A.L. & Goldberger, E. 1994. *Clinical Electrocardiography: A Simplified Approach*, 5th edition, St Louis: Mosby

Guevara, M.R. 1991. Iteration of the human atrioventricular (AV) nodal recovery curve predicts many rhythms of AV block. In *Biomechanics, Biophysics, and Nonlinear Dynamics of Cardiac Function*, edited by L. Glass, P. Hunter, P. & A., McCulloch Theory of Heart: New York: Springer, pp. 313–358

Josephson, M.E. 2002. *Clinical Cardiac Electrophysiology: Techniques and Interpretations*, Philadelphia and London: Lippincott Williams & Wilkins

Katz, L.N. & Hellerstein, H.K. 1964. Electrocardiography. In *Circulation of the Blood: Men and Ideas*, edited by A.P. Fishman & D.W. Richards, Oxford and New York: Oxford University Press, pp. 265–354

van der Pol, B. & van der Mark, J. 1928. The heartbeat considered as a relaxation oscillation, and an electrical model of the heart. *Philosophical Magazine*, 6: 763–765

Winfree, A.T. 2001. *The Geometry of Biological Time*, 2nd edition, Berlin and New York: Springer

CARDIAC MUSCLE MODELS

Cardiac muscle was created by evolution to pump the blood, and contractions of cardiac muscle, as of any muscle, are governed by calcium (Ca) ions. Increased concentration of calcium ions inside a cardiac cell ($[\mathrm{Ca}]_i$) induces a contraction, and diminished concentration induces relaxation (diastole). Calcium ion concentration in a cardiac cell is governed by many mechanisms. Importantly, a signal to increase $[\mathrm{Ca}]_i$ is given by an abrupt increase in membrane potential E, which is called an action potential (AP). The membrane potential E in cardiac cells is described by reaction diffusion equations.

Cardiac models have the mathematical structure of the well-known Hodgkin–Huxley (HH) equations:

$$\partial E/\partial t = -I(E, g_1, \ldots, g_N))/C + \nabla\cdot(D\nabla E), \quad (1)$$

$$\partial g_i/\partial t = (\tilde{g}_i(E) - g_i)/\tau_i(E), \; i = 1, \ldots, N, \quad (2)$$

where g_i are the gating variables (describing opening and closing of the gates of ionic channels), $\tilde{g}_i(E)$ are the steady values of those variables, $\tau_i(E)$ are associated time constants, $I(\cdots)$ is the transmembrane ionic current, and the diffusive term $\nabla \cdot (D\nabla E)$ describes the current flowing from the neighboring cells. In this description of the coupling, the anisotropy of the heart is properly described with an anisotropic diffusivity tensor. Equations (1) and (2) are similar to standard HH equations (with $N = 3$), and this formulation was used for the first cardiac model introduced by Denis Noble in 1962. As more and more ionic channels were discovered, they were incorporated into more detailed models: $N = 6$ for a modified Beeler–Reuter (BR) model, and $N > 10$ for the Luo–Rudy model and for recent Noble models. Although the original HH formulation was based on analytic functions for the $\tilde{g}_i(E)$ and $\tau_i(E)$, experimentally measured functions can be used in the equations, as shown in Figure 1. This results in both physical transparency and faster numerical calculations.

The functions $g_i(E)$ and $\tau_i(E)$ have a simple physical interpretation—the dynamics of each gating variable g_i are governed by the membrane potential E only. For fixed E, Equations (2) are linear and independent; each of them describes relaxation to the steady value $\tilde{g}_i(E)$ with the time constant $\tau_i(E)$. The characteristic times $\tau_i(E)$ scan four orders of magnitude (see Figure 1(a)), leading to qualitative understanding using time-scale separation.

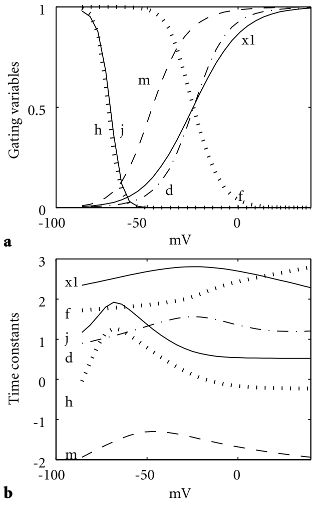

Figure 1. BR model with eight variables. (a) Gating variables $\tilde{g}_i(E)$. (b) Time constants $\tau_i(E)$ (log scale).

Noble's 1962 Model

This original cardiac model (N4 model), which is the key for understanding all subsequent models, consists of the following four equations (Noble, 1962).

$$\partial E/\partial t = -(I_{\mathrm{Na}}(E,m,h) + I_{\mathrm{K}}(E,n))/C + D\nabla^2 E, \tag{3}$$

$$\partial n/\partial t = (\tilde{n}(E) - n)/\tau_n, \tag{4}$$

$$\partial m/\partial t = (\tilde{m}(E) - m)/\tau_m, \tag{5}$$

$$\partial h/\partial t = (\tilde{h}(E) - h)/\tau_h, \tag{6}$$

where

$$I_{\mathrm{Na}}(E,m,h) = (g_{\mathrm{Na}} m^3 h + \varepsilon_1)(E - E_{\mathrm{Na}}), \tag{7}$$

$$I_{\mathrm{K}}(E,n) = (\alpha g_{\mathrm{K}} n^4 + \varepsilon_2)(E - E_{\mathrm{K}}), \tag{8}$$

A cardiac action potential has a duration of about 200 ms, which is about two orders of magnitude longer than that of a typical nerve impulse. To describe cardiac action potential, just one time constant in the HH equations was increased by a factor of 100, two small terms (ε_1 and ε_2) were added, and other time constants were adjusted to incorporate cardiac experimental results.

To observe how the N4 model works, note that it has three gating variables: m, n, and h. The characteristic time τ_n is about 100 ms, while τ_m and τ_h are almost two orders of magnitude faster (shorter). This permits adiabatic elimination of the fast equations (5) and (6), thereby replacing Equations (3)–(6) with a system of equations only (system N2) (Krinsky & Kokoz, 1973):

$$\partial E/\partial t = -(I_{\mathrm{Na}}(E,\tilde{m},\tilde{h}) + I_{\mathrm{K}}(E,n))/C + D\nabla^2 E, \tag{9}$$

$$\partial n/\partial t = (\tilde{n}(E) - n)/\tau_n. \tag{10}$$

Note that the current I_{Na} becomes

$$I_{\mathrm{Na}} = (g_{\mathrm{Na}} \tilde{m}^3(E)\tilde{h}(E) + \varepsilon_1)(E - E_{\mathrm{Na}}), \tag{11}$$

which contains only the known functions $\tilde{m} \equiv \tilde{m}(E)$ and $\tilde{h} \equiv \tilde{h}(E)$. The behavior of this model is illustrated in Figure 2.

Nullclines $\mathrm{d}E/\mathrm{d}t = 0$, and $\mathrm{d}n/\mathrm{d}t = 0$ for a Purkiknje fiber are shown in Figure 2(a). A Purkinje fiber has a pacemaker activity, which the nullclines (in the phase plane) show directly. Note that there is only one fixed point S, which is unstable, thus giving rise to limit cycle oscillations.

The nullclines for a myocyte are shown in Figure 2(b). There is no pacemaker activity in myocytes; thus, the nullclines show the absence of a limit cycle. Although there are two unstable fixed points (S′ and S″), they do not induce a limit cycle because there is a third fixed point (S) that is stable and determines the resting potential.

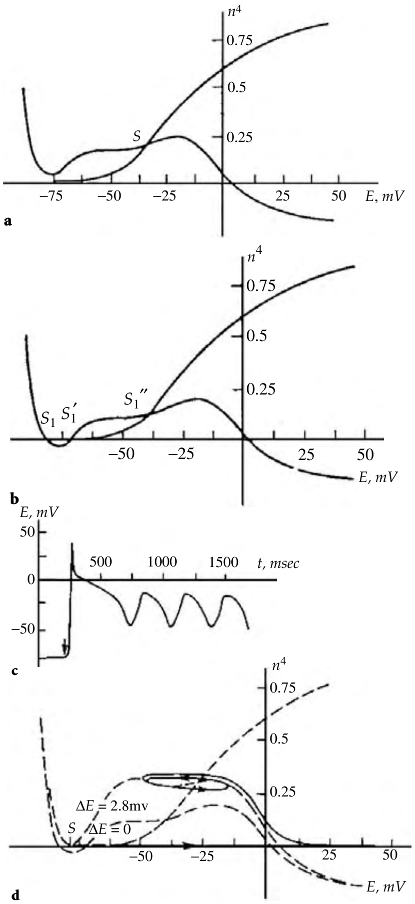

Figure 2. Analysis of the N4 model based on adiabatic elimination of two fast variables. (a) A Purkiknje fiber with a pacemaker activity. The nullclines of the reduced system N2 contain only one fixed point S that is unstable, thus giving rise to the limit cycle oscillations. (b) An excitable myocyte. There are three fixed points: S′ and S″ are unstable, and point S determines the resting potential. (c) Analysis of the effect of an arrhythmogenic drug aconitine, showing oscillations of the plateau of action potential (AP). (d) As in (c), where the nullclines show a small-amplitude limit cycle on the plateau of AP. (Parameters: (a) $\alpha = 1.2$; (b) $\alpha = 1.3$; (c) and (d) $\alpha = 1.3$; h = $\tilde{h}(E - \delta E)$; $\delta E = 2.8$ mV.)

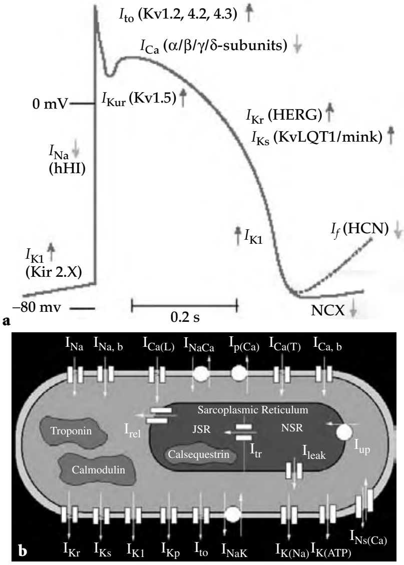
I_{to} (Kv1.2, 4.2, 4.3)
I_{Ca} (α/β/γ/δ-subunits)
0 mV
I_{Kur} (Kv1.5)
I_{Kr} (HERG)
I_{Ks} (KvLQT1/mink)
I_{Na} (hHI)
I_{K1}
I_f (HCN)
I_{K1} (Kir 2.X)
−80 mv
0.2 s
NCX
a
I_{Na}
$I_{Na,b}$
$I_{Ca(L)}$
I_{NaCa}
$I_{p(Ca)}$
$I_{Ca(T)}$
$I_{Ca,b}$
Sarcoplasmic Reticulum
Troponin
I_{rel}
JSR
NSR
Calsequestrin
I_{tr}
I_{leak}
I_{up}
Calmodulin
I_{Kr}
I_{Ks}
I_{K1}
I_{Kp}
I_{to}
I_{NaK}
$I_{K(Na)}$
$I_{K(ATP)}$
$I_{Ns(Ca)}$
b

Figure 3. (a) Action potential showing the principal currents that flow in each phase, with the corresponding subunit clones shown in parentheses. (Courtesy of S. Nattel) (b) Main ionic currents (Courtesy of Y. Rudy).

Analysis of the effect of the arrhythmogenic drug aconitine (inducing oscillations on the plateau of the action potential) is shown in Figures 2(c) and (d). Aconitine induces dangerous oscillations because of a shift of the voltage dependence of Na inactivation variable h. Oscillations on the plateau of action potential are shown in Figure 2(c), and the corresponding nullclines are shown in Figure 2(d). The shift of $h(E)$ dependence results in the disappearance of the stable resting point and appearance of a small-amplitude limit cycle on the plateau of the AP. The electrophysiological characteristics in the full (N4) and reduced (N2) models are the same with an accuracy of 0.1–0.2 mV (Krinsky & Kokoz, 1973).

Contemporary Models

Recent models include more ionic currents (see Figures 1 and 3), and also incorporate a change of intracellular ionic concentrations. For example, the BR model contains an additional equation for the concentration $[\mathrm{Ca}]_i$ of intracellular Ca:

$$\partial Ca_i / \partial t = I_{\mathrm{Ca}}. \qquad (12)$$

The Luo–Rudy (LR) model, and the Noble model are widely known. The LR model (Rudy, 2001) can be downloaded from http://www.cwru.edu/med/CBRTC/LRdOnline. Noble models are described at http://www.cellml.org/ and http://cor.physiol.ox.ac.uk/. The model of human ventricular myocyte (Ten Tusscher et al.

2003) is also available at http://www-binf.bio.uu.nl/khwjtuss/HVM.

In contemporary models the gap between the 100 and 1 ms time scales has been filled by many ionic currents, and (contrary to what was seen in the first cardiac model) the graphs even intersect each other (see Figure 1(b)). This makes the separation for fast and slow variables dependent upon the phase of AP, so results as in Figure 2 cannot be obtained directly. Instead, events on every time scale must be analyzed separately, eliminating adiabatically equations with a faster time scale and considering variables with a larger time scale fixed, or even postulating a model with only two to four equations (Keener & Sneyd, 1998; Fenton & Karma, 1998).

Models that do not follow the HH formalism are needed because HH-type models predict that a point stimulation will create a circular or an elliptical distribution of membrane potential, while in the experiments a quadrupolar distribution was found. Thus, *bidomain models* were created that describe separately potentials inside (E_i) and outside (E_o) of a cell instead of considering $E_\mathrm{o} = 0$ as in the HH formulation. These models correctly reproduce many important electrophysiological effects, and turn out to be useful for understanding the mechanisms of cardiac defibrillation (Trayanova et al., 2002). The integration of these models, however, is computationally more expensive than the integration of HH models. *Markov chain models* are used to describe transitions between states of single ionic channels linking genetical defects with cardiac arrhythmias. They also depart from the HH formalism. *Anatomical models* incorporate cardiac geometry and tissue structure, but they require months of laborious measurements on anatomical slices cut from the heart. A new approach is being developed at INRIA in France, which aims to create models for every cardiac patient and is intended to be used in clinics. Cardiac contractions are measured and incorporated into the model. NMR tomography methods were used that permit obtaining anatomical data in a few minutes, not months, see http://www-sop.inria.fr/epidaure/. More cardiac models and authors can be found at http://www.cardiacsimulation.org/.

Dynamics of Myocardial Tissue

In the past, breakthroughs were due to very simple models beginning with the pioneering work of Norbert Wiener and Arturo Rosenbluth (Wiener & Rosenbluth, 1946). They led the way to understanding rotating waves using a cellular automata model, where a cardiac cell can be in three states only: rest, excitation, and refractory. This model explained anatomical reentry (a wave rotating around an obstacle (Wiener & Rosenbluth, 1946)), and predicted a free rotating wave. Partial differential equations yield more refined refined results (Keener & Sneyd, 1998).

As two- and three-dimensional studies of rotating waves are time consuming, it is often convenient to use a two-variable model, permitting increased speed of calculations by two orders of magnitude. Numerical simulation of ionic models can be accelerated either by adiabatic elimination of fast variable m (Na) activation or by slowing down its dynamics and increasing the diffusion coefficient to keep the propagation velocity unchanged.

Propagation Failure

As cardiac cells are connected via gap junctions, an excitation can be blocked when propagating from one cardiac cell to another, similar to the propagation failure in myelinated nerves. When an excitation propagates from auricles (A) to ventricles (V) via the AV node, a periodic pattern can be observed: for example, from every three pulses, only two pulses propagate, and the third pulse is blocked (3:2 Wenckebach periodicity). Other periodicities $N : (N - 1)$ can also be observed. Propagation block can be observed on any cardiac heterogeneity when the period T of stimulation is shorter than the restoration time R (refractory period). Usually, Wenckebach rhythms with only small periods N are observed because

$$T_N \sim N^{-2}, \tag{13}$$

where T_N is an interval of T that can yield. Wenckebach rhythms with period N. When an excitation block (Wenckebach rhythm) occurs in a two- or three-dimensional excitable medium, it generically gives rise to a wave break that evolves into a rotating wave. For cardiac muscle, initiation of a rotating wave is a dangerous event, often leading to life-threatening cardiac arrhythmias.

A new approach is being developed at INRIA in France, to create patient-specific models to be used in clinics. A 3-d electro-mechanical model of the heart is automatically adpted to a time series of volumetric cardiac images gated on the ECG (Ayache et al., 2001) providing useful quantitative parameters on the heart function in a few minutes, not months, see http://www-sop.inria.fr/epidaure/.

V. KRINSKY, A. PUMIR, AND I. EFIMOV

See also **Hodgkin–Huxley equations; Myelinated nerves; Neurons; Scroll waves; Synchronization; Van der Pol equation**

Further Reading

Ayache, N., Chapelle, D., Clément, F., Coudiére, Y., Delingette, H., Désidéri, J.A., Sermesant, M. Sorine, M. & Urquiza, J. 2001. Towards model-based estimation of the cardiac electro-mechanical activity from ECG signals and ultrasound images. In *Functional Imaging and Modeling of the Heart* (*FIMH'01*), *Helsinki, Finland*, Lecture Notes in Computer Sciences, vol. 2230, Berlin, Springer, pp. 120–127

Chaos, topical issue: Ventricular fibrillation, 8(1), 1998
Fenton, F. & Karma, A. 1998. Vortex dynamics in three-dimensional continuous myocardium with fiber rotation: Filament instability and fibrillation. *Chaos*, 8: 20–47
Journal of Theoretical Biology, topical issue devoted to the work of Winfree, 2004
Keener, J. & Sneyd, J. 1998. *Mathematical Physiology*, New York: Springer
Krinsky, V. & Kokoz, Ju. 1973. Membrane of the Purkinje fiber. reduction of the noble equations to a second order system. *Biophysics*, 18: 1133–1139
Noble, D. 1962. A modification of the Hodgkin–Huxley equations applicable to Purkinje fiber action and pacemaker potential. *Journal of Physiology*, 160: 317–352
Noble, D. & Rudy, Y. 2001. Models of cardiac ventricular action potentials: iterative interaction between experiment and simulation. *Philosophical Transactions of the Royal Society A*, 359: 1127–1142
Rudy, Y. 2001. The cardiac ventricular action potential. In *Handbook of Physiology: A Critical, Comprehensive Presentation of Physiological Knowledge and Concepts. Section 2, The cardiovascular system. vol. 1, the heart*, edited by E. Page, H.A. Fozzard & R.J. Solaro, Oxford: Oxford University Press, pp. 531–547
Pumir, A., Romey, G. & Krinsky, V. 1998. De-excitation of cardiac cells. *Biophysical Journal*, 74: 2850–2861
Sambelashvili, A. & Efimov, I.R. 2002. Pinwheel experiment re-revisited. *Journal of Theoretical Biology*, 214: 147–153
TenTusscher, K.H.W.J., Noble, D., Noble, P.J. & Panfilov, A.V. 2003. A model of the human ventricular myocyte. *American Journal of Physiology*, 10: 1152
Trayanova, N., Eason, J. & Aguel, F. 2002. Computer simulations of cardiac defibrillation: a look inside the heart. *Computing and Visualization in Science*, 4: 259–270
Wiener, N. & Rosenbluth, N. 1946. The mathematical formulation of the problem of conduction of impulses in a network of connected excitable elements, specifically in cardiac muscle. *Archivos del Instituto de cardiologia de Mexico*, 16: 205–265

CASIMIRS

See **Poisson brackets**

CAT MAP

The cat map is perhaps the simplest area-preserving transformation that exhibits a high degree of chaos. In the development of the theory of dynamical systems, it served as a guiding example to illustrate new concepts such as entropy (Sinai, 1959) and Markov partitions (Adler & Weiss, 1967). The cat map owes its name to an illustration by V.I. Arnol'd showing the image of a cat before and after the application of the map. In the mathematical literature it is also referred to as "hyperbolic toral automorphism."

The torus M, which topologically has the shape of a doughnut, may be described by the points in the unit square (see Figure 1) where opposite sides are identified. Alternatively, one may think of a point on M as a point in the plane $\mathbb{R}^2$ modulo integer translations in $\mathbb{Z}^2$. This yields a representation of M as the coset space $\mathbb{R}^2/\mathbb{Z}:=\{x+\mathbb{Z}^2 : x\in\mathbb{R}^2\}$. The cat map is now a mapping $\phi : M \to M$ defined by $x \mapsto Ax \pmod{\mathbb{Z}^2}$, where

$$x=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad A=\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{1}$$

provided $a, b, c, d \in \mathbb{Z}$ are chosen such that $|\det A| = 1$ and A has eigenvalues $\lambda_\pm$ with modulus not equal to one. (This implies that both eigenvalues are real and distinct.) The matrix A used in our illustration is

$$A=\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}. \tag{2}$$

Let us now explore some of the dynamical properties that show that the cat map is indeed completely "chaotic" (*See* **Chaotic dynamics**). Sensitivity on initial conditions is measured by the rate of divergence of two nearby points, x and $x+\delta x$, under iterations of ϕ. For any smooth map ϕ, the Taylor expansion for small δx yields $\phi(x+\delta x)=\phi(x)+D\phi_x\delta x+O(\delta x^2)$ where $D\phi_x$ is the differential of ϕ at x; it may be viewed as a linear map from TM_x to $TM_{\phi(x)}$, the tangent spaces at x and $\phi(x)$, respectively.

Because ϕ is linear in the case of the cat map, the above Taylor expansion is in fact exact, that is, $\phi(x+\delta x)=\phi(x)+D\phi_x\,\delta x$ with $D\phi_x=A$. If $v_\pm$ are the eigenvectors of A corresponding to the eigenvalues $\lambda_\pm$, let us denote by E_x^+ and E_x^- the subspaces of TM_x spanned by v_+ and v_-, respectively. As $\lambda_+ \neq \lambda_-$, we have

$$TM_x = E_x^+ \oplus E_x^-. \tag{3}$$

Furthermore, since $|\lambda_+| = 1/|\lambda_-| > 1$, we find

$$\|D\phi_x(\xi)\| \geq \mathrm{e}^{\lambda}\|\xi\| \quad \text{if } \xi\in E_x^+, \tag{4}$$

$$\|D\phi_x(\xi)\| \leq \mathrm{e}^{-\lambda}\|\xi\| \quad \text{if } \xi\in E_x^-, \tag{5}$$

where $\lambda=\ln|\lambda_+|>0$ and $\|\cdot\|$ is the Euclidean norm. Hence, ϕ is expanding in the direction of v_+, and contracting in the direction of v_-, which will therefore be referred to as the unstable and stable directions, respectively. [Here, inequalities (4) and (5) are in fact equalities; inequalities become necessary in the case of more general Anosov maps, if one seeks uniform bounds with λ independent of x.] For the nth forward or backward iterates of the map (n a positive integer), we have, by the above arguments with ϕ replaced by $\phi^{\pm n}$,

$$\|D\phi_x^n(\xi)\| \geq \mathrm{e}^{n\lambda}\|\xi\|, \quad \|D\phi_x^{-n}(\xi)\| \leq \mathrm{e}^{-n\lambda}\|\xi\|, \quad \text{if } \xi\in E_x^+, \tag{6}$$

$$\|D\phi_x^n(\xi)\| \leq \mathrm{e}^{-n\lambda}\|\xi\|, \quad \|D\phi_x^{-n}(\xi)\| \geq \mathrm{e}^{n\lambda}\|\xi\|, \quad \text{if } \xi\in E_x^-. \tag{7}$$

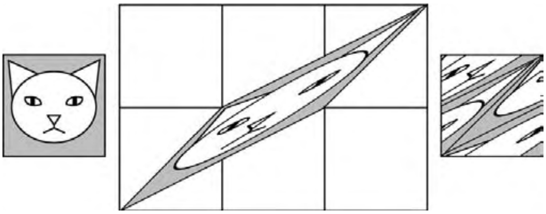

Figure 1. The cat map: the image of a cat in the unit square (left) is stretched by the matrix A (middle) and then re-assembled by cutting and translating (without rotation) the different parts of the cat's face back into the unit square (right). (Illustration by Federica Vasetti.)

The expansion/contraction is thus exponentially fast in time. Relations (3), (6), and (7) are equivalent to the statement that the cat map is an Anosov system. Special features of Anosov systems are ergodicity, mixing, structural stability, exponential proliferation of periodic orbits, and positive entropy h.

There is a particularly simple formula for the entropy h due to Sinai (1959), which states that $h = \lambda = \ln|\lambda_+|$. The number of fixed points of the nth iterate ϕ^n is equal to $|\det(1 - A^n)| = |(1 - \lambda_+^n)(1 - \lambda_-^n)|$ and is therefore asymptotically given by $\exp(nh)$, for large n. The notion of ergodicity implies that for any $f \in L^1(M, \mathrm{d}x)$, for almost every $x \in M$ (that is, for all x up to a set of Lebesgue measure zero), we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(\phi^n x) = \langle f \rangle, \quad \langle f \rangle := \int_M f(x)\,\mathrm{d}x. \tag{8}$$

"Mixing" means that, if $f, g \in L^2(M, \mathrm{d}x)$, then

$$\lim_{n\to\pm\infty} \int_M f(\phi^n x)\, g(x)\,\mathrm{d}x = \langle f \rangle \langle g \rangle. \tag{9}$$

Although the mixing property follows from general arguments for Anosov systems, it can be proved for ϕ directly by means of Fourier analysis. What is more, the rate of convergence in (9) is in fact exponentially fast in n for suitably smooth test functions f, g ("exponential decay of correlations") and super-exponentially fast for analytic ones.

Markov partitions are a powerful tool in the analysis of dynamical systems. In the case of the cat map the torus is divided into a finite collection of non-overlapping parallelograms $P_1, \ldots, P_N$ whose sides point in the directions of the eigenvectors v_+ and v_-, such that if $\phi(P_i)$ (or $\phi^{-1}(P_i)$) intersects with P_j, then it extends all the way across P_j. Let us construct an N by N matrix B whose coefficients are $B_{ij} = 1$ if the intersection of $\phi(P_i)$ with P_j is non-empty, and $B_{ij} = 0$ otherwise. A symbolic description of the dynamics of the cat map can now be obtained as follows. Consider doubly infinite sequences of the form $\cdots b_{-2}b_{-1}\underline{b_0}b_1b_2\ldots$, where b_n is an integer $1, \ldots, N$ with the condition that the number b_n can be followed by b_{n+1} only if $B_{b_n b_{n+1}} = 1$. To each such sequence we can associate a point x on M by requiring that for every $n \in \mathbb{Z}$, the parallelogram R_{b_n} contain the iterate $\phi^n(x)$. The symbolic dynamics of ϕ is now given by shifting the mark by one step to the right: the new word $\cdots b_{-1}b_0\underline{b_1}b_2b_3\cdots$ indeed represents the point $\phi(x)$. The dynamical properties of ϕ are thus encoded in the matrix B. In particular, the higher the rate of mixing of ϕ, the smaller the number of coefficients B_{ij} that are zero. This in turn means that we have fewer restrictions on $b_n \mapsto b_{n+1}$ and a typical orbit will be represented by a more "random" sequence of symbols.

Cat maps, as well as higher-dimensional toral automorphisms, are featured in most textbooks on dynamical systems and ergodic theory. An introduction to the basic concepts may be found in Arnol'd & Avez (1968), and more advanced topics, such as entropy, Markov partitions, and structural stability, are discussed, for example, in Adler & Weiss (1970), Katok & Hasselblatt (1995), Pollicott & Yuri (1998), and Shub (1987).

JENS MARKLOF

See also **Anosov and Axiom-A systems; Chaotic dynamics; Horseshoes and hyperbolicity in dynamical systems; Maps; Markov partitions; Measures**

Further Reading

Adler, R.L. & Weiss, B. 1967. Entropy, a complete metric invariant for automorphisms of the torus. *Proceedings of the National Academy of Sciences of the United States of America*, 57: 1573–1576

Adler, R.L. & Weiss, B. 1970. *Similarity of Automorphisms of the Torus*, Providence, RI: American Mathematical Society

Arnol'd, V.I. & Avez, A. 1968. *Ergodic Problems of Classical Mechanics*, New York and Amsterdam: Benjamin

Katok, A. & Hasselblatt, B. 1995. *Introduction to the Modern Theory of Dynamical Systems*, Cambridge and New York: Cambridge University Press

Pollicott, M. & Yuri, M. 1998. *Dynamical Systems and Ergodic Theory*, Cambridge and New York: Cambridge University Press

Sinai, Ya.G. 1959. On the concept of entropy for a dynamic system. *Doklady Akademii Nauk SSSR*, 124: 768–771

Shub, M. 1987. *Global Stability of Dynamical Systems*, Berlin and New York: Springer

CATALYTIC HYPERCYCLE

The concept of the "hypercycle" was invented in the 1970s in order to characterize a functional entity that integrates several autocatalytic elements into an organized unit (Eigen, 1971; Eigen & Schuster, 1977, 1978a,b). This concept is a key to understanding the dynamics of living organisms.

A catalytic hypercycle is defined as a cyclic network of autocatalytic reactions (Figure 1). Autocatalysts, in general, compete when they are supported by the same source of energy or material. Hypercyclic coupling introduces mutual dependence of elements and suppresses competition. Consequently, the fate of all members of a hypercycle is identical to that of the entire system and, in other words, no element of a hypercycle dies out provided the hypercycle as such survives.

The current view of biological evolution distinguishes periods of dominating Darwinian evolution based on variation, competition, and selection interrupted by rather short epochs of radical innovations often called *major transitions* (Maynard Smith & Szathmáry, 1995; Schuster, 1996). In the course of biological evolution major transitions introduce higher hierarchical levels. Examples are (i) the origin of translation from nucleic acid sequences into proteins including the invention of the genetic code, (ii) the transition from independent replicating molecules to chromosomes and genomes, (iii) the transition from the prokaryotic to the eukaryotic cell, (iv) the transition from independent unicellular individuals to differentiated multicellular organisms, (v) the transition from solitary animals to animal societies, and (vi) presumably a series of successive transitions from animal societies to humans. All major transitions introduce a previously unknown kind of cooperation into biology. The hypercycle is one of very few mechanisms that can deal with cooperation of otherwise competing individuals. It is used as a model system in prebiotic chemistry, evolutionary biology, theoretical economics, as well as in cultural sciences.

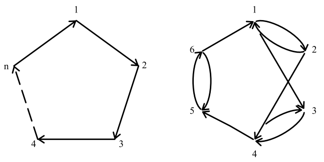

Figure 1. Definition of hypercycles. Replicator equations as described by the differential equation (2) can be symbolized by directed graphs: the individual species are denoted by nodes and two nodes are connected by an edge, $j\ \cdot \longrightarrow \cdot\ i$, if and only if $a_{ij} > 0$. The graphs of hypercycles consist of single Hamiltonian arcs as sketched on the left-hand side of the figure. These dynamical systems are permanent independent of the choice of rate parameters f_i. For $n \leq 5$ they represent the only permanent systems, but for $n \geq 6$ the existence of a single Hamiltonian arc is only a sufficient but not a necessary condition for permanence. The graph on the right-hand side, for example, does not contain a Hamiltonian arc but the corresponding replicator equation is permanent for certain choices of rate parameters (Hofbauer & Sigmund, 1998).

The simplest example of a catalytic hypercycle is the *elementary hypercycle*. It is described by the dynamical system

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i \left(f_i x_{i-1} - \sum_{j=1}^{n} f_j x_{j-1} x_j \right); \quad i, j = 1, 2, \ldots, n;\ i, j = \mathrm{mod}\, n. \qquad (1)$$

The catalytic interactions within a hypercycle form a directed closed loop comprising all elements, often called Hamiltonian arc: $1 \rightarrow 2 \rightarrow 3 \rightarrow \cdots \rightarrow n \rightarrow 1$ (Figure 1). Hypercycles are special cases of replicator equations of the class

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i \left(\sum_{j=1}^{n} a_{ij} x_j - \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} x_j x_k \right); \quad i, j, k = 1, 2, \ldots, n \qquad (2)$$

with $a_{ij} = f_i \cdot \delta_{i-1,j}$; $i, j = \mathrm{mod}\, n$. (The '$\mathrm{mod}\, n$' function implies a cyclic progression of integers, $1, \ldots, n-1, n, 1, \ldots$. The symbol $\delta_{i,j}$ represents Kronecker's symbol: $\delta = 1$ for $i = j$ and $\delta = 0$ for $i \neq j$.)

For positive rate parameters and initial conditions inside the positive orthant (the notion of an *orthant* refers to the entire section of a Cartesian coordinate system in which the signs of variables do not change. In n-dimensions, the positive orthant is defined by $\{x_i > 0\ \ \forall i = 1, 2, \ldots, n\}$.). The trajectory of a hypercycle remains within the orthant:

$$f_i > 0,\ x_i(0) > 0\ \forall i = 1, 2, \ldots, n \Longrightarrow x_i(t) > 0\ \forall t \geq 0.$$

In other words, none of the variables is going to vanish and hence, the system is *permanent* in the sense that no member of a hypercycle dies out in the limit of long times, $\lim_{t \to \infty} x_i(t) \neq 0\ \ \forall i = 1, \ldots, n$. The existence of a Hamiltonian arc, that is, a closed loop of directed edges visiting all nodes once, is a sufficient condition for permanence (Hofbauer & Sigmund, 1998). It is also a necessary condition for low-dimensional systems with $n \leq 5$, but there exist permanent

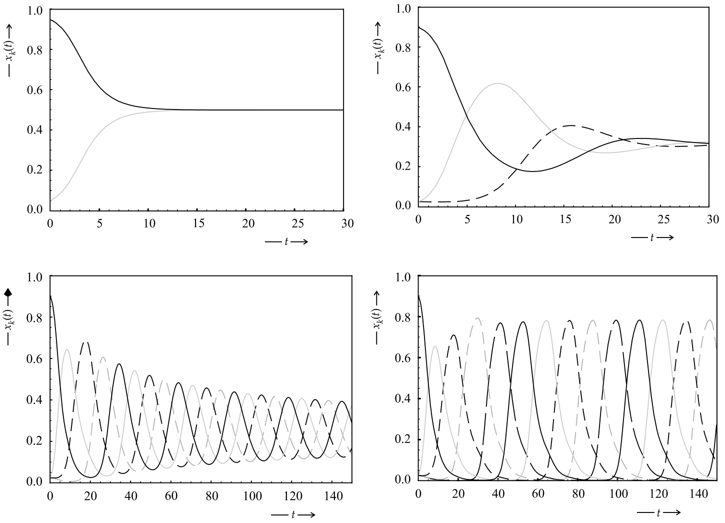

Figure 2. Solution curves of small elementary hypercycles. The figure shows the solution curves of Equation (1) with $f_1 = f_2 = \ldots = f_n = 1$ for $n = 2$ (upper left picture), $n = 3$ (upper right picture), $n = 4$ (lower left picture), and $n = 5$ (lower right picture). The initial conditions were $x_1(0) = 1 - (n - 1) \cdot 0.025$ and $x_k(0) = 0.025 \, \forall \, k = 2, 3, \ldots, n$. The sequence of the curves $x_k(t)$ is $k = 1$ full black line, $k = 2$ full gray line, $k = 3$ hatched black line, $k = 4$ hatched grey line, and $k = 5$ black line with long hatches. The cases $n = 2$, 3, and 4 have stable equilibrium points in the middle of the concentration space $c = (1/n, 1/n, \ldots, 1/n)$; Equation (1) with equal rate parameters, $n = 4$, and linearized around the midpoint c exhibits a marginally stable "center" and very slow convergence is caused by the nonlinear term, which becomes smaller as the system approaches c. For $n = 5$, the midpoint c is unstable and the trajectory converges toward a limit cycle (Hofbauer et al., 1991).

dynamical systems for $n \geq 6$ without a Hamiltonian arc; one example is shown in Figure 1.

The dynamics of Equation (1) remains qualitatively unchanged when all rate parameters are set equal: $f_1 = f_2 = \ldots = f_n = f$, which is tantamount to a *barycentric* transformation of the differential equation (Hofbauer, 1981). The hypercycle is invariant with respect to a rotational change of variables, $x_i \Longrightarrow x_{i+1}$ with $i = 1, 2, \ldots, n$; $i \bmod n$, it has one equilibrium point in the center, and its dynamics depends exclusively on n. Some examples with small n are shown in Figure 2. The systems with $n \leq 4$ converge toward stable equilibrium points, whereas the trajectories of Equation (1) with $n \geq 5$ approach limit cycles. Independent of n, elementary hypercycles do not sustain chaotic dynamics.

Hypercycles have two inherent instabilities, which are easily illustrated for molecular species: (i) The members of the cycle may also catalyze the formation of nonmembers that do not contribute to the growth of the hypercycle, and thus hypercycles are vulnerable to parasitic exploitation (Eigen & Schuster, 1978a,b), and (ii) concentrations of individual species in oscillating hypercycles ($n \geq 5$) go through very small values, and these species might become extinct through random fluctuations. More elaborate kinetic mechanisms can stabilize the system in Case (ii). Exploitation by parasites, Case (i), can be avoided by compartmentalization. Competition between different hypercycles is characterized by a strong nonlinearity in selection (Hofbauer, 2002): once a hypercycle has been formed and established, it is very hard to replace it by another hypercycle. Epochs with hypercyclic dynamics provide explanations for "once for ever" decisions or "frozen accidents."

Peter Schuster

See also **Artificial life; Biological evolution**

Further Reading

Eigen, M. 1971. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58: 465–523

Eigen, M. & Schuster, P. 1977. The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften*, 64: 541–565

Eigen, M. & Schuster, P. 1978a. The hypercycle. A principle of natural self-organization. Part B: The abstract hypercycle. *Naturwissenschaften*, 65: 7–41

Eigen, M. & Schuster, P. 1978b. The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. *Naturwissenschaften*, 65: 341–369

Hofbauer, J. 1981. On the occurrence of limit cycles in the Volterra–Lotka equation. *Nonlinear Analysis, 5: 1003–1007*

Hofbauer, J. 2002. Competitive exclusion of disjoint hypercycles. Zeitschrift für Physikalische Chemie, 216: 35–39

Hofbauer, J. & Sigmund, K. 1998. *Evolutionary Games and Replicator Dynamics, Cambridge and New York: Cambridge University Press*

Hofbauer, J., Mallet-Paret, J. & Smith, H.L. 1991. Stable periodic solutions for the hypercycle system. Journal of Dynamics and Differential Equations, 3: 423–436

Maynard Smith, J. & Szathmáry, E. 1995. *The Major Transitions in Evolution, Oxford and New York: Freeman*

Schuster, P. 1996. How does complexity arise in evolution? Complexity, 2(1): 22–30

CATASTROPHE THEORY

Many natural phenomena (cell division, the bursting of bubbles, the collapse of buildings, and so on) involve discontinuous changes, whereas the majority of applied mathematics is directed toward modeling continuous processes. On the other hand, catastrophe theory is primarily concerned directly with singular behavior and as such deals with properties of discontinuities directly. This approach has found use in many and diverse fields and at one time was heralded as a new direction in mathematics, uniting singularity and bifurcation theories and their applications (Zeeman, 1976).

A simple mechanical system that illustrates the important ideas of discontinuous changes and hysteresis is provided by Zeeman's catastrophe machine (Zeeman, 1972, Poston & Stewart, 1996). A sketch showing its construction is given in Figure 1(a). It is recommended that readers make such devices for themselves and experiment with them.

The lines between Q, P, and the pointer represent rubber bands attached to a disk that rotates around O. Movement of the pointer such that it remains outside of the region ABCD results in a smooth motion of the wheel from one equilibrium state to another. This is illustrated by the path ℓ_1 in Figure 1(b), where x and θ are as defined in Figure 1(a) and k is a measure of the stiffness of the bands. However, starting with the pointer below AB and moving the pointer horizontally to cross AD will cause an anticlockwise jump in the wheel. This is equivalent to following the path ℓ_2 in Figure 1(b). This equilibrium will remain until AB is crossed when the pointer is moved backwards. The loci DAB form a cusp in the parameter space of the system where AD and AB are lines of folds that meet at the cusp point A. The cusp is a projection down onto the plane of a three-dimensional folded surface. The region labeled ABCD comprises four such cusps, each of which can be described by a simple cubic equation. Indeed, the set can be obtained from an approximate model for the machine (Poston & Stewart, 1996).

The term *catastrophe* was introduced by René Thom in 1972 to describe such discontinuous changes in a system where a parameter is changed smoothly. Zeeman (1976) then coined the phrase *catastrophe theory*, and an explosion of applications arose ranging from the physical to the social sciences. An important idea put forward is that there are seven elementary catastrophes that classify most types of observed discontinuous behavior. They are the fold, cusp, swallowtail, and butterfly catastrophes and the elliptic, hyperbolic, and parabolic umbilic catastrophes. These describe all possible discontinuities for potential systems that are controlled by up to four variables. They are ordered according to their typicality of occurrence with the fold being the most common. A path of folds will be represented by a line of singular behavior in parameter space and a cusp will be formed when two such lines meet. Indeed, these two singularities are sufficient to cover most of the observed macroscopic critical behavior in practical applications.

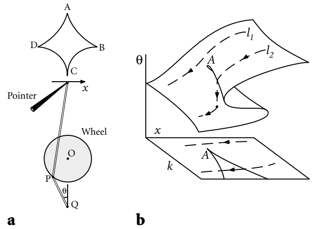

Figure 1. (a) Sketch of Zeeman's catastrophe machine, (b) A sketch of the three-dimensional solution surface and its projection onto two dimensions near the cusp A.

One area of application where catastrophe theory has been used with considerable success is in optical caustics (Nye, 1999). Common experience of this phenomenon is observation of a distant light source through a drop of water on a window pane where a web of bright lines separated by dark regions can often be seen. The bright lines on the bottom of swimming pools are also examples of caustics where the bright sunlight is focused by the surface of the water. In this case, the line caustics are examples of paths of folds in the ray surfaces. An example of an optical cusp is provided by strong sunlight focused on the surface of a cup of coffee where two principal fold lines are made to meet by the curvature of the cup. In an outstanding series of experiments (Berry, 1976), a laser beam was shone through a water drop and a whole sequence of catastrophes was uncovered including swallowtails. All of the observed patterns can be reproduced in detail using the equations of ray optics (Nye, 1999).

Catastrophe theory has also been used to explore in some detail the state selection process in Taylor–Couette flow between concentric cylinders. In this case, even numbers of vortices are generated in the flow field, and the number that is realized depends on the length of the cylinders. For a given length of the cylinder, one state develops smoothly, with control parameter and neighboring states delimited by fold lines in parameter space. The fold lines meet in a cusp that has been observed experimentally (Benjamin, 1978) and calculated numerically (Cliffe, 1988) from the Navier–Stokes equations.

There has been considerable criticism of catastrophe theory on both technical and practical grounds (Arnol'd, 1986; Arnol'd et al., 1994). For example, it is known that critical behavior or bifurcations in some multidimensional gradient systems do not reduce to critical points of potentials (Guckenheimer, 1973). Also, the ideas have been applied to a wide range of social, financial, and biological applications where the governing rules are not known or are very primitive. Very often, it is a case of re-interpretation of common experience in terms of technical mathematical language, which is most often qualitative rather than quantitative (Arnol'd, 1986). Hence, it is often the case that disparate systems superficially appear the same, but closer examination reveals that they are quite different in terms of important details.

TOM MULLIN

See also **Bifurcations; Critical phenomena; Development of singularities; Equilibrium; Taylor–Couette flow**

Further Reading

Arnol'd, V.I. 1986. *Catastrophe Theory*, Berlin and New York: Springer

Arnol'd, V.I., Afrajmovich, V.S. Il'yashenko, Yu.S. & Shil'nikov, L.P. 1994. *Bifurcation Theory and Catastrophe Theory*, Berlin and New York: Springer

Benjamin, T.B. 1978. Bifurcation phenomena in steady flows of a viscous fluid. *Proceedings of the Royal Society of London*, Series A, 359: 1–43

Berry, M.V. 1976. Waves and Thomas theorem. J *Advances in Physics*, 25(1): 1–26

Cliffe, K.A. 1988. Numerical calculations of the primary-flow exchange process in the Taylor problem. *Journal of Fluid Mechanics*, 197: 57–79

Guckenheimer, J. 1973. *Bifurcation and catastrophe*. In *Dynamical Systems*: Proceedings of the Symposium of University of Babia Salvador, 1971, pp. 95–109

Nye, J.F. 1999. *Natural Focusing and Fine Structure of Light: Caustics and Wave Dislocation,* Bristol: Institute of Physics Publishing

Poston, T. & Stewart, I. 1996. *Catastrophe Theory and Its Applications*, New York: Dover

Saunders, P.T. 1980. *An Introduction to Catastrophe Theory*, Cambridge and New York: Cambridge University Press

Thom, R. 1972. *Structural Stability and Morphogenesis.* Reading, MA: Benjamin

Zeeman, E.C. 1976. Catastrophe theory: a reply to Thom. In *Dynamical Systems Warwick 1974*, edited by A. Manning, Berlin and New York: Springer, pp. 405–419

Zeeman, E.C. 1972. A catastrophe machine. In *Towards a Theoretical Biology*, vol. 4, edited by C.H. Waddington, Edinburgh: Edinburgh University Press, pp. 276–282

CAUSALITY

Basic to science as well to common sense is the root notion of causality, that things and processes in the world we experience are not totally random, but ordered in specific ways that allow for rational understanding through explanations of various types. In the transition from a mythological worldview to a rational one, the notion of guilt (in Greek *aitia*), as in a criminal being guilty of a crime, was metaphorically used to describe nonpersonal natural processes whenever one phenomenon would necessarily follow another. As one aim of modern science is to uncover the deep structure of the world beyond our immediate experience, its explanations deal with the different determinants (or causes) of processual order. Three classical conceptions of inquiry, associated with the traditions of Plato, Aristotle, and Archimedes, have provided influential ideas about the role of causal explanations in science.

In the Platonic tradition, certain properties of nature could be derived from a priori given mathematical structures. As discovered by the Pythagorean philosophers of nature, a specific mathematical structure (relations between small whole numbers) could be the key to a part of nature (such as acoustics), so why not see whether that same structure could also describe other areas (such as astronomy)? Although the latter attempt failed, the general idea of using the power of demonstrative or formal necessity in mathematics as a descriptive tool of natural causality is still vital in many areas of science, including cosmology and high-energy physics. Also, the use of analog mathematical structures to describe phenomena has become standard.

The Aristotelian tradition did not refuse mathematical description but saw it only as a tool in a search for the real causes of things. For Aristotle, there were four kinds of causes: the material cause (*hyle* or *causa materialis*) that describes the stuff of which something is made, the formal cause (*eidos* or *causa formalis*) that describes the organization of something, the efficient cause (*to kineti'kon* or *causa efficiens*) that describes the active forces by which the phenomenon comes into being, and the final cause (*to 'telos* or *causa finalis*) that describes the purpose that it serves. Thus, for a house, bricks and mortar are its material cause, the plan of the house is its formal cause, the mason building it is its efficient cause, and the purpose of sheltering a family is its final cause.

In that ancient world of Aristotle, each phenomenon generally served a purpose. Aristotle did not consider the four causes as necessarily separate aspects of nature, but more like principles of explanation that may sometimes merge, as in the sprouting acorn becoming an oak tree where the formal, efficient, and final causes work together to actualize the characteristics of an adult oak. The popular renaissance critique of the final cause as implying the paradox of a future state (a goal) influencing a present state led to a dismissal of any pluralist conception of causes. In the subsequent mechanical world picture, only efficient causes were left as explanatory. The life sciences could not live up to this reduction but continued as a descriptive natural history with an essentially Aristotelian outlook, at least until Darwin would explain the final cause of adaptations by the efficient causes of natural selection and heredity. Yet, even the Darwinian paradigm could not account for the nonlinear mechanisms of self-organization in the organism's embryonic development. Such goal-like (teleological) properties of development and self-reproduction remained necessary yet unexplained preconditions for the mechanics of natural selection.

The Archimedean tradition was founded by disciplines more physical than mathematical, although combining the two, such as optics, astronomy, mechanics, and music theory. The mathematical relations discovered by Archimedes (ca. 287–212 BC) in his books on mechanics were not a priori, as in the Platonic tradition, but derived from experience. However, the Aristotelian pursuit after the causes of the phenomena, especially the final ones, was regarded as metaphysical and so ignored. The Archimedean tradition includes such names as Ptolemaios, Johannes Kepler, Galileo Galilei, and Isaac Newton. Kepler started out as a Platonic, aiming to explain the Copernican system (which placed the Sun at the center of the solar system, in opposition to the Ptolemaic system) by regular polyhedrons, but failed and found the right laws for planetary movements through a mathematical analysis of Tycho Brahe's empirical observations. Galileo found his laws of falling bodies by eschewing the search for a hypothetical cause of gravitational force and instead using measures proportional to the velocity of a moving body for the effect of this force.

Although the mechanical worldview emphasizes only the role of efficient causes as principles of explanation in physics, the very idea of cause gave way for a long period to skepticism about proving any real existence of causes (the positivism of David Hume), eventually seeing the concept of cause as a feature of the observing subject (the transcendental idealism of Immanuel Kant). Yet, in physics, the laws of nature as expressed in terms of mathematics came to play the explanatory role of the causes of a system's movement. It was assumed that any natural system could be encoded into some formalism (e.g., a set of differential equations representing the basic laws governing the system) and that the entailment structure of that formalism perfectly mirrored the (efficient) causal structure of that part of nature. This view was compatible with a micro-determinism where a system's macroscopic properties and processes are seen as completely determined by the behavior of the system's constituent particles, governed by deterministic laws. This view was deeply questioned by quantum physics, and by Rosen's work on fundamental limits on dynamic models of causal systems.

The complexity of causality, especially in goal-directed systems, was presaged by cybernetic research in the 1940s, dealing with negative feedback control (in animals and artifacts such as self-guiding missiles) and the role of information processing for the regulation of dynamic systems. A paradigmatic example is the closed causal loops connecting various physiological levels of hormones in the body, essential for maintaining a constant internal environment (homeostasis)—a modern version of the ancient symbol of uroboros, the snake biting its own tail.

The emergence of nonlinear science in the late 20th century increased interest in the old idea that causal explanations may not all reduce to simple one-to-one correspondences between cause and effect. The realization that complex systems may occupy different areas in phase space characterized by qualitatively distinct attractors, eventually separated by fractal borders, has questioned micro-determinism even more than the fact that many such nonlinear systems have a high sensitivity to the initial conditions (the butterfly effect).

Another insight is that complex things often self-organize as high-level patterns via processes of local interactions between simple entities. This emergence of wholes (or collective behavior of units) may be mimicked in causal explanations. Instead of top-down reductive explanations, nonlinear science provides additional bottom-up explanations of emergent phenomena. Although these explanations are still reductive (in the methodological sense that one can show exactly what is going on from step to step in a simulation of the system), the complexity makes prediction impossible; thus, computational shortcuts to predict a future state can rarely be found.

As an emergent whole is formed bottom-up, its organization constrains its components in a top-down manner, that has been called downward causation (DC). There are three interpretations of DC: in strong DC, the emergent whole (a human mind) effectuates a change in the very laws that govern the lower-level (like free will might suspend what normally determines the action of the brain's neurons). This interpretation is often related to vitalist and dualist conceptions of life and mind and is hard to reconcile with science. In medium DC, lower-level laws remain unaffected; yet, their boundary conditions are constrained by the emergent pattern (a mental representation), which is considered just as real as the components of the system (neuronal signaling). Here, the state of the higher level works as a factor selecting which of the many possible next states of the high level may emerge from the low level. In weak DC, the emergent higher levels are seen as regulated by stable (cyclic or chaotic) attractors for the dynamics of the lower level. The fact that a biological species consists of stable organisms is not solely a product of natural selection, but is a result of such internal, formal properties in the system's organization—the job of natural selection being to sort out the possible stable organisms and find those most fit for the given milieu (Kauffman, 1993; Goodwin, 1994). It should be emphasized that DC is not a form of efficient causation (involving a temporal sequence from cause to effect), rather it is a modern version of the Aristotelian formal and final cause.

Nonlinear science may be said to integrate a Platonic appreciation of universality (as found in the equations governing the passage to chaos in systems of quite distinct material nature), an Aristotelian acceptance of several types of causes, and an Archimedean pragmatism regarding the deeper status of determinism and causality. The latter is reflected in the fact that although deterministic chaos characterizes a large class of systems, this does not imply that these systems (or nature) are fully deterministic. The determinism refers to the mathematical tools used rather than an ontological notion of causality.

CLAUS EMMECHE

See also **Biological evolution; Butterfly effect; Determinism; Feedback**

Further Reading

Depew, D.J. & Weber, B.H. 1995. *Darwinism Evolving: System Dynamics and the Genealogy of Natural Selection*, Cambridge, MA: MIT Press

Emmeche, C., Stjernfelt, F. & Køppe, S. 2000. Levels, Emergence, and three versions of downward causation. In *Downward Causation. Minds, Bodies and Matter*, edited by P.B. Andersen, C. Emmeche, N.O. Finnemann & P.V. Christiansen, aarhus: Aarhus University Press, pp. 13–34

Fox, R.F. 1982. *Biological Energy Transduction: The Uroboros*, New York: Wiley

Goodwin, B. 1994. *How the Leopard Changed Its Spots: The Evolution of Complexity*, New York: Scribner's

Kauffman, S.A. 1993. *The Origins of Order. Self-organization and Selection in Evolution*, Oxford and New York: Oxford University Press

Pedersen, O. 1993. *Early Physics and Astronomy: A Historical Introduction*, Cambridge and New York: Cambridge University Press

Rosen, R. 2000. *Essays on Life Itself*, New York: Columbia University Press

Weinert, F. (editor). 1995. *Laws of Nature: Essays on the Philosophical, Scientific and Historical Dimensions*, Berlin and New York: Walter de Gruyter

CAUSTICS

See **Catastrophe theory**

CAVITY SOLITONS

See **Solitons, types of**

CELESTIAL MECHANICS

Although its origins can be traced back in antiquity to the first attempts of explaining the apparently irregular wandering of the planets, celestial mechanics was born in 1687 with the release of Isaac Newton's *Principia*. In 1799, Pierre-Simon Laplace introduced the term *mécanique céleste* (Laplace, 1799), which was adopted to describe the branch of astronomy that studies the motion of celestial bodies under the influence of gravity. Celestial mechanics is researched and developed by astronomers and mathematicians; the methods used to investigate it including numerical analysis, the theory of dynamical systems, perturbation theory, the quantitative and qualitative theory of differential equations, topology, the theory of probabilities, differential and algebraic geometry, and combinatorics.

Ptolemy's idea of the epicycles—according to which planets are orbiting on small circles, whose centers move on larger circles, whose centers move on even larger circles around the Earth—dominated astronomy in antiquity and the Middle Ages. In 1543, after working for more than 30 years on a new theory, Copernicus finished writing *De Revolutionibus*, a book in which he expressed the motion of the planets with respect to a heliocentric reference system, that is, one with the Sun at its origin. This allowed Kepler to use existing observations and formulate three laws of planetary motion, published in 1609 in *Astronomia Nova*:

(i) *The law of motion*: every planet moves on an ellipse having the sun at one of its foci.
(ii) *The law of areas*: every planet moves such that the segment planet-sun sweeps equal areas in equal intervals of time.
(iii) *The harmonic law*: the squares of the periods of any two planets are to each other as the cubes of their mean distances from the sun.

But all these achievements were empirical, based on observations, not on deductions obtained from a more general physical law.

In 1666, Newton came up with the idea that the attractive force responsible for the free fall of objects might be the same as the one keeping the Moon in its orbit. He conjectured that the expression of this force is directly proportional to the product of the masses and inversely proportional to the square of the distance between bodies. The tools of calculus, which he had invented independent of—and at about the same time as—Gottfried Wilhelm von Leibniz, allowed him to proceed with the computations. Two decades later, in *Principia*, Newton proved the correctness of his theory. Kepler's laws follow as consequences. They are obtained from the differential equations of the Newtonian two-body problem (also called the Kepler problem) given by a potential energy of the form $U(r) = -Gm_1m_2/r$, where G is the gravitational constant and r is the distance between the bodies of masses m_1 and m_2.

After Newton, mathematicians, such as Johann Bernoulli, Alexis Clairaut, Leonhard Euler, such as Jean d'Alembert, Laplace, Joseph-Louis Lagrange, Siméon Poisson, Carl Jacobi, Karl Weierstrass, and Spiru Haretu, attacked various theoretical questions of celestial mechanics (e.g., the 2- and 3-body problem, the lunar problem, the motion of Jupiter's satellites, and the stability of the solar system) mostly with the quantitative tools of analysis, algebra, and the theory of differential equations. On the practical side, the first resounding success in the field was the prediction of the return of Halley's comet, which occurred in 1758—as the calculations had shown. An even more spectacular achievement came in 1846 with the discovery of the planet Neptune on the basis of the perturbation theory through computations independently performed by John Couch Adams and Urbain Jean-Joseph Le Verrier. Having its origin in one of Euler's papers, which applied the calculus of trigonometric functions to the 3-body problem, perturbation theory is now an independent branch of mathematics (see, e.g., Verhulst, 1990; Guckenheimer & Holmes, 1983) that is often used in celestial mechanics.

An important theoretical advance was achieved by Henri Poincaré toward the end of the 19th century, when the questions of celestial mechanics—especially those concerning the Newtonian 3-body problem—received substantial attention. While working on this problem, Poincaré understood that the quantitative methods of obtaining explicit solutions for differential equations are not strong enough to help him make significant progress; thus, he tried to describe the qualitative behavior of orbits (e.g., stability, the motion in the neighborhood of collisions and at infinity, existence of periodic solutions) even when their expressions were too complicated or impossible to derive, which is the case in general. His ideas led to the birth of several branches of mathematics, including the theory of dynamical systems, nonlinear analysis, chaos, stability, and algebraic topology (Barrow-Green, 1997; Diacu & Holmes, 1996).

Today's astronomers working in celestial mechanics are primarily interested in questions directly related to the solar system, such as the accurate prediction of eclipses, orbits of comets and asteroids, the motion of Jovian moons, Saturn's rings, and artificial satellites. The invention of the electronic computer had a significant impact on the practical aspects of the field. The development of numerical methods allowed researchers to obtain good approximations of the planet's motion for long intervals of time. These types of results are also used in astronautics. No space mission, from the Sputnik, Apollo, and Pioneer programs to the space shuttle, the Hubble telescope launch, and the recent international space collaboration projects, could have been possible without the contributions of celestial mechanics.

Contemporary mathematicians active in the field are mostly dealing with theoretical issues, as, for example, the study of the general N-body problem and its particular cases (Wintner, 1947) ($N = 2, 3, 4$, the collinear, isosceles, rhomboidal, Sitnikov, and planetary problems, central configurations, etc.), attempting to answer questions regarding motion near singularities and at infinity, periodic orbits, stability and chaos, oscillatory behavior, Arnol'd diffusion, etc. Some researchers also study alternative gravitational forces like that suggested by Manev (Diacu et al., 2000; Hagihara, 1975; Moulton, 1970), which offers a good relativistic approximation at the level of the solar system.

Celestial mechanics and mathematics have always influenced each other's development, a trend that is far from slowing down today. The contemporary needs of space science bring a new wave of interest in the theoretical and practical aspects of celestial mechanics, making its connections with mathematics stronger than ever before.

FLORIN DIACU

See also ***N*-body problem; Perturbation theory; Solar system**

Further Reading

Barrow-Green, J. 1997. *Poincaré and the Three-Body Problem*, Providence, RI: American Mathematical Society

Diacu, F. & Holmes, P. 1996. *Celestial Encounters—The Origins of Chaos and Stability*, Princeton, NJ: Princeton University Press

Diacu, F., Mioc, V. & Stoica, C. 2000. Phase-space structure and regularization of Manev-type problems, *Nonlinear Analysis*, 41: 1029–1055

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Berlin and New York: Springer

Hagihara, Y. 1975. *Celestial Mechanics*, vol. 2, part 1, Cambridge, MA: MIT Press

Laplace, P.-S. 1799. *Traité de mécanique céleste*. 5 vols. Paris, 1799–1825

Moulton, J.R. 1970. *An Introduction to Celestial Mechanics*, Dover

Verhulst, F. 1990. *Nonlinear Differential Equations and Dynamical Systems*, Berlin and New York: Springer

Wintner, A. 1947. *The Analytical Foundations of Celestial Mechanics*, Princeton, NJ: Princeton University Press

CELL ASSEMBLIES

The term *cell assembly* became well established as a neuropsychological concept with the publication in 1949 of Donald Hebb's book *The Organization of Behavior*. A cell assembly forms when a group of cortical neurons gets wired together through experience-dependent synaptic potentiation induced by nearly synchronous activity in pairs of connected neurons. Once formed, it serves as a cooperative and holistic mental representation. The mutual excitation within the active cell assembly influences network dynamics such that (i) the entire group of cells can be activated by stimulating only a part of it and (ii) once active, the ensemble displays afteractivity outlasting the stimulus triggering it. An important extension of the concept was suggested by Peter Milner, who proposed that negative feedback from lateral inhibition is essential to prevent mass activation of the entire network and the emergence of epileptic-like activity (Milner, 1957). This also introduces an element of competitive interaction between the cell assemblies allowing one assembly at a time to dominate the network. Hebb further proposed that cell assemblies activated in succession would wire together to form "phase sequences" that might be the physiological substrate of chains of association and the flow of thought. The concept of cell assemblies has been extensively elaborated in the context of cortical associative memory (see, e.g., Fuster, 1995).

Notably, abstract attractor network models (Hopfield, 1982; Rolls and Treves, 1997) may be regarded as mathematical instantiations of Hebb's cell assembly theory. They have been useful, for example, to estimate how many assemblies can be stored in a network of a given size, the existence of spurious attractors, and the effect of sparse activity and diluted network connectivity.

The extensive recurrent neuronal circuitry required for the formation of cell assemblies is abundant in the cerebral cortex in the form of horizontal intracortical and cortico-cortical connections. The latter are myelinated and support fast communication over large distances. The existence of experience-dependent synaptic changes in the form of "Hebbian" long-term synaptic potentiation is very well experimentally established today. In light of what we now know about brain circuitry and neuronal response properties, it is reasonable to assume that a cell assembly may extend across large areas of sensory, motor, and association cortices, perhaps even involving subcortical structures like the thalamus and basal ganglia. Rather than individual neurons, the actual nodes of the cell assembly are likely to be cortical modules, that is, minicolumns, comprising a couple of hundred neurons and with a diameter of about 30 μm (Mountcastle, 1998). Feature detectors in primary sensory areas like the visual orientation columns described by David Hubel and Torsten Wiesel are typical examples. Although the modular organization of higher-order cortical areas is less obvious, it has been proposed that the cortical sheet actually comprises a mosaic of such minicolumns. A single-cell assembly would then engage a small fraction of them distributed over a large part of the brain.

The reverberatory afteractivity in an active cell assembly was proposed by Hebb (1949) to correspond to "a single content in perception" and last some hundred milliseconds to half a second. Notably, this is of the same order as the duration of visual fixations and the period of cognitively relevant neurodynamical phenomena evident in EEG (e.g., the theta rhythm) and evoked potentials (Pulvermueller et al., 1994). Persistent activity over seconds but otherwise of the same origin has recently been proposed as a mechanism underlying working memory (Compte et al., 2000).

The cell assembly theory connects the cellular and synaptic levels of the brain with psychological phenomenology. It suggests explanations for the close interaction between memory and perception, including the holistic and reconstructive nature of Gestalt perception, perceptual fusion of polymodal stimuli, and perceptual illusions. Typical examples are perceptual completion and filling in as when looking at the Kaniza triangle and perceptual rivalry as demonstrated by the slowly alternating perception of an ambiguous three-dimensional stimulus like the Necker cube (Figure 1). An analogy to Gestalt perception in the motor domain would be motor synergies, and their existence can also be understood based on the cell assembly theory. In the motor control domain, however, the

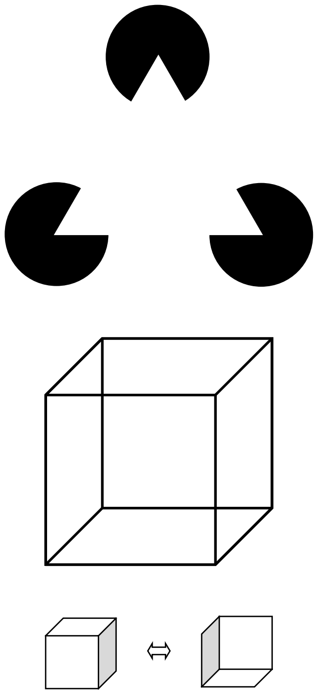

Figure 1. Two perceptual illusions that may be understood in terms of cell assembly dynamics.

temporal component of the underlying neurodynamics is critical, and for instance, finely tuned temporal sensory-motor coordination cannot be explained within this paradigm alone.

The neurodynamics of cell assemblies has been studied in a network with biophysically detailed model neurons (see, e.g., Lansner and Fransén, 1992; Fransén and Lansner, 1998), showing that the cellular properties of cortical pyramidal cells promote sustained activity. Moreover, the experimentally observed level of mutual excitation is sufficient to support such activity, and the measured magnitude of cortical lateral inhibition is effective in preventing coactivation of assemblies. The time to activate an entire assembly from a part of it was found to be within 50–100 ms in accordance with psychological experimental results. Modeling has further shown that neuronal adaptation due to accumulation of slow afterhyperpolarization (presumably together with synaptic depression) is a likely cause of termination of activity in an active assembly. Afteractivity may typically last some 300 ms, but the network dynamics is quite sensitive to endogenous monoamines such as serotonin, which acts by modulating the neuronal conductances underlying such adaptation. Further, simulations demonstrate that a cell assembly can survive even if the average conduction delay between the participating neurons increases to 10 ms, corresponding to a spatial extent of about 50 mm at axonal conduction velocities of 5 m/s. Despite quite powerful mutual excitation, reasonably low firing rates of cortical cells can be obtained in models with saturating synapses with slow kinetics (e.g., NMDA receptor gated channels) together with cortical feedback inhibition.

Although biologically highly plausible and supported by computational models, solid experimental evidence for the existence of cell assemblies is still lacking. Detection of their transient and highly distributed and diluted activity requires simultaneous noninvasive measurement in awake animals of the activity in a large number of neurons with high spatial and temporal resolution. This is still beyond the reach of current experimental techniques. Nevertheless, Hebb's original proposal has remained a vital hypothesis for more than half a century and it continues to inspire much experimental and computational research aimed at understanding how the brain works.

ANDERS LANSNER

See also **Attractor neural network; Electroencephalogram at large scales; Gestalt phenomena; Neural network models; Neurons**

Further Reading

Compte, A., Brunel, N., Goldman-Rakic, P.S. & Wang, X.-J. 2000. Synaptic mechanisms and network dynamics underlying visuospatial working memory in a cortical network model. *Cerebral Cortex*, 10: 910–923

Fransén, E. & Lansner, A. 1998. A model of cortical associative memory based on a horizontal network of connected columns. *Network: Computation in Neural Systems*, 9: 235–264

Fuster, J.M. 1995. *Memory in the Cerebreal Cortex*. Cambridge, MA: MIT Press

Hebb, D.O. 1949. *The Organization of Behavior*, New York: Wiley

Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational properties, *Proceedings of the National Academy of Sciences, USA*, 81: 3088–3092

Lansner, A. & Fransén, E. 1992. Modeling Hebbian cell assemblies comprised of cortical neurons. *Network: Computation in Neural Systems*, 3: 105–119

Milner, P.M. 1957. The cell assembly: Mark II. *Psychological Review*, 64: 242–252

Mountcastle, V.B. 1998. *Perceptual Neuroscience: The Cerebral Cortex*. Cambridge, MA: Harvard University Press

Pulvermueller, F., Preissl, H., Eulitz, C., Pantev, C., Lutzenberger, W., Elbert, T. & Birbaumer, N. 1994. Brain rhythms, cell assemblies and cognition: evidence from the processing of words and pseudowords. *Psycoloquy*, 5(48)

Rolls, E. & Treves, A. 1997. *Neural Networks and Brain Function*. Oxford and New York: Oxford University Press

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*. Berlin and New York: Springer

CELLULAR AUTOMATA

Following a suggestion by Stanislaw Ulam, John von Neumann developed the concept of cellular automata (CA) in 1948. Von Neuman wanted to formalize a set of primitive logical operations that were sufficient to evolve the complex forms of organization necessary for life. In doing this, he constructed a two-dimensional self-replicating automaton and initiated not only the study of CA but also the idea, now popular among students of complexity theory, that highly complex global behavior can be generated from local interaction rules. Much of the interest in CA has been motivated by their ability to generate complex spatial and temporal patterns from simple rules. Because of this, they provide a rich modeling environment, having the twin virtues of mathematical tractability and representational robustness (see Burk, 1970, for discussion of much of the early work on CA).

The non-obvious connection between simple local rules of interaction and emergent complex global patterns offers a possible approach to an explanation of complexity through determination of its generating local interactions. Some researchers have gone so far as to suggest that the universe itself is a CA, or CA-like object, and that sets of local generative interaction rules can replace differential equations as the standard mathematical expression for physical models.

CA are spatially and temporally discrete symbolic dynamical systems defined in terms of a lattice of sites (or cells), an alphabet of symbols that can be assigned to lattice sites, and a local update rule that determines what symbol is assigned to each site at time $t+1$ on the basis of the site values in a local neighborhood of that site at time t. Given the local neighborhood structure, a neighborhood state is defined by an assignment of symbols from the alphabet to each site in the neighborhood. The local update rule can be specified in terms of a look-up table that lists the set of neighborhood states together with the symbol that each one assigns to the designated site at the next time step.

For example, if the lattice is isomorphic to the integers, the alphabet is 0, 1, and the neighborhood of any site s is the set of sites $s-1, s, s+1$, then there are 256 possible update rules, defined by the look-up table:

000	001	010	011	100	101	110	111
x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7

(The updated symbol is entered in the central cell. This defines what are called nearest-neighbor rules.) Here, each of the x_i is either 0 or 1 depending on the specific rule. Thus, they can be thought of as components of the rule.

A cellular automaton can be defined in any number of dimensions. Von Neuman's original automaton was two dimensional, as is what is perhaps the best-known CA—Conway's Game of Life (Gardner, 1970). This is one of the simplest two-dimensional CA known to be equivalent to a universal Turing machine.

If Σ is the full state space for a CA, then the local update rule defines a global mapping $\psi : \Sigma \rightarrow \Sigma$. Much of the analytical work on CA has been directed at determining the mathematical properties of the map ψ. A fundamental paper published by Hedlund in 1969 showed that CA are just the shift-commuting endomorphisms of the shift dynamical system. It is also known that surjectivity (a function is surjective, or onto, if every state has at least one predecessor) of the map ψ is decidable only in dimension one and that for one-dimensional additive rules (i.e., those satisfying the condition $\psi(\mu+\mu') = \psi(\mu) + (\mu')$), injectivity is equivalent to a certain rule-dependent complex polynomial having no roots that are nth roots of unity for any n. (A function is injective, or one to one, if every state has at most one predecessor. A function that is both surjective and injective is called reversible.)

From the early 1960s until the early 1980s, much of the work on CA was either simple applications or mathematical analysis. The terminology was not settled, and work can be found under the names *cellular structures, cellular spaces, homogeneous structures, iterative arrays, tessellation structures*, and *tessellation arrays*.

As computers became powerful enough to support the intense calculations required, however, an experimental mathematics approach became possible. In addition, solution of systems of differential equations by computer makes use of numerical combination rules on a discrete lattice, and CA are the simplest examples of such rules, adding impetus to interest in their study.

Concurrent with the appearance of powerful computers, work was initiated on the physics of computation, and the construction of reversible automata that, it was supposed, would be a discrete equivalent to the time-invariant differential equations of physics. More or less simultaneously, Stephen Wolfram began to publish a series of papers that popularized the study of elementary automata (see Wolfram, 1994), and by the mid-1980s, CA had emerged as a major field of interest among research in the field of complex systems theory.

Because of their generality as a modeling platform, CA have found wide application in many areas of science. In chemistry and physics, they have provided models of pattern formation in reaction-diffusion systems, the evolution of spiral galaxies, spin exchange systems and Ising models, fluid and chemical turbulence (especially as lattice gas automata), dendritic crystal growth, and solitons, among other applications.

Spatially recurring patterns that propagate in the space-time evolution of certain CA have been likened to particles moving in physical space-time. The interactions of these "particles" are important in attempts to use CA for computational tasks (e.g., Crutchfield & Mitchell, 1995). It has also been pointed out that these particles are analogous to the defects, or coherent structures found in pattern formation processes in condensed matter physics, and to solitons in hydrodynamics. The best-known examples of such particles are the so-called "gliders" that occur in Conway's game of life.

Numerous connections have also been shown between fractals and cellular automata. Rescaling the space-time output of a CA often generates a fractal, as, for example, the two-site rule defined by $00, 11 \rightarrow 0, 01, 10 \rightarrow 1$ generates the well-known Sierpinski gasket (Peitgen, Jürgen & Saupe, 1992).

In biology and medicine, CAs have been applied in models of heart fibrillation, developmental processes, evolution, propagation of diseases infectious, plant growth, and ecological simulations.

In computation, CAs have been used as parallel computers, sorters, prime number sieves, and tools for encryption and for image processing and pattern recognition. Some automata have the capacity for universal computation, although how to implement this capacity remains problematic.

Cellular automata have also been used to model social dynamics (Axtell & Epstein, 1996), the spread of forest fires, neural networks, and military combat situations. Extensive references to these applications and others can be found in Voorhees (1995) and Ilachinski (2001).

Work on CA has also stimulated work on other systems that generate complex patterns based on local rules. There are close connections to the fields of artificial life, random Boolean networks, genetic programming and evolutionary computation (Mitchell, 1996), and the general theory of computational mechanics.

A web search under the key word "cellular automata" will turn up literally hundreds of sites devoted to various aspects of their study. A particularly useful program for the study of CA, Boolean networks, and other discrete iterated systems is *Discrete Dynamics Lab* (Wuensche & Lesser, 1992), available for downloading from http://www.ddlab.com.

BURTON H. VOORHEES

See also **Artificial life; Chaotic dynamics; Emergence; Fractals; Game of life; Integrable cellular automata; Lattice gas methods; Neural network models; Solitons**

Further Reading

Axtell, R. & Epstein, J.M. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*, Cambridge, MA: MIT Press

Burk, A.W. (editor). 1970. *Essays on Cellular Automata*, Champaign, IL: University of Illinois Press

Crutchfield, J.P. & Mitchell, M. 1995. The evolution of emergent computation. *Proceedings of the National Academy of Sciences*, 92(10): 10,742–10,746

Doolen, G.D. (editor). 1991. *Lattice Gas Methods: Theory, Applications, and Hardware*, New York: Elsevier

Gardner, M. 1970. The fantastic combinations of John Conway's new solitaire game of life. *Scientific American*, 223: 120–123

Ilachinski, A. 2001. *Cellular Automata*, Singapore: World Scientific

Mitchell, M. 1996. *An Introduction to Genetic Algorithms*, Cambridge, MA: MIT Press

Peitgen, H.-O, Jürgens, H. & Saupe, D. 1992. *Chaos and Fractals: New Frontiers in Science*, New York: Springer

Toffoli, T. & Margolis, N. 1987. *Cellular Automata Machines: A New Environment for Modeling*, Cambridge, MA: MIT Press

Voorhees, B.H. 1995. *Computational Analysis of One-Dimensional Cellular Automata*. Singapore: World Scientific

Wolfram, S. 1994. *Cellular Automata and Complexity*, Reading, MA: Addison-Wesley

Wuensche, A. & Lesser, M. 1992. *The Global Dynamics of Cellular Automata*, Reading, MA: Addison-Wesley

CELLULAR NONLINEAR NETWORKS

The development of cellular nonlinear networks (CNN) is embedded in the history of the electronic and computer industry, which is characterized by three revolutions: cheap computing power via microprocessors (since the 1970s), cheap bandwidth (since the end of the 1980s), and cheap sensors and MEMS (micro-electro-mechanical system) arrays (since the end of the 1990s). These research and technology breakthroughs led the way for several important economic enterprises such as the PC industry of the 1980s, the Internet industry of the 1990s, and the future analog computing industry, which is growing, together with optical and nanoscale implementations on the atomic and molecular level. Analog cellular computers have been the technical response to the sensors revolution, mimicking the autonomy and physiology of sensory and processing organs.

The CNN was invented by Leon O. Chua and Lin Yang in Berkeley in 1988. The main idea behind the CNN paradigm is Chua's so-called "local activity principle," which asserts that no complex phenomena can arise in any homogeneous media without local activity. Obviously, local activity is a fundamental property in microelectronics, where, for example, vacuum tubes and, later on, transistors are locally active devices in the electronic circuits of radios, televisions, and computers. The demand for local activity in neural networks was motivated by practical technological reasons. In 1985, John Hopfield theoretically suggested a neural network, which, in principle, could overcome the failures of pattern recognition in Frank Rosenblatt's perceptron (*See* **Perceptron**). However, its globally connected architecture was highly impractical for technical realizations in VLSI (very-large-scale-integrated) circuits of microelectronics: the number of wires in a fully connected Hopfield network grows exponentially with the size of the array. A CNN only needs electrical interconnections in a prescribed sphere of influence.

In general, a CNN is a nonlinear analog circuit that processes signals in real time. It is a multicomponent system of regularly spaced identical ("cloned") units, called cells, which communicate with each other directly only through their nearest neighbors. However, the locality of direct connections also permits global information processing. Communications between nondirectly (remote) connected units are obtained on passing through other units. The idea that complex and global phenomena can emerge from local activities in a network dates back to John von Neumann's first paradigm of cellular automata (CA). In this sense, the CNN paradigm is a higher development of the CA paradigm under the new conditions of information processing and chip technology. Unlike conventional cellular automata, CNN host processors accept and generate analog signals in continuous time with real numbers as interaction values. Furthermore, the CNN paradigm allows deep insights into the dynamic complexity of computational processes. While the classification of complexity by CA was more or less inspired by empirical observations of pattern formation in computer experiments, the CNN approach delivers a mathematically precise measure of dynamic complexity. The basic idea is to understand cellular automata as a special case of CNNs that can be characterized by a precise code for attractors of nonlinear dynamical systems and by a unique complexity index.

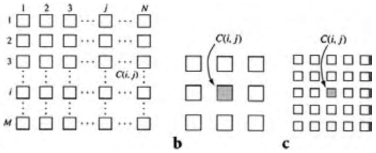

Figure 1. Standard CNN with array (a), 3×3 and 5×5 neighborhood (b,c).

Mathematical Definition

A CNN is defined by (1) a spatially discrete set of continuous nonlinear dynamical systems (cells or neurons) where information is processed into each cell via three independent variables (input, threshold, initial state) and (2) a coupling law relating relevant variables of each cell to all neighboring cells within a predescribed sphere of influence. A standard CNN architecture consists of an $M \times N$ rectangular array of cells $C(i, j)$ with cartesian coordinates (i, j) with $i = 1, 2, ..., M$ and $j = 1, 2, ..., N$ (Figure 1a). Figures 1b–c show examples of cellular spheres of influence as 3×3 and 5×5 neighborhoods. The dynamics of a cell's state is defined by a nonlinear differential equation (CNN state equation) with scalars for state x_{ij}, output y_{ij}, input u_{ij}, and threshold z_{ij}, and coefficients, called "synaptic weights", modeling the intensity of synaptic connections of the cell $C(i, j)$ with the inputs (feedforward signals) and outputs (feedback signals) of the neighboring cells $C(k, l)$. The CNN output equation connects the states of a cell with the outputs.

The majority of CNN applications use space-invariant standard CNNs with a cellular neighborhood of 3×3 cells and no variation of synaptic weights and cellular thresholds in the cellular space. A 3×3 sphere of influence at each node of the grid contains nine cells with eight neighboring cells and the cell in its center. In this case, the contributions of the output (feedback) and input (feedforward) weights can be reduced to two fixed 3×3 matrices, which are called feedback (output) cloning template **A** and feedforward (input) cloning template **B**. Thus, each CNN is uniquely defined by the two cloning templates **A**, **B**, and a threshold z, which consist of $3 \times 3 + 3 \times 3 + 1 = 19$ real numbers. They can be ordered as a string of 19 scalars with a uniform threshold, nine feedforward, and nine feedback synaptic weights. This string is called a "CNN gene" because it completely determines the dynamics of the CNN. Consequently, the universe of all CNN genes is called the "CNN genome." In analogy to the human genome project, steady progress can be made by isolating and analyzing various classes of CNN genes and their influences on CNN genomes.

Applications

In visual computing, the triple **A**, **B**, z, and its 19 real numbers can be considered as a CNN macroinstruction on how to transform an input image into an output image. Simple examples are a subclasses of CNNs with practical relevance such as the class $C(\mathbf{A}, \mathbf{B}, z)$ of space-invariant CNNs with excitatory and inhibitory synaptic weights, the zero-feedback (feedforward) class $C(0, \mathbf{B}, z)$ of CNNs without cellular feedback, the zero-input (autonomous) class $C(\mathbf{A}, 0, z)$ of CNNs without cellular input, and the uncoupled class $C(\mathbf{A}^0, \mathbf{B}, z)$ of CNNs without cellular coupling. In $\mathbf{A}^0$, all weights are zero except for the weight of the cell in the center of the matrix. Their signal flow and system structure can be illustrated in diagrams that can easily be applied to electronic circuits as well as to typical living neurons.

CNN templates are extremely useful for standards in visual computing. Simple examples are CNNs detecting edges either in binary (black-and-white) input images or in gray-scale input images. An image consists of pixels corresponding to the cells of a CNN with binary or gray scale. Logic operators can also be realized by simple CNN templates in order to combine CNN templates for visual computing. The logic NOT CNN operation inverts intensities of all binary image pixels, the foreground pixels becoming the background, and vice versa. The logic AND (logic OR, respectively) CNN operation performs a pixel-wise logic AND (logic OR operation, respectively) on corresponding elements of two binary images. These operations can be used as elements of some Boolean logic algorithms that operate in parallel on data arranged in the form of images.

The simplest form of a CNN can be characterized via Boolean functions. We consider a space-invariant binary CNN belonging to the uncoupled class $C(\mathbf{A}^0, \mathbf{B}, z)$ with a 3×3 neighborhood that maps any static 3×3 input pattern into a static binary 3×3 output pattern. It can be uniquely defined by a Boolean function of nine binary input variables, where each variable denotes one of the nine pixels within the sphere of influence of a cell. Although there are infinitely many distinct templates of the class $C(\mathbf{A}^0, \mathbf{B}, z)$, there are only a finite number of distinct combinations of 3×3 pattern of black and white cells, namely, $2^9 = 512$. As each binary nine input pattern can map to either 0 (white) or 1 (black), there are 2^{512} distinct Boolean maps of nine binary variables. Thus, every binary standard CNN can be uniquely characterized by a CNN truth table, consisting of 512 rows with one for each distinct 3×3 black-and-white pattern, nine input columns with one for each binary input variable, and one output column with binary values of the output variable.

$2^{512} \approx 1.3408 \times 10^{154} > 10^{154}$ is an "immense" number (in the sense proposed by Walter Elsasser), although the uncoupled $C(\mathbf{A}^0, \mathbf{B}, z)$ CNNs are only a small subclass of all CNNs. So, the question arises as to which subclass of Boolean functions exactly characterizes the uncoupled CNNs. In their critique of the perceptron (1969), M. Minsky and S. Papert introduced the concept of linearly separable and nonseparable Boolean functions. It can be proven that the class $C(\mathbf{A}^0, \mathbf{B}, z)$ of all uncoupled CNNs with binary inputs and binary outputs is identical to the linearly separable class of Boolean functions. Thus, linearly nonseparable Boolean functions such as, for example, the XOR function cannot be realized by an uncoupled CNN. But the uncoupled CNNs can be used as elementary building blocks that are connected by CNNs of logical operations. It can be proved that every Boolean function of nine variables can be realized by using uncoupled CNNs with nine inputs and either one logic OR CNN, or one logic AND CNN, in addition to one logic NOT CNN.

Every uncoupled CNN $C(\mathbf{A}^0, \mathbf{B}, z)$ with static binary inputs is completely stable in the sense that any solution converges to an equilibrium point. The waveform of the CNN state increases or decreases monotonically to the equilibrium point if the state at this point is positive or negative. Moreover, except for some degenerate cases, the steady-state output solution can be explicitly calculated by an algebraic formula without solving the associated nonlinear differential equations. Obviously, this is an important result to characterize a CNN class of nonlinear dynamics with robust CNN templates. Completely stable CNNs are the workhouses of the most current CNN applications. But there are also even simpler CNNs with oscillatory or chaotic behavior. Future applications will exploit the immense potentials of the unexplored terrains of oscillatory and chaotic operating regions. Then, Cellular Neural Networks will actually be transformed to Cellular Nonlinear Networks with all kinds of phase transitions and attractors of nonlinear dynamics.

Complexity Paradigm

From the perspective of nonlinear dynamics, it is convenient to think of standard CNN state equations as a set of ordinary differential equations with the components of the CNN gene as bifurcation parameters. Then, the dynamical behavior of standard CNNs can be studied in detail. Numerical examples deliver CNNs with limit cycles and chaotic attractors. The emergence of complex structures in nature can be explained by the nonlinear dynamics and attractors of complex systems. They result from the collective behavior of interacting elements in a complex system. The different paradigms of complexity research promise to explain pattern formation and pattern recognition in nature by specific mechanisms (e.g., Prigogine's chemical dissipation, Haken's work on lasers). From the CNN point of view, it is convenient to study the subclass of autonomous CNNs where the cells have no inputs. In these systems, it can be explained how patterns can arise, evolve, and sometimes converge to an equilibrium by diffusion-reaction processes. Pattern formation starts with an initial uniform pattern in an unstable equilibrium that is perturbed by small, random displacements. Thus, in the initial state, the symmetry of the unstable equilibrium is disturbed, leading to rather complex patterns. Obviously, in these applications, cellular networks do not refer only to neural activities in nerve systems, but to pattern formation in general.

A CNN is defined by the state equations of isolated cells and the cell coupling laws. For simulating reaction-diffusion processes, the coupling law describes a discrete version of diffusion (with a discrete Laplacian operator). CNN state equations and CNN coupling laws can be combined in a CNN reaction-diffusion equation, determining the dynamics of autonomous CNNs. If we replace their discrete functions and operators by their limiting continuum version, then we obtain the well-known continuous partial differential equations of reaction-diffusion processes that have been studied in different complexity approaches. Chua's version of the CNN reaction-diffusion equation delivers computer simulations of these pattern formations in chemistry and biology (e.g., concentric, auto, and spiral waves). On the other hand, for any nonlinear partial differential equation, many appropriate CNN equations can be associated with it. In many cases, it is sufficient to study the computer simulations of associated CNN equations, in order to understand the nonlinear dynamics of these complex systems.

CNN Universal Machine and Programming

There are practical and theoretical reasons for introducing a CNN Universal Machine (CNN-UM). From an engineering point of view, it is totally impractical to implement different CNN components or templates with different hardwired CNNs. Historically, John von Neumann's general-purpose computer was inspired by Alan Turing's universal machine in order to overcome all the different hardware machines of the 1930s and 1940s for different applications. From a theoretical point of view, CNN-UM opens new avenues of analog neural computers. In the CNN-UM, analog (continuous) and logic operations are mixed and embedded in an array computer. It is a complex nonlinear system, which combines two different types of operations, namely continuous nonlinear array dynamics and continuous time with local and global logic. Obviously, the mixture of analog and digital components considerably resembles to neural information processing in living organisms. The stored program, as a sequence of templates, could be considered as a genetic code for the CNN-UM. The elementary genes are the templates.

After the introduction of the architecture with standard CNN universal cells and the global analog programming unit (GAPU), the complete sequence of an analog CNN program can be executed on a CNN-UM. The description of such a program contains the global task, the flow diagram of the algorithm, the description of the algorithm in a high level α (analog) programming language, and the sequence of macro-instructions by a compiler in the form of an analog machine code (AMC). At the lowest level, the chips are embedded in their physical environment of circuits. The AMC code will be translated into hardware circuits and electrical signals. At the highest level, the α compiler generates a macro-level code called analog macro-code (AMC). The input of the α compiler is the description of the flow diagram of the algorithm using the language. In Figure 2, the levels of the software and the core engines are described. The analog macro code is used for software simulations running on a Pentium chip in a PC and for applications in a CNN-UM Chip with a CNN Chip Prototyping System (CCPS).

The CNN-UM is technically realized by analog and digital VLSI implementation. It is well known that any complex system of digital technology can be built from a few implemented building blocks by wiring and programming. In the same way, the CNN-UM, also containing analog building blocks, can be constructed. A core cell needs only three building blocks of a capacitor, resistor, and a VCCS (voltage-controlled current source). If a switch, a logic register, and a logic gate are added to the three building blocks, the extended CNN cell of the CNN-UM can be implemented. In principle, six building blocks plus wiring are sufficient to build the CNN-UM: resistor, capacitor, switch, VCCS, logic register, logic gate. As in a digital computer, stored programmability can also be introduced for analog neural computers, enabling the fabrication of visual microprocessors. Similar to classical microprocessors, stored programmability needs a complex computational infrastructure with high-level language, compiler, macro-code, interpreter, operating system, and physical code, in order to make it understandable for the human user. Using this computational infrastructure, a visual microprocessor can be programmed by downloading the programs onto the chips, as in the case of classical digital microprocessors. Writing a program for an analog CNN algorithm is as easy as writing a BASIC program.

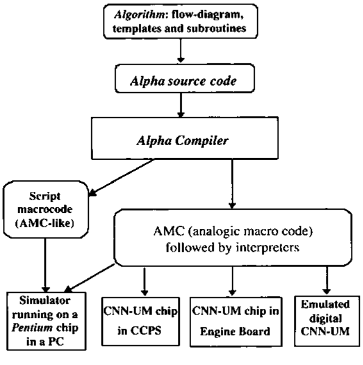
Algorithm: flow-diagram, templates and subroutines
Alpha source code
Alpha Compiler
Script macrocode (AMC-like)
AMC (analogic macro code) followed by interpreters
Simulator running on a *Pentium* chip in a PC
CNN-UM chip in CCPS
CNN-UM chip in Engine Board
Emulated digital CNN-UM

Figure 2. Levels of the software and the core engines in the CNN-UM.

With respect to computing power, CNN computers offer an orders-of-magnitude speed advantage over conventional technology when the task is complex. There are also advantages in size, complexity, and power consumption. A complete CNN-UM on a chip consists of an array of 64 × 64 0.5 μm micron CMOS cell processors. Each cell is endowed not only with a sensor for direct optical input of images and video but also with communication and control circuitries, as well as local analog and logic memories. CNN cells are interfaced with their nearest neighbors, as well as with the outside world. A CNN chip with 4096 cell processors on a chip means more than 3.0 Tera-OPS (operations per second) equivalent of computing power, which is about a 1000 times faster than the computing power of an advanced Pentium processor. By exploiting the state-of-the-art vertical packaging technologies, close to 10^{15} OPS CNN-UM architectures can be constructed on chips with 200 × 200 arrays. Thus, CNN universal chips will realize Tera-OPS or even Penta-(10^{15}) OPS, which are required for high-speed target recognition and tracking, real-time visual inspection of manufacturing processes, and intelligent machine vision capable of recognizing context-sensitive and moving scenes.

KLAUS MAINZER

See also **Attractor neural network; Cellular automata; Integrable cellular automata**

Further Reading

Chua, L.O. 1998. *A Paradigm for Complexity*, Singapore: World Scientific

Chua, L.O., Gulak, G., Pierzchala, E. & Rodriguez-Vázquez (editors). 1998. Cellular neural networks and analog VLSI. *Analog Integrated Circuits and Signal Processing. An International Journal*, 15(3)

Chua, L.O. & Roska, T. 2002. *Cellular Neural Networks and Visual Computing: Foundations and Applications*, Cambridge and New York: Cambridge University Press

Chua, L.O., Sbitnev, V.I. & Yoon, S. 2003. A nonlinear dynamics perspective of Wolfram's new kind of science. Part II: universal neuron. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 13: 2377–2491

Chua, L.O. & Yang, L. 1988. Cellular neural networks: theory. *IEEE Transactions on Circuits and Systems*, 35: 1257–1272; Cellular neural networks: applications. *IEEE Transactions on Circuits and Systems*, 35: 1273–1290

Chua, L.O., Yoon, S. & Dogaru, R. 2002. A nonlinear dynamics perspective of Wolfram's new kind of science. Part I: threshold of complexity. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 12: 2655–2766

Elsassee, W.M. 1987. *Reflections on a Theory of Organisms: Holism in Biology*, Frelighsburg, Quebec: Editions Orbis

Huertas, J.L., Che, W.K. & Madan, R.N. (editors). 1999. *Visions of Nonlinear Science in the 21st Century. Festshrift Dedicated to Leon O. Chua on the Occasion of his 60th Birthday*, Singapore: World Scientific

Mainzer, K. 2003. *Thinking in Complexity: The Computational Dynamics of Matter, Mind, and Mankind*, 4th edition, Berlin and New York: Springer

Tetzlaff, R. (editor). 2002. Cellular Neural Networks and Their Applications. *Proceedings of the 7th IEEE International CNN Workshop*, Singapore: World Scientific

CENTER MANIFOLD REDUCTION

A dynamical system might be difficult to solve, even numerically. To better understand its behavior in the neighborhood of an equilibrium, a reduction can be performed. For this, one starts with the eigenvalue spectrum λ of the linearized system. In the linear evolution $\sim \exp(\lambda t)$, eigenvalues with $\Re\lambda < 0$ ($\Re\lambda > 0$) are called stable (unstable), whereas those with $\Re\lambda = 0$ are called central.

In the neighborhood of an equilibrium point, P, of a dynamical system, in general, three different types of invariant manifolds exist: the trajectories belonging to the stable manifold M^{s} are being attracted by P, whereas those of the unstable manifold M^{u} are being repelled. The dynamics on the center manifold M^{c} depend on the nonlinearities. For the linearized problem, $E^{\mathrm{s}} \equiv M^{\mathrm{s}}$, $E^{\mathrm{u}} \equiv M^{\mathrm{u}}$, and $E^{\mathrm{c}} \equiv M^{\mathrm{c}}$ are uniquely determined linear subspaces that span the whole space. The transition to the nonlinear system causes (only) deformations of the linearly determined manifolds M_{s}, M_{u}, and M_{c}. However, the form of the latter critically depends on the nonlinear terms.

Let us elucidate that the behavior for the very simple system (Grosche et al., 1995)

$$\dot{x} = -xy, \quad \dot{y} = -y + x^2. \tag{1}$$

Here, the dot means differentiation with respect to time t. The equilibrium point P is (0, 0), and the linearized system is

$$\dot{x} = 0, \quad \dot{y} = -y. \tag{2}$$

Here, we have the stable manifold E^{s} being identical to the y-axis and the center manifold E^{c} being identical to the x-axis. The linearized problem can be visualized by the graph shown in Figure 1.

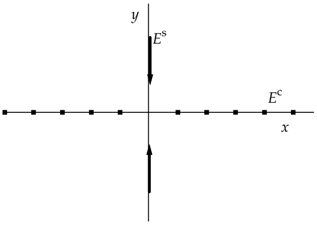

Figure 1. The linear stable (E^{s}) and center manifold (E^{c}) for example (1).

Both manifolds will be deformed in the transition to the nonlinear system. The (nonlinear, perturbed) center manifold can be described in the present example by

$$M^{\mathrm{c}} : y = h(x) \tag{3}$$

with $h(0) = h'(0) = 0$. Using that ansatz for the center manifold in (1), we obtain

$$\dot{x} = -x\, h(x). \tag{4}$$

Differentiating (3) with respect to t, leads to

$$\dot{y} = h'(x)\, \dot{x} \tag{5}$$

or

$$-h(x) + x^2 = h'(x)[-xh(x)], \tag{6}$$

that is, a differential equation for $h = h(x)$. Performing a power series ansatz $h(x) = cx^2 + dx^3 + \cdots$, we find that $c = 1$. The (nonlinear) center manifold is thus given by

$$y = x^2 + \cdots, \tag{7}$$

and the dynamics on it follows from

$$\dot{x} = -x^3; \tag{8}$$

that is, the trajectories are being attracted by P. For an illustration, see Figure 2.

More General Theoretical Background

Let us now generalize the idea and consider a system of ordinary differential equations (ODEs)

$$\dot{a} = Aa + N(a, b), \tag{9}$$

$$\dot{b} = Bb + M(a, b), \tag{10}$$

describing the dynamics of amplitudes $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ of n linear marginal stable modes and m linear stable modes, respectively [$(a_1, \ldots, a_n) := a$,

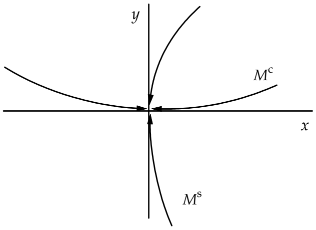

Figure 2. The stable (M^{s}) and center manifold (M^{c}) for example (1).

$(b_1, \dots, b_m) := b]$. This implies that the real parts of eigenvalues of the matrix A vanish, and the real parts of the eigenvalues of the matrix B are negative. The functions $N(a, b), M(a, b) \in C^r$ on the right-hand sides of Equations (9) and (10) represent the nonlinear terms. Let E^{c} be the n-dimensional (generalized) eigenspace of A and E^{s} be the m-dimensional (generalized) eigenspace of B. Under these assumptions, the center manifold theorem provides the following statement (Guckenheimer & Holmes, 1983):

> There exists an invariant C^r manifold M^{s} and an invariant C^{r-1} manifold M^{c} that are tangent at $(a, b) = (0, 0)$ to the eigenspaces E^s and E^{c}, respectively. The stable manifold M^{s} is unique but the center manifold M^{c} is not necessarily unique.

Locally, the center manifold M^{c} can be represented as a graph,

$$M^{\mathrm{c}} = \{(a, b) | b = h(a)\}, \; h(0) = 0, \; Dh(0) = 0, \tag{11}$$

where the C^{r-1} function h is defined in a neighborhood of the origin, and Dh denotes the Jacobi matrix. Introducing (11) in Equations (9) and (10), we obtain

$$\dot{a} = Aa + N(a, h(a)), \tag{12}$$

$$Dh(a)\,[Aa + N(a, h(a))] = Bh(a) + M(a, h(a)). \tag{13}$$

The solution h of Equation (13) can be approximated by a power series. The ambiguity of the center manifold is manifested by the fact that h is determined only modulo C^{∞}, a non-analytic function; thus, the power series approximation of the function h is unique. The importance of the center manifold theory is reflected by the following theorem (Marsden & McCracken, 1976; Carr, 1981):

> If there exists a neighborhood U^{c} of $(a, b) = (0, 0)$ on M^{c}, so that every trajectory starting in U^{c} never leaves it, then there exists a neighborhood U of $(a, b) = (0, 0)$ in $\Re^n \times \Re^m$, so that every trajectory starting in U converges to a trajectory on the center manifold.

Therefore, it is sufficient to discuss the dynamics on the center manifold, described by Equation (12). If all solutions are bounded to some neighborhood of the origin, then we have described all features of the asymptotic behavior of Equations (9) and (10). In order to fulfill the condition, the function $N(a, h(a))$ has to be expanded up to a sufficiently high order. We end up with normal forms, for example, the third order may be adequate.

Very often, the problems contain parameters and, in addition, the systems may be infinite dimensional. In both cases, one can generalize the theory presented so far. Parameters can be taken into account by expanding Equations (9) and (10) to

$$\dot{a} = A(\Lambda)a + N(a, b, \Lambda), \tag{14}$$

$$\dot{b} = B(\Lambda)b + M(a, b, \Lambda), \tag{15}$$

$$\dot{\Lambda} = 0, \tag{16}$$

where $\Lambda = (a_{n+1}, \dots, a_{n+l})$ contains l parameters. The center manifold now has dimension $n + l$.

PDE Reduction and Symmetry Considerations

The theory is also valid in the infinite-dimensional case, if the spectrum of the linear operator can be split into two parts. The first part contains a finite number of eigenvalues whose real parts are zero, and the second part contains (an infinite number of) eigenvalues with negative real parts that are bounded away from zero.

To elucidate the power of center manifold reduction, let us consider the partial differential equation (PDE)

$$\frac{\partial \phi}{\partial t} + \phi\,\frac{\partial \phi}{\partial y} + \alpha\,\frac{\partial^2 \phi}{\partial y^2} + \beta\,\frac{\partial^3 \phi}{\partial y^3} + \frac{\partial^4 \phi}{\partial y^4} + \nu\phi = 0. \tag{17}$$

All coefficients α, β, and ν are nonnegative. In the following, we treat β as a fixed parameter, and consider the dynamics in dependence in α and ν. The linearization with $\phi \equiv 0$ as the equilibrium solution leads to

$$\omega = -\beta k^3 + \mathrm{i}\left(-k^4 + \alpha k^2 - \nu\right) \tag{18}$$

when we assume a unit cell of length 2π with periodic boundary conditions. A typical dependence of the linear growth (or damping) rate $\gamma := \Im\omega$ is shown in Figure 3 for $\alpha = 5.25$ and $\nu = 3.8$.

The case of two unstable modes ($k = 1, 2$) is already highly nontrivial. Let us choose $\alpha = \alpha_{\mathrm{c}} = 5$ and $\nu = \nu_{\mathrm{c}} = 4$. Then, the modes $\phi^{(1)} = \sin y$, $\phi^{(2)} = \cos y$, $\phi^{(3)} = \sin 2y$, and $\phi^{(4)} = \cos 2y$ belonging to $k = 1$ and 2, respectively, are marginally stable. We introduce

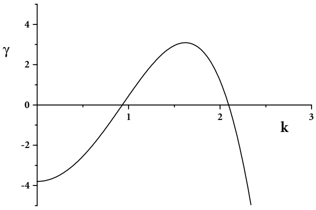

Figure 3. Growth rate curve, with two unstable modes at $k = 1$ and $k = 2$ for the PDE (17).

the four (real) amplitudes a_1, a_2, a_3, and a_4, as well as $\alpha_5 = \alpha - \alpha_c$ and $\alpha_6 = \nu - \nu_c$. The center manifold theory will allow us to derive a closed set of nonlinear amplitude equations

$$\dot{a}_n = f_n(a_1, \ldots, a_6),\ n = 1, \ldots, 6, \tag{19}$$

which are valid in the neighborhood of the critical point α_c, ν_c. One has $f_5 \equiv f_6 \equiv 0$. The other functions f_n are written as a power series in a_n,

$$f_n = \sum_{1\le m\le 6} A_n^m\, a_m + \sum_{1\le m\le p\le 6} A_n^{mp}\, a_m\, a_p + \cdots . \tag{20}$$

The dynamics on the center manifold is characterized by $a_1, \ldots, a_6$. Thus, we can make the ansatz (Carr, 1981)

$$\begin{aligned}\phi(y,t) &= \sum_{1\le n\le 4} a_n(t)\, \phi^{(n)}(y) \\ &\quad + \sum_{1\le n\le m\le 6} a_n(t)\, a_m(t) \phi^{(nm)}(y),\end{aligned} \tag{21}$$

where the $\binom{7}{2} = 21$ new functions $\phi^{(nm)}$ and, of course, the next $\binom{8}{3} = 56$ functions $\phi^{(nmp)}$, and so on, can be chosen orthogonal to $\phi^{(n)}, n = 1, \ldots, 4$.

The technical procedure is now as follows. One inserts ansatz (21) into the basic equation (17) and compares equal orders in the amplitudes. For example, in the second-order, one collects equal powers $a_r a_s$; the "coefficients" (being equated to zero) will determine the unknown functions $\phi^{(nm)}$ via ODEs. Taking into account the (periodic) boundary conditions, we have to satisfy the solvability conditions. Collecting equal powers of the amplitudes a_n, we find the solutions for the coefficients A_r^{np}. With these values, we can solve for $\phi^{(mn)}$. This procedure should be continued to higher orders.

Actually, when written explicitly, one faces considerable work (in second order, we have to solve for 84, and in third order for 224 coefficients $A_r^{\cdots}$, and so on). One can simplify the calculations by making use of symmetries.

Translational invariance implies the following. If $\phi(y)$ is a solution,

$$T_{y_0}\phi(y) := \phi(y + y_0) \tag{22}$$

will also satisfy the dynamical equation (17), where y_0 is a real shift parameter. (In the case $\beta \equiv 0$, we also have the mirror symmetry $\phi(y) = \phi(-y)$.)

Remember the structure of the center manifold reduction: the modes $\phi^{(nm)}, \ \phi^{(nmp)}, \ldots$ have to be determined from inhomogeneous differential equations. The inhomogeneities contain (in nonlinear forms) the marginal modes $\phi^{(r)}, \ r = 1, \ldots, 4$. Thus, the so-called slaved modes can be written in symbolic form as

$$\phi^{(m\ldots)} = h^{(m\ldots)}\left[\{\phi^{(r)}\}\right]. \tag{23}$$

Thus, the following symmetry should hold:

$$T_{y_0} h^{(m\ldots)}\left[\{\phi^{(r)}\}\right] = h^{(m\ldots)}\left[T_{y_0}\{\phi^{(r)}\}\right]. \tag{24}$$

The consequences of the translational symmetry are most easily seen when combing the marginal modes to

$$\varphi := \sum_{r=1}^{4} a_r \phi^{(r)} \equiv \Im\left(c_1 e^{iy} + c_2 e^{2iy}\right) \tag{25}$$

with the complex amplitudes $c_1 := a_1 + ia_2, \ \ c_2 := a_3 + ia_4$. The (complex) amplitude equations are

$$\dot{c} = g_n(c_1, c_2, a_5, a_6),\ n = 1, 2, \tag{26}$$

$$\dot{a}_m = 0,\ m = 5, 6. \tag{27}$$

The translational symmetry (22) requires

$$e^{iny_0} g_n(c_1, c_2, a_5, a_6) = g_n\left(e^{iy_0}c_1, e^{i2y_0}c_2, a_5, a_6\right) \tag{28}$$

for $n = 1, 2$. The vector field (g_1, g_2) is called equivariant with respect to the operation

$$(c_1, c_2) \to \left(e^{iy_0}c_1, e^{i2y_0}c_2\right). \tag{29}$$

The most general form of vector fields being equivariant under operation (28) is

$$(g_1, g_2) = \left(c_1 P_1 + \bar{c}_1 c_2 Q_1,\ c_2 P_2 + c_1^2 Q_2\right), \tag{30}$$

where P_1, P_2, Q_1, and Q_2 are polynomials in $|c_1|^2, |c_2|^2$, and $\Re(c_1^2\bar{c}_2)$; of course, they can also depend on a_5 and a_6.

Keeping in mind the symmetry properties, the general form of the amplitude equations reduces to

$$\dot{c}_1 = \lambda c_1 + \mathcal{A}\bar{c}_1 c_2 + \mathcal{C} c_1|c_1|^2 + \mathcal{E} c_1|c_2|^2 + \mathcal{O}(|c|^4), \tag{31}$$

$$\dot{c}_2 = \mu c_2 + \mathcal{B} c_1^2 + \mathcal{D} c_2|c_1|^2 + \mathcal{F} c_2|c_2|^2 + \mathcal{O}(|c|^4), \tag{32}$$

$$\dot{a}_5 = \dot{a}_6 = 0. \tag{33}$$

A straightforward analysis leads to

$$\begin{aligned}
&\lambda = a_5 - a_6 - \mathrm{i}\beta, \quad \mu = 4a_5 - a_6 + \mathrm{i}8\beta,\\
&\mathcal{A} = \tfrac{1}{2}, \quad \mathcal{B} = -\tfrac{1}{2},\\
&\mathcal{C} = 0, \quad \mathcal{D} = -\tfrac{3}{4(20-\mathrm{i}9\beta)},\\
&\mathcal{E} = \tfrac{1}{2}D, \quad \mathcal{F} = -\tfrac{1}{12(15-\mathrm{i}4\beta)}.
\end{aligned} \tag{34}$$

This completes the center manifold reduction.

Very interesting conclusions result, for example, with respect to the number of modes and their interplay in time, from the systematic treatment with the center manifold theory. For example, one interesting aspect is that the present codimension-two analysis can describe successive bifurcations of one unstable mode, which, in some cases can lead to chaos in time.

KARL SPATSCHEK

See also **Inertial manifolds; Invariant manifolds and sets; Synergetics**

Further Reading

Carr, J. 1981. *Applications of Center Manifold Theory*, New York: Springer

Grosche, G., Ziegler, V., Ziegler, D. & Zeidler, E. (editors). 1995. *Teubner-Taschenbuch der Mathematik, Teil II*, Stuttgart: Teubner

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Berlin and New York: Springer

Marsden, J.E. & McCracken, M. 1976. *The Hopf Bifurcation and Its Applications*, New York: Springer

CENTRAL LIMIT THEOREM

See **Martingales**

CHAOS VS. TURBULENCE

The notion of chaos has its genesis in the work of Henri Poincaré (*See* **Poincaré theorems**) on the three-body problem of celestial mechanics. Poincaré realized that this problem cannot be reduced to quadratures and solved in the manner of the two-body problem. A precise definition of chaos or non-integrability can be given in terms of the absence of conserved quantities necessary to yield a solution. It took several decades for the full significance of non-integrable dynamical systems to be appreciated and for the term "chaos" to be introduced (*See* **Chaotic dynamics**). An important step was the 1963 paper by Edward N. Lorenz, entitled "Deterministic Nonperiodic Flow" (Lorenz, 1963), on a model describing thermal convection in a layer of fluid heated from below. The Lorenz model truncates the basic fluid dynamical equations, written in terms of Fourier amplitudes, to just three modes (*See* **Lorenz equations**):

$$\begin{aligned}
\dot{x} &= -\sigma x + \sigma y,\\
\dot{y} &= -xz + rx - y\\
\dot{z} &= xy - bz.
\end{aligned} \tag{1}$$

In this system, x is the time-dependent amplitude of a stream-function mode, while y and z are mode amplitudes of the temperature field. The parameters σ, r, and b depend on the geometry, the boundary conditions, and the physical parameters of the fluid. Equations (1) are a subset of the full, infinite system of mode amplitude equations, chosen such that it exactly captures the initial instability of the thermally conducting state to convecting rolls when the parameter r, known as the Rayleigh number, is increased.

What Lorenz observed in numerical solutions of (1), and verified by analysis, was that very complicated, erratic solutions would arise when r was increased well beyond the conduction-to-convection transition. In fact, Lorenz had found the first example of what is today called a strange attractor (*See* Figure 1 and **Attractors**). System (1) is clearly deterministic, yet it can produce non-periodic solutions. There were other intriguing aspects of the solutions to (1) in the chaotic regime. Solutions arising from close initial conditions would separate exponentially in time, leading to an apparently random dependence on initial conditions of the solution after a finite time (*See* **Butterfly effect**). Today, this would be associated with the existence of a positive characteristic Lyapunov exponent. A list of "symptoms" can be established that are shared by systems having the property of chaos, including: complex temporal evolution, exponential separation from close initial conditions, a strange attractor in phase space (if the system is dissipative), and positive Lyapunov exponents. An important difference from Poincaré's work was that Lorenz's system described a dissipative system in which energy is not conserved.

From the start, the potential connection between chaos and other concepts in statistical physics, such as ergodicity and turbulence, was of central interest. For example, chaos was thought to imply ergodic

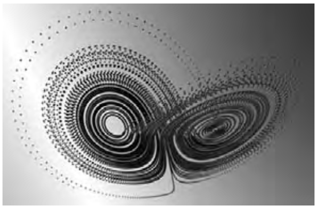

Figure 1. Strange attractor associated with the Lorenz equations. Reproduced with permission from Images by Paul Bourke, http://astronomy.swin.edu.au/ pbourke/fractals/lorenz/.

behavior in the sense of the "ergodic hypothesis" underlying equilibrium statistical mechanics (*See* **Ergodic theory**). Similarly, the connection between chaos and turbulence was sought, particularly appropriate given that Lorenz's model was of a fluid flow. Experiments on other fluid systems by Gollub, Swinney, Libchaber, and later many others established that the transition from laminar to turbulent flow typically takes place through a regime of chaotic fluid motion. The well-known route to chaos via period-doubling bifurcations of Mitchell J. Feigenbaum belongs here as well (Feigenbaum, 1980; Eckmann, 1981). In view of this, it is natural to think that turbulent flow itself is simply some kind of chaotic flow state.

Turbulence is a common state of fluid flow that shares several "symptoms" with chaotic dynamical systems, but also has distinct features not easily duplicated by chaos. The word "turbulence" was apparently first used by Leonardo da Vinci to describe a complex flow. In mathematical terms, turbulent flows should be solutions of the Navier–Stokes equation, usually written in the dimensionless form (*See* **Navier–Stokes equation**)

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} = -\nabla p + R^{-1} \Delta \boldsymbol{u}, \qquad (2)$$

$$\nabla \cdot \boldsymbol{u} = 0. \qquad (3)$$

We have restricted attention to incompressible flows by insisting in (3) that the velocity field $\boldsymbol{u}(\boldsymbol{x}, t)$ be divergence free. In (2) the field p represents the pressure—the constant density has been absorbed in the nondimensionalization. The sole dimensionless parameter R is Reynolds number. In terms of physical variables $R = UL/\nu$, where U is a typical scale of velocity, L a typical length scale of the flow, and n is the kinematic viscosity of the fluid. For small values of R, say $0 < R \leq 1$, the flow is laminar. For moderate R, say $1 < R \leq 100$, various periodic flow phenomena may arise, such as the shedding of vortices from blunt bodies. For large R, the flow eventually breaks down into many interacting eddies—this is turbulent flow. Since most flowing fluid is, in fact, flowing at large R, turbulence is the prevailing flow state of fluids in our surroundings (oceans and atmosphere), in the universe in general, in many industrial processes, and to some extent, within our bodies.

The characterization of what makes a flow turbulent is not nearly so clear as what makes a dynamical system chaotic. First, the issue of whether the particular set of nonlinear partial differential equations (2) and (3) even has a smooth solution for all time, given smooth initial conditions, is still unsettled and is one of the prize challenges set by the Clay Mathematics Institute (http://www.claymath.org). In spite of several attempts, a convincing example of a flow with smooth initial conditions, evolving under (2) and (3), that develops a singularity in a finite time has not been found. Conversely, there is no proof that solutions with the requisite number of derivatives will exist for all time.

Turbulent flows are also recognized by a variety of "symptoms." The flow velocity as a function of time at any given point in a turbulent flow is a random function (roughly a Gaussian). However, the overall nature of the velocity field viewed as a random vector field is not Gaussian. The random nature of turbulent velocity fields is today thoroughly familiar to the flying public. The randomness is not just temporal at a fixed point in space; the spatial variation of the flow field at a given time constitutes a multitude of interacting eddies of different sizes. Because of their random character, turbulent flows stir vigorously, leading to rapid dispersal of a passively advected substance or a field, such as temperature, and to a rapid exchange of momentum with contiguous fluid. In the classic pipe flow experiment of Osborne Reynolds, for example, in which the transition from laminar to turbulent flow was first demonstrated to depend only on the dimensionless number R, a streak of dye introduced at the inlet would remain a thin streak (except for a bit of molecular diffusion) when the flow in the pipe was laminar. When the flow rate was increased and the flow became turbulent, the dye rapidly dispersed across the pipe.

In a turbulent flow, the large scales of motion, which are typically in contact with some kind of forcing from the outside, will generate smaller scales through interactions and instabilities. This process continues through a broad range of length scales, ultimately reaching small scales where molecular dissipation is effective and quells the motion altogether. The repeated process of "handing down" energy from larger scales to smaller scales is a key process in turbulence. It is usually referred to as the Kolmogorov cascade (*See* **Kolmogorov cascade**). The qualitative nature of this process was already envisaged by Lewis Fry Richardson and was described by him in an adaptation of a verse by Jonathan Swift:

> *Big whorls have little whorls,*
> *Which feed on their velocity;*
> *And little whorls have lesser whorls,*
> *And so on to viscosity (in the molecular sense).*

Because of its broad range of length scales, the energy in a turbulent flow may be considered partitioned among modes of different wavenumbers k. The energy spectrum $E(k)$ is defined such that $E(k)\,\mathrm{d}k$ is the amount of kinetic energy of the turbulent flow associated with motions with wavenumbers between k and $k + \mathrm{d}k$. The cascade implies a transfer of energy from scale to scale with a characteristic energy flux per unit mass, ε, which must also be equal to the rate at which energy is fed to the flow from the largest scales, and to the rate at which energy is dissipated by viscosity at the smallest scales. A simple dimensional argument then (*See* **Dimensional**

analysis) gives the dependence of $E(k)$ on ε and k to be

$$E(k) = C\varepsilon^{2/3}k^{-5/3}. \tag{4}$$

This is the well-known Kolmogorov spectrum, predicted by Andrei N. Kolmogorov in 1941 (Hunt et al., 1991; Frisch, 1995) and only subsequently verified by experiments in a tidal channel (see Figure 2).

Turbulence has many further intriguing statistical properties, which remain subjects of active research. A major shift in our thinking on turbulence occurred in the late 1960s and in the 1970s when experiments by Kline and Brown & Roshko demonstrated that even in turbulent shear flows at very large Reynolds number, one can identify coherent structures that organize the flow to some extent (Figure 3). Later investigations have shown that even in homogeneous, isotropic turbulence, the flow is often organized into strong filamentary vortices. The persistence of these organized structures, which can dominate the flow for long times and interact dynamically, forces a strong coupling among the spectral modes, reducing the effective number of degrees of freedom of the problem.

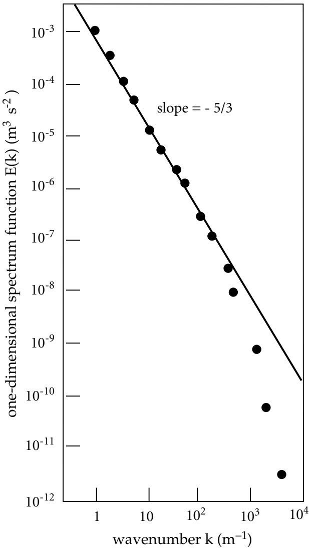

Figure 2. One-dimensional spectrum in a tidal channel from data in Grant et al. (1962).

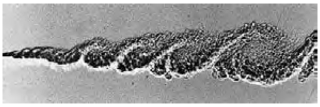

Figure 3. Coherent structures in a turbulent mixing layer. From Brown & Roshko (1974), reprinted from *An Album of Fluid Motion*, M. Van Dyke, Parabolic Press, 1982.

Chaos and turbulence both describe states of a deterministic dynamical system in which the solutions appear random. Our current understanding of chaos is largely restricted to few-degree-of-freedom systems. Turbulence, on the other hand, is a many-degree-of-freedom phenomenon. It seems somewhat unique to fluid flows—related phenomena such as plasma turbulence or wave turbulence appear to be intrinsically different. The emergence of collective modes in the form of coherent structures in turbulence amidst the randomness is an intriguing feature, somewhat reminiscent of the mix between regular "islands" and the "chaotic sea" observed in chaotic, low-dimensional dynamical systems. The coherent structures themselves approximately form a deterministic, low-dimensional dynamical system. However, it seems impossible to fully eliminate all but a finite number of degrees of freedom in a turbulent flow—the modes not included explicitly form an essential, dissipative background, often referred to as an eddy viscosity, that must be included in the description.

Turbulence is intrinsically spatiotemporal, whereas chaotic behavior in a fluid system can be merely temporal with a simple spatial structure. It is possible for the flow field to be perfectly regular in space and time, yet the trajectories of fluid particles moving within the flow will be chaotic. This is the phenomenon of chaotic advection (*See* **Choatic advection**), which points out the hugely increased complexity of a turbulent flow relative to chaos in a dynamical system.

Paul K.A.Newton and Hassan Aref

See also **Attractors; Butterfly effect; Celestial mechanics; Chaotic advection; Chaotic dynamics; Diffusion; Ergodic theory; Kolmogorov cascade; Lorenz equations; Lyapunov exponents; Navier–Stokes equation; *N*-body problem; Partial differential equations, nonlinear; Period doubling; Phase space; Poincaré theorems; Routes to chaos; Shear flow; Thermal convection; Turbulence**

Further Reading

Aref, H. & Gollub, J.P. 1996. Application of dynamical systems theory to fluid mechanics. *Research Trends in Fluid Dynamics, Report of the US National Committee on Theoretical and Applied Mechanics*, edited by J.L. Lumley et al., New York: AIP Press, pp. 15–30

Eckmann, J.P. 1981. Roads to turbulence in dissipative dynamical systems. *Reviews of Modern Physics*, 53: 643–654

Feigenbaum, M.J. 1980. Transition to aperiodic behavior in turbulent systems. *Communications in Mathematical Physics*, 77: 65–86

Frisch, U. 1995. *Turbulence—The Legacy of A. N. Kolmogorov*, Cambridge and New York: Cambridge University Press

Grant, H.L., Stewart, R.W. & Moilliet, A. 1962. Turbulent spectra from a tidal channel. *J. Fluid. Mech.*, 12: 241–268

Hunt, J.C.R., Phillips, O.M., & Williams, D. (editors). 1991. Turbulence and stochastic processes: Kolmogorov's ideas 50 years on. *Proceedings of the Royal Society, London* A, 434: 1–240

Lorenz, E.N. 1963. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20: 130–141

Ruelle, D. 1991. *Chance and Chaos*, Princeton, NJ: Princeton University Press

CHAOTIC ADVECTION

In fluid mechanics, *advection* means the transport of material particles by a fluid flow, as when smoke from a chimney is blown by the wind. The term *passive advection* is sometimes used to emphasize that the substance being carried by the flow is sufficiently inert that it follows the flow entirely, the velocity of the advected substance at every point and every instant adjusting to that of the prevailing flow.

To describe the kinematics of a fluid, two points of view may be adopted: the Eulerian representation focuses on the velocity field **u** as a function of position and time, $\mathbf{u}(\mathbf{x}, t)$; the Lagrangian representation emphasizes the trajectories $\mathbf{x}_P(t)$ of a fluid particle as it is advected by the flow. The two points of view are linked by stating that the value of the velocity field at a given point in space and instant in time equals the velocity of the fluid element passing through that same point at that instant, that is,

$$\dot{\mathbf{x}}_P(t) = \mathbf{u}(\mathbf{x}_P(t), t). \tag{1}$$

The Eulerian representation is used extensively for measurements and numerical simulations of fluid flow since it allows one to fix the points in space and time where the field is to be determined. The Lagrangian representation, on the other hand, is often more natural for theoretical analysis, as it explicitly addresses the nonlinearity of the Navier–Stokes equation.

For a given flow, the equations of motion (1), sometimes called the *advection equations*, are a system of ordinary differential equations that define a dynamical system. These equations can be integrable or non-integrable. Chaotic advection appears when the equations are non-integrable and the trajectories of fluid elements become chaotic. The dynamical system defined by (1) has two or more degrees of freedom. For a two-dimensional time-independent or steady flow, there are just two degrees of freedom and no chaotic motion is possible. However, already for a 2-d time-dependent or a 3-d steady flow, there are enough degrees of freedom to allow for chaotic trajectories. In other words, chaotic advection can appear even for flows that would otherwise be considered laminar.

The phenomenon of chaotic advection is also known as Lagrangian chaos, or sometimes Lagrangian turbulence. Usually, the word turbulence refers to the Eulerian representation and to flows in which the velocity field fluctuates across a wide range of spatial and temporal scales with limited correlations. In such flows, the trajectories of fluid elements are always chaotic. By contrast, chaotic advection or Lagrangian chaos can arise in situations where the velocity field is spatially coherent and the time dependence is no more complicated than a simple periodic modulation.

Many examples have now been given to illustrate the point that the complexity of the spatial structure of material advected by a flow can be much greater than one might surmise from a picture of the instantaneous streamlines of the flow. Thus, in the paper that introduced the notion of chaotic advection (Aref (1984) and Figure 1), the case of two stirrers that act alternately on fluid confined to a disk was considered. Each stirrer was modeled as a point vortex that could be switched on and off. There are several parameters in the system, such as the strengths and positions of the vortex stirrer and the time interval over which each acts. For a wide range of parameter values, the dynamics is as shown in Figure 1; after just a few periods, the 10,000 particles being advected are spread out over a large fraction of the disk.

Chaotic advection gives rise to very efficient stirring of a fluid. Material lines are stretched at a rate given by the Lyapunov exponent. In bounded flows, these exponentially growing material lines have to be folded back over and over again, giving rise to ever finer and denser striations. They are familiar from the mixing of paint or from marbelized paper. On the smallest scales diffusion, takes over and smoothes the steep gradients, giving rise to mixing on the molecular scale. The interplay between stirring and diffusion is the

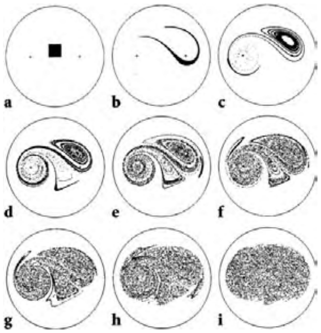

Figure 1. Spreading of 10,000 particles in a cylindrical container (disk) under the alternating action of two stirrers. The positions of the stirrers are marked by crosses. (a) initial distribution; (b)–(g) positions of the particles after 1, 2, …, 6 periods; (h) after 9 periods; (i) after 12 periods. From (Aref, 1984).

source of the efficient mixing in the presence of chaotic advection. This phenomenon is being exploited in various procedures for mixing highly viscous fluids, including applications to materials processing, in micro-fluidics, and even in large-scale atmospheric, oceanographic, and geological flows. It may play a role in the feeding of microorganisms.

In the case of 2-d incompressible flows the equations of motion allow for an interesting connection to Hamiltonian dynamics. The velocity field can be represented through a stream function $\psi(x, y, t)$, so that $\mathbf{u} = \nabla \times \psi \mathbf{e}_z$ and Equations (1) for the trajectories of fluid elements become

$$\dot{x} = \frac{\partial \psi}{\partial y}, \qquad \dot{y} = -\frac{\partial \psi}{\partial x}. \tag{2}$$

The relation to Hamilton's canonical equations is established through the identification x=position, y=momentum, and ψ=Hamilton function. Thus, what is the phase space in Hamiltonian systems can be visualized as the position space in the hydrodynamic situation. The structures that appear in 2-d periodically driven flows are, therefore, similar to the phase space structures in a Poincaré surface of section for a chaotic Hamiltonian system, and the same techniques can be used to analyze the transport of particles and the stretching and folding of material lines.

The phenomena that arise in chaotic advection by simple flows may be relevant to turbulent flows when a separation of length and time scales is possible. Consider, for example, the small-scale structures that appear in the density of a tracer substance when the molecular diffusivity κ of the tracer is much smaller than the kinematic viscosity ν of the liquid, that is, in a situation where the Schmidt number $\mathrm{Sc} = \nu/\kappa$ is much larger than one. Then, the velocity field is smooth below the Kolmogorov scale, $\lambda_K = (\nu^3/\varepsilon)^{1/4}$, where ε is the kinetic energy dissipation, but the scalar field has structures on even smaller scales, down to $\lambda_s = (\mathrm{Sc})^{-1/2}\lambda_K$. These arise from Lagrangian chaos with a randomly fluctuating velocity field. The patterns produced in this so-called Batchelor regime are strikingly similar to the ones observed in laminar flows.

On larger scales, ideas from chaotic advection are relevant when there are large-scale coherent structures with slow spatial and temporal evolution. Typical examples are 2-d or quasi-2-d flows, for example, in the atmosphere or in the oceans. Fluid volumes can be trapped in regions bounded by separatrices or by stable and unstable manifolds of stagnation points and may have very little exchange with their surroundings. Such a reduction in stirring appears to occur in the Wadden sea (Ridderinkhof & Zimmerman, 1992).

Equations (1) apply in this form to fluid elements and ideal particles only. For realistic particles with finite volume and inertia, further terms must be added. A significant change in the qualitative side is that the effective velocity field for inertial particles can have a nonvanishing divergence even for incompressible flows (Maxey & Riley, 1983).

The book by Ottino, (1989) and the two conference proceedings (Aref, 1994; IUTAM, 1991) provide good starting points for entering the many aspects of chaotic advection and Lagrangian chaos in engineering applications, geophysical flows, turbulent flows, and theoretical modeling. Historical remarks may be found in the Introduction to Aref, (1994) and in Aref, (2002). Today, the term chaotic advection designates an established subtopic of fluid mechanics that is used as a classification keyword by leading journals and conferences in the field.

HASSAN AREF AND BRUNO ECKHARDT

See also **Chaotic dynamics; Chaos vs. turbulence; Dynamical systems; Hamiltonian systems; Lyapunov exponents; Turbulence**

Further Reading

Aref, H. 1984. Stirring by chaotic advection. *Journal of Fluid Mechanics*, 143: 1–21

Aref, H. (editor). 1994. Chaos applied to fluid mixing. *Chaos, Solitons and Fractals*, 4: 1–372

Aref, H. 2002. The development of chaotic advection. *Physics of Fluids*, 14: 1315–1325

IUTAM Symposium on fluid mechanics of stirring and mixing. 1991. *Physics of Fluids*, 3: 1009–1496

Maxey, M. & Riley, J. 1983, Equation of motion for a small rigid sphere in a nonuniform flow. *Physics of Fluids*, 26: 883–889

Ottino, J.M. 1989. *The Kinematics of Mixing: Stretching, Chaos and Transport*, Cambridge: Cambridge University Press

Ridderinkhof, H. & Zimmermann, J.T.F. 1992. Chaotic stirring in a tidal system. *Science*, 258: 1107–1111

CHAOTIC BILLIARDS

See **Billiards**

CHAOTIC DYNAMICS

When we say “chaos”, we usually imagine a very complex scene with many different elements that move in different directions, collide with each other, and appear and disappear randomly. Thus, according to everyday intuition, the system's complexity (e.g., many degrees of freedom) is an important attribute of chaos. It seems reasonable to think that in the opposite case, for example, a system with only a few degrees of freedom, the dynamical behavior must be simple and predictable. In fact, this point of view is Laplacian determinism.

The discovery of dynamical chaos has destroyed this traditional view. Dynamical chaos is a phenomenon that can be described by mathematical models for many natural systems, for example, physical, chemical, biological, and social, which evolve in time according to a deterministic rule and demonstrate capricious and seemingly unpredictable behavior. To illustrate such behavior, consider a few examples.

Examples

Hyperion: Using Newton's laws, one can compute relatively easily all future solar eclipses not only for the next few hundred years but also for thousands and millions of years into the future. This is indicative of a real predictability of the system's dynamical behavior. But even in the solar system, there exists an object with unpredictable behavior: a small irregularly shaped moon of Saturn, Hyperion. Its orbit is regular and elliptic, but its altitude in the orbit is not. Hyperion is tumbling in a complex and irregular pattern while obeying the laws of gravitational dynamics. Hyperion may not be the only example of chaotic motion in the solar system. Recent studies indicate that chaotic behavior possibly exists in Jovian planets (Murray & Holman, 1999), resulting from the overlap of components of the mean motion resonance among Jupiter, Saturn, and Uranus. Chaos in Hamiltonian systems, which represent the dynamics of the planets, arises when one resonance is perturbed by another one (*See* **Standard map**).

Chaotic mixing is an example of the complex irregular motion of particles in a regular periodic velocity field, like drops of cream in a cup of coffee; see Figure 1. Such mixing, caused by sequential stretching and folding of a region of the flow, illustrates the general mechanism of the origin of chaos in the phase space of simple dynamical systems (*See* **Chaotic advection; Mixing**).

Billiards: For its conceptual simplicity, nothing could be more deterministic and completely predictable than the motion of a single ball on a billiard table. However, in the case of a table bounded by four quarters of a circle curved inward (Sinai billiard), the future fate of a rolling billiard ball is unpredictable beyond a surprisingly small number of bounces. As indicated by Figure 2, a typical trajectory of the Sinai billiard is irregular and a statistical approach is required for a quantitative description of this simple mechanical system. Such an irregularity is the result of having a finite space and an exponential instability of individual trajectories resulting in a sensitive dependence on initial conditions. Due to the curved shape of the boundary, two trajectories emanating from the same point but in slightly different directions with angle δ between them, hit the boundary $\partial\Omega$ (see Figure 2) at different points that are $c\delta$ apart where $c > 0$. After a bounce, the direction of the trajectories will differ by angle $(1+2c)\delta$, and because an actual difference between the directions is multiplied by a factor $\mu = (1+2c) > 1$, the small perturbation δ will grow more or less exponentially (Sinai, 2000). Such sensitive dependence on initial conditions is the main feature of every chaotic system.

A Markov map: To understand in more detail how randomness appears in a nonrandom system, consider a simple dynamical system in the form of a one-dimensional map

$$x_{n+1} = 2x_n \bmod 1. \tag{1}$$

Since the distance between any two nearby trajectories ($|x_n - x'_n| \ll 1$) after each iteration increases at least two times ($|\mathrm{d}x_{n+1}/\mathrm{d}x_n| = 2$), any trajectory of the map is unstable. The map has a countable infinity of unstable periodic trajectories, which can be seen as fixed points when one considers the shape of the map $x_{n+k} = F^{(k)}(x_n)$; see Figure 3(b). Since all fixed points and periodic trajectories are repelling, the only possibility left for the most arbitrarily selected initial condition is that the map will produce a chaotic motion that never exactly repeats itself. The irregularity of such dynamics can be illustrated using a binary symbolic description ($s_n = 0$ if $x_n < \frac{1}{2}$ and $s_n = 1$ if $x_n \geq \frac{1}{2}$). In this case, any value of x_n can be represented as a binary

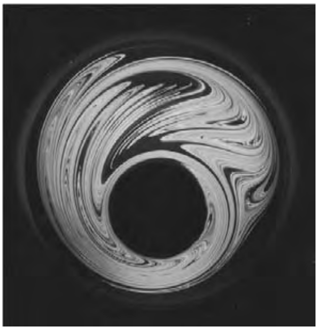

Figure 1. Mixing of a passive tracer in a Newtonian flow between two rotating cylinders with different rotation axes. The rotation speed of the inner cylinder is modulated with constant frequency. The flow is stretched and folded in a region of the flow. The repetition of these operations leads to a layered structure—folds within folds, producing a fractal structure (Ottino, 1989).

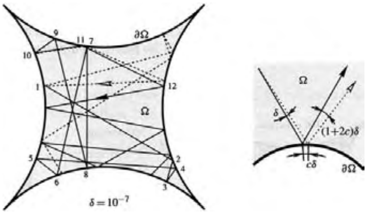

Figure 2. Illustration of the trajectory sensitivity to the initial conditions in a billiard model with convex borders.

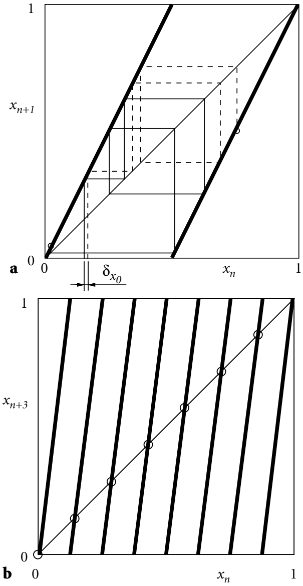

Figure 3. Simple map diagram: (a) two initially close trajectories diverge exponentially; (b) illustration of the increasing of the number of unstable periodic trajectories with the number of iterations.

decimal

$$x_n = 0.s_{n+1}s_{n+2}s_{n+3}\ldots \equiv \sum_{j=n+1}^{\infty} 2^{-j}s_j.$$

If the initial state happens to be a rational number, it can be written as a periodic sequence of 0's and 1's. For instance, 0.10111011101110111… is the rational number $\frac{11}{15}$. Each iteration $x_n \to x_{n+1}$ of map (1) corresponds to setting the symbol s_{n+1} to zero and then moving the decimal point one space to the right (this is known as a *Bernoulli shift*). For example, the iterations of the number $\frac{11}{15}$ yield

$$\begin{aligned}&0.10111011101110111\ldots,\\&0.01110111011101110\ldots,\\&0.11101110111011101\ldots,\\&0.11011101110111011\ldots,\\&0.10111011101110111\ldots,\end{aligned}$$

which illustrates a periodic motion of period 4. Selecting an irrational number as the initial condition, one chooses a binary sequence that cannot be split into groups of 0's and 1's periodically repeated an infinite number of times. As a result, each iteration of the irrational number generates a new irrational number. Since the irrational numbers appear in the interval $x_n \in [0, 1]$ with probability one, one can observe only the aperiodic (chaotic) motions. Random-like behavior of the chaotic motions is illustrated in a separate figure in the color plate section (See the color plate section for a comparison of chaos generated by Equation (1) and a truly random process).

The degree of such chaoticity is characterized by *Lyapunov exponents* that can be defined for one-dimensional maps ($x_{n+1} = f(x_n)$). The stability or instability of a trajectory with the initial state x_0 is determined by the evolution of neighboring trajectories starting at $\tilde{x}_0 = x_0 + \delta x_0$ with $|\delta x_0| \ll 1$. After one iteration

$$\tilde{x}_1 = x_1 + \delta x_1 = f(x_0 + \delta x_0) \approx f(x_0) + \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x=x_0} \delta x_0.$$

Now, the deviation is $\delta x_1 \approx f'(x_0)\delta x_0$. After the nth iteration it becomes $\delta x_n = (\prod_{m=0}^{n-1} f'(x_m))\delta x_0$. The evolution of the distance between the two trajectories is calculated by taking the absolute value of this product. For infinitesimally small perturbations and large enough n, it is expected that $|\delta x_n| = \alpha^n |\delta x_0|$, where

$$\alpha \approx \lim_{n\to\infty}\left(\left|\frac{\delta x_n}{\delta x_0}\right|\right)^{1/n} = \left(\prod_{m=0}^{n-1} |f'(x_m)|\right)^{1/n}$$

or

$$\ln\alpha \approx \lambda = \lim_{n\to\infty}\frac{1}{n}\sum_{m=0}^{n-1}\ln\left(|f'(x_m)|\right). \qquad (2)$$

Limit (2) exists for a typical trajectory x_m and defines the Lyapunov exponent, λ, which is the time average of the rate of exponential divergence of nearby trajectories. For map (1) $f' = 2$ for all values of x and, therefore, $\lambda = \ln 2$ (*See* **Lyapunov exponents**).

Assuming that the initial state cannot be defined with absolute accuracy, the prediction of the state of the map after a sufficiently large number of iterations becomes impossible. The only description that one can use for defining that state is a statistical one. The statistical ensemble in this case is the ensemble of initial conditions. The equation of evolution for the initial state probability density $\rho_{n+1}(F(x))$ can be written as (Ott, 1993, p. 33)

$$\rho_{n+1}(F(x)) = \sum_{j=1,2} \rho_n(x) \Big/ \left|\frac{\mathrm{d}F(x)}{\mathrm{d}x}\right|_j, \qquad (3)$$

where the summation is taken over both branches of $F(x)$. Considering the evolution of a sharp initial distribution $\rho_0(x)$, one can see that at each step this distribution becomes smoother. As n approaches infinity, the distribution asymptotically approaches the steady state $\rho(x) = 1$.

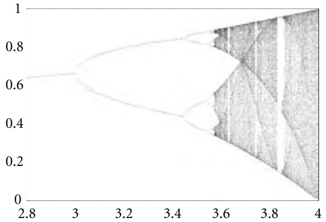

Figure 4. Bifurcation diagram for the logistic map.

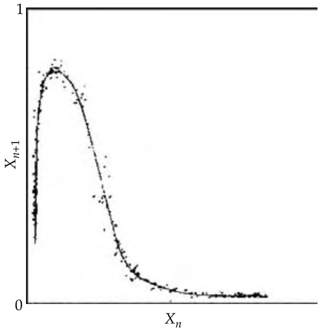

Figure 5. Return map measured in the Belousov–Zhabotinsky autocatalytic reaction.

Population dynamics: A popular model of population growth is the logistic map $x_{n+1} = \alpha x_n(1 - x_n)$, $0 \leq \alpha \leq 4$ (*See* **Population dynamics**). The formation of chaos in this map is illustrated in the bifurcation diagram shown in Figure 4. This diagram presents the evolution of the attracting set as the value of α grows. Below the Feigenbaum point $\alpha_\infty = 3.569\ldots$, the attractor of the map is periodic. Its period increases through a sequence of period-doubling bifurcations as the value of α approaches α_∞ (*See* **Period doubling**). For $\alpha > \alpha_\infty$, the behavior is chaotic but some windows of periodic attractors exist (*See* **Order from chaos**).

Belousov–Zhabotinsky (BZ) autocatalytic reaction: In the BZ reaction (*See* **Belousov–Zhabotinsky reaction**), an acid bromate solution oxidizes malonic acid in the presence of a metalion catalyst and other important chemical components in a well-stirred reactor (Roux et al., 1983). The concentration of the bromide ions is measured and parameterized by the return map (plotting a variable against its next value in time) $x_{n+1} = \alpha x_n \exp[-bx_n]$ (see Figure 5). This map exhibits chaotic behavior for a very broad range of parameter values.

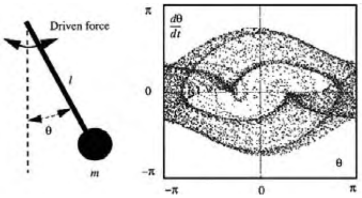

Figure 6. Chaotic oscillation of a periodically driven pendulum, in phase-space plot of angular velocity versus angular position (Deco & Schürmann, 2000).

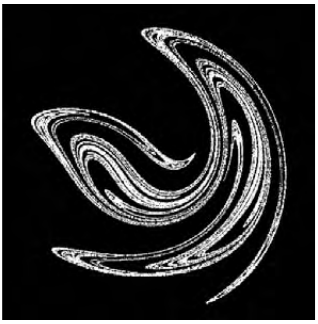

Figure 7. Ueda attractor. The fractal structure of the attractor is typical for all chaotic sets (compare this picture with Figure 1) (Ueda, 1992).

Simple chaotic oscillators: The dynamics of the periodically driven pendulum shown in Figure 6 is described by

$$\frac{\mathrm{d}^2\Theta}{\mathrm{d}t^2} + \nu\frac{\mathrm{d}\Theta}{\mathrm{d}t} + \frac{g}{l}\sin\Theta = B\cos 2\pi f t, \qquad (4)$$

where the term on the right-hand side is the forcing (sinusoidal torque) applied to the pivot and f is the forcing frequency. Chaotic motions of the pendulum computed for $\nu = 0.5$, $g/l = 1$, $B = 1.15$, $f = 0.098$, and visualized with stroboscopic points at moments of time $t = i/f$ are shown in Figure 6.

A similar example of chaotic behavior was intensively studied in an oscillator where the restoring force is proportional to the cube of the displacement (Ueda, 1992, p. 158)

$$\frac{\mathrm{d}^2\Theta}{\mathrm{d}t^2} + \nu\frac{\mathrm{d}\Theta}{\mathrm{d}t} + \Theta^3 = B\cos t. \qquad (5)$$

The stroboscopic image (with $t = i$) of the strange attractor in this forced Duffing-type oscillator computed with $\nu = 0.05$ and $B = 7.5$ is shown in Figure 7.

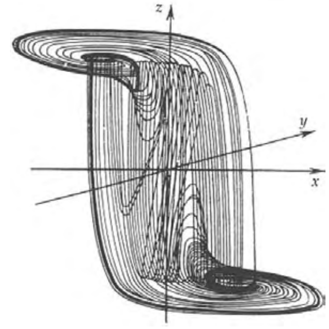

Figure 8. Chaotic attractor generated by electric circuit, which is a modification of van der Pol oscillator: $\dot{x} = hx + y - gz$; $\dot{y} = -x$; $\mu\dot{z} = x - f(x)$; where $f(x) = x^3 - x$ (Pikovsky & Rabinovich, 1978).

Figure 8 presents a chaotic attractor generated by an electronic circuit. Such circuits are a popular topic in engineering studies today.

Characteristics of Chaos

Lyapunov exponents: Consider the Lyapunov exponents for a trajectory $\tilde{\boldsymbol{x}}(t)$ generated by a d-dimensional autonomous system

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{F}(\boldsymbol{x}), \tag{6}$$

with initial condition $\boldsymbol{x}_0$, $\boldsymbol{x} \in \Re^d$. Linearizing Equation (6) about this solution, one obtains a linear system which describes the evolution of infinitesimally small perturbations $\boldsymbol{w} = \boldsymbol{x}(t) - \tilde{\boldsymbol{x}}(t)$ of the trajectory, in the form

$$\frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} = M(\tilde{\boldsymbol{x}})\boldsymbol{w}, \tag{7}$$

where $M(\boldsymbol{x}) = \partial \boldsymbol{F}(\boldsymbol{x})/\partial \boldsymbol{x}$ is the Jacobian of $\boldsymbol{F}(\boldsymbol{x})$ that changes in time in accordance with $\tilde{\boldsymbol{x}}(t)$. In d-dimensional phase space of (7), consider a sphere of initial conditions for perturbations $\boldsymbol{w}(t)$ of diameter l, that is, $|\boldsymbol{w}(0)| \leq l$. The evolution of this ball in time is governed by linear system (7) and depends on trajectory $\tilde{\boldsymbol{x}}(t)$. As the system evolves in time, the ball transforms into an ellipsoid. Let the ellipsoid have d principal axes of different length l_j, $j = 1, d$. Then, the values of Lyapunov exponents of the trajectory $\tilde{\boldsymbol{x}}(t)$ are defined as

$$\lambda_j(\tilde{\boldsymbol{x}}) = \lim_{t\to\infty} \left[\left(\frac{1}{t}\right) \ln \left(\frac{l_j(\tilde{\boldsymbol{x}}, t)}{l(\boldsymbol{x}_0, 0)} \right) \right]. \tag{8}$$

Although limit (8) depends on $\tilde{\boldsymbol{x}}(t)$, the *spectrum of the Lyapunov exponents* λ_j for the selected regime of chaotic oscillations generated by (6) is independent of the initial conditions for the typical trajectories and characterizes the chaotic behavior.

The Lyapunov exponents, λ_j, can be ordered in size: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Self-sustained oscillations in autonomous time-continuous systems always have at least one Lyapunov exponent that is equal to zero. This is the exponent characterizing the stretching of phase volume along the trajectory. The spectrum of λ_j for chaotic trajectories contains one or more Lyapunov exponents with positive values.

Kolmogorov–Sinai entropy is a measure of the degree of predictability of further states visited by a chaotic trajectory started within a small region. Due to the divergence, a long-term observation of such a trajectory gives more and more information about the actual initial condition of the trajectory. In this sense, one may say that a chaotic trajectory creates information. Consider a partitioning of the d-dimensional phase space into small cubes of volume ε^d. Observing a continuous trajectory during T instances of time, one obtains a sequence $\{i_0, i_1, \ldots, i_T\}$, where $i_0, i_1, \ldots$ are the indexes of the cubes consequently visited by the trajectory. As a result, the type of the trajectory observed during the time interval from 0 to T is specified by the sequence $\{i_0, i_1, \ldots, i_T\}$. As Kolmogorov and Sinai showed, in dynamical systems whose behavior is characterized by exponential instability, the number of different types of trajectories, K_T, grows exponentially with T:

$$0 < H = \lim_{T\to\infty} \frac{1}{T} \log K_T.$$

The quantity H is the Kolmogorov–Sinai (KS) entropy.

The number of unique random sequences $\{i_0, i_1, \ldots, i_T\}$ that can be obtained without any rules applied increases exponentially with T. In the case of nonrandom sequences where there is a strict law for the generation of future symbols, like the periodic motion, the number of possible sequences grows in time slower than the exponent. Since the exponential growth takes place for the segments of trajectories in the unstable dynamical system producing chaos, such a dynamical system is capable of generating "random" sequences. The Kolmogorov–Sinai entropy is a measure of such "randomness" in a "nonrandom" system, for example, a dynamical system.

Since both KS entropy and Lyapunov exponents reflect the properties of the divergence of the nearby trajectories, these characteristics are related to each other. The formula describing this relation is given by Ruelle's Inequality

$$H \leq K = \sum_{j=1}^{m} \lambda_j \geq 0 \tag{9}$$

where m is the number of positive λ_i ($K = 0$, when $m = 0$). The equality $H = K$ holds when the system has a physical measure (Sinai–Ruelle–Bowen measure) (Young, 1998). The invariant set of trajectories characterized by a positive Kolmogorov–Sinai entropy is a chaotic set.

Forecasting

If a sufficiently long experimental time series capturing the chaotic process of an unknown dynamical system is available in the form of scalar data $\{x_n\}_{n=0}^{N}$, it is possible, in principle, to predict x_{N+m} with finite accuracy for some $m \geq 1$. Such predictions are based on the assumption that the unknown generating mechanism is time independent. As a result, "what happened in the past may happen again—even stronger: that what is happening now has happened in the past" (Takens, 1991). In classical mechanics (no dissipation), this idea of "what happens now has happened in the past" is related to the Poincaré Recurrence Theorem.

Usually, the prediction procedure consists of two steps: first, it is necessary to consider all values of n in the "past," that is, with $n < N$, such that $\sum_{k=0}^{K} |x_{n-k} - x_{N-k}| < \varepsilon$, where ε is a small constant. If there are only a few of such n, then one can try again with a smaller value of K or a larger value of ε. In the second step, it is necessary to consider the corresponding elements x_{n+l} for all the values of n found in the first step. Finally, taking a union of the ε-neighborhoods of all these elements, one can predict that x_{N+l} will be in this union.

To understand when and why forecasting is possible and when it is not, it is reasonable to use characteristics such as *dimension* and *entropy* that can be computed directly from time series (Takens, 1991). If we want to make a catalog of essentially different segments of length $k+1$ in $\{x_n\}_{n=0}^{N}$, this can be done with $C(k, \varepsilon, N)$ elements. $C(k, \varepsilon, N)$ is a function of N that has a limit $C(k, \varepsilon) = \lim_{N\to\infty} C(k, \varepsilon, N)$, and for prediction, we need $C(k, \varepsilon) \ll N$.

The quantitative measure for the way in which $C(k, \varepsilon)$ increases as ε goes to zero is

$$D = \lim_{k\to\infty} \left(\overline{\lim_{\varepsilon\to 0}} \frac{C(k, \varepsilon, N)}{\ln(1/\varepsilon)} \right). \qquad (10)$$

If D is large, the prediction is problematic. The quantity D defined by (10) is the dimension of the time series.

The quantitative measure for the way in which $C(k, \varepsilon)$ increases with k is

$$H = \overline{\lim_{\varepsilon\to 0}} \left(\lim_{k\to\infty} \frac{C(k, \varepsilon, N)}{k} \right), \qquad (11)$$

This is the entropy of the time series. For the time series generated by a differentiable dynamical system, both the dimension and entropy are finite, but for a random time series they are infinite. Suppose each x_n is taken at random in the interval [0,1] (with respect to the uniform distribution) and for each $n_1, \ldots, n_k$ (different), the choices of $x_{n_1}, \ldots, x_{n_k}$ are independent. For such time series, one can find: $C(k, \varepsilon) = (1 + \lfloor 1/2\varepsilon \rfloor)^{k+1}$, where $\lfloor 1/2\varepsilon \rfloor$ is the integer part of $1/2\varepsilon$. From this formula, it immediately follows that both dimension and entropy in such random time series are infinite.

Models of the Earth's atmosphere are generally considered as chaotic dynamical systems. Due to the unstability, even infinitesimally small uncertainties in the initial conditions grow exponentially fast and make a forecast useless after a finite time interval. This is known as the butterfly effect. However, in the tropics, there are certain regions where wind patterns and rainfall are so strongly determined by the temperature of the underlying sea surface, that they do not show such sensitive dependence on the atmosphere. Therefore, it should be possible to predict large-scale tropical circulation and rainfall for as long as the ocean temperature can be predicted (Shukla, 1998).

History

The complex behavior of nonlinear oscillatory systems was observed long before dynamical chaos was understood. In fact, the possibility of complex behavior in dynamical systems was discovered by Henri Poincaré in the 1890s in his unsuccessful efforts to prove the regularity and stability of planetary orbits. Later on, experiments with an electrical circuit by van der Pol and van der Mark (1927) and the double-disk model experiments of the magnetic dynamo (Rikitake, 1958) also indicated the paradoxically complex behavior of a simple system. At that time, several mathematical tools were available to aid the description of the nontrivial behavior of dynamical systems in phase space, such as homoclinic Poincaré structures (homoclinic tangles). However, at the time, neither physicists nor mathematicians realized that deterministic systems may behave chaotically. It was only in the 1960s that the understanding of randomness was revolutionized as a result of discoveries in mathematics and in computer modeling (Lorenz, 1963) of real systems. An elementary model of chaotic dynamics was suggested by Boris Chirikov in 1959. During the last few decades, chaotic dynamics has moved from mystery to familiarity.

Standard map and homoclinic tangle: The standard map (Chirikov, 1979) is an area-preserving map

$$\begin{aligned} I_{n+1} &= I_n + K \sin \Theta_n, \\ \Theta_{n+1} &= I_n + \Theta_n + K \sin \Theta_n, \end{aligned} \qquad (12)$$

where Θ is an angle variable (computed modulo 2π) and k is a positive constant. This map was proposed as a model for the motion of a charged particle in a magnetic field. For K larger than K_{cr}, map

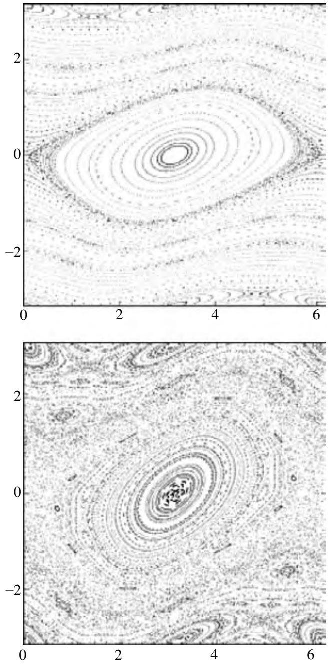

Figure 9. Examples of chaos in the standard map for two different values of K. The coexistence of the "chaotic sea" and "regular islands" that one can see in the panel on the right is typical for Hamiltonian systems with chaotic regimes (Lichtenberg & Lieberman, 1992).

(12) demonstrates an irregular (chaotic) motion; see Figure 9. The complexity of the phase portrait of this map is related to the existence of *homoclinic tangles* formed by stable and unstable manifolds of a saddle point or saddle periodic orbits when the manifolds intersect transversally. The complexity of the manifold's geometry stems from the fact that, if stable and unstable manifolds intersect once, then they must intersect an infinite number of times. Such a complex structure results in the generation of a horseshoe mapping, which persistently stretches and then folds the area around the manifolds generating a chaotic motion. The layers of the chaotic motion are clearly seen in Figure 9.

Lorenz system: The first clear numerical manifestation of chaotic dynamics was obtained in the Lorenz model. This model is a three-dimensional dynamical system derived from a reasonable simplification of the fluid dynamics equations for thermal convection in a liquid layer heated from below. The differential equations $\dot{x} = \sigma(y - x)$, $\dot{y} = rx - y - xz$, $\dot{z} = -bz + xy$ are written for the amplitude of the first horizontal harmonic of the vertical velocity (x), the amplitude of the corresponding temperature fluctuation (y), and a uniform correction of the temperature field (z) (Lorenz, 1963). σ is the Prandtl number, r is the reduced Rayleigh number, and b is a geometric factor. The phase portrait of the Lorenz attractor, time series, and the return mapping generated on the Poincaré cross section computed for $r = 28$, $\sigma = 10$, and $b = \frac{8}{3}$ are presented in Figure 10. A simple mechanical model illustrating the dynamical origin of oscillations in the Lorenz system is shown in Figure 11 (*See* **Lorenz equations**).

Definition of Chaos

As was shown above, dynamical chaos is related to unpredictability. For quantitative measurment of the unpredictability, it is reasonable to use the familiar characteristics *dimension* and *entropy*. These characteristics are independent: it is possible to generate a time series that has a high dimension and at the same time entropy equal to zero. This is a quasiperiodic motion. It is also simple to imagine a low-dimensional dynamical system with high entropy (see, e.g., the map in Figure 3).

Various definitions of chaos exist, but the common feature of these definitions is the sensitive dependence on initial conditions that was formalized above as positive entropy. Thus, dynamical chaos is the behavior of a dynamical system that is characterized by finite positive entropy.

Chaotic Attractors and Strange Attractors

A region in the phase space of a dissipative system that attracts all neighboring trajectories is called an attractor. An attractor is the phase space image of the behavior established in the dissipative system, for example, a stable limit cycle is the image of periodic oscillations. Therefore, the image of chaotic oscillations is a chaotic attractor.

A *chaotic attractor* (CA) possesses the following two properties that define any attractor of the dynamical system:

- There exists a bounded open region U containing a chaotic attractor ($\mathrm{CA} \in U$) in the phase space such that all points from this neighborhood converge to a chaotic attractor when time goes to infinity.
- A chaotic attractor is invariant under the evolution of the system,

In addition, the motion on a chaotic attractor has to be chaotic, for example:

- each trajectory of a chaotic attractor has at least one positive Lyapunov exponent.

Such types of attractors represent some regimes of chaotic oscillations generated by a Lorenz system and

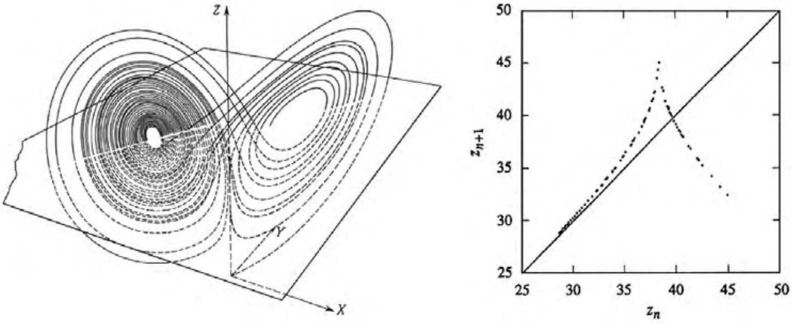

Figure 10. Lorenz attractor (left) and the return map $z_{n+1} = F(z_n)$ plotted for maximum values of variable z for the attractor trajectory (right).

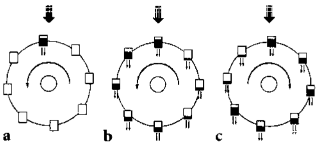

Figure 11. A toy model invented by Willem Malkus and Lou Howard illustrates dynamical mechanisms analogous to oscillations and chaos in the Lorenz system. Water steadily flowing into the top (leaky) bucket makes it heavy enough to start the wheel turning. When the flow is large enough, the wheel can start generating chaotic rotations characterized by unpredictable switching of the rotation direction; see Strogatz (1994, p. 302) for details.

the piece-wise linear maps. However, most of the chaotic oscillations observed in dynamical systems correspond to attractors that do not precisely satisfy the latter property. Although almost all trajectories in such attractors are unstable, some stable periodic orbits may exist within the complex structure of unstable trajectories. Chaos in such systems is persistent both in physical experiments and in numerical simulations because all of these stable orbits have extremely narrow basins of attraction. Due to natural small perturbations of the system, the trajectory of the system never settles down on one of the stable orbits and wanders within the complex set of unstable orbits.

The definition of a *strange attractor* is related to the complicated geometrical structure of an attractor. A strange attractor is defined as an attractor that cannot be presented by a union of the finite number of smooth manifolds. For example, an attractor whose topology can be locally represented by the direct product of a Cantor set to a manifold is a strange attractor. In many cases, the geometry of a chaotic attractor satisfies the definition of a strange attractor. At the same time, the definition of a strange attractor can be satisfied in the case of a nonchaotic strange attractor. This is an attractor that has fractal structure, but does not have positive Lyapunov exponents.

The origin of chaotic dynamics in dissipative systems and Hamiltonian systems in many cases is the same and is related to coexistence in the phase space of infinitely many unstable periodic trajectories as a part of homoclinic or heteroclinic tangles.

The Lorenz attractor, as for many other attractors in systems with a small number of degrees of freedom, can appear through a finite number of easily observable bifurcations. The bifurcation of a sudden birth and death of a strange attractor is called a crisis. Usually, it is related to the collision of the attractor with an unstable periodic orbit or its stable manifold (Arnol'd et al., 1993; Ott, 1993).

Order in Chaos

How does the dynamical origin imprint in chaos? Or in other words, how can the rules or order of the dynamical system be found inside a chaotic behavior? Consider the images (portraits) of the dynamical chaos shown in Figures 7, 8, and 10. The elegance of these images reflects the existence of order in dynamical chaos.

The dynamical origin of such elegance is very similar: different trajectories with close initial conditions have to be close in time $t_l \approx 1/\lambda$, where λ is the maximally positive Lyapunov exponent. The domain occupied by the strange attractor in phase space is finite; thus, the divergence of the phase space flow changes to convergence, and as a result of sequential action of divergence and convergence of the phase flow in the finite domain, the mixing of trajectories occurs. Such mixing can be illustrated with the motions of liquids in the physical space experimentally observed by Ottino (1989; see Figure 1).

Another way to recognize the existence of order in chaos is to analyze its dependence on a control parameter. The macroscopic features of real stochastic

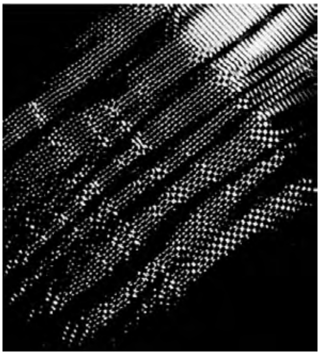

Figure 12. Appearance of spatiotemporal chaos in the extended Faraday experiment: chaotic patterns on the surface of the liquid layer in the oscillating gravitational field. The irregular chain of the localized structures—dark solitons—can be seen beneath a background of the square capillary lattice (Gaponov-Grekhov & Rabinovich, 1988).

processes, for example, Brownian motion or developed turbulence, depend on this parameter and change without any revolutionary events such as bifurcations. But for dynamical chaos, the picture is different. A continuous increase of control parameters of the logistic map does not necessarily gradually increase the degree of chaos: within chaos, there are windows—intervals of control parameter values in which the chaotic behavior of the system changes to stable periodic behavior, see Figure 4.

In a spatially extended system, for example, in convection or Faraday flow, order within chaos is related to the existence of coherent structures inside the chaotic sea (Rabinovich et al., 2001); see Figure 12.

Spatiotemporal Chaos

Similar to regular (e.g., periodic) motions, low-dimensional chaotic behavior is observed not only in simple (e.g., low-dimensional) systems but also in systems with many, and even with infinite number of degrees of freedom. The dynamical mechanisms behind the formation of low-dimensional chaotic spatiotemporal patterns in dissipative and nondissipative systems are different. In conservative systems, such patterns are related to the chaotic motion of particle-like localized structures. For example, a soliton that is described by a nonlinear Schrödinger equation with the harmonic potential

$$\mathrm{i}\frac{\partial a}{\partial t}+\beta\frac{\partial^2 a}{\partial^2 x}+\left(|a|^2+\alpha\sin qx\right)a=0 \qquad (13)$$

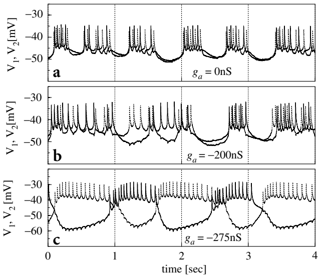

Figure 13. Dynamics of chaotic bursts of spikes generated by two living neurons coupled with an electrical synapse—a gap junction (Elson et al., 1998). Chaotic busts in naturally coupled neurons synchronize (a). When natural coupling is compensated by additional artificial coupling g_a, the chaotic oscillations are independent oscillations (b). The neurons coupled with negative conductivity fire in the regimes of antiphase synchronization (c).

moves chaotically in physical space x and reminds us of the chaotic motion occurring in the phase space of a parametrically excited conservative oscillator (the equations of such an oscillator can be derived from (13) for slow variables characterizing the motion of the soliton center mass).

The interaction of the localized structures (particles) in a finite area, large in comparison with the size of the structure, can also lead to the appearance of spatiotemporal chaos. It was observed that collisions of solitons moving in two-dimensional space result in chaotic scattering similar to the chaotic motion observed in billiards (Gorshkov et al., 1992).

In dissipative nonlinear media and high-dimensional discrete systems, the role of coherent structures is also very important (such as defects in convection, clusters of excitations in neural networks, and vortices in the wake behind a cylinder; see Rabinovich et al., 2001). However, the origin of low-dimensional chaotic motions in such systems is determined by dissipation. There are two important mechanisms of finite dynamics (including chaos) that are due to dissipation: (1) the truncation of the number of excited modes (in hydrodynamic flows) due to high viscosity of the small-scale perturbations and (2) the synchronization of the modes or individual oscillators. Dissipation makes synchronization possible not only among periodic modes or oscillators but even in the case when the interacting subsystems are chaotic (Afraimovich et al., 1986). Figure 13 illustrates the synchronization of chaotic bursts of spikes observed experimentally in two coupled living neurons. In the case of a dissipative lattice of chaotic elements (e.g., neural lattices or models of an extended autocatalytic chemical reaction), complete synchronization leads to the onset

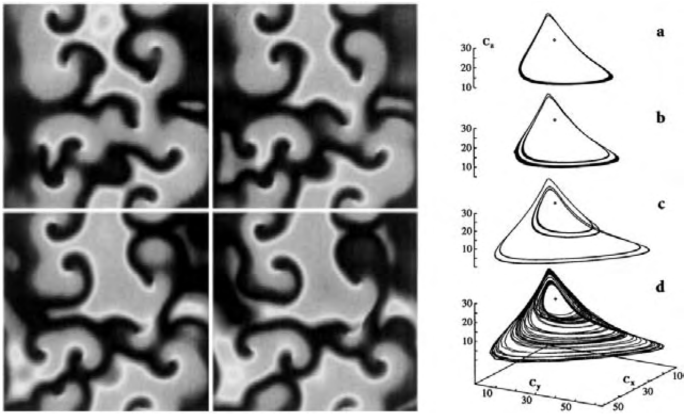

Figure 14. Coherent patterns generated in the chaotic medium with Rössler-type dynamics of medium elements. Left: an example of coherent patterns with defects. Right: evolution of the attractor with increasing distance r from a defect. The attractor changed from the limit cycle of period T at $r = r_1$ to the period $2T$ limit cycle at $r = r_2 > r_1$, then to the period $4T$ limit cycle at $r = r_3 > r_2$, and finally to the chaotic attractor for $r = r_4 > r_3$ (Goryachev & Kapral, 1996).

of a spatially homogeneous chaotic state. When this state becomes unstable against spatial perturbations, the system moves to the spatiotemporal chaotic state. A snapshot of such spatiotemporal chaos, which is observed in the model of chaotic media consisting of diffusively coupled Rössler-type chaotic oscillators, is presented in Figure 14. Figure 14 also illustrates the sequence of period-doubling bifurcations that are observed in the neighborhood of the defect in such a medium.

Edge of Chaos

In dynamical systems with many elements and interconnections (e.g., complex systems), the transition between ordered dynamics and chaos is similar to phase transitions between states of matter (crystal, liquid, gas, etc.). Based on this analogy, an attractive hypothesis named "edge of chaos"(EOC) appeared at the end of the 1980s. EOC suggests a fundamental equivalence between the dynamics of phase transitions and the dynamics of information processing (computation). One of the simplest frameworks in which to formulate relations between complex system dynamics and computation at the EOC is a cellular automaton. There is currently some controversy over the validity of this idea (Langton 1990; Mitchel et al., 1993).

Chaos and Turbulence

The discovery of dynamical chaos has fundamentally changed the accepted concept of the origin of hydrodynamic turbulence. When dealing with turbulence at finite Reynolds number, the main point of interest is the established irregular motion. The image of such irregularity in the phase space could be a chaotic attractor. Experiments in closed systems, for example, one in which fluid particles continuously recirculate through points previously visited, have shown the most common scenarios for the transition to chaos. These are (i) transition through the destruction of quasiperiodic motion that was observed in Taylor–Couette flow (Gollub & Swinney, 1975); (ii) period-doubling sequence observed in Rayleigh–Bénard convection (Libchaber & Maurer, 1980); and (iii) transition through intermittency (Gollub & Benson, 1980). Observation of these canonical scenarios for particular flows proved the validity of the concept of dynamical origin of the transition to turbulence in closed systems. It is possible to reconstruct a chaotic set in the phase space of the flow directly from observed data; see Brandstäter et al. (1982). At present it is difficult to say how dynamical chaos theory can be useful for the understanding and description of the developed turbulence.

The discovery and understanding of chaotic dynamics have important applications in all branches of science and engineering and, in general, to our evolving culture. An understanding of the origins of chaos in the last decades has produced many clear and useful models for the description of systems with complex behavior, such as the global economy (Barkly Russel, 2000), the human immune system (Gupta et al., 1998), animal behavior (Varona et al., 2002), and more. Thus, chaos theory provides a new tool for the unification of the sciences.

M.I. Rabinovich and N.F. Rulkov

See also **Attractors; Billiards; Butterfly effect; Chaos vs. turbulence; Controlling chaos; Dripping faucet; Duffing equation; Entropy; Fractals; Hénon map; Horseshoes and hyperbolicity in dynamical systems; Intermittency; Kicked rotor; Lorenz equations; Lyapunov exponents; Maps; Maps in the complex plane; Markov partitions; Multifractal analysis; One-dimensional maps; Order from chaos; Period doubling; Phase space; Quasiperiodicity; Rössler systems; Routes to chaos; Sinai–Ruelle–Bowen measures; Spatiotemporal chaos; Synchronization; Time series analysis**

Further Reading

Abarbanel, H.D.I. 1996. *Analysis of Chaotic Time Series*, New York: Springer

Afraimovich, V.S., Verichev, N.N. & Rabinovich, M.I. 1986. Stochastic synchronization of oscillations in dissipative systems. *Izvestiya Vysshikh Vchebnykh Zavedenii Radiofizika. RPQAEC*, 29: 795–803

Arnol'd, V.I., Afraimovich, V.S., Ilyashenko, Yu.S. & Shilnikov, L.P. 1993. Bifurcation theory and catastrophe theory. In *Dynamical Systems*, vol. 5, Berlin and New York: Springer

Barkly Russel, J., Jr. 2000. *From Catastrophe to Chaos: A General Theory of Economic Discontinuities*, 2nd edition, Boston: Kluwer

Brandstäter, A., Swift, J., Swinney, H.L., Wolf, A., Doyne Farmer, J., Jen, E. & Crutchfield, P.J. 1982. Low-dimensional chaos in a hydrodynamic system. *Physical Review Letters*, 51: 1442–1445

Chirikov, V.A. 1979. A universal instability of many-dimensional oscillator systems. *Physics Reports*, 52: 264–379

Deco, G. & Schürmann, B. 2000. *Information Dynamics: Foundations and Applications*, Berlin and New York: Springer

Elson, R.C., Selverston, A.I., Huerta, R., Rulkov, N.F., Rabinovich, M.I. & Abarbanel H.D.I. 1998. Synchronous behavior of two coupled biological neurons. *Physical Review Letters*, 81: 5692–5695

Gaponov-Grekhov, A.V. & Rabinovich, M.I. 1988. *Nonlinearity in Action: Oscillations, Chaos, Order, Fractals*, Berlin and New York: Springer

Gollub, J.P. & Benson, S.V. 1980. Many routes to turbulent convection. *Journal of Fluid Mechanics*, 100: 449–470

Gollub, J.P. & Swinney, H.L. 1975. Onset of turbulence in rotating fluid. *Physical Review Letters*, 35: 927–930

Gorshkov, K.A., Lomov, A.S. & Rabinovich, M.I. 1992. Chaotic scattering of two-dimensional solitons. *Nonlinearity*, 5: 1343–1353

Goryachev, A. & Kapral, R. 1996. Spiral waves in chaotic systems. *Physical Review Letters*, 76: 1619–1622

Gupta, S., Ferguson, N. & Anderson, R. 1998. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agent. *Science*, 280: 912–915

Langton, C.C. 1990. Computation at the edge of chaos—phase transitions and emergent computation. *Physica* D, 42: 12–37

Libchaber, A. & Maurer, J. 1980. Une expérience de Rayleigh-Bénard en géométrie réduite; multiplication, accrochage et démultiplication de fréquences. *Journal de Physique Colloques*, 41: 51–56

Lichtenberg, A.J. & Lieberman, M.A. 1992. *Regular and Chaotic Dynamics*, Berlin and New York: Springer

Lorenz, E.N. 1963. Deterministic nonperiodic flow. *Journal of Atmospheric Science*, 20: 130–136

Mitchel, M., Hraber, P. & Crutchfield, J. 1993. Revisiting the edge of chaos: evolving cellular automata to perform computations. *Complex Systems*, 7: 89–130

Murray, N. & Holman, M. 1999. The origin of chaos in the outer solar system. *Science*, 283: 1877–1881

Ott, E. 1993. *Chaos in Dynamical Systems*, Cambridge and New York: Cambridge University Press

Ottino, J.M. 1989. *The Kinetics of Mixing: Stretching, Chaos, and Transport*, Cambridge and New York: Cambridge University Press

Pikovsky, A.S. & Rabinovich, M.I. 1978. A simple generator with chaotic behavior. *Soviet Physics Doklady*, 23: 183–185 (see also Rabinovich, M.I. 1978. Stochastic self-oscillations and turbulence. *Soviet Physics Uspekhi*, 21: 443–469)

Rabinovich, M.I., Ezersky, A.B. & Weidman, P.D. 2001. *The Dynamics of Patterns*, Singapore: World Scientific

Rikitake, T. 1958. Oscillations of a system of disk dynamos. *Proceedings of the Cambridge Philosophical Society*, 54: 89–105

Roux, J.C., Simoyi, R.H. & Swinney, H.L. 1983. Observation of a strange attractor. *Physica* D, 8: 257–266

Shukla, J. 1998. Predictability in the midst of chaos: a scientific basis for climate forecasting. *Science*, 282: 728–731

Sinai, Ya.G. 2000. *Dynamical Systems, Ergodic Theory and Applications*, Berlin and New York: Springer

Strogatz, S.H. 1994. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* Reading, MA: Addison-Wesley

Takens, F. 1991. Chaos, In *Structures in Dynamics: Finite Dimensional Deterministic Studies*, edited by H.W. Broer, F. Dumortier, S.J. van Strien & F. Takens, Amsterdam: North-Holland Elsevier Science

Ueda, Y. 1992. *The Road to Chaos*, Santa Cruz, CA: Aeirial Press

van der Pol, B. & van der Mark, B. 1927. Frequency demultiplication. *Nature*, 120: 363–364

Varona, P., Rabinovich, M.I., Selverston, A.I. & Arshavsky, Yu.I. 2002. Winnerless competition between sensory neurons generates chaos: A possible mechanism for molluscan hunting behavior. *CHAOS* 12: 672–677

Young, L.S. 1998. Developments in chaotic dynamics. *Notices of the AMS*, 17: 483–504

CHARACTERISTICS

Systems of first-order partial differential equations describe many different physical phenomena from the behavior of fluids, gases, and plasmas. To introduce the Method of Characteristics, consider the simple scalar conservation law of the form

$$\frac{\partial U}{\partial t} + A(U)\frac{\partial U}{\partial x} = 0. \qquad (1)$$

Here, $U = U(x, t)$, where x is a spatial coordinate and t is the time coordinate. The function $A(U)$ defines the speed of propagation of a disturbance and either may be independent of U, in which case equation (1) is a linear partial differential equation, or it may depend explicitly on the dependent variable U, in which case the equation is a nonlinear partial differential equation.

It is important to specify the initial or boundary conditions that the solution $U(x, t)$ must satisfy.

Consider the simple case where the function is known at the initial time $t = 0$. Thus,

$$U(x, 0) = F(x). \tag{2}$$

The idea is to simplify Equation (1) by choosing a suitable curve in the x-t plane. This curve can be written in parametric form as

$$x = x(s), \quad t = t(s), \tag{3}$$

where s is the parameter that can be thought of as measuring the distance along the curve. To understand how to select the particular form of the curve, note that, using (3),

$$U(x, t) = U(x(s), t(s)), \tag{4}$$

implies that U is a function of s. Hence, the chain rule gives the derivatives of U along the curve as

$$\frac{\mathrm{d}U}{\mathrm{d}s} = \frac{\partial U}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}s} + \frac{\partial U}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}s}. \tag{5}$$

Comparing the right-hand side of (5) with the left-hand side of (1), it is clear that they are identical, provided the parametric form of the curve is chosen as

$$\frac{\mathrm{d}t}{\mathrm{d}s} = 1, \tag{6}$$

$$\frac{\mathrm{d}x}{\mathrm{d}s} = A(U), \tag{7}$$

and (1) reduces to

$$\frac{\mathrm{d}U}{\mathrm{d}s} = 0, \Rightarrow U = \text{constant along the curve.} \tag{8}$$

The curve satisfied by (6) and (7) is called the *characteristic curve*. Along this curve, U is constant. However, the value of the constant may be different on each characteristic curve. The solution of the characteristic equations requires some initial conditions for x and t. These are taken as

$$x = x_0, \quad t = 0, \quad \text{at } s = 0. \tag{9}$$

Note that x_0 covers the same domain as x. Solving (6)–(8) yields

$$t = s, \quad x = A(U)s + x_0, \tag{10}$$

on using the initial conditions (9). x_0 can be thought of as a constant of integration and so it has a fixed value along the characteristic curve. This implies, using (2), that

$$U(x, 0) = F(x_0). \tag{11}$$

Note that the particular characteristic curve is determined by the value of the parameter x_0. Hence, eliminating the parameter s and solving for x_0 in terms of x, U, and t, the solution given by (11) is

$$U(x, t) = F(x_0) = F(x - A(U)t). \tag{12}$$

If $A(U)$ is a constant, say c, then the solution is simply

$$U(x, t) = F(x - ct), \tag{13}$$

but if $A(U)$ depends explicitly on U, then the solution is an *implicit solution*.

The characteristic curves in this example are straight lines in the x-t plane with a gradient given by $A(U)$. When $A(U) = c$ is a constant, the characteristic curves are parallel straight lines. This means that the shape of the initial disturbance propagates unchanged, to the right if c is a positive constant. If $A(U)$ depends on U, then the characteristic curves are straight lines but with different gradients. There exists the possibility that the characteristic curves may cross, and this corresponds to the formation of a shock. When the characteristic curves diverge, the solution exhibits an expansion fan that can be expressed in terms of a similarity variable.

Note that the method of characteristics can be used when $A = A(U, x, t)$ depends explicitly on the space and time coordinates. In this case, the coupled equations, (6)–(8), may be solved numerically. A detailed description of the method of characteristics for general first-order partial differential equations is given in Rubenstein & Rubenstein (1993).

Example: Burgers Equation

Consider the case when $A(U) = U$. Then, (1) becomes the inviscid Burgers equation

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} = 0, \tag{14}$$

and the solution satisfying the initial condition

$$U(x, 0) = \begin{cases} 0, & x < -1, \\ 1 + x, & -1 \le x < 0, \\ 1 - x, & 0 \le x \le 1, \\ 0, & x > 1 \end{cases} \tag{15}$$

is

$$U = \begin{cases} 0, & x < -1 + Ut, \\ 1 + x - Ut, & -1 + Ut \le x < Ut, \\ 1 - x + Ut, & Ut \le x \le 1 + Ut, \\ 0, & x > 1 + Ut. \end{cases} \tag{16}$$

Thus, solving for U in each region gives the solution as

$$U = \begin{cases} 0, & x < -1, \\ (1 + x)/(1 + t), & -1 \le x < t, \\ (1 - x)/(1 - t), & t \le x \le 1, \\ 0, & x > 1. \end{cases} \tag{17}$$

Note that the solution becomes multi-valued for $t > 1$. This can be understood by considering the characteristic curves defined by (9). In the x-t plane, they are straight lines of the form

$$x = Ut + x_0, \tag{18}$$

so that the gradient depends on the value of U at $t = 0$ and $x = x_0$. Thus, using the initial conditions, the characteristic curves, valid in the region $t \leq x \leq 1$, can be expressed as

$$x = (1 - x_0)t + x_0, \tag{19}$$

for $0 \leq x_0 \leq 1$. Considering two different values of x_0, say x_a and x_b, the straight lines cross when

$$\begin{aligned} x &= (1 - x_a)t + x_a = (1 - x_b)t + x_b, \\ \Rightarrow &\qquad t = 1 \quad \text{and} \quad x = 1. \end{aligned} \tag{20}$$

Hyperbolic Systems of Several Dependent Variables

Systems of first-order hyperbolic equations can be expressed in vector and matrix form as

$$\frac{\partial \boldsymbol{U}}{\partial t} + \boldsymbol{A}(\boldsymbol{U}, x, t)\frac{\partial \boldsymbol{U}}{\partial x} = 0, \tag{21}$$

where $\boldsymbol{U}$ is a column vector of n elements containing the dependent variables and $\boldsymbol{A}$ is an $n \times n$ matrix whose coefficients may depend on the dependent variables. The problem is linear if the matrix $\boldsymbol{A}$ has elements independent of $\boldsymbol{U}$ and nonlinear otherwise. The characteristic curves in this case are given by the equations

$$\frac{\mathrm{d}t}{\mathrm{d}s} = 1, \tag{22}$$

$$\frac{\mathrm{d}x}{\mathrm{d}s} = \lambda_i(\boldsymbol{U}, x, t) \tag{23}$$

for $i = 1, 2, ..., n$ and where λ_i is an eigenvalue of the matrix $\boldsymbol{A}$. Here, it is assumed that the matrix $\boldsymbol{A}$ has n distinct eigenvalues. For the linear problem, and in particular, for the case where the matrix $\boldsymbol{A}$ has constant coefficients, the full solution can be obtained by using a suitable linear combination of dependent variables so that the equations reduce to a set of simple advection equations. Hence, the first step is to determine the eigenvalues, λ_i, of the matrix $\boldsymbol{A}$ and the corresponding eigenvectors $\boldsymbol{z}_i$, where

$$\boldsymbol{A}\boldsymbol{z}_i = \lambda_i \boldsymbol{z}_i, \tag{24}$$

Next, use the change of variable

$$\boldsymbol{U} = \boldsymbol{Q}\boldsymbol{V}, \tag{25}$$

where $\boldsymbol{Q}$ is an $n \times n$ matrix whose jth column is the jth eigenvector $\boldsymbol{z}_j$. Substituting into (21) yields

$$\boldsymbol{Q}\frac{\partial \boldsymbol{V}}{\partial t} + \boldsymbol{A}\boldsymbol{Q}\frac{\partial \boldsymbol{V}}{\partial x} = 0. \tag{26}$$

Finally, pre-multiplying by $\boldsymbol{Q}^{-1}$, the inverse of $\boldsymbol{Q}$, results in a decoupled system of equation, since $\boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{Q}$ is a diagonal matrix whose elements are the eigenvalues λ_i. Thus, the final set of n equations are

$$\frac{\partial V_i}{\partial t} + \lambda_i \frac{\partial V_i}{\partial x} = 0, \tag{27}$$

for $i = 1, 2, ..., n$. The solutions to (27) are simply

$$V_i = F_i(x - \lambda_i t), \tag{28}$$

where F_i is an arbitrary function determined by the initial conditions. Once all the solutions for V_i are determined from the initial conditions, the solution in terms of the original variables is obtained using (25). Note that while the original variables may depend on all the characteristic variables, the V_i solution is *constant* along the ith characteristic curve.

Example: The Second-Order Wave Equation

The second-order wave equation

$$\frac{\partial^2 U}{\partial t^2} = c^2 \frac{\partial^2 U}{\partial x^2} \tag{29}$$

can be expressed as a pair of first-order equations as

$$\begin{aligned} \frac{\partial U}{\partial t} &= -c\frac{\partial p}{\partial x}, \\ \frac{\partial p}{\partial t} &= -c\frac{\partial U}{\partial x}. \end{aligned}$$

Thus,

$$\boldsymbol{A} = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix}. \tag{30}$$

The eigenvalues are simply $\lambda_1 = c$ and $\lambda_2 = -c$ and the corresponding eigenvectors are

$$\boldsymbol{z}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad \boldsymbol{z}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{31}$$

Thus,

$$\boldsymbol{Q} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad \boldsymbol{Q}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}. \tag{32}$$

Equation (27) reduces to the pair of equations

$$\frac{\partial V_1}{\partial t} - c\frac{\partial V_1}{\partial x} = 0, \qquad \frac{\partial V_2}{\partial t} + c\frac{\partial V_2}{\partial x} = 0. \tag{33}$$

The solutions are $V_1 = F_1(x+ct)$ and $V_2 = F_2(x-ct)$ and, in terms of the original variables, the solution is

$$\begin{aligned} U &= F_1(x+ct) + F_2(x-ct), \\ p &= -F_1(x+ct) + F_2(x-ct). \end{aligned} \qquad (34)$$

Riemann Invariants

A Riemann invariant may be thought of as a function of the dependent variables that is constant along a characteristic curve. In the previous example, it is clear that $U + p = 2F_2(x-ct)$, so $U + p$ is a Riemann invariant along the characteristic curve defined by $x - ct = \text{constant}$. Similarly, $U - p$ is constant along $x + ct = \text{constant}$.

Example: Isentropic Flow

The dimensionless equations for *isentropic* fluid flow with $p/\rho^\gamma = 1$ can be expressed in terms of the fluid velocity, u, and the sound speed $c = (\gamma p/\rho)^{1/2}$ as

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{2}{\gamma - 1}c\frac{\partial c}{\partial x} = 0, \qquad (35)$$

$$\frac{\partial c}{\partial t} + \frac{\gamma - 1}{2}c\frac{\partial u}{\partial x} + u\frac{\partial c}{\partial x} = 0. \qquad (36)$$

A detailed derivation of these equations is given in Kevorkian (1989). The matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{pmatrix} u & \frac{2}{\gamma-1}c \\ \frac{\gamma-1}{2}c & u \end{pmatrix} \qquad (37)$$

having eigenvalues $\lambda_1 = u + c$ and $\lambda_2 = u - c$. Thereare two characteristic curves given by the solution of the coupled differential equations

$$\begin{aligned} \frac{\mathrm{d}t}{\mathrm{d}s} &= 1, \\ \frac{\mathrm{d}x}{\mathrm{d}s} &= \lambda_i, \end{aligned}$$

where $i = 1$ or $i = 2$. For $i = 1$, the initial conditions are $t = 0$ and $x = \xi$ at $s = 0$, which implies that the characteristic curve is defined by $\xi = x - \int_0^t (u + c)\,\mathrm{d}s = \text{constant}$. For $i = 2$, $t = 0$, and $x = \eta$ at $s = 0$, the second curve is defined by $\eta = x - \int_0^t (u - c)\,\mathrm{d}s = \text{constant}$. Multiplying (35) by $(\gamma - 1)/2$ and then adding and subtracting (36) gives two equations

$$\begin{aligned} \frac{\partial R}{\partial t} + (u + c)\frac{\partial R}{\partial x} &= 0, \\ \frac{\partial S}{\partial t} + (u - c)\frac{\partial S}{\partial x} &= 0, \end{aligned}$$

where $R = c + (\gamma - 1)u/2$ and $S = c - (\gamma - 1)u/2$. R and S are Riemann invariants since R is constant along the characteristic curve defined by $\xi = \text{constant}$) and S is constant along $\eta = \text{constant}$). A more detailed derivation of Riemann invariants is described in Kevorkian (1989).

ALAN HOOD

See also **Burgers equation; Coupled systems of partial differential equations; Hodograph transform; Shock waves**

Further Reading

Kevorkian, J. 1989. *Partial Differential Equations: Analytical Solution Techniques*, New York: Chapman & Hall

Rubenstein, I. & Rubenstein, L. 1993 *Partial Differential Equations in Classical Mathematical Physics*, Cambridge and New York: Cambridge University Press

CHARGE DENSITY WAVES

A charge density wave (CDW) is a collective transport phenomenon, whose origin lies in the interaction between electrons and phonons in a solid (Grüner & Zettl 1985; Grüner, 1988). As envisioned by Rudolph Peierls in 1930 in some quasi-one-dimensional metals (where the influence of one electron to each other electron is much stronger than in higher dimensions), the elastic energy needed to displace the position of the atoms may be balanced by a lowering of conduction electron energy.

In such cases, the more stable configuration may have a periodic distortion of the lattice; thus, there is a modulation of the electronic charge density, which gives rise to a CDW. The wave vector turns out to be $Q = 2k_\mathrm{F}$, where k_F is the Fermi wave vector, and the electronic density becomes

$$\delta\rho = \rho_0 \cos(2k_\mathrm{F}x + \phi).$$

Due to this periodic lattice distortion, a gap at the Fermi level appears, and the conduction electrons lower their kinetic energy. At high temperatures, thermal excitation of electrons across the band gap makes the normal metallic state stable. When the temperature is sufficiently low, a second-order phase transition (known as the Peierls transition) takes place, and a CDW is formed.

In 1954, Herbert Fröhlich suggested that if Q was not commensurate with the lattice constant, the CDW energy would be independent of the phase ϕ, and thus, an electrical current would appear under any electric field, independent of its intensity. For a while, this phenomenon was speculated to be a possible origin of superconductivity. Interestingly, the interplay and relationship among CDWs, superconductivity, and spin density waves is still a field of study (Gabovich et al., 2002).

If the translational invariance of ϕ is disrupted, there is a phase for which the CDW energy is the lowest, and there is also a minimum threshold field to overcome this energy reduction and to initiate the conduction. A possible cause of the invariance break could be that the CDW is commensurate with the lattice. Although this case is unusual and mostly of theoretical interest, such a CDW (with a period quasimultiple of the lattice constant) may contain solitons in the form of constant phase zones, separated by abrupt change areas. This soliton behavior is modeled by sine-Gordon-like equations. However, empirical evidence suggests that the origin of the pinning of the CDW to the lattice and the appearance of a threshold field stem from impurities.

Experimental evidence of CDW behavior became available in the 1970s, and nowadays several materials show CDW behavior, both inorganic, like $NbSe_3$, NbS_3, or $K_{0.3}MoO_3$ ("blue bronze"), and organic, like $(fluoranthene)_2PF_6$. Evidence of this kind of transition is detected through magnitudes affected by the gap at the Fermi level, including magnetic susceptibility, resistivity, thermoelectric power, scattering experiments where the CDW wave vector manifests itself, and more recently, by means of scanning tunneling microscope images.

Conductivity is among the more interesting properties of CDWs. The dielectric constants for these materials are high, and conductivity suffers an abrupt change from insulating to metallic values of orders of magnitude. The Hall and thermoelectric effects suggest that their conductivity consists of an ohmic linear term and a CDW nonlinear term.

The response of the CDW to a field higher than the threshold value is twofold. First, there appears a high-frequency coherent current, or narrow band noise, which seems to be due to the displacement of the CDW over the pinning potential. Second, a low-frequency broad band noise, incoherent response, is also detected. It is also found that the conductivity saturates for high values of the external field, and it seems that this is due to electrons leaving the CDW region or due to the elimination of $2k_F$ phonons.

When an a.c. field is present, the CDW exhibits a strong dependence on the field frequency, and its conductivity, σ_{ac}, also saturates for high frequencies. There also appear an induced conductivity, σ_{dc}, which increases when V_{ac} increases, and some interference phenomena between the narrow band noise and the a.c. The external field $E_{ac} \cos \omega_e t$ causes oscillations of the current at frequency ω_e, and if there is also a d.c. field, E_{dc}, there are oscillations at frequency ω_i corresponding to the narrow band noise. These two frequencies may interact to produce mode-locking phenomena when they are commensurate $(n\omega_i = m\omega_e)$.

In CDW systems, this locking shows up in the step structure of the differential resistance as a function of the d.c. field. As the external d.c. changes, the nonlinearity of the system keeps the relation between both frequencies constant, ω_i/ω_e, over a finite interval of the external parameter, corresponding to the intervals where $\mathrm{d}V/\mathrm{d}I$ is constant. When the external parameter moves far from the locking region, the system undergoes a transition to an unlocked state, which is quasi-periodic, with two incommensurate frequencies. The interference between the internal frequency ω_i and the external one ω_e is the origin of the coherent and incoherent responses of the system. Usually, the low-frequency region of the power spectra consists of a broad band noise, while the narrow components show up at high frequencies as narrow band noise. A systematic elimination of the broad band noise when the CDW entered mode locking (Sherwin & Zettl, 1985) and a reinforcement of this noise in the unlocked regime have been observed. The interplay between the internal frequency and the external one may give rise to chaotic behavior, with a period-doubling route to chaos.

Studied in the context of self-organized criticality, CDWs are an example of systems that reorganize themselves near the edge of stability, and any small change in the external electrical field gives rise to a drastic change in the response of the CDW (high increase of conductivity).

Although several models have been proposed to explain CDW behavior, none is completely satisfactory. The classical model considers the CDW as a rigid carrier, without any internal degree of freedom, using the forced oscillator equations with some analogy to the Josephson junctions. The tunneling model focuses on the gap in the excitation spectrum of the CDW, explaining the nonlinear conductivity and the scale relationship between $\sigma_{ac}(\omega)$ and $\sigma_{dc}(E)$. However, these models do not explain the interference phenomena between the narrow band noise and the external field frequency ω.

There are other models that consider the internal degrees of freedom of the CDW (segmenting the CDW either through a hydrodynamical description or the Kelmm–Schrieffer model), but none of them completely explains the phenomenology observed in a CDW. Another interesting model is the Fukuyama–Lee–Rice model, which treats the CDW as a classical extended elastic medium, interacting with impurities and an electric field. Discrete versions of these models have also been used. For the commensurate case, Frenkel–Kontorova and soliton models (such as the sine-Gordon equation) have been used.

Several applications have been suggested for these materials, including tunable condensers, optical detectors, memory devices, and switches, among others.

LUIS VÁZQUEZ, P. PASCUAL, AND S. JIMÉNEZ

See also **Coupled oscillators; Frenkel–Kontorova model; Polarons; Sine-Gordon equation; Superconductivity**

Further Reading

Brown, S. & Grüner, G. 1994. Charge and spin density waves. *Scientific American*, April 1994: 50–56

Gabovich A.M., Voitenko, A.I. & Ausloos, M. 2002. Charge-and spin-density waves in existing superconductors: competition between Cooper pairing and Peierls or excitonic instabilities. *Physics Reports*, 367: 583–709

Grüner, G. & Zettl, A. 1985. Charge density wave conduction: a novel collective transport phenomenon in solids. *Physics Reports*, 119: 117–232

Grüner, G. 1988. The dynamics of charge-density waves. *Reviews of Modern Physics*, 60: 1129–1181

Sherwin, M. & Zettl A. 1985. Complete charge density-wave mode locking and freeze-out of fluctuations in $NbSe_3$. *Physical Review B*, 32: 5536–5539

Thorne, R.E. 1996. Charge-density-wave conductors. *Physics Today*, May 1996: 42–47

CHEMICAL KINETICS

Chemical kinetics is a well-defined field of physical chemistry that arose in the 1850s as a complement to the investigation of chemical equilibria. The question of how fast a reactive mixture in a closed vessel reaches equilibrium gave rise to the concept of reaction velocity. The mass action law, enunciated by Cato Guldberg and Peter Waage in 1863, provided a quantitative expression of the velocity of an elementary reaction step in a homogeneous medium in terms of the concentrations or the mole fractions of the reactants involved, and a parameter known as the rate constant. Chemical kinetics is intrinsically nonlinear, since the law of mass action features products of concentrations of the species involved.

Early Developments

Evidence that chemical reactions can generate complex behavior was reported in the early days of chemical kinetics (Pacault & Perraud, 1997). In 1899, Wilhelm Ostwald discovered that in a reaction involving chromium in concentrated acid solution, the release of hydrogen gas was periodic. In 1906, Robert Luther observed propagating chemical reaction fronts in connection with the catalytic hydrolysis of alkyl sulfates. These studies remained isolated for a long time. Possible origins, including the systematic study of reaction mechanisms, were hardly touched upon and there was little or no modeling effort. Not surprisingly, therefore, they came to be regarded by the scientific community as curiosities or even as artifices. On the theoretical side in the 1920s, Alfred Lotka devised a model formally deriving from chemical kinetics and giving rise to sustained oscillations. As the model did not apply to any known chemical system, it was discarded by chemists but was far better received in population dynamics where it played a seminal role. This connection was further enforced in 1926 when Vito Volterra advanced an explanation of ecological cycles in connection with predator-prey systems, using ideas similar to those of Lotka.

The Phenomenology of Nonlinear Chemical Kinetics

Nonlinear chemical kinetics in its modern form owes much to the Belousov–Zhabotinsky reaction (Zhabotinsky, 1964; Field et al., 1972) dealing with the oxidation of a weak acid by bromate in the presence of a metal ion redox catalyst. In addition to the possibility of displaying long records of oscillatory behavior in batch (closed reactor), this reaction gave rise for the first time to a thorough mechanistic study which highlighted the important role of feedback in the onset of complex behavior, in addition to nonlinearity.

Nonlinear phenomena in chemical kinetics have been observed on whole classes of systems giving rise to a large variety of complex behaviors, as reviewed in a Faraday discussion held in 2001. Quantitative phase diagrams have been constructed separating different behavioral modes as some key parameters are varied (Gray & Scott, 1990; Epstein & Pojman, 1998).

Open Well-Stirred Reactors

Simple periodic, multi-periodic, and chaotic oscillations are observed as the residence time of reactants (inversely proportional to their pumping rate into the reactor) is varied. A second type of phenomenon is multistability, the possibility of exhibiting more than one simultaneously stable steady-state concentration level. A third type of phenomenon is excitability whereby, once perturbed, a system performs an extended excursion resembling a single pulse of an aborted oscillation, before settling back to its stable steady state. Finally, an interesting phenomenology pertains to combustion reactions, where the dependence of the rate constant on temperature is the source of a universal (reaction mechanism-independent) positive feedback.

Open Unstirred Reactors

In the absence of stirring, chemical dynamics coexists with transport phenomena. This can give rise to the generic phenomenon of propagating wave fronts. In a bistable system, a typical front may join the two stable states, with one of them progressing at the expense of the other. In two- and three-dimensional reactors undergoing excitable or oscillatory dynamics, the fronts can take the more spectacular form of circles (target patterns), rotating spirals, and scrolls. An exciting form of spatial self-organization is spontaneous symmetry-breaking leading to sustained steady-state patterns, anticipated by Alan Turing in 1952 and realized experimentally by Patrick De Kepper and coworkers; see Figure 1 (Turing, 1952; De Kepper et al., 2000).

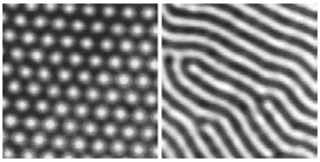

Figure 1. Stationary concentration patterns arising in the chlorite–iodide–malonic acid reaction beyond a symmetry-breaking instability (courtesy of P. De Kepper).

Heterogeneous Systems

Since the late 1980s, a series of novel developments has been initiated following the encounter of nonlinear chemical kinetics with surface science as it is manifested, for instance, in heterogeneous catalysis. Complex behavior in all the above-mentioned forms is observed. Furthermore, the development of sophisticated monitoring techniques, such as field ion microscopy, opens the perspective of monitoring chemical dynamics at the nanoscale level (Hildebrand et al., 1999).

Theoretical Developments: Dynamical Systems and Nonlinear Chemical Kinetics

The essence of nonlinear chemical kinetics is captured by the reaction-diffusion equations (Nicolis & Prigogine, 1977)

$$\frac{\partial c_i}{\partial t} = v_i(\{c_j\}, k_\alpha, \Delta H_\alpha, \cdots) + D_i \nabla^2 c_i, \qquad (1)$$

where $c_i (i = 1, \ldots, n)$ denotes the concentrations or the temperature, k_α and ΔH_α the rate constants and heats of reaction of the steps involved, and D_i the mass or heat diffusivity coefficients. The rate function v_i accounts for the nonlinearities and feedbacks, whereas the contribution of transport processes is linear.

The reaction-diffusion equations (1) exhibit nonlinearity in its simplest expression, as a property arising from intrinsic and local cooperative events. Because of this, complex behavior may arise in the absence of spatial degrees of freedom and persist even when few variables are present. In thermodynamic language, reactions are purely dissipative processes, whereas in nonlinear mechanics, inertia plays a very important role in the onset of complex behavior. Understanding how purely dissipative systems can come to terms with the restrictions imposed by the laws of thermodynamics and statistical mechanics has stimulated several fundamental developments (Glansdorff & Prigogine, 1971; Nicolis & Prigogine, 1977). It has also led to the design of canonical models, such as the Brusselator, that are being used with success to test ideas and to assess the limits of validity of approximations.

The intrinsic parameters k and D in Equations (1) has dimensions of $[\text{time}]^{-1}$ and $[(\text{length})^2/\text{time}]$, respectively. It follows that a reaction-diffusion system possesses intrinsic time (k^{-1}) and space ($(D/k)^{1/2}$) scales. This places nonlinear kinetics at the forefront for understanding the origin of endogenous rhythmic and patterning phenomena as observed, in particular, in biology and in materials science. In thermodynamic equilibrium, these intrinsic time and length scales remain dormant, owing to detailed balance. Nonequilibrium allows for the excitation and eventual stabilization of finite amplitude disturbances bearing these characteristic scales.

Equations (1) form the basis of interpretation of the experiments surveyed above. They also constitute some of the earliest and most widely used models of bifurcation and chaos theories. The classical tools used in their analysis are stability theory and the reduction to normal form (amplitude) equations using perturbation techniques and/or symmetry arguments, complemented by interactive numerical simulations (Nicolis, 1995). A most interesting development is the prediction of an impressive variety of intrinsically generated spatial and spatiotemporal patterns, including spatiotemporal chaos, when two or more mechanisms of instability are interfering.

Nonlinear Chemical Kinetics in the Life Sciences

Research in nonlinear chemical kinetics has led to a semi-quantitative interpretation of a wide spectrum of dynamical behaviors in biochemistry (Goldbeter, 1996). This has been possible thanks to the development of models in which the involvement of cooperative enzymes in some key steps provides the principal source of nonlinearity and feedback. Glycolytic oscillations, calcium oscillations and waves, the cell division cycle, cAMP-induced aggregation in amoebae, and synchronization in cell populations are among the main achievements of this effort that helped to identify the principal mechanisms behind the observed behavior.

Nonlinear kinetics has also been a source of inspiration for approaching dynamical phenomena of crucial importance in biology, in which modeling involving a few variables and/or well-established molecular mechanisms is still not available. Immune response, the electrical activity of the brain, embryonic development, cooperative processes such as food recruitment or building activity in social insects, and, last but not least, chemical and biochemical evolution itself (Eigen & Schuster, 1979; *See* **Biological evolution**) have been explored in one way or the other in the light of the concepts and techniques of nonlinear kinetics.

As a closing remark, Equations (1) anticipate a decoupling between the evolution laws of the macroscopic observables and dynamics at the microscopic level, which may actually break down when reactive systems are embedded on low-dimensional supports owing to the generation of anomalous inhomogeneous fluctuations. This leads to interesting synergies among nonlinear chemical kinetics, statistical mechanics, and computational science (Nicolis, 2001).

G. Nicolis

See also **Belousov–Zhabotinsky reaction; Brusselator; Population dynamics; Reaction-diffusion systems; Turing patterns**

Further Reading

De Kepper, P., Dulos, E., Boissonade, J., De Wit, A., Dewel, G. & Borckmans, P. 2000. Reaction–diffusion patterns in confined chemical systems, *Journal of Statistical Physics*, 101: 495–508

Eigen, M. & Schuster, P. 1979. *The Hypercycle: A Principle of Natural Self-organization*, Berlin and New York: Springer

Epstein, I.R. & Pojman, J.A. 1998. *An Introduction to Nonlinear Chemical Dynamics*, Oxford and New York: Oxford University Press

Field, R.J., Körös, E. & Noyes, R. 1972. Oscillations in chemical systems. II. Thorough analysis of temporal oscillation in the bromate–cerium–malonic acid system. *Journal of the American Chemical Society*, 94: 8649–8664

Glansdorff, P. & Prigogine, I. 1971. *Thermodynamic Theory of Structure, Stability and Fluctuations*, London and New York: Wiley

Goldbeter, A. 1996. *Biochemical Oscillations and Cellular Rhythms*, Cambridge and New York: Cambridge University Press

Gray, P. & Scott, S.K. 1990. *Chemical Oscillations and Instabilities*, Oxford: Clarendon Press and New York: Oxford University Press

Hildebrand, M. Kuperman, M., Wio, H., Mikhailov, A.S. & Ertl, G. 1999. Self-organized chemical nanoscale microreactors. *Physical Review Letters*, 83: 1475–1478

Nicolis, G. 1995. *Introduction to Nonlinear Science*, Cambridge and New York: Cambridge University Press

Nicolis, G. 2001. Nonlinear kinetics: at the crossroads of chemistry, physics and life sciences. *Faraday Discussions*, 120: 1–10

Nicolis, G. & Prigogine, I. 1977. *Self-organization in Nonequilibrium Systems*, New York: Wiley

Pacault, A. & Perraud J.-J. 1997. *Rythmes et Formes en Chimie*, Paris: Presses Universitaires de France

Royal Society of Chemistry. 2002. *Nonlinear Chemical Kinetics: Complex Dynamics and Spatio-temporal Patterns*. Faraday Discussion no. 120, London: Royal Society of Chemistry; pp. 1–431

Turing, A.M. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London, Series B*, 237: 37–72

Zhabotinsky, A.M. 1964. Periodic liquid phase oxidation reactions, *Doklady Akademie Nauk* SSSR, 157: 392–395

CHERENKOV RADIATION

Cherenkov radiation is the electromagnetic radiation of a charged particle moving uniformly in a medium with a velocity exceeding the velocity of light (c) in that medium. It was discovered experimentally by Pavel Cherenkov in 1934 and was theoretically explained by Igor Tamm and Il'ja Frank in 1937. (In 1958, Cherenkov, Tamm, and Frank were awarded a Nobel Prize in physics for this work.) Earlier, Cherenkov radiation was theoretically predicted by Arnol'd Sommerfeld, who, in 1904, solved a formal problem on the radiation of a charged particle moving in vacuum with a velocity $v > c$, and by Oliver Heaviside (Heaviside, 1950) at the end of the 19th century.

At present, all radiation phenomena of waves of any origin, created by a source that moves with a velocity exceeding the phase velocity of the waves, are regarded as Cherenkov radiation. Common examples that can be observed in ordinary life include waves created on a water surface by moving objects, the so-called bow waves according to a theory developed by Lord Kelvin (William Thomson) in the middle of the 19th century, and acoustic shock waves brought about in the atmosphere by a supersonic jet or a rocket, first described by Ernst Mach in 1877.

Essentially, Cherenkov radiation can be understood from the following simple considerations. A source moving steadily with a velocity $\boldsymbol{v}$ and depending on coordinates and time as $f(\boldsymbol{r}-\boldsymbol{v}t)$ can be presented in the form of a Fourier integral $f(\boldsymbol{r},t)=\int f(\boldsymbol{k})\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}-\mathrm{i}\boldsymbol{k}\boldsymbol{v}t}\mathrm{d}^3k$, which suggests that each Fourier harmonic has frequency $\boldsymbol{k}\boldsymbol{v}$. In electrodynamics, the role of the source is played by the distribution of charge density $\rho(\boldsymbol{r}-\boldsymbol{v}t)$ and current $\boldsymbol{j}=\rho\boldsymbol{v}(\boldsymbol{r}-\boldsymbol{v}t)$, and in hydrodynamics by external forces and moving particles.

If a source is to excite the waves in a medium whose wave vector $\boldsymbol{k}$ and frequency ω are related by the dispersion equation $\omega=\omega(\boldsymbol{k})$, a resonance condition must be satisfied, under which the wave frequency $\omega(\boldsymbol{k})$ coincides with the external force frequency $\boldsymbol{k}\boldsymbol{v}$. This equality yields the Cherenkov condition

$$\omega(\boldsymbol{k})=\boldsymbol{k}\cdot\boldsymbol{v}, \tag{1}$$

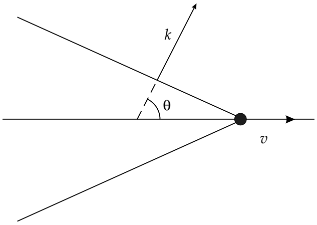

Figure 1. Wave front configuration by Cherenkov radiation, $\boldsymbol{v}$ is the particle velocity, and $\boldsymbol{k}$ is the wave vector of the emitted wave.

which can be fulfilled only if the moving source velocity exceeds the phase velocity of the waves $v > v_{\mathrm{ph}} = \omega(\boldsymbol{k})/k$.

The radiated wave vectors form a characteristic Cherenkov cone, which is similar to the Mach cone in hydrodynamics. Thus, the angle θ between the wave vector of the radiated wave and the direction of particle velocity is determined by the Cherenkov condition of Equation (1), which requires $\cos\theta = v_{\mathrm{ph}}/v$, showing that the frequency distributions of radiation are related to each other. Figure 1 shows the wave front configuration of a wave of frequency ω, emitted by a moving source.

Nowadays, Cherenkov radiation is used in high-energy physics as an important experimental tool (the Cherenkov counter), enabling identification and velocity measurements of fast charged particles, and in electronics for Cherenkov electron oscillators and amplifiers of electromagnetic waves, such as traveling-wave tubes and backward-wave oscillators. Many diverse phenomena can be related to manifestations of Cherenkov radiation, including Landau damping and instabilities in plasma physics, and the Kelvin–Helmholtz instability in hydrodynamics (excitation of waves on a water surface by wind).

Vladislav V. Kurin

See also **Dispersion relations; Shock waves**

Further Reading

Heaviside, O. 1950. *Electromagnetic Theory*, 3 vols, London: The Electrician; reprinted New York: Dover

Jackson, J.D. 1998. *Classical Electrodynamics*, New York: Wiley

Landau, L.D. & Lifshitz, E.M. 1963. *Theoretical Physics*, vol. 6: *Fluid Mechanics*, Oxford: Pergamon Press

Landau, L.D. & Lifshitz, E.M. 1963. *Theoretical Physics*, vol. 8: *Electrodynamics of Continious Media*, Oxford: Pergamon Press

Landau, L.D. & Lifshitz, E.M. 1963. Theoretical Physics, vol. 10: *Physical Kinetics*, Oxford: Pergamon Press

CHI(2) MATERIALS AND SOLITONS

See **Nonlinear optics**

CHIRIKOV MAP

See **Maps**

CHUA'S CIRCUIT

Having witnessed futile attempts at producing chaos in an electrical analog of the Lorenz equations while on a visit to Japan in 1983, Leon Chua was prompted to develop a chaotic electronic circuit. He realized that chaos could be produced in a piecewise-linear circuit if it possessed at least two unstable equilibrium points—one to provide stretching and the other to fold the trajectories. With this insight and using nonlinear circuit theory (Chua et al., 1987), he systematically identified those third-order piecewise-linear circuits containing a single voltage-controlled nonlinear resistor that could produce chaos. Specifying that the V–I characteristic of the nonlinear resistor N_R should be chosen to yield at least two unstable equilibrium points, he invented the circuit shown in Figure 1.

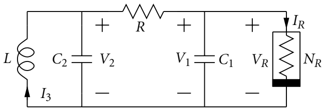

Figure 1. Chua's circuit consists of a linear inductor L, two linear capacitors (C_1 and C_2), a linear resistor R, and a voltage-controlled nonlinear resistor N_R (called a Chua diode).

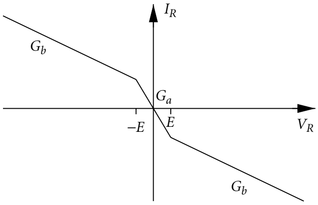

Figure 2. The V–I characteristic of the Chua diode N_R has breakpoints at $\pm E$ and slopes G_a and G_b in the inner and outer regions, respectively.

This circuit is described by three ordinary differential equations

$$\begin{aligned} C_1\frac{\mathrm{d}V_1}{\mathrm{d}t} &= -GV_1 - f(V_1) + GV_2, \\ C_2\frac{\mathrm{d}V_2}{\mathrm{d}t} &= GV_1 - GV_2 + I_3, \\ L\frac{\mathrm{d}I_3}{\mathrm{d}t} &= -V_2, \end{aligned} \tag{1}$$

where $G = 1/R$. Also, $f(\cdot)$ is the V–I characteristic of the nonlinear resistor N_R (known as a Chua diode), which has a piecewise-linear V–I characteristic defined by

$$f(V_R) = G_b V_R + \tfrac{1}{2}(G_a - G_b)(|V_R + E| - |V_R - E|), \tag{2}$$

where $\pm E$ are the breakpoints in the characteristic, and G_a and G_b are the slopes in the inner and outer regions, respectively, as shown in Figure 2.

If the values of the circuit parameters are chosen such that the circuit contains three equilibrium points (two in the outer regions and one at the origin), all of which are unstable with saddle-focus stability, then a homoclinic trajectory can be formed, potentially producing chaos.

Soon after its conception, the rich dynamical behavior of Chua's circuit was confirmed by computer simulation and experiment and in 1986 was proven to exhibit chaos in the sense of Shilnikov (Chua et al., 1986). An intensive effort since then to understand every aspect of the dynamics of this circuit has resulted in its widespread acceptance as a powerful paradigm for learning, understanding, and teaching about nonlinear dynamics and chaos (Madan, 1993; Chua, 1992).

By adding a linear resistor R_0 in series with the inductor, Chua's circuit has been generalized to the Chua oscillator (Chua, 1993), with the last of Equations (1) changing to

$$L\frac{\mathrm{d}I_3}{\mathrm{d}t} = -V_2 - R_0 I_3. \tag{3}$$

Chua's oscillator is canonical in the sense that it is equivalent (topologically conjugate) to a 13-parameter family of three-dimensional ordinary differential equations with odd-symmetric piecewise-linear vector fields. The circuit can exhibit every dynamical behavior known to be possible in a system described by a continuous odd-symmetric three-region piecewise-linear vector field. Unlike the Lorenz or Rössler equations, which have more complex multiplicative nonlinearities, the only nonlinearity in Chua's circuit is a scalar function of one variable. With an appropriate choice of parameters, the circuit can be made to follow the classic period-doubling, intermittency, and torus breakdown routes to chaos; in addition, over 60 different types of strange attractors have been reported in Chua's oscillator.

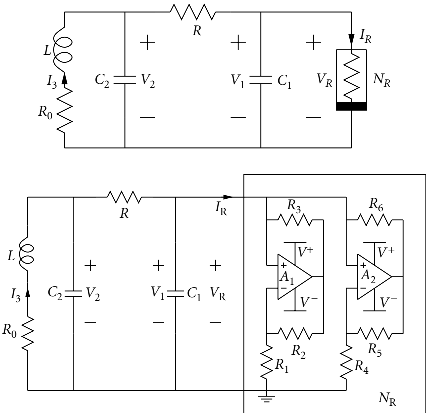

Figure 3. Robust practical implementation of Chua's circuit/oscillator using two op amps and six resistors to realize the Chua diode. In the case of Chua's circuit, R_0 is zero; in Chua's oscillator, R_0 may assume negative or positive values. Component values for Chua's circuit are listed in Table 1.

Chua's circuit oscillator can be realized in a variety of ways by using standard or custom-made electronic components. Since all of the linear elements (capacitors, resistors, and inductor) are readily available as two-terminal devices, only the nonlinear diode must be synthesized using a combination of standard electronic components. The most robust practical realization of Chua's circuit/oscillator, shown in Figure 3, uses two operational amplifiers (op-amps) and six resistors to implement the nonlinear diode (Kennedy, 1992).

The op-amp subcircuit consisting of A_1, A_2, and R_1 through R_6 functions as a Chua diode with V–I

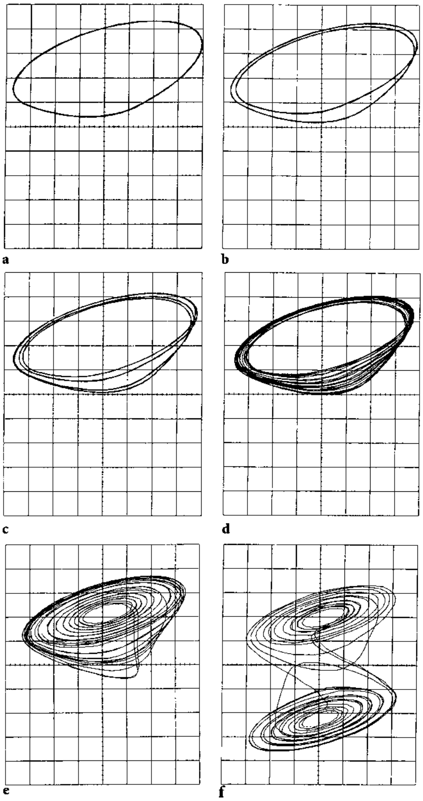

Figure 4. Typical experimental bifurcation sequence in Chua's circuit (component values as in Table 1) recorded using a Hitachi VC-6025 Digital Storage Oscilloscope. Horizontal axis V_2 200 mV/div; vertical axis V_1 1 V/div. (a) $R = 1.83$ kΩ, period 1; (b) $R = 1.82$ kΩ, period 2; (c) $R = 1.81$ kΩ, period 4; (d) $R = 1.80$ kΩ, Spiral attractor; (e) $R = 1.76$ kΩ, Spiral attractor; (f) $R = 1.73$ kΩ, double-scroll attractor [reproduced from Kennedy (1993)].

Element	Description	Value	Tolerance (%)
A_1	Op-amp ($\frac{1}{2}$ AD712 or TL082)		
A_2	Op-amp ($\frac{1}{2}$ AD712 or TL082)		
C_1	Capacitor	10 nF	± 1
C_2	Capacitor	100 nF	± 1
R	Potentiometer	2 kΩ	
R_1	$\frac{1}{4}$W Resistor	3.3 kΩ	± 1
R_2	$\frac{1}{4}$W Resistor	22 kΩ	± 1
R_3	$\frac{1}{4}$W Resistor	22 kΩ	± 1
R_4	$\frac{1}{4}$W Resistor	2.2 kΩ	± 1
R_5	$\frac{1}{4}$W Resistor	220 Ω	± 1
R_6	$\frac{1}{4}$W Resistor	220 Ω	± 1
L	Inductor (TOKO type 10RB)	18 mH	± 5

Table 1. Component list for Chua's circuit.

characteristic as shown in Figure 2. Using two 9 V power supplies for the analog devices AD712, op-amps set their saturation voltages at approximately ± 8.3 V, yielding breakpoints $E \approx 1$V. With $R_2 = R_3$ and $R_5 = R_6$, the V–I characteristic of the Chua diode is defined by $G_a = -1/R_1 - 1/R_4 = -25/33$ mS and $G_b = 1/R_3 - 1/R_4 = -9/22$ mS. Note that the value of the resistance R_0 is ideally zero in Chua's circuit. In practice, the real inductor L has a small parasitic resistance that can be modeled by R_0; the TOKO-type 10RB inductor is preferred because it has a sufficiently low parasitic resistance R_0 for this application.

By reducing the value of the variable resistor R from 2000 Ω to zero, with all other components as in Table 1, the circuit exhibits a Hopf bifurcation from dc equilibrium, a sequence of period-doubling bifurcations to a spiral attractor, periodic windows, a double-scroll strange attractor, and a boundary crisis (Kennedy, 1995).

Although the diode characteristic in Equation (2) is piecewise-linear, qualitatively similar behavior is observed when a smooth nonlinearity such as a cubic is used instead. The piecewise-linear nonlinearity is more convenient for circuit realization, while the smooth nonlinearity is more appropriate for bifurcation analysis.

Michael Peter Kennedy

See also **Attractors; Bifurcations; Chaotic dynamics; Hopf bifurcation; Horseshoes and hyperbolicity in dynamical systems; Period doubling; Routes to chaos**

Further Reading

Chua, L.O. 1992. The genesis of Chua's circuit. *Archiv für Elektronik und Übertragungstechnik*, 46(4): 250–257

Chua, L.O. 1993. Global unfolding of Chua's circuit. *IEICE Transactions Fundamentals*, E76A(5): 704–734

Chua, L.O., Desoer, C.A. & Kuh, E.S. 1987. *Linear and Nonlinear Circuits*, New York: McGraw-Hill

Chua, L.O., Komuro, M. & Matsumoto, T. 1986. The double scroll family—Parts I and II. *IEEE Transactions on Circuits and Systems*, 33(11): 1073–1118

Kennedy, M.P. 1992. Robust op amp realization of Chua's circuit. *Frequenz*, 46(3–4): 66–80

Kennedy, M.P. 1993. Three steps to chaos—Parts I and II. *IEEE Transactions on Circuits and Systems—Part I*, 40(10): 640–674

Kennedy, M.P. 1995. Experimental chaos from autonomous electronic circuits. *Philosophical Transactions of the Royal Society London* A, 353(1701): 13–32

Madan, R.N. 1993. *Chua's Circuit: A Paradigm for Chaos*, Singapore: World Scientific

CIRCLE MAP

See **Denjoy theory**

CLEAR AIR TURBULENCE

Description

In 1966, the National Committee for Clear Air Turbulence officially defined *clear air turbulence* (CAT) as "all turbulence in the free atmosphere of interest in aerospace operations that is not in or adjacent to visible convective activity (this includes turbulence found in cirrus clouds not in or adjacent to visible convective activity)." FAA Advisory Circular AC 00-30B (1997) has simplified this somewhat to "turbulence encountered outside of convective clouds." Thus, CAT is considered to mean turbulence in the clear air, not in or near convective clouds, usually at upper levels of the atmosphere (above 6 km).

CAT was first observed by high-flying fighter aircraft in the mid-to-late 1940s. It was expected that turbulence encounters would be rare at high levels due to the lack of clouds at upper levels. However, it was soon discovered that turbulence encounters in clear air were not only frequent but sometimes quite severe. Since then, CAT encounters have become a significant problem for commercial aircraft flying at cruising altitudes (18,000–45,000 ft above the mean sea level). In fact, various reviews of National Transportation Safety Board (NTSB) reports indicate that in the U.S., turbulence encounters account for approximately 65% of all weather-related accidents or incidents for commercial aircraft; probably more than half of these are due to CAT. Although most turbulence encounters are generally just an annoyance to passengers and crew, on average there are about eight commercial aircraft turbulence-related incidents per year that are significant enough to be reported to the NTSB, accounting for 10 serious and 32 minor injuries. Fortunately, fatalities and substantial damage to the aircraft structure are rare but can occur, as shown in Figure 1.

It should be noted that only a certain range of frequencies or wavelengths of turbulent eddies is felt by aircraft as bumpiness. For most commercial aircraft, this wavelength is anywhere from about 10 m to 1 km.

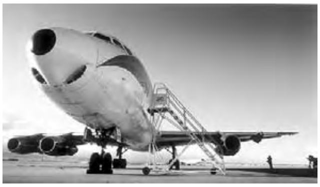

Figure 1. Damage sustained to a DC-8 cargo aircraft in an encounter with CAT on 9 December 1992 at 31,000 ft over the Rocky Mountains near Evergreen, Colorado. Note the loss of left outboard engine and approximately 6 m of the wing. (Photo by Kent Meiries.)

Shorter wavelengths are integrated out over the aircraft structure and longer wavelengths are felt as "waves" and do not generally have vertical accelerations large enough to be felt as "bumps."

CAT can be measured quantitatively with instrumented research aircraft or remotely by instruments such as clear air radar and lidar, but by far the majority of measurements are through pilot reports (PIREPs) of turbulent encounters. PIREPs usually report turbulence on intensity scales of smooth, light, moderate, severe, or extreme. Although definitions of these categories are provided in terms of normal accelerations or airspeed fluctuations, there is still an amount of subjectivity associated with these reports. The pilot reporting system is fairly successful in warning other aircraft of turbulence regions encountered, but to use PIREPs to deduce CAT climatology is difficult, since they are biased by air traffic patterns, non-uniform reporting practices, and the tendency to avoid known turbulence areas. One way to reduce these biases is to examine the ratio of moderate or severe PIREPs to total reports in the three-dimensional airspace averaged over many years of reports. This has been done over the continental U.S. (Sharman et al., 2002). The distribution of this ratio of moderate or greater (MOG) severity PIREPs to total PIREPs by altitude for both CAT and in-cloud encounters is shown in Figure 2. Note that above about 8 km, the majority of reports are in clear air. Similar analyses show the occurrence of CAT to be about twice as frequent in winter as in summer. CAT encounters also tend to be more frequent and more severe over mountainous regions, for example, the Colorado Rockies.

One characteristic of CAT is its patchiness in both time and space. These patches tend to be relatively thin compared with their length (the median thickness is about 500 m, whereas the median horizontal dimension is about 50 km); the median duration is about 6 hours (Vinnichenko et al., 1980). Within a patch, the turbulence may be continuous or may occur in discrete bursts that may be very severe but very narrow (1–2 km).

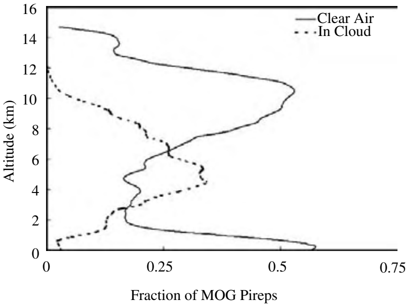

Figure 2. Vertical distribution of the fraction of moderate or greater-intensity (MOG) turbulence pilot reports taken over a two-year period. Solid lines indicate reports in clear air, and dashed lines indicate reports in cloud.

Relation to Kelvin–Helmholtz Instability

From Figure 2, it can be seen that the altitude of maximum occurrence of CAT is at upper levels near the tropopause and jet stream levels. This relation has been known since the 1950s. For example, Bannon (1952) noted that severe CAT tended to occur above and below the jet stream core on the low-pressure side. These areas tend to have large values of the vertical shear of the horizontal wind, and this led to the hypothesis (e.g., Dutton & Panofsky, 1970) that, at least in some cases, CAT may be related to Kelvin–Helmholtz (KH) instability (KHI, *See* **Kelvin–Helmholtz instability**). The KHI process occurs in stably stratified shear flows when dynamic instabilities due to wind shear exceed the restoring forces due to stability. KH waves (also known as "billows") are, in fact, commonly observed in the atmosphere near the top of clouds, where the KH distortions become visible (Figure 3).

Further, the KHI connection to CAT has been verified on occasion by simultaneous measurements of KH billows by high-powered radar and aircraft measurements of turbulence (Browning et al., 1970). Although the figure shows a KH wave train at an instant in time, the KHI process is an evolutionary one, where waves develop, amplify, roll up, and break down into turbulent patches. The resultant turbulent mixing will eventually destroy the wave structure and mix out the shear, and density distributions that created it, but if larger scale processes continue to reinforce the shears, the entire process may reinitiate.

The names associated with KHI derive from the early works of Hermann von Helmholtz (1868), who realized the destabilizing effects of shear, and later of Lord Kelvin (William Thomson) (1871), who posed and solved the instability problem mathematically. Richardson (1920), using simple energy considerations, deduced that a sufficient condition for

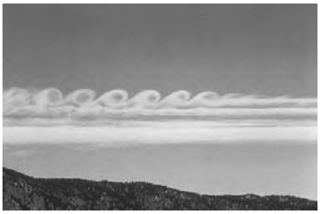

Figure 3. An example of Kelvin–Helmholtz billows observed in the presence of clouds. © 2003 University Corporation for Atmospheric Research.

stability of disturbances in shear flow occurs when the restoring force of stability (as measured by the buoyancy frequency N) is greater than the destabilizing force of the mean horizontal velocity shear (dU/dz) in the vertical (z) direction. Thus, when the ratio $N^2/(dU/dz)^2$ is greater than unity, the flow should be stable. In honor of Richardson's insight, it has become common to refer to this ratio as the Richardson number (Ri). The linear problem for various stratified shear flow configurations is well reviewed in the texts by Chandrasekhar (1961, Chapter 11) and by Drazin and Reid (1981). The sufficient condition for stability to linear two-dimensional disturbances is that $Ri > 0.25$; $Ri < 0.25$ is necessary but not sufficient for instability. More recently, Abarbanel et al. (1984), using the method of Arnol'd, were able to show that the necessary and sufficient condition for Liapunov stability to three-dimensional nonlinear disturbances is $Ri >$ unity, in agreement with Richardson's deduction.

Through these theoretical studies, laboratory experiments (e.g., Thorpe, 1987), and more recently, very high-resolution numerical simulations (e.g., Werne & Fritts, 1999), considerable progress has been made in understanding the intricacies of KHI. One (probably common) method in which KHI is initiated in the atmosphere is through longer wavelength gravity-wave-induced perturbations that lead to local reductions (e.g., in the crest of the wave) in Ri to a value small enough to initiate instability. Gravity waves are ubiquitous in the atmosphere and can be generated in a variety of ways, for example, by flow over mountains or by strong updrafts and downdrafts in convective storms. However, the processes by which KHI may lead to turbulent breakdowns within three-dimensional transient gravity waves is not yet completely understood. It should be mentioned that gravity wave breakdown into turbulent patches may also occur through other mechanisms besides KHI. Examples include convective overturning in large-amplitude waves or nonlinear wave–wave interactions (see the reviews by Wurtele et al. (1996) and Staquet & Sommeria (2002) for discussions of some of these effects). Further, other instability mechanisms besides KHI may lead to CAT, for example, inertial instability or critical-level instability. Thus, the processes by which CAT may be generated at any given time and place are complex, involving many different sources, making its forecasting quite difficult.

One new promising avenue of research is the use of high-resolution numerical simulations of the atmosphere to reconstruct the atmospheric processes that led to particularly severe encounters of CAT (e.g., Clark et al., 2000; Lane et al., 2003). These types of studies have only recently become possible with advances in computing capabilities that allow model runs to contain both the large-scale processes that create conditions conducive to turbulence and the smaller scales that may affect aircraft. Further studies such as these, along with continued theoretical and numerical studies and field measurement campaigns, should lead to a better understanding of CAT genesis and evolution processes.

Until this understanding is available, forecasting of CAT must be accomplished by empirical means. This is done by forecasting various large-scale atmospheric conditions that are known through experience to be related to CAT. Until recently, these diagnostics for likely regions of turbulence had to be performed by laborious weather map analyses of jet streams and upper-level fronts. Nowadays, these diagnostics can be computed from the output of routine numerical weather prediction forecast models. However, the reliability of these turbulence diagnostics is highly variable, and at the moment it seems that better success may be achieved by combining the various diagnostics within an artificial intelligence framework (Tebaldi et al., 2002).

Robert Sharman

See also **Kelvin–Helmholtz instability; Turbulence**

Further Reading

Abarbanel, H.D.I., Holm, D.D., Marsden, J.E. & Ratiu, T. 1984. Richardson number criterion for the nonlinear stability of three-dimensional stratified flow. *Physical Review Letters*, 52: 2352–2355

Bannon, J.K. 1952. Weather systems associated with some occasions of severe turbulence at high altitude. *Meteorological Magazine*, 81: 97–101

Browning, K.A., Watkins, C.D., Starr, J.R. & McPherson, A. 1970. Simultaneous measurements of clear air turbulence at the tropopause by high-power radar and instrumented aircraft. *Nature*, 228: 1065–1067

Chandrasekhar, S. 1961. *Hydrodynamic and Hydromagnetic Stability*, Oxford: Clarendon Press

Clark, T.L., Hall, W.D., Kerr, R.M., Middleton, D., Radke, L., Ralph, F.M., Nieman, P.J. & Levinson, D. 2000. Origins of aircraft-damaging clear-air turbulence during the 9 December 1992 Colorado downslope windstorm: numerical simulations and comparison with observations. *Journal of the Atmospheric Sciences*, 57: 1105–1131

Drazin, P.G. & Reid, W.H. 1981. *Hydrodynamic Stability*, Cambridge: Cambridge University Press

Dutton, J.A. & Panofsky, H.A. 1970. Clear air turbulence: a mystery may be unfolding. *Science*, 167: 937–944

von Helmholtz, H. 1868. On discontinuous movements of fluids. *Philosophical Magazine*, 36: 337–346 (originally published in German, 1862)

Lane, T.P., Sharman R.D., Clark T.L. & Hsu, H.-M. 2003. An investigation of turbulence generation mechanisms above deep convection. *Journal of the Atmospheric Sciences*, 60: 1297–1321

Kelvin, Lord. 1871. Hydrokinetic solutions and observations. *Philosophical Magazine*, 42: 362–377

Richardson, L.F. 1920. The supply of energy from and to atmospheric eddies. *Proceedings of the Royal Society, London*, A97: 354–373

Sharman, R., Fowler, T.L., Brown, B.G. & Wolff, J. 2002. Climatologies of upper-level turbulence over the continental U. S. and oceans. Preprints, 10th Conference on Aviation, Range, and Aerospace Meteorology, Portland OR: American Meteorological Society, J29–J32

Staquet, C. & Sommeria, J. 2002. Internal gravity waves: from instabilities to turbulence. *Annual Reviews of Fluid Mechanics*, 34: 559–593

Tebaldi, C., Nychka D., Brown B.G., and Sharman, R. 2002. Flexible discriminant techniques for forecasting clear-air turbulence. *Environmetrics*, 13(8): 859–878

Thorpe, S.A. 1987. Transitional phenomena and the development of turbulence in stratified fluids: a review. *Journal of Geophysical Research*, 92: 5321–5248

Vinnechenko, N.K., Pinus, N.Z., Shmeter, S.M. & Shur, G.N. 1980. *Turbulence in the Free Atmosphere*, 2nd edition, New York: Consultants Bureau

Werne, J. & Fritts, D.C. 1999. Stratified shear turbulence: evolution and statistics. *Geophysical Research Letters*, 26: 439–442

Wurtele, M.G., Sharman, R.D. & Datta, A. 1996. Atmospheric lee waves. *Annual Reviews of Fluid Mechanics*, 28: 429–476

CLUSTER COAGULATION

In 1916, nine years before he was awarded the Nobel prize for his studies of colloidal solutions, the Austro-Hungarian chemist Richard Zsigmondy (1865–1929) brought forth the first model for cluster coagulation. Interpreting the behavior of aqueous solutions of gold colloidal particles, he posited that each cluster of particles is surrounded by a sphere of influence. According to this model, clusters execute independent Brownian motions when their spheres of influence do not overlap. Whenever the spheres of influence of a pair of clusters touch, the clusters instantaneously stick together to form a new cluster. This kind of non-equilibrium kinetics (*See* **Nonequilibrium statistical mechanics**) has proven to be truly ubiquitous: bond formation between polymerization sites; the coalescence of rain drops, smog, smoke, and dust; the aggregation of bacteria into colonies; the formation of planetesimals from submicron dust grains; the coalescence arising in genetic trees; and even the merging of banks to form ever-larger financial institutions are all examples of cluster coagulation.

Cluster coagulation results in a broad distribution of cluster sizes described by $\{n_i(t)\}_{i=1,\ldots,\infty}$, where $n_i(t)$ is the number of clusters of size i present in the system at time t. The size of a cluster is defined as the number of unit clusters that it comprises. The primary goal of coagulation theory is to determine the evolution of $n_i(t)$ for all i.

The most important theory of coagulation was given by the Polish physicist Marian Smoluchowski (1872–1917) (Smoluchowski, 1916, 1917). In 1916, prompted by a request from Zsigmondy to provide a mathematical description of coagulation, Smoluchowski postulated that (1) clusters are randomly distributed in space and this feature persists throughout the coagulation process, (2) only collisions between pairs of clusters are significant, and (3) the number of new clusters of size $i+j$, formed per unit time and unit volume due to collisions of clusters of sizes i and j, is proportional to the product of the concentrations $c_i = n_i/V$ and $c_j = n_j/V$:

$$\frac{\text{number of new clusters}}{V\Delta t} = K_{i,j}c_i c_j. \qquad (1)$$

Here, V is the volume of the coagulating system and $K_{i,j}$ is the collision frequency factor, also called the *coagulation kernel*.

The rate equation describing the evolution of $c_i(t)$ follows from the balance between the total number of clusters of size i created and annihilated as a result of coagulation:

$$\begin{aligned}\dot{c}_i(t) &= \frac{1}{2}\sum_{j=1}^{i-1} K_{i-j,j}c_{i-j}(t)c_j(t)\\ &\quad -\sum_{j=1}^{\infty} K_{i,j}c_i(t)c_j(t), \quad i = 1, \ldots, \infty. \qquad (2)\end{aligned}$$

Here, $\dot{c}_i(t)$ is the time derivative of the concentration $c_i(t)$.

This equation—in fact, the chain of nonlinear ordinary differential equations—is called the Smoluchowski coagulation equation (SCE). It describes the evolution of homogeneous aggregating systems with the distribution $c_i(t)$, provided knowledge of $K_{i,j}$ and an initial distribution $c_i(0)$. The SCE does not depend on the spatial dimension in which the coagulation process is taking place. According to modern terminology, Smoluchowski theory is a mean field (*See* **Phase transitions**) theory of nonequilibrium growth. It neglects fluctuations of the concentrations c_k; that is, it presumes the existence of the thermodynamic limit: $V \to \infty$, $n_k \to \infty$, $n_k/V \to c_k$. For a broad variety of aggregating systems, this proves to be a reasonable assumption. However, the first assumption of Smoluchowski, that correlations in the distribution of the cluster may be disregarded, is not always satisfied. Aggregating systems fulfilling this assumption are called well-mixed. In low-dimensional systems with no "external" mixing mechanism, the cluster collisions are often not able to provide sufficient mixing. This may result in correlation build-up and, therefore, a breakdown of Equation (2).

The last, essential, albeit obvious, condition for the applicability of the SCE is that interactions between aggregating clusters be treated as instantaneous collisions, that is, $\tau_{\text{col}} \ll \tau_{\text{coag}}$, where τ_{col} is the characteristic time scale of collisions and τ_{coag} is the characteristic time scale of coagulation. Thus, similar to the Boltzmann kinetic equation (*See* **Nonequilibrium statistical mechanics**), the SCE describes a slow evolution of the distribution due to fast collisions.

For a continuous distribution $c(t, v)$, SCE takes the form of an integro-differential equation (Müller, 1928):

$$\partial_t c(t, v) = \frac{1}{2}\int_0^v K(u, v-u)c(t, u)c(t, v-u) \\ -c(t, v)\mathrm{d}u \int_0^\infty K(u, v)c(t, u)\,\mathrm{d}u. \quad (3)$$

Here u, v are the physical sizes of clusters.

Although the SCE establishes a firm foundation for our understanding of a great variety of cluster coagulation processes, other models have also been devised. They include the Oort–van de Hulst–Safronov equation, which describes cluster coagulation as a continuous growth process, and various stochastic models, such as Kingman's coalescent and the Marcus–Lushnikov process. In this article, we limit ourselves to the Smoluchowski coagulation theory.

The mathematical structure of SCE can be traced to the master equation for stochastic processes:

$$\dot{P}_i(t) = \sum_j \left(w_{i,j} P_j(t) - w_{j,i} P_i(t)\right). \quad (4)$$

Here, $P_i(t)$ is the probability of finding the system in a state i at time t, and $w_{i,j}$ is the probability of transition from state j to state i per unit time. Under the aforementioned assumptions of Smoluchowski theory, $P_i = c_i$, $w_{i,j} = K_{i,j} P_i$. Thus, we arrive at the probabilistic interpretation of $K_{i,j}$ as the probability of coagulation of a pair of clusters i and j in unit volume per unit time.

Calculating the coagulation kernel for a particular coagulation mechanism is a separate problem that is beyond the scope of this article. However, a few remarks are appropriate. First, due to the aforementioned time-scale separation, calculation of K can be treated as a stationary problem. Second, the probability of cluster collisions will depend on the cluster geometry. Very often, the aggregates prove to be fractals (*See* **Fractals; Pattern formation**) having no characteristic size. The coagulation kernel K will then depend on their fractal dimension D_f. Third, the probability of coagulation of a pair of clusters is a product of the probability of their collision and the sticking efficiency E. The latter is defined as the probability of clusters merging once they have collided. Two practically important limiting cases are distinguished for coagulation of diffusing clusters: diffusion-limited cluster-cluster aggregation (DLCA), when $E \approx 1$, and reaction-limited cluster-cluster aggregation (RLCA), when $E \ll 1$. DLCA and RLCA produce aggregates of different fractal dimensions and have kinetics of different speeds. Fourth, when the coagulation mechanism is scale-free and the aggregates are fractals, the coagulation kernel K should also be scale-free. In mathematical terms this amounts to the following requirement on the function $K(u, v)$:

$$K(\lambda u, \lambda v) = \lambda^\alpha K(u, v) \quad (5)$$

for any real $\lambda > 0$. Such kernels are called homogeneous, and α is called the homogeneity index. Several examples of widely used kernels are listed in Table 1.

Our present knowledge of when solutions of (2) exist and are unique is limited, as the nonlinearity of SCE presents challenging problems for rigorous mathematical analysis. Existence and uniqueness of solutions for all times have been proven for the kernels $K(u, v) \le C(u + v)$, where C is a constant. This result has recently been extended to the kernels $K(u, v) \le r(u)r(v)$, where $r(v) = o(v)$, as $v \to \infty$ (Norris, 1999; Leyvraz, 2003).

The distribution function is significant because any macroscopic property characterizing a given coagulating system can be calculated in the continuous (discrete) case as an integral (sum) over the distribution. From a mathematical point of view, the distribution function is simply a time-dependent measure on the set of all cluster sizes. Therefore, it is natural to look for *weak solutions* to SCE. Weak solutions can be conveniently defined in the continuous case by means of a Laplace transformed SCE, which describes the time evolution of the Laplace transformation of the concentration.

Weak solutions are inverse Laplace transforms of the solutions to this equation. The discrete case can be treated analogously with the help of generating functions. In fact, most of the presently known exact solutions of SCE were obtained by this approach.

Since the total mass of clusters is, apparently, conserved during collisions, the SCE is expected to conserve the first moment of the distribution

$$M_1(t) = \int_0^\infty v c(t, v)\,\mathrm{d}v. \quad (6)$$

It appears as a deceptively simple exercise to prove this by substituting (6) into SCE. The proof, however, hinges on the condition that the infinite sums involved are convergent. Violation of this condition gives rise to the important phenomenon of *gelation*, also known as

Coagulation mechanism	Kernel	Originator (year)
"Mating"	2	Smoluchowski (1916)
Brownian motion	$\frac{(r(u)+r(v))^2}{r(u)r(v)}$; $r(v) \propto v^{1/D_f}$	Smoluchowski (1916)
Isotropic turbulent shear	$(r(u)+r(v))^3$; $r(v) \propto v^{1/D_f}$	Saffman and Turner (1956)
Gravitational coalescence	$(r(u)+r(v))^2 \, \lvert r(u)^2 - r(v)^2 \rvert$; $r(v) \propto v^{1/D_f}$	Findheisen (1939)
Polymerization (RA_∞ model)	uv	Flory (1953)
Polymerization (ARB_∞ model)	$u+v$	

Table 1. Examples of kernels. All kernels are given up to a non-dimensional prefactor. D_f is the fractal dimension of the coagulates.

the gel-sol transition or runaway growth. It was first predicted for the kernel $K(u, v) = uv$, which serves as a model of polymerization (*See* **Polymerization**) in which new links are formed randomly between polymerization sites. In the mean field approximation, this model also describes random graph growth and bond percolation (*See* **Percolation theory**). An exact solution of this problem shows that starting with monodisperse initial conditions, the first moment $M_1(t)$ begins to decay after a finite time, t_c, whereas the second moment $M_2(t)$, measuring the average cluster size, diverges. This behavior corresponds to the formation of an infinite cluster, or gel, in a finite time due to the coagulation kinetics "accelerating" with growing cluster sizes. The sol mass M_1 decreases, as part of it is being lost to the gel. This kind of kinetics has also proved to be a key to the explanation of the rapid growth of Jupiter and planetesimal growth in the terrestrial planets.

It has been shown that $M_1(t) = M_1(0)$ for all times; that is, the system is nongelling, when $K(u, v) \leq C(u + v)$. A wealth of data suggests that this is, in fact, the exact bound separating gelation at finite times from nongelling behavior, although a rigorous proof has not yet been given.

The nonlinear character of SCE complicates mathematical analysis for arbitrary kernels, and the set of exactly solvable kernels is limited. Considerable progress in understanding coagulation kinetics has been achieved for the wide class of homogeneous (or asymptotically homogeneous) kernels by looking for similarity solutions. This can be done in two different ways, which we shall refer to as the self-preservation theory and the self-similarity theory.

The first approach embodies the notion of a single characteristic size in the system, which can be chosen to be equal to an average cluster size, $s(t)$. The asymptotic solution of SCE is then expected to have a self-preserving (scaling) form:

$$c(t, v) = s(t)^{-2}\Phi\left(v/s(t)\right). \quad (7)$$

The self-preserving form has been further studied with the objective of identifying when it will lead to a power-law distribution asymptotically (Leyvraz, 2003). This approach allows one to obtain sensible results for a large variety of problems, including some extensions of SCE, in an almost automatic manner. However, since the scaling hypothesis is postulated, one cannot estimate a priori the accuracy of these solutions. Experimental data on coagulation often display a power-law distribution over some range of cluster sizes.

The second approach deals with a coagulating system maintained at steady state by an external source of monomers. By analogy with the scaling theories for turbulent flows and the theory of critical phenomena, it may be expected that the steady-state distribution at large sizes will "forget" the forcing scale v_0 and, therefore, will evolve to a scale-free form,

$$c(v) = \text{const} \times v^{-\tau}. \quad (8)$$

A careful mathematical analysis has shown that this is indeed the case for a wide range of homogeneous kernels, and that the asymptotic distribution equals

$$c(v) = \left(\frac{E}{\kappa}\right)^{1/2} v^{-(3+\alpha)/2}. \quad (9)$$

Here, E is the total influx of mass into the system due to the forcing and κ is a kernel-dependent constant.

Dmitri O. Pushkin and Hassan Aref

See also **Brownian motion; Dimensional analysis; Fractals; Nonequilibrium statistical mechanics; Pattern formation; Percolation theory; Phase transitions; Polymerization**

Further Reading

Aldous, D.J. 1999. Deterministic and stochastic methods for coalescence (aggregation, coagulation): review of the mean-field theory for probabilists. *Bernoulli*, 5: 3

Drake, R.L. 1972. A general mathematical survey of the coagulation equation. In *Topics in Current Aerosol Research*, edited by G.M. Hidy and J.R.Brock, part 2, Oxford: Pergamon Press

Family, F. & Landau, D.P. (editors). 1984. *Kinetics of Aggregation and Gelation*, Amsterdam and New York: Elsevier Science

Findheisen, W. 1939. Zur Frage der Regentropfenbildung in reinen Wasserwolken. *Meteorologische Zeitschrift*, 56: 365–368

Flory, P. 1953. Principles of Polymer Chemistry. Ithica: Cornell University Press

Friedlander, S.K. 1960. Similarity consideration for the particle-size spectrum of a coagulating, sedimenting aerosol. 17(5): 479

Friedlander, S.K. 2000. *Smoke, Dust, and Haze: Fundamentals of Aerosol Dynamics*, 2nd edition, Oxford and New York: Oxford University Press

Friedlander, S.K. & Wang, C.S. 1966. The self-preserving particle size distribution for coagulation by Brownian motion. *Journal of Colloid and Interface Science*, 22: 126

Galina, H. & Lechowicz, J.B. 1998. Mean-field kinetic modeling of polymerization: the Smoluchowski coagulation equation. *Advances in Polymer Science*, 137: 135–172

Hunt, J.R. 1982. Self-similar particle-size distributions during coagulation: theory and experimental verification. *Journal of Fluid Mechanics*, 122: 169

Leyvraz, F. 2003. Scaling theory and exactly solved models in the kinetics of irreversible aggregation. *Physics Reports*, 383: 95–212

Müller, H. 1928. Zur algemeinen Theorie der raschen Koagulation. *Kolloid-chemische Beihefte*, 27: 223

Norris, J.R. 1999. Uniqueness, non-uniqueness, and a hydrodynamic limit for the stochastic coalescent. *Annals of Applied Probability*, 9: 78

Pushkin, D.O. & Aref, H. 2002. Self-similarity theory of stationary coagulation. *Physics of Fluids*, 14(2): 694

Saffman, P.G. & Turner, J.S. 1956. On the collision of drops in turbulent clouds. *Journal of Fluid Mechanics*, 1: 16–30

Smoluchowski, M.V. 1916. Drei vorträge über diffusion, *Physikalische Zeitschrift*, 17: 557

Smoluchowski, M.V. 1917. Versuch einer mathematischen theorie der Koagulationskinetik kolloider lösungen. *Zeitschrift für Physikalische Chemie*, 92: 129

van Dongen, P.G.J. & Ernst, M.H. 1985. Dynamic scaling in the kinetics of clustering. *Physical Review Letters*, 54: 1396

van Dongen, P.G.J. & Ernst, M.H. 1988. Scaling solution of Smoluchowski coagulation equation. *Journal of Statistical Physics*, 50: 295

CNOIDAL WAVE

See **Elliptic functions**

COHERENCE PHENOMENA

The word *coherence* comes from the Latin *cohaerens*, meaning "being in relation." Thus, coherence phenomena are those displaying a high level of correlation between several objects.

From the physical point of view, it is necessary to distinguish between two types of coherence: state coherence, which characterizes correlations between static properties of the considered objects, and transition coherence, which describes correlated dynamical processes. These types of coherence are two sides of the same coin, and one obtains a better insight from considering them together.

To gain an intuitive idea of these two types of coherence, imagine a group of soldiers all standing at attention, without moving. This corresponds to state coherence. If the soldiers were all in different positions (some standing, some sitting, some lying down), there would be no state coherence between them. Now, imagine well-aligned rows of soldiers in a parade, moving synchronously with respect to each other. This corresponds to transition coherence. Also, if they were to march with different speeds and in different directions, transition coherence would be absent.

Coherence is related to the existence of a kind of order—be it a static order defining the same positions or an ordered motion of a group. Then, the antonym of *coherence* is *chaos*. Thus, state chaos means the absence of any static order among several objects, and transition chaos implies an absolutely disorganized motion of an ensemble of constituents.

The notion of coherence is implicit in the existence of correlation among several objects (enumerated with the index $i = 1, 2, \ldots, N$). Each object, placed in the spatial point $\boldsymbol{r}_i$, at time t, can be associated with a set $\{Q_\alpha(\boldsymbol{r}_i, t\}$ of observable quantities labeled by α. To formalize the definition of state and transition coherence, one may write $Q_i^\alpha = Q_\alpha(\boldsymbol{r}_i, t)$, where Q_i^z corresponds to a state property of an object, while Q_i^x and Q_i^y describe its motion. As an illustration, assume that Q_i^α are spin components. Another example assumes that Q_i^z is the population difference of a resonant atom, while Q_i^x and Q_i^y are its transition dipoles. Instead of considering the latter separately, it is convenient to introduce the complex combinations $Q_i^\pm \equiv Q_i^x \pm \mathrm{i} Q_i^y$. (In general, Q_i^α are not restricted to classical quantities but may be operators.)

If the system is associated with a statistical operator $\hat{\rho}$, then the observable quantities are the statistical averages

$$\langle Q_i^\alpha \rangle \equiv \mathrm{Tr}\hat{\rho}\, Q_i^\alpha , \qquad (1)$$

expressed by means of the trace operation. A convenient way of describing the system features is by introducing dimensionless quantities, normalized to the number of objects N and to the maximal value $Q \equiv \max\langle Q_i^z \rangle$. Then, one may define the state variable

$$s \equiv \frac{1}{QN} \sum_{i=1}^{N} \langle Q_i^z \rangle \qquad (2)$$

and the transition variable

$$u \equiv \frac{1}{QN}\sum_{i=1}^{N}\langle Q_i^-\rangle. \tag{3}$$

One may distinguish two opposite cases, when the individual states of all objects are the same and when they are randomly distributed. These two limiting cases give

$$|s| = \begin{cases} 1, & \text{state coherence,} \\ 0, & \text{state chaos.} \end{cases} \tag{4}$$

Next consider the transition characteristic (3) and collective motion of an ensemble of oscillators. Again, there can be two opposite situations, when the oscillation frequencies of all oscillators, as well as their initial phases, are identical and when these take randomly different values. For the corresponding limiting cases of completely synchronized oscillations and of an absolutely random motion, respectively, one has

$$|u| = \begin{cases} 1, & \text{transition coherence,} \\ 0, & \text{transition chaos.} \end{cases} \tag{5}$$

In the intermediate situation, one may say that there is partial state coherence if $0 < |s| < 1$ and partial transition coherence when $0 < |u| < 1$.

Accepting that coherence is not necessarily total, it is convenient to define qualitative characteristics for partial coherence by introducing correlation functions. Let $Q_\alpha^+(\boldsymbol{r}, t)$ denote the Hermitian conjugation for an operator $Q_\alpha(\boldsymbol{r}, t)$. When $Q_\alpha(\boldsymbol{r}, t)$ is a classical function, Hermitian conjugation means complex conjugation. For any two operators from the set $\{Q_\alpha(\boldsymbol{r}, t)\}$, one may define the correlation function

$$C_{\alpha\beta}(\boldsymbol{r}_1, t_1, \boldsymbol{r}_2, t_2) \equiv \langle Q_\alpha^+(\boldsymbol{r}_1, t_1) Q_\beta(\boldsymbol{r}_2, t_2)\rangle. \tag{6}$$

The function $C_{\alpha\alpha}(\ldots)$ for coinciding operators is called the autocorrelation function. There is also a shifted correlation function

$$B_{\alpha\beta} \equiv \langle Q_\alpha^+ Q_\beta\rangle - \langle Q_\alpha^+\rangle\langle Q_\beta\rangle,$$

where, for brevity, the spatiotemporal variables are not written explicitly. For describing coherent processes, it is convenient to use the normalized correlation function

$$K_{\alpha\beta} \equiv \frac{\langle Q_\alpha^+ Q_\beta\rangle}{(\langle Q_\alpha^+ Q_\alpha\rangle\langle Q_\beta^+ Q_\beta\rangle)^{1/2}}, \tag{7}$$

which is sometimes termed a "coherence function." Functions (6) and (7) can be specified as second-order correlation functions since, in general, it is possible to define higher-order correlation functions, such as the $2p$-order function

$$C_{\alpha_1\ldots\alpha_{2p}} = \langle Q_{\alpha_1}^+ \cdots Q_{\alpha_p}^+ Q_{\alpha_{p+1}} \cdots Q_{\alpha_{2p}}\rangle,$$

which are closely related to reduced density matrices.

Correlations are usually strongest among nearby spatiotemporal points. Thus, function (7) varies in the interval $0 \le |K_{\alpha\beta}| \le 1$, being maximal for the autocorrelation function $|K_{\alpha\alpha}| = 1$ at the coinciding points $\boldsymbol{r}_1 = \boldsymbol{r}_2$, $t_1 = t_2$. When either the spatial or temporal distance between two points increases, correlations diminish; this is named *correlation decay*. At an asymptotically large distance, the correlation function (6) for two local observables displays the property of correlation weakening (correlation decoupling)

$$\langle Q_\alpha^+(\boldsymbol{r}_1, t_1) Q_\beta(\boldsymbol{r}_2, t_2)\rangle \simeq \langle Q_\alpha^+(\boldsymbol{r}_1, t_1)\rangle\langle Q_\beta(\boldsymbol{r}_2, t_2)\rangle, \tag{8}$$

where either $|\boldsymbol{r}_1 - \boldsymbol{r}_2| \to \infty$ or $|t_1 - t_2| \to \infty$. It is important to stress that property (8) holds only for local observables; thus, for operators representing no observable quantities, correlation decoupling generally has no meaning.

Coherence characteristically implies correlations between similar objects, which require the use of autocorrelation functions. To describe coherence decay, it is also necessary to fix a point from which this decay is measured (usually at $\boldsymbol{r} = 0$ and $t = 0$), whereupon coherence decay is studied by considering an autocorrelation function

$$C_\alpha(\boldsymbol{r}, t) \equiv \langle Q_\alpha^+(\boldsymbol{r}, t)\; Q_\alpha(0, 0)\rangle. \tag{9}$$

In many cases, there exists a spatial direction of particular importance, for example, the direction of field propagation. It is natural to associate this special direction with the longitudinal z-axis and the transverse direction with the radial variable $r_\perp$. The characteristic scale of coherence decay in the longitudinal direction is called coherence length l_{coh}, where

$$l_{\text{coh}}^2 \equiv \frac{\int z^2 |C_\alpha(\boldsymbol{r}, t)|^2 \, \mathrm{d}\boldsymbol{r}}{\int |C_\alpha(\boldsymbol{r}, t)|^2 \, \mathrm{d}\boldsymbol{r}}, \tag{10}$$

and the integration is over the entire space volume. Coherence decay in the transverse direction is classified as transverse coherence radius r_{coh}, where

$$r_{\text{coh}}^2 \equiv \frac{\int r_\perp^2 |C_\alpha(\boldsymbol{r}, t)|^2 \, \mathrm{d}\boldsymbol{r}}{\int |C_\alpha(\boldsymbol{r}, t)|^2 \, \mathrm{d}\boldsymbol{r}}. \tag{11}$$

For isotropic systems, one replaces $r_\perp$ by the spherical radius r and obtains a coherence radius from equation (11). It is natural to call $A_{\text{coh}} \equiv \pi r_{\text{coh}}^2$ the coherence area and $V_{\text{coh}} \equiv A_{\text{coh}} l_{\text{coh}}$ the coherence volume. The typical scale of temporal correlation decay is termed the coherence time t_{coh}, where

$$t_{\text{coh}}^2 \equiv \frac{\int_0^\infty t^2 |C_\alpha(\boldsymbol{r}, t)|^2 \, \mathrm{d}t}{\int_0^\infty |C_\alpha(\boldsymbol{r}, t)|^2 \, \mathrm{d}t}. \tag{12}$$

As seen, the coherence length (10) and coherence radius (11) are related to a fixed moment of time, while the coherence time (12) defines the temporal coherence decay at a given spatial point. Equations (10)–(12) all have to do with a particular coherence phenomenon characterized by the correlation function (9).

Phase transitions in equilibrium statistical systems are collective phenomena demonstrating different types of state coherence arising under adiabatically slow variation of thermodynamic or system parameters (temperature, pressure, external fields, and so on). Phase transitions are conventionally specified by means of order parameters, which are defined as statistical averages of operators corresponding to some local observables. The order parameter is assumed to be zero in a disordered phase and nonzero in an ordered phase. For example, the order parameter at Bose–Einstein condensation is the fraction or density of particles in the single-particle ground state. The order parameter for superconducting phase transition is the density of Cooper pairs or the related gap in the excitation spectrum. Superfluidity is characterized by the fraction or density of the superfluid component. For magnetic phase transitions, the order parameter is the average magnetization. Thermodynamic phases can also be classified by order indices.

Let the autocorrelation function (9) be defined for the operator related to an order parameter. Then, for a disordered phase, the coherence length is close to the interparticle distance and the coherence time is roughly the interaction time. But for an ordered phase, the coherence length is comparable with the system size and the coherence time becomes infinite.

Taking account of heterophase fluctuations in the quasiequilibrium picture of phase transitions, there appear mesoscopic coherent structures, with the coherence length being much larger than interparticle distance, but much smaller than the system size. The coherence time of these mesoscopic coherent structures (their lifetime) is much longer than the local equilibrium time, although it may be shorter than the observation time. Such coherent structures are similar to those arising in turbulence.

Electromagnetic coherent radiation by lasers and masers presents a good example of transition coherence. Such radiation processes are accompanied by interference patterns, a phenomenon that is typical of coherent radiation and can be produced by atoms, molecules, nuclei, or other radiating objects. Interference effects caused by light beams are studied in nonlinear optics. But coherent radiation and related interference effects also exist in other ranges of electromagnetic radiation frequencies, including infrared, radio, or gamma regions. Moreover, there exist other types of field radiation, such as acoustic radiation or emission of matter waves formed by Bose-condensed atoms. Registration of interference between a reference beam and that reflected by an object is the basis for holography, which is the method of recording and reproducing wave fields.

The description of interference involves correlation functions. Let $Q_i(t)$ represent a field at time t, produced by a radiator at a spatial point $\boldsymbol{r}_i$. The radiation intensity of a single emitter may be defined as

$$I_i(t) \equiv \langle Q_i^+(t)\, Q_i(t)\rangle, \tag{13}$$

whereupon the radiation intensity for an ensemble of N emitters is

$$I(t) = \sum_{i,j=1}^{N} \langle Q_i^+(t)\, Q_j(t)\rangle. \tag{14}$$

Separating the sums with $i = j$ and with $i \neq j$ yields

$$I(t) = \sum_{i=1}^{N} I_i(t) + \sum_{i\neq j}^{N} \langle Q_i^+(t)\, Q_j(t)\rangle, \tag{15}$$

which shows that intensity (14) is not simply a sum of the intensities (13) of individual emitters but also includes the interference part, expressed through the autocorrelation functions of type (9). The first term in equation (15) is the intensity of incoherent radiation, while the second term corresponds to the intensity of coherent radiation.

V.I. YUKALOV

See also **Bose–Einstein condensation; Chaotic dynamics; Critical phenomena; Ferromagnetism and ferroelectricity; Lasers; Nonequilibrium statistical mechanics; Nonlinear optics; Order parameters; Phase transitions; Spatiotemporal chaos; Spin systems; Structural complexity; Superconductivity; Superfluidity; Turbulence**

Further Reading

Andreev, A.V., Emelyanov, V.I. & Ilinski, Y.A. 1993. *Cooperative Effects in Optics*, Bristol: Institute of Physics Publishing

Benedict, M.G., Ermolaev, A.M., Malyshev, V.A., Sokolov, I.V. & Trifonov, E.D. 1996. *Superradiance: Multiatomic Coherent Emission*, Bristol: Institute of Physics Publishing

Bogolubov, N.N. 1967. *Lectures on Quantum Statistics*, Vol. 1, New York: Gordon and Breach

Bogolubov, N.N. 1970. *Lectures on Quantum Statistics*, Vol. 2, New York: Gordon and Breach

Coleman, A.J. & Yukalov, V.I. 2000. *Reduced Density Matrices*, Berlin: Springer

Klauder, J.R. & Skagerstam, B.S. 1985. *Coherent States*, Singapore: World Scientific

Klauder, J.R. & Sudarshan, E.C.G. 1968. *Fundamentals of Quantum Optics*, New York: Benjamin

Lifshitz, E.M. & Pitaevskii, L.P. 1980. *Statistical Physics: Theory of Condensed State*, Oxford: Pergamon Press

Mandel, L. & Wolf, E. 1995. *Optical Coherence and Quantum Optics*, Cambridge and New York: Cambridge University Press

Nozières, P. & Pines, D. 1990. *Theory of Quantum Liquids: Superfluid Bose Liquids*, Redwood, CA: Addison-Wesley
Perina, J. 1985. *Coherence of Light*, Dordrecht: Reidel
Scott, A.C. 1999. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, Oxford and New York: Oxford University Press
Ter Haar, D. 1977. *Lectures on Selected Topics in Statistical Mechanics*, Oxford: Pergamon Press
Yukalov, V.I. 1991. Phase transitions anad heterophase fluctuations. *Physics Reports*, 208: 395–492
Yukalov, V.I. & Yukalova, E.P. 2000. Cooperative electromagnetic effects. *Physics of Particles and Nuclei*, 31: 561–602

COHERENT EXCITON

See **Excitons**

COHERENT STRUCTURES

See **Emergence**

COLE–HOPF TRANSFORM

See **Burgers equation**

COLLAPSE

See **Development of singularities**

COLLECTIVE COORDINATES

The soliton equations describe a number of important nonlinear physical phenomena. However, in real life, these phenomena are not precisely modeled by say the sine-Gordon (SG) equation, the nonlinear Schrödinger (NLS) equation, or some of the other pure soliton equations. Corrective and often small terms are added to include, for example, inhomogeneities, dissipation, or energy input. The resulting wave phenomena possess modified solitonic features that can be treated approximately starting with a pure soliton solution and then allow the parameters of the soliton solution to vary slowly with time under the influence of the perturbations instead of being constant. Solution parameters that are chosen to vary with time are called *collective coordinates*. They encompass the influence of the perturbations in the pure soliton equations. The advantage of introducing collective coordinates is to reduce a problem with infinitely many degrees of freedom to a problem with a few degrees of freedom (Kivshar & Malomed, 1989; Sánchez & Bishop, 1998).

To illustrate the use of a collective coordinate approach, we shall investigate the NLS equation with the perturbative term εR

$$\mathrm{i}u_t + u_{xx} + 2|u|^2 u = \varepsilon R. \tag{1}$$

Here, $u = u(x, t)$, where x is the spatial coordinate and t is time. Subscripts x and t denote partial derivatives with respect to these variables. The simple single soliton solution of the pure NLS equation ($\varepsilon = 0$) reads

$$u(x, t) = a\,\mathrm{sech}(a\theta)\exp(\mathrm{i}\xi\theta + \mathrm{i}\sigma), \tag{2}$$

where $\theta = x - x_0$. This solution possesses four parameters $(a, \xi, x_0, \sigma) = (y_1, y_2, y_3, y_4)$, which we shall choose as collective coordinates. For weak perturbations, $\varepsilon \ll 1$, we shall assume that the collective coordinates depend slowly on time t due to the influence of εR. In addition, the perturbation leads to a radiation field of small amplitude, which is neglected. A variational approach is employed to determine the time evolution of the collective coordinates, but this is not the only method available. In the framework of a variational approach, the collective coordinates are the generalized coordinates. Although perturbations may destroy the Hamiltonian property of the pure soliton equations, dissipative effects and external nonconservative forces can be accounted for in the variation of a Lagrangian function by introducing generalized forces associated with the generalized coordinates. Below, this is done by introducing a generalized force for each collective coordinate as in classical mechanics.

The unperturbed NLS equation (and its complex conjugate) can be derived from the Lagrangian density (Caputo et al., 1995; Scott, 2003);

$$\begin{aligned}\mathcal{L}(u, u^*, u_t, u_t^*, u_x, u_x^*) = {} & \frac{\mathrm{i}}{2}(u^* u_t - u_t^* u) \\ & -|u_x|^2 + |u|^4. \end{aligned} \tag{3}$$

The total Lagrangian function is

$$L(y_i, \dot{y}_i) = \int_{-\infty}^{\infty} \mathcal{L}\,\mathrm{d}x, \tag{4}$$

where we denote the collective coordinates $y_i(t)$, $i = 1, 2, 3, 4$, and $\dot{y}_i$ is the time derivative of y_i. Together with Equation (1), the variation of total Lagrangian leads to the Euler–Lagrange equations (Caputo et al., 1995; Scott, 2003)

$$\frac{\partial L}{\partial y_i} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{y}_i}\right) = \varepsilon \int_{-\infty}^{\infty} R \frac{\partial u^*}{\partial y_i}\mathrm{d}x + \text{c.c.}, \tag{5}$$

where c.c. stands for complex conjugate of the preceding term on the right-hand side. The inhomogeneous term on the right-hand side is interpreted as the generalized force associated with the collective coordinate $y_i(t)$, which is a key result as we do not rely on a perturbation with a strict Hamiltonian nature. Another important feature is that the above approach provides as many dynamical equations as we choose collective coordinates; thus it is straightforward to determine the generalized forces associated with each collective coordinate.

To illustrate the method, we calculate the total Lagrangian for the NLS equation using the simple single soliton solution in (2)

$$L = \frac{2}{3}a^3 - 2a\xi^2 + 2a\xi\frac{\mathrm{d}x_0}{\mathrm{d}t} - 2a\frac{\mathrm{d}\sigma}{\mathrm{d}t}. \qquad (6)$$

Consider the perturbation

$$\varepsilon R = -\mathrm{i}\Gamma u + \frac{\mathrm{i}g_0 u}{1 + p/p_\mathrm{s}}, \qquad (7)$$

describing light propagation through an optical fiber amplifier with a loss factor Γ, gain g_0, and power $p = \int_{-\infty}^{\infty} |u|^2 \mathrm{d}x$. The constant p_s is a saturation power level that is characteristic for a given fiber amplifier. From Equation (5), the resulting dynamical equations for the collective coordinates are

$$\frac{\mathrm{d}a}{\mathrm{d}t} = -2\Gamma a + \frac{2g_0 a}{1 + 2a^2/p_\mathrm{s}}, \qquad (8)$$

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = 0, \qquad (9)$$

$$\frac{\mathrm{d}x_0}{\mathrm{d}t} = 2\xi, \qquad (10)$$

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = a^2 + \xi^2, \qquad (11)$$

which is a drastic simplification compared to the original perturbed NLS problem. As for the sine-Gordon system, one can define a power balance condition by requiring $\dot{a}(t) = 0$, implying an equilibrium amplitude a_∞ given as $a_\infty = \sqrt{(g_0 - \Gamma)p_\mathrm{s}/2\Gamma}$.

A number of other strategies have been designed to determine slow time variation of collective coordinates in perturbed soliton solutions. These include slow variation of scattering data of the inverse scattering theory for pure solitons (Kivshar & Malomed, 1989; Lamb, 1980), more direct perturbation approaches (McLaughlin & Scott, 1978), and utilizing Hamiltonian structures in cases where the perturbation leads to Hamiltonian systems (Caputo & Flytzanis, 1991). An important result of perturbations is the formation of trailing radiation fields, which are low-amplitude linear waves created as perturbed solitons propagates (McLaughlin & Scott, 1978; Kivshar & Malomed, 1989; Willis, 1997). The solitons lose energy to the radiation field, and the variational approach used here can be extended to include such radiation.

Mads Peter Sørensen

See also **Constants of motion and conservation laws; Energy analysis; Euler–Lagrange equations; Hamiltonian systems; Inverse scattering method or transform; Nonlinear optics; Nonlinear Schrödinger equations; Perturbation theory; Solitons**

Further Reading

Caputo, J.G. & Flytzanis, N. 1991. Kink-antikink collisions in sine-Gordon and ϕ^4 models: problems in the variational approach. *Physical Review* A, 44(10): 6219–6225

Caputo, J.G., Flytzanis, N. & Sørensen, M.P. 1995. The ring laser configuration studied by collective coordinates. *Journal of the Optical Society of America* B, 12(1): 139–145

Kivshar, Y.S. & Malomed, B.A. 1989. Dynamics of solitons in nearly integrable systems. *Reviews of Moderen Physics*, 61: 763–915

Lamb, G.L. 1980. *Elements of Soliton Theory*, New York: Wiley

McLaughlin, D.W. & Scott, A.C. 1978. Perturbation analysis of fluxon dynamics. *Physical Review A*, 18: 1652–1680

Sánchez, A. & Bishop, A.R. 1998. Collective coordinates and length-scale competition in spatially inhomogeneous soliton-bearing equations. *SIAM Review*, 40(3): 579–615

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Willis, C.R. 1997. Spontaneous emission of a continuum sine-Gordon kink in the presence of a spatially periodic potential. *Physical Review E*, 55(5): 6097–6100

COLLISIONS

An interesting consequence of Hamiltonian structures is that there typically exist symmetries and invariances that allow one to generate mobile localized states from standing ones. For instance, the sine-Gordon (SG) equation (Dodd et al., 1982)

$$u_{tt} = u_{xx} - \sin(u), \qquad (1)$$

where the subscripts denote partial derivatives, is invariant under the Lorentz transformation $x \to x' = \gamma(x - vt)$ and $t \to t' = \gamma(t - vx)$. $\gamma = \sqrt{1 - v^2}$. As a result, static kinks and antikinks, corresponding to the two signs of the solution

$$u(x,t) = \pm 4\tan^{-1}(x - x_0), \qquad (2)$$

can be boosted to any subluminal speed $v < 1$, as

$$u(x,t) = \pm 4\tan^{-1}\left[\gamma(x - x_0 - vt)\right]. \qquad (3)$$

Similarly, standing waves of the nonlinear Schrödinger (NLS) equation

$$\mathrm{i}u_t = -u_{xx} - |u|^2 u \qquad (4)$$

can be boosted by Galilean invariance into traveling ones with any speed v of the form (Sulem & Sulem, 1999)

$$u(x,t) = (2\alpha)^{1/2} \mathrm{e}^{\mathrm{i}(v/2)x + \mathrm{i}(\alpha - (v^2/4))t} \times \mathrm{sech}\,(x - x_0 - vt). \qquad (5)$$

Note that localized solutions of dissipative partial differential equations do not typically share such features, because their traveling wave speeds are determined by the dynamics rather than initial conditions. Hereafter, we will focus on Hamiltonian models.

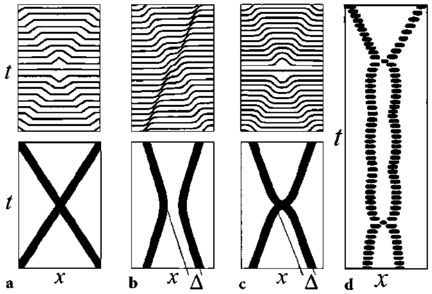

Figure 1. Elastic collisions: (a) linear-shaped antikink and kink in the linear wave equation; (b) kink-kink repulsive collision in SG; (c) antikink-kink attractive collision in SG. Top panels show $u(x, t)$, while bottom panels show the trajectories of soliton cores in the (x, t)-plane. (d) Inelastic collisions between two NLS solitons in the presence of weak perturbation.

Given the mobility of the localized coherent structures, it is natural to consider the outcome of their collisions. In fact, it was the "solitary" nature of such interactions (Zabusky & Kruskal, 1965) that inspired the term *soliton* in the case of integrable systems.

Linear versus Nonlinear Collisions

Consider the collision of two wavepackets that are governed by a linear wave equation. In this case, the superposition principle inherent in linearity guarantees that the two packets do not "feel" each other and survive the collision without change of shape, speed, or trajectory (see Figure 1(a)).

On the other hand, nonlinear dynamics offers more interesting collisions. Here, the result of the interaction of two excitations does not resemble their sum and (at least) a phase shift Δ is present (compare Figures 1(b) and 1(c) with 1(a)). In 1(b) and 1(c), the kink-kink and the antikink-kink collisions in the nonlinear SG equation are shown. The soliton cores do not merge in mutually repulsive collisions 1(b) while they do so in attractive collisions 1(c).

Elastic versus Inelastic Collisions

In fully integrable systems, solitons have the remarkable property (often used to define them) of colliding elastically (Ablowitz & Segur, 1981). In these special systems, the dynamics are severely restricted by the existence of an infinite set of conservation laws. Although realistic systems are typically non-integrable, the non-integrability in many applications is weak and can be treated by including small perturbative terms in integrable equations. Such perturbations are called Hamiltonian if the total energy of the perturbed system remains a dynamical invariant. While collisions in linear systems are much simpler than those in integrable nonlinear systems (such as the SG and NLS equations), the latter can, in turn, also be very different from the much more complex picture of inelastic collisions in near-integrable or more generally non-integrable systems.

In general non-integrable systems, the inelasticity of collisions may be manifested through emission of radiation, excitation of soliton internal modes, and energy exchange between solitons. Internal modes are the long-lived, spatially localized, oscillatory excitations (corresponding to the point spectrum of the linearization around the wave). These can typically be excited only for a particular sign of the perturbation (Campbell et al., 1983).

Small-amplitude radiation waves correspond to an irreversible chunk of energy lost in the collision (radiated toward the boundaries). Such modes are extended, plane waves of the continuous spectrum. Notice that the energy of internal modes can be partly restored to solitons if they collide again, while this is not possible for radiation waves (in the first-order approximation).

Strong energy exchange between nonlinear waves in near-integrable systems can also occur in the absence of the above two mechanisms (and for different types and signs of the original perturbation). There are two necessary conditions for this recently manifested mechanism (Dmitriev et al., 2001, 2002). It can be observed only if the energy exchange is not forbidden by the conservation laws existing in the perturbed system and if the collision is of attractive type, as in Figure 1(c). For example, in the SG equation perturbed by the term εu_{xxxx}, where energy and momentum are conserved for the one (free)-parameter kink-solitons, energy exchange is possible only when more than two solitons participate in the collision. The energy exchange between only kinks or only antikinks is also not possible because SG solitons with the same parity repel each other. Similarly, the effect is not possible in the Korteweg–de Vries (KdV) equation, where soliton interactions are always mutually repulsive.

In the NLS equation, in-phase solitons attract each other, while out-of-phase solitons repel. Each soliton has two parameters (amplitude and phase), and for many practically important perturbations, there are two conserved quantities: the Hamiltonian and L^2 norm of the solution. Thus, energy exchange between two nearly in-phase NLS solitons is possible in the presence of a weak perturbation. The effect of radiationless energy exchange between solitons survives even for a very weak perturbation, as it decreases linearly with a decrease in perturbation amplitude, while other effects of the perturbation decay faster. If the perturbation is not small, the energy exchange effect mingles with radiation emission and possibly with the excitation of internal modes.

Probabilistic Nature of Soliton Collisions

In many examples of perturbed, non-integrable models related to applications in optics, fluid mechanics, or condensed-matter physics (Kivshar & Malomed, 1989), the result of soliton collisions can be predicted only in a probabilistic sense. The following sources of stochastic behavior can be identified. First, chaotic soliton scattering can arise from resonant interaction with the soliton internal modes (Campbell et al., 1986; Gorshkov et al., 1992). Second, in discrete systems, the result of inelastic collisions can be sensitive to the coordinate of collision point, x_c, with respect to the lattice site (Dmitriev et al., 2003; Papacharalampous et al., 2003). Because the coordinate of the collision point usually cannot be controlled, it is natural to describe the result of the collision as a function of the random variable x_c. Finally, the result of the collision can be extremely sensitive to some other uncontrolled characteristics, such as the relative phase of the colliding solitons, as has been demonstrated to dramatically affect the collisions between NLS solitons or between kinks and breathers in SG (Dmitriev et al., 2001, 2002, 2003; Papacharalampous et al., 2003).

This last source of randomness is important when energy exchange between solitons is possible. In Figure 1(d), an example of inelastic interaction between two NLS solitons in the presence of quintic perturbation, $\varepsilon|u|^4u$ in Equation (4), with a small $\varepsilon > 0$ is presented (Dmitriev & Shigenari, 2002). (The regions of the (x, t)-plane are shown, where the real part of the solution exceeds a certain value.) After each collision, the properties of the solitons such as the frequency and amplitude are different depending on the collision phase. With a certain probability, the solitons can attain (after a number of collisions) a velocity sufficient to overcome their weak mutual attraction. In the example presented, the solitons split after the fourth collision. Emission of extended wave radiation is monitored and found to be very weak in this case; as a result, the solitons may continue to collide for a very long time, forming a two-soliton bound state. However, the probability P to obtain a bound state with the lifetime T decays algebraically as $P \sim T^{-\alpha}$ (Dmitriev & Shigenari, 2002), which is a manifestation of the chaotic character of their interaction.

In conclusion, linear waves do not "feel" each other, and nonlinear waves of integrable equations emerge unscathed from collisions, retaining their solitary character. However, solitary wave collisions in more realistic, non-integrable models remain a fascinating topic, where a number of basic mechanisms (such as internal mode resonances, extended wave radiation, and radiationless energy exchange) have been elucidated, but the full picture is far from complete. The probabilistic interpretation of collisions mentioned above may prove a fruitful viewpoint in future studies.

P.G. Kevrekidis and S.V. Dmitriev

See also **Nonlinear Schrödinger equations; Partial differential equations, nonlinear; Sine-Gordon equation; Solitons; Solitons, types of**

Further Reading

Ablowitz, M.J. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM

Campbell, D.K., Schonfeld, J.F. & Wingate, C.A. 1983. Resonance structure in kink–antikink interactions in $\phi/4$ field theory. *Physica* D, 9: 1–32

Campbell, D.K., Peyrard, M. & Sodano, P. 1986. Kink-antikink interactions in the double sine-Gordon equation. *Physica* D, 19: 165–205

Dmitriev, S.V., Kivshar, Yu.S. & Shigenari, T. 2001. Fractal structures and multiparticle effects in soliton scattering. *Physical Review* E, 64: 056613

Dmitriev, S.V., Semagin, D.A., Sukhorukov, A.A. & Shigenari, T. 2002. Chaotic character of two-soliton collisions in the weakly perturbed nonlinear Schrödinger equation. *Physical Review* E, 66: 046609

Dmitriev, S.V. & Shigenari, T. 2002. Short-lived two-soliton bound states in weakly perturbed nonlinear Schrödinger equation. *Chaos*, 12: 324

Dmitriev, S.V., Kevrekidis, P.G., Malomed, B.A. & Frantzeskakis, D.J. 2003. Two-soliton collisions in a near-integrable lattice system. *Physical Review* E, 68: 056603

Dodd, R.K., Eilbeck, J.C., Gibbon, J.D. & Morris, H.C. 1982. *Solitons and Nonlinear Wave Equations*, London: Academic Press

Gorshkov, K.A., Lomov, A.S. & Gorshkov, M.I. 1992. Chaotic scattering of two-dimensional solitons. *Nonlinearity*, 5: 1343–1353

Kivshar, Yu.S. & Malomed, B.A. 1989. Dynamics of solitons in nearly integrable systems. *Reviews of Modern Physics*, 61: 763–915

Papacharalampous, I.E., Kevrekidis, P.G., Malomed, B.A. & Frantzeskakis, D.J. 2003. Soliton collisions in the discrete nonlinear Schrödinger equation. *Physical Review* E, 68: 046604

Sulem, C. & Sulem, P.L. 1999. *The Nonlinear Schrödinger Equation*, New York: Springer

Zabusky, N.J. & Kruskal, M.D. 1965. Interactions of solitons in a collisionless plasma and the recurrence of initial states. *Physical Review Letters*, 15: 240–243

COLOR CENTERS

Gemstones are brightly colored because of the presence of *color centers*, atomic-scale imperfections that absorb light in otherwise transparent crystals. Historically, the term *color centers* has been associated with such imperfections in a special class of crystals called alkali halides, because it was in these relatively simple transparent hosts that the scientific study of color centers flourished. A German research program that began in the 1930s soon led to the recognition that alkali halides are excellent hosts for scientific studies of defects in nonmetallic solids, and the understanding of fundamental properties in these materials has been of great significance in the studies of similar phenomena in more complicated (and sometimes more practical) situations and materials.

Sodium chloride—common table salt—is the most familiar alkali halide. It consists of positively charged sodium and negatively charged chlorine ions, alternating positions in a simple cubic array. In the perfect crystal, each ion has six nearest neighbors of opposite charge. Simple defects may involve chemical impurities, such as a positively charged silver or thallium ion on an alkali site, or the removal of one or more ions from normal positions. The most fundamental of the latter class is the F center (from the German *Farbzentren*), a halogen ion vacancy that has trapped an electron. Defects involving more than one F center, or chemical impurities next to one or more F centers, may also occur.

Understanding the properties of color centers in detail requires some knowledge of quantum concepts. For example, the trapped electron of an F center can exist only in certain quantum states. In order for light to be absorbed, the energy of an incident photon must match the energy difference between the lowest quantum state and a higher one. This energy difference depends on the host crystal, and so F centers induce different colors in different alkali halides.

The situation becomes more complicated when one considers that the electron is not trapped in a static host, but rather that the neighboring ions are vibrating. The electron's interaction with these vibrations, whose time scale is long with respect to that of the (light) electron's motion, means that the energy difference between ground and excited electronic states does not have a single value, but rather a distribution of values about some mean. Thus, for example, whereas the mean photon absorption energy for F centers at low temperatures in NaCl is 2.77 eV, the mean width of the distribution of absorption energies is 0.26 eV.

This *electron-lattice interaction* has other consequences. In most cases, an F center at low temperature that has been excited by light will re-emit light. However, the mean energy of the emitted photon is found to be considerably smaller than that of the absorbed photon. This *Stokes shift* is exemplified in NaCl, where the mean photon emission energy at low temperatures is 0.98 eV.

Why is there a Stokes shift? In simple terms, the energy-level structure of the F center is determined by the mean position of the neighboring ions. However, the mean position of the neighboring ions is in turn determined by the quantum state of the F center electron. After the electron is excited from the ground state to another quantum state, the neighboring ions relax in response to the change in the average force exerted on them by the electron. This relaxation then leads to a change in the energy-level separations, as well as other fundamental aspects of their properties. The emitted photon has an energy smaller than the absorbed photon had, and energy is conserved in the total cycle as the relaxation processes create quanta of lattice vibrations, or phonons, which remove the excess energy.

This cyclic process may be visualized by means of a configuration coordinate diagram. In the simplest case, this diagram consists of two equal parabolas, displaced vertically (in energy, E) and horizontally (in the displacement of neighboring ions, R), as shown in Figure 1. Each parabola represents the vibration of neighboring ions and leads to quantized vibrational states. According to the Franck–Condon principle, electronic transitions take place vertically on the diagram—the massive ions do not respond instantly to the excitation by the photon—so absorption corresponds to a transition from A to B, emission C to D. This picture is highly oversimplified, but it does represent a physical and visual way to understand phenomena associated with the optical properties of F centers (and by extension, many other types of defects in both insulators and semiconductors).

We now consider a perfect alkali halide crystal in which one electron has been removed by light (ionizing radiation) from the array of negative ions. At low temperatures, this is found to lead to the phenomenon of self-trapping, whereas in semiconductors, the removal of an electron from a valence band of occupied states leads to motion of the empty state, or hole, and resulting electrical conduction; in most alkali halides the missing charge becomes localized in space, or self-trapped. What happens in detail is that the halogen ion that has lost one electron (and become a neutral atom) can form a chemical bond with a nearby halogen ion. In the process, both of these move toward each other, the lattice around them relaxes, and the missing charge is equally shared by this halogen molecule-ion.

Self-trapping is not a universal process; it "costs" energy to localize a quantum particle, and the energy

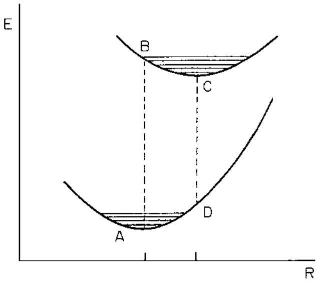

Figure 1. Schematic configuration coordinate diagram.

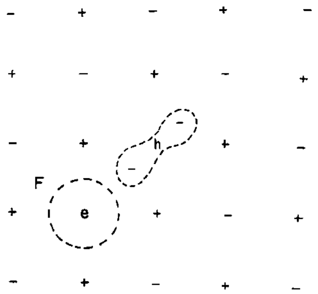

Figure 2. Adjacent F center and self-trapped hole center in an alkali halide. "+" and "−" denote alkali and halogen ions, respectively.

gained by chemical bonding and relaxation must overcome this. Self-trapping of holes occurs in most alkali halides, but not in semiconductors and not in many other insulators. Also of interest is the creation of defects by the self-trapping of excitation energy in alkali halides, leading to a *self-trapped exciton*. To approach this, we consider the trapping of an electron by a self-trapped hole. Since the hole is effectively positive (it resulted from the removal of an electron), we might imagine that the trapped electron would find itself in loosely bound quantum states about the self-trapped hole. However, there is another possibility. If it does not cost too much energy for the halogen molecule-ion to move into one halogen site, rather than be shared by two sites, then the electron may be trapped in the other, empty halogen site. One then has an F center next to a halogen molecule-ion; hence, two defects have been formed. This situation is shown in Figure 2. Since the two defects are adjacent, this may be an unstable arrangement, and indeed there is a finite probability that the system will revert back to the perfect crystal, either before or after the emission of a photon. But, in many cases, it is found that this nearest-neighbor arrangement is the precursor for the creation of a stable defect pair: the halogen molecule-ion may migrate through the halogen sublattice, not by long-range atomic motion, but rather by short-range halogen motion accompanied by motion of the hole. This results in sequential sharing of the hole by the halogens as the hole migrates away from the F center.

W. Beall Fowler

See also **Excitons; Franck–Condon factor; Quantum theory**

Further Reading

Crawford, J.H. & Slifkin, L.M. (editors). 1972 (vol. 1), 1975 (vol. 2). *Point Defects in Solids*, New York, Plenum

Fowler, W.B. (editor). 1968. *Physics of Color Centers*, New York: Academic

Hayes, W. & Stoneham, A.M. 1985. *Defects and Defect Processes in Nonmetallic Solids*, New York: Wiley

Schulman, J.H. & Compton, W.D. 1962. *Color Centers in Solids*, New York: Macmillan

Song, K.S. & Williams, R.T. 1993, 1996. *Self-Trapped Excitons*, Berlin and New York: Springer

COMMENSURATE-INCOMMENSURATE TRANSITION

When some local property in a crystalline solid (atomic positions or orientation of local magnetic moments) develops a spatial modulation with a wavelength b that differs from the underlying lattice spacing a, one speaks of a "modulated phase." A modulated phase is said to be "commensurate" when the ratio b/a is rational and "incommensurate" when b/a is irrational.

Modulated phases are experimentally observed by the appearance of "satellite spots" in X-ray, neutron, or electron diffraction patterns (Janssen & Janner, 1987; Cummins, 1990). Observations of spatially modulated structures are abundant in condensed matter physical systems, such as the ferrimagnetic phases of the rare earths and their compounds, long-period structures of binary alloys, graphite intercalation compounds, or the polytypic phases of spinelloids, perovskites, and micas, among other minerals.

The wavelength b characterizing the modulation varies with external parameters, such as temperature, pressure, or magnetic field. Sometimes, this variation is smooth, but often it remains constant at a rational locking value through some range of values of the external parameter before changing to another rational locking value, and so on.

The ubiquity of modulated phases shows that the physical origin of these behaviors cannot be tied to particularities of specific types of systems, but must be of a general character. It is widely recognized that modulated phases appear whenever different terms in the free energy of the system compete, each one trying to impose its own characteristic length scale. The external parameters control the relative strength of the competing interactions and new compromises are reached as they vary.

An insight into the physics of modulated phases was obtained through detailed analyses of simple model systems of competing interactions. Although these simple models are unlikely to fit experiments on specific materials, they help to understand the complexity of behaviors that emanate from length-scale competition and to discern essential features.

One of the best-studied model systems with competing interactions is the axial next-nearest-neighbors Ising (ANNNI) model (motivated by the magnetic structures of erbium), which was introduced by R.J. Elliot in 1961 (Yeomans, 1988). This is an Ising model with a two-state spin, $S = \pm 1$, on each site of a cubic lattice. Interactions between spins on nearest-neighbor sites are ferromagnetic, but there are second (next-nearest) neighbor antiferromagnetic interactions along the axial direction, z. The Hamiltonian is

$$H = -\frac{1}{2}J_0 \sum_{i,j,j'} S_{i,j}S_{i,j'} - J_1 \sum_{i,j} S_{i,j}S_{i+1,j}$$
$$-J_2 \sum_{i,j} S_{i,j}S_{i+2,j}, \qquad (1)$$

where i indicates the two-dimensional layers perpendicular to the axial direction, and j, j' are nearest-neighbor sites within a layer. Both J_0 and J_1 are positive (ferromagnetic interactions), thus favoring the same value of neighboring spins, but $J_2 < 0$ (antiferromagnetic) so that it favors opposite values of second neighbor spins along the z-direction.

In the absence of thermal effects ($T = 0$), the favored spin configurations along the z-direction are the ferromagnetic alignment ($\cdots + + + + + + + \cdots$) if $\kappa = -J_2/J_1 < \frac{1}{2}$, and the antiphase ($\cdots + + - - + + \cdots$) configuration if $\kappa > 1/2$. Exactly at $\kappa = \frac{1}{2}$, any configuration containing a stripe of two or more spins "+" followed by a stripe of two or more spins "−", etc., has the same energy; thus there is a multiphase point.

A convenient notation is to use $\langle n_1, \ldots, n_m \rangle$ to represent a state in which a set of stripes of width n_i of alternating spins repeat. For example, ($\cdots + + - - + + + - - + + - - - - \cdots$) is denoted by $\langle 223 \rangle$. Ferromagnetic and antiphase configurations are consistently denoted by $\langle \infty \rangle$ and $\langle 2 \rangle$, respectively. When temperature increases from zero, entropic effects regulate the competition among the ferro and antiferro interactions, and a flower (called a "devil's flower" by Per Bak) of petal-like phases of modulated structures opens up from the multiphase point.

Figure 1 shows how new commensurate phases appear as temperature increases, demonstrating qualitatively what is commonly observed in experiments—locking to a few short-wavelength commensurate phases separated either by first-order phase transitions (as in the low-temperature regime of Figure 1) or by regions where the wave vector appears to vary smoothly (as at higher temperatures).

The following question naturally arises: what determines the behavior of the modulation wave vector as parameters vary? The answer to this question is closely tied to a central paradigm of nonlinear science, the soliton concept in the context of its discrete counterpart called discommensuration.

To clarify this issue, consider the simplest model of competing interactions (Griffiths, 1990), the Frenkel–Kontorova model, which can be visualized as an array of atoms at positions (u_n), $-\infty < n < +\infty$, experiencing a periodic substrate potential $V(u) = V(u+1)$ and nearest-neighbor interaction $W(\Delta u)$. The Hamiltonian is

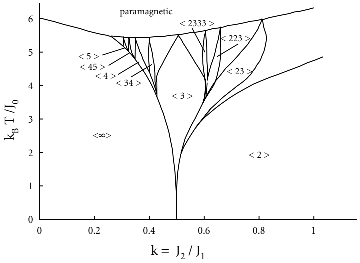

Figure 1. Mean-field phase diagram of the ANNNI model showing the main commensurate phases.

$$H = \sum_n [KV(u_n) + W(\Delta u_n)], \qquad (2)$$

where the parameter K controls the relative strength of the interactions. The standard Frenkel–Kontorova model corresponds to

$$V(u) = \frac{1}{(2\pi)^2}[1 - 2\cos(2\pi u)],$$
$$W(\Delta u) = \frac{1}{2}(\Delta u)^2 - \sigma \Delta u. \qquad (3)$$

Note that V favors an integer value of the interspacing Δu, while W favors a uniform value σ, so both interactions compete to determine the configuration (u_n), characterized here by the average interspacing $\omega = \langle \Delta u \rangle$.

If (as in the standard model) the interaction $W(\Delta u)$ is a convex function, a complete rigorous characterization of the model phase diagram was given by Aubry (1985). Thus, we restrict the analysis to convex $W(\Delta u)$.

In the thermodynamic limit of the system, care must be taken when defining what is meant by configurations of minimum energy or even by the energy of a configuration. Thus, the mean energy ε per particle of a configuration (u_n) is defined as

$$\varepsilon = \lim_{(N-M)\to\infty} \frac{1}{N-M}$$
$$\times \sum_{j=M}^{N-1} [KV(u_j) + W(u_{j+1} - u_j)]. \qquad (4)$$

A minimum energy configuration (MEC) (u_j) is that for which the arbitrary displacement of any finite segment $(u_j + \delta_j)$, $M < j < N$, of the configuration always produces a nonnegative energy change. For an MEC, both limits (ω and ε) exist. Moreover, for every real number ω, there is at least an MEC with average spacing ω, and its mean energy per particle $\varepsilon(\omega)$ is a well-defined function. A ground state is an MEC that is "recurrent," meaning that any arbitrary finite segment of the configuration approximately reappears somewhere else, within arbitrary precision. One often speaks of the ground state that corresponds to fixed values of the parameters (K and σ for the standard model), meaning the ground-state configuration for which $\varepsilon = \min_{\omega} \varepsilon(\omega)$ occurs at the fixed values of the parameters. The minimum ε, which depends on the parameter values, is the ground state energy, and the value of ω where that minimum occurs is the ground-state average spacing.

The ground-state phase diagram of the standard Frenkel–Kontorova model is shown in Figure 2. Some low-order commensurate phases (rational ω) are shown there, but in between any two of them, there are an infinite number of others. The ground-state average spacing ω as a function of the tension σ, for a fixed value of $K = 1$, is shown in Figure 3. This is a continuous function with plateaus of (rational) constant value, which correspond to commensurate phases. This function is called "devil's staircase" or Cantor function. No first-order phase transitions between commensurate phases of different commensurabilities are allowed in convex models.

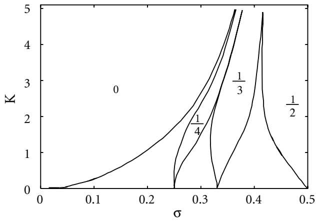

Figure 2. Ground-state phase diagram of the standard Frenkel–Kontorova model. Each point in the diagram is characterized by the value of ω, average spacing of the groundstate configurations for the parameters (K, σ) of the model. The numbers in the figure are values of ω. Only a few commensurate phases are plotted, as there are infinitely many commensurate and incommensurate phases in between any two of them.

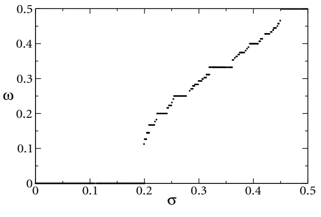

Figure 3. ω as a function of σ at K = 1, a devil's staircase, there being a "plateau" for each rational value of ω.

For rational values of $\omega_c = p/q$, commensurate ground states are periodic configurations: $\sigma_{qp}(u_n) := (u_{n+q} - p) = (u_n)$. But there are also two MECs with average spacing (ω_c) that are nonrecurrent and called elementary "discommensurations" (DC). These can be described as a localized compression (retarded DC) or a localized expansion (advanced DC) over a commensurate ground state. The distortion being localized, they have the same mean energy per particle $\varepsilon(\omega_c)$ of the ground state. However, the difference of the non-averaged energy of a DC and the ground state (the extra energy of the localized compression or expansion) is generally nonzero; this is called the "creation energy" of the DC.

If a configuration contains two elementary DCs superimposed on a commensurate ground state, the energy difference with respect to the ground state can be understood as the sum of the creation energies of the DCs plus the interaction energy. If both DCs are of the same type, the energy decreases with the interdistance (repulsive interaction). The interaction is attractive for DCs of opposite type. Elementary DCs have an *excess length* Δ with respect to the ground state, which is exactly $\Delta = 1/q$ for an advanced DC and $\Delta = -1/q$ for a retarded one.

The creation energy of a DC varies linearly with the tension σ, the slope being the excess length Δ of the DC. Inside the corresponding commensurate plateau, the creation energy of an advanced DC increases with σ and disappears at the left edge $\sigma^{-}_{\mathrm{C-IC}}$ of the plateau, while the creation energy of the retarded DC decreases with σ, disappears at the right edge $\sigma^{+}_{\mathrm{C-IC}}$ (see Figure 4). Inside the commensurate plateau, the creation of an elementary DC is energetically costly, but right after either edge ($\sigma^{-}_{\mathrm{C-IC}}$ or $\sigma^{+}_{\mathrm{C-IC}}$) is crossed, it becomes energetically advantageous to create as many DCs (of the correct type) as possible, were it not for the repulsive interaction between them that prevents condensation of an infinite number of DCs.

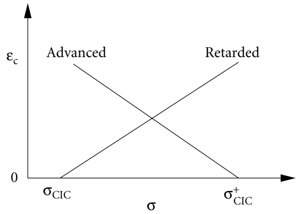

Figure 4. Creation energy ε_c of both types of elementary discommensurations for values of σ inside the tongue associated with a rational ω. At a certain critical value of σ this energy becomes negative and a commensurate–incommensurate transition occurs, changing the average spacing of the ground-state configuration.

The compromise between these two effects determines the density c of DCs, and, as each DC carries an excess length Δ, the average spacing of the new structure is $\omega = \omega_c + c\Delta$. For very small values of $|\sigma - \sigma_{C-IC}|$, where the interdistance (c^{-1}) is so large as to approximate a continuous variable, the scaling behavior of the ground-state average spacing ω in the commensurate-incommensurate transition is then

$$|\omega - \omega_c| \sim -(\ln|\sigma - \sigma_{C-IC}|)^{-1}. \qquad (5)$$

For nonconvex interactions, the basis for an interpretation of the transitions between modulated phases in terms of interacting DCs is less rigorously supported. However, the central idea that the creation of localized defects or discommensurations drives these transitions and that the character (repulsive, attracting, or oscillatory) of their interaction determines the type of transition has been successfully applied, and undoubtedly it provides the most physical way of understanding why and how transitions between modulated phases occur. This fact adds to the many merits of the soliton paradigm in nonlinear science.

Luis Mario Floría

See also **Aubry–Mather theory; Frenkel–Kontorova model; Ising model; Phase transitions**

Further Reading

Aubry, S. 1985. Structures incommensurables et brisure de la symmetrie de translation I [Incommensurate structures and the breaking of traslation simmetry I]. In *Structures et Instabilités*, edited by C. Godréche, Les Ulis: Editions de Physique, pp. 73–194

Cummins, H.Z. 1990. Experimental studies of structurally incommensurate crystal phases. *Physics Reports*, 185: 211–409

Griffiths, R.B. 1990. Frenkel–Kontorova models of commensurate–incommensurate phase transitions. In *Fundamental Problems in Statistical Mechanics VII*, edited by H. van Beijeren, Amsterdam: North-Holland, pp. 69–110

Janssen, T. & Janner, A. 1987. Incommensurability in crystals. *Advances in Physics*, 36: 519–624

Yeomans, J. 1988. The theory and application of axial Ising models. *Solid State Physics*, 41: 151–200

COMPACT SET

See **Topology**

COMPACTON

See **Solitons, types of**

COMPARTMENTAL MODELS

Compartmental models, which model the flow of material between components of a system, have been applied to the study of nutrient or energy flow in an ecosystem, to analyses of tracer kinetics in an organism, and to the theory of epidemics, among other applications.

The construction of a compartmental model is straightforward and intuitive, but the estimation of the associated parameters is not. The system is divided into homogeneous compartments, and the flow of "material" (meaning nutrients, energy, money, goods, electrons, radioactive tracers, information, etc.) between them is traced, keeping account of the level of material in each compartment. The compartments are represented by boxes and the flows are represented by arrows (see Figure 1).

This representation is a directed graph in which the compartments are vertices and the flows are arcs. To analyze a system, however, the model must be quantized by studying the rate of change of the level x_i of the ith compartment in time as

$$x_i = x_i(t). \qquad (1)$$

The flow between the ith and jth compartment, designated f_{ij}, is a rate, measured in terms of quantity of material per unit time.

A differential equation describing the behavior of x_i may be obtained by equating the time rate of change of x_i to the difference between the flows coming in and those going out of the ith compartment. Thus,

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_k f_{ki} - \sum_j f_{ij}.$$
$$\text{rate of change} = \text{flows in} - \text{flows out} \qquad (2)$$

The f_{ij} may be constant or variable in time and are usually functionally dependent on the x_i's.

A widely used assumption (particularly in physiology and medicine) is that

$$f_{ij} = a_{ij}x_i, \qquad (3)$$

which corresponds to a certain fraction a_{ij} of the level of material x_i in compartment i passing to compartment j per unit time. In the ecological literature, this is often

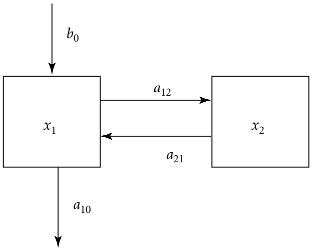

Figure 1. A compartmental model of tracer kinetics.

called donor-controlled flow. (The a_{ij} may be functions of time but are usually assumed to be independent of the x_is.)

Tracer Kinetics

Tracer kinetics refers to a technique for studying the flow of a drug or other material through a biological system (often a human body) near steady state. A small quantity of radioactive isotope or other recognizable form of the material (the tracer) is introduced, so that the tracer flows may be modeled by linear equations.

As a simple example, consider lipoprotein metabolism. One compartment corresponds to the blood plasma and the other corresponds to the extravascular space. A unit dose of tracer is injected into the blood and the flow follows the diagram in Figure 1. The quantity of interest is the flow out of compartment 1 (excretion), which cannot be measured directly, although the level of compartment 1 over time is known.

The differential equations are those obtained from (2) by taking the flow rates to be donor controlled; thus,

$$\begin{aligned} \mathrm{d}x_1/\mathrm{d}t &= -a_{10}x_1 - a_{12}x_1 + a_{21}x_2, \\ \mathrm{d}x_2/\mathrm{d}t &= a_{12}x_1 - a_{21}x_2, \end{aligned} \tag{4}$$

with initial conditions

$$x_1(0) = 1, \quad x_2(0) = 0. \tag{5}$$

These equations can be solved for $x_1(t)$ in terms of the three parameters a_{10}, a_{12}, a_{21}; since x_1 is known for all t, this result can be used to estimate these parameters by fitting a number of nonlinear equations.

General Differential Equations and Solutions

The differential equations for a system of n compartments can be written in vector form as

$$\mathrm{d}\boldsymbol{x}/\mathrm{d}t = A\boldsymbol{x} + \boldsymbol{f}, \tag{6}$$

where A is the matrix given by

$$A = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix} \tag{7}$$

and $\mathbf{x} = [x_1, x_2, ..., x_n]^{\mathrm{T}}$ is the vector of levels of material in the n compartments and $\boldsymbol{f}$ is the input vector. The a_{ij} for $i \neq j$ are the flow rates, while the diagonal elements of A are

$$a_{ii} = -\sum_{j \pm i} a_{ij} - a_{i0}. \tag{8}$$

Although the input to a compartmental system is often to a single compartment, it may be distributed to input compartments. In general, we assume that the input to the system is $\boldsymbol{f} = B\boldsymbol{u}$, where $\boldsymbol{u}$ is a vector function distributed to the various compartments by B. Typically, B has one column but as many rows as there are compartments.

The outputs (measurements) of the system may be similarly combined. If the system is assumed donor to be controlled, the output will have the form

$$\boldsymbol{y} = C\boldsymbol{x}, \tag{9}$$

where the output of the compartments is directed to one or more measuring devices with measurements $\boldsymbol{y}(t)$. (Here, C is a general sampling matrix that is usually composed of 0s and 1s.)

The problem is then translated to the differential equation form

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= A\mathbf{x} + B\mathbf{u}, \\ \mathbf{y} &= C\mathbf{x}, \\ \mathbf{x}(0) &= \mathbf{0}. \end{aligned} \tag{10}$$

where we have assumed that the initial levels are $\mathbf{0}$. (This is no restriction as an initial value can be incorporated into $B\boldsymbol{u}$.) Solutions of the system are determined by matrices A, B, and C as

$$\mathbf{y}(t) = C\mathbf{x}(t) = C\int_0^t \mathrm{e}^{A(t-s)}B\mathbf{u}(s)\mathrm{d}s \tag{11}$$

where $\boldsymbol{x}(0) = 0$. Typically, however, the flow rates in the matrix A are unknown and must be determined from knowledge of $\mathbf{u}(t)$ and $\boldsymbol{y}(t)$, which is called the identification problem. A system is said to be identifiable if there is a unique solution to this problem, which is not always the case, and even if a solution does exist, it may be difficult to find.

Parameter Estimation

The problem of actually estimating the parameters of an identifiable problem is often an ill-posed problem, meaning that the estimated output may be close to the true output, while the parameters are still quite different from the true parameters. Moreover, such parameters are usually sensitive to slight perturbations of the data.

Suppose we are given a discrete set of observations/measurements made at times $t_i, i = 1, \ldots, m$ such that

$$y_i = y(t_i) + \varepsilon_i, \tag{12}$$

where ε_i are assumed to be independent normal random variables with mean 0 and constant variance. The standard approach to estimation is to solve for A in the least-squares sense

$$\min_{a_{ij}} \sum_i |y_i - y(t_i)|^2 \tag{13}$$

subject to the condition that A is a compartmental matrix of the appropriate model. (This is a nonlinear least-squares problem, since $y(t)$ is a linear combination of exponential functions.)

Another approach to estimation involves two stages; the data are first fitted to a sum of exponentials (Anderson, 1983)

$$g(t) = \sum_{k=1}^{n} c_k e^{\lambda_k t}. \tag{14}$$

In most nonpathological cases, the number of exponentials to be considered is the same as the size of the model. The estimates $\hat{c}_k$ and $\hat{\lambda}_k$ are obtained by least squares subject to $\lambda_k \le 0$. Then, the Laplace transform of the fitted function $g(t)$ is compared with the Laplace transform of $y(t)$. As both are rational functions, the numerator and denominator are polynomials whose coefficients must be approximately equal.

There is also the method of peeling, which avoids least squares. This involves fitting an exponential function to the later data values. This corresponds to the eigenvalue of smallest magnitude. The difference between the data and this function is then used to approximate the second eigenvalue, and so on.

Gilbert G. Walter

See also **Continuum approximations; Neural network models**

Further Reading

Anderson, D.H. 1983. *Compartmental Models and Tracer Kinetics,* Berlin and New York: Springer

Bellman, R.E. & Aström, K.J. 1970. On structural identifiability. *Mathematical Biosciences*, 7: 329–339

Dennis, J.E. & Schnabel, R. 1983. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Englewood Cliffs, NJ: Prentice-Hall

Godfrey, K. 1983. *Compartmental Models and Their Applications*, New York: Academic Press

Jacquez, S.A. 1996. *Compartmental Analysis in Biology and Medicine*, Ann Arbor: University of Michigan Press

Walter, E. 1987. *Identifiability of Parametric Models*, Oxford and New York: Pergamon

Walter, G.G. & Contreras, M. 1999. *Compartmental Modelling with Networks*, Boston: Birkhäuser

COMPLEX GINZBURG–LANDAU EQUATION

Ginzburg–Landau (GL) is a generic name for a class of basic models of nonlinear science, with a wide range of applications. They combine most fundamental features that underlie pattern formation in nonlinear media: dissipation (linear and nonlinear) and gain (provided by an intrinsic instability), conservative nonlinearity (accounting for wave mixing), and diffusion and/or dispersion (dependence of the dissipation rate and wave frequency on the wave number). Usually, the diffusion and dispersion are linear, but in more sophisticated models they may be both linear and nonlinear. In most cases, GL equations are local, but nonlocal (integral) generalizations are sometimes used. Both single GL equations and systems thereof play important roles as fundamental models.

The name of this class of equations first appeared in a seminal paper by Vitaly Ginzburg and Lev Landau (1950), which presented a phenomenological theory of superconductivity (V.L. Ginzburg was awarded the Nobel Prize in Physics in 2003 for this work.) In a standard form (Tinkham, 1996), these equations are

$$\begin{aligned}
\eta\left(\partial_t + \mathrm{i}\kappa\phi\right)\psi &= -\left(\mathrm{i}\kappa^{-1}\nabla + \boldsymbol{A}\right)^2\psi \\
&\quad + \left(1 - |\psi|^2\right)\psi, & (1)\\
\boldsymbol{A}_t + \nabla\phi &= -\nabla\times(\nabla\times\boldsymbol{A}) + (2\mathrm{i}\kappa)^{-1} \\
&\quad \times\left(\psi^*\nabla\psi - \psi\nabla\psi^*\right) - |\psi|^2\boldsymbol{A} \\
&\quad + \nabla\times\boldsymbol{H}, & (2)
\end{aligned}$$

where ψ is the wave function of superconducting electrons, also known as the order parameter of the superconductivity, $\boldsymbol{A}$ and $\boldsymbol{H}$ are the vector potential of the intrinsic magnetic field and an externally applied field, ϕ is the electrostatic potential (related to $\boldsymbol{A}$ by an arbitrary gauge condition, e.g., $\phi_t + \nabla\cdot\boldsymbol{A} = 0$), η and κ are the dissipative and so-called GL parameters, ∇ is the gradient, and subscripts stand for the partial derivatives. A solution with $\psi = 0$ represents a normal state with no superconductivity. In their original work, Ginzburg and Landau (1950) introduced a time-independent version of the equations, while the full time-dependent system was proposed later.

In the modern theory of nonlinear pattern formation (Cross & Hohenberg, 1993), the name GL is applied to equations resembling Equation (1) [while there is usually no counterpart of Equation (2)], which govern the evolution of a physically or phenomenologically defined order parameter in a dynamical medium. For instance, in the case of chemical-reaction waves, the order parameter determines local concentrations of reactants (Ipsen et al., 2000); in applications to fluid mechanics, the velocity field is expressed in terms of the order parameter, and in numerous applications to optics the order parameter is a local amplitude of the electromagnetic field (Arecchi et al., 1999).

The first generic example of the GL equation of this type is the complex cubic GL (CCGL) equation ("cubic" refers to the nonlinearity), which is closest in its form to Equation (1), but the name "complex" stresses that its coefficients are complex:

$$\psi_t = g\psi + (a + \mathrm{i}b)\nabla^2\psi - (d + \mathrm{i}c)|\psi|^2\psi. \tag{3}$$

The first term on the right-hand side of Equation (3) accounts for the gain, provided that $g > 0$ (its imaginary part can be trivially removed), a and b are real coefficients accounting for the diffusion and dispersion, while d and c are parameters controlling the nonlinear dissipation and frequency shift. Note that a and d must be positive, otherwise (3) is an ill-posed equation. In the case $|a| \ll |b|$, $|d| \ll |c|$, Equation (3) may be treated as a perturbed version of the nonlinear Schrödinger (NLS) equation.

In the case $b = c = 0$, Equation (3) is called a real GL equation, to stress that all its coefficients are real, although ψ remains complex. The real GL equation may be represented in the gradient form, $\psi_t = -\delta L\{\psi, \psi^*\}/\delta\psi^*$, where $\delta/\delta\psi^*$ stands for the functional derivative, and $\int \left(-|\psi|^2 + |\nabla\psi|^2 + (1/2)|\psi|^4\right) \mathrm{d}V$ is a real Lyapunov functional. A consequence of the gradient representation is that L may only decrease, $\mathrm{d}L/\mathrm{d}t \leq 0$. This fact simplifies the dynamics of the real GL equation.

A fundamental feature of Equation (3) is that its zero solution, $\psi = 0$, becomes unstable as the gain g changes its sign from negative to positive. In this case, a transition from the stable trivial solution to a nontrivial state is called *supercritical.* In particular, the supercritical transition in the one-dimensional (1-d) case yields a solitary-pulse (SP) solution that can be found in an exact analytical form,

$$u = A\ [\cosh(\kappa x)]^{-(1+\mathrm{i}\mu)} \exp(-\mathrm{i}\omega t), \tag{4}$$

where A, κ, μ, and ω are uniquely determined by parameters of Equation (3). If the CCGL equation reduces to a perturbed NLS equation, the SP (4) can be obtained from the NLS soliton by means of the perturbation theory, provided that $bc < 0$ (otherwise, the NLS equation does not have bright-soliton solutions). However, the SP solution (4) is always unstable because the zero background around it is unstable.

In many cases (for instance, in the case of thermal convection in a binary fluid), a nontrivial state may be excited by a finite-amplitude perturbation, while the trivial solution is stable against small perturbations. The simplest model that describes the corresponding *subcritical* transition to a nontrivial state is the cubic-quintic complex GL (CQCGL) equation, first proposed by Petviashvili and Sergeev (1984) as

$$\begin{aligned}\psi_t &= -g\psi + (a + \mathrm{i}b)\,\nabla^2\psi \\ &\quad + (d - \mathrm{i}c)\,|\psi|^2\psi - (f + \mathrm{i}h)\,|\psi|^4\psi\,.\end{aligned} \tag{5}$$

Here, $g > 0$ implies stability of the zero solution, the last term with $f > 0$ guarantees overall stabilization of the system, the coefficient h accounts for a quintic nonlinear correction to the wave frequency, while $d > 0$ provides for a possibility of *nonlinear* gain. The CQCGL equation may give rise to nontrivial states, coexisting with the stable zero solution, if the nonlinear gain coefficient exceeds a value $d_{\min} = 2\sqrt{gf}$. An important result, obtained by means of analytical and numerical methods, is that the 1-d and 2-d versions of the CQCGL equation support SP solutions that may be *stable* (in the 2-d case, the localized pulse may carry vorticity, having a spiral structure). If all the parameters g, a, d, f, and h are small, the 1-d pulse can be constructed on the basis of the NLS soliton by means of the perturbation theory. However, the CQCGL equation does not make it possible to find stable SP solutions in an exact analytical form (one exact solution for a 1-d SP is known, but it is always unstable).

Patterns in nonlinear dissipative media may be supported not only by intrinsic gain; another possibility is to apply an external field, which is, for instance, the case for the pattern formation in laser cavities (Arecchi et al., 1999). In this case, the appropriate CCGL equation is

$$\psi_t = -g\psi + (a + \mathrm{i}b)\,\nabla^2\psi - (d + \mathrm{i}c)\,|\psi|^2\psi + P, \tag{6}$$

where the driving term, induced by the external field, may be of two different types: direct drive, $P = \varepsilon$, or parametric drive, $P = -\mathrm{i}\omega_0\psi + \varepsilon\psi^*$, where the asterisk stands for complex conjugation, ω_0 fixes the frequency, and ε is the drive's amplitude. Equation (6) with either type of drive can support stable SPs in 1-d and 2-d cases (in the case of the direct drive, SP settles on a nonzero background).

An important generalization is to consider systems of coupled GL equations. These may describe counter-propagating waves (for instance, in thermal convection in a binary-fluid layer), or second-harmonic generation in a lossy medium. In the latter case, the nonlinearity is not cubic, but *quadratic*, viz., $\psi_1^*\psi_2$ in the equation for the fundamental-frequency field ψ_1, and ψ_1^2 in the equation for the second-harmonic field ψ_2.

An alternative to the CQCGL equation is a system originating in nonlinear optics, in which the stability of the zero solution is provided by an extra linear equation,

$$\begin{aligned}\psi_t &= g\psi + (a + \mathrm{i}b)\,\nabla^2\psi - (d + \mathrm{i}c)\,|\psi|^2\psi - \mathrm{i}\kappa\chi, \\ \chi_t &= -G\chi - \mathrm{i}\kappa\psi,\end{aligned} \tag{7}$$

where κ is a real coupling constant, and $G > 0$ is the loss coefficient in the additional equation. The zero solution in this system is stable if the loss in the second equation and coupling are strong enough, $G > g$ and $\kappa^2 > Gg$. System (7) has exact SP solutions, in which both fields ψ and χ take the form (4) (with different amplitudes), but, in contrast to the CCGL equation proper, in this case the pulse may be *stable* (Atai & Malomed, 1998).

Yet another type of a system occurs if, due to a specific nature of the underlying physical problem, the complex order parameter ψ is coupled to an extra real-order parameter ϕ, which accounts for the existence of a conserved quantity in the medium (Matthews & Cox, 2000): $\psi_t = \psi + \nabla^2\psi - |\psi|^2\psi - \phi\psi$, $\phi_t = \nabla^2\left[\sigma\phi + \mu\left(|\psi|^2\right)\right]$. In this simplest version of the system, $\sigma > 0$ and μ are real constants.

A common feature of the various GL equations displayed above is their universality. Each is a generic representative of a class of models with given qualitative properties (for instance, super- or subcritical character of the excitation of nontrivial states, and the absence or presence of a conserved quantity). In more specific situations, there arise generic equations of other types. In particular, the complex Swift–Hohenberg (SH) equation describes a situation (for instance, in Rayleigh–Bénard convection) when the instability of the zero solutions appears at a finite wave number k_0 of small perturbations; thus,

$$\psi_t = g\psi - (a+ib)\left(\nabla^2+k_0^2\right)^2\psi - (d+ic)\,|\psi|^2\psi \quad (8)$$

with $g, a, d > 0$. Quasi-1-d solutions of Equation (8) can be looked for as $\psi\,(x, y, t) = \Psi(x, y, t)\,\exp\,(ik_0x)$, where Ψ is a slowly varying function, whose x- and y-dependences are characterized by large scales $X, Y \gg k_0^{-1}$. An asymptotic consideration is then consistent in the case $X \sim Y^2$, reducing the SH equation (8) to an anisotropic complex Newell–Whitehead–Segel equation, $\Psi_t = g\Psi + (a+ib)\left(2i\partial_x + \partial_y^2\right)^2\Psi - (d+ic)\,|\Psi|^2\Psi$.

GL equations generate rich dynamics. The simplest exact solutions are plane waves (PWs). In the case of the CCGL equation (3) (here, it is set by rescaling $g = a = d \equiv 1$), a family of PWs is

$$\psi = A\,\exp\,(iQx - i\omega t),\ A = \sqrt{1-Q^2}, \\ \omega = c + (b-c)Q^2, \quad (9)$$

where Q is a wave number (a parameter of the solution family) taking values $-1 < Q < +1$. Note that the group velocity of this PW is $d\omega/dQ = 2(b-c)Q$.

The CQCGL equation gives rise to two different PWs for given Q, with amplitudes $A_\pm = (2f)^{-1} \times \left[1 \pm \sqrt{1-4f\left(1+Q^2\right)}\right]$ (under the same normalization $g = a = d \equiv 1$). This solution exists in the interval $Q^2 < (4f)^{-1}(1-4f)$, and only A_+ may correspond to a stable PW.

The stability of PWs is an important issue. In the CCGL equation, a condition for stability against long-wave perturbations with a large wavelength takes a simple form, $Q^2 < (1+bc)\,/\left(3+2c^2+bc\right)$, which is called a generalized Eckhaus stability criterion (the original Eckhaus criterion, which applies to the real GL equation with $b = c = 0$, is $Q^2 < 1/3$). Obviously, this criterion cannot hold unless $1+bc > 0$ (the Benjamin–Feir–Newell (BFN) condition). The consideration of finite-wavelength perturbations gives rise to more complex stability conditions. Therefore, following Aranson and Kramer (2002), the structure of a full stability region for the PWs can be shown by means of its cross sections in the space (b, c, Q); see Figure 1. The figure makes a distinction between convective and absolute instabilities, when, respectively, the growing perturbation is traveling away or staying put. The transition from the 1-d CCGL equation to its multidimensional counterpart does not import extra instabilities to the PWs.

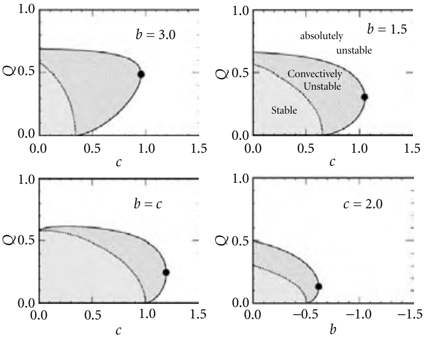

Figure 1. A set of cross sections of the stability region for the plane-wave solutions (9) of the cubic complex GL equation (3) in the space $(b, c, |Q|)$. The two top panels, the left bottom panel, and the right bottom panel show, respectively, the cross sections $b = \text{const}$, $b = -c$, and $c = \text{const}$. The filled circles mark turning points on the border between absolute and convective instabilities.

If the BFN combination $1+bc$ becomes negative, the PW develops phase turbulence, which means that $|\psi|$ remains roughly constant, while the phase of the complex phase ψ demonstrates spatiotemporal chaos. In the 1-d case, close to the BFN instability threshold, the chaotic evolution of the phase gradient $p \equiv \phi_x$ obeys the Kuramoto–Sivashinsky equation, which, in a rescaled form, is $p_t + p_{xx} + p_{xxxx} + pp_x = 0$. Deeper into the instability region, phase-slip points (PSPs) arise, at which the amplitude $|\psi|$ disappears. Multiple creation of PSPs leads to a transition from the phase turbulence to defect turbulence, which is distinguished by random dynamics of the PSP ensemble. Mixed turbulence at the border between these two types also occurs (Aranson & Kramer, 2002).

In the case when PWs are stable, shock waves can be generated by collision between two PWs (9) with different wave numbers Q_1 and Q_2. Although exact solutions for shocks are not available, they can be obtained in an approximate form, provided that $Q_1 - Q_2$ is small, or the coefficients b and c are small. In particular, a transient layer between the two PWs moves at the velocity $v = (b-c)(Q_1+Q_2)$, which is exactly the mean of the group velocities of the colliding PWs. In the 2-d case, PWs can collide obliquely; in this case, shocks take the form of a domain wall. Generally, the shocks are stable.

Besides the shocks (which are sources that emit PWs), the 1-d CCGL equation also gives rise to sinks, that is, localized hole-type structures that absorb PWs. If the CCGL equation (3) is a perturbed version of the NLS equation, sinks can be constructed as perturbed counterparts of NLS dark solitons, provided that $bc > 0$ (particular solutions for the sinks are available in an exact analytical form). A standing sink is actually a PSP, as $|\psi|$ vanishes at its center, and it may be dynamically stable. A moving sink is a finite dip in the profile of $|\psi|$; it is structurally unstable, as it is either decelerated (turning into a standing sink) or accelerated (eventually vanishing) by a quintic term added to the CCGL equation (Lega, 2001; Aranson & Kramer, 2002).

The 2-d CCGL equation displays spiral waves (SWs) in the form $\psi = A(r)\,\exp\,(\mathrm{i}N\theta - \mathrm{i}\omega t)$, where r and θ are the polar coordinates, N is an integer vorticity, and $A(r)$ is a complex function. An asymptotic form of the SW is $A(r) \approx \sqrt{1-Q^2}\exp(\mathrm{i}Qr)$ at $r \to \infty$, where Q is related to ω as in Equation (9), and $A(r) \sim r^N$ at $r \to 0$. (In the case of the real GL equation, the SW has $Q = 0$, which corresponds to a *vortex* solution. Similar solutions are generated by Equations (1) and (2), which represent *Abrikosov vortices* in superconductivity. For the prediction of these vortices, A.A. Abrikosov shared the Nobel Prize for Physics in 2003.) The asymptotic wave number Q is an eigenvalue of the 2-d CCGL equation, as it is uniquely selected by parameters of the equation. All the SWs with $N > 1$ are unstable. The SW with $N = 1$ may be subject to specific instabilities localized near its core (Aranson & Kramer, 2002).

An extension of the SW is a vortex line in three dimensions, which, in particular, may be closed into a ring. A vortex line with an additional wave number directed along its axis (twisted vortex) is also possible. The dynamics of 3-d vortex lines are quite complicated (Aranson & Kramer, 2002).

BORIS MALOMED

See also **Nonlinear Schrödinger equations; Partial differential equations, nonlinear; Pattern formation; Spatiotemporal chaos; Superconductivity**

Further Reading

Aranson, I.S. & Kramer, L. 2002. The world of the complex Ginzburg–Landau equation. *Reviews of Modern Physics*, 74: 99–143

Arecchi, F.T., Boccaletti, S. & Ramazza, P. 1999. Pattern formation and competition in nonlinear optics. *Physics Reports*, 318: 1–83

Atai, J. & Malomed, B.A. 1998. Exact stable pulses in asymmetric linearly coupled Ginzburg–Landau equations. *Physics Letters* A, 246: 412–422

Cross, M.C. & Hohenberg, P.C. 1993. Pattern-formation outside of equilibrium. *Reviews of Modern Physics*, 65: 851–1112

Ginzburg V.L. & Landau, L.D. 1950. On the theory of superconductivity. *Zhurnal Eksperimentalnoy i Teoreticheskoy Fiziki (USSR)*, 20: 1064–1082 (in Russian) [English translation: in *Men of Physics*, vol. 1. 1965. Oxford: Pergamon Press, pp. 138–167]

Ipsen, M., Kramer, L. & Sorensen, P.G. 2000. Amplitude equations for description of chemical reaction-diffusion systems. *Physics Reports*, 337: 193–235

Lega, J. 2001. Traveling hole solutions of the complex Ginzburg–Landau equation: a review. *Physica* D, 152: 269–287

Matthews, P.C. & Cox, S.M. 2000. Pattern formation with a conservation law. *Nonlinearity*, 13: 1293–1320

Petviashvili, V.I. & Sergeev, A.M. 1984. Spiral solitons in active media with an excitation threshold. *Doklady Akademii Nauk SSSR*, 276: 1380–1384 (in Russian)

Tinkham, M. 1996. *Introduction to Superconductivity*. New York: McGraw-Hill

COMPLEXITY, MEASURES OF

See **Algorithmic complexity**

CONDENSATES

See **Bose–Einstein condensation**

CONLEY INDEX

In a five-page paper published in the Proceedings of the 1970 International Congress of Mathematicians (Conley, 1971), Charles Conley gave the first definition of his index. For context, he chose the phase space to be a compact connected metric space X and F to be the space of flows on X with the compact open topology. (A flow is a continuous function such that $f(x, 0) = x$, $f(x, t + s) = f(f(x, t), s)$ for all choices of x, t, s.)

For a compact subset Y of X define the set $\mathrm{Inv}(Y, f) = \{y \in Y : f(y, t) \in Y \text{ for all } t\}$. This set is the maximal invariant set contained in Y. An isolating neighborhood for a flow f is a compact subset N of X such that $\mathrm{Inv}(N, f)$ is contained in the interior of N. The set, $\mathrm{Inv}(N, f)$, is the isolated invariant set associated with N.

It is easy to show that isolating neighborhoods persist. If N is an isolating neighborhood for f, then there exists a neighborhood U of f in F such that N is an isolating neighborhood for every flow g in U. Suppose that M is an isolating neighborhood for g. If $\mathrm{Inv}(M, g) = \mathrm{Inv}(N, g)$ then the isolated invariant set $\mathrm{Inv}(M, g)$ is said to be a local continuation of the set $\mathrm{Inv}(N, f)$. Two invariant sets $\mathrm{Inv}(N, f)$ and $\mathrm{Inv}(M, g)$ are related by continuation if there is a finite sequence of local continuations linking one to the other.

For a flow f and an arbitrary compact subset W of X, one defines forward and backward exit time functions from W into the extended real line $[-\infty, \infty]$ as follows:

$$t^+(x) = \sup\{t \geq 0 : f(x, [0, t]) \subset W\},$$
$$t^-(x) = \inf\{t \leq 0 : f(x, [t, 0]) \subset W\}.$$

Certain subsets of W are associated with these functions. The forward asymptotic set is the set $A^+ = \{x : t^+(x) = \infty\}$, and the backward asymptotic set is the set $A^- = \{x : t^-(x) = -\infty\}$. Note that the maximal invariant set $\mathrm{Inv}(W, f)$ is the set $A^+ \cap A^-$. Forward and backward exit sets are the sets $W^\pm = \{x : t^\pm(x) = 0\}$.

An isolating block B for a flow f is a special type of isolating neighborhood with the following property. The boundary of B is the union of the exit sets B^+ and B^-. The intersection of these sets is the "tangency" set τ of boundary points that immediately exit in both time directions. Thus, a block has no internal tangencies where an orbit comes to the boundary from inside the block and does not exit.

If B is an isolating block for a flow f, one can show that the exit time functions are continuous. Using these functions and the flow, one may define deformation retractions of $B - A^+$ onto B^- and $B - A^-$ onto B^+. For example, a retraction r of $B - A^-$ onto B^+ is defined by the formula $r(x) = f(x, t^+(x))$ (This property was first used by Wazewski (1954).)

Suppose that f is a flow on R^3 and also that T is a solid torus that is an isolating block for the flow. Suppose that the exit set is a disk D on the boundary of T. Then, the invariant set $\mathrm{Inv}(T, f)$ must not be empty. If it were empty, one could define a deformation retraction of T onto D, which is impossible.

Under these definitions, the following theorem is fundamental (Conley & Easton, 1971). Given an isolating neighborhood N for a flow f, there exists an isolating block B contained in N such that $\mathrm{Inv}(N, f) = \mathrm{Inv}(B, f)$.

We use this theorem to define the Conley index of the set $\mathrm{Inv}(N, f)$ to be the homotopy types of the pair of quotient spaces $[B/B^+]$ and $[B/B^-]$. These spaces are obtained by collapsing the exit sets to points and using the quotient topology on the resulting spaces.

Consider, for example, the flow f defined on R^2 by $f((x, y), t) = (\mathrm{e}^{-t}x, \mathrm{e}^{t}y)$. This flow has the origin as a saddle point. Let B be a square centered at the origin. Then B is an isolating block for the flow. The exit set consists of the top and bottom sides of B. The quotient space $[B/B^-]$ has the homotopy type of the pair of spaces consisting of a circle and a point on this circle.

The Conley Index has the following properties (Conley, 1978; Herman et al., 1988; Smoller, 1983; Easton, 1998; Hofer & Zehnder, 1995; Mischaikow, 2002):

(i) The Conley index is well defined. Thus it is independent of the choice of block B.

(ii) If $\mathrm{Inv}(N, f)$ and $\mathrm{Inv}(M, g)$ are two isolated invariant sets that are related by continuation, then they have the same Conley index.

(iii) The index $[B/B^-]$ of a saddle point for a smooth flow on a manifold is a sphere together with a point on the sphere. The sphere has the same dimension as that of the unstable manifold of the saddle point. Thus, the Conley index is a generalization of the Morse index of a saddle point.

(iv) The index of two disjoint isolated invariant sets is the "sum" or "join" of their indices.

Traveling Waves

One of the early applications of the Conley index was to find traveling waves for reaction-diffusion equations (Smoller, 1983). We will use as an example the FitzHugh–Nagumo (FN) equations $u_t = \varepsilon v$, $v_t = v_{xx} + f(v) - u$, which are a simplification of the Hodgkin–Huxley equations used to model nerve impulses. The parameter ε is assumed to be small, and one seeks a solution of the form $U(x, t) = u(s)$, $V(x, t) = v(s)$, where $s = x + \theta t$ and θ is the wave velocity.

Substituting the trial solutions into the FN equations, one obtains the following system of ordinary differential equations: $\mathrm{d}u/\mathrm{d}s = (\varepsilon/\theta)v$, $\theta\, \mathrm{d}v/\mathrm{d}s = \mathrm{d}^2v/\mathrm{d}s^2 + f(v) - u$, where $f(v)$ is assumed have the general shape of a cubic equation that is decreasing, increasing, and then decreasing. To be specific, we take $f(v) = -v(v-1)(v-2)$. The corresponding first-order system of ordinary differential equations is $u' = \sigma v$, $v' = w$, $w' = \theta w + u - f(v)$, where $\sigma = \varepsilon/\theta$. Our goal is to find a periodic solution of this system for small values of the parameters sigma and theta, and thereby to find a periodic solution of the FN equations.

One can completely understand the phase portrait of system when the parameters σ and θ are set to zero. In this case, $u(t)$ is constant and the equations for v, w form a Hamiltonian system with Hamiltonian $H(v, w) = w^2/2 + F(v) - uv$ with $F(v) = \int_0^u f(r)\mathrm{d}r = -v^2(v-2)^2/4$. The phase portrait in the $u = 0$ plane has saddle points at $(v, w) = (0, 0)$, $(v, w) = (2, 0)$ and a center at $(v, w) = (1, 0)$. The saddle points are connected by heteroclinic orbits implicitly defined by the equation $H(v, w) = 0$, $0 < v < 2$. Note that the set of equilibrium points for the system is the set $\{(u, v, w) : u = f(v), w = 0\}$.

Next consider the system with $\sigma = 0$, $\theta > 0$. For this system, we have $(d/dt)H(u(t), v(t), w(t)) = \theta w^2(t)$. Let $a(u) < b(u) < c(u)$ denote the three solutions of the cubic equation $u - f(u) = 0$. One can show that there are values $0 < u_1 < u_2$ such that for $j = 1, 2$, there is a heteroclinic solution joining the equilibrium points $(u_j, a(u_j), 0)$ and $(u_j, c(u_j), 0)$ in the plane $u = u_j$. We now have a cycle consisting of the two heteroclinic orbits together with arcs of equilibrium points $\{(u, a(u), 0) : u_1 < u < u_2\}$ and $\{(u, c(u), 0) : u_1 < u < u_2\}$. This cycle is an invariant set for the system. However, the cycle is not isolated since it intersects the set of equilibrium points in two arcs. The two arcs of equilibrium points may be viewed as normally hyperbolic invariant manifolds, whose stable and unstable manifolds intersect transversally along the two heteroclinic orbits.

Finally, consider the system for small positive values of sigma. In this case, u is increasing when u is positive and decreasing when negative. The hard part is to construct an isolating block, which topologically is a solid torus containing the cycle. The transversal intersection noted above is essential to this construction. The cycle is no longer invariant when sigma is positive. However, one shows that the isolating block must contain a periodic solution of the full system of equations. The periodic solution thus constructed may be viewed as a periodic traveling wave solution of the FN equations.

Applications to Discrete Dynamical Systems

Consider the discrete dynamics generated by iterating a homeomorphism f of a compact metric space X. It is natural to study orbits with "errors" such as truncation or round-off errors in numerical algorithms. Thus, an ε-chain for f is a finite sequence $(y_0, y_1, y_2, \ldots)$ such that $d(f(y_n), y_{n+1}) \leq \varepsilon$ where $d(x, y)$ is the distance function on X. Conley (1978) and Bowen (1975) both understood the importance of studying orbits with errors. Bowen asked when such an orbit could be shadowed by a true orbit of the system.

Conley defined the ε-chain recurrent set $\mathrm{CR}(f, \varepsilon)$ to be the set of points that are contained in periodic ε-chains of length at least 2. The chain recurrent set is the set $\mathrm{CR}(f) = \cap\{\mathrm{CR}(f, \varepsilon){:}\varepsilon > 0\}$. Points in $\mathrm{CR}(f)$ are chain equivalent if for any positive epsilon there is a periodic ε-chain containing both points. He showed that every orbit uniformly approaches a unique chain equivalence class in $\mathrm{CR}(f)$. This result is known as the Conley decomposition theorem, and it generalizes Smale's decomposition of an Axiom A system into basic sets (Bowen, 1975).

An isolating block for f is a compact subset N of X such that whenever $x, f(x), f^2(x) \in N$, then $f(x)$ is contained in the interior of N. This is something like having no internal tangencies. The exit set of N is the set $E = \{x \in N : f(x) \ni \mathrm{int}(N)\}$. Because N is a block, the exit set is compact. Define an equivalence relation on N making all points of E equivalent and all points not in E equivalent only to themselves. Let $N^{\#}$ denote the space of equivalence classes with the quotient topology obtained by projecting N onto $N^{\#}$ by sending a point to the equivalence class to which it belongs. Let $E^{\#}$ denote the image of E. The index space of N is the pair $(N^{\#}, E^{\#})$. Define an index map $f^{\#} : N^{\#} \to N^{\#}$ by $f^{\#}(x) = E^{\#}$ if $x = E^{\#}$ or $f(x) \in E^{\#}$. Otherwise, define $f^{\#}(x) = f(x)$. Note that the index map is continuous. The pair $(N^{\#}, f^{\#})$ plays the role of the Conley index in this context. If the index map is not homotopic to a constant, then one can prove that the set $\mathrm{Inv}(N, f)$ is non-empty.

Smale's horseshoe map in the plane may be used as an example. Suppose that a rectangle B is mapped across itself so that the image crosses the rectangle in a horizontal strip, then curves back and crosses the rectangle again in another horizontal strip above the first. In this case, the exit set consists of three vertical strips, one in the center of the rectangle and the other two on the left and right sides containing the vertical edges of B. The index space $[B/B^-]$ has the homotopy type of a figure of eight and the index map is non-trivial.

Sequences of compact sets (called "windows") may sometimes be constructed to contain an orbit with errors. If each window in the sequence is "correctly" mapped across the next one, then a true orbit runs through the sequence of windows and shadows the orbit with errors (Easton, 1998).

ROBERT W. EASTON

See also **Anosov and Axiom-A systems; FitzHugh–Nagumo equation; Horseshoes and hyperbolicity in dynamical systems**

Further Reading

Bowen, R. 1975. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, New York: Springer

Conley, C.C. 1971. On the continuation of invariant sets of a flow. In *Proceedings of the International Congress of Mathematicians 1970*, Paris: Gauthiers-Villars, pp. 909–913

Conley, C.C. & Easton, R.W. 1971. Isolated invariant sets and isolating blocks. *Transactions of the American Mathematical Society* 158: 35–61

Conley, C.C. 1978. *Isolated Invariant Sets and the Morse Index*, Providence, RI: American Mathematical Society.

Easton, R. 1998. *Geometric Methods for Discrete Dynamical Systems*, Oxford and New York: Oxford University Press.

Herman, M., McGehee, R., Moser, J. & Zehnder, E. (editors). 1988. Charles Conley Memorial Volume, Special Issue of *Ergodic Theory and Dynamical Systems*, 8

Hofer, H. & Zehnder, E. 1995. Symplectic invariants and Hamiltonian dynamics. *The Floer Memorial Volume Progress in Mathematics*, vol. 133, Basel: Birkhäuser.

Mischaikow, K. 2002. Topological techniques for efficient rigorous computations in dynamics, *Acta Numerica*,vol. 11, 435–477

Smoller, J. 1983. *Shock Waves and Reaction-diffusion Equations*, New York: Springer.

Wazewski, T. 1954. Sur un principe topologique de l'examen de l'allure asymptotique des integrales des equations differentielles, *Proceedings of the International Congress of Mathematicians*, vol. 3, 132–139

CONSTANTS OF MOTION AND CONSERVATION LAWS

Although nonlinear spatiotemporal processes may be very complicated, they frequently obey simple constraints in the form of conservation laws. It is sometimes possible to construct one or several constants of motion (also called dynamical invariants, DIs), in the form of spatial integrals of local densities expressed in terms of the physical fields and their derivatives, which are conserved in time, as a consequence of the underlying dynamics. Such commonly known conserved quantities as energy, momentum, and angular momentum belong to this class.

Typically, the existence of conservation laws can be established if the underlying dynamics is dissipation-free; however, a specific DI may sometimes also exist in dissipative systems. Examples of the latter are provided by the diffusion equation, $u_t = u_{xx}$, and its important nonlinear counterparts in the form of the Burgers equation,

$$u_t = u_{xx} + uu_x, \tag{1}$$

and Cahn–Hilliard equation,

$$u_t + \left(u - u^3 + u_{xx}\right)_{xx} = 0 \tag{2}$$

(the subscripts stand for the partial derivatives). They all conserve $\int_{-\infty}^{+\infty} u(x,t)\mathrm{d}x$, which is simply the total mass of the substance in the case of diffusion.

A more sophisticated example of the "dissipative conservation" occurs in physically important models based on the nonlinear Schrödinger (NLS) equation with special additional terms:

$$\mathrm{i}u_t + u_{xx} + F'(|u|^2)u = \varepsilon Q, \tag{3}$$

where the function F' describes conservative nonlinearity of the medium (the prime stands for the derivative with respect to the argument of F; in particular, $F = \left(|u|^2\right)^2$ corresponds to the most generic case of the cubic NLS equation), ε is a real parameter, and the "special perturbation" is, for instance, the nonlinear Landau-damping term in the NLS equation for Langmuir waves in plasmas,

$$Q = -u \int_{-\infty}^{+\infty} \mathrm{d}x' \left(x - x'\right)^{-1} |u(x')|^2, \tag{4}$$

or the stimulated Raman scattering term in the equation for electromagnetic waves in nonlinear optical fibers, $Q = \left(|u|^2\right)_x u$. While these terms are dissipative ones, the corresponding perturbed NLS equation conserves the single DI, namely, the total *wave action* (alias "number of quanta"),

$$N = \int_{-\infty}^{+\infty} |u(x,t)|^2 \mathrm{d}x. \tag{5}$$

In the general case, equations that govern the dissipation-free spatiotemporal dynamics can be derived from the underlying action functional $S\left\{u^{(n)}\right\}$: $\delta S/\delta u^{(n)} = 0$, where $u^{(n)}(\boldsymbol{r}, t)$ is the nth field variable, $\boldsymbol{r}$ is the set of the spatial coordinates, $\delta/\delta u^{(n)}$ stands for the variational (functional) derivative, and the action is expressed in terms of the Lagrangian density $\mathcal{L}$, so that

$$S = \int\int \mathcal{L}\left(u^{(n)}, \nabla u^{(n)}, u_t^{(n)}\right) \mathrm{d}\boldsymbol{r}\mathrm{d}t. \tag{6}$$

For instance, the density

$$\mathcal{L} = (1/2)\left[u_t^2 - (\nabla u)^2\right] - F(u) \tag{7}$$

yields a nonlinear Klein–Gordon (NKG) equation for a single real field u,

$$u_{tt} - \nabla^2 u + F'(u) = 0. \tag{8}$$

The fundamental nature of DIs in Lagrangian systems is established by a theorem that was published by Emmy Noether in 1918: any *continuous symmetry* of the system, that is a family of transformations of the field variables, spatial coordinates, and time, which depend on an arbitrary continuous parameter ξ and leave the action invariant, generates a constant of motion. If the infinitesimal symmetry transformation is written in the form

$$u^{(n)} \to u^{(n)} + U_n\,\mathrm{d}\xi,\ \boldsymbol{r} \to \boldsymbol{r} + \boldsymbol{R}\,\mathrm{d}\xi,$$
$$t \to t + T\,\mathrm{d}\xi, \tag{9}$$

then the main result following from the Noether theorem is the continuity equation $\mathcal{I}_t + \nabla\cdot\boldsymbol{J} = 0$, with the following density and current:

$$\mathcal{I} = \sum_n \frac{\partial\mathcal{L}}{\partial\left(u_t^{(n)}\right)}\left(\boldsymbol{R}\cdot\nabla u^{(n)} + Tu_t^{(n)} - U_n\right)$$
$$+ T\mathcal{L}, \tag{10}$$

$$\boldsymbol{J} = \sum_n \frac{\partial\mathcal{L}}{\partial\left(\nabla u^{(n)}\right)}\left(\boldsymbol{R}\cdot\nabla u^{(n)} + Tu_t^{(n)} - U_n\right)$$
$$+ \boldsymbol{R}\mathcal{L}, \tag{11}$$

($\partial/\partial\left(\nabla u^{(n)}\right)$ is realized as a vector with the components $\partial/\partial\left(u_x^{(n)}\right)$, $\partial/\partial\left(u_y^{(n)}\right)$, $\partial/\partial\left(u_z^{(n)}\right)$). Then, assuming, as usual, that the fields disappear at $|\boldsymbol{r}|\to\infty$, the continuity equation immediately yields the conservation law in the form of $\mathrm{d}I/\mathrm{d}t = 0$, with $I \equiv \int\mathcal{I}\,\mathrm{d}\boldsymbol{r}$. A detailed derivation of this fundamental result can be found in the book by Bogoliubov & Shirkov (1973); for discussion of the Noether theorem in various contexts, see also Sulem & Sulem (1999), and Whitham (1974) If the underlying equations are complex, the Lagrangian density and all the DIs are nevertheless real.

The obvious invariance of the action against arbitrary temporal and spatial shifts, which are described by Equation (9) with $U_n = 0$ and, respectively, $\boldsymbol{R} = 0$, $T = 1$, or $\boldsymbol{R}_j = \boldsymbol{e}_j$, $T = 0$ ($\boldsymbol{e}_j$ is the unit vector corresponding to the jth spatial coordinate) gives rise to the conservation of the energy (Hamiltonian) H and momentum $\boldsymbol{P}$. For the important classes of the NKG and multidimensional NLS, Equations (8) and (3), Equation (10) yields

$$H_{\mathrm{NKG}} = \int \left[\frac{1}{2}\left(u_t^2 + (\nabla u)^2\right) + F(u)\right] \mathrm{d}\boldsymbol{r},$$

$$\boldsymbol{P}_{\mathrm{NKG}} = -\int u_t \nabla u \ \mathrm{d}\boldsymbol{r}, \tag{12}$$

$$H_{\mathrm{NLS}} = \int \left[|\nabla u|^2 - F(|u|^2)\right] \mathrm{d}\boldsymbol{r},$$

$$\boldsymbol{P}_{\mathrm{NLS}} = \mathrm{i}\int \left(u\nabla u^* - u^*\nabla u\right) \mathrm{d}\boldsymbol{r}, \tag{13}$$

where * stands for the complex conjugation (the transition from the Lagrangian to Hamiltonian density as per Equation (10) in the case of the temporal-shift invariance is called the Legendre transformation). The invariance against rotations in the three-dimensional space leads to the conservation of the angular momentum,

$$\boldsymbol{M} = \int (\boldsymbol{r} \times \mathcal{P})\ \mathrm{d}\boldsymbol{r}, \tag{14}$$

where $\mathcal{P}$ is the density in the expressions for the momentum in Equations (12)–(14); in the two-dimensional case, there is only one component of the conserved angular momentum. Additionally, in the NLS-type equations, the invariance against the phase shift (alias *gauge invariance*), $u \to u \exp(\mathrm{i}\xi)$ with an arbitrary constant ξ, generates the conservation of the above-mentioned wave action (5), which is $\int |u|^2\, \mathrm{d}\boldsymbol{r}$ in the multidimensional case.

Another important class of models in one dimension is based on equations of the Korteweg–de Vries (KdV) type for a real function $u(x,t)$,

$$u_t + u_{xxx} + F''(u)u_x = 0 \tag{15}$$

(the most important cases of the KdV equation proper and modified KdV equation correspond to $F = u^3$ and $F = u^4$, respectively). The Lagrangian representation of Equation (15) is possible in terms of the potential field v, defined so that $v_x \equiv u$, but the Hamiltonian and momentum are expressed solely in terms of the original field u,

$$H_{\mathrm{KdV}} = \int_{-\infty}^{+\infty} \left[\frac{1}{2}u_x^2 - F(u)\right] \mathrm{d}x,$$

$$P_{\mathrm{KdV}} = \int_{-\infty}^{+\infty} u^2\, \mathrm{d}x. \tag{16}$$

The invariance of the action, written in terms of the potential v, against the arbitrary shift $v \to v + \xi$ additionally generates the conservation of the "mass," $\int_{-\infty}^{+\infty} u\, \mathrm{d}x$.

Besides being a dynamical invariant, the Hamiltonian gives rise to a canonical representation of the equation(s) in the Hamiltonian form, which is dual to the Lagrangian representation. In particular, for the complex and real equations of the NLS and KdV types, respectively, this representation is

$$u_t = -\mathrm{i}\frac{\delta H}{\delta u^*}, \quad u_t = \frac{\partial}{\partial x}\frac{\delta H}{\delta u}. \tag{17}$$

The conservation of H itself and the conservation of the mass in the KdV-type equations are immediate consequences of the general form of Equations (17). The conservation of the wave action in the NLS-type equation is also a consequence of its representation in the form of Equations (17).

If a multicomponent Lagrangian system possesses an additional ("isotopic") symmetry against linear transformations of the components, this also gives rise to a specific DI. An important example is a system of coupled NLS equations of Manakov's type (Manakov, 1973)

$$\mathrm{i}\begin{pmatrix} u_t \\ v_t \end{pmatrix} + \nabla^2 \begin{pmatrix} u \\ v \end{pmatrix} + F'\left(|u|^2 + |v|^2\right)\begin{pmatrix} u \\ v \end{pmatrix} = 0, \tag{18}$$

which are invariant against rotation in the plane of (u, v). In this case, Equation (10) gives rise to the DI ("isotopic spin") in the form

$$S = \mathrm{i}\int \left(uv^* - u^*v\right) \mathrm{d}\boldsymbol{r}. \tag{19}$$

A very special situation arises for the DIs in the case of integrable equations, that is, those that are amenable to the application of the inverse scattering transform (IST) (Ablowitz & Segur, 1981; Newell, 1984; Zakharov et al., 1984). Integrable equations have an infinite set of hidden dynamical symmetries, which, unlike the above-mentioned elementary invariances against temporal and spatial shifts, spatial rotations, phase shift, etc., do not have a straightforward meaning. In compliance with the Noether theorem, each hidden symmetry generates the corresponding DI, which is an integral expression with a density that, unlike those corresponding to the elementary DIs (see Equations (12), (4), and (16), involves higher-order derivatives. For instance, in the integrable KdV equation (15) with $F = u^3$, the first higher-order DI is

$$I = \int_{-\infty}^{+\infty} \left(u_{xx}^2 + 5u^2 u_{xx} + 5u^4\right) \mathrm{d}x. \tag{20}$$

In fact, it was an empirical discovery of several higher-order DIs in the KdV equation that was a major incentive for the study that had resulted in the discovery of the IST technique.

The IST provides a systematic method to derive the infinite set of the DIs in terms of the corresponding scattering data, into which the original wave field is mapped to make the hidden integrability explicit (Ablowitz & Segur, 1981; Newell, 1984; Zakharov et al., 1984). The use of the scattering data makes it possible to explicitly introduce a full system of the action-angle variables for the integrable equations, and demonstrate that the infinite set of the action variables is in one-to-one correspondence with the set of the DIs. It is also possible to prove; that all the DIs are *in involution* among themselves; that is, the Poisson bracket between any two DIs, defined as per the corresponding symplectic (Hamiltonian) structure, is zero. Thus, integrable equations are direct counterparts, for the case of the infinite number of degrees of freedom, of finite-dimensional Hamiltonian systems that are Liouville-integrable; that is, with a set of DIs that are in involution, their number being equal to the number of the degrees of freedom.

The presence of the infinite set of the DIs in the integrable equations helps to understand such a well-known property as the completely elastic character of collisions between solitons (Ablowitz & Segur, 1981; Newell, 1984; Zakharov et al., 1984): roughly speaking, the necessity to satisfy the infinite set of the conservation laws leaves no room for changes of the solitons, except for phase shifts. On the other hand, some equations amenable to the application of the IST technique, such as, for instance, the standard three-wave system,

$$\left(u_{1,2}\right)_t + c_{1,2}\left(u_{1,2}\right)_x = \mathrm{i}u_{2,1}^{*}u_3, \quad (u_3)_t = \mathrm{i}u_1u_2, \qquad (21)$$

where $u_{1,2}$ and u_3 are, respectively, the "daughter" and pump waves, and $c_{1,2}$ are group velocities, feature nontrivial "soliton reactions"—for instance, a spontaneous split of a pump-wave soliton into separating daughter ones. This possibility is explained by the fact that the above-mentioned one-to-one correspondence between the infinite sets of the degrees of freedom and DIs does not really hold for these equations: the set of the DIs is not "infinite enough" (Fokas & Zakharov, 1992). Such equations are sometimes called "solvable," to stress their difference from the genuinely integrable ones (integrable in the sense of Liouville, as generalized to systems with infinitely many degrees of freedom).

Integrable lattice (discrete) models feature another important property: due to the lack of the continuous translational invariance, lattice systems lack the momentum conservation. Nevertheless, integrable lattice models do possess a conserved momentum, due to their hidden symmetry. For example, the Ablowitz–Ladik equation,

$$2\mathrm{i}\frac{\mathrm{d}u_n}{\mathrm{d}t}+(u_{n+1}+u_{n-1}-2u_n)+|u_n|^2(u_{n+1}+u_{n-1}) = 0, \qquad (22)$$

which is an integrable discretization of the cubic NLS equation, conserves the real momentum in the form of

$$P = \sum_{n=-\infty}^{+\infty} (\psi_n\psi_{n+1}^{*} - \psi_n^{*}\psi_{n+1}). \qquad (23)$$

In fact, conservation of the momentum is a specific integrability feature of discrete models, in contrast to continuum ones.

Elementary DIs find specific applications in systems perturbed by small nonconservative terms to order ε. In that case, the conservation laws no longer hold; however, using evolution (balance) equations for the former DI(s) is a convenient way to derive effective equations of motion for solitons (or other collective nonlinear excitations) in the weakly perturbed model. For instance, in the cubic NLS equation (3) with the above-mentioned terms (Kerr and stimulated Raman scattering ones), $F = \left(|u|^2\right)^2$ and $Q = \left(|u|^2\right)_x u$, an exact soliton solution with arbitrary amplitude η and velocity c, in the case of $\varepsilon = 0$, is

$$\begin{aligned} u_{\mathrm{sol}} &= \eta\ \mathrm{sech}\,(\eta\,(x - ct)) \\ &\quad \times \exp\left[\mathrm{i}(c/2)x + \mathrm{i}\left(\eta^2 - c^2\right)t\right]. \end{aligned} \qquad (24)$$

In the presence of small $\varepsilon > 0$, the wave action (4) remains a DI (see above), and the balance equation for the formerly conserved momentum P_{NLS}, see Equation (13), is

$$\frac{\mathrm{d}P}{\mathrm{d}t} = 2\varepsilon\int_{-\infty}^{+\infty}\left[\left(|u|^2\right)_x\right]^2\mathrm{d}x. \qquad (25)$$

Substitution of the unperturbed soliton (24) into Equation (25), and into the conservation of the wave action, yields evolution equations for the amplitude and velocity:

$$\frac{\mathrm{d}\eta}{\mathrm{d}t} = 0, \quad \frac{\mathrm{d}c}{\mathrm{d}t} = \frac{16}{15}\varepsilon\eta^4. \qquad (26)$$

For further details and references, see the review by Kivshar & Malomed (1989).

Boris Malomed

See also **Hamiltonian systems; Integrability; Integrable lattices; Inverse scattering method or transform; Korteweg–de Vries equation; *N*-wave interactions; Symmetry groups**

Further Reading

Ablowitz, M. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM

Bogoliubov, N.N. & Shirkov, D.V. 1973. *Introduction to the Theory of Quantized Fields*, Moscow: Nauka (in Russian); English translation, 2nd edition: New York: Wiley, 1980

Fokas, A.S. & Zakharov, V.E. 1992. The dressing method and nonlocal Riemann-Hilbert problems. *Journal of Nonlinear Science*, 2: 109–134

Kivshar, Y.S. & Malomed, B.A. 1989. Dynamics of solitons in nearly integrable systems. *Reviews of Modern Physics*, 61: 763–915

Manakov, S.V. 1973. Theory of two-dimensional stationary self-focusing of electromagnetic waves. *Zhurnal Eksperimentalnoy i Teoreticheskoy Fiziki*, 65: 505–516 (in Russian); translated in *Soviet Physics—Journal of Experimental and Theoretical Physics*, 38: 248 (1974)

Newell, A.C. 1984. *Solitons in Mathematics and Physics*, Philadelphia: SIAM

Sulem, C. & Sulem, P.-L. 1999. *The Nonlinear Schrödinger Equation*, New York: Springer

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

Zakharov, V.E., Manakov, S.P., Novikov, S.P. & Pitaevskii, L.P. 1984. *Theory of Solitons*, New York: Consultants Bureau

CONTINUITY EQUATION

See **Constants of motion and conservation laws**

CONTINUOUS SPECTRUM

See **Inverse scattering method or transform**

CONTINUUM APPROXIMATIONS

In general, a physical system belongs to one of three broad classes: (i) media with distributed parameters (electromagnetic fields, fluids, or liquid), (ii) discrete media (crystal lattices, polymers, or macromolecules), and (iii) artificial periodic systems (layered structures, lattices of nano-dots, or Josephson arrays). In the first class, the dynamics of a system is described by partial differential equations (PDE) for the field variable $u(x,t)$ and in the other two cases by discrete differential equations (DDE) for the field variable at the lattice sites $u(n,t)=u_n(t)$. For simplicity, only one-dimensional (1-d) models are discussed here.

Well-known examples of discrete nonlinear dynamical systems include the following:

- The discrete 1-dimensional elastic chain with a nonlinear interaction between the nearest neighbors (generalized Fermi–Pasta–Ulam (FPU) model), whose equation of motion reads

$$m\frac{\mathrm{d}^2u_n}{\mathrm{d}t^2}=\varphi'(u_{n+1}-u_n)-\varphi'(u_n-u_{n-1}), \quad (1)$$

  where u_n is the displacement of the nth atom in a chain; a prime indicates the derivative with respect to the argument, and $\varphi(u_n-u_{n-1})$ is the energy of the interatomic interaction (Fermi et al., 1955). The particular choice $\varphi(\xi)=A\xi^2/2+\alpha\xi^3/3+\beta\xi^4/4$ and $\varphi(\xi)=c\exp(-p\xi)+q\xi$ represents α-FPU (for $\beta=0$), β-FPU (for $\alpha=0$), and Toda models.

- The discrete 1-d chain with linear interatomic interaction exposed to a nonlinear external field (discrete Frenkel–Kontorova (FK) model) with the following equation of motion:

$$m\frac{\mathrm{d}^2u_n}{\mathrm{d}t^2}=A(u_{n+1}+u_{n-1}-2u_n)-w'(u_n), \quad (2)$$

  where $w(u_n)$ is the nonlinear on-site external potential. The particular choice $w=U(1-\cos(2\pi u_n/a))$, where a is the interatomic distance, corresponds to the traditional FK-model (Frenkel & Kontorova, 1939).

- 1-d photonic crystals (periodic arrays of optical waveguides) or the discrete spin lattice, which may be described in the context of the discrete nonlinear Schrödinger equation (DNLS)

$$\begin{aligned}&\mathrm{i}\frac{\mathrm{d}\psi_n}{\mathrm{d}t}+B(\psi_{n+1}+\psi_{n-1}-2\psi_n)\\&+F(|\psi_n|^2)\psi_n=0,\end{aligned} \quad (3)$$

  where in the simplest case F is a linear function of the argument. ψ_n denotes the value of the effective field at the nth element of the discrete system, which can be assigned different physical meanings for various applications.

Even in the 1-d case, the solution of the nonlinear DDE poses a fairly complicated mathematical problem and only a few of them can be solved exactly (the Toda and Ablowitz–Ladik equations). Thus, it is often easier to study discrete problems in the "continuum approximation" (CA) within the framework of PDEs. Clearly, some information about nonlinear dynamics of discrete systems is lost, and some phenomena cannot be described in this continuum limit; but in the long wave limit, this approach provides a good qualitative and even quantitative agreement with the results for a discrete system investigation.

In the CA, the discrete number of the atom site n is replaced with the continuous coordinate $x: na\rightarrow x$, with a being the interatomic equilibrium distance or the period of mesoscopic periodical structure, and $u_n(t)=u(na,t)$ is replaced with $u(x,t)$. The finite differences $u_{n\pm1}-u_n$ have to be expanded in Taylor series

$$\begin{aligned}u_{n\pm1}-u_n&=\pm a\frac{\partial u}{\partial x}+\frac{a^2}{2}\frac{\partial^2u}{\partial x^2}\\&\pm\frac{a^3}{6}\frac{\partial^3u}{\partial x^3}+\frac{a^4}{24}\frac{\partial^4u}{\partial x^4}\cdots.\end{aligned} \quad (4)$$

This expansion is valid under the condition

$$|(u_{n+1}-u_n)/u_n|\ll 1. \quad (5)$$

For linear waves of the form $u_n=u_0\sin(kna-\omega t)$ this expansion agrees with the long wavelength approximation $ak\ll 1$, where k is a wave number.

Substitution of expansion (4) in DDE (1) in the leading approximation yields for the α-FPU and Toda models the Boussinesq equation

$$m\frac{\partial^2 u}{\partial t^2} - Aa^2\frac{\partial^2 u}{\partial x^2} - \frac{Aa^4}{12}\frac{\partial^4 u}{\partial x^4}$$
$$-2\alpha a^3\frac{\partial u}{\partial x}\frac{\partial^2 u}{\partial x^2} = 0. \qquad (6)$$

(In the case of the β-FPU model, the modification of this equation with the nonlinear term $3\beta a^4(\partial u/\partial x)^2(\partial^2 u/\partial x^2)$ is obtained.)

Equations (2) for discrete FK model within continuum limit are transformed in the leading approximation into the nonlinear Klein–Gordon (NKG) equation

$$m\frac{\partial^2 u}{\partial t^2} - A\,a^2\frac{\partial^2 u}{\partial x^2} + w'(u) = 0. \qquad (7)$$

In the particular case of the periodic external potential $w = U(1 - \cos u_n)$, one obtains the sine-Gordon (SG) equation. Finally, the CA for DNLS equation (3) reduces to the usual partial differential nonlinear Schrödinger equation

$$\mathrm{i}\frac{\partial\psi}{\partial t} + Ba^2\frac{\partial^2\psi}{\partial t^2} + F(|\psi|^2)\psi = 0. \qquad (8)$$

Examples (6)–(8) demonstrate that a different number of terms in expansion (4) are to be taken into account in different situations. In the case of Equation (1), the dispersion relation is

$$\omega = 2\sqrt{(A/m)}\,\sin ak/2, \qquad (9)$$

which is of acoustic type with a weak dispersion in the limit $k \to 0$. Within the CA, this weak dispersion is governed by the fourth spatial derivative alone in expansion (4), and it is necessary to retain the dispersion term in (6). The dispersion relation for Equation (6) is consistent with the exact formula (9) only in the long wave limit $ak \ll 1$.

Unfortunately, the dispersion term $-(Aa^4/12)\,\partial^4 u/\partial x^4$ in (6) necessitates an additional boundary conditions for this equation and results in the appearance of additional nonphysical solutions with small frequencies and $ak \sim 1$. To avoid these side effects, a regularization of the expansion over the discreteness parameter a can be performed (Rosenau, 1987). This corresponds to substitution of the relation $\partial^2/\partial x^2 \simeq (m/Aa^2)\partial^2/\partial t^2$ into the dispersion term in (6) and leads to yet another version of the CA for Equation (1) containing the term with mixed derivatives:

$$m\frac{\partial^2 u}{\partial t^2} - Aa^2\frac{\partial^2 u}{\partial x^2} - \frac{ma^2}{12}\frac{\partial^4 u}{\partial x^2\partial t^2}$$
$$-2\alpha a^3\frac{\partial u}{\partial x}\frac{\partial^2 u}{\partial x^2} = 0. \qquad (10)$$

The above estimation of CA application area ($ak \ll 1$) holds for long wave small amplitude envelope solitons in the nonlinear case as well. But in general, this condition can differ for solitons of different types. For example, CA descriptions of Boussinesq solitons of the type $u(x,t) = u(x - vt)$ describe the solutions of FPU model (1) only under the condition $v/v_c - 1 \ll 1$, where $v_c = \sqrt{Aa^2/m}$.

Only the lowest-order terms of expansion (4) have so far been taken into account for discrete systems within the CA. In general, retaining the next terms with higher powers of spatial derivatives exceeds the accuracy of CA. But in some way, such extended versions of the CA also take into account the discreteness of the systems and can lead to interesting and important physical results. Retaining the fourth order derivative in (4) transforms the nonlinear KGE (7) into

$$m\frac{\partial^2 u}{\partial t^2} - Aa^2\frac{\partial^2 u}{\partial x^2} - \frac{Aa^4}{12}\frac{\partial^4 u}{\partial x^4} + w'(u) = 0. \qquad (11)$$

In the case of a sinusoidal external force, the corresponding generalized equation (dispersive SG equation) has steady-state bounded kinks solutions (4π-kink solitons) for some particular values of their velocities (Bogdan et al., 2001). This result obtained within the CA is in agreement with the numerical result for the corresponding discrete system (2) (Alfimov et al., 1993). The inclusion of yet higher terms of expansion (4) in the nonlinear parts of discrete equations in the CA gives rise to the nonlinear dispersion and leads to an existence of exotic solitons such as "compactons" and "peakons."

The CA is not restricted to the long-wave limit. For high-frequency short waves with wave numbers $k \simeq \pi/a$ and $\omega - \omega_{\max} \ll \omega_{\max}$ (where $\omega_{\max} = 2\sqrt{A/m}$ for (1) and (2)), the CA for the slowly varying envelope of antiphase oscillations ($u_n = (-1)^n v_n$) in the β-FPU-model results in a PDE with the Euclidean differential part

$$m\frac{\partial^2 v}{\partial t^2} + Aa^2\frac{\partial^2 v}{\partial x^2} + 4Av + 16\beta v^3 = 0. \qquad (12)$$

The breather solution of this equation (Kosevich & Kovalev, 1975) within the CA describes the "intrinsic modes," which are currently being widely discussed since the pioneering paper of Sievers and Takeno (1988). In more complicated diatomic chains with the gap in the linear waves spectrum at $k = \pi/2a$, the so-called "gap solitons" (breathers with frequencies lying in the gap) can be described in CA for the envelopes of the antiphase oscillations of atoms from two sublattices.

To this point, applications of the CA for DDE models of discrete systems were discussed. Often, the opposite approach is used where the corresponding PDEs are investigated numerically in some discrete schemes as a system of DDE (Dodd et al., 1982). The finite-differences method is one of the most popular in this case: the initial function $u(x,t)$ is defined on the rectangular net of the (x,t) plane at points $x = hn$, $t = h't$. The partial derivatives are replaced by the finite differences

$$\frac{\partial u(x,t)}{\partial x} = \frac{u_{n+1} - u_n}{h}, \quad \frac{\partial^2 u(x,t)}{\partial x^2} = \frac{u_{n+1} + u_{n-1} - 2u_n}{h^2}, \ldots \tag{13}$$

Generally, sampling over all variables is performed, but in some hybrid methods space sampling alone is carried out and the resulting system of ODEs is solved by using standard computer codes.

Space sampling is commonly used for complicated biological systems. In order to simulate the behavior of a single neuron, for example, its continuous structure may be sliced into a large number of small segments (compartments). This procedure is called the "compartmental" approach, and within it the continuous PDEs are replaced by sets of ODEs. The advantage of this modeling approach is that it imposes no restrictions on the properties of each compartment and permits great flexibility at the level of resolution. Compartmental methods make it possible to develop the realistic models that have a close relationship with the relevant experimental data.

ALEXANDER S. KOVALEV

See also **Compartmental models; Delay-differential equations; Discrete nonlinear Schrödinger equations; Dispersion relations; Fermi–Pasta–Ulam oscillator chain; Frenkel–Kontorova model; Partial differential equations, nonlinear; Peierls barrier; Sine-Gordon equation**

Further Reading

Alfimov, G., Eleonskii, V., Kulagin, N. & Mitskevich, N. 1993. Dynamics of topological solitons in models with nonlocal interaction. *Chaos*, 3: 405–414

Bogdan, M., Kosevich, A. & Maugin, G. 2001. Soliton complex dynamics in strongly dispersive medium. *Wave Motion*, 34: 1–26

Dodd, R., Eilbeck, J., Gibbon, J. & Morris, H. 1982. *Solitons and Nonlinear Wave Equations*, London: Academic Press

Eisenberg, H., Silberberg, Y., Marandotti, R., Boyd, A. & Aitchison, J. 1998. Discrete spatial optical solitons in waveguide arrays. *Physical Review Letters*, 81: 3383–3386

Fermi, E., Pasta, J. & Ulam, S. 1955. Studies of nonlinear problems. 1965. *Collected Works of E. Fermi*, vol. II, Chicago: University Chicago Press

Frenkel, J. & Kontorova, T. 1939. On the theory of plastic deformation and twinning. *Physical Journal USSR*, I: 137–149 (originally published in Russia, 1938)

Kosevich, A. & Kovalev, A. 1975. Self-localization of vibrations in 1D anharmonic chain. *Soviet Physics, Journal of Experimental and Theoretical Physics*, 40: 891–896

Rosenau, P. 1987. Dynamics of dense lattices. *Physical Review*, B36: 5868–5876

Sievers, A. & Takeno, S. 1988. Intrinsic localized modes in anharmonic crystals. *Physical Review Letters*, 61: 970–973

CONTOUR DYNAMICS

A wide variety of fluid dynamical problems involve the material advection of a tracer field $q(\boldsymbol{x},t)$, expressed by

$$\frac{\mathrm{D}q}{\mathrm{D}t} \equiv \frac{\partial q}{\partial t} + \boldsymbol{u} \cdot \nabla q = 0, \tag{1}$$

where $\boldsymbol{u}(\boldsymbol{x},t)$ is the fluid velocity. The value of q thus does not change following an infinitesimal element or particle. That is, for a particle at $\boldsymbol{x} = \boldsymbol{X}(\boldsymbol{a},t)$, where $\boldsymbol{a}$ is a vector label (e.g., the initial position of the particle), Equation (1) implies that $q = q(\boldsymbol{a})$, a constant, and $\partial\boldsymbol{X}/\partial t = \boldsymbol{u}(\boldsymbol{X},t)$, which is just the statement that the particle moves with the local fluid velocity. The collective effect of this transport is a rearrangement of the tracer field q by the velocity field $\boldsymbol{u}$.

Depending on the nature of $\boldsymbol{u}$, this may lead to highly intricate distributions of q, even starting from simple initial conditions. Moreover, there are important applications in which $\boldsymbol{u}$ depends on q itself, often in a nonlocal manner, that is, in which the entire field of q contributes to $\boldsymbol{u}$ at any given point $\boldsymbol{x}$. A specific example relevant to the present topic is provided by the two-dimensional Euler equations governing the behavior of an inviscid, incompressible fluid:

$$\frac{\mathrm{D}\boldsymbol{u}}{\mathrm{D}t} = -\frac{\nabla p}{\rho}, \tag{2}$$

$$\nabla \cdot \boldsymbol{u} = 0, \tag{3}$$

where p is the pressure and ρ is the density (here constant), and where now the velocity field is two dimensional: $\boldsymbol{u} = (u, v)$. Taking the curl of Equation (2) gives an equation for the scalar vorticity $\zeta \equiv \partial v/\partial x - \partial u/\partial y$:

$$\frac{\mathrm{D}\zeta}{\mathrm{D}t} = 0, \tag{4}$$

which is identical to Equation (1) if we take $q = \zeta$. Thus, the vorticity is materially conserved in this system. But it also induces the velocity field $\boldsymbol{u}$, that transports it. Equation (3) is satisfied generally by considering

$$u = -\partial\psi/\partial y \quad \text{and} \quad v = \partial\psi/\partial x, \tag{5}$$

where $\psi(\boldsymbol{x},t)$ is called the streamfunction. Substituting these components into the definition of ζ leads to a Poisson equation for ψ:

$$\nabla^2\psi = \zeta. \tag{6}$$

Given the distribution of ζ, this equation (with suitable boundary conditions) may be inverted to find ψ, whose spatial derivatives provide u and v. The inversion of this equation can be formally accomplished by using the Green function $G(\boldsymbol{x}; \boldsymbol{x}')$ of Laplace's operator ∇^2; in two dimensions,

$$G(\boldsymbol{x}; \boldsymbol{x}') = (2\pi)^{-1} \log |\boldsymbol{x}' - \boldsymbol{x}|. \tag{7}$$

(G is the solution to Equation (6) for $\zeta = \delta(\boldsymbol{x}' - \boldsymbol{x})$, a singular delta distribution of vorticity having a unit spatial integral.)

Consider henceforth an unbounded two-dimensional fluid. Then, the formal solution to the inversion problem is

$$\psi(\boldsymbol{x}, t) = \iint G(\boldsymbol{x}; \boldsymbol{x}')\zeta(\boldsymbol{x}', t)\, \mathrm{d}x'\mathrm{d}y', \tag{8}$$

which shows explicitly that the flow field at any point depends on the vorticity field at all points. Moreover, the integration over space implies that the field of ψ is generally smoother than that of ζ.

The evolution of the flow in this case consists of two basic steps:

- *inversion*—the recovery of the velocity field $\boldsymbol{u}$ from the distribution of ζ and
- *advection*—the transport of fluid particles to the next instant of time.

Now, a two-dimensional plane is carpeted by an infinite number of such particles, and therefore, this view of the evolution may appear to be quite complex. However, the material conservation of q (or ζ) affords an enormous simplification. First, note that, if one exchanges two particles labeled $\boldsymbol{a}$ and $\boldsymbol{a}'$ having the same value of q, this does not alter the distribution of q, and as a result the velocity field $\boldsymbol{u}$ remains unchanged. This is a "particle relabeling" symmetry, and in general, it gives rise to an infinite number of globally conserved quantities (the spatial integrals of any functional of q). This symmetry implies that contours of fixed q consist of the same set of fluid particles for all time. They are called material contours.

Contour dynamics arises from representing the distribution of q by a finite set of contours C_k, $k = 1, 2, \ldots, n$, between which q is spatially uniform, and across which q jumps by Δq_k (defined to be the value of q to the left of C_k minus that to the right of C_k). The contours here are still material ones—the particles just on either side of a contour retain their distinct values of q. Between two contours, any two fluid particles may be exchanged without altering q or $\boldsymbol{u}$. This implies that only the contours matter for determining the velocity field. Also, since the contours are material, their advection suffices to evolve the entire distribution of q. This is the basis of contour dynamics.

To see how this works for the two-dimensional Euler equations, it remains to be shown as to how one can calculate the velocity field directly from the contours C_k. The starting point is Equation (8), in which we consider $\zeta = q$ to be a piecewise-uniform field. For the moment, we need only use the property $G(\boldsymbol{x}; \boldsymbol{x}') = g(\boldsymbol{x}' - \boldsymbol{x})$ satisfied by the Green function of the Laplace operator. Also, we need only consider one contour at a time and afterwards linearly superpose the results, since the relation between q and $\boldsymbol{u}$ is linear. We may take $q = 0$ outside of the (closed) contour, so that $q = \Delta q$ inside it (denoted by the region R below). Nonzero exterior q simply gives rise to solid-body rotation, and may be superposed afterwards. Then, from Equation (5), we have

$$\begin{aligned}(u(\boldsymbol{x}, t), v(\boldsymbol{x}, t))\\ &= \left(-\frac{\partial}{\partial y}, \frac{\partial}{\partial x}\right) \Delta q \iint_R g(\boldsymbol{x}' - \boldsymbol{x})\, \mathrm{d}x'\mathrm{d}y' &(9)\\ &= -\Delta q \iint_R \left(-\frac{\partial}{\partial y'}, \frac{\partial}{\partial x'}\right) g(\boldsymbol{x}' - \boldsymbol{x})\, \mathrm{d}x'\mathrm{d}y' &(10)\\ &= -\Delta q \oint_C g(\boldsymbol{X}' - \boldsymbol{x})\, (\mathrm{d}X', \mathrm{d}Y'), &(11)\end{aligned}$$

where we have used the symmetry of g with respect to $\boldsymbol{x}$ and $\boldsymbol{x}'$ in the second line and Stokes' theorem in the third, and where $\boldsymbol{X}'$ denotes a point on the contour C. The velocity field anywhere thus depends only on the shape of C.

For a set of contours, the velocity field is required only on the contours themselves to evolve q. The contours thus form a closed dynamical system, contour dynamics, governed by

$$\frac{\mathrm{d}\boldsymbol{X}_j}{\mathrm{d}t} = \boldsymbol{u}(\boldsymbol{X}_j) = -\sum_{k=1}^{n} \Delta q_k \oint_{C_k} g(\boldsymbol{X}_k' - \boldsymbol{X}_j)\mathrm{d}\boldsymbol{X}_k' \tag{12}$$

for all points $\boldsymbol{X}_j$ on contours C_j, $j = 1, 2, \ldots, n$.

These equations were first derived for the two-dimensional Euler equations by Zabusky et al. (1979), following earlier work by Berk & Roberts (1967), who derived a similar contour-based model for the two-dimensional Vlaslov equations in plasma physics. These authors also developed numerical methods for contour dynamics in which the contours were discretized into a finite set of points or nodes, originally connected by straight line segments. A wide variety of numerical methods have since been developed, many of which are summarized in the review articles of Dritschel (1989) and Pullin (1992). They principally differ in terms of the choice of the interpolation between nodes (linear, quadratic, cubic; local and global splines); the method of numerical quadrature used to evaluate the contour integral over the segment connecting adjacent nodes (trapezoidal, Gaussian, explicit); the method of redistributing, inserting, and removing nodes to maintain an accurate representation of the contour shape; and the procedure used, if any, to remove fine-scale structure (e.g., filaments and thin bridges connecting two separating regions)—a procedure coined "contour surgery" (Dritschel, 1988).

Contour dynamics has been used to study a wide variety of problems, from the interaction of two vortex patches (having just one contour each), to the filamentation and stripping of nested vortices (having many contours to represent a continuum) (Dritschel, 1989; Pullin, 1992). The numerical method illustrated next is described in Dritschel (1988, 1989). This method uses local cubic splines between contour nodes, explicit quadrature to first order in the departure of the contour from a straight line between nodes, node redistribution based on maintaining a local node density proportional to the square root of contour curvature, and automatic surgery whenever contours or contour parts get closer than a prescribed cutoff scale δ. This scale and the precise formula for the node density are chosen to balance the errors arising from surgery and node redistribution. A fourth-order Runge–Kutta scheme is used for the time integration.

An example is presented next of the collapse of three vortex patches (see also (Rogberg & Dritschel, 2000)). The centers of the vortices are initially chosen at points where equivalent delta-distributed point vortices of the same circulation (spatial integral of q) are known to collide in finite time. Two of the vortices have $q = +2\pi$ and radii 1 and $2/\sqrt{5}$, while the third has $q = -2\pi$ and radius $\frac{2}{3}$. The two positive vortices are initially separated by a distance $d = 5$, and the negative vortex is placed at a distance $d\sqrt{17/27}$ and at an angle 225° relative to the joint center of the two positive vortices. The vortices are then all shifted so that the joint center of all three vortices lies at the origin. Starting from this configuration, the collapse time for the point vortices is $7.70059886\ldots$

Figure 1 illustrates the evolution of the vortices—in the upper left-hand frame, the initial conditions are shown, while the remaining frames (to the right and then downwards) are spaced at unit increments in time starting from $t = 5$ and ending at $t = 15$. By $t = 6$, the two positive vortices begin to merge (they are separated by only a thin channel of irrotational fluid). Thereafter, the flow grows rapidly in complexity, as many filaments are generated and small vortices roll up at the tips of some of the filaments. Notably, the negative vortex does not distort significantly, but merely acts to bring the two positive vortices together.

The complexity just illustrated is typical of many vortex interactions. An accurate, robust numerical method must be able to capture this generic behavior, at least over time scales when the flow is reasonably predictable. To see how well the current method performs, we next examine how the results vary with spatial resolution. Two additional simulations were performed at half and double the average point spacing used in Figure 1. The results are compared in Figure 2 at the final time, $t = 15$, when the numbers of nodes in the three simulations are, from low to high resolution, 2740, 10,738 and 27,297 (at $t = 0$, the numbers of nodes are 183, 349, and 682). Note the cutoff scale $\delta = 0.000225$, 0.0009, and 0.0036 in the three simulations—there is a factor of 4 difference in δ between resolutions. The agreement is striking even in the detailed structure. The most visible differences show up in the lengths of the filaments, which are removed more readily at low resolution. These filaments, however, contribute negligibly to the velocity field, and retaining them makes little difference to the evolution of the flow.

Contour dynamics has since been applied in a variety of diverse fields. Its largest growth has occurred in the field of atmospheric and oceanic dynamics, where the potential vorticity plays the role of the materially

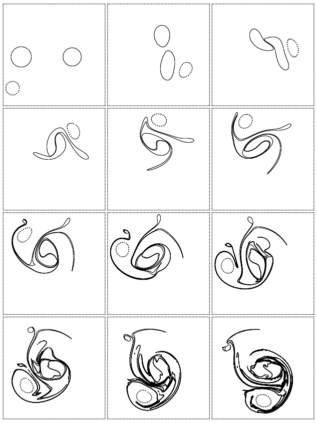

Figure 1. The collapse of three vortex patches. The initial condition is shown in the upper left frame. Time proceeds to the right and downwards in increments of one unit, from $t = 5$ to $t = 15$. The window of view is $-5.0 < x < 5.0$ and $-5.8 < y < 4.2$. The negative vortex is rendered with a short-dashed line (with a dash between each node), while the positive vortices are rendered with a continuous solid line.

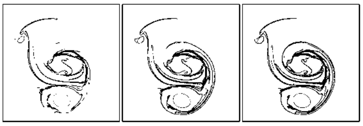

Figure 2. Comparison, at $t = 15$, of three contour dynamics simulations of vortex collapse. Resolution increases from left to right, doubling between each frame (the node spacing parameter is $\mu = 0.12, 0.06$, and 0.03, and the large-scale length $L = 1$ in all cases; consult Dritschel (1989) or Dritschel & Ambaum (1997) for further details). The domain of view is the same as used in the previous figure.

conserved tracer q often to a very good approximation (Hoskins et al., 1985). Indeed, its application to this field is on a much sounder footing than it is to the fields it was originally developed for: plasma physics and aeronautics. The two-dimensional approximation is a particularly severe one in aeronautics, since real flows do not preserve two-dimensional symmetry, unless constrained in some manner. In the atmosphere and oceans, rotation and stratification serve to constrain the flow to be two dimensional, or more appropriately, layerwise two dimensional, and furthermore one may extend contour dynamics to study such flows (indeed, the equations are formally no different than given those by (12); see (Dritschel, 2002)).

Finally, the use of contours to carry out tracer advection—the fast and accurate part of contour dynamics—has been combined with more traditional approaches of computing the velocity field (the inversion step) to produce a particularly fast, accurate, and versatile numerical method called the contour-advective semi-Lagrangian (CASL) algorithm (Dritschel & Ambaum, 1997; Dritschel et al., 1999; Dritschel & Viúdez, 2003). This latest development allows the extension of the contour approach to much more realistic sets of equations, and has significantly widened the applicability of the original contour dynamics method.

DAVID DRITSCHEL

See also **Chaotic advection; Euler–Lagrange equations; Vortex dynamics of fluids**

Further Reading

Berk, H.L. & Roberts. K.V. 1967. The water-bag model. *Methods in Computational Physics*, 9: 87–134

Dritschel, D.G. 1988. Contour surgery: a topological reconnection scheme for extended integrations using contour dynamics. *Journal of Computational Physics*, 77: 240–266

Dritschel, D.G. 1989. Contour dynamics and contour surgery: numerical algorithms for extended, high-resolution modelling of vortex dynamics in two-dimensional, inviscid, incompressible flows. *Computer Physics Reports*, 10: 77–146

Dritschel, D.G. 2002. Vortex merger in rotating stratified flows. *Journal of Fluid Mechanics*, 455: 83–101

Dritschel, D.G. & Ambaum, M.H.P. 1997. A contour-advective semi-Lagrangian numerical algorithm for simulating finescale conservative dynamical fields. *Quarterly Journal of the Royal Meteorological Society*, 123: 1097–1130

Dritschel, D.G., Polvani, L.M. & Mohebalhojeh, A.R. 1999. The contour-advective semi-Lagrangian algorithm for the shallow water equations. *Monthly Weather Review*, 127(7): 1551–1565

Dritschel, D.G. & Viúdez, A. 2003. A balanced approach to modelling rotating stably-stratified geophysical flows. *Journal of Fluid Mechanics*, 488: 123–150. See also: www-vortex.mcs.st-and.ac.uk.

Hoskins, B.J., McIntyre, M.E. & Robertson, A.W. 1985. On the use and significance of isentropic potential-vorticity maps. *Quarterly Journal of the Royal Meteorological Society*, 111: 877–946

Pullin, D.I. 1992. Contour dynamics methods. *Annual Review of Fluid Mechanics*, 24: 89–115

Rogberg, P. & Dritschel, D.G. 2000. Mixing and transport in two-dimensional vortex interactions. *Physics of Fluids*, 12(12): 3285–3288

Zabusky, N.J., Hughes, M.H. & Roberts, K.V. 1979. Contour dynamics for the Euler equations in two dimensions. *Journal of Computational Physics*, 30: 96–106

CONTROL PARAMETERS

See **Bifurcations**

CONTROLLING CHAOS

It may seem paradoxical that chaotic systems—which are extremely sensitive to the tiniest fluctuations—can be controlled; yet, the earliest reference to this idea appears around 1950, when John von Neumann presaged just that. Nowadays, laboratory demonstrations of the control of chaos have been realized in chemical, fluid, and biological systems, and the intrinsic instability of chaotic celestial orbits is routinely used to advantage by international space agencies who divert spacecraft to travel vast distances using only modest fuel expenditures.

A variety of techniques for chaos control has been implemented since around 1990 when the first concrete analyses appeared, including traditional feedback and open-loop methods, neural network applications, shooting methods, Lyapunov function approaches, and synchronization to both simple and complex external signals. These techniques resolve the paradox implied by chaos control in different ways, but they all make use of the fact that chaotic systems can be productively controlled if disturbances are countered by small and intelligently applied impulses. Just as an acrobat balances about an unstable position on a tightrope by the application of small correcting movements, a chaotic system can be stabilized about any of an infinite number of unstable states by continuous application of small corrections.

Two characteristics of chaos make the application of control techniques even more fruitful. First, chaotic systems alternately visit small neighborhoods of an infinite number of periodic orbits. The presence of an infinite number of periodic orbits embedded within a chaotic trajectory implies the existence of an enormous variety of different behaviors within a single system. Thus, the control of chaos opens up the potential for tremendous flexibility in operating performance within a single system.

As an example, Figure 1 depicts the Lorenz attractor, used to model fluid convection. Embedded within the gray attractor are innumerable periodic orbits, such as the solid figure-8 orbit and the more complicated dashed one. For practical systems such as chemical reactors or fluidized beds, the presence of multiple

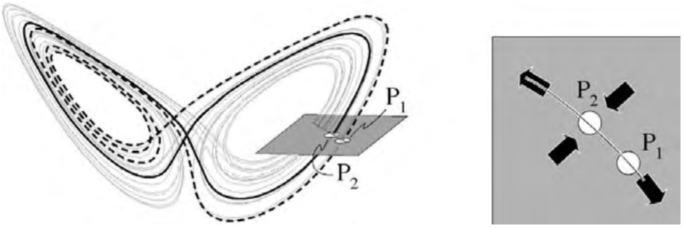

Figure 1. Left: Lorenz attractor with two of its embedded unstable periodic orbits highlighted. Right: unstable points P_1 and P_2 in the surface of section indicated. The unstable direction is denoted by outgoing arrows, and stable direction is denoted by ingoing arrows.

co-existing states implies that one chaotic system could be operated in multiple different states, thus potentially performing the function of several separate units.

A second characteristic of chaos that is important for control applications is the exponential sensitivity of the phenomenon. That is, the fact that the state of a chaotic system can be drastically altered by the application of small perturbations means two things: such a system if uncontrolled can be expected to fluctuate wildly, and if controlled can be directed from one state to a very different one using only very small controls.

Traditional feedback control remains among the most widely used methods of control for chaotic systems. To implement feedback control, one waits until a chaotic trajectory by chance lands near a desired periodic point and then applies small variations to an accessible system parameter in order to repeatedly nudge the trajectory closer to that point. As an example, consider the plot to the right in Figure 1, where we depict two periodic points as they appear on a "surface of section" formed in this case by recording every intersection between the Lorenz chaotic attractor and the half plane, $Z = 0$, $X > 0$. To control the state to remain near point P_2 (so the trajectory stays near the figure-8 trajectory shown to the left), one needs to apply variations in a parameter, p, that directs the state toward P_2 along the unstable direction (or directions in more complicated problems) indicated by outgoing arrows in Figure 1. One can establish the direction in which a parametric control moves the chaotic state either experimentally, by varying the parameter and recording the future variation of the system state, or analytically, by determining the Jacobian of the flow or mapping where available. Nudging the state closer to P_2 amounts to what is termed "pole placement" in traditional control literature, and numerous reports of alternative strategies for selecting parameter variations appear in the literature. Strategies include simple pole placement, optimal control, neural network approaches, simple proportional control, periodic forcing, and control dependent on the distance from P_2. Most of these strategies have proven successful under appropriate conditions, and the choice of strategy depends principally on details of the control goal required and the computational resources available to meet that goal.

All of these strategies require that the system state must lie close to the desired state in order to achieve control. In such a case, the system dynamics can be linearized, making control calculations rapid and effective. Fortunately, in chaotic systems, one can rely on ergodicity to ensure that the system state will eventually wander arbitrarily close to the desired state. By the same token, if it is desired to switch the system between one accessible state (say P_1) and a second (say P_2), one can merely release control from P_1 and re-apply a new control algorithm once the system strays close to P_2, which it is certain to do by ergodicity.

In higher-dimensional or slowly varying systems, the time taken for the state to move on its own from one state to another can be prohibitive, and for this reason fully nonlinear control strategies have been devised that use chaotic sensitivity to steer the system state from any given initial point to a desired state. Since chaotic systems amplify control impulses exponentially, the time needed to steer such a system can be quite short. These strategies have been demonstrated both in systems in which a large effect is desired using very modest parameter expenditures (energy or fuel) and in systems in which rapid switching between states is needed (computational or communications applications).

On the other hand, in both linear and nonlinear control approaches, one needs to repeatedly re-apply control over a time that is short compared with the inverse of the fastest growing growth rate of the system in order to counter the potential amplification of ubiquitous noises. Computational and experimental analyses have demonstrated that this is readily done in typical chaotic systems and that control can be robustly achieved. Because large but rare noise events can occur, however, controlled states occasionally break free when the system encounters an anomalous large noise. In this case, bounds have been established for the frequency and duration of these noise-induced excursions.

Numerous biological control applications have been proposed since the first introduction of the notion of chaotic control. Among the first applications were studies of intrinsic nonlinear control mechanisms involved in autonomic and involuntary functions such as the regulation of internal rhythms and the control of gait and balance. These studies confirm that nontrivial control algorithms are involved in the maintenance of normal physiological function and that provocative insights into pathological conditions can be gained (such as cardiac and breathing arrythmias and motor tremor). Further work has shown that networks of chaotic devices, under prescribed conditions, can be brought into synchronization, and strong indications have been presented that neuronal signaling may rely on nonlinear synchronization. Additional experimental studies are promising for the control of unwanted fluctuations (e.g., during fibrillation of the heart) or for the so-called "anticontrol" of synchronized periodic signals in focal epilepsy. In both studies, the goal is to use feedback control methods to steer a diseased organ using small electrical stimulation: in the former state, toward a stabilized state, and in the latter, away from a synchronized state.

Troy Shinbrot

See also **Chaotic dynamics; Feedback; Lorenz equations**

Further reading

Alekseev, V.V. & Loskutov, A.Y. 1987. Control of a system with a strange attractor through periodic parametric action. *Soviet Physics Doklady*, 32: 1346–1348

Ditto, W.L. & Pecora, L.M. 1993. Mastering chaos. *Scientific American*, 78–84

Garfinkel, A., Spano, M.L., Ditto, W.L. & Weiss, J.N. 1992. Controlling cardiac chaos. *Science*, 257:1230–1235

Glass, L. & Zeng, W. 1994. Bifurcations in flat-topped maps and the control of cardiac chaos. *International Journal of Bifurcation & Chaos*, 4: 1061–1067

Hayes, S., Grebogi, C. & Ott, E. Communicating with Chaos. *Physical Review Letters*, 70: 3031–3014

Hübler, A., Georgii, R., Kuckler, M., Stelzl, W. & Lscher, E. 1988. Resonant Stimulation of nonlinear damped oscillators by Poincaré maps. *Helvetica Physica Acta*, 61: 897–900

Lima, R. & Pettini, M. 1990. Suppression of chaos by resonant parametric perturbations. *Physics Review A*, 41: 726–733

Ott, E., Grebogi, C. & Yorke, J.A. 1990. Controlling chaos. *Physical Review Letters*, 64: 1196–1199

Pecora, L.M. & Carroll, T.J. 1990. Synchronization in chaotic systems. *Physical Review Letters*, 64: 821–824

Schiff, S.J., Jerger, K., Duong, D.H., Chang, T., Spano, M.L. & Ditto, W.L. 1994. Controlling chaos in the brain. *Nature*, 370: 615–620

Shinbrot, T., Ott, E., Grebogi, C. & Yorke, J.A. 1993. Using small perturbations to control chaos. *Nature*, 363: 411–417

COSMOLOGICAL MODELS

Relativistic cosmology—the science of the structure and evolution of the universe—is based on the building and investigation of cosmological models (CMs), which describe geometrical properties of physical space-time, the matter, composition and structure of the universe, and physical processes at different stages of the universe's evolution. Prominent in cosmology is the hot big bang CM, which is based on solutions of Alexander Friedmann's cosmological equations for homogeneous isotropic models deduced in the framework of Einstein's general relativity theory (GR). Because of its large-scale structure (galaxies, clusters of galaxies, etc.), the universe is homogeneous and isotropic only on the largest scales from 100 Mpc. (The pc or parsec is an astronomical unit of distance equal to 3.2616 light years; thus, an Mpc is 3.2616 million light years.)

The most important feature of Friedmann's CM is its nonstationary character, which was confirmed by Edwin Hubble's discovery of cosmological expansion in 1929. In this formulation, the geometrical properties of physical space depend on the value of energy density ρ relative to a critical density

$$\rho_{\text{crit}} = \frac{3H_0^2}{8\pi G}, \tag{0}$$

where H_0 is the expansion rate (Hubble parameter) at the present epoch and G is Newton's gravitational constant. If $\Omega = \rho/\rho_{\text{crit}} = 1$, the 3-space is flat; if $\Omega > 1$, 3-space possesses positive curvature and if $\Omega < 1$, 3-space possesses negative curvature. Corresponding CMs are flat, closed, and open CM, respectively.

All Friedmann CMs have a beginning in time (or cosmological singularity), where energy density and curvature invariants diverge. Their evolutions depend on properties of matter. In the case of ordinary matter with a positive energy density and a nonnegative pressure, the evolution of flat and open models has the character of expansion, and closed models recollapse after an expansion stage. The assumption that the temperature was very high at the initial stage of cosmological explanation (hot CM) was confirmed by the discovery of the cosmic microwave background (CMB) radiation in 1965, with a present epoch temperature of about $T = 2.7$ K. The theory of nucleosynthesis of light elements (hydrogen, helium, deuterium, lithium, etc.) into the first few minutes of cosmological expansion based on the framework of the hot big bang CM is in accord with empirical data.

Advances in both theory and technology during the last 20 years have launched cosmology into a most exciting period of discovery. By using precise instruments (telescopes, satellites, spectroscopes), several cosmological research programs are being carried out, including investigations of the anisotropy of CMB and supernovae observations. Cosmological observations have not only strengthened and expanded the hot big bang CM but they have also revealed surprises.

Recent measurements of the anisotropy of the CMB have provided convincing evidence that the spatial geometry is very close to being uncurved (flat) with $\Omega = 1.0 \pm 0.03$. The currently known components of the universe include ordinary baryonic matter, cold dark matter (CDM), massive neutrinos, the CMB and other forms of radiation, and dark energy. The sum of the values for these densities derived empirically is equal to the critical density (to within their margins of error). The largest contributions to energy density are from two components—CDM and dark energy. About 30% of the total mass-energy is dark matter, composed of particles probably formed early in the universe. Two thirds is in smooth dark energy whose gravitational effects began causing the expansion of the universe to speed up just a few billion years ago. The remarkable fact that the expansion is accelerating can be accounted for within GR, as the source of gravity is proportional to $(\rho + 3p)$, where the pressure p and energy density ρ describe the bulk properties of "the substance." A substance with pressure more negative than one-third its energy has repulsive gravity in GR. Such a situation occurs, for example, for gravitating vacuum (positive cosmological constant), for which $p = -\rho$.

In addition to breakthrough empirical observations, creative theoretical ideas are also driving progress in cosmology. The development of cosmology during the last 20 years shows that profound connections exist between the elementary particles on the smallest scales and the universe on the largest. Using unified gauge theories of elementary particles, an inflationary scenario was formulated, which resolves a number of problems of standard Friedmann cosmology: flatness and the problem of horizon, among others. According to inflation, small bits of the universe underwent a burst of expansion when the universe was extremely young, explaining the homogeneity and isotropy of the universe at initial stages of cosmological expansion. Based on the framework of an inflationary CM, the appearance of quantum fluctuations with a nearly scale-invariant distribution by transition to radiation-dominated era was predicted, explaining the large scale structure of the universe.

The inflationary CM as well as others discussed above are singular, which is an outstanding problem of GR. Assuming that the Planck era (when the universe was sufficiently dense to require a quantum mechanical treatment) existed, some quantum gravitation theory is necessary to construct a regular CM. At present, the superstring theory is a candidate for such a theory.

Some regular CMs have been constructed in a "brane world," under which our universe is thought to exist as a slice (or membrane) through a higher-dimensional space. By using scalar fields with a negative potential, a solution for an oscillating CM was obtained; thus, the Big Crunch takes place in such models before Big Bang. Resolving the problem of cosmological singularity requires that the gravitation theory not only admits regular solutions for CM but also excludes singular solutions. This suggests gauge theories of gravitation (Poincaré gauge theory or metric-affine gauge theory), leading to regular bouncing solutions for CMs.

The building of more realistic CMs requires the resolution of fundamental cosmological problems. According to present knowledge, our universe is flat and 13 Gyr old, and it is expanding at the current rate of $H_0 = 72 \pm 8\,\text{km sec}^{-1}\ \text{Mpc}^{-1}$. Measurements of the past rate reveal that the universe is presently in a period of cosmic acceleration. The contribution of ordinary matter to the overall mass energy is small, with more than 95% of the Universe existing in new and unidentified forms of matter and energy. What is the composition of dark matter (axions, neutralinos, or other exotic particles)? What is the nature of dark energy (quantum vacuum energy or scalar fields)? What is the field that drives inflation? Answers to these and other questions will change the picture presented above.

VIACHASLAV KUVSHINOV AND ALBERT MINKEVICH

See also **Black holes; Einstein equations; Galaxies; General relativity**

Further Reading

Gasperini, M. & Veneziano, G. 2003. The pre-big bang scenario in string cosmology. *Physics Reports*, 373: 1–212

Khoury, J., Ovrut, B.A, Seiberg, N., Steinhardt, P.J. & Turok, N. 2002. From big crunch to big bang. *Physical Review* D, 65: 086007

Kolb, E.W. & Turner, M.S. 1990. *The Early Universe*, Reading, MA: Addison-Wesley

Linde, A.D. 1990. *Particle Physics and Inflationary Cosmology*, Chur: Harwood Academic

Steinhardt, P.J. & Turok, N. 2002. A Cyclic Model of the Universe. hep-th/0111030

CONVECTION

See **Fluid dynamics**

CONVECTIVE INSTABLITY

See **Wave stability and instability**

CORRELATION DIMENSION

See **Dimensions**

CORRESPONDENCE PRINCIPLE

See **Quantum nonlinearity**

COUPLED MAP LATTICE

Originally introduced in the study of spatiotemporal chaos, the coupled map lattice (CML) can be presented as a dynamical model for the evolution of a spatially extended system in time (Kaneko, 1983). CMLs have been widely used, not only as a tool for the study of spatiotemporal chaos but also for pattern dynamics in physics, chemistry, ecology, biology, brain theory, and information processing.

A CML is a dynamical system with discrete time (map), discrete space (lattice), and a continuous state. It consists of dynamical elements on a lattice, which interact (are coupled) with suitably chosen sets of other elements.

The construction of a CML is carried out as follows. First, choose a (set of) field variable(s) on a lattice. This (set of) variable(s) is on a macroscopic, not a microscopic level. Second, decompose the phenomenon of interest into independent units (e.g., convection, reaction, diffusion, and so on). Third, replace each unit by simple parallel dynamics (procedure) on a lattice, where the dynamics consists of a nonlinear transformation of the field variable at each lattice point and/or a coupling term among suitably chosen neighbors. Finally, carry out each unit dynamics (procedure) successively.

As a simple and widely used example, consider a phenomenon that is created by a locally chaotic process and by diffusion, and choose a suitable lattice model on a coarsegrained level for each process. As the simplest choice, we can adopt some one-dimensional map for chaos, and a discrete Laplacian operator for the diffusion. The former process is given by $x'_n(i) = f(x_n(i))$, where $x_n(i)$ is a variable at time n and lattice site i, $(i = 1, 2, \ldots, N)$, whereas $x'_n(i)$ is introduced as the intermediate value. The discrete Laplacian operator for diffusion is given by

$$x_{n+1}(i) = (1-\varepsilon)x'_n(i) + \frac{\varepsilon}{2}\{x'_n(i+1) + x'_n(i-1)\}. \qquad (0)$$

Combining the above two processes, the CML is given by

$$x_{n+1}(i) = (1-\varepsilon)f(x_n(i)) + \frac{\varepsilon}{2}\{f(x_n(i+1)) + f(x_n(i-1))\}. \quad (1)$$

The mapping function $f(x)$ is chosen to depend on the type of local chaos. For example, one can choose the logistic map, $f(x) = rx(1-x)$, as a typical model for chaos. As the map dynamics are well studied, dynamical systems theory can be applied to understand behaviors of the CML.

By adopting different procedures, one can construct models for different types of spatially extended dynamical systems. For problems of phase transition dynamics, it is useful to adopt a map with bistable fixed points (e.g., $f(x) = \tanh x$) as a local dynamics. The choice of a different type of coupling, as well as the extension to a higher-dimensional space is straightforward. By changing the procedures in the CML, one can easily construct a model for dynamical phenomena in space-time. Examples include spinodal decomposition, crystal growth, boiling, convection, and cloud dynamics, among others.

Universality Classes of the Phenomena

Phenomena found in one CML are often observed in a wide variety of systems, and they form a universality class common to such systems. CMLs thus work as a tool to predict novel phenomenology forming such qualitative universality classes.

In the model of Equation (1), the following phenomena have been discovered: (i) spatial bifurcation and frozen chaos, (ii) spatiotemporal intermittency (STI), (iii) Brownian motion of chaotic defects, and (iv) global traveling wave by local phase slips. These phenomena are observed in a wide variety of systems, including experiments. In particular, STI is now regarded as a universal route to fully developed spatiotemporal chaos. In fully developed spatiotemporal chaos, statistical mechanics theory is developed by taking advantage of the discreteness in spacetime.

If one adopts a two-dimensional lattice system, spiral pattern dynamics are often observed. For example, by taking a local map with an excitable state, the formation of spiral waves is studied, including turbulence due to the break-up of a spiral wave pair. Such a model is studied in relation to the pattern dynamics in reaction-diffusion systems as well as wave propagation in cardiac tissue.

Another straightforward extension is a spatially asymmetric coupling. In an open fluid flow, for example, there is coupling from up-flow to down-flow, instead of the diffusion. The CML $x_{n+1}(i) = (1-\varepsilon) f(x_n(i)) + \varepsilon f(x_n(i))$ gives a prototype model for such a case. In this open flow system, it is important to distinguish absolute instability from convective instability. If a small perturbation against a reference state grows in a stationary frame, it is called "absolute instability," while if the perturbation grows only in a frame moving with a specific velocity, it is called "convective instability." This convective instability leads to spatial bifurcation from a homogeneous state to down-flow convective chaos.

Globally Coupled Maps with Applications to Biology

An extension of CML to global coupling is interesting, and often important for biological problems. Thus a globally coupled map (GCM) was introduced as a mean-field-type extension of a CML, written as

$$x_{n+1}(i) = (1-\varepsilon) f(x_n(i)) + (\varepsilon/N) \sum_{j=1}^{N} f(x_n(j)). \qquad (2)$$

One important notion here is clustering. The elements split into several clusters, within which all the elements oscillate in synchronization. Depending on the numbers of clusters in the GCM, there are phase transitions among a coherent phase, an ordered phase, a partially ordered phase, and a desynchronized phase, as the parameter describing the nonlinearity in $f(x)$ is increased. In the partially ordered phase, there are many attractors with different numbers of clusterings and with a variety of partitions. Dynamically, the system spontaneously switches between ordered states through disordered states, known as chaotic itinerancy. In the desynchronized phase, nontrivial collective motion is observed with some hidden coherence among elements. This demonstrates the existence of macroscopic chaos different from microscopic chaos represented by each map $x_{n+1} = f(x_n)$. This observation may shed new light on the origin of collective behavior by an ensemble of cells, such as an electroencephalogram (EEG) in the brain.

Often, a biological system has both internal dynamics and interactions. Chemical dynamics in a cell includes both intra-cellular reactions associated with gene expressions and cell-cell interactions. Since a CML or GCM is a model for such intra-inter dynamics, the concepts developed in this area will be relevant to biological problems. For example, clustering leads to differentiation of the states of elements. The theory for cell differentiation and robust developmental process may be based on this dynamic differentiation.

KUNIHIKO KANEKO

See also **Cellular automata; Cluster coagulation; Maps**

Further Reading

Chaté, H. & Courbage, M. (editiors). 1997. Special issue on lattice dynamics. *Physica* D, 103: 1–612

Kaneko, K. 1986. *Collapse of Tori and Genesis of Chaos in Dissipative Systems* Singapore: World Scientific 1986 (PhD thesis originally published 1983)

Kaneko, K. (editior). 1992. Chaos focus issue on coupled map lattices. *Chaos*, 2(3): 279–408

Kaneko, K. (editor). 1993 *Theory and Applications of Coupled Map Lattices*, Chichester and New York: Wiley

Kaneko, K. & Tsuda, I. 2000. *Complex Systems: Chaos and Beyond—A Constructive Approach with Applications in Life Sciences*, Berlin and New York: Springer

Kaneko, K. & Tsuda, I. 2003. Chaos focus issue on chaotic itinerancy. *Chaos*, 13(3): 926–1164

COUPLED OSCILLATORS

The simplest coupled oscillator is a pair of linearly coupled harmonic oscillators, which is used as a model for a wide variety of physical systems—including the interactions of musical instruments and tuning forks, lattice vibrations, electrical resonances, and so on—in which energy tunnels back and forth between two sites at a difference (beat) frequency. If there are many elementary oscillators that are nonlinear, coupled systems exhibit more varied nonlinear phenomena. There are two types of coupled nonlinear oscillators: those described by Hamiltonian (energy-conserving) dynamics, and systems in which energy is not conserved. In addition to coupled pendula, examples of the first kind include the Fermi–Pasta–Ulam model and the Toda lattice.

Coupled nonlinear oscillators that do not conserve energy can be viewed as coupled limit cycle oscillators. A limit cycle oscillator (also called a self-sustained oscillator) is described as an attractor in a dissipative dynamical system. A typical dissipative dynamical system that exhibits a limit cycle oscillation is van der Pol's equation

$$\frac{d^2x}{dt^2} - \varepsilon\left(1-x^2\right)\frac{dx}{dt} + \omega^2 x = 0 \qquad (1)$$

in which the character of the oscillation varies from sinusoidal and energy-conserving to a strongly dissipative (blocking or relaxation) oscillation through the variation of a parameter (ε) from zero to large values (van der Pol, 1934).

Among the varieties of limit cycle oscillators, the behavior of a quasilinear oscillator (small ε) can be expressed by a sinusoidal wave, $x(t) = A\sin(\omega t + \phi_0)$. The wave shape of a relaxation oscillator (large ε), on the other hand, is composed of alternating fast and slow motions, similar to the spikes and slow recovery motions in a firing neuron, and stick-slip oscillations in frictional motions.

Although the limit cycle oscillation has a certain natural amplitude and frequency, the phase variable, for example, $\phi = \omega t + \phi_0$ for a quasilinear oscillator, is a neutral mode, sensitively perturbed by an external force. If the external force is periodic with a frequency close to the natural frequency of the limit cycle oscillator, the phase of the limit cycle oscillator tends to approach the phase of the external periodic force. If the external force is sufficiently strong, the phase difference $\Delta\phi(t) = \phi(t) - \phi_e(t)$ between the limit cycle oscillator and the external force is fixed. This phenomenon—termed *phase* or *frequency locking*—occurs more easily when ε is large, the frequency of the limit cycle oscillator is close to that of the external force, and the coupling (K) is large.

Regions in the (ω, ε, K) parameter space where frequency locking is observed are termed "Arnol'd

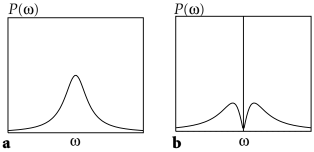

Figure 1. Frequency distribution $P(\omega)$ (a) in an asynchronous state for $K < K_c$ and (b) in a mutually entrained state for $K > K_c$.

tongues” owing to their peculiar shape. The frequency ratio between the limit cycle and the external force is 1:1 in the above frequency locking. In general, n:m frequency lockings are possible, where n and m are small integers.

For a collection of coupled limit cycle oscillators with slightly different natural frequencies, frequency locking (called mutual entrainment) also occurs, as was first observed by Christiaan Huygens in the 17th century. He found that the motions of pendulum clocks suspended from the same wooden beam come to coincide with each other perfectly. Nobert Wiener analyzed such systems in the 1950s, showing that the power spectrum of the waves should have a peak close to 10 Hz, and he inferred that a similar shape of the power spectra of electroencephalogram (EEG) is due to mutual entrainment in coupled neural oscillators (Wiener, 1958). Buck and Buck reported that rhythmical flashes of South Asian fireflies were mutually synchronized (Buck & Buck, 1976).

Mutual entrainment of coupled limit cycle oscillators has been studied by Winfree (2000) and also by Kuramoto, who considered a coupled phase oscillator model, noting the neutrality of phase variables (Kuramoto, 1984). The simplest model with global coupling has the form

$$\dot{\phi}_i = \omega_i + \frac{K}{N}\sum_{j=1}^{N}\sin(\phi_j - \phi_i), \qquad (2)$$

where ϕ_i and ω_i represent the phase and the natural frequency of the ith oscillator, N is the total number of oscillators, and K is a coupling constant.

For $K < K_c$, the motion of each oscillator is independent and the frequency of the ith oscillator is the same as ω_i. However, for $K > K_c$, collective oscillation appears and a number of oscillators are entrained to the collective oscillation. Figure 1 displays a typical frequency distribution for $K < K_c$ and $K > K_c$. The δ-function peak in the frequency distribution implies mutual entrainment and a depression is seen around the deserved frequency for $K > K_c$.

The Josephson junction is a quantum device composed of two weakly coupled superconductors. With the current bias current below a critical value, the superconducting current flows without a voltage drop. If the bias is above the critical current, the phase difference (ϕ) between the Josephson junction is not constant in time, and the voltage drop (V) between the Josephson junction equals $\hbar\dot{\phi} = 2\,\mathrm{eV}$. This is called the AC Josephson effect. Thus the Josephson junction behaves as a kind of limit cycle oscillator above the critical current. If microwaves with frequency ω_0 are applied to the Josephson junction, $n:1$ frequency locking occurs, and the voltage becomes $V = n\hbar\omega_0/2e$. With N Josephson junctions coupled in series, the total voltage across the array is given by $V = Nn\hbar\omega_0/2e$. Such series arrays are currently used to establish the international standard of voltage (*See* **Josephson junction arrays**).

Hidetsugu Sakaguchi

See also **Chaotic dynamics; Phase dynamics; Synchronization; Van der Pol equation**

Further Reading

Buck, J. & Buck, E. 1976. Synchronous fireflies. *Scientific American*, 234: 74–85

Kuramoto, Y. 1984. *Chemical Oscillations, Waves, and Turbulence*, Berlin: Springer

van der Pol, B. 1934. The nonlinear theory of electric oscillations. *Proceedings of the IRE*, 22: 1051–1086

Wiener, N. 1958. *Nonlinear Problems in Random Theory*, Cambridge, MA: MIT Press

Winfree, A.T. 2000. *When Time Breaks Down*, Berlin and New York: Springer

COUPLED SYSTEMS OF PARTIAL DIFFERENTIAL EQUATIONS

Coupled systems of nonlinear partial differential equations (PDEs) are often derived to simplify complicated systems of governing equations in theoretical and applied sciences (Engelbrecht et al., 1988). Nonlinear electromagnetic theory, fluid dynamics, and systems in general relativity are difficult computational problems even with the help of numerical algorithms and the latest computer technologies. Using additional assumptions on properties of nonlinear wave processes in physical systems, however, one can derive coupled systems of nonlinear PDEs from the original governing equations, which simplify the analysis.

The main effects of nonlinear waves (such as nonlinearity, dispersion, diffraction, diffusion, damping and driven forces, and resonances) can be described with coupled nonlinear PDEs. Such systems may exhibit simple solutions such as traveling solitary waves and periodic waves, and some can be solved with the inverse scattering transform methods. Coupled systems comprise various combinations of nonlinear evolution equations that describe long solitary waves (Korteweg–de Vries and Boussinesq equations), envelope waves (nonlinear Schrödinger equations), kinks and breathers (sine-Gordon equations), and traveling fronts and pulses (reaction-diffusion systems). Here, we present a few examples.

Long surface water waves occur in oceans, seas, and lakes. The tsunami wave (Bryant, 2001) is an example of a nonlinear surface wave that arises following underwater earthquakes or underwater volcano eruptions and may reach heights of 20–30 m as it comes ashore. Because tsunamis are as long as tens and hundreds of kilometers, the ocean can be considered as shallow for such waves. This shallow-water approximation reduces the Euler equations for water waves to the Boussinesq system of coupled PDEs (Whitham, 1974):

$$\begin{aligned} u_t + uu_x + g\eta_x - (h^3/3)u_{txx} &= 0, \\ \eta_t + hu_x + \eta u_x + u\eta_x &= 0, \end{aligned}$$

where $\eta = \eta(x,t)$ is the wave surface elevation, $u = u(x,t)$ is the horizontal velocity, h is the water depth, and g is the gravitational acceleration.

The linear Boussinesq equation takes the form of the wave equation: $\eta_{tt} - c^2\eta_{xx} = 0$, which exhibits a two-wave solution $\eta = f(x-ct) + g(x+ct)$, where $f(x)$, $g(x)$ are arbitrary functions and $c = \sqrt{gh}$ is the wave speed. When the two waves are separated in space, small nonlinearity and dispersion are captured in the unidirectional Korteweg–de Vries (KdV) equation (Johnson, 1997):

$$\eta_t + c\left(1 + \frac{3\eta}{2h}\right)\eta_x + \frac{ch^2}{6}\eta_{xxx} = 0.$$

Different modes of long weakly nonlinear waves may travel with the same speed, exchanging energy by means of wave resonances. Because ocean water is stratified in density and shear flow, gravity waves can propagate along internal interfaces of the ocean stratification. Resonant interaction of internal wave modes in stratified shear flows is described by the system of coupled KdV equations (Grimshaw, 2001):

$$\boldsymbol{u}_t + A\boldsymbol{u}_x + B(\boldsymbol{u})\boldsymbol{u}_x + C\boldsymbol{u}_{xxx} = 0,$$

where A, $B(\boldsymbol{u})$, C are matrices and $\boldsymbol{u} = \boldsymbol{u}(x,t)$ is the vector for amplitudes of different internal wave modes.

Optical pulses may consist of electromagnetic waves in optical fibers, waveguides, and transmission lines. The propagation of optical pulses due to a balance between nonlinearity and dispersion is based on the paraxial approximation of the Maxwell equations with nonlinear refractive indices. This perturbation technique results in the nonlinear Schrödinger (NLS) equation (Newell & Moloney, 1992):

$$\mathrm{i}\psi_t + \tfrac{1}{2}\omega''(k_0)\psi_{xx} + \gamma(k_0)|\psi|^2\psi = 0,$$

where $\psi = \psi(x,t)$ is the envelope amplitude of a wave packet with the carrier wave number k_0. Depending on the relative signs of the dispersion coefficient $\omega''(k_0)$ and the nonlinearity coefficient $\gamma(k_0)$, wave perturbations are focused or defocused in the time evolution of the NLS equation. Interactions between waves with two orientations of polarization (ψ_y and ψ_z with propagation in the x-direction) can be represented in a normalized form as

$$\begin{aligned} \mathrm{i}\psi_{y,t} + \psi_{y,xx} + 2(|\psi_y|^2 + |\psi_z|^2)\psi_y &= 0 \\ \mathrm{i}\psi_{z,t} + \psi_{z,xx} + 2(|\psi_y|^2 + |\psi_z|^2)\psi_z &= 0. \end{aligned}$$

These are a pair of coupled NLS equations that are integrable by the inverse scattering transform method and display vector solitons (Manakov, 1974). Under collisions, the polarization vectors of two vector solitons change.

In wavelength-division-multiplexing optical systems, optical signals are transmitted through parallel channels at different carrier wave numbers (up to 40 channels in latest communication lines). Incoherent interaction of optical pulses at nonresonant frequencies is described by the system of coupled NLS equations (Akhmediev & Ankiewicz, 1997):

$$\mathrm{i}\boldsymbol{\psi}_t + D\boldsymbol{\psi}_{xx} + E(|\boldsymbol{\psi}|^2)\boldsymbol{\psi} = 0,$$

where D and $E(|\boldsymbol{\psi}|^2)$ are matrices and $\boldsymbol{\psi} = \boldsymbol{\psi}(x,t)$ is the vector for optical pulses in different channels. If the coupling between optical pulses is coherent (as in birefringent fibers, waveguide couplers, phase mixers, and resonant optical materials), the system of coupled NLS equations takes a general form:

$$\mathrm{i}\boldsymbol{\psi}_t + A\boldsymbol{\psi} + \mathrm{i}B\boldsymbol{\psi}_x + \mathrm{i}C(\boldsymbol{\psi})\boldsymbol{\psi}_x + D\boldsymbol{\psi}_{xx} + E(\boldsymbol{\psi})\boldsymbol{\psi} = \mathbf{0}.$$

The coupled NLS equations describe phase-matching resonance in quadratic χ^2 materials, gap solitons in periodic photonic gratings under Bragg resonance, Alfvén waves in plasmas, and other applications (Newell & Moloney, 1992).

In conservative nonlinear systems, wave dynamics of small amplitudes occur typically in the neighborhood of local minima of potential energy. Wave oscillations in the system of nonlinear massive pendulums are described by the Frenkel–Kontorova dislocation model. In a continuous approximation, the Frenkel–Kontorova lattice model reduces to the sine-Gordon (SG) equation (Braun & Kivshar, 1998):

$$\varphi_{tt} - c^2\varphi_{xx} + \sin\varphi = 0,$$

where φ is the angle between a pendulum and the vertical axis in a mechanical model. The nonlinear pendulums swing on a rigid rod under the gravity force and couple to each other with elastic springs. More complicated models of molecular crystals and ferromagnetics in solid state mechanics, stacked Josephson contacts in superconductivity, and strings in the general relativity theory are formulated as coupled systems of sine-Gordon equations (Maugin, 1999). The coupled Klein–Gordon equations take the form

$$\boldsymbol{\varphi}_{tt} - C\boldsymbol{\varphi}_{xx} + \boldsymbol{f}(\boldsymbol{\varphi}) = 0,$$

where C is a positive-definite matrix and $\boldsymbol{f}(\boldsymbol{\varphi})$ is the nonlinear vector function of components of the vector $\boldsymbol{\varphi} = \boldsymbol{\varphi}(x, t)$.

In more general systems, the energy of a nonlinear wave changes in time due to active and dissipative forces. The simplest system of this type is the nonlinear heat equation, which models the flame propagation (Zeldovich et al., 1985):

$$u_t = Du_{xx} + f(u),$$

where $u = u(x, t)$ is the temperature and D is the diffusivity constant. A complex form of the nonlinear heat equation (known as the Ginzburg–Landau equation) is derived for the amplitude of the most unstable wave mode (Newell & Moloney, 1992).

Active and dissipative systems include typically pairs of coupled activators and inhibitors. Coupled activator-inhibitor equations, known as the reaction-diffusion systems, are derived from the governing equations of thermodynamics in the form (Remoissenet, 1999):

$$\boldsymbol{u}_t = C\boldsymbol{u}_x + D\boldsymbol{u}_{xx} + \boldsymbol{f}(\boldsymbol{u}),$$

where C and D are matrices, and $\boldsymbol{f}(\boldsymbol{u})$ is a nonlinear vector function of components of the vector $\boldsymbol{u} = \boldsymbol{u}(x, t)$. Reaction-diffusion systems exhibit static, traveling, and pulsating nonlinear wave structures such as fronts and impulses. Coupled reaction-diffusion systems include the FitzHugh–Nagumo and Hodgkin–Huxley equations for nerve impulses, ephaptic coupling among nerve impulses, and models of the global dynamics of heart.

DMITRY PELINOVSKY

See also **Ephaptic coupling; Nonlinear optics; Reaction-diffusion systems; Sine-Gordon equation; Water waves**

Further Reading

Akhmediev, N. & Ankiewicz, A. 1997. *Solitons, Nonlinear Pulses, and Beams*, London: Chapman & Hall

Braun, O.M. & Kivshar, Yu.S. 1998. Nonlinear dynamics of the Frenkel–Kontorova model. *Physics Reports*, 306: 1–109

Bryant, T., 2001. *Tsunami: The Underrated Hazard*, Cambridge and New York: Cambridge University Press

Engelbrecht, J.K., Fridman V.E. & Pelinovski E.N., 1988. *Nonlinear Evolution Equations*, London: Longman and New York: Wiley

Grimshaw, R. (editor). 2001. *Environmental Stratified Flows*, Boston: Kluwer

Johnson, R.S. 1997. *A Modern Introduction to the Mathematical Theory of Water Waves*, Cambridge and New York: Cambridge University Press

Manakov, S.V. 1974. On the theory of two-dimensional stationary self-focusing of electromagnetic waves. *Soviet Physics, JETP*, 38: 248–253

Maugin, G.A. 1999. *Nonlinear Waves in Elastic Crystals*, Oxford and New York: Oxford University Press

Newell, A.C. & Moloney, J.V. 1992. *Nonlinear Optics*, Redwood City, CA: Addison-Wesley

Remoissenet, M. 1999. *Waves Called Solitons*. Berlin and New York: Springer

Whitham, G. 1974. *Linear and Nonlinear Waves*. New York: Wiley

Zeldovich, Ya.B., Barenblatt, G.I., Librovich, V.B. & Makhviladze, G.M. 1985. *The Mathematical Theory of Combustion and Explosions*. New York: Consultants Bureau

CRITICAL PHENOMENA

The term *critical phenomenon* is used synonymously with “phase transition,” which involves a change of one system phase to another and occurs at a characteristic temperature (called a transition temperature or a critical temperature: T_c). There are several different kinds of phase transitions such as melting, vaporization, and sublimation, as well as solid-solid, conducting-superconducting, and fluid-superfluid transitions. In systems undergoing phase transitions, an emergence of long-range order is seen in which the value of a physical quantity at one arbitrary point in the system is correlated with its value at another point a long distance away.

A classification scheme of phase transitions which remains the most popular was originally proposed by Paul Ehrenfest. According to this scheme, a transition for which the first derivative of the free energy with respect to temperature is discontinuous is called a first-order phase transition; thus, the heat capacity, C_p, at a first-order transition is infinite. A second-order phase transition is one in which the first derivative of the thermodynamic potential with respect to temperature is continuous, but its second derivative is discontinuous, so the heat capacity is discontinuous but not infinite at the transition. Near a second-order phase transition (due to the reduction of rigidity of the system), critical fluctuations dominate as their amplitudes diverge.

A useful concept in analyzing phase transitions is that of a critical exponent. In general, if a physical quantity $Q(T)$ either diverges or tends to a constant value (see Figure 1) as T tends to T_c, it can be characterized by defining the reduced temperature ε as

$$\varepsilon \equiv \frac{T - T_c}{T_c}. \qquad (1)$$

The associated critical exponent is

$$\mu = \lim_{\varepsilon \to 0} \frac{\ln Q(\varepsilon)}{\ln \varepsilon}. \qquad (2)$$

The most important critical exponents are denoted as α, β, γ, δ, υ, and η and describe the specific heat, order parameter, isothermal susceptibility, response to an external field, the correlation length, and the pair correlation function, respectively. (See Table 1 where

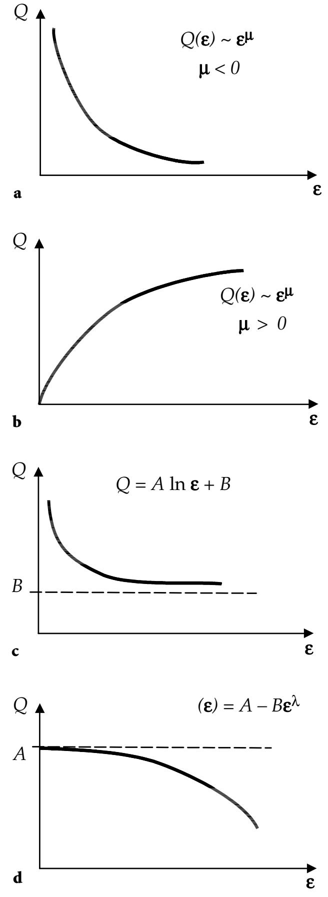

Figure 1. The four generic behaviors near criticality.

the primed exponents are introduced for temperatures below the critical temperature while the unprimed exponents are valid above the critical temperature.)

The mean field approximation (Landau theory) describes the physics of phase transitions well except in the immediate vicinity of the critical point where order parameter fluctuations are large. It is assumed that close to T_c, the free energy F can be expanded in a Taylor series of the order parameter ψ. Introducing the reduced temperature ε as a control parameter, the simplest such expansion is (see Figure 2)

$$F(T, V, \psi) = F_0 + a\varepsilon\psi^2 + A_4\psi^4, \tag{3}$$

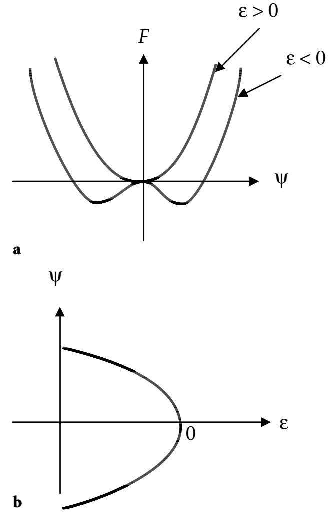

Figure 2. A prototype of second-order phase transitions according to Landau.

where $a > 0$ and $A_4 > 0$. Solving the equilibrium conditions for ψ yields $\psi = 0$ for $\varepsilon > 0$ and

$$\psi = \pm\left(-\frac{a\varepsilon}{2A_4}\right)^{1/2} \quad \text{for } \varepsilon < 0, \tag{4}$$

thus $\beta = 0.5$ is obtained. Calculating the entropy

$$S = \frac{\partial F}{\partial T} = S_0 + \frac{a^2}{2A_4}\frac{\varepsilon}{T_c} \quad \varepsilon \leq 0, \tag{5}$$

where for $\varepsilon > 0$, $S = S_0$ is the entropy of the disordered phase, which gives the specific heat as

$$C_v = T\frac{\partial S}{\partial T} = C_0 + \frac{a^2}{2A_4 T_c}T \quad \varepsilon \leq 0, \tag{6}$$

where for $\varepsilon > 0$, $C_v = C_0$ is the specific heat of the disordered phase. Hence, a discontinuity occurs at T_c (see Figure 3).

$$\Delta C = \frac{a^2}{2A_4 T_c}. \tag{7}$$

Thus, $\alpha = 0$.

Including in F an external field h coupled to ψ

$$F = F_0 + a\varepsilon\psi^2 + A_4\psi^4 - h\psi \tag{8}$$

Exponent	Definition (liquid-vapor) Specific heat at constant volume	Definition (magnetic) Specific heat at constant H
α'	$C_v \sim (-\varepsilon)^{-\alpha'}$	$C_H \sim (-\varepsilon)^{-\alpha'}$
α	$C_v \sim \varepsilon^{-\alpha}$	$C_H \sim \varepsilon^{-\alpha}$
	Density difference	Magnetization
β	$\rho_L - \rho_G \sim (-\varepsilon)^{\beta}$	$M \sim (-\varepsilon)^{\beta}$
	Isothermal compressibility	Isothermal susceptibility
γ'	$\kappa_T \sim (-\varepsilon)^{-\gamma'}$	$\chi_T \sim -\varepsilon^{-\gamma'}$
γ	$\kappa_T \sim \varepsilon^{-\gamma}$	$\chi_T \sim \varepsilon^{-\gamma}$
	Pressure-density critical isotherm	Magnetic field-magnetization
δ	$P - Pc \sim \lvert\rho_L - \rho_G\rvert^{\delta}$	$H \sim \lvert M\rvert^{\delta}$
	$(T = T_c)$	$(T = T_c)$
	Correlation length	Correlation length
ν'	$\xi \sim (-\varepsilon)^{-\nu'}$	$\xi \sim -\varepsilon^{-\nu'}$
ν	$\xi \sim \varepsilon^{-\nu}$	$\xi \sim \varepsilon^{-\nu}$
	Density–density pair correlation	Spin-spin pair correlation
η	$\Gamma(r) \sim \lvert r\rvert^{-(d-2+\eta)}$	$\Gamma(r) \sim \lvert r\rvert^{-(d-2+\eta)}$

Table 1. The definitions of critical exponents for liquid-vapor and magnetic systems. [The primed (unprimed) exponents are for temperatures below (above) T_c.]

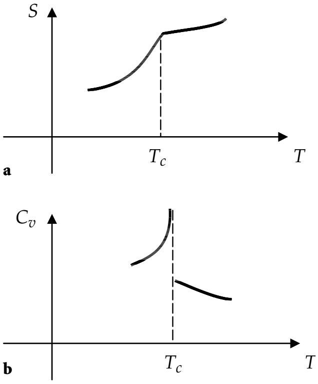

Figure 3. Plots of $S(T)$ and $C_v(T)$ in the Landau model of a second-order phase transition.

and minimizing F with respect to ψ yields an equation of state of the form

$$h = 2\psi\lfloor a\varepsilon + 2A_4\psi^2\rfloor. \tag{9}$$

Because the susceptibility $\chi \equiv \partial\psi/\partial h$, we find

$$\chi = [2a\varepsilon + 12A_4\psi^2]^{-1}. \tag{10}$$

As $\psi \to 0$ as $T \to T_c$, the third exponent is $\gamma = 1$. At $T = T_c$, the equation of state (9) simplifies to:

$$h \cong 4A_4\psi^3, \tag{11}$$

and hence $\psi \sim h^{1/3}$ giving $\delta = 3$. The quartic expansion in the Landau model invariably leads to the classical critical exponents: $\alpha = 0$, $\beta = 0.5$, $\gamma = 1$, and $\delta = 3$.

While the Landau theory cannot describe spatial fluctuations, following Ginzburg and Landau's proposal, it can be extended to consider the free energy to be a functional:

$$F(\psi(\vec{r}), T) = \int d^3\mathbf{r}[A_2\psi^2 + A_4\psi^4 - h\psi + D(\nabla\psi)^2] \tag{12}$$

where D describes the energy due to spatial inhomogeneities. Applying a variational principle to F results in a nonlinear Klein–Gordon equation for the order parameter

$$h = 2A_2\psi + 4A_4\psi^3 - 2D\nabla^2\psi. \tag{13}$$

A linearized solution of Equation (13) in spherical coordinates is

$$\psi = \frac{h_0}{4\pi D}\frac{e^{-r/\xi}}{r}, \tag{14}$$

where $\xi \sim A_2^{-1/2}$ is the correlation length that diverges as $T \to T_c$ so that the critical exponent $\upsilon = 0.5$. Fourier transforming the order parameter according to

$$\psi(\vec{r}) \equiv L^{-d/2}\sum_{k<k_0}\psi_k e^{i\vec{k}\cdot\vec{r}}, \tag{15}$$

where d is the spatial dimensionality and the cutoff wavelength $k_0 = \Lambda^{-1}$ corresponds to the smallest periodicity, F becomes (for $h = 0$)

$$F = \sum_{k<k_0}|\psi_k|^2(A_2 + Dk^2) + L^{-d}\sum_{k,k',k''<k_0} A_4\psi_k^*\psi_{k'}^*\psi_{k''}\psi_{k+k'-k''} \tag{16}$$

Ignoring mode-mode coupling provides the basis for a Gaussian approximation, where

$$F \cong \sum_{k<k_0} |\psi_k|^2 (A_2 + Dk^2). \qquad (17)$$

The Fourier transform of the correlation function is found as

$$\begin{aligned}\Gamma(k) &= L^{-d} \int \mathrm{e}^{-\mathrm{i}\vec{k}(\vec{r}-\vec{r'})} < \psi(\vec{r})\psi(\vec{r'}) > \,\mathrm{d}\vec{r}\,\mathrm{d}\vec{r'} \\ &= \langle |\psi_k|^2 \rangle \cong \frac{n}{2}(A_2 + Dk^2)^{-1} \qquad (18)\end{aligned}$$

Therefore, as $T \to T_\mathrm{c}$, we find that $\Gamma(k) \sim k^{\eta-2}$ with $\eta = 0$ in the Gaussian approximation and also $\Gamma(r) \sim r^{-(d-2)}$. Furthermore, the Gaussian approximation introduces a new value of the specific heat critical exponent $\alpha = 2 - d/2$.

JACK A. TUSZYŃSKI

See also **Bifurcations; Ferromagnetism and ferroelectricity; Order parameters; Phase transitions**

Further Reading

Anderson, P.W. 1984. *Basic Notions of Condensed Matter Physics*, Menlo Park, CA: Benjamin Cummings

Landau, L.D. & Lifshitz, E.M. 1959. *Statistical Physics*, London: Pergamon

Ma, S.-K. 1976. *Modern Theory of Critical Phenomena*, New York: Benjamin

Reichl, L.E. 1979. *A Modern Course in Statistical Physics*, Austin: University of Texas Press

Stanley, H.E. 1971. *Introduction to Phase Transitions and Critical Phenomena*, Oxford: Clarendon Press and New York: Oxford University Press

White, R.H. & Geballe, T. 1979. *Long Range Order in Solids*, New York: Academic Press

Yeomans, J.M. 1992. *Statistical Mechanics of Phase Transitions*, Oxford: Clarendon Press and New York: Oxford University Press

CRITICAL POINTS

See **Critical phenomena**

CRYSTAL DISLOCATIONS

See **Dislocations in crystals**

CRYSTALS, MOLECULAR

See **Local modes in molecular crystals**

CUBIC MAP

See **Maps**

D

DAMPED-DRIVEN ANHARMONIC OSCILLATOR

Many problems in engineering and the applied sciences are modeled by low-dimensional forced oscillators. Examples include vibrating mechanical and structural engineering systems (beams, bridges), marine systems (ships, oil platforms), electronic circuits and devices (Josephson junctions), as well as biological oscillators. While the linear behavior of such systems has long been well understood, the potential dynamics of nonlinear oscillators is far more complex and many open problems remain.

Damped Nonlinear Pendulum

Let us consider a mathematical pendulum and denote with θ the angular displacement of the pendulum from the equilibrium (hanging down) position. The equation of motion for the oscillations of a damped nonlinear pendulum of unitary mass is

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \eta\frac{\mathrm{d}\theta}{\mathrm{d}t} + \frac{g}{l}\sin(\theta) = 0, \qquad (1)$$

where l is the length of the pendulum and g is the gravitational acceleration. This equation expresses Newton's second law with the terms on the left-hand side, respectively, representing acceleration, damping, and gravitation. Equation (1) may be rewritten as two first-order equations:

$$\begin{cases} \dot{\theta} = \omega, \\ \dot{\omega} = -\eta\omega - \frac{g}{l}\sin(\theta), \end{cases} \qquad (2)$$

where ω denotes the angular velocity. Then, one can use (θ, ω) as phase space coordinates.

Fixed points: The dynamics of the nonlinear pendulum can be analyzed as trajectories in the phase space. The rest position $\theta = \omega = 0$ is called a *fixed point*. It is easy to see that almost all phase space trajectories spiral into the fixed point at $(0, 0)$. This is a linearly stable fixed point, since if a small perturbation is made from the fixed point, the perturbation decays in time (exponentially for small enough perturbations). There is a second fixed point at $(\pi, 0)$ corresponding to the pendulum pointing vertically up. The fixed point at $(\pi, 0)$ is linearly unstable because a small perturbation from this fixed point grows exponentially. Only very carefully tuned initial conditions will lead to a trajectory ending on the unstable fixed point, and almost all perturbations to the initial condition will lead to a trajectory that may approach close to the unstable fixed point, but eventually spirals into the stable fixed point. The $(0, 0)$ fixed point is attracting, and in this case the basin of attraction, that is, the set of initial conditions leading to trajectories that approach the fixed point, is the whole phase space except for a set of zero measure.

An example of a trajectory of system (2) converging towards the stable fixed point at $(0, 0)$ is shown in Figure 1.

The Periodically Driven Damped Pendulum

Let us now consider the nonlinear damped pendulum of Equation (1) when a harmonic driving force is applied. The equation of motion for the oscillations of a damped, sinusoidally driven nonlinear pendulum is

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \gamma\frac{\mathrm{d}\theta}{\mathrm{d}t} + \sin(\theta) = A_{\mathrm{D}}\cos(\omega_{\mathrm{D}}t), \qquad (3)$$

where A_{D} is the forcing amplitude and the angular velocity of the forcing, ω_{D}, may be different from the natural frequency of the pendulum. Note, with respect to Equation (1), that the time variable has been rescaled so that the period of small oscillations of the undamped and undriven pendulum is unity, and the (scaled) dissipation coefficient is now denoted by γ. Physically, the situation corresponds to driving the pendulum, feeding in energy to resupply the energy dissipated.

Small-Amplitude Oscillations

For small driving amplitudes, A_{D}, and assuming a small initial condition, one can approximate $\sin\theta \approx \theta$ and

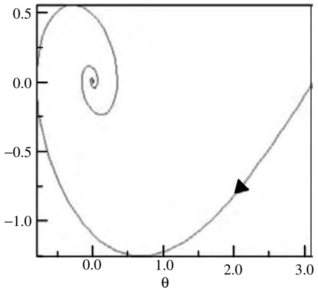

Figure 1. Typical trajectory for the nonlinear pendulum of Equation (2). All initial conditions lead to an orbit spiraling into the fixed point at (0, 0).

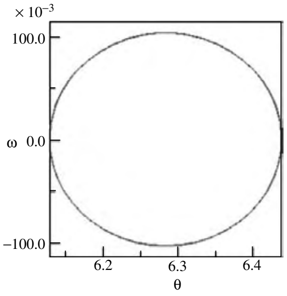

Figure 2. If small driving is turned on ($A_D = 0.1$), the steady-state solution is a stable limit cycle. Note that, for simplicity, the transient dynamics are not shown.

solve Equation (3) analytically:

$$\theta = \frac{A_D}{\sqrt{(1-\omega_D^2)^2 + \gamma^2\omega_D^2}} \cos(\omega_D t + \Phi) + A_0 e^{-\gamma t/2}\cos(\omega t + \varphi_0), \tag{4}$$

where

$$\tan\Phi = -\frac{\gamma\omega_D}{(1-\omega_D^2)}, \quad \omega = \sqrt{1-\frac{\gamma^2}{4}}, \quad \gamma < 2 \tag{5}$$

while the constants A_0 and φ_0 depend upon the initial conditions (θ_0, ω_0), that is, for $t = 0$.

Equation (4) describes the resonant response (the first term) oscillating at the applied frequency, together with decaying free oscillations (the second term), depending on the initial conditions. This kind of solution is called an attracting *limit cycle*.

A typical limit cycle at this small amplitude, for $A_D = 0.1$, is shown in Figure 2 and is seen to be close to circular. From now on, although the transient may be quite complicated, only the asymptotic behavior, that is, the behavior of the attractors will be considered.

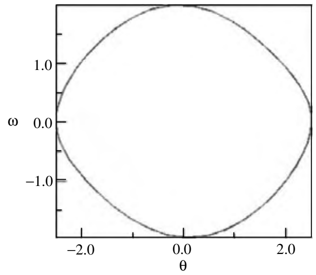

Figure 3. Large-amplitude limit cycle, for $A_D = 0.9$.

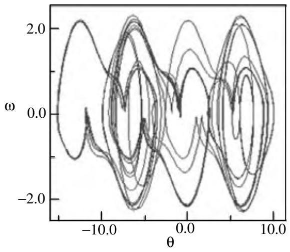

Figure 4. Chaotic attractor, for driving amplitude $A_D = 1.15$.

Large-Amplitude Solutions

What happens for large driving amplitudes? Here, there are no analytic solutions, and one must proceed numerically. To gain some intuition it is more convenient to view the dynamics in the phase space. To this end, Equation (3) may be converted to autonomous form by using three variables, as follows:

$$\begin{cases} \dot{\theta} &= \omega, \\ \dot{\omega} &= -\gamma\omega - \sin(\theta) + A_D\cos(\theta_D), \\ \dot{\theta}_D &= \omega_D, \end{cases} \tag{6}$$

where the phase of the driving, θ_D, has been introduced.

If one increases the driving strength to $A_D = 0.9$, a large-amplitude limit cycle is observed, as shown in Figure 3. This limit cycle bifurcates for increasing values of the driving strength according to the well-known period-doubling route to chaos, leading to a chaotic attractor for $A_D = 1.09$.

Then, increasing the driving to $A_D = 1.15$ results in an even more irregular chaotic behavior, as illustrated in Figure 4. But, perhaps at first sight surprisingly, increasing the driving further, for example, to $A_D = 1.35$ gives periodic motion again. In this case, though, there is a drift along with the oscillation—the pendulum is oscillating but also consistently rotating all the way around.

In summary, numerical simulations show that both periodic and chaotic solutions of the driven damped pendulum are possible, depending on the particular choice of the system parameters. Chaos may indeed occur since the driven pendulum equation is equivalent to a three-dimensional nonlinear autonomous system. Note the contrast with linear theory, where a damped system driven periodically *must* eventually respond periodically.

Gian Mario Maggio and Ljupco Kocarev

See also **Chaotic dynamics; Coupled oscillators; Distributed oscillators; Pendulum; Relaxation oscillators; Semiconductor oscillators**

Further Reading

Andronov, A.A., Vitt, A.A. & Khaikin, S.E. 1966. *Theory of Oscillators*, New York: Dover

Baker, G.L. & Gollub, J.P. 1996. *Chaotic Dynamics*, Cambridge and New York: Cambridge University Press

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, New York: Springer

Hagedorn, P. 1988. *Non-Linear Oscillations*, Oxford: Clarendon Press

Mitropolskii, Y.A. & Van Dao, N. 1997. *Applied Asymptotic Methods in Nonlinear Oscillations*, Dordrecht: Kluwer

Ott, E. 1993. *Chaos in Dynamical Systems*, Cambridge: Cambridge University Press

Parker, T.S. & Chua, L.O. 1989. *Practical Numerical Algorithms for Chaotic Systems*, New York: Springer

Thompson, J.M.T. & Stewart, H.B. 2002. *Nonlinear Dynamics and Chaos*, 2nd edition, New York: Wiley

DARBOUX TRANSFORMATION

From a historical point of view, the Darboux transformation is the specialization of a more general result due to Théodore Moutard. The prior result is of application in the classical theory of surfaces (Darboux, 1887–96) and takes the following form.

Let $\phi_0(x, y)$ be any particular integral of the hyperbolic partial differential equation

$$\phi_{xy} + q(x, y)\phi = 0, \tag{1}$$

$q(x, y)$ being a given function. We may rewrite this equation in the form

$$(D_x^+ D_y^- + D_y^+ D_x^-)\phi = 0, \tag{2}$$

where the Ds are differential operators, $D_u^{\pm} = \partial_u \pm \phi_0^{-1}\phi_{0u}$. Now let $\psi(x, y)$ and $\phi(x, y)$ be functions satisfying the pair of first-order partial differential equations

$$\begin{aligned} D_x^- \phi &= -D_x^+ \psi, \\ D_y^- \phi &= D_y^+ \psi, \end{aligned} \tag{3}$$

then (1) is equivalent to the identity $[D_x^+, D_y^+]\psi = 0$. Alternatively, the identity $[D_x^-, D_y^-]\phi = 0$ yields an equation for ψ, namely,

$$\psi_{xy} + p(x, y)\psi = 0, \tag{4}$$

where $p(x, y) = q(x, y) + 2(\ln \phi_0)_{xy}$. The class of equations of form (1) is thus covariant under maps $\mu_{\phi_0} : (\phi, q) \mapsto (\psi, p)$.

The map μ_{ϕ_0} is the *Moutard* transformation and its value lies in its relating, and generating, hyperbolic equations integrable in finite terms (Darboux, 1887–96).

Darboux (1882) specialized this result to the case where (1) carries a symmetry, that is, where $q(x, y) = Q(x + y)$ is a function of $x + y$ only. Putting $z = x + y$, $\bar{z} = x - y$ and applying the standard technique of separation of variables leads to $\phi(x, y) = \exp(\lambda\bar{z})\Phi(z, \lambda)$ and the equation for Φ,

$$\Phi_{zz} + Q(z)\Phi = \lambda^2\Phi. \tag{5}$$

Given a solution Φ_0 to this equation with $\lambda = \lambda_0$, there is a corresponding $\phi_0(x, y) = \exp(\lambda_0\bar{z})\Phi_0(z, \lambda_0)$ and after a simple, constant rescaling of Φ, Equations (3) become

$$\begin{aligned} \Psi &= (\partial_z - \Phi_0^{-1}\Phi_{0z})\Phi, \\ (\lambda^2 - \lambda_0^2)\Phi &= (\partial_z + \Phi_0^{-1}\Phi_{0z})\Psi \end{aligned} \tag{6}$$

and Equation (5) becomes

$$\Psi_{zz} + P(z)\Psi = \lambda^2\Psi. \tag{7}$$

Here $P(z) = Q(z) + 2(\ln \Phi_0)_{zz}$.

The map $\delta_{\lambda_0} : (\Phi, Q) \mapsto (\Psi, P)$ is the *Darboux* transformation (DT). Note that it creates from the Φ equation one member of a family of equations depending on the free parameter λ_0.

Starting with $Q = 0$ and a specific choice λ_0 for λ, we may take $\Phi_0 = c_1 \cosh \lambda_0(z + c_2)$. Then $P(z) = 2\lambda_0^2 \operatorname{sech}^2 \lambda_0(z + c_2)$ and so, from the trivial ("vacuum") potential, the DT has generated a family of potential barriers involving the chosen λ_0 and a parameter of translation, c_2, in the z-direction.

The linear PDE (5) plays a crucial role in the Inverse Scattering Transform (IST). We suppose the potential Q to depend upon an additional parameter t ("time"): $Q = Q(z, t)$. The pair of linear PDE (Lax pair),

$$L_Q\Phi = \lambda^2\Phi, \tag{8}$$

$$M_Q\Phi = \partial_t\Phi, \tag{9}$$

where $L_Q = \partial_z^2 + Q$ and $M_Q = 4\partial_z^3 + 6Q\partial_z + 3Q_z$, is integrable in the Cauchy sense provided Q satisfies the nonlinear PDE

$$Q_t = Q_{zzz} + 6QQ_z, \tag{10}$$

namely, the Korteweg–de Vries (KdV) equation. This follows from the operator relation (Frobenius condition)

$$L_{Qt} = [M_Q, L_Q] \equiv M_Q L_Q - L_Q M_Q. \qquad (11)$$

The DT, $\Psi = \mathcal{D}\Phi$ for $\mathcal{D} = \partial_z + \alpha$, $\alpha = -\Phi_0^{-1}\Phi_{0z}$, is seen to play the role of an intertwining operator:

$$\mathcal{D} L_Q = L_P \mathcal{D}, \qquad (12)$$

and

$$\partial_t \Psi = M_P \Psi + \{\alpha_t + \mathcal{D} M_Q - M_P \mathcal{D}\}\Phi. \qquad (13)$$

Some calculation shows that the bracketed term is vanishing and we have the Lax pair

$$L_P \Psi = \lambda^2 \Psi, \qquad (14)$$

$$M_P \Psi = \partial_t \Psi, \qquad (15)$$

in the DT transformed variables Ψ and P. Consequently, the transformed potential P will also satisfy the KdV equation (10).

Returning now to the simple example above, Φ_0 will satisfy Equation (9) with $Q = 0$, namely $\Phi_{0t} = 4\Phi_{0zzz}$, so that $\Phi_0 = c_1 \cosh(\lambda_0 z + 4\lambda_0^3 t + c_3)$ and the DT generates the traveling wave (1-soliton) solution to the KdV,

$$P = 2\lambda_0^2 \operatorname{sech}^2(\lambda_0 z + 4\lambda_0^3 t + c_3). \qquad (16)$$

Imagine now that we wish to form a sequence of DTs. Starting with (5), we make a choice Φ_0, λ_0 to form (7) via (6). Now for (7) we choose Ψ_1, λ_1 and use

$$\Theta = (\partial_z - \Psi_1^{-1}\Psi_{1z})\Psi$$

to generate Θ, satisfying

$$\Theta_{zz} + R(z)\Theta = \lambda^2 \Theta,$$

where $R = P + 2(\ln \Psi_1)_{zz}$. But since Ψ_1 satisfies (6) with $\lambda = \lambda_1$, it can itself be constructed as

$$\Psi_1 = \Phi_{1z} - \Phi_0^{-1}\Phi_{0z}\Phi_1,$$

where Φ_1 satisfies (5) with $\lambda = \lambda_1$. Then simple algebra yields formulae for Θ and R:

$$\Theta = \begin{vmatrix} \Phi_0 & \Phi_1 \\ \Phi_{0z} & \Phi_{1z} \end{vmatrix}^{-1} \begin{vmatrix} \Phi_0 & \Phi_1 & \Phi \\ \Phi_{0z} & \Phi_{1z} & \Phi_z \\ \Phi_{0zz} & \Phi_{1zz} & \Phi_{zz} \end{vmatrix}, \qquad (17)$$

$$R = Q + 2\ln\left(\begin{vmatrix} \Phi_0 & \Phi_1 \\ \Phi_{0z} & \Phi_{1z} \end{vmatrix} \right)_{zz}. \qquad (18)$$

From the symmetry of these formulae under interchange of λ_0 and λ_1, it follows that the order of composition of DTs is immaterial. The fundamental result of Crum (1955) is that, starting with $n + 1$ pairs (Φ_i, λ_i), $i = 0, \ldots, n$, associated with (5) one obtains the n-fold DT:

$$\begin{aligned} \delta_0 \ldots \delta_n(\Phi, Q) = {} & (W^{-1}(\Phi_0, \ldots, \Phi_n) \\ & \times W(\Phi_0, \ldots, \Phi_n, \Phi), Q \\ & + 2\ln W(\Phi_0, \ldots, \Phi_n)_{zz}), \end{aligned}$$

where W denotes the Wronskian function of its arguments.

This result, when applied to a set of n single soliton solutions of the KdV, yields an explicit n-soliton solution.

Generally speaking, a transformation of dependent variables that carries solutions of one DE to solutions of another is termed a *Bäcklund* transformation (BT). The term *Darboux* transformation refers to the special case of linear equations. We have described above a DT for a Lax pair that gives rise to a BT for the KdV equation.

For DTs of higher order, ordinary linear differential equations have been developed. For such results the intertwining property can be related to the factorization of differential operators, an approach to solvable Schrödinger potentials pioneered in quantum mechanics. This naturally leads to their algebraic formulation in terms of $\mathcal{D}$-modules. Moutard (sometimes called the Binary DT) and Darboux transformations have also been generalized to higher dimensions and utilized for solving associated analogues of the KdV with three independent variables, the Kadomtsev–Petviashvili, Novikov–Veselov, and Davey–Stewartson equations, to name but three.

The double logarithmic derivative occurring in the expression of the DT transformed potentials is of fundamental importance in the bilinear theory of soliton equations originally due to Hirota.

In the Hamiltonian theory of soliton equations the DT is seen to play the role of a canonical transformation.

A comparatively recent application of DTs has been in the theory of integrable systems with discrete variables, that is, systems defined on lattices (Antoniou & Lambert, 2000).

Finally, it should be remarked that the geometrical insights of Gaston Darboux, Luigi Bianchi, and their contemporaries are being rediscovered and developed within the modern theory of integrable systems and that Darboux and Moutard transformations are playing a key role in this resurgence (Rogers & Schief, 2002).

There are a number of monographs where further details about DTs and BTs may be found, for example, Calogero & Degasperis (1982); Konopelchenko (1987); Matveev & Salle (1991); Rogers & Shadwick (1982). For more recent reviews of work in this area see Coley (2001).

CHRIS ATHORNE

See also **Bäcklund transformations; Inverse scattering method or transform**

Further Reading

Antoniou, I. & Lambert, F. (editors). 2000. Integrability and chaos in discrete systems. Proceedings Brussels Meeting 1997. *Chaos, Solitons and Fractals*, 11(1–3)

Calogero, F. & Degasperis, A. 1982. *The Spectral Transform and Solitons*, Amsterdam: North-Holland

Coley, A. (editor). 2001. Bäcklund and Darboux transformations. The geometry of solitons, *Proceedings AARMS-CRM Workshop, Halifax, Nova Scotia, 1999*, Providence, RI: American Mathematical Society

Crum, M. 1955. Associated Sturm–Liouville systems. *Quarterly Journal of Mathematics*, 6: 121–128

Darboux, G. 1882. Sur une proposition relative aux équations linéaires. *Comptes Rendus de l'Academie des Sciences, Paris*, 94: 1343; 1456–1459

Darboux, G. 1887–96. *Leçons sur la Théorie des Surfaces*, Parts I–IV, Paris: Gauthier-Villars

Konopelchenko, B.G. 1987. *Nonlinear Integrable Equations*, Berlin and New York: Springer

Matveev, V.B. & Salle, M.A. 1991. *Darboux Transformations and Solitons*, Berlin and New York: Springer

Rogers, C. & Shadwick, W. 1982. *Bäcklund Transformations and their Applications*, New York: Academic Press

Rogers, C. & Schief, W.K. 2002. *Bäcklund and Darboux Transformations*, Cambridge and New York: Cambridge University Press

DAVEY–STEWARTSON EQUATION

See **Multidimensional solitons**

DAVYDOV SOLITON

In 1972, a meeting was convened in order to search for the general principles that underlie biological processes such as muscle contraction, active transport, enzyme catalysis, and oxidative phosphorylation (Green, 1974). One of the main questions asked was whether there was a crisis in bioenergetics. While some participants disagreed that a crisis existed, others proposed solutions to the crisis. One solution was the "conformon" of Green and Ji, defined as the free energy associated with a localized conformational strain, which was to store and transport energy (Green, 1974, pp. 419–437). McClare's paper, on the other hand, suggested that vibrational excited states and resonant energy transfer constitute a fundamental step in energy transduction in biological systems (Green, 1974, pp. 74–97). The main objection to McClare's proposal was that the lifetime of vibrational excited states, thought to be in the subpicosecond time range, was considered too short for these states to be useful. Alexander Davydov's model, proposed in connection with a mechanism for muscle contraction (Davydov, 1982), aimed to answer this objection.

The Davydov Model

The Davydov model describes energy transfer in the hydrogen-bonded spines that stabilize protein α-helices:

$$\begin{array}{ccc} \cdots\ \text{H-N-C=O}\ \cdots & \text{H-N-C=O}\ \cdots & \text{H-N-C=O}\ \cdots \\ n-1 & n & n+1 \end{array}$$

The idea is that the energy liberated in the hydrolysis of adenosine triphosphate (ATP) creates up to two quanta of amide-I, an excited vibrational state in the peptide group that is essentially a stretching vibration in the C=O bond (Davydov, 1982). This vibration excitation propagates from one group to the next because of the dipole-dipole interaction between the groups. But it also interacts with the neighboring hydrogen bonds, leading to a deformation of the lattice and a lower energy state. This new state, which is constituted by an amide-I excitation and its associated hydrogen bond distortion, is the Davydov soliton. Structurally, the Davydov soliton is, in fact, a subtle local conformational change of the α-helix, similar to the conformon.

Davydov's Hamiltonian has three parts:

$$\hat{H} = \hat{H}_{\text{ex}} + \hat{H}_{\text{int}} + \hat{H}_{\text{ph}}, \qquad (1)$$

where $\hat{H}_{\text{ex}}$, the exciton Hamiltonian, describes the transfer of an amide-I excitation between adjacent sites

$$\hat{H}_{\text{ex}} = -V \sum_{n=1}^{N} \left[\left(\hat{a}_n^{\dagger} \hat{a}_{n-1} + \hat{a}_n^{\dagger} \hat{a}_{n+1} \right) \right], \qquad (2)$$

$-V$ being the dipole-dipole interaction energy between neighboring sites, $\hat{a}_n^{\dagger}(\hat{a}_n)$ the creation (annihilation) operator for a quantum quasiparticle (the amide-I excitation) at site n, and N being the number of sites (the peptide groups) in the lattice; $\hat{H}_{\text{ph}}$, the phonon Hamiltonian, describes the vibrations of the peptide groups n in the one-dimensional chain:

$$\hat{H}_{\text{ph}} = \frac{1}{2} \sum_{n=1}^{N} \left[\kappa \left(\hat{u}_n - \hat{u}_{n-1} \right)^2 + \frac{\hat{P}_n^2}{M} \right], \qquad (3)$$

$\hat{u}_n$ being the displacement operator from the equilibrium position of site n, $\hat{P}_n$ the momentum operator of site n, M the mass of each site, and κ the elasticity constant of the lattice; $\hat{H}_{\text{int}}$, the interaction Hamiltonian, describes the interaction of amide-I excitation with the motions of the lattice sites:

$$\hat{H}_{\text{int}} = \chi \sum_{n=1}^{N} \left[\left(\hat{u}_{n+1} - \hat{u}_{n-1} \right) \hat{a}_n^{\dagger} \hat{a}_n \right], \qquad (4)$$

χ being an anharmonic parameter arising from the coupling between the vibrational excitation and the lattice displacements.

Davydov's Hamiltonian (Equations (1–4)) is formally similar to the Holstein Hamiltonian for the interaction of electrons with a polarizable lattice. The difference is that while Davydov's Hamiltonian couples quantum particles to acoustic phonons, the Holstein Hamiltonian couples quantum particles to optical phonons.

In Equations (1–4) both the excitation and the lattice are treated quantum mechanically. Exact wavefunctions for the system are yet to be determined in this approximation, and different solutions have been proposed in the literature (see references in Scott, 1992; Cruzeiro-Hansson & Takeno, 1997). On the other hand, when the lattice is treated classically, an approximation that has been called mixed quantum-classical (Cruzeiro-Hansson & Takeno, 1997), it is possible to determine exact wave functions. They are

$$|\psi(t)\rangle = \sum_{i=1}^{N} \varphi_i(\{u_m\},\{p_m\},t)\,\hat{a}_i^{\dagger}\,|0\rangle, \qquad (5)$$

where the dependence of the probability amplitude for an excitation at site *i* on the displacements and momenta of all the sites is not specified *a priori*. The equations of motion for the quantum variables are derived from Schrödinger's equation, while the equations of motion for the classical variables are derived from Hamilton's equations. For a one-quantum state they are

$$\begin{aligned} \mathrm{i}\hbar\frac{\mathrm{d}\varphi_n}{\mathrm{d}t} &= -V\,(\varphi_{n-1}+\varphi_{n+1}) \\ &\quad +\chi\,(u_{n+1}-u_{n-1})\,\varphi_n, \qquad (6) \\ M\frac{\mathrm{d}^2u_n}{\mathrm{d}t} &= \chi\,(|\,\varphi_{n+1}\,|^2-|\,\varphi_{n-1}\,|^2) \\ &\quad +\kappa\,(u_{n+1}+u_{n-1}-2\,u_n) \qquad (7) \end{aligned}$$

The numerical simulation of these equations (and of extensions of these equations to the three interacting chains that are present in an α-helix (Scott, 1992)) confirmed Davydov's analytical studies in the continuum limit by showing that solitons can form and travel along the chains. It was also found that in the discrete chains there is a threshold in the nonlinearity parameter χ above which localized excitations can form. This threshold tends to zero as the length of the system increases.

Thermal Stability of Localized States

While the zero temperature results summarized above allow for the existence of soliton states in proteins, an important question is whether these states are stable at biological temperatures and, if not, how long they last. Coupling to a thermal bath was modeled by extending the equations of motion into Langevin equations, that is, adding stochastic forces $F(n)$ and damping terms $-\Gamma \mathrm{d}u_n/\mathrm{d}t$ to the right-hand side of Equation (7), so that these terms obey the fluctuation-dissipation relations (Lomdahl & Kerr, 1985): $\langle F_n(t)\,F_m(t')\rangle \geq 2M\Gamma k_{\mathrm{B}}T\delta_{nm}\delta(t-t')$, k_{B} being the Boltzmann constant and T the temperature. The result is that soliton solutions disperse in a few picoseconds at biological temperatures. On the other hand, exact quantum Monte Carlo (QMC) simulations (Wang et al., 1989) showed that the distortion induced by the excitation increased with temperature. This conflict was resolved by showing that the coupling of a classical bath to a mixed quantum-classical system leads to a classical behavior of the quantum part (Cruzeiro-Hansson & Takeno, 1997). While the states of a classical excitation at finite temperature are predominantly delocalized, the states of a quantum excitation are predominantly localized. A set of dynamic equations, that leads to the same equilibrium averages as the QMC, leads to the following dynamics (Cruzeiro-Hansson & Takeno, 1997): instead of coherent propagation along the chain, as is expected from solitons, localized amide-I excitations jump stochastically from site to site, changing their shape and their velocity as they travel. It has been suggested that this stochastic propagation is a very robust way of transferring energy in proteins because it can survive mutations more than coherent propagation. Thus these simulations indicate that the Davydov model is indeed a possible mechanism for energy transfer in proteins (Cruzeiro-Hansson & Takeno, 1997).

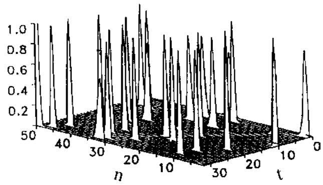

Figure 1. Dynamics of an amide-I excitation at biological temperatures.

Experimental Evidence

The first evidence for a Davydov-like state was obtained in a crystal of acetanilide (ACN), which includes hydrogen-bonded chains identical to those found in proteins. Careri and co-workers found an anomalous line, red-shifted by 15 cm^{-1} with respect to the amide-I vibration in that crystal (Scott, 1992). A series of studies established that this spectral line could not be attributed to a Fermi resonance, a Davydov splitting, or a phase transition. Its interpretation in terms of an interaction of an amide excitation with optical phonons provided good fits of the variation of its intensity with temperature as well as of its overtone frequencies. Careri and Wyman (see Scott, 1992) integrated the Davydov model in the so-called "turning wheel" model of enzyme action in which it is assumed that the binding of ligands leads to the formation of solitons that are used to promote the enzymatic cycle.

An early objection to the role of vibrational excited states in proteins was their short lifetime, which was supposed to be in the subpicosecond range. Recently, Austin and co-workers have measured the lifetime of amide-I vibrations in myoglobin and found it to be 15 ps (Xie et al., 2000). Hamm and co-workers have also recently applied nonlinear spectroscopy to the study of ACN, not only to the amide-I band (Edler & Hamm, 2002) but also to the NH vibrations (Edler et al., 2002) that have approximately twice the energy of amide-I and a binding energy 20 times larger. In a pump-probe femtosecond spectroscopy experiment, they found that the ground state recovery for NH at room temperature is 20 ps and for amide-I at 90 K, it is 35 ps, and that the lifetime of NH stretch in the crystal is 20 times greater than in isolated ACN. Computer simulations show that vibrational excitations can travel tens of nanometers within the lifetime of these excitations (Cruzeiro-Hansson & Takeno, 1997). Edler & Hamm (2002) also find that while the low-temperature amide-I excitations are self-trapped, the higher temperature ones are Anderson-localized, in agreement with the predictions of the computer simulations (Cruzeiro-Hansson & Takeno, 1997).

Future Work

The theoretical work on the Davydov model shows that localized vibrational excited states can form in proteins and that they can be used to transfer energy from the active site to other regions of the proteins where this energy is used for work. The experimental evidence available corroborates this possibility. Further experiments should test the role of excited vibrational states in protein function in a more direct way. A theoretical challenge is to integrate the Davydov model in the full protein cycle and find out how vibrational energy transfer may eventually lead to a conformational change (Cruzeiro–Hansson & Silva, 2001), the well-known way in which proteins fold or transform into their permanent three-dimensional shapes.

LEONOR CRUZEIRO

See also **Biomolecular solitons; Excitons**

Further Reading

Cruzeiro-Hansson, L. & Takeno, S. 1997. Davydov model: the quantum, mixed quantum-classical and full classical systems. *Physical Review*, E 56(1): 894–906

Cruzeiro-Hansson, L. & Silva, P.A.S. 2001. Protein folding: thermodynamic versus kinetic control. *Journal of Biological Physics*, 27: S6–S9

Davydov, A.S. 1982. *Biology and Quantum Mechanics*, Oxford–Frankfurt: Pergamon

Edler, J. & Hamm, P. 2002. Self-trapping of the amide I band in a peptide model crystal. *Journal of Chemical Physics*, 117: 2415–2424

Edler, J., Hamm, P. & Scott, A.C. 2002. Femtosecond study of self-trapped vibrational excitons in crystalline acetanilide. *Physical Review Letters*, 88(6): 067403

Green, D.E. (editor). 1974. The mechanism of energy transduction in biological systems. *Annals of the New York Academy of Sciences*, 227: 1–668

Lomdahl, P.S. & Kerr, W.C. 1985. Do Davydov solitons exist at 300K? *Physical Review Letters*, 55(11): 1235–1238

Scott, A. 1992. Davydov's soliton. *Physics Reports*, 217(1): 1–67

Wang, X., Brown, D.W. & Lindenberg, K. 1989. Quantum Monte Carlo simulation of the Davydov model. *Physical Review Letters*, 62(15): 1796–1799

Xie, A.H., van der Meer, L., Hoff, W. & Austin, R.H. 2000. Long-lived amide I vibrational modes in myoglobin. *Physical Review Letters*, 84: 5435–5438

DELAY COORDINATES

See **Embedding methods**

DELAY-DIFFERENTIAL EQUATIONS

A delay-differential equation (DDE) comprises an unknown function and certain of its derivatives, evaluated at arguments that differ by fixed numerical values. For example, $\dot{x}(t) = F(t, x(t), x(t-r))$ is a retarded DDE for $r > 0$.

DDEs can be characterized by the number of variables, the differential order, the order of the highest derivative appearing, and the difference order (one less than the number of different arguments appearing). DDEs (also called functional differential equations or difference-differential equations) generalize the concept of differential equations by allowing the state of the system to depend on states different from the present one, via, for example, a feedback mechanism.

DDEs appear in many fields, from economics to biology and from chemistry to mathematics, as the following examples suggest.

Bacterial Infection

In a book on the study of malaria published in 1911 (Ross, 1911), Ross considered a differential equation model of malarial epidemics based on the interaction of two populations, that of infected humans, described by $h(t)$, and that of infected mosquitoes, by $m(t)$. The relevant nonlinear DDEs are

$$\dot{h}(t) = m(t-u)\frac{p - h(t-u)}{p} - (M+r)h(t),$$

$$\dot{m}(t) = h(t-v)\frac{q - m(t-v)}{p} - (N+s)m(t), \quad (1)$$

with p and q being the total populations, M and N the death rates, and r and s the recovery rates for human and mosquito populations, respectively. The delay from the time of a bite to the time at which the human or the mosquito is infective is $u = 0.5$ months for humans and $v = 0.6$ months for mosquitoes.

Enzyme Catalysis

Roussel (1996) presents a model of enzyme catalysis, proposed by Adrian Brown in 1902, in which the enzyme-substrate complex has a fixed lifetime τ before giving rise to its reaction products

$$E(t)+S(t)\rightarrow P(t+\tau)+E(t+\tau). \qquad (2)$$

The DDEs generated by this model are

$$\begin{aligned}\frac{\mathrm{d}S(t)}{\mathrm{d}t}&=-kE(t)S(t),\\ \frac{\mathrm{d}E(t)}{\mathrm{d}t}&=-kE(t)S(t)+kE(t-\tau)S(t-\tau), \quad (3)\\ \frac{\mathrm{d}P(t)}{\mathrm{d}t}&=kE(t-\tau)S(t-\tau),\end{aligned}$$

where k is the rate constant and by $E(t)$, $P(t)$, and $S(t)$, we mean the enzyme, product, and substratum concentrations.

Economic Dynamics

Delays can emerge in the dynamics of economic processes in two ways. First, there is a lag between the time economic decisions are taken and the time the decisions bear fruit. Also, the expected future values of a variable are functions of its current and past values. Let us consider a world in which infinitely many houses are inhabited, and let us consider the problem of finding the best time for constructing a new house (Asea & Zak, 1999).

The individuals' preferences are represented by a continuous, strictly increasing, and concave utility function $U(c(t))$. Subjective discount rates are characterized by $\rho>0$. In this model, it takes $r>0$ periods to obtain new capital. The solution of this planning problem is given by

$$\max_{\{c(t)\}}\int_0^\infty U(c(t))\mathrm{e}^{-\rho t}\mathrm{d}t, \qquad (4)$$

subject to the DDE

$$\frac{\mathrm{d}k(t)}{\mathrm{d}t}=f(k(t-r))-\delta k(t-r)-c(t), \qquad (5)$$

with initial condition $k(t)=\phi(t)$, for all $t\varepsilon[-r,0]$. $f(\cdot)$ is the production function; $c(t)$, the rate at which capital depreciates, such that $0<c(t)\leq f(k(t-r))$, $\delta\in[0,1]$; and $k(t)$, the productive capital stock at time t.

Mathematics

If one considers the symmetry reduction of a nonlinear differential-difference equation with respect to a combination of continuous and discrete symmetries, then the initial equation reduces to a DDE. As an example, consider the Toda lattice (Levi & Winternitz, 1993)

$$\ddot{u}_n(t)=\mathrm{e}^{u_{n-1}(t)-u_n(t)}-\mathrm{e}^{u_n(t)-u_{n+1}(t)}, \qquad (6)$$

and assume a reduction with respect to $\partial_n+a\partial_t$ where a is an arbitrary real parameter. Equation (6) then reduces to

$$\frac{\mathrm{d}^2u(\eta)}{\mathrm{d}\eta^2}=\mathrm{e}^{u(\eta+a)-u(\eta)}-\mathrm{e}^{u(\eta)-u(\eta-a)}, \qquad (7)$$

where $\eta=t-an$.

The equations considered in these examples are all instances of a general DDE, which, in the simple case of a linear first-order equation for just one field, can be written as

$$a_0\frac{\mathrm{d}u(t)}{\mathrm{d}t}+a_1\frac{\mathrm{d}u(t-\sigma)}{\mathrm{d}t}+b_0u(t)+b_1u(t-\sigma)=f(t). \qquad (8)$$

An equation of the form (8) is said to be a DDE of *retarded* type if $a_0\neq 0$ and $a_1=0$; it is said to be of *neutral* type if $a_0\neq 0$ and $a_1\neq 0$; and of *advanced* type if $a_0=0$ and $a_1\neq 0$.

In applications, an equation of *retarded* type may represent the behavior of a system in which the rate of change of $u(t)$ depends on its past and present values. A *neutral* equation represents a system in which the present rate of change depends on past rates of changes as well as its present and past values. An *advanced* type equation may represent a system in which its rate of change depends on its present and future values.

If $a_0=a_1=0$, Equation (8) is a pure difference equation, while if $a_0=b_0=0$ or $a_1=b_1=0$, it is a pure differential equation. In either case, $f(t)$ is a forcing function.

Let us compare the solution techniques for DDEs with those of ordinary differential equations (ODEs) and note some of their peculiar features. For simplicity we limit ourselves to retarded DDEs. For more details, see Bellman & Cooke (1963), Hale (1977), Hale & Verduyn Lunel (1993), Driver (1977), Bainov & Mishev (1991), Kuang (1993), and Gyori & Lada (1991).

Because retarded DDEs depend on previous history, the initial condition at one point is not sufficient to obtain the present time behavior. What one needs depends on the discrete order of the equation. If the equation is a DDE of first order, then the initial solution on a whole delay interval is needed.

For constant coefficient DDEs, an algebraic method of solution is provided by the "method of steps," which also provides a constructive proof of the existence of the solution. To illustrate this method, consider a DDE generalization of the logistic equation

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t}=-cx(t-1)[1+x(t)], \qquad t>0, \qquad (9)$$

with the initial condition $x(t) = \phi(t)$ for $t \in [-1, 0]$. To solve Equation (9), we divide the interval $[0, \infty)$ into steps of the size of the delay and solve recursively in each interval. We use the solution obtained in one interval to solve Equation (9) in the next one. For example, the solution in the interval $[0, 1]$ is given by

$$x(t) = [\phi(0) + 1]\mathrm{e}^{-c\int_0^t \phi(s-1)\mathrm{d}s} - 1,$$

which is obtained as a solution of the ODE

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -c\phi(t-1)[1 + x(t)]. \qquad (10)$$

For linear DDEs we can construct, as in the case of linear ODEs, the characteristic equation, by looking at exponential solutions. In this case, however, the characteristic equation is given by a nonlinear algebraic equation. For example in the case of Equation (8), with $a_1 = 0$, we have

$$h(\lambda) = a_0\lambda + b_0 + b_1 e^{-\lambda\sigma} = 0. \qquad (11)$$

Once the characteristic equation is solved, a particular solution of the DDE is obtained by applying the Laplace transform (Bellman & Cooke, 1963).

As we have seen, the nature of the method of solution of a DDE is similar to that of an ODE. Nevertheless, DDEs exhibit more complicated behaviors, even in the linear case. For example, scalar linear first-order homogeneous DDEs with real coefficients can have nontrivial oscillating solutions unlike ODEs (Kalecki, 1935). Moreover, solutions to DDEs may be discontinuous and, depending on the initial conditions, a solution may also not exist (Winston & Yorke, 1969). As in the case of ODEs, series solutions can be used to approximate solutions to nonlinear DDEs (Bellman & Cooke, 1963); however, the solutions obtained are often complicated and obscure.

We can gain a better insight into the solution using qualitative theory and stability analysis to obtain properties of the dynamics of a nonlinear DDE by looking at its linearization.

The stability of a fixed point of a DDE is defined by examining the roots of the characteristic equation $h(\lambda)$. Thus, a fixed point of a DDE is stable if all roots of $h(\lambda)$ have negative real parts. As the characteristic equation (11) is transcendental, it has an infinity of roots, and it is not guaranteed that all roots will have real parts, strictly negative or positive. So fixed points of DDEs will often be saddle points. Moreover, stability may depend crucially on the initial data (Driver, 1977).

The stability of homogeneous scalar DDEs of the first order has been studied by Hayes (Bellman & Cooke, 1963). These results can be extended to nonlinear systems by linearizing the DDE around a stable solution and then using a generalization of the Poincaré–Lyapunov theorem. In such a way, one can show that DDEs often admit periodic solutions after a sequence of Hopf bifurcations. Chaotic orbits may also exist, with the structure of the orbits depending critically on the smoothness of the feedback mechanism.

DECIO LEVI

See also **Bifurcations; Equations, nonlinear; Feedback; Hopf bifurcation; Integral transforms; Ordinary differential equations, nonlinear; Poincaré theorems; Quasilinear analysis; Stability; Symmetry: equations vs. solutions**

Further Reading

Asea, P.K. & Zak, P.J. 1999. Time-to-build and cycles. *Journal of Economic Dynamics & Control*, 23: 1155–1175

Bainov, D.D. & Mishev, D.P. 1991. *Oscillation Theory for Neutral Differential Equations with Delay*, Bristol: Adam Hilger

Bellman, R. & Cooke, K.L. 1963. *Differential-Difference Equations*, New York: Academic Press

Driver, R.D. 1977. *Ordinary and Delay Differential Equations*, New York: Springer

Gyori, I. & Ladas, P. 1991. *Oscillation Theory of Delay Differential Equations: with Applications*, Oxford: Clarendon Press

Hale, J.K. 1977. *Theory of Functional Differential Equations*, New York: Springer

Hale, J.K. & Verduyn Lunel, S.M. 1993. *Introduction to Functional Differential Equations*, New York: Springer

Kalecki, M. 1935. A macroeconomic theory of business cycles. *Econometrica*, 3: 327–344

Kuang, Y. 1993. *Delay Differential Equations with Applications in Population Dynamics*, Boston: Academic Press

Levi, D. & Winternitz, P. 1993. Symmetries and conditional symmetries of differential-difference equations. *Journal of Mathematical Physics*, 34: 3713–3730

Ross, R. 1911. *The Prevention of Malaria*, 2nd edition, London: John Murray

Roussel, M.R. 1996. The use of delay differential equations in chemical kinetics. *Journal of Physical Chemistry*, 100: 8323–8330

Winston, E. & Yorke, J.A. 1969. Linear delay differential equations whose solutions become identically zero. *Académie de la République Popolaire Roumaine*, 14: 885–887

DENJOY THEORY

The theory developed by Arnaud Denjoy (1884–1974) showed that any sufficiently smooth orientation-preserving diffeomorphism T of the unit circle S^1 with an irrational rotation number ρ is topologically equivalent to a linear rotation by the angle $2\pi\rho$ (Denjoy, 1932). Informally, diffeomorphism is a smooth invertible map such that its inverse is also smooth.

Circle diffeomorphisms arise naturally in many physical problems. For instance, in the case of Hamiltonian systems with two degrees of freedom, such diffeomorphisms appear as Poincaré first return maps for the two-dimensional invariant tori. When the rotation number is irrational, circle diffeomorphisms represent an important model for quasi-periodic dynamics (*See* **Quasiperiodicity**). The Denjoy theory implies the following important fact: if two smooth circle maps have the same irrational rotation number then the topological structure of their trajectories is exactly the same.

The topological equivalence means that circle diffeomorphisms are conjugated to a linear rotation with the help of a homeomorphic change of variables. Namely, there exists a homeomorphism, ϕ, which is an invertible map that is continuous together with its inverse, such that $T \circ \phi = \phi \circ T_\rho$, where T_ρ is the linear rotation by the angle $2\pi\rho$ and $\circ$ stands for a composition of two maps. Denjoy's theorem holds if T is absolutely continuous and $\log T'(x)$ has bounded total variation: $V = \mathrm{Var}_{S^1} \log T'(x) < \infty$. The last condition is satisfied if T is C^2-smooth and $T'(x) > 0$. The conjugacy ϕ is defined uniquely up to an arbitrary rotation T_α. In fact, a mapping ϕ of the unit circle S^1 that satisfies condition $T \circ \phi = \phi \circ T_\rho$ can be constructed for any quasi-periodic homeomorphisms T. This means that any homeomorphism T with irrational rotation number ρ is semiconjugate to T_ρ. However, if T is not regular enough, ϕ may not be a homeomorphism. To construct ϕ it is enough to take two arbitrary points x_0 and y_0 and define their forward trajectories by T and T_ρ, respectively: $x_i = T^i x_0$, $y_i = T_\rho^i y_0$, $i \geq 1$. Now one can define ϕ on $\{y_i\}$ by letting $\phi(y_i) = x_i$, $i \geq 0$ and extending ϕ by continuity to the whole unit circle. This can be done since any trajectory of a linear rotation by an irrational angle is everywhere dense. It is easy to see that a conjugacy ϕ is a homeomorphism if and only if T is transitive; that is, all its trajectories are dense on S^1.

When the total variation V is bounded, the transitivity of T follows from the Denjoy inequality:

$$\exp(-V) \leq \prod_{i=0}^{q_n-1} T'(x_i) \leq \exp(V),$$

where q_n are the denominators of the convergents $p_n/q_n = [k_1, k_2, \ldots, k_n]$, and $\rho = [k_1, k_2, \ldots, k_n, \ldots]$ is the continued fraction expansion for ρ. The condition $T \in C^2(S^1)$ that implies topological equivalence is almost sharp. Indeed, Denjoy constructed counterexamples where $T \in C^1(S^1)$ and the derivative $T'(x)$ is a Hölder continuous function with an arbitrary Höder exponent $0 < \alpha < 1$. In these examples T is not transient and, hence, is not conjugate to T_ρ.

An important extension of the Denjoy theory is connected with the problem of smoothness of the conjugacy ϕ. It is natural to ask when the homeomorphism ϕ is at least C^1-smooth, which implies not only topological but also asymptotic metrical equivalence between T and T_ρ. In this case, the unique probability invariant measure for T is absolutely continuous with respect to the Lebesgue measure. The first progress in this direction was made by Arnol'd (1961), who proved that for analytic diffeomorphisms, T that are close enough to the linear rotation T_ρ, a conjugacy ϕ is analytic provided the rotation number ρ is Diophantine, that is, $\|\rho - p/q\| \geq 1/q^{2+\delta}$ for some $\delta > 0$ and all integers p, q. Diophantine numbers form a set of positive Lebesgue measure and, hence, are typical in the Lebesgue sense. Arnol'd has also constructed counterexamples in the case of nontypical rotation numbers, which show that the smooth theory cannot be constructed for all irrational rotation numbers. In these counterexamples, ϕ is not differentiable, and the invariant measure for T is essentially singular with respect to Lebesgue measure. Arnol'd's results are of the KAM-type (Kolmogorov–Arnol'd–Moser) and, hence, have a local character. However, as it was conjectured by Arnol'd, in the one-dimensional case the local condition of T being close to T_ρ should not be necessary, and the global result should hold for all T smooth enough. Such a global result has been proven by Herman (1979) in the case when ρ satisfies certain Diophantine condition and $T \in C^3(S^1)$. Later Herman's results were extended to a wider class of rotation numbers (Yoccoz, 1984) and to diffeomorphisms $T \in C^{2+\varepsilon}(S^1)$ (Khanin & Sinai, 1987; Sinai & Khanin, 1989; Katznelson & Ornstein, 1989).

Finally, we mention another extension of the Denjoy theory to the case of diffeomorphisms with singularities. Such mappings appear, for example, in the case of critical invariant tori in Hamiltonian systems with two degrees of freedom. The extension of the Denjoy theory to this case is a subject of the so-called rigidity theory. The main aim is to find conditions which imply that two topologically equivalent homeomorphisms that have the same local structure of their singular points are, in fact, C^1-smoothly conjugate to each other. Significant progress in this direction has been made in the last 5 years in the case of mappings with one singular point (de Faria & de Melo, 1999, 2000; Yampolsky, 2001; Khanin & Khmelev, 2003). Note that the presence of singularities makes rigidity stronger than in the case of smooth diffeomorphisms. The arithmetical properties of the rotation numbers are less important, and one should expect C^1-rigidity for all irrational rotation numbers.

KONSTANTIN KHANIN

See also **Kolmogorov–Arnol'd–Moser theorem; Maps; Quasiperiodicity**

Further Reading

Arnol'd, V.I. 1961. Small denominators. I. Mapping the circle onto itself. *Izvestiya Akademii Nauk SSSR Seriya Mathematicheskaya*, 25: 21–86

Cornfeld, I. P., Fomin, S. V. & Sinai, Ya. G. 1982. *Ergodic Theory*, New York: Springer

Denjoy, A. 1932. Sur les courbes définies par les équations différentielles à la surface du tore. *Journal des Mathematiques Pures et Appliques*, ser. 9, 11: 333–375 de Faria, E. & de Melo, W. 1999. Rigidity of critical circle mappings. I. *Journal of the European Mathematical Society (JEMS)*, 1: 339–392

de Faria, E. & de Melo, W. 2000. Rigidity of critical circle mappings. II. *Journal of the European Mathematical Society (JEMS)*, 13: 343–370

Herman, M. 1979. Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 49: 5–233

Katznelson, Y. & Ornstein, D. 1989. The differentiability of the conjugation of certain diffeomorphisms of the circle. *Ergodic Theory & Dynamical Systems*, 9: 643–680

Khanin, K.M. & Sinai, Ya.G. 1987. A new proof of M. Herman's theorem. *Communications in Mathematical Physics*, 112: 89–101

Khanin, K. & Khmelev D. 2003. Renormalizations and rigidity theory for circle homeomorphisms with singularities of the break type. *Communications in Mathematical Physics*, 235: 69–124

Sinai, Ya.G. & Khanin, K.M. 1989. Smoothness of conjugacies of diffeomorphisms of the circle with rotations. *Russian Mathematical Surveys*, 44: 69–99

Yampolsky, M. 2001. The attractor of renormalization and rigidity of towers of critical circle maps. *Communications in Mathematical Physics*, 218: 537–568

Yoccoz, J.-C. 1984. Conjugaison différentiable des difféomorphismes du cercle dont le nombre de rotation vérifie une condition diophantienne. *Annales Scientifique de l'École Normale Superierieure*, 4(17): 333–359

DERIVATIVE NLS EQUATION

See **Nonlinear Schrödinger equations**

DERRICK–HOBART THEOREM

The Derrick–Hobart scaling argument concerns certain solutions of nonlinear partial differential equations that arise as models for elementary particles; thus they are mostly of the relativistic variety. To appreciate the context in which the argument arose and the way it is used, some introductory remarks on relativistic quantum field theory are in order.

There are only a few interacting relativistic quantum field theories that have been solved explicitly, in the sense that physically relevant quantities (particle spectrum, scattering, form factors, and so on) are known in closed form. For all of these models the dimension of space-time equals two. To gain more insight into higher-dimensional models, it has become standard practice to study the field theory first as a classical field theory. The underlying idea is that (via the Feynman path integral) one can use classical findings to obtain nonperturbative information on the quantum version. In particular, the presence of nonconstant, smooth, stable, time-independent, finite-energy, classical solutions is believed to signal the presence of an associated stable quantum particle.

The notion of "stability" refers to small fluctuations around such a classical finite-energy solution. To first order, such variations do not change the energy, as expressed by the Euler–Lagrange equation. To second order, however, the energy may become smaller, in which case the corresponding quantum particle is considered to be unstable. (Think of a ball resting on top of a hill. A little push makes it roll down.)

In order to study the existence of nonconstant finite-energy solutions (either stable or unstable), an argument due independently to Derrick (1964) and (in a somewhat different form) to Hobart (1963) is often useful. These authors were concerned with three-dimensional space (four-dimensional space-time), but the argument can be extended without difficulty to an arbitrary space dimension N. Briefly, the argument is as follows.

Assume $\phi(x,t)$, $x \in R^N$, $t \in R$, is a scalar field on $(N+1)$-dimensional space-time, whose dynamics is given by the Lagrangian

$$\mathcal{L} = \tfrac{1}{2}(\partial_t\phi^2 - \nabla\phi\cdot\nabla\phi) - V(\phi), \qquad (1)$$

with $V(y)$ being a potential function. Now let $\phi(x)$ be a (smooth) time-independent nonconstant solution to the Euler–Lagrange equation, with finite energy

$$E = E_{\text{kin}} + E_{\text{pot}}, \quad E_{\text{kin}} = \frac{1}{2}\int_{R^N} \nabla\phi\cdot\nabla\phi\,\mathrm{d}x,$$

$$E_{\text{pot}} = \int_{R^N} V(\phi)\,\mathrm{d}x. \qquad (2)$$

Starting from the above data, Derrick's key idea is to consider the family of scaled functions

$$\phi_\lambda(x) = \phi(\lambda x). \qquad (3)$$

Clearly, the energy associated with ϕ_λ is given by

$$E_\lambda = \lambda^{(2-N)}E_{\text{kin}} + \lambda^{-N}E_{\text{pot}}, \qquad (4)$$

so that

$$(\mathrm{d}E_\lambda/\mathrm{d}\lambda)_{\lambda=1} = (2-N)E_{\text{kin}} - NE_{\text{pot}}, \qquad (5)$$

$$(\mathrm{d}^2E_\lambda/\mathrm{d}\lambda^2)_{\lambda=1} = (2-N)(1-N)E_{\text{kin}} + N(N+1)E_{\text{pot}}. \qquad (6)$$

Since ϕ_λ makes the energy stationary for $\lambda = 1$, we have

$$(\mathrm{d}E_\lambda/\mathrm{d}\lambda)_{\lambda=1} = 0. \qquad (7)$$

Hence (5) yields

$$E_{\text{pot}} = \frac{2-N}{N}E_{\text{kin}}, \qquad (8)$$

which entails

$$(\mathrm{d}^2E_\lambda/\mathrm{d}\lambda^2)_{\lambda=1} = 2(2-N)E_{\text{kin}}. \qquad (9)$$

Let us now draw the relevant conclusions from this simple calculation. Since $\phi(x)$ is nonconstant, its kinetic energy E_{kin} is positive. For $N > 2$, then, (9) says that the finite-energy solution cannot be stable. This is the first consequence, an instability result for $N > 2$. It does not involve restrictions on the potential $V(y)$.

Assuming from now on that $V(y) \geq 0$, far stronger conclusions can be drawn. Indeed, since ϕ_λ is a solution for $\lambda = 1, \phi_1 = \phi$ makes the energy stationary. But since ϕ is nonconstant, we have $E_{\mathrm{kin}} > 0$, and since $V \geq 0$, we also have $E_{\mathrm{pot}} \geq 0$. Therefore, the right-hand side of (5) is negative for $N > 2$, a contradiction. A second consequence, therefore, is the absence of finite-energy nonconstant solutions for $V \geq 0$ and $N > 2$.

Retaining the assumption $V \geq 0$, one can draw a conclusion for $N = 2$, too. Indeed, it then follows that $E_{\mathrm{pot}} = 0$, so that ϕ must satisfy $V(\phi) = 0$; moreover, the second variation (6) vanishes.

For $N = 1$ the variation formulas (5), (6) have no useful consequences. Indeed, in two-dimensional space-time there do exist stable time-independent finite-energy solutions, as exemplified by the one-soliton and one-antisoliton solutions of the sine-Gordon theory.

In applications of Derrick's argument, one usually encounters positive potentials and invokes the latter consequences sketched above. Thus, it is used to the effect that for $N \geq 2$, time-independent finite-energy solutions must be constant (the so-called vacuum solutions). Some caveats should be heeded, however.

First, it is important to keep track of the above steps in models that are not of the above form, since the reasoning may need to be suitably modified. Second, even when this can be done at face value, it should be observed that the above argument, although convincing at first sight, is not a rigorous proof. Indeed, the scaling variation that is involved has a global character, whereas the Euler–Lagrange equation is derived by considering local variations. More in detail, one needs to control boundary terms that can *a priori* spoil the above derivation. (This was already realized in Hobart (1963).)

We exemplify these related issues with two models described by Lagrangians that are different from the above, namely a (special) Yang–Mills/Higgs model in physical space ($N = 3$) and a class of nonlinear σ-models for $N \geq 2$. In the first setting, explicit static finite-energy solutions were obtained in Prasad & Sommerfield (1975) and Bogomolnyi (1976). (These are nowadays called BPS monopoles.) The energy of these solutions is manifestly not scale-invariant, contradicting (7) for the case at hand. Inspection of the solution shows that this is due to poor decay at spatial infinity; it entails that the pertinent boundary term cannot be ignored.

Turning to $O(3)$ σ-models, one can once more study the issue of finite-energy solutions by adapting Derrick's scaling argument. For $N = 2$ (now viewed as Euclidean space-time) this yields no conclusion, since the energy is scale-invariant. In this case, the so-called instanton and anti-instanton solutions do exist, and they are stable for topological reasons.

For $N > 2$, the scaling argument leads to the absence of finite-energy solutions. In this particular setting, the heuristic reasoning can be corroborated. More specifically, the boundary term can be rigorously controlled. The pertinent result (Garber et al. (1979), Theorem 5.1) has later been used by differential geometers to prove the nonexistence of harmonic maps, which are closely related to the above type of solution.

SIMON RUIJSENAARS

See also **Matter, nonlinear theory of; Skyrmions; Virial theorem; Yang–Mills theory**

Further Reading

Bogomolnyi, E.B. 1976. The stability of classical solutions. *Soviet Journal of Nuclear Physics*, 24: 449–454

Derrick, G.H. 1964. Comments on nonlinear wave equations as models for elementary particles. *Journal of Mathematical Physics*, 5: 1252–1254

Garber, W.-D., Ruijsenaars, S.N.M., Seiler, E. & Burns, D. 1979. On finite-action solutions of the nonlinear σ-model. *Annals of Physics*, 119: 305–325

Hobart, R.H. 1963. On the instability of a class of unitary field models. *Proceedings of the Physical Society, London*, 82: 201–203

Prasad, M.K. & Sommerfield, C.M. 1975. Exact classical solution for the 't Hooft monopole and the Julia-Zee dyon. *Physical Review Letters*, 35: 760–762

DETAILED BALANCE

This entry provides a qualitative discussion of equilibrium, a more quantitative discourse of principles such as detailed balance (which are needed in the description of equilibrium phenomenon), and a brief presentation of the Einstein relation between mobility and diffusion, which can be related to the above topics.

The Problem of Time

One often says that a system has reached an equilibrium state if its physical variables are constant in time. Because of fluctuations that cannot be removed, however, it is better to regard the system as in equilibrium when there are no systematic trends in the time averages of its physical parameters. Here, averages are considered over all the microscopic constituents of the system, whether they are elementary particles, atoms, molecules, or larger objects. Equilibrium can be established among these constituents.

Thus, a system that is in equilibrium cannot reveal the time variable among its broad characteristics. In other words, there is no way of telling which way time is running if one's observations are confined to an equilibrium system. Formulated as a philosophical puzzle about the nature of time, this subject has spawned a library of books and papers, with little agreement among the authors (see Landsberg, 1982; Smith, 1993; Price, 1996; Davies, 1995).

Some Relevant Principles of Statistical Mechanics

Here and below, we shall deal with a number of important principles that may or may not hold in any given case and are related to each other. To make these matters quantitative, denote by P_i the probability of finding a system of interest in any one of the ith group of states, G_i in number. The probability per unit time that a transition occurs from a state of group i to a state of group j is denoted by A_{ij}. The transition rate $i \to j$ can be written as

$$R_{ij} = P_i A_{ij} G_j. \qquad (1)$$

If there are W available groups of states, the time rate of change of P_i is

$$\dot{P}_i = \sum_{l=1}^{W} (R_{li} - R_{il}) \quad (i = 1, 2 \ldots W). \qquad (2)$$

To be tractable the A_{ij} have to be independent of time. The first sum gives the transitions into states i and the second sum gives the transitions out of the states i. To simplify the picture one can replace a typical group of states i by a single state, i.e., one can put $G_i = 1$.

Now some additional general principles can be defined. The existence of the A_{ij} can be deduced from quantum mechanical perturbation theory, but it is then valid only for a restricted time interval. One often finds the symmetry relation

$$A_{ij} = A_{ji} \quad (\text{all } i, j) \qquad (3)$$

as a result of the Hermitian character of the perturbation operator. In statistical mechanics, one also uses the principle of Equation (3). It can then be independent of perturbation theory and is regarded instead as resulting from adequate statistical assumptions. It is then called the *principle of microscopic reversibility*.

Next we have the principle of detailed balance which asserts that at a certain time t the forward and reverse transition rates between two groups of states are equal at a certain time; thus,

$$R_{ij} = R_{ji} \quad (\text{all } i, j). \qquad (4)$$

If Equation (4) holds, one sees that $\dot{P}_i$ vanishes for all i. In fact, we can define a steady state by

$$\dot{P}_i = 0 \quad (\text{all } i). \qquad (5)$$

Such a state need not be an equilibrium state since the system may, for example, be continuously raised to a high energy state by some external influence and then drop back continuously, for example, by the emission of radiation. Thus, one sees that Equation (4) implies Equation (5), but not conversely. For more details, see Lifschitz & Pitaewski (1981) and Landsberg (1991).

A Simple Example from the Solid State

One can use detailed balance arguments to infer the form of an unknown emission rate from a known absorption rate, as will now be shown by an example (Landsberg, 1991, p. 391). The idea is to obtain an expression for the equilibrium absorption rate per unit volume of photons of frequency ν_0 in a semiconductor of refractive index μ and, hence, to infer spontaneous emission rates per unit volume.

The probability of a single photon of vacuum wavelength λ_0 being absorbed per unit time per unit volume is

$$P(\lambda_0) = c\alpha(\lambda_0)/V\mu(\lambda_0). \qquad (6)$$

The dimensions $(LT^{-1} \cdot L^{-1} \cdot L^{-3})$ are easily verified to be correct. To find the volume rate of excitation in the solid by photons in the vacuum wavelength range $\mathrm{d}\lambda_0$, $P(\lambda_0)$ has to be multiplied by the number of relevant photon modes $(8\pi\mu^3\lambda_0^{-4}V\,\mathrm{d}\lambda_0)$, and also by their equilibrium occupation probability at temperature T:

$$1/[\exp(ch/\lambda_0 kT) - 1]. \qquad (7)$$

But not all photons of wavelength λ_0 will, when absorbed, produce one electron-hole pair. We denote by $\alpha'(\lambda_0)/\alpha(\lambda_0)(\leq 1)$ the probability of this happening per absorbed photon. Hence, the equilibrium absorption rate (per unit volume) of photons in the wavelength range $\mathrm{d}\lambda_0$ with production of an electron-hole pair is

$$\alpha'/\alpha \frac{8\pi\mu^2\alpha c\lambda^{-4}}{\exp(ch/\lambda_0 kT) - 1} d\lambda_0$$

or

$$\frac{8\pi\alpha'}{h^3}\frac{\mu}{c^2}(kT)^3\frac{x^2\mathrm{d}x}{\exp x - 1}. \qquad (8)$$

Here $x = h\nu_0/kT$, and the second of these expressions is like the first, except that it is in terms of frequencies. According to detailed balance, the new inference is that these expressions give the rate per unit volume of spontaneous radiated recombination of electron-hole pairs with the emission of photons in the range $\mathrm{d}\lambda_0$ or $\mathrm{d}\nu_0$. Note that we have passed from absorption to emission data. This widely used result was first given by W. van Roosbroeck and W. Shockley in 1954. For other examples of the use of the principle of detailed balance in solid state physics, see Landsberg (1991).

The Einstein Relation

The Einstein Relation is basic to solid states physics and rests on the assumption that in a steady state the flux of charged particles due to an electric field must be balanced by diffusion of these particles induced by their density gradients. These two effects are due to well-known and simple forces. The first is a particle flux due to diffusion (with diffusion coefficient D, say). It can be written $-D\mathrm{d}n/\mathrm{d}x$ for one-dimensional motion, where n is the density of particles and $\mathrm{d}n/\mathrm{d}x$ the gradient ("grad n" in three dimensions). The minus sign shows that the force acts to the left if the concentration n increases to the right. The second force on the charged particles is due to a built-in or externally applied electric field E, which is a vector in three dimensions. Here we deal merely with the one-dimensional problem, and note that E can be replaced by $-\mathrm{d}V/\mathrm{d}x$, where V is the electrostatic potential at the point considered. The flux of particles can be written as $n\nu E = -n\mu \mathrm{d}\, V/\mathrm{d}x$, where ν is the so-called mobility of the particles.

In order to obtain the Einstein relation in its simplest form, one has to equate the two forces

$$-n\mu\frac{\mathrm{d}V}{\mathrm{d}x} = D\frac{\mathrm{d}n}{\mathrm{d}x}, \tag{9}$$

which implies that

$$\frac{\mathrm{d}(\ln n)}{\mathrm{d}x} = -\frac{\mu}{D}\frac{\mathrm{d}V}{\mathrm{d}x} \tag{10}$$

giving the simple result

$$n = n_0 \exp(-\mu V/D). \tag{11}$$

As we also know that the stationary state in an electric field at a temperature T is governed by the Boltzmann distribution

$$n = n_0 \exp(-eV/kT), \tag{12}$$

where n_0 is a constant and k is Boltzmann's constant. Comparison yields the Einstein relation

$$\mu = eD/kT. \tag{13}$$

This result connects the mobility of charged particles in a field with their diffusion coefficient. At first sight this seems unexpected because one side of the equation deals with the mechanical characteristic of diffusion.

The extension to three dimensions is not the only generalization that can be made. For example, a similar Einstein relation holds for thermal current density, and generalizations have also been made for large departures from equilibrium (Landsberg, 1991). A further variety of special cases arises for different assumptions about the shape of the semiconductor bands that can occur; for example, they can be degenerate or nondegenerate, parabolic or nonparabolic, etc., and the results can be given in a table of formulae. (Einstein's paper was published in *Annalen der Physik und Chemie* in 1905, the first of three important papers published by him in that year.)

The principle of detailed balance emerged somewhat hesitantly in the 1920s, based on Einstein's 1917 paper on transition possibilities. It was named by Fowler and Milne following other authors and other names. A brief historical survey is given by ter Haar (1955).

Peter Landsberg

See also **Diffusion; Stochastic processes**

Further Reading

Coveney, P. & Highfield, R. 1991. *The Arrow of Time*, London: Allen, 1990 and New York: Fawcett Columbine

Davies, P. 1995. *About Time: Einstein's Unfinished Revolution*, New York: Simon and Schuster

Einstein, A. 1905. Die von der molekularkinetischen Theorie der Wärme geforderte Bewegung. *Annalen der Physik und Chemie*, 17: 549–560

Landsberg, P.T. (editor). 1982. *The Enigma of Time*, Bristol: Adam Hilger

Landsberg, P.T. 1991. *Thermodynamics and Statistical Mechanics*, Oxford and New York: Oxford University Press

Landsberg, P.T. 1991. *Recombination in Semiconductors*, Cambridge and New York: Cambridge University Press

Lifschitz, E.M. & Pitaewski, L.P. 1981. *Physical Kinetics*, Oxford and New York: Pergamon Press

Price, H. 1996. *Time's Arrow and Archimedes' Point*, Oxford and New York: Oxford University Press

Smith, Q. 1993. *Language and Time*, Oxford and New York: Oxford University Press

ter Haar, D. 1955. Foundations of statistical mechanics. *Reviews of Modern Physics*, 27: 289

DETERMINISM

Determinism is a philosophical and scientific notion, and discussions about it are as old as philosophy and science themselves. Richard Taylor writes "Determinism is the general philosophical thesis which states that for everything that ever happens there are conditions such that, given them, nothing else could happen" (Taylor, 1996). This seems to be the most general formulation of determinism. In philosophy, he continues, "There are five theories of determinism to be considered, which can for convenience be called ethical determinism, logical determinism, theological determinism, physical determinism, and psychological determinism." Here we shall confine ourselves only to physical determinism in the natural sciences, except in the concluding section.

In physics, the deterministic view developed along with the experimental approach to research, in the sense that phenomena are reproducible under the same unchanged external conditions, implying that the same cause leads to the same consequences under the same conditions. The quantitative description of physical reality began with Galileo Galilei; although some early developments are due to Pythagoras. However, Isaac Newton was the first to lay down the complete basis of classical mechanics, which at the time was considered to be the origin of all physical phenomena. His laws of mechanics plus the law of gravitation enabled him to reproduce and mathematically derive the motion of the planets, observations of which were empirically well known by the beginning of the 16th century and formulated in Johannes Kepler's laws of celestial mechanics. With the rise and development of classical mechanics the view of determinism developed, with the opinion that all natural laws can be described by dynamical equations, either ordinary differential equations (as, for example, in celestial mechanics) or partial differential equations (as, for example, in the dynamics of fluids). In each case precise knowledge of the initial conditions (all positions and all velocities) completely determines the entire future and entire past of the system. When pushed to its extremum, this view implies complete deterministic evolution of the entire universe, including all its smallest and largest details. The French mathematician Pierre Simon de Laplace, about one century after Newton, wrote (in an often quoted passage):

> We ought then to regard the present state of the universe as the effect of its antecedent state and the cause of the state that is to follow. An intelligence knowing, in any instant of time, all forces acting in nature, as well as the momentary positions of all things of which the universe consists, would be able to comprehend the motions of the largest bodies in the world and those of the smallest atoms in one single formula, provided it were sufficiently powerful to subject all data to analysis: to it, nothing would be uncertain, both future and past would be present before its eyes. (Laplace, 1814)

We can comment on Laplace's statement from our modern perspective. First, to store and process data of *infinite precision* about the state of the entire universe is problematic, as it would require a computer that would be of comparable size and complexity to the entire universe. Thus, its presence has to be taken into account, since—obeying the same mechanical laws as the rest of the universe—it would itself disturb the universe. Therefore, we can conclude that Laplace's "intelligence" (sometimes known as Laplace's daemon) cannot exist, and consequently his idea is fiction. Second, infinite precision of all the initial conditions (positions and momenta) can never be achieved in practice. And when the precision is finite, the existence of chaos (positive Lyapunov exponents) implies sensitive dependence on initial conditions and exponential divergence of nearby trajectories. In other words, there is a finite time horizon exists in general chaotic mechanic systems, beyond which nothing at all can be predicted (Lyapunov time). Therefore, the modern notion of omnipresent chaotic behavior makes Laplace's idea impossible to implement, even in principle. Third, the universe is not described by classical mechanics, but by quantum mechanics, classical mechanics being just a useful or even excellent approximation in observing and describing the motions of sufficiently large bodies. Quantum mechanics tells us, through Heisenberg's principle of uncertainty, that momenta and positions cannot be measured simultaneously with infinite precision, but we have instead the inequality $\Delta x \Delta p_x \geq \hbar/2$, where Δx and Δp_x are the uncertainties of position x and the conjugated momentum p_x. So, Laplace's initial conditions can never be known to arbitrary precision, even in principle.

Quantum mechanics is the correct description of physical reality, with the Schrödinger equation as the starting tool, for nonrelativistic systems. The quantum theory has been further developed by Paul Dirac for relativistic quantum systems and by the quantum field theory up to the unifying field theories, which capture three fundamental interactions (electromagnetic, weak, and strong interactions), but not yet gravity. The Schrödinger equation is a deterministic equation of motion of the wave function ψ, which contains the complete description of the quantum state of a given system. Importantly, ψ itself is a statistical quantity and thus not deterministic: it gives merely probabilities for the given system to be found (by measurement) in a given state.

This is the so-called Copenhagen interpretation of quantum mechanics, initiated by Max Born in 1926 and further developed by Niels Bohr and his colleagues, according to whom there is no determinism in physical reality. This view was strongly opposed by Albert Einstein and colleagues, who accepted the quantum theory as correct but thought that it was an incomplete theory, to be supplemented (through future research) by a more general deterministic theory, uncovering further "hidden variables," which seem to be ignored in present-day quantum mechanics. Many attempts have been made to find such a classical theory of fields to deduce the quantum theory but without success. There are also certain predictions such as Bell's inequalities that are the testing ground of whether quantum theory can in principle be an extended classical deterministic field theory. So far the answer is no, at least for a large class of "local hidden variables theories," and today we do have experimental confirmations where Bell's inequalities are experimentally violated, meaning that the quantum theory and its prediction for the outcome of such experiments is correct. Therefore, the statistical interpretation of quantum mechanics of Bohr's Copenhagen school, together with the strongly counter-intuitive notion of nonlocality, is proven to be correct, and these nondeterministic properties of quantum mechanics are being used in technological applications (such as quantum information theory, quantum teleportation, and quantum computing).

It is, of course, a philosophical shock to learn that the world is not deterministic, but there seems to be no way out. One of the main causes is the process of quantum measurement, which as a process is *not* described by the Schrödinger equation and seems to be the primary source of quantum indeterminism. Quantum measurement is the main source for the generally accepted statistical interpretation of quantum mechanics. Still, the potential of a classical nonlinear field theory (including its turbulent solutions) seems largely unexplored as a description of physical reality. Even classical nonlinear dynamics is not deterministic (even in principle) due to the existence of chaos. A nonlinear classical field theory is even richer, for example the complex turbulent solutions of the Navier–Stokes equations.

In a deterministic world, there would be no place for free will in the lives of human beings or other living creatures. Everything would be predetermined by the initial state before our life, even if we do not have information about that, which implies that we cannot be aware of our predestination. Since the world is not deterministic, there is room for free will and free choice. However, it might be that the world is deterministic, if we do not observe it, and is not deterministic as soon as we "touch" it. Therefore, determinism can never be proved (in analogy with Kurt Gödel's famous incompleteness theorem). Thus, our free will may materialize as soon as we interact with the world, otherwise we would be completely predestined, but isolated from the rest of the world, which is of course not possible. The issue of classical and quantum measurement lies at the bottom of such discussions. It leads to the general conclusion that the world ultimately is not deterministic, but determinism might be a good approximation under certain conditions imposed on the measurement process.

MARKO ROBNIK

See also **Butterfly effect; Causality; Chaotic dynamics; Lyapunov exponents; Quantum theory; Turbulence**

Further Reading

Belavkin, V.P. 2002. Quantum causality, stochastics, trajectories and information. *Reports on Progress in Physics*, 65: 353–420

Bell, J.S. 1987. *Speakable and Unspeakable in Quantum Mechanics*, Cambridge and New York: Cambridge University Press

Edward U. Condon, author. Mechanics, Quantum. 1980. *The New Encyclopaedia Britannica: Macropaedia*, 15th edition, vol. 11, Chicago: Encyclopaedia Britannica: 793

Laplace, P.S. 1814. *Essai philosophique sur le probabilités*, Paris: Courier, 1814; as *Philosophical Essay on Probabilities*, Berlin and New York: Springer, 1995

Peres, A. 1995. *Quantum Theory: Concepts and Methods*, Dordrecht: Kluwer

Philip W. Goetz. (editor). 1980. Determinism. *The New Encyclopaedia Britannica: Micropaedia*, 15th edition, vol. III, Chicago: Encyclopaedia Britannica: 494

Taylor, R. 1996. Determinism. In *The Encyclopedia of Philosophy*, vol. 2, edited by Paul Edwards, New York: Macmillan and London: Simon & Schuster, 359

Wheeler, J.A. & Zurek, W.H. (editors). 1983. *Quantum Theory and Measurement*, Princeton, NJ: Princeton University Press

DETERMINISTIC WALKS IN RANDOM ENVIRONMENTS

A "deterministic walk in a random environment" (DWRE) is the name given to a system generated by the motion of some object (such as, a particle, signal, wave, ant, read/write head of the Turing machine) on a graph. At each time step, the object hops from a vertex to one of its neighboring vertices. The choice of neighbor is completely determined by the type of deterministic scattering rule or scatterer, located at the vertex. A random environment is formed by the scatterers that are assumed to be initially randomly (usually independently) distributed among the vertices. DWREs (in their simplest form and under different names) were introduced in various branches of science (Gunn & Ortuño, 1985; Langton, 1986; Ruijgrok & Cohen, 1988) as paradigms, for example, for propagation of a signal in a random media, evolutionary dynamics, growth processes, and the computational environment.

In the early numerical studies, graphs were regular lattices and usually two types of scatterers were considered in each model. The most studied case was that of the regular quadratic lattice with left and right rotators, which rotate the particle to the left or to the right by an angle $\pi/2$, or left and right mirrors aligned along the two diagonals of the lattice. Two classes of such models have been extensively studied numerically (Cohen, 1992). The first class corresponds to the case when there is no feedback of the moving particle to the environment; that is, a particular type of scatterer is fixed at each site of the lattice forever. Another class is formed by models with flipping scatterers, when a scatterer at a site changes (deterministically) after every visit of a particle to this site.

In statistical physics, these models naturally appear as deterministic Lorentz lattice gases (but with a random distribution of scatterers). The scatterers are not spheres (disks) as in the classical Lorentz gas. Instead, say in the d-dimensional cubic lattice, there are $(2d)^{2d}$ different types of scatterers because each vertex in this case has $2d$ incoming and $2d$ outcoming edges. In theoretical computer science, these models are referred to as many-dimensional Turing machines because the changes of scatterer type at each vertex occur deterministically—according to some program written on an infinite tape divided into commands, for example, to change a given scatterer to some other type (Bunimovich & Khlabystova, 2002a).

Although a DWRE reminds one of random walks, these systems are essentially different. The major difference with random walks is that instead of carrying out a random experiment (like flipping a coin), the particle chooses each step deterministically. Formally, DWREs are deterministic cellular automata, but their behavior reflects a mixture of deterministic dynamics and a random environment. Their dynamics is often counterintuitive (Cohen, 1992) because one's intuition is essentially based on exactly understood (completely solved) systems and models. There are many such models among purely deterministic and purely stochastic systems; however, there were basically no completely understood systems with a mixture of deterministic and stochastic features. Some subclasses of DWREs provide such exactly solvable models (Bunimovich, 2000).

Although closest to stochastic systems, DWREs have fixed environments. This seems counterintuitive, but the evolution of scattering types makes the entire dynamics more deterministic than in the case where an (initially random) distribution of scatterers is frozen.

In many cases, DWRE systems are equivalent to various models from percolation theory (Bunimovich & Troubetzkoy, 1992). Not only the structure of the graph (lattice) but also the types of scatterer in the model determine the corresponding percolation problem. For instance, the mirror's model in the square lattice is reduced to the percolation problem on the square lattice, while the rotator's model is reduced to the percolation problem on some nonplanar graph (Bunimovich & Troubetzkoy, 1992).

Perhaps the most widely known DWRE models are Langton's Ant (Langton, 1986) or the flipping rotators model on the square lattice (Ruijgrok & Cohen, 1988), which are solvable again with rather counterintuitive results (Bunimovich & Troubetzkoy, 1993). If all vertices are occupied with rotators, then all orbits (particle's paths) are unbounded. If, on the other hand, one allows vertices to be empty with positive probability (i.e., the third, straight-ahead scatterer is allowed), then the particle's path becomes bounded with probability one.

The results, both numerical and mathematical, continued to surprise until "Walks in Rigid Environments" (WRE) were introduced and analyzed (Bunimovich, 2000). WREs employ a new integer parameter r, $1 \leq r \leq \infty$, which is called the rigidity of the environment. The rigidity determines how many times the particle must collide with the given scatterer in order to change its type. In other words, the scatterer at a given vertex changes its type immediately after the rth visit of the particle to this site. Therefore, WREs interpolate between DWREs with fixed environments (where $r = \infty$) and DWREs with flipping environments (where $r = 1$).

WREs on a one-dimensional lattice $\mathbb{Z}$ are completely solved (Bunimovich, 2000; Bunimovich & Khlabystova, 2002b). In this case, there are only four types of scatterers. Two of them (forward scatterer and backscatterer) are symmetric with respect to the reflection of $\mathbb{Z}$, which is the only nontrivial symmetry of the one-dimensional lattice. The other two scatterers, which always send the particle to the right (or to the left), do not respect this symmetry. Therefore, the WRE with the last two types of scatterers has the same behavior for all values of the rigidity r. This model demonstrates a diffusive type of behavior, in which the particle eventually visits all vertices and the mean square displacement of the particle is proportional to t. On the contrary, WREs with forward and back scatterers demonstrate totally different behavior depending on the parity of the rigidity r. For even rigidities, the particle eventually visits all vertices again but its motion is subdiffusive. The most interesting behavior occurs for odd values of the rigidity. In this case the particle—after a short initial period of seemingly irregular motion near the origin—starts to propagate in one direction with random velocity. This phenomenon of (eventual) propagation reminds one of "gliders" in Conway's Game of Life. However, in a WRE this propagation occurs for all initial configurations of environment, while in the Game of Life, gliders appear as only very special solutions. The phenomenon of eventual propagation in one direction is not restricted to one-dimensional WREs. For instance, the same behavior is demonstrated by the model with right and left rotators on the triangular lattice (Grosfils et al., 1999).

If the rigidity $r < \infty$, then one can also investigate the dynamics of the environment. Observe that if $r < \infty$, then it makes sense to consider not one but many moving particles as well. Indeed, even though the particles do not interact directly, they do effectively interact by changing environments with each other. The evolution of the environment in such models can be chaotic having a positive and even infinite metric entropy (Bunimovich & Troubetzkoy, 1993). This means, in particular, that the moving particle sees itself surrounded by almost any possible environment of scatterers at different moments of time.

An important difference between DWREs and random walks is that in the continuous limit DWREs become completely deterministic (Bunimovich & Khlabystova, 2002b), whereas biased random walks under proper scaling of probabilities become in the continuous limit stochastic diffusion processes.

LEONID BUNIMOVICH

See also **Game of life; Lorentz gas; Random walks**

Further Reading

Bunimovich, L.A. 2000. Walks in rigid environments. *Physica* A, 279:169–179

Bunimovich, L.A. & Khlabystova, M.A. 2002a. Lorentz lattice gases and many-dimensional Turing machines. In *Collision-Based Computing*, edited by A. Adamatzky, London: Springer

Bunimovich, L. A. & Khlabystova, M.A. 2002b. Walks in rigid environments: continuous limits. *Journal of Statistical Physics*, 108: 905–925

Bunimovich, L. A. & Troubetzkoy, S.E. 1992. Recurrence properties of the Lorentz lattice gases. *Journal of Statistical Physics*, 67: 289–302

Bunimovich, L. A. & Troubetzkoy, S.E. 1993. Topological properties of flipping Lorentz lattice gas models. *Journal of Statistical Physics*, 72: 297–307

Cohen, E.G.D. 1992. New types of diffusion in lattice gas cellular automata. In *Microscopic Simulations of Complex Hydrodynamic Phenomena*, edited by M. Mareschal & B.L. Holian, New York: Plenum

Grosfils, P., Boon, J.P., Cohen, E.G.D. & Bunimovich, L.A. 1999. Propagation and self-organization in lattice random media. *Journal of Statistical Physics*, 97: 575–608

Gunn, J.M.F. & Ortuño, M. 1985. Percolation and motion in a simple random environment. *Journal of Physics* A, 18: 1095–1099

Langton, C.G. 1986. Studying artificial life with cellular automata. *Physica* D, 22: 120–149

Ruijgrok, T.W. & Cohen, E.G.D. 1988. Deterministic lattice gas models. *Physics Letters* A, 133: 415–419

DEVELOPMENT OF SINGULARITIES

For some nonlinear PDEs, solutions of the Cauchy (initial-value) problem may exist only until a finite moment $t = t_c$, and they cannot be continued afterwards. At $t = t_c$, the solution loses its initial smoothness, and a singularity appears, which causes the blow-up of the solution. For conservative systems where the wavefield develops sharply diverging gradients, this blow-up phenomenon is often called collapse.

Historically, the first example of collapse follows from the so-called Hopf equation: $u_t + uu_x = 0$, which describes the one-dimensional velocity of dust with zero pressure. This equation admits the generic implicit solution $u = F(x - ut)$, where F is determined from the initial velocity profile. For any compact distribution F, there exists a couple of coordinates (x_c, t_c), at which the spatial derivative of u reaches infinity.

Collapse also occurs in the high-dimensional solutions to the nonlinear Schrödinger (NLS) equation, consisting of a point-singularity that achieves the fate of nonlinear waves undergoing self-focusing (Kelley, 1965). To describe this process, let us consider the NLS equation

$$\mathrm{i}\partial_t\psi + \nabla^2\psi + |\psi|^2\psi = 0, \tag{1}$$

where t is a time variable and the Laplacian $\nabla^2 = \partial_x^2 + \partial_y^2 + \partial_z^2 + \cdots$ accounts for the dispersion of a wave-packet along D orthogonal spatial axes $[\boldsymbol{r} = (x, y, z, ...)]$. The wave function ψ evolves from the spatially localized initial datum $\psi(\boldsymbol{r}, 0) \equiv \psi_0(\boldsymbol{r})$, assumed to belong to the Hilbert space H^1 with finite norm $\|\psi\|_{H^1} = (\|\psi\|_2^2 + \|\nabla\psi\|_2^2)^{1/2}$, where $\|f\|_p \equiv (\int |f|^p \mathrm{d}\boldsymbol{r})^{1/p}$. Two invariants are associated with ψ, namely, the L^2 norm N (sometimes called mass, power, or number of particles) and the Hamiltonian H:

$$N = \|\psi\|_2^2, \quad H = \|\nabla\psi\|_2^2 - \tfrac{1}{2}\|\psi\|_4^4. \tag{2}$$

From Equation (1), the following virial equality can be established (Glassey, 1977):

$$N d_t^2\langle r^2\rangle = 4\{2H + (1 - D/2)\|\psi\|_4^4\}, \tag{3}$$

where $\langle r^2\rangle = \int r^2|\psi|^2\mathrm{d}\boldsymbol{r}/N$ denotes the mean-squared radius of the solution ψ. By a double integration in time, Equation (3) shows that, whenever $D \geq 2$, there exist initial conditions for which $\langle r^2\rangle$ vanishes at finite time, which is the signature of a wave collapse. For finite norms N, the inequality $N \leq (2/D)^2\langle r^2\rangle \times \|\nabla\psi\|_2^2$ implies that the gradient norm diverges in collapse regimes. As H is finite, the collapse dynamics makes the L^4 norm $\|\psi\|_4^4$ blow up in turn and $\max|\psi|$ diverges, by virtue of the mean-value theorem $\int |\psi|^4\mathrm{d}\boldsymbol{r} \leq \max_r|\psi|^2 \times N$. This leads to a finite-time blow-up, at which the solution ψ ceases to exist in H^1. Blow-up generally occurs before $\langle r^2\rangle$ reaches zero (Rasmussen & Rypdal, 1986).

While $H < 0$ arises from Equation (3) as a sufficient condition for collapse, sharper requirements can be derived by means of the so-called Sobolev inequality

$$\|\psi\|_4^4 \leq C\|\nabla\psi\|_2^D \times \|\psi\|_2^{4-D}. \tag{4}$$

- In the critical case $D = 2$, this inequality can be used to bound H from below, so that the gradient norm blows up only if N fulfills the constraint $N > N_c$. The best constant in Equation (4) is exactly $C_{\text{best}} = 2/N_c$, and it involves the quantity $N_c = \int \phi_0^2\,\mathrm{d}\boldsymbol{r} = 11.68$, where ϕ_0 is the radially symmetric soliton solution of $-\phi_0 + \nabla^2\phi_0 + \phi_0^3 = 0$ (Weinstein, 1983). N_c provides the minimum power that ψ must necessarily contain at $t = 0$ for producing a collapse, justifying the existence of a critical power for the 2-dimensional self-focusing of optical beams in nonlinear Kerr media.
- In the supercritical case $D = 3$, a criterion for collapse, sharper than $H < 0$, can be established from a combination of Equations (3) and (4) as $H < N_c^2/N$ for gradient norms initially above $3N_c^2/N$ (Kuznetsov et al., 1995). N_c again corresponds to the mass of the three-dimensional soliton solution ϕ_0 satisfying $-\phi_0 + \nabla^2\phi_0 + \phi_0^3 = 0$.

Once the collapse is triggered, the solution self-focuses and shrinks isotropically in a self-similar way near the singularity point t_c. It is, thus, convenient to introduce the self-similar substitution:

$$\psi(\boldsymbol{r}, t) = a^{-\alpha}(t)\phi(\boldsymbol{\xi}, \tau)\mathrm{e}^{\mathrm{i}\lambda\tau - \mathrm{i}\beta\xi^2/4}, \tag{5}$$

where $\boldsymbol{\xi} = \boldsymbol{r}/a(t)$, $\tau(t) \equiv \int_0^t \mathrm{d}u/a^2(u)$, and $\beta = -a\dot{a}$ (dot indicates differentiation with respect to time). Here, the parameter λ is positive for localization of ϕ. The function $a(t)$ represents the scale length that vanishes as collapse develops, and ϕ converges to an exactly self-similar form $\phi(\boldsymbol{\xi})$, which no longer depends explicitly on time $[\partial_\tau \phi \to 0]$. Inserting (3) into H shows that the right balance between the gradient and L^4 norms requires $\alpha = 1$, in order to preserve the finiteness of the Hamiltonian. With $\alpha = 1$, $N = a^{D-2}(t) \int |\phi|^2 \mathrm{d}\boldsymbol{\xi}$ and for radial solutions, Equation (1) transforms to

$$\mathrm{i}\partial_\tau \phi + \xi^{1-D}\partial_\xi \xi^{D-1}\partial_\xi \phi + |\phi|^2\phi + \varepsilon[\xi^2 - \xi_\mathrm{T}^2]\phi = 0, \tag{6}$$

where $\xi_\mathrm{T}^2 \equiv \varepsilon^{-1}[\lambda + \mathrm{i}\beta(D/2 - 1)]$ is viewed as a turning point, with $\varepsilon \equiv \frac{1}{4}(\beta^2 + \partial_\tau \beta)$. As $a(t) \to 0$, ϕ can be treated by means of quasi-self-similar techniques (Bergé, 1998). In the limit $\partial_\tau \phi \to 0$, ε converges to $\beta^2/4$ and the solution ϕ is split into a nonlinear core (ϕ_c) extending in the range $\xi < \xi_\mathrm{T}$ and a linear tail (ϕ_T) defined in the complementary spatial domain $\xi > \xi_\mathrm{T}$, where the nonlinearity vanishes. As a result, the wave function ψ reads near the collapse point:

$$\begin{aligned}\psi(\boldsymbol{r}, t) = {} & \frac{\mathrm{e}^{\mathrm{i}\lambda \int_0^t \mathrm{d}u/a^2(u)}}{a(t)} \\ & \times \left\{ \phi_\mathrm{c}\left(\frac{r}{a}, \varepsilon\right) \mathrm{e}^{-\mathrm{i}\beta r^2/4a^2} \Big|_{0 \le r < r_\mathrm{T}} \right. \\ & \left. + \frac{C(\beta)}{(r/a)^{1+\mathrm{i}\lambda/\beta}} \Big|_{r_T < r < r_\mathrm{max}} \right\},\end{aligned} \tag{7}$$

where $|C(\beta)|^2$ evolves like $(2/\beta|\xi_\mathrm{T}|^{D-2})\mathrm{e}^{-\lambda\pi/\beta}$, $r_\mathrm{T} = a(t)|\xi_\mathrm{T}|$ and r_max bounds the self-similarity domain. The length $a(t)$ is identified from the continuity equation describing the mass exchanges between the core and tail parts of ψ. The dynamics of self-similar collapses then vary with the space dimension number as follows.

- *Strong collapse*: For $D = 2$, the size $a(t)$ behaves with a double-logarithmic correction: $a(t) \simeq a_0 \sqrt{t_\mathrm{c} - t}/\sqrt{\ln\ln[1/(t_\mathrm{c} - t)]}$, coming from $\beta \simeq \lambda\pi/\ln[\tau(t)]$ and $\tau(t) \simeq \ln[1/(t_\mathrm{c} - t)]$. As $t \to t_\mathrm{c}$, the exponential contribution in tail (5) thus decreases to zero, while the core boundary ξ_T increases slowly to infinity. Collapse thus takes place with a core solution providing the principal contribution in the wavefunction ψ. N is mainly given by $\int |\phi_\mathrm{c}|^2 \mathrm{d}\boldsymbol{\xi}$, which relaxes to the critical value $N_\mathrm{c} = 11.68$. The mass stays mostly located around the center, which meets the definition of *strong collapse* (Zakharov & Kuznetsov, 1986).
- *Weak collapse*: For $D = 3$, β attains a fixed point $\beta_0 \neq 0$, leading to the scaling law $a(t) \simeq a_0(t_\mathrm{c} - t)^{1/2}$. The power is no longer preserved self-similarly in space, since $N = a(t) \int |\phi|^2 \mathrm{d}\boldsymbol{\xi}$. This integral behaves as $N \simeq N_\mathrm{core}(t) + N_\mathrm{tail}(t)$, where $N_\mathrm{core}(t) \sim a(t)$ vanishes, while $N_\mathrm{tail}(t) \simeq 4\pi|C(\beta)|^2 r_\mathrm{max}$ contains almost all the initial mass as $a(t) \to 0$. A 3-d collapse is thus accompanied by an expulsion of particles towards the large-distance domain $r \gg r_\mathrm{T}(t)$, which characterizes a *weak collapse*. The solution ψ blows up at the center, where N_core becomes zero. Accordingly, ψ extends in the outer domain with the stationary density $r^2|\psi|^2 \to |C(\beta_0)|^2$ = const.

These different collapses can be tested numerically by adding a small nonlinear dissipation term in Equation (1), i.e., by changing ψ_t into $\psi_t + \beta|\psi|^m\psi$ with $\beta \ll 1$ and $m > 1$. For $D = 2$, dissipation removes a substantial amount of energy per collapse event, because N remains mostly confined at the center. This leads to a step-wise decrease of N (Kosmatov et al., 1991). In contrast, for $D = 3$, nonlinear dissipation cannot remove much energy, the major part of $N \simeq N_\mathrm{tail}$ being transferred to large distances. This causes a monotonical decrease of N.

- *Superstrong collapse*: For weak and strong collapses, only the energy captured by the singularity dissipates. A different situation is realized if collapsing solutions can sustain dissipation through power loss into a persistent region of high intensity. Such solutions approach a quasi stationary state, described by $\psi_{rr} + (D - 1)\psi_r/r + |\psi|^2\psi = 0$, in the limit of small $\beta \to 0$. For instance, when $D = 4$, Equation (1) admits the stationary radial state $\psi(r) = B\mathrm{e}^{\mathrm{i}\ln r\sqrt{B^2-1}}/r$ $(B > 1)$, for which the power density $|\psi|^2$ flows into the singularity with a constant energy flux equal to $B^2\sqrt{B^2 - 1}$. This collapse, which received the adjective of *superstrong*, has been numerically detected at high dimension numbers $D > 3$ (Kosmatov et al., 1991).

So far, the discussion has remained within the realm of the one-wave component NLS equation with a cubic nonlinearity. It is thus worth underlining the following.

- The previous results can be generalized to a power-law nonlinearity, when the cubic term $|\psi|^2\psi$ of Equation (1) is replaced by $|\psi|^{2n}\psi$ with $n > 1$ (Rasmussen & Rypdal, 1986; Bergé, 1998). Solutions with $D = 2/n$ follow the route of a strong collapse, while solutions defined for $D > 2/n$ collapse weakly. Superstrong collapses apply to the dimensional configurations $D > 2 + 1/n$.
- Several NLS equations coupled through their cubic nonlinearities often serve to model the self- and cross-interactions of multiple wave components (or different polarizations) in vector systems. Such systems promote blow-up phenomena, which can be examined by means of the above analytical tools (Bergé, 2001).
- Blow-up may take place in solutions of PDEs other than the NLS equation. For example, investigations of the solutions to the generalized D-dimensional Korteweg–de Vries (KdV) equation

$$q_t + q^n q_x + (\nabla^2 q)_x = 0 \tag{8}$$

suggest that, whereas no collapse occurs for values of the product $nD < 4$, collapsing states can arise and adopt a self-similar shape provided that $nD \geq 4$ (Blaha et al., 1989). The mathematical proof for this statement is presently incomplete.

LUC BERGÉ

See also **Filamentation; Kerr effect; Nonlinear Schrödinger equations; Virial theorem**

Further Reading

Bergé, L. 1998. Wave collapse in physics: Principles and applications to light and plasma waves. *Physics Reports*, 303: 259–370

Bergé, L. 2001. Nonlinear wave collapse. In *Spatial Solitons*, edited by S. Trillo & W. Torruellas, Berlin: Springer, pp. 247–267

Blaha, R., Laedke, E.W. & Spatschek, K.H. 1989. Collapsing states of generalized Korteweg–de Vries equations. *Physica* D, 40: 249–264

Glassey, R.T. 1977. On the blowing-up of solutions to the Cauchy problem for nonlinear Schrödinger equations. *Journal of Mathematical Physics*, 18: 1794–1797

Kelley, P.L. 1965. Self-focusing of optical beams. *Physical Review Letters*, 15: 1005–1008

Kosmatov, N.E., Shvets, V.F. & Zakharov, V.E. 1991. Computer simulation of wave collapses in the nonlinear Schrödinger equation. *Physica* D, 52: 16–35

Kuznetsov, E.A., Rasmussen, J., Rypdal, K. & Turitsyn, S.K. 1995. Sharper criteria for the wave collapse. *Physica* D, 87: 273–284

Rasmussen, J. & Rypdal, K. 1986. Blow-up in nonlinear Schrödinger equations-I: a general review. *Physica Scripta*, 33: 481–497

Weinstein, M.I. 1983. Nonlinear Schrödinger equations and sharp interpolation estimates. *Communications in Mathematical Physics*, 87: 567–576

Zakharov, V.E. & Kuznetsov, E.A. 1986. Quasiclassical theory of three-dimensional wave collapse. *Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki* (USSR JETP), 91: 1310–1324 [Trans. in *Soviet Physics JETP*, 64: 773–780]

DEVIL'S STAIRCASE

See **Fractals**

DIFFEOMORPHISM

See **Maps**

DIFFERENTIAL GEOMETRY

A *topological manifold of dimension n* is a topological space M that can locally be identified with an open set in $\mathbb{R}^n$; a superscript on the symbol for M is often used to indicate the dimension. For example, a circle is a one-dimensional manifold, denoted S^1, while a figure-eight is not a manifold. In more detail, M must have a set of *coordinate charts* ϕ_α, each of which is a homeomorphism from an open set $U_\alpha \subset M$ to an open set in $\mathbb{R}^n$, such that every point of M is in the domain of a chart. If the intersection $U_{\alpha\beta}$ of two domains U_α and U_β is nonempty, then the change of coordinates $\phi_\alpha \circ \phi_\beta^{-1}$ is a continuous map from $\phi_\beta(U_{\alpha\beta})$ to $\phi_\alpha(U_{\alpha\beta})$ with a continuous inverse.

A *differentiable manifold* is a topological manifold M^n equipped with charts such that, on the overlaps, $\phi_\alpha \circ \phi_\beta^{-1}$ is differentiable with a differentiable inverse. (Here, "differentiable" or "smooth" functions are those whose partial derivatives exist and are continuous to all orders; however, for C^k *manifolds*, the changes of coordinates are only required to be differentiable up to order k.) If $x^1, \ldots, x^n$ and $\bar{x}^1, \ldots, \bar{x}^n$ are the coordinate functions for two overlapping charts, then the Jacobian determinant $||\partial x^i/\partial \bar{x}^j||$ must be nonzero at each point.

For example, the subset M of $\mathbb{R}^{n+1}$ defined by an equation form $f(x^0, x^1, \ldots, x^n) = 0$ is a differentiable manifold if f is a smooth function and at each point $p \in M$, at least one partial derivative (say, $\partial f/\partial x^0$) is nonzero. Then, by the Implicit Function Theorem, the projection from a neighborhood of p (i.e., an open subset of M containing p) onto the $x^1 \cdots x^n$ coordinate hyperplane is differentiable with a differentiable inverse. So, the n-dimensional spheres S^n are compact differentiable manifolds. Similarly, the group SL_n of $n \times n$ matrices with determinant one is a noncompact manifold of dimension $n^2 - 1$.

A function $f : M \to \mathbb{R}$ is smooth if $f \circ \phi$ is smooth for any chart ϕ. More generally, if M^m and N^n are two differentiable manifolds, we say a mapping $F : M \to N$ is smooth if it is smooth with respect to coordinate charts on both ends, that is, $\psi \circ F \circ \phi^{-1}$ is smooth from $\mathbb{R}^m$ to $\mathbb{R}^n$ for any charts ϕ on M and ψ on N. When $F : M \to N$ also has a smooth inverse, it is a *diffeomorphism*, and M and N are *diffeomorphic*. For example, the hyperboloid $x^2 + y^2 - z^2 = 1$ is diffeomorphic to the cylinder $x^2 + y^2 = 1$, and any open interval on the real line $\mathbb{R}$ is diffeomorphic to all of $\mathbb{R}$.

Vector Fields and 1-Forms

A *tangent vector* at a point $p \in M^n$ is a linear operator v on smooth functions f defined near p, such that (i) $\mathsf{v}(f_1 f_2) = f_1 \mathsf{v}(f_2) + f_2 \mathsf{v}(f_1)$, and (ii) $\mathsf{v}(f_1) = \mathsf{v}(f_2)$ if $f_1 = f_2$ on a neighborhood of p (i.e., f_1, f_2 have the same *germ* at p). For example, if $c : \mathbb{R} \to M$ defines a curve in M such that $c(t) = p$, then we define the tangent vector $c'(t)$ to the curve by $c'(t)(f) := (\mathrm{d}/\mathrm{d}t) f(c(t))$, where the symbol ":=" indicates a definition. If $\mathsf{v} = c'(t)$, then we say that the curve is *tangent to* v at p.

The set of tangent vectors at p form a vector space of dimension n, denoted T_pM. If $x^1, \ldots, x^n$ are local coordinates near p, then the partial derivative operators $\partial/\partial x^i$ are a basis for T_pM. The set of all tangent vectors at all points of M is itself a differentiable manifold of dimension $2n$, since we can adjoin coordinates

$y^1, \ldots, y^n$ and locally write all tangent vectors as

$$\mathsf{v} = \sum_{i=1}^{n} y^i \frac{\partial}{\partial x^i}.$$

This manifold is the *tangent bundle* of M, and is denoted TM.

A *vector field* on M smoothly assigns a tangent vector at each point; in local coordinates, we specify a vector field by giving the y^i as smooth functions of x^i. If f is a smooth function on M and $\boldsymbol{V}$ is a vector field, then $\boldsymbol{V}(f)$ is another smooth function.

Given a vector field $\boldsymbol{V}$ and a point $p \in M$ at which $\boldsymbol{V} \neq 0$, existence theorems for systems of ODE (e.g., Picard's Theorem) imply that there exist a neighborhood U of p and a one-parameter family of smooth, one-to-one mappings $F_t : U \to M$, such that for any fixed $q \in U$, the curve $F_t(q)$ is tangent to $\boldsymbol{V}$ for every t. The F_t's are called the *flow by vector field* $\boldsymbol{V}$.

The *Lie bracket* $[\boldsymbol{V}_1, \boldsymbol{V}_2]$ of two vector fields $\boldsymbol{V}_1, \boldsymbol{V}_2$ is a third vector field defined by

$$[\boldsymbol{V}_1, \boldsymbol{V}_2](f) := \boldsymbol{V}_1(\boldsymbol{V}_2(f)) - \boldsymbol{V}_2(\boldsymbol{V}_1(f)).$$

If y_1^i and y_2^i are the components of $\boldsymbol{V}_1, \boldsymbol{V}_2$ in local coordinates, then

$$[\boldsymbol{V}_1, \boldsymbol{V}_2] = \sum_{i,j=1}^{n} \left(y_1^i \frac{\partial y_2^j}{\partial x^i} - y_2^i \frac{\partial y_1^j}{\partial x^i} \right) \frac{\partial}{\partial x^j}. \qquad (1)$$

The vector fields are said to *commute* if $[\boldsymbol{V}_1, \boldsymbol{V}_2]$ is identically zero; then, flow by $\boldsymbol{V}_1$ commutes with flow by $\boldsymbol{V}_2$.

A *differential 1-form at p* is a linear mapping function from T_pM to $\mathbb{R}$. For example, given a differentiable function f defined near p, we define the 1-form $\mathrm{d}f$ by $\mathrm{d}f(\mathsf{v}) := \mathsf{v}(f)$ for all $\mathsf{v} \in T_pM$. The vector space of 1-forms at p is denoted T_p^*M. Given local coordinates, the differentials $\mathrm{d}x^1, \ldots, \mathrm{d}x^n$, which satisfy $\mathrm{d}x^i(\partial/\partial x^j) = \delta_j^i$, are a basis for T_p^*M. While tangent vectors generalize directional derivatives in $\mathbb{R}^n$, 1-forms generalize gradients; however, 1-forms cannot be identified with tangent vectors without using some nondegenerate bilinear form on tangent vectors, for example, a Riemannian metric or symplectic form on M.

A *differential 1-form* on M assigns a 1-form at each point; in local coordinates, these appear as

$$\omega = \sum_{i=1}^{n} z_i \mathrm{d}x^i,$$

where the z_i are some smooth functions of the x^i. If $\boldsymbol{V}$ is a vector field on M, then $\omega(\boldsymbol{V})$ is a function on M.

Given a smooth mapping $F : M \to N$, we can transform a tangent vector $\mathsf{v} \in T_pM$ to a tangent vector $F_*\mathsf{v}$ at $F(p)$ by defining

$$F_*\mathsf{v}(f) := \mathsf{v}(f \circ F)$$

for any function $f : N \to \mathbb{R}$. This is called the *pushforward* of v. We similarly define the pushforward $F_*\boldsymbol{V}$ of a vector field $\boldsymbol{V}$ on M. We can also transform a 1-form ω on N to a 1-form $F^*\omega$ on M by defining

$$F^*\omega(\mathsf{v}) := \omega(F_*\mathsf{v})$$

for any tangent vector v to M. This is called the *pullback* of ω.

If $\boldsymbol{V}$ is a vector field and F_t is the flow by $\boldsymbol{V}$, we define the *Lie derivative* of a 1-form ω with respect to $\boldsymbol{V}$ by

$$\mathcal{L}_{\boldsymbol{V}}\omega := \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} F_t^*(\omega).$$

If the components of $\boldsymbol{V}$ and ω in local coordinates are y^i and z_i, respectively, then

$$\mathcal{L}_{\boldsymbol{V}}\omega = \sum_{i,j=1}^{n} \left(y^i \frac{\partial z_j}{\partial x^i} - z_i \frac{\partial y^j}{\partial x^i} \right) \mathrm{d}x_j. \qquad (2)$$

Some useful properties of the Lie derivative are that it obeys the product rule and commutes with d, that is,

$$\begin{aligned} \mathcal{L}_{\boldsymbol{V}}(f\boldsymbol{W}) &= (\mathcal{L}_{\boldsymbol{V}} f)\boldsymbol{W} + f\,\mathcal{L}_{\boldsymbol{V}}\boldsymbol{W}, \\ \mathcal{L}_{\boldsymbol{V}}(f\omega) &= f\,\mathcal{L}_{\boldsymbol{V}}\omega + (\mathcal{L}_{\boldsymbol{V}} f)\omega, \\ \mathcal{L}_{\boldsymbol{V}}(\mathrm{d}f) &= \mathrm{d}(\mathcal{L}_{\boldsymbol{V}} f), \end{aligned}$$

where we define $\mathcal{L}_{\boldsymbol{V}} f := \boldsymbol{V}(f)$ and $\mathcal{L}_{\boldsymbol{V}}\boldsymbol{W} := [\boldsymbol{V}, \boldsymbol{W}]$. (Note than one can derive (2) using these properties.)

Higher Degree Differential Forms and Topology

A *differential k-form at p* is a multilinear function on k-tuples of vectors in T_pM that is skew-symmetric; that is, its value is multiplied by -1 whenever two adjacent vectors in the k-tuple are exchanged. For example, a 2-form may be constructed from two 1-forms ω^1, ω^2 using the *wedge product*:

$$\omega^1 \wedge \omega^2(\mathsf{v}_1, \mathsf{v}_2) := \omega^1(\mathsf{v}_1)\omega^2(\mathsf{v}_2) - \omega^1(\mathsf{v}_2)\omega^2(\mathsf{v}_1).$$

(Note that this is zero if ω^1 and ω^2 are linearly dependent.) More generally, the wedge product of 1-forms $\omega^1, \ldots, \omega^k$ is defined by

$$\begin{aligned} &\omega^1 \wedge \ldots \wedge \omega^k(\mathsf{v}_1, \ldots \mathsf{v}_k) \\ &\quad := \sum_{\sigma} (-1)^{\sigma} \omega^1(\mathsf{v}_{\sigma(1)})\omega^2(\mathsf{v}_{\sigma(2)}) \cdots \omega^k(\mathsf{v}_{\sigma(k)}), \end{aligned}$$

where the sum is over all permutations σ of $1, 2, \ldots, k$ and $(-1)^{\sigma}$ is the sign of the permutation. On an n-dimensional manifold, the vector space of k-forms at a point has dimension $\binom{n}{k}$ and is spanned by wedge products of 1-forms; in particular, there are no forms of degree higher than n.

A 2-form may also be constructed from a single 1-form ω by taking the *exterior derivative* $\mathrm{d}\omega$, which satisfies

$$\mathrm{d}\omega(\boldsymbol{V}_1, \boldsymbol{V}_2) = \boldsymbol{V}_1(\omega(\boldsymbol{V}_2)) - \boldsymbol{V}_2(\omega(\boldsymbol{V}_1)) - \omega([\boldsymbol{V}_1, \boldsymbol{V}_2]). \tag{3}$$

(Although the right-hand side is defined using vector fields, its value at a point p depends only on the values of $\boldsymbol{V}_1$, $\boldsymbol{V}_2$ at p.) The exterior derivative of a k-form is a $(k+1)$-form; it can be calculated inductively using linearity and the product rule

$$d(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^{\mathrm{degree}(\alpha)} \alpha \wedge \mathrm{d}\beta.$$

If the exterior derivative of a k-form is identically zero on M, the form is *closed*. If a k-form is an exterior derivative of a $(k-1)$-form, it is *exact*. An important property of the exterior derivative is that $\mathrm{d}(\mathrm{d}\alpha) = 0$, i.e., exact forms are closed. The *Poincaré Lemma* asserts that a closed k-form α is *locally* exact; that is, in the vicinity of any given point a $(k-1)$-form β is defined such that $\alpha = \mathrm{d}\beta$. However, not every closed form is *globally* exact (e.g., the 1-form $\mathrm{d}\theta$ on S^1). Moreover, the *de Rham Theorem* asserts that the dimension of the vector space of closed k-forms modulo exact k-forms is a topological invariant of M; that is, two manifolds cannot be homeomorphic (or diffeomorphic) unless these dimensions, known as the *Betti numbers*, match up. (For example, the Euler characteristic is determined by the Betti numbers.)

THOMAS A. IVEY

See also **Invariant manifolds and sets; Lie algebras and Lie groups; Topology**

Further Reading

Isham, C.J. 1999. *Modern Differential Geometry for Physicists*, Singapore: World Scientific

Ivey, T. & Landsberg, J.M. 2003. *Cartan for Beginners: Differential Geometry via Moving Frames and Exterior Differential Systems*, Providence, RI: American Mathematical Society

Warner, F.W. 1983. *Foundations of Differentiable Manifolds and Lie Groups*, Berlin and New York: Springer

DIFFUSION

When a small drop of ink is poured onto a soft gel, the ink molecules disperse through by diffusion. Similarly, the spread of heat through a medium can also be a diffusive process (called heat conduction). Many other applications of diffusion arise in biology, combustion, economics, chemical engineering, and geophysics, among other fields.

Mathematically, the diffusion equation takes the form (Crank, 1975)

$$u_t = \triangle u, \tag{1}$$

where $u = u(\boldsymbol{x}, t)$ is the state variable, representing, for example, the density of concentration of some substance, at time $t \geq 0$ and position $\boldsymbol{x}$ in R^n, where $\triangle u$ denotes the Laplacian of u with respect to the space variable $\boldsymbol{x}$. Equation (1) is an example of a parabolic equation of evolution. If it holds for all $\boldsymbol{x}$ in R^n, then the problem is fully specified once appropriate initial conditions

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) \tag{2}$$

are known. If Equation (1) holds in a limited domain $\Omega \subset R^n$, then some boundary conditions must be imposed on u at $\partial\Omega$ that are compatible with the physical situation.

The diffusion equation can be viewed as a balance law (Grindrod, 1996). Let $Q(\boldsymbol{x}, t)$ be the net creation rate of particles at $\boldsymbol{x} \in \Gamma \subset \Omega$ and time t, and let $\boldsymbol{J}(\boldsymbol{x}, t)$ be the flux density. For any unit vector $\boldsymbol{n} \in R^n$, the scalar product $\boldsymbol{J} \cdot \boldsymbol{n}$ is the net rate at which particles cross a unit area in a plane perpendicular to $\boldsymbol{n}$ (take the plus sign in the direction of $\boldsymbol{n}$). Assuming that the rate of change of mass in Ω is due to particle creation or degradation inside Γ and the inflow and the outflow of particles through the boundary $\partial\Gamma$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Gamma u\,\mathrm{d}\boldsymbol{x} = -\int_{\partial\Gamma} \boldsymbol{J} \cdot \boldsymbol{n}\,\mathrm{d}S + \int_\Gamma Q\,\mathrm{d}\boldsymbol{x}, \tag{3}$$

where $\int_\Gamma u\,\mathrm{d}\boldsymbol{x}$ denotes the population mass in Γ. If the solution is smooth enough, then applying the divergence theorem to the right-hand side in (3) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Gamma u\,\mathrm{d}\boldsymbol{x} = \int_\Gamma \nabla \cdot \boldsymbol{J}\,\mathrm{d}S + \int_\Gamma Q\,\mathrm{d}\boldsymbol{x}. \tag{4}$$

As Γ was arbitrary in Ω, Equation (4) implies that

$$u_t = -\nabla \cdot \boldsymbol{J} + Q \tag{5}$$

at every point in Ω, which is the required balance law.

In practice, depending on the process studied, one must specify the flux $\boldsymbol{J}$ and the source term Q. For example, in combustion or in chemistry, one uses Fick's law:

$$\boldsymbol{J} = -D\nabla u. \tag{6}$$

Here $D > 0$ is a constant called diffusivity with physical units of $\mathrm{m}^2\mathrm{s}^{-1}$. The minus sign in (6) accounts for the fact that the particles are transported from high to low densities. Using (6) in (5) gives the usual reaction-diffusion equation

$$u_t = D\triangle u + Q(\boldsymbol{x}, t, u, \ldots). \tag{7}$$

A heuristic derivation of the diffusion equation (1) involves the notion of a random walk. Consider a one-dimensional (1-d) random walker that at each time step t_i hops from its position x_n one unit to the left or right, $x_{n\pm1}$. The change in probability $p(x_n, t_i)$ to find the walker at x_n at t_i is equal to the sum of probabilities for it to hop into the point minus the sum of probabilities to hop off the point:

$$p(x_n, t_i) - p(x_n, t_{i-1}) \\ = \frac{1}{2}\left(p(x_{n+1}, t_{i-1}) + p(x_{n-1}, t_{i-1})\right) - p(x_n, t_{i-1}). \qquad (8)$$

Introducing the space and time scales of the motion, this equation can be rearranged as

$$\frac{p(x_n, t_i) - p(x_n, t_{i-1})}{\delta_t} \\ = \frac{\delta_x^2}{2\delta_t}\left[\frac{p(x_{n-1}, t_{i-1}) - 2p(x_n, t_{i-1}) + p(x_{n+1}, t_{i-1})}{\delta_x^2}\right]. \qquad (9)$$

Denoting $D = \delta_x^2/2\delta_t$, Equation (9) becomes a discrete approximation of Equation (1), and by taking the limit as $\delta_t, \delta_x \to 0$, keeping D finite, we recover Equation (1).

A rigorous derivation of the diffusion equation is obtained via stochastic calculus. The motion of individual particles is described by stochastic differential equations with the positions of the particles being modeled as random variables in R^n. The global behavior depends on the type of stochastic process governing the motion of the particles. Typically, one describes this in terms of the probability distribution of the random variable. In many cases, one finds that all its moments of order higher than 2 vanish. Consequently, the distribution of the population density, u, satisfies the second-order Fokker–Planck (or Kolmogorov's forward equation) (Øksendal, 2000)

$$u_t = \nabla(D(\boldsymbol{x}, t)u) - \nabla(C(\boldsymbol{x}, t)u). \qquad (10)$$

Equation (10) is a diffusion equation with a nonconstant, inhomogeneous diffusivity and convective term. It often arises for particles whose individual speeds are random deviations from some externally applied convection velocity. Examples arise in fluid flow and in biology (dispersals of population dispersals, e.g. chemotaxis), among other fields.

Particularly important cases are when the diffusivity depends on the population density such as in biology or ecology (Aronson & Weinberger, 1975). For example, insect dispersal models use the fact that the rate of spread of the population is increased at higher insect density. This is usually modeled with an equation of the form

$$u_t = \Delta(D(u)\nabla u) + f(u), \qquad (11)$$

where typically $D(u) = D_0 u^m$, $m > 0$ (Okubo & Levin, 2001; Murray, 2002). Density-dependent diffusion equations appear in many other fields. In physics, impurities are diffused into semiconductor materials (in the processing of silicon-based electronic devices) so the diffusivity is a function of the density of the semiconductor. Other examples include models of crystal growth, porous media, magnetic flux vortices in superconductors, surface reactions, and so on.

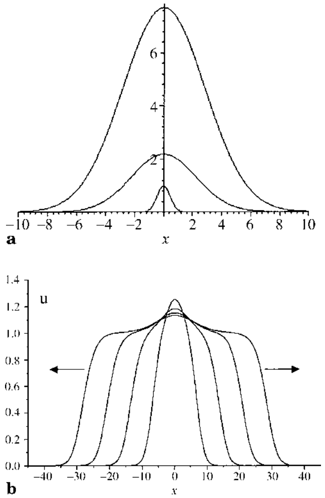

Figure 1. (a) Linear diffusion solutions of Equation (13). (b) Nonlinear diffusion solutions of Equation (14).

The behavior of the solution to Equation (1) is well understood. If Q is linear in u, then the solution is found by Fourier transform or eigenfunction-expansion techniques. For example, the linear diffusion equation

$$u_t = \Delta u + u \qquad (12)$$

with the initial condition $u(x, 0) = \delta(x)$ (Dirac's delta function) has, in (1-d), the fundamental solution

$$u(x, t) = \frac{1}{2\sqrt{\pi t}} \exp\left(t - \frac{x^2}{4t}\right), t > 0. \qquad (13)$$

Figure 1a illustrates the behavior of Equation (13) as a function of x for various times. Due to the linear source term, the solution grows exponentially (is unbounded). Another feature of solution (13) is the "paradox of infinite speed propagation." For all $x \neq 0$, $u(x, 0) = 0$, but $u(x, t) > 0$ for all $t > 0$. However, the diffusion equation describes well the global behavior of mass, as it can be easily verified that the center of mass does propagate with a finite speed.

The behavior of solutions of Equation (7) changes dramatically if Q is no longer linear in u. Consider the typical nonlinear autonomous form now called the Fisher–KPP equation and first investigated by Kolmogorov, Petrovsky, and Piscounoff (1937) and, separately, by Fisher (1937) to model the process of genetic diffusion:

$$u_t = \triangle u + u(1-u). \qquad (14)$$

The only solution evolving from a positive compactly supported initial data is always bounded and propagates in the form of a traveling wave with constant speed $v = 2$, see Figure 1b for an illustration. This is due to the combined action of diffusion and local nonlinearity and has been used in many models applied in biology (genetics, ecology, population dynamics, etc.), chemistry, combustion, economics, physics, etc. The equation similar to (14) but with a cubic nonlinearity (bistable model) instead of the quadratic was proposed in 1938 as a model for a flame front by Zeldovich and Frank-Kamenetsky (1938).

Razvan A. Satnoianu

See also **Fokker–Planck equation; Heat conduction; Reaction-diffusion systems; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Aronson, D.G. & Weinberger H.F. 1975. Nonlinear diffusion in population genetics, combustion and nerve pulse propagation. In *Partial Differential Equations and Related Topics*, edited by J.A. Goldstein, New York: Springer

Crank, J. 1975. *The Mathematics of Diffusion*, Oxford: Clarendon Press

Fisher, R.A. 1937. The wave of advance of advantageous genes. *Annuals of Eugenics*, 7, 353–369

Grindrod, P. 1996. *The Theory and Applications of Reaction–Diffusion Equations*, 2nd edition, Oxford: Clarendon Press and New York: Oxford University Press

Kolmogorov, A., Petrovsky, I. & Piscounoff, N. 1937. Etude de l'equation de la diffusion avec croissance de la quantité de matière et son application á un problème biologique. *Moscow Univ. Bull. Math.*, 1: 1–25

Murray, J.D. 2002. *Mathematical Biology*, vol. I, 3rd edition: *An Introduction*, Berlin and New York: Springer

Øksendal, B. 2000. *Stochastic Differential Equations*, Berlin and New York: Springer

Okubo, A. & Levin, S.A. 2001. *Diffusion and Ecological Problems: Modern Perspectives*, Berlin and New York: Springer

Zeldovich, Ya.B. & Frank-Kamenetsky, D.A. 1938. K teorii ravnomernogo rasprostranenia plameni [Toward a theory of uniformly propagating flames]. *Doklady Akademii Nauk SSSR*, 19: 693–697

DIMENSIONAL ANALYSIS

The dimensions of quantities in any equation (in physics these are mass, length, time, charge, temperature, angle, and so on) must be the same on both sides; otherwise an equality would be violated by changing units. From such reasoning, it is often possible to derive valuable insights without delving into mechanisms. This "dimensional analysis" is an old idea; Lord Rayleigh (John William Strutt), an early vigorous exponent of the method, called it the "principle of similitude," which he extolled in the following terms (Rayleigh, 1915): "It happens not infrequently that results in the form of 'laws' are put forward as novelties on the basis of elaborate experiments, which might have been predicted a priori after a few minutes' consideration."

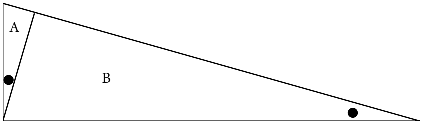

Figure 1. A right triangle is broken into two smaller right triangles of the same proportions.

The principle was familiar to Galileo, Newton, Fourier, Reynolds, and Maxwell, and was widely used in engineering around 1900. Edgar Buckingham formalized it in what is now called the Pi (for "product") Theorem, that any functional relation among N quantities represented by real numbers and collectively involving $U < N$ basic units can be rewritten as a dimensionless, constant function of $N - U$ dimensionless, multiplicative combinations of those variables (Buckingham, 1914). Unless N and U are trivially small, some systematic procedure is helpful for finding all possible ways of combining variables into dimensionless constants (Birkhoff, 1950; Coyle & Ballico-Lay, 1984).

Interestingly, there have been attempts to discover the laws of economics and finance by similar procedures, starting of course from a different list of fundamental units (DeJong, 1967).

As most articles on dimensional analysis expound abstract principles, two examples are presented here, one drawn from the ancient roots of mathematics and one from the nonlinear physics of shock waves.

Pythagorean Theorem

The area of a right triangle is uniquely determined by the length of the "long" side and one of the other ("wrong") angles. Area is some universal function of Long and Angle, let us say the smaller Angle, dotted in Figure 1. Do we have to figure out exactly what function? Maybe not. We know Area has to be proportional to Long^2 to make the dimensions come out right, and the other factor must be some dimensionless function of the dimensionless Angle (or equivalently of the ratio between the two shorter sides). Break this triangle into two smaller ones of exactly the same shape by constructing a perpendicular from the right angle to the Long side. Call the two areas Area A and Area B. So Area = Area A + Area B.

Each of these similar triangles has for its own long side one of the big triangle's shorter sides, and all the angles are the same as in the big one. Thus $\mathrm{Long}^2 f(\mathrm{Angle}) = \mathrm{ShortA}^2 f(\mathrm{Angle}) + \mathrm{ShortB}^2 f(\mathrm{Angle})$, where Angle means the dotted smaller angle, which is the same in all three cases. So unless $f(\mathrm{Angle}) = 0$ (i.e., unless the areas are all 0 anyhow), then

$$\mathrm{Long}^2 = \mathrm{ShortA}^2 + \mathrm{ShortB}^2. \qquad (1)$$

This is the theorem attributed to Pythagoras. We obtained it by merely insisting on dimensional consistency (Goldenfeld, 1992).

Atomic Explosions

From any big explosion in air, shock waves propagate outwards, slowing as the hemisphere of destruction expands. How fast does the hemisphere expand? You might think that an answer to this question necessarily involves a complex variety of considerations about sound, chemistry, thermal physics, and so on. Indeed it does, but let us consider an atomic explosion from the perspective of dimensional analysis.

Suppose the results are pretty much the same for any blast of the same total energy. If that were so, how could the expansion of the shock—its distance from ground zero (R) as a function of time (t)—depend on the energy (E) of the blast?

Let us denote the unit of length by the symbol L and the unit of time by the symbol T. (Thus L might be meters and T seconds.) Then energy (which is mass times velocity squared) has units (ML^2T^{-2}), where M denotes the unit of mass.

Since no combination of the three factors R, T, and E is unitless, we need to include another factor that involves mass. Such a factor is the air density, ρ. This factor is clearly important because without air there would be no shock wave but only bomb parts flying at fixed speeds through the vacuum of space. So throw ρ into the stew with units ML^{-3}. How to relate these four quantities (R, t, E, and ρ) to obtain a unitless result?

From the Pi Theorem, there is only one unitless combination of our four quantities (R, t, E, and ρ) that are expressed in terms of three basic units (L, t, and M). Thus,

$$\frac{R^5\rho}{t^2E} = a, \qquad (2)$$

where a is a dimensionless number that Geoffrey Taylor estimated (in 1941) to be 0.926 (Taylor, 1950a).

In 1949, Taylor verified his analysis by plotting the log of R (in meters) against the log of t (in seconds) for the first atomic explosion—the Trinity blast of 16 July 1945 in Alamogordo, New Mexico. According to Equation (2), log R versus log$t$ should be a straight line with a dimensionless slope $\frac{2}{5}$. Except for the first point at 0.1 ms, the log-log plot in Figure 2 is indeed close to a straight line with slope $\frac{2}{5}$. In dimensional terms,

$$\frac{R^5}{t^2} = 6 \times 10^{13}\mathrm{m}^5\mathrm{s}^{-2} \qquad (3)$$

to an experimental uncertainty of about 10%. Because a and ρ are known, the data in Figure 2 reveal the blast energy, which Taylor computed to be the equivalent of about 20 ktons of TNT (Taylor, 1950b).

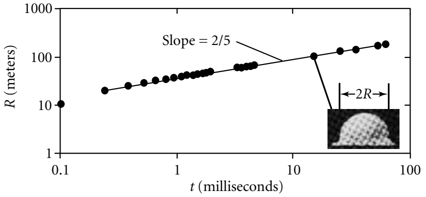

Figure 2. A log-log plot of the expansion of the Trinity fireball. From the origin of the blast, a hemispherical shock of radius R expands as a function of t during the first 62 ms. (The inset shows is a photograph of the blast at 15 ms, and the data are from Taylor, 1950b.)

In 1947, several of the photographs upon which Figure 2 is based were published by *Life*, a popular magazine of the time. Using these data, assuming $a = 1$, and taking $\rho = 1.25\ \mathrm{kg\,m^{-3}}$, Equations (2) and (3) imply a blast energy of $8.4 \times 10^{10}\ \mathrm{kg\,m^2\,s^{-2}}$ (or joules), which is equivalent to about 20 ktons of TNT. It may be presumed that interested parties made this simple estimate without delay.

Although these two examples may seem like magic, dimensional analysis has some limitations (Rayleigh, 1915). For example, this approach is useless in the absence of clear functional relations among mathematical quantities. Furthermore, one may doubt that the list of dimensioned variables presumed to be relevant is complete and does not include superfluous items. Finally, the desired relations may be unknown functions of dimensionless combinations of the pertinent variables.

A.T. WINFREE

See also **Explosions; Nerve impulses; Shock waves**

Further Reading

Birkhoff, G. 1950. *Hydrodynamics: A Study in Logic, Fact, and Similitude*, Princeton, NJ: Princeton University Press (Chapter 3 especially is a primary source and particularly lucid)

Bridgman, P.W. 1922. *Dimensional Analysis*, New Haven, CT: Yale University Press; often reprinted

Buckingham, E. 1914. On physically similar systems: illustrations of the use of dimensional equations. *Physical Review*, 4: 345–376

Buckingham, E. 1915. Model experiments and the form of empirical equations. *Transactions of the Americal Society of Mechanical Engineers*, 37: 263

Coyle, R.G. & Ballico-Lay, B. 1984. Concepts and software for dimensional analysis in modeling. *IEEE Transactions on*

Systems, Man, and Cybernetics SMC-14(3): 478–482 (Other than MatLab, the only source I know for software in this area)

DeJong, F.J. 1967. *Dimensional Analysis for Economists*, Amsterdam: North-Holland (The only effort I know to adapt dimensional analysis to economics and finance)

Goldenfeld, N. 1992. *Lectures on Phase Transitions and the Renormalization Group*, Reading, MA: Addison-Wesley (Contains full analysis of shock wave from explosions, esp. nuclear, following Taylor, 1952)

Rayleigh, Lord, 1915. The principle of similitude. *Nature*, 95: 66–68

Taylor, G.I. 1950a. The formation of a blast wave by a very intense explosion. I. Theoretical discussion. *Proceedings of the Royal Society of London*, 201A: 159–174

Taylor, G.I. 1950b. The formation of a blast wave by a very intense explosion. II. The atomic explosion of 1945. *Proceedings of the Royal Society of London*, 201A: 175–186

DIMENSIONS

The classical, integer-valued definition of dimension (see Hurewicz & Wallman, 1941) is defined inductively: the empty set has dimension -1, and a set has dimension n if n is the smallest integer such that every point has arbitrarily small neighborhoods whose boundaries have dimension less than n. This gives the "right" answer for smooth curves and surfaces, whose dimension we know intuitively.

In order to describe more accurately the complicated fractal sets that arise in nonlinear dynamics, we need to introduce more subtle definitions. Surprisingly, there are several generalizations of dimensions that still assign the intuitively correct dimensions to the above-noted well-behaved sets, and we recall two of them here.

The first of these is the "box-counting" dimension, also known as the Minkowski dimension, the fractal dimension, the entropy dimension, the capacity dimension, and the limit capacity: a litany of names that testifies to its popularity. For a subset X of $\mathbb{R}^n$, take a fixed array of boxes of side δ, and count the number $N_\delta(X)$ of these boxes that intersect with X. If $N_\delta(X) \sim \delta^{-d}$ as $\delta \to 0$, then X has box-counting dimension d. This can be made mathematically precise by defining

$$d_{\text{box}}(X) = \limsup_{\delta\to 0} \frac{\log N_\delta(X)}{-\log \delta}. \qquad (1)$$

(For alternative definitions that give the same quantity see Falconer (1990).) While the box-counting dimension is simple to define, it is not without problems. For example, the set $S = \{0\} \cup \{1/n : n = 1, 2, \ldots\}$, an unlikely candidate for a fractal, has box-counting dimension $\frac{1}{2}$. We now introduce another widely used definition of dimension that does not suffer from this anomaly.

We could try to define a notion of the "d-dimensional volume" of X as the limit of $N_\delta(X)\delta^d$ as $\delta \to 0$, but such a definition does not even agree with the standard definition of volume (Lebesgue measure) when d is an integer. Instead, the proper generalization of Lebesgue measure to non-integer dimensions is d-dimensional Hausdorff measure. Essentially, we cover a set $X \subset \mathbb{R}^n$ by a collection of balls of radii $r_i \le \delta$, and then let $\mathcal{H}^d(X)$ be the limit of $\sum r_i^d$ as δ tends to zero. More precisely, we define

$$\mathcal{H}^d(X) = \liminf_{\delta\to 0} \left\{ \sum_i r_i^d \,:\, r_i \le \delta \text{ and } X \subseteq \cup_i B_{r_i}(x_i) \right\}$$

(the notation $B_r(x)$ denotes an open ball centered at x of radius r). The resulting measure is proportional to Lebesgue measure when d is an integer. The "Hausdorff dimension" of X is the smallest value of d for which $\mathcal{H}^d(X)$ is finite,

$$d_{\text{H}}(X) = \inf\{d > 0 : \mathcal{H}^d(X) < \infty\}.$$

Since $\mu_d(X, \delta) \le N_\delta(X)\delta^d$, we always have $d_{\text{H}}(X) \le d_{\text{box}}(X)$ (and this inequality can be strict: the set S defined above has zero Hausdorff dimension). While harder to estimate in practice, the Hausdorff dimension is easier to deal with theoretically.

If we want to estimate the dimension of the attractor $\mathcal{A}$ of a dynamical system, it is useful to have a method based on dynamical quantities. In 1980, Douady & Oesterlé showed how to obtain a bound on $d_{\text{H}}(\mathcal{A})$, the dimension of the attractor of an iterated C^1 map f on $\mathbb{R}^n$. Denote by $\mathrm{D}f(x)$ the matrix of partial derivatives of f, i.e., $[\mathrm{D}f]_{ij} = \partial f_i/\partial x_j$, and let $\lambda_1(x) \ge \lambda_2(x) \ge \lambda_n(x)$ be the logarithms of the eigenvalues of $[\mathrm{D}f(x)^T \mathrm{D}f(x)]^{1/2}$. Now set

$$d(x) = j + \frac{\lambda_1(x) + \cdots + \lambda_j(x)}{|\lambda_{j+1}(x)|}, \qquad (2)$$

where j is the largest integer for which $\lambda_1(x) + \cdots + \lambda_j(x) \ge 0$ (note that $j \le d(x) < j + 1$). If $d > d(x)$, then any infinitesimal d-volume near x is contracted under the application of f, so

$$d_{\text{H}}(\mathcal{A}) \le \sup_{x\in\mathcal{A}} d(x). \qquad (3)$$

Hunt (1996) showed that the right-hand side of (3) also bounds $d_{\text{box}}(\mathcal{A})$. (A similar approach also works for the attractors of flows by taking f to be the time T map, for some suitable T. Constantin & Foias (1985) have proved a version of (3) for the attractors of infinite-dimensional dynamical systems.)

However, the box-counting and Hausdorff dimensions give equal weighting to all points in the attractor, while it is possible to have regions of the attractor that are visited very rarely. In such a situation, it can be more natural to consider invariant measures rather than attractors. As a (canonical) example of such a measure, suppose that $f : \mathbb{R}^n \to \mathbb{R}^n$ generates a dynamical system on $\mathbb{R}^n$. Then for any set X, we can define

$$\mu(X) = \lim_{m\to\infty} \frac{1}{m} \operatorname{card}\{k : f^k(x) \in X,\ 1 \le k \le m\},$$

where x is a point in the basin of attraction of $\mathcal{A}$. The quantity $\mu(X)$ is the proportion of time spent in X by a "typical trajectory" on the attractor.

There are various ways of defining the dimension of a measure μ. We could define the Hausdorff/box-counting dimension of μ to be the dimension of its support,

$$d_{\text{box/H}}(\mu) = \inf\{d_{\text{box/H}}(E) : \ \mu(E) = 1\},$$

but this still discounts the dynamical information contained in μ. Kaplan & Yorke (1979) defined the *Lyapunov dimension* of μ, $d_{\text{L}}(\mu)$, precisely as in (2), but replacing $\lambda_j(x)$ by the Lyapunov exponents associated with μ (the asymptotic growth rates of infinitesimal displacements about trajectories through μ-almost every choice of initial condition). In 1981, Ledrappier showed that for a very general class of dynamical systems, $d_{\text{H}}(\mu) \le d_{\text{L}}(\mu)$ (the inequality can be strict), while

$$d_{\text{box}}(\mathcal{A}) \le \sup_{\text{all invariant ergodic } \mu} d_{\text{L}}(\mu).$$

(Kaplan & Yorke had originally conjectured that $d_{\text{box}}(\mathcal{A}) = d_{\text{L}}(\mu)$.)

We now give two definitions of dimension that take into account the spatial structure of μ. The correlation at scale δ is defined by

$$C(\delta) = \int_{X\times X:\ |x-y|\le\delta} \mathrm{d}\mu(x)\,\mathrm{d}\mu(y),$$

which gives the probability that two points chosen according to the probability measure μ lie within δ of each other. If $C(\delta) \sim \delta^d$ as $\delta \to 0$, then d is the *correlation dimension* $d_{\text{corr}}(\mu)$. This was introduced by Grassberger & Procaccia (1983), who demonstrated that this quantity is particularly suited to numerical calculation.

Alternatively, define the "δ-entropy" $K_\delta(\mu) = -\sum_i \mu(B_i)\ln\mu(B_i)$, where $\{B_i\}$ is an array of boxes of side δ; the *information dimension* is given by

$$d_{\text{inf}}(\mu) = \lim_{\delta\to 0} \frac{K_\delta(\mu)}{-\log\delta}.$$

(Ruelle (1989) refers to $d_{\text{H}}(\mu)$ as the "information dimension" of μ: this should serve to emphasize how important it is when discussing dimensions to be explicit about the definition.)

Three of these dimensions occur as part of a scale of dimension-like quantities (see Grassberger, 1983). If B_i is an array of boxes of side δ, set

$$K_\delta(q) = \frac{1}{1-q}\log\sum_i \mu(B_i)^q$$

("the Renyi q entropy"). Note that $K_\delta(0) = \log N_\delta$ $(\operatorname{supp}\mu)$,

$$\lim_{q\to 1} K_\delta(q) = -\sum_i \mu(B_i)\ln\mu(B_i) = K_\delta(\mu)$$

and that since $C(\delta) \le \sum_i \mu(B_i)^2 \le C(\delta\sqrt{n})$, we have $\log C(\delta) \le K_\delta(2) \le \log C(\delta\sqrt{n})$. Now define the *Renyi dimensions* $D_q(\mu)$ by

$$D_q(\mu) = \lim_{\delta\to 0}\frac{K_\delta(q)}{-\log\delta}.$$

Then $D_0(\mu) = d_{\text{box}}(\mu)$, $D_1(\mu) = d_{\text{inf}}(\mu)$, and $D_2(\mu) = d_{\text{corr}}(\mu)$. Since D_q is non-increasing in q, we have in particular $d_{\text{corr}}(\mu) \le d_{\text{inf}}(\mu) \le d_{\text{box}}(\mu)$.

The Renyi dimensions are similar to quantities used to define the "multi-fractal spectrum." The theory relates the numbers $\tau(q) = (1-q)D_q$ to the frequency of various scaling behaviors about points on a fractal set: roughly, if for some small ε the number of δ-mesh cubes B_i with

$$\delta^{\alpha+\varepsilon} \le \mu(B_i) < \delta^\alpha$$

scales like $\delta^{-f(\alpha)}$, then

$$f(\alpha(q)) = \tau(q) + q\alpha(q),$$

where $q = f'(\alpha(q))$. The curve $f(\alpha)$ is the multi-fractal spectrum of the measure μ. (As remarked by Falconer (1990), the tempting interpretation of the "fractal spectrum" as the dimension of sets of points x where $\mu(B_\delta(x)) \sim \delta^\alpha$ is incorrect: typically the dimension of such sets will be zero or the same as that of the whole space, see Genyuk (1997/98).) Although these ideas have proved useful in the theory of turbulence (e.g. Frisch, 1995), their mathematical foundations have still to be fully resolved.

JAMES C. ROBINSON

See also **Attractors; Fractals; Lyapunov exponents; Measures**

Further Reading

Constantin, P. & Foias, C. 1985. Global Lyapunov exponents, Kaplan–Yorke formulas and the dimension of the attractor for 2D Navier–Stokes equation. *Communications in Pure and Applied Mathematics*, 38: 1–27

Douady, A. & Oesterlé, J.D. 1980. Dimension de Hausdorff des attracteurs. *Compes Rendus de l'Academies des Sciences, Paris Sèries A–B*, 290: A1135–A1138

Falconer, K. 1990. *Fractal Geometry*, Chichester and New York: Wiley

Frisch, U. 1995. *Turbulence*, Cambridge and New York: Cambridge University Press

Genyuk, J. 1997/98. A typical measure typically has no local dimension. *Real Analysis Exchange*, 23: 525–537

Grassberger, P. 1983. Generalized dimensions of strange attractors. *Physics Letters A*, 97: 227–230

Grassberger, P. & Procaccia, I. 1983. Measuring the strangeness of strange attractors. *Physica* D, 9: 189–208

Hunt, B. 1996. Maximal local Lyapunov dimension bounds the box dimension of chaotic attractors. *Nonlinearity*, 9: 845–852

Hurewicz, W. & Wallman, H. 1941. *Dimension Theory*, Princeton, NJ: Princeton University Press

Kaplan, J.L. & Yorke, J.A. 1979. Chaotic behavior of multi-dimensional difference equations. In *Functional Differential Equations and Approximation of Fixed Points*, Berlin: Springer, 204–227

Ledrappier, F. 1981. Some relations between dimension and Lyapounov exponents. *Communications in Mathematical Physics*, 81: 229–238

Ruelle, D. 1989. *Chaotic Evolution and Strange Attractors*, Cambridge and New York: Cambridge University Press

DIODES

Many of the desirable features of a wide variety of solid-state electronic devices are based on nonlinear current-voltage characteristics. These involve situations where a sufficiently high bias is applied to the device so that it either switches from one conductive state to another or oscillates between different conductive states.

A two-terminal electronic device is called a diode. In general, the current I through a diode depends upon the polarity of the applied voltage U and exhibits a nonlinear voltage dependence. This includes the important special case of a rectifier diode that conducts the current for one polarity of voltage and blocks it for the other. A rectifier can be realized by a p–n junction (a semiconductor that is p-doped on one side and n-doped on the other side), and it can be described approximately by the current-voltage characteristic

$$I = I_s\left(e^{eU/k_BT} - 1\right), \qquad (1)$$

where I_s is the saturation current for reverse bias $U < 0$, k_B is Boltzmann's constant, and T is the temperature (Figure 1a). For forward bias (positive voltage applied to the p side), the majority carriers (holes from the p side and electrons from the n side) flow towards the junction where they recombine, while for reverse bias they are pulled away from the junction.

The Schottky diode is a rectifier diode, which consists of a metal-semiconductor contact. Interface states and to a small degree the difference in work functions between the metal and the semiconductor give rise to a potential barrier (Schottky barrier) and depletion of majority carriers in the barrier region for reverse bias. The current-voltage characteristic is similar to Equation (1).

Depending upon bias conditions, doping profiles, and device geometry, various other terminal functions of diodes are possible, which involve nonlinear behavior of the conductivity, the capacitance, and the inductance of the device.

The Zener diode is a p–n junction that exhibits a sharp increase in the magnitude of the current at a certain well-controlled reverse bias (Figure 1b). Depending upon the specific device structure, it is either due to avalanche breakdown (multiplication of carriers by impact ionization occurring at high voltage) or due to tunneling across the bandgap (occurring at lower voltage). It is used to stabilize and limit the dc voltage in circuits, utilizing the property that the current can vary over a large range at the breakdown threshold without noticeable change in voltage.

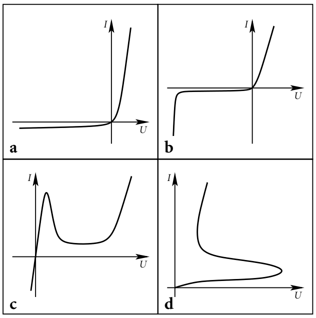

Figure 1. Typical nonlinear current-voltage chareristics of diodes. (a) p–n diode, (b) Zener diode, (c) Tunnel diode, (d) p–i–n diode (schematic).

A class of diodes exhibits a nonmonotonic dependence of the current I upon voltage U. Negative differential conductance $\mathrm{d}I/\mathrm{d}U < 0$ can arise due to various mechanisms and may result in self-generated oscillations and complex self-organized spatiotemporal patterns. A famous example is the Esaki tunnel diode (Figure 1c) that, in its original version, consists of a heavily doped p–n junction. In thermal equilibrium, the Fermi level lies within the conduction band on the n side and within the valence band on the p side. When a small forward bias is applied, electrons can tunnel from the n- to the p-side where they find empty states in the valence band. With increasing bias, the filled conduction band states on the n side move up, and the overlap with empty states in the valence band decreases; consequently, the tunneling current decreases, and $\mathrm{d}I/\mathrm{d}U < 0$. With a further increase of the bias, diffusion of the electrons over the barrier sets in as in a normal p–n junction, and the current increases again.

Modern variants of the tunnel diode are double-barrier resonant tunneling structures and superlattices that consist of alternating layers of different semiconductor materials (heterostructures) forming potential barriers and quantum wells on length scales of a few nanometers. The current density across the barrier between two wells is maximum if there is maximum overlap between the occupied states in one well and the available unoccupied states in the other, that is, if the energies are in resonance. For low bias, equivalent levels in adjacent wells are approximately in resonance, while for higher bias the ground energy level in one quantum well becomes aligned with the second level in the neigboring well. Thus, resonant tunneling produces an $I(U)$ characteristic similar to Figure 1c. High-frequency oscillations up to 150 GHz can be generated in resonant tunneling structures and superlattices.

Gunn diodes are used to generate and amplify microwaves at frequencies typically beyond 1 GHz (these devices are called "diodes" because they are two-terminal devices, but no p–n junction is involved). The mechanism is based upon field-induced intervalley transfer of electrons from a high- to a low-mobility valley in the conduction band, and the manifestation of the current instability (oscillations or switching) is determined primarily by the cathode contact boundary condition. Traveling high-field domains (Gunn domains), which show up as transit time oscillations, represent an important mode of operation.

Multilayered structures with alternating p- and n-doping represent another class of diodes that exhibit negative differential conductance, high-power microwave oscillations well above 30 GHz, and complex spatiotemporal current density patterns. Starting from the basic n^+–p–i–p^+ structure, where "+" denotes high doping and i is an intrinsic (undoped) layer, various modifications like IMPATT (impact ionization avalanche transit time) diodes, TRAPATT (trapped plasma avalanche-triggered transit) diodes, p–i–n diodes (with double injection of electrons and holes), or $n^+p^+np^-p^+$ devices, where "−" denotes low doping, have been studied. These structures typically display bistability and switching from a low- to a high-conductivity state where carrier multiplication and avalanche breakdown set in (Figure 1d). Negative differential conductance is also exhibited by multilayer systems composed of layers of different semiconductor materials (heterojunctions) like the heterostructure hot electron diode or real-space transfer devices.

A p–n diode may also be operated as a nonlinear capacitor. As the depletion-layer width increases with increasing reverse bias, its capacitance decreases in a controlled way depending upon the doping profile. This effect is used in varactor (*var*iable re*actor*) diodes to tune the capacitance and for parametric amplification. State-of-the-art developments include studies on p–n diodes with embedded quantum dot nanostructures, which strongly exhibit nonlinear capacitance-voltage characteristics as signatures of the charging of these quantum dots.

Nonlinear inductors represent another class of two-terminal devices. A Josephson junction consists of two superconductors separated by a thin nonsuperconducting region. The tunneling of superconducting electron pairs produces a current at zero voltage called the Josephson current. Switching can be obtained from this state to a voltage state by either current overdrive or a magnetic field. The dynamic response of the Josephson junction can be represented by an equivalent circuit with an intrinsic field-dependent inductance.

ECKEHARD SCHÖLL

See also **Avalanche breakdown; Josephson junctions; Nonlinear electronics; Semiconductor oscillators**

Further Reading

Böer, K.W. 2002. *Survey of Semiconductor Physics*, 2nd edition, New York: Plenum

Ibach, H. & Lüth, H. 2003, *Solid-State Physics*, Berlin: Springer

Schöll, E. 2001. *Nonlinear Spatio-Temporal Dynamics and Chaos in Semiconductors*, Cambridge and New York: Cambridge University Press

Shaw, M.P., Mitin, V.V., Schöll, E. & Grubin, H.L. 1992. *The Physics of Instabilities in Solid State Electron Devices*, New York: Plenum Press

Sze, S.M. 1981. *Physics of Semiconductor Devices*, New York: Wiley

Sze, S.M. 1998. *Modern Semiconductor Device Physics*, New York: Wiley

DIRAC'S DELTA FUNCTION

See **Generalized functions**

DISCRETE BREATHERS

The study of dynamical nontopological localization in translationary invariant nonlinear Hamiltonian lattices has experienced considerable development during the late 1990s (Sievers & Page, 1995; Aubry, 1997; Flach & Willis, 1998). The discreteness of space—that is, the use of a spatial lattice—is crucial in order to provide structural stability for spatially localized excitations. Spatial discreteness is a very common situation for various applications from, for example, solid-state physics.

To make things precise, consider a d-dimensional hypercubic spatial lattice with discrete translational invariance. Each lattice site is labeled by a d-dimensional vector l with integer components. To each lattice site we associate one pair of canonically conjugated coordinates and momenta X_l, P_l that are real functions of time t. Let us then define some Hamiltonian H being a function of all coordinates and momenta and further require that H has the same symmetries as the lattice. The dynamical evolution of the system is given by the usual Hamiltonian equations of motion. Without loss of generality, let us consider that H is a nonnegative function and that $H = 0$ for $X_l = P_l = 0$ (for all l's). We call this state the classical ground state. Generalizations to other lattices and larger numbers of degrees of freedom per lattice site are straightforward.

When linearizing the equations of motion around $H = 0$, we obtain an eigenvalue problem. Due to translational invariance the eigenvectors will be spatially extended plane waves, and the eigenvalues Ω_q (frequencies) form a phonon spectrum, that is, Ω_q is a function of the wave vector q. Due to the translation symmetry of the Hamiltonian, Ω_q will be periodic in

q. Moreover, the phonon spectrum will be bounded, that is, $|\Omega_q| \leq \Omega_{\max}$. Depending on the presence or absence of Goldstone modes Ω_q might be gapless (zero belongs to the spectrum, spectrum is acoustic) or exhibit a gap ($|\Omega_q| \geq \Omega_{\min}$, spectrum is optical). Increasing the number of degrees of freedom per lattice site induces several branches in Ω_q with possible gaps between them.

Let us search for spatially localized time periodic solutions of the full nonlinear equations of motion, that is, $X_{|l| \to \infty} \to 0$, and

$$X_l(t) = X_l(t + T_{\rm b}) + \lambda k_l,\ P_l(t) = P_l(t + T_{\rm b}), \quad (1)$$

with k_l being integers and λ a spatial period (the equations of motion should be invariant under shifts of X_l by multiples of λ if applicable.) These solutions are called discrete breathers. If $k_l \neq 0$ for a finite subset of lattice sites, the solutions are sometimes called "rotobreathers."

If a solution exists, we can expand it into a Fourier series in time, that is, $X_l(t) = \sum_k A_{k_l} {\rm e}^{{\rm i}k\omega_{\rm b}t}$ ($\omega_{\rm b} = 2\pi/T_{\rm b}$). Spatial localization implies $A_{k_l \to \infty} \to 0$. Inserting these series into the equations of motion results in a set of coupled algebraic equations for the Fourier amplitudes (Flach & Willis, 1998). Consider the spatial tail of the solution where all Fourier amplitudes are small and should further decay to zero with growing distance from the excitation center. Since all amplitudes are small, the equations of motion can be linearized. This procedure decouples the interaction in k-space, and we obtain for each k a linear equation for A_{k_l} with coupling over l. This equation will contain $k\omega_{\rm b}$ as a parameter. It will, in fact, be identical to the above-discussed equation linearized around $H = 0$, and it will contain $k\omega_{\rm b}$ instead of Ω_q (Flach & Willis, 1998). If $k\omega_b = \Omega_q$, the corresponding amplitude A_{k_l} will not decay in space, instead it will oscillate. To obtain localization, we arrive at the nonresonance condition (Flach & Willis, 1998)

$$k\omega_b \neq \Omega_q. \quad (2)$$

This condition has to be fulfilled for all integer k. For an optical spectrum Ω_q, frequency ranges for $\omega_{\rm b}$ exist, which satisfy this condition. For acoustic spectra, $k = 0$ has to be considered separately (Flach et al., 1997).

The nonresonance condition is only a necessary condition for generic occurrence of discrete breathers. More detailed analysis shows that breathers being periodic orbits bifurcate from band edge plane waves (Flach, 1996). The condition for this bifurcation is an inequality involving parameters of expansion of H around $H = 0$ (Flach, 1996). Rigorous existence proofs for weakly coupled anharmonic oscillators use, for example, the implicit function theorem (MacKay & Aubry, 1994).

Discrete breathers (periodic orbits) appear generically as one-parameter families of periodic orbits. The parameter of the family can be, for example, the frequency (or energy, action, etc.). Note that we do not need any topological requirement on H (no energy barriers). Indeed, breather families possess limits where the breather delocalizes and its amplitude becomes zero.

With the help of the nonresonance conditions, we can exclude the generic existence of spatially localized solutions that are quasi-periodic in time. Indeed, in the simplest case, we would have to satisfy a nonresonance condition $k_1\omega_1 + k_2\omega_2 \neq \Omega_q$ for ω_1/ω_2 being irrational and all possible pairs of integers k_1, k_2. This is impossible (Flach, 1994).

The nonresonance condition also explains why breather solutions are nongeneric for nonlinear Hamiltonian field equations, since Ω_q becomes unbounded for a spatially continuous system. Consequently, generically an infinite number of unavoidable resonances destroys the breather existence there.

Note that in many cases breathers can be easily excited by choosing some localized perturbation of the lattice system. Integrating numerically the equations of motion, we find that the energy distribution is not delocalizing but stays essentially localized over several orders of magnitude of the characteristic phonon periods. These numerical results clearly show that breathers are not only interesting solutions but can be rather typical and robust depending on the system's parameters.

Note that breathers can also exist for autonomous forced damped systems (MacKay & Sepulchre, 1998). In these systems, contrary to the Hamiltonian ones, breather periodic orbits do not come in one-parameter families of the frequency $\omega_{\rm b}$, but correspond to limit cycle attractors that are isolated in the system's phase space.

SERGEJ FLACH

See also **Breathers; Discrete nonlinear Schrödinger equations; Discrete self-trapping system; Fermi–Pasta–Ulam oscillator chain; Josephson junction arrays; Local modes in molecular crystals; Multidimensional solitons; Nonlinear Schrödinger equations; Solitons**

Further Reading

Aubry, S. 1997. Breathers in nonlinear lattices: existence, linear stability and quantization. *Physica* D, 103: 201

Flach, S. 1994. Conditions on the existence of localized excitations in nonlinear discrete systems. *Physical Review* E, 50: 3134

Flach, S. 1996. Tangent bifurcation of band edge plane waves, dynamical symmetry breaking and vibrational localization. *Physica* D, 91: 223

Flach, S., Kladko, K. & Takeno. S. 1997. Acoustic breathers in two-dimensional lattices. *Physical Review Letters*, 79: 4838

Flach, S. & Willis, C.R. 1998. Discrete breathers. *Physics Reports*, 295: 182

MacKay, R.S. &. Aubry, S. 1994. Proof of existence of breathers for time-reversible or Hamiltonian networks of weakly coupled oscillators. *Nonlinearity*, 7: 1623

MacKay, R.S. & Sepulchre, J.A. 1998. Stability of discrete breathers. *Physica* D, 119: 148

Sievers, A.J. & Page, J.B. 1995. Unusual Anharmonic Local Mode Systems. In *Dynamical Properties of Solids VII Phonon Physics The Cutting Edge*, edited by G.K. Horton and A.A. Maradudin, Amsterdam: Elsevier

DISCRETE NONLINEAR SCHRÖDINGER EQUATIONS

With a fairly generous definition, a discrete nonlinear Schrödinger (DNLS) equation is any equation that can be obtained from a nonlinear Schrödinger (NLS) equation of general form

$$\mathrm{i}\frac{\partial\phi}{\partial t} + \Delta\phi + f(|\phi|^2)\phi = 0, \tag{1}$$

by employing some finite-difference approximation to the operators acting on the space-time-dependent continuous field $\phi(\boldsymbol{r}; t)$. In (1), $\Delta = \nabla^2$ is the Laplace operator acting in one, two, or three spatial dimensions, and f is a quite general function that, for most purposes, is taken to be differentiable and with $f(0) = 0$. In the most well-known case of cubic nonlinearity, $f(|\phi|^2) = \gamma|\phi|^2$, Equation (1) is often referred to as *the* NLS equation, and is integrable with the inverse scattering method if the number of spatial dimensions is one. Here we use the term DNLS equation to denote the set of coupled ordinary differential equations resulting from discretizing *all* spatial variables in (1), while keeping the time-variable t continuous. However, one may also consider equations with discrete time ("fully discrete NLS equations"), as well as equations with only some of the spatial dimensions discretized ("discrete-continuum NLS equations"). The former are of interest as algorithms for numerical solution of (1), while the latter may describe pulse propagation in *arrays* of coupled nonlinear optical fibers (Aceves et al., 1995).

The simplest example of a DNLS equation can be formally obtained by just replacing the Laplacian operator in (1) with the corresponding discrete Laplacian. Thus, for the one-dimensional (1-d) case, we let $\phi_n(t) \equiv \phi(x = na; t)$ where a is the lattice parameter, so that for the particular case of cubic nonlinearity, the following equation is obtained:

$$\mathrm{i}\frac{\mathrm{d}\phi_n}{\mathrm{d}t} + C(\phi_{n+1} - 2\phi_n + \phi_{n-1}) + \gamma|\phi_n|^2\phi_n = 0, \tag{2}$$

where $C = 1/a^2$. This set of differential–difference equations with purely diagonal ("on-site") nonlinearity is sometimes called the diagonal DNLS (DDNLS) equation, but since it is by far the most studied example of a DNLS equation, it is most commonly referred to as simply *the* DNLS equation. Extensions to higher dimensions are straightforward, so that, for example, for a 2-d lattice with $x = ma$, $y = na$, the DNLS equation reads

$$\mathrm{i}\frac{\mathrm{d}\phi_{m,n}}{\mathrm{d}t} + C(\phi_{m+1,n} + \phi_{m-1,n} + \phi_{m,n+1} + \phi_{m,n-1} - 4\phi_{m,n}) + \gamma|\phi_{m,n}|^2\phi_{m,n} = 0. \tag{3}$$

The study of DDNLS equations has a long and fascinating history, beginning in the 1950s within solid-state physics with Holstein's model for polaron motion in molecular crystals (Holstein, 1959); reappearing in the 1970s within biophysics with Davydov's model for energy transport in biomolecules (see, e.g., Scott, 1999, Chapter 5.6), in the 1980s within physical chemistry in the theory of local modes of small molecules (see, e.g., Scott, 1999, Chapter 5.4), and within nonlinear optics modeling coupled nonlinear waveguides (see, e.g., Hennig & Tsironis, 1999, Chapter 1.4); and most recently around the turn of the century within matter wave physics in the description of a dilute Bose–Einstein condensate trapped in a periodic potential (Trombettoni & Smerzi, 2001). A brief account of experimental verifications of the validity of the DNLS description in the two latter contexts available at the time of writing was given in Eilbeck & Johansson (2003, Chapter 10). In addition, the DDNLS equation has played a central role in the development of the general theory for intrinsic localized modes ("discrete breathers") in systems of coupled anharmonic oscillators during the 1990s (Flach & Willis, 1998). The reader should also note that the DDNLS equation is a particular example of the more general "discrete self-trapping" (DST) systems (described under a separate entry), where the general DST dispersion matrix describing interactions between lattice sites is restricted to nearest-neighbor couplings. Thus, the general theory described for DST systems is also applicable for the DDNLS equation.

The reason for the ubiquity of the DDNLS equation in nonlinear lattice systems is analogous to that of the NLS equation for continuum systems: it takes into account dispersion (through the nearest-neighbor interaction terms) as well as nonlinearity (the term $\gamma|\phi_n|^2\phi_n$) at the lowest order of approximation. It can be derived, for example, from a general system of coupled anharmonic oscillators using a "rotating wave approximation" (RWA), where it is assumed that each oscillator approximately can be described by a complex rotating-wave amplitude as $u_n(t) = \mathrm{Re}(\phi_n \mathrm{e}^{-\mathrm{i}\omega t})$ (see, e.g., Flach & Willis, 1998, Chapter 2.3; Scott, 1999, Chapter 5.3.1). Thus, the RWA assumes time-periodic solutions to have a purely harmonic time dependence, neglecting the generation of all higher harmonics. This approximation can be justified for small-amplitude oscillations in weakly coupled oscillator chains, using perturbational techniques with expansions on multiple time scales (see, e.g., Flach & Willis, 1998, Chapter 2.2, for a general outline, and Morgante et al., 2002, Chapter 2.2, for details).

As for general DST systems, the DDNLS equation has, in addition to the energy (Hamiltonian) $H=\sum_n \left(C|\phi_{n+1}-\phi_n|^2-\frac{\gamma}{2}|\phi_n|^4\right)$ (where $\mathrm{i}\phi_n$ and ϕ_n^* are canonical conjugated variables), a second conserved quantity, which is the excitation number $N=\sum_n |\phi_n|^2$. The conservation of excitation number results (through Noether's theorem) from the invariance of the equation under infinitesimal transformations of the overall phase ($\phi_n \to \phi_n \mathrm{e}^{\mathrm{i}\varepsilon}$). As a consequence, the DDNLS equation is integrable for two degrees of freedom but nonintegrable for larger systems. Still, the existence of a second conserved quantity has some notable consequences, which makes the DDNLS equation nongeneric among general Hamiltonian lattice systems, such as:

- It has purely harmonic time-periodic solutions $\phi_n(t)=A_n\mathrm{e}^{-\mathrm{i}\omega t}$ with time-independent A_n ("stationary solutions").
- It has continuous families of time-quasi-periodic solutions, with two incommensurate frequencies, which may be spatially exponentially localized also in infinite systems ("quasi-periodic breathers") (see, e.g., Eilbeck & Johansson, 2003, Chapter 7; Kevrekidis et al., 2001, Chapter 2.4).

From a mathematical point of view, it is highly interesting that there also exist discretizations of the integrable 1-d cubic NLS equation that conserve its integrability. The most famous integrable DNLS equation is the so-called Ablowitz–Ladik (AL) DNLS equation, the integrability of which was first proven by Ablowitz & Ladik (1976). It is obtained by replacing the nonlinear term $\gamma|\phi|^2\phi$ with an "off-diagonal" discretization $\frac{\gamma}{2}|\phi_n|^2(\phi_{n+1}+\phi_{n-1})$, yielding

$$\begin{aligned}\mathrm{i}\frac{\mathrm{d}\phi_n}{\mathrm{d}t} &+ C(\phi_{n+1}-2\phi_n+\phi_{n-1})\\ &+\frac{\gamma}{2}|\phi_n|^2(\phi_{n+1}+\phi_{n-1})=0.\end{aligned} \tag{4}$$

Due to its integrability, it is possible to obtain exact analytical solutions to (4) describing, for example, traveling waves (in terms of elliptic functions, see, e.g., Scott, 1999, Chapter 5.3.2), solitons, and multisolitons. In particular, for $\gamma/C>0$, there is a (bright) one-soliton solution given by (with rescalings such that $C=1, \gamma=2$):

$$\phi_n(t) = \sinh\beta\,\mathrm{sech}[\beta(n-vt)]\mathrm{e}^{\mathrm{i}(kn+\omega t+\alpha)}, \tag{5}$$

where β, k, and α are free parameters,

$$v=(2/\beta)\sinh\beta\sin k,$$

and $\omega=2(\cosh\beta\cos k-1)$. On the other hand, when $\gamma/C<0$, there are dark-soliton solutions with nonvanishing amplitude as $|n|\to\infty$ (Vekslerchik & Konotop, 1992).

The ALDNLS equation (4) also has a Hamiltonian structure with

$$\begin{aligned}H = \sum_n \Big[&-C\phi_n^*(\phi_{n+1}+\phi_{n-1})\\ &+\frac{4C^2}{\gamma}\log\left(1+\frac{\gamma}{2C}|\phi_n|^2\right)\Big],\end{aligned}$$

although the conjugated variables $\mathrm{i}\phi_n$ and ϕ_n^* are noncanonical and the corresponding Poisson bracket deformed (see, e.g., Faddeev & Takhtajan, 1987, p. 303; Scott, 1999, Chapter 5.3.2). There is, at this writing, no known direct physical application of the ALDNLS equation; however, it is commonly used as a starting point for perturbational studies of physically more relevant equations such as (2). A particularly interesting model allowing interpolations between Equations (2) and (4) is the so-called "Salerno equation," which is described under a separate entry.

As a final example, we mention a rather complicated DNLS equation, which was introduced and proven to be integrable by Izergin and Korepin in 1981. This Izergin–Korepin equation reads

$$\mathrm{i}\frac{\mathrm{d}\phi_n}{\mathrm{d}t} = 4\phi_n + \frac{P_{n,n+1}}{Q_{n,n+1}} + \frac{P_{n,n-1}}{Q_{n,n-1}}, \tag{6}$$

where

$$\begin{aligned}P_{n,n\pm1} = {}&-\phi_n-\phi_{n\pm1}\sqrt{\left(1-\frac{\gamma}{8}|\phi_n|^2\right)\left(1-\frac{\gamma}{8}|\phi_{n\pm1}|^2\right)}\\ &+\frac{\gamma}{4}\phi_n|\phi_{n\pm1}|^2+\frac{\gamma}{16}\left(|\phi_n|^2\phi_{n\pm1}+\phi_n^2\phi_{n\pm1}^*\right)\\ &\times\sqrt{\frac{1-\frac{\gamma}{8}|\phi_{n\pm1}|^2}{1-\frac{\gamma}{8}|\phi_n|^2}}\end{aligned}$$

and

$$\begin{aligned}Q_{n,n\pm1} = {}&1-\frac{\gamma}{8}\left(|\phi_n|^2+|\phi_{n\pm1}|^2\right)\\ &-\frac{\gamma}{8}\left(\phi_n\phi_{n\pm1}^*+\phi_n^*\phi_{n\pm1}\right)\\ &\times\sqrt{\left(1-\frac{\gamma}{8}|\phi_n|^2\right)\left(1-\frac{\gamma}{8}|\phi_{n\pm1}|^2\right)}\\ &+\frac{\gamma^2}{32}|\phi_n|^2|\phi_{n\pm1}|^2.\end{aligned}$$

It is associated with a lattice Heisenberg magnet model, and turns into the cubic NLS equation in the continuum limit (see Faddeev & Takhtajan, 1987, pp. 299, for details).

MAGNUS JOHANSSON

See also **Discrete self-trapping system; Nonlinear Schrödinger equations; Rotating wave approximation; Salerno equation**

Further Reading

Ablowitz, M.J. & Ladik, J.F. 1976. Nonlinear differential-difference equations and Fourier analysis. *Journal of Mathematical Physics*, 17(6): 1011–1018

Aceves, A.B., De Angelis, C., Luther, G.G., Rubenchik, A.M. & Turitsyn, S.K. 1995. All-optical-switching and pulse amplification and steering in nonlinear fiber arrays. *Physica* D, 87(1–4): 262–272

Eilbeck, J.C. & Johansson, M. 2003. The discrete nonlinear Schrödinger equation—20 years on. In *Localization and Energy Transfer in Nonlinear Systems: Proceedings of the Third Conference, San Lorenzo de El Escorial, Spain, 17–21 June 2002*, edited by L. Vázquez, M.P. Zorzano and R. MacKay. Singapore: World Scientific

Faddeev, L.D. & Takhtajan, L.A. 1987. *Hamiltonian Methods in the Theory of Solitons*. Berlin and Heidelberg: Springer

Flach, S. & Willis, C.R. 1998. Discrete breathers. *Physics Reports*, 295(5): 181–264

Hennig, D. & Tsironis, G.P. 1999. Wave transmission in nonlinear lattices. *Physics Reports*, 307(5–6), 333–432

Holstein, T. 1959. Studies of polaron motion. *Annals of Physics*, 8: 325–389

Kevrekidis, P.G., Rasmussen, K.Ø. & Bishop, A.R. 2001. The discrete nonlinear Schrödinger equation: a survey of recent results. *International Journal of Modern Physics* B, 15(21): 2833–2900

Morgante, A.M., Johansson, M., Kopidakis, G. & Aubry, S. 2002. Standing wave instabilities in a chain of nonlinear coupled oscillators. *Physica* D, 162(1–2): 53–94

Scott, A. 1999. *Nonlinear Science: Emergence & Dynamics of Coherent Structures*. Oxford and New York: Oxford University Press

Trombettoni, A. & Smerzi, A. 2001. Discrete solitons and breathers with dilute Bose-Einstein condensates. *Physical Review Letters*, 86(11): 2353–2356

Vekslerchik, V.E. & Konotop, V.V. 1992. Discrete nonlinear Schrödinger equation under non-vanishing boundary conditions. *Inverse Problems*, 8(6): 889–909

DISCRETE SELF-TRAPPING SYSTEM

In the early 1980s, experimental evidence suggested that vibrational energy in natural proteins (specifically the CO stretch oscillation of the peptide unit) might become self-localized, with implications for the storage and transport of energy in biological organisms (Careri et al., 1984). Because the structures of natural proteins take a wide variety of shapes, the following equation—called the discrete self-trapping (DST) equation (Eilbeck et al., 1985)—was proposed to capture the essential features of self-localization:

$$\left(\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t} - \omega_0\right)\boldsymbol{U} + \gamma D(|\boldsymbol{U}|^2)\boldsymbol{U} + \varepsilon M\boldsymbol{U} = 0. \quad (1)$$

Here, $\boldsymbol{U}(t) \equiv \mathrm{col}(u_1, u_2, \ldots, u_f)$ is a column vector representing the amplitudes of f oscillatory modes, each of which is described in the rotating wave approximation by the complex amplitudes: $u_1(t), u_2(t), \ldots, u_f(t)$. With $\gamma = 0$ and $\varepsilon = 0$, these modes (sites) oscillate independently and sinusoidally at the site frequency ω_0.

In the last term of Equation (1), $M = [m_{jk}]$ is an $f \times f$ dispersion matrix, expressing energetic interactions among the f modes stemming from electromagnetic couplings. This matrix is real and symmetric ($m_{ij} = m_{ji}$), and its diagonal elements can be chosen to represent small variations in the site frequencies from ω_0. Thus, with $\gamma = 0$ but with the dispersion parameter ε not zero, there will be f modes of oscillation given by eigenvectors of the matrix $(M - \omega_0 I)$.

In the second term of Equation (1), the parameter γ introduces nonlinearity into the formulation, where $D(|\boldsymbol{U}|^2) \equiv \mathrm{diag}|(|u_1|^2, |u_2|^2, \ldots, |u_f|^2)$ is a diagonal matrix. Thus, with $\varepsilon = 0$ but with the nonlinear parameter γ not equal to zero, each site is an independent (uncoupled) anharmonic oscillator.

From a physical perspective, there are two types of nonlinearity: intrinsic (stemming from weakening of electronic bonding with mode amplitude) and extrinsic (arising from interactions between localized vibrations and the lattice) (Scott, 2003). Intrinsic nonlinearity governs energy localization in small molecules (such as "local modes" of the CH stretch oscillations in benzene), which has been known to physical chemists since the 1920s (Ellis, 1929).

With both γ and ε not equal to zero, the DST equation displays an interesting variety of regular and chaotic motions for various values of the energy

$$H = \sum_j \left(\omega_0|u_j|^2 - \frac{\gamma}{2}|u_j|^4\right) - \varepsilon \sum_{j,k} m_{jk} u_j^* u_k \quad (2)$$

and the "mass" $N = |u_1|^2 + |u_2|^2 + \cdots + |u_f|^2$. Numerical studies of these motions provide insights into the dynamics of vibrational energy in proteins and in small molecules. (For descriptions at small levels of oscillatory energy, Equation (1) is conveniently quantized, Scott et al., 1994.)

The scaling $\boldsymbol{U} \to \boldsymbol{U}\exp(-\mathrm{i}\omega_0 t)$ reduces (1) to the standard form of the DST equation

$$\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} + \gamma D(|\boldsymbol{U}|^2)\boldsymbol{U} + \varepsilon M\boldsymbol{U} = 0. \quad (3)$$

An important class of solutions of the DST equation are the so-called stationary solutions ("stationary" because amplitudes are time-independent), which satisfy the ansatz

$$\boldsymbol{U}(t) = \boldsymbol{y}\exp(\mathrm{i}\omega t).$$

Inserting this into (3), we see that the constant vector $\boldsymbol{y} = \mathrm{col}(y_1, y_2, \ldots, y_f)$ satisfies the nonlinear eigenvalue problem

$$-\omega\boldsymbol{y} + \gamma D(|\boldsymbol{y}|^2)\boldsymbol{y} + \varepsilon M\boldsymbol{y} = 0, \quad (4)$$

where ω plays the role of the eigenvalue. In many cases, the vector $\boldsymbol{y}$ can be chosen to be real. In the limit $\varepsilon \to 0$ (alternatively $\gamma \to \infty$), called the "anti-integrable" limit, we have $y_i = 0$ or $|y_i|^2 = \omega/\gamma$, and in the real case, we have $y_i = \pm\sqrt{N/K}$, where K is the number of nonzero y_i on the chain. Starting from these solutions, branches of solutions for nonzero ε can then be generated numerically by path-following methods (Eilbeck et al., 1984). In the limit $\gamma \to 0$, (4) becomes a linear eigenvalue problem.

The values of y_i in the anti-integrable limit serve as a useful classification scheme for the branch, where we denote the three different (real) limits by the symbols $\cdot$, $\uparrow$, and $\downarrow$, respectively. The simplest stationary localized solution has only one nonzero mode or site amplitude in the anti-integrable limit, with $y_k = \sqrt{N}$, $y_i = 0$, $i \neq k$. For small but nonzero ε the amplitudes on the sites $\neq k$ are small and tend to zero exponentially as $|k - i| \to \infty$.

Such a localized solution was called a soliton in the early work on the DST equation but is now more often referred to as a breather due to its internal degree of freedom ω. A single breather in the center of a 1-dimensional lattice would be written as $(\ldots\cdots\uparrow\cdots\ldots)$, and would appear as in Figure 1. More complicated breather structures are possible, such as $(\ldots\cdots\uparrow\uparrow\cdots\ldots)$ and $(\ldots\cdots\uparrow\downarrow\cdots\ldots)$.

Breathers satisfying (4) are referred to as stationary breathers; more general mobile solutions are of interest. Feddersen (1991b) carried out a numerical study of solutions of the form

$$U_n(t) = \psi(n - ct)\mathrm{e}^{\mathrm{i}(kn - \omega t)}$$

in the case where M is a tridiagonal matrix with constant coefficients (nearest-neighbor interactions). This case is now known as the discrete nonlinear Schrödinger (DNLS) equation. Note that the "carrier wave" speed ω/k is different from the "envelope" speed c. He found branches of localized solutions to high accuracy, but the existence of such solutions is still an open question.

The stability criteria for stationary solutions of the DST equation were studied by Carr & Eilbeck (1985). An important step is to consider perturbations in a frame rotating with the stationary solutions

$$U_n(t) = [y_n + \varepsilon_n(t)]\,\mathrm{e}^{\mathrm{i}\omega t}.$$

Once this trick is carried out, the study of linear stability reduces to an algebraic eigenvalue problem, at least when f is finite, which greatly simplifies the problem. One noteworthy point is that the resulting stability matrix is not self-adjoint, so branches can change stability at points away from bifurcation points.

On a finite lattice, modeling vibrational modes on small molecules, the stationary solutions of the DST equation can often be found exactly. For $f = 3$ and $M_{ij} = 1$ if $i \neq j$ (modeling NH_3 stretching oscillations in ammonia), we obtain the bifurcation diagram shown in Figure 2. Details are discussed in Eilbeck et al. (1985). Note the change of stability of the $\uparrow\downarrow 0$ branch at two places, and the local-mode branch, where the vibrational energy is localized on a single degree of freedom.

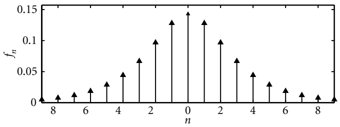

Figure 1. Stationary Breather on a DST (DNLS) lattice.

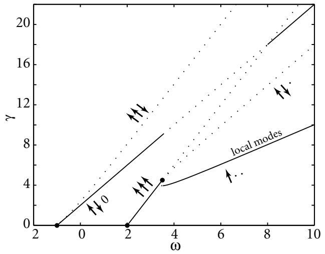

Figure 2. Stationary solutions on a DST lattice with $f = 3$. Solid (dotted) lines indicate stable (unstable) solutions.

In the large f case, in an application of the DST formulation to the dynamics of a globular protein, Feddersen considered interactions among CO stretch oscillations in adenylate kinase, which comprises 194 amino acids ($f = 194$) (Feddersen, 1991a). Since the structure of this enzyme has been determined by X-ray analysis, the $f(f-1)/2 = 18721$ off-diagonal elements of the dispersion matrix (M) were calculated from Maxwell's equations. Also, diagonal elements were selected from a random distribution, and the degree of localization of a particular solution was defined by evaluating the quotient $\sum |u_j|^4 / \sum |u_j|^2$.

In this study at experimentally reasonable levels of nonlinearity (γ), stable localized solutions were observed near some but not all of the amino acids. Interestingly, this anharmonic localization was observed to be distinctly different from Anderson localization, a property of randomly interacting linear systems. Thus, none of the stationary states, that were observed to be highly localized at large γ, remained so as γ was made small. Also, none of the states, that were localized at $\gamma = 0$ (i.e., Anderson localized), remained so as γ was increased to a physically reasonable level.

Other studies and applications of the DST equation include models for Bose–Einstein condensates and coupled optical fiber arrays. On the theoretical side, work has been carried out in the study of nonstationary and chaotic solutions on both small and large lattices. A list of citations to Eilbeck et al. (1985) is presently maintained at http://www.ma.hw.ac.uk/~chris/dst/.

ALWYN SCOTT AND CHRIS EILBECK

See also **Anderson localization; Discrete breathers; Discrete nonlinear Schrödinger equations; Local modes in molecular crystals; Local modes in molecules; Rotating wave approximation**

Further Reading

Careri, G., Buontempo, U., Galluzzi, F., Scott, A.C., Gratton, E. & Shyamsunder, E. 1984. Spectroscopic evidence for Davydov-like solitons in acetanilide. *Physical Review* B 30: 4689–4702

Carr, J. & Eilbeck, J.C. 1985. Stability of stationary solutions of the discrete self-trapping equation. *Physics Letters* A 109: 201–204

Eilbeck, J.C., Lomdahl, P.S. & Scott, A.C. 1984. Soliton structure in crystalline acetanilide. *Physical Review* B, 30: 4703–4712

Eilbeck, J.C., Lomdahl, P.S. & Scott, A.C. 1985. The discrete self-trapping equation. *Physica* D 16: 318–338

Ellis, J.W. 1929. Molecular absorption spectra of liquids below 3 μ. *Transactions of the Faraday Society*, 25: 888–897

Feddersen, H. 1991a. Localization of vibrational energy in globular protein. *Physics Letters* A, 154: 391–395

Feddersen, H. 1991b. Solitary wave solutions to the discrete nonlinear Schrödinger equation. In *Nonlinear Coherent Structures in Physics and Biology*, edited by M. Remoissenet & M. Peyrard, Berlin and New York: Springer, 159–167

Scott, A.C., Eilbeck, J.C. & Gilhøj, H. 1994. Quantum lattice solitons. *Physica* D, 78: 194–213

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*. 2nd edition, Oxford and New York: Oxford University Press

DISLOCATIONS IN CRYSTALS

Dislocations in crystals are linear topological defects near which the regular crystalline atomic arrangement is broken. The dislocation as a structural defect is characterized by the following peculiarity: the regular crystal lattice structure is greatly distorted only in the near vicinity of an isolated line (the dislocation axis), and the region of irregular atomic arrangement has transverse dimensions of the order of the interatomic distance. Nevertheless, deformation occurs even far from the dislocation core, and this is a consequence of a topological property of the dislocation. The deformation at a distance from the dislocation axis may be seen by passing along a closed contour around the dislocation core and considering the displacement vector of each crystalline site from its position in an ideal lattice. Upon calculating the total increment of this displacement vector at the end of the pass, one sees that the total atomic displacement is nonzero and equals one atomic translation period.

Among the many microscopic models of dislocations, the simplest takes the dislocation to be the edge of an extra half-plane present in the crystal lattice. In the conventional atomic scheme of this model (Figure 1), the trace of the half-plane coincides with the upper semiaxis y, and its edge coincides with the z-axis; thus, one has an "edge" dislocation. If one surrounds the dislocation axis with a tube with radius of the order of several interatomic distances, the crystal outside this tube may be regarded as ideal and subject only to elastic deformation (crystal planes are connected to one another almost regularly), and inside the tube the atoms are considerably displaced relative to their equilibrium positions and form the "dislocation core." The atoms of the dislocation core are distributed over the contour of the shaded pentagon in Figure 1. The closed pass is the external contour starting from point 1 and finishing at point 2. The vector connecting atoms 1 and 2 is denoted by $\boldsymbol{b}$ and called the "Burgers vector." Possible values of the Burgers vectors in a crystal are determined by its crystallographic structure and correspond, as a rule, to a small number of certain directions in a crystal. The dislocation lines are arranged arbitrarily, although their arrangement is limited by a set of definite crystallographic planes. The dislocation line cannot end inside a crystal. It must either leave the crystal with each end on the crystal surface or form a dislocation loop.

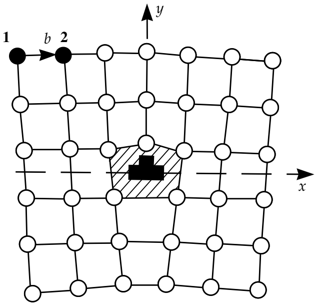

Figure 1. Schematic atomic arrangement in the vicinity of an edge dislocation.

The main topological property of the dislocation implies that a dislocation in a crystal is a specific line D having the following general property. After a circuit around the closed contour L enclosing the line D (see Figure 2), the elastic displacement vector $\boldsymbol{u}$ changes by a certain finite increment $\boldsymbol{b}$ equal to one of the lattice periods. This property can be written as

$$\oint_L \mathrm{d}u_i = \oint_L \frac{\partial u_i}{\partial x_k}\mathrm{d}x_k = -b_i, \qquad (1)$$

assuming that the direction of the circuit is related to a chosen direction of the tangent vector $\boldsymbol{\tau}$ to the dislocation line. The dislocation line is in this case a line of singular points of the elastic fields.

Macroscopic considerations (based on elasticity theory) allow one to find a concentration of strains and stresses near the dislocation. The elastic field created by a dislocation leads to its interaction with

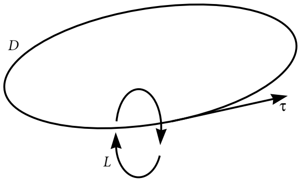

Figure 2. Mutual coordination of the vector $\boldsymbol{\tau}$ and circuit around the contour L.

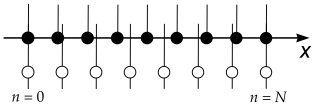

Figure 3. Edge dislocation in the Frenkel–Kontorova model.

other dislocations and other types of crystal defects. In particular, the dislocation can accumulate point defects near its core. If such defects are color centers, the dislocation line can be decorated (*See* **Color centers**).

Displacements of the dislocation line in the plane determined by the vectors $\boldsymbol{b}$ and $\boldsymbol{\tau}$ (called a "slip plane") have no effect on the crystal continuity. Therefore, a comparatively easy mechanical motion of the dislocation is possible in principle in this plane. But any displacement of the dislocation produces some plastic deformation, so moving dislocations are carriers of crystal plasticity.

Macroscopic dislocation theory can describe a contribution of the dislocations to mechanical properties of real crystals and proposes a physical explanation of the crystal plasticity. However, the macroscopic theory cannot propose a structure of the dislocation core and explain possible mechanisms of the dislocation motion. The Soviet scientists Yakob Frenkel and Tatiana Kontorova were the first to present a simple 1-dimensional microscopic dislocation model (Frenkel & Kontorova, 1938). To formulate their model, it is convenient to analyze the following picture. The atomic chain (a set of black circles in Figure 3) is an edge series of one half of a plane crystal ($y > 0$, Figure 1) displaced in a certain way with respect to another half of a crystal (a substrate at $y < 0$). The influence of the nondisplaced half of the crystal on the atoms distributed along the x-axis can be qualitatively described assuming that those atoms are in a given external periodic field whose period coincides with the interatomic distance on the substrate a. The chain energy is then determined not only by a relative displacement of neighboring atoms but also by an absolute displacement of separate atoms u_n (n is the atomic number) in the external field. If the external periodical field has a simple sine-shaped dependence, the equation of chain motion has the form

$$m\frac{\mathrm{d}^2 u_n}{\mathrm{d}t^2} = \alpha(u_{n+1} + u_{n-1} - 2u_n) \\ -U_0 \sin(2\pi u_n/a)\,, \tag{2}$$

where m is the atom mass and α and U_0 are constant parameters. The displacements of the atoms in the chain are performed in such a way that $N+1$ atoms are situated over N crystal sites on the substrate. This atom distribution corresponds to the following boundary conditions for the displacements:

$$u_n = a \quad \text{at} \quad n = -\infty\,, \qquad u_n = 0 \quad \text{at} \quad n = \infty. \tag{3}$$

Equation (2) and the boundary conditions (3) are basic in the Frenkel–Kontorova model.

In a long-wavelength approximation, the function of the discrete number u_n is replaced by a continuous function of the coordinate $u(x)$ where $x = na$, and Equation (2) transforms into the sine-Gordon equation (SG) equation

$$\frac{\partial^2 w}{\partial t^2} = s^2\frac{\partial^2 w}{\partial x^2} - \omega_0^2 \sin w, \quad w = 2\pi u/a, \tag{4}$$

where $s^2 = a^2\alpha/m$ and $\omega_0^2 = U_0/m$.

The SG equation (4) possesses the following (kink-soliton) solution satisfying the boundary conditions (3): $u(x) = \tan^{-1}\exp[-(x - x_0/l)]$, where $l = s/\omega_0$ and $x_0 = \text{constant}$. This kink describes the distribution of atoms in the core of the Frenkel–Kontorova dislocation, the parameter l determines the dislocation semiwidth, and x_0 is the coordinate of the dislocation center. Later Peierls (1940) proposed a semimicroscopic model of a straight dislocation line (directed along the z-axis in Figure 1) in a 3-d isotropic elastic medium. In this model, it was assumed that strains around the dislocation are described by the theory of elasticity; however, the interaction energy of both semi-spaces above (+) and under (−) the slide plane depends periodically on the relative displacements $v(x) = w(x)^+ - w(x)^-$ (the x-axis is a trace of the slide plane) and is described by a sine-shaped periodic function of $v(x)$. The main equation of Peierls's theory has the following form:

$$a\int_{-\infty}^{\infty}\left(\frac{\mathrm{d}v}{\mathrm{d}x}\right)\frac{\mathrm{d}\xi}{x - \xi} + M \quad \sin v = 0\,, \tag{5}$$

where the integral means its principal value and M is some dimensionless combination of elastic modules. Equation (5) has a soliton solution very close to the Frenkel–Kontorova dislocation $w(x) = \pi + 2\tan^{-1}[(x - x_0)/l]$ where $l = \pi a/M$.

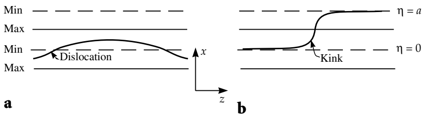

Figure 4. Two types of dislocation motion in the Peierls potential field. (a) Dislocation oscillates in its own trough. (b) The dislocation forms a kink moving along the z-axis.

Both the chain energy in the 1-d case and crystal energy in the 3-d case (calculated for the Frenkel–Kontorova and Peierls models, respectively) do not depend on the coordinate of the dislocation center x_0. Thus, the dislocation can glide, changing its location on the slide plane without action of any external field. This fact is a result of using the continuous approximation. Taking into account discreteness of the crystal and following the Peierls theory, Nabarro (1947) showed that a moving dislocation is influenced by some crystal field periodically depending on the dislocation center x_0 (later called the "Peierls relief"). Therefore, the dislocation can start moving only if an external stress acting on it exceeds a certain value called the "Peierls barrier."

Many dynamic problems in dislocation theory can be analyzed using the so-called "string model." This model treats the dislocation line as a heavy string under tension lying on a corrugated surface. An effective mass per unit length of the dislocation m and its line tension T_D are of a field origin and are associated with the inertia and energy of dislocation dynamic elastic field (Kosevich, 1964). The corrugated surface describes the Peierls relief. Troughs of this surface correspond to the potential minima of the slide plane occupied by a straight-line dislocation in equilibrium (see Figure 4). If the z-axis is directed along the equilibrium position of the dislocation and its transverse displacements η go along the x-axis, then the equation of motion of the dislocation has the form (Seeger, 1956):

$$m\frac{\partial^2\eta}{\partial t^2} - T_D\frac{\partial^2\eta}{\partial x^2} + b\sigma_{\mathrm{p}}\sin(2\pi\eta/a) = b\sigma\,, \qquad (6)$$

where σ_{p} is the Peierls stress and σ is the applied stress. In the case $\sigma = 0$, this equation is equivalent to Equation (4), proposed by Frenkel and Kontorova. Equation (6) describes both small amplitude vibrations of the dislocation effected by an external oscillation force and the so-called kink-mechanism of the transverse displacements of the dislocation along the x-axis effected by a strong stationary driving force.

ARNOLD KOSEVICH

See also **Color centers; Frenkel–Kontorova model; Peierls barrier; Sine-Gordon equation; Topological defects**

Further Reading

Frenkel, J. & Kontorova, T. 1938. On the theory of plastic deformation and twinning. *Physikalische Zeitschrift der Sowjetunion*, 13: 1

Kosevich, A.M. 1964. Dynamical theory of dislocations. *Uspekhi Fizicheskikh Nauk*, 84: 579 (in Russian); translated in 1965. *Soviet Physics, Uspekhi*, 7: 837

Kosevich, A.M. 1979. Crystal dislocations and the theory of elasticity, in *Dislocations in Solids*, edited by F.R.N. Nabarro Amsterdam: North Holland, p. 33

Nabarro, F.R.N. 1947. Dislocations in a simple cubic lattice *Proceedings of Physical Society, London*, A59: 256

Nabarro, F.R.N. 1967. *Theory of Crystal Dislocations*, Oxford: Clarendon Press

Peierls, R.E. 1940. The size of a dislocation. *Proceedings of Physical Society, London*, A52: 34

Seeger, A. 1956. On the theory of the low-temperature internal friction peak observed in metals. *Philosophical Magazine*, 1: 651

DISPERSION MANAGEMENT

The propagation of optical pulses in fiber lines is usually limited by dispersion and nonlinearity. Dispersion causes a broadening of the pulses whereas nonlinearity is the reason for four-wave mixing (FWM). A balance of dispersion and nonlinearity may produce the so-called optical soliton. Dispersion management (manipulations of the chromatic dispersion along the line) is an attractive technique that allows enhancement of the performance of fiber communication links both for soliton and nonsoliton transmission. Dispersion-managed solitons can be viewed as novel kinds of information carriers with many attractive features that will lead to further improvement of the transmission capacity of fiber links. The starting point of all considerations in this area is the cubic nonlinear Schrödinger (NLS) equation for an optical mode in a cylindrical fiber with (constant) dispersion and Kerr nonlinearity (Newell & Maloney, 1992):

$$\mathrm{i}\,\frac{\partial u}{\partial z} + \frac{1}{2}\,\frac{\partial^2 u}{\partial t^2} + |u|^2 u = 0. \qquad (1)$$

This equation is written in dimensionless form (u is the normalized intensity, z the normalized propagation distance, and t is the normalized time); for more details, see, for example, Hasegawa & Kodama (1995).

From FWM calculations (Agrawal, 1995), one knows that the FWM efficiency depends on the phase mismatch and thereby on the dispersion coefficient, D. The larger the phase mismatch (proportional to dispersion), the smaller the efficiency of FWM. On the other hand, small D values are necessary for high bit-rates B. One reason is the so-called Gordon–Haus effect (Gordon & Haus, 1986). Solitons adjust themselves to external perturbations, and thereby they can be considered as quite robust. However, this adjustment also has disadvantages. Any re-shaping, for example, in the presence of amplifier noise, also has consequences

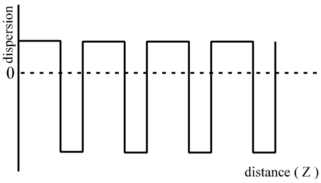

Figure 1. Schematic plot of the dispersion coefficient that varies in space when dispersion management is applied.

for the velocities. Changing velocities cause a timing jitter, and the latter limits the maximal possible bit-rate. Note that the general soliton solution of Equation (1)

$$u_s = A \operatorname{sech}[A(t + \Omega z) - q_0]\, \mathrm{e}^{-\mathrm{i}\Omega t-(\mathrm{i}/2)(A^2-\Omega^2)z+\mathrm{i}\phi_0} \tag{2}$$

is moving with velocity $v = -1/\Omega$. The detailed calculation (Gordon & Haus, 1986) gives an upper limit for the bit-rate B and the length L of the system in the form

$$(BL)^3 \leq \text{upper limit} \sim \frac{1}{D}. \tag{3}$$

One clearly recognizes that low dispersion favors high bit-rates over large distances. On the other hand, as was mentioned already, low dispersion is dangerous with respect to FWM. Dispersion management is an elegant way out of this dilemma.

Principle of Dispersion Management

The idea of using a dispersion-compensating fiber (DCF) to overcome the dispersion of the standard mono-mode fiber (SMF) was proposed in 1980 (Lin et al., 1980). The simplest optical-pulse equalizing system consists of a transmission fiber (SMF or dispersion-shifted fiber (DSF)) and an equalizer fiber with the opposite dispersion (DCF to compensate SMF or, for instance, SMF to compensate DSF with normal dispersion). The incorporation of a fiber with normal dispersion reduces (or, in the ideal case, eliminates) the total dispersion of the fiber span between two amplifiers. Low average dispersion is good for reducing the Gordon–Haus effect while high local dispersion reduces the FWM efficiency. In Figure 1, we indicate the principle of dispersion management by a schematic drawing.

Dispersion management means that the coefficient in front of the second time derivative becomes Z-dependent (we change notation in order to allow for different normalization parameters), that is, the simplest form of the dispersion-managed nonlinear Schrödinger equation becomes (Kodama, 1998)

$$\mathrm{i}\,\frac{\partial q}{\partial Z} + \frac{1}{2}\,\mathrm{d}(Z)\,\frac{\partial^2 q}{\partial T^2} + |q|^2 q = 0. \tag{4}$$

A short remark: if $\mathrm{d}(Z)$ would be $\neq 0$ everywhere in the fiber, the variation of $\mathrm{d}(Z)$ can be eliminated by using a new coordinate Z' via

$$\frac{\mathrm{d}Z'}{\mathrm{d}Z} = d(Z). \tag{5}$$

Then a Z-dependent coefficient would appear in front of the nonlinear term. Again, we would end up with an equation with a rapidly varying coefficient. However, characteristic for nearly all practical dispersion management arrangements is the vanishing of $d(z)$ at certain positions z.

Weak Dispersion Management

For weak dispersion management of small-amplitude pulses, one can start with the equation

$$\mathrm{i}\frac{\partial u}{\partial z} + d(z)\frac{\partial^2 u}{\partial t^2} + \varepsilon|u|^2 u = 0. \tag{6}$$

Here,

$$\begin{aligned} d(z) &= \langle d\rangle + \tilde{d} \quad \text{for } 0 \leq z \leq l, \\ d(z) &= \langle d\rangle - \tilde{d} \quad \text{for } l \leq z \leq 2l, \end{aligned} \tag{7}$$

and so on, is the z-dependent dispersion. The z-coordinate is made dimensionless with the dispersion length of the local dispersion, that is, $\tilde{d} \sim O(1)$. The averaged dispersion $\langle d\rangle$ is assumed to be small, and we introduce the smallness parameter $\varepsilon \ll 1$ via $\langle d\rangle \approx \varepsilon$. Nonlinearity will also be scaled by ε. The strength of the dispersion management is characterized by

$$R(z) = \int_0^z (d - \langle d\rangle)\,\mathrm{d}s. \tag{8}$$

Weak dispersion management means $R \sim O(\varepsilon^{1/2})$; that is, R is assumed to be small. From the definition $R \sim O(\tilde{d}\cdot l)$ holds. Therefore, the length l (of the parts with constant dispersion) is also small compared with the dispersion length, $l \sim O(\varepsilon^{1/2})$.

One can use (at least three different) averaging methods to calculate the average behavior of pulses over large distances. The first one is a direct method, starting with an expansion. The second one uses the Lie-transform technique; it is, therefore, much more systematic than the direct method. The third method makes use of a Bogolyubov transformation, which is very elegant in practice, but the ansatz is not as straightforward as the Lie transformation. The ansatz has to be specifically chosen for each problem. In the direct method (Kivshar & Spatschek, 1995; Yang et al., 1997), one expands u in orders of $\varepsilon^{1/2}$ and applies a multiple scale technique,

$$\begin{aligned} u &= U + u_1 + u_2 + \cdots, \\ \frac{\partial}{\partial z} &= \frac{\partial}{\partial \xi} + \frac{\partial}{\partial z_0} + \frac{\partial}{\partial z_1} + \cdots \end{aligned} \tag{9}$$

with $U = U(z_0, z_1, z_2, \ldots)$ and $u_k = u_k(\xi, U)$. One may use the scaling $\xi \sim O(\varepsilon^{-1/2})$, $z_k \sim O(\varepsilon^{k/2})$, $u_k \sim O(\varepsilon^{k/2})$, where $\langle u \rangle = U$, $\langle u_k \rangle = 0$.

The Lie transformation technique was originally developed for ordinary differential equations but can also be used for a partial differential equation (Neyfeh, 1973). To apply the Lie transformation one rearranges Equation (6) into the form

$$\begin{aligned}\frac{\partial u}{\partial z} &= \mathrm{i}\,\langle d\rangle\, u_{tt} + \mathrm{i}\varepsilon|u|^2 u + \mathrm{i}\tilde{d}u_{tt}\\ &= X_0 + \tilde{d}X_{0D} = X[u, u^*]. \qquad (10)\end{aligned}$$

X depends on the infinite set of variables $(u, u_t, u_{tt}, \ldots, u^*, u_t^*, u_{tt}^*, \ldots)$, indicated by $X[u, u^*]$. Obviously $X_0 \sim O(\varepsilon)$, $X_{0D} \sim O(1)$. The transformation

$$\begin{aligned}u &= \mathrm{e}^{\vec{\phi}\nabla} v = v + \phi + \frac{1}{2!}\left(\vec{\phi}\cdot\nabla\right)\phi\\ &\quad + \frac{1}{3!}\left((\vec{\phi}\cdot\nabla)\phi\cdot\nabla\right)\phi + \cdots \qquad (11)\end{aligned}$$

transforms Equation (10) into

$$\frac{\partial v}{\partial z} = Y[v, v^*]. \qquad (12)$$

The right-hand side can, in principle, be derived appropriately in order to identify v with the averaged intensity. Here, $\vec{\phi} = (\phi, \phi_t, \ldots)$ and $\nabla = (\frac{\partial}{\partial v}, \frac{\partial}{\partial v_t}, \ldots)$ are infinite-dimensional vectors.

In the third method (Arnol'd, 1988), one uses a Floquet–Lyapunov transformation

$$u(z, t) = \mathrm{e}^{\mathrm{i}R(z)\partial_t^2} B(z, t) \qquad (13)$$

as well as the Bogolyubov–Krylov transformation

$$\begin{aligned}B &= v + \mathrm{i}f|v|^2 v + (g + p)N_1(v)\\ &\quad + \mathrm{i}hN_2(v) + q|v|^4 v + O(\varepsilon^{5/2}), \qquad (14)\end{aligned}$$

with coefficients to be determined appropriately during the calculation. All three methods lead to the same result. Over fairly large distances, the averaged solution is still a fundamental soliton, that is, robust. This follows from the fact that *to lowest order*, the basic equation for the averaged pulse intensity is the *integrable* cubic nonlinear Schrödinger equation

$$\mathrm{i}v_z + \langle d\rangle\, v_{tt} + \varepsilon|v|^2 v = \frac{1}{2}\varepsilon\left\langle R^2\right\rangle N_2(v). \qquad (15)$$

The correction term on the right-hand side contains a polynomial N_2 in v, v_t, $v_{tt}, \cdots$. One should bear in mind, on the other hand, that on the short scale (length $2l$ of one of the periodic elements), the exact (non-averaged) solution is oscillating (breathing) with the frequency following from the periodicity in the dispersion map. Besides the scaling $\tilde{d} \sim O(1)$ and $R \sim O(\varepsilon^{1/2})$, other scalings are possible and tractable by the methods mentioned above, as long as R is small (weak dispersion management).

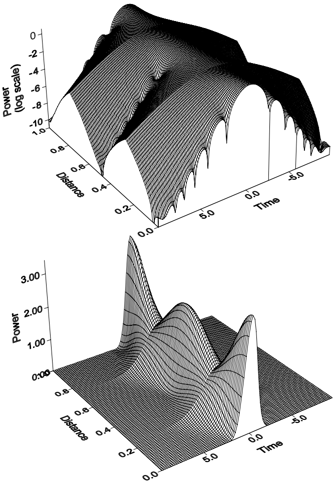

Figure 2. Plot of a DM soliton. The lower part shows the breathing over one element of the dispersion map. The upper graph shows the fine structure on a logarithmic axis, indicating the characteristic deviations from the fundamental soliton.

Strongly Dispersion-Managed Solitons

Recently (Nijhof et al., 1998), the strong dispersion management (R is no longer small) became of increasing interest. The chirp (pulse phase has a quadratic time-dependence) is a new and one of the most important features of the (strongly dispersion managed) DM soliton (recall that the conventional soliton is unchirped). In addition, the DM soliton is not anymore of the sech-type. The DM soliton consists of a self-similar energy-containing core surrounded by oscillating tails. Such tails manifest themselves as non-self-similar modulations of the soliton profile during the compensation period, although their amplitudes are rather small compared with the main peak (see Figure 2).

The most surprising feature of the DM soliton is that it can propagate stably along the line with zero or even with normal average dispersion (in contrast to the fundamental soliton that propagates stably only in the anomalous dispersion region). These observations indicate that an average model describing the evolution of the breathing pulse differs from the nonlinear Schrödinger equation. In other words, strong dispersion management imposes such a strong perturbation that a carrier pulse in this case is no longer the NLS soliton. For more details, see Nijhof et al. (1998).

Experimental Results

One can estimate the potential of dispersion management from successful experiments, for example, Carter et al. (1999).

KARL SPATSCHEK

See also **Nonlinear optics; Optical fiber communications**

Further Reading

Agrawal, G.P. 1995. *Nonlinear Fiber Optics*, San Diego: Academic Press

Arnol'd, V.I. 1988. *Geometrical Methods in the Theory of Ordinary Differential Equations*, New York: Springer

Carter, G.M., Grigoryan, V.S., Mu, R.-M., Menyuk, C.R., Sinha, P., Carruthers, T.F., Dennis, M.L. & Duling, I.N. 1999. Transmission of dispersion-managed solitons at 20 Gbit/s over 20 000 km. *Electronics Letters*, 35(3): 233

Gordon, J.P. & Haus, H.A. 1986. Random walk of coherently amplified solitons in optical fiber transmission. *Optics Letters*, 11: 665

Hasegawa, A. & Kodama, Y. 1995. *Solitons in Optical Communications*, Oxford: Clarendon Press

Kivshar, Yu.S. & Spatschek, K.H. 1995. Nonlinear dynamics and solitons in the presence of rapidly varying periodic perturbations. *Chaos, Solitons, and Fractals*, 5: 2551–2569

Kodama, Y. 1998. Nonlinear pulse propagation in dispersion managed system. *Physica* D, 123: 255–266

Lin, C., Kogelnik, H. & Cohen, L.G. 1980. Optical pulse equalization and low dispersion transmission in single-mode fibers in the 1.3–1.7 mm spectral region. *Optics Letters*, 5: 476–478

Newell, A.C. & Moloney, J.V. 1992. *Nonlinear Optics*, Redwood City, CA: Addison-Wesley

Neyfeh, A.H. 1973. *Perturbation Methods*, New York: Wiley

Nijhof, J., Doran, N., Forysiak, W. & Bernston, A. 1998. Energy enhancement of dispersion-managed solitons and WDM. *Electronics Letters*, 34: 481–482

Yang, T.-S., Golovchenko, A., Pilipetskii, A.N. & Menyuk, C.R. 1997. Dispersion-managed soliton interactions in optical fibers. *Optics Letters*, 22: 793–795

DISPERSION RELATIONS

Dispersion relations (DRs) are associated with wave equations to characterize the nature of their temporal and spatial evolution. For a linear wave equation in one dimension, the DR expresses the wavenumber ($k = 2\pi/\lambda$) as a function of the frequency ($\omega = 2\pi/T$), or vice-versa, where λ and T are, respectively, the wavelength and temporal period of an elementary periodic solution. Through the superposition principle, the behaviors of all solutions of linear equations are determined by the DR. In nonlinear cases, DRs also depend on wave amplitude and can be used in perturbative analyses of quasiharmonic systems. For more than one spatial dimension, the wave number is a vector.

Dispersion Relations for Linear Equations

Fourier analysis establishes a one-to-one correspondence between DRs and equations in such a way that we can obtain one from the other. Consider as an example, the Klein–Gordon equation,

$$u_{tt} - u_{xx} + m^2 u = 0, \tag{1}$$

where m is a constant and the subscripts indicate partial derivatives. Substituting a plane-wave solution of the form $u(t, x) = a\mathrm{e}^{\mathrm{i}(\omega t - kx)}$ into (1), we obtain the DR

$$\omega^2 = m^2 + k^2. \tag{2}$$

For linear equations, the phase velocity $v_\varphi(k)$ and the group velocity $v_\mathrm{g}(k)$ are defined as

$$v_\varphi(k) = \frac{\omega(k)}{k}, \quad v_\mathrm{g}(k) = \frac{\mathrm{d}\omega(k)}{\mathrm{d}k}. \tag{3}$$

The phase velocity of a wave number k is the speed of propagation of the corresponding harmonic wave component or mode. The group velocity corresponds to the propagation speed of a wave packet, and thus is the speed of energy transmission for the system (Whitham, 1974). This distinction is especially important in the propagation of wave trains, where signals travel with the group velocity.

An equation (or the wave that is represented by its solution) is said to be *dispersive* if $v_\varphi(k)$ is not a constant (or equivalently if $v_\varphi \neq v_\mathrm{g}$ for all k). In such cases, the initial profile of the wave is not conserved. For the Klein–Gordon equation we have

$$v_\varphi(k) = \pm\frac{\sqrt{m^2 + k^2}}{k}, \quad v_\mathrm{g}(k) = \pm\frac{k}{\sqrt{m^2 + k^2}}, \tag{4}$$

and the equation is dispersive except in the special case $m = 0$ (which corresponds to the wave equation).

The DR also plays an important role in simulations of Equation (1) by finite differences. The usual (von Neumann) stability conditions are equivalent to requiring that the phase velocity in the mesh be greater or equal to the phase velocity of the equation (Ritchmyer & Morton, 1967). Other aspects that mark deviations of numerical solutions from continuous solutions are governed by discrete DRs (Trefethen, 1982).

Amplitude-Dependent Dispersion Relations

For nonlinear equations, solutions cannot, in general, be expressed as a superposition of plane waves and thus, in principle, there is not dispersion relation. There are nevertheless two possible applications of the concept. The first is in the small-amplitude regime where an approximate DR can be established by neglecting the higher order terms, and the second is when plane waves are particular solutions of the nonlinear equation. In both cases, the DRs are amplitude dependent. We illustrate both cases with a nonlinear Klein–Gordon equation and the cubic Schrödinger equation.

The nonlinear Klein–Gordon equation with cubic term

$$u_{tt} - u_{xx} + m^2 u + g u^3 = 0 \qquad (5)$$

does not have exact plane wave solutions. (The case $m = 1$, $g = -1/3!$ corresponds for instance to the first two terms expansion of the sine-Gordon equation.) Nevertheless, we may assume a term of the form $a\cos(\omega t - kx)$ in the solution. Substituting and retaining only the lower-order terms, we obtain the approximate DR

$$-\omega^2 + k^2 + m^2 + \frac{3}{4} g a^2 + \cdots = 0 \qquad (6)$$

provided $m^2 \gg ga^2$. The presence of coupled higher-order terms together with the approximate DR gives a picture of how such modes get excited and evolve.

The previous analysis can also be applied, for instance, to the cubic nonlinear Schrödinger equation

$$\mathrm{i}\phi_t - \phi_{xx} + |\phi|^2\phi = 0, \qquad (7)$$

but this equation is special in the sense that it has particular solutions in the form of plane waves $\phi = ae^{\mathrm{i}(\omega t - kx)}$. Although the general solution cannot be expressed as a superposition of such plane waves, we have an amplitude-dependent DR of the form

$$\omega = k^2 + a^2. \qquad (8)$$

In this case, wave dispersion can be compensated by nonlinearity and solitons may appear.

L. Vázquez and S. Jiménez

See also **Modulated waves; Wave packets, linear and nonlinear**

Further Reading

Ritchmyer, R.D. & Morton, K.W. 1967. *Difference Methods for Initial-Value Problems*, New York: Wiley-Interscience

Trefethen, L.N. 1982. Group velocity in finite difference schemes. *SIAM Review*, 24: 113–136

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley-Interscience

DISTRIBUTED OSCILLATORS

Two modes of oscillation sharing a common source of energy may have a stationary solution of equal mode amplitudes, but this solution may be unstable for the following reason. One of the modes will inevitably gain a bit of amplitude (due to noise), and if it then consumes more than its share of the available power, it will grow further, causing the other mode to decrease. Interestingly, this effect depends upon the spatial dimension of the oscillating system. Related dynamics arise in diverse areas of engineering and science, including flip-flop circuits, multimode lasers, interacting biological species, and the neocortex of a human brain.

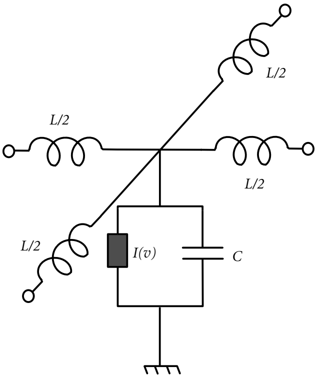

Figure 1. A unit cell of a two-dimensional oscillator array.

As a model for mode interactions, consider a two-dimensional array of identical oscillators, each interacting only with its nearest neighbors. Power flow among the modes of such an array can be studied using the method of harmonic balance, introduced by Balthasar van der Pol in 1934 (van der Pol, 1934) and developed by applied mathematicians and engineers in the Soviet Union during the 1930s and 1940s (Andronov et al., 1966; Kryloff & Bogoliuboff, 1947).

As shown in Figure 1, a unit cell of the array consists of a nonlinear conductance $I(v)$ in parallel with a capacitance C, connecting a lattice point to a ground plane. Each lattice point is attached to its four nearest neighbors through an inductance L. The nonlinear conductance is represented as

$$I(v) = -Gv\left(1 - 4v^2/3V_0^2\right), \qquad (1)$$

so for $-V_0/2 < v < +V_0/2$ the differential conductance $(\mathrm{d}I/\mathrm{d}v)$ is negative.

Assume a large, square $(N \times N)$ lattice of these unit cells with zero-voltage (short circuit) boundary conditions on the edges, as in Figure 2. In a zero-order approximation, the nonlinear conductance can be neglected, and in a continuum approximation, this linear, lossless system supports an arbitrary number (say n) modes. Thus, the total voltage is

$$\begin{aligned} v(x, y, t) = {} & \mathcal{V}_1(x, y)\cos\theta_1 + \mathcal{V}_2(x, y) \\ & \times \cos\theta_2 + \cdots + \mathcal{V}_n(x, y)\cos\theta_n, \end{aligned} \qquad (2)$$

which depends on space (x and y) and time (t) where

$$\mathcal{V}_i(x, y) \equiv V_i \cos k_{xi}x \cos k_{yi}y \quad \text{and} \quad \theta_i \equiv \omega_i t + \phi_i.$$

Noting that the energy in the ith mode is related to its amplitude by $U_i = (N^2C/8)V_i^2$ and averaging the

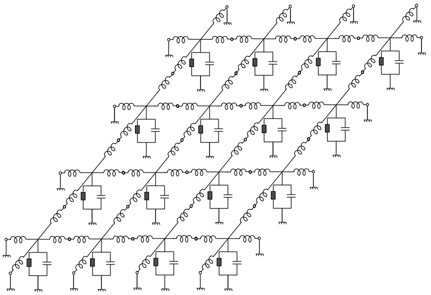

Figure 2. A 4 × 4 array ($N^2 = 16$) of the unit cells in Figure 1.

product of current (given by Equation (1)) and voltage (given by Equation (2)) over space and time gives the rate of change in mode energies as functions of mode energies (Scott, 2003).

The rate of change of energy (or power) into the first mode is

$$\frac{\mathrm{d}U_1}{\mathrm{d}\tau} = U_1\left[1-\alpha\left(\frac{9}{8}U_1+U_2+\cdots+U_n\right)\right], \quad (3)$$

with $\tau \equiv Gt/C$ and $\alpha \equiv 4/(N^2CV_0^2)$. For n excited modes, there is a set of n equations—each similar to Equation (3) but with the indices appropriately altered—for the rates of change of the mode energies as functions of those energies. These nonlinear, autonomous equations have the same form as those introduced by Vito Volterra to describe the interaction of biological species competing for the same food supply (Volterra, 1937), and they are similar to a formulation suggested by Peter Greene (Greene, 1962) to describe interactions between assemblies of neurons in the brain.

Generalizing to a system of d-space dimensions, equations of interacting mode energies become

$$\begin{aligned}\frac{\mathrm{d}U_1}{\mathrm{d}\tau} &= U_1\left[1-\alpha\left(KU_1+U_2+\cdots+U_n\right)\right],\\ &\vdots \\ \frac{\mathrm{d}U_n}{\mathrm{d}\tau} &= U_n\left[1-\alpha\left(U_1+U_2+\cdots+KU_n\right)\right],\end{aligned} \quad (4)$$

where

$$K = 3^d/2^{d+1}.$$

For $K > 1$ (implying $d \geq 2$), analysis of Equations (4) indicates multimode stability. In other words, several modes can stably exist for two or more spatial dimensions. For one spatial dimension, on the other hand, two modes of equal amplitude are unstable, and only a single mode can be stably established. This observation is important in the design of semiconductor lasers, where large transverse dimensions are introduced to increase the power output. Unfortunately, such a design allows several modes to oscillate together, thereby decreasing the coherence and spectral purity of the output beam.

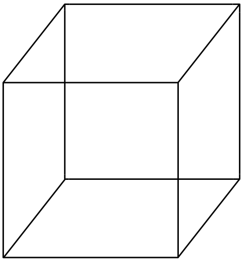

Figure 3. A Necker cube.

Multimode oscillator arrays have been realized using semiconductor tunnel diodes (Scott, 1971), superconductor tunnel diodes (Hoel, et al., 1972), and integrated circuits (Aumann, 1974). In these experiments, a variety of stable multimode oscillations have been observed, some quasiperiodic and others periodic, indicating mode locking. In a manner qualitatively similar to that proposed by Greene in his model of the brain's neocortex (Greene, 1962), the oscillator array can be induced to switch from one stable multi-mode configuration to another.

At the level of subjective perception, similar switchings in the brain are observed as one stares at Figure 3. Constructed by Louis Albert Necker (a Swiss geologist) in the mid-1800s, this image seems to jump from one metastable orientation to another, like the flip-flop circuit of a computer engineer. Defining order parameters (ξ_1 and ξ_2) to represent neural activities corresponding to the two perceptions of the Necker cube, Hermann Haken has recently suggested the equations

$$\begin{aligned}\frac{\mathrm{d}\xi_1}{\mathrm{d}t} &= \xi_1[1-C\xi_1^2-(B+C)\xi_2^2],\\ \frac{\mathrm{d}\xi_2}{\mathrm{d}t} &= \xi_2[1-C\xi_2^2-(B+C)\xi_1^2]\end{aligned} \quad (5)$$

as an appropriate dynamic description (Haken, 1996). With $U_j = \xi_j^2$ ($j = 1, 2$), Equations (5) are identical to Equations (4), showing only single-mode stability for $B > 0$. Study of this system in the (ξ_1, ξ_2) phase plane reveals two stable states: $(1/\sqrt{C}, 0)$ and $(0, 1/\sqrt{C})$, and jumping back and forth between these two states models one's subjective experience of Figure 3.

ALWYN SCOTT

See also **Cell assemblies; Lasers; Population dynamics; Quasilinear analysis; Synergetics; Tacoma Narrows Bridge collapse**

Further Reading

Andronov, A.A., Vitt, A.A. & Khaikin, S.E. 1966. *Theory of Oscillators*, Oxford and New York: Pergamon

Aumann, H.M. 1974. Standing waves on a multimode ladder oscillator. *IEEE Transactions on Circuits and Systems*, CAS-21: 461–462

Greene, P.H. 1962. On looking for neural networks and "cell assemblies" that underlie behavior. *Bulletin Mathematical Biophysics* 24: 247–275 and 395–411

Haken, H. 1996. *Principles of Brain Functioning: A Synergetic Approach to Brain Activity, Behavior and Cognition*, Berlin: Springer

Hoel, L.S., Keller, W.H., Nordman, J.E. & Scott, A.C. 1972. Niobium superconductive tunnel diode integrated circuit arrays. *Solid State Electronics*, 15: 1167–1173

Kryloff, N. & Bogoliuboff, N. 1947. *Introduction to Nonlinear Mechanics*, Princeton, NJ: Princeton University Press

Scott, A.C. 1971. Tunnel diode arrays for information processing and storage. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1: 267–275

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynmaics of Coherent Structures*, 2nd edition. Oxford and New York: Oxford University Press

van der Pol, B. 1934. The nonlinear theory of electric oscillations. *Proceedings of the Institute of Radio Engineers*, 22: 1051–1086

Volterra, V. 1937. Principes de biologie mathématique. *Acta Biotheoretica*, 3: 1–36

DNA PREMELTING

DNA premelting phenomena are spontaneous dynamical processes that occur well below the melting temperature of DNA. They reflect DNA "breathing," a process that combines the transient unstacking of base-pairs (allowing planar drugs and dyes to intercalate into DNA), with the transient breaking and rejoining of hydrogen-bonds connecting base-pairs in limited DNA regions (allowing tritium nuclei in hydrogen bonds connecting base-pairs to exchange with protons in water). The formation and propagation of chain-slippage structures either within or between homologous DNA molecules, along with a variety of structural phase transitions in DNA are also included in discussions of premelting. This entry describes how concepts of kink-antikink-bound states (breather solitons) have been used to assist in understanding a wide range of DNA premelting phenomena.

Conceptually, the lowest energy kink-antikink-bound state in DNA (termed a "premelton") contains a central hyperflexible beta-DNA core region modulated into B-DNA on either side through kink and antikink boundaries. As the kink and antikink move together, for example, the energy density in the central beta-DNA core region rises. This energy is used first to enhance (alternate) base-pair unstacking, and next, to stretch and to eventually break hydrogen-bonds connecting base-pairs. As kink and antikink move apart, the reverse occurs. Energy within this central core region falls, allowing hydrogen-bonds to reform, and (alternate) base-pairs to partially restack. Isoenergetic breather-motions such as these can facilitate the intercalation of drugs and dyes into DNA and allow tritium exchange to occur at temperatures well below the melting temperature.

Beta-DNA is a structural form that differs from either B- or A-DNA. Evidence for its existence comes from studies of intercalation by drugs and dyes into DNA and the binding of certain proteins to DNA. Although double-stranded, beta-DNA is unique in being both hyperflexible and metastable. Its hyperflexibility suggests it be a liquid-like phase, lying intermediate between the more rigid B- and A-forms and the melted single-stranded form. Both properties necessitate beta-DNA to be pinned by an intercalator or held by a protein in order to be studied in detail. The structure is composed of repeating units called beta-structural elements. These are a family of base-paired, dinucleotide structures, each possessing the same mixed sugar-puckering pattern (i.e., C3′ endo (3′–5′) C2′ endo) and having similar backbone conformational angles, but varying in the degree of base-unstacking. Lower energy forms contain base-pairs that are partially unstacked, while higher energy forms contain base-pairs completely unstacked.

Beta-DNA is an intermediate in DNA melting, lying on the minimal energy pathway connecting B-DNA with single-stranded melted DNA. Three distinctly different sources of nonlinearity appear as DNA chains unwind, and these determine the sequence of conformational changes that occur along this pathway.

The first two sources of nonlinearity stem from changes in the sugar-pucker conformations and base-pair stacking. These changes require small energies (ca. kT) and appear as part of the initial structural distortions accompanying DNA unwinding. Starting with B-DNA, the effect of unwinding DNA is to counterbalance this with an equal but opposite right-handed superhelical writhing to keep the linking invariant. This is achieved through a modulated beta-alternation in sugar-puckering along the chains, accompanied by the gradual partial unstacking of alternate base-pairs. The lowest energy beta-DNA structure emerges as an end result. Its metastability reflects the presence of additional energies in its structure that cause the partial unstacking of alternate base-pairs.

The third source of nonlinearity arises from the stretching and ultimate rupture of hydrogen bonds connecting base-pairs. At first, beta-DNA accommodates further unwinding through the gradual loss of superhelical writhing. This reflects the appearance of beta-structural elements having increasingly higher energy (these have base-pairs further unstacked and unwound). Eventually, however, a limit is reached and further unwinding begins to stretch hydrogen-bonds that connect the base-pairs. Continued unwinding results in the disruption of these hydrogen-bonds, and the appearance

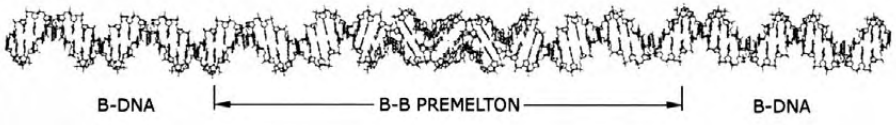

Figure 1. Molecular structure of a B–B premelton.

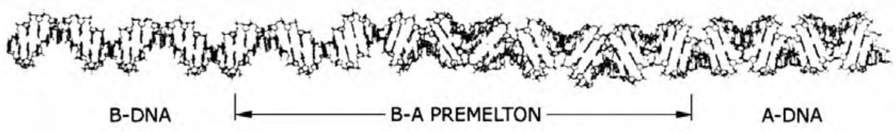

Figure 2. Molecular structure of a B–A premelton.

of single-stranded DNA. This final sequence of conformational change defines the boundary that connects beta-DNA to single-stranded melted DNA.

Using nonlinear least-squares methods, it has been possible to form premelton structures within B-DNA and within A-DNA (i.e., B–B premeltons, and A–A premeltons), as well as hybrid structures that connect the two (i.e., B–A premeltons and A–B premeltons) (see Figures 1 and 2). Such hybrid premeltons are constructed by connecting the central beta-DNA core with either type of kink-antikink boundary. Importantly, B–B and A–A premeltons are nontopological, whereas B–A and A–B premeltons are topological.

The B to A structural phase transition can be understood in the following way. In the presence of suitable thermodynamic conditions, kink and antikink within premeltons in B-DNA structure begin to move apart to form larger and larger core regions, whose centers modulate into A–DNA structure. Eventually, B–A premeltons and A–B premeltons form, and these continue to move apart, leaving a growing A-DNA region within. Such a mechanism is reversible and illustrates how a bifurcation within the central core region of this low-energy kink-antikink bound state structure can give rise to the B- to A-DNA structural phase transition.

Bifurcations arising within premeltons having higher energy (containing longer beta-DNA core regions that undergo more vigorous breather-motions) can lead to the formation of two additional types of higher energy kink-antikink bound states in DNA. These are called "branch-migratons" and "dislocatons." Each gives rise to different types of chain-slippage events, called double- and single-strand branch-migration.

Branch-migratons arise from breathing events that cause DNA chains to come apart transiently, and to then snap back, at nucleotide base sequences having 2-fold symmetry. Weakly hydrogen-bonded hairpin-like structures initially form. These are lengthened by a series of kinetic steps in which hydrogen-bonds connecting base-pairs within dinucleotide elements in vertical stems are broken and simultaneously reformed in horizontal stems (or vice versa) in a concerted 2-fold symmetric process. This phenomenon is called "cruciform-extrusion." A branch-migraton contains four stems; each stem contains kink (or antikink) boundaries connecting beta-DNA core regions with surrounding B–DNA.

Dislocatons arise at repetitive base sequences, for example, in poly-d(G-A): poly-d(C-T). Again, these structures form as a result of particularly energetic breathing events that cause DNA chains to come apart, and to then "snap back," forming small single-stranded bubble-like protrusions on opposite chains. These protrusions can then move apart, leaving growing regions of beta-DNA in between. The centers of these regions modulate into kink and antikink boundaries, and these, in turn, continue to move apart leaving B-DNA. The net result is the appearance of pairs of dislocatons, each moving in opposite directions along DNA. Movement involves single-chain slippage and is remarkably similar to the mechanism underlying moving crystal lattice dislocations, hence, the term dislocaton.

The formation of these two different kinds of higher energy kink-antikink bound state structures is another example of bifurcations emanating within the centers of premeltons. The underlying source of nonlinearity that determines the path of the bifurcation is the breaking and reforming (after chain-slippage) of hydrogen bonds connecting dinucleotide base-pairs. The decision as to whether branch-migratons or dislocaton pairs form is determined by information coded in the nucleotide base-sequence. This information constitutes the bias.

The combined presence of torsional and writhing strain energies found in negatively superhelical circular

DNA increases the probability that branch-migratons and dislocatons form at the appropriate sequences. These energies are first used to form premeltons. They are next used to form dislocatons or branch-migratons, and to propagate chain-slippage events.

Although B- and A-DNA are right-handed double-helical structures, DNA molecules containing the alternating poly-d(G–C): poly-d(G–C) sequence can, under certain conditions, assume a left-handed double-helical conformation (i.e., in the presence of high salt and/or negative superhelicity). This structure, called Z-DNA, contains the dinucleotide (G–C) as the asymmetric unit, held together by base-pairs. Being a left-handed double-helical structure, Z-DNA contains sugar-phosphate backbone conformations radically different from either B- or A-DNA.

The B to Z transition is proposed to occur as a result of a bifurcation that takes place during the formation of the dislocaton. As before, DNA breathing within the premelton takes place, followed by chains snapping back to form pairs of single-stranded, bubble-like protrusions on opposite chains. As pairs of protrusions move apart, Z-DNA forms within. The molecular boundaries that allow the helix to "swing left," capitalize on the additional flexibility and length provided by the single-stranded DNA regions on opposite DNA chains. B to Z boundaries form simultaneously on both ends in a concerted two-fold symmetric process. This is a direct consequence of the nonlinearity that ties the process together. A prediction of the model is the appearance of single-stranded DNA regions juxtaposed to beta-DNA regions within each B–Z junction, which is supported by experimental evidence.

HENRY M. SOBELL

See also **Biomolecular solitons; DNA solitons; Domain walls; Sine-Gordon equation**

Further Reading

Banerjee, A. & Sobell, H.M. 1983. Presence of nonlinear excitations in DNA structure and their relationship to DNA premelting and to drug intercalation. *Journal of Biomolecular Structure and Dynamics*, 1: 253–262

Bond, P.J., Langridge, R., Jennette, K.W. & Lippard, S.J. 1975. X-ray fiber diffraction evidence for neighbor exclusion binding of a platinum metallointercalation reagent to DNA. *Proceedings of the National Academy of Sciences*, 72: 4825–4829

Crothers, D.M. 1968. Calculation of binding isotherms for heterogeneous polymers. *Biopolymers*, 6: 575–583

Jessee, B., Gargiulo, G., Razvi, F. & Worcel, A. 1982. Analogous cleavage of DNA by micrococcal nuclease and a 1,10-phenanthroline-cuprous complex. *Nucleic Acids Research*, 10: 5823–5834

Lerman, L.S. 1961. Structural considerations in the interaction of DNA and acridines. *Journal of Molecular Biology*, 3, 18–30

Pohl, F.M., Jovin, T.M., Baehr, W., & Holbrook, J.J. 1972. Ethidium bromide as a cooperative effector of a DNA structure. *Proceedings of the National Academy of Sciences*, 69: 3805–3809

Printz, M.P. & von Hippel, P. 1965. Hydrogen exchange studies of DNA structure. *Proceedings of the National Academy of Sciences*, 53: 363–367

Sobell, H.M. 1985. Kink–antikink bound states in DNA Structure. In *Biological Macromolecules and Assemblies*, vol. 2: *Nucleic Acids and Interactive Proteins*, edited by F.A. Jurnack & A. McPherson, New York: Wiley, pp. 172–232

Sobell, H.M. 1985. Actinomycin and DNA transcription. *Proceedings of the National Academy of Sciences*, 82: 5328–5331

Stasiak, A., DiCapua, E. & Koller, T. 1983. Unwinding of duplex DNA in complexes with recA protein. *Cold Spring Harbor Symposia on Quantitative Biology*, 47: 811–820

DNA SOLITONS

In the early 1970s, Alexander Davydov and his colleagues began studies of nonlinear conformational waves in α-helix proteins. Englander et al. (1980) extended these ideas to DNA, interpreting open states (also called bubbles or local unwound regions, evidenced in hydrogen exchange experiments) as solitary waves, which can be viewed as kink solutions of the sine-Gordon (SG) equation. To understand the relation between the SG equation and DNA, we begin with models of the structure and dynamics of DNA.

Tightly packed in the nuclei of living cells, DNA consists of two polynucleotide chains (Figure 1). The chemical structure of each chain consists of periodically repeating phosphate groups and sugar rings (shown by dark ribbons), and irregular nitrous bases (shown in gray). Sequences of four types of bases (adenine, thymine, guanine, and cytosine) occur in DNA, forming the genetic code. These two polynucleotide chains interact weakly through hydrogen bonds and wind around each other to produce the double helix.

The internal structure of DNA is rather flexible. Thermal collisions with molecules of the surrounding solution, action of radiation, and local interactions with proteins, drugs, or other ligands induce several types of internal motions. Among these are

- small-amplitude oscillations of individual atoms;
- rotational, transverse, and longitudinal displacements of the atomic groups (phosphate groups, sugars, and bases);
- motions of the double chain fragments, having several base pairs lengths;

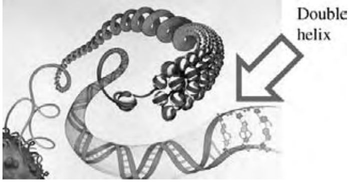

Figure 1. DNA molecule packing into the cell nucleus (courtesy of Nicolas Bouvier).

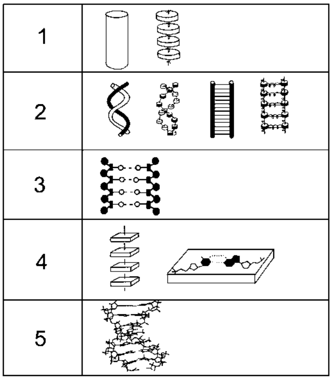

Figure 2. Hierarchy of models of DNA structure and dynamics.

- local unwinding of the double helix; and
- transitions of DNA fragments from one conformational form to another (e.g., from A- to B-form).

Thus, the DNA molecule is a complex dynamical system with many types of internal motions having different energies, velocities, amplitudes, frequencies, and lifetimes (McCammon & Harvey, 1987).

To describe the motions of DNA, different approximate models may be used, as indicated in Figure 2. The simplest model resembles a fragment of elastic thread (or rod), which ignores the internal structure. This model emerges naturally from microphotos where the DNA molecule does indeed resemble a thin elastic thread. Under this formulation only three coupled differential equations are needed: one for longitudinal displacements, another for angular displacements, and a third for transverse displacements. In Figure 2 a discrete analog of the model is placed nearby in the same row.

In the second row of Figure 2, a more complex model recognizes that DNA consists of two polynucleotide chains while ignoring the internal structure. Thus, the model has two weakly interacting elastic threads that wind around each other in a double helix. A mathematical formulation of this model requires six coupled differential equations—three for each of the two helices. The discrete version of the model and two simplified versions where the helical structure is neglected are shown nearby.

In the third row a more complex model is presented, taking into account that each of the polynucleotide chains comprises three types of atomic groups (phosphate groups, sugar rings, and nitrous bases) and representing solid-like motions of each of the atomic groups. Under this model, the required number of equations dramatically increases.

The model shown in the fourth row takes into account the internal motions of the atomic groups while neglecting variations of the base pairs (genetic code). Finally, the most accurate model is shown in the fifth line, which represents positions and motions of all atoms of the DNA molecule. The number of differential equations required to describe the motions in both of these models is inconveniently large at present levels of computer development.

To see the relation between DNA and the SG equation, apply the first model of the second row to study the opening of base pairs. The mathematical formulation of this model comprises six coupled differential equations that are nonlinear because base pair opening is a motion of large amplitude. Following Englander et al. (1980) we assume that rotational motions of bases around the sugar-phosphate chains make the greatest contribution to the process of opening, reducing the number of equations to two. Considering the rotational motions of bases in only one of the chains (accounting for the other chain as some averaged external field) reduces the number of equations from two to one.

To obtain the form of this single equation, an analogy between rotational motions of DNA bases and rotational motions of pendula in a mechanical model of the SG equation is convenient (*See* **Laboratory models of nonlinear waves**). Thus, the pendula of the mechanical model can be considered as analogs of DNA bases, the horizontal spring as an analog of one of the two sugar-phosphate chains, and the gravitational field as an averaged field formed by the second chain interacting with the first chain through hydrogen bonds. The analogy between these two dynamical systems suggests that the equation describing rotational motions bases in DNA should be an SG equation, with parameters defined as follows:

- I is the moment of inertia of bases.
- K is the coefficient of rigidity of the sugar-phosphate chain.
- V_0 is a potential function that represents interactions of bases through hydrogen bonds.
- R_0 is the radius of DNA.
- a is the distance between bases along the chains.

The variable of the equation is Ψ (the angle of rotations of bases around one of the two sugar-phosphate chains), and the equation itself is

$$I\frac{\partial^2}{\partial t^2}\Psi - KR_0^2a^2\frac{\partial^2}{\partial z^2}\Psi + V_0\sin\Psi = 0, \qquad (1)$$

where z indicates distance along the molecule. Kink solutions of this equation describe the opening of base pairs.

Another type of DNA soliton has been obtained by Peyrard & Bishop (1989), who studied melting and denaturation of DNA. Supposing that transverse (rather than rotational) motions of bases make the most contribution to the process, they reduced the number of equations from six to two, and after a simple linear transformation, they obtained two independent equations describing transverse displacements in phase and out of phase. The first of the equations was nonlinear and its solitary wave solutions were interpreted as bubbles appearing in the DNA structure as temperature was increased.

Solitary wave solutions obtained by Englander et al. (1980) and by Peyrard and Bishop (1989) were improved later by investigators who considered other types of internal motions and their interactions, effects of dissipation and inhomogeneity, and interactions with surroundings. These refinements led to additional solitary wave solutions that have been used to interpret experimental data on hydrogen-tritium exchange, resonant microwave absorption, and neutron scattering by DNA. Such studies help us to understand transitions between different DNA forms, long-range effects, regulation of transcription, DNA denaturation, protein synthesis (e.g., insulin production), and carcinogenesis (Scott, 1985; Gaeta et al., 1994; Yakushevich, 1998).

LUDMILA YAKUSHEVICH

See also **Biomolecular solitons; Davydov soliton; DNA premelting; Laboratory models of nonlinear waves; Sine-Gordon equation**

Further Reading

Davydov, A.S. & Kislukha, N.I. 1973. Solitary excitations in one-dimensional chains. *Physica Status Solidi* B, 59: 465–470

Davydov, A.S. & Suprun, A.D. 1974. Configurational changes and optical properties of alpha-helical proteins. *Ukrainian Physical Journal*, 19: 44–50

Englander, S.W., Kallenbach, N.R., Heeger, A.J., Krumhansl, J.A. & Litwin, A. 1980. Nature of the open state in long polynucleotide double helices: possibility of soliton excitations. *Proceedings of the National Academy of Sciences (USA)*, 77: 7222–7226

Gaeta, G., Reiss, C., Peyrard, M. & Dauxois, T. 1994. Simple models of nonlinear DNA dynamics. *Rivista del Nuovo Cimento*, 17: 1–48

McCammon, J.A. & Harvey, S.C. 1987. *Dynamics of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press

Peyrard, M. (editor). 1995. *Nonlinear Excitations in Biomolecules*, Berlin and New York: Springer

Peyrard, M. & Bishop, A.R. 1989. Statistical mechanics of a nonlinear model for DNA denaturation. *Physical Review Letters*, 62: 2755–2758

Scott, A.C. 1969. A nonlinear Klein–Gordon equation. *American Journal of Phyics*, 37: 52–61

Scott, A.C. 1985. Solitons in biological molecules. *Comments on Molecular and Cellular. Biophysics*, 3: 5–57

Yakushevich, L.V. 1998. *Nonlinear Physics of DNA*, Chichester and New York: Wiley

DOMAIN WALLS

In ferromagnetic materials, small regions of correlated magnetic moments formed below the critical temperature are called domains. Domain walls are two-dimensional structures that separate distinct domains of order and form spontaneously when a discrete symmetry (such as time-reversal symmetry in magnets) is broken at a phase transition. With each subdivision of a substance into distinct domains, there is a decrease in the bulk energy because the order parameter value inside each domain minimizes its free energy. However, there is a simultaneous increase in the energy of interaction between differently aligned domains giving rise to an extra surface energy at the boundaries between neighboring domains. Consequently, this competition leads to an average domain size that gives the lowest overall free energy in a material sample, which is quantified below.

The energy of a ferromagnetic domain wall is calculated as arising from the exchange interactions between spins augmented by the anisotropy energy, while the exchange energy for N spins of magnitude S comprising a domain wall varies as

$$E_{\text{exch}} = \frac{\pi^2 J S^2}{N}, \qquad (1)$$

where J is the exchange constant and the anisotropy energy is

$$E_{\text{anis}} = K N a^3, \qquad (2)$$

where a is the lattice constant and K the anisotropy constant. Minimizing the sum with respect to N yields

$$N = \sqrt{\frac{\pi^2 J S^2}{K a^3}} \qquad (3)$$

giving the domain width as $\Delta = Na$.

A Bloch domain wall is a region separating two (magnetic) domains within which magnetization changes gradually by rotating in the plane perpendicular to the line along the direction from one domain to the next. This way the magnetization direction experiences a reversal by 180° without changing its magnitude. The energy associated with a domain wall decreases with the width of the wall, and domain wall thickness is found as a minimization problem involving the anisotropy energy. A Neel domain wall, on the other hand, involves magnetization reversal in the plane perpendicular to the boundary between two domains.

Domains undergo reorganization under the effects of applied fields and can move in space. This occurs especially in the initial phase of remagnetization favoring those domains that are aligned with it and thus setting their boundary in motion to occupy more space, and it is followed by reorientation of the magnetization within each domain that is not aligned with the field.

Domain walls also exist in other systems, including crystals, ferroelectrics, metals, alloys, liquid crystals, and so on. In annealing metals, for example, domain walls appear as the grain boundaries between two sharply different compositions. Generically, the underlying physical quantity is called the order parameter and is specific for a given substance. (For the annealing metal it is a real field, while in superfluid helium it is a complex-valued field.) Over most of the sample, the order parameter has a constant magnitude, but the sign (when it is real) or the phase (when complex) is not fixed and can change from place to place. A real order parameter field may be positive in one region of space and negative in its neighborhood, the continuity of the field implies that it must cross the zero value on a surface between them. This transition region is a domain wall.

In all types of critical systems, domain walls arise due to the competition between the bulk part of the free energy, which in the Landau theory of phase transitions is a quartic polynomial in the order parameter ϕ, and the surface energy term, which is due to inhomogeneities and varies as the square of the order parameter gradient following Ginzburg. Minimizing this type of free energy functional

$$F = \int \left[A_2\phi^2 + A_4\phi^4 + D\left(\frac{d\phi}{dx}\right)^2 \right] dx \quad (4)$$

leads to a stationary nonlinear Klein–Gordon (NLKG) equation for the order parameter as a function of the spatial variable

$$D\phi'' = A\phi + B\phi^3, \quad (5)$$

where $A = A_2$ and $B = 2A_4$. One of its stable solutions is proportional to $\phi_0 \tanh(x/\xi)$ where $\xi = \sqrt{-8D/A}$, which describes a smooth function that interpolates between the two homogeneous phases $\phi = \pm\sqrt{-A/B}$. For magnetization as an order parameter, this solution represents a magnetic domain wall (in one space dimension), and for ferroelectrics, where the order parameter is a polarization vector, this represents a ferroelectric domain wall. For crystals undergoing structural phase transitions, there can also be a kinetic energy term in the free energy functional leading to a standard form of the NLKG equation

$$-m\ddot{\phi} + D\phi'' = A\phi + B\phi^3. \quad (6)$$

This solution is a moving domain wall (or kink)

$$\phi = \phi_0 \tanh[(x - vt)/\xi] \quad (7)$$

as shown in Figure 1 (Krumhansl & Schrieffer, 1975).

Nonlinear traveling solitary waves have also been investigated in ferroelectrics where kinks representing domain walls were shown to carry an electric dipole flip (Benedek et al., 1987). Domain walls in ferroelectrics are typically several unit cells wide, while in ferromagnets their thickness is several hundred or even thousands of unit cells. This difference is due to the exchange interactions between spins that are much stronger than the dipole-dipole interactions in ferroelectric crystals. There also exist cylindrical domains in magnets. As can be seen from Table 1, domain walls are examples of topological defects, and as such, they are very common in all broken-symmetry phenomena that take place slowly enough to allow for the generation of defects. Modern particle physics predicts that phase transitions occurred in the early universe following the Big Bang. Of particular interest to cosmology is the production of topological defects, which may be sheet-like, line-like, or point-like concentrations of energy.

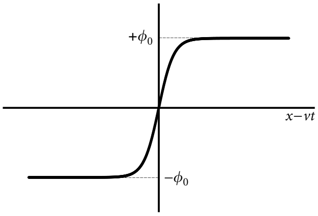

Figure 1. A typical form of a domain wall (or kink).

Geometry	Defect
Sheet-like	domain walls, membranes
Line-like	vortices, strings
Point-like	monopoles, hedgehogs

Table 1. Geometry of space and the corresponding topological defects.

JACK A. TUSZYŃSKI

See also **Critical phenomena; Ferromagnetism and ferroelectricity; Order parameters; Topological defects**

Further Reading

Anderson, P.W. 1984. *Basic Notions of Condensed Matter Physics*, Menlo Park, CA: Benjamin Cummings

Benedek, G., Bussmann-Holder, A. & Bilz, H. 1987. Nonlinear travelling waves in ferroelectrics. *Physical Review*, B36: 630–638

Kittel, C. 1996. *Introduction to Solid State Physics*, 7th edition, New York: Wiley

Krumhansl, J.A. & Schrieffer, J.R. 1975. Dynamics and statistical mechanics of a one-dimensional model Hamiltonian for structural phase transitions. *Physical Review*, B11: 3535–3545

White, R.H. & Geballe, T. 1979. *Long Range Order in Solids*, New York: Academic Press

DOPPLER SHIFT

The Doppler shift (or Doppler effect) is named after the Austrian scientist Christian Doppler, who in 1842 noticed that the observed frequency of waves from a source depends on how it is moving relative to the observer. In its simplest form, applicable to motions that are slow compared with the wave speed, it is quite easy to visualize and understand. If a source of waves at some frequency f is approaching an observer, the frequency f' that the observer detects is higher than would be the case were the two at rest with respect to each other. From the point of the observer, he is "running into" the waves, while from the point of view of the source, the waves are being "bunched up." The opposite effect occurs when a source is moving away from an observer, and the general situation can be expressed as

$$\frac{f'}{f} = 1 + (v/c)\cos\theta, \tag{1}$$

where v is the relative speed, c the speed of the waves, and θ the angle between the wave propagation and the relative velocity. The $\cos\theta$ factor is simply a statement that the only important component of the relative velocity which can contribute to the shift is that along the direction of wave propagation. As long as v is small compared with the speed of light, this expression works well for most applications, and is quite appropriate to discuss the frequency shift of reflected radar signals used to detect speeding motorists (where c would be the speed of light) or the rise and fall in the pitch of a train whistle or an ambulance siren as it approaches and passes (in which case c would be the speed of sound).

One of the most famous experiments ever done to demonstrate the Doppler shift was done by the Dutch meteorologist Buys-Ballot who put a group of musicians on a train and then had them hold a constant note while racing past him as he stood on the platform.

The Doppler shift often appears more than once in a single application. For example, in police radar, electrons in the metal of an approaching car see a higher frequency radar signal than the police officer emits. They then re-emit radio waves at a higher frequency that is in turn seen at a still higher frequency by the officer being approached.

When speeds involved are very large, the effects of special relativity come into play. While it is possible to outrun a sound wave, the speed of light is the same no matter how the source moves. If we take c to be the speed of light and consider the Doppler shift for light (or any other electromagnetic wave), we have

$$\frac{f'}{f} = \frac{1 + (v/c)\cos\theta}{\sqrt{1 - v^2/c^2}}. \tag{2}$$

A good discussion of special relativity including the Doppler shift is that of Rindler (1991), and a particularly elegant geometric derivation of the formula under very general conditions can be found in Burke (1980). The additional term in the denominator of Equation (2) is due to time dilation whereby a moving clock appears to go more slowly than one at rest. In this case there is a Doppler shift even when $\cos\theta = 0$, which is referred to as the "transverse Doppler effect." One might imagine that it would be very difficult to detect, but in the special case where the sources are in random thermal motion, the leading $(v/c)\cos\theta$ term averages to zero and merely broadens spectral lines of atoms without actually shifting them. This "thermal Doppler effect" due to the transverse Doppler effect was measured by Pound and Rebka in 1960 using the Mössbauer effect. Interestingly, the nontransverse Doppler shift itself played an important role in that measurement.

The Doppler effect has numerous practical applications. In addition to its use already mentioned in measuring automobile velocities, these include Doppler radar for studying weather based on measuring the speeds of raindrops blown by wind (see, e.g., Doviak & Zrnic, 1993) and imaging moving tissues such as the heart in echocardiography, or measuring the speed of blood flow through an artery (see, e.g., Evans & McDicken, 2000).

The Doppler effect is also a valuable tool for pure science and has played a key role in many experiments upon which our present view of the world is based. Edwin Hubble used it to determine the velocities of distant objects in space from shifts in their spectral lines (shifted toward the red end of the spectrum, hence redshift), finding that distant objects seem to be moving away from us with speeds that increase with their distance (Christianson, 1995). This discovery is the basis of modern cosmology and is one of the strongest pieces of evidence that the universe had its origins in a Big Bang. The Doppler shift has also been used in delicate and beautiful experiments that demonstrate the time dilation effects of gravity. Both these gravitational effects on time and the relativistic effect in the transverse Doppler shift must be taken into account in order for the Global Positioning System (GPS) to function properly.

JOHN DAVID SWAIN

See also **Einstein equations; Gravitational waves**

Further Reading

Burke, W.L. 1980. *Spacetime, Geometry, Cosmology*, Mill Valley, CA: University Science Books

Christianson, G.E. 1995. *Edwin Hubble: Mariner of the Nebulae*, New York: Farrar Straus & Giroux

Doviak, R.J. & Zrnic, D.S. 1993. *Doppler Radar and Weather Observations*, New York: Academic Press

Evans, D.H. & McDicken, W.N. 2000. *Doppler Ultrasound: Physics, Instrumental, and Clinical Applications*, 2nd edition, New York: Wiley

Rindler, W. 1991. *Introduction to Special Relativity*, 2nd edition, Oxford and New York: Oxford University Press

DOUBLE-WELL POTENTIAL

See **Equilibrium**

DRESSING METHOD

The dressing method is a technique of constructing and solving nonlinear partial differential equations of integrable models. It is based on dressing transformations that are symmetries of nonlinear partial differential equations and act on the Lax operators as gauge transformations on the connection (Zakharov & Shabat, 1974). Accordingly, the form of the linear spectral problem and the zero-curvature conditions are preserved.

The basic concept of the dressing method (Zakharov & Shabat, 1974) is that starting from known solutions of the underlying linear problem, one obtains new solutions of the transformed ("dressed") linear problem. Suppose Ψ_0 is a solution of the following linear problems:

$$\left(\frac{\partial}{\partial t_i} - E_i^{(0)}\right)\Psi_0\,(t_1,\ldots,t_n,\lambda) = 0, \quad i = 1,\ldots,n, \tag{1}$$

where $t_1, t_2, \ldots, t_n$ are independent flow variables and $E_i^{(0)}$ are complex $m \times m$ mutually commuting matrices that depend on the spectral parameter λ. The matrix Ψ_0 is referred to as a "bare" wave function and commuting operators $L_i^{(0)} = \partial/\partial t_i - E_i^{(0)}$ as "bare" Lax operators.

For the soliton equations in $1+1$ (one spatial and one time) dimensions, the dressing transformations are generated by the dressing matrices $\Theta_\pm$ which can be defined by the Riemann–Hilbert factorization problem (Zakharov & Shabat, 1974; Faddeev & Takhtajan, 1987). Given a Lie group of functions $g(\lambda)$ on the contour $\mathcal{C}$ in the complex plane, one finds a new (dressed) wave function through the factorization problem:

$$\Psi_0(t_1,\ldots,t_n,\lambda)g(\lambda)\Psi_0^{-1}(t_1,\ldots,t_n,\lambda) = \Theta_-^{-1}(t_1,\ldots,t_n,\lambda)\Theta_+(t_1,\ldots,t_n,\lambda), \tag{2}$$

where $\Theta_-^{-1} = (\Psi_0 g\Psi_0^{-1})_-$ and $\Theta_+ = (\Psi_0 g\Psi_0^{-1})_+$ have analytic continuation inside or outside the contour $\mathcal{C}$, respectively. The dressing transformation defines a new wave function $\Psi^{(g)} = \Theta_-\Psi_0 g = \Theta_+\Psi_0$, which is a solution of new linear problems:

$$\left(\frac{\partial}{\partial t_i} - E_i\right)\Psi^{(g)}(t_1,\ldots,t_n,\lambda) = 0, \tag{3}$$

where

$$E_i = \frac{\partial\Theta_-}{\partial t_i}\Theta_-^{-1} + \Theta_- E_i^{(0)}\Theta_-^{-1} = \frac{\partial\Psi^{(g)}}{\partial t_i}(\Psi^{(g)})^{-1} \tag{4}$$

satisfy the compatibility conditions also known as zero-curvature conditions:

$$\frac{\partial E_j}{\partial t_i} - \frac{\partial E_i}{\partial t_j} - \left[E_i\,,\,E_j\right] = 0. \tag{5}$$

Comparing the expression for E_i from Equation (4) with $E_i^{(0)} = (\partial\Psi_0/\partial t_i)\Psi_0^{-1}$, following from Equation (1), one sees that the dressing transformations preserve the form of the Lax connections $L_i = \partial/\partial t_i - E_i$. A related approach also exists, which yields the dressing matrix in terms of the Fredholm type of integral operator entering the Gel'fand–Levitan–Marchenko equation (Zakharov & Shabat, 1974). Both dressing approaches are equivalent.

By considering two successive dressing transformations associated with two group elements g_1 and g_2, one naturally arrives at the concept of the group of the dressing transformations with $\Psi^{(g_1 g_2)} = (\Psi^{(g_1)})^{(g_2)}$. The general theory of dressing transformations and its group was developed by Semenov–Tian–Shansky (1985), who also introduced a Poisson bracket covariant under the dressing group action on the phase space of functions E_i. With such a bracket, the group of the dressing transformations induces on a phase space a Lie–Poisson action and turns out to be a symmetry of the phase space. Furthermore, it was observed that the group of the dressing transformations appears as a semi-classical limit of the quantum group symmetry of the two-dimensional integrable quantum field theories. Hence the group of the dressing transformations appears to be a classical precursor of the quantum group structure of an integrable field system in two dimensions (Babelon & Bernard, 1992). In many integrable models, the N-soliton solutions can be thought of as elements of the dressing group orbit of the vacuum state. Accordingly, successive dressing transformations can be used to build N-soliton solutions from the vacuum solution. In Babelon & Bernard (1993), the authors presented the construction of N-soliton solutions by dressing transformations in the sine-Gordon model. The dressing group also admits an elegant interpretation within the Kyoto school (Date et al., 1983) approach to the integrable models appearing in this context as a transformation group of the τ-function.

For equations in $2+1$ (two spatial and one time), dimensions, one applies the dressing technique based on a nonlocal Riemann–Hilbert problem for Kadomtsev–Petviashvili I and Davy–Stewartson I equations as well as the N-wave equations. The nonlocal (D-bar) $\bar{\partial}$ problem is required for the Kadomtsev-Petviashvili II and Davey–Stewartson II equations (Zakharov & Manakov, 1985).

The nonlocal (D-bar) $\bar{\partial}$-problem gives rise to the $\bar{\partial}$-dressing method in which an $N \times N$ quasi-analytical matrix function of $\lambda, \bar{\lambda}\ \chi(\lambda, \bar{\lambda}, x)$, $x \in \mathbb{C}^n$ satisfies the following nonlocal $\bar{\partial}$-problem:

$$\frac{\partial\chi(\lambda,\bar{\lambda})}{\bar{\lambda}} = \hat{R}\chi = \int\!\!\int_C \mathrm{d}\lambda'\mathrm{d}\bar{\lambda}'\chi(\lambda',\bar{\lambda}')R(\lambda',\bar{\lambda}',\lambda,\bar{\lambda}), \tag{6}$$

with R being a kernel of a linear integral operator $\hat{R}$. The set of commuting differential operators defined in terms of commuting rational matrix functions I_i:

$$D_i\chi = \frac{\partial\chi}{\partial x_i} + \chi I_i(\lambda), \qquad i = 1,\ldots,n, \tag{7}$$

defines an integrable system. Let the integral linear operator $\hat{R}$ commute with all differential operators D_i, $[D_i\,, \hat{R}] = 0$. Then, the choice $I_1 = \lambda$, $I_2 = \lambda^2$, $I_3 = \lambda^3$ leads for $N = 1$ to the KP II equation: $(u_t + u_{xxx} + + 6uu_x)_x + 3u_{yy} = 0$ with $u = 2\partial\chi_1/\partial x_1$ where χ_1 is the term in an asymptotic expansion ($\lambda \to \infty$) of $\chi = 1 + \chi_1/\lambda + \chi_2/\lambda^2 + \ldots$.

HENRIK ARATYN

See also **Bäcklund transformations; Darboux transformation; Hirota's method; Inverse scattering method or transform; Multidimensional solitons; *N*-soliton formulas; Riemann–Hilbert problem**

Further Reading

Babelon, O. & Bernard, D. 1992. Dressing symmetries. *Communications in Mathematical Physics*, 149: 279–306

Babelon, O. & Bernard, D. 1993. Affine solitons: a relation between tau functions, dressing and Bäcklund transformations. *International Journal of Modern Physics A*, 8: 507–543

Date, E., Jimbo, M., Kashiwara, M. & Miwa, T. 1983. Transformation groups for soliton equations. In *Nonlinear Integrable Systems-Classical and Quantum Theory*, edited by M. Jimbo & T. Miwa. Singapore: World Scientific

Faddeev, L.D. & Takhtajan, L.A. 1987. *Hamiltonian Methods in the Theory of Solitons*, Berlin and Heidelberg: Springer

Semenov–Tian–Shansky, M.A. 1985. Dressing transformations and Poisson group actions, *Publications of Research Institute of Mathematical Sciences, Kyoto University*, 21: 1237–1260

Zakharov, V.E. & Manakov, S.V. 1985. Construction of multidimensional integrable non-linear systems and their solutions, *Functional Analysis and Applications*, 19 (N2): 11–25

Zakharov, V.E. & Shabat, A.B. 1974. A scheme for integrating the nonlinear equations of mathematical physics by the method of the inverse scattering problem. I. *Functional Analysis and Applications*, 8: 226–235

Zakharov, V.E. & Shabat, A.B. 1979. Integration of the nonlinear equations of mathematical physics by the method of the inverse scattering problem. II. *Functional Analysis and Applications*, 13: 166–174

DRIPPING FAUCET

A dripping faucet may easily be seen in everyday life. Its rhythm is sometimes regular but sometimes not, which sensitively depends on the flow of water. If the rhythm is irregular, one might blame it on noise due to unseen influences such as small air vibrations. However, it is nowadays well known that the irregularity arises from deterministic chaos instead of stochastic noise. In fact, this system is a good example showing that chaos is not only a mathematical product but also a phenomenon ubiquitous in the real world.

Chaotic dripping was originally suggested by Otto Rössler (Rössler, 1977) and the first experimental study was performed by Robert Shaw and his colleagues (Shaw, 1984; Martien et al., 1985). They measured the time interval (T_n) between the nth drip and its successor, and obtained a time series $\{T_1, T_2, \cdots\}$. To detect possible determinism from a nonperiodic time series, they made a return map (plot of (T_n, T_{n+1}) for each n). A return map for nonperiodic dripping is shown in the inset of Figure 1(a). If irregular numbers $T_1, T_2, \cdots$ were generated stochastically by throwing the dice, the plots would look like a set of random points with no particular structure. The observed map actually exhibited a clear structure, which implies a deterministic rule existing in the seemingly random outcomes.

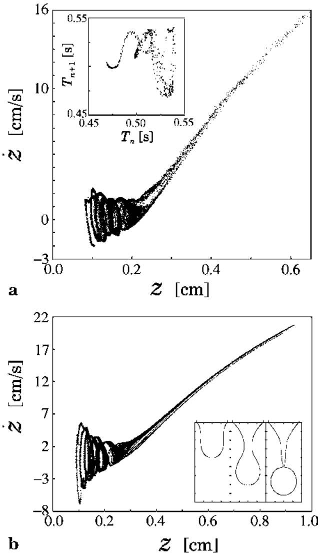

Figure 1. (a) An experimental strange attractor reconstructed from the observation of the oscillation, deformation, and breakup of drops hanging from a nozzle (7 mm inner diameter and 10 mm outer diameter). The projection of the orbit onto the plane (z, $\dot{z}$) is presented, where z is the center of mass and $\dot{z}$ is its velocity. The flow rate $Q = 0.24$ g/s ($\sim$ 2 drips/s on average). Inset of (a): return map of the dripping time interval (T_n). (b) Fluid dynamic simulation for 7 mm nozzle diameter. $Q = 0.32$ g/s. Inset of (b): drop deformation.

As suggested by Rössler, the deterministic randomness, chaos is expected if two oscillating variables couple: (i) damped oscillation of the drop position (z, the center of mass) due to the surface tension of the water and (ii) relaxation oscillation of the mass of the drop (m) due to the filling-and-discharging process. A minimal model including these variables is the so-called mass-spring model described as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(m\frac{\mathrm{d}z}{\mathrm{d}t}\right) = -kz - \gamma\frac{\mathrm{d}z}{\mathrm{d}t} + mg, \qquad (1)$$

where g is the acceleration of gravity, k is the spring constant, and γ is the damping parameter. The mass increases at a constant flow rate Q as

$$\frac{\mathrm{d}m}{\mathrm{d}t} = Q. \qquad (2)$$

The model assumes that when the position z reaches a critical point, a part of the total mass (Δm) breaks away. Shaw used this model to explain the dripping dynamics as low-dimensional chaos.

Since then, dripping faucets have attracted many physicists, and a wide range of nonlinear behavior has been reported, such as strange attractor, period-doubling bifurcation, Hopf bifurcation, intermittency, hysteresis, crisis, and satellite drop formation. On the other hand, theoretical studies mainly rested on the mass-spring model, and any direct link between this simple model and the physics of drop formation was not known. The basic dynamics inherent in the complex behavior of the dripping faucet system was revealed quite recently, owing to detailed analyses of (i) experiments for a wide range of flow rates (Katsuyama & Nagata, 1999), (ii) fluid dynamic simulations using a new algorithm (Fuchikami et al., 1999), and (iii) the improved mass-spring model based on the fluid dynamic simulations (Kiyono & Fuchikami, 1999).

The fluid dynamic simulations clearly visualized how drops are formed and pinched off repeatedly (see figures in color plate section):

- The water under the faucet, increasing at a constant flow rate, forms a drop, which bulges until m reaches $\sim m_{\text{crit}}$ (the maximum mass for the static stable state) and the surface tension is overwhelmed by the gravitational force.
- Then its sides begin to shrink, forming a rapidly narrowing neck (necking process) and the drop is soon pinched off (breakup). In the necking process, the drop undergoes almost free-fall.
- Because the water is stretched downward at the breakup moment, the surface tension works as a restoring force just after the breakup, which causes oscillation of the water. Thus the position z, the center of mass of the water, moves downward with up-and-down oscillation.
- After m reaches m_{crit} again, the necking followed by the breakup is repeated.
- The phase of oscillation at the onset of the necking process affects the breakup moment, the mass at that moment, and so the remnant mass ($m_{\text{r}} = m - \Delta m$).
- If the phase changes periodically, the motion is periodic (for example, in the period-two motion T_1, T_2, T_1, T_2, $\cdots$, the phase changes as θ_1, θ_2, θ_1, θ_2, $\cdots$), while the phase is random in chaotic motion.

The essential information obtained from the fluid dynamic simulations is as follows:

(i) The state point of the water is well described in the limited phase space (z, $\mathrm{d}z/\mathrm{d}t$, m), even if the water, an infinite-dimensional system, deforms its shape in a complex way. (Remember z denotes the center of mass.)

(ii) The instability of the shape of water induces the instability of the chaotic orbit. In other words, stretching of the chaotic attractor mainly occurs in the beginning of the necking process.

(iii) After the breakup, the renewed (i.e., remnant) mass realizes various values, while the renewed position and velocity are confined in a small region, well approximated as constants. Thus the attractor is compressed and becomes low dimensional.

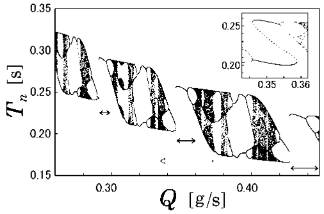

Figure 2. A section of the bifurcation diagram of the improved mass-spring model obtained by decreasing the control parameter Q, the flow rate. Hysteresis is observed in certain ranges of Q (indicated with $\leftrightarrow$). Inset: a hysteresis curve.

These features have also been confirmed experimentally. Figure 1 presents a recent experiment (a) and the corresponding fluid dynamic simulation result (b). The low-dimensional (almost one-dimensional) pattern of the return map (inset of (a)) suggests that the system can be described by a low-dimensional dynamical system. The experimental trajectory in the phase space (z, dz/dt, m) was reconstructed from the continuous change of the drop shape observed. The spiral orbit indicates that the drop oscillates several times and then makes free-fall in the necking process before breakup. The drop shape obtained from the simulation (inset of (b)) is very close to the experimental result.

The observation of the drop formation process made it possible to improve the traditional mass-spring model by taking account of several points ignored so far, which include

- the mass dependence of the spring constant: $k = k(m)$;
- the necking process by setting $k = 0$ for $m > m_{\text{crit}}$; and
- the mass dependence of the remnant mass: $m_{\text{r}} = m_{\text{r}}(m)$, where m is the mass just before the breakup.

Bifurcation diagrams (plot of T_n versus Q) obtained from the improved mass-spring model reproduce the global structure of experimental bifurcation diagrams for a wide range of flow rates, Q. As seen in Figure 2, one distinct feature is a repetition of similar "units." The neighboring units are very similar but gradually become complex as Q is increased. Each unit is characterized by an integer that is the number of oscillations of each drop before breakup. This number decreases with increasing Q because the drop mass reaches the critical value sooner. In Figure 2, for example, there are three units and the corresponding numbers are 6, 5, 4 from left to right. Units in a range of relatively large Q include various types of bifurcations, such as period-doubling cascade to chaos, intermittency, and hysteresis (inset of Figure 2), while units in a range of sufficiently small Q include just period one motion. The bifurcation diagrams for small faucet diameters also exhibit a relatively simple structure.

The improved mass-spring model systematically explains the characteristic complexities of low-dimensional chaos (Ott, 1993) inherent in dripping dynamics. However, the model can be applied only when Q is small enough so that drops are clearly separated from each other. For larger flow rates, experimental results are so complex that approximations used in the fluid dynamic simulations do not work. New theoretical approaches are required, especially to interpret Hopf bifurcation and statistical features of satellite drops.

NOBUKO FUCHIKAMI AND KEN KIYONO

See also **Bifurcations; Chaotic dynamics**

Further Reading

Fuchikami, N., Ishioka, S. & Kiyono, K. 1999. Simulation of a dripping faucet. *Journal of the Physical Society of Japan*, 68: 1185–1196

Katsuyama, T. & Nagata, K. 1999. Behavior of the dripping faucet over a wide range of the flow rate. *Journal of the Physical Society of Japan*, 68: 396–400

Kiyono, K. & Fuchikami, N. 1999. Dripping faucet dynamics by an improved mass-spring model. *Journal of the Physical Society of Japan*, 68: 3259–3270

Kiyono, K., Katsuyama, T., Masunaga, T. & Fuchikami, N. 2003. Picture of the low-dimensional structure in chaotic dripping faucets. *Physics Letters* A, 320: 47–52

Martien, P., Pope, S.C., Scott, P.L. & Shaw, R.S. 1985. The chaotic behavior of the leaky faucet. *Physics Letters* A, 110: 399–404

Ott, E. 1993. *Chaos in Dynamical Systems*, Cambridge and New York: Cambridge University Press

Rössler, O.E. 1977. Chemical turbulence. In *Synergetics: Proceedings of the International Workshop on Synergetics at Schloss Elmau, Bavaria*, edited by Hermann Haken, Berlin: Springer, 174–183

Shaw, R. 1984. *Dripping Faucet as a Model Chaotic System*, Santa Cruz: Aerial Press

DRUDE MODEL

In 1900, Paul Drude developed his theory of metallic conduction of electricity, following the discovery of the electron by Joseph John Thomson in 1897. Applying the kinetic theory of gases to a metal, considered as a gas of electrons, he made the following assumptions:

(i) Each electron moves according to Newton's law of motion in the presence of external fields until it collides with other electrons or ions. Between collisions, interactions with both the other electrons and with the ions are neglected.

(ii) Collisions are instantaneous events that abruptly alter the velocity of an electron. The probability per unit time that an electron experiences a collision is given by $1/\tau$, where τ is called the collision time, mean free time, or relaxation time.

(iii) Electrons are assumed to achieve thermal equilibrium with their surroundings only through collisions. After each collision, the electron emerges with a randomly directed velocity whose magnitude is given by the temperature.

Note that in contrast to a classical gas of neutral molecules, the electrons move against a background of positively charged immobile ions.

The Drude model considers an average electron representative for the whole ensemble. Newton's equation of motion is

$$m\frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{r} + \frac{m}{\tau}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r} = -e\boldsymbol{E}, \qquad (1)$$

where $\boldsymbol{r}$ is the spatial drift coordinate (excluding the random thermal motion), m is the effective mass of the electron, $e > 0$ is the electron charge, and $\boldsymbol{E}$ is the applied electric field. The second term is a friction term arising from collisions. This equation can be rewritten in terms of the electron momentum $\boldsymbol{p} = m\boldsymbol{v} = m\mathrm{d}\boldsymbol{r}/\mathrm{d}t$, where $\boldsymbol{v}$ is the drift velocity,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p} + \frac{\boldsymbol{p}}{\tau} = -e\boldsymbol{E}\,. \qquad (2)$$

In the overdamped case the momentum induced by the electric field is given by

$$\boldsymbol{p} = -e\tau\boldsymbol{E}\,. \qquad (3)$$

The current density is obtained by multiplying the drift velocity of the mean electron by the electron charge and by the density per unit volume n of all electrons

$$\boldsymbol{j} = -en\boldsymbol{v} = en\frac{e}{m}\tau\boldsymbol{E} = en\mu\boldsymbol{E}\,, \qquad (4)$$

where the drift mobility $\mu = (e/m)\tau$ has been introduced. This is Ohm's law, with constant conductivity $\sigma = en\mu$.

The classical Drude model neglects electron-electron interactions, heating of the electrons, non-equilibrium dynamics, and quantum transport effects, and is thus restricted to low fields, moderate carrier densities and temperature ranges. If any of these conditions are violated, nonlinear and non-Ohmic conduction arises. Such effects are abundant in metals and semiconductors and may lead to instabilities and bifurcations of self-sustained oscillations or self-organized spatiotemporal patterns (Shaw et al., 1992, Schöll, 2001).

In the simplest case, a local, instantaneous conductivity $\sigma = en\mu$ still exists, but n or μ depends upon the field E (generation-recombination or drift instability, respectively), leading to a nonlinear or even nonmonotonic current density-field relation $j = en(E)\mu(E)E$, possibly with a range of negative differential conductivity (NDC), where $\mathrm{d}j/\mathrm{d}E = en\mu + e(\mathrm{d}n/\mathrm{d}E)\mu(E)E + en(\mathrm{d}\mu/\mathrm{d}E)E < 0$.

In an extension of the simple Drude picture, those nonlinearities may be due to a dependence of the momentum relaxation time upon the field (drift instability) or upon the electron temperature that, in turn, depends upon E (electron overheating instability, as a result of changes in the dissipation of energy and momentum). They may also be due to dependence of the carrier density upon field (generation-recombination instability, as in avalanche breakdown). More complex situations arise if the (semi)conductor consists of several layers of different materials, and intrinsic inhomogeneities render the notion of a local conductivity inappropriate. In particular, this applies to low-dimensional nanoscale structures (Ferry & Goodnick, 1997) or mesoscopic conductors (Datta, 1995), where the characteristic length scales may be such that the device dimension becomes smaller than the mean free path, the quantum mechanical phase relaxation length, and the de Broglie wavelength of the electron. Then various nonlinear transport regimes occur, where ballistic or non-equilibrium or coherent quantum effects dominate.

ECKEHARD SCHÖLL

See also **Avalanche breakdown; Diodes; Semiconductor oscillators**

Further Reading

Ashcroft, N.W. & Mermin, N.D. 1976. *Solid State Physics*, Philadelphia: Saunders College

Datta, S. 1995. *Electronic Transport in Mesoscopic Systems*, Cambridge and New York: Cambridge University Press

Drude, P. 1900. Zur Elektronentheorie der Metalle. *Annalen der Physik*, 1: 566–613

Ferry, D.K. & Goodnick, S.M. 1997. *Transport in Nanostructures*, Cambridge and New York: Cambridge University Press

Schöll, E. 2001. *Nonlinear Spatio-temporal Dynamics and Chaos in Semiconductors*, Cambridge: Cambridge University Press

Schöll, E. (editor). 1998. *Theory of Transport Properties of Semiconductor Nanostructures*, London: Chapman & Hall

Shaw, M.P., Mitin, V.V., Schöll, E. & Grubin, H.L. 1992. *The Physics of Instabilities in Solid State Electron Devices*, New York: Plenum Press

DUFFING EQUATION

Serious studies of forced nonlinear oscillators appeared early in the 20th century when Georg Duffing (1918) examined mechanical systems with nonlinear restoring forces and Balthasar van der Pol studied electrical systems with nonlinear damping. Subsequently, any equation of the form

$$a\ddot{x} + b\dot{x} + f(x) = F \sin \omega t \qquad (1)$$

was called Duffing's equation, the nonlinearity often being polynomial, usually cubic (Hayashi, 1964). Here, $\dot{x}$ represents the time derivative $\mathrm{d}x/\mathrm{d}t$.

Linear resonance has $f(x) = cx$ $(c > 0)$, and Duffing's extension models many mechanical and electrical phenomena. With symmetry, f is odd, giving the first nonlinear approximation

$$f(x) = cx + dx^3. \qquad (2)$$

If $d > 0$, the system is *hardening* and the resonant peak tilts to the right. A *softening* system ($d < 0$) has a peak tilted to the left (as in Figure 2). Duffing's method of successive approximation and a variety of averaging and perturbation techniques can estimate these tilts for conditions of weak nonlinearity.

Twin-Well Duffing Oscillator

Taking Equations (1) and (2) with $c < 0$, $d > 0$ gives the "twin-well" Duffing oscillator governed by the potential energy $V(x) = \frac{1}{2}cx^2 + \frac{1}{4}dx^4$. Trajectories lie in and across the two symmetric potential wells separated by the hill-top at $x = 0$. This is a useful archetypal model for studies of chaos (Guckenheimer & Holmes, 1983). The undamped, unforced system ($b = F = 0$) has two symmetrically placed orbits that (in infinite time) leave and return to the phase-space saddle at $x = 0$, $\dot{x} = 0$: we say they are "homoclinic to the saddle." For small b and F the corresponding invariant manifolds exhibit a homoclinic tangency on an arc in the (b, F) control space. Beyond this global bifurcation a homoclinic tangle generates "horseshoe dynamics" giving chaos and fractal basin boundaries (see Figure 3). The (b, F) arc can be estimated by Mel'nikov perturbation analysis.

Ueda's Chaos

For hardening with $c = 0$, Ueda (1980, 1992) mapped intricate regimes of subharmonics and chaos. Ueda's equation is

$$\ddot{x} + k\dot{x} + x^3 = B \cos t. \qquad (3)$$

With $k = 0.05$, $B = 7.5$, all solutions settle onto a unique chaotic attractor, irrespective of $x(0), \dot{x}(0)$: its basin of attraction is the whole starting plane.

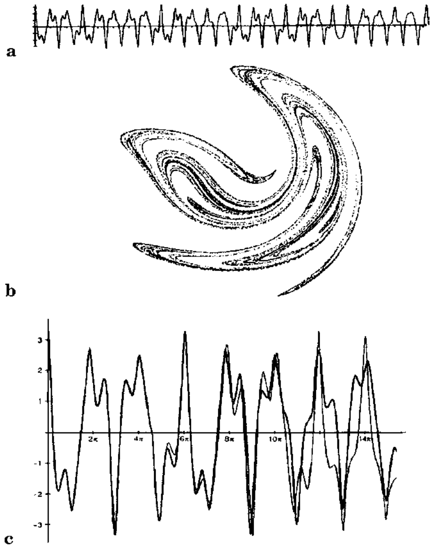

Figure 1. Waveform (a) shows a plot of $x(t)$ at the end of a long computer time integration during which any start-up transient is effectively dissipated, while the large fractal dot diagram (b) shows this chaotic attractor in a Poincaré section sampling $x(t)$ and $\dot{x}(t)$ stroboscopically at the period of the forcing. Two waveforms in (c) show the exponential divergence from two adjacent starts on the attractor where $(x(0), \dot{x}(0))$ are $(3, 4)$ and $(3.01, 4.01)$, respectively. Reproduced from Thompson & Stewart (2002) with the permission of John Wiley.

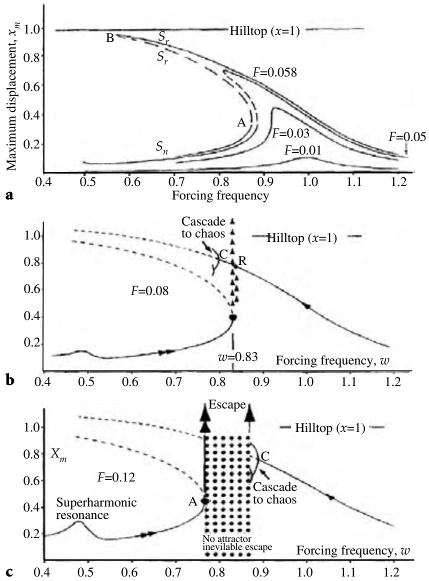

Figure 2. Three resonance response diagrams for Equation (4), where the maximum of the steady-state response, x_m, is plotted against the forcing frequency, ω, for fixed values of F (all with $\beta = 0.1$). At higher F, in (c), there is a regime with no attractor, implying inevitable escape from either fold A or the chaotic crisis ending the cascade from C. Reproduced from Thompson & Stewart (2002) with the permission of John Wiley.

The steady chaotic response is shown in Figure 1. In each cycle, sheets of trajectories are folded and compressed, as in making flaky pastry. This mixing produces divergence (Figure 1(c)) that quickly makes adjacent motions totally uncorrelated, although both remain on the fractal attractor. This divergence serves to identify chaos: it is quantified by Lyapunov exponent techniques (Guckenheimer & Holmes, 1983).

Escape from a Potential Well

Asymmetric models were used by Hermann Helmholtz to model vibrations in the human ear. An archetypal example introduced by Thompson (1989) is

$$\ddot{x} + \beta\dot{x} + x - x^2 = F \sin \omega t. \qquad (4)$$

This Helmholtz–Thompson equation governs escape from the well,

$$V(x) = \tfrac{1}{2}x^2 - \tfrac{1}{3}x^3. \qquad (5)$$

Such escape is a universal problem in science, from activation energies of molecular dynamics to the gravitational collapse of massive stars. Failures in electrical systems are triggered when the underlying system escapes from a well: if power generators slip out of synchronization, an entire city can be blacked out. Naval research (Thompson, 1997) is directed toward the capsizing of vessels under sinusoidal forcing by ocean waves.

Equation (4) is used by Thompson & Stewart (2002) to illustrate a variety of complex phenomena including chaos, fractal boundaries, and indeterminacies. These arise in a wide class of systems involving nonlinear damping, different well shapes, different direct and parametric forcing, and hardening characteristics. They are not detected by perturbation and averaging methods.

Chaos in Nonlinear Resonance

Three response diagrams for Equation (4) are shown in Figure 2. The quadratic form of the restoring force is technically *softening* for $x > 0$ and *hardening* for $x < 0$, but its overall effect is *softening*. Thus in 2(a), the peaks tilt to the left, as for a cubic Duffing oscillator with

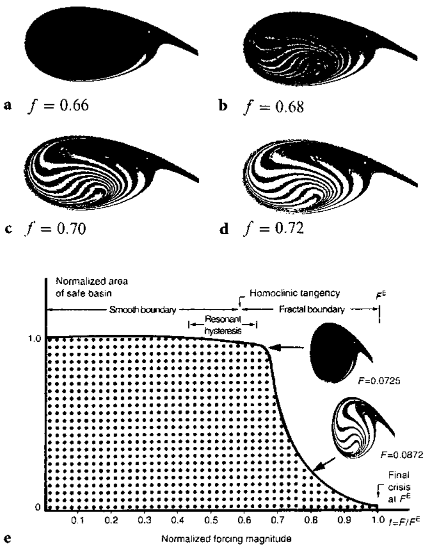

Figure 3. In the space of the coordinates $x(0), \dot{x}(0)$, the four safe basins of attraction, (a)–(d), comprise all those starts that do not give escape: we see how they vary with f, the basin being dramatically eroded by fractal fingers at $f \approx 0.68$. This basin erosion is quantified in (e) by plotting the area of the safe basin against f to give an integrity diagram. Reproduced from Thompson & Stewart (2002) with the permission of John Wiley.

$d < 0$. At the low value of $F = 0.01$, where the small response is almost linear, the untilted peak lies over $\omega = 1$ (the natural frequency). At $F = 0.056$ we see the hysteresis response of a softening oscillator, with a jump to resonance at fold A as ω increases, and a jump from resonance at B as ω decreases. Between these *cyclic folds*, there are three steady periodic solutions with frequency ω: nonresonant attractor, S_{n}; resonant attractor, S_{r}; unstable saddle, D_{r}.

In 2(b), the resonant branch loses stability in a period-doubling cascade to chaos, and the jump at A is indeterminate.

Fractal Boundaries, Indeterminate Bifurcations

Transient motions arise from different starting values of $x(0), \dot{x}(0)$, and Figure 3 shows how the safe basin of attraction varies with $f = F/F^{\mathrm{E}}$. Here F^{E} is the steady-state escape magnitude, at which a slowly evolving system would jump out of the well. The observed fractal structure is generated at a homoclinic tangency. The sudden drop in the integrity diagram at the "Dover cliff" can be used as a design criterion. An engineering system should not be operated at values of $f > 0.68$, even though there are still stable motions within the well up to $f = 1$ (Thompson & Stewart, 2002).

Fractal basin boundaries and their chaotic transients generate great complexities in all forms of Duffing's equation once the response is significantly nonlinear: they are a topic of active research (Stewart et al., 1995). An example is the indeterminate jump from fold A in Figure 2(b). Locally, this is a normal saddle-node fold, but it is located precisely on a fractal boundary. The outcome from such a "tangled saddle-node" is unpredictable (Thompson & Soliman, 1991). Depending sensitively on how A is approached, the jump may settle onto attractor R or escape out of the well; it may also settle onto a co-existing subharmonic motion.

Michael Thompson

See also **Attractors; Chaotic dynamics; Damped-driven anharmonic oscillator; Mel'nikov method; Van der Pol equation**

Further Reading

Duffing, G. 1918. *Erzwungene Schwingungen bei Veränderlicher Eigenfrequenz*, Braunschweig: Vieweg

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, New York: Springer

Hayashi, C. 1964. *Nonlinear Oscillations in Physical Systems*, Princeton, NJ: Princeton University Press

Stewart, H.B., Thompson, J.M.T., Ueda, Y. & Lansbury, A.N. 1995. Optimal escape from potential wells: patterns of regular and chaotic bifurcation. *Physica* D, 85: 259–295

Thompson, J.M.T. 1989. Chaotic phenomena triggering the escape from a potential well. *Proceedings of the Royal Society of London*, A, 421: 195–225

Thompson, J.M.T. 1997. Designing against capsize in beam seas: recent advances and new insights. *Applied Mechanics Reviews*, 50: 307–325

Thompson, J.M.T. & Soliman, M.S. 1991. Indeterminate jumps to resonance from a tangled saddle-node bifurcation. *Proceedings of the Royal Society of London*, A, 432: 101–111

Thompson, J.M.T. & Stewart, H.B. 2002. *Nonlinear Dynamics and Chaos*, 2nd edition, Chichester and New York: Wiley

Ueda, Y. 1980. Steady motions exhibited by Duffing's equation: a picture book of regular and chaotic motions. In *New Approaches to Nonlinear Problems in Dynamics*, edited by P.J. Holmes, Philadelphia: SIAM, pp. 311–322

Ueda, Y. 1992. *The Road to Chaos*, Santa Cruz: Aerial Press

DUNE FORMATION

Dunes are sand formations, found on land, that have heights ranging from 1 to 500 m and have been shaped by the wind. These topographical structures are found typically where large masses of sand have accumulated, which can be in the desert or along the beach; thus, one distinguishes desert dunes and coastal dunes. Dunes can be mobile or fixed. Fixed dunes are older and are either "fossilized," that is, transformed into a cohesive material, a precursor to sandstone, or fixed because of

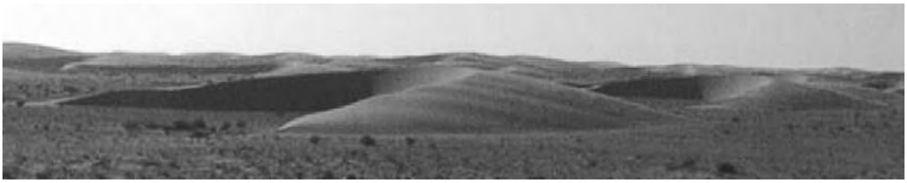

Figure 1. Barchan dunes near Laâyoune, Morocco.

the vegetation or because the average wind over some period at their location is zero. Otherwise the sand moves if the winds are strong enough, which means typically stronger than 4 ms^{-1}.

The beautiful landscapes (Figure 1; see also color plate section) formed by dunes are characterized by very gentle hills interrupted by sharp edges called brink lines, delimiting regions of steeper slope, called slip-faces, lying in the wind shadow. Depending on the amount of available sand and the variation of the wind direction, one distinguishes different typical dune morphologies that have been classified by geographers into over a hundred categories. The most well known are longitudinal, transverse, and barchan dunes. Barchans (from an Arabic word) are crescent-shaped mobile dunes that appear when the wind always comes from the same direction and there is not much sand present. Their movement ranges from 5 to 50 m per year and is inversely proportional to their height. If sufficient sand is available to cover all the surface, then transverse dunes appear (from the merger of many barchans). Longitudinal dunes, that is, along the direction of the wind, are observed when the wind periodically changes its direction over about 30°. Other famous dune types are star dunes, ergs, parabolic dunes, and draas.

The driving force for sand motion is the drag imposed by the wind on the grains at the surface. A given dune morphology can therefore be understood as an aerodynamic instability close to a mobile surface. A complete mathematical description of the problem therefore needs the equation of motion of the wind velocity field coupled with an equation of motion of the granular surface. The right formulation of these equations requires a good understanding of the transport mechanism of sand.

Three types of transport can be distinguished according to the size of a sand grain: creep, saltation (bouncing), and suspension. The only mechanism relevant for dune formation is saltation, which drags grains typically of 100–300 µm diameter. The mechanism of saltation, first described by Ralph Bagnold in his pioneering work (Bagnold, 1941), consists of grains, which once lifted out from their granular bed are accelerated by the wind and then impact against the surface, ejecting new grains. These grains are again accelerated and eject a further splash of grains, so that the number of grains flying above the surface increases exponentially until the total momentum transferred from the air to the grains saturates to its maximum capacity, and the wind can no longer pick up more sand from the dune. On dunes in the field, these saltating grains form a sheet of grains floating typically 5 cm above the surface. The wind typically is turbulent and has a logarithmic profile as a function of height, which (due to the presence of the grains) is strongly modified close to the surface. The height of the boundary layer is less than 1 cm. Using the techniques of Jackson & Hunt (1975), one can calculate the shear force of the wind at the surface in an approximate form and obtain reasonable agreement with measurements on dunes. The saturated flux of sand at the surface is a function of this shear stress and has been described by various phenomenological expressions, the first one given by Bagnold (1941) and subsequent ones by Lettau & Lettau (1969), and by Sørensen (1981). A full description also requires taking into account the transient length before (or after) reaching saturation. Together with mass conservation one can then close the system of equations, giving at the end a full set of equations of motions (Sauermann et al., 2001).

When the local slope exceeds a value of typically 35°, the angle of repose, the sand begins to slide in the form of avalanches giving a second mechanism of sand transport driven by gravity. The slip-faces all have this slope. The edges separating them from the purely wind-driven regions are just given by the brink lines. Over these regions the wind field develops recirculation eddies of velocities typically below the minimum threshold for grain motion. When the saturation length is less than the size of these low-velocity regions, the sand grains get trapped. This is the principal instability underlying dune morphology: the dunes become traps of sand for more sand. The typical saturation length of about 10 m also means that no dune below 1.5 m in height is stable under Earth's meteorological conditions; the dune is too short for the air to reach its saturation point, and the dune suffers erosion.

Dunes have been studied on all continents and their shape, sand flux, velocity, and granulometry have been presented in many publications. Several books review the subject (Pye & Tsoar, 1990). For different morphologies one finds specific shapes and scaling laws, but systematic studies exist only for barchans. Sand fluxes are typically measured with traps, but a more sophisticated metrology (e.g., acoustic, optic) is evolving. The limitation factors are the fluctuations of the wind fields and the climate. In the arid regions of the world, in particular the poor countries in the Sahara, dune motion poses an important threat to housing, roads, and fields, and sand removal constitutes a significant economic factor in these countries. Many empirical techniques of dune fixing and dune destruction have been developed, mostly applied to coastal dunes, which are in fact disappearing in many places, sometimes damaging the fragile dune ecosystem.

H.J. Herrmann

See also **Avalanches; Geomorphology and tectonics; Sandpile model**

Further Reading

Bagnold, R.A. 1941. *The Physics of Blown Sand and Desert Dunes*, London: Methuen

Hersen, P., Douady, S. & Andreotti, B. 2002. Relevant length scale of barchan dunes. *Physical Review Letters*. 89: 264301

Jackson, P. S. & Hunt, J.C.R. 1975. Turbulent wind flow over a low hill. *Quarterly Journal of the Royal Meteorological Society*, 101–929

Lettau, K. & Lettau, H. 1969. Bulk transport of sand by the barchans of the Pampa de La Joya in Southern Peru. Zeitschrift für Geomorphologie, N.F. 13(2): 182–195

Pye, K. & Tsoar, H. 1990. *Aeolian Sand and Sand Dunes*, London: Unwin Hyman

Sauermann, G., Kroy, K. & Herrmann, H.J. 2001. A continuum saltation model for sand dunes. *Physical Review* E, 64: 31305

Sørensen, M. 1991. An analytic model of wind-blown sand transport. *Acta Mechanica*, (Suppl.) 1: 67–81

DYM EQUATION

See **Solitons, types of**

DYNAMIC PATTERN FORMATION

See **Synergetics**

DYNAMIC SCALING FUNCTION

See **Routes to chaos**

DYNAMICAL SYSTEMS

A dynamical system is a time-dependent, multicomponent system of elements with local states determining a global state of the whole system. In a planetary system, for example, the state of the system at a certain time is the set of values that completely describe the system (the position and momentum of each planet). The states can also refer to moving molecules in a gas, excitation of neurons in a neural network, nutrition of organisms in an ecological system, supply and demand of economic markets, or behavior of social groups in human societies.

The dynamics of a system (the change of system states with time) is given by linear or nonlinear differential equations. In the case of nonlinearity, several feedback activities take place between the elements of the system: in the solar system, the movement of the Earth is determined by the gravitation not only of the Sun, but of all the other celestial bodies of the system, which attract each other gravitationally.

For deterministic processes (for example, movements in a planetary system), each future state is uniquely determined by the present state. A conservative (Hamiltonian) system such as an ideal pendulum is characterized by the reversibility of time direction and conservation of energy. Dissipative systems, for example, a real pendulum with friction, are irreversible. The time-dependent development of a system's state is geometrically represented by orbits (trajectories) in a state space or phase space, which is defined by the multidimensional vectors of the nonlinear system (*See* **Phase space**).

In addition to continuous processes, we can also consider discrete processes of changing states at certain points of time. Difference equations are important for modeling measured data at discrete points of time, which are chosen equidistant or defined by other measurement devices.

Random events (Brownian motion in a fluid, mutation in evolution) are represented by additional fluctuation terms. Classical stochastic processes, such as the billions of unknown molecular states in a fluid, are defined by time-dependent differential equations with distribution functions of probabilistic states. Stochastic nonlinear differential equations (such as Fokker–Planck equation, Master equation) are also used to model phase transitions of complex systems, including migration dynamics of populations, traffic dynamics, data dynamics in the Internet, among others.

In quantum systems, the dynamics of quantum states are determined by Schrödinger's equation. Although it is a deterministic differential equation of a wave function, its observables (position and momentum of a particle) depend on Heisenberg's uncertainty principle, which only allows probabilistic forecasts.

During the 17th–19th centuries, classical physics viewed the universe as a deterministic and conservative system. The astronomer and mathematician Pierre-Simon Laplace assumed that all future states of the universe could be computed or determined if all forces acting in nature and the initial states of all celestial bodies are known at one instant of time (Laplacian determinism). Laplace's assumption was correct for linear and conservative dynamical systems such as a harmonic oscillator. However, at the end of the 19th century, Henri Poincaré discovered that celestial mechanics is not a completely computable system even if it is considered as a deterministic and conservative system. The mutual gravitational interactions of more than two celestial bodies (the many-body problem) correspond to nonlinear and non-integrable equations with instabilities and sometimes chaos (*See* ***N*-body problem**).

According to Laplacian determinism, similar causes effectively determine similar effects. Thus, in the phase space, trajectories starting close to each other also remain close to each other during time evolution. Dynamical systems with deterministic chaos exhibit an exponential dependence on initial conditions for bounded orbits: the separation of trajectories with close initial states increases exponentially. (The rate at which nearby orbits diverge from each other after small perturbations is measured by Lyapunov exponents.) Consequently, tiny deviations of initial states lead to exponentially increasing computational efforts for future data, limiting long-term prediction, although the dynamics is, in principle, uniquely determined.

The sensitivity of chaotic dynamics to small changes in initial conditions is known as the "butterfly effect": small and local causes (local perturbations of the weather, for example) can lead to unpredictable large and global effects in unstable states (*See* **Butterfly effect**). According to the famous KAM theorem of Andrei Kolmogorov, Vladimir Arnol'd, and Jürgen Moser, trajectories in the phase space of classical mechanics are neither completely regular nor completely irregular, but depend sensitively on the chosen initial conditions. Changes of states in a dynamical system that change the stability of solutions to its nonlinear equations are associated with bifurcations of orbits in the corresponding phase space (*See* **Bifurcations**).

Dynamical systems can be classified by the effects of their dynamics on a region of the phase space. A conservative system is defined by the fact that, during time evolution, the volume of a region remains constant, although its shape may be transformed. An attractor is a region of a phase space into which all trajectories departing from an adjacent region, the so-called basin of attraction, converge. There are different kinds of more or less complex attractors. Chaotic attractors are highly complex structures of nonperiodic orbits in a bounded region of the phase space between regular behavior and stochastic behavior or noise. Although high-dimensional dynamical systems (such as the stock market, the economy, and society) cannot be formulated in all detail, approximate models may provide qualitative insights into their complex behavior.

KLAUS MAINZER

See also **Chaotic dynamics; Determinism; Equations, nonlinear**

Further Reading

Abarbanel, H.D.I. 1996. *Analysis of Observed Chaotic Data*, New York: Springer

Abraham, R.H. & Shaw, C.D. 1992. *Dynamics: The Geometry of Behavior*, 2nd edition, Redwood City, CA: Addison-Wesley

Arnol'd, V.I. 1978. *Mathematical Methods of Classical Mechanics*, Berlin and New York: Springer

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, New York: Springer

Hirsch, M.W. & Smale, S. 1974. *Differential Equations, Dynamical Systems, and Linear Algebra*. New York: Academic Press

Kaplan, D. & Glass, L. 1995. *Understanding Nonlinear Dynamics*, New York: Springer

Mainzer, K. 2003. *Thinking in Complexity. The Computational Dynamics of Matter, Mind, and Mankind*, 4th edition, New York: Springer

Rand, D.A. & Young, L.S. (editors). 1981. *Dynamical Systems and Turbulence*, Berlin and New York: Springer

Shilnikov, L.P., Shilnikov, A.L., Turaev, D.V. & Chua, L.O. 2001. *Methods of Qualitative Theory of Nonlinear Dynamics*, Singapore: World Scientific

DYNAMICAL ZETA FUNCTIONS

See **Periodic orbit theory**

DYNAMOS, HOMOGENEOUS

Dynamos convert mechanical energy into electromagnetic energy, the most familiar example being the bicycle dynamo. Nearly all electricity consumed by mankind is generated by dynamos in power plants—electrochemical processes also generate electric power, for example, in batteries, but these play only a minor role. The simplest dynamo is the disk dynamo originally invented by Faraday and shown in Figure 1. A metal disk rotates about its axis with an angular velocity $\boldsymbol{\Omega} = \boldsymbol{k}\Omega$. When the disk is permeated by an initial magnetic field $\boldsymbol{B}_0$ parallel to the axis, an electromotive force is generated between the axis and rim of the disk,

$$U = \boldsymbol{\Omega} \cdot \int_{r_1}^{r_0} \boldsymbol{B}_0 \mathrm{d}r, \qquad (1)$$

where r_1 and r_0 denote the radii of axis and disk. The electromotive force U can be used to drive an electric current J through the circuit indicated in Figure 1. Denoting by L and R the inductivity and the ohmic resistance of the circuit, we obtain for the current J

$$L\frac{\mathrm{d}J}{\mathrm{d}t} + RJ = U. \qquad (2)$$

The current J flowing through the circuit is associated with a magnetic field $\boldsymbol{B}_1$, which may replace the initial field $\boldsymbol{B}_0$. The integral $2\pi \int_{r_1}^{r_0} \boldsymbol{k} \cdot \boldsymbol{B}_1 \mathrm{d}r/J$ describes

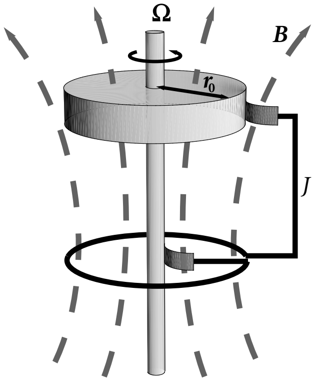

Figure 1. The disk dynamo.

the mutual inductivity M between circuit and disk. Equation (2) for the self-exited disk dynamo can thus be written in the form

$$L\frac{\mathrm{d}J}{\mathrm{d}t} = (\Omega M/2\pi - R)J, \tag{3}$$

which allows for exponentially growing solutions once the dynamo condition

$$\Omega > 2\pi R/M \tag{4}$$

is satisfied.

The disk dynamo is an inhomogeneous dynamo since it depends on an inhomogeneous distribution of electrical conductivity as given by the wiring of the circuit. The dynamo would not work if the sense of wiring around the axis would be opposite to that shown in the figure or if it would be short-circuited by immersion in a highly conducting fluid. While it is generally believed that planetary and stellar magnetic fields are generated by dynamos in the electrically conducting interiors of these celestial bodies, these dynamos must be homogeneous ones because they operate in singly connected finite domains of essentially homogeneous electrical conductivity. In 1919, Larmor first proposed this idea as an explanation for the magnetic field of sunspots. It was doubtful for a long time whether homogeneous dynamos were possible. Cowling proved in 1934 that axisymmetric fields could not be generated by a homogeneous dynamo. But, in 1958, Backus and Herzenberg independently demonstrated in a mathematically convincing way that homogeneous dynamos are indeed possible.

The velocity fields that are required to drive a homogeneous dynamo are necessarily more complex than the simple rotation velocity of the disk dynamo. A new mathematical discipline called dynamo theory has evolved and continues to be an active field of research. One distinguishes between the kinematic dynamo problem based on a linear equation such as (3) with a prescribed velocity field and the magnetohydrodynamic dynamo problem in which the influence of the growing magnetic field on the velocity field is taken into account. Obviously, a magnetic field cannot grow exponentially forever. Just as the Lorentz force produced by the magnetic field together with the current density in the disk opposes the rotation of the disk in Figure 1, it also changes the velocity field of the magnetohydrodynamic dynamo. The external torque T_e applied to the disk in the figure must be increased in the presence of dynamo action in order to sustain the rotation rate Ω. The equilibrium strength of the magnetic field will thus be determined by the available torque T_e.

In the case of the geodynamo driven by convection flows in the liquid iron core of the Earth, as well as in the case of other planetary or stellar dynamos, the equations of motions together with the equation of magnetic induction (*See* **Magnetohydrodynamics**) must be solved to determine the strength of the generated magnetic field as a function of the relevant parameters. Extensive computer simulations have been performed in recent years. Some examples can be found in Jones et al. (2003). The complex numerical simulations of magnetohydrodynamic dynamos share the following properties with the simple disk dynamo:

(i) For given external parameters, there always exists a solution without magnetic field besides the dynamo solution, just as the disk of Figure 1 can rotate with $\Omega > \Omega_\mathrm{c}$ in the absence of any initial field $\boldsymbol{B}_0$.

(ii) The existence of a dynamo solution requires that the magnetic Reynolds number Rm exceeds a critical value Rm_c. Rm is defined by

$$Rm = V\mathrm{d}\sigma\mu, \tag{5}$$

where V is a typical velocity, d is a characteristic length such as the radius of the iron core in the case of the Earth, and σ and μ are the electrical conductivity and the magnetic permeability of the fluid. The inverse product $\lambda = 1/\sigma\mu$ is called the magnetic diffusivity. The dimensionless parameter Rm corresponds to the quantity $\Omega M/R$ in the case of the disk dynamo. The nonmagnetic solution exists, but it is unstable for $Rm > Rm_\mathrm{c}$.

(iii) Magnetohydrodynamic dynamos exist in two forms that are identical except for the sign of the magnetic field $\boldsymbol{B}$. This property reflects the fact that the Lorentz force is given by an expression quadratic in $\boldsymbol{B}$.

Property (iii) is the basic reason that the geomagnetic field has often switched its polarity in the geologic past. These "reversals" have occurred nearly randomly about every 200,000 years on average. In contrast, the solar magnetic field reverses every 11 years in a surprisingly periodic fashion.

For the description of magnetic fields associated with a spherical system, one often uses the general representation for a solenoidal vector field,

$$\boldsymbol{B} = \nabla \times (\nabla h \times \boldsymbol{r}) + \nabla g \times \boldsymbol{r}, \qquad (6)$$

in terms of a poloidal and a toroidal component each of which is described by a scalar function, the poloidal function h, and the toroidal function g. Without loss of generality, the averages of h and g over surfaces $|\boldsymbol{r}| = \text{constant}$ can be assumed to vanish. A homogeneous dynamo usually requires the interaction of both components of the magnetic field. It can be shown (Kaiser et al., 1994) that a magnetic field with vanishing toroidal part cannot be generated. It is also generally believed that a purely poloidal field cannot be generated either. But a proof of this hypothesis has not yet been given. The functions h and g can be separated into their axisymmetric parts $\bar{h}$ and $\bar{g}$ and non-axisymmetric parts, $\check{h} = h - \bar{h}$ and $\check{g} = g - \bar{g}$. The component $\bar{g}$ can easily be generated in a spherical dynamo through a stretching of the axisymmetric poloidal field by a differential rotation. This process is known as the ω-effect. The amplification of $\bar{h}$ requires the interaction of the non-axisymmetric components of magnetic fields and velocity fields. This is often called the α-effect. This latter effect can, of course, also be used for the generation of $\bar{g}$ in the absence of a differential rotation. Accordingly, one distinguishes between $\alpha\omega$- and α^2-dynamos. These concepts were originally introduced within the framework of mean-field magnetohydrodynamics for which the reader is referred to the book by Krause & Raedler (1980).

F.H. BUSSE

See also **Alfvén waves; Fluid dynamics; Magnetohydrodynamics; Nonlinear plasma waves**

Further Reading

Childress, S. & Gilbert, A.P. 1995. *Stretch, Twist, Fold: The Fast Dynamo*, Berlin and New York: Springer

Davidson, P.A. 2001. *An Introduction to Magnetohydrodynamics*, Cambridge and New York: Cambridge University Press

Jones, C.A., Soward, A.M. & Zhang, K. 2003. *Earth's Core and Lower Mantle*, London and New York: Taylor & Francis

Kaiser, R., Schmitt, B.J. & Busse, F.H. 1994. On the invisible dynamo. *Geophysical and Astrophysical Fluid Dynamics*, 77: 91–109

Krause, F. & Raedler, K.-H. 1980. *Mean-field Magnetohydrodynamics and Dynamo Theory*, Oxford and New York: Pergamon Press

Moffatt, H.K. 1978. *Magnetic Field Generation in Electrically Conducting Fluids*, Cambridge and New York: Cambridge University Press

E

EARTHQUAKES

See **Geomorphology and tectonics**

ECKHAUS INSTABILITY

See **Wave stability and instability**

ECONOMIC SYSTEM DYNAMICS

Economic dynamics is concerned with fluctuations in the economy. Most economic variables, such as gross domestic product (GDP), production, unemployment, interest rates, exchange rates, and stock prices, exhibit perpetual fluctuations over time. These fluctuations are characterized by sustained growth of production and employment as well as large oscillations in relative changes or growth rates. The fluctuations vary from fairly regular business cycles in macroeconomic variables to very irregular fluctuations, for example, in stock prices and exchange rates, in financial markets. In this note, we discuss some approaches to the theory of economic fluctuations, emphasizing the role of nonlinear dynamic models.

In contrast to many dynamic phenomena in natural sciences, uncertainty always plays a role in an economy, at least to some extent. Therefore, a purely deterministic model seems inappropriate to describe fluctuations in the economy, and a stochastic dynamic model is needed. Nevertheless, a key question in economic dynamics is whether a simple, nonlinear dynamic model can explain a *significant part* of observed economic fluctuations.

Brief History

There are two contrasting viewpoints concerning the explanation of observed economic fluctuations. According to the first (New Classical) viewpoint, the main source of fluctuations is to be found in exogenous, random shocks (news about economic fundamentals) to an inherently stable, often linear economic system. Without any external shocks to economic fundamentals (preferences, endowments, technology, etc.), the economy would be stable and converge to the unique steady-state (growth) path. According to the second, opposing (Keynesian) viewpoint, economic fluctuations are not caused by chance or random impulses, but should be explained by nonlinear economic laws of motion. Even without any external shocks to the fundamentals of the economy, fluctuations in prices or other economic variables may arise. It is an old Keynesian theme that fluctuations are not determined by economic fundamentals only, but are also driven by volatile, self-fulfilling expectations (“animal spirits,” market psychology).

The view that business cycles are driven by external random shocks was propagated in the 1930s, for example, by Ragnar Frisch and Jan Tinbergen (sharing the first Nobel Prize in Economic Sciences in 1969 “for having developed and applied dynamic models for the analysis of economic processes”). They observed that simple, linear systems buffeted with noise can mimic time series similar to those observed in real business cycle data. To several economists this approach was unsatisfactory, however, because it does not provide an *economic* explanation of business cycles, but rather attributes them to external, random events. In the 1940s and 1950s Nicholas Kaldor, John Hicks, and Richard Goodwin developed nonlinear dynamic models with locally unstable steady states and stable limit cycles as an explanation for business cycles. These early nonlinear business cycle models, however, suffered from a number of serious shortcomings. First of all, the laws of motion were too “ad hoc,” and in particular they were *not* derived from rational behavior, that is, from utility and profit maximizing principles. Secondly, the simulated time series from the models were too regular compared with observed business cycles, even when small dynamic noise was added to the models. Finally, expectation rules were “ad hoc,” and along the regular cycles, agents made “systematic” forecasting errors.

The Role of Expectations

The most important difference between economics and natural sciences is perhaps the fact that an economic system is an *expectations feedback* system. Decisions of economic agents are based upon their expectations and beliefs about the future state of the economy. Through these decisions, expectations feed back into the economy and affect actual realization of economic variables. These realizations lead to new expectations, in turn affecting new realizations, implying an infinite sequence of expectational feedback. For example, in the stock market, optimistic expectations that stock prices will rise will lead to a larger demand for stocks, which will cause stock prices to rise. This process may lead to a self-fulfilling speculative bubble in the stock market. A theory of expectation formation is, therefore, a crucial part of economics, in particular for modeling dynamic asset markets.

In the early business cycle models, simple, ad hoc expectations rules were employed, such as naive expectations (where the forecast of the economic variable is simply the latest observation of that variable) or adaptive expectations (where the forecast is a weighted average of the previous forecast and the latest observation). An important problem with simple forecasting rules is that typically agents make systematic forecasting errors, especially when there are regular cycles. A smart agent would learn from her mistakes and adapt her expectations rule accordingly. Another problem is that if an agent is to use a simple forecasting rule, it is far from clear which simple rule to choose in a particular model. With the development of empirical, econometric analysis of business cycles, it became clear that unrestricted models of expectations preclude a systematic inquiry into business fluctuations. These considerations led to the development of *rational expectations*, a solution to the expectations feedback system proposed by John Muth (1961) and applied to macroeconomics, for example, by Robert Lucas and Thomas Sargent. Rational expectations means that agents use all available information, including economic theory, to form optimal forecasts and that, on average, expectations coincide with realizations. In a deterministic model, without noise and randomness, rational expectations implies perfect foresight (no mistakes at all); in a stochastic model, rational expectations coincides with the conditional mathematical expectations given all available information (no mistakes on average, no systematic bias).

In the 1970s and 1980s, the rational expectations critique culminated in the development of New Classical economics and real business cycle models, based upon rational expectations, intertemporal utility and profit maximization, and perfectly competitive markets. This approach outdated the early Keynesian nonlinear business cycle models of the 1950s. Due to the discovery of deterministic chaos and other developments in nonlinear dynamics, however, the last two decades have witnessed a strong revival of interest in nonlinear endogenous business cycle models.

Nonlinear Dynamics

In mathematics and physics, things changed dramatically in the 1970s due to the discovery of deterministic chaos, the phenomenon that simple, deterministic laws of motion can generate unpredictable time series. This discovery shattered the Laplacian deterministic view of perfect predictability and made scientists realize that long-run prediction may be fundamentally impossible, even when laws of motion are known exactly. Inspired by "chaos theory," economists (e.g., Richard Day and Jean-Michel Grandmont) started looking for nonlinear, deterministic models generating erratic time series similar to the patterns observed in real business cycles. This search led to new, simple nonlinear business cycle models within the paradigm of rational expectations, optimizing behavior and perfectly competitive markets, generating chaotic business fluctuations.

In the 1980s, several economists (e.g., William Brock, Davis Dechert, Jose Scheinkman, and Blake LeBaron) also employed nonlinear methods, such as correlation dimension tests, from the natural sciences to look for evidence of nonlinearity and low deterministic chaos in economic and financial data. This turned out to be a difficult task because the methods employed require very long time series and the methods are very sensitive to noise. One can say that evidence for low-dimensional deterministic chaos in economic and financial data is weak (but it seems fair to add that because of the sensitivity to noise, the hypothesis of chaos buffeted with dynamic noise has not been rejected) but evidence for nonlinearity is strong. In particular, Brock, Dechert, and Scheinkman have developed a general test (the BDS-test), based upon ideas from U-statistics theory and correlation integrals, to test for nonlinearity in a given time series; see Brock et al. (1996) and Brock, Hsieh, & LeBaron (1991) for the basic theory, references, applications, and extensions. The BDS test has become widely used in economics and also in physics.

Bounded Rationality

Already in the 1950s, Herbert Simon pointed out that rationality requires unrealistically strong assumptions about the computing abilities of agents and proposed that *bounded rationality*, with limited computing capabilities and with agents using habitual rules of thumb instead of perfectly optimal decision rules, would be a more accurate description of human behavior. Nevertheless, as noted above, rational expectations became the dominating paradigm in dynamic economics in the 1970s and 1980s. Nonlinear dynamics, the possibility of chaos, and its implications for limited predictability shed important new light on the expectations hypothesis, however. In a simple (linear) stable economy with a unique steady-state path, it seems natural that agents can learn to have rational expectations, at least in the long run. A representative, perfectly rational agent model nicely fits into a linear view of a globally stable and predictable economy. But how could agents have rational expectations or perfect foresight in a complex, nonlinear world, with prices and quantities moving irregularly on a strange attractor and sensitivity to initial conditions? A boundedly rational world view with agents using simple forecasting strategies, perhaps not perfect but at least approximately right, seems more appropriate for a complex nonlinear world. These developments contributed to a rapidly growing interest in bounded rationality in the 1990s (see, e.g., the survey in Sargent (1993)). A boundedly rational agent forms expectations based upon observable quantities and adapts her forecasting rule as additional observations become available. Adaptive learning may converge to a rational expectations equilibrium, or it may converge to an "approximate rational expectations equilibrium," where there is at least some degree of consistency between expectations and realizations (see, e.g., Evans & Honkapohja (2001) for an extensive and modern treatment of adaptive learning in macroeconomics).

Interacting Agents and Evolutionary Models

The representative agent model has played a key role in economics for a long time. An important motivation for the dominance of the rational agent model dates back to the 1950s, to Milton Friedman who argued that nonrational agents will be driven out of the market by rational agents, who will trade against them and simply earn higher profits. In recent years, however, this view has been challenged, and heterogeneous agent models are becoming increasingly popular, especially in financial market modeling (see, e.g., Kirman (1992) for a critique on representative agent modeling).

Many heterogeneous agent models are artificial, computer simulated markets. This work views the economy as a complex evolving system composed of many different, boundedly rational, interacting traders, with strategies, expectations and realizations co-evolving over time (see, e.g., work at the Santa Fe Institute collected in Anderson et al. (1988)). Two typical trader types arising in many heterogeneous agent financial market models are fundamentalists and chartist or technical traders. Fundamentalists base their investment decisions upon market fundamentals such as dividends, earnings, interest rates, or growth indicators. In contrast, technical traders pay no attention to economic fundamentals but look for regular patterns in past prices and base their investment decision upon simple trend following trading rules. An evolutionary competition between these different trader types, where traders tend to follow strategies that have performed well in the recent past, may lead to irregular switching between the different strategies and result in complicated, irregular asset price fluctuations. It has been shown, for example, by Brock and Hommes (1997, 1998), that in these evolutionary systems, rational agents and/or fundamental traders do not necessarily drive out all other trader types, but that the market may be characterized by perpetual evolutionary switching between competing trading strategies. Nonrational traders may survive evolutionary competition in the market (see, e.g., Hommes (2001) for a survey and many relevant references). Lux & Marchesi (1999) show that these types of interacting agent models are able to generate many of the stylized facts, such as unpredictable returns, clustered volatility, fat tails, and long memory, observed in real financial markets.

Future Perspective

A good feature of the rationality hypothesis is that it puts natural discipline on agents' forecasting rules and minimizes the number of free parameters in dynamic economic models. In contrast, the "wilderness of bounded rationality" leaves too many degrees of freedom in modeling, and it is far from clear which out of a large class of habitual rules of thumb is most reasonable. Stated differently in a popular phrase: "there is only one way (or perhaps a few ways) one can be right, but there are many ways one can be wrong." The philosophy underlying the evolutionary approach is to use simple forecasting rules based upon their performance in the recent past. In this type of modeling, "evolution decides who is right." Bounded rationality, heterogeneity, adaptive learning, and evolutionary competition all create natural nonlinearities. Nonlinearity is, therefore, likely to play an increasingly important role in the future of economic dynamics.

CARS HOMMES

See also **Dynamical systems; Forecasting; Game theory; Time series analysis**

Further Reading

Anderson, P.W., Arrow, K.J. & Pines, D. (editors). 1988. *The Economy as a Complex Evolving System*, Redwood City, CA: Addison-Wesley

Brock, W.A. Dechert, W.D., Scheinkman, J.A. & LeBaron, B. 1996. A test for independence based upon the correlation dimension. *Econometric Review*, 15: 197–235

Brock, W.A. & Hommes, C.H. 1997. A rational route to randomness. *Econometrica*, 65: 1059–1095

Brock, W.A. & Hommes, C.H. 1998. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control*, 22: 1235–1274

Brock, W.A., Hsieh, D.A. & LeBaron, B. 1991. *Nonlinear Dynamics, Chaos and Instability: Statistical Theory and Economic Evidence*, Cambridge, MA: MIT Press

Day, R.H. 1996. *Complex Economic Systems*, Cambridge, MA: MIT Press

DeGrauwe, P., DeWachter, H. & Embrechts, M. 1993. *Exchange Rate Theory. Chaotic Models of Foreign Exchange Markets*, Oxford: Blackwell

Evans, G.W. & Honkapohja, S. 2001. *Learning in Macroeconomics*, Princeton, NJ: Princeton University Press

Friedman, M. 1953. The case of flexible exchange rates. In *Essays in Positive Economics*, Chicago: University of Chicago Press

Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic essays in honor of Gustav Cassel*, London: George Allen and Unwin, Ltd; reprinted in *Readings in Business Cycles*, edited by R.A. Gordon and L.R. Klein, Homewook, IL: Richard D. Irwin, Inc, 1965

Gabisch, G. & Lorenz, H.W. 1987. *Business Cycle Theory. A Survey of Methods and Concepts.* Berlin: Springer

Goodwin, R.M. 1951. The nonlinear accelerator and the persistence of business cycles. *Econometrica*, 19: 1–17

Grandmont, J.M. 1985. On endogenous competitive business cycles. *Econometrica*, 53: 995–1045

Hicks, J.R. 1950. *A contribution to the theory of the trade cycle*. Oxford: Clarendon Press

Hommes, C.H. 2001. Financial markets as nonlinear adaptive evolutionary systems. *Quantitative Finance*, 1: 149–167

Kaldor, N. 1940. A model of the trade cycle. *Economic Journal*, 50: 78–92

Kirman, A. 1992. Whom or what does the representative individual represent? *Journal of Economic Perspectives*, 6: 117–136

Lorenz, H.W. 1993. *Nonlinear Dynamical Economics and Chaotic Motion*, 2nd edition, Berlin: Springer

Lucas, R.E. 1971. Econometric testing of the natural rate hypothesis. In *The Econometrics of Price Determination*, edited by O. Eckstein, Washington, DC: Board of Governors of the Federal Reserve System and Social Science Research Council

Lux, T. & Marchesi, M. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397: 498–500

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica*, 29: 315–335

Sargent, T.J. 1993. *Bounded Rationality in Macroeconomics*, Oxford: Clarendon Press

Simon, H.A. 1957. *Models of Man*, New York: Wiley

EFFECTIVE MASS

Effective mass is a physical quantity characterizing the dynamics of a particle or quasiparticle with an energy ($\mathcal{E}$) that is quadratic in the components of the momentum ($\boldsymbol{p}$). With a dispersion relations of the form

$$\mathcal{E} = \mathcal{E}_0 + \tfrac{1}{2}\,\mu_{ik}\,p_i p_k\,, \quad \mathcal{E}_0 = \text{const.}, \tag{1}$$

the tensor $\mu_{ik} = m_{ik}^{-1}$ is a tensor of reciprocal effective masses, and m_{ik} is an effective mass tensor. Considering Equation (1) as the Hamiltonian function of a free particle, one can determine a particle velocity using a canonical equation of motion

$$v_i = \partial\mathcal{E}/\partial p_i, \quad i = 1, 2, 3. \tag{2}$$

From the Hamiltonian function (1) and relations (2) a Lagrangian function of the particle is

$$L = L_0 + \tfrac{1}{2}\,m_{ik}\,v_i v_k, \quad L_0 = \text{const.} \tag{3}$$

As the Lagrangian function of a free particle coincides with its kinetic energy, the effective mass tensor can be associated with kinetic energy (3), which is quadratic in components of the velocity.

Having its origin in the mechanics of particles, the definition of the effective mass can be connected with the dynamics of wave packets. According to the de Broglie principle of wave-corpuscular dualism, the energy $\mathcal{E}$ and momentum $\boldsymbol{p}$ of a particle correspond to the frequency ω and wave vector $\boldsymbol{k}$ of some wave packet, as $\mathcal{E} = \hbar\omega$ and $\boldsymbol{p} = \hbar\boldsymbol{k}$ where $\hbar$ is Planck's constant. From this point of view, any quasiparticle excitation in condensed matter behaves as a particle-like wave packet, and Equation (2) coincides with the definition of a group velocity of the wave packet.

The effective mass tensor for a free Newtonian particle or quasiparticle in an isotropic media (for example, excitations in the superfluid liquid He) has the simple form: $m_{ik} = m\,\delta_{ik}$, $i, k = 1, 2, 3$.

The dispersion relation of type (1) for the quasiparticles described by band theory in a periodic structure (electrons in metals and phonons in crystals) takes place at vicinities of singular points in the **p**-space. These are at the points where energy $\mathcal{E}(\boldsymbol{p})$ has a minimum (then μ_{ik} is positively definite), at the points near the maximum of $\mathcal{E}(\boldsymbol{p})$ (then μ_{ik} is negatively definite), and at the so-called conical points, when the main values of the tensor μ_{ik} have different signs.

In the general band theory of electrons and semiconductors, energy is a more complicated (arbitrary) function of the momentum $\mathcal{E} = \mathcal{E}(\boldsymbol{p})$, and the tensor of the reciprocal effective mass is defined as

$$\mu_{ik} = \frac{\partial^2\mathcal{E}}{\partial p_i\,\partial p_k}, \quad i, k = 1, 2, 3, \tag{4}$$

which can be a function of the momentum. The effective mass tensor allows one to calculate the acceleration of a particle under the action of the external force $\boldsymbol{f}$. As an equation of the particle motion is

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \boldsymbol{f}, \tag{5}$$

Equations (2) and (5) lead to the following equation for the acceleration:

$$m_{ik}\,\frac{\mathrm{d}v_k}{\mathrm{d}t} = f_i, \quad i = 1, 2, 3. \tag{6}$$

The force $\boldsymbol{f}$ is determined, for example, by the electric field effect on a charged particle.

The effective mass has another definition for an electron (or a charged particle) moving in a static magnetic field $\boldsymbol{B}$. In such a case the force $\boldsymbol{f}$ is a Lorentz force, and Equation (5) implies that the electron moves under the following conditions: $\mathcal{E}(\boldsymbol{p}) =$ constant and $p_B =$ constant, where p_B is a projection of the momentum on the magnetic field direction. Thus, an electron trajectory in $\boldsymbol{p}$-space is a section of the isoenergy surface $\mathcal{E}(\boldsymbol{p}) = \mathcal{E}$ with the plane $p_B = p$. If this section is a closed curve and has a sectional area $S(\mathcal{E}, p)$, the electron motion is periodic in time and characterized by the "cyclotron frequency" $\omega_c = eB/(m^*c)$, where e is the electron charge, and c is the velocity of light. Thus, the effective mass m^* is equal to

$$m^* = \frac{1}{2\pi}\frac{\partial S}{\partial \mathcal{E}}. \tag{7}$$

Cyclotron resonance is the most convenient experimental method for measuring the effective mass defined in Equation (6).

ARNOLD KOSEVICH

See also **Dispersion relations; Group velocity; Wave packets, linear and nonlinear**

Further Reading

Haken, H. 1976. *Quantum Field Theory of Solids*, Amsterdam: North-Holland

Kittel, C. 1987. *Quantum Theory of Solids*, New York: Wiley

Slater, J.C. 1951. *Quantum Theory of Matter*, New York: McGraw-Hill

EIFFEL JUNCTION

See **Long Josephson junction**

EIGENVALUES AND EIGENVECTORS (BOUND STATE)

See **Inverse scattering method or transform**

EIKONAL CURVATURE EQUATION

See **Geometrical optics, nonlinear**

EINSTEIN EQUATIONS

After Albert Einstein (1905) published his special theory of relativity, Hermann Minkowski (1908) delivered a seminar in which he said

> Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.

This "kind of union" is called space-time. In special relativity, two observers in inertial frames moving at constant velocity with respect to each other will not, in general, agree on the distance between two objects or the duration of some interval of time they both observe. However, suppose an observer sees a particle at time t and at (cartesian) coordinates (x, y, z) and later observes the particle at time $t + \mathrm{d}t$ to be at the point $(x + \mathrm{d}x, y + \mathrm{d}y, z + \mathrm{d}z)$. If both observers record these measurements, they will both agree on the value of the quantity

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \tag{1}$$

(where c is the speed of light in vacuum), which is called the metric on Minkowski space-time. In tensor notation, $\mathrm{d}s^2 = \sum \eta_{\mu\nu}\,\mathrm{d}x^\mu \mathrm{d}x^\nu$, where the four values of the subscripts μ, ν correspond to the coordinates on the four-dimensional space-time, and $\eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$ is the metric tensor. Minkowski space-time is flat; that is, it can be identified with its tangent space.

In general relativity, described by the Einstein equations, the constants $\eta_{\mu\nu}$ are replaced by more general, coordinate-dependent metric coefficients $g_{\mu\nu}$, and space-time is allowed to have nonzero curvature. This curvature manifests itself as gravitation and is caused by the presence of matter. In any space-time, test particles follow geodesics (paths that minimize distance locally, i.e., generalizations of straight lines) of the space-time. In the now-famous words of the astrophysicist John Archibald Wheeler (Misner, Thorne & Wheeler, 1973),

> Matter tells space-time how to curve and curved space tells matter how to move.

The latter part of the sentence corresponds to the geodesics; the first part is the content of the Einstein equations, which in tensor notation are

$$G_{\mu\nu} = 8\pi\, T_{\mu\nu}, \qquad \mu, \nu = 0, 1, 2, 3, \tag{2}$$

where we are using units in which the speed of light and Newton's gravitational constant both have the value 1. The tensor $T_{\mu\nu}$ is called the stress-energy or energy-momentum tensor, and its components measure several physical properties of continuous matter. In vacuum, $T_{\mu\nu} = 0$. The tensor $G_{\mu\nu}$ is called the Einstein tensor. It is constructed from $g_{\mu\nu}$ and its first two derivatives, and it is a measure of the curvature of space-time. In short, the left-hand side of the Einstein equations (2) measures curvature, which encodes the geometry of the space-time, while the right-hand side of the equations measures energy, momentum, and stress, and so encodes the physical properties of the matter. The Einstein equations can be derived from an action principle and are constructed in such a way as to satisfy a generalization of the law of conservation of energy.

Apart from the flat space-time of Minkowski, the most important solutions of the Einstein equations are those with spherical symmetry. In Schwarzschild coordinates, the metric of any vacuum, asymptotically flat, spherically symmetric space-time has the form

$$ds^2 = \left(1 - \frac{2M}{r}\right) dt^2 - \left(1 - \frac{2M}{r}\right)^{-1} dr^2 \\ -r^2(d\theta^2 + \sin^2\theta \, d\phi^2), \qquad (3)$$

where M is a constant (Schwarzschild, 1916). The Schwarzschild metric (3) is used to describe the external field of a nonrotating spherically symmetric matter distribution (such as a star) of mass M. Expression (3) becomes singular at the so-called Schwarzschild radius $r = 2M$. In ordinary stellar models, the Schwarzschild radius lies deep in the interior of the star where the vacuum metric (3) does not apply. However, if a star undergoes gravitational collapse, then its radius shrinks to zero and the Schwarzschild radius is in the vacuum region.

By computing certain curvature invariants, it can be shown that there is no physical singularity at $r = 2M$. In other words, there is nothing wrong with the space-time manifold at $r = 2M$; it is simply that the coordinates r and t are bad here. This phenomenon is called a coordinate singularity. Use of the so-called Kruskal–Szekeres coordinates allow us to explicitly extend the space-time past this singularity. The surface $r = 2M$ is the event horizon of a black hole. Nothing can escape from the interior of the event horizon, including light, and will eventually fall into the (physical) singularity at $r = 0$ in finite proper time, where the theory breaks down.

Although a great number of exact solutions of the Einstein equations are known (Stephani et al., 2003), they describe very special and often unphysical situations. In order to study more general situations, extensive numerical studies have been undertaken. There are many ways in which the numerical evolution of the Einstein equations is especially difficult. The dependent variables of the Einstein equations are metric coefficients which describe the space-time manifold on which the independent variables (the space-time coordinates) live. Many issues stem from the fact that there is no preferred frame in general relativity meaning that one has to deal with the choice of an appropriate gauge and to recognize and avoid coordinate singularities. The Einstein equations are a system of ten strongly coupled nonlinear PDEs.

The Einstein equations as written in (2) are not in evolution form. For numerical purposes the Einstein equations are usually written in terms of an evolution variable λ. A Cauchy or characteristic approach corresponds to the normal of the $\lambda = \text{const}$ hypersurfaces being time-like or null, respectively. Einstein's equations then project to a set of constraint equations on the hypersurfaces and a set of evolution equations. If the constraint equations are satisfied on an "initial" hypersurface, then they are preserved under evolution to other hypersurfaces.

Numerical studies using the Cauchy approach have been used to model the collision of axisymmetric black holes (Anninos et al., 1995), and they have led to the discovery of critical phenomena in spherical collapse (Choptuik, 1993). One drawback of this method is that boundary conditions are usually artificially imposed on the hypersurfaces to avoid integrating out to infinity (although sometimes conformal methods can be used to include space-like infinity). The characteristic approach has led to the first unlimited evolution of a single black hole space-time (Gomez et al., 1998). While this approach allows for long-time simulations, it is only valid in the absence of crossing characteristics or caustics.

General relativity makes a number of predictions that have been tested. These include the excess advance in the perihelion of Mercury, the bending of light rays near the sun, time delays in radar signals passing near the sun, and gravitational lensing of distant galaxies by nearer ones. The most important prediction still awaiting confirmation is the existence of gravitational waves. A number of detectors have now been built around the world and observations are expected soon.

ROD HALBURD AND GINO BIONDINI

See also **Black holes; Cosmological models; General relativity; Gravitational waves**

Further Reading

Anninos, P., Hobill, D., Seidel, E., Smarr, L. & Suen, W.-M. 1995. Head-on collision of two equal mass black holes, *Physical Review* D, 52: 2044–2058

Choptuik, M.W. 1993. Universality and scaling in gravitational collapse of a massless scalar field. *Physical Review Letters*, 70: 9–12

Einstein, A. 1905. Zur elecktrodynamik bewegter Körper. *Annalen der Physik*, 17: 891–921

Gomez, R., Lehner, L., Marsa, R.L. & Winicour, J. 1998. Moving black holes in 3D, *Physical Review*, D57: 4778–4788

Minkowski, H. 1908. Space and time. Translated in *The Principle of Relativity*, New York: Dover, 1923

Misner, C.W., Thorne, K.S. & Wheeler, J.A. 1973. *Gravitation*, San Francisco: Freeman

Schwarzschild, K. 1916. Über das Gravitationsfeld einer Kugel aus inkompressibler Flussigkeit nach der Einsteinschen Theorie, *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Sitzung der Physikalisch-mathematischen Klasse*, 424–434

Stephani, H., Kramer, D., MacCallum, M., Hoenselaers, C. & Herlt, E. 2003. *Exact Solutions of Einstein's Field Equations*, Cambridge and New York: Cambridge University Press

Wald, R.M. 1984. *General Relativity*, Chicago: University of Chicago Press

ELASTIC AND INELASTIC COLLISIONS

See **Collisions**

ELECTROENCEPHALOGRAM AT LARGE SCALES

Since the first human scalp recordings of the tiny electric currents generated by the brain (electroencephalogram or EEG) were obtained in the mid-1920s, EEG has been recognized as a genuine (if often opaque) window on the mind, allowing observations of brain processes to be correlated with behavior and cognition. EEG has important applications in medicine, including epilepsy, head trauma, drug overdose, brain infection, sleep disorder, coma, stroke, tumor, monitoring anesthesia depth, and fundamental cognitive studies. Most EEG power occurs at frequencies below about 15 Hz in scalp (not cortical) recordings. EEG and magnetoencephalography (MEG) are the only technologies with sufficient temporal resolution to follow the fast dynamic changes associated with cognition; however, EEG spatial resolution is poor relative to modern brain structural imaging methods such as positron emission tomography (PET), computed tomography (CT), and magnetic resonance imaging (MRI). Each scalp electrode records electrical activity at very large scales, involving cortical tissue containing perhaps 10^7–10^9 neurons.

Mesoscopic and Microscopic Sources

Scalp potentials are generated by micro-current sources at cell membranes. Sorting out the complex relations between micro-sources and macroscopic scalp potentials is facilitated by assuming an intermediate (mesoscopic) descriptive scale that recognizes the columnar structure of the neocortex. From this perspective, the mesoscopic source strength of a volume of tissue is its electric current dipole moment per unit volume (microamps/mm^2), designated as $\boldsymbol{P}(\boldsymbol{r}', t)$. This function represents the weighted average of micro-source activity in a volume of the neocortex (near $\boldsymbol{r}'$) that is large compared with the scale of individual neurons yet small compared with the size of a typical EEG electrode. For the idealized case of micro-sources of one sign confined to a superficial cortical layer and micro-sources of opposite sign confined to a deep layer, $\boldsymbol{P}(\boldsymbol{r}', t)$ is roughly the mesoscopic current density across a cortical column (≈ 1 mm^2). The contributing electrical activities are primarily dendritic post-synaptic currents, although currents associated with axonal action potentials (spikes) may also contribute (see Figure 1).

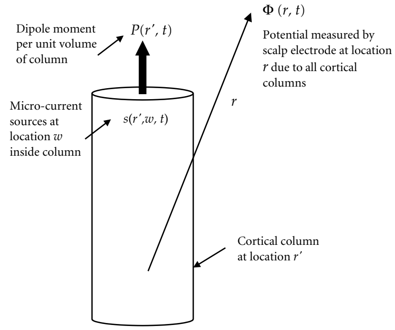
Φ (r, t)
Potential measured by scalp electrode at location r due to all cortical columns
Dipole moment per unit volume of column
P(r′, t)
Micro-current sources at location w inside column
s(r′,w, t)
r
Cortical column at location r′

Figure 1. A mesoscopic tissue mass (for example, a mm-scale cortical column containing millions of volume micro-current sources $s(\boldsymbol{r}', \boldsymbol{w}, t)$) produces a current dipole moment per unit volume $\boldsymbol{P}(\boldsymbol{r}', t)$, or meso-source strength. Cortical or scalp potential due to brain sources is the weighted integral of $\boldsymbol{P}(\boldsymbol{r}', t)$ over the brain volume or, in the case of exclusively cortical sources, the integral over the cortical surface.

Mesoscopic Sources and Scalp Recordings

Human neocortical sources form dipole layers over which the function $\boldsymbol{P}(\boldsymbol{r}', t)$ varies with cortical location ($\boldsymbol{r}'$), measured in and out of cortical folds. In a few special cases, $\boldsymbol{P}(\boldsymbol{r}', t)$ may be approximated by a few cm-scale active regions, consisting of focal sources as in focal epilepsy or mid-latency components of evoked potentials (before tangential spread to other cortical locations). In general, however, $\boldsymbol{P}(\boldsymbol{r}', t)$ is widely distributed, perhaps over the entire cortical surface. Most EEGs are believed to be generated as a linear sum of contributions from cortical sources, in which case cortical or scalp potential may be approximated by the following integral of dipole moment over the cortical surface:

$$\Phi(\boldsymbol{r}, t) = \int_S \boldsymbol{G}(\boldsymbol{r}, \boldsymbol{r}') \cdot \boldsymbol{P}(\boldsymbol{r}', t)\, \mathrm{d}S(\boldsymbol{r}'). \qquad (1)$$

The Green function $\boldsymbol{G}(\boldsymbol{r}, \boldsymbol{r}')$ contains all geometric and conductive information about the head volume conductor. Cortical dipole moment may in turn be expanded in a series of basis functions $\boldsymbol{p}_n(\boldsymbol{r})$:

$$\boldsymbol{P}(\boldsymbol{r}, t) = \sum_{n=0}^{\infty} \xi_n(t) \boldsymbol{p}_n(\boldsymbol{r}). \qquad (2)$$

An idealized model neocortex consists of a thin spherical shell with $\boldsymbol{P}(\boldsymbol{r}', t)$ everywhere normal to the surface, reflecting the columnar structure of the closed neocortical surface. An appropriate choice of basis functions for this idealized cortex is the set of spherical harmonics $Y_{lm}(\theta, \phi)\boldsymbol{a}_r$, where $\boldsymbol{a}_r$ is a unit vector in the radial direction and (θ, ϕ) are the usual spherical coordinates. The single sum over n in Equation (2) may then be expressed as a double sum $l = 0, \infty$; $m = -l, +l$, associated with the spherical geometry. Combining Equations (2) and (1) yields the following series expansion for the cortical $\boldsymbol{r}_\mathrm{C}$ or scalp $\boldsymbol{r}_\mathrm{S}$ surface potential $\Phi(\boldsymbol{r}, t)$ in terms of a new set of basis functions $\phi_n(\boldsymbol{r})$ that are surface integrals of the dot product of the Green's function $\boldsymbol{G}(\boldsymbol{r}, \boldsymbol{r}')$ with the basis functions $\boldsymbol{p}_n(\boldsymbol{r})$. Thus,

$$\Phi(\boldsymbol{r}, t) = \sum_{n=0}^{\infty} \xi_n(t)\phi_n(\boldsymbol{r}). \tag{3}$$

General Dynamic Properties of Cortical and Scalp Potentials

As summarized below, EEG exhibits many dynamic behaviors, depending on recording location, physiologic state, and subject.

(i) Often complex physical or biological systems can be adequately characterized by only a few terms in Equation (3). The time-dependent coefficients $\xi_n(t)$ are called order parameters in the field of synergetics and may be governed by nonlinear differential or integral equations. The basis functions $\phi_n(\boldsymbol{r})$ may be chosen (bottom up) by physiological theory or (top down) by experimental data, for example, by constructing Karhunen–Loeve expansions (also called principal components analysis) in which the $\phi_n(\boldsymbol{r})$ are chosen as the most efficient set of orthogonal functions representing a data record.

(ii) The basis functions $\phi_n(\boldsymbol{r})$ are typically ordered in terms of progressively higher spatial frequencies as in Fourier series. The index n is then a measure of the two-dimensional spatial frequency of the corresponding basis function. Many systems exhibit a correspondence between spatial and temporal frequencies such that Fourier transforms of the order parameters $\xi_n(\omega)$ peak at higher frequencies ω for higher spatial frequencies n. For linear wave phenomena, such a correspondence is called the dispersion relation.

(iii) The large-scale spatiotemporal dynamics of cortical potential $\Phi(\boldsymbol{r}_\mathrm{c}, t)$ are believed to be very similar to the dynamics of the mesoscopic cortical sources $\boldsymbol{P}(\boldsymbol{r}, t)$. The head volume conductor acts as a low-pass spatial filter, resulting mainly from the low electrical conductivity of skull and the physical separation (1–2 cm) between sources and electrodes. Scalp potential $\Phi(\boldsymbol{r}_\mathrm{s}, t)$ is then a spatial low-pass representation of cortical potential $\Phi(\boldsymbol{r}_\mathrm{c}, t)$ or mesoscopic source function $\boldsymbol{P}(\boldsymbol{r}, t)$. Comparisons of cortical potential (ECoG) with scalp potential (EEG) show that this spatial filtering results in temporal filtering. That is, the Fourier transform of cortical potential $\Phi(\boldsymbol{r}_\mathrm{c}, \omega)$ typically contains much more relative power at higher frequencies (say 15–40 Hz) than the scalp potential transform $\Phi(\boldsymbol{r}_\mathrm{s}, \omega)$, recorded in the same brain state, an observation qualitatively consistent with normal wave dispersion relations.

(iv) EEG phenomena typically exhibit larger amplitudes at lower frequencies. In deep sleep and moderate to deep anesthesia, $\Phi(\boldsymbol{r}_\mathrm{S}, t)$ is typically a few 100 μV with nearly all power in the delta range (0–4 Hz) at all scalp locations. The eyes closed, waking alpha rhythm (ca. 40 μV) is normally dominated by one or two spectral peaks in the 8–13 Hz range at widespread scalp locations. Low-amplitude beta activity (ca. 13–20 Hz) superimposed on alpha rhythms is more evident in frontal cortex. Alcohol and hyperventilation typically lower alpha frequencies and increase amplitudes; barbiturates increase beta activity. Scalp EEG activity is more consistent with limit cycle modes $\xi_n(t)$ than low-dimensional chaos.

(v) Scalp EEG is a weighted space average of many cortical rhythms that can look different in different cortical regions. Alpha rhythms have been recorded from nearly the entire upper cortical surface, including frontal and prefrontal areas. Differences in ECoG waveforms between cortical areas are largely eliminated with anesthesia, suggesting shifts from more functional localization to more globally dominated brain states.

(vi) Both globally coherent and locally dominated behavior can occur within the alpha band, depending on narrow band frequency, measurement scale and brain state. Upper alpha (ca. 10 Hz) and theta (ca. 4–6 Hz) phase locking between cortical regions during mental calculations often occurs, consistent with neural network formation. At the same time, quasi-stable alpha phase structures consistent with global standing waves have been observed.

Neocortical Dynamic Theory

The apparent balance among locally and globally dominated dynamic processes has been estimated by phase synchronization among other measures. Cortical or thalamic interactions with time delays due to rise and decay times of post-synaptic potentials (local theory) have been modeled. Network frequencies are then determined only by local tissue properties. In global theories, characteristic frequencies depend on the entire neocortex/cortico-cortical fiber system. Excitatory synaptic action density $F(\boldsymbol{r}, t)$ may be expressed in terms of action potential density $\Gamma(\boldsymbol{r}, t)$ by

$$F(\boldsymbol{r}, t) = \int_0^{\infty} \mathrm{d}v \int_S R(\boldsymbol{r}, \boldsymbol{r}_l, v) \times \Gamma\left(\boldsymbol{r}_l, t - \frac{|\boldsymbol{r} - \boldsymbol{r}_l|}{v}\right) \mathrm{d}S(\boldsymbol{r}_l) \tag{4}$$

The outer integral is over distributed cortico-cortical propagation speeds v, and the inner integral is over the neocortical surface S. The dependence of cortico-cortical fiber density with distance $|\boldsymbol{r} - \boldsymbol{r}_l|$ is expressed by the function $R(\boldsymbol{r}, \boldsymbol{r}_l, v)$. This linear relation between the dependent variables $F(\boldsymbol{r}, t)$ and $\Gamma(\boldsymbol{r}, t)$ may be combined with a variety of nonlinear local equations in the same dependent variables. Such local/global theories include both network-to-global and global-to-network interactions.

PAUL NUÑEZ

See also **Cell assemblies; Electroencephalogram at mesoscopic scales; Gestalt phenomena; Synergetics**

Further Reading

Edelman, G.M. & Tononi, G. 2000. *A Universe of Consciousness*, New York: Basic Books

Freeman, W.J. 1975. *Mass Action in the Nervous System*, New York: Academic Press

Haken, H. 1983. *Synergetics. An Introduction*, 3rd edition, Berlin: Springer

Malmuvino, J. & Plonsey R. 1995. *Bioelectromagetism*, New York: Oxford University Press

Niedermeyer, E. & Lopes da Silva, F.H. (editors). *electroencephalography. Basic Principals, Clinical Applications, and Related Fields*, 4th edition. London: Williams and Wilkins

Nuñez, P.L. 1981. *Electric Fields of the Brain: The Neurophysics of EEG*, New York: Oxford University Press

Nuñez, P.L. 1995. *Neocortical Dynamics and Human EEG Rhythms*, New York: Oxford University Press

Nuñez, P.L., Wingeier, B.M. & Silberstein, S.B. 2001. Spatial-temporal structure of human alpha rhythms: theory, microcurrent sources, multiscale measurements, and global binding of local networks. *Human Brain Mapping*, 13: 125–164

Nuñez, P.L. 2000. Toward a quantitative description of large scale neocortical dynamic function and EEG. *Behavioral and Brain Sciences*, 23: 371–437

Penfield, W. & Jasper, H.D. 1954. *Epilepsy and the Functional Anatomy of the Human Brain*, London: Little, Brown and Co

Scott, A.C. *Stairway to the Mind*, Berlin and New York: Springer, 1995

Uhl, C. (editor). 1999. *Analysis of Neurophysiological Brain Functioning*, Berlin: Springer 1999

Wilson, H.R. & Cowan, J.D. 1973. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13: 55–80

ELECTROENCEPHALOGRAM AT MESOSCOPIC SCALES

The two most dominant theoretical positions that have influenced the development of explanations of behavior are functional modularity, where specific behaviors are believed to reside in distinct cortical locations, and mass action, in which behavior is posited to arise out of the cooperative activity of distributed neural structures comprising the brain. The first of these positions predominates in the medical and cognitive sciences, based upon observations collected over many years in which selective behavioral deficits were observed to occur in response to the specific destruction or stimulation of various areas of the brain. This has led to considering the brain as a collection of interconnected functionally specialized modules. This view is generally known as modularism and in the context of higher functions is also referred to as cortical localizationism. Modules are thought to embody algorithms, acting mechanistically on input to produce output, in a computer-like manner. Mass action, however, considers behavior as best understood as arising out of cooperative neural activity occurring over a number of interacting spatial and temporal scales. The consequence of this view is that behavior can now be characterized by the observation of dynamical patterns of brain activity.

The dynamics of the brain can be observed at a number of different spatial and temporal scales. These range from the level of the ion channel, synapse, and neuron (microscopic), to the level of neuronal population (mesoscopic), up to the level of large aggregates of brain tissue (macroscopic). The associated methods can generally be divided into those that measure some form of electromagnetic activity and those that reflect the metabolic correlates of neural activity. At the mesoscopic scale electromagnetic measures typically have millisecond temporal resolution whereas metabolic measures (e.g., fMRI, PET, near-infrared spectroscopy, diffuse optical tomography) have comparatively poor temporal resolution (s), while both mesoscopic and metabolic measures have comparable spatial resolution (mm). Electromagnetic measurements of brain activity better reflect the time scales of neuronal activity associated with the dynamics of behavior. For these reasons theories of mesoscopic brain dynamics are built around state variables that characterize the electromagnetic activity associated with neural populations. The most important mesoscopic electromagnetic measures are those of the local field potential (LFP) and the electrocorticogram (ECoG) which predominately reflect the quasi-static electromagnetic fields produced in response to the ionic currents generated by synchronized synaptic activity. It is generally thought that in the cortex this synchronized activity is linearly related to the spatially averaged soma membrane potential of populations of excitatory (pyramidal) neurons.

For both excitatory and inhibitory neuronal populations the mean soma membrane potential determines the mean rate of neuronal action potential generation (or firing rate). In the simplest case, the mean population firing rate is a sigmoidal function of the mean soma membrane potential for functionally equivalent members of the same neuronal population. More generally, such a relationship will be time-variant. Because the mean soma membrane potential is a function of neuronal population synaptic input which itself is a function of the mean neuronal population firing rates, the mean soma membrane potential can be used as a canonical state variable to both characterize and develop theories of the dynamics of neuronal populations.

At the level of the neuronal population, the spatially averaged soma membrane potential is typically defined over the characteristic scales of the short-range

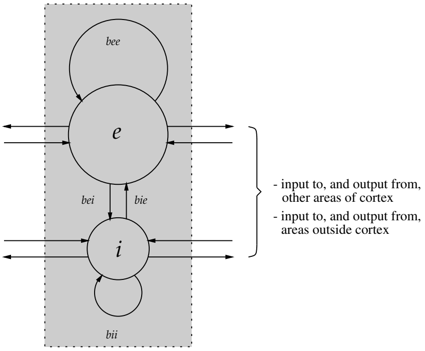
bee
e
bei
bie
i
bii
- input to, and output from, other areas of cortex
- input to, and output from, areas outside cortex

Figure 1. The neurons of mammalian cortex interact with each other locally (by short range connections) and globally (by long-range connections). Short-range connections serve to diffusely interconnect the intermixed excitatory (e) and inhibitory (i) neuronal populations of cortex. The characteristic scales of these connections loosely organize cortex into a sheet of overlapping "modules" or cylinders that span the entire thickness of cortex. The resulting pattern is often referred to as columnar organization with a single one of these cylinders known as a cortical macrocolumn. One way of quantifying the strength of interaction between these local neuronal populations is to specify the mean number of connections (or synapses) neurons of one population receive from neurons of the same or another population (b_{ij}).

intracortical fibers. These intracortical fibers comprise the recurrent axonal branchings of the pyramidal neurons as well as all the axonal fibres of the inhibitory interneurons. Detailed morphometric studies of cortex have established that the typical characteristic spatial scale of these fibers ranges anywhere between 30 μm to 1 mm (Braitenberg & Schüz, 1998). The advantage of using this spatially averaged state variable is that it has a commensurate spatial scale to LFP and ECoG recordings.

Most mesoscopic theories of neuronal dynamics have considered cerebral cortex to consist of two functionally distinct neural populations—excitatory (e) and inhibitory (i)—reflecting the respective pyramidal and interneuron populations. The earliest models of cortical mass action attempted to describe neuronal dynamics exclusively in terms of short- and long-range excitatory interactions between excitatory populations. While a number of interesting analytical solutions were obtained using this approach, they are no longer of any significant physiological relevance because they did not incorporate the activity of inhibitory neurons. Subsequent models, based on anatomical considerations, have addressed this deficiency by the incorporation of all forms of feedforward and feedback connectivity between excitatory and inhibitory neuronal populations. Starting with Wilson & Cowan (1972), most modern theories functionally incorporate local $e \to e$, $e \to i$, $i \to e$, and $i \to i$ connections, as illustrated in Figure 1. However, a notable exception is the influential body of work by Lopes da Silva et al. in which all the significant population neurodynamics arise out of the reciprocal interactions between excitatory and inhibitory neurons.

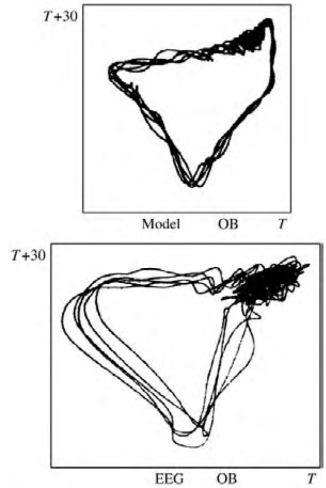

Figure 2. Delay embedded phase plane plots comparing EEG seizure activity recorded from rat olfactory bulb (a phylogenetically primitive form of cortex present in all mammals), top panel, and with a mesoscopic theory of olfactory bulb dynamics (bottom panel). The delay embedding is 30 ms for experimental and simulated time series of 1s duration, with time increasing counterclockwise. Figure reproduced with permission from Freeman (1987), Copyright Springer-Verlag.

In the majority of theories the dynamical response of the mean soma membrane potential to synaptic activity induced by population neuronal firings is generally described using a differential formalism.

$$\dot{h}_n = -a_n h_n + \sum_{r=1}^{N} g_{nr}(h_n) I_{nr}, \qquad (1)$$

$$\mathcal{D}_{nr} I_{nr} = S_{nr}(h_r) \quad n, r = 1 \dots N, \qquad (2)$$

where N is the number of locally interacting neuronal populations and the a_n are real constants which correspond approximately to the reciprocal of the mean neuronal membrane time constant. In most theories $N = 2$. h_n is the mean soma membrane potential of the nth neuronal population, and S_{nr} defines functions that take into account the topology of neuronal population connectivity as well as the sigmoidal relationships between mean soma membrane potential and mean firing rate. $g_{nr}(h_n)I_{nr}$ represents the postsynaptic current induced in local neuronal population n by neuronal population r, whose temporal evolution is determined by the temporal differential operator $\mathcal{D}_{nr}$. This operator takes into account bulk neurotransmitter kinetics and neuronal cable properties.

Theories of mesoscopic dynamics can generally be distinguished by the form and order of $\mathcal{D}_{nr}$, g_{nr} and S_{nr}, and N. For instance, in the original work of Wilson and Cowan, $\mathcal{D}_{nr}$ was of zero order in time, whereas in the theories of Lopes da Silva, $\mathcal{D}_{nr}$ is of first order and $g_{nr}(h_n) = \text{constant}$. More recent theories have considered $g_{nr}(h_n)$ linear in h_n and $\mathcal{D}_{nr}$ second order in time due to more detailed physiological considerations, (Liley et al., 2002). Differing forms and parametrizations of $\mathcal{D}_{nr}$, g_{nr}, and S_{nr} have been shown to give rise to a wide range of dynamics (e.g., limit cycle, chaos, and filtered noise) some of which bear strong similarities to experimental recordings (see Figure 2; see also figures in color plate section).

David Liley and Mathew Dafilis

See also **Cell assemblies; Electroencephalogram at large scales; Nerve impulses; Neurons; Synergetics**

Further Reading

Braitenberg, V. & Schüz, A. 1998. *Cortex: Statistics and Geometry of Neuronal Connectivity*. Berlin and New York: Springer

Freeman, W.J. 1975. *Mass Action in the Nervous System*. New York: Academic Press

Freeman, W.J. 1987. Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biological Cybernetics*, 56(2–3): 139–150

Freeman, W.J. 2000. *Neurodynamics: An Exploration in Mesoscopic Brain Dynamics*. London and New York: Springer

Liley, D.T.J., Cadusch, P.J. & Dafilis, M.P. 2002. A spatially continuous mean field theory of electrocortical activity. *Network: Computation in Neural Systems*, 13(1): 67–113

van Gelder, T. 1999. Dynamic approaches to cognition. In *The MIT Encyclopedia of the Cognitive Sciences*, edited by R.A. Wilson & F.C. Keil, Cambridge, MA: MIT Press

van Rotterdam, A., Lopes da Silva, F.H., van den Ende, J., Viergever, M.A. & Hermans, A.J. 1982. A model of the spatial-temporal characteristics of the alpha rhythm. *Bulletin of Mathematical Biology*, 44(2): 283–305

Wilson, H.R. & Cowan, J.D. 1972. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12: 1–24

ELECTRON BEAM MICROWAVE DEVICES

Electron beam devices have been used in a wide variety of applications since the beginning of vacuum technology early in the 20th century. These include simple diodes and triodes, used in early radio; cathode-ray tubes, used for display in oscilloscopes, televisions, and computers; lithography; high-voltage devices used to produce X-rays; sources for electron accelerators; and the energy source for microwave devices, with applications in communications, microwave ovens, industrial processes, and electronic warfare. For the purposes of this entry, we will limit our discussion to just one application, the production of electromagnetic energy at microwave frequencies, where nonlinear space charge effects tend to play a major role. For a wider range of applications, including the basic theory, texts are available such as Harmon (1953) and Gilmour (1994). It should be noted that solid state devices, not involving electron beams, are now often used at low power, particularly as oscillators to drive microwave electron-beam amplifiers.

Generation of microwaves with electron beams involves the interaction of a beam with electromagnetic fields across gaps of discrete cavities or synchronously with electromagnetic fields propagating on a slow wave structure. The basic physical principle is that a bunched beam interacts with the decelerating phase of an electric field to transfer energy from the beam to the field. This mechanism is the converse of the interaction in linear accelerators and synchrotrons that operate by interaction of a field in an accelerating phase with a bunched beam to transfer energy from the fields to the beam. In a device such as a traveling-wave tube (TWT) or magnetron, in which the beam interacts with a traveling wave, there is a natural bunching action as the wave develops, such that most of the beam particles are decelerated, producing the wave amplification. In a klystron, with a bunching cavity and an energy extraction cavity, the proper phase relation must be established, which is, in fact, equivalent to the synchronous condition in the TWT.

The earliest beam-type device to be extensively used for microwave generation is the magnetron, a coaxial cylindrical device with a potential between inside cathode and outside anode cylinders, and a magnetic field along the cylindrical axis. The electrons accelerated from the cathode toward the anode perform cycloidal motion due to the magnetic field with an average azimuthal drift velocity in the combined electric and magnetic fields given by $\boldsymbol{v}_{\mathrm{d}} = \boldsymbol{E} \times \boldsymbol{B}/B^2$. The anode consists of a series of microwave cavities with the operating frequency, chosen such that the cavities operate in the π-mode, that is, 180° phase shift between cavities. The fields and dimensions are chosen such that the electrons are in synchronism with the fundamental Fourier component of the fields that appear across the cavity openings. Through a complicated nonlinear process, the electrons bunch at a phase with respect to the traveling microwave field to give energy to the field, finally being collected on the anode structure with an energy considerably less than the energy eV, where V is the cathode-anode accelerating voltage. Simplified models describe some of the processes (Hutter, 1965), but most development has been experimental, assisted more recently by detailed numerical calculations (Lemke et al., 2000). The magnetron configuration described above operates naturally as an oscillator. It has high efficiency and usually operates at high power. It was an essential element in the development of radar during World War II, which was important in winning the air war over Great Britain. The magnetron has become ubiquitous in the consumer market, as the power source in the household microwave oven. Magnetrons, despite being inexpensive and robust, have a number of unfavorable characteristics, including excess noise, narrow bandwidth, and difficulty in tuning, which limit their applications.

The klystron, which was developed roughly during the same period of time as the magnetron, is more flexible because it can operate, with small modifications, as either an amplifier or an oscillator, and versions operate over very large ranges of power and frequency. The basic configuration is two cylindrical microwave cavities with an electron beam passing through their center. The cavities have a reentrant shape such that the central hole consists of a narrow gap. The beam is velocity modulated by an alternating electric field across the first cavity gap which becomes density modulated in the second gap. At very low beam density, the electron trajectories can be taken to be kinematic, with the bunching distance related to the beam velocity and the perturbed velocity created by the gap fields. However, the usual operation is in a regime where space charge effects are fundamental to the operation. The excitation produces two waves, fast and slow space charge waves, which give a beat modulation distance $\lambda_b/4 = \pi v_o/2\omega_{pb}$, where v_o is the beam velocity and ω_{pb} is the plasma frequency of the beam. The trajectories are shown schematically in Figure 1. The beam plasma frequency is reduced from the natural electron oscillation $\omega_p = ne^2/m\epsilon_0$ due to the transverse beam dimensions. A second cavity at a position $z = \lambda_b/4$ from the first cavity is excited by a maximum of the radio-frequency current in the beam. The amplifier is converted into an oscillator by either feeding back some of the signal externally from the second cavity to the first cavity or by reflecting the modulated beam, after traversing only a single cavity, to self-excite it as a reflex klystron. See a basic text such as Harmon (1953), or more detail in Hutter (1965), for mathematical analysis.

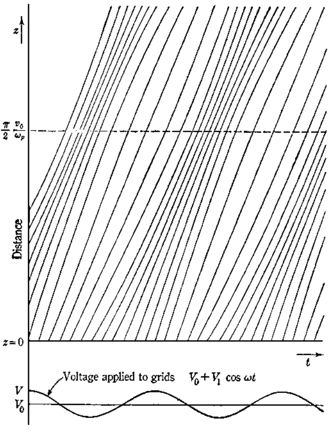
z
$z_1 = \frac{\pi}{2}\frac{v_0}{\omega_p}$
Distance
z=0
t
Voltage applied to grids $V_0 + V_1 \cos \omega t$
V
V_0

Figure 1. Distance-time diagram of a klystron indicating how velocity variations at the input grids result in density variations farther down the electron stream (after Harmon, 1953).

The traveling-wave tube (TWT), developed somewhat later, competes for applications with the klystron at all but the highest powers. The device transfers energy from a beam to a slow wave structure, usually a helix, by a resonant interaction in which the beam becomes naturally modulated in the presence of the wave. A linear, or small signal, analysis is sufficient to obtain the amplifying properties of the device for many low-power applications. However, as the microwave power becomes comparable to the beam power, nonlinear considerations become very important. As in the klystron, space charge is a fundamental source of nonlinearity. If significant power is extracted from the beam, the wave velocity on the circuit must be reduced along its length to maintain the coherent synchronous interaction. Due to the periodicity of slow wave structures, the propagation characteristics have operating regions for which the group velocity $v_g = d\omega/dk$ is opposite in direction to the phase velocity $v_{ph} = \omega/k$ (Brillouin, 1953). If the TWT is operated with such parameters, then wave energy is propagated backward from the direction in which the beam is moving and in which the wave amplitude, by interaction with the beam, is growing. This feedback allows the beam to self-excite the wave and therefore become an oscillator, which is called a backward wave oscillator (BWO). The operating conditions of a BWO are inherently nonlinear. In addition to the general books already mentioned, a detailed account of TWTs by one of the scientists associated with their development, is Pierce (1950). Sketches of the three basic types of devices are shown in Figures 2, 3, 4. There are many variants of these basic configurations.

Because electron beams are the power source and the active interacting medium for both klystron and TWT amplifiers and their associated oscillators, beam design is of great importance in obtaining good operating

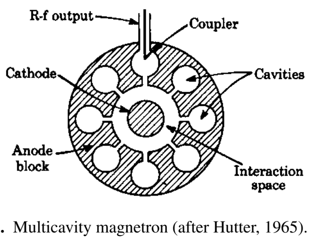

Figure 2. Multicavity magnetron (after Hutter, 1965).

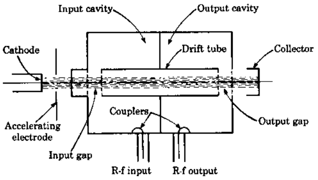

Figure 3. Two-cavity klystron (after Hutter, 1965).

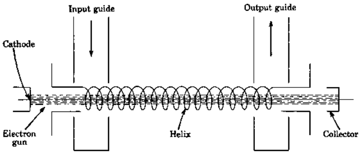

Figure 4. Helix-type traveling-wave tube (after Hutter, 1965).

characteristics. Some of the issues involved are collimation (minimizing transverse emittance), energy spread (minimizing longitudinal emittance), and high current and/or current density. The collimation is often accomplished with a uniform magnetic field along the beam axis, but radial space charge forces also lead to beam rotation. Alternatively, electric or magnetic lenses can be employed. An important consideration is to launch the beam from an accelerating region in a smooth manner; this led to the electrode shape known as the Pierce diode. An important method of confining a beam with significant space charge is to inject the beam into a uniform magnetic field from a magnetic field-free region. The proper choice of magnetic field leads to azimuthal rotation that just takes up the space charge potential, lending to smooth, uniform axial flow, called Brillouin flow. Many of the basic topics of beam dynamics and beam design are covered in Pierce (1954).

Some of the important considerations that have motivated the development of various microwave beam devices have been noise characteristics, particularly achieving low noise to amplify very low signals; frequency ranges, particularly pushing devices to increasingly high frequencies for broadband applications; tunability; and power and efficiency, particularly to obtain high power at high efficiency. For certain applications, there are also other types of constraints such as ruggedness, reliability, and weight limits, which we will not consider here. Low noise requirements have tended to favor TWTs over klystrons on the front end of receivers. Higher frequencies, used to obtain more bandwidth in communication applications, have led to miniaturization of klystrons, but other types of devices such as free electron lasers can obtain even higher frequencies for some applications (Freund & Antonsen, 1996). Some of the key issues motivating nonlinear analysis have arisen from the requirement of obtaining high efficiency at high power. Currently, the state of the art in high power is to produce 1GW in a pulse of 0.1kJ. For studies of more recent analysis and developments the reader is referred to Benford & Swegle (1992) and Barker & Schamiloglu (2001).

To obtain high efficiency the majority of the electrons should be trapped in the wave field, such that they can be decelerated. This requires matching the phase space of the beam emittance to the acceptance of the wave field, usually by some additional bunching mechanisms. Some of the basic ideas are treated in Lichtenberg (1969), but each device needs detailed numerical trajectory calculations to optimize trapping. In addition to trapping, the exiting spent beam must be decelerated to retrieve the excess energy (a depressed collector). Reducing the collector voltage is limited by the requirement that the average beam energy of the exiting beam must be greater than its energy spread so as not to turn electrons around. This condition requires knowledge of the nonlinear characteristics of the longitudinal phase space emittance.

Allan J. Lichtenberg and John P. Verboncoeur

See also **Particle accelerators**

Further Reading

Barker, R.J. & Schamiloglu, E. (editors). 2001. *High Power Microwave Sources and Technologies*. New York: IEEE Press

Benford, J. & Swegle, J. 1992. *High Power Microwaves*, Norwood, MA: Artech House

Brillouin, L. 1953. *Wave Propagation in Periodic Structures*. New York: Dover

Freund, H.P. & Antonsen, Jr., T.M. 1996. *Principles of Free Electron Lasers*, London: Chapman & Hall

Gilmour, A.S. 1994. *Principles of Traveling Wave Tubes*, Boston: Artech House

Harmon, W.W. 1953. *Fundamentals of Electronic Motion*, New York: McGraw-Hill

Hutter, R.G.E. 1965. *Beam and Wave Electronics in Microwave Tubes*, Cambridge, MA: Boston Technical Publishers

Lemke, R.W., Genoni, T.C. & Spencer, T.A. 2000. Effects that limit efficiency in relativistic magnetrons. *IEEE Transactions on Plasma Science*, 28: 887–897

Lichtenberg, A.J. 1969. *Phase-Space Dynamics of Particles*, New York: Wiley

Pierce, J.R. 1950. *Traveling-Wave Tubes*, New York: Van Nostrand

Pierce, J.R. 1954. *Theory and Design of Electron Beams*, 2nd edition, New York: Van Nostrand

---

ELLIPTIC FUNCTIONS

Elliptic functions were first introduced as inverses of elliptic integrals, so called because the integral for the arclength of the ellipse studied by John Wallis in the 17th century (Stillwell, 1989) was the first such example. By the end of the 18th century, in particular after the work of Leonhard Euler, Joseph-Louis Lagrange, and Adrien-Marie Legendre, mathematicians had realized that integrals of the form $\int \mathrm{d}x/\sqrt{P(x)}$, where $P(x)$ is a cubic or quartic polynomial, could not be expressed in terms of elementary functions (or their inverses). Many elliptic integrals arise naturally in the solutions of problems in mechanics (Lawden, 1989). For example, the solution of the simple pendulum equation

$$\dot{\theta} + \omega^2 \sin\theta = 0$$

can be obtained using conservation of energy $\frac{1}{2}(\dot{\theta})^2 - \omega^2 \cos\theta = E, E > -\omega^2$, to arrive at the inversion problem for the integral

$$t - t_0 = \int_0^{\theta} \frac{\mathrm{d}\theta}{\sqrt{2E + \omega^2 - \omega^2 \sin^2(\theta/2)}}\,\mathrm{d}\theta.$$

Substituting $z = \sin(\theta/2)$ and defining $k^2 = \omega^2/(E + \omega^2) \in (0, 1)$, one arrives at the expression

$$K(z; k) = \int_0^z \frac{\mathrm{d}z}{\sqrt{1 - z^2}\sqrt{1 - k^2 z^2}} = \frac{\omega}{k}(t - t_0). \quad (1)$$

In the late 19th century, Niels Henrik Abel, Carl Jacobi, and Carl Friedrich Gauss observed the similarity between $K(z; k)$ and the integral defining the transcendental function arcsin z as

$$\text{arcsine}\, z = \int_0^z \frac{\mathrm{d}z}{\sqrt{1 - z^2}}$$

(which is indeed the limiting case of (1) as $k \to 0$, giving the solution for the small amplitude oscillations of the pendulum), and introduced the *Jacobi elliptic sine function* $\mathrm{sn}(z; k)$ as the inverse of $K(z; k)$

$$z = K(\mathrm{sn}(z; k); k).$$

While the circular (trigonometric) functions $\sin z, \cos z, \tan z$, etc., are singly periodic functions of the complex variable z, satisfying $f(z + 2\pi) = f(z)$, elliptic functions can be characterized as the doubly periodic functions, which only possess pole singularities. Elliptic functions satisfy

$$f(z + 2\omega_1) = f(z), \quad f(z + 2\omega_2) = f(z),$$

where ω_1 and ω_2 (the *periods* of f) are complex numbers such that ω_1/ω_2 is not purely real. The region of the complex plane with vertices $\{0, 2\omega_1, 2\omega_2, 2\omega_1 + 2\omega_2\}$ is called the *fundamental period parallelogram* (see Figure 1).

The simplest elliptic functions are those with two poles: the *Jacobi elliptic functions*, with two simple poles of opposite residues, and the *Weierstrass elliptic functions*, with a single double pole with zero residue, all others can be constructed from these.

Jacobi Elliptic Functions

The three most common Jacobi elliptic functions are $\mathrm{sn}(z; k)$, introduced earlier, and the associated $\mathrm{cn}(z; k) = \sqrt{1 - \mathrm{sn}^2(z; k)}$ (a generalization of the cosine function) and $\mathrm{dn}(z; k) = \sqrt{1 - k^2\mathrm{sn}^2(z; k)}$. Some of their main properties, analogues of the properties of trigonometric functions, are listed below; their proofs are elementary (Whittaker & Watson, 1943) (see Figure 2):

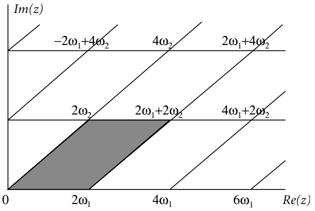

Figure 1. The period lattice of an elliptic function.

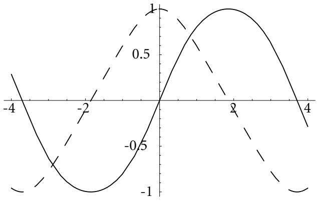

Figure 2. Graphs of $\mathrm{sn}(x; k)$ (solid line) and $\mathrm{cn}(x; k)$ (dashed line), for real x, $k = 0.5$.

Symmetry:

$$\mathrm{sn}(-u) = -\mathrm{sn}\, u, \quad \mathrm{cn}(-u) = \mathrm{cn}\, u,$$
$$\mathrm{dn}(-u) = \mathrm{dn}\, u.$$

Derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{sn}\, u = \mathrm{cn}\, u\, \mathrm{dn}\, u, \quad \frac{\mathrm{d}}{\mathrm{d}u}\mathrm{cn}\, u = -\mathrm{sn}\, u\, \mathrm{dn}\, u,$$
$$\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{dn}\, u = -k^2 \mathrm{sn}\, u\, \mathrm{cn}\, u.$$

Addition formulas: For example,

$$\mathrm{sn}(u + v) = \frac{\mathrm{sn}\, u\, \mathrm{cn}\, v\, \mathrm{dn}\, v + \mathrm{sn}\, v\, \mathrm{cn}\, u\, \mathrm{dn}\, u}{1 - k^2 \mathrm{sn}^2 u\, \mathrm{sn}^2 v.}$$

Periods: Define the *complete elliptic integral*

$$K = \int_0^1 \frac{\mathrm{d}z}{\sqrt{(1 - z^2)(1 - k^2 z^2)}},$$

the *complementary modulus* k' such that $k^2 + k'^2 = 1$, and the constant $K' = K(1; k')$. Then, for example,

$$\mathrm{sn}(u + 4K) = \mathrm{sn}\, u, \quad \mathrm{sn}(u + 2\mathrm{i}K') = \mathrm{sn}\, u.$$

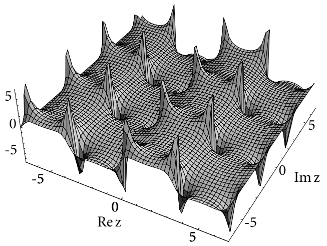

Figure 3. The doubly periodic function $\mathrm{sn}(z;k)$: graph of $Re[\mathrm{sn}(z;0.3)]$.

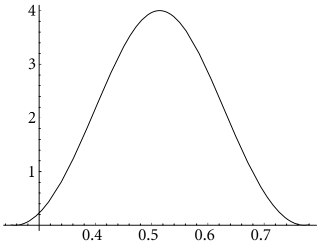

Figure 4. The profile of the cnoidal wave solution of the KdV equation, $A=4$, $k=0.3$.

Poles: For example, $\mathrm{sn}(z;k)$ is analytic except at points congruent to $\mathrm{i}K'$ and $2K+\mathrm{i}K'$ $(\mathrm{mod}\, 2K, 2\mathrm{i}K')$, which are simple poles.

Weierstrass Elliptic Functions

This class of elliptic functions was introduced by Weierstrass by means of infinite partial-fraction and product expansions, rather than as inversions of elliptic integrals. Indeed, such construction exists for all elliptic functions. Given $\omega_1, \omega_3 \in \mathbb{C}$, $\omega_2 = \omega_1 + \omega_2$, we define the most famous of Weierstrass' elliptic functions, the $\wp$-function, on the period lattice $\{m2\omega_1 + n2\omega_3 \mid m, n \in \mathbb{Z}\}$:

$$\wp(z) = \frac{1}{z^2} + \sum_{(m,n)\neq(0,0)} \left\{ \frac{1}{(z-(m2\omega_1+n2\omega_3))^2} - \frac{1}{(m2\omega_1+n2\omega_3)^2} \right\}. \tag{2}$$

$\wp(z)$ is a meromorphic function with double poles at each point of the period lattice and satisfies the following differential equation:

$$\begin{aligned}[\wp'(z)]^2 &= 4[\wp(z)]^3 - g_2[\wp(z)] - g_3 \\ &= 4(\wp(z)-e_1)(\wp(z)-e_2)(\wp(z)-e_3),\end{aligned} \tag{3}$$

where $e_i = \wp(\omega_i)$.

Among applications of the Weierstrass $\wp$-function is the profile of a wave traveling in water (see, e.g., Walker, 1996), as described by the Korteweg–de Vries (KdV) equation

$$u_t + u_x + 12uu_x + u_{xxx} = 0. \tag{4}$$

Seek a traveling-wave solution by setting $u(x,t) = f(x-ct)$, for some constant speed c. Then f satisfies the ordinary differential equation $(1-c)f' + 12ff'$ $-+f''' = 0$ which, after twice integrating, leads to

$$(f')^2 = 4f^3 - (1-c)f^2 + Af + B,$$

A, B constants of integration. By comparison with (3), the form of f must be $f(s) = \alpha\wp(\beta s) + \gamma$, for suitable constants α, β, γ. Since the Weierstrass $\wp$-function can be shown to be related to the square of Jacobi elliptic functions (Whittaker & Watson, 1943), this leads to the notable *cnoidal wave* solution of KdV (see Figure 4):

$$u(x,t) = A\,\mathrm{cn}^2\left(\frac{\sqrt{A}}{k}(x-ct); k\right),$$

$$c = 1 + 4A\,\frac{2k^2-1}{k^2}.$$

Finally, we briefly mention hyperelliptic functions (or abelian functions), as inverses of integrals of rational functions $R(z, \sqrt{P(z)})$, with $P(z)$ a quintic or higher-degree polynomial (Baker, 1995). For example, the equations for the Kovalevsky top were shown by Sophia Kovalevsky to be integrable in terms of hyperelliptic functions (Whittaker, 1960).

ANNALISA M. CALINI

Further Reading

Baker, H.F. 1995. *Abelian Functions: Abel's Theorem and the Allied Theory of Theta Functions*. Reissue of the 1897 edition, Cambridge and New York: Cambridge University Press

Byrd, P.F. & Friedman, M.D. 1954. *Handbook of Elliptic Integrals for Engineers and Physicists*, New York: Wiley

Lawden, D.F. 1989. *Elliptic Functions and Applications*, Berlin and New York: Springer

Rauch, H.E. & Lebowitz, A. 1973. *Elliptic Functions, Theta Functions and Riemann Surfaces*, Baltimore: Williams and Wilkins

Stillwell, J. 1989. *Mathematics and Its History*, 2nd edition, Berlin and New York: Springer, Chapter 12

Walker, P.L. 1996. *Elliptic Functions: A Constructive Approach*, Berlin and New York: Springer

Whittaker, E.T. 1960. *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies: With an Introduction to the Problem of Three Bodies*, Cambridge: Cambridge University Press (originally published 1904)

Whittaker, E.T. & Watson, G.N. 1943. *A Course of Modern Analysis*, 4th edition, Cambridge: Cambridge University Press

EMBEDDING METHODS

The modeling of a deterministic dynamical system relies on the concept of a phase space, which is a theoretical representation of the totality of possible system states. In general, a system state consists of information of positions and velocities needed to specify all future system states. For a system that has a mathematical model, the phase space is known from the equations of motion.

For experimental and natural chaotic dynamical systems, the full state space is unknown. Embedding methods have been developed as a means to reconstruct the phase space and develop new predictive models. One or more signals from the system must be observed as a function of time. The time series are then used to build a proxy of the observed states.

Mathematical theorems show how the observed time series can be used. A famous theorem of Hassler Whitney from the 1930s holds that a generic map from an n-manifold to $2n+1$-dimensional Euclidean space is an embedding. In particular, the image of the n-manifold is completely unfolded in the larger space because $2n+1$ signal traces measured from a system can be considered as a map from the set of states to $2n+1$-dimensional space, Whitney's theorem implies that each state can be identified uniquely by a vector of $2n+1$ measurements, thereby reconstructing the phase space.

The contribution of Floris Takens (1981) was to show that the same goal could be reached with a single measured quantity. He proved that instead of $2n+1$ generic signals, the time-delayed versions $[y(t), y(t-\tau), y(t-2\tau), \ldots, y(t-2n\tau)]$ of one generic signal would suffice to embed the n-dimensional manifold. There are some technical assumptions that must be satisfied, restricting the number of low-period orbits with respect to the time-delay τ and repeated eigenvalues of the periodic orbits. This result was roughly contemporaneous with similar theoretical results by D. Aeyels and a more empirical account by Packard et al. (1980).

The idea of using time delayed coordinates to represent a system state is reminiscent of the theory of ordinary differential equations, where existence theorems say that a unique solution exists for each $[y(t), \dot{y}(t), \ddot{y}(t), \ldots]$. For example, in Newtonian many-body dynamics, current knowledge of the position and momentum of each body suffices to uniquely determine the future dynamics. The time derivatives can be approximated by delay-coordinate terms as

$$\left[y(t), \frac{y(t)-y(t-\tau)}{\tau}, \right.$$
$$\left. \frac{y(t)-2y(t-\tau)+y(t-2\tau)}{\tau^2}, \ldots \right].$$

The emergence of chaos and fractal geometry in physical systems motivated a reassessment of the original theory, which applies to smooth manifold attractors. It was shown (Sauer et al., 1991) that an attractor of (possibly fractional) box-counting dimension d can always be reconstructed with m generic observations, or with m time-delayed versions of one generic observation, where m is any integer greater than $2d$.

Embedding ideas were later extended beyond autonomous systems with continuously measured time series. A version was designed for excitable media, where information may be transmitted by spiking events, extending usage to possible neuroscience applications. An embedding theorem for skew systems by J. Stark explores extensions of the methodology when one part of a system is driving another.

Although the theory implies that an arbitrary time delay is sufficient to reconstruct the attractor, efficiency with a limited amount of data is enhanced by particular choices of the time delay τ. Methods for choosing an appropriate time delay have centered on measures of linear autocorrelation and mutual information (Fraser & Swinney, 1986). Further, in the absence of knowledge of the phase space dimension n, a choice of the number of embedding dimensions m must also be made. A number of ad hoc methods have been proposed that try to estimate whether the image has been fully unfolded by a given m-dimensional map.

The success of embedding in practice depends heavily on the specifics of the application. In particular, the hypothesis of a generic observation function creating the time series is often problematic. A mathematically generic observation monitors by definition all degrees of freedom of the system. The extent to which this is true affects the faithfulness of the reconstruction. If there is only a weak connection from some degrees of freedom to the observation function, the data requirements for a satisfactory reconstruction may be prohibitive in practice. Other factors that limit success are difference in time scales between different parts of the system, as well as system and observational noise.

Applications of embedding time-series data (Ott et al., 1994; Kantz & Schreiber, 1997) have been extensive since Takens's theorem was published. Many techniques of system characterization and identification were made possible, including determination of periodic orbits and symbolic dynamics, as well as approximation of attractor dimensions and Lyapunov exponents of chaotic dynamics. In addition, researchers have focused on methods of time series prediction and nonlinear filtering for noise reduction, the use of chaotic signals for communication, and the control of chaotic systems.

Tim Sauer

See also **Chaos vs. turbulence; Controlling chaos; Fractals; Phase space; Time series analysis**

Further Reading

Eckmann, J.-P. & Ruelle, D. 1985. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57: 617–652

Fraser, A.M. & Swinney, H.L. 1986. Independent coordinates for strange attractors from mutual information. *Physical Review* A, 33: 1134–1140

Kantz, H. & Schreiber, T. 1997. *Nonlinear Time Series Analysis*, Cambridge and New York: Cambridge University Press

Ott, E., Sauer, T. & Yorke, J.A. 1994. *Coping with Chaos: Analysis of Chaotic Data and the Exploitation of Chaotic Systems*, New York: Wiley Interscience

Packard, N., Crutchfield, J., Farmer, D. & Shaw, R. 1980. Geometry from a time series. *Physical Review Letters*, 45: 712–715

Sauer, T., Yorke, J.A. & Casdagli, M. 1991. Embedology. *Journal of Statistical Physics*, 65: 579–616

Takens, F. 1981. *Detecting Strange Attractors in Turbulence*, Berlin: Springer

EMERGENCE

Like many commonly used words, the term *emergence* has several meanings. In its weakest sense, a metaphor is provided by Michelangelo Buonarroti's famous sculptures entitled *The Prisoners*, which show human figures struggling to free themselves from their lithic confines, suggesting the artist's view of the creative process.

In simplest terms, there is little mystery here; the sculptor merely removes the unnecessary marble to expose the finished work, as Michelangelo himself is said to have pointed out. Thus, his prisoners are emerging only in an elementary sense which philosopher Robert Van Gulick calls "specific value emergence" and defines as follows (Van Gulick, 2001):

> SPECIFIC VALUE EMERGENCE: The whole and its parts have features of the *same kind,* but have different *specific subtypes* or *values* of that kind. For example, a bronze statue has a given mass as does each of the molecular parts of which it is composed, but the mass of the whole is different in value from that of any of its material parts.

Moving beyond this limited sense, Van Gulick defines various degrees of "modest emergence" in these terms.

> MODEST EMERGENCE: The whole has features that are *different in kind* from those of its parts (or alternatively that *could* be had by its parts). For example, a piece of cloth might be purple in hue even though none of the molecules that make up its surface could be said to be purple. Or a mouse might be alive even if none of its parts (or at least none of its subcellular parts) were alive.

Modest emergence thus arises in a spectrum of different ways depending upon the degree of difference between a phenomenon and the base out of which it emerges, with the coherent structures of nonlinear science providing many examples (Scott, 2003).

Among the more modest types of emergence, one would include solitons of the Korteweg–deVries (KdV) equation, which emerge out of a nonlinear partial differential equation (PDE) in response to certain initial conditions. Although KdV solitons are independent dynamic entities, their speeds and shapes are determined via the inverse scattering transform (IST) method from the initial conditions applied to the system. Somewhat less modest would be the various solitary wave solutions of Hamiltonian (energy conserving) systems for which IST formulations are not currently known and may not exist, precluding the prediction of solitary wave speeds from initial conditions.

Further decreasing modesty (increasing robustness) of emergence leads to the nerve impulse, which has several model PDEs (Hodgkin–Huxley, FitzHugh–Nagumo, etc.) in addition to those many physiological manifestations (the action potentials of the brain) which compose our mental activity (Scott, 2002). While propagating on a uniform system with constant speed and shape, a nerve impulse differs fundamentally from solitary waves of Hamiltonian systems for the following reason: a nerve impulse (like the flame of a candle) does not conserve energy. The nonlinear dynamics of a nerve impulse involves a balance between the release and dissipation of energy, so the process is open and thus does not have a Hamiltonian formulation. This, in turn, implies that the dynamic behavior of a nerve impulse changes greatly upon reversal of the direction of time, whereas the qualitative behavior of a Hamiltonian system is insensitive to time reversal.

Under this distinction, we can gauge the relative modesty of several other types of emergence that arise in the realms of nonlinear science. Vortex solutions of viscosity-free fluids (superfluids, for example) would be more modestly emergent than those of (more or less) viscous fluids, in which dissipative processes cause the dynamics to (more or less) rapidly forget the information received from the initial conditions. As residents of Tornado Alley in the U.S. midlands know well, tornados are famously ill-behaved, detached from their initial conditions and moving quite wildly in response to local variations of pressure, humidity, temperature, and so on.

A deeper meaning of Michelangelo's metaphor suggests the emergence of living organisms from the oily brine of the Hadean oceans some three thousand million years ago. Life is even less modestly (more robustly) emergent, with the "arrow of time" clearly constraining us all and playing a key role in the unpredictable drama of biological evolution (Gould, 1989). In other words, the emergence of life is far more intricate than the emergence of John Scott Russell's soliton from the prow of his test vessel on the Union Canal.

Yet more robust (less immodest) than the emergence of life is the phenomenon of human consciousness, which philosophers have struggled for centuries to understand and many find qualitatively different from its material substrate. To include the qualitative aspects of emergence at this far end of the scale, Van Gulick introduces the following definition.

> RADICAL EMERGENCE: The whole has features that are both different in kind from those had by its parts, and of a kind whose nature and existence is not necessitated by the features of its parts, their mode of combination and the law-like regularities governing the features of its parts.

Whether human consciousness offers an example of radical emergence is currently controversial among cognitive scientists, neuroscientists, psychologists, cultural anthropologists, philosophers, and others interested in the nature of mind. On the one hand, reductive materialists assert that all of reality must "in principle" reduce to a physical basis (Kim, 1999), whereas substance dualists claim that the human mind differs ontologically (in its nature) from physical reality (Chalmers, 1996). Situated between these two positions, property dualists suggest that the human mind may radically emerge from intricate interactions among the various nonlinear dynamic levels of body, brain, and culture (Scott, 1995; Van Gulick, 2001).

ALWYN SCOTT

See also **Biological evolution; Game of life; Morphogenesis, biological**

Further Reading

Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*, Oxford and New York: Oxford University Press

Gould, S.J. 1989. *Wonderful Life: The Burgess Shale and the Nature of History*, New York and London: Norton

Kim, J. 1999. *Mind in a Physical World*, Cambridge, MA: MIT Press

Scott, A.C. 1995. *Stairway to the Mind: The Controversial New Science of Consciousness*, Berlin and New York: Springer

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Van Gulick, R. 2001. Reduction, emergence and other recent options on the mind/body problem. *Journal of Consciousness Studies*, 8(9–10): 1–34

ENDOMORPHISM

See **Maps**

ENERGY ANALYSIS

In time-reversible Hamiltonian systems, the total energy of the system is constant in time. A local change of the energy density W at a point x in time t is balanced by the energy flux S from/to the point x according to the energy balance equation:

$$\frac{\partial W}{\partial t} + \frac{\partial S}{\partial x} = 0.$$

In irreversible (active and/or dissipative) systems, the total energy changes in time due to the energy sinks/sources of density ρ are given by the extended the extended energy balance equation:

$$\frac{\partial W}{\partial t} + \frac{\partial S}{\partial x} = -\rho.$$

In mathematical physics, energy balance is used for analysis of well-posedness of partial differential equations. Solutions of well-posed differential equations remain bounded in a suitable function space, starting with a bounded initial data (Strauss, 1992). For illustration, we consider the energy balance for the heat equation $u_t = u_{xx}$ that takes the form

$$\frac{\partial}{\partial t}\left(u^2\right) - 2\frac{\partial}{\partial x}\left(uu_x\right) = -2\left(u_x\right)^2.$$

Suppose the initial data $u(x, 0)$ belong to space of real-valued $L^2(\mathcal{R})$ functions such that the initial energy is bounded: $E(0) = \int_{-\infty}^{\infty} u^2(x, 0)\,\mathrm{d}x < \infty$. The energy $E(t)$ decreases at later times such that $0 \le E(t) = \int_{-\infty}^{\infty} u^2(x, t)\,\mathrm{d}x \le E(0)$. These simple estimates of energy analysis immediately imply the following properties of the heat equation:

(i) The zero solution $u(x, t) = 0$ is unique.
(ii) A general solution $u(x, t)$ is asymptotically stable in space of $L^2(\mathcal{R})$ functions.
(iii) A general solution $u(x, t)$ decays uniformly to zero in the L^2-norm sense: $\lim_{t\to\infty} ||u(\cdot, t)||_{L^2(\mathcal{R})} = 0$.

A real-valued energy can be introduced for complex functions, for example, in the Schrödinger equation $\mathrm{i}u_t = u_{xx}$, where the energy balance takes the form

$$\frac{\partial}{\partial t}|u|^2 + \mathrm{i}\frac{\partial}{\partial x}\left(\bar{u}u_x - \bar{u}_x u\right) = 0.$$

The Schrödinger equation is reversible, and the total energy is constant in time such that $0 \le E(t) = \int_{-\infty}^{\infty} |u|^2(x, t)\,\mathrm{d}x = E(0)$. As a result, the solution is well-posed in the space of complex-valued $L^2(\mathcal{R})$ functions with the following properties:

(i) The zero solution $u(x, t) = 0$ is unique.
(ii) A general solution $u(x, t)$ is neutrally stable in the L^2-norm sense.

Partial differential equations can be well-posed in energy space, which is different from the space of square integrable functions. For the wave equation $u_{tt} = c^2 u_{xx}$, two balance equations take the form

$$\frac{\partial}{\partial t}\left(u_t^2 + c^2 u_x^2\right) - 2c^2\frac{\partial}{\partial x}\left(u_t u_x\right) = 0$$

and

$$\frac{\partial}{\partial t}(u_t u_x) - \frac{1}{2}\frac{\partial}{\partial x}\left(u_t^2 + c^2 u_x^2\right) = 0.$$

The first equation prescribes the positive-definite quantity, $E(t) = \int_{-\infty}^{\infty}\left(u_t^2 + c^2 u_x^2\right)\mathrm{d}x$, which defines the energy space $H^1(\mathcal{R})$ for solutions of the wave equation, such that $0 \le E(t) = E(0)$. The second equation prescribes the sign-indefinite quantity, $P(t) = \int_{-\infty}^{\infty} u_t u_x \,\mathrm{d}x$, referred to as the momentum, such that $P(t) = P(0)$.

In modeling of various physical phenomena, momentum balance equations are useful for analysis of integral properties of solutions of underlying equations at infinite or finite intervals. For instance, adiabatic dynamics of localized pulses, envelope wave packets, and radiative wave trains can be studied with the momentum balance equations, when a solitary wave changes under the action of external perturbations, variations of physical parameters, internal instabilities, various resonances, and interactions with other wave structures (Kivshar & Malomed, 1989). In the simplest version of the soliton perturbation theory, effects of external perturbations to a physical system are captured by slow variations of soliton parameters. The dynamical rate of change of soliton parameters is found by substituting the soliton solutions in the momentum balance equations. For illustration, we consider the perturbed sine-Gordon equation with dissipative and external harmonic forces:

$$\begin{aligned} u_{tt} - u_{xx} + \sin u &= \varepsilon R(u) \\ &= \varepsilon\left[-\alpha u_t + \Gamma \sin(\omega t)(1 - \cos u)\right], \end{aligned}$$

where $\varepsilon \ll 1$, α is a damping parameter and Γ is the amplitude of the external force such that $\omega < 1$ (no resonance occurs). With the account of the perturbation, the balance equation for momentum is

$$\frac{\partial}{\partial t}(u_t u_x) - \frac{1}{2}\frac{\partial}{\partial x}\left(u_t^2 + u_x^2 + 2\cos u\right) = \varepsilon R(u) u_x.$$

As a result, the rate of change of momentum $P(t) = \int_{-\infty}^{\infty} u_t u_x \, dx$ is given by

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \varepsilon \int_{-\infty}^{\infty} u_x R[u]\mathrm{d}x.$$

provided that $\lim_{|x|\to\infty}\left(u_t^2 + u_x^2\right) = 0$ and $\lim_{x\to\infty} \cos u = \lim_{x\to-\infty}\cos u = 1$. The unperturbed sine-Gordon equation at $\varepsilon = 0$ has the kink solution:

$$u_k(x,t) = 4\arctan\left(\exp\left(\frac{x - vt - x_0}{\sqrt{1 - v^2}}\right)\right),$$

where $|v| < 1$ is the kink's velocity and x_0 is its position. The momentum of the unperturbed kink is a function of its velocity: $P_k(v) = -8v/\sqrt{1 - v^2}$. The kink solution satisfies the aforementioned vanishing conditions at infinity. Assuming that the velocity of the kink $v = v(t)$ changes due to the external perturbation, we use the momentum balance equation and find a particle-like equation of motion for the kink's adiabatic dynamics:

$$\frac{\mathrm{d}P_k}{\mathrm{d}t} + \varepsilon\,\alpha P_k = 2\pi\,\varepsilon\,\Gamma\sin(\omega t).$$

The equation has a simple solution:

$$\begin{aligned} P_k(t) &= C\mathrm{e}^{-\varepsilon\alpha t} \\ &\quad - \frac{2\pi\varepsilon\omega\Gamma}{\omega^2 + \varepsilon^2\alpha^2}\cos\omega t \\ &\quad + \frac{2\pi\varepsilon^2\alpha\Gamma}{\omega^2 + \varepsilon^2\alpha^2}\sin\omega t, \end{aligned}$$

where C is found from initial condition: $P_k(0) = P_0$.

Adiabatic dynamics of kinks and solitary waves often generate a strong radiation field that takes away part of momentum of the localized solution. The radiation field can be taken into account from other balance equations, such as the mass balance. Radiative effects usually occur in the second order of the perturbation theory, leading to radiative decay of solitary waves or their perturbations (Pelinovsky & Grimshaw, 1996). For illustration, we consider dynamics of solitary waves in the critical KdV equation:

$$u_t + 15u^4 u_x + u_{xxx} = 0.$$

The balance equations for mass and momentum of a nonlinear field are:

$$\frac{\partial}{\partial t}(u) + \frac{\partial}{\partial x}\left(3u^5 + u_{xx}\right) = 0$$

and

$$\frac{\partial}{\partial t}(u^2) + \frac{\partial}{\partial x}\left(5u^6 + 2uu_{xx} - u_x^2\right) = 0.$$

Solitary waves are given by special solutions of the critical KdV equation:

$$u_{\mathrm{s}}(x,t) = \left[\sqrt{v}\,\operatorname{sech}(2\sqrt{v}(x - vt - x_0))\right]^{1/2}.$$

Solitary waves may change adiabatically due to internal perturbation of the initial data, such that v becomes a function of t. The radiation field is generated behind the solitary wave due to the uni-directional property of the dispersion relation for the linear KdV equation: $u_t + u_{xxx} = 0$. The radiation field can be found from the mass balance equation

$$\lim_{x\to-\infty} u(x,t) = -\frac{1}{v}\frac{\mathrm{d}M_{\mathrm{s}}}{\mathrm{d}t},$$

where $M_{\mathrm{s}}(v) = \int_{-\infty}^{\infty} u_{\mathrm{s}}(x,t)\,\mathrm{d}x = M_0/v^{1/4}$ and M_0 is constant. The momentum balance equation leads to a closed particle-like equation of motion for the solitary wave's adiabatic dynamics:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\left(\frac{\mathrm{d}M_s}{\mathrm{d}v}\right)^2 \frac{\mathrm{d}v}{\mathrm{d}t}\right] = -v\left[\lim_{x\to-\infty} u(x,t)\right]^2$$
$$= -\frac{1}{v}\left(\frac{\mathrm{d}M_s}{\mathrm{d}v}\right)^2\left(\frac{\mathrm{d}v}{\mathrm{d}t}\right)^2.$$

The dynamical equation has a unique solution:

$$v(t) = v_0\left(\frac{t_0}{t_0 - t}\right)^2,$$

where v_0 and t_0 are constants. The exact solution defines the scaling law for self-similar blowup in the critical KdV equation. The blowup rate is modified due to generation of the radiation field, which makes the balance between the left- and right-hand sides of the momentum balance equation. To summarize, the balance equations for energy, momentum, and mass can be used for both qualitative and quantitative estimates of the interaction between localized and radiative components of the nonlinear wave in solutions of nonlinear wave equations.

DMITRY PELINOVSKY

See also **Constants of motion and conservation laws; Multisoliton perturbation theory; Power balance**

Further Reading

Kivshar, Yu.S. & Malomed, B.A. 1989. Dynamics of solitons in nearly integrable systems. *Reviews of Modern Physics*, 61: 763–915

Pelinovsky, D.E. & Grimshaw, R.H.J. 1996. An asymptotic approach to solitary wave instability and critical collapse in long-wave KdV-type evolution equations. *Physica* D, 98: 139–155

Strauss, W.A. 1992. *Partial Differential Equations: An Introduction*, New York: Wiley

ENERGY CASCADE

See **Turbulence**

ENERGY OPERATORS

See **Quantum nonlinearity**

ENTRAINMENT

See **Coupled oscillators**

ENTROPY

Entropy is a quantity characterizing disorder, or randomness. This conclusion is the result of successive advances since the pioneering works by Sadi Carnot on the fundamentals of steam engines in 1824 and by Rudolf Clausius, who, between 1851 and 1865, developed the concept of entropy (from $\tau\rho o\pi\eta$ meaning "transformation" in Greek) as a thermodynamic state variable. Called S, this quantity varies as $\mathrm{d}S = \mathrm{d}Q/T$ in an equilibrium system at temperature T exchanging a quantity $\mathrm{d}Q$ of heat and no matter reversibly with its environment.

Thereafter, Ludwig Boltzmann (1896/1898), Max Planck (1901), Josiah Willard Gibbs (1902), and others discovered the statistical meaning of entropy. Boltzmann and Gibbs introduced the concepts of probability and statistical ensemble in the context of mechanics. If P_α is the probability (satisfying $0 \le P_\alpha \le 1$ and $\sum_\alpha P_\alpha = 1$) that the system is found in the microstate α specified by a set of observables such as energy and linear or angular momenta, the thermodynamic entropy is given by

$$S = -k_\mathrm{B}\sum_\alpha P_\alpha \ln P_\alpha, \qquad (1)$$

where $k_\mathrm{B} = 1.38065\ 10^{-23}\,\mathrm{J\,K^{-1}}$ is the so-called Boltzmann constant, although it was originally introduced by Planck (Sommerfeld, 1956).

The entropy S measures the disorder in a statistical ensemble composed of a very large number, $\mathcal{N}$, of copies of the system. Each copy is observed in a microstate α occurring with the probability P_α. The microstates of all the $\mathcal{N}$ copies form a random list $\{\alpha_1, \alpha_2, \ldots, \alpha_\mathcal{N}\}$. Typically, $\mathcal{N}_\alpha = \mathcal{N}P_\alpha$ copies are found in the microstate α, and $\sum_\alpha \mathcal{N}_\alpha = \mathcal{N}$. The copies being statistically independent, the total number of possible lists of the microstates of the ensemble is thus equal to

$$W = \frac{\mathcal{N}!}{\prod_\alpha \mathcal{N}_\alpha!}, \qquad (2)$$

where $\mathcal{N}! = \mathcal{N} \times (\mathcal{N} - 1) \times \cdots \times 3 \times 2$ denotes the factorial of the integer $\mathcal{N}$. In the limit $\mathcal{N} \to \infty$, according to Stirling's formula $\mathcal{N}! \simeq (\mathcal{N}/\mathrm{e})^\mathcal{N}$, where $\mathrm{e} = 2.718\ldots$ denotes the Naperian base, entropy (1) is given in terms of the logarithm of the number W of possible lists (originally called complexions by Boltzmann):

$$S \simeq \frac{k_\mathrm{B}}{\mathcal{N}} \ln W. \qquad (3)$$

If the system is perfectly ordered, all the copies in the ensemble would be in the same microstate, and there would be a single possible complexion $W = 1$ so that the entropy would vanish ($S = 0$). In contrast, if the system were completely disordered with A equiprobable microstates α, the entropy would take the maximum value $S = k_\mathrm{B} \ln A$. For partial disorder, the entropy takes an intermediate value.

In spatially extended homogeneous systems, the thermodynamic entropy is an extensive quantity. If the system is covered by $\mathcal{N}$ disjoint windows of observations of volume V, the microstates $\{\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{N}}\}$ in the $\mathcal{N}$ windows form one among W possible lists (2). If the volume V of the observation window is large enough, the thermodynamic entropy is again given by Equation (3). In this case, the entropy per unit volume obtained by dividing the entropy S by the volume V is a measure of spatial disorder.

Entropy can also be interpreted as a quantity of information required to specify the microstate of the system. Indeed, the recording of the random microstates $\{\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{N}}\}$ of the statistical ensemble requires one to allocate at least $I = \log_2 W$ bits of information in the memory of a computer. This number of bits is related to the entropy (1) by $I \simeq \mathcal{N} S/(k_B \ln 2)$. Such connections between entropy and information have been developed since the works by Leo Szilard in 1929 and Léon Brillouin around 1951.

Equation (1) for the entropy is very general. It applies not only to equilibrium but also to out-of-equilibrium systems provided the states α are understood as coarse-grained states. In a classical system of N particles, the coarse-grained states α should correspond to cells of volume h^{3N} in the phase space of the positions and momenta of the particles, where $h = 6.626\ 10^{-34}$ Js is Planck's constant of quantum mechanics.

In 1902, Gibbs suggested that the second law of thermodynamics is a consequence of a dynamics having the mixing property according to which the statistical averages of observables or the coarse-grained probabilities P_α converge to their equilibrium values. Not all systems are mixing, but for those that are, entropy (1) converges toward its equilibrium value. In mixing systems, the statistical correlations in the initial probability distribution tend to disappear on finer and finer scales in phase space and are shared among more and more particles during a causal time evolution. The approach to the thermodynamic equilibrium may thus be described in terms of asymptotic expansions of the probability distributions in the long-time limits $t \to \pm\infty$. Both limits are not equivalent because, in the limit $t \to +\infty$, the probability distributions remain smooth in the unstable phase-space directions but become singular in the stable ones and vice versa in the other limit $t \to -\infty$, which may appear as an irreversibility or time arrow in the long-time description. The irreversibility in the increase of the entropy is thus closely related to the problems of identifying all the degrees of freedom guaranteeing the causality of the time evolution and of reconstructing the initial conditions, which is of great importance for the understanding of historical processes such as biological and cosmological evolution.

During recent decades, it has been shown that the increase of entropy does not preclude the formation of structures in equilibrium or non-equilibrium systems, nor in self-gravitating systems. At equilibrium, the homogeneity of pressure, temperature, and chemical potentials does not prevent inhomogeneities in the densities as is the case in crystals, in vortex states of quantum superfluids, and in mesomorphic phases of colloidal systems where equilibrium self-assembly occurs. Besides, open non-equilibrium systems can remove entropy to their environment, leading to far-from-equilibrium self-organization into spatial structures such as Turing patterns, self-sustained oscillations in such systems as chemical clocks, or complex processes such as biological morphogenesis.

While the entropy per unit volume characterizes spatial disorder at a given time, a concept of entropy per unit time was introduced in 1949 by Claude Shannon in development of his information theory, as a measure of temporal disorder in random or stochastic processes. It is defined in the same way as standard entropy but replacing space by time. In 1959, Andrei N. Kolmogorov and Yakov G. Sinai applied Shannon's idea to deterministic dynamical systems with an invariant probability measure, and they defined a metric entropy per unit time in analogy with Equation (1), considering the states α as the sequences $\omega_1\omega_2\ldots\omega_n$ of the phase space cells ω_j successively visited at time intervals Δt by the trajectories during the time evolution of the system. In order to get rid of the arbitrariness of the coarse-grained cells ω_j, Kolmogorov and Sinai considered the supremum (least upper bound) of the entropy per unit time with respect to all possible partitions $\mathcal{P}$ of the phase space into cells ω_j, defining

$$\begin{aligned} h_{\mathrm{KS}} &= \mathrm{Sup}_{\mathcal{P}} \lim_{n\to\infty} -\frac{1}{n\Delta t} \\ &\quad \times \sum_{\omega_1\omega_2\ldots\omega_n} P_{\omega_1\omega_2\ldots\omega_n} \ln P_{\omega_1\omega_2\ldots\omega_n}, \qquad (4) \end{aligned}$$

where the probability P is evaluated with the given invariant measure (Cornfeld et al., 1982).

In isolated chaotic dynamical systems, the temporal disorder of the trajectories finds its origin in the sensitivity to initial conditions because the Kolmogorov–Sinai entropy per unit time is equal to the sum of positive Lyapunov exponents λ_i, $h_{\mathrm{KS}} = \sum_{\lambda_i>0} \lambda_i$, as proved by Yakov B. Pesin in 1977 (Eckmann & Ruelle, 1985). We notice that the entropy per unit time h_{KS} differs from the irreversible entropy production defined by the time derivative of the standard thermodynamic entropy S in an isolated system. Indeed, the entropy per unit time may take a positive value for a system of particles already at thermodynamic equilibrium where entropy production vanishes. In spatially extended chaotic systems, the spatiotemporal disorder can be characterized by a further concept of entropy per unit time and unit volume.

A so-called topological entropy has also been introduced as the rate of proliferation of cells in successive partitions iteratively refined by the dynamics (Eckmann & Ruelle, 1985). In finite chaotic systems, the topological entropy is the rate of proliferation of periodic orbits as a function of their period. The topological entropy is not smaller than the Kolmogorov–Sinai entropy: $h_{\text{top}} \geq h_{\text{KS}}$.

PIERRE GASPARD

See also **Algorithmic complexity; Biological evolution; Cosmological models; Information theory; Lyapunov exponents; Measures; Mixing; Morphogenesis, biological; Nonequilibrium statistical mechanics; Pattern formation; Phase space; Stochastic processes; Turing patterns**

Further Reading

Boltzmann, L. 1896/1898. *Vorlesungen über Gastheorie*, 2 vols, Leipzig: Barth; as *Lectures on Gas Theory*, translated by S.G. Brush, Berkeley, University of California Press, 1964

Cornfeld, I.P., Fomin, S.V. & Sinai, Ya.G. 1982. *Ergodic Theory*, Berlin: Springer

Eckmann, J.-P. & Ruelle, D. 1985. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57: 617–656

Gibbs, J.W. 1902. *Elementary Principles in Statistical Mechanics*, New Haven, CT: Yale University Press

Planck, M. 1901. Uber das Gesetz der Energieverteilung im Normalspektrum. *Annalen der Physik*, 4: 553–563 (First historical publication of $S = k_{\text{B}} \ln W$)

Sommerfeld, A. 1956. *Thermodynamics and Statistical Mechanics*, New York: Academic Press

ENVELOPE EQUATIONS

See **Nonlinear Schrödinger equations**

ENVELOPE SOLITONS

See **Solitons, types of**

EPHAPTIC COUPLING

Neurons communicate with each other by different means. The most investigated of these is synaptic transmission, which can be either chemical or electrical. Chemical synapses transmit neural signals from presynaptic to postsynaptic membranes via neurotransmitters and are widely found in vertebrate neurons, while electrical synapses correspond to an electronic coupling through specialized gap junctions and are more common in invertebrates (Jefferys, 1995). Therefore, both transmission processes need a specialized anatomical structure to create points of contact between neurons but there are other possibilities.

Since the work of Elwald Hering in 1882, it has been known that electrical communication can also occur between neurons when neuronal membranes are closely apposed but not contiguous (Scott, 2002). Called ephaptic coupling by Angelique Arvanitaki (who defined the term *ephapse* as "... the locus of contact or close vicinity of the active functional surfaces, whether this contact be experimental or brought by natural means"), this process relies on current spread through the extracellular space, which may influence the membrane dynamics of a second fiber (Arvanitaki, 1942). Early experimental evidence of ephaptic interactions was also provided by Bernhard Katz and Otto Schmitt on a pair of naturally adjacent unmyelinated fibers from the limb nerve of a crab (Katz & Schmitt, 1940). They showed that an impulse traveling on one fiber changes the excitability of the other fiber. Furthermore, impulses on adjacent fibers with similar speed and launched at about the same time become synchronized.

In the 1970s, Markin proposed a theoretical description of ephaptic coupling between two parallel unmyelinated fibers based on the assumption that they share an external series resistance per unit length proportional to the ionic resistivity of the extracellular medium (Markin, 1970). Using a piecewise constant function for the transmembrane ionic current, yielding to a leading-edge analysis, Markin concluded that under normal physiological conditions, ephaptic coupling does not allow transmission of an impulse from a fiber to an adjacent one, as verified in numerous experiments. Nevertheless, he also suggested that a synchronization of two impulses is to be expected with a longitudinal distance δ equal to zero.

A recent leading-edge analysis (Scott, 2002) considered a more appropriate representation of the ionic current (sodium ions), the resulting system corresponding to the Zeldovich–Frank-Kamenetsky (ZF) equation. In the case of a small ephaptic coupling, a perturbation theory showed that $\delta = 0$ corresponds to a stable locking of pairs of impulses if the threshold parameter a of the membrane is above a critical value, otherwise stability occurs with δ increasing when a decreases.

Other possibilities of synchronization have been found when studying the influence of complete impulses including a leading edge and a recovery part. Based on the FitzHugh–Nagumo model, studies have shown that two other locking distances δ are also stable and separated by two unstable ones (Scott, 2002; Eilbeck et al., 1981). Recently, it has also been shown that ephaptic coupling influences the speed of conduction of synchronized impulses, the speed decreasing when the coupling increases (Binczak et al., 2001).

In mammals, motor or sensory nerves are often organized in fiber bundles. These fibers are myelinated, and the active nodes of the membrane are separated

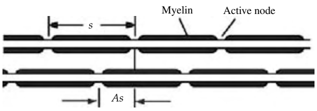

Figure 1. Sketch of two adjacent and parallel myelinated nerve fibers.

by a myelin sheath, which acts as an insulator. Thus, ephaptic coupling may be influenced by the alignment of the active nodes between two adjacent fibers (Binczak et al., 2001), as illustrated in Figure 1 where A is an alignment parameter.

Using a ZF description for the ionic current, a leading-edge analysis suggests that synchronization of impulses occurs whatever the alignment A. Nevertheless, an alignment of the nodes ($A = 1$) leads to a stronger synchronization while staggered nodes ($A = \frac{1}{2}$) allow a broader impulse coupling. Furthermore, when impulses are synchronized, an alignment of adjacent nodes reduces the critical longitudinal internodal distance at which propagation fails. Staggered nodes, on the other hand, lead to a more robust medium in order to prevent propagation failure.

From a functional perspective, synchronization of impulses on parallel and adjacent fibers might provide a means to adjust and organize the timing of impulses necessary for coordination and computation of neuronal information. Furthermore, ephaptic interactions may be important in neurological pathophysiology. Indeed, it has been observed that the transmission of impulses occurs from a fiber to an adjacent one when nerves are damaged, as when the nerves end in a neurisma after a nerve crush injury or a nerve compression (Seltzer & Devor, 1979) or when the ionic composition of the extracellular medium is changed (Ramon & Moore, 1978). Demyelination diseases, such as multiple sclerosis, are also a cause of ephaptic connections leading to pathological synchronization (Jefferys, 1995). Finally, ephaptic phenomena have been reported between muscles and motor nerves causing possible cramps and spasms and between cardiac cells, implying a possible involvement of ephaptic coupling in cardiac arrhythmias (Suenson, 1984).

Stephane Binczak

See also **FitzHugh–Nagumo equation; Myelinated nerves; Neurons; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Arvanitaki, A. 1942. Effects evoked in an axon by the activity of a contiguous one. *Journal of Neurophysiology*, 5: 89–108

Binczak, S., Eilbeck, J.C. & Scott, A.C. 2001. Ephaptic coupling of myelinated nerve fiber. *Physica D*, 148: 159–174

Eilbeck, J.C., Luzader, S.D. & Scott, A.C. 1981. Pulse evolution on coupled nerve fibers. *Bulletin of Mathematical Biology*, 43(3): 389–400

Jefferys, J.G.R. 1995. Nonsynaptic modulation of neuronal activity in the brain: electric currents and extracellular ions. *Physiological Reviews*, 75 (4): 689–723

Katz, B. & Schmitt, O.H. 1940. Electric interaction between two adjacent nerve fibers. *Journal of Physiology*, 97: 471–488

Markin, V.S. 1970. Electrical interactions of parallel non-myelinated fibers. I and II. *Biophysics*, 15: 122–133 and 713–721

Ramon, F. & Moore, J.W. 1978. Ephaptic transmission in squid giant axons. *American Journal of Physiology: Cell Physiology*, 234: C162–C169

Seltzer, Z. & Devor, M. 1979. Ephaptic transmission in chronically damaged peripherical nerves. *Neurology*, 29: 1061–1064

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Suenson, M. 1984. Ephaptic impulse transmission between ventricular myocardial cells in vitro. *Acta Physiologica Scandinavica*, 120: 445–455

EPIDEMIOLOGY

Throughout human history diseases have played an important role. The black death in Europe in the middle ages is one example, and more recent examples include the flu pandemic of 1918 and 1919 and diseases such as AIDS and various childhood diseases. Additionally, the study of population dynamics in ecology has been greatly advanced by the study of diseases for several reasons. First, the data on the incidence of childhood diseases are both extensive and accurate. Second, the processes involved in the dynamics of childhood diseases (essentially, infection and either recovery or death) are relatively simple and straightforward and well understood.

In particular, many questions of scientific or practical interest can be answered using relatively simple models. A fundamental question is why does a disease die out before everyone has the disease? And, what fraction of the population needs to be vaccinated so a disease can be controlled or eliminated? Answering this question can explain why smallpox was more easily eradicated than other so-called childhood diseases.

Perhaps the simplest epidemic model, which also introduces many of the ideas, is the case of a single epidemic of a disease, first studied in detail by Kermack and McKendrick (1927). We focus here on diseases that are caused by microparasites, so individuals either have the disease or do not, as opposed to diseases caused by macroparasites (such as tapeworms) where the number of infectious agents in individuals needs to be considered explicitly in order to understand the disease dynamics. Different models are needed to understand epidemics and diseases which are endemic (Kermack & McKendrick, 1932).

We assume a constant population size and divide the population into three classes: susceptible, infective, and removed. Since the time scale of an epidemic is much shorter than the time scale of changes in human population sizes, we can simplify the system by ignoring any demographic influences in the simplest model for an epidemic, which corresponds to the assumption of a constant population size. In other words, the population size is assumed constant, and births and deaths are ignored (although there are models that do take this into account) because the time scale of an epidemic is short relative to the time scale of human population dynamics.

We assume that the rate at which susceptibles become infected is simply proportional to the product of the number of susceptible and infective individuals, corresponding to an assumption of random encounters. The rate at which infective individuals recover is assumed to be a constant. Then, with S the number of susceptibles, I the number of infectives, and R the number of removed individuals, the dynamics are given by

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI, \tag{1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I, \tag{2}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I. \tag{3}$$

Under the assumption that the total population size N remains constant, we can use the relationship $N = S + I + R$, and reduce the system (1)–(3) to one based on just the first two equations. The phase plane analysis of the resulting system is facilitated by the fact that the formula

$$\frac{\mathrm{d}I}{\mathrm{d}S} = \frac{\gamma}{\beta S} - 1 \tag{4}$$

is so simple. As first discussed by Kermack and McKendrick (1927), one can solve this system explicitly by integration and using approximations. From this one can see that the solution curves which start with I arbitrarily small return to $I = 0$ before all individuals in the population become infected, or in other words the number of susceptibles at $t = \infty$ is positive.

The qualitative behavior of the system (1)–(3) essentially depends on a single nondimensional parameter, the reproductive number for the disease. The reproductive number is defined as the mean number of infective individuals produced by a single infective individual, which can clearly be calculated by multiplying the rate at which a single infective individual produces new infections by the mean period of infectivity for a single individual. Under our assumptions, the mean rate of infection is simply βS and the mean infective period is $1/\gamma$. Thus the reproductive number is simply

$$R_0 = \frac{\beta S}{\gamma}. \tag{5}$$

The dynamics are governed by the observation that the number of infectives will increase if $R_0 > 1$ and will decrease if $R_0 < 1$. For the case of a single infective introduced into a population of susceptibles, we can use the total population N instead of S in the formula for R_0.

The import of the observation about the importance of R_0 (and results from integrating (4)) is typically summarized in the *threshold theorem* which states that there will be an epidemic if the population initially satisfies $R_0 > 1$, which may be a reflection of the population size. From integrating (4) one finds that the total number of individuals who get the disease is dramatically larger if the population is initially above the threshold.

Kermack & McKendrick (1932) further studied the cases of endemic diseases, where it is necessary to take into account the demography of the population, since the time scale is long enough that births and deaths need to be explicitly included. Other more complex systems were also studied (Kermack & McKendrick, 1933).

Modifications of the basic equations have been used to study the dynamics of sexually transmitted diseases, including AIDS. Here, an important modification has been to break up the population into classes based on different encounter rates.

More recent work on disease dynamics has played a central role in population biology, as the data sets for the incidence of childhood diseases in the United Kingdom and many large U.S. cities are among the longest, most accurate, and most detailed records of populations available (Bjornstad & Grenfell, 2001). In particular, the prevaccination dynamics of measles has been carefully analyzed using a variety of approaches, based essentially on extensions of the basic model (Equations (1)–(3)), modified to include a seasonally varying contact rate and, in some cases, stochasticity. Studies of this data has led to important substantial advances in the analysis of the kinds of time series available: relatively short with substantial stochastic influences. Analysis of these time series using nonlinear methods have demonstrated that at least some of the dynamics may be chaotic. Further and more recent efforts have focussed on spatiotemporal dynamics (Rohani et al., 1999) and on applied questions (Keeling et al., 2002) like the recent hoof and mouth epidemic in the UK.

Alan Hastings

See also **Chaotic dynamics; Phase plane; Population dynamics**

Further Reading

Bjornstad, O.N. & Grenfell, B.T. 2001. Noisy clockwork: time series analysis of population fluctuations in animals. *Science*, 293: 638–643

Diekmann, O. & Heesterbeek, J.A.P. 2000. *Mathematical Epidemiology of Infectious Diseases*, Chichester and New York: Wiley

Keeling, M.J., Woolhouse, M.E.J., Shaw, D.J., Matthews, L., Chase-Topping, M., Haydon, D.T., Cornell, S.J., Kappey, J., Wilesmith, J. & Grenfell, B.T. 2002. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, 294: 813–817

Kermack, W.O. & McKendrick, A.G. 1927. Contributions to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, 115: 700–721

Kermack, W.O. & McKendrick, A.G. 1932. Contributions to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, 138: 55–83

Kermack, W.O. & McKendrick, A.G. 1933. Contributions to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, 141: 94–122

Rohani, P., Earn, D.J.D. & Grenfell, B.T. 1999. Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, 286: 968–971

EQUATIONS, NONLINEAR

Nonlinear equations arise in a wide variety of forms in all branches of science, engineering, and technology as well as in other fields such as economics and social dynamics. These include algebraic, differential (ordinary/partial/delay), difference, differential-difference, integro-differential, and integral equations.

Nonlinear algebraic equations such as polynomial (quadratic, cubic, quartic, etc.) and transcendental equations in one or more variables and simultaneous equations have a chequered history. A large number of root-searching algorithms and methods of solutions are available in the literature.

With the advent of differential calculus in the 17th century, differential equations started to play a pivotal role in scientific investigations. This is particularly so for evolutionary problems, as differential equations are obvious candidates for the dynamical description of natural phenomena. Since most natural processes are nonlinear, it is no wonder that nonlinear differential equations arise frequently in theoretical descriptions (Ablowitz & Clarkson, 1991; Murray, 2002; Lakshmanan & Rajasekar, 2003; Scott, 2003).

Nonlinear ordinary differential equations (ODEs) occur in a wide variety of situations depending upon the nature of the problem. Their solution properties, such as integrability, non-integrability, and chaos, depend upon the order of the highest derivative, number of dependent variables, degree of nonlinearity, whether homogeneous or inhomogeneous, and on the values of the parameters. Some ubiquitous nonlinear ODEs and their significance are indicated in Table 1.

Nonlinear partial differential equations (PDEs) also have a long history dating back to the early days of differential calculus. For example, the basic equations of fluid dynamics, such as the Euler equation and the Navier–Stokes equation, are highly nonlinear. Many of the equations describing nontrivial surfaces in geometry are essentially nonlinear. The properties of solutions of nonlinear PDEs depend heavily on the number of dependent and independent variables, the order, the nature of nonlinearity, and the boundary conditions. However, it is also useful to classify the nonlinear PDEs into nonlinear dispersive and nonlinear diffusive/dissipative types. Further, they can

	Name	Equation	Significance
1.	Logistic equation	$\dot{x} = ax - bx^2$	Population growth model
2.	Bernoulli equation	$\dot{x} + P(t)x = Q(t)x^n$	Linearizable
3.	Riccati equation	$\dot{x} + P(t) + Q(t)x + R(t)x^2 = 0$	Admits nonlinear superposition principle
4.	Lotka–Volterra equation	$\dot{x} = ax - xy,$ $\dot{y} = xy - by$	Population dynamics; exhibits limit cycle
5.	Anharmonic oscillator	$\ddot{x} + \omega_0^2 x + \beta x^3 = 0$	Integrable by Jacobian elliptic function(s)
6.	Pendulum equation	$\ddot{\theta} + \frac{g}{L}\sin\theta = 0$	Integrable by Jacobian elliptic function(s)
7.	Painlevé II equation	$\ddot{x} = 2x^3 + tx + \alpha$	Satisfies Painlevé property
8.	Duffing oscillator	$\ddot{x} + p\dot{x} + \omega_0^2 x + \beta x^3 = f\cos\omega t$	Exhibits chaotic dynamics
9.	Damped driven pendulum	$\ddot{\theta} + \alpha\dot{\theta} + \omega_0^2 \sin\theta = \gamma\cos\omega t$	Exhibits chaotic dynamics
10.	Lorenz equation	$\dot{x} = \sigma(y - x),$ $\dot{y} = -xz + rx - y,$ $\dot{z} = xy - bz$	Prototypical example of chaotic motion
11.	Hénon–Heiles system	$\ddot{x} + x + 2xy = 0,$ $\ddot{y} + y + x^2 - y^2 = 0$	Hamiltonian chaos
12.	Mackey–Glass equation	$\dot{x} + \frac{ax_\tau}{(1+x_\tau^{10})} + bx = 0,$ $x_\tau = x(t-\tau),\ \tau$: constant	Delay-differential system

Table 1. Some important nonlinear ordinary differential equations ($\cdot = \mathrm{d}/\mathrm{d}t$)

	Name	Equation	Significance
I	*Dispersive Equations*		
1.	Korteweg–de Vries equation	$u_t + 6uu_x + u_{xxx} = 0$	Integrable soliton eqn. in (1+1) dimensions
2.	Sine-Gordon equation	$u_{xt} = \sin u$	Integrable soliton eqn. in (1+1) dimensions
3.	Nonlinear Schrödinger equation	$\mathrm{i}q_t + q_{xx} + 2\|q\|^2 q = 0$	Integrable soliton eqn. in (1+1) dimensions
4.	Heisenberg ferromagnetic spin equation	$\mathbf{S}_t = \mathbf{S} \times \mathbf{S}_{xx},\quad \mathbf{S} = (S_1, S_2, S_3)$	Integrable soliton eqn. in (1+1) dimensions
5.	Kadomtsev–Petviashvili equation	$(u_t + 6uu_x + u_{xxx})_x + 3\sigma^2 u_{yy} = 0,$ $\sigma^2 = \pm 1$	Integrable soliton eqn. in (2+1) dimensions
6.	Davey–Stewartson equation	$\mathrm{i}q_t + \frac{1}{2}(q_{xx} + q_{yy}) + \alpha\|q\|^2 q + q\phi = 0,$ $\phi_{xx} - \phi_{yy} + 2\alpha(\|q\|^2)_{xx} = 0, \alpha = \pm 1$	Integrable soliton eqn. in (2+1) dimensions
7.	ϕ^4 equation	$u_{tt} - u_{xx} + u - u^3 = 0$	Nonintegrable equation
II	*Diffusive equations*		
8.	Burgers equation	$u_t + uu_x - u_{xx} = 0$	Linearizable through Cole–Hopf transformation
9.	FitzHugh–Nagumo equation	$u_t = u_{xx} + u - u^3 - v,$ $v_t = a(u - b)$	Nerve impulse propagation model
10.	Kuramoto–Sivashinsky equation	$u_t = -u - u_{xx} - u_{xxxx} - uu_x$	Spatiotemporal patterns and chaos
11.	Ginzburg–Landau equation	$u_t = (a + \mathrm{i}b)\nabla^2 u + (c + \mathrm{i}d)(\|u\|^2)u$	Spatiotemporal patterns and chaos

Table 2. Some important nonlinear partial differential equations (suffix denotes partial derivative with respect to that variable)

be classified into integrable and non-integrable PDEs. A select set of such equations along with their significance is given in Table 2.

Another class of interesting nonlinear equations is the so-called difference equations/recurrence equations/iterated maps of the form $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n, n)$, where the independent/time variable n takes discrete values $n = 0, 1, 2, \ldots,$ and $\boldsymbol{x}_n$ stands for the m dependent variables, while $\boldsymbol{F}$ is an m-dimensional nonlinear function. These equations often correspond to various finite-difference schemes of nonlinear ODEs/PDEs, such as the explicit Euler or implicit midpoint. For example, the logistic differential equation $\mathrm{d}x/\mathrm{d}t = ax - bx^2$ can be approximated by Euler's forward difference scheme as $x_{n+1} - x_n = h(ax_n - bx_n^2)$, where $h > 0$. But nonlinear difference equations can arise as dynamical systems on their own merit as in population dynamics or radioactive decay. The most famous example is the logistic map,

$$x_{n+1} = \lambda x_n(1 - x_n), \quad 0 \le x \le 1, \quad 0 \le \lambda \le 4,$$

as a prototypical model exhibiting a period-doubling bifurcation route to chaos. A two-dimensional example is the Hénon map, $x_{n+1} = 1 - ax_n^2 + y_n,\ y_{n+1} = bx_n,\ a, b > 0$ showing self-similar and fractal nature. Spatiotemporal patterns have been identified in coupled logistic map lattices. There also exist several integrable families of maps (Kosmann-Schwarzbach et al., 1997): for example, the McMillan map,

$$x_{n+1} + x_{n-1} = \frac{2\mu x_n}{1 - x_n^2}. \tag{1}$$

Other examples include integrable discrete Painlevé equations; Quispel, Robert, and Thompson (QRT) map; and so on. Several integrable nonlinear differential-difference and difference-difference equations also exist (Ablowitz & Clarkson, 1991). The Toda lattice equation

$$\begin{aligned} \ddot{x}_n &= \exp(-(x_n - x_{n-1})) - \exp(-(x_{n+1} - x_n)), \\ n &= 0, 1, 2, \ldots, N \end{aligned} \tag{2}$$

is an integrable soliton system. Examples of integrable nonlinear difference-difference equations can be constructed from integrable PDEs.

Nonlinear integro-differential equations arose (Davis, 1962) with the work of Vito Volterra when he introduced a hereditary component composed of the sum of individual factors encountered in the past as a modification to the logistic equation as

$$\frac{\mathrm{d}y}{\mathrm{d}t} = ay - by^2 + y\int_c^t K(t - s)y(s)\,\mathrm{d}s, \tag{3}$$

where a, b, c are parameters and $K(t)$ is a given function. A special case of such an equation is the nonlinear integral equation

$$y(x) = f(x) + \lambda \int_a^b K(x, s, y(s))\,\mathrm{d}s, \qquad (4)$$

where λ is a parameter. Some interesting special cases correspond to $f = 0$, $K = \bar{k}(x, t)y^n$ and $f = 0$, $K = G(x, t)\exp y$ (Bratu equation), where $\bar{k}$ and G are specified functions. One of the most important nonlinear integro-differential equations in physics is the Boltzmann transport equation for the distribution function $f(\boldsymbol{r}, \boldsymbol{v}, t)$ of a dilute gas

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v}_1 \cdot \nabla_r + \frac{\boldsymbol{F}}{m} \cdot \nabla_{\boldsymbol{v}_1}\right) f(\boldsymbol{r}, \boldsymbol{v}_1, t) = \int d^3 v_2 \int d\Omega\, \sigma(\Omega) |\boldsymbol{v}_2 - \boldsymbol{v}_1| (f_1' f_2' - f_1 f_2), \qquad (5)$$

where $f_i' = f(\boldsymbol{r}, \boldsymbol{v}_i', t)$, $f_i = f(\boldsymbol{r}, \boldsymbol{v}_i, t)$, $i = 1, 2$. In the integrable soliton case, there exist several interesting integro-differential equations, for example, the Benjamin–Ono equation,

$$u_t + 2uu_x + Hu_{xx} = 0, \qquad (6)$$

where Hu is the Hilbert transform.

MUTHUSAMY LAKSHMANAN

See also **Maps; Ordinary differential equations, nonlinear; Partial differential equations, nonlinear**

Further Reading

Ablowitz, M.J & Clarkson, P.A. 1991. *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, Cambridge and New York: Cambridge University Press

Davis, H.T. 1962. *Introduction to Nonlinear Differential and Integral Equations*, New York: Dover

Ince, E.L. 1956. *Ordinary Differential Equations*, New York: Dover

Kosmann-Schwarzbach, Y., Grammaticos, B. & Tamizhmani, K.M., (editors). 1997. *Integrability of Nonlinear Systems*, Berlin and New York: Springer

Lakshmanan, M. & Rajasekar, S. 2003. *Nonlinear Dynamics: Integrability, Chaos and Patterns*, Berlin: Springer

Murray, J.D. 2002. *Mathematical Biology*, Berlin and New York: Springer

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford: Oxford University Press

EQUILIBRIUM

The word *equilibrium* suggests a balance, typically between antagonistic forces, which in turn often implies absence of motion. Consider a rigid rod, which can freely (i.e., with almost no friction) rotate in a vertical plane about a horizontal axis located near one of its ends. Because of gravity, the rod naturally assumes a position where it is suspended from the horizontal axis, and points downward (see Figure 1a). In this situation, the gravity force acting on the rod is balanced by the reaction force at the suspension point, and no motion occurs. The rod is in equilibrium. If displaced a little bit from this position, the rod oscillates about the horizontal axis and, because of friction, eventually returns to its initial position, which is therefore stable. The other equilibrium position of the rod, which is unstable, is that shown in Figure 1b, where the rod points upward. A slight perturbation will take the rod away from this position, toward the stable equilibrium. Consider now a marble rolling on an uneven floor. Typically, it will keep rolling until it stops in a depression of the floor or reaches a wall or a corner of the room. It is clear from this example that depending on the configuration of the floor, there may be more than one stable equilibrium position. As in the case of the rod, there also exist unstable equilibria, which correspond to local maxima of the floor surface. In both of these examples, motion takes place as to decrease the potential energy of the system, which increases with elevation.

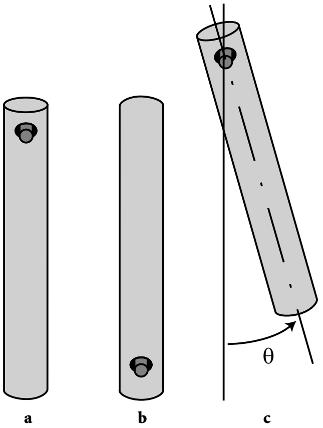

Figure 1. Stable (a) and unstable (b) equilibrium positions of a simple pendulum. (c) The angle θ is measured from the vertical.

From a mathematical point of view, an equilibrium corresponds to a stationary solution of a differential system or a map (see, for instance, Arnol'd, 1992). Quite often, these equations are mathematical models of physical, chemical, or biological systems. Consider the differential equation

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = -\sin(\theta) \qquad (1)$$

describing the motion (in dimensionless form) of a simple pendulum, similar to the rigid rod discussed above but without friction. Here, the variable θ measures the angle between the pendulum and the vertical (see Figure 1c). Time-independent solutions of (1) are obtained by setting the right-hand side of this equation to zero. This gives an infinite number of critical points $\theta = 2p\pi$ and $(2p+1)\pi$, where p is an integer. The first family of solutions can be identified with the stable equilibrium of the pendulum discussed above; the second family corresponds to its unstable equilibria.

Nondissipative mechanical systems conserve their total energy. Typically, the latter can be written in dimensionless form as

$$E(u) \equiv \tfrac{1}{2}(\dot{u})^2 + V(u) = \text{constant}, \qquad (2)$$

where u is a time-dependent variable which describes the position of the system and $\dot{u}$ is its derivative. Taking the derivative of (2) with respect to time and dividing by $\dot{u}$ gives

$$\frac{\mathrm{d}^2 u}{\mathrm{d}t^2} = -\frac{\mathrm{d}V}{\mathrm{d}u}. \qquad (3)$$

In the case of the pendulum, $u = \theta$ and E is obtained by multiplying (1) by $\dot{\theta} = \mathrm{d}\theta/\mathrm{d}t$ and integrating over time. It reads

$$E = \frac{1}{2}\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 - \cos(\theta). \qquad (4)$$

The first term on the right-hand side of E corresponds to the (dimensionless) kinetic energy of the pendulum and the second term to its potential energy V. The graph of V as a function of θ is plotted in Figure 2a. We see that solutions $\theta = 2p\pi$ are minima of V, and solutions of the form $\theta = (2p+1)\pi$ are maxima of V. This can be understood in general as follows. Since the right-hand side of (3) vanishes at an equilibrium point, the latter corresponds to an extremum of the potential energy $V(u)$. Because $-\mathrm{d}V/\mathrm{d}u$ can be interpreted as a force, a minimum of V corresponds to a stable equilibrium, for which the force is restoring. Similarly, maxima of V identify unstable equilibria, as illustrated in Figure 2b. In dynamical systems terminology (see, for instance, Guckenheimer & Holmes, 1990), solutions $\theta = 2p\pi$ of (1) are centers and are *neutrally* stable; solutions $\theta = (2p+1)\pi$ are saddles and are unstable. If we now add a friction term to Equation (1), we get

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = -\sin(\theta) - c\frac{\mathrm{d}\theta}{\mathrm{d}t}. \qquad (5)$$

This equation has the same equilibria as (1) but now the total energy E decreases as a function of time, $\mathrm{d}E/\mathrm{d}t = -c\,(\mathrm{d}\theta/\mathrm{d}t)^2$, and as a consequence, the equilibrium corresponding to solutions of the form $\theta = 2p\pi$, where p is an integer, is *globally asymptotically* stable.

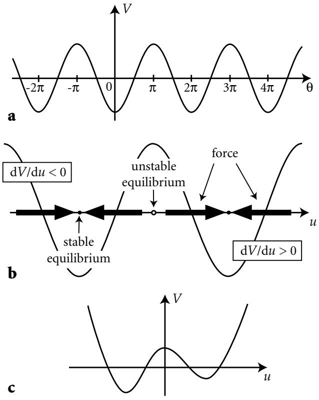

Figure 2. (a) Potential energy $V(\theta) = -\cos(\theta)$ of the simple pendulum. (b) Minima of the potential energy correspond to stable equilibria, whereas maxima describe unstable equilibria. (c) Double-well potential.

In the situation depicted in Figure 2c, V is a double-well potential and the system is *bistable*. Generically, the two minima correspond to different values of the potential energy; the absolute minimum is then the most stable equilibrium of the system; the other minimum corresponds to a metastable equilibrium position.

In the case of maps, equilibria (i.e., stationary solutions) are often called fixed points. For instance, the map

$$u_{n+1} = 2u_n - |u_n|^2 u_n \qquad (6)$$

with $u_n \in \mathbb{R}$ has three fixed points, given by $u_n = u_\mathrm{e}$ for all values of n, where $u_\mathrm{e} = 0$ or $u_\mathrm{e} = \pm 1$. They are obtained by substituting the unknown equilibrium value u_e for u_n in Equation (6) and solving for u_e. If we now consider the same map but let u_n be complex, we see that we have, together with the solution $u_\mathrm{e} = 0$, a whole family of equilibria, given by $u_\mathrm{e} = \exp(\mathrm{i}\varphi)$, where φ is an arbitrary real number. This happens because Equation (6) with u_n complex has the following gauge symmetry: if u_n is a sequence of iterates of the map, then so is $u_n \exp(\mathrm{i}\varphi)$, but with, of course, a different initial condition, $u_0 \exp(\mathrm{i}\varphi)$. More generally, by applying a symmetry of the equation to one of its solutions, one typically obtains another solution. If one starts this iterative process with an equilibrium, a collection of other equilibria is generated. Moreover, if the symmetry is continuous, as it is the case for the gauge invariance of Equation (6) with u_n complex, a continuous family of relative equilibria is created. Similarly, if the symmetry is discrete, a discrete family of equilibria is obtained. For instance, Equation (6) with u_n real has the symmetry $u_n \to -u_n$; as a consequence if $u_e = 1$ is an equilibrium, so is $u_e = -1$.

Similar ideas apply to differential equations. Equations (1) and (5) have the discrete symmetry $\theta \to \theta + 2\pi$. The Landau equation

$$\frac{du}{dt} = u - |u|^2 u, \quad u \in \mathbb{C}, \tag{7}$$

which is the continuous version of Equation (6) with $u_n \in \mathbb{C}$, has the continuous symmetry $u \to u \exp(i\varphi)$, and its equilibria are given by $u = 0$ and $u = \exp(i\varphi)$, where φ is an arbitrary real number.

JOCELINE LEGA

See also **Pendulum; Stability; Symmetry: equations vs. solutions**

Further Reading

Arnol'd, V.I. 1992. *Ordinary Differential Equations*, 3rd edition, Berlin and New York: Springer

Guckenheimer, J. & Holmes, P. 1990. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, New York and London: Springer

EQUIPARTITION OF ENERGY

In 1845, the Scottish scientist John James Waterston submitted a large manuscript outlining a kinetic theory of gases to the Royal Society of London for publication in *Philosophical Transactions*. The central idea of this theory was that heat should be viewed as nothing but the energy associated with the motions of the enormous numbers of molecules, which themselves constitute the gas. Among his results, Waterston noted that "in mixed media the mean square molecular velocity is inversely proportional to the specific weight of the molecules," where the constant of proportionality is to be identified as a measure of thermodynamic temperature. This result is the essence of the Principle of Equipartition of Energy. The manuscript was rejected for publication at the time (as "nonsense, unfit even for reading before the Society") and then published (Waterston, 1893) almost 50 years later when it was (re)discovered by Lord Rayleigh who prefaced the paper with the suggestion that Waterston should be ranked "among the very foremost theorists of all ages."

In the years while Waterston's manuscript lay dormant, the kinetic theory of gases became established, mainly through the collected works of Rudolf Clausius, James Clerk Maxwell, and Ludwig Boltzmann (Brush, 1965, 1966) utilizing statistical methods to describe equilibrium properties. In 1868 Boltzmann obtained one of the most important results of this period, the (Maxwell–Boltzmann) distribution of energy among the molecules for a gas in thermal equilibrium. Let ε denote the energy of a molecule with r degrees of freedom, and position and momentum coordinates in the range $d\omega = \delta q_1 \ldots \delta q_r \delta p_1 \ldots \delta p_r$; the number of such molecules is proportional to the product of the extension $d\omega$ and the Boltzmann factor

$$\exp(-\varepsilon/kT).$$

In the Boltzmann factor, $k = 1.38 \times 10^{-16}$ erg/deg is a constant and T is the absolute temperature. The statistical justification for the Maxwell–Boltzmann distribution is that it is the most probable distribution for the distribution of energy among the possible states for a large number of molecules making up a system at constant energy.

Boltzmann derived the Principle of Equipartition of Energy from the Maxwell–Boltzmann distribution in 1871. In Boltzmann's formulation the mean energy associated with each variable that contributes a quadratic term to the total energy of the molecule has the same value $kT/2$. Thus, the mean energy of a molecule in a system is independent of mass—lighter molecules travel faster. From the mean energy per molecule, it is a simple matter to multiply by the total number of molecules to find the total energy and then to differentiate the total energy with respect to the temperature to find the specific heat at constant volume.

For example, for a perfect monatomic gas of N point atoms where $E = \sum_{i=1}^{N} m v_i^2/2$, there are $3N$ quadratic terms. Thus the internal energy of the gas is $U = 3NkT/2$, and the specific heat at constant volume is $C_V = 3Nk/2$. A simple thermodynamic calculation for the ratio of the specific heat at constant pressure to the specific heat at constant volume yields $\gamma = \frac{5}{3}$ in good agreement with experimental values for the inert gases. In the case of a crystal of N point atoms vibrating about their equilibrium positions according to Hooke's law (harmonic oscillators), there are three quadratic terms for the kinetic energy and three for the potential energy for each particle. Thus, the total internal energy is $U = 3NkT$ and the specific heat is $C_V = 3Nk$, in agreement with the empirical result obtained by Pierre-Louis Dulong and Alexis-Thérèse Petit in 1819.

In 1918 Richard Tolman derived the general Equipartition Principle (Tolman, 1918),

$$\left\langle q_i \frac{\partial \varepsilon}{\partial q_i} \right\rangle = \left\langle p_j \frac{\partial \varepsilon}{\partial p_j} \right\rangle = kT$$

which agrees with Boltzmann's result in the case where ε is a quadratic function of the q_i and p_j but can be applied more generally when ε is not a quadratic function.

Toward the end of the 19th century, some of the predictions from the classical Equipartition Principle were found to be at odds with experimental results.

Examples include the energy distribution from black body radiation, the specific heats of low temperature solids, and the specific heats of diatomic gases. The eventual resolution of these problems played a pivotal role in the upheaval of classical mechanics and the subsequent revolution of quantum mechanics but paradoxically strengthened the atomistic and probabilistic basis on which this principle was founded. The essential new idea of the quantum mechanics is that atoms have sets of discrete energy levels and radiation is absorbed and emitted in discrete quanta. Boltzmann's method of obtaining the Maxwell–Boltzmann distribution as the most probable distribution according to the way that small units of energy could be partitioned was employed by Max Planck in his famous derivation of the radiation law (Planck, 1972) in 1900. The Boltzmann factor is immediately adaptable to the quantum situation. In thermal equilibrium the ratio of the number of particles n_i in one given energy level E_i to the number of particles n_j in another energy level E_j is given by

$$\frac{n_i}{n_j} = \exp\left(-\frac{E_i - E_j}{kT}\right).$$

The word *particle* in this context embraces molecules, electrons, sound waves, light waves, and other dynamical quantities with well-defined energies. As an example, the Boltzmann ratio leads to the average energy of a harmonic oscillator whose quantum energy is an integer multiple of $h\nu$ as

$$\bar{\varepsilon} = \frac{1}{2}h\nu + \frac{h\nu}{e^{h\nu/kT} - 1}.$$

In the low-temperature regime this provides reasonable agreement with the experimental result, but it only agrees with the classical Equipartition Principle $\bar{\varepsilon} = kT$ at high temperatures, $T \gg h\nu/k$.

The Maxwell–Boltzmann distribution is a stationary or steady-state distribution. It can be derived (Tolman, 1948) from the steady-state micro-canonical ensemble distribution for which the density of ensemble copies in phase space is uniform in a narrow energy range $E, E + \delta E$ but zero elsewhere. The dynamical justification for the Maxwell–Boltzmann distribution then rests upon the (quasi-)ergodic hypothesis that the time spent by the phase space trajectory for the dynamical system in a given region of the energy surface will be proportional to the micro-canonical density of ensemble copies in that region. Thus, from a dynamical perspective, ergodicity is a necessary condition for Equipartition of Energy. It follows that in isolated integrable systems Equipartition of Energy has no dynamical justification.

The Equipartition principle emerged again as a catalyst for change in the mid-20th century when its failure observed in computer experiments (Fermi et al., 1955) triggered studies into solitons on the one hand and deterministic Hamiltonian chaos on the other (Ford, 1992). The systems investigated by Fermi, Pasta, and Ulam (FPU) could be considered as collections of harmonic oscillators (harmonic modes of vibration) weakly coupled by nonlinear interactions. The expectation of FPU was that energy initially supplied to one of the harmonic oscillators would become uniformly distributed among all harmonic oscillators. Following the FPU experiments, there has been an enormous literature, attempting to recover the equipartition result, particularly for larger nonlinearities and larger numbers of oscillators. Note however that Boltzmann's principle of equipartition of energy predicts uniform energy sharing among the harmonic modes of an isolated linear chain (a result that cannot be justified dynamically), and it does not make predictions for an isolated nonlinear chain. For energy sharing in an isolated nonlinear chain Tolman's general principle of equipartition can be employed, but the result of uniform energy sharing among the harmonic modes should not generally be expected (Henry & Szeredi, 1995). Moreover, nonlinear chains can sustain intrinsic local nonlinear modes that may persist for very long times, and integrable nonlinear chains support nonlinear normal modes that are more fundamental than the harmonic modes.

The principle of equipartition of energy is one of the most fundamental results in the history of the atomistic description of matter. Inconsistencies between this principle and experimental results in different settings have provided a catalyst for some of the most profound theoretical developments in science in the 20th century. A contemporary problem is the failure of the Equipartition Principle in granular materials (Feitosa & Menon, 2002), the resolution of which may play a fundamental role in the development of a comprehensive kinetic theory of granular systems.

BRUCE HENRY

See also **Ergodic theory; Fermi–Pasta–Ulam oscillator chain; Local modes in molecules; Quantum theory**

Further Reading

Boltzmann, L. 1896. *Lectures on Gas Theory*, New York: Dover (English Translation by S.G. Brush. Originally published in German in two parts by J.A. Barth, Leipzig, 1896 (Part I) and 1898 (Part II), under the title *Vorlesungen über Gastheorie*)

Brush, S.G. 1965, 1966. *Kinetic Theory*, vols. 1(2), Oxford: Pergamon Press

Ehrenfest, P. & Ehrenfest, T. 1959. *The Conceptual Foundations of the Statistical Approach in Mechanics*, Ithaca: Cornell University Press (English Translation of No. 6, vol. IV 2II, *Encyklopädie der mathematischen Wissenschaften*, 1912)

Feitosa, K. & Menon, N. 2002. Breakdown of energy equipartition in a 2D binary vibrated granular gas. *Physical Review Letters*, 88(19): 198301

Fermi, E., Pasta, J.R. & Ulam, S.M. 1955. *Studies of Nonlinear Problems*, Los Alamos Scientific Laboratory Report N. LA-1940 (Electronic Access: http://www.osti.gov/accomplishments/pdf/A80037041/A80037041.pdf)

Ford, J. 1992. The Fermi–Pasta–Ulam problem: paradox turns discovery. *Physics Reports*, 213(5): 271–310

Henry, B.I. & Szeredi, T. 1995. New equipartition results for normal mode energies of anharmonic chains. *Journal of Statistical Physics*, 78: 1039–1053

Planck, M. 1972. On the theory of the energy distribution law of the normal spectrum. In *Planck's Original Papers in Quantum Physics*, edited by H. Kangro with D. ter Haar, & S. Brush (transl.). London: Taylor & Francis (English translation of Planck, M. 1900. Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum. *Verh deutsch phys ges*, 2: 202–237)

Tolman, R.C. 1918. A general theory of energy partition with applications to quantum theory. *Physical Review*, 11: 261–275

Tolman, R.C. 1948. *The Principles of Statistical Mechanics*, Oxford: Oxford University Press, Chapter IV.

Waterston, J.J. 1893. On the physics of media that are composed of free and perfectly elastic molecules in a state of motion. *Philosophical Transactions of the Royal Society London*, 183A: 5–79

ERGODIC THEORY

Ergodic theory is the statistical study of groups of motions of a space, either physical or mathematical, with a measurable structure on it. The origins of ergodic theory can be traced back to the mid-19th century when containers of gas particles were first viewed as sets of randomly moving objects rather than as a collection of individual particles moving under known forces. The word *ergodic* was introduced by Ludwig Boltzmann in the context of the statistical mechanics of gas particles, and it comes from two Greek words *ergon* (work) and *odos* (path).

The mathematical setting in which ergodic theory is studied is as follows. Starting with a space X that represents all possible states of some system which changes over time under known forces, a point $x \in X$ corresponds to one state in the space. The measurable structure consists of a collection of measurable sets $\mathcal{B}$ on X along with a probability measure μ. The measure μ is a function that associates to each set B in $\mathcal{B}$ a number between 0 and 1; this number is the measure of B and we write its measure as $\mu(B)$. A probability measure has the property that $\mu(X) = 1$. Instead of tracking the path of each object in the system, one studies the statistical properties of the motion. The subject evolved from statistical mechanics applied to the study of systems of gas particles moving according to classical laws of physics. In principle, the path of each gas particle can be tracked and its entire history and future can be known; in practice, the complete determination of the paths of the gas molecules is not feasible.

After t units of time, $F_t(x)$ is the point in X corresponding to where x ends up. If $t \in \mathbb{R}$, then the *orbit* of x is the set $\{F_t(x) | t \in \mathbb{R}\}$; F_t is called a flow and defines an action of the group $\mathbb{R}$ on X. Frequently, one uses discrete time intervals and writes $T^n(x) \equiv F_n(x)$ for each integer n, so the orbit of x is a discrete set $\{T^n(x) | n \in \mathbb{Z}\}$ and the group acting on X is $\mathbb{Z}$. In the discrete setting the transformation T is the generating map T^1, and T^n is T composed with itself n times. In classical ergodic theory, the measure μ is preserved under the action; i.e., for any set $A \in \mathcal{B}$, $\mu(A) = \mu(F_t A)$ for all $t \in \mathbb{R}$ or $\mu(A) = \mu(T^n A)$ for all $n \in \mathbb{Z}$. One of the main advantages of the ergodic theoretic point of view is that one can ignore some orbits if they only form a set of measure 0 in the space X; therefore, one uses the terminology *μ-almost everywhere* (or *for μ-a.e. x*) to refer to a property that holds on a set of points of measure 1 in X but perhaps fails to hold on some set of measure 0.

Boltzmann's original ergodic hypothesis has come to be known as the statement that *time average equals space average*; he conjectured (in a certain classical setting) that for any integrable function f, for μ-a.e. x,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \int_X f \mathrm{d}\mu. \tag{8}$$

In this expression, the time average of f for n time steps along the orbit of x is represented by averaging $f(x)$ with $f(Tx)$, $f(T^2x)$, …, and $f(T^{n-1}x)$, since each application of T represents the passage of one unit of time (the left-hand side of the expression). The space average of f is obtained by integrating f over the entire space X, giving the right-hand side of the expression. For a flow the ergodic hypothesis is

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f[F_s(x)] \mathrm{d}s = \int_X f \mathrm{d}\mu. \tag{9}$$

The conjecture is false as stated; it only holds when the action is μ-ergodic, and this realization led to the definition of an ergodic action.

The setting of ergodic theory has been greatly enlarged to include the actions of many other groups (Zimmer, 1984). A nonsingular action Φ of a group G on a space $(X, \mathcal{B}, \mu)$ consists of an action of G on X such that the map $\Phi: G \times X \to X$ is measurable and for each $g \in G$ the map $\phi_g(x) = \Phi(g, x)$ is a nonsingular automorphism of X; that is, ϕ_g is an invertible measurable transformation and for any $B \in \mathcal{B}$, $\mu(B) = 0$ if and only if $\mu(\phi_g^{-1} B) = \mu(\phi_{g^{-1}} B) = 0$. The group operation on G is reflected in the action since $\phi_{g_1 g_2}(x) = \phi_{g_2}(\phi_{g_1} x)$ for all $g_1, g_2 \in G$ and almost every (μ-a.e.) x.

An action is *ergodic* if whenever $\phi_g(A) = A$ for all $g \in G$, then $\mu(A) = 0$ or 1. The study has been extended beyond group actions to the study of ergodic equivalence relations, and the assumption that the measure μ be a probability measure is frequently dropped. One also considers the actions of semigroups when the action being studied is not invertible.

The Birkhoff ergodic theorem for a measure-preserving transformation T states that the limit of the left-hand side in Equation (8) exists for μ-a.e. x, is an integrable function f^*, and $f^*(Tx) = f^*(x)$ for μ-a.e. x. Furthermore, f^* is the constant $\int_X f\,\mathrm{d}\mu$, the space average of f, precisely when the action is ergodic (Birkhoff, 1931). Any point x satisfying the theorem is called a generic point and generates a *typical orbit*. There are many ergodic theorems; perhaps the simplest is the von Neumann ergodic theorem (von Neumann, 1932). By definition, a function $f \in L^2$ if $\int |f|^2 \mathrm{d}\mu < \infty$; moreover defining $f \circ T^k(x) = f(T^k x)$, for measure-preserving T, $f \circ T^k \in L^2$ as well. Von Neumann's theorem is also called the Mean or L^2 ergodic theorem because while no individual orbit is tracked (no x appears in the statement), the average difference (L^2 integral) between the time average and the limit function f^* must be small.

Theorem 1. (von Neumann or L^2 Ergodic Theorem) *If T is a measure-preserving transformation and $f \in L^2$, then there is a function $f^* \in L^2$ such that*

$$\int_X \left| \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k - f^* \right|^2 \mathrm{d}\mu \to 0 \quad as \quad n \to \infty.$$

If $(X, \mathcal{B}, \mu)$ is a probability space and the transformation T (or group action) on X is non-ergodic, there is a disintegration or decomposition of μ into measures μ_y, indexed by points y in another probability space Y, with its own measure ρ, such that $\mu(A) = \int_Y \mu_y(A)\,\mathrm{d}\rho$ for every $A \in \mathcal{B}$. Furthermore, the limit function f^* in the ergodic theorems is constant with respect to each measure μ_y; that is, T is μ_y ergodic for ρ–a.e. y. The decomposition is independent of the function f; this is referred to as the *ergodic decomposition* of the measure μ with respect to T (Rohlin, 1949). One of the central problems in ergodic theory is to determine when two measure-preserving group actions are conjugate via a measure-preserving isomorphism. To this end, invariants such as entropy and spectral properties have been the subject of much study.

There is a hierarchy of statistical properties associated to ergodic actions, including weak mixing, mixing, K-automorphisms, and Bernoulli automorphisms. As a transformation moves up the hierarchy (in the order listed), the more chaotic the behavior of the system is expected to be.

Applications to Dynamical Systems and Chaos

The simplest example of an ergodic transformation is irrational rotation on the circle with respect to Lebesgue measure. The most random transformation is a Bernoulli shift with an independent identically distributed measure on it; this includes coin tosses.

Overlaps of the topological setting of dynamical systems with ergodic theory exist, and much of ergodic theory highlights the interface between the topological and measurable structure of a group G acting on $(X, \mathcal{B}, \mu)$. In 1935, Hedlund proved the ergodicity of the geodesic flow on the unit tangent bundle of a surface of constant negative curvature; in 1940, Hopf extended the result to establish ergodicity of the geodesic flow on arbitrary manifolds with negative sectional curvature. In this setting the invariant measure is the Liouville measure.

Modern ergodic theory was started by Andrei Kolmogorov with the formal development of Boltzmann's notion of entropy, and developed in the 1960's and 1970's to include many differentiable actions. Applications include fluid dynamics, coding theory, number theory, complex dynamics, and cellular automata.

JANE HAWKINS

See also **Chaotic dynamics; Dynamical systems; Entropy; Symbolic dynamics**

Further Reading

Birkhoff, G. 1931. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences USA*, 17: 656–660

Cornfeld, I., Fomin, S. & Sinai, Y. 1982. *Ergodic Theory*, New York: Springer

Furstenberg, H. 1981. *Recurrence in Ergodic Theory and Combinatorial Number Theory*, Princeton: Princeton University Press

Halmos, P. 1956. *Ergodic Theory*, New York: Chelsea Publishing

Katok, A. & Hasselblatt, B. 1995. *Introduction to the Modern Theory of Dynamical Systems*, Cambridge and New York: Cambridge University Press

Petersen, K. 1983. *Ergodic Theory*, Cambridge and New York: Cambridge University Press

Rohlin, V. A. 1949. On the fundamental ideas of measure theory. *American Mathematical Society Translations*, 10 (1): 1–54 (original Russian article 1949)

von Neumann, J. 1932. Proof of the quasiergodic hypothesis. *Proceedings of the National Academy of Sciences USA*, 18: 70–82

Walters, P. 1982. *An Introduction to Ergodic Theory*, New York: Springer

Zimmer, R. 1984. *Ergodic Theory and Semisimple Groups*, Boston: Birkhäuser

ESAKI TUNNEL DIODE

See **Diodes**

EULER'S METHOD

See **Numerical methods**

EULERIAN DESCRIPTION

See **Fluid dynamics**

EULER–LAGRANGE EQUATIONS

In the dynamics of energy conserving systems there are two equivalent formulations: one based on differential equations and the other based on the principle of least action (Goldstein, 1951). At first glance, these two approaches seem to have little in common as one stems from iterating infinitesimal changes along a solution trajectory, whereas the other depends on evaluating all possible paths between two end points and selecting the path with the minimum action.

To see how this goes, consider a simple classical system: the nonrelativistic point particle of unit mass. In three dimensions this particle has three degrees of freedom each labeled by coordinates $q_i(t)$ for $i = 1, 2, 3$. The motion of the particle is determined by a Lagrangian functional $L(q_i, \dot{q}_i)$ which depends on both the position components, $q_i(t)$, and the velocity components, $\dot{q}_i(t)$, of the particle. This functional is

$$L = \sum_i \left[\dot{q}_i^2/2 - V(q_i)\right], \tag{1}$$

where $V(q_i)$ is some potential in which the particle moves; thus, the Lagrangian is defined as the difference between the kinetic and the potential energy of the system. The motion of the particle is determining by minimizing the action, which is the time integral of the Lagrangian

$$S = \int_{t_1}^{t_2} L(q_i, \dot{q}_i)\, \mathrm{d}t, \tag{2}$$

and the path taken by the particle is the one for which the action is a minimum. In other words, $\delta S = 0$, where δ implies a differential change in S under a small change in the path.

In calculating the equations of motion, a small variation in the path of the particle $\delta q_i(t)$ is assumed with the endpoints fixed, that is, $\delta q_i(t_1) = \delta q_i(t_2) = 0$. To calculate δS, the Lagrangian is varied with respect to both changes in the position and the velocity

$$\delta S = \int_{t_1}^{t_2} \sum_i \left(\frac{\partial L}{\delta q_i}\delta q_i + \frac{\partial L}{\partial \dot{q}_i}\delta \dot{q}_i \right) \mathrm{d}t. \tag{3}$$

Integration by parts of the last term gives

$$\delta S = \int_{t_1}^{t_2} \sum_i \left[\frac{\partial L}{\partial q_i}\delta q_i - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\right)\delta q_i \right] \mathrm{d}t, \tag{4}$$

so the condition $\delta S = 0$ implies that

$$\frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = 0 \tag{5}$$

for all i. These are the Euler–Lagrange (EL) equations of motion. For the system of Equation (1), they are $\mathrm{d}^2 q_i/\mathrm{d}t^2 + \partial V/\partial q_i = 0$, which is Newton's second law.

In some cases the motion is constrained, for example to curves or surfaces. Such a constraint can be expressed as a function $f(q_i) = 0$. Augmenting the Lagrangian to $\tilde{L} = L + \lambda f$ (where λ is called a Lagrange multiplier) and minimizing S with respect to $\tilde{L}$ then yields equations for motion under the constraint (Gel'fand & Fomin, 2000). As a simple example, consider the system of Equation (1) with $V = (q_1^2 + q_2^2 + q_3^2)/2$ moving under the constraint $q_1 = q_2$ or $q_1 - q_2 = 0$. Then the EL equations for the augmented Lagrangian $\tilde{L} = L + \lambda(q_1 - q_2)$ are $\mathrm{d}^2(q_1 + q_2)/\mathrm{d}t^2 + (q_1 + q_2) = 0$ and $\mathrm{d}^2 q_3/\mathrm{d}t^2 + q_3 = 0$.

An alternative to the Lagrangian formulation (where positions and velocities are the fundamental variables) is the Hamiltonian formulation. Here the positions and momenta (p_i) are independent variables, where

$$p_i \equiv \partial L/\partial \dot{q}_i \tag{6}$$

and a Hamiltonian is defined through the Legendre transformation

$$H(q_i, p_i) \equiv \sum_i p_i \dot{q}_i - L(q_i, \dot{q}_i). \tag{7}$$

Because this Hamiltonian is conserved under the dynamics, it is often interpreted as the energy of a dynamical system. For the example of Equation (1),

$$H = \sum_i \left[\dot{q}_i^2/2 + V(q_i)\right], \tag{8}$$

the sum of kinetic and potential energies.

The generalization from a point particle to a three-dimensional field may be viewed as the replacement of $q_i(t)$ by a field amplitude: $\phi(x, y, z, , t)$ (Goldstein, 1951). In this case, EL equations can be derived from a variational principle applied to an action integral of the form

$$S = \int \mathcal{L}\left(\phi, \phi_t, \phi_x, \phi_y, \phi_z\right)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\, \mathrm{d}t, \tag{9}$$

where $\mathcal{L}$ is a Lagrangian density and the subscripts indicate partial derivatives. The condition for minimum action then gives

$$\frac{\partial \mathcal{L}}{\partial \phi} - \frac{\partial}{\partial t}\frac{\partial \mathcal{L}}{\partial \phi_t} - \frac{\partial}{\partial x}\frac{\partial \mathcal{L}}{\partial \phi_x} - \frac{\partial}{\partial y}\frac{\partial \mathcal{L}}{\partial \phi_y} - \frac{\partial}{\partial z}\frac{\partial \mathcal{L}}{\partial \phi_z} = 0, \tag{10}$$

which is the EL equation in field theory.

A simple example of this field formulation is provided by the sine-Gordon equation in three-dimensional space. The Lagrangian density is given by

$$\mathcal{L} = \tfrac{1}{2}\left(\phi_t^2 - \phi_x^2 - \phi_y^2 - \phi_z^2\right) + (\cos\phi - 1), \tag{11}$$

for which Equation (10) implies

$$\phi_{xx} + \phi_{yy} + \phi_{zz} - \phi_{tt} = \sin\phi. \tag{12}$$

Because not all dynamics conserve energy, the Lagrangian formulation is not applicable to all physically motivated systems. Reaction-diffusion equations, for example, are excluded from this class (Scott, 2003).

Alwyn Scott

See also **Extremum principles; Hamiltonian systems; Newton's laws of motion; Poisson brackets**

Further Reading

Gel'fand, I.M. & Fomin S.V. 2000. *Calculus of Variations*, New York: Dover

Goldstein, H. 1951. *Classical Mechanics*, Reading, MA: Addison–Wesley

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

EVANS FUNCTION

Traveling waves are one-dimensional patterns that arise in numerous applications, for instance, as defects, flame fronts, nerve impulses, shock waves, and solitons. Typically, traveling waves are only observed in experiments or in numerical simulations if they are stable.

In the early 1970s, Evans developed a number of tools designed to determine the stability of pulses, that is of traveling waves that are localized in space as solutions to a system of partial differential equations (PDEs) (Evans, 1975). The class of PDEs considered by Evans serve as models for the propagation of impulses in nerve axons that include the FitzHugh–Nagumo equation and the Hodgkin–Huxley equations. The strategy to determine the stability of waves consists of two steps: (i) linearize the nonlinear PDE about the wave and calculate the spectrum of the resulting linear operator, and (ii) show that the wave is stable for the nonlinear PDE if the spectrum contains no unstable elements. This entry focuses on the first step, see Sandstede (2002) for references concerning (ii).

The spectrum of the linearization $\mathcal{L}$ about a pulse or a front is the union of two sets, namely the point spectrum (which consists of all isolated eigenvalues) and the essential spectrum (which is also often referred to as the continuous or radiation spectrum). The essential spectrum can be computed easily as it involves only the homogeneous background states of the pulse or front.

The key problem is therefore to find all isolated eigenvalues in the complement Ω of the essential spectrum in the complex plane. To address this issue for his class of nerve-axon models, Evans constructed an analytic function $D(\lambda)$ (now referred to as the Evans function) that maps the region Ω into the complex plane with the property that $D(\lambda)$ vanishes precisely at isolated eigenvalues of the linear operator $\mathcal{L}$. In fact, the order of a root λ_* of the Evans function $D(\lambda)$ coincides with the algebraic multiplicity of λ_* as an eigenvalue of $\mathcal{L}$. Here, ℓ is the order of a root λ_* of $D(\lambda)$ if the first $\ell - 1$ derivatives of $D(\lambda)$ vanish at λ_* but the ℓth derivative does not.

Note that, if the underlying PDE is invariant under space translations, then $\lambda = 0$ is an eigenvalue. The corresponding eigenfunction is given by the spatial derivative of the wave. Hence, if $\lambda = 0$ is not in the essential spectrum, the Evans function $D(\lambda)$ vanishes at $\lambda = 0$. Proving linear stability of a pulse, therefore, amounts to calculating the Evans function $D(\lambda)$ and showing that it does not have any zero in the closed right half-plane except possibly a simple zero at $\lambda = 0$.

The derivative $D'(0)$ is actually given by the Mel'nikov integral, computed with respect to the wave speed, c, of the pulse. This integral measures how the stable and unstable manifolds of the background state intersect as the wave speed c of the pulse is varied. In other words, if the existence problem of the wave is understood well enough, then $D'(0)$ is computable.

This can be exploited as follows. The Evans function $D(\lambda)$ is, in fact, real whenever λ is real, and $D(\lambda)$ can be normalized so that $D(\lambda) > 0$ for all sufficiently large $\lambda \gg 1$. Here, λ is the temporal growth rate so that unstable eigenvalues have positive real part. Thus, if $D'(0) < 0$, then the pulse is necessarily unstable with an odd (and therefore nonzero) number of unstable real eigenvalues. Consequently, the derivative $D'(0)$ of the Evans function provides a parity instability index.

The function $D(\lambda)$ introduced by Evans was then later used by Jones and by Yanagida to prove that the fast pulse arising in the FitzHugh–Nagumo equation is stable. Both proofs used the parity index introduced above: after establishing that there is precisely one eigenvalue besides $\lambda = 0$ that lies close to the imaginary axis or in the right half-plane, a computation of the parity index at $\lambda = 0$ showed that it is positive, so that the other eigenvalue must lie in the left half-plane.

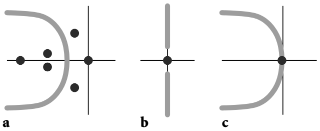

Figure 1. The three illustrations show typical spectra of the linearization about (a) a stable pulse in a dissipative PDE, (b) a soliton in an integrable system, and (c) a viscous shock wave. The horizontal and vertical axes are (Re λ, Im λ). The point and essential spectrum are shown as discs and lines, respectively.

Alexander et al. (1990) developed the Evans function further and also related it to certain topological invariants. Since then, Evans-function calculations have been utilized in various applications, for instance, in nonlinear optics, plasma physics, and thin films (Sandstede, 2002; Zumbrun, 2001).

An important class of PDEs to which the Evans function has been applied is singularly perturbed equations that exhibit two different spatial scales. In this situation, the Evans function can often be calculated from the Evans functions for the slow and the fast subsystem (Doelman et al., 2002).

The Evans function is also a useful tool for investigating the stability of multi-hump pulses, or n-pulses, which are composed of n widely separated copies of a primary pulse. If the primary pulse is stable, so that $\lambda = 0$ is a simple eigenvalue, then such an n-pulse has exactly n critical eigenvalues near the origin which encode the interaction properties of the n individual pulses. These eigenvalues can be calculated using the Evans function (Sandstede, 2002).

The original construction of the Evans function works only outside of the essential spectrum. For Hamiltonian PDEs, the essential spectrum of stable pulses lies on the imaginary axis, see Figure 1(b). Thus, it is possible that eigenvalues move out of the essential spectrum and into the right half-plane once a small perturbation is added to the underlying PDE. The points in the essential spectrum from which such eigenvalues can potentially emerge and the exact location of the bifurcating eigenvalues for concrete perturbations can be found by extending the Evans function across the essential spectrum. For integrable PDEs such as the Korteweg–de Vries or the nonlinear Schrödinger equation, it is possible to compute the extended Evans function explicitly using inverse scattering theory (see Kapitula & Sandstede (2002) for references).

Similar issues occur for the linearization about viscous shock waves where the essential spectrum always touches the origin, see Figure 1(c). An extension of the Evans function across the essential spectrum to a full neighborhood of $\lambda = 0$ allows one to generalize the parity instability index, given by an appropriate sign condition on the derivative $D'(0)$, to viscous shocks. In addition, the extended Evans function plays an important role in the nonlinear stability analysis of viscous shocks (Zumbrun, 2001).

Björn Sandstede

See also **Shock waves; Stability; Wave stability and instability**

Further Reading

Alexander, J.C., Gardner, R.A. & Jones, C.K.R.T. 1990. A topological invariant arising in the stability analysis of travelling waves. *Journal für die Reine und Angewandte Mathematik*, 410: 167–212

Doelman, A., Gardner, R.A. & Kaper, T.J. 2002. A stability index analysis of 1-D patterns of the Gray–Scott model. *Memoirs of the American Mathematical Society*, 155(737): xii+64 pp.

Evans, J. 1975. Nerve axon equations (iv): The stable and unstable impulse. *Indiana University Mathematics Journal*, 24: 1169–1190

Kapitula, T. & Sandstede, B. 2002. Edge bifurcations for near integrable systems via Evans-function techniques. *SIAM Journal on Mathematical Analysis*, 33: 1117–1143

Sandstede, B. 2002. Stability of travelling waves. In *Handbook of Dynamical Systems,* vol. 2, edited by B. Fiedler, Amsterdam: North-Holland

Zumbrun, K. 2001. Multidimensional stability of planar viscous shock waves. In *Advances in the Theory of Shock Waves*, edited by H. Freistühler, A. Szepessy, T.-P. Liu & G. Métiver, Boston: Birkhäuser

EVAPORATION WAVE

Under special conditions a liquid can be superheated to a few hundred degrees Celsius above its nominal boiling point. The thermal energy associated with this excess temperature can be released by rapid evaporation, also called a *vapor explosion*. The largest explosions on Earth are vapor explosions in the form of volcanic eruptions. For example, the Mount St. Helens blast was estimated at 400 Mtons TNT equivalent (Decker & Decker, 1981), which is nearly the sum of all worldwide nuclear explosions to date (about 500 Mtons).

Vapor explosions also pose a hazard in industry. In one incident, 100 lb of molten steel fell into an open trough containing 78 gallons of water. The resulting explosion killed one person and injured many, cracked the 20-in. thick concrete floor, and shattered 6000 panes of glass (Reid, 1976). If a vapor explosion involves a flammable liquid, then the physical explosion can transition to a more powerful chemical explosion.

An *evaporation wave* is a vapor explosion that proceeds across the system with an identifiable propagation velocity. Evaporation waves are an interesting example of self-organized dynamical complexity and comprise a controlled form of vapor explosion that facilitates the detailed study of liquid fragmentation and dispersal mechanisms.

Liquid superheating is caused by cohesion between molecules. Hence, bubbles—even very small

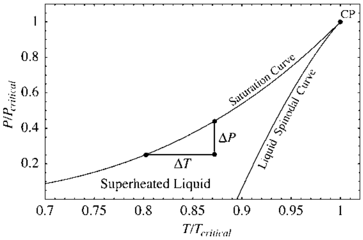

Figure 1. Existence region for superheated liquid from the Redlich–Kwong (R–K) equation of state. R–K predicts universal behavior in a reduced coordinate system.

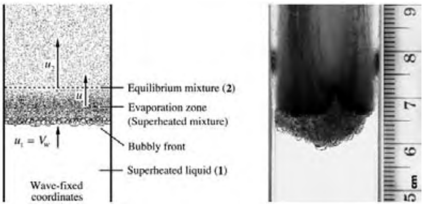

Figure 2. Schematic diagram of an evaporation wave (a), and a photograph of an evaporation wave in freon-12 at 20°C (b). The front is moving down at $0.6\,\mathrm{m\,s^{-1}}$.

ones that reside in most liquids under normal conditions—comprise weak spots that inhibit superheating. Even a completely pure liquid can only be superheated to a point, called the superheat limit, where it becomes mechanically unstable. The theoretical superheat limit occurs at a locus of states called the *spinodal curve* (Debenedetti, 1996). Superheated liquid can exist between the saturation and spinodal curves as shown in Figure 1. The amount of superheat is quantified by the excursion of the liquid state from the saturation curve, for example, the degree ΔT by which the liquid temperature exceeds the equilibrium temperature at its pressure.

The specific enthalpy required to evaporate a liquid is the latent heat of evaporation L. A superheated liquid stores an excess specific enthalpy $C_p \Delta T$, where C_p is the liquid specific heat at constant pressure. The Jakob number, $\mathrm{Ja} \equiv (C_p \Delta T)/L$, is equal to the liquid mass fraction χ that can evaporate adiabatically at constant pressure. If $\mathrm{Ja} > 1$ then complete evaporation can occur. This condition is possible in *retrograde* liquids, which are those possessing a sufficiently large specific heat (Thompson et al., 1986). Nowhere near complete evaporation is required to produce quasisteady evaporation waves. However, sufficient evaporation must occur so that the liquid phase is swept along by the vapor; otherwise, the upstream liquid will cool and the process will quench.

The first mention of an evaporation wave appears to have been by Terner (1962), who performed vertical shock tube experiments on heated water. Terner believed that an evaporation wave followed depressurization, though he did not definitively show it.

The first direct observation appears to have been by Friz (1965), who studied the expulsion of superheated water from a vertical glass tube into the atmosphere. The tube was sealed on top by a diaphragm, and the water was heated to temperatures from 105°C to 125°C. The gas space above the water was pressurized to 3.5 bar, whereupon a cutter was activated to puncture the seal. Friz describes that upon depressurization “a great number of bubbles” formed throughout the water column. An “acceleration front” several centimeters thick then propagated into the bubbly mixture at 1–$2\,\mathrm{m\,s^{-1}}$, transforming it into a spray that was ejected upward. The many bubbles were probably caused by pre-pressurization, which increased the amplitude of the initial expansion wave so that it reflected from the test cell base in tension, causing cavitation.

Grolmes and Fauske (1970) performed similar experiments in heated water, freon-11, and methanol. Their configuration was like Friz’s, except that (i) the initial pressure was the vapor pressure, and (ii) the fluid was expelled into a nearly evacuated reservoir. By paying careful attention to the liquid purity and test cell cleanliness, they were able to suppress *all* nucleation within the liquid column. Upon depressurization boiling erupted on the liquid surface and proceeded with sufficient vigor to expel the generated spray from the container as depicted in Figure 2a. The boiling surface receded at a constant average rate of 0.3–$0.5\,\mathrm{m\,s^{-1}}$, depending on the liquid and conditions.

Thompson et al. (1987) performed similar experiments on the retrograde fluorocarbon perfluoro-*n*-hexane (C_6F_{14}). At the highest (nearly critical) temperatures, the lead expansion wave attempted to lower the pressure past the superheat limit, whereupon homogeneous nucleation erupted and blocked the drop. An approximate plateau was maintained near the spinodal, into which a slower evaporation wave propagated in a manner similar to Friz’s. Such *wave splitting* occurs when material properties abruptly change within the wave, in a way that lowers the sound speed. At lower temperatures, homogeneous nucleation likely erupted upon reflection of the initial expansion wave from the test cell base, and an evaporation wave initiated after the reflected wave reached the free surface. The waves became progressively less distinct as the initial temperature was lowered, presumably because heterogeneous nucleation from the metal side-walls and the transducer ports became increasingly important.

Shepherd and Sturtevant (1982) performed experiments on butane drops that were heated slowly in an ethylene glycol host fluid to the superheat limit

($\Delta T = 105°\text{C}$), whereupon they exploded with a sharp "crack." High-speed photographs showed that evaporation initiated at a single nucleation site at or near the drop surface and spread into pure liquid as a miniature evaporation wave resembling Grolmes and Fauske's. Thus, heating to the superheat limit has produced waves in pure liquid, whereas depressurizing to the superheat limit has produced waves in bubbly liquid.

Shepherd and Sturtevant suggested that the *Landau instability* of premixed flames was the essential fragmentation mechanism. Frost and Sturtevant (Frost, 1988) repeated the exploding droplet experiments in pentane, iso-pentane, and ethyl ether. They obtained several photographs in which the evaporating interface had a mottled texture resembling an orange peel. This suggested that they had caught the interface at the early stages of instability, which they believed grew to the point of pinching off droplets to produce the observed rough-looking interfaces and spray flow. About the same time Mesler (1988) proposed that the *secondary nucleation* mechanism of surface boiling—whereby receding film-caps from ruptured bubbles strike the liquid surface and entrain vapor to create new nucleation sites—was responsible.

Hill and Sturtevant (Hill, 1991) performed Grolmes and Fauske-type experiments in freons 12 and 114, with the goal of observing detailed propagation mechanisms. Figure 2b is a photograph from that study. Several observations suggested that a three-step cycle occurs: (i) bubbles grow at the leading edge in accordance with classical theory; (ii) these are consumed en masse by fragmentation waves that sweep around the bubbly layer transversely to the main propagation at speeds of order the axial flow, and which shatter the superheated interstitial liquid into droplets; (iii) a small fraction of the droplets are propelled forward to strike the leading edge and nucleate new bubbles à la Mesler. Whether such behavior extends toward the superheat limit, or whether a nucleation-free evaporative instability dominates at some point, is unknown.

Simões-Moreira and Shepherd (1999) performed similar experiments in heated dodecane ($C_{12}H_{26}$)—a highly retrograde liquid—with the intention of observing complete evaporation waves. For $\text{Ja} > 1$ a simple plane evaporation wave, or a convoluted but simply connected surface like Frost's "orange peel," would be energetically admissible. Instead they observed a scenario qualitatively like previous experiments, except that the droplets evaporated in the downstream flow.

From a gas-dynamic viewpoint an evaporation wave is analogous to a flame in a combustible mixture. Both are classified as *deflagrations*, or *exothermic discontinuities*. The ideal theory (e.g., Hayes, 1958) considers a plane, steady wave. The conservation equations for mass, momentum, and energy are applied between the upstream state **1**, and an arbitrary state in the reaction or evaporation zone. The flow is assumed to be one-dimensional (no turbulence), and viscosity and heat conduction are neglected. Eliminating the fluid velocities between the three conditions yields the Hugoniot curve $\mathcal{H}$, which specifies the locus of possible downstream states given **1**:

$$h_1 - h(v, P) = \tfrac{1}{2}(P_1 - P)(v_1 + v), \tag{1}$$

where h is the specific enthalpy, P is the pressure, and v is the specific volume. Combining the mass and momentum equations gives the Rayleigh line $\mathcal{R}$:

$$\frac{P - P_1}{v - v_1} = -(\rho_1 V_w)^2 = -\dot{m}^2, \tag{2}$$

a line in P–v space with negative slope equal to the square of the mass flux $\dot{m}$. Since the same value of $\dot{m}$ applies throughout the structure of the steady wave, $\mathcal{R}$ traces the locus of states in the reaction or evaporation zone. Steady wave solutions are given by the intersection of $\mathcal{R}$, with $\mathcal{H}$ evaluated at the end state **2**.

We now specialize the analysis to an evaporation wave with pure (or nearly pure) upstream liquid, with a two-phase downstream mixture described by

$$h = h_l(1 - \chi) + h_v\chi \quad \text{and} \quad v = v_l(1 - \chi) + v_v\chi, \tag{3}$$

where χ is the mass fraction of vapor, and l and v denote the liquid and vapor phases. We neglect slip between the two phases keep to and invoke four additional assumptions that are each usually accurate to a percent or better: (i) the liquid enthalpy is conserved upon depressurization from the initial state **0** to the established upstream state **1** ($h_0 = h_1$); (ii) the liquid density is constant throughout ($\rho_0 = \rho_1 = \rho_l = \rho_{l2}$); (iii) the vapor density is small compared with the liquid density ($\rho_v \ll \rho_l$)—true except near the critical point; (iv) the kinetic energies are small compared with the thermal energy ($\chi_2 = \text{Ja}_2$).

In the evaporation zone the liquid droplets are still somewhat superheated, with bulk enthalpy (averaged over a cross-section) h_l. Motivated by the reaction progress variable used in combustion, we define an *evaporation progress variable*, $0 \le \lambda \le 1$:

$$\lambda \equiv \frac{h_0 - h_l}{h_0 - h_{l\phi}(P)}, \tag{4}$$

where $h_{l\phi}(P)$ is the equilibrium liquid enthalpy. The Jakob number is then written as

$$\text{Ja}(P) \equiv \frac{h_0 - h_{l\phi}(P)}{h_v(P) - h_{l\phi}(P)}. \tag{5}$$

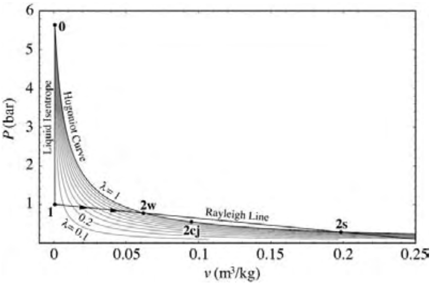

Figure 3. Thermodynamic construction for an evaporation wave—the liquid expansion isentrope, Hugoniot curve, Rayleigh line, and Chapman–Jouguet point.

Combining Equations (3b)–(5) with the energy equation gives an implicit formula for $\mathcal{H}$:

$$v = v_0 + \frac{v_v(P)\lambda \,\mathrm{Ja}(P)}{1 - \mathrm{Ja}(P)(1-\lambda)}. \tag{6}$$

Equation (6) is plotted in Figure 3 for several values of λ. A particular $\mathcal{R}$ is overlaid, which intersects $\mathcal{H}$ in two places. The high-pressure intersection at **2w** is the *weak solution*; the low-pressure intersection at **2s** is the *strong solution*. One can show that the end state **2** is subsonic for the weak branch and supersonic for the strong branch. In progressing along $\mathcal{R}$ from **1** to **2w**, the local state crosses increasing values of λ. This is a physically realizable process. But there is no physical way for evaporation to proceed beyond **2w** to **2s**—which is why strong deflagrations are not observed.

In the special case where $\mathcal{R}$ intersects $\mathcal{H}$ at the tangency point the emerging flow is exactly sonic; this is the Chapman–Jouguet (**cj**) point. It is clear from Equation (2) and Figure 3 that for a given **1**, **2cj** defines the maximum allowable wave speed. In some experiments, the test cell is *choked*, which corresponds to sonic outflow in the laboratory frame. In an evaporation wave, the two-phase flow speed is much faster than the wave speed, such that the difference between the laboratory and wave frames is negligible. Therefore, if an unobstructed test cell of constant cross-sectional area is choked, the wave can be considered **cj**.

Combining $\mathcal{R}$ with $\mathcal{H}$ evaluated at **2** gives an explicit equation for the wave speed:

$$V_w = \sqrt{\frac{v_0 \;\delta P_2}{\mathrm{Jk}(P_2)}}, \quad \text{where}\;\mathrm{Jk}(P) \equiv \mathrm{Ja}(P)\left(\frac{v_v(P)}{v_0}\right)$$

and

$$\delta P \equiv P_1 - P. \tag{7}$$

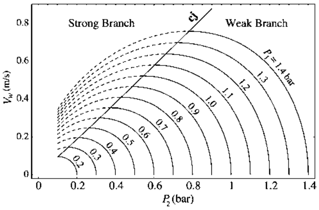

Figure 4. Plot of the evaporation wave speed relation, Equation (7), for freon-12.

δP_2 is the overall wave pressure-amplitude, one measure of its strength. Jk is a useful alternative definition of the Jakob number that arises about as frequently as Ja, and which unfortunately goes by the same name (and usually the same symbol).

The liquid initial conditions v_0 and h_0, and the external reservoir pressure P_r, are set a priori and remain essentially unchanged after depressurization. Assuming that the test cell is long enough for equilibrium to be achieved at or before the exit: if the flow is unchoked (weak solution), then $P_2 = P_\mathrm{r}$; if the flow is choked (**cj** solution), then $P_2 > P_\mathrm{r}$ and a sonic condition applies that specifies P_2. But P_1, and therefore δP_2, is unknown.

Figure 4 plots Equation (7) for various values of P_1. Each curve has a velocity maximum, representing its own **cj** solution, the locus of which is given by

$$V_{w\mathbf{cj}}(P_2) = \sqrt{\frac{v_0}{-\mathrm{d}Jk/\mathrm{d}P_2}}. \tag{8}$$

Predicted and measured wave speeds can deviate by up to 50%. Two issues likely dominate. First, the theory assumes the flow to be strictly one dimensional, but Hill's experiments show that the transient lateral flow velocity in the evaporation zone is of order the mean axial velocity. This results in a higher-than-calculated pressure drop. Second, evaporation is very rapid where fragmentation occurs, but slows downstream where droplets are convecting. Thus, the end state **2** may not exist within the tube.

To close the theory—to calculate the wave speed without measuring P_1—we must add a statement about the physical mechanisms peculiar to evaporation waves. This task is actually simpler for waves in bubbly liquids than for waves in pure ones. For Thompson et al.'s highest-temperature experiments, the upstream state was essentially at the superheat limit, which can be computed from initial conditions. In Friz's experiments, there was time for multiple acoustic reflections in the upstream liquid. He assumed P_1 to be the equilibrium value at the initial temperature, which gave reasonable results.

For waves in pure liquids there is no upstream nucleation to modulate the pressure, and one must address the boiling dynamics in the evaporation zone. The problem is analogous to the calculation of flame speed, which is controlled by the rate of heat conduction (and often convection) from the reaction zone to the reactants. This is where transport properties such as thermal and mass diffusivity enter. No such physically based model exists for evaporation waves, although Reinke and Yadigaroglu (2001) have proposed an empirical correlation based on existing data.

LARRY HILL

See also **Dimensional analysis; Explosions; Flame front; Fluid dynamics**

Further Reading

Debenedetti, P.G. 1996. *Metastable Liquids: Concepts & Principles*. Princeton, NJ: Princeton University Press

Decker, R. & Decker, B. 1981. The eruptions of Mount St Helens. *Scientific American*, 244(3): 68–80

Friz, G. 1965. Coolant ejection studies with analogy experiments. *Proceedings of the Conference on Safety Fuels & Core Design in Large Fast Power Reactors*, US Atomic Energy Commission Report ANL-7120: 890–894

Frost, D.L. 1988. Dynamics of explosive boiling of a droplet. *Physics of Fluids*, 31(9): 2554–2561

Grolmes, M.A. & Fauske, H.K. 1970. Modeling of sodium expulsion with freon-11. *ASME Paper* 70-HT-24

Hayes, W.D. 1958. The basic theory of gasdynamic discontinuities. In *Fundamentals of Gas Dynamics*, edited by H.W. Emmons, Princeton, NJ: Princeton University Press

Hill, L. 1991. An experimental study of evaporation waves in a superheated liquid. Ph.D. Thesis, California Institute of Technology, Pasadena, CA, USA

Mesler, R. 1988. Explosive boiling: a chain reaction involving secondary nucleation. *Proceedings of the ASME National Heat Transfer Conference*, 96: 487–491

Reid, R.C. 1976. Superheated liquids. *American Scientist*, 64: 146–156

Reinke, P. & Yadigaroglu, G. 2001. Explosive vaporization of superheated liquids by boiling fronts. *International Journal of Multiphase Flow*, 27: 1487–1516

Shepherd, J.E. & Sturtevant, B. 1982. Rapid evaporation at the superheat limit. *Journal of Fluid Mechanics*, 121: 379–402

Simões-Moreira, J.R. & Shepherd, J.E. 1999. Evaporation waves in superheated dodecane. *Journal of Fluid Mechanics*, 382: 63–86

Terner, E. 1962. Shock tube experiments involving phase changes. *Industrial and Engineering Chemistry Process Design and Development*, 1(2): 84–86

Thompson, P.A., Carofano, G.C. & Kim, Y.-G. 1986. Shock waves and phase changes in a large-heat-capacity fluid emerging from a tube. *Journal of Fluid Mechanics*, 166: 57–92

Thompson, P.A., Chaves, H., Meier, G.E.A., Kim, Y.-G. & Speckmann, H.-D. 1987. Wave splitting in a fluid of large heat capacity. *Journal of Fluid Mechanics*, 185: 385–414

EXCITABILITY

Nineteen century natural history identified irritability—the active response to stimulation—as a characteristic of living systems; this vague concept has developed into excitability, a characteristic of some nonlinear systems.

An excitable system has a stable resting state, and the response to a small brief perturbation is small, with an amplitude that varies smoothly with the perturbation amplitude. The response to a sufficiently large perturbation, above a threshold, is qualitatively different, undergoing a stereotyped, large amplitude excursion in at least one of the state variables before return to the resting state. For a spatially extended excitable system, the subthreshold response is localized while the suprathreshold response is a traveling wave, or traveling wave train. This dynamics—a stable steady state (a mathematical, but not thermodynamic equilibrium), a threshold, and a return to the steady state—requires at least two processes, a fast nonlinear excitation and a slower recovery process.

The excitation process can be exemplified by the bistability of a candle. The unlit candle is stable, and the threshold for igniting the candle arises from a positive feedback, with heat melting and vaporizing the wax, and the vapor providing the fuel sustaining the flame. This positive feedback loop gives the all-or-none nature of the flame—either it is self-sustaining or not, and the threshold separates these two states. The burnt candle is stable; there is no recovery process.

An example of excitation and recovery is provided by the nerve impulse, where the resting state potential is stable and the energy required for the action potential is obtained from the electrochemical gradients across the membrane. The voltage dependence of the Na^+ conductance provides the positive feedback. Recovery is by the slower, voltage-dependent K^+ conductance, and propagation by spread of depolarization along the fiber (Cole, 1968; Aidley, 1998). In both these examples, the spatial spread of activity can be considered as traveling-wave solution of a reaction-diffusion equation in which there is a balance between the active nonlinear term and the diffusive term. In typical chemical excitable systems, all the variables diffuse, with similar diffusion coefficients; in nerve and muscle excitation only the fast excitation variable (transmembrane voltage) spreads and there is only one nonzero diffusion coefficient. The velocity of the traveling-wave solution varies as the root of the diffusion coefficient.

The simplest excitable dynamical system has two kinetic variables, a fast excitation u and a slower recovery v

$$\varepsilon\, \mathrm{d}u/\mathrm{d}t = f(v,u), \quad \mathrm{d}v/\mathrm{d}t = g(v,u), \tag{1}$$

and the nullcline for the fast excitation system has the characteristics of a cubic that is, for a range of u values the solution of $f(v,u)=0$ has three branches. If, as in Figure 1, the nullclines have a single intersection

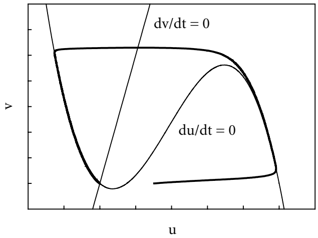

Figure 1. State space of two-variable excitable system, with u the fast excitation variable and v the slower recovery variable. There is a stable equilibrium and a single suprathreshold response.

on the left branch, there is a single globally stable, but excitable, steady state. A large enough perturbation gives a trajectory that jumps to the right branch, moves slowly close to the right branch, and then jumps back to the left branch, returning slowly to rest. If the nullclines intersect on the central branch the resultant equilibrium can be unstable, and the system is no longer excitable but oscillatory.

The key concepts of excitability—threshold and all-or-none response—came from the physiology of nerve and muscle tissue. Nerve and muscle are electrically excitable tissues; their excitability is electrochemical, due to nonlinear membrane current-voltage relations produced by voltage-dependent conductances. The suprathreshold response is the action potential that acts as a trigger, where the intensity of the response is determined by the rate of action potentials. Repetitive activity—that can be idealized as a periodic response—is a biologically significant behavior and may be forced, by inputs to an excitable system, or endogenous, as in a pacemaking system, due to stable limit cycle surrounding an unstable equilibrium when the equilibrium state loses its stability at a Hopf bifurcation as a parameter (say, an applied current or maximal membrane conductance) is changed. Also rate coding phase synchronization effects can be significant—synchronization of bursting discharges in the mammalian central nervous systems provides a candidate mechanism for cognitive effects (Eckhorn, 1999). In spatially extended cells, there can be repetitive traveling-wave trains—effectively one-dimensional in a nerve fiber, and three-dimensional in heart muscle. In the heart three-dimensional scroll waves provide the mechanism for re-entrant arrhythmias that can lead to sudden cardiac death. Some intracellular systems show chemical excitability: a localized increase in intracellular [Ca^{2+}] can trigger calcium-induced calcium release from intracellular organelles, producing intracellular calcium oscillations, or traveling waves. An example is a [Ca^{2+}] wave triggered by fertilization of an oocyte (Dupont & Goldbeter, 1997). Chemical excitability, via cyclic AMP triggering cellular release of cyclic AMP, underlies the wave phenomena seen in the aggregation of the slime mold *Dictyostelium*, a simple example of morphogenesis (Tyson & Murray, 1989). In all these biological examples, high-order, mechanistically detailed stiff systems of equations have been developed, validated, and investigated numerically, but much of the essential phenomenology can be captured by low-order caricatures. Simple models of biological populations can also show excitable behavior, for example, the initiation of an epidemic.

The Belousov–Zhabotinsky reaction, initially developed as an oscillatory organic reaction as an analogue of biochemical oscillations, was the first of several autocatalytic chemical reactions that, in a flow reactor with concentrations that allow bistability, addition of a feedback species can lead to oscillations (Sagues & Epstein, 2003). These chemical oscillatory systems, in thin films and bulk media, also show traveling waves—target patterns, spirals, and scrolls—and with some parameter values are excitable. Models of autocatalytic processes with cubic excitable kinetics are widely used to represent oscillatory and wave phenomena in chemistry, especially combustion problems (Gray & Scott, 1990).

ARUN V. HOLDEN

See also **FitzHugh–Nagumo equation; Hodgkin–Huxley equations; Hopf bifurcation; Nerve impulses; Neurons; Periodic bursting; Threshold phenomena**

Further Reading

Aidley, D.J. 1998. *The Physiology of Excitable Cells*, Cambridge and New York: Cambridge University Press

Cole, K.S. 1969. *Membranes, Ions and Impulses*, Berkley: University of California Press

Dupont, G. & Goldbeter, A. 1997. Modelling oscillations and waves of cytosolic calcium. *Nonlinear Analysis—Theory, Methods and Applications*, 30: 1781–1792

Eckhorn, R. 1999. Neural mechanisms of visual feature binding investigated with microelectrodes and models. *Visual Cognition* 6: 231–265

Gray, P. & Scott, S.K. 1990. *Chemical Oscillations and Instabilities: Non-linear Chemical Kinetics*, Oxford: Clarendon Press and New York: Oxford University Press

Holden, A.V., Markus. M. & Othmer, H.G. 1991. *Nonlinear Wave Processes in Excitable Media*, New York: Plenum

Keener, J.P & Sneyd, J. 1998. *Mathematical Physiology*, New York: Springer

Sagues F. & Epstein, I.R. 2003. Nonlinear chemical dynamics. *Dalton Transactions*, (7): 1201–1217

Tyson, J.J. & Murray, J.D. 1989. Cyclic-AMP waves and aggregation of *Dictyostelium amebas*. *Development*, 106: 421–426

EXCITABLE MEDIUM

See **Reaction-diffusion systems**

EXCITONS

Traditionally, an exciton is defined as a quantum of electronic excitation energy traveling in a periodic structure, whose motion is characterized by a wave vector. It is generated when an electron from a filled electronic orbital of a molecule is transferred through an optical or electrical excitation to a high-energy unoccupied electronic orbital, leaving behind it a hole. Coulomb interaction binds the excited negative electron and the positively charged hole producing an electrically neutral bound state, and this bound electron-hole pair is the exciton. Once created the exciton can be destroyed through emission of light (radiative recombination) or heat (nonradiative recombination), electron-hole separation (exciton dissociation), and subsequent absorption of the electron by an acceptor site (electron transfer). It can change its spin state through intersystem crossing, while at high excitation densities, excitons can mutually collide and dissociate (exciton-exciton annihilation). Furthermore, the whole excitation can move to other sites leading to energy transfer, one of its most important properties.

Excitons occur in molecular crystals, inorganic semiconductors, and conjugated polymers. Three different types of excitons are known. The Frenkel exciton is an electronic excitation of a single molecular unit with both electron and hole located on the same molecule. As any molecule is equally likely to be excited and if there is nonzero coupling between adjacent molecules, this exciton is transferred from one molecular unit to another. Frenkel excitons appear in ionic, molecular, and noble gas crystals having a typical binding energy of 1 eV. In the Wannier exciton the electron-hole distance is much larger than the lattice spacing, and the Coulomb interaction gets screened by the crystal dielectric constant forming a hydrogen-like bound system. They appear mostly in inorganic semiconductors and have a small binding energy of about 0.1 eV and a large radius of typically 50 Å. Finally, the charge-transfer exciton is intermediate between Frenkel and Wannier excitons with an electron-hole distance of a few lattice spacings.

Although excitons are neutral they can be found in definite spin states depending on the relative orientation of electron and hole spins. For antiparallel electron-hole spins a singlet exciton with total spin zero is formed with short lifetime of the order of picoseconds due to fast decay through an optically allowed transition leading to fluorescent emission from this state. There are also three degenerate spin one states leading to a triplet exciton that produces emission termed “phosphorescence.” Optical transitions between triplet excited and single ground state are not allowed due to the forbidden spin flip, resulting in conjugated polymers in triplet exciton lifetime of the order of milliseconds at low temperatures. Thermal energy assists triplet motion and at higher temperatures exciton motion process in an incoherent diffusive mode, although some quantum mechanical coherence may be retained. Coherent exciton dynamics at finite temperatures is investigated theoretically through generalized Langevin equations.

Self-trapped excitons (STEs) form a special class of excitations that involve excitons coupled strongly to vibrational degrees of freedom of molecules. Low-dimensional and, in particular, quasi-one-dimensional materials are ideal systems for studying STEs because reduced dimensionality results in strong electron-phonon interaction. Systems under experimental study include halogen-bridged MX materials, where X is a halogen (chlorine, bromine, etc.) and the metal is, for instance, platinum or crystalline hydrogen-bonded acetanilide (ACN).

The STE can be of electronic or vibrational nature if, in the latter case, specific vibrational modes couple strongly to other phonon system modes leading to vibron excitons. In ACN, C = O as well as NH stretching modes have been linked with STEs through the temperature dependence of specific peaks in their absorption spectra, while in the metal-halogen material PtCl, resonant Raman spectroscopy experiments have indicated the existence of self-localized modes.

From a more formal point of view, STEs can be classified as polarons, solitons, or (discrete) breathers, and depending on their specific nature, they may have some differences in their precise physical properties. A generic mathematical model for exciton self-trapping is provided by the discrete nonlinear Schrödinger equation (DNLS) or discrete self-trapping equation

$$\mathrm{i}\frac{\mathrm{d}\psi_n}{\mathrm{d}t} = \varepsilon_n\psi_n + V(\psi_{n+1} + \psi_{n-1}) - \gamma|\psi_n|^2\psi_n, \quad (1)$$

where ψ_n is the probability amplitude for an exciton to be located in a molecular unit at a crystal site n with local energy ε_n, V is the nearest neighbor overlap, and γ is proportional to exciton-phonon coupling. In this approximate, semiclassical description, the nonlinear term in the equation arises from strong electron-phonon coupling in the original exact problem that can be introduced, for instance, by a model such as the Holstein model for molecular crystals. The conserved number $\sum_n |\psi_n|^2 = N$ of the DNLS equation corresponds to the number of excitation quanta present in the system. In the weak coupling limit where $V \to 0$, the STEs of DNLS are completely localized in a given site with energy equal to

$$E_N = N\varepsilon - \frac{\gamma}{2}N^2, \quad (2)$$

where all local site energies are taken to be identical, that is, $\varepsilon_n = \varepsilon$ for all n. The energy of the STE is thus lower than $N\varepsilon$, the energy of the delocalized exciton and also red-shifted by an amount that grows quadratically in the number of quanta N. These spectral features have been observed experimentally in spectra of both PtCl and ACN as well as in other systems. Precise theoretical analysis of resonant Raman spectra of PtCl as well as femtosecond infrared pump-probe spectroscopy experiments in ACN that go beyond the weak coupling DNLS energy in Equation (2) have demonstrated the existence of STEs in these systems. In the case of ACN, an amide-I exciton lifetime of the order of 2 ps was found. If, as originally conjectured by Davydov, similar modes with larger lifetimes exist in biological macromolecules such as proteins, STEs may participate in energy transfer processes of biological significance.

G.P. TSIRONIS

See also **Davydov soliton; Discrete breathers; Discrete nonlinear Schrödinger equations; Discrete self-trapping system; Polaritons; Polarons**

Further Reading

Dexter, D.L. & Knox, R.S. 1965. *Excitons*, New York: Wiley

Edler, J, Hamm, P. & Scott, A. C. 2002. Femtosecond study of self-trapped vibrational excitons in crystaline acetanilide. *Physical Review Letters*, 88, 067403: 1–4

Eilbeck, J.C, Lomdahl, P.S. & Scott, A.C. 1985. The discrete self-trapping equation. *Physica D*, 16: 318–338

Hennig, D. & Tsironis, G.P. 1999. Wave transmission in nonlinear lattices. *Physics Reports*, 307: 333–432

Kenkre, V.M. & Reineker P. 1982. *Exciton Dynamics in Molecular Crystals and Aggregates*, Berlin: Springer

Pope, M. & Swenberg, C.E. 1982. *Electronic Processes in Organic Crystals*, Oxford: Clarendon Press and New York: Oxford University Press; 2nd edition, 1999

Scott, A.C. 1992. Davydov's soliton. *Physics Reports*, 217: 1–67

Voulgarakis, N.K., Kalosakas, G., Bishop, A.R. & Tsironis, G.P. 2001. Multiquanta breather model for PtCl. *Physical Review* B, 64 (020301): 1–4

EXPLOSIONS

The English word *explode* comes from the Latin word *explodere*, which means to drive an actor from the stage by rhythmic clapping. In applied science, a sudden deposition of energy leads to an explosion. Since on a streamline of a flow we have $\mathrm{d}E/\mathrm{d}t = -p\mathrm{d}v/\mathrm{d}t$ (where E is internal energy, t is time, p is pressure, and v is volume), rapid deposition of energy will give high pressure, or rapid motion as volume expands, or both. How large and how rapid the energy deposition must be to have an explosion depends on external circumstances, so the concept of explosion is investigated here by looking at some examples.

Low Explosives and Propellants

The words *energetic materials* are used to describe substances that have fuel and oxidizer in the same molecule or in separate molecules intimately mixed. Production and purification of potassium nitrate (KNO_3) were the keys to making materials that burned rapidly and could be used for fireworks, frightening enemies, setting enemy ships on fire, and entertainment at fairs. After 1300 AD, these materials were used in guns, and the mixtures were improved for better performance. High-strength black powder is about six parts of potassium nitrate and one part each of charcoal and sulfur. Black powder has been replaced by smokeless powders made from nitrocellulose or nitrocellulose with nitroglycerin, with various additives. Modern rocket motors usually use a coarse mixture of ammonium perchlorate (NH_4ClO_4) as oxidizer, and rubber, aluminum powder, and high explosive molecules as fuels. Black powder was used in mining from 1650 to 1900; it has since been replaced by high explosives.

All of these materials deliver their energy by burning, which is also called deflagration. That is, energy is transferred from the reacting region forward to ignite the unreacted material mainly by heat conduction. The wave speed of the flame is slow compared with the speed of sound, so that the pressure is nearly constant in the flame. The distance from the reacting region to the wave front varies with the pressure, so the speed of the flame and its rate of release of energy depend on the external confinement. In a mine, the apparent power of black powder varies with the strength of the rock and the existing cracks in the rock. The powder in a gun is usually in the form of separate pellets or grains, and the rate of burning depends on the exposed surface area that supports the flame. As the projectile moves along the barrel, the volume to be occupied by the gases increases at an accelerating rate. It is advantageous to have the rate of burning also accelerate, and this is accomplished by shaping the grains. For example, the grains may be in the form of cylindrical pellets with a number of holes parallel to the axis. As the pellet burns, the holes are enlarged, exposing more area and accelerating the rate of gas production.

High Explosives

High explosives are also energetic materials with the fuel and oxidizer in intimate mixture or in the same molecule. They explode by a process called detonation, and the speed of the reaction is 100,000–10,000,000 times faster than burning in the same material. The rapid reaction produces extremely high pressures and speeds. It also makes the detonation process almost independent of confinement or boundary effects. High explosives detonate the same way every time, without being influenced by exterior changes.

Detonation proceeds as a wave traveling at high speed, from 4000 to 9000 ms^{-1} in various explosives, with a shock wave at the front of the wave structure. The compression in the shock heats the explosive enough to make it react rapidly. The energy released in the reaction supports the shock. The shock is supersonic, going as fast as the available energy can drive it, so nothing occurs ahead of the shock. In the reaction zone the flow is subsonic so energy can be transferred, but at the end of the reaction zone the flow becomes supersonic, and anything that happens in that supersonic flow cannot come forward to affect the reaction zone. It is for these reasons that the detonation is almost independent of the external conditions. Only at the edges can it be perturbed.

High explosives were discovered about the middle of the 19th century by chemists developing new dyes. Ascanio Sobrero, an Italian chemist, first prepared nitroglycerin and several other high explosives in 1846 and 1847. Alfred Nobel in 1864 developed a method of initiating nitroglycerin reliably by a fast compression wave, or shock wave, but it was dangerous to handle and ill-suited to mining operations. In 1867 Nobel was granted a patent for an explosive prepared by mixing nitroglycerin with a porous absorbent such as charcoal, diatomaceous earth, or fine sawdust. This explosive, called dynamite, proved to be just what was needed for the developing industrial revolution. In a few years it replaced black powder throughout the world. Many compositions competed with dynamite: a dictionary of explosives published in 1891 listed more than a thousand explosives.

After the middle of the 20th century, dynamite was replaced by explosives prepared from ammonium nitrate and fuel oil as granular mixtures, as slurries, or as emulsions. In the United States alone the annual use of high explosives is more than five billion pounds per year, or nearly 20 pounds for each person. A pound of explosive will break about 3000 pounds of rock, so about 60,000 pounds of rock is broken for each person in the United States each year. A large fraction of the explosive is employed to break the rock overburden of coal seams, other uses are in other mining, road construction, lumbering, and farming. Almost any product we purchase has had explosive used for it at some stage.

The first high-precision application of high explosives was to nuclear weapons, developed during World War II, where explosive was used for the precise dynamic assembly of the fissionable material. The nuclear weapons laboratories around the world have contributed to the detailed theory of detonation, to numerical modeling of detonations, to the development of powerful and safe explosives, and to the experimental techniques for studying explosives.

Nonlinear relations are especially important to detonation. The shock wave forms because the stress-strain behavior of the explosive is nonlinear. Materials become stiffer as they are compressed, so the sound speed increases with pressure, and a compression wave steepens to become a shock. The extremely strong nonlinear dependence of the reaction rate on temperature is also very important. It is the on-off switch for the release of energy. For example, a typical rate law for an explosive might be $\mathrm{d}\lambda/\mathrm{d}t = k(1-\lambda)\mathrm{e}^{(-T^*/T)}$, where k is the limiting rate; λ is a progress variable for the reaction, going from zero to one as the explosive goes to products; t is the time; T is the temperature; and T^* is the activation temperature. If $k = 10^{20}\,\mathrm{s}^{-1}$ and $T^* = 25{,}000$ K, the time required for half of the explosive to react when kept at constant temperature is 341 million years at 300 K (room temperature), 8.64 ms at 600 K, and 8.03 ns at 900 K.

Nuclear Weapons

In nuclear fission weapons, U^{235} or Pu^{239} is bombarded by neutrons that cause it to split into two lighter nuclei, releasing a large amount of energy and also emitting more neutrons. If more neutrons are generated than escape through the surface, the exponential increase in the number of neutrons leads to an explosion. A fusion weapon is initiated by a fission weapon and releases energy as hydrogen, deuterium, tritium, and lithium fuse to make heavier nuclei. These weapons release a million or more times as much energy as the same weight of conventional chemical explosive. The energy is often quoted in kilotons or megatons. These units are based on an assumed yield for TNT of 1000 calg^{-1} or 4187 Jg^{-1}. The metric tonne is 10^6 g and the kiloton is a thousand times that, so a kiloton is 10^{12} cal or 4.187×10^{12} J, and a megaton is 4.187×10^{15} J.

Asteroid Collision with Earth

About 65 million years ago an asteroid hit the earth leaving the crater Chicxulub in Yucatan, Mexico, and perturbing the conditions on earth to the extent that dinosaurs and many other animals and plants became extinct. The energy deposited was the kinetic energy of the asteroid. The exact details cannot be known, but if the asteroid was 12 km in diameter with density 2500 kg m^{-3} and relative speed 20 km s^{-1}, its kinetic energy was 4.5×10^{23} J. An obvious time constant for the energy release is the diameter divided by the speed, 0.6 s. The kinetic energy divided by the time gives the power as 7.5×10^{23} W; the actual value is a few times smaller.

An asteroid of the Chicxulub size is not slowed much by the Earth's atmosphere and reaches the Earth's surface. A smaller asteroid is slowed more, and the compression wave from the air travels through the asteroid. If the asteroid is small enough, the compression wave reaches the rear surface before the asteroid reaches the ground. The high pressure inside the asteroid causes it to expand and break up into small pieces that are then slowed more by the air. The kinetic energy is transferred suddenly to the air, and there is an explosion above the surface of the earth. The Tunguska Event in Central Siberia in 1908 flattened trees over an area of about 2000 km^2. It was caused by an asteroid about 60 m diameter that exploded about 8 km above the surface. Its kinetic energy was about 4×10^{16} J.

Volcanoes

The 1883 eruption of the Indonesian volcano Krakatau, which blew ash to a height of 80 km and ejected 22 km^3 of rock, was heard in Australia 4600 km away. Tsunamis caused by the eruption reached heights of 40 m and killed 34,000 people.

Mount St. Helens, in the state of Washington, erupted in 1980 and has been studied extensively. It ejected 2.7 km^3 of rock, devastated an area of 550 km^2, and blew down an estimated 10^7 trees.

A huge volcanic explosion occurred in Yellowstone, United States, about 2,000,000 years ago. About 3000 km^3 of rock were ejected. No explosion of such a magnitude has been experienced during the period of human civilization.

Novae and Supernovae

A nova occurs in a close binary star system made up of a red giant and a white dwarf. Hydrogen-rich matter from the red giant is pulled onto the surface of the white dwarf. When enough matter is accumulated, a nuclear fusion detonation occurs, causing the ejection of hot surface gases and resulting in an extraordinary increase in luminosity. A supernova occurs when a massive star exhausts its nuclear fuel and its core collapses to become a neutron star. The outer layers of the star are attracted by the gravitational pull of the core, and then rebound from it. A shock wave generated in the collision propagates outward and blows off the surface gases.

W.C. Davis

See also **Dimensional analysis; Flame front; Shock waves**

Further Reading

Baker, W.E., 1973. *Explosions in Air*, Austin: University of Texas Press

Baker, W.E., Cox, P.A., Westine, P.S., Kulesz, J.J. & Strehlow, R.A. 1983. *Explosion Hazards and Evaluation*, Amsterdam and New York: Elsevier

Federoff. B.T. et al. 1960. *Encyclopedia of Explosives and Related Items*, 10 vols, Dover, NJ: US Army Armament Research and Development Command (Picatinny Arsenal), (PATR-2700)

Glasstone, S. & Dolan, P.J. (editiors). 1977. *The Effects of Nuclear Weapons*, 3rd edition, Washington: United States Department of Defences

International Astronomical Union, 1991. *Physics of Classical Novae*, Berlin and New York: Springer

Kinney, G.F. & Graham, K.J. 1985. *Explosive Shocks in Air*, 2nd edition, Berlin and New York: Springer

Sigurdsson, H. et al. (editors). 2000. *Encyclopedia of Volcanoes*, San Diego and London: Academic Press

Wheeler, J. Craig 2000. *Cosmic Catastrophes: Supernovae, Gamma Ray Bursts, and Adventures in Hyperspace*, Cambridge and New York: Cambridge University Press

EXTREMUM PRINCIPLES

Extremum principles are ubiquitous in mathematics, with wide applications in areas ranging from genetic theory (Narain, 1993) to thermodynamics (Velasco & Fernandez-Pineda, 2002). The simplest sort of extremum principle occurs in differential geometry, whenever one considers a differentiable function $f(x_1, \ldots, x_N)$ of N variables and wishes to determine the critical points of f, namely those $\boldsymbol{x}$ that satisfy

$$\nabla f(\boldsymbol{x}) = 0.$$

The Hessian matrix H has entries given by the second derivatives

$$H_{jk} = \frac{\partial^2 f}{\partial x_j \partial x_k}(\boldsymbol{x})$$

(assuming these exist). If all the eigenvalues of the Hessian are positive (negative), then this ensures that $\boldsymbol{x}$ is a local minimum (maximum); eigenvalues of mixed sign correspond to a saddle point. The general characterization of functions in terms of their behavior near critical points is the subject of Morse theory (Morse & Cairns, 1969).

Probably the most widespread type of extremum principles are those arising from the application of the calculus of variations (Gel'fand & Fomin, 1963). One of the first historical examples of this sort is the brachistochrone problem posed by Johann Bernoulli the elder in 1696: given a particle sliding from rest along a path between two fixed points under the influence of gravity, find the plane curve such that the time taken is minimized. The solution curve (a cycloid) was obtained by both Johann and Jakob Bernoulli, and also by Newton, Leibniz, and l'Hôpital. Nowadays, the problem is usually solved directly through the use of the total time functional

$$T = \int_{x_0}^{x_1} \sqrt{\frac{1+(y')^2}{2g(y_0 - y)}}\,\mathrm{d}x \equiv \int_{x_0}^{x_1} F(y, y')\,\mathrm{d}x,$$
$$y' = \frac{\mathrm{d}y}{\mathrm{d}x}, \tag{1}$$

where $y(x)$ is the plane curve joining the points (x_0, y_0) and (x_1, y_1) and g is the acceleration due to gravity. To minimize the time, it is necessary that the first variation of the functional T should vanish,

$$\delta T = 0,$$

which leads to the Euler–Lagrange equations

$$\frac{\partial F}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial y'}\right) = 0 \qquad (2)$$

The solution of (2) is given parametrically in terms of (x_0, y_0) and another constant h by

$$x(\theta) = x_0 + (\theta - \sin\theta)/(2h^2),$$
$$y(\theta) = y_0 - (1 - \cos\theta)/(2h^2).$$

In complete analogy to the Hessian condition in finite dimensions, the positivity of the second variation $\delta^2 T$ ensures that the cycloid yields a minimum of the functional $T[y]$.

In physics perhaps the most important extremum principle of all is the Principle of Least Action. An early proponent of this notion was Pierre Louis de Maupertuis, who employed an extremality argument to solve a problem in statics (Beeson, 1992). Maupertuis, who was tutored by Johann Bernoulli, was in regular correspondence with Leonhard Euler, and the work of the latter laid the foundation for the subsequent development of the subject. The enormous generality of the least-action principle became apparent from the Lagrangian formulation of classical mechanics (Goldstein, 1950), as derived from the action functional

$$S[\boldsymbol{q}] = \int_{t_0}^{t_1} L(\boldsymbol{q}, \dot{\boldsymbol{q}}, t)\,\mathrm{d}t. \qquad (3)$$

The variable t is the time, and the vector $\boldsymbol{q}$ and its time derivative $\dot{\boldsymbol{q}}$ denote the generalized coordinates and velocities, respectively. Requiring that the action should be stationary with respect to variations that vanish at the initial and final times, so

$$\delta S = 0, \quad \text{with } \delta\boldsymbol{q}(t_0) = 0 = \delta\boldsymbol{q}(t_1), \qquad (4)$$

it follows (after an integration by parts) that L must satisfy the Euler–Lagrange equations

$$\frac{\partial L}{\partial \boldsymbol{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\boldsymbol{q}}}\right) = 0.$$

The extremality condition (4) is truly fundamental, since once the Lagrangian L has been specified appropriately, then all the classical equations of motion for a mechanical system (in particular, Newton's Second Law) are a direct consequence. Even the geodesic equations for a Riemannian manifold arise in this way (Choquet-Bruhat, 1968), by taking the purely kinetic (free particle) Lagrangian

$$L(\boldsymbol{q}, \dot{\boldsymbol{q}}) = g_{jk}(\boldsymbol{q})\dot{q}^j\dot{q}^k,$$

where g_{jk} is the metric tensor (the Einstein summation convention is assumed).

In parallel to the development of mechanics, a corresponding extremum principle was employed in optics, namely Fermat's Principle of Least Time. This is more correctly formulated within the theory of geometrical optics (Kline & Kay, 1965), as the requirement that the path taken by light traveling between two points P_1 and P_2 is such that the optical distance is *extremized*, that is,

$$\delta \int_{P_1}^{P_2} n(x, y, z)\,\mathrm{d}s = 0, \qquad (5)$$

where $n(x, y, z)$ is the refractive index at the point (x, y, z) in the medium and s denotes arc length along the path. Contrary to Fermat's original formulation, if the optical distance is extremal, then in general, the time taken for the light to travel between P_1 and P_2 need not be minimized—this only holds in certain special cases.

The precise analogy between geometrical optics and mechanics was fully appreciated and exploited by Hamilton. Choosing a parameter τ along the light path such that $\boldsymbol{q} = (x, y, z)^T$ satisfies

$$|\dot{\boldsymbol{q}}|^2 = n^2, \quad \mathrm{d}s = |\dot{\boldsymbol{q}}|\,\mathrm{d}\tau, \qquad (6)$$

the equations arising from the variational principle (5) are just

$$\ddot{\boldsymbol{q}} = \nabla\left(\tfrac{1}{2}n^2\right), \quad n = n(\boldsymbol{q}),$$

which are equivalent to Newton's equations for a particle of unit mass moving in a potential $-\frac{1}{2}n^2$. In this mechanical analogy, τ should be reinterpreted as the time, while the first equation (6) implies that the total energy of the particle is zero.

The Principle of Least Action is also essential in field theory. A scalar field theory in N dimensions has the action

$$S[\phi] = \int_V \mathcal{L}(\phi, \phi_\mu, \ldots)\,\mathrm{d}^N x$$

with the Lagrangian density $\mathcal{L}$ being a function of the field ϕ, its first derivatives ϕ_μ $(\mu = 1, \ldots, N)$, and possibly higher derivatives. The appropriate variational principle for the classical field equations is

$$\delta S = 0 \quad \text{with } \delta\phi(\boldsymbol{x}) = 0 \quad \text{for } \boldsymbol{x} \in \partial V,$$

so that the variations vanish on the boundary of the space-time volume V. In the path integral formalism of quantum field theory, as pioneered by Richard Feynman (Feynman & Hibbs, 1965), the central object is the partition function, defined in terms of the action by

$$Z = \int \mathrm{e}^{\mathrm{i}S[\phi]}\,\mathcal{D}\phi,$$

where $\mathcal{D}\phi$ = path integral measure. When the action is coupled to an external source, $S \to S + \int J(\boldsymbol{x})\phi(\boldsymbol{x})\,\mathrm{d}^N x$, then the vacuum expectation values of the field operators are obtained by taking successive functional derivatives $\delta/\delta J$ of Z. This is the key technique for calculating Feynman rules and perturbative scattering amplitudes in gauge field theory (Bailin & Love, 1986).

Another sort of extremum principle appears in game theory (Owen, 1982). In the simplest case of two-person zero-sum games, with two players denoted "row" and "column" and strategies labelled by an index j ($j = 1, \ldots, M$) for the row player and an index k ($k = 1, \ldots, N$) for the column player, the payoffs to the row player form an $M \times N$ matrix $\mathcal{A}$. The matrix element $\mathcal{A}_{jk}$ is just the payoff to the row player arising from the pure strategy combination (j, k); since the game is zero-sum, the corresponding payoff to the column player is $-\mathcal{A}_{jk}$. If the maximum of the row minima in the payoff matrix is equal to the minimum of the column maxima, then this is a saddle point (or minimax point), corresponding to an optimum solution of the game. However, a saddle point does not exist in general, and so it is necessary to resort to mixed strategies, taking the set of probability distributions $\mathcal{S}_X \times \mathcal{S}_Y$ on the strategy space, where

$$\mathcal{S}_X = \left\{ \boldsymbol{x} \in \mathbb{R}^M \mid x_j \in [0, 1],\ \sum_{j=1}^{M} x_j = 1 \right\},$$

$$\mathcal{S}_Y = \left\{ \boldsymbol{y} \in \mathbb{R}^N \mid y_k \in [0, 1],\ \sum_{k=1}^{N} y_k = 1 \right\}.$$

Then the Minimax Theorem, due to John von Neumann (von Neumann & Morgenstern, 1947), states that

$$\max_{\boldsymbol{x} \in \mathcal{S}_X} \min_{\boldsymbol{y} \in \mathcal{S}_Y} \boldsymbol{x}^{\mathrm{T}} \mathcal{A} \boldsymbol{y} = v = \min_{\boldsymbol{y} \in \mathcal{S}_Y} \max_{\boldsymbol{x} \in \mathcal{S}_X} \boldsymbol{x}^{\mathrm{T}} \mathcal{A} \boldsymbol{y}. \qquad (7)$$

The quantity v is the expected payoff to the row player corresponding to some optimal mixed strategy combination $(\boldsymbol{x}, \boldsymbol{y})$ and is called the *value* of the game. The existence of a more general sort of equilibrium in n-person, noncooperative games has been proved by John Nash (1951). Nash's proof relies on properties of convex functions (as does the standard proof of the Minimax Theorem). Convex analysis has also been used recently in thermodynamics, for a general proof of the minimum energy principle, starting from the maximum entropy principle and using the concavity of the entropy (Prestipino & Giaquinta, 2003).

ANDREW HONE

See also **Euler–Lagrange equations; Game theory; Geometrical optics, nonlinear; Quantum field theory**

Further Reading

Bailin, D. & Love, A. 1986. *Introduction to Gauge Field Theory*, Bristol: Adam Hilger

Beeson, D. 1992. *Maupertuis: An Intellectual Biography*, Oxford: Voltaire Foundation

Choquet-Bruhat, Y. 1968. *Géométrie Différentielle et Systèmes Extérieurs*, Paris: Dunod

Feynman, R.P. & Hibbs, A.R. 1965. *Quantum Mechanics and Path Integrals*, New York: McGraw-Hill

Gel'fand, I.M. & Fomin, S.V. 1963. *Calculus of Variations*, Englewood Cliffs, NJ: Prentice-Hall

Goldstein, H. 1950. *Classical Mechanics*, Reading, MA: Addison-Wesley

Kline, M. & Kay, I.W. 1965. *Electromagnetic Theory and Geometrical Optics*, New York: Interscience Publishers, Wiley

Morse, M. & Cairns, S.S. 1969. *Critical Point Theory in Global Analysis and Differential Topology*, New York: Academic Press

Narain, P. 1993. On an extremum principle in the genetical theory of natural selection. *Journal of Genetics*, 72(2–3): 59–71

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics*, 54: 286–295

Owen, G. 1982. *Game Theory*, 2nd edition, London: Academic Press

Prestipino, S. & Giaquinta, P.V. 2003. The concavity of entropy and extremum principles in thermodynamics. *Journal of Statistical Physics*, 111(1-2): 479–493

Velasco, S. & Fernandez-Pineda, C. 2002. A simple example illustrating the application of thermodynamic extremum principles. *European Journal of Physics*, 23: 501–511

von Neumann, J. & Morgenstern, O. 1947. *Theory of Games and Economic Behaviour*, Princeton, NJ: Princeton University Press

F

FABRY–PEROT RESONANCES

See **Laser**

FAIRY RINGS OF MUSHROOMS

Much of the mythology and mysticism formerly associated with fairy rings is summed up by the words of Prospero in Shakespeare's *The Tempest*, V.i: "Ye elves of hills. . .that/ By moonshine do the green sour ringlets make,/ Whereof the ewe not bites, and you whose pastime/ Is to make midnight mushrooms. . .." However, in what could be interpreted as a minor triumph of reason over superstition, Prospero's invocation is not to the sorcery that was supposedly imbued in fairy rings, but a declaration that he is giving up magic forever.

Fairy rings are not caused by supernatural beings, witches, moles, snails mating, or lightning—these being some of the early explanations, all the more hilarious *now* for being propounded so seriously back *then*, for the annular rings of dead grass, fringed on both edges by concentric rings of over-lush grass, that are often evident in grassy fields. They are caused by fungi. One of the first scientific investigations to establish this fact was reported by Wollaston (1807), whose explanation of fairy rings is still broadly accepted today. Despite this demystification, their appeal to the human imagination remained strong. Kipling (1906) wrote *Puck of Pooks Hill*, a story in which magical manifestations occur when children perform *A Midsummer Night's Dream* in a fairy ring, and Conan Doyle in *The Coming of the Fairies* (1923) only reluctantly admitted that fairy rings were due to fungal growth.

The mode of life involving progressive radial increase from a central point is not unusual among fungi. As pointed out by Ramsbottom (1953), hundreds of fungi grow in circular patterns, including microscopic ones such as *Penicillium* molds as well as those with macroscopic fruiting bodies such as mushrooms and puff-balls. Although the conditions governing initiation of a ring are still obscure, many scientific studies have documented the kinetics of expansion, the species involved, and the biology and ecology of fairy rings. Two excellent and complementary studies are reported in Shantz & Piemeisel (1917) and Dowson et al. (1989).

The underground body of a fairy ring fungus, consisting of a network of filaments called the mycelium, grows radially outwards as it consumes organic matter in the soil. Behind the fungus front, the mycelial mass dies, so that the advancing live fungus front is actually a strange sort of disconnected organism. The dead filaments form a dense and water-repellant mat. Grass in this advancing annular region dies, due simply to physiological drought. Inside this bare region, the grass can grow luxuriantly because the dead filaments eventually decay, providing nitrogen-rich fertilizer. In advance of the fungal front, the grass also grows dark and luxuriant because of the peculiar (to us) eating habits of fungi: they exude digestive enzymes into the medium, then absorb this pre-digested food. Left-over digested food stimulates the grass forward of the fungus.

The radial growth rate of grassland fairy rings has been measured between 99 and 350 mm/year (Dickinson, 1979) and diameters of tens and hundreds of meters have been recorded. The obvious but rather awesome conclusion is that fairy rings may be among the world's oldest living organisms, since many rings are estimated to be several centuries old and some are believed to be 600 or 700 years old. It would be a fascinating exercise to obtain supporting evidence, such as historical records or results from scientific dating methods, for the ages of large fairy rings.

At favorable times of the year, fruiting bodies—mushrooms or toadstools—may be put forth around the circumference of a fairy ring. In woodlands these are often the only obvious manifestation of a fairy ring, since leaf-litter usually covers the ground (Figure 1). Because they depend on the roots of a tree for nutrient supply, woodland fairy rings are referred to as "tethered" rings, whereas those in grassland, for which the nutrient source is spread though the ground, are called "free" rings. The growth of tethered rings is coupled to the radial growth of the host tree roots, and they tend to reach an equilibrium diameter, determined

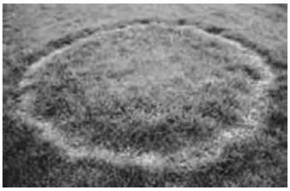

Figure 1. A typical fairy ring in grass.

by the mature size of the tree, rather than increase indefinitely (Gregory, 1982).

What happens when two or more fairy rings meet? Usually, the putative intersecting portions of same-species rings are extinguished, because of direct competition for resources or as each reaches the other's annular dead zone where nutrients and moisture are depleted in any case. Rings of different species either continue growing through each other or only the dominant species may survive. Rings of different species may also form inside one another.

A simple mathematical model for the ecology of fairy ring systems was developed by Parker-Rhodes (1955), from which was derived estimates for the proportion of ground covered by rings at a given time, the age distribution of rings, and geometric factors affecting inter- and intra-specific competition, such as rate of birth of new rings per unit area and distance between their centers. This model was extended by Stevenson & Thompson (1976) to include boundary effects and a more realistic treatment of the rings as annuli rather than discs. Their conclusions had some interesting implications for the management and control of fairy rings.

Why would anyone want to control them? Some people, sadly, consider fairy rings to be a pathogenic nuisance, a disease to be eradicated; and it is true that they can spoil the appearance and functionality of lawns, golf courses, playing fields, and pastures. The modeling results of Stevenson & Thompson (1976) indicated that if one's aim is to control a population of harmful rings by using an antagonistic but innocuous species of rings, it is more effective in the long term to choose the *smallest* growth rate available. In their work, the rings' growth rate was assumed constant.

If we admit, however, that the growth rate is variable, then fairy rings must be modeled as a dynamical system, and since they spread in two spatial dimensions, the appropriate dynamical description is a system of partial differential equations. Davidson et al. (1997) modeled the spatiotemporal dynamics of fungal mycelia by nonlinear reaction-diffusion equations describing the coupled evolution of the mycelial biomass and nutrient substrate concentration. They found that qualitative features of the development of fairy rings, such as the annular advancing front, degeneration of colony centers, and extinction of the interface between two colliding fronts, were reflected in the structure of solutions to the equations.

This type of predictive computational modeling and simulation of radial growth patterns is likely to become more important in future—and not only because it gives fairy rings the ultimate modern imprimatur and authority of the computer. Other organisms are known to exhibit the fairy ring habit of growth. In the semi-arid rangelands of Australia, for example, certain species of saltbush grow radially outwards from a central origin forming a slowly increasing ring of foliage, the interior of which is bare ground. The patterns on the landscape formed by these blue-gray saltbush rings on the red soil are very striking from the air. The saltbush is a nutritious food source for sheep; thus one might be interested in management strategies that *increase* the covering fraction of saltbush and in factors that affect its growth. In this and other similar problems, the results from mathematical modeling of fungus fairy rings will provide valuable insights.

ROWENA BALL

See also **Growth patterns; Pattern formation; Reaction-diffusion systems**

Further Reading

Davidson, F.A., Sleeman, B.D., Rayner, A.D.M., Crawford, J.W. & Ritz, K. 1997. Travelling waves and pattern formation in a model for fungal development. *Journal of Mathematical Biology,* 35: 589–608

Dickinson, C.H. 1979. Fairy rings in Norfolk. *Bulletin of the British Mycological Society,* 13: 91–94

Dowson, C.G., Rayner, A.D.M. & Boddy, L. 1989. Spatial dynamics and interactions of the woodland fairy ring fungus, *Clitocybe nebularis*. *New Phytologist,* 111: 699–705

Gregory, P.H. 1982. Fairy rings: free and tethered. *Bulletin of the British Mycological Society,* 16: 161–163

Parker-Rhodes, A.F. 1955. Fairy ring kinetics. *Transactions of the British Mycologcal Society,* 38 (1): 59–72

Ramsbottom, J. 1953. *The New Naturalist Mushrooms & Toadstools,* London: Collins

Shantz, H.L. & Piemeisel, R.L. 1917. Fungus fairy rings in eastern Colorado and their effect on vegetation. *Journal of Agricultural Research,* 11: 191–247

Stevenson, D.R. & Thompson, C.J. 1976. Fairy ring kinetics. *Journal of Theoretical Biology,* 58: 143–163

Wollaston, W.H. 1807. On fairy-rings. *Philosophical Transactions of the Royal Society of London,* 2: 133–138

FARADAY WAVES

See **Surface waves**

FEEDBACK

Long before the invention of electrical amplifiers and generators, feedback was used widely in mechanical control devices, including the governor that was devised by James Watt in 1784 to stabilize the speed of a steam engine. The theory of such a governor was developed in the latter half of the 19th century by James Clerk Maxwell and by Ivan Alekseevich Vyshnegradskiy. Although the results obtained by Maxwell and Vyshnegradskiy seem contradictory, they present complementary aspects of the analysis. We consider briefly the detailed discussion of these results as given by Neimark et al. (1985).

The Lagrange equations for a steam engine with a governor are

$$I(\vartheta)\ddot{\varphi} + I'_{\vartheta}(\vartheta)\dot{\vartheta}\dot{\varphi} = M_{\mathrm{r}}(\dot{\varphi}, \vartheta) - M_{\mathrm{l}},$$

$$I_{\mathrm{g}}\ddot{\vartheta} - \tfrac{1}{2} I'_{\vartheta}(\vartheta)\dot{\varphi}^2 + U'_{\vartheta}(\vartheta) + h\dot{\vartheta}, \qquad (1)$$

where $I(\vartheta)$ is the moment of inertia of the engine and governor about the engine shaft, φ is the rotation angle of the engine shaft, ϑ and I_{g} are, respectively, the angular deviation and the moment of inertia of the governor balls, $M_{\mathrm{r}}(\dot{\varphi}, \vartheta)$ is the driving torque, M_{l} is the moment of load, and $-U'_{\vartheta}(\vartheta)$ and $\dot{\vartheta}$ are, respectively, the moments of gravity and viscous friction.

For ideal functioning, Maxwell assumed that the friction factor h must be as small as possible and the governor sensitivity $\mathrm{d}\vartheta_0/\mathrm{d}\Omega$ (where $\Omega = \dot{\varphi}_0$, φ_0 and ϑ_0 are the equilibrium values of φ and ϑ) must be as large as possible, and he gave recommendations how these conditions could be fulfilled. Vyshnegradskiy, on the other hand, understood that the friction plays an important role by promoting the stability of control. Therefore, he proposed adding a new element to the governor, called the cataract. In addition, Vysnegradskiy showed that for stability of control $\mathrm{d}\Omega/\mathrm{d}M_{\mathrm{l}}$ must be negative and constructed his famous diagram in the plane $\mathrm{d}\Omega/\mathrm{d}M_{\mathrm{l}}$, Ih/I_{g}, now known as the Vysnegradskiy diagram. Maxwell's ideas have also been realized in practice by a device called an "isodromic" in which feedback is effected by a rod that changes its length due to stretching or compressive force.

In the context of electronic amplifier theory, a detailed mathematical treatment of linear feedback was given by Henrik Bode (1950). Here the basic formula is

$$G = \frac{A}{1 - \mu A}, \qquad (2)$$

which was famously discovered by Harold Black in 1927 (Brittain & Black, 1997). In this equation, A is the (large) forward gain of an amplifier, μ is the loss of a passive feedback circuit, μA is the open-loop gain, and G is the closed-loop gain. Assuming the amplifier does not oscillate and making $\mu A \gg 1$ implies $G \approx 1/\mu$. Thus, the closed-loop gain is essentially determined by the properties of a passive circuit (μ), which is far less sensitive to variations in temperature and aging than the forward gain (A). Long-distance telephone service—which requires a large number of highly stable repeater amplifiers—was made possible by Black's invention.

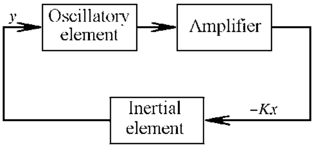

Figure 1. The block diagram of the simplest self-oscillatory system with inertial excitation.

To avoid low-frequency oscillations, the amplifier system of Equation (2) is designed with the open-loop gain having a phase shift of 180° at low frequencies, which corresponds to negative feedback. To avoid oscillations resulting from positive feedback at higher frequencies (where the phase shift approaches 360°), it is necessary and sufficient that the complex mapping of the frequency axis by the open-loop gain function does not surround the point +1 in the μA-plane—the Nyquist criterion (Bode, 1950).

Positive feedback reveals itself in the appearance of negative friction or negative resistance, whereas negative feedback causes the stabilization of device parameters. In Watt's governor, for example, negative feedback stabilizes the speed of rotation, and in other devices it can stabilize frequency variations, the gain factor, and so on. In certain systems, however, negative feedback can result in self-excitation of oscillations. Such systems were called "systems with inertial excitation" by Babitzky & Landa (1982, 1984), and they often arise in physical and engineering studies. The excitation of oscillations in such systems is conditioned by the so-called inertial interaction between dynamical variables occurring as a result of inertia in the feedback loop.

A block diagram of the simplest self-oscillatory system with inertial excitation is shown in Figure 1. The inertial interaction between dynamical variables, like negative friction, can be both linear and nonlinear. A linear interaction can, under certain conditions, lead to spontaneous oscillations, whereas a nonlinear interaction can result in the hard excitation of oscillations (requiring that a threshold be overcome).

A simple system with linear inertial interaction is

$$\ddot{x} + 2\delta\dot{x} + \omega_0^2 x = -ky + f(x, \dot{x}, y),$$

$$\dot{y} + \gamma y = ax + \varphi(x, \dot{x}, y), \qquad (3)$$

where $f(x, \dot{x}, y)$ and $\varphi(x, \dot{x}, y)$ are nonlinear functions free from linear terms, and the parameter a is proportional to the gain of the amplifier. The inertia of the feedback loop is characterized by the parameter γ. The condition of self-excitation of oscillations can be shown to be $\gamma \leq \gamma_{\rm cr}$, where

$$\gamma_{\rm cr} = -\delta + \sqrt{\delta^2 + ak/2\delta - \omega_0^2}\,.$$

Because $\gamma_{\rm cr}$ should be positive, self-excitation can occur only for $ak \geq 2\delta\omega_0^2$. The condition $\gamma \leq \gamma_{\rm cr}$ signifies that the feedback loop must be sufficiently inertial (an inertia-less feedback loop corresponds to $\gamma \to \infty$).

Systems with inertial feedback need not be self-excited. A child on a swing, for example, is an inertial system that is not self-excited. Interestingly, the oscillation of a swing is often inaccurately presented in textbooks as an example of parametric excitation. If oscillations are excited by a child who stands on the swing and lifts her center of gravity up and down at the proper moments, the system is not parametrically excited, because the oscillation frequency of the center of gravity is not constant but tuned according to variations of the frequency of the oscillations of the swing. Thus, a child on a swing is a control system with feedback that is self-oscillatory but not self-excited. For excitation of self-oscillations some finite initial perturbation is necessary (i.e., the excitation of oscillations is hard).

The simplest equations describing oscillations of a swing are

$$\ddot{x} + 2\delta\dot{x} + \omega_0^2 x = -bxy, \quad \dot{y} + \gamma y = ax^2, \qquad (4)$$

where the variable y describes the position of the child's center of gravity and γ is the inertial factor of the control circuit. If the feedback is slightly inertial (i.e., γ is too large to put $y \approx (a/\gamma)x^2$), then the variable x obeys the Duffing equation, which has no self-oscillatory solution. Self-oscillations exist only for modest values of γ, when the feedback loop is sufficiently inertial.

Examples of self-oscillatory systems with inertial excitation include the Lorenz system, the Helmholtz resonator with non-uniformly heated walls, a heated wire with a weight at its center, the Vallis model for nonlinear interaction between ocean and atmosphere, a modified Brusselator, an air-cushioned body, and a lumped model of the "singing" flame (Landa, 1996, 2001). Note that the mechanism of self-excitation of oscillations in some continuous system is similar to the examples considered above. Among these are the Rijke tube and a distributed model of the "singing" flame (Landa, 1996).

It must be emphasized that feedback can be nonlinear as well as linear. Although nonlinear feedback cannot cause self-excitation of oscillations, it is of primary importance in their development, and many examples are known. In addition to the swing, noted above, nonlinear feedback plays a role in population dynamics, and it is related to the development of turbulence in a subsonic submerged jet, where the feedback is realized via an acoustic wave. In general, positive feedback plays a key role in the emergence of coherent structures (Landa, 1996).

POLINA LANDA

See also **Candle; Flame front; Nerve impulses; Population dynamics; Tacoma Narrows Bridge collapse**

Further Reading

Babitzky, V.I. & Landa, P.S. 1982. Self-excited vibrations in systems with inertial excitation. *Soviet Physics, Doklady*, 27: 826–827

Babitzky, V.I. & Landa, P.S. 1984. Auto-oscillation systems with inertial self-excitation. *Zeitschrift fuer Angewandte Mathematik und Mechanik*, 64: 329–339

Bode, H.W. 1950. *Network Analysis and Feedback Amplifier Design*, New York: Nostrand

Brittain, J.E. & Black, 1997. Black and the negative feedback amplifier. *Proceedings of the IEEE*, 85: 1335–1336

Landa, P.S. 1996. *Nonlinear Oscillations and Waves in Dynamical Systems*, Dordrecht and Boston: Kluwer

Landa, P.S. 2001. *Regular and Chaotic Oscillations*, Berlin and New York: Springer

Neimark, Yu.I., Kogan, N.Ya. & Savel'ev, V.P. 1985. *Dinamicheskie Modeli Teorii Upravleniya* [*Dynamical Models of the Control Theory*], Moscow: Science

FEIGENBAUM THEORY

See **Routes to chaos**

FERMI ACCELERATION AND FERMI MAP

In a seminal paper, Enrico Fermi (1949) proposed two related methods of producing cosmic rays by accelerating charged particles by repeated collisions with moving magnetic fields. The particles could either be trapped by magnetic mirroring (*See* **Averaging methods**) or be deflected without trapping. In either case, the charged particles gain energy if the magnetic field is moving opposite to the particle's motion and lose energy if the motions are in the same direction. For random motions of the magnetic fields the net effect is stochastic acceleration; that is, the probability is higher for an accelerating collision, due to the relative velocities. The mechanism could explain the power law of proton energies and was consistent with the magnitude of the cosmic ray energies. However, because of the competition between energy loss due to ionization and energy gain from the stochastic heating, the process could not reasonably explain the existence of high mass cosmic rays. More recently, after the discovery of supernovas, and observing magnetic fields and motions of supernova remnants (SNRs), the trapping version of Fermi's acceleration mechanism has been closely examined. A general understanding, developed over the last few years, is that shock waves in SNRs can repeatedly accelerate trapped charged particles to produce both proton and higher mass cosmic rays (see a recent account by Malkov et al. (2002), and references therein).

Well before the recent interest in cosmic ray production by SNRs, the original basic idea of Fermi acceleration was being investigated with simple models. In particular, the question was asked whether the nonlinear dynamics of the particle motion would lead to stochastic acceleration even if the particles were acted on by periodic forces. A simple mapping model to detect this effect is that of a ball bouncing between a fixed and an oscillating wall. The model, developed by Ulam in 1961, was very straightforward to implement on a computer. The results were explained analytically and confirmed numerically in work during the 1960s and early 1970s by, among others, Zaslavsky and Chirikov, Brahic, and Lieberman and Lichtenberg (see Lichtenberg & Lieberman, 1991). They showed that for smooth forcing functions, the phase plane divides up into three distinct regions with increasing ball velocity: (1) a low-velocity region in which all period 1 fixed points are unstable, leading to stochastic motion over almost the entire region; (2) an intermediate velocity region in which islands of stability, surrounding elliptic fixed points, are imbedded in a stochastic sea; and (3) a higher-velocity region in which bands of stochastic motion, near the separatrices of the island trajectories joining the hyperbolic fixed points, are isolated from each other by regular orbits. The existence of region (3), in which invariant curves span the entire range of phase, bounds the energy gain of the particle. If the forcing function is not sufficiently smooth, then region (3) does not exist, in agreement with the Kolmogorov–Arnol'd–Moser (KAM) theory.

Because the Fermi particle acceleration mechanism was one of the first considered for determining the regions of parameter space where KAM surfaces exist and is easily approximated by simple mappings for which numerical solutions are attainable for "long times," it has become a bellweather problem in understanding the dynamics of nonlinear Hamiltonian systems with the equivalent of two degrees of freedom. Various versions both with analytic and non-analytic wall oscillation functions, and with physically moving walls or walls that just impart momentum, have been analyzed. The basic Ulam model, together with an interesting variant of it, is shown in Figure 1. The model in (a) is homologous to many physical problems (see below) and has therefore been of considerable importance. A simplified version of the model in (b) leads to the *standard map*.

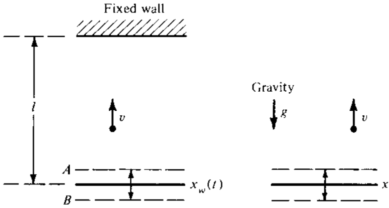

Figure 1. Fermi acceleration models. (a) Version in which a particle bounces between an oscillating wall and a fixed wall. (b) Version in which the particle returns to an oscillating wall under the action of a constant gravitational acceleration.

A simplified form of the Ulam map can be constructed if the oscillating wall imparts momentum to the ball, according to the wall velocity, without the wall changing its position in space. The problem defined in this manner has many of the features of the more physical problem and can be analytically treated with various wall-forcing functions. In this simplified form, the mapping is

$$
\begin{aligned}
u_{n+1} &= |u_n + f(\psi_n)|, \\
\psi_{n+1} &= \psi_n + \frac{2\pi M}{u_{n+1}},
\end{aligned}
\tag{1}
$$

where u is the velocity, normalized to the maximum wall velocity, and ψ is the phase of the oscillating wall at impact, with $2\pi M/u$ a conventionally used form for phase advance. Equation (1) is the product of two involutions and is, therefore area, preserving; that is, it satisfies the Jacobian condition

$$
\frac{\partial(u_{n+1}, \psi_{n+1})}{\partial(u_n, \psi_n)} = 1. \tag{2}
$$

The mapping in (1) serves as an approximation (with suitably defined variables) to many physical systems in which the transit time between kicks is inversely proportional to a velocity. The absolute-value signs correspond to the velocity reversal, at low velocities, $u < 1$ but have no effect on the region $u > 1$, which is the primary region of interest. The forcing function $f(\psi)$ is often a sinusoid in physical problems, but it may have other forms.

The basic method to analyze mapping models is to expand about a fixed point $u_{n+1} = u_n$, $\psi_{n+1} = \psi_n + 2\pi n$, and examine the linear stability. Alternatively, a local linearization of the phase advance equation about the phase-stable value of u linearizes the phase-advance equation but retains the nonlinearity in the forcing term. This expansion leads to the standard map, whose nonlinear stability properties have been extensively analyzed. These procedures have determined the basic phase-plane motion, as described above. A difference between the exact and simplified mappings, which should be noted, is that the canonical variables are different, leading to different variables in which an invariant distribution is uniform. For the simplified problem, a proper canonical set of variables is the ball velocity and phase just before the nth impact with the moving wall. The normalized velocity u then has a uniform invariant distribution. A detailed theoretical and numerical study of various forms of the Fermi map can be found in Lichtenberg & Lieberman (1991, Chapters 3 and 4), where many of the original references can be found. The Fermi map together with the standard map has also led to extensive analysis of phase space diffusion and correlations, which can also be found in the above reference, Chapter 5.

Various elaborated forms of the Fermi map have been applied to a variety of problems. Two such mapping models are electron cyclotron resonance heating (ECRH) in magnetic mirrors (Lieberman & Lichtenberg, 1973) and radio-frequency heating in capacitively driven discharges (Goedde et al., 1988). Although these physical models are generally more complicated than that given in (1), much of the basic analysis is similar. In particular, the expansion around fixed points can be made to investigate linear stability, and nonlinear expansions can obtain the standard mapping from which the KAM barriers to heating can be obtained. Weak dissipation can also be included in models, but this results in a contracting phase space not discussed here (*See* **Chaotic Dynamics**).

Another variant of the basic mapping is one for which the sign in the phase advance equation changes at some value of the action. Such mappings naturally occur in the dynamics of particles in circular accelerators where the sign transition occurs with increasingly relativistic motion in two frequency cyclotron heating and is an underlying cause of the period 3 "catastrophe" of the standard map. When the phase advance equation changes sign, the stable and unstable fixed points exchange their ψ-values, and the phase space structure between these fixed points undergoes a change in topology called reconnection. For further discussion of this phenomenon and references to the original literature, see Lichtenberg & Lieberman (1991, Sections 3.3 and 5.5).

Allan J. Lichtenberg

See also **Averaging methods; Chaotic dynamics; Kolmogorov–Arnol'd–Moser theorem; Maps; Phase space diffusion and correlations; Standard map**

Further Reading

Fermi, E. 1949. On the origin of cosmic radiation. *Physical Review*, 75: 1169–1174

Goedde, C.G., Lichtenberg, A.J. & Lieberman, M.A. 1988. Self-consistent stochastic electron heating in radio frequency discharges. *Journal of Applied Physics*, 64: 4375–4383

Lichtenberg, A.J. & Lieberman, M.A. 1991. *Regular and Chaotic Dynamics*, 2nd edition, New York: Springer

Lieberman, M.A. & Lichtenberg, A.J. 1973. Theory of electron cyclotron resonance heating—II. Long-time and stochastic effects. *Plasma Physics*, 15: 125–150

Malkov, M.A., Diamond, P. & Jones, T.W. 2002. On the possible reason for non-detection of TEV protons in SNRs. In *Bifurcation Phenomena in Plasmas*, edited by S.I. Itoh & Y. Kawai, Fukuoka: Kyushu University

FERMI RESONANCE

See **Harmonic generation**

FERMI–PASTA–ULAM OSCILLATOR CHAIN

A closely examined model of a high-dimensional system comprises a number of masses constrained to move along a line, with a specified force law governing their interaction. In the early 1950s, Enrico Fermi, John Pasta, and Stan Ulam (FPU) (Fermi et al., 1955) numerically examined such a chain of coupled oscillators, with a Hamiltonian of the general form

$$H(p,q) = \sum_{k=1}^{N}\left[\frac{1}{2}p_k^2 + \frac{1}{2}(q_{k+1}-q_k)^2 + \frac{1}{3}\alpha(q_{k+1}-q_k)^3 + \frac{1}{4}\beta(q_{k+1}-qk)^4\right], \tag{1}$$

where each of N particles has unit mass and unit harmonic coupling constant and the end points are either fixed (vertical position $q_1 = q_{N+1} = 0$) or periodic ($q_{N+1} = q_1$). The parameters α, β represent the nonlinearity in the forces between the particles in the chain. In the original FPU numerical computations either α or β was set equal to zero, and fixed end points were used to correspond to a discretized nonlinear spring. Most subsequent work was done with $\alpha = 0$ and is known as the FPU-β model. The FPU-β model can be normalized so that the significant parameter is βE, where E is the total energy, $H(p,q) = E$.

The original idea of Fermi was that the nonlinearity of the springs would lead to mode mixing such that thermodynamic properties in the lattice could be studied. The original numerical work placed the initial energy in the lowest mode of the linearized system, which at low energy becomes a quasimode of the nonlinear system. The N independent normal modes Q_k of the harmonic system with fixed end point boundary conditions give

$$q_i = \sqrt{\frac{2}{N+1}}\sum_{k=1}^{N} Q_k \sin\left(\frac{ik\pi}{N+1}\right), \tag{2}$$

for which the amplitudes q_i of the mass points form a half-sine for the lowest mode and pick up an additional zero for each higher mode. The harmonic mode frequencies are

$$\omega_k = 2\sin\left(\frac{\pi k}{2N+2}\right). \tag{3}$$

The original numerical study surprisingly showed recurrences, rather than a thermodynamic spreading of energy among the modes. Subsequent studies explained these recurrences as resulting from near resonances among various combinations of modes (Bivins et al., 1973). However, the question of whether the nonlinearity led to long-term mixing of the energy remained an open question.

Although the original question of whether the nonlinearity led to equipartition among the degrees of freedom went unanswered during the 1960s, the FPU-β system became rightly famous for having inspired the early development of soliton theory. Zabusky & Kruskal (1965) showed that a Taylor expansion to transform the FPU chain into a continuous differential equation resulted in the Korteweg–de Vries (KdV) or the modified Korteweg–de Vries (mKdV) equation, depending on whether the α or β term in (1) was retained. For the FPU-β chain (mKdV equation), we have

$$u_\tau + 12u^2 u_\xi + u_{\xi\xi\xi} = 0, \qquad (4)$$

where time and length variables have been rescaled by $\tau = h^3 t/24$ and $\xi = x - ht$; subscripts τ, ξ denote differentiation with respect to that variable.

The periodic solutions of (4), stationary in the frame $\xi - c\tau$, may be obtained by integrating to give

$$\tfrac{1}{2}u_\xi^2 + u^4 - \tfrac{1}{2}cu^2 + A \equiv \tfrac{1}{2}u_\xi^2 + P(u) = 0, \qquad (5)$$

where A is a constant of integration. Equation (5) is in the form of a one-degree-of-freedom Hamiltonian, which is therefore integrable. Solutions have been obtained in terms of the Jacobi elliptic functions (or cnoidal waves) $\mathrm{cn}(\xi, q)$, with q (the modulus) taken as a parameter with $0 \le q^2 < 1$. This formalism led to the observations of the stability properties of solitons, and more generally motivated the development of the inverse scattering technique (Gardner et al., 1967). However, the explanation that long-time recurrences were due to initial conditions that were superpositions of stable solitons was not a complete picture and was ultimately misleading.

An explanation for the mechanism leading to stochastic diffusion of energy over the $2N$-dimensional phase space, which also predicted a transition from regular to stochastic motion with increasing energy, was put forward by Izrailev & Chirikov (1966). Using the transformation to modes as in (2), they postulated that if the interaction of pairs of neighboring modes was sufficient to create overlapping beat modes, then the overlap would create global stochasticity. The concept that had been developed by Chirikov for low degree-of-freedom Hamiltonian systems, known as the Chirikov overlap criterion, was applied to the high-dimensional system by isolating a few modes containing the energy. The criterion had been confirmed numerically and later studied in great detail using the *standard map*. The criterion can be shown to be roughly equivalent to requiring that the nonlinear frequency shift with the energy in a single mode be equal to the mode separation (Lichtenberg & Lieberman, 1991, Section 6.5b). Substituting (2) into (1), the nonlinear shift in mode frequency of mode k can be calculated approximately as $\Delta\omega_k \simeq 3\beta E_k \omega_k/4N$. From (3), the frequency separation between low-frequency modes with $\omega_k \simeq \pi k/N$, $k \ll N$, is $\delta\omega_k \simeq \pi/N$, such that for overlap ($\Delta\omega_k/\delta\omega_k \ge 1$)

$$\beta E_k/N \ge 4/3k. \qquad (6)$$

Stochastic energy transfer among modes, leading to approximate equipartition among modes, has been numerically found at much lower values of energy density than given by (6). Numerical studies of equipartition (e.g., Livi et al. (1985); Pettini & Landolphi, 1990) found that (6) roughly corresponds to a transition between an inverse time to obtain some measure of equipartition that scales as $\tau^{-1} \propto (E/N)^2$ at low energy, to a time scaling as $\tau^{-1} \propto (E/N)^{2/3}$ at high energy. The latter scaling models a random process that is very strongly stochastic. Furthermore, the prediction that mode overlap determines the transition between regular motion and stochastic diffusion leading to equipartition does not hold if the energy is initially placed in one or a few high-frequency modes. There is local mode mixing, but the high-frequency modes interact only weakly with the low frequencies, as theoretically predicted by Benettin et al. (1987), such that equipartition is not observed above the mode overlap transition on a fast time scale.

The mKdV soliton was found to become unstable from a low-frequency mode, leading to exponential growth of higher frequencies, which correspond quite closely to similar growth of higher frequencies in an oscillator chain (Driscoll & O'Neil, 1976). However, this low-frequency instability is neither a necessary nor sufficient condition for a transition to equipartition, as the soliton theory does not describe the high-frequency modes, but it does present a physical picture of the process that can hold a low-frequency mode together in the absence of coupling to high-frequency modes and how it can break down.

A more coherent picture of the underlying processes leading to equipartition has been developed recently. Basic phase space arguments indicate that, for relatively high-dimensional systems with not too small a perturbation strength, the probability will be high that a generic initial condition will lie in a stochastic portion of the phase space. Arnol'd diffusion will then transport energy over most of the degrees of freedom, essentially leading to equipartition, on some time scale. However, the time to equipartition can be exponentially slow.

Focusing on the region of weaker stochasticity, a transition was numerically found between power law and exponentially long time scales as the energy density $\varepsilon = E/N$ of the system is decreased. This latter transition is of prime importance for the observation of equipartition, as it essentially separates observable times from those that are not observable. For the FPU chain, the main mechanisms leading to equipartition in this lower energy regime are that resonant interaction of some set of low-frequency modes, in which a significant portion of the energy resides, can lead to local superperiod (very low frequency) beat oscillations that are stochastic. The beat oscillations, which increase with energy, can become comparable to frequency differences between high-frequency modes. This results in Arnol'd diffusion transferring energy to high-frequency modes, on a power law time scale (DeLuca et al., 1995). Furthermore, the appearance of stochasticity corresponds to the onset of the mKdV instability. The driving frequency for diffusion is associated with the libration frequency of the resonance, Ω_B. Using resonant normal form perturbation theory to isolate the most important coupling to the high-frequency modes, the energy transfer to high-frequency modes by Arnol'd diffusion, depends exponentially on the frequency ratio as

$$\frac{\mathrm{d}E}{\mathrm{d}t} \propto \exp(-\pi\delta\omega_h/2\Omega_\mathrm{B}), \tag{7}$$

where $\delta\omega_h$ is the difference frequency between two high-frequency modes. When $\Omega_\mathrm{B} \sim \delta\omega_h$, the exponential factor is of order unity, allowing strong diffusion of energy to high-frequency modes, and equipartition on computationally observable time scales. In a separate work, an estimate of the scaling, with energy density, of the equipartition time for $E \gg E_c$ was found theoretically to be $T_\mathrm{eq} \propto (N/E)^3$, which agrees with numerical computations (De Luca et al., 1999). The somewhat weaker quadratic scaling found in earlier work, as quoted above, was probably due to the use of a less accurate measure of equipartition.

If the energy is initially placed in high-frequency modes, the equipartition process is significantly different from that starting from low-frequency initial conditions. In this case, the dynamics is transiently mediated by the formation of unstable nonlinear structures. First, there is an initial fast stage in which the mode breaks up into a number of breather-like structures. Second, on a slower time scale, these structures coalesce into one large unstable structure. These structures have been called chaotic breathers (CB). Because a single large CB closely approximates a stable breather, the final decay stage, toward equipartition, can be very slow. This behavior has been observed in oscillator chains approximating the Klein–Gordon equation with various force laws and the FPU-β model (Cretegny et al., 1998; Mirnov et al., 2001). In Cretegny et al. (1998), the energy was placed in the highest frequency mode with strict alternation of the amplitudes from one oscillator to the next (periodic boundary conditions). This configuration is stable up to a particular energy, for which there exist exact discrete solutions. Beyond this energy a parametric instability occurs, leading to the events described above. However, the nonlinear evolution does not depend on special initial conditions but will generically evolve from any high-frequency mode initial condition that has predominantly the alternating amplitude symmetry. One does not know, in this generic situation, whether there exists any true energy threshold. As discussed with respect to low-frequency mode initial conditions, the practical thresholds refer to observable time scales. From a phase space perspective, it is intuitively reasonable that for a large number of oscillators and not too low an initial energy, the generic set of initial conditions will lie in a chaotic layer, but the chaotic motion can remain close to a regular orbit for very long times (Lichtenberg & Lieberman, 1991).

Considerable insight into the behavior of a nonlinear oscillator chain, starting from high-frequency mode initial conditions, can be obtained by introducing an envelope function for the displacements of the oscillators. Low-order expansions produce PDEs that have integrable solutions in the form of envelope solitons, analogous to the solitons produced from low-frequency initial conditions (Kosevich, 1993). Higher-order terms usually destroy the integrability.

Substituting the envelope function $\psi_i(t) = (-1)^i q_i(t)$ in (1) (with $\alpha = 0$), using the continuous variable $x = ai$, a Taylor's expansion in a, the lattice period, yields

$$\begin{aligned}
&\psi_{tt} + 4\psi + 16\beta\psi^3 \\
&\quad + a^2\{\psi_{xx} + \beta(12\psi\psi_x^2 + 12\psi^2\psi_{xx})\} \\
&\quad + a^4\{(1/12)\psi_{xxxx} + \beta(3\psi_x^2\psi_{xx} + 3\psi\psi_{xx}^2 \\
&\quad + 4\psi\psi_x\psi_{xxx} + \psi^2\psi_{xxxx})\} + \cdots = 0.
\end{aligned} \tag{8}$$

Comparing (8) with (4) qualitatively explains why relaxation is accompanied by the formation of sharply localized states if energy is initially deposited in the high-frequency part of the spectrum where the effect of dispersion is small, while only broad nonlinear structures are formed if the energy is initially in the low-frequency modes where the dispersion is large. Using a dimensionless variable, introducing the rotating wave approximation (RWA) $\cos^3 \omega t \simeq (\frac{3}{4}) \cos \omega t$ and neglecting terms proportional to a^4 and higher, (8) can be integrated to yield

$$(-\omega^2 + 4)\psi^2 + \psi_x^2 + \beta(6\psi^4 + 9\psi^2\psi_x^2) = C_1, \tag{9}$$

which has integrable trajectories in phase space similar to (5) but includes a high-frequency drive ω. Depending on the boundary conditions, these phase trajectories represent multiple or single breathers.

As with low-frequency initial conditions, the soliton solution of the envelope, obtained from (9), becomes unstable with increasing energy. For most numerical studies of oscillator chains, the initial state imposed on the system is mainly that of a single linear mode. This state is generally not close to an equilibrium. The initial state rapidly relaxes, governed by the nonlinear equations. Numerical studies indicate that the chaotic breathers that are formed in the nonlinear processes are probably marginally stable, which accounts for their long-lived existence.

After a set of chaotic breathers has been formed, on a short time scale by a modulational instability or breakup relaxation, the breathers coalesce, on a longer time scale into a single chaotic breather. This process has been well documented numerically (Cretegny et al., 1998; Mirnov et al., 2001) and an analytic estimate of the process made (Kosevich & Lepri, 2000; Mirnov et al., 2001) with the breather coalescence time found to scale as $\tau_{\mathrm{B}}^{-1} \propto E_{\mathrm{B}}/N$, the breather energy density. The background mode spectrum beats with the breather to transfer energy to low-frequency modes, resulting in equipartition on a slower time-scale as $T_{\mathrm{eq}} \propto (E/N)^{-2}$ (Mirnov et al., 2001).

Although we have concentrated on the seminal problem of the FPU-β lattice, the results are qualitatively connected to other types of oscillator chains. For example, if the cubic, rather than the quartic, potential is retained in (1), then the Hamiltonian is equivalent to the lowest-order nonlinear expansion of the integrable Toda lattice. Consequently, the chain is considerably more stable. Other interesting types of lattices are composed of discretizations of Klein–Gordon equations. In particular, with a quartic potential, the dynamics of the Klein–Gordon Hamiltonian

$$H = \sum_{i=1}^{N} \left[\tfrac{1}{2} p_i^2 + \tfrac{1}{2}(q_{i+1} - q_i)^2 + \tfrac{1}{2} m^2 q_i^2 + \tfrac{1}{4} \beta q_i^4 \right], \tag{10}$$

has been compared with the FPU-β chain, showing both similarities and differences (Pettini & Cerruti-Sola, 1991). The stability of "discrete breathers" has also been examined for various chains (Cretegny et al., 1998).

Allan J. Lichtenberg

See also **Arnol'd diffusion; Breathers; Discrete breathers; Frenkel–Kontorova model; Korteweg–de Vries equation; Phase space diffusion and correlations; Solitons; Solitons, a brief history; Standard map; Toda lattice**

Further Reading

Benettin, G., Galgani, L. & Giorgilli, A. 1987. Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory. *Communications in Mathematical Physics*, 113: 87–103

Bivins, R.L., Metropolis, N. & Pasta, J.R. 1973. Nonlinear coupled oscillators: model equation approach. *Journal of Computational Physics*, 12: 65–87

Cretegny, T., Dauxois, T., Ruffo, S. & Torcini, A. 1998. Localization and equipartition of energy in the β-FPU chain: chaotic breathers. *Physica* D, 121: 109–126

De Luca, J., Lichtenberg, A.J. & Lieberman, M.A. 1995. Time scale to ergodicity in the Fermi–Pasta–Ulam system. *Chaos*, 5: 283–297

De Luca, J., Lichtenberg, A.J. & Ruffo, S. 1999. Finite times to equipartition in the thermodynamic limit. *Physical Review* E, 60: 3781–3786

Driscoll, C.F. & O'Neil, T.M. 1976. Explanation of instabilities observed on the Fermi–Pasta–Ulam lattice. *Physical Review Letters*, 37: 69–72

Fermi, E., Pasta, J.R. & Ulam, S.M. 1955. Studies of nonlinear problems. *Los Alamos Scientific Report*, LA-1940; also in *Collected Works of Enrico Fermi* 1965. Chicago: University of Chicago Press, 2: 978–988

Gardner, C.S., Greene, J.M., Kruskal, M.D. & Mura, R.M. 1967. Method for solving the Korteweg–de Vries equation. *Physical Review Letters*, 19: 1095–1097

Izrailev, F.M. & Chirikov, B.V. 1966. Statistical properties of the nonlinear string. *Soviet Physics-Doklady*, 11: 30–32

Kosevich, Y.A. 1993. Nonlinear envelope-function equation and strongly localized vibrational modes in anharmonic lattices. *Physical Review* B, 47: 3138–3151

Kosevich, Y.A. & Lepri, S. 2000. On modulation instability and energy localization in harmonic lattices at finite energy density. *Physical Review* B, 61: 299–314

Lichtenberg, A.J. & Lieberman, M.A. 1991. *Regular and Chaotic Dynamics*, 2nd edition, New York: Springer

Livi, R., Pettini, M., Ruffo, S., Sparglioni, M. & Vulpiani, A. 1985. Equipartition threshold in nonlinear large Hamiltonian systems: the Fermi–Pasta–Ulam model. *Physical Review* A, 31: 1039–1045

Mirnov, V.V., Lichtenberg, A.J. & Guclu, H. 2001. Chaotic breather formation, coalescence, and evolution to energy equipartition in an oscillator chain. *Physica* D, 157: 251–282

Pettini, M. & Cerruti-Sola, M. 1991. Strong stochasticity threshold in nonlinear large Hamiltonian systems: effect on mixing times. *Physical Review* A, 44: 975–987

Pettini, M. & Landolphi, M. 1990. Relaxation and ergodicity breaking in nonlinear Hamiltonian dynamics. *Physical Review* A, 41: 768–783

Zabusky, N.J. & Kruskal, M.D. 1965. Interactions of solitons in a collisionless plasma, and the recurrence of initial states. *Physical Review Letters*, 15: 240–243

FERROMAGNETISM AND FERROELECTRICITY

Iron, nickel, cobalt, and some rare earths (e.g., gadolinium) are characterized by long-range ferromagnetic order. This originates at the atomic level and causes the unpaired electron spins to line up parallel to each other within a region of space called a magnetic domain. Domains range from 0.1 mm to a few mm in size. Within each domain, the net magnetization is large and homogenous, but over the entire sample it averages out to zero due to random orientations of spins. An externally applied magnetic field can cause the material to become macroscopically magnetized, as the magnetic domains already aligned in the direction of this field grow at the expense of their neighbors and those neighbors reorient their magnetizations towards the field direction.

Ferromagnets have very high magnetic susceptibilities (χ), ranging from 1000 up to 100,000 (meaning the ferromagnet is that much easier to magnetize than free space), and they tend to stay magnetized following the application of an external magnetic field. This tendency to remember magnetic history is called "hysteresis." All ferromagnets have a maximum temperature, called the Curie temperature (T_c), which for iron is about 1043 K. Above T_c the ferromagnetic phase changes into a paramagnetic phase, in which induced magnetism is proportional to the applied field. Ferromagnets also respond mechanically to an applied magnetic field, changing length slightly in the direction of the applied field, a property that is called magnetostriction.

In addition to ferromagnets, there exist other magnetically ordered compounds with parallel-oriented sublattices. The simplest such example is antiferromagnetism with two antiparallel sublattice magnetizations. To measure the degree of order in a complex magnetic phase, as many order-parameter components as there are distinct sublattices may be needed. For ferromagnets, the order parameter is the net magnetization. In antiferromagnets, the order parameter is the staggered magnetization: $\boldsymbol{M}_1 - \boldsymbol{M}_2$, where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are the magnetization vectors for the two sublattices.

In 1907, Pierre Weiss proposed a phenomenological theory of ferromagnetism, building on the model of paramagnetism that Paul Langevin introduced in 1905. The Langevin function: $L = \coth(x) - 1/x$ (where $x = \mu H/kT$) describes the paramagnetic susceptibility of N non-interacting classical spins in a magnetic field of intensity H. Weiss assumed that spins interact with each other through a molecular field proportional to the average magnetization in the sample. So that

$$H_{\text{eff}} = H + \lambda M. \tag{1}$$

By replacing spin-spin interactions with the interaction of a single spin (S) and all its neighbors taken as an average field, the nonlinear problem was approximately solved giving rise to ferromagnetism for T below $T_c = \lambda C$, where

$$C = \frac{N\mu^2 S(S+1)}{3k} \tag{2}$$

with k denoting the Boltzmann constant.

As a consequence, the Langevin expression for susceptibility in paramagnetism

$$\chi = \frac{M}{H} = \frac{C}{T} \tag{3}$$

becomes the Curie–Weiss relation

$$\chi = \frac{C}{T - T_c} \tag{4}$$

above T_c, with spontaneous magnetization below T_c for ferromagnets. This is characteristic of a second-order phase transition. Spin alignment stems from the exchange interactions between spins whose energy can be expressed via the Ising Hamiltonian $\mathcal{H}$ in the case of strong uniaxial anisotropy favoring alignment parallel or antiparallel to the z-axis of quantization depending on the sign of the exchange constant J:

$$\mathcal{H} = -\sum_{i,j} J S_i^z S_j^z, \tag{5}$$

where S_i^z and S_j^z are the z-components of the spin vectors at the lattice sites i and j, respectively. In the absence of anisotropy, one uses the Heisenberg Hamiltonian that couples the spin vectors in a scalar product. Free-energy expansion of the Landau type can be obtained within the Curie–Weiss approximation for a Hamiltonian that includes the Zeeman interaction with an external magnetic field and an Ising-type spin-spin interaction.

Ferroelectricity was discovered in the beginning of the 20th century as a property of ionic, covalent, molecular crystals, and even polymers that possess electrical polarization either spontaneously (e.g., Rochelle salt) or under mechanical stress (piezoelectricity) or temperature changes (pyroelectricty). The net polarization of a ferroelectric crystal can be reoriented by applying an electric field. In ferroelectric phase transitions, a change in the crystal structure is accompanied by the appearance of spontaneous polarization. Ferroelectric phase transitions can be either displacive or order-disorder type. In displacive transitions (e.g., $BaTiO_3$), atoms or molecules exhibit small (compared with the unit cell) positional shifts with long-range correlations. These transitions are caused by phonons and the order parameter is the amplitude of the related lattice distortion giving rise to a change of the lattice structure. Displacive transitions are described using a continuous Landau–Ginzburg model with ensuing solitary waves. In order-disorder transitions (e.g., $NaNO_2$), atoms or molecules order themselves on distances comparable to the unit cell. A transformation between randomly distributed atomic positions of their local double-well potential bottoms ($T > T_c$) and an ordered arrangement ($T < T_c$) takes place. Models of order-disorder transitions use the Ising Hamiltonian with an effective (not real) spin variable.

Ferroelectric phase transitions involve symmetry changes in the crystal structure that are manifested by the emergence of an order parameter: spontaneous polarization vector $\boldsymbol{P}$. For second-order transitions, the symmetry group of the ferroelectric phase is a subgroup of that in the paraelectric phase (in which $\boldsymbol{P}$ is proportional to the applied electric field $\boldsymbol{E}$). In some cases, such as the onset of ferroelectricity with a transverse optical branch, a so-called soft mode is responsible for the transition. The soft-mode's frequency ω_k for the wave vector k tends to 0 as $T \to T_c$. A special type of ferroelectric phase transition involves incommensurate phases where spontaneous polarization develops a spatial modulation with a wavelength that is incommensurate with lattice periodicity. The occurrence of incommensurate phases is usually explained by competition between long- and short-range forces, for example, in the Frenkel–Kontorova model. As in ferromagnets, ferroelectrics develop domains in which a particular orientation of polarization is selected. These domains can range in sizes from submicroscopic to macroscopic, and the region between two neighboring domains is called a domain wall.

JACK A. TUSZYŃSKI

See also **Critical phenomena; Domain walls; Frenkel–Kontorova model; Hysteresis; Ising model; Order parameters**

Further Reading

Bruce, A.D. & Cowley, R.A. 1981. *Structural Phase Transitions*, London: Taylor & Francis

Kittel, C. 1996. *Introduction to Solid State Physics*, 7th edition, New York: Wiley

Landau, L.D. & Lifshitz, E.M. 1959. *Statistical Physics*, London: Pergamon Press

Lines, M.E. & Glass, A.M. 1977. *Principles and Applications of Ferroelectric and Related Materials*, Oxford: Clarendon Press

Stanley, H.E. 1971. *Introduction to Phase Transitions and Critical Phenomena*, Oxford: Clarendon Press and New York: Oxford University Press

White, R.H. & Geballe, T. 1979. *Long Range Order in Solids*, New York: Academic Press

FIBERS, OPTICAL

See **Optical fiber communications**

FIBONACCI SERIES

Leonardo of Pisa (approx. 1175 to around 1250), also known as Leonardo Pisano, referred to himself as Fibonacci (or son of Bonacci), the name by which he is usually called today. As the son of a customs officer, he had opportunities to travel around the Mediterranean coast and observe many commercial practices. He saw that the Hindu-Arabic system of ten digits and its algorithms for arithmetic had many advantages over Roman numerals. His book of 1202 (revised 1228), *Liber Abaci*, meaning *The Book of the Abacus* or *The Book of Reckoning*, described the system in the Italian vernacular, and so the decimal system became common in Europe.

In his book are some problems and puzzles that were meant as arithmetic practice using the new system. One referred to a problem about rabbits in a field: "A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair, which from the second month on becomes productive?" It is based on the assumption that the pairs mate according to the same conditions. Each month, the number of rabbit pairs in the field is 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144. So the answer is 144 pairs after 12 months. The number of rabbits in any month is found by adding the number of rabbits alive in the last month (since we assume none die or are eaten) and adding to it the number of new rabbits born that month, which is one for each pair that was alive in the month before that. The rule is therefore "add the last two numbers to get the next." The next number in the series is 89 plus 144, or 255. The series was probably known before this and Fibonacci merely copied it from another source. The sequence is now called the Fibonacci sequence in his honor, but this name was not given until the late 1800s when Edouard Lucas rediscovered the series and wrote about some of its many properties.

Earlier than this, it had been noticed that the Fibonacci numbers (though not called that) appear in the number of petals of flowers of many plant species, in the arrangements of leaves round a branch, or seed whorls on a seed head. The study of such features of plants is termed phyllotaxis. There are flowers with 3 petals (clover) and many with 5, but none with 7 or 9 leaves or petals. The numbers 4, 6, and 10 also occur but the arrangement is of two sets of 2, 3, or 5. No scientific justification for this was given until Douady and Couder wrote about it in 1993 where they showed that if the growing tip of a plant (the meristem) produces a new primordial cell that becomes a leaf or petal or branch, then the optimal arrangement is for the new cell to be produced at the rate of 1.618 per turn of the growing point. This produces the least overlapping for leaves and the maximum exposure to collect sunlight, or for seeds, it provides the most compact packing (given the simple method of growth of the meristem) with even placing of seeds no matter how large the seed head becomes.

This number, precisely computed as $g \equiv (\sqrt{5}+1)/2 = 1.6180339887\ldots$, is a famous number in geometry called the golden section (also golden ratio, golden mean, divine proportion). It appears as the ratio of diagonals to sides in a pentagon. The golden section as a decimal number never recurs, and so cannot be expressed exactly as a fraction (i.e., it is irrational). It has the property that its reciprocal (1 divided by the golden section) is 0.6180339887…. This is exactly 1 less than the golden section itself, giving rise to the definition of the golden section as "the positive number that is 1 less than its reciprocal." In other words, $1/g = g - 1$ or $1 = g^2 - g$, a quadratic equation that is solved by the value above together with $-1/g = -0.6180339887\ldots$.

The fractions that best approximate g are the ratios of two successive Fibonacci numbers: 3/2, 5/3, 8/5, 13/8, 21/13 in the sense that no fraction with numbers smaller than the numerator of one of these terms will give a better approximation. It is this that explains why the Fibonacci numbers appear in phyllotaxis since the optimal arrangements are best approximated using Fibonacci numbers (or, indeed any series where the last two numbers are added to produce the next).

Since the Fibonacci rule is so simple, it is also found in other mathematical situations. For instance, suppose that houses can be made in two sizes, single (separate) and double (two attached), the latter taking twice the frontage along a road as the single ones. Given a road that is long enough for n houses on one side, how many arrangements are there that the architect can choose from? For instance, a road that can have 3 houses on it could have 3 singles, or else a double followed by a single or a single first and then a double—3 ways. For 4 houses there are 5 arrangements; for 5 houses, 8 possibilities. The number of arrangements is always a Fibonacci number.

The rule also applies to the family tree of a male honeybee. Male bees are produced from unfertilized eggs and female bees from fertilized eggs. So males have one parent and females have two. A male bee thus has one parent (F) and two grandparents (M and F), three great-grandparents and so on with 5, 8, 13, 21 as we go back each generation.

In two dimensions, a rectangle with sides in the golden ratio is called a golden rectangle. It has been observed in the shape of many parts of the Parthenon on the Acropolis in Athens, although none of the original plans of the building are extant. The Greeks knew of the ratio, and Euclid's *Elements* shows how to find and construct the golden section point on any line.

RON KNOTT

See also **Branching laws**

Further Reading

For further information, references, applications, and related mathematics, see www.mcs.surrey.ac.uk/Personal/R.Knott/Fibonacci/

Boncompagni, B. 1852. *Della vita e delle opere di Leonardo Pisano*, Rome: Tipografia delle Belle Arti (The only complete printed version of Fibonacci's 1228 edition of *Liber Abbaci*.)

Gies, J. & Gies, F. 1969. *Leonard of Pisa and the New Mathematics of the Middle Ages*, New York: Crowell (Another book with much on the background to Fibonacci's life and work.)

Grimm, R.E. 1973. The autobiography of Leonardo Pisano. *Fibonacci Quarterly*, 11: 99–104

Horadam, A.F. 1985. Eight hundred years young. *The Australian Mathematics Teacher*, 31: 123–134 (An interesting and readable article on Fibonacci, his names and origins as well as his mathematical works. It refers to and expands upon the article by Grimm.)

Parshall, Karen Hunger. *The Art of Algebra from al-Khwarizmi to Viète: A Study in the Natural Selection of Ideas*, http://www.lib.virginia.edu/science/parshall/algebra.html (Contains a brief biography of Fibonacci if you want to read more about the history of mathematics.)

Smith, D.E. 1923. *History of Mathematics*, vol. 1, Boston and London: Ginn; reprinted, New York: Dover, 1958 (Gives a complete list of other books that he wrote and is a fuller reference on Fibonacci's life and works.)

FIBRILLATION

See **Cardiac arrhythmias and electrocardiogram**

FILAMENTATION

When we look at a window on rainy days, we may observe a curious phenomenon: the rain forms a sheet of water on the glass. A single drop—from time to time—introduces a perturbation on this surface that is sufficient to trigger small-scale channels breaking the layer of liquid and escaping from it. Filamentary structures of water emerge from the unstable flow, destroying the initial expanse. The water sheet has thus decayed into filaments through an instability that is called filamentation.

Filamentation also occurs in plasma physics and nonlinear optics. In this context, an intense optical beam propagating in a focusing (Kerr) medium may break up into several spots due to the small inhomogeneities affecting its initial distribution, as evidenced by self-focusing experiments in solids, liquids, and gases (see, e.g., Bespalov & Talanov, 1966; Campillo et al., 1973). This phenomenon can be understood from the equation for the paraxial self-focusing of optical beams that describes the propagation of the slowly varying complex envelope ψ of a scalar electric field with central frequency ω_0 through a nonlinear medium. The envelope ψ, expressed in the frame moving with the group velocity, obeys the nonlinear Schrödinger (NLS) equation

$$\mathrm{i}\partial_z \psi + \nabla_\perp^2 \psi + f(|\psi|^2)\psi = 0, \qquad (1)$$

where $\nabla_\perp^2 \equiv \partial_x^2 + \partial_y^2$ represents wave diffraction and $f(|\psi|^2)$ models the nonlinear response of the medium. Here, the radial (x, y) and longitudinal (z) variables are normalized with respect to the beam waist w_0 and to the Rayleigh length $z_0 = \pi n_0 w_0^2/\lambda_0$, depending on the linear refractive index n_0 and central wavelength λ_0.

For an unsaturated Kerr medium, $f(s) = s$ and solutions of Equation (1) collapse at finite distance whenever their power $P = \int |\psi|^2 \mathrm{d}\boldsymbol{r}$ exceeds the threshold $P_\mathrm{c} \simeq 11.7$. Nonlinear responses of optical materials are, however, generally saturating, for example, like $f(s) = s/(1+\beta s)$ with $\beta = |\psi|_{\max}^{-2} \ll 1$. Such saturations occur in two-level systems, such as sodium and rubidium atomic vapors (Soto-Crespo et al., 1992).

Modeling filament formation requires a perturbation theory, following which small fluctuations break up steady-state solutions, expressed as $\psi_{\rm s}(\boldsymbol{r}, z) = \phi(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\lambda z}$, where ϕ satisfies the differential equation

$$-\lambda\phi + \boldsymbol{\nabla}_{\perp}^{2}\phi + f(\phi^2)\phi = 0 \tag{2}$$

and $\lambda =$ const. Stability of ϕ can thus be investigated from the perturbed solution

$$\begin{aligned}\psi(x, y, z) \;&=\; [\phi(x, y, \lambda) + \varepsilon\{v(x, y, z)\\ &\quad + \mathrm{i}w(x, y, z)\}]\mathrm{e}^{\mathrm{i}\lambda z},\end{aligned} \tag{3}$$

where v and w are real-valued functions with amplitude parameter $\varepsilon \ll 1$.

Linearizing Equation (1) with respect to these functions yields the eigenvalue problem

$$\partial_z v = L_0 w, \quad -\partial_z w = L_1 v, \tag{4}$$

where L_0 and L_1 represent the self-adjoint operators

$$\begin{aligned}L_0 \;&=\; \lambda - \boldsymbol{\nabla}_{\perp}^{2} - f(\phi^2),\\ L_1 \;&=\; \lambda - \boldsymbol{\nabla}_{\perp}^{2} - [f(\phi^2) + 2f'(\phi^2)\phi^2],\end{aligned} \tag{5}$$

with $f'(\phi^2) = \partial f(u)/\partial u|_{u=\phi^2}$. Combining Equations (4), we then obtain $\partial_z^2 v = -L_0 L_1 v$, and different filamentation-like instabilities may be investigated from this general formalism.

- *Modulational instability (MI):* Originally, Bespalov & Talanov (1966) proposed a modulational instability theory, following which oscillatory perturbations with an exponential growth rate split the beam envelope taken as a background uniform solution into small-scale cells. Perturbative modes are chosen as $v, w \sim \cos(k_x x)\cos(k_y y)\mathrm{e}^{\gamma z}$, and they apply to a plane wave ϕ which, by definition, satisfies $\boldsymbol{\nabla}_{\perp}^{2}\phi = 0$, so that $\lambda = f(\phi^2)$. The growth rate γ is then given by Equation (4):

  $$\begin{aligned}\gamma^2 \;&=\; k^2[2A - k^2], \quad A \equiv uf'(u)|_{u=\phi^2},\\ k^2 \;&=\; k_x^2 + k_y^2.\end{aligned} \tag{6}$$

  Plane waves are unstable with $\gamma^2 > 0$ in the range $0 < k < \sqrt{2A}$ and the maximum growth rate $\gamma_{\max} = A$ is attained for $k = k_{\max} = \sqrt{A}$. This instability promotes the beam breakup into a wave-train of small-scale filaments regularly distributed in the diffraction plane with the transversal spacing $\lambda_{\rm mod} \simeq 2\pi/k_{\max}$ and longitudinal length $\sim \gamma^{-1}$. In practical uses, the number of filaments is close to the ratio $P_{\rm in}/P_{\rm fil}$, where $P_{\rm in}$ is the power of the input beam and $P_{\rm fil}$ the power enclosed in one filament. Considering each filament with radial symmetry, the evaluation $P_{\rm fil} \simeq 2\pi\int_0^{\lambda_{\rm mod}/2} r|\phi|^2\,\mathrm{d}r \simeq 2.65P_{\rm c}$ holds for unsaturated Kerr media [$f(s) = s$].
- *Filamentation on a ring*: For broad beams, self-focusing often takes place as a regular distribution of dots superimposed upon ring-like diffraction patterns, so that filamentation may not develop over the entire surface of the input beam. To model this instability, the Laplacian in Equation (1) can be rewritten as $\boldsymbol{\nabla}_{\perp}^{2} = r^{-1}\partial_r r\partial_r + r^{-2}\partial_\theta^2$, where θ denotes the azimuthal angle. Unstable modes $v, w \sim \cos(m\theta)\mathrm{e}^{\gamma_m z}$ with azimuthal index number m break up a spatial ring, which is modeled by a uniform background solution ϕ lying on a circular path with length $s = \bar{r}\theta$, where $\bar{r}$ is the mean radius of the ring. Equation (4) then yields (Soto-Crespo et al., 1992)

  $$\gamma_m^2 = \left(\frac{m}{\bar{r}}\right)^2\left(2A - \frac{m^2}{\bar{r}^2}\right) \tag{7}$$

  and the maximum number of modulations on the ring is provided by the integer part number $m_{\max} = \mathrm{int}\{\bar{r}^2 A\}^{1/2}$. An example of filamentation on a ring-diffraction pattern emerging from a tenth-order super-Gaussian beam is shown in Figure 1 for $f(s) = s$ in the presence of a random-phase noise. About seven to eight filaments emerge and rapidly collapse.
- *Transverse instability*: Although elementary, MI theory has been considered as the starting point for understanding laser filamentation, until plane waves are replaced by bounded solutions. These can be the soliton modes of the NLS equation, and the resulting instability is termed transverse instability. It appears when a soliton ϕ is perturbed by oscillatory modulations developing along one axis. For instance, when the Laplacian reduces to the 1-dimensional operator $\boldsymbol{\nabla}_{\perp}^{2} = \partial_x^2$ in Equation (2), a soliton solution $\phi(x, \lambda) = \sqrt{2\lambda}\,\mathrm{sech}(\sqrt{\lambda}x)$ can undergo perturbative modes v, w decomposed as $V(x), W(x)\cos(k_y y)\mathrm{e}^{\gamma z}$. These perturbations are local in x and they promote the formation of bunches, periodically distributed over the y-axis. The operators L_0 and L_1 in Equation (5) are transformed as $L_0 = \lambda - \partial_x^2 - f(\phi^2) + k_y^2$ and $L_1 = L_0 - 2f'(\phi^2)\phi^2$. For the cubic nonlinearity ($f(s) = s$), the instability growth rate is close to the theoretical curve $\gamma^2 = 1.08k_y^2(3\lambda - k_y^2)$ (Rypdal & Rasmussen, 1989). The same reasoning holds when the Laplacian of Equation (1) includes a third dimension, for example, a time variable, along which 2-dimensional spatial solitons are subject to periodic fluctuations. Because bounded solutions ϕ are not always accessible analytically, numerical computations are often required for determining γ^2 (Akhmediev et al., 1992).

Besides filament formation, Equations (4) and (5) provide information on the inner stability of solitons, known as orbital stability, which refers to the ability

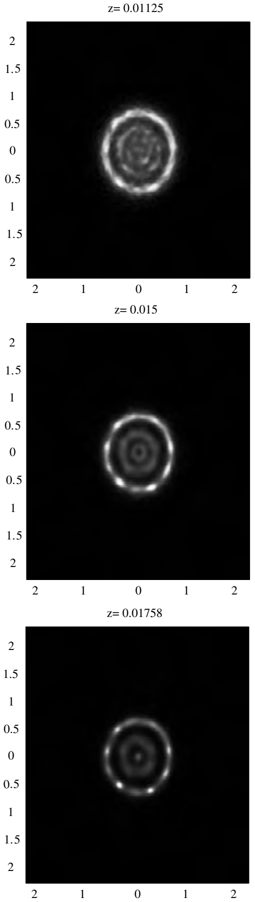

Figure 1. Filamentation pattern numerically computed for a tenth-order super-Gaussian beam with $P_{\text{in}} = 30P_{\text{c}}$ in a cubic medium.

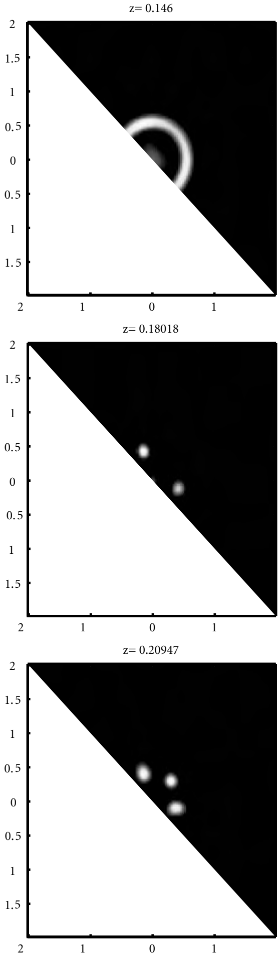

Figure 2. Filamentation pattern numerically computed for a Gaussian beam with $P_{\text{in}} = 20P_{\text{c}}$ in a saturable medium.

of initial solutions near equilibrium states to converge to robust soliton shapes. Standard procedures allow the stability criterion $\mathrm{d}P(\phi)/\mathrm{d}\lambda > 0$ to be established, where $P(\phi) \equiv \int \phi^2 \mathrm{d}\boldsymbol{r}$ is the soliton power (Kuznetsov et al., 1986). Orbital stability applies to filaments formed in saturable media with $f(s) = \alpha s - \beta s^2$, $f(s) = \alpha s/(1 + \beta s)$, or $f(s) = \alpha(1 - \mathrm{e}^{-s})$, where α and β are positive constants. Relaxation of filaments to stable solitons promoted by the nonlinearity $f(s) = s/(1 + 2 \times 10^{-3}s)$ is shown in Figure 2.

On the whole, filamentation follows from linear stability analyses, which are valid as long as $\varepsilon(v + \mathrm{i}w)$ remains smaller than ϕ, thus applying to early stages in the beam propagation. At later stages, filaments develop into a fully nonlinear regime, and they may interact. From the interplay between diffraction and nonlinearity, two filaments with radius ρ and separated by the distance $\Delta > 2\rho$ can fuse whenever Δ is below a critical value, Δ_{c}. This critical distance can be evaluated by the balance between the free and interaction contributions in the Hamiltonian of Equation (1) (Bergé et al., 1997). With no saturation, each filament whose

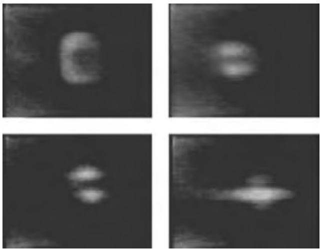

Figure 3. Filamentation and coalescence event in a two-spot pattern produced by a 50 fs pulse with 5 mJ energy propagating in air at the increasing distances (from left to right and from top to bottom): $z = 2.5$, 4.5, 6.5, and 8.5 m (Tzortzakis et al., 2001).

power is above critical creates its own attractor, at which it mostly freely collapses. This constraint is softened by including saturation, so that filaments with powers above critical are able to coalesce into an intense central lobe. As an example, the coalescence of two spots resulting from the propagation of modulationally unstable femtosecond pulses in air is shown in Figure 3 (Tzortzakis et al., 2001).

Modulational instability and multisoliton-like generation take place in various nonlinear media, such as biased photorefractive crystals or quadratically nonlinear optical materials favoring the coupling of fundamental and second harmonic fields (Fuerst et al., 1997). Filamentation moreover occurs in inertial fusion confinement (IFC) experiments as a harmful instability destroying the homogeneity in the beam energy distributed in the focal spot. For IFC, it has detrimental consequences to the hydrodynamics of the plasma created by a laser beam and contributes to the growth of parametric instabilities, which dissipate part of the laser energy. To limit its influence, optical smoothing techniques can be employed. For instance, random phase plates are used to generate a diffraction pattern composed of speckles, whose size λ_{sp} is smaller than the optimal wavelength λ_{mod} that maximizes the filamentation growth rate. This contributes to suppress laser filamentation (Labaune et al., 1992).

Luc Bergé

See also **Development of singularities; Kerr effect; Nonlinear optics**

Further Reading

Akhmediev, N.N., Korneev, V.I. & Nabiev, R.F. 1992. Modulational instability of the ground state of the nonlinear wave equation: optical machine gun. *Optics Letters*, 17: 393–395

Bergé, L., Schmidt, M.R., Rasmussen, J. Juul, Christiansen, P.L. & Rasmussen K.Ø. 1997. Amalgamation of interacting light beamlets in Kerr-type media. *Journal of the Optical Society of America* B, 14: 2550–2562

Bespalov, V.I. & Talanov, V.I. 1966. Filamentary structure of light beams in nonlinear media. *Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki, Pis'ma v Redaktsiyu* (USSR JETP), 3: 471–476 [Translated in *JETP Letters*, 3: 307–310]

Campillo, S.L., Shapiro, S.L. & Suydam, B.R. 1973. Periodic breakup of optical beams due to self-focusing. *Applied Physics Letters*, 23: 628–630

Fuerst, R.A., Baboiu, D.-M., Lawrence B., Torruellas, W.E., Stegeman, G.I., Trillo, S. & Wabnitz, S. 1997. Spatial modulational instability and multisolitonlike generation in a quadratically nonlinear optical medium. *Physical Review Letters*, 78: 2756–2759

Kuznetsov, E.A., Rubenchik, A.M. & Zakharov, V.E. 1986. Soliton stability in plasmas and hydrodynamics. *Physics Reports*, 142: 103–165

Labaune, C., Baton, S., Jalinaud, T., Baldis, H.A. & Pesme, D. 1992. Filamentation in long scale length plasmas: experimental evidence and effects of laser spatial incoherence. *Physics of Fluids* B, 4: 2224–2231

Rypdal, K. & Rasmussen, J. Juul. 1989. Stability of solitary structures in the nonlinear Schrödinger equation. *Physica Scripta*, 40: 192–201

Soto-Crespo, J.M., Wright, E.M. & Akhmediev, N.N. 1992. Recurrence and azimuthal-symmetry breaking of a cylindrical Gaussian beam in a saturable self-focusing medium. *Physical Review* A, 45: 3168–3175

Tzortzakis, S., Bergé, L., Couairon, A., Franco, M., Prade, B. & Mysyrowicz, A. 2001. Break-up and fusion of self-guided femtosecond light pulses in air. *Physical Review Letters*, 86: 5470–5473

FINGERING

See **Hele-Shaw cell**

FINITE ELEMENT METHODS

See **Numerical methods**

FINITE-DIFFERENCE METHODS

See **Numerical methods**

FISHER'S EQUATION

See **Zeldovich–Frank-Kamenetsky equation**

FISKE STEPS

See **Josephson junctions**

FITNESS LANDSCAPE

The notion of landscape is chosen in analogy to terrestrial landscapes, which are functions over a two-dimensional space, $f(x, y)$. Commonly, landscape is altitude h as a function of latitude θ and longitude φ: $h(\theta, \varphi)$. In the history of physics, the landscape concept was first applied to motion in the gravitational field resulting in the concept of a potential energy surface. The gravitational potential V depends on altitude, latitude, the mass of the particle m, and the gravitational acceleration on the surface of the Earth as a function of latitude, $g(\theta)$: $V(\theta, \varphi) = m\, g(\theta)\, h(\theta, \varphi)$. The metaphor of a landscape embedded in the gravitational field suggests an exploration of optimal downhill paths following the negative gradient of the potential, $-\,\mathbf{grad}\ V(\boldsymbol{x}) = -\,(\partial V/\partial x_1, \partial V/\partial x_2, \ldots, \partial V/\partial x_n)$, with infinitesimal step width and vanishing kinetic energy corresponding to infinitesimally slow motion.

The landscape metaphor is used today in many different disciplines where the optimization of a nonsimple cost function is the primary goal. Such a cost function may depend on any number of variables, for example, spatial coordinates and strengths of external fields, and then the space upon which a landscape is built will be high-dimensional. Landscapes describing complex systems exhibit a large number of local minima and maxima. Models of disordered systems, spin glasses in particular, were the first examples studied in detail. Spin glasses are traditionally studied by statistical mechanics and still represent the best-studied cases of statistics on complex landscapes (Binder, 1986; Dotsenko et al., 1990).

The notion of fitness landscape was introduced into evolutionary biology by Sewall Wright in 1932 as a metaphor for the visualization of Darwinian evolution as an optimization process (see Wright, 1967). The Darwinian mechanism operates on populations over the course of many generations. It is based on genetic variation of individuals through mutation and recombination and selection of the variants with largest reproductive success. Reproductive success of a genotype I is measured in terms of its fitness value f, which counts the number of (fertile) descendants in the next generation. The concept of a fitness landscape or mapping ϕ assigns a fitness value f_k to every genotype I_k: $\phi(I_k) = f_k$. Population genetics describes the evolution of a population by means of the time dependence in the distribution of genotypes: At time t the genotype I_k is assumed to be present with frequency $x_k(t)$ in a population of N individuals distributed over n types or variants. The frequencies fulfill $\sum_{k=1}^{n} x_k = 1$, and for asexual reproduction, we have $\mathrm{d}x_k/\mathrm{d}t = x_k(f_k - \bar{f})$ with $\bar{f} = \sum_{k=1}^{n} x_k f_k$ being the mean fitness of the population. The time derivative of the mean fitness,

$$\frac{\mathrm{d}\bar{f}}{\mathrm{d}t} = \overline{f^2} - \left(\bar{f}\right)^2 = \mathrm{var}\{f\} \geq 0, \qquad (1)$$

is equal to the variance of the fitness. The mean fitness of a population is a nondecreasing function of time and, hence, subjected to optimization. The frequencies of all variants with fitness values larger than average, $f_k > \bar{f}$, increase whereas those of genotypes that are less productive than average, $f_k < \bar{f}$ decrease until the genotype becomes extinct. This process continues until the mean fitness $\bar{f}$ has reached its maximum because all variants except the fittest have disappeared. The optimization principle also holds for sexual reproduction with recombination, as long as variants at a single gene locus are considered, and is called Fisher's fundamental theorem in this context. More general models of population dynamics describe variation explicitly. When mutation is included and several genetic loci are considered, the optimization principle is no longer valid. However, it works in the mild form as an optimization heuristic in the sense that optimization is observed in almost all cases.

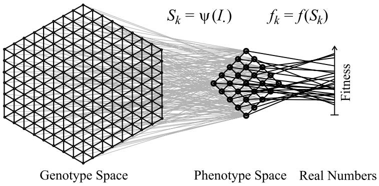

Figure 1. Fitness landscapes illustrated by two consecutive mappings, one from genotype space into phenotype space and the other from phenotype space into the positive real numbers, $\mathbb{R}_+$, representing the fitness values. In agreement with the available data on biopolymer landscapes, we are dealing with the phenomenon of neutrality: There are many more genotypes than phenotypes, and different phenotypes may have fitness values that cannot be distinguished by selection.

The dichotomy of biological evolution—genetic variation in the form of mutation and recombination changes genotypes, whereas selection operates on phenotypes—suggests a splitting of the conventional fitness landscapes into two successive mappings: Genotypes are mapped onto a space of phenotypes, and the phenotypes are evaluated by a second map to yield fitness values (Figure 1):

Biological evolution:
  genotype $\Longrightarrow$ phenotype $\Longrightarrow$ fitness,
RNA evolution:
  sequences $\Longrightarrow$ structure $\Longrightarrow$ replication rate.

Evolution experiments with RNA molecules in the test tube (Biebricher & Gardiner, 1997) are sufficiently simple to allow for a description of the fitness landscape in molecular terms. The RNA sequence forms a molecular structure that determines the replication rate parameters that are the molecular counterparts of fitness values.

In biological landscape metaphors, the genotypes are materialized as polynucleotide sequences, DNA or RNA. These are strings built from four classes of symbols, $\{\mathbf{A}, \mathbf{T}(\mathbf{U}), \mathbf{G}, \mathbf{C}\}$. In RNA **T** is replaced by **U**. All sequences of given chain length are subsumed in a discrete sequence space, $\mathcal{I}$, with the Hamming distance between two sequences I_i and I_j, $d_{ij}^{(h)}$ serving as metric. The fitness landscape can be expressed by

$$\phi \; : \; \{\mathcal{I}; d_{ij}^{(h)}\} \Longrightarrow \mathbb{R}_{(+)}, \qquad (2)$$

where the plus sign implies a restriction to strictly positive fitness values. The structures are points in another discrete metric space $\mathcal{S}$ with some distance measure between structures $d_{ij}^{(s)}$. The fitness landscape is properly split into two mappings

$$\psi : \{\mathcal{I}; d_{ij}^{(h)}\} \Longrightarrow \{\mathcal{S}; d_{ij}^{(s)}\}$$

and

$$f : \{\mathcal{S}; d_{ij}^{(s)}\} \Longrightarrow \mathbb{R}_{(+)}. \tag{3}$$

Sequence-structure maps of RNA have been studied by means of computer models (Schuster, 2001, 2003) on the simplified level of RNA secondary structure. As sketched in Figure 1, the number of sequences is much larger than the number of secondary structures. Hence, many sequences may give rise to the same structure. In population biology this phenomenon is known as neutrality in sequence space, and it was discussed already in the 1980s by Kimura (1983) when the first biopolymer sequences became accessible. Neutrality was also found to be highly relevant for the efficiency of evolutionary processes because it allows for escape from local optima that would otherwise trap populations for long times (Fontana & Schuster, 1998).

The nature of landscapes can be analyzed by means of an expansion of the landscape in eigenfunctions of the Laplacian on the underlying discrete space (Reidys & Stadler, 2002). The expansion coefficients of this expansion allow for estimates of the hardness of optimization. Whenever we have only a single dominating expansion coefficient, the optimization problem is much simpler than in the case of equally important blending of two or more eigenfunctions. Biopolymer landscapes of free energies of conformations or other relevant properties turned out to be rather complex as several eigenfunctions of the Laplacian were found to be important.

PETER SCHUSTER

See also **Biological evolution; Spin systems**

Further Reading

Biebricher, C.K. & Gardiner, W.C. 1997. Molecular evolution of RNA in vitro. *Biophysical Chemistry*, 66: 179–192

Binder, K. 1986. Spin glasses: experimental facts, theoretical concepts and open questions. *Reviews of Modern Physics*, 58: 801–976

Dotesenko, V.S., Feigel'man, M.V. & Ioffe, L.B. 1990. Spin glasses and related problems. *Soviet Science Reviews A. Physics*, 15: 1–250

Fontana, W. & Schuster, P. 1998. Continuity in evolution. On the nature of transitions. *Science*, 280: 1451–1455

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*, Cambridge and New York: Cambridge University Press

Reidys, C.M. & Stadler, P.F. 2002. Combinatorial landscapes. *SIAM Review*, 44: 3–54

Schuster, P. 2001. Evolution *in silico* and in vitro: the RNA model. *Biological Chemistry*, 382: 1301–1314

Schuster, P. 2003. Molecular insights into the evolution of phenotypes. In *Evolutionary Dynamics—Exploring the Interface of Accident, Selection, Neutrality, and Function*, edited by J.P. Crutchfield & P. Schuster, Oxford and New York: Oxford University Press, pp.163–215

Wright, S. 1967. "Surfaces" of selective value. *Proceedings of the National Academy of Sciences USA*, 58: 165–172

FITZHUGH–NAGUMO EQUATION

Around 1960, Jin-ichi Nagumo (from the University of Tokyo) visited with Richard FitzHugh (at the U.S. National Institutes of Health), who was adapting the relaxation-oscillator models of Karl Bonhoeffer and Balthasar van der Pol to provide a more simple formulation of nerve membrane switching than the four-variable Hodgkin–Huxley system. From this visit emerged the FitzHugh–Nagumo (FN) system as a two-variable oscillator model (FitzHugh, 1961), which led to a simplified formulation of neural action potentials (Nagumo et al., 1962).

Neurons display relaxation oscillations par excellence. When the voltage across the cell membrane (in the trigger zone of the soma) exceeds a threshold, they fire an action potential and then gradually decay to the resting state. If the exciting input is constantly above the threshold, the neuron fires repeatedly. The frequency of such firing depends on the intensity of the input (total input coming from neighboring neurons). FitzHugh sought to reduce the Hodgkin–Huxley model of this dynamics to a two-variable model for which phase-plane analysis applies. His general observations were that the gating variables of the Hodgkin–Huxley model, n (potassium activation) and h (sodium inactivation), have slow kinetics relative to m (or sodium activation), and that for typical nerves $n+h \approx 0.8$. Furthermore, FitzHugh noticed that the voltage nullcline had the shape of a cubic function and the n-nullcline could be approximated by a straight line, both within the physiological range of the variables. This led to the following two-variable model that provides a phase space qualitative explanation of the formation and decay of the action potential.

$$\begin{aligned} \dot{v} &= +v(v-\theta)(1-v) - y + I, \\ \dot{y} &= \varepsilon(v - \gamma y), \end{aligned} \tag{1}$$

where the dots indicate time derivatives.

Like all relaxation oscillators, this oscillator has a slow accrual phase and a fast release phase. Here v is the scaled voltage or membrane potential (namely, the output of the neuron), and y the single recovery variable accounting for the slow dynamics (sodium inactivation and potassium activation variables). The difference in time scales between sodium activation (m) and sodium inactivation and potassium activation (n and h) is represented by $\varepsilon \ll 1$. As ε increases, so does the frequency of oscillation. The constant γ

is a shunting parameter, $0<\theta<1$ is a thresholding parameter, and I is an externally applied current.

Such oscillators exhibit the characteristic dynamics of a real neuron. When fed by a constant amount of low current, the oscillator gradually increases its voltage (or membrane charge) until it reaches the threshold. Upon reaching threshold, the neuron fires and quickly releases the accumulated charge. If the current source is sufficient and is constantly applied, this results in stable limit-cycle oscillations.

The FN oscillator is applied in neuroscience to study the synchronization properties of neurons, which is particularly easy for relaxation oscillators such as the FN model. Following the Hebbian rule, neurons that fire closely in time strengthen their synaptic connections, and therefore synchronization in neural activity is important to the way the states of a brain evolve in time. Furthermore, information in neural tissues seems to be coded in the spatiotemporal firing of sets of neurons that fire synchronously, and tissues that are anatomically separated may operate together when they are synchronized.

An ensemble of coupled excitable oscillators of the FN type makes up an excitable medium, through which electrical signals can propagate. This is the case for a nerve and for other excitable media such as muscles. To model the spatiotemporal evolution of membrane potential and its generated current flows in such structure, one treats the axon or dendrite branch of the nerve as a membrane cylinder, with x denoting the distance along the cylinder. If ρ is the intracellular resistivity (in ohm-cm) and d the diameter, then the axial current is

$$I_i = -(\pi d^2/4\rho)\frac{\partial V}{\partial x}\,, \tag{2}$$

which follows directly from Ohm's law. Conservation of current then requires

$$C_m\frac{\partial V}{\partial t} = \frac{d}{4\rho}\frac{\partial^2 V}{\partial x^2} - I_{\text{ion}} + I(x,t), \tag{3}$$

where I_{ion} is the ionic current flowing through a unit area of the membrane surface and $I(x,t)$ and C are an external current and membrane capacitance per unit area.

Applying Equation (3) to the FN case, suggests that the electrical propagation in nerves can be represented as a reaction-diffusion system of the form

$$\begin{aligned}\frac{\partial^2 V}{\partial x^2} - \frac{\partial V}{\partial t} &= +F(V) - Y + I(x,t),\\ \frac{\partial Y}{\partial t} &= \varepsilon(V-\gamma Y)\,.\end{aligned} \tag{4}$$

Here $F(V)$ is a function with cubic shape such as $F(V)=V(V-\alpha)(1-V)$ or some similar function. [Such diffusion models can also be extended to more dimensions with $\partial^2V/\partial x^2+\partial^2V/\partial y^2$ and $V(x,y;t)$.]

Under a traveling-wave analysis, let $I(x,t) = 0$ and $z = x+ct$ and define $\mathrm{d}V/\mathrm{d}z \equiv W$. Then Equations (4) become a system of ordinary differential equations

$$\begin{aligned}\frac{\mathrm{d}V}{\mathrm{d}z} &= W,\\ \frac{\mathrm{d}W}{\mathrm{d}z} &= F(V) - Y + cW,\\ \frac{\mathrm{d}Y}{\mathrm{d}z} &= \frac{\varepsilon}{c}(V-\gamma Y)\,,\end{aligned} \tag{5}$$

where one seeks traveling-waves solutions as trajectories $(V(z), W(z), Y(z))$ that approach the origin $(0,0,0)$ as $z\to\pm\infty$ (Scott, 2003).

With $\varepsilon=0$, Y is constant. In this case, the only stable traveling wave solution is a level change from one of the outer zeros of $F(V)-Y$ to the other; in other words, a moving impulse-like solution does not exist. For $0<\varepsilon\ll 1$, however, there is an impulse-like traveling-wave solution (or solitary wave). At arbitrarily small values of ε, this solitary wave continues to exist but with the front and back edges of the wave becoming far apart. These front and back edges interpolate between the slowly varying regions and are called boundary layers (as in hydrodynamics). As $\varepsilon\to 0$ on a scale where the impulse length is unity, the boundary layers reduce to step functions.

One of the main research areas related to these reaction-diffusion models in electrophysiology focuses on pattern formation and cardiac rhythm disturbances. FN equations have been used to investigate a variety of unusual front-bifurcation and pattern-formation processes. The precise conditions for wave front formation and subsequent wave propagation in an excitable medium are critical for understanding the genesis and possible control of re-entrant (spiral-wave) cardiac arrhythmias.

L. VAZQUEZ AND M.-P. ZORZANO

See also **Boundary layers; Hodgkin–Huxley equations; Neurons; Reaction-diffusion systems; Relaxation oscillators**

Further Reading

FitzHugh, R. 1961. Impulses and physiological states in theoretical models of nerve membranes. *Biophysiscal Journal*, 1: 445–466

Nagumo, J., Arimoto, S. & Yoshizawa, S. 1962. An active pulse transmission line simulating nerve axons. *Proceedings of the Institute of Radio Engineering and Electronics*, 50: 2061–2070

Scott, A. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Scott, A. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

FIXED POINTS

See **Equilibrium**

FLAME FRONT

Combustion waves can propagate over a wide range of burning velocities that differ by more than three orders of magnitude for the same mixture. At one end of the velocity spectrum, we have a laminar flame (deflagration) that propagates at a typical velocity of about half a meter per second for common fuel-air mixtures at normal conditions. At the other end, the combustion wave propagates as a detonation whose speed is of the order of a couple of thousand meters per second in the same mixture. In between these limits, we have an almost continuous range of turbulent burning velocities. Both laminar flames and detonation wave are intrinsically unstable and have the morphology of a transient cellular structure (Figure 1).

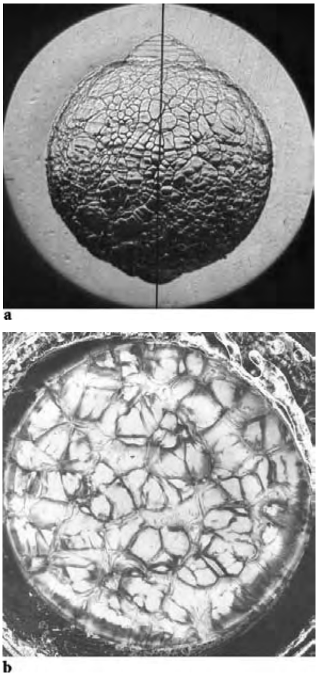

Figure 1. Cellular structures of laminar flame and detonation front. (a) Laminar outward propagating spherical flame (rich dimethyl ether–air at equivalence ratio of 1.2 and 10 atm). (b) Detonation: unstable cellular detonation front as recorder upon reflection from a soot-coated glass plate (C_2H_2–O_2 mixture at 10 mmHg).

A laminar flame is essentially an isobaric diffusion-reaction wave. Its propagation speed is determined by the rate of diffusion transport of heat from the reaction zone to the cold unburned mixture and the characteristic time of heat release of the chemical reactions. A laminar flame speed (S_L) is proportional to the square root of the product of thermal diffusivity (D_{th}) and the reaction rate (w_r) (i.e., $S_L \sim \sqrt{D_{th} w_r}$). The reaction rate is given by $w_r \sim \exp(-E/RT_f)$ in which E is the activation energy and T_f is the flame temperature. Because the activation energy is very large in general, the reaction rate is extremely temperature sensitive. Thus, any fluctuation in the flame temperature will result in a large variation in the reaction rate leading to the development of instability of the flame front. Due to the large density and temperature changes across the flame, a strong thermal expansion of the burned gas results from the conservation of mass. The flame as a strong density interface as well as an expansion wave is subject to a number of dynamic instability mechanisms in the presence of an acceleration field. Furthermore, competition between heat and mass diffusion across the flame results in thermal diffusion instability. Rapid density changes across the flame also give rise to acoustic wave generation that can couple with increase in burning rate to induce acoustic driven instability. Thus, in practice, there is a wealth of instability mechanisms that render laminar flames unstable. Various instability mechanisms can be at work simultaneously and can influence each other. However, historically, each instability mechanism was isolated and studied individually and was, thus, named after the original researchers.

The flame as a density interface is unstable when subjected to an acceleration field (for example, gravity). If the flame propagates upward, the light burned gas (ρ_b) is at the bottom and the heavy unburned gas (ρ_u) is on the top. The lighter fluid will be driven upward (i.e., buoyancy) whereas the heavier fluid is driven downward by gravity (g). This motion will destabilize the interface, and hence, small perturbations on the flame surface will grow with time (t). If the perturbed flame surface (F) is defined as $A\exp(\sigma t + ikx)$, where k is the wave number and x the space coordinate, the disturbance growth rate σ can be determined from normal mode stability analysis as

$$\sigma = \sqrt{gk\gamma/(2-\gamma)}, \quad \gamma = 1 - \rho_b/\rho_u. \tag{1}$$

Therefore, instability occurs at all wavelengths but growth is faster at shorter wavelengths. This phenomenon was first discovered by Lord Rayleigh (1883) and later by Geoffrey Ingram Taylor (1950), and thus, it is referred to as the Rayleigh–Taylor instability. This instability is common to all density gradient fields in the presence of an acceleration field normal to it.

Due to the density change across the flame, the flow velocity increases as the density ratio because the flame is approximately isobaric. The expansion across the front will induce a divergent flow field ahead of a curved flame (Williams, 1985). This divergence slows down the local flow speed ahead of the curved flame front, and assuming the flame speed to be constant, this will result in a growth of the curvature of the flame. This instability was first discovered by G. Darrieus (1938) and Lev Landau (1944). The Darrieus–Landau instability was obtained by treating the flame as a surface of discontinuity moving at a constant speed. In the limit of small density change, stability analysis gives the growth rate as

$$\sigma = \gamma k S_{\mathrm{L}}/2. \tag{2}$$

Therefore, it was concluded that a flame front is unstable to perturbations at all wavelengths, with growth rate proportional to wave number (i.e., perturbations grow faster for small wavelengths). Unfortunately, this conclusion was contrary to later laboratory observations of small-scale stable flames.

The deficiency of this model was the result of neglecting the finite thicknesses of the flame front, which influences the flame speed when the flame is curved. To include the effect of flame thickness, George Markstein (1951) proposed a phenomenological model by adding a modification of flame speed due to curvature and showed that curvature decreases the flame speed and tends to stabilize the flame at short wavelengths inhibiting its growth. A more rigorous derivation of the dispersion equation including diffusion effect on the flame speed (in the limit of small density jump) was given later by Gregory Sivashinsky (1983) as

$$\sigma = \frac{\gamma k S_{\mathrm{L}}}{2} - \left[\frac{\beta}{2}(\mathrm{Le}-1)+1\right] \times D_{\mathrm{th}}k^2 - 4\frac{D_{\mathrm{th}}^3}{S_{\mathrm{L}}^2}k^4 + \sqrt{gk\gamma/(2-\gamma)}, \tag{3}$$

where $D_{\mathrm{th}}k^2$ in the second term of Equation (3) represents the thermal relaxation via transverse thermal diffusion that stabilizes the flame. Le is the Lewis number (i.e., the ratio of thermal diffusivity to mass diffusivity) and $(\mathrm{Le}-1)$ in the second term denotes the competition between heat loss via thermal diffusion and the enthalpy gain via mass diffusion. As a result, flame temperature will increase or decrease if Le is less or larger than unity. Equation (3) shows that if $\mathrm{Le}-1$ is less than $-2/\beta$ (β is the reduced activation energy), diffusion transport will destabilize the flame. This is the mechanism of the cellular instability. The third term in Equation (3) is the thermal relaxation to the modification of flame temperature caused by the heat and mass diffusion (Clavin, 1985). Therefore, for long wavelength disturbance, the hydrodynamic instability dominates (Figure 1a). At short wavelengths, diffusion relaxation stabilizes the flame. At moderate wavelengths, the competition between heat and mass transfer induces cellular instability at small Le. At large Le, flame temperature is very sensitive to the mass diffusion of the deficient reactant. Coupling between the diffusion and temperature sensitive reaction yields pulsating and spinning waves. This traveling wave is often seen in lean propane-air flames.

By further considering the effect of flame curvature on flow field, Equation (3) can be normalized as an evolution equation of flame front

$$F_t + \tfrac{1}{2}(\nabla F)^2 + \nabla^2 F + 4\nabla^4 F = 0 \tag{4}$$

This is the so-called Kuramoto–Sivashinsky equation which was independently developed by Yoshiki Kuramoto (Kuramoto & Tsuzuki, 1976) for the study of phase turbulence in the Belousov–Zhabotinsky reaction and by Gregory Sivashinsky (1983) for thermal diffusive instabilities of flame fronts. This equation has also been used to model directional solidification and weak fluid turbulence. However, the assumption of small density jump used in Equations (3) and (4) is not rigorous in practical flames. A unified model considering large density jump and Le was obtained by Class Andreas et al. (2003).

Heat release by combustion results in an increase in the specific volume of the product gases and thus generates acoustic waves (Chu, 1956). The acoustic waves play two roles in affecting combustion: (1) inducing pressure-heat release coupling via pressure-dependent reactions, and (2) increasing flame surface area via the baroclinic torque (Meshkov instability). If the changes of pressure and chemical heat release are in phase, the acoustic instability occurs (Rayleigh, 1877; Markstein, 1953; Clavin, 2002). Lord Rayleigh first used this criterion (Rayleigh criterion) to explain the singing flame and Rijke's tone (where heating the bottom of a tube causes it to produce sound). The acoustic instability causes the major problems of noise and vibration in combustors (Putnam & Dennis, 1953). On the other hand, volumetric heat loss reduces the flame speed and changes the resident time of the emitting gases. This coupling triggers the radiation induced instability for weak flames (Ju et al., 2000).

At the upper limit of propagation of combustion waves, the propagation mechanism is not due to diffusion. The flame instability mechanisms discussed above are too slow to be relevant in detonation wave instability. A detonation wave is a supersonic compression wave where mixture is ignited by the adiabatic compression of the leading shock. The classical structure of a detonation wave was formulated by Yakov Zeldovich, John von Neumann, and W. Döring (ZND) independently in the early 1940s and consists of a leading shock followed by the reaction zone after a short induction length (Zeldovich, 1940; von Neumann, 1942; Döring, 1943). Gas dynamic theory gives the detonation wave speed as proportional to the square root of the chemical heat release and does not involve any non-equilibrium rate processes. Again due to the high-temperature sensitivity of the reaction rates, small temperature fluctuations due to variation of the leading shock speed will result in large variations in the induction length and reaction rates, hence the coupling between the energy release zone and the leading shock. The instability yields a transient three-dimensional cellular detonation front (Lee, 1984). The unstable cellular detonation front consists of an ensemble of interacting transverse shock waves with the leading shock front. The cell boundaries (Figure 1b) are formed by the intersections of the transverse shocks. Shock interactions (Mach reflections) also give rise to the formation of shear layers which lead to turbulence generation due to Kelvin–Helmholtz instability. Chemical reactions in cellular detonations occur in disjointed piecemeal zones embedded within the complex of interacting shocks and shear layers.

The instability of the laminar ZND detonation structure was demonstrated theoretically by standard normal mode stability analysis using the one-dimensional Euler equation (e.g., Erpenbeck, 1964; Lee & Stewart, 1990). In one dimension, unstable detonations are referred to as pulsating detonations that go from harmonic oscillations near the stability limit to highly nonlinear and eventually to chaotic oscillations with the increase of the activation energy. By examining the bifurcation diagram, it is interesting to find that the path to higher instability mode follows closely the Feigenbaum route (Feigenbaum, 1983) of a period-doubling cascade observed in many nonlinear systems. One-dimensional pulsating detonation as well as two- and three-dimensional cellular detonations have been reproduced qualitatively via numerical simulation using the reactive Euler equations (Bourlioux et al., 1991; Short & Stewart, 1999). However, the detailed description of the turbulent structure and chemical reactions requires resolutions far beyond current computing capabilities.

In between the two limits of laminar flames and detonations, there is a continuous range of flame speeds that depend on turbulence. The morphology of a turbulent flame is a time-dependent cellular or wrinkled surface. Turbulent flame is, in fact, an unstable flame, and the effect of turbulence is to increase the burning rate via faster transport and increase in burning surface area. In the limit of very intense turbulence where mixing and reaction rates are comparable, auto-ignition may result, and thus, the mechanism becomes similar to that of a detonation. It differs only in the manner in which auto-ignition is achieved by turbulent mixing of fresh mixture with hot products or by adiabatic heating of the leading shock. Thus, nature tends to maximize the burning rate of a mixture, and instability is a route to optimize the burning rate for given initial and boundary conditions.

YIGUANG JU AND JOHN LEE

See also **Candle; Explosions; Forest fires; Kuramoto–Sivashinsky equation; Reaction-diffusion systems; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Bourlioux, A., Majda, A.J. & Roytburd, V. 1991. Theoretical and numerical structure for unstable one-dimensional detonations. *SIAM Journal of Applied Maths*, 51: 303–343

Class Andreas G., Matkowsky, B.J. & Klimenko, A.Y. 2003. Stability of planar flames as gas dynamic discontinuities. *Journal of Fluid Mechanics*, 491: 51–63

Clavin, P. 1985. Dynamic behaviour of premixed flame fronts in laminar and turbulent flows. *Progress in Energy and Combustion Science*, 11: 1–39

Clavin, P. 2002. Dynamics of combustion fronts in premixed gases: from flames to detonation. *Proceedings of the Combustion Institute*, 29: 569

Chu, B.T. 1956. Stability of systems containing a heat source—the Rayleigh criterion, National Advisory Committee for Aeronautics research memorandum, RM56D27

Darrieus, G. 1938. Propagation d'un front de flame, unpublished work presented at La Technique Moderne and le Congrès de Méchanique Appliquée, Paris

Döring, W. 1943. On detonation processes in gases. *Annals of Physics, Leipzig*, 43: 421–436

Erpenbeck, J. 1964. Stability of idealized one-reaction detonations. *Physics of Fluids*, 7: 684–696

Feigenbaum, M. 1983. Universal behaviour in nonlinear systems. *Physica* D, 7: 16–39

Ju, Y., Law, Chung K., Maruta, K. & Niioka, T. 2000. Radiation induced instability of stretched premixed flames. *Proceedings of the Combustion Institute*, 28: 1891–1900

Kuramoto, Y. & Tsuzuki, T. 1976. Persistent propagation of concentration waves in dissipative media far from thermal equilibrium. *Progress of Theoretical Physics*, 55: 356–369

Landau, L.D. 1944. On the theory of slow combustion. *Acta Physiocochimica URSS*, 19: 77–85

Lee, J.H.S. 1984. Dynamic parameters of gaseous detonation. *Annual Reviews of Fluid Mechanics*, 16: 311–316

Lee, H. & Stewart, D. 1990. Calculation of linear detonation instability: one-dimensional instability of plane detonation. *Journal of Fluid Mechanics*, 216: 103–132

Markstein, G.H. 1951. Experimental and theoretical studies of flame front instability. *Journal of the Aeronautical Sciences*, 18: 199–209

Markstein, G.H. 1953. Instability phenomena in combustion waves. *Proceedings of the Combustion Institute*, 14: 44–59

Putnam, A.A. & Dennis, W.R. 1953. A study of burner oscillations of the organ-pipe type. *Transaction of the ASME*, 75: 15–28

Rayleigh, L. (John William Strutt). 1877. *The Theory of Sound*, vol.2, London: Macmillan; reprinted New York: Dover, 1945, p. 226

Rayleigh, L. (John William Strutt). 1883. Investigation of the character of the equilibrium of an incompressible heavy fluid of variable density. *Proceedings of the London Mathematical Society*, 14: 170–177

Short, M. & Stewart, D. 1999. Multi-dimensional stability of weak-heat-release detonation. *Journal of Fluid Mechanics*, 382: 103–135

Sivashinsky, G.I. 1983. Instabilities, pattern formation, and turbulence in flames. *Annual Reviews of Fluid Mechanics*, 15: 179–199

Taylor, G.I. 1950. The instability of liquid surfaces when accelerated in a direction perpendicular to their planes.

Proceedings of the Royal Society of London A, 201: 192–196

von Neumann, J. 1942. Theory of detonation waves, Proj. Report No. 238, OSRD report No.549. In *Von Neumann, Collected Works*, vol. 6, edited by A.J. Taub, Oxford: Pergamon Press, 1963

Williams, F.A. 1985. *Combustion Theory*, 2nd edition, New York: Benjamin Cumming

Zeldovich, Y.B. 1940. On the theory of the propagation of detonations in gaseous systems. *Experimental and Theoretical Physics SSSR*, 10: 542

FLIP-FLOP CIRCUIT

A bistable circuit is one that can exist indefinitely in either one of two stable states and that can be induced to make an abrupt transition from one stable state to the other by means of an external excitation. Bistable circuits are known by a variety of names, such as bistable multivibrator, Eccles–Jordan circuit (after the inventors), trigger circuit, binary, and flip-flop, the latter being the term that we adopt. Flip-flops are used for the performance of many digital operations such as counting and storing of binary information.

In general, digital (switching) circuits may be either combinational or sequential. In combinational circuits the Boolean relation that describes the output function is at any moment uniquely determined by the inputs. The output function is independent of all prior input or output conditions and the circuit is without memory. In contrast, the sequential circuit contains feedback and its outputs depend not only on the present inputs but, in general, also on the entire past history of inputs. The flip-flop is a basic sequential circuit that functions as a basic logic memory element. It has two distinct states of equilibrium and may, therefore, be used as a single binary-digit (bit) storage device. The two stable states are referred to by various names such as TRUE and FALSE, HIGH and LOW, or 1 and 0.

Sequential circuits can be presented using a graphical tool known as a state diagram. In a state diagram, nodes are used to represent the different states of a circuit, and connections between nodes are used to show transitions between states with different inputs that act as conditioning signals. Figure 1 represents a general bistable circuit. Let us denote the present state of the circuit as $Q(t)$ and the state that follows after a time interval Δt as $Q(t + \Delta t)$. Then the circuit stays in state 1 as long as condition 1 is applied, but moves to state 0 when condition 2 is applied. Similarly, it stays in state 0 as long as condition 3 is applied, and moves to state 1 when condition 4 occurs.

Four types of flip-flops are presently in use, RS (conditional **S**et and **R**eset), JK (unconditional set and reset), D (**D**elay), and T (**T**rigger), and they have different sets of conditions for transition between the two states. For example, the characteristic equation for the RS flip-flop is given as

$$Q(t + \Delta t) = S + Q(t) \cdot \overline{R} \quad (1)$$

under the constraint

$$S \cdot R = 0. \quad (2)$$

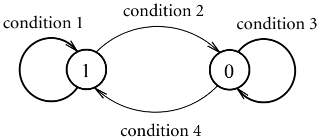

Figure 1. State diagram of a general bistable circuit.

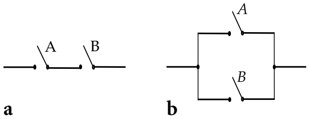

Figure 2. Basic switching elements, (a) AND and (b) OR. An ideal switch consists of a pair of contacts that have zero internal (or switch closed) resistance and infinite leakage (or switch open) resistance. Transitions from one state to the other should be instantaneous. It should attain the open and closed states with equal probability. Although these characteristics have been approached most closely with every new technological generation, we deal in reality with approximations to them.

Here $+$ (OR), $\cdot$ (AND), and $-$ (NOT, also known as inversion and complement) are operations from switching algebra. Switching circuits which represent the AND, $A \cdot B$, and the OR, $A + B$, functions are presented in Figures 2a and b, respectively.

The mathematical theory of switching circuits, first postulated by Shannon (1938, 1949), is based on Boolean algebra. It is defined on a set U that consists of two values, 0 and 1, two basic binary operations, $\cdot$ (which is also called AND, product or conjunction), and $+$ (which is also called OR, sum or disjunction), and a set of basic postulates that were derived and proven by Edward Huntington in 1904, based on the work of George Boole in 1847.

Switching circuits have been designed from various technologies; for example, vacuum tubes were used in their early development. They have become dramatically faster and dramatically smaller in size with every new technology. The advent of integrated circuits (ICs), in which many discrete components (diodes, transistors, and resistors) are fabricated at the same time on one chip of silicon, has led to many different types of switching circuits in IC form. Recent research in nanoelectronics has introduced the concepts of resonant tunneling diodes, electronic quantum cellular automata, single electron transistors, and molecular electronic devices.

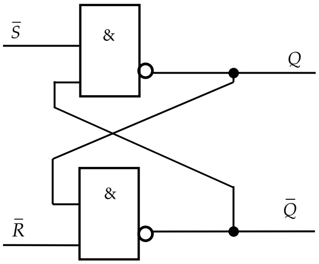

$Q(t)$	$Q(t+\Delta t)$	S	R
0	0	0	×
0	1	1	0
1	0	0	1
1	1	×	0

Figure 3. The RS flip-flop with NAND gates (top) and its truth-table (bottom). Its characteristic equation is $Q(t+\Delta t) = \overline{\overline{S}\cdot\overline{Q}(t)} = \overline{\overline{S}\cdot\overline{Q(t)\cdot\overline{R}}} = S + Q(t)\cdot\overline{R}$. Here we used de Morgan's laws: for any two subsets A and B of the U, we have $\overline{A\cdot B} = \overline{A} + \overline{B}$ and $\overline{A+B} = \overline{A}\cdot\overline{B}$.

The basic switching circuits are called gates. Often NAND and NOR gates are used, as they are easier to implement. The RS flip-flop can be designed simply by interconnecting a pair of two-input NAND gates, with appropriate feedback, as shown in Figure 3. The truth-table in Figure 3 gives input values, R and S, that condition changes from a present state $Q(t)$ to a next state $Q(t+\Delta t)$. It is constructed by considering the physical action of the circuit shown in the same figure. For example, to stay in state 1, the equivalent to input condition 1 in Figure 1 is $R = 0$, and S need not be defined; its value is ×. The entries marked with × correspond to "not allowed" or "don't care" inputs because, under these conditions, when both R and S are present simultaneously, the operation of the circuit becomes uncertain.

The Δt is a delay that occurs between the present state and the next state. This delay is essential in the operation of sequential circuits. With respect to the delay, two types of sequential circuits can be distinguished: synchronous and asynchronous. The synchronous sequential circuits trigger on receiving a certain clock pulse. The maximum frequency of the clock is defined by the operational time of the slowest element in the circuit. State transitions in a synchronous sequential circuit thus occur with a constant frequency. The delay in asynchronous sequential circuits changes from transition to transition, i.e. $\Delta t \neq$ const.

The RS flip-flop, introduced above, is a bistable multivibrator. It stays in one of two states that can be changed only when an external input is applied. If there are no external inputs, but one state causes a transition to the other state and this repeats continuously, the multivibrator is said to be astable. It is not stable in either state and alternates between them at a specific frequency. The astable multivibrator is, in fact, a free-running oscillator. The frequency is usually determined by a capacitor placed between the two gates and a resistor in parallel with each gate. Thus the delay, Δt, and consequently the frequency, are determined by the values of the capacitor and resistors.

In general, every oscillator contains a positive and negative feedback loop at the same time. For example, a unit with one excitatory and one inhibitory neuron that are mutually connected is a neuronal oscillator. The model proposed by Wilson and Cowan (*See* **Inhibition**), who introduced an inhibitory neuron and by this a positive along with the negative feedback loop, is today known as the Wilson–Cowan oscillator. As the exchange of energy and matter is continuous in every biological process that is based on existence of both positive and negative feedback loops, an oscillator can be seen as a basic biological unit.

Flip-flop circuits are today used as a general paradigm for bistability in a very wide range of nonlinear systems, one of many examples being the phenomenon of stochastic resonance (Fauve & Heslot, 1983).

ANETA STEFANOVSKA

See also **Coupled oscillators; Feedback; Inhibition; State diagrams**

Further Reading

Fauve, S. & Heslot, F. 1983. Stochastic resonance in a bistable system. *Physics Letters* A, 97: 5–7

Shannon, C.E. 1938. A symbolic analysis of relay and switching circuits. *Transactions of the AIEE*, 57: 713–723

Shannon, C.E. 1949. The synthesys of two-terminal switching circuits. *Bell Systems Technical Journal*, 28: 59–98

FLOQUET THEORY

See **Periodic spectral theory**

FLUCTUATION-DISSIPATION THEOREM

In the 19th century, there were two schools of thought on the existence of atoms, those that believed in them and those that did not. Remarkably, it was not until the early 20th century that the experiments of Jean Perrin—summarized in his 1913 book *Atoms* (Perrin, 1913)—established once and for all the existence of atoms. One of the laws Perrin experimentally verified was the fluctuation-dissipation relation, which interrelates the physical notions of randomness (through fluctuations) and determinism (through dissipation) (Montroll & West, 1979).

No matter how carefully experiments are done, they never yield the same value of a physical observable from one measurement to the next. The collection of values from such measurements is called an ensemble, and the number of times a measurement falls in an assigned interval divided by the total number of measurements in the ensemble yields a probability, which when all the intervals are taken together yields a probability distribution function. The best representation of this ensemble of measurements is the mode of the distribution (or ensemble average) as first noted by Carl Friedrich Gauss (1809). The deviations of a physical observable from its average value are called fluctuations, which are typically small and random. In physical systems, the source of these fluctuations is thermal agitation of the atoms. Thus, the thermodynamic properties of an equilibrium physical system are determined by the probability density, through the averages, and not by the instantaneous values of the positions and velocities of the atoms.

The changes in macroscopic physical systems over time comprise a combination of deterministic dynamics and microscopically induced macroscopic fluctuations. Albert Einstein, in 1905, was the first to fully appreciate the influence of these fluctuations on macroscopic transport phenomena in his investigations of the phenomenon of equilibrium diffusion (Furth, 1956). Einstein showed that the strength of the fluctuations, as measured by their mean-square level through the diffusion coefficient, D, is directly proportional to the temperature of the ambient fluid, $kT/2$, which is the average kinetic energy per degree of freedom of the ambient fluid particles. The constant of proportionality between the temperature and the diffusion coefficient is the dissipation time per unit mass, λ/m, thereby inter-relating the fluctuations and dissipation of the medium through the particle's motion

$$D = \frac{2\lambda kT}{m}. \tag{1}$$

Equation (1) was the first fluctuation-dissipation relation and demonstrates that macroscopic fluctuations and dissipation have the same microscopic origin.

Two decades before Einstein's analysis of diffusion, Walther Nernst (1884) investigated the combined process of mobility and diffusion to determine the size of the charge on an individual particle. He established that the mobility is proportional to the ratio of the diffusion coefficient to the temperature, and the charge of the particle is the proportionality constant. This ratio is usually called the Einstein equation even though Nernst discovered it. One can more generally define the mobility of a particle as the terminal velocity per unit force, through a generalized Nernst relation

$$D = kT \times \text{mobility} \tag{2}$$

for any sort of particle that is free to move but is subject to frictional drag. The linear transport coefficient (here the mobility) was only partially understood until Harry Nyquist (in considering the fluctuations in the current flowing through an electrical resistance) showed that the circuit impedance can be used to compute the fluctuations arising from the thermal agitation of the electrons (Nyquist, 1928). Nyquist's form of the fluctuation-dissipation theorem gives the mean-square voltage fluctuations as proportional to the product of the temperature and resistance, with the proportionality constant being the bandwidth. Alternatively, the temperature can be written as the ratio of the spectral density of the random fluctuations of the electromotive force $S_e(\omega)$ to the real part of the impedance $Z(\omega)$:

$$S_e(\omega) = 2kT \,\mathrm{Re}\, Z(\omega) \tag{3}$$

at the frequency ω. When the Nyquist relation (3) was first derived, the applicability of the underlying reasoning to other linear transport processes involving thermal noise was not appreciated.

Let us consider the dynamical equation constructed by Paul Langevin concerning the forces acting on a particle in a fluid (Langevin, 1908). The equation of motion for a particle of mass m is

$$m\frac{\mathrm{d}u}{\mathrm{d}t} + \frac{u}{\mu} = F + f, \tag{4}$$

where u is the particle velocity, μ is the mobility (the inverse of the dissipation), F is an external driving force, and f is a fluctuating force with zero mean produced by the thermal agitation of the ambient fluid particles. The Fourier transform of Equation (4) yields the impedance for the particle $Z(\omega) \equiv \bar{F}(\omega)/\bar{u}(\omega) = \mathrm{i}m\omega + \mu^{-1}$. The average of the transformed equation yields the Nernst relation for a constant external force: $\langle \bar{u}(0)\rangle = \mu F$. The diffusion coefficient is given by the integral over the velocity autocorrelation function, which using Parseval's theorem can be written in terms of the velocity spectral density, $S_u(\omega)$, as

$$D = 2S_u(0). \tag{5}$$

This expression for the diffusion coefficient is valid for any form of the velocity autocorrelation function and velocity spectral density, under the condition that the underlying process is stationary in time.

In general, one can conclude that the autocorrelation function of the random force in a physical system is proportional to the dissipation in that system, with the proportionality constant given by the temperature kT. This relation can be summarized as

$$\langle f(t)f(t')\rangle = 2D\delta(t - t'), \tag{6}$$

so that the diffusion coefficient determines the strength of the fluctuations. Further, there are time-dependent generalizations of this form of the fluctuation–dissipation relation, where the δ function is replaced with a memory kernel, as well as extensions from the classical to the quantum domain, see, for example, Lindenberg & West (1990) for a review. However, each elaboration contains essentially the same information, namely, that microscopic dynamics are amplified to macroscopic fluctuations and dissipation. A complete understanding of the macroscopic phenomena of fluctuations and dissipation therefore requires an understanding of microscopic dynamics.

BRUCE J. WEST

See also **Brownian motion; Diffusion; Fokker–Planck equation**

Further Reading

Furth, R. (editor). 1956. *Albert Einstein, Investigations on the Theory of Brownian Motion*, New York: Dover; originally published in German, 1926

Gauss, F. 1809. *Theoria Motus Corporum Coelestrium*, Hamburg:

Langevin, P. 1908. Comptes Rendus Acad. Sci. Paris, 530

Lindenberg, K. & West, B.J. 1990. *The Nonequilibrium Statistical Mechanics of Open and Closed Systems*, New York: VCH

Montroll, E.W. & West, B.J. 1979. An enriched collection of stochastic processes. In *Fluctuation Phenomena*, edited by E.W. Montroll & Lebowitz, J. Amsterdam: North-Holland

Nernst, W. 1884. *Zeitschrift für Physikalische Chemie*, 9: 613

Nyquist, H. 1928. Thermal agitation of electric charge in conductors. *Physical Review*, 32: 110

Perrin, J. 1990. *Atoms*, Woodbridge: Ox Bow Press; originally published as *Les Atomes* in French, 1913

FLUID DYNAMICS

The field of fluid dynamics is devoted to the study of flows of matter in the liquid or gaseous state or in the form of multiple phases including suspensions of solid particles in liquids or gases. While it is in principle possible (and sometimes appropriate) to apply Newton's laws to individual molecules of a fluid and to describe the flow as an average over trajectories of the particles (*See* **Molecular dynamics**), it is usually far more efficient to consider liquids and gases as a continuum and to apply the equations for the dynamics of continuous media.

Conservation Laws of Continuous Media

The laws of the conservation of mass, momentum, and energy can be written in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho\,\mathrm{d}^3V = \int_V \frac{\partial\rho}{\partial t}\,\mathrm{d}^3V = -\oint_{\partial V} \rho n_j v_j\,\mathrm{d}^2A, \tag{1a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho v_i\,\mathrm{d}^3V = -\oint_{\partial V} \rho v_i n_j v_j\,\mathrm{d}^2A, + \int_V \rho F_i\,\mathrm{d}^3V + \oint_{\partial V} S_{ij}n_j\,\mathrm{d}^2A, \tag{1b}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_V &\rho(e+v_jv_j/2)\,\mathrm{d}^3V \\ &= -\oint_{\partial V} v_jn_j\rho(e+v_iv_i/2)\mathrm{d}^2A \\ &+ \int_V \rho F_iv_i\mathrm{d}^3V + \int_V \rho q\,\mathrm{d}^3V \\ &+ \oint_{\partial V} v_iS_{ij}n_j\,\mathrm{d}^2A - \oint_{\partial V} h_jn_j\,\mathrm{d}^2A,\end{aligned} \tag{1c}$$

where V denotes an arbitrary volume fixed in space and ∂V denotes its surface. ρ is the density of the continuous medium, v_i is the velocity vector, and ρF_i is the body force density acting on the medium. A typical example for F_i is the acceleration of gravity. e is the internal energy per unit mass, ρq is the heat source density, and h_i denotes the heat flux. Contributions like absorption and emission of radiation have not been included explicitly in the energy balance (1c). But in optically thick media they can be subsumed under q and h_i. The index notation has been used where the index i refers to the coordinates 1, 2, 3 of a Cartesian system of coordinates and where the summation over indices occurring twice in any term is implied. Relativistic effects have been neglected in writing Equations (1) that are thus valid only for velocities small compared with the velocity of light.

The fundamental quantity of the dynamics of a continuous medium is the stress tensor S_{ij} which describes the surface force $S_{ij}n_j$ exerted on a surface element with the normal unit vector n_i. More exactly, $S_{ij}n_j$ is the force per unit area exerted by the material pierced by n_i on the material pierced by $-n_i$. The concept of the surface force was advanced long before the atomistic nature of materials became generally accepted. Surface forces in the ideal sense do not exist in nature. But since the forces between molecules in materials act only over atomic distances, the concept of surface forces has turned out to be very useful. The limitation should be kept in mind, however, in the fluid dynamical treatment of the dynamics of galaxies, for instance.

For applications of the conservation laws (1) it is convenient to formulate them in differential form by taking the limit $V \to 0$ and using Gauss' theorem,

$$\left(\frac{\partial}{\partial t} + v_j\partial_j\right)\rho + \rho\partial_jv_j = 0, \tag{2a}$$

$$\rho\frac{\partial}{\partial t}v_i + \rho v_j\partial_jv_i = \rho F_i + \partial_jS_{ij}, \tag{2b}$$

$$\frac{\partial}{\partial t}\rho e + \partial_j(\rho v_je) = S_{ij}\partial_jv_i + \rho q - \partial_jh_j. \tag{2c}$$

In deriving (2b) from (1b) and (2c) from (1c), we have used relationships (2a) and (2b), respectively. While Equations (2) apply to all continua, including solid materials, for applications to fluids the stress tensor is separated into two parts,

$$S_{ij} = -p\delta_{ij} + S'_{ij}, \tag{3}$$

where p is the thermodynamic pressure and S'_{ij} is the part of the stress tensor which depends on viscous friction and other dissipative effects. Then $S'_{ij}\partial_j v_i \equiv \rho\Phi$ is the irreversible conversion of work into heat. After using (2b), we may write (2c) in the form

$$\frac{\mathrm{D}e}{\mathrm{D}t} + p\frac{\mathrm{D}\rho^{-1}}{\mathrm{D}t} = \Phi + q - \rho^{-1}\partial_j h_j, \tag{4}$$

where the material derivative

$$\frac{\mathrm{D}}{\mathrm{D}t} \equiv \frac{\partial}{\partial t} + v_j\partial_j \tag{5}$$

has been introduced that describes the change in time with respect to the frame of reference moving with the fluid. This is also referred to as the Lagrangian description in contrast to the Eulerian description where the time derivative at a fixed point in space is taken.

In applications, the pressure p and the temperature T are usually the most readily known thermodynamic variables. We thus use the relationship $e = h - p/\rho$ for simple materials where h is the specific enthalpy and obtain

$$\mathrm{d}h = c_p\,\mathrm{d}T + \left.\frac{\partial h}{\partial p}\right|_T \mathrm{d}p = c_p\,\mathrm{d}T + \frac{1}{\rho}(1-\alpha T)\,\mathrm{d}p, \tag{6}$$

where some simple thermodynamic relationships have been used and $\alpha \equiv \rho(\partial\rho^{-1}/\partial T)_p$ is the coefficient of thermal expansion. As a final result, we thus obtain the energy equation in its most useful form

$$T\frac{\mathrm{D}s}{\mathrm{D}t} \equiv c_p\frac{\mathrm{D}T}{\mathrm{D}t} - \frac{\alpha T}{\rho}\frac{\mathrm{D}p}{\mathrm{D}t} = \Phi + q - \frac{1}{\rho}\partial_j h_j. \tag{7}$$

Fourier's law, $h_i = -\lambda\partial_i T$, is commonly used for the heat flux, and for laboratory applications, it is a good approximation to neglect $\alpha T\rho^{-1}\mathrm{D}\rho/\mathrm{D}t + \Phi$ in comparison with $c_p\,\mathrm{D}T/\mathrm{D}t$. In this case, the familiar heat equation

$$\left(\frac{\partial}{\partial t} + v_j\partial_j\right)T = q + \kappa\partial_j\partial_j T \tag{8}$$

is obtained where $\lambda =$ const. has been assumed and where $\kappa = \lambda/\rho c_p$ is the thermal diffusivity.

Dynamics of Inviscid Fluids

The simplest equation of fluid dynamics is obtained when viscous, elastic, and other effects are neglected and $S_{ij} = -p\delta_{ij}$ is assumed. The resulting Euler equation,

$$\left(\frac{\partial}{\partial t} + v_j\partial_j\right)v_i = -\rho^{-1}\partial_i p + F_i, \tag{9}$$

is most easily solved when the fluid is incompressible and $\rho =$ const. can be assumed. In that case

$$\partial_j v_j = 0 \tag{10}$$

holds and an equation of state connecting ρ and p is not needed.

A more general form of the equations is obtained in the barotropic case when the density is prescribed as a function of the pressure alone as, for instance, in the form of a polytropic relationship

$$p = p_0\left(\frac{\rho}{\rho_0}\right)^{\gamma}. \tag{11}$$

This expression is valid, for example, for adiabatic changes of an ideal gas in which case γ equals the ratio of specific heats, $\gamma = c_p/c_v$. The equation for sound waves can be obtained when Equations (2a) and (9) are linearized around the static equilibrium state with $\rho = \rho_0$, $p = p_0$,

$$\frac{\partial^2}{\partial t^2}\rho = \gamma\frac{p_0}{\rho_0}\partial_j\partial_j\rho, \tag{12}$$

where $\gamma p_0/\rho_0 = c_\mathrm{s}^2$ is the square of the speed of sound. For general fluids, the relationship $c_\mathrm{s}^2 = \partial p/\partial\rho$ holds. The ratio of a characteristic velocity divided by the sound speed is called the Mach number Ma. The domains of fluid flow with $\mathrm{Ma} > 1$ are usually separated from regions with $\mathrm{Ma} < 1$ by shock fronts in which energy is dissipated and frictional processes must be taken into account.

A fundamental consequence of Equation (9) is Kelvin's theorem,

$$\frac{\mathrm{D}}{\mathrm{D}t}\oint_C \boldsymbol{v}\cdot\mathrm{d}\boldsymbol{l} = 0, \tag{13}$$

where the integral is called the circulation and must be taken over a closed curve C moving with the fluid. Theorem (13) holds for barotropic fluids when the force field F_i is conservative; that is, it can be written as the gradient of a single-valued scalar function. The property that the circulation is an invariant of fluid motion represents an elegant formulation of Helmholtz's vortex theorem, which states that the strength of a vortex tube moving with the fluid remains unchanged,

$$\frac{\mathrm{D}}{\mathrm{D}t}\int_S \nabla\times\boldsymbol{v}\cdot\mathrm{d}^2\boldsymbol{A} = 0, \tag{14}$$

where the surface S is bounded by a closed material curve, C. $\nabla\times\boldsymbol{v}$ is called the vorticity field of the velocity field $\boldsymbol{v}$. The manifold of vorticity vectors intersecting the curve C generates as tangential vectors the surface of the vortex tube. Kelvin's theorem and Helmhotz's theorem are connected, of course, by Stokes' theorem.

An important consequence of theorem (13) or (14) is that motions starting in a fluid from rest have vanishing vorticity if the force field is conservative and viscous effects can be neglected. An incompressible fluid with vanishing vorticity obeys the equation

$$\boldsymbol{v} = \nabla\phi \quad \text{with} \quad \nabla^2\phi = 0. \tag{15}$$

Such flows are called potential flows.

Dynamics of Viscous Fluids

In the general case, the stress tensor S'_{ij} may be a rather complex function of the properties of the fluid and its motions. The various problems arising in this connection are the subject of the field of rheology. Fortunately, the most important fluids such as air and water and many others can be described as Newtonian fluids for which a linear homogeneous relationship holds between S'_{ij} and the velocity gradient tensor $\partial_i v_j$,

$$S'_{ij} = \mu(\partial_i v_j + \partial_j v_i - \tfrac{2}{3}\delta_{ij}\partial_k v_k) + \tilde{\mu}\delta_{ij}\partial_k v_k, \tag{16}$$

where μ is the dynamic viscosity and $\tilde{\mu}$ is the bulk viscosity. It is difficult to measure the latter property. Because $\tilde{\mu} = 0$ can be shown to hold for mono-atomic gases, this relationship is often generally applied in which case the Navier–Stokes equations are obtained,

$$\begin{aligned}\rho\left(\frac{\partial}{\partial t} + v_j\partial_j\right) v_i &= -\partial_i p + \rho F_i \\ &\quad + \partial_j\left(\mu(\partial_i v_j + \partial_j v_i - \tfrac{2}{3}\delta_{ij}\partial_k v_k)\right).\end{aligned} \tag{17}$$

For a given force field F_i, this equation together with the equation of continuity (2a) and a barotropic equation of state, $\rho = \rho(p)$, provides a complete set for the determination of the variables v_i, ρ, p (*See* **Navier–Stokes equations**).

An even simpler set of equations can be obtained for incompressible fluids with $\rho =$ const., $\mu =$ const.,

$$\left(\frac{\partial}{\partial t} + v_j\partial_j\right) v_i = -\partial_i \frac{p}{\rho} + F_i + \nu\partial_j\partial_j v_i, \tag{18a}$$

$$\partial_j v_j = 0, \tag{18b}$$

where the kinematic viscosity $\nu = \mu/\rho$ has been introduced. Using U as typical velocity and d as a typical length scale, we find for the ratio of the last terms on the left- and right-hand sides of Equation (18a)

$$\frac{Ud}{\nu} = \mathrm{Re} \tag{19}$$

which is called the Reynolds number. In the low Reynolds number limit, the Stokes equation is obtained in which the term $v_j\partial_j v_i$ in Equation (18a) is dropped. In the high Reynolds number limit, the Euler equation does not become applicable, in general, because of the existence of a turbulent cascade where energy is transferred to smaller and smaller scales such that the term $\nu\partial_j\partial_j v_i$ cannot be neglected in the limit $\nu \to 0$ (*See* **Turbulence**).

Buoyancy-driven Flows

Among the force fields that are not conservative and therefore tend to generate motions with vorticity, buoyancy forces are the most common ones. Especially in atmospheric, oceanic, and astrophysical applications, buoyancy-driven flows are of fundamental importance. In order to keep as much as possible of the convenient properties of the Navier–Stokes equations for incompressible fluids, the Boussinesq approximation is usually introduced in which the temperature dependence of the density, say $\rho = \rho_0(1 - \gamma(T - T_0))$, is taken into account in the gravity force term only, while ρ is replaced by ρ_0 elsewhere in the equations. Together with the heat equation, the following set of equations is obtained:

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v}\cdot\nabla\right)\boldsymbol{v} = -\nabla\pi - \gamma(T - T_0)\boldsymbol{g} + \nu\nabla^2\boldsymbol{v}, \tag{20a}$$

$$\nabla\cdot\boldsymbol{v} = 0, \tag{20b}$$

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v}\cdot\nabla\right) T = \kappa\nabla^2 T, \tag{20c}$$

where $\boldsymbol{g}$ is the gravity vector and where $\nabla\pi$ includes all terms that can be written as a gradient. Because the pressure dependence of the density is neglected in the Boussinesq approximation, the contribution Φ must also be neglected (Busse, 1989).

A simple solution described by these equations is the flow between two parallel plates with separation d that are kept at different temperatures and that are inclined at an angle χ with respect to the horizontal plane (see Figure 1).

$$\boldsymbol{v} = z(z^2 - d^2/4)\gamma g \sin\chi(T_2 - T_1)\boldsymbol{i}, \tag{21a}$$

$$T = T_0 + (T_2 - T_1)z/d. \tag{21b}$$

A Cartesian system of coordinates has been assumed with the z-coordinate normal to the plates and the origin in the middle between the plates. $\boldsymbol{i}$ is the unit vector in the direction of inclination parallel to the x-axis as shown in Figure 1. T_1 and T_2 are the fixed temperatures at the boundaries $z = -0.5$ and $+0.5$, respectively, and $T_0 = (T_2 + T_1)/2$. Since in laboratory realizations of this configuration the space between the plates will be enclosed by side walls, the condition that the total mass transport between the plates vanishes has been imposed. Of special interest is the case $\gamma(T_2 - T_1) < 0$

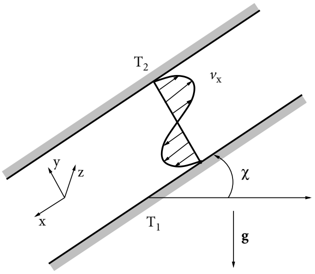

Figure 1. Inclined fluid layer heated from above.

when the density increases opposite to the direction of gravity such that an unstable stratification results. In a horizontal layer where solution (21a) vanishes, the instability occurs in the form of Rayleigh–Bénard convection. The dimensionless control parameter for this case, $\mathrm{Ra} \equiv \gamma g(T_2 - T_1)d^3/\kappa\nu$, is called the Rayleigh number. Convection rolls set in when Ra exceeds a critical value $\mathrm{Ra_c}$ depending on the boundary conditions. Besides buoyancy driven rolls aligned with the x-axis for $\chi \neq 0$ hydrodynamic, instabilities may set in in the form of transverse vortices. Since these forms of secondary states of motion are themselves subject to secondary instabilities, sequences of bifurcations are observed leading to increasingly complex patterns of fluid flow as indicated in Figure 2 for a few examples. The variety of patterns that can be realized continue to be a subject of intense experimental and theoretical research (*See* **Taylor–Couette flow**).

A role similar to that of buoyancy is played by the temperature dependence of surface tension that may give rise to instabilities much like Rayleigh–Bénard convection. Instead of the parameter Ra, the Marangoni number $M = \xi(T_2 - T_1)d/\kappa\eta$ is used as a control parameter, in this case where $\xi = \partial\sigma/\partial T$ denotes the derivative with respect to temperature of the surface tension σ.

Dynamics of Rotating Fluids

The Navier–Stokes equations (17) are invariant with respect to a Galileo transformation; that is, they retain their form in the transition from one inertial frame of reference to another one. When the transformation is made from an inertial to a frame of reference rotating with the constant angular velocity $\boldsymbol{\Omega}$, it must be taken into account that the transformation of a vector field $\boldsymbol{a}$ is different from that of a scalar field ϕ,

$$\frac{\partial}{\partial t'}\boldsymbol{a} = \left(\frac{\partial}{\partial t} + \boldsymbol{\Omega}\times\boldsymbol{r}\cdot\nabla - \boldsymbol{\Omega}\times\right)\boldsymbol{a}, \qquad (22a)$$

$$\frac{\partial}{\partial t'}\phi = \left(\frac{\partial}{\partial t} + \boldsymbol{\Omega}\times\boldsymbol{r}\cdot\nabla\right)\phi; \qquad (22b)$$

that is, a constant vector field in an inertial frame becomes time dependent as seen from the rotating frame. Using $\boldsymbol{v}' = \boldsymbol{v} - \boldsymbol{\Omega}\times\boldsymbol{r}$ where $\boldsymbol{r}$ is the position vector, we find as Navier–Stokes equation in the rotating frame

$$\left(\frac{\partial}{\partial t'} + \boldsymbol{v}'\cdot\nabla\right)\boldsymbol{v}' + 2\boldsymbol{\Omega}\times\boldsymbol{v}' = -\nabla\frac{p}{\rho} - \nabla\Psi + (\boldsymbol{\Omega}\times\boldsymbol{r})\times\boldsymbol{\Omega} + \nu\nabla^2\boldsymbol{v}' \qquad (23a)$$

$$\nabla\cdot\boldsymbol{v}' = 0. \qquad (23b)$$

We have restricted our attention to the incompressible case with $\rho =$ const. and have assumed that the force field $\boldsymbol{F}$ is conservative, that is, $\boldsymbol{F} = -\nabla\Psi$. Because the centrifugal force is also conservative it can be combined with $\boldsymbol{F}$; that is, $\Psi' = \Psi - |\boldsymbol{\Omega}\times\boldsymbol{r}|^2/2$ can be used as the potential. The other new term in Equations (23) is the Coriolis force, $-2\boldsymbol{\Omega}\times\boldsymbol{v}'$, which is responsible for the fact that the dynamics of rotating systems is quite different from that of nonrotating systems. Since the Coriolis force is typically the largest term on the left-hand side of Equation (23a) and since viscous friction is usually negligible in the interior of the fluid, the approximate relationship

$$2\boldsymbol{\Omega}\times\boldsymbol{v} = -\nabla\pi \qquad (24)$$

is obtained which is called the geostrophic balance because of its importance in meteorology and oceanography. Here, $\rho\pi$ is the dynamic pressure, $\pi = p/\rho + \Psi'$, and for simplicity, the prime of $\boldsymbol{v}$ has been dropped. When the curl of balance (24) is taken and Equation (23b) is used, the famous Proudman–Taylor theorem is obtained,

$$2\boldsymbol{\Omega}\cdot\nabla\boldsymbol{v} = 0. \qquad (25)$$

This theorem states that in a rapidly rotating system, a steady velocity field of a nearly inviscid fluid must be independent of the coordinate in the direction of the axis of rotation. As a consequence, $\boldsymbol{v}$ will depend only on the x-, y-coordinates of a cartesian system with the z-coordinate pointing in the direction of $\boldsymbol{\Omega}$. Since $\boldsymbol{v}$ must also satisfy $\nabla\cdot\boldsymbol{v} = 0$, a single scalar function, the stream function ψ, is sufficient to describe $\boldsymbol{v}$. Assuming boundaries perpendicular to $\boldsymbol{\Omega}$, we find

$$\boldsymbol{v} = \nabla\psi\times\boldsymbol{\Omega}/\Omega \quad \text{with} \quad \psi = -\pi(x, y)/2\Omega$$

and

$$\Omega = |\boldsymbol{\Omega}|. \qquad (26)$$

The form of the velocity field indicates that the flow parallels the isobars in contrast to the situation in a nonrotating system where the flow is usually directed perpendicular to the pressure gradient. The

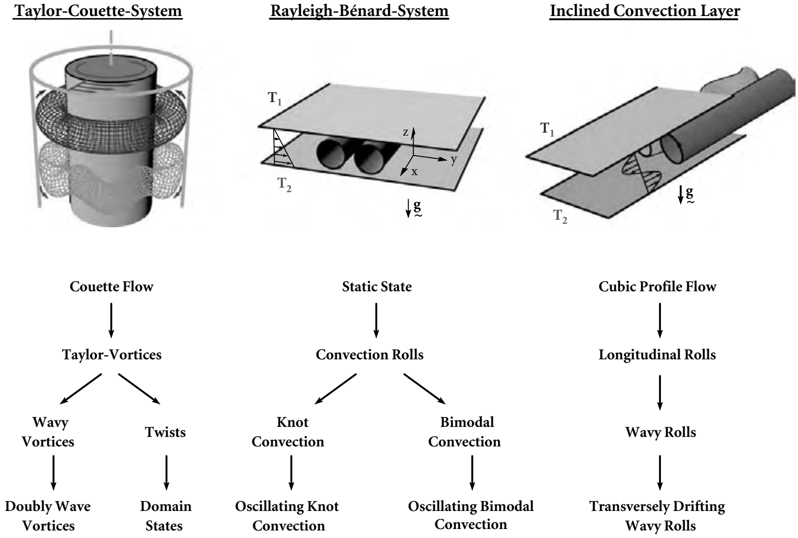
Taylor-Couette-System
Rayleigh-Bénard-System
Inclined Convection Layer
T1
T2
z
y
x
g
T1
T2
g
Couette Flow
Taylor-Vortices
Wavy Vortices
Twists
Doubly Wave Vortices
Domain States
Static State
Convection Rolls
Knot Convection
Bimodal Convection
Oscillating Knot Convection
Oscillating Bimodal Convection
Cubic Profile Flow
Longitudinal Rolls
Wavy Rolls
Transversely Drifting Wavy Rolls

Figure 2. Examples of bifurcation sequences.

geostrophic wind (26) flows cyclonically (i.e., in the sense of rotation) around low-pressure regions and anticyclonically around high-pressure regions, as can be seen on weather maps.

The fact that the function $\pi(x, y)$ has remained undetermined so far is known as the geostrophic degeneracy. Terms of higher order in the equations of motion are needed to remove this degeneracy and to introduce time dependences on scales much longer than the rotation period $2\pi/\Omega$ of the system. A three-dimensional nature of the velocity field enforced, for instance, by boundary conditions can thus be introduced as a perturbation of the geostrophic solution. Rossby waves are a typical example for the quasigeostrophic dynamical response of rotating fluid systems.

The friction term in Equation (23a) is characterized by the highest spatial derivatives and thus can become large wherever strong variations of the velocity field occur, as, for example, near solid boundaries where the tangential velocity component must vanish. Thin Ekman boundary layers are formed here with the typical thickness $\sqrt{\nu/\Omega}$. But internal shear layers in the interior of the fluid can also be realized. The Ekman number defined by

$$E = \nu/\Omega d^2, \tag{27}$$

where d is a typical length scale of the system in the direction of the axis of rotation, is often used as a dimensionless parameter in the dynamics of rotating fluids.

F.H. Busse

See also **Boundary layers; Molecular dynamics; Navier–Stokes equation; Taylor–Couette flow; Thermal convection; Turbulence; Vortex dynamics of fluids**

Further Reading

Acheson, D.J. 1990. *Elementary Fluid Dynamics*, Oxford: Clarendon Press and New York: Oxford University Press

Batchelor, G.K. 1967. *An Introduction to Fluid Dynamics*, Cambridge: Cambridge University Press

Busse, F.H. 1989. Fundamentals of Thermal Convection. In *Mantle Convection*, edited by W. Peltier, London: Gordon and Breach

Greenspan, H.P. 1968. *The Theory of Rotating Fluids*, Cambridge and New York: Cambridge University Press

Swinney, H.L. & Gollub, J.P. (editors). 1985. *Hydrodynamic Instabilities and Transition to Turbulence*, 2nd edition, Berlin and New York: Springer

Tritton, D.J. 1988. *Physical Fluid Dynamics*, 2nd edition, Oxford: Clarendon Press and New York: Oxford University Press

FLUX FLOW OSCILLATOR

See **Long Josephson junction**

FLUX QUANTIZATION

See **Josephson junction**

FLUXON

See **Long Josephson junction**

FOCUSING SYSTEM

See **Nonlinear optics**

FOKKER–PLANCK EQUATION

The Fokker–Planck equation (FPE) plays a role in stochastic systems analogous to that of the Liouville equation in deterministic mechanical systems. Namely, the FPE describes in a statistical sense how a *collection* of initial data evolves in time. To be precise, let the phase space describing the system be parameterized by the state vector $\boldsymbol{y}$, and let $\boldsymbol{Y}(t)$ denote the value of the state vector assumed by the system at time t. Suppose the dynamics of the system is governed by a stochastic differential equation (SDE) of the form

$$\mathrm{d}\boldsymbol{Y}(t) = \boldsymbol{U}(\boldsymbol{Y}(t), t)\,\mathrm{d}t + \sigma(\boldsymbol{Y}(t), t) \cdot \mathrm{d}\boldsymbol{W}(t). \quad (1)$$

The deterministic component of the dynamics is described by the drift vector $\boldsymbol{U}(\boldsymbol{y}, t)$, while the random part is driven by a vector of independent Brownian motions $\boldsymbol{W}(t)$ that are coupled to the system through the matrix $\sigma(\boldsymbol{y}, t)$.

Alternatively, we can define the system dynamics as a "diffusion process" in the relevant phase space such that

$$\begin{cases} \lim_{\tau\downarrow 0} \tau^{-1}\langle \boldsymbol{Y}(t+\tau) - \boldsymbol{Y}(t)\rangle = \boldsymbol{U}(\boldsymbol{Y}(t), t), \\ \lim_{\tau\downarrow 0} \tau^{-1}\langle (\boldsymbol{Y}(t+\tau) - \boldsymbol{Y}(t)) \\ \quad \otimes (\boldsymbol{Y}(t+\tau) - \boldsymbol{Y}(t))\rangle = 2\mathsf{D}(\boldsymbol{Y}(t), t), \\ \lim_{\tau\downarrow 0} \tau^{-1}\langle |\boldsymbol{Y}(t+\tau) - \boldsymbol{Y}(t)|^{\gamma}\rangle = 0 \text{ for } \gamma > 2, \end{cases} \quad (2)$$

where $\langle\cdot\rangle$ denotes an average over the random component of the dynamics. The SDE and diffusion process descriptions are equivalent, with $2\mathsf{D} = \sigma\sigma^{\dagger}$.

One useful way to describe the evolution of stochastic systems over a finite (rather than infinitesimal) time interval is through the probability transition density $p(t, \boldsymbol{y}|t', \boldsymbol{y}')$, which describes the likelihood that if the state variable assumes the value $\boldsymbol{y}'$ at time t', then it will assume a value near $\boldsymbol{y}$ at the later time t. More precisely, p is defined in terms of conditional probability as

$$\mathrm{Prob}(\boldsymbol{Y}(t) \in B|\boldsymbol{Y}(t') = \boldsymbol{y}') = \int_B p(t, \boldsymbol{y}|t', \boldsymbol{y}')\,\mathrm{d}\boldsymbol{y} \quad (3)$$

for all nice (Borel) sets B in the phase space. The FPE, also known as the Kolmogorov forward equation, is a partial differential equation that describes how the probability transition density evolves when the system can be described as an SDE (1) or diffusion process (2):

$$\begin{cases} \dfrac{\partial p(t, \boldsymbol{y}|t', \boldsymbol{y}')}{\partial t} \\ \quad = \dfrac{\partial}{\partial \boldsymbol{y}}\left(-\boldsymbol{U}(\boldsymbol{y}, t)p + \dfrac{\partial}{\partial \boldsymbol{y}}(\mathsf{D}(\boldsymbol{y}, t)p)\right), \\ p(t = t', \boldsymbol{y}|t', \boldsymbol{y}') = \delta(\boldsymbol{y} - \boldsymbol{y}'). \end{cases} \quad (4)$$

This equation is to be solved for $t > t'$ with fixed data for the source variables $(t', \boldsymbol{y}')$ and appropriate boundary conditions in $\boldsymbol{y}$ (Risken, 1989, Chapter 4).

The probability transition density can be used to describe the probability density for the state vector at any moment of time:

$$\mathrm{Prob}(\boldsymbol{Y}(t) \in B) = \int_B \phi(t, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \quad (5)$$

by simply integrating against the prescribed probability distribution of the states at the initial time t':

$$\phi(t, \boldsymbol{y}) = \int p(t, \boldsymbol{y}|t', \boldsymbol{y}')\phi(t', \boldsymbol{y}')\,\mathrm{d}\boldsymbol{y}'. \quad (6)$$

Alternatively, $\phi(t, \boldsymbol{y})$ can be shown to satisfy the FPE (4) but with initial data $\phi(t', \boldsymbol{y}')$ prescribed more generally as a nonnegative function with integral one rather than as a delta function.

The FPE was first applied to describe Brownian motion. In the most idealized case, where inertia is completely neglected, Einstein showed that the statistics of the position $\boldsymbol{x} = \boldsymbol{X}(t)$ of a Brownian particle obeys an FPE that coincides with the ordinary diffusion equation:

$$\frac{\partial \phi(t, \boldsymbol{x}|t', \boldsymbol{x}')}{\partial t} = \kappa \Delta_{\boldsymbol{x}}\phi, \quad (7)$$

where $\Delta_{\boldsymbol{x}}$ is the Laplace operator and the diffusion coefficient $\mathsf{D} = \kappa\mathsf{I}$ is a constant scalar multiple of the identity matrix. The stochastic differential description of this model is simply

$$\mathrm{d}\boldsymbol{X}(t) = (2\kappa)^{1/2}\,\mathrm{d}\boldsymbol{W}(t). \quad (8)$$

The effects of inertia and external forces can be incorporated by passing to a phase space description including both the position $\boldsymbol{x} = \boldsymbol{X}(t)$ and the velocity $\boldsymbol{v} = \boldsymbol{V}(t)$ of the Brownian particle. The equations of motion can then be written in terms of Newton's law with a random forcing component proportional to $\mathrm{d}\boldsymbol{W}$:

$$\begin{cases} \mathrm{d}\boldsymbol{X}(t) = \boldsymbol{V}(t)\,\mathrm{d}t, \\ m\,\mathrm{d}\boldsymbol{V}(t) = -m\xi\boldsymbol{V}(t)\,\mathrm{d}t + \boldsymbol{f}(\boldsymbol{X}(t), t)\,\mathrm{d}t \\ \qquad\qquad + (2k_{\mathrm{B}}Tm\xi)^{1/2}\,\mathrm{d}\boldsymbol{W}(t), \end{cases} \tag{9}$$

where ξ is a friction coefficient, k_{B} is Boltzmann's constant, T is the absolute temperature, and $\boldsymbol{f}(\boldsymbol{x}, t)$ represents the deterministic part of the external applied force. The equivalent Fokker–Planck description for the phase space probability transition density in this system reads

$$\begin{aligned} &\frac{\partial p(t, \boldsymbol{x}, \boldsymbol{v}|t', \boldsymbol{x}', \boldsymbol{v}')}{\partial t} \\ &\quad = \mathbf{grad}_{\boldsymbol{v}} \cdot \left[\left(\xi\boldsymbol{v} - \frac{\boldsymbol{f}(\boldsymbol{x}, t)}{m} + \frac{k_{\mathrm{B}}T\xi}{m}\mathbf{grad}_{\boldsymbol{v}}\right) p\right] \\ &\qquad -\mathbf{grad}_{\boldsymbol{x}} \cdot (\boldsymbol{v}p) \end{aligned} \tag{10}$$

Observe how this equation generalizes the Liouville equation for deterministic mechanics to include Brownian motion. The simplified diffusion equation (7) can be obtained through an asymptotic limit of the full FPE (10) with $\kappa = k_{\mathrm{B}}T/(m\xi)$, when $\boldsymbol{f}(\boldsymbol{x}, t) = 0$ and $(k_{\mathrm{B}}T/(m\xi^2\ell^2))^{1/2} \ll 1$ with ℓ a characteristic length of the system (Bocquet, 1997; Risken, 1989). A similar reduction to a partial differential equation in coordinate space, called the Smoluchowski equation, is also possible for collections of interacting particles where the friction tensor depends on the particle configuration (Titulaer, 1980). From the FPE for Brownian motion, one can compute various statistical properties such as its mean-square displacement, the spectrum of its fluctuations, the rate of relaxation of initial velocity to thermal equilibrium values, and the probability distribution for the time at which the particle first achieves a certain location or surmounts a potential barrier (Wax, 1954).

The FPE has found useful application in computing similar statistical properties in numerous other systems, such as lasers, polymers, particle suspensions, quantum electronic systems, molecular motors, and finance. In some instances, the system is not at first formulated in one of the senses (1) or (2) which immediately imply an FPE. Rather, the systems are more naturally represented in terms of a master equation with a complete description of the rates at which the system (randomly) jumps from one state to another. The FPE is an appropriate approximation to this system when certain asymptotic conditions, such as small jumps, a large system size, or a separation of time scales between resolved and unresolved variables, are met (Grabert, 1982; Risken, 1989; van Kampen, 1981).

A variety of techniques for practical analysis of FPEs can be found in Risken (1989). The main alternative to analyzing stochastic systems through the FPE is through the stochastic differential (or Langevin) formulation (1). The FPE has the merit of being a deterministic partial differential equation that can be solved once through either analytical or numerical methods. The stochastic differential description (1) by contrast consists of ordinary differential equations for which individual realizations (samples) are faster to simulate through Monte Carlo methods (Kloeden & Platen, 1992). But obtaining a statistical description of the general behavior of the system requires the generation of a large number of realizations.

Certain statistical quantities described above can be easily calculated using either the solution of the FPE or by operating directly on the FPE itself. On the other hand, the stochastic differential system (1) must be used if one wishes to characterize particular realizations of the stochastic system. Moreover, the stochastic differential description can be readily generalized to cases in which the random driving has temporal correlations (Kubo et al., 1991; Moss & McClintock, 1989). The FPE formulation can be modified to a "fractional" formalism for certain self-similar temporal correlation structures (Metzler & Klafter, 2000), but further generalizations are much more complicated and usually require some sort of approximation (Moss & McClintock, 1989).

PETER R. KRAMER

See also **Brownian motion; Fluctuation-dissipation theorem; Nonequilibrium statistical mechanics; Phase-space diffusion and correlations; Stochastic processes**

Further Reading

Bocquet, L. 1997. High friction limit of the Kramers equation: The multiple time-scale approach. *American Journal of Physics*, 65(2): 140–144

Grabert, H. 1982. *Projection Operator Techniques in Nonequilibrium Statistical Mechanics*, Berlin and New York: Springer [FPE as approximation to more detailed models when separation of time scales exists]

Kloeden, P.E. & Platen, E. 1992. *Numerical Solution of Stochastic Differential Equations*, Berlin and New York: Springer [Sections 1.7, 2.4, 4.1, 4.7]

Kubo, R., Toda, M. & Hashitsume, N. 1991. *Statistical Physics*, 2 vols, 2nd edition, Berlin and New York: Springer [Chapter 2 has a derivation of hierarchy of statistical descriptions for physical systems]

Metzler, R. & Klafter, J. 2000. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1): 1–77

Moss, F. & McClintock, P.V.E. (editors). 1989. *Noise in Nonlinear Dynamical Systems*, vol. 1. Cambridge and New York: Cambridge University Press [A collection of expository articles on FPE methods with extensions to temporal correlations in noise]

Risken, H. 1989. *The Fokker–Planck Equation*, 2nd edition, Berlin: Springer

Titulaer, U.M. 1980. Corrections to the Smoluchowski equation in the presence of hydrodynamic interactions. *Physica* A, 100(2): 251–265 van Kampen, N.G. 1981. *Stochastic Processes in Physics and Chemistry*, Amsterdam: North-Holland [Chapter 8 has a careful discussion of physical applicability of FPE]

Wax, N. (editor). 1954. *Selected Papers on Noise and Stochastic Processes*, New York: Dover

FORCED SYSTEMS

See **Damped-driven anharmonic oscillator**

FORECASTING

The success of Newtonian classical mechanics led in the 18th and 19th centuries to a strong belief in determinism; that is, the notion that the past determines the present and the future. As a consequence, a superior intellect ("Laplace's Daemon") would be able to forecast the future given enough information about the present and the past. However, the discovery of chaos in a large class of deterministic nonlinear dynamical systems sets a limit as to what Laplace's Daemon can forecast.

One of the characteristics of deterministic chaos is that the future state of a dynamical system depends sensitively on the initial conditions—infinitesimal perturbations will grow exponentially. This effect is also known more poetically as the "butterfly effect" (from a lecture by Edward Lorenz in 1972: "Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?"). Since there will always be some uncertainty—however small—about the observed state of any physical system, a consequence of extreme sensitivity on the initial conditions is that the future state of such a system can only be forecast to lie within a given error tolerance a limited time ahead. As even Laplace's Daemon would not be able to determine the exact present state of the system, she would not be able to forecast the state of the future, but as she would know the uncertainty about the present, she would be able to work out all possible future states.

Uncertainty in the initial conditions is appropriately expressed in terms of a multivariate probability density function (pdf). A forecast is then given by the time evolution of the initial pdf, subject to the underlying dynamical system. Typically, the initial pdf is sharp, centered around the best estimate of the observed state of the physical system, but with time the pdf is smeared out in a complicated fashion until, eventually, all information about the initial condition is lost. For the short-range evolution, when the pdf remains "sufficiently" sharp, a forecast can be given in deterministic terms. For the intermediate-range evolution, when the pdf still has sufficient structure, a forecast can (and should) be given in probabilistic terms. Beyond that, a forecast will be no better than a guess based on the observed long-term behavior of the system.

When the governing dynamical equations are known, the time evolution of the pdf can be expressed in terms of the Liouville equation, which is a special case of the stochastic-differential Fokker–Planck equation. In theory, the Liouville equation could be applied to weather forecasting where dynamical models have been developed of the atmospheric circulation, but in practice, numerical problems and computational limits need to be overcome. An alternative approach is to make a Monte Carlo-type estimation of the time evolution of the pdf, that is, numerical integration of an ensemble of a large number of initial conditions that are generated by random sampling of the initial pdf. Application to weather forecasting of this Monte–Carlo approach is, however, also problematic, primarily due to the very high dimension ($\sim 10^7$) of the phase space of atmospheric circulation models that requires a very large number of Monte Carlo integrations to be performed (beyond the capacity of present-day supercomputers) in order to sufficiently reduce sampling uncertainty.

The practical approach that is applied to today's weather forecasting is to generate a relatively small ensemble (comprising typically 50–100 members) of initial perturbations to the observed initial state of the atmosphere. The perturbations are calculated in a way that ensures that the corresponding trajectories of the dynamical model equations initially diverge rapidly from the observed initial state. Thus, the time evolution of the individual ensemble members is not representative of the time evolution of the "true" pdf, but the ensemble is useful for the weather forecaster who is often most interested in the most likely development of the weather and possible major deviations from this development.

Forecasting becomes increasingly problematic when we also take into account that the forecast model is not perfect; that is, there are errors in the model (perhaps due to incorrect assumptions or poor spatial resolution), the effect of which is that the model does not behave like the real world. In the area of weather forecasting, model errors cause severe problems for long-range forecasts. A long-range weather forecast, or seasonal forecast, valid out to typically six months ahead, aims at predicting slowly varying surface conditions, such as sea surface temperature, sea ice cover, and soil moisture, and the effect they have on the preferred state of the atmospheric circulation in certain regions. A prominent example is prediction of El Niño conditions (increased sea surface temperature in the central and eastern tropical Pacific), which have an influence on worldwide weather.

The effect of model errors is in practice taken into account by basing seasonal weather forecasts on ensembles made up of integrations of more than one model. In this way, a pseudo-stochastic element is introduced into the forecast through the different formulations of the models, particularly for processes that take place on spatial scales that are smaller than the resolution of the model grid.

Forecasting is also of interest in areas in which models are not readily expressed in terms of known physical laws, most notably, in economics. When a physical model is not available, forecasting has to be based on more empirical methods. For example, a statistical forecast model would be based on the forecaster's notion of cause and effect, while an empirically constructed dynamical model could be derived from analysis of the combined behavior of many observed variables.

Due to the nondeterministic nature of economic evolution, economic forecasts will often be associated with substantial uncertainty. However, it is worth noting that a forecast need not be very accurate to be useful. If, in the long run, an economic forecast is just slightly better than a guess, then money can be made from the forecast.

In practical forecasting, no matter in what field, there is almost always an element of human expert judgment involved, and overall the expert judgment probably remains the most widespread forecasting method.

HENRIK FEDDERSEN

See also **Butterfly effect; Chaotic dynamics; Determinism; Economic system dynamics; Fokker–Planck equation; General circulation models of the atmosphere**

Further Reading

Ehrendorfer, M. 1994. The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: theory. *Monthly Weather Review*, 122: 703–713

Houghton, J. 1991. The Bakerian lecture, 1991. The predictability of weather and climate. *Philosophical Transactions of the Royal Society of London* A, 337: 521–572

Lorenz, E. 1963. Deterministic non-periodic flow. *Journal of Atmospheric Sciences*, 20: 130–141

Sivillo, J.K., Ahlquist, J.E. & Toth, Z. 1997. An ensemble forecasting primer. *Weather and Forecasting*, 12: 809–818

Smith, L.A. 2000. Disentangling uncertainty and error: On the predictability of nonlinear systems. In *Nonlinear Dynamics and Statistics* (Chapter 2), edited by A.I. Mees, Boston: Birkhauser

FOREST FIRES

The modeling of forest fires (FF) is based upon the concept of interacting particle systems (Liggett, 1985) for the description of complex systems. Within this framework, attention is on the interactions among particles (i.e., the trees), disregarding any attempt to understand the behavior of each particular tree. Using local interaction rules among neighboring trees, interesting collective behavior can be observed on a global scale.

FF models are defined so that trees may grow on the nodes of a surface network with periodic boundary conditions (Bak et al., 1990; Drossel & Schwabl, 1992, 1993; Christensen et al., 1993; Grassberger, 1993; Albano, 1995). Each node can assume three different states: (s1) occupied by a growing tree, (s2) occupied by a burning tree, and (s3) empty. The last corresponds to the absence of a tree that occurs after the completion of the tree burning process. Local interaction rules govern the dynamics of the system, which is updated as a cellular automation. For a given configuration, the time evolution is obtained by applying the following rules: (r1) a tree grows on an empty node with probability P; (r2) a burning tree evolves into an empty node; and (r3) a tree catches fire from a neighbor burning tree with probability $1 - G$. Thus, P is the growing probability of trees and G is the immunity of each tree to catch fire (Drossel & Schwabl, 1993).

Qualitatively, one expects that if the immunity is nonzero, the fire fronts present for $G = 0$ will become more and more fuzzy when G is increased. Of course, the fire cannot propagate for $G = 1$. For large values of P ($P \to 1$) and if the immunity is low ($G \to 0$), one observes coexistence of fire, trees, and empty nodes. Within this active state, fire fronts exhibit a rich variety of spatiotemporal patterns, including the development of spirals, and the collision between fronts. However, by keeping P constant and increasing G, the fire will eventually cease and the system will become trapped in an inactive state where the network is completely occupied by trees (see Figure 1). Since the spontaneous ignition of trees is not considered, the system will remain in the inactive state forever. So, for some critical values of the parameters $P \equiv P_c$ and $G \equiv G_c$, this simple kind of FF model exhibits irreversible phase transitions (Marro & Dickman, 1999) between active states with fire propagation and an inactive state where the fire becomes irreversibly extinguished. Figure 1 shows the critical curve at the border between these two states. In the case of standard reversible phase transitions and in marked contrast to forest fire models, the state of the system can be reversibly selected by tuning the control parameter (Binder & Heermann, 2002), that is, the temperature or pressure.

In spite of the irreversible nature of the transitions, a statistical treatment of this kind of FF model reveals that they share the relevant features of many phase transitions and critical phenomena occurring under equilibrium conditions. Therefore, FF models can be described by using a well-established framework in the field of statistical physics. In fact, close to the transition point, it is well known that the correlation length among particles is quite large—diverging to infinite at the critical point (Binder & Heermann, 2002). Therefore, at the critical state $P \to P_c$, $G \to G_c$, the burning of a single tree would affect the whole forest. The onset of this collective behavior, characteristic of critical phenomena, washes out the microscopic details (i.e., the mechanisms keeping every tree alive)

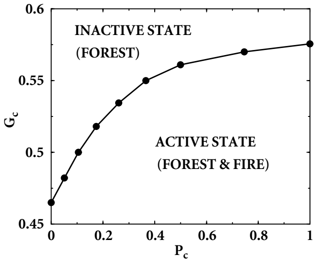

Figure 1. Plot of the critical threshold G_C versus P_C. Filled circles are critical values determined quite accurately by means of epidemic simulations. The critical line (solid curve) is drawn to guide the eye. The upper part of the figure corresponds to the inactive state of the system where the fire becomes irreversibly extinguished, while the lower part corresponds to the stationary state with fire fronts.

allowing the relevant behavior of the whole system to be caught by means of a small set of simple local rules, in agreement with the concept of interacting particle systems (Liggett, 1985).

A powerful method used to characterize and study an irreversible transition consists of performing the so-called epidemic (or spreading) simulations (ES) (Grassberger, 1989). The idea behind ES is to start the simulations from a configuration very close to the inactive state and follow the temporal evolution of the system under consideration. For this purpose, a network fully occupied by trees, except for a small patch of burning trees, is employed. Depending on the values of the parameters P and G, such a small fire would either propagate or become extinguished. During the propagation the following quantities are measured:

(i) The average number of burning trees $N(t)$;
(ii) The survival probability of the fire $P(t)$, that is, the probability that the fire is still ignited at time t;
(iii) The average mean-square distance $R^2(t)$ over which the fire has spread.

Due to the existence of a diverging correlation length, it is usually assumed that the measured quantities obey simple power laws given by

$$N(t) \propto t^{\eta}, \quad P(t) \propto t^{-\delta} \quad \text{and} \quad R^2 \propto t^{z}, \qquad (1)$$

where η, δ, and z are critical exponents. These exponents have been evaluated by means of extensive numerical simulations, yielding $\eta = 0.212 \pm 0.005$, $\delta = 0.459 \pm 0.005$, and $z = 1.115 \pm 0.010$ (Albano, 1995). This result allows us to place the FF model with immune trees within the universality class of directed percolation (DP) (Grassberger, 1989). It should be stressed that DP was previously proposed in the field of physics as a model for the systematic dripping of a fluid through a lattice with randomly occupied bonds (Kinzel, 1983), and similar irreversible behavior is observed in many disciplines, such as catalysis (the interruption of chemical reactions due to the poisoning of the catalysts (Loscar & Albano, 2003)), biology (the epidemic propagation of a disease (Marro & Dickman, 1999)), ecology (the extinction of species when resources become exhausted or the environment is changed (Rozenfeld & Albano, 2001)), and sociology (models for a society of interacting individuals (Monetti & Albano, 1997)).

By introducing an additional parameter given by the lightning probability F, which can be thought of as the probability of spontaneous ignition of a tree, to the above-formulated FF model, an interesting behavior with two characteristic time scales is observed. In the limit $F \to 0$, the onset of fire is quite sporadic and occurs on large time scales. After ignition, however, fire propagates during relatively short periods of time (Drossel & Schwabl, 1992; Christensen et al., 1993; Grassberger, 1993). By analogy to sandpile models, these rapid events are called avalanches. A statistical analysis of the area of the forest reached by fire shows that it lacks a characteristic size. So, in contrast to the standard Gaussian distribution, which describes many natural phenomena exhibiting a typical (average) size, avalanches of all magnitude are expected to be observed. Under these circumstances, it is believed that the time-scale separation effectively tunes the system into a scale-free state. This behavior is the typical sign of criticality observed in studies of the phase transitions of many substances and is due to the existence of a diverging correlation length.

Finally, several attempts to reproduce forest fires in the laboratory should be mentioned. Nahmias et al. (1989, 1996) have built a square network sample containing combustible and noncombustible blocks randomly distributed with a variable concentration. The effect of randomness on the propagation of the fire is studied by considering both the presence and absence of wind. It is found that the results are consistent with critical models of percolation. Also, Zhang et al. (1992) have studied forest-fire propagation by means of paper-burning experiments. It is found that fire fronts exhibit self-affine scaling behavior that can be described by a nonlinear stochastic differential equation of the Kardar–Parisi–Zhang type with Gaussian noise.

EZEQUIEL ALBANO

See also **Cellular automata; Critical phenomena; Flame front; Reaction-diffusion systems; Sandpile model**

Further Reading

Albano, E.V. 1995. Spreading and finite-size scaling study of the critical behaviour of a forest-fire model with immune trees. *Physica* A, 216: 213–226

Bak, P., Chen, K. & Tang, C. 1990. A forest-fire model and some thoughts on turbulence. *Physics Letters*, 147A: 297–300

Binder, K. & Heermann, D. 2002. *Monte Carlo Simulations in Statistical Physics: An Introduction*, 4th edition, Berlin and New York: Springer

Christensen, K., Flyvbjerg, H. & Olami, Z. 1993. Self-organized critical forest-fire model: mean-field theory and simulation results in 1 to 6 dimensions. *Physical Review Letters*, 71: 2737–2740

Drossel, B. & Schwabl, F. 1992. Self-organized critical forest-fire model. *Physical Review Letters*, 69: 1629–1632

Drossel, B. & Schwabl, F. 1993. Forest-fire model with immune trees. *Physica* A, 199: 183–97

Grassberger, P. 1989. Directed percolation in $2 + 1$ dimensions. *Journal of Physics A* (*Mathematical & General*), 22: 3673–3680

Grassberger, P. 1993. On a self-organized critical forest-fire model. *Journal of Physics A* (*Mathematical & General*), 26: 2081–2089

Kinzel, W. 1983. Directed percolation. In *Percolation Structures and Processes.* Annals of the Israel Physical Society. vol. 5, edited by G. Deutscher, R. Zallen & J. Adler Jerusalem: Israel Physical Society, pp.425–445

Liggett, T.M. 1985. *Interacting Particle Systems*, New York: Springer

Loscar, E.S. & Albano, E.V. 2003. Critical behaviour of irreversible reaction systems. *Reports on Progress in Physics*, 66: 1343–1382

Marro, J. & Dickman, R. 1999. *Nonequilibrium Phase Transitions and Critical Phenomena*, Cambridge and New York: Cambridge University Press

Monetti, R.A. & Albano, E.V. 1997. On the emergence of large-scale complex behaviour in the dynamics of a society of living individuals: the stochastic game of life. *Journal of Theoretical Biology*, 187: 183–194

Nahmias, J., Téphany, H. & Duarte, J. 1996. Percolation de sites en combustion. *Comptes Rendues de l´Academie des Sciences Paris*, 332: 113–119

Nahmias, J., Téphany, H. & Guyon, E. 1989. Propagation de la combustion sur un réseau hétérogène bidimensionnel. *Revue Physique Applique*, 24: 773–777

Rozenfeld, A.F. & Albano, E.V. 2001. Critical and oscillatory behaviour of a system of smart preys and predators. *Physical Review* E, 63: 061907-1–061907-6

Zhang, J., Zhang, Y.-C., Alstrom, P. & Levinsen, M.T. 1992. Modeling forest fire by paper-burning experiment, a realization of the interface growth mechanism. *Physica* A, 189: 383–389

FOURIER TRANSFORM

See **Integral transforms**

FOUR-WAVE MIXING

See **N-wave interactions**

FRACTAL BASIC BOUNDARIES

See **Attractors**

FRACTAL DIMENSION

See **Dimensions**

FRACTALS

Many spatial patterns and objects in nature are either irregular or fragmented to such an extreme degree that it is very hard to describe their form. For example, the coastline of a typical oceanic island or of a continent is neither straight, nor circular, nor elliptic, and no other classical curve can serve to describe it. Similarly, there is no Euclidean surface to capture adequately the boundaries of clouds or of rough turbulent wakes. To fill this gap, Benoit Mandelbrot proposed to use a family of shapes called *fractals* for geometric representation of the shapes mentioned above and many other irregular patterns. The word *fractal* derives from the Latin *fractus* (an adjective from the verb *frangere*, meaning "to break"), capturing the irregular characteristics of geometrical objects it describes. One can think of a fractal as an irregular set consisting of parts similar to the whole. It means that a fractal is a rough or fragmented geometrical shape that can be subdivided into parts, each of which is (at least approximately) similar at reduced size to the whole. Roughness present at any resolution of a fractal object distinguishes it from Euclidean shapes. The transition from a Euclidean geometry to a fractal one reflects the conceptual jump from translational invariance of traditional Euclidean shapes to continuous scale invariance of fractal objects. The concept of fractals is useful for describing various natural objects, such as clouds, coasts, and river or road networks. However, the fractal structure of natural objects is only observed over a limited range of scales, beyond which the fractal description breaks down.

During the study of geometry of irregular patterns, it became clear that a proper understanding of irregularity or fragmentation must, among other things, involve an analysis of the intuitive notion of dimension. Topological dimension cannot properly describe strongly irregular fractal structures. Let us consider the simplest example of a fractal, namely the Cantor set, see Figure 1. For the construction of the Cantor set an iteration procedure is used. At the zeroth level, the construction of the triadic Cantor set begins with the unit interval, that is, all points on the line between 0 and 1. The first level is obtained from the zeroth one by deleting all points between $\frac{1}{3}$ and $\frac{2}{3}$, that is the middle third of the initial interval. The second level is obtained from the first level by deleting the middle third of each remaining interval at the first level. In general, the next level is obtained from the previous one by deleting the middle third of all intervals obtained from the previous level. This process continues ad infinitum. The result is a collection of points with topological dimension $d_{\mathrm{t}} = 0$ since its total measure (length) is zero.

It is seen that the notion of dimension from Euclidean geometry is not very useful for the Cantor set and similar objects since it does not distinguish between the rather complex set of elements and a single point that also has a vanishing topological dimension. To cope

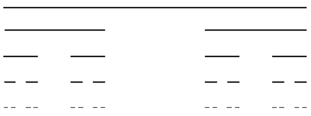

Figure 1. The construction of the Cantor set.

with this degeneracy, concepts of fractal dimensions were introduced for quantifying such fractal sets. Many paradoxical aspects of the Cantor set can be traced to the fact that it is dimensionally discordant. Furthermore, the discordant character of basic fractals is not at all a minor nuisance. Rather, it is such a basic feature that we shall use it to give a tentative definition of the concept of fractal. The simplest nontrivial dimension that generalizes the topological dimension is the so-called fractal dimension, which in this context can be defined as follows:

$$D_{\mathrm{c}} = \lim_{n\to+\infty} \frac{\ln N_n}{\ln 1/l_n}. \tag{1}$$

Here, N_n is the number of observable elements at the nth level, l_n is the measure (length) of the smallest element at the nth level, and $1/l_n$ is the resolution. For the triangle Cantor set at the nth level, the set consists of $N_n = 2^n$ segments, each of which has length $l_n = \frac{1}{3}^n$. Therefore, the capacity dimension of the triangle Cantor set equals

$$D_{\mathrm{c}}^{\mathrm{cantor}} = \frac{\ln 2}{\ln 3} \simeq 0.63. \tag{2}$$

We see that fractal dimension is noninteger. Mathematically, we can say that a fractal is a set of points whose fractal dimension exceeds its topological dimension.

Felix Hausdorff developed another way to quantify a fractal set. He suggested examining the number $N(\varepsilon)$ of small intervals needed to cover the set at a scale ε. The Hausdorff dimension is defined by considering the quantity

$$M \equiv \lim_{\varepsilon\to 0} \inf_{\varepsilon_m<\varepsilon} \sum_m \varepsilon_m^d, \tag{3}$$

which is constructed by summing d-dimensional volumes ε_m^d of balls of radii ε_m not exceeding ε that cover the fractal set. The "inf" means that all partitions of balls of radius less than ε that cover the set are considered and the one that gives the smallest possible value for the sum is kept. Hausdorff demonstrated that there is a special value D_{H} for the exponent d, called the *Hausdorff dimension*, such that for $d < D_{\mathrm{H}}$ measure $M = \infty$ as $\varepsilon \to 0$ and for $d > D_{\mathrm{H}}$ measure $M = 0$ in the limit $\varepsilon \to 0$.

Now the reader may ask, why do we need more than one fractal dimension? Actually, as is seen below, there is an infinite set of fractal dimensions. The reason for this number of quantities describing a single fractal object is the *fractal form.* The form of Euclidean shapes is well described by topology. However, as was noted by Mandelbrot the concept of form possesses mathematical aspects other than topological ones. Topology is the branch of mathematics that teaches us, for example, that all pots with two handles are of the same form because, if they were made of an infinitely flexible clay, each could be molded into any new opening, closing up any old one. Obviously, this particular aspect of the notion of form is not useful in the study of individual coastlines, since it simply indicates that they are all topologically identical to each other (and to a circle). This identity is underlined by the fact that the topological dimension in each case equals 1. By way of contrast, it will be seen that coastlines of different "degrees of irregularity" tend to have different fractal dimensions. Differences in fractal dimensions express differences in nontopological aspects of form, which can be called the *fractal form.*

Now let us consider the cumulative distribution function of mass along the Cantor set. Let the initial line's length and mass both be equal to 1, and define the cumulative distribution function for the abscissa x as being the mass contained between 0 and x. It is assumed that the total mass of the pattern is not changed while building the Cantor set. It means that we do not take away one-third of each line, but divide it in to two parts and compress them until the length of each part equals one-third of the initial line length at the previous step. Iterating this procedure up to infinity, we obtain the massive Cantor set. Because there is no mass in the intermissions, the distribution function remains constant along almost the whole length of the unit interval. However, since hammering does not affect the total mass, the distribution must manage to increase somewhere from the point of coordinates $(0, 0)$ to the point of coordinates $(1, 1)$. It increases over infinitely many, infinitely small, highly clustered jumps corresponding to the points of the Cantor set. The plot of the cumulative distribution function shown in Figure 2 is called the *Devil's staircase*. It is the plot of a continuous function on the unit interval, whose derivative is 0 almost everywhere, but it rises from 0 to 1 (see Figure 2). The cumulative sums of the widths and of the heights of the steps are both equal to 1, and one finds in addition that this curve has a well-defined length equal to 2. A curve of finite length is called *rectifiable* and is of dimension $D = 1$. This example has the virtue of demonstrating that sharp irregularities do not necessarily prevent a curve from being of dimension $D = 1$, as long as they remain sufficiently few and scattered.

You can see from Figure 2 that the Devil's staircase is a function which is not differentiable at the Cantor set points, but has zero derivative everywhere else. Because a Cantor set has measure zero, this function has

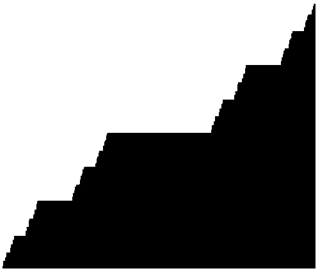

Figure 2. The Devil's staircase.

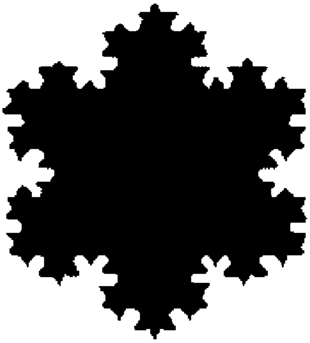

Figure 3. The Koch snowflake.

zero derivative practically everywhere and rises only on Cantor set points.

Now we consider the fractal whose dimension, in contrast to the Cantor set, exceeds unity. As an example we consider the *Koch snowflake*; see Figure 3. This is a fractal, also known as the Koch island, which was first described by Helge von Koch in 1904. It is built by starting with an equilateral triangle, removing the inner third of each side, building another equilateral triangle at the location where the side was removed, and then repeating the process indefinitely (see Figure 4). The border of the snowflake is a curve. Indeed, its area vanishes, but on the other hand each stage of its construction increases its total length by the ratio 4/3. Thus, the limit border is of infinite length. It is also continuous, but it has no definite tangent in almost all of its points because it has, so to speak, a corner almost everywhere. Its nature borders on that of a continuous function without a derivative. This object has a fractal nature. The capacity dimension of the Koch snowflake equals $D_{\mathrm{c}}^{\mathrm{Koch}} = \ln 4/\ln 3 \simeq 1.26$. Though the total length of the border is infinite, the area of the Koch island originated from the initial triangle with unit side is finite and equals $S^{\mathrm{Koch}} = 2\sqrt{5}/5$.

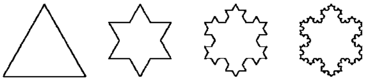

Figure 4. The construction of the Koch snowflake.

Fractals arising in nature usually have a more complex scaling relation than simple fractals (like the Cantor set) described above. The single fractal dimension is not enough to describe their structure, because these fractal sets usually involve a range of scales that depend on their location within the set (the simplest example is the Cantor set with some measure distributed on it). Such fractals are called *multifractals*.

The multifractal formalism is a statistical description that provides global information on the self-similarity properties of fractal objects. For practical realization of the method, one primarily covers the fractal set under study by a regular array of cubic boxes of some given mesh size l. The measure of weight p_n of a given box n is then defined as the sum of the measure of interest within the box. A simple fractal dimension α is defined by the relation:

$$p_n \sim l^{\alpha}. \tag{4}$$

Multifractal analysis is a generalization in which α may change from point to point and is a local quantity. The standard method to test for multifractal properties consists in calculating the moments of order q of the measure p_n defined by

$$M_q(l) = \sum_{n=1}^{n(l)} p_n^q. \tag{5}$$

Here $n(l)$ is the total number of non-empty boxes. Varying q, one can characterize the inhomogeneity of the pattern, for instance, the values of moments with large q are controlled by the densest boxes. If scaling holds, then the generalized dimension D_q is defined by the relation:

$$M_q(l) \sim l^{(q-1)D_q}. \tag{6}$$

For instance, D_0, D_1, and D_2 are called *capacity* (defined above), *information*, and *correlation* dimensions, respectively.

In multifractal analysis, one also determines the number $N_\alpha(l)$ of boxes having similar local scaling characterized by the same exponent α. Using it, one can introduce the multifractal singularity spectrum $f(\alpha)$ as the fractal dimension of the set of singularities of strength α according to the following relation:

$$N_\alpha(l) \sim (1/l)^{f(\alpha)}. \tag{7}$$

There is a general relationship between a moment of order q and a singularity strength α, expressed

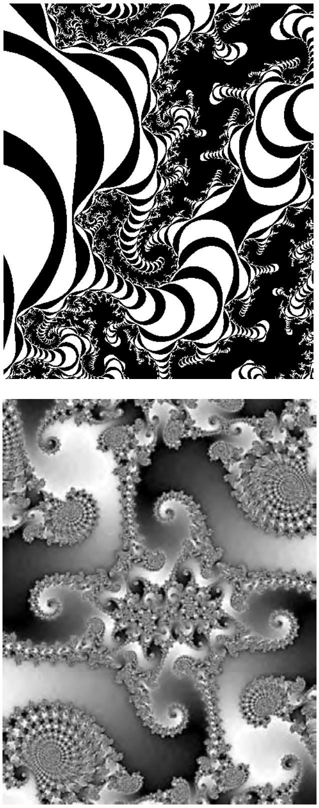

Figure 5. Examples of Julia sets.

mathematically as a Legendre transformation:

$$f(\alpha) = q\alpha - (q-1)D_q. \tag{8}$$

Fractal sets can be impressive. For example, see Figure 5 where a number of so-called *Julia sets* is shown. (See also color plate section.)

Mathematically, the Julia set can be introduced as follows. Take some function $f(z)$ and consider the sequence obtained when $f(z)$ is iterated starting from the point $z = a$:

$$a, f(a), f(f(a)), f(f(f(a))), \text{etc.} \tag{9}$$

Depending on the initial condition a and form of the function f, it may happen that these values stay small or they do not, that is, repeatedly applying f to yield arbitrary large values. So the set of all numbers (initial conditions) is partitioned; into two parts, and the Julia set associated with the function f is the boundary between these sets. The "filled" Julia set includes those numbers $z = a$ for which the iterates of f applied to remain bounded. If one considers complex numbers rather than real ones, it is the complex plane that is partitioned into two sets, and the resulting picture can be quite striking; see Figure 5. That is an example of iterating a quadratic function defined in the complex plane. Linear functions do not yield interesting partitions of the complex plane, but quadratic and higher-order polynomials do.

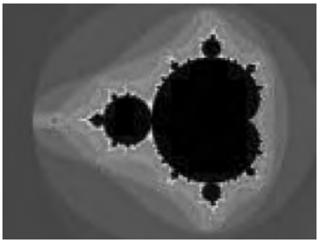

Figure 6. The Mandelbrot set.

Consider the most studied family of quadratic polynomials $f(z) = z^2 + \mu$ parametrized by a complex variable μ. As μ varies, the Julia set will vary on the complex plane. Some of these Julia sets will be connected, and some will be disconnected. Those values of μ for which the Julia set is connected are called the *Mandelbrot set* in the parameter plane. The boundary between the Mandelbrot set and its complement is often called *Mandelbrot separator curve*. The Mandelbrot set is the black shape in Figure 6 (see also color plate section). The Mandelbrot set is the set of points in the complex μ-plane that do not go to infinity when iterating the map $z_{n+1} = z_n^2 + \mu$ starting with $z = 0$. One can avoid the use of the complex numbers by substitution $z = x + \mathrm{i}y$ and $\mu = a + \mathrm{i}b$, and computing the orbits in the ab-plane for two-dimensional mapping:

$$\begin{aligned} x_{n+1} &= x_n^2 - y_n^2 + a, \\ y_{n+1} &= 2x_n y_n + b, \end{aligned} \tag{10}$$

with initial conditions $x = y = 0$ or equivalently $x = a$, $y = b$.

Now let us consider a natural phenomenon exhibiting fractal properties, namely, Brownian motion. Consider a fluid mass in equilibrium, for example, water in a glass; all the parts appear completely motionless. However, if one puts a small enough particle into the water, it is observed that instead of rising or descending regularly (depending on the density of the particle),

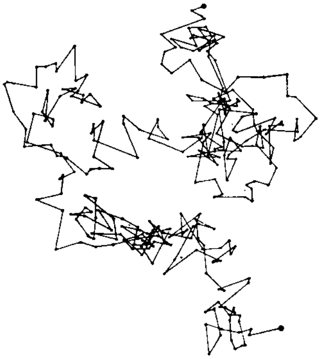

Figure 7. The example of Brownian motion of a colloidal particle.

it moves with a perfectly irregular movement. The segment of motion of a colloidal particle is presented in Figure 7 as it is seen under the microscope. The successive positions of a particle in equal time intervals are marked by points, then joined by straight lines having no physical reality whatsoever. If this particle's position were marked down 100 times more frequently, each segment would be replaced by a polygon relatively just as complicated as the whole drawing, and so on. Here, we have the natural example of the *scaling property* of the colloidal particle's motion. It is easy to see that in practice the notion of tangent is meaningless for such curves. This property of the trajectory reminds one of the Koch's snowflake described above, because when a Brownian trajectory is examined increasingly closely, its length increases without bound. Thus, we see that the trajectory of a colloidal particle has features peculiar to fractal objects, and we can qualify Brownian motion as being*fractal* and giving us a natural example of a fractal pattern.

Vitcheslav Kuvshinov and Andrei Kuzmin

See also **Brownian motion; Dimensions; Multifractal analysis**

Further Reading

Barnsley, M. 1988. *Fractals Everywhere*, Boston: Academic Press

Devaney, R. & Keen, L. (editors). 1989. *Chaos and Fractals: The Mathematics Behind the Computer Graphics*, Providence, RI: American Mathematical Society

Falcone, K.J. 1985. *The Geometry of Fractal Sets*, Cambridge and New York: Cambridge University Press

Feder, J. 1988. *Fractals*, New York and London: Plenum Press

Mandelbrot, B.B. 1988. *Fractal Geometry of Nature*, New York: Freeman

Sornette, D. 2000. *Critical Phenomena in Natural Sciences. Chaos, Fractals, Self-organization, and Disorder: Concepts and Tools*, Berlin and New York: Springer

FRAMED SPACE CURVES

The Frenet–Serret Frame

Let $\gamma : I \to \mathbb{R}^3$ be a C^3 regular parametrized space curve, with I an open interval on the real line; by regular we mean that $\gamma'(t)$ is never zero. We define the unit tangent vector $\boldsymbol{t}$ by $\boldsymbol{t} = \gamma'/|\gamma'|$ and an arclength coordinate s by $s = \int |\gamma'|\,\mathrm{d}t$. An inflection point of the curve is one where $\boldsymbol{t}'$ is zero; away from such points, we may define the unit normal by $\boldsymbol{n} = \boldsymbol{t}'/|\boldsymbol{t}'|$. Then the binormal vector is $\boldsymbol{b} = \boldsymbol{t} \times \boldsymbol{n}$, where $\times$ denotes the cross product of vectors in $\mathbb{R}^3$.

Expressing the derivatives of these mutually orthogonal unit vectors in terms of themselves produces the Frenet equations

$$\begin{aligned}\dot{\boldsymbol{t}} &= \kappa \boldsymbol{n},\\ \dot{\boldsymbol{n}} &= -\kappa \boldsymbol{t} + \tau \boldsymbol{b},\\ \dot{\boldsymbol{b}} &= -\tau \boldsymbol{n},\end{aligned}$$

where the dot denotes derivative with respect to arclength s, and κ and τ are the *curvature* and *torsion* functions of the curve, respectively. Curvature may be defined in terms of $\boldsymbol{t}$ by $\kappa = |\dot{\boldsymbol{t}}|$, so that if κ is identically zero then $\gamma(t)$ is a straight line; more generally, inflection points, where κ vanishes, may be defined as points where the curve has at least second-order contact with a line.

We will say that a regular curve is *nondegenerate* if it has no inflection points. For such curves, we may define τ in terms of the scalar triple product as $\tau = (\boldsymbol{t}, \dot{\boldsymbol{t}}, \ddot{\boldsymbol{t}})/\kappa^2$. Thus, τ depends on the third derivative of γ while κ is a second-order quantity. Moreover, τ is identically zero if and only if $\gamma(t)$ lies in a fixed plane parallel to $\boldsymbol{t}$ and $\dot{\boldsymbol{t}}$.

Curvature and torsion are unchanged if we reverse the orientation of the curve (i.e., the direction of $\boldsymbol{t}$), or if the curve is rotated or translated within $\mathbb{R}^3$. (However, τ is changed by a minus sign if the curve is reflected, and both functions are multiplied by ρ^{-1} when the curve is scaled by a factor $\rho > 0$.) Conversely, uniqueness theorems for linear systems of differential equations, as applied to the Frenet equations, imply that two nondegenerate space curves with the same curvature and torsion, as functions of arclength, are congruent under an orientation-preserving isometry of $\mathbb{R}^3$ (congruence theorem).

Natural Frames

Although the vector $\kappa\boldsymbol{n}$ is well defined for any regular curve, by itself $\boldsymbol{n}$ is undefined at inflection points, and even at simple inflection points (where $\boldsymbol{t} \times \dot{\boldsymbol{t}}$ vanishes only to first order), κ is nondifferentiable and $\boldsymbol{n}$ is discontinuous. One can get around this problem and obtain a C^1 orthonormal framing along the curve, by using *natural frames*, also known as relatively parallel adapted frames (Bishop, 1975). Such frames satisfy equations of the form

$$\begin{aligned}\dot{\boldsymbol{t}} &= k_1\boldsymbol{m}_1 + k_2\boldsymbol{m}_2,\\ \dot{\boldsymbol{m}}_1 &= -k_1\boldsymbol{t},\\ \dot{\boldsymbol{m}}_2 &= -k_2\boldsymbol{t}, \qquad (1)\end{aligned}$$

where $\boldsymbol{t}$ is the usual unit tangent vector, $\boldsymbol{m}_1$ is a unit normal to $\boldsymbol{t}$, $\boldsymbol{m}_2 = \boldsymbol{t} \times \boldsymbol{m}_1$, and k_1, k_2 are the *natural curvatures*. The natural frame is not unique, but any two such frames for the same oriented curve differ only by rotating $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ around $\boldsymbol{t}$ through an angle which is constant along the curve.

Any regular curve that is C^k for $k \geq 2$ has a C^{k-1} natural frame. When the curve also has a Frenet frame, and we suppose $\boldsymbol{n} = \cos\theta\,\boldsymbol{m}_1 + \sin\theta\,\boldsymbol{m}_2$, then $k_1 = \kappa\cos\theta, k_2 = \kappa\sin\theta$, and $\theta = \int \tau\,\mathrm{d}s$. Thus, a space curve is planar when its natural curvatures (k_1, k_2), when graphed in $\mathbb{R}^2$, lie along a fixed line through the origin. More generally, a space curve lies on a fixed sphere of radius ρ when (k_1, k_2) lie along a line distance $1/\rho$ from the origin.

The above congruence theorem is true for regular curves with degeneracies when suitably rephrased in terms of natural frames and natural curvatures.

The Hasimoto Correspondence

The vortex filament flow

$$\frac{\partial\gamma}{\partial t} = \frac{\partial\gamma}{\partial x} \times \frac{\partial^2\gamma}{\partial x^2} \qquad \text{(VFF)}$$

is an evolution equation for space curves that models the self-induced motion of a vortex filament in an ideal fluid taking only local effects into account. This model, originally formulated by L. Da Rios in 1906 (Ricca, 1996), is also known as the *localized induction approximation*. Note that if x is an arclength parameter at a given time, then it remains so for all t. We will take x as such, and then the right-hand side of (VFF) is κB when expressed in terms of the Frenet frame.

Hasimoto (1972) showed that solutions to this equation induce solutions to the focusing cubic nonlinear Schrödinger equation

$$\mathrm{i}q_t + q_{xx} + \tfrac{1}{2}|q|^2 q = 0, \qquad \text{(NLS)}$$

by virtue of letting

$$q = \kappa\mathrm{e}^{\mathrm{i}\theta} \quad \text{with} \quad \theta = \int \tau\,\mathrm{d}s. \qquad \text{(H)}$$

However, the "constant of integration" cannot be taken to be an arbitrary function of t; in fact, substituting (H) into (NLS) shows that we must have $\theta_t = \kappa^{-1}\ddot{\kappa} - \tau^2 + \frac{1}{2}\kappa^2$. Thus, $q(x, t)$ is unique up to multiplication by a unit modulus constant. Unfortunately, definition (H) for the Hasimoto map does not work at points of inflection (unless $\ddot{\kappa}$ vanishes to the same order as κ). However, it may be defined for arbitrary regular curves by using the natural curvatures of an evolving natural frame (see Calini, 2000, Section 1.4), whereupon $q = k_1 + \mathrm{i}k_2$.

DNA and White's Formula

At its lowest level of structure, DNA consists of two right-handed helices connected by base pairs, resembling the rungs of a twisted ladder. At lesser magnification the double helix, seen as a single strand, usually coils around itself in a left-handed fashion called *negative supercoiling*. Moreover, as the base pairs are separated during DNA replication and the ladder is untwisted, the supercoiling relaxes (becomes less negative) or may even reverse itself to become positive supercoiling. The dynamic interplay of these two levels of structure in DNA is governed by White's formula (White, 1970; Pohl, 1980). Although the formula, as stated below, applies only to closed loops of DNA (which occur in bacteria and in the mitochondria of some cells), it also governs the behavior of long strands of DNA, since in living cells the DNA is clamped to a protein scaffold at intervals of roughly 100,000 base pairs (Cozzarelli, 1992).

Let $\mathcal{C} \subset \mathbb{R}^3$ be a closed oriented regular curve and let $\boldsymbol{v}$ be a C^1 unit normal vector field along $\mathcal{C}$. (In DNA, we should regard $\mathcal{C}$ as one of the helices and $\boldsymbol{v}$ as pointing along the base pairs.) Let $\tilde{\mathcal{C}} = \mathcal{C} + \varepsilon\boldsymbol{v}$ have the same orientation. For $\varepsilon > 0$ sufficiently small, the linking number $\mathrm{Lk}(\mathcal{C}, \tilde{\mathcal{C}})$ is independent of ε and will be denoted by Lk. Define the total twist of $\boldsymbol{v}$ by the integral

$$\mathrm{Tw} = \frac{1}{2\pi}\int (\boldsymbol{t}, \boldsymbol{v}, \dot{\boldsymbol{v}})\,\mathrm{d}s.$$

The twisting of $\boldsymbol{v}$ about γ obviously contributes to $\mathrm{Lk}(\mathcal{C}, \tilde{\mathcal{C}})$, but there is also a contribution independent of $\boldsymbol{v}$, due to the "writhing" of $\mathcal{C}$ about itself.

White's formula says that

$$\mathrm{Lk} = \mathrm{Tw} + \mathrm{Wr},$$

where Wr is the *writhe* of $\mathcal{C}$, defined as follows. On $D = \{(x, y) \in \mathcal{C} \times \mathcal{C} | x \neq y\}$, define the vector-valued function $\boldsymbol{e} = (x - y)/|x - y|$. Extend $\boldsymbol{e}$ continuously to the two components of ∂D, as the forward- and backward-pointing unit tangent vector. Then

$$\mathrm{Wr} = \frac{1}{4\pi}\int_{\overline{D}} \boldsymbol{e}^*\,\mathrm{d}S,$$

where $\mathrm{d}S$ is the standard element of area on the unit sphere S^2 and $\overline{D}$ is the closure of D. If s_1 and s_2 are arclength parameters on D, then we may also write

$$\mathrm{Wr} = \frac{1}{4\pi}\int\int \left(\boldsymbol{e}, \frac{\partial\boldsymbol{e}}{\partial s_1}, \frac{\partial\boldsymbol{e}}{\partial s_2}\right)\mathrm{d}s_1\,\mathrm{d}s_2.$$

When $\mathcal{C}$ is nondegenerate and $\boldsymbol{v} = \boldsymbol{n}$, the Frenet normal, Lk is called the *self-linking* or *helicity* of $\mathcal{C}$, and White's formula specializes to

$$\mathrm{SL}(\mathcal{C}) = \mathrm{Wr} + \int \tau \, \mathrm{d}s,$$

a formula which was discovered much earlier by Calagareanu and later extended to regular closed curves (Pohl, 1980).

In DNA, with each base pair twisting the ladder by about 34°, Tw is positive and on the order of 10^5. The fact that DNA supercoiling is left-handed is reflected by a negative value of Wr, leading to a relatively small overall linking number. However, this number is rarely zero; the surgery that unlinks each of the two strands, a necessary step in cell division, is accomplished by enzymes known as *topoisomerases* (Sumners, 1995).

THOMAS A. IVEY

See also **DNA premelting; DNA solitons**

Further Reading

Bishop, R. 1975. There is more than one way to frame a curve. *American Mathematical Monthly*, 82: 246–251

Calini, A. 2000. Recent developments in integrable curve dynamics. In *Geometric Approaches to Differential Equations*, edited by P. Vassiliou & I. Lisle, Cambridge and New York: Cambridge University Press

Cozzarelli, N. 1992. Evolution of DNA topology: implications for its biological roles. In *New Scientific Applications of Geometry and Topology*, edited by D. Sumners, Providence, RI: American Mathematical Society

Hasimoto, H. 1972. A soliton on a vortex filament. *Journal of Fluid Mechanics*, 51: 477–485

Pohl, W. 1980. DNA and differential geometry. *Mathematical Intelligencer*, 3: 20–27

Ricca, R. 1996. The contributions of Da Rios and Levi–Civita to asymptotic potential theory and vortex filament dynamics. *Fluid Dynamical Research*, 18: 245–268

Sumners, D. 1995. Lifting the curtain: using topology to probe the hidden action of enzymes. *Notices of the American Mathematical Society*, 42: 528–537

White, J. 1970. Some differential invariants of submanifolds of Euclidean space. *Journal of Differential Geometry*, 4: 207–223

FRANCK–CONDON FACTOR

The concept of Franck–Condon transitions between electronic energy levels originates from the Born–Oppenheimer approximation, which adiabatically separates the fast electronic degrees of freedom of a molecule from the much slower nuclear degrees of freedom (since the nuclei are much heavier than the electrons, electronic motions occur as if the nuclei were fixed in place). Within the Born–Oppenheimer approximation, the total wave function $\Psi(q, Q)$ can be written as a product of an electronic and a nuclear (vibrational) wave function:

$$\Psi(q, Q) = \Psi^{(\mathrm{el})}(q; Q)\varphi^{(\mathrm{vib})}(Q), \tag{1}$$

where the electronic wave function $\Psi^{(\mathrm{el})}(q; Q)$ depends only parametrically on the nuclear coordinates Q. One first calculates the electronic eigenstates for a particular nuclear configuration Q, which is assumed to be fixed in space, and assembles electronic potential energy surfaces by varying the nuclear coordinates Q step-by-step. The nuclear wave function is subsequently calculated by solving the Schrödinger equation on these potential energy surfaces, assuming that the electrons can adapt adiabatically to a changing nuclear configuration.

As usual, the intensity of a transition between two electronic states k and l with vibrational quanta n and m is given by the transition dipole matrix element

$$\mu_{kn,lm}^2 = \langle \Psi_{k,n} | \hat{\mu} | \Psi_{l,m} \rangle^2 = \left(\frac{\partial \mu}{\partial q} \right)^2 \langle \Psi_{k,n} | \hat{q} | \Psi_{l,m} \rangle^2. \tag{2}$$

When incorporating the product wave function of Equation (1), we obtain

$$\mu_{kn,lm}^2 = \left(\frac{\partial \mu}{\partial q} \right)^2 \langle \Psi_k^{(\mathrm{el})} | \hat{q} | \Psi_l^{(\mathrm{el})} \rangle^2 \langle \varphi_{k,n}^{(\mathrm{vib})} | \varphi_{l,m}^{(\mathrm{vib})} \rangle^2 . \tag{3}$$

The transition dipole matrix element separates into a product because (a) the transition dipole operator $\partial \mu / \partial q \cdot \hat{q}$ acts on the electronic coordinates q only and (b) we have assumed that the transition dipole operator does not depend on the nuclear coordinate Q (*Franck–Condon approximation*). The second term $\langle \varphi_{k,n}^{(\mathrm{vib})} | \varphi_{l,m}^{(\mathrm{vib})} \rangle^2$ is the *Franck–Condon factor*.

Figure 1a shows possible transitions between two electronic potential energy surfaces (S_0 and S_1). At not too high temperatures, only the vibrational ground state $\nu = 0$ in the electronic ground state S_0 will be populated. When both potential energy surfaces have the same shape and their lowest points are not displaced, the Franck–Condon factors $\langle \varphi_n^{(\mathrm{vib})} | \varphi_m^{(\mathrm{vib})} \rangle^2$ reduce to the condition of orthonormality of the vibrational eigenstates (i.e., the vibrational wave functions are the same on both potential energy surfaces). Hence, in that case, only the $\nu = 0 \rightarrow \nu' = 0$ transition (the *zero-phonon transition*) carries oscillator strength, while all others are dark. However, when both potential energy surfaces are displaced (i.e., the excited state has a different nuclear coordinate to the ground state), the absorption spectrum will consist of a series of absorption lines (a Franck–Condon progression) as depicted in Figure 1b. The intensity of these lines scales with the Frank–Condon factors $\langle \varphi_{k,n}^{(\mathrm{vib})} | \varphi_{l,m}^{(\mathrm{vib})} \rangle^2$. In the spectra of Figure 1b, the energy origin relates to the frequency of the zero-phonon transition, and

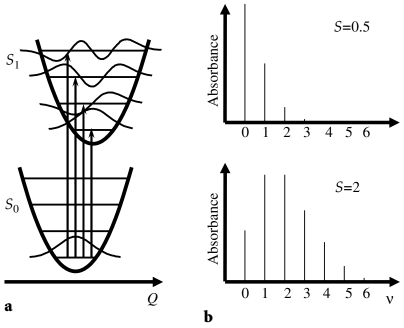

Figure 1. (a) Displaced potential energy surfaces of the electronic ground state S_0 and the first electronic excited state S_1 with the possible transitions depicted at a temperature of 0 K. The overlap integrals $\langle\varphi_i^{(\mathrm{vib})}|\varphi_j'^{(\mathrm{vib})}\rangle^2$ of the vibrational wave functions in both electronic states determine the strength of the individual absorption lines. (b) Absorption spectra for Huang–Rhys parameters $S = 0.5$ and 2, respectively.

the energy spacing between the various lines is given by the frequency ω of the oscillator. We call normal modes with such a displacement of the potential energy surfaces Franck–Condon active modes.

An analytic expression for the Franck–Condon factors can be given when the potential energy surfaces are assumed to be harmonic (as in Figure 1a):

$$\begin{aligned} V_{S_0} &= \frac{1}{2}\hbar\omega Q^2, \\ V_{S_1} &= E_{01} + \frac{1}{2}\hbar\omega Q^2 - \bar{x}\hbar\omega Q, \end{aligned} \tag{4}$$

where Q is the dimensionless oscillator coordinate, E_{01} the energy separation between the electronic ground and first excited states at $Q = 0$, and ω the vibrational frequency. The dimensionless displacement $\bar{x}$ between potential energy surfaces is a measure of the strength of the coupling between electronic and vibrational transition. It is common to introduce an alternative coupling parameter, the Huang–Rhys parameter:

$$S \equiv \frac{x^2}{2} \tag{5}$$

with which the Franck–Condon factors can be expressed as (Fitchen, 1968)

$$\langle\varphi_{S_1,n}^{(\mathrm{vib})}|\varphi_{S_0,m}^{(\mathrm{vib})}\rangle^2 = \mathrm{e}^{-S}\frac{m!}{n!}S^{n-m}\left[L_m^{n-m}(S)\right]^2, \tag{6}$$

where $L_j^{i-j}(S)$ are the Laguerre polynominals. For the important case where $T = 0$ K, the only occupied initial level is $m = 0$ and $L_0^n(S) = 1$, and we obtain

$$\langle\varphi_{S_1,n}^{(\mathrm{vib})}|\varphi_{S_0,0}^{(\mathrm{vib})}\rangle^2 = \frac{S^n}{n!}\mathrm{e}^{-S}. \tag{7}$$

For $S < 1$, the zero phonon line is the most intense transition, while for larger displacements, $S > 1$, the intensity is shifted toward transitions with higher vibrational quanta on the electronically excited state S_1 (Figure 1b). This finding can be explained in simple words with the Franck–Condon principle, which states that in a classical picture, an electronic transition occurs very quickly and is most likely to occur without changes in the positions of the nuclei. Hence, the transition is vertical from the bottom of the S_0 potential energy surface to the S_1 surface. The quantum mechanical formulation of this principle is that the intensity of a vibronic transition is proportional to the square of the overlap integral between the vibrational wave functions of the two states that are involved in the transition.

The concept of a Franck–Condon transition was originally formulated for electronic states. However, it can be used in a more general sense and appears whenever an adiabatic separation of timescales can be made between two molecular degrees of freedom. For example, this is the case when a high-frequency vibrational mode is coupled to a low-frequency vibrational mode. In that case, $\Psi^{(\mathrm{el})}$ in the expressions above is replaced by the wave function of the high-frequency mode, but the rest of the formalism remains unchanged. Most prominent examples are hydrogen bonds XH$\cdots$Y, where the high-frequency X$-$H stretching vibration is coupled to the low-frequency XH$\cdots$Y hydrogen bond vibration. One of the striking features of hydrogen bonds is an extraordinarily large anharmonic coupling between both modes that leads to a Franck–Condon-like progression in the absorption spectrum of the X$-$H stretching band.

Peter Hamm

See also **Color centers; Hydrogen bond**

Further Reading

Fitchen, D.B. 1968. Zero-phonon transitions. In *Physics of Color Centers*, edited by W.B. Fowler, New York: Academic Press, pp. 293–350

FREDHOLM THEOREM

The Fredholm theorem gives a necessary and sufficient condition for the existence of a solution to a linear equation of the form $\boldsymbol{Ax} = \boldsymbol{y}$, where A is a linear operator and $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors. Consider the matrix

$$A = \begin{pmatrix} 1 & 4 & 5 \\ 2 & 5 & 7 \\ 3 & 6 & 9 \end{pmatrix}, \tag{1}$$

and the linear map $\mathcal{F}$, which associates the vector $\boldsymbol{Ax}$ to each vector $\boldsymbol{x}$ in $\mathbb{R}^3$. If $x_i,\ i = 1\ldots3$, are the components of $\boldsymbol{x}$, we have

$$\mathcal{F}\boldsymbol{x} = A\boldsymbol{x} = x_1 \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + x_2 \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + x_3 \begin{pmatrix} 5 \\ 7 \\ 9 \end{pmatrix}$$
$$= (x_1 + x_3) \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + (x_2 + x_3) \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}. \tag{2}$$

The third equality comes from the fact that the last column of A can be written as the sum of its first two columns. It is then clear that the equation $\mathcal{F}\boldsymbol{x} = \boldsymbol{y}$ has a solution if and only if $\boldsymbol{y}$ is a linear combination of the first two columns of A. In this case, there is more than one solution because the kernel or null space of A is not trivial.

If we equip $\mathbb{R}^3$ with the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{3} x_i\, y_i$, the transpose A^{T} of A, defined by $\{A^{\mathrm{T}}\}_{ij} = \{A\}_{ji}$, is such that

$$\langle A\boldsymbol{x}, \boldsymbol{v} \rangle = \langle \boldsymbol{x}, A^{\mathrm{T}}\boldsymbol{v} \rangle, \tag{3}$$

for any two vectors $\boldsymbol{x}$ and $\boldsymbol{v}$ in $\mathbb{R}^3$. This equality shows that $\boldsymbol{v}$ is orthogonal to the range of A (i.e., $\langle A\boldsymbol{x}, \boldsymbol{v} \rangle = \boldsymbol{0}$ for every $\boldsymbol{x}$ in $\mathbb{R}^3$) if and only if $\boldsymbol{v}$ is in the kernel of A^{T}; that is, $A^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}$. As a consequence, the existence of a solution to the equation $\mathcal{F}\boldsymbol{x} = \boldsymbol{y}$ (i.e., $A\boldsymbol{x} = \boldsymbol{y}$), which is equivalent to $\boldsymbol{y}$ being in the range of A, can be restated as $\boldsymbol{y}$ being orthogonal to the kernel of A^{T}. This is the contents of the Fredholm theorem, which is often presented as the following alternative theorem.

Fredholm alternative: For a matrix A, either the equation $A\boldsymbol{x} = \boldsymbol{y}$ always has a solution or the kernel of A^* is nontrivial. Here A^* is the adjoint of A, which is its transpose (resp. Hermitian transpose) if A has real (resp. complex) entries. The Fredholm alternative can be stated in a similar fashion for a special class of infinite-dimensional linear operators. In this case, $A = \lambda I - \mathcal{L}$, where $\mathcal{L}$ is a compact operator, λ is a nonzero complex number, and I is the identity operator (see for instance Reed & Simon, 1980; Kreyszig, 1989; Dunford & Schwartz, 1988).

Applications of this theorem involve multiple scales analysis and perturbation theory. Assume for instance that we want to solve the eigenvalue problem $\mathcal{L}\boldsymbol{x} = \lambda\boldsymbol{x}$, where the linear operator $\mathcal{L}$ can be written as $\mathcal{L} = \mathcal{L}_0 + \varepsilon\mathcal{L}_1$ and ε is a small parameter. If we look for expansions of the eigenvector $\boldsymbol{x} = \boldsymbol{x}_0 + \varepsilon\boldsymbol{x}_1 + \cdots$ and eigenvalue $\lambda = \lambda_0 + \varepsilon\lambda_1 + \cdots$ in powers of ε, substitute in the equation, and equate the left-and right-hand sides at each power of ε, we obtain a hierarchy of linear equations of the form

$$(\mathcal{L}_0 - \lambda_0 I)\, \boldsymbol{x}_i = \mathcal{J}_i, \quad i = 0, 1, \ldots, \tag{4}$$

where I is the identity operator and the right-hand side, $\mathcal{J}_i$, of (4) only depends on $\boldsymbol{x}_k$ for $k < i$. In particular, $\mathcal{J}_0 = \boldsymbol{0}$ and $\mathcal{J}_1 = \lambda_1\boldsymbol{x}_0 - \mathcal{L}_1\boldsymbol{x}_0$. At each step of this iterative process, the linear equation $(\mathcal{L}_0 - \lambda_0 I)\, \boldsymbol{x}_i = \mathcal{J}_i$ has a solution if and only if $\mathcal{J}_i$ is orthogonal to the kernel of the adjoint of $\mathcal{L}_0 - \lambda_0 I$. This typically gives the value of the coefficient λ_i of the expansion of the eigenvalue λ in powers of ε. Once λ_i is known, one can then solve the equation $(\mathcal{L}_0 - \lambda_0 I)\, \boldsymbol{x}_i = \mathcal{J}_i$ for $\boldsymbol{x}_i$. A similar method can be applied to solve equations of the form

$$\mathcal{L}\boldsymbol{x} = \mathcal{G}(\boldsymbol{x}), \tag{5}$$

where $\mathcal{L} = \mathcal{L}_0 + \varepsilon\mathcal{L}_1$ and $\mathcal{G}$ is a nonlinear function of $\boldsymbol{x}$. Such equations are often encountered in bifurcation theory. Assume for instance that $\mathcal{G}$ is polynomial, say $\mathcal{G} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle \boldsymbol{x}$, and that one knows a solution $\boldsymbol{x}_0$ to the linear equation $\mathcal{L}_0\boldsymbol{x}_0 = \boldsymbol{0}$. Small-amplitude solutions to the nonlinear problem can be approximated by introducing a small parameter that measures the size of $\boldsymbol{x}$ and rescaling. For instance, the scaling $\boldsymbol{x} = \sqrt{\varepsilon}\, \boldsymbol{y}$ turns (5) into $(\mathcal{L}_0 + \varepsilon\mathcal{L}_1)\, \boldsymbol{y} = \varepsilon\langle \boldsymbol{y}, \boldsymbol{y} \rangle \boldsymbol{y}$. A solution of the form $\boldsymbol{y} = \boldsymbol{y}_0 + \varepsilon\boldsymbol{y}_1 + \cdots$ can then be sought, using an iterative process similar to the one discussed above.

JOCELINE LEGA

See also **Perturbation theory**

Further Reading

Dunford, D. & Schwartz, J.T. 1988. *Linear Operators. Part I: General Theory*, New York and Chichester: Wiley

Kreyszig, E. 1989. *Introductory Functional Analysis with Applications*, New York and Chichester: Wiley

Reed, M. & Simon, B. 1980. *Methods of Modern Mathematical Physics. I. Functional Analysis*, Boston and London: Academic Press

FREE ENERGY

The concept of "free energy" is a key concept to characterize physically relevant states in statistical mechanics (see, e.g., Mandl, 1988; Reichl, 1998). Given an equilibrium system of statistical mechanics with energy levels E_i of the microstates i, the (Helmholtz) free energy F is defined as

$$F(\beta) = -\frac{1}{\beta} \log Z(\beta), \tag{1}$$

where

$$Z(\beta) = \sum_i \mathrm{e}^{-\beta E_i} \tag{2}$$

is the partition function and β is the inverse temperature. Apparently, the free energy is different from the internal energy U given by

$$U = -\frac{\partial}{\partial\beta} \log Z(\beta) = \langle E_i \rangle. \tag{3}$$

The difference is given by the entropy S times the absolute temperature T:

$$F = U - TS. \tag{4}$$

This equation can also be regarded as describing a Legendre transformation from U to F. Equilibrium states minimize the free energy (rather than the internal energy)—in this sense F is more relevant than U. The minimum of F can be achieved in two contrasting ways, either by making the internal energy U small, or by making the entropy S large (or both). Changes of the free energy are related to the maximum amount of work a system can do at constant pressure and temperature.

The basic principle underlying statistical mechanics, the maximum entropy principle, can also be formulated as a "principle of minimum free energy" (Beck & Schlögl, 1993). In this formulation, one starts from fixed intensive quantities (e.g., the inverse temperature β) and considers a free energy function $F[p]$ as a function of an a priori arbitrary set of probabilities $[p]=\{p_1, p_2, \ldots\}$. The probabilities p_i can describe any state, for example, also non-equilibrium and nonthermal states. One then requires

$$F[p] = \sum_i p_i E_i + \beta^{-1} p_i \log p_i = U - \beta^{-1} S \quad (5)$$

to have a minimum in the space of all possible probability distributions. This minimum is achieved for the canonical distribution

$$p_i = \frac{1}{Z} \mathrm{e}^{-\beta E_i}. \quad (6)$$

There are straightforward generalizations of this principle for systems with several intensive quantities, for example, grand canonical ensembles.

For nonlinear dynamical systems, the concept of free energy is often seen in a much broader sense. Various generalized types of free energies can be defined. A key ingredient for this more general approach is the so-called thermodynamic formalism of dynamical systems (Beck & Schlögl, 1993). One defines partition functions that contain information on certain fluctuating quantities associated with the dynamical system, for example, local singularities of the invariant density or local Lyapunov exponents. To proceed to a (formal) statistical mechanics description, one then defines a free energy function for these partition functions quite analogous to Equation (1).

Examples are "static," "dynamic," and "expansion" free energies of dynamical systems, as well as the so-called "topological pressure." Let us here consider the static free energy in somewhat more detail. This free energy function contains information on the spectrum of singularities of the invariant density of the dynamical system and is closely related to very important quantities that characterize multifractals, the so-called Rényi dimensions. One covers the attractor (or repeller) under consideration with small d-dimensional cubes ("boxes") of volume ε^d, where d is an integer dimension large enough to embed the attractor. The static partition function is then defined as

$$Z(\beta) = \sum_i p_i^\beta = \sum_i \mathrm{e}^{-\beta \alpha_i V}, \quad (7)$$

where p_i is the invariant probability measure associated with box i and β is a formal "inverse temperature" parameter. To illustrate the similarities with statistical mechanics, we have written in the above equation $p_i =: \varepsilon^{\alpha_i} = e^{-\alpha_i \cdot V}$, where α_i is like a fluctuating energy associated with each box i and $V = -\log \varepsilon$ plays the role of a formal volume variable. One can then study the free energy density associated with this partition function,

$$f(\beta) := \lim_{V \to \infty} -\frac{1}{\beta V} \log Z(\beta). \quad (8)$$

This static free energy, up to a trivial factor, is indeed identical with the Rényi dimensions D_β:

$$f(\beta) = \frac{\beta - 1}{\beta} D_\beta. \quad (9)$$

In this way, methods borrowed from statistical mechanics play an important role for the statistical description of dynamical systems.

While the above generalized free energies, of relevance for nonlinear dynamical systems, have only formal analogies with conventional free energies, a more fundamental generalization of statistical mechanics suitable for complex systems has been suggested by Tsallis (1988). In this so-called non-extensive statistical mechanics approach (Abe & Okamoto, 2001; Kaniadakis et al., 2002), the entire formalism of statistical mechanics is generalized by starting from more general entropy measures, the Tsallis entropies. These are defined by

$$S_q = \frac{1}{q-1} \left(1 - \sum_i p_i^q \right). \quad (10)$$

Here the p_i are probabilities associated with the microstates i of a physical system. Note that the Tsallis entropies look somewhat similar to the Rényi information measures but are indeed different (there is no logarithm). The parameter q, called "entropic index," can take on any real value but is in practice often close to 1. For $q \to 1$, one can easily check that the Tsallis entropies reduce to the ordinary Shannon entropy $S_1 = -\sum_i p_i \log p_i$. The principal idea of non-extensive statistical mechanics is to do everything we know from ordinary statistical mechanics, but start from the more general entropy measures. Naturally, ordinary statistical mechanics is contained as a special case ($q = 1$) in this more general formalism.

The maximum entropy principle yields power-law generalizations of the canonical ensemble,

$$p_i = \frac{1}{Z_q(\beta)}(1 + (q-1)\beta E_i)^{-1/(q-1)}, \qquad (11)$$

which reduce to Equation (6) for $q \to 1$. As a consequence, there is then also a more general free energy $F_q(\beta)$, which is parametrized by the entropic index q. One can basically show that the entire formalism of thermodynamics has simple q-generalizations, for example, for the generalized free energy there is a relation of the form

$$F_q = U_q - TS_q, \qquad (12)$$

where the index q indicates that the q-generalized canonical ensemble is chosen. This type of formalism has interesting physical applications, for example, for fully developed turbulence and for scattering processes in high-energy physics (Beck, 2002).

CHRISTIAN BECK

See also **Chaotic dynamics; Dimensions; Entropy; Fractals; Multifractal analysis**

Further Reading

Abe, S. & Okamoto, Y. (editors). 2001. *Nonextensive Statistical Mechanics and Its Applications*, Berlin and New York: Springer

Beck, C. & Schlögl, F. 1993. *Thermodynamics of Chaotic Systems*, Cambridge and New York: Cambridge University Press

Beck, C. 2002. Nonextensive methods in turbulence and particle physics. *Physica A*, 305: 209-217

Kaniadakis, G., Lissia, M. & Rapisarda, A. (editors). 2002. *Nonextensive Thermodynamics and Physical Applications*, Physica A, 305 (special volume)

Mandl, F. 1988. *Statistical Physics*, 2nd edition, Chichester and New York: Wiley

Reichl, L.E. 1998. *A Modern Course in Statistical Physics*, 2nd edition, New York: Wiley

Tsallis, C. 1988. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52: 479-487

FREE PROBABILITY THEORY

A new probabilistic context emerged in the 1980s (Voiculescu, 1985) from studies in operator algebras, a mathematical subject with ties to quantum physics. Free probability theory (FPT) is developing along a parallel to the basics of usual probability theory, drawn from assumptions designed for quantities with the highest degree of noncommutativity. FPT has turned out to have important connections to random matrix theory (Voiculescu, 1991) as well as to some combinatorics (Speicher, 1998) and to certain models in physics (Voiculescu, 1997).

To introduce FPT, one takes the view that "a probability theory" deals with expectation values $\varphi(a)$ for the elements a, called random variables, of a set A endowed with operations of addition, multiplication, and multiplication by scalars. We may describe FPT as noncommutative probability theory (i.e., the products ab and ba can be different for $a, b \in A$) with independence defined by a new relation called free independence. In usual probability theory, A consists of numerical functions on some space of events and clearly $ab = ba$ in this case. By contrast, noncommutative probability occurs typically when A consists of quantum mechanical observables Q. These are linear operators on a vector space $\mathcal{H}$ with an inner product $\langle \cdot, \cdot \rangle$ and $\varphi(Q) = \langle Q\xi, \xi \rangle$ where $\xi \in \mathcal{H}$ is a fixed vector with $\langle \xi, \xi \rangle = 1$. Under quantum mechanics, the usual definition of independence of Q_1 and Q_2 requires commutation $Q_1Q_2 = Q_2Q_1$ and $\varphi(f(Q_1)g(Q_2)) = \varphi(f(Q_1))\varphi(g(Q_2))$ for polynomials f and g. Free independence of Q_1, Q_2, by contrast requires that

$$\varphi(\dots f(Q_1)g(Q_2)h(Q_1)\dots) = 0 \qquad (1)$$

whenever

$$\begin{aligned} &\dots, \varphi(f(Q_1)) = 0, \varphi(g(Q_2)) = 0,\\ &\varphi(h(Q_1)) = 0, \dots, \end{aligned} \qquad (2)$$

where $\dots, f, g, h, \dots$ are polynomials and Q_1, Q_2 alternate as arguments. The definition of free independence extends to sets of variables by replacing the one-variable polynomials $f, g, h, \dots$ by corresponding polynomials in noncommuting variables from those sets. Note that if Q_1, Q_2 satisfy free independence, then in most cases $Q_1Q_2 \neq Q_2Q_1$.

The FPT analogue of the central limit theorem (Voiculescu, 1985; Voiculescu et al., 1992) states $\varphi(Q_j) = 0$, $\varphi(Q_j^2) = 1$ and $|\varphi(Q_j^k)| \leq C_k$, then

$$S_n = n^{-1/2}(Q_1 + \dots + Q_n) \qquad (3)$$

will have in the limit, the moments of a semicircle distribution, that is,

$$\lim_{n\to\infty} \varphi(S_n^k) = (2\pi)^{-1}\int_{-2}^{2} t^k(4 - t^2)^{1/2}\,dt. \qquad (4)$$

Many other topics in classical probability theory have FPT analogs. The list includes the following (Voiculescu et al., 1992; Biane & Speicher, 1998; Voiculescu, 2002):

- Addition and multiplication of freely independent random variables corresponds to highly nonlinear free convolution operations on the distributions.
- The FPT infinitely divisible laws and stable laws have been classified, and there are corresponding processes with free increments.
- Semicircular processes are the FPT analog of Gaussian processes, and in particular there is a concept of free Brownian motion.
- There are stochastic integrals with respect to free Brownian motion.

- There is a free entropy theory that deals with FPT analogues of Shannon's differential entropy and of the Fisher information.

The semicircle distribution, which plays the role of the Gauss law in FPT, was well known from Wigner's work as the limit-distribution of eigenvalues of large symmetric random matrices with independent identically distributed Gaussian entries (the Gaussian orthogonal ensemble). The explanation of this surprising coincidence was found to be that free independence often occurs asymptotically among large random matrices (Voiculescu, 1991). For instance, if for each n we are given two sets $\{A_1^{(n)}, \dots, A_m^{(n)}\}$ and $\{B_1^{(n)}, \dots, B_p^{(n)}\}$ of $n \times n$ Hermitian random matrices, then the requirement that their joint distribution (as matrix-valued random variables) be the same as that of $\{UA_1^{(n)}U^{-1}, \dots, UA_m^{(n)}U^{-1}, B_1^{(n)}, \dots, B_p^{(n)}\}$ for all unitary matrices U is the key to asymptotic free independence of the two sets (Voiculescu, 1998). For $n \times n$ random matrices T, the expectation functional is $\varphi_n(T) = n^{-1}E(TrT)$, where E denotes the expectation of the numerical random variable TrT. Asymptotic free independence means the equalities of free independence hold after taking the limit $n \to \infty$.

A random matrix T is thus a classical matrix-valued random variable and at the same time a noncommutative random variable. Passing to the noncommutative context discards information about the entries, since the expectation values $\varphi_n(T^k)$ remember only the eigenvalue distribution of T.

Asymptotic free independence results for random matrices imply that their limit behavior is governed by FPT. For instance, for the sum or product of random matrices satisfying free independence asymptotically, the limit distribution of eigenvalues is computed by the free convolution operations.

Among the simplest cases of asymptotic free independence are a pair of independent matrices from the Gaussian ensembles or a constant matrix and one from a Gaussian ensemble. Also, if X_n is a rectangular $[\alpha n] \times n$ Gaussian random matrix with independent and identically distributed (i.i.d.) entries, then $X_n^* X_n$ and a constant $n \times n$ random matrix will exhibit asymptotic free independence. Besides these simplest instances of asymptotic free independence, there are many others involving unitary matrices, random permutation matrices or matrices with non-Gaussian i.i.d. entries (Voiculescu, 2000). Also, Gaussian random band-matrices can be handled with more involved FPT techniques (Shlyakhtenko, 1996). Most asymptotic free independence results for complex Hermitian or unitary matrices also hold for real symmetric or orthogonal matrices. The differences between distributions of eigenvalue spacings between the real and complex cases occur at another scale, which is wiped out by taking the limit of the φ_n as $n \to \infty$.

FPT methods have also been used for physics models related to asymptotics of random matrices, like the large N Yang–Mills two-dimensional quantum chromodynamics in papers of I.M. Singer, R. Gopakumar and D. Gross, and M. Douglas (Voiculescu, 1997). Also occurring in physics are multimatrix-models, corresponding to ensembles of k-tuples of $N \times N$-matrices with the densities $c_N \exp(-NTrP(T_1, \dots, T_k))$ that have connections to the analog of entropy in FPT (Voiculescu, 2002). An indication about where FPT and quantum field theory connect is given by the combinatorial approach to FPT (Speicher, 1998) that deals with noncrossing partitions of $\{1, \dots, n\}$, that is, with planar Feynman diagrams.

Dan Virgil Voiculescu

See also **Quantum field theory; Random matrix theory I, II, III, IV**

Further Reading

Biane, P. & Speicher, R. 1998. Stochastic calculus with respect to free Brownian motion and analysis on Wigner space. *Probability Theory and Related Fields*, 112: 373–409

Shlyakhtenko, D. 1996. Random Gaussian band matrices and freeness with amalgamation. *International Mathematics Research Notices*, 20: 1013–1025

Speicher, R. 1998. Combinatorial theory of the free product with amalgamation and operator-valued free probability theory. *Memoirs of the American Mathematical Society*, 627

Voiculescu, D. 1985. Symmetries of some reduced free product C^*-algebras. In *Operator Algebras and Their Connections with Topology and Ergodic Theory*, edited by H. Araki et al., Berlin and New York: Springer

Voiculescu, D. 1991. Limit laws for random matrices and free products. *Inventiones Mathematicae*, 104: 201–220

Voiculescu, D.V. (editor). 1997. *Free Probability Theory*, Providence, RI: American Mathematical Society

Voiculescu, D. 1998. A strengthened asymptotic freeness result for random matrices with applications to free entropy. *International Mathematics Research Notices*: 41–63

Voiculescu, D. 2000. Lectures on free probability theory. In *Lectures on Probability Theory and Statistics, École d'Été de Probabilities de Saint–Flour XXVIII –1998*, edited by M. Emery, A. Nemirovski & D. Voiculescu, New York: Springer

Voiculescu, D. 2002. Free entropy. *Bulletin of the London Mathematical Society*, 34: 257–278

Voiculescu, D.V., Dykema, K.J. & Nica, A.M. 1992. *Free Random Variables*, Providence, RI: American Mathematical Society

FRENKEL–KONTOROVA MODEL

In 1938, Yakob Frenkel and Tatiana Kontorova developed in some detail a model originally proposed by Ulrich Dehlinger (see Nabarro, 1967) for studying stationary dislocations in a crystal, and its motion. A perfect crystalline (regular lattice) solid subjected to small deformations returns to a perfect lattice arrangement of atoms after the forces causing the deformations are released. This regime of small deformations is called “elastic.” Larger deformations can lead to rearrangements of the crystal lattice. In this other regime, called “plastic,” the perfect lattice is not restored after tension is released. The attempt to describe the energetics involved in those irreversible rearrangements (or lattice dislocations, see Figure 1(b)) led to the posing of a theoretical physics problem of a discrete elastic one-dimensional field of displacements (atomic positions) experiencing a spatially periodic potential (Dehlinger, 1929; Frenkel & Kontorova, 1939; Frank & Van der Merwe, 1949). Figure 1(c) shows a schematic representation of Dehlinger’s *Verhakung* (interlocking), as he named his idea.

The Frenkel–Kontorova (FK) model has been used for modeling nanostructural properties of adsorbed monolayers of atoms on crystalline surfaces, nonlinear magnetic excitations in magnets, ionic conductors, and many other solid-state physics problems. Today it has become a textbook mathematical model with direct applications in condensed matter, statistical physics, nonlinear fields, and material science, often under different names, notably the “discrete sine-Gordon” or “periodic Klein–Gordon” equation.

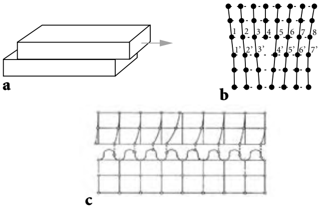

Figure 1. (a) Schematic net outcome of a plastic deformation. Relative shift between layers does not occur uniformly: The slip process can take place sequentially by the mechanism of dislocation motion shown in (b). (c) Dehlinger’s picture of the imperfection, that he named *Verhakung*, where the atoms of the upper row, in this displacement, experience a periodic potential on account of their interaction with atoms in the lower row.

The Generalized Frenkel–Kontorova Model

In the generalized FK model, a discrete one-dimensional field (u_n) $(-\infty < n < +\infty)$, has a potential energy density

$$H(u_{n+1}, u_n) = KV(u_n) + W(\Delta u_n), \qquad (1)$$

where $W(\Delta u)$ is the interaction between neighbor components of the field, $V(u)$ is a periodic function $V(u+1) = V(u)$, and K is a parameter controlling the relative strength of both (V and W) contributions to the energy density.

The standard version of the model corresponds to the simplest choices of sinusoidal on-site potential V and harmonic spring interaction W:

$$V = \frac{1}{(2\pi)^2}\left[1 - \cos(2\pi u)\right],$$
$$W(\Delta u) = \frac{1}{2}(\Delta u)^2 - \sigma \Delta u, \qquad (2)$$

where σ is the unstretched length of the spring, the value at which W is a minimum.

Keeping V as above, the following choice for W

$$W = \frac{1}{(2\pi)^2}\left[1 - \cos(2\pi(\Delta u - \gamma))\right] \qquad (3)$$

(known as a “chiral XY model” in a magnetic field) models a system of magnetic moments with strong planar (XY) anisotropy (here described by the angles $2\pi u_n$), with chiral (preferred relative angle γ) interactions W, and K in Equation (1) measuring the intensity of the external magnetic field. Other choices are motivated by different physical situations of interest.

To complete the formal definition of the generalized FK model, one must define its time evolution and this choice depends on the specific application one has in mind. The standard model with Hamiltonian dynamics

$$\frac{\mathrm{d}^2 u_n}{\mathrm{d}t^2} = u_{n+1} - 2u_n + u_{n-1} - \frac{K}{2\pi}\sin(2\pi u_n) \qquad (4)$$

corresponds to the simplest discretized version of the integrable partial differential equation known as sine-Gordon (SG):

$$\frac{\mathrm{d}^2 u}{\mathrm{d}t^2} = \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} - \sin(2\pi u). \qquad (5)$$

This justifies the name of “discrete SG model” often used for the FK model.

Some physical applications, for example, in Josephson junction devices, suggest introducing dissipative terms in Equation (4) (involving first-time derivatives $\dot{u}_n$) and more complicated functional forms of the right-hand-side dependence on u_n and u_{n+1}. The richness of dynamical behaviors shown by the different dynamical FK models is a continuing source of interest in both theory development and applications (Braun & Kivshar, 1998, 2004; Floría & Mazo, 1996).

The Equilibrium Problem for Convex FK Models

A simpler problem, but quite complex, is the characterization of equilibrium properties of the generalized FK model. The physical motivation for studying this problem comes mainly from the observed abundance of modulated phases in minerals and man-made materials and compounds, and the need to understand the peculiar multiphase diagrams shown by experimental studies.

For the case in which W is not a convex function, like the chiral XY example above, the difficulties of a complete rigorous characterization of equilibria are much greater than in the case of convex W interaction. For the latter, Aubry's work allowed a satisfactory theory in the infinite system size (thermodynamic) limit, by showing a one-to-one correspondence between equilibrium configurations (u_n), which satisfy the (force-equilibrium) equations

$$\frac{\partial(H(u_n, u_{n-1}) + H(u_{n+1}, u_n))}{\partial u_n} = 0 \qquad (6)$$

and orbits of a symplectic twist map of the cylinder.

When an equilibrium configuration is such that any segment of length $L = M - N + 1$, $(u_N, \ldots, u_M)$, can only increase its energy as result of arbitrary fluctuations of $u_{N+1}, \ldots, u_{M-1}$, the field configuration is called a minimum energy configuration (MEC).

There are two types of MEC, which are both recurrent—meaning that given any integer p and any number $\varepsilon > 0$, there are integers m and $q > 0$ such that both inequalities: $|u_{p+q} - u_p - m| < \varepsilon$ and $|u_{p+q+1} - u_{p+1} - m| < \varepsilon$ are satisfied. These two types of recurrent MEC are called commensurate (or periodic) phases and incommensurate (or quasi-periodic) phases, respectively. Remarkably, there are nonrecurrent MECs, which are rigorous discrete versions of the soliton solutions of continuous partial differential equations (like the integrable sine-Gordon model). These are called "discommensurations" or "homoclinic field configurations," and they correspond to the original motivation (Dehlinger's *Verhakung*) of the FK model: localized dislocations over the recurrent (periodic or quasi-periodic) substrate phase.

Commensurate Phases

These are periodic configurations with rational p/q average spacing ω (defined as $\lim_{L\to\infty}(u_M - u_N)/L$), where $u_{n+p} - q = u_n$. Such field configurations are generically pinned, meaning that one has to put some finite energy $E_{\rm PN}$ on the system to displace the configuration over the path-dependent barriers of energy (called "Peierls–Nabarro barriers") separating configurations that are equivalent by translations σ_{mn}.

A pinned configuration has a finite coherence length (or decay range) ξ of fluctuations, meaning that if a field component (the center u_{n_0}) is kept displaced by an external force, the rest of components u_n are displaced with exponentially decaying amplitude $\exp(-|n - n_0|/\xi)$. Also, the linear spectrum of small oscillations (phonon spectrum) of a pinned configuration has a finite gap $\Delta > 0$.

Incommensurate or Quasi-periodic Phases

Recurrent configurations with irrational average spacing ω can be viewed as limits of sequences of periodic phases of mean spacing $\omega_n \to \omega$ as $n \to \infty$. The physical properties of these structures depend on the parameter K of the model: For each irrational ω there is a critical value $K_{\rm c}(\omega)$ such that:

- For $K < K_{\rm c}$, the quasi-periodic MEC is sliding (unpinned), so that $E_{\rm PN} = 0$. Consequently, localized induced fluctuations extend through the system so that the decay range diverge, $\xi \to \infty$. The spectrum of phonons of a sliding configuration is gapless, $\Delta = 0$. The distribution function of the fractional parts Frac(u_n) in this phase is a continuous function.
- For $K > K_{\rm c}$, the incommensurate phase is pinned, $E_{\rm PN} > 0$, the decay range is finite, and a gap Δ appears in the phonon spectrum. The distribution function of the values Frac(u_n) is now a Cantor function; there are infinitely many forbidden intervals (gaps) of values of the field (modulo 1).

MacKay (1993) has characterized this drastic change in the physical properties of quasi-periodic fields as a critical phenomenon (already called by Aubry as a "transition by breaking of analiticity") using renormalization group methods for the case of noble irrational ω and the standard model (2).

Discrete Solitons (DSs) or Discommensurations (DCs)

Field configurations can connect asymptotically ($n \to \pm\infty$) two commensurate configurations (usually one is a translation of the other, though they can be different in general) by an energy minimizing path of exponentially localized length (the width of the DS) around a lattice site (the center of the DS). The width is of the order of the decay range ξ of the recurrent (periodic) substrate.

These structural defects are the elementary excitations of the modulated phase, or localized compressions (advanced DCs) or expansions (retarded DCs) superimposed on the regular substrate modulation. If ω is close to a rational ω_0, the modulated phase corresponding to a mean spacing ω can be correctly described as an array of localized DCs (advanced if $\omega > \omega_0$, retarded if $\omega < \omega_0$) over a periodic substrate of mean spacing ω_0. The interaction energy between neighboring DCs decays exponentially with the quotient interdistance/width, $\exp(-1/(\xi|\omega - \omega_0|))$, so that for ω very close to ω_0 and/or high values of the pinning parameter K (ξ small), the elementary excitations are almost non-interacting (and pinned). However, at lower values of K the tail overlapping of the DCs makes the interactions strong enough, and the array of strongly interacting DCs becomes sliding.

Though with less rigorous formal basis, the DC theory of modulated phases is a well-established formal description of equilibrium for nonconvex FK models. In these cases, interactions between DCs can be of either (attractive or repulsive) sign and even alternating with interdistance. Also, the inner structure of DC can be more complex than in the convex case.

LUIS MARIO FLORÍA AND PEDRO JESÚS MARTÍNEZ

See also **Aubry–Mather theory; Commensurate-incommensurate transition; Dislocations in crystals; Josephson junction arrays; Peierls barrier; Sine-Gordon equation**

Further Reading

Braun, O.M. & Kivshar, Yu. S. 1998. Nonlinear dynamics of the Frenkel–Kontorova model. *Physics Reports*, 306: 1–108

Braun, O.M. & Kivshar, Yu. S. 2004. *The Frenkel–Kontorova Model*, Berlin and New York: Springer

Dehlinger, U. 1929. Zur theorie der rekristallisation reiner metalle [On the theory of recrystallization in pure metals]. *Annalen der Physik*, (5)2: 749–793

Floría, L.M. & Mazo, J.J. 1996. Dissipative dynamics of the Frenkel–Kontorova model. *Advances in Physics*, 45: 505–598

Frank, F.C. & van der Merwe, J.H. 1949. One dimensional dislocations I & II. *Proccedings of the Royal Society* (*London*) A, 198: 205–225

Frenkel, Y. & Kontorova, T. 1939. On the theory of plastic deformation and twinning. *Journal of Physics* (*Moscow*), 1: 137–149 (original Russian publication 1938)

MacKay, R.S. 1993. *Renormalisation in Area-preserving Maps*, Singapore: World Scientific

Nabarro, F.R.N. 1967. *Theory of Crystal Dislocations*, Oxford: Clarendon Press

FREQUENCY DOUBLING

Frequency doubling, or second-harmonic generation, is the process in which an intense laser beam propagating through a nonlinear optical medium produces light at double the frequency of the input beam (Bloembergen, 1969; Boyd, 2002; Shen, 1984; Zernike & Midwinter, 1973). It was the first laser-induced nonlinear optical effect to be reported, and its observation by Franken et al. (1961) marked the beginning of the field of nonlinear optics. This work involved the use of ruby laser radiation at 694.3 nm propagating though a quartz crystal to produce second-harmonic radiation at 347.2 nm (Franken et al., 1961).

Our laboratory demonstration of second-harmonic generation is shown in Figure 1 (also in color plate section). The 800 nm radiation of a titanium-doped sapphire laser is converted into the second harmonic at 400 nm in passing through a lithium niobate crystal. A prism is used to separate the second harmonic from the fundamental radiation exiting of the crystal. A side view showing the ray trajectories is represented in Figure 1c. We can see the red light of the laser beam illuminating the crystal and the blue light of the second harmonic exiting the crystal, together with some portion of unconverted red light. Typically, not all of the incident radiation is converted into the second harmonic. The quantity describing the effectiveness of converting the incident radiation into a harmonic is called the conversion efficiency.

The physical origin of the process of second-harmonic generation can be understood as follows. The electric field of a monochromatic optical wave incident upon a nonlinear medium forces the electrons bound in atoms and molecules to oscillate about their equilibrium positions. Thus, it induces a polarization (dipole moment per unit volume) in the medium that depends on the strength of the electric field. This polarization plays the role of a secondary source of electromagnetic radiation.

The Lorentz model, which treats the atom as a harmonic oscillator, provides a good description of the linear atomic response. The electric force F_{E} exerted on the electron by the electromagnetic field displaces the electron with respect to the nucleus. The restoring force F_{rest} has the opposite action, tending to return the electron to its equilibrium position. If the applied field is weak, the restoring force is linear with respect to the electron displacement x from equilibrium, that is,

$$F_{\mathrm{rest}} = -m\omega_0^2 x.$$

Here m is the mass of the electron and ω_0 is the resonance frequency of the oscillator. Therefore, the potential energy function of the medium is quadratic with respect to x (see Figure 2a, the solid line):

$$U = -\int F_{\mathrm{rest}}\,\mathrm{d}x = \tfrac{1}{2}m\omega_0^2 x^2.$$

The parabolic shape of the potential energy function describes harmonic oscillations of the electron at the applied optical field frequency. In such a way, the induced polarization gives rise to the radiation at the applied field frequency. Figure 2b shows the waveform of the incident monochromatic electromagnetic wave of frequency ω (the top graph). For the case of a medium with linear response, there is no distortion of the waveform associated with the polarization of the medium (the middle graph).

In order to describe the nonlinear response, the model has to be extended by allowing the possibility of a nonlinearity in the restoring force (Boyd, 2002). The lowest-order nonlinear response is modelled by adding a contribution quadratic in the displacement so that

$$F_{\mathrm{rest}} = -m\omega_0^2 x - max^2.$$

Here a characterizes the strength of the nonlinearity. The potential energy function of such a system has an additional term that is cubic with respect to x (Figure 2a, the dashed line):

$$U = -\tfrac{1}{2}m\omega_0^2 x^2 + \tfrac{1}{3}max^3.$$

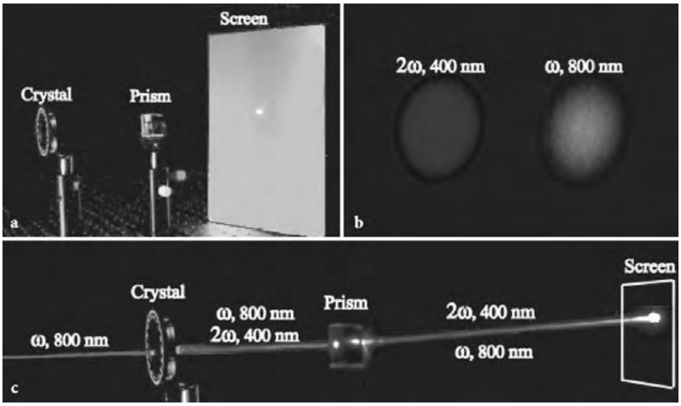

Figure 1. Second-harmonic generation in lithium niobate: (a) Experimental setup; (b) close-up view of the observation screen; (c) side view showing ray trajectories.

Because of the nonlinearity of the restoring force, the atomic response can show a harmonic distortion of the waveform associated with polarization oscillation (Figure 2b the bottom graph). In this case, the induced polarization will give rise to radiation not only at the same frequency as the applied field, but also at higher frequencies (harmonics) as well. A nonlinear effect can occur only if the applied field is sufficiently strong to produce a large displacement x.

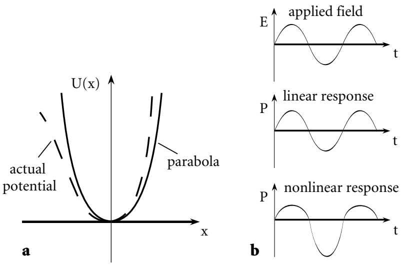

Figure 2. (a) Potential energy function in case of linear response (solid line) and nonlinear noncentrosymmetric response (dashed line); (b) polarization waveforms associated with the atomic response.

More formally, the nonlinear response of a medium to a strong incident field can be described by a nonlinear susceptibility (Bloembergen, 1969; Boyd, 2002; Shen, 1984). In the case of a linear response to the applied field, the induced polarization P depends linearly upon the electric field strength E according to the relationship

$$P = \chi^{(1)}E, \tag{1}$$

where the constant of proportionality $\chi^{(1)}$ is the linear susceptibility. In the case of a nonlinear response, optical response can be described by expanding polarization P as a power series in the field strength E:

$$P = \chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots . \tag{2}$$

The quantities $\chi^{(2)}$ and $\chi^{(3)}$ represent the second- and third-order nonlinear optical susceptibilities, respectively. Here, we will consider only the second-order nonlinear response.

Second-order nonlinear interactions can occur only in noncentrosymmetric media. A typical value of the second-order nonlinear susceptibility in crystals is $\chi^{(2)} \simeq 5 \times 10^{-8}$ esu (6.3×10^{-7} SI). We refer to

$$P^{(2)} = \chi^{(2)}E^2 \tag{3}$$

as the second-order nonlinear polarization. The strength of the applied optical field can be represented by

$$E = E_0 \exp^{-\mathrm{i}\omega t} + \mathrm{c.c.} \tag{4}$$

Substituting (4) into the equation for the second-order nonlinear polarization (3), we can obtain

$$P^{(2)} = 2\chi^{(2)}EE^* + (\chi^{(2)}E^2 \exp^{-\mathrm{i}2\omega t} + \mathrm{c.c.}). \tag{5}$$

We can see from (5) that the second-order polarization consists of a constant contribution (the first term) and a contribution at frequency 2ω (the second term). This doubled frequency contribution leads to the generation of radiation at the second-harmonic frequency 2ω.

Frequency doubling can also be understood by means of a virtual energy level description (Figure 3) (Boyd, 2002). The diagram shows that for a lossless

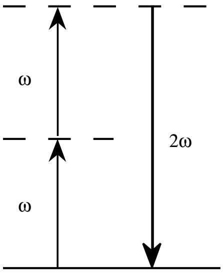

Figure 3. Energy-level diagram describing second-harmonic generation.

medium, the creation of a photon at a doubled frequency 2ω must be accompanied by the destruction of two photons at frequency ω, or according to Manley–Rowe relations: the rate at which a photon at frequency 2ω is created is equal to the rate at which two photons at frequency ω are destroyed.

Second-harmonic generation is a process in which the initial and final quantum mechanical states (energy levels) are identical. Therefore, the energy conservation and momentum conservation laws need to be satisfied. According to the momentum conservation law, the momentum of the two incident photons of frequency ω, $2\hbar\boldsymbol{k}_\omega$, must be equal to the momentum of the radiated photon of frequency 2ω, $\hbar\boldsymbol{k}_{2\omega}$:

$$2\hbar\boldsymbol{k}_\omega = \hbar\boldsymbol{k}_{2\omega}. \tag{6}$$

Here $\boldsymbol{k}_\omega$ and $\boldsymbol{k}_{2\omega}$ are the wave vectors of the incident wave and the second harmonic, respectively. According to (6), in order for the second-harmonic generation to effectively occur, it is necessary for the following phase matching condition to be satisfied:

$$2\boldsymbol{k}_\omega = \boldsymbol{k}_{2\omega}. \tag{7}$$

The dispersion relation for light propagating in a medium is

$$k_\omega = n(\omega)\frac{\omega}{c}, \tag{8}$$

where $n(\omega)$ is the index of refraction for the wave of frequency ω. Thus, condition for perfect phase matching with collinear beams (7) requires that

$$n(\omega) = n(2\omega). \tag{9}$$

In practice, the condition of perfect phase matching is often difficult to achieve because the refractive index of materials displays an effect of normal dispersion: $n(\omega)$ increases monotonically with ω. Thus, condition (9) cannot be satisfied.

There are several ways to achieve phase matching. One of the most common ways is to use the birefringence displayed by many crystals. Birefringence is the dependence of the refractive index on the direction of polarization of the optical radiation. In order to achieve phase matching through the use of birefringent crystals, the highest-frequency wave (2ω) should be polarized in the direction that gives it the lower of two possible refractive indices. There is another technique, known as quasi-phase matching, that can be used when normal phase matching cannot be implemented (Armstrong et al., 1962; Boyd, 2002). This technique involves the use of periodically poled materials, that are fabricated so that the orientation of one of the crystalline axis is inverted periodically as a function of position within the material. It results in periodic alternation of the sign of $\chi^{(2)}$, that can compensate for a nonzero wave vector mismatch $\Delta\boldsymbol{k}$.

Second-harmonic generation can usefully convert the coherent output of a fixed-frequency laser to a different spectral region (Zernike & Midwinter, 1973). For example, the infrared radiation of an Nd:YAG laser operating at 1064 nm can be converted into 532 nm visible radiation with a conversion efficiency of more than 50%.

KSENIA DOLGALEVA, NICK LEPESHKIN, AND
ROBERT W. BOYD

See also **Harmonic generation; Manley–Rowe relations; Nonlinear optics**

Further Reading

Armstrong, J.A., Bloembergen, N., Ducuing, J. & Pershan, P.S. 1962. Interaction between light waves in a nonlinear dielectric. *Physical Review*, 127: 1918–1939

Bloembergen, N. 1969. *Nonlinear Optics*, New York: Plenum Press

Boyd, R.W. 2002. *Nonlinear Optics*, San Diego: Academic Press

Franken, P.A., Hill, A.E., Peters, C.W. & Weinreich, G. 1961. Generation of optical harmonics. *Physical Review Letters*, 7: 118–119

Shen, Y.R. 1984. *The Principles of Nonlinear Optics*, New York: Wiley

Zernike, F. & Midwinter, J.E. 1973. *Applied Nonlinear Optics*, New York: Wiley

FREQUENCY LOCKING

See **Coupled oscillators**

FRÖHLICH THEORY

Starting in the late 1960s and continuing until his death in 1991, Herbert Fröhlich developed a theory of biological coherence that was based on quantum interactions between dipolar constituents of biomolecules. Fröhlich advocated momentum-space correlations within a living system such as a membrane, a cell, or an organism. This dynamic order would be a characteristic feature that distinguishes living systems from inanimate matter. The key assumptions

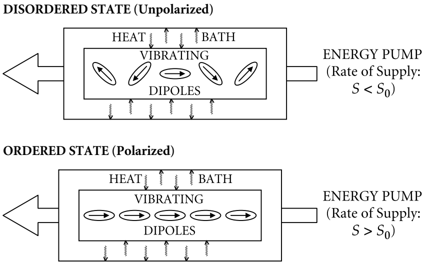

Figure 1. Illustration of the Fröhlich model.

of Fröhlich's theory can be listed as follows:

(i) a continuous supply of metabolic energy (also referred to as energy pumping) above a threshold level,
(ii) the presence of thermal noise due to physiological temperature,
(iii) internal organization of the biosystem that promotes functional features,
(iv) the existence of a large transmembrane potential difference, and
(v) a nonlinear interaction between two or more types of degrees of freedom.

As a result of these nonlinear interactions, in addition to the global minimum that characterizes a biological system in its nonliving state, a metastable energy minimum is achieved in the living state.

The Fröhlich model of biological coherence is based on a condensate of quanta of collective polar vibrations. It is a non-equilibrium property due to the interactions of the system with both the surrounding heat bath and an energy supply (see Figure 1). The energy is channeled into a single collective mode that becomes strongly excited. Most importantly, the Fröhlich model relies on the nonlinearity of internal vibrational mode interactions and in this respect is somewhat reminiscent of the laser action principle. Associated with this dynamically ordered macroscopic quantum state is the emergence of polarization due to the ordering of dipoles in biomolecules such as membrane head groups. Furthermore, Fröhlich predicted the generation of coherent modes of excitation such as dipole oscillations in the microwave frequency range. Nonlinear interactions between dynamic degrees of freedom were predicted to result in the local stability of the polarized state and in the long-range frequency-selective interactions between two systems with these properties.

In the search for empirical support of this model, Fröhlich placed an emphasis on the presence of dipole moments in many biomolecular systems that would then oscillate in synchrony in the frequency range of $10^{11}-10^{12}$ Hz due to their nonlinear interactions. Consequently, due to the resonant dipole-dipole coupling in a narrow frequency range, the entire biological system can be seen as a giant oscillating dipole. An alternative picture developed within the Fröhlich theory was that of a Bose–Einstein condensation in the space of dipole oscillations. The Hamiltonian postulated by Wu and Austin (1977) takes the form

$$\begin{aligned} H = {} & \sum_i \omega_i a_i{}^\dagger a_i + \sum_i \Omega b_i{}^\dagger b_i + \sum\nolimits_i \Theta_i P_i{}^\dagger P_i \\ & + \frac{1}{2} \sum_{i,j,k} (\chi a_i{}^\dagger a_j b_k{}^\dagger + \chi^* a_j a_i{}^\dagger b_k) \\ & + \sum_{i,j} (\lambda b_i a_j{}^\dagger + \lambda^* b_i{}^\dagger a_j) \\ & + \sum_{i,j} (\xi P_i a_j{}^\dagger + \xi^* P_i{}^\dagger a_j), \end{aligned} \tag{1}$$

where $(a_i{}^\dagger, a_i)$, $(b_i{}^\dagger, b_i)$, and $(P_i{}^\dagger, P_i)$ are the cell, heat-bath, and energy-pump creation and annihilation (boson-type) operators, respectively. A kinetic rate equation was derived for this model that indicates Bose-type condensate in the frequency domain with a stationary occupation number dependence of the dipole modes given by

$$N_i = \left[\mathrm{e}^{\beta(\omega_i - \mu)} - 1 \right]^{-1}. \tag{2}$$

The nonlinear coupling comes from the dipole-phonon interaction proportional to χ. Provided that the oscillating dipoles are within a narrow band of resonance frequencies ($\omega_{\min} \le \omega_i \le \omega_{\max}$) and the coupling constants (χ, λ and ξ) are large enough, long-range ($\sim 1\ \mu\text{m}$) attractive forces are expected to act between the dipoles.

The effective potential between two interacting dipoles which initially vibrate with frequencies ω_1 and ω_2 is given by

$$U(r) = \frac{E}{r^3} + \frac{F}{r^6}. \tag{3}$$

The van der Waals coefficient F is negative (attractive) unless the average distance between the interacting particles is very short. However, the van der Waals attraction between particles rapidly decreases with distance making it an unlikely candidate mechanism for biological coherence. The coefficient E can be both positive (repulsion) and negative (attraction) depending on the occupation of the new eigenstates of the system. Population inversion leads to attraction while thermal equilibrium for occupation numbers leads to repulsion between particles. In particular, E is given by

$$E = \frac{\hbar e^2 Z \mid \bar{A} \mid \gamma}{4\bar{\omega} M} \left(\frac{1}{\varepsilon'(\omega_+)} - \frac{1}{\varepsilon'(\omega_-)} \right), \tag{4}$$

where M is the mass of a dipole, e is the electron charge, Z is the number of elementary charges on each dipole, $\bar{A}$ is an angle constant, and $\varepsilon'(\omega)$ is the real part of the frequency-dependent dielectric constant. The quantities $\omega_\pm$ are the new dipole vibration frequencies

$$\omega_\pm = \left[\frac{1}{2}({\omega_1}^2+{\omega_2}^2)\pm\left(\frac{1}{4}\left({\omega_1}^2-{\omega_2}^2\right)+\frac{{\beta_0}^4}{(\varepsilon'_\pm)^2}\right)^{1/2}\right]^{1/2}, \tag{5}$$

where $\beta_0=\gamma^2e^2\sqrt{Z_1Z_2}/Mr^3$. In the resonant frequency case, the effective interaction energy between two oscillating dipoles was found to be long-range, depending on distance as r^{-3}.

Most of the expected condensation of dipolar vibrations was envisaged by Fröhlich to take place in the cell membrane due to its strong potential on the order of 10–100 mV across the thickness of 5–10 nm giving an electric field intensity of 1–20 × 10^6 V/m. The resultant dipole-dipole interactions were calculated to show a resonant long-range order at a high-frequency range of $10^{11}-10^{12}$ Hz with a propagation velocity of about 10^3m/s. In addition to membrane dipoles, several other candidates for Fröhlich coherence were considered, including double ionic layers, dipoles of DNA and RNA molecules, and plasmon oscillations of free ions in the cytoplasm.

Applications of the Fröhlich theory were made to explain cancer proliferation, where a shift in the resonant frequency was seen to affect the cell-cell signaling, the brain waves, and enzymatic chemical reactions.

Various experiments have been reported that appeared to demonstrate the sensitivity of metabolic processes to certain frequencies of electromagnetic radiation above the expected Boltzmann probability level. Raman scattering experiments by Webb (1980) found non-thermal effects in *E. coli* but these could not be reproduced by other labs. Irradiation of yeast cells by millimeter waves showed increased growth at specific frequencies (Grundler, et al., 1983). Rouleaux formation of human erythrocytes was explained in terms of Fröhlich's resonant dipole-dipole attraction (Paul et al., 1983) but did not rule out standard coagulation processes. While some of these experiments illustrate nonthermal effects in living matter that would require nonlinear and non-equilibrium interactions for explanation, no unambiguous experimental proof has been furnished to date to support Fröhlich's hypothesis.

It is interesting to note that while Fröhlich sought evidence for frequency selection in biological systems, another famous physicist, Alexander S. Davydov tried to find spatial localization of vibrational energy in biological systems such as DNA and peptides. There is a conceptual link between these two approaches as shown by Tuszyński et al. (1983) that involves self-focusing in reciprocal (Fröhlich) or real space (Davydov).

JACK A. TUSZYŃSKI

See also **Bose–Einstein condensation; Cluster–coagulation; Davydov soliton; Synergetics**

Further Reading

Davydov, A.S. 1982. *Biology and Quantum Mechanics*, Oxford: Pergamon Press

Fröhlich, H. 1968. Long range coherence and energy storage in biological systems. *International Journal of Quantum Chemistry*, 2: 641–649

Fröhlich, H. 1972. Selective long range dispersion forces between large systems. *Physics Letters*, 39A: 153–155

Fröhlich, H. 1980. The biological effects of microwaves and related questions. In *Advances in Electronics and Electron Physics*, vol. 53, edited by L. Marton and C. Marton, New York: Academic Press, pp. 85–152

Grundler, W., Eisenberg, C.P. & Sewchand, L.S. 1983. *Journal of Biological Physics*, 11: 1

Paul, R., Chatterjee, R., Tuszyński, J.A. & Fritz, O.G. 1983. Theory of long-range coherence in biological systems. I: the anomalous behaviour of human erythrocytes. *Journal of Theoretical Biology*, 104: 169

Tuszyński, J.A., Paul, R., Chatterjee, R. & Sreenivasan, S.R. 1983. Relationship between Fröhlich and Davydov models of biological order. *Physical Review* A, 30: 2666

Webb, S.J. 1980. Laser-Raman spectroscopy of living cells. *Physics Reports*, 60: 201–224

Wu, T.M. & Austin, S. 1977. Bose condensation in biosystems. *Physics Letters* A, 64: 151

FRONT PROPAGATION

See **Reaction-diffusion systems**

FRUSTRATION

Frustration is the inevitable consequence of trying to simultaneously satisfy a number of conflicting constraints. In general, when faced with such conflicting requirements one's only way out is to compromise. Inevitably, there are usually many widely different ways to compromise, some better than others. In physical systems, frustration arises from competition between different contributions to the total energy of a system. Generally, each contribution can be separately minimized by a unique choice of state but the different contributions have different minimum-energy states. The presence of frustration then typically leads to a situation in which there are many, very different, states with similar low energies separated from one another by "barriers" of much higher energy states. This leads to a proliferation of low-energy dynamical modes in the system and, especially in the presence of disorder, slow (glassy) dynamics.

A simple physical example that shows the way in which frustration leads to complexity is afforded by the Ising model (*See* **Spin systems**). Each site of a regular lattice bears a simple binary variable—a spin than can

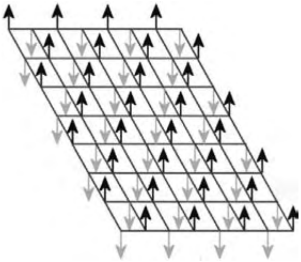

Figure 1. The ground state configuration of the 2-d Ising model on a square lattice.

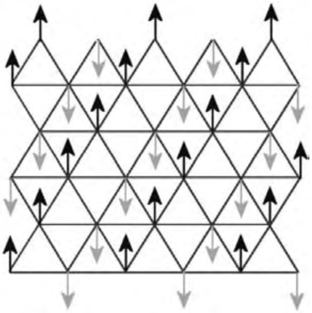

Figure 2. One of the degenerate ground state configurations of the 2-d Isig model on a triangular lattice.

point either up or down. The energy of the whole system has a contribution from each bond on the lattice. A given bond contributes an energy $+J$ if the spins at either end of the bond point in the same direction and an energy $-J$ if they point in opposite directions (this corresponds to the anti-ferromagnetic Ising model). First, consider the model on a two-dimensional square lattice with N sites (this model was solved exactly by Lars Onsager in 1940) as shown in Figure 1.

We can divide the lattice into two sub-lattices such that each site on one sub-lattice is surrounded entirely by sites on the other sub-lattice. The lowest energy state of the system is trivially found—we make all the spins on one sub-lattice point in one direction and all the spins on the other sub-lattice point in the opposite direction (clearly there are two such states with the same energy related by global inversion of all the spins). Each bond of the lattice connects two oppositely oriented spins so that the total energy is $-2NJ$. Any other configuration of spins must result in at least four bonds connecting like spins and so has energy $\geq -2JN + 4J$. This situation is simple and unfrustrated—all bonds can be simultaneously "happy."

Now consider the same simple model, with the same rules for assigning energy, but on a triangular lattice as shown in Figure 2. It is fairly simple to see that there is no way of simultaneously making all the bonds happy. Each elementary plaquette of the lattice has three spins, each of which wants to be in a different state to its neighbors—at least one bond must be unsatisfied (i.e., have energy $+J$). The minimum energy states of this system have alternate "stripes" of aligned spins so that each plaquette has only one unsatisfied bond giving a minimum energy of $-NJ$. A little thought will tell one that there are six equivalent configurations with this minimum energy: three choices for the direction of the stripes as well as the possibility of reversing all spins.

A very simple situation in which there is geometrical frustration arises when one considers the adsorption of one material onto a surface of a different material. Suppose that we have a substrate of material A, which is flat, and we deposit atoms of material B onto it. Further suppose that a bulk sample of pure A atoms has a lattice parameter a, while a bulk sample of pure B atoms has a lattice parameter b. When only a few atoms of material B have been deposited, they will simply bond with A atoms at the surface. Once a large number of B atoms have been deposited, they will begin to come into contact and be able to bond with each other. In the case that $b \approx a$, there is no frustration—the B atoms simply attach themselves to surface A atoms and quite happily bond to other B atoms when they are adjacent. This happy situation is referred to as epitaxial growth and happens for the technologically important situation in which material A is the semiconductor gallium arsenide and material B is the alloy aluminium gallium arsenide. However if $b \not\approx a$, then the growth of B is frustrated—there is a competition between the desire of the B atoms to attach themselves to A atoms and their desire to be a distance b from the nearest B atom. If the energetics are such that one set of bond energies is much greater than the other, the system will either form an adsorbed, strained layer, with B–B neighbor distance set by the substrate (a) or it will detach itself entirely. If the energetics are more closely balanced, then the system will find a compromise, often involving the formation of dislocations in the surface layer. There are, of course, many ways of introducing such dislocations to relieve the strain. A simple model for this effect was proposed by Frenkel and Kontorova (*See* **Frenkel–Kontorova model**).

An interesting system in which the frustration can be continuously tuned experimentally is the Josephson junction array (*See* **Josephson junction arrays**). This is a system composed of a lattice of superconducting islands separated by weak links. Each island is characterized by a phase variable and the total energy of the array can be written

$$E = -\varepsilon_{\mathrm{J}} \sum_{\langle i,j \rangle} \cos\left(\theta_i - \theta_j - A_{ij}\right),$$

where the sum is over nearest-neighbor sites, θ_i is the phase on the ith island, ε_J is the Josephson energy, and A_{ij} is a twist variable along a nearest-neighbor bond. The twist variables must satisfy

$$\sum\!\!\!\!\!\!\bigcirc A_{ij} = 2\pi \frac{\Phi}{\phi_0},$$

where the sum is around all the bonds on an elementary plaquette of the lattice, Φ is the magnetic flux through that plaquette, and ϕ_0 is the flux quantum $h/2e$. When the flux per plaquette is an integer number of flux quanta, the system is unfrustrated and its ground state simply has all phases equal. When Φ/ϕ_0 is a rational fraction, the system is uniformly frustrated and its ground states exhibit periodic spatial ordering. If Φ/ϕ_0 is irrational (or, in finite-sized systems if the commensuration length exceeds the system size), then the system exhibits complex, aperiodic ordering. The introduction of frustration introduces a greater range of elementary excitations and a greater degeneracy in the lowest energy states.

Spin glasses are magnetic alloys in which there are competing interactions between magnetic ions that can be of either sign. The canonical model of a spin glass was devised by S.F. Edwards and P.W. Anderson (Edwards & Anderson, 1975) in which each site of a regular lattice bears a classical vector spin. Nearest-neighbor pairs of spins interact via an exchange coupling drawn from a quenched (i.e., nondynamical) random distribution that includes both ferromagnetic and antiferromagnetic possibilities. The energy of a given configuration of spins is then given by

$$E = -\sum_{\langle i,j\rangle} J_{ij}\mathbf{S}_i\cdot\mathbf{S}_j.$$

The combination of frustration and quenched randomness leads to very rich behavior. A mean field version of this model in which all pairs of spins interact weakly was proposed by Sherrington and Kirkpatrick (1975) and solved by Parisi (1979) by introducing the notion of replica symmetry breaking. An alternative approach based on an equation (the so-called TAP equation) for the number of metastable states was presented by Thouless, Anderson, and Palmer (Mezard et al., 1987; Fischer & Hertz, 1991). The relevance of the notions developed for this mean field model to real finite range spin glasses (the EA model) has been questioned and the so-called droplet picture for these was developed (Fischer & Hertz, 1991). The debate over the relative merits of the droplet picture and the broken replica symmetry picture has resurfaced recently and is still highly contentious. Irrespective of which detailed picture is correct, it is clear that the combination of frustration and quenched randomness leads to spin glass systems having many, closely competing low-energy configurations. This leads to a hierarchy of free energy valleys and barriers causing the dynamics of these systems to show very complex and slow dynamical behavior such as aging in which the dynamical response of the system depends on its thermal history. Much current attention is focussed on quantum mechanical frustrated systems both regular (spin liquid states in quantum antiferromagnets) and disordered (quantum spin glasses, etc.).

KEITH BENEDICT

See also **Ferromagnetism and ferroelectricity; Frenkel–Kontorova model; Ising model; Josephson junction arrays; Spin systems**

Further Reading

Binder, K. & Young, A.P. 1996. Spin glasses: experimental facts, theoretical concepts and open questions. *Reviews of Modern Physics*, 58: 801–976

Edwards, S.F. & Anderson, P.W. 1975. Theory of spin glasses. *Journal of Physics* F, 5: 965–974

Fischer, K.H. & Hertz, J.A. 1991. *Spin Glasses*, Cambridge and New York: Cambridge University Press

Mezard, M., Parisi, G. & Virasoro, M.A. 1987. *Spin Glass Theory and Beyond*, Singapore: World Scientific

Parisi, G. 1979. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43: 1754–1756

Sherrington, D. & Kirkpatrick, S. 1975. Solvable model of a spin glass. *Physical Review Letters*, 35: 1792–1796

Young, A.P. 1997. *Spin Glasses and Random Fields*, Singapore: World Scientific

FUNCTION SPACES

There are classes of spaces that, while maintaining some essential features of the n-dimensional Euclidean space ($\mathbb{R}^n$), are still general enough to include spaces of functions as particular examples. This entry covers both these abstract classes and some particular function spaces.

A *metric space* retains only the notion of distance. To be a metric on a space X, $d(x, y)$ should be nonnegative and satisfy

(i) $d(x, x) = 0$ for every $x \in X$;
(ii) $d(x, y) = d(y, x)$ for every $x, y \in X$; and
(iii) the triangle inequality: $d(x, y) \le d(x, z) + d(z, y)$ for every x, y, and $z \in X$.

More precisely, it is the pair (X, d), rather than X alone, that is the metric space. For $x, y \in \mathbb{R}^n$, the standard metric is just given by $d(x, y) = |x - y|$.

A metric leads to a notion of convergence: $x_j \to x$ in X as $j \to \infty$ if for any $\varepsilon > 0$, there is an N such that $d(x_j, x) < \varepsilon$ whenever $j \ge N$. In Euclidean spaces, a sequence converges if and only if it is a Cauchy sequence. In the context of a metric space, $\{x_j\}_{j=1}^{\infty}$ is Cauchy if for any $\varepsilon > 0$, there exists an N such that $d(x_i, x_j) < \varepsilon$ whenever $i, j \ge N$. A metric space in which every Cauchy sequence converges is called *complete*.

The general version of a length is a norm. Consider a real (or complex) vector space X [which satisfies the property: if x and y are elements of X, then so is $x + \lambda y$ for any $\lambda \in \mathbb{R}$ (or $\mathbb{C}$)]. Then a norm on X is a nonnegative function $\|\cdot\|$ such that

(iv) $\|x\| \geq 0$ with equality if and only if $x = 0$;
(v) $\|\lambda x\| = |\lambda|\|x\|$ for every $x \in X$ and $\lambda \in \mathbb{R}$ (or $\mathbb{C}$); and
(vi) $\|x + y\| \leq \|x\| + \|y\|$ for every $x, y \in X$ (the triangle inequality again).

Note that X needs to be a vector space so that if x and y are in X, so are λx (as in (v)) and $x + y$ (as in (vi)). The pair $(X, \|\cdot\|)$ is called a *normed space*. A complete normed space is called a *Banach space*.

Given a norm, the distance between x and y, $d(x, y) = \|x - y\|$, defines a metric; thus, normed spaces are less general than metric spaces.

The standard notion of the length of a vector in $\mathbb{R}^n$ is not just an ad hoc definition satisfying (iv)–(vi). Vectors are specified relative to n orthogonal coordinate axes, and the expression for the length follows from Pythagoras's theorem. Since the mathematical formalization of orthogonality involves the dot product ($\mathbf{x} \cdot \mathbf{y} = \sum_{j=1}^{n} x_j y_j$), it is natural to consider generalizing this concept.

An *inner product space* is a real (or complex) vector space X equipped with an inner product: that is, a function (x, y) that associates a real (or complex) number with any two elements $x, y \in X$ and satisfies

(vii) $(x, x) \geq 0$, with equality if and only if $x = 0$;
(viii) $(x, y) = \overline{(y, x)}$ for all $x, y \in X$; and
(ix) $(\lambda x + \mu y, z) = \lambda(x, z) + \mu(y, z)$ for every x, y, $z \in X$ and every $\lambda, \mu \in \mathbb{R}$ (or $\mathbb{C}$).

A complete inner product space is called a *Hilbert space*.

Due to (vii), it is possible to set $\|x\| = (x, x)^{1/2}$; the Cauchy–Schwarz inequality

$$|(x, y)| \leq \|x\| \, \|y\| , \tag{1}$$

a consequence of (vii)–(ix), can then be used to show that $\|\cdot\|$ is a norm on X. Thus, every inner product space can also be viewed as a normed space, implying that inner product spaces are more restrictive than normed spaces.

There is one abstract setting that includes all the above: the only concept retained is the notion of an open set. Such *topological spaces* (Sutherland, 1975) consist of a space X and a topology $\mathcal{T}$—the collection of all the open sets in X. This collection must satisfy:

(x) $\emptyset$ and X are elements of $\mathcal{T}$;
(xi) if $O_1, O_2 \in \mathcal{T}$, then $O_1 \cap O_2 \in \mathcal{T}$; and
(xii) the union of any collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$.

Any metric space gives rise to a topological space by taking $\mathcal{T}$ to be the collection of all of its open sets, but there are some notions of convergence (e.g.,weak convergence, *See* **Functional analysis**) that do not correspond to any choice of metric. Because of this, topological vector spaces (vector spaces equipped with a topology) form the basis of advanced functional analysis (Rudin, 1991).

We now give some simple examples of function spaces, in which I denotes any interval (finite or infinite) in $\mathbb{R}$.

The space $C^0(I)$ consists of all real-valued continuous functions on I. The standard norm on this space is the supremum or uniform norm $\|\cdot\|_\infty$, defined as

$$\|f\|_\infty = \sup_{x \in I} |f(x)| \tag{2}$$

(essentially the maximum value of f on the interval I, provided that this is attained). Since the uniform limit of continuous functions is itself continuous, equipped with the uniform norm $C^0(I)$ is complete, and so a Banach space. For functions defined on the whole real line, such uniform convergence is often too strong. More realistic is uniform convergence on compact intervals, that is, equivalent to convergence in the metric

$$\begin{aligned} d_K(f, g) &= \sum_{j=1}^{\infty} 2^{-j} \\ &\quad \times \min\left(\sup_{x \in [-j, j]} |f(x) - g(x)|, 1 \right) \end{aligned} \tag{3}$$

which cannot be derived from a norm.

The space $C^k(I)$ consists of all continuous functions whose first k derivatives are also continuous,

$$\begin{aligned} C^k(I) &= \{f \in C^0(I) : \mathrm{d}^j f/\mathrm{d}x^j \in C^0(I), \\ j &= 1, \ldots, k\}. \end{aligned} \tag{4}$$

Its standard norm (which makes C^k a Banach space) is formed by adding the maximum value of f and its first k derivatives,

$$\|f\|_{C^k} = \sum_{j=0}^{k} \|\mathrm{d}^j f/\mathrm{d}x^j\|_\infty. \tag{5}$$

The theory of generalized functions uses the space $\mathcal{D}(\mathbb{R})$ of infinitely differentiable functions with compact support in $\mathbb{R}$ ("test functions"). A sequence $\{\phi_n\}_{n=1}^{\infty} \in \mathcal{D}(\mathbb{R})$ converges to ϕ in $\mathcal{D}$ if there is a fixed compact set K containing the support of each ϕ_n, and $\mathrm{d}^j \phi_n/\mathrm{d}x^j$ converges uniformly to $\mathrm{d}^j \phi/\mathrm{d}x^j$ for every

$j = 0, 1, 2, \cdots$. This form of convergence gives rise to a topology, but there is no corresponding metric.

Another family of Banach spaces are the "Lebesgue spaces" $L^p(I)$, consisting of all Lebesgue integrable functions on I for which the L^p norm defined by

$$\|f\|_{L^p} = \begin{cases} \left(\int_I |f(x)|^p \,\mathrm{d}x\right)^{1/p}, \ 1 \le p < \infty, \\ \inf\{M : \ |f(x)| \le M \text{ almost everywhere in } I\}, \\ \qquad p = \infty \end{cases} \tag{6}$$

is finite. (See Priestley (1997) for a readable introduction to Lebesgue integration.) The space $L^2(I)$ of square integrable functions is a Hilbert space: for two functions $f, g \in L^2(I)$ one can define an inner product by setting

$$(f, g) = \int_I f(x)\overline{g(x)}\,\mathrm{d}x. \tag{7}$$

Note that this is a very natural space of functions to consider physically: if $u(x)$ denotes a velocity then the L^2 norm of $u(x)$ is proportional to the kinetic energy.

The modern theory of partial differential equations relies heavily on *Sobolev spaces* (Adams, 1975; Evans, 1998; Gilbarg & Trudinger, 1983). These allow discussion of the degree of differentiability of functions that are only weakly differentiable: a function f defined on an open interval I has weak derivative $Df = g$ if there is a function $g \in L^1(I)$ such that for every infinitely differentiable function ϕ with compact support in I

$$\int_I g(x)\phi(x)\,\mathrm{d}x = -\int_I f(x)\phi'(x)\,\mathrm{d}x. \tag{8}$$

(The right-hand side would be the result of integrating $\int f'(x)\phi(x)\,\mathrm{d}x$ by parts if f were a C^1 function. Although this is similar to the definition of the derivative of a generalized function, the weak derivative must be an element of L^1.) By requiring successive weak derivatives $D^j f$ to be in $L^p(I)$, we obtain the Sobolev space $W^{k,p}(I)$:

$$W^{k,p}(I) = \{f \in L^p(I) : \ D^j f \in L^p(I), \\ j = 1, \ldots, k\}. \tag{9}$$

Given the norm

$$\|f\|_{W^{k,p}} = \left(\sum_{j=0}^{k} \|D_j f\|_{L^p}^p\right)^{1/p}, \tag{10}$$

these are Banach spaces. The spaces $H^k(I) \equiv W^{k,2}(I)$ are Hilbert spaces when equipped with the inner product

$$(f, g)_{H^k} = \sum_{j=0}^{k} (D^j f, D^j g)_{L^2}. \tag{11}$$

James C. Robinson

See also **Functional analysis; Generalized functions; Topology**

Further Reading

Adams, R.A. 1975. *Sobolev Spaces*, New York: Academic Press

Evans, L.C. 1998. *Partial Differential Equations*, Providence, RI: American Mathematical Society

Gilbarg, D. & Trudinger, N.S. 1983. *Elliptic Partial Differential Equations of Second Order*, Berlin: Springer

Priestley, H.A. 1997. *Introduction to Integration*, Oxford: Clarendon Press and New York: Oxford University Press

Rudin, W. 1991. *Functional Analysis*, New York: McGraw-Hill

Sutherland, W.A. 1975. *Introduction to Metric and Topological Spaces*, Oxford: Clarendon Press

FUNCTIONAL ANALYSIS

Introduction

Systems that are extended in space are usually modeled by partial differential equations (PDEs). The solution of such an equation will be a function of both space and time, for example, $u(x, t)$, and so the state of the system is specified by the *function* $f(\cdot) = u(\cdot, t)$. Since the resolution of such a function as a Fourier series

$$f(x) = \sum_{k \in \mathbb{Z}} c_k \mathrm{e}^{\mathrm{i}kx}, \qquad c_{-k} = \overline{c_k},$$

requires an infinite number of coefficients, the appropriate phase space for a PDE is generally infinite-dimensional. For example, the stability of a solution (be it stationary, periodic, or more general still) will depend on the eigenvalues of some linear operator that acts on this infinite-dimensional phase space.

Broadly speaking, functional analysis can be viewed as the study of infinite-dimensional spaces and the properties of maps (often linear) defined on them. Functional analysis is required for any rigorous treatment of PDEs (e.g., Evans, 1998, or Robinson, 2001) and many problems in the theory of ordinary differential equations.

In this entry we discuss two topics, both central to the subject. First, we give a very brief outline of spectral theory, which generalizes ideas familiar from linear algebra; and then we discuss the notion of "weak convergence" that goes some way to circumventing the problems arising from a fundamental difference between finite- and infinite-dimensional spaces.

For the sake of simplicity, we will discuss these two topics only in the context of a Hilbert space rather than for a general Banach space (*See* **Function spaces**). We denote this Hilbert space by H, its norm by $\|\cdot\|$, and its inner product by $(\cdot,\cdot)$. Note that $\mathbb{R}^n$ is a finite-dimensional Hilbert space. For more details and a more general treatment, see the suggestions for further reading.

Spectral Theory

Initially motivated by Sturm–Liouville boundary value problems and the related theory of integral equations, spectral theory has become an important part of functional analysis. The theory generalizes ideas from finite-dimensional linear algebra to linear operators on infinite-dimensional spaces.

A map $A : H \to H$ is *linear* if

$$A(x + \lambda y) = Ax + \lambda Ay \quad \text{for every } x, y \in H, \lambda \in \mathbb{R},$$

and is *bounded* if, for some $M > 0$,

$$\|Ax\| \leq M\|x\| \qquad \text{for every } x \in H. \qquad (1)$$

[If $H = \mathbb{R}^n$, then any linear map is bounded, but this is not true when H is infinite-dimensional.]

When $H = \mathbb{R}^n$, the eigenvalues of a matrix A are all those complex numbers λ for which $A - \lambda I$ is not invertible (so that $Ax = \lambda x$ for some nonzero x). When A is a linear operator on an infinite-dimensional space, the *spectrum* of A consists of all the values of λ for which $R_\lambda(A) = (A - \lambda I)^{-1}$ lacks one or more of the following "nice" properties:

(i) $R_\lambda(A)$ exists,
(ii) $R_\lambda(A)$ is bounded, and
(iii) $R_\lambda(A)$ can be defined on a dense subset of H (another "nice" property whose exact meaning is unimportant here).

In general, this spectrum can be divided into three distinct pieces:

- the point spectrum $\sigma_p(A)$ (eigenvalues): (i) does not hold, so that there is a nonzero $x \in H$ (the "eigenfunction") with $Ax = \lambda x$;
- the continuous spectrum $\sigma_c(A)$: (i) and (iii) hold, but not (ii), and
- the residual spectrum $\sigma_r(A)$: (i) holds but (iii) does not.

If $H = \mathbb{R}^n$, then whenever (i) holds so do (ii) and (iii), and the spectrum consists only of the eigenvalues of A.

When $H = \mathbb{R}^n$ and A is a real symmetric matrix, results from linear algebra guarantee that (a) all its eigenvalues are real, (b) the eigenvectors corresponding to distinct eigenvalues are orthogonal, and (c) the eigenvectors form a basis for $\mathbb{R}^n$. To obtain a similar result when H is infinite-dimensional we have to impose certain restrictions on A. The original applications to boundary value problems and integral equations motivated the following two definitions that are useful here: an operator $A : H \to H$ is *compact* if the image $\{Ax_n\}_{n=1}^{\infty}$ of any bounded sequence $\{x_n\}_{n=1}^{\infty}$ has a convergent subsequence [if $H = \mathbb{R}^n$, then any linear map is compact, cf. below]; and a bounded map from a Hilbert space H into itself is *self-adjoint* if

$$(Ax, y) = (x, Ay) \quad \text{for every} \quad x, y \in H$$

[when $H = \mathbb{R}^n$, this means that A is a real symmetric matrix].

If A is a compact, self-adjoint operator that is also invertible, then it behaves much like a real symmetric matrix: (a) all of its eigenvalues are real; (a′) the residual spectrum is empty and there are at most a countable number of eigenvalues, which are bounded and can only have zero as an accumulation point; (b) eigenfunctions corresponding to distinct eigenvalues are orthogonal; and (c) the eigenfunctions form a basis for H. This is the celebrated Hilbert–Schmidt theorem, the rigorous result that justifies the approach of Sturm–Liouville theory, that is, using eigenfunctions as a basis in which to expand solutions as a "generalized Fourier series" (see Kreyszig, 1978; Renardy & Rogers, 1992, for example).

Weak Convergence

A space is finite-dimensional if and only if its unit ball is compact. Put another way, any bounded sequence in a finite-dimensional space has a convergent subsequence (in $\mathbb{R}$ this is the Bolzano–Weierstrass theorem): this result is extremely useful, but it is not true in an infinite-dimensional space.

However, it is possible to define a weaker notion of convergence ("weak convergence") and so recover a form of this compactness property in certain infinite-dimensional spaces. To motivate the definition we make two observations. First, if x and y are two distinct elements of a Hilbert space H, then, it is possible to find a $z \in H$ such that

$$(z, x) \neq (z, y),$$

"inner products can distinguish elements of H." Second, it is also possible to show that if $(z, x) = (z, y)$ for every $z \in H$, then we must have $x = y$: "inner products can determine elements of H."

Because of these two results, it is reasonable to define a notion of convergence based on inner products. In a Hilbert space, a sequence x_n converges weakly to x, written $x_n \rightharpoonup x$, if

$$(z, x_n) \to (z, x) \quad \text{for every element} \quad z \in H.$$

Although a bounded sequence in an infinite-dimensional Hilbert space need not have a convergent subsequence, it will have a subsequence that converges weakly. Such convergence is often sufficient for applications; in particular, it is fundamental to many results in the theory of existence and uniqueness for solutions of PDEs.

[The Riesz Representation theorem guarantees that for any bounded linear map $\rho : H \to \mathbb{R}$, there is a $z \in H$ such that $\rho(x) = (z, x)$ for all $x \in H$. In a general Banach space B, the "inner product with" z used above is replaced by "any bounded linear map from B into $\mathbb{R}$". These maps are known as "linear functionals," and gave rise to the name of the subject to which they are central.]

JAMES C. ROBINSON

See also **Function spaces; Generalized functions; Topology**

Further Reading

Evans, L.C. 1998. *Partial Differential Equations*, Providence, RI: American Mathematical Society

Kreyszig, E. 1978. *Introductory Functional Analysis with Applications*, New York: Wiley

Meise, R. & Vogt, D. 1997. *Introduction to Functional Analysis*, Oxford: Clarendon Press and New York: Oxford University Press

Renardy, M. & Rogers, R.C. 1992. *An Introduction to Partial Differential Equations*, New York: Springer

Robinson, J.C. 2001. *Infinite-dimensional Dynamical Systems*, Cambridge and New York: Cambridge University Press

Rudin, W. 1991. *Functional Analysis*, New York: McGraw-Hill

Yosida, K. 1980. *Functional Analysis*, Berlin: Springer

G

GALAXIES

Galaxies are dense agglomerations of matter in the Universe. They consist of gas, dust, and stars as a major fraction. In addition, for kinematical reasons a hypothetical "dark matter" component is required. Their formation dates back to the early Universe, almost 14 billion years ago. Although some galaxies were already listed as nebulous objects in Charles Messier's catalogue of nebulae and star clusters from 1784, their discovery as detached "island universes" only dates from 1924, when Edwin Hubble resolved the outer parts of Messier Object 31 (M31, the Andromeda Nebula) into stars and measured their distances.

Within 2.2 million light years ($1\,\mathrm{ly} = 9.46\times 10^{12}$ km), M31 is the closest massive galaxy and is similar to our Milky Way galaxy, both belonging to a galaxy type that is characterized by a rotating disk of stars and gas showing spiral patterns. In addition, such spiral galaxies possess a central spheroid of old stars called a bulge and an extended spheroidal halo of old single stars and bound star clusters, called globular clusters.

All spiral galaxy labels today begin with the prefix "*S*". For decreasing bulge-to-disk ratios, Hubble classified the sequence of spirals from Sa to Sc, some of which also show an innermost bar structure where the spiral arms emerge at the ends. Hubble distinguished these as SBs as distinct from normal Ss.

Another morphological galaxy type exists with an elliptical shape, almost no particular substructure, relatively old stars, and depleted in gas: ellipticals. According to their minor-to-major axes ratio b/a, Hubble denoted them by E10$(1 - b/a)$ reaching from E0 (circular) to E7 ($b/a = 0.3$) at most.

Lenticular galaxies (SOs) form the link between spirals and ellipticals in Hubble's famous "tuning-fork diagram" (see Figure 1). Hubble galaxies have masses between several 10^{10} to 10^{12} solar masses ($m_\mathrm{s} = 1.9891 \times 10^{30}$ kg) and diameters of 100,000 light years and more. More refined classification schemes are possible such as those by de Vaucouleurs (1959) and Sandage (1961). Today an extension exists to Sds, and even further to irregularly shaped galaxies (Is) with lower mass and brightness and no central bulge.

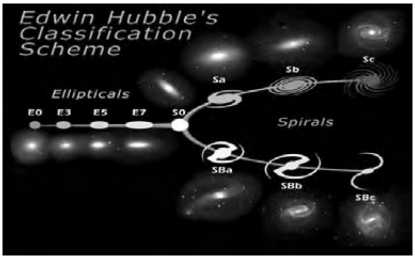

Figure 1. Hubble's "tuning-fork diagram" of galaxies. With courtesy from STScI.

Internal and Dynamical Structure of Galaxies

Ellipticals

According to the stellar mass distribution, any galaxy becomes dimmer with distance R from the geometrical center. The radial dependence however, differs between morphological types and their components. Ellipticals follow in all directions the empirical de Vaucouleurs law with brightness falling off with $\exp(-R^{1/4})$. Although it seems plausible that rotation causes ellipticity, since the 1970s, it has been known that in most ellipticals their regular rotation velocity is smaller than the irregular proper motion of stars with speeds of 200–250 km/s. They are therefore denoted as "hot stellar systems." The elliptical shape is thus formed by the anisotropic velocity dispersion of stars. Moreover, triaxiallity of the figure axes, as for a flattened cigar, can be derived for some elliptical galaxies if the elliptical contours of equal brightness are somehow twisted. Most elliptical galaxies can be divided additionally into boxy- or disky-shaped contours considering the fourth-order cosine in a trigonometric expansion.

Spirals

Spiral galaxies are the most complex systems. While their bulges' brightness profiles resemble ellipticals, they are rotationally supported. Although the halos are not dense enough to show a continuous stellar distribution, one can take, for example, globular clusters as representative and find a power law for the density of $R^{-2} \ldots R^{-3.5}$. While in the halo, the large irregular velocities of stars (determined in our Milky Way) and globular clusters, their age, and the lack of significant amounts of gas are similar to those of ellipticals (without knowing the halo's ellipticity), disks consist of 10–20% gas and of stars, both rotating with velocities of 200–250 km/s. Because the velocity dispersion of stars is much less, in the range of only 10–60 km/s, disks are therefore "cold systems" and rotationally flattened. Their radial face-on brightness drops with $\exp(-R/\alpha)$ where α is the so-called scale length of around 10,000 light years.

Although spiral arms are exceptionally trailing, they are not the result of structures wound by differential rotation because their pitch angles, that is, the angle between an arm and a circle, should be much smaller than observed. Because the arm structure is pronounced in the visual by the brightness of young stars that form out of cold gas condensed within the arms, among different possible processes of arm formation, density waves are the most favored.

Since the characteristic velocities v, for example, the rotation of spiral disks and velocity dispersion in elliptical galaxies, are almost constant with R, although the visible mass decreases, this invokes the existence of dark matter. Because in equilibrium the centrifugal force $F_c = mv^2/R$ acting on the mass element m is balanced by the gravitational force of a mass M_R included within R, $F_g \propto mM_R/R^2$, for observed v=const. M_R has to increase with R. This means that the dark matter contribution dominates in the outermost regime. In contrast, less bright ellipticals have been found recently where a decline of velocities at large radii is measured, so the dark matter content remains controversial.

Formation and Evolution of Galaxies

Galaxies assemble into larger units by means of gravitation, thus forming galaxy groups like our Local Group and galaxy clusters with numerous members, as for example, the Virgo Cluster at a distance of 50 million light years. In the center of the clusters, ellipticals dominate while spirals permeate the whole cluster. This leads to the suggestion that ellipticals are formed by merging events in the densest cluster regions and in the early universe.

While the Hubble-type galaxies are massive, dwarf galaxies with less mass exist. In our vicinity the large and small magellanic clouds (LMC, SMC) as satellites of the Milky Way represent a dwarf irregular type, consisting of gas and stars as spirals but with less regular structures. With masses of $5 \times 10^9 m_s$, the LMC lies at the upper-mass range of dwarf galaxies, while other Milky Way galaxy satellites, dwarf spheroidals, form the low-mass end of the dwarf ellipticals. With 10^6–$10^7 m_s$ their brightness is so low that observations are still incomplete. Dwarf ellipticals are the most frequent galaxy type in the present Universe resembling Hubble's Es.

The class of dwarf galaxies is of substantial importance for our understanding of galaxy formation and the evolution of the Universe, because they serve as the building blocks of massive galaxies in the cosmological picture of hierarchical clustering. Although their accretion by mature galaxies by means of tidal friction is observable, their destruction rate during the course of the universe is yet unclear because it is also possible that the dwarf galaxy types are replenished by other processes such as tidal tails of merging galaxies.

Gerhard Hensler

See also **Cosmological models; Gravitational waves; Hénon–Heiles system; Spiral waves**

Further Reading

Binney, J. & Merrifield, M. 1998. *Galactic Astronomy*, Princeton Series in Astrophysics, Princeton, NJ: Princeton University Press

Binney, J. & Tremaine, S. 1987. *Galactic Dynamics*, Princeton Series in Astrophysics, Princeton, NJ: Princeton University Press

Combes, F., Mazure, A. & Boisse, F. 2001. *Galaxies and Cosmology*, Berlin and Heidelberg: Springer

Sandage, A. 1961. *Hubble Atlas of Galaxies*, Washington, DC: Carnegie Institution of Washington

Sparke, L.S. & Gallagher, J.S. 2000. *Galaxies in the Universe*, Cambridge and New York: Cambridge University Press

de Vaucouleurs, G. 1959. Classification and morphology of external galaxies, vol. 53, *Handbuch der Physik*, Berlin: Springer, p. 275

GAME OF LIFE

The rules underlying Life are simple, according to computer scientists. Biologists are inclined to be skeptical, but they do agree that the cellular automaton known as the Game of Life provides fascinating insights into the phenomena of self-organization and emergence in systems of interacting agents.

Biological life has been around for at least 3.8 billion years, but the Game of Life was invented by mathematician John Conway in 1970 and publicized by Martin Gardner in his "Mathematical Games" column in *Scientific American*. It is probably the best-known example of the class of algorithms known as cellular automata (CA). A CA is a one- or two-dimensional array of cells, each of which can exist in a number of states. Time in a CA is discrete; at each time step, every cell updates itself on the basis of its current state and those of its neighbors. Cellular automata can exhibit surprisingly complex global temporal

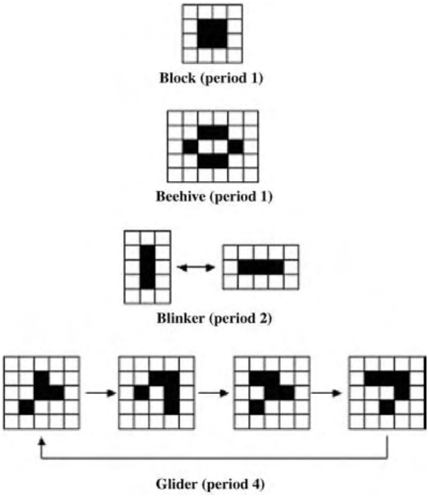

Figure 1. Some commonly seen patterns in the Game of Life.

dynamics, arising from extremely simple rules applied on a local scale.

The standard Game of Life uses a two-dimensional grid. Cells can be either on (alive) or off (dead). The neighborhood of a cell is the eight cells surrounding it. The rules of Life are, indeed, simple; if a live cell has two or three live neighbors, it stays alive. A dead cell with exactly three live neighbors comes alive (is born). In all other cases the cell dies, either of overcrowding (with more than three live neighbors) or loneliness (with fewer than two). At each time step, all cells update their states simultaneously. That's all there is to Life! No wonder the biologists are dubious. But from this simple foundation, complex, consistent patterns of activity emerge. If the grid is seeded at random with live cells, the first few time steps are a turmoil of apparently random activity. However, identifiable patterns of live cells quickly emerge. Interesting patterns exhibit periodic behavior, cycling between a number of different states in a deterministic manner. Many patterns settle into a limit cycle of length one—a stable point attractor, or "still life" in the Life terminology. Others, known to Life practitioners as "oscillators," have longer limit cycles.

So consistent are these patterns that hundreds have been identified, named, and studied by Life enthusiasts all over the world (see e.g., Figure 1). There are several catalogues of Life patterns available on the Web; a particularly nice site is http://hensel.lifepatterns.net/.

Although most initial configurations eventually settle to a stable state or cyclic set of states, this is not always the case. Life aficionados have identified initial states that generate new states indefinitely.

So is Life more than just a generator of interesting patterns? The answer, of course, is "yes." Cellular automata in general, and Life in particular, have interesting theoretical properties. Stephen Wolfram identified four broad classes of dynamic behavior common to one-dimensional (Wolfram, 1984) and two-dimensional (Wolfram, 1985) CAs.

Class 1: Evolution leads to a homogeneous state (analogous to a point attractor in a nonlinear dynamic system),

Class 2: Evolution leads to a set of separated simple stable or periodic structures (analogous to limit cycles),

Class 3: Evolution leads to a chaotic pattern (analogous to the chaotic attractors found in continual dynamic systems),

Class 4: Evolution leads to complex localized structures, sometimes long-lived.

Although the universal applicability and hence usefulness of Wolfram's classification system has been questioned (e.g., Eppstein, 2000), it is still widely accepted. The Game of Life falls, serendipitously, into class 4, the class for which Wolfram hypothesizes that "class-4 cellular automata are generically capable of universal computation, so that they can implement arbitrary information-processing procedures" (Wolfram, 1985). It has, in fact, been proven that Life is a universal cellular automaton—one which can emulate a Turing machine, capable of performing universal

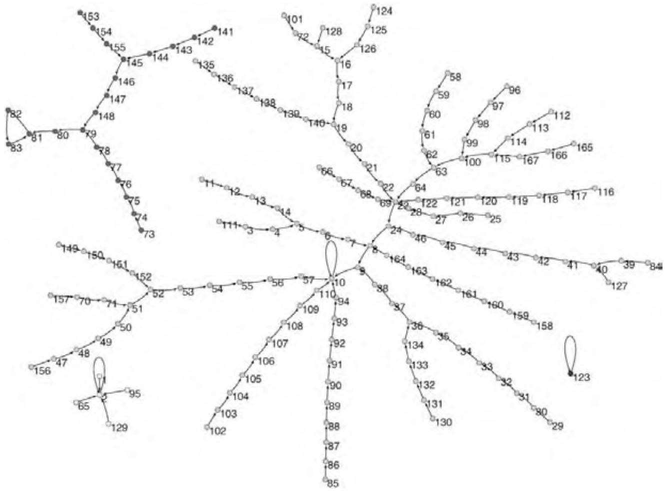
153
154
155
145
144
143
142
141
146
147
148
82
81
80
79
83
78
77
76
75
74
73
101
72
15
128
124
125
126
16
135
136
137
138
139
140
17
18
19
20
21
58
59
60
61
62
63
96
97
98
99
100
112
113
114
115
167
166
165
66
67
68
69
22
64
122
121
120
119
118
117
116
23
28
27
26
25
11
12
13
14
5
111
3
4
6
7
24
46
45
44
43
42
41
40
39
84
127
8
164
163
162
161
160
159
158
149
150
151
152
52
53
54
55
56
57
10
157
70
71
51
110
109
94
9
38
37
50
49
48
47
156
108
107
106
105
104
103
102
93
92
91
90
89
88
87
86
85
36
35
34
33
32
31
30
29
134
133
132
131
130
123
1
2
95
65
129

Figure 2. State transition graph for a cellular automaton. States are arbitrarily numbered. The state space contains four basins of attraction, three of which contain a point attractor, while the fourth leads to a limit cycle of length three (states 81, 82 and 83).

computation. The universality of Life was proved in the early 1980s (Berlekamp et al., 1982), and a Turing machine that can be extended to a universal Turing machine was implemented by Paul Rendell in 2000 (Adamatzky, 2002).

An overview of the temporal dynamics of a two-dimensional CA such as the Game of Life can be provided by a state transition graph, in which each possible pattern of on and off cells in the CA makes up a unique state. If the number of cells in the lattice is s and the number of cell states is k, then there are k^s possible states for that CA. Each state forms a node in the graph. Since the rules are deterministic, each state will map to a single new state in the next timestep, and the transition between the two states is represented by a directed link (an arc) between the two nodes. A state transition graph depicts the course of evolution of the CA from any given starting point (Figure 2).

Inspection of a state transition diagram leads to the conclusion that evolution in a CA is not deterministically reversible; some states have two or more antecedent states. Some states, in contrast, have no preceding states. The latter are known as Garden of Eden states. The existence of Garden of Eden states in the Game of Life was predicted by CA theory, but actually identifying such a state is a nontrivial task. Several Garden of Eden states have been identified by trial and error, but the search for a reliable algorithm for their identification continues.

The rules of Life are simple, but from them arise complex, emergent behaviors—coherent forms arising at different levels of organization and interacting to produce new forms and patterns. This complexity arising from underlying simplicity makes the Game of Life an ideal toy world for the study of many of the fascinating phenomena of complex systems.

JENNIFER HALLINAN

See also **Artificial life; Attractors; Biological evolution; Cellular automata; Order from chaos; State diagrams**

Further Reading

Adamatzky, A. (editor). 2002. *Collision-based Computing*, Heidelberg: Springer

Berlekamp, E.R., Conway, J.H. & Guy, R.K. 1982. *Winning Ways for Your Mathematical Plays*, vol. II. *Games in Particular*, London: Academic Press

Eppstein, D. 2000. Searching for spaceships. In *More Games of No Chance*, edited by R.J. Nowakowski, Cambridge and New York: Cambridge University Press, pp. 433–453

Gardner, M. 1970. Mathematical games: the fantastic combinations of John Conway's new solitaire game "life." *Scientific American*, 223: 120–123

Wolfram, S. 1984. Universality and complexity in cellular automata. *Physica* D, 10: 1–35

Wolfram, S. 1985. Two-dimensional cellular automata. *Journal of Statistical Physics*, 38: 901–946

The Game of Life can be played online at:
http://www.math.com/students/wonders/life/life.html
http://www.bitstorm.org/gameoflife/
http://hensel.lifepatterns.net/

GAME THEORY

Humans play games. From the formalized warfare of chess to the Machiavellian machinations of politics to the subtleties of sexual pursuit, interactions between individuals with desires and priorities that are often conflicting and contradictory lies at the heart of human society. The formal study of games, however, is a relatively recent phenomenon and has moved rapidly from its beginnings as a mathematical tool to aid gamblers to its current status as an essential paradigm in fields as diverse as economics and evolutionary biology.

Today, *game theory* can be defined as "the analysis of rational behavior under circumstances of strategic interdependence, when an individual's best strategy depends upon what his opponents are likely to do" (Varoufakis, 2001). Individuals may be people, corporations, nations, animals, species, or any other entity that can be said to exhibit strategic behavior.

Formal game theory began in 1713, when Pierre-Rémond de Montmort first proposed the concept of a minimax solution to a card game called Le Her. A minimax solution to a two-player game is one in which an individual chooses his strategy so as to minimize the maximum loss or risk that he/she will incur. It was James Waldegrave, the originator of Le Her, who actually produced a minimax solution to the game.

Other concepts of fundamental importance to modern game theory also emerged in the 18th century, a result of work by Daniel Bernoulli in his 1738 analysis of the St. Petersburg paradox (Dimand & Dimand, 1992). These concepts included utility (a measure of the desirability to the player of each possible outcome of the game), the maximization of expected utility, diminishing marginal utility (the decrease in the amount of benefit derived from consuming each additional unit of a product or service), and risk aversion as a parameter of a utility function.

The work on minimax game solutions remained an isolated curiosity until the 1920s, when Émile Borel published a series of short papers on strategic games in the Proceedings of the French Academy of Sciences between 1921 and 1927. In these papers, he defined the normal form of a game: a matrix representation of the game in which each player tries to work out the best strategy independent of the sequence of moves. Borel later claimed to have proven the minimax theorem, but this does not appear to be the case. In fact, he may not even have stated the theorem.

The first formal proof of the minimax theorem for two-person games with any finite number of pure strategies was given by John von Neumann in a paper presented to the Gottingen Mathematical Society on December 7, 1926 (Dimand & Dimand, 1992). He proved that in a zero-sum game (one in which one player's gain is the other's loss), there exists a unique set of mixed strategies, one per player, which equalizes the payoffs that each player can gain regardless of the strategy adopted by the other player.

At about the same time, the Princeton economist Oskar Morgenstern was pondering mixed strategy game theoretic issues, as exemplified by that master of bluff, Sherlock Holmes, and the strategies he should adopt to avoid his arch enemy, Professor Moriarty. In 1944, von Neumann and Morgenstern collaborated to produce *The Theory of Games and Economic Behavior*, the seminal publication in this area. In 1947, they revised the book to include expected utility theory, under which games are expressed in terms of the players' perceptions of the inherent desirability and likelihood of their outcomes, and players never expect other players to hold mistaken beliefs—the assumption of complete rationality, which has been fundamental to much ensuing work. Economists were initially reluctant to accept game theory, but since the publication of von Neumann and Morgenstern's book, the theory has undergone extensive development, and has been applied to an enormous variety of problems in economics, to the point where Leonard (1992) could assert that "game theory plays a central role in economic theory."

Minimax theory assumes that each player has perfect knowledge about the game and that the game is zero-sum. In such a case, the best strategy for each player is independent of the strategy adopted by the other player. In most real-world situations, this is not the case; more often, the best strategy for one player depends on what the other players choose to do. The extension of minimax theory to *n*-player, noncooperative games was achieved by the Princeton mathematician John Nash in a paper published in 1950. In this paper, he defined the Nash equilibrium. A Nash equilibrium is a set of strategies such that no player could improve his/her payoff, given the strategies of all other players in the game, by changing his/her strategy. Nash proved that all noncooperative games have a Nash equilibrium and, thereby, established an analytical structure within which all situations of conflict and cooperation could be studied. For this work, he received the 1994 Nobel Prize for Economics, together with John Harsanyi and Reinhard Selten.

Game theory was also applied with considerable success to other fields of research, perhaps most notably to evolutionary biology. The concepts of game theory transfer readily to evolutionary biology—the values of different outcomes, which in economic theory are measured as utility, are readily interpreted in terms of Darwinian fitness. Moreover, the somewhat sweeping assumption of complete rationality in the behavior of the agents is replaced by the concept of evolutionary stability.

Game-theoretic concepts were first explicitly applied to the study of evolution by Lewontin (1961), who saw the agents in the game as a species on the one hand, against nature on the other. The utility of this approach was quickly recognized, and the focus shifted to modeling interactions between individuals. In this context, the concept of the Nash equilibrium was rediscovered independently by John Maynard Smith and G.R. Price in 1973 as the Evolutionarily Stable Strategy (ESS—"a strategy such that, if all the members of a population adopt it, no other strategy can invade" (Maynard Smith, 1982, p. 204). An ESS represents the solution to a game.

In the last three decades, game theory has been used to provide a framework for the analysis of a wide range of biological phenomena, including the evolution of sex ratios, parental investment in offspring, patterns of animal dispersal, competition for resources (see, e.g., Maynard Smith, 1982) and the evolution of cooperation (Axelrod, 1984). In biology, as in economics, game theory has become an essential tool for the theorist, providing a structured framework for the analysis of a wide range of phenomena.

JENNIFER HALLINAN

See also **Artificial life; Biological evolution; Economic system dynamics**

Further Reading

Axelrod, R. 1984. *The Evolution of Cooperation*, New York: Basic Books

Dimand, R.W. & Dimand, M. 1992. The early history of the theory of strategic games from Waldegrave to Borel. In *Toward a History of Game Theory*, edited by E.R. Weintraub, Durham: Duke University Press

Leonard, R.J. 1992. Creating a context for game theory. In *Toward a History of Game Theory*, edited by E.R. Weintraub, Durham: Duke University Press

Lewontin, R.C. 1961. Evolution and the theory of games. *Journal of Theoretical Biology*, 1: 382–403

Maynard Smith, J. 1982. *Evolution and the Theory of Games*, Cambridge: Cambridge University Press

Nash, J. 1950. Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences*, 36: 48–49

Varoufakis, Y. 2001. General introduction: game theory's quest for a single, unifying framework for the social sciences. In *Game Theory: Critical Concepts in the Social Sciences*, edited by Y. Varoufakis, vol. 1. London and New York: Routledge

von Neumann, J. & Morganstern, O. 1953. *Theory of Games and Economic Behavior*, 3rd edition, New York: Wiley (1st edition, 1944)

GAP SOLITONS

See **Solitons, types of**

GARDNER (GGKM) EQUATION

See **Inverse scattering method**

GAUSSIAN BEAM

See **Nonlinear optics**

GAUSSIAN ENSEMBLES

See **Free probability theory**

GEL'FAND–LEVITAN THEORY

In inverse problems, sometimes also called backward problems, one is given the solution and required to find the underlying equation. Inverse spectral problems deal with the recovery of unknown coefficients of a differential operator from the knowledge of its spectral data. Their importance stems from the fact that coefficients in a differential equation usually model the physical structure or composition of a certain material. Nondestructive testing, geophysical prospecting, radar, and medical imaging are just some of the important applications of inverse problems. By analyzing reflected waves, radar can locate, track, and sometimes identify a target. In geophysical prospecting, the travel time of reflected underground acoustic waves can reveal deposits of petroleum and natural gas. The field of medical imaging has witnessed the development of many non-invasive and safe procedures using the above ideas. For example, ultrasonography using the three-dimensional Doppler effect can help visualize the fetal vascular system for prenatal diagnosis. High-resolution ultrasound cardiovascular imaging not only reconstructs a real-time picture of the heart but also measures the rate of blood flow, thus giving early warnings for clogged arteries, heart attacks, and strokes.

A simple rule of the thumb is that data used should be equivalent to the information to be reconstructed. For example, if we are looking for a square integrable function, the data should be equivalent to its sequence of Fourier coefficients, while an analytic function is equivalent to its Taylor coefficients. Can we recover a matrix of size $n \times n$ from its n eigenvalues? Although in general to define a matrix we would need n^2 entries, possible exceptions would be matrices that are fully determined by their first rows—symmetric Toeplitz or Hankel matrices, for example. In all kinds of inverse problems, one faces two major issues: uniqueness of the solution and the algorithm for its reconstruction. One of the best understood inverse problems, which deals with the Schrödinger operator, was solved by Gel'fand and Levitan in 1951 (see Gel'fand & Levitan, 1955).

The One-dimensional Schrödinger Equation

Consider the one-dimensional Schrödinger operator

$$\begin{cases} L(y) = -y''(x,\lambda) + q(x)y(x,\lambda) = \lambda y(x,\lambda), \\ \quad x \in [0,\ \infty), \\ y'(0,\lambda) - hy(0,\lambda) = 0, \end{cases} \tag{1}$$

where q is a real continuous function and h is a real constant. The continuity of q ensures the existence of a unique solution of the initial value problem at $x = 0$ for every $\lambda \in \mathbb{C}$. The boundary value problem (1) is regular at $x = 0$ and singular at $x = \infty$. The operator L, which acts in the Hilbert space of square integrable functions on the positive half line, is self-adjoint so its eigenvalues are real. We usually normalize solutions of (1) by the condition $y(0, \lambda) = 1$, which leads to $y'(0, \lambda) = h$. To determine the spectrum of L (the set of values λ such that the inverse operator of $L - \lambda I_d$ either does not exist or is unbounded), we need to examine the behavior of the solutions $y(x, \lambda)$ as $x \to \infty$. If the solution decays fast enough and is square integrable, that is, has a finite energy $\alpha_n = \int_0^\infty |y(x, \lambda_n)|^2 \, \mathrm{d}x < \infty$, then λ_n is an eigenvalue and belongs to the discrete part of the spectrum. For other solutions that could be approximated by continuous functions or do not grow faster than a polynomial in x, λ belongs to the continuous spectrum; otherwise λ is not in the spectrum. A more precise result using the theory of distributions (Gel'fand & Shilov, 1967) shows that λ is in the spectrum if and only if the solution $y(x, \lambda)$ is the derivative of a function that cannot grow faster than $x^{3/2+\varepsilon}$ as $x \to \infty$. This characterization threw a new light on the mysterious behavior of the continuous spectrum.

For example, when $q = h = 0$, the spectrum of L is continuous on $[0, \infty]$. Indeed its "eigenfunctions" are $\cos\left(x\sqrt{\lambda}\right)$ which are not square integrable and so there are no eigenvalues, while for $\lambda > 0$ the solutions $\cos\left(x\sqrt{\lambda}\right)$ are bounded and so grow less than $x^{3/2+\varepsilon}$ as $x \to \infty$, which means that $\lambda > 0$ is in the continuous spectrum. On the other hand, when $\lambda < 0$, the solutions $\cos\left(x\sqrt{\lambda}\right) = \cosh\left(x\sqrt{-\lambda}\right)$ have an exponential growth, and so $\lambda < 0$ cannot be not in the spectrum.

Once the operator L is known to be self-adjoint, the solution $y(x, \lambda)$ forms the kernel of a transform, which is similar to the Fourier cosine transform,

$$F(\lambda) = \int_0^\infty f(x) y(x, \lambda) \, \mathrm{d}x,$$

and its inverse transform is defined explicitly by

$$f(x) = \int_\sigma F(\lambda) \overline{y(x, \lambda)} \, \mathrm{d}\rho(\lambda).$$

The function ρ, which is called the spectral function, is nondecreasing, is right-continuous, has jumps at the eigenvalues $\rho(\lambda_n) - \rho(\lambda_n - 0) = 1/\alpha_n$, and is continuous and increasing only on the continuous part of the spectrum. For example, when $q = h = 0$, the spectral function is simply

$$\rho_0(\lambda) = \begin{cases} (2/\pi)\sqrt{\lambda} & \text{if } \lambda \geq 0, \\ 0 & \text{if } \lambda < 0. \end{cases}$$

Obviously when $q \neq 0$, the new spectral function ρ would record all new spectral changes, which leads to the inverse spectral problem: by comparing ρ to ρ_0 can we recover the perturbation q?

Gel'fand–Levitan Theory

If we are given a spectral function ρ, can we find its associated Sturm–Liouville differential operator given in (1)? This is in essence the Gel'fand–Levitan theory, which produced an elegant algorithm for the recovery of the potential q and the boundary condition at $x = 0$.

As explained earlier, the spectrum depends on the growth of solutions and a key point is their representations by the Fourier transform,

$$y(x, \lambda) = \cos\left(x\sqrt{\lambda}\right) + \int_0^x K(x, t) \cos\left(t\sqrt{\lambda}\right) \mathrm{d}t. \quad (2)$$

The kernel $K(x, t)$ is continuous and contains all information about q, namely, that

$$\frac{\mathrm{d}}{\mathrm{d}x} K(x, x) = \frac{1}{2} q(x) \text{ and } K(0, 0) = h. \quad (3)$$

Thus, it is a matter of finding $K(x, t)$ in order to recover q. To do so, we first form the function

$$F(x, t) = \int_{-\infty}^{\infty} \frac{\sin\left(x\sqrt{\lambda}\right)}{\sqrt{\lambda}} \frac{\sin\left(t\sqrt{\lambda}\right)}{\sqrt{\lambda}} \, \mathrm{d}\sigma\left(\lambda\right), \quad (4)$$

where the given spectral function ρ is used in

$$\sigma(\lambda) = \begin{cases} \rho(\lambda) - \frac{2}{\pi}\sqrt{\lambda} & \text{if } \lambda \geq 0, \\ \rho(\lambda) & \text{if } \lambda < 0 \end{cases} \quad (5)$$

and set

$$f(x, t) = \frac{\partial^2 F(x, t)}{\partial x \partial t}.$$

The crux of the theory is that $f(x, t)$ and $K(x, t)$ satisfy a linear integral equation, for each fixed x

$$K(x, t) + \int_0^x K(x, s) f(s, t) \, \mathrm{d}s = -f(x, t) \quad \text{for } 0 \leq t \leq x. \quad (6)$$

Thus given the spectral function ρ, we can form $F(x, t)$ by (4), yielding the kernel $f(x, t)$ and then for each fixed x solve the Fredholm integral equation (6) for $K(x, t)$. As for matrices, a Fredholm integral equation has a solution if the null space contains only the trivial solution. Once the existence of a solution K is guaranteed, its smoothness is examined. It is shown that $f(x, t)$ and $K(x, t)$ have the same degree of smoothness which implies the smoothness of q by (3). The original

Gel'fand–Levitan result in (1951) is based on the linear equation (6) and can be summarized as follows:

Theorem 1 (Gel'fand and Levitan Theory). *If a nondecreasing function $\rho(\lambda)$ satisfies*

(A) *for arbitrary real x the integral*

$$\int_{-\infty}^{0} \exp\left(x\sqrt{|\lambda|}\right) \mathrm{d}\rho(\lambda) \textit{ exists,}$$

(B) *the integral*

$$a(x) = \int_{1}^{\infty} \frac{\cos\left(x\sqrt{\lambda}\right)}{\lambda} \mathrm{d}\sigma(\lambda)$$

exists for all $0 \leq x < \infty$, while $a(x)$ has continuous derivatives up to the fourth order for all $0 \leq x < \infty$ and if the set of points of increase of ρ has at least one finite accumulation point, then there exists just one differential operator of the second order defined by (1) *which has $\rho(\lambda)$ as its spectral function. The function $q(x)$ and the number h are defined by* (3), *where $K(x, t)$ is a solution of* (6).

However, the original theory had a gap between the necessary and sufficient conditions. Ten years later, Levitan and Gasymov proved a stricter version, which we state as the Gel'fand–Levitan–Gasymov theory found in Gasymov & Levitan (1964) that contains necessary and sufficient conditions.

Theorem 2 (Gel'fand–Levitan–Gasymov). *The monotonically increasing function ρ is the spectral function of a Sturm–Liouville of type* (1) *with a function q having m integrable derivatives and a number h if and only if the following conditions are satisfied:*

(A) *If $E(\lambda) = \int_0^\infty f(x)\cos\left(x\sqrt{\lambda}\right) \mathrm{d}x$, where $f \in L^2(0, \infty)$ and of compact support, then*

$$\int |E(\lambda)|^2 \mathrm{d}\rho(\lambda) = 0 \Longrightarrow$$
$$f = 0 \textit{ almost everywhere.}$$

(B) *The limit $\mathbf{\Phi}(x) = \lim_{N\to\infty} \int_{-\infty}^{N} \cos\left(x\sqrt{\lambda}\right)\mathrm{d}\sigma(\lambda)$, where $\sigma(\lambda)$ is defined by* (5), *exists boundedly in every finite range of values of x and $\mathbf{\Phi}$ has $m+1$ locally integrable derivatives with $\mathbf{\Phi}(0) = -h$.*

There are also interesting results for the regular case that is defined by

$$\begin{cases} L(y) = -y''(x,\lambda) + q(x)y(x,\lambda) = \lambda y(x,\lambda), \\ \quad 0 \leq x \leq \pi, \\ y'(0,\lambda) - hy(0,\lambda) = 0, \\ y'(\pi,\lambda) - Hy(\pi,\lambda) = 0, \end{cases} \tag{7}$$

where q is continuous and $q, h, H \in \mathbb{R}$. Let the norming constants be defined by $\alpha_n = \int_0^\pi |y(x, \lambda_n)|^2 \mathrm{d}x$. Note that since q is continuous, its knowledge is equivalent to its sequence of Fourier coefficients. Thus, the data ought to be at least an infinite sequence of numbers in order to recover the potential q.

Theorem 3. *If all the $\alpha_n > 0$,*

$$\sqrt{\lambda_n} = n + \frac{a_0}{n} + \frac{a_1}{n^3} + O\left(\frac{1}{n^4}\right)$$

and

$$\alpha_n = \frac{\pi}{2} + \frac{b_0}{n^2} + O\left(\frac{1}{n^3}\right),$$

where a_0, a_1, and b_0 are constants, then there exists an absolutely continuous function $q(x)$ corresponding to the given λ_n and α_n.

Theorem 4. *The numbers $\{\lambda_n\}_{n=0}^{\infty}$ and $\{\alpha_n\}_{n=0}^{\infty}$ are the spectral data of some boundary value problem (7) with q'' being square integrable if and only if the following asymptotic estimates hold:*

$$\sqrt{\lambda_n} = n + \frac{a_0}{n} + \frac{a_1}{n^3} + \frac{\gamma_n}{n^3}$$

and

$$\alpha_n = \frac{\pi}{2} + \frac{b_0}{n^2} + \frac{\tau_n}{n^3},$$

where $\lambda_n \neq \lambda_m$ and $\alpha_n > 0$ and the series $\sum_{n=1}^{\infty} \gamma_n^2$ and $\sum_{n=1}^{\infty} \alpha_n^2$ are convergent.

The problem of generalizing the theory to higher dimensional operators remains open and depends heavily on the idea of transformation operators that map generalized functions. In Boumenir (1991), it is also proved that the Gel'fand–Levitan theory is based on the factorization of operators whose symbol is given by the spectral functions.

Discrete Case

Consider now a discrete version of the Sturm–Liouville problem

$$Bv = \lambda Av, \tag{8}$$

where B is a Jacobi matrix and A is a diagonal matrix with positive entries (positive definite). Expressing (8) in a vectorial form, we end up with recurrence relation defined by

$$\begin{cases} c_n y_{n+1} = (a_n\lambda + b_n)\, y_n - c_{n-1} y_{n-1}, \\ \quad n = 0, \ldots, m-1, \\ y_{-1} = 0, \quad \text{and} \quad y_m + h y_{m-1} = 0, \end{cases} \tag{9}$$

where $b_n \in \mathbb{R}$, $a_n > 0$, and $c_n > 0$. We look for a nontrivial finite sequence $y_{-1}, \ldots, y_m$ which satisfies (9). To this end, we normalize the solution vectors by $y_0 = 1/c_{-1}$, then knowing the first terms y_{-1} and

y_0 allows us to compute recursively the remaining entries that are now functions of λ. It is easy to see that eigenvalues λ_r are nothing other than the zeros of the last condition $y_m(\lambda) + h y_{m-1}(\lambda) = 0$. For a given eigenvector, define its squared norm by $\alpha_r = \sum_{k=0}^{m-1} a_k |y_k(\lambda_r)|^2$. The following theorem is found in Atkinson (1964) or in Teschl (2000).

Theorem 5. *Assume that we are given* $h \in \mathbb{R}$, $\{a_k > 0\}$, *eigenvalues* $\{\lambda_r\}_{0 \le r \le m-1}$, *norming constants* $\{\rho_r\}_{0 \le r \le m-1}$, *then there exists* $\{c_k\}_{-1 \le k \le m-1}$ *which are positive and constants* $\{b_k\}_{0 \le k \le m-1}$ *such that the boundary value problem has the set* $\{\lambda_r\}_{0 \le r \le m-1}$ *as its eigenvalues.*

AMIN BOUMENIR

See also **Inverse problems; Inverse scattering method or transform; Quantum inverse scattering method**

Further Reading

Atkinson, F.V. 1964. *Discrete and Continuous Boundary Problems*, New York and London: Academic Press

Boumenir, A. 1991. A comparison theorem for self-adjoint operators. *Proceedings of the American Mathematical Society*, 111(1): 161–175

Gasymov, M.G. & Levitan, B.M. 1964. Determination of a differential equation by two of its spectra. *Russian Mathematical Surveys*, 19(2): 1–63

Gel'fand, I.M. & Levitan, B.M. 1955. On the determination of a differential equation from its spectral function. *American Mathematical Society Translations*, (2)1: 253–304

Gel'fand, I.M. & Shilov, G.E. 1967. *Generalized Functions*, vol. 3: *Theory of Differential Equations*, New York and London: Academic Press

Levitan, B.M. 1987. *Inverse Sturm–Liouville Problems*, Utrecht: VNU Science Press

Marchenko, V.A. 1986. *Sturm–Liouville Operators and Applications*, Basel: Birkhäuser

Teschl, G. 2000. *Jacobi Operators and Completely Integrable Nonlinear Lattices*, Providence, RI: American Mathematical Society

GENERAL CIRCULATION MODELS OF THE ATMOSPHERE

Atmospheric general circulation models (AGCMs) simulate the dynamical, physical, and chemical processes of planetary atmospheres. For the Earth's atmosphere (*See* **Atmospheric and ocean sciences**), they are based on the thermo-hydrodynamic equations, which consist of the conservation of momentum, mass, and energy with the ideal gas law in coordinates suitable for the rotating planet.

In the presently used form, they were first derived by Vilhelm Bjerknes in 1904; subsequently, Lewis Fry Richardson (1922) proposed numerical weather prediction (NWP) as a practical application which, in 1950, was successfully performed by Jule Charney, R. Fjørtoft, and John von Neumann on an electronic computer based on a simplified set of these equations. While numerical weather prediction models utilize the atmospheric short-term memory for forecasting, AGCMs developed since the 1960s extend applications to longer time scales simulating seasonal and climate variability (for a personal recollection, see Smagorinsky, 1983).

Since then, numerical weather prediction and atmospheric general circulation modeling have enjoyed continuous advances, which are attributed to the following gains and improvements: (1) observational data accuracy, analysis, and assimilation; (2) insight into dynamical and physical processes, numerical algorithms, and computer power; and (3) the use of model hierarchies to study individual atmospheric phenomena. With simulations of the Earth system and that of other planets envisaged, a broad field of science has been established, which is of vital importance socio-economically, agriculturally, politically, and strategically.

Observations

Since the foundation of the World Meteorological Organization (WMO) and international treaties to monitor and record meteorological and oceanographical data, the collection of global data through local weather and oceanic stations has been systematically organized and has become a truly globalized system through the deployment of satellites and introduction of remote sensing facilities. The availability of extensive global data has enabled the extension of NWP models to complex global GCMs, thus facilitating simulations on larger time scales (months and years instead of hours and days) and validation of NWPs in return. Furthermore, it has become possible to study climate history and make estimates of future climates, which is particularly important due to likely anthropogenic impacts.

General Circulation Models

Atmospheric general circulation models have two basic components. First, the dynamical core consists of the primitive equations (the conservation equations with vertical momentum equation approximated by hydrostatic equilibrium, that is, balancing the vertical pressure gradient and the apparent gravitational forces) under adiabatic conditions. Second, physical processes contribute to the diabatic sources and sinks interacting with the dynamical core. They are incorporated as parameterizations, mostly in a modular format: solar and terrestrial radiation, the hydrological cycle (with phase transitions manifested in evaporation and transpiration, cloud, and precipitation processes), the planetary boundary layer communicating between the free atmosphere and the ground (soil with vegetation, snow and ice cover; ocean with sea ice), and atmospheric chemistry. Most of these parameterizations enter the thermodynamic energy equation as heat sources or sinks.

Dynamical core

State-of-the-art atmospheric general circulation models commonly utilize the so-called primitive equation approximation of the Navier–Stokes equations. In addition to the dry dynamics, equations describing the transport of other constituents such as water vapor, cloud liquid water and ice, trace gases, and particles (aerosols) can be an integral part of the dynamical core. To integrate the equations, they are discretized in space and time where finite differences and spectral methods are the most dominant. For more details on the governing equations; *See* the entries for **Atmospheric and ocean sciences; Fluid dynamics; Navier–Stokes equation**.

(i) Horizontal discretization: In the horizontal, grid point, or spectral representations of the dependent model variables are used. Different grid structures and finite difference schemes have been designed to reduce the error (Messinger & Arakawa, 1976). An alternative approach is the spectral method. The dependent variables are represented in terms of orthogonal functions where appropriate basis functions are the spherical harmonics. The maximum wave number of the expansion defines the resolution of the model. Since the computation of products is expensive, only linear terms are evaluated in the spectral domain. To compute the nonlinear contributions, the variables are transformed into grid point space every time step, where the respective products are computed and transformed back to spectral space. Necessary derivatives are computed during the transformation. This spectral-transform procedure (Eliassen et al.,1970; Orszag, 1970) makes the spectral approach computationally competitive with finite difference schemes. For low resolutions, the spectral method is, in general, more accurate than the grid point method. However, spectral methods are less suitable for the treatment of scalar fields, which exhibit sharp gradients and, for physical reasons, must maintain a positive-definite value (e.g., water vapor, cloud water, chemical tracers). Therefore, selected fields are often treated separately in the grid point domain using, for example, semi-Lagrangian techniques. Recently, with increasing model resolutions and the need for transporting more species (e.g., for chemical submodels), grid point models are attracting more attention again, while novel grid structures are introduced, for instance, the spherical icosahedral grid of the German Weather Service model GME (Majewski et al., 2000).

(ii) Vertical discretization: In general, finite differences and numerical integration techniques are used for the derivatives and integrals in the vertical. The vertical coordinate can be defined in different ways. The isobaric coordinate eliminates the density from the equations and simplifies the continuity equation compared with a z-coordinate system. However, the intersection of low-level pressure surfaces with the orography enforces time-dependent lower boundary conditions which are difficult to treat numerically. This problem can be avoided if terrain following sigma (σ) coordinates are used, where sigma is defined by the pressure divided by the surface pressure $\sigma = p/p_0$. Unfortunately, the sigma coordinate leads to a formulation of the pressure gradient force which, in the presence of steep orography, is difficult to treat. The advantages of both sigma and pressure coordinates are combined by introducing a hybrid coordinate system with a smoothed transition from σ to p with height.

Physical Processes and Parameterizations

Many processes that are important for large-scale atmospheric flow cannot be explicitly resolved by the model due to its given spatial and temporal resolution. These processes need to be parameterized; that is, their effect on the large-scale circulation needs to be formulated in terms of the resolved grid-scale variables. The most prominent processes in building a parameterization package of an atmospheric general circulation model are long- and short-wave radiation, cumulus convection, large-scale condensation, cloud formation and the vertical transport due to turbulent fluxes in the planetary boundary layer, and the effect of different surface characteristics such as vegetation on the surface fluxes. Additional processes such as the excitation of gravity waves and their impact on the atmospheric momentum budget or the effect of vertical eddy fluxes above the boundary layer are often considered. Because the land surface provides a time-dependent boundary condition that acts on time scales comparable to the atmosphere, great effort has been made to include land surface and soil processes in the atmospheric parameterization package. More recently, the effect of the interaction of various chemical species and their reactions with atmospheric circulation are also being considered. In addition to the direct relation between the resolved atmospheric flow and the effect of the parameterized processes, there are various other interactions among the individual processes that have to be taken into account. For typical comprehensive atmospheric general circulation models, Figure 1 displays the interrelations between the adiabatic dynamics, providing the spatial and temporal distribution of the dependent model variables and the various processes being parameterized.

Model Hierarchy

General circulation models of reduced complexity are continuously developed to supplement comprehensive

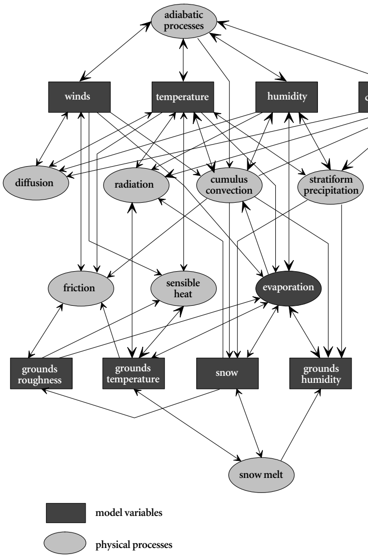

Figure 1. Interactions in comprehensive GCMs (schematic).

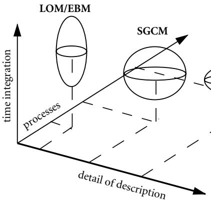

Figure 2. A model hierarchy of general circulation models.

GCMs, to gain insight into atmospheric phenomena (see Figure 2) and for educational purposes: when utilizing the full set of equations, the model spectrum ranges from simple GCMs (SGCMs) with analytic forms of heating and friction to low-order models (LOMs). A prominent LOM is the Lorenz model (*See* **Lorenz equations**), which approximately describes the nonlinear convection dynamics in the vicinity of a critical point for the stream function and temperature in a set of ordinary differential equations. It can be regarded as including first-order nonlinear effects (temperature advection) to a linear model, which leads to chaotic behavior. The Lorenz model is used to study predictability and serves as a paradigm for phase-space behavior of atmospheric GCMs. Utilizing thermal energy conservation only, another spectrum of models (energy balance models, or EBMs) is obtained by averaging in certain spatial directions. This leads to the horizontally averaged one-dimensional radiative-convective models; the one-dimensional energy balance model when averaged vertically and longitudinally for studying climate feedback and stability; and two-dimensional statistical–dynamical models when averaged longitudinally where dynamical processes are being parameterized. A prominent EBM example is the globally averaged or zero-dimensional energy balance model. With ice-albedo and water vapor-emissivity feedbacks included, climate catastrophes leading to a snowball earth and runaway greenhouse can be demonstrated. With random forcing and periodic solar radiation input (e.g., Milankovich cycles), stochastic resonance emerges.

Klaus Fraedrich, Andreas A Aigner,
Edilbert Kirk, and Frank Lunkeit

See also **Atmospheric and ocean sciences; Fluid dynamics; Forecasting; Lorenz equations; Navier–Stokes equation**

Further Reading

Eliassen, E., Machenhauer, B. & Rasmussen, E. 1970. *On a Numerical Method for Integration of the Hydrodynamical Equations with a Spectral Representation of the Horizontal Fields.* Report No. 2, Institut for Teoretisk Meteorologi, University of Copenhagen

Majewski, D., Liermann, D., Prohl, P., Ritter, B., Buchhold, M., Hanisch, T., Paul, G., Wergen, W. & Baumgardner, J. 2000. *The global icosahedral–hexagonal grid point model GME-operational version and high resolution tests*, ECMWF, Workshop Proceedings, Numerical methods for high resolution global models, pp. 47–91

McGuffie, K. & Henderson-Sellers, A. 1997. *A Climate Modelling Primer*, New York: Wiley

Messinger, R. & Arakawa, A. (1976). *Numerical Methods Used in Atmospheric Models*, Geneva: WMO

Orszag, S.A. 1970. Transform method for calculation of vector coupled sums: application to the spectral form of the vorticity equation. *Journal of Atmospheric Sciences*, 27: 890–895

Richardson, L.F. 1922. *Weather Prediction by Numerical Process*, Cambridge: Cambridge University Press

Smagorinsky, J. 1983. The beginnings of numerical weather prediction and general circulation modeling: early recollections. In *Theory of Climate*, edited by B. Saltzman, New York: Academic Press, pp. 3–38

Trenberth, K. 1992. *Climate System Modelling*, Cambridge and New York: Cambridge University Press

GENERAL RELATIVITY

Called *general relativity*, Albert Einstein's theory of gravitation was created as a generalization of his special relativity theory. As special relativity is a theory of physical space-time (neglecting gravitational effects), general relativity is a theory of physical space-time in the presence of gravitation.

While Maxwell's theory of electromagnetism is a relativistic theory that is covariant with respect to Lorentz transformations, Newton's theory of gravitation is incompatible with special relativity. In 1907, two years after proposing special relativity, Einstein was preparing a review of special relativity when he suddenly wondered how Newtonian gravitation would have to be modified to fit in with special relativity. Einstein described this as "the happiest thought of my life." He proposed the *equivalence principle* as "the complete physical equivalence of a gravitational field and the corresponding acceleration of the reference frame" that "extends the principle of relativity to the case of uniformly accelerated motion of the reference frame."

In fact, Einstein's equivalence principle is a generalization of the so-called *weak* equivalence principle, which dates from Galileo and Newton and states that the inertial mass and gravitational mass of any object are equal. Thus (neglecting friction), the acceleration of different bodies in a gravitational field is independent of their masses and other physical characteristics, and hence, with given initial conditions, their motions will be the same.

The next important step was made by Einstein in his 1912 papers, where he concluded that "if all accelerated systems are equivalent, then Euclidean geometry cannot hold in all of them." Further investigations by Einstein to find the correct form of equations for a gravitational field were connected with applications of Riemannian geometry and tensor analysis. The final form of equations of general relativity was given by Einstein in his paper "The Field Equations of gravitation," submitted on 25 November 1915. At about the same time, David Hilbert submitted a paper entitled "Foundations of Physics," which also contains the correct field equations for gravitation, introduced by applying a variation principle.

Physical Space-time and Gravitating Matter in General Relativity

According to general relativity, physical space-time in a gravitational field is non-Euclidean; so to describe the properties of space-time, Einstein applied Riemannian geometry. Without gravitation, physical space-time is a flat pseudo-Euclidean Minkowski 4-continuum, where free particles move uniformly and linearly along geodesic worldlines of Minkowski space-time. Einstein's key idea is that in a gravitational field, particles move along geodesic lines of curved space-time, and in accordance with the Equivalence principle, their movement does not depend on the particles' characteristics. Thus, motion for an observer is motion along curves in 3-space with variable velocity. Curvature of space-time in general relativity is created by sources of gravitational field. In general relativity, the role of sources of gravitational field is played by the energy-momentum tensor describing the distribution and motion of gravitating matter. Energy (or mass) density (the source of gravitation in Newton's theory) is only a component of the energy-momentum tensor. Besides energy (mass), a gravitational field in general relativity is also created by momentum and other components of an energy-momentum tensor. The dependence between geometrical properties of physical space-time and gravitating matter is described by Einstein's gravitational equations.

Einstein Gravitational Equations

The principal geometrical characteristics of a gravitational field in general relativity are given by the metric tensor g_{ik}, which determines the square of distance between two infinitesimally close points of pseudo-Riemannian 4-space-time

$$\mathrm{d}s^2 = g_{ik}\,\mathrm{d}x^i\,\mathrm{d}x^k \quad (i, k = 0, 1, 2, 3). \qquad (1)$$

By means of the metric tensor, time intervals and spatial distances can be defined; thus the formula for proper time fixed by a clock at rest in some reference frame is $\mathrm{d}\tau = (1/c)\sqrt{g_{00}\,\mathrm{d}x^0\,\mathrm{d}x^0}$. Because the value of g_{00} in a gravitational field depends on location, time flow depends on gravitational field.

Einstein's gravitational equations are nonlinear second-order differential equations with respect to the metric tensor and have the form

$$R_{ik} - 1/2 g_{ik} R = 8\pi G/c^4 T_{ik}, \qquad (2)$$

where R_{ik} is the so-called Ricci tensor (a contraction of the curvature tensor), R is the scalar curvature, T_{ik} is the energy-momentum tensor, G is Newton's gravitational constant, and c is the velocity of light in a vacuum. Einstein's equations are covariant with respect to arbitrary coordinate transformations.

Equation (2) can be changed by adding to the right-hand side, the so-called cosmological term Λg_{ik}, where Λ is a cosmological constant introduced by Einstein. This term describes energy density and pressure of the vacuum, and it plays an essential role in cosmology, leading to the effect of gravitational repulsion if $\Lambda > 0$. In the case of weak gravitational fields, when the variation of metric with respect to the Minkowski metric is small, the Einstein equations lead to Newton's law of gravitational attraction and allow one to find first relativistic corrections. In the case of strong gravitational fields, Einstein equations can give new physical results, including black holes and gravitational waves.

Experimental Verification and Bounds of Applicability

Several classical effects of general relativity have been verified observationally, including the bending of light in a gravitational field, gravitational redshift, the advance of the perihelion of the planet Mercury (43 s of arc per century), and retarding of the propagation of light in a gravitational field. The first three effects were discussed by Einstein even before the creation of general relativity.

The *weak* equivalence principle has been verified to high precision (10^{-12}), and general relativity provides a basis for relativistic astrophysics and cosmology. The Hot Big Bang model was built within the framework of general relativity. Over the past two decades, the role of general relativity has grown in connection with discoveries in cosmology—in particular, acceleration of cosmological expansion, dark matter and dark energy, and other problems that need to be resolved.

General relativity is a classical theory, and a consistent quantum theory of gravitation has not yet been developed. In fact, the formulation of a quantum-gravitation theory requires a unified theory of all fundamental physical interactions. At present, the most popular candidate for such a unified theory is the superstring theory. This theory is in higher-dimensional space, and it leads to a generalization of Einstein's gravitation theory. A second problem with general relativity is the presence of gravitational singularities (cosmological singularities, collapsing systems, etc.). According to theorems by Stephen Hawking and Roger Penrose, this problem is connected under certain conditions with internal properties of the gravitational equations of general relativity. As with classical theory, general relativity is inapplicable near singular states.

The creation and development of Einstein's gravitation theory was a triumph of 20th-century science, providing the basis of gravitation theory, relativistic astrophysics, and cosmology. Within the frame of its applicability, general relativity will remain a great achievement of human culture.

VIACHASLAV KUVSHINOV AND ALBERT MINKEVICH

See also **Black holes; Cosmological models; Einstein equations; Gravitational waves; String theory; Tensors; Twistor theory**

Further Reading

Einstein, A. 1989–1996. *The Collected Papers of Albert Einstein*, vol. 2: The Swiss Years: Writings, 1900–1909, vol. 3: The Swiss Years: Writings, 1909–1911; vol. 4: The Swiss years: Writings, 1912–1914, Princeton, NJ: Princeton University Press

Fok, V.A. 1961. *The Theory of Space, Time and Gravitation*, New York: Pergamon Press

Landau, L.D. & Lifshitz, E.M. 1973. *Therie of field*, Moscow: Nauka

Landau, L.D. & Lifshitz, E.M. 1984. *The Classical Theory of Fields*, 4th edition, Oxford and New York: Pergamon Press

Misner, C., Thorne, K. & Wheeler, J. 1973. *Gravitation*, San Francisco: Freeman

Weinberg, S. 1972. *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*, New York: Wiley

Zel'dovich, Ya.B. & Novikov, I.D. 1971. *Relativistic Astrophysics*, 2 vols. Chicago: University of Chicago Press

GENERALIZED FUNCTIONS

Generalized functions were introduced into quantum mechanics by Paul Dirac, who defined the delta function $\delta(x)$ as follows (Dirac, 1958):

$$\delta(x) = 0 \text{ for } x \neq 0$$

and

$$\int_{-\varepsilon}^{+\varepsilon} \delta(x) = 1 \text{ for } \varepsilon > 0. \tag{1}$$

Although $\delta(x)$ is not a true function (from a mathematical perspective) because it is undefined at $x = 0$, the delta function is widely used in physics to approximate functions that are localized in space or time. Examples of such idealizations include the concepts of point mass or point charge and the spatially localized action of a pick on a guitar string.

In an engineering context, the delta function is called the unit impulse function and is used to approximate functions of large amplitude and short duration such as the striking of a golf ball or the instantaneous charging of an electrical capacitor (Guillemin, 1953). Since such point sources are only idealizations, they can be represented by the limiting procedure, $\delta(x) = \lim_{a\to\infty} \delta_a(x)$ for a family of piecewise continuous functions $\delta_a(x)$, where a is a continuous parameter. Among others, the following four functions, $\delta_a(x)$, are popular approximations of the delta function $\delta(x)$ (Pauli, 1973):

- *Unit impulse* (*box*): $\delta_a(x) = a/2$ for $|x| < 1/a$, and $\delta_a(x) = 0$ for $|x| > 1/a$.
- *Finite impulse response filter*: $\delta_a(x) = \sin ax/\pi x$.
- *Gaussian pulse*: $\delta_a(x) = a\mathrm{e}^{-\pi a^2 x^2}$.
- *Lorentzian pulse*: $\delta_a(x) = a/(1 + \pi^2 a^2 x^2)$.

In the limit $a \to \infty$, all representations $\delta_a(x)$ have zero width, infinite peak amplitude, and unit area, that is, all the approximations converge to the delta function $\delta(x)$ (see Figure 1).

Closely related to Dirac's delta function is Oliver Heaviside's step function, which was introduced in the late 19th century and has been widely used in electronics and communications research since the 1920s (Heaviside, 1950; Guillemin, 1953). The Heaviside step $H(x)$ is defined as follows:

$$H(x) = 1 \text{ for } x > 0 \text{ and } H(x) = 0 \text{ for } x < 0. \tag{2}$$

The derivative of the Heaviside step function is recognized as Dirac's delta function because $H'(x) = 0$ for $x \neq 0$ and for $\varepsilon > 0$

$$\int_{-\varepsilon}^{+\varepsilon} H'(x)\,\mathrm{d}x = H(+\varepsilon) - H(-\varepsilon) = 1. \tag{3}$$

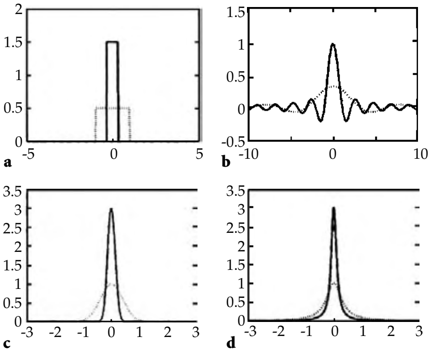

Figure 1. Approximations of the Dirac delta function $\delta(x)$ for $a=1$ (dotted curves) and $a=3$ (solid curves): (a) the box, (b) the impulse response filter, (c) the Gaussian pulse, and (d) the Lorentzian pulse.

In engineering terms, the step function represents an instantaneous jump from zero to unit value of some physical quantity such as the signal voltage at the input terminals of an amplifier.

Because the delta and step are not defined at $x=0$, they are not true functions; thus mathematicians call them distributions, implying rules that assign numbers to integral expressions (Strauss, 1992). To understand this perspective, consider a real function, $f(x)$ which has derivatives for all values of x that approach zero faster than any power of x. In other words, $\lim_{|x|\to\infty} |x|^p f^{(m)}(x) = 0$ for any $m \geq 0$ and $p \geq 0$. Such functions $f(x)$ are called test functions for integral distributions. As is seen from the box approximation of $\delta_a(x)$ in the limit $a \to \infty$, the delta function $\delta(x)$ assigns the number $f(0)$ to the integral distribution associated with a test function $f(x)$:

$$\int_{-\infty}^{\infty} f(x)\delta(x)\,\mathrm{d}x = f(0). \tag{4}$$

More generally, the delta function $\delta(x)$ and its derivatives $\delta^{(m)}(x)$ satisfy the following fundamental property:

$$\begin{aligned} &\int_{-\infty}^{\infty} f(x)\delta^{(m)}(x-\xi)\,\mathrm{d}x \\ &= (-1)^m \int_{-\infty}^{\infty} f^{(m)}(x)\delta(x-\xi)\,\mathrm{d}x \\ &= (-1)^m f^{(m)}(\xi), \end{aligned} \tag{5}$$

where $f(x)$ is a test function. A brief list of useful delta function properties follows from the fundamental property in Equation (5) (Gel'fand & Shilov, 1964):

- *Even function*: $\delta(-x) = \delta(x)$.
- *Scaling transformation*: $|\xi|\delta(\xi x) = \delta(x)$.
- *Factorization*: $2|\xi|\delta(x^2-\xi^2) = \delta(x-\xi) + \delta(x+\xi)$.
- *Projection formula*: $\frac{1}{(x-\xi\mp \mathrm{i}0)} = \mathrm{PV}\left[\frac{1}{(x-\xi)}\right] \pm \pi\mathrm{i}\delta(x-\xi)$ (where PV stands for principal value).
- *Fourier transform*: $\hat{\delta}(k) = \int_{-\infty}^{\infty}\delta(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x = 1$.
- *Spectral representation*: $2\pi\delta(x) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}x$.

Dirac studied orthogonality relations for the wave function $\psi_\lambda(x)$ of the stationary Schrödinger equation with a potential

$$-\frac{h}{2m}\psi''_\lambda(x) + U(x)\psi_\lambda(x) = \lambda\psi_\lambda(x). \tag{6}$$

In this context, square integrable wave functions for different levels of energy λ are orthogonal with respect to the inner product

$$\langle\psi_{\lambda'}|\psi_\lambda\rangle = \int_{-\infty}^{\infty} \bar{\psi}_{\lambda'}(x)\psi_\lambda(x)\,\mathrm{d}x = \delta_{\lambda',\lambda}, \tag{7}$$

where $\delta_{\lambda',\lambda} = 0$ for $\lambda' \neq \lambda$ and $\delta_{\lambda,\lambda} = 1$. Similar orthogonality relations for wave functions of continuous spectra diverge. For example, linear waves in free space (when $U(x) = 0$) take the form $\psi_\lambda(x) = \Psi(x;k) = \mathrm{e}^{\mathrm{i}kx}$ for $\lambda = hk^2/2m$. They are periodic in x with period $L = 2\pi/k$. If $k = k_n = 2\pi n/L$, the periodic function satisfies the following orthogonality relations on the finite interval $x \in [-L/2, L/2]$:

$$\langle\Psi(k_{n'})|\Psi(k_n)\rangle = \int_{-L/2}^{L/2} \mathrm{e}^{\mathrm{i}(k_n-k_{n'})x}\,\mathrm{d}x = L\delta_{n',n}. \tag{8}$$

When linear waves in free space are taken over the whole real axis (i.e., $L\to\infty$), the inner product of the wave function $\Psi(x;k)$ with itself diverges. The delta function $\delta(k)$ replaces this divergence in the closed form

$$\begin{aligned} \langle\Psi(k')|\Psi(k)\rangle &= \int_{-\infty}^{\infty} \bar{\Psi}(x;k')\Psi(x;k)\,\mathrm{d}x \\ &= 2\pi\delta(k'-k), \end{aligned} \tag{9}$$

where $\delta(k) = 0$ for $k \neq 0$ and $\delta(0) = \infty$. The singularity of the delta function $\delta(k)$ is uniquely specified in the distribution (integral) sense by requiring that

$$\int_{-\infty}^{\infty} \delta(k)\,\mathrm{d}k = 1. \tag{10}$$

The delta function $\delta(x)$ represents not only the orthogonality of the wave functions $\psi_\lambda(x)$ of the stationary Schrödinger equation, but also their completeness with the completeness relation:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\bar{\Psi}(x';k)\Psi(x;k)\,\mathrm{d}k+\sum_n\bar{\Psi}_n(x')\Psi_n(x)$$
$$=\delta(x'-x), \qquad (11)$$

where $\Psi(x;k)$ are wave functions of the continuous spectrum and $\Psi_n(x)$ are wave functions of the discrete spectrum. With the use of the delta function $\delta(x)$, a test function $f(x)$ can be expanded into a complete and orthogonal set of wave functions:

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\hat{f}(k)\Psi(x;k)\,\mathrm{d}k+\sum_n\hat{f}_n\Psi_n(x), \quad (12)$$

where coefficients of the expansion are

$$\hat{f}(k)=\langle\Psi(k)|f\rangle, \quad \hat{f}_n=\langle\Psi_n|f\rangle. \qquad (13)$$

Generalized functions are widely used in electric circuit theory, communications, spectral analysis, integral transforms, and Green function solutions of equations of mathematical physics (Guillemin, 1953; Pauli, 1973; Strauss, 1992). For example, the Fourier transform is based on the spectral decomposition above for $\Psi(x;k)=\mathrm{e}^{\mathrm{i}kx}$ and $\Psi_n(x)=0$.

DMITRY PELINOVSKY

See also **Function spaces; Functional analysis; Integral transforms; Quantum theory; Spectral analysis**

Further Reading

Dirac, P.A.M. 1958. *The Principles of Quantum Mechanics*, 4th edition, Oxford: Clarendon Press

Gel'fand, I.M. & Shilov, G.E. 1964. *Generalized Functions,* vol. 1, *Properties and Operations*, New York and London: Academic Press

Guillemin, E.A. 1953. *Introductory Circuit Theory*, New York: Wiley

Heaviside, O. 1950. *Electromagnetic Theory*, New York: Dover

Pauli, W. 1973. *Pauli Lectures on Physics:* Vol. 5, *Wave Mechanics*, Cambridge, MIT Press

Strauss, W.A. 1992. *Partial Differential Equations, An Introduction*, New York: Wiley

GEOMETRICAL OPTICS, NONLINEAR

Nonlinear geometric optics arose in the second half of the 20th century in several areas of research, including nonlinear optics and reaction-diffusion systems. To understand the origin of such problems and the underlying physical phenomena, it is helpful to consider briefly the historical path that has led to them.

Classical Results

Traditionally, geometric optics comprised all of optics. It developed from empirical laws of propagation for light beams, which are bent in a non-uniform medium and refract and reflect at the interfaces of media possessing different optical properties. Remarkable successes were achieved, including inventions of the microscope and the telescope around the beginning of the 17th century. At the same time, some important scientific problems were studied; thus in 1620, Willebrord Snell established the law of refraction at the interface of two transparent media. Assuming angles are measured with respect to a line normal to the boundary, Snell's law states that at any angle of incidence (ψ_1), the ratio

$$\frac{\sin\psi_1}{\sin\psi_2}=\text{constant}\equiv n_{12}, \qquad (1)$$

where ψ_2 is the angle of refraction and the constant n_{12} is the reciprocal refractive index. Interestingly, this important law was independently discovered by René Descartes about a decade later. Pierre Fermat's subsequent formulation of the principle of least time—which governs the propagation of light rays—completed the phenomenological theory of geometric optics.

In the 17th century, two competitive hypotheses were advanced for the physical nature of light: the corpuscular hypothesis of Isaac Newton and the wave theory of Christian Huygens. Two centuries later, after the theoretical studies and experimental works of Thomas Young, Augustin-Jean Fresnel, and François Arago, the wave theory of light triumphed. The true nature of light was revealed in the 19th century to be based on James Clerk Maxwell's theory of electromagnetism, which predicted the existence of electromagnetic waves that propagate in vacuum with constant velocity $c=1/\sqrt{\varepsilon_0\mu_0}\approx 3\times 10^8$ m/s, where ε_0 and μ_0 are the dielectric permittivity and magnetic permeability of vacuum, respectively.

This value—obtained entirely from independent measurements of electric and magnetic fields—is precisely equal to the light velocity in vacuum, which was first measured by Ole Roemer two centuries before (using astronomical data on time intervals between eclipses of Jupiter's satellites). Together with the subsequent experimental discovery and investigation of properties of electromagnetic waves inspired by Hermann Helmholtz and performed by Heinrich Hertz, this result has convinced physicists of the electromagnetic nature of light.

This electromagnetic theory of light explains most optical phenomena. For example, it yields Snell's law of refraction, giving the constant in Equation (1) as

$$n_{12}=v_1/v_2=\sqrt{\varepsilon_2/\varepsilon_1}=n_2/n_1. \qquad (2)$$

Here v_i is the velocity of light in the ith medium ($i = 1, 2$), and ε_i and $n_i = c/v_i = \sqrt{\varepsilon_i/\varepsilon_0}$ are the dielectric permittivity and the absolute refractive index of this medium, respectively.

Physical mechanisms underlying the refractive index were studied by Helmholtz, Paul Drude, and Hendrik Lorentz at the close of the 19th century. They merged the electromagnetic theory with the idea of electrons as charged particles bound in atoms and molecules which are dislocated under the influence of an electromagnetic field of a light wave. (From the modern point of view, such electrons occupy exterior atomic shells; thus, they are referred to as optical electrons.) Further developments of such ideas connect optical phenomena with the dynamic response of an optical medium to the electromagnetic waves propagating through it.

When a light wave propagates through a dielectric medium, its electric field ($\boldsymbol{E}$) displaces the optical electrons, inducing a wave of electric polarization ($\boldsymbol{P}$). The latter generates a secondary electromagnetic wave which is added to the primary wave, modifying the polarization wave, and so on, ad infinitum. In other words, the electromagnetic wave and the response of an optical medium determine each other. As the relations

$$\begin{aligned} \varepsilon \boldsymbol{E} &= \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}, \quad \boldsymbol{P} = \varepsilon_0 \chi \boldsymbol{E}, \\ n &= c/v = \sqrt{\varepsilon/\varepsilon_0} \end{aligned} \tag{3}$$

are always fulfilled (where χ is the dielectric susceptibility of the medium), the absolute refraction index of the medium can be expressed via χ by the formula

$$n = \sqrt{\varepsilon/\varepsilon_0} = \sqrt{1+\chi}. \tag{4}$$

If the light wave is sinusoidal and sufficiently weak, the response of the medium is readily calculated; thus the susceptibility χ is expressed via the oscillatory parameters of optical electrons which, in turn, depend on the frequency (ω), but not on the wave amplitude. Weakness of a light wave means that intensity $\boldsymbol{E}$ of an electric field in the wave must be much smaller than characteristic intensities of intra-atomic electric fields. The greatest intra-atomic intensities (reached in the hydrogen atom) are about 5×10^9 V/cm, whereas the intensities of light fields generated by ordinary (nonlaser) sources of light are about 1 V/cm. Thus, the constitutive equation relating the electric polarization $\boldsymbol{P}$ with the electric field intensity $\boldsymbol{E}$ is linear for ordinary light.

Nonlinear Optics

In the latter half of the 20th century, lasers were invented, which could generate fields with intensities of about 10^7 V/cm, and in 1989, beams of light were produced with electric fields of more than ten times the intra-atomic values. In such strong fields, dielectric susceptibility χ becomes dependent on the electric field intensity, making the constitutive equation nonlinear. Both the dielectric permittivity ε and the absolute refraction index n of the medium also depend on the electric field intensity $\boldsymbol{E}$, and these dependencies are modified by additional influences, such as heating of the medium.

Studies of optical effects in strong light fields use the basic ideas of geometric optics, for example, the concept of refractive index. Consider a uniform isotropic dielectric medium. Adverting to Equation (3), let us write the expansion

$$\boldsymbol{P} = \varepsilon_0 \chi^{(1)} \boldsymbol{E} + \varepsilon_0 \chi^{(2)} |\boldsymbol{E}|^2 + \varepsilon_0 \chi^{(3)} |\boldsymbol{E}|^2 \boldsymbol{E} + \cdots. \tag{5}$$

Here, even powers of $\boldsymbol{E}$ are excluded by symmetry consideration, so $\chi = \boldsymbol{P}/\varepsilon_0 \boldsymbol{E} = \chi^{(1)} + \chi^{(3)} |\boldsymbol{E}|^2 + \cdots$, which implies (see Equation (4))

$$\begin{aligned} n &= \sqrt{1 + \chi^{(1)} + \chi^{(3)} \boldsymbol{E}|^2} \approx n_0 + 2n_3 |\boldsymbol{E}|^2, \\ n_0 &= \sqrt{1 + \chi^{(1)}}, \quad n_3 = \frac{\chi^{(3)}}{4\sqrt{1 + \chi^{(1)}}}, \end{aligned} \tag{6}$$

because the nonlinear term is small. Averaging n on the time period $T = 2\pi/\omega$ of oscillations of electric field $\boldsymbol{E} = \boldsymbol{E}_0(x, y, z) \sin \omega t$ yields

$$n = n_0 + n_3 |\boldsymbol{E}_0|^2, \tag{7}$$

where n_0 is the linear refractive index and $n_3 |\boldsymbol{E}_0|^2$ describes the nonlinear correction to this index.

Equations (3) and (7) indicate that a bounded cylindrical beam of light can create an optical heterogeneity in the medium through which it propagates. Suppose the beam is Gaussian so the amplitude of electric oscillations varies transversely as $|\boldsymbol{E}_0|(r) = A_0 \exp[r^2/(2r_0^2)]$, where A_0 is the value of amplitude on the axis of beam, r is distance from the beam axis, and r_0 is a phenomenological constant determining the typical width of the beam. Then,

$$\begin{aligned} v &= c/(n_0 + n_3 A_0^2) \text{ on the axis of the beam;} \\ v &= c/n_0 \text{ far from the axis of the beam.} \end{aligned}$$

If the inequality $n_3 > 0$ ($n_3 < 0$) is fulfilled, the axial velocity of light is smaller (greater) than the peripheral velocity. In first case, the plane wave front will become concave in the direction of light propagation, and self-focusing of the beam will occur; while in the second case, self-defocusing of the beam will occur. Both these phenomena have been observed; in particular, self-focusing was predicted by Gurgen Askarian in 1962 and then experimentally observed by N.P. Pilipetsky and A.R. Rustamov in 1965.

Nonlinear Reaction-Diffusion Systems

One can image an excitable medium (EM) as a spatial region G occupied by a medium, in which the processes of autocatalytic production, destruction, and diffusion of some substances occur. In the simplest case, when only one substance of concentration $u = u(\mathbf{r}, t)$ is involved (here $\mathbf{r}$ is a point in three-dimensional space and t is time), the dynamics of u are described by two relationships: the continuity equation

$$u_t + \operatorname{div} \boldsymbol{J} = f(u) \tag{8}$$

(where $u_t = \partial u/\partial t$), and the Fick diffusion law

$$\boldsymbol{J} = -D \operatorname{grad} u. \tag{9}$$

For constant diffusivity D, these two equations reduce to the single parabolic reaction-diffusion equation

$$u_t = D\Delta u + f(u). \tag{10}$$

Here, the symbols div, grad, and Δ designate the spatial divergence, gradient operator, and Laplace operator, respectively; $\mathbf{J} = \mathbf{J}(\mathbf{r}, t)$ is the vector of diffusion flux density of the substance; and the function $f(u)$ (called the kinetic function of the active medium) determines the dependence of the production/destruction rate of substance per unit volume of the concentration u.

In general, an EM is able to support the propagation of traveling-wave fronts and impulses. In the steady regime, plane excitation waves propagate with constant speed preserving their spatial profile. Both the speed and the wave profile are determined by physical parameters of the medium in which the waves propagate and are independent of initial conditions; thus such waves are called *autowaves* (AW). This term was coined by Rem Khokhlov (an early Russian specialist on nonlinear physics and president of Moscow State University) in a 1974 presentation as an official opponent during Anatol Zhabotinsky's defense of his doctoral thesis on periodic chemical reactions. (An enthusiastic mountain climber, Khokhlov sadly perished in the Caucasian mountains in 1977.) The significance of AWs stems from the fact that they frequently occur in physical, chemical and biological applications, including muscles, the nervous system, and the heart.

To reveal the qualitative properties of AW propagation, one needs to solve these equations; but even in the spatially homogeneous case—when the medium is described by Equation (10) with constant D—this problem is challenging. Thus, it is of interest to consider an approach in which the general properties of EM are specialized by means of some simple axioms. Interestingly, this approach leads to an analog of traditional geometric optics, as was first shown by Israel Gel'fand and Sergei Fomin in 1961 (Gel'fand & Fomin, 1961).

When either a traveling-wave front or the leading edge of an impulse moves through an EM, the latter switches from the resting state to the excited state (in the case of impulse, the life time of excitable state is finite and equal to the pulse duration). Neglecting the front width, one may imagine the leading edge as a surface that separates the resting and excited zones, and a motion of this surface (which models the motion of an AW) is interpreted as propagation of excitation. Thus, Gel'fand and Fomin introduced two axioms: (i) each spatial point of the EM can be in one of two states: either in the resting state or in the excited one; (ii) if at some time moment t the excitation has reached some spatial point P, then P immediately becomes a source of propagating excitation. In the case of a nonhomogeneous anisotropic EM, the time period $\mathrm{d}t$ of motion of excitation along the infinitesimal path connecting the points $x(s)$ and $x(s + \mathrm{d}s) = x(s) + (\mathrm{d}x/\mathrm{d}s)\,\mathrm{d}s$ (here s is the parameter) depends both on the point $x(s)$ and on the vector $[\mathrm{d}x/\mathrm{d}s\ :\ \mathrm{d}t = f(x, \mathrm{d}x/\mathrm{d}s)\,\mathrm{d}s]$. The function $f(x, \mathrm{d}x/\mathrm{d}s)$ (which is assumed to be a strict convex function of the second argument) is a key element of the theory; thus, if the point $x_1 = x(s_1)$ is excited, then the time period after which the point $x_2 = x(s_2)$ will be excited is given by the expression

$$\min \int_{x_1}^{x_2} f(x, \mathrm{d}x/\mathrm{d}s)\,\mathrm{d}s,$$

where the minimum is taken over all curves connecting the points x_1 and x_2. (Indeed, if the excitation propagating from the point x_1 along all possible paths has already reached the point x_2 along some path, then all paths that connect x_1 and x_2, but take more time, are insignificant.) This is equivalent to Fermat's principle of least time, adapted to dynamic processes in EM. Thus, Gel'fand and Fomin developed a variational theory of propagation of excitation, endowing it with proper Lagrange–Hamiltonian–Jacobian equations that are formally equivalent to the variational formalism of geometric optics based on Fermat's principle.

The Gel'fand–Fomin formulation follows earlier studies of cardiology by Norbert Wiener and Arturo Rosenblueth, whose system of axioms involves an additional (refractory) state (Wiener & Rosenblueth, 1946). Any point goes to the refractory state immediately upon being excited, after which it cannot be excited during the finite time period R. Supposing the velocity v of propagation of excitation to be constant, Wiener and Rosenblueth developed the theory of circulation of an excitation around non-excitable obstacles and determined the least critical length λ of the obstacle, around which the excitation can stationary circulate: $\lambda = Rv$. Further development of these geometric approaches was undertaken by those involved in a seminar on mathematical biology (organized by Gel'fand at Moscow State University in the early 1960s) and led to a prediction of the existence of spiral waves (Balakhovsky, 1965).

Current geometric theories of spiral waves are concerned with properties that were long ignored, including the dependence of an AW leading-edge velocity on the curvature of the edge. Around 1980, Werner Ebeling, Yakov Zeldovich, Yoshiki Kuramoto, and Vladimir Zykov independently showed that the velocity of a curved leading edge must differ from the velocity of a plane edge—if the edge is concave (convex) in the direction of its propagation then it must accelerate (decelerate). This fact is readily understood in the case of a combustion wave where ignition occurs more rapidly (slowly) ahead of concave (convex) parts of the flame front than ahead of the plane parts due to focusing (diverging) the lines of heat flux. The formula for the velocity v of a weakly bent wave front that propagates in two-dimensional EM and obeys Equation (10) is

$$v = v_0 - DK, \tag{11}$$

where v_0 is the velocity of plane front and K is the curvature of the wave front. The domain of applicability of this equation is given by the inequality $Kl_\mathrm{f} \ll 1$, where l_f is a characteristic width of the front.

Equation (11) implies that excitation of sufficiently small circular patches of EM will die out. The radii of such patches are bounded above by the critical value $\rho_\mathrm{c} = |K_\mathrm{c}|^{-1} = Dv_0^{-1}$ which corresponds to the zero value of v in Equation (11). In a more precise theory of curved AW dynamics (Kuramoto, 1980), Equation (11) emerges as an eikonal equation in the "nonlinear geometric optics" of an AW; thus, it is deeply involved in the modern geometric theory of spiral waves (Mikhailov et al., 1994; Elkin et al., 1998).

Interestingly, a version of Snell's law for AWs emerges from such studies which differs from that of classical optics in several important ways:

- Assuming the underlying reaction-diffusion equation to be Equation (10), Equation (1) is replaced by

  $$\frac{\tan \psi_1}{\tan \psi_2} = \text{constant} = \frac{D_1}{D_2}, \tag{12}$$

  where D_1 and D_2 are the diffusion constants in the incident and refractive regions.
- This is a *local* law which determines the structure of refracted concentration fronts only near the interface between homogeneous regions rather than far from it.
- Equation (12) is obeyed for nonstationary as well as stationary refractions.
- In the regime of stationary refraction, when the incident wave and the refracted one shape a planar form and move with constant velocities, the angles of incidence and refraction are determined by the conditions (Mornev, 1984)

  $$\sin \psi_1 = \sqrt{D_1/(D_1 + D_2)}$$

  and

  $$\sin \psi_2 = \sqrt{D_2/(D_1 + D_2)}.$$

If a plane AW is normally incident on the boundary ($\psi_1 = 0$) and D_2 is sufficiently larger than D_1, the wave will be forced to stop. Assuming $f(u)$ to be a cubic-shaped function possessing three zeroes $u_0 < u_b < u_1$ and satisfying the conditions $f(u) > 0$ at $(u < u_0)$ and $(u_b < u < u_1)$, $f(u) < 0$ at $(u_0 < u < u_b)$ and $(u > u_1)$, the condition for the AW to die at the boundary is

$$\frac{D_2}{D_1} > \frac{\sigma_+}{\sigma_-},$$

where σ_- (σ_+) denotes the least (greatest) of two values $|\int_{u_0}^{u_b} f(u)\,\mathrm{d}u|$, $\int_{u_b}^{u_1} f(u)\,\mathrm{d}u$ (Mornev, 1984).

O. A. Mornev

See also **Diffusion; Fairy rings of mushrooms; Nonlinear optics; Reaction-diffusion systems; Spiral waves; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Balakhovsky, I.S. 1965. Nekotorye rezhimy dvizheniya vozbuzhdeniya v ideal'noi vozbudimoi tkani. [Some regimes of motion of excitations in ideal excitable tissue.] *Biofizika*, 10(6): 1063–1067

Born, M. & Wolf, E. 2002. *Principles of Optics*, 7th edition, Cambridge and New York: Cambridge University Press

Elkin, Yu.E., Biktashev, V.N. & Holden, A.V. 1998. On the movement of excitation wave breaks. *Chaos, Solitons, and Fractals*, 9: 1597–1610

Gel'fand, I.M. & Fomin, S.V. 1961. *Variatsionnoe ischislenie* [*Variational Calculus*], Moscow: Gosudarstvennoe Izdatel'stvo Fiziko-Matematicheskoii Literatury

Kuramoto, Y. 1980. Instability and turbulence of wavefronts in reaction–diffusion systems. *Progress of Theoretical Physics*, 63(6): 1885–1903

Mikhailov, A.S., Davydov, V.A. & Zykpv, V.S. 1994. Complex dynamics of spiral waves and motion of curves. *Physics* D, 70: 1–39

Mornev, O.A. 1984. Elements of the "optics" of autowaves. In *Self-Organization of Autowaves and Structurees Far from Equilibrium*, edited by V.I. Krinsky, Berlin: Springer

Vinogradova, M.B., Rudenko, O.V. & Sukhorukov, A.P. 1990. *Teoriya Voln* (*Wave Theory*), 2nd edition, Moscow: Nauka

Wiener, N. & Rosenblueth, A. 1946. The mathematical formulation of the problem of conduction of impulses in a network of connected excitable elements, specially in cardiac muscle. *Archivos del Instituto de Cardiologia de Mexico* XVI (3–4): 205–265

GEOMORPHOLOGY AND TECTONICS

Geomorphology deals with the evolution of Earth's surface by gravity (e.g., landslides) and the sculpting action of wind, river flow, and ice. Tectonics deals with the deformation and uplift of rocks, including the behavior of earthquakes. Nonlinearity arises in many geomorphic systems in the feedbacks between the shape of the surface and the fluid flow above the surface. The shapes of sand ripples and dunes, for example, control the flow of wind above the surface, which, in turn, controls the spatial distribution of erosion and deposition. Over time, erosion and deposition modify the shape of the surface and the flow of wind in a positive feedback loop. The flow of wind over a dune and the sediment transport caused by the wind are both nonlinear processes. The relationship between the sediment flux moved by the wind over a dune, for example, and the shear stress exerted by the wind is strongly nonlinear, including both a threshold shear stress for particle entrainment and a nonlinear power-law relationship between flux and shear stress above that threshold.

Some of the nonlinear feedback relationships in geomorphic systems affect practical issues. The feedback between vegetation density and soil erosion, for example, can lead to dust bowl conditions if low vegetation density promotes wind erosion, stripping the soil to further inhibit vegetation growth in a positive feedback loop. In tectonic systems, nonlinear feedbacks also govern much of the interesting behavior. Deformations in rock or along a fault plane, for example, can become localized by the feedback between shear strength and strain rate. Rocks often become weaker as they are strained, in turn focusing more deformation to the areas of highest strain. The rheology of rocks, including their nonlinear dependence on strain rate and time, thus plays a very important role in our understanding of mountain belts.

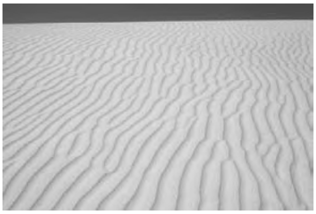

Figure 1. Sand ripples with a characteristic width or spacing. White Sands, New Mexico, courtesy of National Park Service.

Self-organized, periodic landforms have received considerable attention in geomorphology. Sand ripples, dunes, and yardangs (all examples of eolian landforms), flutes (elongated ridges of subglacial sediment), drumlins (sculpted large mounds of subglacial debris), finger lakes, cirques, sorted stone stripes and circles (all glacial and periglacial landforms), discontinuous ephemeral streams and step-pool sequences (fluvial landforms), beach cusps and spits (coastal landforms) all have a characteristic width or spacing (Figure 1) that is controlled in part by the fluid movement on or above the surface and its relationship to erosion and deposition. The dynamic processes governing these systems are disparate and complex, but feedback between the surface and the fluid flow on or above the surface plays a key role in all of these examples. Finger lakes, for example, are elongated glacial troughs that developed on the margins of former ice sheets as glacier flow was focused into incipient bedrock depressions, concentrating glacial erosion in troughs in a positive feedback that forms deeper basins when the ice retreats. The Finger Lakes of central New York are the typical examples, but similar lakes occur along other former ice margins. Step-pool sequences form in mountain river channels by feedback between the roughness of the channel bed, the flow velocity in the channel, and the selective entrainment of particles on the bed. A channel reach with an initially rougher bed than nearby reaches will have slower flow velocity, promoting the deposition of coarse particles along the channel reach and a further decrease in flow velocity. This feedback produces channel reaches characterized by shallow gradients and fine particles alternating with steep gradients and coarse particles. In some cases, numerical models have been developed that enable the wavelengths of these periodic landforms to be predicted.

In many if not most cases, landform evolution does not lead to periodic topography at all. River networks, coastal erosion of a rugged shoreline, and the dissolution of limestone to form cave networks (or karst topography above ground) are all examples of processes that create chaotic landforms with no apparent characteristic scale. Many of these landforms have an underlying order, however, by virtue of their self similarity. Self-similar landforms are those that have a similar appearance and statistical structure at a wide range of spatial scales. Rugged coastlines are the classic example of fractals popularized by Benoit Mandelbrot (1982). Long before the fractal structure of coastlines had been described, however, geomorphologists were interested in the self-similarity of river networks. To examine this self-similarity, streams are first ordered according to their position in a drainage network. The Strahler order defines all channels with no upstream tributaries as first-order channels. Whenever any two streams of like order join at a tributary, they form a stream of the next highest order. Horton's law states the principle of self-similarity of drainage networks mathematically: the ratio of the number of streams of order n to the number of streams of order $n+1$ is equal to approximately 4 and is independent of n (Rodriguez-Iturbe and Rinaldo (2001) provide an excellent review). Similar relationships exist between channel lengths, areas, and the Strahler order.

Horton's laws are satisfied in many different kinds of networks; however, so the modern view is that Horton's laws are not unique properties of drainage networks (Kirchner, 1993). Other relationships, such as the angles of tributary junctions and the relationships between channel slope and the Stahler order, contain important information about the self-organization of channel networks.

The Earth's crust exhibits nonlinear, critical behavior in several ways. First, earthquakes occur over a wide range of sizes with a frequency-size distribution characterized by a power law. This observation, known as the Gutenberg–Richter law, is the most fundamental rule of seismology. Seismicity also exhibits temporal correlations that have self-similar properties. Omori's law, for example, states that the frequency of foreshocks or aftershocks is inversely proportional to the time before or since the mainshock. Theoretical models for fault behavior have been devised based upon a simple model of blocks (representing one fault plane) frictionally coupled to a table (the other fault plane) and elastically coupled to one another and to a driver plate (representing the regional tectonic stress) (e.g., Turcotte, 1997). The model builds up stress until all the elements are near the threshold for slipping. At that point, the slippage of one block can transfer stress to other blocks in a cascade that produces earthquakes that follow the Gutenberg–Richter law. This phenomenon is called stick-slip friction. The discrete nature of the slider-block model appears to have an analog in real faults; the behavior of faults can be characterized according to the size and strength of asperities (sticky spots) on the fault plane, which are like the individual blocks in the slider-block system. The aftershock behavior of earthquakes does not appear to arise naturally in the simplest slider-block models; instead, some viscous coupling between blocks is necessary to reproduce Omori's law (Pelletier, 2000), suggesting that the punctuated process of seismic events is linked to a steady, creeping motion of the fault motion over longer time scales.

The crust also exhibits nonlinear critical behavior in the way that magma makes its way through the crust and is released as volcanic eruptions. The frequency–size distribution of volcanic eruptions appears to follow a power-law distribution analogous to the Gutenberg–Richter distribution in terms of the volume of material released. Magmatic and volcanic activity is also clustered in time in a way that is generally analogous to Omori's law. The processes of fluid movement through the crust are very different compared with the stresses on an earthquake fault, but a model for the spatial interaction of many fluid conduits all near the threshold for eruption produces a model very similar to the slider-block model (Pelletier, 1999). This kind of model appears to be most consistent with the Rayleigh distillation model of geochemical mixing (Turcotte, 1997), which describes the abundances of certain geochemical elements in the crust.

Jon D. Pelletier

See also **Avalanches; Branching laws; Dune formation; Evaporation wave; Feedback; Fractals; Glacial flow; Rheology; Sandpile model**

Further Reading

Kirchner, J.W. 1993. Statistical inevitability of Horton's laws and the apparent randomness of stream channel networks. *Geology*, 21: 591–594

Mandelbrot, B.B. 1982. *Fractal Geometry of Nature*, New York: W.H. Freeman

Pelletier, J.D. 1999. Statistical self-similarity of magmatism and volcanism. *Journal of Geophysical Research*, 104: 15,425–15,438

Pelletier, J.D. 2000. Spring-block models of seismicity: review and analysis of a structurally heterogeneous model coupled to a viscous asthenosphere. In *GeoComplexity and the Physics of Earthquakes*, edited by J.B. Rundle, D.L. Turcotte, & W. Klein, Washington, DC.: American Geophysical Union

Rodriguez-Iturbe, I. & Rinaldo, A. 2001. *Fractal River Basins: Chance and Self-Organization*, Cambridge and New York: Cambridge University Press

Turcotte, D.L. 1997. *Fractals and Chaos in Geology and Geophysics*, 2nd edition, Cambridge and New York: Cambridge University Press

GESTALT PHENOMENA

The Gestalt idea was introduced to science by Ernst Mach (1868) and Christian von Ehrenfels (1890). Mach stated that the spontaneous creation of order, that is, order arising without any external control, can be shown in inanimate nature. Von Ehrenfels characterized Gestalt qualities, that is, higher order qualities emerging from basic elements, by two criteria: (1) supersummativity, which means that the elements of a pattern presented individually to a person give, in the totality of the experience, a poorer impression than the total experience of a person to whom all the elements are presented; and (2) transposition, which means the characteristics of a Gestalt quality are retained even if all the elements which exhibit the Gestalt quality are changed in a certain way (for example, the transposition of a melody).

Wolfgang Köhler (1920) delivered the earliest formulation of a concept of self-organization of perception. The idea that perception must necessarily be understood as a process of autonomous creation of order runs through all his works. Starting from the observation of spontaneous Gestalt tendencies in experience, Köhler made it his primary task to design and test a model of brain function in which the phenomenal organization of the perceptual world is explained as not only stimulus-dependent, but as strongly dependent upon the perceptual system's own inner dynamics. In the development of Gestalt theory, Köhler was mainly oriented to the observations and theoretical concepts of physics. In accordance with the assumption of linear thermodynamics, the general systemic tendency towards final equilibrium was considered in Köhler's time to be the only basic principle of self-organization. This principle can easily be demonstrated in cognition by recursive experiments of serial reproduction of complex patterns. These patterns follow the "principle of prägnanz" towards very simple and stable configurations (Stadler & Kruse, 1990; Kanizsa & Luccio, 1990). In perception, this principle states that people will perceive the most orderly or regular thing they can out of the stimuli that are presented to them. Köhler applied the model of physical fields striving independently to balance forces directly to the way in which the visual system functions. At the time, this almost provocatively contradicted the findings of neuroanatomy and neurophysiology. On the basis of his theoretical model of perception, the brain is not seen as a complex network of many different interacting neurons but as a homogeneous conductor of bioelectric forces.

Köhler's argument was not primarily the postulation of electromagnetic field forces acting in the brain independently of neuroanatomical structures (which has been refuted by most contemporary brain scientists) but the idea of self-organization in the brain. He was fully aware of the fact that the general principle of development of linear thermodynamics has been rather unsuitable for application to biological systems as long as it was oriented exclusively towards the final equilibrium. This was criticized by Köhler himself in 1955: "Although this is a perfectly good principle, it cannot, in its present formulation, be applied to the organism. For the organism is obviously not a closed system; moreover, while the direction indicated by the principle may be called 'downward', the direction of events in healthy organisms is on the whole clearly not 'downward' but, in a good sense, 'upward'."

Modern Developments

The brain is conceived as a self-organizing system that can be treated by means of synergetics (Haken, 1983, 1990). Pattern recognition is understood as pattern formation (of activities of the neural net). Incomplete (visual) data are complemented by a dynamic associative memory. In both cases (pattern formation and recognition), incomplete data generate order parameters that compete with each other. In general, one order parameter wins and generates, according to the slaving principle of synergetics, the complete pattern. In this processes, idealized (Gestalt) patterns may be incorporated. Of particular interest are ambiguous figures, such as Figure 1 (young woman vs. old woman). Here, two or more interpretations (percepts) are possible, and two (or more) order parameters may win the competition. The final outcome is determined by an order parameter dynamics in which the dynamics of attention parameters is included.

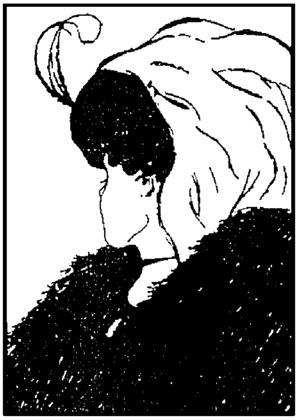

Figure 1. Young woman vs. old woman.

The mathematical approach (algorithm of the "synergetic computer") is as follows: The images of different objects are decomposed into their pixels that are lumped together as pixel vectors

$$\boldsymbol{v}_\mu = (v_{\mu 1}, v_{\mu 2}, ..., v_{\mu N}), \quad \mu = 1, ..., M. \tag{1}$$

The label μ is associated with an interpretation (the name of a person, say), and the adjoint vectors $\boldsymbol{v}_\mu^+$ are defined by

$$(\boldsymbol{v}_\mu^+ \boldsymbol{v}_\nu) = \delta_{\mu\nu}. \tag{2}$$

The activity pattern of the neural net is written as

$$\boldsymbol{q}(t) = \sum_{\mu=1}^{M} \xi_\mu(t) \boldsymbol{v}_\mu + \boldsymbol{r}(t), \tag{3}$$

where $\boldsymbol{r}$ is a residual term that vanishes in the course of time. The order parameters are defined by

$$\xi_\mu(t), \tag{4}$$

where the initial value at the observation time t_0 is given by

$$\xi_\mu(0) = (\boldsymbol{v}_\mu^+ \boldsymbol{q}(0)). \tag{5}$$

The order parameters obey competition equations that can be derived from a potential function V

$$\mathrm{d}\xi_\mu(t)/\mathrm{d}t = -\partial V/\partial \xi_\mu, \tag{6}$$

where

$$V = V(\xi_1, ..., \xi_M; \lambda_1, ..., \lambda_M) \tag{7}$$

does not only depend on the order parameters but also on the attention parameters λ_j. The competition equations read explicitly

$$\mathrm{d}\xi_\mu(t)/\mathrm{d}t = \left(\lambda_\mu - B \sum_{\mu' \neq \mu}^{M} \xi_{\mu'}^2 - D \sum_{\mu'=1}^{M} \xi_{\mu'}^2 \right) \xi_\mu \tag{8}$$

with positive constants λ_μ, B, C. The winning order parameter fixes the activity pattern (3). In the case of ambiguous patterns (such as that of Figure 1), equations for the order parameters ξ_1 and ξ_2 and the attention parameters read

$$\mathrm{d}\xi_1/\mathrm{d}t = (\lambda_1 - C\xi_1^2 - (B+C)\xi_2^2)\xi_1 - \partial V_\mathrm{b}/\partial \xi_1, \tag{9}$$

$$\mathrm{d}\xi_2/\mathrm{d}t = (\lambda_2 - C\xi_2^2 - (B+C)\xi_1^2)\xi_2 - \partial V_\mathrm{b}/\partial \xi_2, \tag{10}$$

$$\mathrm{d}\lambda_j/\mathrm{d}t = \gamma(1 - \lambda_j - \xi_j^2), \quad j = 1, 2. \tag{11}$$

The bias potential V_b is defined by

$$V_\mathrm{b} = 2B\xi_1^2\xi_2^2 \left(1 - 4\alpha \frac{\xi_1^2 - \xi_2^2}{\xi_1^2 + \xi_2^2} \right). \tag{12}$$

The parameter α is determined by the percentage of that perception that is seen first. It also determines the relative length of the perception times that occur in the oscillatory motion of the order parameters ξ_1, ξ_2 that represent the switch from one percept to the other one and back again.

HERMANN HAKEN AND MICHAEL A. STADLER

See also **Cell assemblies; Emergence; Synergetics**

Further Reading

von Ehrenfels, C. 1890. Über Gestaltqualitäten [On Gestalt qualities]. In *Foundations of Gestalt Theory*, edited by B. Smith, München: Philosophia-Verlag, pp. 82–117

Haken, H. 1983. Synopsis and introduction. In *Synergetics of the Brain*, edited by E. Başar, H. Flohr, H. Haken & A.J. Mandell, Berlin and New York: Springer, pp. 3–25

Haken, H. 1990. *Synergetic Computers and Cognition: A Top-down Approach to Neural Nets*, Berlin and New York: Springer

Kanizsa, G. & Luccio, R. 1990. The phenomenology of autonomous order formation in perception. In *Synergetics of Cognition*, edited by H. Haken & M. Stadler, Berlin and New York: Springer, pp. 186–200

Köhler, W. 1920. *Die physischen Gestalten in Ruhe und im stationären Zustand*, Braunschweig: Vieweg

Köhler, W. 1955. Direction of processes in living systems. *Scientific Monthly*, 8: 29–32

Mach, E. 1868. Die Gestalten der Flüssigkeit [The Gestalts of fluids]. In *Populärwissenschaftliche Vorlesungen*, 4th edition, 1910, Leipzig: Barth

Stadler, M. & Kruse, P. 1990. The self-organization perspective in cognition research: historical remarks and new experimental approaches. In *Synergetics of Cognition*, edited by H. Haken & M. Stadler, Berlin and New York: Springer, pp. 32–52

GINZBURG–LANDAU EQUATION

See **Complex Ginzburg–Landau equation**

GLACIAL FLOW

Glaciers are defined as multi-year features, consisting of snow and ice, which flow down-slope under the force of gravity. The broadness of this definition means that there exists a continuum of glaciers that ranges from small, multi-year snow patches with surface areas of the order of 100 square meters, to the Columbia Glacier in Alaska, with a surface area of over 1100 square kilometers (the District of Columbia would fit within its terminus). Yet, in spite of the broad range of features encapsulated in this definition, the same basic physical processes are common to all of the world's roughly 160,000 different glaciers, and most of these processes are nonlinear.

Glaciers deform under their own weight, behaving like highly viscous fluids. Mass input is greatest at the glacier's upper elevations where colder temperatures result in a greater percentage of precipitation falling as snow. Mass loss is greatest at the glacier terminus, the lowest point on the glacier, where temperatures and melting are highest and where ablation (mass loss from melting, evaporation, sublimation, and in special cases, iceberg formation) equals flow. This imbalance results in a continuous mass transfer from the upper reaches to the lower reaches. The basis of the equations governing glacial flow is therefore mass conservation. However, unlike liquid water, where an applied stress, τ, (i.e., a squeeze) causes a linear, proportional deformation or strain, glaciers have a nonlinear stress-strain response. Although measurements in remote mountain locations are difficult and limited, field studies and laboratory experiments have shown that

$$\dot{\varepsilon} = A\tau^n, \tag{1}$$

where n is constant and generally assigned a value of 3 (though values ranging from 1.5 to 4.2 can be found in the literature), $\dot{\varepsilon}$ is the rate of deformation, and $1/A$ is a nonlinear measure of the viscosity. This glacier "flow law" is a version of pseudoplastic flow and is similar to dry sand dune flows (which use a value of 2 for n). Viscoplastic flows, such as clay-water mixtures, and Bagnold macro-viscous flows, such as mud-flows, are also similar, differing primarily in the value of the exponent n.

The nonlinear flow law is the source of many fractal, self-similar, and nonlinear scaling properties. The basic scaling relationships are simple: discharge of ice through a given cross section of the glacier is proportional to glacier depth raised to the power $(n+2)$, and glacier flow velocity is proportional to glacier depth raised to the power $(n+1)$ (Patterson, 1994). However, Bahr (1997) has shown that in conjunction with mass and momentum conservation, the nonlinear flow law implies nonlinear scaling relationships between surface area (a parameter easily measured with satellites) and many difficult to measure but fundamental properties. Glacier thickness, volume, mass balance, velocity, flux, and other parameters relate to the surface area by exponents of 3/8, 11/8, etc. Using these unusual scaling exponents, the volume of ice in the world's glaciers can be predicted based solely on observed surface areas.

Equation (1) applies to basic glacier flow under constant stress. In reality, glaciers are rarely under consistent stress throughout. In areas where glaciers are under tensional stress, if the stress becomes too high, the ice becomes brittle and fractures, resulting in crevassing. Crevassing can be mathematically described using fracture mechanics (Smith, 1976; Sassolas et al., 1995). Crevassing can occur as the glacier flows over large drops in the bed as a result of varying flow speeds (Harper et al., 1998). Particularly dramatic crevassing occurs in the lower reaches of retreating tidewater glaciers as a result of faster flow at the glacier terminus than in the ablation zone (the area of the glacier that is annually losing more mass than it is gaining). Because the lower sections of tidewater glaciers are near or at floatation, unlike land-based glaciers, tidewater glacier crevassing can result in calving (the formation of icebergs from pieces that are broken off the terminus). Calving from tidewater glaciers can be modeled using both fracture mechanics and percolation theory (Bahr, 1995).

Tidewater glacier calving contributes to an unexpected nonlinearity in the tidewater glacier terminus position. Most glaciers move back and forth with changes in climate as the balance between melt and accumulation shifts. For tidewater glaciers, however, there is an additional loss of mass through calving. The tidewater glacier calving rate increases with water depth, but water depth is typically minimized by a pile of debris deposited at the end of the glacier. If the glacier terminus retreats backwards off the debris pile, then the water depth increases and the calving rate increases, further increasing glacier retreat (Meier, 1993; van der Veen, 2002). This is a classic positive feedback loop scenario, and is the cause of the dramatic and rapid retreats recently seen in many glaciers that terminate in water, such as the Columbia Glacier in Alaska and many of the New Zealand glaciers that terminate in lakes. While many of the world's glaciers are slowly retreating due to changes in climate, these tidewater glaciers retreat nonlinearly in response to even the smallest climatic perturbations.

In addition to surface and internal processes, such as flow and crevassing, glaciers exhibit nonlinear behavior in their basal processes. For temperate glaciers (those not frozen to their beds), glacial flow is a combination of ice deformation and sliding at the glacier bed. Motion tends to be stick-slip, very similar to the nonlinear slider-block models of earthquakes (Bahr & Rundle, 1996; Fischer & Clarke, 1997). This gives rise to the grinding of the underlying rock, plucking of rocks out of the bed, and deformation of the bed in places where it is a fine-grained matrix. Lubrication appears to increase flow rates, as it does in sub-surface faults (Patterson, 1994). At the extreme end of lubricated basal flow, we find surging glaciers. These glaciers appear to build up water and water pressure at the glacier bed. Some mechanism or pressure trigger allows this water to be periodically catastrophically released (the Variegated Glacier in Alaska, for example, surged in 1906, 1947, 1964–65, and 1982–83), resulting in rapid flow and over-extension of the glacier (Patterson, 1994).

Finally, the overall structure of large glaciers is fractal. Large glaciers, such as the Talkeetna and Columbia Glaciers in Alaska have multiple upper branches that coalesce into one outlet tongue, similar to a river system or branching tree. Measurements have shown that the structure is statistically self-similar with fractal dimensions ranging from roughly 1.6 for small-to mid-sized mountain glaciers, to 2.0 for large glaciers and space-filling ice sheets such as Greenland (Bahr & Peckham, 1996).

KAREN LEWIS MACCLUNE AND DAVID BAHR

See also **Avalanches; Dune formation; Geomorphology and tectonics; Sandpile model**

Further Reading

Bahr, D.B. 1995. Simulating iceberg calving with a percolation model. *Journal of Geophysical Research*, 100(B4): 6225–6232

Bahr, D.B. 1997. Global distributions of glacier properties: a stochastic scaling paradigm. *Water Resources Research*, 33(7): 1669–1679

Bahr, D.B. & Peckham, S. 1996. Observations of self-similar branching topology in glacier networks. *Journal of Geophysical Research*, 101(B11): 25511–25521

Bahr, D.B. & Rundle, J.B. 1996. Stick-slip statistical mechanics of motion at the bed of a glacier. *Geophysical Research Letters*, 23(16): 2073–2076

Fischer, U.H. & Clarke, G.K.C. 1997. Stick-slip sliding behavior at the base of a glacier. *Annals of Glaciology*, 24: 390–396
Harper, J.T., Humphrey N.F. & Pfeffer, W.T. 1998. Crevasse patterns and the strain rate tensor: a high-resolution comparison. *Journal of Glaciology*, 44(146): 68–76
Hooke, R. 1998. *Principles of Glacier Mechanics*, Saddle River, NJ: Prentice-Hall
Meier, M.F. 1993. Columbia Glacier during rapid retreat: interactions between glacier flow and iceberg calving dynamics. *Workshop on the Calving Rate of West Greenland Glaciers in Response to Climate Change*, Copenhagen
Patterson, W.S.B. 1994. *The Physics of Glaciers*, 3rd edition, New York: Elsevier
Sassolas, C., Pfeffer, T. & Amadei, B. 1995. Stress interaction between multiple crevasses in glacier ice. *Cold Regions Science and Technology*, 24: 107–116
Sharp, R.P. 1988. *Living Ice*, Cambridge and New York: Cambridge University Press
Smith, R.A. 1976. The application of fracture mechanics to the problem of crevasse penetration. *Journal of Glaciology*, 17(76): 223–228
van der Veen, C.J. 2002. Calving glaciers. *Progress in Physical Geography*, 26(1): 96–122

GLOBAL WARMING

Few modern scientific concerns have achieved such notoriety as the possibility of relatively rapid anthropogenic global warming through increased CO_2 emissions. This complex problem became a matter of considerable public attention during the 1980s, and during the 1990s, the first attempt was made at international management of the challenge (the Kyoto Protocol under the United Nations Framework Convention on Climatic Change). However, scientific awareness of CO_2-induced climatic change is not new, and the underlying physical processes were understood from the beginning of studies concerning the absorption of radiation by the atmosphere. Later research resulted in a deeper understanding of the dynamics of the biospheric carbon cycle, and global atmospheric circulation models have been adopted, and adapted, for assessing the future course of tropospheric CO_2 levels. In spite of all of these advances, much remains unclear and uncertain.

Early Studies

Several years before his death in 1830, the French mathematician Joseph Fourier concluded that the atmosphere acts like the glass of a greenhouse, letting light through and retaining the invisible rays emanating from the ground (Fourier, 1822). In modern scientific terms, the atmosphere is highly (though not perfectly) transparent to incoming (shortwave) solar radiation, but it is a strong absorber of certain wavelengths in the outgoing (longwave) infrared spectrum produced by the reradiation of absorbed sunlight.

John Tyndall was the first scientist to study this process in detail by measuring the absorptive properties of air and its key constituent molecules (water vapor and about a dozen different compounds). He used a sensitive galvanometer to measure the electric current passing through gases irradiated by heat. In 1861, Tyndall concluded that water vapor accounts for most of the atmospheric absorption and hence "every variation of this constituent must produce a change in climate. Similar remarks would apply to the carbonic acid diffused through the air. . ." (Tyndall, 1861). The next major contribution to the field came just before the end of the 19th century when Svanté Arrhenius offered the first calculations of the global surface temperature rise resulting from naturally changing atmospheric CO_2.

Arrhenius's conclusions contained all of the key qualitative modern results. He found that geometric increases of CO_2 will produce a nearly arithmetic rise in surface temperatures, that the warming will be smallest near the equator and highest in polar regions, that the Southern hemisphere will be less affected, and that the warming will reduce temperature differences between night and day (Arrhenius, 1896). His quantitative results also resembled those of today's best global climatic models: he predicted that the increase in average annual temperature will be about 50°C in the tropics and just over 6°C in the Arctic. All of these findings applied to natural fluctuations of atmospheric CO_2: Arrhenius concluded (correctly) that future anthropogenic carbon emissions would be largely absorbed by the ocean and (incorrectly, as he grossly underestimated future fossil fuel combustion) that the accumulation would amount to only about 3 ppm in half a century.

The link between CO_2 and climate change was resurrected in 1938 by George Callendar who calculated a more realistic temperature rise with doubling of CO_2 concentrations (1.5°C rise) and documented a slight global warming trend of 0.25°C for the preceding half a century (Callendar, 1938). In his later writings, he also recognized the importance of carbon emissions from land-use changes. In 1956, Gilbert Plass performed the first computerized calculation of the radiation flux in the main infrared region of CO_2 absorption (Plass, 1956). His results (average surface temperature rise of 3.6°C with the doubled atmospheric CO_2) were published a year before Roger Revelle and Hans Suess summarized the problem with continuing large-scale fossil fuel combustion in such a way that the key sentence has become a citation classic:

> Thus human beings are now carrying out a large scale geophysical experiment of a kind that could not have happened in the past nor be reproduced in the future. Within a few centuries we are returning to the atmosphere and oceans the concentrated organic carbon stored in sedimentary rocks over hundreds of millions of years. (Revelle & Suess, 1957)

An almost instant response to this concern was the setting up of the first two permanent stations for the measurement of background CO_2 concentrations, at Mauna Loa in Hawai'i and at the South Pole. Accumulating measurements began showing a steady rise of atmospheric CO_2 at these two remote locations, but, once again, attention to the problem of potential global warming eased during the 1960s and began to grow only in the aftermath of OPEC's two sudden oil price hikes during the 1970s.

Numerical Models of Anthropogenic Global Warming

By the late 1960s, improvements in computer capabilities made it possible to run the first three-dimensional models of global atmospheric circulation and use them to simulate the effects of higher CO_2 levels. Most of these simulations looked at possible effects arising from the doubling of preindustrial CO_2, that is, after reaching levels around 600 ppm. Initial simulations indicated a 2.93°C rise with the doubling of the CO_2 level to 600 ppm (Manabe & Wetherald, 1967). Increases in computing power (subject to Moore's famous law) and better understanding of interactions between the atmosphere, oceans, and the terrestrial biosphere has led to increasingly more realistic models of global climate. Another important refinement was the inexplicably delayed consideration of other greenhouse gases (above all, of CH_4, N_2O, and chlorofluorocarbons) whose combined radiative forcing is now slightly higher than that of carbon dioxide (about 1.5 and $1.4\,\mathrm{W\,m^{-2}}$).

By the late 1990s, the best models coupled the atmosphere's physical behavior with changes on land and in the ocean, and with simulations of some key features of carbon and sulfur cycles and of atmospheric chemistry (Houghton et al., 2001). At the same time, even our best numerical models still represent the atmosphere with a relatively coarse grid and are incapable of reproducing the intricacies and multiple feedbacks that determine the course and the rate of climate change.

One of the most important sources of potential error in the climate models is the treatment of clouds. The best general circulation models represent fairly well some essential gross features of global atmospheric physics but their iterative calculations are done at such widely spaced points of three-dimensional grids that it is impossible to treat cloudiness in a realistic manner. And yet clouds are key determinants of the planetary radiation balance because they have, on balance, a pronounced net cooling effect. Because clouds account for about half of the Earth's albedo (the fraction of incident radiation that is reflected), even relatively small changes in their properties could have an appreciable effect on the course of global warming.

Other unresolved matters include the response of terrestrial biota (Will carbon sequestration take place mostly in short- or long-lived tissues or in soil?), marine algae (especially their role in forming clouds), sudden releases of methane (rising temperatures may lead to catastrophic emissions from methane hydrates), and effects of orbital and solar influences (particularly a very high correlation between the solar cycles shorter than the 11-year mean and higher average land temperature of the Northern Hemisphere, and a link between mid-atmospheric temperature and changing intensity of radiation over the sunspot cycle). If the global forecasts are uncertain, regional predictions are particularly questionable. The most complex coupled models now provide reasonably reliable simulations of climate down to the sub-continental level but their results still have unacceptably large variations on regional scales.

Geological Evidence for Global Warming and Cooling

Indirect or proxy markers (such as isotopic and trace chemical analysis on tree rings, ice, or sediment cores) make it clear that a substantial decline of atmospheric CO_2 preceded the most extensive and longest lasting (some 70 million years, or Ma) glaciation of the entire Phanerozoic era that began about 330 Ma ago. Approximate reconstruction of CO_2 levels for the past 300 Ma—since the formation of the Pangea whose eventual break-up led to the current distribution of oceans and land masses—indicates, first, a pronounced rise (about five times the current level during the Triassic period), followed by a steep decline (Berner, 1998; Figure 1, top). Boron-isotope ratios of planktonic foraminifer shells point to CO_2 levels above 2000 ppm 60–50 Ma ago (with peaks above 4000 ppm), followed by an erratic decline to less than 1000 ppm by 40 Ma ago, and relatively stable and low (below 500 ppm) concentrations ever since the early Miocene 24 Ma ago (Pearson & Palmer, 2000; Figure 1, bottom).

Reliable record of atmospheric CO_2 is available only for the past 420,000 years thanks to the analyses of air bubbles from ice cores retrieved from Antarctica and Greenland. Preindustrial CO_2 levels never dipped below 180 ppm and never rose above 300 ppm (Raynaud et al., 1993; Petit et al., 1999; Figure 2) and their oscillations are highly positively correlated with changing temperatures. But these correlations are not a proof of a clear cause-and-effect relationship as there are no obvious lead-lag sequences. Other recent paleoclimatic studies actually found signs of decoupling of atmospheric CO_2 and global climate during the Phanerozoic eon and particularly during the early to middle Miocene, when a warm period coexisted with low CO_2 levels (Veizer et al., 2000; Pagani et al., 1999). These findings confirm the complexity of climate change where cause and effect are difficult to assign: atmospheric CO_2 may have

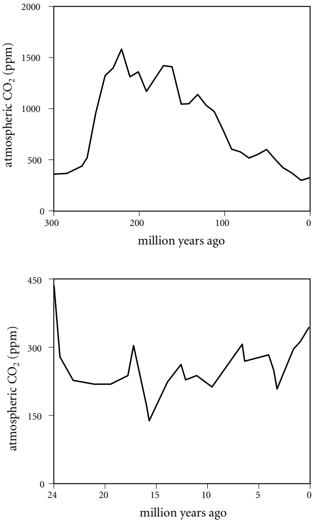

Figure 1. Atmospheric CO_2 concentrations during the past 300 and 24 million years. Based on Berner (1998) and Pearson and Palmer (2000).

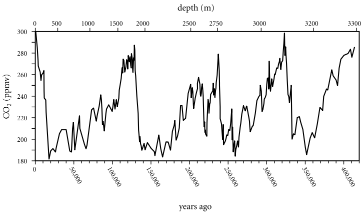

Figure 2. Atmospheric CO_2 concentrations during the past 420,000 years derived from air bubbles in Antarctica's Vostok ice core. Based on Petit et al. (1999).

been a primary climate driver but the evidence is not conclusive (Kump, 2002). The most likely pacemaker during the Pleistocene period was small changes in the Earth's orbit around the Sun; massive methane releases from gas hydrates and volcanic activity must be also considered.

Recent Evidence for Global Warming

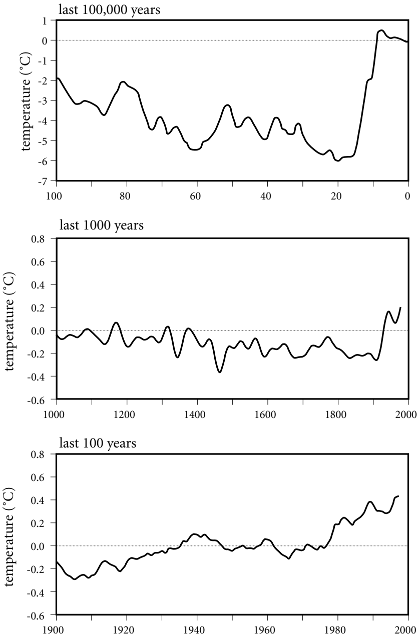

Figure 3. Reconstructed temperature trends during the past 100,000 years (from the Vostok ice core), 1000 years (for the Northern Hemisphere), and a 5-year running mean from instrumental temperature measurements for the past 100 years. Reproduced from Smil (2002).

During the time between the rise of the first high civilizations (5000–6000 years ago) and the beginning of the fossil fuel era, atmospheric CO_2 levels had fluctuated within an even narrower range of 250–290 ppm. Subsequent anthropogenic emissions pushed atmospheric concentrations of CO_2 to a high of 370 ppm by the year 2000. Paleoclimatic studies of the Northern Hemisphere during the last millennium show warming periods during the 12th and 18th centuries and pronounced cooling during the 15th century (the Little Ice Age). A demonstrable cooling trend between the late 18th and the early 20th centuries was followed by an unprecedented rate of warming that has brought the average planetary temperature to levels higher than at any time during the past 1000 years (Figure 3).

The most extensive studies of the existing global record of surface temperatures have detected long-term planetary warming of, respectively, 0.5°C and 0.78°C since the middle of the 19th century (Jones et al., 1986; Hansen & Lebedeff, 1988). Changes in measurement techniques (different thermometers), station locations (from downtowns to suburbs) and station environment (increasing urban heat island effect), and until very recently, highly inadequate coverage of large areas of the Southern Hemisphere complicate the interpretation of this shift, which has distinct spatial patterns with areas of more pronounced warming and regions of slight cooling. However, the most recent (post-1976) spell of warming has been almost global and the 1990s were the warmest decade since the beginning of the

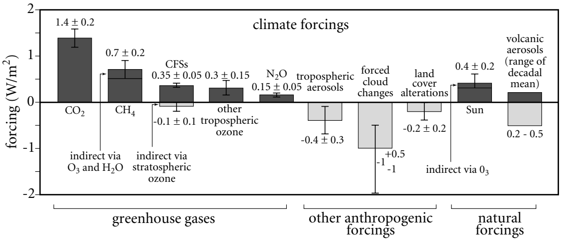

Figure 4. Estimates of cumulative radiation forcings by greenhouse gases and aerosols between 1850 and 2000 according to Hansen et al. (2000).

instrumental record in the 1850s and the warmest ten years of the millennium in the Northern Hemisphere.

According to the general circulation models, the warming should have been more pronounced. The best explanation of the discrepancy between the models and the actual temperature record is that the warming was partially counteracted by sulfate aerosols. The combined direct and indirect effect of all greenhouse gases resulted in a total anthropogenic forcing of about 2.8 W m^2 by the late 1990s (Hansen et al., 2000; Figure 4). This is equal to a little more than 1% of solar radiation reaching the ground.

Future Climate

If the atmospheric warming was primarily the function of radiative forcing, then the level of greenhouse gas emissions would be the key variable. Recent emission scenarios for CO$_2$ alone offer a very large range of concentrations, 540–970 ppm, by the year 2100. CH$_4$ levels may range even wider, from just above 1500 ppb to more than 3600 ppb. The aggregate radiative forcing may thus be anywhere between 4 and 9 W m^{-2} and the climate sensitivity would then range between 1.5°C and 4°C with 2.2–3°C considered to be the most likely by the latest IPCC assessment. Broad consensus from the latest generation of models foresees that this climatic change would cool the stratosphere while raising the tropospheric temperatures in a distinct spatial pattern, with the warming more pronounced on the land (and during nights) and with increases of about two to three times the global mean in higher latitudes in winter than in the tropics, and greater in the Arctic than in the Antarctic.

There are many effective ways to slow down the greenhouse gas emissions and reduce their environmental impact. Most significantly, the affluent countries could largely retain their quality of life while reducing their energy and material consumption by at least a third. While the means are available, the will to act, nationally and internationally, is mostly absent. Global warming is a complex natural process but its anthropogenic enhancement calls for a fundamentally moral solution that runs against some basic human propensities: consume less and do so more efficiently.

VACLAV SMIL

See also **Atmospheric and ocean sciences; Forecasting; General circulation models of the atmosphere**

Further Reading

Alverson, K.D., Bradley, R.S. & Pedersen, T.F. (editors). 2003. *Paleoclimate, Global Change, and the Future*, Berlin and New York: Springer

Arrhenius, S. 1896. On the influence of carbonic acid in the air upon the temperature of the ground. *Philosophical Magazine, Series 5*, 41: 237–276

Berner, R.A. 1998. The carbon cycle and CO$_2$ over Phanerozoic time: the role of land plants. *Philosophical Transactions of the Royal Society of London* B, 353: 75–82

Callendar, G.S. 1938. The artificial production of carbon dioxide and its influence on temperature. *Quarterly Journal of the Royal Meteorological Society*, 64: 223–237

Fourier, J.B.J. 1822. *Théorie Analytique de la Chaleur*, Paris: Firmin Didot

Hansen, J. & Lebedeff, S. 1988. Global surface air temperatures: update through 1987. *Geophysical Research Letters*, 15: 323–326

Hansen J., Sato, M., Ruedy, R., Lacis, A. & Oinas, V. 2000. Global warming in the twenty-first century: an alternative scenario. *Proceedings of the National Academy of Sciences USA*, 97: 9875–9880

Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J. & Xiaosu, D. (editors). 2001. *Climate Change 2001: The Scientific Basis*, Cambridge and New York: Cambridge University Press

Jones P.D., Wigley, T.M.L. & Wright, P.B. 1986. Global temperature variations between 1861 and 1984. *Nature*, 322: 430–434

Kump, L.R. 2002. Reducing uncertainty about carbon dioxide as a climate driver. *Nature*, 419: 188–190

Lozán, J.L., Grassl, H. & Hupfer, P. (editors). 2001. *Climate of the 21st Century: Changes and Risks, Scientific Facts*, Hamburg: Wissenschaftliche Auswertungen

Manabe, S. & Wetherald, R.T. 1967. The effects of doubling CO2 concentration on the climate of a general circulation model. *Journal of the Atmospheric Sciences*, 32: 3–15

Pagani, M., Arthur, M.A. & Freeman, K.H. 1999. Miocene evolution of atmospheric carbon dioxide. *Paleoceanograohy*, 14: 273–292

Pearson, P.N. & Palmer, M.R.. 2000. Atmospheric carbon dioxide concentrations over the past 60 million years. *Nature*, 406: 695–699

Petit, J.R., Jouzel, J., Raynaud, D., Barkov, N.I., Barnola, J.-M., Basile, I., Bender, M., Chappellaz, J., Davis, M., Delaygue, G., Delmotte, M., Kotlyakov, V.M., Legrand, M., Lipenkov, V.Y., Lorius, C., Pepin, L., Ritz, C., Saltzman, E. & Stievenard, M. 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature*, 399: 429–426

Plass, G.N. 1956. The carbon dioxide theory of climatic change. *Tellus*, 8: 140–154

Raynaud, D., Jouzel, J., Barnola, J.M., Chappellaz, J., Delmas, R.J. & Lorius C. 1993. The ice record of greenhouse gases. *Science*, 259: 926–934

Revelle, R. & Suess, H.E. 1957. Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric CO_2 during the past decades. *Tellus*, 9: 18–27

Smil, V. 2002. *The Earth's Biosphere: Evolution, Dynamics, and Change*, Cambridge, MA: MIT Press

Tyndall, J. 1861. On the absorption and radiation of heat by gases and vapours, and on the physical connection of radiation, absorption, and conduction. *Philosophical Magazine and Journal of Science*, 22: 169–194, 273–285

Veizer, J., Godderis, Y. & François, L.M. 2000. Evidence for decoupling of atmospheric CO_2 and global climate during the Phanerozoic eon. *Nature*, 408: 698–701

Woodwell, G.M. & Mackenzie, F.T., eds. 1995. *Biotic Feedbacks in the Global Climatic System*, Oxford and New York: Oxford University Press

GOLDEN MEAN

See **Fibonacci series**

GRADIENT SYSTEM

In the study of dynamic systems, it is often observed that the rate of evolution of some system in its phase space is proportional to the gradient of a state function. A system of this type is called a *gradient system*, and the state function—which governs the course of its evolution—is its *potential*.

If the state s of a system is given by n state variables $s_1, \ldots, s_n$ and $G(s) = G(s_a, \ldots s,_n)$ is the potential of the system, then one can write (in matrix form)

$$\dot{s} = -\hat{k} \cdot \left(\frac{\partial G}{\partial s}\right)^{\mathrm{T}} . \qquad (1)$$

Here the rate of change of state s is given by the column vector $\dot{s} = |\dot{s_1}, \ldots, \dot{s_n}|^{\mathrm{T}}$, where the superscript T indicates transposition; $\partial G(s)/\partial s = |\partial G/\partial s_1, \ldots, \partial G/\partial s_n|$ is the row vector of gradient of potential $G(s)$; and $\hat{k}$ is a coefficient of proportionality, represented by the nonsingular matrix $\hat{k} = |k_{ij}|$, $\det |k_{ij}| \neq 0$. The minus sign in Equation (1) is chosen for convenience. The coordinate representation of this equation is

$$\dot{s_i} = -\sum_{j=1}^{n} k_{ij} \frac{\partial G}{\partial s_j} \quad (i = 1, \ldots, n) . \qquad (2)$$

Equations (1) and (2) are quite general. For example, if the order n is an even number ($n = 2m$, $m \geq 1$), and $\hat{k}$ is a $2m \times 2m$ block-diagonal matrix composed of the skew-symmetric blocks,

$$\hat{k} = \mathrm{diag}|\hat{k}_1, \ldots, \hat{k}_m| , \hat{k}_1 = \hat{k}_2 = \ldots = \hat{k}_m$$
$$= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad (3)$$

then Equations (1) and (2) are the usual Hamiltonian system (Fomenko, 1995).

However, gradient systems differ radically from Hamiltonian systems. Essential restrictions that specialize the definition of gradient systems and predetermine their qualitative properties are on the structure of matrix $\hat{k}$. For a gradient system: (i) The nonsingular matrix $\hat{k}$ is assumed to be symmetric: ($\hat{k} = \hat{k}^{\mathrm{T}}$ or $k_{ij} = k_{ji}$); hence, it possesses exactly n nonzero eigenvalues that the real numbers. (ii) All eigenvalues of $\hat{k}$ are assumed to have the same signs, either positive or negative.

Note that requiring that the rates $\dot{s}$ be proportional to the gradient of potential $G(s)$ does not oblige the coefficient of proportionality $\hat{k}$ to be a constant matrix. In general, the elements of $\hat{k}$ can (smoothly) depend on the current state $s : k_{ij} = k_{ji}(s)$, suggesting a third requirement of a gradient system: (iii) The properties (i) and (ii) must be fulfilled everywhere in the state space of system. Thus, a gradient system is defined as a system, whose evolution follows Equations (1) and (2), with a matrix satisfying conditions (i)–(iii).

The key dynamic difference between Hamiltonian and gradient systems is that Hamiltonian systems preserve the values of system potential, the Hamiltonian function ($\dot{G}(s) = 0$). The behavior of a gradient system is quite different because the matrix $\hat{k}$ in (3) is symmetric rather then skew-symmetric; thus $\dot{G}(s) \neq 0$.

From conditions (i)–(iii), one can demonstrate that the evolution of a gradient system preserves the sign of $\dot{G}(s)$; hence, the potential $G(s)$ either decreases or increases monotonically. To see this, consider the dissipative function of a gradient system, which is introduced by the quadratic form

$$\Gamma = \frac{1}{2}\dot{s}^{\mathrm{T}}\hat{\gamma}\dot{s} = \frac{1}{2}\sum_{i,j=1}^{n} \gamma_{ij}\dot{s}_i\dot{s}_j \quad \text{where} \quad \hat{\gamma} = \hat{k}^{-1}. \qquad (4)$$

The matrix $\hat{\gamma}$ inherits the properties (i)–(iii) of matrix $\hat{k}$, and the constancy of sign $\dot{G}(s)$ under the evolution of a gradient system follows directly from the relation

$$\dot{G}(s) = -2\Gamma. \qquad (5)$$

Relation (5) indicates that the sign of $\dot{G}(s)$ is opposite to the sign of dissipative function; thus, as the latter is positively (negatively) definite, the potential of gradient system strictly decreases (increases).

As an elementary example, the ordinary equation $\dot{u} = f(u)$ is a gradient system because it admits the representation $\dot{u} = -k[\mathrm{d}G(u)/\mathrm{d}u]$, $k = 1$, $G(u) = -\int f(u)\mathrm{d}u$. The corresponding dissipative function is $\Gamma = \dot{u}^2/2$, and under time-dependent solutions of this equation, the potential decreases monotonically.

As a second example, consider a Newtonian particle of mass m that is changing its position $\boldsymbol{r}$ with the speed $\dot{\boldsymbol{r}}$ and acceleration $\ddot{\boldsymbol{r}}$ by the action of forces of two kinds: a potential force $\boldsymbol{Q}^{(\mathrm{P})} = -\,\mathrm{grad}G(\boldsymbol{r})$ (where $G(\boldsymbol{r})$ is the potential energy), and a friction force $\boldsymbol{Q}^{(\mathrm{D})} = -\,\gamma\dot{\boldsymbol{r}}$, $\gamma = \mathrm{const} > 0$. The Newtonian vector equation of motion of this particle is $m\ddot{\boldsymbol{r}} = \boldsymbol{Q}^{(\mathrm{D})} + \boldsymbol{Q}^{(\mathrm{P})} \equiv -\gamma\dot{\boldsymbol{r}} - \mathrm{grad}G(\boldsymbol{r})$, or, equivalently,

$$\dot{\boldsymbol{r}} = \boldsymbol{v}, \quad m\dot{\boldsymbol{v}} + \gamma\boldsymbol{v} = -\mathrm{grad}G(\boldsymbol{r}). \tag{6}$$

This system has two different limits—"Galilean" and "Aristotelian." The Galilean limit corresponds to the situation when dissipative force $\boldsymbol{Q}^{(\mathrm{D})}$ is negligible in comparison with the d'Alembert inertia force $\boldsymbol{Q}^{(\mathrm{I})} = -\,m\dot{\boldsymbol{v}}$. At this limit—which is realized under motion through a vacuum—the term $\gamma\boldsymbol{v}$ in Equations (6) vanishes, and it reduces to the form

$$\dot{\boldsymbol{r}} = \boldsymbol{v}, \quad m\dot{\boldsymbol{v}} = -\mathrm{grad}G(\boldsymbol{r}). \tag{7}$$

We refer to this limit as Galilean because Galileo proposed and experimentally demonstrated that gravitational force determines the acceleration of a falling body rather then its speed.

The Aristotelian limit, on the other hand, describes so-called creeping motions, in which the d'Alembert inertia force $\boldsymbol{Q}^{(I)}$ is much weaker than the dissipative force $\boldsymbol{Q}^{(D)} : |\boldsymbol{Q}^{(D)}| = m|\dot{\boldsymbol{v}}| \ll |\boldsymbol{Q}^{(I)}| = \gamma|\boldsymbol{v}|$. This limit is realized in the viscosity limited motion of particle, which occurs in a strongly viscous medium in the presence of a potential force field. In this limit, Equations (6) asymptotically reduce to

$$\dot{\boldsymbol{r}} = -k\mathrm{grad}G(\boldsymbol{r}), \quad k = \gamma^{-1}. \tag{8}$$

The order of Equation (8) with respect to time is less by half then corresponding order of (6). In this case, the current state s of particle is fully determined by current value of radius-vector $\boldsymbol{r}$. Now, the rate of change of state is proportional to the gradient of potential energy; therefore, the Newtonian particle undergoes creeping motion in a gradient system. The dissipative function of this system is

$$\Gamma = \gamma\boldsymbol{v}^2/2 \quad (\gamma > 0), \tag{9}$$

and the potential $G(\boldsymbol{r})$ always decreases monotonically as $\dot{G} = -\gamma|\mathrm{grad}G(\boldsymbol{r})|^2$. This limit is called Aristotelian because it manifests the Aristotelian principle that "velocity of a body is proportional to a force acting upon the body." (This principle is not wrong; it merely corresponds to one of two possible limits of classical macroscopic dynamics which is realized in the presence of strong friction.)

The Newtonian example can be extended to the more general case of constrained system of n degrees of freedom, which move under stationary holonomic constraints and generalized forces of three types: the potential forces $Q_i^{(\mathrm{P})}$, the dissipative forces $Q_i^{(\mathrm{D})}$ of viscous friction, and the d'Alembert inertia forces ($Q_i^{(\mathrm{I})} = 1, \ldots, n$). A related example is provided by electrical networks that are constructed with linear lumped elements: resistances (R), inductances (L), and capacitances (C). From an electromechanical analogy (Gantmacher, 1975), the considerations discussed above are extended to this case almost automatically. The analogs of Galilean and Aristotelian limits also exist here. The first limit corresponds to the situation when all resistances are negligible; it displays the LC subclass of general RLC networks. The networks belonging to this subclass are described by systems of ordinary differential equations that involve second-order time derivatives; as a rule, they are not gradient systems. The electrical analog of the Aristotelian limit corresponds to the networks with negligible inductances (RC subclass of general RLC networks). The networks making up this subclass are gradient systems.

The gradient systems considered in these examples are characterized by the monotonic diminution of potential G in their motions. Formally, this property follows from positiveness of dissipative functions related to these systems. In the mechanical examples, where the potential G is given by the potential energy of a mechanical system, its diminution can be explained physically by the action of viscous friction, which dissipates the energy of system and fully converts it to heat Q at a rate $\mathrm{d}Q/\mathrm{d}t = 2\Gamma$.

The opposite case, which corresponds to a monotonic *increase* of G, can be viewed as the action of "negative friction" transferring energy from exterior sources to the moving system; for example, processes during evolution of the genetic structure of biological populations which are described by classical Fisher–Haldane–Wright equations. The gradient representations of these equations were developed by Svirezhev in 1972 and by Shahshahani in 1979. (Svirezhev's results now can be found in the comprehensive monograph of Svirezhev & Pasekov, 1990.) In the context of their results, the famous Fisher's Fundamental Theorem of Natural Selection—which asserts that the mean fitness of population always increases with the rate proportional to the genotypical diversity of the population—is a simple consequence of the gradient properties of the equations. In this case, the role of potential of the corresponding gradient system is played by the mean fitness of population, whereas the dissipative function is connected with a certain measure of genotypical diversity.

Owing to the presence of the monotonically varying quantity G, any gradient system is forbidden to return to states it once already left. In particular, such a system cannot perform nontrivial periodic motions (different from the states of rest). These properties are widely used in applied mathematics, including numerical gradient methods of searching for minima and maxima of multivariable functions.

In the above examples, gradient systems with finite degrees of freedom were considered. However, in various problems one deals with nonlocal spatially distributed systems in continuous media, occupying some d-dimensional region X of physical space, whose dynamics inherits the main features of finite-dimensional gradient systems. These objects form a class of gradient systems of an infinite number of degrees of freedom, which are referred to as *continuum gradient systems*. A nontrivial example is the system

$$\dot{u} = -k(u, \boldsymbol{x})\frac{\delta G[u]}{\delta u},$$

$$G[u] = \int_X g(u, \nabla u, \boldsymbol{x})\mathrm{d}X, \qquad (10)$$

where $u = u(\boldsymbol{x}, t)$ is a function describing the state of the system at the spatial point $\boldsymbol{x} \equiv (x_i, \dots, x_d) \in X$ and time t, and the integral functional $G[u]$ is a potential of the continual gradient system. Also ∇u denotes the spatial gradient, $\nabla u = |\partial u/\partial x_1, \dots, \partial u/\partial x_d|$; $\mathrm{d}X = \mathrm{d}x^1 \dots \mathrm{d}x^d$ is the space volume element of the region X occupied by the continuous medium; and $\delta/\delta u$ is a functional derivative, which acts upon the functional $G[u]$ according to the rule

$$\begin{aligned}\frac{\delta G}{\delta u} &= \frac{\partial g}{\partial u} - \operatorname{div}\left[\frac{\partial g}{\partial(\nabla u)}\right] \\ &= \frac{\partial g}{\partial u} - \sum_{\mu=1}^{d} \frac{\partial}{\partial x^\mu}\frac{\partial g}{\partial(\partial u/\partial x^\mu)}. \qquad (11)\end{aligned}$$

Substituting (11) into the second part of (10), one can show that equations (10) are equivalent to the equation

$$\dot{u} + \operatorname{div}\boldsymbol{J} = Q \qquad (12)$$

where $\boldsymbol{J}$ and Q are given by the expressions

$$\boldsymbol{J} = -k\frac{\partial g}{\partial(\nabla u)}, \qquad (13)$$

$$Q = -k\frac{\partial g}{\partial u} - \nabla k \cdot \frac{\partial g}{\partial(\nabla u)}. \qquad (14)$$

Equation (12) has the form of standard continuity equations for some substance filling the spatial region X; therefore, the quantities u and $\boldsymbol{J}$ can be interpreted as the density of this substance and as a vector of density of the corresponding spatial transport flux, respectively. From this standpoint, the function in the second part of continuity equation (12), defined by relationship (14), describes the rate of production of these substances per unit space volume, whereas expression (13) relating the flux $\boldsymbol{J}$ to the density u and its spatial gradient ∇u can be interpreted as a peculiar nonlinear generalization of well-known linear phenomenological laws such as Fick's law of diffusion or Ohm's law of electrical conduction. Hence, the continuum physical systems obeying the functional gradient equation (10) constitute a specific subclass in the class of reaction diffusion systems. This subclass comprises many physically important systems, including the parabolic equation

$$\begin{aligned}\gamma(\boldsymbol{x})\dot{u} &= \operatorname{div}[D(\boldsymbol{x})\nabla u] + f(u, \boldsymbol{x}) \\ &= D(\boldsymbol{x})\Delta u + \nabla D(\boldsymbol{x}) \cdot \nabla u + f(u, \boldsymbol{x}) \qquad (15)\end{aligned}$$

defined in the space region X. Here $\gamma(\boldsymbol{x}) > 0$, $D(\boldsymbol{x}) > 0$, and $f(u, \boldsymbol{x})$ are the given functions of their arguments. Special cases of (15) include the linear diffusion equation $\dot{u} = \Delta u$, as well as the nonlinear reaction diffusion equation

$$\dot{u} = \Delta u + f(u), \qquad (16)$$

which arises in several branches of natural science.

In mathematical genetics, Equation (16) describes the gene exchange waves traveling in the populations of biological organisms (Fisher's equation). In addition, this equation describes flame propagation as well as the switching processes in some physical, chemical, and biological nonlinear media (Zeldovich–Frank-Kamenetsky equation). Also, the time-dependent Ginzburg–Landau equation, which arises in physical kinetics and in synergetics for the phase transitions in spatially distributed self-organizing systems, has the form of Equation (16).

Remarkably, all these examples can be interpreted as processes of time evolution of certain gradient continuum systems. The functional gradient representation (10) is important because it suggests their general qualitative properties by the analogy with finite-dimensional case. For example, the potential $G[u]$ in (10) decreases monotonously in all time-dependent solutions if the latter is endowed with impermeability boundary conditions [see (18)]

$$J_n|_{\partial X} = 0 \leftrightarrow \boldsymbol{n}\,\frac{\partial g}{\partial(\nabla u)}\bigg|_{\partial X} = 0, \qquad (17)$$

where the subscript ∂X indicates that corresponding expressions are considered at the boundary ∂X of space region X enclosing the continual system, $J_n|_{\partial X}$ is the normal component of $\boldsymbol{J}$ at the boundary, and $\boldsymbol{n}$ is unit normal vector at ∂X. Indeed, the full derivative of the energy functional $G[u]$ with respect to time t is given by the elegant relation

$$\frac{\mathrm{d}G}{\mathrm{d}t} = -\int_X k\left(\frac{\delta G}{\delta u}\right)^2 \mathrm{d}X \leq 0\,. \tag{18}$$

Defining the *dissipative functional* by the formula

$$\Gamma[u, \dot{u}] = \int_X \frac{1}{2k}\dot{u}^2 \mathrm{d}X = \frac{1}{2}\int_X \gamma\dot{u}^2 \mathrm{d}X,$$
$$\gamma = k^{-1},$$

puts (18) into the simple form of Equation (5), which appeared in the finite-dimensional case.

As in the finite-dimensional case, the monotone decrease of potential means that time evolution is *unidirectional*—these systems cannot return to the states once left. In particular, they can have neither solutions that are periodic in space and time nor traveling-impulse solutions with complete recovery of the initial state. If they possess moving wave-front solutions, contra directional wave fronts cannot be reflected after collisions with each other. On the contrary, if some spatially distributed system isolated from the outer world by the impermeable boundaries can modify its state supporting the propagation of solitary/periodic traveling pulses, then it cannot be described by continuum gradient equations.

O.A. MORNEV

See also **Diffusion; Flame front; Nerve impulses; Reaction-diffusion systems; Synergetics; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Fomenko, A.T. 1995. *Symplectic Geometry: Methods and Applications*, 2nd edition, New York: Gordon and Breach (original Russian edition 1988)

Gantmacher, F. 1975. *Lectures in Analytical Mechanics*, Moscow: Mir Publications (original Russian edition 1966)

Loskutov, A.Yu. & Mikhailov, A.S. 1996. *Foundations of Synergetics*, 2nd edition, Berlin and New York: Springer (original Russian edition 1990)

Mornev, O.A. 1997. Dynamical principle of minimum energy dissipation for systems with ideal constraints and viscous friction. *Russian Journal of Physical Chemistry*, 71(12): 2077–2081 [Translated from *Zhurnal Fizicheskoi Khimii*, 1997, 71(12): 2293–2298]

Mornev, O.A. 1998. Modification of the biot method on the basis of the principle of minimum dissipation (with an application to the problem of propagation of nonlinear concentration waves in an autocatalytic medium). *Russian Journal of Physical Chemistry*, 72(1): 112–118 [Translated from *Zhurnal Fizicheskoi Khimii*, 1998, 72(1): 124–131]

Mornev, O.A. & Aliev, R.R. 1995. Local variational principle of minimum dissipation in the dynamics of reaction-diffusion systems. *Russian Journal of Physical Chemistry*, 69(8): 1325–1328 [Translated from *Zhurnal Fizicheskoi Khimii*, 1995, 69(8): 1466–1469]

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Svirezhev, Yu.M. & Pasekov, V.P. 1990. *Fundamentals of Mathematical Evolutionary Genetics*, Dordrecht: Kluwer (original Russian edition 1982)

GRANULAR MATERIALS

What do coffee powder, wheat, mustard seeds, granulated sugar, cement, sand, and rocks have in common? They are granular materials, assemblies of solid objects, from tens of micrometers to meter sized, that are generally not bound together by significant attractive forces. Such assemblies of objects possess unique physical and dynamic characteristics. Even though individual grains are solid particles, the assembly of grains behaves distinctly from ordinary solids, fluids, or gases (Jeager et al., 1996; Duran, 1997). For example, a sand dune is reliably solid-like and can sustain the weight of a person. However, if the sand from the dune is placed in an hourglass, it will flow rapidly and at a predictable rate.

The curious behavior of granular matter has long attracted the attention of scientists. In 1885, Osborne Reynolds observed that in order for grains to flow past each other, they have to move out of each other's way, leading to a dilation of granular matter under shear (Reynolds, 1885).

Some of the first modern work on granular matter was inspired by Per Bak, C. Tang, and K. Wiesenfeld (1987), who developed the concept that many systems may self-organize into a critical state. Granular avalanching was thought to be one of the prominent experimental realizations of their theory of self-organized criticality (SOC). The idea was that on a sandpile, grains might self-organize through avalanching to form a heap with a critical angle, which is called the angle of repose of the sandpile.

While experiments showed that SOC does not describe the behavior of real sandpiles (Nagel, 1992), the initial experiments showed some of the puzzling properties of granular matter and inspired a significant resurgence of research into granular matter. Here we describe three of the main questions of recent work:

When and How Does Granular Matter Flow?

In most situations when granular matter flows, such as during emptying of silos or during natural rock avalanches, only part of the material behaves like a fluid. This is illustrated in Figure 1 in a long exposure image of the side of an avalanche flowing down the

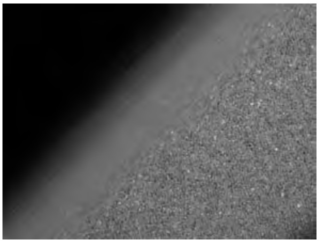

Figure 1. Long exposure image of the side of an avalanche flowing down the side of a granular pile. A clear separation into fluid-like and solid-like regions is visible. (Courtesy of N. Taberlet and P. Richard, University of Rennes).

side of a granular pile. The flow leaves behind a pile with a known surface angle, the angle of repose (Nagel, 1992). Predicting the timing and extent of such partial fluidization is part of the current challenge of modeling dense granular flows.

Once flowing, hydrodynamic equations may describe the flow behavior (Losert et al., 2000), though velocity gradients on length scales of a few particle diameters can call the validity of continuum models with local equilibria into question.

More dilute flows can be modeled as a system of hard spheres in which some energy is lost during each collision. Based on such a picture, kinetic theories of driven granular media have been developed, for example by Goldhirsch & Zanetti (1993). Since energy is continuously lost in collisions and must be added through gravity or shaking, granular flows are non-equilibrium driven, dissipative systems that exhibit interesting instabilities. One example is the clustering instability, in which a dilute system of particles can spontaneously develop localized, dense clusters (Goldhirsch & Zanetti, 1993). The basic mechanism for and clustering is that a local increase in particle density increases the number of collisions in that region, and, thus, slowing particles down and trapping them in the dense region.

How Can Different Kinds of Particles be Mixed or Separated?

During flow, mixtures of grains tend to spontaneously unmix (Shinbrot & Muzzio, 2000), as illustrated in Figure 2. The figure shows unmixing in a mixture of glass particles of three different sizes after several rotations in a half-filled horizontal cylinder. The phenomenology of granular unmixing is complex. In a horizontally rotating drum, for example, materials can segregate both radially and axially by size or by density. There is, to date, no simple model for unmixing, but a wealth of empirical experimental data and several simple physical mechanisms, such as percolation through voids, shear flows, and convection, all play some role in the unmixing process.

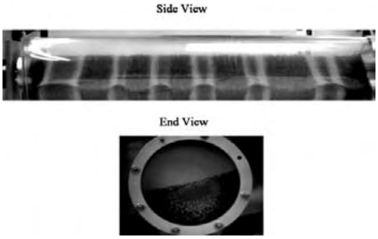

Figure 2. Unmixing of grains of three different sizes in a rotating horizontal cylinder. After 5 min of rotation at 15 rpm, material is separated into axial and radial bands. Both a side view and end view of the cylinder are shown.

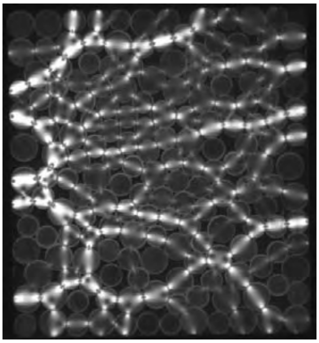

Figure 3. Forces in a 2-dimensional assembly of disks sheared between concentric cylinders. The disks are birefringent and placed between crossed polarizers, so that only points subject to large forces are visible as bright regions. (Courtesy of B. Utter and R.P. Behringer, Duke University).

How Are Forces Transmitted Through Granular Matter?

In civil engineering, the mechanical behavior of granular matter is modeled by solid-like equations (Nedderman, 1992). Zooming in to the scale of grains, forces can only be transmitted at particle contacts, and only repulsive forces are permitted at each contact. This can lead to very inhomogeneous force distributions and locally very large forces as is indeed seen, for example, in grain silos. Experiments with birefringent disks that highlight the location and magnitude of contact forces (Howell et al., 1999) have helped in our microscopic understanding of the way forces are transmitted through granular matter (Figure 3). Strong inhomogeneities in the force magnitude, anisotropies in the direction of stress transmission, and large rearrangements in the force distribution, even for small changes in structure are observed.

Two emerging concepts to explain the properties of such dense granular matter are jamming, the property of a system to get trapped in some intermediate state which may be generic to granular matter as well as thermal systems such as glasses, and the idea of stress chains, lines of particles that carry disproportionately large stress, as can be seen in Figure 3.

To conclude, basic questions about granular matter remain. Granular matter encompasses different kinds of particles with many variables, such as the frictional properties of grain-grain or grain-boundary contacts, the deformability, surface roughness, polydispersity, and grain shape. It remains difficult to predict whether a particular granular parameter, such as the shape of individual grains will qualitatively alter flow, forces, or mixing behavior, and thus, has to be taken into account in modeling. While models for flow, forces, and segregation have been developed that agree with experimental data under various conditions, a broadly valid model of granular flow based on simple physical insights similar to the Navier–Stokes equation for fluids has so far proven elusive. Similarly, no consensus about the most suitable equation for granular solids has yet emerged, nor has a clear separation between the solid-like and liquid-like regime emerged.

WOLFGANG LOSERT

See also **Avalanches; Cluster coagulation; Dune formation; Sandpile model**

Further Reading

Bak, P., Tang, C. & Wiesenfeld, K. 1987. Self-organized criticality: an explanation of the 1/f noise. *Physical Review* A, 59: 381–384

Duran, J. 1997. *Sands, Powders, and Grains: An Introduction to the Physics of Granular Materials*, Berlin and New York: Springer

Goldhirsch, I. & Zanetti, I. 1993. Clustering instability in dissipative gases. *Physical Review Letters*, 70: 1619–1622

Howell, D.W., Veje, C.T. & Behringer, R.P. 1999. Stress fluctuations in a 2D granular couette experiment: a critical transition. *Physical Review Letters*, 82: 5241–5244

Jaeger, H.M., Nagel, S.R. & Behringer, R.P. 1996. *Reviews of Modern Physics*, 68: 1259–1273

Knight, J.B., Jaeger, H.M. & Nagel, S.R. 1993. *Physical Review Letters*, 70: 3728–3731

Losert, W., Bocquet, L., Lubensky, T.C. & Gollub, J.P. 2000. Particle dynamics in sheared granular matter. *Physical Review Letters*, 85: 1428–1431

Nagel, S.R. 1992. Instabilities in a sandpile. *Reviews of Modern Physics*, 64: 321—325

Nedderman, R.M. 1992. *Statics and Kinematics of Granular Materials*, Cambridge and New York: Cambridge University Press

Reynolds, O. 1885. On the dilatancy of media composed of rigid particles in contact. *Philosophical Magazine*, 20: 469

Shinbrot, T. & Muzzio, F.J. 2000. *Physics Today*, March 25

GRAVITATIONAL WAVES

In 1687, Newton published *The Philosophiae Naturalis Principia Mathematica* in which he first proposed his law of gravitation. According to this law, the gravitational force of attraction between two bodies is always proportional to their masses and inversely proportional to the square of their distance apart, and it acts instantaneously through infinite distance. Less than two and half centuries later, Newton's theory was radically revised.

According to Einstein's general theory of relativity (published in 1913), gravitation is not a force of attraction but rather the force required to prevent the natural motion of matter, which is to follow a geodesic in space time. The geodesic or shortest path in four-dimensional space time describes a free-fall trajectory such as the motion of a planet around the sun. The space time is curved by the presence of matter or energy. The predictions of general relativity are very close to those of Newtonian theory as long as gravity is weak and velocities are slow, but they diverge in strong gravity due to two aspects of the theory. The first is the change in geometry due to the non-Euclidean properties of space time. The second stems from the nonlinear aspects of general relativity, which arise because gravitational energy is itself a source of curvature.

Eight years prior to publishing his general relativity theory, Einstein published the special theory of relativity, which predicted that neither matter nor information could ever travel faster than the speed of light. This means that the curvature at a point generated by a mass at another point only achieves its value at a retarded time, a time set by the travel time for light between the two points.

The general theory of relativity changed our vision of a Euclidean flat space that had been assumed since the days of Newton. In non-Euclidean space, the sum of the angles of a triangle does not equal 180° and the area of a circle is not always πr^2. Space time can be considered as an elastic membrane. The deformations are described by the Einstein curvature tensor G, while the sources of curvature (i.e., the mass-energy distribution) are described by the stress-energy tensor T. Einstein's field equations are then expressed by the equation (Misner et al., 1973)

$$\mathsf{T} = \frac{c^4}{8\pi G}\mathsf{G}. \qquad (1)$$

The constant $c^4/8\pi G$ (where c is the speed of light and G is Newton's gravitational constant) is a very large number which can be considered as the "spring constant" of space time. Because it is so large, only very small curvatures are generated even by very large values of mass-energy distribution. A natural consequence of the membrane analogy is the existence of waves; ripples in the curvature of the membrane which propagate through the membrane. This concept can also be deduced from Einstein's field equations.

Equation (1) is a set of ten nonlinear equations. Except in simplified situations, these equations are difficult to solve directly. Matter creates curvature and curvature influences the motion of matter. This mutual influence gives rise to nonlinear phenomena in gravitational wave propagation. Unlike the theory of electromagnetism, the theory of general relativity is intrinsically nonlinear. Since gravity is itself a source of curvature, there can be a gravitational interaction between gravitational waves. While gravitational wave signals are normally expected to be very small (and hence amenable to a linearized theory), at their sources, the nonlinear aspect of the theory makes prediction extremely difficult. The following list offers some examples of expected nonlinear phenomena in gravitational wave propagation:

- For gravitational waves emerging from the birth of a black hole, the gravitational redshift of the waves reduces the total emitted energy. The mass-energy of the gravitational waves themselves constitute the redshift.
- Like electromagnetic radiation, gravitational waves can undergo gravitational lensing. A mass in the path of the waves creates a background curvature. This curvature focuses the waves by modifying the wave front. This effect can enhance the intensity of dim sources. The gravitational lensing effect has successfully been observed for electromagnetic radiation.
- A time-varying curvature can amplify waves. If a wave modulates the space curvature, a second incident wave will be amplified or its frequency shifted. An optical parametric oscillator represents the equivalent phenomena for optical radiation. Similar to the optical Kerr effect, a gravitational wave can interact with its own self-generated background curvature. In the early universe, it is predicted that gravitational waves from the Big Bang may have been parametrically amplified by the action of inflation.
- A strong background curvature (e.g., from a black hole) can scatter a gravitational wave passing in its vicinity.

Under gravitational wave astronomy, it may be possible to find evidence of these predicted phenomena. The first indirect experimental proof of gravitational waves was provided in 1984 by Weisberg and Taylor. By studying the pulsar PSR 1913+16, they showed that the period of the pulsar around its companion star decreased exactly as predicted by the Einstein equations (Weisberg et al., 1981). A part of the pulsar orbital energy is converted to gravitational radiation.

The orbital parameters of binary stars are usually deduced from measurements of the Doppler shift of the radiated electromagnetic waves. According to Newtonian gravitational theory, the mass m_1 of the pulsar and the mass m_2 of its companion cannot be determined with confidence. All results obtained are proportional to an unknown parameter: the sine (normally denoted $\sin i$) of the angle between the orbital plane and the line of sight. This parameter can be evaluated in the framework of general relativity. From this theory, five independent parameters (among them, the advance of the periastron (the point of closet approach of the two stars), the evolution of the period, or the Einstein parameter) which are functions of m_1, m_2, and $\sin i$ can be measured. The overdetermined system of five equations determines with an unprecedented accuracy (error less than 0.5%) the mass of the pulsar. Moreover, the compatible results from the five independent equations confirm the general theory of relativity in the strong field and radiative regime. This high-precision validation of general relativity indirectly implies that the velocity of gravitational waves is equal to the speed of light.

Direct detection of gravitational waves on Earth is one of the most exciting challenges of today's science. In the 1960s, Joseph Weber invented and developed the first gravitational wave detectors, consisting of large (about a ton) vibration-isolated cylinders of aluminium or niobium (Weber, 1960). The gravitational waves excite the longitudinal resonance in the cylinder, which behaves as an extremely low-loss mechanical oscillator. The motion of the cylinder is monitored to very high precision. Most of its motion is thermal vibration but the small effect of the gravitational waves appear as very small perturbations in the amplitude or phase of the vibration. The mechanical oscillations, converted into an electrical signal, are recorded and analyzed by sophisticated predictive filter algorithms. A worldwide network of five resonant bars still uses Weber's technique (see Figure 1).

A new generation of gravity-wave detectors is based on laser interferometry (Blair, 1991). Gravitational waves passing through a Michelson interferometer change the relative length difference between the two perpendicular arms of the interferometer. Thus the variation of the gravitational field can be converted to an optical phase variation. Due to the weakness of gravitational wave interaction with matter, the detectors must be able to measure a length variation of less than 10^{-18} m, which is close to the quantum limit. Compared with the resonant bar, the interferometric devices exhibit a larger bandwidth and higher sensitivity. The sensitivity is limited by seismic and thermal noise at low frequency and photon-counting noise at higher frequency.

Five international projects, similar to the example shown in Figure 2, are using long base interferometry

Figure 1. Photograph of the interior of Niobe, the resonant bar at the University of Western Australia. The bar is a one and half ton cylinder of niobium cooled to 5 K to limit the thermal noise.

Figure 2. Aerial view of the French–Italian gravitational waves detector VIRGO near Pisa in Italy. On this photo, the two perpendicular 3 km vacuum pipes containing the arms of the interferometer are clearly distinct (Bradaschia et al., 1990) (Reproduced with permission from EGO).

to detect gravitational waves. The network of, interferometers (situated in the United States, Europe, Japan, and Australia) will enable the localization of sources of gravitational radiation in space as well as determination of their polarization. By measuring the arrival time of the waves at different detectors, the speed of the gravity waves can be calculated. This speed could be influenced by two independent factors: a coupling of the wave with a strong curvature background (difficult to detect on Earth) and the possibility that the graviton, the particle associated with the gravitational waves, is not massless (Will, 1998).

The direct detection of gravitational waves is expected to occur before 2012. This gravitational spectrum will open a new window to studying the universe, comparable to the revolution fathered by the invention of radio astronomy in the 1950s. The direct mapping of the gravity spectrum will broaden our knowledge about the universe, from the earliest moment of the Big Bang to the end of stars.

JEROME DEGALLAIX AND DAVID BLAIR

See also **Einstein equations; General relativity**

Further Reading

Blair, D.G. 1991. *The Detection of Gravitational Waves*, Cambridge and New York: Cambridge University Press

Bradaschia, C. et al. 1990. The VIRGO project: a wide band antenna for gravitational wave detections. *Nuclear Instruments and Methods in Physics Research* A, 289: 518–528

Misner, C.W., Thorne, K.S. & Wheeler, J.A. 1973. *Gravitation*, San Francisco: W.H. Freemann

Weber, J. 1960. Detection and generation of gravitational waves. *Physical Review*, 117: 306–313

Weisberg, J.M., Taylor, J.H. & Fowler, L.A. 1981. *Gravitational Waves from an Orbiting Pulsar*, Scientific American, p. 74

Will, C.M. 1998. Bounding the mass of the graviton using gravitational waves observations of inspiralling compact binaries. *Physical Review* D, 57: 2061–2068

GRAVITY WAVES

See **Water waves**

GREEN'S FUNCTION

See **Boundary layer problems**

GREY SOLITON

See **Solitons, types of**

GROSS-PITAEVSKII EQUATION

See **Nonlinear Schrödinger equations**

GROUP VELOCITY

Waves often propagate as a packet, or group, within which there are several wave crests. A common illustration is given by the wave pattern formed when a stone is thrown into a pond. An axially symmetric ring of waves on the water surface propagates outwards as a wave group; within the group, however, one can see that the wave crests propagate through the group, apparently from the rear to the front. The speed of the wave crests, called the *phase velocity*, is different from that of the group as a whole, this speed being called the *group velocity*. For water waves, the phase velocity is greater than the group velocity, except for the very short waves dominated by surface tension.

This important distinction between phase and group velocity arises in all physical systems which support waves. The concept of group velocity appears to have been first formulated by William Hamilton in 1839 (quoted by Havelock, 1914). The first recorded observation of the group velocity of a (water) wave is due to John Scott Russell in 1844 (Russell, 1844) (note also his remark that "the sound of a cannon travels faster than the command to fire it" (Russell, 1885)). However, our present understanding of group velocity is usually attributed to George Stokes, who used it as the topic of a Smith's Prize examination paper in 1876.

For simplicity, consider a linearized system, with a single spatial variable x and time represented by t. Then a sinusoidal wave has the representation,

$$u(x,t) = a\cos(kx - \omega t + \phi). \quad (1)$$

Here k is the wave number, ω is the wave frequency, a is the constant amplitude, and ϕ is a constant phase. Thus, this sinusoidal wave has a wavelength $\lambda = 2\pi/k$ and a wave period $T = 2\pi/\omega$. Equation (1) is a kinematic expression, valid for all physical systems which support waves. The dynamics of the system are governed by the dispersion relation

$$\omega = \omega(k), \quad (2)$$

defining the frequency as a function of wave number. The phase velocity $c = \omega/k$ is likewise a function of wavenumber.

To obtain the group velocity, consider an argument first advanced by Stokes (1876) and later in more general form by Lord Rayleigh (1881). Form the linear superposition of two sinusoidal waves, each of the form (1) with frequencies $\omega \pm \Omega$, wave numbers $k \pm K$, and with equal amplitudes a and phases ϕ which can be written as

$$u(x,t) = 2a\cos(kx - \omega t\phi)\cos(Kx - \Omega t). \quad (3)$$

This expression can be viewed as a wave packet, or group, with a dominant wave number k and frequency ω, and with a group velocity of Ω/K. Taking the limit $K \to 0$, one sees that the group velocity is given by

$$c_g = \frac{d\omega}{dk}, \quad (4)$$

which can be obtained by differentiation of the dispersion relation (2).

A more general argument that links the origin of a wave packet to initial conditions uses Fourier superposition to represent the solution of an initial-value problem in the form

$$u(x,t) = \int_{-\infty}^{\infty} F(k)\exp(i(kx - \omega t))\,dk + \text{c.c.}, \quad (5)$$

where $F(k)$ is the Fourier transform of $u(x,0)$ and c.c. stands for complex conjugate. Thus, the initial conditions determine the Fourier transform, each component of which evolves independently with frequency ω related to the wave number k through the dispersion relation (2). To obtain a wave packet, we suppose that the initial conditions are such that $F(k)$ has a dominant component centered at $k = k_0$. The dispersion relation is then approximated by

$$\omega = \omega_0 + b_1(k - k_0) + b_2(k - k_0)^2, \quad (6)$$

where

$$b_1 = \frac{d\omega}{dk} = c_g, \quad \text{and} \quad b_2 = \frac{1}{2}\frac{d^2\omega}{dk^2}. \quad (7)$$

Here, both b_1, b_2 are evaluated at $k = k_0$. Expression (5) then becomes

$$u(x,t) \approx A(x,t)\exp(i(k_0x - \omega_0 t)) + \text{c.c.}, \quad (8)$$

where

$$A(x,t) = \int_{-\infty}^{\infty} F(k_0 + \kappa)\exp[i(\kappa(x - c_g t)) - ib_2\kappa^2 t]\,d\kappa, \quad (9)$$

where the variable of integration has been changed from k to $\kappa = k - k_0$. Here, the sinusoidal factor $\exp(i(k_0x - \omega_0 t))$ is a carrier wave with a phase velocity ω_0/k_0, while the (complex) amplitude $A(x,t)$ describes the wave packet. Since the term proportional to κ^2 in the exponent in (9) is a small correction term, it can be seen that to the leading order, the amplitude A propagates with the group velocity c_g, while the aforementioned small correction term gives a dispersive correction term proportional to $t^{-1/2}$. Indeed, it can be shown that as $t \to \infty$

$$A(x,t)|_{x=c_g t} \sim F(k_0)\left(\frac{2\pi}{|b_2 t|}\right)^{1/2}\exp\left(\frac{i\pi}{4}\operatorname{sign} b_2\right). \quad (10)$$

This same result can be obtained directly from (5) by using the method of stationary phase, valid here in the limit when $t \to \infty$ (see, for instance, Lighthill, 1978); In this case, it is not necessary to also assume that $F(k)$ is centered at k_0, and so wave packets are the generic long-time outcome of initial-value problems.

It is useful to note that the group velocity appears naturally in the kinematic theory of waves (Lighthill, 1978; Whitham, 1974). Thus, let the wave field be defined asymptotically by

$$u(x,t) \sim A(x,t)\exp(i\theta(x,t)) + \text{c.c.}, \quad (11)$$

where $A(x,t)$ is the (complex) wave amplitude, and $\theta(x,t)$ is the phase, which is assumed to be rapidly varying compared with the amplitude. Then it is natural to define the local wave frequency and wave number by

$$\omega = -\frac{\partial\theta}{\partial t}, \qquad k = \frac{\partial\theta}{\partial x}. \tag{12}$$

Note that expression (10) has the required form (11). Then cross-differentiation leads to the kinematic equation for the conservation of waves,

$$\frac{\partial k}{\partial t} + \frac{\partial\omega}{\partial x} = 0. \tag{13}$$

But now, if we suppose the dispersion relation (2) holds for the frequency and wave number defined by (12), then we readily obtain

$$\frac{\partial k}{\partial t} + c_{\mathrm{g}}\frac{\partial k}{\partial x} = 0 \tag{14}$$

with a similar equation for the frequency. Thus, both the wave number and frequency propagate with the group velocity, a fact that can also be seen in (10). Equation (14) is itself a simple wave equation, which can be readily integrated by the method of characteristics, or rays. It is important to note that the group velocity c_{g} is a function of k, so that (14) is a nonlinear equation.

Next, suppose that the physical system contains several spatial variables, represented by the vector $\boldsymbol{x}$. Then the phase variable $(kx - \omega t)$ in (1, 5) is replaced by $(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)$, where $\boldsymbol{k}$ is the vector wave number. The dispersion relation (2) is replaced by

$$\omega = \omega(\boldsymbol{k}). \tag{15}$$

The phase velocity is now a vector ($\boldsymbol{c}$), has a magnitude $\omega/|\boldsymbol{k}|$, and is in the direction of $\boldsymbol{k}$, and the definition of group velocity, replacing (4), is

$$\boldsymbol{c}_{\mathrm{g}} = \nabla_{\boldsymbol{k}}\omega. \tag{16}$$

Thus, the group velocity can differ from the phase velocity in both magnitude and direction. A striking example of the latter arises for internal waves, whose group velocity is perpendicular to the phase velocity (see Lighthill, 1978).

Since the group velocity is the velocity of the wave packet as a whole, it is no surprise to find that it can also be identified with energy propagation. Indeed, it can be shown that in most linearized physical systems, an equation of the following form can be derived:

$$\frac{\partial\boldsymbol{E}}{\partial t} + \nabla\cdot(\boldsymbol{c}_{\mathrm{g}}\boldsymbol{E}) = 0\,, \tag{17}$$

where $\boldsymbol{E}$ is the wave action density, and is proportional to the square of the wave amplitude $|A|^2$, with the factor being a function of the wavenumber $\boldsymbol{k}$. Typically, the wave action is just the wave energy density divided by the frequency, at least in inertial frames of reference.

So far the discussion has remained within the realm of linearized theory, and it remains to mention the consequences of nonlinearity. Analogous definitions and concepts can be developed, and at least for weakly nonlinear waves, the main outcome is that a dependence on wave amplitude (more strictly, on $|A|^2$) appears in the dispersion relation. This has the consequence that the phase and group velocities both inherit a weak dependence on the wave amplitude. It can be shown that at least within the confines of weakly nonlinear theory, in the case of just a single active spatial dimension, the complex wave amplitude A is governed by the nonlinear Schrödinger (NLS) equation,

$$\mathrm{i}\left(\frac{\partial A}{\partial t} + c_{\mathrm{g}}\frac{\partial A}{\partial x}\right) + b_2\frac{\partial^2 A}{\partial x^2} + \mu|A|^2 A = 0. \tag{18}$$

Here recall that b_2 is defined in (7), and μ is a nonlinear coefficient which is system-dependent.

If A is assumed to depend only on t then Equation (18) has the plane wave solution

$$A = A_0 \exp\left(\mathrm{i}\mu|A_0|^2 t\right), \tag{19}$$

which shows that the nonlinear coefficient μ has the physical interpretation that $-\mu|A_0|^2$ is the nonlinear correction to the wave frequency. It can be shown that this plane wave is stable (unstable) according as $\mu b_2 < (>) 0$ (see Whitham, 1974). In the unstable case, $\mu b_2 > 0$, a perturbed plane wave will evolve into one or more solitons, where the soliton family is given by (see Zakharov and Shabat, 1972),

$$A(x,t) = \alpha\exp\left(-\mathrm{i}\mu\alpha^2 t\right)\mathrm{sech}(\gamma\alpha(x - c_{\mathrm{g}}t - x_0)),$$

$$\text{where}\quad \gamma^2 = \frac{\mu}{b_2}. \tag{20}$$

Here, x_0 is a phase constant determining the location of the soliton at $t = 0$, and the amplitude α is a free parameter. In linear theory, the wave packet profile is determined by the initial conditions, whereas the influence of even weak nonlinearity, when it is balanced by weak dispersion, results in the generic sech profile.

Roger Grimshaw

See also **Modulated waves; Nonlinear Schrödinger equations; Wave packets, linear and nonlinear**

Further Reading

Havelock, T.H. 1914. *The Propagation of Disturbances in Dispersive Media*, Cambridge: Cambridge University Press

Lighthill, M.J. 1978. *Waves in Fluids*, Cambridge: Cambridge University Press

Rayleigh, Lord, 1881. On the velocity of light. *Nature*, XXIV: 382

Russell, J.S. 1844. *Report on Waves*, 14th meeting of the British Association for the Advancement of Science, London: BAAS, pp. 311–339

Russell, J.S. 1885. *The Wave of Translation in the Oceans of Water, Air and Ether*, London: Trübner

Stokes, G.G. 1876. Problem 11 of the Smith's Prize examination papers (February 2, 1876). In *Mathematical and Physical Papers of George Gabriel Stokes*, vol. 5, New York: Johnson Reprint Co., 1966, p. 362

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

Zakharov, V.E. & Shabat, A.B. 1972. Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Soviet Physics, JETP*, 34: 62–69

GROWTH PATTERNS

Patterns are ubiquitous in nature. They can range from the broad stripes on a zebra, to the intricate design of a snowflake, to the rib patterns on a saguaro cactus. The idea that naturally occurring patterns might have some underlying and universal governing mathematical structure has long been a source of fascination to researchers. The study of biological patterns was pioneered by D'Arcy Thompson in his landmark treatise *On Growth and Form* (Thompson, 1942). This was followed by a notable contribution by the mathematician (and Enigma code breaker) Alan Turing in a now famous paper "The Chemical Basis of Morphogenesis" (Turing, 1952). A more contemporary account of many biological patterns and their mathematical modeling is given in James D. Murray's textbook *Mathematical Biology* (Murray, 2003). But to what extent are patterns universal? Are there any connections among the three examples cited above? Inevitably, the underlying physical and biological processes governing a particular pattern-forming process can be very different, even if the topology of the resultant patterns and certain "defects" in those patterns look very similar. In some cases, different governing physical processes end up having essentially the same mathematical formulation, and in these cases, commonality of pattern structure is less surprising.

Here, we will concentrate on two physical pattern-forming systems that are popular topics of contemporary research in condensed-matter physics, namely, the fingering patterns at the interface of fluid mixtures and formation of dendritic crystals in solidification. Some of the patterns seen in these two systems are also seen in growing bacterial colonies.

The Hele-Shaw Cell: Fluid Fingering

This classic problem was first studied at the end of the 19th century, but even to this day there are still unresolved issues relating to the stability of the patterns exhibited. The experimental setup is relatively simple, consisting merely of a thin film of fluid sandwiched between two glass plates into which air (or another fluid) is pumped. As the air is pumped into the fluid film, one might imagine that it simply forms an expanding circular planar bubble. Initially this appears to be the case, but the circular perimeter rapidly becomes wavy and these "waves" then grow into long finger-like structures which, in turn can split or side branch forming highly complicated patterns. A typical pattern is shown in Figure 1a. The physical principles governing the growing air-fluid interface are well understood. The key idea is that the normal velocity of the air-fluid boundary (i.e., the velocity in the direction normal to each point of the interface) is governed by the gradient of the pressure across the interface. Because the fluid layer is so thin, the equations of fluid mechanics are greatly simplified, and one ends up with governing equations of the form

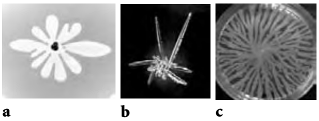

Figure 1. Three different types of growth patterns: (a) Hele-Shaw fingers; (b) dendritic crystals; (c) *Bacillus substilis* bacterial colonies.

$$\nabla^2 p = 0, \tag{1}$$

$$u_{\mathrm{n}} = -\frac{b^2}{12\mu}(\boldsymbol{\nabla} p)\cdot \boldsymbol{n}, \tag{2}$$

$$\Delta p|_{\partial\Omega} = \sigma\kappa. \tag{3}$$

The first equation, in which $p = p(x, y)$ denotes the pressure, is a statement of the incompressibility of the fluid. In the second equation, u_{n} denotes the normal velocity (in the normal direction $\boldsymbol{n}$) of the interface. This is determined by the pressure gradient, the thickness of the fluid film b, and the fluid viscosity μ. The third equation describes the pressure jump Δp across the air-fluid interface $\partial\Omega$, and this depends on the surface tension σ and the interface curvature κ (the Young–Laplace law). Unlike more standard systems of partial differential equations in which the equation and boundary conditions are specified, the solution to this problem involves finding the boundary, $\partial\Omega$, itself. This is an example of a Stefan problem. At first sight the equations might look linear, but the coupling to the boundary dynamics makes them effectively highly nonlinear. Finding an exact solution of these equations is very difficult. A standard "linear stability" analysis is relatively straightforward and can explain why the initial circular air-fluid interface becomes wavy. But beyond this, finding solutions for the finger-like structures, let alone fingers whose tips split, is (still) a difficult problem. A pioneering early study of the Hele-Shaw cell is due to Philip Saffman and Geoffrey Taylor (1958), and many hundreds of papers have been published on the problem in subsequent years.

Solidification and Dendritic Crystal Growth

At first sight this seems like a very different process. Cooling a drop of a molten substance (or a supersaturated salt solution) can produce very striking needle crystals that grow with approximately constant velocity and develop side branches. A typical pattern is shown in Figure 1b. Many questions come to mind: why do finger-like structures form; what determines their propagation speed; why do they form side branches; and is there any connection with the finger-like structures seen in the Hele-Shaw cell? Before addressing these questions it is worth noting that historically the study of dendritic growth had its origins in metallurgical studies associated with the casting of canons. It was discovered that during the casting process the solidifying metal developed a crystalline structure that could weaken the canon—to the point of its shattering when it was fired!

The physical process determining the growth of the solid phase is that of cooling, with the rate of cooling being determined by the rate at which heat can diffuse away from the solid-melt interface. The speed at which the the solid front advances is determined by the temperature gradient across the interface. Using these basic principles, the governing equations are found to be

$$\frac{\partial T}{\partial t} = \nabla^2 T, \tag{4}$$

$$u_{\mathrm{n}} = -\alpha(\mathbf{\nabla} T).\boldsymbol{n}, \tag{5}$$

$$T|_{\partial\Omega} = T_{\mathrm{m}} - \beta\kappa. \tag{6}$$

Here, the first equation is just the diffusion equation for the temperature field T. The second equation for the normal velocity, u_{n}, of the solidification front is exactly analogous to the one for the Hele-Shaw cell. The third equation expresses the temperature drop at the solid-melt interface $\partial\Omega$ in which the normal melting temperature T_{m} is corrected by subtle effects related to the curvature of the interface (the Gibbs–Thompson effect). In more sophisticated theories, there is an additional correction proportional to the velocity of the moving front itself. Here, α and β are certain parameters reflecting the physical properties of the crystalizing substance. Thus, despite the very different physics, the mathematical description of the Hele-Shaw cell and a solidifying melt are almost identical Stefan problems. Again, solving the equations is difficult. Apart from a linear stability analysis (leading to the so-called Mullins–Sekerka instability criterion (Langer 1980)), solutions for the dendritic structures and a description of the side-branching structure in particular are very difficult. A classic review of dendritic growth is Langer (1980) and a more recent textbook is Davis (2001).

Other Pattern-Forming Systems

The formulation of the above problems indicates that two quantities appear to play a key role: the gradient of a field variable (pressure or temperature in the above examples) that drives the evolution of the pattern, and curvature of the interface (Pelcé, 1988; Ben-Jacob & Garik, 1990). In fact, these two quantities also play important roles in other pattern-forming systems. In almost every problem whose pattern is in the form of a propagating front or interface (another classic example is flame propagation), the curvature, and the parameter that multiplies it (e.g., surface tension in the case of the Hele-Shaw cell), plays an important role in determining the basic pattern wavelength.

In the introduction, we cited bacterial colony formation as another example of a pattern-forming system. A typical pattern, of a colony of *Bacillus subtilis* on an agar plate, is shown in Figure 1c. What drives the formation of these patterns? A fundamental concept is that of chemotaxis, namely, the response of the organisms to a concentration gradient. Thus, flux of cells in a medium is typically determined by the nutrient gradient. This principle enables one to write down systems of coupled diffusion and hydrodynamic equations for the concentration of cells responding to the nutrient gradient, and for the nutrient concentration itself, which is clearly going to be depleted in regions of high cell concentration. The ensuing nonlinear feedback between these quantities can lead to a rich, and much-studied, pattern structure (Lega & Passot, 2003).

MICHAEL TABOR

See also **Branching laws; Cluster coagulation; Flame front; Hele-Shaw cell; Navier–Stokes equation; Pattern formation; Turing patterns**

Further Reading

Ben-Jacob, E. & Garik, P. 1990. The formation of patterns in non-equilibrium growth. *Nature*, 343: 523–530

Bensimon, D., Kadanoff, L.P., Liang, S., Shraiman, B.I. & Tang, C. 1986. Viscous flows in two dimensions. *Reviews of Modern Physics*, 58: 977–999

Davis, S.H. 2001. *Theory of Solidification*, Cambridge and New York: Cambridge University Press

Langer, J.S. 1980. Instabilities and pattern formation in crystal growth. *Reviews of Modern Physics*, 52: 1–28

Lega, J. & Passot, T. 2003. Hydrodynamics of bacterial colonies: a model. *Physical Review* E, 67: 031960

Murray, J.D. 2003. *Mathematical Biology*, 3rd edition, Berlin and New York: Springer

Pelcé, P. 1988. *Dynamics of Curved Fronts*, Boston: Academic Press

Saffman, P.G. & Taylor, G.I. 1958. The penetration of a fluid into a porous medium or Hele-Shaw cell containing a more viscous fluid. *Proceedings of the Royal Society of London* B, 245: 312–329

Thompson, D.W. 1942. *On Growth and Form*, 2nd edition, Cambridge: Cambridge University Press

Turing, A.M. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London* B, 237: 37–72

GUTZWILLER TRACE FORMULA

See **Quantum theory**

H

HAMILTONIAN SYSTEMS

A system of $2n$, first order, ordinary differential equations

$$\dot{z} = J\nabla H(z, t), \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \qquad (1)$$

is a Hamiltonian system with n degrees of freedom. (When this system is non-autonomous, it has $n + 1/2$ degrees of freedom.) Here H is the Hamiltonian, a smooth scalar function of the extended phase space variables z and time t; the $2n \times 2n$ matrix J is called the "Poisson matrix"; and I is the $n \times n$ identity matrix. The equations naturally split into two sets of n equations for *canonically conjugate* variables, $z = (q, p)$, as follows.

$$\dot{q} = \partial H/\partial p, \quad \dot{p} = -\partial H/\partial q.$$

Here the n coordinates q represent the configuration variables of the system (positions of the component parts), and their canonically conjugate momenta p represent the impetus associated with movement. These equations generalize Newton's second law: $F = ma = \mathrm{d}p/\mathrm{d}t$, to systems (like particles in magnetic fields, or motion in non-inertial reference frames) where the momentum is not simply mass times velocity. The Hamiltonian usually represents the total energy of the system; thus, if $H(q, p)$ does not depend explicitly upon t, then its value is invariant, and Equations (1) are a conservative system. More generally, however, Hamiltonian systems need not be conservative.

William Rowan Hamilton first gave this reformulation of Lagrangian dynamics in 1834 (Hamilton, 1835). However, Hamiltonian dynamics is much more than just a reformulation. It leads, for example, to Henri Poincaré's geometrical insight that gave rise to symplectic geometry, and it provides a compact notation in which the concept of integrability is most naturally expressed and in which perturbation theory can be efficiently carried out. Moreover, nearly integrable Hamiltonian systems exhibit a remarkable stability expressed by the famous results of Kolmogorov–Arnol'd–Moser (KAM) theory and also Nekhoroshev theory. The Hamiltonian formulation also provides the foundation of both statistical and quantum mechanics.

Importantly, virtually all of the dynamical laws of physics—from the motion of a point particle to the interaction of complex quantum fields—have a formulation based on Equations (1). For example, frictionless mechanical systems are described by a Hamiltonian $H(q, p) = K(p) + V(q)$, where K is the kinetic energy (which is often quadratic in p) and V is the potential energy. For example, an ideal planar pendulum consists of a point particle of mass m attached to a massless rigid rod of length L whose other end is attached to a frictionless pivot. The most convenient configuration variable for the pendulum is $q = \theta$, the angle of the rod from the vertical. The gravitational potential energy of the system is then $V = -mgL\cos\theta$, and its kinetic energy is $K = p^2/(2mL^2)$, where p is the angular momentum about the pivot. For this case Equations (1) become

$$\dot{\theta} = \frac{\partial H}{\partial p} = \frac{p}{mL^2}, \quad \dot{p} = -\frac{\partial H}{\partial q} = -mgL\sin\theta. \quad (2)$$

The point $(0, 0)$ is a stable (elliptic) equilibrium corresponding to the pendulum hanging down, at rest. The point $(\pm\pi, 0)$ is also an equilibrium but is an unstable (saddle) point. The unstable eigenvector of the saddle is the beginning of the unstable manifold, W^{u}, a trajectory that is backwards asymptotic to the saddle. By energy conservation, the unstable manifold corresponds to a branch of the energy contour $E = mgL$, which again joins the saddle point (after the pendulum has undergone one full rotation). Thus, this trajectory is forward asymptotic to the saddle as well; that is, it lies on the stable manifold, W^{s}, of the saddle. Orbits of this kind are called *homoclinic*. In this case the homoclinic orbit separates the trajectories oscillating about the elliptic equilibrium from those in which the pendulum undergoes complete rotations, thus it is called a *separatrix*. Orbits near the elliptic equilibrium oscillate with the frequency of the linearized, harmonic oscillator, $\omega = \sqrt{g/L}$. The frequency of oscillation decreases monotonically as the amplitude increases, approaching zero logarithmically at the separatrix. The frequency of rotation of the solution grows again from zero as the energy is further increased.

Canonical Structure

The geometrical structure of Hamiltonian systems arises from the preservation of the loop action, defined by

$$A[\gamma] = \oint_{\gamma} p\,\mathrm{d}q - H\,\mathrm{d}t, \tag{3}$$

where γ is any closed loop in (q, p, t)-space. A consequence of Equations (1) is that if each point on a loop γ_0 is evolved with the flow to obtain a new loop γ_t, then $A[\gamma_0] = A[\gamma_t]$: the loop action is known as the *Poincaré invariant.*

The Hamiltonian form of Equations (1) is not preserved under an arbitrary coordinate transformation (unlike the Euler–Lagrange equations for a Lagrangian system). A canonical transformation $(q, p) \to (q', p')$ preserves the form of the Equations (1). Canonical transformations can be obtained by requiring that the Poincaré invariant of Equation (3) be the same in the new coordinate system, or locally that

$$\left(p'\,\mathrm{d}q' - H'\,\mathrm{d}t\right) - (p\,\mathrm{d}q - H\,\mathrm{d}t) = \mathrm{d}F$$

is the total differential of a function F. If F is represented as a function of a selection of half of the variables (q, p) and the complementary half of (q', p'), it is a "generating" function for the canonical transformation. For example, a function $F(q, q', t)$ implicitly generates a canonical transformation through

$$p = -\frac{\partial F}{\partial q}, \quad p' = \frac{\partial F}{\partial q'},$$
$$H'(q', p', t) = H(q, p, t) - \frac{\partial F}{\partial t}. \tag{4}$$

In order that this transformation be well defined, the first equation must be inverted to find $q'(q, p)$; this requires that the matrix $\partial^2 F/\partial q \partial q'$ be nonsingular. An autonomous canonical transformation is also called a *symplectic map*. Canonical transformations are often employed to simplify the equations of motion. For example, Hamilton's equations are especially simple if the new Hamiltonian is a function of only the momentum variables, $H'(p')$. If we can find a transformation to such a coordinate system then the system is said to be integrable. In general, such transformations do not exist; one of the consequences is chaotic motion.

Integrability

Loosely speaking, a set of differential equations is integrable if it can be explicitly solved for arbitrary initial conditions (Zakharov, 1991). The explicit solution, when inverted, yields the initial conditions as invariant functions along the orbits of a system—the initial conditions are constants of motion. A Hamiltonian $H(q, p)$ is said to be Liouville integrable if it can be transformed to a canonical coordinate system in which it depends only on new momenta. When the energy surfaces are compact and the new momenta are everywhere independent, Arnol'd showed that it is always possible to choose the momentum variables so that their conjugate configuration variables are periodic angles ranging from 0 to 2π (Arnol'd, 1989). These coordinates are called *action-angle variables*, denoted $(\theta, \mathcal{J})$.

As the Hamiltonian is a function only of the actions, $H(\mathcal{J})$, Equation (1) becomes $\dot{\mathcal{J}} = 0$, and $\dot{\theta} = \Omega(\mathcal{J}) = \partial H/\partial \mathcal{J}$. A system is anharmonic when the frequency vector Ω has a nontrivial dependence on $\mathcal{J}$. Thus, for an integrable system, motion occurs on the n-dimensional tori $\mathcal{J} =$ constant. Orbits helically wind around the torus with frequencies Ω that depend upon the torus chosen. When the frequency is nonresonant (there is no integer vector m for which $m \cdot \Omega = 0$), then the motion is dense on the torus.

Any one degree-of-freedom, autonomous Hamiltonian system is locally integrable. A Hamiltonian with more than one degree of freedom, such as pendulum with an oscillating support, is typically not integrable. Systems that are separable into non-interacting parts are integrable, and there are also a number of classical integrable systems with arbitrarily many degrees of freedom. These include the elliptical billiard, the rigid body in free space, the Neumann problem of the motion of a particle on a sphere in a quadratic potential, the Toda lattice, and the Calogero–Moser lattice (Arnol'd, 1988).

Hamiltonian Chaos

The problem of understanding the motion of a slightly perturbed integrable system originated with the desire to understand the motion of the planets. The Kepler problem corresponding to the gravitational interaction of two spherical bodies is integrable; however, once other effects (such as the mutual forces among planets) are included, there appear to be no general, explicit solutions. Poincaré in particular addressed the question of the stability of the solar system, finally realizing that the convergence of perturbation series for the solutions could not be guaranteed and discovering the phenomenon of transverse homoclinic intersections that is a harbinger of chaos (Poincaré, 1892).

Consider the problem

$$H(\theta, \mathcal{J}) = H_0(\mathcal{J}) + \varepsilon H_1(\theta, \mathcal{J}) + \cdots,$$

where the perturbation H_1 depends periodically on the angles and can be expanded in a Fourier series. When

$\Omega(\mathcal{J})$ is nonresonant, a formal sequence of canonical transformations can be constructed to find a set of coordinates in which H is independent of the angle. The problem is the occurrence of denominators in the coefficients proportional to resonance conditions $m \cdot \Omega(\mathcal{J})$ for integer vectors m. Even for actions where the frequencies are incommensurate, it is always possible to make $m \cdot \Omega(\mathcal{J})$ arbitrarily small by choosing large enough integers m. Thus, a priori bounds on the convergence fail. This is called the *problem of small denominators*.

Chirikov realized that small denominators signal the creation of topologically distinct regions of motion (Chirikov, 1979; MacKay & Meiss, 1987). Near a typical resonance, one can use averaging methods to approximate the motion by an integrable pendulum-like Hamiltonian, effectively discarding all of the terms in H_1 except for those that are commensurate with the resonance, that is, the Fourier modes $H_m(\mathcal{J})$ with $m \cdot \Omega = 0$. Thus, orbits near to a resonance are trapped in an effective potential well. The domain of the trapped motion has the width in action of the corresponding pendulum separatrix; it is typically proportional to the square root of the mth Fourier amplitude of H_1. If we can treat the resonances independently, then each gives rise to a corresponding separatrix. However, as the perturbation amplitude grows, this approximation must break down as it predicts the overlap of neighboring separatrices. In 1959, Chirikov proposed this resonance overlap condition as an estimate of the onset of global chaos. Renormalization theory gives a more precise criterion (*See* **Standard map**).

This picture, together with the fact that rational numbers are dense, leads to the expectation that none of the invariant tori of an integrable system persist when it is perturbed with an arbitrarily small perturbation. Surprisingly, the Fermi–Pasta–Ulam computational experiment in 1955 (Fermi et al., 1965; Weissert, 1997) failed to find this behavior. Indeed, KAM theory, initiated by Andrei Kolmogorov in the 1950s and developed in the 1960s by Vladimir Arnol'd and Jürgen Moser, proves persistence of most of the invariant tori (de la Llave, 2001; Pöschel, 2001). This holds when the perturbation is small enough, provided that the system satisfies an anharmonicity or nondegeneracy condition, it is sufficiently differentiable, and the frequency of the torus is sufficiently irrational. The irrationality condition is that $|m \cdot \Omega| > c/|m|^\tau$ for all nonzero integer vectors m and some $c > 0$ and $\tau \geq 1$; this is a *Diophantine condition*. Each of these conditions is essential, though some systems (such as the solar system for which the frequencies are degenerate) can be reformulated so that KAM theory applies.

Resonant tori and tori whose frequencies are nearly commensurate lie between the Diophantine tori. Generally, these tori are destroyed by a small perturbation and either form new, secondary tori trapped in a resonance or are replaced by a zone of chaotic motion that is found in the neighborhood of the stable and unstable manifolds of the resonance. These generically intersect and give rise to a homoclinic tangle or trellis that contains a Smale horseshoe. In the case that the Hamiltonian is analytic, the size of the chaotic region is exponentially small in ε and, thus, can be difficult to detect (Gelfreich & Lazutkin, 2001).

For small perturbations of an integrable Hamiltonian, it remains an open problem to show that a nonzero volume of initial conditions behaves chaotically, in the sense that they have positive Lyapunov exponents. Numerical investigations indicate that orbits in the chaotic zones do have positive Lyapunov exponents, and that these domains form a "fat fractal" (a fractal with positive measure). There are also many examples of uniformly hyperbolic dynamics (especially for the case of billiards (Bunimovich, 1989)) which can also have properties such as mixing and ergodicity.

The problem of the nonlinear stability of a typical system is also open (*See* **Symplectic maps**). However, N.N. Nekhoroshev showed in 1977 that for an analytic system, the actions drift very little for very long times (at most by an amount that is proportional to a power of ε for times that are exponentially long in ε (Lochak, 1993; Pöschel, 1993)). Thus, while it is possible that a KAM torus is unstable, for most practical purposes, they appear to be stable.

Generalizations

Many partial differential equations (PDEs) also have a Hamiltonian structure. For a PDE with independent variables (x, t), the canonical variables are replaced by fields $(q(x,t), p(x,t))$ and the partial derivatives in (1) by functional or Frechêt derivatives, so that

$$\frac{\partial q}{\partial t} = \frac{\delta H}{\delta p}, \quad \frac{\partial p}{\partial t} = -\frac{\delta H}{\delta q}.$$

The Hamiltonian functional H is the integral of an energy density. For example, the wave equation has the Hamiltonian $H[q,p] = \int \mathrm{d}x \frac{1}{2}\left(p^2 + c^2(\partial_x q)^2\right)$. Other nonlinear wave equations such as the integrable nonlinear Schrödinger, Korteweg–de Vries, and sine-Gordon equations also have Hamiltonian formulations.

James D. Meiss

See also **Adiabatic invariants; Chaotic dynamics; Constants of motion and conservation laws; Ergodic theory; Euler–Lagrange equations; Fermi–Pasta–Ulam oscillator chain; Hénon–Heiles system; Horseshoes and hyperbolicity in dynamical systems; Lyapunov exponents; Mel'nikov method; Pendulum; Phase space; Poisson brackets; Standard map; Symplectic maps; Toda lattice**

Further Reading

Arnol'd, V.I. (editor). 1988. *Dynamical Systems III*, New York: Springer

Arnol'd, V.I. 1989. *Mathematical Methods of Classical Mechanics*, New York: Springer

Bunimovich, L.A. 1989. Dynamical systems of hyperbolic type with singularities. In *Dynamical Systems*, edited by Ya. Sinai, Berlin: Springer, p. 278

Chirikov, B.V. 1979. A universal instability of many-dimensional oscillator systems. *Physics Reports*, 52: 265–379

de la Llave, R. 2001. A tutorial on KAM theory. In *Smooth Ergodic Theory and Its Applications (Seattle, WA, 1999)*, Providence, RI: American Mathematical Society, pp. 175–292

Fermi, E., Pasta, J. & Ulam, S. 1965. Studies of nonlinear problems. In *Collected Papers of Enrico Fermi*, vol. 2, edited by E.Segré, Chicago: University of Chicago Press, pp. 977–988

Gelfreich, V.G. & Lazutkin, V.F. 2001. Separatrix splitting: perturbation theory and exponential smallness. *Russian Mathematical Surveys*, 56(3): 499–558

Hamilton, W.R. 1835. On the application to dynamics of a general mathematical method previously applied to optics. *British Association Report, 1834*, pp. 513–518

Lochak, P. 1993. Hamiltonian perturbation theory: periodic orbits, resonances and intermittancy. *Nonlinearity*, 6: 885–904

MacKay, R.S. & Meiss, J.D. (editors). 1987. *Hamiltonian Dynamical Systems: A Reprint Selection*, London: Adam Hilger

Poincaré, H. 1892. *Les Methodes Nouvelles de la Mechanique Celeste*, Paris: Gauthier–Villars

Pöschel, J. 1993. Nekhoroshev estimates for quasi–convex Hamiltonian systems. *Mathematische Zeitschrift*, 213:187–216

Pöschel, J. 2001. A lecture on the classical KAM theorem. In *Smooth Ergodic Theory and Its Applications* (*Seattle, WA, 1999*), Providence, RI: American Mathematical Society, pp. 707–732

Weissert, T.P. 1997. *The Genesis of Simulation in Dynamics: Pursuing the Fermi–Pasta–Ulam Problem*, New York: Springer

Zakharov, V.E. (editor). 1991. *What Is Integrability?* Berlin: Springer

HARMONIC GENERATION

Harmonic generation is the phenomenon whereby new frequency components of a wave are created upon interaction with a nonlinear medium. The concept of a harmonic frequency is familiar to anyone with a basic knowledge of music: two notes whose frequencies form a ratio of small whole numbers (e.g., 2:1, 3:2, 5:4) produce a sound that is pleasing or "harmonious" compared with two notes with incommensurate frequencies. More generally, any periodic function $f(t+T)=f(t)$ can be written as an infinite Fourier series containing the first harmonic (fundamental frequency) $\omega = 2\pi/T$, the second harmonic, and higher harmonic frequencies $\omega_m = m\omega$:

$$f(t) = \sum_{m=-\infty}^{\infty} c_m \, \mathrm{e}^{-\mathrm{i}\omega_m t}. \tag{1}$$

The function $f(t)$ could represent the displacement of a guitar string, the amplitude of a wave in the ocean, or the electric field strength of a radio wave or a beam of light. If f satisfies a linear dynamic equation, then the principle of superposition holds, and no energy can be exchanged among the different Fourier coefficients c_m. Under nonlinear dynamics, however, the coefficients c_m are coupled to each other, and energy can be transferred back and forth among them.

Harmonic generation refers to the case in which energy originally at frequency ω generates a new wave component at a harmonic frequency ω_m. A closely related phenomenon is that of parametric amplification, in which a frequency ω_m provides the energy source to amplify a signal at a related frequency $\omega_{m'}$.

In optics, harmonic generation is based upon the nonlinear response of some medium to electromagnetic radiation at optical frequencies. The origin of harmonic generation is related to Theodore Maiman's invention of the ruby laser in 1960. Exploiting the high intensities and the coherency properties of the newly available laser output, in 1961 Peter Franken and his colleagues used the nonlinear polarization response of crystal quartz to produce second-harmonic light at a free-space wavelength of 347 nm from the ruby laser output at 694 nm, marking the birth of nonlinear optics.

By the late 1960s, lithium niobate ($LiNbO_3$) had emerged as the nonlinear crystal of choice for parametric amplification and harmonic generation. Much of the work taking place in the following decades was dedicated to improving the efficiency of the generation of harmonics, either by finding materials with better properties (higher transparency, stronger nonlinearity, etc.) or by aiding the energy transfer through phase matching the various waves involved in the process. More recently, improvements in material processing have made possible the technique of quasi-phase-matching (QPM), in which better phase matching is achieved through the engineering of a spatial variation of the nonlinear coefficient of the medium. In contrast to previous phase-matching techniques, QPM allows a wide, continuous range of frequencies to be coupled to their harmonics without imposing constraints on the orientation of the crystal; the orientation can then be chosen to maximize the nonlinear coefficient seen by the fundamental frequency and its harmonic as they pass through the crystal.

In optics, harmonic generation is often used to convert energy from one frequency band to another. The generated frequency may be more suitable for telecommunications or spectroscopy, for example, and may be in a region of the electromagnetic spectrum that is not directly available from high-powered lasers. The generation of a harmonic might also be a way of obtaining a coherent copy of an optical pulse, where the copy can be compressed, broadened, chirped, squeezed, or simply measured to determine the properties of the original pulse. Depending on the type of physical process involved and the particular application considered, one refers to the relevant phenomena as second-harmonic generation, sum- or difference-frequency generation, optical parametric generation, or optical parametric oscillation.

The canonical example for studying harmonic generation is an anharmonic oscillator, which can be taken as a simplified model of coherent light propagating through a nonlinear medium. In optics, the nonlinearity can arise due to the material polarization (electronic response) of the medium, or due to a coupling between the light and lattice vibrations in the medium (optical phonons). Referring to the former phenomenon, the polarization response can be expanded in a power series of the applied field, where the coefficients, the nonlinear susceptibility tensors of the material, can be obtained through quantum-mechanical calculations. In general, the quadratic polarization of a medium will be its leading order nonlinear response unless symmetries of the medium rule it out. In this case, one can then reduce Maxwell's equations to the nonlinear wave equation

$$\frac{\partial^2 E}{\partial z^2} - \frac{n^2}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{\chi^{(2)}}{c^2}\frac{\partial^2 (E^2)}{\partial t^2}, \tag{2}$$

where n is the refractive index of light in the medium, c is the speed of light, and $E(z,t)$ is the electromagnetic field strength which depends on a longitudinal dimension z and on time t.

The nonlinear susceptibility $\chi^{(2)}$ characterizes the quadratic response of the medium to an incident field, where the linear susceptibility has been included in the definition of the refractive index n. To observe the exchange of energy between two frequencies ω_1 and ω_2 mediated by this nonlinearity, the Fourier expansion of Equation (1) can be augmented as

$$E(z,t) = c_1(z)\,\mathrm{e}^{\mathrm{i}\theta_1(z,t)} + c_2(z)\,\mathrm{e}^{\mathrm{i}\theta_2(z,t)} + \text{c.c.}, \tag{3}$$

where "c.c." denotes the complex conjugate, $\theta_j(z,t) = k_j z - \omega_j t$ is a rapidly varying optical phase, and $k_j = n\omega_j/c$ is the linear wave number. Under the assumption that the coefficients $c_j(z)$ vary much more slowly in z than the optical phase (slowly varying envelope approximation), substituting Equation (3) into Equation (2) gives

$$\begin{aligned} &2\mathrm{i}k_1 c_1'(z)\mathrm{e}^{\mathrm{i}\theta_1} + 2\mathrm{i}k_2 c_2'(z)e^{\mathrm{i}\theta_2} + \text{c.c.} \\ &= \frac{\chi^{(2)}}{c^2}\frac{\partial^2}{\partial t^2}(E^2), \end{aligned} \tag{4}$$

where

$$\begin{aligned} E^2 &= c_1^2\mathrm{e}^{\mathrm{i}(2\theta_1)} + c_2^2\mathrm{e}^{\mathrm{i}(2\theta_2)} + 2c_1^* c_2\mathrm{e}^{\mathrm{i}(-\theta_1+\theta_2)} \\ &\quad + 2c_1c_2\mathrm{e}^{\mathrm{i}(\theta_1+\theta_2)} + |c_1|^2 + |c_2|^2 + \text{c.c.} \end{aligned} \tag{5}$$

In order for the terms on the right of Equation (4) to contribute significantly to the growth of $c_1(z)$ or $c_2(z)$ on the left, the rapid phase rotations given by the exponentials must resonate, which occurs if $\theta_2 = 2\theta_1$, or $\omega_2 = 2\omega_1$. The resulting equations will then be

$$c_1'(z) = \frac{\mathrm{i}\omega_1\chi^{(2)}}{n_1 c}\, c_1^* c_2 \mathrm{e}^{\mathrm{i}\Delta kz}, \tag{6a}$$

$$c_2'(z) = \frac{\mathrm{i}\omega_2\chi^{(2)}}{n_2 c}\, c_1^2 \mathrm{e}^{-\mathrm{i}\Delta kz}. \tag{6b}$$

The phase mismatch term $\Delta k = k_1 - 2k_2$ arises from a more careful calculation that takes into account that the medium's linear susceptibility (and therefore the refractive index n) is a function of frequency ω; thus, it is not generally true that $\omega_2 = 2\omega_1$ implies $k_2 = 2k_1$.

Equations (6) can be solved exactly, but it is more instructive to consider the quantities that are invariant in this system of equations. These invariants reflect physical quantities that are conserved during the mixing process (called Manley–Rowe invariants):

$$I_0 = \frac{1}{2}\varepsilon_0 c\left(n_1|c_1|^2 + n_2|c_2|^2\right), \tag{7a}$$

$$I_1 = \frac{1}{2}\varepsilon_0 c\left(\frac{n_1|c_1|^2}{\hbar\omega_1} + 2\frac{n_2|c_2|^2}{\hbar\omega_2}\right). \tag{7b}$$

The first of these, I_0, is the total irradiance, or incident optical power per unit of surface area. That I_0 is invariant follows from conservation of energy, which is to be expected given that no absorption has been accounted for here in the linear susceptibility. The quantity I_1 sums the irradiance from the two frequencies, dividing each by the energy of a single photon at the respective frequency ($E_j = \hbar\omega_j$) before summing. The invariance of I_2, therefore, reflects a law of conservation of photon flux, as the destruction of two photons at the fundamental frequency ω_1 must bring about the creation of a single photon at the second harmonic, $\omega_2 = 2\omega_1$.

In the spectroscopy of molecular crystals, the fundamental frequency of a vibrating mode sometimes lies close to an overtone or combination of other mode frequencies, whereupon both frequencies are pushed away from each another. In such a case, the overtone or combination band borrows intensity from the fundamental in a phenomenon called Fermi resonance.

Richard O. Moore and Gino Biondini

See also **Frequency doubling; Manley–Rowe relations; Nonlinear optics; Parametric amplification**

Further reading

Armstrong, J.A., Bloembergen, N., Ducuing, J. & Pershan, P.S. 1962. Interactions between light waves in a nonlinear dielectric. *Physical Review*, 127: 1918

Bloembergen, N. 1965. *Nonlinear Optics*, Singapore: World Scientific

Boyd, R.W. 2003. *Nonlinear Optics*, 2nd edition, San Diego: Academic Press

Dunn, M.H. & Ebrahimzadeh, M. 1999. Parametric generation of tunable light from continuous-wave to femtosecond pulses. *Science*, 286: 1513

Franken, P.A., Hill, A.E., Peters, C.W. & Weinreich, G. 1961. Generation of optical harmonics. *Physical Review Letters*, 7: 118

Maiman, T.H. 1960. Stimulated optical radiation in ruby. *Nature*, 187: 493

Schubert, M. & Wilhelmi, B. 1986. *Nonlinear Optics and Quantum Electronics*, New York: Wiley

HARMONIC OSCILLATORS

See **Damped-driven anharmonic oscillator**

HARTREE APPROXIMATION

In the quantum-mechanical treatment of a system with many degrees of freedom, the Hartree approximation writes the multidimensional correlated wave function as a simple product of one-dimensional or low-dimensional (orbital) functions. This ansatz reduces the full Schrödinger equation to a set of simpler equations for the orbital functions. The Hartree product approximation, first proposed by Douglas R. Hartree (1928) for many-electron atoms, can in principle be applied to any many-particle system, for example, electrons in an atom, molecule, or crystal. However, the Hartree ansatz is often physically inadequate as it disregards the fundamental permutation symmetry requirements for wave functions of nondistinguishable particles (fermions or bosons). In the case of electron systems, the Hartree approximation soon gave way to the more appropriate Hartree–Fock (HF) approximation that is based on a fully antisymmetrized product representation (Slater determinant) of the many-electron wave function proposed by Vladimir A. Fock (Fock, 1930a). For systems with distinct degrees of freedom, on the other hand, such as the multimode nuclear dynamics in molecules, the Hartree product representation still represents a basic approximation. (see, e.g., Beck et al., 2000).

For the many-electron system of an atom or molecule, the (nonrelativistic) stationary Schrödinger equation can be written in the form (see, e.g., Landau & Lifshitz, 1977)

$$\hat{H}\Phi_n = E_n\Phi_n, \tag{1}$$

where the Hamiltonian $\hat{H} = \hat{H}_o + \hat{V}$ consists of a one-particle part,

$$\hat{H}_o = \sum_i \hat{h}_i, \quad \hat{h}_i = -\frac{\hbar^2}{2m}\nabla^2 - \sum_\alpha \frac{Z_\alpha e^2}{|\boldsymbol{R}_\alpha - \boldsymbol{r}_i|} \tag{2}$$

i.e., $\hat{H}_o$ is the sum of the Hamiltonians $\hat{h}_i$ associated with the motion of the ith electron in the field of the nuclei with charges eZ_α at positions $\boldsymbol{R}_\alpha$, and a two-particle part,

$$\hat{V} = \frac{1}{2}\sum_{i \neq j} \frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} \tag{3}$$

corresponding to the Coulomb repulsion of the electrons. The complexity of the many-electron problem is due to the latter part of the Hamiltonian, as it prevents a separation of the N-electron Schrödinger equation into N independent single-particle equations. The idea underlying the Hartree and the Hartree–Fock approximation is to replace the full electron-electron interaction by a new approximately separable effective Coulomb repulsion (or "mean field"). In the Hartree case, the optimal mean-field formulation is obtained by making use of a variational principle (Fock, 1930b),

$$\delta\langle\Phi|\hat{H}|\Phi\rangle = 0 \tag{4}$$

with the trial wave function given by a Hartree product

$$\Phi(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) = \phi_1(\boldsymbol{x}_1)\phi_2(\boldsymbol{x}_2)\ldots\phi_N(\boldsymbol{x}_N) \tag{5}$$

of normalized single-electron functions (orbitals) $\phi_i(\boldsymbol{x}_i)$. Here $\boldsymbol{x}_i = (\boldsymbol{r}_i, s_i)$ comprises in a short-hand notation the set of spatial and spin variables of the ith electron. The resulting equations for the orbitals read

$$(\hat{h} + \hat{v}_k^H)\phi_k(\boldsymbol{x}) = \varepsilon_k\phi_k(\boldsymbol{x}), \quad k = 1, 2, \ldots, N, \tag{6}$$

where for each orbital $\phi_k(\boldsymbol{x})$ a distinct mean field

$$\hat{v}_k^H = \sum_{l \neq k} \hat{J}_l, \qquad \hat{J}_l = e^2 \int \frac{|\psi_l(\boldsymbol{r}')|^2}{|\boldsymbol{r}' - \boldsymbol{r}|}\,\mathrm{d}\boldsymbol{r}' \tag{7}$$

augments the one-particle Hamiltonian $\hat{h}$. The physical meaning of $\hat{v}_k^H$ is the Coulomb energy arising from the interaction of the electron in orbital $\phi_k(\boldsymbol{x})$ and the charge distribution associated with the other orbitals (the operators $\hat{J}_l$ are referred to as Coulomb operators).

Equations (6) are a set of nonlinear integro-differential equations in the orbitals $\phi_1(\boldsymbol{x}_1)$, $\phi_2(\boldsymbol{x}_2), \ldots, \phi_N(\boldsymbol{x}_N)$. Starting with an initial guess for the orbitals $\phi_k(\boldsymbol{x})$, these equations have to be solved in an iterative way until self-consistency is reached. Here it is essential to select appropriate orbital solutions of the individual one-particle eigenvalue problems (6) so that in the N-electron product wave function the Pauli exclusion principle is at least approximately fulfilled. The resulting potential energy operators (7) constitute the Hartree self-consistent field (SCF).

In the many-electron problem, the Hartree approximation is only of historical or pedagogical interest. It was soon succeeded by the more rigorous and practical Hartree–Fock (HF) approximation (Fock, 1930a) based on the proper antisymmetrized product (Slater determinant) representation

$$\Phi(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) = \begin{vmatrix} \phi_1(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_2) \ldots & \phi_1(\boldsymbol{x}_N) \\ \phi_2(\boldsymbol{x}_1) & \phi_2(\boldsymbol{x}_2) \ldots & \phi_2(\boldsymbol{x}_N) \\ \vdots & \vdots & \vdots \\ \phi_N(\boldsymbol{x}_1) & \phi_N(\boldsymbol{x}_2) \ldots & \phi_N(\boldsymbol{x}_N) \end{vmatrix} \tag{8}$$

of the N-electron wave function. This leads to one common eigenvalue problem for the HF orbitals,

$$(\hat{h} + \hat{v}^{\mathrm{HF}})\phi_k(\boldsymbol{x}) = \varepsilon_k \phi_k(\boldsymbol{x}), \quad k = 1, 2, \ldots, N, \tag{9}$$

where the HF mean field operator

$$\hat{v}^{\mathrm{HF}} = \sum_{l=1}^{N} (\hat{J}_l - \hat{K}_l) \tag{10}$$

now contains in addition to the Coulomb contributions a so-called exchange part $-\sum_{l=1}^{N} \hat{K}_l$, and the nonlocal exchange operators $\hat{K}_l$ are defined by

$$\hat{K}_l \phi(\boldsymbol{x}) = e^2 \int \frac{\phi_l^*(\boldsymbol{x}')\phi(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}'}{|\boldsymbol{r}' - \boldsymbol{r}|} \phi_l(\boldsymbol{x}) \,. \tag{11}$$

Here the integration sums over the spin variable. The HF eigenvalue equation (9) represents a nonlinear problem that again has to be solved in a self-consistent way. It should be noted that because of $(\hat{J}_l - \hat{K}_l)\phi_l = 0$, there arise no self-interaction contributions in the self-consistent HF mean field. Usually, the HF or "molecular" orbitals (MO) are written as linear combinations (LC) of basis functions or "atomic" orbitals (AO). This LCAO representation transforms the integro-differential HF equation (9) into an algebraic eigenvalue problem, and the self-consistency solution is achieved by successive matrix diagonalization. Both with respect to practical applicability and theoretical significance, the HF (or SCF) method is fundamental for electronic structure computations in atoms, molecules, and the solid state. In the LCAO form it is a main ingredient in all existing quantum chemistry program packages. Most of the methods aiming at the description of electron correlation, that is, the effects beyond the HF one-particle picture, are based on the SCF orbital energies and integrals as input data.

Although in electronic structure calculations the Hartree approximation is no longer used, it is still a viable means in the description of nuclear dynamics, that is, the nuclear motion accompanying electronic states and processes. Because one deals here with distinguishable "particles" and not with fermions, the Hartree and not the Hartree–Fock approximation is the appropriate choice. Here one assumes a specific kind of separation of the nuclear and electronic degrees of freedom that has its physical foundation on the large mass differences of electrons and nuclei and correspondingly large differences in the respective velocities (Born & Oppenheimer, 1927). Setting molecular rotations aside, one may write the total wave function as a product of an electronic wave function $\Phi(\boldsymbol{x}; \boldsymbol{R})$ and a wave function $\chi(\boldsymbol{R})$ associated with the vibrational motion of the nuclei:

$$\Psi(\boldsymbol{x}, \boldsymbol{R}) = \Phi(\boldsymbol{x}; \boldsymbol{R})\chi(\boldsymbol{R}) \tag{12}$$

Here $\boldsymbol{x}$ and $\boldsymbol{R}$ denote collectively the electronic and nuclear variables. The electronic function is an eigenfunction of the electronic Hamiltonian at spatially fixed nuclei and, thus, depends parametrically on the nuclear coordinates $\boldsymbol{R}_\alpha$. The product approximation (12) is referred to as Born–Oppenheimer approximation. Having determined the electronic wave function and energy $E(\boldsymbol{R})$, the Schrödinger equation for the vibrational wave function can be formulated as follows:

$$\left(\sum_\alpha -\frac{\hbar^2}{2M_\alpha} \nabla_\alpha^2 + \hat{U}(\boldsymbol{R}) \right) \chi(\boldsymbol{R}) = E_{\mathrm{tot}}\, \chi(\boldsymbol{R}) \tag{13}$$

Here the potential energy term $\hat{U}(\boldsymbol{R})$ is the sum of the nuclear repulsion energy and the electronic energy $E(\boldsymbol{R})$. Since we are here not interested in translational or rotational motion of the molecule as a whole, it is advantageous to replace the Cartesian coordinates used so far by a set of $M = 3N - 6$ nuclear coordinates $Q_1, Q_2, \ldots, Q_M$ describing, for example, the displacements from a reference nuclear configuration. Assuming a separable form of the kinetic energy in the new variables Q_i, that is,

$$\hat{T} = \sum_{i=1}^{M} \hat{t}_i \tag{14}$$

the Hartree ansatz

$$\chi(Q_1, Q_2, \ldots, Q_M) = \chi_1(Q_1)\chi_2(Q_2)\ldots\chi_M(Q_M) \tag{15}$$

leads to the following M nonlinear coupled one-dimensional eigenvalue equations,

$$(\hat{t}_i + \hat{u}_i)\chi_i(Q_i) = e_i \chi_i(Q_i), \qquad i = 1, \ldots, M, \tag{16}$$

where the Hartree mean fields $\hat{u}_i$ are given by

$$\hat{u}_i = \int U(Q_1, Q_1, \ldots, Q_M) \prod_{j \neq i} (|\chi(Q_j)|^2 \,\mathrm{d}Q_j). \tag{17}$$

The Hartree product representation is particularly useful in the context of time-dependent methods which solve the time-dependent Schrödinger equation by propagating an initial wave fuction numerically over a sufficiently long period of time. A very efficient and accurate example of such methods is the multi-configuration time-dependent Hartree (MCTDH) method (Beck et al., 2000). Here the representation of the wave function is given by a linear combination of several Hartree products.

Peter Schmelcher and Jochen Schirmer

See also **Quantum theory**

Further Reading

Beck, M.H., Jäckle, A., Worth, G.A. & Meyer, H.-D. 2000. The multiconfiguration time-dependent Hartree method: a highly efficient algorithm for propagating wavepackets. *Physics Reports*, 324: 1–105

Born, M. & Oppenheimer, R. 1927. On the quantum theory of molecules. *Annalen der Physik*, 84: 457

Fock, V. 1930a. Self-consistent field mit Austausch für Natrium. [Self-consistent field with exchange for Sodium]. *Zeitschrift für Physik*, 62: 795–805

Fock, V. 1930b. Näherungsmethode zur Lösung des quantenmechanischen Mehrkorperproblems. [Approximate method for solution of quantum many body problem]. *Zeitschrift für Physik*, 61: 126–148

Hartree, D.R. 1928. *Proceedings of the Cambridge Philosophical Society*, 24: 111

Landau, L.D. & Lifshitz, E.M. 1977. *Course of Theoretical Physics*, vol. 3: *Quantum Mechanics*, Oxford: Pergamon Press

HAUSDORFF DIMENSION

See **Dimensions**

HEAT CONDUCTION

There are two main aspects of the theory of heat conduction that are intensively treated in nonlinear science. The first is related to the possibility of nonlinear terms in the phenomenological equations of heat conduction. The second aspect is related to the microscopic nature of heat conduction and especially to deriving Fourier's law (proportionality of heat flux to the temperature gradient) from the fundamental laws of motion at an atomic level.

Equations of Heat Conduction

In the phenomenological theory of heat conduction, the nonlinearity is normally manifested in temperature dependence of the thermal conductivity coefficient (or concentration dependence of the diffusion coefficient in the related diffusion problem). Such dependence leads to a nonlinear modification of Fourier's law. In the one-dimensional case, this equation may be written as

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left[\kappa(T)\frac{\partial T}{\partial x}\right], \tag{1}$$

where $T(x, t)$ is the temperature field, $\kappa(T) = \vartheta/\rho C$, ϑ is the heat conductivity coefficient, ρ is the density, and C is the heat capacity of the specimen. All parameters introduced above may be significantly temperature-dependent.

As an example of nonlinear heat conduction, consider heat transfer in a hot plasma driven primarily by radiation. The internal energy of the photon gas is $(4\sigma/c)T^4$, where c is the velocity of light and σ is the Stefan–Boltzmann constant. The heat capacity $C = 16\sigma T^3/3c$. Using results of kinetic theory and plasma physics,

$$\lambda \sim T^{7/2}\rho^{-2}, \tag{2}$$

where λ is the mean free path one can derive the final temperature dependence of the thermal conductivity coefficient: $\kappa(T) \sim T^{13/2}$. In the more common case of molecular heat conduction, the transfer is realized by gas molecules, and therefore, the average velocity of the molecules $v \sim T^{1/2}$ and λ is independent of the temperature ($\lambda \sim 1/\rho$); accordingly, the heat conductivity $\kappa \sim T^{1/2}$. More general models deal with the general form of the temperature dependence, like $\kappa = \kappa_0 T^{\nu}$, $\nu > 0$.

Probably, the most impressive consequence of the temperature dependence of the thermal conductivity coefficient is the phenomenon of so-called "heat inertia" when the heat energy remains localized in the system and does not dissipate away. If $T(x, t)$ is the solution of the Cauchy problem for Equation (1) with initial temperature

$$T(x, t_0) = \begin{cases} T_m(1 - |x|/x_0)^{2/\nu}, & |x| \le x_0, \\ 0, & |x| \ge 0, \end{cases}$$

then $T(x, t) = 0$ for $|x| \ge x_0$, $t_0 < t < t_0 + t^*$, $t^* = x_0^2\nu/2\kappa_0(\nu + 2)T_m^{\nu}$; that is, the heat energy is localized in the initial region for a certain time interval.

Microscopic Nature of Heat Conduction

The problem of heat conduction in dielectric crystals goes back to Rudolph Peierls (1926). A linear model of the crystalline lattice, being useful in many other physical problems, turns out to be insufficient when explaining the finiteness of the coefficient of thermal conductivity. The normal process of the heat conduction implies formation of a linear temperature gradient as the temperatures of the boundaries of the specimen are different. In an ideal linear crystal, however, any excitation is presented as a sum of independent vibrational modes that transfer energy from one boundary to the other without any loss. Therefore, a temperature gradient cannot be formed. This is why an account of the nonlinearity of interactions is necessary to describe the finite heat conductivity. Peierls has suggested that this condition is also a sufficient one because the phonon collision in a weakly nonlinear crystal is governed by the following relationships:

$$\begin{aligned} \omega_1 + \omega_2 &= \omega_1' + \omega_2', \\ \boldsymbol{k}_1 + \boldsymbol{k}_2 &= \boldsymbol{k}_1' + \boldsymbol{k}_2' + m\boldsymbol{p}. \end{aligned} \quad (3)$$

The values with a prime correspond to the state after collision. The relationship for frequencies ω_i expresses the energy conservation. However the conservation of the quasimomenta $\boldsymbol{k}_i$ can be violated by a multiple of Bragg vector $\boldsymbol{p}$. The processes with $m = 0$ are referred to as normal phonon scattering. The case with $m \neq 0$ is referred to as Umklapp process. The latter case means that a certain quantity of the quasimomentum is transferred to the lattice. According to Peierls, the fact that the heat flux is scattered due to the Umklapp processes leads to normal heat conductivity. This picture was accepted until Fermi, Pasta, and Ulam examined it in one of the first computer simulations ever accomplished, finding that a one-dimensional chain with weak anharmonicity of the interaction potential did not demonstrate normal heat conduction (*See* **Fermi–Pasta–Ulam oscillator chain**). Instead, the energy was preserved in a few low-frequency vibrational modes. This famous numerical experiment led to great progress in nonlinear dynamics, including the discovery of lattice solitons, but the problem of the finite heat conductivity of a dielectric crystal again turned out to be unsolved.

It is easy to understand the essence of the problem of finding a sufficient condition for normal heat conductivity. Let us consider a set of equivalent particles arranged along a straight line and interacting only via purely elastic collisions. In every collision the interacting particles mutually exchange their velocities. Any pulse propagates through the system without loss, and therefore, a temperature profile is not formed. Similar properties were proved for the chain with exponential potential of interaction between the nearest neighbors (the so-called Toda lattice).

This problem remains unsolved to date despite numerous efforts. The lack of reliable analytical methods has resulted in many numerical simulations dealing with various lattice models. For one-dimensional systems the results available may be summarized as follows.

- A narrow class of completely integrable systems demonstrates no formation of linear temperature profile at all. Consequently, the coefficient of thermal conductivity cannot be defined. Examples are the linear chain and the Toda lattice.
- For many non-integrable chains, a linear temperature profile is formed and the coefficient of thermal conductivity may be defined for any given size N of the specimen. However, in the thermodynamic limit $N \to \infty$, the coefficient diverges: $\kappa \to \infty$. Therefore, such systems cannot be treated as having normal heat conductivity. Examples are the chain with parabolic and quartic potential among many others.
- A few one-dimensional chains were recently demonstrated to have normal heat conductivity. These examples are the chain of rotators, the Frenkel–Kontorova chain, and some other models.

It should be mentioned that some chains seem to "switch" between different types of behavior with respect to normal heat conductivity with changes of temperature and/or parameters of the potential of interaction.

It seems that different regimes of heat conduction in one-dimensional systems are related to the properties of nonlinear excitations in every concrete model. For instance, the normal heat conduction in the chain of rotators was found to be related to the breakdown of nonlinear acoustic waves and formation of localized breathers, which scatter the heat flux. Similar mechanisms were discovered in other models exhibiting normal heat conduction.

For two-dimensional systems very little information is available, but some investigations suggest that the heat conductivity for different types of potentials is finite or logarithmically divergent. No classification similar to the one-dimensional case can yet be provided.

For the three-dimensional case the situation is as unknown, but the field of molecular dynamics is developing very rapidly and new results are expected to arrive soon.

OLEG GENDELMAN AND LEONID MANEVITCH

See also **Fermi–Pasta–Ulam oscillator chain; Frenkel–Kontorova model; Integrable lattices; Solitons; Toda lattice**

Further Reading

Bonetto, F., Lebowitz, J. & Rey-Bellet, L. 2000. In *Mathematical Physics 2000*, edited by A. Fokas, A. Grigoryan, T. Kibble, & B. Zegarlinski, River Edge, NJ : Imperial College Press

Fermi, E., Pasta, J.R. & Ulam, S.M. 1955. *Studies of Nonlinear Problems*, Los Alamos Scientific Laboratory Report No. LA–1940

Gendelman, O.V. & Savin, A.V. 2000. Normal heat conductivity in one-dimensional chain of rotators. *Physical Review Letters*, 84: 2381–2384

Lepri, S., Livi, R. & Politi, A. 2001. Thermal conduction in classical low-dimensional lattices, preprint: www.lanl.gov, arXiv:cond-mat/0112193

Peierls, R.E. 1955. *Quantum Theory of Solids*, London: Oxford University Press

Samarskii, A.A., Galaktionov, V.A., Kurdiumov, S.P. & Mikailov, A.P. 1987. *Regimes with Sharpening in the Problems for Quasilinear Parabolic Equations*, Moscow, Nauka (in Russian)

Toda, M. 1967. Vibration of a chain with nonlinear interactions. *Journal of the Physical Society of Japan*, 22: 431–36; Wave propagation in anharmonic lattices. *Journal of the Physical Society of Japan*, 23: 501–06

HEAVISIDE STEP FUNCTION

See **Generalized functions**

HELE-SHAW CELL

A very simple device invented by Henry Selby Hele-Shaw in the late 1890s, the Hele-Shaw cell consists of two closely spaced parallel plates, held in place by a frame, between which is a layer of a viscous liquid. Hele-Shaw invented this simple instrument in order to be able to exhibit the streamlines of a flow to students at University College of Liverpool by projecting the flow in the thin liquid layer onto a large screen. He found that minute air bubbles provided excellent flow tracers, and developed the technique to be able to display in a spectacular manner the flow patterns past a variety of objects.

Hele-Shaw discovered that when the distance between the two plates was very small (he used water as the fluid, and glass plates mounted within 0.02 in. of each other), the flow pattern around a circular cylinder perpendicular to the plates closely matched the potential flow solution that was known from the theory of two-dimensional flow of an ideal fluid. Apparently, the viscous liquid between the plates was flowing in this narrow space as though it were an inviscid fluid. George Stokes explained these observations by writing the equation of motion for a viscous fluid—today known as the Navier–Stokes equation—and averaging the flow (today known as plane Poiseuille flow) across the narrow gap between the two bounding plates in the Hele-Shaw cell. One then finds that the averaged, essentially two-dimensional flow velocity, $\boldsymbol{v}$, is proportional to the gradient of the pressure, p, in the fluid:

$$\boldsymbol{v} = -(d^2/12\mu)\nabla p, \tag{1}$$

where d is the separation between the plates and μ is the viscosity of the fluid. Thus, flow in a Hele-Shaw cell is potential flow, with a velocity potential proportional to the pressure, and this explains the coincidence between Hele-Shaw's observations and the theory of two-dimensional, ideal flow. As Stokes wrote, "[Hele-Shaw's experiments] afford a complete graphical solution, experimentally obtained, of a problem which from its complexity baffles mathematicians except in a few simple cases."

Hele-Shaw flows illustrate two-dimensional potential flow around objects with sharp edges and thus show no separation. This is quite remarkable, because even the slightest viscosity would immediately alter the flow pattern were the fluid not constrained by the two plates of the Hele-Shaw cell.

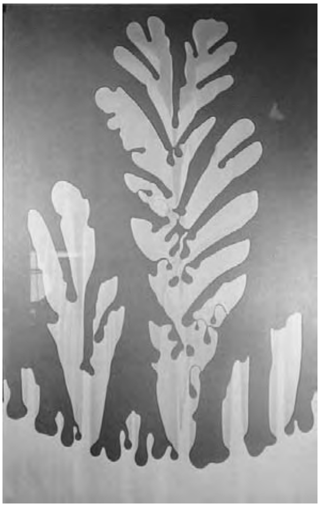

Figure 1. Viscous fingering patterns in a Hele-Shaw cell at a gravitationally unstable interface between two immiscible fluids—viscous silicone oil and air. (Reprinted with permission from Gallery of Fluid Motion, *Physics of Fluids*, 28(9). Copyright 1985, American Institute of Physics. Courtesy of T. Maxwell.)

A controversy with Osborne Reynolds ensued about the validity of the experiments since they appeared to invalidate Reynolds' criterion of a transition to turbulence based solely on the value of "Reynolds number." This led to several acrimonious exchanges of notes and letters in *Nature* in 1898 and 1899 between Reynolds and Hele-Shaw.

The Hele-Shaw cell has found wide application in later years, after it was realized that Equation (1) has the form of the velocity-pressure relationship one would expect for fluid flow in a porous material obeying Darcy's law. Thus, the Hele-Shaw cell became a paradigmatic instrument for studying flows in porous media, a subject of great interest to petroleum engineering.

When two immiscible fluids of different viscosities are present in a Hele-Shaw cell, a number of interesting phenomena occur. In the late 1950s, Geoffrey Taylor and Philip Saffman explored the evolution of "fingers" when a less viscous fluid (air) was pushed into a more viscous fluid (water, glycerin, or oil) in a Hele-Shaw cell. Similar "fingering" occurs when water is forced into oil-carrying porous rock in secondary oil recovery.

The subject of viscous fingering is similar to other pattern-forming instabilities, dendritic crystal growth in particular, and has become a much studied experimental system. The experimental data tend to be very visually appealing (Figure 1). Many variations on the basic setup have been explored and a voluminous literature exists. The degree of structure of the two-fluid interface depends sensitively on the ratio of viscosities of the two fluids in the Hele-Shaw cell, and the pattern is sensitive to small perturbations, such as imperfections on either plate or tiny air bubbles resident in the cell.

The scale of the smallest fingers observed in a two-fluid Hele-Shaw experiment is set by the surface tension at the interface between the two immiscible fluids. When the surface tension is very small, apparently fractal patterns emerge.

HASSAN AREF

See also **Fractals; Navier–Stokes equation; Pattern formation; Rayleigh–Taylor instability**

Further reading

Hele-Shaw, J.H.S. 1898. The flow of water. *Nature*, 58: 34–36

Lamb, H. 1932. *Hydrodynamics*, 6th edition, Cambridge: Cambridge University Press

Saffman, P.G. & Taylor, G.I. 1958. The penetration of a fluid into a porous medium or Hele-Shaw cell containing a more viscous liquid. *Proceedings of the Royal Society of London* A, 245: 312–329

Wooding, R.A. & Morel-Seytoux, H.J. 1976. Multiphase fluid flow through porous media. *Annual Reviews of Fluid Mechanics*, 8: 233–274

HELICAL WAVES

See **Nonlinear plasma waves**

HELICITY

Intuitively, it is clear that a vortex having a component of velocity along its axis is a helical structure. Many structures fall under this category, including such diverse structures as Taylor–Gortler vortices, leading edge and trailing vortices shed from wings and slender bodies, streamwise vortices in boundary layers and free shear flows, Langmuir circulations in the ocean and analogous structures in the atmosphere, tornados, and rotating storms. Similar structures are observed in space and in laboratory plasmas. Such structures lack reflection symmetry. One of the key quantities characterizing reflection symmetry (or its lack) of a fluid flow is the so-called helicity, which roughly measures how parallel velocity and vorticity are in a fluid flow.

Definition

Helicity $\mathcal{H}_B$ of a solenoidal vector field $\boldsymbol{B}$ ($\mathrm{div}\boldsymbol{B}=0$) in a domain D within any surface S on which $\boldsymbol{n}\cdot\boldsymbol{B}=0$ is defined as the integral

$$\mathcal{H}_B=\int_D \boldsymbol{A}\cdot\boldsymbol{B}\,\mathrm{d}V \qquad (1)$$

with $\boldsymbol{A}$ being the vector potential of $\boldsymbol{B}$. Here $\boldsymbol{B}=\mathrm{curl}\boldsymbol{A}$ and the quantity $\mathfrak{h}_B=\boldsymbol{A}\cdot\boldsymbol{B}$ is the helicity density of the field $\boldsymbol{B}$. Both $\mathcal{H}_B$ and $\mathfrak{h}_B$ are pseudoscalars: they change sign under the parity transformation, that is, under change from a right-handed to a left-handed frame of reference. Helicity as defined in Equation (1) is gauge invariant, that is, invariant under replacement of $\boldsymbol{A}$ by $\boldsymbol{A}+\mathrm{grad}\phi$, where ϕ is any single-valued scalar field. This type of invariance should not be confused with time invariance of $\mathcal{H}_B$ in a nondissipative medium (see below). It is important that $\mathcal{H}_B$ is gauge invariant only as defined in Equation (1) for volumes bounded by the field surfaces, that is, $\boldsymbol{n}\cdot\boldsymbol{B}|_s=0$. Generalizations of Equation (1) for open field structures (i.e., such that $\boldsymbol{n}\cdot\boldsymbol{B}|_s\neq 0$) are not gauge invariant. As a rule these generalizations are also not well-defined topologically. Another generalization is the cross-helicity for two solenoidal fields $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ defined as $\mathcal{H}_{B_1,B_2}=\int_D \boldsymbol{A}_1\cdot\boldsymbol{B}_2\,\mathrm{d}V$. For closed structures, that is, $\boldsymbol{n}\cdot\boldsymbol{B}_{1,2}|_s=0$, cross-helicity is gauge invariant and is symmetric, $\mathcal{H}_{B_1,B_2}=\mathcal{H}_{B_2,B_1}$. Of special interest in physics is a magnetic field, $\boldsymbol{B}$, in electrically conducting fluids or plasmas (with $\mathcal{H}_B$ termed magnetic helicity) and vorticity $\boldsymbol{\omega}$ in fluid flows with $\mathcal{H}_\omega=\int_D \boldsymbol{u}\cdot\boldsymbol{\omega}\,\mathrm{d}V$ (kinetic helicity), where $\boldsymbol{u}$ is the velocity and $\boldsymbol{\omega}=\mathrm{curl}\,\boldsymbol{u}$.

Geometrical and Topological Meaning

Helicity is important at a fundamental level in relation with the linkage(s) of field lines, such as magnetic lines or vortex lines of fluid flow (see Figure 1). The simplest example of two linked vortex tubes is shown in Figure 1 with fluxes $\Phi_{1,2}$ associated with each tube, that are equal to their circulations $\kappa_{1,2}$. Hence, $\mathcal{H}_B$ is directly related to the most basic topological invariant of two

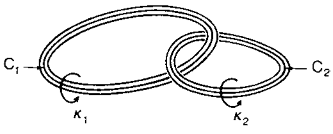

Figure 1. Prototype configuration of linked field tubes with nonzero helicity (Moffatt & Tsinober, 1992). Here the field is assumed to be identical to zero except in two closed tubes with axes $C_{1,2}$ and vanishingly small cross section. The field lines are untwisted within each tube, i.e. each line is a closed curve passing once round the tube, and unlinked with its neighbors in the same tube. In such a case $H_B=\pm 2L_{1,2}F_1F_2$, where $L_{1,2}$ is the linking (or winding) number of the two tubes (in the figure $L_{1,2}=1$), $F_{1,2}$ are the fluxes (equal to their circulations $\kappa_{1,2}$) associated with each tube, and the + or − sign is chosen depending on whether the linkage is right- or left-handed.

linked curves—their linking (winding) number $L_{1,2}$, which was defined by Gauss in 1833 as

$$L_{12} = L_{21} = \frac{1}{4\pi} \oint_{C_1} \oint_{C_2} \frac{\boldsymbol{R} \cdot (\mathrm{d}\boldsymbol{x}_1 \times \mathrm{d}\boldsymbol{x}_2)}{R^3}, \qquad (2)$$

where $\boldsymbol{R} = \boldsymbol{x}_1 - \boldsymbol{x}_2$, $\boldsymbol{x}_1 \varepsilon C_1$, and $\boldsymbol{x}_2 \varepsilon C_2$. In the general case of an infinite number of field lines, one can approximate the field via a finite number of N of small flux tubes each with flux Φ_i ($i = 1, \ldots, N$). Then $\mathcal{H}_B = L_{ij} \Phi_i \Phi_j$, where L_{ij} is the linking number of tubes i and j, and summation is assumed over the repeated indices. In other words, the helicity $\mathcal{H}_B$ can be interpreted as the sum of linking numbers of all pairs of the field lines weighted by the flux. This interpretation remains valid also in the case when the field lines are not closed upon themselves and even may wander chaotically, and the integral of Equation (1) is interpreted as the asymptotic Hopf invariant, that is, asymptotic linking number. The cross-helicity $\mathcal{H}_{B,\omega}$ is interpreted as a measure of mutual linkage of the two fields $\boldsymbol{B}$ and $\boldsymbol{\omega}$.

Invariance in Nondissipative Media

The magnetic field, $\boldsymbol{B}$, in a perfectly conducting fluid and vorticity, $\boldsymbol{\omega}$, in an inviscid fluid is governed by the equations

$$\frac{\partial \boldsymbol{B}}{\partial t} = \mathrm{curl}(\boldsymbol{u} \times \boldsymbol{B}), \qquad (3)$$

$$\frac{\partial \boldsymbol{\omega}}{\partial t} = \mathrm{curl}(\boldsymbol{u} \times \omega), \qquad (4)$$

describing frozen-field transport of the vector fields $\boldsymbol{B}$ and $\boldsymbol{\omega}$. Under this nondissipative evolution both the magnetic helicity $\mathcal{H}_B$ and the helicity of vorticity $\mathcal{H}_\omega$ are conserved. Invariance of helicity is directly associated with invariance of the topology of $\boldsymbol{B}$ and $\boldsymbol{\omega}$.

In cases when the fluid flow is influenced by the presence of the magnetic field, that is, by the Lorenz force (the term curl($\boldsymbol{j} \times \boldsymbol{B}$) added to the left-hand side of Equation (4)), the helicity of vorticity $\mathcal{H}_\omega$ is not conserved whereas the magnetic helicity remains conserved. Another conserved quantity in this latter case is the cross-helicity $\mathcal{H}_{B,\omega} = \int \boldsymbol{u} \cdot \boldsymbol{B} \, \mathrm{d}V$. In principle, one can choose the gauge ϕ in such a way that the helicity density $\mathfrak{h}_B = \boldsymbol{A} \cdot \boldsymbol{B}$ is a Lagrangian invariant, that is, it is pointwise conserved along the paths of fluid particles and therefore for any fluid volume. Such a choice is possible both for magnetic field and for nonconducting fluid flows.

Dissipative effects (finite electrical resistivity and viscosity) are responsible for the reconnection of field lines and thus for the evolution of helicity. In such a case, helicity is either created or destroyed just as if $\boldsymbol{n} \cdot \boldsymbol{B} \neq 0$ on the boundary of the domain.

Kinetic Helicity and Turbulent Dynamo

Helicity plays an important role in the process of growth of magnetic fields in electrically conducting fluids due to fluid motion, called the dynamo process. In turbulent flows with nonzero kinetic helicity, there exists the so-called α-effect, the main ingredient of which is that the large-scale (mean) magnetic field results from currents induced by an electromotive force that is (roughly) proportional to the kinetic helicity $\mathcal{H}_\omega$. An important aspect is that this electromotive force is (generally) nonvanishing for nonvanishing electrical resistivity. There is a buildup of magnetic helicity of the generated large-scale magnetic field, and the importance of finite electrical resistivity is in the dissipation of the small scale (on the energy containing scales of fluid turbulence) magnetic helicity of the opposite sign that allows the build up of the large magnetic helicity (Brandenburg, 2001). Recent laboratory experiments with liquid sodium confirm the possibility of generating a magnetic field by helical fluid flows. The growing magnetic field reacts back on the fluid flow via the Lorenz force that results in a decrease of the electromotive force driving the electrical currents (α-quenching), that is, in a nonlinear saturation in the growth of the mean magnetic field.

Kinetic Helicity and Fluid Turbulence

Present knowledge of the effects of helicity on fluid turbulence is controversial. A difficulty in the assessment of the role of helicity in the dynamics of turbulence stems from the fact that—unlike kinetic energy—helicity is not a positively defined quantity. At a speculative level it is expected that nonvanishing helicity comprises a constraint reducing in some sense the nonlinear processes, and thereby turbulence. The simplest argument is that since $(\boldsymbol{u} \cdot \boldsymbol{\omega})^2 + (\boldsymbol{u} \times \boldsymbol{\omega})^2 = \boldsymbol{u}^2 \boldsymbol{\omega}^2$, larger $\langle |\boldsymbol{u} \cdot \boldsymbol{\omega}| \rangle$ implies reducing of nonlinear processes associated with $\langle |\boldsymbol{u} \times \boldsymbol{\omega}| \rangle$. Indeed, the so-called Beltrami flows with $\boldsymbol{\omega} = \boldsymbol{\lambda} u$, possessing in some sense maximal helicity, are linear. However, existing evidence weakly supports the above argument (Moffatt & Tsinober, 1992; Tsinober, 2001). On the other hand, if $\langle \boldsymbol{u} \cdot \boldsymbol{\omega} \rangle \neq 0$, this is a clear indication of direct and bidirectional coupling of large and small scales. Therefore, in flows with nonzero mean helicity, the coupling between small and large scales is stronger than otherwise. This stronger coupling may aid creation of large scale structures out of small scale turbulence—a process that is frequently called inverse energy cascade. There is little understanding of these processes, although there are suggestions that in the presence of kinetic helicity in compressible fluids there exists an effect (Hα-effect) analogous to the above-mentioned α-effect for a magnetic field.

More subtle issues include spontaneous breaking of relexional symmetry and emergence of helicity in initially helicity-free flows. The evidence is based on indications from stability analysis, laboratory observations and quantitative measurements of helicity, and direct numerical simulations of Navier–Stokes equations (Kholmyansky et al., 2001). Finally, kinetic helicity has a considerable effect on the effective diffusivity of a passive scalar that may be enhanced by the presence of helicity by more than a factor of two. In addition, helicity leads to a new "skew-diffusion" effect, that is, appearance of a component of turbulent flux of passive scalar perpendicular to the local mean gradient of a passive scalar.

Applications

In atmospheric physics, the so-called storm-relative environmental (kinetic) helicity (SREH) is routinely used in the U.S. with moderate success as a parameter for characterizing, interpreting, and forecasting tornado and supercell rotating storms (Weisman & Rotunno, 2000). The origin of these strucures is not understood, though there are suggestions that the Hα-effect (due to the Coriolis force) and intensive heat transport in turbulent convection are among the important physical mechanisms contributing to their formation (Levina et al., 1999). One has to add to this list at least one more factor, the vertical wind shear.

In plasma physics, magnetic helicity plays a central role in magnetic confinement and in the relaxation of plasmas to a minimum-energy state, for example, in toroidal laboratory plasmas and spheromaks. Interestingly, during this process involving reconnection due to small but finite resistivity and turbulence, the total magnetic helicity is approximately conserved and constrains the energy of the magnetized plasma in equilibrium. This minimum energy state is generally a force-free (Beltrami) field, curl $\boldsymbol{B} = \lambda \boldsymbol{B}$ with λ constant.

In astrophysics, magnetic helicity is important in a number of aspects in solar physics (coronal loops, solar wind), the astrophysical dynamo problem (Brown et al., 1999), and cosmology (primordial magnetic field).

ARKADY TSINOBER

See also **Alfvén waves; Magnetohydrodynamics; Nonlinear plasma waves; Topology; Vortex dynamics of fluids; Winding numbers**

Further Reading

Brandenburg, A. 2001. The inverse cascade and nonlinear alpha-effect in simulations of isotropic helical hydromagnetic turbulence. *Astrophysical Journal*, 550: 824–840

Brown, M.R., Canfield, R.C. & Pevtsov, A.A. 1999. *Magnetic Helicity in Space and Laboratory Plasmas*, Washington, DC: American Geophysical Union

Kholmyansky, M., Shapiro-Orot, M. & Tsinober, A. 2001. Experimental observations of spontaneous breaking of reflexional symmetry in turbulent flows, *Proceedings of the Royal Society of London*, 457: 2699–2717

Levina, G.V., Moiseev, S.S. & Rutkevich, P.B. 1999. Hydrodynamic alpha-effect in a convective system. In *Nonlinear Instability, Chaos and Turbulence*, edited by L. Debnath & D.N. Riahi, Southampton and Boston: WIT Press

Moffatt, H.K. & Tsinober, A. 1992. Helicity in laminar and turbulent flows. *Annual Reviews of Fluid Mechanics*, 24: 281–312

Ricca, R.L. (editor). 2001. *An Introduction to the Geometry and Topology of Fluid Flows*, Dordrecht and Boston: Kluwer

Tsinober, A. 1995. Variability of anomalous transport exponents versus different physical situations in geophysical and laboratory turbulence. In *Levy Flights and Related Topics in Physics*, edited by M. Schlesinger, G. Zaslavsky & U. Frisch, Berlin and New York: Springer

Tsinober, A. 2001. *An Informal Introduction to Turbulence*, Dordrecht and Boston: Kluwer

Weisman, M. & Rotunno, R. 2000. The use of vertical wind shear versus helicity in interpreting supercell dynamics. *Journal of Atmospheric Sciences*, 57: 1452–1472

HÉNON MAP

By the mid-1970s, examples such as the Lorenz equations had convinced researchers that strange attractors could arise in differential equations modeling physical systems. Unfortunately, the length of time needed to compute solutions coupled with strong contraction rates made it very difficult to observe fractal structures numerically in these examples with the computers then available. The Hénon map provided the first simple equation in which this fractal structure is easily observed.

Michel Hénon's approach was based on the idea of return maps. The dynamics of differential equations can be modelled by invertible maps, and so evidence of fractal structure in the attractor of an invertible map shows that these objects can exist in differential equations. The Hénon map is a very simple nonlinear difference equation

$$\begin{aligned} x_{n+1} &= y_n + 1 - a x_n^2, \\ y_{n+1} &= b x_n, \end{aligned} \tag{1}$$

where the parameters a and b were chosen as $a = 1.4$ and $b = 0.3$ by Hénon (1976), although other values are also interesting. The attractor, together with some blowups of parts of the attractor, are shown in Figure 1. These pictures are very easy to generate: successive points on an orbit are obtained by evaluating the algebraic expressions on the right-hand side of (1) as the following rough MATLAB program shows:

```
x(1)=0.3; y(1)=0.2;
  %% Initial conditions
N=5000;
  %% N is the number of iterates
a=1.4;b=0.3;
  %% Sets the parameters
for i=1:N
```

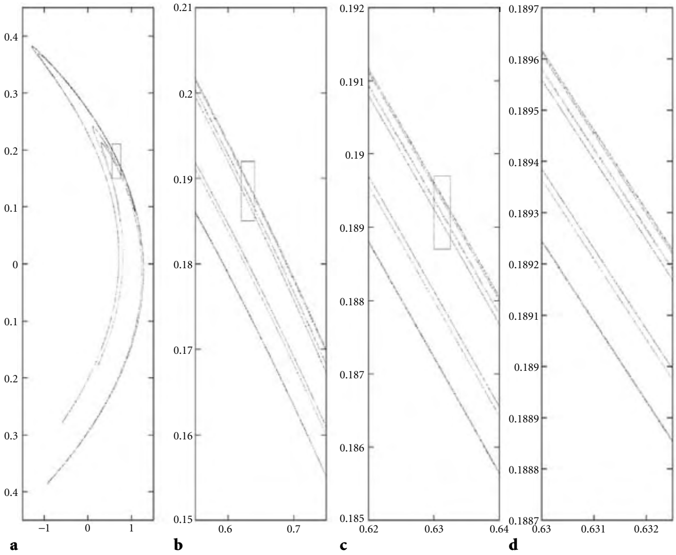

Figure 1. (a) Numerically computed attractor of the Hénon map, (1) with $a = 1.4$ and $b = 0.3$; (b), (c), and (d) are blowups of the boxed regions of (a), (b), and (c), respectively, showing the fractal structure of the attractor. Each figure contains the first 5000 points to land in the displayed region.

```
  %% Begin the iteration loop here
x(i+1)=y(i)+1-a*x(i)^2;
y(i+1)=b*x(i);
  %% That calculated the next
  point
end
plot(x,y,’k.’)
  %% Plot the iterates in the
  %% (x,y) plane
```

Despite the simplicity of this program, Hénon used a mainframe (IBM 7040) to perform the 5 million iterates he needed to get a reasonable number of points in the equivalent of Figure 1(d). Figure 1 uses a slightly more sophisticated program to generate the attractor and zoom in on the rectangular regions indicated so that 5000 points can be plotted in each of the blowup regions. This involved 35,724,657 iterations of the map order to get 5000 points in the smallest blowup region of Figure 1(d). This level of computation would have been almost unthinkable when Hénon wrote his paper.

The Hénon attractor pictured in Figure 1 is computationally cheap, requiring no more than the most simple algebraic operations. Also, the numerical evidence for fractal structure in the attractor is sufficiently convincing that most researchers have come to accept that it is a strange attractor or, at least, to suspend their disbelief. For this reason it has become a canonical example of chaotic motion. Almost every new technique or relevant theoretical result is applied to the Hénon map as part of the evaluation of the method. Early papers on phase space reconstruction, dimension calculations, chaotic prediction, chaotic control and synchronization, periodic orbit expansions, invariant measure algorithms, etc., have all used the Hénon map as an important test example. Given this general level of acceptance, it may come as a surprise to learn that it is still not known whether there really is a strange attractor for the Hénon map at the standard parameter values ($a = 1.4$, $b = 0.3$).

Hénon (1976) gave a number of reasons for looking at orbits of (1):

> ... we try to find a model problem which is as simple as possible, yet exhibits the same essential properties as the Lorenz system. Our aim is (i) to make the numerical exploration faster and more accurate...; (ii) to provide a model which might lend itself more easily to mathematical analysis.

As we have seen, Hénon’s aim of making the numerical exploration of apparently chaotic attractors more straightforward succeeded spectacularly. However, he could not have imagined how hard it would be to answer the theoretical questions posed by this deceptively simple map.

Hénon's intuitive explanation for his map in terms of folding, stretching, and contraction is much closer to the formation of horseshoes rather than to the Lorenz model, which has discontinuities in the natural return map. As such, the Hénon map has become a paradigm for the formation of horseshoes as parameters vary (see Devaney & Nitecki, 1979), and it is the more general question of how the attractors of the Hénon map change as the parameters vary that has occupied most theoretical approaches.

By defining a new y variable $y_{\mathrm{new}} = b^{-1} y_{\mathrm{old}}$, Equation (1) can be written in the form of a more general, Hénon-like map:

$$\begin{aligned} x_{n+1} &= -\varepsilon y_n + f_a(x_n), \\ y_{n+1} &= x_n, \end{aligned} \tag{2}$$

where $\varepsilon = -b$ and $f_a(x) = 1 - ax^2$ gives the Hénon map in the new coordinates. This formulation emphasizes the relationship between Hénon-like maps and one-dimensional maps: if $\varepsilon = 0$ then the x equation decouples and x evolves according to the one-dimensional difference equation $x_{n+1} = f_a(x_n)$, which in the original case, (1), is just the standard quadratic family. The Jacobian of the map is ε, so positive ε corresponds to orientation-preserving maps, which is more natural in the context of return maps, although this means that $b < 0$ in Hénon's original formulation, (1). Early efforts toward proving that strange attractors exist in the Hénon map concentrated on extending results for one-dimensional maps to the two-dimensional case with $\varepsilon > 0$ small. On the negative side, Holmes & Whitley (1984) showed that however small ε is, some periodic orbits of the Hénon map appear in a order different from the order in which they appear in the quadratic family. On the positive side, Gambaudo, van Strien, & Tresser (1989) showed that for sufficiently small $\varepsilon > 0$, the first complete period-doubling cascade is associated with the original period two orbit.

The major breakthrough on the existence of strange attractors was made by Benedicks & Carleson (1991). Using delicate mathematical analysis, they were able to show that if $\varepsilon > 0$ is small enough and a is close to $a = -2$ (the equivalent of $\mu = 4$ for the standard formulation of the quadratic map, $\mu x(1-x)$), then there is a positive measure of parameter values for which the Hénon map has a strange attractor. This result was generalized by Mora & Viana (1993) who showed that Hénon-like maps arise naturally near homoclinic bifurcations of maps. It had long been recognized that these bifurcations occur in the Hénon map (see, for example, Holmes & Whitley (1984)), so this made it possible to deduce the existence of strange attractors at values of ε that are not small. Indeed, this important paper provides a method of proving the existence of strange attractors for a set of parameter values with positive measure in a wide variety of model systems. Despite all these advances, these results only prove that there exist such parameter values and they do not give methods for proving that a strange attractor exists at a *given* parameter value.

Two other avenues of research suggested by the one-dimensional, $\varepsilon = 0$, limit in (2) have led to interesting developments. We have already noted that the one-dimensional order of periodic orbits is not preserved if $\varepsilon > 0$. However, a beautiful theory of partial orders based on period and knot type has emerged, which shows that the existence of some periodic orbits implies the existence of some others in two-dimensional maps (see Boyland (1994) for more details). The second adaptation of one-dimensional approaches is based on the idea of the symbolic dynamics (or *kneading theory*) of unimodal maps. The main idea, introduced by Cvitanovic et al. (1988) and developed by de Carvalho (1999), is to produce symbolic models of the dynamics of the Hénon map by relating the dynamics to modifications of the full horseshoe. This is done by *pruning* the horseshoe, that is, identifying regions of the horseshoe with dynamics that are not present in the Hénon map under consideration and judiciously removing these regions together with their images and preimages, leaving a pruned horseshoe that can still be accurately described.

Much of the current interest in the Hénon map involves the existence and construction of invariant measures for the attractors. This work should lead to a good statistical description of properties of orbits and averages along orbits. So, even now, this simple two-dimensional map with a single nonlinear term is motivating important questions in dynamical systems.

Paul Glendinning

See also **Attractors; Bifurcations; Chaotic dynamics; Horseshoes and hyperbolicity in dynamical systems; Markov partitions; One-dimensional maps; Routes to chaos; Sinai–Ruelle–Bowen measures; Symbolic dynamics**

Further Reading

Benedicks, M. & Carleson, L. 1991. The dynamics of the Hénon map. *Annals of Mathematics*, 133: 73–169

Boyland, P. 1994. Topological methods in surface dynamics. *Topology and Its Applications*, 58: 223–298

de Carvalho, A. 1999. Pruning fronts and the formation of horseshoes. *Ergodic Theory and Dynamical Systems*, 19: 851–894

Cvitanovic, P., Gunaratne, G. & Procaccia, I. 1988. Topological and metric properties of Hénon-type strange attractors. *Physical Review* A, 38: 1503–1520

Devaney, R. & Nitecki, Z. 1979. Shift automorphisms in the Hénon mapping. *Communications in Mathematical Physics*, 67: 137–146

Gambaudo, J.-M., van Strien, S. & Tresser, C. 1989. Hénon-like maps with strange attractors: there exist C^{∞}

Kupka–Smale diffeomorphisms on $\mathbf{S}^2$ with neither sinks nor sources. *Nonlinearity*, 2: 287–304

Hénon, M. 1976. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics*, 50: 69–77

Holmes, P. & Whitley, D. 1984. Bifurcations of one- and two-dimensional maps. *Philosophical Transactions of the Royal Society (London)* A, 311: 43–102

Mora, L. & Viana, M. 1993. Abundance of strange attractors. *Acta Mathematica*, 171: 1–71

HÉNON–HEILES SYSTEM

Although the molecules of a gas move freely and collide frequently, the stars in a galaxy move under the constraint of long-range interactions and do not collide. The motion of any one star in a galaxy can be modeled as if it moves alone under the constraint of a fixed gravitational galactic potential with cylindrical symmetry. The energy of the star is conserved as is the angular momentum with respect to an axis through the center of the galactic plane and at right angles to it. These two conserved quantities are isolating integrals of the motion. In the absence of a third isolating integral, the distribution of stellar velocities in a galaxy with an axisymmetric potential should be circularly symmetric in the meridian plane, a result contrary to observation (Ollongren, 1965).

The astronomer Michel Hénon and a graduate student Carl Heiles introduced a mathematical model to numerically investigate the third integral problem when Hénon visited Princeton University in 1962. In "order to have more freedom of experimentation" in their numerical studies, they considered the simplified model system, with two degrees of freedom, described by the Hamiltonian (Hénon & Heiles, 1964)

$$H = \frac{1}{2}\left(\dot{x}^2 + \dot{y}^2\right) + \frac{1}{2}\left(x^2 + y^2\right) + x^2 y - \frac{1}{3}y^3,$$

where the last three terms are called V. The choice of coefficients in the cubic potential confers a ternary symmetry on the equipotential lines. The equipotential line for $V = \frac{1}{6}$ is an equilateral triangle, and this is the limiting energy at which the motion is bounded. In the Hénon–Heiles model, the original problem of the third isolating integral is transformed into whether there exists an isolating integral other than the constant energy itself. The Hénon–Heiles Hamiltonian also describes the center of mass reduction of a three-particle periodic chain with linear and quadratic nearest neighbor elastic forces and is, thus, a member of the class of systems investigated in the numerical studies of Fermi, Pasta, and Ulam (Ford, 1992). More generally, the Hénon–Heiles Hamiltonian is one of the simplest non-integrable Hamiltonian systems known. It has become a paradigm for studies of chaos and non-integrability in Hamiltonian systems with two degrees of freedom.

The solutions of the equations of motion corresponding to the Hénon–Heiles Hamiltonian can be represented by trajectories in the four-dimensional phase space spanned by $(x, \dot{x}, y, \dot{y})$. If the energy is fixed in the range $0 \le E \le \frac{1}{6}$ then these trajectories are bounded and recurrent on a three-dimensional energy surface. No techniques are known to construct exact algebraic formulae for the trajectories for arbitrary initial conditions. Approximate solutions can be obtained using perturbation methods or numerical methods. In their numerical studies, Hénon and Heiles plotted the points of intersection between phase-space trajectories and a two-dimensional surface of section in the phase-space. For recurrent trajectories, successive intersection points on the surface of section define an area-preserving two-dimensional discrete time dynamical system or map that shares the important dynamical features of the original system (Poincaré, 1993).

A convenient Poincaré map for the Hénon–Heiles Hamiltonian at fixed energy, in the range $0 \le E \le \frac{1}{6}$, is obtained with the surface of section $x = 0$ and $\dot{x} \ge 0$. The coordinate pair $(y, \dot{y})$ in this case uniquely determines an initial point for a trajectory since $\dot{x} = \sqrt{2E - \dot{y}^2 - y^2 + 2y^3/3}$. Thus, the dynamical description is reduced to a study of the Poincaré map from one intersection point $(y_n, \dot{y}_n)$ to the next intersection point $(y_{n+1}, \dot{y}_{n+1})$. The existence of an additional isolating integral in the Hénon–Heiles Hamiltonian would restrict a trajectory to intersect an invariant curve in the surface of section. In the absence of an additional isolating integral, the pattern of intersection points should fill an area on the surface of section. In numerical studies, where the number of intersection points is finite, discrimination between these two signatures is subjective since any finite set of points can be aligned to a smooth curve. Despite this, the interpretation of the intersection points as invariant curves or filled area in the Hénon–Heiles study has been regarded as unambiguous.

Figure 1 shows the intersection points between numerical phase space trajectories and a Poincaré surface of section in the Hénon–Heiles system at four different values of the energy, $E = \frac{1}{24}, \frac{1}{12}, \frac{1}{8}$, and $\frac{1}{6}$.

At $E = \frac{1}{24}$ and $E = \frac{1}{12}$, the intersection points appear to lie on invariant curves. At $E = \frac{1}{8}$, intersection points for some trajectories appear to lie on invariant curves whereas intersection points for other trajectories explore areas on the surface of section. At $E = \frac{1}{6}$, intersection points for most trajectories do not appear to lie on invariant curves (and typically the separation distance between trajectories started from nearby initial conditions diverges exponentially in time, a signature of chaos). A feature that is not revealed in the surface of section portraits at $E = \frac{1}{6}$, but is clear in three-dimensional phase space portraits (Henry & Grindlay, 1994), is intermittency whereby a single trajectory switches in time between regular behavior (close to invariant curves) and irregular behavior (filling areas).

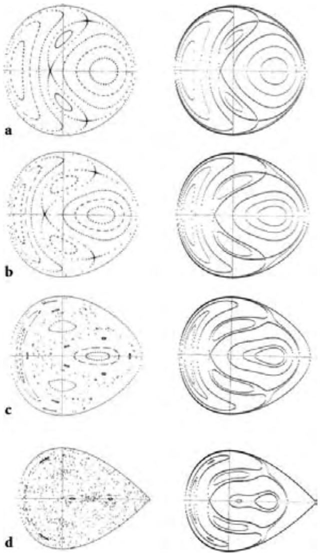

Figure 1. Surface of section portraits for the Hénon–Heiles Hamiltonian at four different values of the energy; (a) $E = 1/24$, (b) $E = 1/12$, (c) $E = 1/8$, (d) $E = 1/6$. The figures on the left are from numerical integrations and the figures on the right are from Birkhoff–Gustavson normal form theory. The figures have been reproduced from Gustavson (1966).

The transition from regular behavior to irregular behavior as the system energy is increased in the Hénon–Heiles model is characteristic of a more general transition from regular behavior to irregular behavior in Hamiltonian systems with two degrees of freedom and in two-dimensional area-preserving maps. A detailed theoretical understanding of this phenomena encompasses collective results from the Birkhoff–Gustavson normal form theory, the Kolmogorov–Arnol'd–Moser (KAM) theory, the Poincaré–Birkhoff theory, the Aubry–Mather theory, the theory of heteroclinic tangles (Poincaré, 1993), and the theory of overlapping resonances (Ford, 1992).

The Birkhoff–Gustavson normal form (Gustavson, 1966) represents the Hamiltonian as an infinite series of harmonic oscillator Hamiltonians through an infinite series of successive canonical transformations. The Hénon–Heiles Hamiltonian is strictly non-integrable at all values of the system energy (Ito, 1985); thus the normal form series for this Hamiltonian does not converge. On the other hand, any finite truncation of the normal form series constitutes an integrable Hamiltonian that can be considered to be "close" in some sense to the Hénon–Heiles Hamiltonian. The additional conserved quantity for an integrable normal form Hamiltonian can be readily calculated and used to plot the associated invariant curves on the Poincaré surface of section. The KAM theory proves that for a non-integrable Hamiltonian sufficiently close to an integrable Hamiltonian, most of the phase space trajectories are confined to invariant tori, similar to the invariant tori of the integrable system. Thus the surface of section portraits for the Hénon–Heiles Hamiltonian should map out the invariant curves from the truncated normal form Hamiltonian whenever this Hamiltonian is sufficiently close to the Hénon–Heiles Hamiltonian. The difference between the truncated normal form series and the Hénon–Heiles Hamiltonian increases with system energy. This can be seen in Figure 1 where the invariant curves from an eighth order truncation (Gustavson, 1966) are compared side by side with the numerical results for the Hénon–Heiles Hamiltonian.

In Hamiltonian systems with two degrees of freedom, the tori are characterized by a winding number ω_1/ω_2. The KAM theory proves the existence of invariant tori whose winding numbers are sufficiently irrational that

$$\left|\frac{\omega_1}{\omega_2} - \frac{r}{s}\right| > \frac{k(\varepsilon)}{s^{5/2}}$$

where r and s are co-prime integers and $k(\varepsilon)$ approaches zero as ε approaches zero. Here ε measures the strength of the non-integrable perturbation on the integrable Hamiltonian. The volume of phase-space not filled by invariant tori approaches zero as ε approaches zero. Tori whose winding numbers do not satisfy the stability condition under perturbation breakup according to the Poincaré–Birkhoff theory (rational winding numbers) or the Aubry–Mather theory (irrational winding numbers). The numerical results for the Hénon–Heiles Hamiltonian are consistent with an increasing breakup of tori with increasing energy as less and less tori satisfy the stability criteria.

The motion of an asteroid around the Sun can also be described approximately by a simplified model Hamiltonian with two degrees of freedom. The simplest nontrivial model is a plane circular restricted three-body problem involving the asteroid, the Sun, and Jupiter (Berry, 1978). The two-body motion of the asteroid around the Sun is a Kepler ellipse with frequency ω_{A}, and the two-body motion of Jupiter about its center of mass with the Sun is taken to be a circle with frequency ω_{J}. In the restricted three-body problem the effect of the asteroid on the motion of the Sun and Jupiter is neglected, the motion of the asteroid is dominated by the Sun, and the motion of Jupiter is considered to be a perturbation on the motion of the asteroid. The Hamiltonian for this restricted three-body problem is non-integrable, and an application of KAM theory to the problem identifies the frequency ratio $\omega_{\mathrm{A}}/\omega_{\mathrm{J}}$ with the winding number for invariant tori. Thus, asteroids should be expected to be found in stable orbits at distances from the Sun where ω_A/ω_J is sufficiently irrational. Indeed, as had been noted by Daniel Kirkwood in 1866, there is an abundance of asteroids at these locations but gaps in the asteroid belt at locations where the frequency ratio is 3:1, 5:2, 7:3, or 2:1; however a detailed understanding of the Kirkwood gaps is still an area of active research (Ferraz-Mello, 1999).

Bruce Henry

See also **Kolmogorov–Arnol'd–Moser theorem; N–body problem; Normal forms theory; Poincaré theorems; Solar system**

Further Reading

Berry, M.V. 1978. Regular and irregular motion. In *Topics in Nonlinear Dynamics*, edited by S. Jorna, New York: American Institute of Physics, pp.16–120

Contopoulos, G. 1963. On the existence of a third integral of motion. *The Astronomical Journal*, 68(1): 1–14

Ferraz-Mello, S. 1999. Slow and fast diffusion in asteroid-belt resonances: a review. *Celestial Mechanics & Dynamical Astronomy*, 73(1–4): 25–37

Ford, J. 1992. The Fermi–Pasta–Ulam problem: paradox turns discovery. *Physics Reports*, 213(5): 271–310

Gustavson, F.G. 1966. On constructing formal integrals of a Hamiltonian system near an equilibrium point. *The Astronomical Journal*, 71(8): 670–686

Gutzwiller, M.C. 1990. *Chaos in Classical and Quantum Mechanics*, New York: Springer

Hénon, M. & Heiles, C. 1964. The applicability of the third integral of motion: some numerical experiments, *The Astronomical Journal*, 69(1): 73–99

Henry, B.I. & Grindlay, J. 1994. From dynamics to statistical mechanics in the Hénon–Heiles model: dynamics. *Physical Review* E, 49(4): 2549–2558

Ito, H. 1985. Non-integrability of Hénon-Heiles system and a theorem of Ziglin. *Kodai Mathematical Journal*, 8: 120–138

Ollongren, A. 1965. Theory of stellar orbits in the galaxy. *Annual Review of Astronomy and Astrophysics*, 3: 113–134

Poincaré, H. 1993. *New Methods of Celestial Mechanics, vol. 3. Integral Invariants and Asymptotic Properties of Certain Solutions*. Woodbury, New York: American Institute of Physics. (English translation of *Les Méthodes nouvelles de la Mécanique Céleste, vol. 3: Invariants integraux. Solutions periodiques du deuxieme genre Solutions doublement asymptotiques*, 1899)

HETEROCLINIC INTERSECTION

See **Phase space**

HETEROCLINIC TRAJECTORY

See **Phase space**

HIERARCHIES OF NONLINEAR SYSTEMS

Although we must deal with many hierarchical structures in the course of life—social, military, and monetary, among others—two are of particular interest to practitioners of nonlinear science: the biological and cognitive hierarchies.

Biological Hierarchy

It is an empirical fact that living creatures are hierarchical in structure, organized according to a scheme that is roughly as follows.

Biosphere
Species
Organisms
Organs
Cells
Processes of replication
Genetic transcription
Biochemical cycles
Biomolecules
Molecules
Atoms

Several comments are in order. First, it is only the general nature of this hierarchy that is of interest here, not the details. Second, the nonlinear dynamics at each level of description generate emergent structures, and nonlinear interactions among these structures provide a basis for the dynamics at the next higher level (Scott, 2003). Thus, molecules emerge from the nonlinear forces among their constituent atoms, biomolecules emerge from molecules, biochemical cycles (such as the Krebs cycle, which processes energy in a living cell) emerge from interactions among biomolecules, and so on. Third, the emergence of a new dynamic entity usually involves presence of a closed causal loop, which leads to positive feedback and exponential growth that is ultimately limited by nonlinear effects.

Finally, it should be noted that philosophers disagree about the ontological status of an emergent entity. Are the levels mere designations, convenient for academic organization, or do they mark qualitatively different realms of reality?

In attempting to answer this question, it is important to know whether the upper levels of the biological hierarchy can be derived from lower levels, which leads to the concept of reductionism. In other words, can the upper levels of the hierarchy (organisms, species, biosphere) be logically derived from the properties of the lower levels? Or are they in some sense independent? In response to such questions, condensed matter physicist Philip Anderson has commented (Anderson, 1972): “The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe.”

Why not? Like the flame of a candle, the Belousov–Zhabotinsky reaction, or a nerve impulse, the dynamics of biological systems arise from nonlinear reaction-diffusion processes, in which the particular atoms participating in the reactions change with time. Thus, exact knowledge of the positions and velocities of the constituent atoms of an organism at one time tells little about the behavior at a later time (Bickhard & Campbell, 2000; Scott, 2002). Of course, a dedicated reductionist would respond that the explanatory net has not been cast far enough—all the atoms and radiation that wander into and out of the organism must be followed—but few seriously propose to do this as it would require computing in detail the dynamics of the entire universe.

Cognitive Hierarchy

Just as the phenomenon of life arises from the hierarchical nature of the biological realm, it has been suggested that the mysteries of mind may stem from an analogous hierarchy, describing levels of activity of the brain. Again ignoring details, this cognitive hierarchy may be sketched as follows.

Human cultures
Phase sequences
Complex assemblies
. . .
Assemblies of assemblies of assemblies
Assemblies of assemblies
Assemblies of neurons
Neurons
Nerve impulses
Nerve membranes
Membrane proteins
Molecules
Atoms

To the previous comments on the biological hierarchy, the following can be added. Although the biological levels are observed—through x-ray diffraction, via electron and light microscopy, or more directly—several levels of the cognitive hierarchy are theoretical constructs. This is particularly so for the cell assemblies and their constituent subassemblies, which Donald Hebb proposed to underlie the dynamic nature of human thought (Hebb, 1949). Thus a particular complex cell assembly that embodies a particular idea (your memory of your grandmother, for example) might involve interactions among several (visual, auditory, conceptual) subassemblies as suggested by Figure 1.

In this figure, the shaded areas do not suggest that all neurons in that region are firing, but only those (perhaps ten thousand or so) that have become specific to this particular memory through many related experiences and interactions. The neurons of this assembly are interconnected in such a manner that they are "capable of interacting briefly as a closed system, delivering facilitation to other such systems and usually having a specific motor facilitation" (Hebb, 1949). As a complex assembly comprises subassemblies that in turn are composed of yet more basic constituent assemblies, the concept is inherently hierarchical.

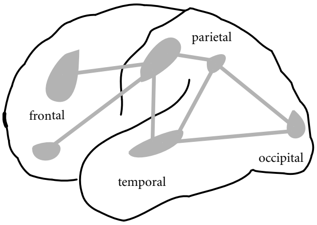

Figure 1. A sketch of the left side of the human brain, suggesting how the active subassemblies (shaded areas) of a complex cell assembly might be distributed over various lobes of the neocortex.

Although Hebb's theory is in accord with a substantial amount of psychological data (Hebb, 1980; Scott, 2002), it is difficult to imagine placing electrodes in the neocortex so that the simultaneous firings of a substantial number of these assembly neurons could be recorded. Nontheless, a substantial amount of experimental and numerical data is accumulating in support of this theory.

At higher levels of the cognitive hierarchy, we recognize the phase sequence, in which the focus of thought moves from one complex assembly to the next in a train of thought or a dream. At yet higher levels, important aspects of cultural dynamics—ingrained social prejudices, for example, or an aversion to eating certain foods—may remain unrecognized as they play important roles in human behavior.

ALWYN SCOTT

See also **Cell assemblies; Electroencephalogram at large scales; Emergence; Neural network models**

Further Reading

Anderson, P.W. 1972. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177: 393–396

Bickhard, M.H. & Campbell, D.T. 2000. Emergence. In *Downward Causation: Minds, Bodies and Matter*, edited by P.B., Andersen, C. Emmeche, N.O. Finnemann & P.V. Christiansen, Aarhus, Denmark: Aarhus University Press

Hebb, D.O. 1949. *Organization of Behavior: A Neuropsychological Theory*, New York: Wiley

Hebb, D.O. 1980. The structure of thought. In *The Nature of Thought*, edited by P.W. Jusczyk & R.M. Klein, Hillside, NJ: Lawrence Erlbaum Associates, pp. 19–35

Scott, A.C. 1995. *Stairway to the Mind*, New York: Springer

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

HIGGS BOSON

Particle physics is devoted to discovering and investigating the fundamental constituents of all the matter in the universe (at all times), and the forces of nature by which these constituents interact with each other. The theoretical framework that currently aids our understanding is called the Standard Model. Although there is no experimental result that significantly disagrees with the Standard Model, it is not a complete theory; indeed, there are several reasons why the Standard Model is not a theory at all in the normal sense of the word. Perhaps the most important of these is that it contains a large number of parameters (such as the fundamental particle masses) that are not predicted but determined experimentally and then incorporated into the theoretical framework.

A principle mission of particle physics is the development of a complete, fundamental theory that unifies all of the known forces into one. Historically, this process seems to be taking place one pair of forces at a time (for example, James Clerk Maxwell famously unified electricity and magnetism in the 19th century), and an important feature of the Standard Model is that it contains the unification of the short-range weak nuclear force and the long-range electromagnetic force. The strong nuclear force is not yet unified and gravity is completely excluded from the standard model.

Although the successful implementation of electroweak unification encourages particle physicists to consider the Standard Model as a low energy approximation of a more fundamental theory, electroweak unification comes at a price. The quantum of the electromagnetic field is the photon, which has zero rest mass. The weak field has a chargeless quantum, known as the Z particle, and a charged quantum, known as the W particle. The rest masses of the Z and the W are both almost 100 times the proton rest mass. This huge difference in the masses of the particles associated with the electromagnetic and weak force fields is known as electroweak symmetry breaking, and in order to interpret it, physicists have introduced into the Standard Model an arbitrary mathematical trick.

In a series of papers some 40 years ago, Englert and Brout (and independently, Higgs) pointed out that a symmetry-breaking idea used in the theory of superconductivity could also be used to break the electroweak symmetry. This mathematical trick is known in particle physics as the Higgs mechanism. In the theory of superconductivity, the idea explains how photons appear to acquire mass inside a superconductor. Introducing it into the Standard Model provides an explanation of where the Z and W masses come from.

The Higgs mechanism is most readily viewed in terms of a constant, ubiquitous energy field that has no preferred direction in space. This field is known as the Higgs field. It is this Higgs field that breaks the electroweak symmetry, giving mass to both Z and W while retaining a massless photon. Just as there are well-defined quanta, known as bosons, associated with each of the known force fields, so it is with the hypothetical Higgs field. In the simplest version of the Higgs mechanism, there is just one quantum and it is known as the Higgs particle or Higgs boson.

So, to recap, the Higgs particle is intimately associated with the Higgs field, and the Higgs field is assumed to be responsible for the mass difference between the quanta of the electromagnetic and weak fields. It is not too much of a stretch to expect that the masses of all matter particles (quarks and leptons) somehow get their masses from the Higgs mechanism. Indeed, the coupling of every particle to the Higgs field has strength proportional to its mass, and so this interpretation seems natural. Unfortunately, the Higgs mechanism does not reduce the number of unknown parameters of the Standard Model—there is still one per particle, plus a few more.

Curiously, the assumed Higgs particle has quite well-determined properties. Its mass, like all other particle masses, is not predicted but must be measured. As a function of an assumed mass, however, one can use the Standard Model to calculate the rate at which it will be produced in a high-energy collision of two elementary particles. The Higgs particle is expected to be highly unstable, and it is also straightforward to calculate, using the Standard Model, how it will decay. In other words, what it will decay into and with what relative rates. These Standard Model predictions provide a basis for experimental searches for the Higgs particle.

The particle experimentalists perform experiments in which they collide, at the highest available energy, protons with protons, or protons with antiprotons, or electrons with positrons. They then examine the debris of the collision for evidence of a Higgs particle whose properties match those suggested by the Standard Model.

In this manner, the experimental techniques of particle physics have been able, over the past decade or so, to eliminate the possibility of a Higgs particle with mass below about 120 proton masses or above about 200 proton masses. When the Higgs mechanism was first introduced to particle physics, the mass of the Higgs particle could have been anywhere between zero and about 1000 times the proton mass. One of the remarkable achievements of particle physics over the past few decades has been the elimination of most of this mass region. Of course, this elimination is of a statistical nature and the best thing that physicists can do is to assign a measure of the statistical confidence to these limits. In late 2002, most particle physicists were confident that if the Standard Model Higgs particle exists, its mass is probably in the mass range of approximately 120–200 proton masses (somewhere between the atomic masses of tin and gold). The Tevatron (proton-antiproton collider accelerator at Fermilab with energy close to 2 TeV) has a chance to investigate the lower 10% of this mass range; the Large Hadron Collider (LHC; proton-proton collider accelerator under construction at CERN with energy 14 TeV, scheduled to start operation in or after 2007) will finish the job. Thus, by 2020 at the latest, we will know whether there is a Higgs particle, and therefore whether Nature uses the Higgs mechanism to generate the masses of fundamental particles.

STEPHEN REUCROFT

See also **Born–Infeld equations; Matter, nonlinear theory of; Particles and antiparticles; Skyrmions; String theory**

Further Reading

Englert, F. & Brout, R. 1964. Broken symmetry and the mass of the gauge vector mesons. *Physical Review Letters*, 13(9): 321

Gunion, J., Haber, H., Kane, G. & Dawson, S. 2000. *The Higgs Hunter's Guide*, Reading, MA: Perseus

Halzen, F. & Martin, A. 1984. *Quarks and Leptons*, New York: Wiley

Higgs, P. 1964. Broken symmetries and the masses of gauge bosons. *Physical Review Letters*, 13(16): 508

Higgs, P. 1964. Broken symmetries, massless particles and gauge fields. *Physics Letters*, 12(2): 132

Kane, G. 1993. *Perspectives on Higgs Physics*, River Edge, NJ: World Scientific

HILL'S EQUATION

See **Periodic spectral theory**

HIROTA'S METHOD

Ryogo Hirota originally proposed his bilinear method as a "direct method" for constructing multisoliton solutions to integrable equations. The method is "direct" in the sense that one can apply it without resorting to deeper mathematical properties, for example, to those that are needed in order to apply the inverse scattering transform (IST). In practice, this means that Hirota's method is applicable to a wider class of equations. Nevertheless, the bilinear method is related to deep mathematical foundations of soliton theory (Sato theory), as has been shown by the members of the Kyoto school (Sato, Date, Jimbo, Miwa, Kashiwara).

Prototypical Result: The KdV Equation

Let us see how Hirota's bilinear method works for the Korteweg–de Vries (KdV) equation

$$u_{xxx} + 6uu_x + u_t = 0. \tag{1}$$

It is known from IST that multisoliton solutions can be written as $u = 2\partial_x^2 \log(\det M)$, where the entries in M are of the form $c_0 + c_1 \mathrm{e}^{c_2 x + c_3 t}$ for some constants c_i. Thus, $\det M$ is a sum of exponentials and, therefore, an entire function. This inspired Hirota to write the KdV equation in terms of a new dependent variable F defined by (Hirota, 1971)

$$u = 2\partial_x^2 \log F. \tag{2}$$

Then the KdV equation (1) becomes

$$\partial_x[F_{xxxx}F - 4F_{xxx}F_x + 3F_{xx}^2 + F_{xt}F - F_xF_t] = 0. \tag{3}$$

Note the derivative outside the brackets, which means that the bilinearization actually works most naturally for the potential KdV equation.

In order to shorten the notation, let us introduce the bilinear derivative operator D as follows:

$$\begin{aligned} D_x^n\, f \cdot g &= (\partial_{x_1} - \partial_{x_2})^n f(x_1)g(x_2)\big|_{x_2 = x_1 = x} \\ &\equiv \partial_\varepsilon^n f(x+\varepsilon)g(x-\varepsilon)\big|_{\varepsilon=0}. \end{aligned} \tag{4}$$

Then the KdV equation may be written in the simple form

$$(D_x^4 + D_xD_t)\, F \cdot F = 0. \tag{5}$$

Since F was assumed to be a sum of exponentials, let us try to construct a one-soliton solution (1SS) to (5) using the ansatz

$$F = 1 + \mathrm{e}^{\eta}, \quad \eta = px + \omega t + \eta^0. \tag{6}$$

A direct computation shows that this is indeed a solution, provided that $\omega = -p^3$; this condition on the parameters is called the *dispersion relation*. From (2) we get the usual soliton solution

$$u = \frac{2p^2\mathrm{e}^{\eta}}{(1+\mathrm{e}^{\eta})^2} = \frac{p^2}{2\cosh^2(\eta/2)}, \tag{7}$$

where $\eta = px - p^3t + \eta^0$.

A particularly nice feature of the bilinear method is that two-soliton solutions (2SSs) are also very easy to obtain. We assume that a 2SS is constructed from two 1SSs perturbatively as follows:

$$F = 1 + e^{\eta_1} + \mathrm{e}^{\eta_2} + a_{12}\mathrm{e}^{\eta_1+\eta_2}, \tag{8}$$

where $\eta_i = p_ix - p_i^3t + \eta_i^0$.

A direct computation then shows that this is indeed a solution of (5), provided that the *phase factor* is given by

$$a_{12} = \left(\frac{p_1 - p_2}{p_1 + p_2}\right)^2. \tag{9}$$

Using (2) one gets the nonlinear form of the 2SS, which is already rather complicated. The above 1SS and 2SS ansätze actually work (with suitable a_{12}) for *any* equation having a bilinear form of type $P(D)F \cdot F = 0$, with $P(0) = 0$.

The above can be extended to a 3SS. By considering various asymptotic limits (where one of the three solitons is far away), one finds that the proper ansatz cannot have new free coefficients, but must have the form

$$\begin{aligned} F &= 1 + \mathrm{e}^{\eta_1} + \mathrm{e}^{\eta_2} + \mathrm{e}^{\eta_3} + a_{12}\mathrm{e}^{\eta_1+\eta_2} + a_{13}\mathrm{e}^{\eta_1+\eta_3} \\ &\quad + a_{23}\mathrm{e}^{\eta_2+\eta_3} + a_{12}a_{13}a_{23}\mathrm{e}^{\eta_1+\eta_2+\eta_3}. \end{aligned} \tag{10}$$

In contrast to the 2SS case, this ansatz only works for *integrable* equations of type $P(D)F \cdot F = 0$, for example, the KdV, Sawada–Kotera, and Kadomtsev–Petviashvili equations. In fact, the existence of multisoliton solutions can be used as a criterion of integrability (Hietarinta, 1990).

Examples of Bilinear Equations

Let us next consider the nonlinear Schrödinger equation (NLS)

$$\mathrm{i}\phi_t + \phi_{xx} + 2|\phi|^2\phi = 0. \tag{11}$$

It is bilinearized with the substitution

$$\phi = G/F, \quad G \text{ complex}, F \text{ real}, \tag{12}$$

which yields

$$F\,[(\mathrm{i}D_t + D_x^2)G \cdot F] - G\,[D_x^2 F \cdot F - 2|G|^2] = 0. \tag{13}$$

For normal (bright) solitons we split this into two equations for two functions:

$$\begin{cases} (\mathrm{i}D_t + D_x^2)G \cdot F = 0, \\ D_x^2 F \cdot F = 2|G|^2. \end{cases} \tag{14}$$

The 1SS ansatz is given by

$$\begin{aligned} &G = \mathrm{e}^{\eta}, \quad F = 1 + a\,\mathrm{e}^{\eta+\eta^*}, \quad \eta = px + \omega t, \\ &p \text{ and } \omega \text{ complex} \end{aligned} \tag{15}$$

and from the equations one finds a dispersion relation for the complex parameters, $\mathrm{i}\omega + p^2 = 0$, and the value of the phase factor, $a = 1/(p + p^*)^2$.

Hirota's method works also for discrete systems. Consider the (semi-)discrete Toda-lattice equation

$$\ddot{y}_n = \mathrm{e}^{-(y_n - y_{n-1})} - \mathrm{e}^{-(y_{n+1} - y_n)}. \tag{16}$$

First one goes over to the dependent variable $r_n = y_n - y_{n-1}$, and then the substitution $\mathrm{e}^{r_n} = 1 + \partial_t^2 \log f_n$ yields

$$(D_t^2 + 2) f_n \cdot f_n = 2 f_{n-1} f_{n+1}. \tag{17}$$

Since $e^{\alpha D_x} f(x) \cdot g(x) = f(x + \alpha) g(x - \alpha)$, we can write (17) also in the form

$$\left[D_t^2 - 4\sinh^2\left(\tfrac{1}{2}D_n\right)\right] f_n \cdot f_n = 0. \tag{18}$$

The 1SS is given by

$$f_n(t) = 1 + \mathrm{e}^{2(pn+\omega t)}, \tag{19}$$

where $\omega = \pm\sinh(p)$.

The above types of bilinearizations work, *mutatis mutandis*, for several classes of equations, but unfortunately there is no algorithmic method of finding the bilinearizing substitution for a given nonlinear equation. In fact, it is not even clear a priori how many dependent or even independent variables one should use. There are some general guidelines; for example, in order to bilinearize higher members in a hierarchy, one usually needs extra independent variables. If the 1SS and 2SS are known, one may use them in order to make an educated guess about the bilinearizing substitution.

The ultimate bilinear equation, from which many other bilinear equations can be derived as particular limits, is the Hirota–Miwa equation

$$(z_1\mathrm{e}^{D_1} + z_2\mathrm{e}^{D_2} + z_3\mathrm{e}^{D_3}) F \cdot F = 0, \tag{20}$$

where z_i are constants and D_i are arbitrary linear combinations of $D_x, D_y, \ldots$.

The General Perturbative Approach for Constructing Soliton Solutions

Once a reasonable bilinear form has been found, one can try to find soliton solutions using a perturbative expansion in a fictitious parameter ε (Hirota, 1976):

$$F = f_0 + \varepsilon f_1 + \varepsilon^2 f_2 + \cdots, \tag{21}$$

and similarly for the other dependent variables appearing in the bilinear equations. One first chooses a suitable "vacuum" solution (no solitons), that is, constant values for the ε^0 terms, with all other terms vanishing. Next for a 1SS, one uses a minimal nontrivial entry (typically $\mathrm{e}^{px+qy+\omega t}$) at order ε, that is, for f_1, and normally one should then be able to truncate the expansion at this or the next level (cf. (6) or (15)). For a 2SS, the previous 1SS solution is generalized at the ε level by taking the sum of two terms (cf. (8)), and as a result the expansion will truncate later (for KdV at ε^2, but for NLS we need even powers up to 4 for F and odd powers up to 3 for G). If the expansion does not truncate, it is probable that the equation is non-integrable. What helps here is that since $P(\boldsymbol{D})\mathrm{e}^{\boldsymbol{a}\cdot\boldsymbol{x}} \cdot \mathrm{e}^{\boldsymbol{b}\cdot\boldsymbol{x}} = P(\boldsymbol{a} - \boldsymbol{b})\mathrm{e}^{(\boldsymbol{a}+\boldsymbol{b})\cdot\boldsymbol{x}}$, the highest order terms in $P(\boldsymbol{D})F \cdot F$ vanish automatically.

In a systematic construction of multisoliton solutions, one uses determinants of Wronskian or Grammian type, or Pfaffians (Nimmo, 1990). Their applicability follows from the fact that determinants often satisfy quadratic identities, of which the Laplace expansion (a Plücker relation) is a typical example. The type of the matrix used is intimately related to the above-mentioned phase factor of the 2SS.

Bäcklund Transformations

In addition to providing a simple method for constructing multisoliton solutions, Hirota's method is also natural for studying Bäcklund transformations (BTs). In general terms, the starting set of equations and variables is doubled (while adding a new free parameter), and their combination is manipulated in order to get a pair of bilinear equations that are *linear* in each set of variables. Furthermore, if variables of one set are eliminated, one should get back the original equations. For example, the BT of the KdV equation is (Hirota, 1980)

$$\begin{cases} (D_x^3 + 3\lambda D_x + D_t)F' \cdot F = 0, \\ (D_x^2 - \mu D_x - \lambda)F' \cdot F = 0, \end{cases} \tag{22}$$

and if one eliminates F' from this pair, one obtains (a derivative of) (5).

Sato Theory

The reason Hirota's bilinear method works so well is that it is a reflection of an elegant fundamental theory called Sato theory, where bilinear identities (expressed using Young or Maya diagrams) play a major role. The theory naturally yields several infinite hierarchies of equations in an infinite number of variables, and their finite-dimensional reductions are the usual soliton equations. For an introduction, see Ohta et al. (1988) and Miwa et al. (2000).

Because there is so much mathematical structure underlying the bilinear approach, it is not surprising that bilinear forms of various equations have appeared before in the literature (although only in passing). For example, in 1902, Painlevé found bilinear forms for the first three of the six Painlevé equations using a substitution like (2); his idea was to express the equations in terms of entire functions. Indeed, the functions that arise in Sato theory are also entire functions, called τ-functions, and so one often refers to the bilinearizing functions (F and G above, for example) as τ-functions.

Summary

Hirota's bilinear method is a very effective technique in constructing multisoliton solutions. For a given equation, one should first find a one-soliton solution, and using it, one can often guess a method for bilinearizing the nonlinear equation. If that is successful, one can search for multisoliton solutions using the perturbative approach.

It is believed that all integrable evolution equations are obtained as reductions from a limited number of hierarchies, as described in Miwa et al. (2000), although the reductions can be highly nontrivial. Conversely, for a given new integrable equation, one should try to understand its origin as a reduction of one of the known hierarchies, and the bilinear form and the phase factor are helpful in this process.

In practical applications, one must nevertheless be careful, first to avoid bilinearizations that trivialize the system. The usual bilinearizing substitution (2) usually yields a multilinear system. One can forcibly bilinearize it by separating it into smaller bilinear parts and requiring each one of them to vanish independently. But if at the end there are more equations than unknown functions, the set of bilinear equations so obtained may include conditions that did not exist in the original equations.

Also, the existence of a bilinear form does not imply integrability. Bilinear forms have also been written for many non-integrable models, and the existence of a bilinear form does not even imply the existence of N-soliton solutions, nor can one always assume that the bilinearizing functions are τ-functions. Indeed, for a large class of bilinear equations (containing non-integrable equations), it is possible to find two-soliton solutions exhibiting elastic scattering, but that does not imply the existence of a *general* N-soliton solution for any $N > 2$.

JARMO HIETARINTA

See also **Bäcklund transformations; Inverse scattering method or transform; N-soliton formulas; Solitons**

Further Reading

Hietarinta, J. 1990. Hirota's bilinear method and integrability. In *Partially Integrable Evolution Equations in Physics*, edited by R. Conte & N. Boccara, Dordrecht: Kluwer, pp. 459–478

Hirota, R. 1971. Exact solution of the Korteweg–de Vries equation for multiple collision of solitons. *Physical Review Letters*, 27: 1192–1194

Hirota, R. 1976. Direct method of finding exact solutions of nonlinear evolution equations. In *Backlund Transformations, the Inverse Scattering Method, and Their Applications*, edited by R. Miura, New York: Springer, pp. 40–68

Hirota, R. 1980. Direct methods in soliton theory. In *Solitons*, edited by R. Bullough & P.J. Caudrey, New York: Springer, pp. 157–176

Miwa, J., Jimbo, M. & Date, E. 2000. *Solitons: Differential Equations, Symmetries and Infinite Dimensional Algebras*, Cambridge and New York: Cambridge University Press

Nimmo, J.J.C. 1990. Hirota's method. In *Soliton Theory: A Survey of Results*, edited by A.P. Fordy, Manchester: Manchester University Press, pp. 75–96

Ohta, Y., Satsuma, J., Takahashi, D. & Tokihiro, T. 1988. An elementary introduction to Sato theory. *Progress in Theoretical physics, Supplementum*, 94: 210–241

HODGKIN–HUXLEY EQUATIONS

Classical experiments on the compound action potential produced by vertebrate nerve trunks provided all the components for a reaction-diffusion model for the propagation of an action potential, with the nerve impulse as a nonlinear traveling wave produced by the nonlinear "reaction" of excitation and the diffusive spread of voltage. Even as late as the 1930s, however, it was not clear if nerve excitation was a membrane phenomenon, or if the the extracellular potential changes were a consequence of some chemical excitation propagating through the axoplasm.

The identification of the 100–500 μm diameter giant axons of the squid as single nerve fibers provided a preparation large enough for excitation and propagation in a single nerve fiber to be analyzed in detail. Wires can be inserted down the axoplasm, one to pass current, one to record potential, and one to short circuit the resistance of the intracellular axoplasm, providing a space clamp under which the membrane potential is spatially uniform. Current can be passed between the internal electrode and external electrodes, exciting a membrane action potential that is recorded as the difference in potential V across the membrane.

The membrane current I_{m} is the sum of a capacitative and an ionic current:

$$I_{\mathrm{m}} = C_{\mathrm{m}}\,\mathrm{d}V/\mathrm{d}t + I_{\mathrm{ion}} \tag{1}$$

with the membrane capacitance $C_{\mathrm{m}} = 1\,\mu\mathrm{F\,cm}^{-2}$. Direct recordings of the membrane action potential show that the potential changes sign, from a resting value of -60 mV to a peak value of $+40$ mV. The peak of the action potential varies logarithmically with the extracellular Na$^+$ concentration, showing the role of sodium ion (Na$^+$) current in the generation of the action potential.

The voltage clamp technique uses a feedback circuit to control the potential across the membrane. Using rectangular command pulses, the membrane potential can be changed from its resting value to an arbitrary value within a fraction of a millisecond. During a voltage clamp, the only current flowing across the membrane is ionic and is equal and opposite to the current that is injected to maintain the clamp. Net ionic currents can be measured, and dissected into their components by ion substitution experiments. From the estimated ionic currents and their electrochemical gradients, the ionic conductances changes can be calculated. Under a depolarizing voltage clamp, the sodium(Na$^+$)-conductance change is fast and transient, while the potassium (K$^+$)-conductance change is slower, delayed, and maintained.

Hodgkin & Huxley (1952) fitted their extensive series of experimental measurements of ionic currents obtained under voltage clamp and synthesized their empirical equations into a quantitative model for the propagating action potential.

For axoplasmic resistance R and membrane capacitance C, both per unit length of axon, the spread of membrane potential V with distance x (cm) and time t (ms) is described by a reaction–diffusion equation:

$$C\frac{\partial V}{\partial t} = \frac{1}{R}\frac{\partial^2 V}{\partial x^2} - I_{\mathrm{ion}}, \tag{2}$$

where I_{ion} is a nonlinear function of V and t .

The membrane ionic current I_{ion} is assumed to be composed of three independent components, a Na$^+$-current I_{Na}, a K$^+$-current I_{K}, and a leakage current density I_{L} ($\mu\mathrm{A\,cm}^{-2}$) that flow through separate conductance pathways that have different ion-selectivities, maximal conductances, and voltage-dependent kinetics. Thus,

$$I_{\mathrm{ion}} = I_{\mathrm{Na}} + I_{\mathrm{K}} + I_{\mathrm{L}}. \tag{3}$$

The instantaneous current-voltage relation for each pathway is linear, so the current is the product of a conductance and a driving force—the difference between the potential V and the equilibrium (Nernst) potential for the pathway. The Nernst potential is determined by the intracellular (axoplasm) and extracellular (artificial sea water) ionic concentrations, and $V_{\mathrm{Na}} = -115$, $V_{\mathrm{K}} = +12$, and $V_{\mathrm{L}} = -10.613$ mV (the unreasonable precision for V_{L} is to fix $V = 0$ as a stable equilibrium solution). The conductance is the product of the maximal conductance $\bar{g}$ and gating variables—activation m and inactivation h for Na$^+$-channels activation n for K$^+$-channels. The activation variables are raised to a certain power (3 for m, 4 for n) that empirically reproduces the delayed, sigmoid increase in current and may be interpreted as the number of gating structures per channel. Thus

$$\begin{aligned} I_{\mathrm{Na}} &= \bar{g}_{\mathrm{Na}} m^3 h (V - V_{\mathrm{Na}}),\\ I_{\mathrm{K}} &= \bar{g}_{\mathrm{K}} n^4 (V - V_{\mathrm{K}}),\\ I_{\mathrm{L}} &= g_{\mathrm{L}} (V - V_{\mathrm{L}}). \end{aligned} \tag{4}$$

Each gating variable obeys first-order kinetics

$$\begin{aligned} \mathrm{d}m/\mathrm{d}t &= \alpha_m(1-m) - \beta_m m,\\ \mathrm{d}h/\mathrm{d}t &= \alpha_h(1-h) - \beta_h h,\\ \mathrm{d}n/\mathrm{d}t &= \alpha_n(1-n) - \beta_n n \end{aligned} \tag{5}$$

with rate coefficients α and β for opening or closing that are empirical functions of voltage:

$$\begin{aligned} \alpha_n &= \frac{0.01(10-V)}{\exp(10-V)/10-1},\\ \beta_n &= 0.125\exp(-V/80),\\ \alpha_m &= \frac{0.1(25-V)}{\exp(25-V)/10-1},\\ \beta_m &= 4\exp(-V/18),\\ \alpha_h &= 0.07\exp(-V/20),\\ \beta_h &= \frac{1}{\exp(30-V)/10+1}. \end{aligned} \tag{6}$$

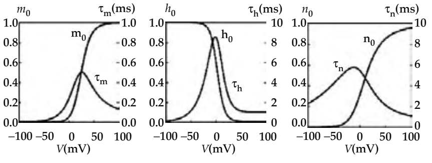

Figure 1. Voltage dependence of time coefficients and steady state values for gating variables m, h, and n of HH equations.

The rate coefficients α and β are derived from the experimentally estimated time constants τ and steady-state values illustrated in Figure 1, for example,

$$\tau_m = 1/(\alpha_m + \beta_m), \quad m_0 = \alpha_m/(\alpha_m + \beta_m). \tag{7}$$

Numerical solution (on a mechanical hand calculator) of the membrane equations reproduced both the subthreshold and action potential responses to initial depolarizations and a decrease in latency seen with increasing amplitude above threshold.

Computations of the total conductance change during the membrane action potential; total ion fluxes, time course of refractory period, and damped oscillatory responses matched experimental observations. The equations were based on experiments performed at 6.3°C; a Q_{10} of 3 (increasing temperature by 10° caused the rates to increase by a factor of 3) allowed simulation of the temperature dependence of the action potential. A solitary propagating solution of Equations (1)–(6) was obtained by assuming a wave solution traveling with a constant velocity of θ and making a coordinate transformation $\xi = x - \theta t$ to reduce Equation (1) to an autonomous system of ordinary differential equations. A value of $\theta = 18.8\,\mathrm{ms}^{-1}$ was found by trial and error to give an appropriately bounded solution in this traveling-coordinate system, which is close to the experimental estimate of $20\,\mathrm{ms}^{-1}$. For this work Alan Hodgkin and Andrew Huxley shared the Nobel Prize in Physiology or Medicine in 1963.

Early digital computations showed the continuous HH membrane equations to have a quasi-threshold, with the size of the action potential increasing smoothly with stimulus intensity. This contradiction of the "all-or-none" nature of the nerve impulse required control of variables to a higher resolution than thermal noise would allow in practice. However, similar behavior was found both numerically and experimentally at high temperatures (Cole et al., 1970). The membrane equations respond to steady injected currents by a repetitive discharge with a rate from 50 to 125 s^{-1}, with a logarithmic relation between rate and injected current density, as found in sensory coding. The numerical and experimental responses to periodic stimulation include phase-locked, entrained, and even chaotic responses. The close agreement between numerical and experimental results has allowed chaos in the periodically forced squid axon to be a test signal for the validation of methods of quantifying chaos.

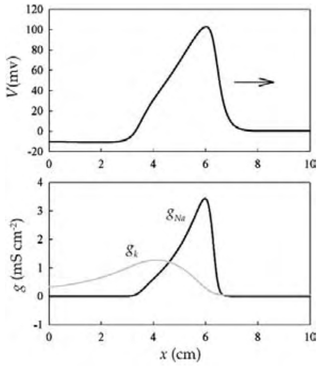

Figure 2. Constant velocity travelling wave solution of HH equations. Transmembrane voltage (upper) and membrane permeability (lower).

The equations successfully reproduced experimental data on membrane potential, ionic currents, and ion fluxes and provided a quantitative theory for the excitation and propagation of the nerve impulse. The equations were an empirical description of the currents seen under voltage clamp that also gave a quantitatively accurate reconstruction of the action potential. The form of the equations suggested a simple model of gated pores, with each pore gated by a system of four independent gates, and each gate opening or closing independently, with voltage-dependent, first-order kinetics. This physical model, of gated ionic conductance channels, has been substantiated by the electrophysiological and molecular biological characterization of single ionic channels, their kinetics and gating (Hille, 2001). The Hodgkin–Huxley formalism has also been successfully applied to the analysis of excitable membranes of nerve, smooth, skeletal, and cardiac muscle.

The introduction of digital computing into bioscience laboratories allowed numerical solution of the full partial differential equation system (Cooley & Dodge, 1966). Both stable and unstable traveling- wave solutions are found (Rinzel & Keller, 1973), and the effects of changes in diameter and axonal branching have been described.

As the first quantitatively accurate description of excitation, the HH equations have been used as a prototype for studying the behavior of excitation cells in general, as well as in circumstances with parameter values that may have no biological relevance for the squid axon. The development of autorhythmicity, by changes in parameters (injected steady current, maximal conductances, extracellular ionic concentrations, temperature), has been shown to be by simple Hopf bifurcation; the bifurcation curves in parameter space have been mapped (Holden & Winlow, 1984) and explained using singularity theory (Golubitsky & Schaeffer, 1985). The HH equations were in fact a case study for the first numerical package for using continuation algorithms to track Hopf bifurcation curves. Phase-resetting behavior and annihilation of repetitive activity by appropriately timed perturbations has been computed and found in experiments (Guttman et al., 1984).

The importance of the HH equations in physiology has led to interest from mathematicians, either by generalizing the equations to characterize the mathematical aspects of excitation (Carpenter, 1977) or to allow exotic behavior such as bursting or by simplifying the equations to allow analysis or to facilitate numerical computations.

Arun V. Holden

See also **Excitability; FitzHugh–Nagumo equation; Hopf bifurcation; Markin–Chizmadzhev model; Nerve impulses; Neurons; Periodic bursting**

Further Reading

Carpenter, G. 1977. A geometric approach to singular perturbation problems with applications to nerve impulse equations. *Journal of Differential Equations*, 23: 335–367

Cole, K.S. 1969. *Membranes, Ions and Impulses*, Berkley: University of California Press

Cole, K.S., Guttman, R. & Bezanilla, F. 1970. Nerve membrane excitation without threshold. *Proceedings of the National Academy of Science of the USA*, 65: 884–891

Cooley, J.W. & Dodge, F.A., Jr. 1966. Digital computer solutions for excitation and propagation of the nerve impulse. *Biophysical Journal*, 6: 583–599

Golubitsky, M. & Schaeffer, D.G. 1985. *Singularities and Groups in Bifurcation Theory: vol. 1*, Berlin and New York: Springer, pp. 382–396

Guttman, R., Lewis, S. & Rinzel, J. 1984. Control of repetitive firing in squid axonal membrane as a model of a neurone oscillator. *Journal of Physiology*, 305: 377–395

Hille, B. 2001. *Ion Channels of Excitable Membranes*, 3rd edition, Sunderland, MA: Sinauer

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117: 500–544

Holden, A.V. & Winslow, W. 1984. Neuronal activity as the behaviour of a differential system. *IEEE Transactions SMC*, 13: 711–719

Rinzel, J. & Keller, J.B. 1973. Traveling wave solutions of a nerve conduction equation. *Biophysical Journal*, 13: 1313–1337

HODOGRAPH TRANSFORM

Consider an N-component system with dependent variables

$$u(x,t) = [u_1(x,t), u_2(x,t), \dots u_N(x,t)]$$

that satisfy the partial differential equations

$$\frac{\partial u_i}{\partial t} = \sum_{j=1}^{N} v_{ij}(u)\,\frac{\partial u_j}{\partial x}, \quad v_{ij}(u) \in \mathbb{R}, \quad i = 1, \dots, N. \tag{1}$$

Under invertible smooth changes of variables of the form

$$u_i = u_i(w_1, w_2, \dots, w_N), \quad i = 1, 2, \dots, N, \tag{2}$$

the coefficients v_{ij} transform as a tensor,

$$v_{ij}(u_1, u_2, \dots, u_N) \to \tilde{v}_{rs}(w_1, w_2, \dots, w_N) = \sum_{i,j=1}^{N} \frac{\partial w_r}{\partial u_i} v_{ij}(u(w)) \frac{\partial u_j}{\partial w_s}.$$

Now let us assume that the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_N$ of the matrix v_{ij} are real and distinct, so system (1) is hyperbolic. Then it is possible to reduce (1), using transformation (2) to a diagonal form

$$\frac{\partial w_i}{\partial t} = \lambda_i(w)\frac{\partial w_i}{\partial x}, \quad i = 1, 2, \dots, N, \tag{3}$$

where $w(x,t) = (w_1(x,t), w_2(x,t), \dots, w_N(x,t))$. The variables $w_1, w_2, \dots w_N$ are called "Riemann invariants," and the coefficients $\lambda_1(w), \lambda_2(w), \dots, \lambda_N(w)$ are the corresponding characteristic velocities.

For $N = 2$ it is always possible locally to reduce (1) to Riemann invariant form, while for $N \geq 3$ this is not true in general. When system (1) is reducible to the diagonal form (3), it is sometimes possible to integrate it through the so-called hodograph transform. For $N \leq 2$, this fact is well known (see Courant & Hilbert, 1962); while for $N \geq 3$, the result was proved by Tsarev (1985). Let us consider the two-component system

$$\frac{\partial w_1}{\partial t} = \lambda_1(w_1, w_2)\frac{\partial w_1}{\partial x}, \quad \frac{\partial w_2}{\partial t} = \lambda_2(w_1, w_2)\frac{\partial w_2}{\partial x}. \tag{4}$$

To integrate (4), we map it into a linear set of equations, by interchanging the role of the dependent and independent variables, the so-called hodograph transform $(x,t) \to (w_1, w_2)$. In this transformation

$$\partial_2 x = -\frac{\partial_t w_1}{J}, \quad \partial_1 x = \frac{\partial_t w_2}{J}, \quad \partial_2 t = \frac{\partial_x w_1}{J}, \quad \partial_1 t = -\frac{\partial_x w_2}{J},$$

where $x = x(w_1, w_2)$, $t = t(w_1, w_2)$, $\partial_i = \partial/\partial w_i$, $i = 1, 2$ and $J = \partial_x w_1 \partial_t w_2 - \partial_t w_1 \partial_x w_2$ is the Jacobian. Then system (4) is mapped into the linear equations

$$\partial_2 x + \lambda_1\, \partial_2 t = 0, \quad \partial_1 x + \lambda_2\, \partial_1 t = 0. \tag{5}$$

The hodograph transform guarantees a linear equation, but it may not be useful in practice. Let us write (5) in the equivalent form

$$\begin{aligned} \partial_2 \underbrace{(x+\lambda_1 t)}_{\chi_1} &= t\,\partial_2\lambda_1, \\ \partial_1 \underbrace{(x+\lambda_2 t)}_{\chi_2} &= t\,\partial_1\lambda_2. \end{aligned} \qquad (6)$$

Then the functions $w_1(x,t)$ and $w_2(x,t)$, defined implicitly by

$$\begin{aligned} \chi_1(w_1,w_2) &= \lambda_1(w_1,w_2)t + x, \\ \chi_2(w_1,w_2) &= \lambda_2(w_1,w_2)t + x, \end{aligned} \qquad (7)$$

solve (4) if

$$\frac{\partial_2\chi_1}{\chi_1-\chi_2} = \frac{\partial_2\lambda_1}{\lambda_1-\lambda_2}, \qquad \frac{\partial_1\chi_2}{\chi_2-\chi_1} = \frac{\partial_1\lambda_2}{\lambda_2-\lambda_1}. \qquad (8)$$

Indeed from (7)

$$t = \frac{\chi_1-\chi_2}{\lambda_1-\lambda_2},$$

and substituting the above into (6), we obtain the linear overdetermined system (8).

As an example, consider the Born–Infeld equation

$$(1-\psi_t^2)\psi_{xx} + 2\psi_x\psi_t\psi_{xt} - (1+\psi_x^2)\psi_{tt} = 0, \quad (9)$$

which was formulated by Max Born in the early 1930s as a nonlinear field model for elementary particles. It is a simple matter to check that either

$$\psi = \Psi(x-t) \qquad \text{or} \qquad \psi = \Psi(x+t)$$

are solutions of (9) for any function Ψ. In particular, the solution can be chosen in the form of a single hump to give the solitary wave appearance. The equation is hyperbolic for solution with $1+\psi_x^2-\psi_t^2>0$. We notice that the solitary waves have constant characteristic velocity ±1 and avoid the usual breaking expected for nonlinear hyperbolic waves. Interacting waves for system (9) were obtained through the hodograph method by Barbishov & Chernikov (1966). First, if new variables

$$\xi = x-t, \quad \eta = x+t, \quad u_1 = \psi_\xi, \quad u_2 = \psi_\eta$$

are introduced, we obtain the equivalent system

$$\begin{aligned} &\partial_\eta u_1 - \partial_\xi u_2 = 0, \\ &u_2^2\partial_\xi u_1 - (1+2u_1u_2)\partial_\eta u_1 + u_1^2\partial_\eta u_2 = 0. \end{aligned} \quad (10)$$

The Riemann invariants are

$$w_1 = \frac{\sqrt{1+4u_1u_2}-1}{2u_2}, \quad w_2 = \frac{\sqrt{1+4u_1u_2}-1}{2u_1},$$

so system (10) reduces to the form

$$\partial_\xi w_1 = w_2^2\partial_\eta w_1, \quad \partial_\xi w_2 = \frac{1}{w_1^2}\partial_\eta w_2. \qquad (11)$$

Performing the hodograph transform $(\eta,\xi)\to(w_1,w_2)$, system (11) is mapped to

$$w_1^2\partial_1\xi + \partial_1\eta = 0, \qquad \partial_2\xi + w_2^2\partial_2\eta = 0,$$

where $\partial_1 \equiv \partial_{w_1}$ and $\partial_2 \equiv \partial_{w_2}$. Eliminating η leads to the simple form

$$\partial_1\partial_2\xi = 0.$$

The above equation is readily integrated as

$$\begin{aligned} x-t = \xi &= F(w_1) - \int w_2^2 G'(w_2)\,\mathrm{d}w_2, \\ x+t = \eta &= G(w_2) - \int w_1^2 F'(w_1)\,\mathrm{d}w_1, \end{aligned}$$

where F and G are arbitrary functions, and it follows that

$$\partial_1\psi = w_1F'(w_1), \quad \partial_2\psi = w_2G'(w_2).$$

It is convenient to introduce

$$\begin{aligned} F(w_1) &= \rho, \quad G(w_2) = \sigma, \\ w_1 &= \Phi_1'(\rho), \quad w_2 = \Phi_2'(\sigma), \end{aligned}$$

so that the corresponding expression for ψ reads

$$\begin{aligned} \psi &= \Phi_1(\rho) + \Phi_2(\sigma), \\ x-t &= \rho - \int_{-\infty}^{\sigma} \Phi_2'^2(\sigma)\,\mathrm{d}\sigma, \\ x+t &= \sigma + \int_{\rho}^{+\infty} \Phi_1'^2(\rho)\,\mathrm{d}\rho. \end{aligned}$$

If $\Phi_1(\rho)$ and $\Phi_2(\sigma)$ are localized, say, they are nonzero in $-1<\rho<0$ and $0<\sigma<1$, then

$$\psi = \Phi_1(x-t) + \Phi_2(x+t), \quad t<0,$$

while for $t\to+\infty$, the solution approaches

$$\begin{aligned} \psi = {} & \Phi_1\left(x-t+\int_{-\infty}^{\infty}\Phi_2'^2(\sigma)\,\mathrm{d}\sigma\right) \\ & +\Phi_2\left(x+t-\int_{-\infty}^{+\infty}\Phi_1'^2(\rho)\,\mathrm{d}\rho\right). \end{aligned}$$

Each wave receives a displacement in the direction opposite to its direction of propagation equal to

$$\int_{-\infty}^{+\infty} \Phi_i'^2(\tau)\,\mathrm{d}\tau, \quad i=1,2.$$

The interaction is remarkably similar to the interaction of solitary waves even though the Born–Infeld equation belongs to a different class of systems. We remark that the above analysis is valid provided that the mapping from the (x,t) plane to the (ρ,σ) plane is nonsingular.

Application to Multiphase Averaging

The generalized hodograph method is used to integrate (Dubrovin & Novikov, 1989) the multiphase averaged equations or Whitham equations (Whitham, 1974; Flaschka et al., 1980). These equations are obtained by averaging conservation laws over the family of multiphase solutions of one-dimensional integrable nonlinear evolution equations like, for example, the Korteweg–de Vries equation. The Whitham equations describe the modulation of the wave parameters and wave numbers of the multiphase solutions over time and space scales of order εt and εx, where t and x are the independent variables of the original nonlinear evolution equations and $0 < \varepsilon \ll 1$. When the Whitham equations are hyperbolic, the technique of integration follows from a result of Tsarev (1985) that generalizes the hodograph transform to many dependent variables. The following condition is sufficient for integrability

$$\partial_i \left(\frac{\partial_k \lambda_j}{\lambda_k - \lambda_j} \right) = \partial_k \left(\frac{\partial_i \lambda_j}{\lambda_i - \lambda_j} \right), \quad i \neq j \neq k, \; i, j, k = 1, \dots, N.$$

The integration of (3) is obtained as follows. If $\chi_1(w), \dots, \chi_N(w)$ solves the linear over-determined system

$$\partial_i \chi_j = \frac{\chi_i - \chi_j}{\lambda_i - \lambda_j} \partial_i \lambda_j, \quad i \neq j, \; i, j = 1, \dots, N, \quad (12)$$

then the solution $w_1(x,t), \dots, w_N(x,t)$ of the generalized hodograph transform

$$x + \lambda_i(w) t = \chi_i(w), \quad i = 1, \dots, N,$$

satisfies (3). Conversely, every solution of (3) can be obtained in this way in the neighborhood of a point (x_0, t_0) where $\partial_x w_i$, $i = 1, 2, \dots, N$, is not vanishing. The solution of the linear overdetermined system (12) corresponding to the Whitham equations can be obtained by algebro-geometric integration.

GREGORIO FALQUI AND TAMARA GRAVA

See also **Born–Infeld equations; Characteristics; Horseshoes and hyperbolicity in dynamical systems; Modulated waves; Periodic spectral theory; Shock waves; Solitons**

Further Reading

Barbishov, B.M. & Chenikov, N.A. 1966. Solution of the two plane wave scattering problem in a nonlinear scalar field theory of the Born–Infeld type, *JETP.*, 23: 1025–1033

Courant, R. & Hilbert, D. 1962. *Methods of Mathematical Physics*, New York: Interscience

Dubrovin, B. & Novikov, S.P. 1989. Hydrodynamic of weakly deformed soliton lattices. Differential geometry and Hamiltonian theory, *Russian Mathematical Surveys*, 44(6): 35–124

Flaschka, H., Forest, M. & McLaughlin, D.W. 1980. Multiphase averaging and the inverse spectral solution of the Korteweg–de Vries equation, *Communications on Pure and Applied Mathematics*, 33: 739–784

Tsarev, S.P. 1985. Poisson brackets and one-dimensional Hamiltonian systems of hydrodynamic type. *Soviet Mathematics Doklady*, 31: 488–491

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

HOLE BURNING

Hole burning is a nonlinear technique for high-resolution spectroscopy, which was first demonstrated in magnetic resonance spectroscopy (Bloembergen et al., 1948). Hole burning employs the spectral purity of a laser to remove from an inhomogeneously broadened spectral line a narrow homogeneous line, causing a dip (or hole) in the spectrum. Thus, the technique is aimed at revealing the homogeneous line shape in the presence of strong inhomogeneous broadening.

The homogenous line shape is usually a Lorentzian whose full width $\Delta\omega_h$ is determined by the phase relaxation time T_2 of the system being considered

$$\Delta\omega_h = 1/T_2 . \quad (1)$$

T_2 is itself composed of two different times T_1 and T_2^*. T_1 is the energy relaxation time, that is, the time by which a non-equilibrium population reaches equilibrium. T_2^* is the pure dephasing time, the time in which an ensemble of molecules randomizes its initial coherent motion. From the Bloch equation, it can be readily shown that

$$1/T_2 = 1/T_1 + 2/T_2^*. \quad (2)$$

Hence, the homogeneous line shape function contains all the information on energy relaxation and dephasing processes (Abragam, 1961).

The inhomogeneous width $\Delta\omega_i$ is often much larger than $\Delta\omega_h$, so that the homogeneous line shape is completely masked. Inhomogeneous line broadening occurs since the microenvironments of the molecules, atoms, etc. under consideration are different, so the respective resonance frequencies experience a dispersion. Accordingly, the associated line shape is, as a rule, a Gaussian. In magnetic resonance this dispersion has its origin in the field inhomogeneities, in optical spectroscopy it comes from structural imperfections of the matrix in which the probe molecule is embedded.

Hole burning is one possible method for revealing the homogenous line shape. It is a special variant of saturation spectroscopy for the case that $\Delta\omega_h \ll \Delta\omega_i$. There are two main branches of this technique, namely saturation via the power and saturation via the radiation dose of the applied field. A two-state system, for example, an electronic two-level system of a probe molecule in an imperfect lattice, can only be saturated

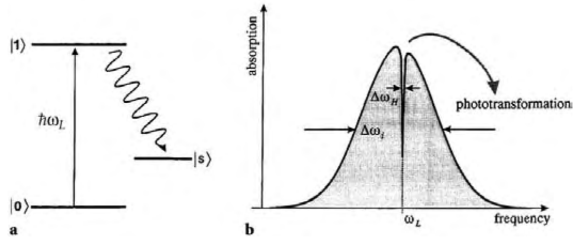

Figure 1. (a) Saturation of a three-level system. The storage level $|s\rangle$ is accessible from the excited state $|1\rangle$ by nonradiative processes. ω_L: laser frequency. (b) Schematic representation of the spectrum associated with saturation of the three-level system. $\Delta\omega_H$: holewidth. In the limit of vanishing saturation $\Delta\omega_H \to 2\Delta\omega_h$.

via the power of the radiation field: power is absorbed as long as there is a population difference between the lower and the upper state. As the power is increased, the population of the two states may equalize. Absorption and stimulated emission are balanced and the sample becomes transparent at the frequency of the radiation field. This phenomenon is called saturation. Note that power saturation of a two-level system is nonpersistent; that is, it decays with the energy relaxation time T_1. If the radiation field is tuned through the saturated transition, a dip or "hole" appears in the spectrum. The width of the hole in a power-saturated transition, $\Delta\omega_p$, however, deviates from the width of the homogenous line-shape function because it depends via the so-called Rabi frequency ω_1 on the power of the saturating radiation:

$$\Delta\omega_p = \Delta\omega_h[1 + \omega_1^2 T_1 T_2]^{1/2}. \tag{3}$$

ω_1 is proportional to the amplitude of the radiation field. Note that although the width becomes power dependent, the shape of the power-saturated transition retains its Lorentzian shape.

A very different saturation process is associated with hole burning involving three levels (Gorokhovskii et al., 1974; Kharlamov et al., 1974). Although power-saturation of an electronic two-level system is possible, it is of much less importance compared with saturation processes involving three levels, say the groundstate of a molecule $|0\rangle$, its first excited singlet state $|1\rangle$, and a long-lived intermediate $|s\rangle$ that acts as a storage state (Figure 1a). This intermediate, for instance, could be a photochemical state, a structural (e.g., a conformational) state or a long-lived spin state. It is not directly accessible to the radiation field, but can only be populated from state $|1\rangle$ via some radiative or radiationless process. Upon irradiating the $|0\rangle \to |1\rangle$ transition with a laser, population from a small range around the laser frequency ω_L is transferred to $|s\rangle$. Accordingly, a hole appears around ω_L in the absorption spectrum (Figure 1b). This hole may be persistent if the storage state $|s\rangle$ lives sufficiently long, as is the case for many photochemical states. In this case, the technique is called "photochemical hole burning." At sufficiently low temperatures, say, a few K, T_2^* becomes very large so that $\Delta\omega_h$ becomes lifetime-limited (Equations (1) and (2)). For a dye molecule with a typical lifetime of 10 ns dissolved in an organic glass, which gives rise to a typical inhomogeneous width of 300 wave numbers, this means that the burnt-in hole can be 4 to 5 orders of magnitude narrower than the inhomogeneous width. Because of this improvement in resolution, spectral hole burning is essentially a low-temperature technique (Friedrich & Haarer, 1984).

Persistent spectral hole burning may be characterized as a zero-power technique because holes can be burnt with vanishingly small power just by increasing the irradiation time. Yet, despite the fact that power saturation can be avoided, there is saturation broadening that, in contrast to the two-level case, depends on the irradiation dose: as population is transferred to the level $|s\rangle$, the number of absorbers in the center of the hole decreases steadily. Hence, the population transfer slows down in the center but still continues in the margins of the hole. This gives rise to an additional broadening in a similar way as has been discussed above for power broadening. Accordingly, $\Delta\omega_h$ is obtained from an extrapolation of the hole width to zero radiation dose.

Since the power of the radiation field can be made vanishingly small, the optics itself is perfectly linear. The nonlinear behavior of optical hole burning comes in via the persistent changes of the complex index of refraction through the irradiated energy. The advantage of the technique lies in the persistency of these changes. This makes it easy to exploit the high resolution, for instance, in measuring the influence of small external perturbations in the presence of strong inhomogeneous broadening, such as the influence of small electric, magnetic, or pressure fields. Persistent hole burning has been one of the major techniques to investigate the dynamics of low-temperature glasses via so-called spectral diffusion experiments

(Friedrich & Haarer, 1986), and it has gained much attraction in high-resolution spectroscopy of biological molecules (Friedrich, 1995). There are also quite a few technical applications, mostly in optical data storage (Moerner, 1988).

JOSEF FRIEDRICH

See also **Nonlinear optics**

Further Reading

Abragam, A. 1961. *Principles of Nuclear Magnetism*, Oxford and New York: Oxford University Press

Bloembergen, N., Purcell, E.M. & Pound, R.V. 1948. Relaxation effects in nuclear magnetic resonance absorption. *Physical Review*, 73: 679–712

Friedrich, J. 1995. Hole burning spectroscopy and physics of proteins. In *Biochemical Spectroscopy*, edited by K. Sauer, San Diego: Academic Press

Friedrich, J. & Haarer, D. 1984. Photochemical hole burning: a spectroscopic study of relaxation processes in polymers and glasses. *Angewandte Chemie*, International Edition, 23: 113–140

Friedrich, J. & Haarer, D. 1986. Structural relaxation processes in polymers and glasses as studied by high resolution optical spectrocopy. In *Optical Spectrocopy of Glasses*, edited by I. Zschokke, Dordrecht: Reidel

Gorokhovskii, A.A., Kaarli, R.K. & Rebane, L.A. 1974. Hole burning in the contour of a purely electronic line in a Shpol'skii system. *JETP Letters*, 20: 216–218

Kharlamov, B.M., Personov, R.I. & Bykovskaya, L.A. 1974. Stable gap in absorption spectra of solid solutions of organic molecules by laser irradiation. *Optics Communications*, 12: 191–193

Moerner, W.E. (editor). 1988. *Persistent Spectral Hole Burning: Science and Application*, Berlin and New York: Springer

HOLONS

In the literature on nonlinear science, the term *holon* has arisen in two quite different senses: as a type of elementary quantum excitation and in the context of emergent biological and social phenomena.

Holons in Physics

In the physics community, holons are particles with zero spin and charge ± 1 (in units of electron charge e) obtained in the fractionalization of an electron or a hole. They are related to the phenomena of spin-charge separation and fractionalization and are the supersymmetric partner of spinons—particles with zero charge and spin $\frac{1}{2}$. Holons and spinons are neither bosons (statistics 0) nor fermions (statistics -1) (statistics α represents the phase that the two-particle wavefunction accumulates when one particle is taken and moved around another), but obey fractional statistics $\frac{1}{2}$ and are, therefore, called semions.

Recently, it has become possible to experimentally realize one-dimensional samples by growing crystals, such as $KCuF_3$, $SrCuO_2$, or Sr_2CuO_3, with strong anisotropy due to antiferromagnetic spin-spin interaction. Photoemission spectroscopy experiments on these crystals show two distinct bands in the energy dispersion, one belonging to the spinon, and the other to the holon. In theory and in experiments, holons have been observed only in one spatial dimension (Kim et al., 1996, 1997; Fujisawa et al., 1998, 1999), with credible efforts to discover them in higher dimensions having remained so far unsuccessful.

The simplest model where holons show up is the supersymmetric t-J model with $1/r^2$ interaction (Kuramoto–Yokohama (KY)-model) (Kuramoto & Yokohama, 1991). This is a system of electrons on a lattice with periodic boundary conditions, where it is forbidden that two electrons be on the same lattice site. Electrons can hop from site to site, there is a Coulomb interaction term, and there are also antiferromagnetic spin-spin interactions. The Hamiltonian is supersymmetric, in the sense that it costs zero energy for a charge to be introduced in the system. The elementary excitations of the system are spinons and holons. The holons have a dispersion relation:

$$E = -\frac{J}{2}\left[\left(\frac{\pi}{2}\right)^2 - q^2\right], \tag{1}$$

where E is the holon energy, q is the holon momentum, and J is an energy scale in the system. The configuration space of positive energy holons is halved with respect to that of an integral particle. Although the equations governing the dynamics of spinons and holons are different, they yield the same form for the interaction between a holon and a spinon and between two spinons. The interaction between the holon and the spinon in one spatial dimension is inversely proportional to the distance between these particles. When the holon and the spinon are at the same point in space, the interaction between them is divergent and together they form the electron. However, the interaction does not diverge fast enough to prevent an instability of the electron toward decay into the two particles, and indeed there is a small but finite matrix element for electron to holon and spinon decay.

An experimentally accessible quantity is the spectral function for a holon-spinon pair:

$$\begin{aligned} A_{\text{holon-spinon}}(\omega, q) = {} & \frac{1}{\pi^2 J q}\sqrt{\frac{J[q+\frac{\pi}{2}]^2 - \omega}{\omega - J[q-\frac{\pi}{2}]^2}} \\ & \times \Theta\left[\omega - J[q - \frac{\pi}{2}]^2\right] \\ & \Theta\left[J[\frac{\pi^2}{4} + q(\pi - q)] - \omega\right], \end{aligned} \tag{2}$$

where ω and q are the energy and momentum at which the measurements are being conducted and J is the relevant energy scale in the system. The physical consequence of these measurements is the proof of instability of the electron, which is no longer a legitimate excitation of the system but breaks up into a holon and a spinon (Bernevig et al., 2001, 2002).

Holons in Biological and Social Systems

Biological and social systems are typically organized as hierarchies, in which nonlinear dynamics at each level of the hierarchy leads to the emergence of new dynamic entities that provide a basis for the nonlinear dynamics of the next higher level (Scott, 1995).

A well-known example of such a hierarchy is the structure of a military unit. Thus, an infantry division comprises four regiments, each of which comprises four battalions, which consist of four companies, and so on, up to 15,000 individual riflemen. Similar hierarchical structure is displayed by living organisms—for example, a human being is composed of a collection of organs, and the organs are composed of cells, which are made up of biochemicals (proteins, DNA, etc.), down to our constituent atoms.

In the 1960s, Arthur Koestler proposed the term *holon* to imply a general feature of a biological or social system that has a unique identity, yet is made up of lower-level features that comprise parts of the whole (Koestler, 1968). In this sense, holons are components of complex dynamic systems that are efficient and adaptable to disturbances, both internal and external. For example, you are a holon, as is your liver and each of your skin cells.

Interestingly, the human neocortex seems also to be hierarchically organized into closely interconnected assemblies of neurons that embody the holonic components of complex thought (Scott, 1995, 2002).

Bogdan A. Bernevig

See also **Cell assemblies; Emergence; Hierarchies of nonlinear systems; Quantum field theory**

Further Reading

Bernevig, B.A., Giuliano, D. & Laughlin, R.B. 2001. Spinon attraction in spin-$\frac{1}{2}$ antiferromagnetic spin chains. *Physical Review Letters*, 86: 3392–3395

Bernevig, B.A., Giuliano, D. & Laughlin, R.B. 2002. Coordinate representation of the one-spinon one-holon wave function and spinon-holon interaction. *Physical Review* B, 65: 195112

Fujisawa, H. et al. 1998. Spin-charge separation in single chain compound Sr_2CuO_3 studied by angle resolved photo emission. *Solid State Communications*, 106: 543

Fujisawa, H. et al. 1999. Angle-resolved photoemission study of Sr_2CuO_3. *Physical Review* B, 59: 7358–7361

Kim, C. et al. 1996. Observation of spin-charge separation in one-dimensional $SrCuO_2$. *Physical Review Letters*, 77: 4054

Kim, C. et al. 1997. Separation of spin and charge excitations in one-dimensional $SrCuO_2$. *Physical Review* B, 56: 15,589–15,595

Koestler, A. 1968. *Ghost in the Machine*, New York: Macmillan

Kuramoto, Y. & Yokohama, M. 1991. Exactly soluble supersymmetric t–J-type model with long-range exchange and transfer. *Physical Review Letters*, 67: 1338–1341

Scott, A. 1995. *Stairway to the Mind*, Berlin and New York: Springer

Scott, A. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

HOLSTEIN MODEL

See **Davydov soliton**

HOLSTEIN–PRIMAKOV TRANSFORMATION

See **Spin systems**

HOMEOMORPHISM

See **Maps**

HOMOCLINIC INTERSECTION

See **Phase space**

HOMOCLINIC TRAJECTORY

See **Phase space**

HOPF BIFURCATION

The term *Hopf bifurcation* (also called Poincaré–Andronov–Hopf bifurcation) refers to the local birth or death of a periodic solution (self-excited oscillation) from an equilibrium as a parameter crosses a critical value. It is the simplest bifurcation not just involving equilibria and, therefore, belongs to what is sometimes called dynamic (as opposed to static) bifurcation theory. In a differential equation, a Hopf bifurcation typically occurs when a complex conjugate pair of eigenvalues of the linearized flow at a fixed point becomes purely imaginary. This implies that a Hopf bifurcation can only occur in systems of dimension two or higher.

That a periodic solution should be generated in this event is intuitively clear from Figure 1. When the real parts of the eigenvalues are negative the fixed point is a stable focus (Figure 1(a)); when they cross zero and become positive the fixed point becomes an unstable focus, with orbits spiralling out. But this change of stability is a local change, and the phase portrait sufficiently far from the fixed point will be qualitatively unaffected. If the nonlinearity makes the far flow contracting, then orbits will still be coming in and we expect a periodic orbit to appear

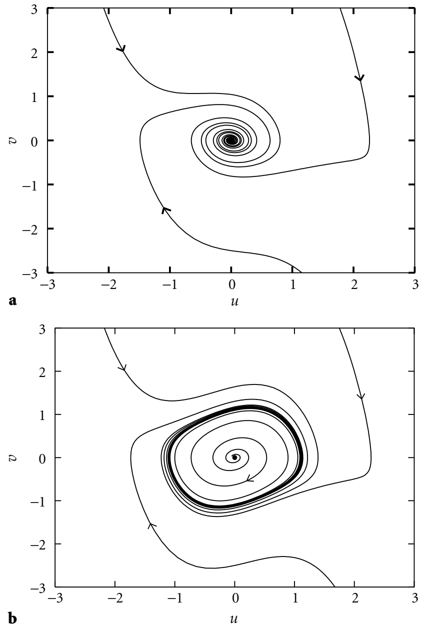

Figure 1. Phase portraits of (2) for (a) $\mu = -0.2$, (b) $\mu = 0.3$. There is a supercritical Hopf bifurcation at $\mu = 0$.

where the near and far flow find a balance (as in Figure 1(b)).

The Hopf bifurcation theorem makes the above precise. Consider the planar system

$$\begin{aligned}\dot{x} &= f_\mu(x, y),\\ \dot{y} &= g_\mu(x, y),\end{aligned} \tag{1}$$

where μ is a parameter. Suppose it has a fixed point that without loss of generality, we may assume to be located at the origin of the (x, y) plane. Let the eigenvalues of the linearized system about this fixed point be given by $\lambda(\mu), \bar{\lambda}(\mu) = \alpha(\mu) \pm \mathrm{i}\beta(\mu)$. Suppose further that for a certain value of μ (which we may assume to be zero), the following conditions are satisfied:

(i) $\alpha(0) = 0, \quad \beta(0) = \omega \neq 0$ (nonhyperbolicity condition: conjugate pair of imaginary eigenvalues);

(ii) $\left.\dfrac{\mathrm{d}\alpha(\mu)}{\mathrm{d}\mu}\right|_{\mu=0} = d \neq 0$ (transversality condition: eigenvalues cross imaginary axis with nonzero speed);

(iii) $a \neq 0$, where

$$\begin{aligned} a = {} & \frac{1}{16}\left(f_{xxx} + f_{xyy} + g_{xxy} + g_{yyy}\right)\\ & + \frac{1}{16\omega}\left(f_{xy}(f_{xx} + f_{yy}) - g_{xy}(g_{xx} + g_{yy})\right.\\ & \left. - f_{xx}g_{xx} + f_{yy}g_{yy}\right), \end{aligned}$$

with $f_{xy} = (\partial^2 f_\mu/\partial x\partial y)\big|_{\mu=0}(0, 0)$, etc. (genericity condition).

Then a unique curve of periodic solutions bifurcates from the origin into the region $\mu > 0$ if $ad < 0$ or $\mu < 0$ if $ad > 0$. The origin is a stable fixed point for $\mu > 0$ (resp. $\mu < 0$) and an unstable fixed point for $\mu < 0$ (resp. $\mu > 0$) if $d < 0$ (resp. $d > 0$), while the periodic solutions are stable (resp. unstable) if the origin is unstable (resp. stable) on the side of $\mu = 0$, where the periodic solutions exist. The amplitude of the periodic orbits grows like $\sqrt{|\mu|}$ while their periods tend to $2\pi/|\omega|$ as $|\mu|$ tends to zero. The bifurcation is called "supercritical" if the bifurcating periodic solutions are stable, and "subcritical" if they are unstable.

This two-dimensional version of the Hopf bifurcation theorem was known to Alexandr A. Andronov and his co-workers from around 1930 (Andronov et al., 1966) and had been suggested by Henri Poincaré (1892). Eberhard Hopf (1942) proved the result for arbitrary (finite) dimensions. Through center-manifold reduction, the higher-dimensional version essentially reduces to the planar one provided that apart from the two purely imaginary eigenvalues no other eigenvalues have zero real part. In his proof (which predates the center-manifold theorem), Hopf assumes the functions f_μ and g_μ to be analytic, but C^5 differentiability is sufficient (a proof can be found in Marsden & McCracken (1976)). Extensions exist to infinite-dimensional problems such as differential delay equations and certain classes of partial differential equations (including the Navier–Stokes equations) (e.g., Marsden & McCracken, 1976, Sections 8 and 9).

Example. Consider the oscillator $\ddot{x} - (\mu - x^2)\dot{x} + x = 0$ (an example of a so-called Liénard system), which, with $u = x$, $v = \dot{x}$, we can write as the first-order system

$$\begin{aligned}\dot{u} &= v,\\ \dot{v} &= -u + (\mu - u^2)v.\end{aligned} \tag{2}$$

The origin is a fixed point for each μ, with eigenvalues $\lambda(\mu), \bar{\lambda}(\mu) = \frac{1}{2}\left(\mu \pm \mathrm{i}\sqrt{4 - \mu^2}\right)$. The system has a Hopf bifurcation at $\mu = 0$. We have $d = \frac{1}{2}$ and $a = -\frac{1}{8}$, so the bifurcation is supercritical and there is a stable isolated periodic orbit (limit cycle) if $\mu > 0$ for each sufficiently small μ (see Figure 1).

Hopf Bifurcation for Maps

There is a discrete-time counterpart of the Hopf bifurcation that occurs when a pair of complex conjugate eigenvalues of a map crosses the unit circle. It is slightly more complicated than the version for flows. The corresponding theorem was first proved independently by Naimark (1959) and Sacker (1965), and the bifurcation is, therefore, sometimes called the Naimark–Sacker bifurcation. A proof can again be found in Marsden & McCracken (1976).

Consider the planar map $f_\mu : \mathbb{R}^2 \to \mathbb{R}^2$, with parameter μ, and suppose it has a fixed point that without loss of generality, we may assume to be located at $(x, y) = (0, 0)$. Suppose further that at this fixed point, Df_μ has a complex conjugate pair of eigenvalues $\lambda(\mu), \bar{\lambda}(\mu) = |\lambda(\mu)|\mathrm{e}^{\pm \mathrm{i}\omega(\mu)}$ and that for a certain value of μ (which we may assume to be 0), the following conditions are satisfied:

(i) $|\lambda(0)| = 1$ (nonhyperbolicity condition: eigenvalues on the unit circle);
(ii) $\lambda^k(0) \neq 1$ for $k = 1, 2, 3, 4$ (nonstrong-resonance condition);
(iii) $\left.\dfrac{\mathrm{d}|\lambda(\mu)|}{\mathrm{d}\mu}\right|_{\mu=0} = d \neq 0$ (transversality condition);

(iv) $a \neq 0$, where

$$a = -\mathrm{Re}\left[\frac{(1-2\mathrm{e}^{\mathrm{i}c})\mathrm{e}^{-2\mathrm{i}c}}{1-\mathrm{e}^{\mathrm{i}c}} c_{11}c_{20}\right] - \tfrac{1}{2}|c_{11}|^2 - |c_{02}|^2 + \mathrm{Re}(e^{-\mathrm{i}c}c_{21}), \qquad c = \omega(0),$$

with

$$c_{20} = \tfrac{1}{8}[(f_{xx} - f_{yy} + 2g_{xy}) + \mathrm{i}(g_{xx} - g_{yy} - 2f_{xy})],$$

$$c_{11} = \tfrac{1}{4}[(f_{xx} + f_{yy}) + \mathrm{i}(g_{xx} + g_{yy})],$$

$$c_{02} = \tfrac{1}{8}[(f_{xx} - f_{yy} - 2g_{xy}) + \mathrm{i}(g_{xx} - g_{yy} + 2f_{xy})],$$

$$c_{21} = \tfrac{1}{16}[(f_{xxx} + f_{xyy} + g_{xxy} + g_{yyy}) + \mathrm{i}(g_{xxx} + g_{xyy} - f_{xxy} - f_{yyy})]$$

(genericity condition).

Then an invariant simple closed curve bifurcates into either $\mu > 0$ or $\mu < 0$, depending on the signs of d and a. This invariant circle is attracting if it bifurcates into the region of μ where the origin is unstable (a supercritical bifurcation) and repelling if it bifurcates into the region where the origin is stable (a subcritical bifurcation).

Note that this result says nothing about the dynamics *on* the invariant circle. In fact, the dynamics on the circle has the full complexity of so-called "circle maps" (including the possibility of having attracting periodic orbits on the invariant circle) and depends sensitively on any perturbation (see the example below). Consequently, unlike the Hopf bifurcation for flows, the Hopf bifurcation for maps is not structurally stable.

Example. Consider the following family of maps:

$$f_\mu \begin{pmatrix} x \\ y \end{pmatrix} = (1 + d\mu + a(x^2 + y^2)) \times \begin{pmatrix} \cos(c + b(x^2 + y^2)) & -\sin(c + b(x^2 + y^2)) \\ \sin(c + b(x^2 + y^2)) & \cos(c + b(x^2 + y^2)) \end{pmatrix} \times \begin{pmatrix} x \\ y \end{pmatrix}. \tag{3}$$

The origin $(x, y) = (0, 0)$is a fixed point for each μ. The Jacobian matrix of f_μ at this fixed point is

$$Df_\mu(0, 0) = (1 + d\mu)\begin{pmatrix} \cos c & -\sin c \\ \sin c & \cos c \end{pmatrix} \tag{4}$$

and the eigenvalues are $\lambda(\mu), \bar{\lambda}(\mu) = (1 + \mathrm{d}\mu)\mathrm{e}^{\pm \mathrm{i}c}$. The map takes a simpler, semi-decoupled form in polar coordinates $r = \sqrt{x^2 + y^2}$, $\theta = \arctan(y/x)$:

$$\begin{pmatrix} r \\ \theta \end{pmatrix} \mapsto \begin{pmatrix} r(1 + d\mu + ar^2) \\ \theta + c + br^2 \end{pmatrix}. \tag{5}$$

This five-parameter map is in fact the *normal form* for the Hopf bifurcation up to cubic terms (i.e., by a smooth change of coordinates, we can bring any f_μ into this form (plus higher-order terms)). The parameters a, c, and d in (3) and (5) are precisely those defined in the conditions above. We choose $a = -0.02$, $b = c = 0.1$, $d = 0.2$. The map then undergoes a supercritical Hopf bifurcation at $\mu = 0$, as can be confirmed by a simple graphical analysis of the decoupled r map (for $a > 0$, it would be subcritical). For sufficiently small $\mu > 0$, we have an attracting invariant circle given by $r = \sqrt{-d\mu/a}$ (see Figure 2). On the circle the map is given by $\theta \mapsto \theta + c - bd\mu/a$. This is simply a rotation through a fixed angle $\phi = c - bd\mu/a$, giving periodic orbits if $2\pi/\phi \in \mathbb{Q}$, or dense (irrational) orbits if $2\pi/\phi \in \mathbb{R} \setminus \mathbb{Q}$.

If the Hopf bifurcation occurs in a map associated with the return map (Poincaré map) near a periodic orbit of an autonomous flow, then the bifurcation is often called a secondary Hopf bifurcation. In this case, the invariant curve corresponds to an invariant torus because the flow and attracting periodic orbits on the circle correspond to mode-locked periodic motion on the torus, while dense orbits correspond to quasi-periodic motion.

Degenerate Hopf Bifurcations

If one or more of the listed conditions for a Hopf bifurcation are not satisfied (for instance, because of symmetry), one may still have the emergence of a

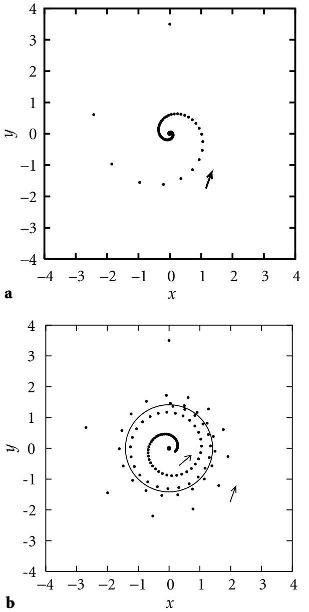

Figure 2. Phase portraits of (3) for (a) $\mu = -0.2$, (b) $\mu = 0.2$. There is a supercritical Hopf bifurcation at $\mu = 0$. ($a = -0.02$, $b = c = 0.1$, $d = 0.2$.)

periodic orbit, but some of the conclusions of the theorem may cease to hold true. The bifurcation is then called a "degenerate Hopf bifurcation." For instance, if the transversality condition is not fulfilled, the fixed point may not change stability, or multiple periodic solutions may bifurcate. An important case is provided by a Hamiltonian system for which complex eigenvalues come in symmetric quadruples, and therefore, the transversality condition cannot be satisfied. This is why the analogous bifurcation in Hamiltonian systems (the so-called "Hamiltonian–Hopf bifurcation" (van der Meer, 1985)) is much more complicated requiring, for one thing, a four-dimensional phase space.

Applications

The balance between local excitation and global damping mentioned above occurs commonly in physical systems. Thus, the Hopf bifurcation underlies many spontaneous oscillations, including airfoil flutter and other wind-induced oscillations (for example, the Tacoma–Narrows bridge collapse) in structural engineering systems, vortex shedding in fluid flow around a solid body at sufficiently high stream velocity, LCR oscillations in electrical circuits, relaxation oscillations (the van der Pol oscillator), the periodic firing of neurons in nervous systems (the FitzHugh–Nagumo equation), oscillations in autocatalytic chemical reactions (the Belousov–Zhabotinsky reaction) as described by the Brusselator and similar models, oscillations in fish populations (as described by Volterra's predator-prey model), and periodic fluctuations in the number of individuals suffering from an infectious disease (as described by epidemic models), among others.

GERT VAN DER HEIJDEN

See also **Bifurcations; Center manifold reduction; Normal forms theory; Phase space; Tacoma Narrows Bridge collapse**

Further Reading

Andronov, A.A., Vitt, A.A. & Khaikin, S.E. 1966. *Theory of Oscillators*, translated from the Russian by F. Immirzi, edited and abridged by W. Fishwick, Oxford: Pergamon Press

Hopf, E. 1942. Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems. *Berichten der Mathematisch-Physischen Klasse der Sächsischen Akademie der Wissenschaften zu Leipzig*, XCIV: 1–22. An English translation, with comments, is included as Section 5 in Marsden & McCracken (1976)

Marsden, J.E. & McCracken, M. 1976. *The Hopf Bifurcation and Its Applications*, Berlin and New York: Springer

Naimark, J. 1959. On some cases of periodic motions depending on parameters. *Doklady Akademii Nauk SSSR*, 129: 736–739

Poincaré, H. 1892. *Les Méthodes Nouvelles de la Mécanique Céleste, vol. 1*, Paris: Gauthier-Villars

Sacker, R.S. 1965. On invariant surfaces and bifurcations of periodic solutions of ordinary differential equations. *Communications on Pure and Applied Mathematics*, 18: 717–732

van der Meer, J.C. 1985. *The Hamiltonian Hopf Bifurcation*, Berlin and New York: Springer

HOPFIELD MODEL

See **Cellular nonlinear networks**

HORSESHOES AND HYPERBOLICITY IN DYNAMICAL SYSTEMS

In dynamical systems, *hyperbolicity* refers to the phenomenon in which nearby orbits diverge exponentially fast. It implies instability and sensitive dependence on initial conditions; when occurring on a wide enough scale, it implies *chaos*. This entry surveys and puts into perspective the core ideas in hyperbolic theory, one of the most developed branches in the mathematical theory of dynamical systems today.

For definiteness, we discuss only discrete-time systems, that is, systems generated by the iteration of a map f of a space (usually Euclidean space or a manifold) to itself, leaving analogous results in the continuous-time case to the reader.

Hyperbolic Fixed Points

A linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ is called hyperbolic if none of its eigenvalues lies on the unit circle. A nonlinear map f is said to have a hyperbolic fixed point at p if $f(p) = p$ and $Df(p)$ is a hyperbolic linear map. Thus, there are three kinds of hyperbolic fixed points: attracting (when the moduli of all the eigenvalues of $Df(p)$ are < 1), repelling (when they are all > 1), and saddle type (when some are > 1 and some are < 1). In the saddle case, p has a stable manifold $W^{\mathrm{s}}(p)$ and an unstable manifold $W^{\mathrm{u}}(p)$ consisting of points the orbits of which tend to p in forward and backward time, respectively.

Smale's Horseshoe

Stephen Smale's horseshoe (Smale, 1967) can be seen as a globalization of the idea of a saddle-type fixed point to an invariant set with complicated dynamics. Geometrically, the presence of a horseshoe implies stretching and folding. In terms of orbit types, it implies the existence of random motion as unpredictable as the repeated flipping of a coin (see below).

A version of the horseshoe map is shown in Figure 1: f stretches the square B in the horizontal direction, compresses it in the vertical direction, bends the resulting rectangle into the shape of a horseshoe, and puts it back on top of B as shown. The two shaded vertical strips are mapped onto the two shaded horizontal strips, the union of which is equal to $B \cap f(B)$. Reasoning inductively, we see that after n iterates, $\cap_{i=0}^{n} f^i(B)$ is the union of 2^n disjoint horizontal strips. Iterating backwards as well as forwards, we see that Λ, the set of points that remain in B in all forward and backward times, is the product of two Cantor sets.

If we label the left vertical strip in B "L" and the right one "R" (or "head" and "tail"), then every point x in Λ can be coded into a bi-infinite sequence of L and R where the ith coordinate is L if and only if $f^i x$ is in the left strip. It is easy to see that this defines a one-to-one correspondence between the points in Λ and the set of all possible bi-infinite strings in L and R. An immediate consequence of this coding is that Λ contains many periodic points, one corresponding to each finite block of L and R.

By *horseshoes*, one generally refers to a much larger class of objects than that depicted in Figure 1. The map f is assumed to be invertible, but it does not have to be linear anywhere. There is a box B (or a region B that can be deformed into a box) that is stretched and compressed by f and mapped to a set, which crosses over B finitely many times. Moreover, for points that remain in B, f has well-defined expanding and contracting directions, that is, it is hyperbolic. For an n-dimensional analog of the linear model in Figure 1, imagine $B = D_k \times D_{n-k}$, where the "horizontal" direction represents a k-dimensional disk and the "vertical" direction an $(n-k)$-dimensional disk (see also the Smale solenoid picture in the color plate section).

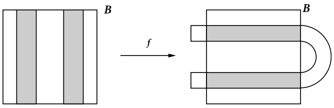

Figure 1. The horseshoe map: f sends the square B to the horseshoe on the right.

Smale's idea for the horseshoe was influenced by the work of Norman Levinson, who in the late 1940s studied a simplified version of the periodically forced van der Pol equation and proved that the resulting oscillator contains infinitely many periodic orbits with distinct periods. This map was shown to have a horseshoe by M. Levi many years later.

Dynamical complexity near homoclinic orbits was noted by Henri Poincaré. (A homoclinic point is a point in $W^s(p) \cap W^u(p)$ where p is a fixed point of saddle type; *See* **Phase space**.) An important result due to Smale says that transverse homoclinic orbits are always accompanied by horseshoes. Thus, locating transverse, homoclinic points is a means of detecting chaos.

Finally, while the presence of horseshoes implies the existence of chaotic behavior, it should be pointed out that from the probabilistic or observational point of view, this may be transient chaos. The reasons are as follows. Horseshoes have Lebesgue measure zero. It is possible to have a horseshoe Λ and at the same time for the orbit of almost every point to go to a stable equilibrium. In such a scenario, a typical orbit may come near Λ, spend some time near Λ mimicking its orbits, before it heads for its eventual destination. An experimenter tracking this orbit will observe chaotic behavior but only for a finite time period.

Uniform and Non-uniform Hyperbolicity

More general than the idea of a horseshoe is that of a uniformly hyperbolic invariant set. Let f be a smooth invertible map. A compact f-invariant set Λ is called uniformly hyperbolic if everywhere on Λ, the tangent space splits into Df-invariant subspaces $E^{\mathrm{u}} \oplus E^{\mathrm{s}}$, with $\|Df(v)\| > \|v\|$ for $v \in E^{\mathrm{u}}$, $\|Df(v)\| < \|v\|$ for $v \in E^{\mathrm{s}}$, $v \neq 0$. From the 1960s to 1970s, a detailed theory was developed for a class of dynamical systems that are uniformly hyperbolic either on their entire phase spaces or on certain important invariant sets. This theory is called Axiom A theory or uniform hyperbolic theory.

A weaker form of hyperbolicity was introduced in the 1970s. The setting here consists of a pair (f, μ) where f is a map and μ is an f-invariant Borel probability measure. Oseledec's multiplicative ergodic theorem says that at μ-almost every x, the limit $\lim_{n\to\infty} \frac{1}{n} \log \|Df^n(x)v\| = \lambda(x, v)$ exists for every tangent vector v. These asymptotic growth rates are called *Lyapunov exponents*. The pair (f, μ) is said to be non-uniformly hyperbolic if there is a positive measure set of x such that $\lambda(x, v) > 0$ for some v and < 0 for other v; that is, there is some expansion and some contraction. Pesin's paper (Pesin, 1977) helped launch a systematic study of these systems. The hyperbolicity here is non-uniform in the sense that there may be points that take arbitrarily long for the expanding and contracting behaviors to manifest themselves. Indeed, on a set of μ-measure zero, the limit above may not exist.

We put into perspective the ideas introduced: Horseshoes are examples of uniformly hyperbolic sets. They occur widely, and their occurrence does not preclude other types of dynamical behaviors. Axiom A is a more stringent condition; it requires that all important parts of the phase space be uniformly hyperbolic. This idealized picture excludes many real-life examples; at the same time it has permitted the development of a rich and extensive theory, one that is useful beyond Axiom A. As for the relation between uniform and non-uniform settings, the latter is clearly more flexible and therefore larger in scope. But the contexts are not identical: the properties of (f, μ) depend on μ, and most maps have many invariant measures. Not all invariant measures are equally important, however (*See* **Sinai–Ruelle–Bowen measures**). To complete this circle of ideas, we mention the following result of Katok: if a non-uniformly hyperbolic system has positive entropy and no zero Lyapunov exponents, then nearly all of its entropy is carried by horseshoes.

For more on uniform hyperbolic theory, *See* **Anosov and Axiom-A systems**.

Highlights from General Non-uniform Theory

The results below are very general. They hold for *all* invertible maps f (for which both f and f^{-1} are twice continuously differentiable) acting on compact domains in finite dimensions. For more detailed expositions, see Eckmann and Ruelle (1985) and Young (1993).

(1) *Local nonlinear theory* (Pesin, 1977): It is shown that corresponding to negative and positive Lyapunov exponents are measurable families of local stable and unstable manifolds.

(2) *Structure of conservative systems*, that is, when μ has a density (Pesin, 1977): Assume there are no zero Lyapunov exponents. Then the phase space is made up of at most countably many ergodic components, each one of which is, up to a permutation of sets, mixing.

(3) *Relation among entropy, Lyapunov exponents, and dimension*: For conservative systems, there is the following entropy formula (Pesin, 1997):

$$h_\mu(f) = \int \sum_{\lambda_i > 0} \lambda_i m_i \, \mathrm{d}\mu.$$

Here $h_\mu(f)$ is Kolmogorov–Sinai entropy, and the λ_i are distinct Lyapunov exponents with multiplicity m_i. In general, i.e. for arbitrary invariant measures, "=" in the formula above is replaced by "$\leq$" (Ruelle, 1978), and dimension enters to give the following equalities (Ledrappier-Young, 1985):

$$h_\mu(f) = \int \sum_{\lambda_i > 0} \lambda_i \delta_i \, \mathrm{d}\mu = -\int \sum_{\lambda_i < 0} \lambda_i \delta_i \, \mathrm{d}\mu.$$

Here δ_i is a notion of *partial dimension*; it gives the dimension of μ in the direction of the subspace of vectors whose growth rates are equal to λ_i; in particular $0 \leq \delta_i \leq m_i$. Together the results in (3) express the following two basic principles: randomness is created entirely from expansion; and dissipation can be measured in "wasted" expansion, that is, by the gap in Ruelle's inequality or, equivalently, by the dimensions of the invariant measure.

(4) *Sinai–Ruelle–Bowen (SRB) measures*: A special invariant measure (to play the role of volume) has been identified for dissipative systems with strange attractors (*See* **Sinai–Ruelle–Bowen measures**).

Examples

A large number of dynamical systems are believed to fit into the framework of non-uniform hyperbolic theory, but relatively few (that are not uniformly hyperbolic) have been rigorously studied. This is due in part to technical difficulties in proving the positivity of Lyapunov exponents in the absence of invariant cones (*See* **Lyapunov exponents**). Among the systems successfully analyzed, the most prominent are (1) Billiards flows, both scattering billiards (including the periodic Lorentz gas) and some focusing billiards (including the stadium), and geodesic flows on manifolds of nonpositive curvature. These systems are conservative and have natural invariant measures. (2) One-dimensional maps (such as the logistic family) and a class of attractors with one direction of instability (including the Hénon maps). For these maps, μ either has a density or is an SRB measure.

LAI-SANG YOUNG

See also **Anosov and Axiom-A systems; Chaotic dynamics; Lyapunov exponents; Sinai–Ruelle–Bowen measures**

Further Reading

Eckmann, J.-P. & Ruelle, D. 1985. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57: 617–656

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, New York: Springer

Pesin, Ya.B. 1977. Characteristic Lyapunov exponents and smooth ergodic theory. *Russian Mathematics Surveys*, 32: 55–114

Smale, S. 1967. Differentiable dynamical systems. *Bulletin of the American Mathematical Society*, 73: 747–817

Young, L.-S. 1993. Ergodic theory of differentiable dynamical systems. In *Proceedings of the NATO Advanced Study Institute, Hillerød, June 20–July 2, 1993*, edited by B. Branner & P. Hjorth, Dordrecht: Kluwer

Young, L.-S. 1999. Ergodic theory of chaotic dynamical systems. In *Proceedings of the 12th International Congress of Mathematics and physics, Brisbane, Australia, 1997*, International Press, pp. 131–143

HURRICANES AND TORNADOES

As examples of nonlinear emergent structures, hurricanes and tornadoes are both violent and destructive atmospheric storms, albeit of different scales: a hurricane is much larger than a tornado and lives longer. The word *hurricane* derives from the Spanish word *huracan* which, in turn, stems from the Caribbean name for the god of evil; *tornado* is a variant of the Spanish *tronada* for thunderstorm. Both are associated with very strong winds circling around a center, called the eye, which is the calmest region of the storm with the lowest pressure.

Due to the Coriolis force (which stems from the Earth's rotation), hurricane winds spiral around the eye in a counter-clockwise direction in the Northern Hemisphere, whereas in the Southern Hemisphere the opposite is true. The term *hurricane* is often restricted to storms occurring in the North Atlantic Ocean; the same phenomenon over the West Pacific Ocean is called a typhoon, and in the South Pacific, a cyclone. For simplicity, in what follows, the term *hurricane* is used for the entire phenomenon.

A hurricane forms as a tropical cyclone in areas over warm ocean waters (at temperatures over 26.5°C) under a number of other preconditions such as a moist atmosphere with a sufficiently strong vertical temperature gradient, and latitudes exceeding approximately 8° from the equator (the latter condition provides sufficient Coriolis force to create a wind rotation). Its energy source is associated with latent heat release. The low-pressure center takes in thermal energy and moisture from the ocean surface, the warm air ascends because of convection, and the higher pressure in the cold upper levels of the atmosphere pushes it outward. When the warm air ascends it cools and the temperature of the air may fall below its dew point, releasing water vapor that is able to condense into water droplets. The condensation releases heat that in turn warms the air, increasing its ascent; at this stage the storm works as a self-sustaining heat engine. When the wind speed exceeds 74 mph (119 km h^{-1}), the tropical cyclone is classified as a hurricane. In some hurricanes wind speeds reach over 280 km h^{-1}. A mature hurricane is nearly circularly symmetrical (see Figure 1) and is moved by the airstreams in which it is embedded.

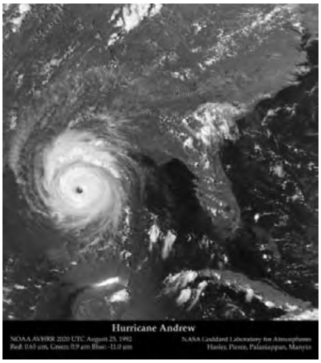
Hurricane Andrew
NOAA AVHRR 2020 UTC August 25, 1992
NASA Goddard Laboratory for Atmospheres
Hasler, Pierce, Palaniappan, Manyin

Figure 1. Satellite view of Hurricane Andrew (1992), 25 August 1992 at 20:20 UT. Courtesy of National Oceanic and Atmospheric Administration.

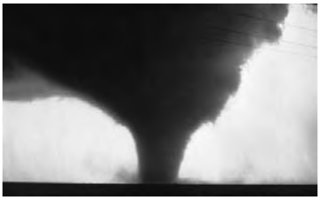

Figure 2. Photo of a tornado in Seymour, Texas (1979). Courtesy of National Oceanic and Atmospheric Administration.

A tornado is a dark, funnel-shaped cloud containing violently rotating air that extends downward to the Earth (Figure 2). It is often accompanied by lightning and a roaring sound. Although it is not yet completely known how tornadoes form, the mechanism of tornado formation is presently attributed to mechanical and thermodynamical processes (the role of atmospheric electricity is disputable). Tornadoes are formed within massive and powerful storms called supercells. The supercells develop inside cumulonimbus clouds, which owe their name to the Latin *cumulus*, meaning mound or heap, and *nimbus*, which refers to rain. Cumulonimbus clouds may reach up to 18 km into the stratosphere, where the distinctive anvil-shaped top of a cloud consisting of ice crystals is formed. Cumulonimbus clouds often form along a squall line (a wall of clouds and harsh weather often associated with a cold front) where, as a result of convective instability, a supercell is generated.

Once a supercell storm is formed, wind shear creates a mesocyclone, a region of rotating updrafts within a mature thunderstorm. Eventually, the mesocyclone reaches below the cumulonimbus cloud, whereupon it is considered a tornado cyclone. As air is pulled into the tornado cyclone, the water vapor condenses into a visible funnel cloud. When the latter touches the surface of the Earth, it is finally considered a full-fledged tornado. As time goes on, the tornado takes on a brownish color because of the dirt and debris it has drawn in. This whole process is termed the organizational stage, which leads to the mature stage when the tornado is at its largest and strongest. Eventually, when its source of warm, humid air is gone, the tornado begins to degenerate. This occurs when the atmosphere is finally stabilized due to convective processes. The tornado enters a shrinking stage and begins gradually weakening as it loses fuel. The funnel decreases to a thin column and becomes fragmented and disorganized in the decaying stage, although it may still remain destructive.

Hurricanes and tornadoes are destructive atmospheric phenomena, and not only because of strong winds. Since both hurricanes and tornadoes are storms, they are typically associated with excessive rainfall that can cause flooding. Another side effect of a hurricane is the so-called storm surge, which occurs when the ocean rises above its normal tide level. In fact, a hurricane can cause even more damage through storm surge than through strong winds. Still another potential side effect is the formation of tornadoes by thunderstorms embedded in the rainbands of a hurricane.

Both hurricanes and tornadoes are categorized by their wind speeds, from weak to devastating or violent, under the Saffir–Simpson scale for hurricanes and the Fujita–Pearson scale for tornadoes. While a hurricane can be hundreds of kilometers in diameter, the diameter of an average tornado is approximately 400 m. Tornadoes generally last from several minutes to several hours, and they travel up to several hundred kilometers before dying, whereas hurricanes last for several days and travel much greater distances. In the United States, about six hurricanes and 600 tornadoes are reported in an average year.

Some typical parameters of hurricanes and tornadoes are given in Table 1.

	Hurricanes	Tornadoes
Wind speed (kph)	119–300	60–over 500
Diameter	200–1300 km	few m–1.5 km
Lifetime	1–30 days	1–180 min

Table 1. Some parameters of hurricanes and tornadoes (by order of magnitude).

Since 1950, hurricanes have been given human names, such as Betty or George. Each year a new list of names is made, with male and female names alternating alphabetically, and the lists are reused every six years. Examples of exceptionally deadly hurricanes are the one that occurred in the Bay of Bengal in 1737, which claimed about 300,000 lives, many lost to the huge sea waves; and an 1881 typhoon in China with a similar death toll. More recent examples include the typhoon Vera (1959) that struck central Japan, causing over 5000 deaths and destroying about 40,000 homes; the hurricane Flora (1963), which presumably killed more than 6000 people in the Caribbean area, and Camille (1969), with 256 casualties and over 5 billion dollars of damage in the southern United States. The most destructive United States hurricane of record was hurricane Andrew. It blasted its way across south Florida on August 24, 1992; Louisiana and the Bahamas were also impacted. The hurricane caused 26.5 billion dollars in damage in the United States, mainly due to the winds.

Regarding tornadoes in the United States, extremely destructive was the Tri-State Tornado of 1925, which killed 689 people and injured nearly 2000. In the United States, tornadoes are most prevalent in "Tornado Alley," a group of central plains states. Since the terrain there is relatively flat, warm air coming up from the Gulf of Mexico and cold air coming down from Canada often clash, to form some of the most fertile tornado-producing storms in the world. One such tornado hit central Oklahoma in May 1999. Actually, it was a tornado outbreak—a collection of tornadoes all striking at the same time in the same general area. More than 70 tornadoes were observed in the region of Kansas, Oklahoma, and Texas. They caused 40 deaths and injured 675 people, with total damages estimated at 1.2 billion dollars.

From the physical viewpoint, hurricanes and tornadoes are examples of environmental vorticity, requiring interdisciplinary, nonlinear studies for their comprehension. Appropriate mathematical models comprise a set of nonlinear hydrodynamic and thermodynamic equations, and for hurricanes it is crucial to account for the Earth's rotation through the Coriolis force. Although a wide network of meteorological stations exists and available models provide some useful insights, the detailed modeling and prediction of hurricanes and tornadoes remains an exciting challenge for future practitioners of nonlinear science.

LEV OSTROVSKY AND MASHA SVERDLOV

See also **Atmospheric and ocean sciences; General circulation models of the atmosphere; Vortex dynamics of fluids**

Further Reading

Elsner, J. & Birol Kara, A. 1999. *Hurricanes of the North Atlantic*, Oxford and New York: Oxford University Press

Grazulis, T. 2001. *The Tornado: Nature's Ultimate Windstorm*, Norman: University of Oklahoma Press

Longshore, D. 1998. *Encyclopedia of Hurricanes, Typhoons, and Cyclones*, New York: Facts on File

Pielke, R., Jr. & Pielke, R., Sr. 1997. *Hurricanes: Their Nature and Impacts on Society*, New York: Wiley

Tufty, B. 1987. *1001 Questions Answered about Hurricanes, Tornadoes, and Other Natural Air Disasters*, New York: Dover

HUYGENS PRINCIPLE

Huygens principle (HP) is a notion that goes back to the classical *Traité de la Lumière* (Treatise on Light) by Christiaan Huygens, which was published in 1690. In its original meaning, HP gives the following geometric description of the wave fronts. Consider a wave front at the moment t as the source of the new (secondary) waves emanating from points of this front. HP states that the new wave front at a later time is the envelope of the fronts of these secondary waves (see Figure 1). This principle was further elaborated by Augustin Fresnel who added the superposition principle for the amplitudes of the secondary waves to explain the phenomenon of diffraction.

In the second half of the 19th century, various mathematical aspects of the HP were discussed in the pioneering works of Gustav Kirchhoff, Eugenio Beltrami, and Vito Volterra. At the same time it, became evident that different authors were using the term *Huygens principle* with different meanings. It was Jacques Hadamard who brought clarity in this area. In his fundamental "Lectures on Cauchy's Problem" published in 1923, Hadamard introduced the notion of the "Huygens principle in the narrow sense" (minor premise) (Hadamard, 1923). In physical language one can formulate this principle in the following way: *an instantaneous signal remains instantaneous for every observer at each later time*. This property implies that a localized disturbance will have an effect localized in time at any point (for example, for sound waves after some time there will be complete silence).

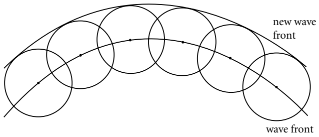

Figure 1. Diagram related to Huygens' principle.

Mathematically, HP in the narrow sense means that the fundamental solution of the corresponding hyperbolic equation vanishes not only outside but also inside the characteristic conoid and, thus, must be located on it. Equivalently, a hyperbolic equation satisfies the HP in the narrow sense if the solution of the Cauchy initial-value problem at a point P depends not on all the Cauchy data, but only on its part on the intersection of the characteristic conoid with vertex P with the Cauchy surface.

In contrast to the original principle ("major premise" in Hadamard's terminology), which holds for a general class of wave propagations, HP in the narrow sense is actually a remarkable property valid only for very special equations. In particular, this never happens in two dimensions, where one has wave diffusion. One can see this when a pebble falls in water, where the front wave is followed by the so-called "residual waves," and any object on the water surface is hit by the subsequent waves many times.

To explain this phenomenon, consider the fundamental solution of the wave equation

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x_1^2} - \cdots - \frac{\partial^2}{\partial x_n^2}\right)\Phi = 0,$$

$$\Phi(x,0) = 0, \qquad \frac{\partial \Phi}{\partial t}(x,0) = \delta(x).$$

This solution has the form

$$\Phi = \frac{1}{2\pi}\frac{\theta(t^2 - |x|^2)}{\sqrt{t^2 - |x|^2}}$$

when $n = 2$, and

$$\Phi = \frac{1}{2\pi}\delta(t^2 - |x|^2)$$

when $n = 3$. Here θ is the Heaviside step-like function and δ is the Dirac delta-function. Thus, we see that indeed for $n = 3$, Φ is located on the characteristic cone $t^2 - |x|^2 = 0$, which explains the sharpness of signal transmission in our three-dimensional world, while in two dimensions this is not the case.

Corresponding formulas for the solutions of the Cauchy problem for the wave equation in two and three spatial dimensions were found by Simeon Poisson and Kirchhoff, and for general n by Orazio Tedone. From these formulas it follows that for the Euclidean spaces, the HP in the narrow sense holds in all odd dimensions starting from 3 and never holds in even dimensions.

Hadamard raised the question of how to describe all second-order hyperbolic equations that satisfy the HP in the narrow sense (*Hadamard's problem*). Hadamard found a criterion for this, but it was not effective enough to answer this question. There was a common belief that the HP in the narrow sense holds only for the wave equations in the Euclidean spaces of odd dimensions until 1953 when Karl Stellmacher found other examples (Stellmacher, 1953). He showed that the hyperbolic equation

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x_1^2} - \cdots - \frac{\partial^2}{\partial x_n^2} + \frac{m(m+1)}{x_1^2}\right)\phi = 0$$

with integer nonnegative m satisfies the HP in the narrow sense if the dimension n is odd and large enough: $n \geq 2m + 3$.

Later in 1967–68, Stellmacher and Lagnese solved the Hadamard problem in the class of the hyperbolic equations of the special form

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x_1^2} - \cdots - \frac{\partial^2}{\partial x_n^2} + u(x_1)\right)\phi = 0,$$

when the potential depends only on one of the coordinates. They showed that the corresponding potentials are rational and (in modern terminology) can be described as the results of the Darboux transformations applied to $u = 0$ (Lagnese & Stellmacher, 1967).

The first examples of the Huygensian potentials $u(x)$ depending on more than one coordinate were found in 1993 by Berest and Veselov, who discovered a close relation of the Hadamard problem with the theory of quantum integrable systems of Calogero–Moser type (Berest & Veselov, 1994; Chalykh et al., 1999).

For the hyperbolic equations on the manifolds with nontrivial metrics, the Hadamard problem is still open even when the number of spatial variables is three. There are known only several particular cases when HP in the narrow sense is satisfied, including the modified wave equation on the symmetric spaces with even multiplicities, in particular on the spheres (Lax & Phillips, 1978) and simple compact Lie groups (Helgason, 1984), and wave equations on the spaces with the plane wave metrics (Günther, 1965). For a review of the results on the HP for other relativistic wave equations (e.g., Maxwell's equations), see Günther (1988).

Because the HP in the narrow sense is a very rare phenomenon, it is natural to ask about a weaker version. Such a version called the "generalized Huygens' principle" was introduced by Lax and Courant and is valid for a general hyperbolic equation. It states that the singularities of the solution at the point P depend only on the singularities of the initial data and only so far as these data are presented on the characteristic conoid of P. This implies that "in an approximate, and for that matter, usually satisfactory, sense any hyperbolic system does preserve the sharp signals, though in general slightly blurred" (Courant & Hilbert, 1962).

Finally, one should mention that the theory of HP is closely related to the deep theory of lacunae in the domains of dependence of hyperbolic equations (Petrovskii, 1945; Atiyah et al., 1970, 1973).

ALEXANDER P. VESELOV

See also **Characteristics; Darboux transformation; Generalized functions; Gravitational waves; Quantum nonlinearity**

Further Reading

Atiyah, M., Bott, R. & Gårding L. 1970 and 1973. Lacunae for hyperbolic differential equations with constant coefficients. *Acta Mathematica,* 124: 109–89 and 131: 145–206

Berest, Yu.Yu. & Veselov, A.P. 1994. Huygens' principle and integrability. *Russian Mathematical Surveys*, 49(6): 5–77

Chalykh, O.A., Feigin, M.V. & Veselov, A.P. 1999. Multidimensional Baker–Akhiezer functions and Huygens' principle. *Communications in Mathematical Physics*, 206: 533–566

Courant, R. & Hilbert, D. 1962. *Methods of Mathematical Physics*, New York: Interscience Publishers

Günther, P. 1965. Ein Beispiel einer nichttrivialen Huygensschen Differentialgleichung mit vier unabhängigen Variablen. *Archive for Rational Mechanics and Analysis*, 18: 103–106

Günther, P. 1988. *Huygens' Principle and Hyperbolic Equations*, Boston: Academic Press

Hadamard, J. 1923. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, New Haven: Yale University Press

Helgason, S. 1984. Wave equations on homogeneous spaces. *Lecture Notes in Mathematics*, 1077: 252–287

Huygens, Ch. 1690. *Traité de la Lumière*, Hague: Pierre vander Aa; as *Treatise on Light*, New York: Dover, 1962, and London: Dawson, 1966

Lagnese, J.E. & Stellmacher, K.L. 1967. A method of generating classes of Huygens' operators. *Journal of Mathematics and Mechanics*, 17: 461–472

Lax, P.D. & Phillips, R.S. 1978. An example of Huygens' principle. *Communications in Pure and Applied Mathematics*, 31: 415–421

Petrovskii, I.G. 1945. On the diffusion of waves and lacunas for hyperbolic equations. *Matematicheskii Sbornik,* 17: 289-370 (Russian) (Also in Petrowsky, I.G. *Selected works.* Part I. Gordon and Breach, London, 1996)

Stellmacher, K.L. 1953. Ein Beispiel einer Huygensschen Differentialgleichung. *Nachrichten der Akademie der Wissenschaften in Goettingen. Mathematisch-physikalisches Klasse,* IIa, 10: 133–138

HYDRAULIC JUMPS

See **Jump phenomena**

HYDRODYNAMIC LIMIT

See **Fluid dynamics**

HYDRODYNAMIC SOLITARY WAVES

See **Water waves**

HYDROGEN BOND

The hydrogen bond is a weak chemical bond, which nevertheless is fundamentally important in stabilizing secondary structure motifs in biological macromolecules such as proteins and DNA. Also, the peculiar properties of liquid water are attributed to a complex hydrogen bond network between neighboring water molecules. A hydrogen bond is mediated by a proton sitting in between two negatively charged atoms, the more negative of which is called the hydrogen bond donor and the other the hydrogen bond acceptor. Figure 1a shows a prototype hydrogen bond, using the N–H and the C = O group of two peptide units in an α-helix as an example (one of the most important secondary structure motif of proteins).

Hydrogen bonding leads to a dramatic distortion of the potential energy surfaces that determine the nuclei positions, giving rise to strongly anharmonic or even double-well potentials. Much of our knowledge about hydrogen bond potentials stems from vibrational spectroscopy, see, for example, Pimentel & McClellan (1976); Hadzi & Bratos (1976); Henri-Rousseau & Blaise (1998). The absorption band of the high-frequency vibration of the donor-proton bond (i.e., the NH bond in the example considered here) changes considerably upon formation of a hydrogen bond. The most evident effects are a strong red-shift with respect to the free group, an intensity increase and band broadening, often accompanied by a peculiar band shape with rich substructure. These effects could be reasonably explained using a model of the electronic (ground state) potential energy surface as a function of two coordinates, the high-frequency N–H stretching coordinate q, and the low-frequency hydrogen bond coordinate Q (see Figure 1a). A Taylor expansion of that potential energy surface yields (Hadzi & Bratos, 1976; Henri-Rousseau & Blaise, 1998)

$$V(Q,q) = \tfrac{1}{2}m\omega^2 q^2 + \tfrac{1}{2}M\Omega^2 Q^2 + \chi' Q q^2 + \cdots, \quad (1)$$

where the first two terms represent the harmonic normal modes of the N–H stretching and the hydrogen bond vibration with oscillation frequencies ω and Ω and effective masses m and M, respectively. Typical vibrational frequencies of N–H stretching modes lie in the range of $\omega = 3000$–$3600\,\mathrm{cm}^{-1}$, while hydrogen bond vibrations are found around $\Omega = 50$–$300\,\mathrm{cm}^{-1}$, depending on the masses of the molecular groups involved. This translates into typical spring constants of $k_{\mathrm{N-H}} \approx 650\,\mathrm{N/m}$ for the N–H bond and $K_{\mathrm{N-O}} \approx 10 - 20\,\mathrm{N/m}$ for the hydrogen bond. In the case of most molecular systems, the harmonic approximation is extremely good, and higher order terms in Equation (1) can be treated as weak perturbation (being responsible, for example, for energy dissipation). Harmonic normal modes have the property to decouple completely. Linear phenomena, such as propagation of linear phonon wave packets, can be fully understood in the harmonic approximation.

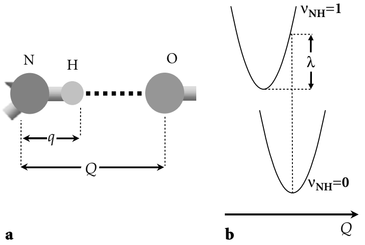

Figure 1. (a) A prototype hydrogen bond. The discussion presented here applies for other hydrogen bonds in essentially the same way. (b) Potential energy surfaces of the $\nu_{\mathrm{NH}} = 0$ and the $\nu_{\mathrm{NH}} = 1$ level.

However, one of the striking features of hydrogen bonds is an extraordinarily large anharmonicity, described by third, and higher, order terms in the expansion Equation (1). This anharmonicity leads to appreciable nonlinear behavior, a property that makes hydrogen-bonded crystals interesting objects to study in nonlinear science.

The problem can be simplified significantly when taking into account the more than one order of magnitude difference in the N–H and the hydrogen bond vibration frequency, allowing an adiabatic separation of timescales to be introduced. This is done in exactly the same way as the Born–Oppenheimer approximation separates off the fast motion of the electrons from the slow motion of the nuclei. In the case of a hydrogen bond, the motion of the fast N–H vibration adiabatically adapts to the position of the coordinate Q of the hydrogen bond. When the coordinate Q is held fixed, one can recast Equation (1) in the form

$$V(Q,q) = \tfrac{1}{2}m\omega_{\mathrm{eff}}^2 q^2 + \tfrac{1}{2}M\Omega^2 Q^2 \quad (2)$$

with

$$\omega_{\mathrm{eff}} \approx \omega + \frac{\chi'}{m\omega}Q \equiv \omega + \chi Q. \quad (3)$$

In this approximation, the vibration frequency of the N–H stretching mode, and hence its excitation energy, varies linearly with the hydrogen bond distance Q. The other third-order term not considered in Equation (1) ($\sim Q^2 q$) leads to a Q-dependent shift of the N–H-bond-length without change of its excitation energy. Such a linear dependence is observed experimentally for a large variety of hydrogen-bonded crystals (Pimentel & McClellan, 1976), which allows direct measurement of the nonlinear coupling constant: $\chi_{\mathrm{N-H}} \approx 500 - 1500\,\mathrm{cm}^{-1}/\text{Å}$ or

$\chi_{\text{N–H}} \approx 1-3 \times 10^{-10}$ N. The nonlinear coupling constant varies significantly with hydrogen bond type.

Starting from Equations (2) and (3), one can construct potential energy surfaces (see Figure 1b), leading to the displaced oscillator picture that is well known from electronic Franck–Condon transitions. However, keep in mind that all this is happening on the *electronic ground state* potential surface! Each potential energy surface corresponds to a vibrational excitation level $\nu_{\text{N–H}} = 0, 1, \ldots$ of the N–H stretching mode, and describes the total energy of the system as a function of the hydrogen bond coordinate Q. It is the nonlinear coupling term χ that gives rise to the displacements of the potential energy surfaces. The reorganization energy λ, that is, the energy the system gains by relaxation towards the bottom of excited state potential energy surface after a vertical "Franck–Condon"-like excitation (see Figure 1b), is given by

$$\lambda = \frac{\chi^2}{2M\Omega^2} = \frac{\chi^2}{2K_{\text{N}\cdots\text{O}}}. \tag{4}$$

Biological macromolecules, such as proteins and DNA, often form one-dimensional quasicrystals with approximate translational symmetry, whose structures are stabilized by hydrogen bonds. This observation motivated Alexander Davydov to speculate about nonlinear collective phenomena in bio-macromolecules (Davydov, 1979). Davydov discussed only the C = O vibration in a N–H···OC-hydrogen bond, which shows the same phenomenon with a smaller, yet still appreciable, nonlinearity $\chi_{\text{CO}} \approx 300\ \text{cm}^{-1}/\text{Å}$ or $\chi_{\text{CO}} \approx 6 \times 10^{-11}$ N (Careri et al., 1984). Dipole–dipole interaction between adjacent C = O and N–H vibrations tends to delocalize the excitation along the crystal, forming a vibrational exciton (vibron). Coupling of the vibron to lattice-deformation modes, mediated through the nonlinearity χ of the hydrogen bonds, subsequently self-localizes the excitation. The hydrogen bond is getting stronger after excitation of the high frequency vibration (see Figure 1b), leading to a contraction of the macromolecular backbone. As a result, solitons and/or polarons may be formed.

PETER HAMM

See also **Color centers; Davydov soliton; Frank–Condon factor**

Further Reading

Careri, G., Buontempo, U., Galluzzi, F., Scott, A.C., Gratton, E. & Shyamsunder, E. 1984. Spectroscopic evidence for Davydov-like solitons in acetanilide *Physical Review* A, 30(8): 4689–4702

Davydov, A.S. 1979. Solitons in molecular systems. *Physica Scripta*, 20: 387–394

Hadzi, D. & Bratos, S. 1976. Vibrational spectroscopy of the hydrogen bond. In *The Hydrogen Bond: Recent Developments in Theory and Experiment*, edited by P. Schuster, G. Zundel & C. Sandorfy, Amsterdam: Elsevier, pp. 565–611

Henri-Rousseau O. & Blaise P. 1978. The infrared spectral density of weak hydrogen bonds within the linear response theory. *Advances in Chemical Physics*, 103: 1–186

Pimentel, G.C. & McClellan, A. 1960. *The Hydrogen Bond*, San Francisco: Freemann

HYDROTHERMAL WAVES

Hydrothermal waves are traveling waves produced by a supercritical instability of the thermocapillary flow of a thin liquid layer with a free surface subjected to a horizontal temperature gradient.

This article presents experiments in one spatial dimension, then gives some indications about the physical basis for wave effects. Extended systems are then considered and applications are described.

Experimental Observation in One Spatial Dimension

Hydrothermal waves were first experimentally observed in hot-wire convection experiments (Vince & Dubois, 1992), and a geometrically simpler system was then imagined by Daviaud & Vince (1993) that led to the first clear experimental observation of hydrothermal waves.

The classical experiment is as follows: a channel whose long sides are thermoregulated copper blocks contains a fluid with free surface. Its typical section is depicted in Figure 1. The horizontal gradient of temperature on the free surface leads to a gradient of surface tension. This results in a stationary flow of the surface from the hot side (low surface tension) toward the cold side (high surface tension): this is the Marangoni effect. Due to mass conservation, there is a return flow in the bottom of the layer. This is the basic thermocapillary flow (Kuhlnan, 1999; Schatz & Neitzel, 2001), the instability of which leads to hydrothermal waves (Smith & Davis, 1983; Davis, 1987). A similar flow is present in the melted wax of a candle (the cold side being the external boundary and the hot side being the wick). For visualization purposes, silicon oil is used because it is transparent and allows shadowgraphy.

For fluid depth h larger than the capillary length λ_c, one observes first when increasing the temperature gradient corotative rolls with their axis parallel to the cold and hot side of the container. Then hydrothermal waves of type 1 (HW1, plane waves) appear on top of the rolls (Figure 2a) (Riley & Neitzel, 1998). HW1 are emitted by "line"-sources that extend over the whole extension between the hot and the cold sides. For smaller depth $h < \lambda_c$, waves of type 2 (HW2) are observed (Figure 2b,c); they are circular and emitted by point sources located on the cold side of the container (Burguete et al., 2001; Garnier & Chiffaudel, 2001).

When the fluid depth is even larger (typically larger than the depth λ_{th} for which thermocapillary

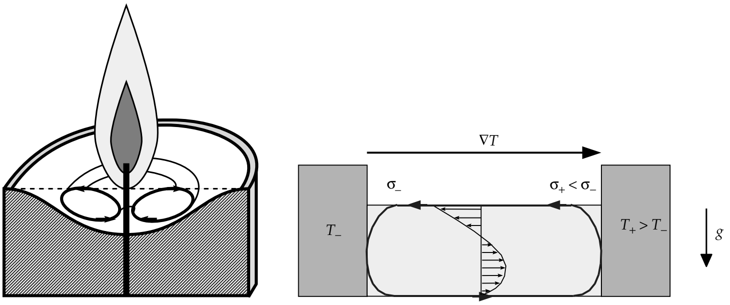

Figure 1. Thermocapillary basic flow created by Marangoni effect and susceptible to instability into hydrothermal waves. Left: the melted wax below the flame of a candle is an example of thermocapillary flow. Right: typical section of a laboratory setup. A horizontal temperature gradient is applied to a thin liquid layer with a free surface. Hydrothermal waves (not drawn) will propagate in the horizontal direction orthogonal to the temperature gradient ∇T, and with a small component along the gradient.

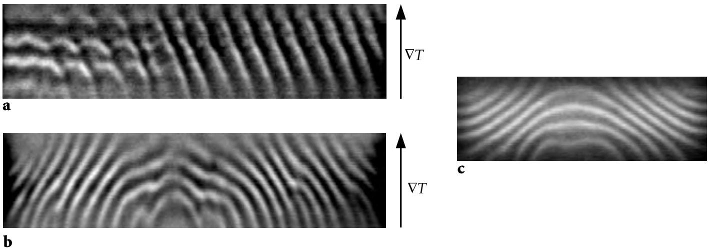

Figure 2. Shadowgraphic images of hydrothermal waves in rectangular cells. (a): $L_x = 30$ mm, $L_y = 180$ mm, $h =$ 2.75 mm, $\Delta T =$ 7 K. A single right-going wave of type 1 is present. The left side is a source, the amplitude of the wave is weak, and stationary corotative rolls orthogonal to the gradient are present. (b): $L_x = 30$ mm, $L_y = 180$ mm, $h = 1$ mm, $\Delta T = 7.3$ K. (c): $L_x = 30$ mm, $L_y = 90$ mm, $h = 1$ mm, $\Delta T = 5$ K. Hydrothermal waves of type 2, with a source located at the center of the cold side of the channel.

forces balances thermogravity forces), hydrothermal wave instability is replaced by a stationary instability: parallel rolls with axis aligned with the temperature gradient.

Physical Mechanisms

Vertical Temperature Gradient: Bénard–Marangoni Instability

Pearson (1958) gave a simple mechanism to explain hexagon formation in Bénard–Marangoni convection. In that case, the temperature gradient is purely vertical (cold at the surface, and hot at the bottom) and the velocity of the fluid is zero before the instability. Let us consider a positive temperature perturbation at the surface of the fluid: a point has a temperature larger than its surroundings. Because surface tension decreases with temperature for simple fluids, the surface tension is smaller at that point. This implies a differential stress on the free surface, and according to the equations of motion, the fluid has to flow away from that point (fluids flow from regions of small surface tension toward regions of large surface tension). But due to mass conservation, some fluid must rise up from the bulk of the fluid to the point on the surface (in the same way that wax climbs around the wick and flows away on the surface towards cooler regions). This ascending fluid is at a higher temperature because of the vertical temperature gradient. In conclusion, any positive temperature perturbation at the free surface is amplified; this means that there is an instability, namely, the Bénard–Marangoni instability into stationary hexagons, as can be shown by a linear stability analysis.

Horizontal Temperature Gradient: Hydrothermal Waves

Basically, the mechanism is the same, but one has to take into account, first, that the basic flow has

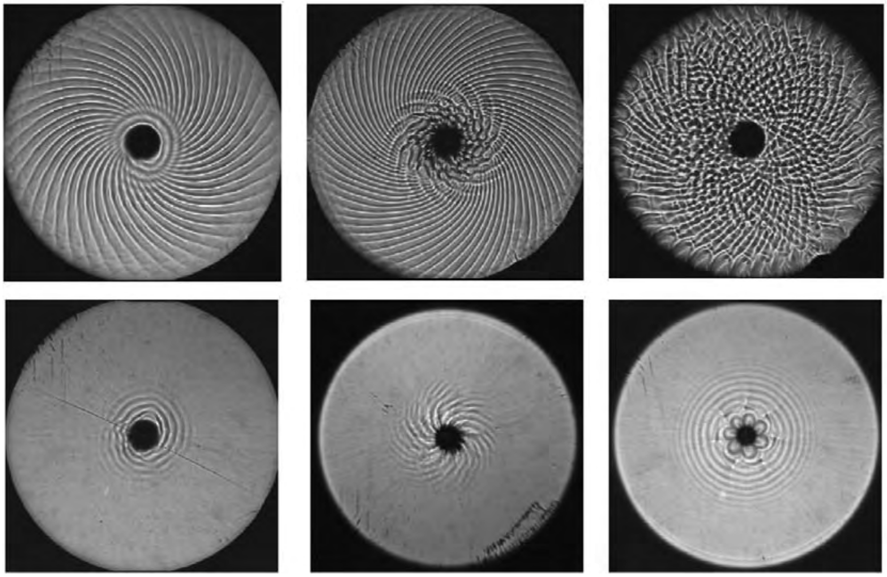

Figure 3. Top three pictures: extended patterns: single HW1 ($h = 1.9$ mm, $\Delta T = 14.25$ K), HW1 and HW2 together ($h = 1.2$ mm, $\Delta T = 20$ K), turbulent HW1 ($h = 1.9$ mm, $\Delta T = 25$ K). Bottom three pictures: localized patterns: HW2 ($h = 1.2$ mm, $\Delta T = 10$ K), inverted spirals ($h = 1.2$ mm, $\Delta T = -9$ K), and flowers ($h = 1.9$ mm, $\Delta T = -7$ K).

a finite velocity (see Figure 1), and second, that the vertical temperature profile is nonmonotonous. This intricacy results in a time-dependent pattern, in contrast to the stationary hexagons described above. Time-dependence is due to the existence of two different timescales: one for the relaxation of thermal perturbations (depending on the thermal diffusion coefficient) and the other for the relaxation of velocity perturbations (depending on kinematic viscosity). The ratio of those two timescales is called the Prandtl number, Pr.

Smith (1988) expressed two different tentative mechanisms depending on Pr in the limit cases of a flow dominated by inertial effects ($\mathrm{Pr} \to 0$) or by viscous effects ($\mathrm{Pr} \to \infty$). The relaxation of temperature and velocity perturbations then occur on very different timescales. Depending on the signs of the temperature and velocity gradients, an oscillatory behavior is shown to be unstable and to propagate along the horizontal temperature gradient (small Pr), or perpendicularly to it (large Pr).

Finally, note that hydrothermal waves are not necessarily associated with surface deflections (in contrast to gravity waves for example): they are an instability mode of the temperature field in the bulk (Pelacho & Burguete, 1999).

Experiments in 2-d

Recent experiments in extended geometries (the two horizontal dimensions are large compared to the fluid height) revealed a large variety of hydrothermal waves instability modes. Several of these obtained in cylindrical geometry (the candle geometry) are reproduced in Figure 3. The control parameter is defined as $\Delta T = T_{\mathrm{ext}} - T_{\mathrm{int}}$. This quantity can be positive or negative, and both cases are not equivalent. This is due to the presence of curvature (Garnier & Normand, 2001) that may also localize patterns near the center.

Applications

Many technological applications in which the Marangoni effect is present will involve hydrothermal waves—even when manufacturing is carried out in low gravity, where gravity-dependent, buoyancy-driven flow is reduced, for example, floating zone purification of silicon crystals, photographic films production, and melting of metals. In all of these processes, the aim is to avoid hydrothermal waves, which are detrimental to the final product, for example, by reducing the homogeneity of crystals.

In the physics laboratory, hydrothermal waves represent an ideal experimental nonlinear waves system. They are well modeled by a complex Ginzburg–Landau equation (Garnier et al., 2003) and can be used as a robust model for the study of the transition to spatiotemporal chaos. For example, as their group velocity is finite, they are subject to the convective/absolute distinction. Modulated amplitude waves have also been reported.

NICOLAS GARNIER AND ARNAUD CHIFFAUDEL

See also **Candle; Complex Ginzburg–Landau equation; Fluid dynamics; Thermo-diffusion effects**

Further Reading

Burguete, J., et al. 2001. Buoyant-thermocapillary instabilities in extended liquid layers subjected to a horizontal temperature gradient. *Physics of Fluids*, 13: 2773–2787

Daviaud, F. & Vince, J-M. 1993. Traveling waves in a fluid layer subjected to a horizontal temperature gradient. *Physical Review* E, 48: 4432–4436

Davis, S.H. 1987. Thermocapillary Instabilities. *Annual Reviews of Fluid Mechanics*, 19: 403–435

Garnier, N. & Chiffaudel, A. 2001. Two dimensional hydrothermal waves in an extended cylindrical vessel. *European Journal of Physics* B, 19: 87–95

Garnier, N. & Normand, C. 2001. Effects of curvature on hydrothermal waves instability of radial thermocapillary flows. *Comptes Rendus de l'Academie des Sciences, Series IV*, 2(8): 1227–1233

Garnier, N., et al. 2003. Nonlinear dynamics of waves and modulated waves in 1D thermocapillary flows. I. General presentation and periodic solutions. *Physica* D, 174: 1–29

Kuhlman, D. 1999. *Thermocapillary Flows*, New York: Springer

Pearson, J.R. 1958. On convection cells induced by surface tension. *Journal of Fluid Mechanics*, 4: 489–500

Pelacho, M.A. & Burguete, J. 1999. Temperature oscillations of hydrothermal waves in thermocapillary-buoyancy convection. *Physical Review* E, 59: 835–840

Riley, R.J. & Neitzel, G.P. 1998. Instability of thermocapillary–buoyancy convection in shallow layers. Part 1. Characterization of steady and oscillatory instabilities. *Journal of Fluid Mechanics*, 359: 143–164

Schatz, M.F. & Neitzel, G.P. 2001. Experiments on thermocapillary instabilities. *Annual Reviews of Fluid Mechanics*, 33: 93–127

Smith, M.K. 1986. Instabilty mechanisms in dynamic thermocapillary liquid layers. *Physics of Fluids*, 29: 3182–3186

Smith, M.K. & Davis, S.H. 1983. Saturation of Rayleigh-Taylor instability. *Journal of Fluid Mechanics*, 132: 119–144 and 132: 145–162

Vince, J-M. & Dubois, M. 1992. Hot wire below the free surface of a liquid: Structural and dynamical properties of a secondary instability. *Europhysics Letters*, 20: 505–510

HYPERBOLIC MAPPINGS

See **Maps**

HYPERCHAOS

See **Rössler systems**

HYPERCYCLE

See **Catalytic hypercycle**

HYPERELLEPTIC FUNCTIONS

See **Elliptic functions**

HYSTERESIS

The word *hysteresis* is derived from the Greek where it meant "shortcoming." In the present physical context it describes a retardation effect when the forces acting upon a body are changed. In particular, in magnetism hysteresis represents a lagging in the values of the net magnetization in a magnetic material due to a changing magnetizing field. In general, hysteresis signifies the history of dependence of physical quantities in systems responding to changes in external conditions. The term is most commonly, but not exclusively, applied to magnetic materials; for example, there is a class of metals called shape memory alloys that can be bent or stretched plastically over large distances back and forth many times without hardening.

Consider a ferromagnetic material that is originally unmagnetized. As the external magnetic field (H) is increased, the induced magnetization (M) also increases. The induced magnetization eventually saturates. Now, if the external field is reduced, the induced magnetization also is reduced, but it does not follow the original curve. Instead, the material retains a certain permanent magnetization called the remanent magnetization M_{r} when $H = 0$. The remanent magnetization is the permanent magnetization that remains after the external field is removed. If the external field is reduced more, the remanent magnetization will eventually be removed. The external field applied in the opposite direction for which the remanent magnetization goes to zero is termed the coercivity H_{c}. The product of M_{r} and H_{c} is termed the strength of the magnet. As the external field continues to reverse, permanent magnetization of the opposite polarity is created in the magnet. A similar curve is traced for the negative direction with saturation, remanent magnetization and coercivity. The hysteresis curve then retraces the previous points as the field cycles, and the shape of the loop after the first cycle is roughly the same as after many cycles. Note that the area under the hysteresis loop corresponds to the work done on the system by an external field that reorients the magnetization in a single cycle.

When an electric field (E) is applied to a ferroelectric crystal, the domains that are favorably oriented with respect to this field grow in size at the expense of those that are misaligned. In addition, favorably oriented domains may nucleate and grow until the whole crystal becomes one domain. The relation between the resulting polarization P and E is described by a hysteresis loop in analogy to the relationship between M and H for ferromagnets.

Suppose now, in general, that the system under consideration can be described by a macroscopic order parameter η. Under the influence of an external field σ coupled linearly to η, the state of the system is determined by an equation of state that expresses a minimum condition of the associated thermodynamic potential V. Assuming the presence of at least one control parameter leads directly to the problem of catastrophes, an area investigated by René Thom. The cusp catastrophe is described by the potential

$$V(\eta) = \frac{\varepsilon}{4}\eta^4 + \frac{1}{2}a\eta^2 - \sigma\eta, \qquad (1)$$

where a is a control parameter. This potential describes second-order phase transitions both in the absence ($\sigma = 0$) and in the presence ($\sigma \neq 0$) of external fields as proposed by Lev Landau. The butterfly catastrophe, on the other hand, is described by

$$V(\eta) = \frac{1}{6}\eta^6 + \frac{a}{4}\eta^4 - \frac{b}{3}\eta^3 + \frac{c}{2}\eta^2 - \sigma\eta \qquad (2)$$

and has been used to model first-order phase transitions.

To illustrate the related phenomenon of hysteresis we first investigate the bifurcation effect by minimizing Equation (1), which yields the equation of state

$$\varepsilon\eta^3 + a\eta = \sigma. \qquad (3)$$

A transition between a single stable solution for $a > 0$ and a bistable situation for $a < 0$ takes place when $a = 0$. However, a new feature is the phenomenon of external-field-induced hysteresis and metastability. Stability corresponds to a solution for which $\partial^2 V/\partial\eta^2 > 0$ and, if more than one solution of the equation of state exists that is stable, we call the higher energy solutions metastable. Figure 1 shows the difference in the response of the order parameter to the application of an external field. In unistable situations, η as a function of σ is a smooth single-valued function. Multistability results in multi-valuedness in some ranges of the external fields. Figure 1c demonstrates the regions of multistability in the parameter space. The dividing line in the cusp catastrophe case is given by the equation

$$-4a^3 + 27\sigma^2 = 0. \qquad (4)$$

This indicates the parameter set for which a triple solution of the cubic equation in (4) is obtained. Figure 2 is an extension of these concepts to the butterfly catastrophe case that exhibits thermal hysteresis (if the control parameter a involves a temperature dependence) and a double hysteresis under the influence of an external field.

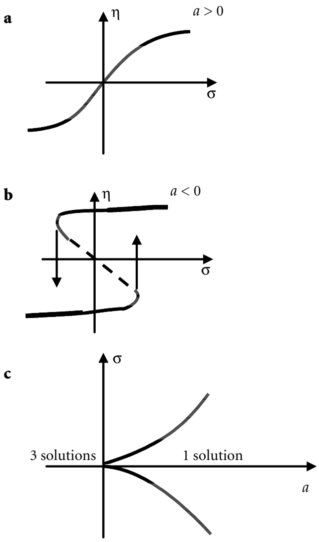

Figure 1. The order parameter's response to an externally applied field (a) in a unistable state, (b) in a bistable state (hysteresis), and (c) separation of the control parameter space into solution's multiplicity regions.

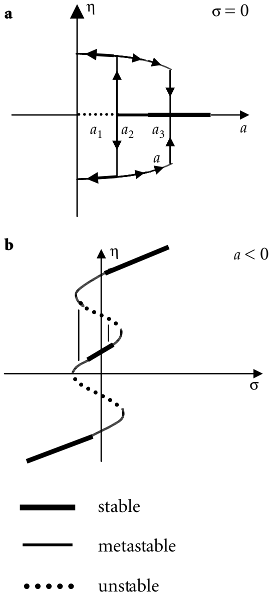

Figure 2. (a) Bifurcation plot (thermal hysteresis) and, (b) double field-induced hysteresis in the butterfly catastrophe case.

The sixth power free energy expansion for the order parameter η

$$F = F_0 + a\varepsilon\eta^2 + A_4\eta^4 + A_6\eta^6 - h\eta \qquad (5)$$

has been widely used to model first-order phase transitions provided $A_4 < 0$. The equation of state gives the order parameter in equilibrium

$$\eta_0 = \left\{\left[-A_4 - \left(A_4{}^2 - 3a\varepsilon A_6\right)^{1/2}\right]/3A_6\right\}^{1/2}. \qquad (6)$$

The transition temperature is

$$T_c^* = T_c + \frac{A_4{}^2}{4aA_6}, \qquad (7)$$

where $\varepsilon = T - T_c$, and the disordered phase ($\eta = 0$) terminates its stability at $T = T_c$ while the ordered one (η_0) does so at $T = T_0{}^*$, where

$$T_0{}^* = T_c + \frac{A_4{}^2}{3aA^6}. \qquad (8)$$

Importantly, the existence of a thermal hysteresis phenomenon is associated with first-order phase transition. Thermal hysteresis means that on cooling, the disordered phase is stable below the equilibrium transition temperature, while on heating, the ordered phase is stable above it. The associated effects are called supercooling and superheating, respectively.

JACK A. TUSZYŃSKI

See also **Catastrophe theory; Critical phenomena; Ferromagnetism and ferroelectricity; Order parameters**

Further Reading

Kittel, C. 1956. *Introduction to Solid State Physics*, New York: Wiley

Landau, L.D. & Lifshitz, E.M. 1959. *Statistical Physics*, London: Pergamon

Thompson, J.M.T. & Stewart, H.B. 1986. *Nonlinear Dynamics and Chaos*, New York: Wiley

I

IDEMPOTENT ANALYSIS

Definition and Introduction

The analyses that we all know and use (differentiation, integration, Fourier analysis, matrix analysis, etc.) are tacitly based on algebra defined by means of the conventional operations: addition, multiplication, and their inverses. In idempotent analysis, the starting point is not conventional algebra but instead the idempotent semiring.

Definition 1. *A semiring is a set $\mathcal{D}$ with two binary operations, $\oplus$ (called "sum" or "addition") and $\otimes$ (called "product" or "multiplication") and two specific elements ε and e such that*

- *$(\mathcal{D}, \oplus, \varepsilon)$ is a commutative monoid with identity element ε,*
- *$(\mathcal{D}, \otimes, e)$ is a monoid with identity element e,*
- *$a \otimes \varepsilon = \varepsilon \otimes a = \varepsilon$, for all $a \in \mathcal{D}$,*
- *$a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$, for all $a, b, c \in \mathcal{D}$.*

The semiring is called idempotent if in addition

- *$a \oplus a = a$, for all $a \in \mathcal{D}$.*

The semiring is called commutative if

- *the multiplication $\otimes$ is commutative.*

A monoid here is a set of elements with one operation. An idempotent semiring is sometimes referred to as a "dioid," which means "twice a monoid." It is the property $a \oplus a = a$ that distinguishes a dioid from conventional algebra. The most well-known example of a dioid is the so-called max-plus algebra, where $\mathcal{D} = R \cup \{-\infty\}$, the $\oplus$ operation is the maximum, with identity element $\varepsilon = -\infty$, and the $\otimes$ operation is (conventional) addition, with the identity element $e = 0$. As a numerical example, $5 \otimes (3 \oplus -2) = 8$. A noticeable difference between max-plus algebra and conventional algebra is that linear equations do not always have a solution (e.g., what is x in $5 \oplus x = 4$?). Addition is not cancellative in the sense that $a \oplus b = a \oplus c$ does not in general imply $b = c$.

The pioneering work in the max-plus algebra is Cuninghame-Green (1979), which has been extended into a system-theoretic direction in Baccelli et al. (1992). The phrase *idempotent analysis* was coined by Kolokoltsov & Maslov (1997). In Gunawardena (1998), it is shown that idempotency shows up in many different areas of mathematics; Gunawardena (1998, pp. 1–49) is an excellent survey. Applications are described in Le Boudec & Thiran (2001); and Heidergott et al. (2004).

We choose the max-plus algebra setting to introduce some analysis. Consider the recurrence relation

$$x(k+1) = A \otimes x(k), \tag{1}$$

which represents a model with state $x \in R^n$ and where A is an $n \times n$ matrix with elements a_{ij}. More explicitly, this relation is defined by its component-wise writing

$$\begin{aligned} x_i(k+1) &= \bigoplus_j (a_{ij} \otimes x_j) = (a_{i1} \otimes x_1) \\ &\quad \oplus (a_{i2} \otimes x_2) \oplus \cdots \oplus (a_{in} \otimes x_n), \\ i &= 1, 2, \ldots, n. \end{aligned}$$

The conventional way of writing this would be $x_i(k+1) = \max_j(a_{ij} + x_j)$. Such a model is sometimes referred to as a discrete event dynamic system.

If $x(0)$ is an eigenvector of A in the max-plus algebra defined as

$$A \otimes x(0) = \lambda \otimes x(0), \tag{2}$$

equivalently written as $\max_j(a_{ij} + x_j(0)) = \lambda + x_j$, with not all components $x_i(0)$ equal to ε, and where λ is called the eigenvalue, then $x_i(k+1) = \lambda + x_i(k) = x_i(0) + (k+1)\lambda$. The new state is obtained by adding λ to the previous state. In graph-theoretic terms, λ can be characterized as follows.

Theorem 1. *If the square matrix A is irreducible, equivalently, its corresponding graph is strongly connected and, there exists one and only one eigenvalue (with possibly several eigenvectors). This eigenvalue is equal to the maximum cycle mean of the graph:*

$\lambda = \max_\zeta |\zeta|_w/|\zeta|_l$, *where* ζ *ranges over the set of circuits in the graph, where* $|\zeta|_l$ *is the number of arcs of* ζ *and* $|\zeta|_w$ *its weight, that is, the addition of all* a_{ij} *which correspond to arcs of* ζ.

This theorem is referred to as the spectral theorem. A practical interpretation for this model and analysis is given below.

Examples

The first example of a dioid already encountered is $R \cup \{-\infty\}$ with $\oplus$ as max and $\otimes$ as addition. Other examples are:

- The set of $n \times n$ matrices with $A \oplus B$ and $A \otimes B$ defined as ($A = \{a_{ij}\},\ B = \{b_{ij}\}$)

$$\{A \oplus B\}_{ij} = \max(a_{ij}, b_{ij}),\ \{A \otimes B\}_{ij} = \max_k(a_{ik} + b_{kj}).$$

- The set $R \cup \{-\infty\}$ with min as $\oplus$ and addition as $\otimes$ is isomorphic to the max-plus algebra.
- The set $R \cup \{-\infty\} \cup \{+\infty\}$ with $\oplus$ defined as max and $\otimes$ as min.
- The set {0, 1} with $\oplus$ defined as max and $\otimes$ as min. This is the Boolean algebra.
- The set of all subsets of the R^2 plane, including Ø and the whole R^2 itself, with $\oplus$ defined as $\cup$ and with $\oplus$ defined as the vector sum of subsets.
- Extension to polynomials or formal power series by defining (a polynomial or formal power series is given by $f(z) = \bigoplus_{k=0,1,\ldots} f_k z^k$):

$$f \oplus g:\ (f \oplus g)_k = f_k \oplus g_k;\ f \otimes g:\ (f \otimes g)_k = \max_{i+j=k} f_i \otimes g_j.$$

Relation to Nonlinear Theory and Practice

One must be careful using the word *nonlinear*. System (2) for instance is nonlinear with respect to conventional algebra, but it is considered linear in the max-plus algebra.

Relationship with Other Mathematical Disciplines

An interpretation in graph-theoretic terms is often possible. The occurrences of events in timed event graphs, the latter forming a subclass of Petri nets, can be modeled by means of an equation of the form (1). If $x \in R^n$, define $||x||$ to be its l^∞ norm, that is, $||x|| = \max_i |x_i|$. A mapping $f : R^n \to R^n$ is called non-expansive if $||f(x) - f(y)|| \leq ||x - y||$. As an example, $f(x) = A \otimes x$ is non-expansive. An important theme is the study of periodic solutions of non-expansive mappings; that is, does λ exist such that $f^k(x) = k\lambda + f(x)$? Another theme is that of the finiteness of the transient behavior (before one reaches a periodic behavior), if one starts with an arbitrary initial condition.

Some Applications

Consider the system (2). Interpret a_{ij} as the traveling time of a train on a track (if existing) between departure station j and arrival station i. Quantity $x_i(k+1)$ refers to the earliest departure time of trains at station i that depart for the $(k+1)$th time. The model states that in order to depart (for the $(k+1)$th time), the train must wait until all trains, which departed for the kth time, have arrived; $\max_j(a_{ij} + x_j)$ reflects the maximum of all arrival times. This allows passengers to change trains (one could add some minutes to all a_{ij} to account for the changeover times). If a track between two stations, say l and m, does not exist, set $a_{ml} = \varepsilon$. If one assumes that the departures take place as soon as possible subject to the restriction of passengers changing over, then the model $x(k+1) = A \otimes x(k)$, with an initial condition $x(0)$, determines all future departure times. If this $x(0)$ is an eigenvector, then one has a timetable with constant interdeparture times, this constant being λ. This constant is a lower bound for the interdeparture times of any periodic timetable obeying the changeover rule.

In the above, trains can be viewed as discrete flows on networks. If one replaces "trains" by "messages" and "all railway tracks" by "computer network" or "internet," one enters another field of application that has been described and analyzed in Le Boudec & Thiran (2001).

In discrete-time optimal control theory, or in decision processes on Markov chains, one encounters the equation

$$V(k, x) = \max_u(V(k+1, f(x, u)) + g(x, u)),$$

which is a consequence of Bellman's principle of optimality. The underlying model is $x(k+1) = f(x(k), u(k))$, and the costs during step k are $g(x(k), u(k))$. The function V is the value function. This equation, with the operations of addition and maximization can be interpreted and analyzed in the max-plus algebra sense (Akian et al., 1998). Applications in automata theory can also be found in Gunawardena (1998).

GEERT JAN OLSDER

See also **Markov partitions**

Further Reading

Akian, M., Quadrat, J.-P. & Viot, M. 1998. Duality between probability and optimization. In *Idempotency*, edited by J. Gunawardena, Cambridge and New York: Cambridge University Press

Baccelli, F., Cohen, G., Olsder, G.J. & Quadrat, J.-P. 1992. *Synchronization and Linearity*, New York: Wiley

Cuninghame-Green, R.A. 1979. *Minimax Algebra*, Berlin: Springer

Gunawardena, J. (editor). 1998. *Idempotency*, Cambridge and New York: Cambridge University Press

Heidergott, B., Olsder, G.J. & van der Woude, J.W. 2004. *Max Plus at Work*, Princeton, NJ: Princeton University Press

Kolokoltsov, V.N. & Maslov, V.P. 1997. *Idempotent Analysis and Its Applications*, Boston: Kluwer

Le Boudec, J.-Y. & Thiran, P. 2001. *Network Calculus, A Theory of Deterministic Queuing Systems for the Internet*, Berlin and New York: Springer

IKEDA MAP

See **Maps**

IMPULSE, NERVE

See **Nerve impulses**

INCOHERENT SOLITONS

Until 1996, solitons were considered to be exclusively coherent entities. However, experimental observations of self-trapping of partially spatially incoherent light (Mitchell et al., 1996) and of white light (Mitchell et al., 1997) in photorefractive media have shown that incoherent beams can also form solitons (Figure 1). These solitons are multimode or speckled beams for which the instantaneous intensity distribution varies randomly in space and/or time. They can exist only in non-instantaneous media whose response time is much longer than the random fluctuation time across the incoherent beam. Such a medium responds to the time-averaged envelope and not to the instantaneous speckles that constitute the incoherent beam. An incoherent soliton forms when the time-averaged intensity induces a multimode wave and traps itself in the waveguide by populating the guided modes in a self-consistent fashion. These random-phase and weakly correlated self-trapped entities exhibit a host of unique properties that have no analog in the coherent regime. Moreover, their existence is related to many other areas of physics, in which nonlinearities, stochastic behavior, and statistical averaging are involved. For example, incoherent modulation instability effects and incoherent pattern formation relate to many systems in nature: from clustering in a cooled atomic gas to self-supported "stripes" of electrons in semiconductors, from high-T_c superconductors to structures in fluids and gravitational effects.

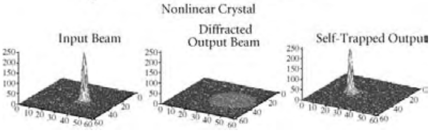

Figure 1. Self-trapping of spatially and temporally incoherent white light from an incandescent light bulb.

Numerous theoretical and experimental works followed the pioneering experiments (see Segev & Christodoulides (2001) and references within). This rapid progress in the area of incoherent solitons offers a number of interesting fundamental ideas as well as possible applications. For example, coherence engineering can be used to control modulation instabilities and to create interesting stable patterns in nonlinear materials for various applications. There is also an exciting possibility of using self-trapped light beams from incoherent sources, such as light-emitting diodes, for creating reconfigurable optical interconnects and beam steering.

Several theories have been developed to describe the propagation of partially incoherent beams and soliton formation in noninstantaneous nonlinear materials (Segev & Christodoulides, 2001). These include the coherent density theory (Christodoulides et al., 1997), the modal theory (Mitchell et al., 1997), and the mutual coherence theory (Shkunov & Anderson, 1998). Other, more approximate theories relying on ray optics (Snyder & Mitchell, 1998) and on the Wigner distribution (Hall et al., 2002) have also been suggested. The first three approaches are rather complete and equivalent to one another (Christodoulides et al., 2001). Here, we briefly overview the two most widely used theories.

Coherent Density Theory

The coherent density method is a dynamic approach suited for studying the evolution dynamics of incoherent solitons, their interactions, instabilities, and correlation statistics. In this formalism, the incoherent field is described by means of a so-called coherent density function, from which one can deduce the optical intensity distribution as well as the associated correlation statistics. Mathematically, the self-trapping process is described by an infinite set of nonlinear Schrödinger-like equations, provided that each coherent component at the input is appropriately weighted with respect to the incoherent angular power spectrum of the source (Christodoulides et al., 1997). These equations can be solved numerically by using modified beam propagation procedures.

Modal Theory

The modal theory is currently widely considered as the method of choice in terms of identifying static incoherent solitons, their range of existence, and correlation properties. The fundamental requirements of incoherent self-trapping were the starting point of this formulation. Incoherent solitons are identified by solving the underlying coupled (stationary) nonlinear Schrödinger-like equations in a self-consistent fashion (Mitchell et al., 1997). To solve these equations self-consistently, one needs to start with an arbitrary index profile, then solve for the guided modes and sum them incoherently to get the intensity profile, and finally use this intensity profile to calculate the new induced-index profile. This process should be repeated until the solutions are stationary. The stationary solutions give the incoherent soliton intensity profile as well as its modal distribution and statistical properties.

1-d and 2-d Dark Incoherent Solitons

The existence of dark incoherent solitons and requirements for their observations were predicted (Coskun et al., 1998) two years after the observation of bright incoherent solitons. Such dark incoherent solitons were experimentally observed soon thereafter (Chen et al., 1998). In that experiment, $(1 + 1)$-dimensional dark incoherent solitons were observed, along with the self-trapping of a $(2 + 1)$-d dark incoherent beam. That is, a self-trapped 2-d void nested in spatially incoherent beam was demonstrated (Chen et al., 1998). As in the coherent case, fundamental dark incoherent solitons were found to exist only in defocusing nonlinear media and to require an initial transverse phase shift at the center of the dark stripe. 2-d dark incoherent solitons were initiated from a vortex beam that was generated by replacing a step mirror with a helicoidal phase mask, which created an optical vortex of unit topological charge. For bright solitons, only guided modes (bound modes) are self-trapped, whereas dark incoherent solitons involve a belt of radiation modes (unbound states). Unlike their coherent counterparts, fundamental dark and vortex solitons were always found to be gray because of the presence of both odd and even radiation modes at the center of the notch (Christodoulides et al., 1998). Furthermore, the grayness depends on the coherence length of the carrier beam. Dark soliton splitting and "phase memory" effects were also observed theoretically and experimentally (Coskun et al., 1999).

Modulation Instability, Pattern Formation, and Soliton Clustering

Both theory and experiment have revealed that modulation instabilities can take place with incoherent light beams provided that the nonlinearity exceeds a well-defined threshold, which depends on the correlation properties of the beam (Soljacic et al., 2000). This shows that modulation instabilities (MI) can be totally suppressed by properly adjusting the correlation distance of the wave front. A natural byproduct of this suppression is the observation of anti-dark incoherent solitons (a bright beam on a constant background) in self-focusing environments. Periodic trains of 1-d filaments and self-ordered 2-d lattices of light spots due to incoherent modulation instabilities were observed by carefully increasing the nonlinearity (Kip et al., 2000). Spontaneous clustering of solitons in partially coherent wave fronts initiated by random noise was also demonstrated experimentally and numerically (Chen et al., 2002). It is, in fact, an outcome of the interplay between random noise, weak correlation, and high nonlinearity. Together, this process leads to incoherent MI, formation of 2-d solitary filaments, and eventually to clustering of 2-d solitons or in other words, aggregates of fine-scale structures. This clustering phenomenon has no counterpart with solitons in coherent systems. The rapid progress in this new area of incoherent solitons brings about many interesting fundamental ideas and possible applications. We have concentrated here on the basic features, but new ideas arise very quickly in this new, dynamic area of soliton science.

TAMER COSKUN, DEMETRIOS CHRISTODOULIDES, AND MORDECHAI SEGEV

See also **Nonlinear optics; Optical fiber communications; Solitons, types of**

Further Reading

Chen, Z., Mitchell, M., Segev, M., Coskun, T.H. & Christodoulides, D.N. 1998. Self-trapping of dark incoherent light beams. *Science*, 280: 889–892

Chen, Z., Sears, S.M., Martin, H., Christodoulides D.N. & Segev, M. 2002. Clustering of solitons in weakly correlated systems. *Proceedings of the US National Academy of Science*, 99(8): 5223–5227

Christodoulides, D.N., Coskun, T.H., Mitchell, M., Chen, Z.C. & Segev, M. 1998. Theory of incoherent dark solitons. *Physical Review Letters*, 80(23): 5113–5116

Christodoulides, D.N., Coskun, T.H., Mitchell, M. & Segev, M. 1997. Theory of incoherent self-focusing in biased photorefractive media. *Physical Review* E, 59(5): R4777–R4780

Christodoulides, D.N., Eugenia, D.E., Coskun, T.H., Segev, M.& Mitchell, M. 2001. Equivalence of three approaches describing partially incoherent wave propagation in inertial nonlinear media. *Physical Review* E, 63: 035601-1–035601-4

Coskun, T.H., Christodoulides, D.N., Chen, Z. & Segev, M. 1999. Dark incoherent soliton splitting and "phase memory" effects: theory and experiment. *Optics Letters*, 23(6): 418–420

Coskun, T.H., Christodoulides, D.N., Mitchell, M., Chen, Z. & Segev, M. 1998. Dynamics of incoherent bright and dark self-trapped beams and their coherence properties in photorefractive crystals. *Optics Letters*, 23(6): 418–420

Hall, B., Lisak, M., Anderson, D., Fedele, R. & Semenov, V.E. 2002. Statistical theory for incoherent light propagation in nonlinear media. *Physical Review* E, 65: 035602-1–035602-4

Kip D., Soljacic M., Segev, M., Eugenieva, E., Christodoulides, D.N. 2000. Modulation instability and pattern formation in spatially incoherent light beams. *Science*, 290: 495–498

Mitchell, M., Chen, Z., Shih, M. & Segev, M. 1996. Self-trapping of partially spatially incoherent light. *Physical Review Letters*, 77(3): 490–493

Mitchell, M. & Segev, M. 1997. Self-trapping of incoherent white light. *Nature*, 387: 880–883

Mitchell, M., Segev, M., Coskun, T.H. & Christodoulides, D.N. 1997. Theory of self-trapped spatially incoherent light beams. *Physical Review Letters*, 79(25): 4990–4993

Segev, M. & Christodoulides, D.N. 2001. Incoherent solitons: self-trapping of weakly-correlated wave-packets. In *Spatial Solitons*, edited by S. Trillo & W.E. Torruellas, New York: Springer

Shkunov, V.V. & Anderson, D.Z. 1998. Radiation transfer model of self-trapping spatially incoherent radiation by nonlinear media. *Physical Review Letters*, 81(13): 2683–2686

Snyder, A.W. & Mitchell, D.J. 1998. Big incoherent solitons. *Physical Review Letters*, 80(7): 1422–1924

Soljacic, M., Segev, M., Coskun, T.H., Christodoulides, D.N. & Vishwanath, A. 2000. Modulation instability of incoherent beams in noninstantaneous nonlinear media. *Physical Review Letters*, 84(3): 467–470

INERTIAL MANIFOLDS

In studying the dynamics of nonlinear systems, it is sometimes observed that all trajectories rapidly converge to a subspace of the phase space, leading to a simplification of the analysis through dimensional reduction. Such subspaces are called inertial manifolds (IMs) and the underlying dynamics are called inertial. In phase spaces of infinite dimension, the question is whether a finite-dimensional dynamical system can reproduce the dynamics of an infinite-dimensional one. In finite-dimensional systems, one tries to find a system of smaller dimension.

Related concepts have been introduced in sciences. In meteorology, for example, the slow manifold introduced and studied by Lorenz and others, is expected to reproduce the slow dynamics, taking into account interaction with the rapidly changing regions of the phase space. From a somewhat different perspective, the Reynolds equations in turbulence and some large eddies simulation models are finite-dimensional systems that attempt to reproduce the dynamics of the large eddies of the fluid, taking into account the action of the small eddies.

Definition

Denote by H the phase space, a Hilbert space of finite or infinite dimension, and consider on H a semigroup of operators $\{S(t)\}_{t\geq 0}$ generating the dynamical system: $u_0 \mapsto S(t)u_0 = u(t)$.

If it exists, an inertial manifold for this system is a smooth finite-dimensional manifold $\mathcal{M}$ in H that enjoys the following properties:

(i) $S(t)\mathcal{M} \subset \mathcal{M}$, $\forall t > 0$ (positive invariance);
(ii) all orbits converge to $\mathcal{M}$, as $t \to \infty$, at an exponential rate;
(iii) the rate of convergence is uniform for all orbits starting in a bounded set of initial data.

If a global attractor $\mathcal{A}$ exists, then necessarily $\mathcal{A}$ is included in $\mathcal{M}$ as well as any stationary solution, periodic orbit, etc. Comparing $\mathcal{A}$ to $\mathcal{M}$ when they both exist, we see that $S(t)\mathcal{A} = \mathcal{A}$, instead of (i), and we know that orbits may converge to $\mathcal{A}$ at arbitrarily slow rates, in contrast with (ii). Finally, attractors are believed to be sometimes fractal sets, whereas smoothness is required for $\mathcal{M}$.

When an inertial manifold exists, the restriction of the semigroup to $\mathcal{M}$ is a group—due to (i)—whose dynamics are the same as that of the initial systems. In particular, their attractor is the same if $\{S(t)\}_{t\geq 0}$ has an attractor, and this reduced dynamical system is called an inertial system. When the dynamical system is defined by a differential equation, as in Equation (1), its restriction to $\mathcal{M}$ is a finite-differential system, also called an inertial system.

Examples

Inertial manifolds have been studied for dynamical systems in infinite dimensions corresponding to dissipative parabolic evolution equations, written in functional form as

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = F(u(t)), \quad u(0) = u_0. \tag{1}$$

Here $u(t) = S(t)u_0$, where u_0 and $u(t)$ belong to H.

The space H is decomposed as the sum of a finite-dimensional space $Y = P_y H$ and of its complement $Z = P_z H$, $u = y + z$. In all examples studied, the inertial manifold is then searched for as a graph:

$$z = \Phi(y), \quad y \in Y. \tag{2}$$

Usually, Y and Z are spectral spaces for the underlying linear dissipative operator A, Y representing hence the low-frequency component of the flow and z its high-frequency component.

The existence of inertial manifolds has been proven for the Kuramoto–Sivashinsky equation, the Ginzburg–Landau equation, reaction-diffusion equations in one and two spatial dimensions, and the Navier–Stokes equations with added (hyper-)viscosity.

The initial hope (Foias et al., 1988) was to find an IM for the Navier–Stokes equation at least in space two, justifying the name coined by Foias, since the role of the inertial terms over the viscous terms would be dominant on this manifold. However, the existence of an inertial manifold for the Navier–Stokes equations has not yet been proven nor disproven.

Several proofs of existence have been derived beyond those of Foias et al. (1988); the key hypothesis and limitation in all the proofs is the so-called spectral gap condition, involving a sufficiently large gap in the spectrum of the dissipative operator A. It is believed (but not proved) that this difficulty could be overcome with IMs that are less smooth and are not graphs. Alternatively, this difficulty has been overcome by removing the condition that $\mathcal{M}$ be a manifold. The corresponding concept was introduced in Eden et al. (1994) where it was called an inertial set or an exponential attractor.

The existence of inertial sets was proven in Eden et al. (1994) for numerous dynamical systems generated by evolutionary dissipative equations, in fact, for all systems for which the existence of a compact finite-dimensional attractor has been established (Hale, 1988; Sell & You, 2002; Temam, 1997).

Despite these limitations, inertial manifolds are a natural concept, giving a precise form to the idea that infinite-dimensional dynamical systems produce a finite-dimensional dynamics. Remember that a less explicit form of this idea is the property of the attractor to be finite-dimensional.

When IMs do not exist or are not easy to construct, approximate inertial manifolds (AIMs) are sometimes used, which are finite-dimensional smooth manifolds that attract the orbits in a small neighborhood. Extensive applications of AIMs in numerical analysis have been developed (Temam, 1997; Sell & You, 2002).

ROGER TEMAM

See also **Center manifold reduction; Invariant manifolds and sets; Synergetics**

Further Reading

Eden, A., Foias, C., Nicolaenko, B. & Temam, R. 1994. *Exponential Attractors for Dissipative Evolution Equations*, Collection Recherches en Mathématiques Appliquées, Paris: Masson, Paris, and New York: Wiley

Foias, C., Sell, G.R. & Temam, R. 1988. Inertial manifolds for nonlinear evolutionary equations, *Journal of Differential Equations*, 73: 309–353

Hale, J. 1988. *Asymptotic Behavior of Dissipative Systems*, Providence, RI: American Mathematical Society

Sell, G.R. & You, Y. 2002. *Dynamics of Evolutionary Equations,* New York: Springer

Temam, R. 1988. *Infinite Dimensional Dynamical Systems in Mechanics and Physics,* New York: Springer (2nd augmented edition, 1997)

INFELD EQUATION

See **Born–Infeld equation**

INFORMATION DIMENSION

See **Dimensions**

INFORMATION THEORY

Although the word *information* is commonly used and its engineering applications influence our lives, the quantitative meaning is often unclear. Among several measures of information, the most common is called Shannon information (denoted as I_S with units of the bit), which was introduced by Claude Shannon in 1948.

As an example, suppose your friend (whose wife has been pregnant for nearly nine months) telephones with the information: It's a girl. Since the prior probability (P) of a girl being born is $\frac{1}{2}$, the Shannon information is computed as

$$I_S = \log_b \left(\frac{1}{P}\right) = \log_2 \left(\frac{1}{1/2}\right) = 1\,\text{bit}, \tag{1}$$

with the logarithmic base $b = 2$. In other words, one bit of Shannon information is gained in a message that resolves one of two equally probable events.

Strings of such events (e.g., 100110100011100) are the language of computers, and their Shannon information can be computed by thinking of a string as a set of N switches, each of which can be on (1) or off (0). Since there are 2^N possible sequences of N switches, the probability of any particular sequence is 2^{-N}, leading to a Shannon information of N bits, which is conveniently equal to the number of switches.

Shannon's information takes the particular form, Equation (1), under the assumption that the received message (an observed switch sequence) is completely reliable. A more general Shannon measure allows for generally unreliable messages; thus

$$\begin{aligned}&(\text{source event } B_1 \text{ or } B_2 \text{ or } \ldots B_N)\\ &\quad\rightarrow \text{observed message } A.\end{aligned} \tag{2}$$

Here a particular source event $B_n \equiv B$ occurs, and the observer receives a message A, providing evidence of what was actually intended. The arrow implies transmission through a real, imperfect communication channel, in which transmission errors may occur. For example, the actual source event B might be the switch sequence given above that is transmitted to an observer. Because of transmission errors, the received message might be $A = 010110100011100$.

As when a detective observes a fingerprint, the message A implies that an event B has occurred, where each B_n is a different possible switch sequence. Shannon information was designed to answer the question: How much evidence does the observed (garbled) message A constitute for the hypothesis that the event B actually occurred?

Intuitively, the evidence provided by A will be strong if the probability of B having occurred is now greater than before receiving A. That is, if the probability $P(B|A)$ (of B in the presence of A) much exceeds $P(B)$ (the probability of B in the absence of A). This suggests that the ratio $P(B|A)/P(B)$ measures the weight of the evidence. Strong evidence would give a large ratio, weak evidence a small ratio. Because any monotonic function of this ratio works as well, taking the logarithm defines the Shannon information

$$I_S(A, B) \equiv I_S \equiv \log_b \frac{P(B|A)}{P(B)}. \tag{3}$$

Although the choice of a logarithmic base b is at the user's discretion, probabilities in computer codes are often powers of 2 for which $b = 2$ is the most convenient choice. Then I_S is assigned the unit of a *bit*. For a fixed channel probability $(P(B|A))$, Equation (3) discloses that the information increases as $P(B)$ becomes smaller. (In editing a newspaper, for example, an article entitled "Dog Bites Man" is not unusual enough to merit publication, but "Man Bites Dog" most definitely is.)

There is a problem with this definition of information—the observer, by hypothesis, does not know which event B_n has occurred. Therefore, he or she cannot know which information quantity, $I_S(A, B_1)$ or … or $I_S(A, B_N)$ to compute. Furthermore, the message A may be one of a set of possibilities A_m, $m = 1, \ldots, M$. All such possibilities can be taken into account by forming the grand average

$$\begin{aligned}&\left\langle \log_b \frac{P(B|A)}{P(B)} \right\rangle\\ &\equiv \sum_{m=1}^{M}\sum_{n=1}^{N} P(A_m, B_n) \log_b \frac{P(B_n|A_m)}{P(B_n)}\\ &\equiv I(A, B). \qquad (4)\end{aligned}$$

Called Shannon's "mutual information," this is the quantity commonly used to evaluate the quality of communication systems. Notice that it depends upon properties of the source (the $P(B_n)$) as well as the channel (the $P(B_n|A_m)$); thus, this is a total system perspective.

If just the quality of the channel (as specified by its fixed channel probabilities $P(B_n|A_m)$) is to be evaluated, the information $I(A, B)$ is maximized through choice of a source $P(B_n)$. The resulting information, then, only depends upon the channel probabilities and is called the channel capacity since it measures the intrinsic capacity of the channel for transmitting information.

Note that Equation (4) may be evaluated in its continuous limit; thus,

$$I(X, Y) = \int\int \mathrm{d}x\,\mathrm{d}y\; p(x, y) \log_b \frac{p(y|x)}{p(y)}. \qquad (5)$$

Is there an information measure that is appropriate for describing growth processes? Consider a generalization of Equation (5) to

$$I_{\mathrm{KL}}(p, r) \equiv -\int \mathrm{d}\boldsymbol{x}'\; p(\boldsymbol{x}') \log_b \frac{p(\boldsymbol{x}')}{r(\boldsymbol{x}')}, \qquad (6)$$

which is called the Kullback–Leibler (K–L) information. The ratio p/r is describes a "cross entropy" between the probability law $p(\mathbf{x}')$ and the reference law $r(\mathbf{x}')$, where the vector $\mathbf{x}'$ is general. In the special case $\mathbf{x}' = (\mathbf{x}, \mathbf{y})$, $r(\mathbf{x}') = p(\mathbf{x})p(\mathbf{y})$, Equation (6) becomes the mutual information in Equation (5). Another important case of the K–L information is when $r(\mathbf{x}') = 1$, whereupon Equation (6) takes the form

$$-\int \mathrm{d}\boldsymbol{x}' p(\boldsymbol{x}') \log_b p(\boldsymbol{x}') \equiv H, \qquad (7)$$

which is called the entropy.

Yet another special case of K–L information is obtained when $\mathbf{x}' \equiv x$, a single coordinate, and the reference function $r(x)$ is made to be a slightly displaced version of the probability law, $r(x) \equiv p(x+\Delta x)$, $\Delta x \to 0$. In this limit, the K–L information is found to obey

$$\begin{aligned}&-\lim_{\Delta x\to 0} \frac{1}{\Delta x^2} \int \mathrm{d}x\; p(x) \log_b \frac{p(x)}{p(x+\Delta x)}\\ &= -\frac{1}{2}\int \mathrm{d}x \frac{p'^2(x)}{p(x)} \equiv -\frac{1}{2} I_{\mathrm{F}}, \qquad (8)\end{aligned}$$

where I_{F} is the Fisher information and the usual derivative notation $p'(x) \equiv \mathrm{d}p(x)/\mathrm{d}x$ is used.

A useful variant of I_{F} is obtained by using a probability amplitude $q(x)$, where $p(x) \equiv q^2(x)$. The right-hand integral of Equation (8) then gives directly

$$I_{\mathrm{F}} = 4\int \mathrm{d}x\, q'^2(x), \quad q'(x) \equiv \mathrm{d}q(x)/\mathrm{d}x. \qquad (9)$$

Thus, Fisher information measures the gradient content of the probability amplitude curve.

Because the smallest possible mean-square error one can achieve in forming an estimate of the parameter is the reciprocal of Fisher's information, this also describes our ability to know a parameter from its measurements. Assuming all knowledge is based upon measurement, Fisher information describes quite generally our ability to know.

Physicists Leon Brillouin and David Bohm long sought a measure of information that would explain the mysteries of physics in general, including those of quantum mechanics. In particular, the question of quantum entanglement has long perplexed physicists. Consider a pair of particles that are created simultaneously as reactants from a mother particle. Measurement of one member of the pair supplies important information (such as spin) about the other member, even without measuring the latter. (This is called the Einstein–Podolsky–Rosen–Bohm, or EPRB, paradox.) In what way could information about the unmeasured particle be transmitted into a measurement of the detected one? This seems to require the realities of the two particles to be somehow "entangled." Hence, an "active information" was sought that could supply the needed information. Of course, they tried the Shannon variety, but this did not work.

Surprisingly, a form of information exists for generating the probability laws of physics that is aptly called the physical information and defined as follows. Let I be the Fisher information that is acquired in an actual measurement. Depending upon the measurement scenario, there may or may not be another, unmeasured particle accompanying the measured particle. The physical information K is the difference $K \equiv I - J$ of two Fisher informations (for simplicity we now drop the subscript F), where

$$I \equiv \sum_n I_n,\; I_n = 4\int \mathrm{d}x\, q_n'^2(x)$$

and

$$J \equiv I/\kappa, \quad 0 \leq \kappa \leq 1. \qquad (10)$$

Notice that the information I in the measurement is actually a sum of individual Fisher terms I_n. It turns out that the additive sum also defines the maximum possible value of the Fisher information. Hence, I is called the Fisher information channel capacity (as in the Shannon case). The integral form of the second Equation (10) shows that I is also a functional of the amplitude functions $q_n(x)$. This is a scalar number whose value depends upon an integral over all values x of the functions $q_n(x)$.

By the third Equation (10), J is likewise a functional of the amplitude functions; however, its form is not fixed but dependent upon the particular measurement scenario. J must be computed in any application.

Recall that I is the information level of the particle in a measurement space. In all measurements of quantum phenomena $\kappa = 1$, so $J = I$ by the third Equation (10). Thus J represents information that arises out of entanglement with that of the measured particle. We, therefore, call J the entangled information. Mathematically, it is I as evaluated in an "entanglement space." If only the measured particle is present, entanglement space is the Fourier conjugate space to measurement space. If another, unmeasured particle is present, entanglement space is the (unused) measurement space of that particle.

In Equations (10), the form of I is a sum of squares of functions q'_n (an L^2 norm or measure). Like any squared length, such a norm is invariant to rotation, and rotations may be performed in either the space of coordinates x (which are more generally of dimension 4 or higher) or of the functions $q'_n(x)$. Because J is linear in I, it also obeys such invariance, and likewise for their difference K. Such invariance suggests a state of stationarity

$$K \equiv I - J = \text{extrem.} \qquad (11)$$

Equation (10), supplemented by Equations (11), is called the principle of extreme physical information or EPI (Frieden, 1998). Its solutions $q_n(x)$ obey corresponding Euler–Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial k}{\partial q'_n}\right) = \frac{\partial k}{\partial q_n}, \quad k \equiv k(x),$$
$$K \equiv \int \mathrm{d}x\, k(x). \qquad (12)$$

These are wave equations whose solutions $q_n(x)$ are the amplitude functions of science. An example is the Schrodinger wave equation. Other examples are provided by growth phenomena, ranging from energy growth to crystal growth to cancer growth, with the preceding formulation leading to exponential, power law, or logarithmic growth depending on the boundary conditions and integration constants that are assumed (Gatenby & Frieden, 2002).

In summary, Shannon information and cross entropy are valuable forms of information in communication theory, and two important outgrowths of these definitions are Fisher information and physical information. Fisher information permits one to estimate the expected error in the estimate of a parameter. Physical information permits one to derive physical laws via the EPI principle.

B. Roy Frieden

See also **Algorithmic complexity; Entropy; Euler–Lagrange equations**

Further Reading

Bohm, D. & Hiley, B.J. 1993. *The Undivided Universe*, London: Routledge

Brillouin, L. 1956. *Science and Information Theory*, New York: Academic Press

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, 222: 309–368

Frieden, B.R. 1998. *Physics from Fisher Information*, Cambridge and New York: Cambridge University Press (see also 2nd edition, 2004)

Gatenby, R. & Frieden, B.R. 2002. Application of information theory and extreme physical information to carcinogenesis. *Cancer Research*, 62: 4675

Kullback, S. 1959. *Information Theory and Statistics*, New York: Wiley

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423; 623–656

INHIBITION

In what is certainly one of the best-known experiments on learning, Ivan Pavlov sounded a bell and shortly after presented food to a hungry dog. After several pairings of the bell and the food, the bell came to elicit salivation, a response that it had not elicited previously. We say that the bell become an excitatory conditioned stimulus. *Excitatory* means that it acquired the ability to excite, arouse, or elicit a new behavior. But Pavlov did not limit his work to excitatory conditioning; he performed and described (Pavlov, 1927) a variety of procedures that would endow the conditioned stimulus with exactly the opposite properties. Such conditioned stimuli acquire the ability to oppose conditioned excitation and are said to be inhibitory.

The term originates from the Latin word *inhibere* and means to restrain. *Inhibere* consists of *in* "in" and *habere* "to hold." The word *inhibition* is widely used in many disciplines, such as medicine, biochemistry, physiology, neuroscience, and psychology, and its general meaning is the arrest or restraint of a process, partial or complete. In psychoanalytic theory, it describes a conscious or unconscious restraining of an impulse or desire. In neurophysiology, to which we will return shortly in more detail, it is the process opposite to

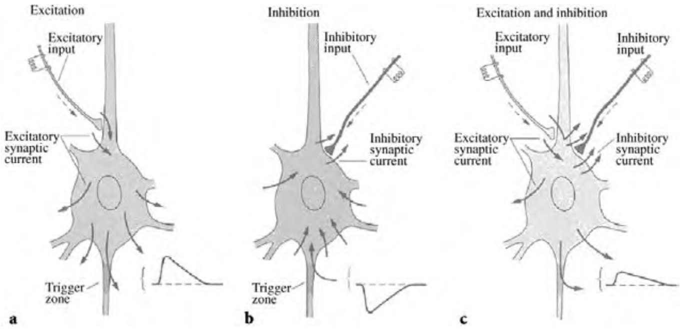
Excitation
Excitatory input
Excitatory synaptic current
Trigger zone
a
Inhibition
Inhibitory input
Inhibitory synaptic current
Trigger zone
b
Excitation and inhibition
Excitatory input
Inhibitory input
Excitatory synaptic current
Inhibitory synaptic current
c

Figure 1. Excitatory and inhibitory currents have competitive effects on a single nerve cell. (a) An excitatory input in the base of a dendrite causes inward current which produces a large synaptic potential that brings the membrane potential close to its threshold. (b) An inhibitory input causes synaptic potential thus generated changes the membrane potential further below its threshold. (c) The shunting action of inhibition. When the cell receives both excitatory and inhibitory synaptic currents, the channels opened by the inhibitory pathway shunt the excitatory current, thereby reducing the excitatory synaptic potential. From Kandel et al. (2000).

excitation, or as Charles Sherrington (1933) wrote "... to refrain from an act is no less an act than to commit one, because inhibition is co-equally with excitation a nervous activity." By "nervous activity" he meant "the activity of the nervous system."

In fact, the function of all living systems and at all levels of complexity can be perceived as continuous adjustment between excitatory and inhibitory processes. The metabolic pathways are, for example, regulated by means of enzyme synthesis and inhibition. Almost all of the thousands of chemical reactions that occur in a living organism require catalysis by a specific enzyme. The character of any cell is based on its particular chemistry, which is determined by its specific enzyme composition. Genetic traits are expressed through synthesis of enzymes, which catalyze reactions that establish the phenotype. Many genetic diseases result from altered levels of enzyme production with changes in enzyme synthesis and/or inhibition. Much current drug therapy is based on the inhibition of a specific enzyme.

The nervous system continuously processes excitatory and inhibitory signals. Sherrington first pointed out that the quintessential action of the nervous system is its ability to weigh the consequences of different types of information and then decide on appropriate responses (Sherrington, 1947). He named it "the integrative action of the nervous system." The point at which one neuron communicates with another is called a synapse, and synaptic transmission is fundamental to processes such as perception, voluntary movement, or learning. Although many synaptic connections are highly specialized, all neurons make use of two basic forms of synaptic transmission: electrical and chemical.

Each neuron in the central nervous system, whether in the spinal cord or in the brain, is continuously bombarded by synaptic inputs from other neurons. Some are excitatory, others inhibitory. The effect of the synaptic potential—whether it is excitatory or inhibitory—is determined, not by the type of transmitter released from the presynaptic neuron, but by the type of ion channels gated by the transmitter in the postsynaptic cell. Most transmitters are recognized by types of receptors that mediate either excitatory or inhibitory potentials (for details see Kandel et al., 2000).

Synaptic contact can occur on the cell body, the dendrites, or the axon of a postsynaptic cell. The location of inhibitory inputs in relation to excitatory ones is critical for their functional effectiveness. Inhibitory short-circuit actions are more significant when they are initiated at the cell body near the initial axon segment. As a result, they can strongly curtail the influence of excitatory current on the membrane potential at the trigger zone, as illustrated in Figure 1c. In contrast, inhibitory action at a remote part of a dendrite is much less effective in shunting excitatory actions or in affecting the more distant trigger zones. Thus, in the brain, significant inhibitory input frequently occurs on the cell bodies of neurons.

Excitatory and inhibitory signals are integrated into a single response by the cell. In motor neurons and most interneurons, the decision to initiate an action potential is made by the initial segment of the axon (Figure 1, trigger zone). This region of cell membrane has a lower threshold for action potentials than the cell body or dendrites because it has a high density of voltage-dependent Na^+ channels. The neural integration involves the summation of synaptic potentials that spread passively to the trigger zone. It is critically affected by two passive membrane

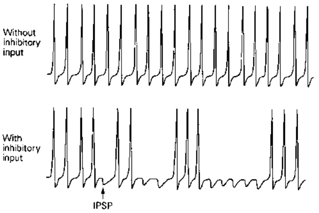

Figure 2. The *sculpturing* effect of inhibition. Without inhibitory input the neuron spontaneously fires at a fixed interval. With inhibitory input that produces an inhibitory postsynaptic potential (IPSP), some action potentials are inhibited, resulting in a distinctive pattern of impulses. From Kandel et al. (2000).

properties. First, the *time constant* affects *temporal summation*, the process by which consecutive synaptic potentials at the same site are added together in the postsynaptic cell. The larger the time constant, the greater the likelihood that two consecutive inputs from an excitatory presynaptic neuron will sum to bring the cell membrane to its threshold for an action potential. Second, the *length constant* affects *spatial summation*. Neurons with a large length constant are more likely to be brought to threshold by two different inputs arising from different sites than are neurons with a short space constant.

In addition to counteracting synaptic excitation, synaptic inhibition can exert control over spontaneously active nerve cells. Many cells in the brain are spontaneously active, as are the pacemaker cells of the heart. By suppressing the spontaneous generation of action potential in these cells, synaptic inhibition can shape the pattern of firing in a cell (Figure 2).

Many mathematical models have been introduced to describe the function of the nervous system. Their complexity ranges from that of individual nerve cells to assemblies of neurons. Hodgkin & Huxley (1952) constructed a model, known as the Hodgkin–Huxley model, to explain the electrical excitability of nerve axons in terms of discrete Na^+ and K^+ currents. The Hodgkin–Huxley model is probably the best-known and the most elegant model of a biological system, exemplifying a balance between experiment and theory that has rarely been matched. To explain how trains of impulses occur in the Hodgkin–Huxley model FitzHugh (1961) used the Hodgkin–Huxley model as one member of a large class of nonlinear systems showing excitable and oscillatory behavior. The model was further analyzed by an approximately distributed pulse transition line using analog simulations (Nagumo et al., 1962) and today is known as FitzHugh–Nagumo model.

Wilson and Cowan (1972) further increased the complexity of the system. The model they proposed was based on the properties of populations, and the single cell activity was represented not by a single spike, but rather by the spike frequency. They based their model on two main assumptions. First, the existence of spatially localized neuronal populations, based on physiological and anatomical evidence that many cells in a small volumes of cortical tissue have very nearly identical responses to identical stimuli. Second, all nervous processes of any complexity are dependent upon the interaction of excitatory and inhibitory cells. Wilson and Cowan were the first to treat inhibition as arising from exclusively inhibitory neurons. Wilson–Cowan oscillators are now widely used to describe higher functions, such as sensory information processing, learning, memory, or pattern recognition. Neural networks and chips are built for signal and image processing.

Much progress has been made in recent years toward a detailed understanding of the complex and nonlinear dynamics that underlie the Hodgkin–Huxley, FitzHugh–Nagumo, and Wilson–Cowan models. In addition, a large number of new models of synchronous oscillations have been introduced, attempting to describe features observed in the electrical activity of the brain as measured by EEG or MEG. They range from those involving details of receptors and inter-inhibitory interactions, to models depending on various nonlinear principles (see, for example, Haken, 2002).

Aneta Stefanovska

See also **FitzHugh–Nagumo equation; Flip-flop circuit; Hodgkin–Huxley equations; Integrate and fire neuron; Multiplex neuron; Myelinated nerves; Nerve impulses; Neurons**

Further Reading

FitzHugh, R. 1961. Impulses and physiological states in models of nerve membrane. *Biophysical Journal*, 1: 445–466

Haken, H. 2002. *Brain Dynamics. Synchronization and Activity Patterns in Pulse-Coupled Neural Nets with Delays and Noise*, Berlin and New York: Springer

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117: 500–544

Kandel, E.R., Schwartz, J.H. & Jessell, T.M. (editors). 2000. *Principles of Neural Science*, 4nd edition, New York: McGraw-Hill

Nagumo, J.S., Arimoto, S. & Yoshizawa, S. 1962. An active pulse transmission line simulating nerve axon. *Proceedings of IRE*, 50: 2061–2071

Pavlov, I.P. 1927. *Conditioned Reflexes*, London: Oxford University Press (original Russian edition 1926)

Sherrington, C. 1933. *The Brain and Its Mechanism*, REDE Lecture, Cambridge: Cambridge University Press

Sherrington, C. 1947. *The Integrative Action of the Nervous System*, 2nd edition, New Haven: Yale University Press

Wilson, H.R. & Cowan, J.D. 1972. Excitatory and inhibitory interactions in localised populations of model neurons. *Biophysical Journal*, 12: 1–24

INSTABILITY

See **Stability**

INSTANTONS

A remarkable feature of the Yang–Mills action is that there are finite-action topological soliton solutions to the classical field equations. These solitons are known as instantons or, in early papers, as pseudoparticles. The Yang–Mills action is important in particle physics; in particular, it describes the behavior of gluons, the particles that carry the strong nuclear force. Before instanton solutions were discovered in 1975 by Belavin et al. (1975), the Yang–Mills theory of the strong force appeared to have a symmetry not found in nature. This was known as the axial U(1) problem and was solved by 't Hooft who realized that one effect of the instanton solutions was to break this unwanted symmetry. This was the first example of an extended classical solution having a physical consequence in a field theory of particle physics.

In four Euclidean dimensions, the pure SU(2) Yang–Mills action is

$$S[A] = -\frac{1}{2}\int \mathrm{d}^4x \,\mathrm{trace}\, F_{\mu\nu}F_{\mu\nu}, \tag{1}$$

where μ and ν are space indices running from 1 to 4 and repeated indices are summed. $F_{\mu\nu}$ is a field strength tensor, a skew Hermitian 2×2 matrix field in four-dimensional space, related to a gauge potential A_μ by

$$F_{\mu\nu} = \frac{\partial}{\partial x_\mu}A_\nu - \frac{\partial}{\partial x_\nu}A_\mu + [A_\mu, A_\nu]. \tag{2}$$

The action is invariant under local gauge transformations

$$A_\mu \to A'_\mu = gA_\mu g^{-1} - \frac{\partial g}{\partial x_\mu}g^{-1}, \tag{3}$$

where g is a special unitary matrix function in four dimensions. Under this transformation $F_{\mu\nu} \to F'_{\mu\nu} = gF_{\mu\nu}g^{-1}$. One key consequence of this is that $F_{\mu\nu}$ not only vanishes when $A_\mu = 0$, but also, whenever A_μ is a gauge transformation of zero. This means that the asymptotic condition required for finiteness of the action is that A_μ approaches a gauge transformation of zero:

$$A_\mu \to -\frac{\partial g}{\partial x_\mu}g^{-1} \tag{4}$$

for $r = \sqrt{x_\mu x_\mu} \to \infty$. Now, the group of special unitary 2×2 matrices is topologically a three-sphere, and so the gauge potential on the large three-sphere at infinity gives a map from that three-sphere to the three-sphere of values of g. Maps between three-sphere are classified topologically by a single integer: the winding number. This integer, therefore, classifies finite-action field configurations; in fact, it is the only gauge-invariant information determined by the asymptotic field behavior. The topologically nontrivial stationary points of the action $S[A]$ are instantons.

Since they are stationary points of the action, instantons obey the Euler–Lagrange equation for the action:

$$\frac{\partial}{\partial x_\mu}F_{\mu\nu} = 0. \tag{5}$$

This is known as the Yang–Mills equation. It can be shown that minimal action solutions obey a first-order equation called the self-dual equation:

$$F_{\mu\nu} = \tfrac{1}{2}\varepsilon_{\mu\nu\lambda\sigma}F_{\lambda\sigma}, \tag{6}$$

where $\varepsilon_{\mu\nu\lambda\sigma}$ is totally skewsymmetric in its indices with $\varepsilon_{1234} = 1$. Solutions to this equation obey the second-order Yang–Mills equation. A solution with winding number one is given by

$$A_\mu = -\frac{r^2}{r^2 + R^2}\frac{\partial g}{\partial x_\mu}g^{-1}, \tag{7}$$

where R is an arbitrary scale and g has the hedgehog form

$$g = \frac{x_4}{r}\mathbf{I}_2 + \frac{i}{r}\left(x_1\sigma_1 + x_2\sigma_2 + x_3\sigma_3\right), \tag{8}$$

where $\mathbf{I}_2$ is the 2×2 identity matrix and the σ_i's are the Pauli matrices. Allowing for gauge equivalence, there is an eight-dimensional space of winding number one instantons, four of these dimensions correspond to the choice of position in space, one to the choice of scale, and three to an overall group orientation. Although these results are particular to SU(2), other groups can be studied; the main difference is in the number of overall group orientation parameters.

The self-dual equations are integrable. In fact, many of the integrable equations of mathematical physics can be derived from the self-dual equations by demanding that the solution has some translational or rotational symmetry. Because of this integrability, solutions to the self-dual equations with winding numbers greater than one may be constructed, in principle, by twistor theory and by algebraic methods (Christ et al., 1978). There is also a simple ansatz (Jackiw et al., 1977), but this does not give a general solution.

In the 1980s, the self-dual equations revolutionized the study of smooth four-dimensional manifolds: by studying the space of solutions to the self-dual equations over different manifolds, it is possible to derive invariants, known as the Donaldson invariants (Donaldson & Kronheimer, 1990). This approach has since been largely superceded by Seiberg–Witten theory.

In the path integral approach to quantum field theory, physical values are derived by certain weighted integrations over all possible field configurations. These path integrals are often intractable, and it is common to expand the integration about the stationary points of the action. This works because of the way the action appears in the integrand of the path integral. Calculations of this type are known as semi-classical calculations because they are effectively an expansion in Planck's constant, $\hbar$. This expansion must include a sum over all possible stationary points, and so, in Yang–Mills theory, it includes a sum over instanton configurations. By studying the symmetry properties of the measure in quantum chromodynamics (QCD), the Yang–Mills theory describing the strong nuclear force, it can be shown that terms in this expansion break the axial U(1) symmetry. This symmetry is unbroken if the instanton terms are omitted. The symmetry breaking allows processes that violate baryon and lepton conservation; however, the amplitudes for these effects are highly suppressed. One useful approach is to consider these processes as tunneling events between different vacua, in fact, instanton calculations in quantum field theory are very similar to Wentzel–Kramers–Brillouin (WKB) calculations in quantum mechanics.

While instantons provide a qualitative explanation for a host of phenomena in quantum chromodynamics, useful quantitative results are not available within the semi-classical approach, and even qualitatively, instanton calculations are not rigorous since they relate only to the semi-classical approximation, a truncation of the full quantum theory. The modern approach to quantum chromodynamics is lattice QCD. It is possible within lattice QCD to verify the original ideas about the role of instantons in the physics of the strong force; however, there are limits to the precision with which the lattice and semi-classical approaches can be compared (Negele, 1998).

Finite action soliton solutions in other equation systems are sometimes referred to as instantons. Examples include the finite action solutions to the forced Burgers equation, which arises in the study of turbulence, and the vortex-like solutions in the abelian Higgs model, which is related to condensed matter physics. Instantons have many similarities to the lump solitons found in certain sigma models.

Conor Houghton

See also **Burgers equation; Higgs boson; Particles and antiparticles; Quantum field theory; Solitons; Yang–Mills theory**

Further Reading

Belavin, A.A., Polyakov, A.M., Schwarz, A.S. & Tyupkin, Yu.S. 1975. Pseudoparticle solutions of the Yang–Mills equations. *Physics Letters* B, 59: 85–87

Christ, N.H., Weinberg, E.J. & Stanton, N.K. 1978. General self-dual Yang–Mills solutions. *Physical Review* D, 18: 2013–2025 (Includes a review of the complete solution to the self-dual equations originally due to Atiyah, Hitchin, Drinfeld, and Manin)

Coleman, S. 1985. *Aspects of Symmetries*, Cambridge and New York: Cambridge University Press (Chapter 7 gives a celebrated description of instantons in QCD)

Donaldson, S.K. & Kronheimer, P.B. 1990. *The Geometry of Four-manifolds*, Oxford: Clarendon Press and New York: Oxford University Press

Jackiw, R., Nohl, C. & Rebbi, C. 1977. Conformal properties of pseudoparticle configurations. *Physical Review* D, 15: 1642–1646 (A large, but incomplete, family of solutions to the self-dual equations.)

Negele, J.W. 1998. Instantons, the QCD vacuum, and hadronic physics. *Nuclear Physics* B *Proceedings Supplements*, 73: 92–104

't Hooft, G. 1976. Symmetry breaking through Bell-Jackiw anomalies. *Physical Review Letters*, 37: 8–11

Weinberg, S. 1996. *Quantum Field Theory*, vol. 2, Cambridge and New York: Cambridge University Press (Chapter 23 concerns extended field configurations in particle physics and has a very clear treatment of instantons.)

INTEGRABILITY

Central problems in the integrability theory for nonlinear ordinary differential equations (ODEs) or partial differential equations (PDEs) are knowing which systems can be solved analytically and developing appropriate solution techniques. Although in some contexts the term *integrability* is not well defined, let us begin by considering two physically motivated nonlinear ODE systems that can be analytically solved.

Nonlinear Pendulum

Here, the nonlinear system is

$$\ddot{x} = ax + bx^2 + cx^3, \quad a, b, c = \text{constants}. \qquad (1)$$

Multiplying both sides of this equation by $\dot{x}$ and integrating yields an intermediate integral $I = \dot{x}^2/2 - ax^2/2 - bx^3/3 - cx^4/4$, which is determined by the initial conditions: $x(t_0)$ and $\dot{x}(t_0)$. In terms of I, the solution $x(t)$ can be written implicitly as

$$t - t_0 = \int_{x_0}^{x} \frac{\mathrm{d}x}{\sqrt{2I + ax^2 + 2bx^3/3 + cx^4/2}}, \qquad (2)$$

where $x(t_0) = x_0$. The inversion of this formula is expressed in terms of Jacobi elliptic functions (Bryd & Friedman, 1954). Note that this system has one degree of freedom, and one intermediate integral was sufficient to obtain a solution. Because it is related to an area, the intermediate integral (I) is sometimes called a quadrature, and by extension, Equation (2) is said to be obtained by the method of quadratures.

Calogero–Moser N-Body Problem

In this example, the nonlinear system is

$$\ddot{q}_n = \sum_{m=1, m\neq n}^{N} 2g^2(q_n - q_m)^{-3}, \quad g = \text{constant}, \quad (3)$$

generalizations of which were studied extensively in the last three decades of the 20th century. Using a Lax isospectral deformation technique, the solution matrix $\operatorname{diag}[q_1(t), \dots, q_N(t)]$ is found to be similar to the matrix $\tilde{Q} = [\tilde{Q}_{nm}(t)]$ defined by

$$\begin{aligned}\tilde{Q}_{nm}(t) &= \delta_{nm}\big[q_n(0) + \dot{q}_n(0)t\big] \\ &\quad + \mathrm{i}(1-\delta_{nm})g\big[q_n(0) - q_m(0)\big]^{-1}t. \quad (4)\end{aligned}$$

Thus, the solution to the initial value (Cauchy) problem is reduced to the algebraic task of finding the N eigenvalues of the matrix $\tilde{Q}$, and the solution formula for initial data, $q_i(0)$ and $\dot{q}_i(0)$, can be constructed from linear functions, through a finite number of algebraic operations and compositions of functions (Calogero, 2001). Equations (1) and (3) are completely integrable.

In the 19th century, Evariste Galois, Niels Henrik Abel, Joseph Liouville, and Sophus Lie tried to rationalize the process of solving differential equations by quadratures, or failing that, massaging them in such a way that useful information might be extracted. Mainly based on geometry and algebra, these efforts initiated much of the mathematics—such as classifications of differential equations in term of symmetries and conservation laws—that has dominated the 20th century. Attempts by Karl Weierstrass and Henri Poincaré to create a systematic theory of integrability are based on complex function theory.

An achievement of 19th-century mathematics was the elaboration of the theory of elliptic and Abelian functions, particularly the introduction of theta functions. A solution comprising such rapidly convergent power series is quite efficient computationally; thus, a natural question for 19th-century mathematicians was this: Which differential equations admit solutions as quotients of power series that converge in large regions, independent of initial or boundary conditions? The Cauchy–Kovalevsky theorem gives local existence for such expressions. This theme traces through the work of Sophia Kovalevsky and Paul Painlevé for nonlinear ODEs and through Bernhard Riemann and Immanuel Fuchs and their successors for linear ODEs (Hermann, 1984).

There are several results about local and global solutions of differential equations that establish existence, uniqueness, smoothness, stability, approximations, and so on; yet attempts to find exact solution formulas in the 19th century were largely unsuccessful. This is because solutions to nonlinear systems may exhibit complex behaviors—such as blow-up, shocks, chaos, fractals, and bifurcation—which are obstacles to integrability. Poincaré put an end to the search for new integrable equations by showing that among all dynamical systems integrable ones are exceptional. Indeed, a small structural perturbation of an integrable system often destroys integrability.

Although Poincaré's results dampened interest in the search for new integrable equations during the first half of the 20th century, the situation changed dramatically with the discovery of the inverse scattering transform as a method of solving the initial value (Cauchy) problem for the Korteweg–de Vries (KdV) equation (Gardner et al., 1967). Rapidly extended to several other nonlinear systems of scientific interest (nonlinear Schrödinger equation, sine-Gordon equation, Toda lattice, and so on), this discovery also led to the emergence of "soliton factories" during the 1980s (Zakharov, 1991).

More generally, integrability theory aims to find global information on solutions and, if possible, to solve differential equations analytically. Integrability itself is an intrinsic characteristic of differential equations, imposing constraints on the way solutions evolve in phase space and suggesting the following working definition.

Definition A differential equation is completely integrable if all solutions to well-posed initial or boundary value problems can be presented beginning with elementary functions, using finitely many algebraic operations and compositions of functions, and evaluating limits.

Algebraic operations include inverting or diagonalizing matrices; compositions of functions include inverses; and the limits can be integrals, infinite series, or asymptotes. Thus, this definition holds for cases in which solutions can be constructed explicitly. It also generalizes the notion of integrability by quadratures (Liouville integrability) that only requires computation of intermediate integrals in addition to algebraic operations and compositions of functions.

For ODEs, the Liouville–Arnol'd theorem on finite-dimensional Hamiltonian systems gives sufficient conditions, called the Liouville conditions, for guaranteeing the integrability by quadratures (Arnol'd, 1989). For a Hamiltonian system of N degrees of freedom, these conditions are:

- the existence of N integrals of motion $\{F_1, F_2, \dots, F_N\}$,
- that are functionally independent on the level surface $\{F_i = a_i\}$ containing the initial data, and
- that commute with each other under the associated Poisson bracket.

If the level surface $\{F_i = a_i\}$ is compact and connected, then the Liouville–Arnol'd theorem says that the underlying Hamiltonian system can be expressed as

$$\dot{I}_i = 0, \; \dot{\varphi}_i = \omega_i(I_1, \dots, I_N), \quad (5)$$

in the action-angle coordinate system (I_i, φ_i) of a neighborhood of the level surface. Thus, the motion is conditionally periodic along an N-dimensional torus with the frequencies ω_i, and the solution to the Cauchy problem can be completely determined by the method of quadratures. The Euler top, the Lagrange top, Kowalewski's top, the Stäckel systems, and geodesic flows on a surface are Liouville-integrable systems (Perelomov, 1990).

For PDEs, there also exist some results on integrability properties from a Hamiltonian perspective, including the bi-Hamiltonian theory (Magri, 1978) from which infinitely many symmetries and conservation laws for PDEs can be deduced. Motivated by the Liouville–Arnol'd theorem, a notion of integrability for PDEs is the existence of infinitely many conservation laws. The KdV equation is integrable in this sense, and it can also be put into a Hamiltonian form in action-angle coordinates like integrable ODEs (Zakharov & Faddeev, 1971).

The existence of infinitely many conservation laws can lead to infinitely many finite-dimensional solution varieties that can be determined analytically (Fuchssteiner, 1992), but it is not clear whether these subvarieties generate the whole infinite-dimensional solution variety. Indeed, the infinitely many degrees of freedom of PDEs introduce large diversity, obscuring their solvability and integrability properties.

The notion of partial integrability arises when systems of differential equations (both ODEs and PDEs) possess more degrees of freedom than conservation laws (Conte & Boccara, 1990). Symmetry constraints developed in soliton theory (Ma & Zhou, 2001) suggest a way to show partial integrability for PDEs through relating PDEs to integrable ODEs. The method was motivated by Moser's work on Hill's equation (Moser, 1980) and nonlinearization of Lax pairs (Cao, 1990). The results generalize the theory of finite-dimensional integrable stationary equations, suggesting the possibility of establishing a Liouville–Arnol'd theorem for infinite-dimensional Hamiltonian systems.

Since the 1970s, several criteria have arisen for testing the integrability of nonlinear PDEs, and corresponding theories include infinitely many symmetries, infinitely many conservation laws, Lax structure, bi-Hamiltonian structure, the Bäcklund transform, Hirota's bilinear form, the inverse scattering transform, and the Painlevé property (Degasperis, 1998; Gu, 1995).

An ODE or a PDE is said to possess the Painlevé property, respectively, if the movable singularities of solutions of the ODE are only poles (Ablowitz & Clarkson, 1991) or if solutions of the PDE are "single-valued" in the neighborhood of noncharacteristic, movable singularity manifolds (Weiss et al., 1983), suggesting that the Painlevé property is an indication of complete integrability (Conte, 1999). One advantage of the Painlevé property is that it can be a tool for computing Lax pairs and conservation laws. Lax pairs, as we have seen in Equation (3), provide linear objects for solving Cauchy problems of nonlinear equations. A thorough understanding of the relationship between Painlevé singularities and integrability may suggest alternatives to the KAM notion of near-integrability (Zakharov, 1991).

The term S-integrability implies integrability of Cauchy problems by the inverse spectral transform technique (Calogero & Degasperis, 1982) or the inverse scattering transform (Ablowitz & Clarkson, 1991; Fokas, 1997). C-integrability, on the other hand, means that a differential equation can be transformed into a linear one by an appropriate change of variables. Examples of C-integrable equations are the Burgers equation and the Calogero–Dagasperis–Ibragimov–Shabat equation:

$$u_t = u_{xx} - 2uu_x, \tag{B}$$

$$u_t = u_{xxx} + 3u^2u_{xx} + 9uu_x^2 + 3u^4u_x. \tag{CDIS}$$

These two equations are, respectively, cast into the heat equation ($v_t = v_{xx}$) and the linear KdV equation ($v_t = v_{xxx}$) under appropriate dependent variable transformations. Calogero and colleagues have made a systematic study of the construction and classification of C-integrable equations, both in $1 + 1$ dimensions and $N + 1$ dimensions (Zakharov, 1991). Although the property of C-integrability often leads to an infinity of conservation laws, the Burgers equation has only one local conservation law of differential polynomial type.

WEN-XIU MA

See also **Burgers equation; Constants of motion and conservation laws; Hamiltonian systems; Hirota's method; Inverse scattering method or transform; Painlevé analysis; Solitons**

Further Reading

Ablowitz, M.J. & Clarkson, P.A. 1991. *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, Cambridge and New York: Cambridge University Press

Arnol'd, V.I. 1989. *Mathematical Methods of Classical Mechanics*, 2nd edition, New York: Springer

Bryd, P.F. & Friedman, M.D. 1954. *Handbook of Elliptic Integrals*, New York: Springer

Calogero, F. 2001. *Classical Many-Body Problems Amendable to Exact Treatments*, Berlin and Heidelberg: Springer

Calogero, F. & Degasperis, A. 1982. *Spectral Transform and Solitons*, vol. I. *Tools to Solve and Investigate Nonlinear Evolution Equations*, Amsterdam and New York: North-Holland

Cao, C.W. 1990. Nonlinearization of the Lax system for AKNS hierarchy. *Science in China, Series A: Mathematics, Physics, Astronomy*, 33: 528–536

Conte, R. (editor). 1999. *The Painlevé Property, One Century Later*, New York: Springer

Conte, R. & Boccara, N. (editors). 1990. *Partially Integrable Evolution Equations in Physics*, Dordrecht: Kluwer

Degasperis, A. 1998. Resource letter Sol-1: solitons. *American Journal of Physics*, 66: 486–497

Fokas, A.S. 1997. A unified transform method for solving linear and certain nonlinear PDEs. *Proceedings of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences*, 453: 1411–1443

Fuchssteiner, B. 1992. Hamiltonian structure and integrability. In *Nonlinear Equations in the Applied Sciences*, Boston, MA: Academic Press, pp. 211–256

Gardner, C.S., Greene, J.M., Kruskal, M.D. & Miura, R.M. 1967. Method for solving the Korteweg–de Vries equation. *Physical Review Letters*, 19: 1095–1097

Gu, C.H. (editor). 1995. *Soliton Theory and Its Applications*, Berlin and Heidelberg: Springer and Hangzhou: Zhejiang Science and Technology Publishing House

Hermann, R. 1984. *Topics in the Geometric Theory of Integrable Mechanical Systems*, Brookline, MA: Math Science Press

Ma, W.-X. & Zhou, Z. X. 2001. Binary symmetry constraints of N-wave interaction equations in $1+1$ and $2+1$ dimensions. *Journal of Mathematical Physics*, 42: 4345–4382

Magri, F. 1978. A simple model of the integrable Hamiltonian equation. *Journal of Mathematical Physics*, 19: 1156–1162

Moser, J. 1980. Various aspects of integrable Hamiltonian systems. In *Dynamical Systems*, Boston: Birkhäuser, pp. 233–289

Perelomov, A.M. 1990. *Integrable Systems of Classical Mechanics and Lie Algebras*, vol. I, Basel: Birkhäuser

Weiss, J., Tabor, M. & Carnevale, G. 1983. The Painlevé property for partial differential equations. *Journal of Mathematical Physics*, 24: 522–526

Zakharov, V. E. (editor). 1991. *What Is Integrability?* Berlin and Heidelberg: Springer

Zakharov, V.E. & Faddeev, L.D. 1971. The Korteweg–de Vries equation is a fully integrable Hamiltonian system. *Functional Analysis and Its Applications*, 5(4): 18–27

INTEGRABLE CELLULAR AUTOMATA

Integrable cellular automata (CAs) are discrete dynamical systems that possess attributes of integrability including conserved quantities, symmetries, and localized solutions. They show how to discretize a given integrable system of differential equations, maintaining the integrability property, suggesting that integrability of discrete systems and their coherent structures are inherently connected with the iterated string processing performed by automata.

The propagating solutions (or coherent objects) on certain cellular automata have been found to exhibit nondestructive collisions similar to those observed for soliton systems like the Korteweg–de Vries equation. In such cases, pulse-like disturbances propagating along a uniform nonlinear medium are represented by strings of symbols (zero-one patterns), passing through a one-way structure—the pipeline of identical automata M. The analysis of these moving objects reduces to investigating the repeated automaton action over strings, also called an iterated automaton map (IAM); $M(a^t) = a^{t+1}, t = 0, 1, \ldots$.

All known models capable of supporting such discrete localized structures are described by automata (Siwak, 2001, 2002), and a method (called ultradiscretization) leads to discrete systems in which solutions of some familiar soliton equations are preserved as soliton-like patterns (Tokihiro et al., 1999). Thus, an IAM is a fundamental discrete mechanism that supports localized soliton-like periodic structures. This is why a new term *iterons* (Siwak, 2001, 2002) has been proposed for these objects. Note also that fractals owe their existence to the iterating process of some maps.

There are two classes of iterons. The first consists of particles well known in cellular automata models (Delorme & Mazoyer, 1999) where the parallel processing of strings occurs. The second class is not widely known and consists of so-called filtrons. The filtrons are emergent coherent objects occurring in serial processing of strings (Siwak, 2001).

In parallel processing of a string a^t at a given time t all symbols a_i^{t+1} of the next string a^{t+1} are updated simultaneously, for example, by the function $a_i^{t+1} = f(a_{i-r}^t, \ldots, a_i^t, \ldots, a_{i+r}^t)$. When the same function f is used for all positions i, one has a 1-dimensional CA, with f being called its local function or rule. The arguments of f are determined by the so-called neighborhood window $N_i = (-r, \ldots, +r)$; here N_i designates r neighbors on both sides of position i.

The listing of consecutive strings a^t for $t = 0, 1, \ldots$ one under another forms what is called a space-time (ST) diagram. This diagram visualizes the evolution of a given string a^0—the dynamics of a CA system in phase space for initial global state a^0. Occasionally, some moving and periodic patterns of symbols or segments of a string are seen on ST diagrams of CA processing. These are just particles or signals of CAs (Delorme & Mazoyer, 1999).

The serial processing of strings is performed by a computational model called a finite (state) automaton. The Mealy type automaton is described by $M = (S, \Sigma, \Omega, \delta, \beta, s_0)$, where S, Σ, and Ω are nonempty finite sets of—respectively—states, inputs, and outputs; $\delta\colon S \times \Sigma \to S$ is called the next state function of M, $\beta\colon S \times \Sigma \to \Omega$ is called the output function of M, and s_0 is the initial (extinction) state of the automaton. The automaton converts sequences of symbols, preserving their length. Any input string is read sequentially from left to right, one symbol at each instant of time (pulse of clock). For all $\tau = 1, 2, \ldots$ the automaton:

(i) reads input symbol $\sigma(\tau)$,
(ii) changes its current state $s(\tau)$ onto the next one according to $\delta(s(\tau), \sigma(\tau)) = s(\tau + 1)$, and
(iii) generates the symbol $\beta(s(\tau), \sigma(\tau)) = \omega(\tau)$ of the resulting string.

Thus the complete one step conversion at $a^t \to a^{t+1}$ requires a series of clock pulses τ. To consider IAMs, one has to assume the unified input-output alphabet $A = \Sigma = \Omega = 0, 1, \ldots, m$. Then the automaton's operation can be described by a

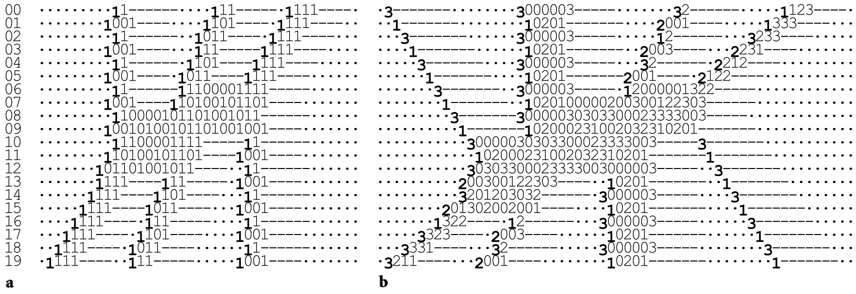

Figure 1. Nondestructive collisions of filtrons ($q = 2$). (a) PST model. (b) Automaton M_1.

sequence of operations $f_s\colon A \to A$, each f_s such that $f_s(a) = \beta(s, a)$.

Coherent objects of serial string processing were not known for a long time. The first model that established a bridge between discrete systems of string processing and soliton-like phenomena was the filter CA (Park et al., 1986). The PST (Park, Steiglitz, & Thurston) model was defined for $A = \{0, 1\}$ by a special space-time window (FCA window) and a nonlinear updating Boolean function $f_{\mathrm{PST}} : A^{2r+1} \to A$. It operates by moving the window to the right along a given string a^t. At each position the function f_{PST} computes consecutive elements

$$a_i^{t+1} = f_{\mathrm{PST}}\left(a_i^t, a_{i+1}^t, \ldots, a_{i+r}^t, , a_{i-r}^{t+1}, a_{i-r+1}^{t+1}, \ldots a_{i-1}^{t+1}\right)$$

of the resulting string a^{t+1} in such a way that $a_i^{t+1} = 1$ if and only if the sum of all window elements (which are arguments of f_{PST}) is even but not zero. It is assumed that the conversion $a^t \to a^{t+1}$ starts somewhere at left with the zero content of window (zero boundary conditions).

The PST model is an automaton and is classified as a digital parity infinite impulse response (IIR) filter. The dynamics of the PST model reveal discrete moving coherent objects capable of colliding in a nondestructive way. An example of such objects is shown in Figure 1(a).

Formal analysis of filter CAs has been given by Goldberg (1988). Goldberg has also given its first automaton-like description by means of a rapid updating rule.

Various generalizations of this model soon appeared. New updating functions were introduced, and the FCA window was modified (Jiang, 1992). A model with an infinite FCA window was given by Takahashi and Satsuma and called a soliton CA. Then, the working alphabet A was enlarged beyond binary (Fokas et al., 1990), and algebraic group operations were considered in computations of updating function.

Another approach was based on deriving the filter-CA-like description from the equations of motion of some systems and from so-called spectral problems (Bruschi et al., 1992).

It was observed that most of the models based on the FCA window perform very special cyclic processing. This led Siwak (2002) to a class of filter automata (FAs). FAs are described by three notions that are useful in physical interpretation of the IAM as a medium, namely:

— the activating function $\iota : A \to \{\mathit{false}, \mathit{true}\}$ such that $\iota(a) = \mathit{true} \leftrightarrow \delta(s_0, a) \neq s_0$;
— the cyclic sequence $(h_j, f_1, \ldots, f_k)$ of mappings f_s (the medium is excited); and
— the extinction conditions (the automaton returns to its extinction state s_0).

Note that the automaton equivalent of PST processing is activated by $a = 1$ (nonzero symbol). Its cyclic sequence of operations f_s is of the form $(N, A, \ldots, A) = NA^r$ where N denotes negate operation, and A is identity id (or accept). The extinction condition is based on a coincidence of some string's current segment $w \in \{a0^r\}$ with the cycle of activity.

The example of IAM dynamics of some cyclic automaton M_1 for $A = \{0, 1, 2, 3\}$ is shown in Figure 1(b). When the automaton M_1 is extinguished, it rewrites input zeroes; $h_0(0) = 0$, but activated by a symbol $j \neq 0$, it performs the cycles of operations $(h_0, f_1, f_2, f_3), (h_j, f_1, f_2, f_3), \ldots$ where

$$h_0 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}, h_1 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 0 & 2 & 2 \end{pmatrix},$$

$$h_2 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 2 & 3 & 0 & 1 \end{pmatrix}, h_3 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 2 & 3 & 0 \end{pmatrix},$$

$$f_1 = f_3 = h_3 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 1 & 2 & 3 \end{pmatrix} = \mathrm{id},$$

$$f_2 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 1 & 3 & 2 \end{pmatrix}.$$

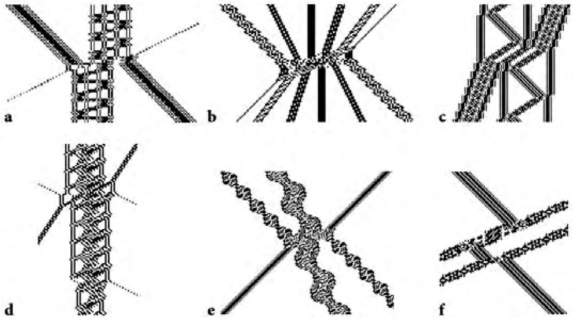

Figure 2. Exemplary images of IAM dynamics—various phenomena of colliding filtrons.

Extinction occurs when the segment $w = 0000$ is read within a cycle.

The symbols of A are shown on ST diagram in order to expose so-called M-segments within strings. Each M-segment begins with an activating symbol, which is printed in bold. The tail zeroes of an M-segment, that is all consecutive zeros immediately preceding the extinction of M are denoted by bar. All zeroes read by the automaton in its state s_0 are shown as dots, and all remaining zeroes as digits 0. With this convention one can recognize whether any two M-segments are distant, adjacent, or overlap and actually form a complex object. Also, a shift q per row to the left is sometimes applied to ST diagrams.

The filtron x of automaton M within the IAM dynamics is defined to be periodic M-segment of the string $a^0 = \ldots 0x0 \ldots$.

Among the discrete models supporting coherent objects there are also the following. For most of them the appropriate equivalent automata have been identified (Siwak, 2001, 2002):

- — iterated arrays of finite state automata (Atrubin, 1965);
- — filter CA, soliton CA, and other generalizations of PST model based on FCA window;
- — digital recursive (IIR) filters and filter automata family (Siwak, 2001);
- — classical CAs and higher order CAs (Delorme & Mazoyer, 1999);
- — CAs with mixed serial and parallel scenarios of their updating schedule;
- — rapid-updating rule-type procedures called also fast rules;
- — integrable CAs derived from so-called spectral problems (Bruschi et al., 1992), and discrete versions of some classical soliton equations;
- — box–ball systems introduced by D. Takahashi;
- — crystal systems and combinatorial data processing;
- — techniques of accelerating the convergence of numerical algorithms;
- — affine Lie algebras.

Let us illustrate the dynamics of IAM with automata $M_{\mathrm{BBSC}}(m, n)$, which are equivalent to the model called the box-ball system with carrier (BBSC). In BBSC each position i represents a box, and each symbol $0 < a_i \le m$ from A is interpreted as the number of balls. A single step of the evolution of a string is performed by a carrier of capacity n that runs to the right once and moves each ball exactly once. At each box, the carrier puts as many balls into the box as possible and simultaneously takes as many balls from the box as possible.

This means that if the carrier contains $0 \le c_j \le n$ balls at box j, then it will take $\min(a_j, n - c_j)$ balls from this box and, simultaneously, will leave $\min(c_j, m - a_j)$ balls. The finite state automaton $M_{\mathrm{BBSC}}(m, n)$ that performs exactly the same string processing as this BBSC is such that: the set of states $S = \{0, 1, \ldots, n\}$ and the alphabet of input/output symbols $A = \{0, 1, \ldots, m\}$ are sets of integers; $s_0 = 0$ is the initial state of $M_{\mathrm{BBSC}}(m, n)$; and the next states of $M_{\mathrm{BBSC}}(m, n)$ are given by $\delta(s, \sigma) = s + \min(\bar{s}, \sigma) - \min(s, \bar{\sigma})$, while the output symbols are determined by $\omega = \beta(s, \sigma) = \sigma + \min(s, \bar{\sigma}) - min(\bar{s}, \sigma)$. Here $\bar{s} = n - s$ and $\bar{\sigma} = m - \sigma$.

An example of colliding filtrons of the automaton $M_{\mathrm{BBSC}}(16, 22)$ is given in Figure 2.

Further examples of IAM dynamics for some other automata are given in Figure 3 where among the multi-filtron nondestructive (solitonic) collisions the following phenomena can be especially seen: complex filtrons (a), (c), jumps of the filtron over a bundle (a), (b), trapped or bouncing filtrons and collision of such complexes (c), decaying object (or quasifiltron) and its earlier solitonic collision (d), vibrating filtrons (e), steady filtrons (a), (b), and attraction and repelling of filtrons (f). Also the fusion, annihilation of filtrons and

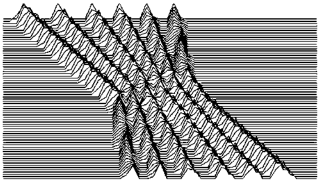

Figure 3. Six filtrons of the automaton $M_{\rm BBSC}(16, 22)$ collide solitonically (shift $q = 1$).

many other events are possible and can be seen in computer simulations—the realm of the behaviors of filtrons of automata is fascinating and amazingly rich.

The automaton approach and explanation of discrete coherent structures phenomena by IAMs leads to many applications, including computing by means of streams of filtrons (solitonic processing idea of Steiglitz), transmission and conversion of information via fiber optical lines, and signals and waves processing (e.g., by an interaction of electromagnetic waves with a shock wave background or another precisely shaped medium).

PAWEL SIWAK

See also **Cellular automata; Cellular nonlinear networks; Game of life; Integrable lattices; Solitons, types of**

Further Reading

Atrubin, A.J. 1965. An iterative one-dimensional real-time multiplier. *IEEE Transactions on Electronic Computers*, EC-14: 394–399

Adamatzky, A. (editor). 2002. *Collision-Based Computing*, London: Springer

Bruschi, M., Santini, P.M. & Ragnisco, O. 1992. Integrable cellular automata. *Physics Letters* A, 169: 151–160

Delorme, M. & Mazoyer, J. (editors). 1999. *Cellular Automata*, Dordrecht: Kluwer

Fokas, A.S., Papadopoulou, E.P. & Saridakis, Y.G. 1990. Coherent structures in cellular automata. *Physics Letters* A, 147(7): 369–379

Jiang, Z. 1992. An energy-conserved solitonic cellular automaton. *Journal of Physics* A: *Mathematical and General*, 25(11): 3369–3381

Park, J.K., Steiglitz, K. & Thurston, W.P. 1986. Soliton-like behavior in automata. *Physica* D, 19: 423–432

Siwak, P. 2001. Soliton-like dynamics of filtrons of cyclic automata. *Inverse Problems*, 17(4): 897–918

Siwak, P. 2002. Iterons of automata. In *Collision-Based Computing*, edited by A. Adamatzky, London: Springer, pp. 299–353

Tokihiro, T., Nagai, A. & Satsuma, J. 1999. Proof of solitonical nature of box and ball systems by means of inverse ultra-discretization. *Inverse Problems*, 15: 1639–1662

INTEGRABLE LATTICES

Lattices, or differential-difference equations (DDEs), are a special class of ordinary differential equations, with an infinite number of dependent variables $q_n = q_n(t)$ numbered by integer indices n, admitting a translational invariance with respect to the shift $n \to n + 1$. Due to this property, such equations are well suited for describing processes in translationally symmetric systems like crystals. By imposing suitable boundary conditions, such as periodic ones ($q_n(t) \equiv q_{n+N}(t)$ for all n), one can also model finite systems. In the notation above the dependent variable t plays the role of time; the range where the dependent variables q_n take values can be a multidimensional vector space or even a more general manifold. Sometimes it is convenient to think of n as the second, discrete independent variable, so that $q_n(t) = q(t, n\Delta x)$ is thought of as a sampling of a function $q(t, x)$ of two continuous variables (time t and space coordinate x) for the discrete sequence of values of $x = n\Delta x$. In such a context differential-difference systems appear as semi-discretizations of systems of partial differential equations (PDEs). For instance, the lattice model

$$\ddot{q}_n = \omega_0^2(q_{n+1} - 2q_n + q_{n-1}) - \sin(q_n), \qquad (1)$$

used by Yakov Frenkel and Tatiana Kontorova in their study (1939) of the relation between dislocation and deformation of crystals, can be considered as a semi-discretization of the sine-Gordon (SG) equation $q_{tt} - q_{xx} = -\sin(q)$ (in this interpretation $\omega_0 = 1/\Delta x$). Subsequent developments revealed that the discrete model (1) is non-integrable, unlike its continuous counterpart SG. Thus, in hindsight, already this early example illustrates the difficult problem of integrable discretization: how to discretize one or several independent variables in a given integrable system, maintaining the integrability property.

Another early example of a physically important system of DDEs is the model used by Enrico Fermi, John Pasta and Stan Ulam in their 1955 study of the dynamics of energy partition in crystals:

$$\begin{aligned}\ddot{q}_n &= -\omega_0^2(q_{n+1} - 2q_n + q_{n-1}) \\ &\quad -\alpha\big((q_{n+1} - q_n)^2 - (q_n - q_{n-1})^2\big). \qquad (2)\end{aligned}$$

In this model, only nearest neighbors interact with the potential depending on their relative displacement, $U(\Delta q) = \frac{\omega_0^2}{2}(\Delta q)^2 + \frac{\alpha}{3}(\Delta q)^3$. Also this model is, strictly speaking, non-integrable. However, by an interesting historical twist, the studies by Fermi, Pasta, and Ulam played the precursor role for the theory of solitons. Presumably, due to the closeness of Equation (2) to integrable models, it demonstrated a lack of thermalization (the FPU recurrence phenomenon), and in attempts to understand this, Kruskal and Zabusky studied the KdV equation and eventually discovered its solitons (*See* **Fermi–Pasta–Ulam oscillator chain**).

The story of integrable lattices began with the discovery by Toda (1989) of the lattice with an exponential interaction of nearest neighbors, which carries nowadays his name (*See* **Toda lattice**):

$$\ddot{q}_n = \mathrm{e}^{q_{n+1}-q_n} - \mathrm{e}^{q_n-q_{n-1}}. \tag{3}$$

He found a number of exact soliton-like solutions of this systems, both in the infinite and in the periodic setting. Multisoliton solutions were constructed in 1973 by Hirota with the help of his ingenious "direct" (or bilinear) method (*See* **Hirota's method**). The integrability of this model in the sense of classical mechanics (existence of an infinite number of independent integrals of motion, which are in involution with respect to the standard Poisson bracket with canonically conjugate momenta $p_n = \dot{q}_n$), as well as in the sense of the solitons theory (Lax representation and applicability of the inverse scattering, or inverse spectral, transformation (IST)) was demonstrated in 1974 independently by Flaschka and by Manakov. Other important early work on this system includes the treatment of scattering in the finite nonperiodic Toda lattice by Moser (1975), finding the Bäcklund transformation by Wadati and Toda (1975), and the solution of the periodic lattice in terms of multidimensional theta-functions given independently by Date and Tanaka and by Krichever (1976). One has a rare possibility to read the first-hand presentation of the whole body of relevant results, including the authentic story of the original discovery, in Toda (1989).

Once the mechanism of integrability via IST for lattice systems was uncovered, new integrable DDEs began to be discovered in quick succession. The next one was the lattice

$$\dot{a}_n = a_n(a_{n+1} - a_{n-1}), \tag{4}$$

having plenty of applications and known today under various names (Volterra, Kac–van Moerbeke, discrete KdV, etc.). Its integrability was established independently by Manakov (1974) and by Kac and van Moerbeke (1975), while the multisoliton solutions for it and for its modification,

$$\dot{a}_n = (1 + a_n^2)(a_{n+1} - a_{n-1}), \tag{5}$$

had been found by Hirota in 1973.

An important paper by Ablowitz & Ladik (1976) contained a lattice version of the zero-curvature (Zakharov–Shabat) representation of integrable systems and integration methods based on it. This deepened understanding of the algebraic structures behind integrable lattices, and opened an opportunity to construct in a unified manner arbitrarily many examples. One of the celebrated systems introduced and studied there is an integrable spatial discretization of the nonlinear Schrödinger equation (DNLS):

$$\begin{aligned} \mathrm{i}\dot{\psi}_n &= -(\psi_{n+1} - 2\psi_n + \psi_{n-1})/(\Delta x)^2 + \nu \mid \psi_n \mid^2 \\ &\quad \times(\psi_{n+1} + \psi_{n-1}). \end{aligned} \tag{6}$$

A different but gauge equivalent discretization of NLS was found by Isergin and Korepin in 1981 (*See* Faddeev & Takhtajan, 1987 and **Discrete nonlinear Schrödinger equations**).

Dozens of integrable lattices have been found since then (which, of course, has led to a certain devaluation of this sort of activity). It seems, however, that only a few of them are of really fundamental importance for theory and applications. Such a "short list" depends naturally on personal tastes, but the following systems (apart from those already mentioned) should be included in any case.

A sort of a relativistic generalization of the Toda lattice is due to Ruijsenaars (1990):

$$\begin{aligned} \ddot{q}_n &= (1 + \alpha\dot{q}_{n+1})(1 + \alpha\dot{q}_n)\, \frac{\mathrm{e}^{q_{n+1}-q_n}}{1 + \alpha^2\mathrm{e}^{q_{n+1}-q_n}} \\ &\quad -(1 + \alpha\dot{q}_n)(1 + \alpha\dot{q}_{n-1})\, \frac{\mathrm{e}^{q_n-q_{n-1}}}{1 + \alpha^2\mathrm{e}^{q_n-q_{n-1}}}. \end{aligned} \tag{7}$$

A Toda-type lattice with nearest neighbors interactions that encounter elliptic functions but depends not only on their relative displacements was discovered by Shabat & Yamilov (1990) and independently by Krichever (2000):

$$\ddot{q}_n = (\dot{q}_n^2 - 1)\big(V(q_n, q_{n+1}) + V(q_n, q_{n-1})\big), \tag{8}$$

where $V(q, q') = \zeta(q + q') + \zeta(q - q') - \zeta(2q)$ is an elliptic function in both arguments q, q' (here $\zeta(q)$ is the Weierstrass ζ-function).

A neat spatial discretization of the Landau–Lifshitz (LL) equation has been derived by Adler (2000):

$$\begin{aligned} \dot{s}_n &= \langle s_n, Js_n\rangle \left(\frac{s_n \times s_{n+1}}{1 + \langle s_n, s_{n+1}\rangle} + \frac{s_n \times s_{n-1}}{1 + \langle s_n, s_{n-1}\rangle} \right) \\ &\quad -2s_n \times (Js_n). \end{aligned} \tag{9}$$

Here $s_n \in S^2$, and J is a constant 3×3 matrix. In the isotropic case $J = I$ and this model reduces to a discrete Heisenberg magnetic discovered independently by Sklyanin and by Ishimori in 1982. Sklyanin found also a discretization of the LL equation different from (9); the relation between both models is discussed in Adler (2000).

In the rest of this article we briefly discuss some mathematical aspects of integrable lattices. We illustrate them mainly with the example of the Toda lattice (3), but almost all of this holds in one form or another for all integrable DDEs.

Lax Representations

The results by Flaschka and by Manakov of 1974 serve as a first application of IST in the differential-difference context. It is based upon the Lax representation of the Toda lattice, that is, the fact that (3) is equivalent to an operator equation

$$\dot{L} = [L, A_\pm], \tag{10}$$

in terms of linear difference operators L and $A_\pm$ with coefficients depending on $q_n, \dot{q}_n$ and on the so-called spectral parameter λ:

$$\begin{aligned} L &= \sum b_n E_{n,n} + \lambda^{-1} \sum a_n E_{n,n+1} \\ &\quad + \lambda \sum E_{n+1,n}, \qquad (11) \\ A_+ &= \sum b_n E_{n,n} + \lambda \sum E_{n+1,n}, \\ A_- &= -\lambda^{-1} \sum a_n E_{n,n+1}\,, \qquad (12) \end{aligned}$$

where the Flaschka–Manakov variables a_n, b_n are expressed through q_n, $p_n = \dot{q}_n$ as

$$a_n = \mathrm{e}^{q_{n+1}-q_n}, \quad b_n = p_n. \qquad (13)$$

Here we represent difference operators as infinite matrices, with $E_{m,n}$ being the matrix with the only nonvanishing element equal to 1 in the position (m, n). In the case of the finite open-end lattice, the operators are $N \times N$ matrices; summation in the above formulas is performed for n from 1 to N, and it is assumed that $E_{N+1,N} = E_{N,N+1} = 0$. The N-periodic case differs in that it is assumed that $E_{N+1,N} = E_{1,N}$ and $E_{N,N+1} = E_{N,1}$. In the infinite and the finite open-end case, the spectral parameter λ can be removed by a conjugation with a diagonal matrix $\mathrm{diag}(\lambda^n)$; it is not so in the periodic case, which makes the latter analytically much more difficult and demanding.

An alternative to the Lax representation is the discrete zero-curvature representation

$$\dot{L}_n = M_n L_n - L_n M_{n-1}, \qquad (14)$$

first developed by Ablowitz and Ladik in 1975. It uses the same second-order operators (11), (12), but in terms of the spectral parameter-dependent 2×2 matrices

$$L_n = \begin{pmatrix} b_n + \lambda & a_{n-1} \\ -1 & 0 \end{pmatrix}, \quad M_n = \begin{pmatrix} -\lambda & -a_n \\ 1 & b_n \end{pmatrix}. \qquad (15)$$

Such a representation is somewhat more flexible from the technical viewpoint.

Inverse scattering

The matrix L is similar to a symmetric tri-diagonal matrix, which means that the operator L is gauge equivalent to a second-order and self-adjoint matrix. The direct and inverse spectral theory of such operators is well-developed (see Teschl, 2000), and parallel, to a large extent, to the direct and inverse spectral theory of second-order differential operators. In the rapidly decaying case ($q_n \to c$, $\dot{q}_n \to 0$ as $n \to \pm\infty$) the standard IST scheme is illustrated by the following diagram.

$$\begin{tikzcd} \boxed{q_n(0), \dot{q}_n(0)} \arrow[r, "\text{direct spectral problem}"] & \boxed{\mu_j, \gamma_j(0), r(\zeta, 0)} \arrow[d, "\text{linear evolution}"] \\ \boxed{q_n(t), \dot{q}_n(t)} & \boxed{\mu_j, \gamma_j(t), r(\zeta, t)} \arrow[l, "\text{inverse spectral problem}"'] \end{tikzcd}$$

The set of spectral data allowing for a solution of the inverse problem consists of the discrete eigenvalues μ_j, the normalizing coefficients γ_j of the corresponding discrete eigenfunctions, and the reflection coefficient $r(\zeta)$ of the continuous spectrum. Equation (10) yields that the evolution of the operator L, induced by the fact that $q_n(t)$ satisfy the Toda lattice equations (3), is isospectral. More precisely, the discrete eigenvalues are constants of the motion, while the evolution of other spectral data is governed by simple linear equations:

$$\begin{aligned} \mu_j &= \text{const}, \quad \gamma_j(t) = \gamma_j(0)\mathrm{e}^{(\mu_j - \mu_j^{-1})t}, \\ r(\zeta, t) &= r(\zeta, 0)\mathrm{e}^{(\zeta - \zeta^{-1})t}. \end{aligned}$$

Solution of the inverse spectral problem is given in terms of the Riemann–Hilbert problem or its variants, such as the Gel'fand–Levitan equation. Multisoliton solutions correspond to reflectionless potentials, for which $r(\zeta, t) \equiv 0$.

In the periodic case, the set of the spectral data is more complicated: it includes a hyperelliptic Riemann surface $\mathcal{R}$ of genus $N - 1$ determined by the eigenvalues of the periodic boundary value problem for the operator L; further, it includes $N - 1$ points P_k on $\mathcal{R}$, which correspond to the eigenvalues of L with vanishing boundary conditions. Due to (10), the Riemann surface $\mathcal{R}$ itself is an integral of motion, and the evolution of points P_k is such that the image of the divisor $P_1 + \ldots + P_{N-1}$ under the Abel map moves along a straight line in the Jacobi variety of $\mathcal{R}$. Solution of the inverse spectral problem is given in terms of multidimensional theta-functions. An important technical device in this type of problems constitute Baker–Akhiezer functions, which have essential singularities at P_k and are meromorphic elsewhere on $\mathcal{R}$; their analytic properties define them uniquely.

Hierarchies

Integrable lattices (and, more generally, integrable systems) do not appear separately, but are organized in hierarchies. To describe the Toda lattice hierarchy, note that the matrices $A_\pm$ in (12) may be seen as $A_\pm = \pm\pi_\pm(L)$, where $\pi_\pm$ stands for the nonnegative (resp. negative) part of the Laurent series with respect to the spectral parameter λ. This is a particular instance of a very general construction, known as the Adler–Kostant–Symes (AKS) method, dealing with a splitting of a Lie algebra g into a direct sum of its two subspaces that are also Lie subalgebras $g_\pm$; in this general setting, $\pi_\pm : g \mapsto g_\pm$ are the corresponding projections. The higher members of the Toda lattice hierarchy (enumerated by $m \in \mathbb{N}$) are characterized by Lax equations of form (10) with the same Lax matrix L as in (11) and $A_\pm = \pm\pi_\pm(L^m)$, see also Kupershmidt (1985).

"Explicit" Solution in Terms of a Factorization Problem

A further important ingredient of the AKS method is a formula for the solution of the initial value problem for such Lax equations. It is given in terms of the factorization problem in the Lie group G corresponding to the Lie algebra g in terms of its Lie subgroups $G_\pm$ corresponding to $g_\pm$: if the elements $U_\pm(t) \in G_\pm$ solve the factorization problem $\exp(tL^m(0)) = U_+U_-$, then $L(t) = U_+^{-1}(t)L(0)U_+(t) = U_-(t)L(0)U_-^{-1}(t)$ is the solution of the mth member of the hierarchy. The factorization problem in G is, in general, a sort of Riemann–Hilbert problem, intimately related to the one participating in the inverse scattering transformation.

Darboux–Bäcklund Transformations and Discretization

The above-mentioned factorization problem in G also leads to an effective construction of Darboux–Bäcklund transformations, which are a further indispensable attribute of integrable systems. For the Toda lattice, a Bäcklund transformation $(q_n, p_n) \mapsto (\widetilde{q}_n, \widetilde{p}_n)$ with a parameter h, which serves also as an integrable discretization with the time step h, is a result of the following factorization: $I + hL = U_+U_-$, with a subsequent flipping of factors: $I + h\widetilde{L} = U_-U_+$, where the factors are given by

$$U_+ = \sum \mathrm{e}^{\widetilde{q}_n - q_n} E_{n,n} + h\lambda \sum E_{n+1,n} \in G_+, \quad (16)$$

$$U_- = I + h\lambda^{-1} \sum \mathrm{e}^{q_{n+1} - \widetilde{q}_n} E_{n,n+1} \in G_-. \quad (17)$$

The formulas for the Bäcklund transformation seen from these factorizations, are

$$\begin{aligned} 1 + hp_n &= \mathrm{e}^{\widetilde{q}_n - q_n} + h^2 \mathrm{e}^{q_n - \widetilde{q}_{n-1}}, \\ 1 + h\widetilde{p}_n &= \mathrm{e}^{\widetilde{q}_n - q_n} + h^2 \mathrm{e}^{q_{n+1} - \widetilde{q}_n}. \end{aligned} \quad (18)$$

This is a canonical transformation, possessing a classical generating function. This result was first obtained by Wadati and Toda in 1975, see also Suris (2003).

Hamiltonian Structure

The last immanent feature of integrable lattices (and, more generally, integrable systems) is their Hamiltonian structure. Often, they are even bi- or multi-Hamiltonian, as it is the case for the Toda lattice. In the Flaschka–Manakov variables (13) the equations of motion of the Toda lattice (3) are rewritten as

$$\dot{a}_n = a_n(b_{n+1} - b_n), \quad \dot{b}_n = a_n - a_{n-1}. \quad (19)$$

The canonical Poisson bracket for the variables q_n, p_n reads in the variables a_n, b_n as follows:

$$\{b_n, a_n\}_1 = -a_n, \quad \{a_n, b_{n+1}\}_1 = -a_n \quad (20)$$

(all other brackets of the coordinate functions vanish), and system (19) is Hamiltonian with respect to this bracket, with the Hamilton function $H_2 = \frac{1}{2}\sum b_n^2 + \sum a_n$. However, one can define also a different Poisson bracket for the variables a_n, b_n:

$$\begin{aligned} \{b_n, a_n\}_2 &= -b_n a_n, \\ \{a_n, b_{n+1}\}_2 &= -a_n b_{n+1}, \\ \{b_n, b_{n+1}\}_2 &= -a_n, \\ \{a_n, a_{n+1}\}_2 &= -a_n a_{n+1}, \end{aligned} \quad (21)$$

with the following properties: it is compatible with the first one (i.e., their linear combinations are again Poisson brackets), and system (19) is Hamiltonian with respect to this bracket, with the Hamilton function $H_1 = \sum b_n$. So, the Toda lattice in the form (19) is a bi-Hamiltonian system. This was found by M. Adler in 1979. The bi-Hamiltonian property, introduced by Magri in 1978 on the example of KdV, has been established since then as an alternative (and highly effective and informative) definition of integrability. Actually, the Toda lattice (19) is even tri-Hamiltonian, since there exists one more local Poisson bracket for the variables a_n, b_n with similar properties, discovered in Kupershmidt (1985).

Higher Dimensions

Up to now, we have considered integrable lattices with one continuous and one discrete independent variables. This clearly allows for a further generalization. Integrable systems with two continuous and one discrete independent variables are well known and widely used as the models in the field theory. For instance, the Toda field theory deals with the system

$$(q_n)_{xy} = \mathrm{e}^{q_{n+1} - q_n} - \mathrm{e}^{q_n - q_{n-1}}, \quad (22)$$

introduced into soliton theory by Mikhailov (1979). Later it was realized that the equivalent equation $(\log v_n)_{xy} = v_{n+1} - 2v_n + v_{n-1}$, which is obtained by setting $v_n = \exp(q_{n+1} - q_n)$, appeared already in studies by Darboux in 1888, as the equation satisfied by the so-called Laplace invariants of the chain of Laplace transformations of a given second-order hyperbolic partial differential equation.

YURI B. SURIS

See also **Bäcklund transformations; Darboux transformation; Discrete nonlinear Schrödinger equations; Hamiltonian systems; Inverse scattering method or transform; Quantum inverse scattering method; Toda lattice**

Further Reading

Ablowitz, M. & Ladik, J. 1976. Nonlinear differential-difference equations and Fourier analysis. *Journal of Mathematical Physics*, 17: 1011–1018

Adler, V.E. 2000. On discretization of the Landau–Lifshits equation. *Teoreticheskaya i Matematicheskaya Fizika*, 124: 48–61

Faddeev, L.D. & Takhtajan, L.A. 1987. *Hamiltonian Methods in the Theory of Solitons*, Berlin and New York: Springer

Flaschka, H. 1974. On the Toda lattice II. Inverse scattering solution. *Progress in Theoretical Physics*, 51: 703–716

Kac, M. & van Moerbeke, P. 1975. On an explicitly soluble system of nonlinear differential equations related to certain Toda lattices. *Advances in Mathematics*, 16: 160–169

Krichever, I. 2000. Elliptic analog of the Toda lattice. *International Mathematics Research Notices*, 8: 383–412

Kupershmidt, B.A. 1985. Discrete Lax equations and differential-difference calculus. *Asterisque*, 123

Manakov, S.V. 1974. On the complete integrability and stochastization in discrete dynamical systems. *Soviet Physics, JETP*, 40: 269–274

Moser, J. 1975. Finitely many mass points on the line under the influence of an exponential potential—an integrable system. In *Dynamic Systems Theory and Applications*, edited by J. Moser (see *Math. Reviews* MR0393426). Berlin and New York: Springer, pp. 467–497

Ruijsenaars, S.N.M. 1990. Relativistic Toda systems. *Communications in Mathematical Physics*, 133: 217–247

Suris, Yu.B. 2003. *The Problem of Integrable Discretization: Hamiltonian Approach*, Basel: Birkhäuser

Teschl, G. 2000. *Jacobi Operators and Completely Integrable Nonlinear Lattices*. Providence, RI: American Mathematical Society

Toda, M. 1989. *Theory of Nonlinear Lattices*, Berlin and New York: Springer

Wadati, M. & Toda, M. 1975. Bäcklund transformation for the exponential lattice. *Journal of the Physical Society of Japan*, 39: 1196–1203

INTEGRABLE TOPS

See **Rotating rigid bodies**

INTEGRAL EQUATIONS

See **Equations, nonlinear**

INTEGRAL TRANSFORMS

Integral transforms have their genesis in the 19th century work of Joseph Fourier and Oliver Heaviside, subsequently set into a general framework during the 20th century. The fundamental idea is to represent a function $f(x)$ in terms of a *transform* $F(p)$, using an integral transform pair,

$$F(p) = \int K(p,x) f(x)\,\mathrm{d}x,$$
$$f(x) = \int L(x,p) F(p)\,\mathrm{d}p. \tag{1}$$

The functions $K(p,x)$ and $L(x,p)$ are *kernels*. If either of $f(x)$, $F(p)$ is discontinuous, further interpretation must be supplied at those points.

Oliver Heaviside invented his operational calculus to solve differential equations, such as those which arise in the theory of electrical transmission lines. A simple example is the diffusion problem (Whittaker, 1928),

$$\partial^2 u/\partial x^2 = k\,\partial u/\partial t, \quad x \geq 0, \quad t \geq 0,$$
$$u(0,t) = U_0. \tag{2}$$

Heaviside treated differentiation as a symbolic operation, $p = \partial/\partial t$, to write the operational solution

$$u(x,t) = \mathrm{e}^{-x\sqrt{kp}} \cdot U_0. \tag{3}$$

Using Euler's formula for differentiation,

$$\frac{\mathrm{d}^n(x^p)}{\mathrm{d}x^n} = \frac{p!}{(p-n)!}\,x^{p-n} \tag{4}$$

with n not restricted to the integers, Heaviside was able to extract all the information he required from such an approach. However, his methods met with general disapproval, leading to lifelong difficulties for him.

Heaviside's work was formalized, using the Laplace integral, to give the Laplace transform pair:

$$F(p) = \int_0^\infty \mathrm{e}^{-px} f(x)\,\mathrm{d}x,$$
$$f(x) = \frac{1}{2\pi\mathrm{i}} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \mathrm{e}^{+px} F(p)\,\mathrm{d}p. \tag{5}$$

In the normal case, the transform integral converges for all p in a complex half-plane $\Re(p) > \alpha$, in which case the constant c in the inversion integral must be chosen to satisfy the restriction $c > \alpha$. Subjecting Equation (2) to the Laplace transform leads to the same formula (3) for the *transform* $U(x,p)$. Everything may be extracted from this using complex variable techniques; the complex variable p is just a placeholder for Heaviside's operational symbol $p = \partial/\partial t$.

The Laplace transform has simple properties with respect to differentiation and convolutions (Schiff, 1999). For the former, if $F(p)$ is the transform of $f(t)$, then the transform of $f'(t)$ is $pF(p) - f(0)$. Again, if $h(t)$ is the convolution of $f(t)$ with $g(t)$, defined as

$$h(t) = \int_0^t f(s) g(t-s)\,\mathrm{d}s, \tag{6}$$

then the transform is given by simple multiplication:

$$H(p) = F(p)\,G(p). \tag{7}$$

An important area of application is systems of ordinary differential equations with constant coefficients, such as (Barnett & Cameron, 1992)

$$\boldsymbol{x}' = A\boldsymbol{x} + B\boldsymbol{u}, \qquad \boldsymbol{y} = C\boldsymbol{x}. \tag{8}$$

Here $\boldsymbol{x}$ is a set of n internal state variables, $\boldsymbol{u}$ a set of r inputs, $\boldsymbol{y}$ a set of s outputs; A, B, and C are matrices. Taking the Laplace transform and using elementary matrix algebra, the response of the system is given by

$$\mathbf{Y}(p) = G(p)\mathbf{U}(p),$$
$$G(p) = C(pI - A)^{-1}B. \tag{9}$$

The transform being a product, the solution is a matrix of convolutions. However, the essential information about the system may be extracted directly, without recourse to inversion.

The Fourier transform represents functions as linear combinations of periodic functions, an idea pioneered by Fourier. A one-dimensional form is

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t,$$

$$f(t) = \frac{1}{2\pi\mathrm{i}}\int_{-\infty}^{\infty} F(k)\,\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}\omega. \qquad (10)$$

The most common uses are (i) with time as the original variable—transformation from the time domain to the frequency domain and (ii) with spatial position as the original variable—transformation from configuration space to momentum space. The Fourier transform also has simple properties with respect to differentiation and convolution (the integral in (6) now extends from $-\infty$ to $+\infty$), making it useful for investigation of partial differential equations, as well as many more complex systems.

Other important transforms are the Hankel transform, which uses Bessel functions for the kernels in pair (1), alternative forms of the Fourier transform using sine or cosine functions, and many other lesser known pairs (Davies, 2002). In fact, every Sturm–Liouville problem generates integral transforms (Antimirov et al., 1993). In addition, there are transforms that involve complex integration in an essential way, such as the Hilbert transform and Cauchy integrals, which use principal values.

The Mellin transform pair

$$F(p) = \int_{0}^{\infty} x^{p-1} f(x)\,\mathrm{d}x,$$

$$f(x) = \frac{1}{2\pi\mathrm{i}}\int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} x^{-p}F(p)\,\mathrm{d}p \qquad (11)$$

has great utility, particularly in asymptotics (Ninham et al., 1992; Davies, 2002). For example, the Mellin transform representation of the exponential integral

$$\mathrm{Ei}(x) = \int_{x}^{\infty} \frac{\mathrm{e}^{-u}}{u}\,\mathrm{d}u, \qquad (12)$$

is

$$\mathrm{Ei}(x) = \frac{1}{2\pi\mathrm{i}}\int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \frac{(p-1)!x^{-p}}{p}\,\mathrm{d}p,\ c>0, (13)$$

$$= -\ln x - \gamma - \sum_{k=1}^{\infty}\frac{(-1)^k x^k}{k!\,k}, \qquad (14)$$

where the infinite sum is obtained by the standard theory of residues.

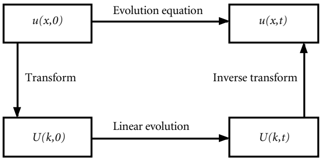

Figure 1. Solution of evolution equations using transforms.

The Fourier transform is particularly well suited to linear evolution problems. Typically, there is an initial configuration, $u(x, 0)$, which is represented as a linear superposition of plane waves,

$$u(x, 0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} U(k, 0)\,\mathrm{e}^{-\mathrm{i}kt}\mathrm{d}k. \qquad (15)$$

The individual waves evolve according to a simple dispersion relation, $\omega = \omega(k)$, which means that the functions $\exp[-\mathrm{i}kx + \mathrm{i}\omega(k)t]$ satisfy the partial differential equation. As a consequence, the solution at later time is

$$u(x, t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} U(k, 0)\,\mathrm{e}^{-\mathrm{i}kx+\mathrm{i}\omega(k)t}\mathrm{d}k. \qquad (16)$$

The process can be visualized as in Figure 1. The top arrow indicates the required evolution of the initial state $u(x, 0)$ to the later state $u(x, t)$. The circuitous route has three steps: (i) map the initial state $u(x, 0)$ to some wave data $U(k, 0)$, (ii) evolve the wave data under the evolution equation to the later wave data $U(k, 0)\mathrm{e}^{i\omega(k)t}$, and (iii) use an inverse transform to reconstruct the new state from the new wave data.

Beginning in the 1970s, it was realized that a large number of nonlinear evolution equations may be solved using an analogous process that has come to be known as the inverse scattering method (Dodd et al., 1982). It is more complicated than the linear methods, but reduces to them in the limit of small nonlinearity. Briefly, the nonlinear equation is associated with a linear one, and there is a mechanism for mapping the data of the former onto the data of the latter. The method again proceeds according to the scheme of Figure 1. However the inverse transform, which recovers the state from the "scattering data," is not so simple, and the method derives its name from it. The first explicit use of this scheme was for the solution of the Korteweg–de Vries (KdV) equation, which is associated with a linear Schrödinger equation in which the function $u(x, t)$ plays the role of the potential. The evolution of the spectrum of the Schrödinger operator, in response to the nonlinear evolution of KdV equation, is linear; furthermore, in the limit of weak nonlinearity, the transform and its inverse reduce to the Fourier transform.

BRIAN DAVIES

See also **Ablowitz–Kaup–Newell–Segur system; Burgers equation; Gel'fand–Levitan theory; Hirota's method; Inverse scattering method or transform; Korteweg–de Vries equation; Nonlinear Schrödinger equations; Sine-Gordon equation; Solitons**

Further Reading

Antimirov, M.Ya., Kolyshkin, A.A. & Vaillancourt, R. 1993. *Applied Integral Transforms*, New York: Springer

Barnett, S. & Cameron, R.G. 1992. *Introduction to Mathematical Control Theory*, 2nd edition, Oxford: Clarendon Press

Davies, B. 2002. *Integral Transforms and Their Applications*, 3rd edition, New York: Springer

Dodd, R.K., Eilberg, J.C., Gibbon, J.D. & Morris, H.C. 1982. *Solitons and Nonlinear Wave Equations*, London and New York: Academic Press

Duffy, D.G. 1994. *Transform Methods for Solving Partial Differential Equations*, Boca Raton, FL: CRC Press

Moore, D.H. 1971. *Heaviside Operational Calculus*, New York: Elsevier

Ninham, B.W., Hughes, B.D., Frankel, N.E. & Glasser, M.L. 1992. Möbius, Mellin, and mathematical physics. *Physica* A, 186: 441–481

Schiff, J.L. 1999. *The Laplace Transform: Theory and Applications*, New York: Springer

Whittaker, E.T. 1928. Oliver Heaviside. *Bulletin of the Calcutta Mathematical Society*, 20: 199–220 (The article is reprinted in Moore, 1971.)

INTEGRALS OF MOTION

See **Constants of motion and conservation laws**

INTEGRATE AND FIRE NEURON

Classical electrophysiology using extracellular recording methods established the all-or-none nature of the nerve impulse—a stimulus has to be above a threshold θ for an action potential to be generated, and for rectangular stimuli, θ depends on stimulus duration T. A plot of θ–T, a strength-duration curve is approximated by

$$\theta = \theta_{\mathrm{rh}}/[1 - \exp(-T/\tau], \tag{1}$$

where θ_{rh} is a constant called the rheobase and τ is a time constant. Lapicque (1907) produced the strength-duration curve of Equation (1) by a model in which subthreshold inputs $x(t)$ are integrated and decay with an exponential time course

$$V(t) = \int_0^t \exp(-(t-u)/\tau)\,\mathrm{d}u, \tag{2}$$

for $0 < V(t) < \theta$, and when $V(t)$ reaches threshold, an idealized action potential (a Dirac δ-function) is triggered and V is instantaneously reset to its resting value (say, $V = 0$). The subthreshold changes in $V(t)$ were represented as the response to an applied current of a resistance and capacitance in parallel. This separates the slow time scales of the subthreshold integrative processes from the fast time scales of the action potential mechanisms. The values of the resistance and capacitance reflect not membrane parameters, but cell parameters, and so depend via surface area on cell size. Guided by the strength-duration curve, nerve or muscle, or different diameter nerve fibers can be excited.

Lapicque's integrate-and-fire model is also known as the leaky integrator and, in the case $\tau \to \infty$, the perfect integrator. Unlike the FitzHugh–Nagumo, Hodgkin–Huxley, and Markin–Chizmadzhev models, it considers only subthreshold, integrative processes, not the processes generating the action potential, and so is more suitable for modeling neuronal information processing than action potential transmission. Originally formulated as a phenomenological model for the conditions for the initiation of an action potential, it is still widely used as a model today, but as a model for the generation of neuronal spike trains and mass activity in neuronal networks. It has been modified to include both absolute and relative refractory periods, obtained as a limiting case for the Hodgkin–Huxley equations, and extended into multiple time constant and spatially extended models (Tuckwell, 1988).

If the input $u(t)$ is a constant c, the interspike intervals t_i are a solution of

$$\theta = \int_0^{t_i} \exp(-(t-u)/\tau)\,\mathrm{d}u, \tag{3}$$

and so the rate of firing is $1/t_i = 1/[-\tau \ln(1 - 1/c\tau)]$, showing an increased sensitivity to small amplitude inputs. Such a logarithmic relation between discharge rate and the magnitude of a constant input is observed in some primary sensory neurons, and is used as an explanation for the Weber–Fechner law of sensory psychophysics.

Gerstein & Mandelbrot (1964) introduced a random walk model for neuronal spike trains, in which the input to a perfect integrator model was a random sequence of small positive and negative steps. This was motivated by the observation that the interspike interval and scaled interval (sums of adjacent intervals) of some auditory, neurons had the same unimodal, asymmetric shape; that is, assuming the spike train was a realization of a renewal point process, the interval density function was invariant under self-convolution or has a stable, or infinitely divisible, distribution. The subthreshold response of the integrate-and-fire model to small and brief positive or negative current pulses could represent the postsynaptic response of a neuron to synaptic inputs. Superposition of a (large) number of independent point processes generates a Poisson process: with appropriate constraints on how the amplitude of the steps decreases while the rate increases, the random walk becomes a white noise, leading to diffusion models for stochastic activity of neurons. The diffusion process models for $V(t)$ are approximations of the discontinuous, random walk models and are constructed to have the same first and second infinitesimal moments. For the perfect integrator, the subthreshold $V(t)$ is a Wiener process; for the leaky integrator, an Ornstein–Uhlenbeck process. The interspike-interval probability density functions for these diffusion models are first passage time densities to an absorbing barrier, the threshold. These densities are obtained as solutions for forward or backward Chapman–Kolmogorov equations (see Holden, 1976; Tuckwell, 1988). These equations have been solved for different diffusion models, mostly numerically, but for some special cases solutions are available in closed form. The diffusion models have been extended to include time-varying threshold, temporal modulation of the rates of the excitatory and inhibitory inputs, and spatial separation filtering of the inputs. In the derivation of these diffusion models the synaptic inputs are assumed to be uncorrelated, producing a white noise input: correlation in the input, producing colored inputs, is more likely for central neurons and has profound effects on the output variability of the integrate-and-fire models (Feng, 2001).

The perfect and leaky integrators respond to a cosinusoidal input by modulated spike trains with a periodically modulated spike density. At large modulation depths both models show rectification, and the leaky integrator also shows phase-locked responses. Both these nonlinear distortions are reduced by adding noise to the input (Stein et al., 1972), and so additive noise can increase the ability of a neuron to transmit signals about narrow-band, high frequencies. This provides a functional explanation for the association between the variability in sensory mechanoreceptors and the frequency of their adequate stimulus: the discharge of auditory neurons is almost random, and velocity sensitive primary muscle spindle afferents have a higher variability than length sensitive secondary muscle spindle afferents, leading to ideas of stochastic resonance. The phase locked and aperiodic responses of the leaky integrator to a periodic input has been characterized in terms of the Arnol'd tongue structure for the dominant modes, where there are low order ratios between the spiking and forcing rhythms (Coombes & Bressloff, 1999).

The success of the integrate-and-fire models in accounting for general properties of spike trains has allowed it, with simple modifications, to be applied to model spike trains from specific neurons, see for example, Herrmann & Gerstner (2002). The simplicity of the model also allows it to be used as an element in constructing population dynamic equations for large systems of coupled neurons, as in Sirovich et al. (2000).

Arun V. Holden

See also **Excitability; FitzHugh–Nagumo equation; Hodgkin–Huxley equations; Markin–Chizmadzhev model; Nerve impulses; Neurons**

Further Reading

Coombes, S. & Bressloff, P.C. 1999. Mode locking and Arnol'd tongues in integrate-and-fire neural oscillators. *Physical Review* E, 60: 2086–2096

Feng. J. 2001. Is the integrate-and-fire model good enough? A review. *Neural Networks*, 14,955–14,975

Gerstein, G.L. & Mandelbrot. 1964. Random walk models for the spike activity of a single neuron. *Biophysical Journal*, 4: 41–68

Herrmann, A. & Gerstner, W. 2002. Noise and the PSTH response to current transients: II. Integrate-and-fire model with slow recovery and application to motoneuron data. *Journal of Computational Neuroscience*, 12: 83–95

Holden, A.V. 1976. *Models of the Stochastic Activity of Neurones*, Berlin: Springer

Lapicque, L. 1907. Recherches quantitatives sur l'excitation electrique des nerfs traite comme une polarization. *Journal de Physiologie et Pathologie Generale*, 9: 620–635

Sirovich, L., Omurtag, A. & Knight, B.W. 2000. Dynamics of neuronal populations: the equilibrium solution. *SIAM Journal on Applied Mathematics*, 6: 2009–2028

Stein, R.B., French, A.S. & Holden, A.V. 1972 The frequency response, coherence, and information capacity of two neuronal models. *Biophysical Journal*, 12: 295–322

Tuckwell, H.C. 1988. *Introduction to Theoretical Neurobiology.* vol. 1. Cambridge and New York: Cambridge University Press

INTEGRO-DIFFERENTIAL EQUATIONS

See **Equations, nonlinear**

INTERMITTENCY

Consider a dynamical system with x as the observed variable. As a function of time t, if $x(t)$ exhibits segments of relative constant values (laminar phase) interspersed by erratic bursts, we say the system dynamics is intermittent (see Figure 1). Pomeau and Manneville (1980) were the first to identify the three bifurcations in low-dimensional maps that are associated with commonly observed intermittencies. An additional mechanism for generating intermittency is related to the transverse instability of chaotic attractors confined to a manifold whose dimension is smaller than that of the full phase space, which is called on-off intermittency (Pikovsky, 1984; Fujisaka & Yamada, 1985; Platt et al., 1993).

Pomeau–Manneville type I intermittency stems from a saddle-node bifurcation. Near the bifurcating point ($x=0$) in one-dimensional maps, the dynamics are described by $x_{n+1}=x_n+\varepsilon+x_n^2$. For positive but small ε, a narrow channel near $x=0$ is formed between the nonlinear map function and the diagonal line $x_{n+1}=x_n$. When a trajectory enters the channel, it moves through it very slowly, giving rise to the appearance of a quiet laminar phase. After exiting the channel, the trajectory can display wild swings governed by nonlinear mechanisms, forming the bursting phase of the dynamics. As time progresses, ergodicity will bring the trajectory back to the channel again, and the process starts anew. The re-injected trajectory can land at different points inside the channel. As a result, the laminar phase has variable length, the mean of which is a statistical quantity of interest. It has been shown that the mean length of the laminar phase scales with the parameter ε as $\sim \varepsilon^{-1/2}$. As ε decreases through zero, a saddle-node bifurcation occurs, creating a stable fixed-point attractor at $-\sqrt{-\varepsilon}$ and an unstable fixed point at $\sqrt{-\varepsilon}$. Intermittency terminates at the bifurcation point $\varepsilon = 0$.

Pomeau–Manneville type II intermittency is related to a subcritical Hopf bifurcation; thus a two-dimensional map is needed for this phenomenon. In polar coordinates, the dynamics near the bifurcation point (assumed to be the origin) is $r_{n+1} = (1+\varepsilon)r_n + r_n^3$, $\theta_{n+1} = \theta_n + \alpha$, where α is a constant rotation. For $\varepsilon > 0$, the origin is an unstable spiral. Viewing the r equation as a one-dimensional map, one again finds a channel for small r. When the trajectory is randomly injected into the channel, a quiet laminar period ensues, which is followed by a burst of chaotic behavior. The mean length of the laminar phase is shown to scale with ε as $\sim \varepsilon^{-1}$. At $\varepsilon = 0$ a subcritical Hopf bifurcation occurs. For $\varepsilon < 0$, intermittency is replaced by a stable fixed-point attractor at the origin that is accompanied by an unstable invariance circle centered at the origin.

Pomeau–Manneville type III intermittency is created near an inverse period-doubling bifurcation. Near the bifurcation point ($x = 0$), the equation of motion is $x_{n+1} = -(1+\varepsilon)x_n + \alpha x_n^2 + \gamma x_n^3$. For this map, the channel dynamics is best seen by examining the second iterate of the above map that is easily obtained (after discarding higher-order terms) to be $x_{n+1} = (1+2\varepsilon)x_n + \beta x_n^3$, where $\beta = -2(\alpha^2+\gamma)$. As ε decreases from positive values to negative values, intermittency is replaced by a stable fixed-point attractor and an unstable period-two orbit. The channel near $x = 0$ here is analogous to the channel near $r = 0$ in the type II intermittency. It is, thus, not surprising that the mean length of the laminar phase scales with the parameter ε as $\sim \varepsilon^{-1}$.

Consider an m-dimensional dynamical system that has an n-dimensional invariant manifold where $n < m$. (By invariant we mean that a trajectory initialized in the manifold stays there for all time.) Suppose that the dynamics on the manifold has a chaotic attractor that is also an attractor for the full phase space. Suppose also that, as a parameter varies across a threshold, a blowout bifurcation takes place (Ott & Sommerer, 1994) and the attractor becomes transversely unstable. If one monitors the distance between the trajectory and the bifurcating invariant manifold, this distance typically exhibits on-off intermittency (see Figure 1) immediately after the blowout bifurcation. During the laminar phase, in other words, the trajectory spends a great deal of time close to the invariant manifold, producing a nearly zero value for the distance variable. The trajectory then departs the manifold rapidly and displays erratic bursting behavior for the distance variable until the dynamics brings it back close to the invariant manifold again, and the process repeats. During the laminar phase, the distance variable can be studied via a random walk model, and the interval between successive bursts follows a distribution akin to that of the first return time (Ding & Yang, 1995).

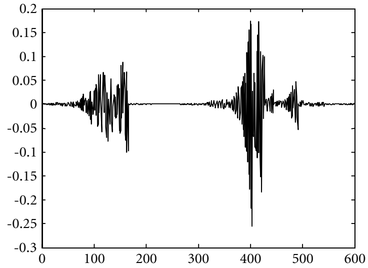

Figure 1. An intermittent function.

All four types of intermittencies have been studied in experimental systems. For example, Jeffries and Perez found Pomeau–Manneville type I intermittency in a nonlinear circuit oscillator (Jeffries & Perez, 1982). Pomeau–Manneville type II intermittency is characterized in a chemical reaction system by Herzel et al. (1991). Dubois et al. (1983) observed Pomeau–Manneville type III intermittency in Rayleigh–Bénard convection. Experimental studies of on-off intermittency are relatively recent. Hammer et al. (1994) are the first to report evidence of on-off intermittency in a nonlinear electronic circuit. On-off intermittency in a laser system was observed by Pisarchik and Pinto-Robledo in 2002.

Mingzhou Ding

See also **Bifurcations; Hénon–Heiles system; Hopf bifurcations; Phase space; Routes to chaos**

Further Reading

Berge, P., Pomeau, Y. & Vidal, C. 1986. *Order Within Chaos, Toward a Deterministic Approach to Turbulence*, New York: Wiley

Ding, M. & Yang, W. 1995. Distribution of the first return time in fractional Brownian motion and its application to the study of on–off intermittency. *Physical Review* E, 52: 207–213

Dubois, M., Rubio, M.A. & Berge, P. 1983. Experimental evidence of intermittencies associated with a subharmonic bifurcation. *Physical Review Letters*, 51: 1446–1449

Fujisaka, H. & Yamada, T. 1985. A new intermittency in coupled dynamical systems. *Progress in Theoretical Physics*, 74: 918–921

Hammer, P.W., Platt, N., Hammel, S.M., Heagy, J.F. & Lee, B.D. 1994. Experimental observation of on-off intermittency. *Physical Review Letters*, 73: 1095–1098

Herzel, H., Plath, P. & Svensson, P. 1991. Experimental evidence of homoclinic chaos and type-II intermittency during the oxidation of methanol. *Physica* D, 48: 340–352

Jeffries, C. & Perez, J. 1982. Observation of a Pomeau-Manneville intermittent route to chaos in a nonlinear oscillator. *Physical Review* A, 26: 2117–2122

Ott, E. 2002. *Chaos in Dynamical Systems*, 2nd edition, Cambridge and New York: Cambridge University Press

Ott, E. & Sommerer, J. 1994. Blowout bifurcation: the occurrence of riddled basins and on–off intermittency. *Physics Letters* A, 188: 39–47

Pikovsky, A. 1984. A new type of intermittent transition to chaos. *Journal of Physics* A, 16: L109–L112

Pisarchik, A.N. & Pinto-Robledo, V.J. 2002. Experimental observation of two-state on–off intermittency. *Physical Review* E, 027203

Platt, N., Spiegel, E.A. & Tresser, C. 1993. On–off intermittency: a mechanism for bursting. *Physical Review Letters*, 70: 279–282

Pomeau, Y. & Manneville, P. 1980. Intermittent transition to turbulence in dissipative dynamical systems. *Communications in Mathematical Physics*, 74: 189–197

INVARIANT MANIFOLDS AND SETS

An invariant set S of a dynamical system has the property that every trajectory (or orbit) that starts in S remains in S. Invariant sets are the basic building blocks of dynamical systems and almost every set that is dynamically interesting has some invariance property. Every trajectory is an invariant set, and so equilibria and periodic orbits are simple examples of such objects. Attractors are also examples of invariant sets.

An invariant set that is also a manifold (i.e., it looks like R^n locally) is called an invariant manifold. Invariant manifolds provide a natural description of the dynamics close to an equilibrium or periodic orbit (using linearization) and can also make it possible to work in lower dimensions than the phase space of the problem, since a smooth dynamical system restricted to an invariant manifold is itself a dynamical system. This reduction in dimension is at the heart of center manifold techniques and inertial manifold methods.

Invariance is a topological property and does not depend upon stability. This is often useful, and the sudden appearance or disappearance of attractors, as parameters are varied, can sometimes be understood as a change in stability of an invariant set that exists throughout the parameter region of interest. Indeed, many invariant sets, and particularly invariant manifolds, have persistence properties under perturbations of the dynamical system that make them useful when working with families of systems.

The basic definitions will be given for differential equations $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ with $\boldsymbol{x} \in R^n$ and solutions $\boldsymbol{x}(t)$, and maps $\boldsymbol{v}_{n+1} = \boldsymbol{g}(\boldsymbol{v}_n)$, $\boldsymbol{v} \in R^n$. In both cases, the complications that arise if solutions do not exist for all time will be ignored.

For differential equations, a set S is a *forward invariant* set (resp. *backward invariant* set) if $\boldsymbol{x}(0) \in S$ implies that $\boldsymbol{x}(t) \in S$ for all $t \geq 0$ (resp. all $t \leq 0$). S is *invariant* if it is both forward invariant and backward invariant.

For maps, a set S is a *forward invariant* set (resp. *backward invariant* set) if $\boldsymbol{v} \in S$ implies that $\boldsymbol{g}(\boldsymbol{v}) \in S$ (resp. $\boldsymbol{g}^{-1}(\boldsymbol{v})$). S is *invariant* if it is both forward invariant and backward invariant.

In the remainder of this entry, some different sorts of invariant sets that arise in dynamics will be considered, illustrated by the Lorenz equations

$$\dot{x} = \sigma(y - x), \quad \dot{y} = -y + rx - xz,$$
$$\dot{z} = -bz + xy, \tag{1}$$

where σ, r, and b are positive constants.

Trapping Regions

A set R is a trapping region if it is forward invariant and if all solutions are eventually contained in R. Trapping regions provide the first approximation to the location of any attractors in the system.

For the Lorenz equations, (1), with $\sigma = 10$ and $b = \frac{8}{3}$ Sparrow (1982) shows that the region

$$R = \{(x, y, x) \mid rx^2 + 10y^2 + 10(z - 2r)^2 \leq 5r^2\} \tag{2}$$

is a trapping region. This can be proved using the Lyapunov function $V(x, y, z) = rx^2 + 10y^2 + 10(z - 2r)^2$.

Stable and Unstable Manifolds of Stationary Points

A stationary point of the differential equation $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ is a solution that is constant for all time, so it must satisfy $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$. The three equations obtained by setting $\dot{x} = \dot{y} = \dot{z} = 0$ in (1) can be solved to show that the origin is always a stationary point of the Lorenz equations, and provided $r > 1$, there are two other stationary points: $(\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1)$. Near a stationary point, the most important terms of the differential equation are the linear terms, so it is natural to ask whether the solutions of the linearized equation at the origin

$$\dot{x} = \sigma(y - x), \quad \dot{y} = -y + rx, \quad \dot{z} = -bz \tag{3}$$

obtained by ignoring the nonlinear terms in (1) reflect properties of the full equations close to the stationary point at the origin. These results are usually stated in terms of the eigenvalues of the matrix that defines the linear equation (3):

$$M = \begin{pmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{pmatrix}. \tag{4}$$

If M has no eigenvalues with zero real parts, then the manifold (vector space) spanned by the eigenvectors with eigenvalues having negative real parts contains all the stable directions and, hence, is called the stable manifold of the linear system, (3), while the manifold spanned by the eigenvectors with eigenvalues having positive real parts contains the unstable directions. The former manifold is called the stable manifold, E^{s}, of the origin, and the latter is the unstable manifold, E^{u}. If $0 < r < 1$, then all the eigenvalues of M are negative, so the stable manifold of the linear system is the whole space, and the unstable manifold is the empty set. If $r > 1$, then M has two negative eigenvalues and one positive eigenvalue, so the unstable manifold of the linear system is a line, and the stable manifold is a plane.

This provides definitions of stable and unstable manifolds for linear systems such as (3) with no eigenvalues with zero real parts. To extend the definition to nonlinear systems, we begin by noting that in the definition used for linear systems, the stable manifold of the origin contains the solutions that tend to the origin in forward time, and the unstable manifold contains those that tend to the origin in backwards time. These properties can be used to define the stable and unstable manifolds of a stationary point $\boldsymbol{x}^*$ for a nonlinear system: the stable manifold of $\boldsymbol{x}^*$, $W^{\mathrm{s}}(\boldsymbol{x}^*)$ is the set of initial conditions for which the solution $\boldsymbol{x}(t)$ tends to $\boldsymbol{x}^*$ as $t \to \infty$, while the unstable manifold, $W^{\mathrm{u}}(\boldsymbol{x}^*)$, is the set of initial conditions for which the solution $\boldsymbol{x}(t)$ tends to $\boldsymbol{x}^*$ as $t \to -\infty$. The stable manifold theorem states that provided the matrix that defines the linearized equation near a stationary point has no eigenvalues with zero real parts, then $W^{\mathrm{s}}(\boldsymbol{x}^*)$ and $W^{\mathrm{u}}(\boldsymbol{x}^*)$ exist, are invariant sets, and are of the same dimension as the corresponding linear stable and unstable manifolds. Moreover, they are each tangential to the corresponding linear manifold at the stationary point itself. See Guckenheimer & Holmes (1983) for more details.

Calculating Stable and Unstable Manifolds

The stable manifold theorem suggests how approximations to the stable and unstable manifolds of a stationary point can be calculated close to a stationary point. First, it is tangential to the corresponding linear manifold, and second, it is invariant. This is enough to obtain a power series solution, and this will be illustrated using the unstable manifold of the origin of the Lorenz equations with $\sigma = 1$ and $r = 4$.

Step 1 (Linear approximation): The eigenvalues of M in equation (4) are 1, -3, and $-b$ if $\sigma = 1$ and $r = 4$. The unstable manifold is one-dimensional, and the unstable manifold of the linear approximation is the eigenvector corresponding to the positive eigenvalue, that is, $(1, 2, 0)$ or $y = 2x$, $z = 0$.

Step 2 (Canonical form): It is good practice to bring the linear part into canonical form at this stage, see Guckenheimer & Holmes (1983). For brevity this step will be omitted.

Step 3 (Power series): Since the unstable manifold is tangential to the linear approximation of Step 1 at the stationary point, it can be represented *locally* as a function of one of the variables. From the form of the linear approximation, either x or y can be used here. If the defining equations are smooth, then the stable and unstable manifolds are smooth, so choosing x as the independent variable we write

$$y = 2x + \alpha x^2 + O(x^3), \quad z = \beta x^2 + O(x^3), \qquad (5)$$

where the linear terms are obtained from the linear approximation, α and β are constants to be determined, and $O(x^3)$ denotes the terms of order x^3 and higher that are small (locally) compared with the linear and quadratic terms. To identify the constants α and β, two different ways of calculating $\dot{y}$ and $\dot{z}$ are used and then compared in the next three steps.

Step 4 (Power series evolution): Differentiating (5) gives $\dot{y} \sim 2\dot{x} + 2\alpha x\dot{x}$ and $\dot{z} \sim 2\beta x\dot{x}$ (ignoring the higher order terms). Now, on the unstable manifold given by (5) $\dot{x} = -x + y \sim x + \alpha x^2$, so

$$\begin{aligned} \dot{y} &\sim 2(x + \alpha a x^2) + 2\alpha x^2 \sim 2x + 4\alpha x^2, \\ \dot{z} &\sim 2\beta x^2, \end{aligned} \qquad (6)$$

where higher-order terms have been ignored again.

Step 5 (Differential equation evolution): From (1) and (5) with $r = 4$ and $\sigma = 1$,

$$\begin{aligned} \dot{y} &= 4x - y - xz \sim 2x - \alpha x^2, \\ \dot{z} &= -bz + xy \sim -b\beta x^2 + 2x^2. \end{aligned} \qquad (7)$$

Step 6 (Equate coefficients): To determine the two constants α and β, we now simply equate coefficients of powers of x in (6) and (7). From the $\dot{y}$ equations $4\alpha = -\alpha$, so $\alpha = 0$, and from the $\dot{z}$ equations $2\beta = -b\beta + 2$, so $\beta = 2/(2 + b)$. This gives the second-order approximation to the unstable manifold:

$$y = 2x + O(x^3), \quad z = \tfrac{2}{2+b}x^2 + O(x^3), \qquad (8)$$

which is valid close to the stationary point at the origin. Higher-order terms can be calculated by including more terms in the power series expansions.

Center Manifolds

If the matrix that defines the linearized differential equation at a stationary point has eigenvalues with zero real parts, then three invariant manifolds can be defined: a strong stable manifold corresponding to eigenvalues with negative real parts, a strong unstable manifold corresponding to eigenvalues with positive real parts, and a center manifold that is tangential at the stationary point to the space spanned by the eigenvectors corresponding to eigenvalues with zero real parts. The motion on the strong stable and unstable manifolds is defined by the dominant linear terms, that is, towards and away from the stationary point, respectively. The motion on the center manifold depends on the nonlinear terms of the differential equation and may be stable or unstable or neutral. Center manifolds can be approximated locally using the same ideas as the approximation of the unstable manifold above, and this leads to the technique of center manifold reduction, which is central to the development of bifurcation theory. The origin of the Lorenz equations has a one-dimensional center manifold if $r = 1$ (M in (4) has an eigenvalue of zero), which signals the bifurcation creating the nontrivial pair of stationary points if $r > 1$.

Unstable Chaotic Sets

The chaotic sets created by horseshoes are not stable, but they can be important from two points of view. First, if an unstable chaotic set exists in a system then it can manifest itself in chaotic transients before a stable attractor is reached. Second, if there are parameters in the system then the chaotic set may gain stability and so the sudden appearance of strange attractors can be explained by understanding the change of stability of a chaotic invariant set. Grebogi et al. (1982) describe some of these mechanisms as *crises*. In the Lorenz equations with $b = \frac{8}{3}$ and $\sigma = 10$, a strange invariant set (an unstable chaotic set with a fractal structure) is created by a homoclinic bifurcation as r is increased through $r \approx 13.93$, but it does not become stable until $r \approx 24.06$; see Sparrow (1982); for details.

There are many other types of invariant sets and properties of these sets that could have been described here—the selection above is only the tip of an invariant iceberg.

PAUL GLENDINNING

See also **Attractors; Bifurcations; Center manifold reduction; Chaotic dynamics; Horseshoes and hyperbolicity in dynamical systems; Inertial manifolds; Linearization; Lorenz equations; One-dimensional maps; Routes to chaos**

Further Reading

Grebogi, C., Ott, E. & Yorke, J.A. 1982. Chaotic attractors in crisis. *Physics Review Letters*, 48: 1507–1510

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamcal Systems, and Bifurcations of Vector Fields*, New York: Springer

Sparrow, C. 1982. *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, New York: Springer

INVARIANT MEASURE

See **Measures**

INVERSE PROBLEMS

Scientists try to interpret the world in terms of mathematical models that depend on parameters by relating their models to the results of measurements. Thus, a model is given when a set $\mathcal{C}^0$ of parameters is mapped into a set $\mathcal{E}^0$ of experimental results. Explicit calculations of this mapping $\mathcal{M}$ are called "solving the direct problem." Providing an exact or approximate way for going back from experimental results to parameters by $\mathcal{M}$ is called "solving the inverse problem" (Sabatier, 2000).

Because the exact limits of $\mathcal{C}^0$ and $\mathcal{E}^0$ are difficult to define, they are embedded in convenient mathematical spaces, $\mathcal{C}$ and $\mathcal{E}$, usually linear normed spaces, where it is understood that $\mathcal{C}$ is the domain of $\mathcal{M}$ and $\mathcal{M}(\mathcal{C}) \subseteq \mathcal{E}$. Because of measurement errors, there may be measurement results belonging to $\mathcal{E}$ but not to $\mathcal{M}(\mathcal{C})$, and if it is so, the inverse problem is said to be overdetermined. To appraise uncertainties on measurements, the norm of $\mathcal{E}$ must be reasonably defined, corresponding to our physical knowledge of errors.

The inverse problem is said to be "well-posed" if the norm of $\mathcal{E}$ is reasonably defined, $\mathcal{C} = \mathcal{C}^0$, $\mathcal{M}(\mathcal{C}) = \mathcal{E}$, and $\mathcal{M}$ has a two-sided inverse $\mathcal{M}^{-1}$, which is continuous for the canonical distance in $\mathcal{E}$. Then given any measurement result $e \in \mathcal{E}$, $\mathcal{M}^{-1}e$ yields parameters $x \in \mathcal{C}$ such that

$$\mathcal{M}x = e. \tag{1}$$

If a small error in the experimental data e implies that the error in the parameters x is small, the solution x is stable. Unfortunately, almost all inverse problems of physics or engineering are "ill-posed" (i.e., not well-posed).

For an ill-posed problem, a regularized solution is a continuous mapping $\overline{\mathcal{M}}$ from $\mathcal{E}$ to $\mathcal{C}^0$ such that $\mathcal{M}(\overline{\mathcal{M}}e)$ is close to $e \in \mathcal{E}$. Because experimental observations define $\mathcal{C}^0$ only approximately, one often imposes additional requirements to ease the regularizing processes, for example, that $\mathcal{C}^0$ is compact (Engl et al., 2000).

Classical examples of regularized solutions for Equation (1) are obtained by minimizing over x in one of two ways: either successively as a Euclidean distance in $\mathcal{E}$ that models the misfit and a Euclidean distance in $\mathcal{C}$ that models the non-reliability or as a combination of their squares into a cost function. Typical is the Tikhonov cost

$$C(x, x_0, \lambda) = \|x - x_0\|_{\mathcal{C}}^2 + \lambda \|\mathcal{M}x - e\|_{\mathcal{E}}^2, \lambda > 0, \tag{2}$$

where two parameters (the “first guess” x_0 and the balance parameter λ) are to be chosen. (λ may be related to an evaluation of the noise/signal ratio or to the stability, whereas “physical” information on x_0 may be suggested by previous results.)

If $\mathcal{M}$ is differentiable, the cost function (2) is also differentiable, and elementary variational calculus applies. Information from statistical analyses is more conveniently represented by quadratic cost functions similar to Equation (2), but involving noncanonical distances in $\mathcal{C}$ and $\mathcal{E}$ that are constructed from covariance matrices (having many more arbitrary parameters). The minimum x_m is unique for a given quadratic cost function and Hilbert spaces $\mathcal{C}$ and $\mathcal{E}$ if $\mathcal{M}$ is linear and continuous and satisfies the requirements of a regularized solution. This minimum can be derived either by solving linear equations or by choosing among many ones an efficient iterative algorithm.

The value $C\,(x_m)$ is a trade-off between requirements on $\mathcal{C}$ and $\mathcal{E}$. Either it is larger than the largest number C_M allowed for this trade-off, so that there is no admissible regularized solution of this form, or it is smaller than C_M, and all values of x such that $C\,(x_m) \leq C(x) \leq C_M$ are also admissible solutions of Equation (2). Hence the range of this non-uniqueness can be appraised along the principal axes of the ellipsoid $C(x) = C_M$, whose center is x_m. In this linear case, the center is given by the formula $(M^*M + \lambda I)^{-1}\, M^*e$, where the regularising process is clearly due to adding λ^{-1} to the eigenvalues of M^*M (the “singular values of M). Extensions along the axes are trivially related to the inverse of $\mathcal{M}$ singular values, damped by means of λ.

Instabilities related to the small singular values (high spatial frequencies) make this kind of regularized solution more sensitive to accidental large deviations of data (less robust) than regularized solutions obtained from cost functions involving L^1 or L^∞ norms. However, choosing the later norms usually leads one to more complicated algorithms and to (bang–bang) solutions piecewise equal to the possible upper and lower bounds. Hence, simple and efficient analyses can be adapted to almost any linear (or linearized) inverse problem, including those important for medical tomography and most medical imaging (Bertero & Bocacci, 1998).

Among many examples of nonlinear inverse problems is the case when the a priori information on a first guess x_0 is so convincing that we seek a unique approximate solution of Equation (1) in a narrow neighborhood of x_0, allowing linearization $\mathcal{M}$. The only difference between such a linearized inverse problem and a linear one is that the conditions limiting $\mathcal{C}_0$ must involve those that ensure the relevance of our approximation.

A quite different case involves nonlinear inverse problems for which $\mathcal{M}$ is so general (with weak a priori conditions) that we cannot escape an exhaustive exploration of $\mathcal{C}^0$ for seeking the exact or approximate solutions of (1). Random walks through $\mathcal{C}^0$ can do the job, for example, Monte-Carlo methods, simulated annealing, or genetic algorithms.

Between these extreme cases are those that correspond to important physical or engineering problems, for which studies either give an exact method of solution or allow solution of the inverse problem by numerical techniques. Defining cost functionals and minimizing them by gradient iterative or other methods (Engl et al., 2000; Ghosh Roy & Couchman, 2002) is then possible, but the occurrence of several secondary minima can also make the job difficult, raising questions about their relevance as admissible solutions.

Solving an inverse problem means that we are able to derive all the admissible regularized solutions. If the non-uniqueness is due not only to small data or parameters deviations but to the structure of $\mathcal{M}$, these equivalent solutions should be classified into identified families. This is essential either to derive ways of restoring uniqueness (for instance, by making other kinds of measurements), to use the non-uniqueness (e.g., to design stealth targets), or to modify the nature of the model. In decisive as opposed to descriptive modeling, well-posed questions seek significant pieces of information instead of trying to get an image of the parameters.

A One-dimensional Example

Inverse scattering in one spatial dimension offers some completely solved examples, the simplest one being the Schrödinger problem on the line (Chadan & Sabatier, 1989; Aktosun & Klaus 2002), which may model observations of quantum, acoustic, or electromagnetic scattering. Experimental results are the scattering coefficients $R(k)$ and $T(k)$, $k \in \mathbb{R}$, related to the parameter $V(x)$ by the following formulas, where indices $\pm$ at the same level correspond :

$$\left[-\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V(x)\right] f^{\pm}(k,x) = k^2 f^{\pm}(k,x)\,, \qquad (3)$$

$$\mathrm{e}^{\mp\mathrm{i}kx f^{\pm}}(k,x) \to 1 \quad \text{as} \quad x \to \pm\infty\,, \qquad (4)$$

$$T(k)f^{-}(k,x) = f^{+}(-k,x) + R(k)f^{+}(k,x)\,, \qquad (5)$$

where $\int_{-\infty}^{+\infty} |V(x)| \left(1+x^2\right) \mathrm{d}x < \infty$. Equation (3) shows that $\mathbb{R}$ is the continuous part of the spectrum of $-\mathrm{d}^2/\mathrm{d}x^2 + V$, which has also a finite number of eigenvalues $k = \mathrm{i}\kappa_n$, $n = 1, 2, \ldots N$, corresponding to eigenfunctions in $L^2(\mathbb{R})$, with normalizing constants

ρ_n. The set $\{R(k),\ \kappa_n,\ \rho_n\}$, called the scattering (or spectral) data, defines the following function on $\mathbb{R}^+$:

$$S(z) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{d}k\, R(k)\mathrm{e}^{\mathrm{i}kz} + \sum_{p=1}^{N} \rho_p \mathrm{e}^{-\kappa_p z}. \quad (6)$$

Although knowledge of $R(k)$ and $T(k)$ is not sufficient to determine V, the spectral data is sufficient to determine everything $\left(V,\ f^{\pm}, T,\ \text{etc}\right)$. In fact, $\int_x^\infty V(t)\,\mathrm{d}t$ is equal to $2K\left(x, x^+\right)$, where $K(x, y)$ has been derived from S by solving the linear integral equation:

$$\begin{aligned} &K(x, y) + S(x + y) + \\ &\quad \int_x^\infty \mathrm{d}z K(x, z)\, S(z + y) = 0. \end{aligned} \quad (7)$$

Only the step $K \to V$ is (weakly) instable; hence, the inverse problem ($R \to V$) is completely solved but with a structural non-uniqueness distributed into families labeled by κ_n and ρ_n.

If we let t be an evolution parameter (time) and assume that the evolution of $S(k, t)$ is described by a linear equation, the evolution of the corresponding $V(x, t)$ is generally not linear. Hence, the couple direct-inverse problem can be used to linearize special nonlinear equations. This observation provides the basis of the inverse scattering transform (IST), where the discrete eigenvalues (κ_n) correspond to solitons.

Three-dimensional Inverse Scattering

The simplest model for inverse scattering in three spatial dimensions obeys the equations:

$$(-\Delta + V)\,\Psi(\boldsymbol{k}, \boldsymbol{x}) = k^2\Psi(\boldsymbol{k}, \boldsymbol{x}) \quad (8)$$

$$\begin{aligned} \Psi(\boldsymbol{k}, \boldsymbol{x}) \;=\; &\exp\,[\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}] + x^{-1}\exp\,[\mathrm{i}kx]\, F\left(\widehat{x}, \widehat{\boldsymbol{k}}, k\right) \\ &+0\left(x^{-1}\right), \end{aligned} \quad (9)$$

where the bold font indicates vectors (italics their length) and the hatted letters are elements of the unit sphere.

The scattering potential V is the parameter to be determined, and F the experimental observation, which in the general case depends on 5 variables, whereas V depends only on 3. Hence, the most general inverse problem is severely ill-posed. On the other hand, spherical symmetry may reduce it to a one-dimensional problem.

If V is a piecewise continuous parameter, the general case at fixed k can be treated by standard optimization techniques. However, in most target identifications (including sonar and radar), the scatterer is not modeled by a medium function $V(\boldsymbol{x})$ but by a reflecting (or absorbing) surface. In some cases, approximate methods are relevant—ray theory, physical optics, Rytov approximation, distorted wave Born approximation (DWBA)—and they often yield linear relations between well-chosen parameters and data, for example, the (inversable) Abel integral, which relates the velocity of an earthquake signal as a function of depth to the measured travel time as a function of the ray parameter. If in mineral prospecting by seismic reflection surveys, geological information and rough identification of the signals suggest that the underground structure may be represented as a homogeneous medium with few reflectors, so-called migration methods may be useful. (Migration methods reposition returned signals that are time-distorted to their true location in time, for example, for dipping structures.)

Unfortunately, the exact relations between scatterers and data are highly nonlinear. Inversion of the DWBA fails in diffraction tomography as the structural complexity increases and so do optimization methods in seismic problems (nonconvexity of cost functions, secondary minima, etc.). Studies of stealth targets also involve nonlinear inverse problems, but of course they are not well documented in the open literature.

Full studies of the inverse problems of target identification are available in well-known works (for approximations, see Langenberg, 1987; Hopcraft & Smith, 1992; for mathematics, see Colton & Kress, 1992; for comprehensive studies, see topics 1.3 to 1.5 in Pike & Sabatier, 2002. For direct and inverse problems of elastic waves, see topics 1.7 and 1.8 in Pike & Sabatier, 2002.) Acoustic soundings used in clinical medicine (i.e., ultrasound scans) show graphs of travel times between source and reflectors (i.e., impedance discontinuities), giving evidence of imaging reflectors without need of solving inverse problems and so do more precisely devices transmitting back a time-reversed version of the recorded wave field (Fink & Prada, 2001).

A few techniques of medical imaging yield inverse problems that are either linearizable or nonlinear, according to a priori information and constraints. Among these, electrocardiography (Johnston, 2001), electroencephalography, and magnetoencephalography try to determine the electrical activity in the heart and the brain as functions of space and time. In brain imaging, the number of data points ($\sim$100) and that of unknowns ($\sim$10, 000) make the linear problem severely underdetermined, leading to arbitrary choices in the space of parameters (Bayesian approaches are commonly used) and to low resolution. From improved a priori knowledge, on the other hand, one can assume relatively few localized active groups at a given time, which are the sources of activity, but the problem becomes nonlinear (Baillet et al., 2001).

PIERRE C. SABATIER

See also **Gel'fand–Levitan theory; Integrability; Inverse scattering method or transform**

Further Reading

Aktosun, T. & Klaus, M. 2002. Inverse theory: problem on the line. In *Scattering*, edited by E.R. Pike & P.C. Sabatier, San Diego and London: Academic Press, pp. 770–785

Baillet, S., Mosher, J.C. & Leahy, R.M. 2001. Electromagnetic brain mapping. *IFEE Signal Processing Magazine*, 18(6): 14–30

Bertero, M. & Bocacci, P. 1998. *Introduction to Inverse Problems in Imaging*, Bristol and Philadelphia: Institute of Physics Publishing

Chadan, K. & Sabatier, P.C. 1989. *Inverse Problems in Quantum Scattering Theory*, 2nd edition, Berlin and New York: Springer

Colton, D. & Kress, R. 1992. *Inverse Acoustic and Electromagnetic Scattering Theory*, Berlin and New York: Springer

Engl, H.W., Hanke, M. & Neubauer, A. 2000. *Regularisation of Inverse Problems*, Dordrecht and Boston: Kluwer

Fink, M. & Prada, C. 2001. Acoustic time-reversal mirrors. *Inverse Problems*, 17(1): R1–R38

Ghosh Roy, D.N. & Couchman, L.S. 2002. *Inverse Problems and Inverse Scattering of Plane Waves*, San Diego and London: Academic Press

Hopcraft, K.I. & Smith, P.R. 1992. *An Introduction to Electromagnetic Inverse Scattering*, Boston and Dordrecht: Kluwer

Johnston, P. (editor). 2001. *Computational Inverse Problems in Electrocardiography*, Southampton: WIT Press

Langenberg, K.J. 1987. Applied inverse problems for acoustic electromagnetic and elastic wave scattering. In *Basic Methods of Tomography and Inverse Problems*, edited by P.C. Sabatier, Bristol: Adams Hilger, pp. 125–467

Pike, E.R. & Sabatier, P.C. (editors). 2002. *Scattering*, San Diego and London: Academic Press

Sabatier, P.C. 2000. Past and future of inverse problems. *Journal of Mathematical Physics*, 41(6): 4082–4124

INVERSE SCATTERING METHOD OR TRANSFORM

The story begins one August day in 1834 on the Union Canal near Edinburgh. John Scott Russell observed a mass of water, displaced by the motion of a barge, race away along the channel and disintegrate into a series of waves, the likes of which he had not seen before. Following on horseback, he noted the lead wave was about 30 ft long and about a foot high. It was a wave purely of elevation above the mean level of the canal, and it traveled at about eight miles/hr without change of shape or loss of speed for a great distance. It took another 40 years before Joseph Boussinesq and Lord Rayleigh verified that such a solitary wave was indeed a solution of the shallow water equations. Twenty years later, in 1895, Diederik Korteweg and Hendrik de Vries (KdV) wrote down the now famous, universal, and ubiquitous KdV equation

$$q_t + 6qq_x + q_{xxx} = 0 \tag{1}$$

for the scaled elevation $q(x, t)$ of a right-going wave in a frame of reference moving with the shallow water speed $\sqrt{gh_0}$ on a layer of water of depth h_0 and its solitary wave solution

$$q(x, t) = 2\eta^2 \operatorname{sech}^2 \eta(x - x_0 - 4\eta^2 t). \tag{2}$$

In the shallow water context, the elevation is

$$Ah_0 \operatorname{sech}^2 \left\{ \frac{\sqrt{3A}}{2h_0} \left(x - \sqrt{gh_0}\,(1 + A/2)\, t - x_o \right) \right\},$$

where g is the acceleration of gravity, A is the arbitrary dimensionless (small) height, and x_0 is the starting position of the wave.

The story continues almost 60 years later in Los Alamos. The question under address was totally different. Why do solids have finite rather than infinite heat conductivity? Peter Debye had suggested in 1916 that the anharmonicity (weak nonlinearity) of the interatomic forces hindered energy transport by coupling all the natural vibrations of the solid; thus, the relaxation time to thermal equilibrium would be a measure of conductivity. Enrico Fermi, John Pasta, and Stan Ulam (FPU) decided in the early 1950s to simulate numerically a model embracing this idea. The model consisted of 63 equal masses m attached by 64 springs with the force in each of them given by $k(\Delta + \alpha\Delta^2)$, a nonlinear modification of Hooke's law where Δ represents displacement from equilibrium. The initial energy was all in the lowest mode. As expected, very soon energy spread to the first few harmonics. But then a surprising thing occurred. The energy did not thermalize! Instead, after a while, it all recollected into the lowest mode. The pattern repeated. The negative outcome of the FPU experiment, one of the first scientific discoveries by computer experiment, was considered a great mystery (*See* **Fermi–Pasta–Ulam oscillator chain**).

Carpe Diem

The moment was seized by two young applied mathematicians, Martin Kruskal and Norman Zabusky (KZ). Because most of the energy of the FPU experiment resided in the low modes, Kruskal and Zabusky argued that they could replace the spring equations for the lateral displacements $y_n(t)$, $n = 1 \ldots 63$ by a continuum approximation $y(x = nh, t) = y_n(t)$, where h was the unstretched spring length. For disturbances traveling in one direction in a frame moving with the lattice velocity $c = \sqrt{k/m}\, h$, the suitably scaled strain $q(x, t) = y_x/6$ obeyed the KdV equation (1). They numerically simulated solutions of (1) on a periodic domain of length L beginning with $q(x, 0) = (\pi a/6) \cos(2\pi x/L)$. They found that the solution evolved toward a sequence of solitary wave pulses each very close to (2) in shape that ran around the interval with the taller, faster ones periodically catching up with the slower ones, thereby producing partial, and occasionally nearly perfect, recurrences. Moreover, they made one startling observation. After a pairwise collision, the two solitary waves would emerge with their former identities (amplitude, width, and speed) intact. The only evidence that the interaction was nonlinear was a phase shift. The faster (slower) pulse seemed to advance (regress) by a fixed amount. Because the solitary waves exhibited such robust, particle-like properties, Kruskal and Zabusky called them solitons (*See* **Solitons, a brief history**).

Such behavior indicated that the KdV equation was somehow special and that its special properties were connected with hidden symmetries and conservation laws, beyond those connected with mass, momentum, and energy. A systematic search led to the Miura–Gardner transformation, $q(x,t) = w(x,t) + \mathrm{i}\varepsilon w_x(x,t) + \varepsilon^2 w^2(x,t)$, a Riccati equation mapping solutions of a modified KdV equation $w_t + 6(w + \varepsilon^2 w^2)w_x + w_{xxx} = 0$ into solutions of (1). The single conservation law $w_t + (3w^2 + 2\varepsilon^2 w^3 + w_{xx})_x = 0$ and the corresponding conserved quantity $\int w\,\mathrm{d}x$ (the integration interval is either infinite with $w(\pm\infty, t) = 0$ or finite with $w(x+L, t) = w(L, t)$) gave rise to an infinite number of conservation laws or conserved quantities for KdV, found by solving w iteratively in power of ε. Solving the Riccati equation by setting $w + 1/2\varepsilon^2 = \mathrm{i}v_x/\varepsilon v$ gives

$$v_{xx} + (q(x,t) + \lambda)v = 0, \tag{3}$$

the Schrödinger equation with energy λ and potential $-q(x,t)$. This connection of solutions of KdV with the Schrödinger operator $L(t) = -\,\mathrm{d}^2/\mathrm{d}x^2 - q(x,t)$ gave rise to the inverse scattering method. By this stage Kruskal had been joined by colleagues Gardner, Greene, and Miura (GGKM).

A Natural Question

In the late 1960s, GGKM began by asking: As the potential $-q(x,t)$, defined on $-\infty < x < \infty$ and decaying to zero as $x \to \pm\infty$, evolves according to (1), how does the spectrum (the set of eigenvalues) of (3) change? The character of the spectrum was already well known. It consists of a discrete number N of negative energy levels $\lambda_n < 0$, $(1 \le n \le N)$, with corresponding eigenfunctions that are square integrable and a positive continuous spectrum $\lambda > 0$. Direct substitution of $q(x,t)$ from (3) into (1) gave $\lambda_t v^2 + (uQ_x - u_x Q)_x = 0$ with $Q \equiv v_t + v_{xxx} - 3(\lambda - q)v_2$. For bound state eigenvalues, the integral of the second term vanishes giving $\lambda_{nt} \int_{-\infty}^{\infty} v_n^2\,\mathrm{d}x = 0$ so that λ_n is a motion constant. It turns out that $-2\lambda_n$ is the amplitude of the nth soliton which is contained in $q(x,t)$. The fact that it is constant is the reason for the special collision property of solitons. Because $\lambda_t = 0$, $uQ_x - u_x Q = \text{constant}$, which can easily be solved for Q to give an equation for the evolution of the eigenfunction $v(x,t)$

$$v_t = B \cdot v = -4v_{xxx} - 6qv_x + (C - 3q_x)v, \tag{4}$$

which is also equal to (remember $\lambda v = -v_{xx} - qv$)

$$v_t = B \cdot v = (q_x + C)v + (4\lambda - 2q_x)v_x. \tag{5}$$

The parameter C is a constant of integration that is used to normalize in a convenient way solutions of (3) called Jost functions. The compatibility $(v_{xx})_t = (v_t)_{xx}$ of (3) and (4) or (5) gives (1).

Peter Lax provided an elegant formalism for this compatibility. Differentiating $L(t)v = [-v_{xx} - q(x,t)v] = \lambda v$ with respect to t and using (4), one obtains $L_t = [B, L]$, which is (1) and is called the Lax equation. L and B are called the Lax pair. The Lax equation guarantees that the flow is an isospectral deformation; namely, that as the potential in (3) deforms according to (1), the spectrum remains unchanged.

The Inverse Scattering Method

The stage is now set to introduce the inverse scattering method (ISM)—or inverse scattering transform (IST)—which is done in three steps. First, one is given the potential $-q(x)$ and from this one defines solutions of (3) and calculates a set of data called the scattering data S. Second, one determines how S evolves as $q(x,t)$ evolves according to (1). Finally, one recovers the potential $-q(x)$ from the time-advanced scattering data S. This is analogous to the set of steps one uses in solving linear PDEs via the Fourier transform, and in a suitable "small" $-q(x)$ limit, S is the Fourier transform. The IST algorithm is

$$\begin{aligned} &\text{Step 1}: \ q(x,0) \to S(0).\\ &\text{Step 2}: \ S(0) \to S(t).\\ &\text{Step 3}: \ S(t) \to q(x,t). \end{aligned}$$

Interestingly, IST can be interpreted as a canonical map from the original coordinates $q(x,t)$ to suitable combinations of S that are action-angle variables.

In the first step, one defines two sets of linearly independent solutions $\{\varphi(x,k),\ \varphi(x,-k)\}$, $\{\psi(x,k),\ \psi(x,-k)\}$, $\lambda = k^2$ for real values of k (called Jost functions) by their asymptotic properties for all t as $x \to \pm\infty$,

$$\varphi(x,k) \to \mathrm{e}^{-\mathrm{i}kx}, \quad x \to -\infty; \ \ \psi(x,k) \to \mathrm{e}^{\mathrm{i}kx}, \quad x \to +\infty. \tag{6}$$

Because of unit (time-independent) coefficients, the choice of C in (5) for $\varphi(x,t,k)$ is $4\mathrm{i}k^3$ and the choice of C in (5) for $\psi(x,t,k)$ is $-4\mathrm{i}k^3$. Since (3) is linear, and second order, the pairs of Jost functions are related

$$\varphi(x,t,k) = a(k,t)\psi(x,t,-k) + b(k,t)\psi(x,t,k), \tag{7}$$

where $2ika(k,t)$ and $2ikb(k,t)$ are, respectively, the Wronskians $\varphi(k)\psi_x(+k) - \varphi_x(k)\psi(+k)$ and $\varphi(k)\psi_x(-k) - \varphi_x(k)\psi(-k)$. For a bound state $\lambda_n < 0$ so that $k_n = \sqrt{\lambda_n} = \mathrm{i}\eta_n$. Since $\psi(x,t,-k)$ diverges as $x \to +\infty$ for $\operatorname{Im} k > 0$, the coefficient $a(k,t)$ must vanish for a bound state so that at $\lambda = \mathrm{i}\eta_n$, $\varphi(x,t,\mathrm{i}\eta_n) = b_n(t)\psi(x,t,\mathrm{i}\eta_n)$. Dividing (7) by $a(k,t)$ and using (6), one sees that

$$\begin{aligned}\frac{\varphi}{a} &\to \frac{1}{a}\,\mathrm{e}^{\mathrm{i}kx} \text{ as } x \to -\infty\\ &\to e^{\mathrm{i}kx} + \frac{b}{a}\,\mathrm{e}^{\mathrm{i}kx} \text{ as } x \to +\infty. \end{aligned} \tag{8}$$

In wave scattering or quantum mechanical language, φ/a is a solution which represents an incoming wave of amplitude unity from $x = \infty$ and waves partially reflected and partially transmitted by the potential $-q(x)$. The coefficients $T(k) = 1/a(k)$ and $R(k) = b(k)/a(k)$ are called the transmission and reflection coefficients, respectively.

It turns out that $\varphi(x,t;k)\mathrm{e}^{\mathrm{i}kx}(\varphi(x,t;-k)\mathrm{e}^{-\mathrm{i}kx})$, $\psi(x,t,k)\mathrm{e}^{-\mathrm{i}kx}(\psi(x,t,-k)\mathrm{e}^{\mathrm{i}kx})$ and $a(k,t)(a(-k,t))$, originally defined for real k, admit analytic continuation to $\operatorname{Im} k > 0$ ($\operatorname{Im} k < 0$). Each tends to unity as $|k| \to \infty$. The zeros of $a(k)$, $k_n = \mathrm{i}\eta_n$, $n = 1, \ldots, N$ are the bound states. There is only a finite number of them. The set of data $S(t) = \{(k_n = \mathrm{i}\eta_n, b_n)_1^N;\ R(k),\ k \text{ real}\}$ is called the scattering data. From S, $a(k)$ can be found.

In step 2, one uses (5) to calculate the time dependence of $S(t)$. Differentiate (7) with respect to t and use (5) with $C = 4\mathrm{i}k^3$ for $\varphi(x,k)$ and $\psi(x,-k)$ and $C = -4\mathrm{i}k^3$ for $\psi(x,k)$ to find

$$a_t = 0, b_t = 8\mathrm{i}k^3 b. \tag{9}$$

For the bound state at $k_n = \mathrm{i}\eta_n$, differentiating $\varphi(x,k_n) = b_n\psi(x,k_n)$ gives $b_{nt} = 8\eta_n^3 b_n$. Therefore, the discrete eigenvalues $k_n = \mathrm{i}\eta_n$ are motion constants and $S(t) = \{(k_n = \mathrm{i}\eta_n,\ b_n(0)\exp 8\eta_n^3 t)_1^N,\ R(k,0)\exp 8\mathrm{i}k^3 t,\ k \text{ real}\}$.

In step 3, one uses $S(t)$ to reconstruct $q(x,t)$ via the Riemann–Hilbert algorithm, which seeks to reconstruct a meromorphic function for $\operatorname{Im} k > 0$ (think $\frac{\varphi(x,k)}{a(k)}\mathrm{e}^{\mathrm{i}kx}$) tending to unity as $|k| \to \infty$, $0 < \arg k < \pi$ and an analytic function for $\operatorname{Im} k < 0$ (think $\psi(x,-k)\mathrm{e}^{\mathrm{i}kx}$) tending to unity as $|k| \to \infty$, $-\pi < \arg k < 0$, whose difference on the real axis $\operatorname{Im} k = 0$ is given (think $R(k)\psi(x,k)\mathrm{e}^{\mathrm{i}kx}$). To achieve this, consider for $\operatorname{Im} k > 0$,

$$I = \frac{1}{2\pi\mathrm{i}}\int_{-\infty}^{\infty} \frac{\varphi(x,k')\mathrm{e}^{\mathrm{i}k'x}}{a(k')(k'+k)}\,\mathrm{d}k' \tag{10}$$

in two ways. First, close the contour at $|k| = \infty$ along $0 < \arg k < \pi$. Second, after using (7), close the contour for that part analytic for $\operatorname{Im} k < 0$ in the lower half-plane. Identifying the two results, one finds a linear integral equation

$$\begin{aligned}\psi(x,k)\mathrm{e}^{-\mathrm{i}kx} &= 1 - \sum_{n=1}^{N} \frac{\gamma_k\psi(x,k_n)}{k+k_n}\,\mathrm{e}^{\mathrm{i}k_n x}\\ &\quad + \frac{1}{2\pi\mathrm{i}}\int_{-\infty}^{\infty} R(k)\frac{\psi(x,k')\mathrm{e}^{\mathrm{i}k'x}}{k'+k}\,\mathrm{d}k', \end{aligned} \tag{11}$$

where $k_n = \mathrm{i}\eta_n$, $\gamma_n = b_n(a'(k_n))^{-1}$. The potential $-q(x)$ is recovered from (11) by the formula

$$q(x) = \lim_{k\to\infty} -2\mathrm{i}k\,\frac{\mathrm{d}}{\mathrm{d}x}\,(\psi(x,k)\mathrm{e}^{-\mathrm{i}kx} - 1). \tag{12}$$

A simple transformation converts (11) into the famous Gel'fand–Levitan equation which, since the mid-1950s, was known as a means for reconstructing potentials from scattering data.

For a special set of initial conditions $q(x,0)$ called reflectionless potentials, the reflection coefficient $R(k)$ is identically zero. In that case, the Riemann–Hilbert algorithm yields a set of linear algebraic equations found by putting $k = \mathrm{i}\eta_r, r = 1, \ldots, N$ in (11). The simplest is the single soliton solution for which $R(k) = 0$, $a(k) = (k - \mathrm{i}\eta)/(k + \mathrm{i}\eta)$, $b_1(0) = \exp 2\eta\, x_0$, $b_1(t) = \exp 2\eta\bar{x}$, $\bar{x} = x_0 + 4\eta^2 t$, $\psi(x,k_1) = \mathrm{e}^{-\eta x}(1 + \mathrm{e}^{2\eta(\bar{x}-x)})^{-1}$ and $q(x,t) = 2\eta^2\operatorname{sech}^2\eta(x - \bar{x})$. Multisoliton solutions can be similarly calculated. They can also be found by a nonlinear superposition principle—the Bäcklund transformation—which is introduced below.

For $q(x,0) = Q\delta(x)$, $a(k) = (Q + 2\mathrm{i}k)/2\mathrm{i}k$ and $R(k) = -Q/(Q + 2\mathrm{i}k)$, and there is one bound state at $k = \mathrm{i}Q/2$. The field consists of a single soliton with $\eta = Q/2s$ and a trailing radiation component. The two components are tied together with the self-similar solution

$$q(x,t) = \frac{1}{(3t)^{2/3}}\,f\left(\frac{x}{(3t)^{1/3}}\right),$$

which we will meet again when we discuss isomonodromic deformations.

The question: "Is the complete integrability of KdV an isolated miracle or are there lots of integrable PDEs?" was quickly resolved. The pioneering GGKM work of the late 1960s was followed within a few years by the discovery of the complete integrability via ISM of the equally ubiquitous nonlinear Schrödinger (NLS) equation by Zakharov and Shabat and the characterization and integrability via ISM of families of integrable PDEs by Ablowitz, Kaup, Newell, and Segur (AKNS), (*See* **Ablowitz–Kaup–Newell–Segur system**).

More Space Dimensions

Completely integrable systems in more than one space dimension were also found. Two of the most important were the two-dimensional analogues of the KdV and NLS equations, the Kadomtsev–Petviashvili (KP) and the Benney–Roskes–Davey–Stewartson (BRDS) equations. The KP equation describes the evolution of slightly oblique waves either on shallow water or in a two-dimensional FPU lattice where the linear spring force in one direction is much weaker than it is in the other. It is given by (1) with $\int_{-\infty}^{x} q_{yy}$ added. It is also the compatibility condition of two linear operators with coefficients depending on $q(x, y, t)$ and its derivatives.

Unexpected connections of integrable PDEs with other exactly integrable models (Ising, Yang–Baxter) were found in the late 1970s and the 1980s. Most notable among them was the discovery by Sato, Miwa, and Jimbo (SMJ) that in the scaling limit, the n-point correlation function of the nearest neighbor, two-dimensional Ising model satisfies a system of very special deformation equations. These express the fact that the monodromy group of an associated linear system, which contains the correlation functions as coefficients, is preserved as the correlation functions change with their arguments. (The monodromy group of matrices encodes information about how the fundamental solution matrix of the linear system changes on taking it through a loop surrounding regular and irregular singular points.) Much of their work has deep connections with classical work on isomonodromic deformations by Ludwig Schlesinger and Paul Painlevé on systems of Fuchsian differential equations (*See* **Monodromy preserving deformations**) and by Albert Bäcklund on superposition principles for nonlinear equations (*See* **Bäcklund transformations**). Indeed, soliton theory has had a major impact in many areas of mathematical physics. It has led, for example, to the solution of the Schottky problem of algebraic geometry, to the discovery of quantum groups, to models for quantum gravity, and to useful connections with arithmetic algebraic geometry and, in particular, Jones' polynomials.

Bäcklund Transforms and Hirota's Method

To see the connection with Bäcklund transformations, it is useful to consider further the KdV family of completely integrable PDEs given by

$$q_{t_{2n+1}} = \frac{\partial}{\partial x} L^n q, \; L = -\frac{1}{4}\partial_x^2 - q - \frac{1}{2}\int_{\infty}^{x} \mathrm{d}x \, q_x, \quad n = 0, 1, \ldots, \tag{13}$$

where $t_{3/4}$ is the t of Equation (1) and $L^0 q = q$. To find them is easy. Write (5) as

$$v_t = A(\lambda; q, q_x, \ldots)v - D(\lambda; q, q_x, \ldots)v_x + Cv, \tag{14}$$

where $A,\ D$ are polynomials in λ with coefficients depending on q and its x derivatives and C is the normalizing constant. The compatibility of (3) and (14), $(v_{xx})_t = (v_t)_{xx}$, tells one that $A = \frac{1}{2} D_x$ and $q_t = (-\frac{1}{2}\partial_x^3 - 2q\partial_x - q_x - 2\lambda\partial_x)D$. Solve for D as $D = -\lambda^n + D_1\lambda^{n-1} + \ldots + D_n$. Comparing powers of λ determines each D_r in terms of its predecessor as $2\partial_x D_{r+1} = (-\frac{1}{2}\partial_x^3 - 2q\partial_x - q_x)D_r, \quad r = 0, \ldots, n$ with $D_0 = 1$ and D_{n+1} defined by $r = n$. The λ^0 balance gives $q_t = 2\partial_x D_{n+1} = \partial_x L^n q$.

Equations (13) are a completely integrable family of Hamiltonian commuting flows containing at level $n = 1$ the KdV equation. They all share the same constants of motion $\{H_{2r+1}\}_{-1}^{\infty}$ with $H_{-1} = \int_{-\infty}^{\infty} q \, \mathrm{d}x, \quad H_1 = \int_{-\infty}^{\infty} \frac{1}{2} q^2 \mathrm{d}x, \quad H_3 = \frac{1}{8}\int_{-\infty}^{\infty} (q_x^2 - 2q^3)\, \mathrm{d}x$, each of which serves as the Hamiltonian for one of the flows $q_{t_{2n+1}} = \partial_x \frac{\delta H_{2n+1}}{\delta q}$ where $\frac{\delta H}{\delta q}[q]$ is the variational derivative of $H[q]$ and ∂_x is the skew symmetric symplectic operator. They are also bi-Hamiltonian in that $\partial_x \frac{\delta H_{2n+1}}{\delta q} = \big(-\frac{1}{4}\partial_x^3 - q\partial_x - \frac{1}{2}q_x\big) \frac{\delta H_{2n-1}}{\delta q}$. They commute under the Poisson bracket

$$\{H_{2m+1},\ H_{2n+1}\} = \int \frac{\delta H_{2m+1}}{\delta q} \frac{\delta}{\delta n} \frac{\delta H_{2n+1}}{\delta q} \, \mathrm{d}x = 0.$$

One consequence of commutativity is that if one begins with the shape $q(x, 0, 0, \ldots)$ and evolves q with respect to t_{2n+1} for a time $t_{2n+1},\ n = 0, \ldots$, in any order, one always obtains the same shape $q(x, t_1, t_3, \ldots, t_{2n+1}, \ldots)$. This means that the phase shift experienced by any two solitons is the same for each member of the family. Why? Start with the larger, faster moving solution behind the smaller, slower one (the velocity of the soliton of flow n is proportional to η^{2n}). Evolve $q(x, t_{2m+1} = 0, t_{2n+1} = 0)$ via the t_{2m+1} flow for t_{2m+1} time units sufficient for the faster soliton to overtake the slower one and then for t_{2n+1} units via the t_{2n+1} flow. Now do it in the reverse order. Because $q(x, t_{2m+1}, t_{2n+1})$ is the same no matter which order is used, the phase shift must be the same.

A second consequence of commutativity is that one can express the vector made up of the string $L^n q$ as the x derivative of the gradient of a potential which we call $2 \ln \tau(x, t_1, t_3, \ldots)$, namely, $L^n q = 2\frac{\partial}{\partial x} \cdot \frac{\partial}{\partial t_{2n+1}} \ln \tau$. The function $\tau(x, t_1, t_3, \ldots)$, called the Hirota τ function, has remarkable properties. It converts each member of the family (13) into bilinear form. For $n = 1$, the KdV equation under $q = 2\frac{\partial^2}{\partial x^2} \ln \tau$ becomes (recall $q_{t_1} = q_x$ and take $4t_3 = t$)

$$D_x^2 \tau \cdot \tau = \tau\tau_{xt} - \tau_x\tau_t + \tau\tau_{xxxx} - 4\tau_x\tau_{xxx} + 3\tau_{xx}^2 = 0, \tag{15}$$

where the symbol D_x acting on the ordered pair σ, τ is defined as the limit as $\varepsilon \to 0$ of $\frac{\partial}{\partial\varepsilon}\sigma(x+\varepsilon)\tau(x-\varepsilon) = \sigma_x\tau - \sigma\tau_x$. For the KP equation, one adds $\tau\tau_{yy} - \tau_y^2$ to the left-hand side of (15). From (15), one can build multisoliton solutions. The solution $\tau = 1$ corresponds to the vacuum state $q = 0$. The solution $\tau = 1 + \exp\theta(x,t)$, $\theta(x,t) = 2k(x - x_0) - 8k^3t$ corresponds to the single solution solution $q(x,t) = 2k^2\mathrm{sech}^2k$ $(x - x_0 - 4k^2t)$. The next solution in the sequence is $\tau = 1 + \exp\theta_1(x,t) + \exp\theta_2(x,t) + \exp(\theta_1(x,t) + \theta_2$ $(x,t) + A_{12})$, $\theta_j(x,t) = 2k_j(x - \overline{x})$, $x_j = 4k_j^2t + x_{j0}$, $j = 1, 2$, $A_{12} = 2\ln\left|\frac{k_1 - k_2}{k_1 + k_2}\right| < 0$, and is a superposition of two solitons with amplitudes $2k_1^2$ and $2k_2^2$, which for large negative t and large positive t are separated with the faster one (assume $k_1^2 > k_2^2$) being on the left(right) when $t = -\infty(+\infty)$. For $t \to -\infty(+\infty)$, near $x \simeq \overline{x}_2$, the center of the slower pulse, it is the second and fourth (first and third) terms that dominate τ so that for large negative (positive) t, the second soliton looks like

$$2k_2^2\mathrm{sech}^2k_2\left(x - \overline{x}_2 - \frac{1}{2k_2}|A_{12}|\right)$$
$$\times(2k_2^2\mathrm{sech}^2k_2(x - \overline{x}_2)).$$

The collision shifts the slower soliton back by $\frac{1}{2k_2}|A_{12}|$ and, by a similar argument, the faster soliton ahead by $\frac{1}{2k_1}|A_{12}|$.

What does the interaction look like? For k_1 much greater than k_2, $|A_{12}| \simeq 0$, and the faster soliton rides adiabatically over the slowly changing slower one. For k_1 greater than but close to k_2, there is an exchange of identities with the slower soliton assuming the form of the faster one as soon as the trailing edge of the former feels the leading edge of the latter.

But a century before Ryogo Hirota's work in the 1970s (*See* **Hirota's method**), the works of Bäcklund, Darboux, and Schlesinger had shown how to build, for certain classes of nonlinear equations, complicated solutions from simple ones. Applied to any member of the KdV family, the idea is this. Define

$$q = -u_x = 2\partial_x^2\ln\tau \quad \text{and} \quad \tilde{q} = -\tilde{u}_x = 2\partial_x^2\ln\tilde{\tau}$$

Then there are two relations, called Bäcklund transformations, which express $\tilde{u}_x + u_x$ and $\tilde{u}_t + u_t$ as functions of $\tilde{u} - u$ and a new free parameter ζ^2. Then, if q satisfies KdV, the enriched solution $\tilde{q}$ also satisfies KdV. These expressions take on a beautifully simple form that reveals the algebraic structure underlying the hidden symmetries of the KdV family when they are expressed in terms of τ functions as $\tilde{\tau} = \tau v$ where $v(x, \varsigma^2)$ solves (3) with the q corresponding to τ and $\lambda = \zeta^2$. But one can also express v as a combination of $\frac{X(\zeta)\tau}{\zeta}$ and $\frac{X(-\zeta)\tau}{\zeta}$, where the operator

$$X(\zeta) = \exp\left(\mathrm{i}\sum_{o}^{\infty}\zeta^{2k+1}t_{2k+1}\right)$$
$$\times\exp\left(\sum_{0}^{\infty}\frac{\mathrm{i}}{(2k+1)\zeta^{2k+1}}\frac{\partial}{\partial t_{2k+1}}\right),$$

so that $\tilde{\tau} = \tau_{\text{new}} = (AX(\zeta) + BX(-\zeta))\tau_{\text{old}}$.

For example, choose $\zeta = \mathrm{i}\eta$. Then if $\tau_{\text{old}} = 1$, $\tau_{\text{new}} = A\mathrm{e}^{\theta} + B\mathrm{e}^{-\theta} = 2\sqrt{AB}\cosh(\theta - \theta_0)$, where $\frac{A}{B} = e^{2\theta_0}$ and $\theta = \sum_0^{\infty}(-1)^k\eta^{2k+1}t_{2k+1}$ and τ_{new} is the one soliton solution. Because of the logarithm, we can also write τ_{new} as $(1 + \beta Y(\zeta))\tau_{\text{old}}$ where $Y(\zeta)$ is called the vertex operator

$$Y(\zeta) = \exp\left(-2\mathrm{i}\sum_{0}^{\infty}\zeta^{2k+1}t_{2k+1}\right)$$
$$\times\exp\left(\sum_{k=0}^{\infty}\frac{2}{\mathrm{i}(2k+1)\zeta^{2k+1}}\frac{\partial}{\partial t_{2k+1}}\right), \tag{16}$$

which has the property that $Y(\zeta)\cdot Y(\zeta')1 = 0$ if $\zeta = \zeta'$. Thus $Y^2 = 0$ and we can replace $1 + \beta Y(\zeta)$ by $\exp\beta Y(\zeta)$. The Bäcklund transformation is therefore

$$\tau_{\text{new}} = \exp\beta Y(\zeta)\cdot\tau_{\text{old}} \tag{17}$$

namely, the action of the "group" (infinite-dimensional groups are not rigorously defined) element corresponding to the algebraic element $Y(\zeta)$, which can be expressed as a Laurent series $\sum_{-\infty}^{\infty}Y_{2k+1}\zeta^{2k+1}$. The coefficients Y_{2k+1} obey a nontrivial set of commutator relations that form an infinite dimensional, graded Lie algebra (Kac–Moody algebra), the central extension of the loop algebra of Sl $(2, C)$. Under repeated application of (17), solutions of KdV trace out the orbit of the highest weight vector $\tau = 1$ in the basic representation of the Kac–Moody algebra. The algebra acts as an algebra of symmetries. There is also a complementary treatment where the algebra is used as the phase space on which there is defined both a natural Poisson bracket and Hamiltonian vector field.

Other Topics

Multisoliton solutions of the KdV family are a special case of finite gap solutions. The latter are defined by adding a nonlinear, constant coefficient ODE called the Lax–Novikov equation

$$\sum_{0}^{N}a_{2r+1}(L^rq)_x = \sum_{0}^{N}a_{2r+1}\,q_{t_{2r+1}} = 0 \tag{18}$$

as a constraint on (3) and (14). To illustrate the basic idea, assume traveling wave solutions

$$q(x,t) = q(x - ct) \text{ satisfy } -cq_x + (q_{xx} + 3q^2)_x = 0,$$

which is (18) with $a_1 = -c$, $a_3 = 4$, and $a_{2k+1} = 0$, $n > 1$. The Lax–Novikov equation, together with (3) and (14) gives rise to a Riemann surface, the analogue of the spectrum for potentials decaying at infinity, $y^2 = \Pi_{j=1}^{2N}(\lambda - \lambda_j)$ which remains invariant on the finite gap solution family. The τ function for the general N gap solution is the Riemann theta function.

Another class of solutions, the multiphase self-similar solutions, is found by attaching a nonlinear non-autonomous ODE

$$a_0 q + \sum_0^N a_{2r+1}\, t_{2r+1} (L^r q)_x = a_0 q + \sum a_{2r+1}\, t_{2r+1}\, q_{t_{2r+1}} = 0 \qquad (19)$$

as a constraint to (3) and (14). It is closely related to the string equation of modern physics, and it leads to a system of linear ODEs for

$$\lambda \frac{\partial}{\partial \lambda} \left\{ \begin{matrix} v(x, t_{2r+1}, \lambda) \\ v_x(x, t_{2r+1}, \lambda) \end{matrix} \right\}$$

with coefficients depending on self-similar scalings of q and its derivatives. Solutions of the multiphase self-similar family are isomonodromic deformations of this system of ODEs in that they leave the monodromic structure invariant. At $\lambda = \infty$, an irregular singular point, the monodromic structure is expressed in terms of Stokes multipliers; at $\lambda = 0$, a regular singular point, the monodromy is that of a simple pole. Members of this family include the Painlevé equations.

For example, (19) with $a_0 = -2$, $a_1 = -1$, $a_3 = 3$, $t_3 = t$ is $-2q - xq_x + 3t(q_{xx} + 3q^2)_x = 0$ which forces $q(x,t)$ to have the form

$$\frac{1}{(3t)^{2/3}} f\left(X = \frac{x}{(3t)^{1/3}}\right),$$

which satisfies an ODE that is a close cousin of the second Painlevé equation. This solution plays a key role in joining the soliton component of the solution with the trailing wave-like radiation. Isomonodromic deformations provide the link between soliton equations and the exactly solvable models of statistical physics. For example, the two-point correlation function for the nearest-neighbor Ising model in the scaling limit satisfies the third Painlevé equation.

The existence of so many completely integrable systems of physical interest (KdV, Boussinesq, nonlinear Schrödinger, derivative nonlinear Schrödinger, massive Thirring, sine-Gordon, Maxwell–Bloch with inhomogeneous broadening, KP, BRDS, three-wave scattering, Raman scattering) leads one to ask how important to applications these integrable models are. In the set of all PDEs, they are of measure zero. On the other hand, there are two reasons for their importance. The first is that many systems such as shallow waves traveling over a channel of slowly varying depth can be treated as perturbations of PDEs integrable by ISM, and there are simple algorithms for computing how the previously constant action variables slowly evolve. The second reason is more subtle and remains to be rigorously proven. Many of the equations of physical interest arise as asymptotic approximations, as reductions of more complicated systems assuming (as one does in deriving KdV) weak nonlinearity and weak dispersion, or (as one does in deriving NLS) weak nonlinearity and strong dispersion of an almost monochromatic wavepacket. If the process of reduction preserves integrability, then as long as there is one integrable equation among the class that reduces to the universal equation of interest, the reduced equation will also be integrable.

Finally, how does one test for integrability? How does one uncover the hidden symmetries, the algebraic structure of a PDE or ODE? There is yet no foolproof method that works in all circumstances. As in the original discovery of the integrability of KdV, serendipity often plays the main role. Two of the more promising semi-algorithmic approaches are the Painlevé test (the location of any algebraic, logarithmic or essential singularity is independent of initial conditions and only the location of poles can depend on arbitrary constants of integration). and the Wahlquist–Estabrook method, which seeks to uncover the hidden symmetries by embedding the equation of interest as an integrability condition of a pair of linear systems whose algebraic structure it is the goal of the method to find. Each has its advantages and each has proven to be successful in extracting the appropriate Lax pair in several contexts. But the challenge of determining a foolproof method to determine whether a particular equation is integrable is still open.

Alan Newell

See also **Ablowitz–Kaup–Newell–Segur system; Bäcklund transformations; Hirota's method; Kadomtsev–Petviashvili equation; N-soliton formulas; Painlevé analysis**

Further Reading

Ablowitz, M.J., Kaup, D.J., Newell, A.C. & Segur, H. 1974. The inverse scattering transform-Fourier analysis for nonlinear problems. *Studies in Applied Mathematics*, 53: 249–315

Ablowitz, M.J. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM

Faddeev, L.D. & Takhtajan, L.A. 1987. *Hamiltonian Methods in the theory of solitons.* Berlin: Springer

Flaschka, H. & Newell, A.C. 1980. Monodromy and spectrum preserving deformations. *Communications in Mathematical Physics*, 76: 65–116

Fokas, A.S. & Zakharov, V.E. 1993. *Important Developments in Soliton Theory*, Berlin and New York: Springer

Gardner, C.S., Greene, J.M., Kruskal, M.D. & Miura, R.M. 1974. The Korteweg de Vries equation and generalization. VI. Methods of exact solution. *Communications in Mathematical Physics*, 27: 97–133

Newell, A.C. 1985. *Solitons in Mathematics and Physics*, Philadelphia: SIAM

ION ACOUSTIC WAVES

See **Nonlinear plasma waves**

ISING MODEL

The Ising or Lenz–Ising model was introduced by Wilhelm Lenz in 1920 as a simplified model of a ferromagnet. Lenz's student Ernst Ising studied the one-dimensional case (a linear chain of coupled magnetic moments) for his doctoral dissertation at the University of Hamburg, and his results published in 1925 showed the absence of a phase transition between a ferromagnetic and paramagnetic state. In 1942, Lars Onsager applied an ingenious mathematical method to the two-dimensional model and showed analytically the existence of a phase transition in the absence of an external magnetic field. Interest in the model has increased in recent years as it became the foundation for modeling changes of state in new areas such as spin glasses and neural networks. In 1975, David Sherrington and Scott Kirkpatrick introduced long-range competing interactions in the model and showed the emergence of a new type of glassy order, while John Hopfield in 1982 used the connection between spins and neurons in extending the Ising model into a model of neuronal networks.

The original motivation for the introduction of the model came from the knowledge that certain metals, such as iron, are ferromagnetic; that is, they form macroscopic domains with nonzero magnetization when the temperature is less than a characteristic Curie temperature (T_c), but this long-range order is lost at temperatures $T > T_c$. Because magnetic properties of atoms are a result of the orbital angular momentum as well as spin of their electrons, the existence of macroscopic size domains with non-zero magnetization stems from mutual alignment of permanent atomic dipole moments induced by dipole-dipole interactions. The Lenz–Ising simplification ignores quantum mechanics and assigns to an atom located in crystal site i a magnetic moment or "spin" S_i that takes only two values, namely, $S_i = \pm 1$. The energy or Hamiltonian for the system of N spins is

$$H = -\frac{1}{2}\sum_{i,j=1}^{N} J_{ij} S_i S_j - \sum_{i=1}^{N} B_i S_i, \tag{1}$$

where J_{ij} is the interaction between spins at sites i and j; respectively; B_i is a local magnetic field at site i; and the sums extend independently over all N sites. Any of the 2^N spin configurations can be thought of as a chemical solution of two species with N_+ and $N_- = N - N_+$ spins in up and down states, respectively, with number densities $n_\pm = N_\pm / N$; the entropy of this state is determined by the number of ways N spins are partitioned in up and down configurations, giving

$$S = -Nk_{\mathrm{B}}\left(n_+ \ln n_+ + n_- \ln n_-\right). \tag{2}$$

Furthermore, the density of $++$, $--$, and $+-$ pairs that determine short-range order is approximated in the Bragg–Williams approximation respectively, by n_+^2, n_-^2, and $2n_+n_-$. Thus, Equation (1) becomes

$$\begin{aligned} H = {} & -\tfrac{1}{2} zNJ\left(n_+^2 + n_-^2 - 2n_+n_-\right) \\ & -NB\left(n_+ - n_-,\right), \end{aligned} \tag{3}$$

where only interactions of strength J between z nearest-neighbor spins are included. Introducing the macroscopic magnetization $M = N_+ - N_- = N(n_+ - n_-)$, expressing the densities as $n_\pm = (1 \pm M/N)/2$, and substituting in the energy and entropy, respectively, one obtains

$$\begin{aligned} H = {} & -\frac{zJM^2}{2N} - MB \\ S = {} & NK\left[\ln 2 - \frac{1}{2}\left(1 + \frac{M}{N}\right)\ln\left(1 + \frac{M}{N}\right)\right. \\ & \left. -\frac{1}{2}\left(1 - \frac{M}{N}\right)\ln\left(1 - \frac{M}{N}\right)\right] \end{aligned} \tag{4}$$

The thermodynamics of the model is now completely determined from the minimization of the Helmholtz free energy $F = H - TS$ with respect to the magnetization M at constant external field B and temperature T. It yields

$$\frac{M}{N} = \tanh\beta\left(B + zJ\frac{M}{N}\right), \tag{5}$$

where $\beta = 1/k_{\mathrm{B}}T$. Equation (5) is identical to that obtained through the Weiss mean field theory predicting, after linearization and for zero external field, a Curie temperature determined by $k_{\mathrm{B}}T_c = zJ$.

If the interaction J_{ij} is not fixed but taken to be random, distributed according to a Gaussian with mean J_0/N and variance J^2/N, then

$$P(J_{ij}) = \left(\frac{N}{2\pi J^2}\right)^{1/2} \exp\left[-\left(J_{ij} - \frac{J_0}{N}\right)^2 \frac{N}{2J^2}\right]. \tag{6}$$

The resulting Ising model is referred to as the Sherrington–Kirkpatrick model, and its low-temperature phase at zero-external-field is a spin glass. This phase is not characterized by the long-range order typical in ferromagnets but by partial system freezing in time showing persistent but spatially random configurations whose thermodynamic signature is a sharp cusp in the zero field cooled susceptibility. A spin glass is typically characterized by metastability, frustration, and slow relaxation resulting from the multiple high barriers (introduced by disorder in the free energy) that force the system to remain trapped in local free energy minima for long times.

Introducing some changes in the Ising spin-glass model and with proper interpretation, one obtains a model for a network of neurons. Ising spins are now states of a neuron found in a firing ($S_i = +1$) or inhibitory state ($S_i = -1$), the interaction elements J_{ij} are synaptic strengths between neurons while the external field depends on a neuron firing threshold potential. In the learning mode, the network stores information on patterns in the synaptic strengths, while in the retrieval mode, these patterns are accessed when the network is exposed to partial information on the desired memories. If $S^\nu \equiv \{S_1^\nu, S_2^\nu, \ldots, S_N^\nu\}$ (for $\nu = 1, \ldots, p$) represent p patterns of N neurons, employment of the following version of Hebb's rule:

$$J_{ij} = \frac{1}{N}\sum_{\nu=1}^{p} S_i^\nu S_j^\nu \tag{7}$$

stores the p patterns in memory. In the Hopfield model for $p/N \le 0.14$, these states, as well as several other spurious ones, become attractors to the evolution flow of neuron updating rules. Whenever a spin-glass phase is present, either alone or in coexistence with the memory states, information retrieval is corrupted.

While the phase diagrams of Ising spin glasses and neural networks are quite complex, the one determined analytically by Onsager for the two-dimensional Ising model with nearest-neighbor interaction J is much simpler, demonstrating the presence of a Curie temperature at $k_B T_c \approx 2.269J$ with a low-temperature ferromagnetic phase. This shows that even though quite simple, the Ising model contains the essential features of the problem that motivated its introduction, the ferromagnetic-paramagnetic transition. The absence of a transition in one dimension as found by Ising can be understood qualitatively by the following argument of Wannier. Energy equal to $2J$ can introduce a spin down defect in an ordered chain of $N + 1$ spins that are all in state up, and this can be done in N ways resulting in entropy $k_B \ln N$. The free energy for this change is $\Delta F = 2J - k_B T \ln N$ and favors always the entropic term except at $T = 0$; thus, there cannot be long-range order in one dimension at finite temperatures.

G.P. Tsironis

See also **Attractor neural network; Ferromagnetism and ferroelectricity; Frustration**

Further Reading

Baxter, R.J. 1982. *Exactly Solvable Models in Statistical Mechanics*, London: Academic Press

Binder, K. & Young, A.P. 1986. Spin glasses: experimental facts, theoretical concepts and open questions. *Reviews of Modern Physics*, 58: 801–976

Müller B., Reinhardt, J. & Strickland, M.T. 1995. *Neural Networks: An Introduction*, Berlin: Springer

Sompolinski H. 1988. Statistical mechanics of neural networks. *Physics Today*, 41 (12 Dec): 70–80

Wannier, G.H. 1966. *Statistical Physics*, New York: Wiley

ITO'S FORMULA

See **Stochastic processes**

IZERGIN–KOREPIN EQUATION

See **Discrete nonlinear Schrödinger equations**

J

JACOBI ELLIPTIC FUNCTIONS

See **Elliptic functions**

JOSEPHSON JUNCTION ARRAYS

Josephson junction arrays (JJAs) consist of islands of superconductor arranged in an ordered lattice, coupled by Josephson junctions. Large JJAs are useful model structures for studying the dynamics of coupled nonlinear oscillators, phase transitions, frustration effects, vortex dynamics, and macroscopic quantum phenomena (Newrock et al., 2000).

JJAs are generally divided into classical arrays ($E_J/E_C \gg 1$) and quantum arrays ($E_J/E_C \ll 1$), depending on the ratio of the Josephson energy, E_J, to the charging energy, E_C (Fazio & van der Zant, 2001). Classical JJAs can be divided into overdamped arrays ($\beta_C \ll 1$) and underdamped arrays ($\beta_C \gg 1$), referring to the fact that the equation of motion for a single Josephson junction is identical to a damped pendulum. The dividing line is determined by the McCumber parameter, β_C, which defines the amount of damping in terms of the junction capacitance and resistance.

Modeling of classical JJAs is essentially based on solving a system of coupled ordinary differential equations for the superconducting phase differences φ_n of Josephson junctions that constitute the array. Typically, JJAs are fabricated using photolithography and the standard fabrication processes for superconductive electronics. They are very well characterized, leading to an interesting synergy between experiments, theory, and simulations.

The properties of JJAs depend on their dimensionality and the way Josephson junctions are connected. In the series one-dimensional (1-d) JJAs shown in Figure 1a, there is no direct interaction between Josephson junctions. The junctions can be coupled, for example, via a load resistor connected between the extremities of the array. Under certain conditions (Jain et al., 1984), when biased by a bias current flowing along the array, Josephson junctions perform coherent oscillations.

Parallel 1-d arrays shown in Figure 1b have inherent mutual coupling between Josephson junctions due to screening currents and magnetic flux quantization in superconducting loops. Such an array is essentially similar to a long Josephson junction, and it is described by the discrete sine-Gordon (SG) model. Due to the similarity between the discrete SG and continuous SG models, properties of vortices in parallel 1-d JJAs are similar to those of Josephson vortices (solitons called fluxons) in long Josephson junctions (Ustinov & Parmentier, 1996; Watanabe et al., 1996).

The two-dimensional (2-d) JJA illustrated in Figure 1c has attracted a great deal of interest as it is isomorphic to a 2-d XY spin system, which is a 2-d lattice of spins free to rotate in the XY plane. In its simplest form, the Hamiltonian of an array can be written as

$$H = -\sum_{\langle n,m\rangle} E_J \cos\left(\varphi_m - \varphi_n - f\right), \qquad (1)$$

where $\langle n, m\rangle$ means summing over nearest neighbors and f is the frustration factor depending on the externally applied magnetic field. It should be noted that model (1) is valid only under the condition that both E_J and the array cell size are small enough, i.e., that inductance effects play no role. In practical JJAs, this is typically not the case and these effects significantly complicate JJA dynamics.

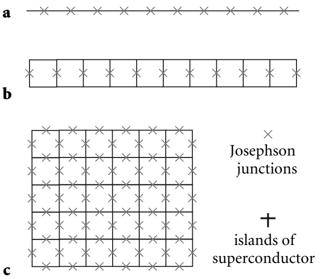

Figure 1. Schematic view of (a) 1-d series JJA; (b) 1-d parallel JJA; (c) 2-d JJA.

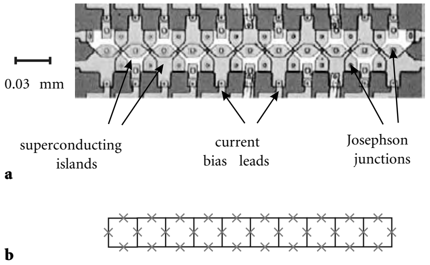

Figure 2. (a) Micro-photograph of a Josephson ladder (Binder et al., 2000). (b) The equivalent circuit of the ladder.

Starting from the mid-1990s, a special type of JJA called the Josephson ladder has received much interest. The ladder is essentially one row of cells cut from the standard 2-d JJA. Floría et al. (1996) studied a ladder driven by an ac bias current and found both oscillating and rotating localized modes. Such intrinsically localized modes, often called discrete breathers, are spatially localized time-periodic dynamical states that occur in various nonlinear lattices. The oscillating modes in Josephson arrays are difficult to detect experimentally as they are accompanied by zero average dc voltage. In contrast to oscillating discrete breathers, rotating discrete breathers induce a localized nonzero dc voltage and can be easily measured. Moreover, rotating discrete breathers can also be supported by a dc bias current. Experiments that followed the theoretical proposals successfully demonstrated dynamical localization in the form of discrete breathers in JJA (Ustinov, 2001).

There are several useful applications of JJA. At microwave frequencies, Josephson junctions can be used as sources and detectors of radiation. The Josephson relation $\Phi_0 \nu = V$ (where $\Phi_0 = h/(2e) = 2.07 \times 10^{-15}$ Vs is the magnetic flux quantum) states that at constant voltage V, the current through the junction will oscillate with a frequency ν, proportional to V. Utilizing this unique property, Josephson junction arrays have for the last two decades been studied as microwave sources (Darula et al., 1999). The advantages of using arrays as opposed to junctions are higher power outputs and better impedance matching to typical loads. Another important application of JJA, based on the same equation, is to define the voltage standard. Since frequency can be measured extremely accurately, locking Josephson junctions to a microwave source provides excellent accuracy for measuring the voltage. Using a series JJA instead of a single Josephson junction allows the upper voltage level to be brought from less than a millivolt up to the range of several volts.

Alexey V. Ustinov

See also **Breathers; Frenkel–Kontorova model; Frustration; Josephson junctions; Long Josephson junctions; Sine-Gordon equation; Superconducting quantum interference device; Superconductivity**

Further Reading

Binder, P., Abraimov, D. & Ustinov, A.V. 2000. Diversity of discrete breathers observed in a Josephson ladder. *Physical Review E*, 62: 2858–2862

Darula, M., Doderer, T. & Beuven S. 1999. Millimetre and sub-mm wavelength radiation sources based on discrete Josephson junction arrays. *Superconductor Science and Technology*, 12: R1–R25

Fazio, R. & van der Zant, H. 2001. Quantum phase transitions and vortex dynamics in superconducting networks. *Physics Reports*, 355: 235–334

Floría, L.M., Marin, P.J., Martinez, J.L., Falo, F. & Aubry, S. 1996. Intrinsic localization in the dynamics of a Josephson junction ladder. *Europhysics Letters*, 36: 539–544

Jain, A.K., Likharev, K.K., Lukens, J.E. & Sauvageau, J.E. 1984. Mutual phase-locking in Josephson junction arrays. *Physics Reports*, 109: 309–426

Newrock, R.S., Lobb, C.J., Geigenmüller, U. & Octavio, M. 2000. The two-dimensional physics of Josephson junction arrays. *Solid State Physics*, 54: 263–412

Ustinov, A.V. 2001. Experiments with discrete breathers in Josephson arrays. In *Nonlinearity and Disorder: Theory and Applications*, edited by F. Abdullaev, O. Bang & M.P. Soerensen, Dordrecht: Kluwer, pp. 183–185

Ustinov, A.V. & Parmentier, R.D. 1996. Coupled solitons in continuous and discrete Josephson transmission lines. In *Nonlinear Physics: Theory and Experiment*, edited by E. Alfinito, M. Boiti, L. Martina & F. Pempinelli, Singapore: World Scientific, pp. 582–589

Watanabe, S., van der Zant, H., Strogatz, S.H. & Orlando, T. 1996. *Physica* D, 97: 429

JOSEPHSON JUNCTIONS

Predicted theoretically by Brian Josephson in 1962, the Josephson junction (JJ) consists of two superconductors that are so weakly coupled that the Cooper pairs may quantum mechanically tunnel between the superconductors without destroying the integrity of their individual macroscopic wave functions, Ψ_1 and Ψ_2. A typical example is the trilayer tunnel junction (see Figure 1) consisting of two overlapping niobium (Nb) thin films separated by a thin (2–5 nm) insulating layer of aluminum oxide (Al_2O_3).

Extremely fast and highly nonlinear, the JJ is described by the two Josephson equations

$$I = I_c \sin\phi, \tag{1}$$

$$\frac{\partial\phi}{\partial t} = \frac{2\pi}{\Phi_0} V, \tag{2}$$

where I is the pair-current through the junction. The critical current I_c is a constant parameter given by the materials, the barrier, and the geometry of the structure. $\Phi_0 = h/2e$ is the flux quantum, $\phi = \theta_1 - \theta_2$ is the difference between the phases θ_1 and θ_2 of Ψ_1 and Ψ_2, respectively, and V is the voltage across the junction. These equations assume a constant current density over the junction area.

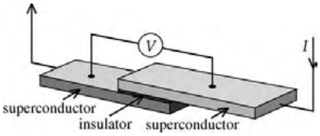

Figure 1. Typical Josephson tunnel junction consisting of two superconducting thin-films separated by a very thin insulating oxide layer.

If one applies a dc bias current $I < I_c$ to the junction, the phase difference ϕ of the macroscopic wave functions automatically adjusts itself so that Equation (1) is satisfied; i.e., the junction remains in the zero-voltage state carrying a supercurrent up to the critical current. For $I \geq I_c$, the junction switches to the voltage state, and ϕ oscillates in time according to the second Josephson equation.

If the junction is supplied with a constant voltage, V_{DC}, the phase difference increases steadily with time, and the junction current oscillates with the frequency

$$f = \frac{1}{\Phi_0} V_{DC}; \tag{3}$$

that is, the junction functions as a voltage controlled oscillator (VCO) that may generate microwave power into the gigahertz range. (The pre-factor $1/\Phi_0 \approx 0.5\,\text{GHz}/\mu\text{V}$.)

The capacitance between the two electrodes in Figure 1 shunts the Josephson tunneling and leads to hysteresis in the current-voltage characteristic (I–V curve) as shown in Figure 2a. Starting at zero and increasing the bias current, there is a vertical supercurrent (zero-voltage state) up to I_c where the junction switches (horizontally) to the steep so-called quasi-particle curve (near 2.7 mV), which reflects the superconducting gap of the two niobium electrodes. The quasi-particle curve is followed both when the bias current is further increased and when the current is returned to $I = 0$. The hysteretic I–V curve is point-symmetric around $(V, I) = (0, 0)$.

The I–V curve near the gap is strongly nonlinear and temperature dependent. Tunnel junctions biased close to the knee are used as low-noise bolometers to detect broad band signals in the millimeter and sub-mm range. Due to its strong nonlinearity, heterodyne receivers based on this SIS-mixer (superconductor-insulator-superconductor) can be operated near the quantum limit ($hf \approx k_B T$). The SIS-mixer may be pumped by the microwave signal emitted from a long JJ (see below). Most modern radio-telescopes employ SIS-mixers for spectral measurements in the frequency range 10–1000 GHz.

Figure 2b shows the I–V curve when a microwave signal is applied to the junction. The supercurrent is suppressed and small equidistantly spaced replica appear as vertical (Shapiro) steps with a voltage difference of $\Delta V = hf/2e$. These zero-crossing steps and the fact that voltage and frequency are related only through fundamental constants allow for the practical realization of the Josephson voltage primary standard. When pumped by a 70 GHz signal, $\Delta V \approx 140\,\mu\text{V}$; thus a small chip with more than 20.000 dc series connected JJs can generate a reference voltage of 10 V with an accuracy of 0.1 nV. Josephson junctions are strongly nonlinear elements which when pumped by external high-frequency signals may not only generate very high-numbered higher harmonics but also via parametric processes generate both even and odd subharmonics (e.g., period-doubling bifurcations) and chaos. The stability of the Josephson voltage standard is limited by chaotic behavior.

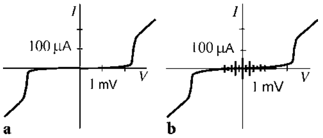

Figure 2. dc current–voltage characteristic of Nb/Al$_2$O$_3$/Nb trilayer Josephson tunnel junction at 4.2 K. Note the supercurrent in the zero-voltage state and the steep gap structure near 2.7 mV. (a) No applied microwaves and (b) with applied microwaves. The small vertical zero-crossing steps are equidistantly spaced with a voltage difference of $\Delta V = hf/2e$. These nonlinear quantum phenomena allow for the practical construction of the Josephson voltage primary standard.

JJs are highly sensitive to magnetic fields. The gradient of the phase difference ϕ is proportional to the magnetic field applied in the plane of the junction, and for constant current density the net critical current is zero each time ϕ has a twist of 2π, which happens when the flux Φ threading the junction is exactly one flux quantum Φ_0. Included in a superconducting ring, JJs constitute the so-called Superconducting Quantum Interference Device (SQUID), which is used, for example, for magnetoencephalographic measurements of electrical activity in the human brain.

The magnetic twist of the phase difference along the junction also leads to the so-called Fiske steps. These are nearly constant-voltage steps in the I–V curve at voltages $V_{FSn} \approx \Phi_0 c_s/2L$, where $n = 1, 2, \ldots$ and L is the junction length perpendicular to the magnetic field. Also, c_s is the Swihart velocity of electromagnetic waves propagating in the junction (about 3% of the light velocity in vacuum). The physical mechanism is as follows. The Josephson oscillations at the voltages V_{FSn} excite standing electromagnetic waves inside the junction (cavity) with wavelengths $\lambda_n = 2L/n$ and resonance frequencies $f_n = c_s/\lambda_n$. Whenever the spatial twist in $\Phi(x)$ fits to the nth mode of the standing wave, there is a strong resonant nonlinear interaction that phase-locks the Josephson oscillation at f_n, giving Fiske steps at voltages $V_{\mathrm{FS}n}$.

Until now we have considered only JJs with small dimensions compared with the so-called Josephson penetration length λ_{J}, which is of the order of 10 μm. Longer tunnel junctions are well modeled by the perturbed sine-Gordon (SG) equation. If we consider a linear one-dimensional (1-d, x-direction) junction with dc current bias, the SG equation in normalized units reads

$$\phi_{xx} - \phi_{tt} = \sin\phi + \alpha\phi_t - \eta, \tag{4}$$

where the normalized magnetic field $\kappa_{1,2}$ (perpendicular to the x-direction) enters as the boundary condition

$$\phi_x(0,t) = \kappa_1 \text{ and } \phi_x(l,t) = \kappa_2, \tag{5}$$

specifying the magnetic field at the two ends of the junction. The normalized bias current η here is assumed to be evenly distributed along the junction (overlap geometry). In the above equations, time t is normalized to the inverse maximum plasma frequency, ω_0, length x to λ_{J}, currents to the maximum critical current, I_{c}, and magnetic fields to the critical field, H_{c}, needed to force the first fluxon into the junction. Magnetic fields can only enter the junction in the form of fluxons, which are individual soliton-like localized 2π phase shifts each containing one flux quantum Φ_0.

Many solutions to the nonlinear SG equation have been found by numerical integration. For zero magnetic field and low values of the damping coefficient α, fluxons oscillate resonantly back and forth inside the junction driven by the Lorentz-like force from the bias current. This gives rise to a 4π phase shift per period and leads to the so-called zero-field steps, which are nearly constant-voltage steps located at voltages $V_{\mathrm{ZFS}n} \approx n\Phi_0 c_{\mathrm{S}} L$, $n = 1, 2, ..$ in the I–V curve. Note that zero-field step voltages are twice as large as the voltages of the Fiske steps discussed above for short junctions.

For every fluxon collision at $x = l$ (where l is the normalized junction length), a small electromagnetic field is emitted and the junction may be used as a microwave oscillator when biased on either a Fiske step or a zero-field step, but applications of this resonant soliton oscillator (RSO) are rather limited. First, the resonance frequencies are fixed (given by L and c_{s}) and the tunability is very small (steep steps). Second, the emitted power is relatively low and has high harmonic content (fluxon collision delta-function-like in time domain). However, due to its frequency stability and rather narrow linewidth, the RSO may be used as an on-chip clock oscillator.

When a stronger external magnetic field is applied (e.g., from a dc current, I_{CL}, in an overlaying control line), the junction will contain many fluxons, and since the fluxons repel each other, they will form an equidistant chain. Under a very strong field, the fluxons are forced so close to each other that the phase gradient becomes nearly uniform and proportional to the field strength. (One may compare the phase variation to a household corkscrew where the pitch is given by the applied magnetic field.) If a dc bias current I_{B} (uniformly distributed along the x-direction) is also applied to the junction, the fluxon chain is forced to move with constant velocity along the junction leading to a unidirectional flux flow through the junction. In our analogy, the bias current forces the corkscrew to rotate with constant angular velocity. The total phase shift per second and, thus, the frequency of the microwave signal emitted at $x = l$ can therefore be adjusted independently by the dc bias current (rotation) and the dc magnetic field (pitch). In the I–V curve, one observes a flux flow step (FFS) where the voltage, V_{FFS}, depends on both I_{B} and I_{CL}. Because the pitch for a dense chain of fluxons increases proportional to the magnetic field, V_{FFS}, the oscillator frequency in this limit depends linearly on I_{CL} for fixed I_{B}.

This explains the dynamics of the important flux flow oscillator (FFO). Its easy tunability not only permits wide band frequency coverage but also allows for accurate phase locking of the FFO. The power emitted from the FFO depends in a complicated way on the junction parameters as well as on I_{B} and I_{CL}, but with appropriate microwave design the power is sufficient to pump an SIS-mixer placed on the same chip. In 2002, a fully superconducting integrated receiver (SIR) with FFO, stripline circuit, SIS-mixer, and antenna (all placed on a $5 \times 5\,\mathrm{mm}^2$ chip) has been operated in phase-locked mode up to 712 GHz with a frequency resolution of less than 1 Hz relative to the reference oscillator used in the phase-locking loop. At 500 GHz the noise temperature was less than 100 K, just above the quantum limit. The low-noise in combination with the high-frequency resolution in the submillimeter frequency range is very promising for spectral investigations in astronomy, chemistry, and biophysics.

Numerical simulations of the SG equation have confirmed that the unidirectional flux flow mode can also be sustained for very large values of the damping parameter α. Often the damping per normalized length $\alpha l > \pi$ is used to define the flux flow range. For low values of αl the FFS consists of a distinct Fraunhöfer pattern of Fiske steps localized in the vicinity of the average FFS voltage given by I_{B} and I_{CL}. The correct boundary conditions, the geometry, and especially the paths along which the two bias currents are supplied to the junction are very important for the free-running linewidth of the FFO and thus for the practical implementation of the SIR.

For simplicity, we have restricted ourselves to a single Josephson junction with linear 1-d geometry. Considerable work has been devoted to 1- and 2-d arrays of both short and long junctions, and vertically stacked junctions (*See* **Josephson junction arrays**; **Long Josephson junctions**). Many structures and circuits have been fabricated in both low-T$_c$ and high-T$_c$ superconductors. The 1-d long Josephson junction with annular geometry is of considerable interest, not least because the cyclic boundary conditions are combined with flux quantization and flux trapping. Whole families of *single flux quantum* (SFQ) and *rapid single flux quantum* (RSFQ) electronics based on propagation of single fluxons in superconducting circuits containing Josephson junctions and SQUIDs (*See* **Superconducting quantum interference device**) have been developed and partially tested for applications. Josephson junction science and technology now is mature, and many applications are waiting to be implemented in future electronics.

JESPER MYGIND

See also **Diodes; Hysteresis; Josephson junction arrays; Long Josephson junctions; Parametric amplification; Period doubling; Sine-Gordon equation; Solitons; Superconducting quantum interference device; Superconductivity**

Further Reading

Duzer, T. Van & Turner, C.W. 1998. *Principles of Superconductive Devices and Circuits*, 2nd edition, Upper Saddle River, NJ: Prentice-Hall, pp. 158–325

Kadin, A.M. 1999. *Introduction to Superconducting Circuits*, New York: Wiley, pp. 178–340

Orlando, T.P. & Delin, K.A. 1990. *Foundations of Applied Superconductivity*, Reading, MA: Addison-Wesley, pp. 393–486

JOST FUNCTIONS

See **Inverse scattering method or transform**

JULIA SETS

See **Fractals**

JUMP PHENOMENA

Jump phenomena or surfaces of discontinuity occur when a field or gradient of a field exhibits finite jumps (discontinuities) due to the system changing nature (such as a liquid-gas interface) or the geometry of the domain changing abruptly in a given hyperplane (such as an acoustical tube with an abrupt widening). A particularly dramatic example of a hydraulic jump (see Figure 1) appeared regularly on the river Seine early in the 20th century.

In nature, systems are often governed by conservation laws of the various quantities. For example, for a fluid, we can write equations for the conservation of mass, momentum, and energy, and in an electrical system, current is conserved. There may be a hierarchy of these laws so when the system is perturbed the lower order ones are approximately maintained while the higher order ones are clearly broken.

Following the discussion of Landau (Landau & Lifschitz, 1959), one can classify discontinuities in two main groups. Let us illustrate this with the example of a continuous medium for which mass flux, momentum flux, and energy flux are continuous across a discontinuity. In the first case, mass flux is zero across the discontinuity, implying that the tangential derivative of the velocity has a jump. This is called a tangential discontinuity, a familiar example being the ocean surface separating water circulation from wind circulation. In the other case, called a shock wave (*See* **Shock Waves**), mass flow is nonzero, implying a jump in the velocity normal to the discontinuity. The existence of such discontinuous fields is strongly connected to the hyperbolic nature of the system of equations for which wave fronts can propagate along definite directions (characteristics).

In reality, the velocity normal to the interface is not discontinuous but varies on a very small scale due to viscosity and damping which are not taken into account in the original model. In particular, energy is dissipated in such an event so that the energy flux is not conserved. A way to remove this difficulty is to assume a given dissipation and modify the energy jump condition. It is remarkable that such modified jump conditions can describe accurately the state of a system where dissipation occurs even without an accurate description of the (microscopic) dissipation mechanism.

Examples

Example (i)

As an example of a normal discontinuity (shock wave), consider a one-dimensional compressible fluid flow. The conservation of mass and momentum of the fluid can be written as

$$\rho_t + (\rho v)_x = 0, \tag{1}$$

$$v_t + v v_x + \frac{c^2}{\rho} h_x = 0, \tag{2}$$

where the subscripts indicate partial derivatives, ρ is the gas density, v the velocity along the x axis, and $c(\rho)$ is the velocity of sound.

In 1860, Bernard Riemann obtained a solution of this system by assuming a single dependence of the two fields $\rho(\alpha(x,t))$ and $v(\alpha(x,t))$ on an unknown function α. He obtained the compatibility condition linking ρ and v:

$$v = \pm \int \frac{c(\rho)}{\rho} \mathrm{d}\rho, \tag{3}$$

and showed that α should satisfy the one-dimensional partial differential equation

$$\alpha_t + (v \pm c)\alpha_x = 0, \tag{4}$$

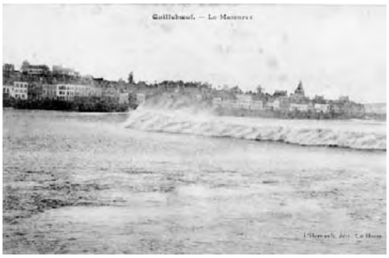

Figure 1. A tidal bore (or *mascaret*) on the Seine at Quillebeuf, France. The site is 19 km from the mouth of the river, where tides up to 10 m occur. Due to dredging in the 1960s, this striking nonlinear phenomenon has disappeared. (Courtesy of J.J. Malandain.)

whose formal solution is $\alpha = f(x - (v \pm c)t)$. This indicates that given values of v and ρ move to the right along the characteristic lines $x = (v \pm c)t$ at speeds $v \pm c$, allowing shock waves across which the velocity and density fields are discontinuous. Consequently, an initial condition with $v > c$ will only propagate to the right, whereas if $v < c$, it will propagate over the whole domain.

Consider how to derive jump conditions; from (1). Assume $c^2 = \rho g$ where g is a constant; this is not physical in the context of gases but it simplifies the derivation. The conservation of the mass flux is obtained by integrating the first equation on a small domain $-\varepsilon < x < \varepsilon$ centered on $x = 0$; thus,

$$\int_{-\varepsilon}^{\varepsilon} \rho_t \, \mathrm{d}x + [\rho v]_{x=-\varepsilon}^{x=\varepsilon} = 0. \qquad (5)$$

Assuming that ρ_t has a finite discontinuity at $x = 0$, we obtain in the limit $\varepsilon = 0$

$$[v\rho] = 0\,. \qquad (6)$$

The momentum conservation law is obtained by examining the time evolution of the average momentum ρv:

$$(\rho v)_t + \left(\rho v^2 + \frac{g\rho^2}{2}\right)_x = 0. \qquad (7)$$

The last conservation law involves the total energy e of the fluid, which is the sum of the kinetic and potential energy

$$(\rho v^2 + g\rho^2)_t + (\rho v^3 + 2\rho^2 v g)_x = 0. \qquad (8)$$

Proceeding in a similar way as for the mass flux, we obtain jump conditions for the momentum and energy of the fluid so the entire set of jump conditions are

$$[v\rho] = 0, \left[\rho v^2 + \frac{g\rho^2}{2}\right] = 0,$$
$$[v^3\rho + 2g\rho^2 v] = 0. \qquad (9)$$

Example (ii)

The second kind of discontinuity occurs when considering compression waves in an inhomogeneous beam made of two beams of different materials soldered at some point. The equation describing the displacement u is the one-dimensional generalized wave equation

$$\rho\sigma u_{tt} - (\sigma E u_x)_x = 0, \qquad (10)$$

where σ is the beam section, ρ its density, and E the Young modulus such that $E = E_\mathrm{l}$ for $x < 0$ and $E = E_\mathrm{r}$ for $x > 0$. To obtain the jump condition at $x = 0$, one integrates equation (10) over a small domain $-\varepsilon < x < \varepsilon$ to obtain

$$\int_{-\varepsilon}^{\varepsilon} \mathrm{d}x \rho\sigma u_{tt} - [\sigma E u_x]_{x=-\varepsilon}^{x=\varepsilon} = 0\,. \qquad (11)$$

Now consider the limit $\varepsilon = 0$, assuming the displacement u to be continuous across $x = 0$. The first term is the acceleration which should tend to zero. The second term is the stress, and relation (11) shows that it too should be continuous. At $x = 0$, we then have the following relations:

$$u_l = u_\mathrm{r}, \quad E_\mathrm{l} u_x|_\mathrm{l} = E_\mathrm{r} u_x|_\mathrm{r}, \qquad (12)$$

so the gradient u_x exhibits a finite jump.

Hydraulic Jumps

Shallow water flows for which the depth h_0 is much smaller than the length l are governed by the following equations for the water elevation h and velocity v in the direction of propagation (Landau & Lifschitz, 1959)

$$h_t + (hv)_x = 0, \qquad (13)$$
$$v_t + v v_x = -g h_x, \qquad (14)$$

which are the same as (1) if we write $h = \rho$ and assume $c^2 = \rho g = gh$.

Consider a so-called hydraulic jump solution occurring, for example, when a dam breaks. The conservation of the mass flux, momentum flux, and energy flux are given by (9). The first two conditions imply the following relations between the fields indexed l (left) or r (right) on each side of the jump.

$$h_\mathrm{l} v_\mathrm{l} = h_\mathrm{r} v_\mathrm{r} \equiv j, \quad h_\mathrm{l} v_\mathrm{l}^2 + \frac{g h_\mathrm{l}^2}{2} = h_\mathrm{r} v_\mathrm{r}^2 + \frac{g h_\mathrm{r}^2}{2}, \qquad (15)$$

where we introduced the (continuous) mass flux j. From these we deduce the velocities $v_{\mathrm{l,r}}$:

$$v_\mathrm{l}^2 = \frac{g}{2} h_\mathrm{r} \left(1 + \frac{h_\mathrm{r}}{h_\mathrm{l}}\right), \quad v_\mathrm{r}^2 = \frac{g}{2} h_\mathrm{l} \left(1 + \frac{h_\mathrm{l}}{h_\mathrm{r}}\right). \qquad (16)$$

To gain further insight consider the energy flux across the hydraulic jump. From the original equations (13) this should be conserved; however (as we discussed above), discontinuities are in general not physical, indicating that the model should be completed by some dissipation process. Without considering the details of this, we just write that energy per unit time (power) is being taken out of the system at the jump. Integrating as above, the energy conservation equation (8), from $-\varepsilon$ to ε and passing to the limit, gives the power dissipation at the jump,

$$w_{\mathrm{lr}} \equiv \lim_{\varepsilon \to 0} \int_{-\varepsilon}^{\varepsilon} e_t \, \mathrm{d}x = -[v^3 h + 2gh^2 v]_{\mathrm{l}}^{\mathrm{r}}. \qquad (17)$$

Using (16) we obtain the final result for the power dissipation at the jump,

$$w_{\mathrm{lr}} = jg \frac{(h_{\mathrm{l}} - h_{\mathrm{r}})^3}{2h_{\mathrm{l}} h_{\mathrm{r}}}. \qquad (18)$$

Assuming that energy is absorbed at the jump so $w_{\mathrm{lr}} < 0$, we get from (18) $h_{\mathrm{l}} < h_{\mathrm{r}}$, and from (16)

$$v_{\mathrm{l}}^2 = gh_{\mathrm{l}} \frac{h_{\mathrm{r}}}{h_{\mathrm{l}}} \frac{h_{\mathrm{r}} + h_{\mathrm{l}}}{2h_{\mathrm{l}}} > gh_{\mathrm{l}}$$

so that $v_{\mathrm{l}} > c_{\mathrm{l}}$, the local wave speed while $v_{\mathrm{r}} < c_{\mathrm{r}}$. These inequalities indicate the stability of the hydraulic jump.

Numerical Problems and Applications

As seen above, jump conditions come from conservation laws so an accurate numerical scheme should satisfy the conservation laws. Because the jump conditions are flux conditions, it is natural to use methods that involve integrating the operator over the spatial domain to satisfy them. One way to do this is the so-called finite volume approach, where the equation is integrated over reference volumes yielding naturally the flux conditions. Another approach is to use finite element methods, where the solution is decomposed on a basis of trial functions. The relation that gives the coefficients is obtained by integrating the equation over the whole domain.

For shock waves it is essential to have numerical schemes that do not artificially smooth the shock but respect the jump conditions. Here, special staggered finite difference schemes have been suggested by Lax, Wendroff, and Godunov.

As a final application, we present the case of a wave guide whose section changes abruptly at some point. This occurs in neurons for which varicosities are known to block the propagation of nerve impulses (Scott, 2003). To simplify the discussion, let us consider the following sine-Gordon nonlinear wave equation

$$\phi_{tt} - \phi_{xx} - \phi_{yy} + \sin(\phi) = 0, \qquad (19)$$

propagating in a two-dimensional T-shaped domain as shown in Figure 2 and assuming a zero normal gradient on the boundary. This describes the electrodynamics of a Josephson junction superconducting device obtained by inserting a small oxide layer between two metallic plates (Scott, 2003).

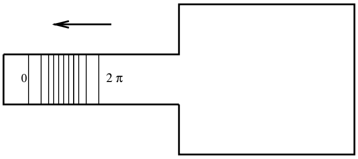

Figure 2. A T-shaped Josephson junction. The kink solution (21) is shown in the left region as contour lines. Kink propagation is easy from right to left but is inhibited in the other direction.

First note that the total energy of the system

$$E = \int \mathrm{d}x \, \mathrm{d}y \frac{1}{2} [\phi_t^2 + \phi_x^2 + \phi_y^2 + 2(1 - \cos(\phi))] \qquad (20)$$

is a constant in time. This can be seen by multiplying (19) by ϕ_t and integrating over the domain using the boundary conditions. Consider the propagation of a front (kink)

$$\phi(x, t) = 4 \arctan \left(\exp \left(\frac{x - vt}{\sqrt{1 - v^2}} \right) \right) \qquad (21)$$

which is an exact solution of the one-dimensional sine-Gordon equation for which ϕ is independent of y. Inserting (21) into (20) and calculating the integrals, it can be shown that the energy is

$$E_k = 8 \frac{w}{1 - v^2}, \qquad (22)$$

where w is the width in the y direction of the domain.

Let us now consider the propagation of such a front in the domain and assume it starts at $t = 0$ from the right-hand side with a negative velocity so that it will encounter the discontinuity. First, notice that because of the boundary conditions, it travels unaffected in the straight section. When it hits the discontinuity, it will cross into the small section with no apparent change in shape apart from reflected waves (see Figure 2). On the contrary, if the kink is started from the left side whose section is narrow, it will get reflected by the discontinuity unless its velocity is quite large.

The explanation of these observations lies in the conservation of the energy of the system, assuming that most of this energy is in the front. For a kink to exist in a domain of lateral extension w, its energy should be greater than $E_k(v = 0) = 8w$. When the kink is started in the right domain, its energy $E/8 = w_{\mathrm{r}}/(1 - v^2) > w_{\mathrm{r}} > w_{\mathrm{l}}$, so we expect it to cross without trouble into the smaller section. Indeed it should gain speed, and it does. On the contrary, if the kink is started in the left domain, its speed v_{l} must be very large (so that $w_{\mathrm{l}}/(1 - v_{\mathrm{l}}^2) > w_{\mathrm{r}}$ where $8w_{\mathrm{r}}$ is the energy of a static kink in the region on the right) in order to cross into the region on the right. This provides a means to realize a rectifying gate, through which signals pass in only one direction.

JEAN GUY CAPUTO

See also **Constants of motion and conservation laws; Long Josephson junctions; Shock waves**

Further Reading

Ames, W. 1992. *Numerical Methods for Partial Differential Equations*, 3rd edition, New York: Academic Press

Caputo, J.G. & Stepanyants, Y. 2003. Bore formation, evolution and disintegration into solitons in shallow inhomogeneous channels. *Nonlinear Processes in Geophysics*, 10: 407–424

Landau, L. & Lifchitz, E. 1959. *Fluid Mechanics*, vol. 6 of *Course in Theoretical Physics*, London: Pergamon Press and Reading, MA: Addison-Wesley

LeVeque, R. 2002. *Finite Volume Methods for Hyperbolic Problems*, Cambridge and New York: Cambridge University Press

Lin, C.C., Segel, L.A. & Handelman, G.H. 1995. *Mathematics Applied to Deterministic Problems in the Natural Sciences*, 4th edition, Philadelphia: SIAM

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

JUPITER'S GREAT RED SPOT

Jupiter's Great Red Spot is a large swirling cloud mass of reddish-brown appearance (see figure in color plate section). Situated in Jupiter's southern hemisphere, it straddles the south tropical zone and, to the north of this, the south equatorial belt. The Great Red Spot (GRS) is roughly elliptical in shape, with the semi-major axis zonally aligned (east-west) and with dimensions approximately 22, 000 km (twice the diameter of the Earth) by 11, 000 km. The atmospheric motions associated with the GRS are visible in the cloud layer near the tropopause. It is generally agreed to be a vortex (Mitchell et al., 1981); and Smith et al. (1979a) give an estimate of the vorticity. This vortex is anticyclonic (rotating in the opposite sense to that induced by the planetary rotation), that is, anticlockwise, but with a weakly counter-rotating, or possibly quiescent inner region. The GRS is at high pressure and low temperature relative to its surroundings. A striking feature associated with the GRS is the turbulent oscillating cloud system to the northwest.

In 1664, Robert Hooke observed a small dark spot in one of the southern belts of Jupiter, reported in Hooke (1665). This is considered by some contemporary authors to be an early manifestation of the GRS (see also Rogers, 1995). By observation of this feature, Hooke was able to make a good estimate of Jupiter's rate of rotation. Many further observations of this and other features of Jupiter were made by Hooke and by Giovanni Cassini over the next five years (e.g., Cassini, 1672), and intermittently by others over the next 50 years (see Denning, 1899). However, it was not until 1831 that Samuel H. Schwabe identified what can be clearly recognized as a feature called the Hollow, in his drawing reproduced in Rogers (1995). As the Hollow is clearly apparent above the GRS in contemporary observations, this is taken to imply the presence of the GRS. Continuous observations began in 1872 (Peek, 1958) and from these it appears that the GRS was considerably larger in the late 1800s than it is now (Beebe & Youngblood, 1979; Rogers, 1995). Very much more detailed observations became available once space probes reached Jupiter; first was Pioneer 10 in 1973, then Pioneer 11 in 1974, Voyagers 1 and 2 in 1979, Galileo in 1996, and then most recently Cassini from October 1, 2000 through until March 22, 2001. The Voyager and Galileo missions (Smith et al., 1979a, b; Vasavada et al., 1998) provided very high resolution still images, while Cassini (Porco et al., 2003) gave more dynamic information. From 1994, after corrections were made to faulty optics by the Endeavour crew, the Hubble telescope has been providing good images of Jupiter and the GRS. These, along with the Cassini sequences, clearly show vortex interactions between the GRS and smaller, intermittent vortices, thus allowing a much better understanding of the underlying dynamical processes (Morales-Juberías et al., 2002).

Attempts have been made since 1961 to model the GRS and to understand the processes giving rise to the motions observed, beginning with Hide (1961), who suggested that the GRS was the visible manifestation of a Taylor column. The quantity of relevance to these calculations is the potential vorticity q. As the atmosphere of Jupiter is stably stratified and there are significant rotational effects at the scale of the GRS, the most commonly used model is the quasi-geostrophic approximation (a simplification of the complete atmospheric dynamic equations, giving a balance of horizontal pressure gradient forces and horizontal Coriolis forces due to Jupiter's rotation). Here q is defined as

$$q = f + \nabla^2\psi + \frac{1}{\rho_0}\frac{\partial}{\partial z}\left(\rho_0 \frac{f_0^2}{N^2}\frac{\partial\psi}{\partial z}\right), \qquad (1)$$

where $\psi(x, y, z, t)$ is the stream function for the flow and (x, y, z) are local cartesian coordinates, with x and y being the zonal and meridional directions, respectively, and z being a measure of the depth. The basic state density profile is $\rho_0(z)$ and the effects of Jupiter's rotation are introduced by the beta-plane approximation, the assumption that we can treat part of Jupiter as a plane, over which the Coriolis parameter f varies linearly: $f = f_0 + \beta y$, where f_0 is the reference value at the latitude of the GRS and β is the ambient potential vorticity gradient. The buoyancy frequency $N(z)$, the frequency with which a parcel of fluid displaced from equilibrium will oscillate, is determined by the vertical temperature variation. The potential vorticity is materially conserved, according to

$$\frac{\partial q}{\partial t} + \boldsymbol{v} \cdot \nabla q = 0, \tag{2}$$

where the horizontal velocity $\boldsymbol{v} = (-\partial\psi/\partial y, \partial\psi/\partial x)$.

Many researchers have modeled the vortical motions associated with the GRS by assuming a thin upper layer of constant density (of order 100 km; Dowling & Ingersoll, 1989) above a much deeper lower layer of slightly higher density in which there is a uniform steady zonal flow, for example, Ingersoll and Cuong (1981), Marcus (1988), and Marcus et al. (1990). The surface of the upper layer is above the visible cloud level. These models assume barotropic motions, that is, uniform in depth with no horizontal component of vorticity, in the upper layer where isolated vortices are assumed to be generated by mechanisms such as zonal shear flow instability. Merger of vortices can lead to the formation of larger vortices, and the relevant issues are then the stability and lifetime of isolated vortices such as the GRS in the observed zonal shear. A statistical mechanics approach to such equilibria is provided in Michel and Robert (1994).

There have also been discussions on the role of vertical density stratification, allowing baroclinic instability, but generally limited to the quasi-geostrophic case of (1) and (2). For example, Achterberg and Ingersoll (1994) studied the barotropic and the first two baroclinic modes, and Cho et al. (2001) used as many as 60 vertical modes. The high horizontal resolution used in Cho et al. (2001) allowed the authors to generate features such as the counter-rotating core (Vasavada et al., 1998, Figure 7, reporting Galileo data and similar Voyager data), but as in all other studies to date, the turbulent cloud system to the northwest of the spot does not appear.

A deficiency of the quasigeostrophic model is that vortices with radius significantly greater than the Rossby radius of deformation, L_{R}, tend to decay by radiation of Rossby waves in the absence of any external forcing. (L_{R} is a measure of the scale at which Coriolis forces become significant.) This would imply a lifetime for the GRS much shorter than the observed period, which is certainly greater than 150 years. In addition, the quasigeostrophic equations have the disadvantage that neither anticyclonic nor cyclonic vortices are preferentially created or destroyed, whereas observations show that the GRS is anticyclonic, as are 90% of the other (typically much smaller) vortices observed on Jupiter (Mac Low & Ingersoll, 1986).

For two-layer models, the zonal flows in the bottom layer can be incorporated in a reduced gravity single-layer shallow water model with meridionally varying topography and a free surface, which is taken to be above the visible cloud layer (Dowling & Ingersoll, 1989). There are various levels of approximation to the shallow-water equations beyond the quasi-geostrophic approximation. The intermediate geostrophic (IG) approximation (Williams & Yamagata, 1984; Nezlin & Sutyrin, 1994) arises when there is significant nonlinearity giving rise to finite free-surface distortion, as is to be expected for vortices with a scale much greater than the Rossby radius, such as the GRS. A more complete discussion of the asymptotic treatment of the nonlinear terms, along with higher-order approximations, is given in Stegner and Zeitlin (1996).

An alternative approach to understanding the GRS as a solitary wave was first introduced by Maxworthy and Redekopp (1976a,b) and Redekopp (1977). They considered the quasigeostrophic equations (1) and (2) and were able to find elementary soliton solutions in the limit of small amplitude disturbance, which obeyed either the Korteweg–de Vries equation (KdV) or the modified KdV equation. These balance zonal shear and dispersion to give isolated disturbances that could be regarded as vortices. However, as noted by several authors (e.g., Ingersoll & Cuong, 1981), these soliton solutions of the KdV equation pass through one another unchanged in form (elastic collisions). This is contrary to the behavior of vortices generally, and more particularly the GRS, where merger with smaller vortices is observed (Morales-Juberías et al., 2002). However, Petviashvili (1980), by going beyond quasigeostrophy, was able to derive solitary axisymmetric vortex solutions, without an imposed zonal shear, for a curved shallow layer, with anticyclones living longer. Nezlin et al. (1990) refer to this anticyclonic vortex as a Rossby soliton, with weak nonlinearity balanced by Rossby wave dispersion. These solitary solutions undergo inelastic collisions and were later thought to be found experimentally in parabolic water tank experiments (Nezlin et al., 1990), although Nycander (1993) and Stegner and Zeitlin (1996) demonstrated that this weakly nonlinear balance could not in fact exist in the parabolic geometry and that stronger nonlinearity was required to explain the laboratory observations. Stegner and Zeitlin (1996) go on to show that axisymmetric soliton-like solutions of a 2-d KdV equation are found for spherical geometry, corresponding to the surface of Jupiter, using a nonlinear model. Furthermore, only an anticyclonic vortex can be supported in this case, which is in accord with the observed rotation of the GRS. However, this cannot explain the preferential formation of anticyclonic vortices smaller than the Rossby radius, where the quasigeostrophic approximation remains appropriate. Finally, as explained by Stegner and Zeitlin (1996), their 2-d KdV solutions do not apply directly to the GRS as neither the elliptical shape nor the counter-rotating inner core can be accounted for.

See also **Atmospheric and ocean sciences; Fluid dynamics; Korteweg–de Vries equation; Solitons**

Further Reading

Achterberg, R.K. & Ingersoll, A.P. 1994. Numerical simulation of baroclinic Jovian vortices. *Journal of the Atmospheric Sciences*, 51(4): 541–562

Beebe, R.F. & Youngblood, L.A. 1979. Pre-Voyager velocities, accelerations and shrinkage rates of Jovian cloud features. *Nature*, 280: 771–772

Cassini, G. 1672. A relation of the return of a great permanent spot in the planet Jupiter, observed by Signor Cassini, one of the Royal Parisian Academy of the Sciences. *Philosophical Transactions of the Royal Society of London*, 7: 4039–4042

Cho, J. Y-K., de la Torre Juárez, M., Ingersoll, A.P. & Dritschel, D.G. 2001. A high-resolution, three-dimensional model of Jupiter's Great Red Spot. *Journal of Geophysical Research—Planets*, 106(E3): 5099–5105

Denning, W.F. 1899. Early history of the Great Red Spot on Jupiter. *Monthly Notices of the Royal Astronomical Society*, 59: 574–584

Dowling, T.E. & Ingersoll, A.P. 1989. Jupiter's Great Red Spot as a shallow water system. *Journal of the Atmospheric Sciences*, 46(21): 3256–3278

Hide, R. 1961. Origin of Jupiter's Great Red Spot. *Nature*, 190: 895–896

Hooke, R. 1665. A spot in one of the belts of Jupiter. *Philosophical Transactions of the Royal Society of London*, 1: 3

Ingersoll, A.P & Cuong, P.G. 1981. Numerical model of long-lived Jovian vortices. *Journal of the Atmospheric Sciences*, 38(10): 2067–2076

Mac Low, M.-M. & Ingersoll, A.P. 1986. Merging of vortices in the atmosphere of Jupiter: an analysis of Voyager images. *Icarus*, 65: 353–369

Marcus, P.S. 1988. Numerical simulation of Jupiter's Great Red Spot. *Nature*, 331: 693–696

Marcus, P.S., Sommeria, J., Meyers, S.D. & Swinney, H.L. 1990. Models of the Great Red Spot. *Nature*, 343: 517–518

Maxworthy, T. & Redekopp, L.G. 1976a. A solitary wave theory of the Great Red Spot and other observed features in the Jovian atmosphere. *Icarus*, 29: 261–271

Maxworthy, T. & Redekopp, L.G. 1976b. New theory of the Great Red Spot from solitary waves in the Jovian atmosphere. *Nature*, 260: 509–511

Michel, J. & Robert, R. 1994. Statistical mechanical theory of the Great Red Spot of Jupiter. *Journal of Statistical Physics*, 77: 645–666

Mitchell, J.L., Beebe, R.F, Ingersoll, A.P. & Garneau, G.W. 1981. Flow fields within Jupiter's Great Spot and White Oval BC. *Journal of Geophysical Research—Space Physics*, 86(NA10): 8751–8757

Morales-Juberías, R., Sánchez-Lavega, A., Lecacheux, J. & Colas, F. 2002. A comparative study of Jovian anticyclone properties from a six-year (1994–2000) survey. *Icarus*, 157(1): 76–90

Nezlin, M.V., Rylov, A. Yu., Trubnikov, A.S. & Khutoreski, A.V. 1990. Cyclonic–anticyclonic asymmetry and a new soliton concept for Rossby vortices in the laboratory, oceans and the atmospheres of giant planets. *Geophysical and Astrophysical Fluid Dynamics*, 52: 211–247

Nezlin, M.V. & Sutyrin, G.G. 1994. Problems of simulation of large, long-lived vortices in the atmospheres of the giant planets (Jupiter, Saturn, Neptune). *Surveys in Geophysics*, 15: 63–99

Nycander, J. 1993. The difference between monopole vortices in planetary flows and laboratory experiments. *Journal of Fluid Mechanics*, 254: 561–577

Peek, B.M. 1958. *The Planet Jupiter*, London: Faber & Faber

Petviashvili, V.I. 1980. Red spot of Jupiter and the drift soliton in a plasma. *JETP Letters*, 32(11): 619–622

Porco, C.C., West, R.A., McEwen, A., et al. 2003. Cassini imaging of Jupiter's atmosphere, satellites and rings. *Science*, 299: 1541–1547

Rogers, J.H. 1995. *The Giant Planet Jupiter*, Cambridge and New York: University Press

Redekopp, L.G. 1977. On the theory of solitary Rossby waves. *Journal of Fluid Mechanics*, 82(4): 725–745

Smith, B.A., Soderblom, L.A., Johnson, T.V., et al. 1979a. The Jupiter system through the eyes of Voyager 1. *Science*, 204: 951–957 and 960–972

Smith, B.A., Soderblom, L.A., Beebe, R., et al. 1979b. The Galilean satellites and Jupiter: Voyager 2 imaging science results. *Science*, 206: 927–950

Stegner, A. & Zeitlin, V. 1996. Asymptotic expansions and monopolar solitary Rossby vortices in barotropic and two-layer models. *Geophysical and Astrophysical Fluid Dynamics*, 83: 159–194

Vasavada, A.R., Ingersoll, A.P., Banfield, D., et al. 1998. Galileo imaging of Jupiter's atmosphere: The Great Red Spot, equatorial region, and White Ovals. *Icarus*, 135(1): 265–275

Williams, G.P. & Yamagata, T. 1984. Geostrophic regimes, intermediate solitary vortices and Jovian eddies. *Journal of the Atmospheric Sciences*, 41(4): 453–478

K

KADOMTSEV–PETVIASHVILI EQUATION

The Kadomtsev–Petviashvili (KP) equation was derived by Kadomtsev & Petviashvili (1970) to examine the stability of the one-soliton solution of the Korteweg–de Vries (KdV) equation under transverse perturbations, and it is relevant for most applications in which the KdV equation arises. After rescaling of its coefficients, the equation takes the form

$$(-4u_t + 6uu_x + u_{xxx})_x + 3\sigma^2 u_{yy} = 0, \qquad (1)$$

where indices denote differentiation, and σ is a constant parameter. If $\sigma^2 = -1$ $(+1)$, the equation is referred to as the KP1 (KP2) equation. All other real values of σ^2 can be rescaled to one of these two cases. In what follows, a reference to KP (as opposed to KP1 or KP2) implies that the result in question is independent of the sign of σ^2.

Depending on the physical context, an asymptotic derivation can result in either the KP1 or the KP2 equation. In all such derivations, the equation describes the dynamics of weakly dispersive, nonlinear waves whose wavelength is long compared with its amplitude, and whose variations in the second space dimension (rescaled y) are slow compared with their variations in the main direction of propagation (rescaled x). Two examples are as follows:

- *Surface waves in shallow water*: In this case, u is a rescaled wave amplitude with a rescaled velocity. The wavelength is long compared with the depth of the water h, which is large compared with the wave amplitude. The sign of σ^2 is determined by the magnitude of the coefficient of surface tension τ. KP1 results for large surface tension $\tau/(gh^2) > 1/3$ (thin films); otherwise, KP2 results. Here g is the acceleration of gravity. For most applications in shallow water, surface tension plays a sufficiently unimportant role, and KP2 is the relevant equation (Ablowitz &, 1979, *See* **Water waves**).
- *Magneto-elastic waves in antiferromagnetic materials*: Here u is a rescaled strain tensor with a rescaled velocity. The sign of σ^2 is determined by the difference between the linear velocities of the magnons and phonons, and the strength and direction of the external magnetic field (Turitsyn & Fal'kovich, 1986).

The KP equation has different classes of soliton solutions. The first class is a generalization of the solitons of the KdV equation. These solutions decay exponentially as $x, y \to \pm\infty$, in all but a finite number of directions along which they approach a constant value. For this reason these solutions are referred to as line solitons. By appropriately choosing their parameters, the direction of propagation of each line soliton can be chosen to be anything but the y-direction. In the simplest case, the solitons all propagate in the x-direction, adding a second dimension to the KdV solitons. Many other scenarios are possible. Two line solitons can interact with different types of interaction regions to produce two line solitons, but two line solitons can also merge to produce a single line soliton. Alternatively, a single line soliton can disintegrate in to two line solitons. The production or annihilation of a line soliton is sometimes referred to as soliton resonance. Although both KP1 and KP2 have line soliton solutions, soliton resonance occurs only for the KP2 equation. Line soliton solutions of the KP2 equation are stable, whereas line soliton solutions of the KP1 equation are unstable. More possibilities exist when more than two line solitons are involved. Two distinct line soliton interactions are illustrated in Figure 1.

Another class of soliton solutions exists for only the KP1 equation and decays algebraically in all directions as $\sqrt{x^2 + y^2} \to \infty$. These soliton solutions are referred to as lumps and are unstable. Individual lumps in multi-lump solutions do interact with each other but leave no trace of this interaction. A lump soliton is shown in Figure 2 (Ablowitz & Segur, 1981).

Another important class of solutions of the KP equation generalizes the exact periodic and quasi-periodic solutions of the KdV equation. A (quasi-) periodic solution with g phases is expressed in terms of

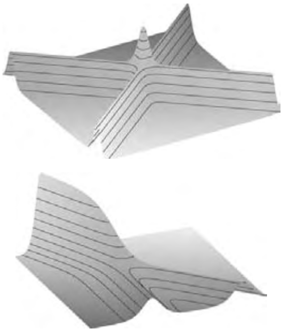

Figure 1. Two types of spatial ($t = 0$) line soliton interactions for the KP2 equation: (a) Two identical line solitons with an interaction that does not change their characteristics; (b) Two line solitons merge to produce one line soliton.

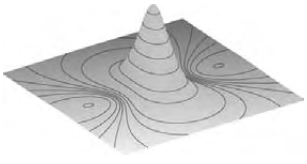

Figure 2. A lump soliton ($t = 0$) solution of the KP1 equation.

the Riemann theta function $\theta(z|\boldsymbol{B})$ by

$$u(x, y, t) = c + 2\frac{\partial^2}{\partial x^2}\ln\theta(\boldsymbol{k}x + \boldsymbol{l}y + \boldsymbol{\omega}t + \boldsymbol{\phi}|\boldsymbol{B}). \quad (2)$$

Here, c is a constant, and $\boldsymbol{k}$, $\boldsymbol{l}$, $\boldsymbol{\omega}$, and $\boldsymbol{\phi}$ are g-dimensional vectors that are interpreted as wave vectors ($\boldsymbol{k}$, $\boldsymbol{l}$), frequencies ($\boldsymbol{\omega}$), and phases ($\boldsymbol{\phi}$). These parameters and the $g \times g$ Riemann matrix $\boldsymbol{B}$ are determined by a genus g compact connected Riemann surface and a set of g points on it.

For $g = 1$, solution (2) generalizes the cnoidal-wave solution of the KdV equation to two spatial dimensions. For $g = 2$, the solution is still periodic in space. Its basic period cell is a hexagon, which tiles the (x, y)-plane. These solutions translate along a direction in the (x, y)-plane. For $g \geq 3$, solution (2) is typically no longer periodic or translating in time. For some values of their parameters, these (quasi-)periodic solutions can be interpreted as infinite nonlinear superpositions of line solitons. Solutions with $g \leq 2$ have been compared with experiments in shallow water, with agreement being more than satisfactory (Hammack et al., 1995). A two-phase solution of the KP2 equation is shown in Figure 3. A good review of the finite-phase solutions of the KP equation is given by Dubrovin (1981).

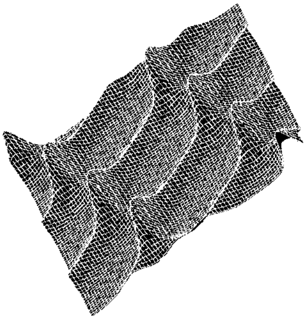

Figure 3. A two-phase periodic solution of the KP2 equation.

Unlike for the KdV equation, where only a restricted class of Riemann surfaces arises, any compact connected Riemann surface gives rise to a set of solutions of the KP equation. The reverse statement is also true: if (2) is a solution of the KP equation, then matrix $\boldsymbol{B}$ is the normalized period matrix of a genus g Riemann surface. This statement is due to Novikov. It provides a solution to the century-old Schottky problem, and its proof is by Shiota (1986).

The KP equation is the compatibility condition $\Psi_{yt} = \Psi_{ty}$ of the two linear equations

$$\begin{aligned} \sigma\Psi_y &= \Psi_{xx} + u\Psi, \\ \Psi_t &= \Psi_{xxx} + \tfrac{3}{2}u\Psi_x + \tfrac{3}{4}(u_x + w)\Psi, \end{aligned} \quad (3)$$

with $w_x = \sigma u_y$. These equations constitute the Lax pair of the KP equation. Using the inverse scattering method, it is the starting point for the solution of the initial-value problem for the KP equation on the whole (x, y)-plane with initial conditions that decay at infinity. The inclusion of line solitons is also possible (Ablowitz & Clarkson, 1991). Although the initial-value problem with periodic boundary conditions for KP2 was solved by Krichever (1989), this approach was unable to solve the same problem for the KP1 equation. In this context, solutions (2) are referred to as finite-gap solutions, as they give rise to operators (3) with spectra that have a finite number of forbidden gaps in them.

More details and different aspects of the theory of the KP equation are found in Ablowitz & Clarkson (1991), Ablowitz & Segur (1981), Dubrovin (1981), Krichever (1989), Shiota (1986), and references therein.

BERNARD DECONINCK

See also **Inverse scattering method or transform; Korteweg–de Vries equation; Multidimensional solitons; Plasma soliton experiments; Theta functions; Water waves**

Further Reading

Ablowitz, M.J. & Clarkson, P.A. 1991. *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, Cambridge and New York: Cambridge University Press

Ablowitz, M.J. & Segur, H. 1979. On the evolution of packets of water waves. *Journal of Fluid Mechanics*, 92: 691–715

Ablowitz, M.J. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM

Dubrovin, B.A. 1981. Theta functions and non-linear equations. *Russian Mathematical Surveys*, 36: 11–92

Hammack, J., McCallister, D., Scheffner, N. & Segur, H. 1995. Two-dimensional periodic waves in shallow water. Part 2. Asymmetric waves. *Journal of Fluid Mechanics*, 285: 95–122

Kadomtsev, B.B. & Petviashvili, V.I. 1970. On the stability of solitary waves in weakly dispersive media. *Soviet Physics Doklady*, 15: 539–541

Krichever, I.M. 1989. Spectral theory of two-dimensional periodic operators and its applications. *Russian Mathematical Surveys*, 44: 145–225

Shiota, T. 1986. Characterization of Jacobian varieties in terms of soliton equations. *Inventiones Mathematicae*, 83: 333–382

Turitsyn, S.K. & Fal'kovich, G.E. 1986. Stability of magneto-elastic solitons and self-focusing of sound in antiferromagnets. *Soviet Physics JETP*, 62: 146–152

KELVIN–HELMHOLTZ INSTABILITY

The Kelvin–Helmholtz (KH) instability was originally investigated by Lord Kelvin (William Thomson) and Hermann von Helmholtz (Kelvin, 1871; Helmholtz, 1868; Chandrasekhar, 1981; Drazin & Reid, 1981; Gerwin, 1968) in the context of wind flowing over water. Their main motivation was to understand how wind ruffles the ocean surface to produce the ripples, eddies, and whitecaps that we see on our trips to the beach.

The KH instability has subsequently been observed in a variety of space, astrophysical, and geophysical settings involving sheared fluid and plasma flows. In addition to the original situations relating to ocean waves, relevant configurations include the interface between the solar wind and the upper magnetosphere, coronal streamers moving through the solar wind, the boundaries between adjacent sectors of the solar wind, the structure of the tails of comets, and the boundaries of the jets propagating from the core of extragalactic double radio sources into their lobes. Another interesting application of KH instability in this age of unexplained air disasters is the phenomenon known as clear air turbulence, where an aircraft flying through a clear sky is suddenly buffeted (*See* **Clear air turbulence**).

With an initial or equilibrium configuration consisting of two superposed incompressible and inviscid fluids streaming in the horizontal or z-direction at differential speeds U_1 and U_2 (see Figure 1), consider how disturbances to this configuration evolve in time. Because the velocity differential occurs at different heights (or the transition in the speed from U_1 to U_2 occurs over some finite range of x), the equilibrium configuration has finite geometrical extent in the x-direction. Note that the inviscid assumption, which holds well for air and water as well as for very low-density interstellar media, allows an arbitrary profile for the velocity transition from U_1 to U_2 with the x transition length chosen to fit experiments. By contrast, the equilibrium is of semi-infinite extent in the horizontal or z-direction, and of infinite extent in y. Thus, one may represent field variables as Fourier series in the directions y and z which are of infinite extent, thereby assuming elementary disturbances of the form

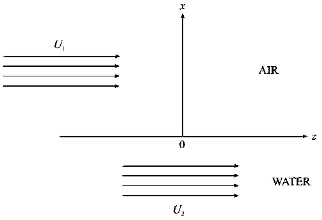

Figure 1. The basic configuration for shear flow across a vortex sheet.

$$\chi \sim f(x)\exp[\mathrm{i}(k_y y + k_z z - \omega t)] \qquad (1)$$

for all variables χ. Here, k_y and k_z are the y and z wave numbers, and ω is the angular frequency. (Note also that only a single mode among the infinite number needed to expand the variable χ as a Fourier series has been displayed in this equation.)

Substitution of Equation (1) into the equations governing the fluid flow yields a functional relationship between the angular frequency and the wave number called the dispersion relation, which provides key information regarding the propagation of the mode of Equation (1). Importantly, there are situations where the dispersion relation yields complex angular frequencies, that is,

$$\omega = \omega_r + \mathrm{i}\omega_i\,. \qquad (2)$$

The temporal exponential growth or decay of the mode is then controlled by ω_i (called the growth rate), while its oscillation frequency is governed by ω_r. Important for us are the unstable cases where $\omega_i > 0$ so the disturbance grows in time. In such cases, χ does not settle back to the original equilibrium, and one must consider the consequences of this growth.

A driving parameter for KH instability is the differential flow speed (velocity shear) of the two layers of fluid, and it suffices to consider one Fourier mode while investigating the stability of a system. The key concept here is that the wave number of the normal mode is left arbitrary, and the system is unstable if $\omega_i > 0$ for any wave number. In general, the system is stable for some range of the driving parameter and becomes unstable beyond a critical value of the parameter. The wave number that first becomes unstable as this critical parameter value is reached selects the most unstable mode, which usually guides the further development of the instability.

Let us apply these concepts by substituting Equation (1) into the dynamical equations governing the flow. In the simplest form of the analysis (linear theory), one neglects all nonlinear terms (where any of the field variables are raised to higher than linear powers, or where there are products of different field variables) to obtain a dispersion relation. This relation satisfies boundary conditions at the discontinuity corresponding to the continuity of the normal (perpendicular) speed, the normal magnetic field (if any), and the normal total (fluid plus electromagnetic) stress. The resulting dispersion relation in the incompressible fluid case is given by (Chandrasekhar, 1981)

$$\begin{aligned}\omega = {} & k_z(\alpha_1 U_1 + \alpha_2 U_2) \\ & \pm \left[gk \left\{ (\alpha_1 - \alpha_2) + \frac{k^2 T}{g(\rho_1 + \rho_2)} \right\} \right. \\ & \left. - k_z^2 \alpha_1 \alpha_2 (U_1 - U_2)^2 \right]^{1/2} .\end{aligned}$$

Here, g is the acceleration of gravity, T is the coefficient of surface tension, and

$$\alpha_1 = \frac{\rho_1}{\rho_1 + \rho_2}, \quad \alpha_2 = \frac{\rho_2}{\rho_1 + \rho_2}, \tag{3}$$

where ρ_i is the density of the i-th fluid. Analysis of Equation (3) reveals instability ($\omega_i > 0$) for (Kelvin, 1871; Chandrasekhar, 1981)

$$(U_1 - U_2)^2 > \frac{2}{\alpha_1 \alpha_2} \sqrt{\frac{Tg(\alpha_1 - \alpha_2)}{\rho_1 + \rho_2}} . \tag{4}$$

For air over sea water, this implies instability at $U_1 - U_2 > 650$ cm/s (14.8 mph) manifested as surface waves of about 1.71 cm wavelength.

The sea, however, is ruffled by winds of smaller speeds than this, and a variety of effects have been advanced to explain this discrepancy. These include continuous velocity variation rather than an abrupt jump at the surface, continuously varying density profiles, rotational effects (due to the Earth's rotation), as well as Lord Kelvin's original suggestion—the effects of viscosity (Chandrasekhar, 1981; Drazin & Reid, 1981).

In addition, compressibility and various electromagnetic fields (Choudhury, 1986; Choudhury & Lovelace, 1984; Miura, 1990; Miura & Pritchett, 1982) become important in magnetospheric, geophysical, and astrophysical settings where there are high-speed shear flows and large fields. An important effect of compressibility is to introduce a second unstable traveling-wave mode with nonzero ω_r, in addition to the standing-wave unstable mode (with zero ω_r) which exists in the incompressible limit.

There have been numerous computer simulation studies of KH instability in a variety of settings (Miura, 1995; Normal et al., 1982). The effect of anisotropic pressure has also been comprehensively considered (Brown & Choudhury, 2000).

Finally, it is important to consider nonlinear effects on the instability since the temporal growth of the linear fields in Equation (1) will eventually reach amplitudes large enough so that product terms can no longer be ignored. From physical considerations, one may also argue that exponential linear growth as predicted by Equation (1) for $\omega_i > 0$ is not tenable indefinitely because the energy is proportional to the square of the field amplitude and this energy is limited. Clearly, nonlinear effects must come into play and limit the linear growth (Debnath & Choudhury, 1997; Drazin & Reid, 1981; Weissman, 1979). Such nonlinear evolution studies reveal a variety of interesting dynamical behaviors from quasi-periodicity and chaos at one extreme (Choudhury & Brown, 2001) to self-organization and coherent structures at the other (Miura, 1999). A formulation of the problem in terms of noncanonical Hamiltonians for the fluid dynamical equations (Benjamin & Bridges, 1997) enables one to follow instability evolution well into its late stages where the interface has become appreciably folded.

S. Roy Choudhury

See also **Clear air turbulence; Coherence phenomena; Dispersion relations; Shear flow; Wave stability and instability; Water waves**

Further Reading

Batchelor, G. K., Moffatt, H. F. & Worster, M. G. (editors) 2000. *Perspectives in Fluid Dynamics*, Cambridge and New York: Cambridge University Press

Benjamin, T. B. & Bridges, T.J. 1997. Reappraisal of the Kelvin–Helmholtz problem I & II. *Journal of Fluid Mechanics*, 333: 301, 327

Brown, K.G. & Choudhury, S.R. 2000. Kelvin–Helmholtz instabilities of high-velocity magnetized shear layers with generalized polytrope laws. *Quarterly Journal of Applied Mathematics*, LVIII, 401–423

Chandrasekhar, S. 1981. *Hydrodynamic and Hydromagnetic Stability*, Oxford: Clarendon Press

Choudhury, S.R. 1986. Kelvin–Helmholtz instabilities of supersonic, magnetized shear layers. *Journal of Plasma Physics*, 35: 375–392

Choudhury, S. R. & Brown, K.G. 2001. Novel dynamics in the nonlinear evolution of the KH instability of supersonic anisotropic tangential velocity discontinuities. *Mathematics and Computers in Simulations*, 55: 377

Choudhury, S.R. & Lovelace, R.V. 1984. On the Kelvin–Helmholtz instabilities of supersonic shear layers. *Astrophysical Journal*, 283: 331–334

Debnath, L. & Choudhury, S.R. 1997. *Nonlinear Instability Analysis*, Southampton: Computational Mechanics

Drazin, P.G. & Reid, W.H. 1981. *Hydrodynamic Stability*, Cambridge and New York: Cambridge University Press

Gerwin, R.A. 1968. Stability of the interface between two fluids in relative motion. *Reviews of Modern Physics*, 40: 652–658

Helmholtz, H. von 1868. Über discontinuirliche Flüssigkeitsbewegungen. *Wissenschaftliche Abhandlungen. Monats. Konigl. Preuss. Akad. Wiss. Berlin*, 23: 215–228; as On discontinuous movements of fluids. *Philosophical Magazine*, 36: 337–36 (1868)

Kelvin, Lord (William Thomson). 1871. Hydrokinetic solutions and observations. *Philosophical Magazine*, 42(4): 362–377; with On the motion of free solids through a liquid; Influence of wind and capillarity on waves in water supposed frictionless. In *Mathematical and Physical Papers*, vol. 4: *Hydrodynamics and General Dynamics*, Cambridge: Cambridge University Press, 1910

Miura, A. 1990. Kelvin–Helmholtz instability for supersonic shear flow at the magnetospheric boundary. *Geophysical Research Letters*, 17: 749–752

Miura, A. 1995. Kelvin–Helmholtz instability at the magnetopause: computer simulations. In *Physics of the Magnetopause, Geophysical Monogr*, edited by P. Song, B.U. Sonnerup & F. Thomsen, Washington, DC: American Geophysical Union

Miura, A. 1999. Self-organization in the two-dimensional Kelvin–Helmholtz instability. *Physical Review Letters*, 83: 1586–1589

Miura, A. & Pritchett, P.L. 1982. Nonlinear stability analysis of the MHD Kelvin–Helmholtz instability in a compressible plasma. *Journal of Geophysical Research*, 87: 7431–7444

Normal, M. L., Smarr, L., Winkler, K.H. & Smith, M.D. 1982. Supersonic jets. *Astronomy and Astrophysics*, 113: 285

Weissman, M. 1979. Nonlinear wave packets in the KH instability. *Philosophical Transactions of the Royal Society, London*, 290: 639

KEPLER PROBLEM

See **Celestial mechanics**

K-EPSILON METHOD

See **Turbulence**

KERMACK–MCKENDRICK SYSTEM

See **Epidemiology**

KERR EFFECT

Discovered by John Kerr in 1875, the electro-optical effect opened new trends in the world of optics. This effect refers to the modifications that a light beam undergoes in amplitude, phase, or direction when it propagates in an optical material that responds to an external electric field. An example is the modulation of optical radiation in crystals in which two possible linearly polarized modes exist with unique directions of polarization and corresponding indices of refraction. For crystals with no inversion symmetry, $(n^2)_{ij}$ denotes the squared ray index along the directions $(i, j) = 1, 2, 3$, such as $x = 1$, $y = 2$, $z = 3$. The linear change in the coefficients $(1/n^2)_{ij}$ due to an applied low-frequency electric field $\boldsymbol{E}(E_x, E_y, E_z)$ is then defined by

$$\Delta\left(\frac{1}{n^2}\right)_{ij} = \sum_k r_{ijk} E_k \quad (k = 1, 2, 3), \tag{1}$$

where the 6×3 matrix with elements r_{ijk} is called the electro-optical tensor. Current applications of this property are electro-optic retardation, amplitude or phase modulation, and beam deflection (Yariv, 1975).

Optical nonlinearities result from an analogous effect arising when the low-frequency electric field applied is replaced by a second optical frequency. Let us consider the electric field of an optical beam along the j-direction with frequency ω_1: $E_j^{\omega 1}(t) = \mathrm{Re}(A_j^{\omega 1} \mathrm{e}^{-\mathrm{i}\omega_1 t})$ while a second field at ω_2 is $E_k^{\omega 2}(t) = \mathrm{Re}(A_k^{\omega 2} \mathrm{e}^{-\mathrm{i}\omega_2 t})$. If the medium is nonlinear, these fields generate polarizations at frequencies $\omega_3 = n\omega_1 + m\omega_2$, where n and m are any integers. The susceptibility tensor $d_{ijk} = -\varepsilon_i \varepsilon_j r_{ijk}/2\varepsilon_0$ then enters the polarization vector as $P_i^{\omega 3} = \sum_j \sum_k d_{ijk} A_j^{\omega 1} A_k^{\omega 2}$, where ε_i and ε_j are the dielectric constants along i and j at ω_1, and ε_0 is the vacuum permittivity.

A single optical beam is able to induce nonlinear effects. The origin of the nonlinear response is then related to anharmonic motions of bound electrons under the influence of a primary optical field. As a result, the induced polarization $\boldsymbol{P}$ created from the electric dipoles is not linear in the electric field $\boldsymbol{E}$, but it satisfies the more general relation

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{P}_{\mathrm{L}} + \boldsymbol{P}_{\mathrm{NL}} \\ &= \varepsilon_0 \{\chi^{(1)} \cdot \boldsymbol{E} + \chi^{(2)} : \boldsymbol{EE} \\ &\qquad + \chi^{(3)} \vdots \boldsymbol{EEE} + \cdots\}, \end{aligned} \tag{2}$$

where $\chi^{(n)}$ $(n = 1, 2, ...)$ is the nth-order susceptibility tensor of rank $n + 1$. The linear susceptibility $\chi^{(1)}$ represents the dominant, linear contribution to $\boldsymbol{P}$,

$$\boldsymbol{P}_{\mathrm{L}} = \varepsilon_0 \int_{-\infty}^{+\infty} \chi^{(1)}(t - t') \cdot \boldsymbol{E}(\boldsymbol{r}, t')\, \mathrm{d}t', \tag{3}$$

including the linear refractive index n_0 and related attenuation coefficient. $\chi^{(2)}$ is the second-order susceptibility responsible, for example, for second-harmonic generation in crystals with a lack of inversion symmetry. For centro-symmetric media such as optical fibers, silica samples, and gases, this contribution is zero and the first nonlinearities are provided by the third-order susceptibility $\chi^{(3)}$ through

$$\begin{aligned} \boldsymbol{P}_{\mathrm{NL}} = \varepsilon_0 \int\!\!\int\!\!\int_{-\infty}^{+\infty} & \chi^{(3)}(t - t_1, t - t_2, t - t_3) \\ & \vdots \boldsymbol{E}(\boldsymbol{r}, t_1) \boldsymbol{E}(\boldsymbol{r}, t_2) \boldsymbol{E}(\boldsymbol{r}, t_3)\, \mathrm{d}t_1\, \mathrm{d}t_2\, \mathrm{d}t_3. \end{aligned} \tag{4}$$

This causes an intensity dependence in the total refractive index of the form

$$n = n_0 + n_2|\boldsymbol{E}|^2 \tag{5}$$

with nonlinear coefficient $n_2 = 3\chi^{(3)}_{xxxx}/8n_0$ for a linearly polarized electric field involving a single $\chi^{(3)}$ component of four-rank tensor. This intensity dependence induced in the refractive index through $\chi^{(3)}$ contributions is called the Kerr effect.

The Kerr effect affects the evolution of the electric field $\boldsymbol{E}$, which is governed by Maxwell's equations. Straightforward combination of the latter, after eliminating the magnetic and electric flux density ($\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}$), provides the wave equation

$$\nabla \times \nabla \times \boldsymbol{E} + \frac{1}{c^2}\partial_t^2 \boldsymbol{E} = -\mu_0 \partial_t^2 \boldsymbol{P}, \tag{6}$$

where $\mu_0\varepsilon_0 = 1/c^2$ and c is the speed of light in vacuum. For treating Equation (6) conveniently, certain assumptions can be made, among which are the following:

- $|\boldsymbol{P}_{\rm NL}|$ is a small perturbation to the total induced polarization.
- Optical losses are weak, so that the imaginary part of the frequency-dependent dielectric constant $\varepsilon(\omega)$ is negligible, that is, $\varepsilon(\omega) = 1 + \tilde{\chi}^{(1)}(\omega) \simeq n^2(\omega)$.
- $\nabla \cdot \boldsymbol{D} \simeq \varepsilon \nabla \cdot \boldsymbol{E} = \boldsymbol{0}$, and thus $\nabla \times \nabla \times \boldsymbol{E} = \nabla \cdot (\nabla \boldsymbol{E}) - \nabla^2 \boldsymbol{E} \simeq -\nabla^2 \boldsymbol{E}$.
- The polarization state is preserved, so that a scalar approach is valid.
- The optical field is quasi-monochromatic; that is, its spectrum centered at $\omega_0 = 2\pi c/\lambda_0$ for the wavelength λ_0 has a spectral width $\Delta\omega = \omega - \omega_0$ satisfying $\Delta\omega/\omega_0 \ll 1$.

It is then useful to separate out the rapidly varying part of the electric field, like $\boldsymbol{E}(\boldsymbol{r}, t, z) = \hat{\boldsymbol{x}} A(\boldsymbol{r}, t, z)\, {\rm e}^{-{\rm i}(\omega_0 t - k_0 z)}$ for a wavefield propagating along the z axis with wave number $k_0 = k(\omega_0) = n_0\omega_0/c$. As a result, the complex envelope A is described by the nonlinear Schrödinger (NLS) model (Brabec & Krausz, 1997)

$${\rm i}\partial_\xi A + \frac{1}{2k_0} T^{-1} \nabla_\perp^2 A + \hat{D} A + \frac{\omega_0 n_2}{c} T(|A|^2 A) = 0, \tag{7}$$

after passing over to the frame moving with the group-velocity, by means of the new variables $\xi = z$, $\tau = t - \beta_1 z$ where $\beta_1 \equiv \partial k/\partial\omega|_{\omega_0}$. Here, the assumptions of paraxiality ($|\partial_\xi A| \ll k_0 |A|$) and slow variations in time ($|\partial_\tau A| \ll \omega_0 |A|$) have been made. The transverse Laplacian $\nabla_\perp^2 = \partial_x^2 + \partial_y^2$ accounts for wave diffraction in the x–y plane [$\boldsymbol{r} = (x, y)$]. $\hat{D} \equiv \sum_{m=2}^{+\infty} \frac{1}{m!} (\partial^m k/\partial\omega^m|_{\omega_0})({\rm i}\partial_\tau)^m$ is the dispersion operator, where only the leading-order contribution ($m = 2$) is considered below, which corresponds to group-velocity dispersion (GVD) with coefficient $\beta_2 \equiv \partial^2 k/\partial\omega^2|_{\omega_0}$. The operator $T = 1 + ({\rm i}/\omega_0)\partial_\tau$ originates from the cross derivatives $\partial^2_{\xi,\tau}$ retained in the derivation of Equation (7), and it reduces to unity, whenever $\Delta\omega/\omega_0 \ll 1$ holds. For $n_2 > 0$, the principal physical processes caused by the Kerr effect can be listed as follows:

- The Kerr effect creates spectral broadening through self-phase modulation (SPM). Solving for ${\rm i}\partial_\xi A = -\omega_0 n_2 |A|^2 A/c$ indeed yields the exact solution $A = A_0 {\rm e}^{{\rm i}\omega_0 n_2 |A_0|^2 \xi/c}$ with $A_0 = A(\xi = 0)$, which describes a self-induced phase shift experienced by the optical field during its propagation. This intensity-dependent phase shift refers to SPM, which is responsible for the spectral broadening of short pulses by virtue of the relation $\Delta\omega/\omega_0 = -\partial_\tau \arg(A)/\omega_0$. The pulse shape does not change in time, but the frequency spectrum is expanded by the nonlinearity.
- In optical fibers for which diffractive effects are negligible, pulses mainly undergo GVD and Equation (7) reduces to the standard cubic NLS equation in the limit $T \to 1$. With $\beta_2 < 0$ (so-called anomalous GVD), this equation describes the balance between 1-d dispersion and nonlinearity, which gives rise to sech-shaped, "bright" temporal solitons. With $\beta_2 > 0$ (normal GVD), "dark" solitons appear as localized dips against a uniform background. Both types of solitons have been detected experimentally (Agrawal, 1989). Solitons in dispersion-managed optical fibers, prepared with alternating spans of normal and anomalous GVD, have important applications in high-rated data communication systems over distances of many kilometers.
- For ultrashort pulses developing sharp temporal gradients, the operator T in front of the Kerr term in (7) induces shock dynamics: the field intensity without GVD and diffraction is governed by the continuity relation $\rho_\xi + 3n_2\rho\rho_\tau/c = 0$ with $\rho = |A|^2$. For localized pulses with compact distribution F, ρ has the exact implicit solution $\rho = F(\tau - 3n_2\xi\rho/c)$, which gives rise to a singular shock profile with $|A_\tau| \to +\infty$ in the trail ($\tau > 0$) of the pulse. This effect is usually called self-steepening (Anderson & Lisak, 1983). A comparable formation of a trailing shock edge, called space-time focusing, follows from the operator T^{-1} acting on diffraction (Rothenberg, 1992).
- For beams with no temporal dispersion ($T = 1$, GVD $= 0$), Equation (7) describes the self-focusing of optical wave packets (Marburger, 1975). Self-focusing arises as a shrinking of the beam in the diffraction plane, which is achieved by a transverse collapse when no saturation of the Kerr nonlinearity takes place. This process develops when the input power $P_{\rm in} = \int |A|^2 \,{\rm d}\boldsymbol{r}$ exceeds the critical value $P_{\rm cr} \simeq \lambda_0^2/2\pi n_0 n_2$. The beam waist then decreases

more and more, as the field amplitude $|A|$ diverges. In media such as CS_2 liquids (Campillo et al., 1973), the optical field distribution is modulationally unstable, and self-focusing often produces filamentation, along which the background beam is broken up into small-scale cells that self-focus in turn. Their growth is, however, generally arrested by higher-order nonlinearities coming from Equation (2), which can form stable spatial solitons.

- By mixing GVD and spatial diffraction when $T \simeq 1$, the Kerr nonlinearity causes defocusing in time for $\beta_2 > 0$ and temporal compression for $\beta_2 < 0$, in addition to wave focusing in the transverse direction. The interplay of these processes results in the symmetric splitting of the pulse along the time axis with normal GVD and to a 3-d spatiotemporal collapse with anomalous GVD (Bergé, 1998).
- Finally, the Kerr response of optical media favors cross-phase modulation (XPM). XPM refers to the nonlinear phase shift of an optical field induced by a copropagating beam at a different wavelength. Noting the total electric field as $\boldsymbol{E} = \hat{\boldsymbol{x}}\,\mathrm{Re}[A_1 \mathrm{e}^{-\mathrm{i}\omega_1 t} + A_2 \mathrm{e}^{-\mathrm{i}\omega_2 t}]$ for two components with different frequencies ω_1 and ω_2, the phase shift for the field at ω_1 over a length L is given by $\phi_{\mathrm{NL}} = n_2 k_0 L(|A_1|^2 + 2|A_2|^2)$. The last term refers to XPM, from which spectral broadening is larger than that for one wave alone. Coupling of several wave components belongs to the category of parametric processes, among which are the third-harmonic generation, four-wave mixing and parametric amplification (Shen, 1984; Agrawal, 1989).

LUC BERGÉ

See also **Development of singularities; Filamentation; Harmonic generation; Nonlinear optics**

Further Reading

Agrawal, G.P. 1989. *Nonlinear Fiber Optics*, San Diego: Academic Press, pp 14–36

Anderson, D. & Lisak, M. 1983. Nonlinear asymmetric self-phase modulation and self-steepening of pulses in long optical waveguides. *Physical Review* A, 27: 1393–1398

Bergé, L. 1998. Wave collapse in physics: Principles and applications to light and plasma waves. *Physics Reports*, 303: 259–370

Brabec, T. & Krausz, F. 1997. Nonlinear optical pulse propagation in the single-cycle regime. *Physical Review Letters*, 78: 3282–3285

Campillo, S.L., Shapiro, S.L. & Suydam, B.R. 1973. Periodic breakup of optical beams due to self-focusing. *Applied Physics Letters*, 23: 628–630

Marburger, J.H. 1975. Self-focusing: theory. *Progress in Quantum Electronics*, 4: 35–110

Rothenberg, J.E. 1992. Space-time focusing: breakdown of the slowly varying envelope approximation in the self-focusing of femtosecond pulses. *Optics Letters*, 17: 1340–1342

Shen, Y.R. 1984. *The Principles of Nonlinear Optics*, New York: Wiley, pp 242–333

Yariv, A. 1975. *Quantum Electronics*, 2nd edition, New York: Wiley, pp 407–507

KICKED ROTOR

The properties of a classical and/or quantum kicked rotor can be understood by examining how nonlinearities influence the dynamics of conservative dynamical systems. The story begins with the French mathematician Henri Poincaré and his famous study of planetary motion (Poincaré, 1892). Unlike the integrable two-body problem, a third body added to the system experiences nonperiodic motion (*See* ***N*-body problem**). Indeed, the trajectory of a light mass moving under the influence of the gravity of two heavy masses is very complex, and we know that such an orbit is fractal. This is one end of the spectrum of dynamical possibilities; on the other end are the integrable Hamiltonian equations of the Kolmogorov–Arnol'd–Moser (KAM) theory.

Thus, in one limit we have the integrable Hamiltonians that are the subject of the KAM theory, and in another we have non-integrable Hamiltonians, an example of which is the three-body problem of Poincaré. So how do we get from one side to the other using a perturbed Hamiltonian? In the unperturbed system, the various degrees of freedom are uncoupled, given by the action-angle variables, and the motion unfolds on KAM tori. Small perturbations leave the motion essentially unchanged, but even then nonlinear resonances are introduced in restricted regions of phase space. These nonlinear resonances change the dynamics of the system and deform the tori on all scales as the strength of the perturbation increases, leading to breakup of the KAM tori and chaos.

The properties of such nonlinear resonances can be examined using maps, which in a Hamiltonian system is equivalent to using a strobe lamp to register the momentum and displacement at equally spaced time intervals. One map that approximates several physical systems is the Taylor–Chirikov or standard map (Chirikov, 1979). For the nth flash we have for this map the momentum p_n and displacement x_n,

$$p_{n+1} = p_n \frac{K}{2\pi} \sin(2\pi x_n), \quad x_{n+1} = x_n + p_{n+1}. \quad (1)$$

Figure 1 shows a sequence of standard map plots for increasing values of the coupling strength K. Each of the continuous curves is the result of a mapping of a single initial condition. The orbits in the vicinity of $p = 0.5$ in Figure 1a are those of a simple pendulum, which is the "harmonic oscillator" of nonlinear dynamics. As the size of K is increased, Figure 1b shows that the nonlinear resonances generate more and more structure in the phase space as well as deforming the already existing KAM tori. The tori break up into sprays of points, beginning in the vicinity of the nonlinear resonances, whose exact locations are sensitively dependent on initial conditions. This is chaos, and as K is further increased, the chaos spreads throughout the phase space among the islands of stability, as in Figure 1c. There is no way to predict the motion of the trajectory in the chaotic sea. Further increases in the coupling parameter result in more tori disintegrating until the phase space appears to be completely dominated by chaos and all but the most stable of the KAM tori have disappeared, as in Figure 1d.

The standard map is also called a kicked rotor because it describes the strobe plot of the phase space trajectories for a rotor kicked every T units of time with a strength that depends on the angular position of the rotor. The Hamiltonian for the classical kicked rotor (CKR) can be written as (Chirikov, 1979; Reichl, 1992)

$$H = \frac{J^2}{2l} + K\cos\theta \sum_{k=-\infty}^{k=\infty} \delta(t - nT)\,, \qquad (2)$$

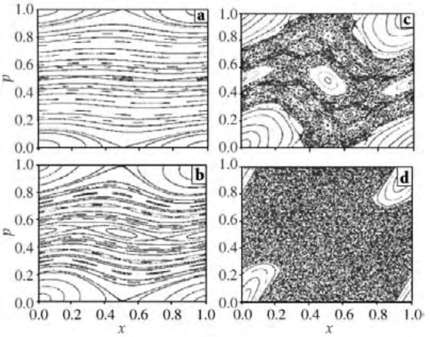

Figure 1. The phase space orbits for the kicked rotor are depicted for four values of the coupling strength: (a) $K = 0.1716354$; (b) $K = 0.4716354$; (c) $K = 1.1716354$, and (d) $K = 3.9716354$ (from Reichl, 1992).

where J is the angular momentum of the rotor, θ is its angular position, and I is the moment of inertia. Reichl (1992) reviews how to obtain the discrete map (2) by integrating the Hamiltonian equation of motion generated by (3) from one pulse to the next.

Hamiltonian (3) provides a classical description of an electron subjected to a periodically pulsed electric field. The nonlinear mechanisms are present in the motion of charged particles in accelerators and storage rings, as well as in the motion of planets and man-made celestial bodies. In the classical case, the particle dynamics in the chaotic sea are interpreted as diffusion, under which the mean-square separation of trajectories increases linearly in time with a diffusion coefficient given by $2D \approx K^2$. Thus, for the coupling strength (K) sufficiently large, the dynamical system behaves like simple Brownian motion. This has led a number of investigators to conjecture that chaos provides the dynamical foundation of thermodynamics, rather than the usual heat bath, with an infinite number of degrees of freedom (Ford & Waters, 1963; Chirikov, 1979). For example, the fluctuation-dissipation relation has been generated using nonlinear dynamical systems (Bianucci et al., 1995).

The quantum kicked rotor (QKR) employs a quantized version of the Hamiltonian, where the angular momentum J is replaced with the operator $\mathrm{i}\hbar\,\partial/\partial\theta$. The mapping is carried out using the Floquet theory on the Schrödinger equation, because the Hamiltonian is periodic in time. There is a remarkable similarity between the QKR and the one-dimensional tight-binding model of Anderson localization in solid state physics. The dynamical evolution of the QKR can be determined analytically in terms of a Floquet map, where the Floquet momentum eigenstates are exponentially localized for irrational kicks and extended for rational kicks (Reichl, 1992). It has been established that the quantum density for the momentum states is a truncated Levy distribution in direct analogy with the classical probability density as shown in Figure 2 (Stefancich et al., 1998). The peaks on either side of the classical and quantum

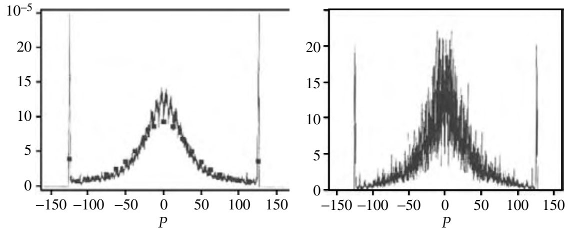

Figure 2. On the left is depicted the classical probability density for the dimensionless momentum of the CKR after 20 iterations of the map (2) and the solid curve is the theoretical prediction of the truncated Lévy distribution. On the right is the quantum probability density after 20 iterations for the QKR (Stefancich et al., 1998).

distributions correspond to the outward propagation of the initial states at maximum velocity. Finally, the CKR and QKR have been used to describe the experimental "quasifractal" structures resulting from a periodically kicked spin system in the semiclassically large-spin regions (Nakamura, 1993).

BRUCE J. WEST

See also **Chaotic dynamics; Fluctuation-dissipation theorem; Hamiltonian systems; Kolmogorov–Arnol'd–Moser theorem; Standard map**

Further Reading

Bianucci, M., Mannella, R., West, B.J. & Grigolini, P. 1995. From dynamics to thermodynamics: linear response and statistical mechanics. *Physical Review* E, 51: 3002–3022

Chirikov, B.V. 1979. A universal instability of many-dimensional oscillator systems. *Physics Reports*, 52: 263–379

Ford, J. & Waters, J. 1963. Computer studies of energy sharing and ergodicity for nonlinear oscillator systems. *J. Math. Phys.*, 4: 1293–1305.

Nakamura, K. 1993. *Quantum Chaos: A New Paradigm of Nonlinear Dynamics*, Cambridge and New York: Cambridge University Press

Ott, E. 2003. *Chaos in Dynamical Systems*, 2nd edition, Cambridge and New York: Cambridge University Press

Poincaré, H. 1892–1899. *Les Methodes nouvelles de la méchanique céleste*, Paris: Gauthier-Villars; as *New Methods of Celestial Mechanics*, 3 vols, Woodbury, NY: American Institute of Physics, 1993

Reichl, L.E. 1992. *The Transition to Chaos*, New York: Springer

Stefancich, M., Allegrini, P., Bonci, L., Grigolini, P. & West, B.J. 1998. Anomalous diffusion and ballistic peaks: a quantum perspective. *Physical Review* E, 57: 6625–6633

KINKS AND ANTIKINKS

See **Sine-Gordon equation**

KIRCHHOFF'S LAWS

Introduction

Kirchhoff's laws are conservation rules for networks; given a particular network topology, they prescribe how its currents and voltages are related. This not only facilitates understanding of the behavior of a network, but also allows an analyst to write down the equations that govern its behavior. These laws were named after Gustav Robert Kirchhoff, who in 1845 published one of the first systematic treatises on the laws of circuit analysis. Kirchhoff's current law (KCL) states that the sum of all currents into any node (or single point) of the network is zero; thus the charge flowing in and out is balanced, which makes intuitive sense. In the network of Figure 1, this means that I_1 and I_4 are equal in magnitude (KCL at node A), and that the algebraic sum of I_1, I_2, and I_3 is zero (KCL at node B).

Kirchhoff's voltage law (KVL) states that the sum of voltages around a closed loop is zero, which is expected because traversing a loop brings one back to the starting point. In Figure 1, this means that V_2 and V_3 are equal (KVL around loop D), and that the algebraic sum of V_1, V_2, and V_4 is zero (KVL around loop C). Signs and directions can be confusing in both cases, and it is important to note these correctly. A positive current entering a node is the same as a negative current leaving a node. Also, moving from the negative to positive voltage (potential) terminal of an element entails a potential rise, while moving in the opposite direction entails a drop. Passive network elements create voltage drops; active ones such as voltage sources can create voltage rises.

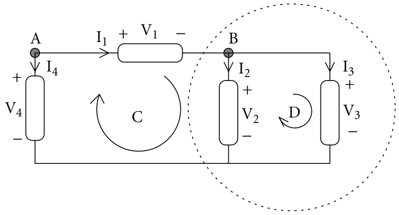

Figure 1. A simple electric network.

As with any such rules, there are some caveats. KCL actually specifies that current into a "cutset," not just a node, is zero; a cutset is any part of the network that is enclosed by a closed curve, such as the dotted circle in Figure 1. Moreover, KCL does not hold if nodes (or cutsets) can store, consume, or produce charge. KVL, on the other hand, is only valid if the voltage (or potential) between two points is independent of the path one takes between the points. If energy is lost or created as it moves around a loop, KVL does not hold.

Applications to Electronic Circuits

Kirchhoff's laws find their most common use in electronics. In Figure 2, for example, KVL around the left- and right-hand loops, respectively, yields $-V + V_1 + V_2 = 0$ and $-V_2 + V_3 = 0$, and KCL written at the point where the three resistors join yields $I_1 - I_2 - I_3 = 0$. These relationships are fairly obvious in such a simple network, assuming one gets the directions and the associated signs right. KVL and KCL become important when the network topology is complex or when one wants to automate the process in a computer algorithm.

If the constitutive relations (voltage-current characteristics) of the elements in a network are known, one can rewrite either set of equations in terms of the other state variables. Resistors, for example, obey Ohm's law ($V = IR$), so the first KVL equation can be rewritten in terms of the currents: $V = I_1R_1 + I_2R_2$.

The governing equations for a simple resistor circuit as in Figure 2 are linear and algebraic. When the

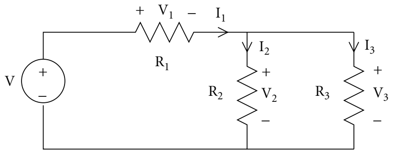

Figure 2. A resistor network.

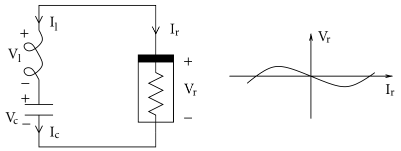

Figure 3. A simple oscillator circuit.

circuit includes inductors and capacitors, the equations become integro-differential, but Kirchhoff's laws work just as well, and they also handle nonlinear components easily. The circuit of Figure 3, for instance, is an example of the van der Pol oscillator. In this case, KCL simply states that the currents through all three elements are identical. KVL implies that $V_l + V_c - V_r = 0$. The constitutive relations for the capacitor and inductor are $I_c = C\,\mathrm{d}V_c/\mathrm{d}t$ and $V_l = L\,\mathrm{d}I_l/\mathrm{d}t$, and the resistor in this circuit is nonlinear: $V_r = -aI_r + bI_r^3$. Thus, KVL can be rewritten as

$$L\frac{\mathrm{d}I_l}{\mathrm{d}t} + \frac{1}{C}\int I_c\mathrm{d}t + aI_r - bI_r^3 = 0.$$

Other Applications

These ideas generalize beyond electronic circuit analysis. Many other systems, ranging from vehicle suspensions to social groups, can be described by networks. Moreover, KVL and KCL are actually instances of a more general set of laws. In the late 1950s and early 1960s, inspired by the realization that the principles underlying KCL and Newton's third law were identical (summation of {forces, currents} at a point is zero, respectively; both are manifestations of the conservation of energy), researchers began combining multi-port methods from a number of engineering fields into a generalized engineering domain with prototypical components (Paynter, 1961). The basis of this generalized physical networks (GPN) paradigm is that the behavior of an ideal two-terminal element—the "component"—can be described by a mathematical relationship between two dependent variables: generalized flow and generalized effort, where flow × effort = power. This pair of variables is different in each domain: (flow, effort) is (current, voltage) in an electrical domain and (force, velocity) in a mechanical domain.

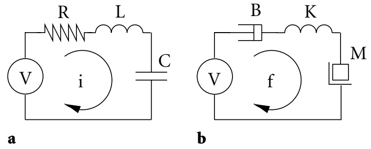

Figure 4. (a) An electrical circuit that is mathematically equivalent to (b) a mechanical circuit.

The GPN representation brings out similarities between components and properties in different domains. Electrical resistors ($v = i\,R$) and mechanical dampers or "dashpots" ($v = f\,B$) are analogous, as both dissipate energy. Both of the networks in Figure 4, for example, can be modeled by a series inertia-resistor-capacitor GPN. Thus network (a) is an electronic RLC circuit (like the van der Pol example of Figure 3), and network (b) is a mechanical mass-spring-damper system that has identical behavior. Similar analogies exist for generalized inertia, capacitance, flow, and effort source components for mechanical rotational, hydraulic, and thermal domains (Karnopp et al., 1990, Sanford, 1965). These correspondences and generalizations allow applications of KVL and KCL to mechanical structures (buildings, vehicle suspensions, aircraft, etc.), which can be modeled as interconnected networks of masses, springs, and dashpots. This approximation gives analytic insight into the vibrational modes of buildings (important for earthquake protection) and of aircraft (to keep engine frequencies from damaging wings).

ELIZABETH BRADLEY

Further Reading

Karnopp, D., Margolis, D. & Rosenberg, R. 1990. *System Dynamics: A Unified Approach*, 2nd edition, New York: Wiley

Paynter, H. 1961. *Analysis and Design of Engineering Systems*, Cambridge MA: MIT Press

Sanford, R. 1965. *Physical Networks*, Englewood Cliffs, NJ: Prentice-Hall

KLEIN–GORDON EQUATION

See **Sine-Gordon equation**

KNOT THEORY

A simple closed curve in three-dimensional space is a knot; more precisely, if M denotes a closed orientable three-manifold, then a smooth embedding of S^1 in M is called a knot in M. A link in M is a finite collection of disjoint knots, where each knot is a component of the link. Knot theory deals with the study and application of mathematical properties of knots and links in pure and applied sciences. As purely mathematical objects, knots are studied for the purpose of classifying three-dimensional surfaces according to the degree of topological complexity, regardless of their specific embedding and geometric properties (Figure 1). In this sense, knot theory is part of topology. In recent years, however, knot theory has embraced applications in dynamical systems, stimulated by the challenging difficulties associated with the study of physical knots (Kauffman, 1995). In this context knots and links are representatives of virtual and numerical objects (given by dynamical flows, phase space trajectories, and visiometric patterns), and are used to model tube-like physical systems, such as vortex filaments, magnetic loops, electric circuits, elastic cords, or even high-energy strings. For physical knots, topological issues and geometric and dynamical aspects are intimately related, influencing each other in a complex fashion. Virtual or numerical knots are studied in relation to the generating algorithms and the probability of forming knots, whereas the study of physical knots addresses questions relating topology and physics, as in the case of the topological quantum field theory (Atiyah, 1990) and topological fluid mechanics (Arnol'd & Khesin, 1998; Ricca, 2001).

Mathematical Aspects

Let us introduce some basic mathematical concepts (see, for example, Adams, 1994). A knot is said to be oriented in M, if it is a smooth embedding of an oriented curve. Two knots K and K' are said to be equivalent if there exists a smooth orientation-preserving automorphism $f: M \to M$ such that $f(K) = K'$; in particular, if the knot K is continuously deformed by f (preserving the curve orientation) to the knot K', then the two knots K and K' are said to be equivalent by ambient isotopy, and the isotopy class of K is represented by its knot type. Since knot theory deals essentially with the properties of knots and links up to isotopy, the knot parametrization, as well as any other geometric information, is irrelevant. A knot diagram of K is a plane projection with crossings marked as under or over; among the infinitely many diagrams representing the same knot K, the minimal diagram is the diagram with a minimum number of crossings. According to the type of crossing, it is customary to assign to each crossing in the knot diagram the value $\varepsilon = +1$ or $\varepsilon = -1$, as shown in Figure 2: by switching one crossing in the knot diagram from positive to negative (or the other way round), we obtain a different knot type, which is identical except for this crossing. By switching all the crossings we obtain the mirror image of the original knot. If the knot is isotopic to its mirror image, then its knot type is said to be achiral, otherwise it is chiral.

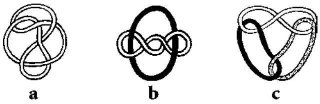

Figure 1. Three examples of knot and link types: (a) the six-crossing knot 6_3; (b) the two-component six-crossing link 6_3^2; (c) the three-component seven-crossing link 7_1^3.

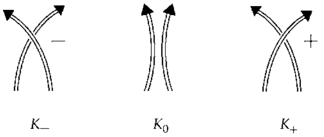

Figure 2. Standard crossing notation and algebraic sign convention for oriented strands: $\varepsilon(K_-) = -1$; $\varepsilon(K_0) = 0$; $\varepsilon(K_+) = +1$.

A knot invariant is a quantity whose value does not change when it is calculated for different isotopic knots. There are many types of invariants of knots and links, but the most common are of numerical or algebraic nature. One of the most important is the genus $g(K)$ of the knot K: recall that closed orientable surfaces are classified by genus, given by the number of handles in a handle-body decomposition. The genus $g(K)$ is defined as the minimum genus over all orientable surfaces S, which span an oriented knot K, where $\partial S = K$. One of the simplest combinatorial invariants of a knot is the minimum number of crossings of a knot K in any projection, called the crossing number $c(K)$. A fundamental invariant of links is the linking number $\mathrm{Lk}(K_1, K_2)$, that measures the topological linking between the knots K_1 and K_2; this invariant, discovered by Carl Friedrich Gauss in 1833, can be easily calculated by the crossing sign convention of Figure 2:

$$\mathrm{Lk}(K_1, K_2) = \frac{1}{2} \sum_{r \in K_1 \sqcap K_2} \varepsilon_r , \qquad (1)$$

where $\varepsilon_r = \pm 1$ and $K_1 \sqcap K_2$ denotes the total number of crossings (not necessarily minimal) between K_1 and K_2. Following the pioneering work of James W. Alexander, who used a Laurent polynomial $\Delta_K(q)$ in q to compute a polynomial invariant for the knot K by using its projection on a plane, many other polynomial invariants have been introduced; most notably the Jones polynomial $V_K(t)$ in $t^{1/2}$, defined by the following set of axioms:

(i) Let K and K' be two oriented knots (or links), which are ambient isotopic. Then

$$V_K(t) = V_{K'}(t). \qquad (2)$$

(ii) If U is the unknotted loop (that is the unknot), then

$$V_U(t) = 1. \tag{3}$$

(iii) If K_+, K_-, and K_0 are three knots (links) with diagrams that differ only as shown in the neighborhood of a single crossing site for K_+ and K_- (see Figure 2), then the polynomial satisfies the following skein relation

$$t^{-1}V_{K_+}(t) - tV_{K_-}(t) = (t^{1/2} - t^{-1/2})V_{K_0}(t). \tag{4}$$

An important property of the Jones polynomial (which is not shared by previous polynomials) is that it can distinguish between a knot and its mirror image. Later work has led to other polynomial invariants, namely, the HOMFLY and Kauffman polynomials, and to a more abstract approach to algebraic invariants (Vassiliev invariants and Lie algebras). There are also invariants of different nature: among these, we mention the fundamental group $\pi_1(S^3/K)$ of the knot complement and its hyperbolic volume $v(K)$. The classification of knots and links has led to the important study of braids: these are given by a set of n interlaced strings, with ends defined on two parallel planes, placed at some distance h apart. According to specific topological characteristics, we may consider special types of knot sub-families, such as torus knots, alternating knots, two-bridge knots, tangles, and many others (see Hoste et al., 1998).

Virtual Knots

Virtual knots arise from dynamical flows, generated by the vector field of a specific ordinary differential equation (Ghrist, 1997), in connection with phase-space dynamics and statistical mechanical models (Millett & Sumners, 1994) or, as recently done, from application of ideas from the quantum field theory with an appropriate Lagrangian. This latter approach, originated in work by E. Witten in 1989, has led to the creation of a new area, called the topological quantum field theory, that has proven to be extremely fruitful in providing new results on invariants of low-dimensional manifolds. Soliton knots are given by solutions to soliton equations for one-dimensional systems: in this context there are intriguing questions relating topological invariants, integrability, and conservation laws. Virtual knots and links are generated in visiometrics by numerical simulations: in this case, smooth knots are replaced by polygonal knots, where the number of segments (or sticks) is the result of numerical discretization. Stimulating questions address the minimum number of sticks of given length for each knot type and the generation of knots and links by minimal random walks. Other questions regard charged knots: these are knots and links charged by potentials that generate self-attraction or repulsion on the knot strands (see Figure 3). Under volume-preserving diffeomorphisms, the knot is led to relax by minimizing the knot energy (defined by an appropriate functional), by shrinking or extending the length as far as possible, depending on the potential, to attain an ideal shape (Stasiak et al., 1998). Questions relating to topology and geometry of ideal shapes, and uniqueness of minimum energy states, pose challenging problems at the crossroads of topology, differential geometry, functional analysis, and numerical simulation.

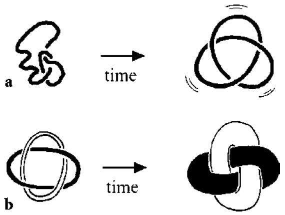

Figure 3. Examples where a topological barrier prevents further relaxation under a volume-preserving diffeomorphism: (a) an electrically charged trefoil knot is maximally extended by the Coulomb repulsion forces to its minimum energy state; (b) a magnetic link attains ground state energy by the action of the Lorentz force on the magnetic volume.

Physical Knots

By physical knots we mean tube models, centered around the knot K, with length $L(K)$, tubular neighborhood of radius $r(K)$, and volume $V(K)$. The tube is filled by vector field lines, whose distribution gives physical properties in terms, for example, of elasticity, vorticity, or magnetic field. A wide variety of filamentary systems present in nature at very different scales can be modeled by physical knots: from DNA molecules, polymer chains, vortex filaments, to elastic cords, strings, and magnetic flux tubes. In fluid systems, the action at a microscopic level of physical processes, such as viscosity and resistivity, may imply changes in knot topology by local recombination of the knot strands (known as knot surgery) and consequential rearrangement of energy distribution. In elastic systems, the material breaking point and internal critical twist are strongly influenced by knot strength and rope length, the latter given by the ratio L/r. All these systems are free to relax their internal energy to states of equilibrium: lower bounds on equilibrium energy for given measures of topological complexity (based, for example, on crossing number information) can be expressed by relationships of the kind

$$E_{\min} \geq h(c, \Phi, V, n), \tag{5}$$

where $E_{\min}$ is the equilibrium energy and $h(\cdot)$ gives the relationship between physical quantities—such as flux Φ, number of components n, knot volume V—and topology, given here by the crossing number c.

Understanding the interplay between topology and energy localization and redistribution can be very important in many fields of science and applications.

Renzo L. Ricca

See also **Differential geometry; Dynamical systems; Structural complexity; Topology**

Further Reading

Adams, C. 1994. *The Knot Book*, New York: W.H. Freeman

Arnol'd, V.I. & Khesin, B.A. 1998. *Topological Methods in Hydrodynamics*, Berlin and New York: Springer

Atiyah, M. 1990. *The Geometry and Physics of Knots*, Cambridge and New York: Cambridge University Press

Ghrist, R., Holmes, P. & Sullivan, M. 1997. *Knots and Links in Three-Dimensional Flows*, Berlin and New York: Springer

Hoste, J., Thistlethwaite, M. & Weeks, J. 1998. The first 1,701,936 knots. *Mathematical Intelligencer*, 20: 33–48

Kauffman, L. (editor). 1995. *Knots and Applications*, Singapore: World Scientific

Millett, K.C. & Sumners, D.W. (editors). 1994. *Random Knotting and Linking*, Singapore: World Scientific

Ricca, R.L. (editor). 2001. *An Introduction to the Geometry and Topology of Fluid Flows*, Dordrecht: Kluwer

Stasiak, A., Katritch, V. & Kauffman, L.H. (editors). 1998. *Ideal Knots*, Singapore: World Scientific

KOCH CURVE

See **Fractals**

KOLMOGOROV CASCADE

The velocity fluctuations of a high Reynolds number flow in a three-dimensional velocity field are typically dispersed over all possible wavelengths of the system, from the smallest scales, where viscosity dominates the advection and dissipates the energy of fluid motion, to the effective size of the system. This is not so bizarre: our everyday experience tells us it is so. On the corner of a city street, one might watch the fluttering and whirling of a discarded tram ticket as it is swept by an updraught, driven by localized thermal gradients from traffic or air-conditioning units; later, on the television news, one might see reports or predictions of storms on the city or district scale, and a weather map with isobars spanning whole continents. If you are a sailor you will know how to sail, or not, the multi-scaled surface of a turbulent ocean (Figure 1). The mechanism for this dispersal is vortex stretching and tilting: a conservative process whereby interactions between vorticity and velocity gradients create smaller and smaller eddies with amplified vorticity, until viscosity takes over (Tennekes & Lumley, 1972; Chorin, 1994).

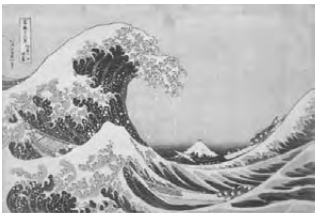

Figure 1. Turbulent action on many different scales in a high Reynolds number flow: woodcut print by Katsushika Hokusai (1760–1849).

An alternative, crude but picturesque, description of multi-scale turbulence was offered by the early 20th century meteorologist Lewis Fry Richardson (1922) in an evocative piece of doggerel: "big whirls have little whirls that feed on their velocity, and little whirls have lesser whirls and so on to viscosity." Richardson's often-quoted rhyme is apparently a parody of Irish satirist Jonathan Swift's verse: "So, naturalists observe, a flea—Has smaller fleas that on him prey—And these have smaller still to bite'—And so proceed *ad infinitum*."

The statistics of the velocity fluctuation distribution in turbulent flows were quantified rather more elegantly and rigorously by the mathematician Andrei N. Kolmogorov (1941b), who derived the subsequently famous "$-5/3$ law" for the energy spectrum of the intermediate scales, or inertial scale subrange, of high Reynolds number flows which are ideally homogeneous (or statistically invariant under translation) and isotropic (or statistically invariant under rotation and reflection) in three velocity dimensions. Two thorough, but different in style and emphasis, accounts of Kolmogorov's turbulence work are Monin & Yaglom (1971) and Frisch (1995).

Kolmogorov's idea was that the velocity fluctuations in the inertial subrange are independent of initial and boundary conditions (i.e., they have no memory of the effects of anisotropic excitation at smaller wave numbers). The turbulent motions in this subrange, therefore, show universal statistics, and the flow is self-similar. From this premise Kolmogorov proposed the first hypothesis of similarity as: "For the locally isotropic turbulence the [velocity fluctuation] distributions F_n are uniquely determined by the quantities ν, the kinematic viscosity, and ε, the rate of average dispersion of energy per unit mass [energy flux]." His second hypothesis of similarity is: "For pulsations [velocity fluctuations] of intermediate orders where the length scale is large compared with the scale of the finest pulsations, whose energy is directly dispersed into heat due to viscosity, the distribution laws F_n are uniquely determined by F and do not depend on ν."

Kolmogorov derived the form of the distribution or energy spectrum, which we denote as $\mathcal{E}(k)$, where k is the wave number given by $k^2 = k_x^2 + k_y^2 + k_z^2$, over the inertial subrange simply by dimensional analysis. By the first and second hypotheses, the spectrum must be a function of the energy flux and wave number and independent of the viscosity or any other parameters:

$$\mathcal{E}(k) = f\,(\varepsilon, k)\,.$$

By reference to the table below (after Vallis, 1999) we find

$$\begin{aligned}\mathcal{E}(k) &\sim \varepsilon^{2/3} g(k) \quad \text{(since } k \text{ is time-independent)}\\ &= C\varepsilon^{2/3}k^{-5/3},\end{aligned} \tag{1}$$

where C is a dimensionless constant which Kolmogorov (and subsequently many others, see Sreenivasan, 1995) deduced from experimental data to be of order 1.

Quantity	Dimension
Wave number	1/length
Energy per unit mass	length^2/time^2
Energy spectrum $\mathcal{E}(k)$	length^3/time^2
Energy flux ε	energy/time $\sim$ length^2/time^3

The physical picture associated with Equation (1) is that the kinetic energy of large-scale motions (whirls or eddies) is successively subdivided and redistributed among stepwise increasing wave number components (or smaller and smaller whirls and eddies), until the action of viscosity becomes competitive. Although this process has come to be known as the "Kolmogorov cascade," the cascade metaphor was not used by Kolmogorov. Its first use in this context is apparently due to Onsager (1945), who also highlights another assumption underlying the $-5/3$ law: that the modulation of a given Fourier component of the velocity field is mostly due to those others that belong to wave numbers of comparable magnitude.

So Kolmogorov's energy distribution says that ε is the only relevant parameter for turbulence in the inertial scale range. Can this really be true? Does the notorious "problem of turbulence" really boil down to such a simple relation for intermediate wavenumbers? (The fabled Problem of Turbulence was well summed up by Horace Lamb in 1932: "When I die and go to Heaven there are two matters on which I hope enlightenment. One is quantum electro-dynamics and the other is turbulence of fluids. About the former, I am really rather optimistic.") Understandably, for a turbulence result that seems so simple and universal, so flimsily derived yet so powerful, much effort has gone into verifying the wave number spectrum, Equation (1). It is difficult to create extremely high Reynolds number flows in the laboratory, but they exist naturally in the ocean. The first and still the most exciting verification of Equation (1) was carried out by Grant et al. (1962), who made a remarkable series of measurements of turbulent velocities from a ship in the Seymour Narrows, part of the Discovery Passage on the west coast of Canada, where the Reynolds number is $\sim 10^8$ (see Figure 2). A spectral exponent close to $-5/3$ has since been measured many times in materially different flows with high Reynolds number (e.g., Zocchi et al. (1994) in helium).

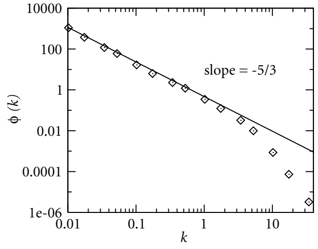

Figure 2. Data re-plotted from Grant et al. (1962), showing a Kolmogorov cascade over nearly three decades. $\phi(k)$ is the measured one-dimensional spectrum function, related to the three-dimensional spectrum function as $\mathcal{E}(k) = k^2\partial^2\phi(k)/\partial k^2 - k\partial\phi(k)/\partial k$.

All this would seem to wrap up the problem of turbulence in the inertial scale range. Or does it? There must surely be a catch somewhere! As usual, the devil is in the details. Kolmogorov himself made a "refinement," as he delicately put it, of his hypotheses (Kolmogorov, 1962). It relates to the problem of small-scale intermittency, or the uneven distribution in space of the small scales. Clearly, intermittency is inherited from initial and boundary conditions, and the side-effects on ε, the assumed-constant rate of energy transfer, are not insignificant.

In fact, there is now quite a log of complaints about the $-5/3$ law, despite its all-pervasive influence on turbulence theoretical and experimental research:

- The hypothesis of local isotropy refers to infinite Reynolds number so is not applicable to a real fluid.
- No-one has ever extracted the $-5/3$ law from the Navier–Stokes equation or vice versa.
- Is it not a circular argument that to define an inertial subrange one has to assume a cascade process, and to postulate a cascade one has to assume that an inertial subrange exists?
- What about stochastic backscatter?
- Direct interaction between large and small scales can short-circuit the cascade.
- Katul et al. (2003) found that the effects of boundary conditions were evident in the inertial subrange of an atmospheric surface layer.
- The $-5/3$ law is demonstrably invalid in two dimensions. And so on.

What is the verdict on the Kolmogorov cascade? Chorin (1994, pp. 55–57) has a bet each way; in the light of experimental verifications of the $-5/3$ law, he considers that it may be correct *despite* flaws in the arguments supporting it. The scenario proposed as being entirely consistent with Kolmogorov's theory is that energy can and does slosh back and forth across the spectrum, once the inertial range has been set up, with the energy dissipation ε being the (presumably average) difference between energy flows in both wave number directions.

The Kolmogorov cascade is starting to sound less and less like a waterfall, which *is* one-way, and more and more like an energy exchange network.

Rowena Ball

See also **Chaos vs. turbulence; Dimensional analysis; Navier–Stokes equation; Turbulence**

Further Reading

Chorin, A.J. 1994. *Vorticity and Turbulence*, New York: Springer

Frisch, U. 1995. *Turbulence: The Legacy of A. N. Kolmogorov*, Cambridge and New York: Cambridge University Press

Grant, H.L., Stewart, R.W. & Moilliet, A. 1962. Turbulence spectra from a tidal channel. *Journal of Fluid Mechanics*, 12: 241–268

Katul, G.G., Angelini, C., De Canditiis, D., Amato, U., Vidakovic, B. & Albertson, J.D. 2003. Are the effects of large scale flow conditions really lost through the turbulent cascade? *Geophysical Research Letters*, 30(4), 1164,doi:10.1029/2002GL015284

Kolmogorov, A.N. 1941a. Local structure of turbulence in an incompressible fluid for very large Reynolds numbers. *Comptes rendus (Doklady) de l'Academie des Sciences de l'U.R.S.S.*, 31: 301–305 Reprinted in, S.K. Friedlander & L. Topper 1961. *Turbulence: Classic Papers on Statistical Theory*, New York: Interscience Publishers

Kolmogorov, A.N. 1941b. On degeneration of isotropic turbulence in an incompressible viscous liquid. *Comptes Rendus (Doklady) de l'Academie des Sciences de l'U.R.S.S.*, 31: 538–540, *ibidem.*

Kolmogorov, A.N. 1962. A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number. *Journal of Fluid Mechanics*, 13: 82–85

Monin, A. S. & Yaglom, A.M. 1971. *Statistical Fluid Mechanics: Mechanics of Turbulence*, vol. 2, Cambridge: MIT Press

Onsager, L. 1945. The distribution of energy in turbulence. *Physical Review*, 68: 286

Richardson, L.F. 1922. *Weather Prediction by Numerical Process*, Cambridge: Cambridge University Press

Sreenivasan, K.R. 1995. On the universality of the Kolmogorov constant. *Physics of Fluids*, 7(11), 2778–2784

Tennekes, H. & Lumley, J.A. 1972. *A First Course in Turbulence*, Cambridge: MIT Press

Vallis, G. 1999. Geostrophic turbulence: the macroturbulence of the atmosphere and ocean. Lecture notes, www.gfdl.gov/~gkv/geoturb/

Zocchi, G., Tabeling, P., Maurer, J. & Willaime, H. 1994. Measurement of the scaling of the dissipation at high Reynolds numbers. *Physical Review* E, 50(5): 3693–3700

KOLMOGOROV, PETROVSKY AND PISCOUNOFF EQUATION

See **Diffusion**

KOLMOGOROV–ARNOL'D–MOSER THEOREM

In 1954, the Russian mathematician Andrei N. Kolmogorov, already famous for his contribution to measure theory and probability theory, delivered to the International Conference of Mathematics in Amsterdam, a lecture based on a paper entitled "On the General Theory of Dynamical Systems and Classical Mechanics." In this lecture he addressed the question of how much structure remains when a "regular" (or integrable) Hamiltonian system is replaced by one that differs from the original by only a small perturbation. Kolmogorov's original paper was not completely detailed, and clarifying work was done independently in the early 1960s by Jürgen Moser and Vladimir Arnol'd (the latter had been a student of Kolmogorov). Since then, a large body of work on this subject within the mathematical theory of Hamiltonian dynamics has been assembled.

Thus, the so-called Kolmogorov–Arnol'd–Moser (KAM) theorem is not a single theorem but rather a body of work with a common theme, which can be roughly described as the persistence of quasi-periodic motions in perturbed systems. These results were a breakthrough in the understanding of the ergodic properties of dynamical systems. The original theorems dealt with Hamiltonian systems and their discrete analogues, while more recent results have been extended to volume-preserving systems, reversible systems, and dissipative systems.

The central ideas in Kolmogorov's original paper can be formulated as follows. Under some general assumptions, the phase space M^{2n} of an integrable Hamiltonian system is foliated by invariant tori: $\mathbb{T}^n$. This system will be referred to as the unperturbed system. In the neighborhood of each torus are defined action-angle variables $(I, \phi (\text{mod}\, 2\pi))$, such that the Hamiltonian H_0 is the function of only action variables I.

The motion on each torus is conditionally periodic with frequency vector $\omega(I) = \partial H_0/\partial I$. A torus is said to be nonresonant if all the n frequencies are rationally independent (incommensurable). The unperturbed system is called nonresonant if the frequencies are functionally independent:

$$\det\left(\frac{\partial \omega}{\partial I}\right) = \det\left(\frac{\partial^2 H_0}{\partial I^2}\right) \neq 0 .$$

In a nondegenerate system, the nonresonant tori form a dense set of full measure. The resonant tori form a set of measure zero, which, however, is also dense.

The system with the Hamiltonian

$$H(I, \phi, \varepsilon) = H_0(I) + \varepsilon H_1(I, \phi, \varepsilon)$$

is called the perturbed system, where the function εH_1 is the perturbation. Note that usually the perturbed system is no longer integrable. Kolmogorov's theorem describes the fate of the nonresonant tori under perturbation.

Theorem (A.N. Kolmogorov). *If the unperturbed system is nondegenerate, then for a sufficiently small ε most nonresonant invariant tori do not vanish but are only slightly deformed, so that in the phase space of the perturbed system there are invariant tori densely filled with conditionally periodic phase curves winding around them, with the number of independent frequencies equal to the number of degrees of freedom. These invariant tori form a majority in the sense that the measure of the complement to their union is small when the perturbation is small.*

The persistence of quasi-periodic motion is consistent with numerical experiments, well-known examples being the standard map (also called the Chirikov map) and the Hénon–Heiles system.

Discussion

KAM theory has several important implications, in particular for the stability theory. Suppose that phase space is four-dimensional. A perturbed system always has one first integral, the Hamiltonian function $H(I, \phi, \varepsilon)$ itself. The energy levels $H = h$ are three-dimensional, while the invariant tori are two-dimensional. Thus, a trajectory that starts in a region between two invariant tori of the perturbed system is forever trapped in the region between the tori. This means that the values of the action variables remain forever near their initial values, which in turn implies stability.

If, however, the number n of degrees of freedom is greater than two, the n-dimensional tori do not separate the $(2n-1)$-dimensional energy level manifold into disjoint regions, and the invariant tori do not prevent a phase curve from wandering far away. There are examples of such drift, when the action variables change from their original value by a quantity of order 1, an effect known as "Arnol'd diffusion" (*See* **Arnol'd diffusion**). Although in the latter case the system is unstable, KAM theory does guarantee "metric stability," that is, stability for most initial conditions.

Examples

1. Assume the masses of the planets are sufficiently small compared with the mass of the Sun in the gravitational n-body problem. Then a large portion of the region of phase space corresponding to unperturbed motion of all planets (on identically oriented Keplerian ellipses having small eccentricities and inclinations) is filled up by conditionally periodic motions. KAM theory provides important estimates for the long-time evolution of such a system. Note that Henri Poincaré's groundbreaking work *Les méthodes nouvelles de la mécanique céleste* (1892) was motivated by questions of the stability of the solar system.
2. Consider the motion of a heavy rigid body fixed at a point. If the kinetic energy of the body is sufficiently large in comparison with its potential energy at the initial moment of time, then the length of the angular momentum vector and its inclination to the horizon remain forever near their initial values, provided that the initial values of the energy and the angular momentum differ sufficiently from values for which the body can rotate around its medium principal axis.
3. For plasma confinement in toroidal chambers, the plasma particles tend to follow magnetic field lines. The equations for these magnetic field lines can be put into Hamiltonian form. When the system is azimutally symmetric, the Hamiltonian is integrable, deviation from azimutal symmetry creates a perturbation, and KAM theory is applicable.

M.V. DERYABIN AND P.G. HJORTH

See also **Hénon–Heiles system; Standard map**

Further Reading

Arnol'd, V.I. 1962. The classical theory of perturbations and the problem of stability of planetary systems. *Soviet Mathematics Doklady*, 3: 1008–1012

Arnol'd, V.I. 1963. Proof of A.N. Kolmogorov's theorem on the preservation of quasiperiodic motions under small perturbations of the Hamiltonian. *Russian Mathematical Surveys*, 18(5): 9–36

Arnol'd, V.I. (editor). 1988. *Dynamical Systems III*, Berlin and New York: Springer

Kolmogorov, A.N. 1954. Théorie générale des systémes dynamiques et mécanique classique. In *Proceedings of the International Congress of Mathematicians*, vol. 1, Amsterdam: North-Holland, pp. 315–333

Moser, J.K. 1962. On invariant curves of area-preserving mappings of an annulus. *Nachrichten der Akademic der Wissenschatten in Goettingen II, Mathematisch-physikalische klasie* 1–20

KORTEWEG–DE VRIES EQUATION

Historical Introduction

The Korteweg–de Vries (KdV) equation, given here in canonical form,

$$u_t + 6uu_x + u_{xxx} = 0, \qquad (1)$$

is widely recognized as a paradigm for the description of weakly nonlinear long waves in many branches of physics and engineering. Here, $u(x, t)$ is an appropriate field variable, t is the time, and x is the space coordinate in the relevant direction. It describes how waves evolve under the competing but comparable effects of weak nonlinearity and weak dispersion. Indeed, if it is supposed that x-derivatives scale as ε where ε is the small parameter characterizing long waves (i.e., typically the ratio of a relevant background length scale to a wavelength scale), then the amplitude scales as ε^2

and the time evolution takes place on a scale of ε^{-3}. The KdV equation is characterized by its family of solitary wave solutions,

$$u = a\,\mathrm{sech}^2(\gamma(x - Vt)),\\ \text{where} \quad V = 2a = 4\gamma^2. \tag{2}$$

This solution describes a family of steady isolated wave pulses of positive polarity, characterized by the wave number γ; note that the speed V is proportional to the wave amplitude a and also to the square of the wave number γ^2.

The KdV equation (1) owes its name to the famous paper of Diederik Korteweg and Hendrik de Vries, published in 1895, in which they showed that small-amplitude long waves on the free surface of water could be described by the equation

$$\zeta_t + c\zeta_x + \frac{3c}{2h}\zeta\zeta_x + \frac{ch^2}{6}\delta\zeta_{xxx} = 0. \tag{3}$$

Here, $\zeta(x,t)$ is the elevation of the free surface relative to the undisturbed depth h, $c = (gh)^{1/2}$ is the linear long wave phase speed, and $\delta = 1 - 3B$, where $B = \sigma/gh^2$ is the Bond number measuring the effects of surface tension ($\rho\sigma$ is the coefficient of surface tension and ρ is the water density). Transformation to a reference frame moving with the speed c (i.e., (x,t) is replaced by $(x - ct, t)$, and subsequent rescaling readily establishes the equivalence of (1) and (3). Although Equation (1) now bears the name KdV, it was apparently first obtained by Joseph Boussinesq (1877) (see Miles (1980) and Pego and Weinstein (1997) for historical discussions on the KdV equation). Korteweg and de Vries found the solitary wave solutions (2), and importantly, they showed that they are the limiting members of a two-parameter family of periodic traveling-wave solutions, described by elliptic functions and commonly called cnoidal waves,

$$u = b + a\,\mathrm{cn}^2(\gamma(x - Vt)|m),\\ \text{where} \quad V = 6b + 4(2m-1)\gamma^2\,,\ a = 2m\gamma^2. \tag{4}$$

Here, $\mathrm{cn}(x|m)$ is the Jacobi elliptic function of modulus m $(0 < m < 1)$. As $m \to 1$, $\mathrm{cn}(x|m) \to \mathrm{sech}(x)$, and then the cnoidal wave (4) becomes the solitary wave (2), now riding on a background level b. On the other hand, as $m \to 0$, $\mathrm{cn}(x|m) \to \cos 2x$, and so the cnoidal wave (4) collapses to a linear sinusoidal wave (note that in this limit, $a \to 0$).

This solitary wave solution found by Korteweg and de Vries had earlier been obtained directly from the governing equations (in the absence of surface tension) independently by Boussinesq (1871, 1877) and Lord Rayleigh (1876), who were motivated to explain the now very well-known observations and experiments of John Scott Russell (1844). Curiously, it was not until quite recently that it was recognized that the KdV equation is not strictly valid if surface tension is taken into account and $0 < B < \frac{1}{3}$, as then there is a resonance between the solitary wave and very short capillary waves.

After the ground-breaking work of Korteweg and de Vries, interest in solitary water waves and the KdV equation declined until the dramatic discovery of the soliton by Zabusky and Kruskal in 1965. Through numerical integrations of the KdV equation, they demonstrated that the solitary wave (2) could be generated from quite general initial conditions, and could survive intact collisions with other solitary waves, leading them to coin the term *soliton*. Their remarkable discovery, followed almost immediately by the theoretical work of Gardner et al. (1967) showing that the KdV equation was *integrable* through an inverse scattering transform, led to many other startling discoveries and marked the birth of soliton theory as we know it today (*See* **Solitons, a brief history**). The implication is that the solitary wave is the key component needed to describe the behavior of long, weakly nonlinear waves.

An alternative to the KdV equation is the Benjamin–Bona–Mahony (BBM) equation in which the linear dispersive term $c\zeta_{xxx}$ in (3) is replaced by $-\zeta_{xxt}$. It has the same asymptotic validity as the KdV equation, and since it has rather better high wave number properties, it is somewhat easier to solve numerically. However, it is not integrable and, consequently, has not attracted the same interest as the KdV equation.

Both the KdV and BBM equations are uni-directional. A two-dimensional version of the KdV equation is the KP equation (Kadomtsev & Petviashvili, 1970),

$$(u_t + 6uu_x + u_{xxx})_x \pm u_{yy} = 0. \tag{5}$$

This equation includes the effects of weak diffraction in the y-direction, in that y-derivatives scale as ε^2 whereas x-derivatives scale as ε. Like the KdV equation it is an integrable equation. When the "+"-sign holds in (5), this is the KP2 equation, and it can be shown that then the solitary wave (2) is stable to transverse disturbances. On the other hand, if the "−"-sign holds, this is the KP1 equation for which the solitary wave is unstable; instead this equation supports "lump" solitons. Both KP1 and KP2 are integrable equations. To take account of stronger transverse effects and/or to allow for bi-directional propagation in the x-direction, it is customary to replace the KdV equation with a Boussinesq system of equations; these combine the long wave approximation to the dispersion relation with the leading-order nonlinear terms and occur in several asymptotically equivalent forms.

Although the KdV equation (1) is historically associated with water waves, it in fact occurs in many other physical contexts, where it arises as an asymptotic multiscale reduction from the relevant governing equations. Typically, the outcome is

$$A_t + cA_x + \mu AA_x + \lambda A_{xxx} = 0. \tag{6}$$

Here, c is the relevant linear long wave speed for the mode whose amplitude is $A(x,t)$, while μ and λ, the coefficients of the quadratic nonlinear and linear dispersive terms, respectively, are determined from the properties of this same linear long wave mode and, like c, depend on the particular physical system being considered. Note that the linearization of (6) has the linear dispersion relation $\omega = ck - \lambda k^3$ for linear sinusoidal waves of frequency ω and wave number k; this expression is just the truncation of the full dispersion relation for the wave mode being considered, and immediately identifies the origin of the coefficient λ. Similarly, the coefficient μ can be identified with the an amplitude-dependent correction to the linear wave speed. Transformation to a reference frame moving with a speed c and subsequent rescaling shows that (6) can be transformed to the canonical form (1). Equations of the form (6) arise in the study of internal solitary waves in the atmosphere and ocean, mid-latitude and equatorial planetary waves, plasma waves, ion-acoustic waves, lattice waves, waves in elastic rods, and in many other physical contexts (see, for instance, Ablowitz & Segur, 1981; Dodd et al., 1982; Drazin & Johnson, 1989; Grimshaw, 2001).

In some physical situations, it is necessary to complement the KdV equation (6) with a higher-order cubic nonlinear term of the form $\nu A^2 A_x$. After transformation and rescaling, the amended equation (6) can be transformed to the so-called Gardner equation

$$u_t + 6uu_x + 6\delta u^2 u_x + u_{xxx} = 0. \tag{7}$$

Like the KdV equation, the Gardner equation is integrable by the inverse scattering transform. Here the coefficient δ can be either positive or negative, and the structure of the solutions depends crucially on which sign is appropriate. Again, in some physical situations, solitary waves propagate through a variable environment which means that the coefficients c, μ, and λ in (6) are functions of x, while an additional term $c(\sigma_x/2\sigma)A$ needs to be included, where $\sigma(x)$ is a magnification factor. After transforming to new variables, $\theta = (\int^x dx/c) - t, x$ with $U = \sigma^{1/2}u$, the variable-coefficient KdV equation is obtained,

$$U_x + \alpha(x)UU_\theta + \beta(x)U_{\theta\theta\theta} = 0. \tag{8}$$

Here, $\alpha = \mu/c\sigma^{1/2}$, $\beta = \lambda/c^3$. In general, this is not an integrable equation and must be solved numerically, although we shall exhibit some asymptotic solutions below. Another modification of the KdV equation occurs when it is necessary to take account of background rotation, leading to the rotation-modified KP equation (see, for instance, Grimshaw, 2001), in which a term $-f^2u$ is added to the left-hand side of Equation (5), where f is a measure of the background rotation.

Solitons

The remarkable discovery of Gardner et al. (1967) that the KdV equation was integrable through an inverse scattering transform marked the beginning of soliton theory. Their pioneering work was followed by the work of Zakharov and Shabat (1972) which showed that another well-known nonlinear wave equation, the nonlinear Schrödinger equation, was also integrable by an inverse scattering transform. Their demonstration that the integrability of the KdV equation was not an isolated result, was followed closely by analogous results for the modified KdV equation (Wadati, 1972) and the sine-Gordon equation (Ablowitz et al., 1973). In 1974, Ablowitz, Kaup, Newell, and Segur provided a generalization and unification of these results in the AKNS scheme. From this point there has been an explosive and rapid development of soliton theory in many directions (see, for instance, Ablowitz & Segur, 1981; Dodd et al., 1982; Newell, 1985; Drazin & Johnson, 1989).

For the KdV equation (1) the starting point is the Lax pair (Lax, 1968) for an auxiliary function $\phi(x,t)$,

$$L\phi \equiv -\phi_{xx} - u\phi = \lambda\phi, \tag{9}$$

$$\phi_t = B\phi \equiv (u_x + C)\phi + (4\lambda - 2u)\phi_x. \tag{10}$$

Here, $C(t)$ depends on the normalization of ϕ. The first of these equations (9), with suitable boundary conditions at infinity (see below) defines a spectral problem for ϕ in the spatial variable x with a spectral parameter λ, and with the time variable t as a parameter. The second equation (10) then describes how the spectral function ϕ evolves in time. If it is now assumed that λ is independent of time (i.e., $\lambda_t = 0$ then the KdV equation (1) is just the compatibility condition for these two equations (9, 10); that is, it emerges as a result of the condition that $(\phi_{xx})_t = (\phi_t)_{xx}$. In terms of the operators L, B defined in the Lax pair (9,10), the KdV equation can be written in the symbolic form $L_t = BL - LB$ (Lax, 1968). This form indicates the path to further generalizations, in that other nonlinear wave equations can be obtained by choosing different operators L, B. The general strategy for integration of the KdV equation now consists of three steps. Here, we will describe the process under the hypothesis that we seek solutions $u(x,t)$ of the KdV equation (1), which decay to zero sufficiently fast as $x \to \pm\infty$ and have the initial condition $u(x,0) = u_0(x)$. First, we insert the initial condition into the spectral problem (9) to obtain the scattering data (these will be defined precisely below). Then (10) is used to move the scattering data forward in time; it transpires that is a very simple process, and note in particular that the spectral parameter λ is independent of t and hence is determined by the initial condition. The third step is to invert the scattering data at time $t > 0$ and so recover $u(x,t)$; this is the most difficult step, but for the KdV equation can be reduced to solution of a linear integral equation. Thus, the three steps constitute a linear algorithm for the solution of the KdV equation, and it is in this sense that it is said that the Lax pair (9,10) constitutes integrability of the KdV equation.

The spectral problem (9) for the KdV equation consists of two parts. The *discrete* spectrum is found by seeking solutions such that $\phi \to 0$ as $x \to \pm\infty$, which requires that $\lambda < 0$. It can be shown that there then exists a *finite* set of discrete eigenvalues $\lambda = -\kappa_n^2, n = 1, 2, \ldots, N$, and corresponding real eigenfunctions ϕ_n such that

$$\phi_n \sim c_n \exp(-\kappa_n x) \quad \text{as} \quad x \to \infty. \tag{11}$$

There is a similar condition as $x \to -\infty$, namely, that $\phi_n \sim d_n \exp(\kappa_n x)$. The real constants c_n, d_n are determined once the normalization condition is satisfied, that is,

$$\int_{-\infty}^{\infty} \phi_n^2 \, \mathrm{d}x = 1. \tag{12}$$

The continuous spectrum consists of all $\lambda > 0$, and so we set $\lambda = k^2$ where k is real. Then we define the scattering problem for solutions $\phi(x; k)$ of (9) by the boundary conditions,

$$\phi \sim \exp(-\mathrm{i}kx) + R(k) \exp(\mathrm{i}kx) \quad \text{as } x \to \infty, \tag{13}$$

$$\phi \sim T(k) \exp(-\mathrm{i}kx) \quad \text{as} \quad x \to -\infty. \tag{14}$$

The scattering data then consists of the set $(\kappa_n, c_n, n = 1, 2, \ldots, N)$ together with the reflection coefficient $R(k)$. It is useful to note that $R(k)$ may be continued into the upper half of the complex k-plane, has there a set of simple poles at $k = \mathrm{i}\kappa_n$, and $R \to 1$ as $|k| \to \infty$.

The next step is to determine from (10) how the scattering data evolves in time (note that the dependence on time t has been suppressed in the preceding paragraph). First, we recall that the discrete eigenvalues κ_n are independent of t. Next, we multiply (10) by ϕ_n and integrate the result over all x; also, on using (9), it is readily found that

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{\infty} \phi_n^2 \, \mathrm{d}x = C_n \int_{-\infty}^{\infty} \phi_n^2 \, \mathrm{d}x,$$

where the constant C in (10) here must be indexed with n to become C_n. But then the normalization condition (12) implies that $C_n = 0$. Now substitute (11) into (10) to show that

$$\frac{\mathrm{d}c_n}{\mathrm{d}t} = 4\kappa_n^3 c_n$$
$$\text{so that} \quad c_n(t) = c_n(0) \exp(\kappa_n^3 t). \tag{15}$$

For the continuous spectrum, the asymptotic expressions (13), (14) are substituted into (10). Now it is found that the constant $C(k) = 4\mathrm{i}k^3$, and that

$$\frac{\mathrm{d}R}{\mathrm{d}t} = 8\mathrm{i}k^3 R$$
$$\text{so that} \quad R(k; t) = R(0; k) \exp(8\mathrm{i}k^3 t). \tag{16}$$

Similarly, it can be shown that $T(k; t) = T(k; 0)$.

The final step is the inversion of the scattering data at time t to recover the potential $u(x, t)$ in 9). This is accomplished through the Marchenko integral equation for the function $K(x, y)$

$$K(x, y) + F(x + y) + \int_x^{\infty} K(x, z) F(y + z) \, \mathrm{d}z = 0. \tag{17}$$

Here the function $F(x)$ is known in terms of the scattering data at time t,

$$F(x) = \sum_{n=1}^{N} c_n^2(t) \exp(-\kappa_n x) + \frac{1}{2\pi} \int_{-\infty}^{\infty} R(k; t) \exp(\mathrm{i}kx) \, \mathrm{d}k. \tag{18}$$

Here the t-dependence of K, F has been suppressed as the linear integral equation (17) is solved with t fixed. Then

$$u(x, t) = 2\frac{\partial}{\partial x}\{K(x, x; t)\}, \tag{19}$$

where the t-dependence in K has been restored.

The inverse scattering transform described by (17), (18) enables one to find the solution of the KdV equation (1) for an arbitrary localized initial condition. The most important outcome is that as $t \to \infty$, the solution evolves into N rank-ordered solitons propagating to the right ($x > 0$), and some decaying radiation propagating to the left ($x < 0$),

$$u \sim \sum_{n=1}^{N} 2\kappa_n^2 \operatorname{sech}^2(\kappa_n(x - 4\kappa_n^2 t - x_n)) + \text{radiation}. \tag{20}$$

Here the N solitons are derived directly from the discrete spectrum, where each eigenvalue $-\kappa_n$ generates a soliton of amplitude $2\kappa_n^2$, while the phase shifts x_n are determined from the constants $c_n(0)$. The continuous spectrum is responsible for the decaying radiation, which decays at each fixed $x < 0$ as $t^{-1/3}$.

The important special case when the reflection coefficient $R(k) \equiv 0$ leads to the N-soliton solution, for which there is no radiation. Indeed, the N-soliton solution can be obtained as an explicit solution of the Marchenko equation (17). We illustrate the procedure for $N = 1, 2$. First, for $N = 1$, $F(x) = c^2 \exp \kappa(x - 4\kappa^2 t)$, where we have omitted the subscript $n = 1$ for simplicity. Then seek a solution of

(17) in the form $K(x, y, t) = L(x, t)\exp(-\kappa y)$, where L can be found by simple algebra. The outcome is that

$$L(x,t) = \frac{-2\kappa c(0)^2 \exp(-\kappa x + 8\kappa^2 t)}{2\kappa + c(0)^2 \exp(-2\kappa x + 8\kappa^2 t)}.$$

Finally u is found from (19),

$$u = 2\kappa^2 \operatorname{sech}^2(\kappa(x - 4\kappa^2 t) - x_1).$$

This is just the solitary wave (2) of amplitude $2\kappa^2$; the phase shift x_1 is such that $c(0)^2 = 2\kappa\exp(2\kappa x_1)$. The procedure for $N = 2$ follows a similar course. Thus, with $R \equiv 0, N = 2$ in F (18), seek a solution of the Marchenko equation (17) in the form $K(x, y, t) = L_1(x, t)\exp(-\kappa_1 y) + L_2(x, t)\exp(-\kappa_2 y)$, and again $L_{1,2}$ can be found by simple algebra. The outcome is the two-soliton solution. For instance, with $\kappa_1 = 1, \kappa_2 = 2$, this is

$$u = 12\frac{3 + 4\cosh(2x - 8t) + \cosh(4x - 64t)}{[3\cosh(x - 28t) + \cosh(3x - 36t)]^2}. \quad (21)$$

It can be readily shown that

$$u \sim 8\operatorname{sech}^2(2(x - 16t \mp x_2) + 2\operatorname{sech}^2(x - 4t \mp x_1) \quad \text{as} \quad t \to \pm\infty, \quad (22)$$

where the phase shifts $x_{1,2} = (-\frac{1}{2}, \frac{1}{4})\ln 3$. Thus, the two-soliton solution describes the elastic collision of two solitons, in which each survives the interaction intact, and the only memory of the collision is the phase shifts; note that $x_1 < 0, x_2 > 0$, so that the larger soliton is displaced forward and the smaller soliton is displaced backward. The general case of an N-soliton is analogous and is essentially a sequence of pair-wise two-soliton interactions.

The integrability of the KdV equation (1) is also characterized by the existence of an infinite set of independent conservation laws. The most transparent conservation laws are

$$\int_{-\infty}^{\infty} u\,\mathrm{d}x = \text{constant}, \quad (23)$$

$$\int_{-\infty}^{\infty} u^2\,\mathrm{d}x = \text{constant}, \quad (24)$$

$$\int_{-\infty}^{\infty} \left(u^3 - \frac{1}{2}u_x^2\right)\mathrm{d}x = \text{constant}, \quad (25)$$

which may be associated with the conservation of mass, momentum, and energy, resepectively. Indeed (23) is obtained from the KdV equation (1) by integrating over x, while (24), (25) are obtained in an analogous manner after first multiplying (1) by u, u^2, respectively. However, it transpires that these are just the first three conservation laws in an infinite set, where each successive conservation law contains a higher power of u than the preceding one. This may be demonstrated using the inverse scattering transform (see Ablowitz & Segur, 1981; Dodd et al., 1982; Newell, 1985). However, here we use the original method based on the Miura transformation as adapted by Gardner. The Miura transformation is

$$u = -v_x - v^2, \quad (26)$$

$$v_t - 6v^2 v_x + v_{xxx} = 0. \quad (27)$$

Here (27) is the modified KdV equation. Direct substitution of (26) into the KdV equation shows that if v solves the modified KdV equation (27), then u solves the KdV equation (1). This discovery was the starting point for the discovery of the inverse scattering transform, since if one considers (26) as an equation for v and writes $v = \phi_x/\phi$, followed by a Galilean transformation for u (i.e., $u \to u - \lambda, x \to x - 6\lambda t$), one obtains the spectral problem (9). Here we follow a different route and write

$$v = \frac{1}{2\varepsilon} - \varepsilon w,$$

which (after a shift $x \to x + 3t/2\varepsilon^2$) converts the mKdV equation (27) into the Gardner equation (7) with $\delta = -\varepsilon^2$. Apart from a constant, which may be removed by a Galilean transformation, the corresponding expression for u is the Gardner transformation,

$$u = w + \varepsilon w_x - \varepsilon^2 w^2. \quad (28)$$

Thus, if w solves the Gardner equation

$$w_t + 6ww_x - 6\varepsilon^2 w^2 w_x + w_{xxx} = 0, \quad (29)$$

then u solves the KdV equation (1).

Next, we observe that the Gardner equation (29) has the conservation law

$$\int_{-\infty}^{\infty} w\,\mathrm{d}x = \text{constant}. \quad (30)$$

Since $w \to u$ as $\varepsilon \to 0$, we write the formal asymptotic expansion

$$w \sim \sum_{n=0}^{\infty} \varepsilon^n w_n.$$

It follows from (30) that then

$$\int_{-\infty}^{\infty} w_n\,\mathrm{d}x = \text{constant},$$

for each $n = 0, 1, 2, \ldots$. But substitution of this same asymptotic expansion for w into (28) generates a sequence of expressions for w_n in terms of u, of which the first few are

$$w_0 = u\,, \quad w_1 = -u_x\,, \quad w_2 = -u^2 + u_{xx}.$$

Thus, we see that $n = 0, 2$ give the conservation laws (23), (24), respectively, while $n = 1$ is an exact differential. It may now be shown that all even values of

n yield nontrivial and independent conservation laws, while all odd values of n are exact differentials.

The KdV equation belongs to a class of nonlinear wave equations, which have Lax pairs and are integrable through an inverse scattering transform. It shares with these equations several other remarkable features, such as the Hirota bilinear form, Bäcklund transformations, and the Painlevé property. Detailed descriptions of these and other properties of the KdV equation can be found in the other entries and referenced texts.

Solitary Waves in a Variable Environment

In a variable environment, the governing equation which replaces (1) is the variable-coefficient KdV equation (8). In general, this is not an integrable equation and is usually solved numerically. However, there are two distinct limiting situations in which some analytical progress can be made. First, let it be supposed that the coefficients $\alpha(x), \beta(x)$ in (8) vary rapidly with respect to the wavelength of a solitary wave, and then consider the case when these coefficients make a rapid transition from the values α_-, β_- in $x < 0$ to the values α_+, β_+ in $x > 0$. Then a steady solitary wave can propagate in the region $x < 0$, given by

$$\begin{aligned} U &= a\operatorname{sech}^2(\gamma(\theta - Wx)), \\ \text{where}\quad W &= \frac{\alpha_- a}{3} = 4\beta_-\gamma^2. \end{aligned} \tag{31}$$

It will pass through the transition zone $x \approx 0$ essentially without change. However, on arrival into the region $x > 0$, it is no longer a permissible solution of (8), which now has constant coefficients α_+, β_+. Instead, with $x = 0$, expression (31) now forms an effective initial condition for the new constant-coefficient KdV equation. Using the spectral problem (9) and the inverse scattering transform, the solution in $x > 0$ can now be constructed; indeed, in this case, the spectral problem (9) has an explicit solution (e.g., Drazin & Johnson, 1989). The outcome is that the initial solitary wave *fissions* into N solitons and some radiation. The number N of solitons produced is determined by the ratio of coefficients $R = \alpha_+\beta_-/\alpha_-\beta_+$. If $R > 0$ (i.e., there is no change in polarity for solitary waves), then $N = 1 + [((8R + 1)^{1/2} - 1)/2]$ ($[\cdots]$ denotes the integral part); as R increases from 0, a new soliton (initially of zero amplitude) is produced as R successively passes through the values $m(m + 1)/2$ for $m = 1, 2, \ldots$. But if $R < 0$ (i.e., there is a change in polarity), no solitons are produced and the solitary wave decays into radiation. For instance, for water waves, $c = (gh)^{1/2}$, $\mu = 3c/2h$, $\lambda = ch^2/6$, $\sigma = c$, and so $\alpha = 3/(2hc^{1/2})$, $\beta = h^2/(6c^2)$, where h is the water depth. It can then be shown that a solitary water wave propagating from a depth h_- to a depth h_+ will fission into N solitons where N is given as above with $R = (h_-/h_+)^{9/4}$; if $h_- > h_+$, $N \geq 2$, but if $h_- < h_+$ then $N = 1$ and no further solitons are produced (Johnson, 1973).

Next, consider the opposite situation when the coefficients $\alpha(x), \beta(x)$ in (8) vary slowly with respect to the wavelength of a solitary wave. In this case a multiscale perturbation technique (see Grimshaw, 1979, or Grimshaw & Mitsudera, 1993) can be used in which the leading term is

$$U \sim A\operatorname{sech}^2\gamma\left(\theta - \int_{x_0}^{x} W\,\mathrm{d}x\right), \tag{32}$$

where

$$W = \frac{\alpha A}{3} = 4\beta\gamma^2. \tag{33}$$

Here the wave amplitude $a(x)$ and, hence, also $W(x), \gamma(x)$ are slowly varying functions of x. Their variation is most readily determined by noting that the variable-coefficient KdV equation (8) possesses a conservation law,

$$\int_{-\infty}^{\infty} U^2\,\mathrm{d}\theta = \text{constant}, \tag{34}$$

which expresses conservation of wave-action flux. Substitution of (32) into (34) gives

$$\begin{aligned} &\frac{2A^2}{3\gamma} = \text{constant}, \\ &\text{so that}\quad A = \text{constant}\left(\frac{\beta}{\alpha}\right)^{1/3}. \end{aligned} \tag{35}$$

This is an explicit equation for the variation of the amplitude $A(x)$ in terms of $\alpha(x), \beta(x)$. However, the variable-coefficient KdV equation (8) also has a conservation law for mass,

$$\int_{-\infty}^{\infty} U\,\mathrm{d}\theta = \text{constant}. \tag{36}$$

Thus, although the slowly varying solitary wave conserves wave-action flux, it cannot simultaneously conserve mass. Instead, it is accompanied by a trailing shelf of small amplitude but long length scale given by U_{s}, so that the conservation of mass gives

$$\int_{-\infty}^{\phi} U_{\mathrm{s}}\,\mathrm{d}\theta + \frac{2A}{\gamma} = \text{constant},$$

where $\phi = \int_{x_0}^{x} W\,\mathrm{d}x$ ($\theta = \phi$ gives the location of the solitary wave) and the second term is the mass of the solitary wave (32). Differentiation then yields the amplitude $U_- = U_{\mathrm{s}}(\theta = \phi)$ of the shelf at the rear of the solitary wave,

$$U_- = \frac{3\gamma_x}{\alpha\gamma^2}. \tag{37}$$

This shows that if the wavelength γ^{-1} increases (decreases) as the solitary wave deforms, then the trailing shelf amplitude U_- has the opposite (same) polarity as the solitary wave. Once U_- is known, the full shelf $U_s(\theta, x)$ is found by solving (8) with the boundary condition that $U_s(\theta = \phi) = U_-$ (see El & Grimshaw, 2002, where it is shown that the trailing shelf may eventually generate secondary solitary waves).

For a solitary water wave propagating over a variable depth $h(x)$, these results show that the amplitude varies as h^{-1}, while the trailing shelf has positive (negative) polarity relative to the wave itself accordingly as $h_x < (>)\ 0$. A situation of particular interest occurs if the coefficient $\alpha(x)$ changes sign at some particular location (note that in most physical systems the coefficient β of the linear dispersive term in (8) does not vanish for any x). This commonly arises for internal solitary waves in the coastal ocean, where typically in the deeper water, $\alpha < 0$, $\beta > 0$ so that internal solitary waves propagating shorewards are waves of depression. But in shallower water, $\alpha > 0$ and so only internal solitary waves of elevation can be supported. The issue then arises as to whether an internal solitary wave of depression can be converted into one or more solitary waves of elevation as the critical point, where α changes sign, is traversed. This problem has been intensively studied (see, for instance, Grimshaw et al., 1998 and the references therein), and the solution depends on how rapidly the coefficient α changes sign. If α passes through zero rapidly compared with the local width of the solitary wave, then the solitary wave is destroyed and converted into a radiating wave train (see the discussion above in the first paragraph of this section). On the other hand, if α changes sufficiently slowly for the present theory to hold (i.e., (35) applies), we find that as $\alpha \to 0$, then $A \to 0$ in proportion to $|\alpha|^{1/3}$, while $U_- \to \infty$ as $|\alpha|^{-8/3}$. Thus, as the solitary wave amplitude decreases, the amplitude of the trailing shelf, which has the opposite polarity, grows indefinitely until a point is reached just prior to the critical point where the slowly varying solitary wave asymptotic theory fails. A combination of this trailing shelf and the distortion of the solitary wave itself then provide the appropriate "initial" condition for one or more solitary waves of the opposite polarity to emerge as the critical point is traversed. However, it is clear that in situations, as here, where $\alpha \approx 0$, it will be necessary to include a cubic nonlinear term in (8), thus converting it into a variable-coefficient Gardner equation (cf. (7)). This case has been studied by Grimshaw et al. (1999), who show that the outcome depends on the sign of the coefficient (ν) of the cubic nonlinear term at the critical point. If $\nu > 0$, so that solitary waves of either polarity can exist when $\alpha = 0$, then the solitary wave preserves its polarity (i.e., remains a wave of depression) as the critical point is traversed. On the other hand, if $\nu < 0$, so that no solitary wave can exist when $\nu = 0$, then the solitary wave of depression may be converted into one or more solitary waves of elevation.

ROGER GRIMSHAW

See also **Inverse scattering method or transform; Kadomtsev–Petviashvili equation; Solitons; Water waves**

Further Reading

Ablowitz, M.J., Kaup, D.J., Newell, A.C. & Segur, H. 1973. Method for solving the sine-Gordon equation. *Physcs Letters*, 30: 1262–1264

Ablowitz, M.J., Kaup, D.J., Newell, A.C. & Segur, H. 1974. The inverse scattering transform–Fourier analysis for nonlinear problems. *Studies in Applied Mathematics*, 53: 249–315

Ablowitz, M.J. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM

Boussinesq, M.J. 1871. Theórie de l'intumescence liquid appellée onde solitaire ou de translation, se propageant dans un canal rectangulaire. *Comptes Rendus Acad. Sci (Paris)*, 72: 755–759

Boussinesq, M.J. 1877. Essai sur la theorie des eaux courantes, *Memoires presentees par diverse savants a l'Academie des Sciences Inst. France (Series 2)* 23: 1–680

Dodd, R.K., Eilbeck, J.C., Gibbon, J.D. & Morris, H.C. 1982. *Solitons and Nonlinear Wave Equations*, London: Academic

Drazin, P.G. & Johnson, R.S. 1989. *Solitons: An Introduction*, Cambridge and New York: Cambridge University Press

El, G.A. & Grimshaw, R. 2002. Generation of undular bores in the shelves of slowly-varying solitary waves. *Chaos*, 12: 1015–1026

Gardner, C.S., Greene, J.M., Kruskal, M.D. & Miura, R.M. 1967. Method for solving the Korteweg–de Vries equation. *Physical Review Letters*, 19: 1095–1097

Grimshaw, R. 1979. Slowly varying solitary waves. I Korteweg–de Vries equation. *Proceedings of the Royal Society*, 368A: 359–375

Grimshaw, R. 2001. Internal solitary waves. In *Environmental Stratified Flows*, edited by Boston: Kluwer, Chapter 1: 1–28

Grimshaw, R. & Mitsudera, H. 1993. Slowly-varying solitary wave solutions of the perturbed Korteweg–de Vries equation revisited. *Studies in Applied Mathematics*, 90: 75–86

Grimshaw, R., Pelinovsky, E. & Talipova, T. 1998. Solitary wave transformation due to a change in polarity. *Studies in Applied Mathematics*, 101: 357–388

Grimshaw, R., Pelinovsky, E. & Talipova, T. 1999. Solitary wave transformation in a medium with sign-variable quadratic nonlinearity and cubic nonlinearity. *Physica* D, 132: 40–62

Johnson, R.S. 1973. On the development of a solitary wave moving over an uneven bottom. *Proceedings of the Cambridge Philosophical Society*, 73: 183–203

Kadomtsev, B.B. & Petviashvili, V.I. 1970. On the stability of solitary waves in weakly dispersive media. *Soviet Physics Doklady*, 15: 539–541

Korteweg, D.J. & de Vries, H. 1895. On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philosophical Magazine*, 39: 422–443

Lax, P.D. 1968. Integrals of nonlinear equations of evolution and solitary waves. *Communications in Pure and Applied Mathematics*, 21: 467–490

Miles, J.W. 1980. Solitary waves. *Annual Review of Fluid Mechanics* 12: 11–43

Newell, A.C. 1985. Solitons in mathematics and physics. In *CBMS-NSF Series in Applied Mathematics*, Vol. 48, edited by Philadelphia: SIAM

Pego, R.L. & Weinstein, M.J. 1997. Convective linear stability of solitary waves for Boussinesq equations. *Studies in Applied Mathematics*, 99: 311–375

Rayleigh, Lord. 1876. On waves. *Philosophical Magazine* 1: 257–279

Russell, J.S. 1844. Report on waves, *14th Meeting of the British Association for the Advancement of Science*, pp. 311–390

Wadati, M. 1972. The exact solution of the modified Korteweg-de Vries equation. *Journal of the Physical Society of Japan*, 32: 62–69

Zabusky, N.J. & Kruskal, M.D. 1965. Interactions of solitons in a collisionless plasma and the recurrence of initial states. *Physical Review Letters*, 15: 240–243

Zakharov, V.E. & Shabat, A.B. (1972). Exact theory of two-dimensional self focussing and one dimensional self-modulation of waves in nonlinear media. *Soviet Physics JETP*, 34: 62–69

KRYLOV–BOGOLYUBOV METHOD

See **Quasilinear analyses**

KURAMOTO–SIVASHINSKY EQUATION

Derived by Yoshiki Kuramoto and Takeo Tsuzuki in the context of reaction-diffusion systems (Kuramoto & Tsuzuki, 1975, 1976), and by Gregory Sivashinsky in the study of flame front propagation (Sivashinsky, 1977), the Kuramoto–Sivashinsky equation (KS) is a prime example of a system that possesses a rich variety of spatial and temporal behaviors. The mathematical and statistical (thermodynamic) analyses and classification of both elementary solutions and observed complex behaviors have generated many advances in the understanding of the often complex patterns that arise in many experimental simulations.

The KS equation in one spatial dimension may be written

$$u_t + uu_x + u_{xx} + u_{xxxx} = 0, \qquad (1)$$

where the subscripts denote the partial derivatives w.r.t. the space variable x, and the time variable t. One important class of solutions is defined by a periodic boundary condition, say, with a $2\pi L$-periodicity. Here we may interpret such solutions as being restricted to a cell of size $2\pi L$ and the cellular state size serving as the control (or bifurcation) parameter. For small values of $L > 0$, the solutions of (1) behave periodically; then as L is increased, the system passes through a wide variety of behavior including exhibiting spatiotemporal chaos. Via a simple rescaling, the parameter dependence can be made explicit in (1), and the solutions then satisfy $u(x,t) = u(x + 2\pi, t)$. Equation (1) possesses both translation and Galilean symmetries.

Physically, the form of Equation (1) models the small perturbations from a reference Poiseuille flow of a film layer on an inclined plane (Pumir et al., 1983). The name KS often refers in the literature to the closely related equation formed by letting $u = w_x$ in (1) and integrating. Higher-dimensional equivalents of this form are found in the independent derivations of Kuramoto and Sivashinsky. The first is in the context of angular-phase turbulence for a system of reaction-diffusion equations modeling the Belouzov–Zabotinsky reaction in three spatial dimensions. Here, the solution is considered to be a small perturbation of a global periodic solution. The Sivashinsky laminar flame front derivation models small thermal diffusive instabilities, where the solution is the perturbation of an unstable planar front in the direction of propagation. Further derivations of the system can be found in the review by Nicolaenko (1986) and references therein.

The KS equation has been shown to posses an inertial manifold, that is, the infinite-dimensional solution space of the system is spanned by the solutions to a coupled system of ordinary differential equations with a low number of degrees of freedom (Foias et al., 1988). Hence, the system may be effectively studied by Fourier mode expansion, where the number of Fourier modes determining the dynamics is proportional to L.

The nontrivial behavior of the KS equation stems from the linear instability of the laminar state $u(x,t) = \text{constant}$—the evolution of the system is governed by the quadratic nonlinear coupling term uu_x and a second-order instability term u_{xx} that is balanced by the dissipation term u_{xxxx}. Long-wavelength k-numbered modes with $k < L$ are unstable. However, the linearly unstable low modes of the system are stabilized by the strong nonlinear coupling while the extremely stable high modes, with intermediate wavelengths with mode number $k \sim L$ play the important role of maintaining a “chaotic dynamical equilibrium.”

For $L \geq 1$, standing and traveling waves may coexist with solutions having complicated oscillatory behavior. For example, antisymmetrically pulsating standing waves and waves that change both form and velocity periodically with time are considered in Demekhin et al. (1991). There exist windows of the parameter L in which many of these cellular states are stable. Windows of intermittency and strange (chaotic) attractors are also observed. For L not too large, these complex motions can be reached quite suddenly after extremely long transients. Figure 1 displays typical evolution once transients have died away. The KS equation also possesses persistent homoclinic and heteroclinic saddle connections which can be effectively explained via symmetry arguments (Kevrekidis et al., 1990).

The KS equation may be damped by the addition of the term νu on the left-hand side of (1), where ν parameterizes the level of damping. For zero damping the KS equation exhibits highly chaotic motions for large L. In the transition to these weakly turbulent motions, temporal intermittency is observed. Here, chaotic motions passing close to simple and weakly unstable states will feel regularizing effects for a

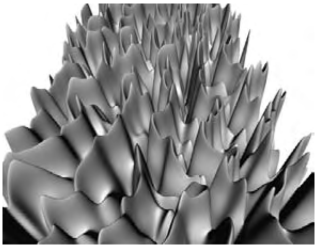

Figure 1. Spatiotemporally chaotic solution $u(x, t)$, computed from a 64-d Fourier truncation at $L = 10$. Coordinate x horizontal, amplitude u vertical, with plots of different times t overlapped at t progresses.

short while, for example. However, for large L and with damping, the intermittent behavior in the transition to weak turbulence also has a spatial element (Chaté & Manneville, 1986). In this scenario, a fluctuating mixture of both regular and chaotic patches with well-defined boundaries are observed in the solution surface $u(x, t)$. Choosing the parameters so that a weakly turbulent solution can be reached by spatiotemporal intermittency, one will observe chaotic domains slowly occupying the system; setting parameter values below such a threshold one will see the domains recede.

Sam Gratrix and John N. Elgin

See also **Belousov–Zhabotinsky reaction; Chaotic dynamics; Pattern formation; Turbulence**

Further Reading

Chang, H.-C. 1986. Travelling waves on fluid interfaces: normal form analysis of the Kuramoto–Sivashinsky equation. *The Physics of Fluids*, 29(10): 3142–3147

Chaté, H. & Manneville, P. 1986. Transition to turbulence via spatiotemporal intermittency. *Physical Review Letters*, 58(2): 112–115

Christiansen, F., Cvitanović, P. & Putkaradze, V. 1997. Spatiotemporal chaos in terms of unstable recurrent patterns. *Nonlinearity*, 10(1): 55–70

Demekhin, Y.A., Tokarev, G.Yu. & Shkadov, Ya.V. 1991. Hierarchy of bifurcations of space-periodic structures in a nonlinear model of active dissipative media. *Physica* D, 52 (2–3): 338–361

Elgin, J.N. & Wu, X. 1996. Stability of cellular states of the Kuramoto–Sivashinsky equation. *SIAM Journal on Applied Mathematics*, 56(6): 1621–1638

Foias, C., Nicolaenko, B., Sell, G.R. & Temam, R. 1988. Inertial manifolds for the Kuramoto–Sivashinsky equation and estimates of their dimension. *Journal de Mathématiques Pures et Appliquèes. Neuvième Série*, 67(3): 197–226

Hooper, P.A. & Grimshaw, R. 1988. Travelling wave solutions of the Kuramoto–Sivashinsky equation. *Wave Motion*, 10(5): 405–420

Kent, P. 1992. *Bifurcations of the travelling-wave solutions of the Kuramoto–Sivashinsky equation*, PhD Thesis, University of London

Kevrekidis, J.G., Nicolaenko, B. & Scovel, J.G. 1990. Back in the saddle again: A computer assisted study of the Kuramoto–Sivashinsky equation. *SIAM Journal on Applied Mathematics*, 50(3): 760–790

Kuramoto, Y. & Tsuzuki, T. 1975. On the formation of dissipative structures in reaction–diffusion systems. *Progress of Theoretical Physics*, 54(2): 687–699

Kuramoto, Y. & Tsuzuki, T. 1976. Persistent propagation of concentration waves in dissipative media far from thermal equilibrium. *Progress of Theoretical Physics*, 55(2): 356–369

Michelson, D. 1986. Steady solutions of the Kuramoto–Sivashinsky equation. *Physica* D, 19(1): 89–111

Nicolaenko, B. 1986. Some mathematical aspects of flame chaos and flame multiplicity. *Physica D*, 20(1): 109–121

Nicolaenko, B., Scheurer, B. & Temam, R. 1985. Some global dynamical properties of the Kuramoto–Sivashinsky equations: nonlinear stability and attractors. *Physica* D, 16(2): 155–183

Pumir, A., Manneville, P. & Pomeau, Y. 1983. On solitary waves running down an inclined plane. *Journal of Fluid Mechanics*, 135: 27–50

Sivashinsky, G.I. 1977. Nonlinear analysis of hydrodynamic instability in laminar flames: I. Derivation of basic equations. *Acta Astronautica*, 4(11–12): 1177–1206

L

LABORATORY MODELS OF NONLINEAR WAVES

In the decade following his 1834 discovery of the hydrodynamic solitary wave on Edinburgh's Union Canal, John Scott Russell constructed a water tank, allowing nonlinear wave phenomena to be studied in a laboratory environment (Russell, 1844). Among other results of these experiments, Russell observed, first, that the speed (v) of a solitary wave is related to its height (h) by the empirical relation $v=\sqrt{g(d+h)}$, where d is the resting depth of the water and g is the acceleration of gravity. Second, two solitary waves pass smoothly through each other without scattering. Third, two or more solitary waves can be generated from a sufficiently large "initial heap" of water. In the case sketched in Figure 1, for example, a volume (V) of water is released (by raising a sliding panel at the left-hand side of the tank) that is sufficiently large to generate two hydrodynamic solitons but not large enough to generate three of them. Also the soliton of larger amplitude is observed to have a higher velocity, leading to a separation between the two components that increases with time.

Since Russell's seminal work, hydrodynamic wave tanks have been widely used to investigate nonlinear wave propagation in a variety of settings, and several tanks suitable for undergraduate laboratories have been developed and described (Bettini et al., 1983; Olsen et al., 1984; Remoissenet, 1999). Tank experiments allow students to quantitatively investigate various properties of the Korteweg–de Vries (KdV) equation, which can be written in normalized form as

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}+u\frac{\partial u}{\partial x}+\frac{\partial^3 u}{\partial x^3}=0, \tag{1}$$

where $u(x,t)$ represents the vertical displacement of the wave from its resting level. Such experiments include quantitative comparisons of the number of solitons produced by an initial amount of water V as determined by eigenvalues of the corresponding time-independent linear Schrödinger equation through the inverse scattering transform method (Olsen et al., 1984). Similar tank experiments on deep water (where the depth is much larger than the lateral extent of the waves) allow quantitative studies of the nonlinear Schrödinger equation (Remoissenet, 1999).

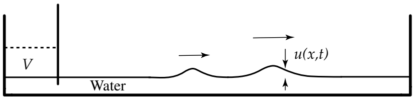

Figure 1. A sketch of John Scott Russell's wave tank, generating two solitons.

In addition to wave tanks, mechanical wave models have been constructed for other nonlinear systems, including mechanical models of the normalized sine-Gordon (SG) equation

$$\frac{\partial^2 u}{\partial x^2}-\frac{\partial^2 u}{\partial t^2}=\sin u. \tag{2}$$

As shown in the model of Figure 2, a number of pendula (dressmaker pins) are connected to a longitudinal spring (elastic band), whereupon $u(x,t)$ is the angle of rotation of the pendulum located at position x as a function of time t. The first term in Equation (2) represents the elastic restoring torque between adjacent pendula, the second term represents their angular acceleration, and the right-hand term is the angular-dependent torque of gravity. With a bit of practice, this simple model allows one to observe and demonstrate kink propagation, kink-kink collisions, breathers, and kink-antikink annihilation (Scott, 1969, 1970). The latter, in turn, is a model for electron-positron annihilation in elementary-particle physics.

More detailed mechanical models of the SG equation have been designed and constructed, which are suitable for undergraduate laboratory experiments (Scott, 1969, 1970; Remoissenet, 1999). As is evident from Figure 3, such models allow quantitative studies of the Lorentz contraction experienced by a kink as it approaches the limiting speed (Mach 1) of the system. For research purposes, Matteo Cirillo developed a mechanical model of fluxon propagation on a long Josephson junction, including an adjustable torque on the pendula (from air jets) that models the bias current acting in a

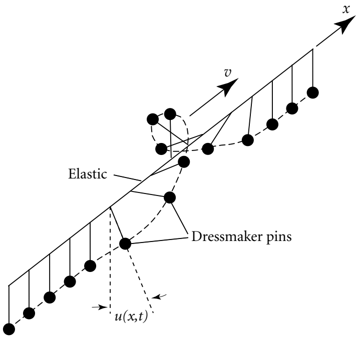

Figure 2. A kink on a simple mechanical model of the sine-Gordon equation that can be made from dressmaker pins and an elastic band.

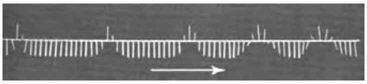

Figure 3. Strobe photos of an SG kink propagating on a mechanical model that was designed for student experiments (Scott, 1969). The kink is traveling to the right and slowing down due to friction. Spacing between the pendula is 1.59 cm and the time between successive images is 0.6 s.

typical experiment (Cirillo et al., 1981). Also, Michel Remoissenet and his colleagues have developed a model of the nonlinear Klein–Gordon equation with a double-well potential (the "phi-fourth model") to study the properties of compactons and the propagation of domain walls in ferroelectric and ferromagnetic materials (Duseul, 1998).

Complementing the family of mechanical models for nonlinear wave phenomena are nonlinear electrical transmission lines (Scott, 1970; Ostrovsky et al., 1972; Lonngren, 1978; Remoissenet, 1999). In energy-conserving versions of these models, nonlinearity is usually introduced through voltage-dependent capacitors (varactor diodes), and models of the KdV equation, Boussinesq equations, and the Toda lattice are readily constructed.

Allowing energy dissipation, it has long been known that the candle (or dynamite fuse or Japanese incense) models the leading edge of a nerve impulse as described by the Zeldovich–Frank-Kamenetsky (ZF) equation

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} = u(u-a)(u-1), \tag{3}$$

where $u(x,t)$ represents temperature of a candle flame or transmembrane voltage of a nerve impulse. This equation can also be modeled by a nonlinear electrical transmission in which the transverse (shunt) conductive element is nonlinear with a region of negative slope (negative differential conductance) as provided by an Esaki tunnel diode or by a Giaever-type superconductive diode (Scott, 1970; Remoissenet, 1999).

In the 1920s, R.S. Lilly showed that nerve impulse propagation can be modeled by a piece of iron wire immersed in a strong nitric- or sulfuric-acid solution (Lilly, 1925). At rest, the wire is stabilized (passivated) by a thin oxide layer that prevents further oxidation. If this passivated layer is disturbed by mechanical or electrical means, however, a deoxidized region propagates along the wire, followed by re-establishment of the stabilizing oxide layer. In this manner, the iron-wire model simulates the recovery property of biological nerves that is missed by Equation (3).

A more complete electrical representation of impulse conduction along a nerve fibre was developed by Jin-ichi Nagumo and his colleagues in collaboration with Richard FitzHugh (Nagumo et al., 1962). Known as the FitzHugh–Nagumo equation, this model augments the ZF equation to allow for recovery to the initial resting state. Finally, the neuristor is an electronic device that functions like a nerve fiber and can be used as the basic element in a family of computing elements.

Alwyn Scott

See also **FitzHugh–Nagumo equation; Korteweg–de Vries equation; Neuristor; Sine-Gordon equation; Solitons, a brief history; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Bettini, A., Minelli, T.A. & Pascoli, D. 1983. Solitons in an undergraduate laboratory. *American Journal of Physics*, 51: 977–984

Cirillo, M., Parmentier, R.D. & Savo, B. 1981. Mechanical analog studies of a perturbed sine–Gordon equation. *Physica* D, 3: 565–576

Duseul, S., Michaux, P. & Remoissenet, M. 1998. From kinks to compacton-like kinks. *Physical Review* E, 57: 2320–2326

Lilly, R.S. 1925. Factors affecting the transmission and recovery in the passive iron wire nerve model. *Journal of General Physiology*, 7: 473–507

Lonngren, K.E. 1978. Obsrvations of solitons on nonlinear dispersive transmission lines. In *Solitons in Action*, edited by K.E. Lonngren and A.C. Scott, New York: Academic Press

Nagumo, J., Arimoto, S. & Yoshizawa, S. 1962. An active pulse transmission line simulating nerve axon. *Proceedings of IRE*, 50: 2061–2071

Olsen, M., Smith, H. & Scott, A.C. 1984. Solitons in a wave tank. *American Journal of Physics*, 52: 826–830

Ostrovsky, L.A., Papko, V.V. & Pelinovsky, E.N. 1972. Solitary electromagnetic waves in nonlinear lines. *Radiophysics and Quantum Electronics*, 15: 438–446

Remoissenet, M. 1999. *Waves Called Solitons: Concepts and Experiments*, 3rd edition, Berlin and New York: Springer

Russell, J.S. 1844. Report on waves. *14th Meeting of the British Association for the Advancement of Science*, London: BAAS: pp. 311–339

Scott, A.C. 1969. A nonlinear Klein–Gordon equation. *American Journal of Physics*, 37: 52–61

Scott, A.C. 1970. *Active and Nonlinear Wave Propagation in Electronics*, New York: Wiley

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

LAGRANGE–EULER EQUATIONS

See **Euler–Lagrange equations**

LAGRANGIAN DESCRIPTION

See **Fluid dynamics**

LAMB DIAGRAM

See **Bäcklund transformations**

LANDAU EQUATION

See **Equilibria**

LANDAU–LIFSHITZ EQUATION

The principal assumption of macroscopic ferromagnetism theory is that a magnetic crystal state is described unambiguously by the magnetization vector $\boldsymbol{M}$, so the dynamics and kinetics of a ferromagnet are determined by variations in its magnetization. The magnetization of a ferromagnet as a function of space coordinates and time $\boldsymbol{M}(\boldsymbol{x}, t)$ is a solution of the Landau–Lifshitz (LL) equation, which was first used by Soviet scientists Lev Landau and Evgeni Lifshitz for describing the dynamics of a small velocity domain wall, the magnetic susceptibility of ferromagnets with a domain structure, and ferromagnetic resonance (Landau & Lifshitz, 1935). Later, the macroscopic theory of spin waves as small vibrations of the magnetization vector was developed on the basis of the linearized LL equation (Akhieser et al., 1968). At present, the LL equation is the theoretical foundation of phenomenological magnetization dynamics in magnetically ordered solids, including ferromagnets, antiferromagnets, and ferrites.

The LL equation in a ferromagnet has the following form:

$$\frac{\partial \boldsymbol{M}}{\partial t} = -\frac{2\mu_0}{\hbar}[\boldsymbol{M} \times \boldsymbol{H}_{\text{eff}}] - \gamma[\boldsymbol{M} \times [\boldsymbol{M} \times \boldsymbol{H}_{\text{eff}}]], \quad (1)$$

where μ_0 is the Bohr magneton and γ is the relaxation constant determining the damping motion of the vector $\boldsymbol{M}$. The effective magnetic field $\boldsymbol{H}_{\text{eff}}$ is equal to the variational derivative of the magnetic crystal energy E with respect to the vector $\boldsymbol{M}$: $\boldsymbol{H}_{\text{eff}} = -\,\delta E/\delta \boldsymbol{M}$. The energy E is assumed to be a function of $\boldsymbol{M}$, its spatial derivatives, and depends on the external magnetic field $\boldsymbol{H}$.

In the case of a many-sublattice magnet, an LL equation for the magnetization of an α-sublattice $\boldsymbol{M}^{\alpha}$ coincides with Equation (1) after the replacement $\boldsymbol{M} \to \boldsymbol{M}^{\alpha}$.

Equation (1) for ferromagnets has an integral of motion $M^2 = M_0 =$ constant and is consistent with the assumption that the $\boldsymbol{M}$-vector length in a ferromagnet is its equilibrium parameter. In the ground state, the quantity M_0 is equal to a so-called spontaneous magnetization $M_0 = 2\mu_0 S/a^3$, where S is the atomic spin and a is the interatomic separation. Conservation of the $\boldsymbol{M}$-vector length allows one to rewrite the LL equation in angular variables that are convenient for describing the magnetization dynamics in ferromagnets with axial symmetry. Assuming the external magnetic field directed along the anisotropy axis (the z-axis), let us define (Figure 1)

$$\begin{aligned} M_x + \mathrm{i}M_y &= M_0 \sin\theta \exp \mathrm{i}\psi, \\ M_z &= M_0 \cos\theta. \end{aligned} \quad (2)$$

Then in a dissipativeless case ($\gamma = 0$)

$$\begin{aligned} \sin\theta \frac{\partial\theta}{\partial t} &= -\frac{2\mu_0}{\hbar M_0}\frac{\delta E}{\delta\varphi}, \\ \sin\theta\left(\frac{\partial\varphi}{\partial t} - \frac{2\mu_0 H}{\hbar}\right) &= \frac{2\mu_0}{\hbar M_0}\frac{\delta E}{\delta\theta}, \end{aligned} \quad (3)$$

where the magnetic energy E is written as a function of the angular variables.

The magnetic energy E of a ferromagnet includes two parts: the exchange energy E_{ex} and the magnetic anisotropy energy E_{a}:

$$\begin{aligned} E_{\text{ex}} &= \frac{1}{2}\alpha \int \frac{\partial \boldsymbol{M}}{\partial x_k}\frac{\partial \boldsymbol{M}}{\partial x_k}\,\mathrm{d}^3x, \\ E_a &= -\frac{1}{2}\beta_1 \int M_x^2\,\mathrm{d}^3x - \frac{1}{2}\beta_3 \int M_z^2\,\mathrm{d}^3x, \end{aligned} \quad (4)$$

where α is the nonuniform exchange energy constant and β_1 and β_3 are uniaxial anisotropy constants. (When all $\beta = 0$, we have an isotropic ferromagnet; when $\beta_1 = 0$, we have a uniaxial ferromagnet, the easy-axis anisotropy corresponds to $\beta = \beta_3 > 0$ and easy-plane anisotropy to $\beta < 0$; and when $\beta_1 \neq 0$ and $\beta_3 \neq 0$, we have biaxial anisotropy.)

Equations (3) for easy-axial ferromagnets can be written in the form

$$\begin{aligned} & l_0^2 \Delta\theta - \left(1 + l_0^2(\nabla\varphi)^2\right)\sin\theta\cos\theta \\ & \quad + \frac{1}{\omega_0}\left(\frac{\partial\varphi}{\partial t} - \omega_H\right)\sin\theta = 0, \end{aligned} \quad (5)$$

$$l_0^2 \operatorname{div}(\sin^2\theta\nabla\varphi) - \frac{1}{\omega_0}\frac{\partial\theta}{\partial t}\sin\theta = 0, \quad (6)$$

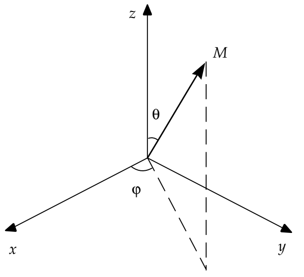

Figure 1. The angular variables for the magnetization $\boldsymbol{M}$.

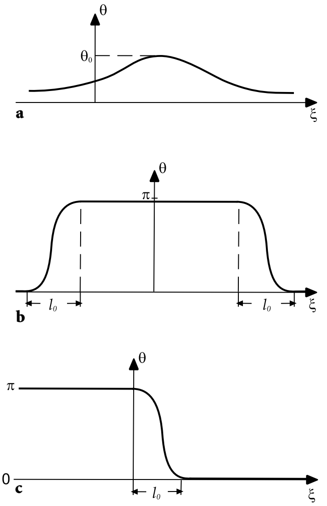

Figure 2. Magnetization distribution in (a) a low-amplitude soliton, (b) a magnetic soliton for small values V and ω, (c) a domain wall.

where $\Delta = \nabla^2$ is the Laplacian, l_0 is the magnetic length ($l_0^2 = \alpha/\beta$), ω is the homogeneous ferromagnetic resonance frequency ($\hbar\omega_0 = 2\mu_0\beta M_0$), and $\omega_H = 2\mu_0 H/\hbar$ is the spin magnetic frequency.

Equations (5) and (6) have two additive integrals of motion, namely, the total momentum of the magnetization field $\boldsymbol{P}$ and z-component (projection) of the total magnetic moment written after normalization as a number of spin deviations N in the excited state of the magnet:

$$\boldsymbol{P} = -\frac{\hbar M_0}{2\mu_0}\int (1-\cos\theta)\nabla\varphi \, \mathrm{d}^3x,$$
$$N = \frac{M_0}{2\mu_0}\int (1-\cos\theta)\, \mathrm{d}^3x. \qquad (7)$$

In a slightly excited state of ferromagnets ($\theta = \theta_0 =$ constant, $\theta_0 \ll 1$, $\psi = \omega t - \boldsymbol{kr}$) the magnetization dynamics is equivalent to a set of precessional spin waves that are solutions of linearized Equations (5), (6) with the following dispersion relation:

$$\omega(\boldsymbol{k}) = \omega_0 + \omega_H + (l_0 k)^2. \qquad (8)$$

The simplest static solution of Equations (5), (6) is a domain wall (Figure 2c)

$$\sin\theta = \mathrm{sech}[(x - x_0)/l_0], \qquad (9)$$

separating two semispace ferromagnet domains at $x < x_0$ and $x > x_0$. Equation (9) presents a topological soliton of the LL equation.

Interesting nontopological soliton solutions of Equations (5) and (6) are two-parametric dynamic magnetic solitons. Magnetic solitons of a general type are given by the following solutions

$$\theta = \theta(\boldsymbol{r} - \boldsymbol{V}t)\,, \varphi = (\omega + \omega_H)t + \psi(\boldsymbol{r} - \boldsymbol{V}t), \quad (10)$$

where the function θ vanishes at infinity ($\theta = 0$ for $r = \infty$) and $\nabla\psi$ is limited for $r = \infty$, $\boldsymbol{V}$ is the translational velocity of the soliton, and $\omega + \omega_H$ is the precessional frequency of the magnetization vector in the frame of reference moving along with the soliton. Typical diagrams for the function θ in the 1-d case are presented in Figure 2. There are the following Hamilton equations for a dynamic soliton (Kosevich et al., 1977):

$$\boldsymbol{V} = \left(\frac{\partial E}{\partial \boldsymbol{P}}\right)_N, \quad \hbar\omega = \left(\frac{\partial E}{\partial N}\right)_{\boldsymbol{P}}. \qquad (11)$$

The 1-dimensional version of Equations (5) and (6) is a totally integrable nonlinear equation in both isotropic (Lakshmanan, 1977; Takhtajan, 1977) and easy-axial (Borovik, 1978) cases and possesses a set of multiple-soliton solutions. The total integrability of 1-d LL equations in the case of biaxial anisotropy was proved by Sklyanin (1979).

If θ is small enough, the 1-d version of Equations (5) and (6) can be reduced to the nonlinear Schrödinger (NLS) equation for the complex function $\psi = M_x + \mathrm{i}M_y$ (Volzhan et al., 1976)

$$\frac{\mathrm{i}}{\omega_0}\frac{\partial\Psi}{\partial t} - l_0^2\frac{\partial^2\Psi}{\partial x^2} + (1+h)\Psi - \frac{1}{2M_0}|\Psi|^2\Psi = 0, \quad (12)$$

which is used for description of small amplitude long-wave excitations in one-dimensional easy-axial ferromagnets.

Another limiting case arises for a ferromagnet with the biaxial anisotropy ($\beta_1 < 0$ and $\beta_3 > 0$) under the condition $\varepsilon = -\beta_1/\beta_3 \gg 1$. Then the plane YOZ plays easy-plane, the magnetization vector lies nearly in this plane, $\chi = \pi/2 - \theta \ll 1$, and the LL equation is transformed to the following equations:

$$l_0^2\frac{\partial^2\phi}{\partial x^2} - \frac{1}{\varepsilon\omega_0}\frac{\partial^2\phi}{\partial t^2} + \sin\phi\cos\phi = 0,$$

$$\omega_0 \chi = -\frac{1}{\varepsilon}\frac{\partial \phi}{\partial t}, \qquad (13)$$

where the new angle variable ϕ is introduced with the equation $M_x + \mathrm{i}M_y = M_0 \cos\chi \exp \mathrm{i}\phi$.

At present, the LL equation is widely used for models of nonlinear macroscopic dynamic phenomena such as spin waves in inhomogeneous or spatially limited magnets, magnetostatic vibrations, magnetization dynamics, interaction of magnetically ordered media with electromagnetic and elastic waves, nonlinear magnetization waves, and magnetic solitons.

Arnold Kosevich

See also **Domain walls; Ferromagnetism and ferroelectricity; Nonlinear Schrödinger equations; Sine-Gordon equation; Solitons**

Further Reading

Akhieser, A.I., Bar'yahtar V.G. & Peletminskii S.V. 1968. *Spin Waves*, Amsterdam: North-Holland

Borovik, A.E. 1978. N-soliton solutions of Landau-Lifshitz equation, *Pis'ma Zhurnal Experimental'noy i Teoreticheskoy Fiziki*, 28: 629 (in Russian); *JETP Letters*, 28: 581

Kosevich, A.M. 1986. Dynamical and topological solitons in ferromagnets and antiferromagnets. In *Solitons*, edited by S.E. Trullinger, V.E. Zakharov & V.L. Pokrovsky, Amsterdam: North-Holland

Kosevich, A.M., Ivanov, B.A. & Kovalev, A.S. 1977. Nonlinear localized magnetization wave of a ferromagnet as a bound-state of a large number of magnons. *Pis'ma Zhurnal Eksperimental'noy i Teoreticheskoy Fiziki* 25: 516 (in Russian); *JETP Letters*, 25: 486

Kosevich, A.M., Ivanov, B.A. & Kovalev, A.S. 1990. Magnetic solitons. *Physics Reports*, 194: 117

Lakshmanan, M. 1977. Continium spin system as an exactly solvable dynamic system. *Physics Letters* A, 61: 53

Landau, L.D. & Lifshitz, E.M. 1935. On the theory of the dispersion of magnetic permability in ferromagnet bodies. *Physikalische Zeischrift der Sowjetunion*, 8: 153

Lifshitz, E.M. & Pitaevskii, L.P. 1980. *Statistical Physics, Part 2*, Oxford: Pergamon Press

Sklyanin, E.K. 1979. On complete integrability of Landau-Lifshitz equation, *Preprit/ Academy of Sciences of USSR*, Leningrad Department Steklov Mathematical Institute, E-3

Takhtajan, L.A. 1977. Integration of the continium Heisenberg spin chain through the inverse scattering method, *Physics Letters* A, 64: 235

Volzhan, E.B., Giorgadze, N.P. & Pataraya A.D. 1976. Weakly nonlinear magnetization density waves in magnetically ordered media. *Fizika Tverdogo Tela*, 25: 516 (in Russian); *Soviet Physics, Solid State*, 18: 1487

LANGEVIN EQUATION

See **Ferromagnetism and ferroelectricity**

LANGMUIR WAVES

See **Nonlinear plasma waves**

LANGMUIR–BLODGETT FILMS

The calming effect of oil on water has been known for centuries, and one early account of the phenomenon, on clay tablets, dates from the 18th century BCE in Babylonia. In 1879, a provisional British patent was filed by John Shields, the proprietor of a Scottish linen mill, for a simple device for spreading oil from valves in undersea pipes to calm the waves at the entrances to harbors. However, the first account of an experiment to investigate this effect was probably that of Benjamin Franklin in 1773, who wrote the following in a letter to a colleague.

> At length being at Clapham, where there is on the common a large pond, which I observed one day to be very rough with the wind, I fetched out a cruet of oil, and dropped a little of it on the water, and there the oil, though not more than a teaspoonful, produced an instant calm over a space several yards square, which spread amazingly, and extended itself gradually till it reached the lee side, making all that quarter of the pond, perhaps half an acre as smooth as a looking-glass.

A quick calculation reveals that Franklin's oil film was 1–2 nm thick—about the same as the size of an oil molecule, but this implication was not realized for many years. Although Lord Rayleigh was the first to propose that such films were only one molecule in thickness, he was not able to make a direct measurement to confirm this. The simple equipment for monolayer studies, now known as a Langmuir trough, was first introduced by Agnes Pockels. In a letter to Lord Rayleigh in 1891, she described the methods that formed the foundation of monolayer research.

Irving Langmuir provided most of the early scientific evidence for the existence of monolayer films. In 1917, he published a substantial paper outlining the properties of such films on a water surface. Some years later, Katharine Blodgett, working with Langmuir at the General Electric Company Research Laboratories in Schenectady, New York, devised a method for transferring the floating monolayers onto solid surfaces. The resulting films now bear the name of these two researchers.

Until the outbreak of World War II, research into the properties of monolayer films flourished, the work being undertaken mainly by surface chemists. However, few uses were found for monolayer and multilayer structures so activity in the area declined. Interest was rekindled in the 1970s following some stimulating experiments on energy transfer in multilayer systems by Hans Kuhn working in Germany. At about this time, organic chemists became aware of the limited range of monolayer forming materials that was available. Novel electroactive compounds were synthesized (dyes, semiconductors, polymers), and a new series of investigations began. This coincided with the birth of molecular electronics, a new interdisciplinary research activity focused on the exploitation of organic materials in electronic and optoelectronic devices. Related thin film technologies, such as self-assembly and layer-by-layer electrostatic deposition, were also developed. By the end of the 20th century, a number of organic thin film deposition technologies and materials, suitable for fabricating organic molecular architectures, was available. The era of molecular nanotechnology had begun.

Materials that produce organized monomolecular layers on the surface of water invariably consist of molecules possessing both water-attracting (hydrophilic) and water-repelling (hydrophobic) chemical groups. Such organic compounds are called amphiphiles. One of the simplest materials suitable for forming such a monomolecular layer is stearic acid, $C_{17}H_{35}COOH$. The molecule consists essentially of 16 CH_2 groups forming a long hydrocarbon chain; one end of the chain terminates in a hydrophilic carboxylic acid COOH group.

Langmuir–Blodgett (LB) films are prepared by first depositing a small quantity of the amphiphilic material, dissolved in a volatile solvent such as chloroform, on the surface of carefully purified water (subphase). When the solvent has evaporated, the organic molecules may be compressed to form a floating two-dimensional solid. The hydrophilic and hydrophobic terminations of the molecules ensure that the individual molecules are aligned in the same way during this process. During compression the monolayer undergoes a number of phase transformations. The different phases are almost analogues of three-dimensional gases, liquids and solids. The phase changes may be readily identified by monitoring the surface pressure Π as a function of the area occupied by the film. This is the two-dimensional equivalent of the pressure versus volume isotherm for a gas/liquid/solid. Figure 1 shows such a plot for a hypothetical long-chain organic monolayer material (e.g., a long-chain fatty acid).

In the "gaseous" state (G in Figure 1), the molecules are far enough apart on the water surface that they exert little force on one another. As the surface area of the monolayer is reduced, the hydrocarbon chains will begin to interact. The "liquid" state that is formed is generally called the expanded monolayer phase (E). The hydrocarbon chains of the molecules in such a film are in a random, rather than a regular orientation, with their polar groups in contact with the subphase. As the molecular area is progressively reduced, condensed (C) phases may appear. There may be more than one of these, and the emergence of each condensed phase can be accompanied by constant pressure regions in the isotherm, as observed in the cases of a gas condensing to a liquid and a liquid solidifying. These regions will be associated with enthalpy changes in the monolayer. In the condensed monolayer states, the molecules are closely packed and are oriented with their hydrocarbon chains pointing away from the water surface. The area per molecule in such a state will be similar to the cross-sectional area of the hydrocarbon chain, that is, $\approx 0.19\,\mathrm{nm}^2\,\mathrm{molecule}^{-1}$.

The LB technique requires that the surface pressure and temperature of the floating monolayer are controlled so that the organic film is in a condensed and stable state. Figure 2 shows the commonest form of LB deposition. The substrate is hydrophilic and the first layer is transferred, like a carpet, as the substrate is raised vertically through the water. Subsequently, a monolayer is deposited on each traversal of the monolayer/air interface. As shown, these stack in a head-to-head and tail-to-tail pattern; this

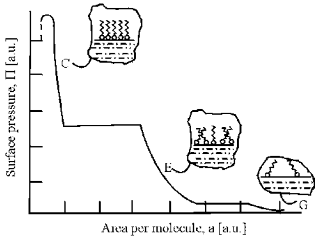

Figure 1. Surface pressure versus area per molecule for a long-chain organic compound. (The surface pressure and area are in arbitrary units (a.u.).)

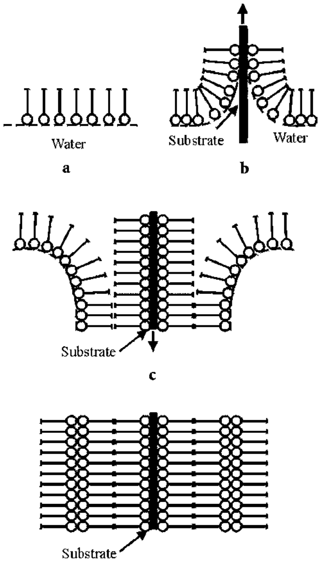

Figure 2. Y-type Langmuir–Blodgett film deposition.

deposition mode is called Y-type. Although this is the most frequently encountered situation, instances in which the monolayer is transferred to the substrate as it is being inserted into the subphase, or only as it is being removed, are often observed. These deposition modes are called X-type (monolayer transfer on the downstroke only) and Z-type (transfer on the upstroke only).

It is also possible to build up thin film architectures containing more than one type of monomolecular layer. In the simplest case, alternate-layer films may be produced by raising the substrate through a monolayer of one material (consisting of molecules of compound A, say) and then lowering the substrate through a monolayer of a second substance (compound B). An asymmetric multilayer structure consisting of ABABAB... layers is produced. This control over the molecular architecture permits the fabrication of organic superlattices with precisely defined symmetry properties. Such molecular assemblies can exhibit pyroelectric, piezoelectric, and second-order nonlinear optical phenomena.

Film transfer is characterized by measurement of the deposition ratio, τ (also called the transfer ratio). This is the decrease in the area occupied by the monolayer (held at constant pressure) on the water surface divided by the coated area of the solid substrate, that is,

$$\tau = A_{\mathrm{L}}/A_{\mathrm{S}}, \qquad (1)$$

where A_{L} is the area occupied by the monolayer on the water surface and A_{S} is the coated area of the solid substrate. Transfer ratios significantly outside the range 0.95–1.05 suggest poor film homogeneity.

A schematic diagram of one experimental arrangement to deposit LB films is shown in Figure 3, together with a photograph of the equipment. The Langmuir trough is made from PTFE (polytetrafluoroethylene) and a working area is defined by a PTFE-coated glass fiber barrier, which can be moved using a low-geared electric motor. The barrier motor is coupled to a sensitive electronic balance, which continuously monitors, via a sensing plate (Wilhelmy plate), the surface pressure of the monolayer. Using a feedback arrangement, this pressure can be maintained at a predetermined value. The physical dimensions of the Langmuir trough arrangement are not critical (the system in the photograph is approximately 30 cm in length) and are governed by the size of the substrate used.

Many analytical techniques, such as X-ray and electron diffraction, and infrared spectroscopy may be used to study the orientation of molecules in an LB assembly. Figure 4 shows an 9 nm × 9 nm atomic force micrograph of the surface of a 12-layer fatty acid LB film. The lighter parts of the image relate to the higher part of the surface and the darker regions correspond to deeper down. Lines of individual molecules are evident at the magnification shown, confirming the highly ordered arrangement of organic molecules in the LB film.

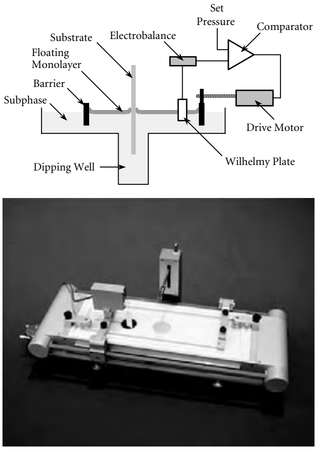

Figure 3. Langmuir–Blodgett trough (courtesy Molecular Photonics).

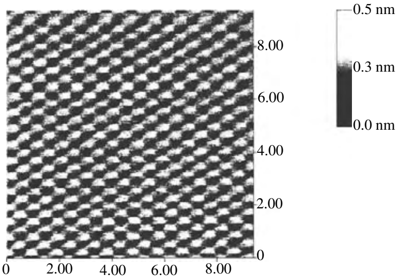

Figure 4. An atomic force micrograph of a fatty acid LB film.

The vertical dipping LB process is not the only way to transfer a floating molecular film to a solid substrate or to build-up multilayer films. Other methods are based on touching one edge of a hydrophilic substrate with the monolayer-covered subphase or lowering the substrate horizontally so that it contacts the hydrophobic ends of the floating molecules. Chemical means, for example, self-assembly of a thiol group onto a gold-coated substrate, can also be used to deposit monolayer organic films. In contrast, electrostatic layer-by-layer assembly relies on the forces between positively and negatively charged polyelectrolytes.

The earliest technical application of organic monolayer films is believed to be the Japanese printing art called *sumi-nagashi*. The dye comprising a suspension of submicron particles and protein molecules is first spread on the surface of water; the application of gelatin to the uniform layer converts the film into a patchwork of colorless and dark domains. These distinctive patterns can then be transferred by lowering a sheet of paper onto the water surface.

There are, of course, many methods available to deposit thin films of organic materials, including: thermal evaporation, sputtering, spaying, painting, dip- and spin-coating, electrodeposition, and molecular beam epitaxy. However, the LB technique is one of the few thin film technologies that actually permits the manipulation of materials at the molecular level. It should, therefore, be appropriate for exploitation by workers in nanotechnology wishing to fabricate interesting material architectures (bottom-up approach to nanotechnology) or to build up novel electronic device structures.

A range of possible applications for LB and related films is evident from the literature. Many of the ideas exploit the physical and chemical properties of the ultra-thin films to provide surface coatings with particular catalytic, adhesive, or mechanical properties (e.g., low friction). The availability of new polymeric amphiphiles has led to an interest in semi-permeable membranes. The extreme thinness of monolayer and multilayer films could provide key benefits in a variety of chemical sensing structures. For gas sensing, adequate sensitivities to some important gases and vapors have already been achieved using a variety of transduction techniques (chemiresistor, surface plasmon resonance, acousto-electric coupling, and so on).

For commercial exploitation, it is imperative to establish those areas in which LB films offer significant advantages over layers produced by other (and perhaps cheaper) means. In the case of second-order nonlinear optics, the LB method offers a means of aligning the molecules in a film of micrometer dimensions. Materials with high second-order hyperpolarizabilities (for example, leading to significant second-harmonic generation) already exist. Further work is needed on the development of practical electro-optic structures with attention to important considerations such as encapsulation and device degradation.

It is also interesting to note that a large number of biological materials form monolayers on a water surface. Chlorophyll a, the green pigment in higher plants; vitamins A, E and K; and cholesterol are all examples. Monomolecular films resemble naturally occurring biological membranes, which are based on a bilayer arrangement of long-chain phospholipid molecules. The LB technique might, therefore, be used as a means to fabricate artificial structures that emulate certain biological functions, such as photosynthesis, molecular recognition, or parallel information processing.

MICHAEL PETTY

See also **Bilayer lipid membranes; Liquid crystals; Nonlinear optics**

Further Reading

Gaines, G.L., Jr. 1966. *Insoluble Monolayers at Liquid-Gas Interfaces*, New York: Wiley-Interscience
Petty, M.C. 1996. *Langmuir–Blodgett Films: An Introduction*, Cambridge: Cambridge University Press
Roberts, G. G. (editor). 1991. *Langmuir–Blodgett Films*, New York: Plenum Press
Tredgold, R.H. 1994. *Order in Organic Films*, Cambridge and New York: Cambridge University Press
Ulman, A. 1991. *Ultrathin Organic Films*, San Diego: Academic Press

LAPLACE TRANSFORM

See **Integral transforms**

LASERS

An acronym for Light Amplification by Stimulated Emission of Radiation, a laser involves interaction of light with matter at the molecular, atomic, or nuclear levels and requires a quantum mechanical description of both light and matter. According to quantum mechanics, when an electron decays from an excited level E_+ to a lower energy level $E_- < E_+$, a photon of energy $E_+ - E_-$ and frequency $\nu = (E_+ - E_-)/h$ is emitted, where $h \doteq 6.672 \times 10^{-34}$ J s is Planck's constant.

The decay of electrons from excited states and the consequent emission of light is stimulated by the interaction of photons (of the appropriate energy) with the electrons (Einstein's stimulated emission). The emitted photons can then interact with excited electrons to produce a cascade of coherent photons, making up the familiar laser beam.

For a laser to work, three basic elements are required: (i) an energy level structure for the electrons (e.g., atomic or molecular fluids (gases or liquids), solids (crystals or glasses), or semiconducting junctures); (ii) a pumping mechanism (such as electrical current or light) to populate the upper energy level; and (iii) an optical cavity to partially confine the photons, enhancing their probability of interacting with the excited electrons.

The most common laser structures are Fabry–Perot cavities (where light is confined between two opposite mirrors) and ring lasers (in which light travels in one direction along the border of a rectangle with mirrors appropriately oriented in the corners) (Siegman, 1986) (see Figure 1). Industrial lasers are normally of Fabry–Perot type while ring lasers are often preferred in research laboratories.

The simplest laser model, describing only the first two required elements, consists of the laser rate equations, which model a laser medium in terms of the number of photons (proportional to the light

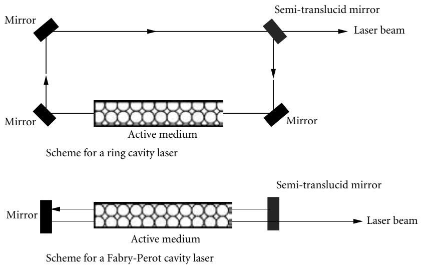

Figure 1. Ring and Fabry-Perot laser cavities.

intensity, I) and the number of excited electrons inside the laser cavity. Because the photon field can induce upward as well as downward transitions, the excited electrons are measured by the population inversion $N \equiv N_+ - N_-$, where N_+ (N_-) denotes the number of electrons in the upper (lower) level.

The photon population (I) is diminished due to the laser output as well as through other losses at a rate εI. Stimulated emission decreases the population inversion and increases the photon number at a rate αN_+, while absorption of photons produce the inverse effect at a rate αN_-. (The equivalence between emission and absorption rate per electron is a quantum mechanical result. Note that for each photon emitted, the population inversion decreases by two as N_+ decreases by one, while N_- increases by the same amount.) Excited electronic states (N_+) are also produced by the pumping mechanism at a rate $A/2$ (which may be a function of time) and diminished by nonradiative processes at a rate $\gamma(N - N_{\mathrm{eq}})$ (where N_{eq}, a negative number, derives from the population inversion at thermal equilibrium). A simple formulation of the laser rate equations is, therefore,

$$\frac{\mathrm{d}I}{\mathrm{d}t} = [-\varepsilon + \alpha N]\, I, \tag{1}$$

$$\frac{\mathrm{d}N}{\mathrm{d}t} = (A + \gamma N_{\mathrm{eq}}) - \gamma N - 2\alpha N I. \tag{2}$$

These laser rate equations have either one or two fixed points corresponding to steady operation of the laser. The laser-off state with $I = 0$ and $N = N_0 = A/\gamma + N_{\mathrm{eq}}$ is always present. The stability of the laser-off solution depends on the sign of $\lambda = -\varepsilon + \alpha N_0$. If this quantity is positive, small disturbances such as those produced by spontaneous emission are amplified and laser action develops. If $\lambda < 0$, disturbances die out, hence $\lambda = 0$ implies a threshold value for the pumping of $A = \gamma(\varepsilon/\alpha - N_{\mathrm{eq}})$. Steady operation of a laser above threshold is represented by the laser-on state $N = N_1 = \varepsilon/\alpha$ and $I = (\gamma/2\alpha)\,(N_0 - N_1)\,/N_1$.

These equations correspond to birth and death processes of photons and electrons that resemble those of predator and prey species in population dynamics. This analogy is somewhat tenuous, however, because quantum theory does not allow a photon or electron to be localized in time and energy simultaneously. Hence, the statistical description of a laser requires of entrained electron-photon states and the descriptions of quantum jump processes (Carmichael, 1999).

A more detailed set of equations is obtained by considering a laser operating in a single longitudinal cavity mode (i.e., having a fixed spatial pattern within the optical cavity). With a simple quantum mechanical model for electron energy levels and a classical description of the electromagnetic field (i.e., a large number of photons), the dynamics of the laser are described by Maxwell–Bloch equations, which involve complex amplitudes of the electric and polarization fields in addition to the population inversion (N) (Narducci & Abraham, 1988).

An aim of laser design has long been to achieve lasers of higher photon energies. While ultraviolet lasers were developed early, X-ray lasers have only recently been demonstrated (Rocca, 1999), and gamma-ray lasers are not yet realized.

Lasers can be classified in various ways according to their dynamics, the active media, the optical cavity, the pumping scheme, optical frequency, among others. Considering the active media, lasers can be grouped as follows.

Solid State Lasers

Solid state lasers, such as the ruby laser (the first laser, announced in 1960), which has a three-level pumping scheme for which Cr^{3+} impurities are essential, require a high-energy source for pumping and operate intermittently. In contrast, Nd-YAG (neodymium-doped yttrium aluminum garnet) and Nd-glass lasers are based in a four-level scheme, operate under moderate pump power, and display different dynamical regimes. Ruby and Nd-YAG lasers represent the two basic pumping schemes for solid state lasers.

In the three-level pumping scheme, the light emitting transition has the ground state as the final state, while in the four-level scheme the state reached by the electron after the transition is an excited state.

There are a large number of solid state lasers operating at different frequencies, each with particular features. Many of them employ a host crystal (or glass) doped with an impurity such as $Al_2O_3(Cr^{3+})$, $Y_2O_3(Eu^{3+})$, $Gd_2O_3(Nd^{3+})$, or $BaF_2(U^{3+})$.

Gas Lasers

Several active media fall in this category: helium-neon, argon, and the powerful CO_2 are among the best-known representatives. The standard method for the excitation of the molecular states is an electric discharge in the gaseous medium where collisions between exited ions provide the pumping mechanism. However, optical pumping and other pumping mechanisms have been demonstrated for some gas lasers.

Gas lasers can be operated in continuous wave and pulsed modes and are particularly important at the low-frequency (infrared) region of the electromagnetic spectrum. Other members of this class include H_2O, HCN, CH_2I, and HF lasers.

Liquid Lasers

The best-known representatives are dye lasers, where a stream of organic liquid is excited using visible or ultraviolet radiation from a flash lamp or a laser beam. The pumping mechanism involves radiative fluorescence at a lower frequency than that of the pumping source, a mechanism known as Stokes-shift.

Other lasers in this class are based on solutions of rare earth elements. In contrast with gas and solid state lasers, dye lasers usually operate in cavities with relatively high losses.

Semiconductor Lasers

Because of their small size, semiconductor lasers play an important role in electronic applications. While earlier lasers were built as semiconductor diodes using p–n junctions as in GaAs diodes (or PbS, PbSe, PbTe, SnTe, InSb, etc.) new generations of lasers, such as the Vertical Cavity Surface Emitting Laser (VCSEL) are produced modifying semiconductor surfaces and reducing the size and power requirements of the laser significantly. Semiconductor lasers are pumped by direct electrical current using the band structure of the semiconductor and electron-hole recomposition to produce the photons.

Nuclear Magnetic Resonance

The unique NMR laser relies on the (permanent) magnetic momentum of nuclei, which are pumped using a radio-frequency field. Optical pumping has also been demonstrated recently.

For gas lasers, some liquid lasers, and solid state lasers, the lasing threshold is reached only with high reflectivity mirrors minimizing the cavity losses (small ε). Such lasers operate just above threshold in a mode that is almost resonant with the frequency of the atomic transition and can be viewed as weakly nonlinear systems. Wide cavities, higher pumping levels, feedback, and other effects can induce multimode dynamics, where several empty cavity modes are required for the spatial description of the dynamics.

When laser devices are based on lossy electromagnetic cavities, as in dye and semiconductor lasers, the relation between cavity modes and monochromatic states (or solutions) is obscured. Nevertheless, these are also called multimode lasers when a description of the spatial extension is required.

Despite the diversity of active media, pumping mechanisms, and lasing frequencies, the dynamics of lasers are often rather similar, the main differences accounted for by different parameter values in Maxwell–Bloch type equations. The dynamics of most lasers are described by simple dynamical attractors (fixed points) under constant pumping. There are, however, situations in which complex behavior emerges; thus, optical reinjection of the emitted light can destabilize a laser. Such a process might occur when the laser is coupled to optic fibers in a communication system. The Lang–Kobayasi equations (Lang & Kobayashi, 1980) (a modification of the rate equations including a delayed field) describe this situation for very small amounts of reinjected signal. In the case of semiconductor lasers, even small amounts of optical feedback require a spatial description of the laser cavity (multimode operation) (Huyet et al., 1998; Duarte & Solari, 1998). Attempts to control, or mode lock, a (usually powerful) slave laser with a (usually weak) master laser also generate a rich dynamical spectrum at the unlocking transition (Zimmermann et al., 2001).

HERNÁN G. SOLARI AND MARIO NATIELLO

See also **Maxwell–Bloch equations; Nonlinear optics; Semiconductor laser**

Further Reading

Carmichael, H.J. 1999. *Statistical Methods in Quantum Optics*, vol. 1, *Master equations and Fokker–Plank equations*, Berlin and New York: Springer

Duarte, A.A. & Solari, H.G. 1998. The metamorphosis of the monochromatic spectrum in a double-cavity laser as a function of the feedback rate. *Physical Review* A, 58: 614–619

Huyet, G., Balle, S., Giudici, M., Green, C., Giacomelli, G. & Tredicce, J.R. 1998. Low frequency fluctuations and multi-mode operation of a semiconductor laser with optical feedback. *Optical Communications*, 149: 341–347

Lang, R. & Kobayashi, K. 1980. External optical feedback effects on semiconductor injection laser properties. *IEEE Journal of Quantum Electronics*, QE-16: 347

Narducci, L.M. & Abraham, N.B. 1988. *Laser Physics and Laser Instabilities*, Singapore: World Scientific

Rocca, J.J. 1999. Table-top soft X-ray lasers. *Review of Scientific Instruments*, 70: 3799

Siegman, A.E. 1986. *Lasers*, Mill Valley, CA: University Science Books

Zimmermann, M., Natiello, M. & Solari, H. 2001. Global bifurcations in a laser with inhected signal: beyond Adler's approximation. *Chaos*, 11: 500–513

LATTICE GAS METHODS

When one is interested in studying the dynamical behavior of fluid systems starting at the microscopic level, it is logical to begin with a molecular dynamics description of the interactions between the constituting particles. This is often a formidable task, as the fluid evolves into a nonlinear regime where chaos, turbulence, or reactive processes take place. But one may question whether a realistic description of the microscopic dynamics is necessary to gain insight on the underlying mechanisms of large-scale nonlinear phenomena. Around 1985, a considerable simplification was introduced (Frisch et al., 1986). These pioneering studies established (theoretically and computationally) the feasibility of simulating fluid dynamics via a microscopic approach based on a new paradigm. A virtual simplified micro-world is constructed as an automaton universe, based not on a realistic description of interacting particles but merely on the laws of symmetry and of invariance of macroscopic physics.

Suppose we implement point-like particles on a regular lattice where they move from node to node at each time step and undergo collisions when their trajectories meet at the same node. As the system evolves, we observe its collective dynamics by looking at the lattice from a distance. A remarkable fact is that—if the collisions occur according to some simple logical rules (satisfying fundamental conservations) and if the lattice has the proper symmetry—this "lattice gas automaton" (LGA) shows global behavior very similar to that of a real fluid. So we can infer that despite its simplicity at the microscopic scale, the LGA should contain the essential features that are responsible for the emergence of complex behavior and, thereby, can help us understand the basic mechanisms involved.

An LGA consists of a set of particles moving on a regular d-dimensional lattice $\mathcal{L}$ at discrete time steps, $t = n\Delta t$, with n an integer. The lattice is composed of V nodes, labeled by the d-dimensional position vectors $\boldsymbol{r} \in \mathcal{L}$. Associated to each node are b channels, labeled by indices $i, j, \ldots$, running from 1 to b. At a given time t, a channel can be either occupied by one particle or empty, so that the occupation variable $n_i(\boldsymbol{r}, t) = 1$ or 0. When channel i at node $\boldsymbol{r}$ is occupied, then the particle at the specified node $\boldsymbol{r}$ has velocity $\mathbf{c}_i$. The set of allowed velocities is such that the condition $\boldsymbol{r} + \boldsymbol{c}_i \Delta t \in \mathcal{L}$ is fulfilled. The "exclusion principle" requirement that the maximum occupation be of one particle per channel allows for a representation of the automaton configuration in terms of a set of bits $\{n_i(\boldsymbol{r}, t)\}$; $\boldsymbol{r} \in \mathcal{L}, i = \{1, b\}$. The evolution rules are thus simply logical operations over sets of bits.

The time evolution of the automaton takes place in two stages: propagation and collision. In the propagation phase, particles are moved according to their velocity vector, and in the (local) collision phase, the particles occupying a given node are redistributed among the channels associated to that node. So the microscopic evolution equation of the LGA reads

$$\begin{aligned} n_i(\boldsymbol{r} + \boldsymbol{c}_i\,\Delta t, t + \Delta t) \\ = n_i(\boldsymbol{r}, t) + \Delta_i(\{n_j(\boldsymbol{r}, t)\}), \end{aligned} \tag{1}$$

where $\Delta_i(\{n_j\})$ represents the collision term that depends on all channel occupations at node $\boldsymbol{r}$. By performing an ensemble average (denoted by angular brackets) over an arbitrary distribution of initial occupations, one obtains a hierarchy of coupled equations for the successive n-body distribution functions. This hierarchy can be truncated to yield the lattice Boltzmann equation for the single particle distribution function $f_i(\boldsymbol{r}, t) = \langle n_i(\boldsymbol{r}, t)\rangle$:

$$\begin{aligned} f_i(\boldsymbol{r} + \boldsymbol{c}_i\,\Delta t, t + \Delta t) - f_i(\boldsymbol{r}, t) \\ = \Delta_i^{\mathrm{Boltz}}(\{f_j(\boldsymbol{r}, t)\}). \end{aligned} \tag{2}$$

The left-hand side is recognized as the discrete version of the left-hand side of the classical Boltzmann equation for continuous systems, and the right-hand side denotes the collision term, where the precollisional uncorrelated state *ansatz* has been used to factorize the b-particle distribution function.

The lattice Boltzmann equation (2) is one of the most important results in LGA theory. It can be used as the starting point for the derivation (via multi-scale analysis) of the macroscopic equations describing the long wavelength behavior of the lattice gas. The LGA macroscopic equations are found to exhibit the same structure as the classical hydrodynamic equations, and under the incompressibility condition, one retrieves the Navier–Stokes equations for nonthermal fluids. Another important feature of the lattice Boltzmann equation is that it can be used as an efficient and powerful simulation algorithm. In practice, one usually prefers to use a simplified equation where the collision term is approximated by a single relaxation time process inspired by the Bhatnagar–Gross–Krook model, known in its lattice version as the LBGK equation:

$$\begin{aligned} f_i(\boldsymbol{r} + \boldsymbol{c}_i\,\Delta t, t + \Delta t) - f_i(\boldsymbol{r}, t) \\ = -\frac{1}{\tau}\left[f_i(\boldsymbol{r}, t) - f_i^{\mathrm{leq}}(\boldsymbol{r}, t)\right], \end{aligned} \tag{3}$$

where the right-hand side is proportional to the deviation from the local equilibrium distribution function.

There is a wealth of applications of the lattice gas methods that have established their validity and their usefulness. LGA simulations, based on Equation (1), are most valuable for fundamental problems in statistical mechanics, such as the study of fluctuation correlations in equilibrium and non-equilibrium systems (Rivet & Boon, 2001; Rothman & Zaleski, 1997). As an example, Figure 1 shows

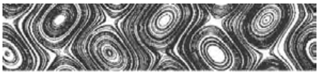

Figure 1. Lattice gas simulation of the Kolmogorov flow: the tracer trajectories reflect the topology of the ABC flow in the regime beyond the critical Reynolds number ($\mathrm{Re} = 2.5 \times \mathrm{Re_c}$).

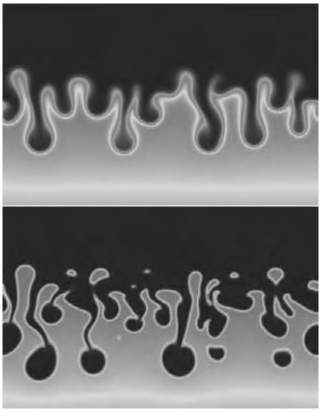

Figure 2. Lattice Boltzmann (LBGK) simulation of viscous fingering in miscible fluids (upper panel) showing the interface sharpening effect of a reactive process between the two fluids (lower panel).

the trajectories of tracer particles suspended in a Kolmogorov flow (above the critical Reynolds number) produced by a lattice gas automaton and from where turbulent diffusion was analyzed (Boon et al., 2000).

Simulations of more direct practical interest, such as, for instance, profile optimization in car design or turbulent drag problems, are most efficiently treated with the lattice Boltzmann method, in particular using the LBGK model. The example given in Figure 2 illustrates the method for the study of viscous fingering in Hele-Shaw geometry, showing the effect of reactivity between the two fluids as a determinant factor in the dynamics of the moving interface (Grosfils & Boon, 2002).

Applications of the LGA approach and of the lattice Boltzmann equation cover a wide variety of theoretical and practical problems ranging from the dynamics of thermal fluctuations and quantum lattice gas automata to multi-phase flow, complex fluids, reactive systems, and inhomogeneous turbulence.

JEAN PIERRE BOON

See also **Cellular automata; Hele-Shaw cell; Molecular dynamics; Navier–Stokes equation**

Further Reading

Boghosian, B.M., Coveney, P.V. & Emerton, A.N. 1996. A lattice gas model of microemulsions. *Proceedings of the Royal Society* A, 452: 1221–1250

Boon, J.P., Dab, D., Kapral, R. & Lawniczak, A. 1996. Lattice gas automata for reactive systems. *Physics Reports*, 273(2): 55–148

Boon, J.P., Hanon, D. & Vanden Eijnden, E. 2000. Lattice gas automaton approach to turbulent diffusion. *Chaos, Solitons and Fractals*, 11: 187–192

Coveney, P.V. & Succi, S. (editors). 2002. Discrete modeling and simulation of fluid dynamics. *Philosophical Transactions of the Royal Society*, 360: 291–573

Frisch, U., Hasslacher, B. & Pomeau, Y. 1986. Lattice gas automata for the Navier–Stokes equation. *Physical Review Letters*, 56: 1505–1508

Grosfils, P. & Boon, J.P. 2002. Viscous fingering in miscible, immiscible, and reactive fluids. *International Journal of Modern Physics* B, 17: 15–20

Meyer, D. 1996. From quantum cellular automata to quantum lattice gases. *Journal of Statistical Physics*, 85: 551–574

Rivet, J.P. & Boon, J.P. 2001. *Lattice Gas Hydrodynamics*, Cambridge and New York: Cambridge University Press

Rothman, D. & Zaleski, S. 1997. *Lattice Gas Cellular Automata*, Cambridge and New York: Cambridge University Press

Succi, S. 2001. *The Lattice Boltzmann Equation for Fluid Dynamics and Beyond*, Oxford: Clarendon Press and New York: Oxford University Press

LATTICE KINK

See **Solitons, types of**

LATTICE SOLITONS

See **Solitons, types of**

LAX OPERATORS

See **Inverse scattering method or transform**

LEADING EDGE DYNAMICS

See **Nerve impulses**

LEBESQUE MEASURE

See **Measures**

LEGENDRE TRANSFORMATION

See **Euler–Lagrange equations**

LENNARD–JONES POTENTIAL

See **Molecular dynamics**

LÉVY FLIGHTS

In 1937, the French mathematician Paul Lévy (1886–1971) introduced statistical descriptions of motion that extend beyond the more traditional Brownian motion discovered over one hundred years earlier. A diverse range of both natural and artificial phenomena is now being described in terms of Lévy statistics, from the flight of an albatross across the Antarctic skies to the trajectories followed by the abstract painter Jackson Pollock as he constructed his famous drip paintings.

In 1828, Robert Brown published his studies of the random motion of soot particles in a dish of water as they were buffeted from random directions by the thermal motion of water molecules. In 1905, Albert Einstein provided a theoretical basis for this diffusion process. A particle's Brownian motion is pictured as a sequence of jumps. For a single jump, the probability dependence on jump size x has a Gaussian distribution. A consequence of Gaussian statistics is that the size distribution for N jumps is also described by a Gaussian.

Paul Lévy generalized beyond Brownian motion by considering other distributions for which one jump and N jumps share the same mathematical form. These Lévy distributions decrease according to the power law $1/x^{1+\gamma}$ for large x values, where γ lies between 0 and 2. Compared with Gaussian distributions, Lévy distributions do not fall off as rapidly at long distances. For Brownian motion, each jump is usually small and the variance of the distribution, $\langle x^2 \rangle$, is finite. For Lévy motion, however, the small jumps are interspersed with longer jumps, or "flights," causing the variance of the distribution to diverge. As a consequence, Lévy jumps do not have a characteristic length scale.

This scale invariance is a signature of fractal patterns. Indeed, Lévy's initial question of "When does the whole look like its parts?" addresses the fractal property of self-similarity. An important parameter for assessing the scaling relationship of fractal patterns is dimension. What, then, is the dimension of the pattern traced out by Lévy motion? The short jumps making up Brownian motion build a clustered pattern that is so dense that area is a more appropriate measure than length—the pattern is actually two-dimensional. In contrast, although the short jumps of Lévy motion produce a clustering, the longer, less frequent jumps initiate new clusters. These clusters form a self-similar pattern with a dimension of less than two. Fractional dimensions are an exotic property of fractals.

Today, Lévy motion is as widely explored in nonlinear, chaotic, turbulent, and fractal systems as Brownian motion is in simpler systems. Following Mandelbrot's research in the 1970s demonstrating the prevalence of fractal patterns in nature, an increasing number of natural phenomena have been described using Lévy statistics (Mandelbrot, 1982). Lévy distributions are also having an impact on artificial systems. A recent example concerns nano-scale electronic devices in which chaotic electron trajectories produce Lévy statistics in the electrical conduction properties (Micolich et al., 2001). Other examples include diffusion in Josephson junctions (Geisel et al., 1985) and at liquid-solid interfaces (Stapf et al., 1995).

It is even possible to picture relatively simple systems in which both Brownian and Lévy motion appear and a transition between the two can be induced. Consider, for example, dropping tracer particles into a container of liquid. This, of course, is Brown's original experiment. In 1993, Harry Swinney extended this experiment by considering a rotating container of liquid shaped like a washer. As turbulence set in, vortices appeared in the liquid and Swinney's group showed that the tracer particles followed Lévy flights between the vortices with $\gamma = 1.3$ (Solomon et al., 1993).

In addition to spatial distributions, Lévy statistics can also be applied to distributions measured as a function of time. A famous example is the dripping faucet (*See* **Dripping faucet**). In 1995, Thadeu Penna's group showed that the fluctuations in the time intervals between drips follow a Lévy distribution with $\gamma = 1.66 - 1.85$. A significant appeal of this result lies in a comparison with earlier medical work by Ary Goldberger's group showing that fluctuations in the human heart beat follow $\gamma = 1.7$ (Goldberger, 1996). This prompted Penna to ask, "Is the heart a dripping faucet?"

Goldberger suggested that the Lévy statistics describing the human heart arise from nonlinear processes that regulate the human nervous system. He has since extended his research of the fractal dynamics of physiology to other examples of involuntary behavior. This includes studies of the human gait that show that fluctuations in stride intervals display fractal variations (Hausdorff et al., 1996). Fractal variations might therefore be a general signature of healthy human behavior, exhibited whenever conscious control is not involved.

It is interesting to consider this speculation within the context of the results of the British Antarctic Survey in 1996, which showed that albatrosses follow Lévy flights. Other species of animals, such as ants and bees, also follow Lévy flights when foraging for food. Due to the diverging variance of the flight distribution, Lévy trajectories represent an efficient way of covering large regions of space, especially when compared with Brownian motion. Significantly, these animal behavioral patterns represent yet another example of Lévy behavior generated by actions that are devoid of intellectual deliberation.

This relationship between unconscious actions and Lévy statistics has even touched on human creativity. In particular, the Surrealist art movement developed a technique called automatic painting, in which artists painted with such speed that any conscious involvement was thought to be eliminated. Jackson Pollock adopted this approach during the 1940–1950s when he dripped paint onto large horizontal canvases (see *Autumn Rhythm*, Figure 1). Remarkably, his paintings are fractal

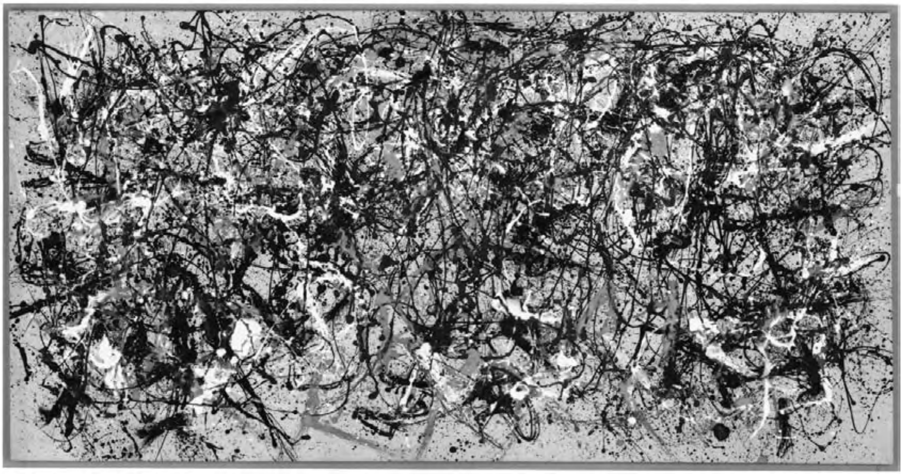

Figure 1. *Autumn Rhythm (Number 30)*, by Jackson Pollock, 1950. © 2003 The Pollock-Krasner Foundation / Artists Rights Society (ARS) New York. Courtesy The Metropolitan Museum of Art, George A. Hearn Fund, 1957. (57.92)

and his motions have been described in terms of Lévy flights (Taylor et al., 1999). This work triggered visual perception tests that identified an aesthetic preference for fractal patterns with dimensions between 1.3 and 1.5 (Taylor, 2001).

Lévy distributions represent a truly interdisciplinary phenomenon that will continue to be useful as novel artificial and natural systems are explored.

RICHARD TAYLOR

See also **Brownian motion; Dimensions; Dripping faucet; Fractals**

Further Reading

Geisel, T., Nierwetberg, J. & Zacherl, A. 1985. Accelerated diffusion in Josephson junctions and related chaotic systems. *Physical Review Letters*, 54: 616–620

Goldberger, A.L. 1996. Non-linear dynamics for clinicians: chaos theory, fractals, and complexity at the bedside. *The Lancet*, 347: 1312–1314

Haussdorff, J.M., Purdon, P.L., Peng, C.K., Ladin, Z., Wei, J.Y. & Goldberger, A.L. 1996. Fractal dynamics of human gait: stability of long-range correlations in stride interval fluctuations. *Journal of Applied Physiology*, 80: 1448–1457

Mandelbrot, B., 1982. *The Fractal Geometry of Nature*, San Francisco: Freeman

Micolich A.P., Taylor, R.P., Davies, A.G., Bird, J.P., Newbury, R., Fromhold, T.M., Ehlert, A., Linke, H., Macks, L.D., Tribe, W.R., Linfield, E.H., Ritchie, D.A., Cooper, J., Aoyagi, Y. & Wilkinson, P.B. 2001. The evolution of fractal patterns during a classical-quantum transition. *Physical Review Letters*, 87: 036802-1–036802-4

Solomon, T., Weeks, E. & Swinney, H. 1993. Observation of anomalous diffusion and Lévy flights in a two-dimensional rotating flow. *Physical Review Letters*, 71: 3975–3979

Stapf, S., Kimmich, R. & Seitter, R. 1993. Proton and deuteron field-cycling NMR relaxometry of liquids in porous glasses: evidence of Lévy-walk statistics. *Physical Review Letters*, 75: 2855–2859

Taylor, R.P., Micolich, A.P. & Jonas, D. 1999. Fractal analysis of Pollock's drip paintings. *Nature*, 399: 422 and *Physics World*, 12: 25–29

Taylor, R.P. 2001. Architect reaches for the clouds. *Nature*, 410: 18

LIE ALGEBRAS AND LIE GROUPS

Lie groups were introduced by the 19th century Norwegian mathematician Sophus Lie through his studies in geometry and integration methods for differential equations (Hawkins, 1999). Further developments by W. Killing, É. Cartan and H. Weyl established Lie's theory as a cornerstone of mathematics and its physical applications. General references include Duistermaat & Kolk (1999), Sattinger & Weaver (1986), and Varadarajan (1984).

An r parameter Lie group is defined as an r-dimensional manifold that is also a group with smooth multiplication and inversion maps. A key example is the $r = n^2$-dimensional general linear group $\mathrm{GL}(n)$ of (either real or complex) $n \times n$ nonsingular matrices, $\det A \neq 0$, under matrix multiplication. Most Lie groups can be realized as matrix groups, that is, subgroups of $\mathrm{GL}(n)$. Important examples include the

- special linear group $\mathrm{SL}(n) \subset \mathrm{GL}(n)$ with $\det A = 1$, and $r = n^2 - 1$;
- orthogonal group $\mathrm{O}(n) \subset \mathrm{GL}(n, \mathbb{R})$ with $A^{\mathrm{T}} A = I$, and $r = n(n-1)/2$;

- unitary group $\mathrm{U}(n) \subset \mathrm{GL}(n, \mathrm{C})$ with $A^\dagger A = I$, and $r = n^2$;
- symplectic group $\mathrm{Sp}(2n) \subset \mathrm{GL}(2n, \mathbb{R})$ with $A^{\mathrm{T}} J A = J = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$, and $r = n(2n+1)$.

A Lie algebra $\mathfrak{g}$ is a vector space equipped with a skew-symmetric, bilinear bracket $[\,\cdot,\cdot\,]\colon \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ that satisfies the Jacobi identity

$$[\,u, [\,v, w\,]\,] + [\,v, [\,w, u\,]\,] + [\,w, [\,u, v\,]\,] = 0.$$

The Lie algebra $\mathfrak{g}$ of left-invariant vector fields on an r-parameter Lie group G can be identified with the tangent space at the identity, and so $\dim \mathfrak{g} = r$. The Lie algebra $\mathfrak{gl}(n)$ of $\mathrm{GL}(n)$ consists of all $n \times n$ matrices under matrix commutator $[\,A, B\,] = AB - BA$. A finite-dimensional Lie algebra with basis $v_1, \ldots, v_r$ is specified by its structure constants c^i_{jk}, defined by the bracket relations $[\,v_j, v_k\,] = \sum_{i=1}^r c^i_{jk} v_i$. Each finite-dimensional Lie algebra corresponds to a unique connected and simply connected Lie group G^*; any other is obtained by quotienting by a discrete normal subgroup: $G = G^*/N$.

A subspace $\mathfrak{h} \subset \mathfrak{g}$ is a subalgebra if it is closed under the Lie bracket: $[\,\mathfrak{h}, \mathfrak{h}\,] \subset \mathfrak{h}$. Lie subalgebras are in one-to-one correspondence with connected Lie subgroups $H \subset G$. The subalgebra is an ideal if $[\,\mathfrak{g}, \mathfrak{h}\,] \subset \mathfrak{h}$. A Lie algebra is simple if it contains no nontrivial ideals, and semi-simple if it contains no nontrivial abelian (commutative) ideals. Semi-simple algebras are direct sums of simple algebras. A Lie algebra is solvable if the sequence of subalgebras $\mathfrak{g}^{(0)} = \mathfrak{g}$, $\mathfrak{g}^{(k+1)} = [\,\mathfrak{g}^{(k)}, \mathfrak{g}^{(k)}\,]$ eventually terminates with $\mathfrak{g}^{(n)} = \{0\}$. The Levi decomposition says that every Lie algebra is the semi-direct sum of a semi-simple subalgebra and its radical—the maximal solvable ideal.

The Killing–Cartan classification of complex simple Lie algebras contains four infinite families, denoted A_n, B_n, C_n, D_n, corresponding to the simple Lie groups $\mathrm{SL}(n+1), \mathrm{O}(2n+1), \mathrm{Sp}(2n+1), \mathrm{O}(2n)$. In addition, there are five exceptional simple Lie groups, G_2, F_4, E_6, E_7, E_8, of respective dimensions 14, 52, 78, 133, 248. The last plays an important role in modern theoretical physics. Extending the classification to infinite-dimensional simple Lie algebras leads to the Kac–Moody Lie algebras, of importance in integrable systems, theoretical physics, differential geometry, and topology (Kac, 1990).

Lie groups typically arise as symmetry groups of geometric objects or differential equations. A (Lie) group G acts on a manifold M, for example, Euclidean space, provided $m \mapsto g \cdot m$ for $g \in G$, $m \in M$, defines a sufficiently smooth invertible map that respects the group multiplication. Lie classified all transformation groups acting on one- and two-dimensional real and complex manifolds (Olver, 1995). According to Klein's Erlanger Program, geometric structure is prescribed by an underlying transformation group; thus, Euclidean, affine, conformal, projective geometries are based on the eponymous Lie groups. If G acts transitively, then $M = G/H$ is a homogeneous space, obtained by quotienting by a closed Lie subgroup. The group orbits—minimal invariant subsets—form a system of submanifolds, and the invariant functions are constant on orbits. The infinitesimal generators of the group action form a Lie algebra $\mathfrak{g}$ of vector fields tangent to the orbits whose flows generate the group action.

A linear action $\rho\colon G \to \mathrm{GL}(V)$ on a vector space V is known as a representation. Representation theory plays a fundamental role in quantum mechanics since linear symmetries of the Schrödinger equation induce actions on the space of solutions, which decompose into irreducible representations. The structure of atoms, nuclei, and elementary particles is governed by the representations of particular symmetry groups (Hamermesh, 1962). Important special functions, for example, Bessel and hypergeometric, arise as matrix entries of representations of particular Lie groups (Vilenkin & Klimyk, 1991). The representation theory of the orthogonal group SO(2) leads to trigonometric functions and, hence, Fourier analysis, as the simplest case of harmonic analysis on semi-simple Lie groups (Warner, 1972).

A Lie group acts on its Lie algebra $\mathfrak{g}$ by the adjoint representation and on the dual space $\mathfrak{g}^*$ by the coadjoint representation. The coadjoint orbits are symplectic submanifolds with respect to the natural Lie–Poisson structure on $\mathfrak{g}^*$, and are of importance in classifying representations (Kirillov, 1999), geometric mechanics, and geometric quantization (Woodhouse, 1992). The Euler equations of rigid body motion and of fluid mechanics are realized as the Lie–Poisson equations on, respectively, the Lie algebra of the Euclidean group and the infinite-dimensional diffeomorphism group (Marsden, 1992).

A transformation group is called a symmetry group of a system of differential equations if it maps solutions to solutions. Symmetry groups are effectively computed by solving the infinitesimal symmetry conditions, which form an overdetermined linear system of partial differential equations, usually amenable to automatic solution by computer algebra packages (Olver, 1993, 1995). Applications include integration of ordinary differential equations, determination of explicit group-invariant (similarity) solutions of partial differential equations, Noether's theorems relating symmetries of variational problems and conservation laws (Noether, 1918), bifurcation theory (Golubitsky & Schaeffer, 1985), asymptotics and blow-up (Barenblatt, 1979), and the design of geometric numerical integration schemes (Hairer et al., 2002).

Classification of differential equations and variational problems admitting a given symmetry group relies on its differential invariants. The simplest examples are the curvature and torsion of space curves, and the mean and Gaussian curvatures of surfaces under the Euclidean group acting on $\mathbb{R}^3$. Cartan's method of moving frames, and its more recent extensions to general Lie group and Lie pseudo-group actions (Olver, 2001), provides a general mechanism for construction and classification of differential invariants, with applications to differential geometry, the calculus of variations, soliton theory, computer vision, classical invariant theory, and numerical methods.

Modern developments in applications of Lie group methods have proceeded in a variety of directions. General theories of infinite-dimensional Lie groups and algebras (Kac, 1990) and Lie pseudo-groups, arising in relativity, field theory, fluid mechanics, solitons, and geometry, remain elusive. Higher order or generalized symmetries, in which the infinitesimal generators also depend upon derivative coordinates, first proposed by Noether (1918), have been used to classify integrable (soliton) systems. Recursion operators are used to generate such higher order symmetries and, via Noether's theorem, higher order conservation laws (Olver, 1993). Most recursion operators are derived from a pair of compatible Hamiltonian structures, and demonstrate the integrability of bi-Hamiltonian systems. The higher-order symmetries also appear in series expansions of Bäcklund transformations in the spectral parameter.

PETER J. OLVER

See also **Bäcklund transformations; Inverse scattering method or transform; Maps; Symmetry groups**

Further Reading

Barenblatt, G.I. 1979. *Similarity, Self-Similarity, and Intermediate Asymptotics*, New York: Consultants Bureau

Duistermaat, J.J. & Kolk, J.A.C. 1999. *Lie Groups*, Berlin and New York: Springer

Golubitsky, M. & Schaeffer, D.G. 1985. *Singularities and Groups in Bifurcation Theory*, 2 vols, New York: Springer

Hairer, E., Lubich, C. & Wanner, G. 2002. *Geometric Numerical Integration: Structure-preserving Algorithms for Ordinary Differential Equations*, Berlin and New York: Springer

Hamermesh, M. 1962. *Group Theory and Its Application to Physical Problems*, Reading, MA: Addison-Wesley, 1962

Hawkins, T. 1999. *The Emergence of the Theory of Lie Groups. An Essay on the History of Mathematics 1869–1926*, New York: Springer

Kac, V.G. 1990. *Infinite Dimensional Lie Algebras*, Cambridge and New York: Cambridge University Press

Kirillov, A.A. 1918. Merits and demerits of the orbit method, *Bulletin of the American Mathematical Society*, 36: 433–488

Marsden, J.E. 1992. *Lectures on Mechanics*, Cambridge and New York: Cambridge University Press

Noether, E. 1918. Invariante variationsprobleme. *Nachricheten von der Konig. Gesellschaft der Wissenschaften zu Gottingen, Mathematisch–Physikalische Klasse*, 235–257

Olver, P.J. 1993. *Applications of Lie Groups to Differential Equations*, 2nd edition, London and New York: Springer

Olver, P.J. 1995. *Equivalence, Invariants, and Symmetry*, Cambridge and New York: Cambridge University Press

Olver, P.J. 2001. Moving frames—in geometry, algebra, computer vision, and numerical analysis. In *Foundations of Computational Mathematics*, edited by R. DeVore, A. Iserles & E. Suli, Cambridge and New York: Cambridge University Press, pp. 267–297

Sattinger, D.H. & Weaver, O.L. 1986. *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*, New York: Springer

Varadarajan, V.S. 1984. *Lie Groups, Lie Algebras, and Their Representations*, New York: Springer

Vilenkin, N.J. & Klimyk, A.U. 1991. *Representation of Lie Groups and Special Functions*, Dordrecht: Kluwer

Warner, G. 1972. *Harmonic Analysis on Semi-simple Lie Groups*, Berlin and New York: Springer

Woodhouse, N.M.J. 1992. *Geometric Quantization*, 2nd edition, Oxford: Clarendon Press and New York: Oxford University Press

LIFETIME

As the word implies, *lifetime* is a measure of the time that something manages to persist and properly function. Often lifetime is a statistically defined quantity, as is evident from thinking about the failure of light bulbs. Clearly, one can exactly measure the lifetime of any particular bulb at the cost of burning it out, but it is more interesting to know the probability of failure in a given interval of time for a particular type of bulb. To obtain this information, it is necessary to measure the average burnout rate for a large number of bulbs that have been manufactured under identical conditions and use that information to characterize bulbs newly manufactured under the same conditions. A related approached is used by actuaries, who compute risks and premiums for human life insurance.

Failure of a Nuclear Power Plant

Statistical ideas are sometimes applied to more special situations, for example, a nuclear power plant, where neighbors ask: "What are the chances of it blowing up or melting down?" Engineers usually answer with soothing words. "It's very safe. This reactor can't blow up, and it has less than a fifty percent chance of melting down in the next thousand years." But the engineers have only built a few dozen plants, all of which seem to be working fine. How do they come up with such an estimate?

One procedure is to consider the various failure mechanisms (onset of corrosion, earthquakes, operators falling asleep, terrorist attacks, and so on), and assign a mean lifetime to each failure mechanism. Then the total failure probability up to time t_1 is defined as

$$P(t_1) = \int_0^{t_1} \frac{\mathrm{d}t}{\tau}, \tag{1}$$

where τ is the mean lifetime, taking all failure mechanisms into account. If these mechanisms are independent (do not influence one another), then

$$\frac{1}{\tau} = \frac{1}{\tau_1} + \frac{1}{\tau_2} + \cdots + \frac{1}{\tau_N}, \qquad (2)$$

where $\tau_1, \tau_2, \ldots, \tau_N$ are mean lifetimes estimated for N failure mechanisms. (Note that the τ_i may be functions of time.) A key result of this simple formulation is that the total lifetime is less than the smallest constituent lifetime, making it imperative to consider all possible failure mechanisms.

Lifetime of a Quantum State

Given a quantum state $\psi(\mathbf{r})$, where $\mathbf{r}$ is a spatial coordinate, it is interesting to know the mean time (τ) before ψ makes a transition (decays) into one of N other states: $\phi_1(\mathbf{r})$, $\phi_2(\mathbf{r})$, ..., $\phi_N(\mathbf{r})$. According to “Fermi’s Golden Rule”, the transition probability into ϕ_i is proportional to the square of an overlap integral (Slater, 1951; Dirac, 1935)

$$|\langle\psi|\phi_i\rangle|^2 \equiv \left[\int \psi\phi^* \, d\mathbf{r}\right]^2; \qquad (3)$$

thus the total lifetime is

$$\tau \propto \frac{1}{|\langle\psi|\phi_1\rangle|^2 + |\langle\psi|\phi_2\rangle|^2 + \cdots + |\langle\psi|\phi_N\rangle|^2}. \qquad (4)$$

Although this equation is simple to write, it is difficult to compute because computation requires detailed knowledge of the wave functions of all states into which ψ can decay. Just as with a nuclear power plant, the total lifetime will be less than the smallest individual decay time, but τ is easier to measure, as it is often a small faction of a second.

Soliton Lifetimes

When John Scott Russell first observed a hydrodynamic soliton on the “happiest day of [his] life” in August of 1834, he was struck by its remarkable stability, giving him time to follow it on horseback until it became “lost in the windings of the canal” (Russell, 1844). Although this and subsequent observations suggest that solitons are long-lived objects, the conclusion is somewhat misleading. It is the spatial localization of soliton energy that persists in time rather than the energy itself. Detailed perturbation calculations for the Korteweg–de Vries equation and for the nonlinear Schrödinger equation show that the times for energy decay under dissipative perturbations are less for nontopological solitons than for delocalized wave packets (Scott, 2003). Solitons of the sine-Gordon (SG) equation, on the other hand, have infinite lifetimes, as their decay is prevented by a topological constraint. Thus, the kinetic energy of a SG kink decays (the soliton slows to a stop), but its rest mass does not disappear. The property of topological stability is shared by Skyrmions, which makes these objects interesting as models for elementary particles.

In studies of nontopological solitons in molecular crystals and of biological solitons, lifetime is a central issue, as the energy must remain localized long enough to be of technical or biological interest. When such solitons are also quantum objects (as is often the case), accurate lifetime calculations require estimates of the type sketched in Equation (4). This is a daunting theoretical task for three reasons: the wave function of a quantum soliton (ψ) is quite complicated (Davydov, 1991; Scott, 2003), biological molecules are often irregular structures for which all the possible final states $\{\phi_i\}$ are not known, and the temperatures at which biological organisms exist (ca. 300 K) introduce structural uncertainties that depend irregularly on time.

Using femtosecond pump-probe spectroscopy, however, it has recently become feasible to measure lifetimes in molecular crystals of the order of a picosecond or more (Elsaesser et al., 2000). Importantly, the response of such observations is null unless nonlinear effects are present, which occurs only for localized energy; thus the measurements focus specifically on soliton-like entities. From data produced by such experiments, the relevance of various theoretical proposals for solitons in molecular crystals and biomolecules should become better understood in coming years.

Alwyn Scott

See also **Biomolecular solitons; Davydov soliton; Pump-probe measurements; Skyrmions**

Further Reading

Davydov, A.S. 1991. *Solitons in Molecular Systems*, 2nd edition, Dordrecht and Boston: Reidel

Dirac, P.A.M. 1935. *The Principles of Quantum Mechanics*, Oxford: Clarendon Press

Elsaesser, T., Mukamel, S., Murnanae, M.M. & Scherer N.F. (editors). 2000. *Ultrafast Phenomena XII*, Berlin: Springer

Russell, J.R. 1844. Report on waves. *14th Meeting of the British Association for the Advancement of Science*, London: BAAS: 311–339

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Slater, J.C. 1951. *Quantum Theory of Matter*, New York: McGraw-Hill

LIGHTHILL CRITERION

See **Wave stability and instability**

LINDSTEDT–POINCARÉ METHOD

See **Perturbation analysis**

LINEARIZATION

Linear systems are much easier to analyze than nonlinear systems because they satisfy the superposition principle, which can be stated as follows: if R_1 is the output of a linear system to input S_1 and R_2 is the output to input S_2, then the output to the cumulative input $aS_1 + bS_2$ is $aR_1 + bR_2$. Applied iteratively, this property allows the inputs and outputs of a linear system to be decomposed in various ways to simplify analysis. From a more general perspective, linear systems do not confuse the lines of causality between the input and output (stimulation and response) of a linear system. Fully nonlinear systems, on the other hand, do not share this useful property; thus, the causal implications of each stimulation must be individually studied.

Linearization is an attempt to carry some of the attractive properties of linear systems into the nonlinear domain. Linearization methods can be classified in two approaches: reductions of nonlinear systems to linearized approximations for small perturbations and extensions of linearized approximations to weakly nonlinear systems. Linearization methods of the first approach are used for stability analyses of special solutions of a nonlinear system, such as critical points, periodic orbits, solitary waves and periodic waves. In the neighborhood of such special solutions of nonlinear systems, linear equations are derived for small perturbations that are then studied using standard methods (Fourier or Laplace transforms, Green functions, etc.). Linearization methods of the second approach are based on specific parameters of a linearized system, such as dominant eigenvalues or resonant wave numbers. Thus, linearized systems are extended to weakly nonlinear systems, such as normal forms and amplitude equations, by means of asymptotic multiple scale expansion methods and Fredholm's alternative theorem for linear nonhomogeneous equations.

For systems of partial differential equations (PDEs), linearization methods can be illustrated with the example of a nonlinear Schrödinger (NLS) equation (Whitham, 1974):

$$\mathrm{i}u_t + u_{xx} - \left(\omega_0 + \omega_2|u|^2\right) u = 0, \tag{1}$$

where ω_0 and ω_2 are parameters such that $\omega_2 > 0$ and $\omega_2 < 0$ occur for defocusing and focusing cases, respectively. The NLS equation (1) has a constant wave solution of the form:

$$u_0(x,t) = a\ \mathrm{e}^{-\mathrm{i}\left(\omega_0+\omega_2 a^2\right)t}, \tag{2}$$

where the wave frequency $\omega = \omega_0 + \omega_2 a^2$ depends on the constant wave amplitude a. Linearizing the NLS equation (1) at the wave solution (2), one expands $u(x,t)$ as

$$u(x,t) = a\ \mathrm{e}^{-\mathrm{i}\left(\omega_0+\omega_2 a^2\right)t}\left[1 + V(x,t) + \mathrm{i}W(x,t)\right], \tag{3}$$

where V and W are real. When the expansion is substituted into the NLS equation and the nonlinear terms in V and W are neglected, the linearized problem can be transformed to the linear Boussinesq equation (4) with constant coefficients:

$$W_{tt} - 2\omega_2 a^2 W_{xx} + W_{xxxx} = 0. \tag{4}$$

The Fourier spectrum of the linear Boussinesq equation consists of two counter-propagating waves:

$$W(x,t) = W_+ \mathrm{e}^{\mathrm{i}kx - \mathrm{i}\Omega(k)t} + W_- \mathrm{e}^{\mathrm{i}kx + \mathrm{i}\Omega(k)t}, \tag{5}$$

where $\Omega(k) = k\sqrt{2\omega_2 a^2 + k^2}$. When $\omega_2 > 0$ (the defocusing case), the wave spectrum is stable, that is, $\Omega(k)$ is real for any k. When $\omega_2 < 0$ (the focusing case), the wave spectrum is unstable, that is, $\Omega(k)$ is complex for an interval of k, namely for $0 < k < \sqrt{2|\omega_2|}a$. Thus, the linearization method of the first approach enables us to check stability of the constant wave solutions of the NLS equation.

The linear Boussinesq equation (4) with $\omega_2 > 0$ is a hyperbolic system. With dispersive effects neglected in a long-wave approximation, the Boussinesq equation reduces to the wave equation: $W_{tt} - c^2 W_{xx} = 0$, where $c^2 = 2\omega_2 a^2$. The wave equation has a solution in the form of two dispersionless counter-propagating waves: $W(x,t) = W_+(x - ct) + W_-(x + ct)$. Unidirectional long-wave, small-amplitude approximation can be captured by the asymptotic multi-scale expansion method in the perturbation series (Zakharov & Kuznetsov, 1986):

$$u(x,t) = a\ \mathrm{e}^{-\mathrm{i}\left(\omega_0+\omega_2 a^2\right)t} \times \left[1 + \mathrm{i}\varepsilon R(\xi,\tau) + \varepsilon^2\left(\frac{1}{c}R_\xi - R^2\right) + \mathrm{O}(\varepsilon^3)\right], \tag{6}$$

where $\xi = \varepsilon(x - ct)$, $\tau = \varepsilon^3 t$, and ε is a small parameter. Applying the Fredholm's alternative theorem to a linear nonhomogeneous equation at order $\mathrm{O}(\varepsilon^3)$, one can derive the Korteweg–de Vries (KdV) equation for $W = R_\xi$:

$$-2cW_\tau - 6cWW_\xi + W_{\xi\xi\xi} = 0. \tag{7}$$

With quadratic nonlinear terms neglected, the KdV equation (7) reproduces the dispersion relation of the linear Boussinesq equation, which has been reduced in the unidirectional long-wave approximation. Thus, the linearized method of the second approach enables us to derive the nonlinear evolution equation for wave propagation over the background of the constant wave solution of the NLS equation (1).

Linearization methods for systems of ordinary differential equations (ODEs) are illustrated with the example of an autonomous nonlinear dynamical system (Glendinning, 1994):

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{u}), \tag{8}$$

where $\boldsymbol{u} = \boldsymbol{u}(t)$ is a vector observable and $\boldsymbol{f}(\boldsymbol{u})$ is a vector nonlinear function. Critical points of the dynamical system occur for $\boldsymbol{u} = \boldsymbol{u}_0$ such that $\boldsymbol{f}(\boldsymbol{u}_0) = 0$. Linearizing the nonlinear system (8) at the critical point $\boldsymbol{u}_0$, one expands $\boldsymbol{u}(t)$ as

$$\boldsymbol{u}(t) = \boldsymbol{u}_0 + \boldsymbol{v}(t). \tag{9}$$

Neglecting quadratic terms in $\boldsymbol{v}(t)$ reduces the linearized problem to the linear system with constant coefficients:

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \mathcal{J}\boldsymbol{v}, \quad \mathcal{J} = \nabla_{\boldsymbol{u}} f\bigg|_{\boldsymbol{u}=\boldsymbol{u}_0}, \tag{10}$$

where $\mathcal{J}$ is the Jacobian matrix. Solutions of the linearized system (10) depend on eigenvalues and eigenvectors of the Jacobian matrix: $\mathcal{J}\boldsymbol{w} = \lambda\boldsymbol{w}$. If all eigenvalues are located in the left half-plane of λ, that is, $\mathrm{Re}(\lambda) < 0$, then the critical point $\boldsymbol{u}_0$ is asymptotically stable and the perturbation $\boldsymbol{v}(t)$ decays to zero exponentially in t. If there exists at least one eigenvalue in the right half-plane of λ, then the critical point $\boldsymbol{u}_0$ is linearly unstable and the perturbation $\boldsymbol{v}(t)$ grows exponentially in t.

Some or all eigenvalues may be located at the imaginary axis of λ, that is, $\mathrm{Re}(\lambda) = 0$. In such systems, when no other eigenvalues exist for $\mathrm{Re}(\lambda) > 0$, the critical point $\boldsymbol{u}_0$ is neutrally (weakly) stable. Perturbations may however grow algebraically in t, if eigenvalues of λ are multiple with algebraic multiplicity exceeding the geometric multiplicity. Local linearization is often insufficient for full description of such weak instability and the nonlinear stability analysis is desired.

When eigenvalues cross or coalesce on the axis $\mathrm{Re}(\lambda) = 0$, bifurcations may occur in the spectrum of a linearized system (Glendinning, 1994). Normal forms for bifurcations can be derived by extending the linearized system into the nonlinear domain. For example, a Hopf bifurcation occurs when two simple eigenvalues $\lambda = \Gamma(\varepsilon) + \mathrm{i}\Omega(\varepsilon)$ and $\bar{\lambda} = \Gamma(\varepsilon) - \mathrm{i}\Omega(\varepsilon)$ cross the imaginary axis at the bifurcation parameter $\varepsilon = 0$, such that $\Gamma(0) = 0$ and $\Omega(0) = \Omega_0$. The normal form for the Hopf bifurcation can be derived by the asymptotic multi-scale expansion method in the perturbation series:

$$\begin{aligned}\boldsymbol{u}(t) &= \boldsymbol{u}_0 + \varepsilon\left[A(T)\boldsymbol{w}\mathrm{e}^{\mathrm{i}\Omega_0 t} + \bar{A}(T)\bar{\boldsymbol{w}}\mathrm{e}^{-\mathrm{i}\Omega_0 t}\right] \\ &\quad + \mathrm{O}(\varepsilon^2),\end{aligned} \tag{11}$$

where $T = \varepsilon^2 t$ and $\boldsymbol{w}$ is the eigenvector of $\mathcal{J}\boldsymbol{w} = \lambda\boldsymbol{w}$ for $\lambda = \mathrm{i}\Omega_0$ at $\varepsilon = 0$. The normal form for Hopf bifurcation follows from the Fredholm's alternative theorem at order $\mathrm{O}(\varepsilon^3)$ in the form

$$\frac{\mathrm{d}A}{\mathrm{d}T} = \left[\Gamma'(0) + \mathrm{i}\Omega'(0)\right]A - \beta(0)|A|^2 A, \tag{12}$$

where $\beta(0)$ is constant. $A(T)$ bifurcates into a periodic orbit at the Hopf bifurcation when $\Gamma'(0)\mathrm{Re}(\beta(0)) > 0$. Thus, linearization methods allow for linear and nonlinear stability analyses, resulting in simplifications of the nonlinear dynamical system (8).

DMITRY PELINOVSKY

See also **Causality; Multiple scale analysis; Quasilinear analysis; Stability**

Further Reading

Glendinning, P. 1994. *Stability, Instability and Chaos: An Introduction to the Theory of Nonlinear Differential Equations*, Cambridge and New York: Cambridge University Press

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

Zakharov, V.E. & Kuznetsov, E.A. 1986. Multi-scale expansions in the theory of systems integrable by the inverse scattering transform. *Physica* D, 18

LIOUVILLE EQUATION

See **Sine-Gordon equation**

LIOUVILLE THEOREM

See **Phase space**

LIPSCHITZ CONDITION

See **Phase space**

LIQUID CRYSTALS

Discovery and Basic Properties

Liquid crystals do not follow the conventional classification of matter into three states: gas, liquid, and solid. In solids the orientation and position of the molecular building blocks are well determined, whereas in liquids the molecules move around in a disordered mixture. In liquid crystals the molecules can change their position, but their orientation is largely determined by neighboring molecules. The liquid crystal phase was discovered by Friedrich Reinitzer in 1888, when he noticed that solid cholesterol benzoate melted at 145°C into a scattering liquid crystal phase which transformed into a clear liquid after further heating to 178°C (Priestly et al., 1976, Chapter 2).

The type of ordering is related to the shape of the molecules or mesogens building the liquid crystal (De Gennes & Prost, 1993, Chapter 1). Calamitic molecules are rod-shaped and form nematic (molecules are aligned) or smectic phases (molecules are aligned and ordered in layers) as shown in Figure 1. Discotic molecules are disc-shaped and form nematic or stacked phases (with the molecules stacked on top of each other). The ordering of the molecules is not perfect due to thermal motion. With increasing temperature, ordering decreases until the liquid crystal is transformed into an isotropic liquid. Because the phase transitions are

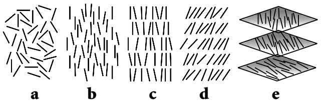

Figure 1. Different phases of liquid crystal materials. (a) isotropic, (b) nematic, (c) smectic-A, (d) smectic-C, (e) cholesteric.

related to temperature, these materials are called thermotropic liquid crystals. The organic molecules have low molecular weight (200–500) and can go through several phases from low to high temperature: solid–smectic-C–smectic-A–nematic–liquid (Figure 1). Chiral (handed) molecules do not have mirror symmetry and this may lead to a cholesteric phase in which the nematic orientation rotates slowly from one layer to the next (Figure 1e). Chiral phases have the property to reflect circularly polarized light that has the same handedness and periodicity as the molecule's structure. In lyotropic liquid crystals which have one hydrophobic and one hydrophilic part, the crystalline ordering depends on the concentration of certain molecules in the solvent. At very low concentrations individual molecules move freely in the liquid, at higher concentrations droplets, cylinders, or planes of attached molecules can be formed. Lyotropic liquid crystals are important in biological systems, in cell membranes, and dispersions (De Gennes & Prost, 1993, Chapter 1).

The nematic and smectic-A phases have uniaxial symmetry with the preferential axis parallel with the molecules. The preferential axis is characterized by a vector, the director $\bar{n}$. Other properties such as permittivity, refractive index, magnetic susceptibility, and conductivity are described by a tensor. The index of refraction of calamitic liquid crystals may, for example, vary between $n_\perp$=1.5 and $n_\parallel$=1.6. The smectic-C phase has biaxial properties. The deformation of a liquid crystal leads to elastic energy. The energy in nematic liquid crystals is due to three types of variations in the orientation: splay, twist, and bend (see De Gennes & Prost, 1993, Chapter 3; Blinov & Chigrinov, 1994, Chapter 4). The elastic energy is then expressed by the following equation:

$$F_{\text{elastic}} = \frac{1}{2}\left(K_{11}\,(\nabla\cdot\bar{n})^2 + K_{22}\,(\bar{n}\cdot\nabla\times\bar{n})^2 + K_{33}\,(\bar{n}\times\nabla\times\bar{n})^2\right). \qquad (1)$$

Surfaces can prefer either homeotropic (perpendicular) or parallel alignment of the director. If the interface is treated by rubbing a polyimide surface layer, this dictates a preferential azimuth for the director. The minimization of the combined elastic energy and surface energy determines the orientation of the director between the surfaces. This can lead to a pure twist as shown in Figure 2.

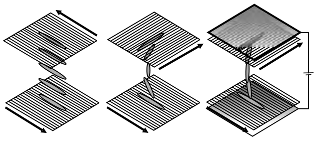

Figure 2. Nematic liquid crystal between two glass plates. The grid represents the alignment layer and the arrow the rubbing direction. Left: antiparallel rubbing with all molecules parallel; Middle: 90° twisted nematic; Right: when a voltage is applied between the electrodes, the molecules tilt.

Nonlinear Properties

In smectic liquid crystals, linear elasticity theory and linear hydrodynamics no longer hold because higher order terms come into play in the elastic energy (De Gennes & Prost, 1993, Chapter 8). The elasticity constants then depend on the deformation strength and this can lead to undulation instabilities, buckling, or focal conic structures. In a similar way, hydrodynamic nonlinearities yield frequency-dependent viscosities. If liquid crystal is illuminated with a high optical power, the physical properties may be modified. The processes leading to optical nonlinearities can be of electronic or non-electronic origin (Khoo, 1995). Electronic nonlinear effects involve processes in which the electronic wave functions of the liquid crystal molecules are perturbed by the optical field. These very fast processes are similar to electronic nonlinearities in solid crystals. These processes lead to second and third harmonic generation and optical phase conjugation (Khoo, 1995). Beside the electronic structure of the molecules, other physical properties can change because of an applied optical field, such as the molecular structure or orientation, the temperature, and the density (Elston & Sambles, 1998, Chapter 6). Slow temperature effects occur when a liquid crystal in a nematic phase is heated by an optical beam. The difference between the ordinary and extraordinary index of refraction decreases until the liquid crystal reaches the transition temperature and becomes isotropic. Another important effect is the molecular reorientation by the optical field. In the static case, the molecules will reorient due to the difference in dielectric constant ($\varepsilon_\parallel$ and $\varepsilon_\perp$); and in the optical case, due to the difference in refractive index ($n_\parallel$ and $n_\perp$).

Applications

Because of the anisotropy in the permittivity (or magnetic susceptibility), calamitic molecules with $\varepsilon_\parallel > \varepsilon_\perp$ prefer to orient their long axis parallel with an applied electric (or magnetic) field. By applying a voltage over a cell filled with liquid crystal, it is possible to make the molecules tilt out of the plane, as illustrated in Figure 2. In a uniaxial liquid crystal, linearly polarized ordinary and extra-ordinary waves propagate with refractive indices $n_\perp$ and n_e (with n_e between $n_\|$ and $n_\perp$) (Iizuka, 2002, Ch.4). Because of the difference in propagation speed, the polarization state of the combined wave changes continuously. In a twisted nematic display (Figure 2), the linearly polarized light roughly follows the orientation of the director (Mauguin regime) and rotates over 90° (Priestly et al., 1976, Chapter 14). If the top polarizer is also rotated over 90° compared to the bottom polarizer, all light will be transmitted. If a voltage is applied to the cell and the molecules orient perpendicularly, the linear polarization is unchanged, and no light passes through the analyser. Cholesteric liquid crystals appear in nature, for example, in the cuticle of insects. The result is a colored and polarized reflection, depending on the angle of observation.

Kristiaan Neyts and Jeroen Beeckman

See also **Dislocations in crystals; Harmonic generation; Nonlinear optics**

Further Reading

Blinov, L.M. & Chigrinov, V.G. 1994. *Electrooptic Effects in Liquid Crystal Materials, Partially Ordered Systems*, New York: Springer

De Gennes, P.G. & Prost, J. 1993. *The Physics of Liquid Crystals*, 2nd edition, Oxford: Clarendon Press and New York: Oxford University Press

Elston, S. & Sambles, R. 1998. *The Optics of Thermotropic Liquid Crystals*, London: Taylor & Francis

Iizuka, K. 2002. *Elements of Photonics, in Free Space and Special Media*, New York: Wiley-Interscience

Khoo, I.C. 1995. *Liquid Crystals, Physical Properties and Nonlinear Optical Phenomena*, New York: Wiley

Priestly, E.B., Wojtowicz, P.J. & Shent, P. 1976. *Introduction to Liquid Crystals*, New York and London: Plenum Press

LOCAL MODES IN MOLECULAR CRYSTALS

Energy transfer in molecular crystals occurs through exciton hopping from molecule to molecule. It is possible, however, that nonlinear localization takes place due to coupling between excitonic and vibrational degrees of freedom, transforming the exciton into a local mode. Such a mode is generated by an effective local nonlinear potential induced by the exciton-phonon coupling, and it has features similar to a polaron, a soliton, and a breather.

The experimental search for nonlinear local modes was initiated by Careri in the early 1970s in crystalline acetanilide ($CH_3CONHC_6H_5$), or ACN, which has structure similar to a polypeptide; thus, positive findings in ACN could be of biological significance. In particular, if nonlinear local modes exist and are sufficiently long-lived, they could be agents of energy transfer in biomolecules.

In ACN and in proteins, one finds hydrogen-bonded quasi-one-dimensional polypeptide chains with the following structure:

$$\cdots H-N-C=O\cdots H-N-C=O\cdots H-N-C=O\cdots. \tag{1}$$

Careri focused on the amide-I (C=O stretching) vibration and found its infrared absorption spectrum to have two peaks: a "normal" one at 1665 cm^{-1} and an "anomalous" one at 1650 cm^{-1}. The basic feature of the latter is its strong temperature dependence and its disappearance at biological temperatures, as shown in Figure 1.

In 1982, Careri and Scott hypothesized that this anomalous peak is a spectral signature of a self-trapped state similar to the Davydov soliton, while the normal peak is the delocalized free exciton of the amide-I vibration. An approximate way to represent the ACN local mode is by assuming an adiabatic separation between the faster amide-I exciton vibration and a slower local (Einstein) oscillator with which the hydrogen bond interacts. The corresponding Hamiltonian has three terms:

$$H = H_{\text{ex}} + H_{\text{ph}} + H_{\text{int}}, \tag{2}$$

where

$$\begin{aligned} H_{\text{ex}} &= \sum_n \left[\varepsilon_0 a_n^\dagger a_n - J(a_n^\dagger a_{n+1} + a_n^\dagger a_{n-1})\right], \\ H_{\text{ph}} &= \frac{1}{2}\sum_n \left[\frac{p_n{}^2}{m} + m\omega_0{}^2 q_n{}^2\right], \\ H_{\text{int}} &= \chi \sum_n q_n a_n^\dagger a_n . \end{aligned} \tag{3}$$

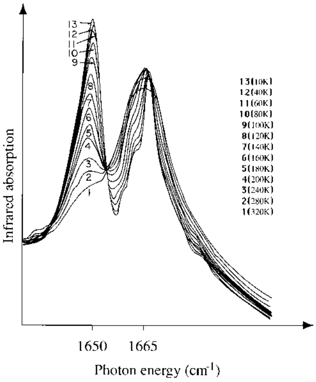

Figure 1. Infrared absorption spectra of crystalline acetanilide in the region of the amide-I mode. The free exciton peak is at 1665 cm^{-1} while the anomalous peak that is at 1650 cm^{-1} has strong temperature dependence and is associated with a self-trapped state (Careri et al., 1984).

Here $a_n^\dagger$ (a_n) creates (destroys) a quantum of amide-I excitation of energy ε_0 at the nth CO unit, J is the energy of exciton-exciton interaction between adjacent CO units, p_n and q_n are the momentum and position, respectively, of the local Einstein oscillator at the nth location representing a low frequency optical mode, and χ determines the strength of the interaction between the exciton and Einstein mode.

One way to analyze this complex problem is by treating the slow ($p_n \approx 0$) Einstein oscillators classically while evaluating the expectation value of the total Hamiltonian with respect to the exciton wave function $|\Psi\rangle = \sum_n c_n(t) a\dagger_n |0\rangle_{\mathrm{ex}}$, where $|0\rangle_{\mathrm{ex}}$ is the exciton system vacuum. Variational minimization of $\mathcal{H} \equiv \langle \Psi | H | \Psi \rangle$ with respect to the oscillator variable q_n leads to $q_n = -\chi/(m{\omega_0}^2)|c_n(t)|^2$, which upon substitution to $\mathcal{H}$ transforms the latter to the classical Hamiltonian

$$\mathcal{H} = \sum_n \left[\varepsilon_0 |c_n|^2 \quad - J(c*_n c_{n+1} + c*_n c_{n-1}) \right. \\ \left. - \frac{\chi^2}{2m\omega_0^2} |c_n|^4 \right]. \tag{4}$$

Hamilton's equations then lead to a discrete self-trapping system or a discrete nonlinear Schrödinger (DNLS) equation:

$$\mathrm{i}\hbar \dot{c}_n + J(c_{n+1} + c_{n-1}) + \frac{\chi^2}{m{\omega_0}^2}|c_n|^2 c_n = 0. \tag{5}$$

This equation admits two types of solutions: one is an extended plane wave corresponding to the free exciton, while the other is localized and corresponds to a self-trapped state. The difference of the energies of these two solutions is

$$E_{\mathrm{b}} = \frac{\chi^2}{2m{\omega_0}^2} - 2J, \tag{6}$$

corresponding to the binding energy of the self-trapped state. A combination of calculations and fitting to data suggests that $J \approx 4\,\mathrm{cm}^{-1}$, $\chi^2/m\omega_0^2 \approx 45\,\mathrm{cm}^{-1}$, and E_{b} is equal to $1665 - 1650 = 15\,\mathrm{cm}^{-1}$.

This picture is in accord with the presence of the two peaks in the amide-I region shown in Figure 1. Use of these parameter values in the context of a more sophisticated model gives also the correct temperature dependence of the anomalous peak. Because in the present case $J \ll \chi^2/m\omega_0^2$, the local mode can also be viewed as a DNLS discrete breather close to the anti-continuous limit (where $J \to 0$).

Although the above proposed was furnished in the 1980s, both the essence of Davydov's idea as well as its specific application to ACN were challenged during this period. The possibility that the anomalous peak is a result of Fermi resonance of the amide-I mode with an overtone of an alternative ACN vibrational mode was suggested as well as a possible structural source for the mode. These possibilities were to a large degree eliminated by the experimental work of Barthes and her collaborators. Definitive experimental information on the ACN local mode was furnished in 2002 by Edler and Hamm.

Using time-resolved pump-probe experiments with infrared beams of femtosecond duration, Edler and Hamm were able to excite the amide-I region and follow the time development of the various modes in that region. Their experiments showed a clear difference in the behavior of the normal and anomalous modes with the latter one being strongly anharmonic. Although the lifetime of the anomalous mode was only about 2 ps, return of this energy to the ground state took about 35 ps. Thus, the local mode initially relaxes into a state that is either spectroscopically dark or outside the spectral window of the probe. Interestingly, similar pump-probe experiments for the region of the NH stretching demonstrate the presence of nonlinear local mode structure in this range as well.

The initial experiments of Careri, the subsequent work of Barthes, and finally the experiments of Edler and Hamm along with Davydov's theory and its application to ACN by Scott present a convincing 30-year story that attributes the source of the anomalous ACN peak to a highly localized, relatively short-lived self-trapped state.

Another molecular crystal with different structure that is believed to form nonlinear local modes is the halide-bridged mixed-valence transition metal complex $\{[\mathrm{Pt(en)_2}][\mathrm{Pt(en)_2Cl_2}](\mathrm{ClO_4})_4\}$, where en = ethylenediamine. In PtCl, essentially one-dimensional chains of platinum alternating with chlorine atoms are formed, and an intra-molecular vibrational excitation of the $\mathrm{PtCl_2}$ trimer caused by energy transfer between Pt^{4+} and Pt^{2+} seems to become effectively localized as a result of coupling with other crystal vibrational modes. Raman spectra show a clear red-shifted PtCl overtone spectrum induced by anharmonicity that reduces in PtBr crystals due to smaller effective anharmonicity and disappears in PtI due to absence of nonlinearity. A multiquanta generalization of the theoretical ACN approach of Equations (1)–(4) demonstrates that the overtone redshifts in the PtCl Raman spectrum can be accounted for quantitatively as well as qualitatively by a nonlinear local mode picture.

G.P. Tsironis

See also **Breathers; Davydov soliton; Discrete nonlinear Schrödinger equations; Discrete self-trapping system; Pump-probe measurements**

Further Reading

Barthes M., Nunzio, G.D. & Ribet, M. 1996. Polarons or protontransfer in chains of peptide groups? *Synthetic Metals*, 76: 337–340

Bushmann, W.E., McGrane, S.D. & Shreve, A.P., 2003. Chemical tuning of nonlinearity leading to intrinsically localized modes in halide-bridged mixed-valence platinum materials. *Journal of Physical Chemistry* A, 107(40): 8198–8207

Careri, G., Buontempo, U., Galluzzi, F., Scott, A.C., Gratton, E. & Shyamsunder, E. 1984. Spectroscopic evidence for Davydov-like solitons in acetanilide. *Physical Review* B, 30: 4689–4702

Edler, J. & Hamm, P. 2002. Self-trapping of the amide I band in a peptide model crystal. *Journal of Chemical Physics*, 117: 2415–2424

Scott, A.C. 1992. Davydov's soliton. *Physics Reports*, 217: 1–67

Swanson, B.I., Brozik, J.A., Love, S.P., Strouse, G.F., Shreve, A.P., Bishop, A.R., Zang, W.-Z. & Salkola, M.I. 1999. Observation of intrinsically localized modes in a discrete low-dimensional material. *Physical Review Letters*, 82: 3288–3291

Voulgarakis, N.K., Kalosakas, G., Bishop, A.R. & Tsironis, G.P. 2001. Multiquanta breather model for PtCl. *Physical Review* B, 64: 020301

LOCAL MODES IN MOLECULES

The local mode model was introduced in the 1970s to describe highly excited vibrational overtone states. It has been used to describe XH stretching overtone spectra where X is some heavier atom like C, N, or O. However, the genesis of the local mode model in molecules dates back to the work of J.W. Ellis (1929). Ellis measured the overtone spectrum of benzene in the liquid phase to $\Delta v_{\mathrm{CH}} = 8$ (a transition from the ground vibrational state to a state with 8 quanta in CH-stretching). He found that the spectra were pseudo-diatomic in character, in that the CH overtone frequencies could be fitted by a simple Birge–Sponer relationship that had been used to describe a diatomic anharmonic oscillator. Thus, if the observed overtone peak frequencies, $\tilde{\nu}_{v0}$, are divided by v, the number of CH-stretching quanta in the excited overtone state, and plotted versus v, a straight line is obtained. The slope yields the local mode anharmonicity, $\tilde{\omega}x$, and the intercept can be used to determine the local mode frequency, $\tilde{\omega}$.

$$\tilde{\nu}_{v0}/v = \tilde{\omega} - (v+1)\tilde{\omega}x. \qquad (1)$$

In 1968, Henry and Siebrand used the localization ideas of Ellis to model the overtone spectrum of benzene on the basis of the assumption that the six local CH-stretching modes were anharmonic but only weakly coupled (Henry & Siebrand, 1968). They derived values for CH-stretching anharmonicity constants expressed in the normal mode basis and calculated the energies of all of the components of the overtone bands expressed in terms of normal modes. At $\Delta v_{\mathrm{CH}} = 6$, there are 462 such components of which 150 (the 75 doubly degenerate E_{1_u} states) are allowed. The spectra were then modeled on the basis of a weighted combination of the allowed normal modes. Subsequently, Hayward and Henry postulated that the radiation field selectively excites only the most anharmonic of the normal mode components (Hayward & Henry, 1975). In such a state, all the vibrational energy is effectively localized in one XH bond (X = C, N, O, etc.) and absorption to this pure local mode state is said to account for the overtone spectra. Some of these ideas had been published in 1936 in Germany by Mecke and his collaborators (Mecke & Ziegler, 1936). However, Mecke's work seemed to have been largely unknown and ignored. The description that emerges is that XH-stretching overtone spectra can be described by a local mode model in which the local oscillators are anharmonic but only weakly coupled. The radiation field selectively excites a state whose components have all of the vibrational energy localized in one of a set of equivalent XH bonds. Such a description predicts only a single transition for a given type of XH oscillator in agreement with experiment. The local mode model is now generally accepted, and because of the localization of energy for spectrally active states in a single chemical bond, XH-stretching overtone spectra are very sensitive to bond properties. Such spectra have been used to investigate molecular structure, molecular conformation, intermolecular forces, nonbonded intramolecular interactions, and intramolecular vibrational energy redistribution.

The Harmonically Coupled Anharmonic Oscillator Model

In a molecule like water, the two OH bonds can be treated as Morse potentials. In accord with the local mode model, the OH oscillators are anharmonic but only weakly coupled. It turns out that this off-diagonal potential and kinetic energy coupling is well approximated by the rules that govern coupling of harmonic oscillators.

The harmonically coupled anharmonic oscillator (HCAO) model was formulated independently by Mortensen et al. (1981) and by Child & Lawton (1981). Coupling is allowed only between states within a given vibrational manifold, that is, between states with the same value of the XH-stretching vibrational quantum number v. Coupling is restricted to the harmonic limit where oscillator population differs by ±1. All of the parameters in this model are obtained either directly from the spectra or from ab initio calculations.

The model has been very successful in accounting for the observed energies of overtone peaks in XH-stretching overtone spectra and, thus, has proved useful in the assignment of these spectra. It has been generalized to more than two equivalent oscillators and to sets of non-equivalent oscillators. No identification of the wave functions is needed for the HCAO model to calculate peak energies. However, if one uses a basis set consisting of products of one-dimensional Morse oscillators, the HCAO model can be used to generate vibrational wave functions. If such wave functions are used along with ab initio calculations of dipole moment functions (expanded in local coordinates), simple but highly successful calculations of overtone intensities are possible (Kjaergaard et al., 1990).

Overtone Intensities

Typically, these calculations: (i) identify transitions to pure local mode states, whose components have all of the vibrational energy localized in one of a set of equivalent XH oscillators, as the dominant peaks for $\Delta v_{XH} \geq 3$; (ii) indicate that the intensity of these pure local mode peaks is determined primarily by second-order diagonal terms involving $\partial^2\mu/\partial R^2$ in the dipole-moment operator μ; (iii) account for the fall-off in intensity with increasing v and the intensity distribution within a given manifold; (iv) account for the intensity distribution at $\Delta v_{XH} = 2$ where coupling between the local oscillators is important; and (v) account for relative intensities for different oscillators within a molecule. For example, calculations on cyclohexane (Kjaergaard & Henry, 1992) and naphthalene (Kjaergaard & Henry, 1995) successfully predict the relative intensities of the two non-equivalent ring hydrogens. These calculations have also been applied to the determination of accurate absolute overtone intensities for a number of species that are important in atmospheric photochemistry so that effects on the absorption of solar radiation can be estimated (Fono et al., 1999; Vaida et al., 2001).

One surprising result in these intensity calculations is that electron correlation in the dipole moment function affects the calculated fundamental intensity but has no significant effect on overtone intensities. One factor in this insensitivity to correlation is likely to be the primary importance of nonlinear terms in the dipole moment operator.

The HCAO intensity model has been generalized to include torsional motion in the Hamiltonian and to include the torsional coordinate in the ab initio calculation of the dipole moment function. As a result of coupling between torsion and stretching, both in the Hamiltonian and through the dipole-moment function, a very large number of transitions carry intensity and contribute to the overall methyl spectral profile. This model has successfully accounted for the methyl spectral profiles in the overtone spectra of methyl substituted aromatic molecules (Kjaergaard et al., 2000).

One early interest in the local mode field was the possibility that these highly localized states could be pumped with high power lasers and the result would be selective bond photochemistry. To date, no practical example of such behavior has arisen. The reason probably lies in the very rapid decay of these highly vibrationally excited states on a subpicosecond timescale. The dynamics of these local mode states continues to be a very active research area.

Bryan R. Henry

Further Reading

Child, M.S. & Lawton, T.R. 1981. Local and normal vibrational states: a harmonically coupled anharmonic oscillator model. *Faraday Discussions of the Chemical Society*, 71: 273

Ellis, J.W. 1929. Molecular absorption spectra of liquids below 3 μ. *Transactions of the Faraday Society*, 25: 888–897

Fono, L., Donaldson, D.J., Proos, R.J. & Henry, B.R. 1999. OH overtone spectra and intensities of pernitric acid. *Chemical Physics Letters*, 311: 131–138

Hayward, R.J. & Henry, B.R. 1975. A general local-mode theory for high energy polyatomic overtone spectra and application to dichloromethane. *Journal of Molecular Spectroscopy*, 57: 221–235

Henry, B.R. & Siebrand, W. 1968. Anharmonicity in polyatomic molecules. The CH stretching overtone spectrum of benzene. *Journal of Chemical Physics*, 49: 5369–5376

Kjaergaard, H.G. & Henry, B.R. 1992. The relative intensity contributions of axial and equatorial CH bonds in the local mode overtone spectrum of cyclohexane. *Journal of Chemical Physics*, 96: 4841–4851

Kjaergaard, H.G. & Henry, B.R. 1995. CH-stretching overtone spectra and intensities of vapor phase naphthalene. *Journal of Physical Chemistry*, 99: 899–904

Kjaergaard, H.G., Rong, Z., McAlees, A.J., Howard, D.L. & Henry, B.R. 2000. Internal methyl rotation in the CH stretching overtone spectra of toluene-α-d$_2$, -α-d$_1$, and -d$_0$. *Journal of Physical Chemistry* A, 104: 6398–6405

Kjaergaard, H.G., Yu, H., Schattka, B.J., Henry, B.R. & Tarr, A.W. 1990. Intensities in local mode overtone spectra: propane. *Journal of Chemical Physics*, 93: 6239–6248

Mecke, R. & Ziegler, R. 1936. Das Rotationsschwingungsspektrum des Acetylens (C_2H_2). *Zeitschrift für Physik*, 101: 405–417

Mortensen, O.S., Henry, B.R. & Mohammadi, M.A. 1981. The effects of symmetry within the local mode picture: a reanalysis of the overtone spectra of the dihalomethanes. *Journal of Chemical Physics*, 75: 4800–4808

Vaida, V., Daniel, J.S., Kjaergaard, H.G., Gross, L.M. & Tuck, A.F. 2001. Atmospheric absorption of near infrared and visible solar radiation by the hydrogen bonded water dimer. *Quarterly Journal of the Royal Meteorological Society*, 127: 1627–1644

LOCALIZATION

See **Solitons, types of**

LOCKING

See **Coupled oscillators**

LOGIC, DENDRITIC

See **Multiplex neuron**

LOGISTIC EQUATION

See **Population dynamics**

LOGISTIC MAP

See **One-dimensional maps**

LONG JOSEPHSON JUNCTIONS

The long Josephson junction is an excellent example of an experimental solid-state system that in a straightforward and measurable way, supports the existence and propagation of solitons (Parmentier, 1978; Pedersen, 1986; Ustinov, 1998). In this system, the soliton is called a fluxon, because it contains one quantum of magnetic flux, $\Phi_0 = h/2e \doteq 2.064 \times 10^{-15}$ webers (h is Planck's constant and e is the electron charge), in accord with the particle nature of solitons (Scott, 2003).

A Josephson junction consists of two weakly coupled superconductors separated by a thin (typically 1–10 nm) insulating layer. An important dynamical variable is the phase difference between the macroscopic quantum mechanical phases of the two superconductors, $\phi(x,t)$. Under an appropriate time normalization, the small Josephson junction obeys approximately the same equation as the damped and driven pendulum. If the Josephson junction is long with respect to an intrinsic length scale (called the Josephson penetration depth, of order 1–1000 μm), the equation becomes the damped and driven sine-Gordon equation (often called the perturbed sine-Gordon equation) (McLaughlin & Scott, 1978):

$$-\phi_{xx} + \phi_{tt} + \alpha\phi_t + \sin\phi = i\,. \tag{1}$$

Here, the damping parameter α (with typical experimental values in the range $0.01 < \alpha < 0.1$) is related to the Josephson junction parameters per unit area: conductance G, capacitance C, and supercurrent J; thus $\alpha = G\sqrt{\hbar/2eCJ}$. Also $i = I_{\text{bias}}/I_0$ is a bias current normalized to the maximum Josephson current, x is the direction of the long side of the junction normalized to the Josephson penetration depth, and time t is normalized to the so-called plasma frequency, which corresponds to the pendulum oscillation frequency. The plasma frequency is of order 1–500 GHz, depending on fabrication parameters (composition and thickness of insulating layer, superconducting materials, etc.).

A particularly interesting solution to Equation (1) is the sine-Gordon soliton (kink), or rather a slightly modified kink because of the damping and bias terms. Such a kink carries one quantum of magnetic flux (an antikink carries a flux quantum pointed in the opposite direction), and it has a steady state velocity determined as a balance between energy input from the bias current (i) and losses stemming from α. This solution has particle-like properties, behaving approximately as a Lorentz invariant, relativistic particle with the velocity of light in the system being the limiting velocity. This so-called Swihart velocity is determined by the junction fabrication parameters and is typically 1–5% of the velocity of light in vacuum.

The simple form of the dynamical equation shown in Equation (1) refers to the so-called overlap geometry shown in Figure 1(a), in which the externally applied current is fed uniformly to the long dimension of the long Josephson junction. In addition to Equation (1), boundary constraints must be satisfied. For an overlap junction of length L, the longitudinal (x-directed) surface current (j) is zero at $x = 0$ and $x = L$ if no external magnetic field is applied.

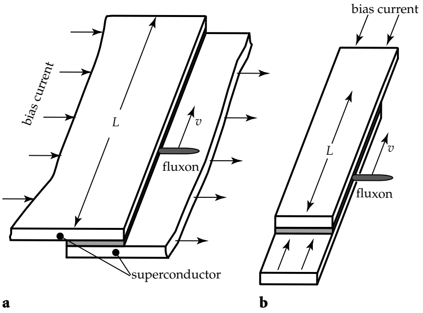

Figure 1. Rectangular-shaped long Josephson junctions (not to scale). (a) Overlap geometry. (b) Inline geometry.

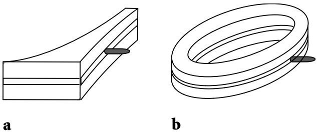

Figure 2. Other long junction geometries (bias not shown). (a) Eiffel junction. (b) Annular junction.

Because the long Josephson junction is typically fabricated using photolithographic and evaporation techniques borrowed from the semiconductor chip industry, it is easy to produce different geometries. Besides the overlap junction, inline junctions, Eiffel junctions, and annular junctions are of experimental and practical interest.

For the inline junction shown in Figure 1(b), the external current is fed from the narrow ends of the junction. The right-hand side of Equation (1) becomes zero, and the bias enter through the boundary conditions at $x = 0$ and L. In the Eiffel junction shown in Figure 2(a), the width of the junction is varied in an exponentially tapered shape approximating that of the Eiffel tower. For reasons of energy conservation, the propagation becomes unidirectional, with the fluxon tending to move in the direction of decreasing junction width (Benabdallah et al., 1996, 2000). The Eiffel junction has been proposed as a dc to ac converter and a microwave oscillator.

The annular junction, shown in Figure 2(b), is like a long overlap junction where the two ends are joined. For topological reasons connected with the superconducting phase quantization, periodic boundary conditions apply; that is, $j(0,t) = j(L,t) + 2p\pi$ with $p = 0, 1, 2, \ldots$, where j is the longitudinal surface current. A relatively clean soliton (fluxon) can be studied here since there are no collisions with boundaries. The total number of fluxons minus antifluxons is a conserved number (p) and fluxons can only disappear in a fluxon-antifluxon annihilation processes, conserving the total flux in the ring. In a variant of the annular junction, periodic changes in the junction width give rise to deceleration and acceleration of the fluxons, thereby producing radiation (McLaughlin & Scott, 1978).

The long Josephson junction has a potential as a microwave oscillator in the hundreds of Gigahertz range (100–1000 GHz). This is a frequency range, interesting for applications such as fast superconducting electronics, radio astronomy, and satellite-to-satellite communication. In this frequency range, competing oscillators are often bulky, noisy, and expensive. Importantly, the line width of the fluxon oscillator with proper stabilization can become very narrow—of the order 50 kHz. Since the line width of a local oscillator sets the frequency resolution for a heterodyne receiver, this is an important property for spectroscopy and for determining the frequency of radio sources in the universe.

For radio astronomy (~100–500 GHz), the standard today is a superconducting superheterodyne receiver, for which all components (mixer, local oscillator, filters, etc.) are fabricated in superconducting electronics, and the local oscillator is typically a long Josephson junction. In practice, two different schemes for a fluxon oscillator have been used. In one scheme, the fluxon propagates back and forth within the junction. At each collision with boundary some radiation is coupled out to an external circuit. The frequency of the radiation is given by the time of flight of the fluxon in the junction.

In another mode—the so-called flux flow mode—a unidirectional stream of fluxons is created. This can be done either by the above mentioned Eiffel junction or by applying a magnetic field that breaks the symmetry, giving rise to energy absorption at one end of the junction and energy creation at the other. With proper parameter adjustment, fluxons will be continuously created at one end of the junction, propagated through the junction, and absorbed in an external load (mixer) at the other end.

The power obtainable in such schemes is at best a few microwatts at 300–500 GHz with a narrow line width (of order 100 kHz). Nonetheless, this power level is sufficient for local oscillators of superconducting receivers.

As a concluding remark, we mention that it has been known since the early days of the long Josephson junction research, that the fluxon has properties related to a relativistic particle. Recent experiments have demonstrated that it also has quantum mechanical properties (Wallraff et al., 2003).

NIELS FALSIG PEDERSEN

See also **Josephson junctions; Sine-Gordon equation; Superconductivity**

Further Reading

Benabdallah, A., Caputo, J.G. & Scott, A.C. 1996. Exponentially tapered Josephson flux flow oscillator. *Physics Review* B, 54: 16139–16146.

Benabdallah, A., Caputo, J.G. & Scott, A.C. 2000. Laminar phase flow for an exponentially tapered Josephson oscillator. *Journal Applied Physics*, 88: 16139–16146

McLaughlin, D.W. & Scott, A.C. 1978. Perturbation analysis of fluxon dynamics. *Physics Review A*, 18: 1652–1680.

Parmentier, R.D. 1978. Fluxons in long Josephson junction. In *Solitons in Action*, edited by K. Lonngren & A.C. Scott, New York: Academic Press, 173–199

Pedersen, N.F. 1986. Solitons in Josephson transmission lines. In *Solitons: Modern Problems in Condensed Matter Sciences*, edited by A.A. Maradudin & V.H. Agranovich, Amsterdam: Elsevier, 469–501

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Ustinov, A.V. 1998. Solitons in Josephson junctions. *Physica* D, 123 (1–4): 315–329

Wallraff, A., Lukashenko, A., Lisenfeld, A., Kemp, A., Fistul, M.V., Koval, Y., & Ustinov, A.V. (2003). Quantum dynamics of a single vortex. *Nature*, 425: 155–158

LONG WATER WAVE APPROXIMATIONS

See **Water waves**

LONG-WAVELENGTH APPROXIMATION

See **Continuum approximation**

LOOP SOLITONS

See **Solitons, types of**

LOOP, CLOSED CAUSAL

See **Causality**

LORENTZ GAS

The Lorentz gas is a model dynamical system generated by the motion of non-interacting point particles ("electrons") in a field of immovable (infinitely heavy) spherical particles, called scatterers. The particles move by inertia according to Newton's law and are reflected from the scatterers elastically. The scatterers may overlap and are assumed to be randomly or periodically distributed.

Hendrik Lorentz introduced this model as a simplified version of the Drude model for metallic conductance of electricity (Lorentz, 1905). Lorentz's goal was to modify the Drude model in such a way that the Boltzmann equation becomes linear. This is so for the Lorentz gas because of the lack of interactions between particles, which do not influence the positions of the scatterers. As it is sufficient to consider just one moving particle, a one particle model is now usually called the Lorentz gas.

The Lorentz gas appeared to be an inadequate model to describe a gas of electrons in metals. However, it became one of the most popular models in non-equilibrium statistical mechanics, which is used to study the most fundamental problems such as irreversibility, the existence of transport coefficients, and the derivation of kinetic and hydrodynamic equations from mechanical laws. The Lorentz gas with random distribution of scatterers also serves as a relevant model of a dilute gas in equilibrium, consisting of a mixture of two types of (spherical) particles with one much larger and heavier than the other. Almost all collisions are between smaller and larger particles, rather than among smaller particles, even though there are many more smaller particles than heavy particles (Dettmann, 2000).

The kinetic stage of the Lorentz gas evolution is described by the Boltzmann equation, which refers to dilute gases. The appropriate limit procedure was suggested by Grad (1958) and is called the Boltzmann–Grad limit. In this limit, the density of a gas (density of scatterers) tends to zero while the mean free path of the moving particle remains constant. For the Lorentz gas, the kinetic Boltzmann equation goes in the Boltzmann–Grad limit into the linear Fokker–Planck–Kolmogorov equation for the corresponding Markov process. It has been proven for a dilute random Lorentz gas in the Boltzmann–Grad limit that the Boltzmann equation holds for almost any (Poisson distributed) configuration of scatterers (Boldrighini et al., 1983). This result allows a comparison of the theory with the results of computer experiments, because it refers to an individual configuration of scatterers rather than to characteristics averaged over the ensemble of all such configurations.

Rigorous results on the hydrodynamics of the Lorentz gas are available only for periodic configurations of scatterers. This problem for a random Lorentz gas currently looks hopeless because it can be reduced to a problem of random walks in random environments that is far beyond the present abilities of this theory. Clearly, the hydrodynamic equation for the Lorentz gas must be the diffusion equation because the momentum is not preserved in this model, and the conservation of energy is just equivalent to the conservation of mass. Therefore, the diffusion coefficient is also the only transport coefficient in the Lorentz gas.

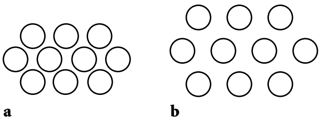

Figure 1. Periodic Lorentz gas: (a) With bounded free path. (b) With unbounded free path.

Consider a periodic Lorentz gas where a free path of the particle is uniformly bounded by some constant. The simplest such configuration of scatterers in the plane is obtained by taking the hexagonal lattice of sufficiently large scatterers (Figure 1(a)). Observe that a periodic Lorentz gas is just Sinai billiards (*See* **Billiards**) when restricted to an elementary cell of a periodic configuration of scatterers. Such a cell can always be chosen to be a rectangle $B = \{q = (q_1, q_2) : 0 \le q_1 \le b_1, 0 \le q_2 \le b_2\}$, where (q_1, q_2) are planar coordinates of the moving particle. Let the particle initially (at $t = 0$) have a random position in B and its velocity (which in a Lorentz gas can always be taken as equal to one) be also randomly distributed in the unit circle. Suppose that this distribution μ (in the direct product Π of B and the unit circle) has a well-behaved density with respect to the volume in Π. Then $x(t) = (q(t), v(t))$ is a random vector, where $q(t) = (q_1(t), q_2(t))$ is a position of the particle at time t and $v(t)$ is its velocity. Rescale trajectories of the particle by setting $q_t(s) = \frac{1}{\sqrt{t}} q(st)$, $0 \le s \le t$. The probability distribution μ on the initial conditions induces a probability distribution on μ_t on the set of all possible trajectories $q_t(s)$, $0 \le s \le t$. To study a hydrodynamic behavior one should consider large times. The measures μ_t converge (weakly) as $t \to \infty$ to a Wiener measure (Bunimovich & Sinai, 1981). This result provides a rigorous derivation of (time non-invertible macroscopic) hydrodynamic diffusion equations from first principles (time invertible microscopic Newtonian equations). It shows that at large scales in space and time, trajectories of the periodic Lorentz gas look the same as trajectories of the stochastic diffusion process. Indeed, the transition probabilities of Wiener processes are the fundamental solutions of the diffusion equation. The same result holds for the periodic Lorentz gas with bounded free path in dimensions ≥ 3 (Chernov, 1994).

According to the Einstein formula,

$$D = \frac{2}{d} \int_0^\infty \int_M [v(x(0)) \cdot v(x(t))\, \mathrm{d}\mu(x(0)]\, \mathrm{d}t, \quad (1)$$

where d is a dimension, M is the phase space, and $v(x(0))$ is the velocity of the particle at the moment $t = 0$. The Einstein formula is the first in the infinite hierarchy of Green–Kubo formulas that relate transport coefficients with the integrals of time correlations of some phase functions. Therefore, the problem of existence of the transport coefficients is reduced to an estimate of the rate of time-correlations decay, which for the periodic Lorentz gas is fast enough (Chernov & Young, 2000).

The condition that a free pass must be bounded is important. Periodic Lorentz gases with an unbounded free path (Figure 1(b)) behave super-diffusively rather than diffusively (Bleher, 1992) and time correlations there decay only according to a power law because the particle gets trapped for a long time in the part of the phase space with arbitrarily long free paths.

Until the 1960s, it was generally believed that time correlations in many-body systems decay exponentially. This opinion was based on the analysis of completing solvable models, such as Markov processes. Therefore, the discovery of long tails of time correlations in numerical experiments came as a surprise (Alder & Wainwright, 1967). It is now believed that time correlations in a random Lorentz gas decay according to a power law. This is confirmed by numerical experiments and physical theories but not by rigorous mathematical results.

Interesting models arise by placing the Lorentz gas in an external field. A periodic Lorentz gas in a magnetic field can demonstrate ergodic as well as integrable behavior (Berglund & Kunz, 1996). The Lorentz gas in a (weak) electric field becomes a non-Hamiltonian system, with the energy of the moving particle increasing indefinitely. One can connect this system to a thermostat to keep the energy fixed and to make the dynamics time reversible, whereupon it is possible to establish rigorously Ohm's law (Chernov et al., 1993). Presently, studies of the thermostated Lorentz gas are popular both in non-equilibrium statistical mechanics and in molecular dynamics (Dettmann, 2000).

LEONID BUNIMOVICH

See also **Billiards; Deterministic walks in random environments; Drude model**

Further Reading

Alder, B.J. & Wainwright, T.E. 1967. Velocity autocorrelations for hard spheres. *Physical Review Letters*, 18: 988–990

Berglund, N. & Kunz, H. 1996. Integrability and ergodicity of classical billiards in a magnetic field. *Journal of Statistical Physics*, 83: 81–126

Bleher, P.M. 1992. Statistical properties of two-dimensional periodic Lorentz gas with infinite horizon. *Journal of Statistical Physics*, 66: 315–373

Boldrighini, C., Bunimovich, L.A. & Sinai, Ya G. 1983. On the Boltzmann equation for the Lorentz gas. *Journal of Statistical Physics*, 32: 477–501

Bunimovich, L.A. & Sinai, Ya.G. 1981. Statistical properties of Lorentz gas with periodic configuration of scatterers. *Communications in Mathematical Physics*, 78: 479–497

Chernov, N.I. 1994. Statistical properties of the periodic Lorentz gas. Multidimensional case. *Journal of Statistical Physics*, 74: 11–53

Chernov, N. & Young, L.S. 2000. Decay of correlations for Lorentz gas and hard balls. In *Hard Ball Systems and the Lorentz Gas*, edited by D. Szasz, Berlin and New York: Springer

Chernov, N.I., Eyink, G.I., Lebowitz, J.L. & Sinai, Ya.G. 1993. Steady state electrical conduction in the periodic Lorentz gas. *Communications in Mathematical Physics*, 154: 569–601

Dettmann, C.P. 2000. The Lorentz gas: a paradigm for nonequilibrium stationary states. In *Hard Ball Systems and the Lorentz Gas*, edited by D. Szasz, Berlin and New York: Springer

Grad, H. 1958. Principles of the kinetic theory of gases. In *Handbuch der Physik* 12, edited by S. Flügge, Berlin: Springer

Lorentz, H.A. 1905. The motion of electron in metallic bodies. *Proceedings of Amsterdam Akademie*, 7: 438–441

LORENTZ TRANSFORM

See **Symmetry groups**

LORENZ ATTRACTOR

See **Attractors**

LORENZ EQUATIONS

A revolutionary development in the realms of nonlinear science over the past four decades has been an understanding of how ubiquitously systems of deterministic equations exhibit complicated behavior. Although the possibility of complicated aperiodic solutions was known much earlier from the work of Henri Poincaré and George D. Birkhoff on the so-called three-body problem (the Earth, Sun, and Moon) in celestial mechanics in the late 19th and early 20th centuries, such behavior was believed to be exceptional (*See* **Poincaré theorems**).

Two crucial pieces of work in the 1960s, the first by Edward Lorenz in 1963 on a continuous model relating to weather prediction (*See* **Butterfly effect**) and the second by Stephen Smale in 1967 in the context of discrete mathematical maps, altered this perception completely, showing that complex aperiodic behavior is generic for most deterministic nonlinear systems.

The basic equations introduced by Lorenz are

$$\begin{aligned}\dot{x} &= \sigma(y - x),\\ \dot{y} &= rx - y - xz,\\ \dot{z} &= xy - bz, \qquad (1)\end{aligned}$$

which were obtained by a drastic simplification of the fluid dynamical equations governing convection currents in the atmosphere. Equations (1) resulted from truncating a Fourier series description after the first few modes (Bergé et al., 1984), and the parameters σ

and r are physically important dimensionless numbers known, respectively, as the Prandtl and Rayleigh numbers in fluid mechanics.

Solutions of the Lorenz equations can be organized around a classification of their attractors (Glendinning, 1994; Strogatz, 1994; Nayfeh & Balachandran, 1995), which are regions of the space of physical variables or phase space (also called state space) to which solutions are attracted if the attractor is stable, or from whence they are repelled if the attractor is unstable.

Nonlinear systems may have several different kinds of attractors, including point attractors (also known as fixed, or critical, or equilibrium, points), isolated periodic attractors (or limit cycles), as well as more complex attractors called quasi-periodic and chaotic attractors. The nature and number of the attractors may change as the parameters ($\sigma, r,$ and b) vary, leading to a qualitative change in behavior referred to as a bifurcation.

A straightforward calculation showed Lorenz that his system contracted volumes in phase space—a dissipative system. This meant that any ball or volume of initial conditions in the phase space must shrink down to zero volume at large times under evolutions governed by the Lorenz equations. In other words, any attractors of the system must have zero volume.

Also, the model has a critical point at the origin ($x = y = z = 0$) and a pair of critical points $C^{\pm}$ at $x = y = \pm\sqrt{b(r-1)}$ for $r > 1$. Lorenz's stability analysis showed that the first fixed point at the origin of the phase space is always stable for $r < 1$ and a global attractor to which all initial points (or initial conditions) in the phase space are attracted for this parameter range. For $r > 1$, he found that the first fixed point at the origin becomes linearly unstable and is longer a bona fide attractor. However, the new fixed points are stable for

$$1 < r < r_{\mathrm{H}} = \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1}. \qquad (2)$$

In this parameter range, calculation shows that these fixed points coexist with an unstable limit cycle. At $r = r_{\mathrm{H}}$, this limit cycle is absorbed by the fixed point in a bifurcation that is known as a subcritical Hopf bifurcation. The fixed points $C^{\pm}$ go unstable in the process, being transformed into saddle points.

For $r > r_{\mathrm{H}}$, there are no simple stable point attractors, so Lorenz considered other stable limit cycles (periodic attractors) or solutions flying off to infinity. He showed that any possible limit cycles should be unstable for $r > r_{\mathrm{H}}$ and also that all trajectories must eventually enter, and be confined within, a certain large ellipsoid.

Clearly, any attractor had to be relatively complex, and the trajectories could not cross while evolving on the attractor. Together with the earlier-mentioned fact of volume contraction (which means that any attractor to which the solution trajectories or orbits are attracted must have zero volume asymptotically in time), Lorenz was left with a perplexing question: What must such an attractor be like?

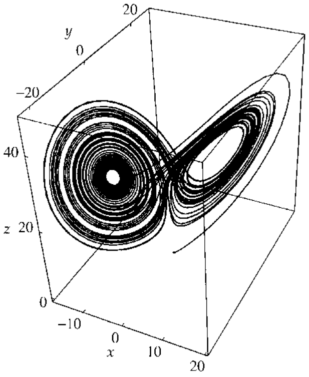

Figure 1. Lorenz attractor in (x, y, z) phase space for $\sigma = 10, r = 28, b = 8/3$.

He numerically studied the case $\sigma = 10, r = 28, b = \frac{8}{3}$ with the $r > r_{\mathrm{H}} = 24.74$ in Equation (2). He also chose initial conditions $(x, y, z) = (0, 1, 0)$ close to the saddle point nearer the origin. The resulting solutions in (x, y, z) phase-space, and for $x(t)$ as a function of time are shown in Figures 1 and 2, respectively, where the solution trajectory settles onto a strange butterfly-shaped attractor set after an initial transient. The clearly aperiodic dynamics on this attractor is emphasized in Figure 2. In addition, the attractor appears to be infinitesimally thin and thus satisfies the requirement of zero volume. Hence, all three requirements of aperiodicity of the dynamics on the attractor, zero volume, and nonself-crossing of the orbit were reconciled on this "strange attractor"—a term subsequently coined by Ruelle and Takens. Among other observations, Lorenz noted that this strange attractor has a self-similar structure at all scales, which is now a well-known feature of chaotic attractors.

Subsequent work has revealed that trajectories on chaotic attractors exhibit sensitivity to initial conditions (SIC) where initially contiguous phase points diverge exponentially in time. Other work, some still in progress (Abraham & Shaw, 1983; Bergé et al., 1984), has elucidated the complex process leading to the structure of strange attractors and the properties of the orbits on them. In particular, there is convergence of orbits along the stable manifolds of saddle fixed points (this enables dissipation or volume contraction to be satisfied), divergence of trajectories along the corresponding unstable manifolds (this accounts for SIC), and foldings leading to bounded strange attractors.

Based on a combination of topological and numerical ideas to reconstruct chaotic attractors from

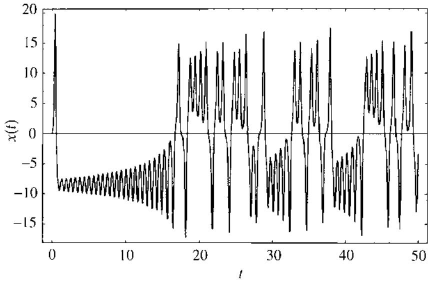

Figure 2. A solution $x(t)$ for $\sigma = 10, r = 28, b = 8/3$.

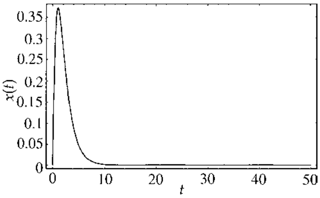

Figure 3. A solution $x(t)$ for $\sigma = 1, b = 2, r = 1/9$.

time series such as that in Figure 2, numerical procedures have been developed to compute various properties such as the fractal dimension (a measure of the global dimension of the attractor in phase space), Lyapunov exponents (measures of sensitivity to initial conditions on the attractor), power spectra and autocorrelation functions, and the so-called Kolmogorov entropy (Nayfeh & Balachandran, 1995). In particular, the chaotic attractor in Figure 1 has a dimension of about 2.05; that is, it occupies a region of the phase space that has infinite area or fills up any area (since the dimension is greater than 2), but has zero volume (since the dimension is less than 3).

The occurrence of sensitivity to initial conditions as indicated by at least one positive Lyapunov exponent or the envelope of the autocorrelation function going to zero in finite time signify the eventual loss of memory of past history (or equivalently the impossibility of prediction far into the future), which are integral features of chaotic dynamics.

Note that the Lorenz equations also exhibit completely deterministic or nonchaotic behavior for isolated parameter sets obtained via Painlevé analysis that are the so-called integrable cases (Tabor & Weiss, 1981). Figure 3 shows the completely regular dynamics for the parameter set $\sigma = 1, b = 2, r = 1/9$. Note how much more regular and orderly, the behavior of $x(t)$ is than that shown in Figure 2.

In conclusion, two features of the Lorenz model continue to make it topical. First, it remains a useful model in which to study essential basic features of chaotic systems. Second, the model has popped up in topical settings as diverse as atmospheric dynamos, lasers, and nonlinear optics, as well as the control of chaos (Ning & Haken, 1990; Pecora & Carroll, 1990; Nayfeh & Balachandran, 1995; Batchelor et al., 2000).

S. Roy Choudhury

See also **Attractors; Butterfly effect; Chaotic dynamics; Lyapunpov exponents; Phase space**

Further Reading

Abraham, R.H. & Shaw, C.D. 1983. *Dynamics: The Geometry of Behavior*, Santa Cruz: Aerial Press

Batchelor, G.K., Moffatt, H.F. & Worster, M.G. (editors). 2000. *Perspectives in Fluid Dynamics*, Cambridge and New York: Cambridge University Press

Bergé, P., Pomeau, Y. & Vidal, C. 1984. *Order Within Chaos*, New York: Wiley

Choudhury, S.R. 1992. On bifurcations and chaos in predator-prey models with delay. *Chaos, Solitons and Fractals*, 2: 393–409

Dodd, R.K., Eilbeck, J.C., Gibbon, J.D. & Morris, H.C. 1982. *Solitons and Nonlinear Wave Equations*, London: Academic Press

Glendinning, P. 1994. *Stability, Instability, and Chaos*, Cambridge and New York: Cambridge University Press

Lorenz, E.N. 1963. Deterministic nonperiodic flow. *Journal of Atmospheric Science*, 20: 130–141

Nayfeh, A.H. & Balachandran, B. 1995. *Applied Nonlinear Dynamics*, New York: Wiley

Ning, C.Z. & Haken, H, 1990. Detuned lasers and the complex Lorenz equations: Subcritical Hopf bifurcations. *Physical Review* A, 41: 3826–3837

Nusse, H. & Yorke, J.A. 1992. *Dynamics: Numerical Explorations*, New York: Springer

Pecora, L.M. & Carroll, T.L. 1990. Synchronization in chaotic systems. *Physical Review Letters*, 64: 821–824

Sparrow, C. 1982. *The Lorenz Equations*, New York: Springer

Strogatz, S.H. 1994. *Nonlinear Dynamics and Chaos*, Reading, MA: Addison-Wesley

Tabor, M. & Weiss, J. 1981. Analytic structure of the Lorenz system. *Physical Review* A, 24: 2157

LOTKA–VOLTERRA EQUATIONS

See **Population dynamics**

LYAPUNOV EXPONENTS

The sequence of powers a^k, $k = 1, 2, \ldots$, of a complex number $a \neq 0$ has one of three types of asymptotic behavior: exponential growth, exponential decay, or constant, according to the modulus $|a|$. A similar situation holds for powers A^k, $k = 1, 2, \ldots$, of an $n \times n$ matrix A. More precisely, for any vector $v \neq 0$, the following limit exists:

$$\lim_{k \to +\infty} \frac{1}{k} \log \|A^k v\| = \lambda(v). \tag{1}$$

The possible values of $\lambda(v)$, $\lambda_1 < \cdots < \lambda_s$ are called Lyapunov exponents (LEs), and in this special case they are equal to $\log |a|$, where a is an eigenvalue of A (real or complex). Moreover, there is a strictly increasing sequence of subspaces in R^n

$$\{0\} = V_0 \subset V_1 \subset \cdots \subset V_s = \mathrm{R}^n, \qquad (2)$$

such that $\lambda(v) = \lambda_i$ for all $v \in V_i \setminus V_{i-1}$, $i = 1, 2, \ldots, s$. The difference of dimensions $m_i = \dim V_i - \dim V_{i-1}$ is called the multiplicity of the Lyapunov exponent λ_i. The rates of growth of k-dimensional volumes of k-dimensional parallelepipeds under the action of A are sums of appropriate Lyapunov exponents. For example, the exponential rate of growth of the n-dimensional volume is equal to $\log |\det A| = m_1\lambda_1 + \cdots + m_s\lambda_s$.

A similar situation occurs for products $A^k = A_k A_{k-1} \ldots A_1$ when the sequence of matrices $A_1, A_2, \ldots$ is periodic (by direct reduction to the constant case we obtain so-called Floquet multipliers, and the LEs are logarithms of their moduli).

In the general nonperiodic case, the key Oseledec Multiplicative Ergodic Theorem (OMET) states that the same scenario holds, when the matrices we multiply $(A_1, A_2, \ldots)$ are supplied by a stationary stochastic process (with matrix values). In this case, the conclusions apply with probability one.

In the context of dynamical systems, the difference equation $x_{n+1} = \Phi(x_n)$, where $\Phi : M \to M$ is a diffeomorphism of a compact smooth manifold M (the phase space). For any ergodic invariant probability measure ν, the Lyapunov exponents are the limits

$$\lim_{k\to+\infty} ||D_x\Phi^k v|| = \lambda(v), \qquad (3)$$

which exist for ν almost every $x \in M$ and all nonzero tangent vectors $v \in T_xM$. Thus, Lyapunov exponents characterize the growth in linear approximation of infinitesimal perturbations of initial conditions. Positive LEs lead to instability and negative LEs to stability for perturbations in the respective directions.

Applying OMET both to the future and the past (i.e., to Φ and to Φ^{-1}), we obtain a splitting of tangent spaces

$$T_xM = W_1 \oplus W_2 \oplus \ldots \oplus W_s, \qquad (4)$$

such that

$$\lim_{k\to\pm\infty} ||D_x\Phi^k v|| = \pm\lambda_i \qquad (5)$$

for ν almost every $x \in M$ and all nonzero tangent vectors $v \in W_i(x)$. It is of direct interest to know if this infinitesimal behavior translates to the behavior of actual orbits of the nonlinear system. We introduce the stable $E^{\mathrm{s}}(x)$ and unstable $E^{\mathrm{u}}(x)$ subspaces, which are the direct sums of subspaces $W_i(x)$ with negative, respectively positive, LEs. It is the content of Pesin Theory that the subspaces E^{s} and E^{u} can be integrated almost everywhere, that is, for ν almost every point $x \in M$ there are smooth submanifolds $W^{\mathrm{s}}(x)$ and $W^{\mathrm{u}}(x)$, which are at every point tangent to the stable and unstable subspaces, respectively.

The stable and unstable manifolds are crucial in understanding chaotic dynamics. Their importance stems from the fact that two orbits starting on one of these submanifolds have the same asymptotic behavior either in the future or in the past.

LEs are associated with the Kolmogorov–Sinai entropy. They enter into the Ruelle Inequality, which states that for any C^1 diffeomorphism Φ and any ergodic invariant probability measure ν

$$h_\nu(\Phi) \le \sum_{\lambda_i>0} m_i\lambda_i, \qquad (6)$$

where $h_\nu(\Phi)$ is the K–S entropy. This inequality turns into equality (the Pesin formula) for absolutely continuous measures ν and C^2 diffeomorphisms. It was proven by Ledrappier and Young that equality holds if and only if ν is the Sinai–Ruelle–Bowen (SRB) measure.

For Hamiltonian systems (i.e., when Φ is a symplectomorphism), the LEs come in pairs of opposite numbers, which is related to the symmetry of spectra of symplectic matrices. Furthermore the subspace V_k is skew-orthogonal to V_{s-k}, $k = 1, 2, \ldots, s$: a property shared with eigenspaces of symplectic matrices.

Although analytic formulas for LEs are almost nonexistent, they can be readily obtained by numerical methods. Analytical estimates of LEs are available where all the matrices in question are J-separated for some quadratic form (field of forms) J. A matrix A is J-separated, if $J(Av) \ge 0$ for all v such that $J(v) \ge 0$.

Maciej P. Wojtkowski

See also **Chaotic dynamics; Sinai–Ruelle–Bowen measures; Phase space**

Further Reading

Barreira, L. & Pesin, Y.B. 2002. *Lyapunov Exponents and Smooth Ergodic Theory*, Providence, RI: American Mathematical Society

Katok, A. & Hasselblatt, B. 1995. *Introduction to the Modern Theory of Dynamical Systems*, Cambridge and New York: Cambridge University Press

Wojtkowski, M.P. 2001. Monotonicity, J-algebra of Potapov and Lyapunov exponents. In *Smooth Ergodic Theory and Its Applications*, Providence, RI: American Mathematical Society, pp. 499–521

Wojtkowski, M.P. & Liverani, C. 1998. Conformally symplectic dynamics and symmetry of the Lyapunov spectrum. *Communications in Mathematical Physics*, 194: 47–60

LYAPUNOV STABILITY

See **Stability**

M

MACH STEM EFFECT

See **Shock waves**

MAGNETOHYDRODYNAMICS

Magnetohydrodynamics (MHD) is the special field of fluid dynamics that is concerned with the motions of electrically conducting fluids in the presence of magnetic fields. Flows of liquid metals or ionized gases (plasmas) are the main areas of applications of MHD.

For the theoretical description of MHD-problems, the basic equations of fluid dynamics must be supplemented by the Lorentz force $\boldsymbol{j} \times \boldsymbol{B}$ where $\boldsymbol{j}$ is the density of electric current and $\boldsymbol{B}$ is the magnetic flux density. To determine these fields Maxwell's equations are used in the magnetohydrodynamic approximation in which the displacement current is neglected. This represents a good approximation as long as the fluid velocity is small compared with the velocity of light. These equations together with Ohm's law for a moving conductor are given by

$$\begin{aligned} &\nabla \cdot \boldsymbol{B} = 0, \quad \frac{\partial}{\partial t}\boldsymbol{B} = -\nabla \times \boldsymbol{E}, \quad \nabla \times \boldsymbol{B} = \mu \boldsymbol{j}, \\ &\boldsymbol{j} = \sigma(\boldsymbol{v} \times \boldsymbol{B} + \boldsymbol{E}), \qquad (1) \end{aligned}$$

where μ is the magnetic permeability of the fluid and σ is its electrical conductivity. These "pre-Maxwell" equations have the property that they are invariant with respect to a Galileo transformation, i.e., the equations remain unchanged in a new frame of reference moving with the constant velocity vector $\boldsymbol{V}$ relative to the original frame of reference. Indicating the variables of the new frame by a prime, we find

$$\begin{aligned} &\boldsymbol{v}' = \boldsymbol{v} - \boldsymbol{V}, \quad \frac{\partial}{\partial t'} = \frac{\partial}{\partial t} + \boldsymbol{V} \cdot \nabla, \\ &\boldsymbol{B}' = \boldsymbol{B}, \quad \boldsymbol{E}' = \boldsymbol{E} + \boldsymbol{V} \times \boldsymbol{B}, \quad \boldsymbol{j}' = \boldsymbol{j}. \qquad (2) \end{aligned}$$

This invariance is the basis for the combination in MHD of Equations (1) with the equations of hydrodynamics in their usual nonrelativistic form. The application of MHD to plasmas is limited to sufficiently low frequency and long wavelength phenomena. If these conditions are not satisfied, two-fluid equations may be used.

Elimination of $\boldsymbol{E}$ and $\boldsymbol{j}$ from Equation (1) yields the equation of magnetic induction

$$\nabla \times \left(\frac{1}{\sigma\mu} \nabla \times \boldsymbol{B} \right) = \frac{\partial}{\partial t}\boldsymbol{B} + \nabla \times (\boldsymbol{v} \times \boldsymbol{B}), \qquad (3)$$

which for a solenoidal velocity field $\boldsymbol{v}$ and a constant magnetic diffusivity $\lambda \equiv (\sigma\mu)^{-1}$ can be further simplified,

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v} \cdot \nabla \right) \boldsymbol{B} - \lambda \nabla^2 \boldsymbol{B} = \boldsymbol{B} \cdot \nabla \boldsymbol{v}. \qquad (4)$$

This equation has the form of a heat equation with the magnetic field line stretching term on the right-hand side acting as a heat source. This interpretation is useful in the dynamo problem of the generation of magnetic fields by fluid motions (*See* **Dynamos, homogeneous**). In order that a magnetic field $\boldsymbol{B}$ may grow, the term on the right-hand side of (3) must overcome the effect of the magnetic diffusion term on the left-hand side. Using a typical velocity U and a typical length scale d, the ratio of the two terms can be estimated by

$$Ud/\lambda \equiv \mathrm{Rm}, \qquad (5)$$

which is called the magnetic Reynolds number in analogy to the ordinary Reynolds number used in fluid dynamics. A necessary condition for dynamo action is $\mathrm{Rm} > 1$, but this is by far not a sufficient condition (*See* **Dynamos, homogeneous**).

A typical problem of MHD is the flow in the channel in the presence of an applied homogeneous magnetic field $\boldsymbol{B}_0$. The equation for the plane parallel flow $\boldsymbol{v} = v_x(y)\boldsymbol{i}$ in the channel between two parallel plates with distance $2d$ is given by

$$\eta \frac{\mathrm{d}^2}{\mathrm{d}y^2} v_x(y) + \frac{1}{\mu} \boldsymbol{i} \cdot ((\nabla \times \boldsymbol{B}) \times \boldsymbol{B}) = -A, \qquad (6)$$

where η is the dynamic viscosity, $\boldsymbol{i}$ is the unit vector in the x-direction along the channel, and A is the constant pressure gradient in the opposite direction. In accordance with the configuration of the problem, we have assumed that the steady solution depends only on the y-coordinate perpendicular to the plates. Using the

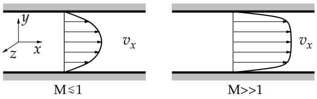

Figure 1. Profile of channel flow for small and large M.

notation $\boldsymbol{B} = \boldsymbol{B}_0 + \boldsymbol{b}$ we find after integrating Equation (4)

$$\lambda \frac{\mathrm{d}}{\mathrm{d}y} b_x(y) + B_{oy} v_x = -D, \qquad (7)$$

where the constant of integration D represents the electric field directed in the z-direction; in other words, Equation (7) is the z-component of Ohm's law. Insertion of this expression into Equation (6) yields

$$\eta \frac{\mathrm{d}^2}{\mathrm{d}y^2} v_x(y) - \sigma B_{0y}^2 v_x(y) = -A + \sigma B_{oy} D \qquad (8)$$

The solution of this equation is given by

$$v_x = v_0 \left(1 - \frac{\cosh(My/d)}{\cosh M} \right), \qquad (9)$$

with

$$M \equiv B_{0y} d \sqrt{\sigma/\eta}$$

and

$$v_0 = (A - \sigma D B_{oy}) d^2 / \eta M^2, \qquad (10)$$

where M is the Hartmann number named after the scientist who first considered this type of problem. The constant D depends on the boundary condition applied in the transverse z-direction. D can be chosen, for example, such that the average electric current in the z-direction vanishes because of the insulating side walls. It is worth noting that only the y-component of $\boldsymbol{B}_0$ enters solution (9). The flow is not affected by the x- and z-components of $\boldsymbol{B}_0$ as long as it does not depend on those coordinates. Depending on the Hartmann number M, the velocity profile $v_x(y)$ varies between the two limiting cases sketched in Figure 1. For $M \to 0$, the parabolic Poiseuille profile is recovered from expression (9), while for $M \gg 1$, boundary layers of thickness d/M develop, which are called Hartmann layers. In geophysical and astrophysical applications of MHD, the Chandrasekhar number $Q = M^2$ is often used in place of the Hartmann number.

In many applications of MHD, it is justified to neglect magnetic diffusion in first approximation. Then the relationship

$$\frac{\partial}{\partial t} \boldsymbol{B} = \nabla \times (\boldsymbol{v} \times \boldsymbol{B}) \qquad (11)$$

can be obtained from Equation (3) in the limit of large magnetic Reynolds numbers. This equation is the same as that obeyed by the vorticity of an inviscid fluid (*See* **Fluid dynamics**). By analogy to the vorticity laws of Kelvin and Helmholtz it can be concluded that the field line that is attached to a fluid element at some initial time continues to be attached to that fluid element at all times; that is, the field lines are "frozen" into the fluid. This statement is known as Alfvén's theorem.

In highly electrically conducting fluids, it is appropriate to consider Equation (11) together with the Euler equation of hydrodynamics,

$$\rho \left(\frac{\partial}{\partial t} \boldsymbol{v} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} \right) = -\nabla p + \mu^{-1} (\nabla \times \boldsymbol{B}) \times \boldsymbol{B}, \quad (12)$$

where the Lorentz force has been included, but viscous friction has been neglected. In the case of an incompressible fluid with $\rho =$ constant, both, $\boldsymbol{v}$ and $\boldsymbol{B}$ are solenoidal vector fields and an exact steady solution of Equations (11) and (12) is given by

$$\boldsymbol{v} = (\rho\mu)^{-1/2} \boldsymbol{B}(x, y, z)$$

$$\text{with} \qquad p = p_0 + \rho \mid \boldsymbol{v} \mid^2 /2, \qquad (13)$$

where the relationship $\boldsymbol{v} \cdot \nabla \boldsymbol{v} = (\nabla \times \boldsymbol{v}) \times \boldsymbol{v} + \nabla \mid \boldsymbol{v} \mid^2 /2$ has been used. The dependence of $\boldsymbol{B}$ on the cartesian coordinates x, y, z is arbitrary except for the condition $\nabla \cdot \boldsymbol{B} = 0$. Let us write $\boldsymbol{B}$ as the sum of its average and the remainder, $\boldsymbol{B} = B_0 \boldsymbol{i} + \boldsymbol{b}$, where it has been assumed without loss of generality that the average magnetic field is parallel to the x-direction given by the unit vector $\boldsymbol{i}$. Using a Galileo transformation with the constant velocity vector $V_A \boldsymbol{i}$, we obtain a time-dependent exact solution relative to the new frame of reference,

$$\boldsymbol{v}'(x - V_\mathrm{A} t, y, z) = (\rho\mu)^{-1/2} \boldsymbol{b}(x - V_\mathrm{A} t, y, z), \quad (14)$$

where the choice $V_\mathrm{A} = (\rho\mu)^{-1/2} B_0$ has been made. This solution describes dispersion-free waves called Alfvén waves. The physical interpretation of these waves of arbitrary amplitude is that the field lines of the basic magnetic field are embedded in the fluid-like rubber strings and provide a restoring force whenever a fluid parcel is displaced. The phase velocity V_A is called the Alfvén velocity. In compressible electrically conducting fluids a more complex spectrum of wave phenomena is found owing to combinations of Alfvén and acoustic modes.

F.H. Busse

See also **Alfvén waves; Dynamos, homogeneous; Fluid dynamics; Nonlinear plasma waves**

Further Reading

Davidson, P.A. 2001. *An Introduction to Magnetohydrodynamics*, Cambridge and New York: Cambridge University Press
Moreau, R. 1990. *Magnetohydrodynamics*, Dordrecht: Kluwer
Roberts, P.H. 1967. *An Introduction to Magnetohydrodynamics*, New York: Elsevier and London: Longmans

MANAKOV EQUATIONS

See **Nonlinear Schrödinger equations**

MANDELBROT SETS

See **Fractals**

MANLEY–ROWE RELATIONS

In lossless linear systems, a harmonic sinusoidal input at a frequency ω remains harmonic, the system's steady-state response to a linear combination of individual harmonic inputs is a linear combination of individual responses, and an average input power is zero at each frequency. In lossless nonlinear systems, on the other hand, harmonic inputs at some frequencies also generate inputs at multiple and combination frequencies. The average input powers at each frequency may not be zero but they are predicted from the general relations found by Jack Manley and Harrison Rowe in 1956 (Manley & Rowe, 1956).

Suppose the input applied to a nonlinear system consists of two harmonic signals of frequencies ω_1 and ω_2. The nonlinear system produces an output at all frequencies $m\omega_1 + n\omega_2$, where m and n are integers. For example, $m = 2, n = 0$ and $m = 0, n = 2$ are signals at the double frequencies $2\omega_1$ and $2\omega_2$, while $m = n = 1$ a signal with the combinational frequency $\omega_1 + \omega_2$. Denote $P_{m,n}$ the total input power at frequency $m\omega_1 + n\omega_2$. In lossless systems, the total power is zero; thus

$$\sum_{m,n} P_{m,n} = 0.$$

In other words, the power $P_{m,n}$ is negative if the signal at $m\omega_1 + n\omega_2$ is generated due to nonlinearity, and the power is positive if the signal is applied from input to the system.

The Manley–Rowe relations are relations between powers $P_{m,n}$ and frequencies $m\omega_1 + n\omega_2$ (Penfield, 1960):

$$\sum_{n=-\infty}^{\infty} \sum_{m=0}^{\infty} \frac{m P_{m,n}}{m\omega_1 + n\omega_2} = 0,$$

$$\sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{n P_{m,n}}{m\omega_1 + n\omega_2} = 0. \quad (1)$$

As an elementary example, we consider a heterodyne system, comprising an oscillator with frequency ω_1, which is mixed with the carrier incident frequency ω such that a combination frequency $\omega_2 = \omega - \omega_1$ occurs in the spectrum of the nonlinear system (Scott, 1970). We neglect signals at other frequencies and denote powers at frequencies ω_1, ω_2 and $\omega = \omega_1 + \omega_2$ as P_1, P_2, and $P = -P_1 - P_2$. The Manley–Rowe relations between the powers and frequencies of the three signals are

$$\frac{P_1}{\omega_1} = \frac{P_2}{\omega_2} = -\frac{P}{\omega}$$

Notice that these relations preserve conservation of powers $P_1 + P_2 + P = 0$ at the three resonant frequencies ω_1, ω_2, and ω. If $\omega > \omega_1, \omega_2$, the Manley–Rowe relations show that the power P_2 at the combinational frequency ω_2 is smaller than the input power P by a factor of ω_2/ω, while the power at the heterodyne frequency ω_1 is smaller than P by a factor of ω_1/ω. Therefore, when an input signal of larger frequency transforms into output signals of smaller frequencies, the powers of output signals become smaller.

As another application, we consider an electro-optical modulator that takes an input signal at frequency 1 GHz and multiplies with another signal at frequency 100 MHz to produce a modulated output signal at combinational frequency 1.1 GHz. If the power of the output signal has to be 1 mW, the Manley–Rowe relations require powers 0.0909 mW at 100 MHz and 0.9090 mW at 1 GHz. The 1 GHz signal is sometimes called the pump since it provides most of the power needed in the modulation process.

Manley–Rowe relations naturally describe constants of motion in the time evolution of the resonant nonlinear wave interactions. If three waves with frequencies ω_1, ω_2 and $\omega = \omega_1 + \omega_2$ and wavevectors $\mathbf{k}_1$, $\mathbf{k}_2$ and $\mathbf{k} = \mathbf{k}_1 + \mathbf{k}_2$ satisfy the phase matching conditions

$$\omega(\mathbf{k}_1 + \mathbf{k}_2) = \omega(\mathbf{k}_1) + \omega(\mathbf{k}_2),$$

then their interaction is resonant in time and is described by the system of three-wave interaction equations:

$$\begin{aligned} \mathrm{i}\left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right) a &= \gamma a_1 a_2 \mathrm{e}^{-\mathrm{i}\Delta t}, \\ \mathrm{i}\left(\frac{\partial}{\partial t} + \mathbf{v}_1\cdot\nabla\right) a_1 &= \gamma a \bar{a}_2 \mathrm{e}^{\mathrm{i}\Delta t}, \\ \mathrm{i}\left(\frac{\partial}{\partial t} + \mathbf{v}_2\cdot\nabla\right) a_2 &= \gamma a \bar{a}_1 \mathrm{e}^{\mathrm{i}\Delta t}, \end{aligned} \quad (2)$$

where $\Delta = \omega(\mathbf{k}) - \omega(\mathbf{k}_1) - \omega(\mathbf{k}_2)$ is the frequency detuning from exact resonance. In other words, $\mathbf{v}$, $\mathbf{v}_1$, and $\mathbf{v}_2$ are group velocities of the three waves, for example, $\mathbf{v} = \nabla\omega(\mathbf{k})$; and γ is a real-valued coupling coefficient in nonlinear systems with quadratic nonlinearities. The system of amplitude equations (2) has integrals of motions:

$$\begin{aligned} Q_1 &= \int \left(|a|^2 + |a_1|^2\right) \mathrm{d}\mathbf{x} = \text{const}, \\ Q_2 &= \int \left(|a|^2 + |a_2|^2\right) \mathrm{d}\mathbf{x} = \text{const}, \\ Q &= \int \left(|a_1|^2 - |a_2|^2\right) \mathrm{d}\mathbf{x} = \text{const}. \end{aligned} \quad (3)$$

The integrals of motions are the Manley–Rowe relations for the powers

$$P_1 = \omega_1 \int |a_1|^2 \, \mathrm{d}\mathbf{x},$$
$$P_2 = \omega_2 \int |a_2|^2 \, \mathrm{d}\mathbf{x},$$
$$P = \omega \int |a|^2 \, \mathrm{d}\mathbf{x}. \qquad (4)$$

The resonant interaction of three waves results in decay of the wave with the larger frequency $\omega > \omega_1, \omega_2$ into waves of smaller frequencies ω_1; ω_2, which alternates with annihilation of the two waves of smaller frequencies ω_1, ω_2 into the wave of larger frequency ω (Gaponov-Grekhov & Rabinovich, 1992). This process has a simple quantum interpretation. Conservation of quantum momentums and energies of wave particles leads to the resonant relations:

$$\hbar\mathbf{k} = \hbar\mathbf{k}_1 + \hbar\mathbf{k}_2, \qquad \hbar\omega = \hbar\omega_1 + \hbar\omega_2,$$

where $\hbar$ is Planck's constant. If the wave of smaller frequency ω_1 is initially larger than the waves of frequencies ω and ω_2, it cannot decay into the other two waves, because there are not enough wave particles with frequencies ω_2 to merge into wave particles with frequency ω. On the other hand, if the wave of larger frequency ω is initially large, it can decay into two wave particles of smaller frequencies ω_1 and ω_2. Manley–Rowe relations follow from the conservation of number of particles in the quantum interpetation above. For instance, when a phonon with energy $\hbar\omega$ is absorbed, two phonons of energies $\hbar\omega_1$ and $\hbar\omega_2$ are emitted, such that the Manley–Rowe relations hold.

Manley–Rowe invariants play an important role in studies of properties of three-wave interactions in system (2). In particular, optical solitons are supported by dispersive and diffraction effects in nonlinear three-wave interactions. The stability of optical solitons is determined by the Vakhitov–Kolokolov criterion, which involves derivatives of the Manley–Rowe invariants (3) with respect to parameters of optical solitons (Buryak et al., 1997).

When the frequencies ω_1 and ω_2 of the two resonant waves coincide, the three-wave interactions degenerate into resonant second-harmonic generation; thus the wave $a_1 = a_2 = a_0$ with the fundamental frequency $\omega_1 = \omega_2 = \omega_0$ generates the wave a at the double frequency $\omega = 2\omega_0$. The system of equations (2) simplifies then to the form (Etrich et al., 2000)

$$\mathrm{i}\left(\frac{\partial}{\partial t} + \mathbf{v}_0 \cdot \nabla\right) a_0 = \gamma a \bar{a}_0 \mathrm{e}^{\mathrm{i}\Delta t},$$
$$\mathrm{i}\left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla\right) a = \gamma a_0^2 \mathrm{e}^{-\mathrm{i}\Delta t}. \qquad (5)$$

System (5) for the second-harmonic generation has only one Manley–Rowe invariant:

$$Q_0 = \int \left(|a_0|^2 + |a|^2\right) \mathrm{d}\mathbf{x} = \mathrm{const}.$$

DMITRY PELINOVSKY

See also **Frequency doubling; Harmonic generation; *N*-wave interactions**

Further Reading

Buryak, A.V., Kivshar, Yu.S. & Trillo, S. 1997. Parametric spatial solitary waves due to type II second-harmonic generation *Journal of the Optical Society of America* B, 14: 3110–3118

Etrich, C., Lederer, F., Malomed, B.A., Peschel, T. & Peschel, U. 2000. Optical solitons in media with a quadratic nonlinearity. *Progress in Optics*, 41: 483–568

Gaponov-Grekhov, A.V. & Rabinovich, M.I. 1992. *Nonlinearities in Action: Oscillations, Chaos, Order, Fractals*, Berlin and New York: Springer

Manley, J.M. & Rowe, H.E. 1956. Some general properties of nonlinear elements. *Proceedings of the Institute of Radio Engineers*, 44: 904–913

Penfield, P., Jr. 1960. *Frequency-power Formulas*, New York: Wiley

Scott, A.C. 1970. *Active and Nonlinear Wave Propagation in Electronics,* New York: Wiley

Weiss, M.T. 1957. Quantum derivation of energy relations analogous to those for nonlinear reactances. *Proceedings of the Institute of Radio Engineers*, 45: 1012–1013

MAPS

A *map* is a dynamical system with discrete time. Such dynamical systems are defined by iterating a transformation $\boldsymbol{\phi}$ of points $\mathbf{x}$

$$\mathbf{x}_{n+1} = \boldsymbol{\phi}(\mathbf{x}_n), \qquad (1)$$

from a space $\mathcal{M}$ of dimension d (or a domain of this space) onto itself. $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$ are thus points belonging to this so-called *phase space* $\mathcal{M}$, which can be a Euclidean space such as the space $\mathbb{R}^d$ of d-tuples of real numbers or a manifold such as a circle, a sphere, or a torus $\mathbb{T}^d$ (or a domain such as an interval or a square).

An *endomorphism* is a surjective (i.e., many-to-one) transformation $\boldsymbol{\phi}$ of the space $\mathcal{M}$ onto itself. Thus, the transformation $\boldsymbol{\phi}$ is not invertible. Examples of endomorphisms are one-dimensional maps of the interval such as the logistic map $\phi(x) = 1 - ax^2$ on $-1 \leq x \leq 1$, the Bernoulli map $\phi(x) = rx$ (modulo 1) with integer r (also called r-adic map), or the Gauss map $\phi(x) = 1/x$ (modulo 1) which generates continuous fractions. The last two maps are defined onto the unit interval $0 \leq x \leq 1$. There also exist multidimensional examples such as the exact map $\boldsymbol{\phi}(x, y) = (3x + y, x + 3y)$ (modulo 1) on the torus $\mathbb{T}^2$ (Lasota & Mackey, 1985).

An *automorphism* is a one-to-one (i.e., invertible) transformation $\boldsymbol{\phi}$ of the space $\mathcal{M}$ onto itself. Automorphisms for which the one-to-one transformation $\boldsymbol{\phi}$ is continuous on $\mathcal{M}$ are called *homeomorphisms*. We speak about $\mathcal{C}^r$*-diffeomorphisms* if $\boldsymbol{\phi}$ is r-times differentiable and $\frac{\partial^r \boldsymbol{\phi}}{\partial \mathbf{x}^r}$ is continuous on $\mathcal{M}$. Examples of automorphisms are the circle maps defined with a monotonously increasing function $\phi(x) = \phi(x+1) - 1$ onto the circle; the baker map:

$$\boldsymbol{\phi}(x, y) = \begin{cases} (2x, \frac{y}{2}) & \text{if} \quad 0 \le x < 1/2, \\ \left(2x - 1, \frac{y+1}{2}\right) & \text{if} \quad 1/2 \le x < 1, \end{cases} \tag{2}$$

onto the unit square (Hopf, 1937); the cat map:

$$\boldsymbol{\phi}(x, y) = (x + y, x + 2y) \quad (\text{modulo } 1), \tag{3}$$

onto the torus $\mathbb{T}^2$ (Arnol'd & Avez, 1968); the quadratic map:

$$\boldsymbol{\phi}(x, y) = (y + 1 - ax^2, bx) \ , \tag{4}$$

onto $\mathbb{R}^2$ also called the *Hénon map* (Hénon, 1976; Gumowski & Mira, 1980), among many others.

Iterating a noninvertible transformation $\boldsymbol{\phi}$ generates a semigroup of endomorphisms $\{\boldsymbol{\phi}^n(\mathbf{x})\}_{n \in \mathbb{N}}$, where $\mathbb{N}$ is the set of nonnegative integers. Iterating an invertible transformation $\boldsymbol{\phi}$ generates a group of automorphisms $\{\boldsymbol{\phi}^n(\mathbf{x})\}_{n \in \mathbb{Z}}$, where $\mathbb{Z}$ is the set of all the integers. Such groups or semigroups are deterministic dynamical systems with discrete time, called maps.

Link Between Maps and Flows

Maps naturally arise in continuous-time dynamical systems (i.e., flows) defined with $d+1$ ordinary differential equations

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \mathbf{F}(\mathbf{X}), \tag{5}$$

by considering the successive intersections of the trajectories $\mathbf{X}(t)$ with a codimension-one Poincaré section $\sigma(\mathbf{X}) = 0$. If $\mathbf{x}$ denotes d coordinates which are intrinsic to the Poincaré section, the successive intersections $\{\mathbf{X}_n = \mathbf{X}(t_n)\}_{n \in \mathbb{Z}}$ of the trajectory correspond to a sequence of points $\{\mathbf{x}_n\}_{n \in \mathbb{Z}}$ and return times $\{t_n\}_{n \in \mathbb{Z}}$ in the Poincaré section. According to Cauchy's theorem which guarantees the unicity of the trajectory $\mathbf{X}(t)$ issued from a given initial condition $\mathbf{X}(0)$ (i.e., by the determinism of the flow), the successive points and return times are related by

$$\begin{cases} \mathbf{x}_{n+1} = \boldsymbol{\phi}(\mathbf{x}_n), \\ t_{n+1} = t_n + T(\mathbf{x}_n), \end{cases} \tag{6}$$

where $\boldsymbol{\phi}(\mathbf{x})$ is the so-called *Poincaré map* and $T(\mathbf{x})$ the return-time (or ceiling) function. The knowledge of the Poincaré map and its associated return-time function allows us to recover the flow and its properties.

Consider a similar construction for ordinary differential equations which are periodic in time

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}(\mathbf{x}, t) = \mathbf{F}(\mathbf{x}, t + T), \tag{7}$$

in which case the return-time function reduces to the period T in Equation (6) and the Poincaré map becomes a *stroboscopic map*.

Examples of Poincaré maps are the Birkhoff maps in the case of billiards. Billiards are systems of particles in free flights (or more generally following Hamiltonian trajectories) interrupted by elastic collisions. The knowledge of the collisions suffices to reconstruct the full trajectories. The Birkhoff map is thus the transformation ruling the dynamics of billiards from collision to collision.

Properties

Maps can be classified according to different properties. An important question is to know if a map is locally volume-preserving or not. If the map is differentiable, the volume preservation holds if the absolute value of its Jacobian determinant is equal to unity everywhere in $\mathcal{M}$:

$$\left| \det \frac{\partial \boldsymbol{\phi}}{\partial \mathbf{x}} \right| = 1. \tag{8}$$

This is the case for the baker map (2), the cat map (3), and the quadratic map (4) if $b = \pm 1$.

Maps that contract phase space volumes on average are said to be *dissipative*. In the limit $b \to 0$, the two-dimensional automorphism (4) contracts the phase space areas so much that it becomes an endomorphism given by the one-dimensional logistic map. This explains why highly dissipative dynamical systems are often very well described in terms of endomorphisms such as the logistic map.

A map is *symplectic* if its Jacobian matrix satisfies

$$\left(\frac{\partial \boldsymbol{\phi}}{\partial \mathbf{x}}\right)^{\mathrm{T}} \cdot \boldsymbol{\Sigma} \cdot \frac{\partial \boldsymbol{\phi}}{\partial \mathbf{x}} = \boldsymbol{\Sigma}, \tag{9}$$

where ^{T} denotes the transpose and $\boldsymbol{\Sigma}$ is an antisymmetric constant matrix: $\boldsymbol{\Sigma}^{\mathrm{T}} = -\boldsymbol{\Sigma}$. Symplectic maps act onto phase spaces of even dimension. Poincaré maps of Hamiltonian systems as well as Birkhoff maps are symplectic in appropriate coordinates. Symplectic maps are volume-preserving. Area-preserving maps are symplectic, but there exist volume-preserving maps which are not symplectic in dimensions higher than two.

A map is symmetric under a group $\mathcal{G}$ of transformations $\mathbf{g} \in \mathcal{G}$ if

$$\mathbf{g} \circ \boldsymbol{\phi} = \boldsymbol{\phi} \circ \mathbf{g}. \tag{10}$$

A map is said to be *reversible* if there exists an involution, that is, a transformation $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}^2 = 1$, which transforms the map into its inverse:

$$\boldsymbol{\theta} \circ \boldsymbol{\phi} \circ \boldsymbol{\theta} = \boldsymbol{\phi}^{-1}. \tag{11}$$

There exist reversible maps which are not volume-preserving (Roberts & Quispel, 1992).

Invariant Subsets

These are subsets of the phase space that are invariant under the action of the map. They include the fixed points $\boldsymbol{\phi}(\mathbf{x}_*) = \mathbf{x}_*$ (which correspond to periodic orbits of a corresponding flow) and the periodic orbits of prime period n defined as trajectories from the initial condition $\mathbf{x}_\mathrm{p}$ such that $\boldsymbol{\phi}^n(\mathbf{x}_\mathrm{p}) = \mathbf{x}_\mathrm{p}$ but $\boldsymbol{\phi}^j(\mathbf{x}_\mathrm{p}) \neq \mathbf{x}_\mathrm{p}$ for $0 < j < n$. Tori may also be invariant as in the case of KAM quasi-periodic motion.

An invariant subset $\mathcal{I}$ of the map is *attracting* if there exists an open neighborhood $\mathcal{U}$ such that $\boldsymbol{\phi}\,\mathcal{U} \subset \mathcal{U}$ and $\mathcal{I} = \cap_{n \in \mathbb{N}} \boldsymbol{\phi}^n \mathcal{U}$. The open set $\mathcal{B} = \cup_{n \in \mathbb{N}} \boldsymbol{\phi}^{-n} \mathcal{U}$ is called the *basin of attraction* of $\mathcal{I}$. An *attractor* is an attracting set which cannot be decomposed into smaller ones. A set which is not attracting is said to be *repelling*.

A closed invariant subset $\mathcal{I}$ is *hyperbolic* if (i) the tangent space $\mathcal{T}_\mathbf{x}\mathcal{M}$ of the phase space $\mathcal{M}$ splits into stable and unstable linear subspaces $\mathcal{E}_\mathbf{x}^{(\mathrm{s})}$ and $\mathcal{E}_\mathbf{x}^{(\mathrm{u})}$ depending continuously on $\mathbf{x} \in \mathcal{I}$,

$$\mathcal{T}_\mathbf{x}\mathcal{M} = \mathcal{E}_\mathbf{x}^{(\mathrm{s})} \oplus \mathcal{E}_\mathbf{x}^{(\mathrm{u})}; \qquad (12)$$

(ii) the linearized dynamics preserves these subspaces; and (iii) the vectors of the stable (resp. unstable) subspace are contracted (resp. expanded) by the linearized dynamics (Ott, 1993). Hyperbolicity implies sensitivity to initial conditions of exponential type, characterized by positive Lyapunov exponents. By extension, a map is said to be *hyperbolic* if its invariant subsets are hyperbolic.

For the baker map (2), the unstable linear subspace is the x-direction while the stable one is the y-direction and the unit square is hyperbolic with a positive Lyapunov exponent. Moreover, the dynamics of the baker map can be shown to be equivalent to a so-called Bernoulli shift, that is, a symbolic dynamics acting as a simple shift on all the possible infinite sequences of symbols 0 and 1, so that most of its trajectories are random. The baker map is thus an example of a hyperbolic fully chaotic map.

A diffeomorphism is said to have the *Anosov* property if its whole compact phase space $\mathcal{M}$ is hyperbolic. Examples of Anosov diffeomorphisms are the cat map (3) and its nonlinear perturbations:

$$\begin{cases} x_{n+1} = x_n + y_n + f(x_n, y_n) \quad (\text{modulo } 1), \\ y_{n+1} = x_n + 2y_n + g(x_n, y_n) \quad (\text{modulo } 1), \end{cases} \qquad (13)$$

with small enough periodic functions $f(x, y)$ and $g(x, y)$ defined on the torus. We notice that these nonlinear perturbations of the cat map are generally not area-preserving.

Dissipative maps may have chaotic attractors (i.e., attractors with positive Lyapunov exponents) which are not necessarily hyperbolic. This is the case for a set of $a > 0$ values of a positive Lebesgue measure in the logistic map (Jakobson, 1981), as well as in the quadratic map (4) if $b > 0$ is sufficiently small (Benedicks & Carleson, 1991).

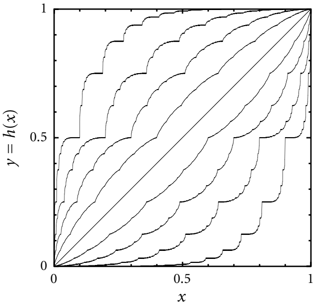

Figure 1. Homeomorphisms $y = h(x)$ transforming the dyadic map into maps (15) with $p = 0.1, 0.2, ..., 0.9$.

Maps may also have sensitivity to initial conditions of stretched-exponential type (with vanishing Lyapunov exponent) as it is the case for the intermittent maps $\phi(x) = x + ax^\zeta$ (modulo 1) with $\zeta > 2$.

Conjugacy Between Maps

In the study of maps, it is often important to modify the analytic form of the map by a change of variables $\mathbf{y} = \mathbf{h}(\mathbf{x})$. Such a conjugacy would transform map (1) into

$$\mathbf{y}_{n+1} = \boldsymbol{\psi}(\mathbf{y}_n), \quad \text{with} \quad \boldsymbol{\psi} = \mathbf{h} \circ \boldsymbol{\phi} \circ \mathbf{h}^{-1}. \qquad (14)$$

The Kolmogorov–Sinai entropy per iteration is known to remain invariant if the conjugacy $\mathbf{h}$ is a diffeomorphism, but it is only the topological entropy per iteration which is invariant if the conjugacy is a homeomorphism. For instance, the logistic map $\phi(x) = 1 - 2x^2$ is conjugated to the tent map $\psi(y) = 1 - 2|y|$ by the conjugacy $y = -1 + \frac{4}{\pi} \arcsin \sqrt{\frac{x+1}{2}}$, both maps having their Kolmogorov–Sinai entropy equal to ln 2. On the other hand, the dyadic map $\phi(x) = 2x$ (modulo 1) is conjugated to the map

$$\psi(y) = \begin{cases} \frac{y}{p} & \text{if} \quad 0 \le y \le p, \\ \frac{y-p}{1-p} & \text{if} \quad p < y \le 1, \end{cases} \qquad (15)$$

with $p \neq \frac{1}{2}$ by a homeomorphism $h(x)$ which is not differentiable (see Figure 1). These two maps have their topological entropy equal to ln 2 but different Kolmogorov–Sinai entropies.

Conjugacies are also important to transform a circle map such as $\phi(x) = x + \alpha + \varepsilon \sin(2\pi x)$ with $|\varepsilon| < \frac{1}{2\pi}$ into a pure rotation $\psi(y) = y + \omega$ of rotation number

$$\omega = \lim_{n \to \infty} \frac{1}{n}(x_n - x_0). \qquad (16)$$

According to Denjoy theory, such a conjugacy is possible if the rotation number is irrational, in which case the motion is quasi-periodic and nonchaotic. The circle map also illustrates the phenomenon of synchronization to the external frequency α, which occurs when the rotation number ω takes rational values corresponding to periodic motions. The motion may become chaotic if $|\varepsilon| > \frac{1}{2\pi}$.

Area-preserving Maps

Periodic, quasi-periodic, and chaotic motions are also the features of area-preserving maps which can be considered as Poincaré maps of Hamiltonian systems with two degrees of freedom. Area-preserving maps can be derived from a variational principle based on some Lagrangian generating function $\ell(x_{n+1}, x_n)$. The variational principle requires that the trajectories are extremals of the action

$$W = \sum_{n\in\mathbb{Z}} \ell(x_{n+1}, x_n). \tag{17}$$

The vanishing of the first variation, $\delta W = 0$, leads to the second-order recurrence equation

$$\frac{\partial \ell(x_{n+1}, x_n)}{\partial x_n} + \frac{\partial \ell(x_n, x_{n-1})}{\partial x_n} = 0. \tag{18}$$

This recurrence can be rewritten in the form of a two-dimensional map by expliciting the equations for the momenta

$$\begin{cases} p_{n+1} = \frac{\partial \ell(x_{n+1}, x_n)}{\partial x_{n+1}}, \\ p_n = -\frac{\partial \ell(x_{n+1}, x_n)}{\partial x_n}. \end{cases} \tag{19}$$

The Birkhoff map of a billiard is recovered if ℓ is the distance traveled by the particle in free flight between collisions and x_n is the arc of the perimeter at which the collision occurs. If a free particle or rotor is periodically kicked by an external driving, the Lagrangian function takes the form

$$\ell = \tfrac{1}{2}(x_{n+1} - x_n)^2 - V(x_n). \tag{20}$$

A famous map is the so-called *standard map*

$$\begin{cases} p_{n+1} = p_n + K \sin x_n, \\ x_{n+1} = x_n + p_{n+1} \quad (\text{modulo } 2\pi), \end{cases} \tag{21}$$

obtained for the kicked rotor with the potential $V(x) = K \cos x$ in Equation (20). The motivation for studying the standard map goes back to works on the origin of stochasticity in Hamiltonian systems (Chirikov, 1979; Lichtenberg & Lieberman, 1983, MacKay & Meiss, 1987).

Phase portraits of an area-preserving map typically present closed curves of KAM quasi-periodic motion, which form elliptic islands. Hierarchical structures of elliptic islands develop on smaller and smaller scales. The elliptic islands are surrounded by chaotic zones extending over finite area (see Figure 2).

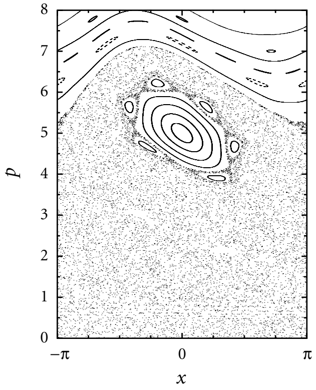

Figure 2. Phase portrait of the Fermi–Ulam area-preserving map, $p_{n+1} = | p_n + \sin x_n |, x_{n+1} = x_n + 2\pi M/p_{n+1}$ (modulo 2π), ruling the motion of a ball bouncing between a fixed wall and a moving wall oscillating sinusoidally in time, in the limit where the amplitude of the oscillations is much smaller than the distance between the walls. p is the velocity of the ball in units proportional to the maximum velocity of the moving wall. x is the phase of the moving wall at the time of collision. The parameter M is proportional to the ratio of the distance between the walls to the amplitude of the oscillations of the moving wall. (See Lichtenberg & Lieberman, 1983, for more details.)

Typical area-preserving maps such as the standard map (21) or the one of Figure 2 display a variety of motions that interpolate between two extremes, namely, the fully chaotic behavior of hyperbolic area-preserving maps such as the baker and cat maps and the fully regular motion of integrable maps such as the one given by the second-order recurrence:

$$x_{n+1} - 2x_n + x_{n-1} = -2\mathrm{i} \ln \frac{\kappa^2 + \mathrm{e}^{-\mathrm{i}x_n}}{\kappa^2 + \mathrm{e}^{+\mathrm{i}x_n}} \tag{22}$$

(Faddeev & Volkov, 1994).

Some area-preserving maps may have a repelling Smale horseshoe as the only invariant subset at finite distance. This is the case in the quadratic map (4) for $b = -1$ and large enough values of $a > 0$. Such horseshoes often arise in open two-degrees-of-freedom Hamiltonian systems describing the chaotic scattering of a particle in some time-periodic potential.

Complex Maps

Such maps are defined with some analytic function $\phi(z)$ of $z = x + \mathrm{i}y \in \mathbb{C}$ or some multidimensional generalizations of it. Complex maps are generally endomorphisms. An example is the complex logistic map $\phi(z) = z^2 + c$. Other examples are given by the Newton–Raphson method of finding the roots of a function $f(z) = 0$:

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}. \tag{23}$$

Complex maps have invariant subsets called Julia sets which are defined as the closure of the set of repelling periodic orbits (Devaney, 1986). The motion is typically chaotic on the Julia set which is repelling and often separates the basins of attraction of the attractors. For instance, the map

$$z_{n+1} = \frac{z_n}{2} + \frac{1}{2z_n}, \tag{24}$$

derived from Equation (23) with $f(z) = z^2 - 1$ has the attractors $z = \pm 1$. Their respective basins of attraction $x > 0$ and $x < 0$ are separated by the line $x = 0$ where the dynamics is ruled by Equation (24) with $z = \mathrm{i}y$. This one-dimensional map is conjugated to the dyadic map

$$\chi_{n+1} = \begin{cases} 2\chi_n - \frac{\pi}{2} & \text{if} \quad 0 \le \chi_n \le \frac{\pi}{2}, \\ 2\chi_n + \frac{\pi}{2} & \text{if} \quad -\frac{\pi}{2} < \chi_n \le 0, \end{cases} \tag{25}$$

by the transformation $y = \tan \chi$, which shows that the dynamics is chaotic on this Julia set.

However, the boundaries between the basins of attraction are typically fractal (Ott, 1993) as is the case for the Newton–Raphson map (23) with $f(z) = \mathrm{e}^z - 1$ (see Figure 3).

Maps and Probability

An important issue is to understand how maps evolve probability in their phase space. The time evolution of probability densities is ruled by the so-called Frobenius–Perron equation (Lasota & Mackey, 1985). The probability density at the current point $\mathbf{x}$ comes from all the points $\mathbf{y}$ that are mapped onto $\mathbf{x}$. Since the inverse of an endomorphism is not unique, the Frobenius–Perron equation is composed of the sum

$$\rho_{n+1}(\mathbf{x}) = \sum_{\mathbf{y}:\, \boldsymbol{\phi}(\mathbf{y}) = \mathbf{x}} \frac{\rho_n(\mathbf{y})}{\left| \det \frac{\partial \boldsymbol{\phi}}{\partial \mathbf{x}}(\mathbf{y}) \right|}. \tag{26}$$

For an automorphism, the sum reduces to the single term corresponding to the unique inverse. An invariant probability measure is obtained as a solution of the Frobenius–Perron equation such that $\rho_{n+1}(\mathbf{x}) = \rho_n(\mathbf{x})$. The study of invariant measures is the subject of ergodic theory (Hopf, 1937; Arnol'd & Avez, 1968; Cornfeld et al., 1982). The knowledge of the ergodic invariant measure provides us with the statistics of the quantities of interest: observables, correlation functions, Lyapunov exponents, the Kolmogorov–Sinai entropy, etc. With these tools, transport properties such as normal and anomalous diffusion can also be studied in maps (Lichtenberg & Lieberman, 1983).

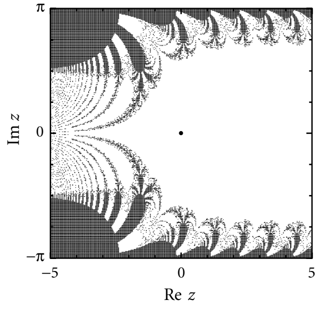

Figure 3. Complex map $z_{n+1} = z_n - 1 + \exp(-z_n)$: basin of attraction of the point at infinity in grey and of the attractors $z = 0, \pm 2\pi\mathrm{i}, \pm 4\pi\mathrm{i}, \ldots$ in white. The dot is the attractor $z = 0$.

Some Applications

Dissipative maps are used to study chaos in hydrodynamics (Lorenz, 1963), chemical kinetics (Scott, 1991), biology (Olsen & Degn, 1985; Murray, 1993), nonlinear optics (Ikeda et al., 1980), and more. In particular, systems with time delay in some feedback can be approximated by maps, as in the nonlinear optics of a ring cavity (see Figure 4).

Dissipative maps are also used to study complex systems composed of many interacting units. The units may form a lattice or a graph and interact with each other by diffusive or global couplings. These high-dimensional maps are often called *coupled map lattices* in reference to their spatial extension.

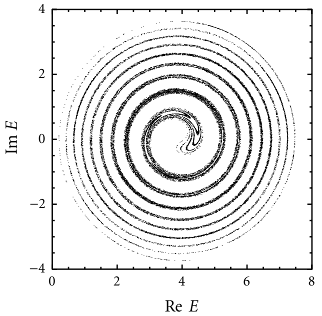

Figure 4. Chaotic attractor of the dissipative Ikeda map $E_{n+1} = a + bE_n \exp(\mathrm{i}\,|E_n|^2 - \mathrm{i}c)$ ruling the complex amplitude $E_n \in \mathbb{C}$ of the electric field of light transmitted in a ring cavity containing a nonlinear dielectric medium, at each passage along the ring (Ikeda et al., 1980). The parameters take the values $a = 3.9$, $b = 0.5$, and $c = 1$.

In addition, area-preserving and symplectic maps have become a fundamental tool to study the long-term evolution of the Solar system (Murray & Dermott, 1999).

PIERRE GASPARD

See also **Anosov and Axiom A systems; Attractors; Aubry–Mather theory; Billiards; Cat map; Chaotic dynamics; Coupled map lattice; Denjoy theory; Entropy; Ergodic theory; Hamiltonian systems; Horseshoes and hyperbolicity in dynamical systems; Kolmogorov–Arnol'd–Moser theorem; Lyapunov exponents; Maps in the complex plane; One-dimensional maps; Phase space; Symbolic dynamics**

Further Reading

Arnol'd, V.I. & Avez, A. 1968. *Ergodic Problems of Classical Mechanics*, New York: Benjamin

Benedicks, M. & Carleson, L. 1991. The dynamics of the Henon map. *Annals of Mathematics*, 133: 73–169

Chirikov, B.V. 1979. A universal instability of many-dimensional oscillator systems. *Physics Reports* 52: 263–379

Cornfeld, I.P., Fomin, S.V. & Sinai, Ya.G. 1982. *Ergodic Theory*, Berlin: Springer

Devaney, R.L. 1986. *An Introduction to Chaotic Dynamical Systems*, Menlo Park CA: Benjamin/Cummings

Faddeev, L. & Volkov, A.Yu. 1994. Hirota equation as an example of an integrable symplectic map. *Letters in Mathematical Physics*, 32: 125–135

Gumowski, I. & Mira, C. 1980. *Recurrences and Discrete Dynamical Systems*, Berlin: Springer

Hénon, M. 1976. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics*, 50: 69–77

Hopf, E. 1937. *Ergodentheorie*, Berlin: Springer

Ikeda, K., Daido, H. & Akimoto, O. 1980. Optical turbulence: chaotic behavior of transmitted light from a ring cavity. *Physical Review Letters*, 45: 709–712

Jakobson, M.V. 1981. Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. *Communications in Mathematical Physics*, 81: 39–88

Lasota, A. & Mackey, M.C. 1985. *Probabilistic Properties of Deterministic Systems*, Cambridge and New York: Cambridge University Press

Lichtenberg, A.J. & Lieberman, M.A. 1983. *Regular and Stochastic Motion*, New York: Springer

Lorenz, E.N. 1963. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20: 130–141

MacKay, R.S. & Meiss, J.D. 1987. *Hamiltonian Dynamical Systems: A Reprint Selection*, Bristol: Adam Hilger

Murray, C.D. & Dermott, S.F. 1999. *Solar System Dynamics*, Cambridge and New York: Cambridge University Press

Murray, J.D. 1993. *Mathematical Biology*, Berlin and New York: Springer

Olsen, L.F. & Degn, H. 1985. Chaos in biological systems. *Quarterly Reviews of Biophysics*, 18: 165–225

Ott, E. 1993. *Chaos in Dynamical Systems*, Cambridge and New York: Cambridge University Press

Roberts, J.A.G. & Quispel, G.R.W. 1992. Chaos and time-reversal symmetry: Order and chaos in reversible dynamical systems. *Physics Reports*, 216: 63–177

Scott, S.K. 1991. *Chemical Chaos*, Oxford: Clarendon Press, and New York: Oxford University Press

MAPS IN THE COMPLEX PLANE

Iterates of rational functions in the complex plane form a class of dynamical systems that can be characterized in amazing detail and completeness. Gaston Julia (1918) and Pierre Fatou (1919/20) proved fundamental theorems on the invariant sets and the relation between critical points and attracting cycles. In some of the first numerical studies of the iterates of a quadratic polynominal, Benoit Mandelbrot discovered the set now named after him (Mandelbrot, 1980). Dennis Sullivan (1985) proved the complete classification of attractors and Adrien Douady & John Hubbard (1984) showed that the Mandelbrot set is connected. Because of their aesthetic beauty, their intricate details, and their omnipresence in iterations of complex functions, Mandelbrot and Julia sets belong to the most fascinating and most widely studied fractal objects. The book by Richter & Peitgen (1986) contains a good survey of the results and many color plates of Julia sets and Mandelbrot sets, as well as personal reflections by Mandelbrot and Douady.

In order to illustrate some of the ideas and concepts, consider the rational maps that result from Newton's method for roots of polynominals (Curry et al., 1983). Their roots can be found using Newton's method, where an initial point z_0 is iterated according to the rational map

$$z_{n+1} = g(z_n) = z_n - f(z_n)/f'(z_n)\,. \tag{1}$$

For initial conditions sufficiently close to a root, the method converges faster than quadratically, hence its popularity in numerical mathematics. But a little numerical exploration shows that when the initial condition is further away from a root, the iterations can behave rather unpredictably.

Helpful for the investigation of the global dynamics of the iterates is a theorem due to Fatou (1919/20), according to which any attracting cycle will have a critical point in its basin of attraction. A point x_p on a cycle of period k returns after k iterations to its starting point, $z_\mathrm{p} = g^{(k)}(z_\mathrm{p})$. At a critical point the derivative vanishes, $g'(z_\mathrm{c}) = 0$. Newton's method for polynominals has $g'(z) = f(z)f''(z)/(f'(z))^2$; therefore, the critical points include the roots of the polynominal and the inflection points where $f''(z_\mathrm{c}) = 0$. Since the roots of the polynominals are at the same time fixed points and critical points for Newton's method, each one of them is attracting. It thus suffices to investigate the dynamics of the inflection points.

Points that do not iterate to a root or to any other attracting object form the Julia set.

In 1879, Arthur Cayley solved the simplest case, Newton's method for the quadratic polynominal $z^2 - 1 = 0$, where the iteration reads

$$z_{n+1} = \frac{z_n^2 + 1}{2z_n}\,. \tag{2}$$

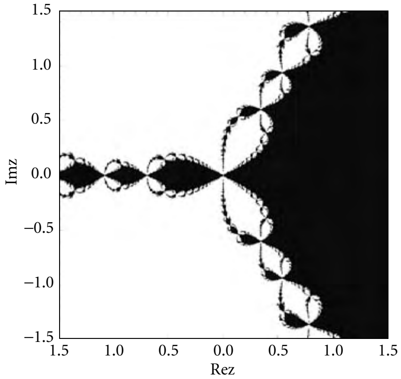

Figure 1. Initial conditions in Newton's method that converge to the root at 1 for the cubic equation $z^3 - 1$. The initial conditions for the other roots follow from symmetry by rotation through an angle of $\pm 2\pi/3$ around the origin.

There are no critical points besides the roots and hence no other attracting regions. The imaginary axis is the Julia set: it is mapped into itself and is the border between the two domains of attraction.

At the end of his paper Cayley comments that "the next succeeding case of the cubic equation appears to present considerable difficulty". Just how difficult is indicated in Figure 1 for the case $f(z) = z^3 - 1$. Evidently, not only the immediate neighborhood of 1 but also many points further away will map into the root $z_0 = 1$. The critical point $z_c = 0$ is mapped to infinity, so that there are no other attractors. The boundary of the black region is the Julia set. It has the interesting feature that in an arbitrarily small neighborhood of every point, initial conditions can be found that iterate to any one of the possible roots. That is to say, at a boundary point, the attracting regions for all roots, and not just for two roots, meet. It has self-similar and fractal features (for instance, its Hausdorff dimension is about 1.429... (Nauenberg & Schellnhuber, 1989)), but it is not of full measure.

For more general third-order polynominals, for example,

$$p_a(z) = z^3 + (a - 1)z - a\,, \qquad (3)$$

it can happen that a set of initial conditions of finite measure will not converge to any one of the roots (Curry et al., 1983). Then the iterates of the critical point $z_c = 0$ remain bounded but do not approach a root. For instance, for the parameter values underlying Figure 2, the critical point maps into a period-2 cycle. For initial conditions in the set shown, Newton's method will not find a root; instead, it will converge to a cycle of period 2.

Concerning the dependence on the parameter a, one can map out regions in the parameter space where the critical point $z_c = 0$ does not iterate toward one of the roots or to infinity: for parameter a inside the black region in Figure 3, it approaches another attracting set and Newton's method fails. Besides the case of an attracting cycle of period 2 as in Figure 2, with parameters from the main cardioid of the set, one can find cycles of period 4 in the circular bud to the left, of period 8 even further to the left, and so on: the attracting periodic orbit undergoes period doubling. For parameters in other parts of the set, other sequences of bifurcations occur.

The object that appears in Figure 3 was first seen by Benoit Mandelbrot (1980) in investigations of iterates of the family of quadratic polynominals

$$z_{n+1} = z_n^2 + c\,. \qquad (4)$$

This map has only one critical point, $z_c = 0$. For points outside the Mandelbrot set z_c iterates to infinity and there are no stable attracting cycles. The different compartments and regions inside the Mandelbrot set then contain parameter values where different cycles are stable, and transitions between regions correspond to various bifurcations. The cycles that are not attracting are dense in the Julia sets. For parameter values inside the Mandelbrot set, the Julia set is connected; for parameters outside, it dissolves into a Fatou dust. For $c = 0$, the Julia set is a circle of dimension 1, and for small $c \neq 0$, it becomes a fractal with dimension

$$d \approx 1 + \frac{|c|^2}{4 \ln 2} \qquad \text{for small } c\,. \qquad (5)$$

For real parameters and real z_n, the map belongs to the class of unimodal maps with quadratic maxima for which the period doubling cascade with its universal scaling laws in parameter and distance between periodic points applies. These relations then translate into scalings of the diameters of the buds in the Mandelbrot set and of structures in the Julia set. A fairly complete description of the structures in both the Mandelbrot set and the associated Julia sets can be achieved using methods from conformal mappings (Douady & Hubbard, 1984/85).

The previous examples already illustrate several kinds of behaviors of iterates. They can map to a fixed point of a periodic orbit, in which case the orbit is stable; if the derivative $M_p = dg^{(k)}(x_p)/dx_p$ along the orbit vanishes, as in the case of the roots for Newton's method, the orbit is superstable. Orbits that have derivatives $|M_p| > 1$ are repelling and belong to the Julia set. To complete the classification of all possible attractors as given by Sullivan (1985), we need to add the marginal cases when the derivative is of modulus one, $M_p = \exp(2\pi\alpha)$: such orbits are called rationally or irrationally indifferent for rational and irrational α, respectively. Near irrationally indifferent

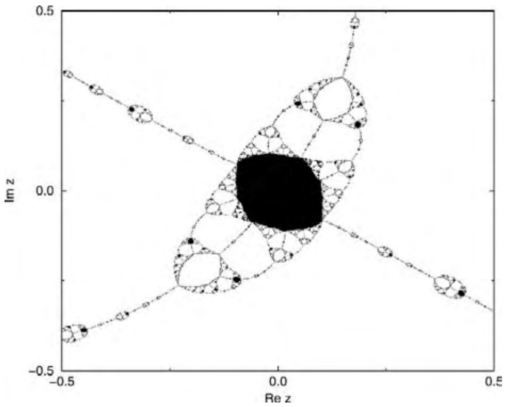

Figure 2. Initial conditions in Newton's method that do not converge to any one of the roots for the cubic polynominal (3) with parameter $a = 0.32 + 1.64i$. In the big blobs initial conditions converge to a period two cycle.

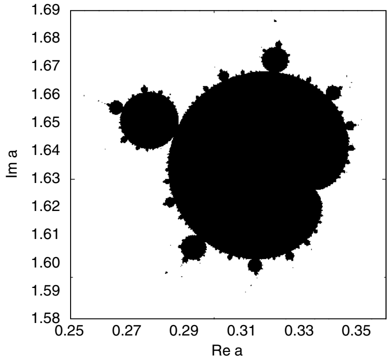

Figure 3. An example of the set of parameters a for which the iterates of the critical points for Newton's method for polynominal (3) do not approach one of the roots. In the main cardioid, the iterates approach a stable period 2 orbit, and in the buds attached to the main cardioid other orbits of higher period are stable. The small speckles outside the main object are also part of the Mandelbrot set that are connected to it by thin hairs and filaments which are not resolved in this plot.

orbits, Siegel disks, and Herman rings can appear. Sullivan's classification theorem now states that there are countably many attracting regions and that they can belong to superstable orbits, stable orbits, Siegel disks, or Herman rings. They can be identified by following iterates of critical points which will bring one to the attracting orbit or to the boundaries for the irrationally indifferent regions.

Bruno Eckhardt

See also **Fractals; Maps**

Further Reading

Curry, J., Garnett, L. & Sullivan, D. 1983. On the iteration of rational functions: computer experiments with Newton's method. *Communications of Mathematical Physics*, 91: 267–277

Douady, A. & Hubbard, J.H. 1984/85. Étude dynamique des polynômes complexes I et II. *Publications Mathematique D'Orsay* 84.02 (154 p.) and 85.04 (75 p.)

Fatou, P. 1919/20. Sur les Equations fonctionnelles. *Bulletin de la Société Mathematique de France*, 47: 161–271; 48: 33-94; 48: 208–314

Julia, G. 1918. Mémoire sur l'itération des fonctions rationelles. *Journal de Mathématiques Pures et Appliquées*, 4: 47–245

Mandelbrot, B.B. 1980. Fractal aspects of $z \to \lambda z(1 - z)$ for complex λ and z. *Annals of the New York Academy of Sciences*, 357: 249–259

Nauenberg, M. & Schellnhuber, H.J. 1989. Analytic evaluation of the multifractal properties of a Newtonian Julia set. *Physical Review Letters*, 62: 1807–1810

Richter, P.H. & Peitgen, H.O. 1986. *The Beauty of Fractals: Images of Complex Dynamical Systems*, Berlin and New York: Springer

Sullivan, D. 1985. Quasiconformal homoemorphisms and dynamics: I. Solution of the Fatou–Julia problem on wandering domains. *Annals of Mathematics*, 122: 401–418

MARANGONI CONVECTION

See **Fluid dynamics**

MARGINAL STABILITY

See **Stability**

MARKIN–CHIZMADZHEV MODEL

The Hodgkin & Huxley (1952) equations provide a quantitatively accurate and detailed model of the currents generating the propagating nerve impulse in the squid giant axon, and as the first such model they formed a prototype for nerve excitation. The equations are nonlinear, with four dynamical variables, and are in the form of a reaction-diffusion partial differential equation. Thus, the model is analytically intractable and numerical solution—now a trivial task on a PC—in the 1960s, required mainframe facilities. Simpler models were needed, both for understanding the mechanisms of propagation, and numerical exploration of propagation; such simple models are still used to simulate propagation in the anisotropic geometry of cardiac muscle (Panfilov, 1997). One approach is the FitzHugh–Nagumo equations, in which the nonlinear current-voltage relation of excitable membranes is caricatured by a cubic function (Rinzel & Keller, 1973). Another approach is to directly specify the currents flowing during the action potential.

Starting with the nonlinear cable equation for a nonmyelinated axon with axoplasmic resistance R and membrane capacitance C (both per unit length of axon), spread of membrane potential V with distance x (cm) and time t (ms) is described by

$$C\frac{\partial V}{\partial t} = \frac{1}{R}\frac{\partial^2 V}{\partial x^2} - I_{\text{ion}}, \qquad (1)$$

where I_{ion} is a nonlinear function of V and t. Markin & Chizmadzhev (1967) assumed the following simple form for the membrane ionic current I_{ion}. This current was assumed to be switched to a constant inward current J_1 at the start of excitation and, after a time τ_1, switched to a smaller, longer constant outward current J_2 for a time period τ_2. The nonlinear diffusion equation (1) is thus replaced by a piecewise linear diffusion equation. Considering a solitary traveling-wave solution with a velocity θ ms^{-1}, moving to the coordinate system $\xi = x - \theta t$ reduces Equation (1) to an autonomous ordinary differential equation

$$\frac{\partial^2 V}{\partial \xi^2} + \theta RC\frac{\partial V}{\partial \xi} - RI(\xi) = 0. \qquad (2)$$

For the Markin–Chizmadzhev model, I_{ion} in Equation (2) is linear in the four regions: (i) $\xi > 0$, (ii) $0 > \xi > -\theta\tau_1$, (iii) $-\theta\tau_1 > \xi > -\theta(\tau_1 + \tau_2)$, and (iv) $\xi < \theta(\tau_1 + \tau_2)$, so the traveling-wave solution can be constructed analytically from four components, as shown in Figure 1.

Analytic estimates for the velocities of the two traveling wave solutions, the larger being faster and stable, were obtained, these being analogous to earlier numerical studies on propagating activity of the Hodgkin–Huxley equations. This simple Markin–Chizmadzhev model for the membrane current generator retains the ratio of inward to outward current magnitudes and the relative time courses of the inward and outward membrane currents. A greater simplification is to consider the action potential as an event that is produced whenever a threshold is exceeded. Such an approach is widely used in modeling periodic and stochastic spike trains, for example, by the integrate and fire model.

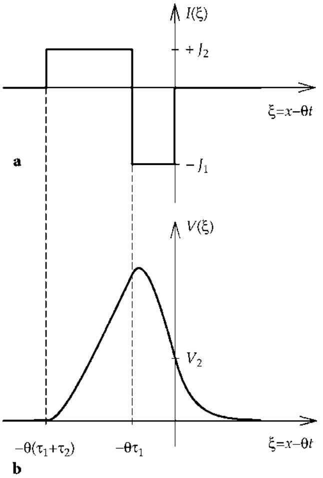

Figure 1. (a) Assumed membrane current and (b) computed potential during a solitary propagating action potential for the Markin–Chizmadzhev model for a nonmyelinated axon, with propagation velocity θ.

Markin et al. (1987) apply the membrane current generator model to provide estimates for the effects of branching and changes in axonal diameter on propagation, and the interaction between propagating activity in axons forming nerve trunks, where extracellular conduction pathways allow the possibility of ephaptic transmission (an action potential in one fiber inducing changes in potential in neighboring fibers), and an increased synchronization of propagating activity in a bundle of nerve fibers. They also applied the model to propagation of activity in syncytia—branching networks of coupled cells. The spread of excitation in such systems depends on the relative cell sizes and connections (it is easier for a large cell to excite a smaller adjacent cell). For systems of similar cells, the syncytium merges into an excitable medium as the cell-to-cell coupling is increased. Such excitable media models are widely used to model propagation in cardiac tissue.

A behavior of excitable media in which a drifting source emits periodic waves was believed to be associated with macroscopic inhomogeneities. Markin & Chizmadzhev (1972) showed that in two homogeneous coupled one-dimensional fibers, activity propagating along one can excite activity in the other and vice-versa, setting up a reverberator that acts as a drifting source of periodic wave trains in a homogeneous system whose components all have stable equilibrium solutions. This provides a prototype, with two coupled one-dimensional fibers, for re-entry, which in two-dimensional excitable media appears as a spiral wave.

The basic phenomenology of propagation in excitable media was first studied with the Markin–Chizmadzhev model, as it allowed both piecewise linear analysis and rapid numerical solution. There is still a need for simple models for rapid simulation in three-dimensional media, but the Markin–Chizmadzhev model has been superceded by an efficient two-variable system for excitable media in general (Dowle et al., 1997), and the Fenton–Karma (1998) three current model for cardiac tissue.

Arun V. Holden

See also **FitzHugh–Nagumo equation; Hodgkin–Huxley equations; Integrate and fire neuron; Nerve impulses**

Further Reading

Dowle, M., Mantel, R.M. & Barkley, D. 1997. Fast simulations of waves in three-dimensional excitable media. *International Journal of Bifurcation and Chaos*, 7: 2529–2546

Fenton, F. & Karma, A. 1998. Vortex dynamics in three-dimesional continuous myocardium with fiber rotation: filamant instability and fibrillation. *Chaos*, 11: 20–47

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve, *Journal of Physiology*, 117: 500–544

Markin, V.S. & Chizmadzhev, Yu. A. 1967. Spread of excitation in a model of the nerve fibre. *Biofizika*, 12: 900–907; *Biophyics GB*, 12: 1032–1040

Markin, V.S. & Chizmadzhev, Y.A. 1972. Properties of a multicomponent medium. *Journal of Theoreretical Biology*, 36: 61–80

Markin, V.V., Parushenko, V.F. & Chizmadzhev, Y.A. 1987. *Theory of Excitable Media*, New York: Wiley

Panfilov, A.V. 1997. Modelling of re-entrant patterns of excitation in an anatomical model of the heart. In *Computational Biology of the Heart*, edited by A.V. Panfilov & A.V. Holden, Chichester and New York: Wiley

Rinzel, J. & Keller, J.B. 1973. Traveling wave solutions of a nerve conduction equation. *Biophyical Journal*, 13: 1313–1337

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

MARKOV CHAIN

See **Stochastic processes**

MARKOV PARTITIONS

To simplify analysis of a dynamical system, we often study a topologically equivalent system using symbolic dynamics, representing trajectories by infinite length sequences using a finite number of symbols. (A simple example of this idea is the writing of real numbers as sequences of digits, a finite collection of symbols.) To represent the state space of a dynamical system with a finite number of symbols, we must partition the space into a finite number of elements and assign a symbol to each one.

Definition. A topological partition of a metric space M is a finite collection $\mathcal{P} = \{P_1, P_2, ..., P_r\}$ of disjoint open sets whose closures cover M in the sense that $M = \overline{P_1} \cup \cdots \cup \overline{P_r}$ (Lind & Marcus, 1995).

In probability theory, the term Markov denotes memorylessness. In other words, the probability of each outcome conditioned on all previous history is equal to conditioning only on the current state; no previous history is necessary. The same idea has been adapted to the dynamical systems theory to denote a partitioning of the state space so that all of the past information in the symbol sequence is contained in the current symbol, giving rise to the idea of a Markov transformation.

One-dimensional Transformations

In the special, but important case that a transformation of the interval is Markov, the symbolic dynamic is simply presented as a finite directed graph. A Markov transformation in $\mathcal{R}^1$ is defined as follows (Góra & Boyarsky, 1997):

Definition. Let $I = [c, d]$ and let $\tau : I \to I$. Let $\mathcal{P}$ be a partition of I given by the points $c = c_0 < c_1 < \cdots < c_p = d$. For $i = 1, \ldots, p$, let $I_i = (c_{i-1}, c_i)$ and denote the restriction of τ to I_i by τ_i. If τ_i is a homeomorphism from I_i onto a union of intervals of $\mathcal{P}$, then τ is said to be Markov. The partition $\mathcal{P}$ is said to be a Markov partition with respect to the function τ.

There are two key elements of the Markov partition that allow the symbol dynamics to accurately represent the system. First, on each interval I_i of the partition, the map must be monotonic (a homeomorphism), which ensures that whenever the preimage of a point is inside an interval, the preimage is unique. Second, whenever the image of a partition element intersects another element, it covers that interval. Therefore, regardless of the trajectory before entering an interval I_j, the orbit may follow any allowed trajectory from I_j. (The future evolves only from the present state.)

As a one-dimensional example, consider map 1 (Figure 1a) which is a Markov map with the associated partition $\{I_1, I_2, I_3, I_4\}$. The symbol dynamics are captured by the transition graph (Figure 1b). Although map 2 (Figure 1c) is piecewise linear and is logically

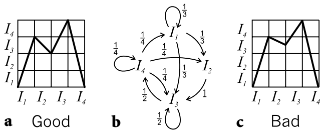

Figure 1. (a) A Markov map with partition shown. Note that on each interval, the map is a one-to-one and the image of the interval covers every interval that it intersects. (b) The transition graph for map 1. (c) The partition is not Markov ("Bad"): the image of I_2 stretches into interval I_3, but it does not completely cover that interval (and similarly for the image of I_3 with I_3).

partitioned by the same intervals as map 1, the partition is not Markov because interval I_2 does not map onto (in the mathematical sense) a union of any of the intervals of the partition. However, we are not able to say that map 2 is not Markov. There may be some other partition that satisfies the Markov condition. In general, finding a Markov partition or proving that such a partition does not exist is a difficult problem.

Higher Dimensions

Any topological partitioning of the state space will create symbol dynamics for the map. In the special case where the partition is Markov, the symbol dynamics capture the essential dynamics of the original system.

Definition. Given a metric space M and a map $f : M \to M$, a Markov partition of M is a topological partition of M into rectangles $\{R_1, \dots, R_m\}$ such that whenever $x \in R_i$ and $f(x) \in R_j$, then (Bowen, 1975; Guckenheimer & Holmes, 1983)

$$f[W^{\mathrm{u}}(x) \cap R_i] \supset W_{\mathrm{u}}[f(x)] \cap R_j$$

and

$$f[W^{s}(x) \cap R_i] \subset W_s[f(x)] \cap R_j. \qquad (1)$$

To determine if the partition is Markov, in other words, we find the stable and unstable manifolds (W^{s} and W^{u}) of each point x and its image $f(x)$, and consider the restriction of these manifolds to the partition rectangles. Thus, whenever an image rectangle intersects a partition element, the image must stretch completely across that element in the expanding (unstable) directions, but the image must be inside that partition element in the contracting (stable) direction. (See Figure 2.)

It is important to use a "good" partition so that the resulting symbolic dynamics of orbits through the partition well represents the dynamical system. If the partition is Markov, then goodness is most easily ensured. However, a broader notion, called generating partition, may be necessary to capture the dynamics. A Markov partition is generating, but the converse is not generally true. See Bollt et al. (2001) and Rudolph (1990) for a discussion of the role of partitions in representing dynamical systems.

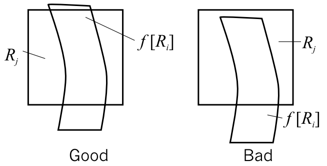

Figure 2. In the unstable (expanding) direction, the image rectangle must stretch completely across any of the partition rectangles that it intersects.

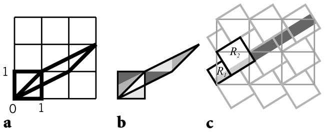

Figure 3. The cat map is a toral automorphism. (a) The operation of the linear map on the unit square. (b) Under the mod operation, the image is exactly the unit square. (c) Tessellation by rectangles R_1 and R_2 forms an infinite partition on $\mathbf{R}^2$. However, since the map is defined on the toral space $\mathbf{T}^2$, only two rectangles are required to cover the space. The filled gray boxes illustrate that R_1 and R_2 are mapped completely across a union of rectangles.

The cat map, defined by

$$\mathbf{x} = (\mathbf{Ax}) \bmod 1, \qquad (2)$$

where

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \qquad (3)$$

yields a map from the unit square onto itself. This map is said to be on the toral space $\mathbf{T}^2$ because the mod 1 operation causes the coordinate $1 + z$ to be equivalent to z. A Markov partition for this map is shown in Figure 3 (see also color plate section). The cat map is part of a larger class of functions called toral Anosov diffeomorphisms and provides a detailed description of how to construct Markov partitions for this class of maps (Robinson, 1995).

Applications

In addition to establishing the link to symbol dynamics, the Markov partition has another direct application in the one-dimensional case. In a dynamical system, we are often interested in the overall behavior of the map—the evolution of an ensemble of initial conditions. The Frobenius–Perron operator is used to describe this evolution. When the map is Markov, this operator reduces to finite-dimensional stochastic transition matrix. Following the same development as in probability theory, the stationary (invariant) density associated with these maps is described by the eigenvector for the eigenvalue $\lambda = 1$. If the system meets certain ergodic conditions, this density will describe the time average behavior of the system.

The analysis of the ensemble behavior of a dynamical system via its transition matrix is such a powerful tool that we would like to apply it to other one-dimensional systems, even when they may not be Markov. A general technique for approximating the invariant density of a map is called Ulam's method, conjectured by Ulam in 1960 and later proven by Li in 1976. The method relies upon the fact that Markov maps are dense in function space (Froyland, 2000).

Erik M. Bollt and Joe D. Skufca

See also **Cat map; Symbolic dynamics**

Further Reading

Bollt, E., Stanford, T., Lai, Y. & Życzkowski, K. 2001. What symbol dynamics do we get with a misplaced partition? On the validity of threshold crossings analysis of chaotic time-series. *Physica* D, 154: 259–286

Bowen, R. 1975. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Berlin and New York: Springer

Froyland, G. 2000. Extracting dynamical behavior via Markov models. In *Nonlinear Dynamics and Statistics: Proceedings, Newton Institute, Cambridge, 1998*, edited by A. Mees, Boston: Birkhäuser

Góra, P. & Boyarsky, A. 1997. *Laws of Chaos, Invariant Measures and Dynamical Systems in One Dimension*, Boston: Birkhäuser

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, New York: Springer

Lind, D. & Marcus, B. 1995. *An Introduction to Symbolic Dynamics and Coding*, Cambridge and New York: Cambridge University Press

Li, T.-Y. 1976. Finite approximation for the Frobenius–Perron operator. A solution to Ulam's conjecture. *Journal of Approximation Theory*, 17: 177–186

Robinson, C. 1995. *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, Ann Arbor, MI: CRC Press

Rudolph, D.J. 1990. *Fundamentals of Measurable Dynamics, Ergodic Theory on Lebesgue Spaces*, Oxford: Clarendon Press, and New York: Oxford University Press

Ulam, S. 1960. *Problems in Modern Mathematics*, New York: Interscience Publishers

MARTINGALES

The classic setting for probability is that of independent random variables. An experiment is repeated many times, and whatever happens in one trial has no effect on what happens in future trials. However, many results of probability also hold true in a much more general setting, that of martingales. The word *martingale* is associated with the concept of a gambling scheme, but its importance in probability is quite general and goes far beyond gambling. The reason is that it is rather easy to find or construct martingales, and these give useful insights into probability problems.

In roulette, the "martingale system" is to double the bet (on black or red) after each bet, until one finally wins. (See Example 4.) The word *martingale* is also used for various arrangements of constraints: part of a horse's reins, lines controlling the jib of a yacht, and the belt at the back of a jacket. The word has been traced back to Middle French before 1600 and, in some accounts, comes from the town of Martigues, presumably a center for horses and gambling.

Definition

The mathematical definition of a martingale involves conditional expectation, and this is a nonlinear concept. If $X_1, \ldots, X_n$ are random variables, then they generate a larger set $\mathcal{F}_n$ of random variables. This set $\mathcal{F}_n$ is defined as the set of all random variables W such that there exists a function f (possibly very nonlinear) with $W = f(X_1, \ldots, X_n)$. If Z is a real random variable with well-defined expectation $E[Z]$, then the "conditional expectation" $E[Z \mid \mathcal{F}_n] = h(X_1, \ldots, X_n)$ is an element of $\mathcal{F}_n$. It is determined by the condition that for every bounded random variable W in $\mathcal{F}_n$ we have $E[E[Z \mid \mathcal{F}_n]\ W] = E[ZW]$. In other words, $E[Z \mid \mathcal{F}_n]$ is the orthogonal projection of Z onto the nonlinearly generated space $\mathcal{F}_n$.

A martingale is a fair game at each step. (There is also a concept of supermartingale, an unfavorable game, and a corresponding concept of submartingale, a favorable game. These are discussed in the references.) Thus, a martingale is a sequence of random variables S_n, one for each time step. These represent the fortune of the gambler at time n. Let $\mathcal{F}_n$ be generated by the random variables that are defined by what happens up to and including time n. Thus, in particular S_n belongs to $\mathcal{F}_n$. However, the fortune S_{n+1} at the next time step typically does not belong to S_n. Past history does not determine future performance.

The martingale condition is that the conditional expectation $E[S_{n+1} \mid \mathcal{F}_n] = S_n$. That is, given the past history up to time n, the expected change in fortune at the next step into the future is zero. It follows easily that the expected fortune $E[S_n]$ does not depend on n and is equal to $E[S_0]$.

A long-held dream of gamblers was to find a gambling scheme to play a fair game and win on the average. The situation is clarified by the following theorems. The dream can be realized, but only if one ignores constraints on time and capital.

Theorems and Examples

Some theorems involving martingales and examples are as follows.

Theorem. *A martingale that is bounded above or bounded below must converge almost surely. Thus there is a random variable* S_∞ *such that* $S_n \to S_\infty$ *as* $n \to \infty$ *with probability one.*

Theorem. *A martingale that is bounded above and below is fair in the limit. That is,* $E[S_\infty] = E[S_0]$.

Example 1. Symmetric random walk. A symmetric random walk is obtained by starting with zero and then making each future step equal to $+1$ or -1, each independently with probability $\frac{1}{2}$. The walk is the sum of the steps. This is a martingale. In fact, it is a sum of independent random variables. It is unbounded above and below, and it does not converge.

Example 2. Stop when ahead. Consider an integer b with $0 < b$. When the random walk first reaches b, make every future step equal to 0. Notice that these future steps are not independent of the past steps. This is a martingale. At each stage it is a fair game. There is no possibility of winning more than b, but one can be very far in debt. Since this martingale is bounded above, it must converge almost surely, and it can only converge to the constant value b. Thus, a gambler with unlimited credit and unlimited time can play a fair game and be almost sure to win a fixed specified amount.

Example 3. Stop when ahead or behind. Let $a < 0 < b$. When the random walk first reaches a or b, make every future step equal to 0. This is a martingale. It is bounded both above and below. This martingale converges almost surely to a random value that is either a or b. Furthermore, the expected value of the eventual winnings is zero.

Example 4. Double the bet and stop when ahead. The first step is ± 1 as before. For a while, each future step is either positive or negative with equal probability, but twice the size of the previous step. After the first positive step, future steps are zero. Thus, for example, the first steps might be $-1, -2, -4, -8, +16$. The result is that the gambler is ahead by one. This is again a martingale. Since it is bounded above, it converges almost surely to 1. This is favorable to the gambler.

Example 5. Double the bet, start again when ahead. The strategy is the same as in the previous example. However, once the gambler is ahead by one, the game is repeated until the gambler is ahead by two. Then it is repeated again, and so on. This too is a martingale. It has the remarkable property that it diverges almost surely to $+\infty$.

The same ideas may be formulated via the concept of martingale difference. Let $X_0, X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables. This is a stochastic process with a discrete time index. (Most of the concepts discussed below extend to continuous time stochastic processes, but this generalization is left to the references.) Let $\mathcal{F}_n$ be the set of all random variables $f(X_0, X_1, X_2, \ldots, X_n)$ that are functions of $X_0, \ldots, X_n$. A sequence of real random variables Y_n belonging to $\mathcal{F}_n$ is called a "martingale difference" if for each $n \geq 0$ the conditional expectation

$$E[Y_{n+1} \mid \mathcal{F}_n] = 0. \tag{1}$$

It is easy to see that if S_0 is in $\mathcal{F}_0$ and the Y_n form a martingale difference, then the sum $S_n = S_0 + Y_1 + Y_2 + \cdots + Y_n$ is a martingale.

The advantage of martingales is that they are easy to create. One general method is to produce martingale differences by subtracting conditional expectations. Thus, if each Z_n is in $\mathcal{F}_n$, then

$$Y_{n+1} = Z_{n+1} - E[Z_{n+1} \mid \mathcal{F}_n] \tag{2}$$

is a martingale difference. However, it is also possible to create martingales by other natural constructions, as shown by the following examples.

Example 6. A branching process. Start with W_0 individuals at stage 0. At each stage n there are W_n individuals. The jth individual in the nth generation has $X_j^{(n+1)}$ children, independently of all other individuals. This number of children is random with expectation $\mu > 0$. Thus the $n + 1$th generation has

$$W_{n+1} = X_1^{(n+1)} + \cdots + X_{W_n}^{(n+1)} \tag{3}$$

individuals. Since

$$E[W_{n+1} \mid \mathcal{F}_n] = \mu W_n, \tag{4}$$

the sequence $S_n = W_n/\mu^n$ is a martingale. In particular, the expected size of the nth generation is

$$E[W_n] = \mu^n E[W_0]. \tag{5}$$

However, this average behavior gives a rather misleading picture of the branching process. The martingale S_n is bounded below by zero, and so it converges to some random value S_∞ almost surely. When $\mu \leq 1$, then the population goes extinct, and so $S_n = W_n/\mu^n \to 0$ almost surely as $n \to \infty$. However, in the case $\mu > 1$, it may be shown that the martingale is fair in the limit. Thus $S_n = W_n/\mu^n \to S_\infty$, where $S_\infty \geq 0$ is random with expectation $E[S_\infty] = E[S_0] = E[W_0]$. If the population does not die out fairly soon, then it has exponential growth: asymptotically $W_n \sim S_\infty \mu^n$.

Example 7. Extinction of a branching process. Let $\mu > 1$ and let ρ be the probability of extinction starting with just one member of the population. Then ρ^{W_n} is a bounded martingale. It converges to 1 when the population goes extinct, and it converges to zero when the population goes to infinity. The fact that it remains fair in the limit says that the probability of extinction starting with w individuals is ρ^w. Each individual's line must die out independently.

Markov Chain Examples

Let $X_0, X_1, X_2, \ldots, X_n, \ldots$ be a Markov chain. (Continuous time Markov processes may also be treated, but that subject is also left to the references.) Let f be a real function of the state of the chain.

Then the present value $f(X_n)$ is a random variable. By the definition of a Markov chain, the expectation of a future value $f(X_{n+1})$, given the past $\mathcal{F}_n$, generated by $X_0, \ldots, X_n$, depends only on the present state X_n. That is,

$$E[f(X_{n+1}) \mid \mathcal{F}_n] = (Pf)(X_n), \tag{6}$$

where Pf is the application of the transition probability operator to the function f. (One can think of f as a column vector and P as multiplication by a square matrix, so Pf is another column vector.) Thus, in the special case when $Pf = f$, the sequence $f(X_n)$ is a martingale. It depends only on the current state. Such functions f are of particular interest when the chain has transient states that can lead to distinct recurrent classes.

Example 8. Asymmetric random walk (gambler's ruin). Let the Markov chain X_n be the random walk that starts at 0 and steps by $+1$ with probability p and steps by -1 with probability q, where $p + q = 1$. To be realistic in the gambling situation, take $0 < p < q < 1$. It is easy to check that the modified game $S_n = (q/p)^{X_n}$ is a martingale.

Example 9. Stop when ahead. Let $0 < b$. When the random walk X_n first reaches b, future steps are zero. Then $S_n = (q/p)^{X_n}$ is a bounded martingale. Therefore, it must converge almost surely and remain fair in the limit. The gambler either wins or goes further and further into debt. It follows from the fact that the modified game is fair in the limit that $1 = (q/p)^b P[X_n \to b]$. Thus, the probability of winning is $P[X_n \to b] = (p/q)^b$.

Example 10. Stop when ahead or behind. Let $a < 0 < b$. When the random walk X_n first reaches a or b, future steps are zero. It follows that $1 = (q/p)^a P[X_n \to a] + (q/p)^b P[X_n \to b]$. From this, it is easy to work out the probabilities of winning or losing. The probability of winning is less than in the last example, but the gambler is protected from catastrophe.

If f is a function of the state of a Markov chain, but $Pf \neq f$, then there is still an associated martingale, but it has a different character. Let

$$Y_{n+1} = f(X_{n+1}) - (Pf)(X_n). \tag{7}$$

Then Y_n forms a martingale difference sequence. Let $S_0 = f(X_0)$ and form the martingale S_n as before. Then the neighboring terms group together, and we get

$$\begin{aligned} S_n &= u(X_0) + u(X_1) + u(X_2) + \cdots + u(X_{n-1}) \\ &\quad + f(X_n), \end{aligned} \tag{8}$$

where $u = f - Pf$. The martingale is a cumulative sum over the entire history.

For this result to be useful, it is necessary to find an interesting function u for which there is a solution f of $u = f - Pf$. This often happens in the context of an irreducible Markov chain with only positive recurrent states. Let π be the invariant probability vector for the Markov chain. (One can think of π as a row vector satisfying $\pi P = \pi$.) Then a necessary condition for a solution is that $\pi u = 0$. For instance, one can take a function h and define $u = h - (\pi h)1$. If f is bounded, then $f(X_n)/n \to 0$, and in suitable circumstances the strong law of large numbers for martingales (see below) gives

$$\frac{u(X_0) + u(X_1) + u(X_2) + \cdots + u(X_{n-1})}{n} \to 0 \tag{9}$$

almost surely. In terms of the function h this says that

$$\frac{h(X_0) + h(X_1) + u(X_2) + \cdots + h(X_{n-1})}{n} \to \pi h \tag{10}$$

almost surely, where πh is the expectation computed with the invariant probability (the product of the probability row vector π with the column vector h). This is a strong law of large numbers for Markov chains. It is the idea underlying the Monte Carlo calculation of unknown invariant probabilities π.

General Results

Many classical results for sums of independent random variables have generalizations to the martingale setting. These include the strong law of large numbers and the central limit theorem.

The Kolmogorov form of the strong law of large numbers says the following. Let $Y_1, \ldots, Y_n, \ldots$ be a sequence of martingale difference random variables with finite variances $\sigma_1^2, \ldots, \sigma_n^2, \ldots$. Assume that the variances are uniformly bounded, or more generally that

$$\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty. \tag{11}$$

Then as $n \to \infty$ the sample means

$$\bar{Y}_n = \frac{Y_1 + \cdots + Y_n}{n} \to 0 \tag{12}$$

almost surely (that is, with probability one). This is the law of averages in a very powerful and general form.

Fluctuations about the average are described by the central limit theorem. The setting for this theorem is a sequence $Y_1, \ldots, Y_n, \ldots$ of martingale difference random variables with finite variances $\sigma_1^2, \ldots, \sigma_n^2, \ldots$. Let

$$s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2 \tag{13}$$

be the variance of the sum $Y_1 + \cdots + Y_n$.

The martingale differences satisfy the conditional variance normalization condition if

$$\frac{1}{s_n^2} \sum_{k=1}^{n} E[Y_k^2 \mid \mathcal{F}_{k-1}] \to 1 \tag{14}$$

in probability as $n \to \infty$.

Fix $\varepsilon > 0$. Consider one of the random variables Y_i. Say that it is large if $|Y_i| > \varepsilon s_n$. Define the large part of Y_i to be the random variable $Y_i^{\varepsilon s_n}$ that is equal to Y_i when Y_i is large and is equal to zero otherwise. The Lindeberg condition is that the contribution of the large values to the total variance is small, in the sense that for each $\varepsilon > 0$

$$\frac{1}{s_n^2}\sum_{i=1}^{n} E[(Y_i^{\varepsilon s_n})^2] \to 0 \qquad (15)$$

as $n \to \infty$.

The central limit theorem states that if $Y_1, \ldots, Y_n, \ldots$ are martingale difference random variables with finite variances that satisfy the conditional variance normalization condition and the Lindeberg condition, then the distribution of

$$Z_n = \frac{Y_1 + \cdots + Y_n}{s_n} \qquad (16)$$

approaches the distribution of a standard normal random variable Z as $n \to \infty$. That is, the Gaussian distribution gives a universal description of fluctuations of martingales.

William G. Faris

See also **Random walks; Stochastic processes**

Further Reading

Doob, J.L. 1953. *Stochastic Processes*, New York: Wiley

Durrett, R. 1991. *Probability: Theory and Examples*, Belmot, CA: Duxbury Press

Hall, P. & Heyde, C.C. 1980. *Martingale Limit Theory and Its Application*, New York: Academic Press

Nelson, E. 1987. *Radically Elementary Probability Theory*, Princeton, NJ: Princeton University Press (This book presents probability through the martingale central limit theorem in the context of finite but non-standard probability spaces. It is an unusual treatment, but inspiring.)

Neveu, J. 1975. *Discrete Parameter Martingales*, Amsterdam: North-Holland

MATHIEU EQUATION

See **Surface waves**

MATTER, NONLINEAR THEORY OF

A nonlinear theory of matter was first proposed in 1912 by Gustav Mie in a prescient series of papers that aimed to derive the elementary particles of matter as localized lumps of energy in a nonlinear field (Mie, 1912). To this end, Mie suggested a nonlinear augmentation of James Maxwell's electromagnetic equations out of which the electron would arise in a natural way.

Specifically, he defined a Lagrangian density ($\mathcal{L}$) depending upon electric field intensity ($\boldsymbol{E}$) and magnetic flux density ($\boldsymbol{B}$) and the four components of the electromagnetic potential ($\boldsymbol{A}, \phi$). Requiring dependence on the parameters $\eta \equiv (\boldsymbol{B}^2 - \boldsymbol{E}^2)$ and $\chi \equiv (\phi^2 - \boldsymbol{A}^2)$ ensured relativistic invariance, and the specific choice $\mathcal{L} = \eta/2 + a\chi^3/6$ led to a static, spherically symmetric electric potential (ϕ) of the form

$$\phi(r) \approx \frac{(3r_0^2/a)^{1/4}}{\sqrt{r^2 + r_0^2}}. \qquad (1)$$

Setting $4\pi(3r_0^2/a)^{1/4} = e$ (the electronic charge) yielded a spherically symmetric model for the electron with a radius of r_0 and electric potential

$$\phi(r) \to e/4\pi r$$

as $r \to \infty$. The Lorentz invariance that is built into the theory permits this solution to travel with any speed up to the limiting velocity of light with an appropriate Lorentz contraction.

Mie's approach to a nonlinear theory of matter was supported by Albert Einstein, who offered the following opinion in the mid-1930s (Einstein, 1954).

> In the foundation of any consistent field theory, the particle concept must not appear in addition to the field concept. The whole theory must be based solely on partial differential equations and their singularity-free solutions.

Empirical support for this perspective was provided in the early 1930s by Carl Anderson's discovery of positron-electron creation from cosmic radiation. In other words, massive particles were observed to emerge from and collapse back into an electromagnetic field, which is not a property of the linear Maxwell equations.

Motivated by Anderson's observation, Max Born revisited Mie's nonlinear electromagnetics. Together with Leopold Infeld, he eliminated the χ-dependence in Mie's functional formulation and chose instead the Lagrangian density (Born & Infeld, 1934)

$$\mathcal{L} = E_0^2\sqrt{1 + (\boldsymbol{B}^2 - \boldsymbol{E}^2)/E_0^2} - E_0^2, \qquad (2)$$

where E_0 sets the field intensities at which nonlinearities arise. At low-field amplitudes, Equation (2) reduces to the classical Lagrangian density for Maxwell's equations: $\mathcal{L} = (\boldsymbol{B}^2 - \boldsymbol{E}^2)/2$. (Even with the currently available, high-intensity lasers, these vacuum nonlinearities would be difficult to observe, as E_0 is estimated to be about 10^{20} V/m in laboratory units.)

Among the solutions of this system, Born and Infeld found a spherically symmetric model electron with $\boldsymbol{E}$ finite everywhere, although the electric displacement (**D**) exhibits a singularity at the origin. Erwin Schrödinger became interested in Born's theory as early as 1935 and continued working on it through the 1940s when—as founding director of the Dublin Institute for Advanced Studies—he attempted to move research in physics toward key areas of nonlinear science (Schrödinger, 1935).

Plane waves derived from Equation (2) obey the equation $(1 - u_t^2)u_{xx} + 2u_x u_t u_{xt} - (1 + u_x^2)u_{tt} = 0$, where u is a component of the vector potential. Called the Born–Infeld equation, this system has been studied as an interesting nonlinear wave equation (Barbashov & Chernikov, 1966).

Einstein's conviction that a consistent theory for particle physics must be based on localized solutions of nonlinear partial differential equations was shared by several of his colleagues. In addition to Mie, Born, Infeld, and Schrödinger, Werner Heisenberg (1966), Louis de Broglie (1960, 1963), and David Bohm (1957) have suggested nonlinear field theories that in their simplest representations, can be viewed as the augmentation of linear field equations by a nonlinear term of the form $|u|^2 u$, as in the nonlinear Schrödinger equation. This nonlinearity conserves the integral

$$\int |u|^2 \, \mathrm{d}\mathbf{r}, \tag{3}$$

which can be interpreted as a mass.

The ideas of de Broglie and Bohm are related to those of the inverse scattering method (ISM). In their "theory of the double solution," the real particle is a localized solution of a nonlinear equation with the form

$$u = U\mathrm{e}^{\mathrm{i}\theta'}.$$

Associated with this localized nonlinear solution is the solution of a corresponding linear equation

$$\psi = \Psi \mathrm{e}^{\mathrm{i}\theta}$$

with

$$\theta = \theta' \tag{4}$$

except in a small region surrounding the real particle. The function ψ is taken to be a solution of Schrödinger's quantum mechanical wave equation, and the phase condition of Equation (4) allows the particle to be guided by ψ. Similarly, in the context of the ISM, the nonlinear solution of a soliton equation is guided through space-time by the linear asymptotic solution of the associated linear operator. Although proposed almost a half century ago, the de Broglie–Bohm theory continues to offer possibilities for further studies (Holland, 1993).

During the 1960s, several investigators proposed the sine-Gordon (SG) equation as a field theory for elementary particles in one space dimension and time (Scott, 2003). This work gained momentum in the 1970s when it became known that special properties of the SG equation allow the corresponding quantum problem to be solved, showing that certain qualitative properties of the classical solution survive quantization (Dashen et al., 1974; Faddeev, 1975; Goldstone & Jakiw, 1975). In particular, the classical field energy was found to be a useful first approximation for the soliton mass, with quantum effects coming in as second-order corrections. More recently, this approach has been developed into the concept of a skyrmion, which is a generalization of the SG kink, carrying its topological stability into three space dimensions.

These examples suggest that it may be possible to develop a nonlinear theory of matter by proceeding as follows. First, guess a classical version of the correct nonlinear field. Second, solve this classical system for salient aspects of localized behavior. Third, analyze the corresponding quantum theory to obtain exact values for the mass spectrum. Finally, compare calculated values of mass with measured mass spectra. As Einstein was aware, however, this is a daunting program because there are no theoretical bounds on the range of conceivable nonlinear theories; thus, it is not surprising to find a proliferation of partially evaluated theories. In addition to the Born–Infeld, de Broglie, and skyrmion formulations, present candidates include string theory and the Yang–Mills equation. What others are out there?

Alwyn Scott

See also **Born–Infeld equations; Hodograph transform; Inverse scattering method or transform; Particles and antiparticles; Skyrmions; String theory; Yang–Mills theory**

Further Reading

Barbashov, B.M. & Chernikov, N.A. 1966. Solution of the two plane wave scattering problem in a nonlinear scalar field theory of the Born–Infeld type. *Soviet Physics JETP*, 23: 1025–1033

Bohm, D. 1957. *Causality and Chance in Modern Physics*, London: Routledge & Kegan Paul

Born, M. & Infeld, L. 1934. Foundations of a new field theory. *Proceedings of the Royal Society* (*London*) A, 144: 425–451

de Broglie, L. 1960. *Nonlinear Wave Mechanics*, Amsterdam: Elsevier

de Broglie, L. 1963. *Introduction to the Vigier Theory of Elementary Particles*, Amsterdam: Elsevier

Dashen, R.F., Hasslacher, B. & Neveu, A. 1974. Particle spectrum in model field theories from semiclassical functional integral techniques. *Physical Review* D, 11: 3424–3450

Einstein, A. 1954. *Ideas and Opinions*, New York: Crown

Faddeev, L.D. 1975. Hadrons from leptons? *JETP Letters*, 21: 64–65

Goldstone, J. & Jakiw, R. 1975. Quantization of nonlinear waves. *Physical Review* D, 11: 1486–1498

Heisenberg, W. 1966. *Introduction to the Unified Field Theory of Elementary Particles*, New York: Wiley

Holland, P.R. 1993. *The Quantum Theory of Motion*, Cambridge and New York: Cambridge University Press

Mie, G. 1912. Grundlagen einer Theorie der Materie. *Annalen der Physil*, 37: 511–534; 39: 1–40; 40 (1913): 1–66

Schrödinger, E. 1935. Contributions to Born's new theory of the electromagnetic field. *Proceedings of the Royal Society* (*London*) A, 150: 465–477

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

MAXWELL–BLOCH EQUATIONS

The Maxwell–Bloch (MB) equations arise in nonlinear optics, where they couple "two-level atoms" to Maxwell's electromagnetic field equations to model a nonlinear dielectric. Although the two-level atom is a pseudo-atom with only two nondegenerate energy levels, it can be realized in practice by real atoms excited by light close to a suitable atomic resonance, and it can be modeled as a two-state system: $|g\rangle$ (for ground) and $|e\rangle$ (for excited). Transitions between the two states are allowed, and there is a quantum mechanical matrix element $p = q\langle g\,|r|\,e\rangle \neq 0$ between them in dipole approximation, where q is the electronic charge and qr is the dipole transition operator.

The outer products of $|e\rangle$ and $|g\rangle$ form a Lie algebra with

$$\begin{aligned} S^z &= \tfrac{1}{2}\left(|e\rangle\langle e| - |g\rangle\langle g|\right), \\ S^+ &= |e\rangle\langle g|, \\ S^- &= |g\rangle\langle e|. \end{aligned}$$

By choosing $\langle e \mid e\rangle = \langle g \mid g\rangle = 1$, $\langle e \mid g\rangle = \langle g \mid e\rangle = 0$ (orthonormality), one finds that the Lie algebra is

$$\begin{aligned} [S^z, S^+] &= S^+, \quad [S^z, S^-] = -S^-, \\ [S^+, S^-] &= 2S^z, \end{aligned}$$

where $[\,,]$ is the Lie bracket or "commutator" (Bullough et al., 1995b). Constructed from $|e\rangle$ and $|g\rangle$, this is a two-dimensional representation of su(2), corresponding to a spin-$\frac{1}{2}$ system in which $|e\rangle$ is spin-up and $|g\rangle$ is spin-down. In a magnetic field $\boldsymbol{B} = (B_x, B_y, B_z)$, a magnetic dipole μ has a Hamiltonian which can be taken in the form of a 2 × 2 Hamiltonian matrix with elements (Feynman et al., 1966, pp. 10–14)

$$\begin{aligned} H_{11} &= -\mu B_z, \quad H_{12} = -\mu(B_x - \mathrm{i}B_y), \\ H_{21} &= -\mu(B_x + \mathrm{i}B_y), \quad H_{22} = +\mu B_z. \end{aligned}$$

There is an exact correspondence for the dipole p of the two-level atom in an electric field $\boldsymbol{E}$, and the two-level atom is a pseudo-spin-$\frac{1}{2}$ system (Bullough et al., 1995b). As any state $|\psi(t)\rangle$ at time t in the two-dimensional Hilbert space can be written $|\psi(t)\rangle = c_1(t)\,|e\rangle + c_2(t)\,|g\rangle$, the correspondence extends to the dynamics. From the Hamiltonian H_{ij}, Schrödinger's equation for the amplitudes $c_i(t)(i = 1, 2)$ becomes

$$\begin{aligned} \mathrm{i}\hbar\dot{c}_1 &= -\mu[B_z c_1 + (B_x - \mathrm{i}B_y)c_2], \\ \mathrm{i}\hbar\dot{c}_2 &= -\mu[(B_x + \mathrm{i}B_y)c_1 - B_z c_2]. \end{aligned}$$

When $B_x = B_y = 0$, $B_z = \text{constant}$ independent of t, $c_1(t), c_2(t)$ evolve as $c(0)\mathrm{e}^{\pm\mathrm{i}\omega_0 t}$, and $\omega_0 = 2\mu B_z\hbar^{-1}$ is a Larmor frequency for the spinning (precessing) magnet of moment μ. The free two-level atom with resonance frequency ω_0 thus *acts* as a true spin-$\frac{1}{2}$ in a fixed magnetic field B_z.

If we construct the Bloch vector $\boldsymbol{r}(t) = (r_1(t), r_2(t), r_3(t))$

$$\begin{aligned} r_1 &= (c_1c_2^* + c_1^*c_2), \; r_2 = -i(c_1c_2^* - c_1^*c_2), \\ r_3 &= (|c_1|^2 - |c_2|^2) \end{aligned}$$

(c_1^* = complex conjugate of c_1), the equations of motion for $c_1(t), c_2(t)$ become the Bloch equation

$$\dot{\boldsymbol{r}} \equiv \mathrm{d}\boldsymbol{r}/\mathrm{d}t = \boldsymbol{\omega} \times \boldsymbol{r} \qquad (1)$$

for a spin-$\frac{1}{2}$ particle, where $\omega \equiv (-2\mu B_x\hbar^{-1}, -2\mu B_y\hbar^{-1}, -2\mu B_z\hbar^{-1})$. (For the two-level atom in a real electric field $E(t)$, $\boldsymbol{\omega} \equiv (-2pE\hbar^{-1}, 0, \omega_0)$.) The normalization $|c_1|^2 + |c_2|^2 = 1$ means $|\boldsymbol{r}|^2 = 1$, and the motion is confined to the surface of a sphere (the Bloch sphere) of unit radius: spin-up = (0,0,1), spin down = (0,0,-1) on this sphere. This description omits a Berry's phase.

Two-level atoms enter laser physics via the Jaynes–Cummings (JC) model, which couples one two-level atom to a single mode of the quantized electromagnetic field of frequency ω. The Hamiltonian is

$$H = \omega_0 S^z + \omega a^\dagger a + g(a^\dagger S^- + S^+ a),$$

where g is a (real and positive) coupling constant, S^z, $S^\pm$ satisfy the su(2) algebra given above, and a, $a^\dagger$ satisfy $[a, a^\dagger] = 1$ and the Heisenberg–Weyl algebra of standard bosons.

The number operator $N = S^z + a^\dagger a + 1/2$ commutes with H. The JC model is thus quantum integrable, there being two degrees of freedom (spin and the quantum oscillator), and there are exactly two commuting constants H and N. The model can be solved exactly in terms of 2 × 2 matrices (Bullough et al., 1995a) and also by the quantum inverse method in Bogoliubov et al. (1996). By coupling the JC model to a heat-bath, one obtains the master equation for a micromaser. (Such a nonlinear quantum device is in operation at the Max Planck Institute, Garching, Germany, using ^{85}Rb atoms which enter a cavity in their upper states $|e\rangle$ but may leave it in $|e\rangle$ or $|g\rangle$.)

The JC model evolves only in time and is a fundamental nonlinear quantum model. One obtains important nonlinear quantum field theories by coupling two-level atoms to Maxwell's electromagnetic field equations

$$\nabla^2\hat{\boldsymbol{E}} - c^{-2}\partial^2\hat{\boldsymbol{E}}/\partial t^2 = 4\pi n c^{-2}\,\partial^2\hat{\boldsymbol{P}}/\partial t^2. \qquad (2)$$

Here, c is the speed of light in a vacuum, n is the number of two-level atoms per unit volume, and $\hat{\boldsymbol{E}} = \hat{\boldsymbol{E}}(\boldsymbol{x}, t)$ and $n\hat{\boldsymbol{P}} = n\hat{\boldsymbol{P}}(\boldsymbol{x}, t)$ are, respectively, the electric field and dipole density operators. This operator Maxwell equation and the operator Bloch equation for the dipole density constitute the quantum operator form of the MB equations.

If the electric fields are strong enough (many photons), both $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{P}}$ can be regarded as classical variables, while quantum mechanics still enters through the atoms. The classical electric field $\boldsymbol{E}(\boldsymbol{x}, t)$ acts on the atom at $\boldsymbol{x}$ at time t through vectors $\boldsymbol{\omega}(\boldsymbol{x}, t)$ given at points $\boldsymbol{x}$ and time t through

$$\boldsymbol{\omega}(\boldsymbol{x}, t) \equiv (-2\boldsymbol{p} \cdot \boldsymbol{E}(\boldsymbol{x}, t)\hbar^{-1}, 0, \omega_0)$$

and the Rabi frequency is $|2\boldsymbol{p} \cdot \boldsymbol{E}(\boldsymbol{x}, t)|\, \hbar^{-1}$. This form of $\boldsymbol{\omega}(\boldsymbol{x}, t)$ shows the coupling between the Maxwell equations and the Bloch equations to form the semiclassical MB equations, and the coupling is nonlinear since $\mathbf{E}(\mathbf{x}, t)$ is driven by $\boldsymbol{p}r_1(\boldsymbol{x}, t)$ for $\boldsymbol{P}(\boldsymbol{x}, t)$ in the (linear) Maxwell equation.

In one space dimension (x) one obtains the standard form of the semiclassical MB equations (Eilbeck et al., 1973). The (linear) Maxwell equations for such an $E(x, t)$ and $P(x, t) = pr_1(x, t)$ is

$$\partial^2 E/\partial x^2 - (1/c^2)\partial^2 E/\partial t^2 = (4\pi np/c^2)\partial^2 r_1/\partial t^2 . \tag{3}$$

With $\omega(x, t) \equiv (-2pE(x, t)\hbar^{-1}, 0, \omega_0)$, the Bloch equation (1) can be written out explicitly as

$$\begin{aligned} \partial r_1/\partial t &= -\omega_0 r_{2,} \\ \partial r_2/\partial t &= \omega_0 r_1 + 2p\hbar^{-1}Er_{3,} \\ \partial r_3/\partial t &= -2p\hbar^{-1}Er_2. \end{aligned} \tag{4}$$

This system of four nonlinear partial differential equations (3) with (4) is not integrable, but it becomes an integrable field theory if the Maxwell equation is replaced by the unidirectional system (valid for small densities in the one space dimension)

$$\partial E/\partial x + (1/c)\partial E/\partial t = (-2\pi np/c)\partial r_1/\partial t. \tag{5}$$

Equations (4) and (5) comprise the reduced Maxwell–Bloch (RMB) equations, which can be explicitly integrated by the AKNS inverse scattering method as in Gibbon et al. (1973).

Under a slowly varying envelope and phase approximation (SVEPA), the envelope equations become the self-induced transparency (SIT) equations

$$\begin{aligned} &\partial\mathcal{E}/\partial x + (1/c)\partial\mathcal{E}/\partial t = \alpha P, \\ &\partial P/\partial t = \mathcal{E}N + \Delta\omega' Q, \\ &\partial N/\partial t = -\mathcal{E}P, \\ &\partial Q/\partial t = -\Delta\omega' P. \end{aligned} \tag{6}$$

In these SIT equations, $\Delta\omega' \equiv \omega_0' - \omega_0$ (called inhomogeneous broadening), and ω_0' is a shifted resonance frequency for any particular atom (induced by Doppler shifts, e.g.) while P, Q, N depend on $(x, t, \Delta\omega')$ and N replaces the inversion r_3 of the atom in the Bloch vector. It is an unfortunate confusion of the literature that what we call the SIT equations are also called the Maxwell–Bloch equations. Here, we reserve MB for the Maxwell equation coupled to the three Bloch equations which are *not* envelope equations.

The SIT equations (6) must be averaged over inhomogeneous broadening $\Delta\omega'$ but remain integrable. The soliton solution is the 2π-pulse of (McCall & Hahn, 1969) and there are multi-soliton solutions. Under a sharp line resonance condition, they further reduce to the standard form of the sine-Gordon (SG) equation

$$\partial^2\phi/\partial x^2 - \partial^2\phi/\partial t^2 = \sin\phi\,. \tag{7}$$

Under an SVEPA, the SG equation becomes the attractive case of the nonlinear Schrödinger (NLS) equation. The RMB, SIT, SG, and attractive NLS equations form a hierarchy of integrable nonlinear field theories in which integrability is handed down by the SVEPA in the fashion described by Calogero in 1995 (see Bullough, 2001).

The two-dimensional representations of su(2) for two-level atoms extend to three-dimensional representations of su(3) for three-level atoms, and the consequent appropriately generalized SIT equations are fundamental to electromagnetically induced transparency (EIT) and the storage of quantum information (Bullough, 2001; Hau, 2001; Haroche & Raimond, 1993; Bullough & Gibbs, 2004).

Robin Bullough

See also **Berry's phase; Lie algebras and Lie groups; Nonlinear optics; Nonlinear Schrödinger equation; Sine-Gordon equation**

Further Reading

Bogoliubov, N.M., Rybin, A.V., Bullough, R.K. & Timonen, J. 1995. Maxwell–Bloch system on a lattice. *Physical Review* A, 52: 1487–1493

Bogoliubov, N.M., Bullough, R.K. & Timonen, J. 1996. Exact solution of generalized Tavis–Cummings models in quantum optics. *Journal of Physics A: Mathematics and General*, 29: 6305–6312

Bullough, R.K. 2001. Optical solitons: twenty-seven years of the last millennium and three more years of the new? In *Mathematics and the 21st Century*, edited by A.A. Ashour & A.-S.F. Obada, Singapore: World Scientific, pp. 69–121

Bullough, R.K. & Gibbs, H.M. 2004. Information storage and retrieval by stopping pulses of light. *Journal of Modern Optics*, 51(2): 255–284

Bullough, R.K., Nayak, N. & Thompson, B.V. 1995a. Cavity quantum electrodynamics: fundamental theory of the micromaser. In *Recent Developments in Quantum Optics*, edited by R. Inguva, New York: Plenum Press

Bullough, R.K., Thompson, B.V., Nayak, N. & Boguliubov, N.M. 1995b. Microwave cavity quantum electrodynamics: I, one and many Rydberg atoms in microwave cavities; and II, fundamental theory of the micromaser. In *Studies in Classical and Quantum Nonlinear Optics*, edited by Ole Keller, New York: Nova Science Publishers

Dirac, P.A.M. 1958. *The Principles of Quantum Mechanics*, 4th edition, Oxford: Oxford University Press

Dodd, R.K., Eilbeck, J.C., Gibbon, J.D. & Morris, H.C. 1982. *Solitons and Nonlinear Wave Equations*, London: Academic Press

Eilbeck, J.C., Gibbon, J.D., Caudrey, P.J. & Bullough, R.K. 1973. Solitons in nonlinear optics I. A more accurate description of the 2π pulse in self-induced transparency. *Journal of Physics A: Mathematics and General*, 6: 1337–1347

Feynman, R.P., Leighton, R.B. & Sands, M. 1965. *The Feynman Lectures on Physics*, vol. III. *Quantum Mechanics*, Reading, MA: Addison-Wesley (second printing, 1966)

Gibbon, J.D., Caudrey, P.J., Bullough, R.K. & Eilbeck, J.C. 1973. An N-soliton solution of a nonlinear optics equation derived by a general inverse method. *Lettre al Nuovo Cimento*, 8: 775–779

Haroche, S. & Raimond, J.-M. 1993. *Scientific American*, vol. 269, April: 26–33

Hau, L.V. 2001. Frozen light. *Scientific American*, vol. 284, July: 52–59

Maimstov, A.I., Basharov, A.M., Elyatin, S.O. & Sklyarov, Yu.M. 1990. The present state of self-induced transparency theory. *Physics Reports*, 191: 1–108

McCall, S.L. & Hahn, E.L. 1969. Self-induced transparency. *Phys. Rev.*, 183(2): 457–485

MCCULLOCH–PITTS NETWORK

In 1943, Warren McCulloch and Walter Pitts published a seminal paper that described the first attempt to provide a mathematical model of a neuron (McCulloch and Pitts, 1943). This work explored the properties of networks of such mathematical neurons in relation to the working of the nervous system. The model arose from the following assumptions based on the knowledge of neurophysiology at the time

(i) The activity of the neuron is an "all-or-none" process.
(ii) A certain fixed number of synapses must be excited within the period of latent addition in order to excite a neuron at any time, and this number is independent of previous activity and position on the neuron.
(iii) The only significant delay within the nervous system is synaptic delay.
(iv) The activity of any inhibitory synapse absolutely prevents excitation of the neuron at that time.
(v) The structure of the net does not change with time.

The McCulloch–Pitts neuron is thus a binary threshold unit whose firing or activation is dependent on the values of its inputs. The neuron receives a number of excitatory (positive) or inhibitory (negative) inputs via "synaptic" connections. Excitatory synapses are equally weighted, and if the sum of these positive input signals exceeds some value, θ, the neuron fires. Otherwise, the neuron does not fire. If a signal is received on any of the negative inputs, the neuron cannot fire, regardless of the positive input values. The threshold value θ is a fixed threshold value unique to the neuron, and the model is assumed to be operating in discrete time steps.

A network of McCulloch–Pitts neurons can be assembled by connecting the outputs of neurons to the inputs of other neurons in some manner. In discussing networks of these artificial neurons,McCulloch and Pitts distinguish between nets with and without "circles." In a network without circles, it is not possible to follow a path of connections in the network from any given neuron and return to that same neuron. In other words, there are no closed causal loops. The activity in such a network will, therefore, have a feed-forward dynamic; any activity imposed on the network will propagate in a unidirectional fashion for a finite amount of time, terminating when neurons are encountered with no outgoing synaptic connections. The perceptron (Rosenblatt, 1962) and multi-layer perceptron (see e.g., Haykin, 1999) networks are examples of this kind of topology. A McCulloch–Pitts network without circles is capable of representing any statement within prepositional logic (i.e., any finite logical expression).

The dynamics of networks with circles are more complex. Activity in these networks may propagate indefinitely around the network in discrete time, so the firing activity of the network at any time may be dependent on the activity of the network at multiple stages in the past.

McCulloch and Pitts were able to show that such nets are equivalent to a Universal Turing Machine in principle. It should, however, be noted that the proofs for McCulloch–Pitts nets are existence proofs only in the sense that they do not imply an algorithm for constructing a network, with appropriate values of threshold and weight parameters, to compute a given function. Also, they do not consider computational time. Interestingly, the work of McCulloch and Pitts was also an influence in the development of the von Neumann stored program computer architecture.

Consider a McCulloch–Pitts net with circles. At any point in time, the state of the network can be defined by the current (binary) activity pattern of the neurons (i.e., which neurons are firing and which are not). The network can be thought of as a point in a configuration space of all possible activity patterns. Under some appropriately chosen scheme for updating the state of the network (synchronously or asynchronously), the activity may converge to an attractor in the configuration space, resulting in a stable state for subsequent time steps. This is the basis for attractor neural nets, as described by Hopfield in the context of simple associative memory networks (Hopfield, 1982).

MARCUS GALLAGHER

See also **Attractor neural network; Cell assemblies; Electroencephalogram at large scales; Electroencephalogram at mesoscopic scales; Neurons; Perceptron**

Further Reading

Haykin, S. 1999. *Neural Networks A Comprehensive Foundation*, 2nd edition, Upper Saddle River, NJ: Prentice-Hall

Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of National Academy of Sciences USA*, 81: 3088–3092

McCulloch, W.S. & Pitts, W. 1943. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5: 115–133

Rosenblatt, F. 1962. *Principles of Neurodynamics*, New York: Spartan

MEAN FIELD THEORIES

See **Phase transitions**

MEASURES

Measures provide the means to introduce probabilistic methods into the study of a dynamical system by supplying a notion of the size or content of sets that is additive—the measure of a union of disjoint sets is the sum of the measures of these sets.

Sets correspond to events and the measure of a set to the probability of that event. The precursor to this general notion is that of the phase volume in a classical mechanical system, which is preserved under the phase flow (Liouville theorem). In other words, if a set of initial conditions is transported by the flow for a fixed time then the set of terminal conditions has the same volume as the set of initial conditions. Put differently, the phase flow consists of volume-preserving maps. In the general context of measures, this corresponds to measure-preserving transformations (see below), which are the subject of ergodic theory. For example, the Birkhoff ergodic theorem establishes a connection between the measure of a set and the proportion of time a typical orbit spends in it. The original result in ergodic theory is the Recurrence Theorem by Poincaré (1890) according to which almost every point in a volume-preserving system has to return arbitrarily close to its initial position.

The *Lebesgue measure* coincides with volume, but it is defined on a larger collection of sets. For example, the one-dimensional volume (length) of the set of rational numbers is not defined, but this set has zero Lebesgue measure. Closely related to the Lebesgue measure are *absolutely continuous* measures, that is, those defined by integrating a nonnegative density function over the set to be measured. The function can be interpreted as a probability density if the total integral is 1. (The bell curve $p(x) := \mathrm{e}^{-(x-b)^2/2a^2}/a\sqrt{2\pi}$ is a standard example.) Taking the Dirac δ-function as a (singular) “density,” one obtains the *Dirac measure* defined by $\delta_p(U)=1$ if $p \in U$, $\delta_p(U)=0$ otherwise.

In general, a measure is a nonnegative additive set function. More precisely, it is a nonnegative function (with $+\infty$ an allowed value) defined on a collection of subsets (which are then said to be measurable) of the space in question, such that the union of any countable or finite union of disjoint measurable sets A_i is measurable (this is a requirement of the collection of measurable sets), and its measure is the sum of the measures of the A_i. (One cannot require an analogous property for unions of uncountably many sets because the real line is a disjoint union of points, each of which has zero Lebesgue measure.) One often is interested in finite measures, that is, those for which the measure of the whole space is finite. These can be rescaled to a *probability measure* (the whole space has measure 1). A set of measure zero is referred to as a *null set*, and a property is said to hold *almost everywhere* if it fails for only a null set of points.

If every open set is measurable and every compact set has finite measure, then the measure is said to be a *Borel measure*. This is a useful notion because a Borel measure μ is regular; that is

$$\begin{aligned}\mu(A) &= \inf\{\mu(O) : A \subset O \text{ and } O \text{ is open}\}\\ &= \sup\{\mu(B) : B \subset A \text{ and } B \text{ is compact}\}.\end{aligned}$$

Up to a scale factor, one defines the Lebesgue measure $\mu_n(A)$ of a set $A \subset \mathbb{R}^n$ by considering collections of balls whose union contains A and minimizing the sums of the volumes of these balls:

$$\mu_n(A) = \inf\left\{\sum r_i^n : A \subset \bigcup_i B(x_i, r_i)\right\}.$$

Replacing the exponent n of the radii by an arbitrary exponent α that is not necessarily related to the dimension of the ambient space, one obtains the α-dimensional *Hausdorff measure* μ_α. For example, for a smooth curve c of length $l(c)$ in the plane, we get $\mu_1(c)=l(c)$ and $\mu_2(c)=0$. Indeed, one can characterize the *Hausdorff dimension* of a set S via Hausdorff measures. It is the number α_0 such that $\mu_\alpha(S)=+\infty$ for $\alpha<\alpha_0$ and $\mu_\alpha(S)=0$ for $\alpha>\alpha_0$.

A measure is *invariant* under a map f and the map is said to be *measure-preserving*, if for any measurable set A the set $f^{-1}(A) := \{x\colon f(x) \in A\}$ is measurable and has the same measure as A. This preimage definition turns out to be the proper one for noninvertible maps. For example, circle rotations $z \mapsto z\mathrm{e}^{2\pi\mathrm{i}\theta}$ preserve the Lebesgue measure because they preserve length. The doubling map $z \mapsto z^2$ of the unit circle in the complex plane preserves the Lebesgue measure because the preimage of an arc consists of two arcs of half the length. The density $1/(1+x)$ on $[0, 1]$ defines a measure μ, which is invariant under the Gauss map $G\colon [0,1] \to [0,1]$ defined by $x \mapsto \{1/x\}$ (fractional part): $a=\{1/x\}$ if and only if $1/x = a+n$ for some $n \in \mathbb{N}$, so

$$\begin{aligned}\mu(G^{-1}([a,b])) &= \sum_{n\in\mathbb{N}} \int_{1/b+n}^{1/a+n} \frac{1}{1+x}\,\mathrm{d}x\\ &= \sum_{n\in\mathbb{N}} \log\left(1+\frac{1}{a+n}\right)\end{aligned}$$

$$
\begin{aligned}
&\quad - \log\left(1 + \frac{1}{b+n}\right) \\
&= \sum_{n\in\mathbb{N}} \log(a+n+1) - \log(a+n) \\
&\quad - \log(b+n+1) + \log(b+n) \\
&= \log(b+1) - \log(a+1) \\
&= \int_a^b \frac{1}{1+x}\,\mathrm{d}x = \mu([a,b]).
\end{aligned}
$$

The Dirac measure δ_p concentrated on a fixed point p is always invariant.

The *Birkhoff ergodic theorem* guarantees the existence of time averages of orbits. If φ is an observable (a continuous or merely measurable scalar function on phase space) and μ is an f-invariant measure, then for almost every point the time (or Birkhoff or ergodic) average

$$\lim_{n\to\infty} \sum_{i=0}^{n-1} \varphi(f^i(x))/n$$

exists. A map f is said to be *ergodic* if the phase space is indecomposable in the following sense. If A is an invariant set (i.e., $f^{-1}(A) = A$), then either A or its complement is a null set. This is the case for the Lebesgue measure and the doubling map or rotations by an irrational angle, but not for rotations by a rational angle. For ergodic systems, the time average over almost every orbit equals the space average $\int \varphi\,\mathrm{d}\mu$.

A particularly interesting measure with respect to the study of hyperbolic attractors and strange attractors is the *Sinai–Ruelle–Bowen measure*. By definition it has a positive Lyapunov exponent almost everywhere and absolutely continuous conditionals (marginals) on unstable manifolds. If such a measure is ergodic and has no zero Lyapunov exponents, then it gives a natural or physical (physically observed) measure, which is defined by the following property. While for any invariant measure μ the Birkhoff ergodic theorem implies that almost every point is μ-equidistributed, the physical measure reflects the asymptotic distribution of Lebesgue-almost every point (or at least that of a set of points of positive Lebesgue measure). This means that if one picks a point at random (with respect to Lebesgue measure), its orbit will be equidistributed uniformly with respect to the physical measure. In other words, a physical (or natural) measure represents the density of points obtained from a computed orbit. For example, the Dirac measure concentrated on an attracting fixed point is a natural measure.

BORIS HASSELBLATT

See also **Ergodic theory; Poincaré theorems; Recurrence; Sinai–Ruelle–Bowen measures**

Further Reading

Hasselblatt, B. & Katok, A. (editors). 2002: *Handbook of Dynamical Systems*, vol. 1A. Amsterdam and New York: Elsevier (See also vol. 1B, 2005)

Hasselblatt, B. & Katok, A. 2003: *Dynamics: A First Course*, Cambridge and New York: Cambridge University Press

Katok, A. & Hasselblatt, B. 1995: *Introduction to the Modern Theory of Dynamical Systems*, Cambridge and New York: Cambridge University Press [Measure theory is summarized in an Appendix chapter.]

Petersen, K. 1983: *Ergodic Theory*, Cambridge and New York: Cambridge University Press

Poincaré, H.J. 1890. Sur le probléme des trois corps et les équations de la dynamique. *Acta Mathematica*, 13: 1–270

MECHANICS OF SOLIDS

Two aspects may be discerned in the historical development of the mechanics of solids. The first of these was motivated by technical requirements, and the second was motivated by studies in theoretical physics. Sometimes these two aspects interacted, but each of them has maintained its specificity.

From the fundamental continuum hypothesis, the mechanics of solids admits an introduction of vector and tensor fields describing the strained state of a solid as a result of its deformation: the displacement vector and tensors of stress and strain. Physical features are manifested by the constitutive relations (algebraic, differential, or integro-differential), which determine the coupling between stress and strain tensors, and Newton's equations of motion can be constructed taking into account all relations mentioned above.

There are two sources of nonlinearity in the mechanics of solids. If the constitutive relations are nonlinear, one deals with an intrinsically nonlinear problem. Nonlinearity also arises from purely geometric reasons (nonlinear dependence of the strains on displacements) and is called geometric nonlinearity. Four important approximations illustrating the basic nonlinear effects in the mechanics of solids are as follows.

Prerequisites of approximations:	*Result of approximation:*
(a) Rotational and translational invariance of infinite undeformed solid	Infinite isotropic media
(b) Discrete group of symmetry	Anisotropic crystal
(c) Details of microstructure	Media with weak or strong dispersion
(d) Smallness of one or two dimensions	Bar, plate, shell

Taking account of the shear, nonlinear dynamic equations for an infinite elastic body can be considered as an extension of the hydrodynamics of a compressible non-viscous fluid. Lagrange's equations of motion in the main nonlinear approximation are written as follows (Blend, 1969):

$$\rho\frac{\partial^2 u_j}{\partial t^2} - \frac{\partial}{\partial u_j}\left(\frac{\partial U}{\partial u_{i,j}}\right) = X_{i,j},$$

$$U = \tfrac{1}{2}\lambda I_1^2 + \mu I_2 + T_0 S - \kappa I_1 S + \text{higher-order terms},$$

$$i, j = 1, 2, 3, \qquad (1)$$

where ρ_0 is the initial density of the elastic media, $X_{i,j}$ are components of external loading, u_j are components of the displacement vector, U is a potential, $I_1 = \varepsilon_{ii}$ and $I_2 = \varepsilon_{ij}\varepsilon_{ij}$ are the first and second invariants of the strain tensor, λ and μ are Lameé coefficients characterizing the elastic properties of isotropic media, S is entropy, κ is the coefficient of heat conductivity, and T_0 is the initial temperature. Lamé coefficients can be expressed via the Young's modulus E and Poisson coefficient ν, describing longitudinal elastic resistance and transversal deformation, respectively, for a specimen subjected to uniaxial longitudinal loading.

In this case, geometric nonlinearity is caused by nonlinear dependence of ε_{ij} on derivatives of displacements, and physical nonlinearity is accounted for by high-order terms involving strains ε_{ij}. A significant consequence of nonlinearity may be the possibility of shock waves manifested by discontinuity of the first or second derivatives of displacements (strong or weak discontinuities), respectively.

The simplest and most important problem admitting discontinuous solutions in nonlinear theory is the description of plane adiabatic shock waves propagation. In this case, nonlinear partial differential equations (1) are replaced by nonlinear algebraic relations. These relations are determined by conservation laws for mechanical and thermodynamical quantities as well as by the second law of thermodynamics. The main difference from shock waves in the fluid is manifested in the presence of shear shock waves.

Taking into account microstructure and intermolecular (physical) nonlinearity in the constitutive relations, one obtains in the lowest approximation for plane waves the Korteweg–de Vries (KDV) equation (Askar, 1985)

$$\frac{\partial w}{\partial \tau} + \frac{1}{2}\alpha_1 w\frac{\partial w}{\partial \zeta} + \frac{1}{24}\frac{\partial^3 w}{\partial \zeta^3} = 0,$$

where τ and ζ are dimensionless time and space coordinates, respectively, and $w(\tau, \zeta)$ is a dimensionless deformation of the media. Due to the presence of nonlinear (second) and dispersion (third) terms, this equation has a stable soliton solution (compression wave) as well as multiple soliton solutions.

A second new aspect, important for media with microstructure, is the possibility of a short wavelength continuum approximation. Such an approximation is valid for the envelopes of short waves for which the wavelengths are comparable with the characteristic length of microstructure (Manevitch, 2001). In this case, equations of motion for the plane waves can be reduced to the dimensionless nonlinear Schrödinger (NLS) equation

$$\frac{\partial \varphi_0}{\partial \tau} - \mathrm{i}\alpha|\varphi_0|^2\varphi_0 - \mathrm{i}\frac{\partial^2 \varphi_0}{\partial \zeta^2} = 0\,,$$

where $\varphi_0 = [\exp \mathrm{i}\tau](w - \mathrm{i}u)$, u is the displacement, w is the velocity, α is a nonlinear parameter, and τ and ζ are dimensionless time and space coordinates. An important manifestation of nonlinearity in this case is the existence of envelope solitons

$$\varphi_0(\zeta, \tau) = \sqrt{2S/\alpha}\,\mathrm{e}^{\mathrm{i}(k\zeta - \omega\tau)}\operatorname{sech}[\sqrt{S}(\zeta - \nu\tau)],$$

where $k = \nu/2$ and $\omega = \nu^2/4 - S$.

Dispersion effects can be also caused by the presence of boundaries as in the case of a rod embedded in another elastic external medium. Along with physical and/or geometric nonlinearity, this may lead to formation of solitons similar to solitons in media with microstructure (Samsonov et al., 1999).

Note that the formation of shock waves in three-dimensional elastic media requires much more energy than in the common case of sound waves. Contrary to this, geometric nonlinearity in thin elastic bodies described by simplified one- and two-dimensional models provides easy manifestation of nonlinear effects due to bifurcation of equilibrium states (elastic instability). Elastic instability may lead to localization of buckling and coupling of linear modes—a sign of strong nonlinearity (Thomson & Hupt, 1973). Thin-walled structures also demonstrate numerous other effects typical of common nonlinear systems, including period doubling and internal resonance, transition to spatial, and temporal chaos. Future developments in nonlinear mechanics of solids will focus on the study of these effects as well as the propagation of nonlinear waves in anisotropic and nonhomogeneous media and on mechanics of solids at the mesoscopic level (Alexander, 1998).

LEONID MANEVITCH

See also **Korteweg–de Vries equation; Nonlinear Shrödinger equations; Shock waves; Solitons**

Further Reading

Alexander, S. 1998. Amorphous solids theory, structure, lattice dynamics and elasticity. *Physics Reports*, 296: 68–256

Askar, A. 1985. *Lattice Dynamical Foundations of Continuum Theories*, Singapore: World Scientific

Blend, D.R. 1969. *Nonlinear Dynamic Elasticity*, Waltham, MA: Blaisdell Publishing

Manevitch, L.I. 2001. Solitons in polymer physics. *Polymer Science* C, 4(2): 117–181

Samsonov, A.M. et al. 1999. Strain solitons in solids: physics, numerics and fracture. In *Dynamics of Vibro-Impact Systems*, edited by V.I. Babitsky, Berlin: Springer, pp. 215–220

Thomson, J.M. & Hupt, G.M. 1973. *A General Theory of Elastic Stability*, New York: Wiley

MEL'NIKOV METHOD

In 1963, V.K. Mel'nikov developed an analytical method for the study of time-periodic perturbations of a planar autonomous system. This method permits determination of persistence and stability of subharmonic motions and measurement of infinitesimal separation of stable and unstable manifolds of hyperbolic fixed points. The Mel'nikov method is one of the few analytical tools for establishing conditions for homoclinic chaos, and it has been generalized to many settings, including Hamiltonian systems with higher degrees of freedom and partial differential equations.

The main ideas can be introduced in the context of a periodically forced planar Hamiltonian system:

$$\begin{aligned}\dot{x} &= \frac{\partial H}{\partial y} + \varepsilon f_1(x, y, t),\\ \dot{y} &= \frac{\partial H}{\partial x} + \varepsilon f_2(x, y, t).\end{aligned}$$

$H(x, y)$ is the Hamiltonian function of the unperturbed system ($\varepsilon = 0$), and $\mathbf{f} = (f_1, f_2)^{\mathrm{T}}$ is a time-periodic perturbation vector field. Both H and $\mathbf{f}$ are assumed to be sufficiently smooth functions of their arguments.

Assume that the unperturbed system possesses a homoclinic orbit $q_{\mathrm{h}}(t)$ to a hyperbolic saddle point p_0, as shown in Figure 1. When $\varepsilon \neq 0$, a perturbation argument guarantees persistence of a hyperbolic fixed point p_ε, and local existence of its stable and unstable manifolds, parametrized, respectively, by solutions $q_\varepsilon^{\mathrm{s}}(t) = q_{\mathrm{h}}(t - t_0) + \varepsilon q_1^{\mathrm{s}}(t) + \mathcal{O}(\varepsilon^2)$ and $q_\varepsilon^{\mathrm{u}}(t) = q_{\mathrm{h}}(t - t_0) + \varepsilon q_1^{\mathrm{u}}(t) + \mathcal{O}(\varepsilon^2)$.

As the unperturbed separatrix is expected to split under perturbation, it is convenient to introduce the distance function

$$\begin{aligned}d(t_0) &= (\partial_x H, \partial_y H)^{\mathrm{T}}|_{q_{\mathrm{h}}(0)} \cdot [q_\varepsilon^{\mathrm{u}}(t_0) - q_\varepsilon^{\mathrm{s}}(t_0)]\\ &= \varepsilon[(\partial_x H, \partial_y H)^{\mathrm{T}}|_{q_{\mathrm{h}}(0)} \wedge (q_1^{\mathrm{u}}(t_0) - q_1^{\mathrm{s}}(t_0))]\\ &\quad + \mathcal{O}(\varepsilon^2),\end{aligned}$$

which measures the displacement of the stable and unstable manifolds of p_ε along a direction normal to the unperturbed separatrix (represented by vector **N** in Figure 1). After deriving an evolution equation for the perturbation expansion of the distance function $d(t)$, we arrive at the following expression:

$$d(t_0) = \varepsilon M(t_0) + \mathcal{O}(\varepsilon^2). \qquad (1)$$

The Mel'nikov function $M(t_0)$ is defined by

$$M(t_0) = \int_{-\infty}^{+\infty} \mathbf{grad}\, H(q_{\mathrm{h}}(t - t_0)) \cdot \mathbf{f}(q_{\mathrm{h}}(t - t_0), t)\,\mathrm{d}t, \qquad (2)$$

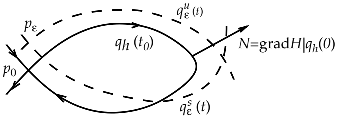

Figure 1. Unperturbed and perturbed separatrices.

where the integrand is the scalar product of vectorfields **grad** H and **f** and is evaluated along the unperturbed homoclinic orbit.

As an example, consider the following planar system of ordinary differential equations with periodic perturbation:

$$\begin{aligned}\dot{x} &= y,\\ \dot{y} &= -x + x^2 + \varepsilon \sin t.\end{aligned}$$

The unperturbed system has Hamiltonian $H(x, y) = y^2/2 + x^2/2 - x^3/3$ and two fixed points: the origin (a center) and the point of coordinates $(1, 0)$ (a saddle point). The unperturbed homoclinic orbit (parametrizing the degenerate stable, unstable manifolds of the hyperbolic fixed point) is given by

$$\begin{aligned}q_{\mathrm{h}}(t - t_0) = \Bigg(&\frac{3}{2}\tanh^2\left(\frac{t - t_0}{2}\right) - \frac{1}{2},\\ &\frac{3}{2}\tanh\left(\frac{t - t_0}{2}\right)\sinh^2\left(\frac{t - t_0}{2}\right)\Bigg)^T,\end{aligned} \qquad (3)$$

and the corresponding Mel'nikov function is computed to be

$$\begin{aligned}M(t_0) &= \frac{3}{2}\int_{-\infty}^{+\infty} \tanh(t/2)\sinh^2(t/2)\sin(t + t_0)\,\mathrm{d}t\\ &= -\frac{3\pi\cos(t_0)}{2\sinh\pi/2\cosh\pi/2}.\end{aligned} \qquad (4)$$

If, as in this example, $M(t_0)$ has an infinite sequence of nondegenerate zeros, an infinite number of transverse homoclinic orbits of the hyperbolic fixed point of the perturbed system is guaranteed to exist for all $\varepsilon \neq 0$ sufficiently small. Thus, the system exhibits homoclinic chaos.

For perturbations containing dissipative or constant forcing terms, the Mel'nikov integral may be nonvanishing. In such cases, the separatrix still splits under perturbation, but no transverse homoclinic points occur.

Generalizations of the Mel'nikov method to higher-dimensional systems have been discussed by various authors (see, for example, Yagasaki (1999), Gruendler (1992), and the monograph by Wiggins (1988, pp. 1–16)).

Autonomous Perturbations in Two Degrees of Freedom

Consider how this method applies to autonomous near-integrable Hamiltonian systems with two degrees of freedom. This case was first treated, and then generalized to integrable equations with $n + 1$ degrees of freedom, by Holmes and Marsden (1982a,b). These authors assume that the unperturbed Hamiltonian has the form $H_0 = F(x, y) + G(I)$, where F is the Hamiltonian of a planar system with a homoclinic orbit to a hyperbolic fixed point, and G is the Hamiltonian of a planar system in action-angle coordinates. By restricting to the energy surface, the original equations are reduced to a single-degree-of-freedom non-autonomous system, to which the usual Mel'nikov technique is applied.

A more direct approach by Robinson (1996) assumes Hamiltonians of the general form

$$H_\varepsilon(\mathbf{x}) = H_0(\mathbf{x}) + \varepsilon H_1(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^4. \tag{5}$$

The unperturbed system is assumed to be completely integrable, with a second constant of motion K, and to possess a hyperbolic periodic orbit $\gamma_{\mathbf{0}}$. The stable and unstable manifolds of γ_0, $\mathcal{W}^{\mathrm{s}}(\gamma_0) = \mathcal{W}^{\mathrm{u}}(\gamma_0)$, are parametrized by a two-dimensional family of homoclinic orbits $q_{\mathrm{h}}(t, \mathbf{x}_0)$, where $\mathbf{x}_0$ is a point on the corresponding level set $H_0(\gamma_0) = h_0$.

Because the perturbation is Hamiltonian, the perturbed hyperbolic periodic orbit γ_ε and its stable, unstable manifolds $\mathcal{W}^{\mathrm{s,u}}(\gamma_\varepsilon)$ continue to lie within the same energy surface $H_\varepsilon^{-1}(h_0)$. Thus **grad** H_0 no longer provides a good measurement of their transversal splitting.

Letting Π_0 be a two-dimensional plane transversal to the unperturbed separatrix at $\mathbf{x_0}$ and denoting by $\zeta^{\mathrm{s,u}}(\mathbf{x}_0, h_0; \varepsilon) = \Pi_0 \cap \mathcal{W}^{\mathrm{s,u}}(\gamma_\varepsilon)$ the corresponding points on the perturbed invariant manifolds, one computes the infinitesimal displacement in terms of the second integral K:

$$\begin{aligned} M(\mathbf{x_0}, h_0) = {} & \frac{\partial}{\partial\varepsilon}[K \circ \zeta^u(\mathbf{x}_0, h_0; \varepsilon) \\ & - K \circ \zeta^s(\mathbf{x}_0, h_0; \varepsilon)]|_{\varepsilon=0}. \end{aligned} \tag{6}$$

The Mel'nikov measurement can be reduced (Robinson, 1996) to the following conditionally convergent integral:

$$\begin{aligned} & M(\mathbf{x}_0, h_0) \\ & = \lim_{j\to\infty} \int_{-T_j^*}^{T_j} \mathbf{grad}\, F \cdot \mathcal{J}\mathbf{grad}\, H_1|_{q_h(t,\mathbf{x}_0)}\, \mathrm{d}t, \end{aligned} \tag{7}$$

where $\mathcal{J}\mathbf{grad}H_1$ is the Hamiltonian vector field of the perturbation, and $-T_j^*$, T_j are chosen so that $q_h(-T_j^*, \mathbf{x}_0)$ and $q_h(T_j, \mathbf{x}_0)$ converge to the same point of the periodic orbit γ_0.

An Elastic Pendulum

Consider the application of these ideas to the following four-dimensional dynamical system modelling an elastic pendulum (Holmes and Marsden, 1982a):

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= \sin x + \varepsilon\left(\sqrt{\frac{2I}{\omega}}\sin\theta - x\right), \\ \dot{I} &= \varepsilon\sqrt{\frac{2I}{\omega}}\left(\sqrt{\frac{2I}{\omega}}\sin\theta - x\right)\cos\theta, \\ \dot{\theta} &= \omega - \varepsilon\frac{1}{\sqrt{2I\omega}}\left(\sqrt{\frac{2I}{\omega}}\sin\theta - x\right)\sin\theta. \end{aligned}$$

The unperturbed equations are completely integrable with constants of motion $H_0(x, y, I, \theta) = \frac{1}{2}y^2 - \cos x + \omega I$ and $K(x, y, I, \theta) = I$. The energy surface $H_0(\mathbf{x}) = 1 + \omega I = h_0$ contains a hyperbolic periodic orbit $\gamma_0 = (\pi, 0, I, \omega t)^{\mathrm{T}}$ and its two-dimensional stable and unstable manifolds, parametrized by the pair of homoclinic orbits

$$\begin{aligned} q_h^{\pm}(t, \theta_0) = {} & (\pm 2\tan^{-1}(\sinh t), \\ & \pm 2\,\mathrm{sech}\, t, \frac{h_0 - 1}{\omega}, \omega t + \theta_0)^{\mathrm{T}}. \end{aligned}$$

As **grad** $K = (0, 0, 1, 0)^{\mathrm{T}}$, the quantities $M^{\pm}(\theta_0, h_0)$ are immediately computed as the following absolutely convergent integrals ($h_0 > 1$):

$$\begin{aligned} & \frac{\sqrt{2(h_0 - 1)}}{\omega}\int_{-\infty}^{+\infty}\left[\frac{\sqrt{2(h_0 - 1)}}{\omega}\sin(\omega t + \theta_0)\right. \\ & \left.\mp 2\tan^{-1}(\sinh t)\right]\cos(\omega t + \theta_0)\,\mathrm{d}t. \end{aligned} \tag{8}$$

The first term is odd and thus vanishes. Using integration by parts to simplify the second term yields the expression

$$\begin{aligned} M^{\pm}(\theta_0, h_0) = {} & \pm\frac{\sqrt{2(h_0 - 1)}}{\omega^2} \\ & \times \int_{-\infty}^{+\infty} \mathrm{sech}(t)\sin(\omega t + \theta_0)\,\mathrm{d}t, \end{aligned}$$

which can be evaluated using the method of residues as

$$M^{\pm}(\theta_0, h_0) = \pm 2\pi\frac{\sqrt{2(h_0 - 1)}}{\omega}\mathrm{sech}\left(\frac{\pi\omega}{2}\right)\sin(\theta_0).$$

An infinite sequence of simple zeros of $M^{\pm}(\theta_0, h_0)$ guarantees the existence of transversal homoclinic orbits on each energy surface $h_0 > 1$, for $\varepsilon \neq 0$ sufficiently small.

Annalisa M. Calini

See also **Chaotic dynamics; Hamiltonian systems; Horseshoes and hyperbolicity in dynamical systems; Phase space**

Further Reading

Gruendler, J. 1992. Homoclinic solutions for autonomous dynamical systems in arbitrary dimension. *SIAM Journal of Mathematical Analysis*, 23: 702–721

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Berlin and New York: Springer

Holmes, P.J. & Marsden, J.E. 1982a. Mel'nikov's method and Arnol'd diffusion for perturbations of integrable Hamiltonian systems. *Journal of Mathematical Physics*, 23(4): 669–675

Holmes, P.J. & Marsden, J.E. 1982b. Horseshoes in perturbations of Hamiltonian systems with two degrees of freedom. *Communications in Mathematical Physics*, 82: 523–544

Mel'nikov, V.K. 1963. On the stability of the center for time periodic perturbations. *Transactions of Moscow Mathematical Society*, 12: 1–57

Robinson, C. 1996. Mel'nikov method for autonomous Hamiltonians. *Contemporary Mathematics*, 198: 45–53

Wiggins, S. 1988. *Global Bifurcations and Chaos: Analytical Methods*, Berlin and New York: Springer

Yagasaki, K. 1999. The method of Mel'nikov for perturbations of multi-degree-of-freedom Hamiltonian systems. *Nonlinearity*, 12(4): 799–822

METASTABILITY

See **Stability**

METEOROLOGY

See **Atmospheric and ocean sciences**

MISSING MASS (FOR KDV SOLITON)

See **Korteweg–de Vries equation**

MIURA TRANSFORMATION

See **Korteweg–de Vries equation**

MIXING

The word *mixing* is used both as a general term defining the operation of putting two or more substances together in order to achieve uniformity and as a mathematical term defining the property of random processes. Mixing as an operation is widespread as both a natural phenomenon and an industrial process. Putting milk into coffee, preparing cement, and pushing a car accelerator pedal involve mixing liquids, granular materials, and gases. On a molecular scale, it is diffusion that provides mixing. When diffusion is caused solely by the gradient of concentration $\theta(\boldsymbol{r}, t)$, it is described by the second-order partial differential equation

$$\frac{\partial \theta}{\partial t} = \operatorname{div}(\kappa \nabla \theta). \tag{1}$$

When the diffusivity κ is constant, (1) is a linear parabolic equation which can be solved by using the Green function: $\theta(\boldsymbol{r}, t) = (4\pi\kappa t)^{-d/2} \times \int \exp[-(\boldsymbol{r} - \boldsymbol{r}')^2/4\kappa t]\theta(\boldsymbol{r}', 0)\,\mathrm{d}\boldsymbol{r}'$.

The diffusivity of gases in gases is of order 10^{-1} cm^{2}s^{-1}, so it would take many hours for an odor to diffuse across the dinner table. Similarly, to diffuse salt to a depth of 1 km in the ocean molecular diffusion would take 10^7 years. It is the motion of fluids that provides large-scale mixing in most cases. In a moving fluid, θ satisfies the advection-diffusion equation:

$$\frac{\partial \theta}{\partial t} + (\boldsymbol{v} \cdot \nabla)\theta = \operatorname{div}(\kappa \nabla \theta). \tag{2}$$

If the velocity gradient is λ then one can define a diffusion scale $r_{\mathrm{d}} = \sqrt{\kappa/\lambda}$ comparing advective and diffusive terms in (2). Fluid motion and molecular diffusion provide for mixing at the scales respectively larger and smaller than r_d. Inhomogeneous flow brings into contact fluid parcels with different values of θ thus producing large gradients that are then eliminated by molecular diffusivity. How fast mixing proceeds and how concentration variance decays in time depends on how inhomogeneous the flow is.

When a velocity field fluctuates, the simplest quantity (and often most important) is the concentration averaged over velocity, $\langle \theta(\boldsymbol{r}, t)\rangle$. The behavior of this quantity is determined by the properties of Lagrangian velocity $\boldsymbol{V}(t) = \boldsymbol{v}[\boldsymbol{q}(t), t]$, which is taken on the trajectory that satisfies $\mathrm{d}\boldsymbol{q}/\mathrm{d}t = \boldsymbol{v}[\boldsymbol{q}(t), t]$. For times longer than the Lagrangian correlation time, $\langle \theta(\boldsymbol{r}, t)\rangle$ also satisfies the diffusion equation

$$\Big[\partial_t - (\kappa\delta_{ij} + D_{ij})\nabla_i\nabla_j\Big]\langle \theta(\boldsymbol{r}, t)\rangle = 0,$$

with so-called eddy diffusivity

$$D_{ij} = \frac{1}{2}\int_0^\infty \langle V_i(0)V_j(s) + V_j(0)V_i(s)\rangle\,\mathrm{d}s.$$

If we release a single spot of, say, a pollutant, then its average position is given by $\langle \theta(\boldsymbol{r}, t)\rangle$. On the other hand, the evolution of the spot itself depends on the spatial properties of the velocity field. In considering hydrodynamic mixing at a given scale, one usually distinguishes between two qualitatively different classes of velocity fields: spatially smooth and nonsmooth. Velocity can be considered spatially smooth on a given scale if the velocity gradient does not change much across the scale. Comparing the inertial term $(\boldsymbol{v} \cdot \nabla)\boldsymbol{v}$ with the viscous term $\nu\Delta\boldsymbol{v}$ in the Navier–Stokes equation for fluid motion, one defines the viscous scale η similarly to r_{d}. Turbulent flows are smooth at scales smaller than η (viscous interval) and nonsmooth at larger scales (intertial interval). Fluid particles separate exponentially with time in smooth flows and according to power laws in nonsmooth flows.

Despite the fact that the fluid viscosity ν (momentum diffusivity) is caused by the same molecular motion as κ (diffusivity of a substance), their ratio varies widely depending on the type of material. That ratio is called the Schmidt number, or Prandtl number when θ is temperature. The Schmidt number is very high for viscous liquids and also for colloids and aerosols, since the diffusivity of, say, micron-size particles (e.g., cream globules in milk and smoke in the air) is six to seven orders of magnitude less than the viscosity of the ambient fluid. In those cases, $r_{\rm d} \ll \eta$. At scales less than η, the flow is spatially smooth, and the velocity difference between two fluid particles can be presented as $\boldsymbol{v}(\boldsymbol{q}_1, t) - \boldsymbol{v}(\boldsymbol{q}_2, t) = \hat{\sigma}(t)\boldsymbol{R}(t)$ so that the separation $\boldsymbol{R} = \boldsymbol{q}_1 - \boldsymbol{q}_2$ obeys the ordinary differential equation

$$\dot{\boldsymbol{R}}(t) = \sigma(t)\, \boldsymbol{R}(t),$$

leading to the linear propagation $\boldsymbol{R}(t) = W(t)\, \boldsymbol{R}(0)$. The main statistical properties of $\boldsymbol{R}(t)$ can be established at the limit when t exceeds the correlation time of the strain matrix $\hat{\sigma}(t)$. The basic idea (going back to the works of Lyapunov, Furstenberg, Oseledec, and many others and developed in the theory of dynamical chaos) is to consider the positive symmetric matrix $W^{\rm T}W$ which determines R. The main result states that in almost every realization of $\hat{\sigma}(t)$, the matrix $t^{-1} \ln W^{\rm T}W$ stabilizes as $t \to \infty$. In particular, its eigenvectors tend to d fixed orthonormal eigenvectors $\boldsymbol{f}_i$. To understand that intuitively, consider some fluid volume, say, a sphere, which evolves into an elongated ellipsoid at later times. As time increases, the ellipsoid is more and more elongated, and it is less and less likely that the hierarchy of the ellipsoid axes will change. The limiting eigenvalues

$$\lambda_i = \lim_{t\to\infty} t^{-1} \ln |W\boldsymbol{f}_i| \qquad (3)$$

define the so-called Lyapunov exponents. The major property of the Lyapunov exponents is that they do not depend on the starting point if the velocity field is ergodic.

Consider now a pollutant spot with size l released within a spatially smooth velocity and assume that the Peclet number $l/r_{\rm d}$ is large. The above consideration shows, in particular, that the spot will acquire an ellipsoid form. The direction that corresponds to the lowest Lyapunov exponent (necessarily negative in an incompressible flow) contracts until it reaches $r_{\rm d}$, and further contraction is stopped by molecular diffusion. Because the exponentially growing directions continue to expand, the volume grows exponentially and the value of θ inside the spot decays exponentially in time. For an arbitrary large-scale initial distribution of θ, the concentration variance decays exponentially in a spatially smooth flow because this is how fast velocity inhomogeneity contracts θ "feeding" molecular diffusion which eventually decreases the variance. Even though it is diffusion that diminishes θ and the rate of decay is independent of κ, it is usually of order of the typical velocity gradient.

If the Schmidt number is small while the Reynolds number of the flow is large then the velocity field at scales larger than $r_{\rm d}$ cannot be considered spatially smooth. That means that the velocity difference $\delta v(r)$, measured between two points distance r apart, scales as r^a with $a < 1$ (of course, δv is random and the statement pertains to the moments). For example, for the energy cascade in incompressible fluids, a is close to $1/3$. The equation $\dot{\boldsymbol{R}} = \delta \boldsymbol{v}(\boldsymbol{R}) \propto R^a$ suggests that interparticle distance grows by a power law: $R(t) \propto t^{1/(1-a)}$. The volume of any spot also grows so that scalar variance decays by a power law: $\langle \theta^2 \rangle \propto t^{d/(1-a)}$. Such estimates are supported by a rigorous theory only for a velocity field short-correlated in time. In this case, one can also show that the probability distribution $P(\theta, t)$ takes the self-similar form $t^{d/2(1-a)}Q(t^{d/2(1-a)}\theta)$ which is likely to be the case for a general scale-invariant velocity. On the contrary, $P(\theta, t)$ does not change in a self-similar way in a spatially smooth flow.

In finite vessels, the long-time properties of fluid mixing are usually determined by slowest parts, namely, the walls, where the velocity gradient may become zero, and corners with recirculating eddies.

In multiphase flows, not only mixing but also segregation can occur. The physical reason for that is a centrifugal force: when fluid streamlines are curved, heavier particles move out while lighter particles move in. It is a matter of everyday experience that air bubbles are trapped inside the sink vortex while heavy particles gather outside the vortices (which is used, in particular, for flow visualization).

Granular mixing is strikingly different from fluid mixing. In a granular flow, collisions of grains are inelastic and friction between grains makes it possible for static configurations (such as arches) to support a load and distribute stresses. As a result, granular motion has nonlocal properties, and no effective hydrodynamic description based on average over local kinetics (such as Equation (2)) is available. When a container partially filled with grains is vertically shaken with accelerations larger than the gravitational acceleration, convective rolls are observed with grains rising at the center and falling along the walls. Contrary to fluid convection, however, the grains move faster and mix better near the walls. Granular flows can also demonstrate segregation. The most celebrated example is the so-called Brazil nut effect whereby large particles (Brazil nuts) rise to the top of shaken container of mixed nuts.

The use of the term *mixing* in mathematics is based on the notion (introduced by Josiah Gibbs) that evolution is mixing when it leads asymptotically in time to some equilibrium invariant measure. Formally, one defines the evolution operator U_t acting on some phase space A and denotes the measure of any subset B as $P(B)$. The evolution is mixing if for any $B, C \in A$ one has $\lim_{t\to\infty} P(A \bigcap U_t B) = P(A)P(B)$. One can also define weak mixing property where

$\lim_{t\to\infty} t^{-1}\int P(A\bigcap U_s B)\,\mathrm{d}s = P(A)P(B)$. Mixing of a random process or dynamical system means ergodicity, that is equality between temporal and phase space average.

Gregory Falkovich

See also **Chaotic advection; Diffusion; Entropy; Granular materials; Intermittency; Kolmogorov cascade; Lyapunov exponents; Turbulence**

Further Reading

Chate, H., Villermaux, E. & Chomaz, J.-M. (editors). 1999. *Mixing: Chaos and Turbulence*, New York: Kluwer

Falkovich, G., Gawȩdzki, K. & Vergassola, M. 2001. Particles and fields in fluid turbulence. *Reviews of Modern Physics*, 73: 913–975

MÖBIUS STRIP

See **Topology**

MODE LOCKING

See **Coupled oscillators**

MODE MIXING AND COUPLING

See **Coupled systems of partial differential equations**

MODIFIED KDV EQUATION

See **Korteweg–de Vries equation**

MODULATED WAVES

The term *modulated waves* implies waves in which some parameters vary slowly in time or in space–time compared with the main (carrier) variables. The term *modulation* comes from acoustics (music) and radio engineering, where signals of the type

$$A(t)\cos[\omega_0 t + \varphi(t)] \qquad (1)$$

have long been used for transmitting and processing information. Here, the carrier frequency ω_0 is constant with respect to time t, whereas the amplitude A and phase φ are slowly varying functions of time as compared with variations of the carrier oscillations described by the function $\cos(\omega_0 t)$. If only one of the parameters A and φ is variable, the terms *amplitude modulation* or *phase modulation* are used. A more general form is $A(t)\cos[\Phi(t)]$, where Φ is the full phase, so that the frequency $\omega(t) = \mathrm{d}\Phi/\mathrm{d}t$ is slowly varying. Although this case is called frequency modulation, the terms *phase* and *frequency modulation* describe essentially the same process.

For spatially varying signals, a natural generalization of that is a modulated quasiharmonic wave

$$A(\mathbf{r},t)\cos[\omega_0 t + \varphi(\mathbf{r},t)] \text{ or } A(\mathbf{r},t)\cos[\Phi(\mathbf{r},t)], \qquad (2)$$

where A, φ, and $\partial\Phi/\partial t$ are slowly varying functions of time. In the spectral representation, the first of these expressions yields a narrow frequency spectrum as compared with the carrier frequency ω_0, whereas the spectrum of the second expression can be arbitrarily wide. The cause of wave modulation can be different: modulation in initial or boundary (signaling) conditions, slow variation of the medium parameters in time and space, attenuation, and amplification.

A typical case is a one-dimensional traveling wave when functions (2) depend on time t and one spatial coordinate x, for example,

$$A(x,t)\cos[\omega_0 t - k_0 x + \varphi(x,t)], \qquad (3)$$

where k_0 is the wave number. At constant A and φ, this wave is a sinusoid propagating with phase velocity $c_{\mathrm{ph}} = \omega_0/k_0$. If the envelope A and phase φ (sometimes called phase envelope, implying all slowly varying parameters of a modulated wave propagate as envelope waves) are slowly varying in space time, and the medium is linear, then up to some distance, they propagate as a wave, with the group velocity, $c_{\mathrm{gr}} = \mathrm{d}\omega_0/\mathrm{d}k_0$, where the dependence between ω_0 and k_0 (dispersion equation) is determined by the properties of the medium. Phase and group velocities are equal only in a nondispersive media, otherwise they are different. At even larger times in a dispersive medium, the wave envelopes are deformed. In particular, a finite-length impulse broadens and turns into a frequency-modulated wave in which each group propagates at its local group velocity depending on its frequency. Such dynamics can be represented by trajectories of groups on the (x, t) plane which are somewhat analogous to spatial rays in geometrical optics or geometrical acoustics and are called space-time rays.

In nonlinear dispersive media, the propagation of envelopes of a modulated dispersive wave is determined not only by its frequency but also by its amplitude (intensity). The interplay of these two factors—dispersion and nonlinearity—results in a variety of scenarios (Ostrovsky & Potapov, 1999; Scott, 1999; Whitham, 1974).

Under definite conditions, first, a non-modulated quasiharmonic wave can be unstable with respect to small modulating perturbations of its amplitude and phase (modulational or Benjamin–Feir instability). In other cases, modulation does not grow, but the envelope distorts with formation of steep (at a modulation scale) fronts (self-steepening). Finally, a steady propagation of envelopes is possible in the form of envelope solitons and envelope shocks.

The notion of modulation can be extended to significantly nonharmonic processes, such as cnoidal

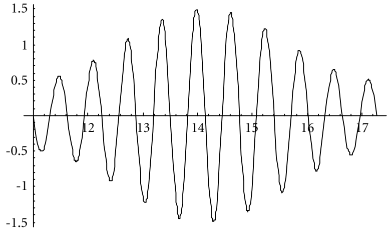

Figure 1. An example of a quasi-harmonic modulated wave.

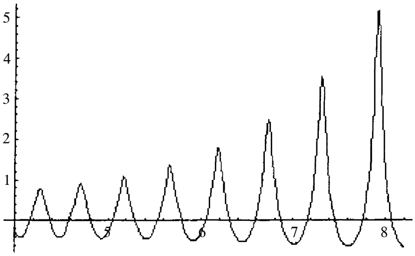

Figure 2. An example of a nonsinusoidal modulated wave.

waves which are typical elliptic function solutions of nonlinear wave equations. Also, in these solutions, slow modulation of their parameters (amplitude and period) in space and time can occur. Let the nonmodulated wave have the form $F(\omega t - kx, A_i)$, where F is a periodic function, the A_i are parameters defining the solution (such as the wave amplitude), and ω and k are effective frequency and wave number defined via the wave period T and its wavelength, Λ, as $\omega = 2\pi/T$ and $k = 2\pi/\Lambda$. Then at slow modulation of these parameters, the wave has the form

$$F[\theta(x,t), A_i(x,t)], \tag{4}$$

where F is a periodic function, A_i are slowly varying functions of x and t, and the phase θ is defined in such a way that $\omega = \partial\theta/\partial t$ and $k = -\partial\theta/\partial x$ are also slowly varying functions. Due to these slow variations, the wave profile can change continuously from an almost sinusoidal wave up to a train of pulses similar to solitary waves or solitons, as the modulus of the elliptic function increases.

Mathematical descriptions of modulated waves are based on asymptotic perturbation theory (Ostrovsky & Gorshkov, 2000) or on more heuristic descriptions. A rather general approach (Whitham's method) starts with a Lagrangian description of a wave field and employs the corresponding variational principle (Whitham, 1974; Ostrovsky & Potapov, 1999; Scott, 1999). For modulated quasi-periodic waves, substituting an expression of type (4) into the Lagrangian (L) and averaging over θ, one obtains an averaged Lagrangian $\mathcal{L}$ depending only on slowly varying parameters, A_i, ω, and k. Considering θ and A_i as new canonical variables, one then obtains from $\mathcal{L}$ variational equations describing the relation among A_i, ω, and k, and their variations in space and time.

This and other averaging methods can work for both nonharmonic nonlinear waves and for weakly nonlinear (or just linear) waves (2) or (3). For narrow-spectrum waves such as (3), a more detailed description can be developed that goes beyond the framework of space-time geometrical optics and takes into account time analogs of wave diffraction. In particular, the nonlinear Shrödinger equation has been obtained to describe such processes. This approach is useful, for example, for the study of envelope solitons having applications in nonlinear optics and for deep water waves.

Moreover, the evolution of a single soliton with slowly varying parameters (due to small losses or smooth inhomogeneities) can be represented in the form of Equation (4), where F is a localized function rather than a periodic one (Ostrovsky & Gorshkov, 2000; Ostrovsky & Potapov, 1999). This class of processes can also be referred to as modulation, and perturbation methods for slowly varying solitary waves have been constructed. Thus, such processes as damping, propagation in inhomogeneous media, and solitary-wave interactions have been analyzed. In the latter case, solitary waves can interact as particles (hence the term *soliton*) with different types of interactions, including repulsion when solitons retain their parameters after interaction and attraction when they may form an oscillating bound state or breather.

Finally, the term *modulation* can be extended to a wave that does not have a prescribed shape but a continuously but slowly distorting profile. A typical example is weakly nonlinear waves in nondispersive media—such as acoustic waves in fluids—where a cumulative steepening of a wave profile can occur up to the formation of weakly nonlinear shock waves. Provided the distortion is small at a wavelength scale, one can also refer to wave profile modulation.

LEV OSTROVSKY

See also **Averaging methods; Collective coordinates; Nonlinear acoustics; Nonlinear optics; Wave stability and instability**

Further Reading

Ostrovsky, L.A. & Gorshkov, K.A. 2000. Perturbation theories for nonlinear waves. In *Nonlinear Science at the Dawn of the 21st Century*, edited by P.L. Christiansen, M.P. Søerensen, & A.C. Scott, New York: Springer, pp. 47–65

Ostrovsky, L.A. & Potapov, A.I. 1999. *Modulated Waves: Theory and Applications*, Baltimore: John Hopkins University Press

Scott, A.C. 1999. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, Oxford and New York: Oxford University Press

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

MODULATIONAL INSTABILITY

See **Wave stability and instability**

MOLECULAR CRYSTALS

See **Local modes in molecular crystals**

MOLECULAR DYNAMICS

Molecular dynamics (MD) refers to a family of computational methods aimed at simulating macroscopic behavior (thermodynamic or hydrodynamic) through the numerical integration of the classical equations of motion of a microscopic many-body system. Macroscopic properties are expressed as functions of particle coordinates and/or momenta, which are computed along a phase space trajectory generated by classical dynamics. At the core of the method is the assumption that over long times the system will reach an equilibrium state, or alternatively a steady state when simulating hydrodynamic conditions, in which temporal averages can be identified with statistical ensemble averages. This requires that mixing and Lyapunov instability be present in the phase space trajectories.

The field of MD simulations was born in the 1950s out of computational studies of the approach to equilibrium in a hard sphere system by B.J. Alder and T.E. Wainwright in 1957 (see Ciccotti et al. (1987) for a collection of seminal works illustrating the development and application of MD simulations). The first MD simulation employing continuous potentials was used to study radiation damage in a two-dimensional solid by Gibson et al. (1960). The first application of the methodology as it is used today, to an atomic fluid modeled by continuous potentials in three dimensions, was a study of structural and dynamical aspects of liquid argon by A. Rahman in 1964. Subsequently, MD simulations have been applied to the study of complex systems such as biomacromolecules and self-assembled systems, and to the elucidation of complex spatiotemporal phenomena in more simple systems. When performed under conditions corresponding to laboratory scenarios, molecular dynamics simulations can provide a detailed view of structure and dynamics at the atomic and mesoscopic levels that is not presently accessible by experimental measurements. The simulations can also be set to perform "computer experiments" that could not be carried out in the laboratory, either because they do not represent natural behavior or because the necessary controls cannot be achieved. Lastly, MD simulations can be implemented as simple test systems for condensed matter theory. In order to realize this wide spectrum of applications, several issues concerning system modeling, dynamical generation of statistical ensembles, and efficient numerical integrators must be considered.

To set the stage, consider the evolution of an isolated system of N point particles. In three dimensions, the canonical description of such system is given by a set of $6N$ first-order differential equations for the particles' positions $\boldsymbol{r}(t) = \{\boldsymbol{r}_1 \ldots \boldsymbol{r}_N\}$ and momenta $\boldsymbol{p}(t) = \{\boldsymbol{p}_1 \ldots \boldsymbol{p}_N\}$. The dynamics of the ith particle can be written in Hamiltonian form as

$$\begin{aligned} \dot{\boldsymbol{r}}_i &= \frac{\partial \mathcal{H}}{\partial \boldsymbol{p}_i} = \frac{\boldsymbol{p}_i}{m}, \\ \dot{\boldsymbol{p}}_i &= -\frac{\partial \mathcal{H}}{\partial \boldsymbol{r}_i} = -\frac{\partial V(\boldsymbol{r})}{\partial \boldsymbol{r}_i} \end{aligned} \qquad (1)$$

with the Hamiltonian given by

$$\mathcal{H}(\boldsymbol{r}, \boldsymbol{p}) = \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m_i} + V(\boldsymbol{r}), \qquad (2)$$

where $V(\boldsymbol{r})$ is the potential. The phase space trajectory generated is constrained to a hypershell parametrized by the conserved total energy of the system $\mathcal{H}(\boldsymbol{r}, \boldsymbol{p}) = E$. Microcanonical ensemble averages are constructed by means of the corresponding probability density function. For any observable A, either explicitly time-dependent or independent, the microcanonical ensemble average is given by

$$\langle A \rangle = \frac{\int \mathrm{d}\mu\, A(\boldsymbol{r}(t), \boldsymbol{p}(t))\, \delta(\mathcal{H} - E)}{\int \mathrm{d}\mu\, \delta(\mathcal{H} - E)}, \qquad (3)$$

where $\mathrm{d}\mu$ is the corresponding invariant measure, in this case $\mathrm{d}\mu = \mathrm{d}\boldsymbol{p}\, \mathrm{d}\boldsymbol{r}$. The chaotic nature of the phase space trajectory, in the sense indicated previously, allows the identification of (3) with the temporal average over an equilibrated simulation of time-length T:

$$\bar{A} = \frac{1}{T} \int \mathrm{d}t A(\boldsymbol{r}(t), \boldsymbol{p}(t)) = \langle A \rangle. \qquad (4)$$

MD simulations use information (positions, velocities or momenta, and forces) at a given instant in time, t, to predict the positions and momenta at a later time, $t + \Delta t$, where Δt is the time step, usually taken to be constant throughout the simulation. Numerical solutions to the equations of motion are thus obtained by iteration of this elementary step. The most popular algorithms for propagating the equation of motion, or "integrators," are based on Taylor expansions of the positions (see Allen & Tildesley (1989) for a survey). For example, the Verlet algorithm uses the sum of third order expansions forward and backward in time of the particle positions, $\boldsymbol{r}_i$,

$$\begin{aligned}\boldsymbol{r}_i(t+\Delta t) &= \boldsymbol{r}_i(t) + \Delta t\,\boldsymbol{v}_i(t) + \frac{\Delta t^2}{2m_i}\boldsymbol{F}_i(t) \\ &\quad + \frac{\Delta t^3}{6}\boldsymbol{b}_i(t) + \mathcal{O}(\Delta t^4), \qquad (5)\\ \boldsymbol{r}_i(t-\Delta t) &= \boldsymbol{r}_i(t) - \Delta t\,\boldsymbol{v}_i(t) + \frac{\Delta t^2}{2m_i}\boldsymbol{F}_i(t) \\ &\quad - \frac{\Delta t^3}{6}\boldsymbol{b}_i(t) + \mathcal{O}(\Delta t^4), \qquad (6)\end{aligned}$$

where $\boldsymbol{v}_i$ is the velocity, $\boldsymbol{F}_i$ is the force, and $\boldsymbol{b}_i$ is the third derivative with respect to time, to obtain an equation for the predicted position that is accurate to third order in time (because of the cancellation of the terms containing odd powers of time):

$$\begin{aligned}\boldsymbol{r}_i(t+\Delta t) &= 2\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t-\Delta t) \\ &\quad + \frac{\Delta t^2}{m_i}\boldsymbol{F}_i(t) + \mathcal{O}(\Delta t^4). \qquad (7)\end{aligned}$$

An equation for the velocities is obtained by subtracting the two expansions,

$$\boldsymbol{v}_i(t) = \frac{\boldsymbol{r}_i(t+\Delta t) - \boldsymbol{r}_i(t-\Delta t)}{2\Delta t} + \mathcal{O}(\Delta t^3). \qquad (8)$$

Note that, according to Equation (7), the velocities are not necessary for prediction of the positions. Because the velocities are determined at time t while the positions are determined at time $t + \Delta t$, calculation of velocity-dependent quantities (e.g., the kinetic energy) coincident with the predicted positions is awkward. The Verlet algorithm can be modified to synchronize the prediction of the positions and velocities. One popular modification, the "velocity Verlet" algorithm, is described below.

A general approach to the development of MD integrators has been formulated based on the Liouville operator formalism of Hamiltonian mechanics, in which Hamilton's equations of motion (Equations (1)) are recast as

$$\dot{\Gamma} = \mathrm{i}L\Gamma. \qquad (9)$$

Here, Γ is a $6N$-dimensional phase vector of the N particle coordinates and momenta (or velocities), and $\mathrm{i}L$ is the Liouville operator

$$\mathrm{i}L = \dot{\Gamma}\cdot\nabla_\Gamma = \sum_{i=1}^{N}\boldsymbol{v}_i\cdot\nabla_{\boldsymbol{r}_i} + \sum_{i=1}^{N}\left[\frac{\boldsymbol{F}_i(\boldsymbol{r})}{m_i}\right]\cdot\nabla_{\boldsymbol{v}_i}. \qquad (10)$$

The formal solution to Equation (10) is

$$\Gamma(t) = \exp(\mathrm{i}Lt)\Gamma(0). \qquad (11)$$

In practice, this is not useful because the action of the propagator, $\exp(\mathrm{i}Lt)$, on the phase vector cannot be determined analytically. However, it is possible to derive accurate short-time approximations to the propagator, whose action can be evaluated analytically, by factorizing the propagator. To this end, Equation (11) is rewritten so that the trajectory from 0 to t is generated in a sequence of P discrete time steps, $\Delta t = t/P$:

$$\Gamma(t) = \prod_{j=1}^{P}\left(\prod_{s}^{n}\exp(\mathrm{i}L_s\Delta t)\right)\Gamma(0). \qquad (12)$$

Then the Trotter factorization may be used to decompose the propagator (Tuckerman & Martyna, 2000):

$$\begin{aligned}\exp(\mathrm{i}L\Delta t) &= \exp\left(\mathrm{i}L_1\frac{\Delta t}{2}\right)\exp(\mathrm{i}L_2\Delta t)\exp\left(\mathrm{i}L_1\frac{\Delta t}{2}\right) \\ &\quad + \mathcal{O}(\Delta t^3) \qquad (13)\end{aligned}$$

giving an approximate, short-time evolution operator that is time-reversible and has the same accuracy as typical MD integrators. The operators $\mathrm{i}L_1$ and $\mathrm{i}L_2$ are chosen so that their action on the phase vector can be determined analytically, with the restriction that $\mathrm{i}L_1 + \mathrm{i}L_2 = \mathrm{i}L$. For example, applying (13) to the phase vector, $\Gamma = (\boldsymbol{r}, \boldsymbol{v})$ (written here and in what follows in terms of velocities), with

$$\begin{aligned}\mathrm{i}L_1 &= \sum_{i=1}^{N}\left[\frac{\boldsymbol{F}_i(\boldsymbol{r})}{m_i}\right]\cdot\nabla_{\boldsymbol{v}_i}, \qquad (14)\\ \mathrm{i}L_2 &= \sum_{i=1}^{N}\boldsymbol{v}_i\cdot\nabla_{\boldsymbol{r}_i} \qquad (15)\end{aligned}$$

gives the velocity Verlet integrator:

$$\begin{aligned}\boldsymbol{r}_i(t+\Delta t) &= \boldsymbol{r}_i(t) + \Delta t\,\boldsymbol{v}_i(t) + \frac{\Delta t^2}{2m_i}\boldsymbol{F}_i(\boldsymbol{r}(t)), \qquad (16)\\ \boldsymbol{v}_i(t+\Delta t) &= \boldsymbol{v}_i(t) + \frac{\Delta t}{2m_i}[\boldsymbol{F}_i(\boldsymbol{r}(t)) + \boldsymbol{F}_i(\boldsymbol{r}(t+\Delta t))]. \qquad (17)\end{aligned}$$

In practice, computing meaningful time averages according to (4) requires sufficient sampling of an already equilibrated trajectory. This determines the total simulation time T. Once a model system and a dynamical system (i.e., an ensemble average) have been chosen, computational efficiency will rely on the choice of the numerical integrator. The total real computational time will depend on how many steps are required to reach the total time-length T. The choice of the time step size depends on the rate of accumulation of numerical errors as evaluated by the tolerance ΔE for the conserved quantity E. In molecular systems, the time step is determined by the highest frequency vibrational mode in the system. A rule of the thumb, is that, for numerical stability, one vibrational period should be covered by roughly 20 time steps. A popular approach to increasing the time step in MD simulations of molecular systems is to use holonomic constraints to freeze the motion of the highest frequency bonds, or all of the bonds altogether. This is considered acceptable practice for most applications because the bond stretching motion is effectively decoupled from the other degrees of freedom that are typically of greater interest. To solve the equations of motion in the presence of constraints, the Lagrangian formalism of classical mechanics is used, and the constraint forces are expressed in terms of undetermined multipliers. There are a variety of algorithms for solving the resulting constraint equations and integrating the equations of motion (see Allen & Tildesley, 1989). The details depend on the choice of the integrator. The most popular are iterative methods known as SHAKE, designed to work with the Verlet integrator, and RATTLE, designed to work with the velocity Verlet integrator.

The utility of the constraint method rests on the fact that the fastest degrees of freedom place an upper bound on the time step, and the time step can therefore be increased if the fastest degrees of freedom are eliminated (constrained). A more general approach is based on the premise that the forces associated with different degrees of freedom evolve on different time scales, and the slower forces do not need to be computed as often as the faster forces. It turns out, for molecular systems, that the fastest forces, those associated with intramolecular interactions, require the least computational effort to evaluate ($\mathcal{O}(N)$), while the slowest forces, those associated with nonbonded interactions, require the most computational effort ($\mathcal{O}(N^2)$). Thus, in the interest of computational efficiency, it is clearly advantageous to calculate the rapidly evolving forces at each elementary time step, and only calculate the slowly evolving forces at a larger time interval that is several times the elementary time step. This is the essence of multiple time step MD (see Tuckerman & Martyna (2000) for a review), which is briefly summarized below.

Suppose that the total force on the particle i may be written

$$\boldsymbol{F}_i = \boldsymbol{F}_i^{\text{fast}} + \boldsymbol{F}_i^{\text{slow}}, \tag{18}$$

where $\boldsymbol{F}_i^{\text{fast}}$ and $\boldsymbol{F}_i^{\text{slow}}$ are the parts of the force that change rapidly and slowly, respectively. Now, according to Equation (13), we may write

$$\begin{aligned}\boldsymbol{\Gamma}(t+\Delta t) &= \{\exp(\mathrm{i}L_3\Delta t/2)[\exp(\mathrm{i}L_1\delta t/2)\exp(\mathrm{i}L_2\delta t) \\ &\quad \times \exp(\mathrm{i}L_1\delta t/2)]^n \exp(\mathrm{i}L_3\Delta t/2)\}\boldsymbol{x}(t) + \mathcal{O}(\Delta t^3),\end{aligned} \tag{19}$$

where $\mathrm{i}L_2$ is as before and

$$\mathrm{i}L_1 = \sum_{i=1}^{N}\left[\frac{\boldsymbol{F}_i^{\text{fast}}(\boldsymbol{r})}{m_i}\right]\cdot\nabla_{\boldsymbol{v}_i}, \tag{20}$$

$$\mathrm{i}L_3 = \sum_{i=1}^{N}\left[\frac{\boldsymbol{F}_i(\boldsymbol{r}) - \boldsymbol{F}_i^{\text{fast}}(\boldsymbol{r})}{m_i}\right]\cdot\nabla_{\boldsymbol{v}_i} \tag{21}$$

with $\delta t = \Delta t/n$. Thus, to evolve the system from time t to $t + \Delta t$ in one step of an MD simulation, the total force needs to be evaluated only once, and the fast component of the force n times. The small time step, δt, is determined by the time scale of the fast forces. As the computational effort required to evaluate the fast component of the force is negligible compared with that required for the slow force, this scheme results in nearly a factor of n decrease in computational effort. The resulting algorithm is straightforward to implement because each action of the evolution operators translates directly into velocity Verlet-like integration steps (Tuckerman & Martyna, 2000).

The majority of microscopic systems modeled by MD simulations are collections of atoms or molecules in a condensed phase. The usual quantum mechanical treatment of molecular systems assumes that the problem is separable between electronic and nuclear (or ion-core) degrees of freedom. Within this context, the most common approach in MD simulations is to consider the classical limit for the nuclear degrees of freedom. Thus, the phase space trajectory describes the dynamics of the atomic nuclei as classical point particles, while the electronic degrees of freedom are accounted for by the interaction potential $V(\boldsymbol{r})$. There is not a single prescription for choosing the form of the potential energy function and determining its parameters. In general, analytical forms for model systems are taken from molecular mechanics, and the specific parameters are dependent on the kind of system that is being modeled. For simple systems, the available information from zero-temperature electronic structure calculations is directly fitted to a convenient analytical function. For complex systems, electronic structure calculations for small molecular moieties are complemented with relevant experimental data pertinent to the phase being modeled. For a detailed overview of commonly used empirical potentials, see Leach (2001).

Although the number and nature of the potential energy terms varies from one application to the next, the potential for an arbitrarily complicated molecule such as a polymer is generally written as the sum of "bonded" and "nonbonded" terms. The bonded terms include energy penalties, usually represented by harmonic potentials, for deforming chemical bonds and the angles between bonds from their equilibrium values, as well as periodic potentials to describe the energy change as a function of the torsion (dihedral) angle about rotatable bonds. The bonded terms are usually taken to be diagonal, that is, there are no terms describing coupling between deformations of bonds and angles, bonds and torsions, etc. These off-diagonal terms are important for accurate reproduction of vibrational properties, as are anharmonic terms (e.g., cubic, quartic). The nonbonded terms, as the name suggests, describe the interactions between atoms in different molecules, or interactions within a molecule that are not completely accounted for by the bonded terms; that is, they are separated by more than two bonds. The nonbonded terms are typically assumed to be pairwise-additive functions of interatomic separation, *r*, and include van der Waals interactions and Coulomb interactions in polar or charged molecules in which the charge distribution is represented by partial charges (usually placed on the atoms). The van der Waals interactions include a term that is strongly repulsive (exponential or inverse 12th power of *r*), and a weakly attractive dispersion term (e.g., inverse 6th power of *r*, representing the induced dipole-induced dipole interaction). A popular form for the van der Waals interactions is the Lennard-Jones ("12-6") potential:

$$V(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \qquad (22)$$

Here, ε is the depth of the minimum, and σ is where the potential crosses zero and is a measure of the van der Waals diameter of the atom.

The number of bonded interactions grows in proportion to N, the number of atoms in the system, while the number of nonbonded interactions grows as N^2. Therefore, most of the computer time used for an MD simulation is spent on the evaluation of the nonbonded energies and forces. To reduce this burden, van der Waals interactions are typically ignored beyond some "cutoff" distance, usually around two or three atomic diameters. However, because the Coulomb potential is long-ranged, it is dangerous to truncate the electrostatic interactions. In condensed phase systems modeled using periodic boundary conditions (to eliminate surface effects in simulations of small systems that are meant to be in a bulk environment), the Ewald method can be used to sum all of the electrostatic interactions in an infinite, periodically replicated system. Details on the implementation of the Ewald sum may be found in Allen & Tildesley (1989).

Although MD simulations are most conveniently carried out in the microcanonical ensemble, there are disadvantages to doing so. An important aspect of performing simulation studies is to calculate experimentally measurable properties for comparison with measurements made at particular thermodynamic state points. Most experiments are carried out under the conditions of constant N and temperature, and either constant pressure or constant volume, that is, in the isobaric-isothermal or canonical statistical mechanical ensembles. In microcanonical simulations, the temperature and pressure are computed via averages of the kinetic energy and virial, respectively, and it is not known until after the simulation has been completed whether it was run at the desired state point. Moreover, in the canonical and isobaric-isothermal ensembles, the characteristic thermodynamic potentials are the Helmholtz free energy, $F(N, V, T)$, and Gibbs free energy, $G(N, P, T)$, respectively, and these free energies are easier to calculate from MD trajectories (see Frenkel & Smit (1996) for a discussion of methods for free energy calculations) than the entropy, $S(N, V, E)$, which is the characteristic thermodynamic potential of the microcanonical ensemble. In addition, constant pressure simulations are especially useful when the density of the system is not known a priori.

Several techniques for generating trajectories in ensembles other than the microcanonical ensemble have been developed (see Allen & Tildesley (1989) for an overview and Ciccotti et al. (1987) for a collection of key papers). The most powerful techniques are based on the concept of an "extended system" (ES), in which the atomic positions and momenta are supplemented by additional dynamical variables representing the coupling of the system to an external reservoir. For example, in the first application of the extended system concept to a simulation at constant pressure by H.C. Andersen in 1980, the system volume was taken to be a dynamical variable to simulate coupling to a pressure reservoir. In the Nosé implementation of constant temperature MD, a time-scaling "thermostat" variable was introduced. M. Parrinello and A. Rahman (1981) generalized Andersen's constant pressure method to allow the simulation cell to change shape as well as size by making the elements of the cell matrix dynamical variables. In these initial formulations of the ES concept, the additional dynamical variables were assigned fictitious kinetic and potential energies, and their equations of motion were derived using the Lagrangian formalism. Note that the choice of the form of the equations of motion for the additional degrees of freedom is not arbitrary, but rather should be made so that the resulting phase space distribution function is that of the desired ensemble. The Lagrangian formalism is not always the most convenient for handling extended systems. A more general approach is based on the formalism of non-Hamiltonian dynamics (Tuckerman & Martyna, 2000), which is sketched below. This approach was pioneered by W.G. Hoover, who rewrote the equations of motion for the Nosé thermostat in non-Hamiltonian form, leading to a much more straightforward implementation of constant temperature MD that was subsequently referred to as the Nosé–Hoover thermostat.

In the non-Hamiltonian dynamics approach, the equations of motion for the ES variables and a conserved Hamiltonian-like quantity, H', which includes terms for the kinetic and potential energies of the ES degrees of freedom, are chosen together such that the microcanonical distribution function generated by the extended system dynamics is formally equivalent (within a normalization constant) to the desired phase space distribution function of the atoms (e.g., canonical, isobaric-isothermal). The "non-Hamiltonian" resides in the fact that the necessary equations of motion cannot be derived from H', as in the case of Hamilton's equations (see Equations (1)). A complication that arises for non-Hamiltonian systems is that the phase space compressibility,

$$\kappa = \nabla_\Gamma \cdot \dot{\mathbf{\Gamma}}, \tag{23}$$

where Γ is now the $6N + 2M$ dimensional vector of coordinates and momenta (or velocities) of the N particles and M extended system variables, does not vanish as it does for Hamiltonian systems. Consequently, a metric factor must be included in the phase space measure,

$$\mathrm{d}\mu = \sqrt{g(\mathbf{\Gamma}, t)}\mathrm{d}\mathbf{\Gamma}, \tag{24}$$

where

$$\sqrt{g(\mathbf{\Gamma}, t)} = \exp[-w(\mathbf{\Gamma}, t)], \tag{25}$$

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \kappa, \tag{26}$$

so that the extended system phase space volume is conserved (i.e., Liouville's theorem is obeyed). Reversible, multiple time step integrators for extended system (Martyna et al., 1996) and non-equilibrium (Mundy et al., 2000) dynamics have been derived based on non-Hamiltonian mechanics using the Liouville operator formalism illustrated above for microcanonical dynamics.

In extended system MD simulations with temperature and pressure control, the instantaneous temperature (kinetic energy) and pressure (virial) fluctuate, and it is their average values that are equal to the values imposed by the thermal and pressure reservoirs. There are two other commonly employed approaches to temperature and pressure control (see Allen & Tildesley (1989) for a survey). In one class of methods, the instantaneous temperature and pressure are constrained to constant values. For large systems, the imposition of constraints is reasonable because fluctuations are small. By introducing Lagrange multipliers and applying Gauss's principle of least constraint, the constraints are applied in such a way so as to minimize the change in dynamics. The resulting phase space distribution functions are equal to the desired ones, multiplied by delta functions that account for the constraints on the temperature and pressure. Another technique is referred to as "coupling to an external bath," where the "bath" can be a temperature and/or pressure reservoir. This scheme drives the system to the desired temperature (pressure) at a rate determined by a preset temperature (pressure) relaxation time by scaling the velocities (volume and coordinates) at each time step. This method has the advantage that it is easily implemented, and the disadvantage that it does not generate states in any known statistical mechanical ensemble.

To date, most MD simulations have been driven by forces that were derived from assumed, empirical potentials. About two decades ago, the scope of MD simulations was greatly expanded by the introduction of a method by Car & Parrinello (1985) that, rather than relying on an empirical potential, utilizes interatomic potentials computed directly from the electronic structure which evolves simultaneously with the nuclear configuration. Thus, changes in electronic polarization and the possibility of chemical transformations are naturally included in the simulation. To this end, the wave function is expanded in a basis set, the expansion coefficients are treated as dynamical variables in an extended system Lagrangian, and the resulting equations of motion adiabatically propagate the electronic orbitals with respect to the nuclei, so that they remain on the ground state Born–Oppenheimer surface at each time step. Consequently, the need to solve the computationally burdensome electronic structure problem (self-consistent field) is avoided, and it is feasible to generate trajectories of sufficient length to study phenomena that were previously impossible to simulate. For technical details on the implementation of this so-called "ab initio MD" method, see Marx & Hütter (2000). Although the size of system (on the order of 100 atoms) and the length of trajectory (roughly tens of picoseconds) that can be simulated presently by ab initio MD are modest compared with simulations based on empirical potentials, it is clear that, as computational resources continue to evolve, applications of ab initio MD will become more and more prevalent and compelling.

DOUGLAS J. TOBIAS AND J. ALFREDO FREITES

See also **Hamiltonian systems; Local modes in molecular crystals; Lorentz gas; Newton's laws of motion; Protein dynamics**

Further Reading

Allen, M.P. & Tildesley, D.J. 1989. *Computer Simulation of Liquids*, Oxford: Oxford Science Publications

Car, R. & Parrinello, M. 1985. Unified approach for molecular dynamics and density-functional theory. *Physical Review Letters*, 55: 2471–2474

Ciccotti, G., Frenkel, D. & McDonald, I.R. (editors). 1987. *Simulation of Liquids and Solids: Molecular Dynamics and Monte Carlo Methods in Statistical Mechanics*, Amsterdam: North-Holland

Frenkel, D. & Smit, B. 1996. *Understanding Molecular Simulation: From Algorithms to Applications*, San Diego: Academic Press

Gibson, J.B., Golan, A.N., Milgram, M. & Vineyard, G.H. 1960. Dynamics of radiation damage. *Physical Review*, 120: 1229–1253

Leach, A.R. 2001. *Molecular Modelling: Principles and Applications*, 2nd edition, Harlow: Prentice-Hall

Martyna, G.J., Tuckerman, M.E., Tobias, D.J. & Klein, M.L. 1996. Explicit reversible integrators for extended systems dynamics. *Molecular Physics*, 87: 1117–1137

Marx, D. & Hütter, J. 2000. *Ab initio* molecular dynamics: theory and implementation. In *Modern Methods and Algorithms of Quantum Chemistry*, edited by J. Grotendorst, Jülich: Forschungszentrum Jülich

Mundy, C.J., Balasubramanian, B., Bagchi, K., Tuckerman, M.E., Martyna, G.J. & Klein, M.L. 2000. Nonequilibrium molecular dynamics. *Reviews in Computational Chemistry*, 14: 291–397

Parrinello, M. & Rahman, A. 1981. Polymorphic phase transitions in single crystals: a new molecular dynamics method. *Journal of Applied Physics*, 52: 7182–7190

Tuckerman, M.E. & Martyna, G.J. 2000. Understanding modern molecular dynamics: techniques and applications. *Journal of Physical Chemistry* B, 104: 159–178

MONODROMY PRESERVING DEFORMATIONS

Suppose $z \to A(z)$ is a $p \times p$ matrix valued meromorphic function (analytic except for a finite number of poles) on the Riemann sphere (also called a "complex projective one space" and denoted as $\boldsymbol{P}^1$) with poles at the points $a_\nu, \nu = 1, \dots, n$. Suppose Ψ is a fundamental solution for the linear system,

$$\frac{\mathrm{d}\Psi}{\mathrm{d}z} = A(z)\Psi, \tag{1}$$

defined in a neighborhood of a regular point, a_0. Analytic continuation of the fundamental solution along a closed curve γ which begins and ends at a_0 produces a change,

$$\Psi \to \Psi M_\gamma,$$

where M_γ is the *monodromy* associated with the path γ and the fundamental solution Ψ. The constant matrix M_γ depends only on the homotopy type of γ, and the map $\gamma \to M_\gamma^{-1}$ is a representation of the fundamental group $\pi_1\left(\boldsymbol{P}^1 \backslash \{a_\nu\}, a_0\right)$ called the monodromy representation of the differential equation.

In the 1850s, Bernhard Riemann solved the connection problem for solutions to the hypergeometric equation by exploiting the monodromy representation. The hypergeometric equation is the family of second-order equations on $\boldsymbol{P}^1$ with three regular singular points, and the connection problem is to understand how the local series solutions defined near the singularities are related to one another by analytic continuation. Riemann made some further speculations about monodromy representations, and (inspired by Riemann's work) David Hilbert asked if any representation of $\pi_1\left(\boldsymbol{P}^1 \backslash \{a_\nu\}, a_0\right)$ might be realized as the monodromy representation of a suitable differential equation. This is the 21st in the list of problems presented by Hilbert at the International Congress of Mathematicians in 1900. It always has a solution for differential equations with regular singular points (Plemelj, 1963), but it does not always have a solution if the differential equation is required to have simple poles. The book by Anasov and Bolibruch (1994) has a careful discussion of the history of this problem and its solution.

Suppose that the matrix $A(z)$ has simple poles,

$$A(z) = \sum_{\nu=1}^{n} \frac{A_\nu}{z - a_\nu}, \text{ with } \sum_{\nu=1}^{n} A_\nu = 0. \tag{2}$$

The second condition guarantees that ∞ is either a simple pole or a regular point. For simplicity, we also suppose that for each ν no pair of eigenvalues for A_ν differ by an integer. In this *nonresonant* case, the differential equation (1) is holomorphically equivalent to the differential equation,

$$\frac{\mathrm{d}\Phi}{\mathrm{d}z} = \frac{A_\nu}{z - a_\nu}\Phi, \tag{3}$$

in a neighborhood of $z = a_\nu$. In 1912, Ludwig Schlesinger asked how the residue matrices A_ν should depend on the pole locations $a = (a_1, a_2, \dots, a_n)$ in order that the monodromy representation remain unchanged. In case ∞ is a regular point, he discovered that the matrices A_ν must satisfy a nonlinear differential equation,

$$\mathrm{d}A_\mu = -\sum_{\nu \neq \mu} \frac{A_\mu A_\nu - A_\nu A_\mu}{a_\mu - a_\nu} d(a_\mu - a_\nu).$$

This equation is known as the Schlesinger equation and is an example of a *monodromy preserving deformation*.

Schlesinger's treatment of the existence question for such deformations was clarified in the work of Malgrange (1983b). Let $D_{\mu\nu} = \{a \in \boldsymbol{C}^n : a_\mu = a_\nu\}$ denote the set of points in $\boldsymbol{C}^n$ where the μ and ν poles collide—evidently such points are singular points for the Schlesinger equations. Malgrange showed that if one starts with a monodromy representation that is realized for a differential equation with simple poles at a_ν^0 and residues A_ν^0, then the deformation equation has a solution $A_\nu(a)$ which is analytic on the simply connected covering of $\boldsymbol{C}^n \backslash \cup D_{\mu\nu}$ *except possibly for poles along a hypersurface in this space*. He also showed that the hypersurface on which the solutions $A_\nu(a)$ has poles consists of those points for which a variant of the Riemann–Hilbert problem fails to have a solution. It is simplest to describe this variant in the nonresonant case (although Malgrange does not make this assumption). The point a will be a pole of $A_\nu(a)$ provided that it is not possible to find a differential equation with simple poles at a_ν for $\nu = 1, 2, \dots, n$ which reproduces the monodromy representation of the original equation,

$$\frac{\mathrm{d}\Psi}{\mathrm{d}z} = \sum_{\nu=1}^{n} \frac{A_\nu^0}{z - a_\nu^0}\Psi,$$

and is in the local holomorphic equivalence class of

$$\frac{\mathrm{d}}{\mathrm{d}z} - \frac{A_\nu^0}{z - a_\nu}$$

near $z = a_\nu$. It is possible that a given monodromy representation can be realized by a differential equation with simple poles without being able to specify the local holomorphic equivalence classes (see Bolibruch, 2002).

Nonlinear differential equations in the complex plane have the property that solutions can develop singularities away from the manifest singularities of the equations. The location of singularities of this sort depends on the initial conditions for the differential equation; thus, such points are called movable singularities. Differential equations with solutions whose only movable singularities are poles are said to have the Painlevé property. Paul Painlevé and Bertrand Gambier classified the simplest such nonlinear equations and determined six new families of transcendental functions that arise in the integration of these differential equations—now referred to as Painlevé I–VI (see Ince, 1956). Schlesinger understood, and as noted above, Malgrange proved that the Schlesinger equation has the Painlevé property. In a related development, Garnier (1912) showed that the Painlevé transcendents arise in the deformation of Fuchsian equations. Later, Jimbo, & Miwa (1981) showed how to obtain all six of the Painlevé equations by suitable specialization of the Schlesinger equations for 2×2 matrix systems. Painlevé VI arises for a linear system with simple poles of the type considered by Schlesinger. The other Painlevé equations arise by successive scaling that introduces irregular singularities and requires an extension of the Schlesinger analysis that was accomplished by Miwa, Jimbo, and Ueno in Jimbo, et al. (1981). Importantly, finding such representations for Painlevé equations is analogous to finding Lax pair representations for nonlinear equations such as the KdV equations (see Kapaev, 2002; Deift et al., 1999).

Flaschka & Newell (1980) first considered a generalization of the notion of monodromy preserving deformations to linear differential equations with higher rank singularities (equations with poles of order 2 or higher are said to have irregular singularities). More subtle local invariants, called Stokes multipliers, are fixed by the deformations, and in addition to the location of the singularities, the deformation parameters include formal expansion parameters that characterize the local asymptotics of solutions near the singularities. Based on Birkhoff's generalization of the Riemann–Hilbert problem (Birkhoff, 1913), Jimbo, Miwa, and Ueno (1981) generalized Flashka and Newell's work to the consideration of nonresonant differential equations with irregular singular points. They obtained the analogue of the Schlesinger equations for generalized monodromy preserving deformations of these equations. They also introduced the notion of a τ-function for such deformations. They conjectured that the τ-function is a holomorphic function of the deformation parameters whose 0 set is the set of pole locations for the solutions to the deformation equations. Miwa (1981) later showed that the τ-function is holomorphic and the Painlevé property is satisfied for the deformation equations in the domain of convergence for his explicit solution of the Birkhoff–Riemann–Hilbert problem. Malgrange (1983b) established the full conjecture for regular singularities without restrictions.

Geometric Perspectives

Röhrl (1957) formulated and solved a version of the Riemann–Hilbert problem on Riemann surfaces. The setting for the result involves vector bundles and holomorphic connections with singularities. Deligne (1970) developed a multidimensional generalization of the Riemann–Hilbert problem which took advantage of rather abstract categorical constructions. Malgrange (1983a) introduced Stokes' sheaves at irregular singular points to describe local holomorphic equivalence classes. More recently, progress in analyzing the local structure of differential equations at irregular singular points has come by introducing sophisticated ideas from algebraic geometry (see Varadarajan, 1996). Part of the impetus for these developments is the effort to achieve clarity in a subject that is subtle enough to have inspired the misinterpretation of fundamental results and even errors in proofs by such luminaries as George D. Birkhoff (see Anasov and Bolibruch (1994) for an account).

Among these developments, there is one geometric idea introduced by Röhrl and more fully realized by Malgrange (1983c), which provides helpful insight into the nature of the Birkhoff–Riemann–Hilbert deformation problem introduced in Jimbo et al. (1981). Suppose that $a^0 = \{a_1^0, a_2^0, \ldots, a_n^0\}$ is a collection of distinct points in $\boldsymbol{P}^1$ and $A^0(z)$ is a rational matrix valued function whose principal part at $z = a_\nu$ is

$$\frac{A_{r+1}^0}{(z - a_\nu)^{r+1}} + \frac{A_r^0}{(z - a_\nu)^r} + \cdots + \frac{A_1^0}{z - a_\nu}.$$

The integer $r = r_\nu$ is called the rank of the singularity. Let

$$\nabla^0 = \frac{\mathrm{d}}{\mathrm{d}z} - A^0(z)$$

denote the associated *connection*. We will say that the connection ∇^0 is *nonresonant* if the leading coefficient A_{r+1}^0 has distinct eigenvalues at each of the singularities with the additional requirement that these eigenvalues do not differ by integers in case $r = 0$. Incidentally, much of the machinery that has been introduced to study irregular singular points is devoted to understanding the substantial complications that arise for *resonant* connections (Varadarajan, 1996).

It is a result of Deligne (1970) that the holomorphic equivalence class of ∇^0 on the *punctured* sphere $\boldsymbol{P}^1\backslash\{a_\nu^0\}$ is determined by its monodromy representation. On the other hand, in a neighborhood of each of the singularities, the connection ∇^0 belongs to a local holomorphic equivalence class of rank r connections. For nonresonant connections, this local equivalence class, is naturally a fiber bundle whose base consists of *diagonal* connections,

$$\frac{\mathrm{d}}{\mathrm{d}z} - \frac{H_r}{(z-a_\nu)^{r+1}} + \cdots + \frac{H_0}{z-a_\nu}, \tag{4}$$

whose leading coefficient, H_r, has distinct eigenvalues which do not differ by integers in case $r = 0$. For $r = 0$ the form (4) determines the local equivalence class, but for $r \geq 1$ the fibers in the bundle are essentially the Stokes multipliers, S. Each nonresonant connection ∇^0 has a *formal reduction* to a type (4) connection and the holomorphic equivalence class for ∇^0 is determined by this formal reduction together with the Stokes multipliers. The Stokes multipliers determine the relations between solutions to $\nabla^0\Psi = 0$ that are sectorially asymptotic to flat sections of the connection (4) (see Sibuya, 1977; Malgrange, 1983c).

Let t denote the collection of diagonal entries for all the matrices H_j. The fiber bundle of local holomorphic equivalence classes has a natural flat connection (it is a so-called local system (Varadarajan, 1996)) and we let $t \to \sigma(t)$ denote the flat (locally constant) section of this bundle normalized to give the class of ∇^0 at the parameter value t^0 which corresponds to ∇^0. The generalized monodromy preserving deformation problem is to find a family of connections $\nabla(a, t)$ which reproduces the monodromy representation of ∇^0 in the punctured plane $\boldsymbol{P}^1\backslash\{a_\nu\}$ and whose local holomorphic equivalence class at each a_j is the equivalence class associated with the value of the flat section $\sigma(t)$ (i.e., locally constant Stokes multipliers).

The version of the Birkhoff–Riemann–Hilbert problem relevant to this deformation problem is: Given a representation of $\pi_1(\boldsymbol{P}^1\backslash a, a_0)$ and an assignment of a compatible local holomorphic equivalence class of a type r_ν connection to each of the points a_ν, find a connection on the trivial bundle over $\boldsymbol{P}^1$ which reproduces this data (compatible means that the local monodromy at a_ν should reproduce the monodromy coming from the representation). Because noncanonical choices must be made to define the Stokes multipliers, this version of the Birkhoff–Riemann–Hilbert problem is even complicated to state if the connection with local holomorphic equivalence classes is not understood. Based on ideas of Malgrange, it has been shown that the τ-function introduced by Miwa, Jimbo, and Ueno (1981) has zeros for the deformation problem described above precisely when the associated Birkhoff–Riemann–Hilbert problem fails to have a solution (Palmer, 1999). It should also be mentioned that a crucial technique used both by Jimbo et al. (1981) and by Malgrange (1983b,c) has the prolongation of the connection originally defined on $\boldsymbol{P}^1$ to an integrable connection defined on $\boldsymbol{P}^1 \times \mathcal{D}$ where $\mathcal{D}$ is the space of deformation parameters. In this multidimensional setting, Malgrange adapts Rörhl's technique to prove the existence of a prolongation for the original connection on a vector bundle. The solution of the original deformation problem depends on the triviality of certain restrictions of this vector bundle.

Applications

The Riemann–Hilbert problem and the Painlevé transcendents have generated a steady stream of papers over the years, but interest in monodromy preserving deformations was largely dormant in the period from the early 1920s to the late 1970s. A rebirth of interest in the subject dates from the surprising discovery by Wu, McCoy, Tracy, and Baruch (WMTB) (1976) that the two-point scaling function for the two-dimensional Ising model can be expressed in terms of a Painlevé transcendent of type III. The two-dimensional Ising model is a statistical model of magnetism which exhibits a phase transition in the infinite volume limit. It is so far the only model with a phase transition in which the correlation functions can be studied exactly in the infinite volume limit. The scaling limit that was investigated in Wu et al. (1976) examines the asymptotics of the spin correlations at the length scale of the correlation length (which tends to ∞) as the temperature approaches the critical point. Not only is this analysis important for an understanding of critical phenomena in statistical physics, but it was understood at about the same time that the mathematics of critical phenomena in statistical physics was recognized as the same as the mathematics of renormalization for quantum field theory.

In a remarkable series of papers (Sato et al., 1978, 1979a,b,c, 1980) starting in 1978, Sato, Miwa, and Jimbo found a context for the WMTB result by showing that the n-point correlation for the Ising model could be expressed in terms of the solutions to "monodromy preserving deformations" of the Euclidean Dirac equation (with a mass term). Note that the equation which is the analogue of the Dirac equation in the earlier development is not the differential equation (1) but the Cauchy–Riemann equation $\bar{\partial}\psi = 0$. The correlation functions in the Ising model were the inspiration for their introduction of the τ-function in other contexts. See Palmer & Tracy (1983), Palmer (1993), Palmer et al. (1994), and Palmer (2002) for further developments of these ideas. In Jimbo et al. (1980), Sato, Miwa, Jimbo, and Mori explained the use of their techniques to analyze the impenetrable Bose gas and incidentally remarked on a connection with an important function in random matrix theory.

The analysis of correlations in integrable models has since been developed by Its et al. (1990), who also codified some of the ideas in a theory of completely integrable integral operators (Korepin et al., 1993). The connection between the random matrix theory and Painlevé functions has been systematically worked out by Tracy & Widom (1994, 1996), who independently developed an integral equation approach to the nonlinear "deformation" equations.

Jimbo (1982) used the monodromy preserving deformation representation of Painlevé functions and the notion of the τ-function to analyze the connection problem for Painlevé functions. The problem is to understand how different families of solutions to the Painlevé equations defined by local asymptotic developments at different singular points are related to one another. Somewhat earlier the connection problem for Painlevé III was worked out by McCoy, Tracy, and Wu (1977) where the answer has consequences for the short distance behavior of the scaling function. Its and Novokshenov (1986) have developed these ideas further.

Although not quite a development in monodromy preserving deformation theory, a "steepest decent" technique based on Riemann–Hilbert ideas has emerged in work of Deift, Zhou and Its that is effective for the asymptotic analysis of solutions to nonlinear equations that arise from spectral or monodromy preserving deformations (Its, 1981; Deift et al., 1997; Bleher & Its, 1999; Deift et al., 1999).

Painlevé functions were discovered in two-dimensional quantum gravity and connections with monodromy preserving deformations are discussed in Moore (1990). Painelevé functions also arise in two-dimensional topological field theory (Dubrovin, 1999).

Further developments are discussed in Harnad & Its (2002).

John Palmer

See also **Painlevé analysis; Riemann–Hilbert problem**

Further Reading

Anasov, D.V. & Bolibruch, A.A. 1994. *The Riemann–Hilbert problem*, Aspects of Mathematics E, vol. 22. Braunschweig; Wiesbaden: Vieweg

Birkhoff, G.D. 1913. The generalized Riemann problem for linear differential equations. *Proceedings of the American Academy of Arts and Sciences*, 32: 531–568

Bleher, P. & Its, A.R. 1999. Semiclassical asymptotics of orthogonal polynomials, Riemann–Hilbert problem, and universality in the matrix model. *Annals of Mathematics*, 150 (2): 185–266

Bolibruch, A. 2002. Inverse problems for linear differential equations with meromorphic coefficients. *CRM Workshop: Isomonodromic Deformations and Applications in Physics*, edited by J. Harnard & A. Its, Providence, RI: American Mathematical Society, pp. 3–25

Deift, P., Its, A., Kapaev, A. & Zhou, X. 1999. On the algebro-geometric integration of the Schlesinger equations. *Communications in Mathematical Physics*, 203: 613–633

Deift, P., Its, A. & Zhou, X. 1997. A Riemann–Hilbert approach to asymptotic problems arising in the theory of random matrix models, and also in the theory of integrable statistical mechanics. *Annals of Mathematics*, 146: 149–237

Deift, P., Kriecherbauer, T., McLaughlin, K., Venakides, S. & Zhou, X. 1999. Uniform asymptotics for polynomials orthogonal with respect to varying exponential weight and applications to universality questions in random matrix theory. *Communications in Pure and Applied Mathematics*, 52: 1335–1425

Deligne, P. 1970. *Équations différentielles à points singuliers réguliers*, Berlin and New York: Springer

Dubrovin, B. 1999. Painlevé transcendents in two-dimensional topological field theory. In *The Painlevé Property, One Hundred Years Later*, edited by R. Conte, New York: Springer, pp. 287–412

Flaschka, H. & Newell, A.C. 1980. Monodromy and spectrum preserving deformations I. *Communications in Mathematical Physics*, 76: 65–116

Garnier, R. 1912. Sur les équationes différentielles du troisième ordre dont l'intégrale générale est uniforme et sur une classe d'équations nouvelles d'ordre supérieur dont l'intégrale générale a ses points critique fixes. *Annales de l'Ecole Normale Supérieure*, (3) 29: 1–126

Harnad, J. & Its, A. 2002. *CRM Workshop: Isomonodromic Deformations and Applications in Physics*, edited by J. Harnard & A. Its, Providence, RI: American Mathematical Society

Ince, E.L. 1956. *Ordinary Differential Equations*, New York: Dover

Its, A.R. 1981. Asymptotics of solutions of the nonlinear Schrödinger equation and isomonodromic deformations of systems of linear differential equations. *Soviet Mathematics. Doklady*, 24: 454–456

Its, A.R., Izergin, A.G., Korepin, V.E. & Slavnov, N.A. 1990. Differential equations for quantum correlation functions. *International Journal of Modern Physics* B, 4: 1003–1037

Its, A.R. & Novokshenov, V.Yu. 1986. *The Isomonodromic Deformation Method in the Theory of Painlevé Equations*, Berlin: Springer

Jimbo, M. 1982. Monodromy problem and the boundary conditions for some Painlevé equations. *Publ. R.I.M.S. Kyoto University*, 18: 1137–1161

Jimbo, M. & Miwa, T. 1981. Monodromy preserving deformations of linear ordinary differential equations with rational coefficients II. *Physica* D, 2: 407–448

Jimbo, M. & Miwa, T. 1983. Monodromy preserving deformations of linear ordinary differential equations with rational coefficients III. *Physica* D, 2: 26–46

Jimbo, M., Miwa, T., Sato, M. & Mori, Y. 1980. Density matrix of an impenetrable bose gas and the fifth Painlevé transcendent. *Physica* D, 1: 80–158

Jimbo, M., Miwa, T. & Ueno, K. 1981. Monodromy preserving deformations of linear ordinary differential equations with rational coefficients I. *Physica* D, 2: 306–352

Kapaev, A.A. 2002. Lax pairs for Painlevé equations, *CRM workshop: Isomonodromic Deformations and Applications in Physics*, edited by J. Harnard & A. Its, Providence, RI: American Mathematical Society, pp. 37–47

Korepin, V.E., Bogoliubov, N.M. & Izergin, A.G. 1993. *Quantum Inverse Method and Correlation Functions*, Cambridge and New York: Cambridge University Press

Malgrange, B. 1983a. *La classification des connections irrégulieéres à une variable*, Boston: Birkhäuser, pp. 381–400

Malgrange, B. 1983b. *Sur les déformations isomonodromiques I, Singularités régulières*, Boston: Birkhäuser, pp. 401–426

Malgrange, B. 1983c. *Sur les déformations isomonodromiques II, Singularités irrégulières*, Boston: Birkhäuser, pp. 427–438

McCoy, B.M., Tracy, C.A. & Wu, T.T. 1977. Painlevé functions of the third kind. *Journal of Mathematical Physics*, 18: 1058–1092

Miwa, T. 1981. Painlevé property of monodromy preserving equations and the analyticity of τ-function. *Publ. R.I.M.S. Kyoto University*, 17: 703–721

Moore, G. 1990. Geometry of the string equations. *Communications in Mathematical Physics*, 133: 261–304

Palmer, J. 1993. Tau functions for the Dirac operator in the Euclidean plane. *Pacific Journal of Mathematics*, 160: 259–342

Palmer, J. 1999. Zeros of the Jimbo, Miwa, Ueno tau function. *Journal of Mathematical Physics*, 40: 6638–6681

Palmer, J. 2002. Short distance asymptotics of Ising correlations, *Journal of Mathematical Physics*, 43(2): 918–953

Palmer, J., Beatty, M. & Tracy, C. 1994. Tau functions for the Dirac operator on the Poincaré disk. *Communications in Mathematical Physics*, 165: 97–173

Palmer, J. & Tracy, C.A. 1983. Two dimensional Ising correlations: the SMJ analysis. *Advances in Applied Mathematics*, 4: 46–102

Plemelj, J. 1963. *Problems in the Sense of Riemann and Klein*, New York: Interscience

Röhrl, H. 1957. Das Riemann–Hilbertsche problem der theorie der linearen differentialgleichungen. *Mathematische Annalen*, 133: 1–25

Sato, M., Miwa, T. & Jimbo, M. 1978. Holonomic quantum fields I. *Publ. R.I.M.S. Kyoto University*, 14: 223–267

Sato, M., Miwa, T. & Jimbo, M. 1979a. Holonomic quantum fields II. *Publ. R.I.M.S. Kyoto University*, 15: 201–278

Sato, M., Miwa, T. & Jimbo, M. 1979b. Holonomic quantum fields III. *Publ. R.I.M.S. Kyoto University*, 15: 577–629

Sato, M., Miwa, T. & Jimbo, M. 1979c. Holonomic quantum fields IV. *Publ. R.I.M.S. Kyoto University*, 15: 871–972

Sato, M., Miwa, T. & Jimbo, M. 1980. Holonomic quantum fields V. *Publ. R.I.M.S. Kyoto University*, 16: 531–584

Schlesinger, L. 1912. Über eine klasse von differentialsystemen beliebiger ordnung mit festen kritischen punkten. *Journal für die reine und angewandte Mathematik*, 14: 95–145

Sibuya, Y. 1977. Stokes' phenomena. *Bulletin of the American Mathematical Society*, 83: 1075–1077

Tracy, C.A. & Widom, H. 1994. Fredholm determinants, differential equations and matrix models. *Communications in Mathematical Physics*, 163: 33–72

Tracy, C.A. & Widom, H. 1996. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177: 727–754

Varadarajan, V.S. 1996. Linear meromorphic differential equations: a modern point of view, *Bulletin of the American Mathematical Society*, 33(1): 1–41

Wu, T.T., McCoy, B., Tracy, C.A. & Barouch, E. 1976. Spin-spin correlation functions for the two dimensional Ising model: exact theory in the scaling region. *Physical Review* B, 13: 316–374

MONTE CARLO METHODS

Monte Carlo methods collectively refer to a set of computational tools based on random samples. These methods have a very long history that can be traced back to biblical times although a systematic foundation was established only in the early days of electronic computing. A group of mathematicians and physicists of the Los Alamos project during World War II, in particular, John von Neumann, Nick Metropolis, Jeffrey Kahn, Enrico Fermi, Stan Ulam, and their collaborators, developed Monte Carlo methods for estimating eigenvalues of the Schrödinger equation and for studying random neutron diffusion in fissile material used in atomic bombs. Since then, Monte Carlo methods have found applications in a vast number of scientific disciplines and are, becoming increasingly popular with the advent of cheap and powerful computers.

Starting with a space $\mathcal{X}$, let $\varphi : \mathcal{X} \to R$ be function and π be a probability density. In essence, Monte Carlo methods are concerned with

- *Integration*: Evaluate the expectation of φ under π; that is,
$$I_\pi(\varphi) = \int_{\mathcal{X}} \varphi(x)\,\pi(x)\,\mathrm{d}x. \tag{1}$$
- *Simulation*: Simulate some random samples $\{X_i\}_{i=1}^N$ distributed according to π.

Integration and simulation are essential elements of many scientific problems arising, for example, in statistical physics, statistics, engineering, and economics. The probability density π may correspond to a posterior density in the framework of Bayesian statistics or a Boltzmann–Gibbs distribution in equilibrium statistical thermodynamics.

In most applications, it is impossible to compute the expectations of interest in closed-form as the dimension of the space $\mathcal{X}$ is large; for example, $\mathcal{X} = R^{1000}$ or $\mathcal{X} = \{-1, 1\}^{5000}$. Moreover, π is often only known up to a normalizing constant, that is,

$$\pi(x) = cf(x), \tag{2}$$

where $f : \mathcal{X} \to R$ is known but c is unknown. Consequently, standard deterministic numerical integration schemes are very inefficient.

The basic idea of Monte Carlo methods is to compute the expectation(s) of interest in Equation (1) via random samples. Assume the random samples $\left\{X^{(i)}\right\}_{i=1}^N$ distributed according to π are available, then an estimate of $I_\pi(\varphi)$ can be given by the empirical average

$$\widehat{I}_\pi(\varphi) = \frac{1}{N}\sum_{i=1}^{N} \varphi\left(X^{(i)}\right).$$

In other words, π is approximated by

$$\widehat{\pi}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta\left(x - X^{(i)}\right),$$

where $\delta(\cdot)$ is the Dirac delta function. This estimate has "good" theoretical properties. First, it is unbiased.

Second, if the samples $\left\{X^{(i)}\right\}_{i=1}^{N}$ are statistically independent, then the variance of this estimate is given by

$$\frac{\int [\varphi(x) - I_\pi(\varphi)]^2 \pi(x)\, \mathrm{d}x}{N}; \tag{3}$$

that is, the convergence rate of the estimation error to zero is independent of the dimension of $\mathcal{X}$. If the numerator of (3) is "small", then only a few hundreds/thousands of samples might be necessary to achieve a good precision even if $\mathcal{X} = R^{1000}$.

Thus far, it has been shown that estimates with good theoretical properties can be easily obtained from a large set of samples distributed according to the probability density of interest, but how to generate these samples? The Monte Carlo integration problem then becomes that of simulation. Generating samples from nonstandard probability distributions/densities only known up to a proportionality factor is a central problem in Monte Carlo methods.

In 1947, von Neumann, in a letter to Ulam, described an algorithm known as the Rejection Method, to simulate samples distributed according to a "target" probability density. Given a target probability density π that is only known up to a normalizing constant, that is, π is of the form in (2) and only f is known, one chooses a candidate density g that is easy to sample from and a number M satisfying $f(x) \leq Mg(x)$ for any $x \in \mathcal{X}$. The algorithm proceeds as follows:

(a) Sample $X^* \sim g$ ("$\sim$" is a standard notation for "distributed according to") and compute

$$\alpha\left(X^*\right) = \frac{f(X^*)}{Mg(X^*)}.$$

(b) Sample a uniform random variable U on [0, 1]. If $U \leq \alpha(X^*)$, return X^* otherwise go back to (a).

It can be easily verified that the accepted sample follows the target distribution π. The problem with this method is the difficulty in finding a candidate density g satisfying $f(x) \leq Mg(x)$ that is easy to sample from. Moreover, even if such a g can be found, the acceptance probability is $(Mc)^{-1}$, which can be extremely low when $\mathcal{X}$ is a high-dimensional space. Consequently, the algorithm may take a long time to generate a single sample.

A more powerful class of algorithms are the Markov chain Monte Carlo (MCMC) techniques. Given a "target" density π, the key idea of MCMC is to build a Markov chain $\{X_i\}_{i\geq 0}$ of transition kernel $K(\cdot|\cdot)$ such that

$$\int K\left(x'\middle| x\right)\pi(x)\, \mathrm{d}x = \pi\left(x'\right), \tag{4}$$

i.e., if $X_{i-1} \sim \pi$ then any sample $X_i \sim K(\cdot| X_{i-1})$ is also distributed according to π. The target density π in (4) is known as the invariant (or stationary) distribution of the Markov chain. Under ergodicity conditions on the kernel $K(\cdot|\cdot)$, the samples $\{X_i\}_{i\geq 0}$ are distributed according to π asymptotically (as $i \to \infty$) regardless of the value of initial state X_0.

There are an infinity of kernels with invariant distribution π. However, the most frequently used are the Metropolis–Hastings kernels, first proposed by Metropolis et al. (1953), and then generalized by Hastings (1970). Green (1995) extended this to the case where $\mathcal{X}$ is a disjoint union of subspaces of different dimensions.

Given a candidate distribution $q(\cdot|\cdot)$ that is easy to sample from, the Metropolis–Hastings kernel can be constructed as follows: given X_{i-1}

(a) Sample a candidate $X^* \sim q(\cdot| X_{i-1})$ and compute

$$\alpha\left(X_{i-1}, X^*\right) = \frac{\pi(X^*)}{\pi(X_{i-1})}\frac{q(X_{i-1}| X^*)}{q(X^*| X_{i-1})}.$$

(b) Sample a uniform random variable U on [0, 1]. If $U \leq \alpha(X_{i-1}, X^*)$, then set $X_i = X^*$ otherwise set $X_i = X_{i-1}$.

It can be verified that the Markov chain $\{X_i\}_{i\geq 0}$ generated by the above kernel has invariant distribution π. Knowledge of the normalizing constant of π is not needed as it disappears in the acceptance ratio $\alpha(X_{i-1}, X^*)$. The major problem with such Markov chain techniques is that only samples from π are obtained as $i \to \infty$. Nevertheless, the Metropolis–Hastings algorithm and MCMC algorithms, in general, have been applied successfully in numerous applications. MCMC methods form the basis of global optimization algorithms such as simulated annealing (Van Laarhoven & Arts 1987). For a thorough treatment of the theory and applications of MCMC algorithms, the reader is referred to Robert & Casella, (1999).

Importance sampling (IS) is based on assigning weights (or importance) to samples generated from a candidate density q so as to approximate the target density π. Suppose that $\pi(x) = w(x)\, q(x)$, where q is a density that is easy to sample from. Then, given random samples $\left\{X^{(i)}\right\}_{i=1}^{N}$ distributed according to q, the expectation $I_\pi(\varphi)$ in Equation (1) can be estimated by the empirical average

$$\widehat{I_\pi}(\varphi) = \frac{1}{N}\sum_{i=1}^{N}\varphi\left(X^{(i)}\right) w\left(X^{(i)}\right).$$

In other words, π is approximated by

$$\widehat{\pi_1}(x) = \frac{1}{N}\sum_{i=1}^{N} w\left(X^{(i)}\right)\delta\left(x - X^{(i)}\right).$$

This strategy assumes full knowledge of π as $w\left(X^{(i)}\right) = \pi\left(X^{(i)}\right)/q\left(X^{(i)}\right)$ needs to be determined. When π is only specified upto normalizing constant, write

$$\pi(x) = \frac{w(x)\,q(x)}{\int_{\mathcal{X}} w(x)\,q(x)\,\mathrm{d}x}$$

and apply importance sampling to both numerator and denominator to yield the following approximation:

$$\widehat{\pi}_2(x) = \sum_{i=1}^{N} \frac{w\left(X^{(i)}\right)}{\sum_{j=1}^{N} w\left(X^{(j)}\right)} \delta\left(x - X^{(i)}\right).$$

An approximate sample from π can be obtained by sampling from the discrete distributions $\widehat{\pi}_1$ or $\widehat{\pi}_2$.

Applications

The range of applications of Monte Carlo methods is vast. Listed below are some of the more well-known areas.

Integral equations: IS methods have been widely used to solve linear systems and integral equations appearing in particle transport problems. The basic idea is to give a probabilistic approximation of operators of the form $(I - H)^{-1} = \sum_{i=0}^{\infty} H^i$; see Sobol (1994) for details.

Computational physics and chemistry simulation: Monte Carlo methods are used in physics and chemistry to simulate from Ising models, simulate self-avoiding random walks, and compute the free energy, entropy, and chemical potential over systems; see, for example, Frenkel & Smith (1996).

Quantum physics: To compute the dominant eigenvalue and eigenvector of a positive operator, it is possible to use a stochastic version of the power method. This is often applied to the Schrödinger equation; see Melik-Alaverdian & Nightingale (1999) for a recent review.

Statistics: Performing inference in complex statistical models invariably requires sampling from high dimensional probability distributions. See Gilks et al. (1996) for applications of MCMC and Doucet et al. (2001) for applications of IS-type methods to such problems.

Arnaud Doucet and Ba-Ngu Vo

See also **Random walks; Stochastic processes**

Further Reading

Doucet, A., De Freitas, J.F.G. & Gordon, N.J. (editors). 2001. *Sequential Monte Carlo Methods in Practice*, New York: Springer

Frenkel, D. & Smith, B. 1996. *Understanding Molecular Simulation: From Algorithms to Applications*, Boston: Academic Press

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (editors). 1996. *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall

Green, P.J. 1995. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82: 711–732

Hastings, W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57: 97–109

Melik-Alaverdian, M. & Nightingale, M.P. 1999. Quantum Monte Carlo methods in statistical mechanics. *International Journal of Modern Physics* C, 10: 1409–1418

Metropolis, N., Rosenblutt, N., Rosenblutt, A.W., Teller, M.N. & Teller, A.H. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics,* 21: 1087–1092

Robert, C.P. & Casella, G. 1999. *Monte Carlo Statistical Methods*, New York: Springer

Sobol, I.M. 1994. *A Primer for the Monte Carlo Method*, Boca Raton, FL: CRC Press

Van Laarhoven, P.J. & Arts, E.H.L. 1987. *Simulated Annealing: Theory and Applications*, Amsterdam: Reidel

MORPHOGENESIS, BIOLOGICAL

One of the central problems in developmental biology is to understand how patterns and structures are laid down. From the initially almost homogeneous mass of dividing cells in an embryo emerges the vast range of pattern and structure observed in animals. For example, the skeleton is laid down during chondrogenesis when chondroblast cells condense into aggregates that lead eventually to bone formation. The skin forms many specialized structures such as hair, scales, feathers, and glands. Butterfly wings exhibit spectacular colors and patterns, and many animals develop dramatic coat patterns.

Although genes play a key role, genetics say nothing about the actual mechanisms that produce pattern and structure—the process known as morphogenesis—as an organism matures from embryo to adult. Tissue movement and rearrangement are the key features of almost all morphogenetic processes and arise as the result of complex mechanical, chemical, and electrical interactions. Despite the recent vast advances in molecular biology and genetics, little is understood of how these processes conspire to produce pattern and form. There is the danger of falling into the practices of the 19th century, when biology was steeped in the mode of classification and there was a tremendous amount of list-making activity. This was recognized by D'Arcy Thompson, in his classic work first published in 1917 (see Thompson (1992) for the abridged version). He was the first to develop theories for how certain forms arose, rather than simply cataloging different forms, as was the tradition at that time.

At the heart of a number of developmental phenomena is the process of convergence-extension, in which a tissue narrows along one axis while extending along another. This process represents the integration of local cellular behavior that produces forces to change the shape of the cell population. In fact, convergence-extension is essentially responsible for the transformation of the spherical egg into the elongated, bilaterally symmetric vertebrate body axis (Keller et al., 1992).

Cell fate and position within the developing embryo can be strongly influenced by environmental factors. Therefore, to investigate the process of morphogenesis, one must really address the issue of how the embryo organizes the complex spatiotemporal sequence of signalling cues necessary to develop structure in a controlled and coordinated manner. Structure can form through tissue movement and rearrangement. Theoretical studies in this area include the early purse-string model (Odell et al., 1981) for tissue folding in which, in response to a large deformation, cells were proposed to actively contract and, in doing so, cause a large deformation in neighboring cells which, in turn, also contract, setting up a propagating contraction wave which leads to tissue folding. This model was applied to a variety of developmental problems and provided the precursor to the mechanochemical theory of developmental patterning developed by Oster, Murray, and coworkers (for review, see Murray, 2003). This approach emphasized the link between tissue mechanics and chemical regulation and has been applied widely in both developmental biology and medicine.

Discrete-cell modeling approaches have subsequently been developed in which morphogenesis is hypothesized to occur via mechanical rearrangement of neighbors in an epithelial sheet, and computational finite elements have been developed to test various theoretical explanations for morphogenesis (Weliky et al., 1991; Davidson et al., 1995).

In all these models, individual cell movements within the tissue are determined by the balance of mechanical forces acting on the cell. Such models can exhibit tissue folding, thickening, invagination, exogastrulation, and intercalation, and have been shown to capture many of the key aspects of processes such as gastrulation, neural tube formation, and ventral furrow formation in *Drosophila*. Cells can also sort out depending on their type, and this has led to the theory of differential adhesion and energy minimization (Steinberg, 1970).

Models for tissue motion are not amenable to a mathematical analysis and tend to be highly computation based. However, models for how cells differentiate can be addressed mathematically. Broadly speaking, there are two classes of such models. In one class, the chemical pre-pattern models, it is hypothesized that a chemical signal is set up in some way and cells respond to this signal by differentiating. In the other class, the cell movement models, it is hypothesized that cells respond to mechanochemical cues and form aggregates. Cells in high density aggregates are then assumed to differentiate (see Murray, 2003, for details).

The fact that such models can lead to the generation of spontaneous order was first realized by Alan Turing (1952), who showed that a system of chemicals, stable in the absence of diffusion, could be driven unstable by diffusion. He proposed that such a spatial distribution of chemicals (which he termed *morphogens*) could set up a pre-pattern to which cells could respond and differentiate accordingly. He was one of the first to postulate the existence of such chemicals, and morphogens have now been discovered. It is still not clear that morphogen patterns in biology are set up by the mechanism proposed by Turing, but Turing patterns have been found in chemistry (see Maini et al., 1997, for a review).

A variety of models based on different biology give rise to mathematical formulations in terms of coupled systems of highly nonlinear partial differential equations. The analysis of these models has, to date, yielded a number of common behaviors. This has led to the idea of using such models to determine developmental constraints. That is, independent of the underlying biology, such models predict that only certain patterns are selected at the expense of others and thus there is a limited variation. This has consequences for evolution. For example, application of mitotic inhibitors to developing limbs produces smaller limbs with reduced elements. Some of the resultant variants look very similar to the pattern of evolution in other species, suggesting that these species may be more closely related than previously thought (Oster et al., 1988). Moreover, the construction rules generated by a study of developmental constraints is another, perhaps more mechanistic, way of describing how different species are related other than the topological deformation approach of D'Arcy Thompson.

Other approaches to morphogenesis and pattern formation include cellular automata models, in which individual entities (cells, for example) behave according to a set of rules. Such models allow one to include much more biological detail and to investigate finer grain patterns than those possible in the continuum approaches discussed above (see, for example, Alt et al., 1997). However, to date they lack a detailed mathematical underpinning.

The recent spectacular advances in molecular genetics raise the issue of how we can combine the enormous amount of data now being generated at this level with the data available from the classical experiments at the cell and tissue level to provide a coherent theory for pattern formation and morphogenesis. This leads to the problem of modeling across a vast range of spatial and temporal scales. The mathematics for this has not yet been developed and is one of the challenges presently being addressed.

Philip K. Maini

See also **Brusselator; Cellular automata; Pattern formation; Reaction-diffusion systems; Turing patterns**

Further Reading

Alt, W., Deutsch, A. & Dunn G. (editors). 1997. *Dynamics of Cell and Tissue Motion*, Basel: Birkhäuser

Davidson, A., Koehl, M.A.R., Keller, R. & Oster, G.F. 1995. How do sea urchins invaginate? Using biomechanics to distinguish between mechanisms of primary invagination. *Development*, 121: 2005–2018

Keller, R., Shih, J. & Domingo, C. 1992. The patterning and functioning of protrusive activity during convergence and extension of the *Xenopus* organiser. *Development Supplement*, 81–91

Maini, P.K., Painter, K.J. & Chau, H.N.P. 1997. Spatial pattern formation in chemical and biological systems. *Faraday Transactions*, 93(20): 3601–3610

Murray, J.D. 2003. *Mathematical Biology II: Spatial Models and Biomedical Applications*, 3rd edition, Berlin and New York: Springer

Odell, G.M., Oster, G., Alberch, P. & Burnside, B. 1981. The mechanical basis of morphogenesis. I. Epithelial folding and invagination. *Developmental Biology*, 85: 446–462

Oster, G.F., Shubin, N., Murray, J.D. & Alberch, P. 1988. Evolution and morphogenetic rules. The shape of the vertebrate limb in ontogeny and phylogeny. *Evolution*, 45: 862–884

Steinberg, M.S. 1970. Does differential adhesion govern self-assembly processes in histogenesis? Equilibrium configurations and the emergence of a hierarchy among populations of embryonic cells. *Journal of Experimental Zoology*, 173: 395–434

Thompson, D.W. 1992. *On Growth and Form*, Cambridge and New York: Cambridge University Press

Turing, A.M. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London* B, 327: 37–72

Weliky, M., Minsuk, S., Keller, R. & Oster, G.F. 1991. Notochord morphogenesis in *Xenopus laevis*: simulation of cell behaviour underlying tissue convergence and extension. *Development*, 113: 1231–1244

MULTIDIMENSIONAL SOLITONS

Although strict analogs of the Korteweg–de Vries soliton (exponentially localized solution with a specific relation between velocity and amplitude and particular scattering properties) have not been found in the multidimensional context, solvable equations with three or more independent variables exhibit a large variety of soliton-like solutions.

As with the Kadomtsev–Petviashvili equation, wide classes of exact explicit solutions have been constructed for other $(2+1)$-dimensional nonlinear equations solvable by the inverse scattering method. We consider here two basic examples, the first being the Davey–Stewartson (DS) equation

$$\begin{aligned} \mathrm{i}q_t + \tfrac{1}{2}\left(\sigma^2 q_{xx} + q_{yy}\right) + |q|^2 q - q\phi &= 0\,, \\ \phi_{xx} - \sigma^2 \phi_{yy} &= 2\left(|q|^2\right)_{xx}\,, \end{aligned} \qquad (1)$$

where $q(x, y, t)$ is a complex-valued function, ϕ is a real-valued function, and the parameter σ^2 takes two values $\sigma^2 = \pm 1$. The DS equation describes propagation of a two-dimensional long surface wave on water of finite depth. In the one-dimensional limit $q_y = \phi_y = 0$, it reduces to the nonlinear Schrödinger equation.

The DS equation (1) has a Lax representation with the two-dimensional Dirac operator as the Lax operator, but it has quite different properties for $\sigma^2 = 1$ (DS-I equation) and for $\sigma^2 = -1$ (DS-II equation). In both cases, there are multi-soliton solutions which do not decay in certain directions on the x, y plane. Similar to the Kadomtsev–Petviashvili equation these solutions describe elastic scattering of line solitons that decay exponentially in the direction of propagation and do not decay in the orthogonal direction. The phase shift can be explicitly calculated.

In addition, the DS equation possesses novel classes of solutions. Thus, the DS-II equation has an infinite set of nonsingular exponential-algebraic solutions, the simplest of which looks like

$$q\,(x, y, t) = \frac{2\nu \exp\left[\lambda(x + \mathrm{i}y) - \bar{\lambda}\,(x - \mathrm{i}y) - \mathrm{i}\left(\lambda^2 + \bar{\lambda}^2\right)t\right]}{|x + \mathrm{i}y + \mu - 2\mathrm{i}\lambda t|^2 + |\nu|^2}, \qquad (2)$$

where λ, μ, and ν are arbitrary complex constants. It decays like $\left(x^2 + y^2\right)^{-1}$ as $x, y \to \infty$.

The DS-I equation also possesses solutions for which q decays exponentially in both space dimensions. The simplest of them is of the form

$$q\,(x, y, t) = \frac{4\rho\sqrt{\lambda\mu}\exp\left[\mu\,(x+y) + \lambda\,(x-y) + \mathrm{i}\left(\mu^2+\lambda^2\right)t\right]}{\left[1 + \mathrm{e}^{2\mu(x+y)}\right]\left[1 + \mathrm{e}^{2\lambda(x-y)}\right] + |\rho|^2}, \qquad (3)$$

where λ, μ are arbitrary real parameters and ρ is an arbitrary complex parameter. The function ϕ has the nontrivial boundary values as $x, y \to \infty$. Called dromions, such solutions exhibit not only a two-dimensional phase shift during interaction but also a change of the form. Basically, these solutions are driven by the boundary conditions on the function ϕ.

Our second example—the Ishimori equation—is of the form

$$\begin{aligned} \boldsymbol{S}_t + \boldsymbol{S} \times \left(\boldsymbol{S}_{xx} + \sigma^2 \boldsymbol{S}_{yy}\right) + \phi_x \boldsymbol{S}_y + \phi_y \boldsymbol{S}_x &= 0\,, \\ \phi_{xx} - \sigma^2 \phi_{yy} + 2\sigma^2 \boldsymbol{S} \cdot \left(\boldsymbol{S}_x \times \boldsymbol{S}_y\right) &= 0, \end{aligned} \qquad (4)$$

where $\boldsymbol{S} = (S_1, S_2, S_3)$ is a unit vector $\boldsymbol{S}^2 = 1$, $\sigma^2 = \pm 1$, and ϕ is a scalar real-valued function. It represents an integrable $(2+1)$-dimensional generalization of the Heisenberg Ferromagnet model equation $\boldsymbol{S} = \boldsymbol{S} \times \boldsymbol{S}_{xx}$. An important feature of the Ishimori equation is that its solutions can be characterized by the topological invariant

$Q = (1/4\pi)\int\int \mathrm{d}x\,\mathrm{d}y\,\boldsymbol{S}\cdot\left(\boldsymbol{S}_x \times \boldsymbol{S}_y\right)$, which is conserved in time and takes integer values $N = 0, \pm 1, \pm 2, \ldots$.

The Ishimori equation for both signs of σ^2 possesses multisoliton solutions that describe elastic scattering of line solitons. The Ishimori-I equation ($\sigma^2 = 1$) has a rich variety of the dromion-like solutions. An interesting feature is that the Ishimori-II equation ($\sigma^2 = -1$) consists of the existence of the topologically nontrivial multi-vortex solutions. The one-vortex solution looks like

$$S_1 + \mathrm{i}S_2 = \frac{2\alpha\rho \mathrm{e}^{-\mathrm{i}\varphi}}{|\alpha|^2 + |\rho|^2}, \qquad S_3 = \frac{|\alpha|^2 - |\rho|^2}{|\alpha|^2 + |\rho|^2}, \quad (5)$$

where $\rho \mathrm{e}^{\mathrm{i}\varphi} = x + \mathrm{i}y - (x_0 + \mathrm{i}y_0)$, α is an arbitrary complex constant, and $Q = 1$. An anti-vortex solution with $Q = -1$ is also given by (5) under the substitution $x + \mathrm{i}y \to x - \mathrm{i}y$, that is, $\varphi \to -\varphi$. General N-vortex solution has $Q = N$ and describes the time dynamics of the spin-vortices and spin-anti-vortices of the form (5) which exhibits no phase shifts. These solutions are all regular and rational functions on x and y; in other words, they are lump solutions of the Ishimori-II equation.

Both the Davey–Stewartson and the Ishimori equations also possess infinite sets of exact solutions parameterized by finite number of arbitrary functions of one variable. As for the NLS equation and the Heisenberg ferromagnet equation, the DS equation and the Ishimori equation are related by the so-called gauge transformation.

Multidimensional nonlinear equations not solvable by the inverse scattering method do not possess multisoliton-like solutions. However, some of them, such as lumps, breathers, or spherical waves, have particular solutions which are similar to one soliton solution in some respects. Such solutions play roles in various nonlinear phenomena arising in classical and quantum field theories, in nonlinear theories of matter, and even in the theory of Jupiter's Great Red Spot.

BORIS B. KONOPELCHENKO

See also **Dressing method; Inverse scattering method or transform; Kadomtsev–Petviashvili equation**

Further Reading

Ablowitz, M.J. & Clarkson, P.A. 1991. *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, Cambridge and New York: Cambridge University Press

Boiti, M. et al. 1995. Multidimensional localized solitons. *Chaos, Solitons and Fractals*, 12: 2377–2417

Coleman, S. 1977. Classical lumps and their quantum descendants. In *New Phenomena in Sub-Nuclear Physics*, edited by A. Zichichi, New York: Plenum Press

Konopelchenko, B. 1993. *Solitons in Multidimensions. Inverse Spectral Transform Method*, Singapore: World Scientific

Rajaraman, R. 1982. *Solitons and Instantons*, Amsterdam: North-Holland

MULTIFRACTAL ANALYSIS

A fractal is a geometrical shape that shows self-similarity, meaning that parts appear similar to the set as a whole (*See* **Fractals**). A multifractal is a fractal with a probability measure attributed to its geometrical support set. A typical multifractal can have different fractal dimensions in different parts of its support set, depending on the multifractal measure chosen. Typical examples of multifractals are attractors of nonlinear mappings, where the relevant probability measure is given by the invariant measure of the map. Often, quite a complicated structure is observed, and the invariant density may diverge at infinitely many points in the phase space with different exponents, that is, an entire spectrum of singularities is generated.

To statistically analyze this complex behavior, it is useful to evaluate the so-called Rényi dimensions D_q (Rényi, 1970). These are generalizations of the usual box dimensions (or capacity) that contain information not only on the topological structure of the fractal but on the probability measure as well. The simplest definition of the Rényi dimensions is as follows. Cover the fractal with small d-dimensional cubes ("boxes") of volume ε^d, where ε is the side length of the box and d is an integer dimension large enough to embed the fractal. For each such box i we consider the probability p_i that is associated with it:

$$p_i = \int_{i\text{th box}} \rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \qquad (1)$$

Here $\rho(\boldsymbol{x})$ is the probability density considered. The Rényi dimensions are defined for any real parameter q as

$$D_q = \lim_{\varepsilon \to 0} \frac{1}{\log \varepsilon} \frac{1}{q-1} \log \sum_i p_i^q, \qquad (2)$$

where the sum is over all i with $p_i \neq 0$. There are also more sophisticated definitions of the Rényi dimensions based on boxes of variable size (analogous to the definition of the Hausdorff dimension), see Beck & Schlögl (1993) for more details.

Generally, for a complicated multifractal (with lots of singularities of the probability density) the Rényi dimensions D_q depend on q, whereas for "simple" multifractals all D_q are the same and equal to the Hausdorff dimension. By changing the parameter q, one "scans" the structure of the multifractal. Large q values give more weight to large probabilities p_i, small q favor small probabilities p_i. Useful special cases of the Rényi dimensions are the box dimension (or capacity) D_0 (usually equal to the Hausdorff dimension, up to pathological cases), the information dimension D_1 (more precisely given by the limit $\lim_{q \to 1} D_q$), and the correlation dimension D_2. The correlation dimension can be easily extracted from experimental time series using the Grassberger–Procaccia algorithm (Grassberger & Procaccia, 1983). It is also of relevance

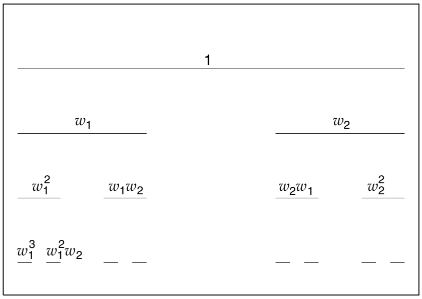

Figure 1. Construction of a classical Cantor set with a multiplicative non-uniform measure.

for the estimation of typical period lengths that are generated due to computer roundoff errors in chaotic dynamical systems (Beck, 1989). Also important are the limit dimensions $D_{\pm\infty}$ obtained for $q \to \pm\infty$. They describe the scaling behavior of the invariant density in the region of the phase space where the measure is most concentrated (D_∞) and least concentrated ($D_{-\infty}$).

From the Rényi dimensions, one can proceed to $f(\alpha)$, the spectrum of local scaling exponents α of the measure, by a Legendre transformation. The basic idea is very similar to thermodynamics. In fact, many of the ideas of the multifractal formalism are related to early work by Sinai, Ruelle, and Bowen on the so-called "thermodynamic formalism" of dynamical systems (Ruelle, 1978). For multifractals, one can regard the function $\tau(q) = (q-1)D_q$ as a kind of free energy, and q as a kind of inverse temperature (Tél, 1988; Beck & Schlögl, 1993). One then defines the variable α (the "internal energy") by $\partial\tau/\partial q =: \alpha$ and proceeds to the $f(\alpha)$ spectrum (the "entropy") by Legendre transformation:

$$f(\alpha) = q\alpha - \tau(q). \tag{3}$$

The advantage of the $f(\alpha)$ spectrum is that it has a kind of "physical meaning." Roughly speaking, it is the fractal dimension of the subset of points for which the probability density scales with a local exponent α, that is, $p_i \sim \varepsilon^{\alpha_i}$ with $\alpha_i \in [\alpha, \alpha + \mathrm{d}\alpha]$. Hence, this is a kind of statistical mechanics of local Hölder indices.

Let us consider a simple example of a multifractal, given by the classical Cantor set with a multiplicative (non-uniform) measure (Figure 1). The classical Cantor set is constructed by cutting the middle third out of the unit interval, then cutting out the middle third out of the remaining two intervals, and so on (*See* **Fractals**). In this way, at the kth construction step there are 2^k intervals of length $(\frac{1}{3})^k$. We now attribute a product measure to each of the intervals, according to the rule sketched in Figure 1, with $w_1 + w_2 = 1$ but $w_1 \neq w_2$. In the limit $k \to \infty$ we obtain a multifractal.

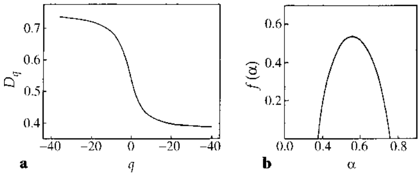

Figure 2. (a) Rényi dimensions and (b) $f(\alpha)$ spectrum of the Feigenbaum attractor.

Let us calculate the corresponding Rényi dimensions. We cover the multifractal with small intervals of size $\varepsilon_k = (\frac{1}{3})^k$. The number of boxes with probability $p_i = w_1^j w_2^{k-j}$ is

$$N_{k,j} = \begin{pmatrix} j \\ k \end{pmatrix} = \frac{j!}{k!(k-j)!}. \tag{4}$$

Hence,

$$\sum_i p_i^q = \sum_{j=0}^{k} N_{k,j} w_1^{jq} w_2^{(k-j)q} = (w_1^q + w_2^q)^k \tag{5}$$

and for the Rényi dimensions, one obtains from definition (2)

$$D_q = \frac{\log(w_1^q + w_2^q)}{(1-q)\log 3}. \tag{6}$$

In particular, the box dimension D_0 is given by the value $D_0 = \log 2/\log 3$, independent of w_1 and w_2. The other dimensions depend on the probabilities w_j, for example, for the information dimension one obtains $D_1 = -(1/\log 3)\sum_{j=1}^{2} w_j \log w_j$ by considering the limit $q \to 1$. The Legendre transform of (6) yields the $f(\alpha)$ spectrum.

Another interesting example is the attractor of the logistic map at the critical point of period doubling accumulation. Figure 2a shows the corresponding Rényi dimensions and Figure 2b the corresponding $f(\alpha)$ spectrum. Generally, the value of α where the function $f(\alpha)$ has its maximum is equal to the Hausdorff dimension of the attractor. The value of $f(\alpha)$ where the function has slope 1 is equal to the information dimension.

In practice, one often wants to evaluate the Rényi dimensions (or the $f(\alpha)$ spectrum) for a given time series (Kantz & Schreiber, 1997) of experimental data without explicitly knowing the underlying dynamics or the invariant measure. Here, various interesting methods are known (Grassberger–Procaccia algorithm (Grassberger & Procaccia, 1983), wavelet analysis (Arneodo et al., 1995), etc.). These algorithms can be implemented without explicitly knowing the underlying dynamics. The wavelet transform of an experimental signal $s(x)$ is defined as

$$W_\Psi(x_0, a) = \frac{1}{a}\int_{-\infty}^{+\infty} \Psi^*\left(\frac{x - x_0}{a}\right) s(x)\,\mathrm{d}x, \quad (7)$$

($*$ indicates complex conjugate), where the analyzing wavelet Ψ is some localized function, often chosen to be the Nth derivative of a Gaussian function. For small a, the wavelet transform extracts local Hölder exponents from the signal s, and qth moments of W_Ψ can then be used to define suitable partition functions which yield the Rényi dimensions.

CHRISTIAN BECK

See also **Dimensions; Fractals; Free energy; Measures; Sinai–Ruelle–Bowen measures; Wavelets**

Further Reading

Arneodo, A., Bacry, E. & Muzy, J.F. 1995. The thermodynamics of fractals revisited with wavelets. *Physica* A, 213: 232–275

Beck, C. 1989. Scaling behavior of random maps. *Physics Letters* A, 136: 121–125

Beck, C. & Schlögl, F. 1993. *Thermodynamics of Chaotic Systems*, Cambridge and New York: Cambridge University Press

Grassberger, P. & Procaccia, I. 1983. Characterization of strange attractors. *Physical Review Letters*, 50: 346–349

Kantz, H. & Schreiber, T. 1997. *Nonlinear Time Series Analysis*, Cambridge and New York: Cambridge University Press

Rényi, A. 1970. *Probability Theory*, Amsterdam: North-Holland

Ruelle, D. 1978. *Thermodynamic Formalism*, Reading, MA: Addison-Wesley

Tél, T. 1988. Fractals, multifractals and thermodynamics – an introductory review. *Zeitschrift für Naturforschung* A, 43: 1154–1174

MULTIFRACTAL MEASURE

See **Dimensions**

MULTIPLE SCALE ANALYSIS

For a number of problems involving differential equations, we know methods that can provide exact solutions. However, the vast majority of modeling problems have a complexity that forbids finding an exact solution by paper and pencil. It may appear that the only way forward is to use numerical analysis. However, this is not necessarily the case since we can resort to finding good approximate solutions by various methods or strategies. In applications we usually encounter systems with dissipation and energy input, where neglecting these energy exchange terms leads to an exact solvable problem. We can then employ a strategy based on the assumption that the exact solution of the unperturbed problem is only modified slightly by adding the perturbation terms and hope that the difference between the two solutions can be estimated sufficiently accurately by assuming the perturbation is weak (Nayfeh, 1973, 1981, 2000).

Let us illustrate such a strategy by considering the damped harmonic oscillator. Denoting the displacement from equilibrium by $x(t)$ at time t, the dynamical equation reads

$$\ddot{x} + x = -\varepsilon\dot{x}. \quad (1)$$

A dot above the dependent variable $x(t)$ denotes differentiation with respect to time t. The damping parameter ε we shall vary, and accordingly it is natural to include ε in the argument list of $x = x(t; \varepsilon)$. The unperturbed oscillator corresponds to $\varepsilon = 0$, and using the initial conditions $x(0) = 1$ and $\dot{x}(0) = 0$, the solution of the unperturbed harmonic oscillator reads

$$x(t; 0) = x_0(t) = A\cos(t), \quad (2)$$

where the amplitude A equals unity. For the same initial conditions, the damped harmonic oscillator (1) with $\varepsilon \neq 0$ possesses the more complicated solution

$$x(t;\varepsilon) = \mathrm{e}^{-\varepsilon t/2}\left[\cos\left(t\sqrt{1-\varepsilon^2/4}\right) + \frac{\varepsilon/2}{\sqrt{1-\varepsilon^2/4}}\sin\left(t\sqrt{1-\varepsilon^2/4}\right)\right]. \quad (3)$$

The exact solution has features which can be used as a guide for developing a systematic way of finding an approximate solution assuming small damping, that is, $|\varepsilon| \ll 1$.

Without surprise, we recognize that the damping leads to a decreasing amplitude as time progresses but on a slow time scale of order $O(\varepsilon)$. Another result of the damping effect is an $O(\varepsilon^2)$ change of the frequency, which is $\omega = \sqrt{1-\varepsilon^2/4} \approx 1 - \varepsilon^2/8$. The damping slows down the oscillation, and accordingly the period increases, hence the frequency decreases. Suitable corrections are therefore to allow for a slow variation of the amplitude A and for an even smaller correction of the frequency. In general terms, this implies that we need to let $x(t;\varepsilon)$ depend on the slower time scales εt and $\varepsilon^2 t$. We can formalize this by introducing (Nayfeh, 1981)

$$T_0 = t\,, \quad T_1 = \varepsilon t\,, \quad T_2 = \varepsilon^2 t, \\ x(t) = x(T_0, T_1, T_2). \quad (4)$$

Thus, the time derivatives expand as

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial}{\partial T_0} + \varepsilon\frac{\partial}{\partial T_1} + \varepsilon^2\frac{\partial}{\partial T_2} + O(\varepsilon^3), \quad (5)$$

$$\frac{d^2}{dt^2} = \frac{\partial^2}{\partial T_0^2} + 2\varepsilon\frac{\partial^2}{\partial T_0\partial T_1} + \varepsilon^2\left(\frac{\partial^2}{\partial T_1^2} + 2\frac{\partial^2}{\partial T_0\partial T_2}\right) + O(\varepsilon^3), \quad (6)$$

by invoking the chain role for partial derivatives. The idea is to introduce slowly time-varying coefficients by regarding the new scaling variables in (4) as independent variables. This is a somewhat mysterious trick, and it is not easy to understand why it works, since the scaling variables are all proportional to time, but it does. In perturbation theory, we often rely on such experience and tricks.

After having introduced the scaling into our original problem, we can add further corrections to the unperturbed solution by introducing a Taylor expansion around $\varepsilon = 0$ with respect to the perturbation parameter ε

$$x(\boldsymbol{T};\varepsilon) = x_0(\boldsymbol{T}) + x_1(\boldsymbol{T})\varepsilon + x_2(\boldsymbol{T})\varepsilon^2 + O(\varepsilon^3), \quad (7)$$

where

$$\mathbf{T} = (T_0, T_1, T_2) \quad \text{and} \quad x_n(\mathbf{T}) = \frac{1}{n!}\frac{\partial^n x(\mathbf{T};0)}{\partial\varepsilon^n}.$$

Inserting the Taylor expansion into Equation (1) results in a polynomial in ε which equals zero. The coefficient of ε^n is a differential operator working on x_n, and as all these coefficients must vanish, we have a differential equation for each x_n, which normally is simple enough that it can be solved analytically, at least for the first few equations. Inserting the Taylor expansion into Equation (1) and ordering according to powers of ε we obtain

$$\begin{aligned} x_{0,T_0T_0} + x_0 &= 0, & (8)\\ x_{1,T_0T_0} + x_1 &= -x_{0,T_0} - 2x_{0,T_0T_1}, & (9)\\ x_{2,T_0T_0} + x_2 &= -x_{1,T_0} - 2x_{1,T_0T_1} - x_{0,T_1} \\ &\quad -x_{0,T_1T_1} - 2x_{0,T_0T_2}. & (10)\end{aligned}$$

Subscript T_0 means a partial derivative with respect to T_0 and similarly for the analogous subscripts. The solution of the first equation reads

$$x_0 = A(T_1, T_2)e^{iT_0} + B(T_1, T_2)e^{-iT_0}. \quad (11)$$

The fact that A and B depend on T_1 and T_2 expresses the slow variation of these parameters with time. The solution of the equation for x_1 can now be found by inserting x_0 from (11) into the right-hand side of Equation (9),

$$\begin{aligned} x_{1,T_0T_0} + x_1 &= -i(A + 2A_{T_1})e^{iT_0} \\ &\quad +i(B + 2B_{T_1})e^{-iT_0}. \quad (12)\end{aligned}$$

Terms proportional to the exponentials $e^{\pm iT_0}$ are resonant forcing terms, leading to a growth of x_1 proportional to time t. When t is of order $1/\varepsilon$, the term $\varepsilon x_1(t)$ has become of order unity, and our analysis breaks down. Such terms are called secular terms. In order to get a uniform and valid expansion, we need to avoid secular terms. However, due to the dependence on T_1 of A and B, we can demand that coefficients proportional to $e^{\pm iT_0}$ vanish thereby providing simple differential equations for determining A

$$A + 2A_{T_1} = 0 \quad \Rightarrow \quad A(T_1, T_2) = A_1(T_2)e^{-T_1/2}, \quad (13)$$

and similarly for B. Equation (12) now possesses the solution $x_1 = 0$. We can proceed by solving the $O(\varepsilon^2)$ equation (10) and invoke the initial conditions $x(0) = 1$ and $\dot{x}(0) = 0$. After rather straightforward calculations, we obtain the approximate solution (Scott, 2003)

$$\begin{aligned} x(t) &= e^{-\varepsilon t/2}\cos[(1 - \varepsilon^2/8)t] + \frac{\varepsilon}{2}e^{-\varepsilon t/2} \\ &\quad \times \sin[(1 - \varepsilon^2/8)t] + O(\varepsilon^3) \quad (14)\end{aligned}$$

in agreement with the exact solution to the desired order in ε. Carrying out the analysis to order $O(\varepsilon^n)$ requires introduction of scaled variables up to the same order for the perturbation analysis to be consistent. The above approach is called the method of multiple scales, and it has been applied to many nonlinear ordinary differential equations as well as perturbed soliton equations (McLaughlin & Scott, 1978; Scott, 2003), including derivations of nonlinear partial differential equations (Dodd, 1984). An interesting extension of the method to nonlinear difference equations is presented in Broomhead & Rowlands (1983).

MADS PETER SØRENSEN

See also **Averaging methods; Damped-driven anharmonic oscillator; Fredholm theorem; Multi-soliton perturbation theory; Perturbation theory**

Further Reading

Broomhead, D.S. & Rowlands, G. 1983. On the analytic treatment of non-integrable difference equations. *Journal of Physics* A, 16: 9–24

Dodd, R.K., Eilbeck, J.C., Gibbon, J.D. & Morris, H.C. 1984. *Solitons and Nonlinear Wave Equations*, London: Academic Press

McLaughlin, D.W. & Scott, A.C. 1978. Perturbation analysis of fluxon dynamics. *Physical Review* A, 18: 1652–1680

Nayfeh, A.H. 1973. *Perturbation Methods*, New York: Wiley-Interscience

Nayfeh, A.H. 1981. *Introduction to Perturbation Techniques*, New York: Wiley-Interscience

Nayfeh, A.H. 2000. *Nonlinear Interactions: Analytical, Computational, and Experimental Methods*, New York: Wiley-Interscience

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

MULTIPLEX NEURON

Under the view that dominated neuroscience around the middle of the 20th century, neural information processing involves three components (McCulloch & Pitts, 1943). First, incoming signals from chemical synapses on the dendritic trees are gathered as a linear, weighted sum of input signals. Second, this sum is compared with a threshold level at the base (initial segment) of the outgoing axonal tree. Finally, if the threshold level is exceeded, an active (all-or-nothing) nerve impulse is launched on the axonal tree, which propagates outward without failure to all of

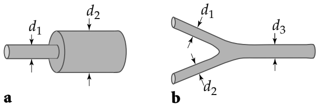

Figure 1. (a) Abrupt widening of a nerve fiber. (b) Branching region.

Membrane model	Transmission	Blockage
H–H	5.5084	5.5126
M–L	2.2521	2.2563

Table 1. Widening ratios at which isolated impulses transmit and become blocked under Hodgkin–Huxley (H–H) and Morris–Lecar (M–L) membrane models (Altenberger et al., 2001).

the distal (distant) twigs of the tree, providing inputs to other neurons or to muscle cells. A simplifying feature of this “McCulloch–Pitts model” is that the incoming dynamics are entirely linear, governed by inhomogeneous diffusion processes on the dendritic trees.

Before the end of the 1960s, there was no compelling evidence to abandon this simple view of dendritic dynamics, and two reasons for clinging to it. If all-or-nothing propagation is supposed to occur on dendrites, then the entire tree might be expected to ignite, preventing the dendrites from integrating incoming information. Additionally, the assumption of linear dendritic dynamics helps the theorist to sort out various causal influences, somewhat easing the difficulties of analyzing neural systems (Scott, 2002).

By the 1970s, three types of evidence began to indicate that dendritic and axonal dynamics are more complex. First, impulse blockage was experimentally observed in the optic nerves of cats, implying that impulses (spikes) can be extinguished at axonal branchings (Chung et al., 1970). Second, experimental studies showed that spikes do indeed propagate on the dendritic trees of some vertebrate neurons (Llinás & Nicholson, 1971). Third, numerical studies by Boris Khodorov and his colleagues in the Soviet Union on realistic models of nerve fibers confirmed that the propagation of active nerve impulses can indeed be blocked at changes in fiber geometry such as the abrupt widening and branching regions shown in Figure 1 (Khodorov, 1974).

These considerations suggested the concept of a “multiplex neuron,” where the term (borrowed from communication engineering) implies the ability to handle two or more messages at the same time. As defined by Steven Waxman, a multiplex neuron has the following salient properties (Waxman, 1972).

- Impulse blockage at the branching regions of dendritic trees allows the possibility of Boolean logic, similar to the elementary operations of a digital computer. If the geometry of the branch is such that an impulse incoming on either daughter branch (one “or” the other) is able to ignite an outgoing impulse on the parent branch, this would be OR logic. If, on the other hand, coincident impulses on both incoming branches (one “and” the other) are required to ignite an outgoing impulse, it would be an example of AND logic. Thus the computational nature of a branching region may depend upon details of its geometry.
- Impulse blockage at branching regions of the axonal tree allows an impulse code transformation under which a time code on the trunk of the tree is transformed to space-time codes at the distal branches (twigs) of the tree.

Seminal Soviet studies of the propagation of impulses through the abrupt widening shown in Figure 1(a) showed that blockage of an isolated Hodgkin–Huxley impulse (Hodgkin & Huxley, 1952) (which corresponds to the dynamics of a squid giant axon) was to be expected at a widening ratio (d_2/d_1) greater than about 5.5. As indicated in Table 1, recent numerical studies by Altenberger et al. (2001) have confirmed the early Soviet results for the Hodgkin–Huxley model and extended them to the Morris–Lecar model (which represents dynamics on barnacle giant muscle fibers and more closely models the calcium dominated dynamics of typical dendritic fibers) (Morris & Lecar, 1981).

Although abrupt widenings of real nerve fibers are not observed, numerical blocking conditions in Figure 1(a) can be related to the corresponding blocking conditions in Figure 1(b) through the concept of a “geometric ratio” (GR) (Goldstein & Rall, 1974). Assuming a discontinuity with several outgoing fibers (of diameters d_{out}) and on which an impulse is incoming on a single fiber of diameter d_{in}, the GR is defined as $\sum d_{\text{out}}^{3/2}/d_{\text{in}}^{3/2}$, where the “3/2 powers” enter because branching currents divide in proportion to the characteristic admittance of the outgoing fibers (Scott, 2002).

Assuming the Morris–Lecar (M–L) model and an incoming impulse on daughter branch # 1 in the branch of Figure 1(b), the AND condition is

$$\frac{d_2^{3/2}+d_3^{3/2}}{d_1^{3/2}} > 2.2563^{3/2} = 3.389, \tag{1}$$

requiring coincident inputs on both incoming branches (# 1 and # 2) for an impulse on the outgoing branch (# 3) to ignite. If this geometric inequality is not satisfied, the branch executes OR logic, in which incoming impulse either on daughter # 1 or on daughter # 2 is able to ignite an outgoing impulse on the parent branch (# 3).

Cell type	GR range
Purkinje	2.3–3.5
Stellate	2.4–3.5
Granule	2.3–4.8
Motoneuron	2.6–3.4
Pyramidal (apical)	2.5–3.4
Pyramidal (basal)	2.4–3.1

Table 2. The GR range for some typical dendrites (Scott, 2002).

Noting the GR values for typical dendritic trees given in Table 2 and considering the several ways in which critical GRs might be lowered in real neurons (changes in ionic concentrations, variations in channel membrane density, incomplete impulse recovery from previous interactions, and so on), it seems prudent to anticipate that dendritic trees can execute logical operations on the incoming (synaptic) codes, as is assumed in the multiplex neuron model (Stuart et al., 1999).

ALWYN SCOTT

See also **McCulloch–Pitts network; Nerve impulses; Neurons**

Further Reading

Altenberger, R., Lindsay, K.A., Ogden, J.M. & Rosenberg, J.R. 2001. The interaction between membrane kinetics and membrane geometry in the transmission of action potentials in non-uniform excitable fibres: a finite element approach. *Journal of Neuroscience Methods*, 112: 101–117

Chung, S.H., Raymond, S.A. & Lettvin, J.Y. 1970. Multiple meaning in single visual units. *Brain, Behavior and Evolution*, 3: 72–101

Goldstein, S.S. & Rall, W. 1974. Changes of action potential, shape and velocity for changing core conductor geometry. *Biophysical Journal*, 14: 731–757

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117: 500–544

Khodorov, B.I. 1974. *The Problem of Excitability*, New York: Plenum

Llinás, R. & Nicholson, C. 1971. Electrophysiological properties of dendrites and somata in alligator Purkinje cells. *Journal of Neurophysiology*, 34: 532–551

McCulloch, W.S. & Pitts, W.H. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5: 115–133

Morris, C. & Lecar, H. 1981. Voltage oscillations in the barnacle giant muscle fibre. *Biophysical Journal*, 35: 198–213

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Stuart, G., Spruston, N. & Häusser, M. (editors). 1999. *Dendrites*, Oxford and New York: Oxford University Press

Waxman, S.G. 1972. Regional differentiation of the axon, a review with special reference to the concept of the multiplex neuron. *Brain Research*, 47: 269–288

MULTIPLICATIVE NOISE

See **Stochastic processes**

MULTIPLICATIVE PROCESSES

Random processes involving accumulation and/or reduction occur widely in nature and in mathematics. At the most basic level, the laws of accumulation (and reduction) may be represented either as the result of arithmetic summation (and differencing) or as the result of multiplications. The latter such processes are the subject of this article.

Any attempt to embody all multiplicative random processes in a single mathematical definition would be futile. However, it is possible to provide large classes of examples for which rich mathematical theories and applications are possible.

The "law of proportionate effect" provides the basis for the evolution of multiplicative stochastic processes $M_0, M_1, \ldots,$ found, for example, in materials science, biology, economics, and finance, in terms of the proportionate changes $r_{n+1} = \Delta M_n / M_n \equiv (M_{n+1} - M_n)/M_n, n = 0, 1, 2, \ldots.$ In finance, M_n may be the value of a unit of stock at the nth period of time when subjected to the hypothesis that the yields $r_1, r_2, \ldots$ are stationary and independent from time period to period. In materials science, on the other hand, M_n may represent the strength of a material after n repeated independent random impacts $R_1 = 1 + r_1$, $R_2 = 1 + r_2, \ldots.$ Similarly, M_n may represent the size of a biological population whose subsequent growth is in (random) proportions $r_1, r_2, \ldots$ to the amount of substance available. The law of proportionate effect is a widely observed statistical law according to which the sequence of proportionate changes $r_{n+1} = \Delta M_n / M_n, n = 0, 1, 2, \ldots,$ is a sequence of independent and identically distributed (iid) random variables. Equivalently, one may iterate this law to obtain the so-called geometric random walk model:

$$M_n = M_0 \prod_{k=1}^{n} R_k, \quad k = 1, 2, \ldots, \tag{1}$$

where

$$R_k = 1 + r_k, \quad k = 1, 2, \ldots. \tag{2}$$

The special case in which R_n has two possible values, corresponding to upward and downward price movements, yields the binomial tree model for stock prices popular in the modern mathematical finance literature (Föllmer & Schied, 2002).

A more general notion of "multiplicative processes", whose origins may be traced back to early 19th century genetics, occurs by consideration of a somewhat finer scale growth rule. In particular, changes to the process occur as M_{n+1} evolves from a previous random number

M_n of constituent members, each of which individually contributes its own growth and decay amounts (Y) to the whole in successive time steps. This more general stochastic law of evolution may be precisely expressed by the following iteration scheme: If $M_n = 0$ then define $M_{n+1} = 0$, else if $M_n > 0$ define

$$M_{n+1} = \sum_{k=1}^{M_n} Y_{n+1,k} \quad n = 0, 1, \ldots, \tag{3}$$

where M_0 is a nonnegative integer valued (counting) random variable representing an initial size, and for each $n \geq 0$, the random variables $Y_{n+1,1}, Y_{n+1,2}, \ldots Y_{n+1,M_n}$ are nonnegative integer valued, representing the respective numbers of offspring of each of the M_n elements composing the nth generation.

Two special cases may be noted. In the case that the $Y_{n+1,k} \equiv R_{n+1},\ k \geq 1$, are the same for each k of the nth generation, but are independent and identically distributed (iid) from generation to generation n, Equation (3) is the law of proportionate effect. On the other hand, if for each n, the offspring sizes $Y_{n+1,1}, Y_{n+2,2}, \ldots$ are iid and independent of M_n, then Equation (3) defines the classical Bieneymé–Galton–Watson branching process (BGW) introduced in the early 19th century as a model for analyzing the survival of family names (Kendall, 1975).

A next generation of multiplicative processes arises out of these models by a still finer coding of the underlying phenomena. For simplicity, consider the BGW branching process and assume a single progenitor $M_0 = 1$. Then one may code the successive offspring as a random tree as follows: The single progenitor is coded by the empty sequence $\emptyset$. Its respective $Y_{1,1}$ offspring are coded (labeled) as $\langle 1 \rangle, \ldots, \langle Y_{1,1} \rangle$. The $Y_{2,k}$ next generation offspring of $\langle k \rangle$ can be coded as $\langle k1 \rangle, \langle k2 \rangle, \ldots, \langle kY_{2,k} \rangle$, and so on. In this way, the successive generations may be represented as a sequence of random family trees

$$\begin{aligned}\tau_0 = \{\emptyset\},\ \tau_1 &= \{\emptyset, \langle 1 \rangle, \ldots, \langle Y_{1,1} \rangle\} \\ &\equiv \tau_0 \cup \{\langle 1 \rangle, \ldots, \langle Y_{1,1} \rangle\}, \\ \tau_2 &= \tau_1 \cup \{\langle 11 \rangle, \ldots, \langle 1Y_{2,1} \rangle, \ldots, \\ &\qquad \langle Y_{1,1}1 \rangle, \ldots, \langle Y_{1,1}Y_{2,Y_{1,1}} \rangle\}, \ldots .\end{aligned}$$

This, in fact, suggests an alternative description of the BGW process as a probability distribution P on a (metric) space of tree graphs. Specifically, let $\boldsymbol{T}$ be the space of (possibly infinite) labeled tree graphs rooted at $\emptyset$. As above, an element τ of $\boldsymbol{T}$ may be coded as a set of finite sequences of positive integers $\langle i_1, i_2, \ldots, i_n \rangle \in \tau$ such that:

(i) The root $\emptyset \in \tau$ is coded as the empty sequence.
(ii) If $\langle i_1, \ldots, i_k \rangle \in \tau$ then $\langle i_1, \ldots i_j \rangle \in \tau$ $\forall\ 1 \leq j \leq k$.
(iii) If $\langle i_1, i_2, \ldots, i_n \langle \in \tau$ then $\rangle i_1, \ldots i_{n-1}, j \rangle \in \tau$ $\forall\ 1 \leq j \leq i_n$.

Neighboring vertices are those of the form $\langle i_1, \ldots, i_n \rangle$ and $\langle i_1, \ldots, i_{n-1} \rangle$. The last condition (iii) specifies a left to right labeling of vertices, and (ii) connects all paths to the root via neighboring vertices without cycles. The space $\boldsymbol{T}$ of such trees may be viewed as a metric space with metric defined by $\rho(\tau, \gamma) = (\sup\{n+1 : \gamma|n = \tau|n\})^{-1}$, and $\tau|n = \{\langle i_1, \ldots, i_k \rangle \in \tau\ :\ k \leq n\}$ is the truncation of τ to the first n generations. This metric is complete and the countable dense subset $\boldsymbol{T}_0$ of finite labeled tree graphs rooted at $\emptyset$ makes $\boldsymbol{T}$ a separable metric space as well.

The BGW model may now be viewed as a probability distribution on the space $\boldsymbol{T}$, wherein the probability assigned to the ball B of radius $1/N$ centered at τ is given by

$$P(B) = \prod_{v \in \tau|N} p(\omega(\tau; v)), \tag{4}$$

where $p(j) = P(Y = j)$, $j = 0, 1, 2, \ldots$ is a prescribed offspring probability distribution, and $\omega(\tau; v) = \max\{j\ :\ \langle v1 \rangle, \ldots, \langle vj \rangle \in \tau\}$ counts the number of offspring of the vertex v in the tree τ.

A number of naturally occurring geophysical and biological structures admit natural tree codings, for example, river networks, lightning discharge patterns, and arterial and neural networks in human organs. Probabilistic models correspond to probability distributions on $\boldsymbol{T}$, leading to further generalizations of multiplicatively branching models; for example, see Barndorff-Nielsen et al. (1998) and references therein.

Random multiplicative cascade models provide another class of multiplicative processes of interest for their intermittency, extreme variability, and multiscaling structure in applications to phenomena ranging from fluid turbulence and spatial oceanic rainfall distributions to internet packet data on the world wide web (Barndorff-Nielsen et al., 1998; Gilbert et al., 1999; Jouault et al., 2000). The origins of this class of models trace back to classic work of Lewis Fry Richardson (1926), Andrei N. Kolmogorov (1962), and A.M. Yaglom (1966) in statistical turbulence theory. A rich geometrical and scaling perspective was subsequently advanced by Benoit Mandelbrot, and a more complete mathematical treatment was initiated by J.-P. Kahane and Jacques Peyrière (Kahane, 1985).

In the context of turbulence, one imagines introducing energy into the fluid by a large-scale stirring motion, whereby smaller-scale eddies split off and dissipate energy in random proportions R to those available. These eddies in turn split off smaller-scale eddies, and so on. In the simplest mathematical formulation, one considers random measures $M_n(\Delta)$ on the one-dimensional unit interval [0, 1] having a piecewise constant density

$\rho_n(x)$ defined by a constant value over a dyadic subinterval

$$\Delta_n(j_1,\ldots,j_n) = \left[\sum_{k=1}^{n} j_k 2^{-k}, \sum_{k=1}^{n} j_k 2^{-k} + 2^{-n}\right],$$
$$j_k \in \{0,1\},\ k \geq 1 \tag{5}$$

with constant value

$$\rho_n(x) = \prod_{k=1}^{n} R_{\langle j_1 j_2 \ldots j_k\rangle},\ \forall x \in \Delta_n(j_1, j_2, \ldots, j_k), \tag{6}$$

where the random factors $R_v, v = \langle j_1 j_2 \ldots j_k \rangle$, are iid nonnegative mean one random variables indexed by the vertices of the binary tree. The mean one condition provides conservation of the mass of M_n on average, where for any Borel subset G

$$M_n(G) = \int_G \rho_n(x)\,\mathrm{d}x. \tag{7}$$

The multiplicative cascade M_∞ is the random measure obtained by passing to the fine scale limit as $n \to \infty$.

The random cascade is defined by the branching number b ($=2$ in this exposition), and the random factors R_v, referred to as cascade generators. Kolmogorov's "log-normal hypothesis" refers to a choice of lognormal distribution for R. Tests of this hypothesis and physical arguments for alternative laws present significant challenges for modern statistical turbulence theory (e.g., She & Waymire, 1995; Jouault et al., 2000). On the mathematical side, the special choice of symmetrically distributed 0–2 valued generators yields the BGW process with binomial offspring distribution $p(j) = \binom{2}{j}\frac{1}{4}$, $j = 0, 1, 2$, for the total masses $M_n([0,1]), n = 1, 2, \ldots$. Refinements for which the branching number b may be viewed as a random parameter are given in Burd & Waymire (2000), and as a continuous parameter in Barral & Mandelbrot (2002).

Branching random walk models of the type illustrated by this next application may also be viewed within the framework of random multiplicative cascades. An explicit representation of the Fourier transform of classes of solutions to 3-d incompressible Navier–Stokes equations

$$\frac{\partial u}{\partial t} + u \cdot \nabla u = \nu \Delta u - \nabla p + g, \quad \nabla \cdot u = 0, \tag{8}$$

in the form of an expected value of a certain product of initial and forcing data evaluated at the nodes of a branching random walk was recently uncovered (LeJan & Sznitman, 1997). While probability models have long enjoyed important connections to partial differential equations, most notable being the heat equation and reaction-diffusion equations, this ranks among the most striking recent connections between multiplicative cascades and nonlinear equations of fluid motion. For a simple illustration of ideas made possible by refinements of the theory developed in Bhattacharya et al. (2003), consider the one-dimensional Burgers equation

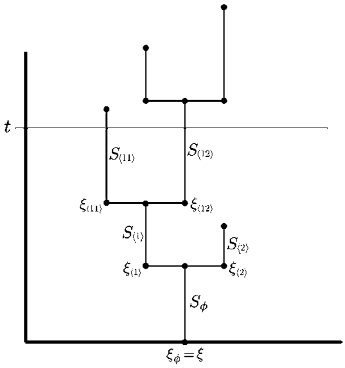

Figure 1. A sample tree graph for the Burgers equation.

$$u_t + uu_x = \nu u_{xx}, \quad u(0,x) = u_0(x). \tag{9}$$

Spatial Fourier transform will be denoted by $\hat{u}$. For simplicity of exposition, assume $\hat{u}_0(\xi) = 0$ for $\xi \leq 0$; that is, the initial data belongs to a Hardy function space. Taking spatial Fourier transforms one obtains, with the aid of an exponential integrating factor and $\xi > 0$ and writing

$$\lambda(\xi) = \nu\xi^2, \qquad m = -\frac{\mathrm{i}}{\nu}, \tag{10}$$

the result

$$\begin{aligned}\hat{u}(t,\xi) &= \mathrm{e}^{-\lambda(\xi)t}\hat{u}_0(\xi)\\ &\quad + \frac{m}{2}\int_0^t \lambda(\xi)\mathrm{e}^{-\lambda(\xi)s}\frac{1}{\xi}\\ &\quad \times \int_0^\xi \hat{u}(t-s,\eta)\hat{u}(t-s,\xi-\eta)\,\mathrm{d}\eta. \end{aligned}\tag{11}$$

Expressed in this form, Equation (11) takes on a probabilistic meaning. In particular, this is a recursive equation for the expected values of a multiplicative stochastic process initiated at $\xi_\emptyset = \xi$. The first term on the right-hand side $\mathrm{e}^{-\lambda(\xi)t}\hat{u}_0(\xi)$ is the product of the initial data $\hat{u}_0(\xi)$ times the probability $\mathrm{e}^{-\lambda(\xi)t}$ that an "exponentially distributed clock" with parameter $\lambda(\xi)$ rings after time t. The second integral term is an expected value in the complementary event of probability density $\lambda(\xi)\mathrm{e}^{-\lambda(\xi)s}$, that the clock rings at time s prior to t. Given that the clock rings at a time s prior to t, a product is formed with the factor $m(\xi)$ and a random selection of a pair of new "offspring" wave numbers (or Fourier frequencies) $\eta, \xi - \eta$ from the interval $[0, \xi]$ with (uniform) probability density $1/\xi$ to complete the recursion over the remaining time $t - s$. That is to say, the unique solution (in the appropriate function space) is furnished by the expected value

$$\hat{u}(t,\xi) = EX(t,\xi), \quad \xi > 0, \tag{12}$$

for a multiplicative cascade $X(t,\xi)$ defined by the following stochastic recursion in Fourier wave number space (see Figure 1): A particle of type $\xi_\emptyset = \xi$ waits for an exponentially distributed time $S_\emptyset$ with mean $1/\lambda(\xi) = \frac{1}{\nu\xi^2}$. If $S_\emptyset > t$ then a value $\hat{u}_0(\xi)$ is assigned to the initial vertex $\emptyset$ and the process terminates. On the other hand, if $S_\emptyset \le t$ then an independent coin flip is made. If the outcome is a tail then the particle dies and a value 0 is assigned, but if a head occurs then the particle branches into two particles $\langle 1 \rangle$, $\langle 2 \rangle$ of respective types $\xi_{\langle 1 \rangle} = \eta$ and $\xi_{\langle 2 \rangle} = \xi - \eta$ selected according to the uniform distribution on $[0, \xi]$. Two independent exponential clocks $S_{\langle 1 \rangle}, S_{\langle 2 \rangle}$, with respective parameters $\lambda(\xi_{\langle 1 \rangle}), \lambda(\xi_{\langle 2 \rangle})$, are set, and the process is repeated independently from each of these two given types for the termination time t reduced to $t - S_\emptyset$. The multiplicative cascade $X(t,\xi)$ is, up to a (random) power of $m \equiv -\sqrt{-1}/\nu$, a product of the assigned values of $\hat{u}_0$ at the selected frequencies; for example, for the sample realization depicted in Figure 1, one has

$$X(t,\xi) = m^2 \cdot \hat{u}_0(\xi_{\langle 11 \rangle}) \cdot \hat{u}_0(\xi_{\langle 12 \rangle}) \cdot \hat{u}_0(\xi_{\langle 2 \rangle}) \cdot 0. \tag{13}$$

In particular, the premature death of $\langle 2 \rangle$ means that this particular sample realization will not contribute a positive value to the mathematical expectation in (12).

By presenting these models as a progression of modifications and extensions built on simpler structures, some sense of a mathematical theory begins to emerge, a large part of which directly rests on martingale theory. For example, for the BGW model one may observe, denoting the mean offspring number by μ, assumed positive and finite, that $M_n/\mu^n, n = 1, 2, \ldots,$ is a martingale. Similarly, for each fixed Borel set G, the cascade measure $M_n(G)$, as a function of (logarithmic) scale $n = 1, 2, \ldots,$ is a martingale. The latter property led Kahane to still further natural generalizations of widespread significance (Kahane, 1985). Such deep mathematical structure has made it possible to precisely analyze many of the singularities, intermittencies, and other critical phenomena associated with these models, as well as to provide precise statistical error bars required for scientifically sound empirical estimations and tests of hypothesis for multiplicative cascades. On the other hand, the overall theory is in its relative infancy and many questions of practical importance remain open (Ossiander & Waymire, 2000). Needless to say, the relationship with the incompressible Navier–Stokes equations provides a link to one of the most outstanding mathematical problems of our times.

Edward C. Waymire

See also **Branching laws; Burgers equation; Martingales; Navier–Stokes equation**

Further Reading

Athreya, K. & Jagers, P. (editors). 1997. *Classical and Modern Branching Processes*, New York: Springer

Barndorff-Nielsen, O.E., Gupta, V.K., Perez-Abreu, V. & Waymire, E. (editors). 1998. *Stochastic Methods in Hydrology: Rain, Landforms and Floods*, Singapore: World Scientific

Barral, J. & Mandelbrot, B.B. 2002. Multifractal products of cylindrical pulses. *Probability Theory and Related Fields*, 124(3): 409–430

Bhattacharya, R., Chen, L., Dobson, S., Guenther, R., Orum, C., Ossiander, M., Thomann, E. & Waymire, E. 2003. Majorizing kernels & stochastic cascades with applications to incompressible Navier–Stokes equations. *Transactions of the American Mathematical Society*, 355: 5003–5040

Bramson, M. 1978. Maximal displacement of branching Brownian motion. *Communications in Pure and Applied Mathematics*, 31: 531–582

Burd, G. & Waymire, E. 2000. Self-similar invariance of critical binary Galton–Watson trees. *Proceedings of the American Mathematical Society*, 128: 2753–2761

Chen, L., Dobson, S., Guenther, R., Ossiander, M., Thomann, E. & Waymire, E. 2002. On Ito's complex measure condition, In *Probability, Statistics and Their Applications: Papers in Honor of Rabi Bhattacharya*, edited by K. Athreya, M. Majumdar, M. Puri & E. Waymire, Beachwood, OH: Institute of Mathematical Statistics

Föllmer, H. & Schied, A. 2002. *Stochastic Finance*, New York: Walter de Gruyter

Gilbert, A.C., Willinger, W. & Feldman, A. 1999. Scaling analysis of conservative cascades with applications to network traffic. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 45: 971–991

Holley, R. & Waymire, E. 1992. Multifractal dimensions and scaling exponents for strongly bounded random cascades. *Annals of Applied Probability*, 2(4), 819–845

Jouault, B., Greiner, M. & Lipa, P. 2000. Fix-point multiplier distributions in discrete turbulent cascade models. *Physica* D, 136: 196–255

Kahane, J.P. 1985. Sur le chaos multiplicatif. *Annales des Sciences Mathématiques du Québec*, 9: 105–150

Kahane, J.P. & Peyrière, J. 1976. Sur certaines martingales de B. Mandelbrot. *Advances in Mathematics*, 22: 131–145

Kendall, D. 1975. Branching processes since 1873, the genealogy of genealogy: branching processes before (and after) 1873. *Bulletin of the London Mathematics Society*, 7: 385–406, 225–253

Kolmogorov, A.N. 1962. A refinement of previous hypothesis concerning the local structure of turbulence in a viscous incompressible fluid at high Reynold number. *Journal of Fluid Mechanics*, 13: 82–85

LeJan, Y. & Sznitman, A.S. 1997. Stochastic cascades and 3-dimensional Navier–Stokes equations. *Probability Theory and Related Fields*, 109: 343–366

Mandelbrot, B.B. 1974. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *Journal of Fluid Mechanics*, 62: 331–333

McKean, H.P. 1975. Applications of Brownian motion to the equation of Kolmogorov, Petrovskii, and Piskunov. *Communications in Pure and Applied Mathematics*, 28: 323–331

Ossiander, M. & Waymire, E. 2000. Statistical estimation theory for multiplicative cascades. *Annals of Statistics*, 28(6): 1–21

Ossiander, M. & Waymire, E. 2002. On estimation theory for multiplicative cascades. *Sankhyā: The Indian Journal of Statistics Series* A, 64(2): 323–343

Resnick, S., Samorodnitsky, G., Gilbert, A. & Willinger, G. 2003. Wavelet analysis of conservative cascades. *Bernoulli*, 9(1): 97–135

Richardson, L.F. 1926. Atmospheric diffusion shown on a distance-neighbour graph. *Proceedings of the Royal Society of London*, A110, 709–737

She, Z.S. & Waymire, E. 1995. Quantized energy cascade and logPoisson statistics in fully developed turbulence. *Physical Review Lettters*, 74(2): 262–265

Waymire, E. & Williams, S.C. 1996. A cascade decomposition theory with applications to Markov and exchangeable cascades. *Transactions of the American Mathematical Society* 348(2): 585–632

Yaglom, A.M. 1966. Effect of fluctuations in energy dissipation rate on the form of turbulence characteristics in the inertial subrange. *Dokladay Akademii Nauk SSSR*, 166: 49–52

MULTISOLITON PERTURBATION THEORY

Solitons appear as robust solutions of several important nonlinear partial differential equations and difference-differential equations, including the Korteweg–de Vries (KdV), nonlinear Schrödinger (NLS), and sine-Gordon (SG) equations and the Toda lattice (TL). It is a remarkable feature of these and other equations, integrable by means of the inverse scattering transform (IST), that one can find exact multisoliton solutions. Among other phenomena, multisoliton solutions describe collisions among several solitons and bound states of solitons (breathers).

Considering systems that conserve energy but lack integrability, multisoliton solutions, and breathers may be only approximate, their dynamics accompanied by emission of radiation. As a result, two colliding solitons may merge into a breather, and the energy of a breather gradually decreases until it completely decays.

The situation is yet more different from that in integrable models if the system is dissipative. The dissipative loss of energy may be offset by an external field (driving force). A well-known example is the damped-driven SG equation

$$\phi_{tt} - \phi_{xx} + \sin\phi = -\alpha\phi_t - \gamma, \tag{1}$$

which models magnetic flux propagation on a long Josephson junction (LJJ) (McLaughlin & Scott, 1978). In this equation, $\partial\phi/\partial x$ is the local magnetic field, α is a coefficient of dissipation, and γ is a bias-current density, which is the driving force. In the absence of perturbations ($\alpha = \gamma = 0$), (1) is the SG equation, whose fundamental soliton solution is the kink (it represents a magnetic-flux quantum, or fluxon, in LJJ),

$$\phi_{\mathrm{k}} = 4\tan^{-1}\left[\exp\left(\sigma\frac{x-\xi(t)}{\sqrt{1-c^2}}\right)\right]. \tag{2}$$

Here, $\sigma = \pm 1, c$, and $\xi(t)$ are, respectively, the polarity, velocity, and central coordinate of the kink (or fluxon). An exact breather solution to the unperturbed SG equation is

$$\phi_{\mathrm{br}} = 4\tan^{-1}\left[\frac{\sin(t\cos\mu)}{\cosh(x\sin\mu)}\tan\mu\right], \tag{3}$$

where the amplitude μ takes values $0 < \mu < \pi/2$. In the limiting case $\mu \to \pi/2$, (3) becomes a solution describing collision between two kinks with opposite polarities.

In the presence of perturbations, Equation (1) still has kink-like solutions, but the kink's steady-state velocity (c_0) is no longer arbitrary. It is determined by the condition that the power input from a constant driving force (γ) is in balance with the dissipation induced by the loss term, which yields

$$c_0 = \frac{\sigma\pi\gamma}{\sqrt{(\pi\gamma)^2 + 16\alpha^2}}. \tag{4}$$

According to Equation (4), kinks with opposite polarities move in opposite directions; hence, they may collide.

If an ac driving force is applied to the system [e.g., $\gamma = \gamma_0\cos(\omega t)$ in (1)], it may compensate the loss and support the breather whose frequency $\cos\mu$ is in resonance with the driving frequency ω. If γ in (1) is a random function of time, and α is small enough, the random force can split the breather into a free kink-antikink (k$\bar{\mathrm{k}}$) pair.

In the general case, basic effects produced by perturbations acting on two- or multi-soliton configurations may be classified as follows. Interaction between two solitons in the presence of conservative perturbations gives rise to emission of radiation, which appears at order ε^2, where ε is the strength of the perturbation. Adiabatic effects (those that neglect radiation) may be generated by conservative perturbations in a *three-soliton* system at order ε, in the form of energy exchange between colliding solitons (this may also occur in a two-soliton system if the conservative perturbation is not spatially uniform—for instance, if it is created by a local defect). In the case of dissipative perturbations, nontrivial effects are possible at order ε for two solitons, typical examples being fusion of a k$\bar{\mathrm{k}}$ pair into a breather or, inversely, breakup of a breather into a pair due to its collision with another kink.

If the model is a perturbed version of an integrable one, multi-soliton processes similar to those outlined above can be investigated by means of perturbation theory (PT), a powerful version of which is based on a perturbed variant of IST (Kaup & Newell, 1978). As in the integrable case, the initial configuration is mapped into scattering data of the corresponding scattering problem, but the time evolution of the scattering data is no longer trivial, discrete eigenvalues being no more time-independent. Using a perturbative expansion, it is possible to derive ordinary differential equations (ODEs) for slow variations of the scattering data, which can be solved to obtain approximate solutions of the perturbed system. A comprehensive review of the IST-based PT for solitons in nearly integrable models was given by Kivshar & Malomed (1989).

An alternative multi-soliton PT is based on using a Green function (GF) for the underlying equation linearized around the unperturbed multi-soliton solution (Keener & McLaughlin, 1977). If the zero-order approximation is integrable, this method is equivalent to the IST-based PT, as the GF can be constructed—by means of IST—around any exact multi-solution solution. Although this approach has the advantage of being directly formulated in terms of physical parameters (soliton speeds, collision delays, breather frequencies, etc.), rather than more abstract IST characteristics (bound-state eigenvalues and reflection coefficients), finding the GF can be computationally demanding. Methods based on the Bäcklund transformation may ease some of these difficulties (McLaughlin & Scott, 1978).

The GF method often works for one-soliton problems in non-integrable models, as a full set of eigenmodes can often be found for an equation linearized about a soliton even if the equation is not integrable. An example is the nonlinear Klein–Gordon equation

$$\phi_{tt} - \phi_{xx} - \phi + \phi^3 = 0, \tag{5}$$

which describes ferroelectric phase transitions, among other applications. A comprehensive account of the GF method for kinks in non-integrable Klein–Gordon equations was given by Flesch & Trullinger (1987).

Another version of PT is based on the variational approximation (VA), which only demands that the full perturbed equation be derivable from a Lagrangian (e.g., the system is conservative), and that a one-soliton solution be available in the absence of perturbations. The method applies to the description of multisoliton effects in the adiabatic approximation, representing the wave field as a linear superposition of unperturbed solitons. Inserting this approximation into the model's Lagrangian and integrating over the spatial coordinate, one arrives at an effective Lagrangian, which is a function of the solitons' parameters (amplitude, velocity, and central coordinate and phase) and their first derivatives in time. Application of the Euler–Lagrange variational procedure to the effective Lagrangian then generates a system of ODEs that govern the evolution of the solitons' parameters. Because the linear-superposition assumption underlying VA is not valid in the general case (when solitons strongly overlap in the course of the interaction), a VA is restricted to cases when the solitons interact staying far apart or when they collide with a large relative velocity.

An early review of results obtained for multisoliton interactions by means of a VA was given by Gorshkov and Ostrovsky (1981), and an up-to-date review, including the interaction problems, was given by Malomed (2002). There is also an intermediate version of PT, which is based on IST for one soliton, but treats the interaction between solitons, even in an integrable equation, as a perturbation, assuming that the solitons are far separated. By means of this technique, Karpman and Solov'ev (1981) analyzed the interaction between solitons in the NLS and SG equations. Their results for NLS solitons were later checked in a direct experiment with solitons in a nonlinear optical fiber.

The approach based on the linear-superposition approximation for widely separated solitons makes it possible to calculate an effective potential of the interaction between solitons. To this end, one should isolate a part of the model's potential energy which depends on the separation between the solitons. For instance, the potential energy corresponding to the unperturbed SG equation (1) is

$$\Pi = \int_{-\infty}^{+\infty} \left[\tfrac{1}{2}\phi_x^2 + (1 - \cos\phi)\right] \mathrm{d}x\,. \tag{6}$$

In the vicinity of the first kink, the field is approximated as $\phi_1(x-\xi_1) + \delta\phi_2\,(x-\xi_2)$, where $\delta\phi_2$ is a small tail of the second kink (both are taken with $c = 0$). Thus, in the lowest approximation, the contribution to the interaction potential from the vicinity of the first kink [which is, say, $-\infty < x < \Xi \equiv (1/2)\,(\xi_1 + \xi_2)$] is

$$\begin{aligned} U_1 &= \int_{-\infty}^{\Xi} \left[(\phi_1)_x\,(\delta\phi_2)_x + (\sin\phi_1)\,\delta\phi_2\right] \mathrm{d}x \\ &\equiv \int_{-\infty}^{\Xi} \left[-(\phi_1)_{xx} + \sin\phi_1\right] \delta\phi_2\,\mathrm{d}x + (\phi_1)_x\,\delta\phi_2|_{-\infty}^{\Xi}, \end{aligned} \tag{7}$$

using integration by parts. A key observation, which holds in much more general situations, is that the integral term in (7) identically vanishes, as $\phi_1(x)$ is an exact stationary solution of the SG equation; hence, all the contribution comes solely from the last term in (7). This yields the interaction potential, $U = U_1 + U_2 = 32\sigma_1\sigma_2 \exp(-|\xi_1 - \xi_2|)$. Noting that each kink is a quasi-particle with mass $m = 8$, this potential provides for full dynamical description of two- or multi-kink systems.

The derivation of the interaction potential based on the same principles applies to several other cases, including two- and three-dimensional solitons (Malomed, 1998). In the case of NLS solitons, the interacting pair is characterized by distance and relative phase between them, a peculiarity being that an effective mass corresponding to the phase difference is *negative*, which strongly affects stability of two-soliton

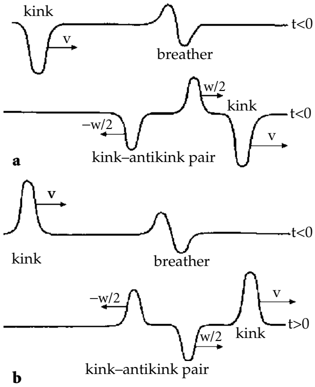

Figure 1. The breather's breakup into a kink-antikink pair as a result of the collision with another fast kink; shown is the field $\partial\phi/\partial x$. Panels (a) and (b) pertain to the polarities $\sigma = -1$ and $+1$ of the fast kink.

bound states induced by perturbations (Malomed, 2002).

The simplest nontrivial example of the two-soliton interaction is fusion of a $k\bar{k}$ pair as a result of collision. An analysis of this problem is outlined here for the case of small γ and α in (1). Because the collision-induced energy loss responsible for fusion is proportional to the dissipation (α) and can be found to be $\Delta E = 8\pi^2\alpha$, fusion occurs if the steady-state velocity given by (4) is small enough. This means that $|\gamma| \ll \alpha$ whereupon Equation (4) yields $c_0 \approx \sigma\pi\gamma\ /\ (4\alpha)$, the corresponding net kinetic energy of the pair being $E_{\mathrm{kin}} = 8c_0^2$. The collision results in pair annihilation if $E_{\mathrm{kin}} < \Delta E$, or $\gamma < \gamma_{\mathrm{cr}} \equiv 4\alpha^{3/2}$.

As shown in Figure 1, the collision of a breather with a moving kink may produce another interesting result in the same model: the breather may be split into a $k\bar{k}$ pair. The final result takes a simple form if the kink is ultrarelavitic, that is, c_0 is close to the limiting value 1: the breakup of the breather (3) into a $k\bar{k}$ pair is possible if $(\pi/2) - \mu < 2.91\sqrt{\alpha}$.

Some special types of conservative perturbations may greatly alter the dynamics. For instance, if the perturbation in the SG equation is $\varepsilon \sin(\phi\ /\ 2)$ (double SG equation), free 2π-kinks do not exist; instead, their pairs form bound states (4π-kinks), $\phi_{4\pi} = 4\tan^{-1}\left(\sqrt{|\varepsilon|\ /\ 2}\sinh x\right)$. It is possible to excite vibrations of this bound state, the eigenfrequency of which is $\sqrt{|\varepsilon|}$, as predicted by PT.

A typical example of two-soliton interactions generating radiation loss is the collision of two kinks (which may have both opposite and equal polarities) with a relative velocity c, the perturbing term in the SG equation being $\varepsilon \sin(2\phi)$. In the ultra-relativistic case, it can be shown directly that the radiation loss appears at order ε^2 and is the same for the $k\bar{k}$ and kk pairs. If the relative velocity of the collision is small, the $k\bar{k}$ pair may annihilate into a breather through the radiation loss, which takes place if $c < c_{\mathrm{cr}} \approx 2.22\,|\varepsilon|$.

Numerical simulations of $k\bar{k}$ collisions in non-integrable models, including Equation (5), reveal an interesting effect. There is a (probably infinite) set of critical relative velocities $c_{\mathrm{cr}}^{(n)}$, $n = 0, 1, 2, \ldots$, such that the annihilation takes place in intervals $0 < c < c_{\mathrm{cr}}^{(0)}$, $c_{\mathrm{cr}}^{(1)} < c < c_{\mathrm{cr}}^{(2)}$, $c_{\mathrm{cr}}^{(3)} < c < c_{\mathrm{cr}}^{(4)}$, …, while in transmission windows between them, $c_{\mathrm{cr}}^{(0)} < c < c_{\mathrm{cr}}^{(1)}$, $c_{\mathrm{cr}}^{(2)} < c < c_{\mathrm{cr}}^{(3)}$, ..., the collision is quasi-elastic.

An explanation for this effect is that the kink has an eigenmode of its internal oscillations, and the dominant collision-induced energy loss is caused not by the emission of radiation but by excitation of the internal mode. An ODE model of a variational type makes it possible to explain the alternating annihilation and transmission windows as a result of resonant transfer of kinetic energy of the colliding kinks into their internal modes and back (Campbell et al., 1986).

Boris A. Malomed

See also **Collisions; Long Josephson junctions; *N*-soliton formulas; Perturbation theory; Solitons**

Further Reading

Campbell, D.K., Peyrard, M. & Sodano, P. 1986. Kink-antikink interactions in the double sine-Gordon equation. *Physica* D, 19: 165–205

Flesch, R.J. & Trullinger, S.E. 1987. Green's function for nonlinear Klein–Gordon kink perturbation theory. *Journal of Mathematical Physics*, 28: 1619–1631

Gorshkov, K.A. & Ostrovsky, L.A. 1981. Interactions of solitons in non-integrable systems: direct perturbation method and applications. *Physica* D, 3: 428–438

Karpman, V.I. & Solov'ev, V.V. 1981. A perturbational approach to the two-soliton systems. *Physica* D, 3: 487–502

Kaup, D.J. & Newell, A.C. 1978. Solitons as particles and oscillators in slowly varying media. *Proceedings of the Royal Society of London* A, 361: 413–466

Keener, J.P. & McLaughlin, D.W. 1977. Solitons under perturbations. *Physical Review* A, 16: 777–790

Kivshar, Y.S. & Malomed, B.A. 1989. Dynamics of solitons in nearly integrable systems. *Reviews of Modern Physics*, 61: 762–915

Malomed, B.A. 1998. Potential of interaction between two- and three-dimensional solitons. *Physical Review* E, 58: 7928–7933

Malomed, B.A. 2002. Variational methods in nonlinear fiber optics and related fields. *Progress in Optics*, 43: 69–191

McLaughlin, D.W. & Scott, A.C. 1978. Perturbation analysis of fluxon dynamics. *Physical Review* A, 18: 1652–1680

MULTISTABILITY

See **Stability**

MUSHROOMS

See **Fairy rings of mushrooms**

MUTUAL ENTRAINMENT

See **Coupled oscillators**

MYELINATED NERVES

While smooth nerve fibers are common to invertebrates, for example, the giant axon of the squid (Hodgkin & Huxley, 1952), myelinated nerve fibers are widely found in vertebrates. A myelinated fiber is mainly composed of a membrane—interfacing the intracellular space and the extracellular one—surrounded by a myelin sheath which acts as an insulator (Waxman et al., 1995; Koch, 1998; Scott, 2002). This sheath, formed of Schwann cells, is periodically absent in gaps called nodes of Ranvier, making a nerve fiber a discrete and periodic structure, as illustrated in Figure 1. Therefore, ion currents through the membrane can only occur at the nodes of Ranvier, which implies a "saltatory" conduction, under which the wave of activity leaps from one node to the next.

One of the main constraints of neural processing is to develop robust and high-speed traveling waves. Compared with smooth fibers, myelinated structures allow an increased velocity of nerve impulses, while decreasing the diameter of the nerve fiber. Thus, the motor nerves of vertebrates may comprise several hundred individual saltatory fibers. For instance, the information transfer rate (bits/s) in a rabbit's sciatic nerve is about 1500 times greater than in a squid axon of the same diameter. In the context of biological evolution, it is interesting to know the link between internode spacing and impulse velocity, and whether this link is optimized in terms of speed and robustness. To this end, a mathematical model can provide some answers.

A single myelinated nerve can be modeled from Kirchhoff's law analysis by the system of difference-differential equations

$$V_n - V_{n+1} = RI_n,$$
$$I_{n-1} - I_n = C\frac{\mathrm{d}V_n}{\mathrm{d}t} + I_{\mathrm{ion},n}. \tag{1}$$

In these equations, the index n indicates successive active nodes, each of which is characterized by a transverse (inside to outside) voltage across the membrane (V_n). A second dynamic variable is the current (I_n) flowing longitudinally through the fiber from node n to node $n+1$. Also, R is the sum of the inside and outside resistances between two nodes and is inversely proportional to the internode spacing s, and C is the membrane capacitance of a node.

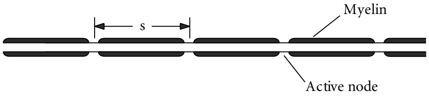

Figure 1. Sketch of a myelinated nerve fiber.

To study the velocity of an impulse, it is sufficient to focus on its leading edge. Thus, we assume that the sodium ion (Na^+) current begins without delay, and the potassium (K^+) permeability remains equal to its resting value. In this case, the expression of the ion currents developed by Alan Hodgkin and Andrew Huxley (*See* **Hodgkin–Huxley equations**) can be reduced to the simple and yet physically reasonable expression:

$$I_{\mathrm{ion},n} = I_{\mathrm{Na},n} = \left(\frac{G}{V_{\mathrm{b}}(V_{\mathrm{b}} - V_{\mathrm{a}})}\right) \times V_n(V_n - V_{\mathrm{a}})(V_n - V_{\mathrm{b}}), \tag{2}$$

where V_{a} is a threshold potential, V_{b} is the Nernst potential, and G is the total ionic conductance near V_{b}.

The parameters of this model can be related to experimental measurements on a real myelinated nerve fiber (Binczak et al., 2001). From studies on frogs' motor nerves with a diameter of 14 μm, $R = 28\,\mathrm{M\Omega}$, $C = 3.7 \pm 1\mathrm{pF}$, $G = 0.57\,\mu\mathrm{S}$, $V_{\mathrm{a}} \sim 25\,\mathrm{mV}$, and $V_{\mathrm{b}} = 122\,\mathrm{mV}$, while the distance between nodes equals 2 mm.

In order to study the influence of the internode spacing s on the impulse velocity, s and therefore R become variable. A discreteness parameter D can be defined by setting $R_{\mathrm{f}} = 28\,\mathrm{M\Omega}$, as

$$D \equiv \frac{2\ \mathrm{mm}}{s} = \frac{R_{\mathrm{f}}}{R}, \tag{3}$$

so that $D = 1$ implies the discreteness of a standard frog nerve. Using dimensionless voltage variables $v_n \equiv V_n/V_{\mathrm{b}}$ and $a \equiv V_{\mathrm{a}}/V_{\mathrm{b}}$, the dynamic equation becomes

$$D(v_{n+1} - 2v_n + v_{n-1}) = R_{\mathrm{f}}C\frac{\mathrm{d}v_n}{\mathrm{d}t} + \frac{R_{\mathrm{f}}G}{1-a}v_n(v_n - a)(v_n - 1), \tag{4}$$

Equation (4) has been used to compute the wave front velocity, which is plotted against D in Figure 2.

If the internodal spacing is small enough, that is, $D \gg 1$, the relative change in voltage between nodes satisfies the inequality $|(v_{n+1} - v_n)/v_n| \ll 1$ and the voltages and currents are smooth functions of distance. Letting $ns \to x$ in this continuum limit, the system can be described by the partial differential equation

$$s^2 D\frac{\partial^2 v}{\partial x^2} - R_{\mathrm{f}}C\frac{\partial v}{\partial t} = \frac{R_{\mathrm{f}}G}{1-a}v(v-a)(v-1), \tag{5}$$

which is the Zeldovich–Frank-Kamenetsky (ZF) equation. Initially formulated as a model for flame front

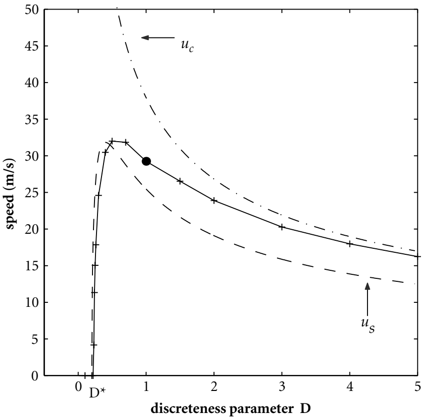

Figure 2. Leading edge impulse velocity on a myelinated axon as a function of the discreteness parameter D. The dot-dashed line indicates the continuum limit of Equation (6). The dashed line indicates the saltatory limit in Equation (8). Crosses indicates numerical results.

propagation, this equation is extensively used to study simple unmyelinated nerve models.

In the continuum limit, the wave speed u tends toward u_c (Zeldovich & Frank-Kamenetsky, 1938), where

$$u_{\mathrm{c}} = s\left(\frac{1-2a}{\sqrt{2(1-a)}}\right)\sqrt{\frac{G}{RC^2}}, \qquad (6)$$

corresponds to the dot-dashed line in Figure 2. As illustrated in this figure, the continuum approximation holds for $D > 5$.

Propagation failure (Keener, 1987) occurs when the distance between the active nodes becomes too large, that is, when D is above a critical value D^*. This critical value of the discreteness parameter is given to lowest order in a as (Erneux & Nicolis, 1993)

$$D^* \approx \frac{R_{\mathrm{f}}Ga^2}{4(1-a)}. \qquad (7)$$

For D slightly larger than D^*, the impulse velocity $u \to u_{\mathrm{s}}$, where

$$u_{\mathrm{s}} = \frac{s}{\pi C}\sqrt{\frac{G(D-D^*)}{R_{\mathrm{f}}(1-a)}}, \qquad (8)$$

which corresponds to the dashed line in Figure 2.

When $D = 1$, the impulse velocity of a normal frog nerve has been calculated to be equal to 29 m/s, which matches the experimental results (Tasaki, 1982). The large dot in Figure 2 shows that the velocity is close to the maximum possible value, suggesting an optimal evolutionary design.

Failure of an impulse is expected to occur at an internode spacing of 1 cm, corresponding to $D = 0.2$, whereas the normal nerve is designed for $D = 1$, corresponding to a spacing of 2 mm. Thus, the process of evolution has provided a comfortable safety margin against accidental failure on the frog's motor nerve.

Because the impulse velocity reaches its maximum value near $D = 1$, we have an explanation for the fact that the conduction velocity of a frog's myelinated motor nerve is insensitive to the internode spacing. From an engineering perspective, operating in this region of parameter space makes the system robust. These observations are extendable to other myelinated fibers (Scott, 2002).

Finally, it should be noted that the transmission of an individual impulse expends much more energy on a smooth fiber, suggesting an additional reason that myelinated nerve fibers may have played an important role in the course of biological evolution.

STEPHANE BINCZAK

See also **Biological evolution; Nerve impulses; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Binczak, S., Eilbeck, J.C. & Scott, A.C. 2001. Ephaptic coupling of myelinated nerve fiber. *Physica* D, 148: 159–174

Erneux, T. & Nicolis, G. 1993. Propagating waves in discrete bistable reaction–difusion systems. *Physica* D, 67: 237–244

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117: 500–554

Keener, J.P. 1987. Propagation and its failure in coupled systems of discrete excitable cells. *SIAM Journal of Applied Mathematics*, 47: 556–572

Koch, C. 1998. *Biophysics of Computation: Information Processing in Single Neurons*, Oxford and New York: Oxford University Press

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Tasaki, I. 1982. *Physiology and Electrochemistry of Nerve Fibers*, New York: Academic Press

Waxman, S.G., Kocsis, J.D. & Stys, P.K. (editors). 1995. *The Axon: Structure, Function and Pathophysiology*, Oxford and New York: Oxford University Press

Zeldovich, Ya.B. & Frank-Kamenetsky, D.A. 1938. K teorii ravnomernogo rasprostranenia plameni (Toward a theory of uniformly propagating flames). *Doklady Akademii Nauk SSSR*, 19: 693–697

N

NATURAL MEASURE

See **Measures**

NAVIER–STOKES EQUATION

In 1822, the French engineer Claude Navier derived the Navier–Stokes equation, as an extension of Euler's equation to include viscosity. Navier was initially interested in blood flow, and he used a molecular approach to arrive at the viscous terms. Navier's equations were generalized to a compressible fluid by Poisson (1829), and one can find fully continuous derivations by De Saint-Venant (1843). A comprehensive treatment was given by George Stokes in 1845, who independently arrived at the results of Poisson and Saint-Venant using a continuous model. Stokes used the common assumption of linear relations between stress and strain rate and discovered "Stokes' law" for the terminal velocity of objects descending in fluids, which he deduced from experiments with slowing pendulums in viscous media.

The Navier–Stokes equation derives from a general equation for the conservation of momentum, balancing forces per unit volume on both sides. It is given by

$$\rho\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \rho\boldsymbol{g} - \nabla p + \mu\left(\nabla^2\boldsymbol{v} + \frac{1}{3}\nabla\nabla\cdot\boldsymbol{v}\right), \qquad (1)$$

where $\boldsymbol{v}(\boldsymbol{r},t)$ is the velocity vector in a cartesian coordinate system, $\rho(\boldsymbol{r},t)$ is the density, $\boldsymbol{g}$ is the acceleration due to external forces (for example, gravitational, magnetic, electro-static forces), $p(\boldsymbol{r},t)$ the pressure, and μ the viscosity coefficient. The material derivative is given by

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \left(\frac{\partial}{\partial t} + \boldsymbol{v}\cdot\nabla\right)\boldsymbol{v}, \qquad (2)$$

where the first and second terms are the local time derivative of the quantity $\boldsymbol{v}(\boldsymbol{r},t)$ with spatial coordinates fixed, and the convective (or advective) term accounting for the change of the quantity $\boldsymbol{v}(\boldsymbol{r},t)$ at $\boldsymbol{r}$ due to the observer following the motion of the fluid with velocity $\boldsymbol{v}$. Both terms together are called the material derivative or convective derivative and are usually denoted by $\mathrm{d}/\mathrm{d}t$.

In an inertial coordinate system, the acceleration $\boldsymbol{g}$ originates solely from the gravitational potential of the Earth. Assuming a homogeneous mass at the center of the Earth, the acceleration $\boldsymbol{g}$ in Equation (1) is given by $\boldsymbol{g} = g\boldsymbol{k}$, where $\boldsymbol{k}$ is the local vertical direction, pointing to the center of gravity of the earth and g is the gravitational acceleration (982.1 cm s^{-2}), which is assumed constant in the first approximation. In a rotating coordinate system, the acceleration $\boldsymbol{g}$ consists of the centrifugal force and Coriolis force in addition to the gravitational acceleration, see Lamb (1906). Also, for large-scale processes, it is often necessary to account for tidal forces of the moon and sun. The term $-\nabla p$ is the force per unit volume due to a pressure gradient of the scalar pressure field $p(\boldsymbol{r},t)$ acting on an infinitesimal volume of fluid with infinitesimal mass.

The last two terms of Equation (1)

$$\frac{\mu}{\rho}\left(\nabla^2\boldsymbol{v} + \frac{1}{3}\nabla\nabla\cdot\boldsymbol{v}\right), \qquad (3)$$

are the frictional terms which derive from a shear stress tensor for a Newtonian viscous fluid with constant viscosity coefficients. A fluid is called Newtonian if there is a linear relation between stress and rate of strain assuming isotropy, which is true for the most common conditions. Non-Newtonian fluids have more complex molecular structure or are mixtures of fluids.

The frictional term $\nabla^2\boldsymbol{v} = \nabla\nabla\cdot\boldsymbol{v} - \nabla\times(\nabla\times\boldsymbol{v})$ contains effects due to compression and rotation. If the fluid is incompressible then the divergence terms are missing from Equations (1) and (3), since $\nabla\cdot\boldsymbol{v} = 0$. Often the coefficient of kinematic viscosity $\nu = \mu/\rho$ is used instead of the viscosity μ. The viscosity of the fluid depends on temperature, and a table for typical values of viscosity μ, kinematic viscosity ν, and density ρ is shown in Table 1 for typical temperatures. If the temperature across the medium is not uniform then a variable viscosity may have to be accounted for (see Batchelor, 1970).

T (°C)	Air ρ (g cm^{-3})	μ (g cm^{-1} s^{-1})	ν (cm^2 s^{-1})	Water ρ (g cm^{-3})	μ (g cm^{-1} s^{-1})	ν (cm^2 s^{-1})
0	1.293×10^{-3}	1.71×10^{-4}	0.132	0.999	1.787×10^{-2}	1.787×10^{-2}
10	1.247×10^{-3}	1.76×10^{-4}	0.141	0.999	1.304×10^{-2}	1.304×10^{-2}
20	1.205×10^{-3}	1.81×10^{-4}	0.150	0.998	1.002×10^{-2}	1.004×10^{-2}

Table 1. Density ρ, viscosity μ and kinematic viscosity ν of air and water.

For an ideal (perfect) fluid without internal shear stress, the momentum equation reduces to Euler's equation of motion for $\mu = 0$. Note that if the flow is irrotational ($\nabla \times \boldsymbol{v} = 0$) and incompressible ($\nabla \cdot \boldsymbol{v} = 0$), it essentially behaves as if it is inviscid ($\mu = 0$), because of the incompressibility condition $\mathrm{d}\rho/\mathrm{d}t = -\rho \nabla \cdot \boldsymbol{v} = 0$. (Note that a fluid may be incompressible yet ρ may not be constant.)

The system of Equations (1) comprises three momentum equations, together with an additional equation for mass conservation; a thermodynamic equation of state; a relation of pressure, density, and either temperature or entropy; and an energy equation for the additional thermodynamic variable (temperature or entropy). These provide six scalar equations for the determination of the six independent variables, velocity $\boldsymbol{v}$, density ρ, pressure p, and temperature or entropy as functions of space $\boldsymbol{x}$ and time t. If the fluid is incompressible, the equation of state becomes obsolete and only four equations are needed, the three momentum equations and the equation for conservation of mass.

Numerical Problems

As with most complicated nonlinear partial differential equations, the Navier–Stokes equation is solved numerically to model a specific fluid flow that is of interest experimentally and theoretically. Because there are many different ways to write the Navier–Stokes equation, the first numerical difficulty is to pick the formulation most suitable to the numerical technique that one wishes to employ and the dimensions of the model. One distinguishes between two-dimensional and three-dimensional models. For two-dimensional incompressible Navier–Stokes equation, for example, there are four different kinds of formulations: the primitive-variable (velocity and pressure), stream function-vorticity, stream function, and velocity-vorticity formulation.

It is important to distinguish between the real dissipation ($\mu \neq 0$), given by the last term of Equation (1) and numerical dissipation, which is introduced by an accumulative error due to the limited order of accuracy of any numerical model. As the numerical dissipation of a good model is often negligible, it is sometimes appropriate to introduce an artificial viscosity that acts to damp growing high wave number modes, which can lead to numerical instability.

Euler's equation conserves linear and angular momentum and energy, so it is necessary to verify that the numerical model conserves these conserved quantities and preserves symmetry. In cases where it is not possible to satisfy all these requirements, properties that are most essential to the physical problem are given priority. In order to satisfy such conserved quantities, it is important that the numerical scheme treats the nonlinear convective term

$$\boldsymbol{v} \cdot \nabla \boldsymbol{\phi} \tag{4}$$

in a conservative manner, where ϕ represents the scalar quantity that is convected and could be one of the three velocity components u, v, w or density ρ in the conservation of mass. Expression (4) is known as the convection form and is nonconservative numerically, since fluxes across mesh boundaries do not cancel. The convection form thus represents a major flaw of every numerical scheme. The most commonly used corresponding conservative formulation is

$$\nabla \cdot (\boldsymbol{v}\phi), \tag{5}$$

which is also known as the divergence form. Recently it was shown that the skewsymmetric form

$$\tfrac{1}{2}\boldsymbol{v} \cdot \nabla\phi + \tfrac{1}{2}\nabla \cdot (\boldsymbol{v}\phi)$$

has advantages, because it reduces aliasing errors (see below) for yet unknown reasons and is preferable to the popular rotational form

$$(\nabla \times \boldsymbol{v}) \times \boldsymbol{v}.$$

It is important to note that although both equations (4) and (5) are formally equivalent, they are not in a corresponding numerical scheme and care must be taken to avoid this common pitfall when modeling equations that contain convective terms such as the Navier–Stokes equation or the equation for mass conservation (see Hirsch, 2000).

Nonlinear terms such as $\boldsymbol{v} \cdot \nabla \boldsymbol{v}$ also generate aliasing errors. These are high-frequency wave number modes appearing or being "aliased" as low-frequency modes that deteriorate the wave number spectrum and eventually lead to numerical blow-up.

Because a numerical model of the Navier–Stokes equation is coupled spatially and temporally, it is helpful to approximate the spatiotemporal scales of the dynamics. One should determine whether the spatial and temporal scales are equal to, smaller, or larger than each other. If, for example, the temporal scale is larger than the spatial scale, it is sufficient to use low order integration in time such as finite-difference methods, and if smaller or equal, then higher-order temporal integration such as spectral methods becomes necessary. In a temporally extensive computation, it is sometimes possible to revert to finite differences, choosing a relatively small time-step.

Especially difficult numerical problems arise when it is necessary to invert nondiagonal, nonsparse, nontrivial matrices, for example, in implicit time integrations or inverting the Laplacian. In such cases iterative techniques using preconditioning and multi-grid methods have been successfully employed. Complications arising from specifying the correct boundary conditions to use along boundaries of the physical domain should also be mentioned.

Although the Navier–Stokes equation is a very good model for a real fluid, it is important to have an understanding of the process or phenomenon that one wishes to model as well as a sound knowledge of the applicability of the numerical methods that one wishes to employ to be able to faithfully represent the physical dynamics of the fluid.

Phenomena

A salient feature of the Navier–Stokes equation is the phenomenon of turbulence. Due to viscous effects the flow of a real fluid can be observed in two very different states: a laminar state and a turbulent state.

Turbulence induces a cascade to fine scales that are eventually dissipated. It is important in many engineering applications, as a loss of energy means increased costs and can possibly cause damage to the structural body that comes in contact with the fluid or which moves through the fluid (e.g., optimization of airfoils, loss of lift force due to turbulence and boundary layer separation).

Closely related to the phenomenon of turbulence is boundary layer theory and boundary layer separation. Fluid flowing past a surface at rest relative to the fluid will experience friction at the surface of the material, and for the velocity field to be continuous, it is required that the normal as well as tangential fluid velocity vanish at the material surface ($\boldsymbol{v} = 0$). For an idealized fluid ($\mu = 0$), only the normal velocity has to vanish.

It was the problem of matching the vanishing flow at the surface boundary to some nonzero fluid flow away from the boundary that led Prandtl (1905) to introduce the concept of boundary layer, which is a small "inner" fluid layer close to the surface where viscous effects dominate the flow. This inner layer is joined to an "outer" layer, where the fluid can be considered inviscid. The concept of boundary layer has led to great progress in fluid mechanics and perturbation methods, where matched asymptotic expansions feature prominently, but a complete understanding of boundary layer separation and its connection to turbulence is still missing.

A useful concept is vorticity, which can be regarded as a beautiful generalization of fluid motion. Formally vorticity is defined as

$$\omega = \nabla \times \boldsymbol{v},$$

which simplifies to twice the angular velocity for solid body rotation. A fluid is called irrotational if its vorticity is zero and rotational if it is not. Taking the curl of Euler's equation ($\mu = 0$) for an incompressible fluid in the absence of baroclinicity (and after some vector algebra) leads to the vorticity equation

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} - \omega \cdot \nabla \boldsymbol{v} = 0.$$

Vortex lines and vortex strength are conserved, vortex lines move with the fluid, and an initially irrotational flow remains irrotational. Associated to vortices and turbulence is the, still not understood, phenomenon of vortex breakdown (Benjamin, 1978).

Among wave phenomena, the most striking is definitely the solitary wave, which appears as a nonlinear wave in a fluid internally as well as at fluid surfaces or interfaces. For free-surface wave phenomena, see Johnson (1997) and the references therein, and for a colorful overview, see Lighthill (1978). A beautiful picture gallery of fluid phenomena can be found in Van Dyke (1982). Other phenomena that can only be listed here include cavitation and thermal convection (Rayleigh–Bénard convection), sub- and supersonic flow, shock waves as well as most instability mechanisms and phenomena (Rayleigh–Taylor instability, Kelvin–Helmholtz instability, barotropic and baroclinic instability).

Andreas A. Aigner

See also **Alfvén waves; Boundary layers; Burgers equation; Chaos vs. turbulence; Fluid dynamics; Hydrothermal waves; Kelvin–Helmholtz instability; Korteweg–de Vries equation; Lattice gas methods; Magnetohydrodynamics; Mixing; Rayleigh–Taylor instability; Solitons; Taylor–Couette flow; Turbulence; Water waves**

Further Reading

Batchelor, G.K. 1970. *An Introduction to Fluid Dynamics*, Cambridge and New York: Cambridge University Press

Benjamin, T.B. 1978. Theory of the vortex breakdown phenomenon. *Journal of Fluid Mechanics*, 14: 593–629

Chandrasekhar, S. 1961. *Hydrodynamic and Hydromagnetic Stability*, Oxford: Clarendon Press and New York: Dover

De Saint-Venant, J.-C. 1843. Note à joindre au mémoire sur le dynamique de fluides. *Comptes Rendu de l'Academie des Sciences Paris*, 17: 1240–1243

Drazin, P.G. 1981. *Hydrodynamic Stability*, Cambridge and New York: Cambridge University Press

Hirsch, C. 2000. *Numerical Computation of Internal and External Flows*, New York: Wiley

Johnson, R.S. 1997. *A Modern Introduction to the Mathematical Theory of Water Waves*, Cambridge and New York: Cambridge University Press

Lamb, H. 1906. *Hydrodynamics*, Cambridge: Cambridge University Press; 6th edition, Cambridge and New York: Cambridge University Press, 1993

Landau, L.D. & Lifshitz, E.M. 1987. *Fluid Mechanics*, vol.6 of *Course of Theoretical Physics*, London: Pergamon Press and Reading, MA: Addison-Wesley

Lighthill, J. 1978. *Waves in Fluids*, Cambridge and New York: Cambridge University Press

Milne-Thomson, L.M. 1968. *Theoretical Hydrodynamics*, 5th edition, New York: Macmillan Press

Navier, M. 1822. Mémoire sur les lois du mouvement des fluides. *Mémoires de l'Academie Royale des Sciences*, 6: 389–440

Poisson, S.D. 1829. Mémoire sur les équations générales de l'équilibre et du nouvement des corps silides élastiques et de fluides. *Journal de l'École Polytechnique*, 13: 1–174

Prandtl, L. 1905. *Verhandlungen des dritten internationalen Mathematiker-Kongresses, Heidelberg 1904*, Leipzig: 484–491

Stokes, G.G. 1845. On the theories of the internal friction of fluids in motion, and of the equilibrium and motion of elastic solids. *Transactions, Cambridge Philosophical Society*, 8: 287–319

Streeter, V.L. & Wylie, E.B. 1985. *Fluid Mechanics*, 8th edition, New York: McGraw-Hill

Van Dyke, M. 1982. *An Album of Fluid Motion*, Stanford: Parabolic Press

Vennard, J.K. 1954. *Elementary Fluid Mechanics*, New York: Wiley

N-BODY PROBLEM

The Newtonian N-body problem of celestial mechanics is a mathematical generalization of the solar system. In the Euclidean space $\mathbb{R}^3$, consider N points of masses $m_i > 0$, $i = 1, \ldots, N$, which attract each other by a force directly proportional to the product of the masses and inversely proportional to the square of the distance. The equations of motion are given by the $6N$-dimensional system of differential equations

$$\begin{cases} \dot{\boldsymbol{q}} = M^{-1}\boldsymbol{p}, \\ \dot{\boldsymbol{p}} = G\nabla U(q), \end{cases} \tag{1}$$

where the upper dot represents differentiation with respect to the time variable; $\boldsymbol{q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ is the configuration of the particle system, with $\boldsymbol{q}_i = (q_i^1, q_i^2, q_i^3)$; giving the coordinates of the point of mass m_i, $\boldsymbol{p} = M\dot{\boldsymbol{q}}$ is the momentum, where M is a $3N$-dimensional square matrix having on the diagonal the elements $m_1, m_1, m_1, \ldots, m_N, m_N, m_N$ and zeros in rest; G is the gravitational constant; and $U(\boldsymbol{q}) = \sum_{1 \le i < j \le N} \frac{m_i m_j}{|\boldsymbol{q}_i - \boldsymbol{q}_j|}$ is the potential function, $-U(\boldsymbol{q})$ representing the potential energy. Standard results of the theory of differential equations ensure the existence and uniqueness of an analytic solution for any initial value problem as long as the initial data do not belong to the collision set $\Delta = \bigcup_{1 \le i < j \le N} \Delta_{ij}$, where $\Delta_{ij} = \{\boldsymbol{q} \in \mathbb{R}^{3N} \,|\, \boldsymbol{q}_i = \boldsymbol{q}_j\}$.

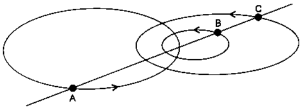

Figure 1. The Eulerian solutions of the three-body problem.

Isaac Newton formulated this problem in his master work *Principia*, but Leonhard Euler was the first to write the equations as we know them today. The case $N = 2$, also called the Kepler problem, is completely solved (see Albouy (2002) for a recent discussion of some early solutions). The relative motion of one body with respect to the other is planar and, depending on the initial conditions, can be a circle, an ellipse, a parabola, a branch of a hyperbola, or a line, in which case a collision takes place in the future or in the past. Kepler's laws (*See* **Celestial mechanics**) can be recovered from Equation (1).

For $N \ge 3$, very little is known about the N-body problem in spite of thousands of research papers written over more than three centuries. The case $N = 3$ was the most studied since many of the results obtained could be generalized to any larger N. The first attempts to understand the three-body problem were quantitative, aiming at finding explicit solutions. In 1767, Euler found the collinear periodic orbits, in which three bodies of any masses move such that they oscillate along a rotating line (Euler, 1767, Figure 1) and in 1772, Joseph-Louis Lagrange discovered some periodic solutions that lie at the vertices of a rotating equilateral triangle that shrinks and expands periodically (Lagrange, 1772, Figure 2). Those solutions led to the study of central configurations, for which $\boldsymbol{q}'' = k\boldsymbol{q}$ for some constant $k > 0$. Each central configuration provides a class of periodic orbits.

Another idea was to reduce the order of the system with the help of first integrals. Ten linearly independent integrals are known: three for the center of mass, three for the momentum, three for the angular momentum, and one for the energy (see Wintner, 1947). Together with a certain symmetry, these integrals allow the reduction of the three-body problem from dimension 18 to 7. But unfortunately the dimension of the problem cannot be further reduced. In 1887, Heinrich Bruns proved that there exist no more linearly independent integrals, algebraic with respect to $\boldsymbol{q}$, $\boldsymbol{p}$ and t (Bruns, 1887), thus showing the limitations of the quantitative methods. This led Henri Poincaré to

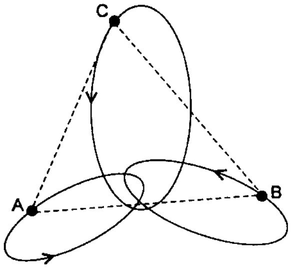

Figure 2. The Lagrangian solutions of the three-body problem.

attempt a qualitative approach. His first prolific ideas appeared in a memoir published in 1890 (Poincaré, 1890), for which he was awarded the King Oscar Prize (see Barrow-Green, 1997; Diacu & Holmes, 1996). There he laid the foundations of several branches of mathematics, including dynamical systems, chaos, KAM theory, and algebraic topology.

Poincaré aimed to understand the geometry of the phase space and the relative behavior of orbits and, thus, to answer questions regarding stability, asymptotic motion at infinity, existence of periodic orbits, etc. An important problem in this direction was that of singular solutions, that is, those that tend to the collision set Δ. It took mathematicians almost a century to prove that for $N \geq 5$, there exist singular solutions that do not end in collisions but in pseudocollisions, which are orbits that become unbounded in finite time (see Diacu & Holmes, 1996). For $N = 4$, the problem is still open.

Recently, a lot of interest has been in finding choreographies, that is, periodic orbits for which all the bodies move on the same closed curve. For more than two centuries the only known example was the class of Lagrangian solutions in the particular case that all the masses are equal and the ellipses are circles. With the help of variational methods, in 2000, a spectacular new class was proved to exist: three bodies of equal mass chase each other along a curve resembling the figure eight (Chenciner & Montgomery, 2000; Montgomery, 2001, Figure 3). There is numerical evidence that this periodic orbit is KAM stable, that is, the solutions through most initial conditions in some sufficiently small neighborhood of the orbit stay close to it for all time, while the other solutions leave the neighborhood very slowly. Unfortunately, the stability region seems to be very small. Numerical experiments suggest that the probability of finding an eight in the universe is somewhere between one per galaxy and one per universe. Hundreds of other choreographies have been numerically put into the evidence.

Still far from fully understood are the questions regarding various restricted three-body problems. In the elliptic one, for example, it is asked to determine the motion of one body, assumed to have zero mass, while the other two move on ellipses as in the Kepler problem with negative energy.

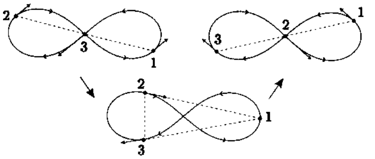

Figure 3. The figure eight solution of the 3-body problem.

Numerical methods are also of help for getting insight into the problem. But due to the apparently chaotic character of the motion, they must be implemented with care. Recently, much progress has been made in successfully applying scientific computation to various aspects of the general and restricted three-body problem.

Many of the ideas of the classical N-body problem can be adapted to related problems for understanding the motion of particle systems given by other potentials, like those of Manev and Schwarzschild (also used in celestial mechanics), Coulomb (atomic and molecular theories), and Lennard-Jones (crystal formation). Based on the Coulomb potential, Niels Bohr's model of the hydrogen atom led to the development of quantum mechanics. Several other branches of science have profited from the study of the N-body problem.

FLORIN DIACU

See also **Celestial mechanics; Poincaré theorems; Solar system**

Further Reading

Albouy, A. 2002. Lectures on the two-body problem. In *Classical and Celestial Mechanics: The Recife Lectures*, edited by H. Cabral & F. Diacu, Princeton, NJ: Princeton University Press, pp. 71–135

Barrow-Green, J. 1997. *Poincaré and the Three-Body Problem*, Providence, RI: American Mathematical Society

Bruns, H. 1887. Über die Integrale des Vielkörper-Problems, *Acta Mathematica* 11: 25–96

Chenciner, A. & Montgomery, R. 2000. A remarkable periodic solution of the three-body problem in the case of equal masses. *Annals of Mathematics*, 152: 881–901

Diacu, F. & Holmes, P. 1996. *Celestial Encounters—The Origins of Chaos and Stability*, Princeton, NJ: Princeton University Press

Euler, L. 1767. De moto rectilineo trium corporum se mutuo attrahentium, *Novo Comm. Acad. Sci. Imp. Petrop.*, 11: 144–151

Lagrange, J.L. 1873. Essai sur le probl eme des trois corps. In *Ouvres de Lagrange*, vol. 6, pp. 229–324, Paris: Gauthier-Villars, 14 vols

Montgomery, R. 2001. A new solution to the three-body problem. *Notices of the American Mathematical Society* May, pp. 471–481

Poincaré, H. 1890. Sur le problème des trois corps et les équations de la dynamique. *Acta Mathematica*, 13: 1–270
Wintner, A. 1947. *The Analytical Foundations of Celestial Mechanics*, Princeton, NJ: Princeton University Press

NEEL DOMAIN WALL

See **Domain walls**

NEGATIVE RESISTANCE AND CONDUCTANCE

See **Diodes**

NEPHRON DYNAMICS

The kidneys play an important role in regulating the blood pressure and maintaining a proper environment for the cells of the body. The mammalian kidney contains a large number of functional units, the nephrons. For a human kidney the number of nephrons is of the order of 1 million, and for a rat kidney the number is approximately 30,000. The nephrons are organized in a parallel structure such that the individual nephron only processes a very small fraction of the total blood flow to the kidney. To distribute the blood that enters through the renal artery, the kidney makes use of a network of arteries and arterioles. Closest to the nephron, we have the afferent arteriole that leads the blood to the capillary network in the glomerulus where water, salts, and small molecules are filtered from the blood and into the tubular system of the nephron. On the other side of the glomerulus, the efferent arteriole leads the blood to another capillary system that receives the water and salts reabsorbed by the tubules.

Figure 1 provides a sketch of the main structural components of the nephron. Note how the terminal part of the loop of Henle passes within cellular distances of the afferent arteriole. As described below, this anatomical feature allows for a special feedback regulation.

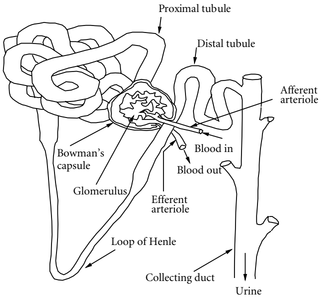

Figure 1. Sketch of the main structural components of the nephron.

In order to protect its function in the face of a varying blood pressure, the individual nephron disposes of a number of mechanisms to control the incoming blood flow. Most important is the tubuloglomerular feedback (TGF) mechanism that regulates the diameter of the afferent arteriole in response to variations in the NaCl concentration in the fluid that leaves the loop of Henle via the distal tubule. If the salt concentration in this fluid becomes too high, specialized cells (macula densa cells) near the terminal part of the loop of Henle elicit a signal that causes the smooth muscle cells at the downstream end of the afferent arteriole to contract. Hence, the incoming blood flow is reduced, and so is the glomerular filtration rate.

The TGF mechanism represents a negative feedback. However, by virtue of a delayed action associated with a finite transit time through the loop of Henle, the flow regulation tends to become unstable and produce self-sustained oscillations with a period of 30–40 s. While for rats with normal blood pressure, these oscillations have the appearance of a regular limit cycle with a sharply peaked power spectrum, highly irregular oscillations, displaying a broadband spectral distribution with strong subharmonic components are observed for spontaneously hypertensive rats. In a particular experiment, clear evidence of a period-doubling of the pressure oscillations has been found, indicating that the nephronic control system is operating close to a transition to chaos.

Figure 2 displays examples of the tubular pressure oscillations observed for normotensive rats (a) and for spontaneously hypertensive rats (b), respectively. The

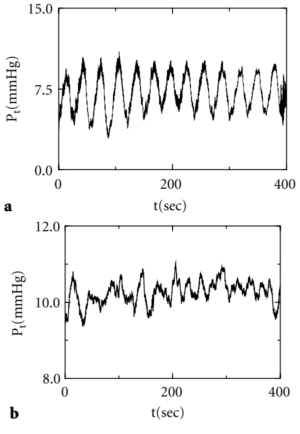

Figure 2. Regular (a) and irregular (b) tubular pressure oscillations from a normotensive and a spontaneously hypertensive rat, respectively.

pressure oscillations can be observed by means of a glass pipette inserted into the proximal tubule of a nephron at the surface of the kidney.

The processes involved in the nephron autoregulation are known in considerable detail. Besides the filtration of water and salts in the glomerulus, these processes include the passive (osmotic) and active (enzymatically controlled) processes by which water and salts are reabsorbed along the loop of Henle, the enzymatic processes through which the smooth muscle cells in the arteriolar wall are activated by the macula densa signal, and the dynamic response of the arteriolar wall to external stimulation.

The steady-state response of the TGF mechanism can be obtained from open-loop experiments in which a block of paraffin is inserted into the middle of the proximal tubule and the glomerular filtration rate is measured as a function of an externally forced flow of artificial tubular fluid into the loop of Henle. Reflecting physiological constraints on the diameter of the arteriole, this response follows an S-shaped characteristic with a maximum at low Henle flows and a lower saturation level at externally forced flows beyond 20−25 nl/min.

Together with the delay in the TGF regulation, the steepness of the response plays an essential role for the stability of the feedback system. The length of the delay can be estimated from the phase shift between the pressure oscillations in the proximal tubule and the oscillations of the NaCl concentration in the distal tubule. A typical value is 10–15 s. In addition, there is a transit time of 3–5 s for the signal from the macula densa cells to reach the smooth muscle cells in the arteriolar wall. The result is a total delay of 14–18 s. The steepness α of the steady state response curve is found to be significantly higher for spontaneously hypertensive rats than for normal rats.

By integrating the different physiological processes into a coherent, nonlinear dynamic model, it has been possible to show how these processes together produce the observed behavior, that is, the emergence of self-sustained oscillations as the slope of the response characteristic exceeds $\alpha \cong 11$ and the transition to chaos via sets of overlapping period-doubling cascades as the feedback slope exceeds $\alpha \cong 20$.

Figure 3 shows a two-dimensional bifurcation diagram for the single-nephron model. The dashed curve is a Hopf bifurcation curve. Period-doubling and saddle-node bifurcations are indicated as fully drawn and dotted curves, respectively. As before, T is the total delay in the TGF regulation, and α is the (maximal) slope of the open-loop feedback characteristic.

The single-nephron model can also be used to simulate the response to an external perturbation, for instance, the infusion of artificial tubular fluid into the loop of Henle or the administration of a drug to the rat. This last possibility is rapidly gaining in significance as the application of simulation models in the development of new drugs becomes more and more important.

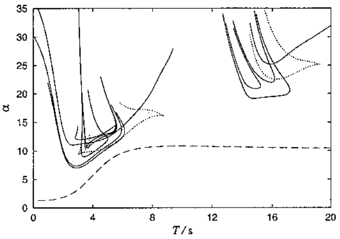

Figure 3. Two-dimensional bifurcation diagram for the single-nephron model. The diagram illustrates the complicated bifurcation structure in the region of 1:1, 1:2, and 1:3 resonances between the arteriolar dynamics and the TGF-mediated oscillations. In the physiologically interesting regime around $T = 16$ s, another set of complicated period-doubling and saddle-node bifurcations occur. Here, the nephron is operating close to a 1:4 (or 1:5) resonance.

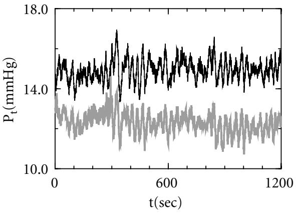

Figure 4. Example of chaotic phase synchronization for a pair of adjacent nephrons in a hypertensive rat.

A variety of cooperative phenomena that can arise from interactions among the nephrons may also be significant. The functional units are typically arranged in couples or triplets with their afferent arterioles branching off from a common interlobular artery. This structure allows neighboring nephrons to interact via signals that propagate along the arteriolar system. As experiments show, this interaction can lead to various forms of synchronization among the nephrons, including in-phase and antiphase synchronization for regularly oscillating nephrons, and chaotic phase synchronization for nephrons with irregular oscillations. By modeling these coupling phenomena in detail, one may be able to predict the typical size of the synchronization domains and to examine the role that synchronization among the nephrons plays in the overall regulation of the kidney.

Figure 4 shows an example of the tubular pressure oscillations observed for two adjacent nephrons in a hypertensive rat. The transition to synchronization can be observed as a locking of the average periods of the two signals in a 1:1 relation to one another. Alternatively, one can define and follow the temporal variation of the instantaneous phases for the two signals.

E. Mosekilde, N.-H. Holstein-Rathlou, and
O. Sosnovtseva

See also **Coupled oscillators**

Further Reading

Blekhman, I. 1988. *Synchronization in Science and Technology*, New York: ASME Press

Fung, Y.-C.B. 1981. *Biomechanics. Mechanical Properties of Living Tissues*, New York: Springer

Glass, L. & Makey, M.C. 1988. *From Clocks to Chaos: The Rhythms of Life*, Princeton, NJ: Princeton University Press

Keener, J. & Sneyd, J. 1998. *Mathematical Physiology, Interdisciplinary Applied Mathematics*, New York: Springer

Layton, H. & Weinstein, A. (editors). 2001. *Membrane Transport and Renal Physiology*, New York: Springer

Mosekilde, E. 1996. *Topics in Nonlinear Dynamics: Applications to Physics, Biology and Economic Systems*, Singapore: World Scientific

Mosekilde, E., Maistrenko, Yu. & Postnov, D. 2002. *Chaotic Synchronization — Applications to Living Systems*, Singapore: World Scientific

Pikovsky, A., Rosenblum, M. & Kurths, J. 2001. *Synchronization: A Universal Concept in Nonlinear Science*, Cambridge and New York: Cambridge University Press

Seldin, D.W. & Giebisch, G. (editors). 1992. *The Kidney: Physiology and Pathophysiology*, New York: Raven

NERVE IMPULSES

The scientific study of nerve impulses goes back to 1791, when Luigi Galvani reported that a frog's leg muscle twitches if the attached nerve is stimulated with a bimetallic contact (Brazier, 1961). This discovery was soon followed by Alessandro Volta's invention of the battery, which launched the science of electrophysiology and raised the question: Does animal electricity differ from chemical electricity?

Among attempts to answer this question was a key experimental study by young Hermann Helmholtz, who cleverly measured the speed of signal propagation on a frog's sciatic nerve (Helmoltz, 1850). In making this measurement, he ignored the advice of his father, a philosopher, who believed that muscular motion was identical to its motivation; thus, any time delay between thinking and doing was theoretically impossible. To the contrary, Helmholtz found a velocity of about 27 m/s, which is much less than the speed at which an electrical signals propagate along conducting wires. Although an outstanding theoretical physicist, Helmholtz was unable to understand why a nerve impulse should move so slowly. Was it the mechanical motion of some molecular substance that he had observed? Interestingly, nonlinear diffusion was suggested as an answer to this puzzle by Robert Luther at the beginning of the 20th century (Luther, 1906).

Among the wonders of electronics that appeared in the 20th century was the cathode-ray oscilloscope (CRO), which Kenneth Cole used to take the first photograph of a nerve impulse on the giant axon of the squid (see Figure 1) (Cole & Curtis, 1938). Time increases to the right as indicated in milliseconds by the marks on the lower margin, showing that horizontal CRO deflections were not yet uniform in the late 1930s. The solid line is the transmembrane voltage (V), which rises rapidly from a resting level, hesitates at a peak level of about 100 mV, and then falls back more slowly. The width of the band indicates the membrane permeability (or conductivity), which evidently increases greatly during passage of the nerve impulse.

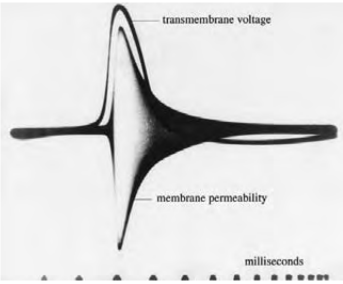

Figure 1. Time course of the transmembrane voltage and membrane permeability of an impulse on a squid nerve. (Courtesy of Kenneth Cole.)

Progress toward explaining these phenomena came soon after the Second World War, taking advantage of the significant advances in electronics during those years. In 1952, Alan Hodgkin and Andrew Huxley presented a series of papers on the squid giant axon, which culminated in a formulation of nerve impulse dynamics based on an empirically determined reaction-diffusion system, from which all details of Figure 1 were computed (Hodgkin & Huxley, 1952). In this Hodgkin–Huxley (HH) model, the initial rise of transmembrane voltage is caused by an inrush of positively charged sodium ions, and the maximum voltage is attained when the inward diffusion of sodium ions is balanced by outward conduction current through the membrane. On a longer time scale, positively charged potassium ions flow out of the nerve, bringing the transmembrane voltage back to its resting value, whereupon it is ready to conduct another impulse. Thus, the means by which squid nerves carry signals are explained by the nonlinear reaction-diffusion system

$$\frac{\partial^2 V}{\partial x^2} - rc\frac{\partial V}{\partial t} = rj_{\text{ion}}, \tag{1}$$

where r is the series resistance per unit length of the axon core, c is the capacitance per unit length, and j_{ion} is a rather complicated expression for the transmembrane ionic current per unit length.

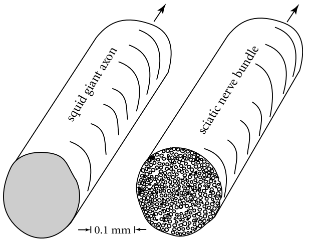

Figure 2. A squid giant axon and the sciatic nerve of a rabbit, to the same scale. (Data from Young, 1951).

In thinking about the HH formulation, it is important to be aware that a squid nerve differs from the sciatic nerve of the frog, which Galvani and Helmholtz studied. A sciatic nerve is actually a bundle of smaller fibers, whereas the squid nerve is a single "giant" axon. The squid nerve is uniform along the propagation direction; thus, the HH system is based on the partial differential equation system of Equation (1). The individual fibers of a sciatic nerve, on the other hand, are covered with an insulating layer (myelin) except at rather widely spaced active nodes (nodes of Ranvier). Thus, impulse propagation on a myelinated axon is described by a difference-differential equation, in which the wave of activity jumps from node to node (saltatory conduction). Among other differences, this means that impulse conduction velocity increases as the square root of diameter of a squid fiber, while it is roughly proportional to the first power of the diameter of a myelinated fiber. Importantly, the energy expended in transmitting a nerve impulse is much less in a myelinated fiber than in a smooth one.

These qualitative differences are emphasized in Figure 2, which compares a typical squid nerve with the sciatic nerve of a rabbit. The rabbit nerve contains about 375 small (myelinated) axons, each of which can conduct an impulse at up to 80 m/s, or about four times faster than a squid axon, leading to an increase in data transmission capacity of about three orders of magnitude. This dramatic increase in information carrying capacity is typical in vertebrate motor neurons, which are myelinated nerve bundles.

From an analytic perspective, the HH formulation is rather complicated, as it involves five dynamic variables, each of which depends on both longitudinal position and time. These are transmembrane voltage, axial current, sodium turn-on and turn-off variables, and a potassium turn-on variable. Thus, it has been of interest to consider other formulations that preserve qualitative properties of the HH system while simplifying their structure.

In the most simple approximation, the transmembrane ionic current is assumed to be a cubic nonlinear function of the transmembrane voltage; thus,

$$j_{\mathrm{ion}} \approx \left(\frac{g}{V_{\mathrm{th}} V_{\mathrm{max}}}\right) V(V - V_{\mathrm{th}})(V - V_{\mathrm{max}}), \qquad (2)$$

where V_{th} is a threshold voltage and V_{max} is the amplitude of the impulse. Under this approximation, the HH dynamics reduce to the Zeldovich–Frank-Kamenetsky (ZF) equation, which describes the leading edge of an impulse but misses the return of the voltage to its resting value. From this perspective, a zero-order estimate of the impulse velocity is of order $\sqrt{g/rc^2}$. Several analytic formulas for the dependence of impulse velocity on axon parameters have been obtained under the ZF approximation (Scott, 2002).

A simple way to represent recovery was developed by Vladislav Markin and Yuri Chizmadzhev in the 1960s (Markin & Chizmadzhev, 1967). Under this MC approximation, a prescribed time course of transmembrane ionic current is assumed to be triggered if the membrane potential reaches a threshold variable. This prescribed current is a negative (inward) current (j_1) maintained for a time τ_1 followed by a positive (outward) current (j_2) maintained for a time τ_2. (The condition $j_1\tau_1 = j_2\tau_2$ then ensures that the net charge crossing the membrane during an impulse is zero.) With appropriate parameters, the MC model yields a recovering impulse having approximately the speed and threshold properties of the HH impulse.

To bring dynamics into the picture without invoking the complexities of the full HH model, Jin-Ichi Nagumo and his colleagues used a membrane model previously developed by Richard FitzHugh, in which the cubic ionic current of Equation (2) is augmented with a single dynamic recovery variable that drives the membrane voltage back to its resting level (Nagumo et al., 1962). This FitzHugh–Nagumo (FN) system was developed as an electronic equivalent of a nerve axon, which can be regarded as a *neuristor*. In the early 1970s, FN became of interest to applied mathematicians, who were beginning to study the theoretical properties of nonlinear reaction-diffusion systems, and in two or three spatial dimensions it has been used to study the emergence of spiral and scroll waves.

In addition to its inherent complexity, the HH system is driven by an initial inrush of sodium ions, whereas the exciting current in many nerve fibers, including dendrites of the human brain, are largely driven by calcium ion current. For many applications, therefore, it is currently of interest to use a version of the model developed by Catherine Morris and Harold Lecar for calcium ion induced membrane switching in the giant muscle fiber of the barnacle (Morris & Lecar, 1981). In this Morris–Lecar (ML) model, $j_{\mathrm{ion}} \approx 2\pi a J_{\mathrm{ml}}$, where a is the axon radius and

$$\begin{aligned} J_{\mathrm{ml}} = {} & G_{\mathrm{K}} n\,(V - V_{\mathrm{K}}) + G_{\mathrm{Ca}} m\,(V - V_{\mathrm{Ca}}) \\ & + G_{\mathrm{L}}(V - V_{\mathrm{L}}) \end{aligned}$$

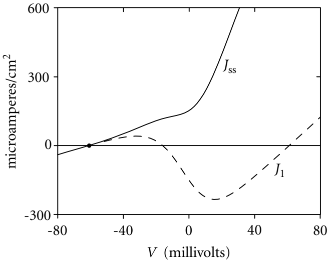

Figure 3. Plots of the initial ionic current (J_1) and the steady state current (J_{ss}) for a typical Morris–Lecar model (Fall et al., 2002; Scott, 2003).

with m and n being calcium and potassium turn-on variables, respectively. Also G_K, G_{Ca}, and G_L are membrane conductances, and V_K, V_{Ca}, and V_L are equilibrium potentials for potassium, calcium, and "leakage" ions, respectively. The turn-on variables are assumed to obey first-order dynamics as (Fall et al., 2002)

$$\frac{dm}{dt} = -[m - m_0(V)]/\tau_m(V),$$
$$\frac{dn}{dt} = -[n - n_0(V)]/\tau_n(V),$$

where

$$m_0(V) = [1 + \tanh((V - V_1)/V_2)]/2,$$
$$n_0(V) = [1 + \tanh((V - V_3)/V_4)]/2$$

and

$$\tau_m(V) = \tau_{m0}\,\mathrm{sech}[(V - V_1)/2V_2],$$
$$\tau_n(V) = \tau_{n0}\,\mathrm{sech}[(V - V_3)/2V_4].$$

At times short compared with τ_n, J_{ml} appears as the cubic function, which is plotted as a dashed line in Figure 3; thus the leading edge of an impulse will propagate as required by the ZF equation. At times long compared with τ_n, $m(t)$ remains equal to $m_0(V)$ and $n(t) \to n_0(V)$, so

$$J_{ml} \to J_{ss} = G_K n_0(V)(V - V_K) + G_{Ca} m_0(V)(V - V_{Ca}) + G_L(V - V_L)\,.$$

This is a steady state membrane current (plotted as the solid line in Figure 3), which forces the system back to its resting state at

$$[V, m, n] = [V_R, m_0(V_R), n_0(V_R)].$$

For sufficiently large values of τ_{n0}, therefore, the ML equation supports a nerve impulse with recovery.

At the dawn of the 21st century, electrophysiology is becoming a highly sophisticated experimental science, generating data on ever more intricate neural structures; thus, it will be of interest to consider using these five approaches (HH, ZF, MC, FN, and ML) to understand the dynamics of nerve impulse propagation on axonal and dendritic branching regions of real neurons.

ALWYN SCOTT

See also **Candle; FitzHugh–Nagumo equation; Hodgkin–Huxley equations; Markin–Chizmadzhev model; Myelinated nerves; Neuristor; Reaction-diffusion systems; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Brazier, M.A.B. 1961. *A History of Electrical Activity of the Brain*, London: Pitman

Cole, K.S. & Curtis, H.J. 1938. Electrical impedance of nerve during activity. *Nature*, 142: 209

Fall, C.P., Marland, E.S., Wagner, J.M. & Tyson, J.J. 2002. *Computational Cell Biology*, New York: Springer

Helmholtz, H. 1850. Messungen über den zeitlichen Verlauf der Zuckung animalischer Muskeln und die Fortpflanzungsgeschwindigkeit der Reizung in den Nerven. *Archiv für Anatomie und Physiologie*, 276–364

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology (London)*, 117: 500–544

Luther, R. 1906. Räumliche Fortpflanzung chemischer Reaktionen. *Zeitschrift fuer Elektrochemie* 12(32): 596–600 [English translation in *Journal of Chemical Education*, 64 (1987):740–742]

Markin, V.S. & Chizmadzhev, Yu.A. 1967. On the propagation and excitation for one model of a nerve fiber. *Biophysics*, 12: 1032–1040

Morris, C. & Lecar, H. 1981. Voltage oscillations in the barnacle giant muscle. *Biophysical Journal*, 71: 193–213

Nagumo, J., Arimoto, S. & Yoshizawa, S. 1962. An active impulse transmission line simulating nerve axon. *Proceedings of the Institute of Radio Engineers*, 50: 2061–2070

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Young, J.Z. 1951. *Doubt and Certainty in Science*, Oxford: Oxford University Press

NEUMANN BOUNDARY CONDITION

See **Partial differential equations, nonlinear**

NEURAL NETWORK MODELS

A neural network consists of units and connections corresponding to neurons and synapses in biological neural networks. We focus here on neural network models as a means to better understand the working principles of nervous systems, in particular the human brain. Quantitative modeling is today accepted as a very important tool in neuroscience that has potential to enable understanding of the very complex dynamical and nonlinear processes underlying the functioning of biological nervous systems. So-called artificial neural networks have been developed mainly for technical applications with little concern for biological plausibility and modeling (Haykin, 1998), and we will not consider them further here.

A computational model may help to explain experimental findings and to make new experimentally testable predictions. By now, most parts of the brain have been modeled, and many different types of models exist of a system, for example, for the hippocampus, an evolutionary old part of cortex important for memory and memory consolidation. When designing a neural network model of a particular system, the constituent neurons and their synaptic connections have to be represented with the desired biophysical detail, for example, as a set of coupled ordinary differential equations that can be solved numerically on a computer (Koch & Segev, 1998; Scott, 2002). The level of biological detail in the models studied ranges from networks of very simple threshold logic units connected by binary weights to networks comprising elaborate compartmental cell models with thousands of compartments and parameters. In addition to the signal transduction and transmission processes of the neuron such a detailed cell model may also represent intracellular processes such as biochemical second messenger cascades and calcium dynamics including, for example, diffusion.

In an accurate network model of a particular system, its constituent neuron types and proportions as well as their synaptic interactions have to be adequately represented. Available data about the system must be collected from literature and experiments and be entered as parameter values. A problem with this approach is that often some part of the information required by the model is lacking, such as the distribution of different kinds of ionic channels over the cell membrane or the details of synaptic plasticity dynamics (for example, augmentation and depression). Thus, some parameters may have to be indirectly inferred, and the models need to be tuned to fit experimental recordings, for example, of the shape of excitatory and inhibitory postsynaptic potentials or neuronal firing patterns at different levels of injected current. This introduces some uncertainty with regard to the validity of the model.

The input to the network from sources outside the model must also be represented in some way. Moreover, conditions and values measured in an *in vitro* preparation like a brain slice, a cell culture, or a piece of isolated spinal cord may differ from those of the intact in vivo system, and the latter may be unaccessible.

Nervous systems, other than the most simple ones, typically comprise a very large number of neurons. With today's computers it is feasible to simulate hundreds of thousands of compartments and millions of synapses at a reasonable level of biophysical detail. For large models the turnaround time for simulations, however, may be in the order of days, so parallel simulators are therefore useful. It is still beyond the capacity of today's supercomputers to handle full scale models of biological networks. A common practice is, therefore, to work with dramatically subsampled network models in which the number of neurons is reduced to a small fraction of those actually present in the target system. As a consequence, the number of input synaptic connections on cells in the model is also dramatically reduced. With cell models tuned to single cell data, for example, with adequate input resistances and thresholds, it will become necessary to compensate by exaggerating the synaptic conductances in order for the model to reproduce the activity seen in the real system. Having few and large synaptic interaction events in the system may, however, distort network dynamics, thus, making the model a poor quantitative representation of the actual system. Such effects should be born in mind when interpreting result from simulations using subsampled models.

Furthermore, one sometimes excludes from the model altogether neuron types known to exist in the actual biological system. For one reason or another a cell type may be considered unimportant for the questions addressed. One example is the neuroglia cells that are on average about ten times more numerous than neurons in the brain. They are thought to mainly serve the purpose of structural support and maintenance of the internal environment and are rarely included in network models. But one cannot entirely exclude that these cells in some situations subserve functions important for signal processing.

At the other extreme from networks of complex multi-compartmental model neurons are neural network models in which the biophysical detail has been reduced to a minimum. Such abstraction serves the important purpose of helping to pinpoint and elucidate the fundamental mechanisms behind the functioning of a complex system, which may enable further theoretical analysis of the phenomena under study.

A network model may be simplified in several different ways. For instance, the neurons can be modeled as point neurons (a single isopotential compartment) lacking geometric extent. Network units may use a graded output (with, for example, a sigmoid transfer function) representing an instantaneous firing frequency, or they may be spiking as real neurons. The former may, in fact, represent the average population activity in a cortical module like a minicolumn rather than an individual neuron. An integrate-and-fire model neuron is a bit more elaborate with a simple membrane dynamics and a spiking threshold (Gerstner, 1999). Whether or not temporal timing at the millisecond range in the spike trains of neurons is fundamental for the information processing in the brain or if a rate code is adequate for most purposes is presently a hotly debated issue. Yet another class of models is neural continuum field models that represent the neural structure as a continuous sheet of excitable neural tissue with a lattice connectivity (Bressloff et al., 2002).

In simplified network models, the details of synaptic transmission and plasticity may be replaced with weighted inputs and simple learning rules. Signal delays along axons are ignored, which may be reasonable at least for millimeter distances in small networks. Instead of an accurate representation of short- and long-term synaptic plasticity, one typically incorporates a correlation-based learning rule, for example, some form of Hebbian learning (Rolls & Treves, 1997). The adiabatic learning hypothesis (Caianiello, 1961) stating that synaptic weight changes occur on a much slower timescale than the neurodynamics itself, simplifies the model and network dynamics considerably. On the other hand, this assumption is now known to be invalid in, for instance, the neocortex where synaptic properties are modulated on a millisecond timescale.

A modeling strategy that has often proven fruitful is to develop a suite of models at different levels of abstraction for the same neuronal network. The most detailed model relates closely to the biological network under study, and the aim is to transform in a well-defined fashion to gradually more abstract descriptions, using the most abstract ones as the starting point for theoretical analysis.

Anders Lansner

See also **Artificial intelligence; Attractor neural network; Cell assemblies; Compartmental models; Hodgkin–Huxley equations; Integrate and fire neuron; McCulloch–Pitts network; Multiplex neuron; Nerve impulses; Neurons; Perceptron**

Further Reading

Bressloff, P.C., Cowan, J.D, Golubitsky, M., Thomas, P.J. & Wiener, M. 2002. What geometric visual hallucinations tell us about the visual cortex. *Neural Computation*, 14: 473– 491

Caianiello, E. 1961. Outline of a theory of thought processes and thinking machines. *Journal of Theoretical Biology*, 1: 204–235

Gerstner, W. 1999. Spiking neurons. In *Pulsed Neural Networks*, edited by W. Maass & C. Bishop, Cambridge, MA:MIT Press, pp. 3–53

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, Upper Saddle River, NJ: Prentice–Hall

Koch, C. & Segev, I. (editors). 1998. *Methods in Neuronal Modeling: From Ions to Networks*, Cambrigde, MA: MIT Press

Rolls, E. & Treves, A. 1997. *Neural Networks and Brain Function*, Oxford and New York: Oxford University Press

Scott, A.C. 2002. *Neuroscience. A Mathematical Primer*, Berlin and New York: Springer

NEURISTOR

Coined by Hewitt Crane in 1962, the term *neuristor* implies "the whole class of lines that exhibit attenuationless propagation with recovery" (Crane, 1962). In principal, this definition was intended to include active nerve fibers and even forest fires, but Crane's motivation was to overcome certain problems associated with the miniaturization of electronic circuits. Based upon a half-dozen reports prepared by Crane and his colleagues at the Stanford Research Institute since the late 1950s, this carefully written paper provides a window into the thinking of electrical engineers at the threshold of the integrated circuit revolution.

As the dimensions of conventional (transistor) computing circuits are greatly reduced, Crane reasoned, at least two design problems must be faced. First, the density of interconnections increases, eventually requiring "a dense set of interconnections at a point." Second, the resistance per unit length of interconnecting wires may become inconveniently large. (Crane pointed out that the resistance of a 1 μm copper wire is about a $1000\,\Omega/\text{in.}$) Both of these problems may be mitigated by basing miniaturized computer design upon an "active wire" in which energy is stored uniformly over the system and continuously dissipated by nonlinear traveling-wave impulses, rather than at discrete and isolated amplifying elements. Some indication that neuristor design is a promising strategy is offered by the fact that it was selected by evolution for the development of our biological brains.

Being uniform in the direction of propagation, Crane's active wire supports a traveling-wave impulse in which the rate of energy release is equal to its rate of dissipation. Upon impulse passage, the line remains inactive for a certain time (the refractory period), after which it is able to transmit a new impulse. Because an impulse cannot propagate through a refractory region, two impulses will destroy one another in a head-on collision, and impulses are not reflected from the end of a neuristor line.

If two interconnection junctions are included—both of which seem feasible from an engineering perspective—a neuristor system is logically complete, meaning that it can realize all possible Boolean switching functions. These two neuristor interconnections are as follows. (a) A *T* junction is shown in Figure 1(a). An impulse entering on one of the branches proceeds outward on the other branches. Such a junction is not difficult to realize, requiring merely that the strength of the incoming impulse divided by the number of outgoing branches is above the threshold of the outgoing branches. (b) An *R* junction is shown in Figure 1(b). In an *R* junction, the refractory (or inhibitory) variable of one line is shared with an adjacent line (over the shaded area), while the excitatory variable is not. Thus, individual impulses traveling from A to B (on the upper line) can block individual impulses traveling from D to C (on the lower line) and vice versa. These two junctions can be interconnected in a variety of useful ways,

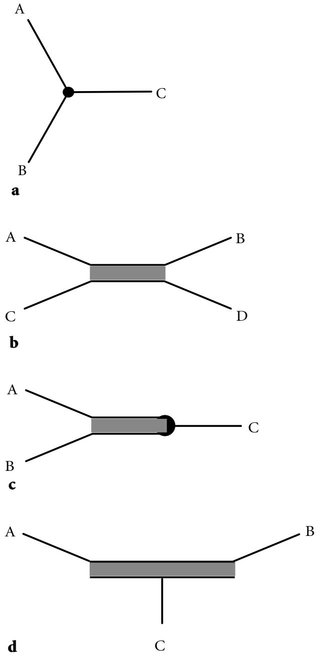

Figure 1. Some simple neuristor interconnections. (a) *T* junction. (b) *R* junction. (c) *T–R* junction. (d) Analog of a relay or transistor.

one being the *T–R* junction, shown in Figure 1(c). Here impulses incoming on line A will proceed through to C but will not be transmitted to B, effectively isolating the input lines (A and B) from each other.

In considering the switching possibilities of such neuristor systems, Boolean variables can be chosen in either of two ways, indicating the presence or absence of individual impulse or the presence or absence of impulse trains. From the latter perspective, the arrangement in Figure 1(d) can be viewed as equivalent to a relay or a transistor because an incoming impulse train on line C will block the transmission between A and B. As computing systems constructed from relays or transistors are known to be logically complete, it follows that neuristor systems are also logically complete.

The length of the refractory zone of an impulse is equal to the product of its velocity times the recovery interval. This is an important design parameter for a neuristor system, setting the scale for many functions. If a circular section of active line is used as a storage ring, for example, the circumference of this ring must be greater than the length of a refractory zone. Thus an insufficiently short refractory length is a major limitation on the degree to which a particular neuristor system can be miniaturized.

In his seminal paper, Crane offered several suggestions for neuristor realizations, taking advantage of the variety of interesting nonlinear diode structures that were being invented in the early 1960s—such as Esaki (tunnel) diodes and four-layer diodes. Also suggested was an ingenious combination of *T* and *R* junctions that allows two signal paths to cross without interference and without being lifted from a common substrate.

Although neuristors have been designed and fabricated in various laboratories (Beretovskii, 1963; Yoshizawa & Nagumo, 1964; Sato & Miyamoto, 1967; Parmentier, 1969; Scott, 1970; Reible & Scott, 1975; Nakajima et al., 1976), the neuristor strategy has not been important for the design of modern computing systems. Among other reasons for this failure must be counted the amazing progress in miniaturization of silicon metal-oxide-semiconductor transistors, which are now far smaller than the refractory length of any conceivable neuristor structure.

Beyond applications to electronic computing systems, however, the neuristor design concept may yet aid in understanding the behavior of neural systems, as was suggested by Crane in the early 1960s (Crane, 1964). Indeed, this possibility is more compelling today because dendritic fibers (in addition to axons) are now known to support action potentials and are, therefore, also neuristors in the original sense of the word (Stuart et al., 1999). Thus, possibilities arise for neuristor-like computations in the axonal and dendritic trees of real neurons (Scott, 2002). Neuroscientists studying the intricate networks of interwoven dendro-dendritic, currently being revealed by electron microscopy, may profit from a review of Crane's early work.

Alwyn Scott

See also **Multiplex neuron; Nerve impulses; Neurons**

Further Reading

Beretovskii, G.N. 1963. Study of single electric model of neuristor. *Radio Engineering and Electronic Physics*, 18: 1744–1751

Crane, H.D. 1962. Neuristor—a novel device and system concept. *Proceedings of the IRE*, 50: 2048–2060

Crane, H.D. 1964. Possibilities for signal processing in axon systems. In *Neural Theory and Modeling,* edited by R.F. Reiss. Stanford: Stanford University Press, pp. 138–153

Nakajima, K., Onodera, Y. & Ogawa, Y. 1976. Logic design of Josephson network. *Journal of Applied Physics*, 47: 1620–1627

Parmentier, R.D. 1969. Recoverable neuristor propagation on superconductive tunnel junction strip lines. *Solid-State Electronics*, 12: 287–297

Reible, S.A. & Scott, A.C. 1975. Pulse propagation on a superconductive neuristor. *Journal of Applied Physics*, 46: 4935–4945

Sato, R. & Miyamoto, H. 1967. Active transmission lines. *Electronics and Communications in Japan*, 50: 131–142
Scott, A.C. 1970. *Active and Nonlinear Wave Propagation in Electronics*, New York: Wiley
Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer
Stuart, G., Spruston, N. & Häusser, M. (editors). 1999. *Dendrites*, Oxford and New York: Oxford University Press
Yoshizawa, S. & Nagumo, J. 1964. A bistable distributed line. *Proceedings of the IEEE*, 52: 308

NEURONS

Nerve cells, or neurons, are typically comprised of a cell body, dendrites, and an axon. The typical cortical neuron depicted in Figure 1 has both an apical dendrite and several basal dendrites. Dendrites and the soma are the sites at which axon terminals from other neurons make contacts and provide stimulation to a neuron. These sites of contact between neurons are termed synapses. The axon emanates from the cell body and contacts other neurons at distances varying from 1.0 mm up to almost 1 m, depending upon the the type and location of the neuron in question. The ends of axons terminate very close to the dendritic and cell body membranes of other neurons to form synpases. When the cell body of a neuron is sufficiently depolarized by its inputs, it reaches the threshold for action potential (or spike) generation, and one or more spikes are triggered. These spikes then propagate along the axon at speeds ranging from less than 1.0 ms^{-1} up to almost 100 ms^{-1}.

Neurons receive both excitatory (depolarizing) and inhibitory (hyperpolarizing) stimulation from the axon terminals that synapse onto them, with a typical cortical neuron having about 10^3–10^4 synapses. When an action potential reaches a synapse, it causes the release of neurotransmitter molecules that rapidly diffuse across the very small extracellular space to the membrane of the postsynaptic cell's dendrite or soma. There they bind to receptor molecules and cause ion channels to open resulting in either depolarization (excitatory synapse) or hyperpolarization (inhibitory synapse). These potential changes propagate down the dendrites to the cell body, where they are combined approximately linearly. If this net potential change depolarizes the cell body past a threshold, the neuron will fire one or more spikes. These spikes in turn propagate along the axon of this cell, and the process repeats itself.

Nonlinear Dynamics of Action Potential Generation

In mathematical terms, a series of action potentials is a limit cycle oscillation in the neural state space. There are several distinct dynamical patterns of action potential generation that have been observed, and these will be characterized in terms of their dynamical foundations.

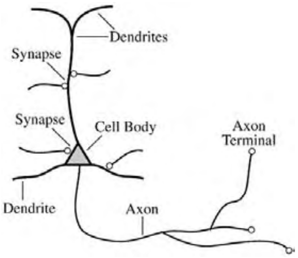

Figure 1. Schematic diagram of a typical neuron in the neocortex. The longest dendrites are between 1–3 mm in length, while the axon can be as short as a few mm or as long as 1.0 m.

As demonstrated by Hodgkin and Huxley (1952) in work that led to the Nobel Prize in 1963, the dynamics of spike generation is based upon two ionic currents: a sodium (Na^+) current that rapidly depolarizes the neuron followed by a slower potassium (K^+) current that repolarizes the neuron. Each current I_j is described by Ohm's law written as the product of a conductance (reciprocal of resistance) g_j and a voltage. For each ion, the voltage term is given by the difference between the membrane potential V and the equilibrium potential for the ion in question, E_j

$$I_j = g_j(V - E_j) \tag{1}$$

Due to the fact that the neural cell membrane functions as a capacitor, the membrane potential V is described by a differential equation of the form

$$\mathrm{C}\frac{\mathrm{d}V}{\mathrm{d}t} = g_{\mathrm{Na}}(V)\,(V - E_{\mathrm{Na}}) + g_{\mathrm{K}}(V)\,(V - E_{\mathrm{K}}) + I_{\mathrm{ext}}, \tag{2}$$

where I_{ext} is an external stimulating current (or a synaptic input). For typical cortical neurons, $E_{\mathrm{Na}} = 55\,\mathrm{mV}$ and $E_{\mathrm{K}} = -95\,\mathrm{mV}$. This equation would be linear except for one crucial observation: the two conductances $g_{\mathrm{Na}}(V)$ and $g_{\mathrm{K}}(V)$ are functions of V. This biophysical discovery by Hodgkin & Huxley (1952) means that ion conductances change with the voltage, thereby generating both positive and negative nonlinear feedback that produces the action potential.

A simple set of equations with normalized variables (Wilson, 1999a) can be used to elucidate the essential dynamics of Equation (2):

$$\frac{\mathrm{d}V}{\mathrm{d}t} = 4\left(V^2 - \frac{V}{10}\right)(V-1) + R\left(V + \frac{1}{5}\right) + I_{\mathrm{ext}},$$

$$5\frac{\mathrm{d}R}{\mathrm{d}t} = R + 3V^2 \tag{3}$$

The term $(V^2 - V/10)$ on the right in the first equation represents the V dependence of the Na^+

conductance as a quadratic function, while the variable R (Recovery variable) represents the K^+ conductance governed by the second equation. Note that the time constant for the second equation is 5 times slower than that in the first equation. Normalization has shifted the resting potential from -70 mV to zero and has set $E_{Na} = 1$ here, which corresponds to 125 mV above the resting potential. Similarly, $E_K = -\frac{1}{5}$, or 25 mV below the resting potential. Finally, the unit of time here corresponds to 0.10 ms to produce spike durations of 1.0 ms. These simplifications make the mathematical analysis significantly easier. A version of these equations with parameter values optimized to describe human neocortical excitatory and inhibitory neurons may be found elsewhere (Wilson, 1999b).

For $I_{ext} = 0$, Equation (3) has three steady states at $V = 0, R = 0$ (resting state); $V = \frac{1}{7}, R = \frac{3}{49}$; and $V = \frac{2}{5}$, $R = \frac{12}{25}$. In order, these are an asymptotically stable node, a saddle point, and an unstable spiral point. The spike firing threshold occurs for $I_{ext} = 0.012$, where the saddle point and node coalesce in a saddle-node bifurcation to a limit cycle. This permits spike firing at arbitrarily low rates for I_{ext} just above threshold, which is characteristic of human and mammalian cortical neurons (Wilson, 1999a). A spike train for $I_{ext} = 0.013$ is depicted in Figure 2a. Above threshold, spike rate increases monotonically as a function of I_{ext}. Neurons that begin firing at arbitrarily low rates due to a saddle-node bifurcation are known as Class I neurons.

A different form of dynamics characterizes Class II neurons. The simplest example is obtained by replacing the $\mathrm{d}R/\mathrm{d}t$ equation above with

$$5\frac{\mathrm{d}R}{\mathrm{d}t} = R + 2V. \qquad (4)$$

For $I_{ext} = 0$, the equations now have a single steady state $V = 0$, $R = 0$. It can be shown that these neural equations undergo a subcritical Hopf bifurcation (Wilson, 1999a) to spiking at $I_{ext} = 0.062$. In this case firing begins at a relatively high spike rate, as arbitrarily low rates are precluded by the nature of the bifurcation to spiking. The original Hodgkin–Huxley (1952) equations, in fact, describe a Class II neuron (the giant axon of the squid).

Neurons in the cortex of humans and other mammals are Class I neurons. Thus, they begin firing at arbitrarily low spike rates (less than one spike per second), and this provides a greatly expanded dynamic range for encoding stimulus intensity into spike frequency. In addition, excitatory cortical neurons (but not most inhibitory neurons) typically have several additional currents that can produce even more complex spiking dynamics. For example, addition of a slow Ca^{2++} current and an even slower Ca^{2++}-driven K^+ current results in a neuron that fires bursts of spikes (Wilson, 1999a,b), as illustrated in Figure 2b.

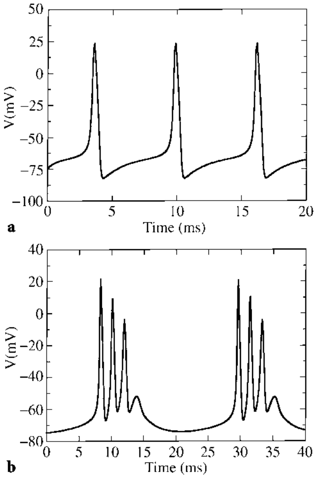

Figure 2. Spikes generated by neural equations. (a) Periodic spike train generated by Equation (3). (b) Pattern of spike bursts generated by a more complex cortical neuron with additional currents. Both graphs have been transformed back from the normalized form described in the text to reflect the mV range and ms time range actually encountered with cortical neurons.

Conclusions

Action potential dynamics are governed by two processes. Once threshold is reached, a rapid influx of Na^+ ions results in depolarization of the neuron causing the upswing of the spike. This process is described by the first term in the $\mathrm{d}V/\mathrm{d}t$ equation in (3) above. Following this, the variable R increases, permitting K^+ to pass out of the neuron, thus hyperpolarizing it back toward its resting potential. The limit cycle is generated because the recovery variable R operates on a slower time scale than the very rapid Na^+ depolarization. There is also an inactivation of the Na^+ ion current that contributes to termination of the spike (Hodgkin & Huxley, 1952), but it is not essential to the dynamics of spike generation.

Virtually all brain function involves a dynamical interaction between excitatory and inhibitory neurons. For example, short-term memory circuits involve groups of neurons that are reciprocally interconnected by excitatory synapses, which enables them to continue firing following the cessation of stimulation. This ongoing neural activity makes up the short-term memory store. Inhibition is necessary both to shut off short-term memory activity and to prevent it from spreading to activate other neurons. Indeed, when there is too little inhibition in a brain area due to an imbalance or injury, the spread of excitation may generate epileptic seizures. Many other examples of excitatory and inhibitory neural circuits may be found elsewhere (Wilson, 1999a; Dayan & Abbott, 2001).

HUGH R. WILSON

See also **Hodgkin–Huxley equations; Integrate and fire neuron; Nerve impulses**

Further Reading

Dayan, P. & Abbott, L.F. 2001. *Theoretical Neuroscience*, Cambridge, MA: MIT Press

Hille, B. 1992. *Ionic Channels of Excitable Membranes*, Sunderland, MA: Sinauer

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117: 500–544

Koch, C. 1999. *Biophysics of Computation: Information Processing in Single Neurons*, Oxford and New York: Oxford University Press

Wilson, H.R. 1999a. *Spikes, Decisions, and Actions: Dynamical Foundations of Neuroscience*, Oxford and New York: Oxford University Press

Wilson, H.R. 1999b. Simplified dynamics of human and mammalian neocortical neurons. *Journal of Theoretical Biology*, 200: 375–388

NEWELL–WHITEHEAD–SEGEL EQUATION

See **Complex Ginzburg–Landau equation**

NEWTON'S LAWS OF MOTION

Isaac Newton's treatise *Philosophiae Naturalis Principia Mathematica* (Mathematical Principles of Natural Philosophy), often simply referred to as the *Principia*, was completed in May 1686. Publication and printing were overseen by the astronomer Edmund Halley, and the work became available to the public in 1687. In this work Newton sets out, in effect, a template for classical mathematical physics, the dynamics of particles and rigid bodies in particular. The text starts by defining certain basic quantities such as mass, momentum, inertia, force, and acceleration. Once these have been defined and some of their basic properties described, Newton states his "axioms" or "laws of motion" thus:

> *Law I*: Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed upon it;
> *Law II*: The change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed;
> *Law III*: To every action there is always opposed an equal reaction: or, the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

(We have quoted the established 1729 translation from the original Latin to English by Andrew Motte.) In modern terms the first law, also known as the law of inertia, and already realized by Galileo, states that uniform translational motion is a "natural" state of motion for a particle or body requiring no cause or outside agent to maintain it. (This was an important departure from Aristotelian physics, which held that all motion required an explanation.) Not so with accelerated motion, according to the second law. If a change in the velocity of motion is to come about, it requires the action of a force acting on the particle or body. Finally, in the third law we have the requirement that if a body, A, is acted upon by a force due to another body, B, then B is subject to a force of the same magnitude but of opposite direction from its interaction with A.

Forces

Newtonian mechanics, then, operates with purely kinematic entities, such as position, velocity, and acceleration, and with a set of new, dynamical entities called forces. Newton's laws of motion instruct us to seek the cause of acceleration, a kinematic quantity, in the total force acting on a particle or body. Force is a dynamic quantity. The mass of the particle or body appears as the constant of proportionality in the relation between force (the cause) and acceleration (the effect). Like accelerations, forces are vectors. If two or more forces act simultaneously, they add vectorially and their resultant gives both the correct magnitude and the correct direction of the net force. Many well-known results about levers, systems with suspended masses, and various simple mechanical machines were immediately subsumed under Newtonian mechanics simply by recognizing that forces behave like vectors upon superposition.

Not only did Newton establish a framework for dynamics—and for much of physics—with his three laws, he provided also a new law of Nature for the attractive force of gravity acting between any two bodies. This development was, in some sense, extraneous to the three laws of motion but without it, the full force of Newton's dynamics might not have been appreciated and embraced. Newton's expression for the force acting on a particle, 1, of mass m_1 located at $\boldsymbol{x}_1$ due to a particle, 2, of mass m_2, located at $\boldsymbol{x}_2$, is

$$\boldsymbol{F}_{12} = Gm_1 m_2 \boldsymbol{e}_{12}/r^2. \qquad (1)$$

Here r is the distance between the two particles, $r = |\boldsymbol{x}_1 - \boldsymbol{x}_2|$ and $\boldsymbol{e}_{12}$ is the unit vector pointing from the location of particle "1" to the location of particle "2", that is, $\boldsymbol{e}_{12} = (\boldsymbol{x}_2 - \boldsymbol{x}_1)/r$. The coefficient G, now known as Newton's universal gravitational constant, has the approximate value $G \approx 6.672 \times 10^{-8}\,\mathrm{cm}^3\,\mathrm{g}^{-1}\,\mathrm{s}^{-2}$.

The first measurements of G were performed by Henry Cavendish in 1798 using a torsion balance. Expression (1) is fully consistent with Newton's third law, which demands that $\boldsymbol{F}_{12} = -\boldsymbol{F}_{21}$. Force (1) represents action-at-a-distance in the sense that it does not result from direct contact between contiguous bodies or particles. Newton was severely criticized by his contemporaries for this notion, which they viewed as fanciful and unsubstantiated.

In modern terminology, in the *Principia* Newton accomplished (among many other things) integration of the two-body problem of celestial mechanics, that is, the following system of differential equations, which result by combining Newton's law of gravitation with his second law of motion:

$$m_1\,\mathrm{d}^2\boldsymbol{x}_1/\mathrm{d}t^2 = Gm_1m_2(\boldsymbol{x}_2 - \boldsymbol{x}_1)/|\boldsymbol{x}_2 - \boldsymbol{x}_1|^3, \quad (2a)$$

$$m_2\mathrm{d}^2\boldsymbol{x}_2/\mathrm{d}t^2 = Gm_1m_2(\boldsymbol{x}_1 - \boldsymbol{x}_2)/|\boldsymbol{x}_1 - \boldsymbol{x}_2|^3. \quad (2b)$$

Because of Newton's third law, addition of Equations (2) yields

$$m_1\,\mathrm{d}^2\boldsymbol{x}_1/\mathrm{d}t^2 + m_2\,\mathrm{d}^2\boldsymbol{x}_2/\mathrm{d}t^2 = 0, \quad (3)$$

which shows that the total momentum of the system, given by

$$\boldsymbol{P} = m_1\mathrm{d}\boldsymbol{x}_1/\mathrm{d}t + m_2\,\mathrm{d}\boldsymbol{x}_2/\mathrm{d}t, \quad (4)$$

is a constant of the motion. Thus, the center of mass of the two particles,

$$\boldsymbol{R} = (m_1\boldsymbol{x}_1 + m_2\boldsymbol{x}_2)/(m_1 + m_2), \quad (5)$$

moves with constant velocity $\boldsymbol{P}/M$ through space, where $M = m_1 + m_2$ is the total mass of the two-body system.

On the other hand, subtracting (2a) from (2b) gives

$$m\mathrm{d}^2\boldsymbol{r}/\mathrm{d}t^2 = -Gm_1m_2\boldsymbol{r}/r^3, \quad (6)$$

where $\boldsymbol{r} = \boldsymbol{x}_2 - \boldsymbol{x}_1$, $r = |\boldsymbol{r}|$, $M = m_1 + m_2$ as before, and $m = m_1m_2/(m_1 + m_2)$ is called the reduced mass. Equation (5) tells us that the problem of the relative motion of the two particles is equivalent to solving for the motion of a single (fictitious) particle of mass m in the same field of force centered at the origin of coordinates. Such a fixed force field pointing toward a given point in space (here chosen as the origin of coordinates) is called a central force.

Once (6) is solved, the original positions of the two gravitationally attracting particles may be reconstructed from the formulae

$$\boldsymbol{x}_1 = \boldsymbol{R} - m_2\boldsymbol{r}/M, \quad \boldsymbol{x}_2 = \boldsymbol{R} + m_1\boldsymbol{r}/M. \quad (7)$$

The differential equation for the vector function $\boldsymbol{r}$ is nonlinear because the magnitude of the force falls off with the square of the distance. Many force laws of interest in applications have the property that they are nonlinear functions of the positions of the constituent particles of the system. The force law for small extensions of a high-quality spring, known as Hooke's law, which states that the force is directly proportional to the deviation from equilibrium, is an important and notable exception.

Newton showed in the *Principia* that the general solution of (6) is that $\boldsymbol{r}$ traces out a conic section (ellipse or hyperbola, with a parabola as the "cross-over" possibility) corresponding to bounded or unbounded relative motions of the two original particles. The bounded motions apply to planets orbiting the Sun or the Moon orbiting the Earth. The effects of bodies farther away are to be treated subsequently by adding perturbations to the above analysis. Unbounded motions are realized by comets. Newton further considered the effects of the finite extension of the attracting body, for example, Earth's finite size relative to the Moon's orbit, and the effect of the deformation of the attracting body due to tidal forces.

HASSAN AREF

See also **Celestial mechanics; Determinism; *N*-body problem**

Further Reading

Chandrasekhar, S. 1995. *Newton's Principia for the Common Reader*, Oxford: Clarendon Press and New York: Oxford University Press

Newton, I. 1934. *Mathematical Principles of Natural Philosophy and System of the World*, translated by A. Motte (1729), edited by F. Cajori, 2 vols, Berkeley: University of California Press

NEWTON'S METHOD

See **Numerical methods**

NODES OF RANVIER

See **Myelinated nerves**

NOISE (WHITE, COLORED ETC.)

See **Stochastic processes**

NONATTRACTING CHAOTIC SETS

See **Chaotic dynamics; Invariant manifolds and sets**

NONAUTONOMOUS SYSTEMS

See **Phase space**

NONEQUILIBRIUM STATISTICAL MECHANICS

Nonequilibrium statistical mechanics aims to describe the behavior of large systems of particles removed from the state of thermodynamic equilibrium, in terms of the properties of the individual constituents and their interactions, as provided by the laws of classical and quantum mechanics. Such systems give rise to irreversible behavior in the form of an approach to thermodynamic equilibrium in the absence of permanent constraints (isolated systems), or to a stationary nonequilibrium state in systems under constraint (referred to as thermostatted systems).

The fundamental problem of non-equilibrium statistical mechanics is to reconcile irreversible behavior at the macroscopic level with the reversible character of the underlying microscopic laws of mechanics. For a long time it was thought that irreversibility could not be understood entirely from mechanics, an idea that led to the systematic introduction of probabilistic ideas in the traditional deterministic description. As a result, the laws of statistical mechanics have frequently been regarded as analogs of the law of large numbers and similar universal laws of probability and statistics, independent of any explicit reference to the nature of the underlying dynamics. Since the 1980s, one has witnessed a change of perspective following the realization that the trajectories of individual particles in an N-body system are typically chaotic.

Boltzmann's Kinetic Theory. Ergodic and Mixing Hypotheses

In 1872, Ludwig Boltzmann derived by what appeared to be completely mechanical arguments his famous kinetic equation for dilute gases (Brush, 1965, 1966). This equation features the time-dependent probability distribution function f, of position $\boldsymbol{r}$ and velocity $\boldsymbol{v}$, for a particle in the gas and allows one to reproduce in a very satisfactory way the transport and flow properties of the system. The kinetic theory culminates in the derivation of the H-theorem, whereby in a homogeneous system the functional $H = k \int f \ln f \, \mathrm{d}\boldsymbol{v}$ (k being the Boltzmann constant) decreases monotonously until f reaches its equilibrium form given by the Maxwell–Boltzmann distribution. The H-theorem should provide, then, a microscopic justification of the second law of thermodynamics. This claim prompted rather negative reactions, crystallized in the famous reversibility (Loschmidt's) and recurrence (Zermelo's) paradoxes. Boltzmann was unable to fully refute these objections since, as he himself recognized, his derivation makes use of a heuristic probabilistic assumption—the *Stosszahlansatz* or the assumption of molecular chaos—which allowed him to express approximately the rate at which binary collisions are taking place only in terms of the one-particle probability density f.

In trying to answer his critics Boltzmann enunciated the ergodic hypothesis, which eventually became a key concept in the entire field of statistical mechanics. Specifically, Boltzmann suggested that (a) in an isolated many-body system the overwhelming part of the phase space consists of regions where the macroscopic properties are very close to the equilibrium properties, (b) the system's trajectory will spend equal times in phase space regions of equal extent, and (c) the macroscopic properties of the system will essentially be constant throughout the allowable part of phase space and will coincide with the long-time averages of the corresponding microscopic quantities over the phase space trajectory. These statements were later completed by the equally ground-breaking discovery by Josiah W. Gibbs of the concept of mixing: the dynamical evolution of an isolated system initially occupying a limited phase space region compatible with some prescribed values of its macroscopic observables will lead it eventually to occupy uniformly (at least in a coarse-grained sense) the entire phase space available.

In their original forms, the ergodic hypothesis and the stronger mixing hypothesis had the serious drawback of relying to a considerable extent on the coarse-grained way one observes the system. Modern ergodic theory starts with the work of Henri Poincaré and George Birkhoff (Arnol'd & Avez, 1968), who were able to relate ergodicity to certain well-defined properties of the underlying deterministic evolution laws. Two examples are provided by Birkhoff's theorem on the existence of the limit $1/T \int_0^{\mathrm{T}} A(t)\, \mathrm{d}t$ as $T \to \infty$ of an integrable phase space function A provided that the motion remains bounded in phase space and by Poincaré's theorem that for systems satisfying suitable resonance conditions there exist no invariants other than total energy that are analytic in some parameter. Since the 1950s, a great deal of effort has been devoted to prove whether a system described by a given Hamiltonian will or will not fulfill these properties and to what extent these properties bear a clear-cut relationship with the type of dynamics going on in phase space. An early (negative) result of considerable historical importance was obtained by Enrico Fermi, John R. Pasta, and Stanislaw Ulam (1955), who showed that in a system of coupled nonlinear oscillators energy may remain localized rather than be equipartitioned among the individual degrees of freedom. At the other extreme, one finds Yakov Sinai's result (Sinai, 1970) on the ergodic and mixing behavior of a system of hard spheres. This work signaled the beginning of a series of developments aimed at relating the foundations of non-equilibrium statistical mechanics to the instability of motion of large classes of non-integrable dynamical systems giving rise to sensitivity to initial conditions and to deterministic chaos, known to be generic since the work of Andrei Kolmogorov.

Generalized Kinetic Theories

In parallel, and largely independently of progress in ergodic theory, the need to go beyond the assumptions underlying Boltzmann's equation became a central preoccupation from the mid-1940s. Three major attempts along this line were initiated by Nikolai Bogolubov, Leon Van Hove, and Ilya Prigogine and his colleagues (see, e.g., Balescu, 1975). Their starting point was a systematic perturbation expansion of the Liouville equation and its quantum counterpart for the N-particle probability distribution or density matrix, or the hierarchy of equations for the reduced n-particle ($n = 1, \ldots, N$) distribution functions obtained by integrating the Liouville equation over the $N - n$ particles, known as the BBGKY hierarchy. This procedure led as a first step to exact, formal non-Markovian equations of the form

$$\frac{\partial \rho_s}{\partial t} + K\rho_s(t) = \int_0^t \mathrm{d}\tau\, G(t-\tau)\rho_s(\tau) + D(t), \quad (1)$$

where ρ_s is a reduced distribution, K stands for the contribution of the mean field, G is a memory kernel, and $D(t)$ depends on the initial correlations in the subspace of phase space complementary to the one of ρ_s. Closed kinetic equations for reduced probability densities could then be obtained under certain assumptions, linked to first principles in a more clear-cut way than Boltzmann's *Stosszahlansatz*: Bogolubov's ansatz of higher-order distributions becoming functionals of the one-particle one, Prigogine's ansatz on initial correlations, van Hove's random phase approximation. But even when these conditions are satisfied, it has so far proved impossible to establish a general H-theorem for the corresponding equations. Still, generalized kinetic equations have been at the foundation of spectacular progress in such fields as the study of dense fluids and plasmas and the microscopic theory of transport coefficients in the linear range of irreversible phenomena close to equilibrium.

Microscopic Chaos and Nonequilibrium Statistic Mechanics

Generalized kinetic theories rest on approximations whose validity is difficult to assess. Furthermore, there is no explicit link between the structure of the kinetic equations and the nature of the microscopic dynamics in phase space. In view of the fundamental importance and the ubiquity of irreversibility in the natural world, it would certainly be desirable to arrive at a description free of both these limitations. This has been achieved since the 1980s by the cross-fertilization between dynamical systems, nonequilibrium statistical mechanics, and microscopic simulation techniques.

Mapping to a Markov Process

A first series of attempts pertains to the class of strongly unstable systems known as Kolmogorov flows, in which each phase space point lies at the intersection of stable and unstable manifolds. It takes advantage of the existence in such systems of special phase space partitions—the Markov partitions—whose boundaries remain invariant under the dynamics, each partition cell being mapped at successive time steps into a union of partition cells. If the operator projecting the full phase space dynamics on such a partition commutes with the Liouville operator, then the Liouville equation can be mapped into an exact Markovian equation exhibiting an H-theorem. This provides a rigorous microscopic basis of coarse-graining (Penrose, 1970; Nicolis et al., 1991). The mapping is, however, not one-to-one as the projection operator is not invertible reflecting the loss of information associated with this process. Ilya Prigogine, Baidyanath Misra, and Maurice Courbage (Prigogine, 1980) succeeded in constructing a non-unitary, invertible, transformation operator Λ, which transforms the unitary evolution operator U_t (essentially the exponential of the Liouville operator) into a Markov semigroup W_t

$$W_t = \Lambda U_t \Lambda^{-1} \quad (2)$$

The idea of non-unitary transformations has also been extended to more general classes of systems, such as non-integrable systems generating Poincaré resonances.

Escape Rate Formalism and Deterministic Thermostats

A second line of approach aims to express transport coefficients and other macroscopic level properties including entropy production in terms of the quantifiers of the microscopic chaos prevailing at the level of individual particle trajectories (in a classical setting). Two different methodologies have been developed.

The escape rate formalism (Gaspard, 1998; Dorfman, 1999). In this formalism, transport in Lorentz gas type systems is linked to the escape from a fractal repellor formed by trajectories for which a certain microscopic quantity associated with the macroscopic flux of interest remains confined in phase space. The corresponding transport coefficients are then expressed in terms of the Lyapunov exponents of the repellor and its Kolmogorov–Sinai entropy or its fractal dimension. No boundary conditions need to be imposed. Furthermore, the hydrodynamic modes associated with the process are computed as generalized eigenmodes of the Liouvillian and turn out to be fractal distributions.

Deterministic thermostats. The question here is how to express in purely mechanical terms a constraint maintaining a system away from equilibrium. One of the most popular ways to achieve this is to augment (in a classical system) the Hamiltonian dynamics by dynamical friction terms preserving the time-reversal symmetry (Hoover, 1999). In this way, one obtains a dissipative dynamical system where the rate of phase space volume contraction and the entropy production are shown to be related to the sum of the Lyapunov exponents. The non-equilibrium steady states generated in this formalism are fractal attractors, contrary to the equilibrium states that extend over the entire phase space available. Furthermore, they are believed to satisfy an interesting fluctuation theorem (Gallavotti &

Cohen, 1995) expressing the probability of fluctuations associated with particles moving in a direction opposite to the applied constraint.

Statistical Mechanics of Dynamical Systems

Yakov Sinai, David Ruelle, and Rufus Bowen (see, for example, Ruelle, 1999) have shown that the long-time behavior of a large class of chaotic dynamical systems is determined by an invariant probability measure characterized by a variational principle, thereby establishing a link with ergodic theory and equilibrium statistical mechanics. A major difference is that contrary to equilibrium measures, Sinai–Ruelle–Bowen (SRB) measures are smooth only along the expanding directions while they possess a fractal structure along the contracting ones. This provides a rationale for characterizing the fractal character of the non-equilibrium steady states generated by deterministic thermostats. Furthermore, it has been the starting point of a series of developments aimed at characterizing low-order deterministic dynamical systems showing complex behavior through probability densities and the spectral properties of their evolution operators, such as the Liouville or the Frobenius–Perron operators (Lasota & Mackey, 1985). In a different vein, the transport induced by the overlap of resonances arising, in particular, around a separatrix has been investigated (Lichtenberg & Lieberman, 1983; Balescu, 1997). These approaches have provided insights to questions motivated by non-equilibrium statistical mechanics that had remained unsolved for a long time, owing to the formidable difficulties arising from the presence of a large number of interacting particles.

Non-equilibrium Statistical Mechanics as a Tool to Understand Cooperative Behavior in Complex Systems

Non-equilibrium statistical mechanics has been a source of inspiration providing interesting ways to analyze complex nonlinear systems arising in chemistry, fluid mechanics, biology, or even finance and sociology. One of the earliest applications was a mesoscopic approach describing the dynamics of fluctuations of the macroscopic observables of such systems around a reference state, using master equation or Langevin–Fokker–Planck equation descriptions (Nicolis & Prigogine, 1977; Haken, 1977). It led to a theory of nonequilibrium phase transitions showing how macroscopic level bifurcation phenomena are reflected at the microscopic level through the anomalous behavior of the non-equilibrium fluctuations. When applied to nanoscale systems, this approach provides interesting insights on energy transduction by non-equilibrium devices such as biological macromolecules, referred to as molecular motors (Astumian, 1997).

Finally, the formalism of non-equilibrium statistical mechanics is well suited to studying the cooperative behavior of interacting multi-agent systems involved in, for instance, food webs and other networks of connected elements, finance, or social phenomena. In this setting, the relevance of power laws describing the statistical behavior of certain systems has attracted considerable interest, the idea being that such laws reflect well the ability of complex systems to show adaptive behavior and to optimize information flows (Albert & Barabasi, 2002). The dynamical origin of these laws and the role of non-equilibrium constraints remain major open questions.

G. Nicolis

See also **Chaotic dynamics; Emergence; Ergodic theory; Fermi–Pasta–Ulam oscillator chain; Markov partitions; Recurrence; Sinai–Ruelle–Bowen measures; Synergetics**

Further Reading

Albert, R. & Barabasi, A.S. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74: 47–97

Arnol'd, V. & Avez, A. 1968. *Ergodic Problems of Classical Mechanics*, New York: Benjamin

Astumian, R.D. 1997. Thermodynamics and kinetics of a Brownian motor. *Science*, 276: 917–922

Balescu, R. 1975. *Equilibrium and Nonequilibrium Statistical Mechanics*, New York: Wiley

Balescu, R. 1997. *Statistical Dynamics*, London: Imperial College Press

Brush, S. 1965, 1966. *Kinetic Theory*, vols. I and II. London: Pergamon Press

Dorfman, J.R. 1999. *An Introduction to Chaos in Nonequilibrium Statistical Mechanics*, Cambridge and New York: Cambridge University Press

Fermi, E., Pasta, J.R. & Ulam S. 1955. *Studies of nonlinear problems*. Los Alamos Scientific Laboratory, Report NoLA-1940

Gallavotti, G. & Cohen, E.G.D. 1995. Dynamical ensembles in nonequilibrium statistical mechanics. *Physics Review Letters*, 74: 2694–2697

Gaspard, P. 1998. *Chaos, Scattering and Statistical Mechanics*, Cambridge and New York: Cambridge University Press

Haken, H. 1977. *Synergetics*, Berlin: Springer

Hoover, W.G. 1999. *Time Reversibility, Computer Simulation, and Chaos*, Singapore: World Scientific

Lasota, A. & Mackey, M. 1985. *Probabilistic Properties of Deterministic Systems*, Cambridge and New York: Cambridge University Press

Lichtenberg, A.J. & Lieberman, M.A. 1983. *Regular and Stochastic Motion*, Berlin: Springer

Nicolis, G., Martinez, S. & Tirapegui, E. 1991. Finite coarse-graining and Chapman–Kolmogorov equation in conservative dynamical systems. *Chaos, Solitons and Fractals*, 1: 25–37

Nicolis, G. & Prigogine, I. 1977. *Self-organization in Nonequilibrium Systems*, New York: Wiley

Penrose, O. 1970. *Foundations of Statistical Mechanics*, Oxford: Pergamon

Prigogine, I. 1980. *From Being to Becoming*, San Fransisco: Freeman

Ruelle, D. 1999. Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. *Journal of Statistical Physics*, 95: 393–468

Sinai, Ya. 1970. Dynamical systems with elastic reflections. *Russian Mathematical Surveys*, 25: 137–189

NONERGODICITY

See **Ergodic theory**

NONINTEGRABLE LATTICES

See **Integrable lattices**

NONLINEAR ACOUSTICS

As an area of physics and mechanics concerned with sound waves of finite amplitude, theoretical nonlinear acoustics stems from advances in classical fluid mechanics made in the 19th century. It was developed originally as a weakly nonlinear limit of gas dynamics and included the study of simple waves, weak shock waves, and their interactions. Nonlinear acoustics as a separate branch of science was formed mainly during and after World War II when, on the one hand, military applications of underwater acoustics became important, and on the other, more powerful sources of sound were developed.

The nature of finite-amplitude acoustic wave propagation in fluids depends on the value of the acoustic Reynolds number $\mathrm{Re} = \beta u_0 \lambda/\nu$, where u_0 is a characteristic amplitude of the particle velocity, λ is the wavelength, ν the kinematic viscosity (or a combination of viscosity and thermal conductivity), and β is the coefficient of nonlinearity (1.2 for air, 3.5 for water). For $\mathrm{Re} \leq 1$, the wave attenuates before significant nonlinear distortion accumulates. For $\mathrm{Re} \gg 1$, higher harmonics are generated, and the wave profile may undergo radical changes.

The physical process underlying waveform distortion for a plane (one-dimensional) traveling wave is the simple-wave characteristic relation $\mathrm{d}x/\mathrm{d}t = c_0 + \beta u$ for the propagation speed of individual points on the waveform, where c_0 is the small-signal (infinitesimal-amplitude) sound speed. Positive values of the particle velocity u (or sound pressure) in the waveform advance on the zero crossings propagating at speed c_0, and negative values recede. For $\mathrm{Re} \gg 1$, this process leads to waveform steepening up to the "breaking" point, beyond which this propagation law predicts a multivalued waveform (see Figure 1). Whereas water waves break, sound waves cannot, and instead an abrupt jump in the wave amplitude, referred to as a shock front, is formed. Shocks substantially increase attenuation of the wave, and their thickness is proportional to a combination of viscosity and thermal conductivity coefficients.

The process just discussed is well described for a plane traveling wave by the Burgers equation

$$\frac{\partial u}{\partial t} + (c_0 + \beta u)\frac{\partial u}{\partial x} = \delta \frac{\partial^2 u}{\partial x^2}, \tag{1}$$

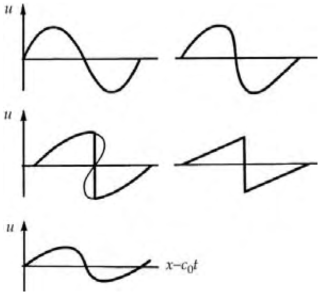

Figure 1. Schematic of the evolution of an initially sinusoidal nonlinear acoustic wave. Each plot corresponds to the one-period waveform as a function of coordinate at the successive time moments.

where δ is a dissipation coefficient that accounts for viscosity and thermal conductivity and equals $2\nu/3$ for shear viscosity alone. The left-hand side accounts for the finite-amplitude propagation speed given above. The Reynolds number, given above, characterizes the ratio of the nonlinear term to the dissipation term. Equation (1) is expressed in a form convenient for initial value problems, when an initial spatial waveform is prescribed. An alternative form of Equation (1), with the roles of x and t reversed, is used in signaling problems, in which a time waveform is prescribed on a boundary. Similar processes occur in spherical waves, such as those at sufficient distances from explosions, where the nonlinearity is weak.

An important role in the development of modern nonlinear acoustics has been played by the parametric array conceived by Westervelt in the USA and Zverev and Kalachev in Russia in the early 1960s. In the parametric array, two acoustic beams close in frequency interact nonlinearly and create a virtual antenna that radiates a secondary, low-frequency beam. Notwithstanding their low efficiency, these antennas prove useful in applications (such as sea bottom profiling) because of their high directivity and absence of side lobes.

Describing the parametric array and nonlinear sound beams, in general, is a complicated problem that requires taking into account the effect of diffraction. The combined effects of nonlinearity, attenuation, and diffraction on a sound beam may be modeled with an augmented form of Equation (1) called the Khokhlov–Zabolotskaya–Kuznetsov (KZK) equation. This equation contains an additional term that accounts for the diffraction of narrow beams. It is widely used for calculations of finite-amplitude sonars, such as the parametric array, and focused beams of high-intensity ultrasound used in imaging and medicine. Further augmentation of Equation (1) permits investigation of nonlinear propagation in inhomogeneous media and leads to a theoretical formalism called "nonlinear geometrical acoustics," which is used to study the evolution of ray patterns.

Acoustic fields in fluids can also generate time-averaged effects such as acoustic streaming (steady flow) and radiation pressure. Acoustic streaming provides a basis for some microfluidic pumps, and in other applications, it influences heat transfer. Radiation pressures generated by standing waves are used to levitate and manipulate particles and bubbles, especially in microgravity environments. Accurate prediction of these time-averaged effects requires special care when defining the problem under consideration.

Nonlinear acoustics of solids includes a variety of interactions between different wave types, such as longitudinal, shear, and surface waves in isotropic solids. Additional types of wave interactions occur in anisotropic media (crystals). The different propagation speeds of the various waves lead to selection rules for particular wave interactions, usually connected with the direction of propagation. Surface (interface) waves exhibit nonlinear propagation effects that differ fundamentally from those of bulk longitudinal and shear waves. Their penetration depth away from an interface is proportional to wavelength, and this frequency-dependent property gives rise to nonlocal nonlinearity that does not occur in bulk waves. With nonlocal nonlinearity, the nonlinear perturbation of the wave velocity at any given point on a waveform is influenced by the entire wave field.

Along with "classical" homogeneous media in which the nonlinearity is due to anharmonicity of atomic forces, there exists an important class of "structurally" nonlinear media in which the nonlinearity is due to the presence of soft inclusions in a harder main body. Examples are small gas bubbles in liquids, pores in rubber-like materials, and grain contacts, microcracks, and dislocations in solids. Acoustic nonlinearities in such media can be several orders stronger than in homogeneous media. Moreover, the nonlinear equation of state (stress-strain relation) in such materials as metals and rock can predict irreversible behavior (hysteresis), which results in some unusual effects; for example, the third-harmonic amplitude in a strain wave can be proportional to the square of the primary-wave amplitude, rather than to its cube.

From a physical viewpoint, classical nonlinear acoustics deals mainly with non- or weakly-dispersive media, in which cumulative waveform steepening and harmonic generation occurs until shocks are formed. Dispersion in acoustics can be introduced by waveguide walls (geometrical dispersion) or by inhomogeneities such as grains and bubbles. As a result, many effects known in nonlinear optics and plasma physics have been realized in acoustics, such as self-focusing, phase conjugation (wave front reversal), and parametric amplification. Acoustic solitons, in which certain effects of nonlinearity and dispersion offset each another, are also possible. Even an acoustical analog of the maser, in which randomly phased oscillators (such as resonant bubbles) self-synchronize and radiate coherently, is possible.

Applications and manifestations of nonlinear acoustics are numerous. Besides the aforementioned processes, acoustic nonlinearities are important in explosion waves, sonic booms, thermoacoustic engines (in which sound waves serve as heat pumps), therapeutic and diagnostic medical ultrasound, materials characterization, and nondestructive testing. Shock waves also form inside collapsing cavitation bubbles, and they are thought to influence the resulting flashes of light referred to as sonoluminescence.

LEV OSTROVSKY AND MARK HAMILTON

See also **Burgers equation; Shock waves; Surface waves**

Further Reading

Beyer, R.T. 1997. *Nonlinear Acoustics*, 2nd edition, New York: Acoustical Society of America

Hamilton, M.F. & Blackstock, D.T. (editors). 1998. *Nonlinear Acoustics*, San Diego: Academic Press

Naugolnykh, K.A. & Ostrovsky, L.A. 1998. *Nonlinear Wave Processes in Acoustics*, Cambridge and New York: Cambridge University Press

Rudenko, O.V. & Soluyan, S.I. 1977. *Theoretical Foundations of Nonlinear Acoustics*, New York: Plenum Press

NONLINEAR ELECTRONICS

In the early part of the 20th century, vacuum tube devices dominated electronics circuitry. However with the advent of junction transistors in 1951 and integrated chips in 1957, semiconductor devices largely replaced vacuum tubes by the 1970s. Present-day electronic circuits make extensive use of active elements/devices that may behave either linearly or nonlinearly, depending upon their applications, operating currents, and voltages (as well as various physical effects including thermal effects, dielectric breakdown, and magnetic saturation). Typical modern nonlinear devices include semiconductor diodes, bipolar junction transistors (BJTs), and field effect transistors (FETs). For analysis and design, the nonlinear devices are often replaced by equivalent basic nonlinear circuit elements such as two-terminal, multiterminal, and multiport resistors, capacitors, and inductors. Thus, a nonlinear electronic circuit is an interconnection of the various circuit elements involving at least one nonlinear element.

The major nonlinear two-terminal circuit elements are typically characterized by nonlinear functional representations in contrast to their linear counterparts: (i) nonlinear resistor: nonlinear voltage $v(t)$ vs.

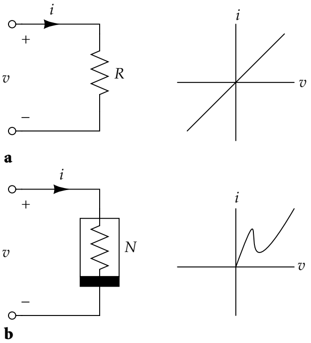

Figure 1. (a) Linear resistor and its v–i characteristic (b) Nonlinear resistor and its v–i characteristic.

current $i(t)$ characteristic curve, (ii) nonlinear capacitor: nonlinear charge $q(t)$ vs. voltage $v(t)$ curve, and (iii) nonlinear inductor: nonlinear $i(t)$ vs. magnetic flux Φ(t) curve (for an illustrative example, see Figure 1). More general circuit elements, such as three terminal transistors and multipliers, and multiterminal elements, such as operational amplifiers (opamps), are also frequently used (Chua et al., 1987; Schubert & Kim, 1996). An innovative nonlinear element is the piecewise-linear device, namely Chua's diode synthesized using opamps/diodes and linear elements (possessing a five-segment v–i characteristic curve, including three negative resistance pieces, see Figure 2), which plays a crucial role in understanding various nonlinear dynamical phenomena (Lakshmanan & Murali, 1996; Lakshmanan & Rajasekar, 2003). The state equations for the currents and voltages underlying a given nonlinear electronic circuit, which are deduced using Kirchhoff's laws for electrical circuits, turn out to be a system of coupled nonlinear differential equations. In particular, when a given circuit includes at least two energy storage elements such as capacitors and inductors and a nonlinear element (which can typically function as an amplifier), it behaves as an oscillator for appropriate feedback and circuit parameters. The underlying system of nonlinear differential equations can then be equivalently considered as a typical nonlinear oscillator dynamical system of dissipative type.

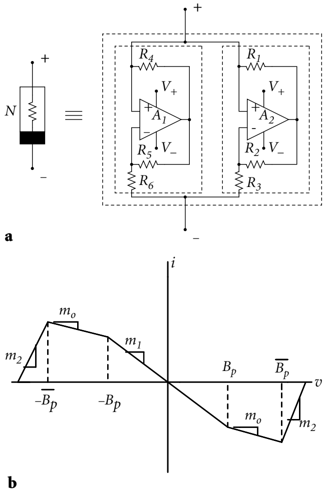

Figure 2. (a) Chua's diode and (b) its characteristic.

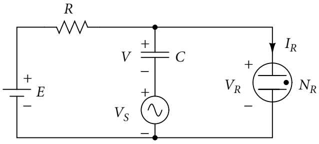

Figure 3. Van der Pol's original oscillator circuit.

Historically, the relaxation oscillator investigated by the Dutch engineer Balthasar der Pol in his seminal paper, "Frequency Demultiplication" (van der Pol & van der Mark , 1927), may be considered as the earliest example of a nonlinear electronic circuit, exhibiting many of the characteristic features underlying bifurcations and chaos, though not fully understood at that time. The circuit typically consists of a high voltage dc source E attached via a large series resistance R to a neon bulb N_R (or a triode valve) and a capacitor C, which are connected in parallel (see Figure 3) along with an external periodic signal V_s. As the capacitance C is increased smoothly, the current exhibits "discrete jumps from one whole-submultiple of the driving frequency to the next." For a critical value of the amplitude of the driving signal, this pattern of mode-lockings has a self-similar fractal structure consisting of an infinite number of steps. In modern jargon, this is just the devil's staircase (Kennedy & Chua, 1986). Van der Pol also noted that "often an irregular noise is heard in the telephone receiver [monitoring the signal in some way] before the frequency jumps to the next level value." Now we know that this "noise" indeed corresponds to chaotic signals. Typically, the circuit is represented by the second-order non-autonomous nonlinear differential equation

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \varepsilon\left(x^2 - 1\right)\frac{\mathrm{d}x}{\mathrm{d}t} + x = f\sin\omega t, \tag{1}$$

where ε is the damping coefficient and f and ω represent the strength and frequency, respectively, of the periodic external forcing.

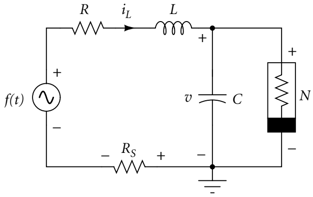

Figure 4. MLC Circuit: N is Chua's diode and $f(t)$ is the periodic signal.

Similar circuits are useful for modeling physical and biological systems; examples include flow of current across Josephson junctions or nerve impulse propagation in neuronal fiber in the form of Hodgkin–Huxley or FitzHugh–Nagumo equations (Scott, 2003). From the point of view of nonlinear dynamics, nonlinear electronic circuits arise either in analog simulation of typical nonlinear oscillators such as the Duffing oscillator or Lorenz system or as new dynamical systems per se constructed using typical nonlinear devices or ingeniously designed elements such as Chua's diode to act as black boxes to understand the various novel nonlinear phenomena. Examples of the latter include Chua's circuit/oscillator and Murali–Lakshmanan–Chua (MLC) circuit (Figure 4). In either case, these circuits are easy to build, analyse, and model, and help to scan the control parameter space quickly, thereby complimenting numerical and analytical studies of nonlinear phenomena.

Just like standard nonlinear dissipative dynamical systems, nonlinear electronic circuits typically exhibit various dynamical phenomena including bifurcations and chaos (Lakshmanan & Murali, 1996; Lakshmanan & Rajasekar, 2003; Thompson & Stewart, 1986).

- *Multistable states:* These correspond to point attractors (nodes and spiral points) as exhibited, for example, by any flip-flop or Schmitt trigger.
- *Periodic oscillations:* Autonomous nonlinear circuits can typically exhibit sinusoidal time response for weak nonlinearity (e.g., LC oscillators) and nearly square wave response (relaxation oscillators) for strong nonlinearity (e.g., square wave generators, Schmitt triggers). On the other hand, nonautonomous circuits often exhibit nonlinear resonant (harmonic, subharmonic, superharmonic, and dissipative) oscillations for weak nonlinearity (e.g., ferroresonant circuit with ac voltage source, hysteresis circuits, power amplifiers).
- *Bifurcations and chaos:* Almost the entire spectrum of bifurcations and chaos phenomena encountered in typical chaotic nonlinear dynamical systems is exhibited by numerous simple nonlinear electronic circuits/oscillators (e.g., Duffing oscillator, van der Pol oscillator, Chua's circuit, MLC circuit, RL diode circuit, Colpitts oscillator, Buck converters in power electronics).
- *Synchronization and controlling:* Nonlinear electronic circuits are versatile models to study and understand the phenomenon of synchronization in all its manifestations (such as identical, phase, lag, and generalized synchronizations) both with periodic and chaotic oscillations between two or a chain of oscillators. Similarly, to study various controlling techniques of periodic as well as chaotic oscillations, nonlinear circuits play a crucial role.
- *Secure communications and cryptography*: Because of the inherent advantage of miniaturization, nonlinear electronic circuits are natural choices for chaos application studies in secure communications, cryptography, and signal processing.

Nonlinear electronic circuits have become indispensable tools to understand a multitude of nonlinear dynamical phenomena, including chaos. In turn, nonlinear electronics has itself been enriched greatly by the advances in understanding various complex nonlinear phenomena, leading to the potential applications mentioned above.

MUTHUSAMY LAKSHMANAN

See also **Chaotic dynamics; Chua's circuit; Diodes; Duffing equation; Van der Pol equation**

Further Reading

Chua, L.O., Desoer, C.A. & Kuh, E.S. 1987. *Linear and Nonlinear Circuits*, New York: McGraw-Hill

Kennedy, M.P. & Chua, L.O. 1986. Van der Pol and chaos. *IEEE Transactions on Circuits and Systems*, 33: 974–980

Lakshmanan, M. & Murali, K. 1996. *Chaos in Nonlinear Oscillators: Controlling and Synchronization*, Singapore: World Scientific

Lakshmanan, M. & Rajasekar, S. 2003. *Nonlinear Dynamics: Integrability, Chaos and Patterns*, Berlin: Springer

Schubert, T.S. & Kim, E.M. 1996. *Active and Nonlinear Electronics*, New York: Wiley

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Thompson, J.M.T. & Stewart, H.B. 1986. *Nonlinear Dynamics and Chaos*, New York: Wiley

van der Pol, B. & van der Mark, J. 1927. Frequency demultiplication. *Nature*, 120: 363–364

NONLINEAR OPTICS

Maxwell's theory of electrical and magnetic fields and his idea that light is an electromagnetic wave were among the great milestones of scientific thought, unifying our understanding of a large and diversified range of physical phenomena. Indeed, by the late 19th century, the success of the classical electromagnetic theory of light led some to believe that there were few fundamental discoveries to be made.

Nonlinear properties of Maxwell's constitutive relations $\boldsymbol{B}=\mu(\boldsymbol{H})\boldsymbol{H}, \boldsymbol{D}=\varepsilon(\boldsymbol{E})\boldsymbol{E}$ had been recognized from the beginning. For example, the nonlinear permeability of ferromagnetic media was of prime concern in the design of electrical machinery in the 19th century. Nonlinear properties in the optical region had to await the discovery of the ruby laser in 1960 by Theodore Maiman (Maiman, 1960). The defining experiment by Peter Franken and coworkers in 1961 (Franken et al., 1961) detected ultraviolet light (l = 3470Å) at twice the frequency of a ruby laser (l = 6940Å) when this beam traversed a quartz crystal. (There is a humorous story about this discovery: the detected radiation was so weak that it appeared as a very faint spot on the photograph submitted to *Physical Review Letters*. The typesetting staff assumed that it was a glitch and erased it in the published document!) This experiment spurred a flurry of activity and many nonlinear optical phenomena were rapidly discovered. A series of monographs by Nicolaas Bloembergen in the 1960s (see Bloembergen, 1996a,b) contain an excellent historical overview of the explosive growth of activity in this field in the early days (see also Marburger, 1977; Shen, 1977).

For the most part, the study of nonlinear optical phenomena assumes a quantum material system coupled to a classical electromagnetic field. Within this approximation, noise-driven phenomena can be adequately described by a Langevin-type driving force. A proper description of quantum effects such as spontaneous emission or scattering requires that the electromagnetic field itself be quantized. Most nonlinear optical materials are nonmagnetic, so one sets the magnetic permeability $\mu(\boldsymbol{H})=\mu_0$, a constant. The procedure then is to expand the dielectric permeability $\varepsilon(\boldsymbol{E})$ as a formal power series in the electric field. This is most commonly done by defining an induced polarization field in terms of the electric displacement vector $\boldsymbol{D}=\varepsilon_0\boldsymbol{E}+\boldsymbol{P}$, where $\boldsymbol{P}=\boldsymbol{P}_{\mathrm{L}}+\boldsymbol{P}_{\mathrm{NL}}$ (linear and nonlinear contributions). The first term on the right-hand side of the expression for the electric displacement vector $\boldsymbol{D}$ is the vacuum contribution. The polarization $\boldsymbol{P}$ describes the coupling of the light to induced dipoles in the material, and this physical quantity is expanded in series in powers of the electric field, $\boldsymbol{E}$, as follows.

$$
\begin{aligned}
\tfrac{1}{\varepsilon_0}\boldsymbol{P} = & \int \overline{\overline{\chi}}_1(\tau)\,\boldsymbol{E}(t-\tau)\,\mathrm{d}\tau \\
& + \int\int \overline{\overline{\chi}}_2(\tau_1,\tau_2)\,\boldsymbol{E}(t-\tau_1) \\
& \times \boldsymbol{E}(t-\tau_2)\,\mathrm{d}\tau_1\,\mathrm{d}\tau_2 \\
& + \int\int\int \overline{\overline{\chi}}_3(\tau_1,\tau_2,\tau_3)\,\boldsymbol{E}(t-\tau_1) \\
& \times \boldsymbol{E}(t-\tau_2)\,\boldsymbol{E}(t-\tau_3)\,\mathrm{d}\tau_1\,\mathrm{d}\tau_2\,\mathrm{d}\tau_3 + \cdots \\
= & \tfrac{1}{2\pi}\int \widetilde{\overline{\overline{\chi}}}_1(\omega)\,\widetilde{\boldsymbol{E}}(\omega)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega \\
& + \tfrac{1}{(2\pi)^2}\int\int \widetilde{\overline{\overline{\chi}}}_2(\omega_1,\omega_2) \\
& \times \widetilde{\boldsymbol{E}}(\omega_1)\,\widetilde{\boldsymbol{E}}(\omega_2)\,\mathrm{e}^{-\mathrm{i}(\omega_1+\omega_2)t}\mathrm{d}\omega_1\mathrm{d}\omega_2 \\
& + \tfrac{1}{(2\pi)^3}\int\int\int \widetilde{\overline{\overline{\chi}}}_3(\omega_1,\omega_2,\omega_3)\,\widetilde{\boldsymbol{E}}(\omega_1)\,\widetilde{\boldsymbol{E}}(\omega_2) \\
& \times \widetilde{\boldsymbol{E}}(\omega_3)\,\mathrm{e}^{-\mathrm{i}(\omega_1+\omega_2+\omega_3)t}\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2\,\mathrm{d}\omega_3 + \cdots
\end{aligned}
\tag{1}
$$

Here, the dielectric susceptibility terms $\overline{\overline{\chi}}_j,\ j=1,2,3$ are second, third, and fourth rank tensors, respectively. The latter are causal functions and tensor product notation is assumed. The second equality is written in terms of Fourier transform variables. Although, in principle, the dielectric susceptibilities at each order could be obtained from fully ab initio quantum mechanical calculations on the relevant materials, in practice, such calculations are too complex and one usually has to rely on experimentally measured data. Higher order processes than those displayed explicitly in this equation are generally unimportant because the leading order accessible nonlinear effect dominates. (Exceptions occur when nonlinear saturation becomes important.) The above induced polarization term acts as a source for the electromagnetic field in Maxwell's vector wave equation,

$$
\frac{\partial^2\boldsymbol{E}}{\partial t^2}-\frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial z^2}-\nabla\cdot(\nabla\cdot\boldsymbol{E})=\frac{1}{\mu_0}\frac{\partial^2\boldsymbol{P}}{\partial t^2}. \tag{2}
$$

The leading order term $\overline{\overline{\chi}}_1$ above is a second rank tensor that describes all linear optical interactions involving propagating optical fields of any arbitrary polarization. For virtually all nonlinear optical interactions, it is appropriate to expand the electric field vector as a linear combination of products of envelope functions and optical carrier waves

$$
\begin{aligned}
\boldsymbol{E}(\boldsymbol{r},t) = & \sum_j \widehat{e}_j A_j(\boldsymbol{r},t)\exp\left(\mathrm{i}\left(\pm\boldsymbol{k}_j\cdot\boldsymbol{r}-\omega_j t\right)\right) \\
& + \text{c.c.}
\end{aligned}
\tag{3}
$$

Here A_j are slowly varying envelope functions, and $\widehat{e}_j$ is a unit vector indicating the direction of polarization. Each wave vector or frequency pair $(\boldsymbol{k}_j,\omega_j)$ satisfies a linear dispersion relation $D(\boldsymbol{k}_j,\omega_j)=0$. A goal in nonlinear optics is to write down evolution equations for the complex slowly varying envelope functions A_j. The structure of these equations, which describe how almost monochromatic, weakly nonlinear, dispersive wave trains interact, is universal and so many of the properties of lightwaves can be inferred from corresponding situations in other fields.

The leading nonlinear behavior in noncentrosymmetric crystals is due to the second order term $\overline{\overline{\chi}}_2$, a third rank tensor (with 9 components). This term is responsible for second harmonic generation as first observed by Franken et al. (1961). Significant quadratic nonlinear interactions between wave packets take place when triads of wave vectors and frequencies $\boldsymbol{k}_j, \omega_j$ obey

$$\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3 = 0, \quad \omega_1 \pm \omega_2 \pm \omega_3 = 0, \tag{4}$$

where each $\boldsymbol{k}_j, \omega_j$ pair satisfies its own dispersion relation. Physical effects arising from this term include second harmonic generation ($2\omega_1 \to \omega_3$), dc rectification ($\omega_1 - \omega_1 \to 0$), and frequency up- and down-conversion ($\omega_1 \pm \omega_2 \to \omega_3$). The first two nonlinear interactions are termed degenerate because $\omega_1 = \omega_2$ and two photons ($2\hbar\omega_1$) in the incident laser beam are needed to create one second harmonic photon. For these three-wave interaction processes to be efficient both energy (i.e., $\hbar\omega_1 \pm \hbar\omega_2 = \hbar\omega_3$ for up-, down-conversion, $2\hbar\omega_1 = \hbar\omega_3$ for second harmonic generation) and momentum ($\hbar\boldsymbol{k}_1(\omega) \pm \hbar\boldsymbol{k}_2(\omega) = \hbar\boldsymbol{k}_3(\omega)$, for up-, down-conversion, $2\hbar\boldsymbol{k}_1(\omega) = \hbar\boldsymbol{k}_3(\omega)$ for harmonic generation) must be nearly conserved. (Here $\hbar\boldsymbol{k}$ is the photon momentum.) Three-wave interactions are generally difficult to observe in isotropic media as material dispersion precludes being able to satisfy the momentum conservation relation. A symmetry argument also shows that χ^2-processes are forbidden in materials possessing a center of symmetry. Noncentrosymmetric uniaxial and biaxial crystals (those in which the refractive index is different in different directions in the crystal) are employed to achieve efficient phase matching. This is because the magnitude of the optical wave vector is related to the material refractive index through the relation $k = n(\omega)\omega/c$, where the dispersion of the refractive index $n(\omega)$ is explicitly displayed. The complex envelopes A_1, A_2, and A_3 of the three wave packets obey the universal three-wave interaction equations of universal type

$$\frac{\partial A_j}{\partial t} + \boldsymbol{c}_j \cdot \nabla A_j = \theta_j A_l^* A_m^* \tag{5}$$

with j, l, m cycled over 1,2,3, where $\boldsymbol{c}_j$ is the linear group velocity of the jth wave packet and θ_j is a coupling coefficient proportional to $\overline{\overline{\chi_2}}$.

Four-wave χ_3-processes are the leading order nonlinear optical processes in a medium possessing a center of symmetry. Here significant exchange of energy only takes place between resonant quartets $\left(\boldsymbol{k}_j, \omega_j\right)_{j=1}^4$, each obeying its dispersion relation. Contrary to the triad resonance condition, energy conservation and phase matching can always be satisfied with several trivial choices. The nonlinear dielectric susceptibility is now a fourth rank tensor although many of its components are either identical or zero, depending on crystal symmetry properties. Degenerate four-wave interactions include self-phase modulation (or Kerr effect, $\omega_1 + \omega_1 - \omega_1 \to \omega_1$) and third harmonic generation ($\omega_1 + \omega_1 + \omega_1 \to 3\omega_1$) (Terhune et al., 1962). Common nondegenerate third-order processes include four-wave-mixing (FWM) interactions.

Four-wave-mixing interactions can be used to create a "phase-conjugated wave." A nonlinear interference grating is created within the nonlinear material through the interaction of two strong pump waves satisfying $\boldsymbol{k}_1 + \boldsymbol{k}_2 = 0$ and $\omega_1 = \omega_2 = \omega$. A weak third wave scatters off this grating and generates a time-reversed phase-conjugated replica of itself. The incident and scattered waves satisfy $\boldsymbol{k}_3 + \boldsymbol{k}_4 = 0$ and $\omega_4 = \omega_3 = \omega$. FWM interactions can also lead to degradation in wavelength-division-multiplexed (WDM) long-haul fiber transmission systems as a result of energy transfer between equally spaced channels. Another important nondegenerate interaction involves the phenomenon of nonlinear birefringence where $\boldsymbol{k}_4, \omega_4 = -\boldsymbol{k}_2, -\omega_2, \boldsymbol{k}_3, \omega_3 = -\boldsymbol{k}_1, -\omega_1$ and the complex slowly varying functions A_1, A_2 represent envelope wave packets of different polarization direction. Coupled nonlinear Schrödinger (NLS) equations for the envelope fields $A_1(\boldsymbol{r}, t)$ and $A_2(\boldsymbol{r}, t)$ can be derived that describe nonlinear self- (i.e., $|A_1|^2 A_1$ in the A_1 equation and $|A_2|^2 A_2$ in the A_2 equation) and cross-phase modulation (i.e., $|A_2|^2 A_1$ in the A_1 equation and $|A_1|^2 A_2$ in the A_2 equation) for co- and counter-propagating laser pulses. The relative weighting of self- and cross-phase modulation nonlinearities depends on the material properties (Marburger, 1977) and plays an important role in polarization switching of soliton pulses in optical fibers. One can write down more general universal four-wave interaction equations along the lines described above for the three-wave interaction case.

A well-known degenerate χ_3-process is the optical Kerr effect. This self-interaction term induces a nonlinear phase change proportional to the intensity $|A|^2$ of the propagating light pulse. This phase change, called self-phase modulation, is an accumulative propagation effect that is central to applications that exploit the optical Kerr effect. A canonical form of the D-dimensional NLS equation, incorporating the Kerr effect, can be written as follows:

$$\mathrm{i}\left(\frac{\partial A}{\partial t} + \sum_{j=1}^{D}\left(\frac{\partial \omega}{\partial k_j}\right)\frac{\partial A}{\partial x_j}\right) - \sum_{j,l=1}^{D}\left(\frac{\partial^2 \omega}{\partial k_j k_l}\right)\frac{\partial^2 A}{\partial x_j x_l} + \left(\frac{\partial \omega}{\partial |A|^2}\right)|A|^2 A = 0. \tag{6}$$

The first two terms combined describe the advection of a pulse (wave packet) with group velocity $v_{\mathrm{g}} = \omega\,(k)$.

The second and third terms represent the balance between dispersion and weak nonlinearity. Higher order corrections terms can be derived as an asymptotic expansion in a small parameter as described in Newell & Moloney (1992). Dispersion in nonlinear optics is manifested in two physically distinct ways: diffraction describing spatial spreading of a tightly collimated beam or pulse in a direction transverse to the direction of propagation, and dispersive spreading in time (group velocity dispersion $\omega''(k)$) of an ultrashort laser pulse. The latter mechanism plays an increasingly influential role as the propagating laser pulse gets shorter and shorter.

In 1-d, the NLSE is integrable and the Kerr nonlinearity can balance group velocity dispersion to form temporal solitons in optical fibers. Both bright and dark temporal soliton pulses have been observed experimentally in fibers. Spatial soliton beams can be created in two-dimensional planar waveguides where there is strong confinement of the optical field in a single spatial dimension transverse to the propagation direction of the light beam. Spatial confinement in two (fibers) or one (planar waveguides) dimension is required to form stable soliton pulses or beams. In unconfined transparent bulk geometries, the optical Kerr effect is responsible for critical self-focusing, leading to extremely intense focused light that causes catastrophic damage in transparent glasses. Critical self-focusing was first observed experimentally by Chiao and coworkers in 1964 (Chiao et al., 1964).

There are other categories of nonlinear optical effects that can be described in terms of envelope functions and that depend sensitively on the finite response of some material oscillation. Stimulated scattering is a case in point. These processes are essentially three-wave interactions where one of the optical waves is replaced by a material oscillation. In transparent glasses for example, there are two characteristic modes of oscillation of the lattice as shown in Figure 1. The lower frequency acoustic phonon mode corresponds to an oscillation of all atoms in unison along a fixed direction in the lattice. The higher frequency optical phonon mode involves neighboring atoms in the lattice oscillating out of phase. The dispersion relation for lattice vibrations, therefore, contains an acoustic and optical branch, both of much lower frequency than the optical frequency.

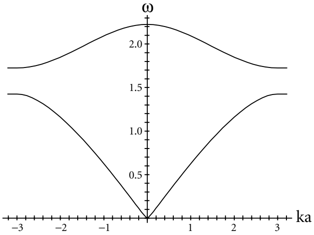

Figure 1. Acoustic (bottom)and optical (top) phonon branches of the dispersion curve for a diatomic lattice. The abscissa is scaled to the lattice constant “a.”

Figure 1 shows a plot of the optical phonon (top curve) and acoustic phonon (bottom curve) dispersion for a diatomic crystal lattice within the first Brillouin zone. On the scale of this graph, the photon branch (light cone) is essentially vertical, making the acoustic and optical phonon dispersion essentially flat. An intense pump laser pulse at frequency ω_p entering a transparent glass can scatter off either a fluctuating low-frequency acoustic phonon (Stimulated Brillouin Scattering, SBS) or a fluctuating optical phonon (Stimulated Raman Scattering, SRS) at finite temperature. Energy conservation ensures that the scattered optical signal has a frequency $\omega_\mathrm{s} = \omega_\mathrm{p} - \omega_\mathrm{v}$, where ω_v is the natural acoustic or optical phonon frequency. Technically, SBS and SRS processes look like a three-wave interaction where phase matching is easy as the dispersion of the material oscillation is locally flat on an optical frequency scale. In fact, it can be shown that these processes are essentially four-wave interactions, as the oscillator model, representing the material oscillation, is driven by the square of the optical field. Typically, SBS is generated in the backward direction and SRS can to be generated in the forward and/or backward direction relative to the propagating pump laser pulse. SRS processes in liquids and gases can generate a cascade of frequency down-shifted (Stokes) and frequency up-shifted (anti-Stokes) waves. Although both SBS and SRS processes are essentially phase matched for all directions, they can only exist by feeding off the intense pump pulse. There also exists a degenerate stimulated scattering process called Stimulated Rayleigh Scattering—this is essentially a Kerr-like process with finite memory. All of these stimulated scattering phenomena can be observed in bulk transparent materials (Marburger, 1977; Shen, 1977) and optical fibers (Agrawal, 1989). They can be stimulated from noise or seeded at the optical signal wavelength by injecting a weak laser field.

Another important class of nonlinear optical phenomena involves direct resonant coupling to some form of material oscillation. The simplest physical manifestation of this category is the two-level atom. Absorbing/amplifying media that give rise to the phenomenon of self-induced transparency (SIT) and lasers belong to this category. The dielectric susceptibility is now complex with the real part corresponding to a refractive index change induced by the optical field and the imaginary part corresponding to absorption or amplification of light. The real and imaginary parts of the dielectric susceptibility are related through a Hilbert transform—the Kramers–Kronig relation. Energy is directly transferred between optical fields and material oscillations. Self-induced transparency (SIT) solitons propagate when this coherent population or energy exchange can occur between the lower (L) and upper (U) level of a two-level atom on a timescale short relative to any irreversible energy decay processes such as spontaneous emission from the excited state (U). This population cycling between two levels occurs at a frequency $\omega_R = |\mu_{12}A|/\hbar$, where μ_{12} is the dipole matrix between the two levels and $\hbar$ is Planck's constant. The frequency ω_R is called the Rabi frequency, from the field of magnetic resonance spectroscopy. Figure 2a depicts this simple two-level atom scheme. For longer optical pulses, nonlinear saturation can occur in the presence of intense propagating optical fields. Population inversion, required for lasing action, cannot be achieved in a two-level atom (*See* **Lasers**). Optical amplification of light requires that a net inversion be achieved between the upper (U) and lower (L) atomic or molecular level involved in the light amplification process. Such inversion can be achieved by pumping atoms or molecules to energetically higher levels with a subsequent rapid nonradiative energy decay down to the upper excited level (U). Classical laser action can be described in terms of three-level and four-level quantum models. Figure 2b and c depict two classical lasing level schemes where the ground level population is pumped ($E = \hbar\omega_P$) into some upper excited state followed by a rapid (nonradiative) population decay (dashed arrows) to the upper excited lasing level (U). The original ruby laser, discovered by Maiman in 1961, is an example of a three-level (actually three band) system schematically shown in Figure 2b. Inversion is more difficult to achieve in a three-level laser as the upper lasing level must have a net population in excess of that in the ground level. This requires pumping of more than half of the atoms to the upper lasing level. In the four-level laser, the lower lasing state is not the ground state and its initial population can be zero or very small due to finite temperature populations. Efficient lasing is achieved by rapidly removing the populations from the lower lasing level (L).

Such energy level schemes are generally gross simplifications of a real laser although they provide useful models of solid state and gas lasers (Siegman, 1986). Fiber lasers with erbium and/or ytterbium-doped cores belong to the category of solid state lasers. Semiconductor lasers, on the other hand, are extremely complex many-body systems involving many bands rather than levels and the above descriptions prove inadequate.

Key current developments in the field of nonlinear optics have benefited from technical developments that have enabled experimentalists to produce ultrashort laser pulses and invent novel materials processing methodologies. We sketch some of the key areas of research where such effects have been exploited in recent years.

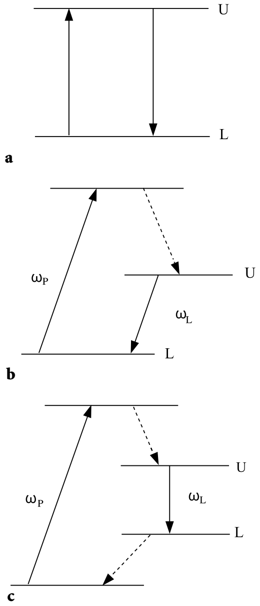

Figure 2. (a). Two-level energy level scheme. (b). Three-level lasing level scheme. (c). Four-level lasing level scheme.

Nonlinear Optical Data Processing

A cumulative intensity-induced phase shift due to the Kerr nonlinearity, of the order of p, provides one of the most useful all-optical data processing capabilities. High-speed optical sampling, switching, amplification and storage applications abound, based on this nonlinear mechanism. Optical switching and data sampling can be implemented in directional couplers and Mach–Zehnder interferometers in both waveguide and fiber arrangements. Optical amplification and storage of pulse trains representing digital information can be achieved by placing a nonlinear transparent Kerr medium in a ring or Fabry–Perot cavity arrangement. So-called optical bistable elements require an injected holding beam to maintain the cavity at some threshold level whereby an incremental input can lead to strong amplification or storage of the system in one of two stable accessible states. The latter situation is depicted in Figure 3. Here the system has a memory (hysteresis) whereby a lower (bit 0) and upper (bit 1) state can be switched by some external perturbation. Adjusting some cavity parameter can also lead to a single-valued but rapidly rising output transmission of the system. This is useful for amplifying injected pulse streams.

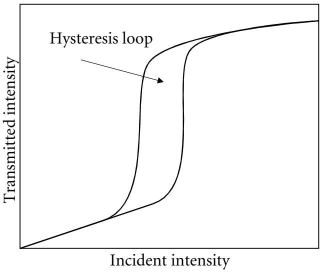

Figure 3. Hysteresis loop for a bistable optical cavity.

Such passive optical bistable cavities have also been shown to exhibit deterministic chaotic dynamics when the finite time delay of the optical signal circulating in the optical resonator has been taken into account. When transverse degrees of freedom of the external pump laser beam are included, so-called transverse spatial solitons have been observed. Although technology applications have been envisioned for both of these dynamical phenomena, there are no specific implementations to date.

Chaotic Synchronization and Communication with Lasers

Semiconductor lasers are highly susceptible to exhibiting chaotic dynamics even in the presence of extremely weak (10^{-4}) feedback from an external reflecting surface. This tendency to readily exhibit deterministic chaos means that semiconductor laser sources require greater than 40 dB optical isolation, making them the most expensive component in an optical fiber telecommunications network. Research in recent years has focused on taking advantage of this chaos by demonstrating that chaotic transmitter and receiver lasers can be synchronized and messages encoded and decoded at the transmitter and receiver end, respectively. Mathematically, these chaotic lasers are described at the simplest level by delay differential equations.

Quasi-phase Matching in Isotropic Media

Quasi-phase matching (QPM) is a means of beating the usual triad resonance condition needed to efficiently transfer energy between waves in second harmonic generation. Figure 4 shows the ideal phase-matched energy transfer to second harmonic as a function of propagation distance in a uniaxial crystal. When detuned from the phase-matching condition, energy transfer occurs less efficiently and there is a periodic transfer back and forth between the fundamental at frequency and the second harmonic signal at frequency 2ω. In this case, one observes the oscillatory nonphase-matched behavior depicted in the picture. This is typically the situation encountered in isotropic materials.

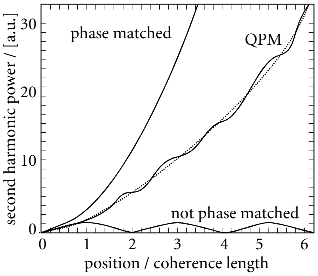

Figure 4. Comparison of energy transfer from the fundamental to the second harmonic for phase-matched, detuned and quasi-phase-matched conditions.

Quasi-phase matching entails taking a crystal with a nonlinear coefficient and periodically reverting the sign of the dipoles in such a way that the phase of the wave is reversed every coherence length along the crystal. The latter is proportional to the refractive index mismatch at the fundamental and second harmonic frequencies. A practical implementation of quasi-phase-matching is through the process of periodic polling whereby a ferroelectric material such as lithium niobate, for example, can effectively have its dipoles reversed periodically with patterned electrodes. QPM soliton propagation is an intense area of modern research and has the advantage that much larger nonlinear optical couplings can be achieved via the χ_2 nonlinearity.

Optical Breakdown in Intense Laser fields

Optical breakdown due to critical self-focusing of continuous wave laser beams or long pulses has been observed since the 1960s. The breakdown phenomenon itself is very complex and can involve nonlinear coupling to thermal, acoustic, vibrational, rotational, and electronic degrees of freedom. Thermally induced lensing, leading to critical self-focusing, is dominant for long microsecond-duration laser pulses. Nanosecond-duration laser pulses couple strongly to hypersonic acoustic waves and critical self-focusing is due to the physical mechanism of electrostriction. Picosecond-duration laser pulses typically generate electron/ion plasmas through the process of avalanche photoionization, whereas femtosecond-duration laser pulses breakdown materials predominantly via the mechanism of multiphoton ionization. The latter process is essentially instantaneous and the nonlinear ionization process is proportional to $|A|^{2N}$, where the minimum energy required to excite the bound electron to the ionization continuum is $E = N\hbar\omega$.

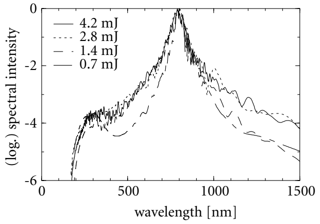

Figure 5. Calculated white-light supercontinuum for a high-power femtosecond laser pulse propagating in air.

Stimulated scattering processes usually accompany optical breakdown processes making it extremely difficult to unravel the relevant physics for longer duration laser pulses. Plasma, generated via avalanche ionization, acts as a shield, preventing the laser pulse from propagating further in the material. Shock waves generated in the nonlinear focal region can cause mechanical rupture for longer duration pulses.

Extreme Nonlinear Optics

The development of high-power femtosecond laser sources has led to some novel nonlinear interactions that exploit the optical Kerr effect. Single cycle pulses have been generated and very recent experimental evidence has been provided for the generation of attosecond pulses. These optical pulses are of such short duration that they interact with the atom or molecule on a timescale comparable to that taken by the electron to orbit the nucleus! The advantage of these laser sources is that many of the detrimental physical effects accompanying longer duration pulses are not operative in the femtosecond regime. Therefore, it is possible to achieve large local field intensities while maintaining low energies in the pulse. Recently, high-power femtosecond laser pulses, focused down to scales on the order of the wavelength of light, have been used to locally change the material properties of glass and write complicated integrated optics waveguiding structures. The actual mechanism for the transformation of the glass to a state with a higher refractive index is not yet fully understood.

High-power femtosecond laser pulses can propagate over anomalously long distances in air, achieving peak field intensities on the order of 10^{13}–10^{14} W/cm^2. These huge intensities break down the air via multiphoton ionization in the vicinity of the nonlinear focus. Plasma, generated within the 100 μm diameter focal spot, can act as a defocusing lens, causing the trailing edge of the laser pulse to defocus and refocus multiple times as it propagates through the atmosphere. Accompanying the extreme self-focusing is white-light supercontinuum generation. The relatively narrow spectral peak of the initial pulse blows out asymmetrically and encompasses wavelengths ranging from the infrared to the ultraviolet. Figure 5 shows the change in the calculated white-light supercontinuum spectrum as a function of increasing pulse energy for a 100 fs-duration laser pulse propagating in air. The experimentally measured back-reflected white-light supercontinuum from a 2.6 TW 100 fs-duration laser pulse launched vertically into the atmosphere (Rairoux et al., 2000) exhibits the same qualitative features. The initial pulse spectrum is localized around the narrow peak at 800 nm. The mechanism, by which this high-power pulse sustains itself as a waveguide, is through a combination of recurring transverse modulational instabilities that create multiple intense chaotic self-focusing light strings and accompanying plasma filament generation that acts to limit the strong focusing effect (Mlejnek et al., 1999).

Another important application of high-power femtosecond laser pulses is in the generation of highly transient X-ray pulsed sources through higher harmonic cascades accompanying extreme pulse focusing. Hundreds of harmonics have been observed experimentally (Bartels et al., 2000), and novel algorithms are being developed to reshape the initial pulse so as to be able to efficiently extract a particular higher harmonic. When the pulse peak intensities exceed around 10^{17} W/cm^2, one enters the nonlinear regime where relativistic effects become important.

The extreme conditions encountered during critical focusing of high-power femtosecond-duration laser pulses have recently spurred theorists to question whether envelope equation models such as the NLS equation, even with higher order corrections included, can adequately describe pulse propagation under such extremes. As NLS describes quasi-monochromatic laser pulses, it is implicitly assumed in its derivation that the generated spectral bandwidth satisfies the inequality, which is the underlying optical carrier wave frequency. Experimental measurements of white-light generation in condensed media show that the spectral extent of the generated supercontinuum can exceed the initial pulse spectral bandwidth by a factor of 5 or more. The picture above for air shows that the supercontinuum extends well beyond the optical carrier frequency. Pulse propagation models that explicitly include the optical carrier wave and consequently allow for optical carrier shocks are currently under active development.

Nonlinear Optics in Photonic Bragg Structures

Periodic modulation of a material refractive index or gain offers a novel means of beating the intrinsic dispersion of available nonlinear optical materials. This idea goes back to the early days of nonlinear optics where 1-dimensional periodic index modulation allowed one to manage or engineer the dispersion of a particular material. This is the photonic analog of the lattice periodicity leading to so-called Bloch states describing electron confinement in a semiconductor material. Just as the optical properties of semiconductor materials can be engineered by modifying the underlying lattice structure through the introduction of quantum wells or dots, so too can the optical confinement of photons be controlled by introducing a refractive index or gain periodicity in a transparent glass. Distributed feedback semiconductor lasers, based on this distributed feedback principal, are the workhorses of modern fiber-based telecommunications systems. 1-d Bragg or gap solitons have been experimentally generated in periodically modulated Kerr glass waveguides—these localized pulses have the property that they can be dramatically slowed down or even be nonpropagating within the nonlinear periodic structure.

2-d and 3-d photonic Bragg and crystal fiber structures have the special property that photonic bandgaps can be introduced in a structure consisting of a periodic lattice of holes or vertical columns. In principle, such structures can guide light around right angle bends in direct contrast to waveguide splitters, trap light in defects, and provide so-called omni-reflecting properties, that is, guide light at all angles and wavelengths. These properties are in marked contrast to conventional index-guided optical waveguides and, when combined with optical nonlinearity, offer many new potential technology applications (Mingaleev & Kivshar, 2002).

Electromagnetically Induced Transparency

Electromagnetically induced transparency (EIT) is a quantum interference effect that acts to reduce the usual absorption of light experienced when its frequency is tuned to the resonance frequency of the sample through which the light is propagating. The transparency is created by a second light (electromagnetic) source tuned to another resonance of the sample. Suppressing absorption through EIT leads to several other effects, including the focusing of one laser beam by another and the production of inversionless laser sources.

JEROME V. MOLONEY

See also **Filamentation; Kerr effect; Lasers; Nonlinear Schrödinger equations; Optical fiber communications; Photonic crystals; Pump-probe measurements; Rayleigh and Raman scattering and IR absorption**

Further Reading

Agrawal, G.P. 1989. *Nonlinear Fiber Optics*, San Diego: Academic Press

Bartels, R., Backus, S., Zeek, E., Misoguti, L., Vdovin, G., Christov, I.P., Murnane, M.M. & Kapteyn, H.C. 2000. Shaped-pulse optimization of coherent emission of high-harmonic soft X-rays. *Nature*, 406: 164

Bloembergen, N. 1996a. *Encounters in Nonlinear Optics: Selected Papers of Nicolaas Bloembergen*, Singapore: World Scientific

Bloembergen, N. 1996b. *Nonlinear Optics*, Singapore: World Scientific

Boyd, R.W. 1992. *Nonlinear Optics*, Boston and San Diego: Academic Press

Chiao, R.Y., Garmire, E. & Townes, C.H. 1964. Self-trapping of optical beams. *Physical Review Letters*, 13: 479–482

Franken, P.A., Hill, A.E., Peters, C.W. & Weinreich, G. 1961. Generation of optical harmonics. *Physical Review Letters*, 7: 118–120

Maiman, T.H. 1960. Stimulated optical radiation in ruby lasers. *Nature*, 187: 493

Marburger, J. 1977. Self focusing: theory. *Progress in Quantum Electronics*, 4: 35–110

Mingaleev, S. & Kivshar, Y. 2002. Nonlinear photonic crystals: toward all-optical technologies. *Optics and Photonics News*, 13: 48–51

Mlejnek, M., Kolesik, M., Moloney, J.V. & Wright, E.M. 1999. Optically turbulent femtosecond light guide in air. *Physical Review Letters*, 83: 2938–2941

Newell, A. & Moloney, J.V. 1992. *Nonlinear Optics*, Redwood City, CA: Addison-Wesley

Rairoux, P., Schillinger, H., Niedermeier, S., Rodriguez, M., Ronneberger, F., Sauerbrey, R., Stein, B., Waite, D., Wedekind, C., Wille, H. & Wöste, L. 2000. Teramobile: a nonlinear terawatt femtosecond lidar. *Applied Physics* B, 71: 573

Shen, Y.R. 1977. Self focusing: experimental. *Progress in Quantum Electronics*, 4: 1–34

Siegman, A.E. 1986. *Lasers*. Mill Valley, CA: University Science Books

Terhune, R.W., Maker, P.D. & Savage, C.M. 1962. Optical harmonic generation in calcite. *Physical Review Letters*, 8: 404–406

NONLINEAR PLASMA WAVES

In studying nonlinear plasma waves, one should recognize that there are several analytical methods, including single particle, fluid, and kinetic. Using the proper model is important, as different models can result in different nonlinear phenomena (Davidson, 1972; Horton & Ichikawa, 1996).

Because plasmas are composed of charged particles and are often immersed in external electric and magnetic fields, it is critical to determine when the self-consistent fields (fields produced by the charged particles directly) must be considered in the presence of externally applied fields. Typically, the self-consistent electric fields need to be considered, but we can often neglect self-consistent magnetic fields, because the forces produced by the self-consistent magnetic fields are usually much lower than those produced by external magnetic fields.

Transverse Plasma Waves

As an example of the single-particle approach, consider transverse electromagnetic (TEM) waves in a plasma. We begin with Maxwell's equations combining them to produce the wave equation below:

$$\nabla \times \nabla \times \boldsymbol{E} = -\mu\sigma \frac{\partial \boldsymbol{E}}{\partial t} - \mu\varepsilon \frac{\partial^2 \boldsymbol{E}}{\partial t^2}, \tag{1}$$

Where σ, ε, and μ are the conductivity, electric permittivity, and magnetic permeability of the plasma, respectively. We shall assume that all of the plasma effects can be collected in the permittivity and that the conductivity is part of the permittivity, so we can set $\sigma = 0$ in Equation (1). If we now assume cartesian coordinates and that we have a transverse electromagnetic wave propagating in the z-direction with the propagation behaving like

$$\exp(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t), \tag{2}$$

then we may substitute this notation into Equation (1) and represent Equation (1) as the matrix shown below:

$$\begin{pmatrix} -k_z^2 + \omega^2\mu\varepsilon & 0 & 0 \\ 0 & -k_z^2 + \omega^2\mu\varepsilon & 0 \\ 0 & 0 & -k_z^2 + \omega^2\mu\varepsilon \end{pmatrix} \times \begin{pmatrix} E_{0x} \\ E_{0y} \\ E_{0z} \end{pmatrix} = 0. \tag{3}$$

There are three possible eigenmodes (two of which are identical) for this equation. The first two are propagating disturbances, while the third mode is a time-invariant nonpropagating mode. In a plasma, this mode represents the electrostatic oscillations generated when the electrons are displaced with respect to the ions and allowed to move under the influence of the self-consistent (charge separation) electrostatic field. Vibrations of gelatin also exhibit this same property. If we assume that E_{0z} is zero, then the remaining two modes look just as they would in an ordinary dielectric. However, the permittivity of a plasma based on single-particle motions with stationary ions and no external d.c. magnetic field is expressed as

$$\varepsilon = \varepsilon_0 \left(1 - \frac{\omega_\mathrm{p}^2}{\omega\,(\omega + \mathrm{i}\nu_\mathrm{m})} \right), \tag{4}$$

where ν_m is the electron momentum transfer collision frequency and ω_p is the electron plasma frequency defined as

$$\omega_\mathrm{p}^2 = \frac{ne^2}{m\varepsilon_0}. \tag{5}$$

In Equation (5), n is the electron density, e is the charge of an electron, m is its mass, and ε_0 is the permittivity of free space. Since Equation (4) now has both real and imaginary parts, an effective conductivity of the plasma can now be obtained. Figure 1 shows a plot of the real part of ω vs. wave number k for two conditions: zero collision frequency and nonzero collision frequency. Because the relationship between k and ω is not a constant as frequency changes, we see that a TEM wave in a plasma exhibits dispersion. Examination of the real part of the propagation vector also shows that the wave amplitude will decrease as long as the collision frequency is not zero.

Nonlinear Ion-acoustic Waves

In this example, we consider the representation of a plasma as a fluid. Under these conditions, it is possible to excite longitudinal acoustic-type oscillations. We assume that the plasma mass is concentrated in its ions but that any electric fields that are generated by charge separation appear because of the higher mobility of the electrons. In this representation, we may write the equation of motion of a fluid element as follows (Chen, 1984):

$$Mn \left[\frac{\partial \boldsymbol{v}_\mathrm{i}}{\partial t} + (\boldsymbol{v}_\mathrm{i} \cdot \nabla)\boldsymbol{v}_\mathrm{i} \right] = en\boldsymbol{E} - \nabla p, \tag{6}$$

where $\boldsymbol{v}_\mathrm{i}$ refers to the ion species fluid velocity, p is the pressure, M is the ion mass, and n is the plasma density. Neglecting the electron mass (for simplicity), linearizing, and assuming propagating waves as previously, $\boldsymbol{E}$ is the negative gradient of a scalar potential ϕ. Using the equation of state gives

$$Mn(\mathrm{i}\omega + \nu_\mathrm{m})v = -en_0\mathrm{i}k\phi - \gamma_i kT_i kn_1, \tag{7}$$

where n_0 is the steady-state density and n_1 is the perturbed density. Assuming that

$$e\phi/kT_\mathrm{e} = n_1/n_0, \tag{8}$$

where T_e is the electron temperature, we obtain

$$(\omega + \mathrm{i}\nu_\mathrm{m})v = kV_\mathrm{s}\frac{n_1}{n_0}. \tag{9}$$

In Equation (9) V_s is the sound speed $(kT_\mathrm{e}/M)^{1/2}$. Applying a similar analysis to the continuity equation and substituting Equation (9) we obtain

$$\begin{aligned} -\mathrm{i}\omega n_1 + \mathrm{i}kn_0 v &= -\mathrm{i}\omega n_1 + \mathrm{i}k\frac{n_0}{(\omega + \mathrm{i}\nu_\mathrm{m})}[kV_\mathrm{s}^2]\frac{n_1}{n_0} \\ &= 0 \end{aligned} \tag{10}$$

or

$$(\omega^2 + \mathrm{i}\nu_\mathrm{m}\omega - k^2V_\mathrm{s}^2) = 0. \tag{11}$$

Longitudinal Plasma Oscillations (Langmuir Waves)

Now consider the kinetic representation of a plasma, in the case where the electric field is parallel to the direction of propagation. Assuming propagation in the z-direction, an electric field in the z-direction can be developed from an electrostatic potential of the form

$$\phi = \phi_0 \exp[\mathrm{i}(kz - \omega t)]. \tag{12}$$

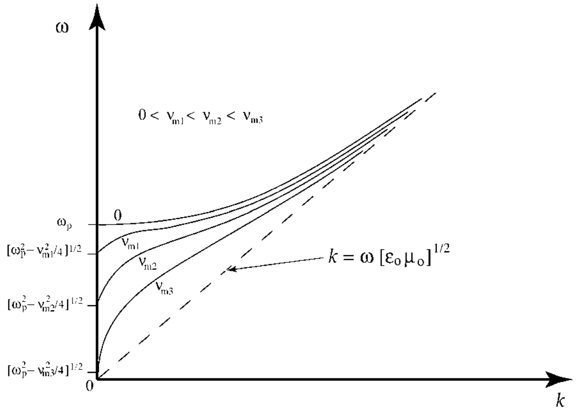

Figure 1. The real part of ω versus wave number k for increasing values of collision frequency ν_{m}.

If there are only electrostatic forces, the velocity distribution f of the electrons is governed by the Vlasov equation:

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla f + \frac{e}{m} \boldsymbol{E} \cdot \frac{\partial f}{\partial \boldsymbol{v}} = 0. \tag{13}$$

Writing a perturbation for f as $f = f_0 + f_1$ in which f_0 is only a function of velocity and f_1 can be a function of both position and velocity, we may write the perturbed part of the Vlasov equation as

$$\frac{\partial f_1}{\partial t} + \boldsymbol{v} \cdot \nabla f_1 + \frac{e}{m} \boldsymbol{E} \cdot \frac{\partial f_0}{\partial \boldsymbol{v}} = 0. \tag{14}$$

Writing Poisson's equation with this ordering results in

$$\nabla \cdot \boldsymbol{E} = \frac{n_0 e}{\varepsilon_0} \int f_1 \, \mathrm{d}\boldsymbol{v}. \tag{15}$$

Solving Equations (14) and (15) simultaneously yields the dispersion relation for these longitudinal plasma oscillations, often called Langmuir waves, as

$$1 = \frac{n_0 \varepsilon_0 e^2}{m} \frac{\boldsymbol{k}}{k^2} \cdot \int \frac{\partial f_0 / \partial \boldsymbol{v}}{(\boldsymbol{k} \cdot \boldsymbol{v} - \omega)} \, \mathrm{d}\boldsymbol{v}. \tag{16}$$

Note that the form of f_0 affects the dispersion relation. Also, the denominator in the integral of Equation (16) can have a singularity at which nonlinear effects can occur. It has been shown, for example, that the velocity distribution can draw energy from the wave or vice versa depending upon its slope. As a result, solitary waves can appear, expressing a balance between the driving and dissipative conditions.

In addition, Zakharov (1972) showed that Langmuir waves can evolve nonlinearly so that the electric field that results from the charge separation can become very large in a very small region. Under these "collapse" conditions, the field does not become infinitely large, but is limited by wave-particle interactions and often bursts of high-energy electrons are observed.

Plasmons

A plasmon is the resulting electrostatic oscillation described above when the velocity distribution can be neglected. For example, if it is assumed the ions in a plasma are stationary and the electrons are each uniformly displaced from their equilibrium position, then a Coulomb restoring force will be produced by the excess and deficiency of electrons at each end of the plasma, causing the electrons to oscillate about the ions. The frequency of oscillation is the electron plasma frequency:

$$\omega_{\mathrm{p}} = \left(\frac{n e^2}{m \varepsilon_0} \right)^{1/2}. \tag{17}$$

If the ion motion is considered, an ion plasma frequency also appears in the calculations. Plasmons, especially in solid state materials, can interact nonlinearly with electromagnetic fields and other radiation.

Plasma Turbulence

Turbulence appears when modeling for the three approaches (single particle, kinetic, and fluid) does not satisfactorily describe the experimental observations. For example, in hot plasmas, such as those appearing in stars and fusion reactors, fluctuations in the plasma produce enhanced effects on mass, momentum and heat transport that must be studied in the context of turbulence.

If we could follow the motion of each and every particle in a plasma and its interactions with other particles and external electric and magnetic fields, we would be able to understand plasma turbulence. As that is not possible, alternative approaches must be undertaken.

Turbulence is said to appear in a plasma when the oscillations and fluctuations exhibit a continuous spectrum of frequencies. As a result of this state of turbulence, interactions between species in the plasma and the fluctuations can occur. For example, electrons, in addition to interacting with charged particles, can also interact with electric fields that are generated by the fluctuations. In addition, turbulent motion in a plasma has a very interesting effect: it can generate strong magnetic fields, producing a dynamo. This has been proposed as the mechanism for the generation of earth's magnetic field.

Numerical simulation has been shown to be a very powerful tool. In this case, the simulation involves a number of assumptions in order to make the problem solvable. The key method for these simulations involves following the trajectory of many simulation particles, as they interact with each other, that represent the plasma under various assumptions which can be applied to particular problems. By use of large supercomputers, it is possible to simulate astronomical and fusion plasma phenomena and examine the structure of the turbulence. A color figure in the color plate section shows a cross section of the results of such a simulation for a toroidal magnetic confinement system called a Tokamak. Calculations of this scale of detail could not be done without numerical simulations.

Whistlers and Helical Waves

Whistler waves have been detected in the ionosphere and have been attributed to the change of phase and group velocities of a right circularly polarized plasma wave that travels along the earth's magnetic field. The dispersion relation for this mode is

$$k^2 = \omega^2 \mu_0 \varepsilon_0 \left[1 - \frac{\omega_{\mathrm{pi}}^2 + \omega_{\mathrm{pe}}^2}{(\omega + \omega_{\mathrm{ci}})(\omega - \omega_{\mathrm{ce}})} \right]. \qquad (18)$$

In the ionosphere, the frequency of the waves is lower than the electron cyclotron frequency. Thus, the phase velocity (ω/k), which can be found from Equation (18) increases with frequency, and the group velocity also increases. When a lightning discharge takes place, an electromagnetic disturbance that has a large continuous spectrum of components is produced. The disturbance will travel to the ionosphere and then propagate along the Earth's magnetic field lines. However, the lower frequency components of the disturbance will take longer to arrive at a receiving station than the higher frequency components; thus, the received signal appears as a signal that decreases in frequency with time.

In examining the behavior of the electric field vector for a right circularly polarized wave, we note that as the wave propagates along an external magnetic field line, the direction of the electric field rotates around the direction of propagation which is, in this case, the direction of the magnetic field. As a result, this or any other mode can be represented as a helical wave in which the rotation of the end of the electric field vector traces out a helical path as it propagates.

MHD Kink Waves in Toroidally Confined Fusion Plasmas

The sine-Gordon equation has been shown to describe the trajectory of the "slinky" mode in a reversed-field-pinch magnetic fusion experiment (Ebraheem et al., 2002). A reversed-field-pinch experiment produces a magnetic field in a torus in which the portion of the field that is in the toroidal direction (the long way around) is reversed in its outer regions from that closer to the minor axis of the torus. Experimentally, it has been seen that during the operation of the system, magnetohydrodynamic (MHD) fluctuations in the center of the plasma can coalesce into a rotating "hot spot" that travels around the torus. Under some circumstances, the rotating mode may lock to other disturbances or to particular locations on the vacuum chamber torus.

As the mode may be envisaged as a helical solitary wave, it lends itself to being represented by the sine-Gordon equation. The mechanical transmission line analogy (Scott, 1971) can be applied here, if it is assumed that the MHD mode is a solenoid that wraps helically around the torus. The gravitational restoring torque can be represented as the gradient of the magnetic energy of the magnetic moment of the mode in the externally generated magnetic fields of the torus due to toroidal effects (the magnetic field on the inside of the torus is stronger than at the outside of the torus). The spring torque due to the adjacent pendula can be modeled by recognizing that the current flowing around the turns of the solenoid tends to make the solenoid align itself along its axis, so any disturbance from the alignment results in an equivalent spring torque.

Parametric Decay Instability

Consider two oscillators that have different resonant frequencies ω_1 and ω_2 that are nonlinearly coupled to each other through a third oscillator of resonant frequency ω_0. We assume that the oscillators are normal modes of various types in the plasma such as described above. If each of these waves is considered from the point of view of quantum theory, interactions are governed by conservation of energy:

$$\hbar\omega = \hbar\omega_1 + \hbar\omega_2 \tag{19}$$

and conservation of momentum

$$k_0 = k_1 + k_2, \tag{20}$$

where k represents the wave number for each mode. The zero subscript refers to the "pump" mode and the other two modes are referred to as the "daughter" modes (Rost et al., 2002).

Assume that two of the oscillators obey the following equation of motion where each is driven by a time-dependent force that is proportional to the amplitude of a pump oscillator E_0:

$$\frac{\mathrm{d}^2 x_1}{\mathrm{d}t^2} + \omega_1^2 x_1 = c_1 x_2 E_0, \tag{21}$$

$$\frac{\mathrm{d}^2 x_2}{\mathrm{d}t^2} + \omega_2^2 x_2 = c_1 x_1 E_0, \tag{22}$$

where c_1 and c_2 are nonlinear coupling constants. If E_0 is made to vary in time at frequency ω_0, it can be seen that both modes can be driven by a frequency not equal to their natural frequency and, in addition, each oscillator's motion is coupled to the other oscillator's equation of motion.

J. Leon Shohet

See also **Alfvén waves; Magnetohydrodynamics; Plasma soliton experiments; Turbulence**

Further Reading

Chen, F.F. 1984. *Plasma Physics and Controlled Fusion*, 2nd edition, New York: Plenum Press.

Davidson, R.C. 1972. *Methods in Nonlinear Plasma Theory*, New York: Academic Press

Ebraheem, H.K., Shohet, J.L. & Scott, A.C. 2002. Mode-locking in reversed-field pinch experiments. *Physical Review Letters*, 88: 235003

Horton, C.W. & Ichikawa, Y.H. 1996. *Chaos and Structures in Nonlinear Plasmas*, River Edge, NJ: World Scientific

Rost, J.C., Porkolab, M. & Boivin, R.L. 2002. Edge ion heating and parametric decay during injection of ion cyclotron resonance frequency power on the Alcator C-Mod tokamak. *Physics of Plasmas*, 9: 1262–1270

Scott, A.C. 1971. *Active and Nonlinear Wave Propagation in Electronics*, New York: Wiley

Zakharov, V.E. 1972. Collapse of Langmuir waves. *Soviet Physics JETP*, 35: 908–912

NONLINEAR RESONANCE

See **Damped-driven anharmonic oscillator**

NONLINEAR SCHRÖDINGER EQUATIONS

The basic nonlinear Schrödinger (NLS) equations are

$$\mathrm{i}u_t \pm \tfrac{1}{2}u_{xx} + |u|^2 u = 0, \tag{1}$$

where $u(t, x)$ is a complex function, the real variables t and x are frequently (but not always, see below) time and space coordinates, and the subscripts indicate partial derivatives. The upper and lower signs in front of the second term in Equation (1) correspond, respectively, to self-focusing and self-defocusing NLS equations, to be denoted below as NLS(+) and NLS(−).

NLS equations attracted much attention after Zakharov & Shabat (1972) demonstrated that Equations (1) are integrable by means of the inverse scattering transform (IST) (Zakharov et al., 1984). Under this formulation, Equation (1) is viewed as a compatibility condition for a system of two auxiliary linear equations for a two-component function $\psi^{(1,2)}(x, t)$. The first of these equations is

$$\begin{pmatrix} -\mathrm{i}\partial_x & u^*(x) \\ -u(x) & \mathrm{i}\partial_x \end{pmatrix} \begin{pmatrix} \psi^{(1)} \\ \psi^{(2)} \end{pmatrix} = \lambda \begin{pmatrix} \psi^{(1)} \\ \psi^{(2)} \end{pmatrix}, \tag{2}$$

where λ is a spectral parameter and the asterisk stands for the complex conjugation. The first step in the application of IST is to solve the direct scattering problem for Equations (2) (also called Zakharov–Shabat (ZS) equations) with a given initial configuration $u_0(x)$ at $t = 0$, that is, to map $u_0(x)$ into a set of scattering data. In the general case, the set contains continuous and discrete components, which correspond, respectively, to real and complex λ. The temporal evolution of scattering data, generated by the temporal evolution of $u(x, t)$ under Equation (1), turns out to be trivial, so that, once the scattering data is known at $t = 0$, it is also known for any $t > 0$. Finally, at given t, one must recover the field $u(t, x)$ from the known scattering data, that is, to solve the inverse scattering problem, which is based on the Gel'fand–Levitan–Marchenko integral equation.

The simplest solution of NLS(+) represents a soliton (exponentially localized pulse)

$$\begin{aligned} u_{\mathrm{sol}} = {} & \eta \operatorname{sech}\left[\eta\left(x - ct - x_0\right)\right] \\ & \times \exp\left(\mathrm{i}cx - \mathrm{i}\omega_{\mathrm{sol}}t + \mathrm{i}\phi_0\right), \end{aligned} \tag{3}$$

where $\omega_{\mathrm{sol}} = \left(c^2 - \eta^2\right)/2$, $\operatorname{sech} z \equiv 2\left(\mathrm{e}^z + \mathrm{e}^{-z}\right)^{-1}$, η and c are arbitrary amplitude and velocity of the soliton, and x_0 and ϕ_0 are the coordinate of the soliton's center and its phase at $t = 0$.

Note that frequencies of the zero-velocity ($c = 0$) solitons are negative ($\omega_{\mathrm{sol}} = -\eta^2/2 < 0$), while the frequencies of radiation modes, that is, solutions to the linearized version of NLS(+), $u_{\mathrm{rad}} = u_0 \exp\left(\mathrm{i}kx - \mathrm{i}\omega_{\mathrm{rad}}t\right)$, are positive ($\omega_{\mathrm{rad}} = k^2/2 > 0$, where k and u_0 are arbitrary wave number and amplitude). Thus, solitons exist in a part of the frequency space where radiation modes are absent, exemplifying a general principle: solitons cannot share frequencies with radiation, as they would otherwise be losing energy through the emission of linear waves.

In terms of the IST, the soliton corresponds to a case when the continuous component of the scattering data is absent, while the discrete component consists of a single eigenvalue $\lambda = (-c + \mathrm{i}\eta)/2$. All the solutions corresponding to an arbitrary set of complex eigenvalues can be found in an explicit form, describing collisions between solitons unless their velocities coincide. The solitons reappear after the collision with the same amplitudes and velocities, that they had prior to the collision, the only change being shifts of the constants x_0 and ϕ_0. If velocities coincide, the solution describes a bound state of solitons, which is dynamically unstable because its binding energy is zero. Thus, the simple soliton given by Equation (3) is the single stable solution to NLS(+) with a permanent wave shape.

Besides the solitons, Equations (1) have spatially uniform solutions in the form of continuous waves (CWs), $u_{\mathrm{CW}}(t) = \eta \exp\left(\pm \mathrm{i}\eta^2 t\right)$, with an arbitrary real amplitude η. An important issue is stability of the CW solutions against small perturbations. To investigate the stability, one can represent a solution as $u(t,x) = a(x,t)\exp(\mathrm{i}\phi(t,x))$, where $a(t,x)$ and $\phi(t,x)$ are real amplitude and phase. Rewriting Equation (1) as a system of real equations for a and ϕ leads to a perturbed CW with $a(t,x) = \eta + a_1(t,x)$, and $\phi(t,x) = \pm\, \eta^2 t + \phi_1(t,x)$, with infinitesimal perturbations a_1 and ϕ_1. Linearizing the equations with respect to a_1 and ϕ_1 yields a solution in the form

$$\begin{aligned} a_1(x,t) &= a_1^{(0)} \exp(\sigma t)\cos(px)\,, \ \phi_1(x,t) \\ &= \phi_1^{(0)} \exp(\sigma t)\cos(px)\,, \end{aligned} \tag{4}$$

where $a_1^{(0)}$ and $\phi_1^{(0)}$ are infinitesimal amplitudes of the perturbation, p is its arbitrary wave number ($-\infty < p < +\infty$), and σ is an instability growth rate. The linearized equations imply the following relation between σ and p:

$$\sigma^2(p) = \pm p^2\left(\eta^2 \mp p^2/4\right). \tag{5}$$

The necessary and sufficient condition for the stability of CW against small perturbations is $\operatorname{Re}\sigma(p) \le 0$, which must hold for all p. As is seen from Equation (5), all the CW solutions to NLS(−) are stable, but in NLS(+) they all are unstable for p^2 in the range $0 < p^2 < 4\eta^2$. It follows from Equation (4) that growing perturbations break the uniformity of the CW, initiating spatial modulation. This is called modulational instability and also Benjamin–Feir instability, after the authors who discovered it in the context of water waves.

Although NLS(−) does not support solitary-wave solutions, it has a stable solution in the form of a dark soliton (DS), which is a region of low amplitude moving on top of a CW. The DS solution is

$$\begin{aligned} u_{\mathrm{DS}} &= \eta \exp\left(\mathrm{i}\eta^2 t\right) \\ &\quad \times \{(\cos\theta)\tanh[\eta(\cos\theta)(x - ct)] + \mathrm{i}\sin\theta\}\,, \end{aligned} \tag{6}$$

where η is the amplitude of the CW background, and θ, which takes values between 0 and $\pi/2$, determines the minimum value of the field at the center of DS and its velocity (Kivshar & Luther-Davies, 1998). The solution of Equation (3) is sometimes called a bright soliton to stress its difference from the DS.

A fundamental property of NLS equations which is not related to the integrability and is therefore more generic, being shared by a broad class of equations, is a possibility to represent the equations in the Lagrangian and Hamiltonian forms. The general Lagrangian representation is $\delta S/\delta u^*(t,x) = 0$, where $S\{u(t,x), u^*(t,x)\}$ is an *action functional*, which must be real, and $\delta/\delta u^*$ stands for the variational (Fréchet) derivative of the functional (u and u^* are treated as independent functional arguments). The action functional that generates Equation (1) is

$$S = \int \mathrm{d}t \int_{-\infty}^{+\infty} \mathrm{d}x \left[\mathrm{i}\left(u^* u_t - u u_t^*\right) \mp |u_x|^2 + |u|^4\right]. \tag{7}$$

The Hamiltonian representation of the NLS equations is

$$\mathrm{i}\frac{\partial u}{\partial t} = \frac{\delta H\{u, u^*\}}{\delta u^*}, \tag{8}$$

where the Hamiltonian (H) is $\int \mathrm{d}x\left(\pm |u_x|^2/2 - |u|^4/2\right)$. The representation in the form of Equation (8) is very general, applying to a large class of equations for complex fields.

For real H, it follows from Equation (8) that the Hamiltonian is conserved. From the perspective of the Noether theorem, conservation of H is a consequence of the fact that Equation (1) is invariant with respect to an arbitrary time shift, $t \to t + t_0$. The invariances with respect to an arbitrary phase shift, $u(t,x) \to u(t,x)\exp(\mathrm{i}\phi_0)$, and to an arbitrary coordinate shift, $x \to x + x_0$, generate the conservation of the number N (alias "number of quanta") and momentum P, which are given by universal expressions, irrespective of the particular form of the equation:

$$\begin{aligned} N &= \int_{-\infty}^{+\infty} \mathrm{d}x\ |u(x)|^2\,, \\ P &= \frac{1}{2}\mathrm{i}\int_{-\infty}^{+\infty} \mathrm{d}x\left(u u_x^* - u^* u_x\right). \end{aligned} \tag{9}$$

If the equation is integrable by means of IST, it possesses, in addition to the dynamical invariants N,

P, and H, an infinite series of higher-order conserved quantities (Zakharov et al., 1984).

Despite the exact integrability of Equation (1) by means of IST, the solution provided by this technique may be too cumbersome; therefore, a more tractable approximate method for the construction of solutions is often desirable. To this end, a variational approximation (VA), whose applicability relies solely on the Lagrangian representation of the equation(s), has been developed for solitons of NLS(+) and for several other equations, including those described below. This method approximates the full PDE by a low-order system of ODEs, which are derived from the same action functional (Malomed, 2002).

NLS Models for Wave Envelopes in Weakly Nonlinear Dispersive Media

Equations (1) also appear as universal equations governing the evolution of slowly varying packets of quasi-monochromatic waves in weakly nonlinear media featuring dispersion (dependence of the wave's group velocity c_{gr} on its frequency ω).

To illustrate this point, one can take the sine-Gordon (SG) equation, $\psi_{tt} - \psi_{xx} + \sin\psi = 0$, for real $\psi(t, x)$. Assuming a small-amplitude wave with a slowly varying complex envelope function $u(t, x)$ carried by a high-frequency wave $\exp(-\mathrm{i}t)$, the solution is $\psi(t, x) = u(t, x)\exp(-\mathrm{i}t) + u^*(t, x)\exp(\mathrm{i}t)$, where the slow variation implies that $|u_t|^2 \ll |u|^2$. Substituting into SG, expanding $\sin\psi$ up to the cubic term, dropping the higher-order small terms u_{tt} and u^*_{tt}, and separately collecting terms in front of two independent rapidly varying functions $\exp(\pm\mathrm{i}t)$ yields two equations: NLS(+) for $u(t, x)$, and its complex conjugate for $u^*(t, x)$.

Thus, NLS equations can be derived as equations governing the evolution of weakly nonlinear wave packets in various media. An important example is light propagation in nonlinear optical fibers and planar waveguides. In the former case, NLS takes the normalized form $\mathrm{i}u_z \pm (1/2)u_{\tau\tau} + |u|^2 u = 0$, where $u(\tau, z)$ is the local amplitude of the electromagnetic wave, which is a function of the propagation distance z and $\tau \equiv t - z/c_{\mathrm{gr}}$ (c_{gr} is the group velocity of light in the fiber). In the context of Equation (1), z plays the role of the evolutionary variable, and τ plays the role of the coordinate. The cubic term in the equation accounts for the Kerr effect (a nonlinear correction to the refractive index), while the upper and lower signs in front of $u_{\tau\tau}$ correspond to the anomalous and normal chromatic dispersion in the fiber, $\mathrm{d}c_{\mathrm{gr}}/\mathrm{d}\omega > 0$ and $\mathrm{d}c_{\mathrm{gr}}/\mathrm{d}\omega < 0$, respectively ($\omega$ is the frequency of the electromagnetic wave). Only anomalous dispersion gives rise to bright optical solitons (also called temporal solitons) as they are localized in τ (Agrawal, 1995). Temporal solitons are important in fiber-optic telecommunications.

In planar waveguides, spatial bright solitons are possible in the form of self-confined light beams. In this case, the propagation coordinate z again replaces t in Equation (1), while the transverse coordinate x appears the same way as in Equation (1), the term u_{xx} accounting for diffraction of light in the waveguide. The corresponding NLS equation is always self-focusing.

Other significant applications in which NLS equations arise are small-amplitude gravity waves on the surface of deep inviscid water and Langmuir waves in hot plasmas (collective oscillations of electrons relative to ions). However, these media are not really one-dimensional, which makes the corresponding NLS solitons unstable against transverse perturbations.

In plasma, NLS(+) approximates a more fundamental model—the Zakharov system

$$\begin{aligned} \mathrm{i}u_t + \tfrac{1}{2}u_{xx} + nu &= 0, \\ n_{tt} - n_{xx} &= -\left(|u|^2\right)_{xx}, \end{aligned} \tag{10}$$

where $n(t, x)$ is a perturbation of the ion density, and $u(t, x)$ is the local amplitude of the Langmuir waves. Although Equations (10) are not integrable, they generate stable solitons and reduce to NLS(+) in the adiabatic limit, when the term n_{tt} may be neglected.

Zakharov's system is a universal model describing weakly nonlinear waves in dispersive media, where a dominating process is the interaction between high- and low-frequency waves (e.g., the Langmuir and ion-acoustic waves, respectively, in plasmas). Another example is the interaction of high-frequency (near infra-red) molecular vibrations and low-frequency (far infra-red) longitudinal deformations in a model of long biological molecules, which gives rise to the Davydov soliton.

Generalized Forms of the Nonlinear Schrödinger Equations

The NLS equation can be viewed as a paradigmatic representative of a large class of nonlinear PDEs, some of which are also integrable. Examples are the derivative NLS equation, $\mathrm{i}u_t + u_{xx} + \mathrm{i}\left(|u|^{2u}\right)_x = 0$, whose integrability was discovered by Kaup and Newell (this equation applies to Alfvén waves in magnetized plasmas), the Hirota equation, $\mathrm{i}u_t + u_{xx} + \mathrm{i}\left(u_{xxx} + 6|u|^2 u_x\right)$, and the Sasa–Satsuma equation, $\mathrm{i}u_t + u_{xx} + \mathrm{i}\left[u_{xxx} + 6|u|^2 u_x + 3\left(|u|^2\right)_x u\right] = 0$.

In applications to fiber optics, an important generalization considers two polarizations of light (or two different wavelengths in one fiber), leading to a system of coupled NLS equations

$$\begin{aligned} \mathrm{i}u_z \pm \tfrac{1}{2}v_{\tau\tau} + \left(|u|^2 + \beta\,|v|^2\right)u &= 0, \\ \mathrm{i}v_z \pm \tfrac{1}{2}v_{\tau\tau} + \left(|v|^2 + \beta\,|u|^2\right)u &= 0. \end{aligned} \tag{11}$$

In the case of linear polarizations, $\beta = \frac{2}{3}$, and in the case of circular polarizations or different wavelengths, $\beta = 2$ (Agrawal, 1995). As demonstrated by Manakov, Equations (11) are integrable if $\beta = 1$ (Manakov, 1974).

Generalized NLS equations are not integrable in other cases of interest, including the multidimensional NLS equations

$$\mathrm{i}u_t \pm \tfrac{1}{2}\nabla^2 u + |u|^2\,u = 0, \tag{12}$$

where ∇^2 is the two- or three-dimensional (2-d or 3-d) Laplacian. Equation (12) describes light propagation in bulk media, and Langmuir waves in plasmas. Both the 2-d and 3-d solitons are unstable because NLS(+) gives rise to collapse, that is, formation of a singularity in finite time (Bergé, 1998) (nevertheless, the 2-d soliton, also called the Townes soliton, plays an important role, as it determines the critical power necessary for the collapse in the 2-d case).

The 2-d NLS(−) gives rise to vortex solutions, with important implications in nonlinear optics, which take the form

$$u = \eta \exp\left(\mathrm{i}\eta^2 t + \mathrm{i}S\chi\right) U\left(\sqrt{2}\eta r\right) \tag{13}$$

with arbitrary η and $U(\rho = \infty) = 1$. Here, χ is the angular coordinate in the 2-d space, and $S \geq 1$ is an integer vorticity. Only the vortex with $S = 1$ is stable, being a 2-d counterpart of the DS solution (6) with $\theta = 0$.

Multidimensional bright solitons (also called light bullets in 3-d optical models) may be stable if the cubic term $|u|^2\,u$ in NLS(+) is replaced by a saturable one, $|u|^2\left(1 + |u|^2/u_0^2\right)^{-1}$ ($u_0^2 = \mathrm{const}$), or by a combination of self-focusing cubic and self-defocusing quintic terms, $\left(|u|^2 - |u|^4\right)u$. In the latter case, bright 2-d solitons with the embedded vorticity $S \geq 1$ and 3-d solitons with the vorticity $S = 1$ may also be stable.

Similar to Equation (12) is the Gross–Pitaevskii (GP) equation for the single-atom wave function $\psi(t, \boldsymbol{r})$, on which the mean-field description of Bose–Einstein condensates is based (Dalfovo et al., 1999):

$$\mathrm{i}\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi + U(\boldsymbol{r})\psi + g\,|\psi|^2\,\psi. \tag{14}$$

Here $\hbar$ is Planck's constant, m is the atom mass, $U(\boldsymbol{r})$ is a trapping potential, and g is proportional to the scattering length of collisions between atoms, which may be both positive and negative.

The simplest approach to GP is to seek a stationary solution, $\psi(t, \boldsymbol{r}) = \exp(-\mathrm{i}\mu t/\hbar)\,\Psi(\boldsymbol{r})$ (μ is called the chemical potential), and apply the Thomas–Fermi approximation to $\Psi(\boldsymbol{r})$, which amounts to dropping the Laplacian in Equation (14). This yields $|\Psi(\boldsymbol{r})|^2 = g^{-1}\left[\mu - U(\boldsymbol{r})\right]$ if it is positive, or $\Psi = 0$ otherwise.

Nonlinear optics motivate examples of coupled NLS equations that are different from Equations (11). Propagation of light in an optical fiber carrying a Bragg grating (a lattice of defects with the period equal to half the wavelength of light, which gives rise to mutual resonant conversion of right- and left-traveling waves) is described by equations for the amplitudes $u(t, x)$ and $v(t, x)$ of the counterpropagating waves,

$$\begin{aligned} \mathrm{i}u_t + \mathrm{i}u_x + \left(|u|^2 + 2\,|v|^2\right)u + v &= 0, \\ \mathrm{i}v_t - \mathrm{i}v_x + \left(|v|^2 + 2\,|u|^2\right)v + u &= 0 \end{aligned} \tag{15}$$

with the linear coupling accounting for the resonant Bragg scattering. The spectrum of the linearized version of Equations (15) is $\omega^2 = k^2 + 1$, hence it has a radiation-free gap, $-1 < \omega < +1$, where solitons may be expected. A family of gap solitons is generated by Equations (15) that can be found exactly although Equations (15) are not integrable and not all members of this family are stable (de Sterke & Sipe, 1994; Malomed, 2002).

Another NLS system describes a basic process in nonlinear optics—second harmonic generation (SHG). If the nonlinearity is quadratic, the system is

$$\begin{aligned} \mathrm{i}u_z + \tfrac{1}{2}u_{xx} + u^*v &= 0, \\ 2\mathrm{i}v_z + \tfrac{1}{2}v_{xx} + \tfrac{1}{2}v^2 + qv &= 0, \end{aligned} \tag{16}$$

where $u(z, x)$ and $v(z, x)$ are amplitudes of the fundamental-frequency and second harmonic waves, and real q is a wave number mismatch between the harmonics. Equations (16) are written for the realistic case of spatial fields in a planar waveguide. They give rise to a family of solitons in the form $u = \exp(\mathrm{i}kz)\,U(x)$, $v = \exp(2\mathrm{i}kz)\,V(x)$, with exponentially localized $U(x)$ and $V(x)$. Except for the case $k = q/3$, the solitons are not available in an analytical form, but most of them are stable (Etrich et al., 2000). Multidimensional generalizations of Equations (16) give rise to stable solitons too, as collapse does not take place in the SHG model.

Ginzburg–Landau (GL) equations, which are formally similar to NLS equations, provide general models of nonlinear dissipative media. The simplest GL equation is Equation (1) with all the coefficients made complex. In general, however, dynamics generated by GL equations is altogether different from that governed by NLS equations (Aranson & Kramer, 2002).

BORIS MALOMED

See also **Complex Ginzburg–Landau equation; Discrete nonlinear Schrödinger equations; Equations, nonlinear; Inverse scattering method or transform; Nonlinear optics**

Further Reading

Agrawal, G.P. 1995. *Nonlinear Fiber Optics*, San Diego: Academic Press

Aranson, I.S. & Kramer, L. 2002. The world of the complex Ginzburg–Landau equation. *Reviews of Modern Physics*, 74: 99–143

Bergé, L. 1998. Wave collapse in physics: principles and applications to light and plasma waves. *Physics Reports*, 303: 259–370

Dalfovo, F., Giorgini, S., Pitaevskii, L.P. & Stringari, S. 1999. Theory of Bose-Einstein condensation in trapped gases. *Reviews of Modern Physics*, 71: 463–512

Etrich, C., Lederer, F., Malomed, B.A, Peschel, T. & Peschel, U. 2000. Optical solitons in media with a quadratic nonlinearity. *Progress in Optics*, 41: 483–568

Kivshar, Y.S. & Luther-Davies, B. 1998. Dark optical solitons: physics and applications. *Physics Reports*, 298: 81–197

Malomed, B.A. 2002. Variational methods in nonlinear fiber optics and related fields. *Progress in Optics*, 43: 69–191

Manakov, S.V. 1974. On the theory of two-dimensional stationary self-focusing of electromagnetic waves. *Soviet Physics JETP*, 38: 248–253

de Sterke, C.M. & Sipe, J.E. 1994. Gap solitons. *Progress in Optics*, 33: 203–260

Zakharov, V.E., Manakov, S.V., Novikov, S.P. & Pitaevskii, L.P. 1984. *Solitons: The Method of Inverse Scattering Transform*, New York: Consultants Bureau

Zakharov, V.E. & Shabat, A.B. 1972. Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Soviet Physics, JETP*, 34: 62–69

NONLINEAR SIGNAL PROCESSING

The field of nonlinear signal processing is an active area of research stimulated by a combination of new applications, advances in technology, and fundamental advances in mathematical theory. The breadth of applications has extended far beyond the original emphasis on communications to disciplines such as brain science, financial analysis, image and voice recognition, cryptography, virtual reality, digital photography and printing technology, and neural computing. Significant accelerations in computing speeds and the increasingly widespread availability of computer clusters have made more complex algorithms computationally feasible and motivated the investigation of a much larger range of algorithms. Included in these new algorithms are techniques extracted, for example, from chaos theory, partial differential equations, computational geometry, support vector machines, numerical analysis on manifolds, and optimization theory. Fundamental advances in nonlinear signal processing have centered around an increased mathematical understanding of the notion of information content and representation.

Nonlinear signal processing began with the realization that all components of a signal are not equal. For example, nonlinear filters have been used for compressing signals associated with telephone voice communications where the fidelity required depends on the frequency of the signal. In general, it is possible to retain fewer bits of high frequency data than lower frequency data and still maintain voice clarity. Additionally, median and min/max filters and histogram equalization represent traditional tools in nonlinear signal (and image) processing. In mathematical terms, the motivation for nonlinear signal processing techniques arises from the geometric structure, or organization, of the data. In some idealized sense, we can view data as either being flat, as possessing some degree of curvature, or having a fractal nature. The nature of this structure in the data dictates the best tool to be used for its analysis. For example, *signal dimension* is an important feature of a signal. Based on the geometry of the signal, one may employ, for example, basis dimension, topological dimension, or Hausdorff dimension. There are now a number of methods that have been proposed to detect nonlinear features in data including poly-spectral analysis, surrogate analysis, and more; see Barnett et al. (1997) for detailed comparisons.

As a consequence of the inherently complicated geometry, modeling nonlinear signals is significantly more challenging than their linear counterparts. For example, an n-dimensional linear signal may be written $f(x(t)) = Ax(t)$ where the model consists of the $m \times n$ matrix of parameters. If there are more points than unknowns to construct the model, then the approach of least squares may be used for solving overdetermined systems. This theory, including the effect of perturbations on the data, is well understood. For nonlinear signals with n-dimensional domains, one encounters the well-known "curse of dimensionality" that suggests that the amount of data required to construct such nonlinear mappings is exponentially dependent on the dimension of the domain. However, the recognition that this problem is overcome by judiciously placing basis functions in the domain of the data has led to a large literature in the theory and applications of modeling nonlinear signals in high dimension. Radial basis functions (RBFs) are a popular choice and have the form

$$f(x) = w_0 + \sum_{k=1}^{N} w_k \phi(\|x - c_k\|),$$

where the vectors c_k are the special sites of the basis functions, x the domain value and the w_i the model parameters. The basis functions may be local, for example, $\phi(r) = \exp(-r^2)$ or global, for example, $\phi(r) = r^3$. Artificial neural networks (ANNs) have also been proposed for such nonlinear approximation tasks. These networks provide a flexible, or adaptive, tool to *learn* a function. Generally, a special type of ANN is employed, that is, a feed-forward sigmoidal network. Mathematically, this may be expressed as

$$y = \sum_i w_i^{(2)} \sigma \left(\sum_j w_{ij}^{(1)} x_j - \theta_i \right),$$

where σ is a monotonic nonlinear squashing function and $\{w_i^{(2)}, w_{ij}^{(1)}, \theta_i\}$ the model parameters. There exist universal approximation theorems that state under which conditions both the RBF and ANN approximations can produce models to essentially arbitrary accuracy. Of course, practical issues such as computation expense and parsimony of the model must be considered in the implementation of such methods. Further details on these issues may be found in Haykin (1994) and Kirby (2001).

Dynamical systems theory has provided a new paradigm for the analysis of nonlinear signals. Fuelled by observation that complex or chaotic signals can arise from simple nonlinear deterministic systems, a suite of new tools has been developed for their empirical detection, characterization, and design. Briefly, a discrete dynamical system may be viewed as the iteration of a mapping, that is

$$x^{(n+1)} = f(x^{(n)}). \tag{17}$$

Often the trajectory $\{x^{(0)}, x^{(1)}, x^{(2)}, \ldots\}$ traces out a complicated path. Here $x^{(n)}$ is viewed as a function of n. However, a *delay embedding* that plots a point as a function of the previous one or more points may produce a picture that reveals the structure of the simple underlying determinisitic system. For example, the well-known logistic map has the form

$$x^{(n+1)} = \alpha(1 - x^{(n)})x^{(n)}.$$

It possesses a complicated, apparently random trajectory with broad Fourier spectrum for $\alpha > 4$. However, clearly the locus of points $(x^{(n)}, x^{(n+1)})$ is a parabola. Thus, one is motivated to look for structure in a sequence of observations of a signal by constructing the delay embeddings $(x^{(n)}, x^{(n-1)})$ (one delay) or $(x^{(n)}, x^{(n-1)}, x^{(n-2)})$ (two delays) and so on. In practice, it is useful to explore delays of multiples of an integer, that is, $N, 2N, 3N, \ldots$.

This heuristic approach gains some credibility with the support of the Takens embedding theorem; see for example, Abarbanel (1996) and references therein. To get a flavor of the power of this theorem, consider a multidimensional discrete dynamical system, that is, the point $\mathbf{x}^{(n)}$ being iterated is a d-tuple. Takens's theorem basically says that if we are able to observe any one component of this system, for example $\{x_1^{(n)}\}$ then the delay embedding of this variable produces a geometrical structure that is in some sense equivalent to the original system. Furthermore, if only a function h of the first component is observed $h(x_1^{(n)})$, then we can still replicate the structure of the original system with the delay embedding approach. The number of delays required to achieve this embedding is prescribed by Takens, for example if the dynamical system resides on an m-dimensional manifold, then the number of delays need only be $2m + 1$ (Abarbanel, 1996).

The dynamical system described by Equation (17) or the delay mapping

$$x^{(n+1)} = f(x^{(n)}, x^{(n-1)}, \ldots),$$

where f is unknown but observed data are available may also be modeled using the RBF or ANN approaches. Note that the domain of this mapping of delay coordinates may be of high enough dimension to preclude the use of other standard methods. A first look at this procedure in the context of nonlinear signal modeling is described in Lapedes & Farber (1987).

Yet another promising direction of the dynamical systems approach is fractal image compression. Here a picture may be stored as an *iterated function system*, that is two or more systems of the form of Equation (17) where f is now an affine transformation. An image can be produced by iterating this system and only a small number of parameters need be stored. See Barnsley & Hurd (1993) for many interesting examples as well as mathematical details.

The richness of the nonlinear signal processing field is in part due to technology. Optical signal processing is based on the application of the characteristics of light to process and transmit information. For example, in optical image processing, holograms are used to store data while lenses are used to compute two-dimensional Fourier transforms and may also perform edge detection and filtering tasks (Poon & Banerjee, 2001). Most recently, the field of fiber optics has arisen from the observation that light emanating from a laser and passing through a tiny pure glass pipe can transmit more information than electrical signals transmitted along a wire. The ability to send and amplify multiple signals simultaneously along an optic fiber has give rise to data transmission rates of 10 Gbps with higher rates in view. This technique is referred to as dense wavelength division multiplexing (DWDM) (Ramaswami & Sivarajan, 2001). New efforts to model and exploit such technologies suggest that the field of nonlinear signal processing is at the dawn of a new era. For example, the emerging area of photonics addresses how light can be used to represent and process signals and the new discipline of biophotonics concerns how light may be used to discover the mechanisms by which the basic building blocks of living matter function. Signal processing problems such as these may prove to have significant impact on the human condition including the battle against diseases such as cancer.

MICHAEL J. KIRBY

See also **Attractor neural network; Embedding methods; McCulloch–Pitts network; One-dimensional maps; Standard map**

Further Reading

Abarbanel, H.D.I. 1996. *Analysis of Observed Chaotic Data*, Berlin and New York: Springer

Barnett, W., Gallant, A.R., Hinich, M.J., Jungeilges, J.A., Kaplan, D.T. & Jensen, M.J. 1997. A single-blind controlled competition among tests for nonlinearity and chaos. *Journal of Econometrics*, 82: 157–192

Barnsley, M.F. & Hurd, L.P. 1993. *Fractal Image Compression*, Wellesley, MA: AK Peters

Haykin, S. 1994. *Neural Networks*, New York: Macmillan

Kirby, M. 2001. *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*, New York: Wiley

Lapedes, A. & Farber, R. 1987. *Nonlinear signal processing using neural neworks: prediction and system modelling*, Techinical Report LA-UR-87-2662, Los Alamos National Laboratory

Poon, T.-C. & Banerjee, P.P. 2001. *Contemporary Optical Image Processing with Matlab*, New York: Elsevier

Ramaswami, R. & Sivarajan, K.N. 2001. *Optical Networks: A Practical Perspective*, 2nd edition, San Francisco: Morgan Kaufmanm

NONLINEAR SUPERPOSITION

See ***N*-solition formulas**

NONLINEAR TOYS

Toys are ubiquitous and some are unique to particular cultures. Many action toys are intriguing because they perform in unexpected ways as a result of underlying nonlinear principles, yielding both amusement and surprises for the unwary. Here we give several examples of toys and devices that operate on linear and nonlinear physical principles.

Frog on a Swing

Several toys are based on the mechanism of *parametric excitation*. Although primarily a linear mechanism, an increase in amplitude in parametric excitation leads to the nonlinear range of amplitude-dependent phenomena. One can demonstrate parametric excitation by taking a string with a weight attached to its end and make it into a pendulum by hanging it over the forefinger (see Figure 1(a)). Now, force the weight into a small amplitude oscillation, and then move the weight up and down with the same phase as the oscillation by pulling on the string periodically. The result is an increase in the amplitude of the oscillation of the pendulum. The phase of the oscillation will vary slowly as the amplitude increases and the inherently nonlinear nature of the pendulum becomes important. The basic linearized equation for parametric excitation of the nonlinear pendulum is the Mathieu equation

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \left(\omega_0^2 - \alpha \sin 2\omega_0 t\right)\theta = 0$$

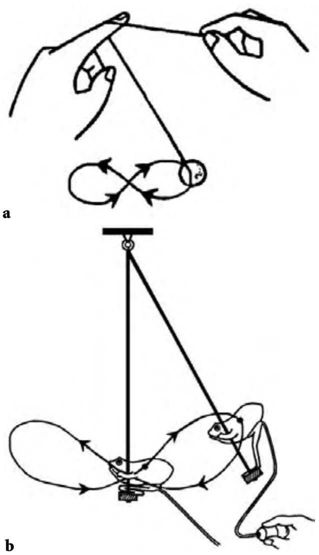

Figure 1. Parametric excitation demonstrated by (a) a pendulum over a finger and (b) a frog on a swing.

where θ is the angle of the pendulum from the vertical, ω_0 is the fundamental small amplitude frequency, and α is the amplitude of the periodically changing pendulum length.

This principle of parametric excitation is active in the toy called the "frog on a swing" (see Figure 1(b)). The frog in this toy is made of rubber and sits on a swing (pendulum). It can be inflated periodically with air using a tube and bulb. This causes the frog to stand up and sit down repeatedly and, hence, moves the center of mass of the frog up and down, as illustrated in Figure 1(b). With proper timing in inflating the bulb at the correct phase during each swing, the amplitude of the swing with the oscillating frog on it gradually increases. The same principle applies to the playground swing where children can increase the amplitude of their swing by "pumping" the swing in a standing position (Wirkus et al., 1998).

Handstand Pendulum

A related, but completely different, phenomenon occurs in the "handstand pendulum." The normal pendulum hangs stably downwards from its fulcrum as a result of gravity. However, applying a high-frequency forced oscillation to the fulcrum can cause the pendulum to be inverted, so that it now "hangs" stably pointing upwards

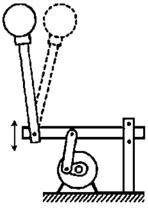

Figure 2. The handstand pendulum.

(Figure 2). Again the governing linearized equation is the Mathieu equation, but with the signs changed:

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \left(-\frac{g}{\ell} + 2\alpha \sin 2\omega t\right)\theta = 0$$

where g is the acceleration due to gravity and ℓ is the length of the pendulum arm.

Shishi-odoshi

In some Japanese gardens, there is a device that was designed originally to scare deer and other animals away from rice fields. Now, its purpose is either to scare away the deer that may intrude into a garden to eat the plants or to serve as an aesthetic fixture in the garden for its novelty and silence breaking "clack." This device, while not strictly a toy, is based on nonlinear principles and is called a *shishi-odoshi* (deer-scarer) or *shika-oi*. The *shishi-odoshi* is made from a thick bamboo culm with a length of three internal cells, that is, four nodes (the raised rings at regular intervals along the bamboo). At one end, the last node has been cut off and the end shaped to receive water (see Figure 3). From this open end, the internal membrane of the first node also has been removed so there is a cylindrical cup that is two cells long. A pivot rod is inserted into the bamboo between the two inner nodes so that when the tube is empty, the closed end of the bamboo rests on a stone. However, the position of the pivot is chosen so that as a steady stream of water flows into the open end of the tube, the center of mass of the bamboo and water shifts across the pivot and the open end of the tube will dip down, thus pouring out the water. The empty tube now returns quickly to its stable position on the stone with a loud "clack," which is the noise that is supposed to scare the deer.

The oscillatory motion of the bamboo tube caused by the steady flow of water is called a *relaxation oscillation* and can be regarded as an example of a self-induced oscillation. There are many toys that use this principle to change their orientation periodically, namely, water filling up a reservoir, and then the water being discharged as a result of the change of orientation of the toy.

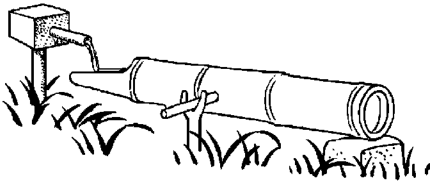

Figure 3. Bamboo deer alarm.

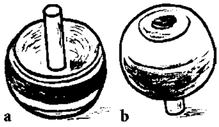

Figure 4. A tippe top in its stable resting position and stable rotating position.

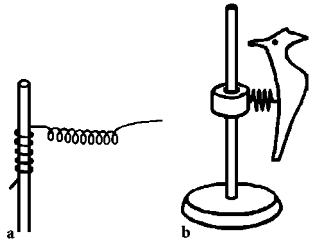

Figure 5. The pecking woodpecker with a detail of the spring in (a).

Tippe Top

A fascinating toy is the tippe top, which does not act like a conventional top. The tippe top consists of a little more than a hemisphere with a short cylindrical stem (see Figure 4(a)). It has a low center of gravity, so when placed on a surface, it will simply sit like an ordinary top on its hemispheric side. However, if the stem is held between a finger and the thumb and given a spin on the hemispheric end, then unlike a conventional top, the tippe top will readily turn over and continue to spin on the stem (see Figure 4(b)). As it turns over, the behavior of the tippe top is unusual in two respects: the center of gravity is raised and the spinning direction with respect to fixed body coordinates is reversed. When the top is spun, the low center of gravity is centrifugally moved away from the vertical spin axis. Before and after the top turns over, the angular momentum of the top about the vertical axis is the dominant momentum component. During the inversion process, the center of gravity is raised and the increase in potential energy reduces the rotational kinetic energy. The torque needed to execute

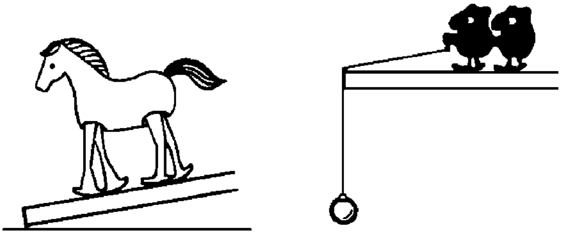

Figure 6. Two other toys that use self-induced oscillations.

this inversion comes from the sliding frictional forces between the top's hemispherical surface and the surface on which it is rotating. The mechanics of this top are described by nonlinear equations of motion that are solvable only in special cases (Cohen, 1977; Or, 1994; Gray & Nickel, 2001).

Pecking Woodpecker

A cute toy providing entertainment for young children is the pecking woodpecker going down a pole. To make this toy, coil a soft wire loosely around a smooth pole to make a soft spring. Now leave several turns of the coil on the pole and the remaining coils extended outwards from the pole (see Figure 5a). The weight of the protruding part of the spring will tilt the spring coiled around the pole so that the static friction is sufficient to keep it fixed on the pole. If one pushes the extended coil downwards and releases it, the extended coil vibrates up and down and the spring around the pole descends in a staccato fashion. The reason for the stuttering movement down the pole is that the coil around the pole oscillates periodically between being stopped by static friction and essentially free falling down the pole when the coil is aligned with the pole. Now attach a small wooden bird to the end of the extended wire, and the bird will peck the pole with its beak as it descends the pole (see Figure 5b).

There are many toys that move by the principle of self-induced excitation (see Figure 6). The basic equations for such oscillations are those of Duffing and van der Pol.

Morikazu Toda and Robert M. Miura

See also **Duffing equation; Equations, nonlinear; Laboratory models of nonlinear waves; Parametric amplification; Pendulum; Relaxation oscillators; Van der Pol equation**

Further Reading

Cohen, R.J. 1977. The tippe top revisited. *American Journal of Physics*, 45: 12–17

Gray, C.G. & Nickel, B.G. 2001. Constants of the motion for nonslipping tippe tops and other tops with round pegs. *American Journal of Physics*, 68: 821–828

Or, A.C. 1994. The dynamics of a tippe top. *SIAM Journal of Applied Maths*, 54: 597–609

Toda, M. 1979. *Toy Seminar.* Tokyo: Nihon-Hyoron-Sha (in Japanese)

Toda, M. 1982. *Toy Seminar (Continued).* Tokyo: Nihon-Hyoron-Sha (In Japanese)

Toda, M. 1983. *Mobile Toys.* Tokyo: Nihonkeizai-Shinbun-Sha (In Japanese)

Toda, M. 1993. *Science of Toys.* (6 vol.), Tokyo: Nihon-Hyoron-Sha, 1995 (In Japanese)

Wirkus, S., Rand, R. & Ruina, A. 1998. How to pump a swing. *The College Mathematics Journal*, 29: 266–275

NONLINEAR TRANSPARENCY

See **Nonlinear optics**

NONLINEARITY, DEFINITION OF

If a system's response to an applied force is directly proportional to the magnitude of that force, the system is said to be linear, otherwise the system is nonlinear. Nonlinear dynamics is concerned with systems whose time evolution equations are nonlinear.

As an example of a linear system, consider the one-dimensional motion of a point mass connected to a spring of force constant k, subject to a force

$$F_x(x) = -kx$$

in the x-direction. If k is changed, then the frequency and amplitude of resulting oscillations will change, but the qualitative nature of the behavior (simple harmonic oscillation in this example) remains the same. In fact, by appropriately rescaling the length and time axes, we can make the behavior for any value of k look just like that for some other value of k. For nonlinear systems, on the other hand, a small change in a parameter can lead to sudden and dramatic changes in both the qualitative and quantitative behavior of the system.

Although almost all natural systems are nonlinear, many of them respond in an approximately linear way provided that the amplitude of the force is small enough. This is one of the reasons why traditional linear approximations are so popular in science. Another reason is that the analytical solution of nonlinear equations is often difficult. With the development and availability of computers, however, numerical solutions of nonlinear equations became relatively easy in the 1970s. This focused attention on important nonlinear phenomena such as deterministic chaos, although the idea of dynamic chaos was first glimpsed by Henri Poincaré (Poincare, 1908). Nonlinearity and the phenomena that result from it have a strong bearing on the concept of determinism.

Given a set of variables (position, velocity, acceleration, pressure, the number of species in a chemical reaction, etc.) or a function of two or more independent variables that describe the state of a system, the subsequent time evolution can be represented as causal relationship between its present state and its next state in the future. The existence of causality in such relationships is suggested by all our experience of experimenting with Nature, and it corresponds to the deterministic perception of the evolution of dynamical systems.

Consider, for example, the oscillatory motion of a damped pendulum in the phase space of two variables, the angle θ and the velocity v

$$\dot{\theta} = v,$$

$$\dot{v} = -\gamma v - \omega_0^2 \sin(\theta),$$

where γ and ω_0^2 are parameters that describe damping rate and frequency of small oscillations. These time evolution equations can be written in operator form as

$$L\theta = \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \gamma\frac{\mathrm{d}\theta}{\mathrm{d}t} + \omega_0^2 \sin(\theta) = 0. \qquad (1)$$

Formally, the dynamics is linear if the causal relation between the present state and the next state is linear, otherwise it is nonlinear. In the above example, the dynamics will be linear if we replace the term $\beta \sin(\theta)$ with its linear approximation $\beta\theta$ near the asymptotically stable equilibrium state $\theta = 0,\ v = 0$. The corresponding linear operator in this case satisfies the property that

$$L^{(\mathrm{lin})}(ax + by) = aL^{(\mathrm{lin})}x + bL^{(\mathrm{lin})}y,$$

where a and b are real numbers and x and y are differentiable functions. The property of linear superposition plays a fundamental role in the analysis of linear equations of motion, and represents one of the basic principles of quantum mechanics (Landau & Lifshitz, 1977). If the property of linear superposition does not hold the dynamics is nonlinear.

On physical grounds, dynamical systems are in general nonlinear. Indeed, if we consider small deviations of a pendulum from its state of unstable equilibrium $\theta = \pi$, $v = 0$ and replace the nonlinear term by its linear approximation $\omega_0^2 \sin(\pi + \delta\theta) \approx -\omega_0^2\delta\theta$, we find unbounded growth of $\delta\theta$ in time. However, the experimentally observed motion of the pendulum is bounded; that is, the character of the motion changes when the deviation from the point of unstable equilibrium is large. Thus, nonlinearity corresponds to the situation that arises when the properties of the system depend directly on its state. The latter property, on one hand, makes a nonlinear system difficult to analyze but, on the other hand, gives rise to a rich diversity of nontrivial phenomena. We now consider briefly two consequences of nonlinearity.

First, the spectrum of oscillations of a nonlinear system is a complex function of its state (see, e.g., Sagdeev et al., 1988). For example, for weakly nonlinear oscillations of the pendulum near the state of stable equilibrium, we find (Hayashi, 1964), neglecting damping, that the frequency ω varies with the amplitude of the oscillations A as

$$\omega \approx \omega_0(1 - 3/64A^3).$$

The dependence of the frequency of oscillations on the state of the system is a key feature of many phenomena, including harmonic generation, synchronization, and the formation of solitons (*See* **Harmonic generation; Solitons; Synchronization**).

Second, a prominent effect of nonlinearity is the onset of deterministic chaos. Linear dynamical systems can have only three types of invariant sets: fixed points, cycles, and quasi-periodic orbits. The occurrence of a chaotic orbit is possible only in nonlinear systems and can be viewed in simple terms as arising from the interplay between the instability and nonlinearity. The instability is responsible for the exponential divergence of two nearby trajectories, whereas the nonlinearity bounds trajectories within a finite volume of the phase space of the system. The combination of these two mechanisms gives rise to a high sensitivity of the system to the initial conditions (*See* **Butterfly effect**). Consequently, the onset of deterministic chaos is restricted to nonlinear systems and involves repeated stretching and folding. The simplest example of a nonlinear system that demonstrates chaotic behavior through this process is a Bernoulli shift: $x_{n+1} = 2x_n \pmod 1$. An important consequence of chaos is that the predictability of even simple nonlinear systems is limited, so that ergodic theory has to be used to describe their statistical properties.

From these two brief examples it is already clear that in nonlinear science, the whole is more than the sum of its parts.

The complexity and diversity of the possible types of behavior in nonlinear systems have been widely explored in physics and chemistry, including analyses of solitons, nonlinear localization, pattern formation, and formulation of the general principles of self-organization.

The formalism of nonlinear systems provides a framework for understanding and modeling of the hierarchy of complexity in the life sciences. However, there are many difficulties to be encountered on the way. In particular, Hermann Haken (in Tschacher & Dauwalder, 1999) notes:

> "While in physics or chemistry it is not too difficult to define the microscopic and macroscopic levels, with respect to the brain we must ask what adequate intermediate levels we have to choose between microscopic and macroscopic."

An example of a practical approach to the solution of this problem arose during the investigation of neurons (*See* **Neurons**). Although the ionic current through the axon membrane is successfully described by the Hodgkin–Huxley equations, heart fibrillation is more conveniently modeled by the continuous FitzHugh–Nagumo equation, which is a simplified version of Hodgkin–Huxley system. On passing to models of neural networks, the hierarchy of connections between neurons structured in time turn out to be of prime importance, but this raises new questions for nonlinear science related to the appearance of closed causal loops. Continuing this hierarchy, one may expect that the application to cognitive science of complexity theory derived from nonlinear analysis will have a major impact on our understanding of reasoning, thinking, behavior, and psychology generally (Tschacher & Dauwalder, 1999).

DMITRY G. LUCHINSKY AND ANETA STEFANOVSKA

See also **Causality; Chaotic dynamics; Determinism; Equations, nonlinear; Linearization; Pendulum; Van der Pol equation**

Further Reading

Hayashi, C. 1964. *Non-linear Oscillations in Physical Systems*, New York: McGraw-Hill

Landau, L.D. & Lifshitz, E. 1977. *Quantum Mechanics: Non-Relativistic Theory*, Oxford: Pergamon (original Russian edition 1947)

Poincaré, H. 1908. *Science et méthode*, Paris: Flammarion

Sagdeev, R.Z., Usikov, D.A. & Zaslavskii, G.M. 1988. *Nonlinear Physics: From the Pendulum to Turbulence and Chaos*, Chur: Harwood Academic

Tschacher, W. & Dauwalder, J.-P. (editors). 1999. *Dynamics, Synergetics, Autonomous Agents: Nonlinear Systems Approaches to Cognitive Psychology and Cognitive Science*, Singapore: World Scientific

NON-NEWTONIAN FLUIDS

See **Fluid dynamics**

NON-TOPOLOGICAL SOLITON

See **Solitons, types of**

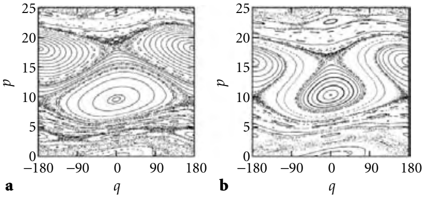

Figure 1. Standard nontwist map for $K = 1.5$ and (a) $\alpha = 0.036$, (b) $\alpha = 0.038$.

NONTWIST MAPS

Two-dimensional area-preserving mappings have become a standard tool for analyzing two-degrees-of-freedom autonomous and one-and-one-half dimensional time-periodic Hamiltonian systems (Lichtenberg & Lieberman, 1990). Prominent among such mappings are the radial twist maps

$$T : \begin{cases} p' = p - DU(q), \\ q' = q + f(p'), \end{cases} \tag{1}$$

where q and p are conjugate variables, $U(q)$ and $f(p)$ are smooth functions, and D is a derivative operator. For example, with $U = -K\cos q$ and $f(p) = p$, we recover the standard map, which has become a paradigm for chaotic Hamiltonian dynamics.

Usually $f(p)$ is monotonic, in which case the twist $\tau = \mathrm{d}f/\mathrm{d}p$ does not change sign. The twist condition, $\tau \neq 0$, is crucial for the validity of such linchpins of stability theory as the Moser twist theorem and the Kolmogorov–Arnol'd–Moser (KAM) theorem Arnol'd. Nevertheless, more and more instances have been found where the twist condition is violated, for example, plasma wave heating, zonal flows in planetary magnetospheres, beam dynamics in particle accelerators, and relativistic charged particle motion. The proliferation of such nontwist maps has challenged theorists to find alternatives to standard tools of nonlinear dynamics.

The Standard Nontwist Map

Many of the essential characteristics of nontwist maps are captured in the standard nontwist map (SNTM) (Howard and Hohs, 1984)

$$\begin{aligned} p' &= p - K\sin q, \\ q' &= q + p' - \alpha p'^2, \end{aligned} \tag{2}$$

where K and α are positive constants. Thus, the twist vanishes along the line $p = 1/2\alpha$. An equivalent version is discussed in del Castillo-Negrete et al. (1997). Generalizations employing higher order polynomial forms for $f(p)$ are discussed in Howard & Humphreys

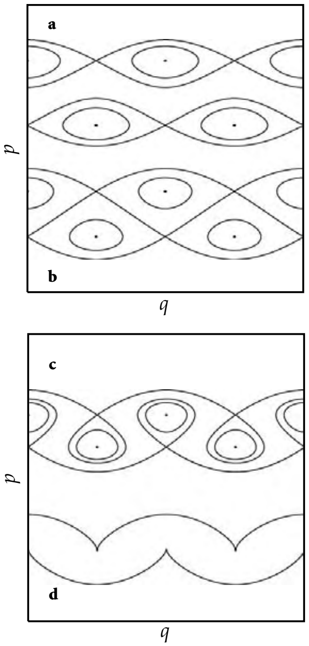

Figure 2. Reconnection scenario for standard nontwist map.

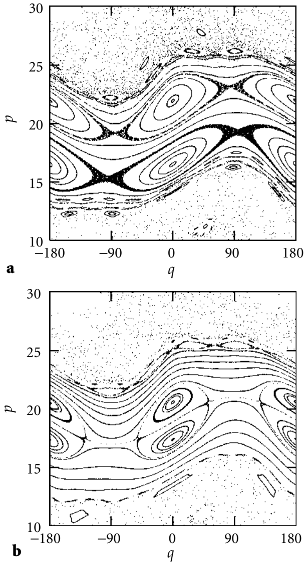

Figure 3. Vortex pair formation in the standard nontwist map for $K = 4.0$ and (a) $\alpha = 0.0260$, (b) $\alpha = 0.02634$.

(1995). Fixed points are located at $q \in (0, \pi)$ and

$$p_n^{\pm} = \frac{1 \pm \sqrt{1 - 8\pi n\alpha}}{2\alpha}. \tag{3}$$

For positive n both roots are positive real for $0 \le 8\pi n\alpha \le 1$, merging in a saddle-node bifurcation when $\alpha^* = (8\pi n)^{-1}$ at $p^* = 4\pi n$. Figure 1 shows the SNTM for $K = 1.5$ and two nearby values of α. Notice that, in contrast to the standard map, stable and unstable fixed points are staggered in phase rather than vertically aligned. This leads to a new kind of global bifurcation, called *reconnection* of KAM curves (by analogy with magnetic reconnection), which occurs when

$$\begin{aligned} K(\alpha) &\approx \frac{1}{2}\int_{p^-(\alpha)}^{p^+(\alpha)} [f(\xi;\alpha) - 2\pi n]\,\mathrm{d}\xi \\ &= \frac{(1 - 8\pi n\alpha)^{3/2}}{12\alpha^2}. \end{aligned} \tag{4}$$

The entire reconnection scenario is depicted schematically in Figure 2. We see that for a given configuration of X- and O-points, there are two topologically distinct arrangements of separatrices. Thus, in going from Figure 2(a) to (c), the upper separatrix is diverted downward, looping around the lower elliptic point. In between (b) the two separatrices merge, a process called *braiding*. Finally, Figure 2(d) shows the characteristic cusps resulting from the merging of X- and O-points via a saddle-node bifurcation. Naturally, the structure of the full mapping (Figure 1) is much more complicated than this approximate depiction. Of particular interest is the band of meandering curves seen in Figure 1(b), which constitute a new kind of barrier to chaotic transport. More complex sequences occur for more complex $f(p)$.

Higher-order fixed points also undergo reconnection. For example, the period-two islands near $p = 6\pi$ exhibit a second kind of reconnection, *vortex pair* formation, as shown in Figure 3. Here the elliptic and hyperbolic fixed points are vertically aligned, in contrast to the staggered pairs of period-one islands seen in Figure 1. An exhaustive classification of possible reconnection scenarios has not yet been accomplished.

As in the case of monotone twist maps, the location of higher order fixed points is facilitated by writing the map as a composition of involutions, $T = I_2 I_1$ ($I_1^2 = I_2^2 = I$). Fixed points of T then lie on the intersections of the invariant curves of I_1 and I_2. For the general twist map (1) we have

$$I_1 : \begin{cases} p' = p - DU, \\ q' = -q, \end{cases} \tag{5}$$

$$I_2 : \begin{cases} p' = p, \\ q' = -q + f(p). \end{cases} \tag{6}$$

Mappings with this decomposition are called *reversible*; that is, there exists a symmetry S such that $T^{-1} = STS$. Explicitly, $S = I_2$, so that $T^{-1} = I_1 I_2$.

Normal Forms

In investigating the structure of a 2-dimensional symplectic map near a stable fixed point (Howard & MacKay, 1987), it is useful to employ Birkhoff normal forms. Thus, the mapping T written in action-angle variables J and θ takes the simple form (Meyer & Hall, 1992)

$$\begin{aligned} J' &= J \\ \theta' &= \theta + 2\pi\Omega(J). \end{aligned} \tag{7}$$

Here the rotation number is given by

$$\Omega(J) = \omega + \tau_0 J + \tfrac{1}{2}\tau_1 J^2 + \cdots \tag{8}$$

and the twist by

$$\tau = \frac{\partial\Omega}{\partial J} = \tau_0 + \tau_1 J + \cdots, \tag{9}$$

where τ_0 and τ_1 are the zero and first-order twists. In this way it has been shown (Moeckel, 1990; Dullin et al., 1999) that a twistless torus is generated for any area-preserving map near a tripling bifurcation, where $\omega = \frac{1}{3}$.

Transition to Global Chaos

In many physical applications of nontwist maps, the transition to global chaos is of paramount importance, as space-spanning KAM curves (invariant rotational circles) act as barriers to particle or energy transport (Meiss, 1992; Lichtenberg & Lieberman, 1990). The breakdown of such barriers is usually studied via some kind of renormalization theory, such as that devised by Greene et al. (1981) for area-preserving maps satisfying the twist condition. Greene's method is based on the conjecture that the last KAM curve to break up with increasing perturbation strength has rotation number equal to the inverse of the golden mean, that is, $\omega_c = 1/\gamma = (\sqrt{5} - 1)/2$. Using the involution decomposition to locate a sequence of periodic orbits whose rotation number $\omega \to 1/\gamma$, he found the critical value $K_c = 0.9716\ldots$. Recently, del Castillo-Negrete et al. (1997) have succeeded in modifying Greene's semi-analytic method to obtain an accurate estimate for the breakup of the $\omega = 1/\gamma$ invariant circle for the SNTM. This calculation is complicated by the facts that (i) periodic orbits come in pairs for nontwist maps and may even fail to exist for certain parameters (K, α for the SNTM), and (ii) the sequence of periodic orbits actually has six convergent subsequences. Preliminary work has also been carried out on developing a suitable renormalization treatment for nontwist maps.

Extension to Higher Dimension

It is of interest to ask whether nontwist maps exist in higher dimension and what sort of reconnection processes might occur. This may be done in a natural way by letting $p, q \in \mathbb{R}^n$. The Birkhoff normal form then depends on the frequency map, $\Omega(J) = DS(J)$, where S is a scalar function of J. The twist, in turn, becomes the $n \times n$ Jacobian matrix $\tau(J) = D\Omega(J) = D^2S(J)$, and the frequency map suffers a singularity whenever $\det\tau$ passes through zero. Dullin & Meiss (2003) have determined the behavior of four-dimensional symplectic maps near such a singularity. Again, a twistless torus appears near a $1:3$ resonance but with interesting topological complications. In four dimensions the frequency map can have a fold or a cusp singularity—in higher dimensions other catastrophes can occur.

JAMES E. HOWARD

See also **Hamiltonian systems; Maps; Standard map; Symplectic maps**

Further Reading

Arnol'd, V.I. 1990. *Mathematical Methods of Classical Mechanics*, 2nd edition, New York: Springer

del Castillo-Negrete, D., Greene, J.M. & Morrison, P.J. 1997. Renormalization and transition to chaos in area preserving nontwist maps. *Physica* D, 100: 311–329

Dullin, H.R. & Meiss, J.D. 2003. Twist singularities for symplectic maps. *Chaos*, 13(1): 1–16

Dullin, H.R., Meiss, J.D. & Sterling, D. 1999. Generic twistless bifurcations. *Nonlinearity*, 13: 203–224

Greene, J., MacKay, R.S., Vivaldi, F. & Feigenbaum, M.J. 1981. Universal behavior in families of area-preserving maps. *Physica* 3D, 13: 468–486

Howard, J.E. & Hohs, S.M. 1984. Stochasticity and reconnection in Hamiltonian systems. *Physical Review* A, 29: 203–224

Howard, J.E. & Humphreys, J. 1995. Nonmonotonic twist maps. *Physics Letters* A, 29: 256–276

Howard, J.E. & MacKay, R.S. 1987. Spectral stability of symplectic maps. *Journal of Mathematical Physics*, 28: 1036–1051

Lichtenberg, A.J. & Lieberman, M.A. 1990. *Regular and Chaotic Dynamics*, 2nd edition, New York: Springer

Meiss, J.D. 1992. Symplectic maps, variational principles, and transport. *Reviews of Modern Physics*, 64: 795–848

Meyer, K.R. & Hall, G.R. 1992. *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, New York: Springer

Moeckel, R. 1990. Generic bifurcations of the twist coefficient. *Ergodic Theory of Dynamical Systems*, 10: 185–195

NORMAL FORMS THEORY

Many nonlinear systems can be modeled by ordinary, differential, or difference equations, and a central problem of dynamical systems theory is to obtain information on the long-time behavior of typical solutions. Because it is not possible, in general, to solve these equations explicitly, we identify particular solutions (such as equilibrium solutions or periodic solutions) and try to infer global information from the behavior of nearby solutions.

As an example, consider an equilibrium state such as the downward position of a pendulum. Looking at small perturbations of this position to identify nearby periodic orbits (corresponding to small oscillations of the pendulum) shows that the downward pendulum is indeed stable. Analysis of the inverted pendulum reveals that it is unstable, in the sense that nearby solutions do not stay close to the equilibrium position.

Mathematically, an equilibrium is a fixed point of a dynamical system, and the stability analysis is carried out by linearizing the system, that is, by replacing (close to the fixed point) the nonlinear equations by linear equations for the perturbations. The resulting linear system can be solved exactly, and the analysis of these solutions may give information about the behavior of the solutions of the nonlinear system around the fixed point. This method is extremely powerful when it works, and is the basis of much dynamical system analysis. In some cases, however, the behavior of the linear system may be entirely different from that of the nonlinear system, and no information can be obtained from the linear analysis. This implies that there is crucial information contained in the nonlinear terms.

Normal forms theory is a general method designed to extract this information. The basic idea is to compute a local change of variables to transform a nonlinear system into a simpler nonlinear system that contains only the essential nonlinear terms—those that cannot be neglected without drastically changing the nature of the system. Hopefully, the new system is either linear or can be solved explicitly.

Consider the normal form analysis of the zero fixed point of systems of two differential equations of the form

$$\begin{aligned}\dot{x}_1 &= a_{11}x_1 + a_{12}x_2 + f_1(x_1, x_2),\\ \dot{x}_2 &= a_{21}x_1 + a_{22}x_2 + f_2(x_1, x_2),\end{aligned} \tag{1}$$

where the dots denote time derivatives and f_1 and f_2 are nonlinear analytic functions whose Taylor expansion about the origin contains no linear term. Linear analysis of the fixed point at the origin is carried out by neglecting the nonlinear terms and considering the linear system $\dot{x}_1 = a_{11}x_1 + a_{12}x_2$, $\dot{x}_2 = a_{21}x_1 + a_{22}x_2$. The dynamics of this system is governed by the eigenvalues λ_1, λ_2 of the matrix (a_{ij}), which is assumed to be diagonalizable. If the eigenvalues have nonvanishing real parts, then the dynamics of the original nonlinear system is qualitatively equivalent to the dynamics of the linear system and the fixed point is stable if both real parts are negative and unstable as soon as one of the real parts is negative. However, if the eigenvalues are imaginary, the behavior of the nonlinear system cannot be inferred from analysis of the linear system, and the information on the stability of the origin is contained in the nonlinear terms. For example, the origin of the system $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1 + \alpha x_1^2 x_2$ is either stable or unstable depending on the sign of α.

The first step in the normal form analysis of system (1) is to introduce a linear change of variables so it reads

$$\begin{aligned}\dot{y}_1 &= \lambda_1 y_1 + g_1(y_1, y_2),\\ \dot{y}_2 &= \lambda_2 y_2 + g_2(y_1, y_2).\end{aligned} \tag{2}$$

Second, look for a near-identity change of variables in the form of power series $y_1 = z_1 + P_1(z_1, z_2)$, $y_2 = z_2 + P_2(z_1, z_2)$ that simplifies system (2). To this end, expand g_1, g_2 in power series and choose the coefficients of the series P_1, P_2 so that in the z_1, z_2 variables, the system $\dot{z}_1 = \lambda_1 z_1 + h_1(z_1, z_2)$, $\dot{z}_2 = \lambda_2 z_2 + h_2(z_1, z_2)$ becomes simpler than the original system. Optimally, $h_1 = h_2 = 0$, which would provide an exact linearization of the original system. In general, however, some nonlinear terms remain after transformations.

It turns out that the ability to exactly linearize the original system is intimately connected to the eigenvalues λ_1, λ_2. If either $(n_1 - 1)\lambda_1 + n_2\lambda_2 = 0$ or $n_1\lambda_1 + (n_2 - 1)\lambda_2 = 0$ for some positive integers n_1, n_2, the eigenvalues are said to be resonant (or in resonance), and in general one of the functions (h_1 or h_2) contains resonant terms of the form $z_1^{n_1} z_2^{n_2}$. In the absence of resonance, therefore, no nonlinear terms remain and exact linearization is achieved.

If the eigenvalues are purely imaginary, z_2 is the complex conjugate of $z_1 = z$, and there are infinitely many resonance conditions. The stability of the fixed point is then decided by the first non-vanishing coefficient of h_1. As an example, consider again the system $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1 + \alpha x_1^2 x_2$. After the linear change of variable $x_1 = y_1 + y_2$, $x_2 = \mathrm{i}(y_1 - y_2)$, we have

$$\begin{aligned}\dot{y}_1 &= \mathrm{i}y_1 + (\alpha/2)\left({y_1}^2 y_2 - y_1 {y_2}^2 - {y_2}^3 + {y_1}^3\right),\\ \dot{y}_2 &= -\mathrm{i}y_2 - (\alpha/2)\left({y_1}^3 - {y_1}^2 y_2 + y_1 {y_2}^2 + {y_2}^3\right).\end{aligned} \tag{3}$$

The normal form transformation to third order reads:

$$\begin{aligned}y_1 &= z_1 + (\mathrm{i}\alpha/8)\left(2\,{z_1}^3 + 2\,z_1 {z_2}^2 - {z_2}^3\right) + \text{h.o.t.},\\ y_2 &= z_2 + (\mathrm{i}\alpha/8)\left({z_1}^3 - 2\,{z_1}^2 z_2 - 2\,{z_2}^3\right) + \text{h.o.t.},\end{aligned} \tag{4}$$

where h.o.t. denotes higher-order terms of degree 5 and higher. Finally, the normal form becomes

$$\begin{aligned}\dot{z}_1 &= \mathrm{i}z_1 + (\alpha/2)z_1 z_2^2 + \text{h.o.t.}, \quad z_2 = z_1^*,\\ \dot{z}_2 &= -\mathrm{i}z_2 + (\alpha/2)z_2 z_1^2 + \text{h.o.t.}.\end{aligned} \tag{5}$$

From Equations (5), $\dot{\rho} = (\alpha/2)\rho^3 + \text{h.o.t.}$ where $\rho^2 = z_1 z_2$, which implies that the origin of the initial system is stable if $\alpha \leq 0$ and unstable otherwise.

If all coefficients of the normal forms vanish identically, then the fixed point is surrounded by an open set of periodic orbits (a nonlinear center). The downward position of the frictionless pendulum is an example of a center. Note, however, that the convergence of the power series P_1, P_2 is not guaranteed, in general, and further conditions on the eigenvalues and the transformation must be satisfied.

For a general N-dimensional system of differential equations, we can compute the linear eigenvalues $\lambda_1, \ldots, \lambda_N$. If the real part of one of these eigenvalues vanishes, there is no guarantee that the dynamics of the linear system is equivalent to the dynamics of the nonlinear system close to the fixed point. Again, one can find explicit near-identity power series change of variables that simplify the original system. If $n_1\lambda_1 + n_2\lambda_2 + \cdots + n_N\lambda_N = \lambda_i$ for any i between 1 and N and for positive integers $n_1, n_2, \ldots, n_N$, then the eigenvalues are in resonance, and the ith new equation will contain some resonant nonlinear terms. This resonance relation is one of the most fundamental relations in nonlinear dynamics. It determines whether linear modes (determined by the eigenvectors of eigenvalues λ) are coupled by the nonlinear terms. It is the same resonance relation that appears in different guises in the analysis of forced linear systems and in the resonance between frequencies in Hamiltonian and celestial mechanics. Essentially, in the absence of resonances, the system evolves following the linear modes and no interaction is possible. When resonances occur, the linear modes interact through the nonlinear terms and create complex dynamics.

Normal forms theory provide a systematic way to include the effect of the nonlinear terms. At the practical level, the method as presented tends to be rather tedious because the number of monomials that have to be taken into account grows rapidly with the dimension of the system and the degree of the normalizing transformation. There are several equivalent alternatives to the computation of normal forms, including the method of "amplitude equation" for ordinary and differential equations and the Birkhoff–Gustavson transformation for Hamiltonian systems. Nevertheless, at the conceptual level, normal forms theory is a central tool to understand the rich dynamics of nonlinear systems and this general framework can be used to study and give a rigorous foundation to many other problems beside stability.

A particularly important use of normal form theory is the theory of bifurcations. In order to identify generic bifurcations, one considers the parameters of the system $\mu_1, \ldots, \mu_M$ as additional variables satisfying trivial differential equations $\dot{\mu}_i = 0, \ i = 1, \ldots, M$ and studies the normal forms of this extended system of dimension $N + M$, revealing the nature of the bifurcation. Other applications of normal forms include the analysis of chaotic systems in systems of differential equations, the formation of patterns for partial differential equations, exponentially small effects in the splitting of separatrices, and the Painlevé theory of integrable systems.

ALAIN GORIELY

See also **Bifurcations; Damped-driven anharmonic oscillator; Painlevé analysis**

Further Reading

Arnol'd, V.I. 1973. *Ordinary Differential Equations*, Cambridge, MA: MIT Press (originally published in Russian, 1971); 3rd edition, Berlin and New York: Springer, 1992

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, New York: Springer

N-SOLITON FORMULAS

A one-soliton solution of an evolutionary nonlinear partial differential equation in one spatial dimension $u_t = F(u, u_x, u_{xx}, \ldots)$ is a solitary wave (pulse) of finite energy and momentum. As an example, the Korteweg–de Vries (KdV) one-soliton is the exact solution

$$u(x,t) = \frac{-2\kappa_1^2}{\cosh^2\left(\kappa_1(x-\xi_1) - 4\kappa_1^3 t\right)}, \tag{1}$$

to the KdV equation

$$u_t = -u_{xxx} + 6uu_x. \tag{2}$$

Note that Equation (1) has the shape of a negative well, moving at a speed that is proportional to the depth.

N-soliton solutions are characterized by the fact that, for $t \to -\infty$ and $t \to +\infty$, the shape of the solution is composed by a quasi-exact superposition of N non-interacting pulses of the form (1). This means that the shape of the solution for $t \to -\infty$ is a sequence of N wells, ordered according to their depth, the deepest being the leftmost. For $t \to +\infty$, the shape is also a sequence of wells, with now the deepest being the rightmost, each of them retaining the original shape. The solution pictorially describes a family of N individual pulses traveling at their own characteristic speeds, the faster ones catching up with the slower, and, after an interaction, reemerging with their individual shapes. The overall effect of the nonlinear interaction is a phase shift between the pulses (that is, a difference in the relative distances of the dips of the wells). It can be proved that, in the collision process of N solitons, the total phase shifts are the sum of phase shifts of two-soliton processes.

The presence of N-soliton solutions, for arbitrary N, is often regarded as the hallmark for integrability of a nonlinear PDE. Three main methods have been devised for constructing and studying N-soliton solutions:

- The inverse scattering method (ISM), based on the Gel'fand–Levitan–Marchenko (GLM) equations
- Hirota's formalism
- Bäcklund transformations

Here we present formulas making explicit reference to the ISM. For direct ways of finding N-soliton solutions, we refer to the entries on Hirota's method and on the Darboux and Bäcklund transformations.

In the ISM, the existence of N-soliton solutions is associated with the following mathematical feature.

A nonlinear integrable evolution equation (say, for one dependent variable u defined on the real line), is presented as a Lax (isospectral) equation for a linear operator $L(u)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}L(u) = [B(u), L(u)], \tag{3}$$

where the bracket indicates a commutator ($[B, L] \equiv BL - LB$). The spectrum of $L(u)$, as an operator in $L^2(R)$, is the union of the point spectrum $\lambda_1, \lambda_2, \ldots$ and the continuous spectrum, consisting of the intervals $I_1 \cup I_2 \ldots$. An N-soliton solution of (3) is a solution associated with an initial data $u(x, 0)$ whose spectrum contains only a finite number N of points. Such initial data are often called reflectionless potentials.

Given N-soliton initial data, the direct spectral problem for $L_0 = L(u(x, 0))$, gives (in addition to the spectrum) normalization coefficients β_i of the eigenfunctions corresponding to the eigenvalues λ_i, $i = 1, \ldots, N$. The evolution equations of the parameters $\beta_i = \beta_i(t)$ are linear equations with coefficients that depend on the constants λ_i. Accordingly, an N-soliton solution describes $2N$-dimensional families of solutions parametrized by the point spectrum of the linear operator and by the normalization coefficients β_i.

Two Soliton Solutions for Korteweg–de Vries and Sine-Gordon Equations

For the KdV equation (2) let $\lambda_1 = -\kappa_1^2$ and $\lambda_2 = -\kappa_2^2 |\kappa_1| > |\kappa_2|$ be the points in the spectrum of the Lax operator $L = -\mathrm{d}^2/\mathrm{d}x^2 + u$, and let ξ_i be quantities associated with the normalization coefficients $\beta_i(0)$ of the eigenvector of L. In other words, κ_i, ξ_i is the scattering data of the reflectionless potential $u(x, 0)$. A two-soliton solution for KdV is

$$u(x,t) = \frac{2\kappa_1^2 (\kappa_1 - \kappa_2)(\kappa_1 + \kappa_2)\left(\cosh^2(\phi_2) + (\kappa_2/\kappa_1)^2 \cosh^2(\phi_1) - 1\right)}{(\kappa_2 \cosh(\phi_1)\cosh(\phi_2) - \kappa_1 \sinh(\phi_1)\sinh(\phi_2))^2},$$

where $\phi_i = \kappa_i\,(x - \xi_i) - 4\,\kappa_i{}^3 t$.

This is an instance of the phenomenon described above. Indeed, for $t \to -\infty$ the solution describes two pulses traveling in the positive x-direction, with the faster one (that is, the one pertaining to the data κ_1, ξ_1) being the leftmost one.

The solution describes the process whereby the fast soliton catches up to the slow one, and finally (for $t \to +\infty$), the two emerge. The shape of the individual pulses remains unchanged, the effect of the nonlinear interaction being a phase shift. The fast soliton experiences a positive phase shift of

$$\Delta\phi_1 = \frac{1}{\kappa_1} \log \left| \frac{\kappa_1 + \kappa_2}{\kappa_1 - \kappa_2} \right|,$$

while the slow soliton loses the same phase (see Figure 1).

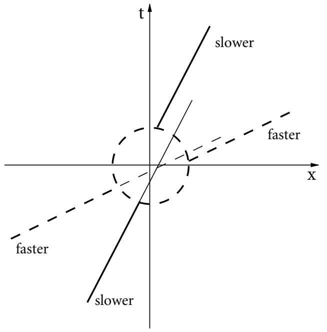

Figure 1. Phase shifts in two soliton collision.

The two-soliton solution of the sine-Gordon (SG) equation

$$\frac{\partial^2 u}{\partial^2 t} - \frac{\partial^2 u}{\partial^2 x} + \sin\, u = 0,$$

in terms of the points λ_1, λ_2 in the spectrum of the corresponding Lax operator and the constants c_1, c_2 is given by

$$u(x,t) = 4\mathrm{arctg}\frac{\frac{c_1}{2\lambda_1}\exp\left(-\frac{x - v_1 t}{\sqrt{1 - v_1^2}}\right) + \frac{c_2}{2\lambda_2}\exp\left(-\frac{x - v_2 t}{\sqrt{1 - v_2^2}}\right)}{1 - \frac{c_1 c_2}{4\lambda_1\lambda_2}\frac{(\lambda_1 - \lambda_2)^2}{(\lambda_1 + \lambda_2)^2}\exp\left(-\frac{x - v_1 t}{\sqrt{1 - v_1^2}} - \frac{x - v_2 t}{\sqrt{1 - v_2^2}}\right)},$$

where $v_i = (4\lambda_i^2 - 1)/(4\lambda_i^2 + 1)$, $i = 1, 2$. The fast soliton undergoes a positive phase shift given by

$$\Delta\phi_1 = \sqrt{1 - v_2^2}\log\frac{(\lambda_1 + \lambda_2)^2}{(\lambda_1 - \lambda_2)^2},$$

while the slow soliton undergoes a negative phase shift:

$$\Delta\phi_2 = -\sqrt{1 - v_1^2}\log\frac{(\lambda_1 + \lambda_2)^2}{(\lambda_1 - \lambda_2)^2}.$$

General Formulas

General formulas for N-soliton solutions of soliton equations are usually expressed in terms of determinants. For certain well-known soliton equations, one has the following results.

(1) Korteweg–de Vries (KdV) equation:

$$u_N(x,t) = -2\frac{\mathrm{d}}{\mathrm{d}x^2}\log\det(A),$$

where A is a $N \times N$ matrix with elements

$$A_{ij} = \delta_{ij} + \frac{\beta_i(t)}{\kappa_i + \kappa_j} \exp(-(\kappa_i + \kappa_j)x),$$

where $\beta_i(t) = \beta_i(0)\exp(8\kappa_i^3 t)$.

(2) Sine-Gordon (SG) equation:

$$u_N(x,t) = 2\mathrm{i}\log\frac{\det(A_+)}{\det A_-},$$

where

$$[A_\pm]_{ij} = \delta_{ij} \pm \frac{c_j}{\lambda_i + \lambda_j}\exp(\gamma_j(x,t)),$$

with $\gamma_j(x,t) = -i(\lambda_j + \frac{1}{4\lambda_j})t + i(\lambda_j - \frac{1}{4\lambda_j})x$.

(3) Nonlinear Schrödinger (NLS) equation:

$$\mathrm{i}\frac{\partial\psi}{\partial t} + \frac{\partial^2\psi}{\partial^2 x} - 2g|\psi|^2\psi\ , \quad g = \pm 1.$$

The N-soliton solutions are given by

$$\psi(x,t) = \frac{\mathrm{i}}{\sqrt{g}}\frac{\det M_1(x,t)}{\det M(x,t)},$$

where

$$M_{ij}(x,t) = \frac{1 + \bar{\gamma}_i(x,t)\gamma_j(x,t)}{\bar{\lambda}_i - \lambda_j}$$

and the $N+1 \times N+1$ matrix $M_1(x,t)$ is described as follows:

$$M_1 = \begin{array}{|ccc|c|}\hline & M & & G \\ & & & \\ \hline 1 & \dots & 1 & 0 \\ \hline\end{array}$$

where G is the vector $(\gamma_1(x,t), \dots, \gamma_N(x,t))$, with $\gamma_i(x,t) = \gamma_i^{(0)}\exp(\mathrm{i}\lambda_i x - \mathrm{i}\lambda_i^2 t)$.

(4) Infinite Toda Lattice (ITL): The ITL equations are compactly written in the Lax form:

$$\frac{\mathrm{d}L}{\mathrm{d}t} = [B, L],$$

where L is the infinite matrix with b_i on the principal diagonal and a_i on the first lower and upper diagonal, and B is the antisymmetric matrix having a_i on the first lower diagonal; the coefficients $\{a_i, b_i\}$ are related to the canonical variables $\{Q_i, P_i\}$ by the transformation

$$a_i = \tfrac{1}{2}\exp(-(Q_{i+1} - Q_i)/2), \quad b_i = \tfrac{1}{2}P_i.$$

N-soliton solutions are given by the formula:

$$\exp(-(Q_i - Q_{i-1})) = \frac{\det B_i \det B_{i-2}}{(\det B_{i-1})^2},$$

$$P_i = \text{const.} + \frac{\mathrm{d}}{\mathrm{d}t}\log\frac{\det B_{i-1}}{\det B_i}.$$

In this formula the B_i's are $N \times N$ matrices with elements given by

$$[B_i]_{j,k} = \delta_{j,k} + \gamma_j(t)\gamma_k(t)\frac{(\lambda_j\lambda_k)^{i+1}}{1 - \lambda_j\lambda_k}, \quad j,k = 1,\dots,N.$$

The λ_j are the N elements of the point spectrum of L, and the γ_j are normalization constants; the time dependence of the solution is given by

$$\gamma_j(t) = \gamma_j(0)\exp(\pm\sinh(\sigma_j)t), \quad \lambda_j = \pm\exp(-\sigma_j).$$

Further Comments

The formulas above can be interpreted as nonlinear superposition formulas for soliton solutions. Indeed, N-soliton solutions are constructed via the definition of suitable matrices determined by the spectral data associated with individual solitons. More precisely, the notion of nonlinear superposition principle can be better understood in the framework of the theory of Darboux–Bäcklund transformations.

In general, N-soliton solutions can be obtained as suitable limits of periodic solutions of the same system as the period tends to infinity. Periodic finite gap solutions are obtained by means of theta functions defined over Riemann surfaces. N-soliton solutions are obtained by letting suitable homology cycles of such Riemann surfaces shrink to a (degenerate) surface (with nodes and cusps) of genus equal to zero.

GREGORIO FALQUI AND TAMARA GRAVA

See also **Bäcklund transformations; Darboux transformation; Hirota's method; Inverse scattering method or transform; Theta functions**

Further Reading

Ablowitz, M.J. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM

Calogero, F. & Degasperis, A. 1982. *Spectral Transforms and Solitons: I*, Amsterdam: North-Holland

Drazin, P.G. & Johnson, R.S. 1989. *Solitons: An Introduction*, Cambridge and New York: Cambridge University Press

Faddeev, L.D. & Takhtajan, L.A. 1987. *Hamiltonian Methods in the Theory of Solitons*, Berlin: Springer

Miura, R.M. (editor). 1976. *Bäcklund Transformations*, Berlin: Springer

Newell, A.C. 1985. *Solitons in Mathematics and Physics*, Philadelphia: SIAM

Novikov, S.P., Manakov, S.V., Pitaevskii, L.P. & Zakharov, V.E. 1984. *Theory of Solitons*, New York: Consultants Bureau

Toda, M. 1988. *Theory of Nonlinear Lattices*, 2nd edition, Berlin: Springer

N-TORI

See **Quasiperiodicity**

NUMERICAL METHODS

Numerical methods are used for approximations of "exact" mathematical solutions, which either do not exist or are very complicated. With fast computer processors and advanced mathematical software, many problems become more attractive for numerical rather than for analytical solutions.

Computational Errors

Numerical computations are always uncertain because the results are only defined within the accuracy of a numerical error. Two main reasons account for errors in numerical approximations: finite precision in computer representation of numbers, and truncation of exact mathematical formulas in numerical algorithms.

Finite precision defines the tolerance interval, within which any further improvement in a numerical solution is impossible. For instance, the precision accuracy of MATLAB on the Windows platform is order of 10^{-16}. Therefore, it is meaningless to initialize irrational numbers (such as π or e) beyond the 16th digit after the period while carrying computations in MATLAB.

Round-off errors can be obstacles for accurate numerical approximations. Numerical differentiation algorithms are typically ill-posed since the round-off error increases with smaller step size of numerical discretization. Catastrophic cancellations can occur due to round-off errors, as in the example below when two large nearly identical numbers are subtracted:

$$f(x) = \sqrt{x+1} - \sqrt{x}.$$

If $x = 1.000000000000000 \times 10^{10}$, then $\sqrt{x+1} = 1.000000000050000 \times 10^5$, $\sqrt{x} = 1.000000000000000 \times 10^5$ and

$$f(x) = 4.999994416721165 \times 10^{-6}.$$

However, if the relative round-off error is order of 10^{-10}, the result is identical to zero. If the precision accuracy cannot be extended, a modification of the numerical procedure is required for a more accurate answer, as in the equivalent representation:

$$f(x) = \frac{1}{\sqrt{x+1} + \sqrt{x}}.$$

Within the same precision accuracy, the result is now $f(x) \approx 4.9999999999 \times 10^{-6}$.

Truncation errors occur due to chopping of an infinite series into a finite number of terms. For example, the Taylor series for analytical functions can be truncated with the Taylor polynomials, as in the example:

$$e^{x^2} = 1 + x^2 + \frac{1}{2!}x^4 + \frac{1}{3!}x^6 + E(x),$$

where $E(x)$ is the truncation error. The integral of e^{x^2} on $x \in [0, 1]$ is given in terms of a special function, called the error function. Equivalently, this integral can be approximated with the Taylor polynomial as

$$\begin{aligned} \int_0^1 e^{x^2}\,\mathrm{d}x &= \left(x + \frac{1}{3}x^3 + \frac{1}{10}x^5 + \frac{1}{42}x^7 \right)\Bigg|_{x=0}^{x=1} + E_{\mathrm{tr}} \\ &= 1.457142857142857 + E_{\mathrm{tr}}, \end{aligned}$$

where the truncation error is $E_{\mathrm{tr}} \approx 5.508914448636659 \times 10^{-3}$.

Numerical algorithms are classified by the rate of convergence and by their numerical stability (i.e., will a small error decay or grow through the successive iterations). Convergence and stability of numerical algorithms are studied with analysis of the truncation error. If the truncation error depends on the step size h of the finite numerical discretization and reduces as h^n, then the numerical algorithm is said to converge to an exact solution as the nth-order algorithm. As an example, we consider a numerical computation of the integral, given by irrational number e:

$$e = \int_0^1 f(x)\,\mathrm{d}x, \qquad f(x) = 1 + e^x.$$

Using the discretization of the unit interval with N equal subintervals of the step size $h = 1/N$, we can use the composite trapezoidal rule to approximate the integral as

$$e = \frac{h}{2}\left(f(0) + 2\sum_{n=1}^{N-1} f(hn) + f(1) \right) + E(h),$$

where the theoretical truncation error is $E(h) = \alpha_2 h^2$ and α_2 is constant, such that $\lim_{h\to 0}\alpha_2 \neq 0$. The composite trapezoidal rule converges to the integral as the second-order method. Indeed, computing the algorithm for $N = 100$ and 200, we have

$$\begin{aligned} N &= 100: \quad e \approx 2.718296147450418, \\ E(0.01) &= 1.431899137260828 \times 10^{-5} \end{aligned}$$

and

$$\begin{aligned} N &= 200: \quad e \approx 2.718285408211362, \\ E(0.005) &= 3.579752316795748 \times 10^{-6}, \end{aligned}$$

and, therefore,

$$\frac{E(0.01)}{E(0.005)} = 3.999995001169598 \approx 2^2 = 4,$$

as predicted by the theoretical formula. An improved numerical algorithm for integration is the composite Simpson's rule, defined as

$$\begin{aligned} e = \frac{h}{3}\Bigg(& f(0) + 4\sum_{n=1}^{N/2} f(h(2n-1)) \\ & + 2\sum_{n=1}^{N/2-1} f(2hn) + f(1) \Bigg) + E(h), \end{aligned}$$

where the theoretical truncation error is $E(h) = \alpha_4 h^4$ and N is even. The composite Simpson's rule converges to the integral as the fourth-order method. Indeed, computing the algorithm for $N = 100$ and 200, we have

$$\begin{aligned} N &= 100: \quad e \approx 2.718281828554504, \\ E(0.01) &= 9.545830792490051 \times 10^{-11} \end{aligned}$$

and

$$\begin{aligned} N &= 200: \quad e \approx 2.718281828465013, \\ E(0.005) &= 5.967226712755291 \times 10^{-12} \end{aligned}$$

and, therefore,

$$\frac{E(0.01)}{E(0.005)} = 15.99709756642108 \approx 2^4 = 16,$$

as predicted by the theoretical formula.

Iteration Methods

Solutions of algebraic, differential, partial differential, and integral equations can be approximated with iteration methods. Numerical errors in iteration methods may grow during iterations. When this happens, numerical approximations cannot give the exact solution because instabilities lead to huge numerical errors. If the numerical iterations are unstable, it does not matter how accurately the truncation error converges to zero with smaller discretization step size. Propagation and growth of numerical errors can be illustrated with the linear iteration map:

$$x_{n+1} = q x_n, \quad |q| \geq 1.$$

The exact solution of the linear iteration map is $x_n = q^n x_0$, where x_0 is the starting value. If two iteration sequences are obtained from almost identical values $x_0^{(1)}$ and $x_0^{(2)}$, so the initial distance $e_0 = |x_0^{(1)} - x_0^{(2)}|$ is small, then the distance grows with larger n as $e_n = |q|^n e_0$. Two iteration sequences diverge from each other, even if the small error e_0 was generated by the round-off error!

Numerical instabilities and divergences often occur in solutions of algebraic equations, which represent the simplest numerical problems. A scalar algebraic equation $f(x) = 0$ can be reformulated as a root finding algorithm:

$$x_*: \quad f(x_*) = 0.$$

If we plot $f(x)$ as a function of x, the root x_* can be immediately found from the graph $y = f(x)$, if it exists. Algorithmically, we can try to approximate the root x_* from the iteration method:

$$x_{n+1} = x_n + f(x_n).$$

If the limit $x_\infty = \lim_{n\to\infty} x_n$ exists, then the root x_* is a fixed point of the iteration method such that $x_\infty = x_*$. If the limit x_∞ does not exist but x_* is known to exist, the numerical method fails and iterations $\{x_n\}_{n=1}^{\infty}$ diverge due to numerical instabilities. Analysis of convergence of the iteration method can be performed with the linearization of the iteration method. Let $e_n = x_n - x_*$ be the small distance of x_n from the fixed point x_* such that

$$\begin{aligned} e_{n+1} &= x_n + f(x_n) - x_* = e_n + f(x_* + e_n) - f(x_*) \\ &= q e_n + \alpha_n e_n^2, \quad q = 1 + f'(x_*), \end{aligned}$$

where α_n is constant, such that $\lim\limits_{n\to\infty} \alpha_n$ exists. If $|q| \geq 1$, the error e_n grows with larger n and iterations $\{x_n\}_{n=1}^{\infty}$ and $\{e_n\}_{n=1}^{\infty}$ diverge. In this case, the fixed point x_* cannot be found from the iteration method. For example, when $f(x) = 3x - 6$, the fixed point $x_* = 2$ cannot be found iteratively, since $q = 4 > 1$.

An improved iteration algorithm is Newton's method which approximates a new point x_{n+1} from the root of the tangent line to the graph $y = f(x)$ at the previous point x_n. Thus

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n) + \mathrm{O}(x_{n+1} - x_n)^2,$$

such that

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Newton's method always converges to a fixed point x_*, if the fixed point x_* exists. The rate of convergence depends on whether the root $x = x_*$ is simple or multiple. For a single root when $f'(x_*) \neq 0$, the rate of convergence is quadratic in terms of the error e_n, since

$$q = \lim_{x\to x_*} \frac{f(x) f''(x)}{(f'(x))^2} = 0.$$

For a multiple root when $f(x) \sim (x - x_*)^m$ and $m > 1$, the rate of convergence is linear in terms of e_n, since $q = (m-1)/m$. When the function $f(x)$ is linear as in the example $f(x) = 3x - 6$, the convergence of the Newton–Raphson method occurs in a single iteration, no matter what x_0 is

$$x_1 = x_0 - \frac{3x_0 - 6}{3} = 2 = x_*.$$

Differential Equations

Numerical algorithms are particularly attractive for solutions of differential equations, which are widely used in all areas of physics and engineering. It is especially important because many applied differential equations cannot be solved in terms of analytical functions. The simplest differential equation is given by a scalar first-order quasilinear equation:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y),$$

starting with the initial condition: $y(0) = y_0$. Since $y(0) = y_0$ and $y'(0) = f(0, y_0)$ are known, it makes sense to step from $t = 0$ along the tangent line to the curve $y = y(t)$ at the point $(0, y_0)$. Using a number of steps from $t = 0$ to T with small step size h, we define a numerical approximation of $y(t_n)$ as y_n, such that $y(t_n) = y_n + e_n$, where e_n is the numerical error, accumulated after n steps. If all steps are taken alone the tangent line segments to the points (t_n, y_n), we have the Euler method (also known as the slope approximation):

$$y_{n+1} = y_n + hf(t_n, y_n).$$

When the time $t = T$ is reached in N steps, such that $T = Nh$, the global error is defined as $E_T = |y(T) - y_N|$. The global error for the Euler method is theoretically given by $E_T = \alpha_1 h$, such that the Euler method is the least accurate, first-order method. Figure 1 shows that oscillations of the undamped pendulum in the nonlinear second-order equation:

$$y'' + \sin y = 0$$

are destroyed in Euler's method. The inaccurate Euler's method introduces an effective damping due to numerical discretization.

Numerical methods for differential equations also produce numerical instabilities. For example, the linear first-order equation for exponential decay:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\lambda y, \quad \lambda > 0,$$

has a simple solution: $y(t) = y_0 \mathrm{e}^{-\lambda t}$. Euler's method applied to this equation is equivalent to the linear iteration map:

$$y_{n+1} = (1 - \lambda h) y_n,$$

which diverges for $h > 2/\lambda$, when $|1 - \lambda h| > 1$. When $\lambda \to 0$, the differential equation for the exponential decay becomes stiff for numerical solution, since the step size h for numerical discretization must be very small to preserve stability of numerical approximation.

Advanced algorithms for numerical solutions of differential equations are developed with numerical integration methods. For example, integrating the first-order quasi-linear equation with the trapezoidal rule on the interval $t \in [t_n, t_{n+1}]$, we derive the implicit iteration method:

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_n, y_n) + f(t_{n+1}, y_{n+1})\right].$$

The global; truncation method of the method is $E_T = \alpha_2 h^2$, that is, it is the second-order method. The method is also implicit since the unknown values of y_{n+1} appear at the left and right sides of the equation. The implicit methods can be solved with iterations at each n, by means of root finding algorithms. However, this modification makes the algorithm complicated and time-consuming. Predictor-corrector methods are used instead, either with single-step (Runge–Kutta) methods or with multi-step (Adams–Bashforth–Moulton) methods. For example, the second-order Runge–Kutta method (also known as the Heun's or improved Euler's method) is based on the trapezoidal rule in the form

$$\begin{aligned} p_{n+1} &= y_n + hf(t_n, y_n), \\ y_{n+1} &= y_n + \frac{h}{2}\left[f(t_n, y_n) + f(t_{n+1}, p_{n+1})\right], \end{aligned}$$

where p_{n+1} is the predicted value at $t = t_{n+1}$ by using the slope approximation and y_{n+1} is the corrected value at $t = t_{n+1}$ by using the trapezoidal rule. The predictor–corrector method above has the same global truncation error $E_{\mathrm{T}} = \tilde{\alpha}_2 h^2$ with a different numerical constant $\tilde{\alpha}_2$. Figure 1 shows that oscillations of the undamped pendulum are well preserved in the Heun's method based on predictions and corrections.

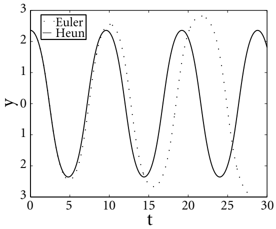

Figure 1. Oscillations of the undamped pendulum in Euler's and Heun's numerical methods.

The most popular and accurate method is the fourth-order Runge–Kutta method, which takes four computations of the function $f(t, y)$ for a single step from $t = t_n$ to $t = t_{n+1}$ and has the global truncation error $E_{\mathrm{T}} = \alpha_4 h^4$.

Higher-order predictor-corrector methods are still affected by numerical instabilities, especially in the case of stiff differential equations. More reliable methods for stiff problems are based on implicit integration formulas. Implicit methods can be rewritten as explicit methods if the function $f(t, y)$ is linear in y. For example, the equation for the exponential decay can be integrated with the implicit second-order method based on the trapezoidal rule, as follows:

$$y_{n+1} = \frac{2 - h\lambda}{2 + h\lambda} y_n.$$

Since $|(2 - h\lambda)/(2 + h\lambda)| \leq 1$ for any h and $\lambda > 0$, the implicit method is unconditionally stable, no matter how large the step size h is chosen. The accuracy of the method is still the same as in the second-order Heun's method, so the step size h must not be too large, to preserve accuracy of numerical approximation.

Partial Differential Equations

Partial differential equations are defined in two- and higher-dimensional domains. They represent most complicated and time-consuming problems for numerical methods. Initial values for partial differential equations at time $t = 0$ are supplemented by boundary values at the boundaries of a physical domain. A simple example of a linear partial differential equation is given by the heat equation, derived for a one-dimensional rod of finite length L:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \quad 0 < x < L, \quad t > 0,$$

such that

$$u(0,t) = u_0(t), \quad u(L,t) = u_L(t), \quad u(x,0) = \phi(x).$$

The heat equation describes temperature $u(x,t)$ as function of time $t \geq 0$ and the point $0 \leq x \leq L$ in the rod. The heat sources $f(x,t)$, temperature values at the end points $u_0(t)$ and $u_L(t)$ and initial temperature $\phi(x)$ are all given as physical parameters of the problem. In the finite-difference methods, numerical approximations of $u(x,t)$ are defined at the equally spaced numerical grid with points at $x_n = nh$, $n = 0, 1, \ldots, N$, where h is the step size, such that $L = Nh$. Numerical approximations of $u(x,t)$ are also evaluated with equal time step τ at points $t_k = k\tau$, $k = 0, 1, \ldots$. Denoting $u_{n,k}$ as numerical approximation of $u(x_n, t_k)$, we approximate the second x-derivative of $u(x,t)$ by the central difference, derived from Taylor series expansions:

$$\begin{aligned} u_{n+1,k} &= u_{n,k} + hu_x(x_n,t_k) + \frac{1}{2}h^2 u_{xx}(x_n,t_k) \\ &\quad + \frac{1}{3!}h^3 u_{xxx}(x_n,t_k) + \mathrm{O}(h^4), \\ u_{n-1,k} &= u_{n,k} - hu_x(x_n,t_k) + \frac{1}{2}h^2 u_{xx}(x_n,t_k) \\ &\quad - \frac{1}{3!}h^3 u_{xxx}(x_n,t_k) + \mathrm{O}(h^4), \end{aligned}$$

such that

$$\frac{\partial^2 u}{\partial x^2}(x_n,t_k) = \frac{u_{n+1,k} - 2u_{n,k} + u_{n-1,k}}{h^2} + \mathrm{O}(h^2).$$

Using the slope approximation, we perform the time step from $u_{n,k}$ to $u_{n,k+1}$ according to the explicit finite-difference scheme:

$$u_{n,k+1} = (1-2r)u_{n,k} + r(u_{n+1,k} + u_{n-1,k}) + \tau f_{n,k},$$

where $r = \tau/h^2$, $n = 1, \ldots, N-1$, and $k = 0, 1, \ldots$. All boundary and initial values $u_{0,k}$, $u_{N,k}$ and $u_{n,0}$ are incorporated in computations of the explicit method for $n = 1, N-1$, and $k = 0$. Numerical approximations of $u_{n,k}$ are only computed at internal points of the grid. The total error of the explicit method is a composition of the truncation error of order $\mathrm{O}(h^2)$ for the central difference approximation and the truncation error of order $\mathrm{O}(\tau)$ for the Euler's method. The explicit method is least accurate with respect to time step size τ. It is also an unstable method for $r > 0.5$ when $\tau > h^2/2$. Figure 2 shows numerical solution of the heat equation with the explicit method for $f(x,t) = 0$, $u_0 = u_L = 0$, $\phi = \sin(\pi x)$, $L = 2$, $h = 0.1$, and $r = 0.55$. Development of instabilities of the explicit method destroys validity of the numerical approximations.

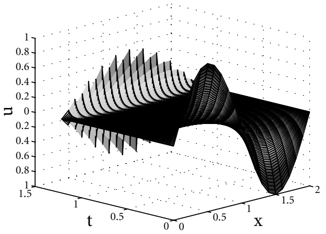

Figure 2. Solutions of the heat equation with the explicit method.

Implicit methods are more reliable in numerical computations. The implicit method, which is based on the trapezoidal rule of integration, is referred to as the Crank–Nicholson method. The method results in a linear system of equations:

$$\begin{aligned} A(r)\mathbf{u}_{k+1} &= A(-r)\mathbf{u}_k + \frac{\tau}{2}(\mathbf{f}_k + \mathbf{f}_{k+1}) \\ &\quad + \frac{r}{2}(\mathbf{b}_k + \mathbf{b}_{k+1}), \qquad k \geq 0, \end{aligned}$$

where

$$\mathbf{u}_k = \begin{bmatrix} u_{1,k} \\ u_{2,k} \\ \vdots \\ u_{N-2,k} \\ u_{N-1,k} \end{bmatrix}, \quad \mathbf{f}_k = \begin{bmatrix} f_{1,k} \\ f_{2,k} \\ \vdots \\ f_{N-2,k} \\ f_{N-1,k} \end{bmatrix},$$

$$\mathbf{b}_k = \begin{bmatrix} u_{0,k} \\ 0 \\ \vdots \\ 0 \\ u_{N,k} \end{bmatrix}$$

and

$$A(r) = \begin{bmatrix} 1+r & -r/2 & 0 & \cdots & 0 \\ -r/2 & 1+r & -r/2 & \cdots & 0 \\ 0 & -r/2 & 1+r & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+r \end{bmatrix}.$$

Although solving of the linear system at each time step is a time-consuming operation, the Crank–Nicholson

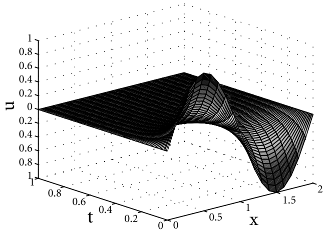

Figure 3. Solutions of the heat equation with the Crank–Nicholson method.

method is more useful compared with the explicit method. In particular, the method has a symmetric numerical error $O(\tau^2 + h^2)$ and unconditional numerical stability for any $\tau > 0$. Figure 3 shows numerical solution of the heat equation with the Crank–Nicholson method for $h = 0.1$ and $r = 1$ when $\tau = h^2$. The numerical solution is stable in time evolution.

Finite-difference methods are used successfully in numerical solution of other boundary-value problems for partial differential equations. Explicit methods have a very straightforward form. Implicit methods are unconditionally stable and result in linear systems for linear differential equations. Implicit methods are usually replaced by the semi-implicit algorithms for nonlinear differential equations. As a drawback, finite-difference methods have low accuracy. More advanced (and also more involved) shooting, finite-element, and spectral methods are used for improved solutions of boundary-value problems.

Shooting methods are based on transformations of the boundary-value problems into the initial-value problems, which are iterated with the root finding algorithms. For simplicity, we illustrate with ordinary differential equations, although multi-dimensional shooting methods are also applied to partial differential equations. We consider a numerical approximation of radially symmetrical bound states of the two-dimensional nonlinear Schrödinger equations, which satisfy the boundary-value problem:

$$\frac{d^2\Phi}{dr^2} + \frac{1}{r}\frac{d\Phi}{dr} - \Phi + 2\Phi^3 = 0, \quad 0 < r < R,$$

such that

$$\Phi'(0) = 0, \quad \Phi(R) = 0,$$

where R is large length of the approximation interval. Using a shooting numerical method, we consider a solution of the initial-value problem for the same equation $\Phi = \Phi_m(r)$, such that

$$\Phi_m(0) = m, \quad \Phi'_m(0) = 0,$$

where the value of m is unknown. Solving the differential equation on $r \in [0, R]$ with an appropriate numerical method, such as the fourth-order Runge–Kutta method, we define the function of parameter m as

$$E(m) = \Phi_m(R).$$

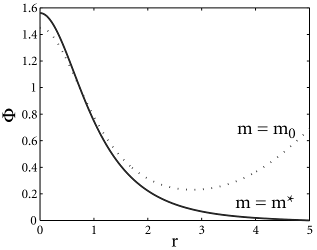

Figure 4. Bound state of the nonlinear Schrödinger equation with the numerical shooting method: starting (dotted) and final (solid) approximations.

The function $E(m)$ has a zero at $m = m_*$ if $\Phi_{m_*}(r)$ is the bound state of the nonlinear Schrödinger equation, such that $\Phi_{m_*}(R) = 0$. Starting with any value m_0, we can approach to the zero of $E(m)$ with a root finding algorithm, such as the Newton's method:

$$m_{k+1} = m_k - \frac{E(m_k)}{E'(m_k)},$$

where the value of

$$E'(m_k) = \left.\frac{\partial \Phi_m(R)}{\partial m}\right|_{m = m_k}$$

can be found from a linearization problem or from a secant method. The starting approximation $\Phi_{m_0}(r)$ and the final solution $\Phi_{m_*}(r)$ are shown on Figure 4 by dotted and solid lines, respectively. The value of $E(m_k)$ converges to zero, such that the sequence $\Phi_{m_k}(r)$ converges to the bound state $\Phi(r)$ of the nonlinear Schrödinger equation.

Finite-element methods approximate solutions of the boundary-value problems from a variational problem, which involves minimization of the energy functionals. The bound state $\Phi(r)$ of the nonlinear Schrödinger equation coincides with a minimizer of the energy functional:

$$H(u) = \frac{1}{2}\int_0^R \left[\left(\frac{du}{dr}\right)^2 - u^4\right] r\, dr,$$

such that the variation of $H(u)$ vanishes if and only if $u = \Phi(r)$:

$$\left.\frac{\delta H}{\delta u}\right|_{u = \Phi(r)} = -\frac{d}{dr}\left[r\frac{d\Phi}{dr}\right] - 2r\Phi^3 = 0.$$

Dividing the interval $r \in [0, R]$ into n subintervals $[0, r_1] \cup [r_1, r_2] \cup \cdots \cup [r_{N-1}, r_N]$, where $r_n = hn$ and $R = hN$, we approximate $\Phi(r)$ by a linear combination of the finite elements $u_n(r)$:

$$u(r) = \sum_{n=0}^{N} \Phi_n u_n(r).$$

The finite elements $u_n(r)$ satisfy the constraints:

$$u_n(r_n) = 1, \quad \text{and} \quad u_n(r_m) = 0, \ m \neq n,$$

such that Φ_n are unknown values for the finite-element approximation of $u(r_n)$. The simplest finite elements (also known as the simplex elements) are given by the piecewise linear interpolation:

$$u_n(r) = \begin{cases} 0, & r \leq r_{n-1}, \\ \frac{r-r_{n-1}}{h}, & r_{n-1} \leq r \leq r_n, \\ \frac{r_{n+1}-r}{h}, & r_n \leq r \leq r_{n+1}, \\ 0, & r \geq r_{n+1}. \end{cases}$$

When $u(r)$ is substituted into the energy functional $H(u)$ and integrated numerically over $0 \leq r \leq R$, the function $H(u) = H(\Phi_0, \Phi_1, \ldots, \Phi_{N-1})$ with $\Phi_N = 0$ can be minimized with respect to parameters $\Phi_0, \Phi_1, \ldots, \Phi_{N-1}$. The minimization leads to a system of nonlinear algebraic equations:

$$\frac{\partial H(u)}{\partial \Phi_n} \equiv h_n(\Phi_0, \Phi_1, \ldots, \Phi_{N-1}) = 0, \quad n = 0, 1, \ldots, N-1.$$

This nonlinear system can be solved with a matrix root finding Newton's method, which completes the finite-element approximation of the bound states $\Phi(r)$. When the differential equations are linear and the exact integrations are replaced with the trapezoidal rule, the finite-element method with simplex elements recovers the finite-difference approximation. However, the finite-element method is more general and suitable for accurate numerical approximations. For example, the finite-element method can be used together with the Simpson's rule of numerical integration.

Spectral methods are based on Fourier series solutions of differential equations in finite intervals, subject to appropriate boundary conditions. Solutions of the homogeneous heat equation $u_t = u_{xx}$ on $x \in [0, L]$ with zero boundary conditions at the ends: $u(0,t) = u(L,t) = 0$ are given by the Fourier sine-series:

$$u(x,t) = \sum_{m=1}^{\infty} B_m(t) \sin \frac{\pi m x}{L},$$

where the set of Fourier amplitudes $B_m(t), m \geq 1$ solves the initial-value problems:

$$\frac{\mathrm{d}B_m}{\mathrm{d}t} = -\frac{\pi^2 m^2}{L^2} B_m, \quad m \geq 1.$$

The initial values of $B_m(0)$ are defined from $u(x,0)$ by the Fourier integrals:

$$B_m(0) = \frac{2}{L} \int_0^L u(x,0) \sin \frac{\pi m x}{L} \, \mathrm{d}x.$$

Truncating the Fourier series with the Galerkin method and defining the Fourier approximation at grid points $x_n = nh$, $n = 1, \ldots, N-1$, we replace the exact Fourier series solution with the discrete Fourier transform:

$$u(x_n, t) = \sum_{m=1}^{N-1} B_m(t) \sin \frac{\pi m n}{N},$$

where $B_m(t)$ solves the same initial-value problem for $m = 1, \ldots, N-1$ but $B_m(0)$ is defined by the inverse discrete Fourier transform:

$$B_m(0) = \frac{2}{N} \sum_{n=1}^{N-1} u(x_n, 0) \sin \frac{\pi m n}{N}.$$

We notice that the inverse discrete Fourier transform follows from Fourier integrals by means of the numerical trapezoidal rule. Solving systems of initial-value problems with Runge–Kutta methods, we define solutions of the heat equation at time steps $t_k = k\tau$, $k = 0, 1, \ldots$.

Spectral methods give more accurate solutions of the heat equation, compared with the finite-difference methods. First, all initial-value problems for Fourier coefficients $B_m(t)$ are uncoupled. Second, the spectral methods have superior accuracy, because the truncation error of the Galerkin approximation decays exponentially with larger number N of the Fourier amplitudes. Third, spectral methods can be applied to nonlinear differential equations. Performing time steps with the Runge–Kutta methods, discrete Fourier transform and its inverse can be employed for computations of the nonlinear terms. As a result, the spectral method is effectively uncoupled for Fourier amplitudes $B_n(t)$ as an explicit single-step method. Split-step spectral methods are especially important for nonlinear evolution equations, such as the nonlinear Schrödinger equations.

As computational power increases, the role of numerical methods will be increasingly important in coming years for efficient solutions of many problems in physical and engineering sciences. Presently, computations of three-dimensional equations for water waves, turbulence, and astrophysical problems can be run on workstations in reasonable time. Modern supercomputing resources based on multiprocessor clusters are designed to help scientists in their theoretical studies of nature. It is common nowadays that many research groups in physics and engineering replace analysis of a problem by numerical methods.

DMITRY PELINOVSKY

See also **Averaging methods; Compartmental models; Extremum principles; Lattice gas methods; Stability**

Further Reading

Akai, T.J. 1994. *Applied Numerical Methods for Engineers*, New York: Wiley

Chilling, R.J. & Harris, S.L. 2000. *Applied Numerical Methods for Engineers using MATLAB and C*, Pacific Grove, CA: Brooks/Cole

Fausett, L.V. 2003. *Numerical Methods: Algorithms and Applications*, Upper Saddle River, NJ: Prentice-Hall

Gerald, C.F. & Wheatley, P.O. 1999. *Applied Numerical Analysis*, 6th edition, Reading, MA: Addison-Wesley

Rao, S.S. 2002. *Applied Numerical Methods for Engineers and Scientists*, Upper Saddle River, NJ: Prentice-Hall

N-WAVE INTERACTIONS

Linear dispersive waves propagate according to a dispersion relation

$$\omega = \omega(\boldsymbol{k}),$$

which relates the frequency ω to the wave vector $\boldsymbol{k}$. In linear systems, waves of different wave vectors and frequencies do not interact due to the superposition principle. In nonlinear systems, the lowest-order nonlinear effect (expanding in the wave amplitudes) is the interaction between N waves of different wave vectors $\boldsymbol{k}_j$ and frequencies ω_j, which satisfy the spatial and temporal resonance conditions (Zaslavsky & Sagdeev, 1988)

$$\sum_{j=1}^{N} \boldsymbol{k}_j = \boldsymbol{0}, \quad \sum_{j=1}^{N} \omega(\boldsymbol{k}_j) = 0.$$

The simplest three-wave resonant interaction occurs if there are nontrivial solutions with

$$\boldsymbol{k} = \boldsymbol{k}_1 + \boldsymbol{k}_2, \quad \omega(\boldsymbol{k}) = \omega(\boldsymbol{k}_1) + \omega(\boldsymbol{k}_2). \tag{1}$$

Resonant wave interactions are governed by the dispersion relation. Some dispersion relations do not exhibit nontrivial solutions of the resonance conditions (1) for any $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$. Others may exhibit resonant three-wave interactions, for example the deep-water gravity-capillary waves satisfying the dispersion relation $\omega^2(\mathbf{k}) = g|\mathbf{k}| + T|\boldsymbol{k}|^3$, where g is the acceleration of gravity and T is the surface tension coefficient. When $T \neq 0$, the resonant configurations of the three waves occur already in one spatial dimension, as shown on a graphical solution on Figure 1. The solid curve on this figure shows the dispersion relation $\omega = \omega(k)$, while the dashed curve shows the same dispersion relation shifted relatively to the point (k_1, ω_1), such that $\omega - \omega_1 = \omega(k - k_1)$ and $\omega_1 = \omega(k_1)$. The intersection of the two curves defines a solution of the resonance equations (1) at the point $k = k_1 + k_2$, where $\omega(k) - \omega(k_1) = \omega(k_2)$.

When $T = 0$, the resonant three-wave configurations do not occur in space of any dimensions, but deep-water gravity waves satisfying the dispersion relation $\omega^2 = g|\boldsymbol{k}|$ exhibit the resonant four-wave interactions of the following type:

$$\begin{aligned} \boldsymbol{k}_1 + \boldsymbol{k}_2 &= \boldsymbol{k}_3 + \boldsymbol{k}_4, \\ \omega(\boldsymbol{k}_1) + \omega(\boldsymbol{k}_2) &= \omega(\boldsymbol{k}_3) + \omega(\boldsymbol{k}_4). \end{aligned}$$

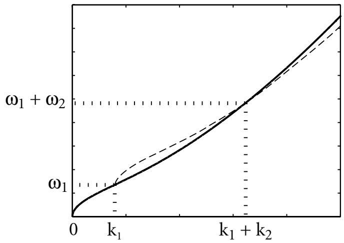

Figure 1. Graphical solution of Equations (1) for three resonant waves.

Given a dispersion relation $\omega(\boldsymbol{k})$ with one or more branches, one needs to compute all lowest-order resonances for $N = 3, 4$, and so on, in order to describe nonlinear dynamics of resonant wave interactions with the normal form analysis (Craig, 1996).

Time evolution of N resonant waves can be studied with an asymptotic multiscale expansion method. In this method, a solution of the system $u = u(\boldsymbol{x}, t)$ is assumed to be close to a linear superposition of N resonant waves with slowly varying amplitudes, where $\boldsymbol{x} = (x, y, z)$. For instance, resonant interaction of three waves is described by the following expansion (Gaponov-Grekhov & Rabinovich, 1992):

$$\begin{aligned} u(\boldsymbol{x}, t) = \varepsilon \Big[& a(\varepsilon \boldsymbol{x}, \varepsilon t) e^{i(\boldsymbol{k}\boldsymbol{x} - \omega(\boldsymbol{k})t)} \\ & + a_1(\varepsilon \boldsymbol{x}, \varepsilon t) e^{i(\boldsymbol{k}_1 \mathrm{x} - \omega(\boldsymbol{k}_1)t)} \\ & + a_2(\varepsilon \boldsymbol{x}, \varepsilon t) e^{i(\boldsymbol{k}_2 \boldsymbol{x} - \omega(\boldsymbol{k}_2)t)} \Big], \end{aligned}$$

where ε is a small parameter and ε^2 terms are dropped. The evolution equations for the wave amplitudes take the form

$$\begin{aligned} i\left(\frac{\partial}{\partial t} + \boldsymbol{v} \cdot \nabla\right) a &= \gamma a_1 a_2 e^{-i\Delta t}, \\ i\left(\frac{\partial}{\partial t} + \boldsymbol{v}_1 \cdot \nabla\right) a_1 &= \gamma_1 a \bar{a}_2 e^{i\Delta t}, \\ i\left(\frac{\partial}{\partial t} + \boldsymbol{v}_2 \cdot \nabla\right) a_2 &= \gamma_2 a \bar{a}_1 e^{i\Delta t}. \end{aligned} \tag{2}$$

Here $\Delta = \omega(\boldsymbol{k}) - \omega(\boldsymbol{k}_1) - \omega(\boldsymbol{k}_2)$ is the frequency detuning from the exact resonance; $\boldsymbol{v}$, $\boldsymbol{v}_1$, and $\boldsymbol{v}_2$ are group velocities of the three waves, for example $\boldsymbol{v} = \nabla \omega(\boldsymbol{k})$; and γ, γ_1, and γ_2 are coupling coefficients in systems with quadratic nonlinearities. In nonlinear dispersive systems with conserved energy, parameters γ, γ_1, and γ_2 are real.

Typical dynamics of the resonant three-wave interaction can be studied from system (2) for space-

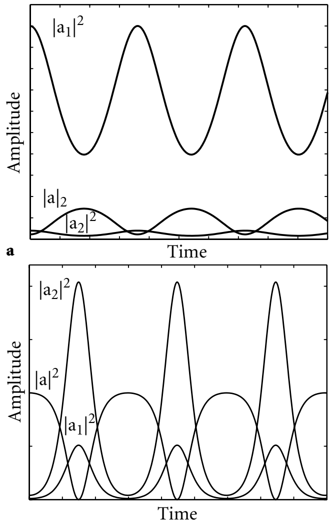

Figure 2. Two typical time evolutions of the three-wave resonant interaction: (a) the high-frequency wave a could not be generated from large low-frequency wave a_1 and small low-frequency wave a_2; (b) the large high-frequency wave a decays into pair of low-frequency waves a_1 and a_2.

independent waves $a(t)$, $a_1(t)$, and $a_2(t)$ in the case of exact resonance, when $\Delta = 0$. Without loss of generality, the real coupling coefficients can be normalized to be equal and positive: $\gamma = \gamma_1 = \gamma_2 > 0$. After these simplifications, system (2) has the following (Manley–Rowe) constants of motion:

$$|a|^2 + |a_1|^2 = C_1, \quad |a|^2 + |a_2|^2 = C_2,$$
$$|a_1|^2 - |a_2|^2 = C. \tag{3}$$

The time evolution of wave amplitudes $a(t)$, $a_1(t)$, and $a_2(t)$ displays two typical scenarios shown in Figure 2(a,b). If the wave with low frequency ($\omega_1, \omega_2 < \omega$) is initially pumped into the system (e.g., either $|a_1|^2 \gg |a|^2, |a_2|^2$ or $|a_2|^2 \gg |a|^2, |a_1|^2$ at $t = 0$), then the other two resonant waves remain small throughout the whole time evolution (see Figure 2(a)). Indeed, it is seen from Equations (3) that when $|a|^2$ grows, $|a_1|^2$ and $|a_2|^2$ decay. If either $|a_1|^2$ or $|a_2|^2$ were initially small, their decay (and the growth of $|a|^2$) is limited by a small value of C_1 and C_2.

If the wave of high frequency ω is initially pumped, however, the waves of low frequencies ω_1 and ω_2 may grow simultaneously. In this case, the resonant three-wave interactions display an exchange of energy between the high-frequency wave and the two low-frequency waves (see Figure 2(b)). This process is referred to as the "splitting" or "decaying" instability of the high-frequency wave. As seen from system (2) for small a_1 and a_2, the amplitude $a(t)$ does not change in time in the first-order approximation, while amplitudes $a_1(t)$ and $a_2(t)$ grow exponentially as $\sim e^{|\gamma a|t}$. The high-frequency wave $|a|^2$ decays into the two low-frequency waves $|a_1|^2$ and $|a_2|^2$, but the two waves merge back to the high-frequency wave later in the time evolution (see Figure 2(b)).

Resonant nonlinear three-wave interactions are known in many physical situations, such as in water waves, nonlinear optics, acoustics, and plasma physics. In nonlinear optics, resonant three-wave interactions are used in parametric amplifiers and oscillators, $\chi^{(2)}$ optical materials, self-induced transparency, and stimulated Raman scattering (Kaup et al., 1977). Four-wave mixing occur in optical communications and leads to growth of ghost pulses in an optical signal sequence. In water waves, resonant wave interactions explain effects of weak turbulence of gravity and gravity-capillary waves, as well as Rossby waves in the atmosphere (Zakharov, 1998). In plasma physics, nonlinear dynamics of high-temperature plasmas in a magnetic fields involve wave-particle and wave-wave interactions. Resonant three-wave interactions occur in ionospheric propagation, plasma heating with high-power electromagnetic sources, microwave sources, and laser sources (Kaup et al., 1977).

DMITRY PELINOVSKY

See also **Dispersion relations; Frequency doubling; Manley–Rowe relations**

Further Reading

Craig, W. 1996. Birkhoff normal forms for water waves. *Contemporary Mathematics*, 200: 57–74

Gaponov-Grekhov, A.V. & Rabinovich, M.I. 1992. *Nonlinearities in Action. Oscillations, Chaos, Order, Fractals*, Berlin: Springer

Kaup, D., Reiman, A. & Bers, A. 1977. Space–time evolution of nonlinear three-wave interactions. *Reviews of Modern Physics*, 51: 915

Zakharov, V.E. (editor). 1998. *Nonlinear Waves and Weak Turbulence*, Providence, RI: American Mathematical Society

Zaslavsky, G.M. & Sagdeev, R.Z. 1988. *Introduction to Nonlinear Physics: From the Pendulum to Turbulence and Chaos*, Moscow: Nauka (in Russian)

O

ONE-DIMENSIONAL MAPS

The term *one-dimensional map* usually indicates a dynamical system with discrete time, generated by some map f of a one-dimensional (1-d) space (the real line or an interval of it, a circle, or a graph) onto itself or, equivalently, the iterations of this map. To specify a 1-d space, the terms *interval map*, *circle map*, or *graph map* are used. The most famous examples of a 1-d maps are the so-called logistic map $x \mapsto \lambda x(1-x)$ of the interval [0, 1] and the rotation $x \mapsto x+\alpha$ of the circle. The theory of 1-d maps can also be considered as the qualitative theory of the difference equations of the form $x_{n+1} = f(x_n)$, $n = 0, 1, 2 \ldots$.

Being part of the general theory of dynamical systems, the theory of 1-d maps includes studies relating to topological dynamics, symbolic dynamics, combinatorial dynamics, differential (or smooth) dynamics, and ergodic theory. The theory of 1-d maps is well developed and contains many deep and beautiful results. Even the simplest non-invertible (for example, quadratic) 1-d maps possess orbits with intricate dynamics, including those behaving like random processes. Thus, this theory offers comparatively simple tools for the understanding of general laws that govern the progression of real dynamic processes from regular to chaotic behaviors.

Many mathematical models, including those arising in population biology, can be reduced to investigations of 1-d maps. For organisms with non-overlapping generations, the population growth can be modeled with the difference equation $x_{n+1} = f(x_n)$, where x_n is the population density of the nth generation. When the density of the population is small and there are plenty of resources, the density x_n increases and follows approximately the linear law $x_{n+1} = \lambda x_n$, where λ is the reproduction coefficient. Because resources are bounded, density cannot increase forever; moreover, if the density x_n is very large, that is, close to 1, then x_{n+1}—the density of the next generation—must be close to 0. The logistic function $f(x) = \lambda x(1-x)$ with $0 < \lambda \leq 4$ is a canonical example.

Interval Maps

Let $f : J \to J$ be a map of the interval J onto itself and f^n be its iterations, defined by $f^n(x) = f(f^{n-1}(x))$, $n = 1, 2, \ldots,$ $f^0(x) = x$. The orbit of a point $x_0 \in J$ under the map f is the sequence of points $x_n = f^n(x_0)$, $n = 0, 1, 2, \ldots$.

A point $\beta \in J$ is a periodic point if $f^m(\beta) = \beta$ for some natural m; the smallest m with this property is the *period* of β; if $m = 1$, β is called a fixed point. The points $f^n(\beta)$, $n = 0, 1, \ldots, m-1$ form a cycle of period m. To characterize the asymptotic behavior of an orbit, the set of its limiting points is used; it is called the ω-limit set of the orbit of x and denoted by $\omega(x)$. A cycle B is attracting if there exists a neighborhood U of B such that for any $x \in U$, $f^n(x) \in U$ for all $n > 0$ and $\omega(x) = B$. The basin of an attracting cycle B is the set $\{x \in J : \omega(x) = B\}$; it consists of a finite or countable number of (open) intervals. A cycle B is repelling if there exists a neighborhood U of B such that for any point $x \in U \setminus B$, there exists n such that $f^n(x) \notin U$. If f is differentiable and has a cycle $B = \{\beta_1, \ldots, \beta_m\}$, the quantity $\mu(B) = f'(\beta_1) \cdot \cdots \cdot f'(\beta_m)$ is the multiplier of B. B is attracting if $|\mu(B)| < 1$, and B is repelling if $|\mu(B)| > 1$. The smallest closed set that contains ω-limit sets of orbits for almost all points of J (with respect to Lebesgue measure), is called the (global) attractor of f and denoted by $\mathbf{A}$ below.

If f is a monotonic (and hence, invertible) continuous function, its dynamics are very simple. If f is monotonically increasing, the ω-limit set of each orbit consists of only a fixed point. If f is monotonically decreasing, the ω-limit set of each orbit is either a fixed point or a cycle of period 2.

The orbits of nonmonotonic (non-invertible) maps can be very complicated. Maps with a single extremum point are called unimodal. Many properties of continuous interval maps can be demonstrated with the simplest representative of unimodal maps—the family of logistic maps $f : x \mapsto \lambda x(1-x)$.

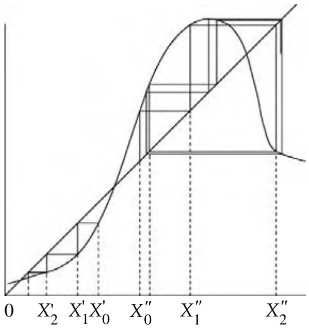

Figure 1. This Königs–Lamerey diagram gives a graphic representation of the orbit of x_0. If the graphs of two functions $y = f(x)$ and $y = x$ in the plane (x, y) are drawn, then the broken line consisting of vertical and horizontal segments that connect the points $(f^n(x_0), f^n(x_0))$ and $(f^n(x_0), f^{n+1}(x_0))$, $n = 0, 1, 2 \ldots$, demonstrates the movement along the orbit. The orbit of x'_0 tends to a fixed point, and the orbit of x''_0 tends to a cycle of period 3.

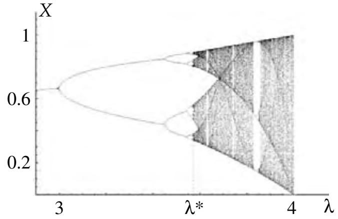

Figure 2. Bifurcation diagram of the attractors for the family of logistic maps.

Logistic Maps

The bifurcation diagram in Figure 2 shows what attractors the logistic map has when the parameter λ is changed.

With increasing λ up to $\lambda^* \approx 3.5699$, the consecutive doubling of period for attracting cycles holds and cycles of periods 2^n, $n = 0, 1, 2, \ldots$, appear: when $\lambda > 3$, the attracting fixed point $\beta_1 = 1 - 1/\lambda$ becomes repelling and a new attracting cycle of period 2 arises; for $\lambda > 1 + \sqrt{6} = 3.449\ldots$, this cycle becomes repelling and generates an attracting cycle of period 4, and so on. The period doubling for attracting cycles occurs with the speed

$$\frac{\lambda_{2^{n+1}} - \lambda_{2^n}}{\lambda_{2^{n+2}} - \lambda_{2^{n+1}}} \to 4.6692\ldots$$

(Feigenbaum–Coullet–Tresser constant), where λ_{2^n} is the parameter value when the cycle of period 2^n arises. The basin of each such attracting cycle is the interval $J = [0, 1]$ except for a countable set of repelling periodic points and their preimages.

If $\lambda = \lambda^*$, f has cycles of periods 2^n, $n = 0, 1, 2, \ldots$, only, and these cycles are repelling. The set K of limiting points for these cycles is an invariant Cantor set. The orbit of each point from K is everywhere dense on K. K is the attractor: the ω-limit set of each point from J, different from a periodic point or its preimages, coincides with K. The restriction of f^2 on the interval $[f^{-1}(\beta_1), \beta_1]$ is topologically conjugated to f on J, and hence, f is infinitely renormalizable.

For $\lambda = 3.83$, f has an attracting cycle of period 3, which is the attractor, and repelling cycles of all periods. The basin of the attracting cycle is J except for a Cantor set. This Cantor set is the closure of the set consisting of points of repelling cycles and their preimages.

If $\lambda = 4$, f is topologically expanding: for any interval $\widehat{J} \subset J$, there exists a number $m = m(\widehat{J})$ such that $f^m(\widehat{J}) = J$. Therefore, f has the property of sensitive dependence on initial conditions (sometimes called the "butterfly effect"). J is an attractor; periodic points lie everywhere dense on J. The map is topologically conjugate to the piecewise linear map

$$g : x \mapsto g(x) = \begin{cases} 2x, & 0 \le x \le 1/2, \\ 2(1-x), & 1/2 < x \le 1 \end{cases}$$

with the help of $h(x) = (2/\pi) \arcsin \sqrt{x}$. Because the Lebesgue measure is invariant under g, f has the invariant measure $\nu(\mathrm{d}x) = \mathrm{d}h(x) = \frac{1}{\pi}\mathrm{d}x \,/\sqrt{x(1-x)}$. This means, in particular, that for almost every orbit of f and for any $a_1, a_2 \in [0, 1]$, the frequency with which the orbit visits the interval (a_1, a_2) to $\int_{a_1}^{a_2} \nu(\mathrm{d}x)$.

The general characteristics of the logistic maps family are formulated as follows: for each $\lambda \in [0, 4]$, the attractor $\mathbf{A}$ of f is a cycle, or a Cantor set (as in the case $\lambda = \lambda^*$), or a cycle of intervals, that is, several closed non-intersecting intervals J_n, $n = 1, \ldots, m$, such that $f(J_n) = J_{(n+1) \bmod m}$ and $\cup_{n=1}^{m} J_n$ contain an everywhere dense orbit (in case $\lambda = 4$, $m = 1$).

The parameter values for which $\mathbf{A}$ is a cycle form an open and dense (on [0, 4]) set, and from the topological standpoint, a map with an attracting cycle is typical for the logistic family. The set of λ for which $\mathbf{A}$ is a Cantor set (and the map is infinitely renormalizable) has Lebesgue measure zero. The set of λ for which $\mathbf{A}$ is a cycle of intervals has positive Lebesgue measure. Moreover, for almost all value of these λ, $\mathbf{A}$ is the support of an ergodic absolutely continuous invariant measure.

The complication of dynamics in the logistic family stems, first of all, from the appearance of new attracting cycles resulting from the period-doubling bifurcation or the tangent bifurcation. With increasing λ, the multiplier of arising attracting cycle decreases from $+1$ to -1, and thereafter, this cycle becomes repelling and does not disappear in the sequel. If λ_m denotes the infinum (greatest lower bound) of the parameter values for which f has an attracting cycle of period m, then $\lambda_{m'} < \lambda_m$ for $m' \prec m$, where "$\prec$" is the following ordering of the natural numbers:

$$1 \prec 2 \prec 2^2 \prec \cdots \prec 2^n \prec \cdots \prec 7 \cdot 2^n \prec 5 \cdot 2^n \prec 3 \cdot 2^n \prec \ldots \prec 9 \prec 7 \prec 5 \prec 3,$$

called "Sharkovsky's ordering." For $\lambda = \lambda_{2^n}$, $n > 0$, the cycle of period 2^n arises as a result of a period-doubling bifurcation; for $\lambda = \lambda_m$, $m \neq 2^n$, $n = 1, 2, \ldots$, the cycle of period m arises as a result of a tangent bifurcation.

If f has an attracting cycle of period m, then with increasing λ, the sequence of period-doubling bifurcations takes place until cycles of periods $2^n m$, $n = 1, 2, \ldots$, appear. For the limiting parameter value for this sequence of bifurcation values, the attractor is a Cantor set (as in the case $\lambda = \lambda^*$).

Smooth Maps

Analytic unimodal maps (briefly, AU-maps) and C^3-smooth unimodal maps with negative Schwarzian

$$Sf = f'''/f' - \tfrac{3}{2}\left(f''/f'\right)^2 \tag{1}$$

(briefly, SU-maps) are the most investigated classes of smooth maps. Any logistic map has negative Schwarzian.

As for logistic maps, the attractor **A** of each AU-map or SU-map includes (in general, not coincides with) one of the previous set: an (attracting) cycle, a Cantor set, or a cycle of intervals and always contains the ω-limit set of the extremum point. For an SU-map, **A** can have, in addition, only one attracting fixed point, and for an AU-map, **A** can have, in addition, any finite number of attracting cycles. If **A** is a cycle or a cycle of intervals, outside of **A**, f has only a finite number of isolated repelling cycles and isolated repelling invariant Cantor sets.

The period-doubling and tangent bifurcations are typical for smooth maps. The birth order of attractive cycles for the logistic family is kept, for example, for families of convex SU-maps of the form $\lambda f(x)$, $\lambda > 0$.

For almost all members of a typical family of SU-maps, the attractor is either a cycle or a cycle of intervals; the cycle of intervals is the support of an ergodic absolutely continuous invariant measure; and there is a closed invariant interval $\widehat{J} \subset J$ containing the extremum point and $m > 0$ such that the restriction of f^m on $\widehat{J}$ is topologically conjugated to some logistic map.

For a typical family of AU-maps, there exists an open and dense set of parameters for which the map attractor consists of a finite number of cycles, and the union of basins of these cycles has full Lebesgue measure.

A smooth (at least C^2) unimodal map f is structurally stable if an attractor consists of cycles; the multiplier of each cycle is not equal to ± 1; the extremum point c is nondegenerate (i.e., $f''(c) \neq 0$); and $f^n(c)$ is not a periodic point for any $n = 0, 1, 2 \ldots$. The class of structurally stable maps is open in the C^2 topology and dense in any smooth topology.

Some properties of unimodal maps (possibly somewhat modified) hold for maps with more than two branches of monotonicity. For example, the number of attracting cycles for any C^3-smooth map with negative Schwarzian is not greater than the number of extremum points plus two.

Continuous Maps

Due to the natural ordering of points on the real line and the continuity of f, the values of f on a finite set of points provides rich information about the behavior of orbits. So, if there exist points $\beta_1 < \beta_2 < \beta_3$ such that $f(\beta_1) = \beta_1$, $f(\beta_2) = \beta_3$, and $f(\beta_3) \le \beta_1$, then f has cycles with any period.

Each finite set $B = \{\beta_1 < \beta_2 < \cdots < \beta_m\}$ such that $f(B) \subseteq B$ generates a permutation π on the set $\{1, 2, \ldots, m\} : \pi(i) = j$, if $f(\beta_i) = \beta_j$, $i = 1, \ldots, m$. If B is a cycle, such a permutation is cyclic. For $J_i = [\beta_i, \beta_{i+1}]$, $i = 1, 2, \ldots m - 1$, $f(J_i) \supseteq J_{k(i)} \cup \cdots \cup J_{K(i)}$, where $k(i) = \min\{\pi(i), \pi(i+1)\}$, $K(i) = \max\{\pi(i), \pi(i+1)\} - 1$. This allows the construction of a Markov matrix of admissible transitions with elements $a_{ij} : a_{ij} = 0$, if $f(J_i) \not\supset J_j$, and $a_{ij} = 1$, if $f(J_i) \supset J_j$, and the corresponding Markov graph. An analysis of loops of Markov graph allows us to obtain the Sharkovsky theorem on the cycles coexistence: if a continuous interval map has a cycle of period m, then it has a cycle of any period m' such that $m' \prec m$.

Research on the coexistence of combinatorial objects, such as cycles, permutations, cyclic permutations, and graphs, is the subject of combinatorial dynamics. An important part of symbolic dynamics for 1-d maps is the kneading theory.

There are different criteria for the chaotic behavior of orbits. In particular, the following properties are equivalent: (i) the *topological entropy* of f is positive; (ii) there is a cycle with period $\neq 2^n$, $n \ge 0$; (iii) there is a *homoclinic orbit*; (iv) there is an ω-limit set containing a cycle, but different from it; and (v) there exist m and a closed invariant set $M \subset J$ such that the restriction of f^m on M is topologically conjugate to the *shift* on the *space of one-side sequences of two symbols*. Any of these involves the following: there is a continuum of many orbits such that for any two orbits $\{x'_n\}$ and $\{x''_n\}$, $\liminf_{n \to \infty} |x'_n - x''_n| = 0$, $\limsup_{n \to \infty} |x'_n - x''_n| > 0$.

If f has a cycle of intervals $\{J_1, J_2, \ldots, J_m\}$, then f on $\widehat{J} = \cup_{i=1}^m J_i$ is topologically expanding and has the property of the sensitive dependence on initial conditions, and the set of periodic points is everywhere dense on $\widehat{J}$.

The typical behavior of continuous maps differs from that of smooth maps. In particular, in the space of continuous maps with C^0-topology, the set of maps that have cycles of all periods contains an open and dense subset. Moreover, almost every continuous (but nowhere differentiable!) interval map has infinitely many minimal Cantor sets that attract almost all orbits.

Discontinuous interval maps constitute another important class of 1-d maps, especially for applications including expanding interval maps and so-called interval exchange maps.

Circle Maps

Many of the properties of interval maps mentioned above are true for circle maps $f : S \to S$. Along with this, there are classes of circle maps with specific properties, for example, circle homeomorphisms semi-conjugated to rotation of S with rotation number α (corresponding to the discontinuous interval map: $\widehat{f} : x \mapsto x + \alpha \bmod 1$). If α is a rational number, the attractor **A** consists of cycles; if α is irrational, **A** is a minimal set and coincides with S or a Cantor set. Bifurcations in a family of circle homeomorphisms are described by Farey's sequences.

Circle maps, semi-conjugated to the discontinuous interval map $\widehat{f} : x \mapsto k\,x \bmod 1$ with an integer $|k| > 1$, are called circle maps of degree k and represent a further important class of circle maps, including expanding circle maps with $|f'(x)| > 1$ at $x \in S$. Any expanding map is topologically conjugate to the shift on the space of one-sided sequences of finite number of symbols; it is structurally stable and possesses an absolutely continuous invariant measure.

A.N. SHARKOVSKY AND V.V. FEDORENKO

See also **Anosov and Axiom-A systems; Bifurcations; Butterfly effect; Chaotic dynamics; Denjoy theory; Horseshoes and hyperbolicity in dynamical systems; Maps in the complex plane; Markov partitions; Measures; Mixing; Period doubling; Population dynamics; Routes to chaos; Sinai–Ruelle–Bowen measures; Symbolic dynamics**

Further Reading

Alseda, L., Llibre, J. & Misiurewicz, M. 2000. *Combinatorial Dynamics and Entropy in Dimension One*, 2nd edition, River Edge, NJ: World Scientific

Argonsky, S.J., Bruckner, A.M. & Laczkovich, M. 1989. Dynamics of typical continuous functions. *Journal of the London Mathematical Society*, 40: 227–243

Avila, A., Lyubich, M. & de Melo, W. 2003. Regular or stochastic dynamics in real analytic families of unimodal maps. *Inventiones Mathematicae*, 3(154): 451–550

Block, L. & Coppel, W.A. 1992. *Dynamics in One Dimension*, Berlin: Springer

Collet, P. & Eckmann, J.-P. 1980. *Iterated Maps on the Interval as Dynamical Systems*, Boston: Birkhäuser

de Melo, W. & van Strien, S.J. 1993. *One-Dimensional Dynamics*, Berlin: Springer

Devaney, R.L. 1992. *A First Course in Chaotic Dynamical Systems: Theory and Experiment*, Reading, MA: Addison-Wesley

Hao, Bai-Lin. 1989. *Elementary Symbolic Dynamics and Chaos in Dissipative Systems*, Singapore: World Scientific

Katok, A. & Hasselblatt, B. 1995. *Introduction to the Modern Theory of Dynamical Systems*, Cambridge and New York: Cambridge University Press

Kozlovski, O.S. 2003. Axiom A maps are dense in the space of unimodal maps in the C^k topology. *Annals of Mathematics*, 1 (157): 1–43

Milnor, J. & Thurston, W. 1988. On iterated maps of the interval. In *Dynamical Systems: Proceedings University of Maryland, 1986–87*, Berlin and New York: Springer, 425–563

Peitgen, H.-O., Jürgens, H. & Saupe, D. 1993. *Chaos and Fractals: New Frontiers of Science*, Berlin and New York: Springer

Sharkovsky, A.N., Kolyada, S.F., Sivak, A.G. & Fedorenko, V.V. 1997. *Dynamics of One-Dimensional Maps*, Boston and Dordrecht: Kluwer

OPEN SETS

See **Topology**

OPTICAL FIBER COMMUNICATIONS

The existence of temporal solitons in optical fibers and their use for optical communications were suggested in 1973 (Hasegawa & Tappert, 1973), and by 1980, such solitons had been observed experimentally (Mollenauer et al., 1980; Hasegawa & Kodama, 1995). Since then, rapid progress has converted temporal solitons into a practical candidate for designing modern communication systems based on optical-fiber technology (Agrawal, 2002; Kivshar & Agrawal, 2003).

Similar to other types of solitons, those in optical fibers emerge from a balance between the group-velocity dispersion (GVD) and self-phase modulation (SPM) induced by the Kerr nonlinearity. The GVD broadens optical pulses during their propagation inside an optical fiber, except when the pulse is initially chirped (compressed with linear frequency modulation) in the right way. More specifically, a chirped pulse can be compressed during the early stage of propagation whenever the GVD parameter β_2 and the chirp parameter C happen to have opposite signs such that $\beta_2 C$ is negative. The optical pulse then propagates undistorted in the form of an optical soliton.

The nonlinear Schrödinger (NLS) equation governing pulse propagation inside optical fibers can be written in the following form:

$$\mathrm{i}\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial \tau^2} + |u|^2 u = 0\,, \tag{1}$$

where τ is a measure of time from the pulse center and is normalized to the input pulse width T_0, and L_D is the dispersion length. Noting that $z = Z/L_\mathrm{D}$, the soliton period Z_0 is defined as

$$Z_0 = \frac{\pi}{2}L_\mathrm{D} = \frac{\pi}{2}\frac{T_0^2}{|\beta_2|}. \tag{2}$$

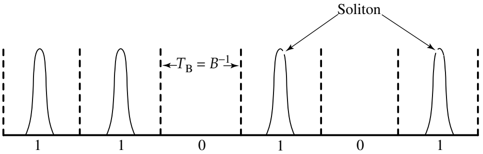

Figure 1. Soliton bit stream in the return-to-zero format where each soliton occupies a fraction of the bit slot representing 1 in a bit stream.

Temporal solitons are attractive for optical communications because they are able to maintain their width even in the presence of fiber dispersion, but their use requires substantial changes in the fiber system design, compared with conventional (nonsoliton) systems.

The basic idea is to use a soliton in each bit slot representing 1 in a bit stream. Figure 1 shows schematically a soliton bit stream in the return-to-zero (RZ) format. Typically, the spacing between two solitons exceeds a few times their full-width at half-maximum (FWHM), and individual solitons are well isolated. This requirement relates the soliton width T_0 to the bit rate B as $B = 1/T_B = 1/(2q_0T_0)$, where T_B is the duration of the bit slot and $2q_0 = T_B/T_0$ is the separation between neighboring solitons in normalized units.

The relatively large spacing necessary to avoid soliton interaction, limits the bit rate of soliton communication systems. The spacing can be reduced by up to a factor of two using unequal amplitudes for the neighboring solitons, so this scheme is feasible in practice and can be useful for increasing the system capacity.

Temporal solitons use the nonlinear phenomenon of SPM to maintain their width even in the presence of fiber dispersion. However, this property holds only if fiber losses are negligible. Optical amplifiers can be used for compensating fiber losses. Two approaches used for the management of losses through amplification of solitons are the lumped- and distributed-amplification techniques.

In the lumped-amplification scheme, optical amplifiers are placed periodically along the fiber link such that fiber losses between two amplifiers are exactly compensated by the amplifier gain. An important design parameter is the spacing L_A between amplifiers, which should be as large as possible to minimize the overall cost. For nonsoliton systems, L_A is typically 80–100 km. For soliton systems, L_A is restricted to smaller values because of the soliton nature of signal propagation.

The physical reason behind smaller values of L_A is that optical amplifiers boost soliton energy to the input level over a length of a few meters without allowing for gradual recovery of the fundamental soliton. The amplified soliton adjusts its width dynamically in the fiber section following the amplifier, but it also sheds a part of its energy as dispersive waves during this adjustment phase. The dispersive part can accumulate to significant levels over a large number of amplification stages and must be avoided. One way to reduce the dispersive part is to reduce the amplifier spacing L_A such that the soliton is not perturbed much over this short length. Numerical simulations show (Hasegawa & Kodama, 1995) that this is the case when L_A is a small fraction of the dispersion length ($L_A \ll L_D$). The dispersion length L_D depends on both the pulse width T_0 and the GVD parameter β_2 and can vary from 10 to 1000 km, depending on their values.

The condition $L_A \ll L_D$, imposed on loss-managed solitons when lumped amplifiers are used, becomes increasingly difficult to satisfy in practice as bit rates exceed 10 GB/s. This condition can be relaxed considerably when distributed amplification is used. The distributed-amplification scheme is inherently superior to lumped amplification because its use provides a nearly lossless fiber by compensating losses locally at every point along the fiber link. In fact, this scheme was used as early as 1985 using the distributed gain provided by Raman amplification (Agrawal, 2002) when the fiber carrying the signal was pumped at a wavelength of about 1.46 μm using a color-center laser (Hasegawa & Kodama, 1995). Alternatively, the transmission fiber can be doped lightly with erbium ions and pumped periodically to provide distributed gain. Several experiments have demonstrated that solitons can be propagated in such active fibers over relatively long distances.

Early soliton experiments on loss compensation used the Raman-amplification scheme. The situation changed with the application of erbium-doped fiber amplifiers from around 1989 for loss-managed soliton systems. In a typical experiment (Agrawal, 2002), 2.5 GB/s solitons were transmitted over 12,000 km by using a 75 km fiber loop containing three amplifiers, spaced 25 km apart (Agrawal, 2002). The bit rate–distance product of $BL = 30$ (TB/s)km is limited mainly by the timing jitter induced by amplifiers.

Dispersion management, which consists of a periodic change of the dispersion β_2 along the fiber length, is employed commonly for modern wavelength-division multiplexed systems. The use of dispersion management forces each soliton to propagate in the normal-dispersion regime of a fiber during each map period. If the map period is a fraction of the nonlinear length, the nonlinear effects are relatively small, and the pulse evolves in a linear fashion over one map period. On a longer length scale, solitons can still form if the nonlinear effects are balanced by the average dispersion. As a result, solitons can survive in an average sense, even though not only the peak power but also the width and shape of such solitons oscillate periodically.

Since 1996, a large number of experiments have shown the benefits of using dispersion-managed solitons for optical communication systems (Agrawal, 2002). Presently, 10 GB/s dispersion-managed solitons can be transmitted over 16 Mm of standard fiber when soliton interactions are minimized by choosing the location of amplifiers appropriately. An important application of dispersion management consists of upgrading the existing terrestrial networks designed with standard fibers and operating in the linear regime.

Yuri Kivshar

See also **Dispersion management; Kerr effect; Nonlinear optics; Nonlinear Schrödinger equations**

Further Reading

Agrawal, G.P. 2002. *Fiber-Optic Communication Systems*, New York: Wiley

Hasegawa, A. & Kodama, Y. 1995. *Solitons in Optical Communications*, Oxford: Clarendon Press

Hasegawa, A. & Tappert, F. 1973. Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers. *Applied Physics Letters*, 23: 142–144

Kivshar, Yu.S. & Agrawal, G.P. 2003. *Optical Solitons: From Fibers to Photonic Crystals*, San Diego: Academic Press

Mollenauer, L.F., Stolen, R.H. & Gordon, J.P. 1980. Experimental observation of picosecond pulse narrowing and solitons in optical fibers. *Physical Review Letters*, 45: 1095–1098

OPTICAL NONLINEARITIES

See **Nonlinear optics**

ORBIT

See **Phase space**

ORDER FROM CHAOS

At first glance, the words *chaos* and *order* seem to be contradictory. Indeed they are contradictory in terms of common definitions of the words, but not in terms of their mathematical usages. The definition of *chaos* in the Oxford American Dictionary (1986) is "great disorder or confusion." In contrast, the mathematical notion of *chaos* explains that data that seems to be random may have been generated by a deterministic and ordered process. In this sense, the field might have profitably been called "order theory" instead of "chaos theory." There are two common usages of the phrase "order from chaos." The first refers to seeking deterministic "chaotic" models to explain complex phenomenon, thus "order *in* chaos." The second refers to the fact that some chaotic dynamical systems can also exhibit regular islands of simplicity within their phase space, or "order *from* chaos." We will discuss each in turn.

Order in Chaos

The fact that seemingly simple and deterministic evolution rules can give rise to extremely complicated motion dates to Henri Poincaré (1892), in the setting of celestial mechanics. Perhaps the most famous example of unpredictable behavior is due to Edward Lorenz (1963), who discovered that even the most simplified models of the weather could produce data for which it is impossible to make long-term forecasts. Lorenz identified the "butterfly effect," or "sensitive dependence to initial conditions," whereby even small measurement errors quickly grow to swamp the signal. In 1975, Tien Yien Li and James A. Yorke published a paper entitled, "Period Three Implies Chaos," (Li & Yorke, 1975). The most famous impact of this paper was that it coined the word *chaos*, the field of study describing the nonlinear effect in which even the simple systems of Poincaré, Lorenz, and others can display sensitive dependence in a bounded domain. The chaos theory has attracted a great deal of popular attention, partly, in fact, due to its sexy title. Part of the appeal is also due to the scientific tradition since the time of Newton of searching for mechanistic explanations of reality.

The concept of phase space is useful for mapping the behavior of a dynamical system. Phase space could be described as the set of all possible relevant variables, which creates a closed description of the time evolution of the system (*See* **Phase space**). For example, for a simple pendulum, we need to specify all possible angular position and angular momentum states to uniquely define all solutions; the phase space for the periodic oscillation is a closed curve, a circle. For an electronic tank oscillator (LRC) circuit, we need to specify capacitor charge and current through the inductor. The dimension of the system is the dimension of the phase space. A billiard table's worth of balls, for example, requires many variables for a complete description—two variables for position and two for momentum for each of the balls.

As a highly interdisciplinary field, chaos theory has been successful in finding deterministic chaos, or order, in what was once thought to be mere noise, or at

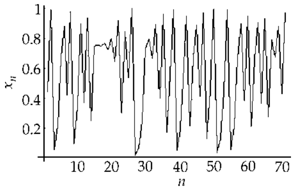

Figure 1. This seemingly stochastic data was actually generated by a "simple," low-dimensional, and deterministic process. This time-series plot of the state x_n versus time n was actually generated by the dynamical system $x_{n+1} = ax_n(1-x_n)$ called the logistic map. ($a = 4.0$, $x_0 = 0.44$).

least an extremely high-dimensional effect. There are numerous explicit examples of chaos in many areas, including biology, electrical engineering, chemistry, celestial mechanics, and brain and heart physiology. Thus, the old idea that extremely complicated data must always be due to extremely complicated effects is false (see Figure 1). One popular misuse of these ideas is a belief that all complicated data must have underlying order in chaos. This can be phrased as "Is what seems complicated always really simple?" The answer, of course, is no; it remains true that noise and other high-dimensional effects can also be responsible for complexity. What is true is that what seems complicated may *sometimes* be simple, in the sense of having a low-dimensional chaotic model. To cite one popular question, the stock market prices unarguably constitute extremely complicated data, but is there underlying order in chaos here, or is the explanation due to intractably high-dimensionality?

Order from Chaos

Some chaotic systems have an occasional tendency to exhibit simple behavior. This is often described as regular islands in an otherwise chaotic phase space. The classic mathematical example of regular islands arises in Hamiltonian systems, which can exhibit a phase space of solutions in which chaotic solutions are both intermingled and bounded by nested islands around islands around islands, and so on, of KAM-like tori (circle-like integrable solutions) (Arrowsmith & Place, 1990; Meiss, 1992).

Put differently, "order from chaos" and "regular islands" are terms commonly used to refer to the presence of an "attractor" (an imaginary point in the phase space about which the trajectories appear to orbit) in a system that one thinks should behave in a complicated manner. For example, consider a thought experiment of a game of pool in which we assume no friction so that the balls never stop moving (a two-dimensional Lorentz gas). Following the path of one specific ball, say the seven-ball, is an extremely complicated problem displaying sensitive dependence, since a small change in the velocity or position of the ball affects the next collision with the wall or with the next ball, an effect that multiplies upon subsequent collisions. However, there are several attractors near which the ball's motion becomes quite simple—the pocket holes. Once within the rim of a pocket (its basin of attraction), falling in becomes inevitable, as it would require a relatively large energy perturbation to prevent it from falling in. In this sense, we can say that the regions of phase space, corresponding to being stuck in the billiards pocket, are regular islands in a chaotic sea.

As analogies, such descriptions of islands in chaotic seas have been extended by some to explain the emergence of a coherent phenomenon from complicated processes. There is perhaps no more intriguing question than the origins of life. Proponents of emergence theories suggest that life in the original Hadean seas (the "primordial" seas on early Earth) gave rise to life through a capture process such as the billiards game (Waldrop, 1992; Kauffman, 1995). While randomly chosen initial conditions in the billiards game may each individually be unlikely to lead to capturing a ball in a pocket (and certainly any analogous attractors in chemical processes must be exceedingly unlikely), the way to win an unlikely bet is to play very quickly, over and over. This may have been the process that led to complex organic molecules, according to proponents of the emergence theory.

Emergence and order from chaos have also been used to describe processes whose attractors are surprisingly complicated themselves, as sets, but arise from surprisingly simple rules. Michael Barnsley has called the following "The Chaos Game" (Barnsley, 1993). First label the vertices of a triangle, 1, 2, 3. Using a random number generator (such as a six-sided die), assign probabilities to each triangle vertex, say die sides 1–2 to vertex 1, die sides 3–4 to vertex 2, and die sides 5–6 to vertex 3. Roll the die to randomly select one of the vertices, and record its planar coordinates (x, y). Then roll the die again to randomly select another vertex, and record the point halfway between the resulting new vertex and the current (x, y) as the new (x, y). Repeat indefinitely. For each newly defined (x, y), we record a dot for a pictorial record. Most people guess that the result will uniformly fill the triangle with a smattering of dots, but the mathematical fact is surprising. The result is an extremely intricate structure, a fractal called the Sierpinski gasket (see Figure 2). Because this simple rule gives rise to such a complicated structure, the argument goes, perhaps many of the other intricacies we see around us might have emerged from other simple rules.

Finally, we mention the meaning of "order from chaos," as developed from theories of the Nobel prize-winning physicist Ilya Prigogine, whose work on dissipative structures of systems held from thermal

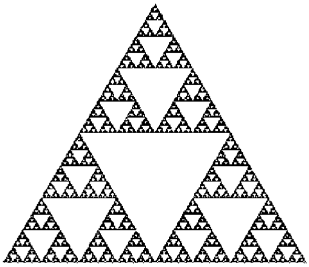

Figure 2. From such a simple algorithm as "The Chaos Game" (Barnsley, 1993) emerges the extremely intricate fractal shown, called the Sierpinski gasket.

equilibrium has been used to study self-organizing systems (Prigogine, 1984). Prigogine has defined complexity as the ability to switch between different modes of behavior as the environmental conditions vary. Thus, he has described a phenomenon in which far-from-equilibrium systems transition "from being to becoming," which some describe as order from chaos.

Erik M. Bollt

See also **Chaotic dynamics; Emergence; Kalmogorov–Arnol'd–Moser theorem; Phase space**

Further Reading

Arrowsmith, D.K. & Place, C.M. 1990. *An Introduction to Dynamical Systems*, Cambridge and New York: Cambridge University Press

Barnsley, M. 1993. *Fractals Everywhere*, 2nd edition, San Francisco: Morgan Kaufmann

Kauffman, S. 1995. *At Home in the Universe, the Search for the Laws of Self-Organization and Complexity*, Oxford and New York: Oxford University Press

Lorenz, E.N. 1963. Deterministic nonperiodic flow. *Journal of Atmospheric Science*, 20: 130–141

Li, T.Y. & Yorke, J.A. 1975. Period three implies chaos. *American Mathematics Monthly*, 82: 985–992

Meiss, J.D. 1992. Symplectic maps, variational principles, and transport. *Reviews of Modern Physics*, 64: 795

Poincaré, H. 1892–99. *Les méthodes nouvelles de la mécanique céleste*, Paris: Gauthier-Villars, 3 vols; as *New Methods of Celestial Mechanics*, New York: American Institute of Physics, 1993

Prigogine, S. 1984. *Order out of Chaos: Man's New Dialogue with Nature,* New York: Bantam Books

Waldrop, M. 1992. *Complexity: The Emerging Science at the Edge of Chaos and Order*, New York: Touchstone Press

ORDER PARAMETERS

Although the concept of an order parameter was introduced by Lev Landau in the 1930s, it does not yet have a precise definition. Broadly, an order parameter is a thermodynamic quantity that is invariant with respect to the symmetry group of the low-temperature phase and zero above a critical point or transition temperature, T_c. It is a measure of the amount of order that is built up in the system in the neighborhood of the critical point. In general, an order parameter has both an amplitude and a phase. To find the equation of state, a minimization procedure is followed for an appropriate thermodynamic potential.

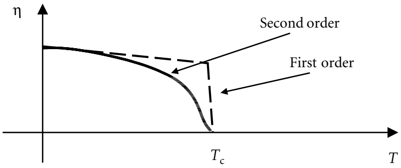

Figure 1. A typical plot of the scalar order parameter as a function of temperature.

From the original application to second-order phase transitions where it changes continuously from T_c to lower temperatures, the idea of an order parameter has been extended to first-order transitions involving an abrupt change at T_c (see Figure 1). The concept of an order parameter has been generalized from a globally defined scalar to a complex time- and space-dependent function. Note that first-order phase transitions are associated with discontinuities of the order parameter and thermal hysteresis effects. Second-order phase transitions have a continuous order parameter and show field-induced hysteresis. Diverse applications of the order parameter concept to both equilibrium and non-equilibrium critical phenomena are listed in Table 1.

Crystal formation is demonstrated by the existence of a regular diffraction pattern associated with the Fourier components of the mass density distribution $\rho(\mathbf{r})$

$$\rho(\boldsymbol{r}) = \sum_{\boldsymbol{g}} \rho(\boldsymbol{g}) \mathrm{e}^{\mathrm{i}\boldsymbol{g}\cdot\boldsymbol{r}}, \qquad (1)$$

where the $\boldsymbol{g}$ are vectors in the reciprocal space. The set of numbers $\rho(\boldsymbol{g})$ can be used as order parameters characterizing the low-temperature (crystal) phase. The nonzero coefficients $\rho(\boldsymbol{g})$ in Equation (1) define a multi-component order parameter.

The order parameter field for a magnet is defined at each position $\boldsymbol{x}$ by a direction of the local magnetization $\boldsymbol{M}(\boldsymbol{x})$ whose length is fixed. By becoming a magnet, this material has broken the rotational symmetry and its order parameter field defines the broken symmetry directions chosen in the material.

A number of metals, alloys, and ceramics below their critical temperature T_c exhibit an ordered state in the conduction electron degrees of freedom manifested by zero resistance. The order parameter for superconductors is the wavefunction of the Cooper pair

Phenomenon	Disordered phase	Ordered phase	Order parameter
Equilibrium			
Condensation	Gas	Liquid	Density difference $\rho_L - \rho_G$
Spontaneous magnetization	Paramagnet	Ferromagnet	Net magnetization M
Antiferromagnetism	Paramagnet	Antiferromagnet	Staggered magnetization $M_1 - M_2$
Superconductivity	Conductor	Superconductor	Cooper pair, wave function η
Alloy ordering	Disordered mixture	Sublattice ordered alloy	Sublattice concentration
Ferroelectricity	Paraelectric	Ferroelectric	Polarization
Superfluidity	Fluid	Superfluid	Condensate wavefunction
Nonequilibrium			
Laser action	Lamp (incoherent)	Laser (coherent)	Electric field intensity
Super-radiant source	Noncoherent polarization	Coherent polarization	Atomic polarization
Fluid convection	Turbulent flow	Bénard cells	Amplitude of mode

Table 1. Examples of order parameters (OPs).

condensate $\eta(\boldsymbol{r})$, and it exhibits a Hopf bifurcation at $T = T_c$.

The superfluid properties in ^{4}He and ^{3}He are manifested by the absence of viscosity. The ^{4}He atoms are bosons that below a transition temperature T_λ, undergo the so-called Bose condensation into a $k = 0$ mode. The associated order parameter is the condensate's quantum wavefunction. Because ^{3}He atoms are fermions, below T_λ they form Cooper pairs.

Liquid crystals are anisotropic fluids composed of strongly elongated molecules. The nematic phase is characterized by the existence of a direction to which most of the molecules are parallel, so the order parameter is a second rank tensor describing correlations along that direction. Numerous other examples of critical phenomena, such as binary fluids, the metal-insulator transition, polymer transitions and spin- and charge-density waves, have their own order parameters.

Landau deduced that second-order phase transitions are associated with symmetry breaking and can be qualitatively described by an order parameter η. Assuming that the free energy F depends on V, T, and η, the equilibrium conditions are:

$$\left.\frac{\partial F(T, V, \eta)}{\partial \eta}\right|_{\eta=\eta_0} = 0$$

and

$$\left.\frac{\partial^2 F(T, V, \eta)}{\partial \eta^2}\right|_{\eta=\eta_0} > 0, \qquad (2)$$

where η_0 is the equilibrium value of η.

The universality hypothesis states that any two physical systems with the same spatial dimensionality, d, and the same number of order parameter components, n, belong to the same universality class having identical critical exponents (see Table 2).

Universality class		System	Order parameter (OP)
$d = 2,$	$n = 1$	Absorbed films	Surface density
$d = 2,$	$n = 2$	Superfluid ^{4}He film	Superfluid wave function
$d = 3,$	$n = 1$	Uniaxial ferromagnets	Magnetization
$d = 3,$	$n = 1$	Fluids	Density difference
$d = 3,$	$n = 1$	Mixtures, alloys	Concentration difference
$d = 3,$	$n = 2$	Planar ferromagnets	Magnetization
$d = 3,$	$n = 2$	Superfluids	Superfluid wave function
$d = 3,$	$n = 3$	Isotropic ferromagnets	Magnetization

Table 2. Examples of universality classes.

Order parameters accompany broken symmetry phenomena where the new ground state of the system does not possess the full symmetry of the Hamiltonian. A classic example is the ferromagnetic-to-paramagnetic phase transition at T_c where the full rotational symmetry of the paramagnetic phase is broken by the axiality of the ground ferromagnetic state below T_c. When a symmetry that is broken is continuous, a vibrational mode appears whose frequency vanishes at long wavelengths. (Quanta of such modes are called "Goldstone bosons.") Examples include ferromagnetic domain walls and acoustic soft modes in structural phase transitions.

There exist several different types of broken symmetries: (a) translational (crystal formation, structural transitions); (b) gauge (superfluidity, superconductivity); (c) time reversal (ferromagnets); (d) local rotational (liquid crystals); (e) rotational (some structural phase transitions); and (f) space inversion (ferroelectricity). Gauge symmetry is a universal property of Hamiltonians whenever the total number of particles is conserved or a generalized charge-like conserved quantity exists. Then, the order parameter η is complex, and its local density $\rho = \eta^* \eta(r)$ is such that a phase shift $\eta \rightarrow \eta e^{i\omega}$ leaves the Hamiltonian invariant.

Defects in the order parameter space can be topological (i.e., kinks, also referred to as domain walls) and nontopological (i.e., solitons, also called nucleation centers). They are obtained as solutions to the equations of motion for the order parameter field. Also, point defects, line defects, vortices, dislocations, vacancies, and interstitials with attendant singularities are seen experimentally in critical systems (systems close to a phase transition).

Finally, note that Haken's separation of modes in synergetic systems into masters (order parameters) and slaves has been influenced by Landau's theory of phase transitions.

Jack A. Tuszyński

See also **Bose–Einstein condensation; Critical phenomena; Domain walls; Ferromagnetism and ferroelectricity; Hysteresis; Liquid crystals; Phase transitions; Solitons; Synergetics**

Further Reading

Anderson, P.W. 1984. *Basic Notions of Condensed Matter Physics*, Menlo Park, CA: Benjamin/Cummings

Haken, H. 1980. *Synergetics*, Berlin and New York: Springer

Landau, L.D. & Lifshitz, E.M. 1959. *Statistical Physics*, London: Pergamon

Ma, S.-K. 1976. *Modern Theory of Critical Phenomena*, Reading, MA: Benjamin

Reichl, L.E. 1979. *A Modern Course in Statistical Physics*, Austin, TX: University of Texas Press

White, R.H. & Geballe, T. 1979. *Long Range Order in Solids*, New York: Academic Press

Yeomans, J.M. 1992. *Statistical Mechanics of Phase Transitions*, Oxford: Clarendon Press and New York: Oxford University Press

ORDINARY DIFFERENTIAL EQUATIONS, NONLINEAR

An ordinary differential equation (ODE) is an equation

$$\Delta(t, f, f', f'', ..., f^{(N)}) = 0 \qquad (1)$$

(with $f^{(j)} = \mathrm{d}^j f / \mathrm{d}t^j$), relating a function $f(t)$ to its derivatives. The order of an ODE is the size of the highest derivative that appears, so Equation (1) is Nth order. An ODE is *linear* if it can be written as a linear combination of f and its derivatives, that is,

$$\begin{aligned}\Delta \equiv\ & a_N(t) f^{(N)} + a_{N-1}(t) f^{(N-1)} + \cdots \\ & + a_1(t) f' + a_0(t) f + b(t) = 0,\end{aligned} \qquad (2)$$

including the possible addition of an inhomogeneous term $b(t)$. All other ODEs, which are not of the form (2), are referred to as *nonlinear*.

More generally, one can consider systems of ODEs relating M functions $f_0, f_1, \ldots, f_{M-1}$ and their derivatives $f_k^{(j)}$ of different orders. In fact, if a system of ODEs can be solved for the highest derivatives appearing (which generally requires the implicit function theorem), then it can always be converted to a system of first-order equations,

$$\begin{aligned} f_0' &= F_0(t, f_0, f_1, \ldots, f_{M-1}), \\ f_1' &= F_1(t, f_0, f_1, \ldots, f_{M-1}), \\ &\vdots \\ f_{M-1}' &= F_{M-1}(t, f_0, f_1, \ldots, f_{M-1}). \end{aligned} \qquad (3)$$

For example, suppose Equation (1) can be solved explicitly for the Nth derivative, as $f^{(N)} = F(t, f, f', f'', \ldots, f^{(N-1)})$. In that case, the single ODE (1) may be rewritten as the first-order system

$$\begin{aligned} f_0' &= f_1, \\ f_1' &= f_2, \\ &\vdots \\ f_{N-2}' &= f_{N-1}, \\ f_{N-1}' &= F(t, f_0, f_1, f_2, \ldots, f_{N-1}). \end{aligned}$$

In most applications of nonlinear ODEs, such as in the physical sciences or biology, they appear as coupled systems of either first or second order. For example, Newton's equations in mechanics (Arnol'd, 1989) are of the form

$$\boldsymbol{q}'' = \boldsymbol{F}(t, \boldsymbol{q}, \boldsymbol{q}'),$$

relating the acceleration $\boldsymbol{q}''$ of a particle (of unit mass) to the force $\boldsymbol{F}$ acting on it, which is a function of its position $\boldsymbol{q}$ and velocity $\boldsymbol{q}'$, and possibly the time t. Mechanical systems are often derived from an action functional $S = \int_{t_0}^{t_1} L(t, \boldsymbol{q}, \boldsymbol{q}', \ldots, \boldsymbol{q}^{(j)})\,\mathrm{d}t$. In particular, if $\boldsymbol{q}'$ is the highest derivative appearing in the Lagrangian, so that $L = L(t, \boldsymbol{q}, \boldsymbol{q}')$, then the corresponding Euler–Lagrange equations form a second-order system of ODEs, viz.

$$\frac{\partial L}{\partial \boldsymbol{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \boldsymbol{q}'}\right) = 0.$$

In Hamiltonian mechanics, on the other hand, the equations of motion are given as a first-order system of ODEs,

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{J}(\boldsymbol{x}) \nabla H(t, \boldsymbol{x}), \qquad (4)$$

describing the evolution of a point $\boldsymbol{x}$ in the phase space with Poisson tensor $\boldsymbol{J}$ and Hamiltonian H. First-order equations are commonly used both in population dynamics, for example, the Verhulst model

$$\frac{\mathrm{d}P}{\mathrm{d}t} = rP(1 - P/K) \tag{5}$$

(with r, K constants), and in reaction kinetics applying the Law of Mass Action (Murray, 1989). The ODEs describing a system of N species or N chemical reagents typically take the form of N coupled first-order ODEs, but usually these are *not* Hamiltonian.

Given an ODE such as (1), the main task is to determine the nature of the solutions, namely, those functions $f(t)$ for which all the derivatives $f^{(j)}$ exist for $j = 1, \ldots, N$ and are related by Equation (1). More specifically, one may wish to solve an initial value problem, where the values $f(t_0), f'(t_0), \ldots, f^{(N-1)}(t_0)$ of the function f and its first $N-1$ derivatives are specified at some initial time $t = t_0$, and f is to be determined at subsequent times $t > t_0$. Alternatively, one might pose a boundary value problem where the values $f(t_0)$ and $f(t_1)$ (and maybe some of the derivatives) are given at the endpoints of the interval $[t_0, t_1]$, and $f(t)$ is to be found for $t_0 < t < t_1$.

For an Nth-order linear homogeneous ODE, of the form (2) with $b \equiv 0$, the general solution is just a linear combination of N-independent solutions s_j, that is, $f(t) = \sum_{j=1}^{N} A_j s_j(t)$ with N arbitrary constants $A_1, \ldots, A_N$. Ideally, one would like to express the general solution of an Nth-order nonlinear ODE as a function of N arbitrary integration constants, but in general this is not possible. However, for systems (3) where the F_j on the right-hand sides are suitably regular functions of all their arguments in the neighborhood of the initial data, the local existence of a solution to the initial value problem near $t = t_0$ can be proved by the Cauchy–Lipschitz method (Ince, 1926). In fact, whenever the F_j are analytic functions, then the existence of a local solution to the initial value problem is guaranteed in some circle around $t = t_0$ in the complex t plane (Hille, 1976). This means that, at least locally, the solution $f(t)$ can be considered as a function of the initial data $f(t_0), f'(t_0), \ldots, f^{(N-1)}(t_0)$. Only for ODEs of Painlevé type can this local solution be extended globally to a single-valued, meromorphic function of t (Hinkkanen & Laine, 1999). In contrast, the solutions of other ODEs can display chaotic behavior, with coalescing branch points in the complex plane (Sachdev, 1991).

For first-order equations ($N = 1$), there are various classes of ODEs that are amenable to exact integration methods. The simplest example is the class of *separable* equations, which are directly solvable by a quadrature:

$$\frac{\mathrm{d}f}{\mathrm{d}t} = T(t)/F(f),$$

whence

$$\int F(f)\,\mathrm{d}f = \int T(t)\,\mathrm{d}t + \text{constant}.$$

The Verhulst model (5) is a particularly simple example for we have

$$\int \frac{\mathrm{d}P}{P(1 - P/K)} = \log\left(\frac{P}{K - P}\right) = rt + c$$
$$\Rightarrow P(t) = \frac{K}{1 + \mathrm{e}^{-rt-c}} = \frac{KP_0}{P_0 + (K - P_0)\mathrm{e}^{-rt}},$$

where the constant $P_0 = K/(1 + e^{-c}) = P(0)$ is the initial population size. Other special classes of first-order equations include the Bernoulli equations $f' = P(t)f + Q(t)f^n$ and Riccati equations $f' = A(t)f^2 + B(t)f + C(t)$; both of these types can be converted to linear equations by a substitution (Ince, 1926).

Among higher-order ODEs, exactly solvable equations are rare, but nevertheless some exact solution methods exist. In particular, if there are first integrals (constants of motion) or symmetries for a system, then it is possible to reduce the order. For autonomous Hamiltonian systems of order $N = 2n$, Liouville's theorem on integrable systems states that if there are n independent first integrals, in involution with respect to the nondegenerate Poisson bracket defined by $\boldsymbol{J}$, then the ODEs (4) can be integrated by quadratures (Arnol'd, 1989). More generally, an ODE of order N with a first integral can be reduced to an equation of order $N-1$, while if it has a Hamiltonian or Lagrangian structure invariant under a symmetry of the system, then (using Noether's theorem) the order can be reduced to $N-2$. The approach to solving ODEs using their symmetries is originally due to Lie; for a modern treatment, see Olver (1993).

As an illustration of some of these ideas, consider the second-order ODE

$$f'' = \left(kt - \frac{(n-1)}{f}\right)(f')^2 + \frac{(n-1)}{t}f' \tag{6}$$

(k, n constants), which is a radial symmetry reduction of an n-dimensional partial differential equation appearing in differential geometry (Abreu, 1998). This can be derived from the action

$$S = \int_{t_0}^{t_1} L(t, f, f')\,\mathrm{d}t,$$
$$L = t^{n-1}\log(f^{n-1}f') - kt^n f. \tag{7}$$

By a Legendre transformation, setting $q = f$ and $p = \partial L/\partial f'$, (6) can also be converted to the first-order Hamiltonian form

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} \partial_q H \\ \partial_p H \end{pmatrix}$$

with $H(t, q, p) = t^{n-1}\log(q^{n-1}/p) - kt^n q$. Because this is a non-autonomous system, with t appearing explicitly, the Hamiltonian H is *not* a constant of motion. However, (6) is invariant under the one-parameter group of scaling symmetries $r \to \mu r$, $f \to \mu^{-1} f$, so by introducing the new scale-invariant independent variable $y = ft$, and the dependent variables $v = v(y) = -\log t$, $w = \mathrm{d}v/\mathrm{d}y$, it reduces to a first-order equation for $w(y)$:

$$\frac{\mathrm{d}w}{\mathrm{d}y} = y(2n - ky)w^3 + (3n - 2ky)w^2 \\ +((n-1)/y - k)w.$$

Unfortunately, it is not possible to reduce this to a quadrature, since the action (7) is *not* invariant under scaling, unless $k = 0$ when the general solution to (6) is $f(t) = (At^n + B)^{1/n}$ (with A, B arbitrary constants).

Other important methods for ODEs include the Painlevé analysis of movable singularities (Kruskal & Clarkson, 1992), and asymptotic expansions around regular or irregular fixed singular points (Wasow, 1965; Tovbis, 1994).

ANDREW HONE

See also **Chaotic dynamics; Constants of motion and conservation laws; Euler–Lagrange equations; Extremum principles; Hamiltonian systems; Integrability; Painlevé analysis; Partial differential equations, nonlinear; Riccati equations**

Further Reading

Abreu, M. 1998. Kähler geometry of toric varieties and extremal metrics. *International Journal of Mathematics*, 9: 641–651

Arnol'd, V.I. 1989. *Mathematical Methods of Classical Mechanics*, 2nd edition, New York: Springer

Hille, E. 1976. *Ordinary Differential Equations in the Complex Domain*, New York: Wiley

Hinkkanen, A. & Laine, I. 1999. Solutions of the first and second Painlevé equations are meromorphic. *Journal d'Analyse Mathématique*, 79: 345–377

Ince, E.L. 1926. *Ordinary Differential Equations*, London: Longmans Green; reprinted New York: Dover, 1956

Kruskal, M.D. & Clarkson, P.A. 1992. The Painlevé–Kowalevski and Poly–Painlevé tests for integrability. *Studies in Applied Mathematics*, 86: 87–165

Murray, J.D. 1989. *Mathematical Biology*, Berlin and New York: Springer

Olver, P.J. 1993. *Applications of Lie Groups to Differential Equations*, 2nd edition, Berlin and New York: Springer

Sachdev, P.L. 1991. *Nonlinear Ordinary Differential Equations and Their Applications*, New York: Marcel Dekker

Tovbis, A. 1994. Nonlinear ordinary differential equations resolvable with respect to an irregular singular point. *Journal of Differential Equations*, 109: 201–221

Wasow, W. 1965. *Asymptotic Expansions for Ordinary Differential Equations*, New York: Wiley

ORGANIZING CENTERS

See **Spiral waves**

OSCILLATOR, CLASSICAL NONLINEAR

See **Damped-driven anharmonic oscillator**

OVERTONES

When a tonal sound, such as a note played on a flute, a human vowel sound, or a bell, is analyzed in the frequency domain by applying a Fourier transform to the acoustic pressure waveform, the spectrum consists of a large number of sharp lines. The component of lowest frequency is termed the "fundamental", and the others are "upper partials" or "overtones." If the frequencies of the overtones are all exact integer multiples of the frequency of the fundamental, then they are termed "harmonics." The partial with frequency $f_n = nf_1$, where f_1 is the frequency of the fundamental, is the nth harmonic, so that the fundamental is the first harmonic.

A one-dimensional simple harmonic oscillator, or linear oscillator, in which the restoring force is proportional to displacement y from the equilibrium position, obeys the equation

$$m\frac{\mathrm{d}^2y}{\mathrm{d}t^2} = -ky, \tag{1}$$

where m is the mass of the moving particle and k is the restoring force constant. A damping term can also be included, but this need not concern us here. Such an oscillator vibrates with a single frequency $f_1 = (1/2\pi)(k/m)^{1/2}$ that is independent of oscillation amplitude. It is useful to think of this oscillator in terms of its potential energy function, which is quadratic as shown in Figure 1(a). In real oscillators, the restoring force is not linear for large displacements, but nonlinear so that

$$m\frac{\mathrm{d}^2y}{\mathrm{d}t^2} = -ky(1 + \alpha_1 y + \alpha_2 y^2 + \cdots), \tag{2}$$

where the α_n are constants. The energy curve then has a distorted parabolic form such as that shown as an example in Figure 1(b). In the absence of damping, the total energy must remain constant so that the magnitude of the velocity is a simple function of the displacement, and the motion repeats cyclically. This means that the spectrum of such a nonlinear oscillator consists of exact phase-locked harmonics of the fundamental frequency, though this fundamental frequency depends upon the amplitude of the motion. For a reason that is derived from molecular physics, as discussed below, such a nonlinear oscillator is often confusingly called an "anharmonic oscillator."

A thin taut string of length L ideally obeys an equation of the form

$$m\frac{\partial^2y}{\partial t^2} = T\frac{\partial^2y}{\partial x^2}, \tag{3}$$

where x measures length along the string, m is the mass per unit length, and T is the string tension.

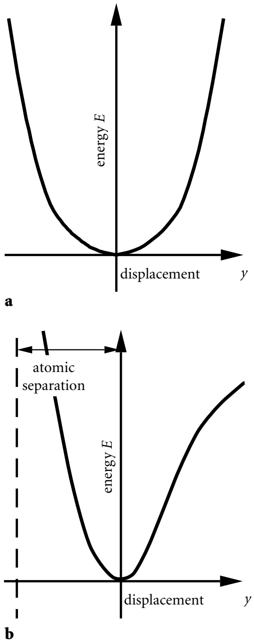

Figure 1. (a) Potential energy curve for a simple harmonic oscillator. (b) Potential energy curve for a typical nonlinear oscillator such as a diatomic molecule.

If its ends are rigidly fixed, then the mode frequencies are exact harmonics of the fundamental so that $f_n = (n/2L)(T/m)^{1/2}$. It is thus a multimode harmonic oscillator. The nonlinear frictional action of the bow on a violin reinforces the harmonicity of the modes and locks them into rigid phase relationship (Fletcher, 1999). Something very similar happens with wind instruments, which also have precisely harmonic spectra.

A thin stiff bar, on the other hand, obeys an equation of the form

$$m\frac{\partial^2 y}{\partial t^2} = K\frac{\partial^4 y}{\partial x^4}\,, \tag{4}$$

where K is the elastic stiffness. If the ends are free or rigidly clamped, then the mode frequencies are approximately $f_n \approx \frac{4}{9}(n+\frac{1}{2})^2 f_1$, and the overtones are very far from being harmonically related. Such an oscillator might be termed "inharmonic." The modes of a three-dimensional object such as a bell are even more complex (Fletcher & Rossing, 1998).

While sustained-tone musical instruments depend upon the nonlinearity of the active generator for their operation (bow, reed, or lip air-flow), the linear resonator (string or air column) determines the oscillation frequency, so that the pitch is nearly independent of loudness, and only the relative amplitudes of the harmonics change (Fletcher, 1999). Some Chinese opera gongs, however, make a virtue of nonlinearity so that, after an impulsive excitation, the pitch either rises or falls dramatically as the vibration dies away (Fletcher, 1985). The frequencies and relative intensities of upper partials determine the tone quality of a musical sound and have dictated the development of musical scales and harmonies (Sethares, 1998).

The human auditory system itself has some nonlinear aspects (Zwicker & Fastl, 1999), and, as in any forced nonlinear oscillator, these lead to the generation of harmonics ("harmonic distortion") and of multiple sum and difference tones ("intermodulation distortion"). In the ear these are chiefly apparent in the generation of the difference tone $|f_1 - f_2|$ when loud tones of frequencies f_1 and f_2 are heard simultaneously.

Optical absorption and emission spectra have many similarities to acoustic phenomena (Herzberg, 1950; Harmony, 1989). Diatomic molecules, for example, have interatomic potentials of the form shown in Figure 1(b) and thus constitute nonlinear oscillators. If the interatomic potentials were simply parabolic, as in Figure 1(a), then the quantum energy levels would have the form $E_n = (n+\frac{1}{2})h\nu$, where h is Planck's constant and ν is the classical vibration frequency labeled f_1 above. The wave functions describing the atomic vibration would then be either symmetric or antisymmetric, and the selection rule would dictate that n could change only by ± 1. There would thus be only a single absorption band consisting of the vibrational transition $0 \rightarrow 1$ surrounded by the allowed rotational transition lines.

For a more realistic model of the interatomic potential, as in Figure 1(b), the energy levels can be written as

$$\begin{aligned} E_n &= (n+\tfrac{1}{2})h\nu[1+\beta_1(n+\tfrac{1}{2}) \\ &\quad +\beta_2(n+\tfrac{1}{2})^2+\cdots]\,, \end{aligned} \tag{5}$$

where β_n are usually called the "coefficients of anharmonicity" and β_1 is always negative in practice. The asymmetry of the potential also relaxes the selection rule so that in addition to the strong allowed absorption transition $0 \rightarrow 1$, there are much weaker transitions from $n=0$ to higher levels. The absorption bands associated with these transitions have frequencies that are in approximate, but not exact, harmonic relationship to the fundamental, and are called "overtone bands."

Although the quantum treatment of a nonlinear oscillator may seem to conflict with the classical treatment, and the term *anharmonic* certainly suggests this, there is not really any disagreement. The infrared spectrum is derived from transitions between two levels of different energies, and therefore different classical amplitudes, and the classical frequency depends upon amplitude, as for the Chinese opera gong.

NEVILLE FLETCHER

See also **Damped-driven anharmonic oscillator; Harmonic generation; Molecular dynamics; Nonlinearity, definition of; Ordinary differential equations, nonlinear; Partial differential equations, nonlinear; Spectral analysis**

Further Reading

Fletcher, N.H. 1985. Nonlinear frequency shifts in quasi-spherical shells: pitch glide in Chinese gongs. *Journal of the Acoustical Society of America*, 78: 2069–2073

Fletcher, N.H. 1999. The nonlinear physics of musical instruments. *Reports on Progress in Physics*, 62: 721–764

Fletcher, N.H. & Rossing, T.D. 1998. *The Physics of Musical Instruments*, 2nd edition, New York: Springer

Harmony, M.D. 1989. Molecular spectroscopy and structure. *A Physicist's Desk Reference*, edited by H.L. Anderson, New York: American Institute of Physics, pp. 238–249

Herzberg, G. 1950. *Molecular Spectra and Molecular Structure*, 2nd edition, New York: Van Nostrand

Sethares, W.A. 1998. *Tuning, Timbre, Spectrum, Scale*, London and New York: Springer

Zwicker, E. & Fastl, H. 1999. *Psychoacoustics: Facts and Models*, 2nd edition, Berlin and New York: Springer

P

PAINLEVÉ ANALYSIS

Most problems of physics, chemistry, biology, engineering, and practically every applied science are nonlinear, in the sense that effects are not related to their causes by simple proportionality. Further, in most cases, these problems are described by differential equations, which provide the rules by which the main observables (such as displacements, velocities, fields, chemical concentrations, or populations) vary with respect to changes in the continuous independent variables (space and/or time). If the independent variables are discrete, the governing rules become difference equations.

Unfortunately, there is no general theory for integrating differential equations globally, that is, for finding their solutions analytically, over the full domain of the independent variables and for arbitrary initial and boundary data. That is why scientists often concentrate their analysis in regimes where the system behaves linearly, for example, small vibrations of lattices and solid structures, Fourier's law of heat conduction, Ohm's and Kirchhoff's laws in electrical circuits, and Laplace's and Maxwell's equations in electromagnetism.

However, even in such problems, the theory of linear differential equations often meets with serious difficulties, when the solutions possess singularities, that is, space and/or time values at which these solutions or their derivatives become infinite. In linear problems, these singularities appear explicitly in the equations, and if they are of special form, the general solution can still be found analytically near the singularity, for example, in terms of convergent series expansions. The theory of Lazarus Fuchs and Ferdinand Frobenius was especially developed for this purpose in the 1860s and led to the remarkable discovery of special functions, associated with such classic second-order ordinary differential equations (ODEs), as those of Bessel, Hermite, Legendre, or the hypergeometric equation.

By contrast, nonlinear ODEs possess singular points that are not evident by inspection of the equation itself. For example, the famous Riccati equation

$$\mathrm{d}x/\mathrm{d}t = a + bx + x^2 \tag{1}$$

(a, b constant parameters) known since the 1700s (*See* **Riccati equations**), has solutions which behave as

$$x(t) = \frac{1}{t-c} + a_0 + a_1(t-c) + a_2(t-c)^2 + \cdots, \tag{2}$$

where c is an arbitrary constant. Thus, this type of singularity, at $t = c$, is called movable, since its location depends on the choice of initial conditions. Furthermore, after the pioneering work of Augustin Cauchy in the 1830s showed the importance of complex variables, a new branch of mathematics developed in which ODEs are studied in the complex domain, with the independent variable taking values $t = t_{\mathrm{R}} + \mathrm{i}t_{\mathrm{I}}$, where $t_{\mathrm{R}}, t_{\mathrm{I}} \in \Re$ are, respectively, the real and imaginary parts of t and $\mathrm{i} = \sqrt{-1}$. As is well known, in this t-plane, the singularity at $t = c$ in (2) is called a pole and limits the convergence region of the Taylor expansion of the function $x(t)$ about any nearby point $t = t_0$ (where $x(t)$ is analytic), to $| t - t_0 | < | c - t_0 |$.

Still, when moving around a circle centered at the pole $t = c$ (and enclosing only this singularity), $x(t)$ always returns to the same values, implying that it is a single-valued function in that region. This is not what happens if $t = c$ is a branch point of $x(t)$ of the type

$$x(t) = (t-c)^{\alpha} + \cdots \quad \text{as } t \to c \tag{3}$$

with α not an integer. For example, if α is a rational number, $\alpha = p\,/\,q$, it takes q turns around $t = c$ for $x(t)$ to return to its starting value, demonstrating that it is a multi-valued function, describing a finitely sheeted Riemann surface, while $t = c$ is called an algebraic singularity. If, on the other hand, α in (3) is irrational or complex, $t = c$ is called a transcendental branch point and $x(t)$ describes an infinitely sheeted Riemann surface. Finally, if an expansion of $x(t)$ about $t = c$ contains logarithmic terms of the form $\log(t - c)$, it is again an infinitely multi-valued function and $t = c$ is called a logarithmic branch point.

Paul Painlevé, the great French mathematician and statesman (he was Prime Minister of France in 1917 and 1925), pursuing the idea that the problems with the simplest singularities would be the easiest to solve, set out to determine which first-order ODEs of the form

$$\mathrm{d}x/\mathrm{d}t = f(t, x) \tag{4}$$

with f rational in x and analytic in t, are free from movable branch points; that is, their only movable singularities are poles. His remarkable discovery was that there is only one such equation: the Riccati equation (1), with the coefficients of 1, x, and x^2 being arbitrary analytic functions of t. Then, in a remarkable series of papers in the late 1890s and early 1900s, Painlevé and his coworkers (notably Bertrand Gambier) studied the same question on all second-order ODEs

$$\mathrm{d}^2x/\mathrm{d}t^2 = f(t, x, \mathrm{d}x/\mathrm{d}t) \tag{5}$$

with f rational in $\mathrm{d}x/\mathrm{d}t$, algebraic in x, and analytic in t. After painstaking analysis, they showed that there are 50 such equations, 44 of which can be explicitly integrated and solved in terms of known functions. For the remaining six equations (called Painlevé I–VI), such reduction was not possible, and new functions had to be introduced called the Painlevé transcendents. The first two of these equations are

$$\begin{aligned} &\text{Painlevé I}: \quad \mathrm{d}^2x/\mathrm{d}t^2 = 6x^2 + t, \\ &\text{Painlevé II}: \quad \mathrm{d}^2x/\mathrm{d}t^2 = 2x^3 + tx + a \end{aligned} \tag{6}$$

with Painlevé III–VI being increasingly more complicated to write down. The interested reader will find them all discussed in the excellent books by Ince (1956) and Davis (1962), where Painlevé's approach to their discovery is also described.

It is important to recall, however, that independent of all this progress, the idea of using singularity analysis to solve systems of ODEs by requiring that their solutions have only poles was also used by the Russian mathematician Sophia Kovalevsky in 1888 to find one more solvable case of the rotating rigid body that bears her name (Kovalevsky's top). Still, the requirement that all solutions of a system of ODEs possess only poles as movable singularities in the complex t-plane has come to be referred to date as the Painlevé property.

Surprisingly enough, the Painlevé theory was not widely appreciated at first and remained largely unknown until it was revived, many decades later, in connection with exactly solvable partial differential equations (PDEs). More specifically, after the discovery that a number of PDEs are integrable by the inverse scattering transform (IST), Ablowitz and Segur observed in the mid-1970s that all ODE reductions of these PDEs had the Painlevé property and led, in fact, to some of the Painlevé I–VI equations, (6) (see Ablowitz & Segur, 1981, as well as Ablowitz & Clarkson, 1991).

Thus, the famous Ablowitz, Ramani, and Segur conjecture was formulated as follows: If a PDE is solvable by IST, all its ODE reductions obey the Painlevé property. Clearly, this conjecture provides a useful criterion to test the integrability of a given PDE by studying its ODE reductions—a much easier task than showing whether or not it is solvable by IST.

This discovery had enormous consequences, as it attracted the attention of many researchers working in the rapidly expanding field of nonlinear dynamics. The fact that most nonlinear dynamical systems of ODEs possess "irregular," "unpredictable," or chaotic solutions evidently suggests that they cannot be integrated and solved in terms of known functions, whose behavior is perfectly regular and predictable. It would thus be extremely interesting to find, in models of physical, chemical, or biological dynamics, integrable cases whose solutions would be clearly distinguished from those displaying chaotic behavior.

Such new, integrable systems could now be identified by means of the Painlevé property. Perhaps one of the most important applications occurred in Hamiltonian systems, like the famous Hénon–Heiles problem

$$H = \frac{1}{2}(\dot{x}^2 + \dot{y}^2 + Ax^2 + By^2) + x^2y + \frac{C}{3}y^3 \tag{7}$$

of two degrees of freedom (dots denote differentiation with respect to t). For $A = B = 1$ and $C = -1$, this problem had been extensively studied since the mid-1960s, as an example of a system whose chaotic regions grow dramatically as the total energy $H = E$ increases from 0 to $\frac{1}{6}$. Using the Painlevé analysis, it was first shown by Bountis et al. (1982) that (7) is completely integrable in exactly three cases:

(i) $C = 1$ and $A = B$ (known to be separable in the variables $s = x + y$, $d = x - y$).
(ii) $C = 6$ and any A and B.
(iii) $C = 16$ and $B = 16A$.

In each of the new cases (2) and (3), the second integral—besides Hamiltonian (7)—was also provided and their complete integrability in the sense of Liouville–Arnol'd was established.

Soon, many novel and highly nontrivial examples of Hamiltonian systems with N degrees of freedom were identified by requiring that their solutions possess the Painlevé property. Rigorous connections between integrability and the Painlevé property were also developed, mainly by Adler and van Moerbeke in the early 1980s, who used algebraic geometry to establish that if a Liouville–Arnol'd integrable Hamiltonian system has rational integrals which continue to describe tori in the complex domain, then its solutions have only movable poles (the converse, though not completely proved, is also believed to hold). A wealth of examples of integrable non-Hamiltonian systems were also discovered, related to models of physical, chemical, or biological interest. Finally, the Painlevé analysis was extended and applied by Weiss, Tabor, and Carnevale to test integrability directly on a PDE, without reference to its ODE reductions (see Ramani et al., 1989).

But the importance of Painlevé analysis does not end here. After the early observation that any deviation from the Painlevé conditions generically introduces logarithmic terms, researchers began to look for connections between the violation of the Painlevé property and non-integrability, in the sense of non-existence of analytic, single-valued integrals such as (7). In that regard, the work of the Russian mathematicians V.V. Kozlov and S.L. Ziglin in the late 1970s turned out to be extremely important (see Kozlov, 1983). They showed, using Mel'nikov's theory, that one of the most fundamental causes of chaotic behavior, that is, the transverse intersection of stable and unstable manifolds of saddle fixed points on a Poincaré map, implies the presence of infinitely multi-valued solutions and non-existence of a second integral in a two-degree-of-freedom Hamiltonian systems.

Kozlov's and Ziglin's results inspired a series of interesting papers by H. Yoshida in the 1980s, where he used them to study the variational equations of simple periodic solutions in a large class of N-degree-of-freedom Hamiltonians. Yoshida was able to prove that when these variational equations do not satisfy certain conditions imposed by the existence of a full set of global single-valued integrals, such a set of constants cannot exist, thus demonstrating non-integrability for vast parameter ranges in many Hamiltonians of physical interest (see Ramani et al., 1989).

Finally, what about systems of ODEs (or PDEs) whose solutions possess only algebraic singularities, that is, points $t = c$, near which solutions have asymptotic expansions of the form

$$x(t) = (t - c)^{p/q} + \sum_{n=0}^{\infty} a_n (t - c)^{n/q} \qquad (8)$$

without any other singularities present? In the 1980s, it was thought that this so-called weak Painlevé property, under some conditions on $p\,/\,q$, could also be related to complete integrability, in the Liouville–Arnol'd sense for N-degree-of-freedom Hamiltonian systems. In the 1990s, however, it was found that the situation is considerably more subtle.

First, it was shown by Goriely that many weak Painlevé examples could be transformed to systems having the usual Painlevé property after some rather general changes of coordinates (see Goriely, 2001). Then it was observed that there were weak Painlevé systems with chaotic solutions, which are evidently not integrable. Thus, the concept of globally finitely sheeted solutions (FSS) was introduced to distinguish integrable systems with algebraic singularities (8) from non-integrable ones, whose solutions are only locally finitely sheeted (see Bountis, 1992, 1995).

This is done as follows: every time a system of weak Painlevé ODEs, integrated numerically around arbitrarily large contours, showed evidence of "lattice" singularity patterns in the complex t-plane and FSS, it has been shown to be transformable to one that is completely integrable and possesses the usual Painlevé property. On the other hand, many such systems were also found which showed "dense" singularity patterns and evidence of globally infinitely sheeted solutions (ISS) for large enough contours. Although not solvable, such systems can still be integrable if they are hyperelliptically separable, that is, if they can be described by N holomorphic differentials on the Jacobian of a hyperelliptic curve of genus $g < N$ (see Abenda & Fedorov, 2000; Abenda et al., 2001).

Finally, in the 1990s, the connection between singularities and integrability was extended to discrete dynamical systems, described by difference equations. Here, since the methods of complex analysis for ODEs no longer apply, a novel criterion was introduced (see Grammaticos et al., 1991). It was based on the observation that if a difference equation such as

$$x_{n+1} = F(x_n, x_{n-1}), \quad n = 0, 1, 2 \ldots \qquad (9)$$

with F rational, possesses an integral of the form

$$I(x_n, x_{n-1}) = \text{const.}, \quad n = 0, 1, 2 \ldots \qquad (10)$$

yielding a family of curves in the x_n, x_{n+1} plane and precluding the presence of chaos, then x_n is infinite at some $n = m$ and becomes finite again at some later $n = m' > m$. This so-called singularity confinement criterion, whose success to date remains a mystery, has yielded a wealth of integrable discrete systems and has enabled many researchers to develop Painlevé difference equations, analogous to the famous Painlevé ODEs Painlevé I–VI (see Conte, 1999).

TASSOS BOUNTIS

See also **Hénon–Heiles system; Inverse scattering method or transform; Mel'nikov method; Riccati equations**

Further Reading

Abenda, S. & Fedorov, Y. 2000. On the weak Kowalevski–Painlevé property for hyperelliptically separable systems. *Acta Applicandae Mathematicae*, 60: 137

Abenda, S., Marinakis, V. & Bountis, T. 2001. On the connection between hyperelliptic separability and Painlevé integrability. *Journal of Physics* A, 34: 3521

Ablowitz, M.J. & Clarkson, P. 1991. *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, Cambridge and New York: Cambridge University Press

Ablowitz, M.J. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM

Adler, M. & van Moerbeke, P. 1988. *Algebraically Completely Integrable Systems*, Boston: Academic Press,

Bountis, T. 1992. What can complex time tell us about real time dynamics? *International Journal of Bifurcation and Chaos*, 2(2): 217

Bountis, T. 1995. Investigating non-integrability and chaos in complex time. *Physica* D, 86 (1995): 256

Bountis, T., Segur H. & Vivaldi, F. 1982. Integrable Hamiltonian systems and the Painlevé property. *Physical Review* A, 25: 1257

Conte, R. (editor). 1999. *The Painlevé Property: One Century Later*, New York: Springer

Davis, H.T. 1962. *Introduction to Nonlinear Differential and Integral Equations*, London: Dover

Goriely, A. 2001. *Integrability and Non-integrability of Dynamical Systems*, Singapore: World Scientific

Grammaticos, B., Ramani, A. & Papageorgiou, V. 1991. Do integrable mappings have the Painlevé property? *Physical Review Letters*, 67(14): 1825–1828

Ince, E.L. 1956. *Ordinary Differential Equations*, London: Dover

Kozlov, V.V. 1983. Integrability and non-integrability in Hamiltonian mechanics. *Russian Mathematical Surveys* (Mathematical Review C) 38: 1

Ramani, A., Grammaticos, B. & Bountis, T. 1989. The Painlevé property and singularity analysis of integrable and non-integrable systems. *Physics Reports*, 180 (3): 160.

Tabor, M. 1989. *Chaos and Integrability in Nonlinear Dynamics*, New York: Wiley Interscience

PARAMETRIC AMPLIFICATION

Parametric generation and amplification of oscillations are based on the phenomenon of parametric resonance, using the work done by an external force under periodic variation of the parameters of an oscillatory system.

A simple mechanical system in which parametric resonance may occur is a pendulum, with the length of the cord (l) changing with time. If l is decreased when the pendulum is in the lower position and increased in its upper position, the average work done by an external force over one period is positive and oscillations of the pendulum build up. This phenomenon is the underlying principle of setting a swing into motion. When one stands on a swing, it is easier to rock it by squatting at the top position of the swing and standing up at the bottom position, thus shifting the center of mass of the person on the swing twice in one period.

An electric analog of a pendulum of variable length is an oscillatory circuit of variable capacitance. The capacitance C can be changed, for example, by mechanically moving the capacitor plates together or apart. For the amplitude of oscillations to build up, energy should be fed to the circuit by performing work against the electrostatic field forces of the capacitor. This means that the plates should be moved apart when the charge q on the capacitor is maximal and moved together when the charge on the capacitor returns to zero (Figure 1).

Such a periodic change in capacitance (pumping-up process) results in growing energy of the oscillations and appearance of instability in the system. The build-up of energy occurs at a pump frequency satisfying the relation

$$\omega_p = 2\omega_0/n, \qquad (1)$$

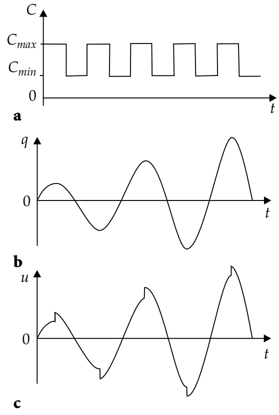

Figure 1. Periodic variations of capacitance C of capacitor (a), capacitor charge q (b), and voltage u (c) at parametric resonance.

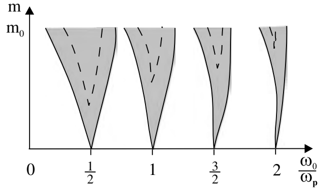

Figure 2. Mathieu instability zones within which parametric resonance may occur (solid line—without damping; dashed line—with damping; ω_0 natural frequency of the oscillatory circuit; ω_p pump frequency.

where n is an integer, ω_0 is the natural frequency of circuit oscillations, and ω_p is the pump frequency. Build-up is most effective for $n = 1$, that is, when the period of capacitance variation is half the period of natural oscillations of the circuit. Interestingly, energy build-up (parametric resonance) is possible not only for the values of ω_p satisfying Equation (1) but also at some deviations from these values within Mathieu instability zones (see Figure 2). The greater the width of Mathieu zones, the greater the extent of parameter variation $m = (C_{max} - C_{min}) / (C_{max} + C_{min})$.

In oscillatory systems with several degrees of freedom (e.g., in a system with two coupled circuits, coupled pendulums, and so on), normal modes with different natural frequencies ω_1 and ω_2 may exist. As parameters of such a system change, oscillations may build up (parametric resonance) not only at a frequency satisfying condition (1) for an arbitrary natural frequency, but also for their linear combination, for instance, when a parameter changes with sum frequency

$$\omega_{\rm p} = \omega_1 + \omega_2. \qquad (2)$$

Resonance coupling is also possible at $\omega_{\rm p} = \omega_1 - \omega_2$. In this case, however, no build-up of oscillations occurs. Instead, energy is periodically pumped between the modes ω_1 and ω_2. For such an oscillatory system, the pump power $P_{\rm p}$ input to the system at frequency $\omega_{\rm p}$ and the powers P_1 and P_2 consumed at frequencies ω_1 and ω_2 are proportional to the corresponding frequencies (the Manley–Rowe relation):

$$\frac{P_{\rm p}}{\omega_{\rm p}} = \frac{P_1}{\omega_1} = \frac{P_2}{\omega_2}. \qquad (3)$$

Parametric resonance may also provide conditions for excitation of normal modes in oscillatory media possessing an infinite number of degrees of freedom. In 1831, Michael Faraday had observed that standing waves are excited at frequency $\omega_{\rm p/2}$ in the liquid in a vessel vibrating in the vertical direction at frequency $\omega_{\rm p}$. The classic experiment staged by Franz Melde in 1859 revealed excitation of transverse oscillations (standing waves) in a string, one end of which was fastened to the prong of a tuning fork. When the tuning fork vibrates and the string is tight, the string performs transverse oscillations with a frequency equal to half the frequency of the tuning fork.

A distinguishing feature of parametric resonance in systems with distributed parameters is that the build up of oscillations depends strongly on the relationship between the variation in space of the system parameters and the spatial structure of the forcing oscillations (waves). For example, if the pump that changes the parameters of the medium is a traveling wave with frequency $\omega_{\rm p}$ and wave vector $\boldsymbol{k}_{\rm p}$, then natural waves with frequencies ω_1 and ω_2 and wave vectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$ are excited, provided the conditions of parametric resonance are fulfilled both in space and time; thus

$$\omega_{\rm p} = \omega_1 + \omega_2, \quad \boldsymbol{k}_{\rm p} = \boldsymbol{k}_1 + \boldsymbol{k}_2. \qquad (4)$$

The potentialities of using parametric resonance for creating a parametric generator and amplifier were studied by Leonid I. Mandelshtam and Nikolai D. Papaleksi (1931–1933). They constructed capacitive and inductive parametric machines that transformed mechanical energy into electrical energy as a result of variation of the magnitudes of capacitance and inductance during shaft rotation. Parametric generators and amplifiers found practical application in the 1950s when semiconductor parametric diodes appeared, the capacitance of which depends on applied voltage, and when variable capacitors using the properties of ferroelectrics and variable inductors using the properties of ferrites and superconductors were developed. Low-noise variable capacitance parametric amplifiers based on high-frequency parametric diodes (varactors) are used in high-frequency receiving microwave devices employed in radar and radio astronomy, among other applications. Parametric generators are capable of generating oscillations whose phase can take one of two values differing by π. This feature of parametric generators was noted by Eiichi Goto who proposed in 1954 to use such generators, which he referred to as "parametrons," as logic switching elements in computers. Parametric generators are widely employed in optics as tunable coherent generators.

Note finally that parametric action on a nonlinear oscillatory system (for example, a nonlinear pendulum with variable-length cord or a nonlinear circuit with periodically varying capacitance and inductance) may also lead to dynamic oscillations.

VLADIMIR SHALFEEV

See also **Diodes; Dynamical systems; Harmonic generation; Manley–Rowe relations; Nonlinear toys**

Further Reading

Leven, R.W. & Koch, B.P. 1981. Chaotic behavior of a parametrically excited damped pendulum. *Physics Letters*, 86(2): 71–74

Rabinovich, M.I. & Trubetskov, D.I. 1989. *Oscillations and Waves in Linear and Nonlinear Systems*, Dordrecht and Boston: Kluver

Scott, A. 1970. *Active and Nonlinear Wave Propagation in Electronics*, New York: Wiley

PARAMETRIC DECAY INSTABILITY

See **Nonlinear plasma waves**

PARTIAL DIFFERENTIAL EQUATIONS, NONLINEAR

A partial differential equation (PDE) is a functional equation of the form

$$F\left(x_1, x_2, \ldots, x_n, z_1, \ldots, z_m, \frac{\partial z_\mu}{\partial x_\nu}, \ldots, \frac{\partial^2 z_\mu}{\partial x_1^2}, \frac{\partial^2 z_\mu}{\partial x_1 \partial x_2}, \ldots\right) = 0 \qquad (1)$$

with m unknown functions $z_1, z_2, \ldots, z_m$ with n independent variables $x_1, x_2, \ldots, x_n$ $(n > 1)$ and at least one of the partial derivatives of z_μ $(\mu = 1, 2, \ldots, m)$ with respect to x_ν $(\nu = 1, 2, \ldots, n)$. In general, there can be a system of coupled PDEs.

For a single equation, the order of the PDE is p if the highest derivative has order p, with an analogous definition for a system of equations. A PDE becomes an ordinary differential equation if the number of independent variables n is one. It is linear if the equation is linear with respect to z and its partial derivatives, meaning it is of the first degree in the unknown function as well as its derivatives. It is called quasilinear if it is linear with respect to the highest-order partial derivatives which have coefficients that depend only on lower-order partial derivatives and the independent variables. In addition, a PDE is called semilinear if the coefficients of the highest-order derivatives depend on the independent variables only. A PDE is called nonlinear if it does not fit into any of the above categories.

Initial value problems (IVPs) for evolution equations arise where one of the independent variables x_i of Equation (1) represents time t and an initial condition at time $t = 0$ is given. Additionally, there are boundary value problems (BVPs), when spatial variables are involved and boundary conditions are specified along the n-dimensional boundary of the domain (where n is the number of independent spatial variables). There can also be mixed initial value-boundary value problems (IVBVPs), where both types of conditions are given.

For a single PDE, the number of initial conditions is generally one less than the highest temporal derivative, while the number of spatial boundary conditions is half the order of the highest spatial derivatives in the equation. For general nonlinear equations, there are no precise definitions.

Second-order semilinear PDEs have the general form

$$a\frac{\partial^2 z}{\partial x_1^2} + 2b\frac{\partial z}{\partial x_1 \partial x_2} + c\frac{\partial^2 z}{\partial x_2^2} + \text{l.o.t.} = 0,$$

where a, b, c are given functions of x_1 and x_2 or constants and l.o.t. stands for lower order terms. The coefficient matrix A of the corresponding second-order system has the determinant

$$\det A = \begin{vmatrix} a & b \\ b & c \end{vmatrix}. \qquad (2)$$

A second-order semilinear PDE is said to be of elliptic type if $\det A \ (= ac - b^2) > 0$, of hyperbolic type if $\det A < 0$, and of parabolic type if $\det A = 0$ for all points x_1, x_2 of the domain D. The determinant of A is positive if both eigenvalues of the matrix A are positive or both negative. The determinant is negative if the eigenvalues are of opposite sign and zero if one of them is zero. The normal forms of the associated differential operators L are given by

$$\det A > 0: \qquad \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} \quad \text{(elliptic)},$$

$$\det A < 0: \qquad \frac{\partial^2}{\partial x_1^2} - \frac{\partial^2}{\partial x_2^2} \quad \text{(hyperbolic)},$$

and

$$\det A = 0: \qquad \frac{\partial^2}{\partial x_1^2} \quad \text{(parabolic)}.$$

Boundary value conditions may be "Dirichlet" (where the value of the solution is given along the boundary ∂D of the domain D), "Neumann" (where the normal derivative $\partial z / \partial n$ is specified along the boundary ∂D), and "Robin" (where both the value and the normal derivative of the solution $\partial z / \partial n + az$ satisfy an equation along the boundary of the domain ∂D).

The archetypical example for an elliptic PDE is the Laplace equation

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0;$$

for a hyperbolic PDE, it is the wave equation

$$\frac{\partial^2 z}{\partial t^2} - c^2\frac{\partial^2 z}{\partial x^2} = 0;$$

and for a parabolic PDE, it is the diffusion equation

$$\frac{\partial z}{\partial t} - \nu\frac{\partial^2 z}{\partial x^2} = 0.$$

The coefficient matrix A (2) can depend on first-order derivatives if the PDE is second-order quasilinear and the classes are analogous. Quasilinear and strictly linear second-order PDEs are classified according to these three classes. For a given nonlinear PDE of higher order or more than two variables, it is not useful to classify according to these three classes, instead, one classifies it according to the physical process it describes. In general, boundary value problems are often associated with elliptic equations and initial value problems with hyperbolic and parabolic equations.

Nonlinear PDEs describe many important processes in nature because often an approximate linear PDE is not sufficient to describe the process. Examples are the dynamics of fluids, gases, and elastic media; the equations from general relativity and quantum electrodynamics; and population dynamics in biology and chemistry.

The two main characteristics of such processes are the nonlinear interactions, due to the nonlinearities of the PDE and the possibility of self-organization into coherent structures.

There are three types of processes that one can distinguish. Reversible processes such as waves; irreversible; processes such as reaction, diffusion and heat flow; and stationary processes. A process is irreversible/reversible if it is impossible/possible to reverse the process in time and stationary processes are independent of time. In general, reversible processes are described by hyperbolic equations, irreversible processes by parabolic equations, and stationary processes by elliptic equations.

Most physical systems comprise a combination of all three types of processes, but often, some processes are dominant and by using a scaling argument, it is possible to neglect less dominant processes and simplify the PDE to the degree that it describes the process adequately but is still easy to work with.

The nature of irreversible processes such as diffusion, is that they tend to smooth the solution in time, whereas for reversible processes singularities can develop, for example, shock waves in gas dynamics. For stationary processes, the solutions depend on the smoothness of the boundary conditions and external forces.

Solutions to nonlinear PDEs are often obtainable only by numerical integration. In integrable cases, on the other hand, analytic solutions can be found using advanced techniques such as the inverse scattering method, the associated Lax theory, and the Bäcklund transformation. There are other techniques that are helpful although not exact, such as Floquet and Mel'nikov theory.

In general, for nonlinear PDEs common tools such as the linear superposition principle are not valid anymore, although it is sometimes possible to find a nonlinear superposition principle. Methods such as the Fourier and Laplace transforms and Green's function, in general, are not applicable to nonlinear PDEs. Alternatively, one can attempt classical solution methods such as the separation of variables or transformation of variables to reduce the nonlinear PDE to a system of equations that is solvable analytically.

ANDREAS A. AIGNER

See also **Burgers equation; Korteweg–de Vries equation; Nonlinear Schrödinger equations; Separation of variables; Sine-Gordon equation**

Further Reading

Courant, R. & Hilbert, D. 1966. *Methods of Mathematical Physics II*, New York: Wiley

Debnath, L. 1997. *Nonlinear Partial Differential Equations*, Basel: Birkhäuser

Garabedian, P. 1964. *Partial Differential Equations*, New York: Wiley

Ockendon, J., Howison, S., Lacey, A. & Movchan, A. 1999. *Applied Partial Differential Equations*, Oxford and New York: Oxford University Press

PARTICLE ACCELERATORS

Particle accelerators have been used for much of the 20th century to investigate the properties of matter. The first accelerators of charged particles used steady (d.c.) electric fields, such as those obtained from a Van de Graaff generator. These devices could accelerate particles to above an energy of a million electron volts (MeV), which was sufficient to probe atomic nuclei. Applications needing higher energies continually emerged, leading to new types of devices, particularly ones having the charged particles move in a circular orbit. In that configuration, with the additional concept that electric fields and particles can be synchronized, such that on each crossing of a gap, particles encounter a field that is accelerating, the basic accelerator configuration was created. Such devices went through a series of developments from a fixed frequency cyclotron to a frequency modulated cyclotron and finally the synchrotron. In the first two devices, the equilibrium orbit spirals out as the energy increases, while in the last device, the orbit is held fixed, and the bending magnetic field is increased.

To keep the particles on a prescribed essentially circular orbit over many traversals of the device, the orbit must be stable to transverse perturbations. Furthermore, to achieve continuous acceleration, the particle motion must be stable about the equilibrium accelerating phase of the electric field. Because the particles are kept close to an equilibrium orbit, the transverse stability can be analyzed from linear stability analysis, using a first-order expansion of the fields about the equilibrium. When the gradients of the magnetic guide field are chosen to give stable oscillations around the equilibrium orbit, they are called betatron oscillations, after the betatron, an early circular induction accelerator for which the oscillations were first analyzed (Kerst & Serber, 1941). For particles traveling in a straight line, the forces resulting from magnetic gradients cannot focus in both transverse directions simultaneously. This is not true for circular orbits, as the gradient in the centrifugal force supplies an extra focusing force to allow focusing in both directions. The resulting homogenous linear equations of motion for the simplest, assumed azimuthally symmetric, device are

$$\frac{d^2z}{d\theta^2} + nz = 0 \qquad (1)$$

and

$$\frac{d^2x}{d\theta^2} + (1-n)x = 0, \qquad (2)$$

where z is the vertical position and x the horizontal deviation from the equilibrium orbit r_0, and θ is the azimuthal angle around the machine. Here, n is the field index

$$n = -\frac{r_0}{B_0}\frac{\partial B_z}{\partial r} \qquad (3)$$

with B_0 the equilibrium magnetic field. From (1) and (2) the stability condition is $0 < n < 1$. The azimuthally symmetric focusing of this type is now called weak focusing because the limitations on stability make the relative frequencies or tunes $\nu_z \equiv \omega_z / \omega_0 = n^{1/2}$ and $\nu_x \equiv \omega_x / \omega_0 = (1-n)^{1/2}$ both less than 1. The weak restoring forces required large vacuum chambers, limiting the size and field strength of the devices and therefore the energy to which particles could be accelerated.

The breakthrough in developing a synchrotron in which particles could be closely contained around their equilibrium orbit was the understanding that much stronger focusing could be achieved by alternating a section which strongly focused vertically while defocusing radially with a section that defocused vertically and focused radially. The net result, from simple lens theory, is that the pair can be strongly focusing. This led to the invention of the alternate gradient (AG) synchrotron (Courant et al., 1952). The simplest analysis of the linearized equations of motion is of the Hill equation

$$\frac{\mathrm{d}^2 y}{\mathrm{d}\theta^2} + K(\theta)y = 0 \tag{4}$$

with $K(\theta + \theta_L) = K(\theta)$, where θ_L is the repeating angular period. The transformation for a period can be analyzed for stability from Floquet theory by setting

$$y(\theta + \theta_L) = y(\theta)\mathrm{e}^{\pm \mathrm{i}\sigma} \tag{5}$$

and solving the transfer matrix, over the repeating period θ_L, for stability. This was done in the original paper, for a simple transfer matrix, leading to a stability diagram. However, even the early AG synchrotrons were more complicated, with bending magnets, focusing quadruple magnets, and straight sections in which resonators supplied the acceleration fields. For various reasons, described in part below, modern machines with very large radii, operating at many billions of electron volts (BeV or GeV), are much more complicated, leading to stability calculations with a very large number of elements.

However, even at the simplest level, calculating the linear stability of orbits is not sufficient. Nonlinear terms, although small, can have very important consequences. The motions given by (1) and (2), or more generally by (4), in each of the transverse directions are independent. The nonlinear terms couple the transverse dimensions and lead to a wide variety of resonances that can degrade the beam cross section. These difference resonances of various strengths must be avoided when finding a proper operating point in the $\nu_x - \nu_z$ parameter space. Additionally, magnet imperfections lead to entirely new classes of resonances. The effect of the increasing energy, coupled with the space charge of the beam of ions or electrons, leads to a movement of the operating point through the parameter space, such that all resonances cannot be avoided.

In addition to the primarily linear transverse oscillations, there are fundamentally nonlinear oscillations associated with the accelerating fields, called synchronous oscillations. For circular devices, these oscillations couple longitudinal and transverse motions. The frequencies of these electric-field-driven oscillations are generally much lower than the transverse betatron oscillations so that it is usually sufficient to consider their radial motion as adiabatic when studying the free oscillations, that is, using averaging methods. Depending on how relativistic the particles are, the phase motion may be either primarily longitudinal or radial. To simplify the treatment, we consider the synchronous motion in a linear accelerator such that transverse motions are unimportant to lowest order. The fields in both linear accelerators and synchrotrons are generated by high-power electromagnetic sources which operate in structures, for example, loaded waveguides in linear accelerators and resonant cavities in synchrotrons, that have a primary traveling-wave field synchronous with the particles. Nonsynchronous harmonic fields also exist, but are usually not important.

The equations of motion along the axis of a linear accelerator are the force equation

$$\frac{\mathrm{d}p}{\mathrm{d}t} = eE\sin\phi \tag{6}$$

and the equation for the change in phase

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \omega\left(1 - \frac{v}{v_0}\right), \tag{7}$$

where v_0 is the wave velocity, ω the r.f. frequency, and E the applied field. The phase stable particle has $v = v_0$, and for stable acceleration, v_0 must be determined from (6), where p and v are relativistically related. The motion of particles around the stable phase is given by trajectories in the $p - \phi$ phase plane with the stable fixed point at $\phi = \phi_\mathrm{s}$, satisfying $\mathrm{d}\phi\,/\,\mathrm{d}t = 0$, and an unstable fixed point on a separatrix orbit satisfying, $\phi_\mathrm{u} = -\phi_\mathrm{s}$. The phase diagram indicates that particles injected between ϕ_u and some limiting ϕ on the separatrix trajectory with velocity v_0 oscillate nonlinearly about ϕ_s, while those outside the phase interval are not stably captured. For a general account of the basic ideas as described above, and also the many types of accelerators and variants on the concepts, the reader is referred to basic accelerator texts. Two such texts, which include many useful original references are Livingood (1961) and Kolomensky & Lebedev (1966). The early dates of publication indicate the maturity of this field.

As AG synchrotrons increased in size and energy, it became increasingly important to match the emittance phase space of the incoming particles to the acceptance phase space of the accelerator. This was to ensure the highest beam brightness, particularly as the long acceleration cycle led to low duty ratios. Phase space matching techniques were developed and exploited at the CERN high-energy physics center (Hereward et al., 1956). The transverse acceptance of a simplified AG synchrotron is illustrated in Figure 1.

The need for higher beam brightness led to the invention of the fixed field alternating gradient (FFAG) synchrotron which injected a number of pulses whose phase spaces were spatially separated by the energy increase. This phase space beam-stacking in the FFAG synchrotron further enhanced the usefulness of employing phase space matching concepts (Symon &

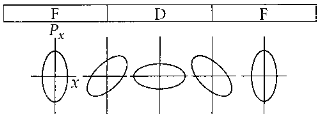

Figure 1. Phase-space transformations in an AG synchrotron.

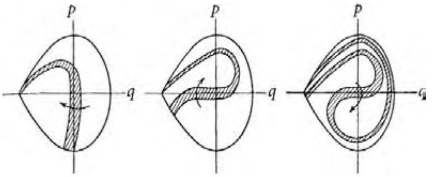

Figure 2. Filamentation of phase space due to nonlinear oscillations.

Sessler, 1956). Phase space matching was also used to increase the efficiency of linear accelerators, and of microwave devices (*See* **Electron beam microwave devices**). The effect of a mismatch of emittance and acceptance is illustrated in Figure 2 in which a beam with small momentum spread (cross-hatched region) is injected into the acceptance phase space of a synchronous oscillation in a linear accelerator or synchrotron, leading to an increase of the effective phase space area by filamentation (Lichtenberg, 1969).

The development of storage rings greatly increased the importance of keeping the phase space of the injected particles to a minimum. Conceived as a device to obtain high beam brightness, it rapidly became essential for operation of proton accelerators at energies many times the proton rest energy for which the energy in the center of mass of collisions with stationary matter was greatly reduced due to momentum conservation. This led to intersecting storage rings of particles traveling in opposite directions and a very severe requirement of high beam brightness (O'Neil, 1956). Storage rings, which store particles over many machine transits, have required the control of beam spreading. In electron storage rings, the synchrotron radiation which set a limit to electron energy is also effective in reducing the beam phase space (Robinson, 1958). For protons, which are little affected by radiation, stochastic cooling was invented to perform a similar function, a requirement that is essential for proton–antiproton colliders (Van der Meer, 1972).

The development of higher energy accelerators and higher intensity beams has led to new methods of analysis and various problems to be solved. The largest AG synchrotrons include many correcting magnets as well as the bending magnets, quadrupoles, and straight sections. Orbit calculations have required symplectic integrators to minimize numerical errors, which have involved the development of a Lie algebraic method (Dragt et al., 1988). Various resonances, coupled with machine errors, have led to exploration of resistive and Arnol'd diffusion explanations for beam spreading. Intense beams in electron linear accelerators have produced a pulse-shortening phenomenon that was analyzed as a nonlinear excitation of an electromagnetic mode in the accelerating structure, and methods of suppressing it were devised (Chao, 1993). Space-charge effects that become negligible at highly relativistic energies are again introduced by the beam-beam interaction in intersecting storage ring beams. Analysis of this subtle phenomenon within the environment of the very complicated beam dynamics has led to new exploration of simple mapping models (*See* **Fermi acceleration and Fermi map**), together with the use of averaging methods to smooth out various more rapid phenomena that do not play an essential role. In linear devices, to go beyond the 30 GeV two-mile long accelerator at SLAC (Stanford), which is field-strength limited, new ideas of structureless acceleration of ions in the wakefields produced by electron beams in plasmas and of electron acceleration, using the wakefield of an intense laser in a plasma, are being actively explored (see Esarey et al. (1996) and other articles in the same issue). The great challenges that face accelerator design above 300 GeV and intersecting beam interactions in storage rings keep a sense of excitement in a mature field. Much of the modern research and new ideas can be reviewed, with references, in Chao & Tigner (1999).

At the same time that the interaction energy has been increased to probe deeper into the fundamental properties of matter, new uses have been found for low-energy accelerators. On the atomic scale, materials are being probed with ultraviolet light and X-rays, using the natural radiation of electrons in circular devices or enhanced radiation employing undulators in combination with accelerators as advanced light sources. These latter devices typically have short period spatially oscillating magnetic fields with the accompanying electron oscillations relativistically doppler shifted to high frequencies (*See* **Electron beam microwave devices**). At still lower energies, accelerators are being used in various medical applications. Some of the technology-oriented work is treated by Scharf (1989).

ALLAN J. LICHTENBERG

See also **Arnol'd diffusion; Averaging methods; Electron beam microwave devices; Fermi acceleration and Fermi map; Hamiltonian systems**

Further Reading

Chao, A. 1993. *Physics of Collective Beam Instabilities in High Energy Accelerators*, New York: Wiley

Chao, A.W. & Tigner, M. (editors). 1999. *Handbook of Accelerator Physics and Engineering*, Singapore: World Scientific

Courant, E.D., Livingston, M.S. & Snyder, H.S. 1952. The strong-focusing synchrotron—a new high energy accelerator. *Physical Review*, 88: 1190–1196

Dragt, A.J., Neri, F. & Rangaranjan, G. 1988. Lie algebraic treatment of linear and nonlinear beam dynamics. *Annual Review of Nuclear and Particle Science*, 38: 455–496

Esarey, E., Sprangle, P., Krall, J. & Ting, A. 1996. Overview of plasma-based accelerator concepts. *IEEE Transactions on Plasma Science*, 24: 252–288

Hereward, H.G., Johnson, K. & Lapostolle, P. 1956. Problems of injection. *CERN Symposium*, 179–191

Kerst, D.W. & Serber, R. 1941. Electronic orbits in the induction accelerator. *Physical Review*, 60: 53–58

Kolomensky, A.A. & Lebedev, A.N. 1966. *Theory of Cyclic Accelerators*, Amsterdam: North-Holland

Lichtenberg, A.J. 1969. *Phase Space Dynamics of Particles*, New York: Wiley

Livingood, J.J. 1961. *Principles of Cyclic Particle Accelerators*, Princeton: Van Nostrand

O'Neil, G.K. 1956. Storage ring synchrotron: device for high energy physics research. *Physical Review*, 102: 1418–1419

Robinson, K.W. 1958. Radiation effects in circular electron accelerators. *Physical Review*, 111: 373–380

Scharf, W. 1989. *Particle Accelerator Applications in Technology and Research*, Taunton, England: Research Studies Press Ltd.

Symon, K.R. & Sessler, A.M. 1956. Methods of radiofrequency acceleration in fixed field accelerators with applications. *CERN Symposium*, 44–58

Van der Meer, S. 1972. Stochastic damping, of betatron oscillations in the ISR. *CERN Report ISR–PO*, pp. 72–81

PARTICLES AND ANTIPARTICLES

For a good understanding of the physical concept of *antiparticle* and the closely related concept of *charge*, it is important to appreciate how these notions emerged. Therefore, we begin with a brief sketch of the associated history.

In the early days of quantum mechanics, the material world was thought to be built from three elementary particles, namely, the electron, the proton, and the neutron. The idea that there might be associated particles with the same mass and opposite charge (now called antiparticles) first arose from the characteristics of the one-particle Dirac equation. Writing it in Hamiltonian form, the resulting Dirac Hamiltonian has spectrum $(-\infty, -m] \cup [m, \infty)$, where m is the particle mass. The negative part of the spectrum was already considered unphysical by Paul Dirac himself: no such negative energies had been observed, and their presence would give rise, for example, to instability of the electron.

This physical inadequacy of the "first-quantized" description quickly led to the introduction of "second quantization," as expressed in quantum field theory. In the quantum field theoretic version of the Dirac theory, the problem of unphysical negative energies is cured by a prescription that goes back to Dirac's hole theory.

Specifically, Dirac postulated that all of the negative energy states of his equation are filled by a sea of unobservable particles. In his picture, annihilating such a negative energy particle with a given charge yields a hole in the sea, which should manifest itself as a new type of positive energy particle with the same mass, but opposite charge. This intuitive idea led Dirac to the prediction that a charged particle should have an oppositely charged partner (its antiparticle).

Ever since, this prediction has been confirmed by experiment, not only for all electrically charged particles, such as the electron and proton, but also for most of the electrically neutral particles, such as the neutron. In the latter case, one still speaks of the particle having a charge, whereas the remaining electrically neutral particles are identical to their antiparticles.

As already mentioned, in the second-quantized Dirac theory no negative energies occur. In the Dirac quantum field, the creation/annihilation operators of negative energy states are replaced by annihilation/creation operators of positive energy holes. This hole theory substitution therefore leads to a physical arena with an arbitrary number of particles and antiparticles with the same positive mass, now called Fock space. (A mathematically precise account can be found in the monograph by Thaller, 1992.)

As it soon turned out, the number of particles and antiparticles in a high-energy collision is not conserved, a phenomenon that can be naturally accommodated in the Fock spaces associated with interacting relativistic quantum field theories. The quantum field theory model that is the most comprehensive description of real-world elementary particle phenomena arose in the early 1970s. During the last few decades, this so-called Standard Model has been abundantly confirmed by experiment.

In spite of these successes, the problem of obtaining nonperturbative insights into the Standard Model remains daunting. This is an important reason why its classical version and various related, but far simpler, classical nonlinear field theories have been, and still are, widely studied. It is a striking and relatively recent finding that within this classical framework, there exist localized, smooth, finite-energy solutions with characteristics that are very reminiscent of particles. The most conspicuous examples in this respect are the soliton field theories, where there exist such particle-like solutions for any given particle number and where the particle numbers and their velocities are preserved in a scattering process.

To be sure, the latter soliton equations arose independently of particle physics. They involve a lower space-time dimension (mostly two), and they have applications in a great many areas that are far removed from high-energy physics. (An early survey that is still one of the best and most comprehensive can be found in Scott et al., 1973.)

Returning briefly to the latter area, there are also various equations, typically within a gauge-theoretic context, where particle-like solutions (instantons, monopoles, vortices, and so on) have been found. A closely related field is classical gravity, where various previously known solutions (such as the Schwarzschild and Kerr black holes) came to be viewed as particle-like solutions, an interpretation strengthened by the occurrence of "many-particle" generalizations.

It should be pointed out that within this nonlinear classical context, there are no explicit many-particle solutions where "creation" and "annihilation" occur. Indeed, it is not even clear how this would manifest itself on the classical level (as compared with the quantum level, where this is a clear-cut matter).

Focusing once again on the notions of charge and particles vs. antiparticles, it is an even more striking fact that these concepts are naturally present within some of the above-mentioned nonlinear field theory models with particle-like solutions. A prime example for gauge theories is given by instanton solutions, which are accompanied by anti-instanton solutions. Roughly speaking, these are distinguished by opposite generalized winding numbers, viewed as charges of a topological nature. More precisely, these localized solutions minimize the energy in certain homotopy classes. This means they are stable under small perturbations.

Turning to soliton field theories, it should be mentioned at the outset that for most soliton equations (for instance, for their most well-known representative, the KdV equation), there exists only one type of soliton, hence no notion of charge. The sine-Gordon field theory is a paradigm for theories where more than one type of soliton occurs. We proceed by using it to exemplify the notions of particle, antiparticle, and charge at the classical level.

The sine-Gordon equation

$$\phi_{xx} - \phi_{tt} = \sin\phi \tag{1}$$

can be obtained as the Euler–Lagrange equation associated with the Lagrangian

$$\mathcal{L} = \tfrac{1}{2}(\phi_t^2 - \phi_x^2) - (1 - \cos\phi). \tag{2}$$

The corresponding energy functional reads

$$E = \int_R \left[\frac{1}{2}(\phi_t^2 + \phi_x^2) + (1 - \cos\phi)\right] \mathrm{d}x. \tag{3}$$

Obviously, the constant solutions

$$\phi_k = 2\pi k, \quad k \in Z, \tag{4}$$

yield $E = 0$. These are the so-called vacuum solutions.

There exist, however, two distinct classes of nonconstant, time-independent, finite-energy solutions, namely,

$$\phi_\pm = 4\arctan(\exp(\pm(x - x_0))), \quad x_0 \in R. \tag{5}$$

They connect the two vacuum solutions ϕ_0 and ϕ_1 for $x \to \pm\infty$. The functions $\exp(\mathrm{i}\phi_\pm(x)) : R \to S^1$ have winding numbers ± 1 and may be viewed as generators of the homotopy group $\pi_1(S^1) = Z$. They are interpreted as one-particle and one-antiparticle solutions (soliton and antisoliton) of the sine-Gordon equation, having charges ± 1. (By Lorentz invariance, they can be boosted to constant velocity v, with $|v|$ smaller than the speed of light, which is 1 for the units chosen in (1).)

The analogy with electrical charge is strengthened by the existence of soliton–antisoliton bound states, the so-called breathers. More generally, there exist solutions with N_+ solitons, N_- antisolitons, and N_0 breathers, where N_+, N_-, N_0 are arbitrary integers. These particle numbers and the velocities are conserved in the collision, the nonlinear interaction showing up only in factorized position shifts.

From the viewpoint of elementary particle physics, these findings are regarded as stepping stones for a better understanding of the associated quantum field theory. In particular, the existence of particle-like solutions stabilized by charges of a topological nature is believed to signal the existence of a corresponding stable quantum particle. A lucid survey in which this scenario is expounded is Coleman (1977).

Returning to the sine-Gordon example (also discussed in Coleman, 1977), the above scenario has not only been confirmed, but considerably enlarged: at the quantum level, the sine-Gordon theory yields a model of interacting solitons and antisolitons, whose scattering preserves particle numbers and other characteristics (such as the set of velocities), yielding an explicitly known factorized S-matrix. Thus, the classical picture is essentially preserved under quantization (cf. the review by Zamolodchikov & Zamolodchikov (1979)). It should be stressed that this absence of particle creation and annihilation is highly nongeneric for relativistic quantum field theories; it occurs for only completely integrable soliton type field theories in two space-time dimensions.

The phenomenon of oppositely charged particles and antiparticles with an attractive interaction between them has also come up in a setting quite different from the above field-theoretic one. Consider the classical Hamiltonian

$$H_N = \frac{1}{2}\sum_{j=1}^{N} p_j^2 + g^2 \sum_{1 \le j < k \le N} 1/\sinh^2(x_j - x_k) \tag{6}$$

on the phase space

$$\Omega = \{(x, p) \in R^{2N} \mid x_N < \cdots < x_1\}. \tag{7}$$

By contrast to the above partial differential equations, most of which are wave equations at face value, the Hamiltonian equations resulting from (6) have, from the outset, a point particle interpretation. To be specific, they describe N nonrelativistic particles on the line with a repulsive pair interaction.

The special character of this Hamiltonian is already manifest from the scattering to which it leads: the particle velocities are conserved and the position shifts are factorized, just as for solitons. Indeed, Hamiltonian (6) is one version of the completely integrable nonrelativistic Calogero–Moser models. (These are surveyed at the classical and quantum levels in Olshanetsky & Perelomov (1981) and Olshanetsky & Perelomov (1983), respectively.) As pointed out first by Calogero, the substitution $x_k \to x_k + \mathrm{i}\pi$, $k = 1, \ldots, N_-$, $N_- \in \{1, \ldots, N-1\}$ has the effect of turning the repulsive $1/\sinh^2$ interaction between particles $1, \ldots, N_-$ and particles $N_- + 1, \ldots, N$ into an attractive $-1/\cosh^2$ interaction. Thus, one obtains an integrable system that can be viewed as describing N_- antiparticles and $N_+ = N - N_-$ particles in interaction.

The nonrelativistic Calogero–Moser systems have been generalized to a relativistic setting, both on the classical and on the quantum levels. Again, there are versions describing particles and antiparticles, with repulsion between like charges and attraction between opposite charges. These versions have been studied at the classical level in Ruijsenaars (1994), where the intimate connection to the sine-Gordon solutions with arbitrary numbers of solitons, antisolitons, and breathers is also detailed. More generally, Ruijsenaars (2001) is a recent survey of this relation between the solitons and antisolitons of the sine-Gordon field theory and the point particles and antiparticles of certain relativistic Calogero–Moser systems, covering both classical and quantum aspects.

SIMON RUIJSENAARS

See also **Higgs boson; Instantons; Quantum field theory; Skyrmions; Solitons, types of; Yang–Mills theory**

Further Reading

Coleman, S. 1977. Classical lumps and their quantum descendants. In *New Phenomena in Subnuclear Physics, Proceedings Erice 1975*, edited by A. Zichichi, New York: Plenum, pp. 297–421

Olshanetsky, M.A. & Perelomov, A.M. 1981. Classical integrable systems related to Lie algebras. *Physics Reports*, 71: 313–400

Olshanetsky, M.A. & Perelomov, A.M. 1983. Quantum integrable systems related to Lie algebras. *Physics Reports*, 94: 313–404

Ruijsenaars, S.N.M. 1994. Action-angle maps and scattering theory for some finite-dimensional integrable systems. II. Solitons, antisolitons, and their bound states. *Publications RIMS Kyoto University*, 30: 865–1008

Ruijsenaars, S.N.M. 2001. Sine-Gordon solitons vs. relativistic Calogero-Moser particles. In *Integrable Structures of Exactly Solvable Two-dimensional Models of Quantum Field Theory*, edited by S. Pakuliak & G. von Gehlen Delft: Kluwer, pp. 273–292

Scott, A.C., Chu, F.Y.F. & McLaughlin, D.W. 1973. The soliton: a new concept in applied science. *Proceedings of the IEEE*, 61: 1443–1483

Thaller, B. 1992. *The Dirac Equation*, Berlin: Springer

Zamolodchikov, A.B. & Zamolodchikov, Al.B. 1979. Factorized S-matrices in two dimensions as the exact solutions of certain relativistic quantum field theory models. *Annals of Physics (NY)*, 120: 253–291

PATTERN FORMATION

Beautiful patterns and complex structures spontaneously emerge in a large variety of natural systems. Examples can be found on all length scales, ranging in size from Jupiter's red spot, to water waves, a zebra's stripes, mussel shell patterns, snowflake structures, and down to the nanoscale patterning of the scaffolding that gives stability to individual cells.

How do patterns form? The basic laws of physics should account for all forces driving pattern formation, but even with knowledge of all equations governing the pattern-forming process, it remains difficult to compute what patterns will emerge. A closer look at how the snowflake pattern develops illustrates the key challenges:

(i) *Patterns form under nonequilibrium conditions.* In order for an initial seed crystal to form, the air generally has to reach temperatures below the dew and freezing point. Once the seed crystal forms, surrounding air is filled with excess water not in thermodynamic equilibrium. Therefore, the crystal grows under non-equilibrium conditions, which means that the shape of the whole crystal may not obey equilibrium thermodynamics principles such as minimization of free energy. Instead of a spherical shape with minimal surface free energy, complex snowflake shapes emerge. This observation is quite general: patterns form while systems are driven out of equilibrium.

(ii) *Patterns can emerge from the competition of processes on various length and/or timescales.* What physical processes govern the pattern-formation process? In the snowflake example, the crystal grows so fast that heat and impurities, rejected from the growing crystal, build up ahead of the crystal-gas interface. They interfere with the growth of the initial spherical crystal and lead to the development of many fine fins. The length scale of the fins is determined by a competition between diffusion, which would lead to many fine fins (for which the diffusion distance becomes shortest) and surface tension, which inhibits very curved interfaces. In general, a variety of physical processes can drive pattern formation including diffusion, chemical reactions, and fluid flow. Patterns emerge from the competition of processes on different length scales and timescales.

(iii) *Patterns are history-dependent.* As a non-equilibrium process, pattern formation is usually history dependent and sensitive to the initial conditions, thus it is difficult to predict patterns accurately. Small differences in temperature or water concentration in the atmosphere are amplified in snowflakes, so each snowflake looks distinct, reflecting the atmospheric conditions along the path it traveled. The theory of pattern formation provides insights into how pattern-forming systems can be analyzed and what one may learn about the underlying physical processes from observations of patterns.

General features of the pattern on the other hand, are often related to underlying symmetries in the system and can be predicted. For example, snowflakes often have six spikes, reflecting the symmetry of the anisotropy in the surface tension of ice.

However, each pattern-forming system may be driven by a unique combination of forces: similarities in the patterns may thus only reflect similar symmetries of the underlying equations, timescales, or length scales, not similarities in the underlying physics.

Driven by new mathematical tools, digital imaging, and powerful image analysis tools, numerous experimental and theoretical investigations of pattern-forming systems have been carried out over the past two decades. Cross & Hohenberg (1993) provide an excellent technical introduction into the mathematical concepts of pattern formation. Ball (1999) introduces the field of pattern formation in a non-mathematical way suitable for a broad audience. An update on the field can be found in Gollub & Langer (1999). Below, we briefly describe the basic principles and mathematical tools.

Model Systems of Pattern Formation

Pattern formation implies a change in symmetry. Spontaneous symmetry breaking is particularly evident in fluids, which are structureless in equilibrium but exhibit a surprising variety of patterns under non-equilibrium conditions. Fluid systems have yielded two model systems for detailed studies of pattern formation: surface waves and convection rolls.

Surface waves develop due to either wind or gravitational forces or, for example, when a stone is thrown into the fluid. Under vertical shaking, fluids can exhibit very ordered patterns, such as hexagonal structures as shown in Figure 1. Michael Faraday was the first to analyze ordered surface wave patterns more than a century ago (Faraday, 1831).

Spontaneous symmetry breaking into a patterned state can also be observed when a fluid is driven out of equilibrium through a thermal gradient. If a fluid is heated from below and cooled from above, the fluid near the bottom expands and rises to the top, while fluid near the top cools, contracts, and due to the increased density, drops back to the bottom. This leads to large-scale convection rolls, the so-called Rayleigh–Bénard convection, where streams of rising and falling fluid arrange into complex patterns, similar to the patterns found in surface waves shown in Figure 1.

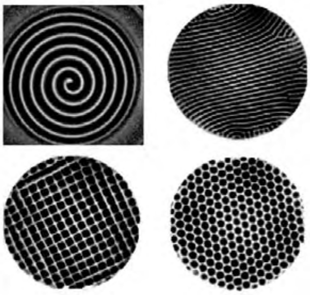

Figure 1. Surface wave patterns in a vertically vibrated fluid. Spirals, stripes, squares, and hexagonal patterns are characteristic of surface waves ("fluid" consists of 170 micron bronze spheres, a few layers deep. Courtesy of D. Goldman, P. Umbanhowar, B. Lewis, and H. Swinney, UT Austin).

While diffusion alone tends to create uniform states, Alan Turing suggested that diffusion coupled with chemical reactions may lead to spatial patterns in chemical composition. He speculated that such mechanisms would be sufficient to explain natural patterns such as zebra stripes or the regular shape and arrangement of leaves. An early experimental realization was the Belousov–Zhabotinsky reaction, which involves several intermediate chemical compounds. The original experiments were carried out in a well-stirred container and generated temporal patterns in which the concentration of intermediate components oscillated. Later experiments showed spatial patterns such as stripes and spirals.

Granular systems such as sand or coffee powder also spontaneously form patterns. Under vertical vibration, they form patterns resembling surface waves (see Figure 1). In addition, localized, stationary waves (so-called oscillons) are observed near the onset of surface waves (Umbanhowar et al., 1996). Unlike localized waves in fluids, oscillons do not spread out or dissipate over time. Mixtures of granular matter are also prone to pattern-forming instabilities. For example, when sand and sugar are mixed in a container and poured into a pile, the mixture spontaneously separates into bands under some conditions. Unlike the fleeting structure of surface waves or convection rolls in fluids, patterns in

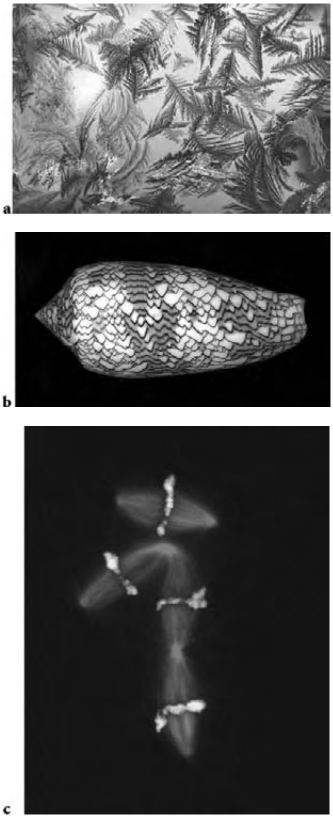

Figure 2. Pattern formation in nature on many scales: (a) Crystal growth pattern (rime on a windshield). (b) Mussel shell pattern (Conus textile, Courtesy of P. Egerton) (c) Scaffolding of a cell during cell division (Spindle assembly in *Xenopus* extracts. Microtubules (long) and DNA (rings), Courtesy of R. Ohi and T. Mitchison, Harvard Medical School, Boston).

granular flows can create lasting impressions such as bands of sugar and sand in a mixture, or sand ripples and sand dunes in wind-driven flows.

In solids, snowflakes are a well-known pattern (see Figure 2a) which we have discussed above. Similar patterns can be found in alloys and polymer blends, though they are hidden within the material micro- or nanostructure. Patterns in solids are technologically important, because they can determine material strength and flexibility.

Biological systems offer a rich variety of patterns. On large scales, patterns such as zebra stripes and mussel shell patterns (Gierer & Meinhard, 1972) (see Figure 2b) help animals blend in or send signals of intimidation or attraction. On a small scale, the microscopic mechanical structure within cells contains ordered arrangements of microtubules and other filaments, which assist, for example, in cell division (see Figure 2c).

Stability and Bifurcations

The first prerequisite for the emergence of patterns from a spatially uniform state is a symmetry-breaking instability. A linear stability analysis indicates whether a given set of equations is prone to a symmetry-breaking instability. For example, we start with a set of functions $\boldsymbol{u}(t)$:

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = f(\boldsymbol{u}). \tag{1}$$

To assess the stability of a state $\boldsymbol{u}_0$, we linearize around that state $\boldsymbol{u} = \boldsymbol{u}_0 + \boldsymbol{\delta}$. This yields an equation for the perturbation $\boldsymbol{\delta}$:

$$\frac{\mathrm{d}\boldsymbol{\delta}}{\mathrm{d}t} = \left.\frac{\mathrm{d}f(\boldsymbol{u})}{\mathrm{d}u}\right|_{\boldsymbol{u}_0} \boldsymbol{\delta}. \tag{2}$$

The state $\boldsymbol{u}_0$ is linearly stable against perturbations if $\boldsymbol{\delta}$ shrinks for any shape of the initial perturbation and unstable if it grows. To test for all possible perturbations, $\boldsymbol{\delta}$ can be decomposed into Fourier modes $\delta = \delta_{\boldsymbol{q}} \exp(\mathrm{i}\boldsymbol{q}x)$. This indicates the range of wave numbers ($\boldsymbol{q}$) that are unstable.

Generally, pattern-forming systems become unstable when they are driven further from equilibrium. The transition point toward instability is called marginal stability. In the marginally stable state, perturbations of one, or a few, wave numbers neither grow nor decay; all other perturbations decay.

When parameters are changed further, the system will become unstable for a range of wave numbers (sometimes including $\boldsymbol{q} = 0$). In the example of crystal growth, a planar crystal growth front is stable for small $\boldsymbol{q}$ (due to diffusion) and for large $\boldsymbol{q}$ (due to surface tension), but unstable for a range of intermediate $\boldsymbol{q}$.

Once the perturbations grow, the linear approximation fails since δ will diverge, and other approaches are needed to predict patterns, but the range of unstable wave number is often a rough indicator of the characteristic length scales of the final patterns.

Model Equations

Beyond linear instability, modeling of pattern-forming systems becomes interesting: rather than going back to the full equations for the pattern-forming system, which are often too complex to be simply studied numerically or analytically, model partial differential equations have been developed to describe the main qualitative features of patterns, and to highlight the universal features of many pattern-forming systems. They are discussed in great detail, for example, in the excellent review by Cross & Hohenberg (1993). Here we list the main equations.

The (complex) Ginzburg–Landau equation is an amplitude equation with an expansion in amplitudes and gradient terms. The equation applied, for example, to a two-dimensional plane can yield similar patterns as the vibrated fluid shown in Figure 1: squares, hexagons, and stripes. The equation provides insight into which gradient or higher-order terms must dominate in specific patterns. Long wavelength instabilities such as the Eckhaus instability are common and limit the range of stable patterns to a smaller range of wave numbers than the range of linearly stable wave numbers.

Other equations include the Swift–Hohenberg equation and the Kuramoto–Sivashinsky equation, which have been solved in various geometries and for different boundary conditions and parameter ranges. They provide a rich set of patterns and dynamical instabilities.

Experimentally and in simulations, the best model system has proven to be surface waves and Rayleigh–Bénard convection. The stability phase diagram of the convecting fluid—the range of stable wave numbers as a function of a measure of fluid convection such as the Rayleigh number—exhibits a region of stability, the so-called Busse balloon, bounded by various instabilities such as the Eckhaus instability, a zig-zag instability, or a cross roll instability. The range of stable patterns and the instability boundaries have been verified through extensive experiments.

Slightly different are equations that describe patterns in excitable media, which are often used to mimic biological processes. Reaction-diffusion equations yield wave fronts or spiral waves. Again, basic symmetries of the underlying processes dictate what patterns can occur, and the stability boundaries are crucial.

Finally, it should be noted that real patterns have boundaries and defects, which can strongly affect pattern selection and stability. Modeling defects and boundaries appropriately can be difficult, since it may only introduce small perturbations that change the system over long timescales (which are often beyond the reach of simulations). Recently, controlled experiments where defects are implanted deliberately into patterns have begun to allow quantitative tests of theories, but this work is only beginning.

Challenges for the future include application of these insights to real pattern forming systems, for example, to suppress or exploit crystal growth instabilities in the semiconductor industry. General theoretical descriptions of pattern formation in networks such as neural networks remains one of the open challenges.

WOLFGANG LOSERT

See also **Belousov–Zhabotinsky reaction; Dune formation; Emergence; Growth patterns; Kuramoto–Sivashinsky equation; Morphogenesis, biological; Rayleigh–Taylor instability; Reaction-diffusion systems; Turing patterns**

Further Reading

Ball, P. 1999. *The Self-made Tapestry: Pattern Formation in Nature*, Oxford and New York: Oxford University Press

Cross, M.C. & Hohenberg, P.C. 1993. Pattern formation out of equilibrium. *Reviews of Modern Physics*, 65: 851–1112

Faraday, M. 1831. On a peculiar class of acoustical figures; and on certain forms assumed by groups of particles upon vibrating elastic surfaces. *Philosophical Transactions of the Royal Society of London*, 52: 299–340

Gierer, A. & Meinhard, H. 1972. A theory of biological pattern formation. *Kybernetik*, 12: 30–39

Gollub, J.P. & Langer, J.S. 1999. Pattern formation in nonequilibrium physics. *Reviews of Modern Physics*, 71: S396–S403.

Turing, A. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B*, 237: 37–72

Umbanhowar, P.B., Melo, F. & Swinney, H.L. 1996. Localized excitations in a vertically vibrated granular layer. *Nature*, 382: 793–796

PATTERN SELECTION

See **Synergetics**

PEIERLS BARRIER

As a localized entity such as an electron, hole, polaron, soliton, or dislocation moves through a discrete lattice, it often experiences a potential barrier between one unit cell and the next. Called a Peierls barrier after the famed German/British physicist Rudolph Peierls, such potential barriers plays a key role in lattice dynamics.

More precisely, a Peierls barrier characterizes the dependence of the energy, $E(\bar{n})$, of a stationary localized entity (LE) on its position, $\bar{n}$, in a periodic lattice. The Peierls barrier, is the function $E(\bar{n}) - E_0$, where $E_0 = \min E(\bar{n})$ is the ground state energy of the LE. Because of lattice periodicity the barrier must be a periodic function with a period equal to the lattice constant a. Minima of the Peierls barrier correspond to stable positions and maxima to unstable positions. The barrier height $\Delta E = \max[E(\bar{n}) - E_0]$ is often called the pinning energy. To propagate along the chain, the LE must be activated with energy exceeding ΔE. For smaller energies, the LE vibrates near minima of the Peierls barrier, and for larger energies it can overcome the Peierls barrier and move through the lattice. The presence of a Peierls barrier also leads to the appearance of new eigenvalues corresponding to periodic oscillations of an LE near a minimum of the Peierls barrier.

The notion of the Peierls barrier goes back to the first quantitative explanation of the mechanism of plastic flow via the motion of dislocations in a lattice (Peierls, 1940; Naborro, 1967). From the perspective of the continuum theory of elasticity, a dislocation is an LE that involves a discontinuity in the displacement field, leading to singularities in strains and stresses. Peierls and Naborro showed how to overcome these artifacts of continuum idealizations by distributing the sharp discontinuity over a finite region, as in the one-dimensional nonlinear continuum Frenkel–Kontorova model of dislocation.

The dynamic influence of a Peierls barrier is greatest when the size of the LE is about equal to the lattice periodicity. Failure to include it in theoretical formulations can lead to erroneous calculations. When modeling the dynamics of protonic solitons in chains of hydrogen bonds, for example, inclusion of the Peierls barrier is necessary to find the temperature dependence of soliton mobility (Savin & Zolotaryuk, 1991).

Let us consider in more detail an autolocalized state of a quantum particle in a lattice (or molecular chain). In a simplified model, the dynamics of such a quasiparticle is described by the system of equations

$$\begin{aligned} \mathrm{i}\hbar\dot{\phi}_n &= \chi u_n \phi_n - J(\phi_{n-1} + \phi_{n+1}), && (1) \\ m\ddot{u}_n &= -Ku_n - \chi|\phi_n|^2, && (2) \end{aligned}$$

where $-\infty < n < +\infty$ is an integer, $\{\phi_n\}$ is a complex-valued wave function normalized by the condition

$$\sum_n |\phi_n|^2 = 1\,, \tag{3}$$

where u_n is the intramolecular displacement of the lattice, and χ, J, m, and K are parameters of the exciton–phonon interaction: energy of resonance or exchange interaction between neighboring lattice sites, reduced mass of the molecules, and stiffness of intermolecular interaction. Neglecting the inertia of intramolecular displacements implies that $u_n = -\chi|\phi_n|^2/K$, whereupon Equations (1) and (2) reduce to a discrete nonlinear Schrödinger equation (DNLS), which can be written as

$$-\mathrm{i}\frac{\mathrm{d}\phi_n}{\mathrm{d}\tau} = \phi_{n-1} + \phi_{n+1} + g\phi_n|\phi_n|^2, \tag{4}$$

where $\tau = Jt/\hbar$ and $g = \chi^2/KJ$ is a dimensionless parameter of nonlinearity.

In the continuum limit, Equation (4) reduces to the nonlinear Schrödinger (NLS) equation, which is integrable. In this limit, the Peierls barrier does not play any role because an envelope soliton can propagate freely along the chain. For the discrete model, which is not integrable, the situation is quite different.

A stationary autolocalized state of the quasiparticle can be found numerically by minimizing

$$H = -\sum_n \varphi_n \varphi_{n+1} + \tfrac{1}{4} g \varphi_n^4,$$

subject to the constraint of Equation (3), where φ_n is the real-valued amplitude of quasiparticle wave function on nth site. If $\{\varphi_n\}_{n=-\infty}^{+\infty}$ is the solution of the minimization problem, the energy and position of the LE are $E = H$ and

$$\bar{n} = \sum_n n\varphi_n^2.$$

respectively.

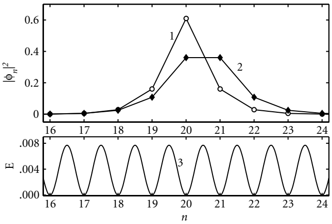

Figure 1. Stationary autolocalized state of the quantum quasiparticle with integer (curve 1) and half-integer (curve 2) centers of symmetry. Curve 3 (below) shows the Peierls barrier, i.e., the dependence of the energy of the stationary state E on the position of the center n. Nonlinear parameter $g = 3$.

Stationary states of quasiparticle have two types of symmetry. First is a solution with integer center of symmetry $\varphi_{n_0+n} \equiv \varphi_{n_0-n}$. In other words, the center of the LE is in one of the lattice sites. Second is a solution with half-integer center of symmetry, where the center of excitation is midway between two lattice sites ($\varphi_{n_0+n} \equiv \varphi_{n_0+1-n}$, $\bar{n} = n_0 + \frac{1}{2}$). The ground state is of the first type and corresponds to a minimum of the Peierls barrier. The second type is unstable and corresponds to a maximum of the Peierls barrier.

Numerical solution of the minimization problem allows one to find the energy profile of a Peierls barrier, as plotted in Figure 1. This profile has a sine-like character with minima at integer points and maxima in half-integer points. The height of the Peierls barrier depends on the value of nonlinear parameter g. If the nonlinearity parameter $g \to 0$, the size of the LE grows without bound, and the pinning energy ΔE goes monotonically to zero. One can calculate Peierls barriers analytically using the method of variational functions (Kuprievich, 1985).

Note that a specific discretization of the continuum NLS equation (Ablowitz & Ladik, 1976) leads to a discrete integrable system. In this case, the ground state energy does not depend on the position of the LE and the Peierls barrier is absent.

LEONID MANEVITCH AND ALEXANDER SAVIN

See also **Bjerrum defects; Discrete nonlinear Schrödinger equations; Dislocations in crystals; Frenkel–Kontorova model; Polarons**

Further Reading

Ablowitz, M.J. & Ladik, J.F. 1976. A nonlinear difference scheme and inverse scattering. *Studies in Applied Mathenatics*, 55(3): 213–229; Nonlinear differential-difference equations and Fourier analysis. *Journal of Mathematical Physics*, 17(6): 1011–1018

Kuprievich, V.A. 1985. Variational study of stationary autolocalized states of molecular chains in discrete model. *Theoretica Materialia Fizika (Soviet Theoretical and Mathematical Physics)*, 64(2): 269–276

Naborro, F.R.N. 1967. *Theory of Crystal Dislocations*, Oxford: Clarendon Press

Peierls, R.E. 1940. *Proceedings of the Physical Society of London*, 52: 34

Savin, A.V. & Zolotaryuk, A.V. 1991. Dynamics of ionic defects and lattice solitons in a thermalized hydrogen-bonded chain. *Physical Review* A, 44: 8167–8183

PENDULUM

With a history covering many centuries, the pendulum has been of primary importance in the development of physics. It is well known that in the 17th century, Dutch experimental physicist Christiaan Huygens discovered the synchronization phenomenon while studying pendulum clocks (Huygens, 1673). Using a pendulum as an example, Huygens formulated the conservation law for mechanical energy. In the 18th century, the French scientist Jean Charles Borda suggested using a pendulum for measuring the acceleration of gravity (Borda, 1792), and Wilhelm Bessel used a pendulum for an accurate verification of the equality of inertial and gravitational masses. There are many other applications of pendula for physical measurements; for example, a pendulum with a randomly vibrated suspension axis has turned out to be a good model for control of turbulence in subsonic submerged jets (Landa, 1996).

Two models of a pendulum are well known—mathematical and physical. The mathematical pendulum is defined as a ball suspended by a weightless thread in vacuum (Figure 1a). It is described by the equation

$$\ddot{\varphi} + \omega_0^2 \sin\varphi = 0, \qquad (1)$$

where φ is the pendulum angle of deviation, $\omega_0 = \sqrt{g/l}$ is the angular frequency of small oscillations, g is the acceleration due to gravity, l is the thread length, and dots indicate derivative with respect to time. The physical pendulum can be defined as a body moving about a fixed horizontal axis O (Figure 1b).

Taking account of air friction, the motion equation for such a body is

$$I\ddot{\varphi} + h\dot{\varphi} + mga\sin\varphi = 0, \qquad (2)$$

where I is the body's moment of inertia, h is the friction factor, m is the body's mass, and a is the distance between the axis and the body's center of gravity.

Equation (1) belongs to the class of nonlinear conservative (Hamiltonian) systems, whereas Equation (2) is a dissipative system. Solutions of these equations can be conveniently analyzed with the phase plane φ, $\dot{\varphi}$. We illustrate this for Equation (1).

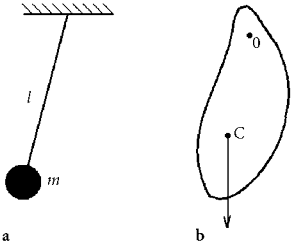

Figure 1. Schematic image of (a) mathematical pendulum and (b) physical pendulum.

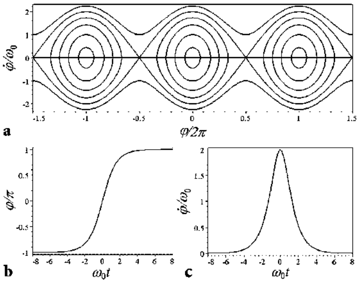

Figure 2. (a) Phase portrait of pendulum oscillations in the phase plane; (b,c) the time dependencies of the pendulum phase coordinates φ and $\dot{\varphi}$ at the motion of a representative point along a separatrix.

Using the energy conservation law, we obtain an equation for a trajectory in the phase plane φ, $\dot{\varphi}$ in the form

$$\frac{ml^2\dot{\varphi}^2}{2} + mgl(1-\cos\varphi) = E, \qquad (3)$$

where E is the total pendulum energy. For $E < 2mgl$, Equation (3) describes closed trajectories corresponding to pendulum oscillations about the stable equilibrium position ($\varphi = 0$), and for $E > 2mgl$, nonclosed trajectories corresponding to the rotation of the pendulum (Figure 2a). These two kinds of trajectories are separated by peculiar trajectories passing through singular saddle points ($\dot{\varphi} = 0$, $\varphi = \pm\pi$, $\pm 3\pi$, ...). Such trajectories are said to be *separatrices.* Because values of φ differing from each other by 2π are physically indistinguishable, the phase plane shown in Figure 2a can be rolled into a cylinder.

In general, solutions of Equations (1) and (3) are Jacobi elliptic functions, but more simple solutions can be obtained in the case of small oscillations ($E \ll mgl$) and in the case when $E = 2mgl$ (this value of E

corresponds to the motion of a representative point along a separatrix). In the latter case, Equation (3) is split into two equations:

$$\dot{\varphi}_{\pm}(t) = \pm 2\omega_0 \cos\frac{\varphi}{2}. \qquad (4)$$

By integrating (4) we find

$$\varphi_{\pm}(t) = \pm\left(4\arctan \mathrm{e}^{\omega_0 t} - \pi\right). \qquad (5)$$

The time dependencies of $\varphi = \varphi_{+}$ and $\dot{\varphi} / \omega_0 = \dot{\varphi}_{+} / \omega_0$ are shown in Figure 2b and c. These solutions play an important part in soliton theory.

In the case of sufficiently small oscillations, when $\sin\varphi \approx \varphi - \varphi^3 / 6$, Equation (1) reduces to the Duffing equation

$$\ddot{\varphi} + \omega_0^2\left(1 - \frac{\varphi^2}{6}\right)\varphi = 0. \qquad (6)$$

The general solution of Equation (6) is

$$\varphi = A\,\mathrm{sn}(\Omega t, k), \qquad (7)$$

where sn is a Jacobi elliptic function of modulus $k = \sqrt{1/12}\,A / \Omega$, and $\Omega = \omega_0\sqrt{1 - A^2/12}$. Solution (7) is valid for $A \le \sqrt{6}$, when $k \le 1$. This constraint is certainly fulfilled because Equation (6) follows from (1) only for $A < 1$. Solution (7) describes periodic oscillations of period $4\mathrm{K}(k) / \Omega$, where $\mathrm{K}(k)$ is the full elliptic integral of the first kind. It follows from this that the period of oscillations increases with increasing amplitude. Such a property inherent in nonlinear systems is called anisochronism. It should be noted that Galileo was the first to discover the isochronism of small pendulum oscillations.

If a pendulum consisting of an iron ball suspended by a thread of length l is placed between the opposite poles of a magnet, its behavior essentially changes (Landa, 1996, 2001). Approximating the magnetic force acting on the ball by $F(\varphi) = ml(a_1\varphi - b_1\varphi^3)$ and restricting ourselves to small oscillations of the ball, then the pendulum angular deviation φ obeys the following approximate equation:

$$\ddot{\varphi} - (a - b\varphi^2)\varphi = 0, \qquad (8)$$

where $a = a_1 - \omega_0^2$, $b = b_1 - \omega_0^2 / 6$, $\omega_0^2 = g / l$. In the case that $a_1 > \omega_0^2$ $(a > 0)$, the equilibrium position $x = 0$ becomes aperiodically unstable (the corresponding singular point $\varphi = 0$, $\dot{\varphi} = 0$ in the phase plane φ, $\dot{\varphi}$ becomes of saddle type). If, in addition to this, the inequality $b_1 > \omega_0^2 / 6$ holds, two stable equilibrium positions with coordinates $\varphi_{1,2} = \pm\sqrt{a / b}$ appear. These equilibrium positions correspond to singular points of center type. But if $b_1 < \omega_0^2 / 6$ and $a > 0$, then the ball adheres to one of the magnet poles. Equation (8) describes a so-called two-well oscillator, which is the subject of recent widespread interest in connection with stochastic and vibrational resonances (Landa, 2001).

If a pendulum is suspended from a uniformly rotating shaft, it can execute self-oscillations. Such a pendulum was discovered by William Froude and mentioned in the famous treatise by Lord Rayleigh (Rayleigh, 1877). Rayleigh showed that oscillations of such a pendulum are approximately described by an equation which came to be known as the Rayleigh equation. A controlled Froude pendulum was suggested by Neimark to be a model of stochastic oscillations (Neimark & Landa, 1992).

POLINA LANDA

See also **Damped-driven anharmonic oscillator; Duffing equation; Elliptic functions; Hamiltonian systems; Solitons**

Further Reading

Borda, J. 1792. *Mémoires sur la mesure du pendule* (*Mesure de la méridienne*), Paris

Huygens, C. 1673. *Christiani Hvgenii Zvlichemii Horologivm oscillatorivm, sive, De motv pendvlorvm ad horologia aptato demonstrationes geometricæ*. Paris: Muguet; as *Christiaan Huygens' The Pendulum Clock, or, Geometrical Demonstrations Concerning the Motion of Pendula as Applied to Clocks*, translated with notes by R.J. Blackwell, Ames: Iowa State University Press, 1986

Landa, P.S. 1996. *Nonlinear Oscillations and Waves in Dynamical Systems*, Dordrecht and Boston: Kluwer

Landa, P.S. 2001. *Regular and Chaotic Oscillations*, Berlin and New York: Springer

Neimark, Yu.I. & Landa, P.S. 1992 *Stochastic and Chaotic Oscillations*, Dordrecht and Boston: Kluwer (original Russian edition 1987)

Rayleigh, Lord (Strutt, J.W.) 1877–78. *The Theory of Sound*, London: Macmillan; reprinted New York: Dover, 1945

PERCEPTRON

The perceptron is one of the earliest computational models inspired by the human brain. Developed by Frank Rosenblatt from 1959–1962, the perceptron model is a parallel, distributed processing network intended to provide a means of modeling the capabilities and properties of the brain at a very simplified level (Rosenblatt, 1962).

Artificial neurons as mathematical processing elements of neurological (or artificial neural) networks were described earlier by Warren McCulloch & Walter Pitts (1943). Also earlier was the work of Donald Hebb, which presented a principal of learning or self-organization in neural networks (Hebb, 1949). The perceptron brought these concepts together by providing an algorithm for adjusting the parameters or weights of the network to learn to perform mappings or predicates. This model was implemented in analog electrical hardware by Rosenblatt and colleagues as the Mark I Perceptron and as such was the first neurocomputer to perform useful functions (Block, 1962). Accepting that biological brains possess "natural" intelligence, developing

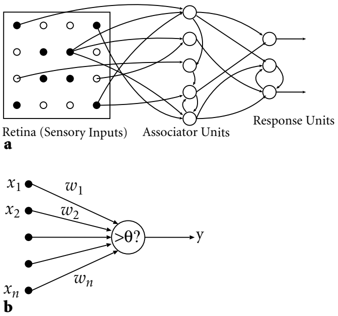

Figure 1. (a) General perceptron model. (b) Single perceptron unit.

computational models inspired by the brain represents one approach toward artificial intelligence.

The general perceptron model consists of a retina, associator units, and response units (see Figure 1a). The retina provides sensory inputs connected (many-to-many and at random) to a layer of associator units. The associator outputs are connected to response units. Connectivity can be sparse, and both associator and response units can be connected to other units in the same layer. An associator unit becomes active if its total input exceeds some threshold value and propagates a signal to the response units. This model appears to have the essential features to begin studying brain function (learning, pattern recognition, memory, generalization) in terms of brain structure.

A perceptron unit computes a threshold function of the values presented via its input connections

$$y = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} w_i x_i > \theta, \\ 0 & \text{otherwise,} \end{cases} \qquad (1)$$

where x_i is the value of the ith input signal. Each input connection is weighted by a real-valued weight w_i, and θ determines the threshold value. Geometrically, the perceptron unit implements a hyperplane in n-dimensional input space, where input vectors lying on one side of the line will result in an output of 1, and input vectors on the other side of the line will result in an output of 0. The position of the hyperplane is determined by the values of the weights and thresholds (zero threshold implies a hyperplane passing through the origin).

The simple perceptron unit provides a useful, simplified realization of the above model which can be implemented and trained to perform useful functions (Figure 1b). A set of inputs $X = \{x_1, \ldots, x_n\}$ are provided to a layer of associator units that compute a fixed value $\varphi(X)$. These values are then combined at the response unit to produce a binary output $\Psi(X)$. The response unit is adaptive in that there is an adjustable weight parameter α_i associated with each of its input connections. The output of the response unit Ψ is calculated as

$$\Psi = 1 \text{ if } \sum_{i=1}^{n} \alpha_i \varphi_i(X) > \theta, \qquad (2)$$

$$\Psi = 0 \qquad \text{otherwise,} \qquad (3)$$

where θ is a threshold value. The simple perceptron can compute any linear threshold function of the $\varphi_i(X)$ given appropriate values of the weights α_i and θ.

Given a set of data points in n-dimensional input space, labeled with desired (0 or 1) output values, Rosenblatt provided a rule for iteratively learning (modifying) the perceptron weights and thresholds, such that the outputs of the perceptron would match the desired output values. He also proved the convergence of the learning rule in a finite number of steps, provided that the perceptron has the capacity (ability) to learn an adequate linear threshold function. That is, if the n-dimensional (0,1)-labeled data is separable by an $(n-1)$-dimensional hyperplane in the input space, the perceptron can learn or be trained to find such a hyperplane.

An adaptive threshold value can be conveniently implemented by replacing θ with an additional input x_0, that always takes the value 1, and an adaptive weight w_0. Given input values and the desired (correct) value d for the output, the output of the model is computed from the inputs and the weights are adapted as follows:

$$\alpha_i = \begin{cases} \alpha_i & \text{if } d = y(X), \\ \alpha_i + x_i & \text{if } d = 1 \text{ and } y(X) = 0, \\ \alpha_i - x_i & \text{if } d = 0 \text{ and } y(X) = n1. \end{cases} \qquad (4)$$

If the threshold function is changed to a continuous, differentiable function, it becomes possible to formulate a cost function for learning that is a differentiable function of the weight values. For example, a sigmoidal function is commonly used to approximate the threshold function. The delta rule can then be used to train the unit via an iterative gradient descent (Widrow & Hoff, 1960). Widrow and Hoff developed the delta (or Least Mean Squared) rule in the context of ADALINEs (ADAptive LInear NEurons), which are equivalent to McCulloch–Pitts neuron models with the bias weight (Widrow & Lehr, 1990). A quadratic error function resulting from the sum of squared error terms is used, making it possible to guarantee convergence of the delta rule under appropriate assumptions. In contrast, the perceptron learning rule does not converge if the data is nonlinearly separable.

In general, a perceptron may consist of a layer of k of the above perceptron units, each of which is fully connected to same set of inputs but has its own output. Both Rosenblatt's perceptron learning rule and the delta rule can be straightforwardly extended to train such a network.

The inability of single-layer perceptron to implement nonlinearly separable mappings and the lack of a procedure for training a multilayer perceptron model led to a lull in progress in the field for many years (Minsky & Papert, 1990). Recently, the back-propagation algorithm led to a resurgence of interest in the multilayer perceptrons and artificial neural network models. Such models have been very successfully applied to a wide range of data-driven learning tasks such as pattern recognition, classification, regression, and prediction (Haykin, 1999).

Marcus Gallagher

See also **Artificial intelligence; Attractor neural network; Cell assemblies; Cellular automata; Cellular nonlinear networks; McCulloch–Pitts network; Neural network models**

Further Reading

Block, H.D. 1962. The perceptron: a model for brain functioning. I. *Reviews of Modern Physics*, 34(1): 123–135

Haykin, S. 1999. *Neural Networks A Comprehensive Foundation*, 2nd edition, Upper Saddle River, NJ: Prentice-Hall

Hebb, D.O. 1949. *The Organization of Behavior*, New York: Wiley

McCulloch, W.S. & Pitts, W. 1943. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5: 115–133

Minsky, M. & Papert, S.A. 1990. *Perceptrons: An Introduction to Computational Geometry*, Cambridge, MA: MIT Press

Rosenblatt, F. 1962. *Principles of Neurodynamics*, New York: Spartan

Widrow, B. & Hoff, M.E., Jr. 1960. Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, Part 4, New York: IRE, pp. 96–104

Widrow, B. & Lehr, M.A. 1990. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78: 1415–1442

PERCOLATION THEORY

For random site percolation, each site of a large lattice is randomly occupied with probability p and empty with probability $1 - p$. A cluster is a set of neighboring occupied sites. A spanning cluster extends from one end of the sample to the opposite end. Percolation theory (PT) deals with the number and properties of spanning and nonspanning clusters as a function of p and of the cluster size s: the number of occupied sites in the cluster. Alternative variants are bond percolation where all sites are occupied but bonds between neighbor sites exist only with probability p. A cluster then is a set of sites connected directly or indirectly by existing bonds. One may also remove the lattice structure and place possibly overlapping circles on a piece of paper (or spheres in three dimensions). Instead of random variables, one can also study correlated percolation, such as Ising models where up spins are identified as occupied and down spins as empty sites. The following results are expected to be also valid for these variants as long as the correlations have a finite spatial extent (i.e., not at the Curie point of Ising models) (Stauffer & Aharony, 1994).

In 1941, Paul Flory published the first percolation theory (ignoring cyclic links and excluded volume effects). Physicists call this the Bethe lattice or mean field approximation, and it becomes exact for very high dimensions but is inaccurate in its critical behavior for the two or three dimensions we are mostly interested in. Percolation theory was continued by Stockmayer (1943), applied to immunology at the beginning of the 1950s, and a few years later, put onto square lattices by Broadbent & Hammersley (1957). The first computer simulations were published around 1960, and in 2001, these reached 4 million times 4 million sites (Tiggemann, 2001) and even larger lattices in Maloney & Pruessner (2003), using the Hoshen–Kopelman algorithm (Stauffer & Aharony, 1994). Exact polynomials for the number $n_s(p)$ of clusters of s sites were found for small and intermediate s, and exact results for many thresholds p_c and critical exponents are known in two dimensions.

For small p, we have only small clusters, and for p near unity, nearly the whole system forms a huge spanning cluster coexisting with rare small clusters. Thus, there is a sharp critical threshold p_c, where with increasing p for the first time, a spanning cluster appears in a very large lattice. For random site percolation on the triangular lattice with six neighbors or for random bond percolation on the square lattice, we have $p_c = \frac{1}{2}$, while for the square site problem, numerical estimates give $p_c \simeq 0.5927462\ldots$ and $p_c \simeq 0.31161\ldots$ for simple-cubic random site percolation. (Dimensions much higher than two and three were recently investigated by Grassberger (2003). Random percolation in one dimension has $p_c = 1$.) Only in infinite lattices are these thresholds defined as infinitely sharp.

Quantities of interest besides n_s are the probability $R(p)$ to have a spanning cluster, the fraction P_∞ of sites in the infinite (spanning) cluster, the mean cluster size

$$S = \frac{\sum_s n_s s^2}{\sum_s n_s} \quad \text{or} \quad \sum_s n_s s^2 , \tag{1}$$

the average cluster radius R_s, and the characteristic length ξ defined best through $\xi^2 = \sum_s R_s^2 n_s s^2 / \sum_s n_s s^2$ among other properties. (For $p > p_c$ in these sums, the spanning cluster is omitted.) Figure 1 shows S versus p, for random site percolation on nearest-neighbor simple-cubic lattices. From scaling laws or renormalization group arguments, the critical phenomena for large clusters and small $p - p_c$ are

$$\begin{aligned} &n_s \propto s^{-\tau} f[(p - p_c)s^{\sigma}] , \\ &P_\infty \propto (p - p_c)^{\beta} , \\ &S \propto |p - p_c|^{-\gamma} , \\ &\xi \propto |p - p_c|^{-\nu} , \\ &R_s(p = p_c) \propto s^{\sigma\nu} \end{aligned} \tag{2}$$

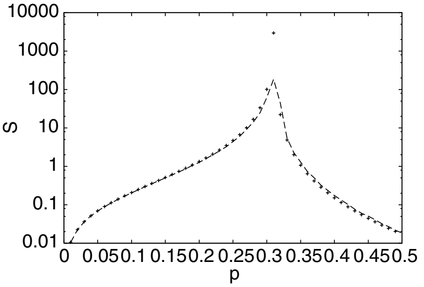

Figure 1. Mean cluster size from Equation (1) for seven nearest-neighbor simple-cubic lattices lattices $1000 \times 1000 \times 1000$ (+) compared with $100 \times 100 \times 100$ (line).

with the critical exponents given by Greek letters.

The first of these equations (Stauffer & Aharony, 1994) leads to Fisher's scaling laws $\beta = (\tau - 2)/\sigma$, $\gamma = (3 - \tau)/\sigma$, while additional hyperscaling assumptions valid for dimensionality d below six give $d\nu = \gamma + 2\beta = (\tau - 1)/\sigma$. The fractal dimensionality D of clusters at the percolation threshold, defined (Bunde & Havlin, 1996) through $s \propto R_s^D$, is then

$$D = 1/(\sigma\nu) = d - \beta/\nu = d/(\tau - 1)\,. \qquad (3)$$

Because an infinite cluster first appears for increasing p at $p = p_c$, it is plausible that for p slightly below p_c, we see finite but very large clusters; thus, the mean cluster size S diverges for $p \to p_c$ in an infinite system. In a finite system, S remains finite but the position of its maximum moves towards p_c and its height towards infinity, for system size $\to \infty$ (see Figure 1). The relation of its critical exponent γ with σ and τ can be easily derived from the above assumption for n_s by approximating the sum for S through an integral over s.

Flory's approximate solution is valid for the critical exponents only above six dimensions: $\beta = \gamma = 2\nu = 2\sigma = 1$, $\tau = \frac{5}{2}$, $D = 4$; the above scaling function f is then a Gaussian. For the more relevant case of $d = 2$, we have $\beta = 5/36$, $\gamma = 43/18$, while in three dimensions $\beta \simeq 0.42$, $\gamma \simeq 1.79$ with the other exponents following from the above scaling laws. The scaling function f is no longer a Gaussian but has a maximum below p_c; the cluster numbers n_s away from p_c decay as $\log(n_s) \propto -s$ above p_c and as $\propto -s^{1-1/d}$ below p_c. For $p < p_c$, the fractal dimension in three dimensions is $D = 2$, one of the rare cases where an exact solution was found in three but not in two dimensions.

Besides computer simulation, small-cell renormalization group methods also give reasonable approximations for thresholds and critical exponents. One divides the infinite lattice into cells of linear dimension b such that the cell centers themselves again form a lattice of the same type as the original lattice. A cell is occupied if the original occupied sites form a cluster spanning this cell. In this way, the original occupation probability p is renormalized into the cell occupation probability $p' = R(p)$. One may iterate this renormalization by computing $p'' = R(p')$ and then $R(p'')$ and so on. The fixed point p^* of this transformation obeys $p^* = R(p^*)$ and approaches, for large b, the true percolation threshold p_c. At this fixed point, the derivative $\mathrm{d}R(p)/\mathrm{d}p$ is related to the critical exponent ν. However, in contrast to many renormalization group publications, $R(p_c)$ is not equal to p_c (Ziff, 1992).

Some mathematical theorists in the 1980s claimed that the infinite cluster is unique, but this is not true at $p = p_c$ where several spanning clusters can coexist (Aizenman, 1997). Even above p_c in very elongated rectangles, one may find several clusters spanning in the short direction.

How do these theoretical and numerical results compare with applications and reality? Because the percolation thresholds p_c depend on the lattice structure, one should not expect them to agree with off-lattice experiments, but the exponents are more "universal" and should agree when comparing reality with PT. There are many applications (Sahimi, 1994) including flow in porous media, conductor-insulator mixtures, elasticity of composite materials, and breakdown of the internet. Sputtering metal drops onto a plane offers a particularly accurate realization of percolation.

Flory invented the percolation theory to describe the sol-to-gel transition of branched polymers, known from boiling your breakfast egg. But only half a century later was it experimentally confirmed that the gelation exponents are indeed those of three-dimensional lattice percolation theory.

Another application of percolation clusters is to speed up computer simulations of Ising models (ferromagnets) near their Curie point, using the clusters of correlated bond percolation formed by up spins connected by bonds with probability $p = 1 - \exp(-2J/k_BT)$. Flipping whole clusters instead of single spins equilibrates the system much faster. (In this Fortuin–Kasteleyn–Coniglio–Klein–Swendsen–Wang theory (Stauffer & Aharony, 1994), T is the absolute temperature, k_B the Boltzmann constant, and J the interaction energy such that $2J$ is the energy to change an isolated pair of parallel spins into antiparallel.) For example, right at the Curie point in a two-dimensional $L \times L$ lattice, the relaxation time no longer increases as $L^{2.17}$ but only as $\log(L)$ if these percolation clusters, instead of single spins, are flipped.

The above application stems from the relation between percolation clusters and the q-state Potts model ($q = 2$ is the Ising model again). The partition function of this Potts model is $Z = \langle q^N \rangle$, where N is the total number of bond percolation clusters at the above $p = 1 - \exp(-2J/k_B T)$ and $\langle \cdots \rangle$ denotes the average over many realizations. Taking the limit $q \to 1$, we get $\ln(Z)/\ln(q) \to \langle N \rangle = \sum_s n_s$, recovering random percolation which is thus a $q \to 1$ Potts model.

Percolation is also a convenient teaching example. Filling a small triangular lattice at $p = p_c = \frac{1}{2}$ by throwing coins reveals in the classroom fractals, statistical errors, and (in the resulting number of isolated sites) strong systematic errors from the lattice boundaries.

Dietrich Stauffer

See also **Cluster coagulation; Polymerization; Sandpile model**

Further Reading

Aizenman, M. 1997. On the number of incipient spanning clusters. *Nuclear Physics* B, 485: 551–582

Broadbent, S.R. & Hammersley, J.M. 1957. Percolation processes. *Proceedings of the Cambridge Philosophical Society*, 53: 629–645

Bunde, A. & Havlin, S. 1996. *Fractals and Disordered Systems*, Berlin: Springer

Flory, P.J. 1941. Thermodynamics of high polymer solutions. *Journal of the American Chemical Society*, 63: 3083–3100

Grassberger, P. 2003. Critical percolation in high dimensions. *Physical Review* E, 67: 036101

Maloney, N.R. & Pruessner, G. 2003. Asynchronously parallelized percolation on distributed machines. *Physical Review* E, 67: 037701

Sahimi, M. 1994. *Applications of Percolation Theory*, London: Taylor & Francis

Stauffer, D. & Aharony, A. 1994. *Introduction to Percolation Theory*, London: Taylor & Francis

Stockmayer, W.H. 1943. Theory of molecular size distribution and gel formation in branched-chain polymers. *Journal of Chemical Physics*, 11: 45–55

Tiggemann, D. 2001. Simulation of percolation on massively-parallel computers. *International Journal of Modern Physics* C, 12: 871–878

Ziff, R.M. 1992. Spanning probability in 2D percolation. *Physical Review Letters*, 69: 2670–2673

PERIOD DOUBLING

Like many terms used in the nonlinear sciences, *period doubling* has more than one meaning. Well known is the response of a system at half the driving frequency, due to nonlinear coupling. Probably the earliest observation of period doubling was by Michael Faraday, in his investigations of shallow water waves (Faraday, 1831, arts. 98–101). The phenomenon was also investigated by Lord Rayleigh (Rayleigh, 1887). He wrote of an example:

> in which a fine string is maintained in transverse vibration by connecting one of its extremities with the vibrating prong of a massive tuning-fork, *the direction of motion of the point of attachment being parallel to the length of the string* the string may settle down into a state of permanent and vigorous vibration *whose period is double that of the point of attachment.*

A more everyday example is the fact that a child may set a swing into transverse motion by standing on the seat and moving up and down at twice the natural frequency. Such phenomena are in the province of *harmonic generation* and *parametric amplification* and are not treated in this entry.

Period doubling (as discussed in this entry) is the most common of several *routes to chaos* for a nonlinear dynamical system.

The dynamics of natural processes, and the nonlinear equations used to model them, depend on externally set conditions, such as environmental or physical factors. These take fixed values over the development of the system in any particular instance, but vary from instance to instance. Expressed as numerical quantities, such factors are called parameters, and their role is vital. Often, increasing a parameter increases the nonlinearity. The simplest example is the logistic model (May, 1976)

$$x_{n+1} = r x_n (1 - x_n), \qquad (1)$$

a discrete system with state variable x_n and parameter r; n is the generation. (If x represents population, then r characterizes the underlying growth rate.)

For period doubling, it is sufficient to vary a single parameter. Over some range, the system may have a periodic attractor, that is, a periodic orbit that attracts neighboring orbits. Typically, the stability (attraction) decreases as the parameter increases, changing to instability (repulsion) at a critical value. This is a bifurcation, or change of structure, of the orbit. In period doubling, this change is accompanied by the "birth" of a new attracting period-doubled orbit, in which the system alternates between two states. In biological systems, these are known as alternans (Glass & Mackey, 1988). In the period-doubling route to chaos, each new periodic attractor loses its stability with increasing parameter value, whereupon the next period-doubled attractor is born. If the original attractor was a fixed point, this generates orbits of period 1, 2, 2^2, 2^3, …, called the main period-doubling cascade.

A surprising *universality* was discovered by Mitchell Feigenbaum. In part, it relates to the sequence of parameter values at which successive period doubling occurs. Label the first by r_1, the second by r_2, and so on, and let the cascade end at the value r_∞. The differences $(r_\infty - r_n)$ decrease to zero; the surprising fact is that the ratios $(r_\infty - r_n)/(r_\infty - r_{n+1})$ converge to a universal constant δ, which takes the same value $\delta \approx 4.669202\ldots$ for all maps with a quadratic maximum (Feigenbaum, 1978). There is a second universal constant, $\alpha \approx 2.502908\ldots$, measuring the relative spatial scale of the orbits, which also becomes increasingly fine.

	Logistic		Hénon	
n	r_n	$\left(\frac{r_{n+1}-r_n}{r_{n+2}-r_{n+1}}\right)$	a_n	$\left(\frac{a_{n+1}-a_n}{a_{n+2}-a_{n+1}}\right)$
1	3.0000000000		0.3675000000	
2	3.4494897428		0.9125000000	
3	3.5440903596	4.751446	1.0258554050	4.807887
4	3.5644072661	4.656251	1.0511256620	4.485724
5	3.5687594195	4.668242	1.0565637582	4.646894
6	3.5696916098	4.668739	1.0577308396	4.659569

Table 1. Estimates of the Feigenbaum constant δ, for the logistic and Hénon maps.

Experiment		Observed number of period doublings	estimated value of δ	estimated value of α
Hydrodynamic:	helium	4	3.5 ± 1.5	
	mercury	4	4.4 ± 0.1	
Electronic:	diode	5	4.3 ± 0.1	2.4 ± 0.1
	Josephson	4	4.5 ± 0.3	2.7 ± 0.2
Laser:	feedback	3	4.3 ± 0.3	
Acoustic:	helium	3	4.8 ± 0.6	

Table 2. Selected experimental data on period doubling.

Numerical data for two main period-doubling sequences is given in Table 1. Columns 2–3 display data for the logistic map,

$$x_{n+1} = 1 - ax_n^2 + y_n, \qquad y_{n+1} = bx_n. \qquad (2)$$

columns 4–5 for its two-dimensional cousin, the Hénon map. For Table 1, the parameter b has been kept constant. Experimental data has also been obtained in quite a few systems (Cvitanović, 1989); a selection is shown in Table 2. To appreciate the experimental difficulty involved, remember that for each successive period doubling, the significance of errors increases five-fold while the complexity of the dynamics doubles.

The mechanism for period doubling is already implicit in the fact that it is also known as a "flip bifurcation." Stability of a fixed point x^* of a smooth one-dimensional map f is the simplest case for theory. Consider a nearby point x that maps to x'. Linear approximation gives

$$(x' - x^*) \approx f'(x^*) \cdot (x - x^*). \qquad (3)$$

This shows (and exact analysis confirms) that the fixed point is stable if $-1 < f'(x^*) < 1$ and unstable if $f'(x^*) < -1$ or $f'(x^*) > 1$. Successive iterates flip from side to side for negative $f'(x^*)$, so period doubling occurs at $f'(x^*) = -1$. For higher-dimensional systems, stability is determined by the eigenvalues of a matrix of partial derivatives; the occurrence of an eigenvalue $\lambda = -1$ leads to period doubling. Because $x_{n+2}^{**} = x_n^{**}$ for the new orbit, period doubling is connected with double iteration, controlled by the second composition map $f_2(x) = f(f(x))$. Elementary calculus shows that if f satisfies the two conditions

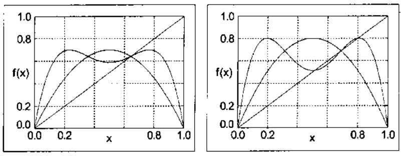

Figure 1. Mechanism: f and f_2 for $r = 2.8$ (left) and $r = 3.2$ (right).

$$f(x^*) = x^*, \qquad f'(x^*) = -1, \qquad (4)$$

then f_2 satisfies three conditions:

$$f_2(x^*) = x^*, \quad f_2'(x^*) = +1, \quad f_2''(x^*) = 0. \qquad (5)$$

As a result, $x - f_2(x)$ must be approximated by a cubic near the bifurcation. One of its zeros is the (now unstable) fixed point of f, the other two constitute the period-doubled orbit, because they are not fixed points of f.

In the case of the logistic map, the first period doubling occurs at $r = 3$. Graphs of f and f_2 show the mechanism (Figure 1). As the cascade proceeds, the pictures and the analysis repeat but with increasing complication. The common thread is that, at each step, conditions (4) for a composition f_n imply conditions (5) for f_{2n}.

BRIAN DAVIES

See also **Attractors; Bifurcations; Dripping faucet; Harmonic generation; Hénon map; Maps; One-dimensional maps; Parametric amplification; Routes to chaos; Stability; Universality**

Further Reading

Cvitanović, P. 1989. *Universality in Chaos*, 2nd edition, Bristol: Adam Hilger

Davies, B. 1999. *Exploring Chaos: Theory and Experiment*, Reading, MA: Perseus Books

Faraday, M. 1831. On a peculiar class of acoustical figures; and on certain forms assumed by groups of particles upon vibrating elastic surfaces. *Philosophical Transactions of the Royal Society of London*, 121: 299–340

Feigenbaum, M.J. 1978. Quantitative universality for a class of nonlinear transformations. *Journal of Statistical Physics*, 19: 25–52

Glass, L. & Mackey, M.C. 1988. *From Chaos to Clocks: The Rhythms of Life*, Princeton, NJ: Princeton University Press

May, R.M. 1976. Simple mathematical models with very complicated dynamics. *Nature*, 261: 459–469

Rayleigh, Lord. 1887. Maintenance of vibrations by forces of double frequency, and propagation of waves through a medium with a periodic structure. *Philosophical Magazine*, 24: 145–159

Schuster, H.G. 1988. *Deterministic Chaos: An Introduction*, 2nd edition, Weinheim: VCH

PERIODIC BURSTING

Bursting systems show intervals of repetitive activity separated by intervals of relative quiescence, and have at least two different time scales; fast for the short period with spike-like oscillations that comprise the burst and slow for the longer period of the burst itself. Laboratory systems that exhibit bursting include the Belousov–Zhabotinsky reaction and neuronal systems. A simple conceptual model would be the bursting that is generated by switching between the active and inactive states, suggesting the coexistence of stable periodic and quiescent states. Bursting can be irregular, in which the periods between bursts are effectively random, or it can be strictly periodic. Some experimental bursting systems exhibit switching between different patterned periodic behaviors, suggesting two or more coexisting periodic states, and models have been shown to exhibit bi- or multi-rhythmicity. The exotic periodic dynamics of bursting systems—with patterning, multistability, and multi-rhythmicity provide a demonstrator for bifurcation and continuation algorithm packages.

A simple two-variable excitable or oscillatory system such as the FitzHugh–Nagumo equations can be driven to bursting by a slow periodic forcing term, so bursting can be viewed as the activity of a fast oscillatory system being modulated by a slower oscillatory system. For an autonomous bursting system, the fast and slow systems are coupled, and the burst is generated by the evolution of the slow system sweeping the dynamics of the fast subsystem between steady state and oscillatory dynamics (Chay & Rinzel, 1985). On the timescale of the fast system, the slow variable acts as a control parameter, so the fast variable approximates its attractors — its stable equilibrium during the period of quiescence and its oscillations during activity.

Bursting in neurons and other electrically excitable secretory cells may be part of their normal pattern generation behavior, as in the generation of the respiratory rhythm, or a sign of pathology, as in the abnormal bursting charges during epileptic fit. Synchronization of bursting activity between widely separate cells in the cortex excited by different features of the same visual input has been observed and proposed to underlay visual binding. Neural central pattern generators are networks of nerve cells, that generate periodic bursts as a result of their interconnections, but some single isolated cells can generate bursts. Both network and isolated cell bursting can be modeled by systems of differential equations. Specific models for isolated cells—the beta-pancreatic cell or the parabolic burster cell R15 of Aplysia—are ordinary differential systems, analogous to the Hodgkin–Huxley equations for the action potential, based on experimental measurements of currents and concentrations (Rinzel & Lee, 1986). They reproduce the observed bursting behaviors and can be reduced to simpler systems in which the dynamical mechanisms producing bursting are obtained by bifurcation analysis of these models.

In these models, the spiking is produced by oscillations of a fast subsystem that is modulated by the slower system. Bifurcation and topological analysis of the models has led to a classification of the types of bursting (Bertram et al., 1995). In type I, the fast subsystem is bistable, and the burst begins at a saddle node and ends at a homoclinic bifurcation. During the burst, the period between spikes increases illustrated by a simple three-variable polynomial model that retains the qualitative features of neuronal bursting, the Hindmarsh–Rose equations (1984) in Figure 1. The stable (thick solid line) and unstable (dotted line) Z-shaped steady state curves exchange stability at a Hopf bifurcation and a saddle-node bifurcation, and the periodic solutions that emerge at the Hopf bifurcation end at a homoclinic point. The xy subsystem has a bistable range, with stable steady state and limit cycle. During the burst, z is approximately constrained to this bistable range, each spike increments z, and the period between spikes increases as the homoclinic point is approached. Between bursts, z decreases slowly close to the lower, stable limb of the steady-state curve, as x increases until the orbit is re-injected into the spiking region.

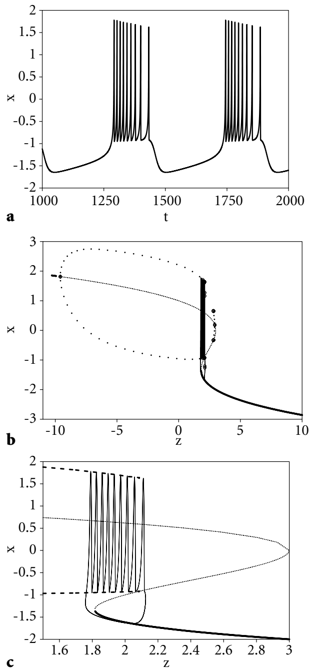

Figure 1. Example of type I burst generation in the three-variable (x, y, z) Hindmarsh & Rose (1984) model for neuronal oscillations, bursting and chaos. (a) Periodic bursting in the fast x variable representing membrane potential; (b) and (c) with z scale magnified: Bifurcation diagram for the fast $x - y$ subsystem with the slow variable z treated as a bifurcation parameter, and the periodic bursting of (a) is overlaid in the $x - z$ plane.

In type II bursting, two slow variables drive the fast system across a homoclinic bifurcation, so the spike period within the burst decreases and then increases parabolically, as the slow variables drive the system toward, and away from, the homoclinic point. This is illustrated by the parabolic burster R15 of Aplysia. The fast system has only a stable periodic limit cycle during the burst.

In type III bursting, as in type I, the fast system is bistable, but the burst ends at a saddle node rather than a homoclinic bifurcation, and the spike period need not increase during the burst.

Examples of all three types are seen in different bursting cells, and many bursting cells show more exotic, and chaotic, behaviors. An alternative approach is to reduce the models to interval maps, and to classify the bifurcations of these maps—see Fan & Holden (1993).

Arun V. Holden

See also **FitzHugh–Nagumo equation; Hodgkin–Huxley equations; Nerve impulses**

Further Reading

Bertram, R., Butte, M.J., Kiemel, T. & Sherman, A. 1995. Toplogical and phenomenological classification of bursting oscillations. *Bulletin of Mathematical Biology*, 57: 413–439

Chay, T.R. & Rinzel, J. 1985. Bursting, beating and chaos in an excitable embrane model. *Biophysical Journal*, 47: 357–366

Fan, Y-S. & Holden, A.V. 1993. Bifurcations, burstings, chaos and crises in the Rose–Hindmarsh model for neuronal activity. *Chaos Solitons and Fractals*, 3: 439–449

Hindmarsh, J.L. & Rose, R.M. 1984. A model for neuronal bursting using three coupled first order differential equations. *Proceedings of the Royal Society of London* B, 221: 87–112

Rinzel, J. & Lee, Y.S. 1986. On different mechanisms for membrane potential bursting. In *Nonlinear Oscillations in Biology and Chemistry*, edited by H.G. Othmer, New York: Springer, pp. 19–33

PERIODIC ORBIT THEORY

Many characteristic quantities for chaotic systems, such as Lyapunov exponents, correlation functions, or fractal dimensions, are defined as averages over the invariant measure (defined as a probability distribution on the phase space). For most chaotic systems, this measure is difficult to represent analytically, so it has to be approximated experimentally or numerically by averages over long trajectories. The periodic orbit theory rests on the idea that periodic orbits provide a skeleton that allows approximation of the invariant measure and that, therefore, ergodic averages can be determined from suitably weighted averages over periodic orbits. It combines mathematically established analytical expressions with well-controlled numerical input for highly accurate quantitative information about chaotic systems. It provides access to Lyapunov exponents, fractal dimensions, escape rates, diffusion constants, resonances, or even correlations between eigenvalues of quantum systems (Eckmann & Ruelle, 1985; Artuso et al., 1990; Chaos Focus Issue, 1992).

In order to motivate the basic expressions and to determine the weight of periodic orbits, consider the evolution of a density $\rho_n(\boldsymbol{x})$ at time n under the action of a discrete dynamical system, $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n)$ (Cvitanović & Eckhardt, 1991). After one time step, the density is given by

$$\rho_{n+1}(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{y}\, U(\boldsymbol{x}, \boldsymbol{y})\rho_n(\boldsymbol{y}) \qquad (1)$$

with the evolution kernel $U(\boldsymbol{x}, \boldsymbol{y})$. In the deterministic map, the points $\boldsymbol{y}$ at time n and $\boldsymbol{x}$ at $n+1$ are connected by a single trajectory, so that U is a δ-distribution, $U = \delta(\boldsymbol{x} - \boldsymbol{F}(\boldsymbol{y}))$. In the presence of noise, the evolution kernel will be more regular and some of the mathematical subtleties that appear when dealing with singular operators will disappear. Noise regularization can also be used to define the "natural" invariant measure for a dynamical system, the Sinai–Ruelle–Bowen measure, as the measure obtained in the limit of vanishing noise (Eckmann & Ruelle, 1985). For an attractor, this invariant measure is the eigenvector to the eigenvalue 1 of the evolution operator.

Periodic orbits appear when traces of U are calculated: the trace requires that $\boldsymbol{x}$ and $\boldsymbol{y}$ coincide, and this singles out a periodic orbit. With p the label for the different periodic orbits, n_p their primitive period (the number of iterations need to first return to the starting point), and $J_p = \mathrm{d}\boldsymbol{x}(n_p) \,/\, \mathrm{d}\boldsymbol{x}(0)$ the Jacobian after one period, the trace for m iterations becomes

$$\begin{aligned} \operatorname{tr} U^m &= \int \mathrm{d}\boldsymbol{x}\, U^m(\boldsymbol{x}, \boldsymbol{x}) \\ &= \sum_p \sum_{r=1}^{\infty} \frac{n_p}{|\det(1 - J_p^r)|} \delta(m - r n_p); \quad (2) \end{aligned}$$

powers of U are defined as multiple mappings so that U^m is the evolution operator after m time steps. The first sum extends over the primitive, nonrepeated orbits p and the second over their multiple traversals r.

The Fredholm determinant for the evolution operator U can be related to traces of the evolution operator via

$$\begin{aligned} \det(1 - zU) &= \exp(\operatorname{tr} \ln(1 - zU)) \quad (3) \\ &= \exp\left(- \sum_{n=1}^{\infty} \frac{z^n}{n} \operatorname{tr} U^n \right) = \sum_{l=0}^{\infty} c_l z^l. \quad (4) \end{aligned}$$

The right-hand side is then an expansion in terms of periodic orbits, first through the traces in the exponent and then through their contributions to the coefficients c_l in the final power series. The left-hand side can formally be expanded in terms of eigenvalues λ_ν of the operator, $\det(1 - zF) = \prod_\nu (1 - z\lambda_\nu)$. It vanishes whenever z equals the inverse of an eigenvalue.

The formal manipulations that lead to (4) can be made precise in the context of nice hyperbolic systems in which periodic orbits are dense ("Axiom A"). Simple maps and flows on surfaces of negative curvature belong to this class. Then the theory can be developed as discussed in Ruelle (1989) and Pollicott (2002). In particular, the eigenvalues with sufficiently small damping can be given an interpretation as experimentally accessible long-lived Ruelle–Pollicott resonances.

For the calculation of averages, the relation between the leading eigenvalue and the invariant density is important. If the average of a given phase space observable A is to be calculated, one can extend the transfer operator U to include the observable, $\tilde{U} = U \exp(-qA)$, and extract the phase space average from the variation of the leading, zero $z_0(q)$ of the Fredholm determinant with q,

$$\langle A \rangle = \frac{\mathrm{d}}{\mathrm{d}q} \ln z_0(q) \bigg|_{q=0}. \quad (5)$$

The contribution to the trace is a term $\exp(q \sum_P A(\boldsymbol{x}_P))$ with the sum extending over all points P of the orbit p: it reflects the sampling of phase space along the orbit.

The dependence of the Fredholm determinant on periodic orbits and their properties can be spelled out completely, so that it can be written as a product over contributions from periodic orbits. For the case of a one-dimensional map, we find

$$\begin{aligned} \det(1 - zU) &= \prod_p \prod_{j=0}^{\infty} \left(1 - \frac{z^{n_p}}{|J_p| J_p^j} \right) \\ &= \prod_j \zeta_j^{-1}(z). \quad (6) \end{aligned}$$

The ζ_j are called dynamical zeta functions in analogy to the Riemann zeta function $\zeta_{\mathrm{R}}(s)$, which by the Euler product representation can be expressed as a product over prime numbers,

$$\zeta_{\mathrm{R}}^{-1}(s) = \left(\sum_{n=1}^{\infty} n^{-s} \right)^{-1} = \prod_p (1 - p^{-s}). \quad (7)$$

The analytic properties of the Fredholm determinant determine the decay of the coefficients c_l in the power series expansion in (4): if the Fredholm determinant is analytic then the coefficients fall off faster than exponentially (Rugh, 1992). This is particularly useful in numerical calculations where highly accurate and quickly converging results can be obtained using the shortest orbits.

For continuous flows, the discrete period z^{n_p} is replaced by the continuous equivalent $\exp sT_p$, and the contributions of observables becomes an integral along the orbit.

When the classical evolution operator is replaced by the quantum one, a relation between the quantum eigenvalues and classical periodic orbits emerges (Gutzwiller, 1990). For instance, the semiclassical periodic orbit expression for the density of states in a chaotic system is

$$\begin{aligned} \rho(E) &= \sum_i \delta(E - E_i) \\ &\sim \operatorname{Re} \frac{1}{\pi\hbar} \sum_p \sum_{r=1}^{\infty} \frac{T_p \exp(-\mathrm{i}\mu_p \pi r/2)}{\sqrt{|\det(1 - J_p^r)|}} \mathrm{e}^{\mathrm{i} r S_p(E)/\hbar}, \quad (8) \end{aligned}$$

where the symbol $\sim$ indicates the omission of a few regularizing terms, and where S_p and μ_p are the classical action and the Maslov index, respectively. This Gutzwiller trace formula is at the heart of many developments in the field of quantum chaos.

Zeta functions for the counting of periodic orbits already appear in Smale (1967). The mathematical theory was developed prominently by David Ruelle, Rufus Bowen, and Yasha Sinai (see Eckmann & Ruelle, 1985; Ruelle, 1989; Pollicott, 2002). Practical aspects for calculating with periodic orbits are discussed by Artuso et al. (1990), by Gaspard (1998), and by many contributions to the Chaos Focus Issue (1992). Applications of this formalism to simple maps and the period-doubling attractor are given by Artuso et al. (1990), to diffusion in Lorentz gases in Chaos Focus Issue (1992), and to the dimension and resonances to the Lorenz model in Eckhardt & Ott (1994). The relation to quantum mechanics is discussed in Gaspard (1998), Gutzwiller (1990), Eckhardt (1993), and the Chaos Focus Issue (1992). As demonstrated by Flepp et al. (1991), a comparison between periodic orbits extracted from experimental data and from a numerical model can be used to fine tune and improve the model.

BRUNO ECKHARDT

See also **Phase space; Quantum chaos**

Further Reading

Artuso, R., Aurell, E. & Cvitanović, P. 1990. Recycling of strange sets: I. Cycle expansion; II. Applications. *Nonlinearity*, 3: 325–360 and 361–386

Chaos Focus Issue on Periodic Orbit Theory. 1992. *Chaos*, 2: 1–158

Cvitanović, P. & Eckhardt, B. 1991. Periodic orbit expansions for classical smooth flows. *Journal of Physics* A, 24: L237–L241

Eckhardt, B. 1993. Periodic orbit theory. *Proceedings of the International School of Physics "Enrico Fermi," Course CXIX, Quantum Chaos*, edited by G. Casati, I. Guarneri & U. Smilansky, Amsterdam: North-Holland, pp. 77–118

Eckhardt, B. & Ott, G. 1994. Periodic orbit analysis of the Lorenz attractor. *Zeitschrift für Physik* B, 94: 259–266

Eckmann, P. & Ruelle, D. 1985. Ergodic theory of chaotic systems. *Reviews of Modern Physics*, 57: 617–656 and 1115 (Addendum)

Flepp, L., Holzner, T., Brun, E., Finardi, M. & Badii, R. 1991. Model identification by periodic-orbit analysis for NMR-laser chaos. *Physical Review Letters*, 67: 2244–2247.

Gaspard, P. 1998. *Chaos, Scattering and Statistical Mechanics*, Cambridge and New York: Cambridge University Press

Gutzwiller, M.C. 1990. *Chaos in Classical and Quantum Mechanics*, New York: Springer

Pollicott, M. 2002. Periodic orbits and zeta functions. In *Handbook of Dynamical Systems*, vol 1A, edited by B. Hasselblatt & A. Katok, Amsterdam: Elsevier, pp. 409–452

Ruelle, D. 1989. *Elements of Differentiable Dynamical Systems and Bifurcation Theory*, New York: Academic Press

Rugh, H.H. 1992. The correlation spectrum for hyperbolic analytic maps. *Nonlinearity*, 5: 1237–1263

Smale, S. 1967. Differentiable dynamical systems. *Bulletin of the American Mathematical Society*, 73: 747–817

PERIODIC SPECTRAL THEORY

The direct problem of periodic spectral theory is that of constructing the spectral data of certain linear operators with periodic coefficients, that is, the determination of the spectrum of this operator and of the associated eigenfunctions. The inverse problem of periodic spectral theory is the problem of the reconstruction of such an operator (and thus its coefficients) from given spectral data. Although these questions can be asked for linear partial differential operators, this article focuses on linear ordinary differential operators.

The history of periodic spectral theory starts with the investigations of Sturm and Liouville on the eigenvalues of certain differential equations of second-order with given boundary conditions, now referred to as Sturm–Liouville theory. In 1836–1837, Charles François Sturm and Joseph Liouville examined independently different aspects of this problem, such as the asymptotics of eigenvalues, different comparison theorems on the solutions of similar equations with different coefficients, and theorems on the zeros of eigenfunctions. For the class of equations Sturm and Liouville considered, these results imply the existence of an infinite sequence of real, increasing eigenvalues, and orthogonality of eigenfunctions corresponding to different eigenvalues. Although their investigations did not as such deal with periodic spectral theory, many of their results carry over to this area.

Consider an ordinary differential operator of order n,

$$\begin{aligned} L &= q_n(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n} + q_{n-2}(x)\frac{\mathrm{d}^{n-2}}{\mathrm{d}x^{n-2}} + \cdots \\ &\quad + q_1(x)\frac{\mathrm{d}}{\mathrm{d}x} + q_0(x), \end{aligned} \tag{1}$$

where the coefficients $q_j(x)$, $j=0,\ldots,n$ are periodic functions of x, sharing a common period: $q_j(x+T)=q_j(x)$, $j=0,\ldots,n$, and $q_{n-1}(x)=0$. They are referred to as potentials. Using this operator L, define the differential equation

$$L\psi = \lambda\psi. \tag{2}$$

The direct periodic spectral problem is the problem of (i) determining the set of all $\lambda \in \mathbb{C}$ for which this differential equation has at least one bounded solution, and (ii) for each such λ, determining all bounded solutions. There are many technical issues to be dealt with: which function space do the potentials belong to? Which function space does ψ belong to? These issues will be ignored here. Sometimes one restricts attention to periodic solutions ψ: $\psi(x+T)=\psi(x)$, or antiperiodic solutions $\psi(x+T)=-\psi(x)$. These and other choices lead to spectra that are subsets of the spectrum as obtained without making these choices.

One approach to solve the direct spectral problem is Floquet theory (Amann, 1990). Rewrite Equation (2) as a first-order linear system:

$$\frac{\mathrm{d}\boldsymbol{\psi}}{\mathrm{d}x} = \boldsymbol{X}(x,\lambda)\boldsymbol{\psi}, \quad \boldsymbol{X}(x+T,\lambda) = \boldsymbol{X}(x,\lambda) \qquad (3)$$

with $\boldsymbol{\psi}_1 = \psi(x)$. Note that from $q_{n-1}(x) = 0$, it follows that $\operatorname{tr}\boldsymbol{X}(x,\lambda) = 0$. Thus, the flow determined by (3) is volume preserving. Define the monodromy matrix of this system as $\boldsymbol{M}(x_0,\lambda) = \boldsymbol{\Psi}(x_0+T, x_0, \lambda)$, where $\boldsymbol{\Psi}(x, x_0, \lambda)$ is a fundamental matrix of (3) such that $\boldsymbol{\Psi}(x_0, x_0, \lambda)$ is the identity matrix. Thus, $\boldsymbol{M}(x_0,\lambda)$ is the operator of translating x by T: $\boldsymbol{M}(x_0,\lambda)\boldsymbol{\psi}(x) = \boldsymbol{\psi}(x+T)$. Note that $\boldsymbol{\psi}(x)$ also depends on x_0 and λ, but this dependence is suppressed here. This operation commutes with $\mathrm{d}/\mathrm{d}x$, since $\boldsymbol{X}(x,\lambda)$ is periodic in x with period T. Thus, (3) has a set of solutions $\boldsymbol{\phi}(x)$ which are also eigenvectors of $\boldsymbol{M}(x_0,\lambda)$. These solutions are known as Bloch functions or Floquet functions. If the eigenvalue of $\boldsymbol{M}(x_0,\lambda)$ for any Bloch function has magnitude greater than one, then this Bloch function is unbounded as $x \to +\infty$ or $x \to -\infty$. Thus, the spectrum of (3) is the set of all λ such that at least one eigenvalue of $\boldsymbol{M}(x_0,\lambda)$ has magnitude one. This spectrum is independent of the choice of x_0 due to the requirement that the flow is volume preserving; that is, $\operatorname{tr}\boldsymbol{X}(x,\lambda) = 0$, or $q_{n-1}(x) = 0$.

An important class of periodic spectral problems is that of self-adjoint operators. These are operators whose spectrum is contained on the real line. For self-adjoint operators, the spectrum of the periodic spectral problem consists of the union of a (possibly infinite) sequence of intervals.

For the sake of explicitness, the remainder of this article will discuss the equation (Magnus & Winkler, 1979)

$$-\psi'' + q(x)\psi = \lambda\psi, \quad q(x+T) = q(x). \qquad (4)$$

Note that many of the results stated below are true for general classes of periodic spectral problems. Depending on the literature source, (4) goes by the name of Hill's equation or (after rescaling) the time-independent Schrödinger equation. This equation is self-adjoint. Its spectrum is bounded from below. It is a collection of intervals such that the length of the separating gaps between intervals $\to 0$ as $\lambda \to \infty$. Using Floquet theory, the condition for λ to be in the spectrum is found to be $|\operatorname{tr}\boldsymbol{M}(x_0,\lambda)| \leq 2$. The endpoints of the intervals are given by $|\operatorname{tr}\boldsymbol{M}(x_0,\lambda)| = 2$. Because (4) is a second-order equation, there are two linearly independent Bloch functions.

The time-independent Schrödinger equation (4) of course plays a fundamental role in quantum mechanics. In this case, $q(x)$ is the potential of the system, and λ plays the role of energy. The context of solid state physics is especially relevant here, because the potential $q(x)$ is periodic. The intervals constituting the spectrum are known as allowed (energy) bands and the gaps between them as forbidden (energy) bands.

The inverse periodic spectral problem for (4) is that of the reconstruction of $q(x)$, given a collection of spectral data. Various choices are possible for the collection of spectral data. In general, the inverse problem does not have a unique solution, using the knowledge of one spectrum. This can be resolved by also providing the eigenfunction. However, now the collection of spectral data is unnecessarily large. It is sufficient to provide two spectra (corresponding to different boundary conditions). Together with the known analyticity properties of the eigenfunction $\psi(x)$, this determines the potential $q(x)$. This is similar to the inverse scattering method where the knowledge of the scattering data and the analyticity properties of the eigenfunction determine the potential (decaying as $|x| \to \infty$) uniquely. A major difference is that in the inverse scattering method, the starting point is the asymptotic behavior as $x \to \pm\infty$. This behavior is simple, because it is governed by a differential equation with constant coefficients. This is one reason why the inverse scattering method is as efficient as it is. In the periodic problem, the role of $x \to \pm\infty$ is taken over by $x = x_0$, but there is no simple asymptotic behavior, resulting in a theory which is more technical and less explicit (Dubrovin et al., 1976; Novikov et al., 1984).

This lack of explicitness for solving the inverse periodic spectral problem is to some extent resolved by the consideration of so-called finite-gap potentials. These are potentials for which the number of intervals constituting the spectrum, and thus, the number of gaps separating these intervals, is finite. The simplest nontrivial example is that of the Lamé equation

$$-\psi'' + n(n+1)\wp(x-x_c)\psi = \lambda\psi. \qquad (5)$$

Here $\wp(x-x_c)$ is the Weierstrass elliptic function, x_c is a fixed complex number, and n is a positive integer. In this case, the number of gaps separating the intervals in the spectrum is n.

This classical example is a special case of a more recent theory of finite-gap potentials, whose development started with the works of Novikov (1974) and Lax (1975). They show that the stationary solutions of the nth member of the KdV hierarchy are n-gap potentials of (4). Here the KdV hierarchy is the collection of equations of the form $u_t = \partial_x(\delta H_n/\delta u)$, where H_n is any conserved quantity of the KdV equation, $\delta H_n/\delta u$ denotes the variational derivative of H_n with respect to u (*See* **Poisson brackets**), and indices denote differentiation. It was soon thereafter shown that all finite-gap potentials of (4) were of this nature. For example, the Lamé potential with $n = 1$ is a stationary solution of the KdV equation. This gives a nonspectral characterization of the finite-gap potentials.

To solve the direct spectral problem with an n-gap potential $q_n(x)$, one first considers the direct periodic spectral problem, as stated above. It is solved using

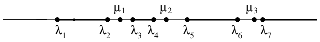

Figure 1. The main spectrum for a three-gap potential (thick solid line) and the auxiliary spectrum μ_1, μ_2 and μ_3.

Floquet theory. The outcome is the main spectrum, consisting of n finite intervals and one infinite interval. The endpoints of these intervals are labeled λ_1, λ_2, etc., in increasing order, as shown in Figure 1. At the endpoints, only one of the eigenfunctions is bounded. These eigenfunctions are periodic with period T or $2T$. There are an infinite number of isolated eigenvalues inside the interval of infinite length for which there are two bounded, periodic eigenfunctions. Next, one considers the Dirichlet problem

$$\begin{cases} -\psi'' + q_n(x)\psi = \lambda\psi, \\ \psi(x_0) = 0, \quad \psi(x_0+T) = 0. \end{cases} \tag{6}$$

The spectrum of this problem is referred to as the auxiliary spectrum. It is discrete, and its points $\mu_k(x_0)$, $k=1,2,\dots$ depend on x_0. All but n of its points lie inside the infinite interval of the main spectrum. Each remaining point lies in a different gap of the main spectrum. This is illustrated in Figure 1. The information contained in the main and auxiliary spectra determines the eigenfunction $\psi(x)$: it is a meromorphic function in the finite λ plane with zeros at $\lambda = \mu_k(x_0)$ and poles at $\lambda = \mu_k(x) = \mu_k(x_0)|_{x_0=x}$ (Dubrovin et al., 1976; Novikov et al., 1984).

Using the main and auxiliary spectra, the inverse periodic spectral problem is solved by (Novikov et al., 1984)

$$q_n(x) = \sum_{j=1}^{2n+1} \lambda_j - 2\sum_{j=1}^{n} \mu_j(x). \tag{7}$$

This is the first of the so-called trace formulae. Other trace formulae give relationships between the potential $q_n(x)$ and its derivatives and the main and auxiliary spectra.

The proposed solution of the inverse periodic spectral problem for finite-gap potentials is not effective. It requires the solution of the direct spectral problem for all x_0 in a period of the potential in order to obtain the auxiliary spectrum as a function of x_0. It is possible to avoid this by determining $\mu_k(x_0)$, $k=1,\dots,n$ as a solution of a set of differential equations (Novikov et al., 1984)

$$\frac{\mathrm{d}\mu_j}{\mathrm{d}x_0} = \frac{\pm 2\mathrm{i}\sqrt{\prod_{k=1}^{2n+1}(\mu_j - \lambda_k)}}{\prod_{k\neq j}^{n}(\mu_j - \mu_k)}, \quad j = 1,\dots,n. \tag{8}$$

The choice of sign gives the direction in which $\mu_j(x_0)$ is going in between its two endpoints.

Another approach, which solves system (8), is the use of abelian functions and Riemann surfaces (Dubrovin, 1981; Belokolos, et al., 1994). An abelian function of n variables is a $2n$-periodic function. As such, abelian functions generalize elliptic functions to more than one variable. All abelian functions are expressible as ratios of homogeneous polynomials of Riemann's theta function. All finite-gap potentials of (4) are abelian functions. For example, the Lamé potentials in (5) are elliptic functions, which are special cases of abelian functions.

In the context of this method, the Bloch function $\phi(x) = \boldsymbol{\phi}_1(x)$ is often referred to as the Baker–Akhiezer function. One of the major results of the theory is the realization that the two Bloch or Baker–Akhiezer functions, regarded as a function of arbitrary complex λ, are distinct branches of one single-valued Baker–Akhiezer function, defined on a two-sheeted Riemann surface covering the complex λ plane (Krichever, 1989). This Riemann surface is already apparent in (8). It is

$$\eta^2 = \prod_{k=1}^{2n+1} (\lambda - \lambda_k), \tag{9}$$

which defines η as a double-valued function of λ. This surface has genus n. It is obtained from Figure 1 by choosing the intervals of the main spectrum as branch cuts and gluing the two resulting sheets together along these cuts. This Riemann surface defines a theta function $\theta(z_1,\dots,z_n|\boldsymbol{B})$ through its normalized period (or Riemann) matrix $\boldsymbol{B}$. In terms of this theta function

$$q_n(x) = c - 2\frac{\mathrm{d}^2}{\mathrm{d}x^2} \ln \theta(k_1 x + \varphi_1, \dots, k_n x + \varphi_n|\boldsymbol{B}). \tag{10}$$

The wave numbers $k_1,\dots,k_n$ are determined as integrals of certain differentials on the Riemann surface (9) with a pole singularity at $\lambda = \infty$. The phase constants $\varphi_1,\dots,\varphi_n$ are determined by the Riemann constants on (9) and the Abel transform. The constant c is determined by a differential on (9) with a double pole singularity at $\lambda = \infty$. Thus, (10) gives an explicit form for all finite-gap potentials of (4), providing a complete solution of the inverse periodic spectral problem.

Some remarks:

(i) The emphasis on finite-gap potentials is justified in the sense that any T-periodic function is approximated arbitrarily well by an infinite sequence of finite-gap potentials with period T and increasing number of gaps (Dubrovin, 1981).

(ii) The periodic spectral problem (4) is the first half of the Lax pair for the Korteweg–de Vries (KdV) equation. As such, it allows the solution of the KdV equation with periodic initial data. The full solution of this problem requires the solution of both the direct and the inverse periodic spectral problems. The schematics is identical to that of the inverse scattering method. First, the initial condition $u(x,t=0) = U(x)$ is used as a potential in (4) to solve the direct periodic spectral problem. This results in the main and auxiliary spectra. The time evolution of these spectra is implied from the second half of the Lax pair: the main spectrum is independent of time, whereas the auxiliary spectrum evolves according to differential equations similar to (8). The spectral data for any time is used to solve the inverse periodic spectral problem of (4). This gives the solution $u(x,t)$ of the KdV equation such that $u(x,t)|_{t=0} = U(x)$ (Novikov et al., 1984).

(iii) The spectral theory for the time-dependent Schrödinger equation is intimately connected to the initial-value problem for the Kadomtsev–Petviashvili equation. A solution of the inverse periodic spectral problem using Riemann's theta function and Riemann surfaces also exists here. However, here there are no restrictions on the form of Riemann surfaces that appear: all compact, connected Riemann surfaces arise. Hence, the periodic spectral theory of the time-dependent Schrödinger equation has important consequences for the theory of Riemann surfaces. It has provided, for instance, a solution to the Schottky problem, which was posed in 1903 (Novikov et al., 1984; Dubrovin, 1981).

(iv)The equation

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + (a - 2k\cos(2x))y = 0 \qquad (11)$$

is Mathieu's equation. It arises from the three-dimensional Helmholtz equation by separation of variables using elliptical coordinates. It is a special case of (4), using a trigonometric potential. One is only interested in period solutions, resulting in a discrete subset of the main spectrum. The periodic solutions of this equation are referred to as Mathieu functions.

BERNARD DECONINCK

See also **Inverse scattering method or transform; Kadomtsev–Petviashvili equation; Korteweg–de Vries equation; Theta functions**

Further Reading

Amann, H. 1990. *Ordinary Differential Eequations: An Introduction to Nonlinear Analysis*, Berlin and New York: de Gruyter

Belokolos, E.D., Bobenko, A.I., Enol'skii, V.Z., Its, A.R. & Matveev, V.B. 1994. *Algebro-geometric Approach to Nonlinear Integrable Equations*, Berlin and New York: Springer

Dubrovin, B.A. 1981. Theta functions and non–linear equations. *Russian Mathematical Surveys*, 36: 11–92

Dubrovin, B.A., Matveev, V.B. & Novikov, S.P. 1976. Nonlinear equations of Korteweg–de Vries type, finite-zone linear operators, and Abelian varieties. *Russian Mathematical Surveys*, 31: 59–146

Krichever, I.M. 1989. Spectral theory of two-dimensional periodic operators and its applications. *Russian Mathematical Surveys*, 44: 145–225. (The introduction is a great overview of the development of periodic spectral theory.)

Lax, P.D. 1975. Periodic solutions of the KdV equation. *Communications on Pure and Applied Mathematics*, 28: 141–188

Magnus, W. & Winkler, S. 1979. *Hill's Equation*, New York: Dover and London: Constable

Novikov, S.P. 1974. The periodic problem for the Korteweg–de Vries equation. *Functional Analysis and Its Applications*, 8: 54–66

Novikov, S.P., Manakov, S.V., Pitaevskii, L.P. & Zakharov, V.E. 1984. *Theory of Solitons: The Inverse Scattering Method*, New York: Consultants Bureau (original Russian edition 1980)

PERMANENT WAVE

See **Wave of translation**

PERTURBATION THEORY

A limited number of nonlinear partial differential equations (PDEs) can be solved analytically for arbitrary initial conditions. Specific applications often lead to nonlinear PDEs that do not fall into the category that can be solved by means of the inverse scattering theory, Bäcklund transformations, or Hirota's bilinear method (Dodd et al., 1984). Examples of such PDEs with physical relevance include soliton equations with added perturbative terms describing energy gain and losses, where the original unperturbed problem is energy-conserving and hence Hamiltonian. Other problems involve nonconservative PDEs, such as nerve fibers and other reaction-diffusion type equations in general. In fluid mechanics, boundary layers possess a special challenge with singular perturbations requiring matching solutions close to a boundary and solutions far from a boundary layer. Perturbations can be additive or parametric (multiplicative), and they can be localized, periodic, quasi-periodic, or random, depending on the nature of the problem to be solved. In the case of a PDE that cannot be solved analytically, we resort to approximate analytical solutions or direct numerical analysis. The approximate solutions can be determined by various perturbation techniques. For soliton equations one procedure is to calculate the variation of the spectral data in the inverse scattering method due to external perturbations (Karpmann, 1979; Kivshar, & Malomed, 1989). Another method introduces slow variation of the parameters entering the soliton solution. Variants of the latter idea include multiple scales, energy methods, and the Lindsted–Poincaré technique. Here we shall illustrate the multiscale method applied to a perturbed nonlinear Schrödinger (NLS) equation. The method follows the ideas in Kaup (1990) with an extension presented in Nguyen et al. (1995). A general perturbed NLS equation is

$$\mathrm{i}\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + 2|u|^2 u = \varepsilon R(t,u). \qquad (1)$$

The complex function $u = u(x, t)$ depends on the spatial coordinate x and time t. The absolute value of the perturbation parameter ε is assumed to be much smaller than unity, and we expect that the perturbation εR on the right-hand side only slightly modifies soliton solutions of the pure NLS equation. We introduce multiple scales of the time variable according to

$$T_0 = t, \quad T_1 = \varepsilon t. \tag{2}$$

In the simplest case, we treat the single soliton solution and invoke the solution ansatz

$$u = q \exp(\mathrm{i}\xi(\Theta - \Theta_0) + \mathrm{i}(\sigma - \sigma_0)), \tag{3}$$

$$q = \eta \operatorname{sech}(\eta(\Theta - \Theta_0)), \quad \text{where} \quad \Theta = x - 2X(T_0, T_1). \tag{4}$$

The functions X and σ are defined by $\partial X/\partial T_0 = -2\xi(T_1)$ and $\sigma = (\eta^2(T_1) + \xi^2(T_1))T_0$. To proceed, we expand q according to $q = q_0 + \varepsilon(\phi + \mathrm{i}\psi)$ and insert into Equation (1), finding

$$Lq_0 = (\partial_{\Theta\Theta} + 2q_0^2 - \eta^2)q_0 = 0, \tag{5}$$

$$M\phi = (\partial_{\Theta\Theta} + 6q_0^2 - \eta^2)\phi = \mathrm{Re}[F], \tag{6}$$

$$L\psi = (\partial_{\Theta\Theta} + 2q_0^2 - \eta^2)\psi = \mathrm{Im}[F]. \tag{7}$$

Here, q_0 is assumed to be real, and F is a lengthy complex expression, omitted for brevity. $\mathrm{Re}[F]$ and $\mathrm{Im}[F]$ denote the real and imaginary parts of F, respectively. First a solution for q_0 is obtained and the next step is to solve the inhomogeneous Equations (6) and (7). In solving these, we shall invoke Fredholm's solvability condition, which states that the null spaces of the operators M and L are orthogonal to the right-hand sides $\mathrm{Re}[F]$ and $\mathrm{Im}[F]$, respectively. The null space of an operator L is the space spanned by the solutions of $L^{\dagger}z(\Theta) = 0$, where $L^{\dagger}$ is the adjoint operator of L. This condition guarantees a solution without secular terms and provides evolution equations for the slowly varying parameters of the form

$$\eta \frac{\partial \xi}{\partial T_1} = \int_{-\infty}^{\infty} \mathrm{Re}[R] \frac{\partial q_0}{\partial \Theta} \, \mathrm{d}\Theta, \tag{8}$$

$$\frac{\partial \eta}{\partial T_1} = \int_{-\infty}^{\infty} \mathrm{Im}[R] q_0 \, \mathrm{d}\Theta. \tag{9}$$

For a given perturbation R, we can solve these equations for the slowly varying soliton amplitude η and ξ.

In the previous discussion, we began with a soliton as the zero-order approximation and adjusted its parameters to achieve an approximate solution of the perturbed soliton equation. In nonlinear diffusion equations, solutions emerge from balancing sources of energy with diffusion and dissipative effects. The perturbation theory outlined above can also be applied to such systems. As an example let us consider the FitzHugh–Nagumo (FN) model for nerve pulse propagation (FitzHugh, 1961):

$$\frac{\partial^2 V}{\partial x^2} - \frac{\partial V}{\partial t} = F(V) + R, \tag{10}$$

$$\frac{\partial R}{\partial t} = \varepsilon V, \tag{11}$$

where $F(V)$ is the cubic-shaped function $V(V - a)(V - 1)$. $V = V(x, t)$ is the voltage across the nerve cell membrane, and R is a recovery variable. The nerve pulse solution is localized in space and propagates by a velocity v to be determined from the model equations. Instead of using a multiscale approach, we invoke the Lindsted–Poincaré technique (Nayfeh, 1973) and expand the velocity parameter according to

$$v = v_0 + \varepsilon v_1 + O(\varepsilon^2). \tag{12}$$

The reason for introducing the above expansion is that nonlinearity will alter the propagation velocity. We also expand the dependent functions V and R

$$V = V_0 + \varepsilon V_1 + O(\varepsilon^2), \tag{13}$$

$$R = R_0 + \varepsilon R_1 + O(\varepsilon^2). \tag{14}$$

It is important to note that terms increasing to infinity for large times t, that is, secular terms in V_1 and R_1, are avoided by using the expansion of the velocity in Equation (12). Introducing a traveling wave assumption $\xi = x - vt$ in the solution and substituting expressions (12)–(14) into the FN equations (10) and (11) and ordering terms in power of ε leads to

$$\frac{\mathrm{d}^2 V_0}{\mathrm{d}\xi^2} + v_0 \frac{\mathrm{d}V_0}{\mathrm{d}\xi} - [F(V_0) - R_0] = 0, \tag{15}$$

$$\frac{\mathrm{d}R_0}{\mathrm{d}\xi} = 0. \tag{16}$$

To order $O(\varepsilon)$ we have

$$LV_1 = \frac{\mathrm{d}^2 V_1}{\mathrm{d}\xi^2} + v_0 \frac{\mathrm{d}V_1}{\mathrm{d}\xi} - F'(V_0)V_1 = R_1 - v_1 \frac{\mathrm{d}V_0}{\mathrm{d}\xi}, \tag{17}$$

$$\frac{\mathrm{d}R_1}{\mathrm{d}\xi} = -\frac{V_0}{v_0}. \tag{18}$$

Here it turns out that we can identify "boundary layers" for the determination of V and R together with appropriate matching. In the limit of $\varepsilon \to 0$, the nerve pulse voltage rises sharply from zero to a plateau value of order $O(1)$. The rapid change is equivalent to a boundary layer in fluid dynamics. Similarly a sharp decrease of the pulse appears at its rear end followed by a slow recovery interval due to a fading R. Matching of the solutions can be done in each region. The correction term v_1 in the expansion of the velocity is determined from Fredholm's solvability condition of

Equation (17). That is, the null space of the operator L should be orthogonal to the right-hand side of (17). This is equivalent to the requirement for the NLS equation in the previous section. For $a < \frac{1}{2}$ the traveling wave velocity v_0 of the leading edge is $v_0 = (1 - 2a) / \sqrt{2}$ and the first-order correction becomes

$$v_1 = \frac{\int_{-\infty}^{\infty} \left[\int_{\xi}^{\infty} V_0(\xi')\,\mathrm{d}\xi' \right] (\mathrm{d}V_0/\mathrm{d}\xi) \mathrm{e}^{v_0 \xi}\,\mathrm{d}\xi}{v_0 \int_{-\infty}^{\infty} (\mathrm{d}V_0/\mathrm{d}\xi)^2 \mathrm{e}^{v_0 \xi}\,\mathrm{d}\xi}. \qquad (19)$$

As we have used a traveling wave ansatz, the obtained solution for V_0 may not be stable. However, a closer analysis reveals that stable traveling wave solutions exist (Scott, 2003). Random perturbations make up a set of conceptually different problems that may be treated by methods of stochastic differential equations and statistics. For soliton equations, studies have been performed on random parameters modeling the stochastic variation of various physical properties, such as geometry or material. The influence of thermal noise and incoherence of amplitude and phase modulations of input pulses are other important examples (Abdullaev, 1994; Abdullaev et al., 2001).

MADS PETER SØRENSEN

See also **Averaging methods; FitzHugh–Nagumo equation; Fredholm theorem; Korteweg–de Vries equation; Multisoliton perturbation theory**

Further Reading

Abdullaev, F.Kh. 1994. *Theory of Solitons in Inhomogeneous Media*, Chichester: Wiley

Abdullaev, F.Kh., Bang, O. & Sørensen, M.P. 2001. *NATO Advanced Research Workshop on Nonlinearity and Disorder: Theory and Applications*, Boston and Dordrecht: Kluwer

Dodd, R.K., Eilbeck, J.C., Gibbon, J.D. & Morris, H.C. 1984. *Solitons and Nonlinear Wave Equations*, London: Academic Press

FitzHugh, R. 1961. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 1: 445–466

Karpmann, V.I. 1979. Soliton evolution in the presence of perturbation. *Physica Scripta*, 20: 462–478

Kaup, D.J. 1990. Perturbation theory for solitons in optical fibers. *Physical Review* A, 24: 5689–5694

Kivshar, Y.S. & Malomed, B.A. 1989. Solitons in nearly integrable systems. *Reviews of Modern Physics*, 61: 763–915

Nayfeh, A.H. 1973. *Perturbation Methods*, New York: Wiley-Interscience

Nguyen, P., Skovgaard, P., Sørensen, M.P. & Christiansen, P.L. 1995. Solitons in fibre lasers and mode-locked systems. *Physical Review* A, 24: 5689–5694

Scott, A. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

PHASE DYNAMICS

Phase as a physical term is most commonly used in reference to oscillatory processes. The concept of phase dynamics is particularly relevant to self-sustained or limit-cycle oscillations. The orbital shape of a limit cycle is more or less rigid against perturbations, while the remaining degree of freedom along the closed orbit (the phase) lacks such rigidity. Consequently, when a limit-cycle oscillator is slightly perturbed, the resulting nontrivial dynamics could well be described in terms of the phase alone. This is the intuition behind the phase description of a variety of oscillatory phenomena, among others biological oscillations (Winfree, 2000).

Mathematically, limit-cycle oscillations are described with a time-periodic solution of an autonomous nonlinear ordinary differential equation $\mathrm{d}X / \mathrm{d}t = F(X)$, $X \in R^n$. Let such a solution with angular frequency ω be given by a 2π-periodic function $X_0(\omega t + \phi_0)$ involving an arbitrary phase constant ϕ_0. One may also introduce a phase variable $\phi(t)$ to represent this one-parameter family of solutions as

$$X(t) = X_0(\phi(t)), \qquad \frac{\mathrm{d}\phi}{\mathrm{d}t} = \omega. \qquad (1)$$

Suppose the oscillator is slightly perturbed, by which F is slightly modified. As the resulting deformation of the closed orbit would be negligibly small, the first equation in (1) still holds approximately, while the correction to the instantaneous frequency $\mathrm{d}\phi/\mathrm{d}t$, however small it may be, would give rise to a large difference in ϕ in the long run and hence is indispensable.

Further, suppose that the perturbation represents the influence from another oscillator with state vector X' coupled weakly to the first one. If these oscillators are identical in nature, one may also use the approximation $X'(t) = X_0(\phi'(t))$ for the second oscillator, so that the current state of our oscillator pair may well be specified only with their phases $\phi(t)$ and $\phi'(t)$. Therefore, the small term to be added to the phase equation in (1) should generally be given by some function $G(\phi, \phi')$ which is 2π-periodic both in ϕ and ϕ'. Due to the assumed smallness of G relative to ω, it turns out that the effect of G over one cycle of oscillation depends only on the phase difference $\phi - \phi'$. Thus, using a 2π-periodic function Γ, one may obtain

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \omega + \Gamma(\phi - \phi') \qquad (2)$$

and, if we need, a similar equation for the second oscillator also. The function Γ can be computed in principle with the knowledge of the original dynamical-system model. Note, however, that the formula for Γ requires an extension of the definition of phase slightly outside the limit-cycle orbit (Kuramoto, 1984a).

The above equation may readily be extended to larger assemblies of weakly coupled oscillators. Furthermore, the oscillators may be slightly dissimilar, with their effect appearing only through a small difference in natural frequency ω. Thus, for such an assembly of N oscillators, we generally have

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \omega_i + \sum_{j=1}^{N} \Gamma_{ij}(\phi_i - \phi_j), \quad i = 1, 2, ..., N. \quad (3)$$

This is called the phase oscillator model which is extremely useful for the study of complex collective behavior of large assemblies of coupled oscillators appearing in large varieties of fields ranging from physics to brain science (Pikovsky et al., 2001).

Phase dynamics are also useful when the oscillators constitute a continuous field. As a representative class of such continua, we will focus on oscillatory reaction-diffusion systems given by

$$\frac{\partial X}{\partial t} = F(X) + \hat{D}\nabla^2 X, \quad (4)$$

where $X(r,t)$ represents a space–time-dependent composition vector of dimension n and $\hat{D}$ is a diagonal matrix of n diffusion constants. If the spatial variation of X is of long wavelength, the diffusion term can be regarded as a small perturbation driving each local oscillator, so that the previous idea of phase dynamics should work. The approximation $X(r,t) = X_0(\phi(r,t))$ would again be valid at each spatial point. Its application to the diffusion term produces terms proportional to $\nabla^2\phi$ and $(\nabla\phi)^2$. According to a systematic perturbation theory, such terms actually appear in the phase equation. Specifically, we obtain a nonlinear phase diffusion equation

$$\frac{\partial \phi}{\partial t} = \omega + \nu\nabla^2\phi + \mu(\nabla\phi)^2 \quad (5)$$

to the nontrivial lowest order approximation with respect to the smallness in the spatial gradient of ϕ, where positive ν has been assumed. If ν is negative, the uniform oscillation is unstable. For small negative ν, the idea of phase dynamics still works. Then (5) is modified to give an equation equivalent to the Kuramoto–Sivashinsky equation, whose solution describes spatiotemporal chaos called phase turbulence.

In a more general context, phase is a degree of freedom appearing when a certain continuous symmetry of the system has been broken spontaneously (Mori & Kuramoto, 1998). For instance, the phase of a limit-cycle oscillator reflects the fact that the oscillation breaks the invariance of the governing evolution equation (assumed autonomous) with respect to temporal translations $t \to t + t_0$. Such symmetry-breaking solutions, not restricted to time-periodic solutions, generally involve an arbitrary phase constant, thus forming an infinite family recovering as a whole the original symmetry. This also means that phase has no prescribed value to which it is bound to relax or, equivalently, phase represents a neutrally stable dynamical variable. When the system is weakly perturbed, such neutrality is slightly violated causing the phase to evolve slowly, possibly apart from a constant drift. In contrast, all the other degrees of freedom will damp quickly, thus following adiabatically the slow motion of the phase. This explains why the phase so often dominates the pattern dynamics in nonlinear dissipative systems.

The general argument given above implies that the applicability of phase dynamics is by no means restricted to oscillatory systems. Two important non-oscillatory applications of phase dynamics, again in reference to reaction-diffusion systems with large spatial extension, will be touched upon next.

Reaction-diffusion systems may develop a spatially periodic static pattern out of a uniform state through the Turing instability, by which the spatial translational symmetry has been broken. When the local phase of the periodic pattern is subject to a large-scale spatial modulation, phase dynamics is applicable and describes the pattern dynamics through a slow evolution of the phase (Kuramoto, 1984b). For instance, the transient dynamics of recovery of a regular pattern can be described with a simple phase-diffusion equation.

Another important class of symmetry-breaking patterns arising in reaction-diffusion systems is localized structures such as moving domain boundaries and solitary pulses. In these cases, phase corresponds to the location of a moving front. For a planar front, the phase forms a uniform field advancing at a constant speed. When the planar front is deformed slightly in large spatial scale, the phase becomes slightly non-uniform and starts to evolve slowly. It is concluded from a systematic theory of phase dynamics that when the planar front is stable, the equation for the phase takes the same form as (5) with ∇ being replaced with a gradient of reduced dimension (Kuramoto, 1984a). A planar front that is weakly unstable due to small negative ν leads to phase turbulence described essentially by the Kuramoto–Sivashinsky equation.

Yoshi Kuramoto

See also **Coupled oscillators; Kuramoto–Sivashinsky equation; Synchronization**

Further Reading

Kuramoto, Y. 1984a. *Chemical Oscillations, Waves, and Turbulence*, Berlin: Springer

Kuramoto, Y. 1984b. Phase dynamics of weakly unstable periodic structures. *Progress of Theoretical Physics*, 71(6): 1182–1196

Mori, H. & Kuramoto, Y. 1998. *Dissipative Structures and Chaos*, Berlin: Springer

Pikovsky, A., Rosenblum, M. & Kurths, J. 2001. *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge and New York: Cambridge University Press

Winfree, A.T. 2000. *The Geometry of Biological Time*, 2nd edition, Berlin and New York: Springer

PHASE LOCKING

See **Coupled oscillators**

PHASE MATCHING

See **Nonlinear optics**

PHASE PLANE

The term *phase plane*, naturally enough, refers to the phase space of a two-dimensional dynamical system. The term is worthy of consideration in its own right because of the various special techniques that exist for the analysis of two-dimensional continuous-time dynamical systems (i.e., flows).

As an example, consider the two-dimensional flow induced by the following differential equations:

$$\begin{aligned} \dot{x}_1 &= x_1\,(1 - x_2)\,, \\ \dot{x}_2 &= x_2\,(x_1 - r)\,. \end{aligned} \tag{1}$$

Each of these equations specifies one component of a two-dimensional vectorfield on the plane, tangent at each point to the induced flow. Equations (1) were first employed by Alfred Lotka (1920) to describe a hypothetical bimolecular chemical reaction capable of sustained oscillations, assuming mass action kinetics. At about the same time and independently, Vito Volterra (1926) made use of the same equations to model the interaction of biological predator and prey populations. In Equations (1), x_1 denotes the density of the prey species and x_2 that of the predator. Figure 1a shows the relevant (positive) quadrant of the phase plane of Equations (1).

The most important aspects of the dynamics of smooth flows are determined by their α- and ω-limit sets. The former are those sets approached by an orbit as $t \to -\infty$; the latter, those approached as $t \to +\infty$. In smooth planar flows, these limit sets are of only three types: singular points (or equilibria), closed (or periodic) orbits, and homo- and heteroclinic chains. The last are made up of singular points together with bounded orbits, each of which has one singular point as α-limit and another (possibly identical) singular point as ω-limit. It should be noted that more complex behaviors such as quasi-periodicity and chaotic dynamics are not possible in two-dimensional flows.

Singular points are determined as the solutions of a system of two algebraic equations and classified, on the basis of linear stability analysis, into six types: stable and unstable nodes, stable and unstable foci, centers, and saddles. Useful information about the simultaneous existence of singular points and closed orbits can be obtained from consideration of the Poincaré index, an integer associated with each singular point (*See* **Winding numbers** and Andronov et al., 1966). For example, the Poincaré index can be used to show that every closed orbit encircles at least one singular point.

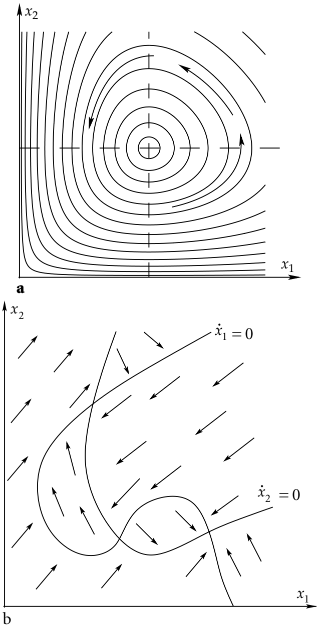

Figure 1. (a) Phase plane of the Lotka–Volterra equations (1). Representative orbits are shown as solid lines; the direction of flow is indicated by arrows. The isoclines are shown using dashed lines. The positive quadrant is filled with a family of closed orbits, corresponding to the fact that equations (1) are Hamiltonian. (b) Diagram showing a generic isocline analysis: many qualitative features of the dynamics are made plain. The isoclines show the loci where the vectorfield is horizontal or vertical. Intersections of the isoclines occur at singular points. From left to right, the three singular points are evidently of focus, saddle, and stable node types. It is not clear from the isocline analysis alone whether the focus is stable or unstable.

The method of isoclines is frequently invaluable in deriving qualitative information about the dynamics of a given smooth flow. An isocline is the locus where one of the vector field components vanishes. The two isoclines together split the phase plane into four divisions: within each such division each vector field component has a uniform sign. Figure 1b depicts a typical isocline analysis.

Closed orbits which attract or repel nearby orbits are called limit cycles. Systems possessing stable limit cycles are said to exhibit auto-oscillations. The method of Liénard can be used to demonstrate the existence of limit cycles in certain systems (Hartman, 2002). Sometimes, a limit cycle can also be found

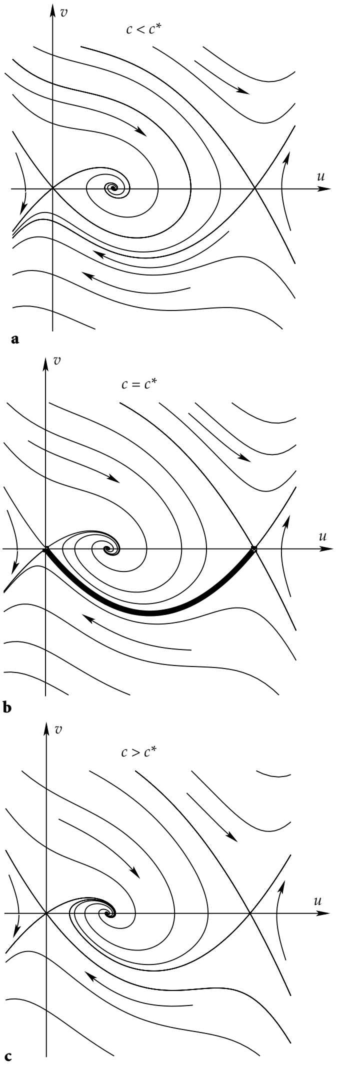

Figure 2. Determining a heteroclinic connection by shooting. Adjusting a parameter (in this case the wave speed c) causes the unstable manifold of the rightmost saddle point to move. For small positive values of c, one branch of the unstable manifold crosses the negative v-axis, never to return (a). At larger values of c, it is drawn into the stable focus point at $u = a$, $v = 0$ (c). There is precisely one intermediate value $c = c^*$ for which a branch of the unstable manifold of the rightmost saddle coincides with a branch of the stable manifold of the leftmost saddle, i.e., for which there is a heteroclinic connection (the emboldened orbit in b).

by association with a Hopf bifurcation (Arnol'd, 1973; Hale & Koçak, 1991). The Poincaré–Bendixson theorem can also be used to demonstrate the existence of limit cycles. It states that any non-empty compact α- or ω-limit set of a smooth planar flow which does not contain a singular point is a closed orbit.

Homo- and heteroclinic chains are, in general, more difficult to determine than equilibria. So-called shooting methods may sometimes be used to prove the existence of homoclinic or heteroclinic connections. For example, consider the following planar system, which arises as the traveling wave reduction of the Zeldovich–Frank-Kamenetsky equation

$$\begin{aligned} \dot{u} &= v, \\ \dot{v} &= u\,(u-a)\,(u-1) - c\,v. \end{aligned} \tag{2}$$

Here, the parameter c is a wave speed (*See* **Wave of translation**), and we assume (without loss of generality) that $0 < a < \frac{1}{2}$. It is readily seen that there are precisely three singular points (at $u = 0$, a, 1 and $v = 0$); the first and third are saddles; the intermediate is of focal type (see Figure 2). If we seek a rightward moving wave ($c > 0$), this focus is stable. For small values of c, the lower branch of the rightmost saddle's unstable manifold crosses the negative v-axis, where as for sufficiently large values of c, it crosses the positive u-axis before winding about the stable focus. It follows that there is an intermediate wave speed c^* for which a heteroclinic connection exists. It should be noted that this shooting argument depends on the fact that deleting a point from an interval in two-dimensional space topologically disconnects the interval. Shooting methods become more complicated in higher dimensions (see, e.g., Dunbar, 1984).

Aaron A. King

See also **Bifurcations; Chaotic dynamics; Chemical kinetics; Dynamical systems; Hamiltonian systems; Hopf bifurcation; Phase space; Poincaré theorems; Population dynamics; Wave of translation; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Andronov, A.A., Vitt, A.A. & Khaikin, S.E. 1966. *Theory of Oscillators*, Oxford: Pergamon Press (translated from Russian)

Arnol'd, V.I. 1973. *Ordinary Differential Equations*, Cambridge, MA: MIT Press

Arnol'd, V.I. & Il'yashenko, Y.S. 1988. *Ordinary Differential Equations and Smooth Dynamical Systems. Dynamical Systems*, vol. 1, Berlin: Springer (original Russian edition 1985)

Dunbar, S.R. 1984. Traveling wave solutions of diffusive Lotka–Volterra equations: a heteroclinic connection in R^4. *Transactions of the American Mathematical Society*, 286: 557–594

Hale, J.K. & Koçak, H. 1991. *Dynamics and Bifurcations*, New York: Springer

Hartman, P. 2002. *Ordinary Differential Equations*, 2nd edition, Philadelphia: SIAM (corrected republication of edition published by Birkhäuser, Boston, 1982)

Hirsch, M.W. & Smale, S. 1974. *Differential Equations, Dynamical Systems, and Linear Algebra*, New York: Academic Press

Lotka, A.J. 1920. Undamped oscillations derived from the law of mass action. *Journal of the American Chemical Society*, 42: 1595–1598

Volterra, V. 1926. Fluctuations in abundance of a species considered mathematically. *Nature*, 118: 558–560

PHASE SPACE

The equations of motion of a mechanical system are usually of second order, and they determine the entire future from the initial *state*, which consists of both positions and velocities. The space of all states is called the *phase space*. For the mathematical pendulum (a point mass in the plane at one end of a massless rod whose other end is fixed), the configuration space is a circle, at each point of which one can choose any tangent vector as the velocity. Therefore, the phase space is a circle with a line attached to each point, that is, a cylinder. The pendulum equation for the position x and velocity v, $\ddot{x} + \sin x = 0$ converts to $(\dot{x}, \dot{v}) = (v, -\sin x)$, so the dynamics is described by integral curves of the vector field $R(x, v) = (v, -\sin x)$ on the cylinder.

That the notion of phase space is natural is also suggested by the Liouville theorem: the skew-symmetry of the Hamilton equation makes the Hamiltonian vector field divergence-free, and accordingly, a Hamiltonian flow preserves the volume on phase space.

Generally, a dynamical system consists of a phase space and a time-evolution of first order. The phase space is a set with some structure, such as differentiable (in the case of differential equations, this belongs to smooth dynamics), topological (one then speaks of topological dynamics), or measurable (this is the subject of ergodic theory, which arose from the Liouville theorem), and the time evolution is a one-parameter family of transformations that preserve this structure and that map initial states to states at another time. The time parameter may run through real numbers (continuous-time system) or integers (discrete-time system, iterations of a map and possibly its inverse). Specifically, a continuous-time system is given by a family $(f^t)_{t\in\mathbb{R}}$ of maps. If $f^t(f^s(x)) = f^{t+s}(x)$ for every s, t, x, then we say that this family is a flow. In the discrete-time case, one considers the iterates $(f^n)_{n\in\mathbb{Z}}$, where $f^0(x) = x$, $f^{n+1}(x) = f(f^n(x))$ for $n \geq 0$, and $f^n(x) = (f^{-1})^n(x)$ for $n < 0$, or if the map is not invertible, only positive iterates $(f^n)_{n\in\mathbb{N}}$. The maps $f_a(x) = ax(1-x)$ are a popular example of the latter (the so-called logistic map).

The long-term behavior of flows in the plane is well understood (*See* **Phase plane**): In the long run, any orbit either approaches fixed points or is asymptotically periodic (Poincaré–Bendixson theorem). This is ultimately due to the fact that a closed curve, such as a periodic orbit, divides the plane into separate regions. Already in three-space, one gets chaotic behavior, such as in the Lorenz attractor.

Qualitative Theory of Differential Equations and Dynamical Systems

On the phase space of a smooth continuous-time dynamical system, the time evolution is given by a first-order differential equation $\dot{x} = R(x, t)$. Suppose the right-hand side R satisfies a Lipschitz condition in x. This means that there is a constant M such that $d(R(x, t), R(x', t)) \leq Md(x, x')$ for all x, x'. The basic Picard theorem then guarantees existence and uniqueness of solutions for any initial condition. Otherwise solutions may not be unique ($\dot{x} = \sqrt[3]{x}$ has infinitely many solutions with initial value 0) or may not exist for any uniform amount of time (the solutions of $\dot{x} = x^2$ have singularities).

If solutions $x(t)$ exist for all time, as we henceforth assume, they define time-t-maps by $f^t(x(0)) = x(t)$. Each map f^t is as smooth as the right-hand side R of the differential equation (smooth dependence on initial conditions).

If R is independent of t, then the differential equation is said to be "autonomous," and R gives a vector field on the phase space that prescribes the velocity vectors of solutions. The family of time-t-maps is then a flow. The iterates of the time-1-map of a flow produce a discrete-time dynamical system whose study may yield useful information about the flow. If R depends on t, the system is said to be non-autonomous. Explicit time dependence can, for example, arise from forcing terms (forced pendulum $\ddot{x} + \sin x = \sin \omega t$) or from varying parameters (driving of a swing by parametric forcing, $\ddot{x} + \rho(t) \sin x = 0$).

An "orbit" or *trajectory* of a continuous-time system is a parametrized curve $(f^t(x))_{t\in\mathbb{R}}$. An orbit or trajectory of a map consists of the sequence of images of a point under iteration of the map: $(f^n(x))_{n\in\mathbb{Z}}$.

A singular point of a differential equation $\dot{x} = R(x, t)$ is a point x for which the right-hand side is zero for all t, that is, an equilibrium, or constant solution, or fixed point. Fixed points of a map f are those points x for which $f(x) = x$. A periodic point is a state that repeats at some positive time. For differential equations, this corresponds to solutions that are periodic functions of time; for maps, these are fixed points of an iterate. For the flow on the cylinder generated by $\ddot{x} + \sin x = 0$, all but four orbits are periodic; the point $(5 + \sqrt{5})/8$ is two-periodic for the map $4x(1-x)$.

Fixed and periodic points can be anchors for the study of the global orbit structure, and therefore, it is important to understand the behavior of nearby orbits. A fixed point is said to be attracting if orbits of nearby

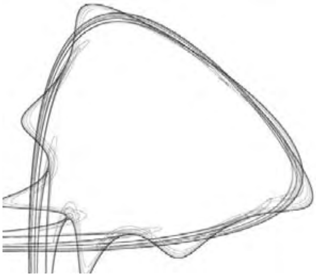

Figure 1. Stable and unstable manifold of a hyperbolic fixed point.

points stay nearby (Poisson stability) and converge to it for large positive time (asymptotic stability). (The example of a circle map like this ↻ with a fixed point at the top illustrates that the second condition does not imply the first.) This is the case if the differential of the map (or time-1-map in the case of a flow) at that point has only eigenvalues of absolute value less than 1. If all eigenvalues have absolute value greater than 1 then the point is repelling: There is a neighborhood which every other point leaves in positive time. The map $f_2(x) = 2x(1-x)$ has 0 and $\frac{1}{2}$ as fixed points. 0 is repelling and $\frac{1}{2}$ is attracting. In fact, $\frac{1}{2}$ is superattracting: $f_2'(\frac{1}{2}) = 0$ and orbits near $\frac{1}{2}$ approach $\frac{1}{2}$ faster than exponentially.

If eigenvalues of the differential are allowed to lie both inside the unit circle and outside it but not on it, then the fixed point is said to be "hyperbolic." In this case the Hartman–Grobman theorem states that there is a continuous coordinate change that maps orbits near the fixed point to orbits of the linearized map. Moreover, tangent to the contracting and expanding subspaces of the linearization, there are the stable and unstable manifold of points positively and negatively asymptotic to the fixed point, respectively. These are smooth subspaces without self-intersections, but they may be packed into the phase space in a complicated way.

For periodic points of maps, the analysis of stability can be carried out by studying the appropriate iterate; for a flow, one likewise studies Poincaré return maps as follows. Take a small hypersurface through the periodic point transverse (for example, orthogonal) to the flow. The orbit of every point sufficiently near the periodic point returns to this surface at a time close to the period, and this defines a map from a neighborhood of the periodic point in the hypersurface into the hypersurface, with the original periodic point as a fixed point.

If, for example, a periodic orbit is an attracting fixed point for the return map, then it is a limit cycle: All nearby orbits are asymptotic to it. For the mathematical pendulum from the introduction, the circle $\{v = 0\}$ is a section of the cylinder, and the return map is defined for all of its points. It has two fixed points, and all other points are two-periodic.

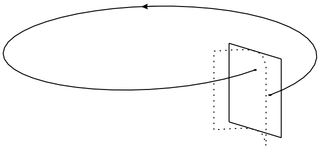

Figure 2. A section.

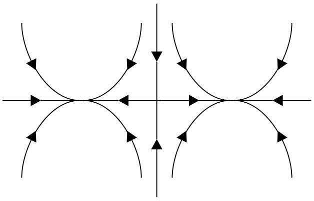

Figure 3. Two attracting points.

A property complementary to stability of a fixed or periodic point as defined in terms of the behavior of nearby orbits (i.e., perturbations of the initial condition) is that of stability under perturbations of the dynamical system. An easy way to guarantee this is *transversality*, which is weaker than hyperbolicity: A fixed point $x = f(x)$ is said to be *transverse* if the derivative of f at x does not have 1 as an eigenvalue. (This implies that there are no other fixed points nearby.) In this case, any C^1-perturbation of f (i.e., one that changes derivatives only a little) also has a (transverse) fixed point near x. The origin is a nontransverse fixed point of $x(1-x)$, and indeed, it is absent for $x(1-x)+\varepsilon$ with $\varepsilon < 0$. The creation of two (hyperbolic) fixed points as ε changes from negative to positive is a basic local *bifurcation*. For differential equations, transversality corresponds to invertibility of the differential of the right-hand side.

An "invariant set" is a union of orbits; for example, [0, 1] is invariant under $4x(1-x)$. It is a repeller if it has a neighborhood in which only orbits of points in the invariant set stay for all positive time. It is an "attractor" if there is a neighborhood that is mapped into itself and the intersection of whose positive-time iterates is the invariant set. (Usually, one also requires that there is no proper subset with the same property. Thus, Figure 3 shows two attracting fixed points, and the interval with these as endpoints is not considered an attractor.)

The "basin of attraction" is the set of points that are asymptotic to the attractor. For example, the interval (0, 1) is the basin of attraction of $\frac{1}{2}$ for the map $2x(1-x)$.

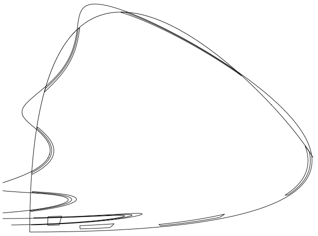

Figure 4. The Birkhoff–Smale theorem.

If two hyperbolic fixed points (saddles) in the plane are connected by a curve segment that lies in the unstable manifold of one of them and in the unstable manifold of the other, then this segment is called a "separatrix" (because it often separates two basins of attraction). More generally, the intersection of the stable manifold of one hyperbolic point with the unstable manifold of another is called a "heteroclinic intersection," and the orbit of every intersection point is called a "heteroclinic orbit." If the two fixed points coincide then one uses the terms "homoclinic intersection" and *homoclinic orbit* instead. The Birkhoff–Smale theorem asserts that if a homoclinic intersection is transverse (or if there is a pair of transverse heteroclinic intersections, that is, two hyperbolic points such that the unstable manifold of each of them intersects the stable manifold of the other point transversely), then there is a "horseshoe," that is, a rectangle that (under an iterate) gets mapped across itself in a horseshoe-like fashion as illustrated in Figure 4 and in Anosov and Axiom-A systems. This implies directly that there is an invariant Cantor set on which the dynamical system exhibits deterministic chaos.

There are several ingredients that make up chaotic behavior. One of these is recurrence. There are several recurrence properties. A point is said to be *recurrent* if it returns arbitrarily near to its initial condition. For a rigid rotation of a circle by an irrational number of degrees, all points have this property. By contrast, a point is said to be *transient* or *wandering* if it has a neighborhood all of whose images are pairwise disjoint. For the circle map $\circlearrowleft$ with a fixed point on top, all nonfixed points are wandering. The set of nonwandering points is called the nonwandering set. Nonwandering orbits can be closed by a localized C^1 perturbation of the map (Pugh closing lemma). A dynamical system is said to be "topologically transitive" if it has a dense orbit and "minimal" if every orbit is dense. Irrational circle rotations have both properties. Minimality does not reflect chaotic behavior. Existence of a dense orbit is equivalent to the condition that for any two open sets there are arbitrarily large times at which the image of one of these sets overlaps the other. A strengthening of this property is that such overlap occurs for all sufficiently large times; this is called topological mixing and implies sensitive dependence on initial conditions. Following Devaney, one can say that a dynamical system is *chaotic* if it is topologically transitive and the set of periodic points is dense. This also implies sensitive dependence. A condition stronger than sensitive dependence is "expansivity": There is a universal positive constant by which the images of any two points, no matter how close initially, are separated at some time. The cat map and horseshoes are good examples of dynamical systems with these properties.

The Poincaré return map is not the only construction that produces a new dynamical system with a different phase space. Another straightforward one is the product of two dynamical systems. For example, the flow of rotations of the unit circle given by $x_1 = \alpha \cos t$, $x_2 = \alpha \sin t$ can be combined with a similar flow $y_1 = \omega \cos t$, $y_2 = \omega \sin t$ to a flow on the two-torus in $\mathbb{R}^4$ defined by all four equations. (Note that the plane defined by $y_1 = y_2 = 0$ is a section on which the return map is a time-$2\pi/\omega$-map of the first flow.) If one projects this to the $x_1 y_1$-plane, one gets Lissajous figures. In some applications, these readily show whether modes in weakly nonlinear oscillators are locked together.

Boris Hasselblatt

See also **Anosov and Axiom-A systems; Attractors; Cat map; Hamiltonian systems; Horseshoes and hyperbolicity in dynamical systems; Maps; Phase plane; Poincaré theorems; Population dynamics; Stability**

Further Reading

Arnol'd, V.I. 1989. *Mathematical Methods of Classical Mechanics*, 2nd edition, Berlin and New York: Springer (original Russian edition 1974)

Arnol'd, V.I. 1992. *Ordinary Differential Equations*, Berlin and New York: Springer (original Russian edition 1971, translated from 3rd edition 1984)

Fielder, B. (editor). 2002. *Handbook of Dynamical Systems*, vol. 2, Amsterdam and New York: Elsevier

Hasselblatt, B. & Katok, A. (editors). 2002. *Handbook of Dynamical Systems*, vol. IA, Amsterdam and New York: Elsevier

Hasselblatt, B. & Katok, A. 2003. *Dynamics: A First Course* (see also vol. 1B, 2005) Cambridge and New York: Cambridge University Press

Katok, A. & Hasselblatt, B. 1995. *Introduction to the Modern Theory of Dynamical Systems*, Cambridge and New York: Cambridge University Press

Katok, A., de la Llave, R., Pesin, Y. & Weiss, H. (editors). 2001. *Smooth Ergodic Theory and Its Applications (Summer Research Institute Seattle, WA, 1999), Proceedings of Symposia in Pure Mathematics*, vol. 69. Providence, RI: American Mathematical Society

Robinson, R.C. 1995. *Dynamical Systems, Stability, Symbolic Dynamics, and Chaos*, Boca Raton, FL: CRC Press

Strogatz, S. 1994. *Nonlinear Dynamics and Chaos*, Reading, MA: Addison-Wesley

PHASE-SPACE DIFFUSION AND CORRELATIONS

For problems concerned with particle acceleration and heating, which can be described in two degrees of freedom, surfaces of section in a two-dimensional phase space are usually appropriate for representing the motion (Lichtenberg & Lieberman, 1991, Chapter. 3). For periodically driven systems, a mapping representation is often convenient to describe the motion (*See* **Fermi acceleration and Fermi map; Standard map**). In the surface of section there is a characteristically divided phase space in which regular Kolmogorov–Arnol'd–Moser (KAM) curves and stochastic trajectories are intermingled. In the globally stochastic region of the phase space for a system with two degrees of freedom, in which KAM curves spanning the phase coordinate do not exist, a complete description of the motion is generally impractical. We can then seek to treat the motion in a statistical sense; that is, the evolution of certain average quantities can be determined, rather than the trajectory corresponding to a given set of initial conditions (Wang & Uhlenbeck, 1945). Such a formulation in terms of average quantities is also the basis for statistical mechanics (see Penrose, 1970, for example). The mathematical foundations for many of the results can be found in Arnol'd & Avez (1968).

In regions in which the trajectories are stochastic, nearby trajectories diverge exponentially in time. The divergence is usually measured by calculating the Lyapunov characteristic exponent σ of a trajectory $\boldsymbol{x}$ and a nearby trajectory $\boldsymbol{x} + \boldsymbol{w}$,

$$\sigma(\boldsymbol{x}, \boldsymbol{w}) = \lim_{t\to\infty,\ d(0)\to 0} \left(\frac{1}{t}\right) \ln \frac{\mathrm{d}(t)}{\mathrm{d}(0)}, \qquad (1)$$

where d is the distance separating the trajectories. Analytical and numerical calculations of the σ's (especially the maximum value of $\sigma = \sigma_1$ with respect to variations of $\boldsymbol{w}$) are widely used as measures of the degree of stochasticity in near-integrable systems. The commonly used numerical procedure for calculating the Lyapunov exponents was developed by Benettin et al. (1976).

In many problems, the density distribution in action space is the important quantity. Its dynamics is simplified by an average over phases to find the dynamical friction and diffusion coefficients for the action, which can then be used for determining its time evolution. If the phases are decorrelated after each mapping step, then a random phase approximation can be used which greatly simplifies the calculation. If the averaging over phase can be performed for a localized value of the action, then the resulting evolution is a Markov process, leading to the Fokker–Planck equation for the evolution of the distribution (Wang & Uhlenbeck, 1945)

$$\frac{\partial P}{\partial n} = -\frac{\partial}{\partial I}(BP) + \frac{1}{2}\frac{\partial^2}{\partial I^2}(DP), \qquad (2)$$

where P is the probability distribution in the action I, B and D are the friction and diffusion coefficients, appropriately averaged over the phases, and n is a characteristic time over which the averaging can be performed. For example, using the standard map

$$\begin{aligned} I_{n+1} &= I_n + K \sin\theta_n, \\ \theta_{n+1} &= \theta_n + I_{n+1}, \end{aligned} \qquad (3)$$

then one step of the mapping gives

$$\Delta I_1 = K \sin\theta_0, \qquad (4)$$

and the transport coefficients, with phases randomized on each step, are

$$F_{\mathrm{QL}} = F = \frac{1}{2\pi}\int_0^{2\pi} \Delta I_1 \,\mathrm{d}\theta_0 = 0, \qquad (5)$$

$$D_{\mathrm{QL}} = \frac{D}{2} = \frac{1}{4\pi}\int_0^{2\pi} (\Delta I_1)^2 \,\mathrm{d}\theta_0 = \frac{K^2}{4}, \qquad (6)$$

where the subscripts QL refer to quasilinear (phase randomized) values with no higher-order correlations. (The factor of 2 difference between D_{QL} and D is a convention.)

Phase correlations, which always exist in a phase space with both regular and stochastic regions, complicate the calculation procedures. Close to the borders between stochastic and regular regions, the correlations become pronounced, requiring entirely different procedures for determining the diffusion. The existence of accelerator modes in the standard map also leads to nondiffusive behavior. Other phenomena of interest for diffusion calculations are the effect of noise and the effect of slow changes of system parameters. For weak correlations, for example, large K in the standard map, corrections to the single-step transport coefficients can be obtained. The corrections were obtained using Fourier techniques by Rechester et al. (1981). In the opposite limit for which a phase-spanning KAM curve (torus) has just been broken, resulting in a cantorus (the KAM curve becomes a cantor set), an approximate rate of local diffusion can be calculated. These techniques were developed to analyze the standard map, but can be used in various approximations to describe other two-degree-of-freedom systems. For a review of the various techniques and limitations see Lichtenberg & Lieberman (1991, Chapter 5).

ALLAN J. LICHTENBERG

See also **Diffusion; Fermi acceleration and Fermi map; Kolmogorov–Arnol'd–Moser theorem; Standard map**

Further Reading

Arnol'd, V.I. & Avez, A. 1968. *Ergodic Problems of Classical Mechanics*, New York: Benjamin

Benettin, G., Galgani, L. & Strelcyn, J.M. 1976. Kolmogorov entropy and numerical experiments. *Physical Review* A, 14: 2338–2345

Lichtenberg, A.J. & Lieberman, M.A. 1991. *Regular and Chaotic Dynamics*, 2nd edition, New York: Springer

Penrose, O. 1970. *Foundations of Statistical Mechanics*, Oxford: Pergamon

Rechester, A.B., Rosenbluth, M.N. & White, R.B. 1981. Fourier-space paths applied to calculation of diffusion in the Chirikov–Taylor map. *Physical Review* A, 23: 2664–2672

Wang, M.C. & Uhlenbeck, G.E. 1945. On the theory of Brownian motion II. *Reviews of Modern Physics*, 17: 323–342

PHASE SPACE RECONSTRUCTION

See **Embedding methods**

PHASE SYNCHRONISM OF CHAOS

See **Synchronization**

PHASE TRANSITIONS

A phase transition is a change in the degree or nature of order in a system and/or a change in its symmetry caused by a change in a thermodynamic variable such as temperature or pressure. When water freezes as the temperature is lowered, it undergoes a phase transition from a liquid to a solid. In the liquid state, the positions of the water molecules are disordered and the system is isotropic. In the solid state, however, the molecules are fixed at the sites of a crystal lattice, so the system has positional order and the symmetry of the ice lattice. Other examples of phase transitions include the liquid-vapor transition, in which the ordering is related to a change in density of the material; the transition from paramagnet to ferromagnet, which involves the appearance of a net magnetization as the temperature is lowered through the Curie temperature; the isotropic-nematic transition in liquid crystals, in which the rod-like liquid crystal molecules all become oriented in the same direction; and the superfluid transition in liquid helium, in which a macroscopic fraction of the helium atoms condense into a single quantum mechanical energy level. There are many others.

For any phase transition, we can define an order parameter. This is a quantity related to the degree of ordering which changes as the system goes through the transition. One measure of the order in a system is its entropy: a more ordered system has a lower entropy. While the entropy could be used as an order parameter, it is normally difficult to measure. An order parameter is usually physically measurable and is typically chosen to be zero in the less-ordered phase and nonzero in the more-ordered phase. For the liquid-vapor transition, the order parameter is proportional to the difference in density between the liquid and vapor phases. The order parameter for the ferromagnetic transition is simply the magnetization. For the superfluid transition in liquid helium, the order parameter is a complex number describing the macroscopic wave function of the superfluid.

Phase transitions can be classified as first order or second order. The terminology refers to the fact that certain derivatives of the free energy are discontinuous at a phase transition. In a first-order transition, it is the first derivatives (entropy, density, magnetization) that are discontinuous, while in a second-order transition, the discontinuity occurs in the second derivatives. A more practical classification is in terms of the behavior of the order parameter. At a first-order transition, the order parameter changes discontinuously. These transitions exhibit hysteresis and a latent heat resulting from the discontinuous change in entropy. At a second-order transition, the order parameter changes continuously; there is no hysteresis and no latent heat. In equilibrium, a thermodynamic system has a well-defined free energy and stable states correspond to absolute minima of this free energy. At a phase transition, the position of the free energy minimum changes, corresponding to a change in phase.

A phase diagram for a typical liquid-vapor system is shown in Figure 1(a). In the pressure-temperature plane, there is a line of first-order phase transitions which ends at a point at which the transition is second order. The same system is shown in the temperature-density plane in Figure 1(b). The region under the coexistence curve in Figure 1(b) is forbidden; there are no stable states in this region. When the mean properties of the system would place it inside this curve, the system phase separates into coexisting liquid and vapor phases with different densities lying on the coexistence curve itself. In the center of this region, the high-temperature phase is globally unstable, while closer to the edge, it is metastable. A supersaturated vapor would lie in this metastable region.

Points in the phase diagram at which second-order transitions occur are called critical points. At these points, the equation of state of the system takes on a special form, and the order parameter and other thermodynamic quantities show a power law dependence on the distance from the critical point. For example, if we consider a magnetic system with magnetization M, temperature T, and Curie temperature T_c, then the magnetization behaves as $M \sim [(T_c - T)/T_c]^{\beta}$ for $T \leq T_c$. The magnetic susceptibility $\chi = \mathrm{d}M/\mathrm{d}H$, where H is an applied magnetic field, diverges at the critical point as

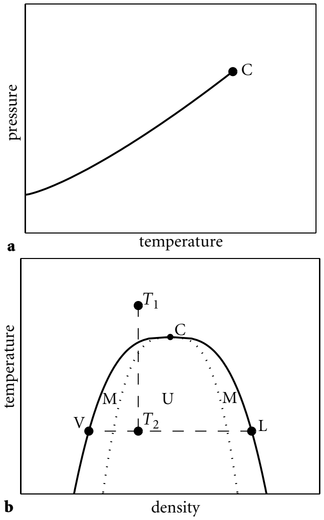

Figure 1. A typical liquid–vapor-phase diagram. (a) In the pressure–temperature plane a first-order phase transition occurs on crossing the solid line. The line of first-order transitions ends at the critical point C, at which the transition is second order. (b) The same phase diagram is shown in the temperature-density plane. The solid curve is the coexistence curve and corresponds to the solid line in (a). Outside the coexistence curve, the system exists as a single phase. In the region labeled **U** the one-phase state is unstable, and in the region labeled M it is metastable. If the system is cooled from the one-phase state at point T_1 into the unstable region at T_2, it will phase separate into a liquid component at the point L coexisting with a vapor component at V.

$\chi \sim |(T_c - T)/T_c|^{-\gamma}$. The quantities β and γ are called critical exponents. Interestingly, while T_c is different for different materials, the exponents β and γ have the same values for many different systems, including Ising ferromagnets and liquid-vapor systems, among others. This is known as "universality."

The simplest theoretical models used to describe phase transitions are called mean field theories. The van der Waals model for the liquid-gas transition and the Curie–Weiss model for the ferromagnetic transition are examples. Such theories assume that the order parameter at a particular point is determined by the average properties of the system—that is, by a mean field due to all other points of the system. Mean field theories are capable of describing first- and second-order-phase transitions qualitatively, but the values of the critical exponents are, in general, wrong. This is because mean field theories explicitly neglect the effects of spatial fluctuations in the order parameter, whereas it is these fluctuations which determine the behavior of the system near the critical point. Renormalization group theory provides a more complete description of critical point behavior.

When a system is quenched below a phase transition by suddenly changing a thermodynamic variable—for example, by suddenly dropping the temperature below the phase transition in a liquid-vapor system—it separates into two distinct phases. If the system is quenched into a region where the high-temperature phase is metastable (Figure 1), then macroscopic phase separation can only occur once large enough droplets of the unstable phase nucleate and grow. If it is quenched in the unstable region, then phase separation occurs by a process known as spinodal decomposition, whereby initially small perturbations to the local order parameter grow and coarsen with time. In a system with a conserved order parameter (e.g., a binary fluid mixture, in which the amount of each substance is fixed), the size of domains grows with time t as $L \sim t^{1/3}$, while if there is no conservation law (e.g., a ferromagnet) then $L \sim t^{1/2}$.

Phase transitions can also occur in systems out of equilibrium, for example, in the presence of a temperature gradient or a time-varying field. An important and interesting aspect of non-equilibrium systems is that the state of the system is no longer governed by the minima in the free energy, and indeed, a free energy functional may not even exist. Phase transitions can be significantly modified by non-equilibrium effects. In the case of a magnet in a time-varying field, for example, a competition between the time scale of the variations in the external field and the intrinsic relaxation time of the system itself leads to new behavior including spontaneous ordering that is not observed in the static case. Rather different phenomena, such as pattern-forming bifurcations in driven fluid systems, can also be thought of as nonequilibrium phase transitions.

John R. de Bruyn

See also **Critical phenomena; Ising model; Order parameters; Renormalization groups**

Further Reading

Chaiken, P.M. & Lubensky, T.C. 1995. *Principles of Condensed Matter Physics*, Cambridge and New York: Cambridge University Press

Chakrabarti, B.K. & Acharyya, M. 1999. Dynamic transitions and hysteresis. *Reviews of Modern Physics*, 71: 847–859

Gunton, J.D., San Miguel, M. & Sanhi, P.S. 1983. The dynamics of first order phase transitions. In *Phase Transitions and Critical Phenomena*, vol. 8, edited by C. Domb & J.L. Lebowitz, New York: Academic Press (Chapter 3). (The other volumes of this series contain review articles on many aspects of phase transitions.)

Stanley, H.E. 1971. *Introduction to Phase Transitions and Critical Phenomena*, Oxford: Oxford University Press

Tolédano, J.C. & Tolédano, P. 1987. *The Landau Theory of Phase Transitions*, Singapore: World Scientific

Uzunov, D.I. 1993. *Introduction to the Theory of Phase Transitions*, Singapore: World Scientific

PHASE TURBULENCE

See **Phase dynamics**

PHASE WINDING

See **Winding numbers**

PHI-FOUR EQUATION

See **Sine-Gordon equation**

PHOTONIC CRYSTALS

Photonic crystals are periodic dielectric (and/or metallic) structures, with periodicity comparable to a wavelength λ of interest, forming a *designable* optical medium where light propagation can exhibit unusual properties (Joannopoulos et al., 1995). The most important such property is a photonic band gap: a range of wavelengths in which there are no propagating modes in the crystal. Light in the band gap decays exponentially in the crystal, which acts like an optical "insulator." Nonlinear devices can exploit photonic crystals for enhanced phase sensitivity as well as to reduce power requirements, by means of tight spatial and long temporal confinement using the band gap and/or slow-light phenomena. Below are outlined two archetypical devices that greatly benefit from these properties of photonic crystals: Mach–Zender interferometers and bistable switches.

The simplest photonic crystals are one-dimensionally periodic multilayer films, or "Bragg mirrors," which were first studied in crystalline minerals by Lord Rayleigh (1887) (who observed that *any* periodic index variation will induce a band gap along that direction) and have since been the basis for a wide variety of applications: from reflective dielectric coatings, to distributed Bragg feedback (DFB) lasers, to fiber Bragg gratings for dispersion compensation and filters. It was not until 1987, however, that Yablonovitch (1987) and John (1987) applied the full principles of solid-state physics to electromagnetism and suggested that a three-dimensional (3-d) crystal could produce an omnidirectional gap. Since then, many photonic-crystal structures have been studied, both theoretically and experimentally, in two and three dimensions.

As a particular example, we consider a *two-dimensional* photonic crystal consisting of a square lattice of dielectric cylinders in air. This structure has a band gap for transverse-magnetic (TM) light (electric field perpendicular to the plane) and can, therefore, confine cavity and waveguide modes in point-like and linear defects of the crystal, as in Figures 1 and 2. For example, a single "defect" rod can be increased or decreased in size to trap a cavity mode within a diameter $\sim\lambda/2$, or a row of defect rods to form a waveguide. Bloch's theorem from solid-state physics (Ashcroft & Mermin, 1976) implies that modes in such a periodic waveguide propagate *without reflections*. Moreover, the waveguides form an effectively one-dimensional system, since the gap prohibits lateral scattering—this property allows photonic-crystal waveguides and cavities to be combined into complex device networks by adhering to simple design rules and symmetries. Alternatively, one can make a periodic sequence of cavities to form a coupled-cavity waveguide (CCW)—light slowly leaks from one cavity to the next, again trapping a guided mode (Yariv et al., 1999). The same principles apply in other photonic crystals, including those in three dimensions, and a direct analogue of this 2-d crystal can be made in 3-d (with an omnidirectional gap) by a stack of 2-d-like layers (Johnson & Joannopoulos, 2002).

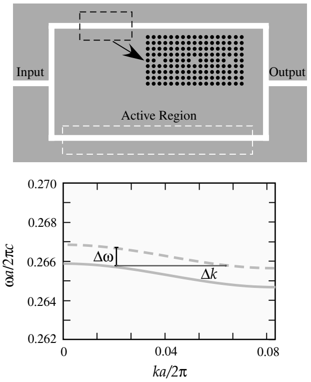

Figure 1. (Top) Schematic of a Mach–Zender interferometer: a waveguide that is split and recombined, with the relative phase modified by an active region to control the output transmission. Inset: a CCW slow-light waveguide to take advantage of photonic crystals. (Bottom) Band diagram of the CCW waveguide showing typical cosine dispersion curve. Because of the low group velocity, a slight shift of the curve (dashed) will cause a large change Δk in the wave number (horizontal axis). (The units are in terms of the distance a, the period of the square lattice.)

Mach–Zender Interferometers

The foundation of many optical device designs, from modulators to optical logic to switching, is the well-known Mach–Zender interferometer (Saleh & Teich, 1991) (Figure 1, Top).

Here, light in an incident waveguide is split into two branches and then recombined. If the two branches have the same optical path length, the recombination is in phase and light is transmitted; if their paths differ by half a period, then the recombination is out of phase and light is reflected. The relative optical path length is altered, for example, by changing the index by some Δn in the branches with a linear electro-optic (external) or Kerr (self-induced) modulation, allowing the transmission to be switched continuously from "on" to "off." By using photonic-crystal line-defect waveguides in the interferometer, one can take advantage of their low group-velocity capability to significantly lower the power and size requirements of the device by a factor of the group velocity or better (Soljačić et al., 2002b).

To attain a low group-velocity waveguide in a photonic crystal, one simple strategy is to employ a coupled-cavity waveguide (CCW) like that in Figure 1 (top). The guided band of such a waveguide has a characteristic cosine-curve dispersion relation (Figure 1, bottom) with two important properties: (i) the mid-band group velocity is low and decreases exponentially with the cavity separation, and (ii) the group-velocity dispersion (frequency derivative of the group velocity) is zero at the center of the bandwidth, minimizing signal distortion.

Because the group velocity v_g is low, when the dispersion curve $\omega(k)$ (frequency vs. wave number) is slightly shifted by altering the index n of a waveguide branch, there will be a large change in k (Figure 1, bottom), causing a corresponding large phase shift $\Delta\phi = \Delta k L$ (where L is the propagation distance). Mathematically, if the curve is shifted by $\Delta\omega \sim \Delta n\omega$, then the phase shift is $\Delta\phi \cong L\Delta\omega/v_\mathrm{g}$, inversely proportional to group velocity $v_\mathrm{g} = \mathrm{d}\omega/\mathrm{d}k$. This means that the device size L to achieve a fixed phase shift $\Delta\phi = \pi$ (with a given Δn) varies proportionally to v_g. Moreover, the power required to modulate the index is proportional to L, so the switching power is also proportional to v_g. Alternatively, one could keep L fixed and reduce Δn by a factor of v_g; for a linear (Pockels) modulation, this reduces the modulation power by v_g^2.

If the Mach–Zender interferometer is modulated by the optical waveguide signal itself, say for all-optical logic, then there is an analogous benefit: the light is compressed in time by a factor of v_g/c, causing a greater field $|\boldsymbol{E}|^2$ for the same input power. Combined with the above mentioned power savings from length reduction, this means that Kerr self-modulated devices have their power reduced proportionally to v_g^2.

Of course, the low group velocity comes at a price: the bandwidth of the waveguide is reduced proportional to v_g. In optical telecommunications systems, however, the required bandwidth is relatively small; for example, a 40 Gbit/s channel has a bandwidth $\Delta\omega/\omega \sim 1/3000$, meaning that the group velocity could potentially be lowered by several hundred times without limiting the bandwidth, with corresponding decreases in the device size and power relative to conventional waveguides.

Optical Bistability

Optical bistability is a dramatic nonlinear phenomenon that can be exploited to implement all-optical transistors, switches, logic gates, amplifiers, and other functions (Saleh & Teich, 1991). Bistability stems from nonlinear feedback combined with resonant transmission through a cavity, produces an output power that is a sharp nonlinear function of input power, and may even display a hysteresis loop (Figure 2, right).

In this context, the ability of photonic crystals to minimize both cavity modal volume V and lifetime Q (number of temporal periods for energy to decay by $\mathrm{e}^{-2\pi}$) simultaneously allows them to greatly reduce the power threshold for optical bistability, in principle even down to the milliwatt level (Soljačić et al., 2002a).

An archetypical bistable device consists of an input waveguide coupled symmetrically into an output waveguide via a resonant cavity; this is shown in a photonic crystal setting of line and point defects in Figure 2 (left). In any such system, the transmission spectrum as a function of frequency will be a Lorentzian-like curve, peaked at 100% (in the absence of other loss mechanisms) at the resonant frequency (Figure 2, middle). In order to achieve bistability, one must include nonlinear feedback: the index (and thus the frequency) of the cavity depends on the field strength (e.g., via a Kerr nonlinearity $\Delta n \sim |\boldsymbol{E}|^2$). Furthermore, one must operate at a frequency ω_0 that lies *below* the linear resonant frequency ω_res. This combination, for continuous-wave (CW) sources, results in the bistable power-response curve shown in Figure 2 (right), which here includes a hysteresis. There are two stable system states for input powers in the bistable region (between the dashed vertical lines), and which one is realized depends upon whether one started from low or high power. (The middle, dashed, branch of the "S" curve is unstable.)

Intuitively, as the input power grows, the increasing index due to the nonlinearity will lower the resonant frequency through ω_0, as depicted by the dashed line in Figure 2 (middle), causing a rise and fall in transmission. This simple picture, however, is modified by feedback: as one moves into the resonance, coupling to the cavity is enhanced (positive feedback), creating a sharper "on" transition; and as one moves out of the resonance, the coupling is reduced (negative feedback), causing a delayed "off" transition.

The power threshold for the onset of bistability depends upon the power required to shift the cavity index sufficiently, which in turn depends upon the nonlinearity of the materials and the field strength $|\boldsymbol{E}|^2$ inside the cavity for a given input power. This field strength is inversely proportional to the modal volume V and is proportional to the lifetime Q (over which time the field builds up in the cavity). On the other hand, the required index shift of the cavity is proportional to the frequency width $1/Q$ of the Lorentzian transmission spectrum. Ultimately, therefore, the threshold power is proportional to V/Q^2; these simple arguments are confirmed by a more detailed analytical

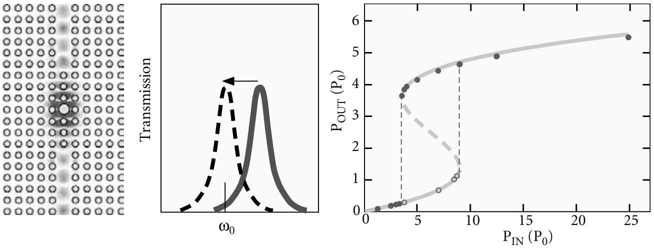

Figure 2. (Left) 100% (peak) resonant transmission from an input to an output waveguide through a cavity, formed by line and point defects, respectively, in a photonic crystal. Shaded regions indicate alternating positive/negative fields. (Middle) Lorentzian transmission spectrum (solid curve) of the linear resonant system. Increasing the power will nonlinearly shift the resonance curve (dashed) towards the operating frequency ω_0. (Right) Bistable transmission curve of output vs. input power resulting from the resonant transmission plus nonlinear feedback. The curve is an analytical theory and the dots are numerical calculations, with the vertical dashed lines indicating the region of hysteresis—for open dots the power was increased from a low value, while for closed dots the power was decreased from a high value.

theory that accurately predicts the bistability curve from the cavity characteristics (Soljačić et al., 2002a). Unlike traditional cavities such as ring resonators, photonic crystals impose no tradeoff between V and Q—the lifetime Q can be increased arbitrarily (up to the required signal bandwidth) while V is maintained near its minimum of $\sim(\lambda/2n)^3$, where n is the index of refraction. Indeed, in the example system of Figure 2, assuming reasonable material parameters and a $Q = 4000$ determined by the telecommunications bandwidth, one obtains a theoretical operating power of only a few milliwatts.

We conclude by presenting an analytical formula, derived from coupled-mode theory and perturbation theory, for the CW bistability relation in Figure 2 (right). The input/output power relation is given by

$$\frac{P_{\text{out}}}{P_{\text{in}}} = \frac{1}{1 + \left(P_{\text{out}}/P_0 - \delta\right)^2},$$

where δ is the frequency-detuning parameter $\delta \equiv 2Q(\omega_0 - \omega_{\text{res}})/\omega_{\text{res}}$ and P_0 is the characteristic power for bistability:

$$P_0 \equiv \frac{1}{\kappa Q^2(\omega_{\text{res}}/c)^{d-1}\max(n_2)}.$$

Here, d is the dimensionality of the system, n_2 is the Kerr coefficient (index change per unit intensity of light), and κ ($\sim 1/V$) is a dimensionless, scale-invariant "nonlinear feedback parameter" quantifying the concentration of the cavity field $\boldsymbol{E}$ in the nonlinear material.

Steven G. Johnson, Marin Soljačić, and
J.D. Joannopoulos

See also **Hysteresis; Nonlinear optics**

Further Reading

Ashcroft, N.W. & Mermin, N.D. 1976. *Solid State Physics*, Philadelphia: Holt Saunders

Joannopoulos, J.D. Meade, R.D. & Winn, J.N. 1995. *Photonic Crystals: Molding the Flow of Light*, Princeton, NJ: Princeton University Press

John, S. 1987. Strong localization of photons in certain disordered dielectric superlattices. *Physical Review Letters*, 58: 2486–2489

Johnson, S.G. & Joannopoulos, J.D. 2002. *Photonic Crystals: The Road from Theory to Practice*, Boston: Kluwer

Rayleigh, Lord (Strutt, W.J.). 1887. On the maintenance of vibrations by forces of double frequency, and on the propagation of waves through a medium endowed with a periodic structure. *Philosophical Magazine*, 24: 145–159

Saleh, B.E.A. & Teich, M.C. 1991. *Fundamentals of Photonics*, New York: Wiley

Soljačić, M., Ibanescu, M., Johnson, S.G., Fink, Y. & Joannopoulos, J.D. 2002a. Optimal bistable switching in non-linear photonic crystals. *Physics Review E Rapid Communincations*, 66: 055601(R)

Soljačić, M., Johnson, S.G., Shanhui Fan, Ibanescu, M., Ippen, E. & Joannopoulos, J.D. 2002b. Photonic-crystal slow-light enhancement of non-linear phase sensitivity. *Journal of the Optical Society of America B*, 19: 2052–2059

Yablonovitch, E. 1987. Inhibited spontaneous emission in solid-state physics and electronics. *Physical Review Letters*, 58: 2059–2062

Yariv, A., Xu, Y., Lee, R.K, & Scherer, A. 1999. Coupled-resonator optical waveguide: a proposal and analysis. *Optics Letters*, 24: 711–713

PIECEWISE LINEARITY

See **Ratchets**

PINNING TO LATTICE

See **Peierls barrier**

PITCHFORK BIFURCATION

See **Bifurcations**

PLASMA SOLITON EXPERIMENTS

Comprising a very large number of charged particles within a confined volume, a plasma is a convenient laboratory facility in which both linear and nonlinear phenomena can be experimentally investigated. The overall plasma is electrically neutral in that the density of positive particles is equal to the density of negative particles. If the negative particles are electrons, this is called a two-component or normal plasma. If a certain fraction ε of the electrons is replaced with negative ions, this is called a three-component or a negative ion plasma.

Washimi and Taniuti (1966) were the first to demonstrate that the evolution of perturbations of the charge density in a normal plasma can be described by the nonlinear fluid equations for the ions. This description includes a Boltzmann assumption for the electrons and Poisson's equation to reflect the local charge nonneutrality in a density perturbation that can be described with the Korteweg–de Vries (KdV) equation,

$$\frac{\partial \psi}{\partial t} + \psi^{\nu}\frac{\partial \psi}{\partial x} + \beta\frac{\partial^3 \psi}{\partial x^3} = 0. \qquad (1)$$

Here the dependent variable ψ represents the perturbations in the ion density, ion velocity, or the electric potential; β is a constant; and the parameter $\nu = 1$.

In 1984, Watanabe showed that perturbations in a negative ion plasma can also be described by the same equation if the fraction ε has certain values. In particular, if this parameter is very large, then the negative ions predominate and rarefactive solitons evolve from a negative ion density perturbation. Because the mass of the negative ions could be comparable with the mass of the positive ions, he found that the parameter ε had a critical value ε_c at which the derivation led to a modified Korteweg–de Vries (mKdV) equation with the parameter $\nu = 2$. Both of these equations describe solitons that propagate in one direction.

The first extension to include effects of higher dimensions was performed in 1970 by Kadomtsev and Petviashvili, who included weak effects in a direction that was perpendicular to the dominant direction of propagation of the ion acoustic soliton. This equation is now called the Kadomtsev–Petviashvili (KP) equation for $\nu = 1$ in a normal plasma and the modified Kadomtsev–Petviashvili (mKP) equation for $\nu = 2$ in a negative ion plasma; thus

$$\frac{\partial}{\partial x}\left(\frac{\partial \psi}{\partial t} + \psi^{\nu}\frac{\partial \psi}{\partial x} + \beta\frac{\partial^3 \psi}{\partial x^3}\right) + \frac{\partial^2 \psi}{\partial y^2} = 0\,. \qquad (2)$$

Both of these equations have certain predictions that have been experimentally verified in a plasma.

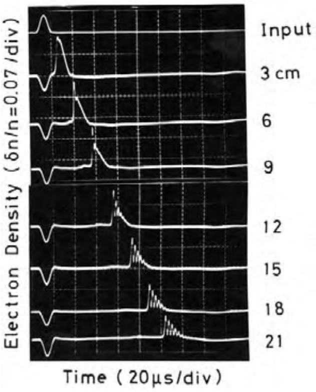

Figure 1. Evolution of a positive ion perturbation in a normal two component plasma. The pictures are taken at increasing distances from the source and illustrate the evolution of a burst of ions into a number of KdV solitons.

Laboratory experiments require the creation of a large volume of uniform collisionless plasma in which localized density perturbations are launched and movable probes monitor the evolution of the perturbation. Collisions between the particles are reduced with the evacuation of the chamber to a low pressure. Typically, a gas such as argon is inserted into the chamber and ionized to create a plasma with a volume of approximately $1\,\mathrm{m}^3$. The density perturbations are created by applying a voltage signal to a fine-mesh grid or by introducing a charge density perturbation from one plasma into a second plasma. This is called a double-plasma (DP) machine and plasma solitons were first observed in a DP machine (Ikezi et al., 1970).

The spatial and temporal evolution of a compressive density perturbation in a normal plasma is illustrated in Figure 1. As this perturbation moves in the plasma, a number of solitary waves emerge for which the following KdV soliton properties have been observed: the product of the soliton amplitude (ψ) times the square of its width (W) is constant, and the soliton velocity $c_s = [1 + \psi/3]c$ where c is the linear ion acoustic velocity. In addition, the nondestructive collision of two solitons with different amplitudes was verified in the initial experiment performed in a normal plasma that existed in the DP machine.

By replacing a certain fraction of the free electrons in the normal plasma with negative ions to which these free electrons become attached, it is possible to realize a negative ion plasma. As a gas such as sulfur hexafluoride has a large attachment coefficient, a negative ion plasma can be created that consists of positive argon ions and negative fluorine ions whose masses are comparable. The parameter ε can be altered to have values of $\varepsilon < \varepsilon_c$, $\varepsilon = \varepsilon_c$, or $\varepsilon > \varepsilon_c$. In the first case, the positive ions are compressed and the normal KdV soliton is excited. KdV solitons are also excited in the third case due to compression of the negative ions. In the second case, mKdV solitons have been excited and their properties have been verified. In particular, it was observed that the product of ψW^2 is constant. The first experimental detection of solitons in a negative ion plasma was in a DP machine (Ludwig et al., 1984; Nakamura & Tsukabayashi, 1984), and both solitons were later detected in a negative ion plasma using the grid excitation mechanism (Cooney et al., 1991).

Two solitons that propagate in a normal plasma or in a negative ion plasma but in directions that are not collinear can still interact and are described by the KP equation. In particular, Miles (1977) noted that at a particular angle, the amplitude of the soliton after such an interaction would be greater than the sum of the amplitudes of the two solitons that preceded the interaction. If the amplitudes of these two initial solitons were equal, the amplitude of the new soliton would be 4 times this amplitude. The critical amplitude and the amplitude enhancement predicted from this resonant interaction was first experimentally examined in a normal plasma (Ze et al., 1979). In the negative ion plasma, an amplitude enhancement of two was anticipated and has been experimentally verified (Nakamura et al., 1999).

A laboratory plasma has been found to be a convenient venue in which several of the fundamental properties of solitons that are described with the KdV, mKdV, KP, or mKP equations can be experimentally studied and verified. A summary of other experiments has recently appeared (Lonngren, 1998).

Karl E. Lonngren and Yoshiharu Nakamura

See also **Kadomtsev–Petviashvili equation; Korteweg–de Vries equation; Multidimensional solitons; Nonlinear plasma waves; Nonlinear Schrödinger equations**

Further Reading

Cooney, J.L., Gavin, M.T. & Lonngren, K.E. 1991. Experiments on Korteweg–de Vries solitons in a positive ion-negative ion plasma. *Physics of Fluids* B, 3: 2758–2766

Ikezi, H., Taylor, R.J. & Baker, D.R. 1970. Formation and interaction of ion acoustic solitons. *Physical Review Letters*, 25: 11–14

Kadomtsev, B.B. & Petviashvili, V.I. 1970. On the stability of solitary waves in weakly dispersing media. *Soviet Physics Doklady*, 15: 539–541

Lonngren, K.E. 1998. Ion acoustic soliton experiments in a plasma. *Optical Quantum Electronics*, 30: 615–630

Ludwig, G.O., Ferreira, J.L. & Nakamura, Y. 1984. Observation of ion acoustic rarefaction solitons in a multicomponent plasma with negative ions. *Physical Review Letters*, 52: 275–278

Miles, J.W. 1977. Resonantly interacting solitary waves. *Journal of Fluid Mechanics*, 79: 171–179

Nakamura, Y., Bailung, H. & Lonngren, K.E. 1999. Oblique collision of mKdV ion-acoustic solitons. *Physics of Plasmas*, 6: 3466–3470

Nakamura, Y. & Tsukabayashi, I. 1984. Observation of modified Korteweg–de Vries solitons in a multicomponent plasma with negative ions. *Physical Review Letters*, 52: 2356–2359

Washimi, H. & Taniuti, T. 1966. Propagation of ion acoustic solitary waves of small amplitude. *Physical Review Letters*, 17: 996–998

Watanabe, S. 1984. Ion acoustic solitons in plasma with negative ions. *Journal of the Physical Society of Japan*, 53: 950–956

Ze, F., Hershkowitz, N., Chan, C. & Lonngren, K.E. 1979. Inelastic collision of spherical ion acoustic solitons. *Physical Review Letters*, 42: 1747–1750

PLASMA TURBULENCE

See **Nonlinear plasma waves**

PLASTIC DEFORMATION

See **Frenkel–Kontorova model**

PLUME DYNAMICS

Plumes and jets are naturally and frequently occurring transport phenomena arising in a variety of settings, ranging from dry convecting atmospheric motion on hot days through explosive volcanic eruptions, for example, the 1915 eruption of Lassen Peak shown in Figure 1. The fluid dynamical purpose of a plume is dynamic equilibration of a localized unstable distortion of the fluid density, which results in vertical, coherent motion of a parcel of fluid seeking an equilibrium density. Viscosity couples and draws fluid along with the parcel on its voyage (turbulent entrainment). If the parcel is miscible with the ambient fluid, turbulent mixing will result, accelerating the equilibration process.

A further complication is that ambient fluids typically develop stable stratifications in which the fluid density is higher at the bottom. Such is the case with the Earth's atmosphere, whose density drops to zero in outer space, and the steady-state density profile is merely the thermodynamic response of an air layer under gravitational compression. Equally interesting stratification processes occur with much sharper gradients in convective boundary layers and in the thermoclines found in lakes and oceans. In these situations, stable transitions from a high-density bottom fluid layer to a low-density upper fluid layer may occur across a very sharp, nearly interfacial, layer. Typical stratifying agents include localized high temperature and/or concentration gradients (e.g., salt in the ocean). The modifications introduced by such layers can be both naturally dramatic and socio-economically challenging. The discharge of pollutants into the atmosphere, lakes, and oceans

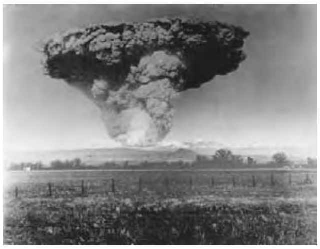

Figure 1. May 22, 1915, eruption of Lassen Peak, taken 50 miles away in Anderson, California, by photographer, R.I. Meyers. (Thanks to Cari Kreshak and Lassen Volcanic National Park for providing the high quality image.)

is frequently accompanied by trapping phenomena directly attributed to the formation of such stable density layers (thermal inversions), in which the discharged pollutants are confined away from mixing flows and may lead to hazardous air and water quality.

Plume Mixing and Entrainment: Modifying the Large Scale Observables

A light plume of fluid in a constant density environment is expected to rise continually until the Archimedian buoyancy force is reduced through mixing of the plume with the ambient to levels at which viscous balances occur. The complete evolution requires, at minimum, the solution of the Navier–Stokes equations, with an evolving density anomaly (the plume) allowed to mix with the ambient. The mixing is turbulent, and the computational simulations of these nonlinear partial differential equations are both difficult, and necessary in making first principle predictions. Modelers have turned to alternative, somewhat ad hoc, yet nonetheless fundamental attempts to describe the evolution with fewer degrees of freedom than the complete fluid equations. As discussed below, the pioneering work of Morton et al. (1956) utilized an entrainment hypothesis with a single entrainment coefficient in an attempt to describe jet (plume) profiles by reduced, nonlinear ordinary differential equations (involving only a few degrees of freedom).

The plume dynamics in stratified fluids are dramatically different (Morton et al., 1956; Morton, 1967; Turner, 1995). In such a situation, the buoyancy of a plume of light fluid is strongly height-dependent, and an initially (low altitude) light fluid parcel may well rise to a height at which a buoyancy reversal occurs and the parcel becomes neutrally buoyant. Such a situation was originally noted by Morton et al. to cause an arrestment of vertical jets of light fluid. Their experiments and modeling for a fluid with a linear stratification (linearly decreasing density with increasing height) indeed show vertical jets arresting. The implications for functioning smokestacks, modeling of volcanic plumes (Sparks et al., 1997), and the mixing of oceanic pollutants (Fischer et al., 1979) is clear.

When the density transitions sharply between two distinct values, one finds mixing between low-density jet (or plume) fluid and the ambient fluid, which may dramatically affect the large scale observables. Experiments, performed in the UNC Applied Mathematics Fluid Lab, further exhibit the need for improved modeling that is specifically designed to better understand turbulent mixing and entrainment.

Figure 2 shows two vertical jets, fired at approximately the same volumetric flow rate (roughly 0.2 gal/min) into two identically stratified fluid tanks with a prepared sharp transition from 1.06 g/cm^3 at the bottom to 1.015 g/cm^3 using varying salt concentrations, with a transition of approximately 1 in, thickness, centered around the 14 in. tick on the tape. The left jet fluid is a gauge oil, with density 0.8 g/cm^3 (lighter than all ambient tank fluid). Recall that oil and water do not mix. The right jet fluid is a dyed alcohol–water mixture, with density 0.8 g/cm^3, also (initially) lighter than everything in the tank. In this case, the alcohol-water mixture may mix with the ambient fluid.

Observe the striking difference in large-scale observables: The nonmixing case penetrates clear to the free surface, whereas the mixing case does not penetrate, but forms, at altitudes in the vicinity of the sharp density transition, a cloud. The alcohol jet, fired in nonstratified cases of either 1.06 g/cm^3 or 1.015 g/cm^3 constant density tanks will reach the free surface at these flow rates and does not form a cloud, which demonstrates the powerful effect that an ambient sharp stratification can have upon plume dynamics, and the effect of the turbulent mixing and entrainment. There has been considerable work on developing plume models for studying the types of behavior shown with the alcohol jet following the original work of (Morton et al., 1956; Sparks et al., 1997; Caulfield & Woods, 1998), and some attempts have been directed at the multi-phase aspects of the oil jet example (Asaeda & Imberger, 1993; Socolofsky et al., 2001). A successful and complete modeling approach handling a full range of cases in which the mixing properties between jet fluid and ambient fluid may be continuously varied is an open challenge.

Dynamic Plumes and Solid Wall Interactions: Transient Levitation of Falling Bodies

As an extreme example in which the injected quantity cannot mix with the ambient fluid, consider recently obtained experimental results concerning the motion of falling bodies through stratified fluids (similar to

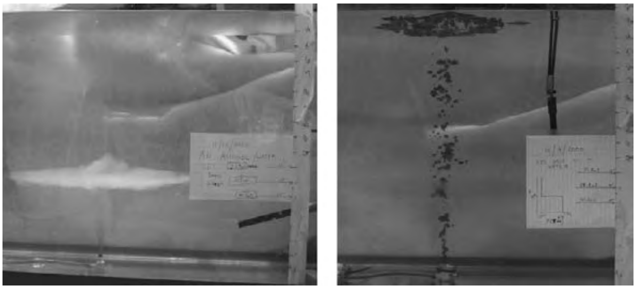

Figure 2. Vertical buoyant jets through a strong stable density step: Left is oil (0.8 g/cm^3), right is alcohol-water mixture (0.8 g/cm^3). (Thanks to former UNC undergraduates Ryan McCabe and Daniel Healion for assistance with the experiment.)

the tank setup in Figure 2) (Abaid, Adalsteinsson, Agyapong, McLaughlin, 2004). This study has focused upon the effect of self-generated plumes upon the falling body and has documented situations in which a falling body may generate a dynamic plume that through hydrodynamic coupling, may temporarily arrest the body. Of course, any body moving through a fluid experiences a hydrodynamic drag (which sets terminal velocities of falling bodies) in which the viscous boundary condition of vanishing fluid flow at the solid boundary necessarily drags a blob of ambient fluid along the moving body. In a constant density fluid, there is no potential energy cost associated with moving such a parcel of ambient fluid vertically. However, in strongly stratified fluids, a parcel of fluid moved from one altitude to another may develop a potential energy (buoyancy), as when the body falls through a sharp density transition layer. The momentum of the attached blob of fluid thrusts it into the lower (heavier) fluid, at which point the blob becomes a density anomaly and rises sharply. This motion in turn drags the falling body along with it.

Figure 3 shows three montages of a descending sphere at uniformly spaced times. The (5 mm radius) sphere in this case has a density of 1.04 g/cm^3 and is falling in a stratified tank whose top is fresh water (0.997 g/cm^3) and whose bottom is salt water (1.039 g/cm^3), again with a transition thickness of approximately 1 in. The top montage demonstrates the arrest and transient rise of the initially falling bead, and subsequent return to slow descent, each image uniformly spaced 1.5 s apart. The bead ultimately comes to rest at the tank bottom. The middle montage is the same as the top, only uniformly spaced at 0.1 s intervals. The lower montage has the same time sequence as the middle row, only focusing upon the shadow on the back of the tank, which highlights the entrained, plume-forming fluid.

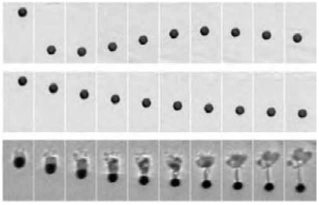

Figure 3. Top: Digital snapshots of bead position on uniform 1.5 s intervals, Middle: uniformly spaced on 0.1 s intervals, Bottom: shadowgraph depicting the dynamic plume on same time interval as middle row (Abaid, Adalsteinsson, Agyapong, McLaughlin, 2004). (Thanks to David Adalsteinsson for help with formatting the collage in his DataTank program and thanks to former UNC undergraduate Nicole Abaid for assistance with the experimental effort.)

The nature of this phenomenon is both nonlinear and dynamic. The nonlinear effect of such plumes upon the motion of solid bodies has been incorporated in a reduced system of ordinary differential equations in which the drag law for the falling body is modified to account for the dynamics of the plume which may modify the relative velocity of the falling sphere (Abaid, Adalsteinsson, Agyapong, McLaughlin, 2004). To describe the detailed dynamics of such transient plumes is quite difficult. Historically, there has been more success in the modeling of plume geometries under steady-state geometries. In pioneering work, Morton, Turner, and Taylor (Morton et al., 1956; Morton, 1967; Turner, 1995) were the first to model maintained plumes using an entrainment hypothesis which has become a standard in many fields (Fisher et al., 1979; Sparks et al., 1997).

Steady Plume Models in Stratified Environments

In 1956, Morton, Turner, and Taylor introduced what has become the standard maintained plume models for the shape of jet plumes (plumes emanating from a maintained source of buoyancy and momentum) (Morton et al., 1956; Morton, 1967; Turner, 1995; Fischer et al., 1979; Sparks et al., 1997; Socolofsky et al., 2001). The entrainment hypothesis assumes that the rate of inflow of diluting, ambient fluid is proportional to the vertical velocity of the jet along its centerline. Much empirical data has been collected exploring the exact dependence of the constant of proportionality upon the various physical parameters describing the jet configuration (stratified profile, jet speed, jet fluid density). A considerable effort since the original work of Morton et al. has addressed the many algebraic fits for the entrainment coefficient as a function of Richardson numbers, etc. (Fischer et al., 1979; Turner, 1995; Socolofsky et al., 2001).

Armed with this entrainment assumption, Morton et al. (1956) developed the following system of nonlinear ordinary differential equations, the solution of which yields the jet (plume) profile in steady state:

$$\frac{\mathrm{d}(b^2w)}{\mathrm{d}z} = 2\alpha bw, \tag{1}$$

$$\frac{\mathrm{d}(b^2w^2)}{\mathrm{d}z} = 2g\lambda^2b^2Q, \tag{2}$$

$$\frac{\mathrm{d}(b^2wQ)}{\mathrm{d}z} = \left(\frac{(1+\lambda^2)b^2w}{\lambda^2\bar{\rho}}\right)\frac{\mathrm{d}\rho_0}{\mathrm{d}z}. \tag{3}$$

Here, the plume parameters are the center-line vertical velocity, $w(z)$, the plume radius $b(z)$, and the nondimensional plume density $Q(z)$. All are functions of the height variable z. The entrainment coefficient is α, which in neutrally stratified cases is empirically seen to be approximately 0.08 (Fischer et al., 1979; Turner, 1995). The gravitational constant is g, $\bar{\rho}$ denotes some constant reference density, and the ambient stratification is contained within the given profile $\rho_0(z)$.

This system is based on the following assumptions. First, vertical derivatives of certain horizontally averaged, low-order moments (for plume mass, momentum, and buoyancy) are simplified in terms of single point, centerline field variables. Second, radial profiles for plume vertical velocity and buoyancy are postulated in terms of "collective variables": $w(z) = \hat{w}(z)f(r/(ab(z))$ and $Q(z) = \hat{Q}(z)f(r/(cb(z)))$. Through these, the integrals defining the moments may be directly calculated, leading to the closed system of differential equations given above. For both velocity and plume density, the functional forms are taken to be Gaussians, following empirical observations (Morton, 1967; Fischer et al., 1979; Turner, 1995). The ratio of the velocity length scale a, to plume length scale c is $\lambda = c/a$ is taken to be approximately 1.2, but this must certainly vary considerably upon the mixing properties of the plume fluid with the ambient. Each of these steps involves numerous approximations, an excellent list of which may be found discussed in Chapter 9 of the text by Fischer et al., along with asymptotic solutions for limiting cases (Fischer et al., 1979).

The solution of these equations gives a rough picture for plume shapes in the environment and typically shows plumes arresting at heights below their heights of static neutral buoyancy (in the absence of any mixing) in stratified environments. A more systematic mathematical reduction of this system from the complete equations, along with a numerical simulation of the complete fluid equations for multiphase fluid flow would be valuable.

RICHARD M. MCLAUGHLIN

See also **Atmospheric and ocean sciences; Mixing; Navier–Stokes equation; Turbulence; Vortex dynamics of fluids**

Further Reading

Abaid, N., Adalsteinsson, D., Agyapong, A. & McLaughlin, R.M. 2004. An internal splash: Levitation of falling spheres in stratified fluids. *Physics of Fluids*, 16(5): 1567–1580

Asaeda, T. & Imberger, J. 1993. Structure of bubble plumes in linearly stratified environments. *Journal of Fluid Mechanics*, 249: 35–57

Caulfield, C.P & Woods, A.W. 1998. Turbulent gravitational convection from a point source in a non-uniformly stratified environment. *Journal of Fluid Mechanics*, 360: 229–248

Csanady, G.T. 1973. *Turbulent Diffusion in the Environment*, Dordrecht: Reidel

Fernando, H.J.S. 1991. Turbulent mixing in stratified fluids. *Annual Reviews of Fluid Mechanics*, 23: 455–493

Fischer, H.B., List, E.J., Koh, R.C.Y., Imberger, J. & Brooks, N.H. 1979. *Mixing in Inland and Coastal Waters*, New York: Academic Press

Larsen, L.H. 1969. Oscillations of a neutrally buoyant sphere in a stratified fluid. *Deep-Sea Research*, 16: 587–603

Morton, B.R., 1967. Entrainment models for laminar jets, plumes, and wakes. *Physics of Fluids*, 10(10): 2120–2127

Morton, B.R., Taylor, G.I. & Turner, J.S. 1956. Turbulent gravitational convection from maintained and instantaneous sources. *Proceedings of the Royal Society*, A, 234: 1–23

Socolofsky, S.A., Crounse, B.C. & Adams, E.E. 2001. Multi-phase plumes in uniform, stratified, and flowing environments. In *Environmental Fluid Mechanics–Theories and Applications*, edited by H. Shen, A. Cheng, K.-H. Wang & M.H. Teng, Reston, Va: American Society of Civil Engineers, Ch. 3

Sparks, R.S.J., Burksik, M.I., Carey, S.N., Gilbert, J.S., Glaze, L.Sl, Sigurdsson, H. & Woods, A.W. 1997. *Volcanic Plumes*, New York: Wiley

Srdic-Mitrovic, A.N., Mohamed, N.A. & Fernando, H.J.S. 1999. Gravitational settling of particles through density interfaces. *Journal of Fluid Mechanics*, 381: 175–198

Torres, C.R., Hanazaki, H., Ochoa, J., Castillo, J. & Van Woert, M. 2000. Flow past a sphere moving vertically in a stratified diffusive fluid. *Journal of Fluid Mechanics*, 417: 211–236

Turner, J.S. 1995. *Buoyancy Effects in Fluids*, Cambridge and New York: Cambridge University Press

POINCARÉ INDEX

See **Winding numbers**

POINCARÉ THEOREMS

One of the greatest of all French mathematicians, Jules Henri Poincaré (1854–1912) graduated from the École Polytechnique in Paris and later studied at the École des Mines. In 1879, he became a docteur es sciences at the University of Paris, where he was appointed as a professor in 1881. Poincaré became a member of Académie des Sciences in 1887.

A mathematical genius of rare power, Poincaré's approach to science was to solve concrete problems arising from mathematics, mechanics, and physics, rather than to present his results in a "pure axiomatic form." However, a complete understanding of Poincaré's scientific works has yet to be achieved. Outside mathematics, Poincaré is also known for his works in theoretical physics including his seminal contribution in the special theory of relativity (1904–1905) and for his works on the philosophy of science.

Many of Poincaré's papers gave birth to whole new branches of mathematics, a prime example being algebraic topology, but the matter that occupied his life throughout was the geometrical approach to nonlinear differential equations—in particular, the long-time behavior of orbits of the Newtonian N-body problem in celestial mechanics.

The qualitative approach to nonlinear dynamics, introduced by Poincaré in his seminal papers "Mémoire sur les courbes définies par une équation différentielle" (1881–1886), which focuses on orbits rather than formulas was of a geometric and global nature. This is how the qualitative theory of ordinary differential equations was born.

Studying smooth vector fields on the plane, he classified their simplest equilibria (i.e., points where the given vector field vanishes): foci, nodes, saddle points, and centers. The typical smooth planar vector field has only the first three types of equilibria, but those of a more complicated nature are not excluded.

Poincaré outlined the proof that if a half trajectory γ of the planar vector field v is confined in a compact domain K in which v is free of equilibrium points, but the whole trajectory γ is not confined in K, then K contains a closed orbit of v to which γ is asymptotically attracted (the Poincaré–Bendixson theorem in its simplest form). This type of closed orbit is called a limit cycle.

To generalize, let v be a smooth vector field on a two-dimensional compact manifold M. To each isolated equilibrium p of v, Poincaré associated an integer $\mathrm{Ind}(v, p)$, called the index of v at p, which is defined as follows. Let J be a small loop surrounding the isolated equilibrium p, and let $\Delta\theta$ be the total change of the angle θ that the vector of v makes with some fixed direction when one runs counterclockwise along the loop J. This number is independent of the choice of loop. The index of p is the number $\Delta\theta/2\pi$ which is always an integer. (The index of a focus, center, or node is $+1$, and the index of a saddle point is -1.)

If v has only a finite number of equilibria $p_1, \ldots, p_r$, then Poincaré showed that

$$\sum_{i=1}^{r} \mathrm{Ind}(v, p_i) = \chi(M), \tag{1}$$

where $\chi(M)$ is the Euler–Poincaré characteristic of M and no restrictions are imposed on the nature of equilibria $p_1, \ldots, p_r$. This result was generalized later by Heinz Hopf to the case of vector fields on compact manifolds of arbitrary dimension and is now called the Poincaré–Hopf theorem.

For a sphere, we have $\chi(M) = 2$, that is, an arbitrary smooth vector field on a two-dimensional sphere must vanish in at least one point—one cannot evenly comb the hair on a sphere! Another consequence of this theorem is that, for example, on the two-dimensional sphere, one cannot have a smooth vector field having as equilibria only two saddle points or having only three centers.

Poincaré's investigations of celestial mechanics led him to the study of Hamiltonian systems with n degrees of freedom

$$\frac{\mathrm{d}q_i}{\mathrm{d}t} = \frac{\partial H}{\partial p_i}, \quad \frac{\mathrm{d}p_i}{\mathrm{d}t} = -\frac{\partial H}{\partial q_i}, \quad i = 1, \ldots, n \tag{2}$$

with an analytic Hamiltonian function $H(q, p)$, $(q, p) \in \mathbb{R}^{2n}$. His researches in this area were summarized in his epoch-making three-volume treatise "Les méthodes nouvelles de la mécanique céleste" (1892, 1893, 1899).

The study of Hamiltonian systems close to integrable ones was called the "general problem of dynamics" by Poincaré. Specifically, he studied Hamiltonian equations (2) with a perturbed Hamiltonian of the form

$$\begin{aligned} H(q, p, \varepsilon) &= H_0(p) + \varepsilon H_1(q, p) \\ &\quad + \varepsilon^2 H_2(q, p) + \cdots , \end{aligned} \tag{3}$$

where $H_1, H_2, \ldots$ are periodic functions with respect to q of the same period. The system obtained by setting $\varepsilon = 0$ is integrable, and the phase space is foliated by n-dimensional invariant tori $p =$ const. For small nonzero ε, system (2) usually becomes non-integrable. In the case of two degrees of freedom, this means that the Hamiltonian function H is the only first integral of Equations (2) that is an analytic and uniform function of variables q, p and ε.

Poincaré's participation in a mathematical competition organized in 1885 by Oskar II, King of Sweden and Norway, led him to the discovery of homoclinic orbits and related phenomena explaining how a very complicated behavior now called dynamical chaos, occurs in nonlinear dynamical systems. In his prize-winning work (1889), Poincaré discusses mainly the restricted three-body problem, where one mass is negligible compared with the other two, executing circular Keplerian motion.

Within the three-dimensional constant-energy manifold of this problem, he considered a two-dimensional surface Π transversal to most of the orbits. An orbit starting on this surface at a point x will pierce it again for the first time at some point y. The map $x \to y$ is the Poincaré return map induced on a surface of section Π.

It was in this framework that he discovered the existence of homoclinic orbits, that is, orbits which are asymptotically attracted by some periodic orbit γ when $t \to +\infty$ and $t \to -\infty$.

During his 1912 investigations on the restricted three-body problem, Poincaré conjectured that if one has a closed plane annulus Ω bordered by two concentric circles Γ_1 and Γ_2, then any area-preserving homeomorphism ϕ of Ω, such that $\phi(\Gamma_1) = \Gamma_1$, $\phi(\Gamma_2) = \Gamma_2$ and rotating these circles in opposite directions, has in Ω at least two different fixed points. This assertion, known as Poincaré's last geometric theorem, was proved in 1913 by George D. Birkhoff and is known also as the Poincaré–Birkhoff theorem.

Inspired by the analogy between a flow induced by a vector field and a flow of an incompressible fluid, Poincaré developed a theory of integral invariants, which was later refined by Elie Cartan.

Considering time t as an independent variable, we can study Hamiltonian system (2) in the extended phase space (q, p, t). A tube of trajectories is a two-dimensional cylindrical surface formed by the segments of trajectories of the vector field defined by (2) and bounded by two disjoint smooth closed curves. According to the Poincaré–Cartan theorem, in the extended phase space (q, p, t), the action integral $\oint_\gamma (p\,\mathrm{d}q - H\,\mathrm{d}t)$ has the same value for two different closed paths γ_1 and γ_2 encircling the same tube of trajectories and lying on it.

Poincaré also proved an important property of the long-time behavior of dynamical systems. In contemporary formulation, this proof shows that for any measure-preserving mapping of a measure space with a finite total measure, almost all trajectories starting from a given subset of positive measure eventually return to it. This is known as the Poincaré recurrence theorem, and it lies at the foundations of ergodic theory.

In his works on celestial mechanics, Poincaré provided the first formal definition of asymptotic series: divergent series giving nevertheless good numerical approximations for functions they represent.

Poincaré is the founder of the concept of normal forms in the theory of ordinary differential equations and of the contemporary bifurcation theory. His work was also at the beginning of modern variational methods in mathematics, in particular of the Morse theory which strongly links mathematical analysis to geometry and topology.

Other nonlinear problems studied by Poincaré include the problem of the existence of geodesics on convex surfaces, the problem of tides, and the stability of rotating fluid bodies.

Poincaré's impact on the theory of ordinary differential equations and dynamical systems is described in the books by Birkhoff (1927), Nemytskii & Stepanov (1960), Coddington & Levinson (1955), and Guckenheimer & Holmes (1990).

JEAN-MARIE STRELCYN AND ALEXEI TSYGVINTSEV

See also **Celestial mechanics; N-body problem; Phase plane; Phase space; Recurrence**

Further Reading

Barrow-Green, J. 1997. Poincaré and the three body problem, *History of Mathematics*, vol. 11, Providence, RI: American Mathematical Society and London: London Mathematical Society

Birkhoff, G.D. 1927. *Dynamical Systems*, Providence, RI: American Mathematical Society

Coddington, E.A. & Levinson, N. 1955. *Theory of Ordinary Differential Equations*, New York: McGraw-Hill

Guckenheimer, J. & Holmes, P. 1990. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, 3rd edition, Berlin and New York: Springer

Milnor, J.W. 1965. *Topology from the Differentiable Viewpoint*, based on notes by David W. Weaver, Charlottesville, VA: University Press of Virginia; revised and reprinted, Princeton, NJ: Princeton University Press, 1997 Contains an elegant and simple proof of the Poincaré–Hopf theorem.

Nemytskii, V.V. & Stepanov, V.V. 1960. *Qualitative Theory of Differential Equations*, Princeton, NJ: Princeton University Press

Poincaré, H. 1904. L'état actuel et l'avenir de la physique mathématique. Conférence lue le 24 Septembre 1904 au Congrés d'arts et de sciences de Saint-Louis. *Bulletin des Sciences Mathématiques*, 28: 302–324; translation in *Bulletin of the American Mathematical Society*, 12 (1905–1906): 240–260, also published in Poincaré's book *La valeur de la science*, Paris: Flammarion, 1905, Chapters VII, VIII, and IX

Poincaré, H. 1905. Sur la dynamique de l'électron, *Comptes Rendus de l'Académie des Sciences*, 140: 1504–1508; also in *Œuvres*, vol. 9, pp. 489–493

Poincaré, H. 1906. Sur la dynamique de l'électron, Rendiconti del Circolo matematico di Palermo, 21: 129–176; also in *Œuvres*, vol. 9, pp. 494–550

Poincaré, H. 1916–1956. *Œuvres de Henri Poincaré*, 11 vols, Paris: Gauthier-Villars (This paper (Poincaré, 1904), which is not included in Poincaré (1916–1956), is in fact the first Poincaré's paper anticipating the special theory of relativity. The annoucement (Poincaré, 1905) and its full presentation in Poincaré (1906) is Poincaré's contribution to what is now called the special theory of relativity.)

Symposium on the Mathematical Heritage of Henri Poincaré. 1983. The Mathematical Heritage of Henri Poincaré, edited by Felix E. Browder, 2 vols., Providence, RI: American Mathematical Society

POINCARÉ–BENDIXSON THEOREM

See **Phase plane**

POISSON BRACKETS

Let M be an n-dimensional manifold (referred to as the phase space), and let f, g, and h denote analytic functions on M. A Poisson bracket of any two analytic functions on the phase space is defined as an operation which satisfies

(i) $\{\alpha f + \beta g, h\} = \alpha\{f, h\} + \beta\{g, h\}$, (linearity in the first component);
(ii) $\{f, g\} = -\{g, f\}$ (skew-symmetry);
(iii) $\{f, \{g, h\}\} + \{g, \{h, f\}\} + \{h, \{f, g\}\} = 0$ (Jacobi identity);
(iv) $\{f, gh\} = g\{f, h\} + \{f, g\}h$ (Leibniz property),

where α, β are numbers. The first two properties ensure that a Poisson bracket is a bilinear operation on M. Properties (i)–(iii) imply that the analytic functions on M form a Lie algebra with respect to the Poisson bracket.

If local coordinates z_i, $i = 1, \ldots, n$ are chosen on M, then the Poisson bracket has the coordinate representation

$$\{f, g\} = \sum_{j,k=1}^{n} J_{jk}(z) \frac{\partial f}{\partial z_j} \frac{\partial g}{\partial z_k} = (\nabla f)^T \boldsymbol{J} \nabla g, \qquad (1)$$

where $\nabla f = (\partial f / \partial z_1, \ldots, \partial f / \partial z_n)$, and the Poisson matrix $\boldsymbol{J}(z) = (J_{jk}(z))_{j,k=1}^{n}$ is a skewsymmetric square matrix, satisfying a technical condition enforced by the Jacobi identity.

Any nonconstant function C on M that Poisson commutes with all other functions on M is called a Casimir of the Poisson bracket. From (1) it follows that the existence of a Casimir requires $\boldsymbol{J}$ to be singular, and ∇C is in the null space of $\boldsymbol{J}$. Furthermore, the number of independent Casimirs is the corank of $\boldsymbol{J}$. For a Poisson bracket with r Casimirs $C_1, \ldots, C_r$, Darboux's theorem states that it is always possible to find coordinates $(q_1, \ldots, q_N, p_1, \ldots, p_N, C_1, \ldots, C_r)$ on M such that in these coordinates

$$\boldsymbol{J} = \begin{pmatrix} \mathbf{0} & \boldsymbol{I}_N & \mathbf{0} \\ -\boldsymbol{I}_N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad (2)$$

where $\boldsymbol{I}_N$ is the N-dimensional identity matrix, and $\mathbf{0}$ is the zero matrix of the appropriate dimensions. In these coordinates,

$$\{f, g\} = \sum_{j=1}^{N} \left(\frac{\partial f}{\partial q_j} \frac{\partial g}{\partial p_j} - \frac{\partial f}{\partial p_j} \frac{\partial g}{\partial q_j} \right). \qquad (3)$$

This representation of the Poisson bracket is called the canonical Poisson bracket, and the coordinates $(q_1, \ldots, q_n, p_1, \ldots, p_n)$ are called canonical coordinates.

The importance of Poisson brackets is derived from their relationship to Hamiltonian systems: let H be a function on M. Hamiltonian dynamics with Hamiltonian function H are defined on any function f on M by

$$\dot{f} = \{f, H\}. \qquad (4)$$

Using the coordinate representation (1), the Hamiltonian dynamics for the coordinates is

$$\dot{z}_j = \{z_j, H\} = \sum_{k=1}^{n} J_{jk} \frac{\partial H}{\partial z_k}, \qquad (5)$$

which reduces to the standard definition of a Hamiltonian system if canonical coordinates are used. From (4) it is clear that any function that Poisson commutes with the Hamiltonian is conserved for the Hamiltonian system defined by the Poisson bracket and the Hamiltonian H. In particular, H is conserved. Also, any Casimir is conserved. Because the conservation of the Casimirs is independent of the choice of H, they do not contain dynamical information. Rather, as is obvious from Darboux's theorem, they foliate the phase space and represent geometric restrictions on the possible motions in phase space. A Hamiltonian system can also be defined using the Hamiltonian function H and a symplectic two-form, of which $\boldsymbol{J}^{-1}$ (if it exists) is the coordinate representation (Weinstein, 1984).

As an example, consider Euler's equations of a free rigid body (Weinstein, 1984). Denote the angular momentum by (M_1, M_2, M_3) and the moments of inertia by I_1, I_2, I_3. The Poisson matrix is

$$\boldsymbol{J} = \begin{pmatrix} 0 & M_3 & -M_2 \\ -M_3 & 0 & M_1 \\ M_2 & -M_1 & 0 \end{pmatrix}. \qquad (6)$$

The Hamiltonian is $H = (M_1^2 / I_1 + M_2^2 / I_2 + M_3^2 / I_3) / 2$. The Poisson matrix has rank 2 (except at the origin), and there is one Casimir: $C_1 = M_1^2 + M_2^2 + M_3^2$.

The notion of Poisson brackets extends to infinite-dimensional phase spaces, so as to describe dynamics governed by evolution (partial differential) equations (Marsden & Morrison, 1984). In this case, the Poisson matrix $\boldsymbol{J}$ is replaced by a skew-adjoint differential operator B. If the evolution equation is first order in the dynamical variable t, this operator is scalar. Otherwise it is a matrix operator of the same dimension as the order of the evolution equation. Instead of functions on phase space, we consider functionals

$$F[\boldsymbol{u}] = \int f[\boldsymbol{u}]\, \mathrm{d}x. \qquad (7)$$

Here $\boldsymbol{u}(x, t)$ is an infinite-dimensional coordinate on the phase space, indexed by the independent variable x. The square brackets denote that $f[\boldsymbol{u}]$ depends not only on $\boldsymbol{u}$, but possibly also on its derivatives with respect to x: $\boldsymbol{u}_x, \boldsymbol{u}_{xx}, \ldots$. The limits of integration depend on the boundary conditions imposed on the evolution equation. In the above, the variable x is assumed to be one dimensional. This is extended to higher dimensions in obvious fashion.

The Poisson bracket between any two functionals on phase space is the functional given by

$$\{F, G\} = \int \frac{\delta F}{\delta u} B \frac{\delta G}{\delta u}\, \mathrm{d}x, \qquad (8)$$

where $\delta F / \delta \boldsymbol{u}$ is the variational (or Fréchet) derivative of F with respect to $\boldsymbol{u}$:

$$\frac{\delta F}{\delta \boldsymbol{u}} = \frac{\partial f}{\partial \boldsymbol{u}} - \frac{\partial}{\partial x}\frac{\partial f}{\partial \boldsymbol{u}_x} + \frac{\partial^2}{\partial x^2}\frac{\partial f}{\partial \boldsymbol{u}_{xx}} - \cdots . \tag{9}$$

The Poisson bracket defined this way satisfies properties (i), (ii), and (iv). The operator B is chosen so that the Jacobi identity (iii) is also satisfied.

A functional H on the phase space defines Hamiltonian dynamics on any functional by

$$\frac{\partial F}{\partial t} = \{F, H\} \quad \Leftrightarrow \quad \frac{\partial \boldsymbol{u}}{\partial t} = \{\boldsymbol{u}, H\} = B\frac{\delta H}{\delta \boldsymbol{u}}. \tag{10}$$

As an example, consider the Korteweg–de Vries equation $u_t = uu_x + u_{xxx}$ (Gardner, 1971; Zakharov & Faddeev, 1971). This equation with $-\infty < x < \infty$ is Hamiltonian with $B = \partial/\partial x$ and $H = \int_{-\infty}^{\infty}(u^3/6 - u_x^2/2)\,\mathrm{d}x$. Thus, the Poisson bracket is

$$\{F, G\} = \int_{-\infty}^{\infty} \frac{\delta F}{\delta u}\frac{\partial}{\partial x}\frac{\delta G}{\delta u}\,\mathrm{d}x, \tag{11}$$

and $\int_{-\infty}^{\infty} u\,\mathrm{d}x$ is its only Casimir.

The Poisson bracket formulation of a Hamiltonian system is especially significant when a quantum description of the dynamics is required. Dirac's principle of canonical quantization postulates that such a quantum description is obtained by replacing all classical quantities by their quantum mechanical operator counterparts (generalized coordinates $q \to \hat{q}$, the operation of multiplying by q, momentum $p \to -\mathrm{i}\hbar\partial/\partial q$, etc., and all Poisson brackets by commutators/$(\mathrm{i}\hbar)$). Then, in the classical limit as $\hbar \to 0$, the quantum mechanical equations reduce to classical equations, as desired by the correspondence principle.

Poisson brackets were introduced by Siméon-Denis Poisson (1809) during his investigations on perturbation theory in classical mechanics. Poisson's *Traité de mécanique* (two volumes, 1811 & 1833) were standard texts for many years.

BERNARD DECONINCK

See also **Constants of motion and conservation laws; Hamiltonian systems; Korteweg–de Vries equation; Lie algebras and Lie groups**

Further Reading

Gardner, C.S. 1971. The Korteweg–de Vries equation and generalizations. IV. The Korteweg–de Vries equation as a Hamiltonian system. *Journal of Mathematical Physics*, 12: 1548–1551

Marsden, J.E. & Morrison, P. 1984. Noncanonical Hamiltonian field theory and reduced MHD. In *Fluids and Plasmas: Geometry and Dynamics*, edited by J.E. Marsden, Providence, RI: American Mathematical Society

Poisson, S.D. 1809. Mémoire sur la variation des constantes arbitraires dans les questions de mécanique [Memoir on the variation of arbitrary constants in mechanics]. *Journal de l'École Polytechnique*, cahier XV: 266–298

Weinstein, A. 1984. Stability of Poisson–Hamilton equilibria. In *Fluids and Plasmas: Geometry and Dynamics*, edited by J.E. Marsden, Providence, RI: American Mathematical Society

Zakharov, V.E. & Faddeev, L.D. 1971. Korteweg–de Vries equation: a completely integrable Hamiltonian system. *Functional Analysis and Its Applications*, 5: 280-287

POLARITONS

A simple view of a dielectric material is that it is just an insulator made from atoms that have bound positive and negative charges. The ratio of an electric field measured in vacuum to that measured inside a dielectric material is called the dielectric constant. The opposite signs of the charges make the atoms electrically neutral, but these charges can be pulled apart by an applied electric field and they are displaced in opposite directions. If this happens, the material is said to be polarized. A metal, on the other hand, appears to be more complicated; it is a conductor and a solid-state plasma. Often called the fourth state of matter, a plasma is an electrically neutral assembly of separated electrons and positively charged ions. A metal is precisely like this but, unlike a gaseous plasma, it has a highly mobile "sea" of electrons carrying charges that are exactly balanced by a virtually immobile set of positive charges residing on the ionic background making up the crystal lattice. The same can be said for a heavily doped semiconductor. In spite of the plasma nature of a metal, its interaction with electromagnetic waves is well modeled by a dielectric function that turns out to be frequency-dependent (Kittel, 1995).

To understand the physical interaction between an electromagnetic wave and a dielectric, it is instructive to look at a (deceptively) simple model of how electrons behave in a metal or a heavily doped semiconductor when exposed to an electromagnetic wave. The core, realistic assumption in this model is that the electrons are highly mobile and free to move against an immobile ionic background. Suppose that the electric field carried by the electromagnetic wave is $\boldsymbol{E} = (0, 0, E)$, with a time variation $\mathrm{e}^{\mathrm{i}\omega t}$, where ω is an angular frequency. For a free electron with mass m, charge $-e$, and velocity $\boldsymbol{v}$, moving in one dimension, the equation of motion becomes

$$\mathrm{i}\omega m\boldsymbol{v} = -e\boldsymbol{E}. \tag{1}$$

If N is the number density of the electrons, this particle motion creates a current density with a magnitude $\boldsymbol{j}$,

$$\boldsymbol{j} = -Ne\boldsymbol{v} = \mathrm{i}\left(\frac{Ne^2}{\omega m}\right)\boldsymbol{E} = \mathrm{i}\omega\boldsymbol{P}, \tag{2}$$

where $\boldsymbol{P}$ is called the polarization. The latter should be thought of as being excited as a dipole moment per unit volume by the electromagnetic wave. Defining the electric displacement vector of free space as $\boldsymbol{D}$, and the magnetic field as $\boldsymbol{H}$, the appropriate Maxwell equation to incorporate (2) is

$$\begin{aligned}\operatorname{curl}\boldsymbol{E} &= \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{j} = \mathrm{i}\omega\varepsilon_0\boldsymbol{E} - \mathrm{i}\frac{Ne^2}{\omega m}\boldsymbol{E}\\ &= \mathrm{i}\omega\varepsilon_0\left[1 - \frac{\omega_\mathrm{p}^2}{\omega^2}\right]\boldsymbol{E} \qquad (3)\end{aligned}$$

in which t is time and ε_0 is the permittivity of free space. Equation (3) shows that a metal is neatly modeled by the frequency-dependent relative permittivity

$$\varepsilon(\omega) = 1 - \frac{\omega_\mathrm{p}^2}{\omega^2}, \qquad (4)$$

where $\omega_\mathrm{p}^2 = \left(\sqrt{Ne^2/\varepsilon_0 m}\right)$ is called the plasma frequency. For this to also be a useful description of a semiconductor, or to take into account that the ionic background of a metal may contribute a frequency-independent permittivity, all that is necessary is to replace unity in Equation (4) with a quantity ε_L, which can be much greater than unity for semiconductors. This model gives good agreement with experiment, and it also serves to introduce the polariton concept. Excitations other than those of a free-electron gas are easy to capture, once this basic example is accepted. In a one-dimensional system stretching along the x-axis, the electrons move a distance $\boldsymbol{x} = e\boldsymbol{E} / (\omega^2 m)$ and the polarization is $\boldsymbol{P} = -\,Ne\boldsymbol{x}$.

If a plane electromagnetic wave has a wave vector $\boldsymbol{k}$ and (4) is the dielectric function, then

$$\begin{aligned}\boldsymbol{k}\times\boldsymbol{E} &= \omega\mu_0\boldsymbol{H}, \qquad (5)\\ \boldsymbol{k}\times\boldsymbol{H} &= -\omega\varepsilon_0\varepsilon(\omega)\boldsymbol{E},\end{aligned}$$

where μ_0 is the magnetic permeability of free space. As a specific example, let the propagation be along the x-axis with wave vector $\boldsymbol{k} = (k, 0, 0)$ and let the electric field vector be $\boldsymbol{E} = (E_x, 0, E_z)$. The following modes are then implied (Boardman, 1982) by (5):

$$\text{Transverse}: \quad (\omega^2 - \omega_\mathrm{p}^2 - c^2k^2)E_z = 0, \qquad (6)$$

$$\text{Longitudinal}: \quad (\omega^2 - \omega_\mathrm{p}^2)E_x = 0, \qquad (7)$$

where $c^2 = 1 / (\varepsilon_0\mu_0)$ is the velocity of light in vacuum and there are two independent solutions. The longitudinal mode is sustained when $E_x \neq 0$ and $\omega = \omega_p$. For this mode, the electrons move collectively in an oscillation that resembles the wobbling of a jelly.

At this stage, it should be pointed out that physicists are prone to call things by "ons" (Walker & Slack, 1970). All this started with the word *electron*. Unfortunately, the latter word does not, in itself, explain too much about the "on" part of the word, which is derived from the Greek word for "amber",—a fact that Benjamin Franklin would readily appreciate. Nevertheless, the electron is recognized to have a particle nature so physicists have become used to defining an "on" in association with an excited field, such as an electromagnetic wave or the jelly-like plasma oscillation. Thus, a *photon* is a particle, or quantum of energy, of light and the quantum of the wobbling electronic jelly, or plasma oscillation, is called the *plasmon*. The name *polariton* (Hopfield, 1958) is reserved for the quantum of an excitation in which electromagnetic waves cannot disentangle themselves from excitations such as plasmons.

The coupling of an electromagnetic field to the polarization excitations associated with electrons can be seen from the transverse solution (6). This shows that a finite value for E_z requires the dispersion equation to be

$$\omega^2 = \omega_\mathrm{p}^2 + c^2k^2 \qquad (8)$$

Furthermore, this equation shows that the excitation is mixed because it involves not only the plasma frequency but also the velocity of light. It is for this reason that the quantum of this photon-plasmon field is called a polariton, as indicated earlier. The total energy is shared over the system because the mode has both plasmon and photon content, as can be seen by looking at the low and high wave number limits of (8). As $\boldsymbol{k} \to 0$, the frequency tends towards the plasma frequency, and the plasmon content grows at the expense of the photon contribution. The opposite is true as $\boldsymbol{k} \to \infty$, and the wave becomes a pure electromagnetic wave. The "on" labeling is not complete, however, unless the type of polariton is identified, so, in this case, they are referred to as plasmon-polaritons. This labeling is a reminder that only plasmons have been used here to introduce the concept of polaritons. The description is generic, however, and there are many types of polaritons. All that is necessary to generalize the discussion given here is to decide which polarization excitation the photons couple with and capture the material polarization properties through an effective permittivity. Well-known examples involve plasmons, excitons, and phonons.

After several decades of research, polaritons are still attracting attention (Baher & Cottam, 2003), in both guided and surface wave form, and the study of nonlinear polaritons in particular has opened new horizons (Boardman & Egan, 1985). The inclusion of nonlinearity means extending the permittivity to depend upon the amplitude of the electromagnetic wave through a functional dependence upon E, such as

$$\varepsilon(\omega, E) = \varepsilon(\omega) + \varepsilon^{\mathrm{NL}}(|E|) \qquad (9)$$

Here $\varepsilon(\omega)$ is the relative permittivity for the excitation of choice, such as phonons, plasmons, or excitons. To reduce (9) to a tractable form, a Kerr-type of optical nonlinearity can be adopted, in which the total permittivity is expressed as

$$\varepsilon(\omega, E) = \varepsilon(\omega) + a\,|E|^2, \qquad (10)$$

where a is a nonlinear coefficient.

This type of nonlinear process is not only popular in the literature but is genuinely applicable when the nonlinearity is well away from resonance and saturation. Immediately an interesting feature appears because the form of (10) permits an investigation of the kind of transverse electric (TE) waves that have no linear ($a \to 0$) limit. This dramatic conclusion culminates in the generation of an entirely new type of polariton. The TE waves can be guided or surface waves and for this discussion are assumed to be traveling along the x-axis. The field components are, therefore,

$$\begin{aligned} \boldsymbol{E} &= [0, E_y(k, \omega, z), 0]\mathrm{e}^{\mathrm{i}(kx-\omega t)}, \\ \boldsymbol{H} &= [H_x(k, \omega, z), 0, H_z(k, \omega, z)]\mathrm{e}^{\mathrm{i}(kx-\omega t)}. \end{aligned} \tag{11}$$

Substituting (11) into Maxwell's equations produces the following equations (Boardman et al., 1991):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}z}E_y(k, \omega, z) &= -\mathrm{i}\omega\mu_0 H_x(k, \omega, z), \\ kE_y(k, \omega, z) &= \omega\mu_0 H_z(k, \omega, z), \\ \frac{\mathrm{d}}{\mathrm{d}z}H_z(k, \omega, z) &= \mathrm{i}kH_z(k, \omega, z) \\ &\quad +\mathrm{i}\omega\varepsilon_0\varepsilon(\omega, \mid E_y \mid^2)E_y(k, \omega, z). \end{aligned} \tag{12}$$

In general, E_y is complex so that $E_y = E(k, \omega, z)\,\mathrm{e}^{\mathrm{i}\phi(k,\omega,z)}$. Using $E \equiv E(k, \omega, z)$, $\phi \equiv \phi(k, \omega, z)$, and $\varepsilon \equiv \varepsilon(\omega)$ in (12) produces the following basic nonlinear equations:

$$\begin{aligned} &\frac{\mathrm{d}^2}{\mathrm{d}z^2}E - E\left(\frac{\mathrm{d}\phi}{\mathrm{d}z}\right)^2 + \left(\frac{\omega^2}{c^2}\varepsilon - k^2\right)E + \frac{\omega^2}{c^2}aE^2 = 0, \\ &\frac{\phi}{\mathrm{d}z} = \frac{K}{E^3}. \end{aligned} \tag{13}$$

The last equation expresses the conservation of energy flux along the z-axis, which is perpendicular to the propagation direction, and K is a constant of the integration process. The z-component of the time-averaged Poynting vector is

$$\langle S\rangle_z = -\frac{1}{2}\mathrm{Re}(E_y^* H_x) = \frac{c^2}{2\omega^2}\varepsilon_0 K. \tag{14}$$

Hence, $K = 0$ implies $\langle S\rangle_z = 0$. Finally,

$$\left(\frac{\mathrm{d}E}{\mathrm{d}z}\right)^2 + \left(\frac{\omega^2}{c^2}\varepsilon(\omega) - k^2\right)E + \frac{a}{2}\frac{\omega^2}{c^2}E^4 = C, \tag{15}$$

where C is another constant of integration. For a semi-infinite medium, $C = 0$, otherwise $C \neq 0$.

Consider now a $C = 0$ case describing a semi-infinite nonlinear medium interfaced to a semi-infinite linear medium. As it is well known that linear surface plasmon–polaritons are transverse magnetic (TM) polarized waves, the nonlinearity has added the extra functionality of making it possible to have TE-polarized surface plasmon-polariton waves. The field profile of these TE nonlinear waves is readily obtained from (15), since it factorizes when $C = 0$. The solution is

$$\begin{aligned} E = \frac{c}{\omega}\sqrt{\frac{2}{a}}&\sqrt{\left(k^2 - \frac{\omega^2}{c^2}\varepsilon(\omega)\right)} \\ &\times \operatorname{sech}\left[\sqrt{k^2 - \frac{\omega^2}{c^2}\varepsilon(\omega)}\right](z - z_0). \end{aligned} \tag{16}$$

This is a self-focused (self-guided) beam that has a peak in the nonlinear medium below the interface. This nonlinear surface wave looks exactly like a spatial soliton. Furthermore, there is no linear limit for the nonlinear TE-polarized surface plasmon-polaritons. The frequency dependence for guided modes is more complex and can have a linear limit.

Polaritons are being studied in a wide variety important of contexts. Some examples include nonlinear exciton-polariton dynamics in the experimental study of semiconductor microcavities (Savvides et al., 2000), two-dimensional electron systems in quantizing magnetic fields (Beletskii & Bludov, 2002), and plasmon-polaritons in dielectric films (Baher & Cottam, 2003). The idea that underpins the polariton concept can be put to dramatic use in an unusual way. This is possible because to create polaritons, photons must associate themselves with the quanta or the quasiparticles of collective excitations, be they plasmons, phonons, or related excitatons. The polariton is a genuine collaborative effect between the photons and particles that are acting in a prepared organized manner, which means that the particles are not absorbing or emitting the photons. If a stream of photons (a light beam) is sent through an atomic gas (Lukin et al., 2000), for example, then provided the gas atoms are not resonant with the laser wavelength, the atoms should swarm around the photons rather like bees around a honey pot. In other words, polaritons should be formed. Furthermore, these polaritons could be organized, through an appropriate choice of laser intensity, to travel at a speed that is much less than the velocity of light. The particle swarm that has associated itself with the photon will slow down the light dramatically. This is an elegant idea and is a striking illustration of the character of the polariton. It is not surprising that the phrase "to catch a moonbeam" has been used (Sincelli, 2000) to describe the polariton-based possibility of slowing down light to walking speed.

ALLAN BOARDMAN

See also **Alfvén waves; Drude model; Excitons; Nonlinear plasma waves; Plasma soliton experiments; Polarons**

Further Reading

Baher, S. & Cottam, M.G. 2003. Theory of nonlinear guided and surface plasmon-polaritons in dielectric films. *Surface Review and Letters*, 10: 13–22

Beletskii, N.N. & Bludov, Yu.V. 2002. Nonlinear surface polaritons in a two-dimensional electron system placed into a quantising magnetic field. *Telecommunications and Radioengineering*, 57: 67–78

Boardman, A.D. (editor). 1982. *Electromagnetic Surface Modes*, Chichester and New York: Wiley

Boardman, A.D. & Egan, P. 1985. s-polarised waves in a thin dielectric film asymmetrically bounded by optically nonliner media. *IEEE Quantum Electronics*, 21: 1701–1713

Boardman, A.D., Egan, P., Lederer, F., Langbein, U. & Mihalache, D. 1991. Third-order nonlinear electromagnetic TE and TM guided waves. In *Nonlinear Surface Electromagnetic Phenomena*, edited by H.-E. Ponath & G.I. Stegeman, Amsterdam and New York: North-Holland

Hopfield, J.J. 1958. Theory of the contribution of excitons to the complex dielectric constant of crystals. *Physical Review*, 112: 1555–1567

Kittel, C. 1995. *Introduction to Solid State Physics*, New York and London: Wiley

Lukin, M.D., Yelin, S. & Fleischhauer, M. 2000. Entanglement of atomic ensembles by trapping correlated photon states. *Physical Review Letters*, 84: 4232–4235

Savvides, P.G., Baumberg, J.J., Stevenson, R.M., Skolnick, M.S., Roberts, J.S. & Whittaker, D.M. 2000. Angular-asymmetric nonlinear polariton dynamics in semiconductor microcavities. *Physica Status Solidi* B, 221: 77–83

Sincelli, M. 2000. How to catch a moonbeam. *Physical Review Focus* 5, story 19

Walker, C.T. & Slack, G.A. 1970. Who named the -ON's? *American Journal of Physics*, 38: 1380–1389

POLARONS

The polaron concept goes back to 1933 when Lev Landau suggested the phenomenon of self-localization (self-trapping) of an electron (or a hole), slowly moving in a polar crystal (Landau, 1933). The electron locally polarizes and, therefore, distorts the ionic lattice in the region where it is located. This local lattice distortion, in turn, creates a potential well that traps the electron, lowering its energy. Therefore, the electron and the accompanying self-consistent polarization field can move in the crystal as a whole entity and may be considered as a quasi-particle. The term *polaron* was introduced in 1946 by Solomon Pekar, who studied the limiting (adiabatic) case of a sufficiently strong electron-lattice interaction and described the most-important properties of a stationary polaron, using a continuum model of electron motion (Pekar, 1946).

Pekar's theory of a polaron is based on the assumption that the polar coupling between electrons and optical phonons can be very large in ionic crystals. In ionic crystals, some of the ions are positively charged, while others are negatively charged, and an optical phonon has the different ions in the crystal vibrating out of phase. When the positive ions and negative ions oscillate in opposite directions (at a point $\boldsymbol{r}$ of the crystal), they create a polarization field $\boldsymbol{P}(\boldsymbol{r})$. This causes an electric field $\boldsymbol{E}(\boldsymbol{r})$, which is the source of the polar coupling. Polar coupling is only to longitudinal optical (LO) phonons, because only LO phonons set up a strong electric field when they vibrate. (This electric field is in the direction of the phonon wave vector $\boldsymbol{q}$.) Writing the expansion in normal modes:

$$\boldsymbol{E}(\boldsymbol{r}) = N^{-1/2} \sum_{\boldsymbol{q}} \boldsymbol{E}_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}$$
$$\boldsymbol{P}(\boldsymbol{r}) = N^{-1/2} \sum_{\boldsymbol{q}} \boldsymbol{P}_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \qquad (1)$$

with N being the number of ions in the crystal and using the fact that there are no free charges, that is, $\nabla \cdot (\boldsymbol{E} + 4\pi\boldsymbol{P}) = 0$, one obtains the relation

$$\boldsymbol{E}_{\boldsymbol{q}} = -4\pi\boldsymbol{P}_{\boldsymbol{q}}. \qquad (2)$$

Each $\boldsymbol{P}_{\boldsymbol{q}}$ is proportional to the (quantized) displacement field

$$\boldsymbol{P}_{\boldsymbol{q}} = U_{\boldsymbol{q}} e \left(\frac{\hbar}{2M\omega_{\boldsymbol{q}}}\right)^{1/2} \mathrm{i}\frac{\boldsymbol{q}}{|\boldsymbol{q}|}\left(a_{\boldsymbol{q}}^{\dagger} + a_{-\boldsymbol{q}}\right), \qquad (3)$$

where the coefficient $U_{\boldsymbol{q}}$ is to be determined, e is the electron charge, M the ion mass, $\omega_{\boldsymbol{q}}$ is the frequency of the $\boldsymbol{q}$th mode, and $a_{\boldsymbol{q}}$ ($a_{\boldsymbol{q}}^{\dagger}$) is the annihilation (creation) operator of the $\boldsymbol{q}$th mode.

On the other hand, the electric field can be represented as a gradient of a potential field:

$$\boldsymbol{E}(\boldsymbol{r}) = -\nabla\phi(\boldsymbol{r}) = -\mathrm{i}\sum_{\boldsymbol{q}} \boldsymbol{q}\mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}\phi_{\boldsymbol{q}}\,. \qquad (4)$$

Using Equations (1)–(4), the potential field for the electron is

$$\phi(\boldsymbol{r}) = 4\pi e \sum_{\boldsymbol{q}} U_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \left(\frac{\hbar}{2NM\omega_{\boldsymbol{q}}}\right)^{1/2} \times \frac{1}{|\boldsymbol{q}|}\left(a_{\boldsymbol{q}}^{\dagger} + a_{-\boldsymbol{q}}\right). \qquad (5)$$

Because each normal mode $\boldsymbol{q}$ assumes a new equilibrium configuration and oscillates in the vicinity of these new equilibria, the coupling constant $U_{\boldsymbol{q}}$ can be evaluated from the potential energy between two fixed electrons. For a dispersionless polar crystal, when the frequency $\omega_{\boldsymbol{q}}$ is nearly a constant ($\omega_{\boldsymbol{q}} \doteq \omega_0$), Pekar has derived the value

$$U_0^2 = \frac{\rho\omega_0^2}{4\pi}\left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_0}\right), \qquad (6)$$

where ρ is the mass density of the crystal (so $\rho V = NM$, with V being the crystal volume). The dielectric constants ε_0 and ε_∞ are both measurable: ε_0 is measured by putting the crystal between the parallel plates of a capacitor at low frequency and ε_∞ is the square of the refractive index.

Thus, it is possible to write the total Hamiltonian of a single electron that interacts with the phonons in the form

$$H = \frac{\boldsymbol{p}^2}{2m} + \omega_0 \sum_{\boldsymbol{q}} a_{\boldsymbol{q}}^{\dagger} a_{\boldsymbol{q}} + V^{-1/2} \sum_{\boldsymbol{q}} \frac{\chi_0}{|\boldsymbol{q}|} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \left(a_{\boldsymbol{q}}^{\dagger} + a_{-\boldsymbol{q}}\right). \qquad (7)$$

Here the unperturbed electron is taken to have free-particle motion with a momentum $\boldsymbol{p}$ and an effective mass m. Since there is only one electron in the problem, the results are independent of the statistics of the particle. The same results are obtained for any fermion or boson in the crystal, such as holes, excitons, or other particles, as long as they are free to move. Owing to Equations (5) and (6), the electron-phonon coupling constant χ_0 in Hamiltonian (7) is given by

$$\chi_0^2 = 4\pi\alpha\hbar(2m)^{-1/2}(\hbar\omega_0)^{3/2} \tag{8}$$

with the dimensionless polaron constant

$$\alpha = \frac{e^2}{\hbar}\left(\frac{m}{2\hbar\omega_0}\right)^{1/2}\left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_0}\right). \tag{9}$$

In Pekar's analysis of this model, the method of calculation is basically a variational procedure using a Gaussian wave function. The total wave function is assumed to be the product of electron and phonon coordinates (adiabatic approximation):

$$\begin{aligned}\Phi(\boldsymbol{r}; \boldsymbol{Q_q}) &= \varphi(\boldsymbol{r})\psi_n(\boldsymbol{Q_q} + \delta\boldsymbol{Q_q}),\\ \varphi(\boldsymbol{r}) &= \frac{\beta^{3/2}}{\pi^{3/4}}\exp\left(-\frac{\beta^2\boldsymbol{r}^2}{2}\right),\end{aligned} \tag{10}$$

where β is a variational parameter to be determined. The phonon wave functions ψ_n are the wave functions for harmonic oscillators centered about new equilibrium displacements, which also need to be determined. Having carried out the variational procedure, one obtains the minimum value of β and the total energy $E = E(\beta)$ of the coupled electron–phonon system:

$$\begin{aligned}\beta_0 &= \frac{2\alpha}{3}\sqrt{\frac{m\omega_0}{\pi}},\\ E(\beta_0) &= -\frac{\alpha^2\omega_0}{3\pi} \simeq -0.106\alpha^2\omega_0.\end{aligned} \tag{11}$$

This result shows the existence of a dynamically stable, self-localized state of an electron interacting with optical phonons in ionic crystals and having a finite binding energy, under the condition that the size of the self-trapped state is large compared with the lattice spacing constant (so the lattice discreteness becomes irrelevant in the theory).

Pekar's continuum polaron was the first model of a self-trapped state admitting a self-consistent theoretical treatment. The importance of this problem was soon understood by many theoretical and mathematical physicists working in solid state physics and quantum field theory, including N.N. Bogolyubov, S.V. Tyablikov, T.D. Lee, H. Fröhlich, R.P. Feynman, E.I. Rashba, T. Holstein, Y. Toyozawa, I.G. Lang, and Yu.A. Firsov. As a result, several pioneering works were published in the 1950s, favoring use of the methods of quantum field theory in solid state physics. Thus, the "polaron" concept has lost its original meaning as an electron interacting with long-wave-length polarized optical phonons and has been applied to a significantly wider class of self-trapped states, resulting in a large body of work that includes exciton-phonon interactions (Rashba, 1957) and even applications in biology (Scott, 1992), with very recent important contributions to the classical polaron theory (Romero et al., 1999). In this context, the following comments are relevant:

(i) The strong coupling limit in Pekar's theory is calculated in the adiabatic approximation, when the electron has a sufficient binding energy and its oscillatory motion in the potential well is much faster than the vibrational frequency of the phonons. Thus, the phonons do not have time to adjust to the individual oscillations of the electron. Instead, they adjust to the average motion of the electron. Quite a different picture applies to the weak coupling limit. In this case, the phonon energy is larger than that of the electron, so that the phonons (or ion polarization) follow the electron during its motion. The weak coupling limit was studied by Fröhlich within his polaron model, in which Hamiltonian (7) was generalized to (Fröhlich, 1954)

$$\begin{aligned}H = &\sum_{\boldsymbol{k}} \frac{\boldsymbol{k}^2}{2m} c_{\boldsymbol{k}}^\dagger c_{\boldsymbol{k}} + \omega_0 \sum_{\boldsymbol{q}} a_{\boldsymbol{q}}^\dagger a_{\boldsymbol{q}}\\ &+ V^{-1/2} \sum_{\boldsymbol{q},\boldsymbol{k}} \frac{\chi_0}{|\boldsymbol{q}|} c_{\boldsymbol{k}+\boldsymbol{q}}^\dagger c_{\boldsymbol{k}} \left(a_{\boldsymbol{q}}^\dagger + a_{\boldsymbol{q}}\right),\end{aligned} \tag{12}$$

where $c_{\boldsymbol{k}}$ ($c_{\boldsymbol{k}}^\dagger$) is the annihilation (creation) operator of the electron with the wave number $\boldsymbol{k}$. However, the most sophisticated study of the "large" (or "continuum") polaron, using the Fröhlich Hamiltonian (12), is due to Feynman (1955) with the path-integral method, substantially extended in the past decade. For an extensive review of the large polaron studies in both the strong and weak coupling limits, see Mitra et al. (1987).

(ii) When the electron-phonon coupling constant is large, all the states in the Brillouin zone are involved in the formation of the polaron wave function. In this case, the polaron radius becomes comparatible with the lattice constant and the continuum approximation is no longer valid. For a "small" polaron, the theory should recognize the periodicity of the crystal and thereby assume that the motion of the electron, or another particle, is no longer translationally continuous. Therefore, one assumes that the electron may occupy an orbital state $\phi(\boldsymbol{r} - \boldsymbol{R}_j)$ centered on an atomic site $\boldsymbol{R}_j$. The orbital states are identical on each site, so that there is periodicity. The electron may move from site to site as in the tight-binding model. This motion may be caused by the overlap of the orbitals on adjacent sites. Thus, the following Hamiltonian is suitable for the small polaron theory:

$$H = J \sum_{j,l} C^{\dagger}_{j+l} C_j + \sum_{\boldsymbol{q}} \omega_{\boldsymbol{q}} a^{\dagger}_{\boldsymbol{q}} a_{\boldsymbol{q}} + \sum_{j,\boldsymbol{q}} \chi_{\boldsymbol{q}} e^{i\boldsymbol{q}\cdot\boldsymbol{R}_j} C^{\dagger}_j C_j \left(a^{\dagger}_{\boldsymbol{q}} + a_{-\boldsymbol{q}} \right), \quad (13)$$

where the factor J describes the overlap of the orbitals and C_j ($C^{\dagger}_j$) is the annihilation (creation) operator of the electron on site $\boldsymbol{R}_j$. Basic features of the small polaron were well recognized a long time ago by Tyablikov, Holstein, Lang, and Firsov, among others, and are described in several review papers and textbooks (Appel, 1968; Böttger & Bryksin, 1985; Mahan, 1990; Alexandrov & Mott, 1995).

(iii) Another interesting case is a problem of the lattice polaron with a long-range Fröhlich interaction. This polaron has a small (atomic) size of wave function but a large radius of lattice deformation. As shown recently (Alexandrov & Kornilovitch, 1999), this "small Fröhlich polaron" can propagate in a narrow band with the effective mass much smaller than that of the Holstein small polaron of the same binding energy. Alexandrov and Kornilovitch argue that small as well as large Fröhlich (bi)polarons are relevant quasiparticles in the cuprates, describing holes in the CuO_2 plane coupled with the lattice distortion created by a long-range interaction.

(iv) The polaron concept appeared so ubiquitous that it was used even in biology to describe a possible mechanism of the localization and transport of vibrational energy in molecular chains, particularly in proteins. As was suggested by Davydov, the energy of the C=O stretching (amide-I) vibrations, being localized on a peptide group of the α-helix, distorts the structure of the helix in this place. In its turn, a (contractive) distortion of the helical chain creates a potential well that traps the amide-I oscillation energy preventing its dispersion. This mobile self-localized (or self-trapped) state that can travel along the chain, carrying the energy released under the adenosine triphosphate (ATP) hydrolysis, is called a "Davydov soliton" (Scott, 1992).

ALEXANDER V. ZOLOTARYUK

See also **Davydov soliton; Discrete self-trapping system; Local modes in molecular crystals**

Further Reading

Alexandrov, A.S. & Kornilovitch, P.E. 1999. Mobile small polaron. *Physical Review Letters*, 82: 807–810

Alexandrov, A.S. & Mott, N.F. 1995. *Polarons and Bipolarons*, Singapore: World Scientific

Appel, J. 1968. In *Solid State Physics*, edited by F. Seitz, D. Turnbull & H. Ehrenreich, New York: Academic Press

Böttger, H. & Bryksin, V.V. 1985. *Hopping Conduction in Solids*, Deerfield Beach, FL: VCH

Feynman, R.P. 1955. Slow electrons in a polar crystal. *Physical Review*, 97: 660–665

Fröhlich, H. 1954. Electrons in lattice fields. *Advances in Physics*, 3: 325

Landau, L.D. 1933. Über die Bewegung der Elektronen in Kristallgitter. *Physik Z. Sowjetunion*, 3: 664–665

Mahan, G.D. 1990. *Many-Particle Physics*, New York and London: Plenum Press

Mitra, T.K., Chatterjee, A. & Mukhopadhyay, S. 1987. Polarons. *Physics Reports*, 153: 91–207

Pekar, S.I. 1946. Autolocalization of an electron in a dielectric inertially polarized medium. *Soviet Physics, JETP*, 16: 335–340

Rashba, E.I. 1957. Theory of the strong interaction of electron excitations with lattice vibrations in molecular crystals. I & II. *Soviet Physics, Optika i Spektroskopiya (Optics and Spectroscopy)*, 2: 75–98

Romero, A.H., Brown, D.W. & Lindenberg, K. 1999. Polaron effective mass, band distortion, and self-trapping in the Holstein molecular-crystal model. *Physical Review* B, 59: 13728–13740; The self-trapping line of the Holstein molecular crystal model in one dimension. *Physical Review*, 60: 4618–4623

Scott, A.C. 1992. Davydov's soliton. *Physics Reports*, 217: 1–67

POLYMERIZATION

Polymers are long molecules made from one or more types of repeating units, called monomers. The simplest synthetic polymer consists of hundreds to even millions of monomers that are connected end to end in a linear chain. Polymers are essential to our modern world and our very life. Biochemical systems involve polymers because all enzymes are proteins, a type of polymer composed of about 20 different amino acids. DNA and RNA are also polymers but formed from four different bases (Nelson & Cox, 2000). Goldbeter (1995) discusses well the type of nonlinear phenomena that can occur. We focus only on synthetic polymers here.

A distribution of chain lengths always exists in a synthetic system. This molecular weight distribution can be quite broad, often spanning several orders of magnitude. However, biopolymers consist of a single molecular weight.

Linear polymers are often thermoplastic, meaning that they can flow at some temperature, which depends on the molecular weight. Common examples are poly(styrene) and poly(methyl methacrylate), whose trade name is Plexiglass.

Polymers need not be simple chains but can be branched or networked. Linear polymers are analogous to strings of spaghetti. Crosslinked polymers can be gels that swell in a solvent or thermosets, which form rigid three-dimensional networks. Home epoxy glue is an example of a crosslinked polymer. Figure 1 illustrates these variations.

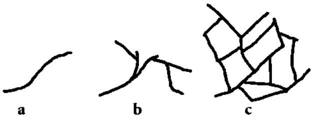

Figure 1. Illustration of three types of polymers: (a) linear, (b) branched, (c) crosslinked.

Polymer Kinetics and Mechanisms of Feedback

The mechanism of polymerization usually falls into one of two categories: chain growth or step growth. In chain-growth polymerization, a monomer is converted into a reactive species by an initiating species, either an ion or a free radical (a molecule containing an unpaired electron), that rapidly reacts with other monomers while always maintaining the reactive character. The reactive chain can be broken through some termination process. In most cases, even though relatively little of the monomer has reacted, high molecular weight molecules are produced in a few seconds. They usually do not participate further in the reaction. Plexiglass is produced via free-radical chain-growth polymerization.

For step-growth polymers, the monomers can usually react from both ends, and thus, all the monomers and polymers can react with each at any time from both ends. The molecular weight increases continuously throughout the reaction as monomers react to form dimers, and then dimers can react with each other or monomers, etc.

The physical properties of the polymer medium change dramatically during reaction. For example, the viscosity almost always increases orders of magnitude as the polymer concentration and/or molecular weight increases. This can be understood if you consider stirring a pot of penne versus a pot of spaghetti: the longer pasta pieces intertwine and resist flowing. Such physical changes will often affect the kinetic parameters of the reaction and the transport coefficients of the medium.

Synthetic polymer systems can exhibit feedback through several mechanisms. The simplest is thermal autocatalysis, which occurs in any exothermic reaction. The reaction raises the temperature of the system, which increases the rate of reaction through the Arrhenius dependence of the rate constants. Most chain-growth polymerizations, including free-radical polymerizations, are highly exothermic. Some step-growth polymerizations are highly exothermic (epoxies), but some are not (polyesters).

Free-radical polymerizations of certain monomers exhibit autoacceleration at high conversion via an additional mechanism, the isothermal "gel effect" or "Trommsdorff effect" (Odian, 1991). These reactions occur by the creation of a radical that attacks an unsaturated monomer, converting it to a radical, which can bond to another monomer, propagating the chain. The chain growth terminates when two radical chains ends encounter each other, forming a stable chemical bond. As the polymerization proceeds, the viscosity increases. The diffusion-limited termination reactions are thereby slowed down, leading to an increase in the overall polymerization rate. The increase in the polymerization rate induced by the increase in viscosity builds a positive feedback loop into the polymerizing system.

Some polymer hydrogels exhibit "phase transitions" as the pH and/or temperature are varied. Disposable diapers contain hydrogels that absorb great amounts of water because the polymers contain charged groups that bind the water. These charged groups can be affected by acid or base and are also temperature sensitive. The gel can swell significantly as the conditions are changed and can also exhibit hysteresis (Addad, 1996). Most polymers are immiscible, meaning that they will not dissolve with each other but will separate into different phases like oil and water. Introducing chemical reactions to an initially miscible polymer mixture often leads to phase separation. Autocatalytic behavior driven by chemical reactions and concentration fluctuations in miscible polymer mixtures can occur in photo-crosslinked polymer mixtures (Pojman & Tran-Cong-Miyata, 2003, Chapter 22). Concentration fluctuations increase as the reaction proceeds, leading to the reaction of photoreactive groups attached to one of the polymer components. This leads to an increase in the reaction yield that, in turn, accelerates the concentration fluctuations.

Finally, the polymer melts and the solutions are usually non-Newtonian fluids (Gupta, 2000), which means that there is no linear relationship between the stress applied to the fluid and strain that is measured. They often exhibit shear thinning, which means that the viscosity decreases as the shear is increased, but can also exhibit shear thickening. These properties make polymer additives important in the food and cosmetic industry.

Applications

There are two main scenarios by which nonlinear phenomena of possible utility arise with polymers. The first approach is to use a nonlinear system to drive the nonlinear behavior of a polymer or polymerization. Acrylonitrile will polymerize periodically when added to the Belousov–Zhabotinsky (BZ) reaction (Washington et al., 1999).

Yoshida et al. created a self-oscillating gel by coupling a pH oscillating reaction with a polymeric gel that expands and contracts with changes in pH (Yoshida et al., 1995). They have also used a gel in which the ruthenium catalyst of the BZ reaction is chemically incorporated into polymer, causing local swelling and deswelling as chemical waves pass through the gel (Pojman & Tran-Cong-Miyata, 2003, Chapter 3).

The second approach is to use the inherent nonlinearities in a polymeric system. Polymerizations in a continuously stirred flow tank reactor (CSTR) can show nonlinear phenomena through the interaction of thermal feedback and the gel effect. Teymour and Ray showed in laboratory-scale CSTR experiments on vinyl acetate polymerization that oscillations with periods of 200 min could occur (Teymour & Ray, 1992). The gel effect causes the rate of polymerization to increase, but the increase in temperature lowers the viscosity, returning the polymerization to its lower rate.

Frontal Polymerization

One of the most promising applications of nonlinear dynamics to polymer science is the phenomenon of frontal polymerization. Frontal polymerization is a process of converting monomers into a polymer via a localized reaction zone that propagates through the monomer, much like a "liquid flame."

Thermal frontal polymerization involves the coupling of thermal diffusion and Arrhenius reaction kinetics of an exothermic polymerization (Pojman et al., 1996). Front temperatures are typically 200°C, and front velocities are from 1–20 cm/min. Because of the large composition and thermal gradients created in the front, convection can be a significant issue (Pojman et al., 1996). Simple convection can occur for thermosets in ascending fronts, but for thermoplastics, the Rayleigh–Taylor instability is a significant issue unless a filler is added to increase the viscosity.

Such systems exhibit many modes of nonplanar front propagation. Particularly fascinating are the spin modes in which the front propagates as a helix, and hot spots can be seen on the surface with an infrared camera (Figure 2). Even more complicated modes can be observed depending on the chemical composition.

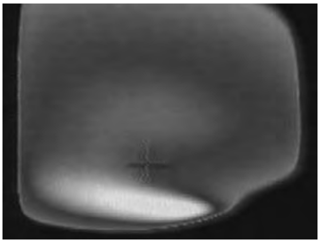

Figure 2. Infrared image of a spin mode in the free-radical-frontal polymerization of an acrylate in a 1.5 cm diameter tube. (Image courtesy of J. Pojman.)

Isothermal frontal polymerization (IFP), also called interfacial gel polymerization, is a slow process in which polymerization occurs at a constant temperature and a localized reaction zone propagates because of the gel effect (Epstein & Pojman, 1998, Chapter 11). Using IFP, one can control the gradient of an added material to generate materials for optical applications.

JOHN A. POJMAN

See also **Belousov–Zhabotinsky reaction; DNA premelting; DNA solitons; Fluid dynamics; Morphogenesis, biological; Protein structure; Rayleigh–Taylor instability; Rheology**

Further Reading

Addad, J.P.C. 1996. *Physical Properties of Polymer Gels*, Chichester: Wiley

Epstein, I.R. & Pojman, J.A. 1998. *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns and Chaos*, Oxford and New York: Oxford University Press

Goldbeter, A. 1995. *Biochemical Oscillations and Cellular Rhythms*, Cambridge and New York: Cambridge University Press

Gupta, R.K. 2000. *Polymer and Composite Rheology*, New York: Marcel Dekker

Nelson, D.L. & Cox, M.M. 2000. *Lehninger Principles of Biochemistry*, 3rd edition, New York: Worth

Odian, G. 1991. *Principles of Polymerization*, New York: Wiley

Pojman, J.A., Ilyashenko, V.M. & Khan, A.M. 1996. Free-radical frontal polymerization: Self-propagating thermal reaction waves. *Journal of the Chemical Society Faraday Transactions*, 92: 2825–2837

Pojman, J.A. & Tran-Cong-Miyata, Q. (editors). 2003. *Nonlinear Dynamics in Polymeric Systems*, Washington, DC: American Chemical Society

Teymour, F. & Ray, W.H. 1992. The dynamic behavior of continuous polymerization reactors — V. Experimental investigation of limit-cycle behavior for vinyl acetate polymerization. *Chemical Engineering Science*, 47: 4121–4132

Washington, R.P., Misra, G.P., West, W.W. & Pojman, J.A. 1999. Polymerization coupled to oscillating reactions: I. A mechanistic investigation and numerical simulation of acrylonitrile polymerization in the Belousov–Zhabotinsky reaction in a batch reactor. *Journal of the American Chemical Society*, 121: 7373–7380

Yoshida, R., Ichijo, H., Hakuta, T. & Yamaguchi, T. 1995. Self-oscillating swelling and deswelling of polymer gels. *Macromolecular Rapid Communications*, 16: 305–310

POPULATION DYNAMICS

Biological populations have a propensity for exponential growth. This Malthusian principle is expressed by the linear, ordinary differential equation $\mathrm{d}n/\mathrm{d}t = rn$, $r > 0$, where $n = n(t)$ is a measure of population size or density (e.g., individual numbers, biomass, or dry weight) as a function of time t. Although populations can, under appropriate conditions, exhibit exponential growth over a finite period of time, no population can grow indefinitely according to this law. The field of nonlinear population dynamics involves the study of how population numbers are regulated by biological and environmental effects. The equation

$$\frac{\mathrm{d}n}{\mathrm{d}t} = rn\,(1 - n/K) \qquad (1)$$

is an example of a modification of the exponential growth equation. It is based on the assumption that per capita growth is a decreasing, linear function of population size. In the 1920s, many researchers, including Alfred J. Lotka, Lowell J. Reed, and, most notably, Raymond Pearl, considered Equation (1) as a fundamental law of population growth. It was later discovered that a Belgian mathematician, Pierre-François Verhulst, had used the equation in his studies of population growth in the 1840s. Equation (1) is called the Pearl–Reed equation, the Pearl–Verhulst equation, or more commonly, the "logistic equation" (a name adapted from Verhulst's name *logistique* for the equation).

Equation (1) implies all population densities monotonically approach the equilibrium K, called the carrying capacity. Before leveling off at this carrying capacity, low population numbers initially grow exponentially at the inherent growth rate r, and the resulting S-shaped trajectory is called a logistic growth curve. In their studies of human population dynamics, Pearl and his colleagues fit logistic curves to census numbers from many countries and states. Although today considered rather simplistic as a model of population growth, the logistic equation nonetheless encapsulates the notion of density self-regulation in population dynamics and has inspired many basic notions in ecology (e.g., habitat carrying capacity, the classification of species as "r or K selectors").

The logistic equation is a first-order differential equation of the Kolmogorov type

$$\frac{\mathrm{d}n}{\mathrm{d}t} = nf(n) \qquad (2)$$

with per capita growth rate $f(n) = r\,(1 - n/K)$. Many theories of population growth use the Kolmogorov equations with other specialized forms for $f(n)$. One notable class of models arises from a modification of the negative feedback density regulation assumption of the logistic Equation (1) to include a positive feedback at low population numbers. This notion was put forth by W.C. Allee in order to account for increased growth rates associated with increased population size experienced by some (many biologists would say most) populations when at low numbers (although a negative feedback still comes into play at high numbers). An example is the equation

$$\frac{\mathrm{d}n}{\mathrm{d}t} = rn\left(1 - \frac{n}{K}\right)(n - \alpha K), \quad \alpha < 1.$$

The "Allee effect" entails a threshold αK such that populations go extinct unless their initial state $n(0)$ exceeds αK, in which case they equilibrate at the carrying capacity K.

Ordinary differential equation models based on equations of form (2) neglect important factors affecting the dynamics of real biological populations. Examples include temporal inhomogeneity (nonconstant vital rates and environment fluctuations), spatial inhomogeneity (diffusion and dispersal), and population inhomogeneity (differences among individuals). The inclusion of temporal inhomogeneities in Equation (2) produces a non-autonomous differential equation with $f = f(t, n)$. For example, a periodic carrying capacity K in the logistic Equation (1) reflects a seasonal habitat.

The movement of populations in a spatially extended habitat can be modeled by adding a diffusion or dispersal term to a growth model. For example, Fisher's equation

$$\frac{\mathrm{d}n}{\mathrm{d}t} = c\Delta n + rn\,(1 - n/K)$$

has been used to study the spread of populations by means of traveling waves. Other types of diffusion operators, including integral operators described by dispersal kernels, are often considered more appropriate for the movement of biological populations (Murray, 2003).

Differences in chronological age, body size, and other physiological characteristics of individual organisms can significantly affect birth and death rates and consequently, the entire population's dynamics. The McKendrick equation

$$\frac{\partial\rho}{\partial t} + \frac{\partial\rho}{\partial a} = -\delta\rho$$

describes the death processes of a population in terms of an age-specific density $\rho = \rho(t, a)$ and death rate δ. This equation is accompanied by a renewal or birth equation

$$\rho(t, 0) = \int \beta\rho\,\mathrm{d}a.$$

If the age-specific birth and death rates β and δ are independent of population density ρ, then these equations are linear and imply (under suitable conditions) the Fundamental Theorem of Demography. This theorem states that regardless of whether the population grows or decays exponentially, the normalized density approaches the positive, normalized eigensolution associated with the dominant eigenvalue. If β and/or δ are dependent on ρ, then the McKendrick model is nonlinear. This model can be used to study the effects that various life-cycle characteristics have on population level dynamics, effects such as maturation delays, gestation periods, and nonreproductive quiescent stages (Webb, 1985; Metz & Diekmann, 1986; Cushing, 1998).

Populations are often modeled using discrete time maps in which population numbers are predicted from one census time to the next. The discrete time exponential growth is described by the linear map (or difference equation) $x_{t+1} = \lambda x_t$, $\lambda > 1$. An analog of logistic equation for density regulated growth is the Beverton–Holt equation

$$x_{t+1} = \lambda x_t \frac{1}{1 + (\lambda - 1)\, x_t/K},$$

whose solutions monotonically approach K as $t \to +\infty$. Other types of density-dependent, population growth models can lead to non-equilibrium asymptotic states. For example, in the Ricker model

$$x_{t+1} = \lambda x_t \exp(-cx_t),$$

larger population numbers, of sufficient magnitude, produce smaller population numbers at the next census. This type of density-dependent "depensation" produces a period-doubling, bifurcation route to chaos as the inherent growth rate λ increases. In the 1970s, Robert May stimulated the interest in complex and chaotic dynamics that flowered during the last decades of the 20th century by his studies of the Ricker model and similar one-dimensional maps as models of population growth (May, 1974).

There is a long tradition of using discrete time models to model structured populations (Caswell, 2001). The Leslie matrix model

$$\boldsymbol{x}_{t+1} = L\boldsymbol{x}_t \tag{3}$$

is an example. The components of the vector $\boldsymbol{x}_t$ are the number of individuals, at time t, in a finite collection of age classes. The nonnegative "projection" matrix L contains birth and survival rates per unit time. In other models, the structuring classes are based on other categories (such as body size and life cycle stage). If L is a constant in time, matrix model (3) implies (under suitable mathematical conditions) the Fundamental Theorem of Demography. If the entries of L depend on $\boldsymbol{x}_t$, then the model is nonlinear. Nonlinear matrix models can be used to study the effects of class specific density effects on birth and death rates. For example, a model of a population with three life cycle stages and Ricker-type negative feedback nonlinearities has been used in conjunction experimental studies to document the occurrence of a bifurcation route to chaos in a laboratory population of insects (Cushing et al., 2002).

Most populations interact with other populations in ways that affect each other's vital rates. A natural extension of the logistic equation to the interaction of two species' assumes per capita growth rates are linear expressions of species densities, an assumption that results in the Volterra–Lotka system

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = n_1 (r_1 + c_{11}n_1 + c_{12}n_2),$$
$$\frac{\mathrm{d}n_2}{\mathrm{d}t} = n_2 (r_2 + c_{21}n_1 + c_{22}n_2)$$

of differential equations. A fundamental classification of two species' interactions can be based on the signs of the six coefficients in these equations. Two classical examples are

$$\begin{aligned}\frac{\mathrm{d}n_1}{\mathrm{d}t} &= rn_1 (1 - n_1/K) - c_1n_1n_2,\\ \frac{\mathrm{d}n_2}{\mathrm{d}t} &= -dn_2 + c_2n_1n_2\end{aligned} \tag{4}$$

and

$$\begin{aligned}\frac{\mathrm{d}n_1}{\mathrm{d}t} &= r_1n_1 (1 - n_1/K_1) - c_1n_1n_2,\\ \frac{\mathrm{d}n_2}{\mathrm{d}t} &= r_2n_2 (1 - n_2/K_2) - c_2n_1n_2.\end{aligned} \tag{5}$$

The Lotka–Volterra predator-prey model (4) describes a logistically growing prey population n_1 that is preyed upon by a predator n_2 that dies out exponentially in the absence of n_1. In this model, the per predator uptake rate of prey is proportional to the amount of prey present, a mass-action type interaction that gives rise to the quadratic nonlinearity in (4). If $K < d/c_2$, this predator-prey model predicts predator extinction; if $K > d/c_2$, the model predicts coexistence (in terms of an asymptotically stable equilibrium state). The mass-action assumption is often replaced by one based on a predation rate that saturates (or even decreases) as the amount of prey increases. For example, the MacArthur–Rosenzweig predator-prey model,

$$\begin{aligned}\frac{\mathrm{d}n_1}{\mathrm{d}t} &= rn_1 (1 - n_1/K) - c_1 \frac{n_1}{a + n_1} n_2,\\ \frac{\mathrm{d}n_2}{\mathrm{d}t} &= -dn_2 + \frac{1}{\alpha} c_1 \frac{n_1}{a + n_1} n_2,\end{aligned}$$

assumes a per predator prey uptake rate $c_1n_1/(a + n_1)$ of the so-called Holling II (or Monod, or Michaelis–Menten) type. This predator-prey model predicts a destabilization of the coexistence equilibrium, accompanied by a Hopf bifurcation to a stable limit cycle, for K large. Thus, an enrichment of the habitat results, paradoxically, in a destabilization of the predator-prey interaction and, for this reason, this phenomenon is called the "paradox of enrichment."

The Lotka–Volterra competition model (5) describes two logistically growing populations that adversely affect each other's growth rate through an interference competitive interaction measured by the quadratic mass–action terms. This model predicts a limited number of competitive outcomes. If the interspecific competition is weak compared with the intraspecific competition ($c_1c_2 < r_1r_2/(K_1K_2)$), then the species coexist in an asymptotically stable equilibrium state. If interspecific competition is too strong ($c_1c_2 > r_1r_2/(K_1K_2)$), then one species becomes extinct and the other survives (in an asymptotically stable equilibrium state). The Lotka–Volterra competition model gave rise to the many fundamental ecological notions, including competitive exclusion, ecological niche (there can be no more species than there are limiting resources), and limiting similarity of competitors. Although not universal for theoretical competition models, these concepts are well supported by a large number of other competition models. For example, the exploitative competition model (the chemostat model)

$$\begin{aligned}\frac{dn_1}{dt} &= -dn_1 + \frac{1}{\alpha_1}c_1\frac{R}{a_1+R}n_1,\\ \frac{dn_2}{dt} &= -dn_2 + \frac{1}{\alpha_2}c_2\frac{R}{a_2+R}n_2,\\ \frac{dR}{dt} &= d\,(R_0-R) - c_1\frac{R}{a_1+R}n_1 - c_2\frac{R}{a_2+R}n_2\end{aligned}$$

describes the competition of two species for a (prey) resource R. This model also predicts the survival of only one species (Smith & Waltman, 1995).

Large-dimensional ecosystems can be modeled by coupling together any number of single species such as those described above (May, 2001). Such multispecies models can include temporal, spatial, and/or demographic inhomogeneities. For example, the dispersal and diffusion of many interacting species can be modeled by systems of reaction-diffusion equations, such as Fisher's equation, coupled together so as to describe predator-prey or competition interactions. Similarly, coupled systems of McKendrick equations or Leslie matrix models describe interactions among structured species.

J.M. Cushing

See also **Biological evolution; Brusselator; Epidemiology**

Further Reading

Caswell, H. 2001. *Matrix Population Models: Construction, Analysis and Interpretation*, 2nd edition, Sunderland, MA: Sinauer Associates

Cushing, J.M. 1998. *An Introduction to Structured Population Dynamics*, Philadelphia: Society for Industrial and Applied Mathematics

Cushing, J.M., Costantino, R.F., Dennis, B., Desharnais, R.A. & Henson, S.M. 2002. *Chaos in Ecology: Experimental Nonlinear Dynamics*, New York: Academic Press

Kingsland, S.E. 1995. *Modeling Nature: Episodes in the History of Population Ecology*, Chicago: University of Chicago Press

May, R.M. 1974. Biological populations with nonoverlapping generations: stable points, stable cycles and chaos. *Science*, 186: 645–647

May, R.M. 2001. *Stability and Complexity in Model Ecosystems*, Landmarks in Biology, Princeton: Princeton University Press

Metz, J.A.J. & Diekmann, O. 1986. *The Dynamics of Physiologically Structured Populations*, Berlin: Springer

Murray, J.D. 2003. *Mathematical Biology*, 3rd edition, Berlin and New York: Springer

Smith, H.L. & Waltman, P. 1995. *The Theory of the Chemostat: Dynamics of Microbial Competition*, Cambridge and New York: Cambridge University Press

Webb, G.F. 1985. *Theory of Nonlinear Age-Dependent Population Dynamics*, New York: Marcel Dekker

POWER BALANCE

As the dynamics of any real system includes dissipation, steady dynamic regimes are observed if energy loss is permanently compensated by energy input from some external or intrinsic source, which is a condition of power balance (PB). In the flame of a candle, for example, power input is supplied by the burning wax and dissipation by the emission of heat and light. Starting from the seminal theoretical work by Yakov Zeldovich & David Frank-Kamenetsky (1938), the modern combustion theory has developed a detailed mathematical analysis of the flame propagation in various combustible media, based on the PB concept (Williams, 1964).

The PB analysis in its general form underlies a variety of dynamical phenomena in physics, chemistry, biophysics, and engineering. A simple model is the van der Pol oscillator, which is described by the equation

$$\ddot{\xi} + \omega^2\xi = \alpha\dot{\xi} - \beta\xi^2\dot{\xi}. \qquad (1)$$

Here, $\xi(t)$ is a dynamical variable, ω is the eigenfrequency of linear oscillations in the system, and the small positive coefficients α and β account for the linear intrinsic gain and nonlinear loss, respectively. As the energy (E) of this system is $(\dot{\xi}^2+\omega^2\xi^2)/2$, Equation (1) implies

$$dE/dt = (\alpha - \beta\xi^2)\dot{\xi}^2. \qquad (2)$$

Assuming that $\xi = A\cos\omega t$, averaging Equation (2) over a cycle, and assuming that $\langle dE/dt\rangle = 0$, one finds that PB is established at the amplitude $A = 2\sqrt{\alpha/\beta}$.

Another generic case is the balance between intrinsic loss and energy supplied from an external source. An example is furnished by an equation for an ac-driven damped pendulum with sinusoidal nonlinearity,

$$\ddot{\xi} + \sin\xi = -\alpha\dot{\xi} + \gamma\cos(\omega_0 t), \qquad (3)$$

where $\alpha > 0$ is a friction coefficient and γ and ω_0 are the amplitude and frequency of the drive. (An important realization of this equation is a Josephson junction. In this case, ξ is the phase difference of the superconducting wave function across the junction, α accounts for ohmic loss, and the ac drive is induced by bias current applied to the junction (Barone & Paternó, 1982).) The corresponding PB equation takes the form (cf. Equation (2))

$$dE/dt = -\alpha\dot{\xi}^2 + \gamma\dot{\xi}\sin(\omega_0 t). \qquad (4)$$

In the lowest-order approximation, $\alpha = \gamma = 0$, two different types of exact solution are known: oscillating and rotating ones with zero and finite average velocity $\langle\dot{\xi}\rangle$, respectively. Substitution of the rotating solution into Equation (4) (where α and γ are treated as small parameters) and averaging over the period shows that PB is established for the solution whose frequency is locked to the driving frequency ω_0, so that $\omega = \omega_0/n$, $n = 1, 2, \cdots$. In other words, the PB condition selects the solution's frequency. Further, it follows from Equation (4) that the PB condition cannot be met unless the drive's strength exceeds a finite threshold value, γ_{thr}.

At $\gamma > \gamma_{\text{thr}}$, the PB between the ac drive and loss selects a constant value of the phase shift ϕ_0 between the ac drive and the phase of the oscillating part of the solution. Indeed, averaging Equation (4) and demanding $\langle \mathrm{d}E/\mathrm{d}t \rangle = 0$ yield a general result, $\cos\phi_0 = \gamma_{\text{thr}}/\gamma$; that is, there are two power-balanced solutions, $\phi_0 = \pm\cos^{-1}(\gamma_{\text{thr}}/\gamma)$. Further analysis demonstrates that one solution is stable and the other one is unstable.

A similar mechanism underlies propagation of a magnetic-flux quantum (fluxon, or topological soliton) in a long weakly damped ac-driven Josephson junction with periodic spatial modulation, which is described by the following sine-Gordon (SG) equation:

$$\xi_{tt} - \xi_{xx} + [1 + \varepsilon \sin(2\pi x/\lambda)] \sin\xi = -\alpha\xi_t + \gamma\cos(\omega_0 t). \quad (5)$$

Here x is the coordinate along the Josephson junction, the subscripts stand for partial derivatives, and ε and λ are the amplitude and period of the spatial modulation. In this case, the average velocity $\langle v \rangle$ of the ac-driven fluxon is determined by the condition of locking the frequency of the periodic passage of the spatially periodic relief by the fluxon to the ac-drive's frequency, which yields a spectrum of PB velocities: $\langle v \rangle = \lambda\omega_0/(2\pi n), n = \pm 1, \pm 2, \ldots$. Note that the sign of the velocity is determined by an initial push setting the fluxon in motion.

The latter phenomenon was observed in a direct experiment, in the form of an *inverse Josephson effect*, that is, dc voltage induced by ac bias current (Ustinov & Malomed, 2001). The periodic spatial modulation is necessary for the effect, as it opens a way to establish the PB between the ac drive and dissipation. Nevertheless, in an annular (ring-shaped) Josephson junction of a long but finite length, the same effect is possible without any spatial modulation, due to the interaction of the fluxon with its own tail (Goldobin et al., 2002). In that case, the average velocity is not "quantized," as above, but may take any value. A noteworthy feature, specific to the ring system, is the dependence of the threshold value γ_{thr} on the ring's length L (see Figure 1). (If the driving signal is slightly nonmonochromatic, progressive motion is still possible over a long time, but it is eventually destroyed by decoherence accumulating due to small fluctuations of the driving frequency (Filatrella et al., 2002.) PB for breathers in long lossy Josephson junctions is also possible (Lomdahl & Samuelsen, 1986).

Another class of PB problems can be formulated in terms of the complex Ginzburg–Landau equation, which may be regarded as a perturbed version of the nonlinear Schrödinger (NLS) equation:

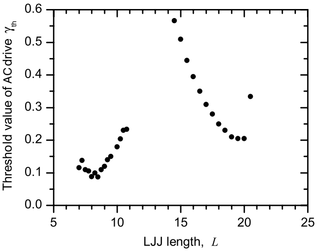

Figure 1. The minimum (threshold) value of the ac-drive's strength γ, in the case $\varepsilon = 0$, $\omega = 1.12$, and $\alpha = 0.1$ [see Equation (5)], which is necessary to support progressive motion (in either direction) of a fluxon in a long Josephson ring, vs. the length (L) of the ring. In regions without data, the ac-driven motion of the fluxon is impossible; in particular, the effect vanishes if L exceeds $L_{\max} \approx 21$.

$$\mathrm{i}u_t + \tfrac{1}{2}u_{xx} + |u|^2u = -\mathrm{i}\alpha_0 u + \mathrm{i}\alpha_1 u_{xx} + \mathrm{i}\alpha_2|u|^2u - \mathrm{i}\alpha_3|u|^4u, \quad (6)$$

where all the coefficients $\alpha_{0,1,2,3}$ are assumed to be positive and small. In this model, linear and quintic nonlinear losses are accounted for by α_0, α_1, and α_3, respectively, while α_2 is a cubic-gain coefficient. Equation (6) describes transmission of soliton signals in nonlinear fiber-optic telecommunication links with loss, gain, and filtering (the last being represented by the term α_2 term) (Iannone et al., 1998) and subcritical pulses observed in binary-fluid thermal convection in narrow channels (Kolodner, 1991).

The energy of the optical field in the fiber is $E = \int_{-\infty}^{+\infty} |u(x)|^2\,\mathrm{d}x$, and the loss and gain terms on the right-hand side of Equation (6) give rise to the corresponding PB equation,

$$\mathrm{d}E/\mathrm{d}t = 2\left[-\alpha_0 E + \int_{-\infty}^{+\infty} \left(-\alpha_1|u_x|^2 - \alpha_3|u|^6 + \alpha_2|u|^4\right)\mathrm{d}x\right]. \quad (7)$$

Approximating solutions by the NLS soliton, $u = \eta\,\mathrm{sech}(\eta x)\exp(\mathrm{i}\phi)$, where η is an amplitude, substituting this in Equation (7) and equating $\mathrm{d}E/\mathrm{d}t$ to zero lead to a PB condition in the form

$$\eta\left[8\alpha_3\eta^4 - 5(2\alpha_2 - \alpha_1)\eta^2 + 15\alpha_0\right] = 0. \quad (8)$$

The trivial solution, $\eta = 0$, is stable. The remaining factor in Equation (8) yields two nontrivial solutions if threshold conditions are satisfied: $\alpha_2 > \alpha_1/2$ and $5(2\alpha_2 - \alpha_1)^2 > 96\alpha_0\alpha_3$. The solution with a larger value of η gives a stable SP, and the one with smaller η is unstable.

Another manifestation of PB occurs in reaction-diffusion systems, such as the FitzHugh–Nagumo model (Cross & Hohenberg, 1993). Besides chemical systems—the Belousov–Zhabotinsky (BZ) reaction, heterogeneous catalysis, and so on—models belonging to this class find numerous other applications, including neural networks, cardiac tissue in biophysics, electron-hole plasmas in semiconductors, and gas-discharge plasmas. In this case, the PB takes place between energy release due to the chemical reaction, which is described by local terms of the model, and loss due to diffusion. As a result, various patterns may be supported as stable dynamical equilibria, including standing or traveling waves, fronts (shock waves) and periodic wave trains in one dimension, spiral vortices and localized spots in the two-dimensional case, and vortex rings in three dimensions, among others. An interesting experimental finding is the observation of two drastically different regimes of propagation of a BZ chemical wave in aqueous solution, which resemble deflagration and detonation waves in gas dynamics: a slow wave, which is driven by diffusion of chemical reactants, and a fast "big wave" (Inomoto et al., 1997), which is coupled to surface deformation and flow in the solution. The two modes of the chemical-wave propagation realize the PB in different forms, the choice between them being determined by the initial perturbation.

BORIS MALOMED

See also **Candle; Complex Ginzburg–Landau equation; Damped-driven anharmonic oscillator; FitzHugh–Nagumo equation; Flame front; Long Josephson junctions; Solitons; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Barone, A. & Paternó, G. 1982. *Physics and Applications of the Josephson Effect*, New York: Wiley

Cross, M.C. & Hohenberg, P.C. 1993. Pattern-formation outside of equilibrium. *Reviews of Modern Physics*, 65: 851–1112

Filatrella, G., Malomed, B.A. & Pagano, S. 2002. Noise-induced dephasing of an ac-driven Josephson junction. *Physical Review* E, 65: 051116

Goldobin, E., Malomed, B.A. & Ustinov, A.V. 2002. Progressive motion of an ac-driven kink in an annular damped system. *Physical Review* E, 65: 056613

Iannone, E., Matera, F., Mecozzi, A. & Settembre, M. 1998. *Nonlinear Optical Communication Networks*, New York: Wiley

Inomoto, O., Kai, S., Ariyoshi, T. & Inanaga, S. 1997. Hydrodynamical effects of chemical waves in quasi-two-dimensional solution in Belousov–Zhabotinsky reaction. *International Journal of Bifurcation and Chaos*, 7: 989–996

Kolodner, P. 1991. Drifting pulses of traveling-wave convection. *Physical Review Letters*, 66: 1165–1168

Lomdahl, P.S. & Samuelsen, M.R. 1986. Persistent breather excitations in an ac-driven sine-Gordon system with loss. *Physical Review* A, 34: 664–667

Ustinov, A.V. & Malomed, B.A. 2001. Observation of progressive motion of ac-driven solitons. *Physical Review* B, 64: 020302(R)

Williams, F.A. 1964. *Combustion Theory*, Reading, MA: Addison-Wesley

Zeldozich, Ya.B. & Frank-Kamenetsky, D.A. 1938. K teorii ravnomernogo rasprostrane-niya plameni [On the theory of uniform propagation of frame]. Doklady Akademi Nauk SSSR, 19(10: 693–697.

POWER SPECTRA

See **Spectral analysis**

PRANDTL NUMBER

See **Fluid dynamics**

PREDATOR-PREY SYSTEMS

See **Population dynamics**

PREDICTABILITY OF FORECASTING

See **Forecasting**

PROTEIN DYNAMICS

Proteins carry out structural, catalytic, and molecular recognition roles that require a wide variety of motions over a broad range of timescales, from 10^{-15} to 10^3 s, as sketched in Figure 1. Motions can be as mundane as coupled vibrations or as sophisticated as gene transcription by an RNA polymerase protein. The broad range of timescales is a principal difficulty in understanding protein dynamics. We will focus here on atomic motions and conformational changes of single proteins.

Although there are a wide variety of proteins displaying many remarkable dynamical phenomena, photoactive proteins play a special role in experimental studies of protein dynamics. This is because experiments can be initiated with a femtosecond laser pulse at any temperature from absolute zero to the unfolding temperature of the protein, allowing the wide variety of motions to be observed. Proteins which have been extensively studied this way are myoglobin (Austin et al., 1975), the photosynthetic reaction center, rhodopsin, photoactive yellow protein, and cytochrome coxidase.

An essential result of such studies is that there are three fundamentally different processes involved in modeling atomic motions of proteins. This distinction can be seen in the temperature-dependent rates of various processes observed in myoglobin, sketched in Figure 2 and described in Fenimore et al. (2002). These processes are labeled as fast fluctuations, bond formation, and conformational motions.

Fast fluctuations occur on the 100 picosecond timescale over the temperature range from 200 to 300 K, but their amplitude (or number) diminishes by a factor of ten over this range. Because these motions

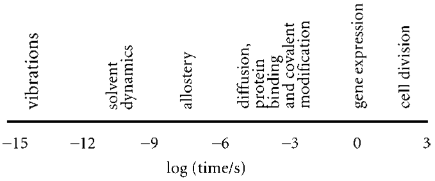

Figure 1. Types of protein motions which occur on timescales from femtoseconds to hours.

occur in the absence of bulk solvent motions, they must involve relatively local rearrangements between the various possible meta-stable configurations possible in a protein. The network properties of water at the surface of the protein and the voids created by thermal expansion play an important role in these types of motion; the motions disappear when a protein sample contains less than about 30% water.

It is possible to capture the essence of these motions with a relatively simple potential energy function and Newton's law of motion, $F = ma$, $V =$ bond stretch + bond angle + dihedral twist + atom-centered point charge electrostatics + Lennard–Jones.

The bond stretch and bond angle terms are simple harmonic potentials which allow the protein to vibrate. The dihedral twist term is a cosine function which allows transitions between rotamer states of single and double bonds. Dihedral twist parameters of the backbone will also influence the relative free energy of alpha helixes and beta sheets, two of the common secondary structure motifs described in the (*See* **Protein structure**). Much of the subtlety in this potential energy function lies in the electrostatics and Lennard–Jones terms. Both are pairwise additive interactions between atoms that are not bonded to each other or through another atom. The electrostatic interaction strength falls off as $1/d$, while the the Lennard–Jones term is of the form $A/d^{12} - B/d^{6}$, where d is the distance between atoms.

When the solvent is treated explicitly, parameters can be developed which reproduce important dynamic and thermodynamic properties that contribute to protein function. These include solvent properties, such as dielectric response time and amplitude, self-diffusion coefficients, and heat of vaporization. For amino acids and other small molecules, heats of solvation and heats of transfer from octanol to water are accurately reproduced. Heats of solvation and some aspects of the hydration shells of ions are also correctly modeled.

When this potential energy function is applied to protein folding and dynamics, it produces a prediction for the ensemble of protein conformations, as well as rates of water penetration, motions of loops, folding temperature, and the ensemble of unfolded states.

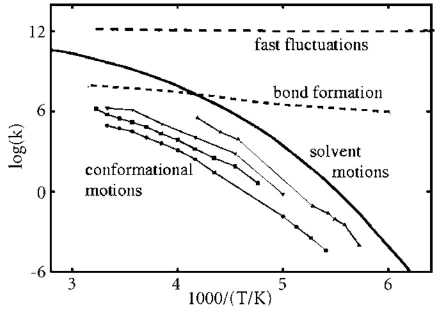

Figure 2. Arrhenius plot, showing the temperature-dependent rates of three types of experimentally measured protein motions, compared with the rate of solvent motions. (Taken from Fenimore et al., 2002, and references therein.)

In practice, it is difficult or impossible to obtain the necessary sampling to unambiguously quantify these properties. The reason is that a solvated protein simulation involves about 30,000 atoms; thus the integration time for the dynamics is limited to about 2 fs, while many of the motions of interest occur on the nanosecond to microsecond timescales (see Figure 2). Much effort in this field over the past two decades has centered on methods to improve the sampling through efficient implementation, parallelization, and statistical physics.

Covalent bond formation is the primary function of enzymes and involves many of the complexities of fast fluctuations combined with the need for a quantum-mechanical description of electronic polarizability, bond distortion, and bond formation. Such descriptions are available through all-electron quantum chemistry techniques, such as density functional theory, but calculations require approximately 1 cpu/day to evaluate energies and forces for 200 atoms. Calculations can be efficiently parallelized, and when more than 200 protein atoms are included, computation time scales linearly with system size. It is also possible to treat the bond-formation region quantum mechanically and the rest of the protein and solvent classically in a single calculation.

The difficulties involved in computing reaction rates include proper treatment of the vibrational dynamics of bond formation, sampling over hydration states, and sampling over the ensemble of protein conformations. One useful simplification is that many enzymes exclude water from their active site, considerably reducing the disorder.

Conformational motions of proteins can involve simple shifts of helixes, repacking of amino acid side chains, or more extensive changes between active and inactive conformations. The molecular dynamics model described above makes a prediction for conformational changes of proteins, but it is only recently that computer power has reached the long times necessary to test these predictions (Karplus & McCammon, 2002; Mayor et al., 2003). It is likely that some tuning of parameters will be necessary to predict protein properties with the same reliability with which these models currently predict small molecule properties.

Because the solvent provides a cage which prevents the protein from moving freely, the fraction of possible configurations explored increases by several orders of magnitude when the solvent is considered implicitly (Takada, 1999). While it is doubtful that a single parameter set can be developed to treat all proteins with an implicit solvent, it is clear from models of protein folding that much can be learned from implicit solvent models which are explicitly, but weakly, biased toward a known folded conformation.

BENJAMIN H. MCMAHON AND PAUL W. FENIMORE

See also **Biomolecular solitons; Local modes in molecules; Molecular dynamics; Pump-probe measurements**

Further Reading

Austin, R.H., Beeson, K.W., Eisenstein, L., Frauenfelder, H. & Gunsalus, I.C. 1975. Dynamics of ligand-binding to myoglobin. *Biochemistry*, 14: 5355–5373

Fenimore, P.W., Frauenfelder, H., McMahon, B.H. & Parak, F.G. 2002. Slaving: solvent fluctuations dominate protein dynamics and functions. *Proceedings of the National Academy of Sciences, USA*, 99: 16047–16051

Karplus, M. & McCammon, J.A. 2002. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9: 646–652

Mayor, U., Guydosh, N.R., Johnson, C.M., Grossmann, J.G., Sato, S., Jas, G.S., Freund, S.M.V., Alonso, D.O.V., Daggett, V. & Fersht, A.R. 2003. Complete folding pathway of a protein from nanoseconds to microseconds. *Nature*, 421: 863–867

McMahon, B.H., Muller, J.D., Wraight, C.A. & Nienhaus, G.U. 1998. Electron transfer and protein dynamics in the photosynthetic reaction center. *Biophysical Journal*, 74: 2567–2587

Takada, S. 1999. Go-*ing* for the prediction of protein folding mechanisms. *Proceedings of the National Academy of Sciences, USA*, 96: 11698–11700

PROTEIN STRUCTURE

A protein's structure, flexibility, and activity depend on the primary sequence of amino acids encoded by DNA, the presence of small-molecule ligands, the presence of protein or nucleic acid binding partners, and its history of covalent modification by other enzymes. Figure 1 (in color plate section) shows the catalytic site of one protein, RNA polymerase. This protein is responsible for transcribing an organism's DNA into messenger RNA, initiating the process of protein production, an essential and highly regulated cellular process. The complex geometry of interactions which stabilize the DNA, RNA, and nucleotide being added to the RNA are highly conserved among polymerases from all organisms, from microbes to humans. It is typical of enzymes that they are precisely (within about 0.1 Å) constructed near the catalytic site in order to exclude water and polarize and orient the reactants, yet still maintain the flexibility to allow reactants to enter and exit the active site.

One reason proteins can adopt highly rigid active sites, yet still undergo the large motions necessary for substrate binding, is that the peptide backbone is stabilized by multiple interactions in particular structural motifs, such as helixes, sheets, and particular types of hairpin turns. Thus, not only are the residues directly responsible for catalysis highly conserved among organisms, but also the arrangement of helixes, sheets, and tightly packed hydrophobic residues which determine the large-scale dynamics of the protein. This is illustrated in page 7 of the color plate section, which shows the entire catalytic domain of RNA polymerase, color-coded according to the polarity and charge of the amino acids. It is useful, looking at this figure, to reflect on the numerous cooperative interactions among the 1034 amino acids of this protein which cause it to fold up in this shape, rather than the shape of any of the other 30,000 proteins, or a glob of tangled polymer. These interactions are predominantly hydrophobicity, shape, hydrogen bond formation, and charge complementarity.

Protein function entails not only catalysis, but also the ability to be regulated by other proteins. This is done either by covalent modification, such as phosphorylation (covalent attachment of PO_3^{2-}) or acetylation (covalent attachment of C_2O), or by complex formation with other proteins. In either case, all the residues at the surface and the geometry of rigid motifs communicating the surface interactions to the active site become important determinants of protein structure. This is illustrated in the color plate section, showing the complex of twelve proteins, DNA, and RNA that is required for RNA polymerase to transcribe DNA to messenger RNA, as an early step in protein synthesis.

We have used the polymerase protein to illustrate several aspects of protein structure. There are between 10,000 and 100,000 types of protein structures used in various organisms, and a significant fraction of them have had their structure experimentally determined by diffraction of X-rays from a protein crystal. These structures are collected in a publicly accessible protein data bank, and several free software programs exist to aid in visualizing various aspects of protein structures.

The single, most important determinant of a protein's structure is the primary sequence of amino acids, encoded by an organism's DNA. Advances in microbiology have allowed automated sequencing of complete genomes of numerous organisms. While tens of thousands of protein structures are known, several million protein sequences are known, from all types of living creatures, from mammals to plants, microbes, and viruses. Proteins can be identified from their sequence of amino acids by searching for patterns of amino acids with BLAST (basic local alignment search tool), or tailoring a search algorithm based on numerous examples of a particular protein with a so-called hidden Markov model. Sequence analysis has been used to annotate both complete and partial genomes, and the results are available in several public databases.

The great excess of sequence data over structural data, combined with the insight into protein function and interactions provided by knowledge of the protein structure, provokes the question of whether it is possible to predict a protein structure from its primary sequence of amino acids. The essential reason this problem is difficult is that amino acids have an average of four dihedral angles which are equally stable in any of two or three positions (0°, $+120°$ and $-120°$), creating $3^{4\times 150}$ possible structures for each protein. The second reason is that when proteins were optimized by evolution, they exploited subtleties of interactions which are not necessarily captured by simple potential energy functions. Thus, while it has long been clear what forces stabilize folded proteins (hydrophobic interactions, hydrogen bonding, and Coulomb interactions), it has only recently been possible to relate the folding of a sequence of 20 amino acids to interaction potentials derived from small-molecule thermodynamic properties.

When it is possible, the most reliable method to predict a protein's structure is to find a protein of known structure which has greater than about 30% sequence identity. The particular conformations of most amino acids can then be inferred from the known protein structure, and resulting problems can usually be corrected by a Monte Carlo search algorithm or by hand. This process is known as homology modeling. Systematic efforts have produced enough experimentally determined structures to cover most of the observed sequences.

If a suitable template protein is not available, the protein structure must be constructed from observed structural motifs. Three of the most useful rules are: (1) hydrophobic residues tend to occupy the interior of the protein; (2) beta sheets, alpha helices, and other types of turns have characteristic amino acid content; and (3) triplets of amino acids tend to make particular configurations. Progress in this so-called ab initio folding can be evaluated by observing the result of the CASP competition, where a dozen experimentally determined structures of novel proteins are held back, while competing research groups are given several months to enter predictions of their structures.

Benjamin H. McMahon and Montiago X. LaBute

See also **Hydrogen bond; Molecular dynamics; Polymerization; Protein dynamics**

Further Reading

Alberts, B. et al. 2002. *Molecular Biology of the Cell*, 4th edition, New York: Garland Science

Branden, C. & Tooze, J. 1999. *Introduction to Protein Structure*, New York: Garland Science

Cramer, P., Bushnel, D.A. & Kornberg, R.D. 2001. Structural basis of transcription: RNA polymerase II at 2.8 Angstrom resolution. *Science*, 292: 1876

Petsko, G. A. & Ringe, D. 2003. *Protein Structure and Function*, Sunderland, MA: Sinauer Associates and Oxford: Blackwell Publishing

Valuable websites to persue:

THE PROTEIN DATA BANK: www.rcsb.org

CASP5 COMPETITION: predictioncenter.llnl.gov/casp5/Casp5.html.

BASIC LINEAR ALIGNMENT SEQUENCE TOOL (BLAST): www.ncbi.nlm.nih.gov/

MODELLER: www.salilab.org/modeller/modeller.html

VMD (A PROTEIN VISUALIZATION): www.ks.uiuc.edu/Research/vmd/

(SEARCHABLE GENE DATABASES): www.ensemble.org, www.ncbi.nlm.nih.gov/Entrez

PROTEIN FAMILY CATALOG: www.sanger.ac.uk/Software/Pfam/index.shtml.

PSEUDO-DIFFERENTIAL EQUATIONS

See **Equations, nonlinear**

PULSONS

See **Solitons, types of**

PUMP-PROBE MEASUREMENTS

Figure 1 shows a prototype pump-probe experiment in which short pump-laser pulse excites the vibrational oscillator (used here as a simple example) from the $\nu = 0$ vibrational state into the $\nu = 1$ vibrational state. A subsequent probe-laser pulse tests the $\nu = 1$ population by probing the $\nu = 1 \rightarrow \nu = 2$ excited state absorption, the $\nu = 1 \rightarrow \nu = 0$ stimulated emission, and/or the $\nu = 0 \rightarrow \nu = 1$ bleach. The time resolution of such an experiment is given by the pulse duration of the laser pulses, while the detector can be slow. With recent progress in laser technology, one can now perform pump-probe experiments in almost any spectral range, starting from far-IR (10 cm^{-1}) to soft X-ray radiation (1 keV) with a time resolution of much less than a picosecond (2000).

Pump-probe spectroscopy is a special form of third-order nonlinear spectroscopy. The theory of linear and nonlinear optical spectroscopy is generally performed in a perturbative expansion, where the electric fields of the light pulses act as weak perturbation onto the molecular Hamiltonian. The linear response, which

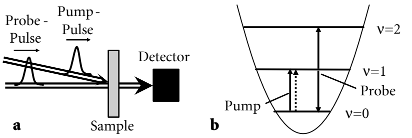

Figure 1. (a) A prototype pump-probe setup and (b) a prototype pump-probe experiment of a vibrator.

describes linear absorption spectroscopy, is given by

$$P^{(1)}(t) = \int_0^\infty \mathrm{d}t_1 R^{(1)}(t_1) \times E(t - t_1). \qquad (1)$$

Expressed in simple words, the electric field E of the probe light interacts once with the sample through the first-order response function $R^{(1)}$, generating a first-order polarization $P^{(1)}$, which subsequently is measured in the detector. The first-order response function $R^{(1)}$ is a material property, which can be tested by this type of spectroscopy. For instance, if the electric field E is chosen to be a monochromatic wave, one can measure the absorption spectrum by tuning its frequency.

Third-order spectroscopy is given by

$$P^{(3)}(t) = \int_0^\infty \mathrm{d}t_3 \int_0^\infty \mathrm{d}t_2 \int_0^\infty \mathrm{d}t_1 R^{(3)}(t_1, t_2, t_3) \\ \times E_1(t - t_3) E_2(t - t_3 - t_2) E_3(t - t_3 - t_2 - t_1). \qquad (2)$$

Again expressed in simple words, the electric fields E_1, E_2, and E_3 originating from a maximum of three laser pulses (in the example of Figure1a, both E_2 and E_3 originate from one pulse, i.e., the pump pulse) interact with the sample at three different time points $t - t_3$, $t - t_3 - t_2$, and $t - t_3 - t_2 - t_1$ through the third-order response function $R^{(3)}$, generating a third-order polarization $P^{(3)}$. The third-order response function $R^{(3)}$ contains significantly more information about the molecular system than the linear response function. Both can, in principle, be calculated once all eigenstates, transition dipole moments, and relaxation pathways of a system are known.

There are different types of third-order spectroscopies, the most important of which being pump-probe and photon echo spectroscopy. Both exist in many variations. They all measure the same third-order response function $R^{(3)}$ but differ by the choice of the three field interactions E_1, E_2, and E_3 (timing, frequency, beam direction, etc.). The different methods project different aspects of the complicated third-order response function $R^{(3)}$. A comprehensive discussion of the principles of nonlinear optical spectroscopy is given by (Mukamel, 1995). In the context of nonlinear science, the following advantages of third-order spectroscopies should be noted.

- In condensed phase systems, spectroscopic transitions are often significantly broadened. The most common broadening mechanisms are (i) lifetime broadening (T_1-relaxation), (ii) homogeneous broadening (T_2-relaxation), and (iii) inhomogeneous broadening. In solution phase systems, the situation becomes even more complicated since the distinction between homogeneous and inhomogeneous broadening becomes a question of the timescale of the fluctuating forces giving rise to dephasing. As a consequence, a continuous transfer between both regimes does, in general, occur. It is a common practice to fit absorption lines to certain line profiles such as a Lorentzian profile (homogeneous and/or lifetime broadening), a Gaussian profile (inhomogeneous broadening), a Voigt profile (a mixture of both), or a Kubo profile (intermediate regime). However, the assignment to different broadening mechanisms is model-dependent and often extremely questionable. Linear absorption spectroscopy principally cannot distinguish between different broadening mechanisms.

  The information boost of third-order nonlinear spectroscopy just makes that distinction possible. For example, pump-probe spectroscopy observes T_1-relaxation directly by populating a spectroscopic state with the pump-pulse and subsequently observing its relaxation back into the ground state with the probe-pulse (see Figure 1b). Photon echo experiments and transient hole burning experiments (a special form of pump-probe experiment with a spectrally narrow pump-pulses) can distinguish between the homogeneous, inhomogeneous, and the intermediate dephasing regime.
- When considering nonlinear vibrational states, such as the Davydov soliton and/or local modes in molecular crystals, a more subtle point is the following: The nonlinear third-order response of a system of linearly coupled harmonic oscillators (i.e., a crystal described in the harmonic approximation) vanishes exactly. This is because the three contributions to the nonlinear response function (bleach, stimulated emission, and excited state absorption, see Figure 1(b) all occur at the same frequency, and the transition dipoles are such that they cancel completely. Only anharmonicity (nonlinearity) of the molecular potential surfaces gives rise to a nonzero pump-probe signal. Hence, nonlinear pump-probe spectroscopy is specifically sensitive to that part of the molecular Hamiltonian which is also responsible for collective nonlinear phenomena. A linear absorption spectrum is, of course, nonzero also in the harmonic limit.

Using these properties, one can test predictions from soliton and self-trapping theories in a much more direct way than linear absorption spectroscopy can do. For example, Davydov has speculated about vibrational solitons in protein secondary structures, with the aim to explain the capability of proteins to efficiently store small quanta of energy (Davydov, 1979). As a first experimental attempt to verify this prediction, Austin and coworkers (Xie et al., 2000) have investigated the lifetime of the amide-I band (mostly the C=O stretching mode of the protein backbone) of myoglobin, which typically relaxes on an ultrafast 1 ps timescale. Interestingly, they found a somewhat prolonged lifetime (15 ps) in the blue wing of the absorption spectrum. Pump-probe experiments on crystalline acetanilide (ACN), a simple model system used to study nonlinear collective phenomena in protein structures (Scott, 1992), have revealed useful information, including the following (Edler et al., 2002): (i) The NH free exciton self-traps on an ultrafast 400 fs timescale. (ii) The phonons that mediate self-trapping can be identified. (iii) The anharmonicity of the amide-I states is a measure of their degree of delocalization. The free exciton is delocalized at a temperature of 90 K but Anderson (disorder) localizes with increasing temperature. On the other hand, the band that has been assigned to a self-trapped state is localized at all temperatures. (iv) The lifetimes of the initially excited states are short (1–2 ps), yet the ground state recovery is somewhat longer (20–40 ps). Apparently, energy is stored in the meantime by mechanisms that still need to be explored. None of this information can be obtained from linear absorption spectroscopy.

PETER HAMM

See also **Biomolecular solitons; Davydov soliton, Local modes in molecular crystals; Nonlinear optics**

Further Reading

Davydov, A.S. 1979. Solitons in molecular systems *Physica Scripta*, 20: 387–394

Edler, J., Hamm, P. & Scott, A.C. 2002. Femtosecond study of self-trapped vibrational excitons in crystalline acetanilide. *Physical Review Letters*, 88: 067403. See also Edler, J. & Hamm, P. 2002. Self-trapping of the amide I band in a peptide model crystal. *Journal of Chemical Physics*, 117: 2415–2424

Elsaesser, T. Mukamel, S., Murnanae, M.M. & Scherer, N.F. (editors). 2000. *Ultrafast Phenomena XII*, Berlin: Springer and previous books of the same series

Mukamel, S. 1995. *Principles of Nonlinear Optical Spectroscopy*, New York: Oxford University Press

Scott, A.C. 1992. Davydov's soliton. *Physics Reports*, 217: 3–67

Xie, A., van der Meer, L., Hoff, W. & Austin, R.H. 2000. Long-lived amide I vibrational modes in myoglobin. *Physical Review Letters*, 84: 5435–5438

Q

Q-DEFORMATION

See **Salerno equation**

QUADRATIC FAMILY

See **One-dimensional maps**

QUANTUM BILLIARDS

See **Billiards**

QUANTUM CHAOS

Classical equations of motion can have arbitrary nonlinearities and thus can show a wide variety of chaotic behavior. The Schrödinger equation for quantum systems, in contrast, is a linear partial differential equation, and the initial value problem cannot show chaotic behavior in the usual sense. However, in the transition regions between classical and quantum descriptions, especially for highly excited systems or short de Broglie wavelengths, chaos in classical dynamics will have its effects on quantum dynamics, the eigenstates, and the eigenenergies. The field of quantum chaos deals with the characteristic properties that emerge in this transitional region. Some features are typical for the time evolution of wave functions and closely related to the classical dynamics of trajectories and phase space densities. Others deal with the properties of eigenenergies and eigenstates. Both aspects can nicely be illustrated for hydrogen atoms in external fields.

A first example for the close relation between classical and quantum dynamics is provided by excitations of hydrogen in microwave fields. Exposed to a microwave field of a frequency of 9.923 GHz, a hydrogen atom starting from a state with main quantum number $n = 66$ needs to absorb about 75 photons in order to ionize. Experiments show little ionization for small intensities of the field and a rapid increase to almost complete ionization for larger fields (Figure 1a). A full quantum calculation is prohibitively complicated because each photon adds one order of perturbation theory. However, an analysis of the classical dynamics of an electron in a Coulomb field and the external microwave field, with Hamiltonian

$$H = \frac{\mathbf{p}^2}{2m} - \frac{1}{4\pi\varepsilon_0}\frac{1}{r} + Ff(t)\cos\omega t, \qquad (1)$$

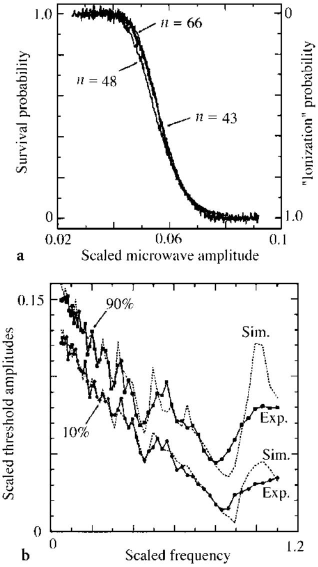

Figure 1. Microwave ionization of hydrogen atoms. The combinations $n_0\omega^3$ for the frequency and n_0^4F for the field strength are suggested by classical scaling. (Upper) Survival probability and ionization probability for different frequencies ω and different initial states n_0 that give the same scaled frequency as a function of field strengths. (Lower) Scaled threshold amplitude for the 10% and the 90% levels vs. scaled frequency. The comparison with classical simulations (dotted line) is very good and attests to the reliability of classical modelling of this process. For scaled frequencies $n_0^3\omega$ larger than 1, quantum effects become more important. (From Koch & van Leeuwen, 1995, with permission from Elsevier.)

where $f(t)$ describes the envelope of the field, reproduces the observed threshold amplitudes including many of the structures evident in Figure 1b very accurately.

The classical explanation for the fairly sharp onset is that for low field intensities, not all tori are destroyed, and the chaotic regions are separated by tori that block the transitions between them. For higher field intensities, the tori break up, the chaotic regions grow together, and electron trajectories can explore large regions of phase space, eventually escaping to infinity. This break up of tori is classically a rather dramatic event and can thus explain easily a fairly sharp onset of ionization. Many experiments since have verified the dependence on the initial state and the field intensities. Peculiar structures in the quantum ionization curves were also noted and could be explained as quantum effects due to almost degenerate quasienergy states. For the history of the problem and a careful analysis of the experiments and theories, see the review by Koch & van Leeuwen (1995).

Closely related to the microwave ionization of hydrogen is the dynamics of cold atoms in a standing light field. This problem is an experimental realization of the quantized kicked rotator, a standard model in classical and quantal chaos. Many properties of the quantized kicked rotator are described in Casati & Chirikov (1995); the first experimental realization is described in Moore et al. (1994).

For stationary problems, time can be separated and the eigenfunctions and eigenvalues of the Schrödinger equation can be analyzed. A good example is hydrogen in strong magnetic fields when the quadratic part of the vector potential cannot be neglected: the Hamiltonian then becomes

$$H = -\frac{\hbar^2}{2m}\Delta - \hbar\omega L_z - \frac{1}{4\pi\varepsilon_0}\frac{1}{r} + \frac{m\omega^2}{2}(x^2+y^2) \quad (2)$$

with the magnetic field expressed through its cyclotron frequency $\omega_c = eB/(2m)$. In order for the energy content in the cyclotron motion to be compatible to the spacing between the ground state and the first excited state, the magnetic field has to be of the order of 10^5 T, values reached and exceeded on the surface of neutron stars only. However, if the initial state is an excited state with main quantum numbers of about $n = 40$, already a few Tesla suffice to bring the system way out of the linear Zeeman regime and to scramble the spectra completely (Figure 2). Many properties of the system are described in Friedrich & Wintgen (1989), and in the contribution by Delande to Giannoni et al. (1991).

In such strong fields, there are not enough quantum numbers anymore to allow one to label the eigenstates uniquely. In a similar situation, nuclear physicists introduced statistical measures for the distribution of eigenvalues (*See* **Random matrix theory**).

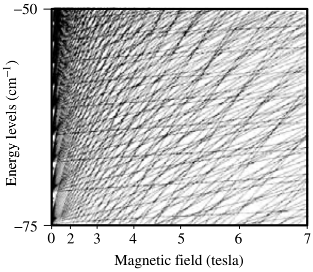

Figure 2. Energy levels for hydrogen in a magnetic field, for field strengths up to 7 T and initial quantum numbers between 38 and 45. The linear splitting within the Zeeman regime is limited to a region with very small fields. Most of the apparent crossings between eigenvalues are in fact avoided. (From Friedrich & Wintgen, 1989, with permission from Elsevier.)

Many investigations have shown that the statistics of eigenvalues for hydrogen in sufficiently strong magnetic fields, such as the level spacing distribution or two point correlation functions, are in very good agreement with random matrix theory (Friedrich & Wintgen, 1989). The same applies to many other classically chaotic systems, including vibrations of molecules, particles trapped in billiards, and model Hamiltonians with cubic and higher-order potentials, see the contribution by Bohigas to Giannoni et al. (1991) and the books by Brack & Bhaduri (1997), Haake (2001), and Stöckmann (1999).

While the distribution of eigenvalues shows statistical features, there are also deterministic ones: the Gutzwiller trace formula relates in the semiclassical limit eigenvalues E_i of the quantum system to classical periodic orbits and their properties (Gutzwiller, 1990; Chaos Focus Issue, 1992; Friedrich & Eckhardt, 1997),

$$\rho(E) = \sum_i \delta(E - E_i) \sim \operatorname{Re} \sum_p A_p \mathrm{e}^{\mathrm{i}S_p(E)/\hbar}, \quad (3)$$

$S_p = \oint \mathbf{p}\, \mathrm{d}\mathbf{q}$ is the classical action. The amplitude A_p carries information about caustics (through their phase shifts) and the instability of the classical orbits: the more unstable the smaller the weight. Because the density of states follows from a quantum amplitude (rather than a probability), the quantum amplitude is in magnitude the square root of the classical probability amplitude (*See* **Periodic orbit theory**). Expanding linearly around some reference energy E_r, the action becomes $S(E) \approx S(E_r) + T(E_r)(E - E_r)$, and a Fourier transform in E will results in peaks at the periods of periodic orbits in the system. This was noted in experiments by Karl Heinz Welge and his group and in numerical calculations by Friedrich & Wintgen (1989) (see also the comparison in Figure 3). Thus, on energy scales larger than a mean spacing

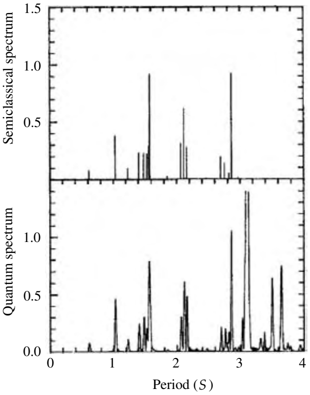

Figure 3. (Upper) The periodic orbits with periods *S* less than 3 and their semiclassical amplitudes and (Lower) a Fourier transform of a cross section of the quantum spectra in Figure 2.

and compatible with classical dynamics, the density of states is modulated by periodic orbits. Such periodic orbit spectroscopy can be used to extract the orbits and, thus, system specific properties.

In the presence of large-scale classical chaos and the accessibility of large parts of phase space by a single trajectory, one might expect that the wave functions are fairly uniformly spread out over the classically ergodic region. For billiards, this is the content of Shnirelman's theorem. However, as first discussed by Heller (1984) (see also Heller's contributions to Giannoni et al., 1991), periodic orbits leave their imprint on wave functions. One can find an enhanced intensity near a periodic orbit (a "scar") that can be detected in enhanced cross sections for certain processes.

The possibilities of quantum chaos are not limited to the realm of quantum phenomena proper: all that is needed is a wave theory, ideally with a classical dynamics for the propagation of wave fronts in the short wavelength limit. Examples outside the quantum world include resonances in microwave cavities, vibrations of solids, and acoustic waves in concert halls. In all cases level repulsion and other features can be found.

Many of the general properties of chaotic wave systems are discussed in the review Eckhardt (1988); in the textbooks by Brack & Bhaduri (1997), Gaspard (1998), Gutzwiller (1990), Haake (2001), and Stöckmann (1999); and in the contributions to Casati & Chirikov (1995), Giannoni et al. (1991), Chaos Focus Issue (1992), and Friedrich & Eckhardt (1997).

BRUNO ECKHARDT

See also **Periodic orbit theory; Random matrix theory I–IV; Regular and chaotic dynamics in atomic physics**

Further Reading

Brack, M. & Bhaduri, R.K. 1997. *Semiclassical Physics*, Reading, MA: Addison-Wesley

Casati, G. & Chirikov, B.V. 1995. *Quantum Chaos: Between Order and Disorder*, Cambridge and New York: Cambridge University Press

Cvitanoic, P. (editor). 1992. Chaos Focus Issue on Periodic Orbit Theory, *Chaos*, 2: 1–158

Eckhardt, B. 1988. Quantum mechanics of classically non-integrable systems. *Physics Reports*, 163: 205–297

Friedrich, H. and Eckhardt, B. (editors). 1997. *Classical, Semiclassical and Quantum Dynamics in Atoms*, Berlin and New York: Springer

Friedrich, H. & Wintgen, D. 1989. The hydrogen atom in a uniform magnetic field — an example of chaos. *Physics Reports*, 183: 37–79

Gaspard, P. 1998. *Chaos, Scattering and Statistical Mechanics*, Cambridge and New York: Cambridge University Press

Giannoni, M.J., Voros, A. & Zinn-Justin, J. (editors). 1991. *Chaos and Quantum Physics*. Proceedings Les Houches summer school 1989, Amsterdam and New York: North-Holland

Gutzwiller, M.C. 1990. *Chaos in Classical and Quantum Mechanics*, New York: Springer

Haake, F. 2001. *Quantum Signatures of Chaos*, Berlin: Springer

Heller, E.J. 1984. Bound-state eigenfunctions of classically chaotic Hamiltonian systems: scars of periodic orbits. *Physical Review Letters*, 53: 1515–1518

Koch, P.M. & van Leeuwen, K.A.H. 1995. The importance of resonances in microwave "ionization" of excited hydrogen atoms. *Physics Reports*, 255: 289–403

Moore, M.L., Robinson, J.C., Bharucha, C.F., Williams, P.E. & Raizen, M.G. 1994. Observation of dynamical localization in atomic momentum transfer: a new testing ground for quantum chaos. *Physical Review Letters*, 73: 2974–2977

Stöckmann, H.J. 1999. *Quantum Chaos: An Introduction*, Cambridge and New York: Cambridge University Press

QUANTUM FIELD THEORY

As our most fundamental description of physical phenomena, quantum field theory (QFT) is the natural culmination of classical mechanics, classical field theory, and quantum mechanics. To understand what QFT is and what makes it so special, it is worthwhile to briefly look at each of these structures in turn.

Classical mechanics describes the motion of objects in space and time in terms of well-defined positions and momenta that, taken together, form the phase space of the system. The location of an object in phase space can, in principle, be determined at any time, and once known suffices to predict its location at all future times given a complete enough knowledge of the forces at work. There are two main formulations of classical mechanics: the Hamiltonian formulation, which concentrates on how to find quantities at later times in terms of their known values at earlier times, and the Lagrangian one, which derives the same information from variational principles. These principles state that an object's trajectory in phase space extremizes a certain quantity called the action, which is the integral over the whole phase space history of a functional called a Lagrangian. The dimension of the phase space is usually finite, but can be enlarged beyond the usual 3 + 3 dimensions for the location and velocity of a point particle to allow for rotations of a rigid body or changes of shape of a deformable body. Relativity can be accommodated, but the ideas of quantum mechanics do not appear.

The move to an infinite number of phase space dimensions takes one from classical mechanics to classical field theory. Fluid mechanics, for example, allows for an infinite variety of shapes that a fluid can take making the phase space infinite. Sums that arose in classical mechanics are now integrals, and the general problems of analysis become much more challenging. Nevertheless, many problems are tractable, and theories based on classical fields with an infinite number of degrees of freedom have had great success. Notable among these is Maxwell's electrodynamics, which describes the electromagnetic field in terms of electric and magnetic fields that can vary in space and time and are present at all points of space and time. The phase space of classical electromagnetism is then infinite dimensional, and this accounts for much of the richness of the physics that it can describe.

Another shift away from classical mechanics is to maintain a finite number of degrees of freedom but to allow that points in phase space cannot be arbitrarily well known. The Heisenberg uncertainty principle placed constraints on the accuracy with which position and momentum could be known simultaneously, and the points of the phase space acquired a certain fuzziness. For example, in place of a point with well-defined position x and momentum p_x in the x-direction, one had something like a fuzzy disk of area roughly equal to the square of the Planck length, and shape determined by whether one tried to accurately determine x or p_x—one could only know one with a consequent sacrifice of information about the other. A system cannot now be thought of as a well-defined classical point moving around in phase space. Rather, information about the system is encoded in a function on the phase space called the wave function Ψ, and the theory allows all that can be known about the system to be recovered by applying various linear operators to Ψ.

The step to QFT involves allowing an infinite number of degrees of freedom in quantum mechanics. Thus, one might say that QFT is to quantum mechanics what classical field theory is to classical mechanics. The shift is profound and carries a number of important implications.

The first is that QFT takes as its phase space classical fields upon which an uncertainty paralleling that in quantum mechanics has been imposed. For example, the role of x could be taken by a classical field ϕ, and the role of momentum p_x by the rate of change of something (the Lagrangian) with respect to the time derivative of ϕ. This quantity is called the momentum conjugate to ϕ, and it is impossible to know both it and ϕ with arbitrary precision.

Time and space are now just dummy labels (parameters) in the theory and are integrated over, their main use being to enforce the notion of locality in a Lagrangian. This point is often misunderstood as people look for the "locations" of field quanta in physical space. In QFT, one looks at ϕ as a dynamical quantity as opposed to x. An excellent discussion of this point can be found in Schwinger (1970).

It turns out, perhaps somewhat tautologically, that one can often expand fields in Fourier series and identify each of the modes as a "particle" with momentum given by the mode numbers. It is a remarkable fact that experiments to detect fields in nature, when performed with sufficient resolution, always find them occurring in (or at least interacting as) discrete chunks or quanta. In this sense, QFT is a theory of particles, and the degree to which these quanta can be identified as isolated, discrete objects determines how well they can be thought of as particles.

As a field changes in time so too do the various Fourier modes and this is interpreted as saying that particles appear and disappear; in other words, particles can be created and destroyed in QFT. This is something foreign to both quantum and classical mechanics and was needed with the advent of high-energy physics experiments in which particles are routinely created and destroyed.

Two of many excellent texts describing how QFT is used in the calculation of physical processes are those by Peskin & Schroeder (1995) and by Weinberg (1995). The text by Itzykson and Zuber (1980) is also very good. A more philosophically inclined and less calculationally oriented reader will find much to think about in the book by Teller (1995). It is easy to suppose that QFT is intrinsically, or somehow must be, relativistic. In fact, the concepts can be very successfully applied to problems in condensed matter physics in nonrelativistic situations, and the interested reader will find the delightful introductory book by Mattuck (1992) and the more advanced book by Abrikosov et al. (1975) excellent places to start.

Several caveats are in order. First of all, in the relativistic case (the real world), these quanta do not generally admit a localization in the sense of having well-defined classical-like positions. That is, they do not behave like billiard balls. Second, if interactions between them are very strong they may not act as intuition might suggest. For example, while electrons and photons are in many ways like what one might expect particles to be, protons and neutrons seem to be comprised of particles called quarks, which cannot be removed. One says that the interactions between quarks are so strong that they are "confined." Phonons are quantized vibrations in crystals that can be detected in neutron scattering experiments and act in many ways

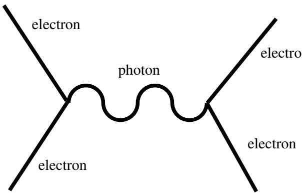

Figure 1. A Feynman diagram representing a first approximation to the scattering of one electron from another (the two straight lines) via the exchange of a photon (the wiggly line). Each diagram like this corresponds to a physically clear picture as well as to a well-defined mathematical expression contributing to the quantum mechanical amplitude for the process to take place. Only the topology of the diagram is important. Bending the lines around represents processes involving electrons and their antiparticles as different aspects of the same basic phenomenon, and connects scattering in space with scattering forward and backward in time, while not violating causality.

like particles, but they are only there insofar as there is a crystal lattice to vibrate. Any attempt to isolate a phonon by pulverizing a crystal and looking for one among the fragments is doomed to failure.

When it makes sense to think of weakly coupled particles, there is a well developed calculational framework called "perturbative quantum field theory" (pQFT), which allows one to make systematic estimates of physical quantities, often using sketches called Feynman diagrams. Figure 1 shows a Feynman diagram contributing to the scatting of one electron by another (the straight lines) via the exchange of a photon (the wiggly line).

The mathematical procedures involved are fraught with difficulties, and the series that arise are often divergent on physical grounds. In addition, the individual terms also tend to be infinite, and several procedures have been developed for handling such problems. The general approach is to first make ill-defined formal expressions finite by changing them ("regularizing" them) in a way controlled by some parameter, then comparing such regularized expressions only to one another and linking only such comparisons to observations ("renormalization"). Despite the feelings of many of the founders that there might be something deeply wrong with these procedures, it turns out that many quantities calculated in this way have given answers accurate to parts per million or better.

For strongly interacting theories the state of the art is much less well-developed. Powerful computers can be used to treat theories that are based on restricting fields to live on a grid or lattice that is meant to approximate space–time, but with a finite number of points. Analytic techniques also offer some hope, but the problems are formidable.

One of the most dramatic consequences of QFT is that the vacuum becomes a very complex object. Recalling the notion of the uncertainly principle applied to fields, one cannot set both a field and its conjugate momentum equal to zero. This, in fact is the case for each possible mode of a field and leads to an infinite energy in the vacuum due to these fluctuating, uncertain fields. As mentioned earlier, infinities like this are basically swept under the rug by insisting that only comparisons are meaningful, but in this case there are remarkable physical consequences.

For example, there are more modes that can exist for the electromagnetic field between two parallel metal plates that are far apart than for ones that are closer together. The infinite energies between the plates in these two cases can be compared and the result is finite—there is less energy between two plates that are close together than two that are far apart. This implies that two parallel metal plates, in order to reduce the energy between them (in empty space), will pull together. This is called the "Casimir effect" and has actually been observed in the laboratory, making it clear that QFT and its strange vacuum are more than theoretical fantasies.

Although it is often not made clear, one final aspect of QFT that differs from quantum mechanics is that QFT admits an infinite number of distinct vacua, all of zero energy. None of these vacua can be reached from another by a unitary transformation, making them physically distinct. This richness of "empty" space is a critical part of what makes QFT able to describe so much. To do justice to this point would require more space than is available, but the interested reader would do well to start with the book by Umezawa (1993). One consequence of this existence of unitarily inequivalent vacua is the amazing fact that the vacuum of one observer can be devoid of particles, while that of another may actually have particles present.

JOHN DAVID SWAIN

See also **Born–Infeld equations; Higgs boson; Quantum inverse scattering method; Skyrmions; String theory; Yang–Mills theory**

Further Reading

Abrikosov, A.A., Gorkov, L.P. & Dzyaloshinski, I.E. 1975. *Methods of Quantum Field Theory in Statistical Physics*, revised edition, translated and edited by Richard A. Silverman, New York: Dover

Itzykson, C. & Zuber, J.-B. 1980. *Quantum Field Theory*, New York: McGraw-Hill

Mattuck, R.D. 1992. *A Guide to Feynman Diagrams in the Many-Body Problem*, 2nd edition, New York: Dover

Peskin, M.E. & Schroeder, D.V. 1995. *An Introduction to Quantum Field Theory*, Reading, MA: Addison-Wesley

Schwinger, J. 1970. *Particles, Sources, and Fields*, Reading, MA: Addison-Wesley

Teller, P. 1995. *An Interpretative Introduction to Quantum Field Theory*, Princeton, NJ: Princeton University Press

Umezawa, H. 1993. *Advanced Field Theory*, New York: American Institute of Physics Press
Weinberg, S. 1995–2000. *The Quantum Theory of Fields*, 3 vols., Cambridge and New York: Cambridge University Press

QUANTUM INVERSE SCATTERING METHOD

The extension of the inverse scattering method to quantum theory, called the quantum inverse scattering method (QISM), was introduced at the end of the 1970s by the Leningrad (now St Petersburg) group and has been developed extensively since the 1980s. The QISM is significant in many respects. First, similar to the classical inverse scattering method, it provides a powerful method to solve the problem of quantum integrable systems. Second, it has a wide applicability to quantum particle systems, quantum field theoretic models, and quantum spin systems. Importantly, the QISM provides a unified view on exactly solvable models in physics. That is, it places completely integrable models in many-body theory and field theory and solvable models in statistical mechanics in a unified framework. Third, it has been the source of new mathematical objects and concepts such as Yangian algebra, quantum groups, q-deformation, and a revived knot theory. In these mathematical developments, the Yang–Baxter relation plays a central role.

As the simplest example, we begin with the quantum nonlinear Schrödinger (QNLS) model,

$$\mathrm{i}\phi_t + \phi_{xx} - 2\kappa\phi^\dagger\phi\phi = 0. \tag{1}$$

Here, $\phi(x,t)$ is the boson field operator, satisfying the equal-time commutation relation, $[\phi(x,t),\phi^\dagger(y,t)] = \delta(x-y)$. The Hamiltonian of the system is

$$H = \int \mathrm{d}x(\phi_x^\dagger\phi_x + \kappa\phi^\dagger\phi^\dagger\phi\phi). \tag{2}$$

The QNLS model (1) is considered to be the second-quantized theory of boson particles interacting via the pair-wise, delta-function potential. Motivated by the Zakharov–Shabat eigenvalue problem in the classical theory, we associate a linear auxiliary problem to (1),

$$\begin{aligned}\psi_{1x} + \frac{\mathrm{i}}{2}\lambda\psi_1 &= \mathrm{i}\alpha\phi^\dagger\psi_2,\\ \psi_{2x} + \frac{\mathrm{i}}{2}\lambda\psi_2 &= \mathrm{i}\varepsilon\alpha\psi_1\phi,\end{aligned} \tag{3}$$

where $\alpha = |\kappa|^{1/2}$ and ε denotes the sign of κ. We further assume the boundary condition $\phi(x) \to 0$ as $|x| \to \infty$, which is interpreted as the weak relation (a relation for the matrix elements). The Jost function operator $\Psi(x,\lambda)$ is defined by

$$\begin{aligned}\Psi(x,\lambda) &= \begin{bmatrix}1\\0\end{bmatrix}\exp\left(-\tfrac{\mathrm{i}}{2}\lambda x\right), \qquad x\to\infty,\\ &= \begin{bmatrix}A(\lambda)\exp\left(-\tfrac{\mathrm{i}}{2}\lambda x\right)\\ B(\lambda)\exp\left(\tfrac{\mathrm{i}}{2}\lambda x\right)\end{bmatrix}, \qquad x\to-\infty.\end{aligned} \tag{4}$$

The coefficients $A(\lambda)$ and $B(\lambda)$, corresponding to the transmission coefficient $a(\lambda)$ and the reflection coefficient $b(\lambda)$ in the classical theory, are called scattering data operators. The auxiliary problem (3) with (4) is solved perturbatively to express $A(\lambda)$ and $B(\lambda)$ in terms of the operators ϕ and $\phi^\dagger$ (direct problem). As a result, we obtain

$$[H, A(\lambda)] = 0, \qquad [H, B(\lambda)] = -\lambda^2 B(\lambda). \tag{5}$$

The former relation proves that the QNLS model has an infinite number of conserved operators. The latter indicates that the state

$$|k_1, k_2, \ldots, k_n\rangle = B^\dagger(k_1)B^\dagger(k_2)\ldots B^\dagger(k_n)|0\rangle, \tag{6}$$

where $|0\rangle$ is the vacuum state $\phi(x)|0\rangle = 0$, is an eigenstate of the Hamiltonian. Similar to the classical theory, the quantum Gel'fand–Levitan equation can be derived (inverse problem).

The state $|k_1, k_2, \ldots, k_n\rangle$ in (6) is valid irrespective of the sign of κ and describes the continuous state (Bethe ansatz state). For the attractive case ($\kappa < 0$), there appear bound states. A bound state (n-string state) is made of the complex momenta, $k_j = P/n - \mathrm{i}(n-2j+1)\kappa/2$, $j = 1, 2, \ldots, n$ and the energy is $E = \sum_{j=1}^n k_j^2 = P^2/n - \kappa^2 n(n^2-1)/12$. This bound state corresponds to the classical bright soliton in the $n \to \infty$ limit and may be called a quantum soliton.

To make the mathematical structure clear, it is useful to consider an operator version of an auxiliary linear problem defined on a one-dimensional lattice

$$\psi_{m+1} = L_m(\lambda)\psi_m, \qquad \mathrm{d}\psi_m/\mathrm{d}t = M_m\psi_m, \tag{7}$$

where $L_m(\lambda)$ and M_m are $M \times M$ matrix operators, and λ is the spectral parameter. For a quantum integrable system, it is found that direct products of two L_n operators with different spectral parameters satisfy a similarity relation

$$R(\lambda,\mu)[L_n(\lambda)\otimes L_n(\mu)] = [L_n(\mu)\otimes L_n(\lambda)]R(\lambda,\mu). \tag{8}$$

Here, the symbol $\otimes$ denotes the direct product of matrices, and $R(\lambda,\mu)$, called R-matrix, is an $M^2 \times M^2$ c-number matrix. Relation (8) is the Yang–Baxter relation for a quantum system on a lattice. If the $L_n(\lambda)$ operators on different sites commute, (8) leads to

$$R(\lambda,\mu)[\mathcal{T}_N(\lambda)\otimes\mathcal{T}_N(\mu)] = [\mathcal{T}_N(\mu)\otimes\mathcal{T}_N(\lambda)]R(\lambda,\mu), \tag{9}$$

where $\mathcal{T}_N(\lambda) = L_N(\lambda)L_{N-1}(\lambda)\cdots L_1(\lambda)$ is called the transition matrix or monodromy matrix. From (9), the transfer matrix

$$T_N(\lambda) = \mathrm{Tr}\mathcal{T}_N(\lambda) = \sum_{i=1}^{M} [\mathcal{T}_N(\lambda)]_{ii}, \qquad (10)$$

satisfies the commutation relation

$$[T_N(\lambda), T_N(\mu)] = 0. \qquad (11)$$

Relation (11) indicates the transition matrix $T_N(\lambda)$ is a generator of conserved operators. The λ (or λ^{-1}) expansion of $T_N(\lambda)$ gives a set of conserved operators $\{I_j\}$, which are involutive, $[I_i, I_j] = 0$. In addition, commutation relations among elements of the monodromy matrix offer an algebraic formulation of the Bethe ansatz method (algebraic Bethe ansatz method).

For quantum field theoretical models, the subscript N of the operator $T_N(\lambda)$ is understood as the system size that is to be sent to infinity. Then relation (11) proves the existence of an infinite number of involutive conserved operators. It is remarkable that we may consider $L_n(\lambda)$ and $R(\lambda, \mu)$ as vertices (local Boltzmann weights) of a vertex model in statistical mechanics. In this context, the Yang–Baxter relation (8) is a sufficient condition for the commutativity of the transfer matrix in statistical mechanics. Thus, quantum integrable models in $(1+1)$-dimension and solvable statistical models in 2-dimensions have a common property, a family of commuting transfer matrices.

Further, the quantum inverse scattering method and the Yang–Baxter relation have produced interesting developments in mathematics. Solutions of the Yang–Baxter relation with the difference property $R(\lambda, \mu) = R(\lambda - \mu)$ are classified into three classes, (i) elliptic, (ii) trigonometric, and (iii) rational functions. The trigonometric (rational) model is the critical case of the elliptic (trigonometric) model. Quantum groups and knot theory are due to the rich mathematical properties of the exactly solvable models at criticality. Extensive studies on the elliptic case are in progress.

Miki Wadati

See also **Bethe ansatz; Inverse scattering method or transform; Quantum field theory; Solitons**

Further Reading

Faddeev, L.D. 1981. Quantum completely integrable models in field theory. *Soviet Science Review of Mathematics and Physics,* C1: 107–155

Korepin, V.E., Bogoliubov, N.M. & Izergin, A.G. 1993. *Quantum Inverse Scattering Method and Correlation Functions*, Cambridge and New York: Cambridge University Press

Thacker, H.B. 1981. Exact integrability in quantum field theory and statistical systems. *Reviews of Modern Physics*, 53: 253–285

Wadati, M., Deguchi, T. & Akutsu, Y. 1989. Exactly solvable models and knot theory. *Physics Reports*, 180: 247–332

Wadati, M., Kuniba, A. & Konishi, T. 1985. The quantum nonlinear Schrödinger model: Gel'fand–Levitan equation and classical soliton. *Journal of the Physical Society of Japan*, 54: 1710–1723

QUANTUM NONLINEARITY

Although quantum mechanics is a linear theory, it is often used to analyze classical problems that are nonlinear. In the course of such analyses, it is interesting to consider how the nonlinear features of the classical dynamics are represented in the context of linear quantum theory, as is expected from the correspondence principle of Niels Bohr. In this entry, we see how the correspondence principle asserts itself to describe a variety of interesting nonlinear behaviors, including anharmonic oscillation, local modes, soliton binding energy, and soliton pinning.

A Nonlinear Oscillator

Consider the nonlinear mass-spring oscillator shown in Figure 1, where the spring potential is $V(x) = Kx^2/2 - \alpha x^4/4$. Under quantum theory, the oscillator energy (E) is not a continuous variable; it takes only discrete values that are eigenvalues of the time-independent Schrödinger equation

$$\left(\frac{\hbar^2}{2M}\right)\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + [E - V(x)]\psi = 0, \qquad (1)$$

where M is the oscillating mass and $\hbar$ is Planck's constant divided by 2π (Schiff, 1968). If α is zero, the restoring force of the spring is a linear function of its extension, and the classical oscillator of Figure 1 is linear. In this case, the energy eigenvalues are given by

$$E_n = (n + 1/2)\hbar\omega, \qquad (2)$$

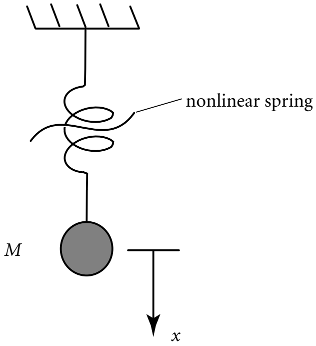

Figure 1. A nonlinear oscillator.

where $n = 0, 1, 2, \ldots$ indicates the number of quanta (called bosons) in the oscillation and $\omega = \sqrt{K/M}$ is the oscillation frequency of the classical oscillator. First derived by Erwin Schrödinger, the corresponding eigenfunctions (ψ_n) in Equation (1) are nth-order Hermite polynomials multiplied by a Gaussian factor—$\exp(-M\omega x^2/2\hbar)$—the first few of which are shown in Figure 2 (Schrödinger, 1926). A restriction of the linear ($\alpha = 0$) case is that transitions are allowed only between adjacent quantum levels. Thus, energy can only enter or leave the oscillator in quantum units of $\hbar\omega$, corresponding to Planck's radiation law.

If $\alpha \neq 0$, the classical oscillator is nonlinear, and two aspects of the quantum picture change. First, the energy eigenvalues are no longer given by the simple expression of Equation (2), and second, energetic transitions are not restricted to jumps between adjacent quantum levels. Both of these new features allow packets of energy to enter and leave the oscillator in amounts that differ from $\hbar\omega$ and, therefore, (according to Planck's law) at frequencies that differ from $\omega = \sqrt{K/M}$, in accord with observations of classical nonlinear oscillators.

In general, Equation (1) must be solved numerically to find its eigenvalues and eigenfunctions in the nonlinear case, but useful results can be obtained if the classical oscillator is considered in the rotating-wave approximation. Under this approximation, the energy eigenvalues are given by Scott (2003)

$$E_n = (\hbar\omega - \gamma/2)(n + 1/2) - \gamma n^2/2, \qquad (3)$$

where $\gamma \equiv 3\alpha\hbar^2/4M^2\omega^2$, and the corresponding eigenfunctions are identical to those of the linear oscillator (as in Figure 2). Equation (3) is an example of the Birge–Sponer relation, which is widely observed for nonlinear interatomic oscillations.

To represent the quantum dynamics of an oscillator, the (linear) superposition theorem allows construction of a wave packet of the form

$$\Psi(x, t) = \sum_{n=1}^{\infty} c_n \psi_n(x) e^{-iE_n t/\hbar}, \qquad (4)$$

where the c_n are complex constants that determine the initial position of the oscillator, $\Psi(x, t)$ is a solution of the linear, time-dependent Schrödinger equation, and the probability density of finding a particular position (x) is $\int |\Psi(x, t)|^2 \, dx$ (Schiff, 1968).

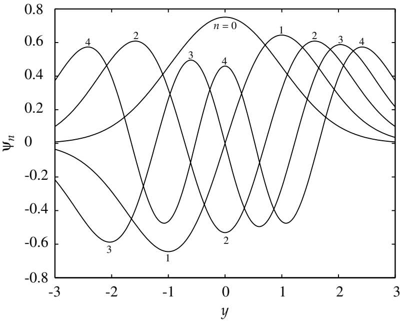

Figure 2. Harmonic oscillator wave functions, where $y = x\sqrt{M\omega/\hbar}$.

As the mass and spring constant become larger and the total energy of the oscillator is increased, the energy difference between adjacent levels becomes an ever smaller fraction of the total energy, and the discrete energy levels (the E_n) are better approximated as a continuous function. If the c_n in Equation (4) are chosen so that Ψ is a localized wave packet, the center of mass of this packet will follow the classical nonlinear trajectory. Thus the quantum model—which is necessary for describing interatomic oscillations—merges smoothly into the classical description that we use in our daily lives.

Local Modes in Molecules

Although empirical evidence for localization of vibrational energy in molecules has been available since the 1920s, such observations have been questioned because they seem to be at variance with the predictions of quantum theory. The benzene molecule (C_6H_6), for example, is invariant under a 60° rotation about its main axis; thus eigenfunctions of the (linear) quantum theory must be similarly invariant, but a local mode of CH-stretching vibration is not (see Figure 3).

This seeming contradiction between theory and experiment can be understood by noting that a local mode is represented not by a single eigenfunction but by a wave packet of them. For a local mode of n quanta, it turns out that the energy levels lie within a range of order (Scott, 2003)

$$\Delta E \sim \frac{n\varepsilon^n}{(n-1)!\gamma^{n-1}},$$

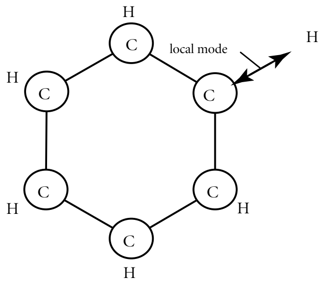

Figure 3. The planar structure of a benzene molecule, showing a local mode of the CH stretching oscillation.

where ε is the coupling energy between CH oscillations and γ is the anharmonicity as defined below Equation (3). For CH-stretching oscillations, $\varepsilon < \gamma$, so this energy range decreases very rapidly with increasing n, making the local-mode wave packet quasidegenerate at moderate quantum levels. This means that the wave packet will remain localized for times of order $\hbar/\Delta E$, which can be much longer than the time required to make an experimental observation.

Quantum Solitons

Consider next a chain (or one-dimensional lattice) of f classical nonlinear oscillators of the form

$$\left(\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}-\omega\right)A_j+\varepsilon(A_{j+1}+A_{j-1})+\gamma|A_j|^2A_j = 0, \tag{5}$$

where $A_j = A_{j+f}$ is a complex amplitude and j is an index running over periodic boundary conditions. In this formulation, ω is the oscillation frequency at one of the lattice points, ε is the interaction energy between adjacent oscillators, and γ is the anharmonicity of each oscillator in the rotating wave approximation. (With $f = 6$, Equation (5) provides a model for the benzene oscillations of Figure 3.)

This discrete nonlinear Schrödinger (DNLS) equation has solitary wave solutions that approach solitons of the continuum nonlinear Schrödinger (NLS) equation as $\gamma/\varepsilon \to 0$. (Be careful not to confuse the *linear* Schrödinger equation of quantum theory in Equation (1) with Equation (5), which is a *nonlinear, classical* equation. They are quite different.) Because Equation (5) is in the rotating wave approximation, the corresponding quantum problem can be formulated and solved (Scott et al., 1994).

For two quanta ($n = 2$) in the large f limit, energy eigenvalues are arranged as in Figure 4, where each eigenfunction changes by a factor of $\mathrm{e}^{\mathrm{i}k}$ under translation by one unit of j (lattice spacing); thus k is the "crystal momentum" of an eigenstate, (Haken, 1976). This figure shows both a continuum band (the shaded area) and a soliton band given by

$$\begin{aligned} E_2(k) &= \sqrt{\gamma^2 + 16\varepsilon^2\cos^2(k/2)} \\ &= E_2(0) + k^2/2m^* + \mathrm{O}(k^4), \end{aligned} \tag{6}$$

where m^* is the "effective mass" near the band center. The soliton band is characterized by two features. First, it is displaced below the continuum band by a binding energy $E_\mathrm{b} = \sqrt{\gamma^2 + 16\varepsilon^2} - 4\varepsilon$. Second, inspection of the corresponding eigenfunctions shows that the two quanta are more likely to be on the same site for the soliton band than in the continuum band.

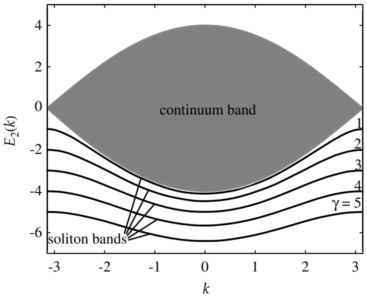

Figure 4. Energy eigenvalues of the DNLS equation at the second ($n = 2$) quantum level with $\varepsilon = 1$. There is one soliton band for each value of γ, which is plotted for several different values of γ.

For arbitrary n, $\gamma \ll \varepsilon$ and sufficiently large f, the quantum binding energy is

$$E_\mathrm{b} = \frac{\gamma^2}{48\varepsilon}n(n^2 - 1), \tag{7}$$

which corresponds to the binding energy of a classical NLS soliton under the identification $n = \sum |A_j|^2 \gg 1$ (Makhankov, 1990).

In the classical DNLS system with $\gamma \gg \varepsilon$, numerical studies show that the soliton becomes pinned to the lattice. Under quantum theory, this classical, nonlinear phenomenon is reflected by the fact that the effective mass,

$$m^* = \frac{(n-1)!\gamma^{n-1}}{2n\varepsilon^n},$$

becomes very large. Because the classical Ablowitz–Ladik (AL) system is completely integrable for all parameter values, its soliton is not pinned for $\gamma \gg \varepsilon$. This classical fact is reflected by an effective mass that approaches zero under the same conditions. Interestingly, the Salerno equation interpolates between the DNLS and AL limits.

Quantum Representation of a Classical Soliton

Although the quantum formulation can, in principle, be used to represent a classical soliton (on an optical fiber, for example), this is a rather complex dynamical object. The closest quantum representation of a classical soliton is provided by the coherent (or Glauber) state (Makhankov, 1990), which is constructed so that the complex wave amplitude—A_j in Equation (5)—is an eigenfunction of the bosonic lowering operator. This requires a wave packet comprising components with all values of the principal quantum number (n) (Scott, 2003). For a system with f degrees of freedom, a component at the nth level, in turn, has

$$\frac{(n+f-1)!}{(f-1)!\,n!}$$

terms, each of which is a complex construction of the eigenfunctions shown in Figure 2. (These are all of the different ways that n quanta can be placed on f freedoms.) As the number of terms at each level grows very rapidly with n, it becomes difficult to construct an exact quantum representation.

One way out of this computational dilemma is to go to the continuum approximation ($f \to \infty$) and employ the methods of quantum field theory (Lai & Haus, 1989). Another approach is to use a Hartree approximation (HA), which assumes that each boson feels the same mean-field potential as all of the other bosons (Scott, 2003). When this approximation procedure is carried through, it is found that the binding energy at the nth level is

$$E_{\mathrm{b}}^{\mathrm{HA}} = \frac{\gamma^2}{48\varepsilon} n(n-1)^2, \qquad (8)$$

which is less than the exact value by the factor $(n-1)/(n+1)$.

ALWYN SCOTT

See also **Discrete nonlinear Schrödinger equations; Hartree approximation; Local modes in molecules; Quantum field theory; Quantum theory; Rotating wave approximation; Salerno equation**

Further Reading

Haken, H. 1976. *Quantum Field Theory of Solids*, Amsterdam: North-Holland

Lai, Y. & Haus, H. 1989. Quantum theory of solitons in optical fibers. I. Time-dependent Hartree approximation. II. Exact solution. *Physical Review* A, 40: 844–866

Makhankov, V.G. 1990. *Soliton Phenomenology*, Dordrecht: Amsterdam

Schiff, L.I. 1968. *Quantum Mechanics*, 3rd edition, New York: McGraw-Hill

Schrödinger, E. 1926. Quantisierung als Eigenwertproblem. *Annalen der Physik*, 79: 361–376

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Scott, A.C., Eilbeck, J.C. & Gilhøj, H. 1994. Quantum lattice solitons. *Physica* D, 78: 194–213

QUANTUM THEORY

Quantum theory was born in 1900 with an important paper by Max Planck on black-body radiation. In this paper, he introduced quanta of vibrational energy E which are proportional to their frequency ν, and his famous constant currently known to take the value (Mohr & Taylor, 2000)

$$\frac{E}{\nu} = h = 6.62606876(52)\ 10^{-34}\ \mathrm{J\,s}.$$

Ten years after its formulation, Planck's hypothesis had been extraordinarily successful, explaining the light spectrum of black-body radiation, the photoelectric effect, and the low-temperature reduction of heat capacity in solids.

Using Rutherford's planetary model of atoms, Niels Bohr showed in 1913 that Planck's hypothesis helps to understand the spectral lines of hydrogen by deriving their frequencies in terms of h, and the electron mass and charge. The success of Bohr's atomic model led to the formulation of the Bohr–Sommerfeld quantization rule for classically integrable systems possessing as many invariants of motion as degrees of freedom:

$$\oint p_j\,\mathrm{d}q_j = h\left(n_j + \frac{\mu_j}{4}\right), \qquad j = 1, 2, ..., f,$$

where (q_j, p_j) are the canonically conjugate position-momentum variables in terms of which the system is integrable (also said separable), n_j are integers, μ_j are the so-called Maslov indices that characterize the topology of the motion (e.g., $\mu_j = 0$ for rotation, $\mu_j = 2$ for libration), and f is the number of degrees of freedom.

In 1917, Albert Einstein pointed out that the Bohr–Sommerfeld quantization rule cannot be applied to classically non-integrable systems (such as the helium atom), and it slowly became apparent that radically new ideas were required. In 1923, Louis de Broglie suggested that massive particles should behave as waves and he completed Planck's hypothesis by his famous relation, $p\lambda = h$, between the particle's momentum p and its wavelength λ.

Finally in 1925 and 1926, Werner Heisenberg and Erwin Schrödinger established quantum mechanics in two equivalent formulations. The first represents the observable quantities as matrices and the second is based on the famous Schrödinger equation

$$\mathrm{i}\,\hbar\,\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial \boldsymbol{r}^2} + V(\boldsymbol{r}, t)\,\psi, \qquad (1)$$

for the wave function $\psi(\boldsymbol{r}, t)$ of a particle of mass m moving in the energy potential $V(\boldsymbol{r}, t)$, with $\hbar = h/2\pi$ and $\mathrm{i} = \sqrt{-1}$. In 1927, Max Born proposed to interpret the square of the modulus of the wave function, $|\psi(\boldsymbol{r}, t)|^2$, as the probability density to observe the particle at position $\boldsymbol{r}$ and time t if the wave function is normalized to unity over the whole configuration space according to

$$\int |\psi(\boldsymbol{r}, t)|^2\,\mathrm{d}\boldsymbol{r} = 1. \qquad (2)$$

In this formulation, physical observables are represented by linear Hermitian operators $\hat{A}$ acting on the wave function. The expectation value of an observable is given by

$$\langle A\rangle = \int \psi^*(\boldsymbol{r}, t)\ \hat{A}\psi(\boldsymbol{r}, t)\ \mathrm{d}\boldsymbol{r},$$

if the particle is in the state described by the normalized wave function ψ. The wave property prevents a quantum particle from having simultaneously well-defined position and momentum as in the classical world. This impossibility (which follows from Fourier transform theory) is expressed by Heisenberg's uncertainty relation between the uncertainties on position Δx and momentum Δp

$$\Delta x \ \Delta p \geq \hbar/2,$$

which finds its origin in the noncommutativity of position and momentum operators.

The normalization condition (2) has the pivotal role of selecting the physically acceptable wave functions among all the possible solutions of the Schrödinger equation (1). In particular, it is the normalization condition that leads to the quantization of energy into well-defined eigenvalues associated with the stationary states. Without the normalization condition, the wave function would present spatial instabilities that are not physically meaningful.

Because of their spatial extension, wave functions are allowed to penetrate into classically forbidden regions where the classical kinetic energy of the particle would be negative. In these regions, the normalization condition (2) forces the wave function to decrease exponentially and precludes its instability. A consequence is the phenomenon of quantum tunneling, which manifests itself in cold electronic emission or α radioactivity and finds technological applications in Leo Esaki's semiconductor tunneling diode, Ivar Giaever's superconducting tunneling diode, and the electron tunneling microscope.

Together with the normalization condition, the Schrödinger equation shares with nonlinear systems the general scheme:

$$\text{instability} \ \rightarrow \ \text{saturation} \ \rightarrow \ \text{structure}.$$

In quantum mechanics, the instability mechanism is spatial and the saturation is provided by the normalization condition that selects spatially stable wave functions such as the electronic orbitals of atoms, molecules, and solids. The selection of normalizable wave functions generates the molecular structures of stereochemistry.

However, Schrödinger's equation is linear and thus obeys the principle of linear superposition (Dirac, 1930). Accordingly, the linear combination

$$\psi = \sum_n c_n \psi_n$$

(with complex numbers c_n) is a physically acceptable solution of Equation (1). Following studies by Steven Weinberg and others (Weinberg, 1989), extremely stringent limits have been put on hypothetical nonlinear corrections to quantum mechanics. These limits have been obtained by searching for nonlinearly induced detuning of resonant transitions between two atomic levels, putting the following upper bounds on a hypothetical nonlinear term $\Delta E(\psi, \psi^*)\psi$ supposed to correct the right-hand side of Schrödinger Equation (1):

$$|\Delta E| < 2.4 \ 10^{-20}\,\text{eV in } {}^{9}\text{Be}^{+} \text{ ion}$$
(Bollinger et al., 1989),
$$|\Delta E| < 1.6 \ 10^{-20}\,\text{eV in } {}^{201}\text{Hg atom}$$
(Majumder et al., 1990).

The principle of linear superposition of quantum mechanics has thus been confirmed by these investigations.

Nevertheless, in low-temperature many-body quantum systems, effective nonlinearities may arise if some wave function describing a subset of degrees of freedom has a feedback effect onto itself. Examples of effectively nonlinear quantum equations include the Hartree–Fock equation for fermionic systems, the Ginzburg–Landau equation for superconductors, the Gross–Pitaevskii equation for Bose–Einstein condensates, and the Ginzburg–Landau equation coupled to a Chern–Simon gauge field for the fractional quantum Hall effect.

Another problem where nonlinearities are associated with Schrödinger's equation is the theory of the optimal control of quantum systems by an external electromagnetic field. The optimal external field can be obtained as the solution of coupled nonlinear Schrödinger equations, where nonlinearity arises by a feedback mechanism of the quantum system onto itself through the external field and the desired control. Such feedback mechanisms have been experimentally implemented for laser control of chemical reactions (Rice & Zhao, 2000).

Nonlinear effects also emerge out of wave mechanics in the semiclassical limit, where the motion can be described in terms of classical orbits (solutions of nonlinear Hamiltonian equations). In the 1970s, starting from Schrödinger's equation, Gutzwiller derived a semiclassical trace formula that expresses the density of energy eigenvalues in terms of periodic orbits (Gutzwiller, 1990). The periodic orbits are unstable and proliferate exponentially in chaotic systems, where semiclassical quantization can be performed, thanks to the Gutzwiller trace formula as an alternative to the Bohr–Sommerfeld quantization rule.

In summary, nonlinear effects manifest themselves in quantum systems as phenomena emerging out of the linear wave mechanics in particular limits such as the semiclassical limit or the many-body limit at low temperature.

PIERRE GASPARD

See also **Bose–Einstein condensation; Hartree approximation; Nonlinear Schrödinger equations; Quantum chaos; Superconductivity; Superfluidity**

Further Reading

Bollinger, J.J., Heinzen, D.J., Itano, W.M., Gilbert, S.L., & Wineland, D.J. 1989. Test of the linearity of quantum mechanics by rf spectroscopy of the $^9\mathrm{Be}^+$ ground state. *Physical Review Letters*, 63: 1031–1034

Dirac, P.A.M. 1930. *Principles of Quantum Mechanics*, Oxford: Clarendon Press

Gutzwiller, M.C. 1990. *Chaos in Classical and Quantum Mechanics*, New York: Springer

Majumder, P.K., Venema, B.J., Lamoreaux, S.K., Heckel, B.R. & Fortson, E.N. 1990. Test of the linearity of quantum mechanics in optically pumped $^{201}\mathrm{Hg}$. *Physical Review Letters*, 65: 2931–2934

Mohr, P.J. & Taylor, B.N. 2000. CODATA recommended values of the fundamental physical constants: 1998. *Reviews of Modern Physics*, 72: 351–495

Rice, S.A. & Zhao, M. 2000. *Optical Control of Molecular Dynamics*, New York: Wiley

Weinberg, S. 1989. Precision tests of quantum mechanics. *Physical Review Letters*, 62: 485–488; and references therein

QUASILINEAR ANALYSIS

Analytical methods in the present-day theory of nonlinear oscillations originated in the investigations by Henri Poincaré, George D. Birkhoff, and Aleksandr Lyapunov, who laid the mathematical foundations for this theory. However, it should be noted that direct application of these mathematical methods to oscillation theory as such did not occur till much later, primarily owing to the work of Alexander Andronov (Andronov et al., 1966).

An important contribution to the development of the quantitative theory of nonlinear oscillations, especially of the applied part, was made by Balthasar van der Pol (van der Pol, 1934), who studied the operation of an electronic generator and proposed his own investigative method, namely, the method of slowly time-varying amplitudes. A rigorous justification of this method was later given by Osip Mandelshtam and N. Papaleksi (Mandelshtam & Papaleksi, 1934).

Almost independently of Mandelshtam, Andronov, and other physicists, the mathematical groundwork for nonlinear oscillation theory was laid by Nikolai Krylov, Nikolai Bogolyubov, Yuri Mitropol'sky (Krylov & Bogolyubov, 1947; Bogolyubov, 1950; Bogolyubov & Mitropol'sky, 1961; Mitropol'sky, 1971), and their disciples. They worked out the most important methods for the analysis of quasilinear oscillations: the asymptotic method, the averaging method, and the method of equivalent linearization. The last can serve as a theoretical justification of the heuristic methods of harmonic balance and statistical linearization (Landa, 1980; Pervozvansky, 1962), which are well known in mechanical and electrical engineering. Indeed, in accordance with the method of equivalent linearization, for a nonlinear function $f(x)$, we substitute a linear function λx, where λ is determined from the condition of minimization of the mean-square error $\varepsilon = \overline{\left(f(x) - \lambda x\right)^2}$. Differentiating ε with respect to λ and equating the derivative to zero, we find

$$\lambda = \frac{\overline{f(x)x}}{\overline{x^2}}, \tag{1}$$

where over-line denote the averaging operation. Depending on the averaging technique, we obtain harmonic linearization (harmonic balance) or statistical linearization. The former takes place if we put $x = \cos\omega t$ and average the numerator and denominator of expression (1) over t for the period $T = 2\pi/\omega$. The latter takes place if we suppose x to be distributed according to a certain probability distribution (for example, Gaussian) and find the statistical average.

The van der Pol method, or the method of slowly time-varying amplitudes, is applicable to near-linear, near-conservative self-oscillatory systems. The method was suggested by van der Pol as applied to self-oscillatory systems with one degree of freedom. However, the method may be easily generalized to self-oscillatory systems with n degrees of freedom, which are described by equations of the form

$$\ddot{y}_k + \omega_k^2 y_k = \mu Y_k(t, y, \dot{y}) \quad (k = 1, 2, \ldots, n), \tag{2}$$

where μ is a small parameter, $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, $\dot{\boldsymbol{y}} = \{\dot{y}_1, \ldots, \dot{y}_n\}$. For $\mu = 0$, Equations (2) are the equations of a linear oscillatory system with n degrees of freedom written in terms of normal coordinates.

As mentioned above, a rigorous justification of the van der Pol method was given by Mandelshtam and Papaleksi (Mandelshtam & Papaleksi, 1934). They showed that the truncated van der Pol equations can be found by averaging the initial equations over "fast time" for the period $T = 2\pi/\omega$. By doing so, Mandelshtam and Papaleksi pioneered the averaging method, the rigorous theory of which was worked out by Bogolyubov (1950) and developed by Mitropol'sky (1971). This theory concerns so-called equations in a standard form. By means of a certain change of variables, any equations describing oscillations in near-conservative systems can be reduced to equations in a standard form. More general theory is related to systems incorporating fast and slow variables (Volosov & Morgunov, 1971; Vasilyeva & Butuzov, 1990). In particular, this theory is used for analysis of relaxation oscillators (Mischenko & Rozov, 1975).

The asymptotic Krylov–Bogolyubov method is conceptually a generalization of the van der Pol method that allows us to calculate the higher approximations. The method has two modifications, depending on the form of the original equations and on the problem in question (Landa, 1980). If we are interested in the calculation of multi-frequency oscillations, we conveniently set the equations in the form of (2).

Provided we are interested in the calculation of single-frequency oscillations, we can set the original equations in the following form:

$$\dot{\boldsymbol{y}} + \mathcal{B}\boldsymbol{y} = \mu \boldsymbol{Y}(t, \boldsymbol{y}), \tag{3}$$

where $\boldsymbol{y}$ is a vector with n components, $\mathcal{B}$ is a square matrix with elements b_{jk}, μ is a small parameter, and $\boldsymbol{Y}(t, \boldsymbol{y})$ is a nonlinear vector function of time and of all components of the vector $\boldsymbol{y}$. We assume that for $\mu = 0$, Equations (3) describe a linear conservative system.

First, we consider the case when the original equations of a system are set in the form of (2). For $\mu \neq 0$, let us represent a solution of Equations (2) as a power series in μ:

$$\begin{aligned} y_k = {} & y_k^{(0)} + \mu u_{1k}(A, \psi, \mu t) \\ & + \mu^2 u_{2k}(A, \psi, \mu t) + \cdots, \end{aligned} \tag{4}$$

where $y_k^{(0)} = A_k \cos \psi_k$, $\psi_k = \omega_k t + \varphi_k$, A_k and φ_k are slowly time-varying functions obeying the equations

$$\begin{aligned} \frac{\mathrm{d}A_k}{\mathrm{d}t} &= \mu f_{1k}(A, \varphi, \mu t) + \mu^2 f_{1k}(A, \varphi, \mu t) + \ldots, \\ \frac{\mathrm{d}\varphi_k}{\mathrm{d}t} &= \mu F_{1k}(A, \varphi, \mu t) + \mu^2 F_{1k}(A, \varphi, \mu t) + \ldots. \end{aligned} \tag{5}$$

Here $u_{1k}(A, \psi, \mu t)$, $u_{2k}(A, \psi, \mu t), \ldots$ and f_{1k} $(A, \psi, \mu t)$, $f_{2k}(A, \psi, \mu t), \ldots,$ $F_{1k}(A, \psi, \mu t)$, F_{2k} $(A, \psi, \mu t), \ldots$ are unknown functions, which should be found.

Demanding the absence of resonant constituents in functions u_{1k}, we find the unknown functions f_{1k} and F_{1k}:

$$\begin{aligned} f_{1k} &= -\frac{1}{2\omega_k} X_{1k}(\mu t, A, \varphi), \\ F_{1k} &= -\frac{1}{2\omega_k A_k} Z_{1k}(\mu t, A, \varphi). \end{aligned} \tag{6}$$

Thus, we obtain the equations of the first approximation for the amplitudes and phases:

$$\begin{aligned} \frac{\mathrm{d}A_k}{\mathrm{d}t} &= -\frac{\mu}{2\omega_k} X_{1k}(\mu t, A, \varphi), \\ \frac{\mathrm{d}\varphi_k}{\mathrm{d}t} &= -\frac{\mu}{2\omega_k A_k} Z_{1k}(\mu t, A, \varphi). \end{aligned} \tag{7}$$

These equations coincide with those found by using the van der Pol method. The functions u_{1k} describe the higher harmonics and combination frequencies in the solution of the first approximation.

Using the next terms in the expansions found above, we can obtain the equations of the second and higher approximations.

Let us consider further the case when the original equations of a system are set in the form of (3) and we are interested in finding single-frequency oscillations described by

$$\boldsymbol{y} = A\Big(\boldsymbol{V}\mathrm{e}^{\mathrm{i}(\omega t + \varphi)} + \text{c.c.}\Big), \tag{8}$$

where ω is one of the system fundamental frequencies, $\boldsymbol{V}$ is the eigenvector of the matrix $\mathcal{B}$ corresponding to the frequency ω, and c.c. means the complex conjugate value.

Using the procedure of the method described, we obtain the equations for the unknown vector functions $\boldsymbol{k}_j$:

$$\begin{aligned} \omega\frac{\partial \boldsymbol{k}_1}{\partial \psi} + \mathcal{B}\boldsymbol{k}_1 = {} & -f_1\Big(\boldsymbol{V}\mathrm{e}^{\mathrm{i}\psi} + \text{c.c.}\Big) \\ & -\mathrm{i}AF_1\Big(\boldsymbol{V}\mathrm{e}^{\mathrm{i}\psi} - \text{c.c.}\Big) + \boldsymbol{Y}_1, \ldots \end{aligned} \tag{9}$$

Let us further expand the vector functions $\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots,$ $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ into the Fourier series with slowly time-varying coefficients:

$$\begin{aligned} \boldsymbol{k}_j(A, \psi, \mu t) &= \sum_{k=-\infty}^{\infty} \boldsymbol{U}_j^{(k)}(A, \varphi, \mu t)\mathrm{e}^{\mathrm{i}k\psi}, \\ \boldsymbol{Y}_j(A, \psi, \mu t) &= \sum_{k=-\infty}^{\infty} \boldsymbol{Y}_j^{(k)}(A, \varphi, \mu t)\mathrm{e}^{\mathrm{i}k\psi}. \end{aligned} \tag{10}$$

Substituting (10) into (9) and equating the coefficients of the same harmonics, we obtain, for each j, a system of non-uniform equations for the components of the vector functions $\boldsymbol{U}_j^{(k)}$. For $k \neq 1$, the determinant of this system is nonzero, and, hence, all of $\boldsymbol{U}_j^{(k\neq 1)}$ can be determined uniquely. For $k = 1$, the system determinant is zero. In this case we should require, for all j, the fulfillment of the compatibility conditions $\mathcal{A}\boldsymbol{R}_j = 0$, where $\boldsymbol{R}_j$ is the right-hand side of the jth system and $\mathcal{A}$ is the adjoint matrix. These conditions allow us to find $f_1(A, \varphi, \mu t)$, $F_1(A, \varphi, \mu t)$, $f_2(A, \varphi, \mu t)$, $F_2(A, \varphi, \mu t), \ldots.$

Let us consider the first approximation. From (9) and (10) we obtain the following equation for the vector function $\boldsymbol{U}_1^{(k)}$:

$$(\mathrm{i}k\omega\mathcal{E} + \mathcal{B})\boldsymbol{U}_1^{(k)} = -(f_1 + \mathrm{i}AF_1)\delta_{k1}\boldsymbol{V} + \boldsymbol{Y}_1^{(k)}, \tag{11}$$

where $\mathcal{E}$ is an identity matrix and δ_{k1} is the Kronecker delta. For $k = 1$, the compatibility condition of system (11) is

$$-f_1\mathcal{A}\boldsymbol{V} - \mathrm{i}AF_1\mathcal{A}\boldsymbol{V} + \mathcal{A}\boldsymbol{Y}_1^{(1)} = 0. \tag{12}$$

Splitting the real part and the imaginary part in Equation (12) we find

$$\begin{aligned} & f_1(A, \varphi, \mu t) = \mathrm{Re}\left\{\frac{1}{\mathrm{Sp}\,\mathcal{A}}\sum_j \frac{A_{jj}}{V_j} Y_{1j}^{(1)}(A, \varphi, \mu t)\right\}, \\ & F_1(A, \varphi, \mu t) \\ & \quad = \frac{1}{A}\,\mathrm{Im}\left\{\frac{1}{\mathrm{Sp}\,\mathcal{A}}\sum_j \frac{A_{jj}}{V_j} Y_{1j}^{(1)}(A, \varphi, \mu t)\right\}, \end{aligned} \tag{13}$$

where $\mathrm{Sp}\,\mathcal{A}$ is the spur of the matrix $\mathcal{A}$.

The second and higher approximations can be found in much the same way.

Polina Landa

See also **Averaging methods; Distributed oscillators; Linearization; Perturbation theory; Relaxation oscillators**

Further Reading

Andronov, A.A., Vitt, A.A. & Khaykin, S.E. 1966. *Theory of Oscillations*, Oxford and New York: Pergamon Press (original Russian edition 1959)

Bogolyubov, N.N. 1950. *Teoriya Vozmuscheniy v Nelineynoy Mechanike [Perturbation Theory in Nonlinear Mechanics].* Sbornik Instituta Stroitel'noy Mekhaniki AN USSR, No 14, 9–34

Bogolyubov, N.N. & Mitropol'sky, Yu.A. 1961. *Asymptotic Methods in the Theory of Nonlinear Oscillations*, New York: Gordon and Breach

Krylov, N.M. & Bogolyubov, N.N. 1947. *Introduction to Nonlinear Mechanics*, Princeton, NJ: Princeton University Press (original Russian edition 1937)

Landa, P.S. 1980. *Avtokolebaniya in Systemakh s Konechnym Chislom Stepeney Svobody [Self-Oscillations in Systems with a Finite Number of Degree of Freedom]*, Moscow: Science

Mandelshtam, L.I. & Papaleksi, N.D. 1934. Obosnovanie metoda priblizhennogo resheniya differentsial'nykh uravneniy [On justification of a method of approximate solving differential equations]. *ZhETF* 4: 117–121

Mischenko, E.F. & Rozov, N.Kh. 1975. *Differentsial'nye Uravneniya s Malym Parametrom i Relaksatsionnye Kolebaniya* [*Differential Equations with a Small Parameter and Relaxation Oscillations*], Moscow: Science

Mitropol'sky, Yu.A. 1971. *Metod Usredneniya v Nelineynoy Mechanike* [*The Averaging Method in Nonlinear Mechanics*], Kiev: Naukova Dumka

Pervozvansky, A.A. 1962. *Sluchaynye Processy v Nelineynykh Upravlyaemykh Sistemakh* [*Random Processes in Nonlinear Control Systems*], Moscow: Fizmatgiz

van der Pol, B. 1934. The nonlinear theory of electric oscillations. *Proceedings of IRE*, 22:1051–1086

Vasilyeva, A.B. & Butuzov, V.F. 1990. *Asimptoticheskie Metody v Teorii Singulyarnykh Vozmuscheniy* [*Asymptotic Methods in the Theory of Singular Perturbations*, Moscow: High School

Volosov, V.M. & Morgunov, B.I. 1971. *Metod Usredneniya v Teorii Nelineynykh Kolebatel'nykh Sistem [The Averaging Method in the Theory of Nonlinear Oscillatory Systems]*, Moscow: MSU Publ

QUASIPERIODICITY

The term *quasi-periodicity* implies a type of motion that is regular (nonchaotic) but consists of a combination of periodic motions with a trajectory that—after sufficient time—passes arbitrarily close to an earlier value.

For flow systems, a trajectory $x(t) \in \mathbf{R}^n$ is called k-quasi-periodic if it can be written in the form

$$x(t) = f(\omega_1 t, \ldots, \omega_k t).$$

Here f is a smooth nonlinear function of period 2π in each of its k arguments separately. The function f belongs to a class of almost periodic functions if frequencies $\omega_1, \ldots, \omega_k$ are not rationally related (are incommensurate); that is, $\sum_1^k l_i \omega_i \neq 0$ when $l_1, \ldots, l_k \in \mathbf{Z}$ and $\sum_1^k |l_i| > 0$ (Levitan & Zhikov, 1982). In other words, the ratio between two frequencies ω_i and ω_j, $i \neq j$ is irrational:

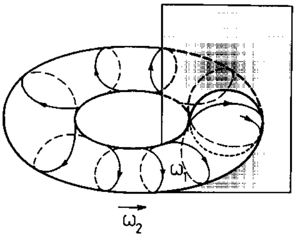

Figure 1. Schematic diagram of motion on the T^2 (two-dimensional) torus and the Poincaré cross section.

$$\frac{\omega_i}{\omega_j} \neq \frac{p}{q}, \quad p \in \mathbf{Z}, \quad q \in \mathbf{Z}. \tag{1}$$

A quasi-periodic trajectory lies on a k-dimensional torus (T^k), which can be an attractor, a repeller, or a saddle of a dynamic system. The term *quasi-periodicity* describes motion on the 2-dimensional torus (T^2) whereas *k-quasi-periodicity* is used for motion on tori of larger dimension. A two-dimensional closed surface in phase space corresponds to the two-dimensional torus T^2, and asymptotically (as $t \to \infty$), the trajectory covers its entire surface.

To describe quasi-periodic motion, it is convenient to use a Poincaré cross section (Figure 1) or a map. The T^2 torus is given by a closed invariant curve in Poincaré cross section. Maps are widely used in practice to investigate the properties of quasi-periodic motion. A typical map that serves for this purpose, as well as for investigation of the transition from quasi-periodicity to chaos, is the circle map:

$$\theta_{n+1} = \theta_n + \Omega - \frac{K}{2\pi} \sin 2\pi\theta_n, \quad \text{mod } 1. \tag{2}$$

Here θ is a phase angle defined in the interval [0; 1], while $\Omega \in [0; 1]$ and $K \geq 0$ are parameters of the map. Map (2) for $K < 1$ describes the dynamics of a non-autonomous self-oscillating system, for example, the van der Pol oscillator under external periodic forcing. The variable θ in this case represents the phase difference between the self-oscillations and the external force, the parameter Ω corresponds to the mismatch between the forcing frequency and the natural frequency of the oscillator, and the parameter K corresponds to the forcing amplitude.

A number of problems arise during numerical simulations and experimental investigations of attractors in quasi-periodic systems. First, it is impossible to simulate an irrational ratio in Equation (1) because each number stored in a computer is limited to a finite set of digits. Second, as a consequence of the measurement error in experimental investigations, it is, in practice, impossible to determinate reliably whether a frequency ratio is rational or irrational. Third, because the presence of noise in real systems leads to a smearing of the system's characteristics, it is possible to interpret experimental results wrongly. Thus, if a calculated invariant curve in a Poincaré cross section is smeared (has a finite width), then a noisy quasi-periodic motion can become entangled with chaos.

To obtain reliable information about the type of motion, it is necessary to use a set of characteristics such as the spectrum of Lyapunov exponents (SLE), a power spectrum, an autocorrelation function (ACF), or the smoothness of the invariant curve in the Poincaré cross section. In quasi-periodic motion, the SLE has no positive values, and the number of zeros is equal to the number of incommensurate base frequencies defined by the quasi-periodic motion. For example, the T^2 torus of a dissipative system is characterized by two Lyapunov exponents that are zero and others that are negative. The SLE of chaotic behavior includes at least one positive value. The power spectrum of quasi-periodic motion is discrete and consists of peaks at frequencies $n\omega_1 \pm m\omega_2$, with n and m taking arbitrary integer values. In reality, the peaks need not be mathematically sharp, but can be instrumentally sharp. The spectrum of a chaotic motion, on the other hand, is continuous and consists of an infinite number of base functions and their combinations. The ACF of a trajectory on a torus is an infinitely oscillating function that does not fall to zero as $t \to \infty$; in contrast, the ACF of a chaotic trajectory tends to zero as $t \to \infty$. In Poincaré cross section, a smooth curve corresponds to a torus, whereas a chaotic regime is characterized by a fractal structure.

Quasi-periodic motion can be observed in conservative and dissipative systems characterized by two or more incommensurate frequencies generated by the system and/or external sources (Glazier & Libchaber, 1988). The transition from a stationary state to quasi-periodic motion is realized as the result of at least two consecutive Hopf bifurcations, which add two incommensurate frequencies to the system, hence giving rise to a T^2 torus. Further Hopf bifurcations lead to a torus of larger dimension. Landau (1965) conjectured that turbulence arises through an infinite series of such Hopf bifurcations, increasing the degree of quasi-periodicity step-by-step up to infinity.

The possibility of such a scenario is problematic because the Kolmogorov–Arnol'd–Moser (KAM) theorem and the results of Newhouse et al. (1979) show that T^k tori can be structurally unstable when $k \geq 3$. This instability is found to increase with nonlinearity, which can lead to synchronization if an irrational frequency ratio becomes rational, or can transform a low-dimensional torus to chaos. However, there are also models (Ott, 2002) and experiments (Gollub & Benson, 1980), which prove that the tori T^k still exist for $k \geq 3$. The flow of blood through the cardiovascular system seems to be an example of dynamics characterized by a hypertorus, with $k \geq 3$ (Stefanovska & Bračič, 1999).

Several scenarios of transition from quasi-periodicity to chaos are known. The transition from T^2 to a strange chaotic attractor has been thoroughly investigated (Afraimovich & Shilnikov, 1991; Anishchenko et al., 2002). First, the torus loses smoothness, and then stable and saddle resonances (cycles corresponding to a rational frequency ratio) arise on the torus. Nonlocal bifurcations (homoclinic tangencies of manifolds) of these saddles lead to chaos later on.

A similar scenario is also observed for the T^3 torus (Anishchenko et al., 2002). The occurrence of torus doublings can precede the appearance of resonances on the torus: a finite number of bifurcations occur, after which the torus is destroyed as described above. The possible reality of an infinite series of torus doubling bifurcations (the Feigenbaum scenario of the transition to chaos) remains an open question, although renormalization group analysis (Feigenbaum et al., 1982) allows such a possibility theoretically.

Newhouse et al. (1979) have shown that small perturbations of the T^3 torus can lead to a strange chaotic attractor on the torus. Bifurcations, which lead from the torus to chaos, have not yet been studied in detail, but investigations indicate that the transition is connected with the appearance and destruction of resonances on the torus. Furthermore, the transition from T^3 to chaos has been observed only in maps, and the possibility of its realization in flow systems remains unclear.

A third scenario is the transition from quasi-periodicity to a nonstrange chaotic attractor (Arnol'd, 1983). If a chaotic attractor on T^3 covers all the surface of the torus, then the capacity of the attractor is integer and the largest Lyapunov exponent is positive.

Another way in which the transition from quasi-periodicity to chaos can occur is through the appearance of a strange nonchaotic attractor (SNA), which has a fractal structure and no positive values in the SLE (Grebogi et al., 1984). Such an evolution is observed in systems under external quasi-periodic forcing that drives the system with an irrational frequency ratio, which does not depend on parameters and properties of the system itself. The bifurcation mechanisms involved in the transition from torus to SNA, and the subsequent transition from SNA to chaos, are subjects of active investigation.

Igor A. Khovanov, Natalya A. Khovanova, and
Aneta Stefanovska

See also **Attractors; Bifurcations; Cat map; Kolmogorov–Arnol'd–Moser theorem; Lyapunov exponents; One-dimensional maps; Period doubling;**

Phase space; Recurrence; Routes to chaos; Synchronization

Further Reading

Afraimovich, V.S. & Shilnikov, L.P. 1991. Invariant two-dimensional tori, their breakdown and stochasticity. In *American Mathematical Society Translations*, Series 2, 149: 201–212 (original Russian edition 1983)

Anishchenko, V.S. et al. 2002. *Nonlinear Dynamics of Chaotic and Stochastic Systems: Tutorial and Modern Developments*, Berlin and New York: Springer

Arnol'd, V.I. 1983. *Geometrical methods in the Theory of Ordinary Differential Equations*, New York: Springer (original Russian edition 1978)

Feigenbaum, M.J., Kadanoff, L.P & Shenker, S.J. 1982. Quasiperiodicity in dissipative systems: a renormalization group analysis. *Physica* D, 5: 370–386

Glazier, J.A. & Libchaber A. 1988. Quasi-periodicity and dynamical systems: an experimentalist's view. *IEEE Transactions of Circuits and Systems*, 35: 790–318

Gollub, J.P. & Benson, S.M. 1980. Many routes to turbulent convection. *Journal of Fluid Mechanics*, 100: 449–470

Grebogi, C., Ott, E., Pelikan, S. & Yorke, J.A. 1984. Strange attractors that are non chaotic. *Phisica* D, 13: 261–268

Landau, L.D. 1965. On the problem turbulence. In *Collected Papers*, edited by D. ter Haar, Oxford: Pergamon (original Russian edition 1944)

Levitan, B.M. & Zhikov, V.V. 1982. *Almost Periodic Functions and Differential Equations*, Cambridge: Cambridge University Press (original Russian edition 1978)

Newhouse, S., Ruelle, D. & Takens, F. 1979. Occurence of strange Axiom A attractors near quasi-periodic flows on T^m, $m \geq 3$. *Communications in Mathematical Physics*, 64: 35–40

Ott, E. 2002. *Chaos in Dynamical Systems*, 2nd edition, Cambridge and New York: Cambridge University Press

Stefanovska, A. & Bračič, M. 1999. Physics of the human cardiovascular system. *Contemporary Physics*, 40: 31–55

R

RAMAN SOLITONS

See **Rayleigh and Raman scattering and IR absorption**

RANDOM MATRIX THEORY I: ORIGINS AND PHYSICAL APPLICATIONS

The textbook examples of quantum systems typically have a complete set of quantum numbers. For hydrogen-like atoms in a weak magnetic field, the set of radial, angular, and magnetic quantum numbers, together perhaps with the spin state, uniquely determines the energy levels of the atom. Nuclear physicists were among the first to encounter a situation where this is not possible. Scattering protons and neutrons off nuclei revealed a huge number of resonances that could no longer be completely labelled by their quantum numbers. In view of this fact and the uncertainty in the detailed interaction between the neutrons, Eugene Wigner advanced a statistical approach to the observed spectra: instead of trying to characterize each resonance individually, he proposed to look for a statistical description of their distribution and their strengths (see Wigner (1967) for an account of the early developments). Typical quantities of interest are then the mean density of resonances, probabilities for their spacings, the two- or more point correlation functions, and the distribution of transition strengths. The Hamiltonian was modeled as a random matrix, and by ingenious mathematical techniques a complete characterization of the spectral properties of certain ensembles of random matrices could be achieved. Wigner's John von Neumann lecture (Wigner, 1967), and the books by Mehta (1991) and Porter (1965) survey many of the results and methods. These sources also contain references to work in mathematics before the widespread use of random matrices in physics.

The most important ensembles are the Gaussian ones. Consider ensembles of $N \times N$ matrices H with statistically independent real or complex entries and a distribution function

$$P_N(H_{11}, H_{12}, \ldots, H_{NN}) = \prod_{i,j} p_{ij}(H_{ij}). \qquad (1)$$

It is natural to demand invariance of the distribution under transformation of bases, since none of them can be singled out a priori. If the allowed transformations are orthogonal, unitary, or symplectic, the Gaussian orthogonal ensemble (GOE), Gaussian unitary ensemble (GUE), or the Gaussian symplectic ensemble (GSE) is obtained, respectively. The distribution functions are

$$P_{N,\beta}=\left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\beta}{\pi}\right)^{\beta N(N-1)/2}\exp\left(-\frac{\beta}{2}\mathrm{tr}\,H^2\right) \qquad (2)$$

with the parameters $\beta = 1$ for GOE, $\beta = 2$ for GUE, and $\beta = 4$ for GSE.

Of the many quantities that can be studied and for which in many cases exact results can be obtained, we list only two: the level spacing distribution and the number variance (Figure 1). For others, see Brody et al. (1981), Guhr et al. (1998), Haake (2001), Mehta (1967), and Stöckmann (1999).

The level spacing distribution is the probability density to find two neighboring levels a distance s apart (in units of the mean level spacing). The expression that results from the random matrix ensembles is somewhat involved, but a very good approximation was given by Wigner:

$$P_{\mathrm{GOE}} = \frac{\pi}{2} s \mathrm{e}^{-(\pi/4)s^2}, \qquad (3)$$

$$P_{\mathrm{GUE}} = \frac{32}{\pi^2} s^2 \mathrm{e}^{-(4/\pi)s^2}, \qquad (4)$$

$$P_{\mathrm{GSE}} = \frac{2^{18}}{3^6\pi^3} s^4 \mathrm{e}^{-(64/9\pi)s^2}. \qquad (5)$$

The characteristic feature of these expression is that for small s the distribution increases like s^β, so that small spacings are suppressed.

The number variance measures the deviation between the true and the expected number of eigenvalues in an interval. In units of the mean spacing

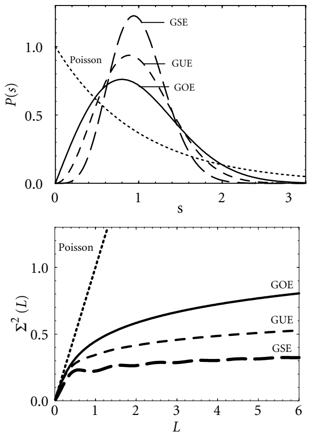

Figure 1. Level spacing distribution and number variance for the different ensembles.

the expected number equals the length L of the interval, and with $N(L)$ the true number, variance is

$$\Sigma^2(L) = \langle (N(L) - L)^2 \rangle = \langle (N(L))^2 \rangle - L^2 . \quad (6)$$

In a Poisson process for uncorrelated levels, the number variance will increase linearly, $\Sigma^2(L) = L$. In the random matrix ensembles, the increase is slower,

$$\Sigma^2(L) \sim \frac{2}{\beta\pi^2} \ln L + c_\beta \quad (7)$$

with constants c_β (for complete expressions, see, for example, Brody et al. (1981)). The slow logarithmic increase reflects the interactions between levels.

Between about 1975 and the mid-1980s, Michael Berry, Oriol Bohigas, and others applied such statistical measures to systems with a chaotic classical limit and observed level distributions and correlators in agreement with random matrix theory. They conjectured that chaotic systems will show random matrix behavior consistent with the global real symmetric, complex Hermitian, or symplectic symmetry of the Hamiltonian. The eigenvalue distribution for integrable systems, as shown earlier by Berry and Tabor, should show Poissonian statistics (Haake, 2001; Stöckmann, 1999). The argument for the two situations is roughly as follows: if the classical system is integrable, the quantum eigenstates are localized on tori and the energies can be approximated by Bohr–Sommerfeld quantization of the appropriate actions. Different tori give independent sequences of quantum eigenvalues, so that the collection of eigenvalues reflects a random appearance of eigenvalues, without correlations. As a result, the distribution of mean level spacings becomes exponential. In a chaotic system the eigenstates are spread over the connected chaotic components. They thus have the same support, but since they have to be orthogonal, they have to interact. This interaction then leads to level repulsion and the observed suppression of small spacings in $P(s)$. The conjecture is supported by semiclassical arguments for the two-point correlation function (Berry, 1985) and a large body of empirical evidence from experimental data for nuclei, atoms, and molecules and from numerical data for various systems (Stöckmann, 1999). The few exceptions (as, for instance, the arithmetical billiards on surfaces of negative curvature (Bogomolny et al., 1997)) have led to clarifications of the necessary requirements. But so far no proof has emerged, neither for integrable nor for chaotic systems.

The connection between random matrices and the chaotic behavior of the underlying wave behavior indicates why random matrix statistics should also appear in many nonquantum situations. Indeed, acoustic resonances in irregularly shaped containers, electromagnetic resonances in cavities, and mechanical vibrations of plates and other solid blocks all show statistical properties in good agreement with random matrix theory expectations when the short wavelength dynamics of wave fronts is chaotic (Stöckmann, 1999).

The statistical measures developed within random matrix theory have propagated into other fields as well, most notably into the theory of the Riemann zeta function. The Riemann zeta function as defined by

$$\zeta_{\mathrm{R}}(s) = \sum_{n=1}^{\infty} n^{-s} \quad (8)$$

can be analytically continued into the complex plane. It then has a pole at $s = 1$ and so-called trivial zeroes at $s = -2, -4, \ldots$. Bernhard Riemann's longstanding conjecture is that all its other zeroes lie along the line $s = \frac{1}{2} + \mathrm{i}t$. David Hilbert proposed to identify an eigenvalue problem that has these zeroes as eigenvalues. While that system is still elusive, analysis of the statistics suggests that it belongs to the universality class of the Gaussian unitary ensemble (see Berry & Keating (2000) for an account of the fascinating relations between random matrix theory, primes, and the statistics of the Riemann zeroes).

While all previous examples dealt with individual systems having a fixed Hamiltonian, one can also consider disordered systems, where as an additional statistical element, certain variations in the Hamiltonian enter. For instance, in a solid the mean density of impurities can be controlled experimentally, but the detailed positions and their effects on a specific experiment are difficult to fix. It is then possible to again set up random matrix models and to derive measurable quantities that are universal in that they depend on the global symmetry of the system only, see Beenakker (1997), Guhr et al. (1998), and Efetov (1997).

Through the connection to disordered systems, techniques from field theory, Grassmann algebras, and supersymmetry entered random matrix theory (Zirnbauer, 1996). Those tools allow analytical calculations of many quantities both in the extreme quantum limit and in the semiclassical limit. More importantly, they have indicated a link to group theory and homogeneous spaces. The classification into three universality classes (GOE, GUE, and GSE) above has to be extended to a total of 10 cases. The realizations of the other cases require additional symmetries. Three cases can be found by extending the three classical ensembles to Dirac Hamiltonians with a particle-hole symmetry. Examples for the remaining four can be found in certain normal-superconducting systems.

As suggested in the introduction to Guhr et al. (1998), the field of random matrices has opened up a new class of stochastic systems with a wide range of applications in physics and elsewhere and with intriguing mathematical connections. The unifying feature is that stochasticity combined with general symmetries leads to universal laws not based on dynamical principles. Indications are that it can also play a role in the analysis of financial data, in wireless communications, or in extracting correlations in neural signals (see, e.g., Forrester et al., 2003).

BRUNO ECKHARDT

See also **Free probability theory; Periodic orbit theory; Quantum chaos**

Further Reading

Beenakker, C.W.J. 1997. Random-matrix theory of quantum transport. *Reviews of Modern Physics*, 69: 713–808

Berry, M.V. 1985. Semiclassical theory of spectral rigidity. *Proceedings Royal Society of London* A, 400: 229–251

Berry, M.V. & Keating, J.P. 2000. The Riemann zeroes and eigenvalue asymptotics. *SIAM Review*, 41: 236–266

Bogomolny, E.B., Georgeot, B., Giannoni, M.J. & Schmit, C. 1997. Arithmetical chaos. *Physics Reports*, 291: 219–324

Brody, T.A., Floris, T., French, J.B., Mello, P.A., Pandey, A. & Wong, S.S.M. 1981. Random-matrix physics: spectrum and strength fluctuations. *Reviews of Modern Physics*, 53: 385–479

Efetov, K.B. 1997. *Supersymmetry in Disorder and Chaos*, Cambridge and New York: Cambridge University Press

Forrester, P.J., Snaith, N.C. & Verbaarschot, J.J.M. (editors). 2003. Special issue: random matrix theory. *Journal of Physics* A, 36: R1–R10 and 2859–3645

Guhr, T., Müller-Groeling, A. & Weidenmüller, H.A. 1998. Random-matrix theories in quantum physics: common concepts. *Physics Reports*, 299: 190–425

Haake, F. 2001. *Quantum Signatures of Chaos*, 2nd edition, Berlin and New York: Springer

Mehta, M.L. 1967. *Random Matrices and the Statistical Theory of Energy Levels*, New York: Academic Press

Porter, C.E. (editor). 1965. *Statistical Theory of Spectra*, New York: Academic Press

Stöckmann, H.J. 1999. *Quantum Chaos: An Introduction*, Cambridge and New York: Cambridge University Press

Wigner, E.P. 1967. Random matrices in physics. *SIAM Review*, 9: 1–23

Zirnbauer, M.R. 1996. Riemannian symmetric superspaces and their origin in random-matrix theory. *Journal of Mathematical Physics*, 37: 4986–5018

RANDOM MATRIX THEORY II: ALGEBRAIC DEVELOPMENTS

It was hypothesized by Eugene Wigner in the 1950s that the highly excited states of complex nuclei would have the same statistical properties as the eigenvalues of a large random real symmetric matrix. In pure mathematics, one finds a random matrix hypothesis in the theory of the celebrated Riemann hypothesis. Thus the Montgomery–Odlyzko law states that the statistics of the large zeros of the Riemann zeta function on the critical line (Riemann zeros) coincide with the statistics of the eigenvalues of a large complex Hermitian matrix. To test such hypotheses (for definiteness, the Montgomery–Odlyzko law), one computes a large sequence of consecutive Riemann zeros, scales the sequence so that locally the mean spacing is unity, and then empirically computes statistical quantities. A typical example of the latter is the distribution of the spacing between consecutive zeros. This must be compared against the same statistical quantity for the eigenvalues of large random complex Hermitian matrices. How then does one compute the eigenvalue spacing distribution for random matrices?

Hermitian random matrices with real, complex, and quaternion real Gaussian elements form matrix ensembles referred to as the Gaussian orthogonal ensemble (GOE) ($\beta = 1$), the Gaussian unitary ensemble (GUE) ($\beta = 2$), and the Gaussian symplectic ensemble (GSE) ($\beta = 4$), respectively, where β is a convenient label. Consider the bulk eigenvalues of such large matrices, scaled to have unit density. Denote by $p_\beta(k; s)$ the probability density that there are exactly k eigenvalues in between two eigenvalues of spacing s, and denote by $E_\beta(k; s)$ the probability that there are exactly k eigenvalues in an interval of size s. Define the generating functions $p_\beta(s; \xi) = \sum_{k=0}^{\infty}(1-\xi)^k p_\beta(k; s)$ and $E_\beta(s; \xi) = \sum_{k=0}^{\infty}(1-\xi)^k E_\beta(k; s)$. Note from the definitions that these quantities are related by $p_\beta(s; \xi) = \frac{1}{\xi^2}\frac{\mathrm{d}^2}{\mathrm{d}s^2} E_\beta(s; \xi)$.

Gaudin observed that the determinantal form of the correlations in the case $\beta = 2$ allows $E_2(s; \xi)$ to be written as a Fredholm determinant,

$$\begin{aligned} E_2(s; \xi) &= \det(1 - \xi K_2) \\ &= \det(1 - \xi K_2^{+}) \det(1 - \xi K_2^{-}), \qquad (1) \end{aligned}$$

where K_2, $K_2^{\pm}$ are integral operators on $(0, s)$ with kernels

$$\frac{\sin \pi(x - y)}{\pi(x - y)},$$

and
$$\frac{1}{2}\left[\frac{\sin\pi(x-y)}{\pi(x-y)} \pm \frac{\sin\pi(x+y)}{\pi(x+y)}\right],$$
respectively. The second equality is noted in (1) because both factors therein are related to E_1. Thus, with $E_1(-1;s)=0$, define

$$E_1^{\pm}(s;\xi) = \sum_{n=0}^{\infty}(1-\xi)^n\Big(E_1(2n;s)+E_1(2n\mp 1;s)\Big).$$

Then an inter-relationship between large GOE and large GUE matrices due to Dyson implies

$$E_2(s;\xi) = E_1^{+}(s;\xi)E_1^{-}(s;\xi).$$

This factorization turns out to be the same as in (1), so one obtains Mehta's result

$$E_1^{\pm}(s;\xi) = \det(1-\xi K_2^{\pm}).$$

For the case $\beta=4$, one uses Mehta's and Dyson's inter-relationship between large GSE matrices and large GOE matrices to conclude

$$E_4(n;s) = \\ E_1(2n;2s)+\tfrac{1}{2}\Big(E_1(2n-1;2s)+E_1(2n+1;2s)\Big).$$

A new line of study of E_β was initiated by Jimbo, Miwa, Môri, and Sato in 1980, which related the Fredholm determinant in (1) to integrable systems theory, resulting in the formula

$$E_2(s;\xi) = \exp\int_0^{\pi s}\sigma(t;\xi)\,\frac{\mathrm{d}t}{t},$$

where σ satisfies a particular example of the σ-form of the Painlevé V equation,

$$(s\sigma'')^2+4(s\sigma'-\sigma)\Big(s\sigma'-\sigma+(\sigma')^2\Big)=0,$$
$$\sigma(s;\xi)\underset{s\to 0}{\sim} -\frac{\xi s}{\pi}-\Big(\frac{\xi s}{\pi}\Big)^2.$$

The quantities $E_1^{\pm}$ can also be expressed in terms of Painlevé transcendents. Thus, combining results of the present author with results of Tracy and Widom, one has

$$E_1^{\pm}(s;\xi) = \exp\int_0^{(\pi s/2)^2} v(t;\xi;a)\,\frac{\mathrm{d}t}{t}\Big|_{a=\pm 1/2},$$

where $v(t;\xi;a)$ satisfies a particular example of the σ-form of the Painlevé III$'$ equation

$$(tv'')^2-a^2(v')^2-v'(4v'+1)(v-tv')=0,$$
$$v(t;\,;\xi;a)\underset{t\to 0^+}{\sim} -\frac{\xi t^{a+1}}{2^{2a+2}\Gamma(a+1)\Gamma(a+2)}.$$

Another bulk spacing distribution with a Painlevé type evaluation is the nearest-neighbor spacing between eigenvalues (i.e., the minimum of the distances to the left neighbor and the right neighbor) for large GUE matrices, $nn(t)$ say. Forrester and Odlyzko (1996) have shown that

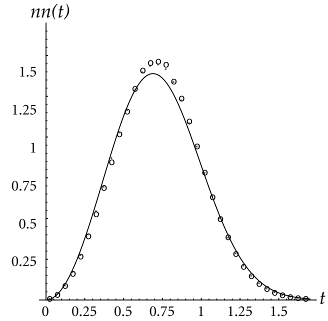

Figure 1. Comparison of $nn(t)$ for the GUE (continuous curve) and for 10^6 consecutive Riemann zeros, starting near zero number 1 (open circles), 10^6 (asterisks), and 10^{20} (filled circles).

$$nn(t) = -\frac{y(2\pi t;a)}{t}\exp\int_0^{\pi t}\frac{y(2s;a)}{s}\,\mathrm{d}s\Big|_{a=1},$$

where $y(s;a)$ satisfies the differential equation

$$(sy'')^2+4(-a^2+sy'-y)\\ \Big((y')^2-\{a-(a^2-sy'+y)^{1/2}\}^2\Big)=0$$

subject to the boundary condition

$$y(s;a)\underset{s\to 0^+}{\sim} -\frac{2(s/4)^{2a+1}}{\Gamma(1/2+a)\Gamma(3/2+a)}.$$

Comparison of a plot of this statistic, obtained from the formula in Forrester and Odlyzko (1996), against the same statistic computed, empirically for large sequences of Riemann zeros starting at three different positions along the critical line is given in Figure 1.

P.J. FORRESTER

See also **Random matrix theory I, III, IV**

Further Reading

Forrester, P.J. Eigenvalue probabilities and Painlevé theory, Chapter 6 of *Log-gases and random matrices*, www.ms.unimelb.edu.au/~matpjf/matpjf.html

Forrester, P.J. & Odlyzko, A.M. 1996. Gaussian unitary ensemble eigenvalues and Riemann ζ function zeros: a non-linear equation for a new statistic, *Physical Review* E, 54: R4493–R4495

Tracy, C.A. & Widom, H. 1993. Introduction to random matrices. In *Geometric and Quantum Aspects of Integrable Systems*, edited by G.F. Helminck, New York: Springer, pp. 407–424

van Moerbeke, P. 2001. Integrable lattices: random matrices and random permutations. In *Random Matrices and Their Applications*, edited by P. Bleher & A. Its, Cambridge and New York: Cambridge University Press, pp. 321–406

RANDOM MATRIX THEORY III: COMBINATORICS

As the two previous entries show, random matrix theory (RMT) is a field of research devoted to the statistical analysis of the eigenvalues of matrices selected at random from a given set of matrices (Mehta, 1991). Recently there have been surprising discoveries that connect RMT to a branch of mathematics called combinatorics.

One fundamental topic within combinatorics is the study of permutations. A permutation is a rearrangement of the elements of a set. The set is usually finite, in which case it may be taken to be the first n positive integers, and a permutation may also be represented as a one-to-one invertible mapping from $\{1, 2, \ldots, n\}$ into itself. For each positive integer n, the number of permutations of length n is $n!$. If permutations of length n are selected at random, and each permutation is equally likely, then the probability of selecting a particular permutation is $1/n!$. In thinking about permutations probabilistically, researchers have discovered new connections to random matrix theory.

Random Matrix Theory: Gaussian Unitary Ensemble

The Gaussian unitary ensemble (GUE) is a fundamental example of a RMT. The collection of matrices is all Hermitian (self-adjoint, complex) matrices of size n. The diagonal matrix entries M_{jj}, $j = 1, 2, \ldots, n$, are independent normal random variables; that is, each has probability measure $\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-m^2/2}\,\mathrm{d}m$. The off-diagonal matrix entries $M_{jk} = M_{jk}^{(\mathrm{R})} + \mathrm{i}M_{jk}^{(\mathrm{I})}$, $1 \le j \ne k \le n$, are complex, and $\left\{M_{jk}^{(\mathrm{R})}, M_{jk}^{(\mathrm{I})}\right\}_{1 \le j < k \le n}$ is a collection of independent identically distributed Gaussian random variables with probability measure $\frac{1}{\sqrt{\pi}}\mathrm{e}^{-m^2}\,\mathrm{d}m$. The matrix entries below the diagonal are determined by the entries above the diagonal because the matrices satisfy the self-adjointness criterion $M^* = M$, where $(M^*)_{jk} = \overline{M_{jk}}$ and $\overline{(\cdot)}$ represents the complex conjugate of $(\cdot)$. Thus, the GUE of random matrices is merely a probability measure defined on the standard Euclidean space of dimension n^2. In standard probability terminology, the probability of finding a matrix in an "infinitesimal volume element" of matrices centered at a given matrix $M = (M_{jk})_{1 \le j,k \le n}$ is

$$\frac{1}{2^n \pi^{n^2}} \mathrm{e}^{-\frac{1}{2}\sum_{j=1}^{n} M_{jj}^2 - \sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(M_{jk}^{(\mathrm{R})})^2 + (M_{jk}^{(\mathrm{I})})^2} \times \prod_{1}^{n} \mathrm{d}M_{jj} \prod_{1 \le j < k \le n} \mathrm{d}M_{jk}^{\mathrm{R}}\, \mathrm{d}M_{jk}^{\mathrm{I}}.$$

This is often more conveniently written in the form

$$\frac{1}{Z_n}\mathrm{e}^{-\frac{1}{2}\,\mathrm{Tr}(M^2)}\,\mathrm{d}M, \tag{1}$$

where $Z_n = 2^n \pi^{n^2}$, $\mathrm{d}M = \prod \mathrm{d}M_{jj} \prod \mathrm{d}M_{jk}^{(I)}\,\mathrm{d}M_{jk}^{(\mathrm{R})}$, and $\mathrm{Tr}(A)$ represents the trace of the matrix A, $\mathrm{Tr}(A) = \sum_{j=1}^{n} A_{jj}$.

This is just one example of an RMT. For example, one may replace $\mathrm{Tr}(M^2)$ in (1) by $\mathrm{Tr}(M^r)$, or $\mathrm{Tr}(V(M))$ for any reasonable function $V : \mathbb{R} \to \mathbb{R}$.

The eigenvalues $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ of a matrix M selected at random according to (1) are fundamental random variables, and quite a lot is known about their statistical properties. One example is the largest eigenvalue λ_N. Because of many different applications, their statistical behavior is particularly interesting when n goes to ∞. This random variable's behavior when $n \to \infty$ is as follows:

$$\lim_{n\to\infty} \mathrm{Prob}\left(\lambda_n < \sqrt{2n} + \frac{s}{\sqrt{2}n^{1/6}}\right) = F(s),$$

where

$$F(s) = \exp\left(-\int_s^\infty (x - s)^2 q(x)\,\mathrm{d}x\right) \tag{2}$$

and q is the unique solution to the Painlevé II equation

$$q'' = sq + 2q^3$$

satisfying the condition

$$q(s) \sim \mathrm{Ai}(s) \quad \text{as } s \to \infty.$$

This result is due to Craig Tracy and Harold Widom (1994), and the function F is often referred to as the Tracy–Widom distribution. Surprisingly, the Painlevé equation appears in a fundamental way in random matrix theory. It also appears, in an equally fundamental way, in combinatorics.

Increasing Subsequences of Permutations

A permutation, or an invertible mapping from the set $\{1, 2, \ldots, n\}$ to itself may be represented as a sequence of numbers. For example, the sequence 3,1,2 represents the permutation $\sigma(1) = 3$, $\sigma(2) = 1$, $\sigma(3) = 2$. An increasing subsequence of a permutation is obtained by representing a permutation as a sequence, and then selecting a subsequence in which the elements are increasing. For example, the sequence 1, 2 is an increasing subsequence of the permutation σ considered above.

Each permutation has many increasing subsequences. In combinatorics the length of the largest increasing subsequence of a permutation is an important object of study. For the permutation σ considered above, that length is 2. As another example, if π is the permutation 1,3,2,4,5, that is, $\pi(1)=1$, $\pi(2)=3$, $\pi(3)=2$, $\pi(4)=4$, $\pi(5)=5$, the length of π's longest increasing subsequence is 4. If σ is a permutation of length n, then $\ell_n(\sigma)$ denotes the length of σ's longest increasing subsequence.

If permutations of length n are selected at random, and each permutation has equal probability $1/n!$, then $\ell_n(\sigma)$ is a natural random variable arising from combinatorics. Its statistical behavior for n large has been an object of study since at least the 1960s. Interest in ℓ_n stems from basic connections between permutations and games of chance, such as "solitaire," also known as "patience sorting." In their mathematical review, Aldous & Diaconis (1999) give a precise description of such a game:

> Take a deck of cards labeled $1, 2, 3, \cdots, n$. The deck is shuffled, cards are turned up one at a time and dealt into piles on the table, according to the rule
>
> - A low card may be placed on a higher card (e.g., 2 may be placed on 7) or may be put into a new pile to the right of the existing piles.
>
> At each stage we see the top card on each pile. If the turned up card is higher than the cards showing then it *must* be put into a new pile to the right of the others. *The object of the game is to finish with as few piles as possible.* (Aldous & Diaconis, 1999, p. 413)

Suppose our deck has 5 cards, numbered 1–5, and that they are shuffled so that they are ordered as 5 3 1 4 2. During the game, the piles might evolve as follows:

		1	1	1
	3	3	3	3 2
5	5	5	5 4	5 4

and at the end of the game, we have two piles. Another (non-optimal) possibility is

			4	4 2
5	5 3	5 3 1	5 3 1	5 3 1

and we wind up with 3 piles. There is a simple algorithm (referred to as the "greedy algorithm") which always yields the minimum number of piles (see Aldous & Diaconis, 1999). The connection between the "solitaire" game described above and longest increasing subsequences is this: for any given shuffle, the minimum number of piles is exactly the longest increasing subsequence of the shuffle (when the shuffle is viewed as a permutation).

Over the past 40 years, mathematicians have studied the statistical behavior of the random variable $\ell_n(\sigma)$ when n, the size of the permutations, tends to infinity. Let $\mathbb{E}(\ell_n)$ denote the mean (or average) value of $\ell_n(\sigma)$, and let $\mathrm{Var}(\ell_n)$ denote its variance, $\mathrm{Var}(\ell_n) = \mathbb{E}\left(\ell_n^2\right) - (\mathbb{E}(\ell_n))^2$. Some natural questions are: How does $\mathbb{E}(\ell_n)$ grow with n? How fast does the variance grow? Is there a notion of a central limit theorem, or is there a limiting distribution for ℓ_n that emerges when n tends to ∞? Here are some results, both old and new, which answer these questions, and also demonstrate a striking connection between combinatorial questions and random matrix theory.

Result 1 (Logan & Shepp, 1977, see also Diaconis, 1995, Vershik & Kerov, 1997):

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \mathbb{E}(\ell_n) = 2, \tag{3}$$

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \mathrm{Var}(\ell_n) = 2. \tag{4}$$

Result 2 (Baik et al., 1999):

$$\lim_{n\to\infty} \mathrm{Prob}\left[\frac{\ell_n(\sigma) - 2n^{1/2}}{n^{1/6}} < x\right] = F(x), \tag{5}$$

where F is the Tracy–Widom distribution function (2). Result 2 was established in 1999. One remarkable aspect of this result is that it provides a mathematically precise connection between random permutations and random matrix theory (and Painlevé equations). There are now several proofs of this result as well as generalizations (see, e.g., Borodin et al. (2002)), and each proof contains different interconnections between areas of mathematics.

Further Developments

There have been a number of recent results that connect combinatorics to random matrix theory. Interesting generalizations of longest increasing subsequence problems have appeared. For example, Baik & Rains (2001) considered only those permutations that are their own inverses (also called involutions) and established results analogous to those mentioned above. Further generalizations to the study of "random words" have appeared (Tracy & Widom, 2001).

The Tracy–Widom distribution F has appeared in a variety of other contexts as well. For example, in the study of random domino tilings of the so-called Aztec diamond, the same distribution function appears in a natural scaling limit (Johansson, 2001). There is a notion of a "frozen region" of a domino tiling, and the boundary of this frozen region may be interpreted as a random variable, which is described, asymptotically, by the Tracy–Widom distribution function. This phenomenon is somewhat universal within random tiling problems as well—it has recently been established (Baik et al., 2004) that the same behavior appears in random rhombi tilings of hexagonal domains.

KENNETH MCLAUGHLIN

See also **Random matrix theory I, II and IV**

Further Reading

Aldous, D.J. & Diaconis, P. 1995. Hammersley's interacting particle process and longest increasing subsequences. *Probability Theory and Related Fields*, 103: 119–213

Aldous, D. & Diaconis, P. 1999. Longest increasing subsequences: from patience sorting to the Baik–Deift–Johansson theorem. *Bulletin of the American Mathematical Society (N.S.)*, 36(4): 413–432

Baik, J., Deift, P. & Johansson, K. 1999. On the distribution of the length of the longest increasing subsequence of random permutations. *Journal of the American Mathematical Society*, 12(4): 1119–1178

Baik, J., Kriecherbauer, McLaughlin, K.D.T.R. & Miller, P. 2004. Uniform asymptotics for polynomials orthogonal with respect to a general class of discrete weights and universality results for associated ensembles. *Annals of Mathematics*, submitted; Preprint CA/0310278

Baik, J. & Rains, E. 2001. The asymptotics of monotone subsequences of involutions. *Duke Mathematical Journal*, 109(2): 205–281

Borodin, A., Okounkov, A. & Olshanski, G. 2000. Asymptotics of Plancherel measures for symmetric groups. *Journal of the American Mathematical Society*, 13(3): 481–515

Johansson, K. 2001. Discrete orthogonal polynomial ensembles and the Plancherel measure. *Annals of Mathematics* (2), 153(1): 259–296

Logan, B.F. & Shepp, L.A. 1977. A variational problem for random Young tableaux. *Advances in Mathematics*, 26: 206–222

Mehta, M.L. 1991. *Random Matrices*, 2nd edition, San Diego: Academic Press

Tracy, C.A. & Widom, H. 1994. Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159: 151–174

Tracy, C.A. & Widom, H. 2001. On the distributions of the lengths of the longest monotone subsequences in random words. *Probability Theory and Related Fields*, 119: 350–380

Vershik, A.M. & Kerov, S.V. 1977. Asumptotics of the Plancherel measure of the symmetric group and the limiting form of Young tables. *Soviet Mathematics, Doklady*, 18: 527–531. Translation of *Doklady Academii Nauk, SSSR*, 233(1977): 1024–1027

RANDOM MATRIX THEORY IV: ANALYTIC METHODS

In this article, we address various analytical issues that arise in random matrix theory (RMT) using techniques from the theory of Riemann–Hilbert problems (RHPs). We refer the reader to the previous three entries for more information on the physical applications and various algebraic developments in RMT, and to the entry on Riemann–Hilbert problem for more information on RHPs. The classic text on RMT is Mehta (1991); see also Deift (2000) for a pedagogic presentation of some recent analytical results.

We discuss only the so-called unitary ensembles (UEs), that is, ensembles of $N \times N$ Hermitian matrices $M = M^* = (M_{jk} = M^{\mathrm{R}}_{jk} + \mathrm{i}M^{\mathrm{I}}_{jk})$ with probability distributions

$$P_N(M)\,\mathrm{d}M = \frac{1}{Z_N}\mathrm{e}^{-\operatorname{tr} V(M)}\,\mathrm{d}M$$
$$= \frac{1}{Z_N}\mathrm{e}^{-\operatorname{tr} V(M)}\prod_{j=1}^{N}\mathrm{d}M_{jj}\prod_{j<k}\mathrm{d}M^{\mathrm{R}}_{jk}\prod_{j<k}\mathrm{d}M^{I}_{jk}, \tag{1}$$

where $V : \mathbb{B} \to \mathbb{R}$ grows sufficiently rapidly at ∞ and Z_N is a normalization constant. The name UE refers to the fact that the distribution is invariant under unitary conjugation $M \to \tilde{M} = UMU^*$, with U unitary, that is, $P(\tilde{M})\,\mathrm{d}\tilde{M} = P(M)\,\mathrm{d}M$. For information on other distributions such as orthogonal ensembles, see Tracy & Widom (1996, 1998) for some additional analytical developments. The special case in which $V(x) = x^2$ is called the Gaussian unitary ensemble (GUE), with $P_N(M)\,\mathrm{d}M = \frac{1}{Z_N}\mathrm{e}^{-\operatorname{tr} M^2}\,\mathrm{d}M$.

Under (1) the eigenvalues $\lambda_1(M) \geq \lambda_2(M) \geq \ldots \geq \lambda_N(M)$ of a matrix M in the ensemble become random variables with distribution given by the Weyl integration formula (see Mehta, 1991)

$$P_N(\lambda)\,\mathrm{d}^N\lambda = \frac{1}{Z'_N}\mathrm{e}^{-\sum_i V(\lambda_i)}\prod_{i<j}(\lambda_i - \lambda_j)^2\chi\,\mathrm{d}^N\lambda, \tag{2}$$

where χ denotes the indicator function on $\mathbb{R}^N$ for the set $\lambda_1 \geq \ldots \geq \lambda_N$. A statistic of basic importance is the *gap probability*

$$P_N(\theta;\,E) = \text{Prob}\ \{\ M \text{ has no eigenvalues in } (E-\theta,\,E+\theta)\} \tag{3}$$

for $E \in \mathbb{R}$, $\theta > 0$. Scaling the eigenvalues in the standard way around the energy E, $\lambda \mapsto \lambda' = \gamma_N(\lambda - E)$, so that the expected number of eigenvalues per unit λ'-interval is 1, one considers the limit

$$P(y;\,E) = \lim_{N\to\infty} P_N(y/\gamma_N;\,E). \tag{4}$$

In the case of GUE, a calculation of Gaudin and Mehta (Mehta, 1991) using classical estimates showed that

$$\gamma_N \sim N^{1/2} \tag{5}$$

and that

$$P(y;\,E=0) = \det(1 - S_y), \tag{6}$$

where S_y denotes the trace class operator with kernel

$$S_y(\xi,\eta) = \frac{\sin\pi(\xi-\eta)}{\pi(\xi-\eta)} \tag{7}$$

acting on $L^2(-y,y)$, the space of square integrable functions on the interval $(-y,y)$. The *universality conjecture* for the gap probability is the claim that under the correct scaling, (6) is true with the same right-hand side for all suitable V and E. The only thing that depends on V is the scaling parameter γ_N. Universality conjectures for other statistical quantities are discussed briefly below.

In this article we present a sketch of how to prove the universality conjecture for (6) following Deift et al. (1999a,b) in which the authors consider two kinds of potentials V

(i) $V(x) = t_{2m}x^{2m} + \ldots + t_0,\ t_{2m} > 0$,
(ii) $V(x) = NQ(x)$, where

(a) $Q(x)$ is real analytic in a neighborhood of $\mathbb{R}$

(b) $|Q(x)|/|\log x| \to +\infty$ as $|x| \to \infty$.

In order to prove (6), one must choose E in the support of the so-called equilibrium measure for V (see Step 4 Analytical considerations (ii) below).

The universality conjecture was first considered in the physics literature in Brézin & Zee (1993), in the mathematical literature in Pastur and Scherbina (1997), and in the special case $V(x) = x^4 + tx^2$, $t < 0$, in Bleher and Its (1999). The method in Bleher & Its has common elements with Deift et al. (1999a,b), but in addition Bleher & Its uses certain isomonodromy ideas.

Step 1: Representation in terms of orthogonal polynomials

Using an earlier calculation of Gaudin and (2), Gaudin and Mehta (Mehta, 1991) showed that

$$P_N(\theta, E = 0) = \det(1 - K_N), \tag{8}$$

where K_N is the finite rank operator acting on $L^2(E - \theta, E + \theta)$ with kernel

$$K_N(x, y) = \sum_{j=0}^{N-1} \phi_j(x)\phi_j(y), \tag{9}$$

where $\phi_j(x) = \mathrm{e}^{-(1/2)V(x)} p_j(x)$ and $p_j(x) = k_j x^j + \cdots$, $k_j > 0$, $j \geq 0$, are the orthonormal polynomials generated by the measure $\mathrm{e}^{-V(x)}\,\mathrm{d}x$,

$$\int_{\mathbb{R}} p_j(x) p_l(x) \mathrm{e}^{-V(x)}\,\mathrm{d}x = \delta_{jl}. \tag{10}$$

By the Christoffel–Darboux formula (Szegö, 1975), K_N can be expressed in terms of p_N and p_{N-1} and so the proof of (6) reduces to the evaluation of the asymptotics (so-called *Plancherel–Rotach* asymptotics) for $p_N(E + x/\gamma_N)$, $p_{N-1}(E + x/\gamma_N)$ as $N \to \infty$, for polynomials orthonormal with respect to a weight as in (10). For GUE, the p_N's are Hermite polynomials and such asymptotics can be found in the classical literature (Plancherel & Rotach, 1929). This is why Gaudin and Mehta were able to derive (6) for GUE, but for general weights $\mathrm{e}^{-V(x)}\,\mathrm{d}x$ the derivation of Plancherel–Rotach asymptotics is crucial.

Step 2: Introduction of Riemann–Hilbert Problem (RHP)

Write $p_j(x) = k_j \pi_j(x)$, so $\pi_j(x) = x^j + \ldots$. Following Fokas et al. (1991), the orthonormal polynomials p_j can be recovered by solving an RHP. Let the z-plane contour $\Sigma = \mathbb{R}$ be oriented from left to right so that the $(+)$ side (respectively, $(-)$ side) lies to the left (respectively, right) as one traverses Σ in the direction of the orientation. Define the jump matrix

$$v(z) = \begin{pmatrix} 1 & \mathrm{e}^{-V(z)} \\ 0 & 1 \end{pmatrix} \text{ for } z \in \Sigma,$$

and for any nonnegative integer q, let $Y = Y^{(q)} = (Y_{ij})_{1 \leq i,j \leq 2}$ be the (unique) 2×2 solution of the RHP (Σ, v).

(i) $Y(z)$ is analytic in $\mathbb{C} \setminus \Sigma$,

(ii) $Y_+(z) = Y_-(z)v(z)$, $z \in \mathbb{R}$, where $Y_\pm$ denotes the plus/minus boundary values of $Y(z)$, normalized so that

(iii) $Y(z)\begin{pmatrix} z^{-q} & 0 \\ 0 & z^q \end{pmatrix} \to \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ as $z \to \infty$.

Then

$$\pi_q(z) = Y_{11}(z),\ \pi_{q-1}(z) = -Y_{21}(z)/2\pi\mathrm{i}k_{q-1}^2 \tag{11}$$

and

$$k_q = \sqrt{-(Y_{21})_1/2\pi\mathrm{i}}, \tag{12}$$

where $Y_{21}(z) = (Y_{21})_1 z^{q-1} \ldots$.

Thus, the evaluation of Plancherel–Rotach asymptotics for p_q, p_{q-1} reduces to the asymptotics of the RHP (Σ, v) as $q = N \to \infty$.

Step 3: Steepest-descent method applied to the RHP

We use the steepest-descent method for the RHPs introduced in Deift & Zhou (1993) and extended in Deift et al. (1997) to include fully nonlinear oscillations (*See* **Riemann–Hilbert problem**). We consider the case where $V(x) = NQ(x)$ with properties (a) and (b) on the left above and as in Deift et al. (1999a). The case where $V(x) = t_{2m}x^{2m} + \ldots + t_0$ can be reduced to this case with lower-order errors by scaling (see Deift et al., 1999b).

Suppose there exists a union of $n + 1$ disjoint intervals $J = \cup_{i=0}^{n} (b_i, a_{i+1})$ in $\mathbb{R}$, $b_0 < a_1 < \ldots < b_n < a_{n+1}$, a real number l, and a function $g(z)$ with the following properties:

$$g(z) \text{ is analytic in } \mathbb{C} \setminus \mathbb{R}, \tag{13}$$

$$g(z) = \log z + \mathrm{o}(1) \text{ as } z \to \infty, \tag{14}$$

$$g_+(z) + g_-(z) - V(z) - l \leq 0 \text{ for } z \in \mathbb{R} \setminus \bar{J}, \tag{15}$$

where $\bar{J}$ denotes the closure of J, and where equality holds for at most a finite number of points,

$$g_+(z) - g_-(z) \in \mathrm{i}\mathbb{R}, \tag{16}$$

for all real z and is constant in each component of $\mathbb{R} \setminus \bar{J}$,

$$g_+(z) + g_-(z) - V(z) - l = 0 \text{ for } z \in \bar{J}\,, \quad (17)$$

$$\mathrm{i}(g'_+ - g'_-) \geq 0 \text{ for } z \in J, \quad (18)$$

where equality holds for at most a finite number of points. As before, $g_\pm$ denotes the boundary values of $g(z)$.

Now define

$$M(z) \equiv \mathrm{e}^{-Nl/2\sigma_3} Y(z) \mathrm{e}^{-(g(z)-l/2)\sigma_3},$$
$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (19)$$

A simple calculation using (13) and (14) shows that $M(z)$ solves a RHP normalized at infinity.

(i) $M(z)$ is analytic in $\mathbb{C} \setminus \mathbb{R}$
(ii) $M_+(z) = M_-(z) v_M(z)$, where $v_M(z) =$

$$\begin{pmatrix} \mathrm{e}^{-N(g_+(z)\,-\,g_-(z))} & \mathrm{e}^{N(g_+(z)\,+\,g_-(z)\,-\,V(z)\,-\,l)} \\ 0 & \mathrm{e}^{N(g_+(z)\,-\,g_-(z))} \end{pmatrix},$$
$z \in \mathbb{R}$,
(iii) $M(z) \to I$ as $z \to \infty$.

As we now show, the properties of $g(z)$ are chosen carefully so that the jump matrix v_M takes a specific form. By (17)

$$v_M(z) = \begin{pmatrix} \mathrm{e}^{-N(g_+(z)-g_-(z))} & 1 \\ 0 & \mathrm{e}^{N(g_+(z)-g_-(z))} \end{pmatrix},$$
$$z \in J, \quad (20)$$

and by (16)

$$v_M(z) = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}N\Omega_j} & \mathrm{e}^{N(g_+(z)+g_-(z)-V(z)-l)} \\ 0 & \mathrm{e}^{\mathrm{i}N\Omega_j} \end{pmatrix},$$
$$z \in (a_j, b_j)\,,\ j = 1, \ldots, n, \quad (21)$$

for some real constants Ω_j. Furthermore, if the inequality in (15) is strict then

$$v_M(z) = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}N\Omega_j} & o(1) \\ 0 & \mathrm{e}^{\mathrm{i}N\Omega_j} \end{pmatrix}, z \in (a_j, b_j)$$
$$\text{as } N \to \infty. \quad (22)$$

By (17), for $z \in J$,

$$G \equiv g_+ - g_- = 2g_+ - V(z) - l = -2g_- + V(z) + l$$

and so has an analytic continuation to a suitable neighborhood U of J. If the inequality in (18) is strict, it follows by the Cauchy–Riemann conditions (make U smaller if necessary) that $\operatorname{Re} G(z) \gtrless 0$ for $z \in \{\operatorname{Im} z \gtrless 0\} \cap U$, respectively. Extend $\Sigma = \mathbb{R}$ to a new contour $\tilde{\Sigma}$ consisting of $\mathbb{R}$ together with the boundary of $n + 1$ lens-shaped regions, one region surrounding each interval (b_{j-1}, a_j) in J, as indicated in Figure 1.

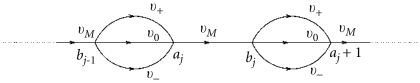

Figure 1. The Riemann–Hilbert problem $(\tilde{\Sigma}, \tilde{v})$.

Note the factorization for v_M on J

$$\begin{pmatrix} \mathrm{e}^{-NG} & 1 \\ 0 & \mathrm{e}^{NG} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \mathrm{e}^{NG} & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 0 \\ \mathrm{e}^{-NG} & 1 \end{pmatrix} \equiv v_- v_0 v_+. \quad (23)$$

The final algebraic step in the procedure is to define

(i) $\tilde{M}(z) \equiv M(z)$ outside the lens-shaped regions,
(ii) $\tilde{M}(z) \equiv M(z) \begin{pmatrix} 1 & 0 \\ -\mathrm{e}^{-NG} & 1 \end{pmatrix}$ in the upper lens regions,
(iii) $\tilde{M}(z) \equiv M(z) \begin{pmatrix} 1 & 0 \\ \mathrm{e}^{NG} & 1 \end{pmatrix}$ in the lower lens regions.

Then $\tilde{M}$ solves the normalized RHP $(\tilde{\Sigma}, \tilde{v})$

(i) $\tilde{M}(z)$ is analytic in $\mathbb{C} \setminus \tilde{\Sigma}$,
(ii) $\tilde{M}_+(z) = \tilde{M}_-(z)\tilde{v}(z),\ z \in \tilde{\Sigma}$,
(iii) $\tilde{M}(z) \to I$ as $z \to \infty$,

where

(i) $\tilde{v}(z) = v_M(z)$ for $z \in \mathbb{R} \setminus \bar{J}$,
(ii) $\tilde{v}(z) = v_\pm(z)$ for $z \in \tilde{\Sigma} \cap \mathbb{C}_\pm$,
(iii) $\tilde{v}(z) = v_0(z)$ for $z \in J$

as indicated in Figure 1.

Now the situation is clear. The above computations show that as $N \to \infty$, $\tilde{v}(z) \to v^\infty(z)$, where

(i) $v^\infty(z) = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}N\Omega_j} & 0 \\ 0 & \mathrm{e}^{\mathrm{i}N\Omega_j} \end{pmatrix}$ for $z \in (a_j, b_j)$, $j = 1, \ldots, n$,
(ii) $v^\infty(z) = I$ for $z \in (-\infty, b_0) \cup (a_0, \infty)$,
(iii) $v^\infty(z) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ for $z \in J$,
(iv) $v^\infty(z) = I$ for z on the boundary of any of the lens-shaped regions.

Thus, we expect that as $N \to \infty$, $\tilde{M}(z) \to M^\infty(z)$, where $M^\infty(z)$ solves the RHP $(\tilde{\Sigma}, v^\infty)$

(i) $M^\infty(z)$ is analytic in $\mathbb{C} \setminus \tilde{\Sigma}$,
(ii) $M_+^\infty(z) = M_-^\infty(z) v^\infty(z).\ z \in \tilde{\Sigma}$,
(iii) $M^\infty(z) \to I$ as $z \to \infty$.

It turns out that the RHP $(\tilde{\Sigma}, v^\infty)$ can be solved explicitly in terms of theta functions (Deift et al., 1999a), thus yielding asymptotic formulae for $p_N(E + x/\gamma_N)$, $p_{N-1}(E + x/\gamma_N)$ as $N \to \infty$, which then leads to the proof of the universality conjecture (6) as indicated above.

Step 4: Analytical considerations

In order to make the above schema rigorous, many analytical difficulties must be addressed:

(i) One must show that J, l, and $g(z)$ with properties (13)–(18) actually exist. This can be done by considering the minimization problem

$$E_V = \inf_{\mu\in\mathcal{M}} \left(\int\int \log|s-t|^{-1}\,\mathrm{d}\mu(s)\mathrm{d}\mu(t) + \int V(t)\,\mathrm{d}\mu(t) \right), \qquad (24)$$

where $\mathcal{M}$ is the space of probability measure on $\mathbb{R}$. It turns out (Saff & Totik, 1997) that the infinum is achieved at a unique measure $\mu = \mu_V$, the so-called *equilibrium measure* for V. For $V = NQ$ satisfying (i) and (ii) above, a result of Deift, Kriecherbauer and McLaughlin shows that $\mathrm{d}\mu_V(s) = \psi_V(s)\,\mathrm{d}s$ is absolutely continuous with respect to Lebesgue measure and is supported on a finite union of intervals J_V. Then quite remarkably

$$g(z) \equiv \int_{\mathbb{R}} \log(z-s)\,\mathrm{d}\mu_V(s) \qquad (25)$$

has the desired properties (13)–(18) on $J = J_V$, provided l is chosen to be the Lagrange multiplier associated with the constraint $\int \mathrm{d}\mu = 1$ in the minimization problem E_V.

(ii) The scaling parameter γ_N must be determined. In order to obtain the universality limit (6) one must consider only points E in the support of μ_V, $\psi_V(E) > 0$. For such E, one takes $\gamma_N \equiv N\psi_V(E)$. If $\psi_V(E) = 0$, the limit in (4) (if it exists!) is *different* from (6). In this connection, see (26) below.

(iii) The convergence of $\tilde{v}(z)$ to $v^\infty(z)$ is not uniform on $\tilde{\Sigma}$ as $N \to \infty$, leading in particular to difficulties at the end points of J.

(iv) The inequalities in (15) and (18) may not be strict.

The analytical heart of the problem lies in addressing the latter three points, and we refer the reader to Deift et al. (1999a) for further details.

The asymptotic behavior of $P(y, E = 0) = \det(1 - S_y)$ as $y \to \infty$ can itself be evaluated using Riemann–Hilbert techniques (see Deift et al., 1997). The work in Deift et al. (1997) follows on the earlier work of Widom (1994, 1995), who uses more classical techniques. The paper by Deift et al. (1997) is the first application of RHP techniques to random matrix theory.

Universality conjectures for other statistical quantities, such as the k-point correlation function and the nearest-neighbor spacing distribution, can also be proved using the above methods (see Deift et al., 1999a; Deift, 2000). For GUE, Tracy & Widom (1994) proved the beautiful result that if $\lambda_1(M)$ is the largest eigenvalue of a GUE matrix M, then

$$\lim_{N\to\infty} \mathrm{Prob}((\lambda_1(M) - \sqrt{2N})2^{1/2}N^{1/6} \le t) = F(t), \qquad (26)$$

where $F(t)$ is the so-called *Tracy–Widom distribution*

$$F(t) = \mathrm{e}^{-\int_t^\infty (s-t)u^2(s)\,\mathrm{d}s},$$

where $u(s)$ is the unique solution (Hastings & McLeod, 1980) of the Painlevé II equation $u'' = 2u^3 + su$ with asymptotics $u(s) \sim \mathrm{Ai}(s)$ as $s \to \infty$. Here Ai(s) is the classical Airy function. As long as $\psi_V(s)$ behaves "generically" near the right end point of J, $\psi_V(s) \sim \sqrt{s - a_{N+1}}$, $s \sim a_{N+1}$, one can use the methods of Deift et al. (1999a) to show that the behavior of $\lambda_1(M)$ is also universal as $N \to \infty$; that is, once the correct scaling parameters at a_{N+1} are inserted, the same limit (26) is true for all UEs. As noted in the entry on Riemann–Hilbert problem, the distribution $F(t)$ arises, in particular, in the solution of Ulam's problem in Baik et al. (1999).

Percy Deift and Xin Zhou

See also **Painlevé analysis; Random matrix theory I, II, and III; Riemann–Hilbert problem**

Further Reading

Baik, J., Deift, P. & Johansson, K. 1999. On the distribution of the length of the longest increasing subsequence of random permutations. *Journal of the American Mathematical Society*, 12(4): 1119–1178

Bleher, P. & Its, A.R. 1999. Semiclassical asymptotics of orthogonal polynomials, Riemann–Hilbert problems, and universality in the matrix model. *Annals of Mathematics*, 150(2): 185–266

Brézin, E. & Zee, A. 1993. Universality of the correlations between eigenvalues of large random matrices. *Nuclear Physics* B, 402(3): 613–627

Deift, P. 2000. *Orthogonal Polynomials and Random Matrices: A Riemann–Hilbert Approach*, Providence, RI: American Mathematical Society

Deift, P., Its, A. & Zhou, X. 1997. A Riemann–Hilbert approach to asymptotic problems arising in the theory of random matrix models, and also in the theory of integrable statistical mechanics. *Annals of Mathematics*, 145: 1–88

Deift, P., Kriecherbauer, T., McLaughlin, K., Venakides, S. & Zhou, X. 1999a. Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory. *Communications in Mathematical Physics*, 52: 1335–1425

Deift, P., Kriecherbauer, T., McLaughlin, K., Venakides, S. & Zhou, X. 1999b. Strong asymptotics of orthogonal polynomials with respect to exponential weights. *Communications in Mathematical Physics*, 52: 1491–1552

Deift, P., Venakides, S. & Zhou, X. 1997. New results in small dispersion KdV by and extension of the steepest descent method for Riemann–Hilbert problems, *Internat. Math. Res. Notices*, No. 6, 285–299

Deift, P. & Zhou, X. 1993. A steepest descent method for oscillatory Riemann–Hilbert problems–asymptotics for the MKdV equation. *Annals of Mathematics*, 137: 295–368

Fokas, A.S., Its, A.R. & Kitaev, A.V. 1991. Discrete Painlevé equations and their appearance in quantum gravity. *Communications in Mathematical Physics*, 142(2): 313–344

Hastings, S.P. & McLeod, J.B. 1980. A boundary value problem associated with the second Painlevé transcendent and the Korteweg de Vries equation. *Archive for Rational Mechanics and Analysis*, 73: 31–51

Mehta, M.L. 1991. *Random Matrices*, 2nd edition, San Diego: Academic Press

Pastur, L. & Scherbina, M. 1997. Universality of the local eigenvalue statistics for a class of unitary random matrix ensembles. *Journal of Statistical Physics*, 86(1–2): 109–147

Plancherel, M. & Rotach, W. 1929. Sur les valeurs asymptotiques des polynomes d'Hermite $H_n(x) = (-1)^n e^{x^2/2} d^n (e^{-x^2/2})/dx^n$. *Commentarii Mathematica Helvetica*, 1: 227–254

Saff, E. & Totik, V. 1997. *Logarithmic Potentials and External Fields*, Berlin and New York: Springer

Szegö, G. 1975. *Orthogonal Polynomials*, 4th edition, Providence, RI: American Mathematical Society

Tracy, C.A. & Widom, H. 1994. Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159: 151–174

Tracy, C.A. & Widom, H. 1996. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177: 727–754

Tracy, C.A. & Widom, H. 1998. Correlation functions, cluster functions, and spacing distributions for random matrices. *Journal of Statistical Physics*, 92(5/6): 809–835

Widom, H. 1994. The asymptotics of a continuous analogue of orthogonal polynomials. *Journal of Approximation Theory*, 77: 51–64

Widom, H. 1995. Asymptotics for the Fredholm determinant of the sine kernel on a union of intervals. *Communications in Mathematical Physics*, 171: 159–180

RANDOM PROCESSES

See **Stochastic processes**

RANDOM WALKS

As conceptually simple stochastic processes, random walks have a plethora of practical modeling applications, leading to elegant and sometimes counterintuitive mathematics. Consider a particle constrained to move in one dimension, moving one unit to the left or right at each unit time step, and with the direction of motion for each time step chosen at random and independently of the particle's position or history. If the particle's position after n time steps is $S(n)$, one can assert

$$S(n) = S(0) + \sum_{i=1}^{n} X_i, \tag{1}$$

where each X_i is an independent identically distributed (IID) random variable taking values $+1$ with probability (say) p and -1 with probability $1-p$. If $p = 1 - p = \frac{1}{2}$, then the simple random walk (SRW) has no "preferred direction" and is unbiased. If $p > \frac{1}{2}$, then the particle is more likely to move right than left at each time step; the SRW is biased. The traditional pedagogical example imagines a drunken student's erratic progress along a street, where bias might be caused by the street being inclined or having a pub at one end.

Basic Results

The statistics of Equation (1) are simple and intuitively reasonable; thus

$$\begin{aligned} \mathrm{E}(S(n)) &= S(0) + n(2p-1) \\ \mathrm{Var}(S(n)) &= 4np(1-p), \end{aligned} \tag{2}$$

so the SRW moves with an average speed of $2p-1$ and spreads about the mean as $\sqrt{n}$ (characteristic of diffusive processes). This spread is greatest when the SRW is unbiased.

One can readily prove that the SRW is temporally and spatially homogeneous and Markovian (i.e., it does not matter where one places the origin of time or space, and conditional on the present, the future is independent of the past). Assuming without loss of generality $S(0) = 0$,

$$\begin{aligned} P(S(n) = a) &= C^n_{(n+a)/2} p^{(1/2)(n+a)} \\ &\quad \times (1-p)^{(n-a)/2}, \end{aligned} \tag{3}$$

the particle moves $(n+a)/2$ steps right and $(n-a)/2$ steps left to give a total of n steps and a net right displacement of a, and there are $C^n_{(n+a)/2}$ equally likely paths achieving this. N.B. for a to be attainable in n steps one requires $|a| \le n$ and $(n+a)/2$ to be an integer i.e., only odd positions are attainable at odd times, hence, the mysterious factor of 2 in Equation (5).

Results exploiting symmetries, homogeneities, and combinatorics, and also applications of probability- and moment-generating functions, are summarized in Grimmett & Stirzaker (2001, Chapters 3 and 4). More interesting questions involving absorption or reflection at barriers, dealt with using linear difference equations; where the nature of the barriers, dictates the appropriate boundary conditions, are also addressed therein.

Continuous Limit: Diffusion

For an unbiased SRW, manipulation of Equation (3) reveals, for sufficiently large a and n,

$$P(S(n) = a) \approx \sqrt{\frac{2}{\pi n}} \exp\left(\frac{-a^2}{2n}\right), \tag{4}$$

resembling a normal (Gaussian) probability distribution (Murray, 2002, Chapter 9). Indeed, if one lets the discrete space and time steps have sizes Δx and Δt, respectively, and forces these to zero such that

$$\lim_{\Delta x \to 0, \Delta t \to 0} \frac{(\Delta x)^2}{2\Delta t} = D \tag{5}$$

for some finite constant D (the diffusion coefficient), then one arrives at a diffusion equation,

$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial x^2}, \tag{6}$$

where $P(x,t)$ is the probability density function for the particle's position x at time t. Solutions of Equation (6) take the form indicated by Equation (4):

$$P(x,t) = \frac{1}{\sqrt{4\pi Dt}}\exp\left(\frac{-x^2}{4Dt}\right). \tag{7}$$

An alternative derivation, using Taylor series expansions of $P(x,t)$ (Murray, 2002, Chapter 9), lacks beauty and elegance but is more readily generalizable to the nonlinear RWs discussed below (Othmer & Stevens, 1997). If the SRW is biased, then an advection term,

$$v\frac{\partial P}{\partial x}$$

(where v represents the average velocity in the preferred direction) is added to the left-hand side of (6). If one considers $P(x,t)$ to represent a density of particles (rather than a probability density for a single particle), then one can incorporate local dynamics (e.g., chemical reactions or biological reproduction) at each point; a reaction term

$$f(P,x,t)$$

appears on the right-hand side of (6). The description of SRWs using diffusion (or more generally reaction-diffusion-advection) equations allows standard partial differential equation methods to be applied, often elucidating information concerning wave-like and spatially structured solutions (Murray, 2002; Okubo & Levin, 2001).

The extension of the SRW into higher dimensions presents no great conceptual, nor mathematical, challenges. If the processes governing steps in orthogonal directions are mutually independent, then one applies the preceding results to each direction separately. More realistically, suppose that at each time step a unit distance is moved in a direction chosen randomly from the infinite number of possibilities. Again, one can show that the process is, on average, diffusive (Denny & Gaines, 2000, Chapter 6):

$$E(S(n)) = S(0) \text{ and } \mathrm{Var}(S(n)) = n\,. \tag{8}$$

Importantly, motion in more than one spatial dimension allows more realistic random walk models to be developed, some of which are mentioned below.

Random Walks in Biology

Random walks have been widely applied in the life sciences. While ostensibly describing a particle's position, the modeled quantity might equally well refer to the amount of food consumed by an animal, the frequency of an allele in a population, or the contentious notion of "fitness." Crucially, random walk models are developed using local rules of "motion" at the level of the "individual" (organism/cell/molecule, etc.) before any scaling-up to a population level description is undertaken. This encourages the modeller to adhere to realistic assumptions and provides a link between concrete experimental observation and theory.

Several accessible applications are presented in Denny & Gaines (2000, Chapters 5 and 6). For example, Equation (6) can be solved to indicate the rate at which nutrients reach an idealized organism, D being strongly dependent on the organism's environment. Such considerations help to explain relative sizes of aquatic and aerial organisms and underly the varying mechanisms of nutrient transport in plant root systems. More complicated applications deal with the distribution of receptors on cell walls, the random drift to fixation of alleles in finite populations, and the energetics of protein folding. For more mathematically demanding applications involving spatial patterns and reaction-diffusion equations, see Murray (2002) and Okubo & Levin (2001).

Nonlinear Random Walks

The basic concept of a particle "deciding" (using IID random variables) its next direction at every step may be excessively restrictive, particularly in biology. For example, a particle moving in two dimensions may be likely to choose a direction similar to that used during the previous time step or may have some fixed preferred direction toward which its trajectory is biased. The mathematics of such "correlated" RWs is discussed in Othmer et al. (1988).

Alternatively, motion may be influenced by interactions with the local environment (reinforced RWs) leading to nonlinear generalizations of the simple diffusion equation; under reasonable assumptions, Equation (6) may be more properly expressed

$$\frac{\partial P}{\partial t} = D\frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}\left[\ln\frac{P}{\tau(W)}\right]\right), \tag{9}$$

where $\tau(W)$ represents a transition rate dictated by the environment W. Such systems can exhibit properties not found in diffusive SRW formulations; particles may tend to aggregate rather than disperse and may reach locally infinite or zero densities in finite time (Othmer & Stevens, 1997).

Numerical simulation of such systems can be built up from individual-level rules as in Sleeman & Wallis (2002) and Figure 1, where the authors model angiogenesis (the movement of endothelial cells to create a blood supply to an avascular tumor). The cells' motions are dictated by chemical gradients between the tumor and the blood vessel; at first the attractive signal from the tumor is weak and the motion tortuous, but

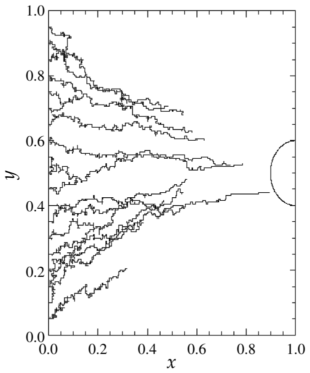

Figure 1. Tumor angiogensis modeled as a reinforced random walk. (Used with kind permission of I.P. Wallis and B.D. Sleeman, Department of Mathematics, University of Leeds, U.K.)

once the tumor's signal is sufficiently strong a more directed motion is observed.

Jon Pitchford

See also **Brownian motion; Deterministic walks in random environments; Martingales**

Further Reading

Denny, M. & Gaines, S. 2000. *Chance in Biology*, Princeton, NJ: Princeton University Press

Grimmett, G. & Stirzaker, D. 2001. *Probability and Random Processes*, Oxford and New York: Oxford University Press

Murray, J.D. 2002. *Mathematical Biology*, 3rd edition, Berlin and New York: Springer

Okubo, A. & Levin, S.A. 2001. *Diffusion and Ecological Problems*, Berlin and New York: Springer

Othmer, H.G., Dunbar, S.R. & Alt, W. 1988. Models of dispersal in biological systems, *Journal of Mathematical Biology*, 26(3): 263–298

Othmer, H.G. & Stevens, A. 1997. Aggregation, blowup and collapse: the ABC's of taxis in reinforced random walks. *SIAM Journal on Applied Mathematics*, 57(4): 1044–1081

Sleeman, B.D. & Wallis, I.P. 2002. Tumour induced angiogenesis as a reinforced random walk: modelling capillary network formation without endothelial cell proliferation. *Mathematical and Computer Modelling*, 36 (3): 339–358

RATCHETS

An essential and striking feature of living cells is their ability to generate mechanical motion and forces. Important examples are cell motility, muscle contraction, and active mass transport within cells, among other active phenomena in biology. These motions and forces are generated at the molecular level by protein molecules (called molecular motors or pumps) that are driven by chemical reactions in situations far from equilibrium (Astumian, 1997; Jülicher et al., 1997).

Two types of molecular motors are classified as linear (or translational) and rotary (or rotatory). Within the first class, motor proteins are classified into several families: myosins, kinesins, and dyneins. Each of these interacts specifically with a certain type of filament along which it is able to move in the presence of adenosine triphosphate (ATP), which is a chemical fuel. The filaments serve as guides or tracks for the motor motion. Two types of filaments play this role: microtubules and actin filaments. Both are formed by a polymerization process from identical monomers (actin and tubulin monomers, respectively), resulting in a regular and periodic one-dimensional structure. An important feature is their polarity; thus the asymmetry of the monomers that form a polar filament structure is essential for motor operation as it defines the direction of motion.

A typical representative of rotary molecular motors is ATP synthase composed of two coupled rotary engines: a membrane-embedded unit F_0 and a water-surrounded part F_1 (Oster & Wang, 2000) ATP synthase works as a reversible motor-pump machine. Thus, the proton flow through F_0 is believed to generate a mechanical torque driving the F_1 motor to synthesize ATP, and using the hydrolysis energy of ATP, the F_1 motor can drive the F_0 motor in reverse to pump protons. Direct observation of the rotation of the F_1 motor has been demonstrated in fascinating experiments by Japanese scientists (Noji et al., 1997).

Thus, molecular motors are microscopic objects that unidirectionally move along one-dimensional periodic structures. Because the motor operation is considered on the molecular level, the motor must be a Brownian object, subject to random fluctuations. A conversion (rectification) of random fluctuations $F(t)$ into useful work (i.e., a directed motion) is called a "ratchet effect." (The names *thermal ratchet*, *Brownian motor*, *Brownian rectifier*, *mechanical diode*, *stochastic ratchet*, or simply *ratchet* are also in use.) A comprehensive review of the work on ratchets and their practical applications is available in Reimann (2002).

Historically, the problem of converting Brownian motion (in general, unbiased random fluctuations) into useful work and its consistency with the second law of thermodynamics was introduced in a conference talk by Marian Smoluchowski in 1912 (Smoluchowski, 1912) and later popularized and extended by Richard Feynman (Feynman et al., 1963), using a so-called "ratchet and pawl" gadget. A simple stochastic model of this rectifying process (called a Smoluchowski–Feynman's ratchet) can be described using the overdamped Langevin equation for a Brownian particle moving in a one-dimensional viscous medium with friction coefficient η in the presence of a constant force F:

$$\eta\dot{x}(t) = -V'(x(t)) + F + \xi(t). \tag{1}$$

Here, the overdot and the prime denote differentiation with respect to the time t and the particle coordinate x, respectively. The potential $V(x)$ is a periodic function with period L, $V(x+L) = V(x)$, which has a broken parity symmetry. More precisely, this breaking of symmetry means that there is no Δx such that $V(-x) = V(x+\Delta x)$ for all x. A typical example of such a ratchet potential is a piecewise-linear (saw-tooth) function, consisting of two continuously matched linear pieces per period L, one with negative and one with positive slope, one being steeper than the other. The stochastic force $\xi(t)$ is a Gaussian white noise of zero mean, $\langle\xi(t)\rangle = 0$, satisfying the fluctuation-dissipation relation (Risken, 1989)

$$\langle\xi(t)\xi(s)\rangle = 2\eta k_{\mathrm{B}} T\delta(t-s), \tag{2}$$

where k_{B} is Boltzmann's constant, $2\eta k_{\mathrm{B}}T$ is the noise intensity or strength (T is absolute temperature), and $\delta(t)$ is Dirac's delta function.

The probability densities induced by the above equations obey a Fokker–Planck (FP) equation, which can be written in the form of a conservation law for probability, namely, a continuity equation (Risken, 1989)

$$\partial_t P(x,t) + \partial_x J(x,t) = 0. \tag{3}$$

Here the probability density $P(x,t)$ satisfies the normalization condition

$$\int_{-\infty}^{\infty} P(x,t)\,\mathrm{d}x = 1, \tag{4}$$

whereas the probability current is defined by

$$J(x,t) \doteq -\eta^{-1}[U'(x) + k_{\mathrm{B}}T\partial_x]P(x,t), \tag{5}$$

with the ratchet potential $V(x)$ and the force F being incorporated into a single effective potential

$$U(x) \doteq V(x) - Fx. \tag{6}$$

The quantity of foremost interest in the context of transport in periodic systems is the particle current, defined as the time-dependent ensemble average over the velocities, $\langle\dot{x}(t)\rangle$ (where for convenience the argument t is omitted). From this definition, the general connection between the probability current and the particle current is obtained as

$$\langle\dot{x}\rangle = \int_{-\infty}^{\infty} J(x,t)\,\mathrm{d}x. \tag{7}$$

Instead of solving the FP equation (3) describing the particle transport in periodic systems but with zero boundary conditions at infinity, it is convenient to rewrite it in terms of reduced probabilities (Reimann, 2002):

$$\hat{P}(x,t) \doteq \sum_{n=-\infty}^{\infty} P(x+nL,t), \quad \hat{J}(x,t) \doteq \sum_{n=-\infty}^{\infty} J(x+nL,t), \tag{8}$$

which are evidently periodic functions with period L. Furthermore, Equations (4) and (7) are transformed to

$$\int_0^L \hat{P}(x,t)\,\mathrm{d}x = 1 \tag{9}$$

and

$$\langle\dot{x}\rangle = \int_0^L \hat{J}(x,t)\,\mathrm{d}x, \tag{10}$$

respectively. Indeed, because $V(x)$ is periodic with period L, with $P(x,t)$ being a solution of the FP equation (3), $\hat{P}(x+nL,t)$ is also a solution for any integer n. Next, since the FP equation is linear, Equations (3) and (5) are also satisfied by the reduced quantities (8). Therefore, it is sufficient to solve an FP equation with periodic boundary conditions.

If the reduced dynamics assumes a steady state $\hat{P}^{\mathrm{st}}(x)$ in Equation (3), then the reduced probability current $\hat{J}(x,t) = \hat{J}^{\mathrm{st}}$ becomes independent of x and t. Multiplying this equation by x and integrating both its sides from x to $x+L$, one finds that the particle current takes the form

$$\langle\dot{x}\rangle = L\hat{J}^{\mathrm{st}}, \tag{11}$$

as expected. Moreover, using Equations (3), (5), and (11), it can easily be checked that the steady-state solution is given by (Stratonovich, 1969; Reimann, 2002)

$$\hat{P}^{\mathrm{st}}(x) = \mathcal{N}\frac{\eta}{k_{\mathrm{B}}T}\mathrm{e}^{-U(x)/k_{\mathrm{B}}T}\int_x^{x+L}\mathrm{e}^{U(y)/k_{\mathrm{B}}T}\mathrm{d}y, \tag{12}$$

$$\langle\dot{x}\rangle = L\mathcal{N}\Big[1 - \mathrm{e}^{[U(L)-U(0)]/k_{\mathrm{B}}T}\Big], \tag{13}$$

$$\mathcal{N} \doteq \frac{k_{\mathrm{B}}T}{\eta}\left[\int_0^L \mathrm{d}x\int_x^{x+L}\mathrm{e}^{[U(y)-U(x)]/k_{\mathrm{B}}T}\mathrm{d}y\right]^{-1}. \tag{14}$$

This solution is valid for a general potential $U(x)$ provided $U'(x+L) = U'(x)$. For the specific form (6), $U(L) - U(0) = -LF$, and therefore, as expected, the sign of the particle current (13) coincides with the sign of F. The absence of an average particle current ($\langle\dot{x}\rangle = 0$) if $F = 0$, in spite of the broken spatial symmetry, agrees with the second law of thermodynamics, like in the original ratchet and pawl gadget (Feynman et al., 1963). However, if the force

F in Equation (1) depends on time, a rectified motion (ratchet effect) may occur in an asymmetric system.

For the occurrence of the ratchet effect, it is sufficient to show that $\langle \dot{x} \rangle \neq 0$ provided $\langle F(t) \rangle = 0$. If the changes of $F(t)$ are extremely slow, then at any given instant t, the particle current has practically the same value as the steady-state current (13) with a static tilt F. In this adiabatic approximation, the time t plays the role of a parameter and using result (13) for a periodic driving force $F(t + \mathcal{T}) = F(t)$, one finds the time-averaged particle current:

$$\langle \dot{x} \rangle = \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \frac{L k_{\mathrm{B}} T \left[1 - \mathrm{e}^{-LF(t)/k_{\mathrm{B}}T}\right]}{\eta \int_0^L \mathrm{d}x \int_x^{x+L} \mathrm{e}^{[V(y)-V(x)-F(t)(y-x)]/k_{\mathrm{B}}T} \mathrm{d}y} \, \mathrm{d}t. \tag{15}$$

For general analytical conclusions, the adiabatic expression (15) is still too complicated. Only in simple special cases may one predict the direction of the current. As an example, consider the case of when the tilting force $F(t)$ takes only two possible values $\pm F_0$ with very rare deterministic or random flips for which $\langle F(t) \rangle = 0$, and $V(x)$ is a piecewise linear potential with two continuously matched linear pieces per period L, the slopes of which are different (one steeper and one flatter). Outside this fundamental cell of length L, the potential is periodically continued. In this case, the integrals in Equation (15) can be calculated analytically, yielding an explicit expression (Magnasco, 1993), which is positive if the flatter slopes are arranged in the positive direction. Intuitively, this ratchet effect can be understood as follows. For F_0 within a certain range, one of the two tilted asymmetric potentials $V(x) \mp F_0 x$ does not exhibit any extrema and, therefore, supports a permanent downhill motion (even if $T = 0$), while the other still exhibits extrema acting as motion-blocking barriers. Below this window, both $V(x) \mp F_0 x$ exhibit barriers and, thus, prohibit deterministic motion. Because the barrier induced by the steeper slope of $V(x)$ is higher than that induced by the flatter slope, a weak thermal noise induces the current in the direction of the flatter slope.

Alternatively, the ratchet effect in a spatially periodic system can be achieved under a time-dependent variation (pulsation) of the potential shape without affecting its spatial periodicity. Therefore, Equation (1) can be generalized to include both a pulsating $f(t)$ and tilting $F(t)$ forcing as follows (Jülicher et al., 1997; Reimann, 2002):

$$\eta \dot{x}(t) = -V'[x(t), f(t)] + F(t) + \xi(t), \tag{16}$$

where $V'(x, f) \doteq \partial_x V(x, f)$. Here $F(t)$ may contain a load force or torque, generally depending on time t.

Thus, two fundamental classes of ratchet models arise from Equation (16). The first comprises models with $F(t) \equiv 0$, which are called "pulsating" ratchets. Within these, the first main subclass, called "fluctuating potential" ratchets (Astumian & Bier, 1994; Prost et al., 1994), is obtained when $f(t)$ in Equation (16) is additive, i.e., $V(x, f(t)) = V(x)[1 + f(t)]$. This subclass contains as a special case the "on-off" ratchets (Ajdari & Prost, 1992) when $f(t)$ takes only two values, one of them being -1 (potential "off"). The on-off ratchets are suitable for experimental realizations and have been demonstrated (Rousselet et al., 1994) by means of colloidal polystyrene latex spheres suspended in solution and exposed to a dielectric ratchet potential, created by a series of "Christmas-tree" electrodes, which were turned on and off periodically. A similar experimental setup was used by Faucheux & Libchaber (1995) but with solutions containing *two* different species of particles at a time. As a result, it was demonstrated that they can be separated. A further experimental verification of on-off ratchets has been performed (Faucheux et al., 1995) using polystyrene spheres confined to an effective one-dimensional ratchet potential by laser-optical trapping techniques (optical tweezers). The second subclass of pulsating ratchets, called "traveling potential" ratchets, have potentials of the form $V(x, f(t)) = V(x - f(t))$.

The second class are models with $f(t) \equiv 0$, called "tilting" ratchets (Magnasco, 1993), so that $V(x, f(t)) = V(x)$ in Equation (16). When $V(x)$ is a ratchet potential, then $F(t)$ is mostly considered as a symmetric function. If $F(t)$ is a stochastic process, one speaks of a "fluctuating force" ratchet. The case of a tilting ratchet with a periodic driving $F(t)$ is of particular experimental relevance and carries the obvious name "rocking" ratchet.

Clearly, there are many possible combinations and generalizations of these classes; for example, a simultaneously pulsating and tilting ratchet or the simultaneous breaking of more than one symmetry (Reimann, 2002). For various applications in physics, chemistry, and biology, see Astumian (1997), Jülicher et al. (1997), and Reimann (2002).

ALEXANDER V. ZOLOTARYUK

See also **Brownian motion; Fokker–Planck equation; Nonequilibrium statistical mechanics**

Further Reading

Ajdari, A. & Prost, J. 1992. Mouvement induit par un potentiel périodique de basse symétrie: diélectrophorèse pulsée. *Comptes Rendus de l'Academie des Sciences, Paris, Série II*, 315: 1635–1639

Astumian, R.D. 1997. Thermodynamics and kinetics of a Brownian motor. *Science*, 276: 917–922

Astumian, R.D. & Bier, M. 1994. Fluctuation driven ratchets: molecular motors. *Physical Review Letters*, 72: 1766–1769

Faucheux, L.P., Bourdieu, L.S., Kaplan, P.D. & Libchaber, A. 1995. Optical thermal ratchet. *Physical Review Letters*, 74: 1504–1547

Faucheux, L.P. & Libchaber, A. 1995. Selection of Brownian particles. *Journal of the Chemical Society—Faraday Transactions*, 91: 3163–3170

Feynman, R.P., Leighton, R.B. & Sands, M. 1963. *The Feynman Lectures on Physics*, vol. 1, Reading, MA: Addison-Wesley (Chapter 46)

Jülicher, F., Ajdari, A. & Prost, J. 1997. Modeling molecular motors. *Reviews of Modern Physics*, 69: 1269–1281

Magnasco, M.O. 1993. Forced thermal ratchets. *Physical Review Letters*, 71: 1477–1481

Noji, H., Yasuda, R., Yoshida, M. & Kinosita, K. 1997. Direct observation of the rotation of F$_1$-ATPase. *Nature*, 386: 299–302

Oster, G. & Wang, H. 2000. Reverse engineering a protein: the mechanochemistry of ATP synthase. *Biochimica et Biphysica Acta*, 1458: 482–510

Prost, J., Chauwin, J.-F., Peliti, L. & Ajdari, A. 1994. Asymmetric pumping of particles. *Physical Review Letters*, 72: 2652–2675

Reimann, P. 2002. Brownian motors: noisy transport far from equilibrium. *Physics Reports*, 361: 57–265

Risken, H. 1989. *The Fokker–Planck Equation: Methods of Solution and Applications*, Berlin: Springer

Rousselet, J., Salome, L., Ajdari, A. & Prost, J. 1994. Directional motion of brownian particles induced by a periodic asymmetric potential. *Nature*, 370: 446–448

Stratonovich, R.L. 1969. *Theory of Random Noise*, London: Gordon and Breach

von Smoluchowski, M. 1912. Experimentell nachweisbare, der üblichen Thermodynamik widersprechende Molekularphänomene. *Physikalische Zeitschrift*, 13: 1069

RAYLEIGH AND RAMAN SCATTERING AND IR ABSORPTION

When an electromagnetic wave, $E = E_0 \cos 2\pi\nu t$, irradiates a molecule, a crystal, or a disordered substance, the energy may be scattered, absorbed, or transmitted. In the scattering process, the irradiated medium becomes polarized and the induced dipole moment (or electric polarization),

$$P = \alpha E + \cdots = \alpha E_0 \cos 2\pi\nu t + \ldots, \qquad (1)$$

oscillates synchronously with the electromagnetic field. The dipole radiates energy (the scattering), and α is the polarizability (usually a second-order tensor).

Irradiating with ultraviolet or visible light of frequency ν induces an electronic polarization that is modulated by the vibrational frequencies ($\nu_i, \nu_j, \nu_k, \ldots$) of the substance. Assuming a simple diatomic molecule, vibrating with frequency ν_1, the nuclear displacement is $q = q_0 \cos 2\pi\nu_1 t$, and $\alpha = \alpha_0 + (\partial\alpha/\partial q)_0 q$, then

$$\begin{aligned} P = {} & \alpha_0 E_0 \cos 2\pi\nu t \\ & + \frac{1}{2}\left(\frac{\partial a}{\partial q}\right)_0 q_0 E_0 \{\cos[2\pi(\nu + \nu_1)t] \\ & + \cos[2\pi(\nu - \nu_1)t]\}. \end{aligned} \qquad (2)$$

The emitted intensity is proportional to the square of the polarization $I \approx P^2$.

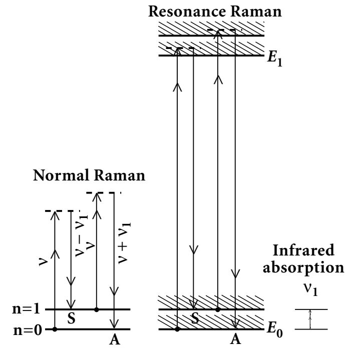

Figure 1. Level schemes for electronic and vibrational transitions involved in Raman scattering, resonance Raman scattering, and infrared absorption.

Rayleigh scattering occurs if the energy, ν, of the scattered photon is equal to that of the incident one (first term of Equation (2)). Rayleigh (1871) demonstrated that the Earth's blue sky was due to scattering of light by atmospheric particles, with a scattering intensity proportional to the inverse fourth power of the wavelength ($I \approx A/\lambda^4$). So, the blue (shorter wavelength) end of the visible spectrum is scattered more strongly, giving the sky its characteristic color.

As shown in Figure 1, the process is called Raman scattering, if the energy of the scattered photon is different from that of the incident one (the process is referred to as Brillouin scattering when the phonon emitted or absorbed is acoustic, and Raman scattering when the phonon is optical) (inelastic scattering, second term of Equation (2)). As a result, light with frequencies ($\nu \pm \nu_i$), ($\nu \pm \nu_j$), ($\nu \pm \nu_k$) ... is emitted. In order that the total energy may be conserved, the vibrational energy of the sample is changed. If the irradiated molecule gains energy, the scattered lines in the Raman spectrum are at ($\nu - \nu_i$) and are called "Stokes lines." If the sample loses energy, then the scattered lines are at larger frequencies than ν (at ($\nu + \nu_i$)...) and are called "anti-Stokes lines."

Classical theory shows that a vibration is Raman active if the polarizability changes during the vibration, and the Stokes lines are always stronger than the anti-Stokes lines. (This latter fact is explained by quantum theory as a consequence of the Maxwell–Boltzmann distribution law.) In crystals, the photon wave vectors are small compared with the dimensions of the Brillouin zone; thus, the momentum conservation law can be obeyed only if the wave vectors of the scattered phonons are also very small (so only the zone center phonons participate in the Raman scattering). It is difficult to determine which of the phonons are Raman active, as

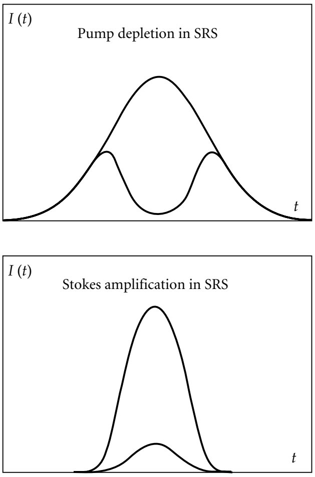

Figure 2. Stimulated Raman scattering: depletion at the pump frequency, and amplification of the Stokes beam.

changes of the components of the polarizability tensor are not obvious from inspection of the normal modes of vibration. Considering the symmetry of the molecule or the crystal and the polarization properties of the Raman lines, application of group theory gives a clear-cut solution to this problem (Poulet & Mathieu, 1970). If the polarization P contains terms of higher order (neglected in Equation (1)), these subsequent terms give hyper-Raman effects.

When the energy of the exciting line coincides with that of an allowed electronic transition of the compound, the intensity of the Raman spectra is enhanced and a series of overtones is observed—this mechanism is referred to as "resonance Raman scattering." This technique has been used in solid state and biophysics studies to identify intrinsic localized modes (Swanson et al., 1999).

Stimulated Raman scattering (SRS) is a nonlinear process in which a medium is irradiated by an intense laser beam at ν_p (pump frequency) and by a weaker beam at ν_s (Stokes frequency). The frequency spacing $\nu_\mathrm{p} - \nu_\mathrm{s}$ corresponds to a Raman frequency of the medium. In the simpler case, SRS amplifies the weak Stokes beam and depletes the pump beam (Raymer & Walmsley, 1991; see Figure 2).

A Raman soliton has been predicted (Chu & Scott, 1975) as a solution of the equations for transient SRS, and an experimental observation of transient SRS in H_2 gas was reported (Drühl et al., 1983). Recently, a complete interpretation of the above experiment has been given (Claude et al., 1995). The Raman soliton is still to be observed.

Picosecond time-resolved coherent anti-Stokes Raman spectroscopy (CARS) is a method for studying vibrational dephasing in the time domain. It is a special case of four-wave mixing (Levenson, 1982). Two intense simultaneous pulses at ν_p (pump) and ν_s (Stokes) specifically excite each Raman active mode at $\nu_\mathrm{p} - \nu_\mathrm{s}$. The third beam (probe pulse), also at the frequency ν_p, is delayed by the variable time t and stimulates coherent anti-Stokes emission. This technique is used to study the vibrational lifetime in anharmonic crystals and biomolecules to identify breathers excitations (Kosic et al., 1984).

Infrared (IR) absorption occurs when a compound is irradiated with frequencies that match its natural vibrational (or rotational) modes (Nakamoto, 1986). Absorption spectra in the infrared region originate in transitions between two vibrational levels of a sample in the electronic ground state. The absorption intensity depends on the change in dipole moment that occurs as a result of molecular vibration. If a molecule at equilibrium has a center of symmetry, the vibrations for which the center of symmetry is shared will be infrared inactive. Symmetry selection rules allow prediction of infrared activity when the associated normal coordinate belongs to the same species as one of the components of the dipole moment. When internal vibrations are coupled with phonons, nonlinear characters are observed in the infrared absorption bands: splitting, corresponding to the creation of a vibrational polaron or breather, intensity strongly dependent on the temperature, or an unusual number of overtones (Careri et al., 1984; Barthes et al., 2002). Time-resolved IR spectroscopy is of crucial importance in this field (Hamm et al., 1998; Xie et al., 2000).

MARIETTE BARTHES

See also **Nonlinear optics; Pump-probe measurements; Spectral analysis**

Further Reading

Barthes, M., Vik, A.F., Spire, A., Bordallo, H.N. & Eckert, J. 2002. Breathers or structural instability in solid L-Alanine: a new IR and inelastic neutron scattering vibrational spectroscopic study. *Journal of Physical Chemistry*, A 106(21): 5230–5241

Careri, G., Buontempo, U., Galluzzi, F., Gratton, E., Shyamsunder, E. & Scott, A.C. 1984. Spectrographic evidence for Davydov-like solitons in acetanilide. *Physical Review* B, 30: 4689–4702

Chu, F.Y.F. & Scott, A.C. 1975. Inverse scattering transform for wave-wave scattering. *Physical Review* A, 12: 2060–2064

Claude, C., Ginovart, F. & Léon, J. 1995. Nonlinear theory of transient stimulated Raman scattering and its application to long-pulse experiments. *Physical Review* A, 52: 767–782

Drühl, K., Wenzel, R. & Carlsten, J. 1983. Observation of solitons in stimulated Raman scattering. *Physical Review Letters*, 51: 1171

Hamm, P., Lim, M. & Hochstrasser, R.M. 1998. The structure of the amide I band of peptides measured by femtosecond nonlinear IR spectroscopy. *Journal of Physical Chemistry* B, 102: 6123–6138

Kosic, T., Cline, R. & Dlott, D. 1984. Picosecond coherent Raman investigation of the relaxation of low frequency vibrational modes in amino acids and peptides. *Journal of Chemical Physics*, 81: 4932

Levenson, M. 1982. *Introduction to Nonlinear Laser Spectroscopy*, New York: Academic Press

Nakamoto, K. 1986. *Infrared and Raman Spectra of Inorganic and Coordination Compounds*, New York: Wiley

Poulet, H. & Mathieu, J.P. 1970. *Spectres de Vibration et Symétrie des Cristaux*, London: Gordon & Breach

Rayleigh, Lord (John William Strutt). 1871. On the light from the sky, its polarization and colour. *Philosophical Magazine*, 41: 107–120, 274–279

Raymer, M. & Walmsley, I.A. 1991. The quantum coherence properties of stimulated Raman scattering. *Progress in Optics*, 28: 216

Swanson, B.I., Brozik, J.A., Love, S.P., Strouse, G.F., Shreve, A.P., Bishop, A.R., Zang, W.-Z. & Salkola, M.I. 1999. Observation of intrinsically localized modes in a discrete low-dimensional material. *Physical Review Letters*, 82: 3288–3291

Xie, A.H., van der Meer, L., Hoff, W. & Austin, R.H. 2000. Long-lived amide I vibrational modes in myoglobin. *Physical Review Letters*, 84: 5435–5438

RAYLEIGH–BÉNARD CONVECTION

See **Thermal convection**

RAYLEIGH–TAYLOR INSTABILITY

The Rayleigh–Taylor instability arises when a layer of heavy fluid is accelerated into a lighter one. This occurs, for instance, if a heavy fluid is put on top of a lighter one in a uniform gravity field. Under these unstable conditions, small disturbances of the interface between the two layers grow, at first exponentially with a growth rate proportional to $\sqrt{k}$ where k is the wave number, as long as the perturbation is small compared with the wavelength. The subsequent nonlinear growth has been mostly investigated numerically or experimentally. The Rayleigh–Taylor instability was first discovered by Lord Rayleigh (John William Strutt) in 1883 for the gravitational case and then extended to the case of any accelerating fluid by Geoffrey Taylor in 1950.

One of Taylor's major observation is related to the Manhattan Project (September 1942 to January 1945), which ultimately led to the successful development of atomic bombs. Indeed, in May 1944, when he was invited to join the project, Taylor pointed out problems due to implosion instabilities (especially of the Rayleigh–Taylor type). This led to a very conservative design to minimize possible instabilities. At the same time, James Tuck brought the idea of explosive lenses for detonation wave shaping, suggesting the use of three-dimensional (3-d) lenses to create a spherical implosion and reduce instabilities.

To study an example of the Rayleigh–Taylor instability, let us look at the interface between a heavy fluid of density ρ_2 on top of a lighter fluid of density ρ_1 subject to gravity. The interface is assumed to be horizontal and both fluids are assumed to be at rest initially. For simplicity, we look at the two-dimensional case. We wish to show that a small disturbance of the interface will grow exponentially fast. Indeed, a small disturbance can engender small potential velocities $v_1 = -\nabla\phi_1$ and $v_2 = -\nabla\phi_2$ in the lower and upper fluid respectively, such that

$$\phi_2 = A\mathrm{e}^{-ky+nt}\cos(kx), \tag{1}$$

$$\phi_1 = -A\mathrm{e}^{ky+nt}\cos(kx). \tag{2}$$

Then the interface that is given by the equation $y = \eta(t,x)$ satisfies $\partial_t\eta = -\partial_y\phi_1(y=0) = -\partial_y\phi_1$ $(y=0)$. Hence,

$$\eta = \frac{Ak}{n}\mathrm{e}^{nt}\cos(kx). \tag{3}$$

Neglecting the nonlinear term, the Euler equation reads $-\rho_2\partial_t\nabla\phi_2 + \nabla p_2 = -\rho_2 g\boldsymbol{j}$, which yields

$$p_2 = p - \rho_2 gy + n\rho_2\phi_2, \tag{4}$$

where p is the mean pressure at the interface. Similarly, the pressure in the lower fluid is given by

$$p_1 = p - \rho_1 gy + n\rho_1\phi_1. \tag{5}$$

At the interface, the pressure should be continuous, namely, $p_1 = p_2$ and, hence,

$$(\rho_2 - \rho_1)g\eta = n(\rho_2 + \rho_1)A\mathrm{e}^{nt}\cos(kx). \tag{6}$$

Combining (3) and (6), we get

$$n^2 = -kg\frac{(\rho_1 - \rho_2)}{(\rho_2 + \rho_1)}. \tag{7}$$

The Atwood number is defined by $At = (\rho_1 - \rho_2)/(\rho_2 + \rho_1)$. In the case $\rho_2 > \rho_1$ (heavy fluid on top of lighter fluid), $-1 < At < 0$, and

$$n = \pm\sqrt{-At\,kg}. \tag{8}$$

The unstable mode corresponds to $n = \sqrt{-At\,kg}$ and yields exponential growth as long as the nonlinear terms can be neglected, namely, as long as η is small compared with the wavelength. In experiments, surface tension and viscosity have a regularizing effect on the motion by damping the waves with large wave numbers. Indeed, for small Weber (and/or) large Reynolds numbers, the Rayleigh–Taylor instability has a short-wave character due to a balance between the exponential growth and the viscous or the surface tension effect, and hence, there exists a wave number that has a maximum growth rate. This is similar to the Rayleigh number, which represents the ratio of the destabilizing effect of buoyancy to the stabilizing effect of viscous force in the Rayleigh–Bénard instability.

Experiments and numerical simulations have given much insight into the different stages of the instability formation. However, they have yielded few quantitative results. Different numerical methods have been used such as the 3-d front tracking and the lattice Boltzmann method. In all cases, the geometric complexity of the interface is a source of difficulties for most of the algorithms. As explained in Sharp (1984), we can divide the growth of the instability into four stages. In the first, we have an exponential growth that can be quantitatively computed using the linear perturbation theory, which is no longer valid if the amplitude is about 20% of its wavelength. During the second stage, the perturbation grows nonlinearly to form bubbles of light fluid rising into the heavy one and spikes of heavy fluid falling into the light one. This stage is strongly influenced by the three-dimensional effects. Experiments and numerical simulations have shown that the bubbles rise at an approximately constant velocity during this stage. The third stage is characterized by the development of additional structures on the spikes. Moreover, the Kelvin–Helmholtz instability begins to develop due to the jump in the tangential velocity at the interface between the two fluids. The heavy fluid begins to roll up along the sides of the spikes to form "mushrooms." This phenomenon is more pronounced when the Atwood number is small and is a 3-d phenomenon. Eventually, the flow evolves into turbulent or chaotic mixing, which dominates the fourth stage.

In engineering applications, the Rayleigh–Taylor instability plays an important role in inertial confinement fusion (ICF). It is also present in the extraction of oil. Indeed, to extract oil from the Earth, water is accelerated into oil, and since oil is lighter, the interface presents the Rayleigh–Taylor instability. Some of the engineering solutions consist in adding polymers to the water. Moreover, there are many astrophysical and geophysical objects that present the Rayleigh–Taylor instability such as the supernova explosions, mantle convection, and deep convection in the ocean.

Another instability that is similar to the Rayleigh–Taylor one is the Ritchtmeyer–Meshkov instability, which arises in the case of an impulsive or instant loading as opposed to a constant acceleration (such as gravity) in the Rayleigh–Taylor case. In many applications the situation is intermediate between these two limiting cases.

NADER MASMOUDI

See also **Kelvin–Helmholtz instability; Wave stability and instability**

Further Reading

Chandrasekhar, S. 1961. *Hydrodynamic and Hydromagnetic Stability*, Oxford: Oxford University Press

Rayleigh, Lord (John William Strutt). 1883. Investigation of the character of the equilibrium of an incompressible heavy fluid of variable density. *Proceedings of the London Mathematical Society*, 14: 170–177

Sharp, D.H. 1984. An overview of Rayleigh–Taylor instability. *Physica* D, 12: 3

Taylor, G.I. 1950. The instability of liquid surfaces when accelerated in a direction perpendicular to their planes. I. *Proceedings of the Royal Society of London, Series* A, 201: 192–196

REACTION-DIFFUSION SYSTEMS

How did life appear and develop on Earth? Although there is not complete agreement, a convincing answer is that given the required conditions (which were almost surely satisfied billions of years ago), life spontaneously emerged through the endless battle of survival of the fittest from a primordial chemical soup. The theory of random catalytic networks (Kauffman, 1995) shows that autocatalytic reactions are likely in this context; thus, the theory of interactions between chemical reactions and molecular diffusion takes center stage in emergence of biological life from atoms and molecules in a system.

Today, reaction-diffusion systems have found many applications ranging from chemical and biological phenomena to medicine (physiology, diseases, etc.), genetics, physics, social science, finance, economics, weather prediction, astrophysics, and so on (Aronson & Weinberger, 1975; Grindrod, 1996; Murray, 2002; Scott, 2003). Even for phenomena that bear no initial resemblance to these processes, it is sometimes useful and productive to use the reaction-diffusion metaphor in order to gain insight into their dynamics.

An important contribution to this subject comes from the theory of pattern formation in nature. Many physical phenomena giving rise to natural patterns can be understood in terms of the interaction of a short-range self-enhancing reaction and a long-range antagonistic reaction. Take a fire, for example. It is a self-enhancing process: more heat is released as more fuel is burned. In the process oxygen and fuel which act as antagonistic factors are consumed and this may lead to a fire's extinction if fuel is not replenished. Also, the heat produced is transported from its local source through diffusion. Thus, we have the two main ingredients for pattern formation, namely, a local antagonistic interaction between two species (or reactions) coupled with a means of transport of their products. Historically, this is also one of the first examples of reaction-diffusion systems studied scientifically, for obvious reasons given the necessity for improved heating and lighting at the beginning of the industrial age. In his "Christmas Lectures" at the Royal Society in London, Michael Faraday discussed the importance of understanding the candle flame and its analogy with the process of respiration of biological organisms (Faraday, 1861). The flame of the candle is an archetypical nonlinear reaction-diffusion system that today, after decades of research, provides a basis for the science of combustion. In this article, we review the essential topics in the field of reaction-diffusion systems from a theoretical point of view and also consider some of their numerous applications.

Theoretical Aspects

A reaction-diffusion (RD) equation is typically obtained by combining Fick's law of diffusion with the chemical reaction rate law. Although the theory can be made rigorous by using the theory of stochastic differential equations leading to the Fokker–Planck equation (Øksendal, 2003), we present here a heuristic argument. If we consider for simplicity a small domain interval on the line inside which we have a concentration c of some reacting species, then the diffusive flux J_{in} of c into one side of the small region will depend on the concentration gradient, $\partial c/\partial x$, at that boundary and the diffusion coefficient, D, with $J_{\mathrm{in}} = -D(\partial c/\partial x)_{\mathrm{in}}$. The parameter $D > 0$ is called diffusivity with physical units of m^2/s. The diffusive flux out of the region at the other side J_{out} will similarly be given by $J_{\mathrm{out}} = -D(\partial c/\partial x)_{\mathrm{out}}$, where the concentration gradient is now evaluated at the other boundary. The rate at which the concentration grows due to diffusion then depends on the difference between these two fluxes—and so involves the second derivative $\partial^2 c/\partial x^2$. If we add a kinetic reaction rate term $r(c)$, then the reaction-diffusion equation, which gives the rate of change of the concentration c in time at any spatial point, has the general form

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} + r(c). \tag{1}$$

This can be extended to any number of spatial variables to read

$$\frac{\partial c}{\partial t} = D\Delta c + Q(\boldsymbol{x}, t, c, \ldots), \tag{2}$$

where Δ denotes the n-Laplacian ($1 \le n \le 3$) and Q accounts for other influences including sources or sink terms. Other more complicated formulations for the flux terms are possible in diffusive processes; see Okubo & Levin (2001) for an account of these ideas in biology. In many sciences, the motion of particles or living organisms is subjected to both internal and external effects often acting simultaneously. For example, in biology, bacteria are known to move randomly (akin to diffusion) but are also able to follow a chemical gradient (chemotaxis). Mathematically, this leads to a description involving reaction-diffusion-chemotaxis equations. For example, chemical attractant signaling can give rise to spiral wave pattern formation at a certain life stage in colonies of the slime mold *Dictyostelium discoideum*. Other extensions include advection, electric and/or magnetic field effects, and so on. In many of these cases, the resulting mathematical description is far more complicated than the simple RD systems above and, for many situations, their exact formulation is still an open question.

In writing an equation such as (2), we will consider that c is a vector; thus, Equation (2) will describe a system of RD equations. A simple archetypical example for a RD system is a quadratic autocatalytic reaction between two chemicals according to the rule $\mathrm{A} + \mathrm{B} \to 2\mathrm{B}$ with rate $r = kab$. Denoting by a the concentration of A and by b that of B, the two species satisfy (after suitable scaling) the equations:

$$\begin{aligned}\frac{\partial a}{\partial t} &= D_1\Delta a - ab,\\ \frac{\partial b}{\partial t} &= D_2\Delta b + ab.\end{aligned} \tag{3}$$

From a mathematical point of view, systems of equations such as (3) must be well posed in order to exhibit appropriate solutions. To specify the problem, let the state variables $c(\boldsymbol{x}, t), \ldots$ represent the density or concentration of some substance at time $t \ge 0$ and position $\boldsymbol{x}$ in R^n. Then Δc denotes the Laplacian of c with respect to the space variable $\boldsymbol{x}$, and Equation (1) is an example of a parabolic equation of evolution. If (1) holds for all $\boldsymbol{x}$ in R^n, then the problem is fully specified once appropriate initial conditions

$$c(\boldsymbol{x}, 0) = c_0(\boldsymbol{x}) \tag{4}$$

are known. If (1) holds in a limited domain $\Omega \subset R^n$, then we must impose boundary conditions on c at $\partial\Omega$ (the domain boundary) compatible with the physical situation. Neumann (or no-flux), Dirichlet, or Robin conditions are usually employed in applications. Although these are linear conditions on the variable c, nonlinear conditions could also be applied. The above system of differential equations with specified initial and boundary conditions is called an initial value problem or IVP in short.

For an IVP, one naturally asks whether there are solutions. In the case of reaction-diffusion systems, there are two different aspects to consider; local existence and global existence of the solutions. By local existence we mean the existence of the solutions over a short time interval. Global existence properties are exhibited by the solutions of the IVP when they are known to exist for all positive time. In some applications global existence is precluded because the solution exhibits blow-up in finite time. This means that there is $(\boldsymbol{x}_0, t_0)$ such that $c(\boldsymbol{x}, t) \to \infty$ as $(\boldsymbol{x}, t) \to (\boldsymbol{x}_0, t_0)$. These questions are difficult to deal with in general for RD systems although there is a well-developed body of results available in the literature (Grindrod, 1996). Moreover, even if the existence of a particular type of solution is established, further important theoretical questions involve the uniqueness of the solution and the stability of this solution to small perturbations. To approach these questions, one usually transforms the given system of RD equations

$$\frac{\partial c}{\partial t} = D\Delta c + f(\boldsymbol{x}, t, c, \nabla c) \tag{5}$$

into a differential equation in an abstract Banach space (for example, $L^2[\Omega]$) as in

$$c_t + Ac = f, \quad t > 0, \quad c = c_0 \quad \text{at}\, t = 0. \tag{6}$$

Then, the above questions reduce the problem to studying the properties of the (linear) operator A, which are mainly resolved if one knows its spectrum. However, in practice this is a difficult question as A has an infinite-dimensional spectrum.

For many RD systems of interest, in practice, there are several rather specific and powerful analytical methods that give detailed insights into the solutions, including comparison principles, invariant regions, matched asymptotic expansions, nonlinear bifurcation, group invariant symmetries, and so on (Grindrod, 1996).

Applications of Reaction-Diffusion Systems

Historically, some of the first applications of reaction-diffusion equations were in population dynamics, combustion, and nerve impulse conduction. Thus, one of the simplest reaction-diffusion equations is the Fisher–KPP equation

$$\frac{\partial u}{\partial t} = D\Delta u + f(u), \tag{7}$$

which was proposed independently by Ronald Fisher (1937) and Andrei Kolmogorov, Ivan Petrovsky and N. Piscounoff (1937) to explain the spread of genetic influences. In those papers a quadratic function of the form

$$f(u) = ku(1-u) \tag{8}$$

was used with $k > 0$ a parameter. A year later, Yakov Zeldovich and David Frank-Kamenetsky used the same equation but with f being a cubic polynomial as a model that represented flame front propagation (Zeldovich & Frank-Kamenetsky, 1938). Due to its generic form, Equation (7) soon found new interesting applications, for example, as a model of impulse conduction along an active nerve fiber, with the solution $u(x,t)$ representing the voltage across a cell membrane.

An important class of similarity solutions in the form of traveling waves (TW) is common to both the RD system (3) (and many of the type at (5)) and scalar equation (7). A traveling wave is a solution of the form $u(x,t) = u(x - vt) = u(y)$, where $y = x - vt$ is the traveling-wave coordinate with v being the wave speed. The problem is simplified because in the TW coordinate we have now to solve an ordinary differential equation (rather than a partial differential equation). For example, (7) generates the following TW equation to solve

$$\frac{\mathrm{d}^2u}{\mathrm{d}t^2} + v\frac{\mathrm{d}u}{\mathrm{d}y} + f(u) = 0, \tag{9}$$

which must be supplemented with appropriate boundary conditions as $|y| \to \infty$.

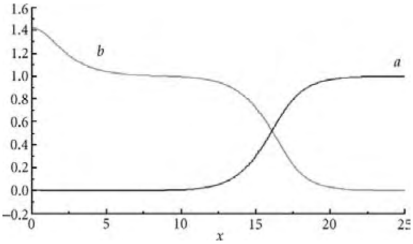

Figure 1. Typical concentration profiles for solutions a, b obtained by numerically solving the RD system (3).

Propagating waves are an important dynamical feature of many physical systems, hence, TW problems have been carefully studied. Key mathematical questions in these cases are the existence and possibly uniqueness of the waves. After the existence of a TW solution is established, it is of physical significance to study its stability under perturbations. Other questions are the influence of the initial conditions on the selection of a particular type of TW solution and the study of the shape of the TW solution. An extensive theory for TW solutions has been developed that deals with many interesting classes of RD systems (see Volpert et al., 1994; Grindrod, 1996; Scott, 2003). One particularly striking property of TW solutions for RD equations such as (3) or (7) is the existence of either a unique speed or a semi-infinite spectrum of speeds with a positive threshold boundary. The first case corresponds to a cubic nonlinearity, as in the nerve conduction application above, whereas the latter applies to a quadratic nonlinearity. These features can be contrasted with the paradox of infinite speed behavior exhibited by the classical diffusion equation. For example, with the quadratic form (8), Equation (7) has a TW solution for all $v \geq v_0 = 2\sqrt{kD}$. A similar property applies to system (3) although in practice only for a limited class of RD equations is an analytical expression known for v_0. In Figure 1, we show a typical profile of a TW solution to system (3) obtained from numerical simulations on a semi-infinite one-dimensional spatial domain via an implicit finite-difference scheme.

One of the oldest but still active fields of application is the modeling of nerve impulse conduction along a

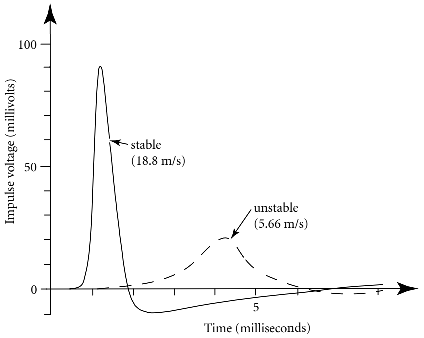

Figure 2. A full-sized action potential and an unstable threshold impulse for the HH axon: (Courtesy of A.C. Scott.)

responsive fiber (Scott, 2003). In 1952, Alan Hodgkin and Andrew Huxley formulated a detailed description of the dynamics of membrane ionic current for a dozen giant squid axons (Hodgkin & Huxley, 1952). They produced the following differential system describing the dynamics of the voltage $V(x,t)$ across the nervous cell (assumed one-dimensional):

$$rc_a \frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2} + rj_i, \qquad (10a)$$

where

$$j_i = 2\pi a J_i\,,\ J_i = G_{\mathrm{K}} n^4 (V - V_{\mathrm{K}}) + G_{\mathrm{Na}} m^3 h (V - V_{\mathrm{Na}}) + G_{\mathrm{L}}(V - V_L). \qquad (10b)$$

Here, a is the radius of the fiber, r is the longitudinal resistance per unit length of the fiber, c_a is the membrane capacitance per unit length, and j_i is the total ionic current flowing across the membrane per unit length. The expressions appearing in Equation (10b) are given in terms of differential first-order rate equations describing the concentrations of potassium, sodium "turn-on" and "turn-off" variables (n, m, and h). These equations are determined from experimental data. The RD system (10a,10b) is now known as the Hodgkin–Huxley (HH) system. Using TW theory, one can show that HH is an excitable RD system for parameter ranges of interest that admit traveling pulse solutions called action potentials (see Figure 2).

From a theoretical perspective, the HH system is complicated; thus, simplified theoretical models have been proposed to capture qualitative features. One such model is the system proposed by FitzHugh and Nagumo (FitzHugh, 1961), which reads

$$\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2} + F(V) + R\,,\ \frac{\partial R}{\partial t} = \varepsilon(V - c_a - b_a R), \qquad (11)$$

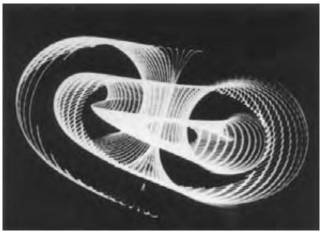

Figure 3. A computer generated scroll wave. (Courtesy of A. Winfree.)

where R is the slowly changing recovery variable and F has a cubic shape. The FitzHugh–Nagumo system admits periodic traveling waves and pulses both stable and unstable. In 2 or 3 spatial dimensions, target (ring) and spiral waves have been numerically produced that today are known to be characteristic types of TW solutions of an excitable system (Scott, 2003).

The most striking illustrations of target and spiral waves can be seen in excitable chemical reactions. Historically, the first such demonstration was produced in the chemical medium proposed by Belousov and Zhabotinsky, now known as the BZ reaction (Zhabotinsky, 1991). A dynamical system is called excitable when it responds in a qualitatively different way to perturbations that are below and above some threshold value, above which it typically amplifies those perturbations. After a pulse has been generated, the system takes some time (refractory state) until it can again support an impulse. In two dimensions, one can find target (ring) and spiral waves as in a typical BZ experiment (*See* photo in **Belousov–Zhabotinsky reaction** and also in color plate section).

To describe these solutions analytically, one can use a simplified model of the BZ reaction, such as the two-species Oregonator model developed by Field and Noyes (1974):

$$\begin{aligned} \varepsilon\frac{\partial a}{\partial t} &= \varepsilon^2 \Delta a + a(1-a) - f\frac{a-q}{a+q}b, \\ \frac{\partial b}{\partial t} &= \varepsilon \Delta b + b - a, \end{aligned} \qquad (12)$$

where ε, f, q are kinetics parameters related to the reagents rate reactions and a, b are proportional with the concentrations of $\mathrm{HBrO_2}$ and a metal ion. Here the diffusion operator Δ denotes one-, two-, or three-dimensional diffusion. In the latter case, more complicated patterns are found such as toroidal scroll waves and multi-armed spiral waves. Figure 3 shows a computer generated scroll ring obtained in a model of the BZ reaction, which has also been observed experimentally.

To understand such solutions a geometrical theory of traveling waves propagating on arbitrary two- and three-dimensional manifolds was developed by Tyson, Keener, Grindrod, and others around the mid-1980s, (see Grindrod, 1996). The idea is to compare the problem of three-dimensional wave propagation with that for wave propagation in one dimension. Consider the equation

$$\varepsilon \frac{\partial u}{\partial t} = \varepsilon^2 \Delta u + f(u), \tag{13}$$

which is in the Oregonator model of the BZ reaction (see Equation (12)). Suppose that u_1, u_2 are two solutions to $f(u) = 0$ and that (13) has a TW solution of the form $u = u(y)$ where $y = (x - vt)/\varepsilon$, so that

$$u'' + vu' + f(u) = 0 \tag{14}$$

satisfying $u \to u_1$ as $y \to \infty$ and $u \to u_2$ as $y \to -\infty$ for some speed $v > 0$. For a TW structure to exist in three dimensions, there should be an oriented surface M moving through R^3 such that $u \sim u_1$ ahead of the surface and $u \sim u_2$ behind it. Introducing a normal stretched variable depending on a curvilinear system of coordinates, one can show that a solution $u = \phi(\xi)$ exists subject to $\phi \to u_1$ as $\xi \to \infty$ and $\phi \to u_2$ as $\xi \to -\infty$ if and only if ϕ satisfies the TW equation

$$\phi_{\xi\xi} + (N + \varepsilon K)\phi_\xi + f(\phi) = 0. \tag{15}$$

From (14) and (15) we find that $u = \phi(\xi)$ is the inner solution of the problem when

$$v = N + \varepsilon K, \tag{16}$$

which is called the eikonal equation (Grindrod, 1996). Here N denotes the normal velocity, and K is twice the mean curvature of M. For example for circular waves, the speed of the curved wave v should be smaller than that of the planar wave v_0. The eikonal equation gives $v = v_0 - 2D/r$, where r is the radius and D is the diffusion coefficient. Hence we have a condition for the initiation of circular waves that is purely geometric, namely, $r_{\text{crit}} = 2D/v_0$ for the critical radius of wave initiation. When two circular waves collide, they create cusp-shaped regions with positive curvature. The theory can also be applied to armed spirals and toroidal scroll waves (Grindrod, 1996).

Spiral waves generated by reaction-diffusion processes also occur in surface reactions as tiny spiral waves, visible only through a microscope, that form out of catalytic nitric oxide reduction by hydrogen on rhodium or platinum surfaces. Such pattern formation in surface reactions can be imaged by photoemission electron microscopy (PEEM).

Flames on a spinning disk have been studied at NASA where it is shown that target and spiral waves appear in a mixture of butane in He–O_2. A simple such model is the Salnikov RD system (Scott et al., 1997):

$$a_t = \nabla^2 a + \mu - af(b), \tag{17a}$$

$$b_t = Le\nabla^2 b + \frac{1}{\kappa}(af(b) - b). \tag{17b}$$

Figure 4. A spiral galaxy in Centaurus. VLT, European Southern Observatory.

Here $f(b) = \exp[b/(1 + \gamma b)]$ and $Le, \mu, \kappa, \gamma > 0$ are physical parameters with $\kappa \ll 1$.

Spiral forms are omnipresent throughout the visible and invisible universe in galaxies, accretion disks around black holes, coalescing interstellar clouds, and many other forms of matter and energy (see Figure 4). Lee Smolin (1996) has suggested that the formation of some spiral galaxies can be regarded as an RD process, and he produced a theory of cosmos as a self-sufficient organically developing natural system. In this scenario the galactic spirals (and everything inside them) are gargantuan relatives of the BZ waves.

Waves of spreading depression appear as electrochemical waves associated with the depolarization of the neuronal membrane of the brain as they spiral around lesions of the cortex. There is speculation that this wave activity is linked to epilepsy. Since Art Winfree's pioneering work, scroll waves have been linked quite firmly with cardiac arrhythmias (Keener & Sneyd, 2002). In particular, the onset of fibrillation seems to be connected with the development of scroll type waves (high-frequency waves of electrical activation that recirculate repeatedly, preventing normal function) in the heart muscle. In this case, the heart contractions are much smaller in amplitude than the normal coordinated contraction and also aperiodic, which has potentially fatal consequences.

In early embryo development, circular calcium waves propagate and occasionally annihilate on the surface of a fertilized egg cell as a precursor to cell division. However, in frog eggs they take the form of spiral waves whose purpose is still unclear.

The dynamics of biological populations can self-organize in propagating traveling structures akin to

RD waves (Okubo & Levin, 2001). This is a classical topic in ecology with ramifications for studies in biodiversity. Among the pioneers of the field were Alfred Lotka in 1910 (Lotka, 1925) and Vito Volterra (1926), who have shown that equations similar to those describing reacting chemicals can provide a crude description of interactions between a predator population and its multiplying prey population. Thus, Lotka–Volterra RD type interactions are the heart of modeling in much of ecological dynamics and modeling of epidemics. The spatial spread of epidemics can be usefully approached with a "prey-predator" type dynamics involving the three species: S (the susceptibles), I (infectives), and R (removed individuals) in a population (Kermack & McKendrick 1932; Murray, 2002):

$$\begin{aligned} S_t &= -\beta IS, \quad I_t = \beta IS - \gamma I + D\Delta I, \\ R_t &= \gamma I. \end{aligned} \tag{18}$$

Equations (18) give a simple model of the spatial spread of rabies by fox, where D is the diffusion coefficient of the infectious foxes and β represents the rate of infection per susceptible per unit density of infectives. More sophisticated versions of such models that are used today to model epidemics (both in humans and in animals) can capture other effects such as secondary outbreaks.

The above discussion has mainly concentrated on the applications of the traveling-wave solutions of the RD systems. However, a much richer solution structure exists in these systems. In fact, the spectrum of possibilities ranges from simple stationary (or time independent) solutions up to complicated spatiotemporal chaotic solutions.

Perhaps one of the most spectacular applications of RD systems is in morphogenesis. The basic question here is to understand how the complicated process of shaping and patterning in an early embryo takes place starting from a uniform structure, the initial egg. In 1952, Alan Turing proposed that this process could be accounted for in terms of the underlying chemical reactions taking place in the embryo. In his paper Turing proposed some hypothetical reactions that could generate spontaneous symmetry breaking, leading to stable spatial patterns (Turing, 1952). Today this mechanism for spatial pattern formation is called Turing instability (or diffusion-driven instability) in recognition of Turing's seminal work. A counterintuitive feature of Turing's prediction is that diffusion can act as a destabilizing force. Murray (1988, 2002) showed that activator-inhibitor RD systems can model the beautiful array of patterns seen on animal skin markings. The idea Murray presented is that at a very early stage, the embryo acquires a "pre-pattern" of chemical morphogens (in Turing's parlance) that is later read out by melanocytes (pigment producing cells). In this way, spotted or striped or a combination of both patterns can be accounted for. Furthermore, Murray showed how the geometrical aspects of the RD domain can considerably alter the final outcome of the patterns. This theory explains why spotted animals (such as cheetah, jaguars, and leopards) can have striped tails, but striped animals (such as zebras and genets) cannot have spotted tails. An equally spectacular application of Turing's theory was proposed by Meinhardt (1998) to account for many patterns seen on seashells.

The book by Murray (2002) contains further references to expanding the above RD framework on modeling crucial processes in morphogenesis, such as feather germs, the initiation of teeth primordia, cartilage and condensation in limb, and epidermis development. Most of them require expansion of the RD framework to incorporate mechanics leading to the study of mechanochemical models.

With growing evidence of biological morphogens in living tissues, it is now clear that Turing's theory may play a crucial role in explaining biological development. Furthermore, a recent extension of the applications of RD systems in biological pattern formation has strengthened considerably the theoretical arguments in favor of the role of RD systems in biological modeling. Satnoianu et al. (2001) have shown that coupling of diffusion by advection in the presence of autocatalytic reactions can lead to an improved recipe for pattern formation, which is able to account also for biological growth. The new mechanism of patterning, now termed flow and diffusion distributed structures (or FDS), produces both stationary and traveling structures and, thus, can be viewed as an unifying theoretical construct for many of the patterning processes described above (Satnoianu, 2003). The FDS mechanism is a robust patterning process that can explain somitogenesis in vertebrates (Kaern et al., 2001).

Reaction-diffusion-advection equations have been also used in astrophysics, geochemistry, fluid dynamics, and finance, for example. Other applications of RD systems arise in chemotaxis, modeling bacterial growth patterns, the process of wound healing, or the spread of cancer cells in healthy tissue to enumerate only a few such ideas in mathematical biology.

RAZVAN SATNOIANU

See also **Belousov–Zhabotinsky reaction; Brusselator; Cardiac arrhythmias and electrocardiogram; Chemical kinetics; Diffusion; FitzHugh–Nagumo equation; Heat conduction; Morphogenesis, biological; Pattern formation; Population dynamics; Scroll waves; Spiral waves; Turing patterns; Zeldovich–Frank-Kamenetsky equation; Vortex dynamics in excitable media**

Further Reading

Aronson, D.G. & Weinberger, H.F. 1975. Nonlinear diffusion in population genetics, combustion and nerve pulse propagation. In *Partial Differential Equations and Related Topics*, edited by J.A. Goldstein, New York: Springer

Faraday, M. 1861. *A Course of Six Lectures on the Chemical History of a Candle*. Reprinted as *Faraday's Chemical History of a Candle*, Chicago: Chicago Review Press, 1988

Field, R.J. & Noyes, R.M. 1974. Oscillations in chemical systems, IV. Limit cycle behaviour in a model of a real chemical reaction. *Journal of Chemical Physics*, 60: 1877–1884

Fisher, R.A. 1937. The wave of advance of advantageous genes. *Annals of Eugenics*, 7: 353–369

FitzHugh, R. 1961. Impulse and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 1: 445–466

Grindrod, P. 1996. *The Theory and Applications of Reaction-Diffusion Equations*, Oxford: Clarendon Press

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117: 500–544

Kaern, M., Menzinger, M., Satnoianu, R.A. & Hunding, A. 2001. Chemical waves in open flows of active media: their relevance to axial segmentation in biology. *Faraday Discussions*, 120: 295–312

Kauffman, S.A. 1995. *At Home in the Universe. The Search for the Laws of Self-organization and Complexity*, New York: Oxford University Press

Keener, J. & Sneyd, J. 2002. *Mathematical Physiology*, New York: Springer

Kermack, W.O. & McKendrick, A.G. 1932. Contributions to the mathematical theory of epidemics. *Proceedings of the Royal Society, London*, A 115: 700–721

Kolmogoroff, A., Petrovsky, I. & Piscounoff, N. 1937. Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Moscow University Bulletin Mathematics*, 1: 1–25

Lotka, A.J. 1925. *Elements of Physical Biology*, Baltimore: Williams and Wilkins

Meinhardt, H. 1998. *The Algorithmic Beauty of Seashells*, Berlin: Springer

Murray, J.D. 1988. How the leopard gets its spots. *Scientific American*, 258: 62

Murray, J.D. 2002, 2003. *Mathematical Biology*, 3rd edition, vols. 1 and 2, Berlin and New York: Springer

Øksendal, B. 2003. *Stochastic Differential Equations*, 6th edition, Berlin and New York: Springer

Okubo, A. & Levin, S.A. 2001. *Diffusion and Ecological Problems: Modern Perspectives*, Berlin and New York: Springer

Satnoianu, R.A. 2003. Coexistence of stationary and traveling waves in reaction-diffusion-advection systems. *Physical Review* E, 68: 032101

Satnoianu, R.A., Maini, P.K. & Menzinger, M. 2001. Parameter space analysis, pattern sensitivity and model comparison for Turing and stationary flow and diffusion distributed structures (FDS). *Physica* D, 160: 79–102

Scott, A. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structure*, 2nd edition, Oxford and New York: Oxford University Press

Scott, S.K., Wang, J. & Showalter, K. 1997. Modeling studies of spiral waves and target patterns in premixed flames. *Journal Chemical Society of Faraday Transactions*, 93: 1733–1739

Smolin, L. 1996. Galactic disks as reaction-diffusion systems. eprint arXiv.org astro-ph/9612033, 22 pages

Turing, A. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London* B, 327: 37–72

Volpert, A.I. & Volpert, V.A. 1994. *Travelling Wave Solutions of Parabolic Systems*. Translations of mathematical monographs, vol. 140, Providence, RI: American Mathematical Society

Volterra, V. 1926. Variations and fluctuations of a number of individuals in animal species living together. In *Animal Ecology*, New York: McGraw Hill, pp. 409-448, 1931. Translation by R.N. Chapman

Zeldovich, Ya.B. & Frank-Kamenetsky, D.A. 1938. K teorii ravnomernogo rasprostranenia plameni [toward a theory of uniformly propagating flames]. *Doklady Akademii Nauk SSSR*, 19: 693–697

Zhabotinsky, A.M. 1991. A history of chemical oscillations and waves, *Chaos*, 1: 379–386

RECURRENCE

Recurrence means repetition, and this, in turn, implies the presence of two actors: the particular event that is coming back after having occurred in the past and the law prescribing how events unfold in time.

The crossing of the vertical position by a well-serviced pendulum is a recurrent event. Here the event coincides with a particular value of the position coordinate, and the intervening law is Newton's second law in the presence of gravity. We deal here with the simplest version of recurrence, namely, strict periodicity. But one could also say that in a capitalistic economy, economic growth or slow down are recurrent events. The event is now an attribute of a whole succession of outcomes interrupted by irregular periods where this attribute is absent, and the intervening laws are no longer directly reducible to nature's dispassionate fundamental interactions but involve, rather, competing human agents each one of whom wishes to maximize his profit. In a still different vein, the need for survival has made man aware of the repeated appearance (recurrence) of weather patterns associated with different winds and/or precipitation patterns. The Greeks constructed an octagonal structure, the Tower of Winds, which can still be visited in Athens. It depicts the wind from each cardinal point and comprises bas-relief figures representing the typical type of weather associated with such a wind. The event is now a lumped, coarse-grained characterization of the state of the atmosphere that came to be known later on as *Grosswetterlagen* or "atmospheric analogs," and the laws are those of fluid mechanics and thermodynamics as applied to a rotating frame.

In the physical sciences, the description usually adopted views the system of interest as a deterministic dynamical system. At first sight this entails that an event is to be associated with the point that the system occupies in phase space at a given time. But in this view exact recurrence of a state is impossible except for the trivial case of periodic dynamics, because it would

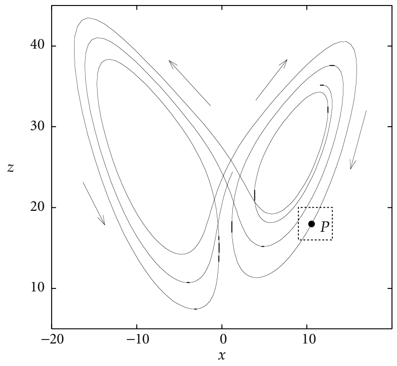

Figure 1. Illustrating recurrence on a two-dimensional projection of the Lorenz chaotic attractor. Recurrence is here reflected by the fact that the trajectory emanating from point P in the dotted square will re-enter this cell (which plays the role of a coarse-grained state) after some time.

violate the basic property of uniqueness of the solutions of the evolution equations. To cope with this limitation, one associates the concept of recurrence with repeated re-entries of the dynamical trajectory in a phase space region possessing finite volume (more technically, "measure") as indicated in Figure 1. This is, precisely, the coarse-grained state referred to earlier, as opposed to the point-like states considered in many (or even most) nonlinear science-related problems. Probabilistic description is a typical instance in which one deals essentially with coarse-grained states.

Poincaré Recurrence

In his celebrated memoir on the three-body problem Henri Poincaré, a founding father of nonlinear science, established a result that may be considered as the first quantitative and rigorous statement concerning recurrence (Poincaré, 1890). He showed that in a system of point particles under the influence of forces depending only on the spatial coordinates, a typical phase space trajectory will visit infinitely often the neighborhood, however small, of a prescribed initial state, provided that the system remains in a bounded region of phase space. The time between two such consecutive visits is referred as a "Poincaré recurrence time."

In modern terms (Kac, 1959), let A be a subset of space Ω within which the system is confined such that $\mu(A) > 0$ where μ designates the measure of A, and T_t a family of one-to-one measure preserving transformations in Ω. Then, for almost every initial point $P \in A$ (i.e., except for a set of Ps of zero μ-measure), there exists a sufficiently large t such that $T_t P \in A$. Using the machinery of the proof, one may derive an expression for the mean recurrence time,

$$\langle\theta_\tau\rangle = \tau/\mu(A), \tag{1}$$

where τ is the time resolution between the successive observations. Actually, as observed by Marian Smoluchowski, this expression needs to be adapted, in the limit of continuous time, to

$$\begin{aligned}\langle\theta_\tau\rangle = {} & \tau(1-\mu(A))/(\mu(A) \\ & -\mu(P \in A, T_\tau P \in A)),\end{aligned} \tag{2}$$

where one has introduced the measure of the intersection of A and its pre-image $T_\tau^{-1}A$.

Part of the historical role of Poincaré's theorem relates to the famous *Wiederkehreinwand*, an early objection by Ernst Zermelo challenging Ludwig Boltzmann's derivation of the second law of thermodynamics on the grounds that since states recur, the H-function (the microscopic analog of entropy) is bound to reverse its course at some stage. Boltzmann made a valiant attempt to respond, based essentially on the enormity of recurrence times in a molecular system (of the order of ten to the power of the Avogadro number). A more pertinent observation is that thermodynamic states are associated with a Gibbs ensemble rather than with single phase space points. Poincaré's theorem does not apply under these conditions; hence, Poincaré recurrences are not necessarily precluding a microscopic interpretation of irreversibility.

From Ehrenfest to Fermi–Pasta–Ulam

A first implementation of the idea that it is indeed possible to reconcile recurrence with irreversible behavior was provided by the, by now famous, "dog flea" model of Paul and Tatiana Ehrenfest. $2N$ numbered balls are distributed in boxes A and B. An integer from 1 to $2N$ is drawn at random and the ball corresponding to that number is moved from the box in which it is found to the other one. The process is then repeated to any desired number of times. One can show (Kac, 1959) that every initial state recurs with probability one, although the system possesses an H-function. Actually, this model belongs to a general class of stochastic processes known as Markov chains, whose theory was developed intensely in the first part of the 20th century (Feller, 1968, 1971). These developments allow one to put the compatibility between recurrence of states and irreversible approach to an asymptotic regime on a firm basis for this very general class of processes amenable to a Markov chain-like description. An important insight is that while recurrence is a property of trajectories (which being now stochastic and thus coarse-grained are not bound to satisfy the uniqueness theorem stated above and can therefore recur as such), the approach to an asymptotic regime is a property of probability distributions. The theory of stochastic processes has also provided a wealth of further information on recurrence independent of the problem of irreversibility, including connections with the important class of renewal processes.

Despite their power, the above results overlooked the nature of the underlying deterministic dynamics, except that the last had to be complex enough to generate somehow a stochastic game. This shortcoming became clearly recognized in the 1950s, when Poincaré's recurrence was viewed again as a problem of strictly deterministic dynamics, thanks partly to the new possibilities afforded by the availability of electronic computers. From the standpoint of nonlinear science, the most relevant among these early attempts is that by Fermi, Pasta, and Ulam (Fermi et al., 1955) who numerically examined a set of coupled differential equations modeling the motion of a linear chain of equimass particles connected by nonlinear springs

$$\begin{aligned}\frac{d^2x_j}{dt^2} &= \{1+\beta[(x_{j+1}-x_j)^2+(x_j-x_{j-1})^2\\ &\quad +(x_{j+1}-x_j)(x_j-x_{j-1})]\}\\ &\quad \times(x_{j+1}+x_{j-1}-2x_j),\\ j &= 1,2,\ldots,N-1, x_0=x_N=0. \qquad (3)\end{aligned}$$

Fermi et al. showed that with 64 particles at low energy there was no equipartition of energy among the oscillators (contrary to their expectations). Instead, a beat phenomenon was observed with ongoing recurrences of initial conditions. Recurrence seemed once again to be at odds with the tendency toward thermal equilibrium, to which is associated the stronger property of mixing and equipartition. A closer analysis shows that Fermi–Pasta–Ulam recurrence is not universal: for higher energies a phenomenon of overlap of resonances is taking place, leading to stochasticity in the sense that energy transfer between modes becomes possible. To what extent this stochasticity can invade the entire phase space as N and the energy are increased is still an unsettled question.

Modern Nonlinear Dynamics and Recurrence

The most significant advance on recurrence since the 1960s has been to dissociate the concept from the approach to equilibrium of a many-body system (and, in particular, to realize that the two concepts are not to be opposed), to extend it to cover the class of dissipative dynamical systems, and to establish a link between the deterministic and the stochastic approaches to it.

What made this advance possible is the realization, at the heart of modern nonlinear dynamics, that simple-looking, low-dimensional models may give rise to complex behavior that emulates, to a large extent, the behavior of real world, multivariate systems. Furthermore, in the presence of deterministic chaos, sensitivity to initial conditions raises the problem of long-term prediction and prompts one to adopt a probabilistic approach, perfectly compatible with (and actually induced by) the underlying dynamics, in which quantities of interest are expressed as statistical averages. In this setting, recurrence became a powerful tool providing new insights along two directions.

Recurrence as an Indicator of the Dynamical Complexity

The main point here is that dynamical systems having different ergodic properties show different recurrence patterns. An interesting example is uniform quasi-periodic motion on a two-dimensional torus or its equivalent discrete-time version given by the shift map. It can be shown (Sös, 1958; Theunissen et al., 1999) that, typically, such a system (known to be ergodic) possesses exactly three possible values of recurrence times in any prescribed phase space cell. On the other hand, in a wide class of uniformly hyperbolic systems such as one-dimensional maps, one shows that recurrence times in a sufficiently small phase space cell are generically exponentially distributed, while a sequence of successive recurrences has a Poissonian limit distribution (Collet, 1991). The situation is different in non-uniformly hyperbolic systems. In particular, in systems showing intermittent behavior the above laws are replaced, respectively, by power laws and by Lévy stable distributions (Balakrishnan et al., 1997). Work along similar lines has also led to connections between recurrence patterns and Lyapunov exponents, unstable periodic orbits, or generalized dimensions. There is, however, no firm indication about the universality of these latter results.

Recurrence as a Tool for Prediction

The question whether, and if so how frequently, the future states of a system will come close to a certain initial configuration has a direct bearing on prediction, an issue of central importance in atmospheric sciences, hydrology, and other environment-related disciplines. A dynamical approach to the recurrence of atmospheric analogs, the lumped states of the atmosphere introduced above reveals a strong dependence of recurrence times on the local properties of the attractor and a pronounced variability around their mean (Nicolis, 1998). Of crucial interest is also the recurrence of extreme events such as natural disasters, a problem approached traditionally at the level of a statistical description (Gumbel, 1958). The extension of the statistical theory of extremes and their recurrences to deterministic dynamical systems is still in its infancy.

G. NICOLIS AND C. ROUVAS-NICOLIS

See also **Dynamical systems; Fermi–Pasta–Ulam oscillator chain; Intermittancy; Nonequilibrium statistical mechanics**

Further Reading

Balakrishnan, V., Nicolis, G. & Nicolis, C. 1997. Recurrence time statistics in chaotic dynamics I. Discrete time maps. *Journal of Statistical Physics*, 86: 191–212

Collet, P. 1991. Some ergodic properties of maps of the interval, Lectures given at the CIMPA summer school *Dynamical Systems and Frustrated Systems*, edited by J.M. Gambaudo, Temuco

Feller, W. 1968, 1971. *An Introduction to Probability Theory and Its Applications*, 2 vols, 3rd edition, New York: Wiley

Fermi, E., Pasta, J.R. & Ulam, S.M. 1955. Studies of nonlinear problems. Los Alamos Scientific Laboratory Report No. LA-1940

Gumbel, E.J. 1958. *Statistics of Extremes*, New York: Columbia University Press

Kac, M. 1959. *Probability and Related Topics in Physical Sciences*, London and New York:

Nicolis, C. 1998. Atmospheric analogs and recurrence time statistics: toward a dynamical formulation. *Journal of Atmospheric Sciences*, 55: 465–475

Poincaré, H. 1890. Sur le problème des trois corps et les équations de la dynamique. *Acta Matematica*, 13: 1–270

Sös, V.T. 1958. On the distribution mod 1 of the sequence nα. *Annales Universitatis Scientiarum Budapestinensts de Rolando Eotovos Nominatae Sectio Mathematica*, 1: 127–134

Theunissen, M., Nicolis, C. & Nicolis, G. 1999. Recurrence times in quasi-periodic motion: statistical properties, role of cell size, parameter dependence. *Journal of Statistical Physics* 94: 437–467

REGULAR AND CHAOTIC DYNAMICS IN ATOMIC PHYSICS

Many important areas of atomic physics dwell in the intriguing world of semiclassical quantum mechanics, where classical concepts such as orbits and phase space are used to calculate purely quantal quantities such as quantum numbers, wave functions, and ionization thresholds. The past two decades have witnessed renewed interest in classical interpretations of quantum phenomena (Berry & Mount, 1988; Casati et al., 1987; Blümel & Reinhardt, 1997; Jensen, 1992), a view sometimes referred to as postmodern quantum mechanics (Heller & Tomsovic, 1993). This development has led to the much debated term *quantum chaos* (*See* **Quantum chaos**). A central issue is the relevance of chaotic dynamics to the quantum world, where the sensitivity to initial conditions characteristic of classical chaos is obscured by the Heisenberg uncertainty principle. Moreover, the Schrödinger equation is linear and does not normally exhibit extreme sensitivity to initial conditions. Nevertheless, classical dynamics plays an important role in the quantum world, even in chaotic regions. Thus, the ionization of hydrogen by a microwave field is often well described in terms of chaotic diffusion of classical electron orbits, while the quantum mechanical wave function exhibits traces of classical (unstable) periodic orbits, called scarring.

When the classical dynamics are regular, semiclassical quantum numbers can be calculated according to the Einstein–Brillouin–Keller (EBK) prescription (Gutzwiller, 1990),

$$J_i = \oint p_i \mathrm{d}q_i = (n_i + \alpha_i/4)h, \qquad (1)$$

where J_i is the classical action, calculated from the loop integral over coordinate q_i and canonically conjugate momentum p_i, n_i is the corresponding quantum number, h is Planck's constant, and α_i is the Maslov index, an integer that depends on the topology of the invariant torus. In this way, approximate semiclassical wave functions can also be constructed. The semiclassical description is found to work well under two conditions: (i) the quantum numbers are large, and (ii) the number of participating states is also large. Under chaotic conditions (when the classical action does not generally exist), one still has recourse to periodic orbits (Eckhardt, 1988), which can be used to organize energy levels and the form of the wave function (Gutzwiller, 1990). The principal tool under chaotic conditions is the Gutzwiller trace formula, based on the Feynman propagator.

Our unifying theme is the hydrogen atom in an electromagnetic field, for which the motion of the single electron is well described by a Hamiltonian model. We consider first two closely related integrable cases (*See* **Constants of motion and conservation laws; Integrability**), the classical problem of two fixed centers as a model for the hydrogen molecule ion H_2^+, and the H-atom in a homogeneous electrostatic field. In both cases axial symmetry implies the constancy of the azimuthal angular momentum, $p_\phi = m\rho^2\dot{\phi}$ (in cylindrical coordinates $\rho,\ \phi,\ z$), allowing a reduction in dimensionality from three to two, via an effective potential formed by including the azimuthal part of the kinetic energy with the potential $U(\rho, z)$:

$$U^e(\rho, z) = U + \frac{p_\phi^2}{2m\rho^2}\,. \qquad (2)$$

Thus, in both cases the motion is described by a two degree of freedom autonomous (time-independent) Hamiltonian.

Next we take up the non-integrable case of the H-atom in a homogeneous magnetostatic field, which has been extensively studied, both theoretically and experimentally (Born, 1960; Friedrich & Wintgen, 1989), and which is also described by a two-degree-of-freedom Hamiltonian. Then we describe fruitful experiments on microwave ionization of H-atoms, which have yielded an abundance of information on quantum physics (Blümel & Reinhardt, 1997; Koch & van Leeuwen, 1995). Various Hamiltonian models have been investigated, from one to three degrees of freedom. Here, we shall limit ourselves to a one-dimensional time-periodic model, which captures most of the relevant physics. Finally, we briefly mention other atomic systems in which a semiclassical treatment sheds light on quantal behavior.

Two Fixed Centers

The two fixed centers model (TC) consists of a test mass orbiting in the gravitational field of two massive bodies rotating about their center of mass at a fixed distance. In atomic physics the attractive force is electrostatic, furnishing a useful model for the hydrogen molecule-ion (Born, 1960; Strand & Reinhardt, 1979; Howard & Wilkerson, 1995). It also enjoys the distinction of being one of the small minority of completely integrable systems, owing to separability in confocal elliptic coordinates. With the two protons fixed at $z = \pm a$, the nonrelativistic motion of the electron is described by the Hamiltonian,

$$H = \frac{1}{2m}(p_\rho^2 + p_z^2) + U^e \tag{3}$$

with effective potential

$$U^e(\rho, z) = \frac{p_\phi^2}{2m\rho^2} - q^2\left(\frac{1}{r_1} + \frac{1}{r_2}\right), \tag{4}$$

where q is the electronic charge,

$$r_1^2 = \rho^2 + (z-a)^2, \quad r_2^2 = \rho^2 + (z+a)^2, \tag{5}$$

and $p_\rho = m\dot{\rho}$, $p_z = m\dot{z}$, and $p_\phi = m\rho^2\dot{\phi}$. All trapped orbits are confined by the two-dimensional effective potential, shown in Figure 1 for two nearby values of the control parameter $\mu = p_\phi^2/a^2$. In the first case, a single well centered on an equatorial critical point of U^e exists, which in the second case has bifurcated into a double well. Quantizable stable circular orbits exist at the elliptic critical points of U^e. Because there is no chaos in this system, it might seem straightforward to construct quantum numbers from the two conserved actions, taking care to get the correct Maslov indices. However, there is a rub: there are two disjoint classes of orbits, a situation referred to as "monodromy," resulting in separate actions, which need to be smoothly joined. This process of "uniform quantization" has been carried out by Strand & Reinhardt (1979), who obtained greatly improved values for energy levels for varying intermolecular separation $R = 2a$.

Hydrogen Atom in a Strong Electric Field

The behavior of the H-atom in a homogeneous electrostatic field is one of the oldest problems of quantum mechanics, the weak-field case giving rise to the Stark effect, traditionally treated using perturbation theory (Born, 1960). For larger fields, ionization occurs (Koch, 1978; Howard, 1995), the ionization threshold field F_c depending on the particle energy as well as a quantity called the Runge–Lenz invariant.

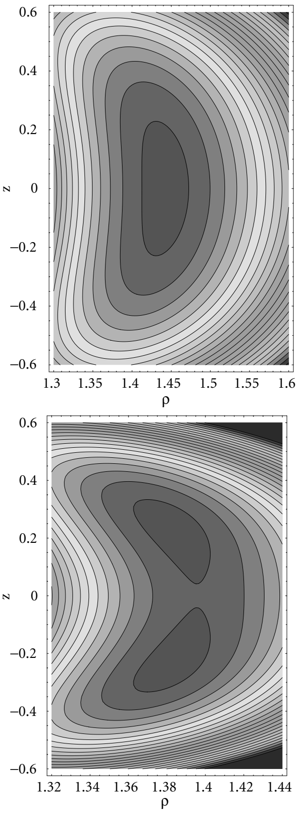

Figure 1. Contour plot of effective potential for the hydrogen molecule-ion before and after pitchfork bifurcation.

This is an interesting problem for two reasons: (i) the required fields are too large for perturbative methods, and (ii) the motion is completely integrable, since the Hamiltonian and the Schrödinger equations both separate in parabolic coordinates. Spectral lines are then conveniently labeled by quantum numbers n, n_1, n_2, and $|m|$, where $n = n_1 + n_2 + |m| + 1$ is the principal quantum number, m is the azimuthal quantum number, and n_1 and n_2 are parabolic quantum numbers.

Field ionization may be treated either classically or via quantum mechanical tunneling (Gallas et al., 1982). With the electric field along the z-axis, the nonrelativistic motion of the electron is again described by Hamiltonian (3), with effective potential

$$U^e(\rho, z) = -\frac{q^2}{\sqrt{\rho^2 + z^2}} + \frac{p_\phi^2}{2m\rho^2} - qFz\,, \tag{6}$$

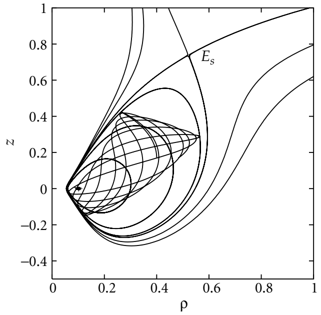

Figure 2. Electron orbit with energy $E < E_{\mathrm{S}}$ trapped in effective potential for the hydrogen atom immersed in a homogeneous electric field.

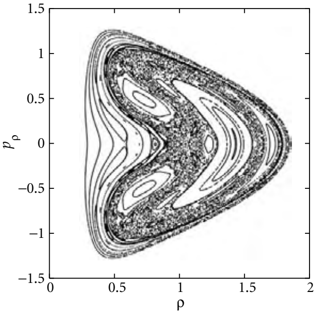

Figure 3. Poincaré section for H-atom in a uniform magnetostatic field.

where p_ϕ gives the azimuthal quantum number $m = p_\phi/\hbar$ and F is the electric field strength. Figure 2 shows an orbit trapped in the two-dimensional potential well formed by the effective potential. All orbits having $E < E_{\mathrm{S}}$, the saddle point energy, are classically trapped. If the system were non-integrable, the condition $E > E_{\mathrm{S}}$ would constitute a necessary and sufficient condition for ionization. However, the existence of the Runge–Lenz invariant allows a subclass of trapped orbits with $E > E_{\mathrm{S}}$. Owing to separability in parabolic coordinates (ξ, η), the actions J_ξ, J_η, and $J_\phi = p_\phi$ always exist so that the EBK formula applies; there is no classical chaos in the Stark problem. In laboratory experiments the electric field is usually gradually increased from zero until ionization occurs. For $m = 0$, the saddle point criterion leads to the ionization threshold $F_{\mathrm{c}} \approx 1/9n^4$, in excellent agreement with experiment.

Hydrogen Atom in a Strong Magnetic field

In contrast to the Stark problem, the electron motion in an H-atom immersed in a homogeneous magnetostatic field can be chaotic, providing a useful testing ground for ideas of quantum chaos (Friedrich & Wintgen, 1989). There are two integrable limits, (i) $B = 0$, for which the orbits are simple Kepler ellipses, and (ii) $B \to \infty$, the helical gyration of a free charged particle in a constant magnetic field. With $\boldsymbol{B} = B_0\hat{\boldsymbol{z}}$, the nonrelativistic motion of a single electron is governed by the effective potential

$$U^e(\rho, z) = \frac{p_\phi^2}{2m\rho^2} - \frac{q^2}{\sqrt{\rho^2 + z^2}} + \frac{1}{8}m\omega_c^2\rho^2 + \frac{1}{2}\omega_c p_\phi, \qquad (7)$$

where $\omega_c = qB_0\,/\,mc$ is the cyclotron frequency and now $p_\phi = m\rho^2\dot{\phi} + \frac{1}{2}\rho^2\omega_c$. The constant paramagnetic term may be removed by transforming to a rotating frame. For small magnetic field strength, perturbation theory may be used to calculate energy levels (Zeeman effect); for very large fields, the diamagnetic term proportional to ρ^2 dominates (quadratic Zeeman effect) and the spectrum resembles that of a free electron. In this case, the existence of chaotic trajectories means that a set of classical actions does not always exist, precluding EBK quantization. For small fields, however, canonical perturbation theory can be employed to calculate perturbed actions. This yields useful energy level curves $E_n = E_n(B)$, where the levels are labeled in terms of the unperturbed $(B = 0)$ states. Furthermore, at low fields an approximate constant of the motion exists, facilitating classification of the very complex spectrum and explaining avoided crossings of nearby energy level curves. This recipe fails for larger fields, where the classical motion is mainly chaotic. At very large fields, intriguing regularities persist, called quasi-Landau resonances, close to the spectrum of a free electron in a magnetostatic field. In this case, it is possible to derive a very simple formula for the ionization threshold energy.

The degree of classical chaos is conveniently revealed by the Poincaré section, which is a two-dimensional slice in the four-dimensional phase space, defined by fixing the total energy and recording unidirectional intersections with a coordinate plane (Lichtenberg & Lieberman, 1990). Figure 3 shows a section for a moderately large B-field, generated by calculating particle orbits and recording $(\rho,\ p_\rho)$ each time an orbit crosses the $z = 0$ plane with $p_z > 0$. Regular orbits appear as smooth closed curves, whose centers represent periodic orbits. A single period-one orbit appears near $\rho = 1.45$; all other fixed points are at least period-three. The scattered dots in between represent chaotic orbits (possibly just one), which ergodically fill a connected region of phase space. In this case, classical actions do not exist, and it is not possible to assign quantum numbers to the very complex spectrum.

Microwave Ionization of Rydberg Atoms

Another fruitful source of information on connections between classical dynamics and quantal behavior is the extensive experimental program on microwave ionization of highly excited Rydberg (hydrogenic) atoms (Blümel & Reinhardt, 1997; Koch & van Leeuwen, 1995). By careful control of experimental conditions, it has been found possible to produce a full range of principal quantum numbers, $n_0 = 34$ to more than 90. Ionization from such excited states can then be readily accomplished using a variety of pulsed or continuous wave sources; linearly polarized, circularly polarized, or elliptically polarized. Remarkably, classical scaling accurately describes much of the experimental results. Thus, the microwave frequency is expressed as a ratio to the Kepler frequency, which according to the Bohr–Sommerfeld model is proportional to n_0^{-3}, giving the dimensionless parameter $\hat{\omega} = n_0^3\omega$. The electric field strength may also be expressed as a ratio to the Coulomb force, giving the dimensionless parameter $\hat{F} = n_0^4 F$. Koch distinguishes several distinct frequency regimes, depending on $\hat{\omega} = n_0^3\omega$; by fixing $\omega/2\pi$ at one of several frequencies between 7 and 36 GHz and varying n_0, a range of scaled frequencies from $\hat{\omega} = 0.02$ to 2.8 can be achieved. For very low $\hat{\omega} \lesssim 0.02$ quantum tunneling dominates, significantly lowering the ionization threshold $\hat{F}_i$. At somewhat higher $\hat{\omega}$ the ionization curves (fraction ionized vs. $\hat{F}$) are primarily classical, with occasional "bumps" deriving from quantum resonances. For $0.1 \lesssim \hat{\omega} \lesssim 1.2$ classical dynamics prevails. Above this value quantum mechanisms increasingly raise the ionization thresholds, by about a factor of two at $\hat{\omega} = 2.8$. As for the Zeeman effect, scarring of the wave function occurs along classical unstable period orbits.

Theoretical and numerical analyses range from crude one-dimensional (1-d) models to elaborate 3-d Monte Carlo simulations. The simplest case to analyze theoretically is linear polarization, in which a simple one-dimensional time-periodic Hamiltonian explains much of the relevant physics,

$$H = H_0 + qFx\sin(\omega t), \tag{8}$$

where

$$H_0 = \begin{cases} \dfrac{p^2}{2m} - \dfrac{q^2}{x}, & x > 0, \\ \infty, & x \le 0. \end{cases} \tag{9}$$

Physically this represents the limiting case of a pencil-thin orbit along the x-axis. In order to compare with experiment, we first transform to action-angle variables $J = \sqrt{a} = \sqrt{-1/2H_0}$, $\theta = u - \sin u$, where a is the semimajor axis and the eccentric anomaly u is defined by $x = a(1 - \cos u)$. The principal quantum number is then given by $J = n_0\hbar$. Poincaré sections can be generated by strobing the orbit at a period $\tau = 2\pi/\omega$ and recording J and θ at $t = \tau, 2\tau, \ldots$. In experiments the microwave field is gradually increased until ionization takes place. For example, for $f_{\mathrm{RF}} = 9.923$ GHz, we obtain the two sections near the $n_0 = 51$ state shown in Figure 4 for $F = 62$ and 66.4 V/cm. At the lower field strength all orbits initialized with $n_0 < 51$ are trapped by invariant circles; at the higher field strength these curves have been destroyed, allowing most of the same orbits to escape. The experimental result for 10% ionization of the $n_0 = 51$ state is about 71 V/cm (Koch & van Leeuwen, 1995), in reasonable agreement with our numerical calculation. Other polarizations have also been investigated (Koch, 1998; Howard, 1992). The most general state is elliptic polarization, which includes circular and linear polarization as special cases. In all cases valuable insight into experiment has been obtained using classical dynamics.

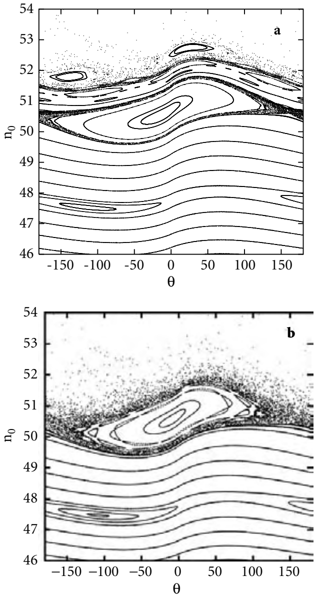

Figure 4. Poincaré section for 1-d H-atom perturbed by a time-periodic microwave field of frequency $f_{\mathrm{RF}} = 9.923$ GHz and strength (a) $F = 62$ V/cm and (b) $F = 66.4$ V/cm.

Discussion

In addition to the above four examples, there are several other systems under active investigation. Combinations of fields, particularly crossed and parallel fields, $\boldsymbol{E} \times \boldsymbol{B}$ and $\boldsymbol{E} \parallel \boldsymbol{B}$, offer challenging structures that can be partially explained by semiclassical quantization (Farrelly, 1991). Three-body systems, such as the helium atom (Tanner et al., 2000) and the hydrogen negative ion H^-, long considered beyond semiclassical methods, have recently yielded important insights.

JAMES E. HOWARD

See also **Constants of motion and conservation laws; Hamiltonian systems; Quantum chaos**

Further Reading

Berry, M.V. & Mount, K.E. 1972. Semiclassical wave mechanics. *Reports on Progress in Physics*, 35: 315–800

Blümel, R. & Reinhardt, W.P. 1997. *Chaos in Atomic Physics*, Cambridge and New York: Cambridge University Press

Born, M. 1960. *The Mechanics of the Atom*, New York: Unger

Casati, B., Chirikov, B.V., Shepelyansky, D.L. & Guarneri, I. 1987. Relevance of classical chaos in quantum-mechanics—the hydrogen atom in a monochromatic field. *Physics Reports*, 154: 77–123

Eckhardt, B. 1988. Quantum mechanics of classically non-integrable systems. *Physics Reports*, 163: 205–297

Farrelly, D. 1991. Semiclassical mechanics of bounded and unbounded states of atoms and molecules, *Advances in Molecular Vibrations*, 1B, 49–79

Friedrich, H. & Wintgen, D. 1989. The hydrogen atom in a uniform magnetic field—an example of chaos. *Physics Reports*, 183: 37–79

Gallas, J.A.C., Walther, H. & Werner, E. 1982. Simple formula for the ionization rate of Rydberg states in static electric fields. *Physical Review Letters*, 49: 867–891

Gutzwiller, M.C. 1990. *Chaos in Classical and Quantum Mechanics*, New York: Springer

Heller, E.J. & Tomsovic, S. 1993. Postmodern quantum mechanics. *Physics Today*, 46: 38–46

Howard, J.E. 1992. Stochastic ionization of hydrogen atoms in a circularly polarized microwave field. *Physical Review* A, 46: 364–372

Howard, J.E. 1995. Saddle point ionization and the Runge-Lenz invariant. *Physical Review* A, 51: 3934–3946

Howard, J.E. & Wilkerson, T.W. 1995. Problem of two fixed centers and a finite dipole: a unified treatment. *Physical Review A*, 52: 4471–4492

Jensen, R.V. 1992. Quantum chaos. *Nature*, 352: 311–315

Koch, P.M. 1978. Resonant states in the nonperturbative regime: the hydrogen atom in an intense electric field, *Physical Review Letters*, 41: 99–103

Koch, P.M. 1998. Polarization dependence of microwave "ionization" of excited hydrogen atoms. *Acta Physica Polonica*, 93: 105–132

Koch, P.M. & van Leeuwen, K.A.H. 1995. The importance of resonances in microwave "ionization" of excited hydrogen atoms. *Physics Reports*, 255: 289–403

Lichtenberg, A.J. & Lieberman, M.A. 1990. *Regular and Chaotic Dynamics*, 2nd edition, New York: Springer

Strand, M.P. & Reinhardt, W.P. 1979. Semi-classical quantization of the low-lying electronic states of H_2^+. *Journal of Chemical Physics*, 70: 3812–3827

Tanner, G., Richter, K. & Rost, J. 2000. The theory of two-electron atoms: between ground state and complete fragmentation. *Reviews of Modern Physics*, 72: 497–544

RELAXATION OSCILLATORS

The first work related to relaxation oscillators was published by Balthasar van der Pol in 1926 (Van der Pol, 1926). Van der Pol, in collaboration with J. van der Mark, suggested an electrical model of the heart consisting of three relaxation generators (van der Pol & van der Mark, 1928).

An oscillator is said to be of the relaxation type if its period is inversely proportional to a relaxation time and nearly independent of other parameters. Classical examples of relaxation oscillators are a thyratron oscillator (Teodorchik, 1952) and an alternating water source known as Tantal's vessel (Panovko & Gubanova, 1964; Strelkov, 1964). Schematic images of these devices are illustrated in Figures 1a and b. The shape of oscillations of the amount of water in Tantal's vessel and of the voltage across the capacitor C is shown in Figure 1c.

An oscillator described by the Rayleigh equation (Rayleigh, 1877–78)

$$\ddot{x} - \mu(1 - \dot{x}^2)\dot{x} + x = 0 \tag{1}$$

is also of relaxation type for $\mu \gg 1$. It should be noted that the Rayleigh equation can be obtained from the van der Pol equation

$$\ddot{y} - \mu(1 - y^2)\dot{y} + y = 0 \tag{2}$$

by the substitution $y = \sqrt{3}\,\dot{x}$.

The phase portrait and the shape of oscillations for Equation (1) are illustrated in Figure 2 for $\mu = 10$. For this value of μ, the equilibrium state is an unstable node and the limit cycle has a shape close to a parallelogram. The time required for the representative point to approach the limit cycle is very short; thus the oscillations are almost discontinuous. For $\mu \ll 1$, the oscillation period is practically independent of μ and approximately equal to the period of free oscillations of the oscillatory circuit $T_0 = 2\pi$. For $\mu \gg 1$, on the other hand, the oscillation period is completely determined by the value of the parameter μ, namely $T \approx 8\mu/3\sqrt{3}$ (see below). As the relaxation time is of the order of $1/\mu$, it follows that for $\mu \gg 1$ the oscillation period is really inversely proportional to the relaxation time.

Introducing the small parameter $\varepsilon = 1/\mu$, we can rewrite Equation (1) in the form of two equations of

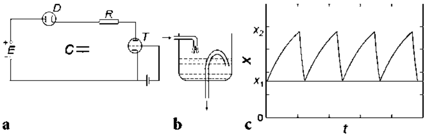

Figure 1. Schematic images (a) of a thyratron oscillator, (b) of Tantal's vessel, and (c) the shape of oscillations of the amount of water in Tantal's vessel and of the voltage across the capacitor C.

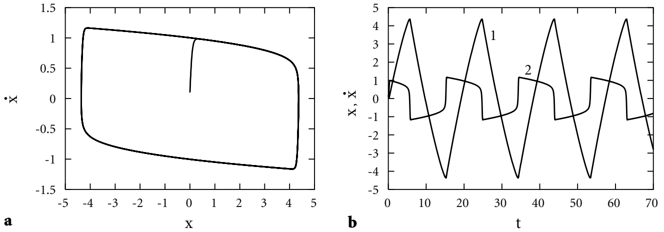

Figure 2. (a) The phase portrait for the Rayleigh equation for $\mu = 10$, and (b) the solution of this equation for $\mu = 10$: curves 1 and 2 show $x(t)$ and $\dot{x}(t)$, respectively.

first order as

$$\dot{x} = y, \tag{3}$$

$$\varepsilon\dot{y} = (1 - y^2)y - \varepsilon x. \tag{4}$$

To find an approximate solution of Equations (3) and (4) we apply the averaging method in systems incorporating fast and slow variables. It follows that the variable x is slow and the variable y is fast. Consequently, in solving Equation (4) the variable x can be considered as a constant. Stationary solutions of Equation (4) under this condition are real roots of the cubic equation

$$y^3 - y + \varepsilon x = 0. \tag{5}$$

Over the range

$$-\frac{2}{3\mu\sqrt{3}} < x < \frac{2}{3\mu\sqrt{3}} \tag{6}$$

Equation (5) has three real roots, whereas outside this range it has only one real root. For $x = \pm 2/(3\mu\sqrt{3})$ Equation (5) has two real roots, $y_1 = \pm 1/\sqrt{3}$, $y2 = \mp 2/\sqrt{3}$. Approximate solutions of Equation (5) can be found analytically under the condition $\varepsilon x \ll 1$. Expanding y sequentially about $y_{10} = 1$, $y_{20} = -1$ and $y_{30} = 0$ we find three solutions:

$$y_1 = 1 - \frac{\mu}{2}x, \quad y_2 = -1 - \frac{\mu}{2}x, \quad y_3 = \mu x. \tag{7}$$

It is seen that the approximate solutions (7) are close to the exact ones over the range (6).

Investigation of the stability of these shows that only those solutions are stable that satisfy the constraint $y_i \geq 1/\sqrt{3}$, which is fulfilled for the upper and the lower parts of the dependence $y(x)$ and is not fulfilled for its middle part.

Substituting $y_{1,2}$ from (7) into Equation (3) and integrating the latter, we find solutions of this equation under the initial condition $x(0) = x_0$:

$$x_{1,2}(t) = x_0 e^{-\varepsilon t/2} \pm \frac{2}{\varepsilon}\left(1 - e^{-\varepsilon t/2}\right). \tag{8}$$

For $\varepsilon t \ll 1$ solutions (8) can be simplified as

$$x_{1,2}(t) = (x_0 \pm t)(1 - \varepsilon t/2). \tag{9}$$

Let the initial conditions be $x_0 = -2/(3\varepsilon\sqrt{3})$, $y_0 = y_1(x_0)$. Then, as follows from (9), x will increase with an approximately constant rate, and at the instant $t = 4/(3\varepsilon\sqrt{3})$, it will attain the value $2/(3\varepsilon\sqrt{3})$. At this instant, the solution $y_1(x)$ becomes unstable because it merges with the unstable solution $y_3(x)$. As a result, y will jump to the value $y_2\left(2/(3\varepsilon\sqrt{3})\right)$. Subsequently, x will decrease to the value $-2/(3\varepsilon\sqrt{3})$ and then a jump will occur again; that is, y will take on the value $y_1\left(-2/(3\varepsilon\sqrt{3})\right)$. As the process is repeated, we obtain a limit cycle on the phase plane x, y close to the one calculated numerically (Figure 2).

If $y_0 \neq y_{1,2}(x_0)$, y takes on one of these values very rapidly, practically for a constant x. This means that the corresponding phase trajectories are nearly vertical.

An approximate expression for the oscillation period T can be found on the assumption that the jumps occur instantly and x varies with the constant velocity ± 1. Thus

$$T \approx \frac{8}{3\varepsilon\sqrt{3}} = \frac{4}{3\pi\sqrt{3}}\mu T_0, \tag{10}$$

where $T_0 = 2\pi$ is the period of natural oscillations for $\mu = 0$. Thus, in the relaxation regime, the oscillation period is approximately $\mu/4$ times the period of natural oscillations.

Importantly, one of the characteristic features of relaxation oscillators is the ease of frequency synchronization (Landa, 2001). This phenomenon was apparently first studied by Teodorchik (1943, 1945). Relaxation oscillators have long been used in radio and electrical engineering, mainly as multivibrators, but recently interest in these oscillators has quickened in connection with biological and medical applications. As early as 1953, the German physicist Karl Bonhoeffer suggested that a neuron might be modeled by equations similar to the van der Pol equation for a relaxation oscillator. The Bonhoeffer–van der Pol equations describe oscillations of the voltage across a neural membrane, including refractoriness (difficulty of refiring). These equations are

$$\dot{x} = x - x^3/3 - y + I_0, \quad \dot{y} = \varepsilon(x + a - by), \quad (11)$$

where x is the voltage across the neural membrane; y is a quantity of refractoriness; a, b, ε are membrane radius, specific resistivity of the fluid inside the membrane, and temperature factor, respectively; and I_0 is a direct component of the current across the membrane.

POLINA LANDA

See also **FitzHugh–Nagumo equation; Neurons; Nonlinear electronics; Synchronization; Van der Pol equation**

Further Reading

Bonhoeffer, K.F. 1953. Modelle der Nervenerregung. *Naturwissenschaften*, 40: 301–311

Landa, P.S. 2001. *Regular and Chaotic Oscillations*, Berlin and New York: Springer

Panovko, Ya.G. & Gubanova, I.I. 1964. *Ustoichivost' i Kolebaniya Uprugikh Sistem [Stability and Oscillations of Elastic Systems]*, Moscow: Science

Rayleigh, Lord (Strutt, J.W.) 1877–78. *The Theory of Sound*, 2 vols., London: Macmillan; reprinted New York: Dover, 1877–78

Strelkov, S.P. 1964. *Vvedenie v Teoriyu Kolebanii [Introduction to the Oscillation Theory]*, Moscow: Science

Teodorchik, K.F. 1943. On the theory of synchronization of relaxation generators. *DAN SSSR*, 40: 63–66 (in Russian)

Teodorchik, K.F. 1945. Theory of synchronization of relaxation self-oscillatory systems. *Journal of Physics*, Moscow, 9: 139–146 (in Russian)

Teodorchik, K.F. 1952. *Avtokolebatel'nye Sistemy [Self-Oscillatory Systems]*, Moscow: Gostekhizdat

van der Pol, B. 1926. On relaxation oscillation. *Philosophical Magazine*, 2: 978–992

van der Pol, B. & van der Mark, M. 1928. The heartbeat considered as a relaxation oscillation and an electrical model of the heart. *Philosophical Magazine*, 6: 763–775

RENORMALIZATION GROUPS

Systems undergoing phase transitions at a critical temperature T_c are conveniently described in terms of order parameters and critical exponents. Above T_c, for example, the specific heat is found experimentally to vary as $\varepsilon^{-\alpha}$ where $\varepsilon \equiv (T - T_c)/T_c$, and below T_c, it varies as $(-\varepsilon)^{-\alpha'}$. Thus, α and α' are called critical exponents for specific heat. Similar critical exponents are found for order parameters, isothermal susceptibility, response to an external field, the correlation length ξ, and the pair correlation function $\Gamma(r)$. These are denoted as β, γ, δ, ν, and η, respectively.

Relations among these critical exponents can be derived from scaling assumptions such as the static scaling hypothesis, which asserts that the Gibbs potential $G(T, H)$ for magnetic systems (as an example) is a generalized homogeneous function

$$G(\lambda^{a_\varepsilon}\varepsilon, \lambda^{a_H}H) = \lambda G(\varepsilon, H). \quad (1)$$

With an arbitrary value of λ and selecting a_ε, a_H appropriately, relations among critical exponents are found through thermodynamic identities, for example, $M(\varepsilon, H) = -\,\mathrm{d}G(\varepsilon, H)/\mathrm{d}H$. Thus, Equation (1) leads to the following relations:

$$\beta = \frac{1 - a_H}{a_\varepsilon}, \ \delta = \frac{a_H}{1 - a_H}, \ \gamma' = \frac{2a_H - 1}{a_\varepsilon},$$
$$\gamma = \frac{2a_H - 1}{a_\varepsilon}, \ \alpha' = 2 - \frac{1}{a_\varepsilon}. \quad (2)$$

These can be recast as critical exponent equations:

$$\begin{aligned} \alpha + \beta(\delta + 1) &= 2\,, \\ \alpha' + 2\beta + \gamma' &= 2\,, \\ \gamma(\delta + 1) &= (2 - \alpha)(\delta - 1)\,, \\ \gamma' &= \beta(\delta - 1)\,, \\ \gamma &= \gamma'\,, \\ \alpha &= \alpha'\,, \end{aligned}$$

reducing the number of independent critical exponents to just four (a_ε, a_H, ν, and η).

The two exponents that describe spatial correlations (ν and η) also involve a scaling relationship as was shown through dynamic scaling arguments by Leo Kadanoff and Michael Fisher, among others. In the two-dimensional (2-d) Ising model above T_c, for example, spin-spin correlations show short-range order, whereas below T_c, the system exhibits long-range order. As T tends to T_c from either side, $\Gamma(r - r')$ becomes long ranged and decays slowly as $|r - r'| \gg 1$, with the correlation length diverging. This is typical of most critical systems for which long-range fluctuations accompany the onset of a second-order phase transition. However, in the 2-d XY model, the correlation function exhibits an algebraic fall-off as $\Gamma(r) \sim r^{-\eta(T)}$, demonstrating lack of long-range order in an unusual critical system—the so-called Kosterlitz–Thouless transition whose hallmark is the formation of vortex-antivortex pairs. Mermin and Wagner (1966) proved that any 2-d system with short-range interactions whose ordered phase has a continuous symmetry does *not* support long-range order. This is because at least one branch of collective excitations—called Goldstone bosons—has energy that tends to zero continuously as its wave vector (crystal momentum) vanishes.

With the exception of such unusual systems, spatial scaling applies to criticality, which according to Kadanoff is to be understood as coarse graining. This means that the essential features of the system remain unchanged as the lattice length scale is increased by a factor λ such that $1 \ll \lambda \ll \xi/a$, where ξ is the correlation length and a is the lattice spacing (see Figure 1).

If N is the number of lattice sites and d the dimensionality of the system, then $m = N/\lambda^d$ is the number of blocks obtained in the first rescaling process. Simultaneously, the lattice variables (e.g., spins) and the interaction parameters are redefined. The effective spins now represent each block and interaction parameters refer to the interacting blocks. Taking the Ising model as an example, the cell Hamiltonian is

$$H_{\text{cell}} = -J \sum_{\langle i,j \rangle}^{N} S_i S_j - h \sum_{i}^{N} S_i. \tag{3}$$

Then the block Hamiltonian becomes

$$H_{\text{block}} = -\tilde{J} \sum_{\langle \alpha,\beta \rangle}^{m} \tilde{S}_\alpha \tilde{S}_\beta - \tilde{h} \sum_{\alpha}^{m} \tilde{S}_\alpha, \tag{4}$$

where the tilde quantities refer to blocks. Crucial assumptions are that thermodynamic potentials scale with block size as

$$F_{\text{block}}(\tilde{\varepsilon}, \tilde{h}) = \lambda^d F_{\text{cell}}(\varepsilon, h), \tag{5}$$

where $\tilde{h} = \lambda^x h$, $\tilde{\varepsilon} = \lambda^y \varepsilon$, and F is a homogeneous function. Hence the exponents x and y are calculated as $y = a_\varepsilon d$ and $x = a_{\text{H}} d$. The Kadanoff construction applied to the pair correlation function,

$$\Gamma(i, j, \varepsilon) = \langle (S_i - \langle S \rangle)(S_j - \langle S \rangle) \rangle, \tag{6}$$

gives $y = \nu^{-1}$, and since $y = da_\varepsilon$ and $a_\varepsilon = (2 - \alpha')^{-1}$, we find that: $\alpha = \alpha'$ and $d\nu = 2 - \alpha$. Using $x = a_{\text{H}} d$ with $a_{\text{H}} = \delta/(\delta + 1)$ results in $(2 - \mu)\nu = \delta$.

Thus, static and dynamic scaling hypotheses reduce the number of independent critical exponents to just two. The renormalization group (RG) theory was developed to calculate the values of these exponents for particular models by carrying out the scaling procedure up to ξ, and because $\xi \to \infty$ as $T \to T_{\text{c}}$, scaling should continue "ad infinitum" as T approaches T_{c}.

Denoting the cell Hamiltonian by H_0 and the Hamiltonian after the nth step of rescaling as H_n, the chain of scaling transformations R is

$$\begin{aligned} R(H_0) &= H_1,\ R(H_1) = H_2, \ldots, R(H_n) \\ &= H_{n+1}, \ldots, R(H^*) = H^*, \end{aligned} \tag{7}$$

where H^* denotes a fixed point Hamiltonian characteristic of the critical state. Each step in the RG transformation chain reduces the number of degrees of freedom by λ^d. In statistical mechanics, we need to ensure that the partition function retains the same symmetry and ground-state, and hence we must re-scale the coupling constant at each step, for example, $K = J/kT$ in the Ising model. We then develop a recursion relation to compute the partition function.

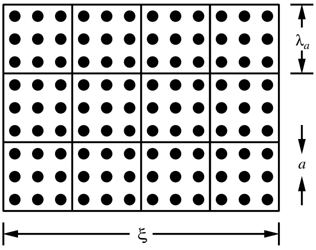

Figure 1. A schematic illustration of the Kadanoff construction.

Model	α	β	γ	ν	δ	η
Classical (MFT)	0 (disc.)	$\frac{1}{2}$	1	—	3	—
Spherical: $d = 3$		$\frac{1}{2}$	2	1	5	0
Spherical: $\varepsilon > 0$	$-\varepsilon/(2-\varepsilon)$	$\frac{1}{2}$	$2/(2-\varepsilon)$	$1/(2-\varepsilon)$	$1 + 4/(2-\varepsilon)$	0
Ising: $d = 2$ (exact)	0 (log)	$\frac{1}{8}$	$\frac{7}{4}$	1	15	$\frac{1}{4}$
Ising: $d = 3$	0.12	0.33	1.25	0.64	4.8	0.04
Heisenberg: $d = 3$	−0.12	0.36	~1.39	0.71	4.8	0.04
S^4-model: $d > 4$	$\varepsilon/2$	$\frac{1}{2} - \varepsilon/4$	1	$\frac{1}{2}$	$3 + \varepsilon$	0
S^4-model: $d = 4$	0	$\frac{1}{2}$	1	$\frac{1}{2}$	3	0
S^4-model: $d < 4$	$\varepsilon/6$	$\frac{1}{2} - \varepsilon/6$	$1 + \varepsilon/6$	$\frac{1}{2} + \varepsilon/12$	$3 + \varepsilon$	0
S^4-model: $d = 3$	0.17	0.33	1.17	0.58	4	0
XY-model: $d = 3$	0.01	0.34	1.30	0.66	4.8	0.04

Table 1. Summary of critical exponents for key models. Note that $\varepsilon = (4 - d)$.

An iterative solution of this recursion relation for the partition function yields "roots" or fixed points that correspond to resultant critical behaviors in the model. The partition function is preserved in the RG procedure via the condition

$$Z_N(H_n) = Z_m(H_{n+1}), \tag{8}$$

where $m = N/\lambda^d$.

It is conjectured that the values of critical exponents are characteristics not of individual Hamiltonians but their sets, with numerous models leading to the same fixed point. The universality hypothesis states that any two physical systems with the same dimensionality, d, and the same number of order parameter components, n, belong to the same universality class, and each fixed point corresponds to one universality class. Table 1 is a summary of the critical exponent values obtained from RG calculations for key theoretical models.

RG ideas extend into many areas of physics, chemistry, biology, and engineering. Based on the work of Kadanoff, Kenneth Wilson proposed an algorithmic approach to the scaling problem by formulating it in reciprocal space. Although less intuitive than the real-space RG of Kadanoff, it leads to exact results for the removal of divergencies in theories of elementary particles. For this work, Wilson was awarded the 1982 Nobel Prize in Physics.

Jack A. Tuszyński

See also **Critical phenomena; Ising model; Order parameters; Phase transitions**

Further Reading

Creswick, R.J., Farach, H.A. & Poole, C.P. 1992. *Introduction to Renormalization Group Methods in Physics*, New York: Wiley

Kosterlitz, J.M. & Thouless, D.J. 1973. Ordering, metastability and phase transitions in two-dimensional systems. *Journal of Physics* C, 6: 1181–1203

Ma, S.-K. 1976. *Modern Theory of Critical Phenomena*, New York: Benjamin

Mermin, N.D. & Wagner, H. 1966. Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic Heisenberg models. *Physics Review Letters*, 17: 1133–1136

Reichl, L.E. 1979. *A Modern Course in Statistical Physics*, Austin: University of Texas Press

Stanley, H.E. 1972. *Introduction to Phase Transitions and Critical Phenomena*, Oxford: Clarendon Press and New York: Oxford University Press

Wilson, K.G. 1972. Feynman-graph expansion for critical exponents. *Physics Review Letters*, 28: 548–551

Wilson, K.G. 1983. The renormalization group and critical phenomena. *Reviews of Modern Physics*, 55: 583–600

Yeomans, J.M. 1992. *Statistical Mechanics of Phase Transitions*, Oxford: Clarendon Press and New York: Oxford University Press

REYNOLDS NUMBER

See **Fluid dynamics**

RHEOLOGY

Rheology is the study of the deformation and flow of materials in general, although most real applications relate to liquids (Barnes, 2000). Deformation and flow processes are linear with respect to extent and rate of deformation only in a limited number of circumstances, and most of the liquids usually considered in practical rheology show strong nonlinear effects in typical deformation and flow situations.

Newtonian Liquids

The simplest flow behavior is that of Newtonian liquids where the viscosity, ν, is a function only of temperature and pressure, and not of the flow conditions (although the dependencies of viscosity on the pressure and temperature are themselves very nonlinear). In simple shear flow (Barnes, 2000), the behavior at any point can be adequately described by the linear equation $\sigma = \nu\dot{\gamma}$, where σ is the shear stress and $\dot{\gamma}$ is the shear rate, or velocity gradient. However, nonlinearities soon set in for the flow of low-viscosity liquids such as water, even at low flow rates. First, smooth, streamline flow is augmented by secondary flows, but these eventually give way to chaotic turbulent flow at higher flow rates. The governing factor for the onset of secondary and turbulent flows is the effect of fluid inertia, arising from the fluid's density. The ratio of inertial to viscous forces is the Reynolds number, Re, and this parameter is important for all fluid flows, whether they are Newtonian or otherwise.

Secondary flows often appear in simple flow geometries such as viscometers and are manifested as an apparent increase in viscosity, due to the extra energy dissipated in the secondary flows cells (Barnes, 2000) (see Figure 1).

The resulting increased couple in the inertial case, T_i, for a cone-and-plate instrument, compared with that for no inertia T, is given for Newtonian liquids by

$$\frac{T_i}{T} \approx 1 + \frac{6}{10^4}\left(\frac{\rho\omega\theta^2 a^2}{\nu}\right)^2, \tag{1}$$

where ρ is the density of the liquid of viscosity ν and a is the cone radius whose angle in radians is θ. From this equation it is easy to see that the correction

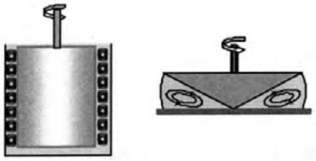

Figure 1. Inertial secondary flow patterns in concentric-cylinder and cone-and-plate geometries.

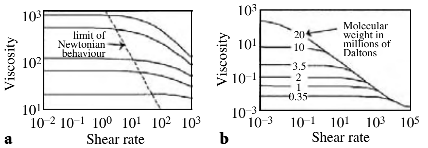

Figure 2. Viscosity (in pascal-seconds) vs. shear rate (1/seconds) for (a) a range of silicone oils, and (b) 3% by weight polystyrene in toluene.

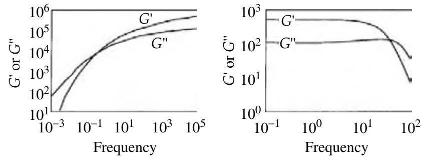

Figure 3. Storage modulus (G') and viscous behavior (G'') (in Pa) vs. Oscillatory frequency (1/s) for (a) low molecular weight polyethylene melt at 150°C, and (b) commercial ketchup.

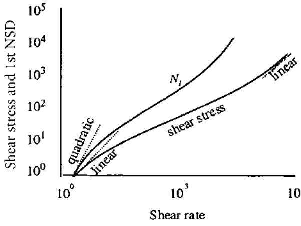

Figure 4. The shear stress (Pa) and first normal-stress difference (1st NSD) N_1(Pa) as a function of shear rate (1/s) for a 0.5% hydroxyethyl cellulose aqueous solution.

becomes important when $\rho\omega\theta^2a^2/\nu$—which is a form of Reynolds number—exceeds 10.

Turbulent flow ensues in pipe flow when the Reynolds number given by $\rho dV/\nu$ (where V is the average velocity in a pipe of diameter d) is well above 2000, and then the pressure drop is no longer a linear function of the flow rate but approaches a quadratic dependence.

Non-Newtonian Liquids

The independence of viscosity to flow conditions is no longer valid for non-Newtonian liquids, but because their microstructure changes with the flow environment, the shear stress shows a nonlinear dependence on the shear rate. The dependence of viscosity on shear rate is such that the typical behavior is a decrease of viscosity with increase in shear rate from a value ν_0 at low shear rates, to a much lower value ν_∞ at much higher shear rates. This can often be described by the nonlinear Cross-type function

$$\frac{\nu - \nu_\infty}{\nu_0 - \nu_\infty} = \frac{1}{1 + (K\dot{\gamma})^m}, \qquad (2)$$

where (written in this particular way) K has the dimensions of time and m is dimensionless. When this model is used to describe non-Newtonian liquids, the degree to which the viscosity decreases with shear rate is dictated by the value of m, with m tending to zero describing more Newtonian liquids, while the most "shear-thinning" liquids have values of m tending to unity. Figure 2 shows the departure from linearity for a number of non-Newtonian liquids.

Viscoelasticity

As well as the nonlinear behavior of viscosity, non-Newtonian liquids also display extra forces. These arise either as time or frequency effects in the linear region of behavior and are manifested, for instance, as an elastic component called the storage modulus G' in oscillatory flow as well as the viscous behavior as G''. In the simplest form of viscoelastic liquid, the parameters G' and G'' are related by a relaxation time, λ, of the liquid. These parameters are only constant over a limited region, but for deformations greater than a few percent for most non-Newtonian liquids, we enter the nonlinear region. The behavior of these G' and G'' with frequency is shown in Figure 3, along with typical departure from linearity.

For steady flows of viscoelastic liquids, the extra forces that become important are the so-called normal stresses. These arise essentially because the microstructure has a preferred configuration, which can become aligned by the flow. These normal forces manifest themselves in the linear region of slow flow as forces proportional to the square of the shear rate, but they also soon show nonlinear behavior in the region where the viscosity also becomes nonlinear. Figure 4 shows the typical behavior of both the shear stress and the first normal stress difference for a viscoelastic liquid.

Rheology of Polymeric Liquids

The microstructure that dictates the behavior of polymeric liquids is usually the state of entanglement of the polymeric chains. The chains are in incessant motion due to Brownian thermal energy, and their motions among each other give rise to viscous and elastic effects. However, as the shear rate increases, there is a gradual loss of entanglements and development of alignment, so that at high enough shear rates, linear polymer coils are transformed into aligned strings. The consequences of this change in microstructure are some very startling viscoelastic effects (Barnes, 2000).

Rheology of Suspensions

Suspension rheology depends on the relative disposition of particles relative to one another. Sometimes the particles are flocculated, and these flocs break down in size as the shear rate increases. This nonlinear phenomena can give rise to apparent yield stresses, where the viscosity is very high at low shear rates, but drops dramatically over a narrow range of stress.

Theoretical Rheology

The aim of mathematicians working in this area is to capture these complicated, and usually nonlinear, physical phenomena in constitutive equations (CEs). At their best, these describe nonlinear behaviour in viscosity—as above with the Cross-type equations—but also in the viscoelastic areas. CEs are written in the form of either differential or integral equations (see Barnes et al., 1989). A typical example of a nonlinear differential-type equation is the White–Metzner equation, where the rate-dependent viscosity $\nu_s(II_{\mathrm{D}})$ and the relaxation time $\lambda(II_{\mathrm{D}})_I I$ are written as

$$\nu_s(II_{\mathrm{D}}) = \frac{\nu_0}{[1+(K_1 II_D)^n]}$$

and

$$\lambda(II_{\mathrm{D}}) = \frac{\lambda_0}{[1+K_2 II_{\mathrm{D}}]}, \qquad (3)$$

where K_1 and K_2 are parameters with the dimensions of time, n is a numerical constant, and II_{D} is the second invariant of the deformation tensor (Barnes & Roberts, 1992). Barnes & Roberts (1992) have shown that this type of equation can describe many nonlinear rheological situations.

Howard A. Barnes

See also **Cluster coagulation; Fluid dynamics; Granular materials; Liquid crystals; Polymers**

Further Reading

Barnes, H.A. 2000. *A Handbook of Elementary Rheology*, Aberystwyth: The University of Wales Institute of Non-Newtonian Fluid Mechanics

Barnes, H.A., Hutton, J.F. & Walters, K. 1989. *An Introduction to Rheology*, Amsterdam: Elsevier

Barnes, H.A. & Roberts, G.P. 1992. A simple model to describe the steady state shear and extensions viscosity of polymer melts. *Journal of Non-Newtonian Fluid Mechanics*, 44: 113–126

RICCATI EQUATIONS

Count Jacopo Francesco Riccati (1676–1754) was an Italian nobleman who spent most of his life in the little town of Castelfranco, Veneto. Born in Venice and schooled at a Jesuit college in Brescia, he proceeded to the University of Padua where he read law. However, following his natural inclination and talent, he also took mathematical courses at Padua, and befriended Father Stefano degli Angeli, an astronomy lecturer. Angeli introduced him to Newton's *Principia*, which they read and studied together.

Dedicated to intellectual inquiry of every sort, Riccati is mostly remembered for his significant results in mathematics, particularly separation of variables and other solution methods for ordinary differential equations. He wrote extensively on such diverse areas as physics, architecture, and philosophy. Preferring to remain in Castelfranco, where he served as mayor for several years, he turned down numerous academic offers, including an invitation from Peter the Great to become President of the St. Petersburg Academy of Science.

Throughout his life he maintained an active correspondence with Italian scholars such as Agnesi and Rizzetti, as well as with European mathematicians such as the Bernoullis. In a letter of 1720 addressed to Rizzetti, he introduced two new differential equations, which in modern notation would be written

$$\frac{\mathrm{d}y}{\mathrm{d}x} = ay^2 + bx^m, \quad \frac{\mathrm{d}y}{\mathrm{d}x} = \alpha y^2 + \beta x + \gamma x^2, \qquad (1)$$

where a, b, m and α, β, γ are constants. His first published results on such equations appeared four years later (Riccati, 1724), and this paper is reproduced in Bittanti (1989) along with later related work of Euler and Liouville. Michieli's biographical work (Michieli, 1943) includes a detailed bibliography of Riccati's publications, letters, and manuscripts.

The scalar first order differential equation that today bears Riccati's name takes the general form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = A(x)y^2 + B(x)y + C(x), \qquad (2)$$

where $A \not\equiv 0$, B, C are arbitrary functions of x. Thus, Equations (1) correspond to special cases of the Riccati equation (2) with constant A, zero B, and particular choices of C. There are two main solution methods (Ince, 1926). First, supposing that a particular solution y_1 is known, the general solution of (2) is found by the substitution

$$y = y_1 + \frac{1}{u}, \qquad (3)$$

which yields the linear equation

$$\frac{\mathrm{d}u}{\mathrm{d}x} + (2y_1 A + B)u + A = 0$$

for u. Hence u is found by quadratures, and then y is obtained (3) in terms of y_1 and u. The second method involves the substitution

$$y = -\frac{1}{A}\frac{\mathrm{d}}{\mathrm{d}x}\log\psi, \qquad (4)$$

which transforms the Riccati equation to a homogeneous second-order linear equation for ψ, namely,

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + f(x)\frac{\mathrm{d}\psi}{\mathrm{d}x} + g(x)\psi = 0 \qquad (5)$$

with

$$f = -B - \frac{\mathrm{d}}{\mathrm{d}x}\log A, \quad g = AC.$$

The scalar- second order Equation (5) can be converted to a 2×2 matrix linear system of the form

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}. \tag{6}$$

Riccati himself considered such a system as describing the trajectory of a point (ψ_1, ψ_2) in the plane. Requesting the equation satisfied by the slope

$$y = \frac{\psi_2}{\psi_1}, \tag{7}$$

he found the answer was precisely (2) with

$$A = -M_{12}, \quad B = M_{22} - M_{11}, \quad C = M_{21}.$$

The left action of the matrix group $SL(2)$ on the linear system (6) induces a corresponding action via Möbius transformations on the projective coordinate y. This generalizes to Nth-order projective Riccati systems with nonlinear superposition principles (Anderson, 1980; Shnider & Winternitz, 1984). The latter are solved in terms of $(N+1)$th-order linear systems, with a projective action of $SL(N+1)$.

Due to this link with linear systems, (coupled) Riccati equations naturally appear in the theory of soliton-bearing integrable partial differential equations (Fordy, 1990). Within the framework of the inverse scattering method, nonlinear soliton equations (in $1+1$ dimensions) arise as the compatibility condition for a pair of matrix linear systems,

$$\Psi_x = U\Psi, \quad \Psi_t = V\Psi. \tag{8}$$

Without loss of generality the matrices U, V are taken as $sl(N+1)$ valued functions of the dependent variables and their derivatives as well as a spectral parameter, λ, say, and subscripts denote partial derivatives.

In the simplest case, $N = 1$, we consider the standard example of the Korteweg–de Vries (KdV) equation, with the first matrix in (8) given by

$$U = \begin{pmatrix} 0 & 1 \\ \lambda - u & 0 \end{pmatrix},$$

and the spatial (x) part of pair (8) is just a 2×2 system (6). The corresponding scalar equation (5) reduces to the Schrödinger equation with potential u and spectral parameter λ; that is,

$$\psi_{xx} + u\psi = \lambda\psi,$$

and in terms of the projective variable y this yields the Riccati equation

$$y_x = -y^2 + \lambda - u. \tag{9}$$

In the special case of zero eigenvalue ($\lambda = 0$), with $y = v$ formula (9) becomes the Miura map

$$u = -v_x - v^2$$

relating a solution u of the KdV equation to a solution v of the modified KdV equation. Riccati equations are further distinguished in the theory of integrable systems by the fact that (for analytic A, B, C) they are the only first- order equations with the Painlevé property (Ince, 1926).

The fundamental problem of the calculus of variations (Gel'fand & Fomin, 1963) is to extremize a functional

$$S[q] = \int_{x_0}^{x_1} L(x, q, \dot{q})\,\mathrm{d}x, \quad \dot{q} \equiv \frac{\mathrm{d}q}{\mathrm{d}x}, \tag{10}$$

subject to $q(x_0) = q_0, q(x_1) = q_1$. Under an arbitrary small change $h(x)$ in $q(x)$,

$$q \to q + h,$$

with $h(x_0) = 0 = h(x_1)$, for a scalar function q the requirement $\delta S[h] = 0$ leads to the Euler–Lagrange equation

$$L_q - \frac{\mathrm{d}}{\mathrm{d}x}L_{\dot{q}} \equiv \frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial L}{\partial \dot{q}} = 0.$$

If we further require to minimize the functional S, then a necessary condition is that the second variation should be nonnegative, that is,

$$\delta^2 S[h] = \frac{1}{2}\int_{x_0}^{x_1} \left(L_{\dot{q}\dot{q}}\dot{h}^2 + 2L_{q\dot{q}}h\dot{h} + L_{qq}h^2\right)\,\mathrm{d}x \geq 0.$$

After an integration by parts, this takes the form

$$\delta^2 S[h] = \int_{x_0}^{x_1} (P\dot{h}^2 + Qh^2)\,\mathrm{d}x, \tag{11}$$

where

$$P = \frac{1}{2}L_{\dot{q}\dot{q}}, \quad Q = \frac{1}{2}\left(L_{qq} - \frac{\mathrm{d}}{\mathrm{d}x}L_{q\dot{q}}\right).$$

Since h vanishes at the endpoints, there is the freedom to add any function of the form

$$\frac{\mathrm{d}}{\mathrm{d}x}(wh^2)$$

to the integrand in (11). If w is chosen to satisfy the Riccati equation

$$\frac{\mathrm{d}w}{\mathrm{d}x} = \frac{w^2}{P} - Q, \tag{12}$$

then the integrand for the second variation can be written in terms of a perfect square,

$$\delta^2 S[h] = \int_{x_0}^{x_1} P\left(\dot{h} + \frac{wh}{P}\right)^2\,\mathrm{d}x.$$

Thus, to ensure nonnegativity of $\delta^2 S$, we arrive at Legendre's necessary condition

$$P(x) \geq 0, \quad x \in [x_0, x_1],$$

for a minimum of the functional $S[q]$. However, some care is needed to make this argument rigorous, since the Riccati equation (12) may not have a solution on the whole interval $[x_0, x_1]$.

In the case where q has n components, $q = (q^1, q^2, \ldots, q^n)$, the analogue of (12) is the matrix Riccati equation

$$\dot{W} = W P^{-1} W - Q,$$

where W, P, Q are symmetric matrices. Matrix Riccati equations are treated extensively in Reid (1972), and together with their discrete versions they find important applications in optimal filtering and control (Bittanti et al., 1991; Zelikin, 2000).

An n-component example is provided by the geodesics on a (pseudo-)Riemannian n-manifold with metric g_{jk} (Hughston & Tod, 1990), derived from the action

$$S[q] = \int_{\tau_0}^{\tau_1} g_{jk}(q) \dot{q}^j \dot{q}^k \, \mathrm{d}\tau$$

with affine parameter τ along the geodesics. The Euler–Lagrange equations for the geodesics, giving the trajectories of free particle motion, are

$$\frac{\mathrm{d}^2 q^j}{\mathrm{d}\tau^2} + \Gamma^j_{kl} \frac{\mathrm{d}q^k}{\mathrm{d}\tau} \frac{\mathrm{d}q^l}{\mathrm{d}\tau} = 0,$$

where Γ^j_{kl} are the Christoffel symbols. In the case of a Riemannian manifold, the metric tensor is positive definite and the Legendre condition is satisfied. The geodesics are the paths of minimum length.

ANDREW HONE

See also **Euler–Lagrange equations; Extremum principles; General relativity; Integrability; Inverse scattering method or transform; Korteweg–de Vries equation; Painlevé analysis**

Further Reading

Anderson, R.L. 1980. A nonlinear superposition principle admitted by coupled Riccati equations of the projective type. *Letters in Mathematical Physics*, 4: 1–7

Bittanti, S. (editor). 1989. *Count Riccati and the Early Days of the Riccati Equation*. Bologna: Pitagora Editrice

Bittanti, S., Laub, A.J. & Willems, J.C. (editors). 1991. *The Riccati Equation*. Berlin: Springer

Fordy, A.P. (editor). 1990. *Soliton Theory: A Survey of Results*. Manchester: Manchester University Press

Gel'fand, I.M. & Fomin, S.V. 1963. *Calculus of Variations*. Englewood Cliffs, NJ: Prentice-Hall

Hughston, L.P. & Tod, K.P. 1990. *An Introduction to General Relativity*. Cambridge and New York: Cambridge University Press

Ince, E.L. 1926. *Ordinary Differential Equations*. 7th edition, Edinburgh: Oliver and Boyd and New York: Interscience, 1959

Michieli, A.A. 1943. Una famiglia di matematici e di poligrafi trivigiani: i Riccati. I. Iacopo Riccati. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, 102(2): 535–587

Reid, W.T. 1972. *Riccati Differential Equations*. New York: Academic Press

Riccati, J.F. 1724. Animadversiones in aequationes differentiales secundi gradus. *Supplementa Acta Eruditorum Lipsiae*, 8(2): 66–73

Shnider, S. & Winternitz, P. 1984. Classification of nonlinear ordinary differential equations with superposition principles. *Journal of Mathematical Physics*, 25(11): 3155–3165

Zelikin, M.I. 2000. *Control Theory and Optimization I: Homogeneous Spaces and the Riccati Equation in the Calculus of Variations*. Berlin and New York: Springer

RIEMANN INVARIANTS

See **Characteristics**

RIEMANN SURFACES

See **Periodic spectral theory**

RIEMANN WAVE

See **Shock waves**

RIEMANN ZETA FUNCTION

See **Random matrix theory I: Origins and physical applications**

RIEMANN–HILBERT PROBLEM

We begin by defining what is meant by a Riemann–Hilbert problem (RHP). Let Σ be an oriented contour in $\mathbb{C}$.

By convention, if we move along the contour in the direction of the orientation, we say that the $(+/-)$ side lies to the (left/right), as indicated in Figure 1. A map v from Σ to the space of the invertible $k \times k$ matrices with complex entries is called a *jump matrix* for Σ if v and v^{-1} are bounded on Σ. We say that a

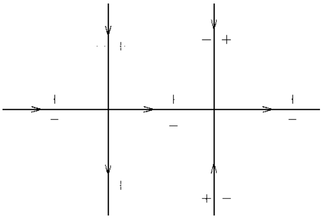

Figure 1. The contour Σ.

$j \times k$ matrix-valued function $m = m(z)$ is a *solution* of the RHP (Σ, v) if

- m is analytic in $\mathbb{C} \setminus \Sigma$,
- $m_+(z) = m_-(z)v(z)$ for $z \in \Sigma$, where $m_\pm(z) = \lim_{\substack{z' \to z \\ z' \in (\pm)\text{–side of } \Sigma}} m(z')$.

  If in addition $j = k$ and
- $m(z) \to I$ as $z \to z_0$ for some $z_0 \in \mathbb{C} \cup \{\infty\}$, we say that the RHP is *normalized* at z_0.

A discussion of technical restrictions on Σ, and the precise sense in which the boundary values, as well as the limit at z_0 in the normalized case, are attained, can be found in many texts (see, e.g., Clancey & Gohberg (1981) and the references therein). Many problems in pure and applied mathematics, and also theoretical physics, can be expressed in terms of an RHP. Riemann–Hilbert problems are closely related to the Wiener–Hopf method. We refer the reader to Clancey & Gohberg (1981) for a history of RHPs and also to the classic text by Muskhelishvili (1946). The goal of this article is to describe some recent developments.

Remark. Included in David Hilbert's famous list of 23 problems, was the following:

> Show that there always exists a linear differential equation of the Fuchsian class with given singularities and a given monodromy group.

A Fuchsian system is an nth-order, linear system of the form

$$\mathrm{d}m/\mathrm{d}z = A(z)m,$$

where A is an $n \times n$ matrix with entries that are analytic except at a finite number of simple poles. Roughly speaking, the monodromy group is the group of transformations of solutions to the above Fuchsian system under analytic continuation about these poles.

The above monodromy problem turns out to be a special case of a Riemann–Hilbert problem with jump matrices that are (essentially) piece-wise constant. In the mathematical literature, the monodromy problem is sometimes identified with the Riemann–Hilbert problem, but this is incorrect. The work of Bernhard Riemann, and later of David Hilbert, on such problems, predates Hilbert's monodromy problem and has very different origins.

Our point of departure here is the observation of Alexey Shabat in 1976 that the inverse scattering method for the Schrödinger operator can be rephrased as a RHP (see Faddeev & Takhtajan, 1987). By way of illustration, we apply Shabat's observation to a simpler case, the self-adjoint ZS–AKNS scattering problem (Zakharov & Shabat, 1971; Ablowitz et al., 1974)

$$\frac{\mathrm{d}\psi}{\mathrm{d}x} = \left(\mathrm{i}z\sigma + \begin{pmatrix} 0 & q(x) \\ \bar{q}(x) & 0 \end{pmatrix}\right)\psi, \tag{1}$$

where $\sigma = \begin{pmatrix} 1/2 & 0 \\ 0 & -1/2 \end{pmatrix}$ and $q(x)$ is a given function on $\mathbb{R}$, which decays sufficiently fast as $|x| \to \infty$. For fixed $z \in \mathbb{C} \setminus \mathbb{R}$, one seeks so-called Beals–Coifman solutions $\psi = \psi(x, z; q)$ of (1) with the properties

$$\psi \mathrm{e}^{-\mathrm{i}xz\sigma} \to I \text{ as } x \to -\infty, \tag{2}$$

$$\psi \mathrm{e}^{-\mathrm{i}xz\sigma} \text{ is bounded as } x \to +\infty. \tag{3}$$

It turns out, conversely, that for fixed $x \in \mathbb{R}$, $m(z) = m(x, z; q) \equiv \psi(x, z; q)\mathrm{e}^{-\mathrm{i}xz\sigma}$ solves the RHP $(\Sigma = \mathbb{R}, v_x)$ normalized at $z_0 = \infty$, where

$$v_x(z) = \begin{pmatrix} 1 - |r(z)|^2 & r(z)\mathrm{e}^{\mathrm{i}xz} \\ -\bar{r}(z)\mathrm{e}^{-\mathrm{i}xz} & 1 \end{pmatrix}, \quad z \in \mathbb{R} \tag{4}$$

and $r(z) = r(z; q)$ is the *reflection coefficient* for q. Define the *reflection map* $\mathcal{R}$ via $\mathcal{R}(q) \equiv r$. Given q, the direct scattering problem is given by

$$q \mapsto m(x, z; q) \mapsto v_x \mapsto r = \mathcal{R}(q).$$

On the other hand, for a given r and a fixed $x \in \mathbb{R}$, let $m = m(x, z; r)$ be the solution of the RHP $(\mathbb{R}, v_x)$ normalized at ∞. Expand $m(z)$ as $z \to \infty$,

$$m(z) = I + m_1(x; r)/z + O(z^{-2}), \tag{5}$$

and define

$$(Q(r))(x) \equiv -\mathrm{i}(m_1(x; r))_{12}. \tag{6}$$

Given r, the inverse scattering problem is given by

$$r \mapsto v_x \mapsto m(x, z; r) \mapsto m_1(x; r) \mapsto Q(r).$$

The basic analytical result in the subject is that $\mathcal{R}$ is bijective from a suitable space of q's onto a suitable space of r's (Beals & Coifman , 1984; Zhou, 1998) and that Q is the inverse of $\mathcal{R}$.

In terms of the classical method of scattering and inverse scattering (Zakharov & Shabat, 1971; Ablowitz et al., 1974), $r \equiv r(z; q)$ is the standard reflection coefficient. Moreover, if m solves the RHP (Σ, v_x), then the Fourier transform of the first row of $m_- \begin{pmatrix} 1 & r\mathrm{e}^{\mathrm{i}xz} \\ 0 & 1 \end{pmatrix} - I$ gives the solution of the Faddeev–Marchenko equation.

In the case of the non-self-adjoint ZS–AKNS scattering problem, the Beals–Coifman solutions $\psi(z)$ may have singularities at points $\{z_j\} \subset \mathbb{C} \setminus \mathbb{R}$, which correspond to eigenvalues with square integrable eigenfunctions. In addition, there may be singularities for $\psi_\pm(z)$ on the contour $\Sigma = \mathbb{R}$. One now associates an RHP to the scattering problem by augmenting the contour $\Sigma \to \Sigma_{\mathrm{aug}}$ so as to enclose these singularities: information about the singularities is then contained in the augmented jump matrix v_{aug} defined on $\Sigma_{\mathrm{aug}} \setminus \Sigma$ (see Zhou, 1989). The above ideas apply to very general first-order systems of size $N \times N$ and also to differential operators of order M. The underlying contour Σ is now more complicated and may have points of self-intersection, which leads in turn to significant new technical complications. Also, the singularities of ψ in $\mathbb{C} \setminus \Sigma$ need not correspond to eigenvalues with square integrable eigenfunctions.

We now consider the relation of RHPs to the theory of integrable systems. Again by the way of illustration, we consider the normalized RHP $(\mathbb{R}, v_x)$ associated with the self-adjoint ZS–AKNS system. If $q(t) = q(x,t)$ solves the defocusing nonlinear Schrödinger (NLS) equation

$$\begin{aligned} &\mathrm{i}q_t + q_{xx} - 2|q|^2 q = 0, \\ &q(x, t=0) = q_0(x) \to 0 \text{ as } |x| \to \infty, \end{aligned} \tag{7}$$

then by Zakharov & Shabat (1971)

$$r(t,z) = r(z)\mathrm{e}^{-\mathrm{i}tz^2}, \quad z \in \mathbb{R}. \tag{8}$$

This leads to the following solution procedure for NLS. Let $m(x,z; r(\Diamond)\mathrm{e}^{-\mathrm{i}t\Diamond^2}) = I + m_1(x; r(\Diamond)\mathrm{e}^{-\mathrm{i}t\Diamond^2})z^{-1} + O(z^{-2})$ be the solution of the normalized RHP $(\mathbb{R}, v_\theta)$, where

$$v_\theta(z) = \begin{pmatrix} 1 - |r(z)|^2 & r\mathrm{e}^{\mathrm{i}\theta} \\ -\bar{r}\mathrm{e}^{-\mathrm{i}\theta} & 1 \end{pmatrix}, \quad \theta = xz - tz^2. \tag{9}$$

Then

$$q(x,t) = -\mathrm{i}(m_1(x; r(\Diamond)\mathrm{e}^{-\mathrm{i}t\Diamond^2}))_{12}. \tag{10}$$

An extraordinarily broad spectrum of problems, both dynamical and nondynamical, can be solved via a RHP as in (10) above. The list includes, in particular, the Painlevé equations (see Flaschka & Newell, 1980; Jimbo et al., 1981) where the underlying contour may be a union of intersecting lines, and also the orthogonal polynomial problem as formulated by Fokas, Its, and Kitaev, and the theory of Hankel and Toeplitz determinants (see e.g., Deift, 1999). In practice, RHPs often arise via associated *integrable operators* (Its et al., 1990; see also Deift, 1999). Let Σ be an oriented contour in $\mathbb{C}$. We say that an operator K acting in $L^p(\Sigma)$, $1 < p < \infty$, is integrable if it has a kernel of the form

$$K(z,z') = (z - z')^{-1} \sum_{j=1}^{k} f_j(z) g_j(z'), \quad z, z' \in \Sigma.$$

Special examples of integrable operators began to appear in field theory and in statistical mechanics in the late 1960s and 1970s, particularly in the work of Wu, McCoy, Tracy, and Baruch, and later Sato, Miwa, and Jimbo, but integrable operators as a distinguished class were first identified by Its et al. (1990) and later Sato-Miwa-Jimbo. Some general results on integrable operators were obtained by Lev A. Sakhnovich in the late 1960s, but the full theory of integrable operators is due to Its et al. (1990).

Integrable operators have many remarkable properties, the most important being that if $(1-K)^{-1} = I + R$, then R is also an integrable operator, $R(z,z') = z - z')^{-1} \sum_{j=1}^{k} F(z)G_j(z')$, and F_i, G_j can be constructed by solving an RHP $(\Sigma, v_{f,g})$ naturally associated with f_i, g_j. It turns out that many quantities H of physical interest can be expressed in the form $H = \det(1-K)$, where K is an integrable operator. For example, let K_x be the operator acting on $L^2(0,1)$ with kernel $\frac{\sin \pi x(z-z')}{\pi(z-z')}$, $z, z' \in (0,1)$. Then $H_x = \det(1 - K_x)$ is the probability that in the bulk scaling limit, a Hermitian matrix chosen from the Gaussian Unitary Ensemble has no eigenvalues in $[0, x]$. Differentiating with respect to x, one derives an expression for $\frac{\mathrm{d}}{\mathrm{d}x} \log H_x$ in terms of $R_x = (1 - K_x)^{-1} - 1$, and hence we see that the analysis of H_x reduces to the analysis of an RHP. Similar examples arise in many different scientific areas, including statistical mechanics, percolation theory, combinatorics, representation theory of large groups, and approximation theory. Relatively recent bibliographies can be found in Deift et al. (1993, 1999).

Returning to the NLS equation, one expects from (9) and (10) that the leading order contribution to the solution $q(x,t)$ as $t \to \infty$ should arise from the stationary phase point $z_0 = x/2t$, $\theta'(z_0) = 0$, as in the classical steepest-descent method for the asymptotic evaluation of (scalar) integrals. In Deift & Zhou (1993), the authors introduce a general noncommutative steepest-descent method to analyze such microlocal problems in the context of RHPs. In the case of NLS, the method proceeds as follows (Deift, Its & Zhou, 1997; Deift & Zhou, 1994). Let δ solve the scalar RHP $((-\infty, z_0), 1 - |r|^2)$ normalized at ∞ and set $\tilde{m} = m\delta^{-\sigma_3}$, $\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Then $\tilde{m}$ solves the normalized RHP $(\mathbb{R}, \tilde{v}_\theta)$, where $\tilde{v}_\theta$ is as in Figure 2.

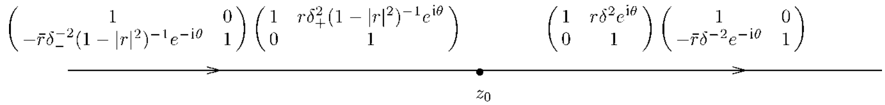
$\begin{pmatrix} 1 & 0 \\ -\bar{r}\delta_-^{-2}(1-|r|^2)^{-1}e^{-\mathrm{i}\theta} & 1 \end{pmatrix} \begin{pmatrix} 1 & r\delta_+^2(1-|r|^2)^{-1}e^{\mathrm{i}\theta} \\ 0 & 1 \end{pmatrix}$
$\begin{pmatrix} 1 & r\delta^2 e^{\mathrm{i}\theta} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\bar{r}\delta^{-2}e^{-\mathrm{i}\theta} & 1 \end{pmatrix}$
z_0

Figure 2. The jump matrix $\tilde{v}_\theta$.

Assuming (as we may, after suitable approximation) that the entries of $\tilde{v}_\theta$ have analytic extensions, define m_θ^d as in Figure 3. Note that, with this definition, m_θ^d solves the normalized RHP $(\Sigma_{z_0}^d, v_\theta^d)$, where v_θ^d is shown in Figure 4.

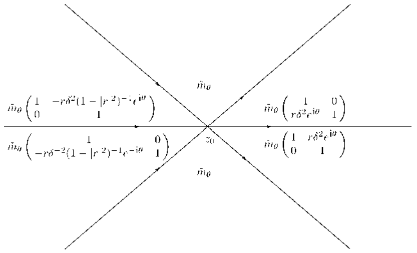
$\tilde{m}_\theta$
$\tilde{m}_\theta \begin{pmatrix} 1 & -r\delta^2(1-|r|^2)^{-1}e^{i\theta} \\ 0 & 1 \end{pmatrix}$
$\tilde{m}_\theta \begin{pmatrix} 1 & 0 \\ r\delta^2 e^{i\theta} & 1 \end{pmatrix}$
$\tilde{m}_\theta \begin{pmatrix} 1 & 0 \\ -r\delta^{-2}(1-|r|^2)^{-1}e^{-i\theta} & 1 \end{pmatrix}$
z_0
$\tilde{m}_\theta \begin{pmatrix} 1 & r\delta^2 e^{i\theta} \\ 0 & 1 \end{pmatrix}$
$\tilde{m}_\theta$

Figure 3. The solution m_θ^d.

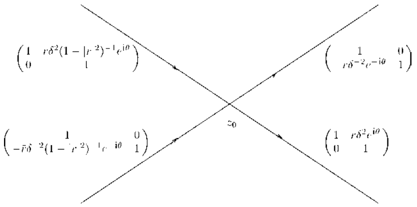
$\begin{pmatrix} 1 & r\delta^2(1-|r|^2)^{-1}e^{i\theta} \\ 0 & 1 \end{pmatrix}$
$\begin{pmatrix} 1 & 0 \\ r\delta^{-2}e^{-i\theta} & 1 \end{pmatrix}$
z_0
$\begin{pmatrix} 1 & 0 \\ -\bar{r}\delta^{-2}(1-|r|^2)^{-1}e^{-i\theta} & 1 \end{pmatrix}$
$\begin{pmatrix} 1 & r\delta^2 e^{i\theta} \\ 0 & 1 \end{pmatrix}$

Figure 4. The jump matrix v_θ^d.

Observe that the definition of δ and the choice of the lower/upper and upper/lower factorizations of $\tilde{v}_\theta$, were made precisely to take advantage of the signature table of $\mathrm{Re}\, \mathrm{i}\theta = -t\, \mathrm{Re}\, \mathrm{i}\,(z-z_0)^2$ in Figure 5.

As $t \to \infty$, $v_\theta^d \to I$ uniformly on $\Sigma_{z_0}^d \setminus \{|z-z_0| \geq \varepsilon\}$ for any $\varepsilon > 0$, and the RHP localizes at z_0 to an RHP that can be solved explicitly. We obtain as $t \to \infty$

$$q(x,t) = t^{-1/2}\alpha(z_0)\mathrm{e}^{\mathrm{i}[x^2/(4t)-\nu(z_0)\log 2t]} + O(\log t/t), \tag{11}$$

where

$$\nu(z) = -\frac{1}{2\pi}\log(1-|r(z)|^2), \quad |\alpha(z)|^2 = \nu(z)/2 \tag{12}$$

and

$$\arg\alpha(z) = \frac{1}{\pi}\int_{-\infty}^{z}\log(z-s)d\log(1-|r(s)|^2) + \frac{\pi}{4} + \arg\Gamma(\mathrm{i}\nu(z)) + \arg r(z). \tag{13}$$

The above asymptotic form was first obtained by Zakharov and Manakov but without the error estimate. The RHP/steepest-descent method, suitably extended, has also been used by Deift and Zhou to obtain precise long-time asymptotics of solutions of perturbed NLS equations.

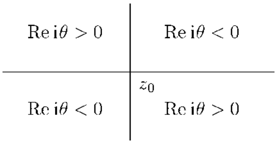
$\mathrm{Re}\, \mathrm{i}\theta > 0$
$\mathrm{Re}\, \mathrm{i}\theta < 0$
z_0
$\mathrm{Re}\, \mathrm{i}\theta < 0$
$\mathrm{Re}\, \mathrm{i}\theta > 0$

Figure 5. Signature of Re, $\mathrm{i}\theta$.

In the above problem, the leading order asymptotic contribution arises from the stationary phase point, as in the classical method of stationary phase for scalar integrals. In many problems, in fact in most problems, the leading order contribution no longer arises from one (or more) critical points, but is given instead by a "critical contour." Such a contour arises, for example, in the asymptotic analysis as $x \to \infty$ of the Painlevé equation by Deift and Zhou, and also in the analysis of the collisionless shock region by Deift, Venakides and Zhou. In Deift, Venakides and Zhou (1997), in the context of the zero-dispersion limit of the Kortweg–de Vries (KdV) equation, the method in Deift & Zhou (1993) was extended significantly to include a prescription for the determination of the critical contour, making it possible in turn to describe the asymptotic development of fully nonlinear oscillations. This extension of Deift & Zhou (1993) made it possible to obtain Plancherel–Rotach-type asymptotics for a general class of orthogonal polynomials, leading in turn to a solution of the so-called universality conjecture for unitary ensembles (see Deift et al., 1999, and **Random Matrix Theory IV: Analytic methods**; see also the work of Pastur and Scherbina for a different approach, and also the work of Bleher and Its for an approach related to Deift et al. (1999), in the special case of a quartic potential). Many other long-standing problems in mathematics and in physics have been solved since the 1990s using RHP techniques and the (extended) steepest-descent method. For example, in combinatorics, Ulam's problem for increasing subsequences in random permutations was solved completely by Baik, Deift, and Johansson in terms of the Tracy–Widom distribution of random matrix theory, giving rise in turn to an explosive growth in applications linking combinatorics, statistical models, and random matrix theory. Many people are contributing to these developments, but there is as yet no adequate review. Fortunately, a comprehensive review, by T. Kriecherbauer, is due to appear in *Nonlinearity* in early 2005. Further extensions of the RHP/steepest-descent method have been given recently in the work of Kamvissis, McLaughlin and Miller on the semiclassical limit for the defocusing nonlinear Schrödinger equation.

Finally, we return to the classical Faddeev–Marchenko approach to inverse scattering theory mentioned earlier and make some comparisons with the RHP method. First, as noted above, the Faddeev–Marchenko equation arises by taking a Fourier transform of objects arising in the RHP theory. This means, in particular, that the classical method only applies in cases where the underlying contour is a group such as a line or a circle; problems such as Painlevé II, for example, cannot be handled in general by the classical method. Second, even in cases where the underlying contour is a group, such as NLS or KdV, the leading order contribution to the solution as $t \to \infty$ arises from the stationary phase points. Taking a Fourier transform "smears out" this feature and the microlocal nature of the problem is obscured.

Beals & Coifman(1984) were the first to use RHPs for the rigorous analysis of scattering and inverse scattering theory.

PERCY DEIFT AND XIN ZHOU

See also **Gel'fand–Levitan theory; Inverse scattering method or transform; Nonlinear Schrödinger equations; Random matrix theory IV: Analytic methods**

Further Reading

Ablowitz, M.J., Kaup, D.J., Newell, A.C. & Segur, H. 1974. The inverse scattering transform-Fourier analysis for nonlinear problems. *Studies in Applied Mathematics*, 53: 249–315

Beals, B. & Coifman, R. 1984. Scattering and inverse scattering for first order systems. *Communications in Pure and Applied Mathematics*, 37, 39–90

Clancey, K. & Gohberg, I.C. 1981. *Factorizations of Matrix Functions and Singular Integral Operators*, Basel and Boston: Birkhäuser

Deift, P. 1999. Integrable operators. *American Mathematical Society Translations*, 2(189): 69–84

Deift, P., Its, A. & Zhou, X. 1993. Long–time asymptotics for integrable nonlinear wave equations. In *Important Developments in Soliton Theory 1980–1990*, edited by A.S. Fokas & V.E. Zakharov, Berlin and New York: Springer, pp.181–204

Deift, P., Kriecherbauer, T., McLaughlin, K., Venakides, S. & Zhou, X. 1999. Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory. *Communications in Pure and Applied Mathematics*, 52: 1335–1425

Deift, P., Venakides, S. & Zhou, X. 1997. New results in small dispersion KdV by an extension of the steepest descent method for Riemann-Hilbert problems, *Internat. Math. Res. Notices*, No. 6, 285–299

Deift, P. & Zhou, X. 1993. A steepest descent method for oscillatory Riemann–Hilbert problems–asymptotics for the MKdV equation. *Annals of Mathematics*, 137: 295–368

Deift, P. & Zhou, X. 1994. *Long-term Behavior of the Non-focusing Nonlinear Schrödinger Equation—A Case Study*, Tokyo: University of Tokyo (Lectures in Mathematical Sciences, vol. 5)

Faddeev, L.D. & Takhtajan, L.A. 1987. *Hamiltonian Methods in the Theory of Solitons*, Berlin and Heidelberg: Springer

Flaschka, H. & Newell, A.C. 1980. Monodromy and spectrum preserving deformations I. *Communications in Pure and Applied Mathematics*, 76: 67–116

Its, A. 2003. The Riemann–Hilbert problem and integrable systems. *Notices of the AMS*, 50: 1389–1400

Its, A.R., Izergin, A.G., Korepin, V.E. & Slavnov, N.A. 1990. Differential equations for quantum correlation functions. *International Journal of Physics* B4: 1003–1037; The quantum correlation function as the τ function of classical differential equations. In *Important Developments in Soliton Theory*, edited by, A.S. Fokas & V. E. Zakharov, Springer, Berlin, pp. 404–417, 1993

Jimbo, M., Miwa, T. & Ueno, K. 1981. Monodromy preserving deformation of linear ordinary differential equations with rational coefficients. I. *Physica* D, 2: 306–352

Muskhelishvili, N.I. 1946. *Singular Integral Equations*, Moscow (in Russian); translation by Groningen. Leiden: P. Noordhoff, 1953

Zakharov, V.E. & Shabat, A.B. 1971. Exact theory of two-dimensional self-focusing and one-dimensional self–modulation of waves in nonlinear media. *Zhurnal Eksperimental'noi Teoreticheskoi Fiziki*, 61: 118–134 [Russian]; translated in *Soviet Physics, JETP*, 34: 62–69 (1972)

Zhou, X. 1989. Direct and inverse scattering transforms with arbitrary spectral singularities. *Communications in Pure and Applied Mathematics*, 42, 895–938

Zhou, X. 1998. The L^2-Sobolev space bijectivity of the scattering and inverse scattering transforms. *Communications in Pure and Applied Mathematics*, 51: 697–731

RING SOLITONS

See **Solitons, types of**

ROBUSTNESS

See **Stability**

ROSSBY WAVES

See **Atmospheric and ocean sciences**

RÖSSLER SYSTEMS

Rössler systems were introduced in the 1970s as prototype equations with the minimum ingredients for continuous-time chaos.

Since the Poincaré–Bendixson theorem precludes the existence of other than steady, periodic, or quasi-periodic attractors in autonomous systems defined in one- or two-dimensional manifolds such as the line, the circle, the plane, the sphere, or the torus (Hartman, 1964), the minimum dimension for chaos is three. On this basis, Otto Rössler came up with a series of prototype systems of ordinary differential equations in three-dimensional phase spaces (Rössler, 1976a,c, 1977a, 1979a). He also proposed four-dimensional systems for hyperchaos, that is, chaos with more than one positive Lyapunov exponent (Rössler, 1979a,b).

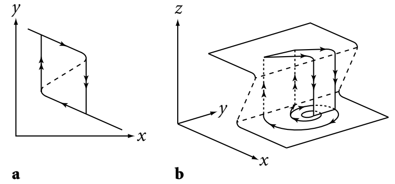

Figure 1. Illustration of the reinjection principle between the two branches of a Z-shaped slow manifold allowing (a) periodic relaxation oscillations in dimension two and (b) higher types of relaxation behavior in dimension three.

Rössler was inspired by the geometry of flows in dimension three and, in particular, by the reinjection principle, which is based on the feature of relaxation-type systems to often present a Z-shaped slow manifold in their phase space. On this manifold, the motion is slow until an edge is reached whereupon the trajectory jumps to the other branch of the manifold, allowing not only for periodic relaxation oscillations in dimension two (see Figure 1a), but also for higher types of relaxation behavior (see Figure 1b) as noted by Rössler (1979a). In dimension three, the reinjection can induce chaotic behavior if the motion is spiraling out on one branch of the slow manifold (see Figure 1b). In this way, Rössler invented a series of systems, the most famous of which is probably (Rössler 1979a):

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= -y - z,\\ \frac{\mathrm{d}y}{\mathrm{d}t} &= x + ay,\\ \frac{\mathrm{d}z}{\mathrm{d}t} &= bx - cz + xz.\end{aligned} \tag{1}$$

This system is minimal for continuous chaos for at least three reasons: its phase space has the minimal dimension three, its nonlinearity is minimal because there is a single quadratic term, and it generates a chaotic attractor with a single lobe, in contrast to the Lorenz attractor which has two lobes. In Equation (1), (x, y, z) are the three variables that evolve in the continuous time t, and (a, b, c) are three parameters. The linear terms of the two first equations create oscillations in the variables x and y. These oscillations can be amplified if $a > 0$, which results into a spiraling-out motion. The motion in x and y is then coupled to the z variable ruled by the third equation, which contains the nonlinear term and which induces the reinjection back to the beginning of the spiraling-out motion.

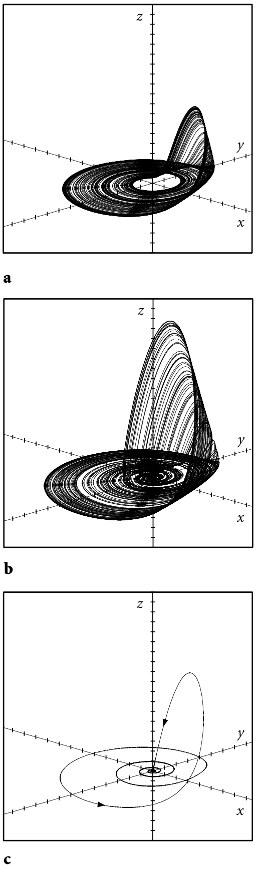

Figure 2. Phase portraits of the Rössler system (1) in the phase space of the variables (x, y, z): (a) spiral-type chaos for $a = 0.32$, $b = 0.3$, and $c = 4.5$; (b) screw-type chaos for $a = 0.38$, $b = 0.3$, and $c = 4.820$, in which case there also exists the Shil'nikov-type homoclinic orbit (c). The ticks are separated by unity. (Adapted from Gaspard & Nicolis, 1983.)

System (1) possesses two steady states: one at the origin $x = y = z = 0$, around which the motion spirals out, and another one at some distance from the origin due to the quadratic nonlinearity. This system presents stationary, periodic, quasi-periodic, and chaotic attractors depending on the value of the parameters (a, b, c). These attractors are interconnected by bifurcations, in particular, a Hopf bifurcation from the stationary to periodic attractors and a period-doubling cascade from periodic to chaotic attractors. The resulting chaotic attractor has a single lobe and is referred to as spiral-type chaos, which mainly manifests itself in irregular amplitudes for the oscillations (see Figure 2a).

A transition occurs to a screw-type chaos in which the oscillations are irregular not only in their amplitudes but also in the reinjection times (see Figure 2b). The screw-type chaos is closely related to the presence of a Shil'nikov homoclinic orbit (see Figure 2c). This homoclinic orbit is attached to the origin $x = y = z = 0$, which is a saddle-focus with a one-dimensional stable manifold for the reinjection and a two-dimensional unstable manifold where the motion is spiraling out. The Shil'nikov criterion for chaos is that the reinjection is faster than the spiraling-out motion (Shil'nikov, 1965), and it is satisfied in the attractor of Figure 2b. As a consequence, the homoclinic system contains periodic and nonperiodic orbits belonging to multiple horseshoes that can be described in terms of symbolic dynamics. Away from homoclinicity, the system undergoes complex bifurcation cascades generating successive periodic and chaotic attractors.

Chaotic behavior and Shil'nikov homoclinic orbits in the Rössler system (1) can also be understood as originating from an oscillatory-stationary double instability taking place around the origin $x = y = z = 0$ and the parameter values $b = 1$, $-\sqrt{2} < a = c < +\sqrt{2}$.

In his work on continuous chaos, Rössler was motivated by the search for chemical chaos, that is, chaotic behavior in far-from-equilibrium chemical kinetics (Ruelle, 1973; Rössler, 1976b, 1977b; Rössler & Wegmann, 1978). With Willamowski, Rössler proposed the following chemical reaction scheme:

$$\begin{aligned} A_1 + X &\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} 2X, \\ X + Y &\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} 2Y, \\ A_5 + Y &\underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} A_2, \\ X + Z &\underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} A_3, \\ A_4 + Z &\underset{k_{-5}}{\overset{k_5}{\rightleftharpoons}} 2Z, \end{aligned} \tag{2}$$

which features two autocatalytic steps (reactions 1 and 5) involving the species X and Z coupled to another autocatalytic step (reaction 2) involving another species Y and two further steps (reactions 3 and 4) (Willamowski & Rössler, 1980). The concentrations of the species $A_1, \ldots, A_5$ are held fixed by large chemical reservoirs that maintain the system out of thermodynamic equilibrium. The time evolution of the concentrations (x, y, z) of the three intermediate species X, Y, and Z is ruled by a system of three coupled differential equations deriving from the mass action law of chemical kinetics. These equations have quadratic nonlinear terms because of the binary reactive steps and keep the concentrations positive as a consequence of mass action kinetics. The chemical reaction scheme (2) leads to a chaotic attractor very similar to that of the abstract system (1) (see Figure 3), and thus provides a mechanistic understanding of chemical chaos in terms of colliding and reacting particles.

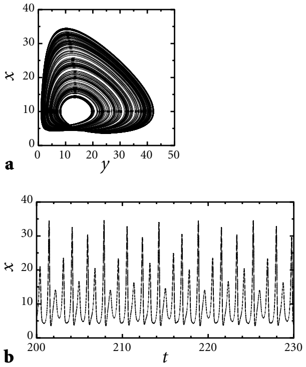

Figure 3. Chaotic time evolution of the concentrations of the Willamowski–Rössler chemical reaction scheme (2) for $k_1a_1 = 30$, $k_{-1} = 0.5$, $k_2 = 1$, $k_{-2} = 0$, $k_3a_5 = 10$, $k_{-3} = 0$, $k_4 = 1$, $k_{-4} = 0$, $k_5a_4 = 16.5$, and $k_{-5} = 0.5$: (a) phase portrait in the plane of the concentrations of species X and Y; (b) concentration of species X versus time t.

In conclusion, Rössler systems are minimal models for continuous-time chaos. The chaotic attractors of Rössler systems are prototypes for a large variety of chaotic behavior, notably, in chemical chaos (Scott, 1991).

P. GASPARD

See also **Attractors; Bifurcations; Brusselator; Chaotic dynamics; Chemical kinetics; Hopf bifurcation; Horseshoes and hyperbolicity in dynamical systems; Invariant manifolds and sets; Period doubling; Phase space; Poincaré theorems; Symbolic dynamics**

Further Reading

Gaspard, P. & Nicolis, G. 1983. What can we learn from homoclinic orbits in chaotic dynamics? *Journal of Statistical Physics*, 31: 499–518

Hartman, P. 1964. *Ordinary Differential Equations*, New York: Wiley

Rössler, O.E. 1976a. An equation for continuous chaos. *Physics Letters* A, 57: 397–398

Rössler, O.E. 1976b. Chaotic behavior in simple reaction systems. *Zeitschrift Naturforschung* A, 31: 259–264

Rössler, O.E. 1976c. Different types of chaos in two simple differential equations. *Zeitschrift für Naturforschung A* 31: 1664–1670

Rössler, O.E. 1977a. Continuous chaos. In *Synergetics: A Workshop*, edited by H. Haken, New York: Springer, pp. 184–199

Rössler, O.E. 1977b. Chaos in abstract kinetics: Two prototypes. *Bulletin of Mathematical Biology*, 39: 275–289

Rössler, O.E. 1979a. Continuous chaos—four prototype equations. *Annals of the New York Academy of Science*, 316: 376–392

Rössler, O.E. 1979b. An equation for hyperchaos. *Physics Letters* A, 71: 155–157

Rössler, O.E. & Wegmann, K. 1978. Chaos in the Zhabotinskii reaction. *Nature*, 271: 89–90

Ruelle, D. 1973. Some comments on chemical oscillators. *Transactions of the New York Academy of Science*, 35: 66–71

Scott, S.K. 1991. *Chemical Chaos*, Oxford: Clarendon Press

Shil'nikov, L.P. 1965. A case of the existence of a countable number of periodic motions. *Soviet Mathematics Doklady*, 6: 163–166

Willamowski, K.-D. & Rössler, O.E. 1980. Irregular oscillations in a realistic abstract quadratic mass action system. *Zeitschrift für Naturforschung* A, 35: 317–318

ROTATING RIGID BODIES

Rigid body motion about a fixed point is a classical mechanics problem associated with the greatest names in 18th-century mathematics. The basic equations of the motion were derived by Leonhard Euler in the 1750s and subsequently solved by him for two special cases (Euler and Lagrange tops). Development of the analytical theory of differential equations in the middle of the 19th century by Augustin Cauchy, Georg Riemann, and Karl Weierstrass inspired Sophia Kovalevsky (1889) to determine all the cases for which the integrals of rigid body dynamics were *single-valued*, *meromorphic* functions in the entire *complex* plane of the time variable. Pursuing this remarkable idea, she obtained a classification of all such (solvable) cases and found a new case, the Kovalevsky top, which she then integrated in quadratures. Considering time as a complex variable and imposing the above conditions upon the integrals of the equations of motion outside the real axis was a revolutionary approach to treating a mechanics problem. It appeared to be a fruitful method that led to a unified theory of (algebraically) completely integrable systems (Adler & van Moerbeke, 1989), a century after it was proposed.

The dynamics of a rigid body rotating about a fixed point can be considered in a fixed (nonmoving) frame of axes or in the moving body frame that has its origin at the fixed point and whose axes are along the principal axes of the ellipsoid of inertia. In 1750, Euler obtained the equations of motion in a fixed frame, but they appeared not to be very useful. In a series of papers during 1758–1765, he introduced the body frame, derived corresponding equations of motion, and also used Eulerian angles to relate the motion of a body frame with respect to a fixed system of coordinates.

Consider first a fixed frame $(\boldsymbol{i}', \boldsymbol{j}', \boldsymbol{k}')$. A fundamental dynamical theorem says that the time derivative of the angular momentum $\boldsymbol{J}$ of a body is equal to the moment $\boldsymbol{L}$ of the forces acting on it,

$$\frac{\mathrm{d}\boldsymbol{J}}{\mathrm{d}t} = \boldsymbol{L}. \tag{1}$$

In the body frame, $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$, the tensor of inertia is diagonal, $\boldsymbol{I} = \mathrm{diag}(A, B, C)$, and links the vector of angular velocity $\boldsymbol{\Omega} = p\,\boldsymbol{i} + q\,\boldsymbol{j} + r\,\boldsymbol{k}$ to the angular momentum $\boldsymbol{J}$ through the linear relation

$$\boldsymbol{J} = \boldsymbol{I\Omega} = Ap\,\mathbf{i} + Bq\,\boldsymbol{j} + Cr\,\boldsymbol{k}. \tag{2}$$

The time derivative of vector (2) is

$$\frac{\mathrm{d}\boldsymbol{J}}{\mathrm{d}t} = \frac{\delta \boldsymbol{J}}{\delta t} + \boldsymbol{\Omega} \times \boldsymbol{J}, \tag{3}$$

where $\frac{\delta \boldsymbol{J}}{\delta t}$ is the relative time derivative, evaluated assuming that the frame $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$ is stationary. The first three Euler equations, therefore, are

$$\begin{aligned} A\frac{\mathrm{d}p}{\mathrm{d}t} + (C - B)qr &= Mg(y_0\gamma'' - z_0\gamma'), \\ B\frac{\mathrm{d}q}{\mathrm{d}t} + (A - C)rp &= Mg(z_0\gamma - x_0\gamma''), \\ C\frac{\mathrm{d}r}{\mathrm{d}t} + (B - A)pq &= Mg(x_0\gamma' - y_0\gamma), \end{aligned} \tag{4}$$

where $\boldsymbol{Mg} = Mg(\gamma\,\boldsymbol{i} + \gamma'\,\boldsymbol{j} + \gamma''\,\boldsymbol{k})$ is the vector of the gravitational force and $\boldsymbol{r}_0 = x_0\,\boldsymbol{i} + y_0\,\boldsymbol{j} + z_0\,\boldsymbol{k}$ is the vector originating in the fixed point and pointing at the center of mass of the body. An extra set of three equations follows from the fact that the vector $\boldsymbol{Mg}$ is stationary in the fixed frame; hence, $\frac{\delta \boldsymbol{Mg}}{\delta t} = -\,\boldsymbol{\Omega} \times \boldsymbol{Mg}$ and

$$\begin{aligned} &\frac{\mathrm{d}\gamma}{\mathrm{d}t} = r\gamma' - q\gamma'', \qquad \frac{\mathrm{d}\gamma'}{\mathrm{d}t} = p\gamma'' - r\gamma, \\ &\frac{\mathrm{d}\gamma''}{\mathrm{d}t} = q\gamma - p\gamma'. \end{aligned} \tag{5}$$

The Eulerian angles, namely, the angle of precession $0 \le \psi < 2\pi$, the angle of nutation $0 \le \vartheta \le \pi$, and the angle of the self rotation $0 \le \varphi < 2\pi$, define the body's position by fixing the moving frame with respect to the fixed one. We have (see Golubev, 1960)

$$\begin{aligned} &\gamma = \sin\varphi\,\sin\vartheta, \qquad \gamma' = \cos\varphi\,\sin\vartheta, \\ &\gamma'' = \cos\vartheta, \end{aligned} \tag{6}$$

$$\begin{aligned} &p = \dot{\psi}\sin\vartheta\,\sin\varphi + \dot{\vartheta}\cos\varphi, \\ &q = \dot{\psi}\sin\vartheta\,\cos\varphi - \dot{\vartheta}\sin\varphi, \qquad r = \dot{\psi}\cos\vartheta + \dot{\varphi} \end{aligned} \tag{7}$$

and, therefore,

$$\frac{\mathrm{d}\psi}{\mathrm{d}t} = \frac{p\gamma + q\gamma'}{\gamma^2 + \gamma'^2}. \tag{8}$$

The problem of studying the motion of a rigid body about a fixed point is solved by integration of the system of differential equations (4), (5), and (8).

Conserved Quantities

It is always useful to know a system's first integrals, for instance for controlling numerical calculations or for performing integration in quadratures. There are three important physical integrals for an arbitrary top (an arbitrary set of the six parameters $\{A, B, C, x_0, y_0, z_0\}$ in Equations (4)). They are the following.

(i) Energy integral: $E = \frac{1}{2}\left(Ap^2 + Bq^2 + Cr^2\right) - Mg(x_0\gamma + y_0\gamma' + z_0\gamma'')$.
(ii) Length of the unit vector along $M\boldsymbol{g}$: $C_1 = \gamma^2 + \gamma'^2 + \gamma''^2 = 1$.
(iii) Projection of the angular momentum $\boldsymbol{J}$ on this vector: $C_2 = Ap\gamma + Bq\gamma' + Cr\gamma''$.

These three first integrals provide three constants of motion (since $C_1 = 1$, there are, in fact, only two arbitrary constants, the third one being the integration constant that appears when integrating (8), that is, the initial value of ψ). In addition, an autonomous system always has an extra constant corresponding to shifting of the time variable t. A remarkable feature of Equations (4) and (5) is that yet another explicit integration can be carried out using the method of Jacobi's last multiplier (Golubev, 1960), bringing in another explicit constant of integration. Therefore, for the case of a complete integration in quadratures of the equations of motion only one first integral is missing. Subsequent analysis showed that an extra algebraic integral exists only in the three special cases of Euler, Lagrange, and Kovalevsky tops. These integrable tops are given by specializing the parameters $\{A, B, C, x_0, y_0, z_0\}$.

Integrable Tops

The simpler cases of the Euler and Lagrange tops were studied, respectively, by Euler and Louis Poinsot and by Joseph-Louis Lagrange and Siméon Poisson, and later on by Gustav Jacobi, who expressed the general integrals for both systems in terms of elliptic functions of time.

Euler Top

In the Euler top $x_0 = y_0 = z_0 = 0$, and the body's center of mass is at the fixed point, so that there is no gravitational effect, meaning a free rotation of an asymmetric top. The fourth integral of motion in this case is the square of the angular momentum $\boldsymbol{J}$:

$$\boldsymbol{J}^2 = A^2p^2 + B^2q^2 + C^2r^2. \tag{9}$$

Poinsot gave a geometric interpretation of the motion based on the intersection of two quadratics: (9) and the energy integral $2E = Ap^2 + Bq^2 + Cr^2$. The vector $\boldsymbol{J}$ is stationary in the fixed frame $(\boldsymbol{i}', \boldsymbol{j}', \boldsymbol{k}')$ and the problem is made simpler if one chooses $\boldsymbol{k}'$ along $\boldsymbol{J}$. Then

$$\cos\vartheta = \frac{Cr}{|\boldsymbol{J}|} \text{ and } \tan\varphi = \frac{Ap}{Bq}. \tag{10}$$

As for time dependence, the variables p, q, and r can be expressed in terms of Jacobi's elliptic functions sn, cn, and dn of time and they are periodic, so are ϑ and φ obtained from (10). The dynamics of the angle of precession ψ is determined by integration of $\dot{\psi} = |\boldsymbol{J}|\,\frac{Ap^2 + Bq^2}{A^2p^2 + B^2q^2} > 0$, derived from (8). It follows that after each period, the value of ψ, in general, changes, so that the body never comes back to its initial orientation, thereby undergoing quasi-periodic motion.

Stationary rotations are those when the vector of angular velocity $\boldsymbol{\Omega}$ is constant. They correspond to uniform rotations about the principal axes of inertia. Rotation around the middle axis is unstable.

A symmetric $(A = B)$ Euler top performs regular precession. The body's symmetry axis draws a circular conic with the axis $\boldsymbol{J}$ and the angle $2\vartheta = \text{const}$. The symmetry axis goes around $\boldsymbol{J}$ with a constant angular velocity while the body rotates with a constant angular velocity around the symmetry axis.

Lagrange Top

This describes the dynamics of a symmetric body, $A = B$, with the fixed point at the symmetry axis and off the center of mass, that is, $x_0 = y_0 = 0$, $z_0 \neq 0$. The fourth constant of motion is simply r. The integration of this top can again be done in terms of elliptic functions, leading to a variety of complicated quasi-periodic motions including pseudo-regular precession, double-asymptotic, and "sleeping" tops. Simpler motions correspond to degenerations of elliptic functions into elementary functions.

The motion of the symmetry axis, described by the angles ϑ and ψ, looks like a perturbation of the corresponding uniform rotation of a symmetric Euler top in the regular precession. Illustrations of this familiar motion are available in many textbooks. Notice that now $\dot{\psi}$ can change sign, thereby giving three kinds of trajectories.

Denote $m = 1 - C/A$ and notice that $m = 0$ for a spherically symmetric body $(A = B = C)$. Consider the Lagrange top with $m \neq 0$ and the variables $(p, q, r, \gamma, \gamma', \gamma'')$ and a fully symmetric Lagrange top with $m = 0$ and the variables $(p_1, q_1, r_1, \gamma_1, \gamma_1', \gamma_1'')$. It is possible to show that the motions of these two bodies coincide when the variables are identified as follows:

$$\begin{aligned} &p_1^2 + q_1^2 = p^2 + q^2, \quad \tan^{-1}\tfrac{q_1}{p_1} = \tan^{-1}\tfrac{q}{p} + mrt, \\ &r_1 = \tfrac{C}{A}\,r, \\ &\varphi_1 = \varphi - mrt, \qquad \vartheta_1 = \vartheta, \\ &\quad \psi_1 = \psi. \end{aligned} \tag{11}$$

See Golubev (1960) for details.

Kovalevsky Top

This is the third general case of complete integrability of the Euler equations (4) and (5), which is characterized by restricting $A = B = 2C$ and $y_0 = z_0 = 0$. Hence, the Kovalevsky top is a special symmetric top whose center of mass lies in the equatorial plane of the ellipsoid of inertia ($x_0 \neq 0$). The fourth constant of motion is of order 4:

$$K = \left(p^2 - q^2 + \tfrac{Mgx_0}{C}\gamma\right)^2 + \left(2pq + \tfrac{Mgx_0}{C}\gamma'\right)^2 . \tag{12}$$

Kovalevsky found this case (Kovalevsky, 1889) and integrated it in terms of hyper-elliptic functions. Later on, Kötter (1893) simplified the formulae.

There is a great difference in complexity between this top and the other two. This is why there was such a long period between the discovery and integration of Euler and Lagrange tops and that of Kovalevsky. The Paris Academy of Sciences established a special Borden Prize to promote major expansions of the theory. It was finally claimed in 1888 by Sophia Kovalevsky who made major progress toward solution of the problem. Her work required usage of theta-functions in two variables and provided an enormous boost to the creation of the modern theory of completely integrable systems. There is a large body of literature about the Kovalevsky top, of which we mention only three items. In Adler & van Moerbeke (1988), a detailed study of the algebraic geometry of the model is given. In Bobenko et al. (1989), the authors find alternative theta-function formulae by making use of the corresponding Lax matrix and the finite-gap integration technique. In Kuznetsov (2002), connections with a representation of the quadratic r-matrix algebra and with the method of separation of variables are presented.

All other general tops corresponding to other choices of the six parameters (A, B, C, x_0, y_0, z_0) are not integrable, which means that they generally exhibit a chaotic dynamics, and it is impossible to find analytic solutions valid for arbitrary initial conditions.

VADIM B. KUZNETSOV

See also **Constants of motion and conservation laws; Integrability; Newton's laws of motion; Nonlinear toys**

Further Reading

Adler, M. & van Moerbeke, P. 1988. The Kowalevski and Hénon–Heiles motions as Manakov geodesic flows on SO(4)—a two-dimensional family of Lax pairs. *Communications in Mathematical Physics*, 113: 659–700

Adler, M. & van Moerbeke, P. 1989. The complex geometry of the Kowalevski–Painlevé analysis. *Inventiones Mathematicae*, 97: 3–51

Bobenko, A.I., Reyman, A.G. & Semenov-Tian-Shansky, M.A. 1989. The Kowalevski top 99 years later: a Lax pair, generalizations and explicit solutions. *Communications in Mathematical Physics*, 122: 321–354

Golubev, V.V. 1960. *Lectures on Integration of the Equations of Motion of a Rigid Body about a Fixed Point*, Moscow, 1953. Published for the National Science Foundation by the Israel Program for Scientific Translations

Kötter, F. 1893. Sur le cas traité par Mme Kowalevski de rotation d'un corps solide autour d'un point fixe. *Acta Mathematica*, 17(1–2)

Kovalevsky, S. 1889. Sur le problème de la rotation d'un corps solide autour d'un point fixe. *Acta Mathematica*, 12: 177–232

Kuznetsov, V.B. 2002. Kowalevski top revisited. In *The Kowalevski Property*, edited by V.B. Kuznetsov, Providence, RI: American Mathematical Society

ROTATING WAVE APPROXIMATION

Energy conserving oscillators comprise pairs of energetic variables (let us call them P and Q), each being the cause of the other. In a mechanical (spring-mass) oscillator shown in Figure 1, for example, Q would be the displacement of a mass and P its momentum (mass times velocity). For a radio engineer's tank oscillator, Q is the voltage across a capacitor and P is the current through an inductor. In a laser mode, Q is electric field energy and P is magnetic field energy.

Consider the mechanical oscillator of Figure 1, assuming the spring to be linear. From Newton's second law (force equals time derivative of momentum), the dynamic equation describing this system is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(M\frac{\mathrm{d}x}{\mathrm{d}t}\right) = -Kx, \tag{1}$$

where x is the vertical displacement of the mass from its resting position and M is the oscillator mass. On the right-hand side of this equation is the vertical force acting on the mass caused by extension or compression of the spring, and on the left-hand side is the time derivative of the vertical momentum: $p \equiv M\mathrm{d}x/\mathrm{d}t$. In this case, a solution is evidently $x = a\sin(\omega_0 t)$, where a is an arbitrary amplitude and $\omega_0 = \sqrt{K/M}$ is the frequency of oscillation.

Often the force of a spring is not quite a linear function of its extension but slightly sublinear. (This

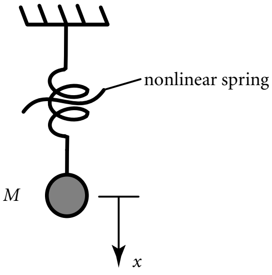

Figure 1. A spring-mass oscillator.

is the case, for example, in molecular vibrations, where the force between a pair of atoms becomes relatively weaker as the distance between them is increased because the electronic contribution to interatomic bonding is lessened.) In such cases, it is convenient to introduce anharmonicity into the formulation through the rotating wave picture (Louisell, 1960; Scott, 2003).

In the rotating wave picture of a linear oscillator, the momentum (p) and extension (x) are combined as real and imaginary parts of a single, complex amplitude

$$A \equiv \frac{(p - \mathrm{i}M\omega_0 x)}{\sqrt{2M\omega_0}}, \tag{2}$$

which obeys the first order equation

$$\mathrm{i}\frac{\mathrm{d}A}{\mathrm{d}t} = \omega_0 A \tag{3}$$

with the solution $A(t) = A(0)\mathrm{e}^{-\mathrm{i}\omega_0 t}$. Under this formulation, the total energy of the oscillator is

$$H = \omega_0 |A|^2 = \frac{1}{2}\left(\frac{p^2}{M} + Kx^2\right), \tag{4}$$

where the first term is the kinetic energy of the moving mass and the second is the potential energy stored in the spring.

If the restoring force of the spring is slightly sublinear, its potential energy may depend upon x as

$$\text{potential energy} = \tfrac{1}{2}Kx^2 - \tfrac{1}{4}\alpha x^4, \tag{5}$$

where $\alpha > 0$ is an anharmonicity parameter. With $\alpha = 0$, $x = \mathrm{i}\,(A - A^*)/\sqrt{2M\omega_0}$, so

$$x^4 = \frac{1}{4M^2\omega_0^2}\left[A^4 - 4A^3A^* + 6|A|^4 \right.$$
$$\left. - 4A(A^*)^3 + (A^*)^4\right]. \tag{6}$$

Except for the middle ($6|A|^4$) term, each term in Equation (6) varies sinusoidally with time at frequencies $\pm 4\omega_0$ or $\pm 2\omega_0$. Thus, the average of the energy over a cycle of oscillation is

$$\langle H \rangle = \omega_0 |A|^2 - \frac{\gamma}{2}|A|^4, \tag{7}$$

where $\gamma = 3\alpha/4M^2\omega_0^2$.

Under the rotating wave approximation (RWA), the energy H is taken as equal to its time average $\langle H \rangle$, whereupon the complex amplitude A is governed by the first order ODE

$$\mathrm{i}\frac{\mathrm{d}A}{\mathrm{d}t} \doteq \omega_0 A - \gamma |A|^2 A. \tag{8}$$

In this expression, the symbol "$\doteq$" indicates that only terms of frequency ω_0 are included. In other words, all nonresonant terms are neglected in the RWA.

Since Equation (8) conserves $N = |A|^2$, a general solution of this equation is

$$A(t) = \sqrt{N}\mathrm{e}^{-\mathrm{i}[(\omega_0 - \gamma N)t + \varphi]}, \tag{9}$$

where φ is an arbitrary phase angle. Equation (9) gives the frequency ($\omega = \omega_0 - \gamma N$) of the nonlinear oscillation as a function of its amplitude in the RWA.

Thus, RWA accounts for all components generated by a weak nonlinearity that resonate with the fundamental oscillator frequency. (As with your radio receiver, those components not in resonance are neglected.) Because the RWA is motivated by a general oscillator formulation, it is widely employed as the initial assumption in a variety of nonlinear analyses (various versions of the nonlinear Schrödinger equation, the discrete self-trapping equation, molecular vibrations, nonlinear optics, etc.). Interestingly, odd terms in the potential function (x^3, x^5, and so on) do not generate resonant components and are, thus, neglected in the RWA.

Finally, the RWA is a convenient formulation because quantum theories are easy to construct and solve for general systems of interacting oscillators, including discrete nonlinear Schrödinger equations and the discrete self-trapping system (Louisell, 1962; Scott et al., 1994; Scott, 2003). This is so for two reasons. First, under quantum theory in the RWA, the complex amplitude (A) becomes the lowering (annihilation) operator for oscillator quanta (bosons), and its complex conjugate (A^*) becomes the raising (creation) operator. Second, the classical fact that $N = \sum |A|^2$ is conserved implies that the corresponding quantum number operator commutes with the energy operator, allowing eigenfunctions of the energy operator to be constructed as sums of linear oscillator eigenfunctions.

Alwyn Scott

See also **Damped-driven anharmonic oscillator; Discrete nonlinear Schrödinger equations; Discrete self-trapping system; Quantum nonlinearity; Salerno equation**

Further Reading

Louisell, W.H. 1960. *Coupled Mode and Parametric Electronics*, New York: Wiley

Louisell, W.H. 1962. Correspondence between Pierce's coupled mode amplitudes and quantum oscillators. *Journal of Applied Physics*, 33: 2435–2436

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Scott, A.C., Eilbeck, J.C. & Gilhøj, H. 1994. Quantum lattice solitons. *Physica* D, 78: 194–213

ROTATION-MODIFIED KORTWEG-DE VRIES EQUATION

See **Korteweg–de Vries equation**

ROUTES TO CHAOS

If a nonlinear systems has chaotic dynamics, then it is natural to ask how this complexity develops as parameters vary. For example, in the logistic map

$$x_{n+1} = rx_n(1 - x_n), \qquad (1)$$

it is easy to show that if $r = \frac{1}{2}$, then there is a fixed point at $x = 0$ that attracts all solutions with initial values x_0 between 0 and 1, while if $r = 4$, the system is chaotic. How, then, does the transition to chaos occur as the parameter r varies? Indeed, is there a clean transition to chaos in any well-defined sense? The identification and description of routes to chaos has had important consequences for the interpretation of experimental and numerical observations of nonlinear systems. If an experimental system appears chaotic, it can be very difficult to determine whether the experimental data comes from a truly chaotic system, or if the results of the experiment are unreliable because there is too much external noise. Chaotic time series analysis provides one approach to this problem, but an understanding of routes to chaos provides another. In many experiments there are parameters (ambient temperature, Raleigh number, etc.) that are fixed in any realization of the experiment, but which can be changed. If recognizable routes to chaos are observed when the experiment is repeated at different values of the parameter, then there is a sense in which the presence of chaotic motion has been explained.

By the early 1980s, three "scenarios" or "routes to chaos" had been identified (e.g., Eckmann, 1981): Ruelle–Takens–Newhouse, period doubling, and intermittency (which has several variants). As we shall see, in their standard forms each of these transitions uses the term "route to chaos" in a different way, so care needs to be taken over the interpretation of experimental or numerical observations of these transitions.

Ruelle–Takens–Newhouse

In 1971, Ruelle and Takens published a mathematical paper with the provocative title "On the Nature of Turbulence." In this paper and a subsequent improvement with Newhouse (1978), they discuss the Landau scenario for the creation of turbulence by the successive addition of new frequencies to the dynamics of the fluid. They show that if the attractor of a system has three independent frequencies (four in the 1971 paper), then a small perturbation of this system has a hyperbolic strange attractor—a Plykin attractor (a solenoid in the 1971 paper). The result became known colloquially as "three frequencies implies chaos," a serious misinterpretation of the mathematical result that has been the cause of a number of misleading statements. First, the result proves the existence of chaos in systems arbitrarily close to the three frequency system in an infinite dimensional function space but gives no indication of the probability of finding chaos in any given example. Second, I know of no experimental situation where a Plykin attractor has been shown to exist even when chaotic behavior has been observed close to systems with three frequency attractors. Numerical experiments suggest that it is much more likely that the system evolves by frequency locking.

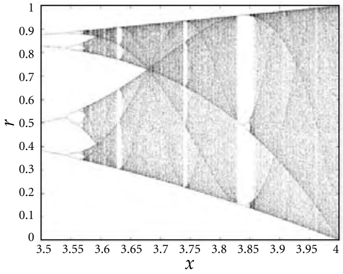

Figure 1. The attractor of the logistic equation as a function of the parameter r.

The Ruelle–Takens–Newhouse route to chaos remains somewhat of an enigma, and more work needs to be done to understand precisely how and when the strange attractors predicted by the theory come into being.

Period Doubling

Figure 1 shows the attractor of the logistic map (1) as a function of the parameter, r, for $3.5 < r < 4$. Thus, the set of points plotted on any vertical line of constant r represents the attractor of the map for that value of r, and if the set is finite, then the (numerically computed) attractor is a periodic orbit that cycles through the finite collection of points. Figure 1 suggests that for small r, the attractor is always periodic and has period 2^n, with n increasing as r increases. Beyond some critical value $r = r_c$, with $r_c \approx 3.569946$, the attractor may be more complicated. There are clearly intervals of r for which the attractor is periodic, and the attractor seems to be contained in 2^n bands that merge as r increases (the final band merging from 2 bands to one band with r just below 3.68 is particularly clear).

As r increases, the periodic orbit of period 2^n is created from the orbit of period 2^{n-1} by a period-doubling bifurcation. If this bifurcation occurs with $r = r_n$, then $r_n \to r_c$ geometrically as $n \to \infty$, with

$$\lim_{n\to\infty} \frac{r_n - r_{n-1}}{r_{n+1} - r_n} = \delta \approx 4.66920, \qquad (2)$$
$$\text{i.e.,} \quad r_n \sim r_c - \kappa\delta^{-n}.$$

The really surprising feature of this period-doubling cascade, as shown in Feigenbaum (1978), is that the cascade can be observed in many maps and the accumulation rate δ of the period-doubling cascade (2)

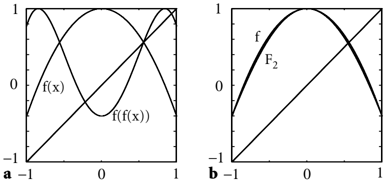

Figure 2. (a) The quadratic map $f(x) = 1 - mx^2$ restricted to the interval $[-1, 1]$ and the second iterate, $f(f(x))$ for $m = 1.40115$ which is just below the critical value m_c. (b) The original map, f, and the rescaled map, $F_2(x) = \mathcal{T} f(x) = -a^{-1} f(f(-ax))$ on $[-1, 1]$ with $a = -f(1) = m - 1$. Note that $a = 0.40115$ is close to the universal value of $\alpha \approx 0.3995$.

is the same although the constants r_c and κ depend on the map. In fact, the universal value of δ depends on the nature of the maximum of the map: $\delta \approx 4.66920$ for maps with quadratic maximum. A complete cascade of band merging from 2^n bands to 2^{n-1} bands occurs at parameter values $\tilde{r}_n$ above r_c, and $\tilde{r}_n \to r_c$ as $n \to \infty$ at the same universal geometric rate δ.

The quantitative universality in parameter space described by the scaling δ has a counterpart in phase space. If x_n denotes the point on the periodic orbit of period 2^{n-1} that is closest to the critical point (or turning point) of the map with $r = r_n$, then

$$\lim_{n\to\infty} \frac{x_{n+1} - \frac{1}{2}}{x_n - \frac{1}{2}} = -\alpha, \tag{3}$$

where α is another universal constant, which, for maps with a quadratic turning point, takes the value $\alpha \approx 0.3995 = 1/2.50\ldots$.

In families of one hump (unimodal) maps this universality can be explained by a renormalization argument. Restrict attention to families of one hump maps with critical point (maximum) at $x = 0$, parametrized by μ and normalized so that $f(0)c = 1$. As shown in Figure 2a, for parameter values near μ_c (the accumulation of period doubling) the second iterate of the map, $f(f(x))$, restricted to an interval about the critical point is a one hump map with a minimum. So, after a rescaling (and flipping) of the coordinates, it is another one hump map with the same normalization as shown in Figure 2b. Mitchell Feigenbaum was able to show (by arguments that have been made rigorous since 1990) that the universal properties described above are due to the structure of the doubling operator $\mathcal{T}$, which is a map on one hump maps $f : [-1, 1] \to [-1, 1]$, with critical point at $x = 0$ and $f(0)c1$, defined by

$$\mathcal{T} f(x) = -a^{-1} f(f(-ax)), \tag{4}$$

where $a = -f(1)$ so that the normalization $\mathcal{T} f(0) = 1$ is preserved. This operator does the rescaling and flipping referred to above. In the appropriate universality class, for example, quadratic critical point together with some further technical conditions, there is a fixed point f_*, of $\mathcal{T}$, so $f_* = \mathcal{T} f_*$, and the universal scaling of phase space is given by $\alpha = -f_*(1)$. Furthermore, the universal accumulation rate δ of (2) is an unstable eigenvalue of the (functional) derivative of $\mathcal{T}$ at f_*.

We can now consider measures of chaos such as the topological entropy or the Lyapunov exponents of the map. If $\{f_\mu\}$ is a family of one hump maps that undergoes period doubling, then the parameter μ can be chosen so that the period-doubling cascade is for $\mu < \mu_c$. Coullet & Tresser (1980) show that the universal structure described above implies that if $H(\mu)$ is either the topological entropy (which can be thought of as the growth rate of the number of periodic orbits) or the Lyapounov exponent of f_μ with $\mu > \mu_c$, then

$$H(\mu) \sim C(\mu - \mu_c)^{(\log 2)/(\log \delta)}. \tag{5}$$

The Lyapunov exponent is a very poorly behaved function of the parameters and this scaling provides only an envelope for the graph of the exponent, but the topological entropy is continuous. Indeed, the proof of Sharkovsky's theorem (*See* **One-dimensional maps**) shows that if a continuous map of the interval has a periodic orbit which is not a power of two, then there is a horseshoe for some iterate of the map, and hence the map has positive topological entropy if $\mu > \mu_c$. The entropy is zero if $\mu < \mu_c$, so if by chaos we mean positive topological entropy, then the period-doubling route is a true route to chaos.

Intermittency

The first stable periodic orbit of each of the windows of periodic motion in $r > r_c$, which can be seen in Figure 1, is created in a saddle-node (or tangent) bifurcation. Throughout the parameter interval for which such orbits are stable, there is a repelling strange invariant set, but most solutions tend to the stable periodic orbit. Just before the creation of the stable periodic orbit, chaotic solutions spend long periods of time near the points at which the stable periodic orbit will be created (the "laminar" phase), then move away and behave erratically before returning to the laminar phase. This behavior is called intermittency by Pomeau & Manneville (1980), who were the first to describe the scaling of the time spent in the laminar phase. They looked at the average time T_A spent by solutions in the laminar phase as a function of the parameter r close to the value r_{sn}, at which the saddlenode bifurcation occurs. A simple argument based on the passage time of a trajectory of a map close to a tangency with the diagonal (the condition for the saddle-node bifurcation) establishes that the average time in the laminar phase diverges as a power law:

$$T_A \sim |r - r_{sn}|^{-1/2}. \tag{6}$$

Other types of intermittency (involving period-doubling bifurcations, etc.) can be analyzed using the same ideas. Note that a strange invariant set exists throughout the parameter regions being considered here, so in this case the term "route to chaos" refers to the stability of the chaotic invariant set, not the creation of a chaotic set. Moreover, in any open neighborhood of r_{sn}, there are parameters for which the map has other stable periodic orbits, so the full description of parameters with stable chaotic motion is much more complicated than the description above suggests.

Other Routes to Chaos

Since the pioneering work of the late 1970s, a number of other routes to chaos have been identified. New routes to chaos are still being identified, and the list provided here is by no means complete. Arnéodo et al. (1981) show that there can be cascades of homoclinic bifurcations to chaos via a mechanism closely related to period doubling. This gives the less standard convergence rates involving nonquadratic turning points immediate relevance. The bifurcation that creates the strange invariant set of the Lorenz model is another type of homoclinic bifurcation, and this strange invariant set becomes stable by a "crisis" in which the strange invariant set collides with a pair of unstable periodic orbits. Ott (2002) contains a good account of such transitions. More complicated transitions involving maps of the circle are detailed in MacKay & Tresser (1986), and Newhouse et al. (1983) give another transition.

PAUL GLENDINNING

See also **Attractors; Bifurcations; Chaotic dynamics; Intermittency; Lorenz equations; One-dimensional maps; Period doubling; Time series analysis**

Further Reading

Arnéodo, A., Coullet, P. & Tresser, C. 1981. A possible new mechanism for the onset of turbulence. *Physics Letters* A, 81: 197–201

Coullet, P. & Tresser, C. 1980. Critical transition to stochasticity for some dynamical systems. *Journal de Physique Lettres*, 41: L255–L258

Eckmann, J.-P. 1981. Roads to turbulence in dissipative dynamical systems, *Reviews of Modern Physics*, 53: 643–654

Feigenbaum, M.J. 1978. Quantitative universality for a class of nonlinear transformations. *Journal of Statistical Physics*, 19: 25–52

MacKay, R.S. & Tresser, C. 1986. Transition to topological chaos for circle maps. *Physica* D, 19: 206–237

Newhouse, S., Palis, J. & Takens, F. 1983. Bifurcations and stability of families of diffeomorphisms. *Publications Mathématiques de l'IHES*, 57: 5–72

Newhouse, S., Ruelle, D. & Takens, F. 1978. Occurrence of strange Axiom A attractors near quasi-periodic flows on $\mathbf{T}^m$, $m \geq 3$. *Communications in Mathematical Physics*, 64: 35–40

Ott, E. 2002. *Chaos in Dynamical Systems*, 2nd edition, Cambridge and New York: Cambridge University Press

Pomeau, Y. & Manneville, P. 1980. Intermittent transition to turbulence in dissipative dynamical systems. *Communications in Mathematical Physics*, 74: 189–197

Ruelle, D. & Takens, F. 1971. On the nature of turbulence. *Communications in Mathematical Physics*, 20: 167–192

RUELLE–TAKENS–NEWHOUSE

See **Routes to chaos**

RUNGE–KUTTA METHOD

See **Numerical methods**

S

SADDLE POINT

See **Phase space**

SAFFMAN–TAYLOR PROBLEM

See **Hele-Shaw cell**

SALERNO EQUATION

The Salerno equation is a q-deformed lattice model that includes, as particular cases, two known discrete versions of the continuous nonlinear Schrödinger equation (NLS): the non-integrable discrete NLS equation (DNLS) with on-site nonlinearity and the integrable Ablowitz–Ladik (AL) equation with intra-site nonlinearity. Here by q-deformed, we mean the existence, both in the Poisson bracket (commutator in the quantum case) and in the Hamiltonian, of a free parameter q that allows "tuning" the nonlinearity (interaction) of the lattice model. From a physical point of view this equation represents a generalization of the tight-binding Schrödinger model

$$\mathrm{i}\frac{\mathrm{d}A_j(t)}{\mathrm{d}t} + \Omega_j A_j(t) + J_j\ \left(A_{j+1}(t) + A_{j-1}(t)\right) = 0, \qquad (1)$$

for the propagation of a molecular excitation in a crystal. Here A_j denotes the quasi-classical complex mode amplitude of a particular molecular vibration, Ω_j is the on-site frequency of this vibration, and J_j is the next-neighbor resonance interaction energy. Equation (1) was considered by Richard Feynman as a starting point for an alternate formulation of quantum mechanics in terms of coupled probability amplitudes (Feynman, 1965) (also it is equivalent to the Schrödinger equation for the description of the wave function of an electron in a perfect crystal in the tight-binding approximation). By assuming the coupling of the mode amplitudes to low-frequency phonons (lattice distortions), one obtains, in the adiabatic and small field amplitude approximation, a dependence of the local energy and intra-site interaction upon A_j of the type

$$\Omega_j \to \Omega_{0j} + \Omega_{1j}|A_j|^2, \\ J_j \to J_{0j} + J_{1j}|A_j|^2. \qquad (2)$$

The Salerno equation follows by substituting the above relations into Equation (1) and redefining parameters as

$$\varepsilon = \frac{\Omega_1}{J_0}, \eta = \frac{\Omega_1}{J_1}, \omega_j = \frac{2J_0 + \Omega_0}{J_0}, \gamma = \frac{2J_1 + \Omega_1}{J_0},$$

thus giving

$$\mathrm{i}\frac{\mathrm{d}A_j(t)}{\mathrm{d}t} - (2 - \omega_j - \varepsilon|A_j|^2)A_j \\ +(1 + \frac{\varepsilon}{\eta}|A_j|^2)(A_{j+1} + A_{j-1}) = 0, \qquad (3)$$

where time has been rescaled by a factor J_0 and equal local energies and intra-site resonance interactions were assumed (note that ω_j can be eliminated from Equation (3) by a rescale of time, so that in the following, we will put $\omega_j = 0$). With the above parametrization the following relationship between ε and η was introduced:

$$q \equiv \frac{\varepsilon}{\eta} = \frac{\gamma - \varepsilon}{2}. \qquad (4)$$

This allows interpolatation between the DNLS and the AL lattice as discussed below. To this end, it is worth noting that Equation (3) is a Hamiltonian system with respect to the following q-deformed Poisson bracket:

$$\{f, g\}_q = \sum \left(\frac{\partial f}{\partial A_j}\frac{\partial g}{\partial A_j^*} - \frac{\partial g}{\partial A_j}\frac{\partial f}{\partial A_j^*} \right) \\ \times (1 + q\,|A_j|^2), \qquad (5)$$

and with Hamiltonian

$$H = -\sum_j \left\{ A_j^*(A_{j+1} + A_{j-1}) + \eta|A_j|^2 \right. \\ \left. - \frac{(2+\eta)}{\log(1 + \frac{\varepsilon}{\eta})}\ \log(1 + \frac{\varepsilon}{\eta}|A_j|^2) \right\}. \qquad (6)$$

It is easy to check that Equation (3) is obtained from Equations (5) and (6), as $\mathrm{i}(\mathrm{d}A_j/\mathrm{d}t) = \{A_j, H\}_q$. Two values of the deformation parameter are of particular interest.

Case $q = 0$. This corresponds to the limit $\varepsilon \to \gamma$ for which the Poisson bracket acquires canonical form and Hamiltonian (6) becomes

$$-\sum_j \left\{ A_j^*(A_{j+1} + A_{j-1}) - 2|A_j|^2 - \frac{\gamma}{2}|A_j|^2) \right\}.$$

Equation (3) then reduces to the DNLS equation

$$\begin{aligned} &\mathrm{i}\frac{\mathrm{d}A_j(t)}{\mathrm{d}t} + A_{j+1} - 2A_j + A_{j-1} + \gamma A_j |A_j|^2 \\ &\quad = 0, \end{aligned} \tag{7}$$

which conserves the number $N = \sum_j |A_j|^2$ and allows solitary wave propagation (Scott, 2003). Although it is non-integrable, the DNLS equation is linked to many physical problems, from propagation of self-trapped modes in biomolecules and in arrays of nonlinear optical fibers, to the tight-binding description of Bose–Einstein condensates in optical lattices.

Case $q = \frac{\gamma}{2}$. This corresponds to the limit $\varepsilon \to 0$ for which Equation (3) reduces to the Ablowitz–Ladik equation

$$\begin{aligned} &\mathrm{i}\frac{\mathrm{d}A_j(t)}{\mathrm{d}t} + A_{j+1} - 2A_j + A_{j-1} + \frac{\gamma}{2} \\ &\times (A_{j+1} + A_{j-1})|A_j|^2 = 0, \end{aligned} \tag{8}$$

which is known to have exact soliton solutions, being integrable by the Inverse Scattering Method (Scott, 2003).

The property of Salerno's equation to incorporate the above discrete versions of the NLS equation makes it an ideal general model to investigate the interplay between on-site and intra-site nonlinearity, discreteness and continuum, integrability and non-integrability. Studies performed along these lines during the past decade have shown the existence of a rich and wide range of behaviors, ranging from the existence of states localized on few lattice sites (intrinsic localized modes or discrete breathers), to the possibility of shock wave formation (Cai et al., 1994; Kivshar & Salerno, 1994; Konotop & Salerno, 1997; Mackay & Sepulchre, 2002). The existence of an integrable limit of the model also allowed clarification of the relation between intrinsic localized modes and exact discrete solitons of the AL lattice.

Signatures of these classical properties exist in the corresponding quantum model whose Hamiltonian is defined from Equation (6) as

$$\begin{aligned} \hat{H} = -\sum_j \Bigg\{ &\hat{b}_j^\dagger(\hat{b}_{j+1} + \hat{b}_{j-1}) + \eta\, \hat{b}_j^\dagger \hat{b}_j \\ &- \frac{2+\eta}{\log(1+\frac{\varepsilon}{\eta})} \log[1 + \frac{\varepsilon}{\eta}\, \hat{b}_j^\dagger \hat{b}_j] \Bigg\}, \end{aligned} \tag{9}$$

where complex mode amplitudes A_j, A_j^* were replaced by creation and annihilation operators $\hat{b}_j$, $\hat{b}_j^\dagger$. Equation (5) implies the following deformed Heisenberg algebra:

$$\begin{aligned} [\hat{b}_i, \hat{b}_j^\dagger] &= \delta_{i,j}(1 + q\, \hat{b}_i^\dagger \hat{b}_i), \\ [\hat{b}_i, \hat{b}_j] &= [\hat{b}_i^\dagger, \hat{b}_j^\dagger] = 0, \end{aligned} \tag{10}$$

with the same q as in Equation (4). Note that the on-site algebra in Equation (10) is the same as the algebra of a q-oscillator (MacFarlane, 1989), thus providing an example of occurrence of a quantum deformation algebra in a physical model. An explicit representation of the q-algebra associated with the Salerno equation can be constructed as (Salerno, 1992; Bogoliubov & Bullough, 1992)

$$\begin{aligned} b^\dagger\,|n\rangle &= \sqrt{[n+1]_q}\,|n+1\rangle, \\ \hat{b}\,|n\rangle &= \sqrt{[n]_q}\,|n-1\rangle \end{aligned} \tag{11}$$

with $[n]_q = ((1+q)^n - 1)/q$. From this equation, it follows that the number operator is given by

$$\hat{N} = \sum_j \frac{\log(1 + \frac{\varepsilon}{\eta}\hat{b}_j^\dagger \hat{b}_j)}{\log(1 + \frac{\varepsilon}{\eta})}. \tag{12}$$

The conservation of $[\hat{N}, \hat{H}] = 0$, and the translational invariance of the system, $[\hat{T}_j, \hat{H}] = 0$ (T_j are translation operators by j sites along the lattice), allows one to block diagonalize the Salerno Hamiltonian into subspaces with fixed $\hat{N}$ eigenvalues and with fixed crystal momentum k. As for the classical model, the two limiting cases $\varepsilon = \gamma$ and $\varepsilon = 0$ correspond to the quantum DNLS lattice and to the quantum Ablowitz–Ladik system, respectively, the latter being integrable by means of the algebraic Bethe ansatz (Scott, 2003). By tuning the deformation parameter, one can get interesting physical behaviors. Thus, for example, for $q = -2$, the first commutator in Equation (10) becomes an anticommutator, so that the model describes a system of bosons with hard-core interactions (no more than one boson on a site). In this case, it was proved that in the mean field approximation (unconstrained hopping) and in the thermodynamic limit, the case $\varepsilon = 0$ displays a Bose–Einstein condensation (Salerno, 1994).

Exact diagonalizations of Equation (9) have been performed mainly for finite size and for finite number of quanta. Figure 1a depicts a typical band structure of Salerno's model in the reduced zone scheme for a chain of 15 sites, $N = 5$, $\gamma = 10$, and $\varepsilon = 5$. For

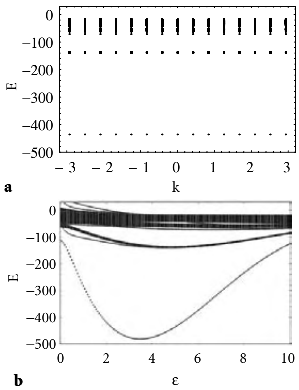

Figure 1. (a) Band structure of Salerno's model in the reduced zone scheme for a chain of 15 sites with $N = 5$, $\gamma = 10$, and $\varepsilon = 5$. (b) Energy of the states with crystal momentum $k = 0$, corresponding to the translational symmetry of the ground state, as a function of ε. Other parameters are fixed as (a).

these parameter values, the lower band corresponds to a bound state in which the five quanta are, with high probability, all on the same site. Figure 1b shows the energy of the states with crystal momentum $k = 0$ (corresponding to the translational symmetry of the ground state) as a function of ε. Note that the minimum ground state energy is achieved for an ε value in between the AL ($\varepsilon = 0$) and the DNLA case ($\varepsilon = 10$).

The presence of an integrable limit of the Salerno model gives the possibility of exploring complicated properties of this quantum many-body system starting from the exact knowledge of its integrable limit. An interesting open problem in this direction is the characterization of the quantum analogue of a discrete breather of the corresponding classical model (discrete quantum breather). Work in this direction is presently under investigation.

Mario Salerno

See also **Discrete nonlinear Schrödinger equations; Quantum nonlinearity**

Further Reading

Bogoliubov, N.M. & Bullough, R.K. 1992. A q-deformed completely integrable Bose gas model. *Journal of Physics* A, 25: 4057–4071

Cai, C., Bishop, A.R. & Grønbech-Jensen, N. 1994. *Physical Review Letters*, 72: 591

Feynman, R.P. 1965. *Lectures on Physics*, vol. III, Reading, MA: Addison-Wesley

Kivshar, Yu.S. & Salerno, M. 1994. *Physical Review* E, 49: 3543

Konotop, V.V. & Salerno, M. 1997. *Physical Review* E, 56: 3611

MacKay, R.S. & Sepulchre, J.A. 2002. *Journal of Physics* A, 35: 3985

MacFarlane, A.J. 1989. On q-analogues of the quantum harmonic oscillator and the quantum group SU$(2)_q$. *Journal of Physics* A, 22: 4581

Salerno, M. 1992. Quantum deformations of the discrete nonlinear Schrödinger equation. *Physical Review* A, 46: 6856

Salerno, M. 1994. Bose Einstein condensation in a system of q-bosons. *Physical Review* E, 50: 4528

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press, and references therein

SANDPILE MODEL

The concept of self-organized criticality (SOC) was introduced by Bak et al. (1987) using the example of a sandpile. If a sandpile is formed on a horizontal circular base with any arbitrary initial distribution of sand grains, a sandpile of fixed conical shape (steady state) is formed by slowly adding sand grains one after another (external drive). In the steady state, the surface of the sandpile makes on average a constant angle with the horizontal plane, known as the angle of repose. The addition of each sand grain results in some activity on the surface of the pile: an avalanche of sand mass follows, which propagates on the surface of the sandpile. In the stationary regime, avalanches are of many different sizes, and Bak et al. (1987) argued that they would have a power law distribution. If one starts with an initial uncritical state, initially most of the avalanches are small, but the range of sizes of avalanches grows with time. After a long time, the system arrives at a critical state, in which the avalanches extend over all length and time scales (Bak, 1996; Jensen, 1998; Dhar, 1999; Sornette, 2004).

Laboratory experiments on sandpiles, however, have not in general shown evidence of criticality in sandpiles due to the effects of inertia and dilatation (moving grains require more space) (Nagel, 1992), except for small avalanches (Held et al., 1990) or with elongated rice grains (Malte-Sørensen et al., 1999) where these effects are minimized. Small avalanches have small velocities (and thus negligible kinetic energy), and they activate only the first surface layer of the pile. Elongated rice grains slip at or near the surface as a result of their anisotropy (thus minimizing dilatational effects), and they also build up scaffold-like structures, which enhance the threshold nature of the dynamics.

On the theoretical front, a large number of discrete and continuous sandpile models have been studied. Among them, the so-called abelian sandpile model is the simplest and most popular (Dhar, 1999). Other variants include Zhang's model, which has modified rules for sandpile evolution (Zhang, 1989), a model for abelian distributed processors and other stochastic rule models (Dhar, 1999), and the Eulerian Walkers model (Priezzhev et al., 1996).

In the abelian sandpile model, each lattice site is characterized by its height h. Starting from an arbitrary initial distribution of heights, grains are added one at a time at randomly selected sites n: $h_n \to h_n + 1$. The sand column at any arbitrary site i becomes unstable when h_i exceeds a threshold value h_c and topples to reduce its height to $h_i \to h_i - 2d$, where d is the space dimension of the lattice. The $2d$ grains lost from the site i are redistributed on the $2d$ neighboring sites $\{j\}$, which gain a unit sand grain each: $h_j \to h_j + 1$. This toppling may make some of the neighboring sites unstable. Consequently, these sites will topple themselves, possibly making further neighbors unstable. In this way, a cascade of topplings propagates, which finally terminates when all sites in the system become stable. When this avalanche has stopped, the next grain is added on a site chosen randomly. This condition is equivalent to assuming that the rate of adding sand is much slower than the natural rate of relaxation of the system. The large separation of the driving and of the relaxation time scales is usually considered to be a defining characteristic of SOC. Finally, the system must be open to the outside; that is, it must dissipate energy or matter, for instance. An outcoming flux of grains must balance the incoming flux of grains, for a stationary state to occur. Usually, the outcoming flux occurs on the boundary of the system: even if the number of grains is conserved inside the box, it loses some grains at the boundaries. Even in a very large box, the effects of the dissipating boundaries are essential: increasing the box size will have the effect of lengthening the transient regime over which the SOC establishes itself; the SOC state is built from the long-range correlations that establish a delicate balance between internal avalanches and avalanches that touch the boundaries (Middleton & Tang, 1995).

The simplicity of the abelian model is that the final stable height configuration of the system is independent of the sequence in which sand grains are added to the system to reach this stable configuration (hence the name "abelian" referring to the mathematical property of commutativity). On a stable configuration $\mathcal{C}$, if two grains are added, first at i and then at j, the resulting stable configuration $\mathcal{C}'$ is exactly the same as in the case where the grains were added first at j and then at i. In other sandpile models, where the stability of a sand column depends on the local slope or the local curvature, the dynamics is not abelian, since toppling of one unstable site may convert another unstable site to a stable site. Many such rules have been studied in the literature (Manna, 1991; Kadanoff et al., 1989).

An avalanche is a cascade of topplings of a number of sites created by the addition of a sand grain. The strength of an avalanche can be quantified in several ways:

- size (s): the total number of topplings in the avalanche,
- area (a): the number of distinct sites that toppled,
- lifetime (t): the duration of the avalanche, and
- radius (r): the maximum distance of a toppled site from the origin.

These four different quantities are not independent and are related to each other by scaling laws. Between any two such measures x, y belonging to the set $\{s, a, t, r\}$, one can define a mutual dependence by the scaling of the expectation of one quantity y as a function of the other x:

$$\langle y \rangle \sim x^{\gamma_{xy}}, \tag{1}$$

where γ_{xy} is called a critical exponent, index, or dimension. This equation quantifies a nonlinear generalized proportionality between the two observables x and y ($\ln\langle y \rangle$ is proportional to $\ln x$). The exponents γ_{xy} are related to one another, for example,

$$\gamma_{ts} = \gamma_{tr}\gamma_{rs}. \tag{2}$$

For the abelian sandpile model, it can be shown that the avalanche clusters cannot have any holes and in addition that $\gamma_{rs} = 2$ in two dimensions, that is,

$$\langle s \rangle \sim r^2. \tag{3}$$

In words, the size (number of toppling grains) of an avalanche is proportional to its surface. It has also been shown that $\gamma_{rt} = \frac{5}{4}$:

$$\langle t \rangle \sim r^{\frac{5}{4}}; \tag{4}$$

that is, the average duration $\langle t \rangle$ of an avalanche grows with its typical radius r faster than linearly. However, averages reflect only a part of the rich behavior of sandpile models. A significant information is provided by the full distribution function $P(x)$ for any measure $x \in \{s, a, t, r\}$. Associated with the above scaling laws (1) and (2), one often finds the finite size scaling form for $P(x)$:

$$P(x) \sim x^{-\tau_x} f_x\left(\frac{x}{L^{\sigma_x}}\right). \tag{5}$$

The exponent σ_x determines the variation of the cut-off of the tail of the distribution of the quantity x with the system size L. As long as $x < L^{\sigma_x}$, expression (5) describes a power law distribution of x, reflecting a self-similar structure of the set of avalanches. When x becomes comparable with L^{σ_x}, the function f_x ensures a fast fall-off of $P(x)$ describing the impact of the finite size L of the system on the statistics of the fluctuations of x. Scaling relations like $\gamma_{xy} = (\tau_x - 1)/(\tau_y - 1)$ connect any two measures. Scaling assumptions (5) for the avalanche sizes have not been demonstrated and may be open to doubt (Kadanoff et al., 1989). This seems to be due to the effect of rare large avalanches dissipating at the border, which strongly influence the statistics.

Many different sandpile models have been studied. However, the precise classification of various models into different universality classes in terms of their critical exponents is not yet available. Exact values of all the critical exponents of the most widely studied abelian model are still not known in two dimensions. Some effort has also been made toward the analytical calculation of avalanche size exponents (Ktitarev & Priezzhev, 1998). Blanchard et al. (2000) have developed a dynamical system theory for a certain class of SOC models (like Zhang's model, 1989), for which the whole SOC dynamics can either be described in terms of iterated function systems, or as a piecewise hyperbolic dynamical system of skew-product type where one coordinate encodes the sequence of activations. The product involves activation (corresponding to a kind of Bernouilli shift map) and relaxation (leading to contractive maps).

In summary, the sandpile model of Bak et al. (1987) and its many extensions have helped found the new concept of self-organized criticality, which is now a useful item in the toolbox and set of concepts used to study complex systems involving triggered activities.

Didier Sornette

See also **Avalanches; Critical phenomena; Granular materials; Fractals; Nonequilibrium statistical mechanics**

Further Reading

Bak, P. 1996. *How Nature Works: The Science of Self-organized Criticality*, Berlin and New York: Springer

Bak, P., Tang, C. & Weisenfeld, K. 1987. Self-organized criticality: an explanation of $1/f$ noise. *Physical Review* A, 38: 364–374

Blanchard, P., Cessac, B. & Kruger, T. 2000. What can one learn about self-organized criticality from dynamical systems theory? *Journal of Statistical Physics*, 98: 375–404

Dhar, D. 1999. The abelian sandpile and related models. *Physica* A, 263: 4–25

Held, G.A., Solina, D.H., Keane, D.T., Haag, W.J., Horn, P.M. & Grinstein, G. 1990. Experimental study of critical-mass fluctuations in an evolving sandpile. *Physical Review Letters*, 65: 1120–1123

Jensen, H.J. 1998. *Self-Organized Criticality*, Cambridge and New York: Cambridge University Press

Kadanoff, L.P., Nagel, S.R., Wu, L. & Zhou, S. 1989. Scaling and universality in avalanches. *Physical Review* A, 39: 6524–6537

Ktitarev, D.V. & Priezzhev, V.B. 1998. Expansion and contraction of avalanches in the two-dimensional abelian sandpile. *Physical Review* E, 58: 2883–2888

Malte-Sørensen, A., Feder, J., Christensen, K., Frette, V., Josang, T. & Meakin, P. 1999. Surface fluctuations and correlations in a pile of rice. *Physical Review Letters*, 83: 764–767

Manna, S.S. 1991. Critical exponents of the sandpile models in two dimensions. *Physica* A, 179: 249–268

Middleton, A.A. & Tang, C. 1995. Self-organized criticality in non-conserved systems. *Physical Review Letters*, 74: 742–745

Nagel, S.R. 1992. Instabilities in a sandpile. *Review of Modern Physics*, 64: 321–325

Priezzhev, V.B., Dhar, A., Krishnamurthy, S. & Dhar, D. 1996. Eulerian Walkers as a model of self-organized criticality. *Physical Review Letters*, 77: 5079–5082

Sornette, D. 2004. *Critical Phenomena in Natural Sciences*, 2nd edition, Berlin and New York: Springer

Zhang, Y.-C. 1989. Scaling theory of self-organized criticality. *Physical Review Letters*, 63: 470–473

SCATTERING OPERATORS

See **Inverse scattering method or transform**

SCHEIBE AGGREGATES

Scheibe aggregates (also called J-aggregates) are a type of Langmuir–Blodgett thin films that were independently discovered in 1936 by Günther Scheibe (Scheibe, 1936) in Germany and E.E. Jelly (Jelly, 1936) in England—hence their names. They comprise compact and regular arrangements of dye molecules composed of chromophores and fatty acids that can be designed to pre-selected specifications. These molecular monolayers exhibit very efficient light capture followed by energy transfer to acceptor molecules present in the film at very low acceptor-to-donor ratios (as low as 1:10,000), giving the efficiency of energy transfer to acceptors as high as 50% (Moebius, 1989).

In the late 1960s and in the 1970s, Kuhn, Moebius and co-workers (Czikkely et al., 1969) studied the effects of irradiation with ultraviolet or visible light and measured the strongly quenched donor fluorescence in these aggregates. An acceptor fluorescence line appeared whose amplitude was almost equal to that of the primary donor spectral line, but its peak was slightly redshifted, which was interpreted as evidence that the aggregate acts as a cooperative molecular array. After absorbing a photon, the energy moves in the form of an exciton traveling laterally over distances of up to 100 nm to a particular acceptor dye in the vicinity. The efficiency of this energy transfer mechanism is strongly linked with the rigidity of the aggregate and its regular order. Surprisingly, the capture of energy by an acceptor molecule improves with the ambient temperature. Optimal efficiency is achieved at fairly low acceptor-to-donor ratios as mentioned above.

Numerous technological applications of these molecular systems have been developed, including photographic and photo-detection processes as well as solar energy cell components (Inoue, 1985).

Several theoretical models of Scheibe aggregates have been developed over the years. The Frenkel exciton picture with the inclusion of diagonal disorder in molecular chains with nearest-neighbor interactions was adopted to describe the pseudo-isocyanine aggregate (Knapp, 1984) and was later extended to account for off-diagonal disorder. Subsequently, linear models of exciton propagation in the presence of shallow impurity potentials at acceptor sites showed (Bartnik et al., 1992) close quantitative agreement with experiment. These impurity potentials were usually assumed to be of a square-well type, and their depth and spatial extent were parameters subject to optimization schemes. Interestingly, shallow impurity levels gave better capture properties than deep ones.

There have also been a number of theoretical studies advocating a role played by nonlinearity in the propagation of excitonic modes in Scheibe aggregates. The main reason for this claim is due to exciton-phonon coupling whose strength for mero-cyanine was experimentally determined as approximately $\chi = 29$ meV (Inoue, 1985) and theoretically estimated as $\chi = 26$ meV (Spano et al., 1990).

In general, three types of excitonic models can be applied: (a) a nearly free delocalized case, (b) a small polaron limit, and (c) the nonlinear self-trapped exciton state. The applicability criterion for these models involves a phase diagram in terms of two characteristic parameters: $g = \pi \chi^2/\omega J$ and $\gamma = h\omega/2\pi J$, where h is Planck's constant, ω is the phonon frequency, and J is the exciton hopping constant. With the phonon energy approximately 30 meV, J in the range between 50 and 150 meV, and the exciton-phonon coupling constant ranging from 25 to 100 meV, the outcome of this parameter determination is yet inconclusive (Tuszyński et al., 1999).

It is important to stress that these systems should be modeled as two-dimensional structures. Continuous two-dimensional models with nonlinear terms have been developed that treat phonons classically and effectively eliminate them via an adiabatic approximation (Huth et al., 1989), leading to a 2-d radially symmetric cubic nonlinear Schrödinger equation, where the energy transfer takes place through soliton-like ring waves with a characteristic collapse time signifying the exciton lifetime. More recent models involving nonlinear equations of the cubic Schrödinger type also account for the presence of Gaussian impurity potentials (Christiansen et al., 1998).

Finally, it is necessary to include radiative losses in modeling the propagation of an exciton domain. Formally, this can be accomplished by adding an imaginary term to the exciton-phonon part of the Hamiltonian that describes the corresponding loss of energy of excitons as they collide with phonons in the thin-film lattice. Consequently, it has been shown that the rate of excitonic energy loss by a coherent exciton domain covering an area composed of N molecules is proportional to its size, N, and can be expressed as: $k_{\text{rad}} = N$ nanoseconds^{-1}. It has also been demonstrated that the characteristic time for radiative losses is approximately proportional to the absolute temperature T divided by 3000 K, and the result is expressed in nanoseconds (Moebius & Kuhn, 1988). In other words, the radiative decay time in nanoseconds is given by $\tau_{\text{rad}} = (T/3000)$ K. For example, at room temperature, a coherent exciton domain composed of 100 lattice sites has a decay time of $0.1\,\text{ns} = 100\,\text{ps}$ compared with a flight time of only 2 ps, which means that radiative losses will not destroy the coherence of the exciton domain unless it is very large (approximately 10,000 lattice sites). A comprehensive review of the key processes taking place in exciton energy transfer in Scheibe aggregates (time of flight, radiative losses, exciton-phonon interaction, and diffusion on the 2-d lattice) can be found in Tuszyński et al. (1999).

JACK A. TUSZYŃSKI

See also **Excitons; Langmuir–Blodgett films; Nonlinear Schrödinger equations**

Further Reading

Bartnik, E.A., Blinowska, K.J. & Tuszyński, J.A. 1992. Stability of quantum capture in Langmuir–Blodgett monolayers against positional disorder. *Physics Letters* A, 169: 46–50

Christiansen, P.L., Gaididei, Yu.B., Johansson, M., Rasmussen, K.O., Mezentsev, V.K. & Rasmussen, J.J. 1998. Solitary excitations in discrete two-dimensional nonlinear Schrödinger models with dispersive dipole-dipole interactions. *Physical Review* B, 57: 11–303

Czikkely, V., Dreizler, G., Försterling, H., Kuhn, H., Sondermann, J., Tillmann, P. & Wiegand, J. 1969. Lichtabsorption von Farbstoff-Molekülpaaren in Sandwichsystemen aus monomolekularen Schichten. *Zeitschrift für Naturforschung*, 24A: 1823

Czikkely, V., Försterling, H. & Kuhn, H. 1970. Extended dipole model for aggregates of dye molecules. *Chemical Physics Letters*, 6: 207

Huth, G.C., Gutmann, F. & Vitiello, G. 1989. Ring solitonic vibrations in Schiebe aggregates. *Physics Letters* A, 140: 339

Inoue, T. 1985. Optical absorption and luminescence in Langmuir films of merocyanine dye. *Thin Solid Films*, 132: 21

Jelly, E.E. 1936. Spectral absorption and fluorescence of dyes in the molecular state. *Nature*, 138: 1009–1010

Knapp, E.W. 1984. Lineshapes of molecular aggregates—exchange narrowing and intersite correlation. *Chemical Physics*, 85: 73

Kuhn, H. 1979. Synthetic molecular organizates. *Journal of Photochemistry*, 10: 111

Moebius, D. 1989. Monolayer assemblies. *Berichte der Bunsen-Gesellschaft—Physical Chemistry*, 82: 848

Moebius, D. & Kuhn, H. 1988. Energy transfer in monolayers with cyanine dye Scheibe aggregates. *Journal of Applied Physics*, 64: 5138–5141

Scheibe, G. 1936. Variability of the absorption spectra of some sensitizing dyes and its cause. *Angewandte Chemie*, 49: 567

Spano, F.C., Kuklinski, J.R. & Mukamel, S. 1990. Temperature dependent superradiant decay of excitons in small aggregates. *Physical Review Letters*, 65: 211

Tuszyński, J.A., Joergensen, M.F. & Moebius, D. 1999. Mechanisms of exciton energy transfer in Scheibe aggregates. *Physical Review* E, 59: 4374

SCHLESINGER EQUATIONS

See **Monodromy preserving deformations**

SCHOTTKY DIODE

See **Diodes**

SCHRÖDINGER EQUATION, LINEAR

See **Quantum nonlinearity**

SCROLL WAVES

Introduction

Scroll waves, which are observed in excitable media, can be imagined as a continuation of two-dimensional spiral waves to three dimensions. Spiral and scroll waves distribute frequency behavior to the two- and three-dimensional space, respectively, with the exception of a small subset of that space. For spiral waves this subset comprises the vicinity of a point and is called a core, whereas in three dimensions it is formed by the vicinity of a curve and is called a filament. In both cases, this exceptional subset acts as a source of waves that organizes the exhibited periodic or quasi-periodic frequency behavior; therefore, it is often referred to as an organizing center.

The intersection of the three-dimensional excitable medium containing a scroll wave with a plane that is locally perpendicular to the filament corresponds to a sheet in which a spiral wave rotates, showing the close relationship between these two types of excitation waves. But because curves (the filaments) can have complex geometrical and topological properties, the behavior of scroll waves turns out to be much more complicated than that of spirals.

Early work on rotating excitation waves was performed on physical rings of heart tissue (quasi-one-dimensional systems), where an electrical impulse can propagate around a hole without attenuation for days. The importance of this observation lies in the complex anatomy of the heart: in particular, the orifices from arteries make the surface of the heart locally similar to physical rings and, therefore, allow for the so called circus-movement reentry, that is, electrical excitation waves circulating around inhomogeneities. Modeling of this phenomenon was undertaken first by Norbert Wiener and Arturo Rosenblueth in 1946. Periodic excitation was found to be also possible in media without holes, that is, in quasi-two-dimensional media. The electrical excitation wave (see Figure 1) then assumes the shape of a spiral (Tyson & Keener, 1988). The importance of the third dimension was recognized when performing experiments with the excitable Belousov–Zhabotinsky reaction system (*See* **Belousov–Zhabotinsky reaction**), since distortions and other unexpected behavior of spiral waves in presumed two-dimensional shallow layers could be readily explained when assuming a non-uniform excitability along the small but nonnegligible vertical direction (see references in Winfree, 1987).

For the theoretical treatment of scroll waves there are several models that are most commonly used.

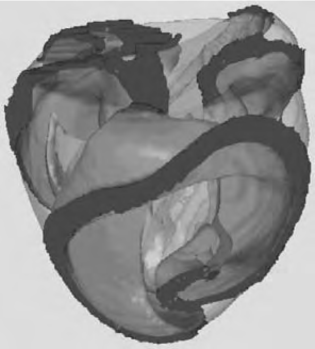

Figure 1. A scroll wave on the surface of a heart. (Image from Panfilov & Pertsov, 2001.)

Since scroll waves are structures appearing in excitable media, it is not surprising that they have been investigated by using three-dimensional FitzHugh–Nagumo equations, that is, a two-variable model designed for the simulation of neural action potentials (*See* **FitzHugh–Nagumo equation**). It was found that scroll waves always undergo twist (see below) when entering inhomogeneous media (Mikhailov et al., 1985).

Another model for the investigation of scroll waves is the Oregonator model derived from the reaction kinetics of the Belousov–Zhabotinsky reaction. The temporal development of a scroll ring, a structure with a closed filament not touching the boundary (see Figure 2, and also Welsh et al., 1983), was investigated by Winfree and Jahnke in 1989. It was found that a scroll ring decreases its size in the course of time and eventually vanishes, showing that scroll waves are topologically distinct from spirals.

A now very popular model is the so-called Barkley model, well known for its computational efficiency. Originally introduced for the investigation of the meander instability of spiral waves (Barkley, 1990), it is also used for investigating scroll waves. With this model it was possible to classify the instabilities of scroll waves in isotropic excitable media (Henry & Hakim, 2002).

While these three models give a description of the full three-dimensional concentration distribution, a fourth one deals with the geometry of an iso-concentration level. Its formulation is based on the eikonal equation, which expresses the relationship between the curvature and the normal velocity of a surface defined by such a level. Thus, the surface is described in specific coordinate systems that reflect the topological situation of the filaments involved in the

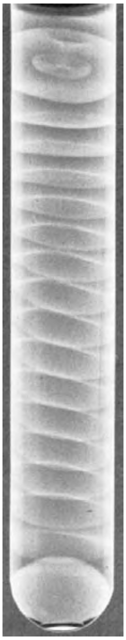

Figure 2. A scroll ring in a test tube, inner diameter: 10 mm. (Image from Winfree, 1987.)

wave structure. Investigating this type of model allows one to estimate the stability of complex, linked wave structures (McDermott et al., 2002).

Complexity of Scroll Waves

The complexity of a scroll wave becomes apparent when one imagines such a wave structure extending into a nonhomogenous excitable medium, that is, when in each slice along the filament the rotation period of the spirals is intrinsically different. Without a coupling mechanism, the phase of the spirals in these adjacent slices would evolve independently, and the differences between the phases would diverge. This is not possible, and the wave structure circumvents this inconvenience by tilting the wave fronts emanating from the organizing center. In terms of the rotation phases of the spirals, this corresponds to a phase shift along the filament, which is called twist.

While the twist takes arbitrary values in the case of scroll waves (organized by filaments reaching from one boundary of the medium to another), it has to fulfill quantization conditions for scroll rings. The first limitation is that the twist of a scroll ring along one turn along the filament has to be a multiple of 2π. A further limitation for the possible structures arises from the constraint that the surface must not have any intersections (i.e., it has to be a Seifert-surface, see Figure 3 for an impossible surface). This, for example, excludes single scroll rings of twist 2π. Instead, these always must appear as linked pairs, although in the limiting case one of them may be infinitely large or may reach from one boundary to the other (Winfree, 1987). More elaborate work on the topology and geometry of filaments has been summarized by Tyson and Strogatz (1991).

Although the computational investigation of scroll waves started in the 1980s, about ten years after their discovery, rigorous experimental research on

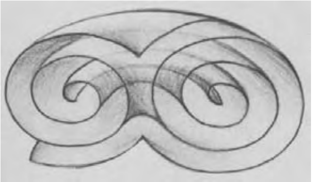

Figure 3. An example of an impossible surface of a wave structure. (Image from Winfree, 1987.)

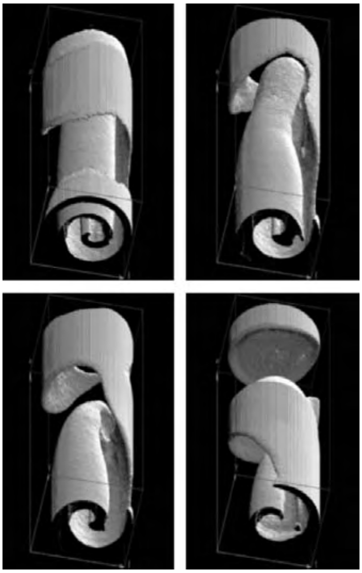

Figure 4. Decomposition of a scroll wave due to a gradient of excitability. (Image from Strob et al., 2003.)

three-dimensional wave structures did not begin before the 1990s. First measuring techniques were restricted to simple projections from one, two, or three pairwise orthogonal directions.

Three-dimensional, fully resolved observations of scroll waves and rings have been performed by optical tomography since about 1995. This technique allows one to record and evaluate time and space resolved data sequences. For instance, the decomposition of a scroll wave due to a gradient of excitability was observed in satisfactory detail (see Figure 4). Thus, experimental and theoretical tools are now available to investigate the complex interaction of the organizing centers of scroll waves.

ULRICH STORB AND STEFAN C. MÜLLER

See also **Belousov–Zhabotinsky reaction; Cardiac arrhythmias and electrocardiogram; Excitability; Geometrical optics, nonlinear; Reaction-diffusion systems; Spiral waves; Vortex dynamics in excitable media**

Further Reading

Barkley, D. 1990. Spiral-wave dynamics in a simple model of excitable media: the transition from simple to compound rotation. *Physical Review* A, 42(4): 2489–2492

Henry, H. & Hakim, V. 2002. Scroll waves in isotropic excitable media: linear instabilities, bifurcations and restabilized states. *Physical Review* E, 65: 046235-1–046235-21

McDermott, S., Mulholland, A.J. & Gomatam, J. 2002. Knotted reaction–diffusion waves. *Proceedings of the Royal Society of London Series A—Mathematical Physical and Engineering Sciences*, 458: 2947–2966

Mikhailov, A.S., Panfilov, A.V. & Rudenko, A.N. 1985. Twisted scroll waves in active three-dimensional media. *Physics Letters* A, 109(5): 246–250

Panfilov, A.V. & Pertsov, A. 2001. Ventricular fibrillation: evolution of the multiple-wavelet hypothesis. *Proceedings of the Royal Society of London Series* A, 359: 1315–1325

Storb, U., Rodrigues Neto, C., Bär, M. & Müller, S.C. 2003. A tomographic study of desynchronization and complex dynamics of scroll waves in an excitable chemical reaction with a gradient. *Physical Chemistry Chemical Physics*, 5: 2344–2353

Tyson, J.J. & Keener, J.P. 1988. Singular perturbation theory of travelling waves in excitable media. *Physica* D, 32: 327–361

Tyson, J.J. & Strogatz, S.H. 1991. The differential geometry of scroll waves. *International Journal of Bifurcation and Chaos*, 1: 723–744

Welsh, B.J., Gomatam, J. & Burgess, A.E. 1983. Three-dimensional chemical waves in the Belousov–Zhabotinskii reaction. *Nature*, 304: 611–614

Wiener, N. & Rosenblueth, A. 1946. The mathematical formulation of the problem of conduction of impulses in a network of connected excitable elements. *Specifically in Cardiac Muscle Archivos del Instituto de Cardiologica de Mexico*, 16: 205–265

Winfree, A.T. 1987. *When Time Breaks Down*, Princeton, NJ: Princeton University Press

Winfree, A.T. & Jahnke, W. 1989. Three-dimensional scroll ring dynamics in the Belousov-Zhabotinsky reagent and in the 2-variable Oregonator model. *Journal of Physical Chemistry*, 93: 2823–2832

SECOND HARMONIC GENERATION

See **Harmonic generation**

SELF-ORGANIZATION

See **Synergetics**

SELF-SIMILARITY

See **Fractals**

SEMICLASSICAL APPROXIMATIONS

See **Quantum theory**

SEMICONDUCTOR LASER

The semiconductor laser is today the most important and widespread type of laser, being a central component in many common household appliances (CD and DVD players) as well as in major industrial areas, such as measurement and sensing, materials manufacturing, and medical surgery. Not least, the semiconductor laser has enabled the rapid evolution of the Internet by providing a means for efficient and cheap conversion of digital electrical signals into optical signals, which can be transmitted at very high data rates and over very long distances in hair-thin optical fibers.

The success of the semiconductor laser to a large degree relies on its many similarities with electronic semiconductor devices such as transistors and diodes. The laser is manufactured by standard semiconductor crystal growth and processing techniques, allowing for small and cheap devices. Furthermore, the semiconductor laser distinguishes itself from other types of lasers by being electrically activated. Thus, the energy needed for pumping the laser to an excited state, from which the energy can be released by emission of photons, is achieved simply by putting direct electrical current through the device. Other kinds of lasers typically need some kind of optical pumping for populating the laser-active states.

Figure 1 shows a schematic of a typical semiconductor laser. The laser is a p-i-n structure; that is, p- and n-doped semiconductor materials sandwich an undoped intrinsic ("i") region. The structure acts as a standard pn-junction diode; when forward biased (with a positive voltage on the metallized p-side relative to the n-side), an electrical current flows. This leads to the build-up of significant electron and hole densities in the intrinsic region, and by recombination of these excited-state (conduction-band) electrons with ground-state (valence-band) holes, photons can be generated. An incoming photon may thus stimulate the emission of a new photon with identical properties. This process of stimulated emission provides optical gain, which is

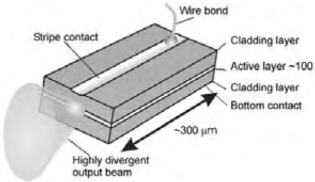

Figure 1. Schematic of a semiconductor laser. Light is confined to a waveguide of transverse dimension typically of the order of 0.2 μm × 2 μm. For a quantum well structure, stimulated emission occurs in a narrow region of width ∼ 100 Å. The laser is pumped electrically by incorporating the active region in a pn-junction.

one of the two key requirements for implementing a laser.

As for any oscillator, the second requirement—besides gain—is the existence of feedback. For a laser, this is usually achieved by incorporating the gain medium in a mirror cavity. In the case of a semiconductor laser, the mirrors are particularly simple since cleaving along one of the crystal planes provides a naturally flat mirror with an intensity reflection coefficient of the order of 30%. Due to the large material gain achievable in semiconductor lasers, a laser-active region length of the order of a few hundred μm is sufficient to compensate for the corresponding 70% outcoupling loss, as well as other losses in the material, thus enabling laser oscillation.

In addition to providing feedback, the laser cavity also must confine the optical laser mode in the transverse plane. This is achieved by establishing a transverse waveguide through index-guiding. The design thus needs to ensure a larger effective index in the active region of the laser as compared with the surrounding regions. In the growth direction, index guiding is provided by use of a so-called semiconductor heterostructure. The intrinsic, "i"-region is thus composed of a material with a smaller bandgap than the surrounding materials, which leads to a larger refractive index. The incorporation of a heterojunction structure was a major achievement in the early development of semiconductor lasers and earned the inventors, Zhores Alferov and Herbert Kroemer, the Nobel Prize in Physics in 2000. Besides providing index confinement, the heterostructure also ensures efficient collection of electrons and holes in the active region.

Index guiding in the plane (lateral direction) of the semiconductor layers is more difficult and is achieved in a number of different ways, the two most important ones employing a ridge waveguide structure and a buried heterostructure. The ridge waveguide structure is obtained by processing a narrow ridge (1–2 μm wide) in the doped semiconductor material topping the active region. The material that is etched away on either side of the ridge is replaced by a material (e.g., polyimide) that is isolating and has an index of refraction less than that of the semiconductor (ca. 3.5). The ridge provides current confinement and leads to an effective refractive index that varies in the lateral direction along the active region, reaching a maximum value right below the high-index ridge, thus imposing lateral waveguiding. The refractive index contrast thus obtained is, however, modest, and ridge waveguide lasers belong to the class of weakly index-guided structures. The other class of strongly index-guided structures is exemplified by the buried heterostructure laser. By employing several growth steps, the active waveguide region can thus be surrounded, in the lateral direction, by materials with higher bandgap and lower refractive index. Furthermore, these regions, adjacent to the active region, are doped to block the current from entering.

By analogy with a standard laser cavity (etalon), a semiconductor laser using cleaved facets to define the laser cavity is denoted a Fabry–Perot laser. Due to the small difference in material gain between the longitudinal modes of such a laser cavity, the laser may oscillate in several closely spaced modes and the single-mode suppression ratio remains modest. By incorporating a grating structure in the laser, either distributed over the entire waveguide length (distributed feedback or DFB laser) or localized close to the facets (distributed Bragg reflector or DBR laser), spectral selection can be achieved, resulting in high-quality single-mode lasers. DFB lasers, in particular, have been successfully applied in optical communications systems. A more recent type of laser is the so-called vertical cavity surface emitting laser (VCSEL), where the laser end mirrors are provided by Bragg gratings parallel to the semiconductor substrate and the laser emits out of the plane of the substrate.

The choice of the materials for the active region determines the wavelength of the output optical beam. Two materials systems are particularly important: GaAlAs lasers, which cover the wavelength range of 700–900 nm, and InGaAsP lasers, which cover the wavelength range of 1000–1600 nm. The latter laser type is the most important for long-distance optical communication systems due to the low loss and/or dispersion of optical fibers in the range of 1300–1550 nm. An additional degree of freedom comes from employing a so-called quantum well structure of the active layer. Thus, by incorporating thin (about 5–10 nm) layers of semiconductor with a bandgap lower than the surrounding (barrier) material in the active region, quantum confinement effects change the allowed energy levels of electrons. This leads to a change in the effective bandgap of the laser (wavelength under lasing), as well as the number of electrons needed to reach population inversion.

Quantum well lasers have achieved low-threshold and high-power operation. Quantum dot lasers, employing three-dimensional quantum confinement of electrons, offer excellent electron control and have led to lasers with record low-threshold current density. However, semiconductor growth technology has not yet (in 2004) matured to the point of offering full control of quantum dot sizes.

The laser threshold condition and the basic features of the laser dynamics are captured by a simple set of rate equations describing the temporal evolution of the photon density, P, and the carrier density, N (Coldren & Corzine, 1995):

$$\frac{\mathrm{d}P}{\mathrm{d}t} = (\Gamma v_{\mathrm{g}} g - \tau_{\mathrm{p}}^{-1})P + \Gamma \beta_{\mathrm{sp}} R_{\mathrm{sp}}, \tag{1}$$

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{I}{eV} - \frac{N}{\tau_{\mathrm{s}}} - v_{\mathrm{g}} g P. \tag{2}$$

Here, v_{g} is the group velocity; τ_{p} is the photon lifetime, its inverse being the rate at which photons are lost from the laser cavity; β_{sp} is the rate of spontaneous emission; ending up in the lasing mode; I is the injected current; e is the electronic charge; V is the active region volume; and τ_{s} is the carrier lifetime. The confinement factor, Γ, expresses the fraction of the optical mode that overlaps with the active region; it may also be expressed as the ratio between the active region volume, V, and the effective volume of the optical mode, V_{opt}. In the form of the equations stated above, the photon density is normalized with respect to V_{opt}, whereas N is normalized with respect to V. Finally, g is the material gain. When considering a laser operating at the peak of the gain curve, it is, at least for lasers with bulk active regions, a good approximation to parameterize the gain as

$$g \cong g_{\mathrm{N}}(N - N_0) \tag{3}$$

with g_{N} being the differential gain, and N_0 the carrier density at transparency. For $N = N_0$, we have $g = 0$, corresponding to the case where stimulated emission and absorption exactly balance. For a further increase of the carrier density, population inversion is achieved and the material gain is positive.

The rate equations (1) and (2) are basically bookkeeping equations. Equation (2) expresses the effective pumping of electrons from valence to conduction band through the applied current. Due to spontaneous emission as well as nonradiative recombination (Auger processes are particularly important in long-wavelength lasers), the excited carrier density has a lifetime τ_{s} of the order of a nanosecond or less. Furthermore, stimulated emission, proportional to the product of the gain and the photon density, depletes the population of excited carriers, as expressed by the last term in Equation (2). That same process leads to a generation term in the rate equation for the photon density. Also, a certain fraction, β_{sp}, which may typically be of the order of 10^{-5} of the total rate of spontaneous emission, R_{sp}, ends up in the laser mode and is accounted for by the last term in Equation (1). Finally, the drain term in Equation (1) describes all mechanisms by which photons are lost from the cavity, including output coupling at the mirror facets and internal loss (due to free-carrier absorption, waveguide scattering loss, etc.).

The steady-state solution of the rate equations, with the gain given by Equation (3), yields the light-current characteristics, laser power that is, expressed in terms of the photon density as a function of the applied current. A simple solution, yet accurate except very close to threshold, is obtained by neglecting the rate of spontaneous emission into the lasing mode, that is, the last term in Equation (1). A small-signal analysis shows that the above-threshold solution is a stable focus. The characteristic frequency is the so-called laser relaxation oscillation frequency, which is the natural frequency at which energy is exchanged between the carrier population and the photon population. The relaxation frequency is also a measure of the order of the maximum bit-rate at which a laser can be efficiently current-modulated to produce an intensity-modulated optical output signal. The square of the relaxation oscillation is approximately proportional to the laser output power, although the oscillations become strongly damped as the frequency increases. High-speed lasers have relaxation frequencies of the order of 30 GHz.

From the Kramers–Kronig relations, any change of the gain of a material (proportional to the imaginary part of the susceptibility) implies a change in the refractive index (proportional to the real part of the susceptibility). Due to the asymmetric nature of the gain spectrum of a semiconductor laser—with a transparent region below the bandgap of the material and a strongly absorbing region at large photon energies—the coupling between index and gain is particularly strong for semiconductor lasers, with profound consequences for the dynamics. This coupling is described by the so-called linewidth enhancement factor (or alpha-parameter) α. It was thus realized that the gain-index coupling gives rise to an enhancement of the linewidth of a semiconductor laser by a factor of $1 + \alpha^2$ (Henry, 1983), and also imposes a chirp on the optical signal under current modulation. These effects are not described by the rate equation (1), which only governs the magnitude squared of the electromagnetic (optical) field and is independent of the phase. Rather the dynamics need to be described by an equation for the complex electric field amplitude, E:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{1}{2}(1 + \mathrm{i}\alpha)\Gamma v_{\mathrm{g}} g_{\mathrm{N}}(N - N_0)E + F_E(t), \tag{4}$$

where $P \propto |E|^2$. The term $F_E(t)$ is a stochastic Langevin noise term accounting for the random nature of spontaneous emission. The amplitude and phase noise properties of single-mode semiconductor lasers can be analyzed based on Equations (3) and (2).

Addition of a feedback term, proportional to $E(t-\tau)$, to the right-hand side of Equation (3) leads to the famous Lang–Kobyashi equations (Lang & Kobyashi, 1980), which govern the dynamics of a semiconductor laser coupled to an external mirror, τ being the roundtrip time in the external cavity. The semiconductor laser with feedback displays very rich dynamics, including mode-hopping, various instabilities and chaos, that to a remarkable degree are explained by Equations (4) and (2) (Mørk et al., 1992).

The rate equation model outlined above is limited to the case of single transverse mode lasers. Wide aperture lasers, in contrast, have an additional degree of freedom in the transverse direction. It has been shown that a complex semiconductor Swift–Hohenberg equation may describe the dynamics of such lasers in a single longitudinal mode mean field limit (Mercier & Moloney, 2002).

JESPER MØRK

See also **Diodes; Lasers; Nonlinear optics; Optical fiber communications**

Further Reading

Coldren, L.A. & Corzine, S.W. 1995. *Diode Lasers and Photonic Integrated Circuits*, New York: Wiley

Henry, C.H. 1983. Theory of the phase noise and power spectrum of a single mode injection laser. *IEEE Journal of Quantum Electronics*, QE-19: 1391–1397

Lang, R. & Kobyashi, K. 1980. External optical feedback effects on semiconductor injection laser properties. *IEEE Journal of Quantum Electronics*, QE-16: 347–355

Mercier, J.-F. & Moloney, J.V. 2002. Derivation of semiconductor laser mean-field and Swift-Hohenberg equations. *Physical Review* E, 66: 036221-1–036221-19

Mørk, J., Tromborg, B. & Mark, J. 1992. Chaos in semiconductor lasers with optical feedback: theory and experiment. *IEEE Journal of Quantum Electronics*, 28: 93–108

SEMICONDUCTOR OSCILLATORS

The electric transport properties of a semiconductor show up most directly in its current-voltage characteristic. It is related to a relationship between the current density j and the electric field E that is determined in a complex way by the microscopic properties of the material. Although a local, static, scalar $j(E)$ relation need not exist, it does in many cases.

Close to thermodynamic equilibrium (at sufficiently low bias voltage), the $j(E)$ relation is linear (Ohm's law), but under practical operating conditions it often becomes nonlinear and may even display a regime of negative differential conductivity (NDC), where $\sigma_{\text{diff}} = \mathrm{d}j/\mathrm{d}E < 0$. The global current-voltage characteristic $I(U)$ of a semiconductor can in principle be calculated from the local $j(E)$ relation by integrating the current density j over the cross section A of the current flow and the electric field E over the length L of the sample. Unlike the $j(E)$ relation, the $I(U)$ characteristic is not only a property of the semiconductor material, but depends also on the geometry, the boundary conditions, and the contacts of the sample. Only for the idealized case of spatially uniform states, are the $j(E)$ and the $I(U)$ characteristics identical, up to re-scaling. The $I(U)$ relation is said to display negative differential conductance if $\mathrm{d}I/\mathrm{d}U < 0$.

In the case of negative differential conductance, the current decreases with increasing voltage, and vice versa, which may lead to instability. The actual electrical response depends upon the attached circuit, which in addition to resistors may comprise capacitors and inductors. These reactive components give rise to additional degrees of freedom that are described by ordinary differential equations for I and U. If a semiconductor element with NDC is operated in such a reactive circuit, oscillatory instabilities may be induced by these reactive components. Self-sustained semiconductor oscillations, where the semiconductor itself introduces an internal unstable temporal degree of freedom, can be distinguished from those circuit-induced oscillations. The self-sustained oscillations under time-independent external bias are discussed here. Examples of internal degrees of freedom are the charge carrier density, the electron temperature, or a junction capacitance within the device.

Two important examples of NDC are described by an N-shaped or an S-shaped $j(E)$ characteristic and denoted by NNDC and SNDC, respectively (Figure 1). However, more complicated forms such as Z-shaped, loop-shaped, or disconnected characteristics are also possible. NNDC and SNDC are associated with voltage or current-controlled instabilities, respectively. In the NNDC case the current density is a single-valued function of the field, but the field is multivalued; in other words, the $E(j)$ relation has three branches in a certain range of j. The SNDC case is complementary in the sense that E and j are interchanged.

In case of NNDC, the NDC branch is often but not always (depending upon external circuit and boundary conditions) unstable against the formation of non-uniform field profiles along the charge transport direction (electric field domains). In the SNDC case, on the other hand, current filamentation generally occurs, in which the current density becomes non-uniform over the cross section of the current flow and forms a conducting channel (Ridley, 1963). These primary self-organized spatial patterns may themselves become unstable, leading to periodically or chaotically breathing, rocking, moving, or spiking filaments or domains, or even solid-state turbulence and spatiotemporal chaos (Schöll, 2001). Alternatively, the spatially uniform steady state may already become unstable with respect to uniform oscillations in a Hopf bifurcation.

Semiconductor oscillators may be classified as dominated by a bulk mechanism (drift instability,

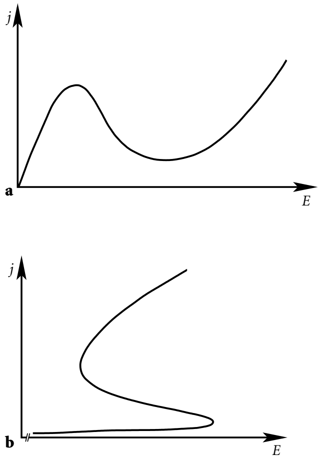

Figure 1. Current density j versus electric field E for two types of negative differential conductivity (NDC): (a) NNDC; (b): SNDC (schematic).

generation-recombination instability) or by heterojunctions and potential barriers and wells (resonant tunneling across, or thermionic emission over, barriers in nanostructures).

The first class of semiconductor oscillators includes drift instability. In the simplest extension of the Drude model, the current density is given by $j(E) = -env(E)$ where $e > 0$ is the electron charge, n is the electron density, and $v(E)$ is the field-dependent drift velocity, which may give rise to negative differential conductivity if $\mathrm{d}|v|/\mathrm{d}|E| < 0$. The best-known example is the Gunn effect, which is based upon intervalley transfer of electrons in k-space from a state of high mobility to a state of low mobility under a strong electric field in direct semiconductors like GaAs or other III–V compounds (Gunn, 1964; Ridley & Watkins, 1961). This phenomenon is used in real devices (Gunn diodes) to generate and amplify microwaves at frequencies typically beyond 1 GHz.

The class of generation-recombination (GR) instabilities is distinguished by a nonlinear dependence of the steady-state carrier concentration n upon the field E that yields a non-monotonic relation $j = en(E)\mu E$ of either NNDC or SNDC type, where μ is the mobility. This dependency is due to a redistribution of electrons between the conduction band and bound states with increasing field. The microscopic transition probabilities of the carriers between different states, and hence the GR coefficients, generally depend upon the electric field. Models of this type are relevant for a variety of materials and in various temperature ranges (Schöll, 1987) and can explain SNDC and current filamentation in the regime of low-temperature impurity breakdown and self-generated current oscillations including chaotic behavior, as observed experimentally (Peinke et al., 1992). Two important devices are also based upon GR-induced bulk negative differential conductivity, but the coupling with junction effects is essential in these cases: p-i-n diodes and impact ionization avalanche transist-time (IMPATT) diodes.

A variety of instabilities can arise due to the specific transport properties of semiconductor heterostructures. One mechanism for NNDC, which is the real space analog of the k-space intervalley transfer in the Gunn effect, uses electron transfer between a high-mobility layer and a low-mobility layer in a modulation-doped semiconductor heterostructure under a time-independent bias applied parallel to the layers (Gribnikov et al., 1995). In the NNDC regime, current oscillations of 2–200 MHz have been experimentally observed and theoretically explained.

Another class of oscillatory instabilities occurs under vertical electrical transport in layered semiconductor structures, for example, in the heterostructure hot-electron diode (HHED) or the double-barrier resonant tunneling diode (DBRT), which are associated with S-shaped and Z-shaped current-voltage characteristics, respectively (Schöll, 2001). A resonant tunneling structure is composed of alternating layers of two different semiconductor materials with different bandgaps. The energy diagram shows a modulation of the conduction band edge on a nanometer scale, forming potential barriers and quantum wells. The current density across the barrier between two wells is due to quantum mechanical tunneling and exhibits a strongly nonlinear dependency upon the electric field. It is maximum if there is maximum overlap between the occupied states in one well and the available unoccupied states in the other (the energies are in resonance). For low-fields, equivalent levels in adjacent wells are approximately in resonance. With increasing field, the energies of the two wells are shifted with respect to each other, and the available states in the collecting well are lowered with respect to the emitting well, and hence the current density drops as the overlap between the energy levels decreases, thereby displaying NDC. Upon further increase of the field, the current density rises again up to a sharp resonance peak when the ground energy level in one quantum well is aligned with the second level in the neighboring well. Thus, resonant tunneling produces NNDC.

The simplest system of this type consists of a double-barrier structure with one embedded quantum well in between, sandwiched between a highly doped emitter and collector region. However, the situation becomes more complicated if the nonlinear feedback between space charges and transport processes is taken into account. The built-up charge in the well leads to an electrostatic feedback mechanism that increases the energy of the well state supporting resonant tunneling conditions for higher applied voltages. This may result in bistability and hysteresis where a high current and a low current state coexist for the same applied voltage, and the current-voltage characteristic becomes Z-shaped. Switching between the two stable states as well as self-sustained current oscillations may occur under appropriate external circuit conditions. The bistability also provides a basis for lateral pattern formation (current filamentation) and spatiotemporal bifurcation scenarios including chaotic breathing and spiking.

Sequential resonant tunneling in a periodic structure of multiple quantum wells, a semiconductor superlattice, likewise displays NNDC (Esaki & Chang, 1974). Now, along the growth direction, the uniform field distribution may break up into a low-field domain, where the field is near the first peak of the $j(E)$ characteristic, and a high-field domain, where the field is close to the second, resonant-tunneling peak. Depending upon doping, the applied voltage, structural parameters, and the emitter contact conductivity, stationary or traveling domains occur—the latter leading to self-sustained current oscillations ranging from several hundred MHz up to 150 GHz at room temperature (Wacker, 2002).

ECKEHARD SCHÖLL

See also **Avalanche breakdown; Diodes; Drude model; Nonlinear electronics**

Further Reading

Aoki, K. 2000. *Nonlinear Dynamics and Chaos in Semiconductors*, Bristol and Philadelphia: Institute of Physics

Esaki, L. & Chang, L.L. 1974. New transport phenomenon in a semiconductor superlattice. *Physical Review Letters*, 33(8): 495–498

Gribnikov, Z.S., Hess, K. & Kosinovsky, G. 1995. Nonlocal and nonlinear transport in semiconductors: Real-space transfer effects. *Journal of Applied Physics*, 77: 1337–1373

Gunn, J.B. 1964. Instabilities of current in III–V semiconductors. *IBM Journal of Research and Development*, 8: 141–159

Peinke, J., Parisi, J., Rössler, O. & Stoop, R. 1992. *Encounter with Chaos*, Berlin and New York: Springer

Ridley, B.K. 1963. Specific negative resistance in solids. *Proceedings of the Physical Society*, 82: 954

Ridley, B.K. & Watkins, T.B. 1961. The possibility of negative resistance effects in semiconductors. *Proceedings of the Physical Society*, 78: 293–304

Schöll, E. 1987. *Nonequilibrium Phase Transitions in Semiconductors*, Berlin and New York: Springer

Schöll, E. 2001. *Nonlinear Spatio-temporal Dynamics and Chaos in Semiconductors*, Cambridge and New York: Cambridge University Press

Schöll, E., Niedernostheide, F.J., Parisi, J., Prettl, W. & Purwins, H. 1998. Formation of spatio-temporal structures in semiconductors. In *Evolution of Spontaneous Structures in Dissipative Continuous Systems*, edited by F.H. Busse & S.C. Müller, Berlin: Springer, pp. 446–494

Shaw, M.P., Mitin, V.V., Schöll, E. & Grubin, H.L. 1992. *The Physics of Instabilities in Solid State Electron Devices*, New York: Plenum Press

Wacker, A. 2002. Semiconductor superlattices: a model system for nonlinear transport. *Physics Reports*, 357: 1–111

SEMI-LINEAR PDES

See **Quasilinear analysis**

SENSITIVE DEPENDENCE ON INITIAL CONDITIONS

See **Butterfly effect**

SEPARATION OF VARIABLES

Separation of variables is the name of a general method for finding particular solutions of partial differential equations (PDEs) as a product of functions, where each factor depends on only one of the independent variables and satisfies an ordinary differential equation (ODE).

In the study of linear equations, a familiar example of the method is eigenfunction analysis. Using the superposition principle, this approach leads to expansions in products of orthogonal functions. The Fourier transform method is a special limiting case.

For a particular PDE, there may be a family of coordinate systems that admits separation of variables. The problem of finding such coordinate systems is closely connected with the group properties of differential equations. Methods from Lie group theory can be used to describe all the separable solutions of many equations from mathematical physics (Miller, 1977). In Morse and Feshbach (1953), the separable orthogonal coordinate systems for the Laplace, Helmholtz, Schrödinger, heat (diffusion), and wave equations in two and three dimensions are listed.

For example, Laplace's equation

$$u_{xx} + u_{yy} = 0 \tag{1}$$

has product solutions $u(x, y) = X(x)Y(y)$, where X (respectively, Y) is a function of only the independent variable x (respectively, y). Here X and Y satisfy the ordinary differential equations $X'' + \lambda X = 0$ and $Y'' - \lambda Y = 0$, where λ is the separation constant and prime denotes differentiation with respect to the independent variable. X and Y may thus be expressed in terms of trigonometric and exponential functions.

Separability depends on the boundary conditions. As a second example, consider the Helmholtz equation

$$u_{xx} + u_{yy} + k^2 u = 0 \tag{2}$$

in two dimensions with Dirichlet or Neumann boundary conditions ($u = 0$ or $u_n = 0$, where subscript n denotes the derivative in the direction normal to the boundary). This boundary value problem is separable in rectangular, parabolic, polar, and elliptic coordinates.

However, with the mixed (Robin) or impedance boundary condition $u_n + cu = 0$ (where the impedance c is a constant), the problem is only separable in rectangular and polar coordinates.

For nonlinear wave equations a method for separation of variables is sometimes called Lamb's method, stemming from an early analysis of the sine-Gordon (SG) equation (Lamb, 1971)

$$u_{xx} - u_{tt} - \sin u = 0. \tag{3}$$

This equation has solutions of the form $u(x,t) = 4\tan^{-1}(X(x)T(t))$, where $(X')^2 = a_1X^4 + a_2X^2 + a_3$ and $(T')^2 = -a_3T^4 + (a_2 - 1)T^2 - a_1$, and a_1, a_2, and a_3 are separation constants. In general, therefore, X and T are Jacobi elliptic functions.

In special cases simpler SG soliton solutions such as kinks, antikinks, colliding kinks and antikinks, breathers or bions, and plasma waves, are obtained. For example, choosing $a_1 = a_2 = 1/(1 - v^2)$ and $a_3 = 0$ yields colliding kinks, traveling with velocities v and $-v$, respectively, while $a_2 = -a_3 = 1/(1 - v^2)$ and $a_1 = 0$ yields colliding kinks and antikinks. Since the SG equation models long Josephson junctions, fluxons traveling on Josephson junctions of infinite length are obtained as derivatives of the kinks (Scott, 2003), and nonlinear oscillations or standing waves on junctions of finite length can also be found for boundary conditions of Dirichlet or Neumann type (Costabile et al., 1978).

Separation of variables provides solutions to other nonlinear partial differential equations such as the nonlinear Schrödinger equation (NLS)

$$\mathrm{i}u_t + u_{xx} + 2|u|^2u = 0. \tag{4}$$

Writing $u(x,t) = \phi(x,t)\mathrm{e}^{\mathrm{i}\theta(x,t)}$, the amplitude function, $\phi(x,t)$, and the phase function, $\theta(x,t)$, satisfy nonlinear coupled ordinary differential equations that permit traveling wave solutions of the form $\phi(x,t) = \phi(x - vt)$, where v is the velocity. These are the NLS envelope solitons (Scott, 2003).

Standing wave solutions to the so-called improved Boussinesq equation

$$u_{xx} - u_{tt} - (u^2)_{xx} + u_{xxtt} = O \tag{5}$$

of the form $u(x,t) = 1/2 + X(x)T(t)$, where $X(x)$ and $T(t)$ satisfy uncoupled nonlinear ordinary differential equations, have been obtained in Rosenau and Schwarzmeier (1986) and Christiansen et al. (1990).

The Hamilton–Jacobi equation is separable when Hamilton's characteristic function, W, can be written as a sum of functions where each function depends on only one of the independent variables (Goldstein, 1980). Let us look at a particle with mass m moving in central force with potential $V(r)$. Using polar coordinates (r, ϕ) in the plane of the orbit, the Hamiltonian has the form

$$H = \frac{1}{2m}\left(p_r^2 + \frac{p_\phi^2}{r^2}\right) + V(r), \tag{6}$$

and is cyclic in ϕ. Consequently, Hamilton's characteristic function

$$W = W_r(r) + W_\phi(\phi) \equiv W_r(r) + \alpha_\phi\phi, \tag{7}$$

where the momentum conjugated to r is $p_r = \partial W_r(r)/\partial r$ and the angular momentum conjugated to ϕ, $p_\phi = \partial W_\phi(\phi)/\partial\phi = \alpha_\phi$ is a constant. The Hamilton–Jacobi equation now becomes

$$\left(\frac{\partial W_r(r)}{\partial r}\right)^2 + \frac{\alpha_\phi^2}{r^2} + 2mV(r) = 2mE, \tag{8}$$

where E is the total energy of the system. Integrating (8) with respect to r, $W_r(r)$ and thus W given by (7) are obtained. The canonical equations then yield the orbitals in the form $r = r(t)$ or $r = r(\phi)$.

A new approach has been developed by Sklyanin (1995), who argues that separation of variables, understood generally enough, could be a universal tool to solve integrable models of classical and quantum mechanics. Standard construction of the action-angle variables for the poles of the Baker–Akhiezer function can be interpreted as a variant of separation of variables. The new approach has been applied to magnetic chains, the Toda lattice, the nonlinear Schrödinger equation, the sine-Gordon model, and other systems (Skyanin, 1995).

Peter L. Christiansen

See also **Boundary value problems; Integral transforms; Long Josephson junctions**

Further Reading

Christiansen, P.L., Lomdahl, P.S. & Muto, V. 1990. On a Toda lattice model with a transversal degree of freedom. *Nonlinearity*, 4: 477–501

Costabile, G., Parmentier, R.D., Savo, B., McLaughlin, D.W. & Scott, A.C. 1978. Exact solutions of the sine-Gordon equation describing oscillations in a long (but) finite Josephson junction. *Applied Physics Letters*, 32: 587–589

Goldstein, H. 1980. *Classical Mechanics*, 2nd edition, Reading, MA: Addison-Wesley

Lamb, G.L., Jr. 1971. Analytical descriptions of ultrashort optical pulse propagation in a resonant medium. *Reviews of Modern Physics*, 43: 99–124

Miller, U. 1977. *Symmetry and Separation of Variables*, Reading, MA: Addison-Wesley

Morse, P.M. & Feshbach, H. 1953. *Methods of Theoretical Physics*, New York and London: McGraw-Hill

Rosenau, P. & Schwarzmeier, J.L. 1986. On similarity solutions of Boussinesq-type equation. *Physics Letters* A, 115: 75–77

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Sklyanin, E.K. 1995. Separation of variables: new trends. *Progress in Theoretical Physics Supplement*, 118: 35–60

SEPARATRIX

See **Phase space**

SHAPIRO STEPS

See **Josephson junction**

SHARKOVSKY'S THEOREM

See **One-dimensional Maps**

SHEAR FLOWS

Shear flows arise whenever two bodies of fluid move relative to each other, as in a jet of liquid entering a fluid at rest or the flow past a solid obstacle. When the flow speeds become too high, these flows undergo instabilities that are at the root of the vortical structures that dominate turbulent flows. For the Kelvin–Helmholtz; instability of two layers of fluid in relative motion; the instabilities of a fluid sheared between two concentric, rotating cylinders (Taylor–Couette flow); the dynamics of boundary layers; and viscous Hele-Shaw experiments, see the corresponding *Encyclopedia* entries.

Energy balance: The significance of shear as a source of energy for perturbations follows from the analysis of the energy content of a perturbation to a shear flow. Let $\boldsymbol{U}$ be a prescribed stationary flow in a domain V and let $\boldsymbol{u}$ be the perturbation added to $\boldsymbol{U}$. The Reynolds number is defined in terms of the velocity scale U of the reference flow, a length scale L, and the kinematic viscosity ν, that is, $\mathrm{Re} = UL/\nu$. For a divergence free perturbation with boundary conditions $\boldsymbol{u} = 0$ on surfaces, the energy content $E(t) = \int_V \mathrm{d}V(\boldsymbol{u}^2/2)$ satisfies the Orr–Reynolds equation

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\int_V u_i u_j \frac{\partial U_i}{\partial x_j}\mathrm{d}V - \frac{1}{\mathrm{Re}}\int_V \frac{\partial u_i}{\partial x_j}\frac{\partial u_j}{\partial x_i}\,\mathrm{d}V. \quad (1)$$

Since the last term on the right-hand side is negative semidefinite, all the energy input has to come from the shear $\partial U_i/\partial x_j$. If the shear is too small, then $E(t)$ will decay monotonically. The Reynolds number up to which $\dot{E} \le 0$ for all perturbations defines the energy stability limit Re_E.

Parallel flows: A necessary condition for an instability of parallel inviscid flows was derived by Lord Rayleigh (John William Strutt) in 1880; he found that instability requires an inflection point. If $\boldsymbol{u} = U(y)\boldsymbol{e}_x$ is the unperturbed profile, then instability can occur only if there is a point y_s with $U''(y_s) = 0$. This criterion was later improved by R.Fjørtoft, who found the requirement $U''(y)(U(y) - U(y_s)) < 0$. Necessary and sufficient conditions are more complicated, as discussed in Balmforth & Morrision (1999).

For the viscous stability, it is useful to represent the perturbation with components (u, v, w) in terms of the vertical velocity component v and the vertical vorticity, $\eta = \partial_z u - \partial_x w$, and to expand in terms of Fourier modes,

$$v(x, y, z, t) = \tilde{v}(y)\mathrm{e}^{\mathrm{i}(\alpha x + \beta z - \omega t)}, \quad (2)$$

$$\eta(x, y, z, t) = \tilde{\eta}(y)\mathrm{e}^{\mathrm{i}(\alpha x + \beta z - \omega t)}. \quad (3)$$

Then the amplitudes satisfy

$$\left[(-\mathrm{i}\omega + \mathrm{i}\alpha U)(D^2 - k^2) - \mathrm{i}\alpha U'' - \frac{1}{\mathrm{Re}}(D^2 - k^2)^2\right] \times \tilde{v} = 0, \quad (4)$$

$$\left[(-\mathrm{i}\omega + \mathrm{i}\alpha U) - \frac{1}{\mathrm{Re}}(D^2 - k^2)\right] \times \tilde{\eta} = -\mathrm{i}\beta U'\tilde{v}. \quad (5)$$

The first equation is the Orr–Sommerfeld equation, the second Squire's equation. In 1933, H.B. Squire showed that if there is an instability for a mode with spanwise wave number $\beta \neq 0$, then there is another one with $\beta = 0$ that has a lower Reynolds number, where instabilities set in without spanwise modulations.

The linearized problem can show transient growth: a decaying eigenstate of the Orr–Sommerfeld equation can drive the Squire equation and cause an amplification of the normal vorticity (if $\beta \neq 0$). The most effective modes often have the form of downstream vortices, with only slight modulations in spanwise and downstream direction. They drive spanwise modulations in the downstream velocity component, so-called streaks. Fabian Waleffe (1995, 1997) pointed out that the streaks can undergo an instability themselves, forming normal vortices, which can then be fed back into downstream vortices to close the loop. This self-sustaining regeneration mechanism plays a major role in the turbulence of parallel flows. The results for the transition to turbulence in viscous planar shear flows are summarized in Table 1. The flows are shear flow between parallel plates with a linear profile in the laminar case (plane Couette), pressure-driven flow between parallel plates with a parabolic profile (plane Poiseuille), and pressure-driven flow down a pipe, also with a parabolic profile (Hagen–Poiseuille). The various Reynolds numbers are defined as follows:

- Re_E *Energy stability*: as above.
- Re_G *Global stability*: up to this Reynolds number the flow is globally stable, and any perturbation will decay, perhaps after a long transient during which the energy can grow above the initial energy.
- Re_T Reynolds number near which experiments indicate a transition to turbulence.
- Re_L Reynolds number for linear instability.

For sufficiently low Re, all flows are linearly stable. For plane Couette flow and pipe flow, we know that the flows cannot be globally stable at Reynolds numbers above the given ones because of the existence of 3-dimensional stationary states or traveling waves

Flow	Re_E	Re_G	Re_T	Re_L
Plane Couette flow	20.7	125	310	∞
Plane Poiseuille flow	49.6	≈ 1000	1000	5772
Hagen–Poiseuille flow	81.5	1250	2250	∞

Table 1. Critical Reynolds numbers for parallel shear flows.

(Nagata, 1990; Ehrenstein & Koch, 1991; Clever & Busse 1997; Faisst & Eckhardt, 2003; Wedin & Kerswell, 2004). Two of the flows do not show a linear instability, and for the third one, it appears at values far above the ones where the transition to turbulence occurs. Incidentally, plane Poiseuille flow provides an example of a flow profile that is inviscidly stable by Rayleigh's criterion but becomes unstable when viscosity is included.

In all three cases, the value for transition to turbulence is somewhat uncertain because of the absence of a sharp transition (Boberg & Brosa, 1988; Darbyshire & Mullin, 1995; Bottin et al., 1998). There is evidence that this is connected with the formation of a chaotic saddle in phase space (Schmiegel & Eckhardt, 1997). Scanning an amplitude–Reynolds number plane for perturbations in a low-dimensional model indeed shows a sensitive dependence on initial conditions and huge variations in lifetimes for neighboring trajectories (see figure in color plate section). The lifetimes of perturbations are exponentially distributed, a clear signature of a chaotic saddle.

Turbulent shear flows: The energy dissipation in a turbulent shear flow can be related to the velocity difference U and the width L of the flow as

$$\varepsilon_V = c_\varepsilon(\mathrm{Re})U^3/L \qquad (6)$$

with a dissipation factor $c_\varepsilon(\mathrm{Re})$ that depends on the Reynolds number $\mathrm{Re} = UL/\nu$. For laminar parallel flows, $c_\varepsilon \sim 1/\mathrm{Re}$. For turbulent shear flows, the variational theories of Busse (1970), Doering & Constantin (1992), Kerswell (1997), and Nicodemus et al. (1997) show that in the limit of infinite Reynolds number, $c_\varepsilon(\mathrm{Re})$ can be rigorously bounded from above. The theory bounds c_ε through a variational functional with a background profile $\phi(y)$, with the constraint that the background profile has to be energy stable. The best bound is $c_\varepsilon \leq 0.0109$. Comparing with experiment, one notes that the observed values are lower and that for smooth boundaries, they tend to decrease for increasing Reynolds number.

The presence of a large-scale shear introduces an anisotropy into the flow, which also affects the turbulent statistics. While in an isotropic turbulent flow the odd moments of the normal derivative of the downstream component vanish, this is no longer the case in a turbulent shear flow. Dimensional estimates by Lumley (1967) suggest that the anisotropy should vanish like $1/\mathrm{Re}$, but experimental and numerical evidence indicates that the decay is much slower. Current efforts aim at extracting information about the relevant process from the dynamics of passive scalars: the scalar fields develop characteristic ramps and cliffs that can be related to the asymmetry in the distributions of the gradients (Schumacher & Sreenivasan, 2003).

Other shear flows: In viscoelastic fluids, the interaction between shear and internal degrees of freedom can also give rise to much more complicated instabilities for which Squire's theorem does not hold. The reduction of turbulent drag by small additions of long flexible polymers remains a fascinating phenomenon awaiting explanation (Lumley, 1969).

In astrophysics a combination of differential motion in the plasma and the presence of a magnetic field can give rise to instabilities that are responsible for the transport of angular momentum, thus solving a long-standing puzzle about the angular momentum distribution in galaxies (Balbus & Hawley, 1998).

BRUNO ECKHARDT

See also **Boundary layers; Hele-Shaw cell; Kelvin–Helmholtz instability; Taylor–Couette flow; Turbulence**

Further Reading

Balbus, S.A. & Hawley, J.F. 1998. Instability, turbulence, and enhanced transport in accretion disks. *Reviews of Modern Physics*, 70: 1–52

Balmforth, N.J. & Morrison, P.J. 1999. A necessary and sufficient instability condition for inviscid shear flow. *Studies in Applied Mathematics*, 102: 309–344

Boberg, L. & Brosa, U. 1988. Onset of turbulence in pipe. *Zeitschrift für Naturforschung*, 43a: 697–726

Bottin, S., Daviaud, F., Manneville, P. & Dauchot, O. 1998. Discontinuous transition to spatiotemporal intermittency in plane Couette flow. *Europhysics Letters*, 43: 171–176

Busse, F.H. 1970. Bounds for turbulent shear flow. *Journal of Fluid Mechanics*, 41: 219–240

Clever, R.M. & Busse, F.H. 1997. Tertiary and quaternary solutions for plane Couette flow. *Journal of Fluid Mechanics*, 344: 137–153

Darbyshire, A.G. & Mullin, T. 1995. Transition to turbulence in constant-mass-flux pipe flow. *Journal of Fluid Mechanics*, 289: 83–114

Doering, C.R. & Constantin, P. 1992. Energy dissipation in shear driven turbulence. *Physical Review Letters*, 69: 1648–1651

Drazin, P.G. & Reid, W.H. 1981. *Hydrodynamic Stability*, 2nd edition, Cambridge and New York: University Press, 2004

Ehrenstein, U. & Koch, W. 1991. Three-dimensional wavelike equilibrium states in plane Poiseuille flow. *Journal of Fluid Mechanics*, 228: 111–148

Faisst, H. & Eckhardt, B. 2003. Travelling waves in pipe flow. *Physical Review Letters*, 91: 224502

Grossmann, S. 2000. The onset of shear flow turbulence. *Reviews of Modern Physics*, 72: 603–618

Joseph, D.D. 1976. *Stability of fluid motion*, part I and II, New York: Springer

Kerswell, R.R. 1997. Variational bounds on shear-driven turbulence and turbulent Boussinesq convection. *Physica* D, 100: 355–376

Lumley, J.L. 1967. Similarity and the turbulent energy spectrum. *Physics of Fluids*, 10: 855–858

Lumley, J.L. 1969. Drag reduction by additives. *Annual Reviews of Fluid Mechanics*, 1: 367–384

Nagata, M. 1990. Three-dimensional finite-amplitude solutions in plane Couette flow: bifurcation from infinity. *Journal of Fluid Mechanics*, 217: 519–527

Nicodemus, R., Grossmann, S. & Holthaus, M. 1997. The background flow method, part I and II. *Journal of Fluid Mechanics*, 363: 281–300 and 301–323

Rosenhead, L. (editor). 1963. *Laminar Boundary Layers*, Oxford: Oxford University Press

Schmid, P. & Henningson, D.S. 2001. *Stability and Transition in Shear Flows*, Berlin and New York: Springer

Schmiegel, A. & Eckhardt, B. 1997. Fractal stability border in plane Couette flow. *Physical Review Letters*, 79: 5250–5253

Schumacher, J. & Sreenivasan, K.R. 2003. Geometric features of the mixing of passive scalars at high Schmidt numbers. *Physical Review Letters*, 91: 174501

Waleffe, F. 1995. Transition in shear flows. Nonlinear normality versus non-normal linearity. *Physics of Fluids*, 7: 3060–3066

Waleffe, F. 1997. On a self-sustaining process in shear flows. *Physics of Fluids*, 9: 883–900

Wedin, H. & Kerswell, R.R. 2004. Exact coherent structures in pipe flow: travelling wave solutions. *Journal of Fluid Mechanics*, in press

SHOCK WAVES

In classical gas dynamics, a shock wave is a sharp, stepwise increase of density, pressure, and temperature that propagates at a supersonic speed with respect to the fluid ahead of it while remaining subsonic with respect to the fluid behind it. The fluid entropy increases after passing through the shock. Shock waves can be formed as a result of numerous processes such as supersonic motion of bodies (aircraft, meteors, bullets) in the atmosphere (see Figure 1), explosions in atmosphere and ocean, collapse of bubbles in the course of cavitation, and gas flow out of a nozzle in rockets. Also, the propagation of a nonlinear nondispersive wave (simple wave), in which each point of the profile propagates at its own velocity, generally results in the shock formation.

The changes (jumps) of different physical (thermodynamic) quantities at a shock satisfy specific relations (boundary conditions) following from the conservation of mass, momentum, and energy (mechanical plus thermodynamical). In the reference frame of the shock front, they can be presented in the form

$$[\rho v_n] = 0, \ \left[p + \rho v_n^2\right] = 0, \left[\frac{v_n^2}{2} + w\right] = 0. \qquad (1)$$

Here the square brackets denote the difference of the corresponding values at the shock; v_n is fluid velocity component normal to the front; w is enthalpy; p and ρ are, respectively, pressure and density. In the reference frame in which the fluid before the shock is immovable, v_n is $-V$ where V is the shock propagation velocity. The velocity of tangential fluid motion at the shock is continuous.

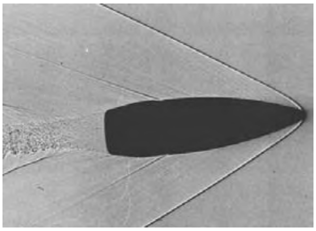

Figure 1. Shadowgraph showing shock waves produced by a Winchester 0.308 caliber bullet traveling through air at about Mach 2.5. (With courtesy from Ruprecht Nennstiel, Wiesbaden, Germany.)

Equations (1) give, in particular,

$$w_2 - w_1 = \tfrac{1}{2}(U_2 + U_1)(p_2 - p_1), \qquad (2)$$

where $U = 1/\rho$ is the specific volume, and subscripts 1 and 2 refer to the gas in front of and behind the shock. The enthalpy w can be expressed in terms of p and ρ via the thermodynamic equation of state. As a result, if the gas parameters p_1 and U_1 before the shock are given, Equation (2) determines the dependence between p_2 and U_2 called the Hugoniot adiabat. It differs from the Poisson adiabat that relates p and ρ in a perfect gas in which entropy is constant.

In a stable shock wave the entropy increases due to dissipative processes occurring inside the shock front, which is actually a transient layer of a finite thickness that grows with viscosity and thermal conductivity in the medium. If the transition region remains thin compared with the outer motion scale, it can be considered a discontinuity at which the boundary conditions (1) (which do not depend on the specific dissipation mechanism) remain valid. Note that if only thermal conductivity determines the shock front width (i.e., viscosity is neglected), the shock front can contain a discontinuity (isothermal jump) inside it.

General relations concerning shock waves are simplified in the important case of a polytropic gas in which p is proportional to ρ^γ, where $\gamma = c_p/c_v$, and c_p and c_v are heat capacities at constant pressure and volume, respectively. For example, in a very strong shock when $p_2 \gg p_1$ and the Mach number, $M = V\,.\,c \gg 1$ (c is the speed of sound), the ratios of gas densities and temperatures (T) behind and before the shock are

$$\frac{\rho_1}{\rho_2} = \frac{U_2}{U_1} = \frac{\gamma - 1}{\gamma + 1}, \ \frac{T_2}{T_1} = \frac{\gamma - 1}{\gamma + 1}\frac{p_2}{p_1}. \qquad (3)$$

Note that the maximal gas compression remains finite.

Among the dynamic problems associated with shock waves is that of shock reflection from a hard wall. If the incidence is normal to the wall, the pressure increases to more than twice that in an incident wave (for a linear wall the pressure is doubled). If a shock of moderate strength is incident obliquely under a sufficiently large grazing angle, the shock is reflected before reaching the wall (Mach stem effect).

Another important problem is a spherical shock from an explosion. For very strong spherical waves, a self-similar solution exists, in which the shock front radius R increases as $t^{2/5}$, whereas the pressure and particle velocity decrease as R^{-3} and $R^{-3/2}$, respectively. For a curved shock front, the solution can be found by an approximation: local shock curvature at each moment is related to the rays normal to the front, and the fluid flows along these ray tubes as in channels of variable width (Whitham, 1974).

A more general description of nonstationary flows containing shocks can be achieved for one-dimensional motions such as the waves in a tube excited by a piston moving at its end. The one-dimensional equations of motion for an ideal gas can be rewritten in terms of two Riemann invariants, $J_\pm$:

$$\left[\frac{\partial}{\partial t} + (v \pm c)\frac{\partial}{\partial x}\right] J_\pm = 0,$$

$$J_\pm = v \pm \int \frac{\mathrm{d}p}{\rho c} = v \pm \frac{2c}{\gamma - 1} \qquad (4)$$

(the latter expressions are written for a polytropic gas). Here $c(\rho) = (\mathrm{d}p/\mathrm{d}\rho)^{1/2} = (\gamma p/\rho)^{1/2}$ is the local speed of sound. Correspondingly, there are two families of trajectories-characteristics, $\mathrm{d}x/\mathrm{d}t = v \pm c$ on the (x, t) plane, along which small perturbations that are linear sound waves propagate. In a particular case when one of the invariants is constant, there is a progressive simple (Riemann) wave that travels at a local velocity $v \pm c$. In such a wave, the variables are related by $v = \pm \int \mathrm{d}p/\rho c$. Each point of the simple wave profile propagates at its own, constant velocity, and if the velocity decreases with x, at some moment the motion becomes multivalued (the wave breaks). In gas dynamics this is physically impossible, which means the formation of a shock wave at which the energy of the continuous part of the wave dissipates.

An important particular case is that of small nonlinearity when $p(\rho) \approx p_0 + c_0^2\rho'$, where p_0 and c_0 refer to the equilibrium state (in the absence of the wave), and ρ' is density variation in the wave. This corresponds to nonlinear acoustics where the velocity of a weak shock wave is approximately the average of the linear sound speeds in front of and behind the shock. In particular, an initially sinusoidal wave transforms into a sawtooth one.

Simple waves may exist in different nondispersive, nonlinear physical systems. One example is a nonlinear surface wave in shallow water when the characteristic wavelength is much larger than the depth of the water layer, h. Such a wave propagates at the local velocity of $c(\eta) = \sqrt{g(h + \eta)}$, where g is gravitational acceleration and $\eta(x, t)$ is the local water surface displacement with respect to its unperturbed level, h. In general, this wave breaks at a finite distance so that the long-wave approximation becomes invalid in the vicinity of the breaking point. The breaking can have different outcomes. Due to dispersion, the wave can generate oscillations that eventually become solitons, or, at larger steepness, the wave crest may break, possibly forming a turbulent front (bore), somewhat similar to a shock.

Magnetohydrodynamic Shock Waves

Shock waves can form in plasma. If the plasma is sufficiently dense, it can be treated as a compressible fluid that has electrical conductivity. When such a fluid moves in a magnetic field, the motion induces electric currents. Interaction between the currents and the magnetic field can significantly affect the fluid motion. Interaction between hydrodynamic and electromagnetic phenomena is the subject of *magnetohydrodynamics*. These phenomena are important in, for example, astrophysics, where at large-scale motions, space plasma typically behaves as a conducting fluid.

In a linear approximation, magnetohydrodynamic (MHD) waves are classified as "slow" and "fast" magnetic sound in which fluid compressibility is significant, and Alfvén waves depending only on magnetic field. In these waves magnetic perturbations are polarized perpendicular to the basic constant magnetic intensity vector $\boldsymbol{H}$ and group velocity is directed along $\boldsymbol{H}$. Finite-amplitude magnetic sound is distorted with a possible formation of MHD shocks.

As in nonmagnetic gas dynamics, parameters of MHD shock waves can be determined from boundary conditions at the discontinuity, which differ from the Rankine–Hugoniot conditions in that they include the magnetic field at both sides of the shock. The corresponding generalization of the Hugoniot adiabat reads (in the CGS system)

$$(\varepsilon_2 - \varepsilon_1) + \frac{1}{2}(p_2 + p_1)(U_2 - U_1) + \frac{1}{16\pi}(U_2 - U_1)(H_{\tau 2} - H_{\tau 1})^2 = 0, \qquad (5)$$

where H_τ is the component tangential to the shock front. In fluids with a positive thermal expansion coefficient, both pressure and density always increase at the shock, as in classical gas dynamics.

The nonshock discontinuities in MHD include tangential discontinuities, in which the vectors of fluid velocity and magnetic field are parallel to the discontinuity plane (in the comoving reference frame), and rotational (Alfvén) discontinuities in which the normal velocity component of velocity is nonzero but continuous whereas the vector $\boldsymbol{H}$ rotates around the normal with constant absolute value.

In plasma, dispersion can cause an oscillating shock structure and solitary wave formation. If plasma is rarefied so that the free path of electrons exceeds the shock front thickness, the so-called noncollisional shocks may exist.

Electromagnetic Shock Waves

Another class of shock waves occurs in media that can be macroscopically immovable but have nonlinear electromagnetic parameters; for example, the dependence between the vectors of magnetic induction and the magnetic field, $\boldsymbol{B}(\boldsymbol{H})$, or between their electric counterparts, $\boldsymbol{D}(\boldsymbol{E})$, are nonlinear. If the nonlinearity is relatively strong and dispersion effects are small, electromagnetic waves can propagate as simple waves, resulting in the formation of electromagnetic shocks.

Boundary conditions at a shock can be obtained by integrating Maxwell's equations over the shock transition layer, to give

$$[\boldsymbol{n} \times (\boldsymbol{E}_2 - \boldsymbol{E}_1)] = \frac{U_n}{c_0}(\boldsymbol{B}_2 - \boldsymbol{B}_1), \qquad (6)$$

$$[\boldsymbol{n} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1)] = \frac{U_n}{c_0}(\boldsymbol{D}_2 - \boldsymbol{D}_1),$$

$$B_{n2} = B_{n1}, D_{n2} = D_{n1},$$

where $\boldsymbol{n}$ is a normal to the discontinuity surface, c_0 is the light speed in vacuum, and U_n is the normal component of shock velocity. Typically only either magnetic or electric properties of the medium are nonlinear. In the former case, an expression for the shock velocity is

$$\frac{\varepsilon U_n^2}{c^2} = \frac{H_{\tau 2} - H_{\tau 1}}{B_{\tau 2} - B_{\tau 1}}. \qquad (7)$$

Electromagnetic shock waves are observed in ferrites, ferroelectrics, and semiconductors, and they have been used for construction of powerful impulse generators. In nonlinear dispersive systems, such as nonlinear transmission lines, the shock transition can be oscillating, and as dissipation approaches zero, it becomes a solitary electromagnetic wave.

LEV OSTROVSKY

See also **Alfvén waves; Characterisitics; Cherenkov radiation; Dimensional analysis; Explosions; Jump phenomena; Magnetohydrodynamics; Nonlinear acoustics**

Further Reading

Courant, R. & Friedrichs, K.O. 1948. *Supersonic Flow and Shock Waves*, reprinted New York: Springer, 1992

Gaponov, G., Ostrovsky, L. & Freidman, G. 1967. Electromagnetic shock waves (a review). *Radiophysics and Quantum Electronics*, 10: 9–10

Landau, L. & Lifshitz, E. 1960. *Electrodynamics of Continuous Media*, Oxford and New York, Pergamon Press

Landau, L. & Lifshitz, E. 1987. *Fluid Mechanics*, 2nd edition, Oxford: Pergamon Press

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

SHOOTING METHOD

See **Phase plane**

SIERPINSKI GASKET

See **Order from chaos**

SIGNALING PROBLEM

See **Wave stability and instability**

SINAI BILLIARD

See **Billiards**

SINAI–RUELLE–BOWEN MEASURES

What is a natural ivariant measure? In a probabilistic approach to dynamical systems, one seeks to understand average or almost-sure behavior, particularly in the limit as time tends to infinity. In these studies, it is often fruitful to have a stationary process, or equivalently, an *invariant probability measure*. Consider a map f on a bounded domain U of a Euclidean space or a manifold with $f(U) \subset U$. (We will restrict ourselves to the discrete-time case, but the discussion is equally valid for continuous-time.) In general, f admits infinitely many mutually singular invariant probability measures, and the ergodic theory of (f, μ) depends on the choice of μ. The purpose of this entry is to discuss what constitutes a "natural" invariant measure.

Throughout this discussion, we will adopt the view that only properties that hold on positive Lebesgue measure sets are observable. A first attempt to characterize natural invariant measures is to require that they have densities. Thus for a Hamiltonian system, Liouville measure is regarded as natural. This criterion runs into difficulties with dissipative systems. Suppose that f is volume-decreasing with an attractor $\Lambda = \cap_{n \geq 0} f^n(U)$. Since there is no recurrence behavior on $U \setminus \Lambda$, by the Poincaré recurrence theorem, all the invariant probability measures are supported on Λ. Observe that with f decreasing volume, Λ must necessarily have Lebesgue measure zero. Thus f cannot have an invariant density.

For dissipative dynamical systems, then, we must relax the idea of a natural invariant measure from one that has a density to one that reflects the properties of positive Lebesgue measure sets. We discuss below three closely related sets of ideas that go in this direction. In less rigorous discussions, the first two are often confused, even though as we will see, they are different substantively and mathematically.

SRB Measures

Sinai–Ruelle–Bowen measures or *SRB measures* were first introduced for Anosov systems and Axiom A attractors; see Sinai (1972), Bowen (1975), and Ruelle (1978). The idea was later extended to the non-uniform hyperbolic setting (*See* **Horseshoes and hyperbolicity in dynamical systems**), and the name SRB measure was introduced in the review article by Eckmann-Ruelle (1985).

The idea is as follows: SRB measures live on strange attractors. The situation being chaotic, there are positive Lyapunov exponents and unstable manifolds. In the same way that the next best scenario to having a differentiable function is to have partial derivatives, if the dissipative nature of a map prevents it from having an invariant measure that is smooth, then the closest approximation would be one that is smooth in certain directions. The expanding property of a map smoothes out invariant measures along unstable manifolds.

We now give the precise definition. Consider a C^2 invertible map f and an f-invariant Borel probability measure μ. We assume that f has positive Lyapunov exponents and hence unstable manifolds μ-almost everywhere. On each k-dimensional unstable leaf γ, let m_γ denote the k-dimensional Lebesgue measure. Then we say μ is an SRB measure if the conditional measures of μ on unstable leaves have densities with respect to the measures m_γ.

The geometric definition of SRB measures given above turns out to be equivalent to the following "variational principle": SRB measures are exactly those invariant measures for which the Kolmogorov–Sinai entropy of the system is equal to the sum of its positive Lyapunov exponents. (For other invariant measures, entropy is smaller.) If there are no zero Lyapunov exponents, there is a structure theorem that says that SRB measures have at most a countable number of ergodic components, and each ergodic component is mixing up to certain permutations (*See* **Horseshoes and hyperbolicity in dynamical systems**). Relations to other notions of natural invariant measures will be mentioned as we go along.

We explained earlier that SRB measures is a notion associated with strange attractors. This does not mean that every strange attractor necessarily has an SRB measure. In practice, one often assumes that it does, but mathematically, this question is far from resolved. In fact, not many attractors have been rigorously proved to have SRB measures. The main examples are Axiom A attractors and a class of attractors with one direction of instability including certain Hénon attractors (see Young, 2002).

Physical Measures

Let f be a map and μ an invariant probability measure. We say μ is a *physical measure* if there is a positive Lebesgue measure set V in the phase space such that for every continuous observable φ, we have

$$\frac{1}{n}\sum_{i=0}^{n-1}\varphi(f^i x) \to \int \varphi \, d\mu$$

for every $x \in V$. Thus, physical measures are those invariant measures that can be *observed*: Suppose we pick an initial condition $x \in V$ and plot the first n points of its trajectory. The resulting picture can be seen as that of a measure giving weight $1/n$ to each one of the points $x, f(x), f^2(x), \cdots, f^{n-1}(x)$. For n large, it is a good approximation of μ.

One of the reasons—some would say the main reason—why SRB measures are important is that all ergodic SRB measures with no zero Lyapunov exponents are physical measures.

Not all physical measures are SRB measures, however. For examples, point masses on attractive fixed points are physical measures. Unlike SRB measures, which are associated with chaotic behavior and have rich, well-defined structures, physical measures have no identifiable characteristics aside from the fact that they are observable.

Zero-noise Limits

If one subscribes to the view that the world is inherently noisy, then the following notion proposed by Kolmogorov is perhaps the most relevant notion of observability. Let f be a map on a bounded domain, and let $P^\varepsilon(\cdot\,|\,\cdot)$, $\varepsilon > 0$ be a family of Markov chains representing random perturbations of f; that is, the transition probabilities $P^\varepsilon(\cdot\,|\,x)$ have the following interpretation: instead of jumping from x to $f(x)$, $P^\varepsilon(\cdot\,|\,x)$ gives the distribution of possible images starting from x. We may think of it as the uniform distribution on an ε-disk centered at $f(x)$ or a Gaussian with ε variance. Let μ_ε be the marginal of the stationary measure of the process defined by $P^\varepsilon(\cdot\,|\,\cdot)$, and let $\varepsilon \to 0$. Limit points of μ_ε are called *zero-noise limits*.

Unlike SRB measures or physical measures, the existence of which leads to unresolved questions, systems defined on compact regions always have zero-noise limits. It is hoped that in most situations, they coincide with the other notions of natural invariant measures. This has been proved for Axiom A attractors and a handful of other examples.

LAI-SANG YOUNG

See also **Cat map; Horseshoes and hyperbolicity in dynamical systems; Measures; Phase space**

Further Reading

Bowen, R. 1975. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Berlin and New York: Springer

Eckmann, J.-P. & Ruelle, D. 1985. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57: 617–656

Ruelle, D. 1978. *Thermodynamic Formalism*, Reading, MA: Addison-Wesley

Sinai, Ya. G. 1972. Gibbs measures in ergodic theory. *Russian Mathematical Surveys*, 27: 21–69

Smale, S. 1967. Differentiable dynamical systems. *Bulletin of the American Mathematical Society*, 73: 747–817

Young, L.-S. 1993. Ergodic theory of differentiable dynamical systems. *Proceedings of the NATO Advanced Study Institute* held in Hillerød, June 20–July 2, 1993, edited by B. Branner & P. Hjorth, Dordrecht: Kluwer

Young, L.-S. 2002. What are SRB measures, and which dynamical systems have them? *Journal of Statistical Physics*, 108: 733–754

SINE-GORDON EQUATION

In normalized units, the classical sine-Gordon (SG) equation is the nonlinear partial differential equation

$$\frac{\partial^2\phi}{\partial x^2} - \frac{\partial^2\phi}{\partial t^2} = \sin\phi, \tag{1}$$

where $\phi(x,t)$ is a field in two dimensions—space and time. Since its introduction in 1939 by Yakov Frenkel and Tatiana Kontorova as a model for the dynamics of crystal dislocations (Frenkel & Kontorova, 1939), this equation has found a variety of applications, including Bloch wall dynamics in ferromagnetics and ferroelectrics, fluxon propagation in long Josephson (superconducting) junctions, self-induced transparency in nonlinear optics, spin waves in the A-phase of liquid ^{3}He at temperatures near to 2.6 mK, and a simple, one-dimensional model for elementary particles (Bullough, 1977; Scott, 2003). The name stems from the linear Klein–Gordon equation $\partial^2\phi\,/\,\partial x^2 - \partial^2\phi\,/\,\mathrm{d}t^2 = \phi$, where the "joke" may be due to Martin Kruskal or it may not (Bullough & Caudrey, 1980).

The SG equation can be physically modeled as a linear array of weakly coupled pendula, suggesting traveling-wave solutions of the form

$$\phi = 4\arctan\exp\left\{\pm\,(x - vt)\,/\sqrt{1-v^2}\right\} \tag{2}$$

with $v < 1$. These solutions are called a kink and an antikink, respectively, since at $x = -\infty, \phi \to 0$, and at $x = +\infty$, $\phi \to 2\pi$ for the $+$ sign, while at $x = -\infty, \phi \to 2\pi$ and at $x = +\infty$, $\phi \to 0$ for the $-$ sign. In the context of the physical model, ϕ twists from 0 to 2π for the kink and untwists from 2π to 0 for the antikink. Alternatively $\phi\,/\,2\pi$ has "topological charge" $+1$ for each kink and -1 for each antikink. Each of these is a one-soliton solution, but there are also N-soliton solutions for the boundary conditions $\phi \to 0$ (mod 2π) as $x \to \pm\infty$. (In "light-cone coordinates" (see below) the N-soliton solutions were given by Caudrey et al. (1973); see the detailed history in Bullough & Caudrey (1980)).

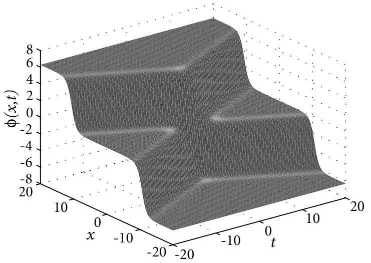

Figure 1. A kink-kink solution of the SG equation plotted from Equation (3) with $v = 0.5$.

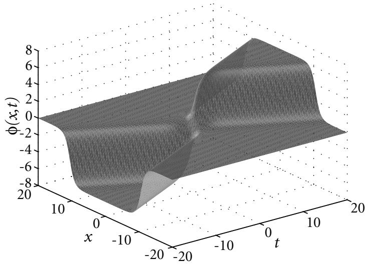

Figure 2. A kink-antikink solution of the SG equation plotted from Equation (4) with $v = 0.5$.

The simplest 2-soliton solution of SG Equation (1) is the kink-kink collision given by Perring & Skyrme (1962) as

$$\phi(x,t) = 4\arctan\left[\frac{v\sinh(x/\sqrt{1-v^2})}{\cosh\,(vt/\sqrt{1-v^2})}\right] \tag{3}$$

as shown in Figure 1 for $v = 0.5$. Similarly, Figure 2 shows a kink-antikink collision, which is plotted from

$$\phi(x,t) = 4\arctan\left[\frac{\sinh(vt/\sqrt{1-v^2})}{v\cosh\,(x/\sqrt{1-v^2})}\right] \tag{4}$$

also with $v = 0.5$.

The kink-antikink equation takes an interesting form if the velocity parameter (v) is allowed to be imaginary. For example, setting

$$v = \mathrm{i}\omega/\sqrt{1-\omega^2}\,, \quad \omega < 1,$$

Equation (4) becomes the stationary breather (or "bion")

$$\phi(x,t) = 4\arctan\left[\left(\frac{\sqrt{1-\omega^2}}{\omega}\right)\frac{\sin\,\omega t}{\cosh\sqrt{1-\omega^2}x}\right], \tag{5}$$

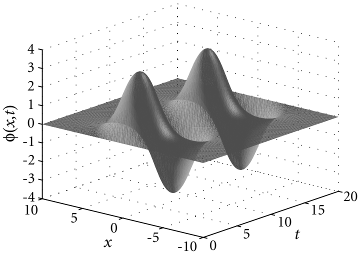

Figure 3. A stationary breather plotted from Equation (5) with $\omega = \pi/5$.

a localized, oscillating solution that is plotted in Figure 3 for $\omega = \pi/5$. Because Equation (1) is invariant under the Lorentz transformation

$$x' = (x - vt)/\sqrt{1 - v^2}$$

and

$$t' = (t - xv)/\sqrt{1 - v^2}, \tag{6}$$

this stationary breather can be boosted into a moving frame. Thus,

$$\phi(x,t) = 4\arctan\left\{\frac{\sqrt{1-\omega^2}}{\omega}\sin\left[\frac{\omega(t - v_{\mathrm{e}}x)}{\sqrt{1 - v_{\mathrm{e}}^2}}\right] \times \operatorname{sech}\left[\frac{\sqrt{1-\omega^2}(x - v_{\mathrm{e}}t)}{\sqrt{1 - v_{\mathrm{e}}^2}}\right]\right\}$$

is an exact solution of the SG equation moving with an envelope velocity (v_{e}) that is equal to the reciprocal of its carrier velocity $v_{\mathrm{c}} = 1/v_{\mathrm{e}}$.

Because $\phi(x,t)$ is defined for each value of x, the Hamiltonian description is infinite dimensional. A standard form of the Hamiltonian (or energy) is (Bullough & Timonen, 1995)

$$H = \int\left[\frac{1}{2}\Pi^2(x,t) + \frac{1}{2}(\partial\phi/\partial x)^2 + (1 - \cos\phi)\right]\mathrm{d}x. \tag{7}$$

Hamilton's equations then take the form

$$\frac{\delta H}{\delta\phi} = -\dot{\Pi} \text{ and } \frac{\delta H}{\delta\Pi} = \dot{\phi}, \tag{8}$$

where the δs indicate functional (or Frechêt) derivatives. The second of Equations (8) yields $\dot{\phi} = \Pi$ and the first yields $-\ddot{\phi} = -\dot{\Pi} = -\phi_{xx} + \sin\phi$, which is the SG equation. Computing H in the rest frame, one finds rest masses equal to 8 for both kinks and antikinks, while for stationary breathers the rest masses are $16\sqrt{1-\omega^2}$. Evidently, a stationary breather is a bound pair of a kink and an antikink oscillating at frequency ω.

The SG equation is a completely integrable Hamiltonian system, which means that the number of independent constants of motion is equal to the number of degrees of freedom (Liouville's theorem). As the number of degrees of freedom is infinite (through the labels x), the theorem is not obvious, but an infinite number of commuting and independent constant action variables can be found. Equivalently, solutions of the SG equation can be obtained through a Bäcklund transformation (BT) or by using an inverse scattering method (ISM).

In BT and ISM studies, it is convenient to transform the independent variables as

$$x \to \xi = (x + t)/2 \text{ and } t \to \tau = (x - t)/2,$$

whereupon Equation (1) becomes

$$\frac{\partial^2\phi}{\partial\xi\,\partial\tau} = \sin\phi. \tag{9}$$

These new variables (ξ and τ) are sometimes called "light-cone" coordinates because they point in the directions of characteristics in the (x,t)-plane.

The SG equation first appeared in light-cone coordinates in the mathematical study of surfaces of constant negative Gaussian curvature by Albert Bäcklund (Lamb, 1976). In 1883, he discovered the first BT, which can be written as

$$\begin{aligned}\phi'_x &= \phi_x + 2k\sin(\phi + \phi')/2,\\ \phi'_t &= -\phi_t + 2k^{-1}\sin(\phi - \phi')/2\end{aligned} \tag{10}$$

for any real value of the parameter k. The integrability condition $\phi'_{xt} = \phi'_{tx}$ implies both $\phi_{xt} = \sin\phi$ and $\phi'_{xt} = \sin\phi'$; thus, Equations (10) transform a known solution ϕ of the SG in light-cone coordinates to a second solution ϕ'. As $\phi = 0$ is a solution, the single-kink solution is found to be $\phi' = 4\arctan\exp(kx - k^{-1}t)$ for this real value of k.

Bullough (1980), for example, shows how the BT Equation (10) becomes the Lax pair for an ISM analysis first written down by Ablowitz et al. (1973) (AKNS). Independently, Takhtajan & Faddeev (1974) gave the corresponding expressions for the covariant SG Equation (1). AKNS's Lax pair is

$$\frac{\partial}{\partial\xi}\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix} = \begin{bmatrix}-\mathrm{i}\lambda & -\phi_\xi/2\\ \phi_\xi/2 & \mathrm{i}\lambda\end{bmatrix}\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix},$$

where ψ_1 and ψ_2 are components of the scattering solution and λ is a complex scattering parameter together with the τ-dependence

$$\frac{\partial}{\partial\tau}\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix} = \frac{\mathrm{i}}{4\lambda}\begin{bmatrix}\cos\phi & \sin\phi\\ \sin\phi & -\cos\phi\end{bmatrix}\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix}. \tag{11}$$

It is readily checked that the cross-derivative condition

$$\frac{\partial}{\partial\tau}\frac{\partial}{\partial\xi}\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix} = \frac{\partial}{\partial\xi}\frac{\partial}{\partial\tau}\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix}$$

implies Equation (9), thus providing the basis for an ISM analysis.

Assuming that $\phi \to 0 \pmod{2\pi}$ as $\xi \to \pm\infty$, the time evolution matrix defined in Equation (11) takes the simple asymptotic form

$$\frac{\mathrm{i}}{4\lambda}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix};$$

thus the reflection coefficient and the residues r_n of its upper-half-plane-poles (in the λ plane) evolve with time as

$$\begin{aligned} b(\lambda,\tau) &= b(\lambda,0)\mathrm{e}^{-\mathrm{i}\tau/2\lambda}, \\ r_n(\tau) &= r_n(0)\mathrm{e}^{-\mathrm{i}\tau/2\lambda_n}, \end{aligned}$$

where $b(\lambda,\tau)$ is the reflection coefficient of the solution and n is an index that runs over the number (N) of soliton components.

The evolved solution of Equation (9) turns out to be

$$\frac{\partial\phi}{\partial\xi}(\xi,\tau) = 4K(\xi,\xi;\tau), \tag{12}$$

where $K(\xi,z;\tau)$ is a solution of the (Gel'fand–Levitan–Marchenko) integral equation

$$\begin{aligned} K(\xi,z;\tau) &= B^*(\xi+z;\tau) \\ &\quad -\int_\xi^\infty\int_\xi^\infty K(\xi,y;\tau)B(y+y';\tau) \\ &\quad \times B^*(z+y';\tau)\,\mathrm{d}y\,\mathrm{d}y' \end{aligned} \tag{13}$$

with $z > \xi$, and

$$\begin{aligned} B(\xi+z;\tau) &\equiv \frac{1}{2\pi}\int_{-\infty}^{\infty} b(\lambda,0)\mathrm{e}^{\mathrm{i}\lambda(\xi+z)-\mathrm{i}\tau/2\lambda}\,\mathrm{d}\lambda \\ &\quad -\mathrm{i}\sum_{n=1}^{N} r_n(0)\mathrm{e}^{\mathrm{i}\lambda_n(\xi+z)-\mathrm{i}\tau/2\lambda_n} \end{aligned} \tag{14}$$

is determined from a scattering analysis of the initial conditions.

The simplest example of this ISM formulation is obtained by assuming that the initial potential is reflectionless ($b(\lambda,\tau)=0$) and has but one bound state ($N=1$), corresponding to a single pole at ($\lambda_1 = \mathrm{i}\kappa$) on the imaginary axis of the upper half λ-plane with residue $r_1 = \pm 2\mathrm{i}\kappa$. Thus, from Equation (14)

$$B(\xi+z;\tau) = \pm 2\kappa\mathrm{e}^{-\kappa(\xi+z)-\tau/2\kappa},$$

so Equation (13) takes the form

$$\begin{aligned} K(\xi,z;\tau) &= \pm\, 2\kappa\mathrm{e}^{-\kappa(\xi+z)-\tau/2\kappa} \\ &\quad -4\kappa^2\mathrm{e}^{-\tau/\kappa}\int_\xi^\infty\int_\xi^\infty K(\xi,y;\tau)\mathrm{e}^{-\kappa(y+y')} \\ &\quad \times\mathrm{e}^{-\kappa(z+y')}\,\mathrm{d}y\,\mathrm{d}y'. \end{aligned}$$

As $K(\xi,z;\tau) \propto \exp(-\kappa z)$, this integral equation is solved for

$$K(\xi,z;\tau) = \pm\frac{2\kappa\exp[-\kappa(\xi+z)-\tau/2\kappa]}{1+\exp(-4\kappa\xi-\tau/\kappa)}.$$

Thus, from Equation (12)

$$\begin{aligned} \frac{\partial\phi}{\partial\xi} &= \pm 4K(\xi,\xi;\tau) \\ &= \pm 4\kappa\,\mathrm{sech}(2\kappa\xi+\tau/2\kappa), \end{aligned}$$

which integrates to

$$\phi(\xi,\tau) = 4\arctan\{\exp[\pm(2\kappa\xi+\tau/2\kappa)]\}.$$

Finally, one can transform back to the laboratory (x,t) coordinates, whereupon the corresponding solution of Equation (1) is the kink or antikink of Equation (2) with its laboratory velocity identified as

$$v = \frac{1-4\kappa^2}{1+4\kappa^2} = \frac{1+4\lambda_1^2}{1-4\lambda_1^2}. \tag{15}$$

With an appropriate choice of the constants c_1 and c_2, the kink-antikink solution of Equation (4) can be generated in a similar manner from the assumption that

$$B(\xi+z;0) = c_1\mathrm{e}^{-\kappa_1(\xi+z)} + c_2\mathrm{e}^{-\kappa_2(\xi+z)}, \tag{16}$$

where

$$\kappa_1 = \frac{1}{2}\sqrt{\frac{1+v}{1-v}} \quad\text{and}\quad \kappa_2 = \frac{1}{2}\sqrt{\frac{1-v}{1+v}}.$$

These integrable properties of the SG equation are shared by the sinh-Gordon equation and the Liouville equation for which the sine function in Equation (1) is replaced by a hyperbolic-sine function and by an exponential function, respectively—although the Liouville equation has no inverse scattering solution. Perhaps the most striking property of SG solitons, however, is their topological stability, which is evidenced by their nonzero rest masses. This property is carried into three space dimensions by the theory of skyrmions.

Equation (1) can be embedded in a more general framework of inverse-scattering methods, with the Hamiltonian expressed in terms of finite numbers of kinks, antikinks, and breathers, in addition to a continuous spectrum of radiation (Bullough, 1980). In the mid-1970s, Roger Dashen, Brosl Hasslacher, and Andre Neveu used semiclassical quantum methods to study this system, finding the mass (or energy) spectrum of the breather of Equation (5) to be

$$M_n = 16\sin\left(\frac{n}{16}\right),$$

where n is a positive integer less than 8π (Dashen et al., 1975). They correctly conjectured this "DHN spectrum" to be exact—a result of theoretical significance because it shows that semiclassical analyses of nonlinear field theories can provide useful information about exact mass spectra.

The statistical mechanics of the SG equation is carried out in both the quantum and classical cases in Bullough & Timonen (1995), with a further appraisal of the action-angle variables in one space dimension. Bullough (1977) develops the theory of double and multiple SG equations (with $\sin\phi$ in Equation (1) replaced by $\pm[\sin\phi + (\sin\phi\,/\,2)\,/\,2]$, $\sin\phi + (\sin\phi\,/\,3)\,/\,3 + 2(\sin 2\phi\,/\,3)\,/\,3$, etc.) in connection with models for degenerate self-induced transparency of metal vapors and of spin waves of the B-phase of liquid ^{3}He near 2.6 mK.

Robin Bullough

See also **Bäcklund transformations; Hamiltonian systems; Inverse scattering method or transform; Laboratory models of nonlinear waves; Long Josephson junctions; Maxwell–Bloch equations; Pendulum; Skyrmions**

Further Reading

Ablowitz, M.J., Kaup, D.J., Newell, A.C. & Segur, H. 1973. The initial value solution for the sine-Gordon equation. *Physical Review Letters*, 30: 1262–1264

Bullough, R.K. 1977. Solitons. In *Interaction of Radiation with Condensed Matter*, vol.1, Vienna: International Atomic Energy Agency; pp. 381–469

Bullough, R.K. 1980. Solitons: inverse scattering and its applications. Five lectures in the Proceedings of the NATO Advanced Study Institute *Bifurcation phenomena in mathematical physics and related phenomena Cargèse, Corsica, June 1977*, edited by D. Bessis and C. Bardos, Dordrecht: Reidel, pp. 295–349

Bullough, R.K. 2000. The optical solitons of QE1 are the BEC of QE14: has the quantum soliton arrived? Special Foundation Lecture: a personal view. 14th National Quantum Electronics and Photonics Conference, Owens Park, University of Manchester, 8 September 1999. *Journal of Modern Optics*, 47(11): 2029–2065 (Erratum. 2001. *Journal of Modern Optics*, 48(4): 747–748)

Bullough, R.K. & Caudrey, P.J. 1980. The soliton and its history. *Topics in Current Physics*, 17: 1–64

Bullough, R.K. & Timonen, J. 1981. Breather contributions to the dynamical form factors of the sine-Gordon systems CsNiF$_3$ and (CH$_3$)$_4$(NiMnCl)$_3$ (TMMC). *Physics Letters*, 82A(82): 182–186

Bullough, R.K. & Timonen, J.T. 1995. Quantum and classical integrable models and statistical mechanics. In *Statistical Mechanics and Field Theory*, edited by V.V. Bazhanov and C.J. Burden, Singapore: World Scientific, pp. 336–414 (Note the 'Solitons' map on pages 358–359)

Caudrey, P.J. 1989. In *Soliton Theory: A Survey of Results*, edited by A.P. Fordy, Manchester: Manchester University Press, and references therein

Caudrey, P.J., Gibbon, J.D., Eilbeck, J.C. & Bullough, R.K. 1973. Exact multi-soliton solutions of the self-induced transparency and sine-Gordon equations. *Physical Review Letters*, 30: 237–238

Dashen, R.F., Hasslacher, B. & Neveu, A. 1975. Semiclassical bound states in an asymptotically free theory. *Physical Review D*, 12: 2443–2458

Dodd, R.K. & Bullough, R.K. 1979. The generalised Marchenko equation and the canonical structure of the AKNS inverse method. *Physica Scripta*, 20: 514–530

Frenkel, J. & Kontorova, T. 1939. On the theory of plastic deformation and twinning. *Journal of Physics (USSR)* 1: 137–149

Lamb, G.L., Jr. 1976. Bäcklund transforms at the turn of the century. In *Bäcklund Transforms*, edited by R.M. Miura, New York: Springer

Perring, J.K. & Skyrme, T.R.H. 1962. A model unified field equation. *Nuclear Physics*, 31: 550–555

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Takhtajan, L.A. & Faddeev, L.D. 1974. Essentially nonlinear one-dimensional model of classical field theory. *Theoretical and Mathematical Physics*, 47: 1046–1057

SINGULAR PERTURBATION THEORY

See **Perturbation theory**

SINGULAR POINTS

See **Phase space**

SINGULARITY THEORY

Does the view from your window encompass a range of hills or mountains? If you are so blessed, you will see immediately both of the persistent singularities that can be observed generically; they are indicated on the mountains sketched in Figure 1, a section of a remote but important range called the Transversal Alps. The points marked "country" are known as fold singularities. Actually the outline of the mountains, which in reality is a mapping of a surface contour onto the retina of the eye, is composed of an infinite number of fold singularities. The two points marked "western" are known as cusp singularities. They occur where a line of folds along the edge of a mountain meets another line of folds along the gully between the mountains.

If the view from your office is more urban in character—of faces in the street, say—you will probably see many more of these two fundamental singularities. You must search a lot harder to find other types of singularities, or purposefully create one, because all other singularities dissolve under the slightest perturbation into either a fold or a cusp. This remarkable fact was first proved by Whitney (1955) whose fundamental discoveries about singularities of differentiable mappings were developed into catastrophe theory and toolkits for treating bifurcation problems with parameters, by mathematicians such as Mather (in a series of very technical papers from 1968 to 1971), Thom (1972), Martinet (1982), Arnol'd et al. (1985), and Golubitsky & Schaeffer (1985).

Singularity theory is not secret mathematicians' business though, and a more apt name for the whole business would be "theory and applications of

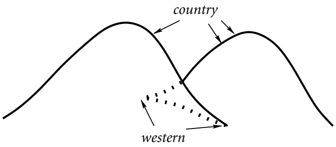

Figure 1. The Transversal Alps.

singularities." It is one of the more accessible entry points both to highly abstract areas of mathematics and to applied fields such as dynamical systems and bifurcations, because singularities can arise in almost any problem. The prerequisites are standard fare in first- and second-year mathematics courses: knowledge of Taylor's formula, the implicit function theorem, the theorem of existence and uniqueness, and some basic group theory, and willingness to learn some terminology.

Naturally, we should begin with a definition of singularity; (after Lu, 1976): Let f be a differentiable mapping from $\mathcal{M}$ to $\mathcal{N}$, where $\mathcal{M}$ and $\mathcal{N}$ are differentiable manifolds. A point $x_0 \in \mathcal{M}$ is a singular point of f if rank $\mathrm{d}f(x_0) < \min\{\dim \mathcal{M}, \dim \mathcal{N}\}$, where $\mathrm{d}f(x_0)$ is the Jacobian matrix of f at x_0. Otherwise x_0 is a regular point of f.

Singularity theory solves three key related problems: Given a mapping f, (i) it determines what types of singularities any good approximation $\bar{f}$ to f must have; (ii) it tells us how can we perturb f slightly to obtain a nicer and simpler, but in some sense equivalent, mapping; and (iii) it provides a taxonomy of singular objects and a binary key to identify them; hence, it is a *classification* science.

At the heart of singularity theory is a concept that is profound and yet somehow ingenuous; it is the concept of transversality. The naive (but not impercipient) version says that two curves intersect transversally if a small deformation of either one would not change the type of intersection. It is transversality that allows us to boil things down to classified normal forms.

Example 1: The mapping f from $\mathbb{R}^2$ into $\mathbb{R}^2$

$$u = x^2, \quad v = y \tag{1}$$

has a fold along $x = 0$ in the xy plane, the set of singular points is $\{(x, y) \in \mathbb{R}^2 \mid y = 2x\}$. Points along the fold remain fixed as y changes, that is, under perturbation. Equations (1) are called the *normal form* for a fold.

Example 2: The equations

$$u = xy - x^3, \quad v = y \tag{2}$$

define a mapping g from $\mathbb{R}^2$ into $\mathbb{R}^2$ for which the set of singular points is $\{(x, y) \in \mathbb{R}^2 \mid y = 3x^2\}$. The arms of the parabola are two lines of folds that come together and disappear at the cusp (0, 0). Equations (2) are the normal form for the cusp, and Whitney proved that any other mapping containing a regular point satisfying the conditions

$$\begin{aligned} &u_x = u_y = v_x = 0, \quad v_y = 1, u_{xx} = 0, \\ &u_{xy} \neq 0, u_{xxx} - 3u_{xy}v_{xx} \neq 0 \end{aligned} \tag{3}$$

can be transformed by coordinate changes into the normal form (2). In catastrophe theory this normal form becomes the *universal unfolding* $G(x, y, u)$ of the germ $g(x) = x^3$:

$$G(x, y, u) \equiv x^3 - yx + u, \tag{4}$$

where G is the gradient of a governing potential V.

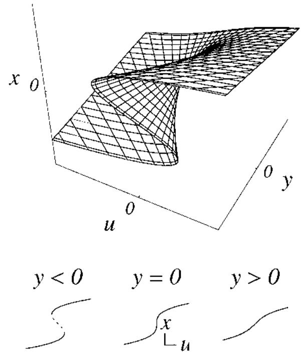

Figure 2. Slices through the cusp manifold (top) yield the three possible bifurcation diagrams (bottom).

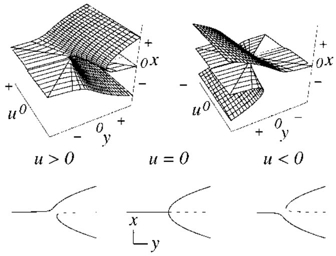

Figure 3. An orthogonal path through that cusp opens into a manifold around the pitchfork.

One may understand the cusp by studying the surface $G(x, y, u) = 0$ shown in Figure 2. By taking slices of this surface at constant y we recover the three qualitatively different bifurcation diagrams, as shown. Now visualize a *projection* of this surface onto the (u, y) plane. We find that the two lines of folds meet at a cusp singularity, the very same that we saw in the Transversal Alps in Figure 1.

An instructive and fascinating lesson on the properties and classification of simple singularities is to parameterize the surface $G(x, y, u) = 0$ differently. In a study of singular surfaces by Ball (2001), it was observed that although the cusp is generic in the sense that all other singularities may be perturbed to either a fold or a cusp, the surface in Figure 2 is not a unique manifold of the cusp. Because all paths through the unfolding (4) are equally valid, we may choose a path in the (x, y) plane. Any such path unfurls laterally into the u-dimension to form a different surface. It is shown from two points of view in Figure 3. Constant-u slices show up the *pitchfork* singularity (center bifurcation diagram) and two of its perturbations. From this point of view Equation (4) is a *partial* unfolding of the pitchfork (but it is not a universal unfolding of the pitchfork).

In applications, the singularity theory approach has been most successful in qualitative studies of the equilibria of dynamical systems dependent on parameters. Given a dynamical system that can be reduced to a set of ordinary differential equations (by a procedure such as Lyapunov–Schmidt reduction), a general approach is to apply defining algebraic criteria systematically to the equilibria until one discovers the highest-order or most degenerate singularity, defined by its normal form. (The binary key given in Golubitsky & Schaeffer (1985, p. 201) is extremely useful for this task.) From a universal unfolding of this *organizing center*, one can "read off" all of the possible qualitatively different bifurcation behavior of the equilibria. A great many dynamical models of physical systems have been given the singularity theory treatment, often yielding results having important implications for the prediction and control of such systems. For a few but varied examples see Ball (1999) (chemical reactions), Ball et al. (2002) (plasma physics), and Broer et al. (2003) (periodic dynamics).

ROWENA BALL

See also **Bifurcations; Catastrophe theory; Development of singularities**

Further Reading

Arnol'd, V.I., Gusein-Zade, S.M. & Varchenko, A.N. 1985. *Singularities of Differentiable Maps*, vol. 1. Boston and Basel: Birkhäuser

Ball, R. 1999. The origins and limits of thermal steady state multiplicity in the continuous stirred tank reactor. *Proceedings of the Royal Society of London Series* A, 455: 141–161

Ball, R. 2001. Understanding critical behaviour through visualisation: a walk around the pitchfork. *Computer Physics Communications*, 142: 71–75

Ball, R., Dewar, R.L. & Sugama, H. 2002. Metamorphosis of plasma shear flow–turbulence dynamics through a transcritical bifurcation. *Physical Review* E, 66: 066408-1–066408-9

Broer, H.W., Golubitsky, M. & Vegter, G. 2003. The geometry of resonance tongues: a singularity theory approach. *Nonlinearity*, 16: 1511–1538

Golubitsky, M. & Schaeffer, D.G. 1985. *Singularities and Groups in Bifurcation Theory*, vol. 1, New York: Springer

Lu, Y.-C. 1976. *Singularity Theory and an Introduction to Catastrophe Theory*, Berlin and New York: Springer

Martinet, J. 1982. *Singularities of Smooth Functions and Maps*, Cambridge and New York: Cambridge University Press

Thom, R. 1972. *Stabilité Structurelle et Morphogénése: Essai d' une Théorie Générale des Modèles*, Reading, MA: W.A. Benjamin

Whitney, H. 1955. On singularities of mappings of Euclidean spaces. I. Mappings of the plane into the plane. *Annals of Mathematics*, 62: 374–410

SINK

See **Attractors**

S-INTEGRABILITY AND C-INTEGRABILITY

See **Integrability**

SKYRMIONS

As described by quantum field theories, elementary particles are associated with the quantization of small fluctuations around the vacuum and have masses that are proportional to Planck's constant, $\hbar$; thus, in the classical limit ($\hbar \to 0$), the mass goes to zero. The properties of a particle follow from the linearization of the field equations and any nonlinear terms are responsible for the interactions between particles, which are often treated perturbatively in the coupling constants. However, the 1960s saw the emergence of a new approach to quantum field theory in which the fully nonlinear classical field equations were investigated. It was found that for certain theories, the classical nonlinear field equations had static solutions that were particle-like, in the sense that they described stable, localized, finite energy field configurations. To describe elementary particles, the field theory must be Lorentz invariant, so such a static solution can simply be Lorentz boosted to describe a particle in uniform motion. These solutions, which are known as topological solitons (or sometimes just solitons for short, though they should not be confused with the solitons of integrable systems), have some novel features in comparison with elementary particles. Because they are classical solutions, then the mass of a soliton, which is identified with the energy of the static solution, is nonzero even in the limit $\hbar = 0$. Also, because the soliton is a solution of the full nonlinear field equation, it is automatically treated nonperturbatively in the coupling constants of the theory. Quantum corrections to the classical properties of a soliton can be addressed, mainly using semiclassical methods, but the important point is that these corrections are often small, so that the classical nonlinear equations largely determine the properties of a soliton. Despite the obvious differences between elementary particles and solitons, impressive research during the last decade has revealed a great deal of evidence for a remarkable connection between these two kinds of particles, with solitons in some weakly coupled theories appearing to be dual to elementary particles in strongly coupled theories (Olive, 1996). These developments have led to a renewed interest in solitons in field theory, particularly in supersymmetric theories.

As the name suggests, topological solitons arise in theories where there is a topological classification of field configurations, with the soliton being in a different topological sector to the vacuum and, hence, preventing its decay, since time evolution cannot change the conserved topological charge. This is perhaps most easily explained using a simple example. As a toy model we shall consider the static sine-Gordon theory in one space dimension, although our presentation will be slightly unusual in order to make the extension to three spatial dimensions a little more transparent. Consider a two-component unit vector $\boldsymbol{\phi}(x) = (\phi_1, \phi_2)$, so that $\boldsymbol{\phi}$ lies on the circle $\boldsymbol{\phi} \cdot \boldsymbol{\phi} = 1$. If we restrict to static fields, then the sine-Gordon model can be defined by the energy expression

$$E = \int_{-\infty}^{\infty} \left(\frac{1}{2} \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}x} \cdot \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}x} + 1 - \phi_2 \right) \mathrm{d}x. \qquad (1)$$

For finite energy the boundary condition is clearly that the field must take the constant vacuum value $\boldsymbol{\phi} = (0, 1)$ at spatial infinity. The fact that the field $\boldsymbol{\phi}$ takes the same value at the two points at spatial infinity implies a compactification of space from $\mathbb{R}$ to S^1, arising from the identification of the points $x = -\infty$ and $x = +\infty$. Therefore, $\boldsymbol{\phi}$ is a map $\boldsymbol{\phi} : S^1 \mapsto S^1$, where the domain is compactified space and the target circle is the set of two-component unit vectors. Such maps have an associated degree, or winding number, N, due to the homotopy group relation $\pi_1(S^1) = \mathbb{Z}$. This integer-valued winding number is simply the number of times (counted with orientation) that the field winds around the target space circle as x ranges over the circle obtained from the compactification of space. The vacuum solution, with energy $E = 0$, is given by $\boldsymbol{\phi}(x) = (0, 1)$ and clearly has $N = 0$. The soliton is the minimal energy field configuration in the $N = 1$ sector and has the explicit expression

$$\boldsymbol{\phi} = (\sin\psi, \cos\psi) \quad \text{with} \quad \psi = 4\tan^{-1} \mathrm{e}^{x-a}, \qquad (2)$$

where a is an arbitrary real constant. The soliton has energy $E = 8$ and the energy density (the integrand in Equation (1)) is a localized lump centered around the point $x = a$, which is the position of the soliton. The full time-dependent sine-Gordon theory follows from the Lorentz invariant Lagrangian associated with the energy (1). The static soliton solution (2) can be Lorentz boosted to provide a moving soliton solution (with any speed less than the speed of light) of the full second-order time-dependent field equations. The antisoliton solution, with $N = -1$, is simply obtained from the soliton solution by making the replacement $(\phi_1, \phi_2) \mapsto (-\phi_1, \phi_2)$. There are no static multisoliton solutions, that is, with $N > 1$, but time-dependent multisoltion solutions can be found in closed form and describe the elastic scattering of several individual solitons (Drazin, 1983). This remarkable fact follows from the integrability property of the time-dependent sine-Gordon equation, a feature which is not shared by more realistic theories with topological solitons in three space dimensions.

The most important aspect of the sine-Gordon model on the line is the topological classification of finite energy field configurations, which follows from the homotopy group relation $\pi_1(S^1) = \mathbb{Z}$. This is the first property that needs to be generalized when searching for a three-dimensional theory with topological solitons. If we consider a theory for which finite energy implies that the field must be constant at spatial infinity, then Euclidean space $\mathbb{R}^3$ becomes compactified to the three-sphere S^3. This is the analogue of the compactification of the line to the circle described above, and hence, we see that the simplest way to obtain a model with topological solitons is to also make the target space S^3. Then the homotopy group relation $\pi_3(S^3) = \mathbb{Z}$ ensures that there is again an integer-valued topological charge, N, which divides finite energy field configurations into distinct topological sectors. To achieve the target space S^3, the field can be taken to be a four-component unit vector, although an equivalent formulation is to take the field to be an $SU(2)$-valued matrix, since S^3 is also the manifold of the group $SU(2)$. The Skyrme model (Skyrme, 1961) is such a theory and has the static energy

$$E = \int \left\{ \mathrm{Tr}(\partial_i U \partial_i U^{-1}) - \frac{1}{8} \mathrm{Tr}\left([(\partial_i U) U^{-1}, (\partial_j U) U^{-1}]^2 \right) \right\} \mathrm{d}^3x, \qquad (3)$$

where $U(\mathbf{x}) \in SU(2)$ is the field of the model and $\partial_i = \partial/\partial x_i$ with x_i $(i = 1, 2, 3)$ is the Cartesian coordinates in Euclidean space, and we have adopted the Einstein summation convention where repeated indices are summed over. Also, the square brackets denote the commutator, $[A, B] = AB - BA$. The first term in (3), called the sigma model energy by physicists and the harmonic map energy by mathematicians, is the higher-dimensional analog of the simple gradient energy given by the first term in (1) for the toy model. In order to provide a finite nonzero scale size for the soliton, the sigma model energy needs to be balanced against an additional term with an appropriate scaling behavior under spatial dilations, as required by the Derrick–Hobart theorem. In one spatial dimension, this requires a term which contains

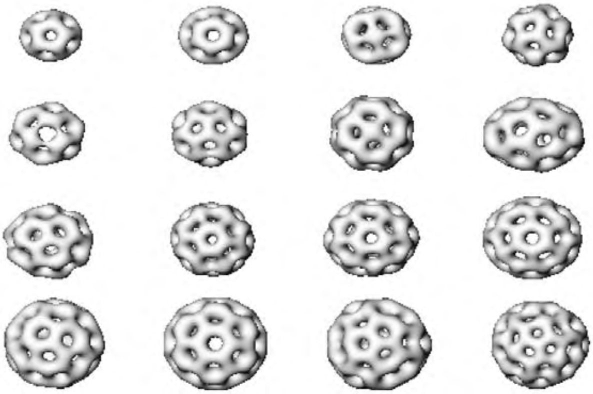

Figure 1. Skyrmions with N from 7 to 22.

no spatial derivatives of the field, such as the final term in (1), but in three spatial dimensions, the appropriate term must be at least fourth-order in the spatial derivatives. The final term in (3), known as the Skyrme term, is the unique order-four expression whose relativistic extension provides a Lagrangian in $(3+1)$-dimensions, which yields a nonlinear field equation that is only second order in time derivatives. The Skyrme model has a stable topological soliton, the solution with minimal energy in the $N=1$ sector, and this is known as a skyrmion. The solution is spherically symmetric but cannot be written in closed form and is only known numerically. Unlike the sine-Gordon toy model, the Skyrme model has static stable bound-state multi-solitons for all $N>0$. These multi-skyrmions are not spherically symmetric but often have surprising discrete symmetries, including the symmetries of the Platonic solids (Battye & Sutcliffe, 2002). Figure 1 displays skyrmions with $7 \leq N \leq 22$ by plotting surfaces around which the topological charge density is concentrated.

The Skyrme model was originally introduced (Skyrme, 1961) in the early 1960s as a model for the strong interactions of hadrons. It is a nonlinear theory of pions, with the pion particles being described in the usual quantum field theory approach by the quantization of the three degrees of freedom associated with the small fluctuations of the field around the vacuum, where U is the identity matrix, with $N=0$. Skyrme identified the conserved topological charge N with baryon number and, hence, within the nonlinear pion theory baryons appear for free as the classical soliton solutions. The Skyrme model was set aside after the advent of quantum chromodynamics (QCD), but much later it was revived by Witten (1983), who showed that it could arise from QCD as a low energy effective description in the limit in which the number of quark colors is large. Semiclassical quantization of the $N=1$ skyrmion reproduces the properties of the nucleon to within an accuracy of around 30% (Adkins et al., 1983), which is quite an achievement. There has been considerable recent progress in computing classical multi-skyrmion solutions, but it still remains to be seen whether their quantization provides a good description of nuclei.

Finally, it should be noted that in Yang–Mills–Higgs gauge theories the topological classification of field configurations arises in a slightly different way than described above for skyrmions and other modified sigma models. In gauged theories, the Higgs field is not required to be constant at infinity, and indeed, it is a nontrivial winding at infinity that provides the topological charge. In two spatial dimensions this yields vortices, and in three-dimensional space it leads to magnetic monopoles. When the Higgs field is massless, which is the situation that arises in interesting supersymmetric theories with monopoles, then the classical field equations have a particularly rich mathematical structure, and many exact results on monopole solutions are known (for a review see, e.g., Sutcliffe, 1997).

PAUL SUTCLIFFE

See also **Derrick–Hobart theorem; Quantum field theory; Sine-Gordon equation; Solitons, types of; Yang–Mills theory**

Further Reading

Adkins, G.S., Nappi, C.R. & Witten, E. 1983. Static properties of nucleons in the Skyrme model. *Nuclear Physics* B, 228: 552–566

Battye, R.A. & Sutcliffe, P.M. 2002. Skyrmions, fullerenes and rational maps. *Reviews in Mathematical Physics*, 14: 29–85
Drazin, P.G. 1983. *Solitons*, Cambridge and New York: Cambridge University Press
Olive, D.I. 1996. Exact electromagnetic duality. *Nuclear Physics Proceedings Supplement* 45A: 88–102
Skyrme, T.H.R. 1961. A nonlinear field theory. *Proceedings of the Royal Society* A, 260: 127–138
Sutcliffe, P.M. 1997. BPS monopoles. *International Journal of Modern Physics* A, 12: 4663–4705
Witten, E. 1983. Current-algebra, baryons, and quark confinement. *Nuclear Physics* B, 223: 433–444

SLAVING PRINCIPLE

See **Synergetics**

SMALE HORSESHOE

See **Horseshoes and hyperbolicity**

SNOWFLAKES

See **Pattern formation**

SOLAR SYSTEM

The solar system shows a multitude of periodic phenomena: from the daily rotations of the Earth, to the monthly phases of the Moon, the seasonal changes during the course of a year, the phases of Venus and other planets, the regular recurrences of comets such as Halley's (every 76 years), the 25,700 year precession of the Earth's rotation axis, all the way up to changes in the parameters of the Earth's orbit, with periods of 40,000 years and more. Closer inspection reveals small modulations on top of these periods, but it is tempting to describe these by additional periods, as in the Ptolemaic theory of cycles, epicycles, and equants. However, the dynamics of the planets is governed by gravitational forces that vary as the inverse square of the distance and, hence, are strongly nonlinear. The question arises whether this nonlinearity results in some chaos.

Johannes Kepler concluded in the early 17th century that an isolated planet would move along a Kepler ellipse (such that a line joining the planet to the Sun sweeps out equal areas in equal times). For several planets their mutual perturbations leads to perihelion precessions and other variations of orbital parameters. Taking such perturbations into account was a major issue in solar mechanics in the following centuries and led to remarkable achievements. For instance, observations of the orbit of Uranus (discovered in 1781 by William Herschel) revealed significant differences to the orbit calculated in the presence of the known planets. When interpreted as due to the influence of another planet, the position of the missing planet could be predicted, and soon thereafter the efforts of Urbain V. LeVerrier, James C. Adams, and Johannes G. Galle were rewarded with the discovery of Neptune in 1846.

These results were obtained using perturbation theory. But does it actually converge? Advances in analytical mechanics in the 19th century suggested that an answer could be found and Gösta Mittag-Leffler included the issue of the stability of the solar system among the list of problems for a prize to be awarded in 1889 by King Oskar II of Sweden and Norway. Henri Poincaré was awarded the prize for his announcement that he could prove convergence, but in the course of revising his paper he noticed that there was a gap in the proof which he could not patch up. Instead, he discovered the intricate motions near a weakly perturbed hyperbolic fixed point, the so-called hyperbolic tangle, which effectively prevents quantitative continuation of trajectories that pass near a hyperbolic fixed point. Barrow-Green (1996) and Diacu & Holmes (1996) give vivid accounts of the events surrounding the prize.

A practical answer to the question of the stability of the solar system emerged in the late 20th century with the advent of powerful computers that allow integration of the equations of motions for several million years. It was then discovered that the inner planets are most susceptible to chaos, and that the Lyapunov time (inverse of the Lyapunov exponent) is about 5 million years. To illustrate the consequences of that, we quote from Laskar (1995):

> A 15 m uncertainty in the position of the Earth will grow to about 150 m in a time of 10 million years. But it will increase to 150 million km or the mean distance of Earth from the Sun within 100 million years.

As a consequence, we have difficulty predicting the Earth's orbit for times much larger than a few tens of millions of years. The consequences such an uncertainty can have are illustrated by numerical simulations for Mercury: by suitably selecting continuations, it is argued in Laskar (1995) that over a time of about 10^9 years the orbit of Mercury could change so that the planet collides with Venus and/or escapes from the solar system. It should be noted that this trajectory was especially tailored in order to show that escape is possible in principle; it does not give a clue as to how likely such an event may be.

The issue of the orbital parameters of the Earth is interesting because variations in the major rotation axis and the distance to the Sun influence our climate. In 1920, Milutin Milankovich calculated insolation data (incident solar radiation) for the Earth for long-term variations in the Earth's orbit and suggested some relationship to the appearance of past ice ages. The uncertainty in orbital parameters over periods of more than 40 million years thus indicate that reconstructing paleoclimates will be problematic (Laskar, 1999).

Besides planets there are many other objects in the solar system. Direct evidence for a chaotic trajectory has been found, for example, for the tumbling motion of Phobos and Deimos (moons of Mars) and for Saturn's moon Hyperion (Wisdom, 1987). Similarly, efforts to retrace the trajectory of Halley's comet back to its earliest recorded sighting in 163 BC gave wildly diverging results. As explained by Chirikov and Vecheslavov (1989), the orbit of Halley is chaotic with an inverse Lyapunov time of about 29 returns, so that the earliest observation is just beyond predictability. The distribution of asteroids between Mars and Jupiter shows conspicuous gaps (named after the astronomer Daniel Kirkwood) near orbits with rotation periods rationally related to the 11.9 year period of Jupiter. Such resonant interactions have strong effects on the orbits and can easily lead to collisions and escape from the resonance (Laskar, 1995; Wisdom, 1987).

The relation between the Moon, the Earth, and the Sun also holds surprises. The history of the problem, and the contributions of Babylonian, Greek, and modern astronomers to the observations and mathematical tools, is reviewed in Gutzwiller (1998). Gutzwiller also gives a quantitative example for the significance of small denominators: in order to calculate the distance between the Earth and Moon with the accuracy of 10^{-10} achievable within the Lunar Laser Ranging project, amplitudes as small as 10^{-17} have to be kept because of a significant resonance. While this does not result in positive Lyapunov exponents, it is a precursor to it, a mild form of chaos, as Gutzwiller calls it.

More surprisingly, the presence of the Moon is very important for the stability of the rotation axis of Earth: with the Moon the obliquity stays within about ±1.3° of 23.3°. Without it, the obliquity ends up in a resonance and can become as large as 60° to 90°, with catastrophic consequences for our climate and the evolution of life (Laskar et al., 1993).

Finally, it is worthwhile to point out that the presence of hyperbolic orbits in the solar system was used in connection with the satellite GENESIS (launched in August 2001) to bring it along a stable orbit close to a Lagrange point, where it will remain for a few years to collect particles in the solar wind, before being brought back to Earth along an unstable manifold (Koon et al., 2000). Exploiting trajectories that exist in the dynamical system allows the mission to be completed with minimal requirements on fuel. This mission may thus also be considered an example of chaos control, where similar ideas are being investigated.

BRUNO ECKHARDT

See also **Celestial mechanics; Controlling chaos; Hénon–Heiles system; N-body problem**

Further Reading

Barrow-Green, J. 1996. *Poincaré and the Three-body Problem*, Providence, RI: American Mathematical Society

Chirikov, B.V. & Vecheslavov, V.V. 1989. Chaotic dynamics of comet Halley. *Astronomy and Astrophysics*, 221: 146–154

Diacu, F. & Holmes, P. 1996. *Celestial Encounters*, Princeton, NJ: Princeton University Press

Gutzwiller, M.C. 1998. Moon–Earth–Sun: the oldest three-body problem. *Reviews of Modern Physics*, 70: 589–639

Koon, W.S., Lo, M.W., Marsden, J.E. & Ross, S.D. 2000. Heteroclinic connections between periodic orbits and resonance transitions in celestial mechanics. *Chaos*, 10: 427–469

Laskar, J. 1995. Large scale chaos and marginal stability in the solar system. In *Constructive Methods and Results*, Proceedings of the XIth ICMP Conference, edited by D. Iagolnitzer, Boston: International Press, pp. 75–120

Laskar, J. 1999. The limits of Earth orbital calculations for geolocical time-scale use. *Philosophical Transactions Royal Society of London*, Series A, 357: 1735–1759

Laskar, J., Joutel, F. & Robutel, P. 1993. Stabilization of the Earth's obliquity by the moon. *Nature*, 361: 615–617

Wisdom, J. 1987. Chaotic motion in the solar system. *Icarus*, 72: 241–275

SOLIDIFICATION PATTERNS

See **Growth patterns**

SOLITON WAVE PACKET

See **Quantum nonlinearity**

SOLITONS

A *soliton* is a localized nonlinear wave that maintains its shape and speed as it travels, even through interaction with other waves. Another name for a localized traveling wave is a *solitary wave*. The term *soliton* was coined (by Zabusky & Kruskal, 1965) to reflect both the solitary-wave-like character and the particle-like interaction properties. Their surprising discovery has had an enormous impact on the field of nonlinear mathematics and science.

In many physical applications, however, the use of the word *soliton* has come to rely on observation of long-lived solitary waves, with little emphasis placed on interaction properties. Jupiter's Great Red Spot is often described as a soliton due to its long-lived identity (observed over hundreds of years) and the fact that it appears to maintain its identity through interaction with other disturbances. However, recent studies indicate that such Jovian vortices do not keep their identity through intrazonal interactions. In nonlinear optics, the deduction of stable, long-lived solitary waves is of great physical interest for their application as coherent light signals propagating through optical fibers, while their interaction properties are secondary. The definition of the word *soliton* in some dictionaries has come to reflect such usage by physicists, omitting the requirement of interaction. However, for mathematicians, the many surprising and miraculous discoveries of soliton theory are tied fundamentally to the special interaction properties so its definition always contains elastic interaction as an essential requirement.

The preservation of identity through interaction would not be surprising if we were describing solutions of linear nondispersive wave equations. A simple example is the linear wave equation $u_{tt} + c^2 u_{xx} = 0$, where c is a constant (assumed positive). The general solution is given by $u(x,t) = f(x - ct) + g(x + ct)$, where f and g are determined by initial conditions. The first part, $f(x - ct)$, is a wave traveling to the right with speed c, while the second part, $g(x + ct)$, is a wave traveling to the left with speed c. Two such wave profiles interact when they meet head-on but they both come out of the interaction with the same shape and speed.

However, until the discovery of solitons, it was generally believed that no such property could hold for nonlinear equations. Common understanding in mathematics and physics in the 1950s suggested that nonlinear wave solutions either break, dissipate, or thermalize, that is, distribute initial energy between different solutions over time, and, therefore, lose their identities with time. Much of this understanding was based on prototypical examples, such as the inviscid Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0, \qquad (1)$$

which describes the one-dimensional propagation of compression waves in gas or dust or automoblie traffic. For the initial condition $u(x,0) = u_0(x)$, the solution is a wave, given implicitly by $u(x,t) = u_0\big(x - t\,u(x,t)\big)$. If the initial wave profile $u_0(x)$ has a part with negative slope, the wave front steepens and eventually breaks in finite time, just like ocean waves do at a beach. (The break up time is easy to find by differentiating the solution with respect to x.)

In 1965, Zabusky and Kruskal published numerical studies of the solutions of the Korteweg–de Vries (KdV) equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \delta^2\frac{\partial^3 u}{\partial x^3} = 0, \qquad (2)$$

which changed the above-described common beliefs about nonlinear waves forever (*See* **Solitons, a brief history**). If $\delta \to 0$, the limiting equation is the inviscid Burgers equation. Zabusky and Kruskal labeled Equation (2) as (1), chose $\delta = 0.022$ and the initial condition $u(x,0) = \cos(\pi x)$ and studied its periodic solutions numerically. The word *soliton* was used for the first time in their paper of 1965, in the following extract.

> (I) Initially, the first two terms of Eq. (1) dominate and the classical overtaking phenomenon occurs; that is, u steepens in regions where it has negative slope. (II) Second, after u has steepened sufficiently, the third term becomes important and serves to prevent the formation of a discontinuity. Instead, oscillations of small wavelength (of order δ) develop on the left of the front. The amplitudes of the oscillations grow, and finally *each* oscillation achieves an almost steady amplitude (that increases linearly from left to right) and has the shape of an individual solitary-wave of (1). (III) Finally, each "solitary wave pulse" or *soliton* begins to move uniformly at a rate (relative to the background value of u from which the pulse rises) that is linearly proportional to its amplitude. Thus, the solitons spread apart. Because of the periodicity, two or more solitons eventually overlap spatially and interact nonlinearly. Shortly after the interaction they reappear virtually unaffected in size or shape. In other words, solitons "pass through" one another without losing their identity.

A standard form of the KdV equation is

$$\frac{\partial u}{\partial t} + 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0. \qquad (3)$$

This is equivalent to Equation (2) under a scaling transformation; that is, $u(x,t) \mapsto 6\delta^{2/3}\,u\big(x\,\delta^{-2/3},\,t\big)$ maps any solution of (2) to one of (3). If we consider the initial value problem for the KdV equation on the whole real x-line and look for solutions that vanish at infinity, then solitons are characterized by the initial condition

$$u(x,0) = N(N+1)\,k^2\,\mathrm{sech}^2\big(k\,x\big), \qquad (4)$$

where N is any nonnegative integer. The corresponding solutions are called N-soliton or multisoliton solutions. The case $N = 1$ gives the traveling wave solution

$$u(x,t) = 2k^2\,\mathrm{sech}^2\big(k\,(x - 4k^2 t)\big), \qquad (5)$$

often called the one-soliton solution of the KdV equation. It is equivalent to the solitary wave observed by John Scott Russell in 1834 and deduced mathematically by Diederik Korteweg and Hendrik de Vries in 1895 (*See* **Solitons, a brief history**).

When $N \geq 2$ and time t is considered to be large negative (a long time into the past) or large positive (a long time into the future), the solution separates into a chain of N distinct, localized one-soliton solutions each having the form (5) for some value of k. The chain is positioned far to the left if t is negative or far to the right if t is positive. Each wave in the chain has a distinct height related to its speed. If their phases are arranged so that a taller soliton is to the left of a shorter one at some time in the past, then the taller one overtakes the shorter one, and reappears to the right with its distinctive height, shape, and speed unchanged. The only explicit sign that interaction has occurred is a phase shift, visible asymptotically as x and t become large, in each soliton. As in the one-soliton case, explicit formulae can be found for the N-soliton solution for all x and t. Moreover, the phase shifts due to pairwise interactions can be calculated exactly (*See* ***N*-soliton formulas**).

The soliton solutions of the KdV equation can be obtained explicitly by using the inverse scattering method. This method associates each solution $u(x,t)$ of Equation (3), that vanishes (faster than $1/x$) as $x \to \pm\infty$, to the stationary Schrödinger equation

$$\psi_{xx} + \big(u(x,t) + \lambda\big)\psi = 0, \qquad (6)$$

where $u(x,\cdot)$ plays the role of the potential function and λ is a parameter. If we let $\lambda = \zeta^2$, we can characterize the space of solutions $\psi(x;\zeta)$ near infinity (in x) in terms of linear combinations of $\exp(\pm \mathrm{i}\,\zeta\, x)$. (Since u vanishes there, the equation becomes simply $\psi_{xx} - \zeta^2\,\psi = 0$, which has such exponentials as solutions.) Quantum mechanical convention regards the solution with behavior $\exp(\mathrm{i}\,\zeta\, x)$ at $x = +\infty$ as an *incoming wave* from $+\infty$. Part of this incoming wave reflects off the potential barrier u while part of it is transmitted through to the other side. Not all boundary conditions at infinity can be satisfied without imposing conditions on ζ. A discrete set of eigenvalues $\{\zeta_i\}$ arises when we demand that the solutions vanish at infinity. The corresponding solutions are called bound states and their amplitudes are usually normalized. The collected information about discrete eigenvalues, the normalization constants, the reflection coefficient, and the transmission coefficient is called the set of spectral or scattering data. A beautiful and fundamental result of soliton theory is that the time evolution of the spectral data can be obtained explicitly as the potential evolves according to the KdV equation. Another fundamental result is that such evolved spectral data can be inverted to give the solution $u(x,t)$ of the KdV equation at a later time.

If all waves ψ are transmitted, with none being reflected, the potential u is called a *reflectionless* potential. It is an amazing fact that the N-soliton solutions of the KdV equation are precisely the reflectionless potentials of Equation (6).

Another way to find the N-soliton solutions of the KdV is through its Bäcklund transform (*See* **Bäcklund transformations**) or equivalently the Darboux transform of the Schrödinger equation (6). The latter was introduced by Gaston Darboux in 1882, and expanded in his four-volume lecture notes on the geometry of surfaces (Darboux, 1915). The Darboux transform relates the potentials of two copies of the Schrödinger equation whose solutions are related linearly. His method shows how to change a potential so that a new discrete eigenvalue is added to those of the old potential (*See* **Darboux transformation**).

Many other nonlinear PDEs are now known to have soliton solutions. Solitons come in different shapes and many flavors (*See* **Solitons, types of**). The sine-Gordon equation

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial t^2} = \sin u \qquad (7)$$

has N-soliton solutions that asymptote to constants as $x \to \pm\infty$. One-soliton solutions can be one of two types, called a "kink" or an "anti-kink," according to whether it decays to zero to the left or the right, respectively. The explicit expression for such solutions is

$$u(x,t) = 4\arctan\left(\exp\left(\pm\frac{x - \eta\, t - x_0}{\sqrt{1-\eta^2}}\right)\right). \qquad (8)$$

The "+" choice gives a kink and the "−" choice an antikink. Multisoliton solutions decompose into a sequence of kinks and antikinks as x and t approach $\pm\infty$.

There is another type of soliton admitted by the SG equation, which appears to make it unique within the class of nonlinear Klein–Gordon equations: $u_{xx} - u_{tt} = F(u)$. This is the "breather" solution:

$$u(x,t) = 4\arctan\left(\frac{\sqrt{1-\omega^2}}{\omega}\,\frac{\sin(\omega\, t)}{\cosh(\sqrt{1-\omega^2}\,(x - x_0))}\right). \qquad (9)$$

This solution oscillates (or "breathes") while staying in the same location as time evolves. However, because SG is invariant under the Lorentz transform: $(x,t) \mapsto (X,T)$, $X = (x - c\,t)/\sqrt{1-c^2}$, $T = (t - c\,x)/\sqrt{1-c^2}$, where c is constant, the breather can be transformed to one that moves with speed $|c| < 1$. The existence of such spatially localized, temporally periodic solutions is rare for nonlinear wave equations.

The nonlinear Schrödinger (NLS) equation

$$\mathrm{i}\,\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + 2\,|u|^2 u = 0 \qquad (10)$$

provides a third example of a nonlinear PDE with N-soliton solutions. Its one-soliton solution is given by

$$\begin{aligned} u(x,t) &= a\,\exp\big(\mathrm{i}\,\eta x + (a^2 - \eta^2)t\big) \\ &\quad \times \operatorname{sech}\Big(a\,(x - 2\eta\, t - x_0)\Big). \end{aligned} \qquad (11)$$

Solitons are not confined to partial differential equations. A differential-difference equation such as the Toda equation

$$\begin{aligned} \frac{\partial^2 u_n}{\partial t^2} &= \exp\,(-(u_n - u_{n-1})) \\ &\quad - \exp\,(-(u_{n+1} - u_n))\,, \end{aligned} \qquad (12)$$

also has N-soliton solutions for arbitrary positive integers N. See Ablowitz & Segur (1981) for inverse scattering theory extended to such discrete evolution equations.

There also exist nonlinear PDEs that have solutions that can be interpreted as 2-soliton solutions, but which do *not* have N-soliton solutions for integer $N > 2$. Hirota pointed this out after inventing a method of calculating solitons explicitly and directly without using the inverse scattering method or Bäcklund transformations. (For details, *see* Ablowitz & Segur (1981) and **Hirota's method**.) An example is the two-dimensional version of the SG equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \frac{\partial^2 u}{\partial t^2} = \sin u\,. \tag{13}$$

Hirota's method shows that there is a traveling wave solution that can be interpreted as a two-soliton. However, no N-soliton solution with higher N appears to exist.

There are many more equations with solitary waves than those with N-soliton solutions. Traveling waves may also be known to exist without having an exact expression. An example is provided by the reduced model of nerve transmission given by the Fitzhugh–Nagumo equation (*See* **FitzHugh–Nagumo equation**).

It can be shown that (for a certain range of parameter values) this equation admits traveling waves that are solitary waves. However, energy is not conserved for nerve impulses while it is for solitons.

Despite their rarity in the class of differential equations, soliton equations arise ubiquitously as models of nature. The KdV equation arises as a canonical model of water waves, under certain constraints. Consider the motion of surface waves in a fluid that is inviscid and nondispersive to leading order. Assume that the height of the waves is small compared with the depth of fluid, which in turn is much smaller than the length scale in one direction (x, say) and, moreover, that these two small ratios of scales balance. Then the resulting model is always given by the KdV equation (or its relations, such as the modified KdV equation). See Ablowitz & Segur (1981) for a detailed derivation. The soliton solutions, in the case when dimensionless surface tension is less than $\frac{1}{3}$, are the waves that travel on and raise the free surface. The general solutions are composed of N-solitons and a part called *radiation* that decays in amplitude as it moves away to infinity. When the water waves are nearly two-dimensional, the model equation becomes the Kadomtsev–Petviashvili equation

$$\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t} + 6\,u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3}\right) + \frac{\partial^2 u}{\partial y^2} = 0\,, \tag{14}$$

which also possesses N-soliton solutions. Since such models always arise in an asymptotic sense (e.g., the balance of small wave height to depth of fluid), stability of solutions under perturbation is important. Solitons are stable under perturbations.

Stratified fluids are another context in which the KdV equation arises as a universal model. Consider a situation where a lower, heavier fluid rests on an impermeable solid with a lighter fluid on top that has a free surface above. A common example is the stratified ocean where colder, denser water lies underneath a warmer, lighter layer. If the sea is contained by a boundary that restricts the colder layer but there are forces present, such as tides, that force the lighter layer, interface waves appear between the two layers. Such waves are also governed by the KdV equation under appropriate assumptions. The soliton solutions are then interface waves that thicken the upper layer. The KdV and its solitons also arise in collision-free hydromagnetic waves, ion-acoustic waves, plasma physics, and lattice dynamics.

Like the KdV equation, the SG and NLS equations also arise as universal models in certain contexts. Some of the applications of the SG equation include the study of propagation of crystal defects, that of domain walls in ferromagnetic and ferroelectric materials, as a one-dimensional model for elementary particles, self-induced transparency of short optical pulses, and propagation of quantum units of magnetic flux on long Josephson (superconducting) transmission lines. See Scott (2003) for detailed descriptions. The NLS equation appears as a model of the propagation of packets of hydrodynamic waves on deep water, nonlinear pulses of light in an optical fibre, two-dimensional self-focusing of a plane wave, one-dimensional self-modulation of a monochromatic wave, propagation of a heat pulse in a solid, and Langmuir waves in plasmas.

NALINI JOSHI

See also **Bäcklund transformations; Inverse scattering method or transform; Solitons, types of; Zero-dispersion limit**

Further Reading

Ablowitz, M. & Segur, H. 1981. *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM

Darboux, G. 1915. *Leĉons sur la théorie générale des surfaces et les applications géometriques du calcul infinitésimal*, vol. 2, 2nd edition, Paris: Gauthier-Villars

Scott, A. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Zabusky, N.J. & Kruskal, M.D. 1965. Interactions of solitons in a collisionless plasma and the recurrence of initial states. *Physical Review Letters*, 15: 240–243

SOLITONS, A BRIEF HISTORY

In 1834, a young engineer named John Scott Russell was conducting experiments on the Union Canal (near Edinburgh, Scotland) to measure the relationship between the speed of a canal boat and its propelling force, with the aim of finding design parameters for conversion from horse power to steam. One August day, a rope parted in his apparatus and (Russell, 1844)

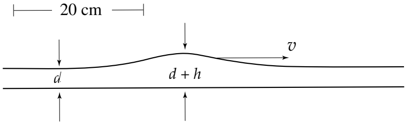

Figure 1. A hydrodynamic solitary wave (or soliton) in a tank similar to that described by John Scott Russell (1844). The wave is generated by suddenly releasing (or displacing) a mass of water at the left-hand side of the tank.

> the boat suddenly stopped—not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel without change of form or diminution of speed.

Russell did not ignore this serendipitous phenomenon, but “followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height” until the wave became lost in the windings of the channel. He continued to study the solitary wave in tanks and canals over the following decade, finding it to be an independent dynamic entity moving with constant shape and speed.

Using a wave tank, he demonstrated four facts Russell (1844). First, solitary waves have a hyperbolic secant shape. Second, a sufficiently large initial mass of water produces two or more independent solitary waves. Third, solitary waves cross each other “without change of any kind.” Finally, a wave of height h and traveling in a channel of depth d has a velocity given by the expression $\sqrt{g(d+h)}$ (where g is the acceleration of gravity), implying that a large amplitude solitary wave travels faster than one of low amplitude.

Although soon confirmed by observations on the Canal de Bourgogne near Dijon, most subsequent discussions of the hydrodynamic solitary wave missed the physical significance of Russell’s observations. Evidence that Russell maintained a deeper appreciation of the importance of his discovery is provided by a posthumous work where—among several provocative ideas—he correctly estimated the height of the Earth’s atmosphere from the fact (well known to military engineers of the time) that “the sound of a cannon travels faster than the command to fire it” (Russell, 1885).

In 1895, Diederik Korteweg and Hendrik de Vries published a theory of shallow water waves that reduced Russell’s problem to its essential features. One of their results was the nonlinear partial differential equation (PDE)

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} + \varepsilon\frac{\partial^3 u}{\partial x^3} + \gamma u\frac{\partial u}{\partial x} = 0, \tag{1}$$

which would play a key role in soliton theory (Korteweg & de Vries, 1895). In this equation, $u(x,t)$ is the wave amplitude, $c = \sqrt{gd}$ is the speed of small amplitude waves, $\varepsilon \equiv c(d^2/6 - T/2\rho g)$ is a dispersive parameter, $\gamma \equiv 3c/2d$ is a nonlinear parameter, and T and ρ are, respectively, the surface tension and the density of water. In general, Equation (1) is nonlinear with exact traveling-wave solutions

$$u(x,t) = h\,\mathrm{sech}^2[k(x - vt)], \tag{2}$$

where $k \propto \sqrt{h}$ implying that higher amplitude waves are more narrow. With this shape, the effects of dispersion balance those of nonlinearity at an adjustable value of the pulse speed; thus, the hydrodynamic solitary wave is seen to be an independent dynamic entity.

Bäcklund Transformations

Although unrecognized at the time, such an energy conserving solitary wave is related to the existence of a transform technique that was proposed by Albert Bäcklund in 1855 (Lamb, 1976). Under this Bäcklund transformation (BT), a known solution generates a new solution through an integration, after which the new solution can be used to generate yet another new solution, and so on. It is straightforward to find a BT for any linear PDE, which introduces a new eigenfunction into the total solution with each application of the transformation. Only special nonlinear PDEs have BTs, but 19th century mathematicians knew that these include

$$\frac{\partial^2 u}{\partial \xi\, \partial \tau} = \sin u, \tag{3}$$

which arose in research on the geometry of curved surfaces (Steuerwald, 1936).

In 1939, Yakov Frenkel and Tatiana Kontorova introduced a seemingly unrelated problem arising in solid state physics to model dislocation dynamics in a crystal (Frenkel & Kontorova, 1939). From this study, an equation describing dislocation motion is

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial t^2} = \sin u, \tag{4}$$

where $u(x,t)$ is atomic displacement in the x-direction and the sine function represents periodicity of the crystal lattice. A traveling-wave solution of Equation (4), corresponding to the propagation of a dislocation, is

$$u(x,t) = 4\arctan\left[\exp\left(\frac{x - vt}{\sqrt{1 - v^2}}\right)\right], \tag{5}$$

with velocity v in the range $(-1,+1)$. Because Equation (4) is identical to Equation (3) after an independent variable transformation, exact solutions involving arbitrary numbers of dislocation components as in Equation (5) can be generated through a succession of Bäcklund transformations, but this was not known to Frenkel and Kontorova.

Numerical Discoveries of the Soliton

In the late 1940s, Enrico Fermi, John Pasta, and Stan Ulam (FPU) suggested one of the first scientific problems to be assigned to the Los Alamos MANIAC computing machine: the dynamics of energy equipartition in a slightly nonlinear crystal lattice, which is related to thermal conductivity. The system they chose was a chain of equal mass particles connected by slightly nonlinear springs, and it was expected that if all the initial energy were put into a single vibrational mode, the small nonlinearity would cause a gradual progress toward equal distribution of the energy among all modes (thermalization). But the numerical results were surprising. If all the energy is originally in the mode of lowest frequency, it returns almost entirely to that mode after a period of interaction among a few other low frequency modes. In the course of several numerical refinements, no thermalization was observed (Fermi et al., 1955).

Pursuit of an explanation for this “FPU recurrence” led Zabusky and Kruskal to approximate the nonlinear spring-mass system by the KdV equation. In 1965, they reported numerical observations that KdV solitary waves pass through each other with no change in shape or speed, and coined the term *soliton* to suggest this particle-like property (Zabusky & Kruskal, 1965).

Zabusky and Kruskal were not the first to observe nondestructive interactions of energy conserving solitary waves. Apart from Russell’s tank measurements, Perring and Skyrme had studied solutions of Equation (4) comprising two solutions as in Equation (5) undergoing a collision. In 1962, they published numerical results showing perfect recovery of shapes and speeds after a collision and went on to discover an exact analytical description of this phenomenon (Perring & Skyrme, 1962).

This result would not have surprised 19th-century mathematicians; it is merely the second member of the hierarchy of solutions generated by a BT. Nor would it have been unexpected by Seeger and his colleagues, who had noted in 1953 the connections between the 19th-century work (Steuerwald, 1936) and the studies of Frenkel and Kontorova (Seeger et al., 1953). Because Perring and Skyrme were interested in Equation (4) as a nonlinear model for elementary particles of matter, however, the complete absence of scattering may have been disappointing.

Throughout the 1960s, Equation (4) arose in a variety of problems, including the propagation of ferromagnetic domain walls, self-induced transparency in nonlinear optics, and the propagation of magnetic flux quanta in long Josephson transmission lines. Eventually it became known as the “sine-Gordon” (SG) equation—a nonlinear version of the Klein–Gordon equation: ($u_{xx} - u_{tt} = u$).

Perhaps the most important contribution made by Zabusky and Kruskal in their 1965 paper was to recognize the relation between nondestructive soliton collisions and the riddle of FPU recurrence. Viewing KdV solitons as independent and localized dynamic entities, they explained the FPU observations as follows. The initial condition generates a family of solitons with different speeds, moving apart in the x–t plane. Since the system studied was of finite length with perfect reflections at both ends, the solitons could not move infinitely far apart; instead, they eventually reassembled in the x–t plane, approximately recreating the initial condition after a surprisingly short “recurrence time.”

By 1967, this insight had led Gardner, Greene, Kruskal, and Miura (GGKM) to devise a nonlinear generalization of the Fourier transform method for constructing solutions of the KdV emerging from arbitrary initial conditions (Gardner et al., 1967). Called the inverse scattering method (ISM), this approach proceeds in three steps. First, the nonlinear KdV dynamics are mapped onto an associated linear scattering problem, where each eigenvalue of the linear problem corresponds to the speed of a particular KdV soliton. Second, the time evolution of the associated linear scattering data is computed. Finally, an inverse scattering calculation determines the time evolved KdV dynamics from the evolved scattering data. Thus, the solution of a nonlinear problem is found from a series of linear computations.

Toda Lattice Solitons

Another development of the 1960s was Morikazu Toda’s discovery of exact two-soliton interactions on a nonlinear spring-mass system (Toda, 1967). As in the FPU system, equal masses were assumed to be interconnected with nonlinear springs, but Toda chose the potential

$$\left(\frac{a}{b}\right)[\mathrm{e}^{-bu_j} - 1] + au_j, \tag{6}$$

where $u_j(t)$ is the longitudinal extension of the jth spring from its equilibrium value and both a and b are adjustable parameters. (In the limit $a \to \infty$ and $b \to 0$ with ab finite, this reduces to the quadratic potential of a linear spring. In the limit $a \to 0$ and $b \to \infty$ with ab finite, it describes the interaction between hard spheres.) Thus by the late 1960s, it was established that solitons were not limited to PDEs; local solutions of difference-differential equations could also exhibit the unexpected properties of unchanging shapes and speeds after collisions.

A Seminal Workshop

These events are only the salient features of a growing panorama of nonlinear wave activities that became gradually less parochial during the 1960s. Solid state physicists began to see relationships between their solitary waves (magnetic domain walls, self-shaping pulses of light, quanta of magnetic flux, polarons, etc.), and those from classical hydrodynamics and oceanography, while applied mathematicians began to suspect that the ISM (originally formulated by GGKM for the KdV equation) might be used for a broader class of nonlinear wave equations. It was amid this intellectual ferment that the first soliton research workshop was organized during the summer of 1972 (Newell, 1974). Interestingly, one of the most significant contributions to this conference came by post. From the Soviet Union arrived a paper by Vladimir Zakharov and Alexey Shabat formulating the ISM for the nonlinear PDE Zakharov & Shabat (1972)

$$\mathrm{i}\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + 2|u|^2 u = 0. \tag{7}$$

In contrast to KdV, SG, and the Toda lattice, the dependent variable in this equation is complex rather than real, so the evolutions of two quantities (magnitude and phase of u) are governed by the equation. This reflects the fact that Equation (7) is a nonlinear generalization of a linear equation $\mathrm{i}u_t + u_{xx} + u = 0$, solutions of which comprise both an envelope and a carrier wave. As this linear equation is a Schrödinger equation for the quantum mechanical probability amplitude of a particle (like an electron) moving through a region of uniform potential, it is natural to call Equation (7) the nonlinear Schrödinger (NLS) equation. When the NLS equation is used to model classical wave packets in such fields as hydrodynamics, nonlinear acoustics, and plasma waves, however, its solutions are devoid of quantum character.

Upon appreciating the Zakharov and Shabat paper, participants left the 1972 workshop aware that four nonlinear equations (KdV, SG, NLS, and the Toda lattice) display solitary wave behavior with the special properties that led Zabusky and Kruskal to coin the term *soliton* (Newell, 1974). Within two years, ISM formulations had been constructed for the SG equation and also for the Toda lattice.

Since the mid-1970s, the soliton concept has become established in several areas of applied science, and dozens of nonlinear systems are now known to be integrable through the ISM. Thus, one is no longer surprised to find stable spatially localized regions of energy, balancing the opposing effects of nonlinearity and dispersion and displaying the essential properties of objects.

ALWYN SCOTT

See also **Bäcklund transformations; Fermi–Pasta–Ulam oscillator chain; Inverse scattering method or transform; Laboratory models of nonlinear waves; Solitons**

Further Reading

Fermi, E., Pasta, J.R. & Ulam, S.M. 1955. Studies of nonlinear problems, Los Alamos Scientific Laboratory Report No. LA–1940 (Reprinted in Newell, 1974.)

Frenkel, J. & Kontorova, T. 1939. On the theory of plastic deformation and twinning. *Journal Physics (USSR)* 1: 137–149

Gardner, C.S., Greene, J.M., Kruskal, M.D. & Miura, R.M. 1967. Method for solving the Korteweg–de Vries equation. *Physical Review Letters*, 19: 1095–97

Korteweg, D.J. & de Vries, H. 1895. On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philosophical Magazine*, 39: 422–443

Lamb, G.L., Jr. 1976. Bäcklund transforms at the turn of the century. In *Bäcklund Transforms*, edited by R.M. Miura, Berlin and New York: Springer

Newell, A.C. (editor). 1974. *Nonlinear Wave Motion*, Providence, R.I: American Mathematical Society

Perring, J.K. & Skyrme, T.R.H. 1962. A model unified field equation. *Nuclear Physics*, 31: 550–555

Russell, J.S. 1844. *Report on Waves*, 14th meeting of the British Association for the Advancement of Science, London: BAAS, 311–339

Russell, J.S. 1885. *The Wave of Translation in the Oceans of Water, Air and Ether*, London: Trübner

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Seeger, A., Donth, H. & Kochendörfer, A. 1953. Theorie der Versetzungen in eindimensionalen Atomreihen. *Zeitschrift für Physik*, 134: 173–193

Steuerwald, R. 1936. Über Enneper'sche Flächen und Bäcklund'sche Transformation. *Abhandlungen der Bayerischen Akademie der Wissenschaften* München, pp. 1–105

Toda, M. 1967. Vibration of a chain with nonlinear interactions. *Journal of the Physical Society of Japan*, 22: 431–436; Wave propagation in anharmonic lattices. *Journal of the Physical Society of Japan*, 23: 501–506

Zabusky, N.J. & Kruskal, M.D. 1965. Interactions of solitons in a collisionless plasma and the recurrence of initial states. *Physical Review Letters*, 15: 240–243

Zakharov, V.E. & Shabat, A.B. 1972. Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Soviet Physics, JETP*, 34: 62–69

SOLITONS, TYPES OF

Solitons are localized solutions of nonlinear partial or difference-differential equations that preserve their integrity under collisions with other such solutions. In general, energy is conserved for soliton systems; thus, they are a special class of Hamiltonian systems. Localization may result from a dynamic balance between the effects of nonlinear and dispersion (nontopological solitons) or from a topological constraint.

Nontopological Solitons

Perhaps the most famous example among equations admitting solitary wave solution is the Korteweg–de Vries (KdV) equation

$$u_t + \alpha u u_x + u_{xxx} = 0, \tag{1}$$

which has the solitary-wave solution

$$u = a\,\mathrm{sech}^2\left\{\sqrt{a\alpha/12}[x - (a\alpha/3)t]\right\}, \tag{2}$$

where a is the wave amplitude. The quadratic nonlinear term ($\alpha u u_x$) is balanced by the dispersive term (u_{xxx}) in the solution of Equation (2). A feature of such solitary waves of this type is that they preserve their shapes and speeds under collisions (act as particles); thus they are called "solitons."

In general, Equation (1) approximates a more detailed physical description. When the quadratic nonlinear term is small, the cubic nonlinear term may be taken into account, leading to the following equation.

$$u_t + \alpha u u_x + \beta u^2 u_x + u_{xxx} = 0, \tag{3}$$

which is known as the extended Korteweg-de Vries (eKdV) equation. With $\alpha = 0$ and $\beta = 6$, Equation (3) is known as the modified Korteweg–de Vries (mKdV) equation with negative dispersion,

$$u_t + 6u^2 u_x + u_{xxx} = 0. \tag{4}$$

Equation (4) admits solitary wave solutions tending to u_0 at infinity (Grimshaw et al., 1999),

$$u(x,t) = u_0 + \frac{2\nu^2}{u_0 + \sigma\lambda\cosh[2\nu(x - \mu t)]}, \tag{5}$$

where $\lambda > |u_0|, \nu = \sqrt{\lambda^2 - u_0^2}, \sigma = \pm 1, \mu = 2u_0^2 + 4\lambda^2$, and

$$u(x,t) = u_0 - \frac{4u_0}{1 + 4u_0^2(x - 6u_0^2 t)^2}. \tag{6}$$

Equation (5) reduces to a soliton solution as $u_0 \to 0$,

$$u = 2\sigma\lambda\mathrm{sech}\{2\lambda(x - 4\lambda^2 t)\}, \tag{7}$$

which is different from the KdV soliton solution (2). Equation (6) shows that the wave shape ($u - u_0$) vanishes algebraically as $|x| \to \infty$. We call it an *algebraic soliton* or *rational soliton*.

The eKdV equation (3) has the following solitary-wave solution ((Grimshaw et al., 1999),

$$u(x,t) = \frac{(6\gamma^2/\alpha)}{1 + B\cosh\{\gamma(x - \gamma^2 t)\}}, \tag{8}$$

where $B^2 = 1 + (6\beta\gamma^2/\alpha^2)$ and γ is an arbitrary parameter, characterizing the inverse width of the solitary wave. In the case $\beta < 0$ ($0 < B < 1$), at small wave amplitudes ($B \to 1$), Equation (8) transforms into the KdV soliton solution. On the other hand, as the wave amplitude increases ($B \to 0$), it approaches the critical value $a_{cr} = \alpha/|\beta|$. In the limit ($B \to 0$), the width of the solitary wave increases to infinity and it becomes the so-called *thick soliton*. The thick soliton can be viewed as a kink-antikink combination.

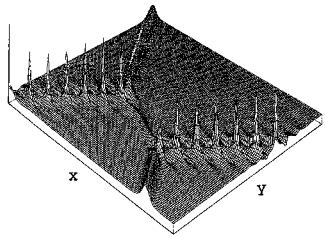

Figure 1. Interaction between line soliton and periodic soliton.

An extension of the KdV equation to motion in two space dimensions was given by Kadomtsev and Petviashvili in order to discuss the stability of the KdV soliton (line soliton in two-dimensional space) against a long transverse disturbance. This two-dimensional extension is known as the Kadomtsev–Petviashvili (KP) equation,

$$(u_t + 6uu_x + u_{xxx})_x + 3su_{yy} = 0, \tag{9}$$

which corresponds to the case of negative and positive dispersion when $s = +1$ and $s = -1$, respectively (Kadomtsev & Petviashvili, 1970). They have shown that the line soliton is stable in the case of negative dispersion and is unstable for the positive dispersion. This leads to the conjecture that a localized soliton in two-dimensional space should be formed in the positive dispersion case, since the line soliton is unstable. Such a soliton solution has been found by Manakov et al. (1977) and Ablowitz & Satsuma (1978), which is no longer exponential in character, but a rational function of space variables,

$$u = 4\frac{(L + L^*)^{-2} - \xi^2 + \eta^2}{\left[(L + L^*)^{-2} + \xi^2 + \eta^2\right]^2}, \tag{10}$$

with

$$\begin{aligned}\xi &= \mathrm{Re}\{x - 2\mathrm{i}Ly - 12L^2 t\} + \xi_0,\\ \eta &= \mathrm{Im}\{x - 2\mathrm{i}Ly - 12L^2 t\} + \eta_0,\end{aligned}$$

where ξ_0 and η_0 are arbitrary constants and $*$ indicates the complex conjugate. This solution decays like $(x^2 + y^2)^{-1}$ as $(x^2 + y^2)^{1/2} \to \infty$; thus, it is also called an *algebraic soliton* or *rational soliton* or *lump*. Another type of localized soliton appears in the positive dispersion case as a sequence of infinite algebraic solitons, called a *periodic soliton* or *soliton chain* (Tajiri & Murakami, 1989). Figure 1 shows a typical interaction between line soliton and periodic soliton.

The Benjamin–Ono (BO) equation,

$$u_t + 4uu_x + \frac{1}{\pi}P\int_{-\infty}^{\infty}\frac{u_{\xi\xi}(\xi,t)}{\xi - x}\mathrm{d}\xi = 0, \qquad (11)$$

describes a weakly nonlinear, long internal wave in a stratified fluid of great depth, where P stands for the principal value of the integration and the third term is dispersive (Ono, 1975). The BO equation has an algebraic soliton solution,

$$u = \frac{a}{a^2(x - at - x_0)^2 + 1}, \qquad (12)$$

which is called a *Benjamin–Ono soliton.*

Envelope solitons

Consider next the propagation of a modulated plane wave in a nonlinear and dispersive medium where the dispersion relation is amplitude dependent, $\omega = \omega(k, |u|^2)$. Expanding around the carrier wave number k_0 and frequency ω_0, we have

$$\omega - \omega_0 = \left(\frac{\partial\omega}{\partial k}\right)_0 (k - k_0) + \frac{1}{2}\left(\frac{\partial^2\omega}{\partial k^2}\right)_0 (k - k_0)^2 + \left(\frac{\partial\omega}{\partial|u|^2}\right)_0 |u|^2 ... \qquad (13)$$

Replacing $\omega - \omega_0$ by $\mathrm{i}\partial/\partial t$ and $k - k_0$ by $-\mathrm{i}\partial/\partial x$ and operating on u gives

$$\mathrm{i}\left\{\frac{\partial u}{\partial t} + \left(\frac{\partial\omega}{\partial k}\right)_0 \frac{\partial u}{\partial x}\right\} + \frac{1}{2}\left(\frac{\partial^2\omega}{\partial k^2}\right)_0 \frac{\partial^2 u}{\partial x^2} - \left(\frac{\partial\omega}{\partial|u|^2}\right)_0 |u|^2 u = 0, \qquad (14)$$

which is the nonlinear Schrödinger (NLS) equation. If

$$(\partial^2\omega/\partial k^2)_0(\partial\omega/\partial|u|^2)_0 < 0,$$

the plane wave is unstable for the modulation, otherwise it is stable.

Under appropriate scaling and variable transformations, Equation (14) takes the standard form

$$\mathrm{i}u_t + u_{xx} + 2\varepsilon|u|^2 u = 0, \quad (\varepsilon = \pm 1). \qquad (15)$$

With $\varepsilon = 1$, this is called the focusing NLS (FNLS) equation and has an *envelope soliton* solution,

$$u = K\mathrm{e}^{\mathrm{i}(\tilde{k}x - \tilde{\omega}t + \theta)}\mathrm{sech}(Kx - \Omega t + \sigma), \qquad (16)$$

where $-(\Omega + \mathrm{i}\tilde{\omega}) + (K + \mathrm{i}\tilde{k})^2 = 0$ and θ and σ are arbitrary constants. This is called a *bright soliton* in nonlinear optics (see Figure 2a). If the phase velocity decreases (increases) with increased amplitude, nonlinear effect results in a decrease (increase) of the wave number in the leading half of the envelope and a increase (decrease) in wave number in the trailing half. Thus, the envelope is compressed, and a soliton finds a balance between the effect of compression caused by nonlinearity and broadening caused by dispersion.

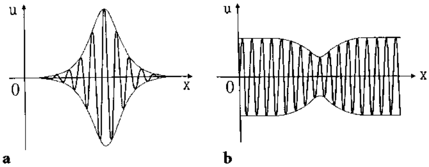

Figure 2. (a) Bright soliton and (b) dark soliton.

The defocusing NLS (DNLS) equation (with $\varepsilon = -1$) has a *dark soliton* solution (shown in Figure 2b)

$$|u|^2 = u_0^2[1 - a^2\mathrm{sech}^2\{u_0 a(x - vt)\}], \qquad (17)$$

which appears as an intensity dip in an infinitely extended constant background. The dark soliton solution with $a = 1$

$$|u|^2 = u_0^2\tanh^2 u_0(x - vt), \qquad (18)$$

is called a *black soliton* and the solution with $a < 1$ is sometimes referred to as a *gray soliton.*

Optical communication systems are degraded by the spreading out of pulses as they travel along a fiber, which is caused by dispersion. Hasagawa and Tappert proposed using the nonlinear change of dielectric constant of the fiber to compensate for this dispersive effect (Hasagawa & Tappert, 1973). The optical pulse was shown to form a bright soliton in the case of anomalous dispersion ($\partial^2\omega/\partial k^2 > 0$), a prediction that was verified by Mollenauer et al. (1980).

Writing $k = k(\omega, |u|^2)$ as the solution of the dispersion relation again leads to the NLS equation but with different variables (ω and k interchanged). The NLS equation in these variables is important in nonlinear optics.

More Space Dimensions

Spatial dark-soliton stripes are experimentally found in the transverse cross section of a continuous-wave (cw) optical beam propagating through material with a self-defocusing nonlinearity. The spatial evolution of a monochromatic transverse electric field $E(x, y, z)$ in a self-defocusing medium with Kerr nonlinearity (the intensity-dependent refractive index $n = n_0 - n_2|E|^2$) is described by the NLS equation

$$\mathrm{i}u_z + \tfrac{1}{2}\Delta_\perp u - |u|^2 u = 0, \qquad (19)$$

where $\Delta_\perp = \partial_x^2 + \partial_y^2$ is the transverse diffraction operator and z is the propagation coordinate. Although dark-soliton stripes are unstable to transverse long-wavelength modulation, linear analysis predicts stability to transverse modulation having a short period.

Kivshar & Yang (1994) showed that self-defocusing nonlinear media can support a *ring soliton*, which is a dark-solitary wave with ring symmetry. If the ring radius is small enough, it is not subject to the transverse instability characteristic of dark-stripe solitons. It has also been demonstrated that in the small-amplitude limit such solitons are governed by a cylindrical KdV equation. Frantzeskakis and Malomed also showed that *anti-dark ring solitons* (humps on top of the cw background rather than dips) can exist too but only for non-Kerr saturable nonlinearities (Frantzeskakis & Malomed, 1999).

Spatial solitons are optical beams that are self-trapped in space due to a balance between the Kerr nonlinearity and diffraction, a well-studied phenomenon. A new type of spatial soliton, which was proposed by Segev et al. (1992), occurs in a photorefractive (PR) crystal biased with an external dc electric field. The presence of optical beams in a crystal leads to photoexcitation of electric charges. The space charge screens the externally applied electric field in the illuminated area of the crystal. The non-uniform screening of the electric field modifies the refractive index in such a way that the beam becomes self-trapped and propagates in a form of a *PR spatial soliton*. PR spatial solitons can occur at microwatt power levels, whereas the observation of Kerr solitons requires much higher power. Steady solitons in PR crystals are called *screening solitons*.

With the application of a slightly detuned driving field, a slightly lossy Kerr medium supports *cavity solitons*, which are then induced by writing pulses (Firth et al., 2002). This provides an optical means for writing patterns of light onto the cavity cross section and subsequently erasing them, in other words an optical memory.

It is known that the diffusion effect of photoexcited charges leads to a bending of the trajectories of PR solitons. Królikowski et al. (1996) showed by using numerical simulations that *self-bending solitons* in the presence of the diffusion processes are stable and withstand relatively large perturbations. They also showed that even a small contribution of diffusion effect leads to strong energy exchange between colliding PR solitons. In general, spatial solitons in PR media do not satisfy the mathematical definition of solitons, even when they propagate as solitary waves with an unchanging beam profile.

Self-trapping has been studied in Kerr-type, PR, quadratic, and resonant atomic nonlinear media. All of these studies have investigated self-trapping of spatially coherent light beams only. Diffraction of a spatially incoherent beam is larger than that of a coherent beam of the same width. Therefore, a spatially incoherent beam diverges much faster than a coherent beam, and self-trapping of an incoherent beam requires stronger optical nonlinearities than self-trapping a coherent beam. Recently, self-trapping of incoherent light beams has been demonstrated experimentally (Mitchell et al., 1996) and theoretically (Mitchell et al., 1997). In general, spatial *incoherent solitons* are multimode, self-trapped entities, found only in materials with non-instantaneous (internal) nonlinearity (Mitchell et al., 1997).

Another interesting phenomena is the formation of solitons by the mutual trapping of the interacting waves. Torner et al. (1996) showed that soliton-like propagation occurs in the presence of walk-off between the interacting waves. The solitons in the presence of walk-off are called *walking solitons*. Temporal walk-off is due to different group velocities of the waves forming the soliton, while spatial walk-off is due to different propagation directions of energy and phase fronts in anisotropic media.

The possibility of creating three-dimensional (3-d) localized pulses in a self-focusing medium with anomalous group-dispersion was demonstrated by Silberberg (1990), who showed that they pulses are robust in the sense that remain as separate solitary formations even after collisions. These 3-d-pulses propagating without changes in space or time are called *light bullets*.

Topological Solitons

Another mechanism leading to the emergence of solitons is a topological constraint, exemplified by the sine-Gordon (SG) equation

$$u_{tt} - u_{xx} + m^2 \sin u = 0. \tag{20}$$

The SG equation is characterized by two properties: Lorentz invariance and multiple ground states of the energy. The presence of multiple ground states aids soliton formation because the field takes different asymptotic values on both sides of the isolated wave, so the wave is unable to disperse. (It cannot decay by spreading out for the same reason a twist sealed in the boundary condition at the end of a band cannot be removed.) Thus, wave fields can incorporate *topological solitons* if there are multiple ground states with an appropriate topology. Such a soliton of the SG equation is given by

$$u = 4 \arctan \left[\exp \left(\pm m \frac{x - vt - x_0}{\sqrt{1 - v^2}} \right) \right]. \tag{21}$$

The solution represents a twist in the configuration of the field connecting one ground state $2n\pi$ to the another ground state $(2n \pm 2)\pi$. The + sign and − sign solutions are called a *kink* and an *antikink*, respectively, as shown in Figure 3. They are often termed *soliton* and *antisoliton* instead of kink and antikink. The energy of the SG kink solution is given by $8m/\sqrt{1 - v^2}$,

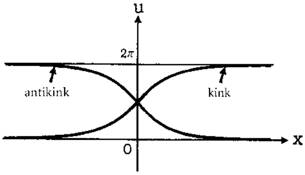

Figure 3. Kink and antikink.

which has relativistic dependence on the velocity v. There is an analogy between kink/antikink and positively/negatively charged elementary particles. The time evolution of an initial state consisting of two kinks shows that the two kinks move away from each other. On the other hand, the kink and antikink move toward each other. The twist (Q) of a kink ($+2\pi$) or antikink (-2π) is known as its *topological charge*.

Perring and Skyrme were interested in the SG equation as a model equation for physical elementary particles (Perring & Skyrme, 1962). They regarded the kink as nucleon and the linear wave (with $Q=0$) as a meson. The soliton solution of a nuclear model for which topological charge is the baryon number is called a *skyrmion*. Topological charge is a conserved quantity stemming from the geometric configuration of the field—an important feature of many field theories that admit kink solutions.

Numerous attempts were made to find particle-like solutions of relativistically invariant nonlinear field equations in higher dimensions. Under radial symmetry, ring and spherical solitons to the SG and Higgs field equations have been studied numerically by Bogolyubskii and Makhan'kov who called such waves *pulsons* due to their pulsating behavior (Bogolyubskii & Makhan'kov, 1976).

Lattice-solitons

Solitons and kinks in discrete systems are sometimes called *lattice-solitons* and *lattice-kinks*, respectively. The Toda lattice is a one-dimensional lattice of equal masses, interconnected by exponetial interaction potentials of the form (Toda, 1981)

$$\phi(r) = \frac{a}{b}\exp(-br) + ar + \text{const.}, \tag{22}$$

where $a, b > 0$ and r is the change of a distance between adjacent masses from its equilibrium value. The equation of motion for Toda lattice is an integrable system, admitting exact N-soliton solutions, whose form for $N=1$ is given by

$$\exp(-br_n) - 1 = \sinh^2(\kappa)\operatorname{sech}^2\left(\kappa n \pm t\sqrt{ab/m}\,\sinh\kappa\right), \tag{23}$$

where κ is an amplitude parameter and m is the mass of the particles.

The theory of wave propagation in periodic structures shows the existence of forbidden frequency bands or band gaps, located around the (Bragg) reflection frequencies. A wave with overtone frequencies within the forbidden frequency bands will have little means of interaction with the periodic structure. The solitons that emerge from a balance between nonlinear self-phase modulation and dispersion associated with the periodic structure are called *gap solitons*.

Other Types of Solitons

A variable coefficient KdV equation arises in the study of a solitary water waves as it enters a region where the bottom is no longer level (Johnson, 1980). A soliton on non-uniform background generally moves with variable speed. When a soliton comes from left (or right) and goes back to left (or right) with opposite velocity—sailing back like an Australian boomerang—it is called a *boomeron*. When a soliton oscillates without ever escaping to infinity, the soliton is referred to as a *trappon* (Calogero & Degasperis, 1982).

The *Dym* equation ($r_t = r^3 r_{xxx}$) was discovered in unpublished work by Harry Dym and rediscovered in a more general form by Sabatier within the classical string problem. This equation belongs to a wide class of nonlinear equations (WKI equations) found by Wadati, Konno, and Ichikawa to be completely integrable (Wadati et al., 1979). Reciprocal links between the Dym and KdV and MKdV equations provide implicit solutions, since they include the simultaneous change of both dependent and independent variables (Kawamoto, 1985). The Dym equation has *cusp soliton* solutions. Some new approaches, which allow construction of multisoliton solutions almost explicitly have been developed for the Dym equation by (Dmitrieva, 1993). Recently, it was shown that the Dym equation on the complex plane is relevant to such physical problems as the Hele-Shaw problem and the Saffman–Taylor problem (Constantin & Kadanoff, 1991).

The nonlinear transverse oscillation of an elastic beam subject to an end-thrust is described by one of the WKI equations. If the beam is flexible enough, it deforms into a loop, with the upper half portion having negative curvature. In this case, the nonlinear oscillation can be described by an equation of the following form (Konno et al., 1981):

$$y_{xt} + \operatorname{sgn}\left(\frac{\mathrm{d}s}{\mathrm{d}x}\right)\left(\frac{y_{xx}}{(1+y_x^2)^{3/2}}\right)_{xx} = 0, \tag{24}$$

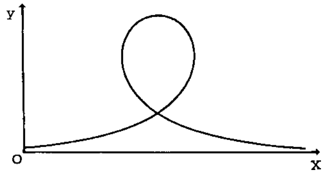

Figure 4. Loop soliton.

where s denotes arc length measured around the loop. Konno et al. showed that such a *loop soliton* (shown in Figure 4) propagating along a stretched rope can be obtained as a one soliton solution to Equation (24).

Solitary waves with finitely extended (compact) support were recently studied in various equations with nonlinear dispersion. Rosenau and Hyman showed that solitary-wave solutions may have compact support under the influence of nonlinear dispersion in various generalizations of the KdV equations with nonlinear dispersion (Rosenau & Hyman, 1993). The nonlinear dispersion is weaker for small amplitude than the linear dispersion in the KdV equation, leading to compactification. Such robust soliton-like pulses—characterized by the absence of the infinite tail—are called *compactons*.

MASAYOSHI TAJIRI

See also **Korteweg–de Vries equation; Multidimensional solitons; Nonlinear optics; Nonlinear Schrödinger equations; Sine-Gordon equation; Skyrmions; Solitons; Toda lattice**

Further Reading

Ablowitz, M.J. & Satsuma, J. 1978. Solitons and rational solutions of nonlinear evolution equations. *Journal of Mathematical Physics*, 19: 2180–2186

Bogolyubskii, I.L. 1976. Oscillating particle-like solutions of the nonlinear Klein–Gordon equation. *Soviet Physics JETP Letters*, 24: 535–538

Bogolyubskii, I.L. & Makhan'kov, V.G. 1976. Lifetime of pulsating solitons in certain classical models. *Soviet Physics JETP Letters*, 24: 12–14

Calogero, F. & Degasperis, A. 1982. *Spectral Taransform and Solitons I*, North-Holland, Amsterdam

Constantin, P. & Kadanoff, L. 1991. Dynamics of a complex interface. *Physica* D, 47: 450–460

Dmitrieva, L.A. 1993. Finite-gap solutions of the Harry Dym equation. *Physics Letters* A, 182: 65–70

Firth, W.J., Harkness, G.K., Lord, A., McSloy, J.M., Gomila, D. & Colet, P. 2002. Dynamical properties of two-dimensional Kerr cavity solitons. *Journal of Optical Society of America* B, 19(4): 747–752

Frantzeskakis, D.J. & Malomed, B.A. 1999. Multiscale expansions for a generalized cylindrical nonlinear Schrödinger equation. *Physics Letters* A, 264: 179–185

Grimshaw, R., Pelinovsky, E. & Talipova, T. 1999. Solitary wave transformation in a medium with sign-variable quadratic nonlinearity and cubic nonlinearity. *Physica* D, 132: 40–62

Hasagawa, A. & Tappert, F. 1973. Taransmission of stationary nonlinear optical pulses in dispersive dielectric fibers. I. Anomalous dispersion, II. Normal dispersion. *Applied Physics Letters*, 23: 142–144, 171–172

Johnson, R.S. 1980. Water waves and Korteweg–de Vries equations. *Journal of Fluid Mechanics*, 97: 701–719

Kadomtsev, B.B. & Petviashvili, V.I. 1970. On the stability of solitary waves in weakly dispersing media. *Soviet Physics-Doklady*, 15: 539–541

Kawamoto, S. 1985. An exact transformation from the Harry Dym equation to the modified K-dV equation. *Journal of the Physical Society of Japan*, 54: 2055–2056

Kivshar, Y.S. & Yang, X. 1994. Ring dark solitons. *Physical Review* E, 50: R40–R43

Konno, K., Ichikawa, Y.H. & Wadati, M. 1981. A loop soliton propagating along a stretched rope. *Journal of the Physical Society of Japan*, 50: 1025–1026

Królikowski, W., Akhmediev, N., Luther-Davies, B. & Gronin-Golomb, M. 1996. Self-bending photorefractive solitons. *Physical Review* E, 54: 5761–5765

Manakov, S.V., Zakharov, V.E., Bordag, L.A., Its, A.R. & Matveev, V.B. 1977. Two-dimensional solitons of the Kadomtsev–Petviashvili equation and their interaction. *Physics Letters*, 63A: 205–206

Mitchell, M., Chen, Z., Shih, M. & Segev, M. 1996. Self-trapping of partially spatially incoherent light. *Physical Review Letters*, 77: 490–493

Mitchell, M., Segev, M., Coskun, T.H. & Christodoulides, D.N. 1997. Theory of self-trapping spatially incoherent light beams. *Physical Review Letters*, 79: 4990–4993

Mollenauer, L.F., Stolen, R.H. & Gordon, J.P. 1980. Experimental observation of picosecond pulse narrowing and solitons in optical fibers. *Physical Review Letters*, 45: 1095–1098

Ono, H. 1975. Algebraic solitary waves in stratified fluids. *Journal of the Physical Society of Japan*, 39: 1082–1091

Perring, J.K. & Skyrme, T.H.R. 1962. A model unified field equation. *Nuclear Physics*, 31: 550–555

Rosenau, P. & Hyman, J. M. 1993. Compactons: solutions with finite wavelength. *Physical Review Letters*, 70: 564–567

Segev, M., Crosignani, B., Yariv, A. & Fischer, B. 1992. Spatial solitons in photorefractive media. *Physical Review Letters*, 68: 923–936

Silberberg, Y. 1990. Collapse of optical pulses. *Optics Letters*, 15: 1282–1284

Tajiri, M. & Murakami, Y. 1989. The periodic soliton resonance: solutions to the Kadomtsev–Petviashvili equation with positive dispersion. *Physics Letters* A, 143: 217–220

Toda, M. 1981. *Theory of Nonlinear Lattices*, Springer-Verlag, Berlin

Torner, L., Mazilu, D. & Mihalache, D. 1996. Walking solitons in quadratic nonlinear media. *Physical Review Letters*, 77: 2455–2458

Wadati, M., Konno, K. & Ichikawa, Y. H. 1979. New integrable nonlinear evolution equations. *Journal of the Physical Society of Japan*, 47: 1698–1700

SOLUTION TRAJECTORY

See **Phase space**

SPATIOTEMPORAL CHAOS

Spatiotemporal chaos is a dynamical regime developing in spatially distributed systems lacking long-time, large-distance coherence in spite of an organized regular behavior at the local scale. It is, so to speak, located in the middle of a triangle, the corners of which are temporal chaos, which is prevalent for a few spatially frozen degrees of freedom, spatial chaos, in disordered time-independent patterns, and turbulence, with cascading processes over a wide range of space and time scales.

Short-term local coherence is usually the result of some instability mechanism that generates dissipative structures of different kinds depending on whether or not a specific frequency ($\omega_c = 0$ or $\neq 0$) and/or a spatial periodicity ($k_c = 0$ or $\neq 0$) is introduced in the system (Cross & Hohenberg, 1993). Examples of spatiotemporal chaos may be found in every combination of these elementary cases: Rayleigh–Bénard convection produces a time-independent cellular pattern ($\omega_c = 0, k_c \neq 0$), the Belousov–Zhabotinsky (BZ) reaction-diffusion system is unstable against a homogeneous oscillatory mode ($\omega_c \neq 0, k_c = 0$), convection in binary fluid mixtures develops in the form of dissipative waves ($\omega_c \neq 0, k_c \neq 0$), and the same holds for hydrothermal waves. Transitions to spatiotemporal chaos have been observed and studied in many experimental systems, including parametrically excited surface waves, electro-hydrodynamic instabilities in nematic liquid crystals (Kramer & Pesch, 1996), and liquid films flowing down inclines (Chang, 1994). Figure 1 (left) illustrates the case of spiral defect chaos in convection.

In practice, confinement effects enter and compete with instability mechanisms in the ordering process. Their intensity can be appreciated through aspect ratios, measuring the physical size of the system in units of the instability wavelength. In small aspect-ratio experiments, confinement effects are effective in the three space directions and chaos is purely temporal. Spatiotemporal chaos develops when confinement is partially relaxed. Accordingly, phenomena developing at surfaces or in thin layers can be understood as (quasi-)two-dimensional (Gollub, 1994). In the same way, narrow channels or oriented media along a specific direction exemplify the case of (quasi-)one-dimensional systems (Daviaud, 1994).

A second important classification results from the nature of the bifurcation, either supercritical or subcritical (continuous or discontinuous), which implies either substitution or coexistence of bifurcating and bifurcated states. To a large extent, this feature dictates the type of theory most appropriate to understand the growth of spatiotemporal chaos.

Consider first the supercritical case. When confinement effects are unable to maintain order everywhere, reduced universal descriptions are generically obtained as "envelope equations" (Newell, 1974). Standing as a paradigm, the cubic complex Ginzburg–Landau (CGL3), reviewed by Aranson & Kramer (2002), spatially unfolds a local supercritical Hopf bifurcation signalling the emergence of uniform oscillations ($\omega_c \neq 0$, $k_c = 0$). In one dimension (1-d), this equation reads

$$\partial_t A = A + (1 + \mathrm{i}\alpha)\partial_{xx} A - (1 + \mathrm{i}\beta)|A|^2 A, \qquad (1)$$

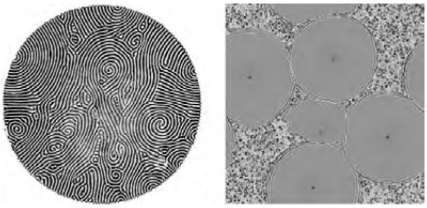

Figure 1. Left: Spiral defect chaos near threshold in Rayleigh–Bénard convection at $\mathcal{O}(1)$ Prandtl number. (Courtesy Ahlers, 1998). Right: Coexistence of spirals and defect mediated chaos in the 2-d CGL equation for $\alpha = -2$ and $\beta \simeq 0.75$ (Manneville, unpublished).

where the parameters α and β measure linear and nonlinear dispersion effects, respectively. The CGL3 equation admits trivial exact solutions in the form of plane waves $A = A_q \exp[\mathrm{i}(qx - \omega_q t)]$, with amplitude $A_q = (1 - q^2)^{1/2}$ and angular frequency $\omega_q = \alpha q^2 + \beta(1 - q^2)$, which are stable or unstable depending on the values of α, β, and q. Other nonlinear solutions can exist; in one dimension, solitary waves called Bekki–Nozaki holes are the best known. In two dimensions (2-d) they take the form of spiral waves that are topological defects of the complex order parameter A. Figure 1 (right) illustrates the coexistence of spirals and defect-mediated chaos (see below) in the 2-d CGL equation. Figure 2 displays the different possible steady-state regimes of the 1-d CGL equation in the $(1/\beta, -\alpha)$ plane. In region I, plane waves attract most initial conditions. Phase turbulence is present in region IV slightly beyond the Benjamin–Feir instability (BF) line, as given by Newell's criterion for "$q = 0$" oscillations (Newell, 1974). In the vicinity of this line, the solution can be written as $A(x, t) = (1 + \varrho(x, t)) \exp(\mathrm{i}\theta(x, t))$. The amplitude modulation ϱ, enslaved to the gradient of the phase perturbation θ, remains small, while θ is governed at lowest order by the Kuramoto–Sivashinsky equation (Kuramoto, 1978)

$$\partial_t \theta = D\partial_{xx}\theta - K\partial_{xxxx}\theta + g(\partial_x\theta)^2, \qquad (2)$$

where $D = 1 + \alpha\beta$ is an effective diffusion coefficient that is negative in the unstable range (Chaté & Manneville, 1994a). Deeper in the unstable domain, in region V, a "revolt" of $|A|$ ends in the formation of defects (phase singularities at zeroes of $|A|$) and amplitude turbulence or defect-mediated turbulence sets in (Coullet et al., 1989). Defects analogous to Bekki–Nozaki holes are observed to evolve in a spatiotemporal intermittent fashion in region II. As suggested by its position in the diagram, the "bi-chaos" regime in region III presents itself as a fluctuating mixture of states in regions IV and V.

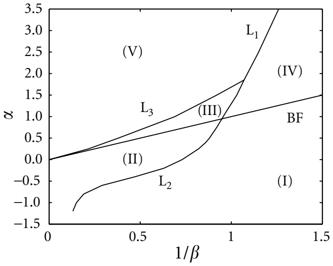

Figure 2. Bifurcation diagram of the 1-d CGL equation. (Data from Chaté, 1994.)

By contrast, subcritical instabilities are characterized by the possibility of finding the system in one of several (usually two) states at a given point in space. Short-range coherence then implies the formation of homogeneous domains of each state, separated by fronts (Pomeau, 1986). In gradient systems, these fronts move regularly so as to decrease the potential, but in nongradient systems, they may have more complicated behaviors. A particularly interesting scenario develops when one of the competing local states is a chaotic transient while the other is regular. At a given time, the whole system can be divided into so-called laminar and turbulent domains, and at a given point in space, the system is alternatively laminar or turbulent, hence the name "spatiotemporal intermittency" (STI). According to Pomeau (1986), front propagation is then akin to a time-oriented stochastic process known as "directed percolation" (DP) used to model epidemic processes (Kinzel, 1983). Directed percolation defines a critical phenomenon, with an associated universality class (a specific set of scaling exponents governing the statistical behavior of the system as a function of the distance to threshold).

The existence of separated domains and sharp fronts supports the idea of modeling extended systems in terms of identical subsystems arranged on a regular lattice, each with its own phase space, and coupled to its neighbors. In addition to space discretization, time discretization leads to the definition of coupled map lattices that have served to illustrate several transition scenarios such as cascades of spatial period doublings, defect-mediated regimes (Kaneko, 1993), or STI (Chaté & Manneville 1994b) that make explicit how local transient temporal chaos is converted into sustained spatiotemporal chaos. A last step can be taken by also discretizing the local phase space, which yields cellular automata (Wolfram, 1986). Further randomization of the dynamics then points towards an understanding of the transition to turbulence in statistical physics terms (Kinzel, 1983). Such approaches should help in understanding plane Couette flow, the simplest shear flow produced by two parallel plates sliding in opposite directions. This configuration is a particularly intriguing example of a subcritical hydrodynamic system where domains of laminar flow, known to be stable for all flow conditions, coexist in a wide range of Reynolds numbers and in continuously varying proportions with domains of small-scale turbulence.

Spatio-temporal chaos definitely relates to the process of transition to turbulence when confinement effects are weak. It appears to occupy a central position at the crossroads of nonlinear dynamics, mathematical stability theory, and statistical physics of many-body systems and non-equilibrium processes, with a wide potential for applications (Rabinovich et al., 2000).

Paul Manneville

See also **Cellular automata; Chaos vs. turbulence; Coherence phenomena; Coupled map lattice; Development of singularities; Gradient system; Hydrothermal waves; Kuramoto–Sivashinsky equation; Modulated waves; Multiple scale analysis; Navier–Stokes equation; Nonequilibrium statistical mechanics; Order parameters; Pattern formation; Phase dynamics; Reaction-diffusion systems; Spiral waves; Surface waves; Topological defects; Turbulence**

Further Reading

Ahlers, G. 1998. Experiments on spatio-temporal chaos. *Physica* A, 249: 18–26

Aranson, I.S. and Kramer, L. 2002. The world of the complex Ginzburg–Landau equation. *Reviews of Modern Physics*, 74: 99–143

Chang, H.-C. 1994. Wave evolution on a falling film. *Annual Reviews of Fluid Mechanics*, 26: 103–136

Chaté, H. 1994. Spatiotemporal intermittency regimes of the one-dimensional complex Ginzburg–Landau equation. *Nonlinearity* 7: 185–204

Chaté, H. & Manneville, P. 1994a. Phase turbulence. In *Turbulence, A Tentative Dictionary*, edited by P. Jabeling & O. Cardoso, NewYork: Plenum press

Chaté, H. & Manneville, P. 1994b. Spatiotemporal intermittency. In *Turbulence A Tentative Dictionary*, edited by P. Jabeling & O. Cardoso, NewYork: Plenum press

Cross, M.C. & Hohenberg, P.C. 1993. Pattern formation outside equilibrium. *Reviews of Modern Physics*, 65: 851–1112

Coullet, P., Gil, L. & Lega, J. 1989. Defect-mediated turbulence. *Physical Review Letters*, 62: 1619–1622

Daviaud, F. 1994. Experiments in 1D turbulence. In *Turbulence A Tentative Dictionary*, edited by P. Jabeling & O. Cardoso, NewYork: Plenum press

Gollub, J.P. 1994. Experiments on spatiotemporal chaos (in two dimensions). In *Turbulence A Tentative Dictionary*, edited by P. Jabeling & O. Cardoso, NewYork: Plenum press

Kaneko, K. (editor). 1993. *Theory and Application of Coupled Map Lattices*. Chichester and New York: Wiley

Kinzel, W. 1983. Directed percolation. In *Percolation Structures and Processes*, edited by G. Deutscher, R. Zallen & J. Adler, Annals of the Israel Physics Society, 5: 425–445

Kramer, L. & Pesch, W. 1996. Electrohydrodynamic instabilities in nematic liquid crystals. In *Pattern Formation in Liquid*

Crystals, edited by A. Buka & K. Kramer, New York: Springer

Kuramoto, Y. 1978. Diffusion-induced chaos in reaction systems. *Suppl. Progress in Theoretical Physics*, 64: 346–367

Newell, A.C. 1974. Envelope equations *Lectures in Applied Mathematics*, 15: 157–163

Pomeau, Y. 1986. Front motion, metastability and subcritical bifurcations in hydrodynamics. *Physica* D, 23: 3–11

Rabinovich, M.I., Ezersky, A.B. & Weidman, P.D. 2000. *The Dynamics of Patterns*. Singapore: World Scientific

Tabeling, P. & Cardoso, O. (editors). 1994. *Turbulence: A Tentative Dictionary*. New York: Plenum Press

Wolfram, S. (editor). 1986. *Theory and Applications of Cellular Automata*. Singapore: World Scientific

SPECTRAL ANALYSIS

Spectral analysis is a central method in studies of linear systems (which obey the superposition principle), and spectral representations are widely used in acoustics, quantum mechanics, wave propagation, optical spectroscopy, harmonic analysis, and signal processing. If $f(x)$ is a square integrable function on real line, it can be represented in a dual (spectral) space k with the integral transforms (Hildebrand, 1976)

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(k)\mathrm{e}^{\mathrm{i}kx}\mathrm{d}k \tag{1}$$

and

$$\hat{f}(k) = \int_{-\infty}^{\infty} f(x)\mathrm{e}^{-\mathrm{i}kx}\mathrm{d}x. \tag{2}$$

The Fourier transform $\hat{f}(k)$ gives the spectral density of harmonic oscillations $\mathrm{e}^{\mathrm{i}kx}$ between the wave numbers k and $k + dk$. The inverse Fourier transform $f(x)$ corresponds to a spectral decomposition of a function $f(x)$ over a continuous linear combination of the harmonic oscillations $\mathrm{e}^{\mathrm{i}kx}$.

For periodic functions with a period L, such that $f(x+L) = f(x)$, the spectral density has peaks at wave numbers $k = k_n = 2\pi n/L$; that is, the spectral decomposition of a function $f(x)$ becomes a discrete sum of the harmonic oscillations $\mathrm{e}^{\mathrm{i}k_n x}$:

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}_n \mathrm{e}^{\mathrm{i}k_n x} \tag{3}$$

and

$$\hat{f}_n = \frac{1}{2L}\int_{-L}^{L} f(x)\mathrm{e}^{-\mathrm{i}k_n x}\mathrm{d}x, \tag{4}$$

where $\hat{f}_n$ are Fourier coefficients of the complex Fourier series for $f(x)$. Thus, depending on properties of a function $f(x)$, it can be decomposed over continuous or discrete spectrum in the Fourier spectral representation.

As an example, the rectangular wave defined as

$$\Delta_\varepsilon(x) = \begin{cases} 1/2\varepsilon, & |x| < \varepsilon, \\ 0, & |x| > \varepsilon \end{cases} \tag{5}$$

can be expressed in the continuous spectral representation with the Fourier transform:

$$\hat{\Delta}_\varepsilon(k) = \frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon} \mathrm{e}^{-\mathrm{i}kx}\mathrm{d}x = \frac{\sin \varepsilon k}{\varepsilon k}. \tag{6}$$

The power spectrum of a signal is defined as the squared amplitude of its Fourier spectrum: $\hat{R}(k) = |\hat{f}(k)|^2$. Equivalently, the power spectrum is the Fourier spectrum of the autocorrelation function $R(x)$ defined as

$$\begin{aligned} R(x) &= \int_{-\infty}^{\infty} f(x')f(x'-x)\,\mathrm{d}x' \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |\hat{f}(k)|^2\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{R}(k)\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k. \end{aligned} \tag{7}$$

The autocorrelation function $R(x)$ represents similarities between the function $f(x)$ and itself, and its Fourier transform can be really measured in applications.

A power spectrum measures the energy of a signal between the wave numbers k and $k + \mathrm{d}k$. The total energy of a signal is related to the power spectrum by the Parseval formula:

$$E = \int_{-\infty}^{\infty} f^2(x)\,\mathrm{d}x = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\hat{f}(k)|^2\,\mathrm{d}k = R(0). \tag{8}$$

As a result, the autocorrelation function is bounded as $|R(x)| \le R(0) = E$.

Linear differential equations, especially initial-value problems for wave equations, can be easily solved with the use of spectral decompositions such as Fourier transforms. In the linear limit, for example, small-amplitude long water waves are described by the linearized KdV equation:

$$\frac{\partial u}{\partial t} + \frac{\partial^3 u}{\partial x^3} = 0, \tag{9}$$

with initial data: $u(x,0) = f(x)$. The Fourier transform method represents the solution to this problem as (Ablowitz & Fokas, 1997)

$$u(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{u}(k,t)\mathrm{e}^{\mathrm{i}kx}\mathrm{d}k, \tag{10}$$

where $\hat{u}(k,0) = \hat{f}(k)$. Since the Fourier transform $\hat{f}(k)$ has the property $\hat{f}'(k) = \mathrm{i}k\hat{f}(k)$, the time evolution of the spectral density $\hat{u}(k,t)$ is trivial,

$$\frac{\partial \hat{u}}{\partial t} - \mathrm{i}k^3\hat{u}(k,t) = 0, \tag{11}$$

with the solution $\hat{u}(k,t) = \hat{f}(k)\mathrm{e}^{\mathrm{i}k^3 t}$. As a result, the exact solution $u(x,t)$ of the initial-value problem for the linearized KdV equation is a spectral superposition with density $\hat{f}(k)$ of waves $\mathrm{e}^{\mathrm{i}(kx-\omega(k)t)}$, where

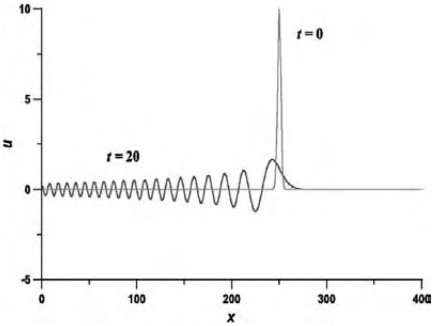

Figure 1. Exact solution $u(x, t)$ of the linear KdV equation in the Fourier spectral representation for $t > 0$, compared with the initial (Gaussian) form $u(x, 0) = \mathrm{e}^{-x^2}$.

$\omega(k) = -k^3$ is the dispersion relation. Figure 1 shows the exact solution $u(x, t)$ at a time $t > 0$, which evolves from the initial (Gaussian) form $u(x, 0) = f(x) = \mathrm{e}^{-x^2}$.

The inverse scattering transform (IST) extends spectral analysis for solutions of initial-value problems for nonlinear differential equations. For example, a solution of the nonlinear KdV equation

$$\frac{\partial u}{\partial t} + \frac{\partial^3 u}{\partial x^3} - u\frac{\partial u}{\partial x} = 0, \tag{12}$$

such that $u(x, 0) = f(x)$ is related to the spectrum of the stationary Schrödinger equation on a real line of x with a t-dependent potential $u(x; t)$ (Ablowitz et al., 1974)

$$-\psi''_\lambda(x) + u(x; t)\psi_\lambda(x) = \lambda\psi_\lambda(x). \tag{13}$$

If $f(x)$ is a square integrable function on real line, the spectral data consists of the continuous spectrum at $\lambda = k^2 \geq 0$ with eigenfunctions $\psi_\lambda = \Psi(x; k)$ and a finite discrete spectrum at $\lambda = -p_n^2 < 0$ with eigenfunctions $\psi_\lambda = \Psi_n(x)$. The spectral decomposition of the potential $u(x; t)$ over eigenfunctions of the continuous and discrete spectrum takes the form (Pelinovsky & Sulem, 2000)

$$\begin{aligned} u(x; t) = {} & \frac{1}{2\pi}\int_{-\infty}^{\infty} \rho(k, t)\Psi(x; k)\,\mathrm{d}k \\ & + \sum_n \rho_n(t)\Psi_n(x), \end{aligned} \tag{14}$$

where $\rho(k, t)$ and $\rho_n(t)$ are coefficients of the spectral decomposition. The time-evolution of the spectral data has the simple solution

$$\begin{aligned} \rho(k, t) &= \rho(k, 0)\mathrm{e}^{\mathrm{i}k^3 t}, \\ \rho_n(t) &= \rho_n(0)\mathrm{e}^{p_n^3 t}. \end{aligned} \tag{15}$$

The initial values for the spectral data are found from spectral analysis of the stationary Schrödinger equation with the initial potential $u(x; 0) = f(x)$.

Dmitry Pelinovsky

See also **Generalized functions; Integral transforms; Inverse scattering method or transform; Quantum theory**

Further Reading

Ablowitz, M.J. & Fokas, A.S. 1997. *Complex Variables: Introduction and Applications*, Cambridge and New York: Cambridge University Press

Ablowitz, M.J., Kaup, D.J., Newell, A.C. & Segur, H. 1974. The inverse scattering transform-Fourier analysis for nonlinear problems. *Studies in Applied Mathematics*, 53: 249–315

Hildebrand, F.B. 1976. *Advanced Calculus for Applications*, 2nd edition, Englewood Cliffs, NJ: Prentice-Hall

Pelinovsky, D.E. & Sulem, C. 2000. Eigenfunctions and eigenvalues for a scalar Riemann–Hilbert problem associated to inverse scattering. *Communications in Mathematical Physics*, 208: 713–760

SPIN SYSTEMS

Spin systems generally refer to ordered magnetic systems. A feel for how spins behave can be obtained by observing the behavior of two or more small compasses or magnetic needles pinned close to each other but free to rotate in a plane (perpendicular to the north–south direction), so that they align in parallel corresponding to a minimum energy configuration.

Even though magnetism is one of the oldest natural phenomena known, it took many centuries to identify its origin. In 1820, the Danish physicist Hans Oersted observed that electric current flowing through a wire affects a nearby magnet, and it is now known that the motion of an electron in an orbit around the nucleus of an atom is equivalent to a small electric current loop that behaves as an atomic magnet called a magnetic dipole moment. Each electron also possesses a rotation about its own axis, which is identified as electron spin, and is equivalent to a circulating electric current with its own dipole magnetic moment.

More generally, spin angular momentum or spin is an intrinsic property, associated with quantum particles, which does not have a classical counterpart. The magnetic dipole moment due to the spinning motion with spin angular momentum operator $\boldsymbol{S}$ is given by $\boldsymbol{m}_\mathrm{S} = -(g\mu_\mathrm{B}/\hbar)\boldsymbol{S}$, where g is the gyromagnetic ratio, μ_B is the Bohr magneton and $\hbar = h/2\pi$ is Planck's constant.

Macroscopically, all substances are magnetic to some extent and every material when placed in a magnetic field acquires a magnetic moment. Magnetic materials can be classified in terms of magnetic moment in the absence of an applied field. The magnetic moment is zero for each atom in diamagnetic materials. In the case of paramagnetic materials, for each atom the moment is nonzero but still averages to zero over many atoms. In ferromagnetic materials, the moment of each atom and even the average is not zero (Mattis, 1988).

Model	Directional Coefficient
Ising/Potts	$a = b = 0$
XY	$c = 0$
Easy axis exchange anisotropy	$c > a = b$
Easy plane exchange anisotropy	$c < a = b$
Heisenberg	$a = b = c = 1$
XYZ	$a \neq b \neq c$

Table 1. Ferromagnetic models.

Ferromagnetic materials, which are made up of domains, exhibit long-range ordering that causes the spins of the atomic ions to line up parallel to each other in a domain. The underlying interaction originates from an exchange interaction that is caused by the overlapping of electronic wave functions that has no classical analog. The exchange interaction is a quantum phenomenon of electrostatic origin first proposed by Heisenberg, and the Hamiltonian representing spin-spin exchange interaction is given by

$$H_{\mathrm{ex}} = -\sum_{\langle ij\rangle} J_{ij} \boldsymbol{S}_i \cdot \boldsymbol{S}_j. \qquad (1)$$

Here $\boldsymbol{S}_i$ is the total spin of all the electrons bound to the atom or ion at the lattice site i. J_{ij} is the phenomenological exchange constant and is positive for ferromagnetic coupling. It is often sufficient to consider only interaction of nearest neighbors, represented by the bracket $\langle ij\rangle$. In the isotropic case J_{ij} is treated as a constant, J.

Different magnetic spin models can be identified by generalizing the Heisenberg Hamiltonian by introducing directional coefficients,

$$H_{\mathrm{ex}} = -\sum_{\langle ij\rangle} \left[a S_{ix} S_{jx} + b S_{iy} S_{jy} + c S_{iz} S_{jz} \right]. \qquad (2)$$

The underlying models are indicated in Table 1. Among these, the Ising model is the simplest, where the interacting spins can be either parallel or antiparallel. The Potts model is a generalization of the Ising model where the number of spin states is extended to k equally spaced directions in a plane.

Other models involve additional interactions: (i) *magnetocrystalline anisotropy* is due to interaction between the crystal field and the spin-orbit coupling, either of easy axis or of easy plane. (ii) *Zeeman energy*: When a strong external magnetic field $\boldsymbol{H}$ is applied, the Hamiltonian is modified with the addition of a term $H_Z = -g\mu_{\mathrm{B}} \sum_i \boldsymbol{S}_i \cdot \boldsymbol{H}$. (iii) One can also consider interactions such as biquadratic exchange, or weak interactions such as Dzyaloshinski–Moriya type.

There are certain materials for which the exchange integrals J_{ij} become negative. Consequently, the spins align antiparallel to each other so that the net magnetization is zero. These materials are called antiferromagnets and are normally modeled to be made up of two sublattices. However, if the resulting total magnetization, which is the difference between that of the two sublattices, is nonzero, the material is a ferrimagnet.

A spin glass is another state of magnetism that is different from the long-range ordered ferro- and antiferromagnetic phases. A spin glass phase can be considered as consisting of randomly interacting spins with ferro- or antiferromagnetic coupling between them. The combination of the randomness with the competing or mixed interactions causes frustration. Consequently, the exchange integrals J_{ij} are normally distributed random variables corresponding to the so-called Ruderman, Kittel, Kasuya, and Yoshida (RRKY) interaction. The two famous models that describe spin glass are the Edwards–Anderson (EA) model and the Sherrington–Kirkpatrick (SK) mean field model.

The spin glass theory can also be considered as a statistical model of neural networks. The simple spin glass theory based on the SK model provides a qualitative description of memory. In a neural network, the minimal structure element is a neuron whose state can be described by one real variable, and a simple Ising representation $\boldsymbol{S}_i = \pm 1$ is often used. The problem of neural network systems can be formulated by writing the Hamiltonian in Ising form with the so-called Hebbian learning rule for J_{ij}, which is the Hopfield model (Dotsenko, 1994).

The Bethe ansatz is extremely useful in solving quantum spin chain models and to obtain exact results including ground state and excited states. A chain of spin one-half with spin-spin coupling has been solved using the Bethe ansatz method. However, the quantum mechanical description corresponding to the spin Hamiltonian with higher spin values is extremely complicated because it is difficult to know in which particular state one of the individual spins finds itself. Even in the ground state this is impossible. Hence in these cases, one looks for limiting procedures, either classical or semiclassical.

Owing to the presence of the exchange interaction, a small disturbance in one of the spins will be propagated in the form of spin waves. It can also be regarded as oscillations in the magnetic moment density, propagating through a magnetically ordered crystal. The spin wave energy must be equal to the excitation energy of the crystal required to cause a change in the orientation of the spin. The basic entity of the quantized form of the spin wave is called a magnon.

In semiclassical theories of magnetism it is common to approximate the unwieldy spin operators or matrices (which obey the standard spin commutation relations) by the analogous harmonic oscillator operators. Among the machineries established to handle the spin operators, namely, (i) Holstein–Primakoff (H–P) transformation, (ii) Schwinger–Boson representation, (iii) Dyson–Maleev transformation, and (iv) Jordon–

Wigner transformation. The H–P transformation is the most useful one.

Just like photons and phonons, spin waves are bosons describing collective excitations that involve all the spins to the same degree. The operators $S_n^{\pm} = S_{nx} \pm \mathrm{i}S_{ny}$ act on the ground state as field theoretical creation and annihilation operators. It is, thus, reasonable to change by means of a unitary transformation to operators a and a^{+} that satisfy the Bose commutation relation $[a_i, a_j^{+}] = \delta_{ij}$. In this way, one can obtain the H–P transformation of spin operators, which transforms the quantum mechanical spin operators into standard boson operators. In the low-temperature limit, one can use semiclassical expansion for S_n^{+} and S_n^{-}. The Heisenberg equation of motion for the boson operator $\mathrm{i}\hbar \mathrm{d}a_j/\mathrm{d}t = [a_j, H]$ when combined with Glauber's coherent state representation for Bose operators in the long wavelength, low temperature limit (continuum limit) leads to nonlinear evolution (partial differential) equations. Specific one-dimensional cases correspond to integrable spin models exhibiting soliton type spin excitations in addition to linear magnons (Kosevich et al., 1991).

Alternatively, in the large spin limit, the quantum description goes over to the classical vector model for spin, so that the spin operator can be replaced by an ordinary vector of length $s(s+1) > s$, which can be treated as a dynamical/canonical variable. For a ferromagnet described by Heisenberg Hamiltonian (Equation (1)) or its generalization, the classical canonical equations for the spins can be obtained in analogy with a spinless nonrelativistic particle, which is a typical nonlinear dynamical system. The resultant equation of motion $\mathrm{d}\boldsymbol{S}_n/\mathrm{d}t = \{\boldsymbol{S}_n, H\}$, where $\{\ \}$ represents spin Poisson bracket defined for functions of spins under continuum approximation, reduces to the appropriate Landau–Lifshitz (LL) equation. In real magnets, dissipative effects also play an important role, which is included in the LL equation via Gilbert damping. Thus, the LL equation with Gilbert damping (originally derived phenomenologically) is written as

$$\begin{aligned} \frac{\partial}{\partial t}\boldsymbol{S}(\boldsymbol{r},t) &= \boldsymbol{S} \times \boldsymbol{F} + \lambda \boldsymbol{S} \times (\boldsymbol{S} \times \boldsymbol{F}), \\ \boldsymbol{S} &= (S_x, S_y, S_z), \quad \boldsymbol{S}^2 = 1, \qquad (3) \end{aligned}$$

where $\boldsymbol{F}$ is the effective field and λ is the Gilbert damping parameter. A typical form for $\boldsymbol{F}$ is $\boldsymbol{F} = J\nabla^2\boldsymbol{S} + 2A(\boldsymbol{S} \cdot \boldsymbol{n})\boldsymbol{n} + \mu\boldsymbol{H}$, where A is the anisotropy parameter, $\boldsymbol{H}$ is the external magnetic field, and $\boldsymbol{n}$ is a unit vector.

The Landau–Lifshitz equation is a highly nontrivial vector nonlinear partial differential equation in three spatial dimensions. Several one-dimensional cases (without damping) corresponding to quasi-one-dimensional ferromagnetic spin chains (Mikesksa & Steiner, 1991; Kosevich et al., 1991) have been found to be completely integrable soliton systems. These include the isotropic case, uniaxial, and biaxial as well as circularly symmetric cases and are related to the nonlinear Schrödinger equation (Lakshmanan, 1977), (Lakshmanan et al., 1990) and its generalizations. In two dimensions, instanton and vortex-type structures have been identified. Magnetic spatiotemporal patterns have also been identified in the presence of damping and external oscillating fields.

The study of spin excitations in magnetic systems has numerous potential applications. These include (i) magnetic storage devices such as magnetic disks in computers, (ii) spintronics or spin electronics (microelectronic devices that function using the spin of the electrons), (iii) magnetic microchips (magnetic dots, magnetic waves, and magnetic computers), and (iv) magneto-optical recording. (For details, refer to the special issue on Magneto-electronics, *Physics Today*, April 1995).

MUTHUSAMY LAKSHMANAN

See also **Attractor neural network; Bethe ansatz; Ferromagnetism and ferroelectricity; Ising model; Landau–Lifshitz equation; Poisson brackets; Solitons**

Further Reading

Dotsenko, V. 1994. *An Introduction to the Theory of Spin Glasses and Neural Networks*, Singapore: World Scientific

Kosevich, A.M. Ivanov, B.A. & Kovalev, A.S. 1991. Magnetic Solitons. *Physics Reports*, 194: 117–238

Lakshmanan, M. 1977. Continuum spin systems as an exactly solvable dynamical system. *Physics Letters* A, 61: 53–54

Lakshmanan, M., Porsezian, K. & Daniel, M. 1990. On the nonlinear dynamics of the one-dimensional classical Heisenberg ferromagnetic spin chain. In *Nonlinear Evolution Equations: Integrability and Spectral Methods*, edited by A. Degasperis, A.P. Fordy & M. Lakshmanan, Manchester: Manchester University Press

Mattis, D.C. 1988. *Theory of Magnetism I: Statics and Dynamics*, Berlin and New York: Springer

Mikeska, H.J. & Steiner, M. 1991. Solitary excitations in one-dimensional magnets. *Advances in Physics*, 40: 191

SPIRAL WAVES

Rotating spiral waves occur in a wide range of chemical, biological, and physical systems that can be considered as effectively two-dimensional excitable or oscillatory media in which the local nonlinear dynamics exhibits threshold behavior, separating an excited state from a recovered state (Holden et al., 1991). In such a medium, waves of excitation spread without decrement through the stationary medium, with diffusive spread of the excitation from a region triggering excitation of neighboring regions.

A single spiral wave in a large homogeneous medium acts as an organizing center, imposing its pattern throughout the medium. It has the characteristics of

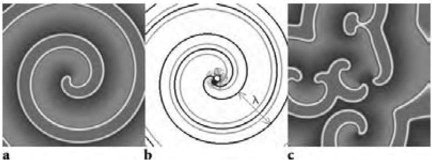

Figure 1. Spiral wave solutions of the FitzHugh–Nagumo equations for an excitable medium: (a) snapshot of solitary spiral wave; (b) intersection of excitation and recovery isolines defining the tip (solid circle), which meanders biperiodically around the central corel; and (c) evolving field of wavebreaks.

a field, in that it fills the medium, but it also has a location. The location of a spiral wave is that of its center, variously defined as the central core that is not invaded by the propagating wave, and around which the spiral rotates; or as a phase singularity; or by the tip, defined by where the wave front meets the wave back; or by the intersection of two isolines. For a rigidly rotating spiral in an isotropic medium, the tip encloses a circular core, but even in a homogeneous medium, the tip motion can be complex, meandering quasi-periodically within a bounded region, or hypermeandering in an unbounded region, and any inhomogeneities and boundaries interactions can produce drift (Barkley, 1994). Along any radius far from the core, activity appears as a periodic wave train with a temporal frequency ω (for a rigidly rotating spiral equal to the rotation frequency of the spiral) and spatial wavelength λ. These are generally single-valued functions of the medium parameters. The wave front (or any other isoline) approximates an Archimedean spiral (the involute of a circle) with a pitch λ, as illustrated in Figure 1b. The simple (one-armed) Archimedean rotating spiral in polar coordinates (θ, r) is a periodic function of phase ϕ

$$\phi = \omega t \pm \theta + a(r - r_0) \qquad (1)$$

with $a > 0$ giving the wavelength λ and pitch $\mathrm{d}r / \mathrm{d}\theta|_{\phi = \mathrm{const}}$ equals to $1 / a$.

If a stable spiral wave is subject to a parametric perturbation, the form of the spiral will be restored, but its location and phase may be changed: this allows periodic perturbations, at the same frequency as the rotation frequency of the spiral, to produce a directed, resonant drift. A single spiral need not be stable—even in a parametrically homogeneous medium, dynamical heterogeneities produced by the rotating spiral wave can lead to wave breaks, resulting in pairs of new phase singularities that may develop into new spirals, collapse, or be extinguished, giving a spiral wave turbulence as illustrated in Figure 1c. In such an evolving field of wave breaks, the singularities may be considered as defects; are born and annihilate on collision, in pairs with opposite chirality; or exit the medium at boundaries.

The best studied laboratory example of spiral waves is the Belousov–Zhabotinsky reaction, a chemical excitable or oscillatory reaction originally constructed as an organic chemical analogue for oscillatory reactions in a biochemical system, in which the phase of the redox oscillation is monitored by color changes. In a shallow layer of solution in a Petri dish, spirals and interacting spirals can be initiated (Winfree, 1972), and spatiotemporal changes in activity can be monitored optically and quantified (Müller et al., 1987). Winfree's experimental and numerical modeling of simplified models and caricatures (Winfree, 1991) of two- and three-dimensional excitable systems established much of the phenomenology of spiral wave behavior.

Spiral waves occur in several biological systems. In many cell types there are intracellular waves of calcium, where calcium-induced calcium release provides the positive feedback necessary for excitability. In the slime mold, where the intercellular relaying of a chemotactic signal—cyclic AMP—between individual amoebae by cAMP-induced cAMP release provides the excitability, the social amoebae collect together in a rotating spiral wave and aggregate into a slug, which can then metamorphose into a fruiting body. An important example is from cardiac patho-physiology, where the three-dimensional analogue of a spiral, the scroll, provides the re-entrant propagation that produces high frequency arrhythmias in cardiac muscle and provides the route to ventricular fibrillation and sudden cardiac death.

In physics, spiral waves are found in surface catalysis, nonlinear optical media, and it has been suggested that spiral galaxies represent a rotating spiral wave of star formation and death.

In all these examples, the excitable medium can be modeled by a reaction-diffusion equation, in which only the excitatory variable (as in most biological examples) or both excitatory and recovery variables (as in most chemical examples) have nonzero diffusion coefficients. Spiral wave solutions of mechanistically detailed, high order models, or simplifications, can be obtained by initiation from artificial initial conditions, or by imposing a wave break or wavebreaking dynamical or parametric heterogeneity. Much of the phenomenology can be seen in the simple, two-variable FitzHugh–Nagumo system, or computationally faster simplifications (Zykov, 1987), and reproduced by eikonal equations that describe the dependence of the propagation velocity of the wave front on its curvature (Mikhailov et al., 1994).

Arun V. Holden

See also **Belousov–Zhabotinsky reaction; Excitability; FitzHugh–Nagumo equation; Geometrical optics, nonlinear; Reaction-diffusion systems; Vortex dynamics in excitable media**

Further Reading

Barkley, D. 1994. Spiral meandering. In *Chemical Waves and Patterns*, edited by R. Kapral & K. Showalter. Dordrecht: Kluwer, pp. 163–194

Holden, A.V., Markus, M. & Othmer, H.G. 1991. *Nonlinear Wave Processes in Excitable Media*, New York: Plenum Press

Mikhailov, A.S., Davydov, V.A. & Zykov, V.S. 1994. Complex dynamics of spiral waves and motion of curves. *Physica* D, 70: 1–39

Müller, S.C., Plesser, T. & Hess, B. 1987. Two-dimensional spectrophotometry of spiral wave propagation in the Belousov–Zhabotinskii reaction. *Physica* D, 24: 71–96

Winfree, A.T. 1972. Spiral waves of chemical reaction. *Science*, 175: 634–636

Winfree, A.T. 1991. Varieties of spiral wave behavior: an experimentalist's approach to the theory of excitable media. *Chaos* 1: 303–334

Zykov, V.S. 1987. *Simulation of Wave Processes in Excitable Media*, Manchester: Manchester University Press

STABILITY

The concept of stability lies at the heart of mathematical physics. Although in a perfect environment it might be thought possible to balance a pencil on its point, this is not possible in the real, noisy world—a pencil standing on its point is unstable. Meanwhile a hexagonal pencil lies easily on its side in a stable equilibrium. Intermediate is the pencil standing on its flat end, which can be achieved, albeit with difficulty, because this configuration is metastable, meaning stable against infinitesimally small disturbances, $\boldsymbol{D}_\mathrm{S}$, and unstable against large disturbances, $\boldsymbol{D}_\mathrm{L}$. Notice that under $\boldsymbol{D}_\mathrm{S}$, the pencil has more than one stable equilibrium (multistability). A special case is a circular pencil on its side: if it rolls a little, and it will not come back and so is marginally stable. Generalizing these experiences, a state is declared stable if it is physically observable under $\boldsymbol{D}_\mathrm{S}$: otherwise, it is unstable.

In 1892, Aleksandr Lyapunov formalized this intuitive idea of stability against $\boldsymbol{D}_\mathrm{S}$ by considering free motions of a dynamical system in phase space. Although this ignores external noise, the results agree with more realistic stochastic criteria. Consider a system of n first-order differential equations, $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$, giving a stationary vector field in the phase space spanned by the n components of vector $\boldsymbol{x}$. Trajectories fill this space with a fluid-like flow. A driven oscillator is put into this form by identifying velocity as a second phase coordinate and time as a third (with $\dot{t} = 1$).

A fixed (equilibrium) point with $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ is asymptotically stable if all local trajectories flow into it (Figure 1). This point attractor is exemplified by the hanging state of a damped pendulum. The equilibrium of a Hamiltonian system (undamped pendulum) can be at most neutrally stable with all trajectories staying close, though not converging. A point is unstable if any single adjacent motion moves out of the neighborhood (inverted pendulum).

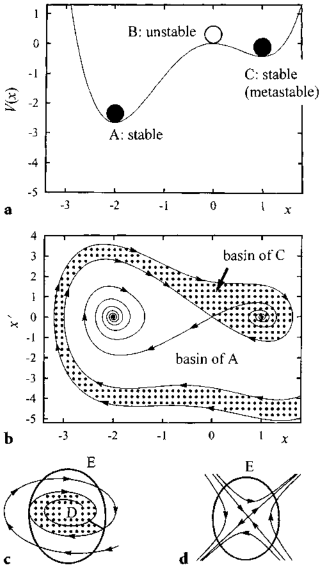

Figure 1. Potential well (a), and phase portrait (b), for $\ddot{x} + 0.4\dot{x} + \mathrm{d}V(x)/\mathrm{d}x = 0$, where $V(x) = x^4/4 + x^3/3 - x^2$, showing two asymptotically stable minima and an unstable maximum. In the enlarged focus (c), representing either minimum, we can devise a region D from which no trajectories leave the small region E, guaranteeing stability; in (d) we cannot construct a closed neighborhood around the saddle from which no trajectories leave E, proving instability.

Linear Stability

Stability analysis starts with the linearized equations describing small variations about the state. For a fixed point of a flow, stability hinges on the real parts, α_i, of the eigenvalues. Typically, the point will be hyperbolic (noncritical) with no $\alpha_i = 0$, and we then have (asymptotic) stability if and only if (*iff*) all $\alpha_i < 0$. Any one $\alpha_i > 0$ gives instability.

For a cycle we use an $(n - 1)$-dimensional Poincaré mapping, $\boldsymbol{x}_{i+1} = \boldsymbol{F}(\boldsymbol{x}_i)$, where the cycle is a fixed point with $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i$. Stability hinges on the moduli, ρ_i, of the eigenvalues. Stability of a (hyperbolic) periodic orbit requires all $\rho_i < 1$.

An unstable state, attracting in some directions, repelling in others, is a saddle. The instability index (degree of instability) is the number of $\alpha_i > 0$ (or $\rho_i > 1$). In nonlinear dissipative systems, saddles with index 1 play an important role organizing basin boundaries.

Nonlinear Stability

A conservative system has zero divergence in phase space (Liouville's theorem), like an incompressible fluid. With dissipation, the negative divergence ensures that an ensemble of starts will shrink onto an attracting set of zero volume. Such a post-transient can be a point, periodic, quasi-periodic, or chaotic attractor. Generically, each attractor is surrounded by its own basin of attraction. All transients from a vanishingly small neighborhood move asymptotically back to it, making it asymptotically stable.

To assess how stable an attractor is against finite disturbances, we estimate the size of its basin. A brute-force method is to make computer simulations from a grid of starts; more efficient is cell-to-cell mapping (Hsu, 1980). A refined approach is to locate a governing saddle and run simulations backwards in time to trace out the boundary formed by its stable manifold (inset). Boundaries can, however, be fractal, making their detailed location impossible (McDonald et al., 1985; Thompson & Stewart, 2002). Analytical techniques, notably Lyapunov's second (direct) method (La Salle & Lefschetz, 1961), can identify contracting (trapping) regimes of guaranteed return, but they may be excessively conservative.

Local nonlinear analysis is necessary to test the Lyapunov stability of a nonhyperbolic state at which a system is about to lose its stability. An oscillator with exclusively nonlinear negative damping is, for example, linearly (non-asymptotically) stable but nonlinearly unstable. In contrast to Lyapunov's local concept, a point is globally stable if every trajectory tends to it.

Loss of Stability

When linear theory indicates a loss of stability, the nonlinear system exhibits a bifurcation. Consider first an α_i passing through zero along a path. If a real eigenvalue becomes zero, we have a saddle-node bifurcation. Typically this gives a fold, at which the path folds smoothly back at a maximum of the control parameter, μ. A saddle-node coalescence at $\mu = \mu^C$ leaves no local solution; the system jumps to a distant, unrelated attractor (Figure 2). With symmetry the saddle-node manifests itself as a pitchfork (Figure 3). Two α_i of a complex conjugate pair passing through zero give the Hopf bifurcation generating stable or unstable cycles.

A mapping loses stability when a ρ_i reaches one. If a real eigenvalue equals $+1$, we have a saddle-node bifurcation, typically a cyclic fold. If it equals -1, we have a flip (period-doubling) bifurcation. If a complex conjugate pair penetrates the unit circle, we have the Neimark–Sacker (secondary Hopf) bifurcation.

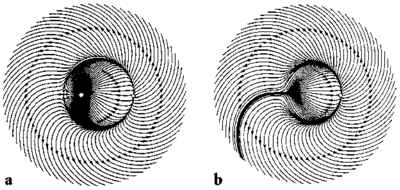

Figure 2. Two phase portraits illustrating the ABC of nonlinear dynamics. Here A stands for attractor, an example of which we see in the center of (a). B stands for basin, and we see the circular basin of attraction in (a); its boundary defined by the flow into an unstable saddle point. C stands for catastrophe, conveniently used here to mean bifurcation. Under variation of a control, (a) has changed qualitatively to become the portrait (b); the central stable node has collided with the saddle in a fold bifurcation, leaving no attractor and all trajectories flowing out to infinity. (Figure reproduced with permission of John Wiley from Thompson & Stewart 2002.)

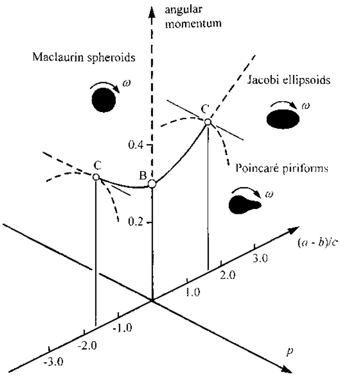

Figure 3. Stability of a rotating mass of a self-gravitating liquid is a classical problem in astrophysics, relevant to planetary evolution and the origin of satellites. It dates back to Poincaré (1899), and following the buckling studies of Euler, it gave fresh impetus to the development of stability and bifurcation theory. Stable (unstable) paths are denoted by solid (broken) lines. As the angular momentum parameter is slowly increased, the rotationally symmetric spheroids become unstable at the supercritical pitchfork at B. Emerging from B is a rising path of stable ellipsoids (with semi-axes a, b, c), which then exhibit a secondary instability at the subcritical pitchforks at C, from which unstable pear-shaped forms of "amplitude" p descend. (Figure reproduced with permission of John Wiley from Thompson, 1982.)

Structural Stability

Because parameters of a physical system are never known precisely and may vary, a mathematical model should ideally have a phase portrait that is robust against small changes in the model itself. Small changes of parameters and functions should not qualitatively change the topology. A portrait having this robustness is structurally stable. Structurally unstable portraits encountered under variation of μ signal a bifurcation.

If we have m controls, μ_i, bifurcations observed under variation of one μ_i will lie in control space on surfaces of dimension $m-1$. A bifurcation in phase-control space is structurally stable if the whole event is topologically robust. Catastrophe theory (Poston & Stewart, 1978) shows how bifurcations of equilibria that are structurally unstable under one μ_i can be made structurally stable (unfolded) by embedding in a higher-dimensional control space. The codimension is the number of μ_i needed to achieve this. An example is the symmetry-breaking imperfection needed to unfold a pitchfork into the codimension-two cusp. The resulting saddle-node locus gives the imperfection-sensitivity of elastic stability theory (Thompson & Hunt, 1973).

MICHAEL THOMPSON

See also **Attractors; Equilibrium; Phase space; Wave stability and instability**

Further Reading

Hsu, C.S. 1980. A theory of cell-to-cell mapping for nonlinear dynamical systems. *ASME Journal of Applied Mechanics*, 47: 931–939

La Salle, J. & Lefschetz, S. 1961. *Stability by Liapunov's Direct Method with Applications*, New York: Academic Press

Lyapunov, A. 1892. *Problème général de la stabilité du mouvement*, Kharkov (in Russian); French translation in *Annals of the Faculty of Science, University of Toulouse*, vol. 9 (1907); reprinted, Princeton, NJ: Princeton, University Press (1949)

McDonald, S.W., Grebogi, C., Ott, E. & Yorke, J.A. 1985. Fractal basin boundaries. *Physica* D, 17: 125–153

Poincaré, H. 1899. *Les Méthodes nouvelles de la mécanique céleste*, vols. 1–3, Paris: Gauthier-Villars; as *New methods of celestial mechanics*, 3 vols., Woodbury, NY: American Institute of Physics, 1991

Poston, T. & Stewart, I. 1978. *Catastrophe Theory and Its Applications*, London: Pitman

Thompson, J.M.T. 1982. *Instabilities and Catastrophes in Science and Engineering*, Chichester: Wiley

Thompson, J.M.T. & Hunt, G.W. 1973. *A General Theory of Elastic Stability*, London: Wiley

Thompson, J.M.T. & Stewart, H.B. 2002. *Nonlinear Dynamics and Chaos*, 2nd edition, Chichester and New York: Wiley

STABLE AND UNSTABLE MAINFOLDS

See **Phase space**

STANDARD MAP

The standard or Taylor–Chirikov map is a family of area-preserving maps: $z' = f(z)$ where $z = (x, y)$ is the original position and $z' = (x', y')$ the new position after application of the map, which is defined by

$$\begin{aligned} x' &= x + y - \frac{k}{2\pi}\sin(2\pi x), \\ y' &= y - \frac{k}{2\pi}\sin(2\pi x). \end{aligned} \tag{1}$$

Here x is a periodic configuration variable (usually computed modulo 1), $y \in \mathbb{R}$ is the momentum variable, and the parameter k represents the strength of a nonlinear kick. This map was first proposed in 1968 by Bryan Taylor and then independently obtained by Boris Chirikov to describe the dynamics of magnetic field lines. The standard map and Hénon's area-preserving quadratic map are extensively studied paradigms for chaotic Hamiltonian dynamics.

The standard map is an "exact symplectic" map of the cylinder. Because $x'(x, y)$ is a monotone function of y for each x, it is also an example of a monotone twist map (*See* **Aubry–Mather theory**). Every twist map has a Lagrangian generating function, and the standard map is generated by $F(x, x') = \frac{1}{2}(x' - x)^2 + (k/4\pi^2)\cos(2\pi x)$, so that $y = -\partial F/\partial x$ and $y' = \partial F/\partial x'$. The map can also be obtained from a discrete Lagrangian variational principle as follows. Define the discrete action for any configuration sequence $\ldots, x_{t-1}, x_t, x_{t+1}, \ldots$ as the formal sum

$$A[\ldots, x_{t-1}, x_t, x_{t+1}, \ldots] = \sum_t F(x_t, x_{t+1}). \tag{2}$$

Then an orbit is a sequence that is a critical point of A. This gives the discrete Euler–Lagrange equation

$$x_{t+1} - 2x_t + x_{t-1} = -\frac{k}{2\pi}\sin(2\pi x_t). \tag{3}$$

This second difference equation is equivalent to (1) upon defining $y_t = x_t - x_{t-1}$.

The standard map is an exact or approximate description of many physical systems, including the "kicked rotor." Consider a rigid body with moment of inertia I that is free to rotate in a horizontal plane about its center of mass. Suppose that an impulsive torque $\Gamma(\theta) = -A\sin(\theta)$ is applied to the rotor at times nT, $n \in \mathbb{Z}$. Let (θ_j, L_j) be angular position and angular momentum at time $jT - \varepsilon$ for $\varepsilon \to 0^+$. At time T later, these become $(\theta_{j+1}, L_{j+1}) = (\theta_j + (T/I)L_{j+1}, L_j + \Gamma(\theta_j))$. Scaling variables appropriately gives (1).

The standard map also describes the relativistic cyclotron and is the equilibrium condition for a chain of masses connected by harmonic springs in a periodic potential—the Frenkel–Kontorova model introduced in 1938 (Meiss, 1992). Similar maps include Chirikov's separatrix map (valid near the separatrix of a resonance), the Kepler map (describing the motion of comets under the influence of Jupiter as well as a classical hydrogen atom in a microwave field), and the Fermi map (for a ball bouncing between oscillating walls) (Lichtenberg & Lieberman, 1992). The higher-dimensional version is the Froeshlé map (*See* **Symplectic maps**).

Symmetries

The standard map f has a number of symmetries that lead to special dynamical behavior. To see these, it is convenient to lift the map from the cylinder to the plane by extending the angle variable x to $\mathbb{R}$.

Let $T_{m,n}(x, y) = (x + m, y + n)$ be the translation by an integer vector (m, n). As f is periodic, its lift has a discrete translation symmetry $f \circ T_{m,0} = T_{m,0} \circ f$. More unusually, the standard map also has a discrete vertical translation symmetry $f \circ T_{0,n} = T_{0,n} \circ f$. Identifying orbits equivalent under these symmetries implies that standard map can be thought of as acting on the torus $\mathcal{T} = \{-\frac{1}{2} \le x, y < \frac{1}{2}\}$.

The standard map also commutes with the reflection $S(x, y) = (-x, -y)$. This can be used to identify the lower half plane with the upper one and to restrict the map to the space $\mathcal{S} = \{(x, y) : -\frac{1}{2} \le x < \frac{1}{2},\ 0 \le y \le \frac{1}{2}\}$ identifying $(-\frac{1}{2}, y) \equiv (\frac{1}{2}, y)$ and each half of the upper and lower boundaries: $(x, 0) \equiv (-x, 0)$, $(x, \frac{1}{2}) \equiv (-x, \frac{1}{2})$. The map on the two sphere $\mathcal{S}$ is singular at the corners $(\pm\frac{1}{2}, 0)$ and $(\pm\frac{1}{2}, \frac{1}{2})$.

The standard map is also reversible; it is conjugate to its inverse $RfR^{-1} = f^{-1}$ (Lamb & Roberts, 1998). One reversor is $R_1(x, y) = (-x, y - (k/2\pi) \sin(2\pi x))$; this generates a family of reversors $R = f^n \circ R_1$. These reversors are involutions, $R^2 = id$, thus f can be written as the composition of two involutions $f = (f \circ R) \circ R$. Finally, the composition of a symmetry and a reversor is also a reversor, so that, for example, $R_2 = SR$ is also a reversor.

Symmetric orbits are invariant under a symmetry or a reversor. This is particularly interesting since symmetric orbits must have points on the fixed sets of the reversor, $\mathrm{Fix}(R) = \{z : z = R(z)\}$, or on $\mathrm{Fix}(fR)$. Because these fixed sets are curves, symmetric orbits are particularly easy to find. Rimmer showed that the bifurcations of symmetric orbits are special; for example, they undergo pitchfork bifurcations (*See* **Bifurcations**).

Dynamics

When $k = 0$, the dynamics of the standard map are integrable: the momentum y is an invariant. On each invariant circle $\mathcal{C}^0_\omega = \{(x, y) : y = \omega\}$, the angle after t iterates is given by $x_t = x_0 + \omega t \bmod 1$; thus, the dynamics are that of the constant rotation, $R_\omega(\theta) = \theta + \omega$, on the circle with rotation number ω. When ω is rational, every orbit on $\mathcal{C}^0_\omega$ is periodic; otherwise, they are quasiperiodic and densely cover the circle.

When $|k| \ll 1$, Moser's version of the KAM theorem implies that most of these invariant circles persist; that is, there is a rotational invariant circle $\mathcal{C}_\omega$ on which the dynamics is conjugate to the rotation R_ω (*See* **Hamiltonian systems**). KAM theory applies to circles with Diophantine rotation number, that is, $\omega \in \{\Omega : |n\Omega - m| > c/n^\tau;\ \forall m, n \in \mathbb{Z}, n \ne 0\}$ for some $\tau \ge 1$ and $c > 0$. This excludes, of course, all of the rational rotation numbers as well as intervals about each rational, but still leaves a positive measure set. While it is difficult to obtain reasonable estimates for the interval of k for which all Diophantine circles (with given c and τ) persist, in 1985 Herman showed analytically that there is at least one invariant circle when $|k| \le 0.029$, and de la Llave & Rana (1990) used a computer-assisted proof to extend this result up to 0.91.

Some of the periodic orbits on the rational circles $\mathcal{C}^0_{m/n}$ also persist for nonzero k. Indeed, the Poincaré–Birkhoff theorem implies that there are at least two period n orbits (with positive and negative Poincaré indices). Aubry–Mather theory implies that orbits with rotation number m/n can be found variationally; one is a global minimum of the action (2), and the other is a minimax point (a saddle of A with one downward direction). For example, when $k > 0$, $(\frac{1}{2}, 0)$ is a minimizing fixed point, and $(0, 0)$ is a minimax fixed point. The reversibility of the standard map implies that there must be symmetric periodic orbits for each $\omega = m/n$ as well. Indeed, it is observed that the minimax periodic orbits always have a point on the line $\mathrm{Fix}(R) = \{y = 0\}$, the "dominant" symmetry line.

The minimax orbits are elliptic when k is small enough. A convenient measure of stability of a period-n orbit is Greene's residue

$$R = \frac{1}{4}(2 - Tr(M)), \quad M = \prod_{t=0}^{n-1} Df(z_t).$$

An orbit is elliptic when $0 < R < 1$. For example, the fixed point (0, 0) has residue $k/4$. Perturbation theory shows that the residues of the minimizing and minimax orbits are $\mathcal{O}(k^n)$.

Each nondegenerate minimum of the action (2) is a hyperbolic orbit and has unstable and stable manifolds. For each minimizing m/n orbit, these intersect and enclose the minimax orbit, forming an island chain or resonance. The intersection of the manifolds is transverse, though the angle between them is exponentially small in k (Gelfreich & Lazutkin, 2001).

A number of island chains are easily visible in computer simulations. In the color figure (*See* **Standard map**: on page 8 of color plate section), we show a number of orbits of the standard map for $k = 0.6$ on the torus $\mathcal{T}$. In the figure, each of the blue curves is formed from many iterates on a rotational invariant circle like those predicted by the KAM theorem. The green orbits are secondary and tertiary circles arising from resonances.

When stable and unstable manifolds intersect transversely, some iterate of the map has a Smale horseshoe. This implies that there is, at least, a Cantor set of chaotic orbits. Umberger & Farmer (1985) showed numerically that there is a fat fractal set on which the dynamics has a positive Lyapunov exponent. The proof of this statement is still illusive. The regions occupied by chaotic orbits appear to grow in measure as k increases. Numerically, it appears that a single initial condition densely covers each "zone of instability," a chaotic zone bounded by invariant circles. Chaotic trajectories that were only slightly visible in the previous color figure (gold orbits near the stable and unstable manifolds of the resonances) dominate the dynamics when $k = 2.0$ (*See* **Standard map**: page 8 in the color plate section). At this value of k, there are no rotational invariant circles. In the figure, the gold region is filled by a single trajectory with 1.5×10^6 iterates. It appears to densely cover most of phase space, though there are still a number of secondary and tertiary islands visible.

There are also many elliptic periodic orbits that are created for nonzero k. For example, the $(0, 0)$ fixed point undergoes a period-doubling bifurcation at $k = 4$, creating a period-two orbit. More generally, when the eigenvalues of any elliptic period-n orbit are $\lambda_\pm = \mathrm{e}^{\pm 2\pi \mathrm{i}\omega}$, then new orbits are born that encircle the original orbit and have relative rotation number ω. When $\omega = m'/n'$, these correspond to a chain of nn' islands. The color figure (*See* **Standard Map**: page 8 in the color plate section) shows red and blue island chains that encircle the fixed point with rotation number $\frac{1}{5}$; that there are two such chains is due to the reflection symmetry S. As Birkhoff realized, the newly created elliptic orbit also will undergo similar bifurcations, so that the phase space shows a structure of islands-around-islands, ad infinitum. This structure can even exhibit self-similarity (Meiss, 1992) just like the Feigenbaum period-doubling sequence for dissipative systems.

The Last Invariant Circle

In 1968, John Greene began studying the destruction of invariant circles in the standard map. He showed that sequences of periodic orbits, namely, the minimizing and minimax m/n orbits, whose rotation numbers converge on a given irrational, can be used to determine the existence of a circle with that frequency. Suppose that ω has a continued fraction expansion $[a_0, a_1, \ldots]$, $a_j \in \mathbb{Z}^+$, and let $m_j/n_j = [a_0, a_1, \ldots, a_j]$ be the jth convergent of ω. Greene conjectured that when the residues of these orbits $R_j \to 0$, as $j \to \infty$, then the invariant circle $\mathcal{C}_\omega$ exists—MacKay (1992) gives a proof of much of this.

For the standard map, it appears that each rotational invariant circle exists only up to a critical value, $k = k_{\mathrm{cr}}(\omega)$; this graph was called the "fractal diagram" by Schmidt and Bialek in 1982. The critical k vanishes at every rational and appears to have local maxima for each *noble* irrational ω. Percival called a number noble if its continued fraction expansion has a tail that is eventually all ones. By this criterion the "most irrational" number is the golden mean $\gamma = (1 + \sqrt{5})/2 = [1, 1, 1, \ldots]$. Indeed for the standard map, Greene discovered that the invariant circles with rotation numbers $\gamma \pm m$, $m \in \mathbb{Z}$ appear to be the last circles destroyed (all such circles are destroyed simultaneously due the symmetries). Numerically it is known that the golden circle is destroyed at $k_{\mathrm{cr}}(\gamma) \approx 0.971635406$.

This value is most efficiently computed by renormalization theory (MacKay, 1993). At the critical parameter for the destruction of a noble invariant circle, the phase space exhibits a self-similar structure. The geometric scaling of this self-similarity can be used to compute k_{cr} from the residues of the m_j/n_j orbits. This is more accurate than iteration methods—pioneered by Chirikov—which rely on finding an orbit that crosses the region containing the circle, and frequency methods—developed by Laskar—which rely on the irregularity of the numerically computed rotation number. While none of these methods proves that k_{cr} corresponds to the last invariant circle, "converse KAM theory" leads to a computer proof that there are no rotational circles for $k > \frac{63}{64}$ (MacKay & Percival, 1985). This is based on Birkhoff's theorem that every rotational invariant circle is a Lipschitz graph (Meiss, 1992). The color figure (*See* **Standard map**: page 8 in the color plate section) shows the dynamics on the sphere $\mathcal{S}$ at $k_{\mathrm{cr}}(\gamma)$. Here the the golden circle (purple) is on the threshold of destruction. Also shown in the figure are 1.5×10^6 iterates of two chaotic trajectories (light blue and light green); the stable (blue) and unstable (red) manifolds of the $(m, n) = (0, 1)$, $(1, 2)$, and (1.3) orbits; and a number of orbits trapped in these island chains as well as the $(2, 5)$ and $(3, 8)$ chains.

Transport

Transport theory studies the motion of ensembles of trajectories from one region of phase space to another. When there are invariant circles separating the regions, then there is no transport. A Birkhoff "zone of instability" is an annular region bounded by, but otherwise not containing any, rotational invariant circles. Birkhoff showed there are orbits that traverse each zone of instability, and Mather (1991) extended this to show that there are orbits future and past asymptotic to the upper and lower bounding rotational invariant circles, respectively.

Aubry–Mather theory implies that for each irrational rotation number there is a minimizing trajectory that is dense on a circle or a cantor set. Percival proposed calling the latter sets "cantori." Thus, for $k > \frac{63}{64}$, every rotational invariant circle has become a cantorus, and vertical transport between any two momentum levels occurs. The color figure (in the color plate section) shows two such cantori (brown) with rotation numbers $(1+\gamma)/(3+4\gamma)$ and $(1+2\gamma)/(2+5\gamma)$. The rate of transport is locally governed by the flux, the area that crosses a closed loop upon iteration. The flux across a cantorus or a separatrix is given by Mather's ΔW, the difference in action between the corresponding minimax and minimizing orbits (MacKay et al., 1984). Renormalization theory shows that the flux through a noble cantorus goes to zero as $(k-k_{\rm cr})^{3.01}$; this can be very small, well beyond $k_{\rm cr}$. For example in the color figure 3 (page 8 in the plate color section), the blue chaotic trajectory is bounded below by a low flux cantorus even for tens of millions of iterates. Geometrically, the flux is the area contained in a "lobe" bounded by pieces of stable and unstable manifolds. All transport occurs through lobes in two-dimensional maps (Wiggins, 1992); unfortunately, the higher-dimensional generalization is not clear.

JAMES D. MEISS

See also **Aubry–Mather theory; Cat map; Chaotic dynamics; Ergodic theory; Fermi acceleration and Fermi map; Hamiltonian systems; Hénon map; Horseshoes and hyperbolicity in dynamical systems; Lyapunov exponents; Maps; Measures; Mel'nikov method; Phase space; Symplectic maps**

Further Reading

Gelfreich, V.G. & Lazutkin, V.F. 2001. Separatrix splitting: perturbation theory and exponential smallness. *Russian Mathematical Surveys*, 56(3): 499–558

de la Llave, R. & Rana, D. 1990. Accurate strategies for small divisor problems. *Bulletin of the American Mathematical Society*, 22: 85–90

Lamb, J.W.S. & Roberts, J.A.G. 1998. Time-reversal symmetry in dynamical systems: a survey. *Physica* D, 112: 1–39

Lichtenberg, A. & Lieberman, M. 1992. *Regular and Chaotic Motion*, *Applied Mathematical Sciences*, vol. 38, 2nd edition, New York: Springer-Verlag

MacKay, R., Meiss, J. & Percival, I. 1984. Stochasticity and transport in Hamiltonian systems. *Physical Review Letters*, 52: 697–700

MacKay, R.S. 1992. Greene's residue criterion. *Nonlinearity*, 5: 161–187

MacKay, R.S. 1993. *Renormalisation in Area-Preserving Maps*, *Advanced Series in Nonlinear Dynamics*, vol. 6, Singapore: World Scientific

MacKay, R.S. & Percival, I.C. 1985. Converse KAM: theory and practice. *Communications in Mathematical Physics*, 98: 469–512

Mather, J.N. 1991. Variational construction of orbits of twist diffeomorphisms. *Journal of the American Mathematical Society*, 4(2): 207–263

Meiss, J.D. 1992. Symplectic maps, variational principles, and transport. *Reviews of Modern Physics*, 64(3): 795–848

Umberger, D.K. & Farmer, J.D. 1985. Fat fractals on the energy surface. *Physical Review Letters*, 55: 661–664

Wiggins, S. 1992. *Chaotic Transport in Dynamical Systems*, *Interdisciplinary Applied Mathematics*, vol. 2, New York: Springer

STATE DIAGRAMS

Consider a system comprising N elements, each of which can be set at one or the other of two values. Thinking of the elements as switches, the values (positions or settings) can be defined as one of the two Boolean variables: off (0) and on (1) (Birkhoff & MacLane, 1953). Now suppose that time progresses in discrete units and the values of the switches at time t_{n+1} depend on their values at t_n. In addition to being a rather general computer model (Fowler, 2004), such a system describes certain relay and electronic switching circuits used by electrical engineers, and it provides an approximate description of biological brains (Scott, 2002), including the "nets with circles" of Warren McCulloch and Walter Pitts (McCulloch & Pitts, 1943). If the elements are placed in a geometrical pattern (e.g., a square array), one has a cellular automaton, such as John Conway's famous "Game of life" (Gardner, 1970).

To understand how the model evolves in time, it is sometimes instructive to sketch a state diagram, which displays all states of the system and shows how these states change from one time increment to the next.

In Figure 1(a), for example, this is done for a system comprising a single switch ($N=1$), where the arrows indicate what will happen in the next increment of time. Reading from left to right, the first diagram shows a system in which the switch turns on if it is off and stays on if it is on. The second diagram shows a system that turns off if it is on and stays off if it is off. In the third diagram, the system stays off if it is off and stays on if it is on. (This corresponds to a typical wall switch.) Finally, the last diagram represents a system that turns off if it is on and turns on if it is off. (This models a "blocking oscillator" in electronic engineering.)

The number of possible state diagrams increases very rapidly with the number of elements (N) in a system. To see this, note that the number of states of the system is 2^N, and from each of these states we must choose an arrow that goes to (possibly) another state. Thus, the total number of diagrams $\mathcal{N}$ is 2^N raised to the 2^Nth power, or

$$\mathcal{N} = 2^{N2^N}. \tag{1}$$

A few calculations from this formula are given in Table 1, demonstrating a growth of $\mathcal{N}$ with N, which mathematicians term "combinatorial."

For a system comprising only six switches, the total number of state diagrams is greater than 10^{110}, which in turn equals the atomic mass of the universe (the mass of the universe measured in units of the hydrogen atom) times the age of the universe in picoseconds. In his analysis of biological organisms, the physicist Walter Elsasser has called such a finite number "immense," which implies that (although the number of them is finite) it is not—and never will be—physically possible to examine all possible diagrams (Elsasser, 1998).

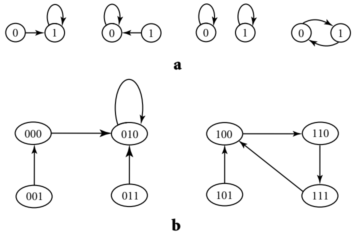

Figure 1. (a) The four Boolean state diagrams that can be constructed from a single switch. (b) One of the more than sixteen million state diagrams that can be constructed from three switches. Note that several arrows can converge on each state, but only one arrow leaves any one state.

N	$\mathcal{N} = 2^{N2^N}$
1	$2^2 = 4$
2	$4^4 = 256$
3	$8^8 \doteq 1.7 \times 10^7$
4	$16^{16} \doteq 1.8 \times 10^{19}$
5	$32^{32} \doteq 1.5 \times 10^{48}$
6	$64^{64} \doteq 3.9 \times 10^{115}$

Table 1. The number of state diagrams ($\mathcal{N}$) that can be constructed for various numbers (N) of Boolean switches.

In Figure 1(b) is shown one of the more than sixteen million state diagrams for a system of three switches. Here we find the following behaviors.

- *Transients:* (001) → (000) → (010), (011) → (010), and (101) → (100).
- *Periodicities:* (010) → (010) and (100) → (110) → (111) → (100).
- *Disjoint dynamics:* States (000), (010), (011), and (001) are dynamically disjoint from states (100), (110), (111), and (101).

The shortest transient or period is a single time unit, whereas the longest is 2^N time units, but this is unlikely for larger values of N.

Consider how the number of state diagrams depends on the way that the state of a system in the next time increment is computed. If each switch responds to the present positions of all N switches (including itself), there are

$$2^{2^N}$$

possible Boolean functions that can perform the computations. In the most general case, any one of the N switches can be assigned one of these Boolean functions. Thus, the number of possible diagrams is equal to

$$\left(2^{2^N}\right)^N = 2^{N2^N},$$

just as in Equation (1). In choosing the arrows for a particular diagram—as in Figure 1(b)—one defines a particular Boolean function.

When state diagrams are used to model biological brains, we may wish to limit the computations to threshold functions rather than Boolean functions. The total number of threshold functions of n inputs is

$$2^{kn^2},$$

where k lies between $\frac{1}{2}$ and 1 (Yajima et al., 1968). In a brain model, furthermore, it is unrealistic to assume that each neuron receives inputs from all other neurons. If each of N neurons receives inputs from n neurons, then the total number of state diagrams is

$$\mathcal{N}_{\text{th}} \geq \left(2^{n^2/2}\right)^N = 2^{n^2N/2}. \qquad (2)$$

Taking $N = 10^{10}$ as a conservative estimate for the number of neurons in the human neocortex and $n = 10^4$ as an estimate for the number of synaptic inputs to each neuron, implies that the number of possible brains is about $10^{10^{17}}$.

ALWYN SCOTT

See also **Attractor neural network; Cellular automata; Game of life; McCulloch–Pitts network; Neural network models**

Further Reading

Birkhoff, G. & MacLane, S. 1953. *A Survey of Modern Algebra*, New York: Macmillan
Elsasser, W.M. 1998. *Reflections on a Theory of Biology: Holism in Biology*, Baltimore: Johns Hopkins University Press (first published in 1987)
Fowler, M. 2004. *UML Distilled: A Brief Guide to the Standard Object Modeling*, Boston: Addison-Wesley
Gardner, M. 1970. Mathematical games. *Scientific American*, 223 (October): 120–123
McCulloch, W.S. & Pitts, W.H. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5: 115–133
Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer
Yajima, S., Ibaraki, T. & Kawano, I. 1968. On autonomous logic nets of threshold computers. *Transactions on IEEE Computers*, 17: 385–391

STATIONARY SOLUTIONS

See **Discrete self-trapping system; Equilibria**

STEADY STATE

See **Equipartition of energy**

STEREOSCOPIC VISION AND BINOCULAR RIVALRY

Most vertebrate predators, including humans, have their two eyes positioned in the front of the head

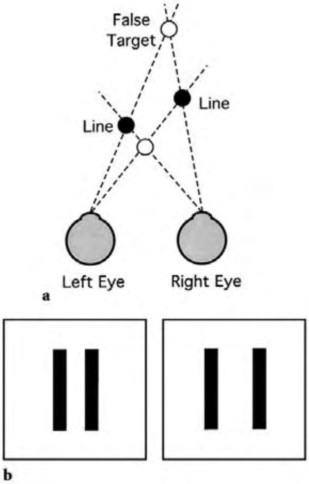

Figure 1. (a) The geometry of stereopsis. (b) The simplest two line stereogram.

looking forward. In contrast, most prey species, such as rabbits, have eyes positioned on the sides of the head. This confers on prey almost panoramic vision, thus enabling them to see predators creeping up behind them. In contrast, predators receive two slightly different views of the scene in front of them due to the horizontal offset between the two eyes (approximately 6.7 cm in humans), which confers depth information. Stereoscopic vision may be defined as the perception of depth that arises when the brain compares the images falling on the two retinas. Indeed, there are neurons in the visual cortex that receive stimulation from both eyes and consequently are sensitive to the depth of lines and contours in the world.

The geometry of stereopsis is illustrated in Figure 1a, which provides a horizontal view of two eyes viewing a scene containing two vertical rods at different depths relative to the eyes. As shown by the dashed sight lines, the retinal images of these two lines will be closer together in one eye's view than in the other. Figure 1b contains a simple two line stereogram, namely a pair of views, one for each eye, with vertical lines closer together in one view than in the other. If the viewer can either cross or diverge their eyes so that each eye fixates a different one of these views, the percept will be one in which one line lies closer to the viewer than the other line. (It should be noted that several percent of humans lack stereoscopic ability, usually due to amblyopia or "lazy eye" during early childhood.) Rather than requiring viewers to cross or diverge their eyes, laboratory research into stereopsis utilizes some version of a stereoscope, which was invented by Charles Wheatstone in 1838. A stereoscope utilizes either prisms or mirrors in front of each eye to optically diverge the views so that each eye views a different visual stimulus. Indeed, stereoscopes and stereoscopic photographs were a Victorian fascination, and many excellent stereoscopes may still be found in antique shops.

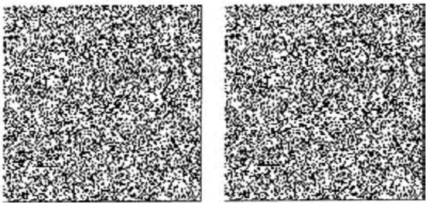

Figure 2. Random dot stereogram.

The two line stereogram in Figure 1b is the simplest possible, as two objects are the minimum for which different depths can be defined. A much more complex example is the random dot stereogram (RDS), which was developed by Bela Julesz (1971) as a tool for the study of stereopsis. To make an RDS, a computer first generates a completely random pattern of black and white dots (little squares), such as that depicted in Figure 2. This will be the stimulus for one eye. An exact copy of this is then made as the basis for the other eye's view. To generate a stimulus for depth, a rectangular (or other shaped) region of dots in the middle of this second random pattern is now shifted horizontally left or right by a distance of several dots. The space vacated by shifting this rectangular region is then filled in with random black and white dots. When each eye views one of the two random dot patterns in an RDS, the brain extracts depth information and produces a percept of a rectangular region of random dots floating in depth in front of (or behind, depending on which eye views which half of the stereogram) a random dot background. Readers who can converge or diverge their eyes while viewing Figure 2 can experience this.

Neural Basis of Stereopsis

Information originating in the two retinas remains segregated until it reaches primary visual cortex at the back of the brain. In the cortex, many individual neurons respond best when they are stimulated through both eyes at once. However, each cortical neuron receives stimulation only from a very small area on each retina, which is known as a "receptive field." If one imagines there being a coordinate system affixed to each retina with the center of the retina (the fovea) being the origin, then each cortical cell receives information from a pair of receptive fields, one from each retina, whose coordinates may be the same or different. If a cortical neuron receives stimulation from receptive fields with the same relative coordinates in each retina, it will be sensitive to lines and edges that lie in the plane on which the eyes are converged. Therefore, such a neuron will be sensitive to objects with zero depth relative to fixation. Other neurons, however, have their receptive fields at different points on the two retinas. For most of these cells, this means that the receptive field in one eye is shifted a short distance horizontally left or right relative to its receptive field in the other eye. Such a difference in receptive field coordinates on the two retinas is termed a horizontal disparity. Neurons with a particular horizontal disparity will respond to objects at a particular depth relative to the fixation plane. For example, if a cortical neurons has its receptive field in the left eye shifted left relative to that in the right eye, its disparity will correspond to lines and edges nearer than fixation. Conversely, a neuron with a left eye receptive field shifted to the right relative to the other eye will be sensitive to disparities corresponding to lines and edges farther away than the plane of fixation. Thus, different cortical neurons, in virtue of the relative placement of their two retinal receptive fields, signal the location of lines and edges in three-dimensional space, including their depth relative to the fixation plane. Although only horizontal disparities are discussed here, vertical disparities are now known to play a role in stereopsis as well, and details can be found in Howard & Rogers (2002).

The RDS demonstrates dramatically the computational problems confronted by cortical neurons in comparing the two retinal images to extract depth information. To make the point, consider a single black dot in one horizontal line of the pattern in one eye's view. The corresponding horizontal line of dots in the other eye's view will contain roughly 50% black dots, any one of which might be the appropriate match for the black dot in question. Thus, the problem has an extremely large number of potential depth solutions, only one of which is actually perceived. In other words, the existence of neurons sensitive to different disparities and, therefore, different depths only represents the initial part of the solution to the problem of extracting depth information from two retinal images.

The types of interactions among disparity-tuned cortical neurons that must be involved in the extraction of accurate depth information may be appreciated by reconsidering the simple two line stereogram in Figure 1. As shown by the dashed lines in Figure 1a, sight lines from the two eyes intersect not only at the locations of the two targets (black circles) but also at two other points in depth (open circles). These are known as "false targets," and neural interactions in the cortex are required to eliminate them. The types of interactions believed to occur include facilitation between neurons tuned to similar disparities in neighboring regions plus inhibition between neurons tuned to very different disparities at adjacent points. Examination of Figure 1a shows that if the neurons tuned to the disparities or depths represented by the black circles were mutually excitatory, whereas both of these inhibited neurons tuned to the much more extreme disparities indicated by the open circles, the neurons representing the black circles would win the competition and suppress the responses to the false targets. Many variations on this theme of facilitation between neurons tuned to similar disparities and inhibition between neurons tuned to radically different disparities have been proposed, and these have been summarized elsewhere (Blake & Wilson, 1991). Such interactions may be conceptualized as implementing an evolutionary constraint that most real world objects are bounded by surfaces that vary smoothly in depth except at their edges.

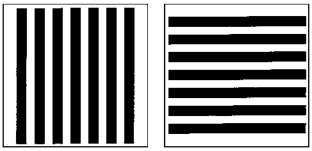

Figure 3. Stereogram illustrating binocular rivalry. Try viewing and perceiving both images simultaneously.

Binocular Rivalry

The preceding discussion assumes that the two eyes are looking ahead and viewing the same scene, albeit from slightly different viewpoints that give rise to disparities. In the laboratory, however, it is possible to generate stereograms that present radically different views to the two eyes. A simple example of this is depicted in Figure 3, where one eye is presented with horizontal lines, while the other is presented with vertical lines. Readers who can converge or diverge their eyes will discover that these lines cannot be fused to produce a depth percept. Rather, one perceives an alternation between horizontal and vertical lines such that the two are almost never seen at the same point at the same time. This perceptual alternation, which occurs at a rate of about one alternation every 2–4 s, is known as binocular rivalry. Rivalry results when the two eyes view images that could not have arisen from a real three-dimensional world in the absence of mirrors or other optical devices. It appears, therefore, that rivalry must reflect neural interactions that evolved to process depth information but that default into a perceptual oscillation when the depth task is impossible.

The neural basis of binocular rivalry requires two components. First, the different groups of rivaling neurons responding to orthogonal stimuli must be strongly and mutually inhibitory. This guarantees that only one set of these neurons can be active at any given place and time. The second ingredient is fatigue or self-adaptation. Self-adaptation, which is well documented among excitatory cortical neurons, causes neurons that are firing rapidly to slowly decrease their response rate over the course of a second or more. As they adapt, their ability to inhibit other neurons weakens. Eventually, these previously inhibited neurons overcome the inhibition and begin to respond, and they then inhibit the adapted neurons so that the percept switches. This neural mechanism has been developed into nonlinear dynamical models that accurately explain many aspects of binocular rivalry (Wilson, 2004).

A further aspect of binocular rivalry is that the reversals are not precisely periodic. Scientists have measured the duration of intervals during which the pattern presented to one eye is visible (e.g., the vertical stripes in Figure 3). The results are well described by a gamma distribution in which both short and long intervals recur in a seemingly random sequence. Recent research using neural network modeling suggests that this gamma distribution results from chaotic dynamics among cortical neurons (see Wilson, 2004).

The final and most important question concerning binocular rivalry is how the inhibitory neural interactions underlying rivalry change their role when stimuli permitting depth perception are viewed. This remains unanswered, but it is at least clear from the discussion of depth perception that inhibition is a necessary neural component. Perhaps, future research will show that the inhibition involved in elimination of false targets during depth perception is related to the inhibition causing perceptual alternations in binocular rivalry.

HUGH R. WILSON

See also **Attractor neural network; Cell assemblies; Gestalt phenomena; Neural network models; Perceptron; Synergetics**

Further Reading

Alais, D. & Blake, R. (editors). 2003. In *Binocular Rivalry and Perceptual Oscillations*, Cambridge, MA: MIT Press

Blake, R. & Wilson, H.R. 1991. Neural models of stereoscopic vision. *TINS*, 14: 445–452

Howard, I. & Rogers, B. 2002. *Seeing in Depth*, Toronto: I. Porteous

Julesz, B. 1971. *Foundations of Cyclopean Perception*, Chicago: University of Chicago Press

Wilson, H.R. 2004. Rivalry and perceptual oscillations: a dynamical synthesis. In *Binocular Rivalry and Perceptual Oscillations*, edited by D. Alais & R. Blake, Cambridge, MA: MIT Press

STIMULATED BRILLOUIN AND RAMAN SCATTERING

See **Rayleigh and Raman scattering and IR absorption**

STOCHASTIC ANALYSIS OF NEURAL SYSTEMS

Oscillatory activity abounds in the central nervous systems of animals and humans (Freeman, 1975; Steriade et al., 1990). Single neurons may act as oscillators by rhythmically generating action potentials or bursts. Central pattern generators (Cohen et al., 1988) are networks of neurons located in the spinal cord and the brain stem, which produce oscillatory activity intrinsically (Sherrington, 1947; von Holst, 1954; Cohen et al., 1988). Central pattern generators are essential for repetitive motor behavior like walking, running, flying, swimming, and chewing (Sherrington, 1947; von Holst, 1954; Cohen et al., 1988). Phase synchronization (PS) between different central pattern generators is the mechanism that controls the different types of locomotion such as walk, trot, or gallop (von Holst, 1954; Cohen et al., 1988).

In the cortex, the well-coordinated action of neural oscillations is indispensable to physiological information processing (Freeman, 1975; Steriade et al., 1990; Haken, 1996). In-phase synchronization might be one aspect of such processes (Singer & Gray, 1995).

In several neurological diseases, brain function is severely perturbed by pathologically enhanced synchronization. Resting tremor in Parkinson's disease (PD) appears to be caused by a cluster of neurons located in the thalamus and the basal ganglia, which fires synchronously at a frequency similar to that of the tremor (3–6 Hz). While under physiological conditions these neurons fire incoherently (Nini et al., 1995); in patients with PD, this cluster acts like a pacemaker and activates cortical areas, which causes the peripheral shaking.

In patients with advanced Parkinson's disease or with essential tremors that do not respond to drug therapy, depth electrodes are chronically implanted in target areas like the thalamic ventralis intermedius nucleus or the subthalamic nucleus (Benabid et al., 1991). Electrical deep brain stimulation (DBS) is performed by administering a permanent high-frequency (>100 Hz) periodic pulse train via the depth electrodes. DBS suppresses the activity of the pacemaker-like cluster, which, in turn, suppresses the peripheral tremor (Benabid et al., 1991). DBS has been developed empirically, mainly based on observations during stereotaxic neurosurgery. The advent of nonlinear dynamics and statistical physics in the field of clinically oriented neuroscience has led to an improvement of data analysis as well as a model-based development of stimulation techniques.

Stochastic Synchronization

Phase synchronization is a type of synchronization where two oscillators adjust their phases, while their amplitudes need not be correlated. Phase

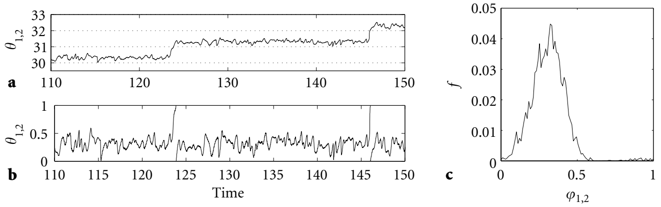

Figure 1. 1 : 2 phase difference $\theta_{1,2}$ (a) and its cyclic version $\varphi_{1,2}$ (b) from Equation (1) between the activity of the right flexor muscle of a patient with Parkinsonian resting tremor and the magnetic field measured above the corresponding left sensorimotor cortex. Corresponding distribution of $\varphi_{1,2}$ (c). (Adapted from Tass et al., 1998).

synchronization is observed in many physical, chemical, and neural systems (von Holst, 1954; Haken, 1996; Tass et al., 1998; Pikovsky et al., 2001; Tass et al., 2003). As neural systems are noisy, it is convenient to use the general concept of PS, which is valid in noisy periodic oscillators as well as in chaotic oscillators (Pikovsky et al., 2001).

To study $n:m$ phase synchronization (PS) we introduce the $n:m$ phase difference $\theta_{n,m}$ and the cyclic $n:m$ phase difference $\varphi_{n,m}$ with

$$\theta_{n,m}(t) = \frac{n\phi_1(t) - m\phi_2(t)}{2\pi},$$
$$\varphi_{n,m}(t) = \frac{n\phi_1(t) - m\phi_2(t)}{2\pi} \mod 1, \qquad (1)$$

where ϕ_1 and ϕ_2 are the phases of the two oscillators. The time course of $\theta_{n,m}$ typically exhibits horizontal periods intersected by 2π jumps (Figure 1a), that vanish when the coupling is sufficiently strong compared to the noise. By $n:m$ PS is meant that the two oscillators maintain a stable relationship of $\varphi_{n,m}$ in a statistical sense (Figure 1b), and thus it is characterized by the appearance of one or more distinct peaks in the distribution of $\varphi_{n,m}$ (Figure 1c). Absence of $n:m$ PS corresponds to a uniform distribution of $\varphi_{n,m}$, whereas perfect $n:m$ PS is related to a Dirac-type distribution of $\varphi_{n,m}$. To quantify the strength of $n:m$ PS, the actual distribution is compared to a uniform distribution with synchronization indices that are based on the Shannon entropy or the Fourier transformation (Tass et al., 1998; Pikovsky et al., 2001). The phase of an oscillatory signal is determined with the Hilbert transform or the wavelet transform (Tass et al., 1998; Pikovsky et al., 2001). Already by the 1930s, PS was studied in coordinated movements using a similar, Poincaré-like approach, but limited phase estimation (von Holst, 1954).

In Parkinsonian patients with resting tremor, a 1 : 2 PS was found between the activity of the shaking muscles and brain areas such as the sensorimotor cortex and premotor areas (Tass et al., 1998) (Figure 1). The dominant frequency of the activity in the brain areas is approximately twice as large as that of the peripheral tremor.

Compared to PS analysis, the cross coherence method cannot distinguish between PS and amplitude modulation as well as between PS and simple linear superposition. Furthermore, cross coherence is not appropriate for the detection of $n:m$ PS with $n/m \neq 1$ (Tass et al., 1998; Pikovsky et al., 2001, 2003).

Stochastic Phase Resetting

In both single neurons and neural populations, the effect of a stimulus depends on the phase of the neural oscillation at which the stimulus is administered. To study phase-dependent effects of pulsing stimuli on neural populations and, in particular, to analyze desynchronizing effects of such stimuli, the concept of phase resetting (Winfree, 1980) has been extended to populations of non-interacting and interacting oscillators in the presence of random forces (Tass, 1999). The single neuron is modeled by a phase oscillator (Kuramoto, 1984). The dynamics of a cluster of N interacting phase oscillators subject to a stimulus S and to random forces F_j is governed by

$$\dot{\psi}_j = \omega_j + \frac{1}{N}\sum_{k=1}^{N}\Gamma(\psi_j - \psi_k)$$
$$+X(t)\,S(\psi_j) + F_j(t), \qquad (2)$$

where ψ_j is the phase of the jth phase oscillator, Γ is a 2π-periodic global coupling, and S is a 2π-periodic, time-independent function modeling the stimulus, where $X(t) = 1$ during stimulation and $X(t) = 0$ without stimulation (Tass, 1999). The random forces $F_j(t)$ are modeled by Gaussian white noise fulfilling $\langle F_j(t)\rangle = 0$, $\langle F_j(t)\ F_k(t')\rangle = D\delta_{jk}\delta(t - t')$ with constant noise amplitude D.

Via the Fokker–Planck equation, one obtains an evolution equation for the average number density $n(\psi, t) = \int_0^{2\pi} \ldots \int_0^{2\pi} d\psi_1 \cdots d\psi_N\ \tilde{n}(\Psi; \psi) f(\Psi; t)$, where $f(\Psi, t)$ is the the probability density with $\Psi = (\psi_1, \ldots, \psi_N)$ and $\tilde{n}(\Psi; \psi) = N^{-1}\sum_{j=1}^{N}\delta(\psi - \psi_j)$. $n(\psi, t)$ tells us how many oscillators most probably have phase ψ at time t. Detailed analytical studies of that equation were performed for $X = 0$ (Kuramoto, 1984; Tass, 1999), whereas the transient stimulation-induced dynamics was analyzed numerically, mainly for the case when all oscillators have the same eigenfrequency (Tass, 1999). Since the single model neuron fires whenever its phase equals 0 (modulo 2π) the average number density of firing neurons is given by the firing density

$$p(t) = n(0, t)\,. \tag{3}$$

It is possible to desynchronize a synchronized cluster of oscillators with a single pulse of the right intensity and duration. To this end, however, the pulse has to hit the cluster in a vulnerable phase range, which corresponds to only a small fraction (5% or even less) of a period of the oscillation. To compound matters, different stimulation parameters have to be used to desynchronize a cluster that is not in its fully synchronized state (Tass, 1999). Hence, complex stimulation techniques have been developed that make it possible to effectively desynchronize a cluster of phase oscillators independently of the cluster's dynamical state at the beginning of the stimulation (Tass, 2002). These methods share one particular feature: the stimulus consists of two qualitatively different stimuli. The first stimulus is stronger and resets (restarts) the cluster, whereas the second, weaker stimulus is a single pulse that is administered after a constant time delay and desynchronizes by hitting the cluster in a vulnerable state. The goal of the reset is to control the dynamics of the cluster by restarting the cluster in a stereotypical way. The reset may be achieved by means of a strong single pulse, a high-frequency pulse train (HFPT) (with a pulse rate 20 times larger than the mean eigenfrequency $\bar{\omega}$) or a low-frequency pulse train (LFPT) (with a pulse rate similar to $\bar{\omega}$) (Tass, 2002). A strong pulse and an HFPT cause a hard (abrupt) reset within less than one period, whereas an LFPT causes a soft (slow) reset, where in the course of an entrainment the influence of the initial dynamic state fades away.

Since the desynchronizing effect of these stimulation techniques does not depend on the initial dynamical state of the neural population, resynchronization can effectively be blocked by administering a desynchronizing stimulus whenever the population tends to resynchronize (Figure 2a). In contrast, high-frequency stimulation not combined with a desynchronizing single pulse causes only a suppression of the firing during stimulation but no desynchronization (Figure 2b). After a simple HFPT, the extent of synchronization is particularly high, and the cluster restarts in a rebound-like manner.

Transient stimulus-induced responses of coupled neural oscillators are a key approach not only for manipulating synchronization processes, but also for studying neural information processing in basic research as well as in diagnosis (Dawson, 1954; Steriade et al., 1990). To improve the signal-to-noise ratio, one typically averages across an ensemble of responses to reveal what is supposed to be the actual response to a stimulus (Dawson, 1954). But noise inherent in the neural dynamics makes coupled neural oscillators react to a stimulus in more than one way, which may be essential for sensory short-term adaptation (Tass, 2003). In this case, averaging causes severe artifacts.

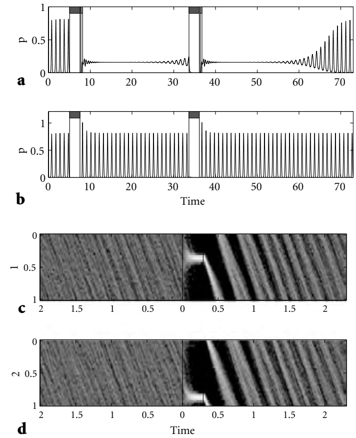

Figure 2. (a) Firing density p from Equation (3) of a neural population stimulated with two subsequent identical composite stimuli consisting of a high-frequency pulse train followed by a single pulse. (b) Two successively administered identical high-frequency pulse trains with parameters as in (a) but without following single pulses. Begin and end of a single pulse are indicated with vertical lines connected by a shaded region ((a) and (b) from Tass, 2001). CT distributions from Equation (5) of ϕ_1 (c) and ϕ_2 (d) show that the post-stimulus responses of each oscillator split into two antiphase ensembles (0 is black and maximal values are white). Begin (at $t = 0$) and end of stimulation are indicated by vertical lines. ((c) and (d) from Tass, 2003).

For illustration we consider Equation (2) for two coupled oscillators ($N = 2$) undergoing an antiphase reset caused by stimuli $S_1(\psi_1) = I\cos(\psi_1)$ and $S_2(\psi_2) = I\cos(\psi_2 + \pi)$. We use the normalized cyclic phases

$$\phi_j(t) = \frac{\psi_j(t)}{2\pi} \bmod 1 \quad (j = 1, 2) \tag{4}$$

and the normalized cyclic $n:m$ phase difference $\varphi_{n,m}$ from Equation (1). To detect whether, in an ensemble of responses, there are epochs during which the phases ϕ_j and/or the phase difference $\varphi_{n,m}$ display, a stereotypical, tightly stimulus-locked time course; the stimuli S_1 and S_2 are simultaneously administered at l random times $\tau_1, \tau_2, \ldots, \tau_l$. Time-dependent cross-trial (CT) distributions of ϕ_j and $\varphi_{n,m}$ are calculated across trials for each time t relative to stimulus onset with

$$\{\phi_j(t+\tau_k)\}_{k=1,\ldots,l}\,,\ \{\varphi_{n,m}(t+\tau_k)\}_{k=1,\ldots,l}\,. \tag{5}$$

The time course of ϕ_j and $\varphi_{n,m}$ is perfectly stimulus-locked at time t if the corresponding CT distributions are Dirac-type distributions, whereas these distributions are uniform if ϕ_j and $\varphi_{n,m}$ are not stimulus-locked at all at time t. The extent of stimulus locking is quantified for each time t with indices characterizing the corresponding CT distributions with Shannon entropy or Fourier transformation (Tass, 1999, 2003). In this way, it is possible to detect stimulus-locked transient synchronization/desynchronization ($\varphi_{n,m}$) as well as transient cross-trial response clustering (ϕ_j). The latter is illustrated in Figures 2c and d, where the pre-stimulus CT distribution of ϕ_j is nearly uniform because of the randomized stimulus application. In all of the $l = 200$ trials, the stimuli S_1 and S_2 reset the oscillators to the phases $\phi_1^{\text{res}} \approx 0.36$ and $\phi_2^{\text{res}} = \phi_1^{\text{res}} + 0.5$, so that Dirac-like CT distributions evolve. Each oscillator reacts in two qualitatively different ways: The ensemble of post-stimulus responses starts at ϕ_j^{res} and splits into two antiphase ensembles, so that two antiphase peaks of the corresponding CT distribution emerge. The mechanism that induces cross-trial response clustering is a stochastic resonance (Tass, 2003).

Coordinated transient responses of neural oscillators such as antiphase cross-trial response clustering typically escape detection with standard methods like cross-trial averaging (Tass, 2003).

Deep Brain Stimulation Techniques

Permanent high-frequency DBS is an unphysiological type of stimulation, which in a number of patients leads to side effects like dysarthria and psychiatric sympotms. Furthermore, the therapeutic effect of DBS may vanish in the course of the treatment, possibly due to adaptation.

Demand-controlled DBS with the stimulation techniques explained in the former section has been suggested as an alternative and milder therapeutic approach (Tass, 2002). Instead of simply suppressing the neural firing of the pacemaker-like cluster, these novel stimulation techniques aim at desynchronizing the pacemaker's pathologically synchronized firing in a demand-controlled way (Tass, 2002). For this, either the depth electrode or an epicortical electrode measures the feedback signal, that is, the local field potential of the pacemaker population or the electrical activity of a cortical area that is strongly coupled to it. A desynchronizing stimulus is administered only whenever the pacemaker-like cluster becomes synchronized (Figure 2a). Repeated administration of desynchronizing stimuli may effectively maintain an uncorrelated firing. Demand-controlled DBS should, in principle, be less aggressive to the brain tissue stimulated and is currently evaluated for clinical use.

PETER A. TASS

See also **Attractor neural network; Cell assemblies; Coupled oscillators; Electroencephalogram at large scales; Electroencephalogram at mesoscopic scales; Stochastic processes; Synchronization**

Further Reading

Benabid, A.L., Pollak, P., Gervason, C., Hoffmann, D., Gao, D.M., Hommel, M., Perret, J.E. & De Rougemont, J. 1991. Long-term suppression of tremor by chronic stimulation of the ventral intermediate thalamic nucleus. *The Lancet*, 337: 403–406

Cohen, A.H., Rossignol, S. & Grillner, S. (editors). 1988. *Neural Control of Rhythmic Movements in Vertebrates*, New York: Wiley

Dawson, G.D. 1954. Summation technique for the detection of small evoked potentials. Electroencephalography and Clinical Neurophysiology, 44: 153–154.

Freeman, W.J. 1975. *Mass Action in the Nervous System*, New York: Academic Press

Haken, H. 1996. *Principles of Brain Functioning. A Synergetic Approach to Brain Activity, Behavior and Cognition*, Berlin and New York: Springer

Kuramoto, Y. 1984. *Chemical Oscillations, Waves, and Turbulence*, Berlin and New York: Springer

Nini, A., Feingold, A., Solvin, H. & Bergman, H. 1995. Neurons in the globus pallidus do not show correlated activity in the normal monkey, but phase-locked oscillations appear in the MPTP model of parkinsonism. *Journal of Neurophysiology*, 74: 1800–1805

Pikovsky, A., Rosenblum, M. & Kurths, J. 2001. *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge and New York: Cambridge University Press

Sherrington, C. 1947. *The Integrative Action of the Nervous System*, 2nd edition, New Haven, CT: Yale University Press

Singer, W. & Gray, C.M. 1995. Visual feature integration and the temporal correlation hypothesis. *Annual Reviews of Neuroscience*, 18: 555–586

Steriade, H., Jones, E.G. & Llinás, R. 1990. *Thalamic Oscillations and Signaling*, New York: Wiley

Tass, P.A. 1999. *Phase Resetting in Medicine and Biology: Stochastic Modelling and Data Analysis*, Berlin: Springer

Tass, P.A. 2002. Effective desynchronization with bipolar double pulse stimulation. *Physical Review* E, 66: 036226

Tass, P.A. 2003. Stochastic phase resetting of two coupled phase oscillators stimulated at different times. *Physical Review* E, 67: 051902

Tass, P.A., Fieseler, T., Dammers, J., Dolan, K., Morosan, P., Majtanik, M., Boers, F., Muren, A., Zilles, K. & Fink, G.R. 2003. Synchronization tomography: a method for three-dimensional localization of phase synchronized neuronal populations in the human brain using magnetoencephalography. *Physical Review Letters*, 90: 088101

Tass, P.A., Rosenblum, M.G., Weule, J., Kurths, J., Pikovsky, A., Volkmann, J., Schnitzler, A. & Freund, H.-J. 1998. Detection of $n : m$ phase locking from noisy data: application to magnetoencephalography. *Physical Review Letters*, 81: 3291–3294

von Holst, E. 1954. Relations between the central nervous system and the peripheral organs. *British Journal of Animal Behavior*, 2: 89–94

Winfree, A.T. 1980. *The Geometry of Biological Time*, Berlin: Springer

STOCHASTIC PROCESSES

Complex processes in physics, chemistry, biology, engineering, and finance often exhibit uncertainties, fluctuations, or noises in their structures as a rule rather than as an exception. It is known that low-dimensional nonlinear dynamical systems can exhibit chaotic phenomena that appear random in nature. Nevertheless, stochastic processes involve an unmanageably large number of variables and their temporal fractal dimensions approach infinity. To interpret, analyze, model, and simulate these noises, several stochastic algorithms developed in the last century yield not only the structure or pattern of these processes, but also the statistical characteristics of the underlying uncertainties that the conventional deterministic models cannot offer.

Analysis of a stochastic process usually starts with the assumption of a Markov property. A Markov process, $X(t)$, is a stochastic process with the property that, given the value of $X(t = t_r)$, the values of $X(t = t_s)$ for $t_s > t_r$ are not influenced by the values of $X(t = t_q)$ for $t_q < t_r$. Depending on the nature of the process and the method of analysis, the random variable, $X(t)$, may assume continuous values or only discrete values. Moreover, the random variable can be analyzed with a continuous or a discrete time scale.

Algorithms have been developed for analyzing two distinct types of noise, internal and external. Internal noise is caused by the fact that the system itself consists of discrete particles and many variables associated with the particles are ignored; it is inherent in the very mechanism by which the process evolves. Small discrete systems governed by a large number of variables often exhibit notable internal fluctuations. The energy state of an electron in a molecule, the spread of an epidemic within a population, the mutation of a gene, and diffusion of a molecule into a medium cannot be predicted precisely due to their inherent complexities; thus, these processes exert internal fluctuations. The discrete state, continuous-time stochastic processes can be analyzed by the master equation in the following general form (Oppenheim et al., 1977; Gardiner, 1985; van Kampen, 2001):

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = \sum_{n'} \{W_{nn'}p_{n'}(t) - W_{n'n}p_n(t)\}, \qquad (1)$$

where $p_n(t)$ denotes the probability of the system to be in state n at time t, and $W_{nn'}$ the transition probability per unit time from state n' to state n. The master equation is a gain-loss equation for the probabilities of the separate state n. The first term is the gain of state n due to transitions from other states n', and the second term is the loss due to transition from n into other states. For a system involving a set of random variables or a random vector, n is a vector and $W_{n'n}$ is a matrix.

When all elements of the transition intensity functions $W_{n'n}$ are either constant or linear functions of n, Equation (1) is considered a linear master equation; otherwise, it is considered nonlinear. Simple linear master equations such as the Poisson equation and birth and death processes can be solved by induction (see, e.g., Karlin & Taylor, 1975). For processes involving large numbers of random variables, full solutions of their joint probability distribution function are usually difficult, if not impossible, and it often suffices to determine only the first and second moments of the resultant probability distribution. These moments can be found by taking the averages of the master equations or by the method of joint probability generating function (Karlin & Taylor, 1975, 1981; Chiang, 1980). For nonlinear master equations, solutions can be approximated by resorting to the system-size expansion, a rational linearization approximation technique based on the power-series expansion (van Kampen, 1976, 2001). The system size, Ω, has been proposed as an expansion parameter because it measures the relative importance of the noises. For a linear system, fluctuations or variances are of the order of $\Omega^{1/2}$ in a collection of Ω entities. As a result, their effect on the macroscopic properties is of the order of $\Omega^{1/2}$. It is expected that the joint probability for the nonlinear system, $p_n(t)$, will have a sharp maximum around the macroscopic value with a noise of width of the order of $\Omega^{1/2}$. Thus, when the higher-order moments are ignored, the approximated joint probability becomes a time-dependent Gaussian distribution. The technique gives rise to the deterministic macroscopic equations and the equations of noises for the master equation; the latter is a set of linear Fokker–Planck equations whose characteristics and solutions will be discussed below.

For the interpretation of its relation with other stochastic algorithms, the master equation has been expressed in terms of a continuous random variable $Y(t)$ with its realization $y(t)$

$$\frac{\partial P(y,t)}{\partial t} = \int \{W(y \mid y')P(y',t) - W(y' \mid y)P(y,t)\}\,\mathrm{d}y'. \qquad (2)$$

Taylor expansion of the above master equation yields the Kramers–Moyal expansion

$$\frac{\partial P(y,t)}{\partial t} = \sum_{v=1}^{\infty} \frac{(-1)^v}{v!}(\frac{\partial}{\partial y})^v \{a_v(y)P(y,t)\}, \qquad (3)$$

where $a_v(y)$ denotes the jump moments:

$$a_v(y) = \int_{-\infty}^{\infty} r^v W(y, r)\, \mathrm{d}r \tag{4}$$

and

$$W(y \mid y') = W(y', r), \quad r = y - y'. \tag{5}$$

A particular important class of continuous Markov processes is that for which $a_v(y) = 0$ for $v \geq 3$ in Equation (3). It is called the Fokker–Planck equation or the diffusion process (Risken, 1996):

$$\frac{\partial P(y,t)}{\partial t} = -\frac{\partial}{\partial y} a_1(y) P(y,t) + \frac{1}{2}\frac{\partial^2}{\partial y^2} a_2(y) P(y,t), \tag{6}$$

where a_1 and $a_2/2$ are called the drift and diffusion coefficients, respectively. Thus, the Fokker–Planck equation is a truncated master equation. If a_1 is a linear function of y and a_2 is constant, Equation (6) is called a linear Fokker–Planck equation, a linearity definition consistent with that of the master equation. Linear Fokker–Planck equations have been used by Rayleigh, Einstein, Smoluchowski, and Fokker for special cases. Taking the moments of Equation (6) yields the equations governing the mean trajectory and the noise around the mean. The solutions of these equations are sufficient in the construction of the macroscopic trajectory and its time-dependent Gaussian noise. An Ornstein–Uhlenbeck process is characterized by the jump coefficients, $a_1(y) = -ky$, and $a_2(y) = D$ where the constants satisfy $k > 0$ and $D \geq 0$ (Gillespie, 1992). For an Ornstein–Uhlenbeck process with the initial condition $P(y, t = 0) = \delta(y - y_0)$, where δ denotes Dirac's delta function, its evolution of the probability distribution function can be written as

$$P(y,t) = \frac{1}{[\pi \frac{D}{k}(1 - \exp[-2kt])]^{0.5}} \times \exp\left\{ -\frac{(y - y_0 \exp[-2kt])^2}{\frac{D}{k}(1 - \exp[-2kt])} \right\}. \tag{7}$$

The evolution of an Ornstein–Uhlenbeck process and its statistical features are illustrated in Figure 1.

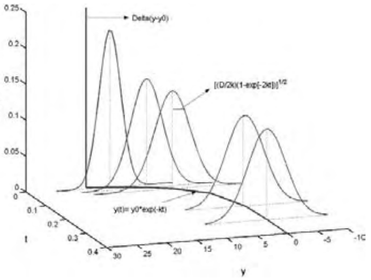

Figure 1. The evolution of an Ornstein–Uhlenbeck process, and its statistical features. The process starts with a given deterministic, initial condition, $P(y, t = 0) = \delta(y - y_0)$, and its fluctuations grow over time and eventually stabilize. The plot is drawn for $y_0 = 20$, $k = 20$, and $D = 1$. The solid line on the (y, t)-plane, $y(t) = y_0 \exp(kt)$, denotes the macroscopic trajectory and its limiting distribution when $t \to \infty$ depicts the Brownian motion.

A discrete state, discrete time stochastic process can be designated by X_i, where $X_i = X(t = t_i)$ and $i = 0, 1, 2, \ldots$. X_i is called a Markov chain if it follows the Markov property. Examples of these processes include a random walk on a line, the number of persons in a waiting line at 1 minute intervals, and the number of catalyst particles of different sizes during attrition. The propagation of a Markov chain can be expressed in the following matrix form (Taylor & Karlin, 1998):

$$p(X_n) = p(X_0) P^n, \tag{8}$$

where $p(X_k)$ is a row vector representing the probability distribution at t_k and P is the one-step transition probability matrix with the following elements:

$$P_{ij} \equiv Pr\{X_{k+1} = j \mid X_k = i\}. \tag{9}$$

Equation (9) suggests that the probability distribution at a given time t_k can be estimated by successive matrix multiplication.

External noises are the fluctuations created in an otherwise deterministic system by the application of an external random force, whose stochastic properties are supposed to be known. The displacement of a mass attached to a spring with a fluctuating external force, an animal population under the influence of a fluctuating environment, the response of an electronic circuit equipped with a noise generator, the random loading of a bridge, and the yield from a chemical reactor with a fluctuating feed rate are examples of external noises. The Langevin approach is widely used for the purpose of finding the effect of noise in macroscopically known systems. The noise is introduced by adding a random term, $L(t)$, to the deterministic equation (Karlin & Taylor, 1981; van Kampen, 2001; Øksendal, 2003)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = A(y) + C(y)L(t), \tag{10}$$

where $A(y)$ denotes the macroscopic characteristics of the system and $C(y)$ characterizes the interaction of the noise and the system. When $A(y)$ is a linear function of y and $C(y)$ a constant, Equation (10) is called a linear Langevin equation. When $A(y)$ is a nonlinear function of y and $C(y)$ a constant, Equation (10) is called a quasilinear Langevin equation. When $C(y)$ is a function of y, Equation (10) is called a nonlinear Langevin equation. These linearity definitions are consistent with those for the master and the Fokker–Planck equations after transformations. This consistency is important because some difficulties emerge during the integration of a Langevin equation that yields the Fokker–Planck equation (van Kampen, 2001). The treatment of the discontinuous stochastic noise, $L(t)$, is the source of the complexity. The added external noise, or the Langevin force, $L(t)$, is irregular and unpredictable, but its averaged properties over an ensemble of similar systems are simple. White noise is a commonly adopted simplification, which satisfies two averaged properties. Its ensemble has zero; that is,

$$E[L(t)] = 0\,. \tag{11}$$

Moreover, the individual noise takes place instantaneously with strength $\sqrt{\Gamma}$, and two successive noises arrive uncorrelated; that is,

$$E[L(t)L(t')] = \Gamma\delta(t - t'). \tag{12}$$

The spectral distribution of white noise, or the Fourier transform of Equation (12), is independent of the frequency ω. If the noise $L(t)$ is not δ correlated, the spectral density depends on the frequency and one uses the term *colored noise*.

Taking the first and second moments of Δy associated with a Langevin equation yields the jump moments of the Fokker–Planck equation. For a linear Langevin equation with $C(y) = 1$, the averaging process give rise to a linear Fokker–Planck equation (see, e.g., van Kampen, 2001)

$$\frac{\partial P(y,t)}{\partial t} = -\frac{\partial}{\partial y}A(y)P(y,t) + \frac{\Gamma}{2}\frac{\partial^2}{\partial y^2}P(y,t), \tag{13}$$

and it represents an Ornstein–Uhlenbeck process. For a quasilinear Langevin equation, the process gives rise to an expression identical to Equation (13). Nevertheless, the nonlinear nature of $A(y)$ suggests its Fokker–Planck equation is also nonlinear. Analytical solutions of the nonlinear Fokker–Planck equation are discussed by Risken (1996) and Grasman and van Herwaarden (1999).

Due to the discontinuous nature of $L(t)$, integrating $C(t)L(t)$ of a nonlinear Langevin equation in the averaging process, however, results in two different interpretations. If the averaged value of $C(y)$ during $(t, t + \Delta t)$ is used, the averaging process yields

$$\begin{aligned}\frac{\partial P(y,t)}{\partial t} &= -\frac{\partial}{\partial y}[A(y)P(y,t)] \\ &\quad + \frac{\Gamma}{2}\frac{\partial}{\partial y}\left[C(y)\frac{\partial}{\partial y}[C(y)P(y,t)]\right]\end{aligned} \tag{14}$$

and the integration procedure is called the Stratonovich calculus. If the value of $C(y)$ before the arrival of the delta peak during $(t, t + \Delta t)$ is used, the averaging process yields

$$\begin{aligned}\frac{\partial P(y,t)}{\partial t} &= -\frac{\partial}{\partial y}[A(y)P(y,t)] \\ &\quad + \frac{\Gamma}{2}\frac{\partial^2}{\partial y^2}[C^2(y)P(y,t)]\end{aligned} \tag{15}$$

and the integration procedure is called the Ito calculus. For practical processes experiencing external noise, the noise $L(t)$ is never infinitely sharp, and it lasts for a finite period of time; this consideration suggests that the Stratonovich interpretation is appropriate (van Kampen, 2001).

WEI-YIN CHEN

See also **Averaging methods; Deterministic walks in random environments; Ergodic theory; Fokker–Planck equation; Lévy flights; Martingales; Monte Carlo methods; Nonlinearity, definition of; Random walks; Stochastic analyses of neural systems; Time series analysis**

Further Reading

Chiang, C.L. 1980. *An Introduction to Stochastic Processes and Their Applications*, Huntington, NY: Krieger

Gardiner, C.W. 1985. *Handbook of Stochastic Methods for Physics, Chemistry, and Natural Sciences*, 2nd edition, Berlin: Springer

Gillespie, D.T. 1992. *Markov Processes*, San Diego: Academic Press

Grasman, J. & van Herwaarden, O.A. 1999. *Asymptotic Methods for the Fokker–Planck Equation and the Exit Problem in Applications*, Berlin: Springer

Karlin, S. & Taylor, H.M. 1975. *A First Course in Stochastic Processes*, 2nd edition, New York: Academic Press

Karlin, S. & Taylor, H.M. 1981. *A Second Course in Stochastic Processes*, New York: Academic Press

Oppenheim, I., Shuler, K.E. & Weiss, G.H. 1977. *Stochastic Processes in Chemical Physics: The Master Equation*, Cambridge, MA: MIT Press

Øksendal, B. 2003. *Stochastic Differential Equations: An Introduction with Applications*, 6th edition, Berlin: Springer

Risken, H. 1996. *The Fokker–Planck Equation: Methods of Solution and Applications*, 2nd edition, Berlin: Springer

Taylor, H.M. & Karlin, S. 1998. *An Introduction to Stochastic Modeling*, 3rd edition, San Diego: Academic Press

van Kampen, N.G. 1976. The expansion of the master equation. *Advances in Chemical Physics*, 34: 245–309

van Kampen, N.G. 2001. *Stochastic Processes in Physics and Chemistry*, revised edition, Amsterdam: Elsevier

STOKES SCATTERING

See **Rayleigh and Raman scattering and IR absorption**

STRANGE ATTRACTORS

See **Chaotic dynamics**

STRING THEORY

String theories (Green et al., 1987; Davies & Brown, 1988; Polchinski, 1998; Gauntlett, 1998; Kaku, 1999) were developed in the 1980s and 1990s in the hope of finding a unified description of all forces in nature. The aim is to combine electroweak, strong, and gravitational interactions into one consistent theory. The basic idea is that subatomic particles are not point-like but one-dimensional extended objects: strings. These vibrate (similarly to a violin string) in various modes and can move in space-time. The various particle states that we observe in nature should then all be explained as suitable excitations of suitable strings. Interaction processes between particles are described by splitting and merging of strings. Due to the finite size of strings, many of the divergences that occur for point-like particles can be avoided. The size of the string is very small, typically of the order of a few Planck lengths (1.6×10^{-35} m), so that on large scales string-like particles look like point particles.

Different types of boundary conditions are possible; there are open strings and closed strings. The simplest example is a bosonic string. In suitable coordinates, the equation of motion is simply a wave equation of the form

$$\frac{\partial^2 X^\mu(\sigma,\tau)}{\partial\tau^2} = C^2 \frac{\partial^2 X^\mu(\sigma,\tau)}{\partial\sigma^2}. \qquad (1)$$

Here X^μ is the coordinate of the string-like particle in d-dimensional space-time ($\mu = 0, 1, \ldots, d-1$), σ and τ denote an internal position and time variable parametrizing the string, and C is a constant.

The general solution can be written as a mode expansion of the form

$$X^\mu(\sigma,\tau) = x^\mu + \dot{x}^\mu\tau + \mathrm{i}\sqrt{2\alpha'}\sum_{n\neq 0}\frac{1}{n}\alpha_n^\mu \mathrm{e}^{-\mathrm{i}\frac{n\pi C\tau}{L}}\cos\frac{n\pi\sigma}{L}. \qquad (2)$$

Here x^μ and $\dot{x}^\mu$ are constants independent of σ and τ that essentially represent the position and velocity components of a constantly moving point particle, whereas the sum over n describes the oscillating part of the dynamics. The string parameter $\sqrt{\alpha'}$ describes the approximate scale where string effects become relevant, and L is the length of the string. Unlike a violin string, this string is not tied down at the end points but it can move freely through space-time while it oscillates.

So far this is a classical string. First quantization is done by letting the coefficients α_n^μ of the above normal mode expansion satisfy suitable commutator relations. Second, quantization of strings, up to now, is still an active area of research with many open questions (Kaku, 1999).

The quantization of bosonic strings yields unpleasant surprises, so-called ghosts. These are quantum states with negative norm. One wants to avoid these unphysical states, which is only possible if the dimension of space-time is $d = 26$. Consequently, it has been necessary to assume that the 22 extra dimensions are "compactified," that is, curled up on small circles that are so small that we do not notice them in daily life.

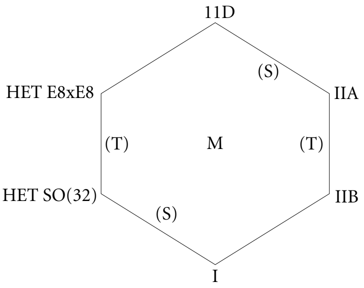

Figure 1. A schematic diagram of M-theory with its known string theory limits and the 11-dimensional theory. Some of the dualities are also shown.

But bosonic string theories have another problem. The lowest excitation mode (the ground state) is a so-called "tachyon," a state for which the mass squared is negative. To avoid this unphysical (unstable) ground state, one proceeds to supersymmetric string theories, in short "superstring theories." In these theories, for each boson there is a corresponding fermionic partner and vice versa. Technically, one introduces anti-commuting coordinates in addition to commuting ones. For supersymmetric theories, ghost states occur as well, but can now be avoided if the number of space-time dimensions is $d = 10$. So to describe low-energy physics, again six extra dimensions have to be compactified. Five different superstring theories have been found, which are denoted as type I, IIA, IIB, HET $E_8 \times E_8$, and HET SO(32). They differ in the degree of supersymmetry, the underlying gauge symmetry (if any), and whether the strings are closed or open.

Nowadays the five superstring theories are regarded as special limit cases of a more fundamental theory, called M-theory. This contains the five superstring theories and an 11-dimensional theory as special cases. These special cases are marked as the edge points of the diagram in Figure 1. Unfortunately, it is not known how M-theory should be formulated in full generality and how it should be second quantized. It is not even clear what the fundamental objects are that this theory describes. Much work needs to be done.

Some of the string theories contained in M-theory are connected by the so-called duality transformations. S-duality essentially means that a string theory with a small coupling constant α describes the same physics as another string theory with a large coupling given by $1/\alpha$. T-duality means invariance of the physics if a compactified dimension of radius R is replaced by another one of radius $1/R$.

In recent years, much research emphasis has concentrated on higher-dimensional objects contained in string and M-theory, so-called D-branes. Open strings can end and start on D-branes. Remember that in electromagnetism there are point charges, which are the sources of the electromagnetic field. D-branes are in a sense higher-dimensional generalizations of these point charges.

As already mentioned, to proceed to our four-dimensional world, six of the dimensions of superstring theory (or 7 of those of M-theory) have to be compactified. The way this should be done is unclear. One needs to compactify on special types of compact manifolds, so-called Calabi–Yau manifolds, to preserve what is generally believed to be the right amount of supersymmetry. The compactification step is clearly necessary for string theory to make contact with the real world, but almost all relevant low energy predictions depend on the numerous possibilities of how to compactify. This severely undermines the predictive power of string theories.

In principle, the ultimate theory should predict everything we want to know about particle physics, such as the values of the coupling constants of the four different forces or the masses of all fermions and bosons. In string theories, the numerical values of gauge couplings (e.g. the fine structure constant α_{el}, which is about $\frac{1}{137}$ at low energies) can be related to vacuum expectations of a scalar field contained in string theories, the dilaton field. But this vacuum expectation depends in an unknown way on second quantization effects and has not been predicted so far. As 't Hooft puts it in his book ('t Hooft, 1997), string theory, at least in its present stage, has similarities with a very uncomplete piece of furniture: "Imagine that I give you a chair while explaining that the legs are still missing, and that the seat, back, and armrest will perhaps be delivered soon; whatever I did give you, can I still call it a chair?"

While typical equations for (free) strings, such as Equation (1), are linear, a recent proposal is to amend ordinary strings (evolving in a regular way) by nonlinear versions of strings, so-called chaotic strings (Beck, 2002). These evolve in a deterministic chaotic way. In this approach, each ordinary string is shadowed by a chaotic string, which yields the "noise" for second quantization via the so-called stochastic quantization method. Mathematically, chaotic strings consist of one-dimensionally coupled Tchebyscheff maps, a very nonlinear and strongly self-interacting theory, which describes a kind of "turbulent quantum state" on a small (quantum gravity) scale. It turns out that the vacuum energy of chaotic strings is minimized for observed standard model parameters; that is, in this extended approach to second quantization, concrete predictions for vacuum expectations of dilaton-like fields and hence on masses and coupling constants can be given.

CHRISTIAN BECK

See also **General relativity; Particles and antiparticles; Quantum field theory**

Further Reading

Beck, C. 2002. *Spatio-temporal Chaos and Vacuum Fluctuations of Quantized Fields*, Singapore: World Scientific

Davies, P.C.W. & Brown, J. 1988. *Superstrings: A Theory of Everything?* Cambridge and New York: Cambridge University Press

Gauntlett, J.P. 1998. M-theory: strings, duality and branes. *Contemporary Physics*, 39: 317–328

Green, M.B., Schwarz, J.H. & Witten, E. 1987. *Superstring Theory*, 2 vols, Cambridge and New York: Cambridge University Press

't Hooft, G. 1997. *In Search of the Ultimate Building Blocks*, Cambridge and New York: Cambridge University Press

Kaku, M. 1999. *Introduction to Superstrings and M-theory*, Berlin and New York: Springer

Polchinski, J. 1998. *String Theory*, 2 vols, Cambridge and New York: Cambridge University Press

STRONG COLLAPSE

See **Development of singularities**

STRUCTURAL COMPLEXITY

By *structural complexity* we mean the study of morphologically complex patterns associated with many physical entities of dynamical systems, using methods based on algebraic, geometric, and topological information (Ricca, 2001). In this respect, structural complexity offers an alternative and complementary route to study complex systems, based on morphological characteristics rather than the governing rules generating complexity. Structural complexity arises from spontaneous growth and self-organization of interacting components present in the bulk of physical systems, such as in turbulence, in biological and chemical reactions, in growth phenomena, across scales of space, time, and organizational complexity (Badii & Politi, 1997). Much of our understanding comes from observations of natural phenomena, from simulations and computational visualizations. Mathematical methods are used to extract and synthesize the relevant information, and to explore possible relationships between morphological complexity, energy localization, and other physical properties.

Dynamical systems are often characterized by the emergence and interaction of *coherent structures* (Nicolis & Prigogine, 1998; Scott, 2003). These latter are produced through filamentation mechanisms and space-time localization of characteristic physical properties. Examples of coherent filamentary structures may be as disparate as DNA macromolecules, synthetic polymer chains, nerve fibers, actin filaments, vortex filaments, magnetic flux tubes, or massive cosmic

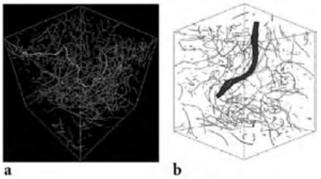

Figure 1. Computational field visualization of a pattern of vortex filaments produced by turbulent flow in a numerical box: (a) low-pressure line and (b) iso-pressure surface, obtained by numerical extraction techniques (Courtesy of S. Kida, Theoretical Division, National Center for Fusion Research, Toki, Japan).

strings. Dislocations and discontinuities in mesoscopic and solid-state physics, fronts and cell membranes in chemistry, fluid dynamics, and biology provide other examples of complex structuring. The formation and continuous re-arrangement of coherent structures are due to local interactions and recombination, governed by specific dynamics given by nonlinear differential equations $\dot{\boldsymbol{x}} = F(\mathbf{x}, \lambda)$, with a nonlinear function F of the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, which depends on an external control parameter λ. At certain critical values, this parameter causes phase transitions and pattern changes (Levin, 2002). *Pattern recognition* techniques aim at providing mathematical methods for diagnostics of the morphological complexity of the emerging pattern, regardless of the governing fundamental physical mechanism that is responsible for complexity.

One of the great benefits of modern capability to perform complex three-dimensional simulations is the possibility to effectively analyze and visualize huge sets of three-dimensional data. Direct volume rendering, iso-surface extraction, integral convolution, and many other computational techniques allow sophisticated multi-field visualizations of complex physical patterns (Johnson et al., 2001). Progress in *visiometrics*, based on the exploitation of such techniques for an accurate identification and representation of complex three-dimensional structures, provides powerful tools and an almost inexhaustible source of information for analysis of structural complexity (see Figure 1; also see **Visiometrics**).

Physical as well as computational domains are decomposed into *tropicity* domains, defined by the characterizing properties of nulls, tubes, sheets, and blobs. These are reference sets defined by the reference physical field distribution, that allow direct measurement of the degree of *tubeness*, *sheetness*, and *bulkiness* of the coherent set by analyzing the aspect ratio of the distribution.

Algebraic, geometric, and topological measures are used to quantify structural complexity. In case of high filamentation, the original network of filaments is reduced (by appropriate threshold filtering) to a tangle $\mathcal{T}$ of space curves (representing the filament axes), hence reducing the problem to the analysis of the mutual positioning of the system of curves in space. By using methods based on crossing number information, we can evaluate the *average crossing number* $\bar{C}$ of the tangle that provides a fundamental algebraic measure of structural complexity. This quantity is defined by the formula

$$\bar{C} = \sum_{\mathcal{L}_i, \mathcal{L}_j \in \mathcal{T}} \bar{C}_{ij}, \tag{1}$$

where

$$\bar{C}_{ij} = \left\langle \sum_{r \in \mathcal{L}_i \sqcap \mathcal{L}_j} |\varepsilon_r| \right\rangle, \qquad \varepsilon_r = \pm 1 \tag{2}$$

denotes the un-signed crossings between the curve $\mathcal{L}_i$ and $\mathcal{L}_j$ of the tangle $\mathcal{T}$, and summation is made over all the crossings resulting from $\mathcal{L}_i \sqcap \mathcal{L}_j$; the average number of crossings is then evaluated by projecting the tangle onto projection planes; here the angular brackets denote averaging over all directions of projections (in the simple case of one oriented space curve, Figure 2 shows an application of this analysis). Geometric measures are based on integral measures of curvature and torsion of the reference curves and surfaces; for tangles of curves, these include *total curvature*, *average twist*, and *writhing number* which give information on the amount of coiling and entwinement of the corresponding filaments. Topological information comes from different types of measures: number of *nulls* (zero-value field singularities), *linking numbers* for open or closed curves, winding numbers and invariants associated with handle-body decomposition of reference sets are all calculable quantities. Further information may come from analysis of cell-decomposition of the network. In particular situations, when a prevailing type of

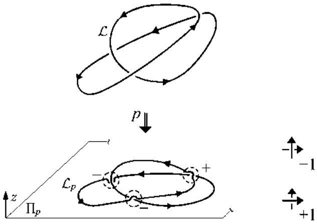

Figure 2. Example of analysis of an oriented space curve $\mathcal{L}$, based on crossing number information of the projected curve $\mathcal{L}_p$: we can calculate the total number of un-signed crossings (3, in the example shown in the figure) and, according to the sign convention shown on the right-hand side of the diagram, the algebraic sum of the ± signs ($-1 - 1 + 1 = -1$, in the example shown). By averaging these two quantities over all directions p of projections, we obtain estimates of, respectively, the average crossing number and the writhing number of $\mathcal{L}$.

geometry emerges, ad hoc analysis is used; for example, spiral spectrum analysis can be employed for scroll-wave reactions in Belousov–Zhabotinsky-type chemical systems and braid theory can be applied to study dynamics of particulates, fluid mixing, trailing vortices, astrophysical plasma loops, and tangle diagrams for polymers. Other information may come from study of dynamical systems, where we have information, for example, on Lyapunov exponents, topological entropy, multifractal properties, and topological scaling (see, for example, de Gennes, 1979; Jensen, 1998), whereas artificial intelligence provides us tools based on algorithmic complexity for neural-type networks. Measures of structural complexity are implemented computationally as time-dependent variables that change with the evolution of the physical system.

Collected information on structural complexity is analyzed against physical information. Since the evolution of any physical system is driven by variational principles that take account of energy and entropy redistributions, information on structural complexity may disclose some useful information on the physics and dynamics of the process. It is known that geometric properties strongly influence the dynamics: curvature forces (such as bending force in elastic rods or surface tension in liquid films) cause telephone cords to coil-up and foam bubbles to coalesce. We, therefore, expect relationships between geometry, dynamics, and ultimately, energy. Possible relationships between energy, entropy, and structural complexity in terms of algebraic information have, however, a more subtle rationale: network restructuring is due to local interactions and recombinations that are induced by acting potentials. These are associated with localized coherent structures, acting either on the strands of neighboring filaments (as in the case of reconnection of vortex tubes) or on the surface elements of sheet discontinuities (as in the formation of elastic films). In this context, local interaction of neighboring structures, resulting in the emergence of apparent crossings (possibly weighted by some distribution function) is a measure of the localization of internal energy. Hence, the resulting growth rate of the average crossing number provides an estimate of energy and entropy variations in the system. In absence of dissipation, topology is conserved; however, even in fully dissipative systems, topological information provides indications to detect preferred paths of energy depletion. In general, if χ denotes a measure (or a family of measures) of structural complexity, and E the energy of the system, we may expect *energy-complexity relations* of type $E = G(\chi)$, where G is a nonlinear function, where the nonlinearity is prescribed by the growth rate exponents of the physical process (see, e.g., Barenghi et al., 2001). Algebraic, geometric, and topological measures of structural complexity are, therefore, not only useful for computational implementation of diagnostics and pattern recognition, but they can also provide new tools to investigate localization and transfer of energy in complex physical systems.

RENZO L. RICCA

See also **Algorithmic complexity; Filamentation; Knot theory; Pattern formation; Topological defects; Vortex dynamics of fluids**

Further Reading

Badii, R. & Politi, A. 1997. *Complexity*, Cambridge and New York: Cambridge University Press

Barenghi, C., Ricca, R.L. & Samuels, D.C. 2001. How tangled is a tangle? *Physica* D, 157: 197–206

de Gennes, P.G. 1979. *Scaling Concepts in Polymer Physics*, Ithaca, New York: Cornell University Press

Jensen, M.H. 1998. *Self-Organized Criticality*, Cambridge and New York: Cambridge University Press

Johnson, C.R., Livnat, Y., Zhukov, L., Hart, D. & Kindlmann, G. 2001. Computational field visualization. In *Mathematics Unlimited – 2001 and Beyond*, edited by B. Engquist & W. Schmid, New York: Springer

Levin, S.A. 2002. Complex adaptive systems: exploring the known, the unknown and the unknowable. *Bulletin of the American Mathematical Society*, 40(1): 3–19

Nicolis, G. & Prigogine, I. 1998. *Exploring Complexity*, New York: W.H. Freeman

Ricca, R.L. (editor) 2001. *An Introduction to the Geometry and Topology of Fluid Flows*, Dordrecht: Kluwer

Scott, A. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, Oxford and New York: Oxford University Press

SUBHARMONIC GENERATION

See **Harmonic generation**

SUPERCONDUCTING QUANTUM INTERFERENCE DEVICE

The Superconducting QUantum Interference Device (SQUID) is the most sensitive detector for magnetic flux, Φ. It relies on the fundamental concept of flux quantization in a multiply connected superconductor, for example, a ring (torus). Flux quantization is a clear demonstration of the quantum mechanical properties of a superconductor in which the conduction electrons form (Cooper) pairs and condense in the ground state: the so-called "Bose condensation." Below the transition temperature, T_c, all the bosonic pairs can be described by a single macroscopic many-body wave function

$$\Psi(\mathbf{r}) = [n_s(\mathbf{r})]^{1/2} e^{i\theta(\mathbf{r})}, \qquad (1)$$

where n_s is the density of Cooper pairs. The phase $\theta(\mathbf{r})$ is a scalar function of the position. For small fields one may ignore the spatial dependence of n_s.

A direct consequence of the condensation is that for $T < T_c$ a superconductor exhibits zero electrical resistance and that magnetic fields are expelled from its interior. However, this diamagnetism (Meissner effect)

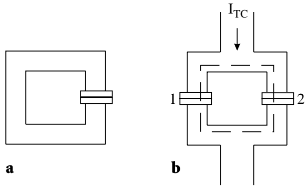

Figure 1. (a) The RF-SQUID and (b) the DC-SQUID. Dashed line is an integration contour deep inside the superconductor.

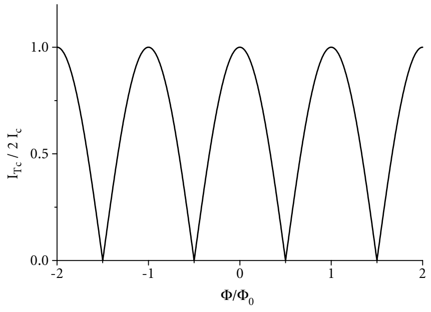

Figure 2. Magnetic interference pattern of symmetric DC-SQUID. Total critical current versus applied magnetic flux, $\beta_L = 0$.

is not perfect as currents may penetrate into a thin surface layer (about 50 nm) called the London penetration depth (λ_L).

Early superconductors were metals (Hg, Al, Sn, Pb, Nb) or alloys (NbTi, Nb_3Sn) with T_c below ≈ 20 K. Since 1986, a new class of ceramic materials has been discovered with transition temperatures in the range 50–135 K, meaning that instead of requiring cooling by liquid helium (4 K), they only need cooling with liquid nitrogen (77 K) or by traditional cryocoolers. The terms low-T_c (LTS) and high-T_c (HTS) are used for these two types of superconducting materials.

The gradient of the phase θ of the macroscopic wave function describing all Cooper pairs depends on the pair-current density, $\boldsymbol{J}_\mathbf{s}$, and the magnetic vector potential, $\boldsymbol{A}$, as

$$\hbar \nabla \theta = 2e(\Lambda \vec{\boldsymbol{J}}_\mathbf{s} + \vec{\boldsymbol{A}}), \tag{2}$$

where the pre-factor, $\Lambda = 2m/n_s(2e)^2$, reflects the double electron charge $2e$ and mass $2m$ of the Cooper pair. If Equation (2) is integrated along a closed contour deep inside the superconducting ring where $J_s = 0$, the fact that the wave function has to be single-valued with phase (modulo 2π) leads to quantization of the magnetic flux contained in the ring so that

$$\Phi_{\text{encl}} = n\Phi_0\,, \tag{3}$$

where $n = 0, 1, 2, 3, \ldots$ is an integer and $\Phi_0 = h/(2e) = 2.07 \times 10^{-15}$ Wb is the "flux quantum." Φ_0 is very small, corresponding to the flux generated by the Earth's magnetic field ($\approx$40 μT) in a circular loop with radius $\approx$4 μm.

A SQUID is a superconducting ring including either one or two Josephson junctions as shown in Figure 1. Named after their operational modes, these are called the RF-SQUID and the DC-SQUID. In both configurations, the response to an external magnetic flux is periodic in the flux quantum Φ_0. For simplicity, we consider only the two-junction DC-SQUID shown in Figure 1b.

Upon integrating Equation (2) along the closed contour, one may show that the difference between the phases, φ_1 and φ_2, of the macroscopic wave functions of the two Josephson junctions varies periodically with the applied flux Φ as

$$\varphi_1 - \varphi_2 = 2\pi\Phi/\Phi_0, \tag{4}$$

giving a total critical (super-)current, I_{Tc}, of the SQUID

$$I_{Tc}(\Phi) = I_{c1} \sin\varphi_1 + I_{c2} \sin(\varphi_1 - 2\pi\Phi/\Phi_0)\,. \tag{5}$$

For two identical ($I_c = I_{c1} = I_{c2}$) junctions, $I_{Tc}(\Phi)$ varies from $2I_c$ to zero with a period of Φ_0 (see Figure 2). The first zero is at $\Phi = \Phi_0/2$. We say that the depth of modulation, $M = I_{Tc,\max}/I_{Tc,\min} = I_{Tc}/2I_c$, is 100%. M is reduced if the two junctions are different.

The variation in $I_{Tc}(\Phi)$ reflects the fact that the applied flux generates a circulating shielding supercurrent, $I_{\text{circ}}(\Phi)$ in the SQUID ring. Given the loop inductance, L, it generates a self-flux $\Phi_{\text{self}} = LI_{\text{circ}}$, which counteracts the applied flux. The effective flux in the ring is only $\Phi = \Phi_{\text{appl}} - \Phi_{\text{self}}$ leading to a reduced depth of modulation. A practical measure for a symmetric SQUID is the inductance parameter $\beta_L = \pi LI_c/\Phi_0$. For $\beta_L = 0$, $M = 100\%$ (as above) while for $\beta_L \to \infty$, $M \to 0$. Practical DC-SQUIDs are optimized with $M \approx 50\%$ corresponding to $\beta_L \approx 1$.

In order to avoid complications caused by hysteresis and flux contained in finite size junctions, one usually keeps junctions small and nonhysteretic (i.e., with relatively large damping and small capacitance). DC-SQUIDs with such junctions can be stably biased in the voltage state using a fixed DC current $I_b > I_{Tc}$. When an external flux is applied, the I-V curve is shifted following $I_{Tc}(\Phi)$, given in Equation (5), and a sensitive measurement of the oscillating voltage, $V(\Phi)|_{I_b}$, across the SQUID enables determination of Φ to a small fraction of Φ_0. Generally, the flux resolution is limited by thermal noise.

The use of a SQUID as a detector is complicated by the highly nonlinear dependence of the voltage response on the applied flux (see Figure 2). This may be overcome by operating the SQUID in the so-called "flux locked mode," where the SQUID output voltage via a feedback loop is returned as a current in a coil placed near the SQUID ring. By proper design, the coil produces a flux that exactly compensates the applied flux in the ring. The SQUID now acts as a null detector where the feedback current is a linear function of Φ over a very large range of magnetic flux.

Its impressive flux resolution does not imply that the SQUID can resolve such small magnetic fields, because the area of the SQUID ring is normally very small. An ingenious way to effectively increase the loop area without increasing L (keeping $\beta_L \leq 1$) is to use a superconducting flux transformer. This is a closed superconducting loop consisting of a series connection of a pickup coil placed in the ambient magnetic field and an *input coil* inductively coupled to the SQUID ring. When an external magnetic field is applied to the pickup coils, a shielding current is set up in the loop so that the total flux in the flux transformer is conserved (flux quantization). The current generates a flux in the input coil, which in turn couples the flux into the SQUID. To increase sensitivity, the input coil often is a multiturn thin-film coil (Washer type) placed on top of the SQUID ring. Usually, the input coil and the SQUID are enclosed in a magnetic shield to suppress external noise.

A well-designed LTS DC-SQUID magnetometer with proper pickup coil has an unsurpassed magnetic field sensitivity of $\approx 10\,\mathrm{fT}$ ($10^{-14}\,\mathrm{T}$), about one billionth (10^{-9}) of the Earth's magnetic field. LTS RF-SQUIDs and HTS DC-SQUIDs reach 100 fT. Most field measurements, therefore, are limited by magnetic noise from the environment and not from the magnetometer itself. Magnetic noise may be dramatically reduced by shaping the pickup coil as two matched counter-wound coils, so that a uniform magnetic field (e.g., from a distant source) generates zero net flux in the flux transformer. Effectively it measures the gradient of the magnetic field working as a first-order gradiometer with baseline equal to the distance between two counter-wound coils. Higher order gradiometers with three (or more) balanced coils have also been constructed. Modern multi-SQUID magnetometers use first-order gradiometers with advanced digital noise reduction and background field nulling.

Because of their high sensitivity, SQUIDs are now used as transducers for virtually any response that can be converted to magnetic flux, including magnetic fields, field gradients, currents, voltages, electromechanical positioning, and movements. A prominent biomedical application is three-dimensional magnetic source imaging, where an array of several hundred SQUIDs placed in a helmet fitted around the head can localize magnetic sources in the human brain to about a millimeter. This is important for precise location before or during brain surgery and for diagnosis of mental disorders. Many hospitals already use commercial SQUID instruments to record the very weak fields generated by nerve pulses in the brain (MEG, magnetoencephalogram) and heart, even the fetal heart beat (MCG, magnetocardiogram), as supplements or improved alternatives to the established electrical counterparts, EEG (electroencephalogram) and ECG (electrocardiogram) using electrodes placed on the skin.

Dedicated SQUID magnetometers are also used in geomagnetic surveying as well as for nondestructive testing (NDT), where defects in conducting samples can de detected by monitoring magnetic fields of induced eddy currents. For detection of the Earth's magnetic field, signal frequencies are typically a few millihertz, but SQUIDs may also be used for ultra-low noise microwave amplifiers as well as picosecond electronics and terahertz oscillators based on propagation at high speed of single quanta of magnetic flux in SQUID arrays.

JESPER MYGIND

See also **Josephson junctions; Superconductivity**

Further Reading

Duzer, T. Van & Turner, C.W. 1998. *Principles of Superconductive Devices and Circuits*, Upper Saddle River, NJ: Prentice-Hall, pp. 256–283

Kadin, A.M. 1999. *Introduction to Superconducting Circuits*, New York: Wiley, pp. 207–225

Orlando, T.P. & Delin, K.A. 1990. *Foundations of Applied Superconductivity*, New York: Addison-Wesley, pp. 244–251 and 410–420

SUPERCONDUCTIVITY

The phenomenon of superconductivity was discovered in 1911 by the Dutch scientist Heike Kamerlingh Onnes. In connection with work on the liquefaction of helium (normal boiling point at 4.2 K), he measured the resistance of mercury and noticed that the resistance dropped gradually with temperature as was expected. However, near 4.2 K, the resistance apparently dropped to a very small value consistent with zero. This, of course, must have been a very dramatic discovery, changing completely the general picture of the properties of solid-state matter. Today, the simple and direct consequence of zero resistance—that a large current can run indefinitely in a closed wire without energy input—intrigues many people. Even though zero resistance may be counter-intuitive, it is not forbidden by the laws of nature, and it does not imply a "perpetuum mobile." The second important property of a superconductor is that it is a perfect diamagnet. This property also has a feature that at first may seem quite unnatural. If a magnet is lowered towards a superconductor, the magnetic field of the magnet cannot penetrate into the superconductor. The magnetic field lines become compressed between the magnet and the superconductor, and eventually the magnet will levitate. This exclusion of the magnetic field is not a consequence of the

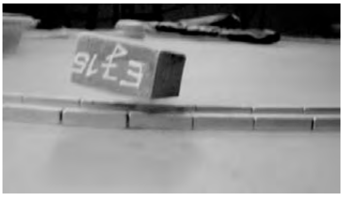

Figure 1. A levitated HTS superconductor.

induction law in connection with zero resistance of the superconductor, but an independent phenomenon. The phenomenon was discovered in 1933 by Walter Meissner and Robert Ochsenfeld long after the discovery of zero resistance. It is called the Meissner effect.

Superconductivity as discussed above occurs in many ordinary metallic elements from the periodic table. Examples are mercury, lead, tin, and niobium. The transition temperatures T_c (i.e., the highest temperature where superconductivity occur) are typically below 10 K. If the external magnetic field is too high, it will quench the superconductivity completely. This field is called the critical field H_c, and it is typically of order some hundreds of Gauss, too low for practical applications in electric machinery. The type of superconductor described above is called a type I superconductor.

Type II superconductors were discovered much later—in the 1950s. They are typically alloys of metals such as NbTi or Nb_3Sn and have transition temperatures typically in the 10–20 K range up to 23 K. They can carry much higher currents and can withstand much higher magnetic fields, thus, being much more useful for applications. A major difference from type I superconductors is the way they behave in a magnetic field. The Meissner effect as described above exists up to a so-called lower critical field H_{c_1}. At H_{c_1}, the magnetic field in the form of magnetic flux lines starts to penetrate the superconductor. As the external field is increased up to the upper critical field H_{c_2}, more and more flux lines penetrate the superconductor until at H_{c_2}, the flux lines completely fill the superconductor, which then becomes normally conducting.

In 1986, a new type of superconductor was found with a complicated structure. These superconductors are called High Temperature Superconductors (HTS, see below) and may, for lack of better understanding, be described as extreme type II. Examples are $YBa_2Cu_3O_7$ (known as YBCO) and $Bi_2Sr_2Ca_1Cu_2O_8$ (BSCCO), both of which have transition temperatures above the boiling point of liquid nitrogen at 77 K. These materials are quite useful for applications since liquid nitrogen—a cheap industrial product—may be used as a coolant. These important superconductors will be discussed separately below.

Figure 1 shows a levitated HTS superconductor. Since it is an extreme type II superconductor, magnetic flux lines penetrate the superconductor. These flux lines are "pinned" on imperfections of the crystal lattice, such as dislocations, and provide a stability not seen in type I superconductors. Thus, the HTS superconductors may even "levitate" below the magnet—hanging on the flux lines from the magnet. By symmetry a magnet can alsobe levitated over a superconductor.

Ginzburg–Landau Theory

Superconductivity is a complicated quantum mechanical phenomenon. The particles giving rise to the supercurrents may be described by a quantum mechanical wave function $\psi(r) = |\psi(r)|\exp(\mathrm{i}\Theta)$. In a normal metal, the phases of all the electrons are different. In the superconducting state, all carriers have the same quantum mechanical phase of the wave function. This is referred to as macroscopic phase coherence. Thus, a simply connected macroscopic superconductor is characterized by a single-phase angle. The difference between a normal metal and a superconductor has some similarity to the difference between the coherent light of a laser, where all the atomic oscillators oscillate in phase and the "white" light of an ordinary bulb, where all the exited atoms are incoherent.

The interpretation of superconductivity as a macroscopic quantum phenomena and the first useful theory of superconductivity were due to Fritz and Heinz London in 1935.

A real breakthrough for applications is the still very often-used phenomenological theory published by the Russian scientists Vitaly Ginzburg and Lev Landau in 1950. Ginzburg and Landau introduced a position-dependent order parameter with a conceptual similarity to a quantum mechanical wave function. The Ginzburg–Landau theory is a masterpiece of intuition. It postulates that as the temperature is lowered and T_c is crossed from above, a (complex) order parameter ($\psi(r)$) for an ordered, superconducting phase appears—increasing from the value zero at T_c. Gibbs free energy is then expressed as a series expansion in terms of this order parameter ($\psi(r)$). Finding the equilibrium value of the Gibbs free energy leads to two coupled differential equations that connect the order parameter and the electromagnetic fields.

Variational methods are used and the Ginzburg–Landau equation is derived. It may be written

$$\begin{aligned}&\alpha\psi(\boldsymbol{r}) + \beta|\psi(\boldsymbol{r})|^2\psi(\boldsymbol{r})\\&\quad +1/2m((h/2\pi\mathrm{i})\nabla - q\boldsymbol{A})2\psi(\boldsymbol{r}) = 0\end{aligned}$$

Here α and β are coefficients to be determined by the particular system, $\boldsymbol{A}$ is the vector potential and m is a particle mass. The last term on the left side has the appearance of a kinetic energy, and we note the similarity to the Schrödinger equation (*See* **Nonlinear Schrödinger equations**).

As mentioned above, the nonlinear Ginzburg–Landau equation can be applied to other systems with an order parameter characterizing a phase transition. Examples are magnetic materials, liquid helium, and liquid crystals. Because of the series expansion, which is only good when the expansion parameter is small, the theory is formally only valid near the transition temperature.

BCS Theory

Although rather successful, the above theories are phenomenological in nature. They do not specify the nature of the physical system or why superconductivity occurs. For this, a full microscopic quantum mechanical theory must be derived. This was done by John Bardeen, Leon Cooper, and Robert Schrieffer in 1957, and their theory carries the name BCS theory. The derivation of the BCS theory is too complicated to be described here, but we note that in most practical cases a full understanding is not needed. Close to T_c the BCS theory approaches the phenomenological Ginzburg–Landau theory as it should. Central to BCS theory is the notion that there are two types of currents. Normal currents are carried by normal electrons, and supercurrents are carried by so-called "superelectrons." Just below T_c, there are only a few superelectrons and many normal electrons. At $T = 0$, all charge carriers are superelectrons. These "superelectrons" are bound pairs of normal electrons called Cooper pairs. Accordingly they have two electron masses and two electron charges. Intuitively, one would think that two electrons could not form a pair since Coulomb repulsion would rip them apart. However, the pairing mechanism has its origin in the electron-phonon interaction, which is responsible for overwhelming the Coulomb repulsion and providing the attractive potential binding the two electrons together.

The binding energy of the pair, Δ, is of order a few meV and is related to the energy scale of the phonons involved in the pairing process. Naturally in the BCS theory, there is a direct proportionality between the transition temperature T_c and the pair binding energy Δ. In fact, this way it has been suggested that T_cs cannot exceed about 30 K. Until recently the highest T_c observed experimentally was 23 K in general agreement. This is not true any more as will be described in the next section.

High T_c Superconductors

In 1986, K. Alex Muller and J. Georg Bednorz from the IBM Zurich research laboratory discovered superconductivity in a new class of materials involving typically the so-called rare earths. Within a few months after the discovery was known, researchers found a T_c of 92 K in the material $YBa_2Cu_3O_7$ and related compounds. YBCO is typical of the so-called HTS although now many other families are known, including the widely used $Bi_2Sr_2Ca_2Cu_3O_8$ compounds. To become superconducting, these materials must undergo a heat treatment at about 1000 K in an oxygen atmosphere. After this baking, the materials appear as a ceramic and are thus hard and brittle—which causes some engineering problems for making the superconductors into wires and coils. Now, many more HTSs are known and the highest T_c found so far is about 150 K. New HTS materials are found regularly, and it cannot be predicted what the limit to higher T_c will be. Maybe we will one day have room-temperature superconductors.

The theory for HTSs is not known, in spite of the efforts of many theoreticians. Obviously, the BCS theory is not directly applicable since it predicts a maximum T_c of about 30 K. An HTS typically has a layered structure with copper-oxide planes separated by isolating atoms at atomic distances, and it is speculated that the superconductivity somehow has its explanation in this structure.

Particularly following the discovery of HTS, the potential for useful applications of superconductivity is quite large. A significant benefit of the high transition temperatures is that liquid nitrogen at 77 K can be used as the coolant. Applications are both small scale (electronics, sensors, and communications) and large scale (power cables, motors, transformers, and large magnets.)

Starting with small scale, we note that Josephson junctions (*See* **Josephson junctions**) are the main component in superconducting electronics. The general advantage over semiconductors is higher speed and lower dissipation. Computers can be constructed that are much faster and much smaller than conventional computers. Sensors, in general, can be made much more sensitive with superconductor technology. This is because of the quantum mechanical nature of superconductivity and, of course, the low temperatures where thermal noise is reduced. A so-called super conducting quantum interference device (SQUID) is a commercial superconducting magnetic field sensor that is so sensitive that it can measure the magnetic field from nerve impulses in the brain. It can also be used to detect internal cracks in the aluminum-alloy plates of aircraft wings. Superconducting photon detectors can detect single photons from space. High-frequency microwave detectors can detect signals down to the quantum limit and in some spectacular cases even lower. Filters for mobile telephone ground stations can be made much sharper than conventional ones, implying more telephones, improved quality, and lower prices.

Large-scale applications include large magnets, levitated trains, power cables, motors, energy storage, and fault current limiters. Large magnets are, for example, used in the Magnetic Resonance Imaging

(MRI) systems at hospitals where a patient's whole body is introduced into the magnet. This is a very large market, and it is only feasible with superconducting magnets. A levitated train with superconducting magnets has already been built in Japan (Maglev). Superconducting (HTS) power cables have already been built in several countries. In Denmark, a 35 m section has already been installed in the Copenhagen grid. In general, the benefits of the superconducting systems is that energy savings can be obtained because there is no—or at least very low—dissipation. Thus superconducting machinery can be made much smaller than similar conventional machinery, which is important in many cases in the automotive industry. Energy can be stored in large superconducting coils or friction-free flywheels. A fault current limiter is an important protection device for the utility grid that can be made using only superconducting technology.

In the coming generation, superconductivity may play the revolutionary role that semiconductors did in the previous generation.

NIELS FALSIG PEDERSEN

See also **Bose–Einstein condensation; Josephson junctions; Nonlinear Schrödinger equations; Superconducting quantum interference device**

Further Reading

Barone, A. & Paterno, G. 1982. *Physics and Applications of the Josephson Effect*, New York: Wiley

Seeber, B. (editor). 1998. *Handbook of Applied Superconductivity*, 2 vols, Bristol and Philadelphia: Institute of Physics Publishing

Sheahen, T.B. 1994. *Introduction to High Temperature Superconductivity*, New York: Plenum Press

Tinkham, M. 1996. *Introduction to Superconductivity*, 2nd edition, New York: McGraw-Hill

SUPERFLUIDITY

Superfluidity is the state of a quantum liquid in which the liquid has the ability to flow through fine slits and capillaries without friction or production of a pressure gradient. The superfluidity of liquid helium He^4 (called "He-II") at temperatures below $T_\lambda = 2.17$ K was discovered by Peter Kapitza in 1938. The transformation of normal liquid helium into the superfluid state is a phase transition of the second order. Nevertheless, the superfluid He-II cannot be considered as a liquid without viscosity. Experiments on rotational vibrations of a disk suspended in He-II show that damping is observed at temperatures below ($T < T_\lambda$), decreasing as the temperature is lowered.

The first qualitative explanation of superfluidity phenomena was proposed by Lazlo Tisza (1938), who introduced a two-fluid model of superfluidity. Lev Landau (1941) later formulated a so-called two-fluid hydrodynamics of He-II. According to this explanation, He-II consists of two interpenetrating fluids, a normal fluid with viscosity and a superfluid fluid with zero viscosity. It is the superfluid that flows without friction through the finest of channels, while the finite viscosity of the normal component is responsible for the observed damping of the oscillating disk. Each component is assigned a density ρ_n and ρ_s, whose sum is the total density $\rho = \rho_n + \rho_s$, and each is assigned its velocity field v_n and v_s.

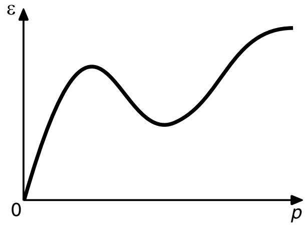

Figure 1. Dependance of the energy of elementary excitations on momentum.

Landau's theory is based on the idea of a possible description of physical properties of a weakly excited macroscopic system at low temperatures by means of elementary excitations or "quasiparticles." The elementary excitations have energy $\varepsilon(p)$ and are regarded as particles of momentum p. The normal component represents a gas of elementary quasiparticles with a dispersion curve as sketched in Figure 1. There are two types of the quasiparticles at low temperatures: sound quanta or phonons of long wavelength (with dispersion relation $\varepsilon = sp$, where s is the sound velocity) and excitations of short wavelength called "rotons" (near the minimum of the curve in Figure 1 their energy is $\varepsilon = \Delta + \frac{1}{2}(p - p_0)^2/\mu$), where μ is the effective mass of the roton). Interactions of quasiparticles with each other and with the vessel's walls cause the viscosity of the normal component. At zero temperature, the normal density $\rho_n = 0$ and the system is in the ground state.

The remaining part of He-II is the superfluid component. It has zero viscosity and can move without friction through narrow slits and capillaries if the energy spectrum of the quasiparticles satisfies the following condition: $\varepsilon(p)/p > 0$. At $T = 0$, the superfluid density $\rho_s = \rho$. The quasiparticle concentration grows when the temperature increases and ρ_s decreases turning into zero at $T = T_\lambda$. According to the Landau theory, the liquid loses the superfluidity property if the velocity of its flux exceeds a critical velocity,

$$v_c = \min(\varepsilon(p)/p); \qquad (1)$$

this condition is called the "Landau criterion." Real critical velocities are much less than (1) because the superfluidity is broken by vortices arising in the superfluid liquid at even smaller velocities. Quantum vortices have been observed during experiments on the motion of the He-II in a rotating cylindrical container.

In a quantum Bose gas and He-II at low temperatures, a large number of particles (a finite proportion of all particles) turns into one coherent state called a "condensate," which can have a stationary velocity different from zero. Such a phenomenon is called "Bose–Einstein condensation," and superfluidity is often connected with it. When the superfluidity component is a correlated many-particle quantum state; however, it may not be connected directly with Bose–Einstein condensation.

The energy spectrum of excitations in a weakly non-ideal Bose gas with repulsive interactions between particles was first derived by Nikolai Bogolubov (1947) assuming existence of the condensate

$$\varepsilon(p) = [(p^2/m)\rho_0\nu(p) + p^4/(4m^2)]^{1/2}, \qquad (2)$$

where ρ_0 is the density of particles in the state $\boldsymbol{p}=0$, $\nu(p)$ is the Fourier component of the potential of interparticle interaction ($\nu(0)>0$), and m is the mass of the particle. Equation (2) determines a spectrum of small vibrations of the condensate in the non-ideal Bose gas. This dispersion relation is of the phonon type at small p, where the Landau criterion is satisfied. Therefore, even the weakly non-ideal Bose gas possesses the superfluid property.

A macroscopic theory of the superfluid dynamics of the weakly non-ideal Bose gas was developed by Gross (1961) and Pitaevskii (1961). In this theory, a condensate wave function Ψ is described by the nonlinear Schrödinger equation ($\hbar=1$):

$$\mathrm{i}\frac{\partial\Psi}{\partial t} = -\frac{1}{2m}\Delta\Psi + \nu(0)\left(|\Psi|^2-\rho\right)\Psi. \qquad (3)$$

A solution of the linearized Equation (3) has a form of harmonic vibrations whose frequency spectrum is determined by Equation (2). Thus, Equation (3) in a total form possesses a solution in the form of the straight-line vortex

$$\Psi = \sqrt{n}f\left(\frac{r}{r_0}\right)\mathrm{e}^{\mathrm{i}\phi}, \quad r_0^2 = \frac{h^2}{2m\nu(0)n}, \qquad (4)$$

where r and ϕ are the distance from the vortex axis and the polar angle with respect to it, and the function $f(r)$ has a shape sketched in Figure 2.

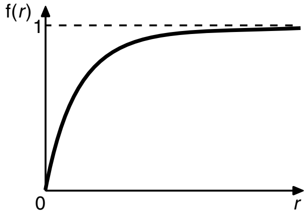

Figure 2. Distribution of the condensate near the vortex.

Superfluidity is also possible in quantum Fermi liquids both in neutral (e.g., liquid He^3) and charged (conducting electrons in metals) liquids. Then a generation of pairs of coupled fermions is necessary (pairs of He^3 atoms or electrons with an integer spin) to create bosons. Such pairing is a result of any attraction (either van der Waals interactions between the atoms or interaction of the electrons with vibrations of the crystal lattice). Being bosons, these "Cooper pairs" can form coherent superfluid condensates. The attractive forces between He^3 atoms are very small; therefore, conditions for generation of the Cooper pairs of quasiparticles and origin of the superfluidity of He^3 appear only at superlow temperatures of the order of 10^{-6}K. Interestingly, there are two superfluid phases (A and B) of the He^3 at superlow temperatures.

The emergence of electron Cooper pairs in a metal corresponds to the appearance of superconductivity. Concepts from superfluidity theory have influenced the theory of superconductivity in metals and also the development of models of the superfluidity of the atomic nucleus. As the atomic nucleus presents a Fermi system with strong interparticle interaction, both the pair correlation effects and also manifestations of some superfluid properties of Bose systems should take place.

Superfluid liquids possess several unusual properties. Besides ordinary sound (i.e., vibrations of the liquid's density), a so-called "second sound" can propagate in such a liquid. The second sound represents sound propagation in the quasiparticle gas (vibrations of the quasiparticle density and, therefore, vibrations of the temperature). Also, a thin film of superfluid He-II forms on a solid wall, rising easily up to a large height. This film can equalize liquid levels of the helium in vessels having a common wall.

Arnold Kosevich

See also **Bose–Einstein condensation; Nonlinear Schrödinger equations; Superconductivity**

Further Reading

Bogolubov, N.N. 1947. On the theory of superfluidity. *Izvestia Akademii Nauk SSSR*, Seriya Fiziko, 11: 77

Donnelly, R.J. 1967. *Experimental Superfluidity*, Chicago: University of Chicago Press

Feynman, R.P. 1957. Appliction of quantum mechanics to liquid helium. In *Progress in Low Temperature Physics*, edited by C.J. Gorter, vol. I: pp. 17–53

Gross, E.P. 1961. Structure of a quantized vortex in boson systems. *Nuovo Cimento*, 20: 454

Kapitza, P.L. 1938. Viscosity of liquid helium at temperatures below the lambda point. *Doklady Akademii Nauk SSSR*, 18: 21 (in Russian); translated in *Nature*, 141: 74

Kapitza, P.L. 1977. *Experiment. Theoriya. Practika*, Moscow: Nauka (in Russian)

Khalatnikov, I.M. 1971. *Teoriya Sverkhprovodimosti*, Moscow: Nauka (in Russian)

Landau, L.D. 1941. Theory of superfluidity of helium II. *Zhurnal Experimental'noj i Teoreticheskoj Fiziki*, 11: 592 (in Russian)

Landau, L.D. 1947. On the theory of superfluidity in helium II. *Journal of Physics of the USSR*, 11: 91

Lifshitz, E.M. & Pitajevski, L.P. 1978. *Statisticheskaya Fizika*, p.2 Moscow: Nauka (in Russian)

Pitaevskii, L.P. 1961. Vortex lines in an imperfect Bose gas. *Zhurnal Experimental'noj i Teoreticheskoj Fiziki*, 40: 646 (in Russian); translation in *Soviet Physics JETP*, 13: 451

Tisza, L. 1938. Transport phenomena in helium II. *Nature*, 141: 913

Tisza, L. 1940. On the theory of quantum liquids. Application to liquid helium. *Journal de Physic et Radium*, 1: 350 (in French)

SUPERLATTICES

Superlattices (SLs) are solid periodic structures usually made artificially and having a period much larger than the interatomic distance. The SL is called one-dimensional (1-d) if it is periodic along one axis. In such a case, the SL has the form of alternating plane-parallel layers of two or several materials differing in their physical properties (dielectric, elastic, magnetic, etc., depending on the implementation of interest) (see Figure 1). Modern techniques of crystal growth allow the creation of artificial crystals with atomic smooth boundaries between layers, which can be arranged in any order. A two-dimensional SL is periodic in two directions and homogeneous in the third, and a three-dimensional SL is periodic along three axes. There are many types of SLs, including semiconducting and superconducting SLs, dielectric or optical SLs (called photonic crystals), and elastic or acoustic SLs (called phononic crystals).

The main physical property of an SL important for applications is a band structure of the energy or frequency spectrum caused by the macroscopic periodicity analogous to the structure of electron energy or phonon frequency spectra in monocrystals.

The physical origin of the band structure can be understood by analyzing a one-dimensional SL constructed from two types of materials, where the alternating layers have thicknesses d_1 and d_2 (see Figure 1). The SL period equals $d = d_1 + d_2$, and the electromagnetic or elastic fields inside each layer are described by wave equations with phase velocities c_1 and c_2. Thus, eigenfunctions of the wave equations are characterized by a quasiwave number k, and the energy or frequency of any excitation is a periodic function of k with the period $2\pi/d$, so values of k can be restricted to the one-dimensional Brillouin zone $-\pi/d < k < \pi/d$.

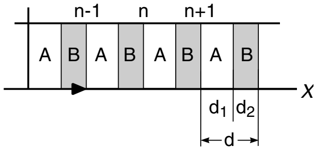

Figure 1. 1D SL made of substances of two types A and B.

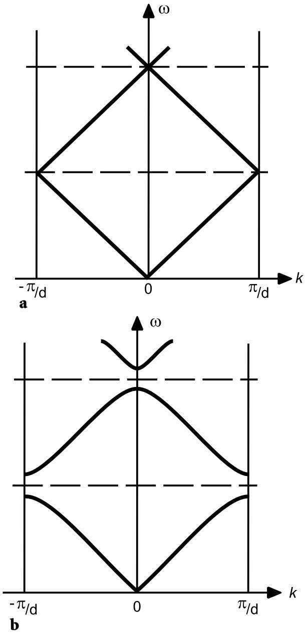

Figure 2. Dispersion relations for the SL (a) in the limiting case $c_1 = c_2$ and (b) in the case $|c_1 - c_2| \ll c_1 \approx c_2$.

If the velocities c_1 and c_2 are close, one can assume $c_1 = c_2 = c$ in the first approximation, so the dispersion relation for the wave solution is that for a homogeneous material. In the case of an elastic medium or a dielectric transparent for electromagnetic waves, $\omega(k) = ck$ where ω is the frequency and c is either the sound or light velocity (the straight lines in Figure 2a). Because the function $\omega(k)$ repeats itself outside the Brillouin zone, the lines fold back into the zone when they reach edges. A small perturbation provided by the small difference between c_1 and c_2 takes off the degenerations at the points $k = 0, \pm\pi/d$ and as a result, a small frequency gap appears between the upper and lower branches of the lines (Figure 2b). This is called a forbidden gap as no excitation mode of the SL can exist in the gap. As the difference $c_1 - c_2$ is increased, the forbidden gap widens considerably.

Dispersion equations for the frequencies of arbitrary one-dimensional SLs made of two dielectric or elastic materials were obtained by the Soviet scientist Sergei Rytov (1955, 1956) using the standard boundary conditions at the layer boundaries

$$\cos kd = \cos k_1 d_1 \cos k_2 d_2 - \frac{1}{2}\left(\frac{k_1}{k_2} + \frac{k_2}{k_1}\right) \sin k_1 d_1 \sin k_2 d_2, \quad (1)$$

where $k_1 = \omega / c_1$ and $k_2 = \omega / c_2$. When the right-hand side of Equation (1) has values from -1 to $+1$, the equation has real roots, which give the values of ω.

If one of the alternating layers is opaque for electromagnetic waves, the corresponding wave vector is imaginary: $k_2 = \mathrm{i}\kappa_2$ and Equation (1) transforms to

$$\cos kd = \cos k_1 d_1 \cosh \kappa_2 d_2 - \frac{1}{2}\left(\frac{k_1}{\kappa_2} - \frac{\kappa_2}{k_1}\right) \sin k_1 d_1 \sinh \kappa_2 d_2. \quad (2)$$

There are two limiting cases of Equations (1) and (2) when the roots of interest can be easily obtained either graphically or analytically. The first one concerns (2) and assumes $\kappa_2 = \text{constant} \to \infty$ and $d_2 \to 0$ at $\kappa_2 d_2 \to 0$ under the condition $\kappa_2^2 d_2 = \text{constant} = q$ (Kronig & Penny, 1931). The simplified version of Equation (2) has the form

$$\cos kd = \cos kd + M \frac{\sin k_1 d}{k_1 d}, \quad M = \frac{1}{2} qd, \quad (3)$$

which can be solved graphically. In Figure 3a, $z = k_1 d$ and the roots of the equation run over values within intervals marked off on the abscissa axis. These intervals become wider and the forbidden gaps become more narrow when the frequency increases.

The second limiting case concerns Equation (1) and corresponds to the limit $d_2 \to 0$, $c_2 \to 0$ under the conditions $d_2 / c_2 = 0$ and $d_2 / c_2^2 = P = \text{constant}$, leading to (Kosevich, 2001)

$$\cos kd = \cos k_1 d - Q\, k_1 d \sin k_1 d,$$
$$Q = c_1^2 P / (2d). \quad (4)$$

Equation (4) can also be solved graphically (see Figure 3b), and a set of permitted intervals contracting with increasing frequency can be obtained. However, there is an analytical expression for frequencies at large numbers m of these intervals ($m^2 Q \gg 1$)

$$\omega = \frac{m\pi c}{d} + \frac{2\Omega}{m} \begin{cases} \sin^2(kd/2), & m = 2p; \\ \cos^2(kd/2), & m = 2p + 1, \end{cases} \quad (5)$$

where $\Omega = c/(\pi Q d)$ and p is an integer.

The frequencies of forbidden bands correspond to a solution of the wave equation inside n-th layer of the type $\exp \pm \kappa n d$ (when $k = \mathrm{i}\kappa$) or $(-1)^n \exp(\pm \kappa n d)$ (when $k = \mathrm{i}\kappa + \pi$), which decrease (grow) with increasing n. The frequency dependence of the parameter κ for solutions of the first (second) type is found from Equation (1) when its right-hand side is larger then $+1$ (then $\cos kd = \cosh \kappa d$ in the left-hand side) or less than -1 (then $\cos kd = -\cosh \kappa d$). Such solutions have a physical meaning only in the x semiaxis under the condition that a solution vanishing at infinity and corresponding to certain boundary conditions at the origin is selected. The necessity of using exponentially decreasing solutions arises in describing localized states of fields under consideration in the vicinity of a local SL defect. Of course, a one-dimensional SL can have only localized states near a given plane. In certain circumstances, a localized mode can exist at the face of the SL (called a surface state).

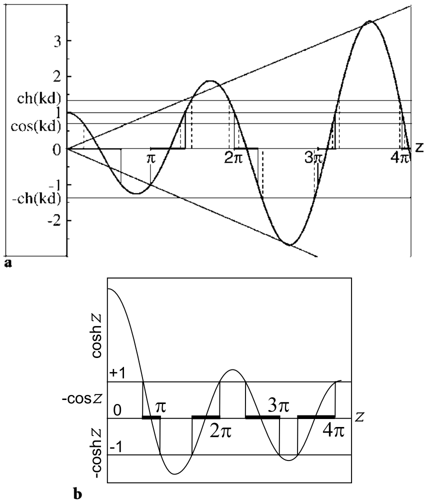

Figure 3. Graphical solutions of (a) Equation (3) and (b) Equation (4). Eigenfrequency bands are shown in heavy lines on the z-axis ($z = k_1 d$).

In the general case, the spectrum of an SL is complicated and contains a system of a great number of both eigenfrequency bands and gaps corresponding to forbidden frequencies of eigenmodes. No extended states are allowed in the gap, but there exist evanescent modes, decaying exponentially with increasing the distance from the defect. In a three-dimensional SL, they are localized states.

Dependence of the physical parameters of materials (the velocities c_1 and c_2, or the parameters Q and M in the above equations) on the field strength leads to nonlinear effects in SLs. In optical SLs, a source of nonlinear effects is the dependence of the refraction coefficient on the electric field of a light wave. In the simplest case, which corresponds to the dispersion relations (5) in the linear approximation, an electromagnetic wave Ψ_n in the nth layer of the SL is governed by the discrete NLS equations for $m=2p$

$$\begin{aligned} \mathrm{i}\frac{\partial \Psi_n}{\partial t} &= \frac{m\pi c}{d}\Psi_n - U_0|\Psi_n|^2\Psi_n \\ &\quad -\frac{2\Omega}{m}(2\Psi_n - \Psi_{n+1} - \Psi_{n-1}) \qquad (6) \end{aligned}$$

and for $m=2p+1$

$$\begin{aligned} \mathrm{i}\frac{\partial \Psi_n}{\partial t} &= \frac{m\pi c}{d}\Psi_n - U_0|\Psi_n|^2\Psi_n \\ &\quad +\frac{2\Omega}{m}(2\Psi_n + \Psi_{n+1} + \Psi_{n-1}), \qquad (7) \end{aligned}$$

where U_0 is a parameter of the nonlinearity. Actually, Equations (6) and (7) are equations for the envelope curve of SL vibrations taken in discrete points. Such equations can be used for describing nonlinear effects and particularly dynamics of solitons in the optical SL.

Arnold Kosevich

See also **Discrete nonlinear Schrödinger equations; Nonlinear optics; Photonic crystals; Solitons**

Further Reading

Hasegawa, A. 2003. *Optical Solitons in Fibers*, 3rd edition, Berlin and New York: Springer

Jin, B.Y. & Ketterson, J.B. 1989. Artificial metallic superlattices. *Advances in Physics*, 38: 189–366

Joannopoulos, J.D., Meade, R.D. & Winn, J.N. 1995. *Photonic Crystals*, Princeton, NJ: Princeton University Press

Kivshar, Y. & Agrawal, G. 2003. *Optical Solitons: From Fibers to Photonic Crystals*, Amsterdam and New York: Academic Press

Kosevich, A.M. 2001. On a simple model of the photonic or phononic crystal. *JETP Letters*, 74: 559–563

Kronig, R. de L. & Penny, W.G. 1931. Quantum mechanics of electrons in crystal lattices. *Proceedings of the Royal Society, London, Series* A, 130: 499–513

Rytov, S.M. 1955. Electromagnetic properties of laminated media. *Zhurnal Experimental'noj i Teoreticheskoj Fiziki*, 29: 605–616; *Soviet Physics, JETP*, (1956), 2: 66

Rytov, S.M. 1956. Properties of laminated medium. *Soviet Physics-Acoustics*, 2: 68

Yeh, P. 1988. *Optical Waves in Layered Media*, New York: Wiley

SUPERSYMMETRY

See **String theory**

SURFACE WAVES

The surface of a fluid such as water lends itself most easily to the demonstration of wave phenomena. At the same time, waves on the surface of shallow water have played a very important role in the development of nonlinear science. The entries in this encyclopedia on water waves and solitons cover this aspect. Following narrower definitions, the term *surface waves* is often used for standing waves that form on the surface of a vessel filled with a liquid and moved up and down periodically in the direction normal to the liquid's surface. When adding the adjective *acoustic* (surface acoustic waves, SAW), one thinks of waves propagating on the surface of a solid elastic medium. These two types of surface waves will be discussed here. The former are called Faraday waves, and the most common species of the latter is named after Lord Rayleigh (John William Strutt).

Faraday Waves

In 1831, Michael Faraday reported on a series of experiments with various liquids that he made oscillate on glass plates and in a basin attached to a vibrating strip of wood (or lath) (Faraday, 1831). On the surface of the vibrating liquids, he observed the formation of standing wave patterns. With increasing amplitude of vibration, he found these standing waves to evolve into a regular square pattern (Figure 1). Faraday also noticed that the frequency of the surface waves was half of that of the driving force.

A proper mathematical explanation for the instability of the flat fluid surface, when being accelerated periodically, was given more than a century later, in 1954, by T.B. Benjamin and F. Ursell. They considered a sinusoidal acceleration with frequency ω. Let the initially planar surface be parallel to the x, y-plane and define the two-dimensional position vector $\boldsymbol{R}=(x, y)$. Benjamin and Ursell expanded the local excursion $\zeta(\boldsymbol{R}, t)$ of the surface from the planar initial position in real eigenfunctions S_m, $m=1, 2, \ldots$, of the two-dimensional Laplace operator, satisfying Neumann boundary conditions at the boundary of the fluid surface with vanishing normal derivative,

$$\zeta(\boldsymbol{R}, t) = \sum_m a_m(t) S_m(\boldsymbol{R}). \qquad (1)$$

Linearizing the equations of hydrodynamics, they arrived at the Mathieu equation for the amplitudes a_m,

$$\frac{\partial^2 a_m}{\partial t^2} = [p_m + q_m \cos(\omega t)]a_m \qquad (2)$$

with coefficients q_m and p_m that depend on the corresponding eigenvalue of the Laplace operator and, hence, on the shape of the surface, on the depth of the liquid, as well as on gravity and surface tension. In addition, q_m is proportional to the amplitude of the periodic driver. The exponentially increasing solutions of the Mathieu equation that occur for certain combinations of system parameters, correspond to the instability. In particular, it becomes clear from (2) why the dominant instability is subharmonic.

The Faraday instability can thus be understood within linear theory. However, linear theory is not

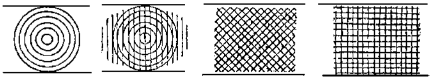

Figure 1. Sequence of surface wave patterns observed by Faraday on a quadrangular pool (Faraday, 1831).

sufficient to decide on which pattern will actually form after the instability has occurred. This can be determined by application of the modern theory of pattern formation to this damped-driven system (Miles & Henderson, 1990; Zhang & Viñals, 1997). Here, the nonlinearity comes into play. For sufficiently large spatial extension of the surface, boundary conditions at the edge of the surface become unimportant, and we may expand ζ in plane waves. Near the onset of the subharmonic instability, one may write

$$\zeta(\boldsymbol{R}, t) = \cos(\omega t/2) \sum_{j=1}^{N} \exp(\boldsymbol{k}_j \cdot \mathbf{R}) A_j(t) + \text{c.c.} \quad (3)$$

to a good approximation. (In (3) c.c. stands for the complex conjugate and a finite number N of wave vectors has been considered.) The modulus of the wave vectors $\boldsymbol{k}_j$ is fixed by the dispersion relation of linear surface waves at frequency $\omega / 2$. Viñals and his coworkers have shown that the slow variations of the amplitudes A_j are governed by a nonlinear equation of the form (Zhang & Viñals, 1997)

$$\frac{\partial A_j}{\partial \tau} = -\partial \mathcal{F} / \partial A_j^* \quad (4)$$

with the Lyapunov function

$$\mathcal{F} = \sum_{j=1}^{N} |A_j|^2 + \frac{1}{2} \sum_{j,j'=1}^{N} g(c_{jj'}) |A_j|^2 |A_{j'}|^2. \quad (5)$$

Here, τ is a stretched time, $c_{jj'}$ is the cosine of the angle between wave vectors $\boldsymbol{k}_j$ and $\boldsymbol{k}_{j'}$, and the function g depends on the system parameters. Stable stationary wave patterns correspond to minima of the Lyapunov function. For given system parameters, the pattern corresponding to the lowest value of $\mathcal{F}$ should be attained. Zhang and Viñals tested configurations with amplitudes A_j, 1, ..., M nonzero and $|A_j|$ independent of j, while the remaining $N - M$ amplitudes are zero. If the amplitudes corresponding to three or fewer different wave vectors are nonzero ($M \leq 3$), regular periodic patterns are obtained. The square pattern, already observed by Faraday, corresponds to two orthogonal wave vectors. With amplitudes of three different wave vectors nonzero, hexagonal or triangular patterns are generated. Interestingly, the theory of Zhang and Viñals also predicts patterns with more than three different wave vectors to be stable for certain system parameters. These correspond to quasi-periodic structures analogous to two-dimensional quasicrystals such as the Penrose tiling, for example. It has been a remarkable success of this nonlinear theory of pattern selection to predict the appearance of such quasi-periodic patterns of eight-fold (four wave vectors) and ten-fold (five wave vectors) symmetry, which have subsequently been found in experiments (Binks & van de Water, 1997) (see Figure 2).

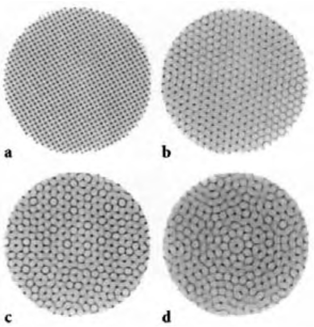

Figure 2. Faraday wave patterns of square (a), hexagonal (b), eight-fold quasi-periodic (c), and ten-fold quasi-periodic (d) symmetry observed by Binks & van de Water (1997).

In systems with length scales such that the boundary of the fluid's surface is relevant for the pattern formed, one has to consider the eigenfunctions $S_m(\boldsymbol{R})$ of the Laplace operator satisfying the correct conditions at the boundary and, hence, depending on the shape of the surface. This opens up the possibility of using Faraday waves to visualize wave-mechanical modes of stadium billiards and relating them to classical trajectories (Kudrolli et. al., 2001). At high values of the periodic acceleration of the fluid, spatiotemporal chaos has been observed.

In addition to the wave patterns arising on the surface of a moving liquid, Faraday, in his 1831 report, describes two other types of standing waves on the surfaces of liquids. The first one concerns patterns that he observed when a wave-maker was oscillating right on top of the surface or partly immersed in the liquid. Ridges are formed that are directed away from the wave-maker. The second type of standing waves occurs on a sloping beach when wind is blowing along the surface against the beach. Here, the ridges are parallel to the direction of the wind. These waves have been studied in detail in theory and by further experiments (Miles & Henderson, 1990). They are relevant in the context of oceanography.

Surface Acoustic Waves

In his landmark paper of 1885, Lord Rayleigh showed that the equations of linear elasticity theory for an isotropic homogeneous elastic half-space with a planar

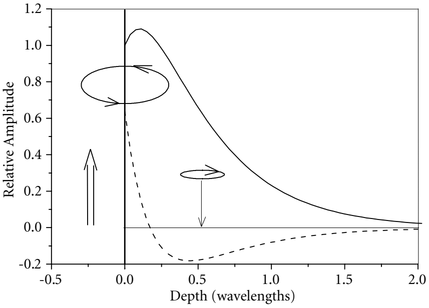

Figure 3. Displacements associated with a Rayleigh wave as function of depth. The surface wave is propagating in the direction of the double arrow. Displacement component normal (solid) and parallel (dashed) to the surface.

stress-free surface admit solutions that correspond to straight-crested waves propagating along the surface (Rayleigh, 1885). The deformations associated with these waves decay exponentially with depth. Such a Rayleigh wave is polarized in the plane spanned by the propagation direction and the surface normal (sagittal plane). It causes the mass elements of the solid to move on ellipses. This motion is retrograde near the surface and changes to prograde at greater depth (Figure 3).

Rayleigh himself thought of these waves in connection with earthquakes and their consequences, pointing out that surface wave amplitudes decrease more slowly with distance from the epicenter than those of bulk waves.

Since then, surface acoustic waves have found important applications in many branches of science and technology with wavelengths ranging from the kilometer scale in geophysics down to micrometers in ultrasonic nondestructive testing, acoustic microscopy, sensors, and micro-electronics. They are also used for basic research in solid-state physics with wavelengths down to the nanometer scale. The advent of the interdigital transducer on surfaces of piezoelectric media led to a number of signal processing devices based on SAW, especially frequency filters that are nowadays used in television sets and cell phones.

A distinctive feature of surface waves on solids as compared with ordinary liquids is the possibility of the medium to be anisotropic. For a long time it had not been known whether surface acoustic waves exist at all on surfaces of anisotropic media apart from special highly symmetric geometries. In the meantime, a mathematical theorem was established by Lothe and Barnett that guarantees their existence in most cases. However, generalized Rayleigh waves that occur in anisotropic media need no longer be polarized in the sagittal plane and may have more complicated depth profiles than the one shown in Figure 3. Their associated displacement field $\boldsymbol{u}$ is a superposition of generalized plane waves,

$$\boldsymbol{u}(x,z,t) = \sum_{j=1}^{n} \boldsymbol{W}_j \exp[\mathrm{i}k(x - vt) - |k|\alpha_j z] + \mathrm{c.c.}\,, \qquad (6)$$

when the wave vector $\boldsymbol{k} = k\hat{\boldsymbol{x}}$ is along the x-direction and the elastic medium fills the half-space $z > 0$. (v is the phase velocity.) The (often complex) coefficients α_j have positive real part. In isotropic media, $n = 2$; in nonpiezoelectric media, $n \leq 3$. In very special circumstances, n may even be 1. In piezoelectric substrates, $n \leq 4$.

In anisotropic media, the group velocity of a surface acoustic wave need no longer be parallel to its wave vector. This can lead to focusing of acoustic energy propagating away from a local excitation of the surface. This effect, which is well known for acoustic bulk waves, has been demonstrated for surface waves in a comparatively simple and efficient way by Kolomenskii and Maznev. Anisotropy also gives rise to leaky waves (pseudo-surface waves) with a very high degree of localization at the surface.

When accounting for piezoelectricity and/or considering an elastic substrate in contact with other media, a whole zoo of various species of surface waves is found (Maradudin, 1985; Auld, 1990), including waves

- of shear-horizontal polarization in homogeneous piezoelectric substrates with or without metal coating (Bleustein–Gulyaev waves),
- in layered structures with shear-horizontal (Love waves) or sagittal polarization (generalized Lamb waves, Sezawa waves),
- at the interface of a solid and a liquid (Scholte waves),
- at the interface between two different solids (Stoneley waves).

The power flow associated with a Rayleigh wave is localized at the surface within a layer of approximately a wavelength. The high spatial localization favors effects generated by the elastic nonlinearity of the medium. In particular, higher harmonics are easily observed. Because neither the equations of nonlinear elasticity nor the geometry of a homogeneous half-space with a planar surface define a length scale, surface acoustic waves are nondispersive in this system. Therefore, shock formation can be observed. Theoretically, the evolution of nonlinear wave forms is studied on the basis of the following equation which follows from derivations by Lardner and Parker (Parker, 1988):

$$\mathrm{i}\frac{\partial}{\partial \tau} B(k) = k \int_0^k F(q/k)\, B(q)\, B(k-q)\, \mathrm{d}q$$
$$+2k \int_k^\infty (k/q)\, F^*(k/q)\, B(q)\, B^*(q-k)\, \mathrm{d}q. \qquad (7)$$

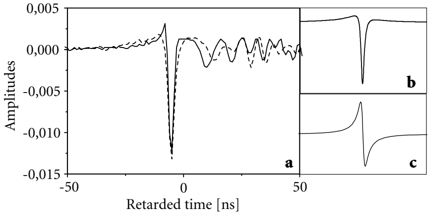

Figure 4. Solitary pulse shapes on the coated Si(111) surface (propagation along the $[\bar{1}\bar{1}2]$ direction) (a,b) and on a coated surface of isotropic quartz (c). Left: Experimental results by Lomonosov and Hess (solid) and numerical simulation (dashed). Right: Stationary pulse shapes. The local slope of the surface is shown as function of arrival time. (Compiled from data published in Lomonosov et al., 2002).

Here, τ is a stretched time and the function B is the Fourier transform of the local surface slope,

$$\frac{\partial}{\partial x}u_3(x,0,t) = \int_0^\infty B(k,\tau)\exp[\mathrm{i}k(x-vt)]\,\mathrm{d}k + \text{c.c.} \quad (8)$$

The complex function F depends on the second-order and third-order elastic moduli of the medium and may often be well approximated by its value at the argument 1 / 2. When transforming (7) into real space, the nonlinearity turns out to be strongly nonlocal. This is the price one has to pay for eliminating the depth dependence of the displacement field from the evolution equation.

On the surface of homogeneous isotropic media, the velocity component of mass elements normal to the surface develops a cusp, while the component parallel to the surface steepens in a way similar to the pressure in the formation of shocks in gases and liquids. This behavior has been found in simulations starting with a sinusoidal wave form and in experiments with pulsed laser excitation. In order to generate acoustic pulses of high intensity at solid surfaces, this experimental technique has proven to be most effective (Lomonosov et al., 2001). A laser pulse focused on a line on the surface leads to rapid local heating and generates an explosive evaporation of a highly absorbing coating. Similar to the epicenter of an earthquake in the macroscopic world, the explosion leads to an acoustic pulse that is localized at the surface after having propagated some distance from the line of excitation.

Depositing on the substrate surface a thin film of a material different from that of the substrate leads to a linear dispersion term in (7). The evolution equation then has solitary wave solutions that may be regarded as solid state analogs of the solitons on shallow water. They have shapes strongly influenced by the anisotropy of the substrate, as demonstrated by the examples in Figure 4. The existence of solitary surface acoustic pulses has been verified by Lomonosov and Hess using laser excitation to generate an initial pulse that subsequently evolves into a solitary pulse and a quasilinear background (Figure 4).

ANDREAS MAYER

See also **Billiards; Nonlinear acoustics; Pattern formation; Shock waves; Solitons; Tessellation; Water waves**

Further Reading

Auld, B.A. 1990. *Acoustic Fields and Waves in Solids*. 2nd edn., Malabar, Florida: Krieger

Binks, D. & van de Water, W. 1997. Nonlinear pattern formation of Faraday waves. *Physical Review Letters*, 78: 4043–4046

Faraday, M. 1831. On a peculiar class of acoustical figures; and on certain forms assumed by groups of particles upon vibrating elastic surfaces. *Philosophical Transactions of the Royal Society of London*, 121: 299–340

Kudrolli, A., Abraham, M.C., & Gollub, J.P. 2001. Scarred patterns in surface waves. *Physical Review* E, 63: 026208/1–8

Lomonosov, A., Mayer, A.P. & Hess, P. 2001. Laser-based surface acoustic waves in materials science. In *Experimental Methods in the Physical Sciences*, vol. 3: *Modern Acoustical Techniques for the Measurement of Mechanical Properties*, edited by M. Levy, H.E. Bass, & R. Stern, San Diego: Academic, pp. 65–134

Lomonosov, A., Hess, P. & Mayer, A.P. 2002. Observation of solitary elastic surface pulses. *Physical Review Letters* 88: 076104/1–4

Maradudin, A.A. 1985. Surface acoustic waves. In *Nonequilibrium Phonon Dynamics*, edited by W.E. Bron, New York: Plenum, pp. 395–599

Miles, J. & Henderson, D. 1990. Parametrically forced surface waves. *Annual Review of Fluid Mechanics*, 22: 143–165

Parker, D.F. 1988. Waveform evolution for nonlinear surface acoustic waves. *International Journal of Engineering Science*, 26: 59–75

Rayleigh, Lord (John William Strutt). 1885. On waves propagating along the plane surface of an elastic solid. *Proceedings of the London Mathematical Society*, 17: 4–11

Zhang, W. & Viñals, J. 1997. Pattern formation in weakly damped parametric surface waves. *Journal of Fluid Mechanics*, 336: 301–330

SYMBOLIC DYNAMICS

Dynamical systems on manifolds are usually described by equations using local Euclidean variables. Symbolic dynamics has its origins in an alternative description of the orbits of a dynamical system by Jacques Hadamard at the end of the 19th century and later by Marston Morse, where it was shown that complex orbital behavior of a dynamical system could be described using sequences of symbols (Hadamard, 1898; Morse, 1921).

A simple example that shows the usefulness of this approach is the iteration $x_{n+1} = f(x_n)$ on the interval of real numbers $\mathbb{I} = [0, 1]$ defined by a piecewise linear map

$$x_{n+1} = \begin{cases} 2x_n, & x_n < 0.5, \\ 2x_n - 1, & x_n \geq 0.5. \end{cases}$$

This map can also be described by

$$x_{n+1} = 2x_n \bmod 1$$

for $x_n \in [0, 1)$ together with $f(1) = 1$. It is easy to *solve* the modular equation formally for the orbital sequence $x_n = f^n(x_0)$, $0 \leq qx_0 < 1$, as

$$x_n = 2^n x_0 \bmod 1.$$

However, the formal solution is no more illuminating than the defining iteration, because it is not at all clear how the repeated exponentiation of the solution interacts with the process of reducing mod 1. To overcome this problem and unlock the dynamical secrets held within the formula, note that any real number $x \in \mathbb{I}$ can be written in the form

$$x = \sum_{n=1}^{\infty} \frac{b_n}{2^n},$$

where $b_n = 0$ or 1. Let Σ_2 denote the set of all such binary sequences. Thus, every binary sequence $\mathbf{b} = \{b_n\}_{n=1}^{\infty} \in \Sigma_2$ represents a point $\pi(\mathbf{b}) = x \in \mathbb{I}$ by a map $\pi : \Sigma_2 \to \mathbb{I}$.

Relative to this new representation of points in $\mathbb{I}$ the map f takes the form of a shift $\sigma : \Sigma_2 \to \Sigma_2$ on binary sequences as

$$\{b_1, b_2, \ldots, b_n, \ldots\} \longmapsto \{b_2, b_3, \ldots, b_{n-1}, \ldots\},$$

that is,

$$\sigma(\{b_n\}_{n=1}^{\infty}) = \{b_{n+1}\}_{n=1}^{\infty}$$

simply because

$$2\left(\sum_{n=1}^{\infty} \frac{b_n}{2^n}\right) \bmod 1 = \sum_{n=1}^{\infty} \frac{b_{n+1}}{2^n}.$$

It is easy to check that the map π provides a *semi-conjugacy* $f\pi = \pi\sigma$ for f and σ. The map π is not a conjugacy because of the redundancy in the binary representation of the reals. This can be resolved by always choosing binary sequences representing reals in [0,1] that *do not* have infinite sequences of 1s. So, for example, 0.25 has the infinite binary representations $\{01\overline{0}\}$ and $\{00\overline{1}\}$, but we choose the former. Nevertheless, the semi-conjugacy implies that for every positive integer m, $f^m\pi = \pi\sigma^m$. In particular, if $\mathbf{b} = \{b_n\} \in \Sigma_2$ is a period-m orbit of σ, then

$$f^m(\pi(\mathbf{b})) = \pi\sigma^m(\mathbf{b}) = \pi(\mathbf{b}),$$

and so π maps the σ-periodic point $\mathbf{b}$ to an f-periodic point $\pi(\mathbf{b})$. Thus, we can obtain the periodic orbits of f by investigating the periodic orbit structure of σ. But this is straightforward because the period-m periodic points of σ in Σ_2 are precisely those binary sequences that repeat after m-digits and no fewer. Thus, period-1 points of σ are given by the repeating expansions

$$\mathbf{b} = \{\overline{0}\ldots\} = \{00000\ldots\}$$

and

$$\mathbf{b} = \{\overline{1}\ldots\} = \{11111\ldots\},$$

which both correspond by π to the fixed points $x = 0$ and $x = 1$, respectively, of the map f.

Period-2 points of σ are given by

$$\mathbf{b}_1 = \{\overline{01}\ldots\}, \qquad \mathbf{b}_2 = \{\overline{10}\ldots\}.$$

Note that $\pi(\mathbf{b}_1) = \frac{1}{3}$ and $\pi(\mathbf{b}_2) = \frac{2}{3}$ and $\left\{\frac{1}{3}, \frac{2}{3}\right\}$ is the only period-2 orbit of f. We can immediately see that periodic points of all orders can be constructed for σ, and thus also for f, in this way. Points that are on *eventually* periodic orbits of f can also be obtained by delaying the introduction of the binary recurrences in the symbol sequences. Thus, any map f defined on $\mathbb{I}$ that is described by a shift on symbol sequences has a rich periodic structure. We can extract further properties about orbits of f when we observe that π can be seen as a continuous function by taking the usual Euclidean metric on $\mathbb{I}$ and the metric

$$d(\mathbf{b}, \mathbf{b}') = \sum_{n=1}^{\infty} \frac{(b_n - b_n')}{2^n},$$

where $\mathbf{b} = \{b_n\}_{n=1}^{\infty}$ and $\mathbf{b}' = \{b_n'\}_{n=1}^{\infty}$ are elements of Σ_2. It is now possible to consider a *topology* on Σ_2.

A key feature of chaos is the requirement that the orbits of the map f should in some sense be bound together, thus, making the dynamics indecomposable. This can be achieved by finding an orbit of f that *densely* fills out the set $\mathbb{I}$. Such an orbit is given by constructing a sequence $\mathbf{b}$ obtained by listing all symbol sequences of length 1, then of length 2, and so on for all positive integers. Let $\mathbf{b}'$ be any prescribed binary sequence. Given any positive integer n, by construction $\mathbf{b}$ has within it the symbol block $\mathbf{b}'^{(n)}$ consisting of the first n symbols of $\mathbf{b}'$. Suppose this block commences at the point b_{k+1} of $\mathbf{b}$. Then $\mathbf{b}'$ and $\sigma^k(\mathbf{b})$ have the same first n entries and thus $d(\mathbf{b}', \sigma^k(\mathbf{b})) < 2^{-n}$. The positive integer n was chosen arbitrarily, and so the orbit of $\mathbf{b}$ approaches arbitrarily close to $\mathbf{b}'$. Thus, the orbit of $\mathbf{b}$ is *dense* on $\mathbb{I}$ and binds the dynamical behavior together preventing a decomposition into closed subsets with simpler dynamical behavior.

The dense orbit was, of course, produced to a special recipe. Obviously, the orbit is not unique as other dense orbits in Σ_2 can be constructed by simple permutations of the ordering of the blocks used to construct the sequence $\mathbf{b}$. What is astonishing is that "almost all" orbits are dense. This follows from a classical result of Emil Borel around 1900 (Hardy, 1983), which, when paraphrased, says that the set of real numbers in $\mathbb{I}$, which have all possible finite sequences of “0”s and “1”s in their binary expansions, called *normal* numbers, has full measure.

Thus, the doubling map can be described by a shift map σ on the space of symbolic sequences on two symbols, Σ_2. The simple ideas discussed here can be extended to provide symbolic dynamics for various types of maps. The *coding* involving two or more symbols has to be carried out so that each point x of the map domain is given by a symbol string that provides the orbit of x by merely shifting the symbols (Devaney, 2003). Note that the binary sequence gives the itinerary of an orbit where if “0” represents the interval $[0, \frac{1}{2}]$ and “1” represents $[\frac{1}{2}, 1]$, the sequence gives the sequence of intervals visited by the orbit of the map f. This approach can be applied to a wide class of one-dimensional maps (Guckenheimer, 1979; Guckenheimer et al., 1977), but it also has much greater importance in dynamical systems.

A key dynamical construction is the *Smale Horseshoe* (Smale, 1967). The simplest example is a diffeomorphism h of the square $Q \in \mathbf{R}^2$ whose image in $\mathbf{R}^2$ is folded in the shape of a horseshoe to overlay the square. The map h has an invariant Cantor set $\Lambda \subset Q$ such that $h|\Lambda$ can also be coded with *bi-infinite* binary sequences together with a shift map. This results in a complex periodic orbit structure and dense orbits within Λ, which gives some of the ingredients of chaos. The key components of the set Λ are the *homoclinic* points, that is, those points that asymptotically approach a periodic orbit of h in both forward and reverse iterations.

The Smale–Birkhoff Theorem proves that “folding maps” of this type, and the associated complex dynamical behavior, exist in the neighborhood of such homoclinic points. This is precisely stated as follows (Smale, 1967).

Theorem. Let $\mathbf{f}$ be a Kupka–Smale diffeomorphism of a compact manifold M and $x^\dagger$ be a transverse homoclinic point of a periodic point x^* of $\mathbf{f}$. Then there is a closed subset Λ containing $x^\dagger$, such that (i) Λ is a Cantor set; (ii) $\mathbf{f}^p(\Lambda) = \Lambda$ for some $p \in \mathbf{Z}^+$; and (iii) $\mathbf{f}^p|\Lambda$ is topologically conjugate to a shift on a symbol space.

Note that all possible binary sequences are allowed in the above symbol spaces and, thus, we have been considering *full* shifts. In many practical applications of data storage, there are practical restrictions on the stored patterns of sequences. The problem of transferring codes that satisfy one set of constraints into other restricted symbol sets involve *sub-shifts of finite type* and allow applications to communications coding and algebra (Coornaert & Papadopoulos, 1993; Lind & Marcus, 1996; Kitchens, 1997). Symbolic coding has been directly useful in modeling in engineering and biology, see (Daw et al., 1997) and (Voss et al., 2000).

David D. Arrowsmith

See also **Chaotic dynamics; Horseshoes and Hyperbolicity in dynamical systems; Maps; One-dimensional maps; Topology**

Further Reading

Coornaert, M. & Papadopoulos, A. 1993. *Symbolic Dynamics and Hyperbolic Groups*, Berlin & New York: Springer

Crutchfield, J.P. & Packard, N. 1982. Symbolic dynamics of one-dimensional maps: entropies, finite precision, and noise. *International Journal of Theoretical Physics*, 21: 433–466

Daw, C.S., Finney, C.E.A., Kennel, M.B. & Connolly, F.T. 1997. Cycle-by-cycle combustion variations in spark-ignited engines. *Proceedings of the Fourth Experimental Chaos Conference*, Boca Raton, FL, Singapore: World Scientific

Devaney, R.L. 2003. *An Introduction to Chaotic Dynamical Systems*, 2nd edition, Boulder, CO: Westview Press

Guckenheimer, J. 1979. Sensitive dependence to initial conditions for one-dimensional maps. *Communications in Mathematical Physics*, 70: 133–160

Guckenheimer, J., Oster, G. & Ipatchki, A. 1977. The dynamics of density dependent population models. *Journal of Mathematical Biology*, 4: 101–147

Hadamard, J. 1898. Les surfaces á courbures opposeés et leur lignes géodesiques. *Journal de Mathématiques Pures et Appliquées*, 4: 27–73

Hao, B.-L. 1986. *Elementary Symbolic Dynamics and Chaos in Dissipative Systems*, Singapore: World Scientific

Hao, B.-L. & Zheng, W.-M. 1998. *Applied Symbolic Dynamics and Chaos*, Singapore: World Scientific

Hardy, G.H. 1983. *An Introduction to the Theory of Numbers*, 5th edition, Oxford and New York: Oxford University Press (first published, 1938)

Kitchens, B.P. 1997. *Symbolic Dynamics: One-Sided, Two-Sided, and Countable State Markov Shifts*, Berlin and New York: Springer

Lind, D. & Marcus, B. 1996. *An Introduction to Symbolic Dynamics and Coding*, Cambridge and New York: Cambridge University Press

Milnor, J. & Thurston, W. 1988. On iterated maps of the interval. In *Dynamical Systems Proceedings, University of Maryland 1986–87*, edited by J.C. Alexander, Berlin and New York: Springer, pp. 425–563

Morse, M. 1921. A one-to-one representation of geodesics on a surface of negative curvature. *American Journal of Mathematics*, 43: 33–51

Smale, S. 1967. Differentiable dynamical systems. *Bulletin American Mathematical Society*, 73(6): 747–817.

Voss, A., Wessel, N., Kurths, J., Schirdewan, A., Osterziel, K.J., Malik, M. & Dietz, R. 2000. Symbolic dynamics—a powerful tool in non-invasive biomedical signal processing. In *Advances in Noninvasive Electrocardiographic Monitoring Techniques*, edited by H.H. Osterhues, V. Mombach & A.J. Moss, Dordrecht and Boston: Kluwer, pp. 429–437

SYMMETRY BREAKING

See **Bifurcations**

SYMMETRY GROUPS

Mathematically, a group is a set G together with a pair-wise “group multiplication” on elements of G that satisfies the following axioms. (i) Multiplication is associative, meaning that $xy(z) = x(yz)$ for x,

y, and z any elements of G. (ii) G contains an identity element e for which $ex = x = xe$ where x is any element of G. (iii) Each element of G has an inverse (x^{-1}), satisfying $x^{-1}x = e = xx^{-1}$ (Armstrong, 1988; Birkhoff & MacLane, 1953). Groups for which all of the multiplications commute are said to be abelian.

Easily visualized examples of finite symmetry groups are provided by operations on familiar geometrical objects, such as squares, hexagons, cubes, and tetrahedrons, which carry vertices into themselves. In general, there will be a finite set of f linearly independent functions (in the space of the geometrical object) that are carried into linear combinations of themselves by the group operations. Thus, each group element corresponds to a matrix, and the corresponding set of $f \times f$ matrices is called a representation of the group (Landau & Lifschitz, 1958).

For a square that is lying on the plane with its center at the origin, group elements are rotations about the origin (by 90°, 180°, and 270°) and reflections (in horizontal and vertical axes and in the two diagonals). Group multiplication is defined as performing two operations in order, and the identity element leaves all vertices unchanged. The reader may find it helpful to cut out a square of cardboard, number the four corners, and construct a group multiplication table for these eight elements, noting that it is not abelian.

It has long been known that symmetries of a dynamic system influence the nature of possible characteristic solutions or natural modes of behavior (*See* **Symmetry: equations vs. solutions**). To appreciate this restriction, consider a system with reflection symmetry, and note that any solution can be resolved into two components that are respectively symmetric and antisymmetric about the plane of symmetry. In a linear system, the symmetric and antisymmetric components do not interact; thus, modes must be either symmetric or antisymmetric about the plane. In a nonlinear system, however, there can also be natural modes of behavior that are neither symmetric nor antisymmetric. Such symmetry breaking is a fundamental feature of nonlinear dynamic systems, leading to the formation of solitons on optical fibers, impulses on nerve axons, and local modes in chemical molecules and molecular crystals.

Applications to Physical Chemistry

A typical chemical molecule is a simple geometric structure for which a symmetry group multiplication table has a finite number of entries (Herzberg, 1991; Wilson et al., 1980). Thus, the water molecule (H_2O) has a single reflection plane; methane (CH_4) is described by the same symmetry group as the regular tetrahedron, benzene (C_6H_6) by a planar hexagon, and so on. The physical chemist uses these finite symmetry groups to organize measurements of electronic and vibrational spectra, using the following notations: reflection group (σ), n-fold rotation symmetry (C_n), inversion symmetry (I), tetrahedral symmetry (T), and so on. A particularly lucid introduction to point symmetry notations is given by Landau & Lifschitz (1958).

Because quantum theory is linear, vibrational and electronic wave functions have the following property. If n successive applications of a symmetry operation return a molecule to its original orientation, the amplitude of the corresponding quantum wave function must change by an nth root of unity under the same operation. Under reflections, in other words, amplitudes change by factors of either $+1$ or -1 (symmetric or antisymmetric), whereas CH-stretching modes of benzene (C_6) change under rotations of 60° by factors of $\exp(\pm \mathrm{i}m\pi/3)$, where $m = 0, \pm 1, \pm 2, \pm 3, \pm 4$, or ± 5. Physical chemists label their spectral lines with notations that correspond to these factors (Landau & Lifschitz, 1958).

Although molecular vibrations can be quite nonlinear (especially those involving hydrogen atoms), this nonlinearity does not play a role in transitions from the ground state to first quantum levels of small molecules. (This is because the nonlinear operator of lowest order contains a product of two lowering operators, which annihilates first quantum states.) Thus, the excitations to the first quantum level are governed by linear symmetry considerations, and local modes in molecules are observed only for transitions to higher quantum levels.

The linear modes of a molecule with a center of inversion (I) appear either in infrared absorption or Raman scattering measurements, but not both. This "principle of mutual exclusion" helps to sort out the components of linear vibrational spectra. Local modes, which are nonlinear, can appear in both infrared and Raman measurements.

A periodic solid (or molecular crystal) with periodic boundary conditions can be viewed as a very large molecule, for which the number of elements of the corresponding symmetry group is also very large. Of particular interest are translations by lattice constants (a, b, and c) along the crystal axes, which bring the crystal back to its original configuration. The phase shift of an electronic or vibrational wave function under such a translation is $\exp(\mathrm{i}\boldsymbol{k})$, where $\boldsymbol{k}$ is called the crystal momentum. As the number of unit cells in the model approaches infinity, the three components of $\boldsymbol{k}$ vary, respectively, from $-\pi/a$ to π/a, $-\pi/b$ to π/b, and $-\pi/c$ to π/c. In vibrational modes of such systems, nonlinearity can arise from local lattice distortion, which allows symmetry breaking (local mode formation) to be observed at the first quantum level.

Applications to Field Theories

The above definition of a group does not require the number of group elements to be finite, and many partial differential equations provide examples of infinite-order symmetry groups.

The sine-Gordon (SG) equation, for example, is invariant under the independent variable transformation $(x, t) \to (\xi, \tau)$, where $\xi = (x - vt)\,/\,\sqrt{1 - v^2}$ and $\tau = (t - vx)\,/\,\sqrt{1 - v^2}$. Taking this invariance as the property defining elements of the group, a symmetry group comprises all such transformations, parameterized by the continuous variable v with $|v| < 1$. This is a one-dimensional version of the Lorentz transformation, which is shared by Maxwell's equations. Interestingly, if $u(x)$ is a time-independent solution of SG, then $u\left[(x - vt)\,/\,\sqrt{1 - v^2}\right]$ is also a solution that demonstrates Lorentz contraction (becomes smaller as $v \to 1$).

In 1915, Emmy Noether established the following important result on the application of symmetry groups to field theories (José & Saletan, 1998).

> **Noether's theorem.** *If a system is described by a Lagrangian that remains invariant under some continuous symmetry transformation, then there is a corresponding conservation law and constant of the motion.*

This theorem has immediate implications. As fundamental descriptions of nature are assumed to be Lagrangian and independent of time displacements (the science of today is the same as it was yesterday), the corresponding conserved quantity is energy. In other words, the law of energy conservation stems from time invariance of scientific laws. Similarly, conservation of momentum and conservation of angular momentum arise, respectively, from the assumptions that scientific laws are independent of spatial displacements and angles of rotation.

Nowadays, it is widely assumed that the fundamental fields of nature are derived from a Lorentz invariant Lagrangian density or more generally based on the Poincaré group, which incorporates independence with respect to displacements in time and space (Kim & Noz, 1986). As physicists attempt to formulate a fundamental description of nature, they use Noether's theorem to build in additional constants (charge, spin, rest mass, etc.) that have been empirically observed.

The pure soliton equations (Korteweg–de Vries, nonlinear Schrödinger, sine-Gordon, and so on) are Lagrangian systems that each have a countably infinite number of conservation laws and constants of the motion. It would be interesting to better understand the corresponding symmetries.

ALWYN SCOTT

See also **Bäcklund transformations; Dimensional analysis; Lie algebras and Lie groups; Local modes in molecules; Symmetry: equations vs. solutions**

Further Reading

Armstrong, M.A. 1988. *Groups and Symmetry*, New York: Springer

Birkhoff, G. & MacLane, S. 1953. *A Survey of Modern Algebra*, New York: Macmillan

Herzberg, G. 1991. *Molecular Spectra and Molecular Structure: Infrared and Raman of Polyatomic Molecules*, Melbourne, FL: Krieger Publishing Company

José, J.V. & Saletan, E. 1998. *Classical Dynamics: A Contemporary Approach*, Cambridge: Cambridge University Press

Kim, Y.S. & Noz, M.E. 1986. *Theory and Applications of the Poincaré Group*, Dordrecht: Reidel

Landau, L.D. & Lifshitz, E.M. 1958. *Quantum Mechanics: Non-relativistic Theory*, Reading, MA: Addison-Wesley

Weyl, H. 1952. *Symmetry*, Princeton, NJ: Princeton University Press

Wilson, E.B., Jr., Decius, J.C. & Cross, P.C. 1980. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*, New York: Dover

SYMMETRY: EQUATIONS VS. SOLUTIONS

The importance of symmetries in various fields of science is well appreciated, much beyond the scope of this entry. Here we focus on one particular aspect—how knowledge about discrete symmetry operations allows one to conclude certain properties of solutions of differential equations.

Let us start with one of the simplest examples: one oscillator that is governed by the equation

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = -\frac{\mathrm{d}V(x)}{\mathrm{d}x}, \tag{1}$$

where t is time, x the coordinate of the oscillator, and $V(x)$ its nonnegative potential with $V(0) = 0$. All solutions to this equation are time-periodic. If an equation is invariant under a certain symmetry operation, then the same symmetry operation applied to a solution of the equation either generates a new solution or reproduces the same old solution it is then said that the solution is also invariant.

All symmetries considered below, in fact, do not change the energy of the oscillator; thus, the solution is always invariant up to a trivial shift of the origin of time $t \to t + t_0$. One symmetry that leaves Equation (1) invariant is time reversal $t \to -t$, which implies that for each solution the origin of time can be chosen in such a way that the solution is also invariant under time reversal, that is, $x_{\mathrm{s}}(t) = x_{\mathrm{s}}(-t)$. For practical purposes, it means that numerical expansions of solutions into Fourier series may be restricted to cosine Fourier series. Another symmetry may hold if $V_{\mathrm{s}}(x) = V_{\mathrm{s}}(-x)$. Then the equation is invariant under space reflection $x \to -x$.

In order to invert the sign of a periodic function $x(t)$, we may assume either antisymmetry $x_{\mathrm{a}}(t) = -x_{\mathrm{a}}(-t)$ or shift symmetry $x_{\mathrm{sh}}(t) = -x_{\mathrm{sh}}(t + T/2)$ where T

is the period. Thus, for symmetric potentials, $x(t)$ is symmetric in time, antisymmetric in time (around a different origin), and shift symmetric. In particular, this implies that all even components of the Fourier expansion of this function vanish, including the dc component.

Consider next a more sophisticated model of a particle subject to a space-periodic force $f(x) = -\,\mathrm{d}V/\mathrm{d}x$ (period λ) and a time-periodic field $E(t)$ (period T) with zero mean:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = f(x) + E(t)\,. \tag{2}$$

Equation (2) corresponds to a non-integrable system with 1.5 degrees of freedom (a three-dimensional phase space). The more complicated the equations become, the more symmetries one might consider. Assume symmetry operations that change the sign of the velocity $v = \mathrm{d}x/\mathrm{d}t$. There are two symmetries: time reversal $t \to -t$ if $E(t) \equiv E_{\mathrm{s}}(t)$ and a combined space reflection with time shift one $x \to -x,\ \ t \to t + T/2$ if $f(x) \equiv f_{\mathrm{a}}(x),\ \ E(t) \equiv E_{\mathrm{sh}}(t)$. For both symmetries to be valid, certain properties of the functions f and E are required. In particular, the time-periodic field $E(t)$ (e.g., an ac electric field) enters both symmetry requirements. That means that by choosing an $E(t)$ dependence that is neither symmetric nor shift symmetric, we lose both symmetries. What are the consequences?

Note first that due to the non-integrability of (2), the separatrix for $E = 0$ is replaced by a chaotic or stochastic layer in the phase space. Ergodicity inside the layer implies that the time average of a certain quantity over a trajectory, if existing, does not depend on the concrete trajectory choice. Returning to the above symmetries and choosing the time average of the velocity of the particle, we conclude that (i) the velocity average exists because the stochastic layer is bounded, and (ii) if the average velocity is nonzero for one trajectory from the stochastic layer and if any of the above symmetries apply, the velocity average will be opposite for a corresponding symmetry related trajectory from the same stochastic layer—with the only consequence that the average velocity vanishes exactly. However, choosing an $E(t)$ dependence that violates both symmetry requirements, we may expect that the average velocity of any trajectory inside the stochastic layer will become nonzero. In other words, we can predict how to generate a dc current of many non-interacting particles by a symmetry breaking choice of the field $E(t)$. This has been, indeed, predicted and demonstrated theoretically and numerically (Flach et al., 2000) and verified experimentally for directed diffusion of ultracold rubidium atoms in symmetric optical lattices (Schiavoni et al., 2003).

Next, consider applications of the above concepts to systems that are characterized by the presence of dissipation and fluctuations. Usual additive Gaussian white noise terms possess all symmetries of periodic functions; that is, if $\xi(t)$ is a realization of a stochastic process, so is $\xi(-t),\ \ \xi(t+\tau),\ \ -\xi(t)$, and so on. When starting with a corresponding Langevin type equation

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = f(x) + E(t) - \gamma\frac{\mathrm{d}x}{\mathrm{d}t} + \xi(t), \tag{3}$$

it is then possible to drop the noise term and study the symmetries of the deterministic yet dissipative equation

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = f(x) + E(t) - \gamma\frac{\mathrm{d}x}{\mathrm{d}t}. \tag{4}$$

While the strategy of the analysis is similar to the one outlined above, for the case of zero dissipation $\gamma = 0$, the phase space of Equation (4) is again three-dimensional, but is now composed of basins of attraction with each basin corresponding to a certain attractor. Existing symmetries of the equation imply symmetry relationships between attractors and their basins of attraction. Violation of symmetries of the equations implies desymmetrization of basins of attraction of (previously symmetry related) attractors. Thus, adding again the stochastic noise to return to the starting Equation (3), intuition suggests that the noise when leading to an average over various basins of attraction will either make certain averages vanish (in the presence of symmetries) or not vanish (in the absence of symmetries). Special care is required due to the fact that some symmetries of (4) may relate attractors with repellors. Another approach is to consider partial differential equations for the probability distributions and their symmetry properties (see Denisov et al., 2002).

The above approach has been successfully used to predict such diverse phenomena as rectification of heat currents (Flach et al., 2002) and induction of magnetizations (Flach & Ovchinnikov, 2000). Finally, note that it is easier to predict a nonzero average by breaking symmetries than to obtain an understanding for the concrete microscopic mechanisms of symmetry breaking and thus to obtain reliable estimates for the expected quantity of rectification and its dependence on essential parameters of the system (Reimann, 2002).

SERGEJ FLACH

See also **Ratchets; Symmetry groups; Stochastic processes**

Further Reading

Denisov, S., Flach, S., Ovchinnikov, A.A., Yevtushenko, O. & Zolotaryuk, Y. 2002. Broken space-time symmetries and mechanisms of rectification of ac fields by nonlinear (non)adiabatic response. *Physical Review* E, 66: 041104

Flach, S. & Ovchinnikov, A.A. 2000. Static magnetization induced by time-periodic fields with zero mean. *Physica* A, 292: 268–276

Flach, S., Yevtushenko, O. & Zolotaryuk, Y. 2000. Directed current due to broken time-space symmetry. *Physical Review Letters*, 84: 2358–2361

Flach, S., Zolotaryuk, Y., Miroshnichenko, A.E. & Fistul, M.V. 2002. Broken symmetries and directed collective energy transport. *Physical Review Letters*, 88: 184101

Reimann, P. 2002. Brownian motors: noisy transport far from equilibrium. *Physics Reports*, 361: 57–265

Schiavoni, M., Sanchez-Palencia, L., Renzoni, F. & Grynberg, G. 2003. Phase control of directed diffusion in a symmetric optical lattice. *Physical Review Letters*, 90: 094101

SYMPLECTIC MAPS

First used mathematically by Hermann Weyl, the term *symplectic* arises from a Greek word that means "twining or plaiting together." This is apt, as symplectic systems always involve a pair of n-dimensional variables, the configuration q, and momentum p, which are intertwined by the symplectic two form

$$\omega = \mathrm{d}p \wedge \mathrm{d}q. \qquad (1)$$

This antisymmetric, bilinear form acts on a pair of tangent vectors and computes the sum of the areas of the parallelograms formed by projecting the vectors onto the planes defined by each canonical pair (q_i, p_i), $i = 1, \ldots, n$, giving

$$\omega(v, w) = \sum_{i=1}^{n} (v_{p_i} w_{q_i} - v_{q_i} w_{p_i}).$$

A diffeomorphism $f : X \to X$ on a $2n$-dimensional manifold X with coordinates $z = (q, p)$ is symplectic if it preserves the symplectic form, that is, if $f^*\omega = \omega$ (Arnol'd, 1989; McDuff & Salamon, 1995). If we write $z' = (q', p') = f(q, p)$, the symplectic condition becomes

$$Df^t J Df = J, \quad \text{where} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \qquad (2)$$

Here $Df_{ij} = \partial f_i / \partial z_j$ is the Jacobian matrix of f, J is the Poisson matrix, and I is the $n \times n$ identity matrix. Equivalently, Stokes' theorem can be used to show that the loop action, $A[\gamma] = \oint_\gamma p\mathrm{d}q$, is preserved by f for any contractible loop γ on X. If f preserves the loop action for all loops, even those that are not contractible, then it is *exact symplectic*.

When $n = 1$, the symplectic condition is equivalent to $\det(Df) = 1$, so that the map is area- and orientation-preserving. Examples include the much studied *standard map* and the area-preserving Hénon quadratic map $f(q, p) = (p + a - q^2, -q)$ (Meiss, 1992). When $n > 1$, the symplectic condition implies volume and orientation preservation, but as we will see, it is stronger than this. A generalization of the standard map to higher dimensions is the map

$$\begin{aligned} q' &= q + p - \nabla V(q), \\ p' &= p - \nabla V(q), \end{aligned} \qquad (3)$$

where $q \in \mathbb{T}^n$ is an angle, $p \in \mathbb{R}^n$ is its conjugate momentum, and $V(q)$ is a periodic potential. This map is exact symplectic for any V. Beginning in 1972, Claude Froeschlé studied the case $n = 2$ and $V(q) = a \cos q_1 + b \cos q_2 + c \cos(q_1 + q_2)$. Similarly, the natural generalization of the Hénon map is the quadratic symplectic map whose normal form has been given by Moser (1994).

Applications

Symplectic maps arise from Hamiltonian dynamics, because these preserve the loop action. Thus, for example, the time t map of any Hamiltonian flow is symplectic, as is a Poincaré return map defined on a cross section. It is often easier to study the Poincaré map instead of the flow, because the dimension is reduced. Even though explicit construction of the map is typically impossible, approximation methods often suffice.

For example, the time T map of a periodically forced system $H(q, p, t) = H(q, p, t + T)$, such as a pendulum with an oscillating support, is symplectic (*See* **Hamiltonian systems**). An extreme example is $H = \frac{1}{2} p^2 - k \cos(q) \bar{\delta}(t)$, where $\bar{\delta}$ is the periodic Dirac delta function; the corresponding map is the standard map.

As Birkhoff showed, an ideal billiard (a free particle moving inside a rigid, convex table) is naturally written as a symplectic map. The canonical coordinates are the position and the tangential momentum at a collision point. Symplectic maps also arise naturally in systems where the forces are localized in time or space. For example, a circular particle accelerator or storage ring has a sequence of accelerating and focusing elements that can be modeled by a composition of symplectic maps, providing the damping effects of radiation can be neglected (Forest, 1998).

Area-preserving maps also arise in the study of the motion of Lagrangian tracers in incompressible fluids or of particles tightly gyrating around magnetic field lines. In particular, when one component of the field is particularly strong, such as in the plasma device called a tokamak, the transverse dynamics reduces to an area-preserving map.

Autonomous canonical transformations are symplectic maps. For example, if $F(q, q')$ is a generating function for a canonical transformation, then it generates a symplectic map. In particular, the Froeschlé map (3) is generated by $F(q, q') = \frac{1}{2}(q' - q)^2 - V(q)$.

An algorithm that respects the symplectic nature of Hamiltonian dynamics is called a symplectic integrator. A first-order symplectic algorithm with time step Δt for the Hamiltonian $H(q, p)$ is generated by $F(q, p') = qp' + \Delta t H(q, p')$, where

$dF = q'dp' + p\,dq$, giving the map

$$q' = q + \Delta t \frac{\partial H}{\partial p'}(q, p'), \quad p' = p - \Delta t \frac{\partial H}{\partial q}(q, p'). \quad (4)$$

Note that the map is implicit because H is evaluated at p'. However, for the case that $H = K(p) + V(q)$ this becomes a leap-frog Euler scheme, an example of a "splitting" method. Symplectic versions of many standard algorithms—such as Runge–Kutta—can be obtained (Marsden et al., 1996). While there is still some controversy on the utility of symplectic methods versus methods that, for example, conserve energy and other invariants or have variable time-stepping, they are superior for stability properties because they respect the spectral properties of the symplectic group.

The Symplectic Group

The stability of an orbit $\{...z_t, z_{t+1}, ...\}$, where $z_{t+1} = f(z_t)$, is governed by the Jacobian matrix of f evaluated along the orbit, $M = \prod_t Df(z_t)$. When f is symplectic, M obeys (2), $M^t JM = J$. The set of all such $2n \times 2n$ matrices form the symplectic group $Sp(2n)$. This group is an $n(2n+1)$-dimensional Lie group, whose Lie algebra is the set of Hamiltonian matrices—matrices of the form JS where S is symmetric. Thus, every near-identity symplectic matrix can be obtained as the exponential of a Hamiltonian matrix and corresponds to the time t-map of a linear Hamiltonian flow. There are symplectic matrices, however, that are not the exponentials of Hamiltonian matrices; for example, $-I$. As a manifold, the symplectic group has a single nontrivial loop (its fundamental group is the integers). The winding number of a loop in the symplectic group is called the Maslov index (McDuff & Salamon, 1995); it is especially important for semi-classical quantization.

If M is a symplectic matrix and λ is an eigenvalue of M with multiplicity k, then so is λ^{-1}. Moreover $\det(M) = 1$, so M is volume and orientation preserving. A consequence of this spectral theorem is that orbits of a symplectic map cannot be asymptotically stable. There are four basic stability types for symplectic maps: an eigenvalue pair (λ, λ^{-1}) is

- *hyperbolic*, if λ is real and larger than one;
- *hyperbolic with reflection*, if λ is real and less than minus one;
- *elliptic*, if $\lambda = e^{2\pi i \omega}$ has magnitude one;
- part of a *Krein quartet*, if λ is complex and has magnitude different from one, for then there is a quartet of related eigenvalues $(\lambda, \lambda^{-1}, \bar{\lambda}, \bar{\lambda}^{-1})$.

Thus, a periodic orbit can be linearly stable only when all of its eigenvalue pairs are elliptic. For this case, the linearized motion corresponds to rotation with n rotation numbers ω_i.

Symplectic Geometry

Every symplectic map is volume- and orientation-preserving, but the group Symp(X) of symplectic diffeomorphisms on X is significantly smaller than that of the volume-preserving ones. This was first shown in 1985 by Gromov in his celebrated "nonsqueezing" (or symplectic camel) theorem. Let $B(r)$ be the closed ball of radius r in $\mathbb{R}^{2n}$ and $C_1(R) = \{(q, p) : q_1^2 + p_1^2 \leq R^2\}$ be a cylinder of radius R whose circular cross section is a symplectic plane. Because the volume of C_1 is infinite, it is easy to construct a volume-preserving map that takes $B(r)$ into $C_1(R)$ regardless of their radii. What Gromov showed is that it is impossible to do this symplectically whenever $r > R$. This is one example of a *symplectic capacity*, leading to a theory of symplectic topology (McDuff & Salamon, 1995).

Another focus of this theory is to characterize the number of fixed points of a symplectic map, that is, to generalize the classical Poincaré–Birkhoff theorem for area-preserving maps on an annulus. Arnol'd conjectured in the 1960s that any Hamiltonian diffeomorphism on a compact manifold X must have at least as many fixed points as a function on X must have critical points. A Hamiltonian map is a symplectic map that can be written as a composition of maps of the form (4). Conley and Zender proved this in 1985 for the case that X is the $2n$-torus: f must have at least $2n+1$ fixed points (at least 2^{2n} if they are all nondegenerate) (Golé, 2001).

Dynamics

In general, the dynamics of a symplectic map consists of a complicated mixture of regular and chaotic motion (Meiss, 1992). Numerical studies indicate that the chaotic orbits have positive Lyapunov exponents and fill sets of positive measure that are fractal in nature. Regular orbits include periodic and quasi-periodic orbits. The latter densely cover invariant tori whose dimensions range from 1 to n. Near elliptic periodic orbits, the phase space is foliated by a positive-measure cantor set of n-dimensional invariant tori. There are chaotic regions in the *resonant* gaps between the tori, but the chaos becomes exponentially slow and exponentially small close to the periodic orbit. Some of these observations, but not all, can be proved.

The simplest case is that of an integrable symplectic map, which can be written in Birkhoff normal form: $f(\theta, J) = (\theta + \nabla S(J), J)$. Here (θ, J) are angle-action coordinates (each n-dimensional) and $\Omega = \nabla S$ is the rotation vector. Orbits for this system lie on invariant tori; thus, the structure is identical to that for integrable Hamiltonian systems.

The Birkhoff normal form is also an asymptotically valid description of the dynamics in the neighborhood of a nonresonant elliptic fixed point, one for which

$m \cdot \Omega(0) \neq n$ for any integer vector m and integer n. However, the series for the normal form is not generally convergent. Nevertheless, KAM theory implies that tori with Diophantine rotation vectors do exist near enough to the elliptic point, providing the map is more than C^3 and that the twist, det $D\Omega(0)$, is nonzero. Each of these tori is also a Lagrangian submanifold (an n-dimensional surface on which the restriction of the symplectic form (1) vanishes). The relative measure of these tori approaches one at the fixed point.

Nevertheless, the stability of a generic, elliptic fixed point is an open question. Arnol'd showed by example in 1963 that lower-dimensional tori can have unstable manifolds that intersect the stable manifolds of nearby tori and thereby allow nearby trajectories to drift "around" the n-dimensional tori; this phenomenon is called Arnol'd diffusion (Lochak, 1993). When the map is analytic, the intersection angles become exponentially small in the neighborhood of the fixed point, and the existence of connections becomes a problem in perturbation theory *beyond all orders*.

Aubry–Mather theory gives a nonperturbative generalization of KAM theory for the case of monotone twist maps when $n = 1$. These are symplectic diffeomorphisms on the cylinder $\mathbb{S} \times \mathbb{R}$ (or on the annulus) such that $\partial q'/\partial p \geq c > 0$. For this case, Aubry–Mather theory implies that there exist orbits for all rotation numbers ω. When ω is irrational, these orbits lie on a *Lipschitz graph*, $p = P(q)$, and their iterates are ordered on the graph just as the iterates of the uniform rotation by ω. They are either dense on an invariant circle or an invariant Cantor set (called a *cantorus* when discovered by Percival). These orbits are found using a Lagrangian variational principle and turn out to be global minima of the action.

Aubry–Mather theory can be partially generalized to higher dimensions, for example, to the case of rational rotation vectors, where the orbit is periodic (Golé, 2001). Moreover, Mather (1991) has shown that action-minimizing invariant measures exist for each rotation vector, though they are not necessarily dynamically minimal. The existence of invariant cantor sets with any incommensurate rotation vector can also be proven for symplectic maps near an anti-integrable limit (MacKay & Meiss, 1992). Finally, converse KAM theory, which gives parameter domains where there are no invariant circles for the standard map, implies that, for example, the Froeschlé map has no Lagrangian invariant tori outside a closed ball in the space of its parameters (a, b, c) (MacKay et al., 1989).

James D Meiss

See also **Aubry–Mather theory; Cat map; Chaotic dynamics; Constants of motion and conservation laws; Ergodic theory; Fermi acceleration and Fermi map; Hamiltonian systems; Hénon map; Horseshoes and hyperbolicity in dynamical systems; Lyapunov exponents; Maps; Measures; Mel'nikov method; Phase space; Standard map**

Further Reading

Arnol'd, V.I. 1989. *Mathematical Methods of Classical Mechanics*, New York: Springer

Forest, E. 1998. *Beam Dynamics: A New Attitude and Framework* (*The Physics and Technology of Particle and Photon Beams*), Amsterdam: Harwood Academic

Golé, C. 2001. *Symplectic Twist Maps: Global Variational Techniques*, Singapore: World Scientific

Lochak, P. 1993. Hamiltonian perturbation theory: periodic orbits, resonances and intermittancy. *Nonlinearity*, 6: 885–904

MacKay, R.S. & Meiss, J.D. 1992. Cantori for symplectic maps near the anti-integrable limit. *Nonlinearity*, 5: 149–160

MacKay, R.S., Meiss, J.D. & Stark, J. 1989. Converse KAM theory for symplectic twist maps. *Nonlinearity*, 2: 555–570

Marsden, J.E., Patrick, G.W. & Shadwick, W.F. (editiors). 1996. *Integration Algorithms and Classical Mechanics*, Providence, RI: American Mathematical Society

Mather, J.N. 1991. Action minimizing invariant measures for positive definite Lagrangian systems. *Mathematische Zeitschrift*, 207: 169–207

McDuff, D. & Salamon, D. 1995. *Introduction to Symplectic Topology*, Oxford: Clarendon Press

Meiss, J.D. 1992. Symplectic maps, variational principles, and transport. *Reviews of Modern Physics*, 64(3): 795–848

Moser, J.K. 1994. On quadratic symplectic mappings. *Mathematische Zeitschrift*, 216: 417–430

SYNAPSES

See **Neurons**

SYNCHRONIZATION

In a classical context, synchronization means adjustment of rhythms of self-sustained periodic oscillators due to their weak interaction, which can be described in terms of phase locking and frequency entrainment. The modern concept also covers such objects as rotators and chaotic systems, in which one distinguishes between different forms of synchronization, including complete, phase, and master–slave.

The history of synchronization goes back to the 17th century when the Dutch scientist Christiaan Huygens reported on his observation of synchronization of two pendulum clocks, which he had invented shortly before (Hugenii, 1673).

> . . . It is quite worth noting that when we suspended two clocks so constructed from two hooks imbedded in the same wooden beam, the motions of each pendulum in opposite swings were so much in agreement that they never receded the least bit from each other and the sound of each was always heard simultaneously. Further, if this agreement was disturbed by some interference, it reestablished itself in a short time. For a long time I was amazed at this unexpected result, but after a careful examination finally found that the cause of this is due to the motion of the beam, even though this is hardly perceptible. The cause is that the oscillations of the pendula, in

> proportion to their weight, communicate some motion to the clocks. This motion, impressed onto the beam, necessarily has the effect of making the pendula come to a state of exactly contrary swings if it happened that they moved otherwise at first, and from this finally the motion of the beam completely ceases. But this cause is not sufficiently powerful unless the opposite motions of the clocks are exactly equal and uniform.

Despite being among the oldest scientifically studied nonlinear effects, synchronization was understood only in the 1920s when Edward Appleton and Balthasar van der Pol theoretically and experimentally studied synchronization of triode oscillators.

The synchronization properties of periodic self-sustained oscillators are based on the existence of a special variable, phase ϕ. Mathematically, ϕ can be introduced as the variable parametrizing motion along the stable limit cycle in the state space of an autonomous continuous-time dynamical system. One can always choose phase in a way that it grows uniformly in time,

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \omega_0, \qquad (1)$$

where ω_0 is the natural frequency of oscillations. The phase is neutrally stable, meaning that its perturbations neither grow nor decay. This corresponds to the invariance of solutions of autonomous dynamical systems with respect to time shifts. Thus, a small perturbation (for example, an external periodic forcing or coupling to another system) can cause large deviations of the phase—contrary to the amplitude, which is only slightly perturbed due to the transversal stability of the cycle. This property allows description of the effect of small forcing/coupling via the phase approximation.

Considering the simplest case of a limit cycle oscillator driven by a periodic force with frequency ω and amplitude ε, one can write the equation for the perturbed phase dynamics in the form

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \omega_0 + \varepsilon Q(\phi, \omega t), \qquad (2)$$

where the coupling function Q is 2π-periodic in both its arguments and depends on the form of the limit cycle and the forcing. Close to the resonance $\omega \approx \omega_0$, the function Q contains fast oscillating and slow varying terms, the latter can be written as $q(\phi - \omega t)$. Upon averaging over a cycle, one obtains the following basic equation for the phase dynamics

$$\frac{\mathrm{d}\Delta\phi}{\mathrm{d}t} = -(\omega - \omega_0) + \varepsilon q(\Delta\phi), \qquad (3)$$

where $\Delta\phi = \phi - \omega t$ is the difference between the phases of the oscillations and of the forcing. The function q is 2π-periodic, and in the simplest case $q(\cdot) = \sin(\cdot)$ Equation (3) is called the Adler equation.

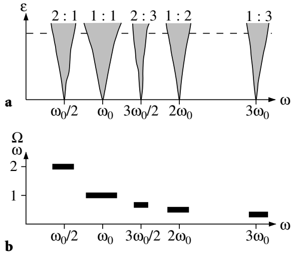

Figure 1. (a) A sketch of Arnol'd tongues. (b) Devil's staircase for a fixed amplitude of the forcing (dashed line in (a)).

One can see that on the plane of parameters of the external forcing, there is a region ($\varepsilon q_{\min} < \omega - \omega_0 < \varepsilon q_{\max}$) where Equation (3) has a stable stationary solution that exactly corresponds to phase locking (the phase ϕ just follows the phase of the forcing, so $\phi = \omega t + \text{constant}$) and frequency entrainment (the observed frequency of the oscillator $\Omega = \langle\dot{\phi}\rangle$ exactly coincides with the forcing frequency ω). Everyday examples of synchronization by external forcing are radio-controlled clocks, cardiac pacemakers, and circadian systems (the internal clocks of living objects that are synchronized to the exact 24-h periodic rhythm of sunlight).

Generally, synchronization is also observed for higher-order resonances $n\omega \approx m\omega_0$. In this case, the dynamics of the generalized phase difference $\Delta\phi = m\phi - n\omega t$ is described by an equation similar to Equation (3), namely, by $\mathrm{d}(\Delta\phi)\,/\,\mathrm{d}t = -\,(n\omega - m\omega_0) + \varepsilon\tilde{q}(\Delta\phi)$. The term *synchronous regime* then means perfect entrainment of the oscillator frequency at the rational multiple of the forcing frequency, $\Omega = n\omega\,/\,m$, as well as phase locking $m\phi = n\omega t + \text{constant}$. The overall picture can be shown on the (ω,ε) plane, where a family of triangular-shaped synchronization regions exists touching the ω-axis at the rationals of the natural frequency $m\omega_0\,/\,n$. These regions are called "Arnol'd tongues" (see Figure 1a). This picture is preserved for moderate forcing, although now the shape of the tongues generally differs from being exactly triangular.

For a fixed amplitude of the forcing ε and variable driving frequency ω, one observes different phase locking intervals where the motion is periodic, whereas in between them it is quasi-periodic. The curve Ω vs. ω, thus, consists of horizontal plateaus at all possible rational frequency ratios; this fractal curve is called a "devil's staircase" (Figure 1b).

An experimental example of such a curve is the voltage–current plot for a Josephson junction in an ac electromagnetic field, where synchronization plateaus are called Shapiro steps. As a junction can be considered as a rotator (rotations are maintained by a dc current), this example demonstrates that synchronization properties of rotators are very close to those of oscillators.

Synchronization of two coupled self-sustained oscillators can be described in a similar way. A weak interaction affects only the phases of two oscillators ϕ_1 and ϕ_2, and Equation (1) generalizes to

$$\frac{\mathrm{d}\phi_1}{\mathrm{d}t} = \omega_1 + \varepsilon Q_1(\phi_1, \phi_2),$$
$$\frac{\mathrm{d}\phi_2}{\mathrm{d}t} = \omega_2 + \varepsilon Q_2(\phi_2, \phi_1)\,. \tag{4}$$

For the phase difference $\Delta\phi = \phi_2 - \phi_1$, one obtains after averaging an equation of the type of (3). Synchronization now means that two non-identical oscillators start to oscillate with the same frequency (or, more generally, with rationally related frequencies). This common frequency usually lies between ω_1 and ω_2. Note that locking of the phases and frequencies implies no restrictions on the amplitudes, in fact, the synchronized oscillators may have very different amplitudes and wave forms; for example, oscillations may be relaxation (integrate-and-fire) or quasiharmonic.

The mutual synchronization in a large population of oscillators (the Kuramoto transition in a population of globally coupled phase oscillators, for example) can be treated as a nonequilibrium phase transition—the mean oscillating field serving as an order parameter. Examples of synchronization in large ensembles include rhythmic applause and simultaneous flashing of fireflies, adjustment of menstrual cycles in women's dormitories, and so on. Synchronization in lattices of coupled self-sustained oscillators usually sets in via formation of clusters, that is, groups of oscillators (neighbors in a lattice) having the same frequency, and with increase of coupling the clusters grow and merge.

The concept of synchronization has been extended to include chaotic systems. One effect, called phase synchronization, is mostly close to the classical locking phenomena. Indeed, many chaotic self-sustained oscillators admit determination of the instantaneous phase and the corresponding mean frequency. Often one can find a projection of the strange attractor that looks like a smeared limit cycle, so phase is then introduced as a variable that gains 2π with each rotation. These rotations are non-uniform due to chaos, which can be modeled by an effective noise in phase dynamics. If this noise is small (i.e., the rotations are rather uniform), the mean frequency of the system can be entrained by a periodic forcing while the chaos is preserved. If two or more chaotic oscillators with different natural frequencies interact, their mean frequencies can be adjusted while the amplitudes remain chaotic and only weakly correlated.

Another type of chaotic synchronization—complete synchronization—can be observed for identical chaotic systems of any type (maps, autonomous or driven time-continuous systems). In the simplest case of two diffusively coupled in all variables systems, the dynamics is described by

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{F}(\boldsymbol{x}) + \varepsilon(\boldsymbol{y} - \boldsymbol{x}),$$
$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{F}(\boldsymbol{y}) + \varepsilon(\boldsymbol{x} - \boldsymbol{y}), \tag{5}$$

where ε is the coupling parameter. The regime when $\boldsymbol{x}(t) = \boldsymbol{y}(t)$ for all t is called complete synchronization; because in this state the diffusive coupling vanishes, the dynamics is the same as if the systems were uncoupled. Although such symmetric solution exists for all ε, it is stable only if the coupling is sufficiently strong. To find the critical value of the coupling, one linearizes Equations (5) near the synchronized state and obtains for the mismatch $\boldsymbol{v}(t) = \boldsymbol{y}(t) - \boldsymbol{x}(t)$, the linearized system

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = J(t)\boldsymbol{v} - 2\varepsilon\boldsymbol{v}, \tag{6}$$

where $J(t)$ is the Jacobian at the chaotic solution $\boldsymbol{x}(t)$. The ansatz $\boldsymbol{v} = \mathrm{e}^{-2\varepsilon t}\boldsymbol{u}$ removes the last term on the right-hand side of (6), and the resulting equation coincides with the linearized equation for small perturbations of the solutions of an individual chaotic oscillator. Thus, $\boldsymbol{u}$ grows in proportion to the maximum Lyapunov exponent λ of a single system, and the critical coupling is $\varepsilon_c = \lambda / 2$. Complete synchronization occurs if $\varepsilon > \varepsilon_c$, that is, when the divergence of trajectories of interacting systems due to chaos is suppressed by the diffusive coupling. For weak coupling $\varepsilon < \varepsilon_c$, the states of two systems are different, $\boldsymbol{x}(t) \neq \boldsymbol{y}(t)$. Some other forms of synchronization in chaotic systems (generalized, master–slave) are similar to the complete one, and in all these cases synchronization appears if the coupling is strong enough.

MICHAEL ROSENBLUM AND ARKADY PIKOVSKY

See also **Commensurate-incommensurate transition; Coupled oscillators; Phase dynamics; Van der Pol equation**

Further Reading

Blekhman, I.I. 1988. *Synchronization in Science and Technology*, New York: ASME Press

Glass, L. 2001. Synchronization and rhythmic processes in physiology. *Nature*, 410: 277–284

Huygens (Hugenii), Ch. 1673. *Horologium Oscillatorium*, Paris: Apud F. Muguet; as *The Pendulum Clock*, Ames: Iowa State University Press, 1986

Kuramoto, Y. 1984. *Chemical Oscillations, Waves and Turbulence*, Berlin: Springer

Mosekilde, E., Maistrenko, Yu. & Postnov D. 2002. *Chaotic Synchronization: Applications to Living Systems*, Singapore: World Scientific

Pikovsky, A., Rosenblum, M. & Kurths, J. 2001. *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge and New York: Cambridge University Press

SYNERGETICS

Synergetics deals with the spontaneous formation of spatiotemporal or functional structures in complex open systems by means of self-organization. The word *synergetics* is taken from Greek and means "science of cooperation." The central aim of synergetics, comprising both theoretical and experimental studies, is the search for basic principles that govern self-organization in both the physical and life sciences. To this end, a comprehensive mathematical theory has been developed that comprises both deterministic and stochastic processes.

Historical Background

Synergetics was initiated by Hermann Haken in 1969 in lectures at Stuttgart University (see also Haken & Graham, 1971). It originated from laser physics, where a pronounced transition from the disordered light of a lamp to the highly ordered light of a laser takes place. This transition can be interpreted as a nonequilibrium phase transition, on the one hand (Graham & Haken, 1968, 1970), and as a typical event of self-organization, on the other. Synergetics shows common features with and differences from a variety of interdisciplinary research fields:

1. In common with cybernetics as introduced by Norbert Wiener (1948), synergetics looks for general laws common to physical and biological systems. While cybernetics focuses on the control of a system in order to achieve its specific performance, synergetics studies the various dynamical structures a complex system can acquire.
2. Synergetics shares with general system theory as introduced by Ludwig von Bertalanffy (1968) the aim of finding general laws, in particular by seeking analogies. But while von Bertalanffy was seeking such analogies between otherwise different systems at the level of their individual elements, synergetics establishes close analogies at the level of order parameters (macroscopic field variables; see below).
3. Synergetics has used and developed methods belonging to dynamical systems theory (see, e.g., Guckenheimer & Holmes, 1983; Haken, 1983). Here in particular, emphasis is laid on the qualitative changes of the behavior of dynamical systems close to their points of instability (singular points, bifurcation points). In contrast to, for example, bifurcation theory, synergetics takes into account the pivotal role of random processes (Haken, 1983).
4. Synergetics has used and developed methods from statistical physics, such as various types of (generalized) Fokker–Planck equations, Langevin equations, and master equations (see, e.g., Stratonovich, 1963, 1967).
5. Synergetics shares with the theory of dissipative structures as introduced by Ilya Prigogine the goal of general laws (Nicolis & Prigogine, 1977). But while Prigogine's theory of dissipative structures is mainly based on thermodynamic principles, such as the entropy production principle or excess entropy production principle, synergetics is based on an approach that is close to statistical physics.
6. The concepts of synergetics have similarities with ideas developed in Gestalt theory founded by Max Wertheimer and Wolfgang Köhler (1924, 1969).

Theoretical Approach

All systems are considered subject to fixed internal or external conditions that are described by control parameters α. At particular values of α the behavior of the system may change macroscopically and qualitatively ("instability" of the previous state). As synergetics shows, close to such instability points the behavior of the system is determined by a small number of dynamical quantities, the order parameters. According to the "slaving principle" of synergetics, the order parameters determine the behavior of the individual parts of the system. In turn, the individual parts generate the order parameters by their cooperation (circular causality). Close to the instability points, nonequilibrium phase transitions occur that are characterized by symmetry breaking, critical fluctuations, and critical slowing down of the order parameters. This approach requires knowledge of the microscopic dynamics (*microscopic synergetics*). If this knowledge is absent, in *phenomenological synergetics* an order parameter dynamics is postulated.

Outline of Micoscopic Synergetics

The multi-component system is described by its state vector $\boldsymbol{q}$, whose components are labeled by the subsystem ℓ and the state or component j of each of them, $q_{\ell j},\ \ell = 1, ..., N;\ j = 1, ..., J$. In continuous media, $\boldsymbol{q}$ becomes a function of a spatial coordinate $\boldsymbol{x}$, $\boldsymbol{q}(\boldsymbol{x})$. The state vector obeys an evolution equation

$$\mathrm{d}\boldsymbol{q}(t)/\mathrm{d}t = \boldsymbol{N}(\boldsymbol{q}(t), \triangledown, \boldsymbol{x}, \alpha) + \boldsymbol{F}(\boldsymbol{q}, \boldsymbol{x}, t). \quad (1)$$

$\boldsymbol{N}$ is a nonlinear vector-valued, nonlinear function of $\boldsymbol{q}$, $\triangledown$ indicates spatial derivatives, α control parameters, and $\boldsymbol{F}$ fluctuating forces. $\boldsymbol{F}$ is mostly assumed to be δ-correlated in time ("white noise").

It is assumed that for a value $\alpha = \alpha_0$, the solution to (1) is known and the stability of $\alpha_0 : \boldsymbol{q}_0(t)$ is studied by *linear stability analysis*, putting $\boldsymbol{q}(t) = \boldsymbol{q}_0(t) + \boldsymbol{\xi}(t)$ and $\boldsymbol{F} = 0$. This leads to

$$\mathrm{d}\boldsymbol{\xi}(t)/\mathrm{d}t = L(\boldsymbol{q}_0(t), \alpha)\boldsymbol{\xi}, \quad (2)$$

where the linear operator L depends on $\boldsymbol{q}_0(t)$. The solutions to (2) are of the form

$$\boldsymbol{\xi}_k(t) = \exp(\lambda_k t)\boldsymbol{v}_k(t), \quad k = 1, ... \quad (3)$$

If λ_k has a discrete spectrum, then (Haken, 1983):

(a) $\boldsymbol{q}_0$ is time-independent (fixed point); $\boldsymbol{v}_k$ is time-independent and contains powers of t if λ_k is degenerate;

(b) $\boldsymbol{q}_0$ is time-periodic (limit cycle); $\boldsymbol{v}_k$ is time-periodic with the same period and contains powers of t if λ_k is degenerate and

(c) $\boldsymbol{q}_0(t)$ is on torus, or arbitrary; $|\boldsymbol{v}_k(t)|$ increases in time more slowly than exponential.

Close to instability points, where $\mathrm{Re}\lambda_k \geq 0$ for some k, we distinguish between unstable modes, $\boldsymbol{v}_u$, and stable modes, $\boldsymbol{v}_s$. In the case of (a), the wanted solution is assumed in the form

$$\boldsymbol{q}(t) = \boldsymbol{q}_0 + \sum_u \xi_u(t)\boldsymbol{v}_u + \sum_s \xi_s(t)\boldsymbol{v}_s, \tag{4}$$

where the amplitudes ξ_u and ξ_s are still to be determined by inserting (4) into the complete Equations (1) and projecting onto the modes $\boldsymbol{v}_k$. This yields the equations

$$\mathrm{d}\boldsymbol{\xi}_u/\mathrm{d}t = \Lambda_u\boldsymbol{\xi}_u + \hat{\boldsymbol{N}}_u(\boldsymbol{\xi}_u, \boldsymbol{\xi}_s) + \boldsymbol{F}_u, \tag{5}$$

$$\mathrm{d}\boldsymbol{\xi}_s/\mathrm{d}t = \Lambda_s\boldsymbol{\xi}_s + \hat{\boldsymbol{N}}_s(\boldsymbol{\xi}_u, \boldsymbol{\xi}_s) + \boldsymbol{F}_s. \tag{6}$$

In the cases, (b) and (c), the hypothesis (4) must be extended to include phase angles, $\boldsymbol{\phi}$. In such a case, Equations (5) and (6) must be supplemented by equations for $\boldsymbol{\phi}$

$$\mathrm{d}\boldsymbol{\phi}/\mathrm{d}t = \hat{\mathbf{N}}_\phi(\boldsymbol{\xi}_u, \boldsymbol{\xi}_s, \boldsymbol{\phi}) + \boldsymbol{F}_\phi, \tag{7}$$

where $\hat{\boldsymbol{N}}_\phi$ is 2π-periodic in $\boldsymbol{\phi}$. In this case, the nonlinear functions on the right-hand side of (5) and (6) become also $\boldsymbol{\phi}$-dependent with periodicity 2π. The slaving principle of synergetics allows us to express the enslaved amplitudes $\boldsymbol{\xi}_s$ by means of the order parameters $\boldsymbol{\xi}_u, \boldsymbol{\phi}$ at the same time

$$\boldsymbol{\xi}_s(t) = \mathbf{f}(\boldsymbol{\xi}_u(t), \boldsymbol{\phi}(t), t). \tag{8}$$

The explicit time dependence stems from the action of the fluctuating forces $\boldsymbol{F}$. By means of (8), the originally high-dimensional system (5)–(7) can be reduced to a low-dimensional system

$$\mathrm{d}\tilde{\boldsymbol{\xi}}_u/\mathrm{d}t = \tilde{\boldsymbol{N}}(\tilde{\boldsymbol{\xi}}_u) + \tilde{\boldsymbol{F}}_u, \tag{9}$$

where $\tilde{\boldsymbol{\xi}}_u$ comprises both $\boldsymbol{\xi}_u, \boldsymbol{\phi}$. In continuous media, where

$$\boldsymbol{\xi}_u(t) \to \boldsymbol{\xi}_u(\mathbf{x}, t), \tag{10}$$

the order parameter equations (9) are replaced by generalized Ginzburg–Landau equations

$$\begin{aligned}\mathrm{d}\boldsymbol{\xi}_u(\mathbf{x}, t)/\mathrm{d}t = {} & \Lambda_u(\bigtriangledown)\boldsymbol{\xi}_u(\boldsymbol{x}, t) \\ & + \tilde{\boldsymbol{N}}(\boldsymbol{\xi}_u(\boldsymbol{x}, t)) + \boldsymbol{F}_u,\end{aligned} \tag{11}$$

where the matrix of eigenvalues Λ_u depends on differential operators. In higher approximation, $\tilde{\boldsymbol{N}}$ also depends on differential operators, $\tilde{\boldsymbol{N}}(\boldsymbol{\xi}_u, \bigtriangledown)$.

While Equations (1), (9), and (11) are of Langevin type, the stochastic problems can also be formulated by means of the Fokker–Planck equation, where the distribution function depends on the vector $[\boldsymbol{\xi}, f(\boldsymbol{\xi}, t)]$ and f obeys the Fokker–Planck equation

$$\begin{aligned}\partial f/\partial t &= Lf, \\ L &= -\sum_j \frac{\partial}{\partial \xi_j}(\tilde{N}_j f) \\ &\quad + \sum_{ij} Q_{ij}\partial^2/\partial\xi_i\partial\xi_j f.\end{aligned} \tag{12}$$

Furthermore, another approach is based on the master equation for the distribution function P obeying

$$\begin{aligned}\mathrm{d}P(\boldsymbol{m}, t)/\mathrm{d}t &= \sum_{\boldsymbol{m}'} w(\boldsymbol{m}, \boldsymbol{m}')P(\boldsymbol{m}', t) \\ &\quad - P(\boldsymbol{m}, t)\sum_{\boldsymbol{m}'} w(\boldsymbol{m}', \boldsymbol{m}),\end{aligned} \tag{13}$$

where $\boldsymbol{w}(\boldsymbol{m}, \boldsymbol{m}')$ are the transition probabilities from state $\boldsymbol{m}' \to \boldsymbol{m}$.

Phenomenological Synergetics

In the case where microscopic dynamics of the system are not known, the analysis is based on phenomenological equations of the type (9), where the order parameters are characterized by those experimentally determined macroscopic quantities that change qualitatively when the control parameters are changed.

Applications of Concepts and Methods of Synergetics

Physics and Chemistry

Synergetics approaches are used to study stochastic properties and spatiotemporal patterns of laser light and of fluids and plasmas, current distributions in semiconductors, crystal growth, and meteorological structures, for example, baroclinic instability.

Synergetics is also used to study the formation of spatiotemporal patterns at macroscopic scales in chemical reactions, for example the Belousov–Zhabotinsky reaction.

Biology

Based on Turing's ideas of morphogenesis, synergetics calculates spatial density distributions, in particular gradients, stripes, hexagons, etc., in dependence on boundary and initial conditions. In initially undifferentiated omnipotent cells, molecules are produced as activators or inhibitors that diffuse between cells and react with each other and thus can be transformed. At places of high concentration, the activator molecules switch on genes that, eventually, lead to cell differentiation.

By means of synergetics, new kinds of analogies between evolution in biological and physical systems have been unearthed. For instance, the equations established by Manfred Eigen (1971) for prebiotic, that is, molecular evolution, turn out to be isomorphic to specific rate equations for laser light (photons), where a specific kind of photon wins the competition between different kinds.

In population dynamics, the resources, such as food, nesting places for birds, or light intensity for plants, serve as control parameters for synergetic analyses of self-organization. The numbers or densities of the individuals of species serve as order parameters. Specific examples are provided by the Verhulst equation or the predator-prey relation of the Lotka–Volterra equations. Of particular interest are dramatic changes, for instance the dying out of species under specific control parameter values. This has influences on environmental policy. If specific control parameters exceed critical values, the system's behavior can change dramatically. For instance, beyond a specific degree of evolution, the fish population of a lake may die out.

Nearly all biological systems show more or less regular rhythms—periodic oscillations or fluctuations. These can be imposed on the system from the outside, for instance, by the day/night cycle or seasons (exogen), or produced by the system itself (endogen). Endogenous rhythms that may proceed on quite different spatial and temporal scales are widely researched in synergetics. Examples are cell metabolism, circadian rhythms, brain waves in different frequency bands (see below), menstrual cycles, and cardiovascular rhythms. For instance, in the last, Stefanovska et al. (2000) were able to identify five order parameters. For certain time intervals, these order parameters can show phase and frequency couplings.

Rhythmical movements of humans and animals show well-defined patterns of coordination of the limbs, for instance, walking or running in humans or gaits of quadrupeds. Synergetics studies especially transitions between movement patterns, for instance the paradigmatic experiment by Kelso (1995). If subjects move their index fingers in parallel at a low frequency; increasing the frequency results in an abrupt involuntary transition to a new symmetric movement. The control parameter is the prescribed finger movement frequency, the order parameter is the relative phase between the index fingers. The experimentally proven properties of a nonequilibrium phase transition (critical fluctuations, critical slowing down, hysteresis) substantiate the concept of self-organization and exclude that of a fixed motor program. Numerous further coordination experiments between different limbs can be represented by the Haken–Kelso–Bunz model (Haken et al., 1985). Gaits of quadrupeds and transitions between them have been modeled in detail (Schöner et al., 1990).

In visual perception, the recognition of patterns, for example, faces, is interpreted as the action of an associative memory in accordance with usual approaches. Here incomplete data (features) with which the system is provided from the outside are complemented by means of data stored in the memory. A particular aspect of the synergetic approach is the idea that pattern recognition can be conceived as pattern formation. This is not only meant as a metaphor, but means also that specific activity patterns in the brain are established. In pattern formation, a partly ordered pattern is provided to ths system whereby several order parameters are evoked that compete with each other dynamically. The control parameters are so-called attention parameters that in cases without bias are assumed to be equal. The winning order parameter imposes a total pattern on the system according to the slaving principle. This process is the basis of the synergetic computer for pattern recognition (Haken, 1991). By means of appropriate preprocessing an invariance of the recognition process against scales, displacements and rotations, and even deformations can be achieved.

Gestalt Psychology

As is shown in Gestalt psychology (Wertheimer, Köhler), *Gestalt* is conceived as a specific organized entity to which in synergetics an order parameter with its synergetic properties (slaving principle) can be attached. The cognition or perception process of Gestalt proceeds in principle according to the synergetic process of pattern recognition. The winning order parameter generates, according to the slaving principle, an ideal percept that is the corresponding Gestalt. In ambiguous patterns, an order parameter is attached to each percept of an object. Because in ambiguous figures two or more possible interpretations are contained, several order parameters participate in the dynamics whereby the attention parameters become dynamical quantities. As already assumed by Köhler and as is shown by the synergetic equations, the corresponding attention parameter saturates; that is, it becomes zero if the corresponding object has been recognized and the other interpretation now becomes possible, where again the corresponding saturation process starts, etc. The model equations allow us also to take into account bias. (*See* the article on **Gestalt phenomena**.)

Psychology

According to the concept of synergetics, psychological behavioral patterns are generated by self-organization of neuronal activities under specific control parameter conditions and are represented by order parameters. In important special cases, the order parameter dynamic can be represented as the overdamped motion of a ball in mountainous terrain. By means of changes in the control parameters, this landscape is deformed and allows new equilibrium positions (stable behavioral patterns). This leads to new approaches to psychotherapy: destabilization of unwanted behavioral patterns by means, for example, of new external conditions or new cognitive influences and measures that support the self-organization of desired behavioral patterns. The insights of synergetics have been applied in the new field of psychosynergetics with essential contributions by Schiepek, Tschacher, Hansch, and others (Tschacher et al., 1992; Hansch, 1997; Ciompi, 1998, 1999).

Brain Theory

According to a proposal by Haken (1983), the brain of humans and animals is conceived as a synergetic, that is, a self-organizing system. This concept is supported by experiments and models on movement coordination, visual perception, and Gestalt psychology and by EEG and MEG analysis (see below). The human brain with its 10^{11} neurons (and glia cells) is a highly interconnected system with numerous feedback loops. In order to treat it as a synergetic system, control parameters and order parameters must be identified. While in synergetic systems of physics, chemistry, and partly biology the control parameters are fixed from the outside, for instance, by the experimenter, in the brain and in other biological systems, the control parameters can be fixed by the system itself. In modeling them it is assumed, however, that they are practically time-independent during the self-organization process. Such control parameters can be, among others, the synaptic strengths between neurons that can be changed by learning according to Hebb (1949), neurotransmitters, such as Dopamin, Serotonin, and drugs that block the corresponding receptors (e.g., Haloperidol, Coffein), and hormones (influencing the attention parameters). Furthermore, the control parameters may be more or less permanent external or internal stimuli.

In the frame of the given control parameters, self-organization takes place in neuronal activity whereby the activity patterns are connected with the corresponding order parameters by means of circular causality. The order parameters move for a short time in an attractor landscape whereby the attractor and also the order parameter disappear (concept of quasi-attractors). An example is the disappearance of a percept in watching ambiguous figures. The origin and disappearance of quasi-attractors and the corresponding order parameters can happen on quite different time scales, so that some of them can act as attractors practically all the time or are hard to be removed (psychotherapy in the case of behavioral disturbances). The activity patterns can be stimulated by external stimuli (exogenous activity) but can also be created spontaneously (endogenous activity), for instance in dreams, hallucinations, and, of course, thinking.

Synergetics throws a new light on the mind–body problem, for instance the percepts are conceived as order parameters, whereas the parts of a system are represented by electrochemical activities of the individual neurons. Because of circular causality, the percepts as order parameters and the neural activity (the "enslaved parts") condition each other. Beyond that, the behavior of a system can be described at the level of order parameters (information compression) or at the level of the activities of individual parts (large amount of information).

Analysis of Electroencephalograms (EEG) and Magnetoencephalograms (MEG)

Neuronal activity is accompanied by electromagnetic brain waves that cover the brain over large areas. The corresponding electric and magnetic fields are measured by the EEG and MEG, respectively. According to the ideas of synergetics, at least in situations where the macroscopic behavior changes qualitatively, the activity patterns should be connected with few order parameters. Typical experiments are the above-described finger coordination experiments by Kelso and closely related experiments, for instance, the coordination between the movement of a finger and a sequence of acoustic signals. In a typical experiment, parts of the brain or the whole brain are measured by an array of SQUIDS (superconducting quantum interference devices) that allows the determination of spatiotemporal field patterns. By means of appropriate procedures, these patterns are decomposed into fundamental patterns. As the analysis shows, two dominant basic patterns appear, whose amplitudes are the order parameters. If the coordination between finger movement and the acoustic signal changes dramatically, the dynamics of the order parameters also does so.

Sociology

Here we may distinguish between the more psychological and the more systems theoretical schools, where synergtics belongs to the second approach. We can distinguish between a qualitative and a quantitative synergetics (for a quantitative approach, see Weidlich, 2000). In the latter case, a number of sociologically relevant order parameters are identified. One example is the language of a nation. After his/her birth, a baby is exposed to the corresponding language and learns it (in technical terms of synergetics: the baby is enslaved) and then carries on this language as an adult (circular causality). These language order parameters may compete, where one wins (e.g., in the United States, the English language), they may coexist (e.g., in Switzerland), or they may cooperate (for instance, popular language and technical language). Whereas in this case the action of the slaving principle is evident, in the following examples its applicability is critically discussed by sociologists so that instead of slaving, some sociologists like to speak of binding or consensualization. Corresponding order parameters are type of state (e.g., democracy, dictatorship), public law, rituals, corporate identity, social climate in a company, and ethics. The latter example is particularly interesting, because order parameters are not prescribed from the outside or ab initio, but originate through self-organization and need not be uniquely determined.

Ethics

Conceived as order parameter means that it originates from the formation of a consensus so that the slaving principle becomes valid and also that there may exist several ethics.

Epistemology

An example for order parameters is provided by the scientific paradigms in the sense of Thomas S. Kuhn (1970), where a change of paradigms has the properties of a nonequilibrium phase transition, such as critical fluctuations and critical slowing down. Synergetics as a new scientific paradigm is evidently self-referential. It explains its own origin.

Management

The concept of self-organization is increasingly used in management theory and management praxis. Instead of fixed order structures with many hierarchical levels, now flat organizational structures with new hierarchical levels are introduced. In the latter case, a hierarchical level makes its decisions by means of its distributed intelligence. For an indirect steering of these levels by means of a higher level, specific control parameters in the sense of synergetics must be fixed, for instance, by fixing special conditions and goals. The order parameters are, for instance, the self-organized collective labor processes. In this context, the slaving principle—according to which the order parameters change slowly, whereas the enslaved parts react quickly (adaptability)—gains a new interpretation. For instance, the employees that are employed for a longer time determine the climate of labor and the style of work whereby it can also be possible that undesired cliques are established. This trend can be counteracted by job rotation.

Development of Cities

So far the development of cities was based on the concept of city planning with detailed plans for areas, but new approaches use concepts of self-organization according to synergetic principles. Instead of a detailed plan, now specific control parameters, such as a general infrastructure (streets, communication centers, and so on), are fixed. For details consider the book by Portugali (1999).

HÉRMANN HAKEN

See also **Dynamical systems; Emergence; Gestalt phenomena; Lasers; Turing patterns**

Further Reading

von Bertalanffy, L. 1968. *General System Theory*, New York: Braziller and London: Allen Lane, 1971

Ciompi, L. 1999. *Emotionale Grundlagen des Denkens*, Göttingen: Van den Hoeck; 1998. *Affektlogik*. Stuttgart: Klett-Cotta

Eigen, M. 1971. Molekulare Selbstorganisation und Evolution (Self-organization of matter and the evolution of biological macro molecules). *Naturwissenschaften*, 58: 465–523

Graham, R. & Haken, H. 1968. Quantum theory of light propagation in a fluctuating laser-active medium. *Zeitschrift für Physik*, 213: 420–450

Graham, R. & Haken, H. 1970. Laser light—first example of a second-order phase transition far away from thermal equilibrium. *Zeitschrift für Physik*, 237: 31–46

Guckenheimer, J. & Holmes, P.J. 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Berlin: Springer

Haken, H. 1983. Synopsis and introduction. In *Synergetics of the Brain*, edited by, E. Başar, H. Flohr, H. Haken & A.J. Mandell, Berlin: Springer, pp. 3–25

Haken, H. 1991. *Synergetic Computers and Cognition*, Berlin: Springer

Haken, H. 1993. *Advanced Synergetics*, 3rd edition, Berlin: Springer

Haken, H. & Graham, R. 1971. Synergetik—Die Lehre vom Zusammenwirken. *Umschau*, 6: 191

Haken, H., Kelso, J.A.S. & Bunz, H. 1985. A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51: 347–356

Hansch, D. 1997. *Psychosynergetik*, Wiesbaden: Westdeutscher Verlag

Hebb, D.O. 1949. *The Organization of Behavior*, New York: Wiley

Kelso, J.A.S. 1995. *Dynamic Patterns: The Self-organization of Brain and Behavior*, Boston: MIT Press

Köhler, W. 1924. *Die physischen Gestalten in Ruhe und im stationären Zustand*, Erlangen: Philosophische Akademie

Köhler, W. 1969. *The Task of Gestalt Psychology*, Princeton, NJ: Princeton University Press

Kuhn, T.S. 1970. *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press

Nicolis, G. & Prigogine, I. 1977. *Self-organization in Non-equilibrium Systems*, New York: Wiley

Portugali, J. 1999. *Self-organization in the City*, Berlin: Springer

Schöner, G., Yiang, W.Y. & Kelso, J.A.S. 1990. A synergetic theory of quadrupedal gaits and gait transitions. *Journal of Theoretical Biology*, 142: 359–391

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Stefanovska, A., Haken, H., McClintock, P.V.E., Hožič, M., Bajrović, F. & Rivarič, S. 2000. Reversible transitions between synchronization states of the cardiorespiratory system. *Physical Review Letters*, 85(22): 4831–4834

Stratonovich, R.L. 1963, 1967. *Topics in the Theory of Random Noise*, vols. 1, 2, New York: Gordon and Breach

Tschacher, W., Schiepek, G. & Brunner, E.J. 1992. *Self-organization and Clinical Psychology*, Berlin: Springer

Weidlich, W. 2000. *Sociodynamics. A Systematic Approach to Mathematical Modelling in the Social Sciences*, Amsterdam: Harwood Academic Publishers

Wiener, N. 1948. *Cybernetics, or Control and Communication in the Animal and the Machine*, Cambridge, MA: MIT Press

T

TACHYONS AND SUPERLUMINAL MOTION

With a name derived from the Greek *tachys* (for swift), the tachyon is a hypothetical elementary particle that travels at speeds exceeding that of light (superluminal speed). In his 1905 paper on special relativity, Albert Einstein showed that the light speed in vacuum (c) is invariant with respect to all inertial observers, constituting a limiting value for the speed (v) of a moving mass. Consequently, tachyon research was delayed until the 1950s and 1960s, in particular till the appearance of the papers by Sudarshan and coworkers (Bilaniuk et al., 1962), and later by Recami and coworkers, among others.

If the special relativity theory is not a priori restricted to subluminal speeds, however, it seems able to accommodate tachyons. Such an extended relativity (ER) is based on the ordinary postulates of special relativity, and, therefore, does not appear to imply violations of causality (Recami, 1986). Just as photons are born, live, and die, always at the speed of light (without any need of accelerating from rest to the light speed), particles or waves may exist endowed always with speeds v larger than c.

Several areas of experimental physics suggest superluminal group velocities and energy propagation although it is not yet clear whether these phenomena imply signal and information transmissions with faster-than-light speeds (Recami, 2001).

Evanescent Waves and Tunneling Photons

In quantum mechanics, the tunneling time does not depend on the potential barrier width, implying arbitrarily large group velocities inside long barriers (Olkhovsky et al., 1992). Analogously, evanescent electromagnetic waves that travel with superluminal speeds have been predicted by extended relativity, confirmed by computer simulations based on Maxwell equations, and empirically observed (Chiao et al., 1993).

The most interesting experiment of this series was performed in 1994 by Günter Nimtz and his colleagues using two classical barriers separated by an intermediate region of width R. For nonresonant tunneling of microwaves, it was claimed that the total crossing time did not depend on R, implying the speed of transmission to be therein practically infinite. This result agrees with theoretical predictions for nonresonant tunneling through two successive opaque barriers (Olkhovsky et al., 2002) that were experimentally confirmed and have been verified by an experiment with two gratings in an optical fiber, suggesting potentially important applications.

Superluminal Localized Solutions (SLS) to the Wave Equation

Although the simplest subluminal object is a small sphere (or a point), a result of ER is that the simplest superluminal objects appear as "X-shaped" waves (double cones), rigidly moving in a homogeneous medium. Beams of this sort have been constructed (as superpositions of Bessel beams) in experiments with acoustic waves (Lu and Greenleaf, 1992), electromagnetic waves, and visible light (Saari & Reivelt, 1997). A single Bessel beam is the following solution of the

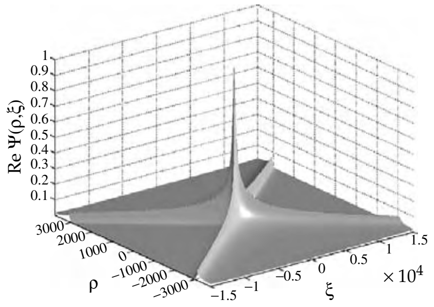

Figure 1. Illustration of the real part of a classical X-shaped wave (as a function of $\zeta \equiv z - vt$ and the radial coordinate ρ), evaluated for $v = 5\,c$ and $a = 5 \times 10^{-7}$ m, plotted from Equation (3).

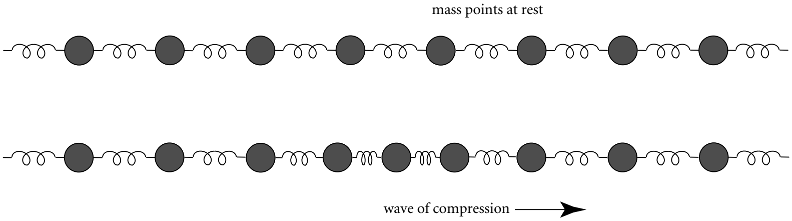

Figure 2. A spring–mass system with nonlinear springs. (Upper) Masses at rest. (Lower) A supersonic compression wave.

Maxwell equations in vacuum:

$$\psi(\rho,\zeta) = J_0[(\omega\rho/c)\sin\theta]\ \exp[\mathrm{i}(\omega\zeta/c)\cos\theta], \tag{1}$$

where J_0 is a Bessel function of the first kind of order zero, $\zeta \equiv z - vt$, $v = c/\cos\theta$, and θ is the axicone angle (angle between the direction of propagation and the propagating cone). This solution depends on z and t only through the quantity ζ; thus, it propagates without dispersion along the z-axis with speed $v > c$. By superposition of Bessel beams, for example, with variable angular frequency ω but constant θ and, therefore, constant v, one can obtain additional solutions to the Maxwell equations with any degree of transverse and longitudinal localization and centered at the desired angular frequency ω with the desired bandwidth (Zamboni-Rached et al., 2002); thus

$$\Psi(\rho,\zeta) = \int_0^\infty S(\omega) J_0[(\omega\rho/c)\sin\theta] \times \exp[\mathrm{i}(\omega\zeta/c)\cos\theta]\,\mathrm{d}\omega. \tag{2}$$

By choosing the exponential spectrum $S(\omega) = \exp[-a\omega]$, one gets the ordinary X-wave (with $v \equiv c/\cos\theta > c$)

$$\Psi(\rho,\zeta) = v\left[\sqrt{(av - \mathrm{i}\zeta)^2 + \rho^2(v^2/c^2 - 1)}\right]^{-1}, \tag{3}$$

which is a superluminal localized solution (SLS) to the wave equation. The plot of Equation (3) in Figure 1 can be understood as follows. For $\zeta < 0$, one has the familiar bow wave of a moving boat or shock wave of a supersonic aircraft. The branches for $\zeta > 0$, on the other hand, prepare the medium for the superluminal disturbance.

SLSs have also been constructed which travel without distortion along cylindrical waveguides and coaxial cables (Zamboni et al., 2002). X-shaped waves keep their localization and superluminality properties only to a certain depth of field (whose length can be determined a priori), decaying abruptly thereafter. As suggested by extended relativity, the simplest means for experimentally producing SLSs employs dynamic antennas consisting of a set of circular rings (or axicons or holographic elements). Acoustic localized supersonic beams have been used in a 3-dimensional ultrasound scanner for medical purposes: high-resolution scanning of the heart (Lu & Greenleaf, 1992).

Tachyons and SLSs in Nonlinear Media

Localized solutions for nonlinear partial differential equations that travel faster than the limiting speed for small amplitude waves are not uncommon (Conti et al., 2003). The sine-Gordon equation, for example, is Lorentz invariant and has tachyonic soliton solutions (Scott, 2003). Although dynamically unstable over long distances, these solutions may influence dynamics over shorter spans.

Another well-known example of a SLS goes back to the birth of nonlinear science, when John Scott Russell measured the speed of a hydrodynamic solitary wave in a wave tank to be $\sqrt{g(d+h)}$, where g is the acceleration of gravity, d the depth of the channel, and h the wave height. Clearly, the solitary wave speed is larger than the speed $\sqrt{gd}$ of low-amplitude waves.

Similar phenomena are observed for lattice solitary waves on spring-mass ladders (see Figure 2), which include the Fermi–Pasta–Ulam system, the Toda lattice, and various generalizations: Existence of supersonic compression-wave solutions has been mathematically proven for a general class of nonlinear intermass potentials (Friesecke & Wattis, 1994). As an example of superluminal propagation that was well known to military engineers in the early 19th century, Russell cited the fact that "the sound of a cannon travels faster than the command to fire it".

ERASMO RECAMI AND ALWYN SCOTT

See also **Cherenkov radiation; Sine-Gordon equation; Skyrmions; Solitons, a brief history**

Further Reading

Bilaniuk, O.M, Deshpande, V.K. & Sudarshan, E.C.G. 1962. Meta relativity. *American Journal of Physics*, 30: 718

Chiao, R.Y., Kwiat, P.G. & Steinberg, A.M. 1993. Faster than light. *Scientific American*, 269(2): 52–60

Conti, C., Trillo, S., Di Trapani, P., Valiulis, G., Piskarskas, A., Jedrkiewicz, O. & Trull, J. 2003. Nonlinear electromagnetic X-waves. *Physical Review Letters*, 90: no.170406

Friesecke, G. & Wattis, J.A.D. 1994. Existence theorem for travelling waves on lattices. *Communications in Mathematical Physics*, 161: 391–418

Lu, J.-Y. & Greenleaf, J.F. 1992. Experimental verification of nondiffracting X-waves. *IEEE Transactions on Ultrasonic Ferroelectric Frequency Control*, 39: 441–446

Olkhovsky, V.S. et al. 1992. Recent developments in the time analysis of tunneling processes. *Physics Reports*, 214: 339–356 and references therein

Olkhovsky, V.S. et al. 2002. Superluminal tunneling through two successive barriers. *Europhysics Letters*, 57: 879–884

Recami, E. 1986. Classical tachyons and possible applications. *Rivista del Nuovo Cimento*, 9(6): 1–178 and references therein

Recami, E. 2001. Superluminal motions? A bird's-eye view of the experimental situation. *Foundations of Physics*, 31: 1119–1135

Saari, P. & Reivelt, K. 1997. Evidence of X-shaped propagation-invariant localized light waves. *Physical Review Letters*, 79: 4135–4138

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures.*, 2nd edition, Oxford and New York: Oxford University Press

Zamboni-Rached, M. et al. 2002. New localized superluminal solutions to the wave equations with finite total energies and arbitrary frequencies. *European Physical Journal D*, 21: 217–228

Zamboni, M. et al. 2002. Superluminal X-shaped beams propagating without distortion along a coaxial guide. *Physical Review* E, 66: no. 046617

TACOMA NARROWS BRIDGE COLLAPSE

The Tacoma Narrows suspension bridge in Washington state was opened to traffic on July 1, 1940. At the time, it was the latest in a trend in suspension bridge design toward building ever longer, lighter, and more flexible bridges. At least two other suspension bridges in America, the Golden Gate bridge in San Francisco and the Bronx–Whitestone bridge in New York, had at that time, already experienced some problems with unwanted oscillations; these bridges were stabilized by adding dampers, stiffening girders, and additional cables. The Tacoma Narrows bridge was far lighter and more flexible than its predecessors.

From its opening, the bridge experienced relatively small vertical oscillations. These were observed, and carefully recorded and the record survives in Amann et al. (1941), which is our primary source for what follows.

These oscillations were purely vertical, with no torsional component. They could be anything from no-noded (with a comparatively low frequency of about 8 cycles/min) to as many as seven-noded (with a frequency of 30 cycles/min). Amplitudes were as much as 5 f from top to bottom. Motions as large as 4 f with a frequency of 16 min^{-1} were observed in winds of 3 mi/h, while at other times, the bridge remained stationary in winds of 35 mi/h. The structure had already survived winds of 48 mi/h.

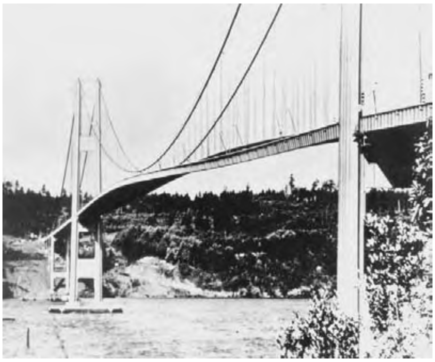

Figure 1. The Tacoma Narrows Bridge twisting, November 7, 1940. (Courtesy, Manuscripts, Special Collections, University Archives, University of Washington Libraries, UW2143.)

These motions were not considered grounds for alarm, and it was apparently expected that with some modifications, they would eventually be eliminated. By October, hold-down cables had been installed in the side-spans, and these had effectively stopped the oscillations in these spans although, of course, they did not effect the center span. Hold-down cables for the center span had already been ordered.

On the morning of November 7, 1940, a qualitatively different phenomenon occurred. The wind was about 42 mi/h. The bridge began oscillating with somewhat more violence than usual in the vertical direction. The motion appears to have been eight- or nine-noded, with an amplitude of 4–5 f and a frequency of about 36–38 cycles/min. Although this seems rather violent, it was not initially viewed as cause for alarm. People and cars were on the bridge.

Suddenly, without warning, and virtually instantaneously, the bridge switched from the vertical motion to the famous torsional motion most often associated with the bridge (see Figure 1).

This new motion was one-noded, with an angle of rotation of nearly 45° each way, and a frequency of about 14/min. The sides of the bridge were moving with a double amplitude of 28 f, with accelerations in excess of gravity. Amazingly, this motion continued for about 45 min, after which the bridge finally began to break up. Occasionally, it would switch from one-noded to no-noded and back, but it remained primarily one-noded. The side-spans remained motionless, until the collapse of the center span, after which a torsional oscillation built up and then died down.

Thus, we are presented with at least three distinct nonlinear phenomena.

(i) The vertical oscillations: why so many and under such widely different wind conditions?
(ii) The instantaneous transition from vertical to torsional motion.
(iii) The persistence of large torsional periodic oscillation for 45 min under comparatively small aerodynamic forcing.
(iv) The peculiar shifting from one-noded torsional motion to no-noded and back.

A commission, including Othmar Amann and Theodore von Karman, tried to investigate the cause of the collapse. They made some preliminary conclusions, including "it is very improbable that resonance with alternating vortices plays an important role in the oscillations of suspension bridges." However, the main recommendation was simply to avoid designs that were light and flexible, This advice has largely been followed since that time, with the notable exception of the Millennium Bridge over the Thames in London. Moreover, other light, flexible, long-span bridges such as the Golden Gate were substantially re-engineered at great cost in the 1950s to eliminate similar unwanted oscillations.

Recent advances in nonlinear science have cast light on some of the phenomena mentioned above. The vertical oscillations may have been the result of the existence of multiple periodic solutions of a nonlinearly supported beam equation (Lazer & McKenna, 1990). They may also have been the result of some sort of negative damping induced by the aerodynamic interaction of the wind and the structure. Wind tunnel experiments on scale models (Scanlan & Tomko, 1971), predict the existence of small torsional oscillations (e.g., in torsion $0 \leq \alpha \leq \pm 3°$). However, this type of motion was never observed on the Tacoma Narrows.

The violent transition from vertical to torsional motion is now well understood. In McKenna (1999), a two-degree-of-freedom oscillator with physical constants chosen to match the bridge exhibits exactly this type of instability, as soon as the cables start alternately slackening. Also in McKenna (1999), this oscillator shows how small torsional forcing can induce large torsional periodic solutions, once one takes into account the pendulum-like nonlinearity of the torsional oscillator. The solutions are remarkably similar, in terms of amplitude and frequency, to the behavior observed on the Tacoma Narrows.

The transition from one-noded to no-noded and back has been observed in Moore (2002), where the torsional oscillations in a nonlinear beam are shown to obey a sine-Gordon equation whose long-term solutions are investigated. One can also see a numerical simulation of the bridge at www.math.lsa.umich.edu/ksmoore.

A good source of interesting behavior of early suspension bridges is Bleich et al. (1950).

JOE MCKENNA

See also **Bifurcations; Distributed oscillators; Stability**

Further Reading

Amann, O.H., von Karman, T. & Woodruff, G.B. 1941. *The Failure of the Tacoma Narrows Bridge*, Federal Works Agency Washington, D.C.

Bleich, F., McCullough, C.B., Rosecrans, R. & Vincent, G.S. 1950. *The Mathematical Theory of Suspension Bridges*, U.S. Dept. of Commerce, Bureau of Public Roads Washington, D.C.

Lazer, A.C. & McKenna, P.J. 1990. Large amplitude periodic oscillations in suspension bridges: some new connections with nonlinear analysis. *SIAM Review*, 32: 537–578

McKenna, P.J. 1999. Large torsional oscillations in suspension bridges revisited: fixing an old approximation. *American Mathematics Monthly*, 106: 1–18

Moore, K.S. 2002. Large torsional oscillations in a suspension bridge: multiple periodic solutions to a nonlinear wave equation. *SIAM Journal of Mathematical Analysis* 33(6): 1411–1429

Scanlan, R.H. & Tomko, J.J. 1971. Airfoil and bridge deck flutter derivatives. *ASCE, Journal of the Engineering Mechanics Division*, EM6: 1717–1737

TANGENT SPACE

See **Differential geometry**

TAYLOR–COUETTE FLOW

Taylor–Couette flow is generated in the gap between a pair of concentric cylinders by the rotary motion of one or both of the cylinders. If the outer cylinder alone is rotated, the torque induced by viscous drag on the inner one can be used to measure the viscosity of the fluid in the gap. This technique (studied by Arnulph Mallock in 1888 and independently by M.M. Couette in 1890) is still in use today as the basis of some commercial viscometers.

In 1916, Lord Rayleigh (John William Strutt) proposed general arguments on the stability of rotating fluids, showing that when the centrifugal force gradient decreases outwards in a rotating body of fluid, this will be an unstable situation for an inviscid fluid. In the case of Mallock's viscometer, the flow will be stable. But suppose the inner cylinder is rotated and the outer is held fixed. Here the flow will be unstable according to Rayleigh's criterion, and this question intrigued Geoffrey Taylor. In 1923, he published a brilliant theoretical and experimental investigation of a viscous version of the Rayleigh criterion, using rotating cylinders to generate what has come to be called Taylor–Couette flow. Taylor's work forms the cornerstone of much modern research on hydrodynamic stability

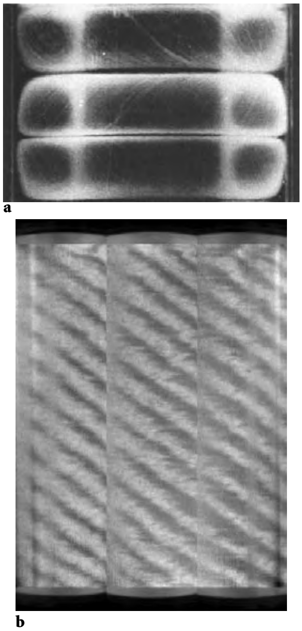

Figure 1. (a) Front view of steady Taylor–Couette cells (G. Pfister). (b) Front view of turbulent spiral flow (O. Dauchot).

and contains important ideas such as the exchange of stability between states.

In the simplest case, the outer cylinder is stationary and the inner cylinder alone rotates. When a critical Reynolds number is exceeded, the initially featureless rotary Couette flow has secondary vortices superposed on it. An example of such a Taylor vortex state is shown in Figure 1(a) where a front view of a Taylor vortex state is shown. The pattern is repeated along the length of the cylinder and has the appearance of a set of doughnuts stacked along the length the inner cylinder.

Increasing the speed of the inner cylinder leads to an instability (Davey et al., 1968) in the form of traveling waves that move at a fixed fraction of the speed of the inner cylinder (Coles, 1965). Yet, further increases in rotation rate give rise to more complicated quasi-periodic motion until low-dimensional temporal chaos ensues within each Taylor cell (Gollub & Swinney, 1975). Nowadays, it is recognized that many routes to chaos exist within this flow (Tagg, 1994), which has proved to be a rich research field for dynamical systems and bifurcation theory.

Taylor considered the general case with both cylinders rotating in co- and counter-directions, obtaining remarkable agreement between theory and experiment for the stability boundary. When the cylinders are rotated in opposite directions, new states of spiraling turbulence are sometimes found (Coles, 1965). An example of such a flow from a more recent study (Prigent et al., 2002) is shown in Figure 1(b). The general case of co- and counter-rotation produces a plethora of interesting dynamical states (Andereck et al., 1986) that can now be classified in terms of equivariant bifurcation theory (Tagg, 1994).

Coles also demonstrated dynamical non-uniqueness when the outer cylinder is stationary, finding more than 20 states at a particular Reynolds number. This multiplicity of solutions has also been observed in the steady vortex states (Burkhalter & Koschmieder, 1973; Benjamin & Mullin, 1982) where evidence was found for 42 different solutions at a single point in parameter space in the latter study.

Most of the above work is based on the notion of a theoretical model where the cylinders are considered to be infinitely long. This permits the use of periodic boundary conditions which allows analytical and numerical progress to be made in the nonlinear regime. However, the importance of end effects and their influence on global aspects of the steady flows was first recognized by Benjamin in 1978. He showed their importance in the selection process for steady flows and discovered the so called "anomalous modes" where all cells rotate in the opposite direction. These modes form an essential part of the solution set. The ideas have been developed and the role of end effects in determining the dynamics is now understood in some detail (Abshagen et al., 2001).

One clear advantage of studying fluid dynamical systems such as Taylor–Couette flow is that quantitative comparisons can be made between the results of controlled laboratory experiments and numerical calculations of the governing equations of motion—the Navier–Stokes equations (Cliffe et al., 2000). This is so not only for steady flows but also for time-dependent states, although disordered motion remains an outstanding challenge.

TOM MULLIN

See also **Bifurcations; Catastrophe theory; Chaos vs. turbulence; Hopf bifurcation; Stability**

Further Reading

Abshagen, J., Pfister, G. & Mullin, T. 2001. Gluing bifurcations in a dynamically complicated extended fluid flow. *Physical Review Letters*, 87: 4501–4505

Andereck, C.D., Liu, S.S. & Swinney, H.L. 1986. Flow regimes in a circular Couette system with independently rotating cylinders. *Journal of Fluid Mechanics*, 164: 155–183

Benjamin, T.B. & Mullin, T. 1982. Notes on the multiplicity of flows in the Taylor experiment. *Journal of Fluid Mechanics*, 121: 219–230

Benjamin, T.B. 1978. Bifurcation phenomena in steady flows of a viscous fluid. *Proceedings of the Royal Society of London*, A359, 1–43

Burkhalter, J.E. & Koschmieder, E.L. 1973. Steady supercritical Taylor vortex flow. *Journal of Fluid Mechanics*, 58: 547–560

Cliffe, K.A., Spence, A. & Tavener, S.J. 2000. The numerical analysis of bifurcation problems with application to fluid mechanics. *Acta Numerica*, 9: 39–131

Coles, D. 1965. Transition in circular Couette flow. *Journal of Fluid Mechanics*, 21: 385–425

Couette, M.M. 1890. Etudes sur le frottement des liquides. *Annales de Chimie et de Physique*, 6: 433–510

Davey, A., diPrima, R.C. & Stuart, J.T. 1968. On the instability of Taylor vortices. *Journal of Fluid Mechanics*, 31: 17–52

Gollub, J. & Swinney, H.L. 1975. Onset of turbulence in a rotating fluid. *Physical Review Letters*, 35: 927–930

Mallock, A. 1888. Determination of the viscosity of water. *Proceedings of the Royal Society of London*, A45: 126–132

Prigent, A., Gregoire, G., Chate, H., Dauchot, O. & van Saarloos, W. 2002. Large-scale finite-wavelength modulation within turbulent shear flows. *Physical Review Letters*, 89: 1501–1504

Rayleigh, Lord. 1916. On the dynamics of revolving fluids. *Proceedings of the Royal Society of London*, A93: 148–154

Tagg, R. 1994. The Couette–Taylor problem. *Nonlinear Science Today*, 4: 2–25

Taylor, G.I. 1923. Stability of a viscous liquid contained between two rotating cylinders. *Philosophical Transactions of the Royal Society of London*, 223: 289–343

TENSORS

Tensors are mathematical representations of objects that have intrinsic, geometric significances. This is a rather wider definition than many which can be found in textbooks, often referring to "sets of quantities" that "transform according to hideous formulae." The best way to understanding is likely to follow a few examples.

Readers interested in learning more will find the books of Simmon (1994) and Dodson & Poston (1991) helpful, as well as that of Bishop & Goldberg (1968). The main caveat, at the risk of repetition, is to beware of books that define tensors as some set of quantities tied to a coordinate system and then describe how they transform—tensors are there whether or not you have coordinates.

Suppose you have some apples on a table. The quantity of apples present is described completely by one number, and that number is well defined and meaningful no matter how you might orient any coordinate axes in the room. The fact that it is meaningful independent of the coordinates makes it a tensor, and the fact that it is the same no matter what coordinates might be introduced makes it a *scalar* or zeroth-rank tensor—the simplest kind of tensor.

Now, consider the weight of one of the apples. This weight is a force, equal to the mass of the apple, m, times the gravitational acceleration ($g = 9.8\,\mathrm{m/s^2}$), and it is directed downwards. That is, its weight is mg directed downwards towards the table top. Let us call this weight $\boldsymbol{W}$, which has a clear physical meaning independent of coordinates, although it is helpful to describe its direction as "downwards." The fact that $\boldsymbol{W}$ is well defined even in the absence of coordinates makes it a tensor, and because it has a direction, it is called a *vector*, or first-rank tensor. If I introduce a set of coordinates x, y, z, I may place these axes in any way I choose. If I put the z-axis pointing downwards, the vector will have components relative to that coordinate system, with only the z component nonzero, and I might describe it as "the vector $(0, 0, mg)$." This (old) point of view is legitimate but carries with it the need to understand that the *components* are referred to as a coordinate system and as such have no intrinsic meaning.

More complicated objects can be easily imagined. For example, there could be a wind blowing across the table, and that could be described by another vector $\boldsymbol{V}$ describing its velocity. Referred to a coordinate system, it would be described by three components, and the presence of these two vectors is clearly something of intrinsic geometric significance. We can think of a product of the two vectors, meaning simply all the information needed to describe them. In terms of coordinates, if the weight is described by components w_i with i ranging from 1 to 3 and the wind velocity by v_j with j ranging from 1 to 3, we can think of a single geometric object $\boldsymbol{V} \bigotimes \boldsymbol{W}$ called the *tensor product* of $\boldsymbol{V}$ and $\boldsymbol{W}$ with 9 components $(\boldsymbol{V} \bigotimes \boldsymbol{W})^{ij} = v^i w^j$. The object $(\boldsymbol{V} \bigotimes \boldsymbol{W})$ is called a second-rank tensor, and this construction can be repeated to create more and more complex objects. (The reason for the upper placement of indices will be made clear shortly.)

The example given above was chosen to make it clear that tensors can turn up anytime one has vectors, but it is easy to find more physically motivated examples. Suppose one has a perfect fluid of density ρ moving with velocity $\boldsymbol{V}$ with components v^i. Density is a scalar if we allow only "proper" rotations which cannot change the signs of volumes. The flux or mass per unit volume per unit area perpendicular to direction i flows has components ρv^i. The flux of momentum in direction j is $\rho v^i v^j$. This motivates defining the second-rank "momentum tensor" tensor $\boldsymbol{T} = \rho \boldsymbol{V} \bigotimes \boldsymbol{V}$ with components $T^{ij} = \rho v^i v^j$.

All the tensors of a given rank have the structure of a vector space which they inherit from the operations allowed on vectors: they can be added, or multiplied by scalars. In fact, all the machinery of linear algebra carries over directly to them.

There has been an implicit indication that tensors here have something to do with symmetry or a concept of allowed transformation of the coordinates. The class of acceptable coordinate systems defines the geometry in any given situation, and we can speak then of tensors with respect to a symmetry group. For example, if we allow arbitrary rotations of a given orthogonal coordinate system, we can speak of "Cartesian tensors" or tensors under the group of orthogonal transformations in three dimensions. This means that one can pass from a representation of a tensor in terms of components with respect to one coordinate system to another by making the coordinates transform according to a matrix which represents that coordinate change.

This final step in thinking clarifies the notion of what a tensor must be. Given some space, one considers the geometry to be defined by the group of transformations which leave it invariant. For example, in flat Euclidean three-dimensional space, one might take the group of rotations, which are orthogonal matrices of determinant one, or SO(3). The vectors and tensors we have been talking about so far are then elements of spaces in which SO(3) acts linearly, i.e., they lie in representation spaces of SO(3).

Depending on the space in question and the structures imposed upon it, special tensors and operations can be defined. For example, in three-dimensional flat space, we have an operation that takes two vectors and produces from them a (coordinate-independent!) scalar called the "dot product," "inner product," or "scalar product." This is defined using the Kronecker delta δ_{ij} which is defined to have the value 1 when $i = j$ and 0 otherwise. It is a geometric object with a well-defined meaning independent of coordinates and is, thus, a tensor but of a different kind. One says that while $v^i w^j$ are components of a "contravariant tensor," δ_{ij} are the components of a "covariant tensor." The expression $\sum_i \sum_j \delta_{ij} v^i w^j$ is the scalar product of $\boldsymbol{v}$ and $\boldsymbol{w}$ and is a scalar. The terms *covariant* and *contravariant* are historical and come from the idea that for a quantity such as $\sum_i \sum_j \delta_{ij} v^i w^j$ to be invariant, the components with upper and lower indices should transform differently under a change of coordinates. The summation signs are often omitted, with a summation over repeated indices implied following the so-called "Einstein summation convention." An inner product automatically provides a dual space to that of the tensors, and one can write $v_i = \delta_{ij} v^j$ and speak of "covariant" as opposed to "contravariant" components of $\boldsymbol{v}$, and of the use of δ_{ij} to "lower indices."

The concept of a tensor is independent of any notion of a dot product, but many spaces of interest have such a product naturally present. In special relativity, for example, we consider "four-vectors" labeling differences in space and time and requiring four components. In units where the speed of light is unity, the dot product is provided by the Minkowski metric tensor η_{ab}: (a, b running from 0 to 3) where η_{ab} is 1 when $a = b = 0$, -1 when $a = b \neq 0$, and zero otherwise. This dot product is preserved by a larger group than just rotations, and this group is the Lorentz group of rotations and boosts, also called SO(3,1). Note that it is not positive-definite, so it is not an inner product in the strict sense of the word, and it defines a pseudo-Euclidean metric. In this case, we have SO(3,1) vectors as opposed to SO(3) vectors.

The most general (pseudo-)Riemannian geometry provides a dot product in terms of a metric tensor g_{ab} which is similar to η_{ab} but unrestricted other than to be nonsingular. In general relativity, we allow all invertible linear transformations in four dimensions, the symmetry group is GL(4), and we have GL(4) tensors for physical quantities.

John David Swain

See also **Einstein equations; Symmetry groups**

Further Reading

Bishop, R.L. & Goldberg, S.I. 1968. *Tensor Analysis on Manifolds*, New York: Macmillan

Dodson, C.T.J. & Poston, T. 1991. *Tensor Geometry*, 2nd edition, Berlin and New York: Springer

Simmon, J.G. 1994. *A Brief on Tensor Analysis*, 2nd edition, New York: Springer

TESSELLATION

A tessellation (or tiling) is a covering of a surface with tiles so that there are no gaps or overlaps. The tile shapes can be all different, as in the chips used to produce a complex Byzantine mosaic, or they may consist of a limited number of shapes, each congruent to one or more "prototiles" which serve as templates. Since ancient times, almost every culture has produced tessellations for utilitarian or decorative purposes—on walls, ceilings, and roofs of buildings, for pavements and plazas, and for designs to be woven, painted, printed, incised, or inlaid on every variety of surface. Tessellations adorn many churches, temples, palaces, and mosques; perhaps the most celebrated geometric tessellations are found in the Alhambra, in Granada, Spain. Tessellations also occur in nature, in the designs formed by scales or packed cells on a living surface.

Although artisans have produced tessellations for thousands of years, only recently have mathematicians undertaken a methodical study of the subject. Grünbaum and Shephard's book is the most complete reference. Intuitively, a shape "tiles" (is a prototile for a tessellation) if congruent copies of that shape can be fitted together exactly to fill a surface. Every triangle and every quadrilateral can tile the plane, and convex polygons with seven or more sides can never tile the plane. All convex hexagons that tile the plane have been determined, and 14 classes of convex pentagons that tile have been discovered, but it is not known if there are others. Tessellations by regular polygons appear frequently; those known as Archimedean tessellations are composed of regular polygons with edges matched, and have the same arrangement of tiles occurring at every vertex of the tiling (Figure 1).

An infinite variety of shapes tile the plane, many of them classified by special properties. Given an arbitrary set of shapes, many tests can attempt to answer the question: Will these tile the plane? But there is no guarantee of an answer. The question is mathematically

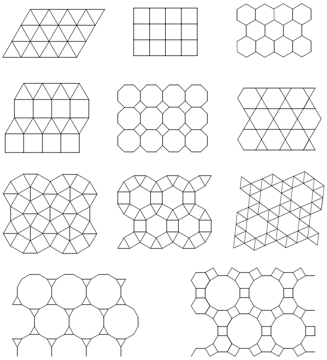

Figure 1. The 11 Archimedean tessellations. The three with only one prototile are called regular tessellations; the others are often called semiregular.

undecidable; that is, there is no algorithm that can deliver an answer of yes or no for every possible set of shapes.

M.C. Escher, a Dutch graphic artist (1898–1972), is the best-known creator of tessellations using whimsical figures as tiles (Figure 2). Inspired by the geometric tessellations in the Alhambra, he sought to answer the question: What shapes can tile the plane so that every tile is surrounded in the same way? As he discovered some answers, he developed a system of tile types that enabled him to create imaginative shapes that fit together in a prescribed manner. In his lifetime, he produced more than 150 finished tessellations; many have been used by scientists, particularly crystallographers, to illustrate their theories.

Almost all tessellations with repetition (including those in Figures 1 and 2) are periodic, that is, there is a smallest patch of the tiling that can be translated repeatedly by two independent vectors to fill out the whole tiling. It is natural to want repetition in a predictable manner to lay tile, to print, or to weave a pattern. Symmetry groups of periodic tilings are known as two-dimensional crystallographic groups, because crystals, by definition (until very recently), were defined by their periodic molecular structure. These groups consist of the translations, rotations, reflections, and glide-reflections that can act on the tiling in such a way that each tile moves to fit exactly onto another, leaving the tiling invariant. There are only 17 distinct symmetry groups for periodic tilings in the Euclidean plane; these are frequently used to classify tilings. Colored tilings (such as an extended checkerboard or Escher's tilings) can be analyzed according to color symmetries, which permute colors of tiles as well as positions of tiles in the tiling.

Nonperiodic tessellations have no translation symmetry. Although a regular tiling by squares can be made nonperiodic by shifting a few rows a small distance, the most interesting nonperiodic tessellations are produced by an aperiodic set of prototiles—every tiling formed by such a set is nonperiodic. It is not known if there is a single aperiodic tile, but there are several known aperiodic sets of two or more prototiles. The most well-known pairs were discovered by Roger Penrose in the 1970s: a kite and a dart (or a thick and a thin rhombus). In each Penrose tiling formed by these pairs (Figure 3), every patch of the tiling repeats infinitely often, but never by a translation that leaves the tiling invariant. Some properties of Penrose tilings are similar to those exhibited by unusual alloys discovered in the 1980s. These were named quasicrystals because their X-ray diffraction patterns displayed a crystal-like orderly repetition of bright spots, but also exhibited rotation symmetry forbidden in a periodic structure. Many unusual properties of Penrose and other aperiodic tilings have been discovered, and various techniques have been developed to study their structure (Senechal, 1995). Yet the study of aperiodic tilings and of quasicrystals is in its infancy, and it remains to be seen if the connections between them are more than superficial.

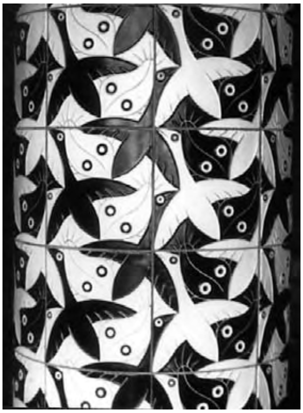

Figure 2. An Escher tessellation covers a column in a school in Baarn, Holland. (All M.C. Escher works © Cordon Art B.V., Baarn, The Netherlands.)

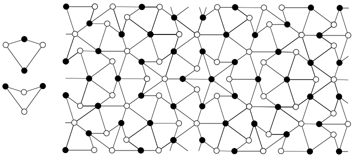

Figure 3. A Penrose tiling by kites and darts requires matching vertices of the same color.

A Voronoï (or Dirichlet) tessellation is determined by a given discrete set S of points in a surface. Each point x of S is surrounded by the region of points that are closer to x than to any others in S. The boundaries of these regions consist of those points that lie equidistant between at least two points of S. These regions with their boundaries (called Voronoï polygons or Dirichlet domains) are the tiles of the Voronoï tessellation determined by S. Such tessellations arise naturally in a wide variety of applications in physics (Wigner–Seitz regions); crystallography (Wirkungsbereich); physiology (capillary domains); urban planning (regions of service for schools or fire stations); biology (modeling cell arrangement); statistical spatial data analysis; and many other areas. Such tilings are almost always nonperiodic with many differently shaped tiles. Mathematical properties of such tilings and various algorithms to construct them are vigorous areas of research (Okabe et al., 1992).

Tessellations on surfaces such as spheres or the hyperbolic plane and tessellations of space of three or higher dimensions are equally important. Which shapes can fill space, and in what manner, is of great interest to scientists (and manufacturers). Little is known in this area; there is not even a complete list of space-filling tetrahedra. Other topics of investigation are tessellations with special tiles (e.g., polyominoes, or fractal tiles), relationships between local and global properties, classification, and construction of tessellations with special properties.

DORIS SCHATTSCHNEIDER

See also **Symmetry groups**

Further Reading

Bezdek, K. 2000. Space filling. In *Handbook of Discrete and Combinatorial Mathematics*, edited by K.H. Rosen, Boca Raton: CRC Press, pp. 824–830

Goodman, J.E. & O'Rourke, J. (editors). 2004. *Handbook of Discrete and Computational Geometry*, 2nd edition, Boca Raton: CRC Press (Relevant chapters: Tiling, D. Schattschneider & M. Senechal; Polyominoes, D. Klarner & S.W. Golomb; Voronoï diagrams and Delaunay triangulations, S. Fortune; Crystals and Quasicrystals, M. Senechal)

Gruber, P.M. & Wills, J.M. (editors). 1993. *Handbook of Convex Geometry*, Amsterdam: North-Holland (Relevant chapters: Geometric algorithms, H. Edelsbrunner, vol. A, pp. 699–735; Tilings, E. Schulte, vol. B, pp. 899–932); Geometric crystallography, P. Engel, vol. B, pp. 989–1041)

Grünbaum, B. & Shephard, G.C. 1987. *Tilings and Patterns*, New York: Freeman

Okabe, A., Boots, B. & Sugihara, K. 1992. *Spatial Tessellations: Concepts and Applications of Voronoï Diagrams*, 2nd edition, New York: Wiley (Okabe, A., Boots, B., Sugihara, K. & Chiu, S.N. 2000)

Patera, J. (editor). 1998. *Quasicrystals and Discrete Geometry*, Providence, RI: American Mathematical Society

Schattschneider, D. 2004. *M.C. Escher: Visions of Symmetry*, new edition, New York: Abrams

Schulte, E. 2002. Tilings. In *Encyclopedia of Physical Science and Technology*, 3rd edition, vol. 16, New York: Academic Press, pp. 763–782

Senechal, M. 1995. *Quasicrystals and Geometry*, Cambridge and New York: Cambridge University Press

Washburn, D.K. & Crowe, D.W. 1988. *Symmetries of Culture: Symmetry and Practice of Plane Pattern Analysis*, Seattle: University of Washington Press

THERMAL CONVECTION

Thermal convection is the transfer of heat by flow. In general, heat can be transported by convection, conduction, or radiation. In all cases, heat transfer requires the presence of a temperature gradient, and heat is transported from high to low temperature. Conduction is a diffusive process that involves the exchange of energy via collisions between molecules in a gas or a liquid, or interactions between lattice waves or electrons in a solid. Heat transport by radiation results from the emission and absorption of electromagnetic waves. Convection, finally, involves the transport of heat by the bulk physical motion of a fluid medium. The flow can arise naturally, due to thermal expansion—hot fluid is less dense than cold fluid and so tends to rise, while cold fluid is more dense and tends to fall—or it can be forced by some externally applied means.

In a layer of fluid heated from below, thermal convection is known as Rayleigh–Bénard convection, which is one of the most important systems for the study of pattern formation. This is because precise, well-controlled experiments are possible and because the equations describing the system (the Navier–Stokes equations coupled with an equation for heat transport and the appropriate boundary conditions) are well known, allowing close contact among experiment, theory, and simulations.

Figure 1(a) shows schematically an experiment for the study of Rayleigh–Bénard convection. The top plate is maintained at a temperature T_t, while the bottom plate is at a higher temperature $T_b = T_t + \Delta T$. The separation between the plates is d. For ΔT small, there is no flow and heat transport across the fluid layer is by conduction. Because of thermal expansion, however, the hotter fluid near the bottom plate is less dense than the cooler fluid above it. This is gravitationally unstable, but initiating flow costs energy due to viscous dissipation. As a result, convection does not begin until ΔT is large enough that the energy gained by starting flow offsets the cost due to dissipation. It is convenient to write ΔT in dimensionless form as the Rayleigh number, given by $\mathrm{Ra} = g\alpha d^3 \Delta T/\nu\kappa$, where g is the acceleration due to gravity, α is the thermal expansion coefficient of the fluid, ν is the fluid's viscosity, and κ is its thermal diffusivity. The onset of convection occurs when Ra reaches a critical value Ra_c which depends on the nature of the top and bottom boundaries. For rigid, isothermal boundaries, $\mathrm{Ra}_c = 1708$. For a rectangular cell with the same boundaries, the flow appears as a pattern of straight convection rolls oriented parallel to the short side of the cell, with a

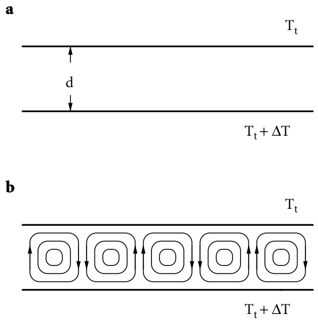

Figure 1. A schematic illustration of Rayleigh–Bénard convection. A fluid layer is bounded above and below by plates separated by a distance d, with the bottom plate at a temperature higher by ΔT than the top plate. When ΔT is below a critical value, there is no flow (a), but when ΔT is higher than the critical value, convective flow develops in the form of parallel, counter-rotating convection rolls (b).

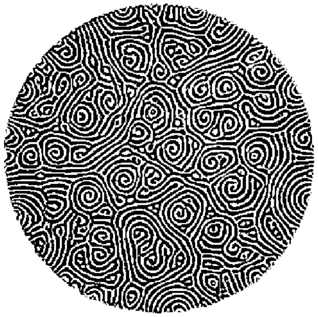

Figure 2. Spiral defect chaos visualized in a convection experiment in a pressurized gas. Spiral defect chaos is a time-dependent steady state that appears near onset at low Prandtl number. Here dark regions indicate upflow and light regions indicate downflow. (Image courtesy of S.W. Morris.)

wavelength equal to $2.016d$. This situation is shown in Figure 1(b).

Heat transport in a convecting fluid is measured by the Nusselt number Nu, defined as $\mathrm{Nu} = \lambda_{\mathrm{eff}}/\lambda$, where λ is the thermal conductivity of the fluid and λ_{eff} is the effective thermal conductivity taking into account the heat transported by the flow. The Prandtl number, $\mathrm{Pr} = \nu/\kappa$, affects the dynamics of the flow pattern above onset.

The onset of Rayleigh–Bénard convection is an example of a bifurcation from a uniform (no-flow) state to one characterized by a spatial pattern. It can be described by the Ginzburg–Landau equation, which can be derived from the full equations describing the system. For the case of two rigid, isothermal boundaries, the system is up-down symmetric and the bifurcation is a supercritical pitchfork bifurcation. If we define $\varepsilon = (\mathrm{Ra} - \mathrm{Ra_c})/\mathrm{Ra_c}$, then the bifurcation occurs at $\varepsilon = 0$. The Ginzburg–Landau equation predicts that close to onset, the amplitude of the convective flow field will grow as $\varepsilon^{1/2}$, while the correlation length of the pattern (the length scale over which the amplitude changes) behaves as $\varepsilon^{-1/2}$. These predictions have been confirmed experimentally.

If the up-down symmetry of the system is broken, either by making top and bottom boundaries different or due to a variation in the fluid properties with temperature across the cell (non-Boussinesq effects), then the symmetry of the bifurcation is also broken and convection appears via a subcritical bifurcation. In this case, the convection pattern at onset takes the form of an array of hexagonal cells (see photo of atmospheric hexagonal convection cells on page 3 of color plate section).

Above onset, the straight-roll pattern is stable within a range of wave numbers and Rayleigh numbers. The range of stability (known as the "Busse balloon") is limited by a variety of secondary instabilities which depend on Pr. The original straight–roll pattern becomes unstable to perturbations in phase or amplitude which lead to new flow patterns, which can have different wave numbers, orientations, and symmetries. Certain of these secondary instabilities can make the flow pattern two- or three-dimensional or time dependent.

In larger systems, when the size of the system becomes much greater than the correlation length of the pattern, the convection patterns are more complex than ideal straight rolls. The tendency of the rolls to orient perpendicular to the sidewalls results in patterns with curved rolls and a spatially varying wave number. Even close to onset, defects and localized instabilities can lead to time-dependent patterns. At low Pr, a transition to a complex, spatiotemporally chaotic state known as spiral defect chaos occurs (Figure 2). Interestingly, spiral defect chaos exists in exactly the regime where ideal straight rolls are expected to be stable, and experiments have shown that the two states can coexist. This suggests that there are two stable states in this regime, with different basins of attraction. Spiral defect chaos is obtained for most conditions, while specially chosen initial or boundary conditions are required to obtain ideal straight rolls.

At high Ra, the convective flow becomes turbulent. In this case, most of the temperature drop occurs in boundary layers near the top and bottom plates of the cell. Coherent plumes of hot and cold fluid can form at the lower and upper boundary layers, respectively, and a coherent large scale flow can exist in the convection cell. Simple theories based on the stability of the boundary layers predict that heat flow for turbulent convection should scale as $\mathrm{Nu} \sim \mathrm{Ra}^{1/3}$, while more sophisticated models give different values of the exponent. Recent research indicates that the exponent is itself a function of Ra.

There are many variants of Rayleigh–Bénard convection that have been studied in an effort to elucidate particular aspects of nonlinear dynamics. In appropriately chosen binary mixtures, one can obtain traveling-wave convection patterns. Thermal convection in anisotropic materials (e.g., liquid crystals) also leads to complex dynamics. Convection with an imposed mean flow has been used to study the transition from convective to absolute instability. Convection with rotation exhibits spatiotemporally chaotic patterns close to onset.

JOHN R. DE BRUYN

See also **Bifurcations; Pattern formation; Turbulence**

Further Reading

Bodenschatz, E., Pesch, W. & Ahlers, G. 2000. Recent developments in Rayleigh–Bénard convection. *Annual Review of Fluid Mechanics*, 32: 709–778

Busse, F. 1981. Transition to turbulence in Rayleigh–Bénard convection. In *Hydrodynamic Instabilities and the Transition to Turbulence*, edited by H.L. Swinney & J.P. Gollub, Berlin: Springer

Cross, M.C. & Hohenberg, P.C. 1993. Pattern formation outside of equilibrium. *Reviews of Modern Physics*, 65: 851–1112

de Bruyn, J.R, Bodenschatz, E., Morris, S.W., Trainoff, S.P., Hu, Y., Cannell, D.S. & Ahlers, G. 1996. Apparatus for the study of Rayleigh–Bénard convection in gases under pressure. *Review of Scientific Instruments*, 67: 2043–2067

Manneville, P. 1990. *Dissipative Structures and Weak Turbulence*, London: Academic Press

Siggia, E. 1994. High Rayleigh number convection. *Annual Review of Fluid Mechanics*, 26: 137–168

THERMO-DIFFUSION EFFECTS

In fluid mixtures, a temperature gradient can drive a concentration current or generate a concentration gradient depending on boundary conditions. This thermo-diffusion effect—nowadays referred to as the Soret effect or Ludwig–Soret effect—was first reported by Carl Ludwig in 1856 and by Charles Soret in 1879. They observed an increase (decrease) of salt concentration at the cold (hot) end of a tube filled with salty water (Ludwig, 1856; Soret, 1879). The reciprocal effect that a concentration gradient drives a heat flow or generates a temperature gradient was first reported by Louis Dufour in 1872 for gas mixtures in a porous medium. Theoretically, these and similar effects are most conveniently captured within the Onsager theory of irreversible macroscopic processes, in which generalized thermodynamic forces and resulting fluxes are linearly related to each other (de Groot & Mazur, 1962; Landau & Lifshitz, 1959).

In the last 20 years or so, it has become clear that the linear Soret effect plays a dominant role in the nonlinear behavior of convective pattern formation in binary fluid mixtures (Platten & Legros, 1984; Cross & Hohenberg, 1993; Lücke et al., 1998). Consider the typical Bénard configuration of a horizontal fluid layer of height d that is heated from below in a homogeneous gravitational field, $\boldsymbol{g} = -g\,\boldsymbol{e}_z$. Strongly heat-conducting impermeable horizontal plates impose a vertical temperature difference ($\Delta T > 0$) such that $T = T_0 \pm \Delta T/2$ at $z = \mp d/2$. T_0 is the mean temperature of the fluid layer. At small ΔT, the laterally homogeneous quiescent conductive state is stable with the linear temperature profile $T_{\text{cond}}(z) = T_0 - \Delta T z/d$. In a mixture like, for example, ethanol dissolved in water, this conductive temperature gradient generates as a consequence of the Soret effect a concentration gradient, so that

$$C_{\text{cond}}(z) = C_0 + S_T C_0(1 - C_0)\Delta T z/d. \tag{1}$$

Here, $C = \rho_1/(\rho_1 + \rho_2)$ is the mass concentration of the solute which is in our example the lighter component. C_0 is its mean, S_T the Soret coefficient, and $k_T = T_0 C_0(1 - C_0)S_T$ the thermodiffusion ratio. The Soret coupling between temperature and concentration fields (cf. below) is most conveniently measured in terms of the separation ratio $\psi = -S_T C_0(1 - C_0)\beta/\alpha = -(k_T/T_0)\beta/\alpha$. Here, α and β are the thermal and solutal expansion coefficients of the total mass density $\rho_1 + \rho_2 = \rho = \rho_0[1 - \alpha(T - T_0) - \beta(C - C_0)]$ of the mixture for small deviations of T and C from their means. Positive S_T corresponding to negative ψ (for mixtures such as ethanol-water where α and β are positive) implies a concentration increase (of the lighter component) near the cold upper plate and a decrease near the warmer lower plate and vice versa for $S_T < 0$ ($\psi > 0$). Note that in experiments, ψ can easily be varied, say, between -0.6 and 0.25, by varying T_0 and C_0.

Convection is described by the balance equations

$$\nabla\cdot\boldsymbol{u} = 0, \tag{2a}$$

$$(\partial_t + \boldsymbol{u}\cdot\nabla)\boldsymbol{u} = \sigma\nabla^2\boldsymbol{u} + R\sigma\,(\delta T + \delta C)\,\boldsymbol{e}_z - \nabla p, \tag{2b}$$

$$(\partial_t + \boldsymbol{u}\cdot\nabla)T = \nabla^2 T, \tag{2c}$$

$$(\partial_t + \boldsymbol{u}\cdot\nabla)C = L\nabla^2 C - L\psi\nabla^2 T \tag{2d}$$

for mass (2a), momentum (2b), heat (2c), and concentration (2d) in the Oberbeck–Boussinesq approximation. δT and δC in (2b) denote deviations from the mean T_0 and C_0, respectively. Lengths are scaled with d, time with the vertical thermal diffusion time d^2/κ, and the velocity field $\boldsymbol{u} = (u, v, w)$ with κ/d, where κ is the thermal diffusivity of the mixture. Temperatures are reduced by ΔT, concentration by $\Delta T\alpha/\beta$, and pressure p by $\rho_0(\kappa/d)^2$.

The Dufour effect that provides a coupling of concentration gradients into the heat balance is discarded because it is relevant only in a few gas mixtures and possibly in liquid mixtures near the liquid–vapor critical point.

Besides the Rayleigh number $R = (\alpha g d^3/\nu\kappa)\Delta T$ measuring the thermal driving force, three additional numbers enter into the field equations: the Prandtl number $\sigma = \nu/\kappa$, which is of order 10 for ethanol–water mixtures at room temperature, the Lewis number $L = D/\kappa \simeq 0.01$, and the separation ratio ψ. Here, ν denotes the kinematic viscosity and D the concentration diffusivity.

The concentration field is responsible for the significantly larger complexity of binary mixture convection compared with pure fluids. It causes the richness of spatiotemporal properties of the convective structures, of the bifurcation behavior, and of the transient growth of convection. The Soret-generated concentration variations δC influence the buoyancy, that is, the driving force for convective flow in (2b). The flow in turn mixes by advectively redistributing concentration. This nonlinear advective mixing in developed convective flow is typically much larger than the smoothing by linear diffusion—the Péclet number measuring the strength of advective concentration transport relative to diffusion is easily of the order of a few thousand. Thus, the concentration balance is strongly nonlinear giving rise to boundary layer behavior and strongly anharmonic concentration field profiles in the horizontal direction, as in Figure 1. In contrast, the momentum and heat balances remain weakly nonlinear close to onset as in pure fluids, implying only smooth and basically harmonic variations, $\sim e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$, as the critical modes (cf. Figure 1).

To summarize, the feedback interplay among (i) the Soret-generated concentration variations that are sustained against mixing and diffusion by externally imposed and internal temperature gradients, (ii) the resulting changes in the buoyancy, and (iii) the strongly nonlinear advective transport and mixing causes binary mixture convection to be rather complex not only with respect to its spatiotemporal properties but also concerning its bifurcation behavior.

Take, for example, $\psi < 0$, where the Soret-induced separation requires higher heating to destabilize the conductive state than for a pure fluid characterized by $\psi = 0$ (for a review of the multitude of convection states appearing for destabilizing positive ψ see, for example, Huke et al., 2000). Then the off-diagonal coupling between solutal buoyancy and advection of Soret-induced concentration variations described above generates oscillations—traveling waves (TWs) of horizontally propagating rolls occur via a subcritical Hopf bifurcation whenever ψ is sufficiently negative. The bifurcation properties of such oscillatory TW states are shown in Figure 2 for different ψ as a function of the reduced Rayleigh number $r = R/R_c^0$, where $R_c^0 = 1707.76$ marks the convective onset in pure fluids. With increasing flow intensity (Figure 2a), the fluid gets more mixed while simultaneously the TW frequency decreases (Figure 2b) as the flow intensity and the Nusselt number (Figure 2d) approach the pure fluid reference values. Here the mixing is measured by the reduced spatial variance $M = \sqrt{\langle \delta C^2 \rangle / \langle \delta C_{\text{cond}}^2 \rangle}$ of the concentration.

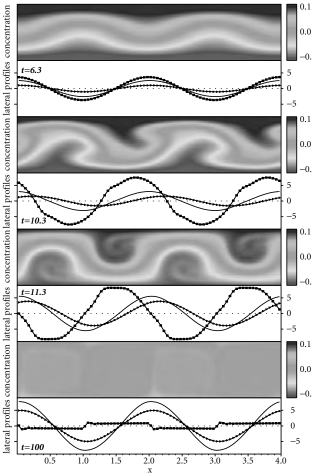

Figure 1. Evolution of convection after perturbing the quiescent conductive state. The concentration distribution in a vertical cross section of the fluid layer is displayed by color-coded plots where highest concentration was initially at the top, and lowest at the bottom. Wave profiles at midheight, $z = 0$, are shown for the fields of vertical velocity w (thin lines), $40\delta T$ (lines with triangles), and $400\delta C$ (lines with squares). The final TW propagates to the left. Parameters are $L = 0.01$, $\sigma = 10$, $\psi = -0.25$, $r = 1.42$, and wavelength $\lambda = 2$. For better visibility two wavelengths are shown. (This figure is also reproduced on page 3 of the color plate section.)

Figure 1 shows the complex spatiotemporal concentration redistribution during the growth of oscillatory convection at slightly supercritical heating. The growth starts generically from perturbations of the conductive state that contain the two critical Hopf modes for counterpropagating TWs with roughly equal amplitudes. First, they linearly superimpose to form SW-like oscillations of growing amplitude with the large Hopf frequency. But then they compete via nonlinear advection with each other; at a critical SW amplitude, advective breaking of the concentration wave triggers a very fast flow-induced transition from SW to TW convection with anharmonic profile, large phase velocity, and large amplitude of the concentration wave. Finally, advective mixing and diffusive homogenization slow down the TW as the concentration differences between left and right turning rolls slowly decrease.

In mixtures with sufficiently negative ψ, there are also uniquely selected stable LTW states of

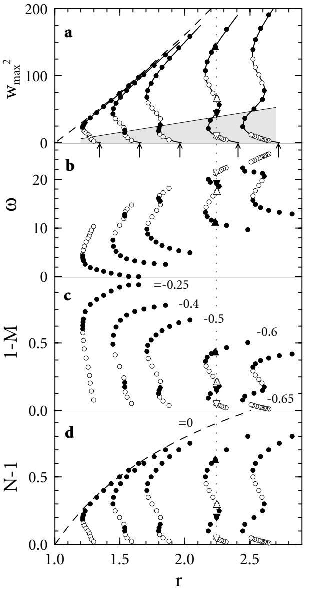

Figure 2. TW–bifurcation diagrams for different Soret coupling strength ψ: (a) squared maximal vertical flow w^2_{max}, (b) frequency ω, (c) degree of mixing $1 - M$, and (d) convective contribution to the Nusselt number $N - 1$ vs reduced Rayleigh number r. Stable (unstable) TW states are marked by filled (open) symbols. Arrows mark Hopf thresholds r_{osc} for onset of TW convection. The $\psi = 0$ pure fluid limit is included in (a) and (d) by the dashed line. TW states on the vertical line are discussed in more detail in Hollinger et al. (1997). Only states in the shaded region of (a) are weakly nonlinear. Parameters are $L = 0.01$, $\sigma = 10$, and $\lambda = 2$.

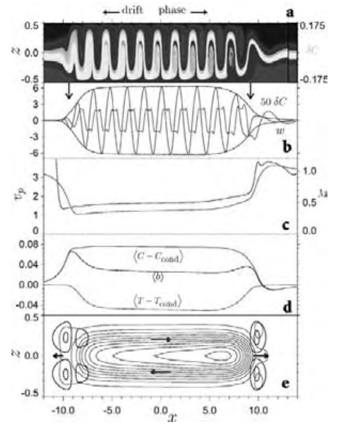

Figure 3. Broad LTW of length $l = 17.4$: (a) Concentration deviation δC from global mean (pale gray) in a vertical cross section of the layer. (b) Lateral wave profiles at midheight, $z = 0$, of δC (gray), vertical velocity w (black), and its envelope. At the arrows, $w_{\mathrm{max}} = v_{\mathrm{p}}$. (c) Mixing number M (gray) and phase velocity v_{p} (black). The variation of the wavelength $\lambda(x) = 2\pi v_{\mathrm{p}}(x)/\omega$ is the same because the LTW frequency ω is a *global* constant. (d) Time averaged deviations from the conductive state at $z = -0.25$ for concentration (upper), temperature (lower), and their sum ($\langle b \rangle$) measuring the convective contribution to the buoyancy. (e) Streamlines of time averaged concentration current $\langle \boldsymbol{J} \rangle = \langle \boldsymbol{u}\delta C - L\nabla(\delta C - \psi \delta T) \rangle$ (gray) and velocity field $\langle \boldsymbol{u} \rangle$ (black). The latter results from $\langle b \rangle$ and documents roll shaped contributions of $\langle \boldsymbol{u} \rangle \langle \delta C \rangle$ to $\langle \boldsymbol{J} \rangle$ under the fronts and the associated $\langle \delta C \rangle$ redistribution. Thick black and gray arrows indicate $\langle \boldsymbol{u} \rangle$ and transport of positive δC (alcohol surplus), respectively. Thus, in the lower half of the layer, negative δC (water surplus) is transported to the right. Parameters are $L = 0.01$, $\sigma = 10$, $\psi = -0.35$, $r = 1.346$. (This figure is also reproduced on page 3 of the color plate section.)

localized, that is, spatially confined, TWs. They occur at small subcritical heating where extended TWs cannot exist and where the conductive state is strongly stabilized by the Soret effect. Such a strongly nonlinear LTW (Figure 3) is robustly sustained by a complex concentration redistribution process. Therein flow-induced mixing locally reduces the Soret separation and thereby increases the buoyancy to levels that suffice to drive well-mixed fluid flow there. In Figure 3, positive (negative) δC is sucked from the top (bottom) boundary layer into right (left) turning rolls as soon as they become nonlinear under the trailing LTW front. This happens when the vertical velocity w roughly exceeds the local phase velocity v_p (left arrow in Figure 3b) so that regions with closed streamlines appear (Hollinger et al., 1997; Lücke et al., 1998). Within them "dark" ("gray") concentration is transported predominantly in the upper (lower) part of the layer to the right. Mean concentration, on the other hand, migrates mostly to the left along open streamlines that meander between the closed roll regions and that follow the global mean in Figure 3a. The *time-averaged* current of δC (gray lines in Figure 3e) reflects the mean properties of this transport. Since positive and negative (zero) δC is transported away from (towards) the left trailing front, mean concentration accumulates there and causes a strong drop of $M(x)$. In the same way, the leading front's concentration varies and with it, $M(x)$ are strongly increased even beyond the conductive state's values. Thus, unlike TWs, LTWs do not reach a balance among δC injection, advective mixing, and diffusive homogenization on a constant level of small M. Rather LTW rolls collapse under the leading front when v_{p} has grown up to w (right arrow in Figure 3b). Thereafter, concentration is discharged and sustains ahead of the leading front a barrier of $\langle \delta C \rangle$ that prevents the expansion of the conductive state into the LTW.

M. LÜCKE

See also **Thermal convection**

Further Reading

Cross, M.C. & Hohenberg, P.C. 1993. Pattern formation outside of equilibrium. *Reviews of Modern Physics*, 65: 851

de Groot, S.R. & Mazur, P. 1962. *Non-equilibrium Thermodynamics*, Amsterdam: North-Holland

Hollinger, St., Büchel, P. & Lücke, M. 1997. Bistability of slow and fast traveling waves in fluid mixtures. *Physical Review Letters*, 78: 235

Huke, B., Lücke, M., Büchel, P. & Jung, Ch. 2000. Stability boundaries of roll and square convection in binary fluid mixtures with positive separation ratio. *Journal of Fluid Mechanics*, 408: 121

Jung, D. & Lücke, M. 2002. Localized waves without the existence of extended waves: Oscillatory convection of binary mixtures with a strong Soret effect. *Physical Review Letters*, 89: 054502

Köhler, W. & Wiegand, S. (editors). 2002. *Thermal Nonequilibrium Phenomena in Fluid Mixtures*, Berlin and New York: Springer

Landau, L.D. & Lifshitz, E.M. 1959. *Fluid Mechanics*, Oxford: Pergamon Press and Reading, MA: Addison-Wesley (originally published in Russian)

Lücke, M., Barten, W., Büchel, P., Fütterer, C., Hollinger, St. & Jung, Ch. 1998. Pattern formation in binary fluid convection and in systems with throughflow. In *Evolution of Structures in Dissipative Continuous Systems*, edited by F.H. Busse & S.C. Müller, Berlin and New York: Springer, p. 127

Ludwig, C. 1856. Diffusion zwischen ungleich erwärmten Orten gleich zusammengesetzter Lösungen. *Sitzungsberichte der Kaiserliche. Akademie der Wissenschaften (Mathematisch-Naturwissenschaftlicheclasse), Wien*, 65: 539

Platten, J.K. & Legros, J.C. 1984. *Convection in Liquids*, Berlin: Springer

Soret, C. 1879. Sur l'état d'équilibre que prend, au point de vue de sa concentration, une dissolution saline primitivement homogène, dont deux parties sont portées à des températures différentes. *Archives des Sciences: Physiques et Naturelles, Genève*, 2: 48–61

THETA FUNCTIONS

The n-dimensional *theta function* is a function of $n + n(n+1)/2$ complex variables, $\theta(z|\tau)$. The first n coordinates form a vector in an n-dimensional complex space $z = (z_1, \dots, z_n)$ while the remaining $n(n+1)/2$ variables are entries in a symmetric n-dimensional matrix τ_{ik}; $(i, k = 1, \dots, n)$. The theta function is defined by a Fourier series as

$$\theta(z|\tau) = \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \mathrm{e}^{\{\mathrm{i}\pi \boldsymbol{m}\tau\boldsymbol{m}^t + 2\mathrm{i}\pi \boldsymbol{z}\boldsymbol{m}^t\}},$$

where the n-tuple summation runs over the whole n-dimensional set of integers $\mathbb{Z}^n$, and the imaginary part of the matrix τ is supposed to be positive definite to provide convergence of the series. The defining properties of the theta function are its periodicity and modular properties. The periodicity property is given by the relation

$$\begin{aligned}\theta(z + \boldsymbol{e}_k|\tau) &= \theta(z|\tau), \quad \theta(z + \boldsymbol{e}_k\tau|\tau)\\ &= \mathrm{e}^{-\mathrm{i}\pi\tau_{kk} - 2\mathrm{i}\pi z_k}\theta(z|\tau), \quad k = 1, \dots, n,\end{aligned}$$

where only the kth component of the vector $\boldsymbol{e}_k$ is nonzero and equal to unity.

Consider the modular group Γ the group of all $2n \times 2n$ integer matrices γ such that $\gamma J \gamma^t = J$ where

$$J = \begin{pmatrix} 0 & -1_n \\ 1_n & 0 \end{pmatrix},$$

and 1_n is the unit n-dimensional matrix. Under the action of the modular group, the theta-function transforms as

$$\theta(z|\tau) = \frac{\varepsilon}{\sqrt{\det M(\tau)}} \exp\left\{\frac{1}{2}\sum_{i,k=1}^{n} z_i z_k \frac{\partial}{\partial \tau_{ik}} \det M(\tau)\right\} \theta\left(z'|\tau'\right),$$

where

$$\begin{aligned} M(\tau) &= c\tau + d, z' = (a\tau + d)^{-1}z,\\ \gamma &= \begin{pmatrix} a & b \\ c & d \end{pmatrix},\\ \tau' &= (a\tau + b)(c\tau + d)^{-1}, \end{aligned}$$

a, b, c, d are $n \times n$ matrices and $\varepsilon^8 = 1$.

The theta functions are introduced to construct modular functions in τ-variables of order k defined to satisfy $f(\gamma \circ \tau) = \det(c\tau + d)^k f(\tau)$ and abelian functions in z-variables. Abelian functions $F(z)$ are functions of n complex variables with $2n$ complex periods, $\boldsymbol{T}_i$, $i = 1, \dots, 2n$, $F(z + \boldsymbol{T}_i) = F(z)$. The advantage of using theta functions to define modular and abelian functions comes from the rapid convergence of the $theta$-series.

The most important class of abelian functions are abelian functions whose τ-variables are constructed from an algebraic curve X of genus n, given by a polynomial equation $P(\lambda, \mu) = 0$. The introduction of local coordinates turns X into a one-dimensional complex analytical variety called the Riemann surface of the curve X. The Riemann surface of genus n can be topologically described as a sphere with n handles. It is always possible to draw, on such a torus, a basis of $2n$-cycles $a_1, \dots, a_n; b_1, \dots, b_n$ with intersection numbers $a_i \circ a_j = b_i \circ b_j = 0$ and $a_i \circ b_j = -b_j \circ a_i = \delta_{ij}$, where $\circ$ means intersection of corresponding cycles.

Differential and integral calculus can also be developed on X. In contrast to the case of an extended complex plane, or Riemann sphere, which can be considered as a Riemann surface of genus zero, there exists n linearly independent holomorphic differentials $\mathrm{d}w_1, \dots, \mathrm{d}w_n$ which can be normalized by the conditions $\oint_{a_k} \mathrm{d}w_i = \delta_{ik}$. The period matrix τ of the curve is then given by $\tau_{ik} = \oint_{b_k} \mathrm{d}w_i$. Meromorphic differentials $\mathrm{d}\omega_k$, that is, differentials with poles of order k, can be also defined on X; usually these differentials are normalized by the condition $\oint_{a_l} \mathrm{d}\omega_k = 0, l = 1, \dots, n$.

The abelian function $u(\boldsymbol{t}) = u(t_1, \dots, t_n)$ of n complex variables associated with the curve X of genus n is then defined as

$$u(t_1, \dots, t_n) = -2\frac{\partial^2}{\partial t_1^2} \log \theta \left(\sum_{k=1}^{n} \boldsymbol{U}_k t_k + \boldsymbol{U}_0 | \tau \right) + c, \tag{1}$$

where the vector $\boldsymbol{U}_k$ is the vector of b-periods of the normalized meromorphic differential $\mathrm{d}\omega_k$, $\boldsymbol{U}_0$ is a vector through which initial data are introduced and c is a constant.

The most developed case is the case of hyperelliptic curves, when the polynomial P is given by

$$P(\lambda, \mu) = \mu^2 - \prod_{k=1}^{2n+1} (\lambda - \lambda_k),$$

where n is the genus and the branching points λ_k, $k = 1, \dots, 2n+1$ are supposed to be distinct. The holomorphic differentials described above are given in this case by the formula $\mathrm{d}w_k = \lambda^{k-1}\mathrm{d}\lambda/\mu$, $k = 1, \dots, n$. In the case $n = 1$, the abelian function of the curve is the well-known elliptic function.

The remarkable role of θ functions in the spectral theory of the Schrödinger equation was discovered by Its & Matveev (1975). Consider the spectral problem

$$\left\{ \frac{\partial^2}{\partial x^2} - u(x) \right\} \Psi(x; \lambda, \mu) = \lambda \Psi(x; \lambda, \mu),$$

where $u(x)$ is smooth and real potential, $\Psi(x; \lambda, \mu)$ is an eigenfunction, and λ is the spectral parameter. Suppose that the spectrum consists of $n+1$ continuous segments $[\lambda_1, \lambda_2], \dots, [\lambda_{2n+1}, \infty]$. Then the potential is given by the formula (1) with $t_1 = x$ and $t_k = \mathrm{const}$ for $k > 1$, while the eigenfunction $\Psi(x; \lambda, \mu)$ is given by the formula

$$\Psi(x; \lambda, \mu) = C \frac{\theta \left(\int_{(\infty,\infty)}^{(\lambda,\mu)} \mathrm{d}\boldsymbol{w} + \boldsymbol{U}_2 x + \boldsymbol{U}_0 \middle| \tau \right)}{\theta \left(\boldsymbol{U}_2 x + \boldsymbol{U}_0 \middle| \tau \right)} \times \exp \left\{ x \int_{(\lambda_0,\mu_0)}^{(\lambda,\mu)} \mathrm{d}\omega_2(\lambda, \mu) \right\}, \tag{2}$$

where C is a normalizing constant, $\mathrm{d}\omega_2(\lambda, \mu)$ is the second kind abelian differential with second-order pole at infinity and zero a_i-periods, $\boldsymbol{U}_2$ is a vector of b_i-periods of $\mathrm{d}\omega_2(\lambda, \mu)$, and the constant vector $\boldsymbol{U}_0$ and point (λ_0, μ_0) are defined by initial conditions. The isospectral deformation of the potential $u(x)$ when the second variable $t_2 = t$ in formula (1) is switched on, while other variables $t_k, k > 2$ remain constant, turns the function $u(x, t)$ into the n-gap solution of the Korteweg–de Vries equation. Equations (1) and (2) form the foundation of the theory of finite-gap solutions of soliton equations. In the limit when the branch points collide in pairs, $\lambda_{2k-1} \to \lambda_{2k}$ $k = 1, \dots, n$, these formulae become N-soliton formulae. Krichever (1977) generalized the whole theory to other soliton equations and nonhyperelliptic curves.

Introductions to the subject are given in Farkas & Kra (1998) for the theory of Riemann surfaces, Novikov, Chapter 11 in Zakharov et al. (1980), Dubrovin (1981). Mumford (1983, 1984) for theta functions and completely integrable equations, and Belokolos et al. (1994) for algebro-geometric methods of integration of nonlinear equations.

VICTOR ENOLSKII

See also **Elliptic functions; Inverse scattering method or transform; *N*-soliton formulas**

Further Reading

Belokolos, E.D., Bobenko, A.I., Enolskii, V.Z., Its, A.R. & Matveev, V.B. 1994. *Algebro-geometric Approach to Nonlinear Integrable Equations*, Berlin and New York: Springer

Dubrovin, B.A. 1981. Theta functions and nonlinear equations. Russian Mathematical Surveys, 36: 11–80

Farkas, H. & Kra, I. 1998. *Riemann Surfaces*, New York: Springer

Its, A.R. & Matveev, V.B. 1975. Sohrödinger operators with a finite-band spectrum and the N-soliton solutions of the Korteveg–de Vries equation. *Teoreticheskaya i Matematicheskaya Fizika*, 23: 51–68

Krichever, I.M. 1977. The method of algebraic geometry in the theory of nonlinear equations. *Russian Mathematical Surveys*, 32: 180–208

Mumford, D. 1983, 1984. *Tata Lectures on Theta*, vols. 1, 2, Boston: Birkhäuser

Zakharov, V.E., Manakov, S.V., Novikov, S.P. & Pitaevskii, L.P. 1980. *Soliton Theory: Inverse Scattering Method*, Moscow: Nauka (in Russian)

THREE-BODY PROBLEM

See ***N*-body problem**

THREE-WAVE INTERACTION

See ***N*-wave interactions**

THRESHOLD PHENOMENA

Defined in the dictionary as an "intensity below which a mental or physical stimulus cannot be perceived and can produce no response," the term *threshold* has deep roots in our language, representing a collective awareness of strong nonlinearity. Above threshold, the effect of a stimulus changes dramatically from that below, as is indicated by such common phrases as the "tipping point" and the "last straw." Examples of threshold phenomena abound in physics, engineering, nonlinear mathematics, chemistry, biology, and neuroscience.

Applications of the Threshold Concept

In nonlinear optics, the threshold for laser oscillation specifies a level of pump power below which only a small amount of incoherent light is emitted from the laser, while above threshold a brilliant, highly directed beam of output light is observed. Beginning with the wall switch, electrical engineers have devised many varieties of threshold circuits that change rapidly from one voltage level to another when an input variable exceeds a certain value; indeed, a digital computer can be viewed as a large collection of interacting threshold devices. Also the threshold logic unit (TLU), comprising a switch with input channels having adjustable weights, is a useful component of learning systems for pattern recognition (Nilsson, 1990).

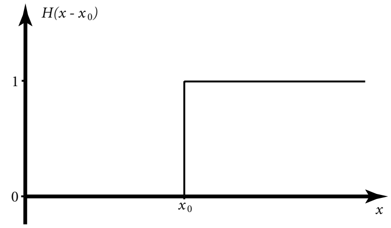

Figure 1. A Heaviside step function.

Mathematical representations of switching devices often involve the Heaviside step function,

$$H(x - x_0),$$

a generalized function that is zero for $x < x_0$ and one for $x > x_0$, where x_0 is the threshold as shown in Figure 1. In phase space models of nonlinear dynamic systems, one often finds a *separatrix*, or critical surface across which solution trajectories have very different behaviors. Noting this, mathematician cum meteorologist Edward Lorenz famously asked: “Does the flap of a butterfly’s wings in Brazil set off a tornado in Texas?”—not only coining a dramatic metaphor for threshold phenomena and launching modern studies of chaos, but also leading philosophers to examine what is meant by causality. Is it only the *last* straw that should be blamed for breaking the camel’s back? Or are *all* the straws that were loaded onto the beast to some degree complicit?

In the realms of physical chemistry, a reaction-diffusion (excitable) system rests quietly until stimulated above its threshold for self-supporting activity, whereupon traveling waves, spirals, and scroll waves are typically observed to emerge. In biology, examples of threshold phenomena include the germination of a seed at a critical (threshold) level of ambient moisture and the insemination of an ovum, in addition to the birth of life itself from the chemical components of the Hadean oceans some 3.5 billion years ago.

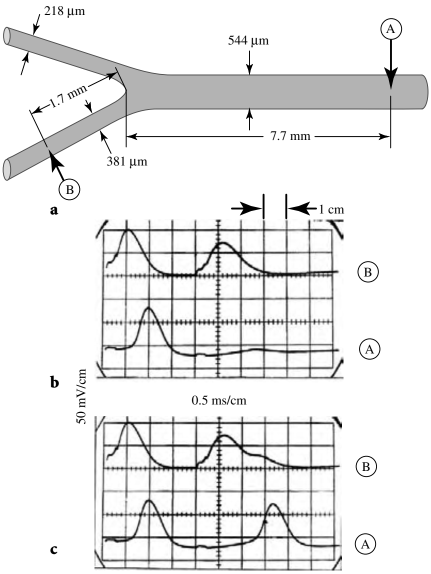

Figure 2. Threshold action in a branching region of a squid giant axon. The interpulse interval is slightly longer in (c) than in (b).

Threshold Effects in Neuroscience

As one would suppose from introspection, the field of neuroscience offers many examples of threshold phenomena which have become known since the observation of "all-or-none" response of a neuron was proposed by Edgar Adrian in 1914 (Adrian, 1914). This threshold behavior of a neuron was used by psychiatrist Warren McCulloch and mathematician Walter Pitts in 1943 to formulate the first computer model of the human brain McCulloch & Pitts (1943). Their model assumes each neuron to be represented by a single Heaviside step function, which jumps from the off (zero) to the on (one) state when a linear sum of input (dendritic) signals exceeds a threshold value. In 1958, Frank Rosenblatt introduced the "perceptron" which employs TLUs as the basic emenents of a brain model (Rosenblatt, 1958).

Recent studies of dendritic dynamics suggest far greater complexity, with the possibility of threshold effects occurring at the branching regions of incoming fibers (Stuart et al., 1999). In Figure 2 are displayed some experimental measurements of nerve impulse transmission through a branching region in the giant axon of a squid (Scott, 2002). Figure 2(a) shows the geometry of the preparation (not to scale), indicating the electrode positions where upstream (B) and downstream (A) recordings were made. The preparation was stimulated with two pulses, which are seen in the upper traces of both Figures 2(b) and 2(c). In these experiments, the spacing between the two incoming impulses was under experimental control, and below a certain threshold value of this impulse interval—that shown in Figures 2(b) and 2(c)—it is seen that the second impulse no longer makes it through. Besides interpulse interval, several other neural parameters can result in such a threshold effect, including branch geometry, ionic concentrations, fatigue, and narcotization level.

As the dendrites of many neurons are known to carry active impulses (action potentials) (Stuart et al., 1999), it appears that the branching regions of dendrites can act as threshold devices (switches), greatly increasing our estimates of the ability of a neuron to process information (Scott, 2002).

ALWYN SCOTT

See also **Butterfly effect; Excitability; Flip-flop circuit; Lasers; McCulloch–Pitts network; Multiplex neuron; Nerve impulses; Phase space; Reaction-diffusion systems**

Further Reading

Adrian, E.D. 1914. The all-or-none principle in nerve. *Journal of Physiology (London)* 47: 460–474

McCulloch, W.S. & Pitts, W.H. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology* 5: 115–133

Nilsson, N.J. 1990. *Learning Machines*, 2nd edition, San Mateo, CA: Morgan Kaufmann

Rosenblatt, F. 1958. A probabilistic model for information storage and organization in the brain. *Psychology Review* 65: 386–408

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, New York: Springer-Verlag

Stuart, G., Spruston, N. & Häusser, M. 1999. *Dendrites*, Oxford: Oxford University Press

TIME-SERIES ANALYSIS

Predicting total eclipses of the moon or the sun is an art that dates back to mankind's oldest civilizations. Nonetheless, it represents a class of very modern analytic tools: the analysis of time-series data. A time-series is a set of measurements s_n, $n = 1, \ldots, T$, whose index n refers to the time t when the measurement is recorded by $n\tau = t$, τ is the sampling interval. For time-dependent phenomena, this index carries part of the information which is destroyed, for example, by a random reshuffling of the temporal order of the measurements. If one wishes to make use of this property in data analysis, particular statistical methods are required which are able to characterize temporal correlations inside the data. Typical goals in time-series analysis are predictions, data classification, signal manipulation, and system identification.

Today's technical facilities for data acquisition and data storage call for efficient time-series tools. To mention a few examples, in medicine (particularly cardiology and neurology), there is a need for automated diagnostics (i.e., data classification). In finance or weather and climate research, data-driven prediction methods are relevant, since model equations are either lacking (economy) or expensive to solve (weather). In modern telecommunications and automatic speech recognition, noise reduction is an essential issue, where time-series analysis should supply the background for signal separation.

The above examples show that unlike data of the positions of the planets, whose analysis enabled Johannes Kepler to derive the laws of planetary motion, modern time-series problems are concerned with data that have a complicated, strong aperiodic component. All data analysis methods start from some paradigm about the origin of the observed signatures and irregularities. This background is indispensable for a sound interpretation of the statistical quantities thus obtained, or for an estimate of the validity of the consequences drawn from the results. However, the large number of different approaches in time-series modeling supplies a corresponding diverse set of time-series analysis methods. We focus here on the two most general approaches, both of which represent a whole theoretical framework and not just a particular aspect.

Linear stochastic models are a well-developed class of time-series models. Being linear, their properties can be fully and rigorously derived from their model equations. However, in order to generate aperiodic time-series data, such models require stochastic inputs. Auto regressive models AR(M),

$$s_{n+1} = \sum_{k=0}^{M-1} a_k s_{n-k} + \xi_n,$$

where ξ_n is Gaussian distributed white noise ($\langle \xi_l \xi m \rangle = \delta_{l,m}$) with unit variance, represent the time discretized motion of the superposition of $M/2$ damped harmonic oscillators driven by noise, if the coefficients a_k fulfill certain stability conditions. The output is therefore characterized by $M/2$ frequencies and the corresponding damping coefficients, such that a spectral analysis is the most suitable analysis tool (*See* **Spectral analysis**). Thus, AR-models are suitable for data sets that have a few pronounced peaks in their power spectrum. If the observed power spectrum instead is broad band, another linear model, the moving average model MA(N), can be more reasonable:

$$s_{n+1} = \sum_{k=0}^{N-1} b_k \xi_{n-k},$$

where ξ_n is again white Gaussian noise. Notice that there is no feedback of the observable s, so that this model just averages over the independent noise inputs and hence creates colored noise. In the limits $N, M \to \infty$, both model classes are equivalent. For practical purposes where a small number of coefficients is desirable, a combination of both, the ARMA(M,N)-model, is often used. The coefficients of such models can be fit to observed data, for example, by solving least-squares problems (Box & Jenkins, 1976).

The well-known sunspot number time series of solar activity with its pronounced 11-year period can be well captured by an AR-model. ARMA models are also rather suited to describe many noise-dominated signals such as sound emission signals in technical environments, and they are used to model single phonemes of human speech in automatic speech recognition systems. ARMA-models are hence employed for data-driven predictions and signal classification tasks. They are often useful if the signal is either dominated by some few frequencies or when it is really noisy. However, the linearity of the model behind translates into the fact that the observables should be Gaussian random variables themselves, and that all higher-order statistics beyond the power spectrum and the auto-correlation function are fully determined by either of these two.

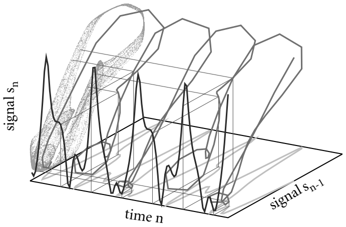

Figure 1. Sketch of the time-delay embedding procedure: The time series s_n (plotted in the frontal plane) combined with its time shifted version (in the bottom plane) forms a sequence of vectors (the curve in space), whose projection along the time axis accumulates to a set of points representing (in this particular case) a strange attractor together with an invariant measure on it. The data are voltages measured in a chaotic electric resonance circuit.

From many model systems and physical laboratory experiments arises a different class of sources for aperiodic time-series data: so-called chaotic dynamical systems. Aperiodicity here comes from intrinsic instabilities without random inputs, and nonlinearity destroys the superposition principle, such that a Fourier decomposition of the signal is not useful. Higher-order statistics is not trivially related to second-order statistics. Instead of a statistical characterization, one would here try to reconstruct the deterministic and dynamic origin of the signal.

If one assumes that a given experimental observable represents the deterministic dynamics in some non-observed and even unknown phase space, the concept of phase space reconstruction by embedding is needed: As proven by Takens (1981), the set of delay vectors obtained from a scalar time series by joining successive observations, $\boldsymbol{s}_n = (s_n, s_{n-1}, s_{n-2}, \ldots, s_{n-m+1})$, is equivalent to a set of phase space vectors of the underlying dynamical system, provided that the embedding dimension m fulfills $m > 2D_f$ (see Figure 1). Here, D_f is the attractor dimension of the underlying dynamical system (*See* **Attractors; Dimensions; Fractals**). Hence, successive vectors $\boldsymbol{s}_n$ are deterministically related to each other by an unknown function $\boldsymbol{s}_{n+1} = \boldsymbol{F}(\boldsymbol{s}_n)$, which reduces to an unknown scalar function $s_{n+1} = f(\boldsymbol{s}_n)$, since the other components of $\boldsymbol{s}_{n+1}$ are just copied from $\boldsymbol{s}_n$.

In a practical analysis, one represents the time-series data in embedding spaces with increasing dimension m and searches for signatures of determinism. One model free approach for this is the estimation of the fractal dimension of the set of delay vectors in the embedding space, which should saturate at the value D_f for $m > D_f$. Of interest is predictability: one can try to extract the unknown function $f(\boldsymbol{s}_n)$ from the time series, by selecting a suitable model for f and then solving the least-squares problem $\sum (s_{n+1} - f(\boldsymbol{s}_n))^2 = \min$, where the minimization is done with respect to parameters in f. In the simplest case, f is approximated by a constant in a neighborhood of the actual observation $\boldsymbol{s}_n$. Then, the predictor is the average over the future values of all neighboring vectors

$\boldsymbol{s}_k$ of $\boldsymbol{s}_n$, $\hat{s}_{n+1} = \langle s_{k+1} \rangle$, where $||\boldsymbol{s}_k - \boldsymbol{s}_n|| < \varepsilon$ with some suitable norm and some suitably small ε (also called the "Lorenz method of analogues"). Regardless of how the function f is represented, the relation $\hat{s}_{n+1} = f(\boldsymbol{s}_n)$ will yield only good predictions for s_{n+1} if the embedding dimension m is large enough for the delay vectors to be equivalent to the unknown phase space vectors of the underlying dynamical system. In such an embedding space, the delay vectors form a finite sample according to the underlying invariant measure of the dynamical system, so that, in principle, all characteristics of the latter (Lyapunov exponents, entropies) can be determined (*See* **Chaotic dynamics**).

Such methods have been successfully applied to many physical laboratory experiments, but more recently successful applications to real-world data have been reported, such as noise reduction for human voice, wind speed prediction for wind farms, diagnostics on human heart rate data, epilepsy prediction, and machine wear detection in technical production systems. The extension of the theory to noise-driven nonlinear systems and to nonstationary data has recently been tackled, but highly nonrecurrent data such as economics data will probably remain untractable.

HOLGER KANTZ

See also **Attractors; Chaotic dynamics; Dimensions; Fractals; Spectral analysis**

Further Reading

Abarbanel, H.D.I. 1996. *Analysis of Observed Chaotic Data*, New York: Springer

Box, G.E.P. & Jenkins, G.M. 1976. *Time Series Analysis*, San Francisco: Holden-Day

Farmer, J.D. & Sidorowich, J.J. 1987. Predicting chaotic time series. *Physical Review Letters*, 59: 845

Kantz, H. & Schreiber, T. 1997. *Nonlinear Time Series Analysis*, Cambridge and New York: Cambridge University Press

Lorenz, E.N. 1969. Atmospheric predictability as revealed by naturally occuring analogues. *Journal of Atmospheric Sciences*, 26: 636

Takens, F. 1981. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, edited by, D.A. Rand & L.S. Young Berlin and New York: Springer, p. 366

TODA LATTICE

The Hamiltonian system governing the longitudinal oscillations of a chain of unit point masses connected by identical springs is

$$\begin{aligned} \dot{q}_n &= \dot{p}_n, \\ \dot{p}_n &= V'(q_{n+1} - q_n) - V'(q_n - q_{n-1}), \end{aligned} \tag{1}$$

where $-V'$ is the restoring force of the springs and q_n is the displacement of the nth mass from equilibrium. Morikazu Toda (Toda, 1981) introduced the potential (scaled here so that two physical parameters disappear)

$$V(r) = e^{-r} + r - 1, \tag{2}$$

and discovered explicit solutions of (1). He obtained periodic traveling waves, in terms of elliptic functions, and two-soliton solutions in terms of sech^2. (The term $(r-1)$ makes $V(r) \sim r^2/2$ for small r; it cancels out in the problems considered below.)

This mass-spring chain is the Toda lattice. It is valuable as a solvable model of one-dimensional lattice dynamics; it is frequently used to illustrate fundamental constructions common to many integrable systems; and it appears, sometimes in surprising ways, in a variety of physical and mathematical problems.

The Infinite and Periodic Lattices

All the special properties of the Toda lattice flow from a Lax representation, $\dot{\mathsf{L}} = [\mathsf{B}, \mathsf{L}]$, which implies that eigenvalues and certain other spectral data of L are independent of time. For many purposes, it is convenient to write L and B in new coordinates. With the definition

$$a_n = \tfrac{1}{2}\,\mathrm{e}^{-(q_{n+1}-q_n)/2}, \quad b_n = \tfrac{1}{2}\,p_n, \tag{3}$$

the exponential nonlinearity in (2) becomes polynomial:

$$\dot{a}_n = a_n(b_n - b_{n+1}), \quad \dot{b}_n = 2\,(a_{n-1}^2 - a_n^2). \tag{4}$$

The standard Poisson bracket on q_n, p_n induces a (nonstandard) Poisson bracket on a_n, b_n, in which the only nonzero relations are

$$\{a_n, b_n\} = -\frac{1}{4}a_n, \quad \{a_n, b_{n+1}\} = \frac{1}{4}a_n. \tag{5}$$

The Lax operators are discretized Schrödinger operators with potentials a_n, b_n, acting on sequences $\mathbf{y} = (\ldots, y_{-1}, y_0, y_1, \ldots)$:

$$\begin{aligned} (\mathsf{L}\mathbf{y})_n &= a_{n-1}y_{n-1} + b_n y_n + a_n y_{n+1}, \\ (\mathsf{B}\mathbf{y})_n &= a_{n-1}y_{n-1} - a_n y_{n+1}. \end{aligned} \tag{6}$$

A (finite or infinite) tridiagonal symmetric matrix, such as L, is called a "Jacobi matrix." In terms of the shift operator $\Delta z_n = z_{n+1}$, which is a convenient abbreviation,

$$\mathsf{L} = \Delta^{-1}a + b + a\Delta, \tag{7}$$

where a and b are the diagonal matrices $\mathrm{diag}(a_n)$, $\mathrm{diag}(b_n)$. Constants of motion derived from the Jacobi operator L will be in involution, that is, they have zero Poisson bracket.

Thanks to its Lax representation, methods used in the study of other soliton equations also apply to the Toda lattice.

The Bi-infinite Lattice with Decaying Initial Condition (Toda, 1981)

According to (3), $(q_{n+1} - q_n) \to 0$ and $p_n \to 0$ translates to $a_n \to \frac{1}{2}$, $b_n \to 0$. The initial value problem is studied by a discrete version of the inverse scattering method. The unperturbed eigenvalue operator at $n = \pm\infty$ is $\mathsf{L}_0 = \frac{1}{2}(\Delta^{-1} + \Delta)$, which has the bounded "plane wave" eigenfunctions $y_n = z^{\pm n}$, $|z| = 1$, for $\lambda = (z + z^{-1})/2 \in [-1, 1]$. This relation between "energy" λ and "momentum" z is the analog of $\lambda = k^2$ in the Schrödinger equation. The scattering matrix is a function of momentum and is defined on $|z| = 1$. Bound states may occur for $\lambda < -1$ and $\lambda > 1$; these correspond to solitons moving to the left and to the right, respectively. In the absence of solitons (and with a certain condition on the reflection coefficient), the large-time behavior consists of a decaying wave train for $|n/t| < 1$, vanishingly small motion when $|n/t| > 1$, and a connecting regime described by a Painlevé transcendent (Kamvissis, 1993).

Constants of motion of the infinite lattice can be computed explicitly as $I_k := \mathrm{Trace}\,\frac{1}{k}(\mathsf{L}^k - \mathsf{L}_0^k)$ (L_0^k removes a divergent term). The Hamiltonian equations generated by the I_k are found by use of the Poisson bracket (5).

The Periodic Lattice (Toda, 1981)

The potential in the Lax operator is periodic, say, period N. As in the theory of Hill's equation, one introduces a Floquet multiplier and Floquet solutions by imposing the condition

$$y_{N+n} = \rho y_n. \tag{8}$$

L and B then reduce to matrices:

$$\mathsf{L}(\rho) = \begin{pmatrix} b_1 & a_1 & 0 & \dots & \rho^{-1} a_N \\ a_1 & b_2 & a_2 & \dots & 0 \\ 0 & a_2 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_{N-1} \\ \rho\, a_N & \dots & \dots & a_{N-1} & b_N \end{pmatrix}, \tag{9}$$

$$\mathsf{B}(\rho) = \begin{pmatrix} 0 & -a_1 & 0 & \dots & \rho^{-1} a_N \\ a_1 & 0 & -a_2 & \dots & 0 \\ 0 & a_2 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -a_{N-1} \\ -\rho\, a_N & \dots & \dots & a_{N-1} & 0 \end{pmatrix}. \tag{10}$$

The parameter ρ cancels from the Lax equation. The characteristic polynomial $\det(\mathsf{L}(\rho) - \lambda\mathbb{I})$ is a rational function whose coefficients are independent of time. The function $\rho(\lambda)$ obtained by solving the characteristic equation for ρ is defined on the λ–plane with branch cuts, and determines a compact Riemann surface. For Toda's traveling wave, there are only two branch cuts; this fact is responsible for the occurrence of elliptic functions in his solution formula.

Every energy surface is compact, so that the solution curves lie on tori; abstractly, these are the (real) Jacobian varieties associated with the Riemann surface. Action variables can be written explicitly as loop integrals. The solution of the periodic Toda lattice is expressed in terms of theta functions.

The Free Toda Lattice and Lie Theory

Many features common to integrable systems are illustrated by an example due to Jürgen Moser.

The Free Toda Lattice (Toda, 1981)

Particles number 0 and $(N + 1)$ are pulled to $-\infty$ and $+\infty$, respectively. (3) implies that $a_0 = a_N = 0$, and the corner entries in $\mathsf{L}(\rho)$ and $\mathsf{B}(\rho)$ disappear. The resulting matrices will be written simply as L, B.

Springs governed by potential (2) resist compression and are encouraged to expand. Therefore, there will now be no confining force on the remaining springs. The masses move apart as $t \to \pm\infty$, and behave asymptotically like free particles. As $t \to \pm\infty$, $(q_{n+1} - q_n) \to \infty$, whence $a_n \to 0$, the matrix $\mathsf{L}(t)$ becomes diagonal, and the entries $b_n = p_n/2$ approach the eigenvalues of $\mathsf{L}(0)$. Thus, the eigenvalues are the momenta of the free particles. As $t \to -\infty$, these are arranged with the slowest particle farthest to the right; after interaction and a phase shift, the particles emerge as $t \to +\infty$ with the fastest to the right. This is soliton interaction reduced to bare essentials. The diagonalization of $\mathsf{L}(t)$ as $t \to +\infty$, with the eigenvalues (momenta) ordered, is called the "sorting property," and points to a connection between isospectral flows and numerical linear algebra.

For tridiagonal L, the number of eigenvalues is exactly half the number of entries, and they suffice for integrability. Remarkably, the Lax equation $\dot{\mathsf{L}} = [\mathsf{B}, \mathsf{L}]$ (with appropriate B), is completely integrable even when L is a generic symmetric matrix. The additional constants of motion required are constructed in Deift et al. (1986). The sorting property again holds; it is related to the QR diagonalization algorithm in numerical linear algebra.

The free Toda lattice is perhaps the simplest integrable system with a clear Lie-algebraic generalization. If one sets $\mathrm{Trace}\,\mathsf{L} = 0$ in (9), that is, total momentum = 0, then L belongs to $\mathfrak{sl}_N$, the Lie algebra of $N \times N$ matrices with trace zero. Let $\{h_i\}$ be a basis for the diagonal matrices in $\mathfrak{sl}_N$. The matrices $e_{\pm i}$ with a single 1 in the $(i, i+1)$, resp, $(i+1, i)$ entry, and zeros elsewhere, are called "raising and lowering operators." There are analogs of $e_{\pm i}$, h_i in the class of split real semisimple Lie algebras, such as the algebra of symplectic matrices. The Lax pair, written as

$$\mathsf{L} = \sum_{i=1}^{N-1} [\beta_i h_i + a_i(e_i + e_{-i})],$$

$$\mathsf{B} = \sum_{i=1}^{N-1} a_i(-e_i + e_{-i}),$$

makes sense in those algebras. In the corresponding generalized mass–spring systems, some particles will be governed by modifications of (2) (Olshanetsky & Perelomov, 1994; Guest, 1997; Reyman & Semenov-Tyan-Shansky, 1994; Semenov-Tyan-Shansky, 1994 a,b).

The extension to Lie algebras is not just generalization for sake of generalization. It illuminates concrete matrix calculations. There is an abstractly defined Poisson bracket, the Kostant–Kirillov bracket, of which (5) is a special case. Hamiltonian equations with respect to this Poisson bracket always have the Lax form, $\dot{\mathsf{X}} = [\mathsf{Y}, \mathsf{X}]$. One of the fundamental properties of Lax equations, the representation of the solution by means of a factorization in the Lie group, is a general Lie-theoretic phenomenon that specializes to many different integrable systems and, in particular, yields involutivity of the constants of motion TraceL^k (Guest, 1997; Semenov-Tyan-Shansky, 1994 a,b)

For the free Toda lattice, the factorization method is simple linear algebra, but it already conveys the basic idea.

Here, as well as later, it is convenient to redefine

$$a_n = \exp(q_n - q_{n+1}), b_n = -p_n. \tag{11}$$

In terms of the shift operator (7), L and B then become

$$\mathsf{L} = \Delta^{-1}a + b + \Delta, \quad \mathsf{B} = -\Delta^{-1}a. \tag{12}$$

($\mathsf{L} = b + \Delta$ or $\mathsf{B} = b + \Delta$ will give the same $\dot{\mathsf{L}}$.) Given the initial value $\mathsf{L}(0)$, write

$$\exp(t\mathsf{L}(0)) = n_-(t)^{-1}d(t)n_+(t) := n_-(t)^{-1}b_+(t) \tag{13}$$

with $n_{\pm(t)}$ upper/lower triangular having 1's on the diagonal and $d(t)$ diagonal; this is just Gaussian elimination. Then

$$\mathsf{L}(t) = b(t)\mathsf{L}(0)b(t)^{-1} \text{ satisfies } \dot{\mathsf{L}}(t) = [B(t), \mathsf{L}(t)], \tag{14}$$

where $\mathsf{B}(t) = \mathsf{L}(t)_+$ is the upper triangular part of $\mathsf{L}(t)$.

Similarly, one finds the solutions of the Hamiltonian equations generated by the constants of motion I_{k+1} (referred to as "higher Toda flows"). One introduces a separate time variable t_k for each of these systems and now factors $\exp(t_k\mathsf{L}(0)^k)$. The resulting $\mathsf{L}(t_k)$ satisfies

$$\frac{\partial \mathsf{L}}{\partial t_k} = [\mathsf{B}_k, \mathsf{L}], \quad \text{with } \mathsf{B}_k = (\mathsf{L}^k)_+. \tag{15}$$

(L^k arises because $\nabla I_{k+1} = \mathsf{L}^k$). The simultaneous solution of (15) is a multi-parameter flow, $\mathsf{L}(t_1, \ldots, t_{N-1})$, on phase space ($t_1$ is the original t). Compatibility of these equations, that is, equality of mixed partial derivatives of L, is equivalent to the involutivity of the I_k. For an infinite lattice, there are infinitely many equations (15), and infinitely many time variables $\mathbf{t} = (t_1, t_2, \ldots)$. Many soliton hierarchies have the same general form; for the higher KdV equations, for example, $\mathsf{L} = \mathrm{d}^2/\mathrm{d}x^2 + u(x)$ and $\mathsf{B}_k = (\mathsf{L}^{(2k+1)/2})_+$.

In Gaussian elimination, as in (13), d_{nn} is known to be τ_n/τ_{n-1}, where τ_n is the upper left $n \times n$ minor determinant of $\exp[t\mathsf{L}(0)]$. Tracing through the factorization steps, one finds the Hirota formula

$$a_n(t) = a_n(0)\frac{\tau_{n+1}(t)\tau_{n-1}(t)}{\tau_n(t)^2}, \tag{16}$$

which implies

$$a_n = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\ln \tau_n, \quad b_n = \frac{\mathrm{d}}{\mathrm{d}t}\ln \frac{\tau_n}{\tau_{n-1}}.$$

These τ_n are prototypes of a fundamental object in the theory of soliton hierarchies, the "τ-function," which elsewhere occurs in far more complex settings.

The τ-functions have a representation-theoretic meaning. They are the matrix elements

$$\tau_n(t) = \langle \exp(t\mathsf{L}(0))\, \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \cdots \wedge \mathbf{e}_n \mid \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \cdots \wedge \mathbf{e}_n \rangle \tag{17}$$

in the representation of the Lie group SL_N on the nth exterior power of $\mathbb{R}^N$ (totally antisymmetric covariant tensors of order n). This formula gives precisely the $n \times n$ minor determinant of $\exp(t(0))$. The skew-vector $\mathbf{e}_1 \wedge \ldots \wedge \mathbf{e}_n$ is the highest weight vector of the representation; it is annihilated by all raising operators. One may think of it as a vacuum vector and of (17) as a vacuum expectation value.

Loop Algebras and Affine Lie Algebras

The lower/upper factorization (13), solution (14), the τ-functions arising from the diagonal part in this factorization, and their interpretation in terms of group representations are fundamental features of integrable systems. The free Toda lattice affords a transparent illustration; more sophisticated versions of these ideas, set in other Lie algebras and groups, apply to a wide variety of equations. "Loop algebras" and "affine Lie algebras" are particularly useful in the theory of the periodic Toda lattice.

The loop algebra, $\widetilde{\mathfrak{sl}_N}$, is an infinite-dimensional Lie algebra whose elements are trace zero matrices $\mathsf{X}(\rho)$ with entries that are polynomials in ρ and ρ^{-1}. Such $\mathsf{X}(\rho)$ are called "loops," because the mapping $\mathsf{X} : \{|\rho| = 1\} \mapsto \mathfrak{sl}_N$ gives a closed curve, that is, "loop," of matrices. The Lie bracket is still the matrix commutator.

The Lax operators (9) and (10), for the periodic Toda lattice, belong to $\widetilde{\mathfrak{sl}}_N$. To solve the periodic Toda lattice by factorization, one prescribes the initial value $\mathsf{A}(\rho)$ of $\mathsf{L}(\rho)$ and seeks matrix functions $g_\pm(t,\rho)$ for which

$$\exp(t\mathsf{A}(\rho)) = g_-(t,\rho)^{-1}g_+(t,\rho). \qquad (18)$$

These $g_\pm(\cdot,\rho)$ are required to be analytic inside (resp. outside) the circle $|\rho|=1$; this is the analog of lower/upper in (13). The eigenvector $v(\lambda,\rho(\lambda))$ of $\mathsf{A}(\rho)$, which is a function on the Riemann surface $\det(\mathsf{A}(\rho)-\lambda\mathbb{I})=0$, determines $g_\pm(t,\rho)$. The former is expressed in terms of theta functions; hence, so are $g_\pm$ and also $\mathsf{L}=g_+\mathsf{A}g_+^{-1}$ (Reyman & Semenov-Tyan-Shansky, 1994).

To obtain the τ-functions from a Lie algebra representation, one must introduce an extension of the loop algebra, the "affine" Lie algebra $\widehat{\mathfrak{sl}}_N$. (Affine Lie algebras give the name to affine Toda field theory, described below.)

The affine algebra has one extra element $\hbar$, which, in some respects, acts like an identity matrix. It has zero bracket with all loops; however, the ordinary matrix commutator in $\widetilde{\mathfrak{sl}}_N$ is modified so that $\hbar$ can arise as a Lie bracket of loops. This would not be possible if $\hbar$ were truly the identity, since a matrix commutator $[\mathsf{X}(\rho),\mathsf{Y}(\rho)]$ must have trace zero. The extended algebra contains loops that behave as annihilation and creation operators: they satisfy $[a_k^*,a_k]=\hbar$. Thanks to this modification, the familiar realization of the Heisenberg commutation relations by multiplication and differentiation operators becomes available in the study of representations of $\widehat{\mathfrak{sl}}_N$.

The τ-functions for the periodic Toda lattice are obtained as vacuum expectation values, analogous to (17), in certain representations of the affine Lie algebra; they turn out to be theta functions and again satisfy the Hirota equations (16). For the free Toda lattice, τ-functions also arose from the diagonal factor in n_-dn_+. There is a similar factorization in the affine group $\widehat{SL}_N$. This group and the diagonal factor in the lower/upper factorization are rather complicated objects. In particular, the determinant of d will be an infinite determinant.

The matrices e_0 (resp. f_0) with the corner elements $\rho\, e_{N1}$, resp. $\rho^{-1}e_{1N}$ in (9), (10) have an intrinsic meaning (they are raising resp. lowering operators in $\widetilde{\mathfrak{sl}}_N$). Therefore, the periodic Toda lattice will generalize to the algebra of loops with values in a semisimple Lie algebra.

An affine Lie algebra also provides the setting for a remarkable unification of two of the most important soliton systems, the Toda lattice and the Ablowitz–Kaup–Newell–Segur (AKNS) equations. From the solution $a_n(\mathbf{t}), b_n(\mathbf{t})$ of the Lax equations (15) for the infinite lattice, one can build functions $q_n(\mathbf{t}), r_n(\mathbf{t})$, which, for every n, satisfy the general AKNS hierarchy, with t_1 playing the role of x. For example, the simultaneous solution $a_n(t_1,t_3), b_n(t_1,t_3)$ of the Toda equations and the higher Toda flow generated by Trace L^4 will be transformed into solutions $q_n(x,t_3), r_n(x,t_3)$ of the modified KdV (MKdV) equation. A kind of Bäcklund transformation (precisely, a Schlesinger transformation) connects q_n, r_n and $q_{n\pm1}, p_{n\pm1}$. As a special case, solutions of the free Toda lattice correspond to soliton-like potentials for AKNS.

This result can be obtained by operations on formal series, but the Lie-theoretic explanation is more illuminating (Bergvelt & ten Kroode, 1988). The 2 × 2 matrix in the AKNS scattering problem depends on a (spectral) parameter and so belongs to the loop algebra $\widetilde{\mathfrak{sl}}_2$. There is a representation of the affine extension $\widehat{\mathfrak{sl}}_2$ in which the Heisenberg subalgebra, mentioned above, has infinitely many vacuum vectors killed by the a_k^*. The vacuum expectation values are the Toda τ-functions τ_n, and the AKNS variables are determined by $q_n=-\tau_{n+1}/\tau_n, r_n=\tau_{n-1}/\tau_n$.

The Two-dimensional Toda Lattice

The Toda-AKNS family has a sweeping generalization related to the Kadomtsev–Petviashvili equation (KP). It consists of four Lax equations which involve two infinite families of time variables, t_k, x_k. A continuum limit of this system was encountered in the study of deformations of a two-dimensional oil–water interface (Hele-Shaw flow) and has revealed an integrable structure of conformal mappings. This is sketched in the next section.

The operator L in (12) is replaced by a formal series, and a second operator M is introduced:

$$\mathsf{L} = \Delta + u_0 + u_1\Delta^{-1} + u_2\Delta^{-2} + \cdots, \qquad (19)$$

$$\mathsf{M} = \Delta^{-1} + v_0 + \Delta v_1 + \Delta v_2 + \cdots. \qquad (20)$$

The u_k, v_k are infinite diagonal matrices. Two of the Lax equations are

$$\frac{\partial\,\mathsf{L}}{\partial t_k} = [\mathsf{B}_k,\mathsf{L}],\ \mathsf{B}_k = (\mathsf{L}^k)_+,$$

$$\frac{\partial\,\mathsf{L}}{\partial x_k} = [\mathsf{C}_k,\mathsf{L}],\ \mathsf{C}_k = (\mathsf{M}^k)_-. \qquad (21)$$

The subscripts ± denote projections on the positive/negative powers of Δ (analog of upper/lower triangular). In the other two equations, L is replaced by M (Ueno & Takasaki, 1984).

Equations (21) specialize to the usual Toda lattice, the equations $\dot{\mathsf{L}}=[\mathsf{B},\mathsf{L}]$ for banded matrices L, and an extension of the Toda lattice in which the q_n depend on two variables. This is the "two–dimensional Toda lattice" (2DTL).

The 2DTL is obtained as follows. The compatibility conditions of the two systems in (21) are "zero curvature equations." With the abbreviations $x_1=x, t_1=t$, the first of these is

$$\frac{\partial \mathsf{B}}{\partial x} - \frac{\partial \mathsf{C}}{\partial t} + [\mathsf{B}, \mathsf{C}] = 0. \tag{22}$$

Following the one-dimensional case (12), let $\mathsf{B} = -\Delta^{-1}a$ and $\mathsf{C} = b + \Delta$. a_n, b_n, still given by (11), are now functions of t and x. (22) becomes

$$(a_n)_x = a_n(b_{n+1} - b_n), \ (b_n)_t = a_n - a_{n-1}$$

or

$$(q_n)_{xt} = \exp(q_{n-1} - q_n) - \exp(q_n - q_{n+1}). \tag{23}$$

A standard change of variables converts $(q_n)_{xt}$ to the wave operator $(q_n)_{\xi\xi} - (q_n)_{\tau\tau}$.

The free and periodic boundary conditions on the 2DTL are of primary interest. The free boundary condition for 2×2 matrices B, C yields the Liouville equation

$$q_{xt} + \exp q = 0. \tag{24}$$

Under the periodic boundary condition $q_0 = q_2, q_1 = q_3$, system (23) becomes the sinh-Gordon equation for $\Theta = q_2 - q_1$,

$$\Theta_{xt} = -4 \sinh \Theta,$$

or, if q_1, q_2 are taken to be imaginary, the sine-Gordon equation

$$\Theta_{xt} = -4 \sin \Theta.$$

Since the sine-Gordon equation is solved by the inverse scattering method, it is natural to introduce a spectral parameter ζ and an eigenvalue problem for the N-component free 2DTL:

$$\Psi_t = \mathsf{B}(\zeta)\Psi, \quad \Psi_x = \mathsf{C}(\zeta)\Psi. \tag{25}$$

Sine-Gordon theory, in which ζ^{-1} appears, further suggests that for the free 2DTL

$$\mathsf{B}(\zeta) = -\zeta^{-1}\Delta^{-1}a, \quad \mathsf{C}(\zeta) = b + \zeta\Delta. \tag{26}$$

Compatibility of the two equations in (25) implies (22). The free 2DTL is referred to as "Toda field theory."

Periodic boundary conditions are handled by an adaptation of (26). ζ is put in the lower left corner of $\mathsf{C}(\zeta)$ and $-\zeta^{-1}a_N$ in the upper right corner of $\mathsf{B}(\zeta)$. The inverse scattering method can be applied to (25). Indeed, under two-periodic boundary condition, (25) is precisely the AKNS system used to solve the sine-Gordon equation.

The periodic 2DTL is called "affine Toda field theory" (ATFT), because, as is the case for the one-dimensional periodic lattice, it is set in the context of loop and affine Lie algebras and groups. The structure of ATFT, however, is richer because of the new x–dependence.

In complex coordinates $t = w, x = \overline{w}$, ATFT is a system of elliptic partial differential equations which is encountered in the theory of harmonic maps. Solutions of the sinh-Gordon equation define Riemannian metrics on surfaces $z = f(x, y),\ (x, y) \in U$, of constant mean curvature 2. The Gauss map, which sends the unit normal vector to the unit sphere S^2, is a harmonic map, meaning that it is a critical point of the "energy" $\frac{1}{2}\int_U \|\nabla\phi\|^2 \, \mathrm{d}x \, \mathrm{d}y$ of maps $\phi : U \to S^2$. (For real–valued maps, a critical ϕ is an ordinary harmonic function.) Solutions of the multi-component elliptic ATFT define harmonic maps into other symmetric spaces. A certain group of loops is the infinite–dimensional symmetry group of this 2DTL: given a solution of 2DTL and a loop, one can construct a new solution via a factorization problem (Guest, 1997; Fordy & Woods, 1994).

In wave equation form, $(q_n)_{\xi\xi} - (q_n)_{\tau\tau}$, the 2DTL equations are hyperbolic; they describe fields on two-dimensional Minkowski space. As a first step toward a quantum theory of these fields, one can ask whether they are conformally invariant. Conformal transformations scale the (indefinite) metric. In light cone coordinates $x_\pm = \xi \pm \tau$, they are given by $(x_+, x_-) \mapsto (f_+(x_+), f_-(x_-)) = (\bar{x}_+, \bar{x}_-)$. Then $(q_n)_{x_+x_-}$, the left side of (23), becomes

$$\frac{\partial x_+}{\partial \bar{x}_+}\frac{\partial x_-}{\partial \bar{x}_-}(q_n)_{\bar{x}_+\bar{x}_-}, \text{ abbreviated } \quad J \cdot (q_n)_{\bar{x}_+\bar{x}_-}. \tag{27}$$

The 2DTL equations will be conformally invariant under a transformation of the fields, $q_n(x_\pm) \mapsto \bar{q}_n(\bar{x}_\pm)$, for which the right-hand side of (23) is also multiplied by J.

The Liouville equation (24) is conformally invariant if q transforms according to $q \mapsto \bar{q} + \ln J$. The sinh-Gordon equation, whose right side is $-2(\exp(\Theta) + \exp(-\Theta))$, admits no such transformation of Θ and is not conformally invariant. A particle described by sinh-Gordon theory has "mass": if $\sinh\Theta$ is linearized about the vaccuum state $\Theta = 0$, one obtains $\Theta_{x_+x_-} = -4\Theta$; the mass is $\sqrt{4}$. The sinh-Gordon equation is a perturbation of the conformal Liouville field equation. For example, under the change of variables $q \mapsto \Theta + \ln \varepsilon$, the equation

$$q_{x_+x_-} = -\mathrm{e}^q + \varepsilon^2\mathrm{e}^{-q},$$

becomes the sinh-Gordon equation with mass $\sqrt{2\varepsilon}$. As $\varepsilon \to 0$, which recovers the Liouville equation, the mass tends to zero. Similarly, Toda field theory is conformally invariant, while affine Toda field theory is not. Affine Toda field theory is massive in the sense described. The actual fields of interest are linear combinations of the q_n that are suitable for generalization to affine Lie algebras. This generalization is important, because the quantizations of ATFTs associated to different types of affine Lie algebras can have very different properties (Corrigan, 1999).

Continuum Limits

There are two natural continuum limits of the Toda mass–spring chain. Keeping both nonlinearity and dispersion to first order, one gets the KdV equation. The "zero dispersion limit" results in a hyperbolic system, called the "dispersionless Toda lattice."

The approximation leading to the KdV equation is taken in a right-moving coordinate system, so that the left-moving solitons of the Toda lattice disappear. The difference operator (6) becomes the Schrödinger operator, and the discrete Gel'fand–Levitan–Marchenko inverse theory limits to the inverse scattering formalism for KdV (Toda, 1981).

The zero dispersion limit retains only the quadratic nonlinearity. In the Toda equations $\dot{a}_n = a_n(b_n - b_{n+1})$, $\dot{b}_n = 2\,(a_{n-1}^2 - a_n^2)$, take $q = n\varepsilon$ and $T = t\varepsilon$. These are slow scales. With $b_n(T) \sim b(q, T)$, the difference $(b_n - b_{n+1})$ is approximately $-\varepsilon\, b_q$. The inconvenient minus signs are removed by redefining a_n, b_n. One then finds the DTL equations

$$a_T = ab_q, \quad b_T = 2(a^2)_q. \qquad (28)$$

There are again Lax equations. In the Toda eigenvalue problem (6), write $\exp(\pm\varepsilon\partial/\partial n)$ for the shift by $\pm\varepsilon$, and assume a WKB ansatz for $y_n(T) \sim y(q, T)$,

$$y(q, T) = \exp[\varepsilon^{-1} S(q, T)].$$

In the WKB approximation to the Schrödinger wave function, S_q is the inverse wavelength, which is proportional to momentum, by the deBroglie relation. Set $p = S_q$. Thus, p and q are canonically conjugate variables. As $\varepsilon \to 0$, the eigenvalue problem $\mathsf{L}\mathbf{y} = \lambda\mathbf{y}$ reduces to an eikonal equation, and $\mathbf{y}_t = \mathsf{B}\mathbf{y}$ becomes the time-evolution of the momentum S_q:

$$\begin{aligned} \mathfrak{L}(p, q) : &= a\,(\mathrm{e}^p + \mathrm{e}^{-p}) + b = \lambda, \\ p_T &= \frac{\partial}{\partial q}\,[a(\mathrm{e}^p - \mathrm{e}^{-p})] := \frac{\partial}{\partial q}\,\mathfrak{B}(p, q). \end{aligned} \qquad (29)$$

The commutator $[\mathsf{L}, \mathsf{B}]$ is replaced by the Poisson bracket with respect to p and q, and then

$$\dot{\mathcal{L}} = \{\mathcal{B}, \mathcal{L}\}$$

yields Equations (28). Their solutions may develop shocks, but as long as they are smooth, the dispersionless limits of the Toda lattice constants of motion, I_k, are constants of motion, say H_k, for (28). For example, $\sum(2a_n^2 + b_n^2) \to H_2(a, b) = \int(2a^2 + b^2)\,\mathrm{d}q$.

For smooth solutions, the eigenvalue sorting property of the free Toda lattice remains valid. The "free" boundary conditions are $a(q, T) = 0$ for $q = 0, q = 1$. Let $a_0(q), b_0(q)$ be the initial values. As $T \to \infty$, $a(q, T) \to 0$, while $b(q, T)$ tends to a decreasing function $b^*(q)$. The conserved quantities H_k have the same values for b^* as for a_0, b_0, for example, $H_2(a_0, b_0) = H_2(0, b^*)$. In this sense, the initial "matrix" a_0, b_0 becomes diagonal, and the "eigenvalues" are sorted (Brockett & Bloch, 1990).

The two-dimensional dispersionless Toda lattice hierarchy arises in an idealized model of viscous fingering in a Hele-Shaw cell. This sketch follows Kostov et al. 2001 and references therein. Two plates confine water (zero viscosity) and oil (viscous) in the complex z-plane. The water occupies a bounded region D_+, which is surrounded by oil in the exterior domain D_-. There is a source of water at $z = 0$ and a compensating sink for the oil at $z = \infty$. The object is to find the motion of the interface $\Gamma(t)$. Under some simplifying assumptions, the time development of Γ is governed by the "Laplacean growth equation" (LGE), so called because the velocity potential satisfies $\triangle\phi = 0$.

The shape of Γ is determined by the harmonic moments

$$\begin{aligned} t_k &= -\frac{1}{k\pi}\int_{D_-} z^{-k}\,\mathrm{d}x\mathrm{d}y, \ k \ge 1, \\ t_0 &= \frac{1}{\pi} \times \text{area of } D_+. \end{aligned} \qquad (30)$$

It is known that the t_k are constant under the LGE, while the area t_0 changes linearly. One, therefore, takes t_0 as time variable. Finding $\Gamma(t_0)$ amounts to finding the time-dependent conformal map $\mathfrak{L}(t_0)$ from $\{w \mid |w| > 1\}$ to $\{z \mid z \in D_-\}$. The dispersionless limit of the two-dimensional Toda hierarchy (21) enters when one allows the moments $\boldsymbol{t} = (t_1, \ldots)$ to vary and considers a family $\mathfrak{L}(t_0, \boldsymbol{t})$ of conformal maps; the LGE is then recovered as a constraint on this family.

The time-like variables t_k, x_k in (21) are taken to be the moments t_k and their conjugates $\bar{t}_k$. The area t_0 plays the role of the spatial coordinate q in (28); also set $w = \exp p$. The conformal map $\mathfrak{L}(w, \boldsymbol{t})$ and its conjugate $\bar{\mathfrak{L}}(w^{-1}, \bar{\boldsymbol{t}})$ are the dispersionless limits of L and M. The expansion of L in powers of Δ^{-1} becomes the Laurent expansion $\mathfrak{L}(w) = \text{const} \cdot w + u_0 + u_1 w^{-1} + \ldots$. The dependence of $\mathfrak{L}$, $\bar{\mathfrak{L}}$ on the deformation parameters $\boldsymbol{t}$, $\bar{\boldsymbol{t}}$ is given by

$$\frac{\partial\mathfrak{L}}{\partial t_k} = \{\mathfrak{B}_k, \mathfrak{L}\}, \quad \frac{\partial\mathfrak{L}}{\partial \bar{t}_k} = -\{\bar{\mathfrak{B}}_k, \mathfrak{L}\}, \qquad (31)$$

plus similar equations for $\bar{\mathfrak{L}}$. As in (21), $\mathfrak{B}_k$ is $(\mathfrak{L}^k)_+$. The Poisson bracket is still $\{p, q\} = 1$ or in the new notation, $\{\ln w, t_0\} = 1$. The Laplacean growth equation can be written in the form

$$\{\mathfrak{L}(w, t_0), \bar{\mathfrak{L}}(w^{-1}, t_0)\} = 1,$$

with all $t_k, \bar{t}_k$ fixed. The constraint $\{\cdot, \cdot\} = 1$ is known as the "dispersionless string equation."

Other Topics

The topics chosen give only a hint of the importance of the Toda lattice and its generalizations. Other aspects and applications that deserve an expanded description include the following.

The Toda mass-spring chain can model dispersive lattice shocks. Two unstretched halves of the lattice move toward each other at constant speed $2c$. In the scattering problem, the boundary condition is not $b_{|n|} \to 0$, but $b_n \to \mp c$ for $n \gtrless 0$. The spectrum changes with c, and this is reflected in the shockwave behavior, which is analyzed by means of the powerful steepest descent method for Riemann–Hilbert problems (Deift et al., 1995).

The quantized Toda lattice is solvable; the eigenstates of the multi-particle Toda Hamiltonian are matrix elements of infinite-dimensional group representations. The construction generalizes the one-dimensional case, $-\mathrm{d}^2/\mathrm{d}q^2 + \exp(-2q)$, whose eigenfunctions are Whittaker functions (Semenov-Tyan-Shansky, 1994a).

Orthogonal polynomials satisfy a three-term recurrence relation such as (6). For this reason, the Toda lattice arises in random matrix theory. A probability measure

$$\mathrm{d}\mu_N(H) = Z_N^{-1} \mathrm{e}^{-N \mathrm{Trace}\, V(H)}\, \mathrm{d}H$$

on Hermitean matrices is given. It determines a family $p_j(x)$ of orthogonal monic polynomials, $\int p_i(x) p_j(x)\, \mathrm{d}\mu_N(x) = 0$ if $i \neq j$. In this basis, the shift operator $Lp(x) = xp(x)$ acting on polynomials $p(x)$ has the Jacobi form (12). If $V(x)$ depends on parameters t_k, for example, $V(x) = x^2 + t_4 x^4$, then $\mathrm{d}\mu_N$ and the p_j depend on the t_k. The change of L in the moving basis $\{p_j\}$ is described by a Lax equation (Witten, 1991).

HERMANN FLASCHKA

See also **Hele-shaw cell; Integrable lattices; Lie algebras and lie groups; Zero-dispersion limits**

Further Reading

Bergvelt, M. & ten Kroode, F. 1988. τ-functions and zero curvature equations. *Journal of Mathematical Physics*, 29: 1308–1320

Brockett, R. & Bloch, A. 1990. Sorting with the dispersionless limit of the Toda lattice. In *Proceedings of the Conference on Hamiltonian Systems, Transformation Groups and Spectral Transform Methods*, edited by J. Harnad & J. Marsden, Montréal: Publications CRM

Corrigan, E. 1999. Recent developments in affine Toda field theory. In *Particles and Fields*, edited by G.E. Semenoff & L. Vinet, New York: Springer

Deift, P., Kriecherbauer, T. & Venakides, S. 1995. Forced lattice vibrations, I, II, *Communications in Pure and Applied Mathematics*, 48: 1187–1249, 1251–1298

Deift, P., Li, L.-C., Nanda, T. & Tomei, C., 1986. The Toda lattice on a generic orbit is integrable, *Communications in Pure and Applied Mathematics*, 39: 183–232

Fordy, A. & Woods, J.C. (editors). 1994. *Harmonic Maps and Integrable Systems*. Braunschweig/Wiesbaden: Viehweg

Guest, M.A. 1997. *Harmonic Maps, Loop Groups, and Integrable Systems*, Cambridge and New York: Cambridge University Press

Kamvissis, S. 1993. On the long time behavior of the doubly infinite Toda lattice. *Communications in Mathematical Physics*, 153: 479–519

Kostov, I.K., Krichever, I., Mineev-Weinstein, M., Wiegmann, P.B. & Zabrodin, A. 2001. τ-function for analytic curves. In *Random Matrix Models and Their Applications*, edited by P. Bleher & A. Its, Mathematical Sciences Research Institute Publications, Cambridge University Press

Olshanetsky, M.A. & Perelomov, A.M. 1994. Integrable systems and finite-dimensional Lie algebras. In *Encyclopaedia of Mathematics, Dynamical Systems VII*, edited by V.I. Arnol'd & S. P. Novikov, New York: Springer (original Russian edition 1987)

Reyman, A.G. & Semenov-Tyan-Shansky, M.A. 1994. Group theoretical methods in the theory of finite dimensional integrable systems. In *Encyclopaedia of Mathematics, Dynamical Systems VII*, edited by V.I. Arnol'd & S.P. Novikov, New York: Springer (original Russian edition 1987)

Semenov-Tyan-Shansky, M.A. 1994a. Lectures on R–matrices, Poisson–Lie groups, and integrable systems. In *Lectures on Integrable Systems, In Memory of Jean-Louis Verdier*, edited by O. Babelon, P. Cartier & Y. Kosmann-Schwarzbach, Singapore: World Scientific

Semenov-Tyan-Shansky, M.A. 1994b. Quantization of open Toda lattices. In *Encyclopaedia of Mathematics, Dynamical Systems VII*, edited by V.I. Arnol'd & S.P. Novikov, New York: Springer (original Russian edition 1987)

Takasaki, K. & Takebe, T. 1995. Integrable hierarchies and dispersionless limits. *Reviews in Modern Physics*, 7: 743–808

Toda, M. 1981. *Theory of Nonlinear Lattices*, Berlin and Heidelberg: Springer (original Japanese edition 1978)

Ueno, K. & Takasaki, K. 1984. Toda lattice hierarchy. In *Group Representations and Systems of Differential Equations*, edited by K. Okamoto, Tokyo: Kinokuniya

Witten, E. 1991. Two-dimensional gravity and intersection theory on moduli space. In *Surveys in Differential Geometry*, Bethlehem, PA; Lehigh University

TOPOLOGICAL CHARGE

See **Sine-Gordon equation**

TOPOLOGICAL CONJUGACY

See **Maps**

TOPOLOGICAL DEFECTS

A *topological defect* (or *topological soliton*) represents a spatially non-uniform configuration of an order parameter field that offers topological stability and cannot be transformed into the ground state of a system under finite deformation of a field (Mermin, 1979). The structure and properties of a topological defect (TD) depend essentially on the dimensionality of a system, its symmetry, and degeneration of the ground state. Historically, vortices in liquid were the first example of a TD to be investigated (Lugt, 1996). In a two-dimensional (2-d) incompressible liquid, the equation

$\Delta\varphi = 0$ for the velocity potential φ (velocity $\boldsymbol{v} = \nabla\varphi$) has the evident solution $\varphi = \kappa\chi$ for the vortex, where κ is an arbitrary parameter and χ is the azimuthal angle in cylindrical coordinates (ρ, χ) in the x, y plane with its origin in the vortex center.

Pitaevskii vortices in superfluidity theory (Lifshitz & Pitaevskii, 1980), *magnetic vortices* in easy-plane ferro- and anti-ferromagnets (FMs and AFMS) (Mertens & Bishop, 2000), magnetic *disclinations*, and disclinations in nematic liquid crystals (de Gennes, 1974) exemplify the similar one-dimensional TD in 3-d continuous media with continuous degeneration of the ground state. The order parameters in the above examples are of the two-component type: the complex thermodynamical wave function $\psi(\boldsymbol{r}, t) = \psi_0 \exp(\mathrm{i}\phi)$ of a Bose condensate for the superfluid liquid, which, within the approach of a weakly nonideal Bose gas, satisfies the Gross–Pitaevskii equation

$$\mathrm{i}\hbar\frac{\partial\psi}{\partial t} + \frac{\hbar^2}{2m}\Delta\psi + U\{\psi - |\psi|^2\psi\} = 0, \qquad (1)$$

and two angle variables (θ, ϕ) in polar coordinates in magnetic space (associated with the hard axis z) of a magnetization vector $\boldsymbol{M} = M_0(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, which satisfies the Landau–Lifshitz equation (similar in structure to (1)), or the director vector $\boldsymbol{L}$ in antiferromagnets and liquid crystals. These parameters are continuously degenerate in the phase of wave function ψ or in the direction of spins (or elongated molecules), defined by the angle ϕ in the easy-plane.

The solutions for the discussed TD map a plane (x, y), perpendicular to the defect line in coordinate space, onto the 2-d manifold of the order parameter (complex plane of ψ, half-sphere of radius $|\boldsymbol{M}|$, or sphere of radius $|\boldsymbol{L}|$). For example, the solution of (1) has the form $\psi = \psi_0(\rho)\exp(\mathrm{i}n\chi)$, where $n = \pm 1, \pm 2, \cdots$ and $(\pm)$ correspond to the vortex and antivortex. The density of a Bose gas tends to zero in the center of the vortex, $\psi^2(0) = 0$, and the solution has no singularity. A magnetic vortex has the same properties: $\phi = n\chi (n = \pm 1, \pm 2\cdots)$, $\theta(\rho = 0) = 0$ or π, $\theta(\rho \to \infty) \to \pi/2$ (Kosevich et al., 1990). In AFM- and nematic disclinations $\phi = k\chi/2$, where $k = \pm 1, \pm 2\cdots$, is called the Frank index. The mapping degree of TD is usually characterized by some integral *topological invariant* (or *topological charge*) related to the solution under consideration. The hydrodynamic vortex in an incompressible liquid is defined by its total vorticity $(1/2\pi)\oint \boldsymbol{v}\, \mathrm{d}\boldsymbol{l}$, where integration is performed over the contour enclosing the vortex center. This value coincides with κ and can be arbitrary. Other types of vortices can be characterized by the same integral $\oint \nabla\phi\, \mathrm{d}\boldsymbol{l}$. (Since a contour can be chosen with an infinite radius, these topological solitons are nonlocalized.) But in some cases, the topological charge is more conveniently defined as the 2-d integral over the coordinate plane. For example, in a ferromagnet, this invariant is

$$\begin{aligned} Q &= \frac{1}{4\pi}\int m\left(\frac{\partial m}{\partial x_\alpha} \times \frac{\partial m}{\partial x_\beta}\right)\varepsilon_{\alpha\beta}\, \mathrm{d}^2x \\ &= \frac{1}{2\pi}\int \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\phi, \end{aligned} \qquad (2)$$

where $m = \boldsymbol{M}/M_0$. For a magnetic vortex $Q = pn$, where $p = m_z(\rho = 0)$, is the polarity of the vortex.

Another situation appears in systems with discrete degeneracy of the ground state. For instance, a localized magnetic 2-d topological soliton (TS) (*magnetic skyrmion*) of the type $\phi = n\chi$, $\theta = \theta(\rho)$, and $\theta(0) = 0$, $\theta(\infty) = \pi$ can exist in the easy-axis ferromagnet with two equivalent ground states $m_z = \pm 1$. At fixed n, the topological charge for this skyrmion is twice that for the vortex. Moreover, in such a system, TDs can exist in the 1-d and 3-d cases. In 3-d easy-axis FM, the TS, localized in all 3 dimensions, corresponds to the solution with the nonzero integer Hopf invariant. The 1-d TS can exist in 1-d and quasi-1-d systems or in 2-d and 3-d media as solutions depending on one spatial coordinate. Such a TD describes the *domain wall* in ferromagnets. In the framework of the Landau–Lifshitz equation, the corresponding *kink* solution is written as

$$\theta(x) = \arccos(\tanh(x/l_0)), \quad \phi = \text{const.}, \qquad (3)$$

where l_0 is the magnetic length. The domain boundary (kink) separates two half-spaces in different ground states. The topological charge, analogous to (2), can be defined in this case as follows: $Q = (1/2\pi)\int \sin\theta(x)\, \mathrm{d}\theta(x)\, \mathrm{d}\phi(x)$. If an additional anisotropy in the plane perpendicular to the easy axis is taken into account (in orthorhombic ferromagnets), more complicated TDs can exist inside the domain walls: *Bloch lines* and *Bloch points*. Similar to (3), the kink-like TD can exist in other systems with discrete degenerate ground states: kinks in antiferromagnets, kinks of incommensurate surface structures, fluxons in a long Josephson junction, phase boundaries in the problem of structural phase transitions, solitons in polyacetylene, in 1-d metals in the Peierls–Fröhlich phase, and so on. Usually these TDs are investigated within the framework of the φ^4-model or sine-Gordon equation

$$\frac{\partial^2 w}{\partial t^2} - s^2\frac{\partial^2 w}{\partial x^2} + \omega_0^2 \sin w = 0, \qquad (4)$$

where the field variable w and coefficients have different physical meanings for various systems. The topological solution to this equation is well known: $w = 4\arctan\exp((x - vt)/l\sqrt{1 - v^2/s^2})$ with $l = s/\omega_0$.

A crystal *dislocation* represents one of the most important examples of a TD in a crystal lattice. Dislocations exist due to the translational symmetry of the lattice: it transforms into itself under the translation by the interatomic distance a. (Rotational symmetry of the crystal lattice leads to the existence of another TD—crystal *disclination*.) A dislocation represents a one-dimensional TD with the following properties: the regular lattice structure is distorted only in the core of the dislocation line and the total displacement, as a closed contour goes around the dislocation line, is equal to the translation period. (In two-sublattice antiferromagnets, this translation period is twice as small as that for a magnetic lattice.) This fact leads to the appearance of the magnetic disclination and formation of the complex magnet-structural TD (Kovalev & Kosevich, 1977). In the simplest case of a screw dislocation or in a scalar model, the deformation is characterized solely by one component of displacement, u. Then, within the elasticity theory approximation, the deformation of a crystal is governed by the equation $\triangle u = 0$ with the dislocation solution $u = a\chi/2\pi$. (Here, a plays the role of the topological charge of the dislocation: $\oint du = a$.) The above solution is singular and does not describe the discrete structure of the dislocation core. It can be investigated in the simple one-dimensional model proposed by Yakov Frenkel and Tatiana Kontorova (1938) (*See also* **Frenkel–Kontorova model**). The FK model was the first 1-d model for 2-d topological defects. Within this model, the relative displacements u of the atoms from two atomic rows above and below the dislocation center are described by Equation (4), where $w = 2\pi u/a$, $\omega_0 = (2\pi/a)\sqrt{U/m}$, m is the atomic mass, s is the velocity of sound, U is the energy of the interaction between the rows, and l is the width of the dislocation core. The energy of a stationary dislocation (kink) is $E_0 = (4/\pi)\sqrt{ms^2U}$. The integral $\int \partial u/\partial x \, dx$ now plays the role of the topological charge. The FK model describes more adequately other TDs in the crystal lattice: *crowdions*, kinks of incommensurate surface structures, and dislocation kinks.

The dynamics of TDs are of great variety and depend strongly on the nature of the TD. The dynamics of dislocations, dislocation kinks, crowdions, domain walls, and phase boundaries are very simple within the framework of the Lorentz-invariant 1-d FK model: the kink can move with velocities below the velocity of sound in a system and its effective mass can be considerably smaller than the atomic mass. Actually, the dynamics of TDs in a lattice is much more complicated. Dislocations move mainly in some preferred direction (in slip planes), and crowdions propagate in closely packed atomic rows in a "relay race" manner. The discreteness of a lattice gives rise to Peierls relief for TP, and its dynamics assume diffusive features. Dislocation creeps over this relief, producing the dislocation kinks. The dynamics of vortices in media with distributed parameters has some interesting peculiarities. An isolated vortex in an infinite ideal medium cannot move: it is frozen in a liquid or superfluid flow or spin flux. The vortex can move in a bounded area or in the presence of other defects but these dynamics are non-Newtonian. Within the collective coordinates approach, the effective equation of motion for the center of a vortex $\boldsymbol{X}(t)$ is (Thiele, 1973)

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t} \times \boldsymbol{G} = \boldsymbol{F}_{\mathrm{G}}, \tag{5}$$

where the force $\boldsymbol{F}_{\mathrm{G}}$ is formally equivalent to the Magnus force in fluid dynamics and the gyrocoupling vector $\boldsymbol{G}$ is parallel to the line of a vortex and depends on its topological charge.

In all the examples above, topological defects with the opposite sign of the topological charge exist (vortex-antivortex, kink-antikink, dislocations with opposite signs, etc). This implies that TDs can emerge from the ground state of a medium in pairs with zero total topological charge. At a nonzero temperature, these pairs dissociate and the finite density of TDs can be observed. Seeger & Schiller (1966) were the first to develop the thermodynamics of topological solitons in the framework of the 1-d FK model. As they showed, the equilibrium density of kinks and antikinks at a temperature T is

$$n \approx \frac{a}{l}\sqrt{\frac{2}{\pi\tau}}\exp\left(-\frac{1}{\tau}\right), \tag{6}$$

where $\tau = k_{\mathrm{B}}/E_0$, E_0 is the soliton energy, and l is the width of this kink. The situation with a TD in the 2-d case (vortices and 2-d dislocations) is essentially different. The energy of vortices and dislocations in infinite 2-d systems is infinite, but in systems of size R, it is of the order of $E_v \sim E_* \ln(R/a)$, where E_* is a characteristic energy, being specific for different TDs. The contribution from a vortex or dislocation to configuration entropy in a 2-d crystal is $\delta S = \ln(R/a)^2$, and hence, the change in free energy, if one TD is added, is $\delta F = (E_* - 2T)\ln(R/a)$. At the temperature $T_{\mathrm{c}} = E_*/2$, the value δF becomes negative and a Berezinskii–Kosterlitz–Thouless phase transition takes place: the crystal melts or magnetic ordering is broken (Berezinskii, 1971).

As indicated above, topological defects play an important role in the kinetic and thermodynamic properties of condensed matter. Dislocations and dislocation kinks cause plasticity and strengthening of a crystal, and their behavior under radiation depends to a large extent on the crowdions. Recently, the influence of dislocations on the concentration and mobility of current carriers has been widely studied experimentally in pure semiconductor and alkali-halide crystals (Osip'yan et al., 2000). Topological defects (charge-density-wave solitons) make an essential contribution to the conductivity and electrodynamics of quasi-1-d conductors such as $(CH)_x$, TaS_2, and $NbSe_2$ (Krive et al., 1986).

ALEXANDER S. KOVALEV

See also **Collective coordinates; Dislocations in crystals; Domain walls; Frenkel–Kontorova model; Landau–Lifshitz equation; Liquid crystals; Long Josephson junctions; Multidimensional solitons; Nonlinear Schrödinger equations; Sine-Gordon equation; Spin systems; Superfluidity; Topology; Vortex dynamics of fluids**

Further Reading

Berezinskii, V.L. 1971. Violation of long range order in one-dimensional and two-dimensional systems with a continuous symmetry group. I. Classical systems. *Soviet Physics-JETP*, 34: 610

de Gennes, P.G. 1974. *The Physics of Liquid Crystals*, Oxford: Clarendon Press

Frenkel, J. & Kontorova, T. 1938. On the theory of plastic deformation and twinning. *Physikalishe Zeitschrift der Sowjetunion*, 13: 1–12

Kosevich, A.M., Ivanov, B.A. & Kovalev, A.S. 1990. Magnetic solitons. *Physics Reports*, 194: 117–238

Kovalev, A.S. & Kosevich, A.M. 1977. Dislocation and domains in an antiferromagnet. *Soviet Journal of Low Temperature Physics*, 3: 125–126

Krive, I.V., Rozhavsky, A.S. & Kulik, I.O. 1986. Nonlinear conductivity mechanisms and electrodynamics of quasi-1D conductors in Peierls dielectric phase. *Soviet Journal of Low Temperature Physics*, 12: 635

Lifshitz, E.M. & Pitaevskii, L.P. 1980. *Statistical Physics*, part 2, Oxford: Pergamon Press

Lugt, H.J. 1996. *Introduction to Vortex Theory*, Potomac, MD: Vortex Flow Press

Mermin, N.D. 1979. The topological theory of defects in ordered media. *Review of Modern Physics*, 51: 591–648

Mertens, F.G. & Bishop, A.R. 2000. Dynamics of vortices in two-dimensional magnets. In *Nonlinear Science at the Dawn of the 21st Century*, Berlin and New York: Springer

Osip'yan, Yu.A. et al. 2000. *Electronic Properties of Dislocations in Semiconductors*. Moscow: Editorial USSR (in Russian)

Seeger, A. & Schiller, P. 1966. *Physical Acoustics*, New York: Academic Press

Thiele, A.A. 1973. Steady-state motion of magnetic domains. *Physical Review Letters*, 30: 239–233

TOPOLOGICAL ENTROPY

See **Entropy**

TOPOLOGICAL SOLITONS

See **Solitons, types of**

TOPOLOGY

Continuity is conventionally associated with functions defined on the real line or higher-dimensional Euclidean spaces. Topology is concerned with the abstraction of continuity to maps between more general sets. The subject is vast and its study can take many forms depending on the nature of the structures considered. They include the areas of point-set, combinatorial, algebraic, and differential topology.

A topological space has a distinguished collection of subsets known as *open sets*. An open subset U of the real line $\mathbb{R}$ is one for which every point of U is a subset of some real interval wholly contained in U. Thus, every point of U is the *interior* of U. Openness of a set U can also be expressed using the Euclidean distance, or *metric*, by saying that for every point $x \in U$, all points within a sufficiently small distance of x also lie in U. Thus, metrics can be used to create open sets. Complements of open sets are said to be *closed*. The simplest examples of open and closed sets in $\mathbb{R}$ are, respectively, the "open interval" (a, b) of all points between the numbers a and b excluding the end points, and the "closed interval" $[a, b] = (a, b) \cup \{a, b\}$. Also, sets can be neither open nor closed, for example $[a, b) = (a, b) \cup \{a\}$.

A *topology* on a set X is defined by its collection of open subsets, τ, which then makes X a *topological space* (X, τ). The collection of open subsets must include both X and the empty set ϕ, any union of elements of τ, and the intersection of any finite collection of elements of τ. The conditions for a topology can equally well be cast in terms of closed sets. Also, a set can have many different topologies.

In any sophisticated mathematical structure, there is usually a way of relating two objects. For example, if we are only considering sets, say X and Y, we consider maps $f : X \to Y$. The natural equivalence for sets X,Y would be the existence of a map $f : X \to Y$ which is both one–one and onto, that is, a bijection. When topologies are placed on X and Y, it is natural to consider maps $f : X \to Y$ that are continuous. The metric definition of continuity for maps $f : \mathbb{R} \to \mathbb{R}$, or more generally $f : \mathbb{R}^m \to \mathbb{R}^n$, can be shown to be equivalent to the topological definition: $f : X \to Y$ is continuous if V is an open subset of Y; the set $f^{-1}(V)$ is an open subset of X.

This alternative definition of continuity is the one that makes topology a key mathematical discipline of widespread importance.

The corresponding equivalence for topological spaces X and Y requires a map $h : X \to Y$, which is both (i) a bijection, and (ii) bicontinuous; that is, both h and its inverse h^{-1} are continuous. Such a map is called a *homeomorphism*. The spaces X and Y are said to be topologically equivalent or homeomorphs. Any subset S of a topological space X can be made into a topological space by declaring the intersections of open sets of X with S to be open sets of S. In fact, this collection of subsets of S forms a topology on S, called the *subspace topology*. Thus, important geometrical objects which are subsets of Euclidean spaces such as the circle, the sphere (and therefore all classical polyhedra), the torus, the pretzel, and the Klein bottle (see Figure 1) can all be seen as topological spaces when endowed with the subspace topology.

An important property of continuous functions defined on the real numbers is that a continuous function defined on a bounded closed interval attains its bounds; that is, there exist points at which the function takes its maximum and minimum value. This is not true if the "closed, bounded" condition is relaxed. For example, $f(x) = x$ is not bounded on the real line $\mathbb{R}$ which is a closed (but not bounded) set. Also, $f(x) = x$ does not attain its bounds on the bounded (but not closed) set $(0, 1)$. The bounded closed interval on $\mathbb{R}$ is called a *compact* set. Again, such a set can be defined solely in terms of open sets, and so we can define the concept of a *compact topological space* (Munkres, 2000).

Another key result in elementary analysis is that the continuous image of an interval of real numbers is also an interval. This property is often used to find roots of a continuous function $f: \mathbb{R} \to \mathbb{R}$ by finding values $a, b \in \mathbb{R}$ for which $f(a) \cdot f(b) < 0$, the so-called "intermediate value" theorem. In the generalization of this result to topological spaces, the key property of the interval is its *connectedness*. The analogous result for topological spaces is that the continuous image of a connected set is also a connected set. The characterization of connectedness in the real numbers can be described purely in terms of the properties of the open sets on the real line, that is, in terms of the Euclidean topology on $\mathbb{R}$.

Thus, both compactness and connectedness are topological properties in the sense that they can be described purely in terms of properties of the open sets of a topological space (Munkres, 2000).

In some areas of topology, the importance of the open sets is not so apparent and other features of the topological space are considered. For example, polyhedra such as the cube, tetrahedron, and dodecahedron are finite ways of building homeomorphic images of the standard sphere. We note that for all such constructs, the number of vertices (V), edges (E), and faces (F) satisfy the condition $V - E + F = 2$, the *Euler characteristic* of the sphere. The torus, when built up in terms of faces, edges, and vertices, has the property that its Euler characteristic is zero. Given that the numbers V, E, and F are conserved by homeomorphism, we see that the torus and sphere having different Euler characteristics not only makes them look "different," but ensures that they are not homeomorphic; that is, they are topologically distinct.

Note that not all spaces can be easily distinguished using topology. For example, the subsets of rational numbers, $\mathbb{Q}$, and the irrationals, I, of the real line $\mathbb{R}$, are both topologically *dense* sets in $\mathbb{R}$; that is, for both sets, the smallest closed superset is the whole interval $\mathbb{R}$. Also, both are neither open nor closed. However, note that there is no bijection between the sets $\mathbb{Q}$ and I because they have different cardinalities and, thus, cannot be homeomorphs.

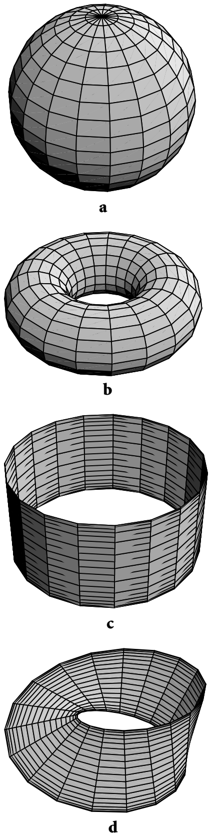

Figure 1. Four distinct topological spaces: (a) sphere; (b) torus; (c) cylinder; (d) Möbius band. See http://library.wolfram.com/graphics/

The topological spaces mentioned above, such as the torus, sphere, and pretzel, can be easily visualized within the three-dimensional Euclidean space $\mathbb{R}^3$ in which we locally live. However, it is not difficult to see how we can start to construct spaces in $\mathbb{R}^3$ which are more geometrically demanding. The simplest is the Möbius strip, M, which is derived from the rectangle by pasting together one opposite pair of edges of a band with a half-twist, see Figure 1(d). Note that the Möbius band has only one "side"; just draw a pen line along the spine of the band and the line arrives on the opposite side of the paper from its initial point. Continuing the line brings a return to the initial point of the curve. By comparison, one cannot get from one side of a cylinder to the other without passing across one of its edges. The Möbius band is topologically different from the cylinder for several reasons. For instance, the boundary of the cylinder consists of two disjoint circles whereas that of the Möbius band is a single circle.

DAVID ARROWSMITH

Further Reading

Dieudonné, J. 1985. The beginnings of topology from 1850 to 1914. *Proceedings of the Conference on Mathematical Logic 2 (Siena, 1985)*, 585–600

Dieudonné, J. 1989. *A History of Algebraic and Differential Topology, 1900–1960*, Boston: Birkhäuser

Lefschetz, S. 1970. The early development of algebraic topology. *Boletim da Sociedade Brasileira de Matematica*, 1(1): 1–48

Mendelson, B. 1990. *Introduction to Topology*, New York: Dover

Munkres, J.R. 2000. *Topology*, Upper Saddle River, NJ: Prentice-Hall

Stillwell, J.C. 1993. *Classical Topology and Combinatorial Group Theory*, 2nd edition, New York: Springer

Weil, A. 1979. Riemann, Betti and the birth of topology. *Archive for the History of Exact Sciences*, 20(2): 91–96

MATHEMATICA (http://library.wolfram.com/graphics/) Graphics of various surfaces

TRAFFIC FLOW

Most automobile drivers have encountered the widespread phenomenon of so-called "phantom traffic jams," for which there is no visible reason such as an accident or a bottleneck. Why are vehicles sometimes stopped although everyone likes to drive fast? Due to the finite adaptation time (= reaction time + acceleration time), a small disturbance in the traffic flow Q_* can cause an overreaction (overbraking) of a driver, if the safe vehicle speed $V_*(\rho)$ drops too rapidly with increasing vehicle density ρ. At high enough densities ρ, this will give rise to a chain reaction of the followers, as other vehicles will have approached before the original speed can be regained. This feedback can eventually cause the unexpected standstill of vehicles known as traffic jam.

Lighthill & Whitham (1955) described traffic flow as a function of space x and time t by means of a fluid-dynamic conservation law of vehicles, reflecting the fact that vehicles are not generated or lost in the absence of ramps, intersections, or accidents. The traffic flow $Q(x,t)$ (= vehicle density $\rho(x,t) \times$ average velocity $V(x,t)$) was specified as a function of the density $\rho(x,t)$. The corresponding "fundamental diagram" $Q_*(\rho) = \rho V_*(\rho)$ is obtained as a fit to empirical data. The conservation law $\partial\rho/\partial t + \partial Q/\partial x = 0$ leads to the nonlinear wave equation

$$\frac{\partial\rho}{\partial t} + C(\rho)\frac{\partial\rho}{\partial x} = 0\,, \tag{1}$$

according to which the propagation velocity

$$C(\rho) = V_*(\rho) + \rho\frac{\mathrm{d}V_*(\rho)}{\mathrm{d}\rho} \le V_*(\rho) \tag{2}$$

of kinematic waves depends on the vehicle density. Thus, while a density profile on a ring road keeps its amplitude, its shape is changing until shock waves (i.e., discontinuous changes in the density) have developed. The densities ρ_+ and ρ_- immediately upstream and downstream of a shock front determine its propagation speed

$$S(\rho_+,\rho_-) = \frac{Q_*(\rho_+) - Q_*(\rho_-)}{\rho_+ - \rho_-}\,. \tag{3}$$

As discontinuous density changes are not fully consistent with empirical observations and a problem for efficient numerical integration, Whitham (1974) has suggested adding a diffusion term $D\,\partial^2\rho/\partial x^2$ with $D > 0$ to the right-hand side of the Lighthill–Whitham equation. For the linear velocity-density relation $V_*(\rho)$ suggested by Greenshields (1935), the resulting equation is equivalent with the Burgers equation and can be transformed into the linear heat or diffusion equation; that is, it is analytically solvable.

Experimental observations of traffic patterns show some additional features that cannot be reproduced by the above models. While traffic flow appears to be stable with respect to perturbations at small and large densities, there is a linearly unstable range at medium densities, where already small disturbances of uniform traffic flow give rise to traffic jams. Between these three density ranges, one finds meta- or multistable ranges, since there exists a density-dependent, critical amplitude $A(\rho)$, so that the resulting traffic pattern is path- or history-dependent (Kerner et al., 1994–1997). While subcritical perturbations fade away, supercritical perturbations cause a breakdown of traffic flow (nucleation effect). Consequently, traffic flows display critical points, nonequilibrium phase transitions, noise-induced transitions, and fluctuation-induced ordering phenomena.

One may view this situation as nonequilibrium analogue of the phase transitions between vapor, water, and ice. However, the breakdown and structure formation phenomena, when the "temperature" (i.e., the fluctuation strength) is increased, are sometimes counterintuitive due to the repulsive nature of vehicular interactions. From classical many-particle systems with attractive interactions, we are rather used to the idea that increasing temperature breaks up structures and destroys patterns (fluid structures are replaced by gaseous ones, not by solid ones).

The above observations in freeway traffic can be described by microscopic, mesoscopic, or macroscopic models, which are theoretically connected by means of a micro-macro link. Microscopic models are usually follow-the-leader models specifying the acceleration $\mathrm{d}v_i/\mathrm{d}t$ of the single vehicles i as a function of their distance headway $d_i = x_{i-1} - x_i$, their speed v_i, and/or their relative velocity $\Delta v_i = v_i - v_{i-1}$:

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = f(d_i, v_i, \Delta v_i)\,. \tag{4}$$

A typical example is the non-integer car-following model:

$$\frac{\mathrm{d}v_i(t+\Delta t)}{\mathrm{d}t} = -\frac{\Delta v_i(t)}{T}\frac{[v_i(t+\Delta t)]^m}{[d_i(t)]^l} \tag{5}$$

with the reaction time $\Delta t \approx 1.3$ s and the parameters $T \approx \Delta t/0.55$, $m \approx 0.8$, and $l \approx 2.8$ (Gazis et al., 1961). It has a linearly unstable range for $\Delta t/T > \frac{1}{2}$. A simpler model is the optimal velocity model

$$\frac{\mathrm{d}v_i(t)}{\mathrm{d}t} = \frac{1}{\tau}\Big[v\big(d_i(t)\big) - v_i(t)\Big], \tag{6}$$

where $v(d_i)$ is the "optimal" velocity-distance relation and τ the adaptation time (Bando et al., 1994, 1995). This model has an unstable range for $\mathrm{d}v(d_i)/\mathrm{d}d_i > 1/(2\tau)$. The respective nonlinearly coupled differential equations (or stochastic differential equations, if fluctuations are taken into account) are numerically solved as in molecular dynamics. An alternative approach is rule-based cellular automata, which discretize space and time in favor of numerical efficiency: $t = i\,\Delta t, x = j\,\Delta x, d = \hat{d}\,\Delta x, v = \hat{v}\,\Delta x/\Delta t$. The Nagel–Schreckenberg model (1992), for example, can be written in the form

$$\hat{v}_{i+1} = \max\Big(0, \min(\hat{v}_{\max}, \hat{d}_i - 1, \hat{v}_i + 1) - \xi_i^{(p)}\Big), \tag{7}$$

where $\hat{v}_{\max}\Delta x/\Delta t$ is the maximum velocity and $\xi_i^{(p)}$ a Boolean random variable which is 1 with probability p and 0 otherwise. Typical parameters are $\Delta t = 1$ s, $\Delta x = 7.5$ m, $\hat{v}_{\max} = 5$, and $0.2 < p \leq 0.5$.

Mesoscopic models describe the spatiotemporal change of the phase space density (= vehicle density × velocity distribution). This approach has been introduced by Prigogine et al. (1960, 1961, 1971) and is inspired by kinetic gas theory. The related equations are either of Boltzmann type (for point-like vehicles or low densities) or of Enskog type, if vehicular space requirements at moderate and high densities are taken into account (Helbing et al., 1995–1999). The equations allow a systematic derivation of a hierarchy of macroscopic equations for the vehicle density $\rho(x,t)$, the average velocity $V(x,t)$, the velocity variance $\Theta(x)$, etc. This hierarchy is usually closed after the velocity or variance equation, although the separation of time scales assumed by the underlying approximations is weak. Nevertheless, the observed traffic dynamics are rather well reproduced by the resulting coupled partial differential equations. The density equation is just the continuity equation

$$\frac{\partial\rho}{\partial t} + \frac{\partial(\rho V)}{\partial x} = \nu_+ - \nu_-\,, \tag{8}$$

where ν_+ and ν_- denote on- and off-ramp flows, respectively. The velocity equation can be cast into the form

$$\frac{\partial V}{\partial t} + V\frac{\partial V}{\partial x} = -\frac{1}{\rho}\frac{\partial P}{\partial x} + \frac{1}{\tau}(V^* - V). \tag{9}$$

In theoretically consistent macroscopic traffic models such as the gas-kinetic-based traffic model, the "traffic pressure" P and the velocity V^* are nonlocal functions of the density ρ, the average velocity V, and the variance Θ (Helbing et al., 1998, 1999). The Lighthill–Whitham model (1955) results in the unrealistic limit $\tau \to 0$ of vanishing adaptation times τ. Payne's macroscopic traffic model (1971, 1979) is obtained for $P(\rho) = [V_0 - V_*(\rho)]/(2\tau)$ and $V^* = V_*(\rho)$, where V_0 denotes the (average) desired velocity (the average velocity at very low densities). Kerner and Konhäuser's model (1993) is a variant of Kühne's model (1984) and corresponds to the specifications $P = \rho\Theta_0 - \eta_0\partial V/\partial x$, and $V^* = V_*(\rho)$, where Θ_0 and η_0 are positive constants. The corresponding equation is a Navier–Stokes equation with a viscosity term $\eta_0\partial^2 V/\partial x^2$ and an additional relaxation term $[V_*(\rho) - V]/\tau$ describing the delayed adaptation to the velocity–density relation $V_*(\rho)$. The condition for linear instability reads

$$\rho\left|\frac{\mathrm{d}V_*(\rho)}{\mathrm{d}\rho}\right| > \sqrt{\frac{\mathrm{d}P(\rho)}{\mathrm{d}\rho}}\,; \tag{10}$$

that is, the Payne model and the Kerner–Konhäuser model have linearly unstable ranges, if $\mathrm{d}V_*(\rho)/\mathrm{d}\rho$ is large, while the Lighthill–Whitham model is marginally stable. The Burgers equation, by the way, is always stable.

According to Krauß (1998), traffic models show the observed hysteretic phase transition related with metastable traffic and high flows only, if the typical maximal acceleration is not too large and the deceleration strength is moderate. In such models, the outflow Q_{out} from traffic jams is a self-organized constant of traffic flow (Kerner et al., 1994, 1996). It corresponds approximately to the intersection point of the linear jam line

$$J(\rho) = \frac{1}{T}\left(1 - \frac{\rho}{\rho_{\text{jam}}}\right) \tag{11}$$

with the free branch of the flow-density diagram, where T denotes the time headway in congested traffic and

ρ_{jam} the density inside traffic jams. The jam line corresponds to the flow-density relation for moving traffic patterns with a self-organized, stationary profile (Kerner & Konhäuser, 1994). These propagate with the velocity $C = -1/(T\rho_{\text{jam}}) \approx -15$ km/h, which is another traffic constant. (Once a traffic jam is fully developed, it moves upstream with constant velocity, as vehicles leave the downstream jam front at a constant rate, while new ones join it at the upstream front.)

With this knowledge, one can understand the various congested traffic states observed on freeway sections with bottlenecks (Helbing et al., 1998–2002). Let us assume a bottleneck due to ramp flows $\nu_+ = Q_{\text{rmp}}/(nL)$, where L is the used length of the on-ramp and n the number of freeway lanes. The corresponding bottleneck strength is, then, $\Delta Q = Q_{\text{rmp}}/n$. If Q_{up} denotes the traffic flow upstream of the bottleneck and $Q_{\text{tot}} = (Q_{\text{up}} + \Delta Q)$ is the total capacity required downstream of the ramp, we will eventually find a growing vehicle queue upstream of the bottleneck, if Q_{tot} is greater than the dynamic capacity Q_{out}. The traffic flow Q_{cong} resulting in the congested area plus the inflow or bottleneck; strength ΔQ are normally given by the outflow Q_{out}; that is,

$$Q_{\text{cong}} = Q_{\text{out}} - \Delta Q \tag{12}$$

(if vehicles cannot enter the freeway downstream of the congestion front). One can distinguish the following cases: If the density ρ_{cong} associated with the congested flow

$$Q_{\text{cong}} = Q_*(\rho_{\text{cong}}) \tag{13}$$

lies in the stable range, we find homogeneous congested traffic (HCT) such as typical traffic jams during holiday seasons or after serious accidents. For a smaller on-ramp flow or bottleneck strength ΔQ, the congested flow Q_{cong} is linearly unstable, and we either find oscillating congested traffic (OCT) or triggered stop-and-go traffic (TSG). In contrast to OCT, stop-and-go traffic is characterized by a sequence of moving jams, between which traffic flows freely. This state can either emerge from a spatial sequence of homogeneous and oscillating congested traffic (Koshi et al., 1983; called "pinch effect" by Kerner, 1998), or it can be caused by the inhomogeneity at the ramp. In the latter case, each traffic jam triggers another one by inducing a small

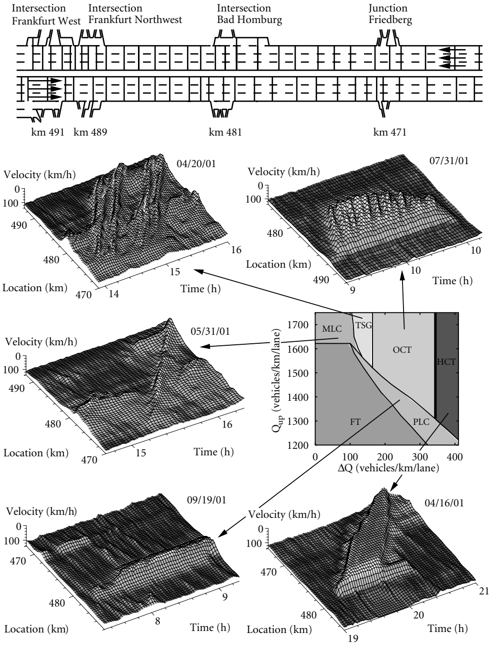

Figure 1. Numericlly determined phase diagram of traffic states in the presence of one bottleneck as a function of the upstream flow Q_{up} and the bottleneck strength ΔQ (center right), and empirical representatives for the related kinds of congested traffic (by Martin Schönhof and D. Helbing. Note that the outflow Q_{out} in the applied simulation model depends on the bottleneck strength ΔQ.

perturbation in the inhomogeneous freeway section (see Figure 1), which propagates downstream as long as it is small, but turns back when it has grown large enough (boomerang effect). This, however, requires the downstream traffic flow to be linearly unstable. If it is (meta-)stable instead (when the traffic volume Q_{tot} is further reduced), a traffic jam will usually not trigger a growing perturbation. In that case, one finds either a single moving localized cluster (MLC) or a pinned localized cluster (PLC) at the location of the ramp. The latter requires the traffic flow in the upstream section to be stable, so that no traffic jam can survive there. Finally, for sufficiently small traffic volumes Q_{tot}, we find free traffic (FT), as expected. For freeways with a single bottleneck and a large perturbation of traffic flow, these facts can be summarized by the phase diagram in Figure 1, which is universal for all microscopic and macroscopic, stochastic and deterministic traffic models with the same instability diagram (stable, metastable, and unstable density ranges). Results for more complex freeway geometries, other initial or boundary conditions, and other instability diagrams are available as well.

Current research focuses on the following open questions: Are fluctuations and psychological concepts necessary to understand the empirical observations in traffic flows? Can the large individual variation of time headways fully account for the large scattering of flow-density data in synchronized flow? (In congested traffic flow, the velocities in neighboring lanes are usually synchronized, as different speeds are balanced by lane changes.) What are the site- and country-dependent differences in traffic dynamics, and can they be adequately reflected by different model parameters for the driver-vehicle units? How can the insights regarding the laws of traffic dynamics be used for traffic optimization by variable speed limits, intelligent on-ramp controls, dynamic re-routing, and driver assistance systems? How can they be transferred to the explanation of breakdown and obstruction phenomena in socioeconomic systems?

DIRK HELBING

See also **Burgers equation; Constants of motion and conservation laws; Phase transitions; Shock waves**

Further Reading

Helbing, D. 2001. Traffic and related self-driven many-particle systems. *Reviews of Modern Physics*, 73(4): 1067–1141 [All the authors mentioned in the text above are found as references in this article.]

Lighthill, M.J. & Whitham, G.B. 1955. On kinematic waves: II. A theory of traffic on long crowded roads. *Proceedings of the Royal Society, London* A, 229: 317–345

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

TRAJECTORIES

See **Phase space**

TRANSITION TO CHAOS

See **Chaotic dynamics**

TRAVELING WAVE

See **Wave of translation**

TRIAD INTERACTION

See ***N*-wave interactions**

TUNNEL DIODE ARRAYS

See **Distributed oscillators**

TURBULENCE

Turbulence is a state of a nonlinear physical system that has energy distribution over many degrees of freedom strongly deviated from equilibrium. Turbulence is irregular both in time and in space. Turbulence can be maintained by some external influence or it can decay on the way to relaxation to equilibrium. The term first appeared in fluid mechanics and was later generalized to include far-from-equilibrium states in solids and plasmas.

If an obstacle of size L is placed in a fluid of viscosity ν that is moving with velocity V, a turbulent wake emerges for sufficiently large values of the Reynolds number

$$\mathrm{Re} \equiv VL/\nu.$$

At large Re, flow perturbations produced at scale L experience a viscous dissipation that is small compared with nonlinear effects. Nonlinearity then induces motions at smaller and smaller scales until viscous dissipation terminates the process at a scale much smaller than L, leading to a wide (so-called inertial) interval of scales where viscosity is negligible and nonlinearity plays a dominant role.

Examples of this phenomenon include waves excited on a fluid surface by wind or moving bodies and waves in plasmas and solids that are excited by external electromagnetic fields. The state of such a system is called turbulent when the wavelength of the waves excited greatly differs from the wavelength of the waves that dissipate. Nonlinear interactions excite waves in the interval of wavelengths (called the transparency window or inertial interval as in fluid turbulence) between the injection and dissipation scales.

The ensuing complicated and irregular dynamics require a statistical description based on averaging over regions of space or intervals of time. Because nonlinearity dominates in the inertial interval, it is natural to ask to what extent the statistics are universal, in the sense of being independent of the details of excitation and dissipation. The answer to this question is far from evident for nonequilibrium systems. A fundamental physical problem is to establish which statistical properties are universal in the inertial interval of scales and which are features of different turbulent systems.

Constraints on dynamics are imposed by conservation laws, and therefore, conserved quantities must play an essential role in turbulence. Although the conservation laws are broken by pumping and dissipation, these factors do not act in the inertial interval. Under incompressible turbulence, for example, the kinetic energy is pumped by external forcing and is dissipated by viscosity. As suggested by Lewis Fry Richardson in 1921, kinetic energy flows throughout the inertial interval of scales in a cascade-like process. The cascade idea explains the basic macroscopic manifestation of turbulence: the rate of dissipation of the dynamical integral of motion has a finite limit when the dissipation coefficient tends to zero. In other words, the mean rate of the viscous energy dissipation does not depend on viscosity at large Reynolds numbers. That means that symmetry of the inviscid equation (here, time-reversal invariance) is broken by the presence of the viscous term, even though the latter might have been expected to become negligible in the limit $\mathrm{Re} \to \infty$.

The cascade idea fixes only the mean flux of the respective integral of motion, requiring it to be constant across the inertial interval of scales. To describe an entire turbulence statistics, one has to solve problems on a case-by-case basis with most cases still unsolved.

Weak Wave Turbulence

From a theoretical point of view, the simplest case is the turbulence of weakly interacting waves. Examples include waves on the water surface, waves in plasma with and without a magnetic field, and spin waves in magnetics. We assume spatial homogeneity and denote by a_k the amplitude of the wave with the wave vector $\boldsymbol{k}$. When the amplitude is small, it satisfies the linear equation

$$\frac{\partial a_k}{\partial t} = -\mathrm{i}\omega_k a_k + f_k(t) - \gamma_k a_k. \tag{1}$$

Here, the dispersion law ω_k describes wave propagation, γ_k is the decrement of linear damping, and f_k describes pumping. For the linear system, a_k is different from zero only in the regions of $\boldsymbol{k}$-space where f_k is nonzero. To describe wave turbulence that involves wave numbers outside the pumping region, one must account for the interactions among different waves. Considering the wave system to be closed (no external pumping or dissipation), one can describe it as a Hamiltonian system using wave amplitudes as normal canonical variables (Zakharov et al., 1992). At small amplitudes, the Hamiltonian can be written as an expansion over a_k, where the second-order term describes noninteracting waves and high-order terms determine the interaction

$$\begin{aligned} H = \int \omega_k |a_k|^2 \, \mathrm{d}\boldsymbol{k} + \int \Big(V_{123} a_1 a_2^* a_3^* + \text{c.c.} \Big) \\ \delta(\boldsymbol{k}_1 - \boldsymbol{k}_2 - \boldsymbol{k}_3) \, \mathrm{d}\boldsymbol{k}_1 \, \mathrm{d}\boldsymbol{k}_2 \, \mathrm{d}\boldsymbol{k}_3 + \mathrm{O}(a^4). \end{aligned} \tag{2}$$

Here, $V_{123} = V(\boldsymbol{k}_1, \boldsymbol{k}_2, \boldsymbol{k}_3)$ is the interaction vertex, and c.c. denotes complex conjugate. In this expansion, we presume every subsequent term smaller than the previous one, in particular, $\xi_k = |V_{kkk} a_k| k^d / \omega_k \ll 1$. Wave turbulence that satisfies that condition is called weak turbulence. Also, space dimensionality d can be 1, 2, or 3.

A dynamic equation that accounts for pumping, damping, wave propagation, and interaction thus has the following form:

$$\frac{\partial a_k}{\partial t} = -\mathrm{i}\frac{\delta H}{\delta a_k^*} + f_k(t) - \gamma_k a_k. \tag{3}$$

It is likely that the statistics of the weak turbulence at $k \gg k_f$ is close to Gaussian for wide classes of pumping statistics (this has not been shown rigorously). It is definitely the case for a random force with the statistics close to Gaussian. We consider here and below a pumping by a Gaussian random force statistically isotropic and homogeneous in space and white in time. Thus,

$$\langle f_k(t) f_{k'}^*(t') \rangle = F(k)\delta(\boldsymbol{k} + \boldsymbol{k}')\delta(t - t'), \tag{4}$$

where angular brackets imply spatial averages, and $F(k)$ is assumed nonzero only around some k_f. For waves to be well defined, we assume $\gamma_k \ll \omega_k$.

Because the dynamic equation (3) contains a quadratic nonlinearity, the statistical description in terms of moments encounters the closure problem: the time derivative of the second moment is expressed via the third one, the time derivative of the third moment is expressed via the fourth one, and so on. Fortunately, weak turbulence in the inertial interval is expected to have the statistics close to Gaussian, so one can express the fourth moment as the product of two second ones. As a result, one gets a closed kinetic equation for the single-time pair correlation function $\langle a_k a_{k'} \rangle = n_k \delta(\boldsymbol{k} + \boldsymbol{k}')$ (Zakharov et al., 1992):

$$\begin{aligned} &\frac{\partial n_k}{\partial t} = F_k - \gamma_k n_k + I_k^{(3)}, \\ &I_k^{(3)} = \int \big(U_{k12} - U_{1k2} - U_{2k1} \big) \, \mathrm{d}\boldsymbol{k}_1 \, \mathrm{d}\boldsymbol{k}_2 , \\ &U_{123} = \pi \big[n_2 n_3 - n_1 (n_2 + n_3) \big] |V_{123}|^2 \\ &\qquad \times \delta(\boldsymbol{k}_1 - \boldsymbol{k}_2 - \boldsymbol{k}_3) \delta(\omega_1 - \omega_2 - \omega_3). \end{aligned} \tag{5}$$

This is called the kinetic equation for waves. The collision integral $I_k^{(3)}$ describes three-wave interactions: the first term in the integral corresponds to a decay of a given wave while the second and third terms correspond to a confluence with other waves.

One can estimate from (5) the inverse time of nonlinear interaction at a given k as $|V(k,k,k)|^2 n(k) k^d/\omega(k)$. We define k_d as the wave number where this inverse time is comparable with $\gamma(k)$ and assume nonlinearity to dominate over dissipation at $k \ll k_d$. As has been noted, wave turbulence appears when there is a wide (inertial) interval of scales where both pumping and damping are negligible, which requires $k_d \gg k_f$, the condition analogous to $\mathrm{Re} \gg 1$.

The presence of the frequency delta-function in $I_k^{(3)}$ means that wave interaction conserves the quadratic part of the energy $E = \int \omega_k n_k \, \mathrm{d}\boldsymbol{k} = \int E_k \, \mathrm{d}k$. For the cascade picture to be valid, the collision integral has to converge in the inertial interval which means that energy exchange is small between motions of vastly different scales, a property called interaction locality in k-space. Consider now a statistical steady state established under the action of pumping and dissipation. Let us multiply (5) by ω_k and integrate it over either interior or exterior of the ball with radius k. Taking $k_f \ll k \ll k_d$, one sees that the energy flux through any spherical surface (Ω is a solid angle),

$$P_k = \int_0^k k^{\mathrm{d}-1} \mathrm{d}k \int \mathrm{d}\Omega \, \omega_k I_k^{(3)},$$

is constant in the inertial interval and is equal to the energy production/dissipation rate:

$$P_k = \varepsilon = \int \omega_k F_k \, \mathrm{d}\boldsymbol{k} = \int \gamma_k E_k \, \mathrm{d}k. \tag{6}$$

Let us assume now that the medium (characterized by ω_k and V_{123}) can be considered isotropic at the scales in the inertial interval. In addition, for scales much larger or much smaller than a typical scale in the medium (like the Debye radius in plasma or the depth of the water), the Hamiltonian coefficients are usually scale invariant: $\omega(k) = ck^\alpha$ and $|V(\boldsymbol{k}, \boldsymbol{k}_1, \boldsymbol{k}_2)|^2 = V_0^2 k^{2m} \chi(\boldsymbol{k}_1/k, \boldsymbol{k}_2/k)$ with $\chi \simeq 1$. Remember that we presumed statistically isotropic force. In this case, the pair correlation function that describes a steady cascade is also isotropic and scale invariant:

$$n_k \simeq \varepsilon^{1/2} V_0^{-1} k^{-m-d}. \tag{7}$$

One can show that (7) reduces $I_k^{(3)}$ to zero (see Zakharov et al., 1992).

If the dispersion relation $\omega(k)$ does not allow for the resonance condition $\omega(k_1) + \omega(k_2) = \omega(|\boldsymbol{k}_1 + \boldsymbol{k}_2|)$, then the three-wave collision integral is zero and one has to account for four-wave scattering which is always resonant; that is, whatever $\omega(k)$ one can always find four wave vectors that satisfy $\omega(k_1) + \omega(k_2) = \omega(k_3) + \omega(k_4)$ and $\boldsymbol{k}_1 + \boldsymbol{k}_2 = \boldsymbol{k}_3 + \boldsymbol{k}_4$. The collision integral that describes scattering,

$$\begin{aligned} I_k^{(4)} = \frac{\pi}{2} \int & |T_{k123}|^2 \big[n_2 n_3 (n_1 + n_k) \\ & - n_1 n_k (n_2 + n_3) \big] \delta(\boldsymbol{k} + \boldsymbol{k}_1 - \boldsymbol{k}_2 - \boldsymbol{k}_3) \\ & \times \delta(\omega_k + \omega_1 - \omega_2 - \omega_2) \, \mathrm{d}\boldsymbol{k}_1 \, \mathrm{d}\boldsymbol{k}_2 \, \mathrm{d}\boldsymbol{k}_3, \end{aligned} \tag{8}$$

conserves the energy and also the wave action $N = \int n_k \, \mathrm{d}\boldsymbol{k}$ (which can also be called the number of waves). Pumping generally provides for an input of both E and N. If there are two inertial intervals (at $k \gg k_f$ and $k \ll k_f$), then there should be two cascades. Indeed, if $\omega(k)$ grows with k, then absorbing finite amount of E at $k_d \to \infty$ corresponds to an absorption of an infinitely small N. It is thus clear that the flux of N has to go in the opposite direction, that is, to small wave numbers. A so-called inverse cascade with a constant flux of N can thus be realized at $k \ll k_f$. A sink at small k can be provided by wall friction in the container or by long waves leaving the turbulent region in open spaces (as in sea storms).

The collision integral $I_k^{(3)}$ involves products of two n_k, so that flux constancy requires $E_k \propto \varepsilon^{1/2}$ while for the four-wave case, one has $E_k \propto \varepsilon^{1/3}$. In many cases (when there is complete self-similarity), that knowledge is sufficient to obtain the scaling of E_k from a dimensional reasoning without actually calculating V and T. For example, short waves in deep water are characterized by the surface tension σ and density ρ, so the dispersion relation must be $\omega_k \sim \sqrt{\sigma k^3/\rho}$, which allows for the three-wave resonance and thus $E_k \sim \varepsilon^{1/2}(\rho\sigma)^{1/4} k^{-7/4}$. For long waves in deep water, the surface-restoring force is dominated by gravity, so that the gravitational acceleration g replaces σ as a defining parameter and $\omega_k \sim \sqrt{gk}$. Such a dispersion law does not allow for three-wave resonance, so that the dominant interaction is four-wave scattering which permits two cascades. The direct energy cascade corresponds to $E_k \sim \varepsilon^{1/3} \rho^{2/3} g^{1/2} k^{-5/2}$. The inverse cascade carries the flux of N which we denote Q; it has the dimensionality $[Q] = [\varepsilon]/[\omega_k]$ and corresponds to $E_k \sim Q^{1/3} \rho^{2/3} g^{2/3} k^{-7/3}$.

Because the statistics of weak turbulence is near Gaussian, it is completely determined by the pair correlation function, which is in turn determined by the respective flux. We thus conclude that weak turbulence is universal in the inertial interval.

Strong Wave Turbulence

One cannot treat wave turbulence as a set of weakly interacting waves when the wave amplitudes are large ($\xi_k \geq 1$) and also in the particular case of linear (acoustic) dispersion where $\omega(k) = ck$ for arbitrarily small amplitudes. Indeed, there is no dispersion of wave velocity for acoustic waves, so waves moving in the same direction interact strongly and produce shock waves when viscosity is small. Formally, there is a singularity due to the coinciding arguments of delta-functions in (5) (and in the higher terms of perturbation expansion for $\partial n_k/\partial t$), which is thus invalid at however small amplitudes. Still, some features of the statistics of acoustic turbulence can be understood even without a closed description.

Consider a one-dimensional case which pertains, for instance, to sound propagating in long pipes. Because weak shocks are stable with respect to transverse perturbations (Landau & Lifshitz, 1987), quasi-one-dimensional perturbations may propagate in two and three dimensions as well. In a reference frame that moves with the sound velocity, weakly compressible 1-d flows ($u \ll c$) are described by the Burgers equation (Landau & Lifshitz, 1987)

$$u_t + uu_x - \nu u_{xx} = 0. \qquad (9)$$

The Burgers equation has a propagating shock-wave solution $u = 2v\{1 + \exp[v(x - vt)/\nu]\}^{-1}$ with the energy dissipation rate $\nu \int u_x^2 \,\mathrm{d}x$ independent of ν. The shock width ν/v is a dissipative scale, and we consider acoustic turbulence produced by a pumping correlated on much larger scales (i.e., pumping a pipe from one end by frequencies much less than cv/ν). After some time, the system will develop shocks at random positions. Here we consider the single-time statistics of the Galilean invariant velocity difference $\delta u(x, t) = u(x, t) - u(0, t)$. The moments of δu are called structure functions $S_n(x, t) = \langle [u(x, t) - u(0, t)]^n \rangle$. Quadratic nonlinearity allows the time derivative of the second moment to be expressed via the third one:

$$\frac{\partial S_2}{\partial t} = -\frac{\partial S_3}{3\partial x} - 4\varepsilon + \nu \frac{\partial^2 S_2}{\partial x^2}. \qquad (10)$$

Here $\varepsilon = \nu\langle u_x^2\rangle$ is the mean energy dissipation rate. Equation (10) describes both a free decay (then ε depends on t) and the case of a permanently acting pumping which generates turbulence statistically steady at scales less than the pumping length.

In the first case, $\partial S_2/\partial t \simeq S_2 u/L \ll \varepsilon \simeq u^3/L$ (where L is a typical distance between shocks); while in the second case, $\partial S_2/\partial t = 0$ so that $S_3 = 12\varepsilon x + \nu \partial S_2/\partial x$. Consider now the limit $\nu \to 0$ at fixed x (and t for decaying turbulence). Shock dissipation provides for a finite limit of ε at $\nu \to 0$, then

$$S_3 = -12\varepsilon x. \qquad (11)$$

This formula is a direct analog of (6). Indeed, the Fourier transform of (10) describes the energy density $E_k = \langle |u_k|^2 \rangle /2$: $(\partial_t - \nu k^2)E_k = -\partial P_k/\partial k$ where the k-space flux

$$P_k = \int_0^k \mathrm{d}k' \int_{-\infty}^{\infty} \mathrm{d}x\, S_3(x) k' \sin(k'x)/24.$$

It is thus the flux constancy that fixes $S_3(x)$ which is universal (determined solely by ε) and depends neither on the initial statistics for decay nor on the pumping for steady turbulence. On the contrary, other structure functions $S_n(x)$ are not given by $(\varepsilon x)^{n/3}$. Indeed, the scaling of the structure functions can be readily understood for any dilute set of shocks (that is, when shocks do not cluster in space) which seems to be the case for both smooth initial conditions and large-scale pumping in Burgers turbulence. In this case, $S_n(x) \sim C_n|x|^n + C'_n|x|$, where the first term comes from the regular (smooth) parts of the velocity while the second comes from $O(x)$ probability to have a shock in the interval x. The scaling exponents, $\xi_n = \mathrm{d}\ln S_n/\mathrm{d}\ln x$, thus behave as follows: $\xi_n = n$ for $n \le 1$ and $\xi_n = 1$ for $n > 1$. That means that the probability density function (PDF) of the velocity difference in the inertial interval $P(\delta u, x)$ is not scale-invariant; that is, the function of the rescaled velocity difference $\delta u/x^a$ cannot be made scale-independent for any a. As one goes to smaller scales, the lower-order moments decrease faster than the higher-order ones, that means that the smaller the scale the more probable are large fluctuations. In other words, the level of fluctuations increases with the resolution. When the scaling exponents ξ_n do not lie on a straight line, this is called an anomalous scaling since it is related again to the symmetry (scale invariance) of the PDF broken by pumping and not restored even when $x/L \to 0$. As an alternative to the description in terms of structures (shocks), one can relate the anomalous scaling in Burgers turbulence to the additional integrals of motion. Indeed, the integrals $E_n = \int u^{2n}\,\mathrm{d}x/2$ are all conserved by the inviscid Burgers equation. Any shock dissipates the finite amount of E_n at the limit $\nu \to 0$, so that similar to (11), one denotes $\langle \dot{E}_n \rangle = \varepsilon_n$ and obtains $S_{2n+1} = -4(2n+1)\varepsilon_n x/(2n-1)$ for integer n.

Note that $S_2(x) \propto |x|$ corresponds to $E(k) \propto k^{-2}$, which is natural since every shock gives $u_k \propto 1/k$ at $k \ll v/\nu$; that is, the energy spectrum is determined by the type of structures (shocks) rather than by energy flux constancy. Similar ideas were suggested for other types of strong wave turbulence assuming them to be dominated by different structures. Weak wave turbulence, being a set of weakly interacting plane waves, can be studied uniformly for different systems (Zakharov et al., 1992). On the contrary, when nonlinearity is comparable with or exceeds dispersion, different structures appear in different systems. Identifying structures and the role they play in determining different statistical characteristics of strong wave turbulence remains to be investigated for most cases. Broadly, one distinguishes conservative structures (like solitons and vortices) from dissipative structures which usually appear as a result of finite-time singularity of the nondissipative equations (like shocks, light self-focusing, or wave collapse). For example, nonlinear wave packets are described by the nonlinear Schrödinger equation,

$$\mathrm{i}\Psi_t + \Delta\Psi + T|\Psi|^2\Psi = 0. \qquad (12)$$

Weak wave turbulence is determined by $|T|^2$ and is the same for both $T < 0$ (wave repulsion) and $T > 0$ (wave attraction). At high levels of nonlinearity, different signs of T correspond to dramatically different physics: At $T < 0$, one has a stable condensate, solitons, and vortices, while at $T > 0$, instabilities dominate and wave collapse is possible at $d = 2, 3$. No analytic theory is yet available for strong turbulence described by (12).

Because the parameter of nonlinearity $\xi(k)$ generally depends on k, then there may exist a weakly turbulent cascade until some k_* where $\xi(k_*) \sim 1$, and strong turbulence beyond this, wave number; thus weak and strong turbulence can coexist in the same system. Presuming that some mechanism (for instance, wave breaking) prevents the appearance of wave amplitudes that correspond to $\xi_k \gg 1$, one may hypothetize that some cases of strong turbulence correspond to the balance between dispersion and nonlinearity local in k-space so that $\xi(k)$ is constant throughout its domain in k-space. That would correspond to the spectrum $E_k \sim \omega_k^3 k^{-d}/|V_{kkk}|^2$ which is ultimately universal, that is, independent even of the flux (only the boundary k_* depends on the flux). For gravity waves, this gives $E_k = \rho g k^{-3}$, the same spectrum one obtains presuming the wave profile to have cusps (another type of dissipative structure leading to whitecaps in stormy seas—see Phillips, 1977). It is unclear if such flux-independent spectra are realized.

Incompressible Turbulence

Incompressible fluid flow is described by the Navier–Stokes equation

$$\partial_t \boldsymbol{v}(\boldsymbol{r},t) + \boldsymbol{v}(\boldsymbol{r},t)\cdot\nabla\boldsymbol{v}(\boldsymbol{r},t) - \nu\nabla^2\boldsymbol{v}(\boldsymbol{r},t)$$
$$= -\nabla p(\boldsymbol{r},t), \quad \operatorname{div}\boldsymbol{v} = 0.$$

We are again interested in the structure functions $S_n(\boldsymbol{r},t) = \langle [(\boldsymbol{v}(\boldsymbol{r},t) - \boldsymbol{v}(0,t))\cdot\boldsymbol{r}/r]^n\rangle$ and treat first the three-dimensional case. Similar to (10), one considers distance r smaller than the force correlation scale for a steady case and smaller than the size of the turbulent region for a decay case. For such r, one can derive the Karman–Howarth relation between S_2 and S_3 (see Landau & Lifshitz, 1987):

$$\frac{\partial S_2}{\partial t} = -\frac{1}{3r^4}\frac{\partial}{\partial r}(r^4 S_3) + \frac{4\varepsilon}{3} + \frac{2\nu}{r^4}\frac{\partial}{\partial r}\left(r^4\frac{\partial S_2}{\partial r}\right). \quad (13)$$

Here $\varepsilon = \nu\langle(\nabla\boldsymbol{v})^2\rangle$ is the mean energy dissipation rate. Neglecting the time derivative (which is zero in a steady state and small compared with ε for decaying turbulence), one can multiply (13) by r^4 and integrate $S_3(r) = -4\varepsilon r/5 + 6\nu\,\mathrm{d}S_2(r)/\mathrm{d}r$. Andrei Kolmogorov in 1941 considered the limit $\nu \to 0$ for fixed r and *assumed* nonzero limit for ε, which gives the so-called $\frac{4}{5}$ law (see Landau & Lifshitz, 1987; Frisch, 1995):

$$S_3 = -\tfrac{4}{5}\,\varepsilon\, r. \qquad (14)$$

This relation is a direct analog of (6) and (11). It also means that the kinetic energy has a constant flux in the inertial interval of scales (the viscous scale η is defined by $\nu S_2(\eta) \simeq \varepsilon\eta^2$). Law (14) implies that the third-order moment is universal; that is, it does not depend on the details of the turbulence production but is determined solely by the mean energy dissipation rate. The rest of the structure functions have not yet been derived. Kolmogorov (and also Werner Heisenberg, Karl von Weizsacker, and Lars Onsager) presumed the pair correlation function to be determined only by ε and r which would give $S_2(r) \sim (\varepsilon r)^{2/3}$ and the energy spectrum $E_k \sim \varepsilon^{2/3}k^{-5/3}$. Experiments suggest that $\zeta_n = \mathrm{d}\ln S_n/\mathrm{d}\ln r$ lie on a smooth concave curve sketched in Figure 1. While ζ_2 is close to $2/3$, it has to be a bit larger because experiments show that the slope at zero $\mathrm{d}\zeta_n/\mathrm{d}n$ is larger than $\frac{1}{3}$ while $\zeta(3) = 1$ in agreement with (14). As in Burgers turbulence, the PDF of velocity differences in the inertial interval is not scale-invariant in 3-d incompressible turbulence. No one has yet found an explicit relation between the anomalous scaling for 3-d Navier–Stokes turbulence and either structures or additional integrals of motion.

While not exact, the Kolomogorov approximation $S_2(\eta) \simeq (\varepsilon\eta)^{2/3}$ can be used to estimate the viscous scale: $\eta \simeq L\mathrm{Re}^{-3/4}$. The number of degrees of freedom involved in 3-d incompressible turbulence can thus be roughly estimated as $N \sim (L/\eta)^3 \sim \mathrm{Re}^{9/4}$. That means, in particular, that detailed computer simulation of water or oil pipe flows ($\mathrm{Re} \sim 10^4 - 10^7$) or turbulent clouds ($\mathrm{Re} \sim 10^6 - 10^9$) is out of question for the foreseeable future. To calculate correctly at least the large-scale part of the flow, it is desirable to have some theoretical model to parametrize the small-scale motions, the main obstacle being our lack of qualitative understanding and quantitative description of how turbulence statistics changes as one goes downscale.

Large-scale motions in a shallow fluid can be approximately considered two dimensional. When the velocities of such motions are much smaller than the velocities of the surface waves and the velocity of sound, such flows can be considered incompressible. Their description is important for understanding atmospheric and oceanic turbulence at the scales larger than atmosphere height and ocean depth.

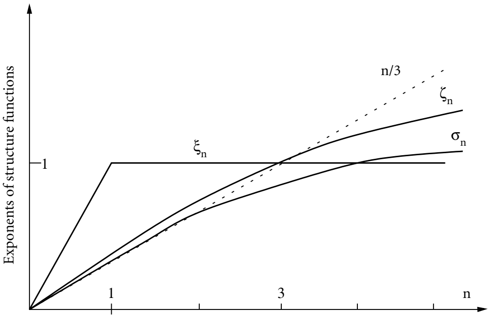

Figure 1. The scaling, exponents of the structure functions ξ_n for Burgers, ζ_n for Navier–Stokes, and σ_n for the passive scalar. The dotted straight line is the Kolmogorov hypothesis $n/3$.

Vorticity $\omega = \text{curl}\,\boldsymbol{v}$ is a scalar in a two-dimensional flow. It is advected by the velocity field and dissipated by viscosity. Taking the curl of the Navier–Stokes equation, one gets

$$\partial_t \omega + (\boldsymbol{v} \cdot \nabla)\omega = \nu \nabla^2 \omega. \qquad (15)$$

Two-dimensional incompressible inviscid flow just transports vorticity from place to place and thus conserves spatial averages of any function of vorticity. In particular, we now have the second quadratic inviscid invariant (in addition to energy) which is called enstrophy: $\int \omega^2 \, \mathrm{d}\boldsymbol{r}$. Since the spectral density of the energy is $|\boldsymbol{v}_k|^2/2$ while that of the enstrophy is $|\boldsymbol{k} \times \boldsymbol{v}_k|^2$, Robert Kraichnan suggested in 1967 that the direct cascade (towards large k) is that of enstrophy while the inverse cascade is that of energy. Again, for the inverse energy cascade, there is no consistent theory except for the flux relation that can be derived similar to (14):

$$S_3(r) = 4\varepsilon r/3. \qquad (16)$$

The inverse cascade is observed in the atmosphere (at scales of 30–500 km) and in laboratory experiments. Experimental data suggest that there is no anomalous scaling; thus, $S_n \propto r^{n/3}$. In particular, $S_2 \propto r^{2/3}$ which corresponds to $E_k \propto k^{-5/3}$. It is ironic that probably the most widely known statement on turbulence, the $\frac{5}{3}$ spectrum suggested by Kolmogorov for the 3-d case, is not correct in this case (even though the true scaling is close), while it is probably exact in Kraichnan's inverse 2-d cascade. Qualitatively, it is likely that the absence of anomalous scaling in the inverse cascade is associated with the growth of the typical turnover time (estimated, say, as $r/\sqrt{S_2}$) with the scale. As the inverse cascade proceeds, the fluctuations have enough time to get smoothed out as opposed to the direct cascade in three dimensions, where the turnover time decreases in the direction of the cascade.

Before discussing the direct (enstrophy) cascade, we describe a similar yet somewhat simpler problem of passive scalar turbulence, which allows one to introduce the necessary notions of Lagrangian description of the fluid flow. Consider a scalar quantity $\theta(\boldsymbol{r}, t)$ that is subject to molecular diffusion and advection by the fluid flow but has no back influence on the velocity (i.e., is passive):

$$\partial_t \theta + (\boldsymbol{v} \cdot \nabla)\theta = \kappa \nabla^2 \theta + \varphi. \qquad (17)$$

Here κ is molecular diffusivity. In the same 2-d flow, ω and θ behave in the same way, but vorticity is related to velocity while the passive scalar is not. Examples of passive scalar are smoke in air, salinity in water, and temperature when one neglects thermal convection. If the source φ produces fluctuations of θ on some scale L, then the inhomogeneous velocity field stretches, contracts, and folds the field θ producing progressively smaller and smaller scales. If the rms velocity gradient is Λ, then molecular diffusion is substantial at scales less than the diffusion scale $r_d = \sqrt{\kappa/\Lambda}$. The ratio

$$\mathrm{Pe} = L/r_d$$

is called the Péclet number. It is an analog of the Reynolds number for passive scalar turbulence. When $\mathrm{Pe} \gg 1$, there is a long inertial interval where the flux constancy relation derived by A.M. Yaglom in 1949 holds,

$$\langle (\boldsymbol{v}_1 \cdot \nabla_1 + \boldsymbol{v}_2 \cdot \nabla_2)\theta_1\theta_2 \rangle = 2P, \qquad (18)$$

where $P = \kappa \, \langle \, (\nabla\theta)^2 \, \rangle$ and subscripts denote the spatial points. In considering the passive scalar problem, the velocity statistics is presumed to be given. Still, the correlation function (18) mixes $\boldsymbol{v}$ and θ and does not generally allow one to make a statement on any correlation function of θ. The proper way to describe the correlation functions of the scalar at scales much larger than the diffusion scale is to employ the Lagrangian description, that is, to follow fluid trajectories. Indeed, if we neglect diffusion, then Equation (17) can be solved along the characteristics $\boldsymbol{R}(t)$ which are called Lagrangian trajectories and satisfy $\mathrm{d}\boldsymbol{R}/\mathrm{d}t = \boldsymbol{v}(\boldsymbol{R}, t)$. Presuming zero initial conditions at $t \to -\infty$, we write

$$\theta\Big(\boldsymbol{R}(t), t\Big) = \int_{-\infty}^{t} \varphi\Big(\boldsymbol{R}(t'), t'\Big)\,\mathrm{d}t'. \tag{19}$$

In that way, the correlation functions of the scalar $F_n = \langle\, \theta(\boldsymbol{r}_1, t) \ldots \theta(\boldsymbol{r}_n, t)\,\rangle$ can be obtained by integrating the correlation functions of the pumping along the trajectories that satisfy the final conditions $\boldsymbol{R}_i(t) = \boldsymbol{r}_i$.

Consider first, the case of pumping which is Gaussian, statistically homogeneous, and isotropic in space and white in time: $\langle\, \varphi(\boldsymbol{r}_1, t_1)\varphi(\boldsymbol{r}_2, t_2)\,\rangle = \Phi(|\boldsymbol{r}_1 - \boldsymbol{r}_2|)$ $\delta(t_1 - t_2)$ where the function Φ is constant at $r \ll L$ and goes to zero at $r \gg L$. The pumping provides for symmetry $\theta \to -\theta$ which makes only even correlation functions F_{2n} nonzero. The pair correlation function is

$$F_2(r, t) = \int_{-\infty}^{t} \Phi\Big(R_{12}(t')\Big)\,\mathrm{d}t'. \tag{20}$$

Here $R_{12}(t') = |\boldsymbol{R}_1(t') - \boldsymbol{R}_2(t')|$ is the distance between two trajectories and $R_{12}(t) = r$. The function Φ essentially restricts the integration to the time interval when the distance $R_{12}(t') \le L$. Simply speaking, the stationary pair correlation function of a tracer is $\Phi(0)$ (which is twice the injection rate of θ^2) times the average time $T_2(r, L)$ that two fluid particles spend within the correlation scale of the pumping. The larger r, the less time it takes for the particles to separate from r to L and the smaller is $F_2(r)$. Of course, $T_{12}(r, L)$ depends on the properties of the velocity field. A general theory is available only when the velocity field is spatially smooth at the scale of scalar pumping L. This so-called Batchelor regime happens, in particular, when the scalar cascade occurs at the scales less than the viscous scale of fluid turbulence. This requires the Schmidt number ν/κ (called the Prandtl number when θ is temperature) to be large, which is the case for very viscous liquids. In this case, one can approximate the velocity difference $\boldsymbol{v}(\boldsymbol{R}_1, t) - \boldsymbol{v}(\boldsymbol{R}_2, t) \approx \hat{\sigma}(t)\boldsymbol{R}_{12}(t)$ with the Lagrangian strain matrix $\sigma_{ij}(t) = \nabla_j v_i$. In this regime, the distance obeys the linear differential equation

$$\dot{\boldsymbol{R}}_{12}(t) = \hat{\sigma}(t)\boldsymbol{R}_{12}(t). \tag{21}$$

The theory of such equations is well developed and related to what is called Lagrangian chaos, as fluid trajectories separate exponentially as is typical for systems with dynamical chaos (see, e.g., Falkovich et al., 2001): At t much larger than the correlation time of the random process $\hat{\sigma}(t)$, all moments of R_{12} grow exponentially with time and $\langle\, \ln[R_{12}(t)R_{12}(0)]\,\rangle = \lambda t$, where λ is called a senior Lyapunov exponent of the flow (note that for the description of the scalar we need the flow taken backwards in time which is different from that taken forward because turbulence is irreversible). Dimensionally, $\lambda = \Lambda f(\mathrm{Re})$ where the limit of the function f at $\mathrm{Re} \to \infty$ is unknown. We thus obtain

$$F_2(r) = \Phi(0)\lambda^{-1}\ln(L/r) = 2P\lambda^{-1}\ln(L/r). \tag{22}$$

In a similar way, one shows that for $n \ll \ln(L/r)$, all F_n are expressed via F_2 and the structure functions $S_{2n} = \langle\, [\theta(\boldsymbol{r}, t) - \theta(0, t)]^{2n}\,\rangle \propto \ln^n(r/r_d)$ for $n \ll \ln(r/r_d)$. This can be generalized for an arbitrary statistics of pumping as long as it is finite-correlated in time (Falkovich et al., 2001).

One can use the analogy between passive scalar and vorticity in two dimensions as has been shown by Falkovich and Lebedev in 1994 following the line suggested by Kraichnan in 1967. For the enstrophy cascade, one derives the flux relation analogous to (18):

$$\langle(\boldsymbol{v}_1 \cdot \nabla_1 + \boldsymbol{v}_2 \cdot \nabla_2)\omega_1\omega_2\rangle = 2D, \tag{23}$$

where $D = \langle\, \nu(\nabla\omega)^2\,\rangle$. The flux relation along with $\omega = \operatorname{curl} \boldsymbol{v}$ suggests the scaling $\delta v(r) \propto r$, that is, velocity being close to spatially smooth (of course, it cannot be perfectly smooth to provide for a nonzero vorticity dissipation in the inviscid limit, but the possible singularitites are indeed shown to be no stronger than logarithmic). That makes the vorticity cascade similar to the Batchelor regime of passive scalar cascade with a notable change in that the rate of stretching λ acting on a given scale is not a constant but is logarithmically growing when the scale decreases. Since λ scales as vorticity, the law of renormalization can be established from dimensional reasoning, and one gets $\langle\omega(\boldsymbol{r}, t)\omega(0, t)\,\rangle \sim [D\ln(L/r)]^{2/3}$ which corresponds to the energy spectrum $E_k \propto D^{2/3}k^{-3}\ln^{-1/3}(kL)$. Higher-order correlation functions of vorticity are also logarithmic, for instance, $\langle\, \omega^n(\boldsymbol{r}, t)\omega^n(0, t)\,\rangle \sim [D\ln(L/r)]^{2n/3}$. Note that both passive scalar in the Batchelor regime and vorticity cascade in two dimensions are universal, that is, determined by the single flux (P and D, respectively) despite the existence of higher-order conserved quantities. Experimental data and numeric simulations support these conclusions.

Zero Modes and Anomalous Scaling

Let us now return to the Lagrangian description and discuss it when velocity is not spatially smooth, for example, that of the energy cascades in the inertial interval. One can assume that it is Lagrangian statistics that are determined by the energy flux when the distances between fluid trajectories are in the inertial interval. That assumption leads, in particular, to the Richardson law for the asymptotic growth of the interparticle distance:

$$\langle R_{12}^2(t)\rangle \sim \varepsilon t^3, \tag{24}$$

which was first established from atmospheric observations (in 1926) and later confirmed experimentally for energy cascades both in 3-d and in 2-d. There is no consistent theoretical derivation of (24), and it is unclear whether it is exact (likely to be in 2-d) or just approximate (possible in 3-d). The semi-heuristic argument usually presented in textbooks is based on the mean-field estimate: $\dot{\boldsymbol{R}}_{12} = \delta\boldsymbol{v}(\boldsymbol{R}_{12}, t) \sim (\varepsilon R_{12})^{1/3}$, which upon integration gives $R_{12}^{2/3}(t) - R_{12}^{2/3}(0) \sim \varepsilon^{1/3} t$. For the passive scalar it gives, by virtue of (20), $F_2(r) \sim \Phi(0)\varepsilon^{-1/3}[L^{2/3} - r^{2/3}]$ which was suggested by S. Corrsin and A.M. Oboukhov. The structure function is then $S_2(r) \sim \Phi(0)\varepsilon^{-1/3} r^{2/3}$. Experiments measuring the scaling exponents $\sigma_n = \mathrm{d}\ln S_n(r)/\mathrm{d}\ln r$ generally give σ_2 close to $2/3$ but higher exponents deviating from the straight line are even stronger than the exponents of the velocity in 3-d. Moreover, the scalar exponents σ_n are anomalous even when advecting velocity has a normal scaling like in the 2-d energy cascade.

To better understand the Lagrangian dynamics (and passive scalar statistics) in a spatially nonsmooth velocity, Kraichnan suggested considering the model of a velocity field as having the simplest statistical and temporal properties, namely Gaussian velocity which is white in time:

$$\begin{aligned} \langle v^i(\boldsymbol{r}, t) v^j(0, 0)\rangle &= \delta(t)\Big[D_0\delta_{ij} - d_{ij}(\boldsymbol{r})\Big], \\ d_{ij} &= D_1 r^{2-\gamma}\Big[(d + 1 - \gamma)\,\delta^{ij} + (\gamma - 2) r^i r^j r^{-2}\Big]. \end{aligned} \tag{25}$$

Here the exponent $\gamma \in [0, 2]$ is a measure of the velocity nonsmoothness with $\gamma = 0$ corresponding to a smooth velocity and $\gamma = 2$ corresponding to a velocity very rough in space (distributional). Richardson–Kolmogorov scaling of the energy cascade corresponds to $\gamma = 2/3$. Lagrangian flow is a Markov random process for the Kraichnan ensemble (25). Every fluid particle undergoes a Brownian random walk with the so-called eddy diffusivity D_0. The PDF for two particles to be separated by r after time t satisfies the diffusion equation (see, e.g., Falkovich et al., 2001)

$$\begin{aligned} \partial_t P(r, t) &= L_2 P(r, t), \\ L_2 = d_{ij}(\boldsymbol{r})\nabla^i\nabla^j &= D_1(d - 1) r^{1-d}\partial_r r^{d+1-\gamma}\partial_r, \end{aligned} \tag{26}$$

with the scale-dependent diffusivity $D_1(d - 1)r^{2-\gamma}$. The asymptotic solution of (26) is lognormal for the Batchelor case while for $\gamma > 0$

$$P(r, t) = r^{d-1} t^{d/\gamma} \exp\Big(-\mathrm{const}\, r^\gamma / t\Big). \tag{27}$$

For $\gamma = 2/3$, it reproduces, in particular, the Richardson law. Multiparticle probability distributions also satisfy diffusion equations in the Kraichnan model as well as all the correlation functions of θ. Multiplying equation (17) by $\theta_2 \ldots \theta_{2n}$ and averaging over the Gaussian statistics of $\boldsymbol{v}$ and φ, one derives

$$\begin{aligned} \partial_t F_{2n} &= L_{2n} F_{2n} + \sum_{l,m} F_{2n-2}\Phi(\boldsymbol{r}_{lm}), \\ L_{2n} &= \sum d_{ij}(\boldsymbol{r}_{lm})\nabla_l^i \nabla_m^j. \end{aligned} \tag{28}$$

This equation enables one, in principle, to derive inductively all steady-state F_{2n} starting from F_2. The equation $\partial_t F_2(r, t) = L_2 F_2(r, t) + \Phi(r)$ has a steady solution $F_2(r) = 2[\Phi(0)/\gamma d(d - 1)D_1][dL^\gamma/(d - \gamma) - r^\gamma]$, which has the Corrsin–Oboukhov form for $\gamma = 2/3$. Further, F_4 contains the so-called forced solution having the normal scaling 2γ but also, remarkably, a zero mode Z_4 of the operator L_4: $L_4 Z_4 = 0$. Such zero modes necessarily appear (to satisfy the boundary conditions at $r \simeq L$) for all $n > 1$, and the scaling exponents of Z_{2n} are generally different from $n\gamma$ that is anomalous. In calculating the scalar structure functions, all terms cancel out except a single zero mode (called irreducible because it involves all distances between $2n$ points). Calculations of Z_n and their scaling exponents σ_n were carried out analytically at $\gamma \ll 1$, $2 - \gamma \ll 1$ and $d \gg 1$, and numerically for all γ and $d = 2, 3$ (Falkovich et al., 2001).

That gives σ_n lying on a convex curve (as in Figure 1) which saturates to a constant at large n. Such saturation (confirmed by experiments) is a signature that most singular structures in a scalar field are shocks (as in Burgers turbulence), the value σ_n at $n \to \infty$ is the fractal codimension of fronts in space. Interestingly, the Kraichnan model enables one to establish the relation between the anomalous scaling and conservation laws of a new type. Thus, the combinations of distances between points that constitute zero modes are the statistical integrals of Lagrangian evolution. To give a simple example, in a Brownian walk, the mean distance between every two particles grows with time, $\langle R_{lm}^2(t)\rangle = R_{lm}^2(0) + \kappa t$, while $\langle R_{lm}^2 - R_{pq}^2\rangle$ and $\langle 2(d + 2)R_{lm}^2 R_{pq}^2 - d(R_{lm}^4 + R_{pq}^4)\rangle$ (and an infinity of similarly built harmonic polynomials) are conserved. Note that the integrals are not dynamical, they are conserved only in average. In a turbulent flow, the form of such conserved quantities is more complicated, but the essence is the same: the increase of averaged distances between fluid particles is compensated by the decrease in shape fluctuations. The existence of statistical conserved quantities breaks the scale invariance of scalar statistics in the inertial interval and explains why scalar turbulence knows more about pumping than just the value of the flux. Note that both symmetries, one broken by pumping (scale invariance) and another by damping (time reversibility) are not restored even when $r/L \to 0$ and $r_d/r \to 0$.

For the vector field (like velocity or magnetic field in magnetohydrodynamics), the Lagrangian statistical integrals of motion may involve both the coordinate of the fluid particle and the vector it carries. Such integrals of motion were built explicitly and related to the anomalous scaling for the passively advected magnetic field in the Kraichnan ensemble of velocities (Falkovich et al., 2001). Doing the same for velocity that satisfies the Navier–Stokes equation remains a task for the future.

Gregory Falkovich

See also **Burgers equation; Chaos vs. turbulence; Development of singularities; Intermittency; Kolmogorov cascade; Lagrangian chaos; Magnetohydrodynamics; Mixing; Navier–Stokes equation; Nonlinear Schrödinger equations; Water waves; Wave packets, linear and nonlinear**

Further Reading

Falkovich, G., Gawędzki, K. & Vergassola, M. 2001. Particles and fields in fluid turbulence, *Reviews of Modern Physics*, 73: 913–975

Frisch, U. 1995. *Turbulence: The Legacy of A.N. Kolmogorov*, Cambridge and New York: Cambridge University Press

Landau, L. & Lifshitz, E. 1987. *Fluid Mechanics*, 2nd edition, Oxford and New York: Pergamon Press

Phillips, O. 1977. *The Dynamics of the Upper Ocean*, 2nd edition, Cambridge and New York: Cambridge University Press

Zakharov, V., L'vov, V. & Falkovich, G. 1992. *Kolmogorov Spectra of Turbulence*, Berlin and New York: Springer

TURBULENCE, IDEAL

Ideal turbulence (IT) is a mathematical phenomenon that occurs in certain infinite-dimensional deterministic dynamical systems. The attractor of an IT system lies off the phase space, and among the attractor points there are fractal or even random functions. IT is observed in various idealized models of real distributed systems (electrodynamics, acoustics, radiophysics, etc.), and it helps to understand the mathematical scenarios for features of real turbulence. Cascade processes in IT are capable of giving birth to structures of arbitrarily small scale and even causing stochastization of the systems.

A mathematically rigorous definition of ideal turbulence is based on notions of dynamical systems theory and chaos theory. Spatiotemporal chaotization in dynamical systems on spaces of smooth or piecewise smooth functions is perceived as a cascading evolution of such functions with the result that their behavior becomes more and more intricate (see Figure 1), whereupon the limiting states cannot be described with smooth functions. This implies that the attractor of the dynamical system is not contained entirely in the phase space; thus, the dynamical system needs to be extended on a wider functional space so that this new space contains whole "ω-limit" sets of all or almost all trajectories. (The ω-limit set of a trajectory is defined as the attractor of the trajectory or, more precisely, as the set of limit points of the trajectory.) The spaces of fractal and random functions are particularly appealing for use as a wider space.

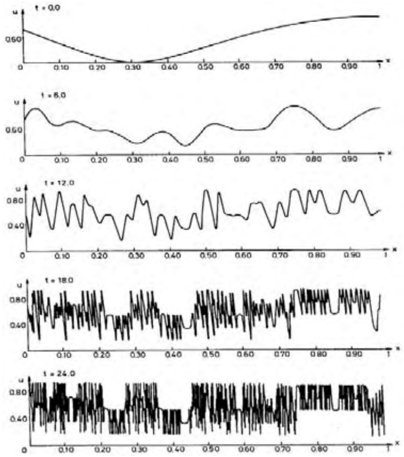

Figure 1. Start of ideal turbulence: Typical instantaneous distributions of current in a lossless transmission line described by the boundary-value problem $i_x = -Cv_t$, $v_x = -Li_t$, and $v(0,t) = 0$, $i(1,t) = G(v(1,t))$, where i and v are the current and voltage along the line, and G specifies the v-i characteristic of an Esaki (tunnel) diode fixing the boundary condition at $x = 1$.

If, with such an extension, the ω-limit set of the trajectory corresponding to some initial state contains a "point" that is a fractal function, then this initial state is said to generate IT. Similarly, if a dynamical system can be extended on a space containing both deterministic and random functions and for some initial state its associated ω-limit set contains a "point" that is a random function, then this initial state is said to generate stochastic ideal turbulence (SIT). If initial states generate IT or SIT, then IT or SIT is said to occur in the dynamical system.

For a space containing fractal functions, one may take the space of multivalued functions with the metric $\rho(\zeta_1, \zeta_2) = \mathrm{dist}_H(\mathrm{gr}\,\zeta_1, \mathrm{gr}\,\zeta_2)$, where $\mathrm{dist}_H(\cdot, \cdot)$ is the Hausdorff distance between sets and $\mathrm{gr}\,\zeta$ denotes the graph of ζ. As any function (deterministic or random) can be interpreted as the collection of all its finite-dimensional distributions, a metric for spaces containing random and deterministic functions is conveniently chosen to compare the distributions of functions (Sharkovsky & Romanenko, 1992).

This classification can be deepened. For instance, an initial state is said to generate weak ideal turbulence (WIT) if it does not generate IT but its associated

ω-limit set contains a function that is multivalued at an infinite number of points.

A simple example of a system with turbulence is the discrete dynamical system acting on the space of smooth functions $\varphi : D \to E$ according to the rule

$$S : \varphi(x) \mapsto f(\varphi(x)), \tag{1}$$

where $f : E \to E$ is a smooth function, and D and E are regions of Euclidean spaces. The trajectory through a point φ is the sequence $f^n(\varphi(x))$, $n = 0, 1, 2, \ldots$, where the superscript n denotes the nth iteration. Thus, the dynamics of the trajectory can be treated as the dynamics of a continuum of uncoupled oscillators. At every point $x \in D$, there is a "pendulum," oscillating under the law $z_n \mapsto z_{n+1} = f(z_n)$ with $z_0 = \varphi(x)$ and independently of the pendula at other points of D. The independence of the oscillators causes IT in the dynamical systems (1), and moreover, when f has the property of sensitive dependence on initial data on some open set $E' \subset E$, those φ such that $\varphi(D) \supset E'$ often generate SIT. In more general situations that occur in applied problems, the oscillation law depends on initial data φ and/or a point $x \in D$; it can also be time-dependent.

A description of long-term properties for the dynamical systems (1) is most advantageous when f is a one-dimensional map, and D and E are intervals, whereupon one has the following result:

> **Theorem.** *There occur (i) weak ideal turbulence, if f has periodic trajectories of periods 2^i, $0 \le i \le l$, with some $l > 1$, and no other periodic trajectories; (ii) ideal turbulence, if f has a periodic trajectory of period $\neq 2^i, i = 0, 1, \ldots$; and (iii) stochastic turbulence, if f possesses an ergodic smooth invariant measure.*

The map $f : z \mapsto 4z(1 - z)$, $z \in [0, 1]$, has an invariant measure with the density $p(z) = 1/\pi\sqrt{z(1 - z)}$, and almost each $\varphi : [0, 1] \to (0, 1)$ generates SIT. Its associated ω-limit set consists of a single point which is the random function with the distribution $F(x, z) = \int_0^z p(z)\,dz = (2/\pi)\arcsin\sqrt{z}$. Thus the attractor of the dynamical system (1) consists of just this one point. For $f = f_\lambda : z \mapsto \lambda z(1 - z)$, $0 < \lambda \le 4$, there exists a set of positive Lebesgue measure $\Lambda \subset (3, 4]$ such that f_λ with $\lambda \in \Lambda$ has an ergodic smooth invariant measure on $[0, 1]$; hence SIT in the dynamical systems (1) is a non-exclusive phenomenon.

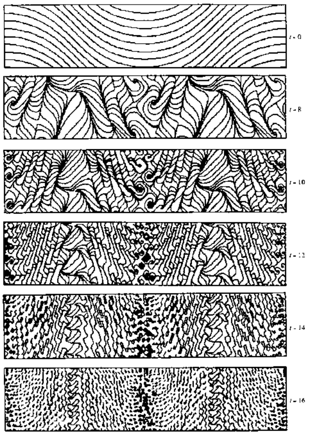

Figure 2. Towards stochastic turbulence: Typical evolution of flow lines for the vector field (w^1, w^2) given by $w_t^1 = w_x^1 + w_y^1$, $w_t^2 = -w_x^2 - w_y^2$, $-\infty < x < +\infty$, $0 \le y \le 1$, and $w^1 = w^2|_{y=0}$, $w^1 = f(w^2)|_{y=1}$ with $f(z) = 1 - 2z^2$. The attractor of the BVP consists of one point—the random vector field $(\hat{w}^1, \hat{w}^2)$ whose components have the same (x, y)-independent distribution density $1/2\pi\sqrt{1 - z^2}$, $-1 < z < 1$.

Many evolutionary boundary-value problems (BVPs) for partial differential equations induce dynamical systems of shifts along solutions. It is natural to say that there arises IT or SIT in a BVP if such turbulence occurs in the corresponding dynamical system. The simplest example is the BVP

$$w_t - w_x = 0, 0 \le x \le 1, t \ge 0, \tag{2}$$

$$w(1, t) = f(w(0, t)), \tag{3}$$

where f is a smooth function from some interval E into itself. On the space of smooth functions $\varphi : [0, 1] \to E$, the BVP induces the dynamical system of shifts $S^t : \varphi(x) \mapsto w_\varphi(x, t)$, $t \ge 0$, where $w_\varphi(x, t)$ is the solution meeting the initial condition $w(x, 0) = \varphi(x)$. For the BVP considered, the shift operator S^t is represented as

$$S^t : \varphi(x) \mapsto f^{\langle x+t \rangle}(\varphi(\{x + t\})) \tag{4}$$

with $\langle \cdot \rangle$ and $\{\cdot\}$ being the integer and fractional part of a number, respectively. Thus the dynamical system (4) is a continuous analog of the dynamical system (1), and one can formulate conditions for turbulence in the BVP according to the above theorem.

Replacing (3) with $w_t(1, t) = g(w(0, t))\, w_t(0, t))$ leads to the dynamical system (4) where f is replaced with $f_{\gamma[\varphi]} = f + \gamma[\varphi]$, f is an antiderivative of g and $\gamma[\varphi] = f(\varphi(0)) - \varphi(1)$. The type of the turbulence arising in a solution $w_\varphi(x, t)$ follows the theorem, as applied to $f_{\gamma[\varphi]}$.

BVPs for the wave equation (and related systems) provide examples of ideal turbulence. For the BVP

$$w_{tt} - w_{xx} = 0, \quad 0 \le x \le 1, \tag{5}$$

$$w(0, t) = 0, \quad w_t(1, t) = h(w_x(1, t)), \tag{6}$$

its associated shift operator S^t is expressed through a one-dimensional map $f: z_n \mapsto z_{n+1}$, defined implicitly by $z_{n+1} - z_n = h\,(z_n + z_{n+1})$. This allows one to find conditions for IT or SIT for particular functions h (Sharkovsky, 1994; Sharkovsky et al., 1995). When the conditions of (6) are replaced with $w(0,t) = 0$, $w_x(1,t) = h(w_x(0,t))$, there arises a two-dimensional map defined by $z_{n+1} - z_{n-1} = h(z_n)$.

There are many other one- and many-dimensional BVPs whose dynamics are described in terms of low-dimensional maps, as in the above examples. In these cases, the theory of maps suggests why and how turbulence occurs in the BVP and presents scenarios for self-structuring and self-stochastization. Of importance here are the following properties of maps: the intricate dynamical structure of the basins of attracting cycles, the local self-similarity of the set of points with unstable trajectories, and the occurrence of a smooth invariant measure. Figure 2 is an example of how processes of self-structuring lead to stochastic turbulence.

A.N. SHARKOVSKY AND E.YU. ROMANENKO

See also **Attractors; Butterfly effect; Chaotic dynamics; Dimensions; Dynamical systems; Ergodic theory; Maps; Measures; Mixing; One-dimensional maps; Phase space; Routes to chaos; Sinai–Ruelle–Bowen measures; Turbulence**

Further Reading

Romanenko, E.Yu. & Sharkovsky, A.N. 1996. From one-dimensional to infinite-dimensional dynamical systems: ideal turbulence. *Ukrainian Mathematical Journal*, 48(12): 1817–1842

Romanenko, E.Yu., Sharkovsky, A.N. & Vereikina, M.B. 1995. Self-structuring and self-similarity in boundary value problems. *International Journal of Bifurcation and Chaos*, 5(5): 145–156

Sharkovsky, A.N. 1994. Ideal turbulence in an idealized time-delayed Chua's circuit. *International Journal of Bifurcation and Chaos*, 4(2): 303–309

Sharkovsky, A.N., Deregel, Ph. & Chua, L.O. 1995. Dry turbulence and period-adding phenomena from a 1-D map. *International Journal of Bifurcation and Chaos*, 5(5): 1283–1302

Sharkovsky, A.N. & Romanenko, E.Yu. 1992. Ideal turbulence: attractors of deterministic systems may lie in the space of random fields. *International Journal of Bifurcation and Chaos*, 2(1): 31–36

TURING PATTERNS

As noted by D'Arcy Wentworth Thompson (1917) people in the early 1900s (when the study of symmetry-breaking instabilities was still in its infancy) had already considered the possibility of generating stationary regular concentration patterns through the interplay of diffusion and chemistry. At mid-century, the British mathematician and computing pioneer, Alan Turing, was the first to formulate necessary conditions for the occurrence of space symmetry breaking in the context of biological morphogenesis (Turing, 1952). Following the emergence of the self-oscillating Belousov–Zhabotinsky reaction (Epstein & Pojman, 1998) in the mid-1960s, Ilya Prigogine and coworkers revived Turing's concept, put it on sound thermodynamic and kinetic grounds, and showed that it could only be sustained in continuously fed reactors at a finite distance from equilibrium (Nicolis & Prigogine, 1977).

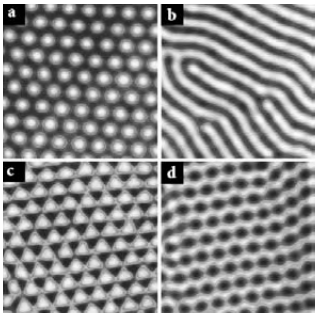

Figure 1. (a)–(d) Turing structures of different symmetries obtained with the chlorite-iodide-malonic acid reaction. Dark and light regions, respectively, correspond to high and low iodide concentration. The wavelength, a function of kinetic parameters and diffusion coefficients, is of the order of 0.2 mm. All patterns are at the same scale: view size 1.7mm × 1.7mm (Courtesy P. De Kepper, CRPP).

This work opened up a whole new field of physical chemistry. Many theoretical studies followed, and the diffusive instability that generates such dissipative structures has popped up in other domains of physics and chemistry (Ball, 1999). However, experiments in the chemical realm lagged behind, and it was only in 1989 that the first experimental evidence was obtained by De Kepper and his group (Castets et al., 1990) using the chlorite-iodide-malonic acid reactive system in so-called gel reactors. A recent detailed status of Turing patterns and other symmetry-breaking instabilities in solution chemistry is presented in Borckmans et al. (2002).

The Turing–Prigogine mechanism consists in the spontaneous instability of a homogeneous mixture of chemically reacting species, when some parameter threshold is crossed as one moves away from equilibrium conditions. It leads to stationary, space-periodic patterns for the concentrations of reactants (see Figure 1). In its minimal form, the description of all the systems that exhibit such diffusive instability can formally be cast in the common language of reaction-diffusion systems governed by the set of equations

$$\frac{\partial \boldsymbol{c}(\boldsymbol{r},t)}{\partial t} = \boldsymbol{f}(\boldsymbol{c},b) + \nabla \cdot D\nabla \boldsymbol{c}(\boldsymbol{r},t), \qquad (1)$$

where $\boldsymbol{c}(\boldsymbol{r}, t) \equiv (\ldots, c_i, \ldots)$ is the local concentration vector, $\boldsymbol{f}(\boldsymbol{c}, b)$ is a vector function representing the reaction kinetics wherein lies the source of nonlinearity, b stands for a set of control parameters, and D is the matrix of diffusive transport coefficients. Appropriate initial and boundary conditions, in relation with the experimental setup are added to complete the mathematical formulation.

To support such symmetry-breaking instability, the chemical kinetics must involve some type of positive feedback loop controlled at least by an activator species that reinforces its own changes, the latter being counterbalanced by an inhibitory process. Spatial structures can form when the inhibitory effects are transported by diffusion over a larger space range than that of the activating mechanism. An intuitive picture may be obtained when a single activator (A) and inhibitor (H) are present. A autocatalytically promotes its own production and that of H, while the latter opposes the production of A. Consider such system in a nonequilibrium homogeneous steady state (hss) and quench it beyond the instability threshold. The hss then becomes very sensitive. A slight local fluctuation of the concentration of A will increase while it also spreads to the surroundings through diffusion. It will also start producing some H that, however, will diffuse away much faster from the point where the fluctuation occurred as $D_{\mathrm{H}} > D_{\mathrm{A}}$. H thereby hinders the propagation of A. A localized peak of activator surrounded by a barrier of H is thus created. In extended systems, such peaks tend to emerge everywhere, randomly distributed, and their interactions lead to the periodic concentration patterns. The beauty of Turing's idea lies in the counterintuitive organization role of diffusive processes when they compete with the proper autocatalytic chemistry, while diffusion still locally strives to erase any concentration of inhomogeneity.

Theoretical work uses nonlinear kinetic models for $\boldsymbol{f}(\boldsymbol{c}, b)$ with a limited number of chemical species, typically two or three (Brusselator, Oregonator, CDIMA, etc.). These models stand as a compromise between a minimum of chemical realism and mathematical tractability. For their part, the experimental kinetic schemes usually involve many species, often not fully determined (Epstein & Pojman, 1998).

Analytical work that relies heavily on bifurcation theory (Nicolis & Prigogine, 1977; Manneville, 1990) allows one to determine, through the solution of amplitude equations, which structures of given symmetry are stable for specific conditions (pattern selection). The calculated bifurcation diagrams help to organize the results obtained by straightforward numerical integration of the reaction-diffusion equations. Both types of information may be used to interpret the experimental results. This pattern selection problem was already on Turing's mind when he stated (Turing, 1952): "Most of an organism, most of the time, is developing from one pattern into another, rather than from homogeneity into a pattern. One would like to be able to follow this more general process mathematically also."

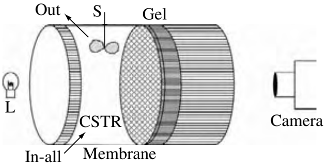

Figure 2. Schematic representation of a disc-shaped one side fed reactor (OSFR): CSTR (continuous stirred tank reactor), Membrane (mineral disc, pore size 0.02 mm) often placed to protect the gel from mechanical stress produced by the stirrer of the CSTR, Gel, In and Out (input and output ports of chemicals), L (light source), CCD camera.

The experimental work takes place in so-called open spatial reactors (Borckmans et al., 2002) which are specifically designed to control the reaction and the structures that eventually develop at a fixed distance from equilibrium and allow probing of the true asymptotic states of the reaction-diffusion systems. Experiments are now usually performed in a one-side fed reactor (OSFR) sketched in Figure 2. The core consists of a piece of soft hydrogel fed by diffusion through one of its faces with chemicals contained in a continuous stirred tank reactor (CSTR), the contents of which are continuously renewed by pumps. The other faces of the gel are pressed against impermeable transparent walls (Plexiglas). Viewing can be practiced both along the feeding axis or orthogonal to it (Ball, 1999; Ouyang & Swinney, 1991). The gel is used to avoid all perturbations induced by the hydrodynamic flows as those associated with the constant supply of fresh reactants, so that only reactive and diffusive processes compete. The necessary diffusion differential between activator and inhibitor species is obtained through the reversible binding of the activator molecules to the large molecular weight color indicator species that is included for visualization purposes. An advantage of such reactors is that they allow for direct correlations to be made between the dynamics of the CSTR, the bifurcation behaviors of which have been extensively studied in the past (Epstein & Pojman, 1998), and that of the gel.

Although scores of papers have been devoted to the application of Turing's idea to biological problems, this speculation remains to be confirmed (Epstein & Pojman, 1998; Borckmans et al., 2002).

Pierre Borckmans and Guy Dewel

See also **Belousov–Zhabotinsky reaction; Brusselator; Morphogenesis, biological; Pattern formation; Reaction-diffusion systems**

Further Reading

Ball, P. 1999. *The Self-made Tapestry*, Oxford and New York: Oxford University Press

Borckmans, P., Dewel, G., De Wit, A., Dulos, E., Boissonade, J., Gauffre, F. & De Kepper, P. 2002. Diffusive instabilities and chemical reactions. *International Journal of Bifurcation and Chaos*, 12: 2307–2332

Castets, V., Dulos, E., Boissonade, J. & De Kepper, P. 1990. Experimental evidence of a sustained standing Turing-type nonequilibrium chemical pattern. *Physical Review Letters*, 64: 2953–2956

Cross, M.C. & Hohenberg, P.C. 1993. Pattern formation outside of equilibrium. *Reviews of Modern Physics*, 65: 851–1112

Epstein, I.R. & Pojman, J.A. 1998. *An Introduction to Nonlinear Chemical Dynamics*, Oxford and New York: Oxford University Press

Manneville, P. 1990. *Dissipative Structures and Weak Turbulence*, New York: Academic Press

Nicolis, G. & Prigogine, I. 1977. *Self-Organization in Nonequilibrium Systems*, New York: Wiley

Ouyang, Q. & Swinney, H.L. 1991. Transition from a uniform state of hexagonal and striped Turing patterns. *Nature*, 352: 610–612

Thompson, D'A.W. 1917. *On Growth and Form*, 2nd edition, Cambridge: Cambridge University Press (2nd edition, 1942); see quotes to S. Leduc

Turing, A. 1952. The chemical basis of morphogenesis *Philosophical Transactions of the Royal Society*, B 237: 37–42

Turing, A. 1992. *Morphogenesis*, Collected works of A. Turing, vol. 3, edited by P.T. Saunders, Amsterdam: Elsevier

http://www.turing.org.uk/, Alan Hodges's pages on Alan Turing

http://data.archives.ecs.soton.ac.uk/turing/, the Turing digital archive

http://www.swintons.net/jonathan/turing.htm, Alan Turing and morphogenesis

TWIST MAP

See **Nontwist maps**

TWISTOR THEORY

Introduced by Roger Penrose as a geometrical framework for the unification of quantum theory and general relativity (gravity), twistor theory brings out the complex (holomorphic) geometry that underlies real space-time. In general relativity, space-time is a four manifold with metric g. When $g = \mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2$, where (t, x, y, z) are coordinates on $\mathbb{R}^4$, g is said to be flat with signature (1,3) and is called Minkowski space.

The first appearance of a complex structure arises from the fact that at a given event, the celestial sphere of light rays (null directions with respect to g) naturally has the structure of the Riemann sphere, $\mathbb{CP}^1$, in such a way that Lorentz transformations (linear transformations of the tangent space preserving the metric) act on this sphere by Möbius transformations. Twistor space extends this idea to the whole of Minkowski space. Denoted $\mathbb{PT}$, the twistor space for Minkowski space is complex projective three space, $\mathbb{CP}^3$, the space of one-dimensional subspaces of $\mathbb{C}^4$; it is a three-dimensional complex manifold obtained by adding a "plane at infinity" to $\mathbb{C}^3$. Physically, points of twistor space correspond to spinning massless particles in Minkowski space. Mathematically, the correspondence can be understood as the Klein correspondence.

The Klein Correspondence

The correspondence between $\mathbb{PT}$ and Minkowski space can be extended first to complexified Minkowski space, so that the coordinates are allowed to take on values in $\mathbb{C}$, and then to its conformal compactification by including some points at infinity. It then coincides with the classical complex Klein correspondence. The Klein correspondence is the one-to-one correspondence between lines in $\mathbb{CP}^3$ and points of a four complex-dimensional quadric, $\mathbb{CM}$, in $\mathbb{CP}^5$. The four-quadric $\mathbb{CM}$ can be understood as conformally compactified complexified Minkowski space. Introducing affine coordinates (λ, z_1, z_2) on $\mathbb{PT}$, we find that a line in $\mathbb{PT}$ corresponds to a point (t, x, y, z) by

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} t - z & x + \mathrm{i}y \\ x - \mathrm{i}y & t + z \end{pmatrix} \begin{pmatrix} 1 \\ \lambda \end{pmatrix}.$$

Alternatively, fixing (λ, z_1, z_2) in these equations gives a two-plane in complex Minkowski space corresponding to all the lines in $\mathbb{PT}$ through (λ, z_1, z_2). Such two-planes are called α-planes. They are totally null (i.e., the tangent vectors not only have zero length but are also mutually orthogonal) and also self-dual (under the differential geometer's notion of Hodge duality).

This complex correspondence can also be restricted to give correspondences for $\mathbb{R}^4$ with metrics of positive definite signature or ultra-hyperbolic (2, 2) signature.

The Penrose Transform

A basic task of twistor theory is to transform solutions to the field equations of mathematical physics into objects on twistor space. This works well for linear massless fields such as the Weyl neutrino equation, Maxwell's equations for electromagnetism, and linearized gravity. In its general form, this transform has become known as the Penrose transform. Such fields correspond to freely prescribable holomorphic functions $f(\lambda, z_1, z_2)$ (or, more precisely, analytic cohomology classes) on regions of twistor space. The field can be obtained from this function by means of a contour integral. The simplest of these integral formulae is

$$\phi(x^a) = \oint f\,(\lambda, t - z + \lambda(x + \mathrm{i}y), x - \mathrm{i}y + \lambda(t + z))\,\mathrm{d}\lambda, \tag{1}$$

and differentiation under the integral sign leads to the fact that ϕ satisfies the wave equation

$$\frac{\partial^2\phi}{\partial t^2} - \frac{\partial^2\phi}{\partial x^2} - \frac{\partial^2\phi}{\partial y^2} - \frac{\partial^2\phi}{\partial z^2} = 0\,.$$

Equation (1) was originally discovered by Bateman (1910).

The Penrose transform has found important applications in representation theory and integral geometry. For a review, the reader is referred to Baston and Eastwood (1989), the relevant survey articles in Bailey & Baston (1990) or Chapter 1 of Mason et al. (1990).

Twistor Theory and Nonlinear Equations

The Penrose transform for the Maxwell equations and linearized gravity turns out to be linearizations of correspondences for nonlinear versions of these equations, the Einstein vacuum equations and the Yang–Mills equations, but only in the case that these fields are anti-self-dual. This is the condition that the curvature two-forms satisfy $F^* = -\mathrm{i}F$ where $*$ denotes the Hodge dual (which, up to a change of sign, has the effect of interchanging electric and magnetic fields); it is a nonlinear generalization of the right-handed circular polarization condition. In Minkowski signature, the i factor in the anti-self-duality condition implies that real fields cannot be anti-self-dual. Thus, these extensions are not sufficient to fulfill twistor theory's aim of incorporating real basic physics in Minkowski space. However, the factor of i is not present in Euclidean and ultrahyperbolic signature, so the anti-self-duality condition *is* consistent with real fields in these signatures, and this is where the main applications of these constructions have been.

The Nonlinear Graviton Construction and Its Generalizations

The first nonlinear twistor construction was due to Penrose (1976) and was inspired by Newman's construction of "heavens" from the infinities of asymptotically flat space-times in general relativity (Newman, 1976).

The nonlinear graviton construction proceeds from the definition of twistors in flat space-time as α-planes in complexified Minkowski space. It is natural to ask which complexified metrics admit a full family of α-surfaces, that is, two surfaces that are totally null and self-dual. The answer is that a full family of α-surfaces exists if and only if the conformally invariant part of the curvature tensor, the Weyl tensor, is anti-self-dual. In this case, twistor space can be defined to be the (necessarily three-dimensional) space of such α-surfaces. A remarkable fact is that the twistor space (together with its complex structure) is sufficient to determine the original space-time, and that the data defining the twistor space is effectively freely prescribable, see Penrose (1976) or Atiyah et al. (1978), for a discussion specialized to Euclidean signature.

There are now large families of extensions, generalizations and reductions of this construction. They are all based on the idea of realizing a space with a given complexified geometric structure as the parameter space of a family of holomorphically embedded submanifolds inside a twistor space. In general, the most useful of these constructions are those in which the "space-time" is obtained as the space of rational curves in a twistor space. This is because the equations that are solved on the corresponding space-time can be thought of as a completely integrable system. See Chapter 13 of Mason & Woodhouse (1996) for a more detailed discussion from this point of view.

The Anti-self-dual Yang–Mills Equations and Its Twistor Correspondence

The anti-self-dual Yang–Mills equations extend Maxwell's equations for electromagnetism in the right circularly polarized case. They are really a family of equations depending on a choice of Lie group G, usually taken to be a group of complex matrices, and Maxwell's equations arise from the case in which $G = U(1)$.

Introduce coordinates x^a, $a = 0, 1, 2, 3$, on $\mathbb{R}^4$ with metric $\mathrm{d}s^2 = \mathrm{d}x^0\,\mathrm{d}x^3 + \mathrm{d}x^1\,\mathrm{d}x^2$. The dependent variables are the components A_a of a connection $D_a = \partial_a - A_a$, where $\partial_a = \partial/\partial x^a$ and $A_a = A_a(x^b) \in \mathrm{Lie}\,G$, the Lie algebra of G. This connection defines a method of differentiating vector valued functions s in some representation of G. The freedom in changing bases for the vector bundle induce the gauge transformations $A_a \to g^{-1}A_a g - g^{-1}\partial_a g,\ g(x) \in G$ on A_a, and two connections that are related by a gauge transformation are deemed to be the same.

The self-dual Yang–Mills equations are the condition

$$[D_0, D_2] = [D_1, D_3] = [D_0, D_3] + [D_1, D_2] = 0\,.$$

They are the compatibility conditions $[D_0 + \lambda D_1, D_2 + \lambda D_3] = 0$ for the linear system of equations

$$(D_0 + \lambda D_1)s = (D_2 + \lambda D_3)s = 0, \qquad (2)$$

where $\lambda \in \mathbf{C}$ and s is an n-component column vector. These last equations form a Lax pair for the system.

The Ward construction (Ward, 1977) provides a one-to-one correspondence between gauge equivalence classes of solutions of the self-dual Yang–Mills equations and holomorphic vector bundles on regions in twistor space. The key point here is that Equation (2) defines parallel propagation along α-planes. To each point Z in twistor space, we can associate the vector space E_Z of solutions to Equation (2) along the corresponding α-plane. These vector spaces vary holomorphically with Z, and that is what one means by a vector bundle $E \to \mathbb{PT}$. A remarkable fact is that the anti-self-dual Yang–Mills field can be reconstructed up to gauge from E, and E is effectively freely prescribable. See Penrose (1984, 1986); Ward & Wells (1990), or Mason & Woodhouse (1996) for a full discussion, and Atiyah (1979) for a discussion in Euclidean signature.

The Connection with Completely Integrable Systems

In effect, the twistor constructions amount to providing a geometric general local solution to the anti-self-duality equations in the sense that the twistor data is (for a local solution) freely prescribable. Thus, they demonstrate complete integrability of the anti-self-duality equations. The reconstruction of a solution on space-time from twistor data can be hard. In the anti-self-dual Yang–Mills case, it involves the solutions of a Riemann–Hilbert problem, and in the case of the anti-self-dual Einstein equations, the construction of a family of rational curves inside a complex manifold. Nevertheless, such constructions are a familiar part of the apparatus of the theory of integrable systems.

In Ward (1985), this connection with integrable systems was developed further, and the anti-self-dual Yang–Mills equations were shown to yield many important integrable systems under symmetry reduction. Ward's list has been extended and now includes many of the most famous examples of integrable systems such as the Painlevé equations, the Korteweg–de Vries equation, the nonlinear Schrödinger equation, the N-wave equations, among others (see Ablowitz & Clarkson (1992) and Mason & Woodhouse (1996) for a review). There are some notable omissions from the list such as the Kadomtsev–Petviashvili and Davey–Stewartson equations (at least if one restricts oneself to finite-dimensional gauge groups), but the list remains impressive.

One can impose symmetries on the twistor constructions for the anti-self-duality equations to obtain a reduced twistor correspondence for solutions to any of these integrable equations (see Mason & Woodhouse (1996) and Chapter 1 of Mason et al. (1995)), so there are many many twistor correspondences.

Applications

Twistor constructions have the effect of reducing problems in nonlinear differential equations to problems in complex holomorphic geometry, where there are many powerful tools. Twistor theory underlies the many appearances of algebraic geometry, loop groups, and Riemann–Hilbert problems in the theory of integrable systems, even though these structures were, for systems in one and two dimensions, usually discovered without knowledge of the twistor theory, see Mason & Woodhouse (1996). The most impressive applications of twistor theory have been in three and four dimensions, where it is difficult to imagine making such significant progress without the twistor theory.

The Ward construction was used by Atiyah, Drinfeld, Hitchin, and Manin to construct the Yang–Mills instantons on S^4; see Atiyah (1979). Its symmetry reduction was also used by both to obtain monopoles on $\mathbb{R}^3$ and to study the hyperkähler metric on their moduli spaces, see Ward & Wells (1990) and Atiyah and Hitchin (1988) (see Mason & Woodhouse (1996) for further applications).

The nonlinear graviton construction and its generalizations have been used for many constructions of Einstein manifolds and more general anti-self-dual manifolds (see Hitchin (1979) for the construction of asymptotically locally Euclidean hyperkähler spaces in four dimensions). The twistor constructions have also been an important tool in studying general properties, for a twistor construction of an anti-self-dual conformal structure on any manifold that is a connected sum of two other such manifolds (Donaldson & Friedman, 1989). Further applications and developments can be found in Mason et al. (2001).

An application that goes beyond complete integrability is the twistor framework of Merkulov for studying arbitrary geometric structures. This has led to the remarkable classification of all possible irreducible holonomies of torsion-free affine connections (Merkulov & Schwachhöfer, 1999).

It is to be hoped that these many applications will one day feed back into Penrose's original program and provide a unification between quantum theory and gravity.

LIONEL MASON

See also **Einstein equations; General relativity; Instantons; Integrability; Integral transforms; Inverse scattering method or transform; Riemann–Hilbert problem; Yang–Mills theory**

Further Reading

Ablowitz, M.J. & Clarkson, P.A. 1992. *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, Cambridge and New York: Cambridge University Press

Atiyah, M.F. 1979. *Geometry of Yang-Mills Fields*, Pisa: Accademia Nazionale dei Lincei Scuola Normale Superiore

Atiyah, M.F. & Hitchin, N.J. 1988. *The Geometry and Dynamics of Monopoles*, Princeton, NJ: Princeton University Press

Atiyah, M.F., Hitchin, N.J. & Singer, I.M. 1978. Self-duality in four-dimensional Riemanninian geometry. *Proceedings of the Royal Society* A, 362: 425

Bailey, T.N. & Baston, R. (editors). 1990. *Twistors in Mathematics and Physics*, Cambridge and New York: Cambridge University Press

Baston, R.J. & Eastwood, M.G. 1989. *The Penrose Transform: Its Interaction with Representation Theory*, Oxford and New York: Oxford University Press

Bateman, H. 1910. *Partial Differential Equations of Mathematical Physics*, New York: Dover

Donaldson, S. & Friedman, R. 1989. Connected sums of self-dual manifolds and deformations of singular spaces. *Nonlinearity*, 2(2): 197–239

Hitchin, N. 1979. Polygons and gravitons. *Mathematical Proceedings of the Cambridge Philosophical Society*, 85: 456–476

Mason, L.J. & Hughston, L.P. (editors). 1990. *Further Advances in Twistor Theory, Vol. I: The Penrose Transform and Its Applications*, Harlow: Longman

Mason, L.J., Hughston, L.P. & Kobak, P.Z. 1995. *Further Advances in Twistor Theory, Vol. II: Integrable Systems, Conformal Geometry and Gravitation*, Harlow: Longman

Mason, L.J., Hughston, L.P., Kobak, P.Z. & Pulverer, K. (editors). 2001. *Further Advances in Twistor Theory, Vol. III: Curved Twistor Spaces*, Boca Raton, FL: Chapman & Hall

Mason, L.J. & Woodhouse, N.M.J. 1996. *Twistor Theory, Self-duality and Twistor Theory*, Oxford and New York: Oxford University Press

Merkulov, S., & Schwachhöfer, L. 1999. Classification of irreducible holonomies of torsion-free affine connections. *Annals of Mathematics*, (2), 150(1): 77–149

Newman, E.T. 1976. Heaven and its properties. *General Relativity and Gravitation*, 7(1): 107–111

Penrose, R. 1976. Nonlinear gravitons and curved twistor theory. *General Relativity and Gravitation*, 7: 31–52

Penrose, R. 1984, 1986. *Spinors and Space-time*, 2 vols., Cambridge and New York: Cambridge University Press

Ward, R.S. 1977. On self-dual gauge fields. *Physics Letters*, 61A: 81–82

Ward, R.S. 1985. Integrable and solvable systems and relations among them. *Philosophical Transactions of the Royal Society* A, 315: 451–457

Ward, R.S. & Wells, R.O. 1990. *Twistor Geometry and Field Theory*, Cambridge and New York: Cambridge University Press

TWO SOLITON COLLISION

See ***N*-soliton formulas**

U

UEDA EQUATION

See **Duffing equation**

UNIVERSALITY

The concept of universality serves to emphasize like behavior in seemingly unrelated systems. It became popular in the context of phase transitions when it was noted that after a suitable mapping between the thermodynamic variables, the scaling behavior of spin systems near the critical point depends on the number of spin directions, the dimensionality of the system, and the type and range of interactions only (Griffiths, 1970; Binney et al., 1992). Thus, there is no "universal" universality, but only a restricted one among systems that belong to the same universality class.

A deeper understanding of the empirically established relations emerged within the renormalization group treatment of phase transitions, which allowed identification of the parts of the interactions that are relevant for an assignment to a universality class (Wilson, 1983; Binney et al., 1992).

An elementary example of thermodynamic universality is provided by the law of corresponding states for interacting gases. They can be described over a wide range of temperature T, pressure p, and volume V by the van der Waals equation of state. For one mole of the gas it is

$$(p + a/V^2)(V - b) = RT \tag{1}$$

with R being the gas constant and a and b parameters characteristic of the gas. This equation of state has a critical point where both $\partial p/\partial V = 0$ and $\partial^2 p/\partial V^2 = 0$, given by

$$RT_\mathrm{c} = \frac{8a}{27b}, \qquad p_\mathrm{c} = \frac{a}{27b^2} \text{ and } V_\mathrm{c} = 3b. \tag{2}$$

Because of the dependence on the material parameters a and b, these quantities differ from gas to gas. However, when p, T, and V are expressed in terms of the critical values, the material constants disappear. With $\tilde{p} = p/P_\mathrm{c}$, $\tilde{V} = V/V_\mathrm{c}$, and $\tilde{T} = T/T_\mathrm{c}$ the reduced quantities, the van der Waals equation of state becomes

$$(\tilde{p} + 3/\tilde{V}^2)(3\tilde{V} - 1) = 8\tilde{T}. \tag{3}$$

Interestingly, this mapping between states of different gases remains useful even when the gases do not follow the predictions from van der Waals theory, for example, near the critical point and in the liquid-gas coexistence region (Stanley, 1971; Binney et al., 1992).

Universality in thermodynamic systems is often characterized quantitatively by the exponents in the power laws with which quantities, such as the specific heats, susceptibilities, and spatial or temporal correlation functions, diverge near a critical point (if the exponent vanishes, then there can be logarithmic variations or discontinuities). The ideal behavior in the thermodynamic limit is often clouded by corrections because of the finite size of the sample, and these very often show power law dependencies too. The search for universality in other systems and situations is, hence, often accompanied by investigations of scaling laws and comparisons of exponents.

The success of universality considerations in thermodynamics has triggered many investigations in nonlinear dynamical systems. Universality arguments can help to link behavior in one-dimensional maps, in chemical reactions, in electrical circuits, or in hydrodynamic systems, for instance. Here are some situations where universality and scaling appear in nonlinear systems.

Period-doubling cascade. Grossmann & Thomae (1977), Feigenbaum (1978), and Coullet & Tresser (1978) noted that the sequence of parameters λ_n in quadratic maps, like $x_{i+1} = \lambda_n(1 - 2x_i^2)$ where bifurcations from orbits of period 2^n to 2^{n+1} occur, that is, $\lambda_1 = 0.70716$, $\lambda_2 = 0.80953$, $\lambda_3 = 0.83112$, $\lambda_4 = 0.83574$, etc., form a geometric sequence, $\lambda_n = \lambda_\infty - a\delta^{-n}$ (a is a constant). The ratio

$$\delta = \frac{\lambda_n - \lambda_{n-1}}{\lambda_{n+1} - \lambda_n} \tag{4}$$

converges to a value

$$\delta = 4.669201609\ldots . \tag{5}$$

This values appears for all maps with a single quadratic maximum and negative Schwartzian derivative ($f'''/f' - (\frac{3}{2})(f''/f')^2) < 0$), as a renormalization of the map due to Cvitanovic (1984), Feigenbaum (1979), Coullet & Tresser (1978), and Lanford (1982) shows. It also appears in higher-dimensional dissipative maps, in continuous differential equations, and even in partial differential equations. Experimental evidence was first found in experiments on thermal convection (Libchaber & Maurer, 1980) and in acoustical, electrical, chemical, and many other cases since. Examples are given in Schuster (1988) and Cvitanovic (1984).

Further universal scaling laws have been established for period n-tupling, for period doubling with other forms of the maximum, for the scaling of the splitting of iterates in period doubling, for the amplitudes of higher harmonics in Fourier spectra, and for the influence of noise on period-doubling bifurcations (Cvitanovic, 1984; Schuster, 1988). Period-doubling in conservative systems has different scaling exponents (MacKay & Meiss, 1987).

Other situations with universal scaling laws arise in the case of intermittency (saddle-node bifurcations) and near the break-up of tori in conservative systems (Cvitanovic, 1984; Schuster, 1988).

Pattern formation. In many pattern forming systems, the equation for the amplitude of the pattern is of Ginzburg–Landau type, with the interactions dictated by continuous and discrete symmetries of the system (Golubitsky & Schaeffer, 1985; Golubitsky et al., 1988). A simple example for a system with a real amplitude A and invariance under $A \to -A$ is

$$\partial_t A = \varepsilon A - A^3 + \Delta A\,, \qquad (6)$$

If the control parameter ε is negative, there is no pattern. For positive ε the amplitude increases like $\varepsilon^{1/2}$. This behavior is widely observed (Cross & Hohenberg, 1993).

Singularity formation. The final stages during the formation of singularities often show asymptotic scaling behavior. For instance, two gravitating bodies starting at rest will collide with distance vanishing like $t^{2/3}$ and velocities diverging like $t^{-1/3}$. The formation of the pinching off of a cylindrical column of liquid is an example of a process that is universal not only in its scaling behavior but also in the prefactors. The formation of the pinch is due to surface tension and is counteracted by viscous effects. From the relevant material parameters surface tension γ (of dimension kg/s^2), viscosity η (kg/(m s)), and density ρ (kg/m^3), one can form a scale of length $l_p = \eta/(\rho\gamma)$ and one of time $t_p = \eta^3/(\gamma^2\rho)$. The behavior of an axisymmetric column is universal in the sense that the minimal diameter $d_{\min}(t)$ and the maximal velocity $v_{\max}(t)$ at a time t before pinch off are given by (Eggers, 1993)

$$\frac{d_{\min}(t)}{l_p} = 0.0608\left(\frac{t}{t_p}\right),$$

$$\frac{v_{\max}(t)}{l_p/t_p} = 3.07\left(\frac{t}{t_p}\right)^{-1/2}. \qquad (7)$$

Random matrix theories. The Hamiltonian operator that describes a given quantum system has a specific form that may be poorly known, as in the case of impurities in a solid or the interactions between nucleons in an atomic nucleus, but it certainly is not arbitrary or random. Nevertheless, for very many systems, the statistical properties of eigenvectors or the statistics of neighboring eigenvalues behave in a universal way that does not depend on the specific system anymore. When rescaled by the mean distance, neighboring energy eigenvalues of hydrogen in strong magnetic fields, electrons in quantum dots, scattering resonances in nuclei such as ^{26}Al, resonance frequencies in microwave resonators, and vibrating quartz blocks can all show the same spacing distribution (Stöckmann, 1999). The only requirement is that the systems are disordered or classically chaotic. The specific form of the distributions then depends on the global symmetry properties of the Hamiltonian, that is, on whether it is real symmetric, complex Hermitian, or symplectic (Guhr et al., 1998; Stöckmann, 1999). System-specific properties appear on much larger scales of separation between eigenvalues; they can be understood within semiclassical periodic orbit theory.

Universality in turbulence. Lewis Fry Richardson described a turbulent flow as a hierarchical arrangement of vortices of different sizes that appear and disappear in an unpredictable fashion. The famous scaling laws of Kolmogorov, Obhukov, Weizäcker, and Onsager hold that in homogeneous, isotropic turbulence, the square of the velocity difference between two points a distance r apart scales like $(\varepsilon r)^{2/3}$, where ε is now the energy dissipation (Frisch, 1995). The expectation is that this is independent of the precise mechanism of stirring, as long as it is confined to large scales and becomes exact as the Reynolds number of the flow approaches infinity. Turbulent fluctuations behind grids in wind or water tunnels or in turbulent jets support this observation.

BRUNO ECKHARDT

See also **Critical phenomena; Dimensional analysis; Fractals; Free probability theory; Kolmogorov cascade; Period doubling; Periodic orbit theory; Random matrix theories: I, II, III, IV; Renormalization groups; Turbulence**

Further Reading

Binney, J.J., Dowrick, N.J., Fisher, A.J. & Newman, M.E.J. 1992. *The Theory of Critical Phenomena: An Introduction to the Renormalization Group*, Oxford: Clarendon Press

Coullet, P. & Tresser, C. 1978. Iterations d'endomorphismes et groupe de renormalization. *Journale de Physique* 39: Colloque C5–C25

Cross, M.C. & Hohenberg, P.C. 1993. Pattern formation outside of equilibrium. *Reviews of Modern Physics*, 65: 851–1112

Cvitanovic, P. 1984. *Universality in Chaos*, Bristol: Adam Hilger

Eggers, J. 1993. Universal pinching of 3D axisymmetric free-surface flow. *Physical Review Letters*, 71: 3458–3460

Feigenbaum, M.J. 1978. Quantitative universality for a class of nonlinear transformations. *Journal of Statistical Physics*, 19: 25–52

Feigenbaum, M.J. 1979. The universal metric properties of nonlinear transformations. *Journal of Statistical Physics*, 21: 669–706

Frisch, U. 1995. *Turbulence: The Legacy of A.N. Komlogorov*, Cambridge and New York: Cambridge University Press

Griffiths, R.B. 1970. Dependence of critical indices on a parameter. *Physical Review Letters*, 24: 1479–1482

Golubitsky, M. & Schaeffer, D.G. 1985. *Singularities and Groups in Bifurcation Theory*, vol. 1, New York: Springer

Golubitsky, M., Stewart, I.N. & Schaeffer, D.G. 1988. *Singularities and Groups in Bifurcation Theory*, vol. 2, New York: Springer

Grossmann, S. & Thomae, S. 1977. Invariant distributions and stationary correlation functions of the one-dimensional discrete processes. *Zeitschrift für Naturforschung*, 32a: 1353–1363

Guckenheimer, J. & Holmes, P. 1983. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, New York: Springer

Guhr, T., Müller-Groeling, A., & Weidenmüller, H.A. 1998. Random-matrix theories in quantum physics: common concepts. *Physics Reports*, 299: 190–425

Lanford, O.E. 1982. A computer assisted proof of the Feigenbaum conjectures. *Bulletin of the American Mathematical Society*, 6: 427–434

Libchaber, A. & Maurer, J. 1980. Une experience de Rayleigh-Bénard de geometrie reduite: multiplication, acchroage, et demultiplication de frequences. *Journal de Physique*, 41: Colloque C3–51

MacKay, R.S. & Meiss, J.D. (editors). 1987. *Hamiltonian Dynamical Systems: A Reprint Collection*, Bristol: Adam Hilger

Schuster, H.G. 1988. *Deterministic Chaos*, 2nd edition, Weinheim: Physik-Verlag and New York: VCH

Stanley, H.E. 1971. *Introduction to Phase Transitions and Critical Phenomena*, Oxford: Clarendon Press and New York: Oxford University Press

Stöckmann, H.J. 1999. *Quantum Chaos: An Introduction*, Cambridge and New York: Cambridge University Press

Wilson, K.G. 1983. The renormalization group and critical phenomena. *Reviews of Modern Physics*, 55: 583–600

UNSTABLE MANIFOLD

See **Phase space**

V

VAN DER POL EQUATION

In 1926, Balthasar van der Pol derived the equation (now named after him) to describe self-sustained oscillations in a triode circuit. To solve the equation, he developed a method that is based on the separation of fast and slow time dependencies and on averaging, an idea that provides the basis of various analytic approaches to nonlinear problems (van der Pol, 1926).

In dimensionless variables, the van der Pol (VDP) equation reads

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} - \mu(1-x^2)\frac{\mathrm{d}x}{\mathrm{d}t} + x = 0. \tag{1}$$

The variable x represents the triode plate voltage in the oscillating circuit, and the frequency of the circuit is normalized to 1 by the appropriate change of the time scale. The parameter $\mu > 0$ gives the growth rate of small linear oscillatory perturbations, and the nonlinear term, approximating the nonlinear current-voltage characteristic of the triode, is normalized by scaling x. Physically, the VDP equation describes growth and saturation of oscillatory perturbations with eventual onset of periodic self-sustained oscillations. With a change of variables $x = \mathrm{d}y/\mathrm{d}t$, the VDP equation transforms to the Rayleigh equation

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} - \mu\left[1 - \frac{1}{3}\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2\right]\frac{\mathrm{d}y}{\mathrm{d}t} + y = 0.$$

In the case of weak instability and nonlinearity ($\mu \ll 1$), the VDP equation can be treated analytically. Here the oscillations are nearly sinusoidal with slowly (on the time scale of order $1/\mu$) varying amplitude and phase. This key observation constitutes the essence of the method of averaging developed by van der Pol.

The first step in the solution is a transformation from $(x, \dot{x})$ to new variables: the amplitudes A, B; thus,

$$\begin{aligned} x(t) &= A(t)\cos t + B(t)\sin t,\\ \dot{x}(t) &= -A(t)\sin t + B(t)\cos t. \end{aligned} \tag{2}$$

In these variables, Equation (1) reads

$$\begin{aligned} \frac{\mathrm{d}A}{\mathrm{d}t} &= -\mu\dot{x}(1-x^2)\sin t,\\ \frac{\mathrm{d}B}{\mathrm{d}t} &= \mu\dot{x}(1-x^2)\cos t, \end{aligned} \tag{3}$$

where $\dot{x}$, x on the right-hand side are to be expressed through A, B. One can see that the time variations of A, B are slow, and the major contribution to them is given by non-oscillating terms on the right-hand side of (3). Keeping only the non-oscillating terms is equivalent to averaging over the oscillation period 2π, and under the averaging procedure, the slowly varying amplitudes $A(t)$ and $B(t)$ are considered as constants. Having performed the averaging, van der Pol obtained the approximate equations in the form

$$\begin{aligned} \frac{\mathrm{d}A}{\mathrm{d}t} &= \frac{\mu}{2}A\left(1 - \frac{A^2+B^2}{4}\right),\\ \frac{\mathrm{d}B}{\mathrm{d}t} &= \frac{\mu}{2}B\left(1 - \frac{A^2+B^2}{4}\right). \end{aligned} \tag{4}$$

This equation can be readily solved by means of a transformation to the slow amplitude and phase $A = R\cos\phi$, $B = R\sin\phi$:

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{\mu}{2}R\left(1 - \frac{R^2}{4}\right), \quad \frac{\mathrm{d}\phi}{\mathrm{d}t} = 0.$$

For any initial condition $R \neq 0$, the stationary amplitude $R_0 = 2$ is established as $t \to \infty$. In the original variables of Equation (1), the corresponding periodic solution reads $x(t) = 2\cos(t - \phi)$ (see Figure 1). In terms of the dynamical systems theory, the van der Pol equation (1) possesses a limit cycle; for small μ, the cycle is approximately circular with radius 2. Physically, it corresponds to periodic weakly nonlinear, self-sustained oscillations.

The solution of the VDP equation can be treated analytically also in the limiting case $\mu \gg 1$, when the equation describes relaxation oscillations. Introducing new time $\tau = t/\mu$, we can rewrite Equation (1) as a

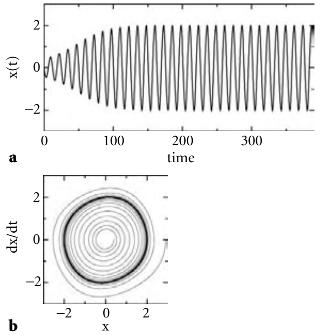

Figure 1. A solution of the van der Pol equation for $\mu = 0.1$. (a) The time dependence of x. (b) The limit cycle on the phase plane $(x, \mathrm{d}x/\mathrm{d}t)$ attracts other trajectories; its form for this small value of μ is nearly circular.

system

$$\mu^{-2}\frac{\mathrm{d}x}{\mathrm{d}\tau} = -y + x - \frac{x^3}{3}, \qquad (5)$$

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = x. \qquad (6)$$

This system has a small parameter μ^{-2} at the derivative, and it belongs to the class of singularly perturbed equations. All motions in the phase space (x, y) are divided into fast motions, when variable x jumps with the rate $\sim \mu^2$ while variable y remains nearly constant, and slow motions, when both variables x and y vary with rates of ~ 1. The slow motions are restricted to the slow manifold, where the right-hand side of Equation (5) vanishes (the curve $y = x - x^3/3$ shown as a dotted line in Figure 2b). More precisely, they are restricted to the stable branches of the curve, where $\mathrm{d}(x - x^3/3)/\mathrm{d}x < 0$. The direction of motion on these stable branches is determined by Equation (6) and is depicted in Figure 2b by arrows. When the phase point moving along a stable branch arrives at its border, it jumps to the other branch. Thus, the limit cycle consists of two pieces of slow motion connected by two pieces of fast motion (solid line in Figure 2b).

The VDP equation serves as a prototype for different dynamical models with a limit cycle. For example, an equation with a more complex nonlinearity and $\mu > 0$,

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} - \mu(-1 + x^2 - \beta x^4)\frac{\mathrm{d}x}{\mathrm{d}t} + x = 0,$$

describes the so-called "hard excitation" of self-sustained oscillations. (This equation was derived by van der Pol and Appleton for the description of a triode generator, whose operating point is shifted from the inflection point of the current-voltage characteristics of the triode.) In this case, the averaging method above described leads to the following amplitude equation:

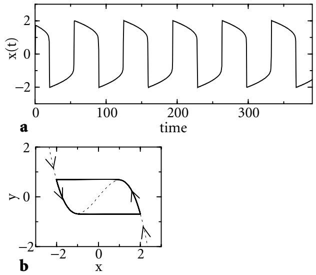

Figure 2. A solution of the van der Pol equation for $\mu = 20$. (a) The time dependence of x. (b) The limit cycle on the phase plane (x, y) consists of pieces of the slow manifold (shown with a dotted line) and fast jumps of x.

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{\mu}{2}R\left(-1 + \frac{R^2}{4} - \beta\frac{R^4}{8}\right), \quad \frac{\mathrm{d}\phi}{\mathrm{d}t} = 0.$$

For $\mu > 0$ and $0 < \beta < \frac{1}{8}$, this equation possesses coexisting stable steady state $R = 0$, an unstable limit cycle at $R_{\mathrm{un}} = \beta^{-1/2}(1 - \sqrt{1 - 8\beta})^{1/2}$, and a stable limit cycle at $R_{\mathrm{st}} = \beta^{-1/2}(1 + \sqrt{1 - 8\beta})^{1/2}$. The basin of attraction of the stable limit cycle is $R > R_{\mathrm{un}}$, while the circle $R < R_{\mathrm{un}}$ is the basin of the stable fixed point. For the triode generator this means that self-sustained oscillations cannot develop from small fluctuations, but appear only if a relatively large (larger than R_{un}) perturbation is applied.

Another generalization of the van der Pol equation is often called the van der Pol–Duffing oscillator:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} - \mu(1 - x^2)\frac{\mathrm{d}x}{\mathrm{d}t} + x + \gamma x^3 = 0. \qquad (7)$$

This model combines the dissipative nonlinearity of the van der Pol equation (term proportional to μ) with the conservative nonlinearity of the Duffing equation (term proportional to γ). For $\mu \ll 1$ and $\gamma \ll 1$, the equations for the slowly varying amplitude and phase can be obtained by the method of averaging as

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{\mu}{2}R\left(1 - \frac{R^2}{4}\right), \quad \frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{3\gamma}{4}R^2.$$

The difference to the van der Pol case is in the phase dynamics. Now the oscillations are non-isochronous, because the dynamics of the slow phase depends on the amplitude. In particular, the frequency of the self-sustained oscillations differs from the frequency of linear oscillations by 3γ. In the VDP equation (1), such a frequency shift appears only in the second approximation (by taking into account the terms $\sim \mu^2$).

Another model related to Equation (1) is the Bonhoeffer–van der Pol (BVDP) oscillator, with equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -y + x - \frac{x^3}{3} + I_0,$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = c(x + a - by).$$

Writing this system as one second-order equation, one gets a model similar to Equation (7), but with additional terms; thus,

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} - (1 - bc - x^2)\frac{\mathrm{d}x}{\mathrm{d}t} + c(1 - b)x + \frac{bc}{3}x^3 + c(a - bI_0) = 0\ . \qquad (8)$$

In the case $b = I_0 = 0$, the BVDP model reduces to the FitzHugh–Nagumo model of neuron spiking. Depending on the parameters, both these models demonstrate self-sustained oscillations, excitability (a stable steady state that responses to a finite perturbation by generating a spike), or bistability.

A periodically forced VDP equation,

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} - \mu(1 - x^2)\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x = E\sin\omega t,$$

displays the phenomenon of frequency locking. For ω close to the natural frequency of the autonomous system ω_0, the forced systems starts to oscillate with the frequency of the forcing term or becomes entrained. Entrainment occurs even for relatively small E, especially when $\mu \gg 1$. For large E, the forced VDP equation can demonstrate chaotic regimes.

ARKADY PIKOVSKY AND MICHAEL ROSENBLUM

See also **Attractors; Averaging methods; Duffing equation; FitzHugh–Nagumo equation; Nonlinear electronics; Relaxation oscillators; Synchronization**

Further Reading

Andronov, A.A., Vitt, A.A. & Khaykin, S.E. 1966. *Theory of Oscillators*, Oxford and New York: Pergamon Press (original Russian edition, 1937)

Bogoliubov, N.N. & Mitropolsky, Yu.A. 1961. *Asymptotic Methods in the Theory of Nonlinear Oscillations*, New York: Gordon and Breach

van der Pol, B. 1926. On "relaxation oscillations." *Philosophical Magazine*, 2: 978–992

VECTOR FIELD

See **Phase space**

VERHULST EQUATION

See **Population dynamics**

VIRIAL THEOREM

First established by Rudolf Clausius in 1870, the virial theorem relates the average potential energy $\langle V\rangle$ to the average kinetic energy $\langle T\rangle$ of a system of particles. The particles can have arbitrary potential interactions, and the theorem holds both in classical and quantum mechanics.

For a single classical nonrelativistic particle, the theorem can be derived as a consequence of Newton's second law: $\boldsymbol{F} = m\mathrm{d}^2\boldsymbol{x}/\mathrm{d}t^2$. Thus,

$$\frac{\mathrm{d}(m\boldsymbol{x}\cdot\mathrm{d}\boldsymbol{x}/\mathrm{d}t)}{\mathrm{d}t} = \boldsymbol{x}\cdot\boldsymbol{F} + m\left(\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right)\cdot\left(\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right), \quad (1)$$

which in terms of the momentum $\boldsymbol{p} = m(\mathrm{d}\boldsymbol{x}/\mathrm{d}t)$ and the kinetic energy $T = m(\mathrm{d}\boldsymbol{x}/\mathrm{d}t)^2/2$ reads

$$-\boldsymbol{x}\cdot\boldsymbol{F} = -\frac{\mathrm{d}(\boldsymbol{x}\cdot\boldsymbol{p})}{\mathrm{d}t} + 2T. \qquad (2)$$

In taking the average over time, the first term on the right-hand side of the equation does not contribute; thus, one obtains

$$-\langle\tfrac{1}{2}\boldsymbol{x}\cdot\boldsymbol{F}\rangle_{\mathrm{av}} = \langle T\rangle_{\mathrm{av}}, \qquad (3)$$

which is the classical virial theorem. The expression on the left-hand side of this equation (called the virial by Clausius) is a measure of the net attractiveness of a system.

In the special case when the force is derivable from a potential $\boldsymbol{F} = -\nabla V$ and the potential varies as the nth power of the distance to the origin $V = C|\boldsymbol{x}|^n$, the term $\boldsymbol{x}\cdot\boldsymbol{F} = -nV$, and the virial theorem reduces to

$$n\langle V\rangle_{\mathrm{av}} = 2\langle T\rangle_{\mathrm{av}}\ . \qquad (4)$$

The theorem can also be obtained from a variational derivation.

Applications

- For the harmonic oscillator, $V(\boldsymbol{x}) = \frac{1}{2}mw^2|\boldsymbol{x}|^2$ then $n = 2$ and $E \equiv \langle V\rangle + \langle T\rangle = 2\langle T\rangle = 2\langle V\rangle$.
- For a bouncing ball, $V(\boldsymbol{x}) = mg|\boldsymbol{x}|$ and $n = 1$, then $E = 3\langle T\rangle = 3/2\langle V\rangle$.
- If the forces vary inversely as the square of the distance (as in atomic and planetary systems), then $V(\boldsymbol{x}) = e^2/|\boldsymbol{x}|$ and $n = -1$. The average kinetic energy is then numerically equal to the total energy $E = -\langle T\rangle$ but of opposite sign.
- In astrophysics, the virial theorem is a valuable tool for studying static, non-evolving (relaxed) systems such as stars (systems of gas particles), gas clouds, star clusters, galaxies, and galaxy clusters. This theorem can be used to estimate the mass of a given system since for the gravitational case $\langle T\rangle = -\frac{1}{2}\langle V\rangle$ and

$$\frac{1}{2}M_{\mathrm{tot}}\left\langle\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\right\rangle^2 = +\frac{1}{4}\frac{GM_{\mathrm{tot}}^2}{\langle R_{\mathrm{tot}}\rangle}\ .$$

Thus, the virial mass of the system can be estimated as

$$M_{\rm tot} \approx 2\frac{\langle R_{\rm tot}\rangle \langle \mathrm{d}\boldsymbol{x}/\mathrm{d}t\rangle^2}{G}.$$

- Another application of this theorem is in the classical calculation of the state equation of gases. The system considered is that of a large number N of particles with coordinates $\boldsymbol{x}_i$ acted on by external and internal forces $\boldsymbol{F}_i = \boldsymbol{F}_{i,\rm ext} + \boldsymbol{F}_{i,\rm int}$ and confined within a box of volume $\mathcal{V}$. As the contribution from the external forces can be related to the pressure of the gas on the walls of the box, the virial theorem takes the form

$$P\mathcal{V} = Nk_{\rm B}T + \frac{1}{3}\left\langle \sum_i^N \boldsymbol{x}_i \cdot \boldsymbol{F}_{i,\rm int} \right\rangle_{\rm av},$$

where $k_{\rm B}$ is the Boltzmann constant. This was the application originally considered by Clausius.
- The virial theorem can be extended to a quantum system described by the linear dimensionless Schrödinger equation

$$\mathrm{i}\psi_t = -\Delta\psi + V(\boldsymbol{x})\psi\,.$$

In this case, $2T_m = nV_m$, where T_m and V_m are the kinetic and potential energy of the mth eigenstate, respectively.
- The virial theorem can also be applied to the study of the solitary waves. As an illustration, consider the one-dimensional nonlinear Schrödinger (NLS) equation

$$\mathrm{i}u_t + u_{xx} + 2(u^*u)u = 0\,, \qquad (5)$$

which has solitary wave solutions that preserve their shape after collision with other solitary waves. The general form of a solitary wave is

$$u(x,t) = a\exp\mathrm{i}\left[\frac{v_{\rm e}}{2}x + \left(a^2 - \left(\frac{v_{\rm e}}{2}\right)^2\right)t - \phi_0\right] \times \mathrm{sech}\,[a(x - v_e t - x_0)]\,, \qquad (6)$$

where a is the wave amplitude and $v_{\rm e}$ the envelope velocity, while ϕ_0 and x_0 are the initial phase and position.

One of the infinite conservation laws associated to (5) is the energy

$$E = \int u_x^* u_x\,\mathrm{d}x + \int -(u^*u)^2\,\mathrm{d}x,$$

where the first term is the kinetic energy (T) and the second term is the potential energy (V).

The NLS equation is derived from the Lagrangian density $\mathcal{L} = \frac{1}{2}(uu_t^* - u_tu^*) + u_x^*u_x - (u^*u)^2$ and the corresponding action is $\mathcal{S} = \int \mathrm{d}t\,\mathrm{d}x\mathcal{L}$. Consider a stationary localized solution $u = \exp(\mathrm{i}\Omega t)\;u(x)$, which is the case of solution (6) with $v_{\rm e} = 0$ and $\Omega = a^2$. As the Lagrangian density is static, we can apply a dilation transformation to get the global condition $\int |u_x|^2\,\mathrm{d}x = \int(\Omega|u|^2 - |u|^4)\,\mathrm{d}x$. From Equation (5), $\int(\Omega|u|^2 + |u_x|^2 - 2|u|^4)\,\mathrm{d}x = 0$, which in turn implies that $\int |u_x|^2\,\mathrm{d}x = \frac{1}{2}\int |u|^4\,\mathrm{d}x$. Thus the total energy is $E = -T = \frac{1}{2}V$, and in the case of the stationary soliton, $E = -2a^3/3$.

Thus, the virial theorem provides nontrivial relations between quantities of physical interest, which can be used to test the accuracy of numerical simulations and to find variational solutions.

LUIS VAZQUEZ AND M.P. ZORZANO

See also **Damped-driven anharmonic oscillator; Nonlinear Schrödinger equations; Rotating wave approximation**

Further Reading

Goldstein, H. 1980. *Classical Mechanics*, 2nd edition, Reading, MA: Addison-Wesley

Saslaw, W.C. 1985. *Gravitational Physics of Stellar and Galactic Systems*, Cambridge and New York: Cambridge University Press

Scott, A. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

VISIOMETRICS

With the advent of robust and rapid sensing devices and increasing computer speeds, memories and transfer rates, the technology for observation, measurement, and computer simulation of nonlinear processes in physical and biological systems has improved rapidly, allowing massive experimental and simulated data sets to be produced. In order to gain physical insight into the evolving phenomena and construct reduced mathematical models, the mathematical essences in this sea of data need to be determined.

This can be accomplished through the process of visiometrics, which involves visualization, projection, identification and classification, extraction, tracking, quantification, and juxtaposition of evolving amorphous coherent structures and statistical backgrounds in massive multidimensional data sets. The goal is to produce cogent images and specific, parameter-scaled (normalized) graphs for intuitive understanding and mathematical analyses (Bitz & Zabusky, 1990; Zabusky et al., 1993; Feher & Zabusky, 1996; Fernandez et al., 1996; Zabusky, 1999).

Consider the process of dealing with evolving simulation data. Most simulations use continuum partial or integro-differential equations or discrete particle descriptions and produce numerous fields, particle locations, and trajectories. (In hybrid codes both are present.) Examples of continuum fields mainly from the compressible Euler equations of fluid motion are presented here. For particle data (e.g., from Monte Carlo or plasma simulations), appropriate local averages can be used to convert to continuum quantities.

In fluid flows, we deal with scalars such as density, pressure, and temperature, vectors such as velocity $\boldsymbol{u}$ and vorticity $\boldsymbol{\omega}$, and tensors such as the rate of strain tensor $\nabla \boldsymbol{u}$. Very thin high-gradient regions such as shock waves and diffusing interfacial transition layers (ITLs) (Zabusky et al., 2003) between species with different properties may be embedded in the flows. To quantify the ITLs, we may extract a medial axis (i.e., some center line) or surface, so we can determine tangents and normals to these curves in two-dimensions (2-d) and surfaces (in 3-d), respectively. In 3-d, the curvatures of these medial surfaces may be important tensors. The visiometrics operations of visualization, projection, identification and classification, extraction, tracking, quantification, and juxtaposition are each considered below.

Visualization: Numerical fields are displayed in a variety of 2-d and 3-d images. The data are first preprocessed or filtered to make them more accessible visually. High wave number incoherent modes are removed with a filter (e.g., the wavelet process of Farge et al., 2001). The choice of appropriate color maps (palettes) (Farge, 2000) is a nontrivial operation. Farge (1992) represented the vorticity scalar with a grayscale format for lower magnitudes, a yellow intermediate contour for the region near zero, and a few graded colors for the higher magnitudes. In DAVID (see http://www.caip.rutgers.edu/~nzabusky/vizlab_cfd/david/david.html), an interactive color domain system has been developed with colors chosen, for example, with hue, saturation, and value (HSV), and with optional black/white grading superimposed in each color domain. The user sees the image automatically colored as the color map is modified. Simultaneously, at each color transition, the numerical value of the function and between neighboring color transitions, the integrated content, and the corresponding underlying pixel-area of the image are seen. (This gives the user a qualitative feeling about the magnitudes of the objects that the color renders.)

Standard visualization techniques include continuous or discrete contour maps in 2-d; volume rendering a function in 3-d as it would appear when radiating or reflecting light from various sources (Upson & Keeler, 1988); and displaying isosurfaces (contours) from connected polygons that bound regions of functions extracted by thresholding (Lorensen & Cline, 1987). There are many excellent computer vision and scientific visualization contemporary texts in the literature (both web and printed).

Projection: A general process produces abstract images and graphical information in lower dimensions, for example, the projection to 2-d from 3-d by integration with respect to a kernel. If it is a planar delta function, one extracts 2-d surfaces or curves from 3-d or 2-d spatial data, respectively. Also, integration of an appropriately weighted field variable transverse to some initial axis or flow direction (like that behind a planar shock in a shock tube) will produce a space–time diagram from a 2d+1 data set (see color plate section; Figure 2 of Hawley & Zabusky, 1989; see also Zabusky, 1999).

Identification and classification: The geometry, topology, content, moments, and distribution functions of an extracted region are examined. Are the domains in 3-d layer-like or tube-like? Are the tube-like domains right- or left- handed helices?

Extraction and setting of thresholds: Complex data sets may have a hierarchy of embedded coherent structures in a sea of incoherent very small-scale structures. Farge and colleagues (Farge et al., 2001) have applied a threshold prescription to wavelet amplitude coefficients and have developed high-compression techniques to systematically extract coherent vortex objects. Samtaney and Zabusky (2000) have examined the extraction of shocks and species ITLs in 2-d. For ITLs, one looks at absolute values of gradients and Laplacians of the density, temperature, and so on. Various heuristic and analytical methodologies have been presented by Villasenor & Vincent (1992), Melander & Hussain (1994), Jeong & Hussain (1995), Kida & Miura (1998), and Miura (2002). The results of Melander, Jeong, and Hussain are particularly noteworthy.

Tracking: Structures in space-time that have been extracted and identified are followed. One must allow for objects to collide and amalgamate, split, be created, and disappear (Samtaney et al., 1994). Post et al. (2002) and colleagues (e.g., Vrolijk et al., 2003) have carried this work further. With this information, we may be able to formulate kinematic and dynamical models.

Quantification: Graphs of projected and tracked structures are plotted and underlying physical, mathematical, and numerical parameters are varied. In the vicinity of extrema (e.g., magnitude of vorticity $\boldsymbol{\omega}$), second- and third-order moments are useful (e.g., ellipsoids in 3-d). Distribution functions of one or more variables are evaluated, and statistical characterizations of incoherent domains, after structures have been extracted, are examined.

Juxtaposition: These comparisons are of similar or different functions at the same or different times from simulations of the same or linked mathematical models in a relevant domain of parameters. Coherent structures of different functions (e.g., density or vorticity) and their quantifications with runs from different resolutions (validation) and parameters (scaling physical behavior) are compared. One often has to remove a background translation or rotation.

As nonlinear phenomena are simulated at increasingly high resolution and for longer times on adaptive meshes across parallel processors, the validity of results becomes an important issue. Numerical errors accumulate from round-off, truncation (e.g., higher-order and nonlinear dissipative and dispersive regularizations), spatially and temporally adjusting meshes, and ad-hoc filters. The visiometric process will help find these defects and provide a more rigorous basis for model building and prediction. Visiometrics will also produce new art forms (Zabusky, 2000).

N.J. ZABUSKY

See also **Contour dynamics; Fluid dynamics; Vortex dynamics of fluids; Wavelets**

Further Reading

Bitz, F. & Zabusky, N. 1990. DAVID and “Visiometrics”: visualizing and quantifying evolving amorphous objects. *Computers in Physics*, 4: 603–14

Farge, M. 1992. Wavelet transforms and their application to turbulence. *Annual Reviews of Fluid Mechanics*, 24: 395–457

Farge, M. 2000. Choice of representation modes and color scales for visualization in computational fluid dynamics. In *Proceedings of the Science and Art Symposium*, edited by A. Gyr, P.D. Koumoutsakos & U. Burr, Dordrecht: Kluwer, pp. 91–100

Farge, M.G., Pellegrino, G. & Schneider, K. 2001. Coherent vortex extraction in 3D turbulent flows using orthogonal wavelets. *Physical Review Letters*, 87(5): 054501-1–054501-4

Feher, A. & Zabusky, N. 1996. An interactive imaging environment for scientific visualization and quantification. *International Journal of Imaging System Technology*, 7: 121–30

Fernandez, V.M., Silver, D. & Zabusky, N.J. 1996. Visiometrics of complex physical processes: diagnosing vortex-dominated flows. *Computers in Physics*, 10: 463–470

Hawley, J. & Zabusky, N. 1989. Vortex paradigm for shock accelerated density stratified interfaces. *Physical Review Letters*, 63: 1241–44

Jeong, J. & Hussain, F. 1995. On the identification of a vortex. *Journal of Fluid Mechanics*, 285: 69–94

Kida, S. & Miura, H. 1998. Identification and analysis of vortical structures. *European Journal of Mechanics*, B/Fluids, 17(4): 471–488

Lorensen, W.E. & Cline, H.E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics*, 21(3): 163–169

Melander, M.V. & Hussain, F. 1994. Topological vortex dynamics in axisymmetric viscous flows. *Journal of Fluid Mechanics*, 260: 57–80

Miura, H. 2002. Analysis of vortex structures in compressible isotropic turbulence. *Computer and Physics Communications*, 147: 552–554

Post, F.H., Vrolijk, B., Hauser, H., Laramee, R.S. & Doleisch, H. 2002. Feature extraction and visualization of flow fields. In *Eurographics* 2002 *State-of-the-Art Reports*, edited by D. Fellner, & R. Scopigno, pp. 69–100. See also http://visualization.tudelft.nl/

Samtaney, R., Silver, D., Zabusky, N. & Cao, J. 1994. Visualizing features and tracking their evolution. *IEEE Computer*, 27(7): 20–27

Samtaney, R. & Zabusky, N.J. 2000. Visualization, feature extraction and quantification of numerical visualizations of high-gradient compressible flows. In *Flow Visualization, Techniques and Examples*, edited by A. Smits & T.T. Lim, London: Imperial College Press, pp. 317–344

Upson, C. & Keeler, M. 1988. V-BUFFER: visible volume rendering. *Computer Graphics*, 22(4): 59–64

Villasenor, J. & Vincent, A. 1992. An algorithm for space recognition and time tracking of vorticity tubes in turbulence. *CVGIP: Image Understanding*, 55: 27–35

Vrolijk, B., Reinders, F. & Post, F.H. 2003. Feature tracking with skeleton graphs. In *Data Visualization: The State of the Art*, edited by F.H. Post, G.M. Nielson & G.P. Bonneau, Boston: Kluwer, pp. 37–52

Zabusky, N.J. 1999. Vortex paradigm for accelerated inhomogeneous flows: visiometrics for the Rayleigh–Taylor and Richtmyer–Meshkov environments. *Annual Review of Fluid Mechanics*, 31: 495

Zabusky, N.J. 2000. Scientific computing visualization—a new venue in the arts. In *Proceedings of the Science and Art Symposium*, edited by A. Gyr, P.D. Koumoutsakos & U. Burr, Dordrecht: Kluwer, pp. 1–11

Zabusky, N.J., Gupta, S. & Gulak, Y. 2003. Localization and spreading of contact discontinuity layers in simulations of compressible dissipationless flows. *Journal of Computational Physics*, 188: 347—363

Zabusky, N.J., Silver, D. & Pelz, R. 1993. Vizgroup ’93. Visiometrics, juxtaposition and modeling. *Physics Today*, 46(3): 24-31

VOLCANOS

See **Geomorphology and tectonics**

VOLTERRA SERIES AND OPERATORS

In addition to playing an important role in the development of theoretical biology, the Italian mathematician Vito Volterra (1860–1940) also strongly influenced the development of modern calculus. We deal here with the Volterra functional series (VFS) and associated Volterra differential operators (VDO) which provide a consistent mathematical framework for stating material properties in nonlinear wave propagation systems, for example, in nonlinear optics (Censor & Melamed, 2002; Censor, 2000; Sonnenschein & Censor, 1998). For simplicity, the present introduction is restricted to the temporal domain, adequate for time signals. Waves require a spatiotemporal domain.

Linear Systems

Modeling of physical systems requires material (or constitutive) relations. In electromagnetics, we have relations like $D = \varepsilon E$, $B = \mu H$; in acoustics, the compressibility (relation of pressure to volume) is needed; in elastodynamics, we include Hooke’s law (relation of stress to strain). Here $D = \varepsilon E$ and its nonlinear extensions are treated as prototypes.

Materials are dispersive, depending (in the restricted sense of temporal dispersion discussed here) on frequency f. Consider the linear case first:

$$D(\omega) = \varepsilon(-\mathrm{i}\omega)E(\omega), \quad \omega = 2\pi f, \qquad (1)$$

where ω is the angular frequency, and defining $\varepsilon(-\mathrm{i}\omega)$ instead of $\varepsilon(\omega)$ is convenient for subsequent applications.

The Fourier transform pair

$$F(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega F(\omega)\mathrm{e}^{-\mathrm{i}\omega t},$$
$$F(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\, F(t)\mathrm{e}^{\mathrm{i}\omega t} \qquad (2)$$

relates the spectral and temporal domains. We use the same symbol F, although $F(t)$ and $F(\omega)$ are different functions. Accordingly, (1) becomes a convolution integral

$$D(t) = \int_{-\infty}^{\infty} \mathrm{d}t_1 \varepsilon(t_1)E(t-t_1), \qquad (3)$$

where $D(t)$, $\varepsilon(t)$, $E(t)$, are related to $D(\omega)$, $\varepsilon(-\mathrm{i}\omega)$, $E(\omega)$, respectively, according to (2). Note that (3) can be viewed as an integral operation, acting on $E(t)$. Also (3) is the simplest form of a VFS.

Formally we can start with (1), transform according to (2), and note that in the integral $E(-\mathrm{i}\omega)$ (if it can be represented or approximated by a polynomial in $-\mathrm{i}\omega$) can be considered as a polynomial operator $\varepsilon(\partial_t)$ acting on the exponential, in which every time derivative ∂_t replaces a term $-\mathrm{i}\omega$ in $\varepsilon(-\mathrm{i}\omega)$. Note that $\varepsilon(t)$ and $\varepsilon(\partial_t)$ are different functions, but $\varepsilon(-\mathrm{i}\omega)$ and $\varepsilon(\partial_t)$ possess the same functional structure. Thus, instead of (3), we now have the VDO representation

$$D(t) = \varepsilon(\partial_t)E(t) = \varepsilon(\partial_\tau)E(\tau)\,|_{\tau\to t}\,. \qquad (4)$$

The last expression in (4) with the instruction $\tau \to t$ is superfluous here but will be important for the nonlinear case below. The possibility of using this technique for $\varepsilon(\partial_t)$ a rational function (ratio of polynomials) is discussed elsewhere (Censor, 2001).

As a trivial example for (3) and (4), consider a harmonic signal

$$E(t) = E_0\mathrm{e}^{-\mathrm{i}\omega t},$$
$$D(t) = E_0\mathrm{e}^{-\mathrm{i}\omega t}\int_{-\infty}^{\infty} \mathrm{d}t_1\, \varepsilon(t_1)\mathrm{e}^{\mathrm{i}\omega t_1}$$
$$= \varepsilon(-\mathrm{i}\omega)E_0\mathrm{e}^{-\mathrm{i}\omega t} = \varepsilon(\partial_t)E_0\mathrm{e}^{-\mathrm{i}\omega t} \qquad (5)$$

clarifying the role of the VDO in (4).

Nonlinear Systems and the Volterra Series and Operators

In nonlinear systems, the material relations involve powers and products of fields. Can we simply replace (1) by a series involving powers of $E(\omega)$? A cursory analysis reveals that this leads to inconsistencies. Instead, we ask if (3) can be replaced by a "super convolution" and what form that should take. Indeed, the Volterra series provides a consistent mathematical answer to these questions. It is given by

$$D(t) = \sum_m D^{(m)}(t),$$
$$D^{(m)}(t) = \int_{-\infty}^{\infty} \mathrm{d}t_1 \ldots \int_{-\infty}^{\infty} \mathrm{d}t_m$$
$$\times \varepsilon^{(m)}(t_1,\ldots,t_m)E(t-t_1)\ldots E(t-t_m). \qquad (6)$$

Typically, the VFS (6) contains the products of fields expected for nonlinear systems, combined with the convolution structure (3). Various orders of nonlinear interaction are indicated by m.

Theoretically, all the orders co-exist (in practice, the series will have to be truncated within some approximation), and therefore, we cannot inject a time harmonic signal as in (5). If instead, we start with a periodic signal, $E(t) = \sum_n E_n\mathrm{e}^{-\mathrm{i}n\omega t}$ and substitute in (6), we find

$$D^{(m)}(t) = \sum_{n_1,\ldots,n_m} \varepsilon^{(m)}(-\mathrm{i}n_1\omega,\ldots,-\mathrm{i}n_m\omega)$$
$$\times E_{n_1}\ldots E_{n_m}\mathrm{e}^{-\mathrm{i}N\omega t} = \sum_N D_N\mathrm{e}^{-\mathrm{i}N\omega t},$$
$$N = n_1 + \ldots + n_m, \qquad (7)$$

with (7) displaying the essential features of a nonlinear system, namely, the dependence on a product of amplitudes, and the creation of new frequencies as sums (including differences and harmonic multiples) of the interacting signals frequencies. In addition, (7) contains the weighting function $\varepsilon^{(m)}(-\mathrm{i}n_1\omega,\ldots,-\mathrm{i}n_m\omega)$ for each interaction mode.

The extension of (4) to the nonlinear VDO is given by

$$D^{(m)}(t) = \varepsilon^{(m)}(\partial_{t_1},\ldots,\partial_{t_m})$$
$$\times E(t_1)\ldots E(t_m)\,|_{t_1,\ldots,t_m\to t}\,. \qquad (8)$$

In (8), the instruction $t_1,\ldots,t_m \to t$ guarantees the separation of the differential operators, and finally renders both sides of the equation to become functions of t.

The VFS (6), including the convolution integral (3), is a global expression describing $D(t)$ as affected by integration times extending from $-\infty$ to ∞. Physically, this raises questions about causality, that is, how future times can affect past events. In the full-fledged four-dimensional generalization, causality is associated with the so-called "light cone" (Bohm, 1965). It is noted that the VDO representation (4, 8) is local, with the various time variables just serving for bookkeeping of the operators, and where this representation is justified, causality problems are not invoked.

In general, the frequency constraint of (7) is obtained from the Fourier transform of (6), having the form

$$
\begin{aligned}
D^{(m)}(\omega) &= \frac{1}{(2\pi)^{m-1}} \int_{-\infty}^{\infty} \mathrm{d}\omega_1 \ldots \int_{-\infty}^{\infty} \mathrm{d}\omega_{m-1} \\
&\quad \times \varepsilon^{(m)}(-\mathrm{i}\omega_1, \ldots, -\mathrm{i}\omega_m) E(\omega_1) \ldots E(\omega_m), \\
\omega &= \omega_1 + \ldots + \omega_m. \qquad (9)
\end{aligned}
$$

It is noted that in (9), we have $m-1$ integrations, one less than in (6). This tallies with the linear case where (1) and (3) involve zero and one integration, respectively. Consequently, the left- and right-hand sides of (9) are functions of ω and ω_m, respectively. The additional constraint $\omega = \omega_1 + \cdots + \omega_m$ completes the equation and renders (9) self-consistent.

Summary

The modeling of nonlinear media using the VFS and VDO provides a mathematically consistent framework which includes linear media as a limiting case. The model displays the typical ingredients of nonlinear circuits and wave systems, where the nonlinear terms are proportional to the product of the amplitudes of the interacting fields, and the newly created frequencies are sums (or differences, or harmonic multiples) of the interaction frequencies, given by $\omega = \omega_1 + \cdots + \omega_m$. In the quantum-mechanical context this is an expression of the conservation of energy. Not shown here is the associated wave propagation vector constraint $\boldsymbol{k} = \boldsymbol{k_1} + \cdots + \boldsymbol{k_m}$, which in the quantum-mechanical context expresses conservation of momentum. Schetzen (1980) is an excellent source of further reading on VFS in nonlinear systems and has many early references.

Dan Censor

See also **Harmonic generation; Manley–Rowe relations; Nonlinear acoustics; Nonlinear optics**

Further Reading

Bohm, D. 1965. *The Special Theory of Relativity*, New York: Benjamin

Censor, D. 2000, A quest for systematic constitutive formulations for general field and wave systems based on the Volterra differential operators. *Progress in Electromagnetics Research*, 25: 261–284

Censor, D. 2001, Constitutive relations in inhomogeneous systems and the particle–field conundrum. *Progress in Electromagnetics Research*, 30: 305–335

Censor, D. & Melamed, T. 2002. Volterra differential constitutive operators and locality considerations in electromagnetic theory. *Progress in Electromagnetic Research*, 36: 121–137

Schetzen, M. 1980. *The Volterra and Wiener Theorems of Nonlinear Systems*, Chichester and New York: Wiley

Sonnenschein, M. & Censor, D. 1998. Simulation of Hamiltonian light beam propagation in nonlinear media. *Journal of the Optical Society of America* B, 15: 1335–1345

VOLTERRA–LOTKA EQUATIONS

See **Population dynamics**

VOLUME-PRESERVING MAPS

See **Measures**

VORONOÏ DIAGRAMS

See **Tessellation**

VORTEX DYNAMICS IN EXCITABLE MEDIA

Vortices in excitable media include spiral waves in two spatial dimensions and scroll waves in three spatial dimensions. They are described by reaction-diffusion systems of equations,

$$
\begin{aligned}
&\partial_t \boldsymbol{u} = \boldsymbol{f}(\boldsymbol{u}) + \mathbf{D}\nabla^2 \boldsymbol{u} + \varepsilon \boldsymbol{h}, \\
&\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{h} \in \mathbb{R}^{\ell},\ \mathbf{D} \in \mathbb{R}^{\ell\times\ell},\ \ell \geq 2, \qquad (1)
\end{aligned}
$$

where $\boldsymbol{u}(\boldsymbol{r}, t) = (u_1, u_2, \ldots)^{\mathrm{T}}$, is a column-vector of the reagent concentrations, $\boldsymbol{f}(\boldsymbol{u})$ is a column vector of the reaction rates, $\mathbf{D}$ is the matrix of diffusion coefficients, $\varepsilon\boldsymbol{h}(\boldsymbol{u}, \boldsymbol{r}, t)$ is some small perturbation, and $\boldsymbol{r} \in \mathbb{R}^2$ or $\mathbb{R}^3$ is the vector of coordinates on the plane or in space.

In an unbounded two-dimensional medium with $\varepsilon\boldsymbol{h} = 0$, a spiral wave solution rotating with angular velocity ω has the form

$$
\begin{aligned}
\boldsymbol{u} &= \boldsymbol{U}(\boldsymbol{r}, t) = \boldsymbol{U}[\rho(\boldsymbol{r}), \vartheta(\boldsymbol{r}) + \omega t] \\
&\approx \left. \boldsymbol{P}[\rho(\boldsymbol{r}) - \frac{\lambda}{2\pi}(\vartheta(\boldsymbol{r}) + \omega t)] \right|_{\rho\to+\infty}, \qquad (2)
\end{aligned}
$$

where $\rho(\boldsymbol{r})$ and $\vartheta(\boldsymbol{r})$ are the polar coordinates corresponding to the cartesian coordinates $\boldsymbol{r}$. $\boldsymbol{P}(\xi; \omega, \lambda)$ is a periodic wave solution with frequency ω and spatial period λ, so the $\rho \to +\infty$ asymptotic means that isolines are approximately Archimedian spirals with pitch λ. Solutions (2) are typically possible for isolated values of ω and corresponding λ.

Note that system (1) with $\varepsilon\boldsymbol{h} = 0$ is invariant with respect to the Euclidean group of motions of the plane $\{\boldsymbol{r}\}$. Solution (2) is a relative equilibrium, meaning that the states of the wave at all moments of time are equivalent to each other up to a Euclidean motion, namely, a rotation around the origin. If (2) is a solution, then from symmetry

$$
\tilde{\boldsymbol{U}} = \boldsymbol{U}(\rho(\boldsymbol{r} - \boldsymbol{R}_{\odot}), \vartheta(\boldsymbol{r} - \boldsymbol{R}_{\odot}) + \omega t - \Phi_{\odot}) \qquad (3)
$$

is another solution for any constant displacement vector $\boldsymbol{R}_{\odot} = (X_{\odot}, Y_{\odot})^{\mathrm{T}}$ and initial rotation phase $\Phi_{\odot}$. Thus, we have a three-dimensional manifold, parameterized by coordinates $X_{\odot}, Y_{\odot}, \Phi_{\odot}$, of spiral wave solutions neutrally stable with respect to each other. By "dynamics" of the vortices, we understand any deviation of the solutions from the stationary rotation (2).

Meander

A nonstationary rotation of a spiral wave accompanied by constant change of its shape is called meander. It is convenient to describe this phenomenon in terms of the spiral tip, which can be defined as an intersection of selected isolines of two components of the nonlinear field $\boldsymbol{u}$,

$$\begin{aligned} &u_{j_1}(X_\bullet, Y_\bullet, t) = v_1, \quad u_{j_2}(X_\bullet, Y_\bullet, t) = v_2, \\ &\Phi_\bullet = \arg(\partial_x + \mathrm{i}\partial_y) u_{j_3}(X_\bullet, Y_\bullet, t), \end{aligned} \qquad (4)$$

$j_1 \neq j_2$, where $X_\bullet(t)$, $Y_\bullet(t)$ are the coordinates of the tip and $\Phi_\bullet(t)$ is its orientation angle. Typically, a spiral wave in a given system develops the same kind of meander pattern $X_\bullet(t)$, $Y_\bullet(t)$ independent of the initial conditions. Changes of parameters in the same system change the meander pattern, and types of patterns can be qualitatively similar in very different excitable media models.

Possible types of meander can be classified using an orbit manifold decomposition of (1) by the Euclidean group. Evolution of the shape of the wave can be described in coordinates (ξ, η) in a moving frame of reference attached to the spiral tip,

$$\begin{aligned} \partial_t \boldsymbol{u} = \mathbf{D}(\partial_\xi^2 + \partial_\eta^2)\boldsymbol{u} + \big[C_1(t)\partial_\xi + C_2(t)\partial_\eta \\ + \omega(t)\left(\xi\partial_\eta - \eta\partial_\xi\right)\big]\boldsymbol{u} + \boldsymbol{f}(\boldsymbol{u}) \\ u_{j_{1,2}}(0,0) = 0, \quad \partial_\eta u_{j_3} = 0, \end{aligned} \qquad (5)$$

and the movement of the tip is described by ordinary differential equations

$$\begin{aligned} &\frac{\mathrm{d}\Phi_\bullet}{\mathrm{d}t} = \omega(t), \\ &\frac{\mathrm{d}X_\bullet}{\mathrm{d}t} + \mathrm{i}\frac{\mathrm{d}Y_\bullet}{\mathrm{d}t} = (C_1(t) + \mathrm{i}C_2(t))\mathrm{e}^{\mathrm{i}\Phi_\bullet}. \end{aligned} \qquad (6)$$

Equations (5) define a dynamic system with the phase space $\{(\boldsymbol{u}(\xi, \eta), C_1, C_2, \omega)\}$, devoid of the Euclidean symmetry of the original system (1). Knowing the attractor in (5), one can deduce the properties of the meander patterns by integrating the ODE system (6) (see Figure 1).

Forced Drift

Another kind of deviation from (2) is drift of spirals due to perturbations $\varepsilon\boldsymbol{h} \neq 0$. As solutions of family (3) are neutrally stable with respect to each other, a small perturbation of a spiral wave caused by an $\varepsilon\boldsymbol{h}$ limited in time will die out, but it will typically result in a small change in the spiral wave coordinates $X_\odot$, $Y_\odot$ and $\Phi_\odot$.

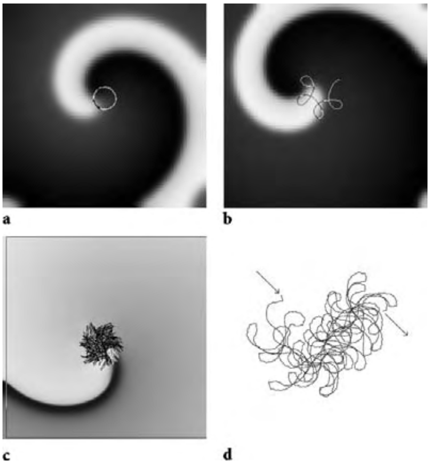

Figure 1. Typical meander patterns. Shown are snapshots of the excitation field with pieces of preceding tip paths superimposed. (a) Stationary (rigid) rotation: equilibrium in the base system (5). (b) Classical biperiodic "flower" meander: limit cycle in the base system (5). (c) Quasi-periodic hypermeander: invariant torus in the base system (5). (d) Pseudorandom walk hypermeander: chaotic attractor in the base system (5) (only the tip path shown).

If similar perturbations are applied repeatedly with a period equal to the period of the spiral, then small shifts of $X_\odot$ and $Y_\odot$ accumulate, which is called a resonant drift. Another type of slow drift is inhomogeneity-induced drift, occurring when $\varepsilon\boldsymbol{h}$ depends explicitly on spatial coordinates (medium properties are slightly inhomogeneous, as in Figure 2(b)). In the first order of perturbation theory, this is equivalent to a time-dependent perturbation synchronized with the spiral rotation and is therefore resonant. A third type of drift occurs if the medium is bounded, and the boundary influence on the spiral wave is not negligible. Although a boundary is not a slight perturbation, the effects of passive (nonflux) on the spiral wave can be small and similar to that of small spatial inhomogeneity (Figure 2(c)). Other kinds of perturbations breaking the Euclidean symmetry of (1) can also cause drift.

Being a first-order effect, the slow drift of a spiral due to small forces of different types obeys a superposition principle. It leads to motion equations

$$\begin{aligned} &\partial_t(X_\odot + \mathrm{i}Y_\odot) = C(X_\odot, Y_\odot) + v(X_\odot, Y_\odot)\mathrm{e}^{\mathrm{i}\Theta}, \\ &\partial_t\Theta = \Omega(t) - \omega(X_\odot, Y_\odot), \end{aligned} \qquad (7)$$

where $C(X_\odot, Y_\odot)$ is the velocity of the inhomogeneity and boundary induced drift, $v(X_\odot, Y_\odot)$ is the velocity of the resonant drift, Θ is the phase difference between the spiral and the resonant forcing, $\Omega(t)$ is the

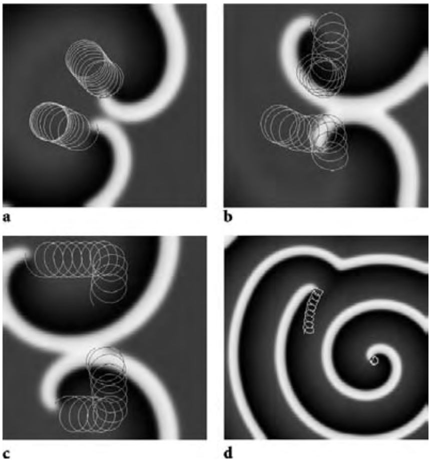

Figure 2. Different drifts of spiral waves. Shown are snapshots of the excitation field with pieces of preceding tip paths superimposed. The right half of the medium is slightly "stronger" than the left half; (a)–(c) are consecutive stages of the same numeric experiment. (a) Two close oppositely charged spirals attract with each other and form a pair drifting in SE direction. (b) The spirals have reached the inhomogeneity and are being driven apart by it. (c) The spirals have reached the medium boundary and now drift along it. (d) In a bigger medium: the right spiral has subdued the left spiral into an induced drift.

perturbation frequency, and $\omega(X_\odot, Y_\odot)$ is the own spiral angular frequency possibly depending on the current spiral location.

These are motion equations for rigidly rotating spirals, and $X_\odot$ and $Y_\odot$ are sliding period averages of $X_\bullet$, $Y_\bullet$. Dynamics of forced meandering spirals are more complicated because of possible resonances.

Spiral Waves as Particles

Motion equations (7) are obtained by summation of the effects of elementary perturbations of different modalities localized in different sites and occurring at different moments of time, onto the spiral's location and phase. These elementary responses are described by response functions, which are critical eigenfunctions of the adjoint linearized operator. An interesting property of the response function is their localization in the vicinity of the spiral core (see Figure 3). The spiral will only drift if the perturbation is applied not too far from its core.

A paradox is thus created. Although a spiral wave appears as a significantly nonlocal process, involving in its rhythm all of the excitable medium, it behaves as a localized, particle-like object in its response to perturbations.

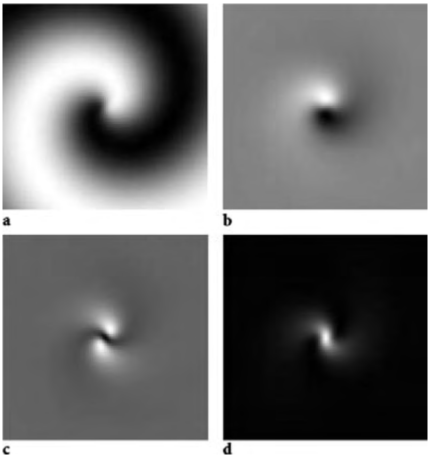

Figure 3. A spiral wave solution (a) and its temporal (b) and spatial (c,d) response functions, as density plots. Monotone gray periphery on (b–d) corresponds to zero. Thus the spiral wave is a non-local process, but its response functions are well localized.

A spatial response function, defining the proportionality between drift velocity and inhomogeneity magnitude, typically has scalar (drift along the parameter gradient or toward the boundary) and pseudo-scalar (across the parameter gradient or along the boundary) components. The sign of the latter depends on the direction of the spiral rotation.

Bending and Twisting of Scroll Waves

As a scroll wave is a three-dimensional analogue of a spiral wave, all comments about spiral wave dynamics remain valid for the scrolls, but there are new aspects arising from the third dimension. The simplest 3-d vortex is the straight scroll wave, a spiral wave continued unchanged in the third dimension. The spiral tip, a point in the plane, becomes the edge of the scroll, a line in space, and the spiral core, a circle in the plane, becomes the scroll filament, a tube. The term *filament* also sometimes denotes the center line of the tube filament.

More interesting regimes are scrolls with bent filaments (Figure 4(a)), and with rotation phase varying along the filament (twisted scrolls, as in Figure 4(b)). Both bending and twist of scrolls are factors of their dynamics. A vortex ring will collapse or expand and, at the same time, drift along its symmetry axis, and twisted vortex will usually spread the twist evenly along its filament or, if possible, untwist.

The asymptotic motion equation for the scroll waves can be derived using response functions. If the twist is

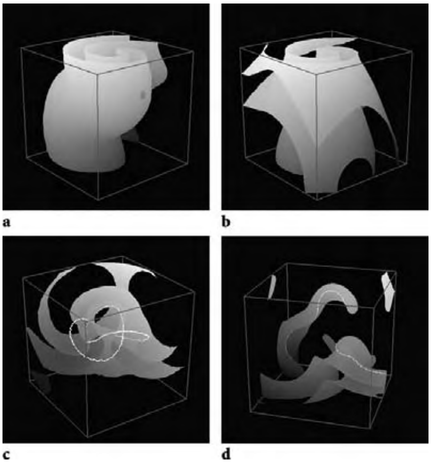

Figure 4. Scroll waves. (a) Scroll with a curved filament. (b) Twisted scroll with straight filament. (c) Scroll with a knotted filament. (d) "Turbulent regime": many scrolls developed from one via negative filament tension multiplication mechanism. On panels (c) and (d), part of the wavefronts is cut out, to make the filaments (white lines) visible.

not too strong, then the dynamics of the scroll due to bending and due to twist are decoupled. The motion equation of the filament is

$$\partial_t \boldsymbol{R}_\mathrm{f} = b_2 \partial_\mathrm{s}^2 \boldsymbol{R}_\mathrm{f} + c_3 \left[\partial_\mathrm{s} \boldsymbol{R}_\mathrm{f} \times \partial_\mathrm{s}^2 \boldsymbol{R}_\mathrm{f} \right], \qquad (8)$$

where $\boldsymbol{R}_\mathrm{f} = \boldsymbol{R}_\mathrm{f}(p, t) \in \mathbb{R}^3$ is the position of the filament as function of a length parameter p and time, and ∂_s is arclength differentiation, $\partial_\mathrm{s} \equiv \left| \partial_p \boldsymbol{R}_\mathrm{f} \right|^{-1} \partial_p$. At $b_2 = 0$, Equation (8) is completely integrable; in particular, the total length of the filament is conserved. Otherwise, the total filament length decreases if $b_2 > 0$ and increases if $b_2 < 0$, and in the latter case, a straight filament is unstable.

If twist is high, it changes the filament tension and may make it negative. This causes an instability of the straight filament shape, leading to "sproing": a sudden transition from a strongly twisted scroll with straight filament to a less twisted scroll with a helical filament.

Competition and Interaction *(Divide et impera)*

Normally, two colliding excitation waves annihilate each other. Thus, if there are many periodic sources of waves (e.g., vortices), then the medium splits into domains, or regions of influence, each domain receiving waves from its source. The domains are separated by shock structures where the waves collide (see Figure 2(a–c)).

The domain boundaries work like non-flux boundaries. Thus, two spiral waves can interact with each other (cause each other's drift and frequency shift), whereas each of them actually interacts with the boundary between their domains. Such interaction between spirals may lead to the formation of linked pairs (see Figure 2(a)).

Different scroll filaments or different parts of the same filament can also interact with each other. If this interaction is repulsive, it may compensate positive tension normally causing closed filaments to contract and collapse, leading to stable "particle-like" 3-d scrolls with compact filaments (see Figure 4(c)).

Induced Drift

If colliding waves annihilate in a ratio of one-to-one, continuity of phase applies. If two vortices have different frequencies (e.g., because of a spatial inhomogeneity of the medium), then by continuity of phase the domain boundary between them moves toward the slower vortex. When it reaches its core, the slower vortex loses its identity and turns into a dislocation in the wave field emitted by the faster vortex. This dislocation, appearing as a free end of an excitation wave, periodically rejoins from one wave to another with some overall drift depending on the frequency and direction of the incident waves (see Figure 2(d)). If the incident wave packet ceases, the dislocation can develop into a vortex again.

As a dislocation is very different from a vortex, this induced drift is an example of hard, non-perturbative dynamics.

Hard Dynamics: Births, Deaths, and Multiplication of Vortices

Another kind of hard dynamics involves complete elimination of a vortex. This may happen if the wave propagation around the vortex becomes impossible, for example, if the vortex has been driven too close to a medium boundary. Alternatively, two spiral waves with opposite topological charges may annihilate if driven too close to each other. For a scroll wave, annihilation may happen to a piece of its filament, which then appears as splitting of a scroll wave into two.

Birth of a vortex may occur as a result of a temporary local block of excitation propagation. Unless this happens near the medium boundary, this means birth of a pair of oppositely rotating spirals in the plane, or a scroll with a closed filament around the perimeter of the propagation block. The block may occur as a result of external forcing or special initial conditions, or develop as a result of an instability of an existing vortex. Such instability can underlie a chain reaction of the vortex multiplication, which may lead to a turbulence of excitation vortices—a spatiotemporal chaotic state where generation of new vortices is balanced by their annihilation when they get close to each other due to overcrowding.

Several mechanisms of such instabilities have been identified, including those working in two or three dimensions, such as Eckhaus instability, zigzag/lateral instability or imposed mechanical movement of the medium, and those possible only in three dimensions. The latter include instability due to the negative "tension" b_2 of the vortex filament (see Figure 4(d)), or caused by spatially inhomogeneous anisotropy of the medium such as that observed heart ventricular muscle. Some of these types of instabilities may be responsible for the phenomenon of fibrillation of the heart (*See* **Cardiac arrhythmias**).

Vadim N. Biktashev

See also **Complex Ginzburg–Landau equation; Framed space curves; Reaction-diffusion systems; Scroll waves; Spiral waves**

Further Reading

Gaponov-Grekhov, A.V., Rabinovich, M.I. & Engelbrecht, J. (editors). 1989. *Nonlinear Waves II. Dynamics and Evolution*, Berlin and New York: Springer

Holden, A.V. & El Naschie, M.S. (editors). 1985. *Nonlinear Phenomena in Excitable Physiological Systems*. Special issue: *Chaos Solitons & Fractals*, 5(3/4)

Holden, A.V., Markus, M. & Othmer, H.G. (editors). 1989. *Nonlinear Wave Processes in Excitable Media*, New York and London: Plenum Press

Kapral, R. & Showalter, K. (editors). 1995. *Chemical Waves and Patterns*, Dordrecht and Boston: Kluwer

Panfilov, A.V. & Holden, A.V. (editors). 1997. *Computational Biology of the Heart*, Chichester and New York: Wiley

Panfilov, A. & Pertsov, A. 2001. Ventricular fibrillation: evolution of the multiple wavelet hypothesis. *Philosophical Transactions of the Royal Society of London* A, 359: 1315–1325

Swinney, H.L. & Krinsky, V.I. (editors). 1991. *Waves and Patterns in Chemical and Biological Media*, Cambridge, MA: MIT Press

Winfree, A.T. (editor). 1998. *Fibrillation in normal ventricular myocardium*. Focus issue. *Chaos*, 8(1)

Winfree, A.T. 1991. Varieties of spiral wave behavior in excitable media. *Chaos*, 1(3): 303–334

Zykov, V.S. 1987. *Modelling of Wave Processes in Excitable Media*, Manchester: Manchester University Press

VORTEX DYNAMICS OF FLUIDS

The swirling and chaotic behavior of vortex-dominated fluid flows has inspired philosophers, poets, and artists from antiquity to the present, including Leonardo da Vinci, the 15th-century artist, scientist, and engineer who captured complex fluid motions as part of his applied and creative work (van Dyke, 1982; Lugt, 1983; Minahen, 1992). The traditional woodblock prints (*Ukioy-e*) of Hiroshige and Hokusai also show spiral motion produced by a vortex and "curls" on breaking ocean waves in various locations around Japan (Zabusky, 2000).

Vorticity is a measure of the "spin" or rotation of the fluid and is usually represented by a continuous vector variable $\boldsymbol{\omega}(\boldsymbol{x}, t)$. Regions of flow that do not contain vorticity are called irrotational. Vorticity can be defined in two equivalent ways. First, as the curl of the velocity $\boldsymbol{u}$, $\boldsymbol{\omega} = \nabla \times \boldsymbol{u}$. Second, and more geometrically, as a limit (when it is finite) of a closed circuit integral over a domain boundary D' (or equivalently the integral over domain D within boundary D') whose enclosed area $A \to 0$, or $\boldsymbol{\omega} = \boldsymbol{e}_{\tan}[\lim |_{A \to 0} \ (\Gamma/A)]$ where for finite A, Γ is the circulation or line integral of velocity around the domain,

$$\Gamma = \int_{D'} \boldsymbol{u} \cdot \mathbf{d}s = \int_{D} \boldsymbol{\omega} \cdot \mathbf{d}A \qquad (1)$$

and $\boldsymbol{e}_{\tan}$ is the tangent to the vorticity vector at the point where $A \to 0$. Note that the scalars $\frac{1}{2}\boldsymbol{\omega} \cdot \boldsymbol{\omega}$ (enstrophy) and $\boldsymbol{\omega} \cdot \boldsymbol{u}$ (helicity) arise frequently in discussions of turbulence.

Leonhard Euler derived continuum mathematical equations for an ideal (non-dissipative) fluid. In the 19th century, Claude Navier, George Stokes, and others included dissipative processes for a realistic viscous fluid (Lamb, 1932; Batchelor, 1967; Kiselev et al., 1999). These equations provide models of fluids under usual (for example nonrelativistic) conditions. Asymptotic (reduced) mathematical equations, particularly those emphasizing inviscid vortex-dominated flows (Saffman, 1992) were derived in the 19th century with works of William Thomson (Lord Kelvin) and Hermann von Helmholtz, among others (*See* **Kelvin–Helmholtz instability**). Vortex models are used today to predict the behavior of geophysical and astrophysical fluids (Nezlin, 1993), and in engineering applications (Green, 1995).

If $\boldsymbol{\omega}(\boldsymbol{x}, t)$ is modeled by one or more line filaments in three dimensions (3-d) on which the vorticity is concentrated, then the velocity produced by this vorticity can be calculated from a (Biot–Savart) line integral over the filaments. With this velocity, one can obtain the trajectories of points in the fluid merely by solving $\mathrm{d}\boldsymbol{x}/\mathrm{d}t = \boldsymbol{u}(\boldsymbol{x}, t)$. In 2-d planar flows, the vorticity vector has one component perpendicular to the plane, $(\omega = \partial_x v - \partial_y u)$, and it may be modeled by one or more points of strength Γ_i.

With these definitions, one can prove some fundamental theorems, including the following (Kiselev et al., 1999):

- *Thomson's theorem*: In an ideal barotropic moving fluid, Γ for any closed contour does not depend on time.
- *Helmholtz's theorem*: If the particles of a liquid satisfy Thomson's theorem and form a vortex filament at some moment of time, then these particles form a vortex filament at all subsequent and previous

moments of time and the vortex tube of circulation Γ will be time invariant and constant along its length.

For 2-d, the simplicity of the kinematic description for homogeneous incompressible fluids allows for simple computer models and valuable insights into the motion of ideal fluids. For example, point circulations Γ_i are usually positive and negative constants; vortex sheets along a line described by the circulation per unit length $\gamma(\boldsymbol{s}, t)$; or vorticity $\omega(\boldsymbol{x}, t)$ is constant in a domain (*See* **Contour dynamics**). (Note that for points and sheets, the equations are ill-posed mathematically, and regularization techniques must be used when computing numerically long times.) Interesting effects observed in 2-d include the merger of like-signed vortex domains if initially placed sufficiently close, binding of opposite-signed domains into "vortex projectiles" (a generic translating form whose simplest example is the dipole composed of two opposite-signed points). For vortex sheets, a linear (Kelvin–Helmholtz) stability analysis shows growth of a perturbation for all wavelengths of the disturbance. These evolve into finite-amplitude rolls with a characteristic spiral depending on their sign (see photo of an atmospheric vortex sheet on page 1 of color plate section). If one relaxes the sheet to a very thin layer then small wavelength perturbations are stabilized whereas wavelengths larger than the initial thickness remain unstable. Placing more than one harmonic perturbation on the thin layer, one observes roll-up and merger (energy cascade to large scales) as well as entrophy cascade to smaller scales; that is, the process becomes turbulent. Many details are found in the references below.

In 3-d, $\boldsymbol{\omega}$ may be concentrated on one or more filamentary lines or in finite area tubes, for example, the well-known vortex ring which translates at uniform speed and is unstable (Shariff & Leonard, 1992). If the ring is perturbed, this can lead to collapse and reconnection (Kida & Takaoka, 1994; Fernandez et al., 1995). In a dramatic experiment, Lim and Nickels (1992, 1995) fired a blue and red vortex ring toward each other in water and observed them interact and split into many blue-red smaller vortex ringlets.

The vorticity evolution equation is obtained if one applies the curl operation to the fluid momentum equation. This shows clearly many fundamental and interesting results on vorticity creation, modification, and dissipation. For the zero viscosity (ideal or Euler equation) limit, we have

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{\omega} + \boldsymbol{\omega}(\nabla \cdot \boldsymbol{u}) = \boldsymbol{\omega} \cdot \nabla \boldsymbol{u} + \rho^{-2}(\nabla \rho \times \nabla p). \qquad (2)$$

On the right-hand side the terms are: (1) vorticity "stretching" by the rate of strain tensor, which is only present in 2-d-axisymmetric and 3-d motions and is essential for turbulence seen in 3-d; and (2) baroclinic terms, which stem from the misalignment between the gradient of density and either the gravitation vector or the gradient of pressure. Both are important in incompressible inhomogeneous multispecies fluids (e.g., the Rayleigh–Taylor instability) and the latter term is important when shock waves interact with density inhomogeneities in compressible fluids (known as the Richtmyer–Meshkov (RM) instability environment). Recently, Ghoneim and colleagues (e.g., Soteriu & Ghoneim, 1995; Reinaud et al., 2000) used a point vortex model to represent inhomogeneous fluids where Γ_i and $\nabla\rho$ vary in time and Zabusky and colleagues have quantified the vortex-accelerated vorticity deposition in the compressible RM environment (Peng et al., 2003) relevant to laser fusion and supersonic combustion.

For real fluids with small dissipation or at high Reynolds number,

$$\mathrm{Re} = \frac{UL}{\nu} = \frac{\text{velocity} \times \text{length}}{\text{kinematic viscosity}}, \qquad (3)$$

these insights are for short times. Vorticity is generated in a thin boundary layer in the vicinity of rigid or compliant objects in the flow—for example, cylinders (Williamson, 1996), spheres, or airfoils—or within channels because of viscosity (non-slip boundaries). Note, that the larger the Reynolds number, the more "unstable" is the fluid motion and the more likely that one finds it in a turbulent state. Cogent discussions of realistic turbulent effects including vortex phenomena are given in recent books (Pope, 2000; Tsinober, 2001).

Vortices are audible. When we hear the wind howling or the crack of a whip, we are sensing vortices in action. Aeroacoustics, the branch of fluid dynamics concerned with sound generated from vortices, is being applied to noise from jet engines, sounds in music and speech, and so on.

Currently, active fields of study include separating turbulent flows into their coherent and incoherent vortex structures (thereby simplifying prediction and control of flows) (*See* **Visiometrics**). These studies arise in geophysics (e.g., hurricanes, the Gulf stream, and Jupiter's red spot) and astrophysics (planetary nebulae, supernova, and galaxy collisions).

N.J. ZABUSKY

See also **Chaotic advection; Contour dynamics; Fluid dynamics; Turbulence; Visiometrics; Vortex dynamics in excitable media**

Further Reading

Batchelor, G. 1967. *An Introduction to Fluid Dynamics*, Cambridge: Cambridge University Press

Fernandez, V.M., Zabusky, N.J. & Gryanik, V.M. 1995. Vortex intensification and collapse of the Lissajous-elliptic ring: single and multiple-filament Biot–Savart simulations and visiometrics. *Journal of Fluid Mechanics*, 299: 289–331

Green, S.I. 1995. (editor). *Fluid Vortices*, Boston: Kluwer

Kida, S. & Takaoka, M. 1994. Vortex reconnection. *Annual Reviews of Fluid Mechanics*, 26: 169–189

Kiselev, S.P, Vorozhtsov, E.V. & Fomin, V.M. 1999. *Foundations of Fluid Mechanics with Applications*, Boston: Birkhauser

Lamb, H. 1932. *Hydrodynamics*, 6th edition, Cambridge: Cambridge University Press

Leonard, A. 1985. Computing three-dimensional incompressible flows with vortex elements. *Annual Reviews of Fluid Mechanics*, 17: 523–559

Lim, T.T. & Nickels, T.B. 1992. Instability and reconnection in the head-on collision of two vortex rings. *Nature*, 357: 225

Lim, T.T. & Nickels, T.B. 1995. Vortex rings. In *Fluid Vortices*, edited by S.I. Green, Boston: Kluwer, pp. 95–153

Lugt, H.J. 1983. *Vortex Flow in Nature and Technology*, New York: Wiley

Minahen, C. 1992. *Vortex/t: The Poetics of Turbulence*, University Park: Pennsylvania State University Press

Nezlin, M.V. 1993. *Rossby Vortices, Spiral Structures and Solitons: Astrophysics and Plasma Physics in Shallow Water Experiments*, New York: Springer

Peng, G., Zabusky, N.J. & Zhang, S. 2003. Vortex-accelerated secondary baroclinic vorticity deposition and late-intermediate times of a two-dimensional Richtmyer–Meshkov interface. *Physics of Fluids*, 15: 3730–3744

Pope, S.B. 2000. *Turbulent Flows*, Cambridge and New York: Cambridge University Press

Reinaud, J., Joly, L. & Chassaing, P. 2000. The baroclinic secondary instability of the two-dimensional shear layer. *Physics of Fluids*, 12(10): 2489–2505

Saffman, P.G. 1992. *Vortex Dynamics*, Cambridge and New York: Cambridge University Press

Shariff, K. & Leonard, A. 1992. Vortex rings. *Annual Reviews of Fluid Mechanics*, 24: 235-279

Soteriu, M.C. & Ghoneim, A. 1995. Effects of the free-stream density ratio on free and forced spatially developing shear layers. *Physics of Fluids*, 7: 2036–2051

Tsinober, A. 2001. *An Informal Introduction to Turbulence*, Boston: Kluwer

van Dyke, M. 1982. *An Album of Fluid Motion*, Stanford, CA: Parabolic Press

Williamson, C.H.K. 1996. Vortex dynamics in the cylinder wake. *Annual Reviews of Fluid Mechanics*, 28: 477–539

Zabusky, N.J. 1999 Vortex paradigm for accelerated inhomogeneous flows: visiometrics for the Rayleigh–Taylor and Richtmyer–Meshkov environments. *Annual Reviews of Fluid Mechanics*, 31: 495–536

Zabusky, N.J. 2000. Scientific computing visualization—a new venue in the arts. In *Science and Art Symposium 2000*, edited by A. Gyr, P.D. Koumoutsakos & U. Burr, Boston: Kluwer

W

WATER WAVES

Water waves have attracted the attention of scientists for many centuries. Although much is now understood, they are a continuing source of fascination, as many aspects of their often-complicated nonlinear behavior remain to be fully elucidated. We describe first the linear theory, much of which was developed in the 19th century, before discussing the more modern developments concerning weakly nonlinear waves. Throughout, the theory is based on the traditional assumptions that water is inviscid, incompressible with a constant density ρ, and in irrotational flow; that is, it has zero vorticity. It follows that for water waves, the governing equation is Laplace's equation, which is linear, and so all the nonlinearity in the problem resides in the free-surface boundary conditions. There is the kinematic condition that the free surface is a material surface at all times and the dynamic condition that the free surface has constant pressure at all times. The weakly nonlinear theory for water waves is described, which employs some of the basic paradigms for nonlinear waves. The special features that emerge when one considers finite-amplitude water waves are described in the review articles of Schwarz & Fenton (1982) and Dias & Kharif (1999).

Linear Waves

When the governing equations of motion are linearized about the rest state, it is customary to make a Fourier decomposition and seek solutions in which the wave elevation (deviation of the water surface from its rest position) is given by

$$\zeta = a\exp\left(\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)\right) + \text{c.c.}, \qquad (1)$$

where $\boldsymbol{x} = (x, y)$ denotes the horizontal coordinates, and t is the time variable, while $\boldsymbol{k} = (k, l)$ is the wave number, ω is the wave frequency, and a is the wave amplitude. Here, c.c. denotes the complex conjugate. This elementary disturbance represents a sinusoidal wave propagating in the direction $\boldsymbol{k}/\kappa$ where $\kappa = |\boldsymbol{k}| = \sqrt{k^2 + l^2}$, with a phase speed $c = \omega/\kappa$, a wavelength $\lambda = 2\pi/\kappa$, and a period $T = 2\pi/\omega$. The corresponding fluid velocity is $\boldsymbol{u} = (u, v)$ and is obtained from a velocity potential (so that $\boldsymbol{u} = \operatorname{grad}\phi$), where

$$\phi = -\,\mathrm{i}ca\frac{\cosh\kappa(z+h)}{\sinh\kappa h}\exp\left(\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)\right) + \text{c.c.} \qquad (2)$$

Here, z is the vertical coordinate, and h is the total depth of water in the rest state (that is, the water occupies the region $-h < z < 0$). Note that the horizontal velocity $\boldsymbol{u}$ is then in phase with the surface elevation, but that the vertical velocity w is $\pi/2$ out of phase, while both velocity components decrease with depth away from the free surface.

Equations (1) and (2) provide a kinematic description of water waves, which to this point means that the conditions of incompressibility and irrotational flow have been satisfied, that the vertical velocity is zero at the bottom, and that the (linearized) kinematic boundary condition that the free surface remains a material surface for all time has been satisfied. To obtain the dynamics, these expressions are substituted into the remaining boundary condition that the free surface is one of constant pressure. This linearized formulation then yields the dispersion relation determining the wave frequency in terms of the wave number,

$$\omega^2 = (g\kappa + \sigma\kappa^3)\tanh(\kappa h), \qquad (3)$$

where $\rho\sigma$ is the coefficient of surface tension, which has a value of 74 dyn/cm at 20°C (*See* **Dispersion relations**). Detailed derivations of (3) and discussions of the consequences for the properties of water waves can be found in many classical and modern texts, see for instance, Lamb (1932), Whitham (1974), Lighthill (1978), or Mei (1983). Indeed, water waves have formed the paradigm for much of our present-day understanding of linear dispersive waves.

There are two branches of the dispersion relation (3), corresponding to waves running to either the right or the left. Note that (3) is isotropic, in that the wave frequency, and hence the phase speed, depend only on the magnitude of the wave number and not its direction. It is apparent from (3) that the effect of surface tension is significant only when $\kappa > (g/\sigma)^{1/2}$; using the above value for σ this corresponds to $\lambda < 1.73$ cm. Such waves are then usually called capillary waves, while waves with $\kappa < (g/\sigma)^{1/2}$ are called gravity waves. Another useful measure of the effect of surface tension is the Bond number $B = \sigma/gh^2$. When $B < 1/3$, the phase speed $c = \omega/\kappa$ decreases from its long wave value $c_0 = (gh)^{1/2}$, achieved at $\kappa = 0$, to a minimum value of $c_{\rm m}$ at a wave number $\kappa = \kappa_{\rm m}$, and then increases without limit as κ increases to infinity. For the case when the Bond number $B > 1/3$, the phase speed increases monotonically from the long wave value c_0 as κ increases from zero. However, the critical depth below which this regime is realized occurs when the Bond number $B = 1/3$, which corresponds to $h = 0.48$ cm. In practice, this is too shallow to ignore the effects of friction and, hence, is usually not regarded as being of any practical interest.

From expressions (2) for the velocity potential and the dispersion relation (3), we see that water waves do not feel the effect of the bottom if $\kappa h \gg 1$. More precisely, if $\kappa h > 2.65$ (or $\lambda/h < 2.37$), then there is less than 1% error in supposing that $h \to \infty$ and that $\tanh(\kappa h) \to 1$. This case describes deep water waves, for which the dispersion relation (3) collapses to

$$\omega^2 = g\kappa + \sigma\kappa^3. \tag{4}$$

In this deep water limit, the Bond number $B \to 0$ ($<1/3$), so that the dispersion relation for the phase speed $c = \omega/\kappa$ has a minimum value at $\kappa = \kappa_{\rm m}$ ($= (g/\sigma)^{1/2}$ here) or $c_{\rm m} = (4g\sigma)^{1/4} = 23.2$ cm/s. There are no sinusoidal waves of the form (1) for any phase speed $c < c_{\rm m}$, while for any $c > c_{\rm m}$ there are two classes of deep water waves, gravity waves with $\kappa < \kappa_{\rm m}$ and capillary waves with $\kappa > \kappa_{\rm m}$.

Long waves are characterized by the limit $\kappa h \to 0$, that is, the limit when $\lambda/h \to \infty$. In this limit, the wave frequency tends to zero, and the wave phase speed tends to $c_0 = (gh)^{1/2}$. In this case, for the reasons discussed above, it is customary to ignore the effects of surface tension, so that there is only one class of waves, namely, gravity waves, which in this limit of $\kappa h \to 0$ are often referred to as shallow water waves. The fluid velocity is then approximately horizontal and independent of the vertical coordinate z, while at $O(\kappa h)$, the vertical component of the fluid velocity is a linear function of $(z + h)$.

In this linearized system, more general solutions can be built up by Fourier superposition of the elementary solutions (1) over the wave number $\boldsymbol{k}$ in which the dispersion relation (3) is satisfied and the wave amplitude a is allowed to be a function of $\boldsymbol{k}$. From this process, it can readily be established that a localized initial state will typically evolve into wave packets, with a dominant wave number $\boldsymbol{k}$ and corresponding frequency ω given by (3), within which each wave phase propagates with the phase speed c, but whose envelope propagates with the group velocity, given by

$$\boldsymbol{c}_{\rm g} = \nabla_{\boldsymbol{k}}\omega. \tag{5}$$

The group velocity is the velocity of energy propagation, where the wave energy density is $2\rho(g + \sigma\kappa^2)|a|^2$, being composed of equal parts of potential and kinetic energy. Because water waves are isotropic, the group velocity is in the direction of the wave number $\boldsymbol{k}$, and has a magnitude

$$c_{\rm g} = \partial\omega/\partial\kappa = \frac{c}{2}\left\{\frac{g + 3\sigma\kappa^2}{g + \sigma\kappa^2} + \frac{2\kappa h}{\sinh 2\kappa h}\right\}. \tag{6}$$

Assuming that the Bond number $B < \frac{1}{3}$, it can be shown for gravity waves (defined by the wave number range $0 < \kappa < \kappa_{\rm m}$) that $c_{\rm g} < c$, with equality in the long wave limit, $\kappa \to 0$ when $c_{\rm g}$ and c tend to c_0, and at the value $\kappa = \kappa_{\rm m}$, $c_{\rm g} = c = c_{\rm m}$. For capillary waves (defined by the wave number range $\kappa > \kappa_{\rm m}$), on the other hand, $c_{\rm g} > c$. In the absence of surface tension (i.e., $B \to 0$), $c_{\rm g} < c$ for all wave numbers $\kappa > 0$, and in the deep water limit $c_{\rm g} \approx c/2$.

Weakly Nonlinear Waves

The theory of linearized waves described in the previous section is valid when initial conditions are such that the waves have sufficiently small amplitudes. However, after a sufficiently long time (or if the initial conditions describe waves of moderate or large amplitudes), the effects of the nonlinear terms in the free surface boundary conditions need to be taken into account. There are three different areas where weak nonlinearity needs to be taken into account, namely, long waves, wave packets, and wave resonances.

Long Waves: Korteweg–de Vries Equation

Initially, we consider unidirectional waves propagating in the positive x-direction, so that the wave number is $\boldsymbol{k} = (k, 0)$. In the long wave limit, $kh \to 0$, the dispersion relation (3) can be approximated by

$$\omega = c_0 k - \frac{c_0 h^2}{6}\delta k^3 + \dots, \qquad \delta = 1 - 3B,$$

where we recall that B is the Bond number measuring the effect of surface tension. To leading order, the waves propagate with the linear long wave phase speed $c_0 = (gh)^{1/2}$. But, after a long time the cumulative effects of weak nonlinearity must be taken into account. When these are balanced against the leading order linear dispersive terms (the $O(k^3h^3)$ terms above), the result is the well-known Korteweg–de Vries (KdV) equation for the wave elevation

$$\zeta_t + c_0\zeta_x + \frac{3c_0}{2h}\zeta\zeta_x + \frac{c_0 h^2}{6}\delta\zeta_{xxx} = 0. \tag{7}$$

This equation was derived by Diederik Korteweg and Hendrik de Vries in a now very famous paper published in 1895 (Korteweg & de Vries, 1895), although in fact Joseph Boussinesq had obtained it earlier in 1877 (Boussinesq, 1877; *See* **Korteweg–de Vries equation**). Relative to the leading order propagation with the speed c_0, the time evolution occurs on the nondimensional scale ε^{-3}, where the wave amplitude ζ/h scales with ε^2 and the wave dispersion kh scales with ε (that is, spatial derivatives scale with ε).

Korteweg and de Vries found a family of traveling-wave solutions, periodic waves described by elliptic functions and commonly called cnoidal waves, and the solitary wave solution

$$\zeta = a\,\mathrm{sech}^2(\gamma(x - Vt)),$$
$$\text{where} \quad V - c_0 = \frac{c_0 a}{2h} = \delta\frac{2c_0h^2}{3}\gamma^2. \tag{8}$$

The solitary wave has a free parameter, for instance, the amplitude a. When $\delta > 0$ (that is, $B < 1/3$), then the amplitude a is always positive and $V > c_0$. Further, as a increases, the wave speed V and the wave number γ also increase. Although the case when $B > \frac{1}{3}$ has little practical application, it is interesting to note that these conclusions are all reversed as then $\delta < 0$, now, $a < 0$ and $V < c_0$. This solitary wave had earlier been obtained directly from the governing equations independently by Boussinesq (1871) and Rayleigh (1876), who were motivated to explain the now very well-known observations and experiments of John Scott Russell (1844). Curiously, it was not until quite recently that it was recognized that the KdV equation is not strictly valid if surface tension is taken into account and $0 < B < \frac{1}{3}$, as there is then a resonance between the solitary wave and very short capillary waves.

After this ground-breaking work of Korteweg and de Vries, interest in solitary water waves declined until the dramatic discovery of the "soliton" by Zabusky and Kruskal in 1965 (*See* **Solitons, a brief history**). Through numerical integrations of the KdV equation, they demonstrated that the solitary wave (8) could be generated from quite general initial conditions, and could survive intact collisions with other solitary waves, leading them to coin the term *soliton*. Their remarkable discovery, followed almost immediately by the theoretical work of Gardner, Greene, Kruskal, and Miura (Gardener et al., 1967) showing that the KdV equation was integrable through an inverse scattering transform, led to many other startling discoveries and marked the birth of the soliton theory as we know it today. The implication for water waves is that the solitary wave is the key component needed to describe the behavior of long, weakly nonlinear waves.

An alternative to the KdV equation is the Benjamin–Bona–Mahony (BBM) equation in which the linear dispersive term $c_0\zeta_{xxx}$ in (7) is replaced by $-\zeta_{xxt}$. It has the same asymptotic validity as the KdV equation and, since it has rather better high wave number properties, is somewhat easier to solve numerically. However, it is not integrable and, consequently, has not attracted the same interest as the KdV equation.

Both the KdV and BBM equations are unidirectional. A two-dimensional version of the KdV equation is the KP equation (Kadomtsev & Petviashvili, 1970; *See* also **Kadomtsev–Petviashvili equation**),

$$\left(\zeta_t + c_0\zeta_x + \frac{3c_0}{2h}\zeta\zeta_x + \frac{c_0h^2}{6}\delta\zeta_{xxx}\right)_x + \frac{c_0}{2}\zeta_{yy} = 0. \tag{9}$$

This equation includes the effects of weak diffraction in the y-direction, in that y-derivatives scale as ε^2 whereas x-derivatives scale as ε. When $\delta > 0$ ($0 \le B < \frac{1}{3}$), this is the KPII equation, and it can be shown that then the solitary wave is stable to transverse disturbances. On the other hand, if $\delta < 0$ ($B > \frac{1}{3}$), this is the KPI equation for which the solitary wave (8) is unstable; instead this equation supports "lump" solitons. Both KPI and KPII are integrable equations. To take account of stronger transverse effects and/or to allow for bi-directional propagation in the x-direction, it is customary to replace the KdV equation with a Boussinesq system of equations. These combine the long wave approximation to the dispersion relation with the leading order nonlinear terms and occur in several asymptotically equivalent forms.

Wave Packets: Nonlinear Schrödinger Equation

The linear theory of water waves predicts that a localized initial state will typically evolve into wave packets, with a dominant wave number $\boldsymbol{k}$ and corresponding frequency ω given by (3), within which each wave phase propagates with the phase speed c, but whose envelope propagates with the group velocity c_g (5, 6). After a long time, the packet tends to disperse around the dominant wave number, which tendency may be opposed by cumulative nonlinear effects. In the absence of surface tension, the outcome for unidirectional waves is described by the nonlinear Schrödinger (NLS) equation, for the wave amplitude $A(x,t)$; that is, the wave elevation is $\zeta = A\exp\mathrm{i}(kx - \omega t) +$ c.c. to leading order,

$$\mathrm{i}(A_t + c_gA_x) + \tfrac{1}{2}\lambda A_{xx} + \mu|A|^2A = 0, \tag{10}$$

and the coefficients are given by

$$\lambda = \frac{\partial^2\omega}{\partial k^2},$$
$$\mu = -\frac{\omega k^2}{16S^4}(8C^2S^2 + 9 - 2T^2) + \frac{\omega}{8C^2S^2}\frac{(2\omega C^2 + kc_g)^2}{gh - c_g^2},$$

where

$$C = \cosh(kh)\,, \quad S = \sinh(kh)\,, \quad T = \tanh(kh).$$

For water waves, the NLS equation was first derived by Zakharov (1968) for the case of deep water, and then by Hasimoto and Ono (1972) for finite depth. An analogous equation can be derived for nonzero surface tension in which the nonlinear coefficient μ will take a different value, but for reasons discussed below, it is not so useful in that case. In deep water ($kh \to \infty$) the coefficient $\mu \to -\omega k^2/2 < 0$. In general, $\mu < 0 (> 0)$ according as $kh > (<) 1.36$).

Like the KdV equation, the NLS equation is integrable with an associated inverse scattering transform, a result first shown by Zakharov and Shabat (1972). There are two cases, the so-called focusing NLS equation when $\lambda\mu > 0$ and the defocusing NLS equation when $\lambda\mu < 0$. For water waves, $\lambda = \partial c_g/\partial k < 0$, and so we have the focusing (defocusing) NLS equation according as $kh > (<) 1.36$. The focusing NLS equation has solitary wave solutions (bright solitons), given by

$$A = a \operatorname{sech}(\gamma(x - c_g t)) \exp(-i\Omega t),$$
$$\text{where } \mu a^2 = \lambda\gamma^2, \; \Omega = -\tfrac{1}{2}\mu a^2. \tag{11}$$

On the other hand, the defocusing NLS equation has no such solitary wave solutions which decay to zero at infinity. Instead, it has solitary waves riding on a nonzero background (dark solitons). A key property of the NLS equations is that plane waves are modulationally unstable (stable) in the focusing (defocusing) case. That is, the NLS equation has the exact plane wave solution,

$$A = A_0 \exp(i\mu|A|^2 t), \tag{12}$$

which is then perturbed with a small-amplitude modulation proportional to $\exp(iKx - \sigma t)$. It is readily found that the growth rate σ is given by

$$\sigma = K^2 \left(\lambda\mu A_0^2 - \frac{\lambda^2 K^2}{4} \right). \tag{13}$$

Thus, in the focusing NLS case when $\lambda\mu > 0$, there is a positive growth rate for modulation wave numbers K such that $K < 2(\lambda/\mu)^{1/2}|A_0|$. On the other hand, σ is purely imaginary for all K in the defocusing case when $\lambda\mu < 0$. The implication for water waves is that plane Stokes waves in deep water ($kh > 1.36$) are unstable. This remarkable result was first discovered by Benjamin and Feir in 1967, by a different theoretical approach, and has since been confirmed in experiments. The maximum growth rate occurs for $K = K_m = (\lambda/\mu)^{1/2}|A_0|$, and the instability is due to the generation of side bands with wave numbers $k \pm K_m$. As the instability grows, the full NLS equation (10) is needed to describe the long-time outcome of the collapse of the uniform plane wave into several soliton wave packets, each described by (11).

When the effects of modulation in the transverse y-direction are taken into account, so that the wave amplitude is now given by $A(x, y, t)$, the NLS equation is replaced by the Benney–Roskes system (Benney & Roskes, 1969), also widely known as the Davey–Stewartson equations (Davey & Stewartson, 1974),

$$i(A_t + c_g A_x) + \tfrac{1}{2}\lambda A_{xx} + \tfrac{1}{2}\delta A_{yy}$$
$$+ \mu|A|^2 A + UA = 0, \tag{14}$$
$$\alpha U_{xx} + U_{yy} + \beta(|A|^2)_{yy} = 0, \tag{15}$$

where the coefficients μ and λ are those defined in (10), while

$$\delta = \frac{c_g}{k}, \; \alpha = 1 - \frac{c_g^2}{gh},$$
$$gh\beta = \frac{\omega}{8C^2S^2}(2\omega C^2 + kc_g)^2.$$

Here, the surface tension has been set to zero. If surface tension effects are included, a similar equation holds but with different values for the coefficients (Djordjevic & Redekopp, 1977). Note that $\lambda\delta < 0$ and $\alpha > 0$, so that the equation for A is hyperbolic, but that for U it is parabolic. The variable U which appears here is a wave-induced mean flow, which tends to zero in the limit of deep-water waves, $kh \to \infty$. This system (14) again has the plane wave solution (12), whose stability can be analyzed in a manner similar to that described above in the context of the NLS equation. The outcome is that now instability can occur for all values of kh and occurs in a band in the $k - l$ plane where k and l are the modulation wave numbers. The instability is purely two-dimensional when $kh < 1.36$, and the band becomes narrower and the growth rate weaker as $kh \to 0$. For more details, see Benney and Roskes (1969) or Mei (1983).

Wave Resonant Interactions

A superposition of weakly nonlinear waves, each of which is given by (1) with a wave number $\boldsymbol{k}_n, n = 1, 2, \ldots, N$, and a corresponding frequency $\omega_n(\boldsymbol{k}), n = 1, 2, \ldots, N$, each satisfying the dispersion relation (3), interact resonantly whenever

$$\boldsymbol{k}_1 + \boldsymbol{k}_2 + \cdots + \boldsymbol{k}_N = 0,$$

and

$$\omega_1 + \omega_2 + \cdots + \omega_N = \Omega.$$

Here, Ω is a detuning term, so that exact resonance is achieved whenever $\Omega = 0$. The most prominent interactions are triad interactions; that is, $N = 3$, followed by the quartet interactions when $N = 4$, and so on. In a resonant interaction, energy is exchanged between the Fourier components in a periodic manner, assuming that dissipation is absent. For instance, the set of equations describing a triad interaction is (Craik, 1985)

$$A_{1t} + \boldsymbol{c}_{g1} \cdot \nabla A_1 = \frac{\partial D}{\partial \omega_1} \mu A_2^* A_3^* \exp(-i\Omega t), \ldots,$$

where $A_n(\boldsymbol{x}, t)$ is the amplitude of the nth mode, $D(\omega, \boldsymbol{k}) = 0$ is the dispersion relation, μ is a coefficient, and the superscript "*" denotes the complex conjugate. Remarkably, these equations are integrable for exact resonance ($\Omega = 0$) (Zakharov & Manakov, 1973).

For water waves in the absence of surface tension, the dispersion relation does not allow for triad interactions. Hence, the dominant resonance is a quartet interaction, as first shown by Phillips (1960). The four wave numbers making up the quartet are two-dimensional (i.e., the y-wave number components are generally not zero), and the allowed wave number vectors can be determined graphically from Phillips' "figure-of-eight" diagram. The interaction equations for a discrete quartet of waves is analogous to that displayed above for a triad interaction. Further, in deep water ($kh \to \infty$), Craig and Worfolk (1995) have shown that in the Birkhoff normal form for these interaction equations, the coefficients vanish for all nongeneric resonant terms and the remaining system is then integrable. However, the same is not the case for quintet interactions.

When considering a continuous spectrum of gravity waves, the resulting evolution of a spectral component is described by Zakharov's integral equation, first derived in the deep water limit by Zakharov (1968). Krasistskii (1994) later employed canonical transformations to obtain a more desirable Hamiltonian form (see also the review by Dias & Kharif, 1999). In this equation, usually truncated at the third order in wave amplitude, the evolution of the spectral component $A(\boldsymbol{k}, t)$, which is the spatial Fourier transform of $\eta(\boldsymbol{x}, t)$, is determined essentially by quartet interactions, since the nonresonant triad interactions are removed by a canonical transformation. This integral evolution equation contains much of the previous weakly nonlinear theory, in that the discrete interaction equations and the NLS equation can be derived from it. It also forms the basis for a statistical description of water waves and can then be used to describe the ocean wave spectrum.

When surface tension is taken into account, then triad interactions are possible (i.e., $N = 3$ in (**??**)). The most well-known example occurs when $\boldsymbol{k}_1 = \boldsymbol{k}_2 = -\boldsymbol{k}_3/2$, and is a second harmonic resonance in that then $\omega_1 = \omega_2 = -\omega_3/2$. It was first noted by Wilton (1915) and leads to a phenomenon known as Wilton's ripples. In general, the existence of triad resonances implies that capillary-gravity waves undergo wave–wave interactions on a faster time scale than pure gravity waves.

ROGER GRIMSHAW

See also **Dispersion relations; Group velocity; Korteweg–de Vries equation; Modulated waves; Nonlinear Schrödinger equations; Solitons, a brief history; Wave packets, linear and nonlinear**

Further Reading

Benjamin, T.B. & Feir, J.E. 1967. The disintegration of wave trains on deep water. *Journal of Fluid Mechanics*, 27: 417–430

Benney, D.J. & Roskes, G. 1969. Wave instabilities. *Studies in Applied Mathematics*, 48: 377–385

Boussinesq, M.J. 1871. Theórie de l'intumescence liquid appellée onde solitaire ou de translation, se propageant dans un canal rectangulaire. *Comptes Rendus de l'Académie des Sciences (Paris)*, 72: 755–759

Boussinesq, M.J. 1877. Essai sur la theorie des eaux courantes, *Memoires presentees par diverse savants a l'Academie des Sciences Inst. France (Series 2)*, 23: 1–680

Craig, W. & Worfolk, P.A. 1995. An integrable normal form for water waves in infinite depth. *Physia* D, 84: 513–531

Craik, A.D.D. 1985. *Wave Interactions and Fluid Flows*, Cambridge and New York: Cambridge University Press

Davey, A. & Stewartson, K. 1974. On three-dimensional packets of surface waves. *Proceedings of the Royal Society of London* A, 338: 101–110

Dias, F. & Kharif, C. 1999. Nonlinear gravity and capillary-gravity waves. *Annual Reviews of Fluid Mechanics*, 31: 301–346

Djordjevic, V.D. & Redekopp, L.G. 1977. On two-dimensional packets of capillary-gravity waves. *Journal of Fluid Mechanics*, 79: 703–714

Gardner, C.S., Greene, J.M., Kruskal, M.D. & Miura, R.M. 1967. Method for solving the Korteweg–de Vries equation. *Physical Review Letters*, 19: 1095–1097.

Hasimoto, H. & Ono, H. 1972. Nonlinear modulation of gravity waves. *Journal of the Physical Society of Japan*, 33: 805–811

Kadomtsev, B.B. & Petviashvili, V.I. 1970. On the stability of solitary waves in weakly dispersive media. *Soviet Physics Doklady*, 15: 539–541

Korteweg, D.J. & de Vries, H. 1895. On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philosophical Magazine*, 39: 422–443

Krasitskii, V.P. 1994. On reduced equations in the Hamiltonian theory of weakly nonlinear surface waves. *Journal of Fluid Mechanics*, 272: 1–30

Lamb, H. 1932. *Hydrodynamics*, Cambridge and New York: Cambridge University Press; 6th edition, Cambridge and New York, Cambridge University Press, 1993

Lighthill, M.J. 1978. *Waves in Fluids*, Cambridge and New York: Cambridge University Press

Mei, C.C. 1983. *The Applied Dynamics of Ocean Surface Waves*, New York: Wiley

Phillips, O.M. 1960. On the dynamics of unsteady gravity waves of finite amplitude. Part 1. The elementary interactions. *Journal of Fluid Mechanics*, 9: 193–217

Rayleigh, Lord (Strutt, W.J.) 1876. On waves. *Philosophical Magazine*, 1: 257–279

Russell, J.S. 1844. Report on Waves, *14th meeting of the British Association for the Advancement of Science*, London: BAAS, 311–390

Schwarz, L.W. & Fenton, J.D. 1982. Strongly nonlinear waves. *Annual Reviews of Fluid Mechanics*, 14: 39–60

Wilton, J.R. 1915. On ripples. *Philosophical magazine*, 29: 688–700

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

Zabusky, N.J. & Kruskal, M.D. 1965. Interactions of solitons in a collisionless plasma and the recurrence of initial states. *Physical Review Letters*, 15: 240–243

Zakharov, V.E. 1968. Stability of periodic waves of finite amplitude on the surface of a deep fluid. *Journal of Applied Mechanics and Technical Physics*, 2: 190–194

Zakharov, V.E. & Manakov, S.V. 1973. Resonant interactions of wave packets in nonlinear media. *Soviet Physics, JETP*, 18: 243–247

Zakharov, V.E. & Shabat, A.B. 1972. Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Soviet Physics, JETP*, 34: 62–69

WAVE OF TRANSLATION

Translational invariance is one of the fundamental symmetries of continuum nonlinear partial differential equations (PDEs). Such a symmetry is present when the equation of interest remains unchanged under the transformation $x \to x + \varepsilon$. Typically, partial differential equations that contain only derivatives with respect to the spatial variable x (but no explicit dependence on x) possess this symmetry.

The topic of symmetries and their role in bifurcation theory is a very large subject, a detailed discussion of which can be found, for example in Golubitsky & Schaeffer (1995) and Golubitsky (1988). Here, our scope is much narrower in giving a perspective of translational symmetry and its role in the existence, bifurcation, or absence of traveling wave solutions in some Hamiltonian and dissipative classes of PDEs.

To make our discussion of this subject more definitive, we will introduce a rather general class of dissipative or Hamiltonian PDEs of the form

$$\{u_t, u_{tt}\} = u_{xx} + f(u). \tag{1}$$

The paradigm of Equation (1) is dissipative (that is, of the reaction-diffusion type), if a one-time derivative of the field u is used in the left-hand side. On the other hand, the model is Hamiltonian if the second time derivative is used on the left-hand side. While the model is written as a single component model in a one-dimensional setting, it can be easily generalized in multiple components/dimensions. In the former case, u becomes a vector, while in the latter, the spatial second partial derivative is substituted by the Laplacian operator. We should note in passing that while here we consider the Hamiltonian and dissipative cases, there are many models which lie between the conservative and diffusion limits. Numerous examples of this type can be found in Kivshar & Malomed (1989).

It can be immediately observed that the translational symmetry mentioned above is present in the model of Equation (1). A straightforward example of its absence would be the case of a "reaction term" $f(u, x)$, i.e., one explicitly dependent on x. Let us now explore the implications of this symmetry for the static problem of Equation (1) and then for the corresponding dynamic problem.

For the static problem (i.e., for solutions $u = g(x)$ that satisfy $u_{xx} + f(u) = 0$), the presence of the invariant direction signifies that the solution can be translated along the group orbit of this invariance. Otherwise stated, there is a one parameter infinity of solutions (alternatively, a degeneracy of solutions) due to the available freedom to select solutions along the symmetry direction. In practical terms, this means that if $u = g(x)$ is a solution, then $u = g(x - x_0)$ is also a solution and arbitrary translations of the solution satisfy the original equation. Furthermore, this feature bears consequences on the linear stability around the solution. In particular, looking for the linear stability of the solution $u = g(x)$, we use the linearization $u = g(x) + \varepsilon \exp(\lambda t) w(x)$, which to $O(\varepsilon)$, yields the eigenvalue problem:

$$\{\lambda, \lambda^2\} w = w_{xx} + f'(g(x)) w. \tag{2}$$

Given that $g(x)$ satisfies the $u_{xx} + f(u) = 0$, differentiation of the latter immediately yields that (due to the absence of the explicit spatial dependence and hence due to the symmetry) $w = u_x \equiv g'(x)$ satisfies Equation (2), with $\lambda = 0$ (for the dissipative system) or $\lambda^2 = 0$ (for the Hamiltonian system). Hence, the existence of the symmetry generates a single (for the dissipative) or a pair (for the Hamiltonian) of eigenvalues at the origin of the spectral plane (i.e., with $\lambda = 0$). This is the neutral eigendirection connected with the symmetry that is often referred to as a Goldstone mode.

Furthermore, in the Hamiltonian version of the system, the presence of such eigenmodes, and of their corresponding symmetries, is intimately connected with conservation laws (through Noether's theorem; see for example the discussion in Arnol'd (1989); Sulem & Sulem (1999)). The invariance with respect to translations is directly related with the conservation of linear momentum, which in the case of Equation (1) is of the form

$$P = \int_{-\infty}^{\infty} u_t u_x \, \mathrm{d}x. \tag{3}$$

We now turn to the dynamic consequences of the symmetry. The translational symmetry is directly related to the traveling of solutions. In the case of Hamiltonian (especially continuum) systems, there may be extra symmetries that may allow the construction of traveling solutions from stationary ones. In the case of Equation (1), such a symmetry is the Lorentz invariance $[x \to x' = \gamma(x - vt),\ t \to t' = \gamma(t - vx/c^2),\ \gamma = 1/\sqrt{1 - v^2/c^2}]$, which allows one to boost the solutions to any "subsonic" speed. In other cases, such as the one of the nonlinear Schrödinger equation, the corresponding symmetry, is the Galilean invariance. Hence, in Hamiltonian systems, due to the additional symmetry, traveling and standing solutions are often, in some sense, equivalent. On the other hand, for dissipative systems, such equivalence is typically absent. In the latter case, we look for traveling wave solutions in the form $u = u(\xi)$, where $\xi = x - ct$ is the traveling wave variable. This transformation leads to the so-called traveling wave frame (TWF) equation of motion (i.e., we travel together with the solution and, hence, observe it as a steady one) of the form

$$u_t = u_{\xi\xi} + cu_\xi + f(u). \tag{4}$$

By solving the steady-state problem of Equation (4) (notice: an ordinary differential equation (ODE) problem to find special solutions of the PDE), we can identify the traveling-wave solutions of Equation (1). In dissipative systems, the absence of additional symmetries typically allows for isolated solutions of the TWF ODE rather than monoparametric families of such solutions. Notice that here we have in mind fixed model parameter values (but not fixed initial condition parameters, such as, for example, energy). Furthermore, we do not discuss the mechanisms (in dissipative systems) of selection of a given speed (which is related to issues of stability). Such examples and a detailed discussion can be found in Xin (2000). A typical example is the bistable nonlinearity where $f(u) = 2u(u-1)(\mu - u)$, for which a front solution of the form

$$u(x,t) = \frac{1}{2}\left[1 - \tanh\left(\frac{x - x_c(t)}{2}\right)\right], \tag{5}$$

exists, where $x_c = x_c^{\text{in}} + ct$ and x_c^{in} is the original position of the center, while the speed c is connected to the parameter μ through

$$c = 1 - 2\mu. \tag{6}$$

In view of the above comments, in energy conserving systems, typically standing and traveling solutions coexist, while in dissipative systems one can (locally) have solutions either of the former or of the latter type. In fact, as parameters are varied in dissipative settings, traveling solutions may bifurcate from standing ones, through the so-called drift pitchfork bifurcation (Kness et al., 1992, see also Malomed & Tribelsky (1984) and Coullet & Iooss, 1990). It is worth noting that recently, a template-based technique has been proposed that dynamically factors out translational invariance (and other continuous symmetries) (Rowley & Marsden, 2000; Rowley et al., 2003).

Discrete Systems and Symmetry Breaking

An interesting variation of the above presented scenario occurs in (spatially) discrete systems. In this case, generic discretizations of the original problem of Equation (1) no longer preserve the symmetry with respect to continuum translations. Instead, only an integer shift invariance persists in this discrete limit of the equation:

$$\{\dot{u}_n, \ddot{u}_n\} = \Delta_2 u_n + f(u_n), \tag{7}$$

where $\Delta_2 u_n = (u_{n+1} + u_{n-1} - 2u_n)/h^2$ is the discrete Laplacian for a lattice of spacing h. In this case, from the informative Taylor expansion of the form

$$\Delta_2 u_n = \sum \frac{2h^{2j-2}}{(2j)!}\frac{\mathrm{d}^{2j}u}{\mathrm{d}x^{2j}}, \tag{8}$$

we deduce the following conclusions:

- discreteness is a singular higher-order derivative perturbation to the continuum limit;
- to all orders in this asymptotic expansion, the right-hand side is translationally invariant. Hence, the breaking of the translational symmetry can only occur beyond all algebraic orders and thus has to be an exponentially small effect.

A manifestation of this exponentially small symmetry breaking effect is given by the fact that there are now two (lattice shift invariant) stationary wave states in the lattice context. One of these solutions is centered on a lattice site and one is centered between two consecutive lattice sites (instead of a single translationally invariant steady state in the continuum limit). One of these solutions is stable and one is unstable. The energy difference between the two (which mirrors the amount of symmetry breaking and hence should be exponentially small, that is, $\Delta E \sim \exp(-\pi^2/h)$) is the celebrated Peierls–Nabarro barrier. For a detailed discussion of these issues, see, for example, the review by Kevrekidis et al. (2001).

From the above, we can infer that the generic effect of discreteness in breaking translational invariance is to generate an exponentially small (in the natural lattice spacing parameter) shift periodic (in fact approximately trigonometric) potential. The center of mass of the waves propagates inside the Peierls–Nabarro barrier. There is considerable interest in shearing this potential with an external (constant) field.

A constant external force F introduces a term in the potential energy $\sim Fu$, generating a washboard potential for the motion of the wave. If the external field becomes sufficiently strong, then one can infer that a saddle-node (in fact, infinite period; the so-called SNIPER) bifurcation will occur in which the stationary states (the maxima and minima of the potential, that is, the saddles and nodes) will disappear and traveling waves will arise. Scaling analysis predicts and numerical results verify (Kaldko et al., 2000; Keverkidis et al., 2001; Carpio & Bonilla, 2001) that the relevant bifurcation will yield waves of speed

$$c \sim (F - F_{\mathrm{c}})^{1/2}, \tag{9}$$

where F is the constant external field and F_{c} is its critical value. The energy landscapes for the cases of $F = 0$, $F < F_{\mathrm{c}}$, and $F > F_{\mathrm{c}}$ are shown in Figure 1. This scenario happens in the dissipative case, but one can also analyze the Hamiltonian case in the same manner.

Finally, it should be noted that while the above scenario will be the generically relevant one, there

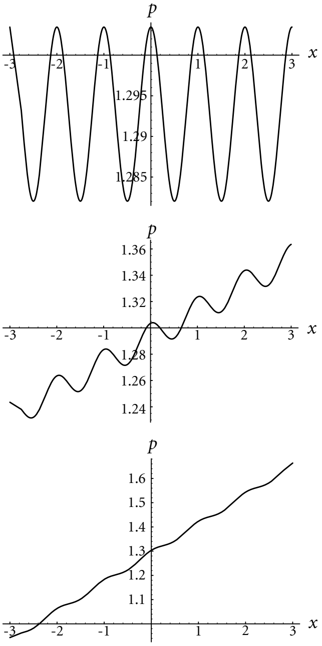

Figure 1. The (potential) energy (P) landscape for generic DDE models in the case of $F = 0$ (top panel), $F < F_c$ (middle panel) and $F > F_c$ (bottom panel), as a function of the position of the kink center.

are exceptions to the rule of absence of a continuum-like symmetry in the lattice setting. Let us consider, for example, the discretization of the Hamiltonian PDE:

$$\ddot{u}_n = \Delta_2 u_n + \frac{F(u_{n+1}) - F(u_{n-1})}{u_{n+1} - u_{n-1}}. \tag{10}$$

It can be seen that the equation of motion has a conservation law of the form

$$P = \sum \dot{u}_n (u_{n+1} - u_{n-1}) \tag{11}$$

which is the discrete analog of Equation (3). In this case, the discrete equation preserves a "ghost" of the continuum symmetry and maintains the multiplicity of Goldstone modes of the continuum problem (Kevrekidis, 2003).

In conclusion, translational invariance is a symmetry that plays a significant role in the context of both partial-differential as well as differential-difference (i.e., lattice) equations. In the former, it is typically present (unless an explicit spatial dependence occurs) and has both static as well as kinematic consequences on the solutions of the problem and their stability. In the discrete setting, the symmetry is generically absent and its absence plays a critical role in diversifying kinematic and dynamic phenomena on the lattice (see, for example, Kevrekidis, et al. (2001)) from their continuum siblings. However, it is possible to construct discretizations that respect a discrete conservation law reminiscent of the one imposed by the continuum symmetry group. In the nongeneric case, the lattice dynamics can be significantly closer to their continuum counterparts.

P.G. KEVREKIDIS AND I.G. KEVREKIDIS

See also **Partial differential equations, nonlinear; Peierls barrier; Sine-Gordon equation; Zeldovich–Frank-Kamenetsky equation**

Further Reading

Arnol'd, V.I. 1989. *Mathematical Methods in Classical Mechanics*, New York: Springer

Carpio, A. & Bonilla, L.L. 2001. Wave front depinning transition in discrete one-dimensional reaction–diffusion systems. *Physical Review Letters*, 86: 6034–6037

Coullet, P. & Iooss, G. 1990. Instabilities of one-dimensional cellular patterns. *Physical Review Letters*, 64: 866–869

Dodd, R.K., Eilbeck, J.C., Gibbon, J.D. & Morris, H.C. 1982. *Solitons and Nonlinear Wave Equations*, London and New York: Academic Press

Golubitsky, M. & Schaeffer, D.G. 1985. *Singularities and Groups in Bifurcation Theory*, vol. I, New York: Springer

Golubitsky, M., Stewart, I.N. & Schaeffer, D.G. 1988. *Singularities and Groups in Bifurcation Theory*, vol. II, New York: Springer

Kevrekidis, P.G. 2003. On a class of discretizations of Hamiltonian nonlinear partial differential equations. *Physica* D, 183: 68–86

Kevrekidis, P.G., Kevrekidis, I.G. & Bishop, A.R. 2001a. Propagation failure, universal scalings and Goldstone modes. *Physics Letters* A, 279: 361–369

Kevrekidis, P.G., Rasmussen, K.Ø. & Bishop, A.R. 2001b. The discrete nonlinear Schrodinger equation: a survey of recent results. *International Journal of Modern Physics* B, 15: 2833–2900

Kivshar, Yu.S. & Malomed, B.A. 1989. Dynamics of solitons in nearly integrable systems. *Reviews of Modern Physics*, 61: 763–915

Kladko, K., Mitkov, I. & Bishop, A.R. 2000. Universal scaling of wave propagation failure in arrays of coupled nonlinear cells. *Physical Review Letters*, 84: 4505–4508

Kness, M., Tuckerman, L.S. & Barkley, D. 1992. Symmetry-breaking bifurcations in one-dimensional excitable media. *Physical Review* A, 46: 5054–5062

Malomed, B.A. & Tribelsky, M.I. 1984. Bifurcations in distributed kinetic systems with aperiodic instability. *Physica* D, 14: 67–87

Rowley, C.W. & Marsden, J.E. 2000. Reconstruction equations and the Karhunen-Loeve expansion for systems with symmetry. *Physica* D, 142: 1–19

Rowley, C.W., Kevrekidis, I.G., Marsden, J.E. & Lust, K. 2003. Reduction and reconstruction for self-similar dynamical systems. *Nonlinearity*, 16: 1257–1275

Sulem, C. & Sulem, P.L. 1999. *The Nonlinear Schrödinger Equation*, New York: Springer

Xin, J. 2000. Front propagation in heterogeneous media. *SIAM Review*, 42: 161–230

WAVE PACKETS, LINEAR AND NONLINEAR

Because linear wave systems have elementary solutions of the form $e^{i(kx-\omega t)}$, it is often convenient to write the general solution of an initial value problem as an integral of Fourier components. Thus,

$$u(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(k)e^{i(kx-\omega t)}\,dk, \qquad (1)$$

where $F(k)$ is the Fourier transform of $u(x,0)$. Initial conditions thus determine the Fourier transform, each component of which evolves independently with frequency ω related to wave number k through the dispersion relation

$$\omega = \omega(k)\,. \qquad (2)$$

Unless $\omega = k$, different components in Equation (1) travel at different speeds (ω/k), and an initially localized wave spreads out or "disperses," hence the name.

A wave packet is a special form of Equation (1) with the largest Fourier components lying close to some wave number (k_0) and the corresponding frequency (ω_0). In other words, the initial conditions $u(x, 0)$ are selected so that $F(k)$ has its maximum value at $k = k_0$, falling rapidly with increasing $|k - k_0|$. This suggests writing the dispersion relation as a power series about k_0. With the notation

$$\omega = \omega_0 + b_1(k-k_0) + b_2(k-k_0)^2 \qquad (3)$$

(which assumes that the system has no higher than second derivatives with respect to x), Equation (1) becomes

$$u(x,t) = e^{i(k_0x-\omega_0t)}\frac{1}{2\pi} \times \int_{-\infty}^{\infty} F(k)e^{i[(k-k_0)x-b_1(k-k_0)t-b_2(k-k_0)^2t]}dk, \qquad (4)$$

where the factor $e^{i(k_0x-\omega_0t)}$ is a carrier wave with a velocity $v_c = \omega_0/k_0$, shown in Figure 1(a). Riding over (or multiplying) the carrier is an envelope wave

$$\phi(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\kappa+k_0)e^{i(\kappa x-b_1\kappa t-b_2\kappa^2t)}\,d\kappa, \quad (5)$$

where the variable of integration has been changed from k to $\kappa \equiv k - k_0$.

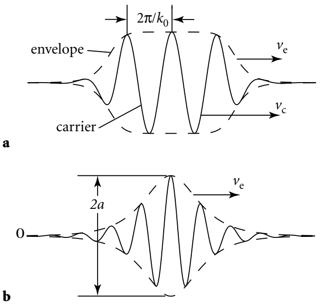

Figure 1. (a) The real part of a linear wave packet, showing the envelope (dashed lines) and the carrier (full line), as in Equation (4). (b) The real part of a soliton solution of Equation (7).

Taking the time derivative of Equation (5), one finds

$$\begin{aligned}\frac{\partial\phi}{\partial t} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} -i(b_1\kappa + b_2\kappa^2)F(\kappa+k_0) \\ &\quad\times e^{i(\kappa x-b_1\kappa t-b_2\kappa^2t)}\,d\kappa \\ &= -b_1\frac{\partial\phi}{\partial x} + ib_2\frac{\partial^2\phi}{\partial x^2},\end{aligned}$$

which can be written as

$$i\left(\frac{\partial\phi}{\partial t} + b_1\frac{\partial\phi}{\partial x}\right) + b_2\frac{\partial^2\phi}{\partial x^2} = 0. \qquad (6)$$

Equation (6) is a partial differential equation that governs time evolution of the envelope for a linear wave packet solution of a second-order equation. Assuming b_2 is not too large (weak dispersion), the envelope moves with the velocity $v_e \doteq b_1 = d\omega/dk|_{k=k_0}$, as in Figure 1(a).

Up to this point, the discussion has remained within the realm of linear theory, but now assume that nonlinear effects alter Equation (6) to

$$i\left(\frac{\partial\phi}{\partial t} + b_1\frac{\partial\phi}{\partial x}\right) + b_2\frac{\partial^2\phi}{\partial x^2} + \alpha|\phi|^2\phi = 0, \qquad (7)$$

with the nonlinear (amplitude-dependent) dispersion relation $\omega = b_1\kappa + b_2\kappa^2 - \alpha|\phi|^2$. Equation (7) is the nonlinear Schrödinger (NLS) equation, for which the following comments are relevant:

- If ϕ is assumed independent of x, Equation (7) has the plane wave solution

  $$\phi = \phi_0 e^{i\alpha|\phi_0|^2t}, \qquad (8)$$

  which may or may not be stable.
- With $\alpha b_2 < 0$, this plane wave is stable (Whitham, 1974). If $\alpha b_2 > 0$, on the other hand, the plane wave experiences Benjamin–Feir instability, out of which emerge stable NLS solitons (Benjamin & Feir, 1967; Ostrovsky, 1967).

- The term $b_2(\partial^2\phi/\partial x^2)$ introduces wave dispersion into the problem at the lowest order of approximation. Similarly, the term $\alpha|\phi|^2\phi$ introduces nonlinearity at the lowest order of approximation. Thus, the NLS equation is generic, arising whenever one wishes to consider lowest order effects of dispersion and nonlinearity on a wave packet, including nonlinear optics (Kelley, 1965), deep water waves (Benney & Newell, 1967), and acoustics (Ostrovsky & Potapov, 1999).
- Unstable NLS wave packets decay into one or more solitons. Choosing $b_1 = 0$, $b_2 = 1$, and $\alpha = 2$ in Equation (7), for example, a family of NLS solitons is (Zakharov & Shabat, 1972)

$$u(x,t) = a \exp\left[\mathrm{i}\frac{v_{\mathrm{e}}}{2}x + \mathrm{i}\left(a^2 - \frac{v_{\mathrm{e}}^2}{4}\right)t\right] \times \operatorname{sech}\left[a(x - v_{\mathrm{e}}t - x_0)\right], \qquad (9)$$

one of which is sketched in Figure 1(b).

Beyond the superficial similarities between Figures 1(a) and (b), the differences are profound. In the linear wave packet of Figure 1(a), the shape of the envelope is determined by initial conditions and their subsequent time evolution, as in Equation (6). In the NLS soliton of Figure 1(b), on the other hand, the envelope shape is determined through a dynamic balance between the influences of dispersion and nonlinearity, as expressed by the last two terms of Equation (7).

ALWYN SCOTT

See also **Dispersion relations; Modulated waves; Nonlinear Schrödinger equations; Wave stability and instability**

Further Reading

Benjamin, T.B. & Feir, J.E. 1967. The disintegration of wave trains in deep water. *Journal of Fluid Mechanics*, 27: 417–430

Benney, D.J. & Newell, A.C. 1967. The propagation of nonlinear wave envelopes. *Journal of Mathematical Physics*, 46: 133–139

Kelley, P.L. 1965. Self-focusing of optic beams. *Physical Review Letters*, 15: 1005–1008

Ostrovsky, L.A. 1967. Propagation of wave packets and space-time self-focusing in a nonlinear medium. *Soviet Physics, JETP*, 24: 797–800

Ostrovsky, L.A. & Potapov, A.I. 1999. *Modulated Waves: Theory and Applications*, Baltimore: Johns Hopkins University Press

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

Zakharov, V.E. & Shabat, A.B. 1972. Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Soviet Physics, JETP*, 34: 62–69

WAVE PROPAGATION IN DISORDERED MEDIA

Although nonlinear partial differential equations with constant coefficients well describe main features of numerous nonlinear systems, understanding of many natural phenomena or experimental data requires taking into account the imperfectness of media which often have random character of two main types: time-dependent fluctuations and random spatial inhomogeneities (both may appear simultaneously). A thermal bath and quantum laser fluctuations are examples of the first type while impurities in crystal lattices, irregular variations of dielectric permittivity, and imperfectness of optical fibers are of the second type. Modeling of such phenomena requires nonlinear evolution equations with stochastic terms.

If a physical process is described by a field $u(t, \boldsymbol{r})$ that in an idealized situation is governed by a nonlinear equation $N[u] = 0$, then including irregularities (fluctuations) of the medium in the consideration will result in an equation of the type $N[u] = \varepsilon(t, \boldsymbol{r})R[u]$, where $\varepsilon(t, \boldsymbol{r})$ is a random field and a (possibly nonlinear) operator $R[u]$ depends on physical nature of randomness. Such problems are difficult and still not well understood. The main difficulty arises because nonlinearity invalidates powerful methods of the linear theory (such as the Fourier transform) which allow one to relate stationary and nonstationary problems or introduce a Gaussian approach, allowing decoupling of high-order moments.

From the physical point of view, we can distinguish several important factors that must be taken into account, including generation of high harmonics resulting in nontrivial changes of the field statistics, creation of rather stable localized excitations (which have no analogy in the linear theory), and the multistability phenomenon.

Let us illustrate these issues with two examples: the nonlinear Schrödinger equation for a complex field $u(t, x)$:

$$\mathrm{i}\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + 2|u|^2 u = \varepsilon(t, x)R[u] \qquad (1)$$

and the nonlinear Klein–Gordon equation for a real scalar field $u(t, x)$

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} + u - u^3 = \varepsilon(t, x)R[u]. \qquad (2)$$

These models represent some essential differences. First, at $\varepsilon = 0$ Equations (1) and (2) are, respectively, integrable and nonintegrable. Second, the former model allows a solution in a form of a monochromatic wave and hence admits formulation of a stationary problem while harmonic generation is an indispensable property of the second model. Finally, their solitary wave solutions at $\varepsilon(t, x) = 0$ are of different types: a dynamical soliton (in the restricted mathematical sense) in case (1) and a topological kink in case (2).

The random field $u(t, x)$, being a functional of $\varepsilon(t, x)$, is completely determined by a set of all $(n + m)$-

order moments

$$\langle u(t_1, x_1) \cdots u(t_n, x_n)\bar{u}(t_{n+1}, x_{n+1}) \cdots \bar{u}(t_{n+m}, x_{n+m})\rangle, \qquad (3)$$

where averaging is designated by the angular brackets and is provided over all realizations of $\varepsilon(t, x)$, when they are known. Because a Gaussian approximation is not applicable in nonlinear theory, the generic situation is that an $(n + m)$-order momentum cannot be expressed only through lower-order correlation functions. Moreover, temporal evolution of the mean field may drastically differ from the evolution of the field itself.

This is illustrated by Equation (1) with $\varepsilon(t, x)R[u] \equiv -f(t)xu(t, x)$, $f(t)$ being a Gaussian random process with

$$\langle f(t)\rangle = 0 \quad \text{and} \quad \langle f(t)f(t')\rangle = \sigma^2\delta(t - t'), \qquad (4)$$

where σ is the dispersion of the fluctuations (Besieris, 1980). This equation is exactly solvable with the one-soliton solution

$$u(t, x) = \eta \frac{\exp\left\{\mathrm{i}\left[\mu(t)x + \eta^2 t - \beta(t)\right]\right\}}{\cosh\{\eta[x - 2\xi(t)]\}}, \qquad (5)$$

where $\dot{\mu}(t) = f(t)$, $\mu(t) = \dot{\xi}(t)$, and $\dot{\beta} = \mu^2$ (a dot indicates a time derivative). Being interested in the evolution of the soliton intensity, which depends on $\xi(t)$ only, one can calculate the distribution

$$\begin{aligned} P(\xi, t) &= \int_{-\infty}^{\infty} \langle \delta(v - \dot{\xi}(t))\delta(\xi - \xi(t))\rangle \,\mathrm{d}v \\ &= \sqrt{\frac{3}{2\sigma^2 t^3}} \exp\left(-\frac{3\xi^2}{2\sigma^2 t^3}\right). \end{aligned} \qquad (6)$$

Although the solution undergoes Brownian motion without any distortion, it follows from Equations (5) and (6) that its mean intensity is described by the Gaussian asymptotics

$$\langle u(t, x)\bar{u}(t, x)\rangle \approx \eta\sqrt{\frac{3}{2Dt^3}} \exp\left(-\frac{3x^2}{8\sigma^2 t^3}\right)$$
$$\text{at} \quad t \gg \frac{1}{(\sigma\eta)^{2/3}}.$$

In a general situation when exact solutions are not available, the problem becomes dependent on its statement and on the physical characteristics to be determined. The main statements of the problem are listed below.

Stationary Wave Scattering

This is a generalization of the problem of a wave transmission through a random slab to the nonlinear case. In the linear theory, the wave intensity decays exponentially with the slab width (Anderson localization). In the nonlinear case, one considers a monochromatic solution of Equation (1): $u(t, x) = (k/\sqrt{2}) \exp(-\mathrm{i}k^2 t)\phi(x)$, where $\varepsilon(t, x) \equiv k^2\varepsilon_0(x)$, $\varepsilon_0(x)$ being a random function in the interval $0 < x < L$ and zero outside this interval, and $R[u] \equiv u(t, x)$. Then $\phi(x)$ satisfies the equation

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} + k^2(1 + \varepsilon_0(x) + |\phi|^2)\phi = 0. \qquad (7)$$

In order to describe an incidence of a plane wave on the random layer from the right, one imposes

$$\phi(x) = \begin{cases} a\left[\mathrm{e}^{\mathrm{i}k(L-x)} + R(L, w)\mathrm{e}^{\mathrm{i}k(x-L)}\right], & x \geq L, \\ T(L, w)\mathrm{e}^{-\mathrm{i}kx}, & x \leq 0 \end{cases} \qquad (8)$$

and the conditions of the continuity of the field $\phi(x)$ and of its derivative $\mathrm{d}\phi(x)/\mathrm{d}x$ at the boundaries of the layer, $x = 0$ and $x = L$ (see Figure 1).

The coefficients a, $R(L, w)$ and $T(L, w)$ are amplitudes of the incident wave, reflection and transmission coefficients. The last two quantities depend on the slab width L and on the intensity of the incident wave $w = a^2$. Using the imbedding method (Klyatskin, 1988), one can obtain a nonlinear partial differential equation for the reflection coefficient as a function of L and w, which (upon applying the method of characteristics) results in a system (Doucot & Rammal, 1987)

$$\frac{\mathrm{d}R_1}{\mathrm{d}L} = -2kR_2 - kR_2(1+R_1)(\varepsilon(L)+w|1+R|^2), \qquad (9)$$

$$\begin{aligned} \frac{\mathrm{d}R_2}{\mathrm{d}L} &= 2kR_1 + \frac{k}{2}\left[(1 + R_1)^2 - R_2^2\right] \\ &\quad \times(\varepsilon(L) + w|1 + R|^2), \end{aligned} \qquad (10)$$

where $R_1(L) = \Re(R(L, w))$, $R_2(L) = \Im(R(L, w))$,

$$w(L) = \frac{w(0)}{1 - |R(L, w)|}, \qquad (11)$$

and the initial conditions are $R_1(0) = R_2(0) = 0$. Formula (11) allows us to understand two essentially different statements of the scattering problem. Indeed, (9) and (10) are dynamical equations for the characteristics that pass either through the point $(L = 0, w(0) = w_0)$, if the output intensity of the wave w_0, is given, or through the point $(L, w(L))$ if the input intensity is given (called fixed input and fixed output problems). In the former case, while increasing L, one follows the characteristic starting with its initial position. To solve a fixed output problem, one must determine a characteristic or characteristics which cross the point $(L, w(L))$ (having different starting points). As there may be more than one characteristic, multistability occurs (for another way of understanding the multistability phenomenon, see Knapp et al. (1991)).

Consider the fixed output problem (w_0 is given) in the case of weak Gaussian fluctuations, when

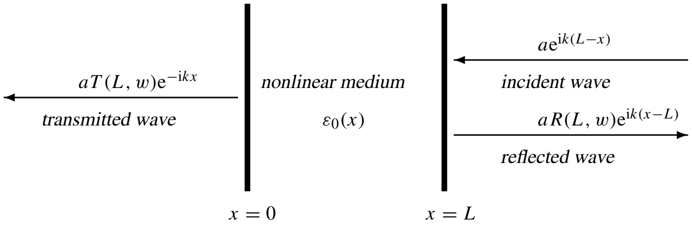

Figure 1. To the statement of the stationary scattering problem.

$\langle\varepsilon_0(x)\rangle = 0$, $\langle\varepsilon_0(x)\varepsilon_0(x')\rangle = D\delta(x - x')$, and $L_{\text{loc}} \gg \lambda, L_{\text{nl}}$ where $L_{\text{loc}} = 1/(Dk^2)$ is the Anderson localization length of the underlying linear system, $\lambda = 2\pi/k$ is the incident wavelength, and $L_{\text{nl}} = 1/(w_0 k)$ is the effective nonlinear length characterizing change of the reflection coefficient due to the nonlinearity. Starting with the region of a relatively strong nonlinearity, $L_{\text{nl}} \ll \lambda$, one can define a period of motion $L_{\text{p}} = \int_0^{2\pi}(\mathrm{d}\theta/\mathrm{d}L)^{-1}\mathrm{d}\theta \approx L_{\text{nl}}$ (here $\theta = \arg R$) and study weak drift of trajectories of "dynamical system" (9), (10) due to the random perturbations over one period L_{p} (providing averaging over θ). The main result can be formulated as $\langle 1/T^2\rangle = O(L^2)$ and decay of the fluctuations of $\langle \ln T^2\rangle$ at $L \to \infty$ (Doucot & Rammal, 1987). Similar results, obtained for a step-like random function $\varepsilon_0(x)$ (Fröhlich et al., 1986). Thus, compared with the linear theory, the nonlinearity does not change the property of $\ln T^2$ to be a self-averaging value, but results in change of the decay law for the decay of the transmission coefficient from the exponential one to a power one. However, when the intensity of the wave becomes small enough and the effect of back-scattering becomes dominating, one can recover the exponential decay law $T \sim \exp(-L/L_{\text{loc}})$. The transition between two regions with the different laws of the decay of the reflection coefficient happens at $L_0 \sim L_{\text{loc}} \ln(1/w_0)$: the law is power at $L > L_0$ and is exponential at $L < L_0$.

Interaction of a Wave Packet with a Random Layer

As this case cannot be reduced to the study of a stationary problem, the statement of the problem and physical characteristics describing wave scattering must be redefined (compared with the previous case of stationary scattering). Consider the Cauchy problem with a given initial field distribution that decays sufficiently rapidly at infinity. The main task can be formulated as a description of the evolution of the wave-packet characteristics during its propagation. Such a statement acquires a special meaning in systems that possess stable solitary wave solutions in the unperturbed ($\varepsilon = 0$) limit and in which the randomness is weak enough. For relatively large temporal intervals, the nonlinear field can be represented in the form $u(t, x) \approx u_{\text{ad}}(t, x) + u_{\text{rad}}(t, x)$, where $u_{\text{ab}}(t, x)$ is a function having the same form as the unperturbed solitary wave but now with slowly varying parameters (adiabatic approximation) and $u_{\text{rad}}(t, x)$ is a small component compared with $u_{\text{ab}}(t, x)$ describing deformation of the soliton shape and radiative losses.

An advantage of the adiabatic approximation is that it reduces a stochastic partial differential equation to a system of ordinary differential equations for the soliton parameters. For Equation (1), the adiabatic approximation corresponds to the substitution (5) with time-dependent coefficients $\eta(t)$, $\xi(t)$, $\mu(t)$, and $\beta(t)$. If $R[u] \equiv u(t, x)$, the standard perturbation theory for solitons shows that (Karpman, 1979)

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = \eta \int_{-\infty}^{\infty} \frac{\tanh z}{\cosh^2 z}\, \varepsilon\left(t, \frac{z}{\eta} + 2\xi\right) \mathrm{d}z, \tag{12}$$

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = 2\mu, \qquad \eta = \text{const}, \tag{13}$$

$$\frac{\mathrm{d}\beta}{\mathrm{d}t} = \mu^2 + \eta \int_{-\infty}^{\infty} \frac{1 - z\tanh z}{\cosh^2 z}\, \varepsilon\left(t, \frac{z}{\eta} + 2\xi\right) \mathrm{d}z. \tag{14}$$

If $\varepsilon(t, x) = f(t)x$, one recovers the exact solution. As the perturbation is random, this is a system of stochastic equations; it allows rather complete analysis when $\varepsilon(t, x) \equiv V_\varepsilon(x)$ is an ergodic process with a finite support localized on the interval $[0, L/\varepsilon^2]$ (Garnier, 1998). The soliton dynamics are then governed by deterministic equations for almost every realization of $V_\varepsilon(x)$. For a soliton of small amplitude (weak nonlinearity), one can define $\nu_0 \ll \mu$ such that decay of the soliton amplitude follows either an exponential, for $\nu < \nu_0$, or a power, for $\nu_0\nu \ll \mu$, law. When the soliton has large amplitude and small velocity, $\mu \ll \nu$, its amplitude experiences rather weak changes, while the velocity decreases.

Solitary Wave Propagation in Media with Fluctuating Parameters

Consider $\varepsilon(t, x) \equiv f(t)$ to be a stationary Gaussian process (4). Then the adiabatic approximation allows us to obtain a Fokker–Planck equation for the distribution function of the soliton parameters, similar to Equation (6), where typical behavior is a kind of Brownian motion (Konotop & Vazquez, 1994). At large times, the adiabatic approximation fails and fluctuations of the medium may result in resonant parametrical processes. Such processes are especially interesting in systems having topological solitons which cannot be destroyed by fluctuations. If, for example, $R[u] \equiv u - u^3$ in Equation (2), long-time numerical simulations of the kink dynamics show anomalous diffusion, and the dispersion of the fluctuation of its center grows as $t^{2.087}$ while the energy of the system increases exponentially. This phenomenon is a manifestation of the stochastic parametrical resonance of linear modes.

Vladimir V. Konotop and Luis Vázquez

See also **Brownian motion; Characteristics; Nonlinear Schrödinger equations; Stochastic processes**

Further Reading

Besieris, I.M. 1980. Solitons in randomly inhomogeneous media. In *Nonlinear Electromagnetics*, edited by P.L.E. Uslenghi, London and New York: Academic Press, pp. 87–116

Doucot, B. & Rammal, R. 1987. On Anderson localization in nonlinear random media. *Europhysics Letters*, 3: 969–974

Fröhlich, J., Spencer, Th. & Wayne, C.E. 1986. Localization in disordered, nonlinear dynamical systems. *Journal of Statistical Physics*, 42: 247–274

Garnier, 1998. Asymptotic transmission of solitons through random media. *SIAM Journal of Applied Mathematics*, 58: 1969–1995

Karpman, V.I. 1979. Soliton evolution in the presence of perturbations. *Physica Scripta*, 20: 462–478

Klyatskin, V.I. 1988. *Imbedding Approach in the Theory of Wave Propagation*, Moscow: Nauka (in Russian)

Knapp, R., Papanicolaou, G. & White, B. 1991. Transmission of waves by a nonlinear random medium. *Journal of Statistical Physics*, 63: 567–583

Konotop, V.V. & Vázquez, L. 1994. *Nonlinear Random Waves*, Singapore: World Scientific

WAVE STABILITY AND INSTABILITY

As with other dynamic entities, a solution of a wave system is considered to be stable if it does not deviate greatly under small perturbations, and unstable if it does.

Consider first the stability of the null solution of a partial differential equation (PDE) system in an infinite and uniform medium. Small perturbations of the null solution can be taken in the form $A \exp[\mathrm{i}(kx - \omega t)]$. To satisfy the PDE, k and ω are related by a dispersion relation, which is denoted as $D(k, \omega) = 0$. The dispersion relation may be solved in the form

$$\omega = \omega_n(k), \quad (n = 1, 2, \ldots). \tag{1}$$

Thus, there may be several solutions (or modes), with different functions $\omega_n(k)$, which are referred to as different modes. In general, ω may be complex for real k, leading to real solutions of the form

$$\begin{aligned} &A \exp[\mathrm{i}(kx - \omega t)] + \text{c.c.} \\ &\quad = \{A \exp[\mathrm{i}(kx - \omega_r t)] + \text{c.c.}\} \exp(\omega_i t), \end{aligned} \tag{2}$$

where $\omega = \omega_r + \mathrm{i}\omega_i$. Therefore, if $\omega_i > 0$ for any mode, that mode will grow with time, indicating instability. If all ω_i are negative, all modes damp away, indicating stability.

Convective and Absolute Instability

In practice, an initial perturbation is often localized and can be represented as a (Fourier) wave packet of various normal modes, in which each component propagates with its own phase velocity ($v_{\mathrm{p}} = \omega/k$). The collective motion of a wave packet, on the other hand, is governed by its group velocity ($v_{\mathrm{g}} = \mathrm{d}\omega/\mathrm{d}k$). If a perturbative wave packet is localized and moving with a certain group velocity, its amplitude at a certain point will at first begin to rise and then eventually fall back to zero. This leads to two distinct concepts: convective and absolute instability, which originally arose in studies of wave instabilities in plasmas (Briggs, 1964).

A wave system is convectively unstable if the maximum amplitude of a perturbing wave packet grows without bound, but at any fixed point of the system, the disturbance eventually relaxes back to zero as the wave packet propagates away. Such behavior is useful for the design of distributed amplifiers, such as the traveling-wave tube or optical (laser) amplifiers.

If the dispersion equation contains unstable modes with zero group velocity, however, a more robust instability arises. In this case, called absolute instability, perturbations grow without bound at every point of the system. Absolutely unstable dynamics can be employed for the design of distributed oscillators (backward-wave oscillators, for example) but not distributed amplifiers.

Modulational Instability

Consider next the stability of a small amplitude component of a wave system, which is stable at infinitesimal amplitude. At finite amplitude, however, the wave is not necessarily stable against wave modulation, because the finite intensity of wave modifies the propagation properties of medium. In this case, a nonlinear dispersion relation that takes into account the finite intensity of wave may be written as $\omega = \omega(k, a)$, where a is the amplitude of wave. For a slowly modulated plane wave having a small but finite amplitude, the frequency ω may be expressed approximately as

$$\omega(k, a) = \omega_0(k) + \alpha(k)a^2 + \mathrm{O}(a^4). \tag{3}$$

(Odd terms in a are excluded from this formulation because they would imply different values of ω when a merely changes its algebraic sign.) Modulations on a linear and weakly nonlinear wave train can be described by the equations (Whitham, 1974)

$$\frac{\partial k}{\partial t} + \frac{\partial \omega}{\partial x} \doteq 0, \tag{4}$$

$$\frac{\partial a^2}{\partial t} + \frac{\partial}{\partial x}\left(\frac{\partial \omega}{\partial k}a^2\right) \doteq 0. \tag{5}$$

Equation (4) follows from the relations $\omega = -\partial\theta/\partial t$ and $k = \partial\theta/\partial x$, where $\theta(x, t)$ is the phase. (This equation can also be viewed as a conservation law for wave crests.) As a^2 is proportional to energy density and $a^2\partial\omega/\partial k = a^2 v_\mathrm{g}$ is proportional to power flow, Equation (5) is the law of energy conservation.

Substituting Equation (3) into these equations and assuming a to be sufficiently small leads to the following equations:

$$\frac{\partial k}{\partial t} + v_\mathrm{g}\frac{\partial k}{\partial x} + \alpha(k_0)\frac{\partial a^2}{\partial x} \doteq 0, \tag{6}$$

$$\frac{\partial a^2}{\partial t} + v_\mathrm{g}' a^2\frac{\partial k}{\partial x} + v_\mathrm{g}\frac{\partial a^2}{\partial x} \doteq 0. \tag{7}$$

Here $v_\mathrm{g} = \mathrm{d}\omega_0(k_0)/\mathrm{d}k_0$, the group velocity of the linear wave with wave number k_0, and $v_\mathrm{g}' = \mathrm{d}^2\omega_0(k_0)/\mathrm{d}k_0^2$. Linearizing k and a by

$$k = k_0 + \bar{k}\exp\{\mathrm{i}(Kx - \nu t)\},$$
$$a = a_0 + \bar{a}\exp\{\mathrm{i}(Kx - \nu t)\},$$

yields the modulational dispersion relation

$$\nu = K\left(v_\mathrm{g} \pm \sqrt{\alpha a_0^2 v_\mathrm{g}'}\right). \tag{8}$$

If $\alpha v_\mathrm{g}' < 0$, ν becomes complex, which implies that the modulation becomes unstable and grows (Lighthill, 1965; Whitham, 1974). Equation (8) is called Lighthill's theorem, and the corresponding modulational instability (which has been studied analytically and experimentally in the context of deep water waves by Benjamin & Feir (1967)) is called the Benjamin–Feir (BF) instability. The nonlinear evolution of modulated envelopes is described by the nonlinear Schrödinger equation. Interestingly, the BF instability can lead to formation of stable traveling waves of modulation—envelope solitons. For further discussions, see Whitham (1974), Infeld & Rowlands (2000), Longuet-Higgins (1978), and McLean (1982).

Soliton Stability

The stability problem of traveling-wave solutions with respect to small perturbations has been studied by various methods, including the normal-mode approach (Infeld and Rowlands, 2000). Consider the solution $u(x, t)$ of a nonlinear PDE to be the stationary traveling wave $u_\mathrm{s}(\xi)$ plus a small perturbation $p(\xi, t)$

$$u(x, t) = u_\mathrm{s}(\xi) + p(\xi, t), \tag{9}$$

where $\xi = x - v_\mathrm{s}t$ and v_s is a propagation velocity. Transforming the original nonlinear PDE to a coordinate system moving at v_s and linearizing, the resulting equation has parameters that are functions of ξ. This equation can then be solved for the time development of the perturbation subject to an appropriate boundary condition.

When such problems are uniform in time (as they often are), it is convenient to express $p(\xi, t)$ as product solution in the form $p(\xi, t) = f(\xi)\exp(\sigma t)$ (separation of variables). Traveling-wave solutions are linearly unstable if any product solution has $\mathrm{Re}(\sigma) > 0$ and asymptotically stable if all product solutions have $\mathrm{Re}(\sigma) < 0$. A stationary solution that is neither unstable nor asymptotically stable is said to be neutrally stable, and there is always such a case because $\sigma = 0$ corresponds to a simple displacement of the traveling wave in the direction of propagation.

As examples, consider the stability of solitons of the Korteweg–de Vries (KdV) and its two-dimensional generalization. Writing the KdV equation as $u_t + 6(u^2)_x + u_{xxx} = 0$, a solitary wave solution is given by

$$u_s = A + k^2\mathrm{sech}^2\{k(x - v_\mathrm{s}t) + \delta\} = 0, \tag{10}$$

where $v_\mathrm{s} = 4k^2 + 12A$ and A and k are both arbitrary. Using the method of normal modes, Jeffrey and Kakutani (1972) and also Berryman (1976) have investigated the stability of this soliton, showing that small localized perturbations do not grow without bound; thus, the KdV soliton is linearly stable. Benjamin formulated a nonlinear stability theory for the KdV soliton and also showed that the KdV soliton is stable against small but finite perturbations (Benjamin, 1972).

A soliton in a two-dimensional (2-d) space often appears as a line soliton (LS). Kadomtsev and Petviashvili studied a 2-d-generalization of the KdV equation in order to discuss the stability of the line soliton with respect to long and small transverse perturbations. They obtained the KP equation (Kadomtsev & Petviashvili, 1970)

$$(u_t + 6(u^2)_x + u_{xxx})_x + su_{yy} = 0, (s = \pm 1), \tag{11}$$

which corresponds to the case of negative and positive dispersion when $s = +1$ and -1, respectively. The line soliton is unstable in the case of positive dispersion and is stable for negative dispersion. The KP equation with positive dispersion also has a periodic soliton (PS) solution. A spatially periodic resonance exist between the LS and PS solutions, as indicated in Figure 1. From this

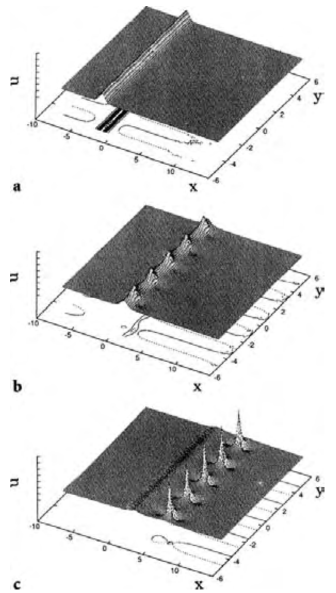

Figure 1. The sequence of snapshots of quasiresonant solution.

figure, we see how a transversely perturbed LS decays into a small LS and a PS, where the instability of the LS may be relaxed by the emission of a PS (Infeld and Rowlands, 2000, Chapter 10). Importantly, the nonlinear stage of instability of soliton is, in general, different from the conclusion of the linear stability theory. For more detailed studies of 2-d soliton stability, see Infeld & Rowlands (2000) and Zakharov et al. (1986).

Eckhaus Instability

In the neighborhood of a bifurcation (the appearance of a new solution as a parameter λ is changed), the description of a dynamical system can be greatly reduced, where the only relevant variable is a complex normalized amplitude: Z. Studies of the dissipative structures that emerge beyond the instability is often facilitated by using amplitude equations, such as the Newell–Whithead (NW) equation (Newell & Whitehead, 1969)

$$\frac{\partial Z}{\partial t} = (\lambda - \lambda_c)Z + \left(\frac{\partial}{\partial x} - \frac{i}{2k_c}\frac{\partial^2}{\partial y^2}\right)^2 Z - |Z|^2 Z. \tag{12}$$

This is the normal form of a symmetry-breaking bifurcation leading to roll or stripe patterns with wave vector parallel to the x-axis, where λ_c is the critical control parameter and k_c is the critical wave number. In the supercritical region, $(\lambda - \lambda_c) > 0$, the NW equation has the two parameter $(\delta k, \theta_0)$ family of solutions:

$$Z_k = A_k e^{i(\delta k x + \theta_0)}, \tag{13}$$

where $A_k = \sqrt{(\lambda - \lambda_c) - (\delta k)^2}$, δk is a small wave number describing the modulation of the basic dissipative structure at the critical wave number k_c, and θ_0 is an arbitrary phase. This solution expresses a stationary roll pattern with wave number $k = k_c + \delta k$.

Introducing a solution that is perturbed by small changes in amplitude and phase leads to the instability criterion (Eckhaus, 1965)

$$|\delta k| > \sqrt{\frac{\lambda - \lambda_c}{3}}. \tag{14}$$

If perturbations transverse to the basic structure are allowed, a zig-zag instability emerges (Nicolis, 1995; Mori & Kuramoto, 1998).

Nerve Impulse Stability

The stability of nerve impulses can be studied by the reduced version of the Hodgkin–Huxley system called the FitzHugh–Nagumo (FN) equation (FitzHugh, 1961; Nagumo et al., 1962):

$$\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2} - f(V) - R,$$
$$\frac{\partial R}{\partial t} = \varepsilon(V - bR),$$

where V is the nerve membrane potential, $f(V) = V(V-1)(V-a)$, and ε is a temperature parameter which controls the rate of change of the recovery variable R.

If $0 < \varepsilon < \varepsilon_c$, the FN equation has slow and fast impulse solutions with propagation speeds $c_s(\varepsilon)$ and $c_f(\varepsilon)$: $0 < c_s < c_f$. The relation between these two speeds and the temperature parameter is sketched in Figure 2, where the fast solution is stable and the slow solution is unstable (Rinzel & Keller, 1973). Thus, small positive perturbations of the slow solution will eventually grow into the fast solution, whereas small negative perturbations will cause the slow solution to collapse to zero (Scott, 2003).

At each value of ε, the FN equation has also two periodic traveling-wave solutions with the same wavelength but different propagation speeds. Viewed as a traveling wave, the fast periodic solution is stable and the slow solution is unstable, just as for a single impulse.

Suppose that a time-periodic boundary condition is imposed at $x = 0$ (where $V(0,t) = V(0,t+T)$) and periodic solutions are sought (for $x > 0$) of the form $V(x,t) = V(x+\lambda,t) = V(x,t+T)$. In this "signaling problem," an important question is whether (or not)

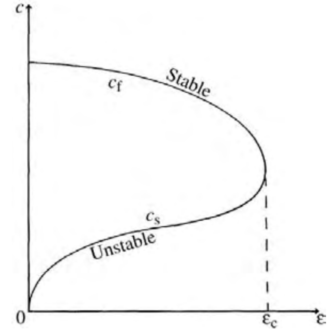

Figure 2. The typical relation between the pulse speed c and the temperature parameter ε.

small perturbations grow with x. Because $c(\lambda) = \lambda/T$, the dependence of the nonlinear frequency $(1/T)$ upon the nonlinear wave number $(1/\lambda)$ is readily calculated, and Rinzel has shown that the condition for solutions not to grow with increasing x is (Rinzel, 1975)

$$\frac{\mathrm{d}(1/T)}{\mathrm{d}(1/\lambda)} > 0.$$

Masayoshi Tajiri

See also **Equilibrium; Kelvin–Helmholtz instability; Modulated waves; Rayleigh–Taylor instability; Stability**

Further Reading

Benjamin, T.B. 1972. The stability of solitary waves. *Proceedings of the Royal Society of London*, 328A: 153–183

Benjamin, T.B. & Feir, J.E. 1967. The disintegration of wave trains in deep water, I. *Journal of Fluid Mechanics*, 27: 417–430

Berryman, J.G. 1976. Stability of solitary waves in shallow water. *Physics of Fluids*, 19: 771–777

Briggs, R.G. 1964. *Electron-Stream Interaction with Plasmas*, Cambrige, MA: MIT Press

Eckhaus, W. 1965. *Studies in Nonlinear Stability Theory*, Berlin: Springer

FitzHugh, R. 1961. Impulses and physiological states in theoretical models of nerve membranes. *Biophysical Journal*, 1: 445–466

Infeld, E. & Rowlands, G. 2000. *Nonlinear Waves, Solitons and Chaos*, Cambridge and New York: Cambridge University Press

Jeffrey, A. & Kakutani, T. 1972. Weakly nonlinear dispersive waves: a discussion centered around the Korteweg–de Vries equation. *SIAM Review*, 14: 582–643

Kadomtsev, B.B. & Petviasvili, V.I. 1970. On the stability of solitary waves in weakly dispersive media. *Soviet Physics-Doklady*, 15: 539–541

Lighthill, J. 1965. Contributions to the theory of waves in nonlinear dispersive system. *Journal of the Institute of Mathematics and Its Applications*, 1: 269–306

Longuet-Higgins, M.S. 1978. The instabilities of gravity waves of finite amplitude. *Proceeding of Royal Society of London*, *I Superhamonics*, A360: 471–488; *II Subharmonics*, A360: 489–505

McLean, J.W. 1982. Instabilities of finite-amplitude water waves, *Journal of Fluid Mechanics*, 114: 315–330

Mori, H. & Kuramoto, Y. 1998. *Dissipative Structures and Chaos*, Berlin: Springer

Nagumo, J., Arimoto, S. & Yoshizawa, S. 1962. An active pulse transmission line simulating nerve axon. *Proceedings of the Institute of Radio Engineering*, 50: 2061–2070

Newell, A.C. & Whitehead, J.A. 1969. Finite bandwidth, finite amplitude convection. *Journal of Fluid Mechanics*, 38: 279–303

Nicolis, G. 1995. *Introduction to Nonlinear Science*, Cambridge and New York: Cambridge University Press

Rinzel, J. 1975. Spatial stability of traveling-wave solutions of a nerve conduction equation. *Biophysical Journal*, 15: 975–988

Rinzel, J. & Keller, J.B. 1973. Traveling-wave solutions of nerve conduction equation. *Biophysical Journal*, 13: 1313–1337

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

Zakharov, V.E., Kuznetsov, E.A. & Rubenchik, A.M. 1986. Soliton stability. In *Solitons*, edited by S.E. Trullinger, V.E. Zakharov & V.L. Pokrovsky, Amsterdam and New York: North-Holland, pp. 503–554

WAVE TANKS

See **Laboratory models of nonlinear waves**

WAVELETS

The wavelet transform is a tool that divides up data, functions, or operators into different frequency components and then studies each component with a resolution matched to its scale. The idea emerged independently in many different fields, including mathematics, quantum physics, electrical engineering, and seismology. By the end of the 1980s, French researchers Grossmann & Morlet (1984), Meyer (1990), and later Daubeschies (1992) had laid the mathematical foundations of the wavelet transform technique. Their work was motivated by the problem Morlet (a geophysicist) was facing while analyzing seismic data which comprised different features in time and frequency.

The frequency content of a signal can be obtained by taking its Fourier transform. However, in transforming to the frequency domain, time information is lost and it is impossible to tell when a particular event took place. To correct this deficiency, the Fourier transform was adapted to analyze only a small section (or window) of the signal at a time. Such a short-time Fourier transform (STFT) maps a signal into a two-dimensional function of time and frequency and provides information about both when and at what frequencies a signal event occurs. This information can only be obtained with limited precision, and that precision is determined by the size of the time window. The main drawback of STFT is that once a particular size for the time window is chosen, that window is the same for all frequencies. With Morlet's data, this approach failed either to follow the time evolution of rapid events or to estimate the frequency content in the low-frequency band.

Wavelet analysis represents the next logical step: signal cutting is performed by a window of variable length. Short windows are used at high frequencies and long windows at low frequencies. Thus, the time-frequency resolution is no longer constant but changes with frequency, allowing good time resolution for high frequencies and good frequency resolution for low frequencies to be achieved.

In Fourier analysis, the signal is decomposed into sine waves of various frequencies. Similarly, wavelet analysis is a decomposition of a signal into a set of basis functions $\Psi_{s,\tau}(t)$ called wavelets

$$\hat{f}(s,\tau) = \int f(t)\Psi^{*}_{s,\tau}\,\mathrm{d}t, \tag{1}$$

where $\hat{f}(s,\tau)$ is the wavelet transform of $f(t)$ and * denotes complex conjugation. The wavelets are generated from a single basic wavelet $\psi(t)$ called the mother wavelet, by scaling and translation:

$$\Psi_{s,\tau} = \sqrt{|s|}\,\psi\Big(\frac{t-\tau}{s}\Big). \tag{2}$$

As the scaling parameter s changes, the wavelets cover different frequency ranges: large values of s correspond to low frequencies, small values of s correspond to high frequencies. Changing the translation parameter τ allows the time localization center to be moved: each $\Psi_{s,\tau}(t)$ is localized around $t=\tau$. The factor $\sqrt{|s|}$ is for energy normalization across different scales.

An important difference between wavelet and Fourier transforms is that wavelet basis functions are not specified. The theory deals with general properties of wavelets and defines a framework within which different wavelets can be designed. Numerous families of wavelets have been proposed and proven to be useful in different applications. In order for a function to be used as the mother wavelet, it must allow for analysis and reconstruction of the signal without loss of information. This so-called admissibility condition implies that the average value of a wavelet in the time domain must be zero and it must, therefore, be oscillatory, a wave. The other important property of wavelets (the regularity condition) states that wavelets are smooth and concentrated both in time and frequency. This makes them suitable for capturing local features of a signal. Figure 1 presents an example of a wavelet, the Morlet wavelet, in time and frequency domains for two scales.

We can distinguish between the continuous and discrete wavelet transform. In a continuous transform, the parameters s and τ vary continuously. It maps a one-dimensional signal to a two-dimensional time-scale joint representation that is highly redundant. This redundancy can be either exploited or removed. To reduce the redundancy, discrete wavelets have been introduced, which can only be scaled and translated in discrete steps. For the scale, we choose integer powers of fixed dilation parameter $s_0 > 1$, $s = s_0^j$, and j is an integer. The discretization of the translation parameter τ depends on j: narrow wavelets are translated by small steps, while wider wavelets are translated by larger steps, $\tau = n\tau_0 s_0^j$, with j and n integers and $\tau_0 > 0$ is fixed. For some special choices of mother wavelets, the discrete wavelets can be made orthogonal to their own translations and dilations. In this case, they behave exactly like an orthonormal basis and redundancy is removed.

Wavelets have a bandpass-like spectrum and can be viewed as bandpass filters. Compression in time stretches the spectrum and shifts it upwards, while stretching in time compresses the bandwidth and shifts it toward zero (Figure 1). The series of dilated wavelets can be used to cover the spectrum of a signal. However, an infinite number of wavelets is needed to reach zero frequency. This problem was solved by introducing a low-pass or averaging filter with a spectrum that belongs to the scaling function $\phi(t)$. If we analyze the signal using a combination of scaling functions and wavelets, the scaling function takes care of the spectrum otherwise covered by all the wavelets up to a chosen scale, while the rest is done by wavelets. The family of scaling functions and wavelets allows for wavelet multiresolution analysis. In multiresolution analysis, the signal is split into an approximation on a coarser scale, obtained using the scaling function, and details on a current scale, obtained by wavelets. This process is repeated, giving a sequence of approximations and details removed at every scale. After N iterations, the original signal can be reconstructed by summing up the last approximations and details on all previous scales.

Interactions between the fields where wavelets were first introduced have led to many wavelet applications. Wavelets are of particular interest for the analysis of nonstationary signals with broad spectra because they permit time-frequency presentation with logarithmic resolution. Wavelet analysis is capable of revealing aspects of signals that other signal analysis techniques miss, such as trends, breakdown points, discontinuities in higher derivatives, and self-similarity. Furthermore, wavelet analysis can often compress (or reduce noise in) a signal without appreciable degradation. The above-described one-dimensional aspect can be generalized to more dimensions, for example, to handle image analysis. Wavelets are a very powerful tool for image compression, since wavelet transform clearly separates high-pass and low-pass information on a pixel-by-pixel

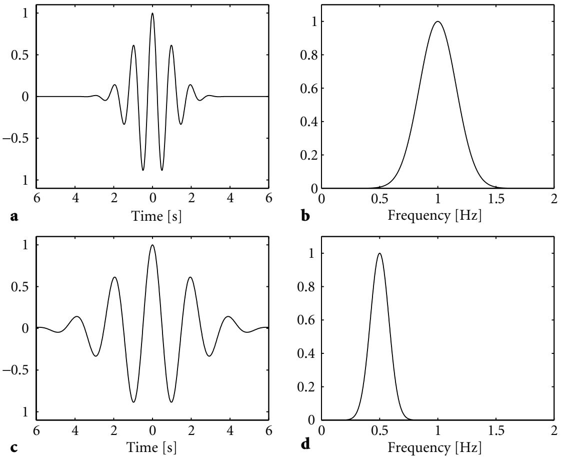

Figure 1. Morlet wavelet in time and frequency domain for scales $s = 1$ (a,b) and $s = 2$ (c,d).

basis. Among the successes of wavelet representation are the compression of digitized fingerprints by the US Federal Bureau of Investigation and image compression using the JPEG 2000 standard (the older JPEG standard uses a non-wavelet compression).

Maja Bračič Lotrič and Aneta Stefanovska

See also **Integral transforms; Nonlinear signal processing**

Further Reading

Daubeschies, I. 1992. *Ten Lectures on Wavelets*, Philadelphia, PA: Society for Industrial and Applied Mathematics

Grossmann, A. & Morlet, J. 1984. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal of Mathematical Physics*, 27: 2437–2479

Meyer, Y. 1990. *Ondelettes et Operateurs*, 3 vols, Paris: Herrmann; vol. 1 as *Wavelets and Operators*, Cambridge and New York: Cambridge University Press, 1992 and vols. 2 and 3 as *Wavelets: Calderón-Zygmund and Multilinear Operators*, Cambridge and New York: Cambridge University Press, 1997

WAVE NUMBER SELECTION

See **Pattern formation**

WEAK COLLAPSE

See **Development of singularities**

WEAK TURBULENCE

See **Turbulence**

WEAKLY NONLINEAR ANALYSIS

See **Quasilinear analysis**

WEATHER FRONTS

See **Atmospheric and ocean sciences**

WEIERSTRASS ELLIPTIC FUNCTIONS

See **Elliptic functions**

WESTERVELT EQUATION

See **Nonlinear acoustics**

WHISTLERS

See **Nonlinear plasma waves**

WHITHAM'S METHOD

See **Modulated waves**

WIGNER STATISTICS

See **Random matrix theory I, II**

WINDING NUMBERS

If you wrap a rubber ring (rubber band) around a pencil, the intuitive idea of an integer invariant for the wrapping process arises. The number of oriented turns around the pencil is an integer, and it is independent of how tight or creased the ring is. The only way to change this integer is to wrap/unwrap one loop of the ring at the endpoints of the pencil. Using a rubber string, the number of turns could be some real number instead.

How does this fact connect with dynamics? Consider a three-dimensional (3-d) dynamical system having a torus as phase space. Let the hole in the torus represent the pencil and the rubber ring represent a closed trajectory within the torus. Small modifications to the trajectory will not alter the wrapping property.

The underlying feature in the above examples is the forbidden region given by the pencil or the hole of the torus. To further analyze its structure, decompose 3-d space ($\mathbb{R}^3$) as the product of the pencil direction times a perpendicular plane (equivalent to $\mathbb{R}^2$). The role of the pencil is to identify a special point on that plane given by the projection along the pencil direction.

Similarly, consider a periodic orbit in $\mathbb{R}^3$, as the forbidden region. Since the points in the periodic orbit are regular points in a small tube around the periodic orbit, the flow can be decomposed in a component parallel to the periodic orbit and a projected flow onto a perpendicular section. Hence, any other closed trajectory sufficiently near to the orbit will wind around it. A similar situation arises around a period-one orbit of a periodically forced flow in $\mathbb{R}^2$.

The winding number characterizes the topological properties of "the plane minus one point." Moreover, the topology of the plane minus n points gives a deeper characterization of the periodic orbit structure of 3-d dynamical systems admitting a Poincaré section.

Definition

Consider a simple continuous closed curve γ in the complex plane ($\mathbb{C}$) and a point $z_0 \in \mathbb{C} - \gamma$. The winding number n is defined as

$$n(\gamma, z_0) = \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{\mathrm{d}z}{z - z_0}. \qquad (1)$$

We may regard our wrapped rubber ring as a suitable complex function γ of the unit circle, thus connecting our motivating idea with the formal definition (similarly for the second example if we project the torus along the direction perpendicular to the hole when seen as a disc). Where appropriate in the sequel, we will let $z_0 = 0$ for simplicity and recast γ as a map of the unit circle $\gamma : \mathbb{S}^1 \to \mathbb{S}^1$. It follows from Equation (1) that $n(\gamma, 0)$ is a real integer given by

$$n(\gamma, 0) = \int_{\mathbb{S}^1} \frac{\gamma'(\theta)}{\gamma(\theta)}\, \mathrm{d}\theta, \qquad (2)$$

which is called the degree of γ (Rotman, 1988, p. 50).

Applications

Homotopy classes of the circle. A loop is a continuous map g of the circle to itself such that $g(0) = g(2\pi) = 2\pi$. Two loops α and β are homotopic if there exists a continuous map $H : \mathbb{S}^1 \times [0, 1] \to \mathbb{S}^1$ such that $H(\cdot, 0) = \alpha$, $H(\cdot, 1) = \beta$ and for each t, $H(\cdot, t)$ is a loop. In other words, two loops are homotopic if one can continuously deform one of them into the other, keeping it as a loop all the way throughout the deformation. The winding number classifies the homotopy classes of loops, namely, if σ is homotopic to γ, then $n(\sigma, 0) = n(\gamma, 0)$ (Rotman, 1988, p. 52).

Braids and periodic orbits. While the plane minus one point produces the winding number as a class invariant under homotopies, the homotopy classes of the plane with n special points requires a more elaborated structure which connects nicely with the dynamical properties of 3-d flows admitting a Poincaré section. In fact, periodic orbits of such flows can be regarded as imbeddings of the unit disk in phase space parametrized with time in units of $2\pi/T$ where T is the minimal period. On the Poincaré section, these special trajectories appear as an invariant set of n periodic points. The homotopy classes of loops on the plane with an invariant set of n points are classified by elements of the *Braid group on n strands* (Thurston, 1988; Hall, 1994; Natiello & Solari; 1994).

Linking number. In the same lines, given a pair of periodic orbits in phase space, we may think of the number of turns that one orbit does around the other when completing one excursion along itself. Such number is a link invariant which has a natural interpretation in terms of winding number, and it is called *linking number* (Uezu & Aizawa, 1982; Solari et al., 1996).

For 3-d flows admitting a Poincaré section, the periods of the orbits are commensurate and one may compute the average rotation per period of one orbit around another. This is called the *relative rotation rate* (Solari et al., 1996) and helps in understanding the orbit organization of such flows.

Poincaré index (PI). Consider a planar dynamical system $\dot{x} = f(x, y)$, $\dot{y} = g(x, y)$ and a simple closed counterclockwise curve C not passing through any equilibrium points. The PI k computed along C is defined as

$$\begin{aligned} k &= \frac{1}{2\pi} \int_C \mathrm{d} \left\{ \arctan \left(\frac{\mathrm{d}y}{\mathrm{d}x} \right) \right\} \\ &= \frac{1}{2\pi} \int_C \frac{f\, \mathrm{d}g - g\, \mathrm{d}f}{f^2 + g^2}. \end{aligned} \qquad (3)$$

(See below for a discussion of the PI in terms of complex analysis and winding number.)

In the context of planar dynamics, the PI of a node or a center is $+1$, of a hyperbolic saddle point is -1, and of a closed orbit is $+1$. Also the PI of a closed curve not containing fixed points is zero, and the PI of a closed curve equals the sum of the indices of the fixed points within (Guckenheimer & Holmes, 1983, p. 51).

Fixed point theorems. The degree of a map can be generalized to higher dimensions. In fact, this property (or the winding number when adequate) is a basic ingredient in the proof of Brouwer's fixed point theorem. An interesting discussion of this fact along with some philosophical considerations can be found in www.mathpages.com.

Complex analysis. The computation of the winding number is a standard tool in the proof of the Fundamental Theorem of Algebra.

Also, let C be a closed contour on the complex plane not passing through any singularities or zeroes of the complex function f, which is analytic inside C except at most at a finite number of poles. Then

$$\frac{1}{2\pi \mathrm{i}} \int_C \frac{f'(z)}{f(z)} \,\mathrm{d}z = N - P, \tag{4}$$

where N is the number of zeroes of f and P the number of poles inside C. This is called the Principle of the Argument in standard textbooks (Wunsch, 1994, p. 458).

This result is related to Equation 2 and to the PI. Concerning Equation (2), taking the special point of the plane to be the origin (or any point inside the unit circle), n counts how many turns γ performs around this point when running along the unit circle. Assume now that f has only one zero inside C with multiplicity n and no poles. Then f restricted to C is exactly the same as γ with a suitable choice of parametrization for C and $\mathbb{S}^1$. Concerning the PI, let $z = x + \mathrm{i}y$ and $F(z) = f(x, y) + \mathrm{i}g(x, y)$, regarding the xy-plane as the complex plane. If the vector field (f, g) is continuous, F will not have poles within C and the Poincaré index reduces to the Principle of the Argument calculation for F.

MARIO NATIELLO AND HERNÁN SOLARI

See also **Conley index; Phase space; Poincaré theorems**

Further Reading

Guckenheimer, J. & Holmes, P.J. 1983. *Nonlinear Oscillators, Dynamical Systems and Bifurcations of Vector Fields*, New York and London: Springer

Hall, T. 1994. Fat one-dimensional representatives of pseudo-Anosov isotopy classes with minimal periodic orbit structure. *Nonlinearity*, 7: 367–384

Natiello, M.A. & Solari, H.G. 1994. Remarks on braid theory and the characterisation of periodic orbits. *Journal of Knot Theory and Its Ramifications*, 3: 511–539

Rotman, J.J. 1988 *An Introduction to Algebraic Topology*. New York: Springer

Solari, H.G. & R. Gilmore. 1988. Relative rotation rates for driven dynamical systems. *Physical Review* A, 37: 3096

Solari, H.G., Natiello, M.A. & Mindlin, B.G. 1996. *Nonlinear Dynamics: A Two-way Trip from Physics to Math*, Bristol: Institute of Physics Publishing

Thurston, W.P. 1988. On the geometry and dynamics of diffeomorphisms of surfaces. *Bulletin of American Mathematical Society*, 19: 417

Uezu, T. & Aizawa, Y. 1982. Topological character of a periodic solution in three-dimensional ordinary differential equation system. *Progress of Theoretical Physics*, 68: 1907

Wunsch, A.D. 1994. *Complex Variables with Applications*, Reading, MA: Addison-Wesley

www.mathpages.com see http://www.mathpages.com/home/kmath262/kmath262.htm or do a search for "Brouwer" on http://www.mathpages.com

Y

YANG–BAXTER EQUATION

See **Quantum inverse scattering method**

YANG–MILLS THEORY

Modern particle theories, such as the Standard Model, are quantum Yang–Mills theories. In a quantum field theory, space-time fields with relativistic field equations are quantized and, in many calculations, the quanta of the fields are interpreted as particles. In a Yang–Mills theory, these fields have an internal symmetry: they are acted on by space-time-dependent non-abelian group transformations in a way that leaves physical quantities, such as the action, invariant. These transformations are known as local gauge transformations, and Yang–Mills theories are also known as non-abelian gauge theories.

Yang–Mills theories, and especially quantum Yang–Mills theories, have many subtle and surprising properties and are still not fully understood, either in terms of their mathematical foundations or in terms of their physical predictions. However, the importance of Yang–Mills theory is clear; the Standard Model has produced calculations of amazing accuracy in particle physics, and in mathematics, ideas arising from Yang–Mills theory and from quantum field theory are increasingly important in geometry, algebra, and analysis.

Consider a complex doublet scalar field ϕ_a; a scalar field is one that has no Lorentz index, but, as a doublet, ϕ_a transforms under a representation of SU(2), the group represented by special unitary 2×2 matrices:

$$\phi_a(x) \to g_{ab}\phi_b(x), \tag{1}$$

where $g \in \mathrm{SU}(2)$ and the repeated index are summed over. If this is a global transformation; that is, if g is independent of x, then derivatives of ϕ_a have the same transformation property as ϕ_a itself:

$$\frac{\partial \phi_a}{\partial x_\mu} \to \frac{\partial g_{ab}\phi_b}{\partial x_\mu} = g_{ab}\frac{\partial \phi_b}{\partial x_\mu} \tag{2}$$

However, this is not true for a local, or space-time-dependent, transformations where

$$\frac{\partial \phi_a}{\partial x_\mu} \to \frac{\partial g_{ab}\phi_b}{\partial x_\mu} = g_{ab}\frac{\partial \phi_b}{\partial x_\mu} + \frac{\partial g_{ab}}{\partial x_\mu}\phi_b. \tag{3}$$

In order to construct an action that includes derivatives and that is invariant under local transformations, a new derivative is defined that transforms the same way as ϕ_a:

$$D_\mu \phi_a = \frac{\partial \phi_a}{\partial x_\mu} + (A_\mu)_{ab}\phi_b, \tag{4}$$

where A_μ is a new two-indexed space-time field, called a gauge field or gauge potential, defined to have the transformation property

$$(A_\mu)_{ab} \to g_{ac}(A_\mu)_{cd}g^{-1}_{db} - \frac{\partial g_{ac}}{\partial x_\mu}g^{-1}_{cb}. \tag{5}$$

Now, under a local transformation

$$D_\mu \phi_a \to g_{ab} D_\mu \phi_b \tag{6}$$

and so, $D_\mu \phi_a$ transforms in the same way as ϕ_a. This derivative is called a covariant derivative.

A physical theory that includes the gauge field A_μ should treat A_μ as a dynamical field, and so the action should have a kinetic term for A_μ. In other words, the action should include derivative terms for A_μ. These terms are found in the field strength

$$F_{\mu\nu} = \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu} + [A_\mu, A_\nu] \tag{7}$$

that has the covariant transformation property

$$(F_{\mu\nu})_{ab} \to g_{ac}(F_{\mu\nu})_{cd}g^{-1}_{db}, \tag{8}$$

where $[A_\mu, A_\mu]$ is the normal matrix commutator. In fact, the simplest Yang–Mills theory is pure Yang–Mills theory with action

$$S[A] = -\frac{1}{2}\int \mathrm{d}^4x \, \mathrm{trace}\, F_{\mu\nu}F^{\mu\nu} \tag{9}$$

and corresponding field equation

$$\frac{\partial F_{\mu\nu}}{\partial x_\mu} = 0. \tag{10}$$

Solutions to this equation are known as instantons (*See* **Instantons**).

More generally, Yang–Mills theories contain gauge fields and matter fields like ϕ and fields with both group and Lorentz or spinor indices. Also, the group action described here can be generalized to other groups and to other representations. In the case of the Standard Model of particle physics, the gauge group is $\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$, and the group representation structure is quite intricate.

Yang–Mills theory was first discovered in the 1950s. At that time, quantum electrodynamics (QED) was known to describe electromagnetism. Quantum electrodynamics is a local gauge theory but with an abelian gauge group. It was also known that there is an approximate global non-abelian symmetry called isospin symmetry that acts on the proton and neutron fields as a doublet and on the pion fields as a triplet. This suggested that a local version of the isospin symmetry might give a quantum field theory for the strong force with the pion's fields as gauge fields (O'Raifeartaigh, 1997). This did not work because pion fields are massive whereas gauge fields are massless, and the main thrust of theoretical effort in the 1950s and 1960s was directed at other models of particle physics.

However, it is now known that the proton, neutron, and pion are not fundamental particles, but are composed of quarks and that there is, in fact, a quantum Yang–Mills theory of the strong force with quark fields and gauge particles called gluons. Furthermore, it is now known that it is possible to introduce a particle, called a Higgs boson, to break the non-abelian gauge symmetry in the physics of a symmetric action and give mass terms for gauge fields. This mechanism is part of the Weinberg–Salam model, a quantum Yang–Mills theory of the electroweak force, that is a component of the Standard Model and that includes both massive and massless gauge particles.

These theories were only discovered after several key experimental and theoretical breakthroughs in the late 1960s and early 1970s. After it became clear from collider experiments that protons have a substructure, theoretical study of the distance-dependent properties of quantum Yang–Mills theory led to the discovery that Yang–Mills fields are asymptotically free (Gross, 1999). This means that the high-energy behavior of Yang–Mills fields includes the particle-like properties seen in experiments, but the low-energy behavior may be quite different, and in fact, the quantum behavior might not be easily deduced from the classical action. Confinement and the mass gap are examples of this. The strong force is a local gauge theory with quark fields. The quark structure of particles is observed in collider experiments; but free quarks are never detected, instead, at low energies, they appear to bind together to form composite particles, such as neutrons, protons, and pions. This is called confinement. It is possible to observe this behavior in simulations of the quantum gauge theory of the strong force, but it has not been possible to prove mathematically that confinement is a consequence of the theory. The same is true of the mass gap; it is known that particles have nonzero mass, and this is observed in simulations, but there is no known way of deriving the mass gap mathematically from the original theory (Clay, 2002).

The symmetries of Yang–Mills theory can be extended to include a global symmetry between the bosonic and fermionic fields called supersymmetry. While there is no direct evidence for supersymmetry in physics, the indirect case is very persuasive, and it is commonly believed that direct evidence will be found in the future. Often, supersymmetric theories are more tractable; for example, Seiberg and Witten have found an exact formula for many quantum properties in $N = 2$ super-Yang–Mills theory (Seiberg & Witten, 1994). It is also commonly believed by theoretical physicists that the quantum Yang–Mills theories in particle physics are in fact a limit of a more fundamental string theory.

CONOR HOUGHTON

See also **Higgs boson; Instantons; Matter, nonlinear theory of; Particles and antiparticles; Quantum field theory; String theory; Tensors**

Further Reading

Clay Mathematics Institute, Millennium Prize, 2002. The Clay Institute has offered a prize for a rigorous formulation of a quantum Yang–Mills theory in which there is a mass gap. The Clay Institute web site has a description of the problem along with an essay by A. Jaffe & E. Witten http://www.claymath.org/Millennium_Prize_Problems

Davies, C. 2002. *Lattice QCD*, Bristol: Institute of Physics Publishing

Gross, D. 1999. Twenty-five years of asymptotic freedom. *Nuclear Physics Proceedings Supplements*, 74: 426–446

O'Raifeartaigh, L. 1997. *The Dawning of Gauge Theory*, Princeton, NJ: Princeton University Press

Seiberg, N. & Witten, E. 1994. Electric-magnetic duality, monopole condensation and confinement in $N = 2$ supersymmetric Yang–Mills theory. *Nuclear Physics* B, 426: 19–52, Erratum-ibid 430: 485–486 and Monopoles, duality and chiral symmetry breaking in $N = 2$ supersymmetric QCD. *Nuclear Physics* B, 431: 484–558

Weinberg, S. 1996. *Quantum Field Theory*, vol. 2, Cambridge and New York: Cambridge University Press

Z

ZAKHAROV–SHABAT EQUATION

See **Nonlinear Schrödinger equations**

ZELDOVICH–FRANK-KAMENETSKY EQUATION

In 1938, Yakov Zeldovich and David Frank-Kamenetsky published a brief theoretical paper devoted to flame propagation, presenting one of the first nonlinear traveling-wave front solutions (Zeldovich & Frank-Kamenetsky, 1938). Although both scientists later played outstanding roles in Soviet H-bomb and nuclear projects (and then both performed remarkable works in different fields of physics, including the theory of elementary particles, plasma physics, astrophysics, and cosmology), in the third decade of the last century, they were engaged with the theory of combustion and detonation and attendant problems of chemical kinetics. Their paper was intended for experts and had a rather specialized character, but one of the problems that they examined can be presented as follows.

Consider the autocatalytic production, destruction, and diffusion transfer of a substance proceeding in a homogeneous active chemical medium that occupies some region of physical space. Such processes obey two fundamental macroscopic relationships, the continuity equation

$$u_t + \operatorname{div} \boldsymbol{J} = f(u)$$

and the phenomenological Fick diffusion law

$$\boldsymbol{J} = -D \operatorname{grad} u \qquad (D = \text{const}).$$

Here, $u = u(\boldsymbol{r}, t)$ is the concentration of the substance at the point $\boldsymbol{r} = \{x, y, z\}$ and moment of time t; the literal subscripts symbolize the derivatives with respect to the corresponding variables; the symbols div and *grad* designate the spatial divergence and gradient operators; $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{r}, t)$ is the vector of diffusion flux density of the substance; the function $f(u)$ is the kinetic function of th active medium, which determines the dependence of the production/destruction rate of substance per unit volume on the concentration u; and D is diffusivity.

Generally speaking, in a medium of this type, self-sustaining nonlinear concentration waves can propagate, and their velocities will not be arbitrary but will be determined by the balance between two types of processes: the active processes of production/destruction of a substance at each local patch of medium and the passive processes of diffusion transfer between the patches. With the diffusivity taken as unity (this can always be achieved by a proper choice of units of measurements), the wave velocity will depend only on the parameters of the function $f(u)$. The problem consists in finding both the profile of a propagating wave and the wave velocity.

Zeldovich and Frank-Kamenetsky solved the last problem for the case of a one-dimensional infinite medium extending along the x-axis, whose kinetic function is described by a cubic polynomial with three zeros (see below). In this case, the diffusion flux density $\boldsymbol{J}$ has only one nonzero component, the x-component J, and the two equations written above look (at $D = 1$) like

$$u_t + J_x = f(u),$$

$$J = -u_x,$$

or, equivalently,

$$u_t = u_{xx} + f(u), \qquad (1)$$

where $f(u)$ is given as

$$f(u) = -Ku(u - b)(u - 1), \qquad (2)$$

where K is a positive constant and b is a constant $(0 < b < 1)$. The reaction-diffusion equation (1) endowed with the kinetic function (2) is referred today to as the Zeldovich–Frank-Kamenetsky (ZF) equation. As its authors have noted, the cubic polynomial structure of (2) corresponds to autocatalysis of the second order.

The active chemical medium described by the ZF equation is bistable: it has two homogeneous stable states described by two trivial solutions $u(x,t) \equiv 0$, $u(x,t) \equiv 1$ of Equation (1), which are determined by zeros $u = 0, 1$ of the kinetic function (2). Its third homogeneous state $u(x,t) \equiv b$, which is determined by intermediate zero b of $f(u)$ and located between 0 and 1, is unstable; it plays the role of a threshold. Testing the stability of these states proceeds as follows.

Let $u = u_*$, $u_* \in \{0, b, 1\}$ be a coordinate of one of three zeros of the kinetic function, and $k_* = f_u(u_*)$, $k_* \in \{k_0, k_b, k_1\}$ be the slope of the kinetic function at this zero; note that the values of k_* satisfy the inequalities

$k_0 = f_u(0) = -Kb < 0,$
$k_b = f_u(b) = Kb(1-b) > 0,$
$k_1 = f_u(1) = -K(1-b) < 0.$

Next, find the solution to Equation (1) in form of $u(x,t) = u_* + w(x,t)$, where

$$\begin{aligned} &w(x,t) = \omega(x,t)\exp(k_* t)\\ &w(x,0) = \omega(x,0) = w_0(x),\\ &|w_0(x)| \ll 1 \end{aligned}$$

is the perturbation, supposed to be small at moments of time close to $t = 0$. Substituting these expressions into (1) and linearizing the kinetic function, one comes to the usual diffusion equation $\omega_t = \omega_{xx}$. The solution of the latter, defining on the infinite x-axis and satisfying the initial condition $\omega(x,0) = w_0(x)$ is well known to be given by Poisson's formula. Recopying this solution and multiplying it by the factor $\exp(k_* t)$ yields the expression

$$\begin{aligned} w(x,t) &= 2^{-1}(\pi t)^{-1/2}\exp(k_* t)\\ &\times \int_{-\infty}^{+\infty} \exp\left[-\left(x - x'\right)^2/4t\right] w_0(x')\,\mathrm{d}x', \end{aligned}$$

which describes the time evolution of perturbation. Examination of this expression indicates that $w(x,t)$ decreases with time at the values $k_* = k_0 < 0$ and $k_* = k_1 < 0$, that is, near the states determined by zeros 0, 1 of function $f(u)$. At $k_* = k_b > 0$ (near the state determined by intermediate zero b of function $f(u)$), the perturbation increases because the exponent before the integral rises with time faster than the preexponential factor $2^{-1}(\pi t)^{-1/2}$ falls.

For physical reasons, a bistable medium obeying the ZF equation must maintain the propagation of nonlinear wave fronts, which switch the medium from one of its two stable states to the other. For example, a wave of this type can be obtained numerically by setting the initial conditions $u(x,0) = 1$ at $x < 0$, $u(x,0) = 0$ at $x > 0$. In the steady-state regime, which is established after some transition time, the switching wave moves along x with constant velocity v and possesses a steady spatial profile, which is described by the traveling-wave front solution of a ZF equation of the kind

$$u = u(\xi), \quad \xi = x - vt, \tag{3}$$

$$u(-\infty) = 1, \quad u(+\infty) = 0, \tag{4}$$

$$|u(\xi)| < \infty. \tag{5}$$

Obtaining this solution is a principal goal of the ZF analysis. It includes both the problem of derivation of $u(\xi)$ and the relevant problem of determination of v. Of course, the latter problem is a central one: as a rule, it appears every time when one deals with wave propagation in nonlinear reaction-diffusion systems. Zeldovich and Frank-Kamenetsky found the expression for v in the case of (1), (2) to be

$$v = \sqrt{K/2}(1-2b) = \sqrt{2K}\,(1/2 - b). \tag{6}$$

This beautiful formula connects the wave front velocity v with the position b of intermediate zero of the cubic kinetic function. In particular, it indicates that v is proportional to the deviation of b from the middle value of u, which is equal to 1/2. If $b < 1/2$ ($b > 1/2$), then v is positive (negative), and the wave front, which obeys the boundary conditions (4), (5), moves in the positive (negative) direction of the x-axis. If b is exactly equal to 1/2, then the front does not move: it is stationary.

A short derivation of (6) along with the expression for $u(\xi)$ proceeds as follows. First, substituting (3) directly into two equations written immediately before (1) (they are evidently equivalent to (1)) and allowing for the relations $\partial/\partial t = -v\,(\mathrm{d}/\mathrm{d}\xi)$, $\partial/\partial x = \mathrm{d}/\mathrm{d}\xi$, which follow from (3), one obtains

$$u_\xi = -J, \tag{7a}$$

$$J_\xi = -vJ + f(u). \tag{7b}$$

Second, dividing (7b) by (7a), one excludes the independent variable (thereupon, the evident equality $J_\xi/u_\xi = J_u$ is used) and reduces Equation (7) to the single equation

$$JJ_u = vJ - f(u). \tag{8}$$

To be integrated correctly, this differential equation must be provided with the proper boundary conditions at the points $u = 0$, 1, that is, in the equilibrium states that are achieved by the traveling wave front solution at $\xi \longrightarrow \pm\infty$. To set these conditions, one should know the asymptotic behavior of diffusion flux $J = u_\xi(\xi)$ generated by the traveling wave near the states $u = 0$, 1. To recognize it, one represents the unknown solution near these states in the form of $u(\xi) = u_* + w(\xi)$, $u_* \in \{0, 1\}$, where $w(\xi)$ is a small perturbation necessarily satisfying the conditions $w(\pm\infty) = 0$. Substituting this expression directly into (1) and linearizing the kinetic function yields

$$\begin{aligned} &w_{\xi\xi} + vw_\xi - |k_*|w = 0, \quad (k_* = f_u(u_*),\\ &k_* \in \{k_0, k_1\}, k_0 < 0, k_1 < 0), \end{aligned}$$

where the negative parameter k_* is presented as the positive constant $|k_*|$ taken with the sign "minus." The solutions of the last linear ordinary equation, singled out by the conditions $w(\pm\infty) = 0$ look like

$$\begin{aligned} w &= A_1 \exp(\lambda_1 \xi), \quad A_1 = \text{const}, \\ \lambda_1 &= -(v/2) + \sqrt{(v/2)^2 + |k_*|} > 0 \\ &\quad (\text{at}\, \xi \longrightarrow -\infty), \end{aligned}$$

$$\begin{aligned} w &= A_2 \exp(\lambda_2 \xi), \quad A_2 = \text{const}, \\ \lambda_2 &= -(v/2) - \sqrt{(v/2)^2 + |k_*|} < 0 \\ &\quad (\text{at}\, \xi \longrightarrow +\infty). \end{aligned}$$

Thus, irrespective of the unknown value of velocity v, the traveling-wave front solution $u(\xi)$ approaches its limit values 0, 1 exponentially and therefore, the diffusion flux $J = -u_\xi = -w_\xi$ tends to zero near these values. Hence, the correct boundary conditions to the solution $J(u)$ of Equation (8), which are compatible with the desired traveling wave front solution (3)–(5), look like

$$J(0) = J(1) = 0. \qquad (9)$$

Next, Zeldovich and Frank-Kamenetsky assumed the solution to Equation (8) to have the form of a quadratic parabola $J = -\alpha u(u-1)$ (α is a positive constant to be determined), which satisfies conditions (9) automatically. Substituting this expression into (8) and performing the cancellation yields an equation of the sort $P(\alpha, v)u + Q(\alpha, v) = 0$, where $P(\alpha, v)$ and $Q(\alpha, v)$ are set by calculations. Here, u can take any value belonging to the segment [0, 1]. Fixing the variable u on its limit value $u = 0$ yields the equation $Q(\alpha, v) = 0$; taking into account the latter and setting $u = 1$ yields the equation $P(\alpha, v) = 0$. Solving these two equations with respect to α and v leads directly to Equation (6) and to the desired expression $J = -\sqrt{K/2}u(u-1)$. Substituting the latter into (7a) and taking the integral (at the condition $u(0) = \frac{1}{2}$) yields the desired profile of the traveling-wave front:

$$\begin{aligned} u &= 1/\left[1 + \exp\left(\sqrt{K/2}\xi\right)\right] \\ &= (1/2)\left[1 - \tanh\left(\sqrt{K/8}\xi\right)\right]. \end{aligned} \qquad (10)$$

To appreciate the significance of this result in the context of current knowledge, we should stress that the problem of finding a traveling-wave solution to the parabolic reaction-diffusion equations dramatically differs from that arising in the case of nonlinear hyperbolic equations. The latter correspond to conservative physical systems and usually possess the first integrals of the kind of integrals describing energy conservation. Such equations have a Hamiltonian structure, which helps to integrate them analytically using different powerful methods (for example, by methods based on the inverse scattering problem). But the ZF equation is dissipative rather than conservative. Thus it is not Hamiltonian: it describes a gradient physical system that shows the dissipation of its free energy during its time evolution. In these circumstances, the analytical integration of reaction-diffusion equation (1) with the arbitrary kinetic function was a challenge. Zeldovich and Frank-Kamenetsky were the first to have recognized the integrable case of this equation and presented its nontrivial solution.

We should emphasize a distinction of the ZF equation from Fisher's equation, which was investigated in 1937, one year before Zeldovich and Frank-Kamenetsky's work.

The form of the Fisher equation is identical to (1), but the corresponding kinetic function is a quadratic polynomial, which possesses only two zeros. Of course, these zeros correspond to two stationary states of an active medium, but only one of them is a stable one, whereas the second is unstable. As a consequence, the Fisher equation admits not only one traveling-wave front solution, but also a continuum of such waves, and either of them is very responsive to the initial conditions. This equation is applicable only to those media in which the processes of spontaneous production of substances occur against the unstable background state.

In contrast to the Fisher equation, the ZF equation describes active media, which possess two stable states, separated by a third, unstable, state playing the role of a threshold of excitation. The natural field of application of the ZF equation covers the class of bistable active media displaying threshold properties. The linear stability analysis, which was first carried out by Zeldovich and Barenblatt in 1959, and subsequent nonlinear stability analyses performed independently by Lingren and Buratti and by Maginu show that traveling-wave fronts propagating in such media are stable (Scott, 2002).

After Zeldovich and Frank-Kamenetsky's work, decades passed before new analytical traveling-wave front solutions to the ZF-like reaction-diffusion equations appeared. They were constructed with the use of different representations of three-zero kinetic function including the piecewise linear and sinusoidal approximations (Scott, 2002, 2003). The significance of these solutions is predetermined by the fact that all of them describe various physical, chemical, and biological phenomena, which, at first glance, have no common ground. Among these are:

- Electric signals propagating along bistable transmission lines of nerve fibers and neuristors (Scott, 2002, 2003);
- Thermal waves switching boiling regimes from nucleate to film boiling near one-dimensional fuel elements (Zhukov et al., 1980);

- Waves of resistance modification in normal metals (Barelko et al., 1983) and superconductors (Gurevich & Mints, 1984), caused by thermal change;
- Gene flows and population waves in spatially distributed biological populations (Svirezev & Pasekov, 1990) and
- Nonlinear processes arising in synergetics (Loskutov & Mikhailov, 1996).

In view of these applications, the importance of Zeldovich and Frank-Kamenetsky's result is established, yet the destiny of their paper of 1938 is strange. The result obtained in it was used in the Soviet Union for processing experimental data on chemical kinetics even before World War II. After the war, when rapid development of research in the field of physiology of nervous impulses and nonlinear physical chemistry took place, Equation (6) for the velocity of a traveling-wave front became familiar to a broad audience of researchers and appeared frequently in papers and monographs. But the manuscript from which this formula was derived for the first time seemed to have been forgotten: the paper has not been cited until recently! Surprisingly, it is absent even in the two-volume edition of Zeldovich's selected works issued in Russia in 1984, in the lifetime of their author. However, as Mikhail Bulgakov has written (in his classic *Master and Margarita*): "manuscripts do not burn." One could add: even if they are devoted to the theory of combustion.

O.A. Mornev

See also **Diffusion; Flame front; Gradient system; Nerve impulses; Reaction-diffusion systems**

Further Reading

Barelko, V.V., Beibutian, V.M., Volodin, Yu.E. & Zeldovich, Ya.B. 1983. Thermal waves and non-uniform steady states in a $Fe + H_2$ system. *Chemical Engineering Science*, 38(11): 1775–1780

Frank-Kamenetskii, D.A. 1969. *Diffusion and Heat Transfer in Chemical Kinetics*, New York: Plenum Press (original Russian editions, 1947, 1967, 1987)

Gurevich, A.V. & Mints, R.G. 1984. Localized waves in inhomogeneous media. *Soviet Physics – Uspekhi,* 27(1): 19–41

Loskutov, A.Yu. & Mikhailov, A.S. 1996. *Foundations of Synergetics*, 2nd edition, Berlin: Springer (original Russian edition 1990)

Scott, A.C. 2002. *Neuroscience: A Mathematical Primer*, Berlin and New York: Springer

Scott, A.C. 2003. *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, 2nd edition, Oxford and New York: Oxford University Press

Svirezhev, Yu.M. & Pasekov, V.P. 1990. *Fundamentals of Mathematical Evolutionary Genetics*, Dordrecht: Kluwer (original Russian edition 1982)

Zeldovich, Ya.B. & Frank-Kamenetsky, D.A. 1938. K teorii ravnomernogo rasprostraneniya plameni [On the theory of uniform propagation of flame]. *Doklady Akademii Nauk SSSR*, 19(9):693–697

Zhukov, S.A., Barelko, V.V. & Merzhanov, A.G. 1980. Wave processes on heat generating surfaces in pool boiling. *International Journal of Heat Mass Transfer*, 24: 47–55

ZENER DIODE

See **Diodes**

ZERO FIELD STEPS

See **Josephson junctions**

ZERO-DISPERSION LIMITS

The Korteweg–de Vries (KdV) equation with small dispersion

$$\begin{cases} u_t - uu_x + \varepsilon^2 u_{xxx} = 0, \quad t, x \in R, \\ u(x, t = 0; \varepsilon) = u_0(x), \end{cases} \tag{1}$$

is a model for the formation and propagation of dispersive shock waves in one dimension. Let $u(x, t; \varepsilon)$ denote the solution of Cauchy problem (1), where the initial data $u_0(x)$ is smooth and decreases at infinity sufficiently fast. It is known that for $\varepsilon > 0$, no matter how small, the solution of (1) remains smooth for all $t > 0$. For $\varepsilon = 0$, (1) becomes the Cauchy problem for the Hopf equation

$$\begin{cases} u_t - uu_x = 0, \\ u(x, t = 0) = u_0(x). \end{cases} \tag{2}$$

The solution of the Hopf equation can be obtained by the method of characteristics. If the initial data $u_0(x)$ is somewhere increasing, the solution $u(x, t)$ of Equation (2) always has a point (x_c, t_c) of gradient catastrophe where an infinite derivative develops.

After the time of gradient catastrophe t_c, the solution $u(x, t, \varepsilon)$ of (1) develops in the neighborhood of x_c an expanding region filled with rapid modulated oscillations of wavelength of order $1/\varepsilon$. These oscillations are called dispersive shock waves.

Lax and Levermore (1998), performing the zero-dispersion asymptotics for the inverse-scattering problem for the KdV equation, showed that as ε tends to zero, $u(x, t; \varepsilon)$ tends uniformly to the smooth solution $u(x, t)$ of (2) as long as $t < t_c$. For $t > t_c$, the solution $u(x, t; \varepsilon)$ converges weakly in the oscillation region to a limit $\bar{u}(x, t)$ that is not a solution of conservation law (2).

The first example describing dispersive shock waves was proposed by Gurevich and Pitaevski (1973). Their description was rigorously proved by Venakides (1990) who derived the general form of the rapid oscillations. The oscillation zone is approximately described for a short time $t > t_c$ by a modulated periodic wave solution of the KdV equation:

$$u(x,t,\varepsilon) \simeq V(x,t) + \frac{1}{3}\alpha(x,t) + \Theta\left(\frac{x - Vt + \phi}{\varepsilon}\right), \tag{3}$$

$$V(x,t) = \tfrac{1}{6}(u_1(x,t) + u_2(x,t) + u_3(x,t)) .$$

In the above formula, the term $V(x,t) + \frac{1}{3}\alpha(x,t)$ is the weak limit $\bar{u}(x,t)$ of $u(x,t,\varepsilon)$ as $\varepsilon \to 0$, while the remaining term describes the rapid oscillations. The function Θ is 2π-periodic with zero average, and it can be expressed in terms of elliptic functions. The quantity α defined below and the phase ϕ depend on some functions $u_i(x,t)$, $i = 1,2,3$. The functions $u_1(x,t) > u_2(x,t) > u_3(x,t)$ solve the Whitham (1974) modulation equations

$$\partial_t u_i(x,t) - \lambda_i(u_1,u_2,u_3)\partial_x u_i(x,t) = 0, \quad i = 1,2,3 , \tag{4}$$

where

$$\lambda_i(u_1,u_2,u_3) = \frac{1}{3}(u_1 + u_2 + u_3) + \frac{2}{3}\frac{\prod_{j\neq i, j=1}^{3}(u_i - u_j)}{u_i + \alpha}, \tag{5}$$

$$\alpha = -u_3 + (u_3 - u_1)\frac{E(s)}{K(s)} \tag{6}$$

and $K(s)$ and $E(s)$ are the complete elliptic integrals of the first and second kind with modulus $s = (u_2 - u_1)/(u_3 - u_1)$.

The solution $u_1(x,t) > u_2(x,t) > u_3(x,t)$ of the Whitham equations can be plotted in the (x,u) plane as branches of a multivalued function. The solutions of the Hopf equation and the Whitham equations are connected to one another as illustrated in Figure 1(a). The function $u_2(x,t)$ can vary from $u_3(x,t)$ to $u_1(x,t)$. On the $(x, t \geq 0)$ plane, the oscillation region is bounded on one side by the curve $x^-(t)$ where $u_2(x,t) = u_1(x,t)$, and on the other side by the curve $x^+(t)$ where $u_2(x,t) = u_3(x,t)$ (see Figure 1(a)). For $x^-(t) < x < x^+(t)$, the solution of (1) for small ε is approximately given by (3) while outside the interval $[x^-(t), x^+(t)]$ is given by the solution $u(x,t)$ of the Hopf equation (2). At edge $x = x^-(t)$ of the oscillation region, the amplitude of the oscillations vanishes and (3) goes to $u(x,t,\varepsilon)|_{u_1=u_2} \simeq u_3(x,t)$. When $x = x^+(t)$, solution (3) goes to the one-soliton solution of the KdV equation. In general, the oscillation zone grows with time. For generic analytic initial data with a cubic inflection point, the growth in the (x,t) plane of the oscillation zone near the point of gradient catastrophe (x_c, t_c) is described, up to shifts and rescaling, by the semi-cubic law

$$x^{\pm}(t) = x_c \pm a^{\pm}(t - t_c)^{3/2},$$

where $a^{\pm}$ are two positive numbers. A completely different behavior appears in the zero-diffusion case. The simplest equation that combines nonlinearity and diffusion is the Burgers equation

$$u_t - uu_x = \varepsilon u_{xx} , \tag{7}$$

where $\varepsilon > 0$. For smooth initial data $u_0(x)$, the Burgers equation can be integrated through the Cole–Hopf transformation to

$$u(x,t,\varepsilon) = -\frac{\int_{-\infty}^{\infty} \frac{x-\xi}{t} e^{-\frac{G(\xi)}{2\varepsilon}}\, d\xi}{\int_{-\infty}^{\infty} e^{-\frac{G(\xi)}{2\varepsilon}}\, d\xi}, \tag{8}$$

where

$$G(\xi) = -\int_0^{\xi} u_0(\eta)\, d\eta + \frac{(x-\xi)^2}{2t}.$$

The behavior of the exact solution (8) as $\varepsilon \to 0$ can be obtained by observing that the dominant contributions

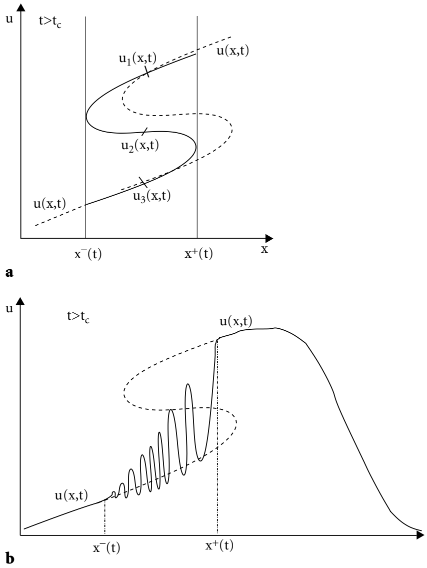

Figure 1. (a) The dashed line represents the formal solution of the Hopf equation; the continuous line represents the solution of the Whitham equations. The solution $(u_1(x,t), u_2(x,t), u_3(x,t))$ of the Whitham equations and the position of the boundaries $x^-(t)$ and $x^+(t)$ are to be determined from the conditions $u(x^-(t),t) = u_3(x^-(t),t)$, $u(x^+(t),t) = u_1(x^+(t),t)$, where $u(x,t)$ is the solution of the Hopf equation. (b) The oscillations in the region $x^-(t_1) < x < x^+(t_1)$.

to the integrals in (8) come from the neighborhood of the stationary points of G where

$$\frac{\partial G}{\partial \xi} = -u_0(\xi) - \frac{x-\xi}{t} = 0. \tag{9}$$

If (9) has only one stationary point, by the application of the steepest descent method, the asymptotic solution $u(x, t; \varepsilon)$ as $\varepsilon \to 0^+$ converges strongly to

$$u(x, t) = u_0(\xi), \quad x = \xi - u_0(\xi)t. \tag{10}$$

The above is exactly the solution of the Cauchy problem for the Hopf equation (2). The stationary point $\xi(x, t)$ of (9) becomes the characteristic variable in (10). For bump-like initial data the solution of the Hopf equation (2) has a point of gradient catastrophe $(x_\mathrm{c}, t_\mathrm{c})$. After the time $t = t_\mathrm{c}$ of gradient catastrophe, (10) gives a multivalued solution: the characteristics of the Hopf equation begin to intersect. For a typical initial pulse, there are usually three characteristics that intersect at each point of the multivalued region; that is, (9) has three solutions $\xi_1(x, t), \xi_2(x, t), \xi_3(x, t), \xi_1 > \xi_2 > \xi_3$. The dominant behavior of the solution of the Burgers equation will be given by the following contributions:

$$u \simeq -\frac{\sum_{i=1}^{3} \frac{x-\xi_i}{t} |G''(\xi_i)|^{-\frac{1}{2}} \mathrm{e}^{-\frac{G(\xi_i)}{2\varepsilon}}}{\sum_{i=1}^{3} |G''(\xi_i)|^{-\frac{1}{2}} \mathrm{e}^{-\frac{G(\xi_i)}{2\varepsilon}}}. \tag{11}$$

Let us suppose that for $x_\mathrm{c} < x < x_\mathrm{s}$ and $t > t_\mathrm{c}$, the function $G(\xi_1(x, t))$ is less than $G(\xi_2(x, t))$ and $G(\xi_3(x, t))$. Then the above expression for u in the limit $\varepsilon \to 0$ reads

$$u \simeq -\frac{x-\xi_1}{t}, \quad x_\mathrm{c} < x < x_\mathrm{s}, \tag{12}$$

while assuming that for $x > x_\mathrm{s}$ $G(\xi_2(x, t)) < G(\xi_1(x, t)), G(\xi_3(x, t))$, we have

$$u \simeq -\frac{x-\xi_2}{t}. \tag{13}$$

In each case, (10) applies to both ξ_1 and ξ_2. Therefore, the solution of the Burgers equation converges as $\varepsilon \to 0^+$ to the solution of the Hopf equation (2) almost everywhere except at the points (x, t) where $G(\xi_i(x, t)) = G(\xi_j(x, t)), \ i \neq j, \quad i, j = 1, 2, 3.$ For example, in the case treated above, the change over from ξ_1 to ξ_2 occurs when $x = x_\mathrm{s}$ where $G(\xi_1) = G(\xi_2)$. Near $x = x_\mathrm{s}$ the solution of the Burgers equation as $\varepsilon \to 0^+$ has a transition from (12) to (13) which is called a shock wave. In other words, the solution of the Burgers equation in the zero viscosity limit is given by two different branches of the solution of the Hopf equation joined by a jump at the point x_s. The condition $G(\xi_1) = G(\xi_2)$ reads

$$\begin{aligned} &-\int_0^{\xi_1} u_0(\eta)\,\mathrm{d}\eta + \frac{(x-\xi_1)^2}{2t} \\ &\quad = -\int_0^{\xi_2} u_0(\eta)\,\mathrm{d}\eta + \frac{(x-\xi_2)^2}{2t}. \end{aligned}$$

Because of (10), the above relation is equivalent to

$$\frac{1}{2}(u_0(\xi_1) + u_0(\xi_2)) = \frac{1}{\xi_1 - \xi_2} \int_{\xi_2}^{\xi_1} u_0(\eta)\,\mathrm{d}\eta, \tag{14}$$

which describes the shock wave. Since the shock occurs at $x = x_\mathrm{s}(t), t > t_c$, we also have

$$x_\mathrm{s}(t) = \xi_1 - u_0(\xi_1)t, \quad x_\mathrm{s}(t) = \xi_2 - u_0(\xi_2)t.$$

The above three equations determine the functions $x_\mathrm{s}(t)$, $\xi_1(t)$, and $\xi_2(t)$. The values of $u(x, t)$ on the two sides of the shock are $u^-(x, t) = u_0(\xi_1(x, t))$ and $u^+(x, t) = u_0(\xi_2(x, t))$. The shock speed can be derived by taking the time derivative of the above two equations and reads

$$\frac{\mathrm{d}x_\mathrm{s}(t)}{\mathrm{d}t} = -\frac{1}{2}(u_0(\xi_1) + u_0(\xi_2)).$$

Comparison of the above relation with (14) shows that the modulus of the shock speed is equal to the average value of the characteristics velocity $u_0(\eta)$ over the interval $[\xi_1, \xi_2]$.

While the zero-dispersion limits have been studied only for integrable equations such as the KdV or the nonlinear Schrödinger equation, the zero-viscosity limits have been studied for the parabolic equation of the form

$$u_t + [f(u)]_x = \varepsilon u_{xx}, \quad u \in R^n, \ (t, x) \in R \times R^m.$$

The scalar case in several spatial dimensions was investigated by Kruzhkov (1970). The two-component case in one spatial dimension has been studied by DiPerna (1983), while the n-component case in one spatial dimension has been investigated by Bressan (2002).

TAMARA GRAVA

See also **Burgers equation; Constants of motion and conservation laws; Inverse scattering method or transform; Jump phenomena; Modulated waves; Shock waves**

Further Reading

Bressan, A. 2002. Hyperbolic systems of conservation laws in one space dimension. In *Proceedings of the International Congress of Mathematicians,* Beijing, vol. I, Beijing: Higher Education Press, pp. 159–178

DiPerna, R. 1983. Convergence of approximate solutions to conservation laws. *Archive for Rational Mechanics and Analysis*, 82: 27–70

Gurevich, A.G. & Pitaevskii, L.P. 1973. Non-stationary structure of a collisionless shock waves. *JEPT Letters*, 17: 193–195

Kamvissis, S., McLaughlin, K.D.T.-R. & Miller, P.D. 2003. *Semiclassical Soliton Ensembles for the Focusing Nonlinear Schrödinger Equation*, Princeton, NJ: Princeton University Press

Kruzhkov, S. 1970. First order quasi-linear equations with several space variables. *Mathematics of the USSR Sbornik*, 10: 217–243

Lax, P.D. & Levermore, C.D. 1983. The small dispersion limit of the Korteweg de Vries equation, I, II, III. *Communications in Pure and Applied Mathematics*, 36: 253–290, 571-593, 809–830

Novikov, S., Manakov, S.V., Pitaevski, L.P. & Zakharov, V.E. 1984. *Theory of Solitons: The Inverse Scattering Method*, New York: Consultants Bureau

Venakides, S. 1990. The Korteweg–de Vries equations with small dispersion: higher order Lax–Levermore theory. *Communications in Pure and Applied Mathematics*, 43: 335–361

Whitham, G.B. 1974. *Linear and Nonlinear Waves*, New York: Wiley

ZETA FUNCTIONS

See **Randum matrix theory II**

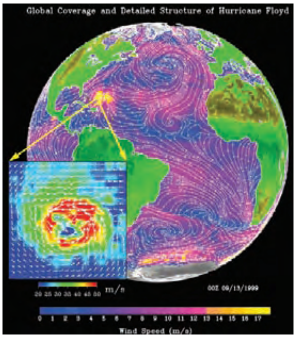
Global Coverage and Detailed Structure of Hurricane Floyd
m/s
00Z 09/13/1999
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Wind Speed (m/s)

Synoptic view of surface winds over the Atlantic Ocean on September 13, 1999. The background colors indicate wind speeds and the white arrows show the direction of the wind. The high winds of Hurricane Floyd can be seen east of the Bahamas. The images are obtained from the NASA/NOAA sponsored data system Seaflux, at JPL through the courtesy of W. Timothy Liu and Wenqing Tang. (*See* **Atmospheric and ocean sciences; Hurricanes and tornadoes; Vortex dynamics in fluids**.)

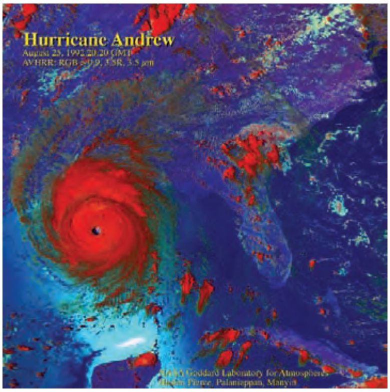
Hurricane Andrew
August 25, 1992 20:20 GMT
AVHRR: RGB = 0.9, 3.5R, 3.5 μm
Goddard Laboratory for Atmospheres

Overhead view of Hurricane Andrew on 25 August 1992 at 20:20 Universal Time (UT). Different wavelengths are assigned to the red, green, and blue channels to make the hurricane appear red. Courtesy of National Oceanic and Atmospheric Administration. (*See* **Atmospheric and ocean sciences; Hurricanes and tornadoes; Vortex dynamics of fluids**.)

Wake Vortex Study at Wallops Island
NASA Langley Research Center
5/4/1990
Image # EL-1996-00130

Aircraft wake vortex visualized by colored smoke. NASA's Wake Vortex Study at Wallops Island, 1990. Courtesy of NASA Langley Research Center. (*See* **Fluid dynamics; Vortex dynamics in fluids.**)

The planet Jupiter showing the Great Red spot, photographed by the NASA Cassini satellite on December 29, 2000. Courtesy NASA/JPL/ Space Science Institute. (*See* **Fluid dynamics; Jupiter's Great Red spot; Vortex dynamics of fluids.**)

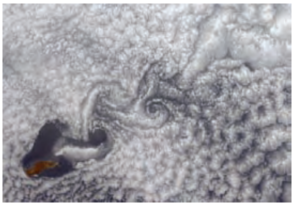

An atmospheric flow pattern known as the von Karman vortex street—a linear chain of spiral eddies. The vortex pattern is made visible by the marine stratocumulus clouds around Guadalupe Island, which is about 22 miles long and lies about 250 miles southwest of San Diego. Multi-angle Imaging Spectroradiometer (MISR) images from June 11, 2000: NASA/ GSFC/JPL, MISR Team. (*See* **Fluid dynamics; Turbulence; Vortex dynamics of fluids**.)

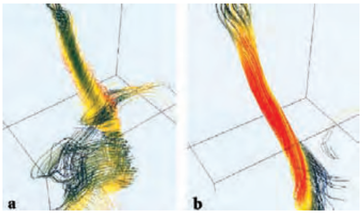

Close-up view of coherent vortex structure from direct numerical simulation of homogeneous isotropic turbulence: (a) streamlines, and (b) vortex lines—color-coded according to intensity—forming a tangle of space curves in a reference box (tropicity domain). Courtesy of Z.-S. She, E. Jackson and S.A. Orszag, Princeton University, Princeton, NJ, USA. (*See* **Structural complexity; Turbulence**.)

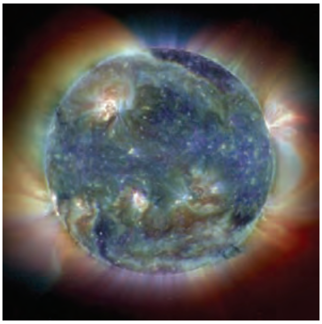

The Sun, imaged in three color-coded wavelengths (171, 195 and 284 angstrom), showing solar flares (sun spots) comprising coronal mass ejections which launch solar winds. Courtesy of SOHO/Extreme Ultraviolet Imaging Telescope (EIT) consortium. SOHO is a project of international cooperation between ESA and NASA. (*See* **Alfvén waves; Nonlinear plasma waves**.)

Pumice and ash erupting from Mount St. Helens on May 18, 1980 is a powerful example of turbulent dispersion of small particles. The plume eventually reached 12 to 15 miles (20–25 km) above sea level. Photograph by Donald A. Swanson, USGS/CVO. (*See* **Evaporation waves; Plume dynamics**.)

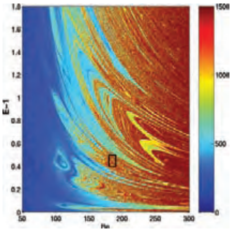

Lifetimes of perturbations in a planar shear flow for different values of initial energy (E) and Reynolds number (Re). In the transition region the lifetimes show a sensitive dependence on initial conditions and huge fluctuations. Similar behavior has been observed in plane Couette flow and in pipe flow. (*See* **Fluid dynamics; Shear flow**.)

Red and green colors predominate in this view of the Aurora Australis photographed from the Space Shuttle in May 1991. Long a source of myth, auroras are known from late nineteenth-century studies by Kristian Birkeland to be driven by solar wind and stabilized by Earth's magnetosphere. Photograph by the STS-39 Crew, courtesy of Earth Sciences and Image Analysis Laboratory, NASA Johnson Space Center. (*See* **Alfvén waves; Nonlinear plasma waves**.)

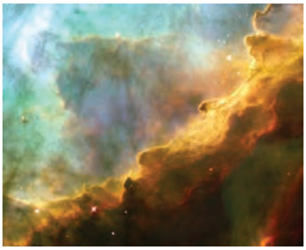

Turbulent clouds of hydrogen gas mix with small amounts of oxygen, sulfur, and other elements in this Hubble Space Telescope image of a region of M17, a nebula about 5,500 light-years away that is giving birth to new stars. Ultraviolet energy from hot, young stars in the nebula causes the gas to glow, and the intense heat and pressure cause turbulence. The different colors represent different gases. Courtesy of NASA/ESA/J. Hester (ASU). (*See* **Galaxies**.)

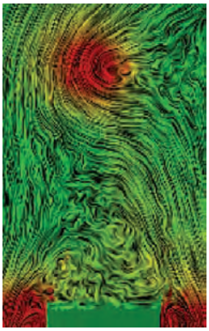

A slice from a three-dimensional numerical simulation of turbulent flow around a square cylinder (visible at the bottom of the image). The data is visualized using the spot noise technique to show flow structure, color shows the pressure. Simulation by R. Verstappen and A. Veldman of the University of Groningen in the Netherlands, and visualization by W. de Leeuw and R. van Liere of the Center of Mathematics and Computer Science in Amsterdam. (*See* **Fluid dynamics; Turbulence**.)

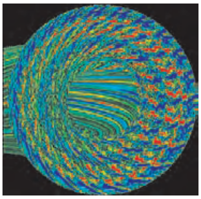

Snapshot of the results of a simulation for a toroidal magnetic confinement system called a tokamak. Contours of **E**×**B** reveal ballooning turbulence. Turbulence is on a relatively short scale perpendicular to the magnetic field and very extended along the magnetic field lines. The eddies tend to point preferentially in the direction of positive (bad) magnetic curvature. Photograph courtesy of Gary Kerbel, National Energy Research Supercomputer Center. (*See* **Nonlinear plasma waves**.)

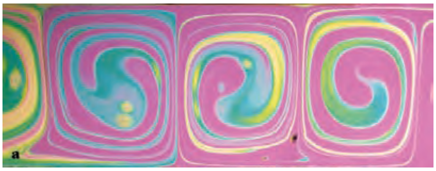

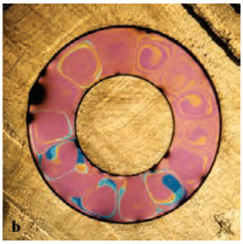

Electrically driven convection in a smectic (the most ordered phase) liquid crystal film. (a) A nonuniformly thick film is freely suspended between two wires about 1mm apart. The film itself is less than 1 micron thick, and shows bright interference colors under reflected white light, like a soap bubble. The film convects when a DC voltage of a few volts is applied to the wires. The vortex flow is visualized as it advects the thickness variations. The driving mechanism involves the field coupling to the surface charge which develops on the film's two free surfaces. This system forms a simple one-dimensional nonlinear pattern which has been quantitatively studied in films with perfectly uniform thickness, for which no color variation is visible. (b) As in the previous figure, now with annular electrodes and a radial driving force. One can also rotate the inner electrode and thus study convection superposed on a two-dimensional circular Couette flow. Images courtesy of Stephen W. Morris, University of Toronto. (*See* **Liquid crystals; Spatiotemporal chaos; Thermal convection**.)

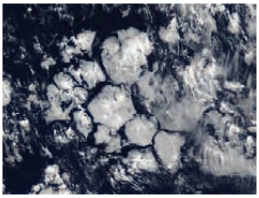

Closed Rayleigh–Bénard convection cells in clouds over the South Atlantic Ocean. NASA/GSFC. (*See* **Atmospheric and ocean sciences; Fluid dynamics; Pattern formation; Thermal convection**.)

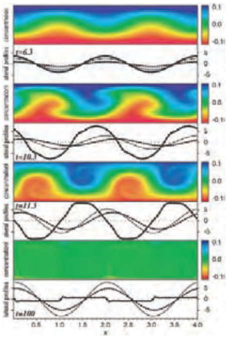

Evolution of oscillatory convection in a binary fluid layer heated from below. The concentration distribution in a vertical cross section of the fluid layer is displayed by color coded plots with blue and red denoting high and low concentration, respectively. Wave profiles at midheight are shown for the fields of vertical velocity (thin lines), temperature (lines with triangles), and concentration (lines with squares). Initially a standing wave grows ($t = 6.3$). It breaks at $t = 10.3$ and is thereby transformed into a fast traveling wave propagating to the left which then slows down. For better visibility two wavelengths are shown. (*See* **Fluid dynamics; Rayleigh–Taylor instability; Thermo-diffusion effects**.)

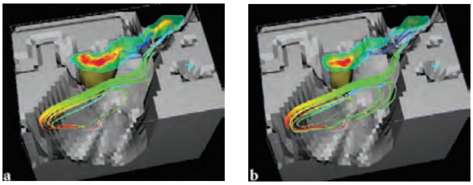

Numerical results from a coupled atmosphere-ocean model of the North Atlantic circulation including bottom topography (a) now and (b) the prediction for 100 years from now. Noticeable are the divergence of the Gulf Stream in the North Atlantic, reduction in the production of deep cold water masses near Greenland, and slowing of part of the global conveyor belt. Relevant streamlines of the general circulation are colored according to their temperature (red=warm, blue=cold). Courtesy of the Deutsches Klimarechenzentrum (Michael Böttinger, DKRZ, Hamburg, Germany). (*See* **Atmospheric and ocean sciences**.)

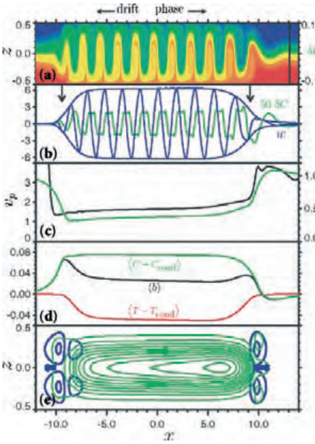

Spatially confined convection in binary fluid mixtures in the form of a localized traveling wave. (a) Concentration deviation (δC) from global mean (light green/yellow) in a vertical cross section of the layer. (b) Lateral wave profiles at midheight, $z = 0$, of δC (green), vertical velocity w (blue), and its envelope. (c) Phase velocity v_p (black) and mixing number M (green) measuring rms deviations of concentration. (d) Time averaged deviations from the conductive state at $z = -0.25$ for concentration (green), temperature (red), and their sum ($\langle b \rangle$) measuring the convective contribution to the buoyancy. (e) Streamlines of time averaged concentration current $\langle \boldsymbol{J} \rangle$ (green) and velocity field $\langle \boldsymbol{u} \rangle$ (blue). Thick blue and green arrows indicate $\langle \boldsymbol{u} \rangle$ and transport of positive δC (alcohol surplus), respectively. Thus, in the lower half of the layer negative δC (water surplus) is transported to the right. (*See* **Fluid dynamics; Thermo-diffusion effects**.)

Snowflakes on a window pane. Ice crystals are now recognized as examples of nonlinear pattern formation. (*See* **Growth patterns; Pattern formation**.)

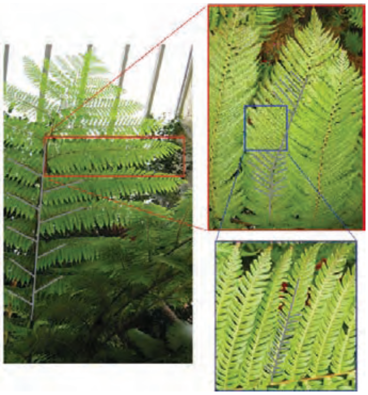

Self similarity of a fern frond. There is a limit to the range of scales at which the self similarity is maintained, and it occurs at only a few discrete scales. Courtesy of Paul Bourke. (*See* **Fractals; Turing patterns**.)

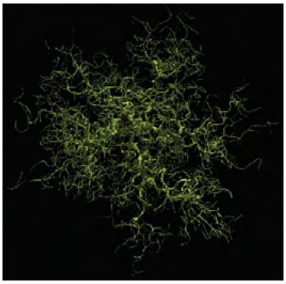

Computer simulation of growth in three dimensions by diffusion-limited aggregation. Such growth patterns are natural fractals. Courtesy of Paul Bourke. (*See* **Cluster coagulation; Fractals; Growth patterns**.)

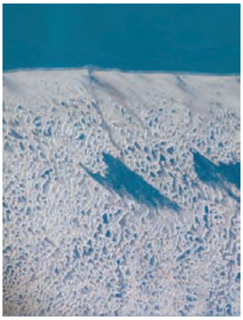

Dune fields with a regular pattern, Lençóis Maranhenses National Park, on Brazil's north coast. Image courtesy of Earth Sciences and Image Analysis Laboratory, NASA Johnson Space Center, ISS007-E-15177. (*See* **Dune Formation; Geomorphology and tectonics; Pattern formation**.)

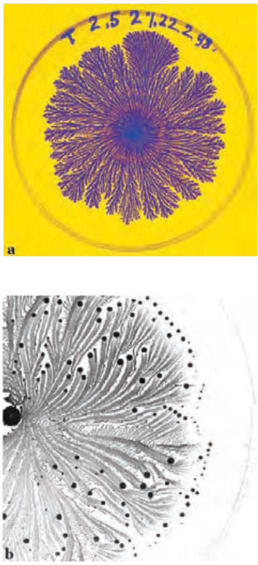

Bacterial colonies develop complex spatiotemporal patterns in response to adverse conditions such as nutrient starvation. (a) Branching pattern in a Petri dish of *Paenibacillus dendritiformis*. Growth rate is limited by diffusion of nutrients towards the colony, giving a branching or fractal pattern. (b) Vortex branching exhibited by *Paenibacillus vortex* bacteria. At the dark dot at the tip of each of the branches, millions of bacteria rotate in a highly organized vortex. Chemorepellents push the vortices outwards, allowing the colony to expand. Images courtesy of Eshel Ben Jacob, Tel Aviv University. (*See* **Cluster coagulation**; **Growth patterns**.)

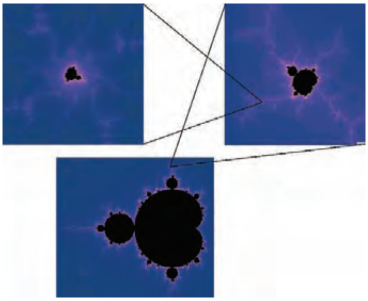

Approximate self similarity in the Mandelbrot set. The Mandelbrot set is the set of all c for which the iteration $z \rightarrow z^2+c$, starting from $z=0$, does not diverge to infinity. This set is named after the mathematician Benoit Mandelbrot, who discovered their fractal and self replicating structure. Three successive magnifications are shown, and at each level a structure similar, but not exactly the same, as the entire set is observed. Courtesy of Paul Bourke. (*See* **Fractals; Multifractal analysis**.)

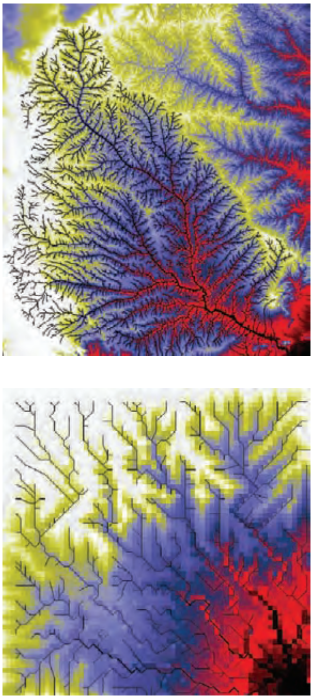

Drainage networks have scale-invariant (fractal) symmetry. Top: Color image of elevation (white/yellow = high elevation black/red = low elevation) and drainage network (black lines) for a river basin in the Loess Plateau region of the Shanxi Province, P.R.C. River basins in homogeneous material such as loess (a silty soil deposited by wind storms) form particularly symmetric and ordered networks. Bottom: Color image of elevation for a fluvial landscape evolution simulation. Courtesy of Jon Pelletier, University of Arizona. (*See* **Geomorphology and tectonics**.)

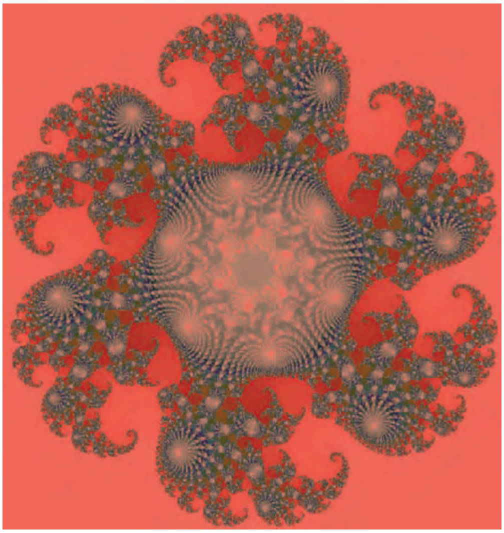

The elephants image is a Julia set with an unusual initialization scheme. The image plane is mapped to the 3rd power of the pixel coordinate. This has the effect of increasing symmetrical order. Maximim iterations/ pixel = 1024, center coordinates = $(+3.4861\times10^{-14}, -4.88498\times10^{-15})$, magnification = 0.9662319. The Julia set coordinate, p1 = (0.28767393, 0.01500042), $z = \text{pixel}^3$: $z = z^z + p1$, ($|z| \leq 4$.) Image courtesy of Noel Giffin. (*See* **Maps in the complex plane.**)

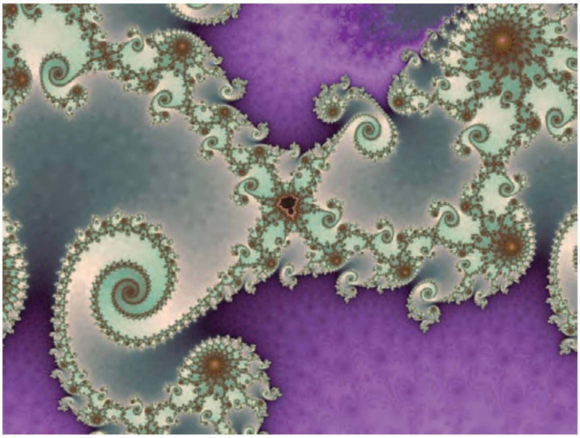

A classic Mandelbrot image created with high iteration limits. Maximum iterations/pixel = 300,000, minimum iterations/pixel = 9587, center coordinates = (+0.37873904887513700, +0.22838619041160610), magnification = 8.279722×10^{10}. High iteration fractals can reveal striking fractal detail as the iteration count and zoom depth increase. A logarithmic coloring algorithm is used to distribute the color pallette over the large iterative range. Image courtesy of Noel Giffin. (*See* **Maps in the complex plane.**)

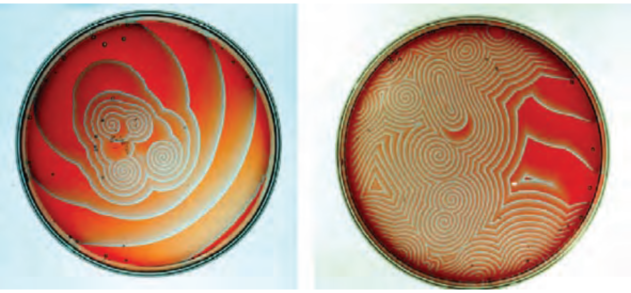

Spiral wave patterns in a BZ reaction, confined to a thin layer in a petri dish. Early on in the reaction, the fronts of the target pattern are far apart, but as the reaction proceeds they become closer and additional wave sources appear, producing a complicated pattern. Courtesy of S.K. Scott, University of Leeds. (*See* **Belousov–Zhabotinsky reaction; Chaotic dynamics; Reaction-diffusion systems**.)

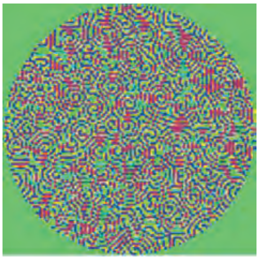

Simulation of spiral defect chaos in Rayleigh–Bénard convection. Color corresponds to temperature, increasing from violet to red. Local spirals are erratic in time—changing position and size. Courtesy of James D. Gunton, Lehigh University. (*See* **Spatiotemporal chaos; Spiral waves; Thermal convection**.)

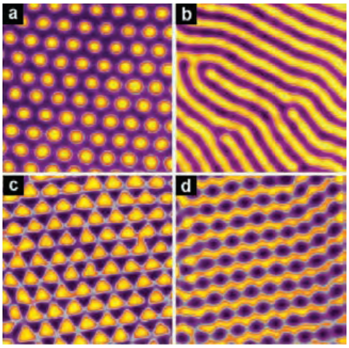

Turing structures of different symmetries obtained with the chlorite-iodide-malonic acid reaction. Dark and light regions respectively correspond to high and low iodide concentration. The wavelength (a function of kinetic parameters and diffusion coefficients) is of the order of 0.2 mm. All patterns in (a)–(d) are at the same scale: view size 1.7 mm × 1.7 mm. Courtesy P. De Kepper, CRPP. (*See* **Turing patterns**.)

A montage of planetary images taken by spacecraft managed by the Jet Propulsion Laboratory in Pasadena, CA. (Pluto is not shown.) The emergence of our planets from a cloud of gas and dust about 4.5 billion years ago was a highly nonlinear event. In his work to solve the three-body problem of planetary motion, Henri Poincaré first formulated the basic concepts of chaotic dynamics. Courtesy of NASA. (*See* ***N*-body problem; Solar system**.)

Re-creation of the famous 1834 first sighting of a soliton or solitary wave on the Union Canal near Edinburgh, 12 July 1995. Copyright Heriot-Watt University, used with permission. (*See* **Solitons, a brief history.**)

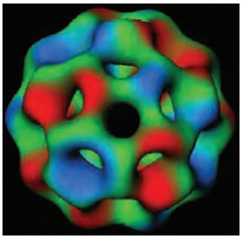

The baryon density isosurface of a 17-Skyrmion (topological charge of 17). The color represents the value of the third component of the pion field. (*See* **Skyrmions**.)

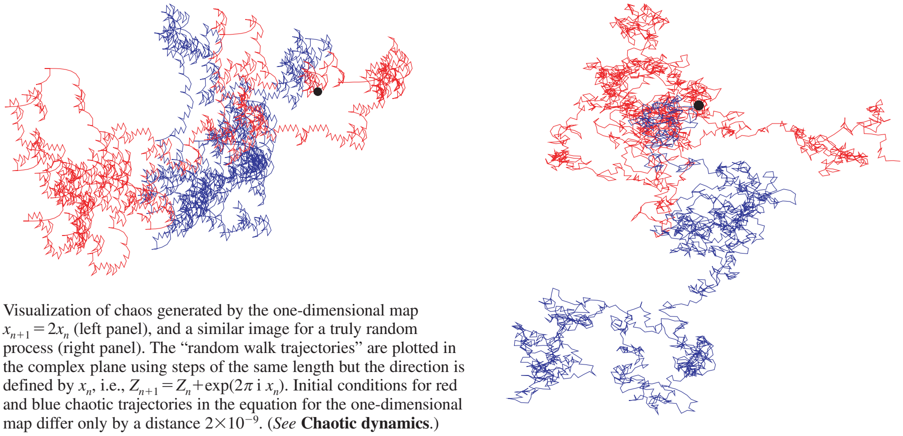

Visualization of chaos generated by the one-dimensional map $x_{n+1} = 2x_n$ (left panel), and a similar image for a truly random process (right panel). The "random walk trajectories" are plotted in the complex plane using steps of the same length but the direction is defined by x_n, i.e., $Z_{n+1} = Z_n + \exp(2\pi\, \mathrm{i}\, x_n)$. Initial conditions for red and blue chaotic trajectories in the equation for the one-dimensional map differ only by a distance 2×10^{-9}. (*See* **Chaotic dynamics**.)

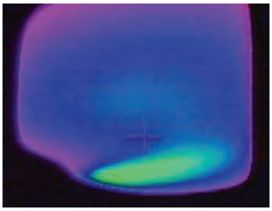

Infrared image of a spin mode in the free-radical-frontal polymerization of an acrylate in a 1.5 cm diameter tube. Image courtesy of J. Pojman. (*See* **Polymerization**.)

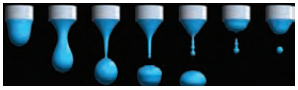

Numerical simulations of a dripping water faucet. The dripping faucet is a familiar example of chaotic dynamics. (*See* **Chaotic dynamics; Dripping faucet**.)

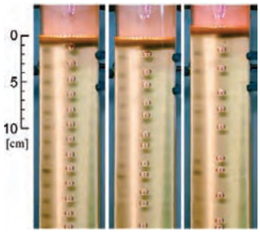

Dripping patterns of water drops falling from a nozzle (7 mm inner diameter and 10 mm outer diameter) into a transparent cylindrical container of salad oil, which sensitively depends on the flow rate. The salad oil was used only to decrease the acceleration of the drops to obtain successive pictures. Examples of period-one (left), period-two (middle) and chaos (right) are presented. (*See* **Dripping faucet**.)

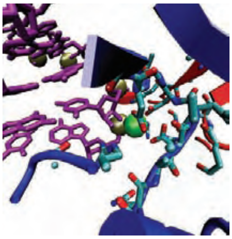

RNA polymerase active site. The DNA template and growing RNA strands are shown in purple with gold phosphorous atoms; the protein backbone is shown in blue; protein atoms are colored red for oxygen, blue-green for carbon, and blue for nitrogen. The green sphere is a magnesium ion, which coordinates both the protein and the nucleotide tri-phosphate which is being added to the growing DNA strand. Atomic coordinates are available using the protein data bank accession code 1I6H. Produced with VMD. (*See* **Protein structure**.)

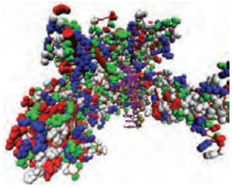

RNA polymerase catalytic subunit with DNA bound at the catalytic site. The nucleic acids are shown in purple, as in active site figure. The protein backbone is shown in red, with amino acid side chains colored according to charge and polarity; negatively charged amino acids are red, positively charged amino acids are blue, non-polar amino acids are white, and polar but uncharged amino acids are green. The patterns of charge, hydrogen bonds, and shape are recognized by other proteins in order to regulate activity of this protein. (*See* **Protein dynamics**.)

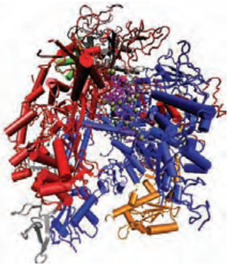

RNA polymerase transcription initiation complex, shown with several of the additional proteins necessary for it to function. Different colors represent different proteins interacting with the catalytic subunit. The interactions are quite specific. (*See* **Protein dynamics**.)

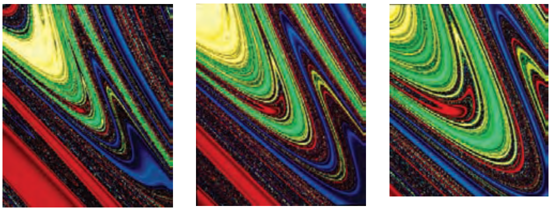

A periodically forced damped pendulum has complex behavior. These computer-generated images show initial positions that map to one of several different behaviors (one color for each). For example, orbits starting at points in the blue region would yield a different type of asymptotic motion than orbits starting in the red region. The brighter the shade of color, the longer it takes to settle into the corresponding motion. The different regions are separated by fractal basin boundaries. The pictures are made at increasing magnification levels. Courtesy, University of Maryland Chaos Group. (*See* **Attractors; Damped-driven anharmonic oscillator; Pendulum**.)

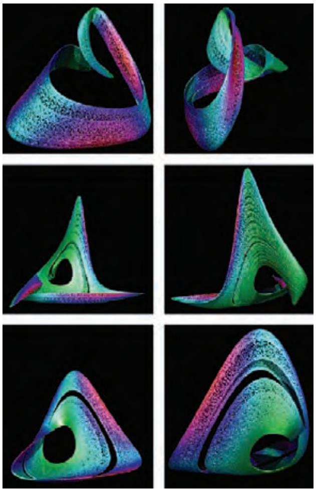

Three chaotic attractors arising from numerical solutions of a spatially homogeneous model of the mammalian electroencephalogram. Each attractor was constructed by delay-embedding a single state variable corresponding to the electroencephalographic observable. Attractors were rendered from two different viewpoints (left and right images), from 100,000 point time series of 10 seconds total duration and colored by velocity. All attractors have a dominant frequency of oscillation in the electroencephalographic alpha band (8–13 Hz). The top attractor has a Lyapunov dimension of 2.086 ± 0.003, representing an upper bound for the correlation (D_2) dimension. Graphics by Paul Bourke, Centre for Astronomy and Supercomputing, Swinburne University of Technology, Melbourne, Australia. (*See* **Electroencephalogram at mesoscopic scales**.)

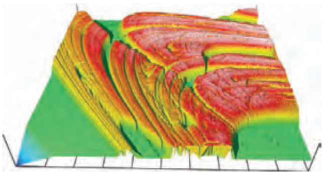

Chaos in a spatially homogeneous version of Liley et al.'s theory of the mammalian electroencephalogram. The two independent axes represent neuronal population input parameters, whereas the dependent variable (indicated by height and color) represents the largest Lyapunov exponent (LLE) of the system dynamics. Each point in this plane can be thought of as indexing a particular attractor (see previous figure). Green represents limit cycle (LLE = 0) activity, whereas points above and below this plane correspond to chaos (LLE > 0) and point attractor dynamics (LLE < 0) respectively. (*See* **Electroencephalogram at mesoscopic scales**.)

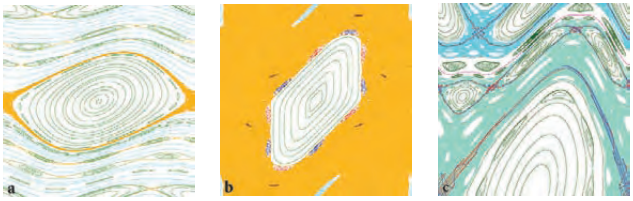

Standard Map (a): Dynamics of the standard map for $k = 0.6$;
(b). Dynamics of the standard map for $k = 2.0$; (c). Dynamics of the standard map for $k = k_{cr}(\gamma) = 0.971635406$. (*See* **Standard map.**)

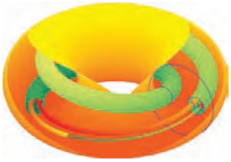

Smale's solenoid, an Axiom-A attractor. At each iteration, the fat torus is mapped into a long, thin torus which is then put back into the original torus, winding around twice. Reprinted from *A First Course in Dynamics*, Hasselblatt & Katok with the permission of Cambridge University Press. (*See* **Anosov and Axiom A systems**.)

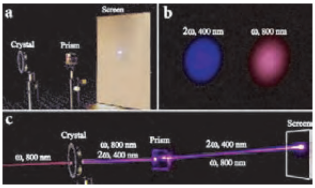

Second harmonic generation in lithium niobate: (a) Experimental setup; (b) Closeup view of the observation screen showing the fundamental and second-harmonic light; (c) Side view showing the partial conversion of the red 800-nm incident beam into the blue 400-nm beam; the separation of the colors by the prism is relatively small in the plane of this photograph. (*See* **Frequency doubling.**)

The Universal Book of Mathematics

From **A**bracadabra to **Z**eno's Paradoxes

David Darling

WILEY
John Wiley & Sons, Inc.

The Universal Book of Mathematics

The Universal Book of Mathematics

From **A**bracadabra to **Z**eno's Paradoxes

David Darling

WILEY
John Wiley & Sons, Inc.

This book is printed on acid-free paper. ♾

Copyright © 2004 by David Darling. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Limit of Liability/Disclaimer of Warranty: While the publisher and the author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor the author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information about our other products and services, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Darling, David J.
The universal book of mathematics : from abracadabra to Zeno's paradoxes / David Darling.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-27047-4 (cloth: alk. paper)
1. Mathematics–Encyclopedias. I. Title.
QA5 .D27 2004
510'.3–dc22 2003024670
Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Mathematics is not a careful march down a well-cleared highway, but a journey into a strange wilderness, where the explorers often get lost. Rigor should be a signal to the historian that the maps have been made, and the real explorers have gone elsewhere.

–William S. Anglin

But leaving those of the Body, I shall proceed to such Recreation as adorn the Mind; of which those of the Mathematicks are inferior to none.

–William Leybourn (1626–1700)

The last thing one knows when writing a book is what to put first.

–Blaise Pascal (1623–1662)

Contents

Acknowledgments ix

Introduction 1

Mathematics Entries A to Z 3

References 359

Solutions to Puzzles 369

Category Index 373

Acknowledgments

Many people have helped me enormously in assembling this collection of mathematical oddities, delights, whimsies, and profundities. Thanks especially go to Jan Wassenaar (www.2dcurves.com) for drawing many of the plane curves that are featured in the book; Robert Webb (www.software3d.com) for numerous photos of his wonderful, homemade polyhedra; Jos Leys (www.josleys.com) for his mesmerizing fractal artwork; Xah Lee (www.xahlee.org) for a variety of ingenious digital imagery; Sue and Brian Young at Mr. Puzzle Australia (www.mrpuzzle.com.au) and Kate and Dick Jones at Kadon Enterprises (www.gamepuzzles.com) for their advice and photos of puzzles from their product lines and personal collections; Gideon Weisz (www.gideonweisz.com) and Istvan Orosz for stunning recursive and anamorphic art images; my good friend Andrew "Dogs" Barker for stimulating discussions and the solution to one of the problems; William Waite for pictures from his antique math collection; and Peter Cromwell, Lord & Lady Dunsany, Peter Knoppers, John Lienhard, John Mainstone, David Nicholls, Paul and Colin Roberts, Anders Sandberg, John Sullivan, and others for their valuable contributions.

I'm greatly indebted to Stephen Power, senior editor, and to Lisa Burstiner, senior production editor, at John Wiley & Sons, for their encouragement and unfailing attention to detail, and even proffering of alternative, clever solutions to some of the problems in the book. Any errors that remain are entirely my own responsibility. Thanks also to my marvelous agent, Patricia Van der Leun. And last but most of all, thanks to my family for letting me pursue a career that is really a fantasy.

Introduction

You are lost in a maze: How do you find your way out? You want to build a time machine, but is time travel logically possible? How can one infinity be bigger than another? Why can't you drink from a Klein bottle? What is the biggest number in the world to have a proper name, and how can you write it? Who claimed he could see in the fourth dimension? And what does "iteration" mean? And what does "iteration" mean?

Mathematics was never my strong point in school, but because I wanted to become an astronomer, I was told to stick with it. Fortunately, in my last two years before heading off to university, I had a wonderful old-fashioned, eccentric teacher (he actually wore a black gown when teaching), called Mr. Kay (known to one and all as "Danny"), who would suddenly divert from the chalk and blackboard to ask, "But how did the universe come to be *asymmetric*–that's what I want to know," or "These imaginary numbers are *very* interesting; in part, because they are so remarkably real." During lunchbreak, Danny and the senior chemistry teacher, Mr. Erp (whose nickname I need hardly spell out), would always meet in the chemistry prep room for a game of chess. They looked and acted very much like characters from a Wellsian science fiction tale, and I sometimes imagined them musing on formulas for invisibility or doorways to higher dimensions. At any rate, though I was never a shining student, I realize what a profound effect those two deeply imaginative, thoughtful men had on my future career. I did become an astronomer. I did persevere with math to a certain level of competence. But, much more than that, my curiosity was fired by the wonderful and weird possibilities of these subjects: curved space, Möbius bands, parallel universes, patterns in the heart of chaos, alternative realities. These strange possibilities, and a thousand others, make up the stuffing of this book. If you want a comprehensive, academic dictionary of mathematics, look elsewhere. If you want rigor and proof, try the next shelf. Herein you will find only the unusual and the outrageous, the fanciful and the fantastic: a compendium of the mathematics they *didn't* teach you in school.

Entries range from short definitions to lengthy articles on topics of major importance or unusual interest. These are arranged alphabetically according to the first word of the entry name and are extensively cross-referenced. Terms that appear in **bold** type have their own entries. A number of puzzles are included for the reader to try; the answers to these can be found at the back of the book. Also at the back are a comprehensive list of references and a category index. Readers are invited to visit the author's Web site at www.daviddarling.info for the latest news in mathematics and related subjects.

A

abacus

A counting frame that started out, several thousand years ago, as rows of pebbles in the desert sands of the Middle East. The word appears to come from the Hebrew *âbâq* (dust) or the Phoenician *abak* (sand) via the Greek *abax,* which refers to a small tray covered with sand to hold the pebbles steady. The familiar frame-supporting rods or wires, threaded with smoothly running beads, gradually emerged in a variety of places and mathematical forms.

In Europe, there was a strange state of affairs for more than 1,500 years. The Greeks and the Romans, and then the medieval Europeans, calculated on devices with a **place-value system** in which **zero** was represented by an empty line or wire. Yet the written notations didn't have a symbol for zero until it was introduced in Europe in 1202 by **Fibonacci**, via the Arabs and the Hindus.

abacus A special form of the Chinese abacus (c. 1958) consisting of two abaci stacked one on top of the other. *Luis Fernandes*

The Chinese *suan pan* differs from the European abacus in that the board is split into two decks, with two beads on each rod in the upper deck and five beads, representing the digits 0 through 4, on each rod in the bottom. When all five beads on a rod in the lower deck are moved up, they're reset to the original position, and one bead in the top deck is moved down as a carry. When both beads in the upper deck are moved down, they're reset and a bead on the adjacent rod on the left is moved up as a carry. The result of the computation is read off from the beads clustered near the separator beam between the upper and lower decks. In a sense, the abacus works as a 5-2-5-2-5-2 . . . –based number system in which carries and shifts are similar to those in the decimal system. Since each rod represents a digit in a decimal number, the capacity of the abacus is limited only by the number of rods on the abacus. When a user runs out of rods, she simply adds another abacus to the left of the row.

The Japanese *soroban* does away with the dual representations of fives and tens by having only four counters in the lower portion, known as "earth," and only one counter in the upper portion, known as "heaven." The world's largest abacus is in the Science Museum in London and measures 4.7 meters by 2.2 meters.

Abbott, Edwin Abbott (1838–1926)

An English clergyman and author who wrote several theological works and a biography (1885) of Francis Bacon, but is best known for his standard *Shakespearian Grammar* (1870) and the pseudonymously written ***Flatland: A Romance of Many Dimensions*** (by A Square, 1884).[1]

ABC conjecture

A remarkable **conjecture**, first put forward in 1980 by Joseph Oesterle of the University of Paris and David Masser of the Mathematics Institute of the University of Basel in Switzerland, that is now considered one of the most important unsolved problems in **number theory**. If it were proved correct, the proofs of many other famous conjectures and theorems would follow immediately–in some cases in just a few lines. The vastly complex current proof of **Fermat's last theorem**, for example, would reduce to less than a page of mathematical reasoning. The ABC conjecture is disarmingly simple compared to most of the deep questions in number theory and, moreover, turns out to be equivalent to all the main problems that involve **Diophantine equation**s (equations with integer coefficients and integer solutions).

Only a couple of concepts need to be understood to grasp the ABC conjecture. A *square-free number* is an integer that isn't divisible by the square of any number. For example, 15 and 17 are square-free, but 16 (divisible by 4^2) and 18 (divisible by 3^2) are not. The *square-free part* of an integer n, denoted $\mathrm{sqp}(n)$, is the largest square-free number that can be formed by multiplying the prime factors of n. For $n = 15$, the prime factors are 5 and 3, and $3 \times 5 = 15$, a square-free number, so that $\mathrm{sqp}(15) = 15$. On the other hand, for $n = 16$, the prime factors are all 2, which means that $\mathrm{sqp}(16) = 2$. In general, if n is square-free, the square-free part of n is just n; otherwise, $\mathrm{sqp}(n)$ represents what is left over after all the factors that create a square have been eliminated. In other words, $\mathrm{sqp}(n)$ is the product of the distinct **prime number**s that divide n. For example, $\mathrm{sqp}(9) = \mathrm{sqp}(3 \times 3) = 3$ and $\mathrm{sqp}(1{,}400) = \mathrm{sqp}(2 \times 2 \times 2 \times 5 \times 5 \times 7) = 2 \times 5 \times 7 = 70$.

The ABC conjecture deals with pairs of numbers that have no common factors. Suppose A and B are two such numbers that add to give C. For example, if $A = 3$ and $B = 7$, then $C = 3 + 7 = 10$. Now, consider the square-free part of the product $A \times B \times C$: $\mathrm{sqp}(ABC) = \mathrm{sqp}(3 \times 7 \times 10) = 210$. For most values of A and B, $\mathrm{sqp}(ABC) > C$, as in the prior example. In other words, $\mathrm{sqp}(ABC)/C > 1$. Occasionally, however, this isn't true. For instance, if $A = 1$ and $B = 8$, then $C = 1 + 8 = 9$, $\mathrm{sqp}(ABC) = \mathrm{sqp}(1 \times 8 \times 9) = \mathrm{sqp}(1 \times 2 \times 2 \times 2 \times 3 \times 3) = 1 \times 2 \times 3 = 6$, and $\mathrm{sqp}(ABC)/C = {}^6\!/_9 = {}^2\!/_3$. Similarly, if $A = 3$ and $B = 125$, the ratio is ${}^{15}\!/_{64}$.

David Masser proved that the ratio $\mathrm{sqp}(ABC)/C$ can get arbitrarily small. In other words, given any number greater than zero, no matter how small, it's possible to find integers A and B for which $\mathrm{sqp}(ABC)/C$ is smaller than this number. In contrast, the ABC conjecture says that $[\mathrm{sqp}(ABC)]^n/C$ reaches a minimum value if n is any number greater than 1–even a number such as 1.0000000001, which is only barely larger than 1. The tiny change in the expression results in a huge difference in its mathematical behavior. The ABC conjecture in effect translates an infinite number of Diophantine equations (including the equation of Fermat's last theorem) into a single mathematical statement.[144]

Abel, Niels Henrik (1802–1829)

> *The divergent series are the invention of the devil, and it is a shame to base on them any demonstration whatsoever. By using them, one may draw any conclusion he pleases and that is why these series have produced so many fallacies and so many paradoxes.*

A Norwegian mathematician who, independently of his contemporary Évariste **Galois**, pioneered **group** theory and proved that there are no algebraic solutions of the general **quintic** equation. Both Abel and Galois died tragically young–Abel of tuberculosis, Galois in a sword fight.

While a student in Christiania (now Oslo), Abel thought he had discovered how to solve the general quintic algebraically, but soon corrected himself in a famous pamphlet published in 1824. In this early paper, Abel showed the impossibility of solving the general quintic by means of radicals, thus laying to rest a problem that had perplexed mathematicians since the mid-sixteenth century. Abel, chronically poor throughout his life, was granted a small stipend by the Norwegian government that allowed him to go on a mathematical tour of Germany and France. In Berlin he met Leopold Crelle (1780–1856) and in 1826 helped him found the first journal in the world devoted to mathematical research. Its first three volumes contained 22 of Abel's papers, ensuring lasting fame for both Abel and Crelle. Abel revolutionized the important area of elliptic integrals with his theory of **elliptic function**s, contributed to the theory of **infinite series**, and founded the theory of commutative groups, known today as **Abelian group**s. Yet his work was never properly appreciated during his life, and, impoverished and ill, he returned to Norway unable to obtain a teaching position. Two days after his death, a delayed letter was delivered in which Abel was offered a post at the University of Berlin.

Abelian group

A **group** that is **commutative**, that is, in which the result of multiplying one member of the group by another is independent of the order of multiplication. Abelian groups, named after Niels **Abel**, are of central importance in modern mathematics, most notably in **algebraic topology**. Examples of Abelian groups include the **real number**s (with addition), the nonzero real numbers (with multiplication), and all cyclic groups, such as the **integer**s (with addition).

abracadabra

A word famously used by magicians but which started out as a cabalistic or mystical charm for curing various ailments, including toothache and fever. It was first mentioned in a poem called "Praecepta de Medicina" by the Gnostic physician Quintus Severus Sammonicus in the second century A.D. Sammonicus instructed that the letters be written on parchment in the form of a triangle:

A B R A C A D A B R A
A B R A C A D A B R
A B R A C A D A B
A B R A C A D A
A B R A C A D
A B R A C A
A B R A C
A B R A
A B R
A B
A

This was to be folded into the shape of a cross, worn for nine days suspended from the neck, and, before sunrise, cast behind the patient into a stream running eastward. It was also a popular remedy in the Middle Ages. During the Great Plague, around 1665, large numbers of these amulets were worn as safeguards against infection. The origin of the word itself is uncertain. One theory is that it is based on Abrasax, the name of an Egyptian deity.

PUZZLE

A well-known puzzle, proposed by George Polya (1887–1985), asks how many different ways there are to spell abracadabra in this diamond-shaped arrangement of letters:

A
B B
R R R
A A A A
C C C C C
A A A A A A
D D D D D
A A A A
B B B
R R
A

Solutions begin on page 369.

abscissa

The x-coordinate, or horizontal distance from the y-axis, in a system of **Cartesian coordinates**. Compare with **ordinate**.

absolute

Not limited by exceptions or conditions. The term is used in many different ways in mathematics, physics, philosophy, and everyday speech. Absolute space and absolute time, which, in Newton's universe, form a unique, immutable frame of reference, blend and become deformable in the **space-time** of Einstein. See also **absolute zero**. In some philosophies, the absolute stands behind the reality we see–independent, transcendent, unconditional, and all-encompassing. The American philosopher Josiah Royce (1855–1916) took the absolute to be a spiritual entity whose self-consciousness is imperfectly reflected in the totality of human thought. Mathematics, too, reaches beyond imagination with its absolute **infinity**. See also **absolute value**.

absolute value

The value of a number without regard to its sign. The absolute value, or *modulus,* of a **real number**, r, is the distance of the number from zero measured along the real **number line**, and is denoted $|r|$. Being a distance, it can't be negative; so, for example, $|3| = |-3| = 3$. The same idea applies to the absolute value of a **complex number** $a + ib$, except that, in this case, the complex number is represented by a point on an **Argand diagram**. The absolute value, $|a + ib|$, is the length of the line from the origin to the given point, and is equal to $\sqrt{(a^2 + b^2)}$.

absolute zero

The lowest possible temperature of a substance, equal to 0 Kelvin (K), –273.15°C, or –459.67°F. In classical physics, it is the temperature at which all molecular motion ceases. However, in the "real" world of **quantum mechanics** it isn't possible to stop all motion of the particles making up a substance as this would violate the Heisenberg uncertainty principle. So, at 0 K, particles would still vibrate with a certain small but nonzero energy known as the *zero-point energy.* Temperatures within a few billionths of a degree of absolute zero have been achieved in the laboratory. At such low temperatures, substances have been seen to enter a peculiar state, known as the *Bose-Einstein condensate,* in which their quantum wave functions merge and particles lose their individual identities. Although it is possible to approach ever closer to absolute zero, the third law of thermodynamics asserts that it's impossible to ever attain it. In a deep sense, absolute zero lies at the asymptotic limit of low energy just as the speed of light lies, for particles with mass, at the asymptotic limit of high energy. In both cases, energy of motion (kinetic energy) is the key quantity involved. At the high energy end, as the average speed of the particles of a substance approaches the speed of light, the temperature rises without limit, heading for an unreachable ∞ K.

abstract algebra

> *To a mathematician, real life is a special case.*
>
> –Anonymous

Algebra that is not confined to familiar number systems, such as the **real numbers**, but seeks to solve equations that may involve many other kinds of systems. One of its aims, in fact, is to ask: What other number systems are there? The term *abstract* refers to the perspective taken on the subject, which is very different from that of high school algebra. Rather than looking for the solutions to a particular problem, abstract algebra is interested in such questions as: When does a solution exist? If a solution does exist, is it unique? What general properties does a solution possess? Among the structures it deals with are **group**s, **ring**s, and **field**s. Historically, examples of such structures often arose first in some other field of mathematics, were specified rigorously (axiomatically), and were then studied in their own right in abstract algebra.

Abu'l Wafa (A.D. 940–998)

A Persian mathematician and astronomer who was the first to describe geometrical constructions (see **constructible**) possible only with a straightedge and a fixed compass, later dubbed a "rusty compass," that never alters its radius. He pioneered the use of the tangent function (see **trigonometric function**), apparently discovered the secant and cosecant functions, and compiled tables of sines and tangents at 15′ intervals—work done as part of an investigation into the orbit of the Moon.

abundant number

A number that is smaller than the sum of its **aliquot part**s (proper divisors). Twelve is the smallest abundant number; the sum of its aliquot parts is 1 + 2 + 3 + 4 + 6 = 16, followed by 18, 20, 24, and 30. A *weird number* is an abundant number that is not *semiperfect;* in other words, n is weird if the sum of its divisors is greater than n, but n is not equal to the sum of any subset of its divisors. The first few weird numbers are 70, 836, 4,030, 5,830, and 7,192. It isn't known if there are any odd weird numbers. A *deficient number* is one that is greater than the sum of its aliquot parts. The first few deficient numbers are 1, 2, 3, 4, 5, 8, and 9. Any divisor of a deficient (or perfect) number is deficient. A number that is not abundant or deficient is known as a **perfect number**.

Achilles and the Tortoise paradox

See **Zeno's paradoxes**.

Ackermann function

One of the most important **function**s in computer science. Its most outstanding property is that it grows astonishingly fast. In fact, it gives rise to **large numbers** so quickly that these numbers, called *Ackermann numbers,* are written in a special way known as **Knuth's up-arrow notation**. The Ackermann function was discovered and studied by Wilhelm Ackermann (1896–1962) in 1928. Ackermann worked as a high school teacher from 1927 to 1961 but was also a student of the great mathematician David **Hilbert** in Göttingen and, from 1953, served as an honorary professor in the university there. Together with Hilbert he published the first modern textbook on mathematical logic. The function he discovered, and that now bears his name, is the simplest example of a well-defined and total function that is also computable but not primitive recursive (PR). "Well-defined and total" means that the function is internally consistent and doesn't break any of the rules laid down to define it. "Computable" means that it can, in principle, be evaluated for all possible input values of its variables. "Primitive recursive" means that it can be computed using only *for loops*—repeated application of a single operation a predetermined number of times. The recursion, or feedback loop, in the Ackermann function overruns the capacity of any *for loop* because the number of loop repetitions isn't known in advance. Instead, this number is itself part of the computation, and grows as the calculation proceeds. The Ackermann function can only be calculated using a *while loop,* which keeps repeating an action until an associated test returns false. Such loops are essential when the programmer doesn't know at the outset how many times the loop will be traversed. (It's now known that everything computable can be programmed using while loops.)

The Ackermann function can be defined as follows:

$$
\begin{aligned}
A(0, n) &= n + 1 \text{ for } n = 0\\
A(m, 0) &= A(m - 1, 1) \text{ for } m = 1\\
A(m, n) &= A(m - 1, A(m, n - 1)) \text{ for } m, n = 1.
\end{aligned}
$$

Two positive integers, m and n, are the input and $A(m, n)$ is the output in the form of another positive integer. The function can be programmed easily in just a few lines of code. The problem isn't the complexity of the function but the awesome rate at which it grows. For example, the innocuous-looking $A(4,2)$ already has 19,729 digits! The use of a powerful large-number shorthand system, such as the up-arrow notation, is indispensable as the following examples show:

$$
\begin{aligned}
A(1, n) &= 2 + (n + 3) - 3\\
A(2, n) &= 2 \times (n + 3) - 3\\
A(3, n) &= 2\uparrow(n + 3) - 3\\
A(4, n) &= 2\uparrow(2\uparrow(2\uparrow(\ldots\uparrow 2))) - 3 \;(n + 3 \text{ twos})\\
&= 2\uparrow\uparrow(n + 3) - 3\\
A(5, n) &= 2\uparrow\uparrow\uparrow(n + 3) - 3, \text{ etc.}
\end{aligned}
$$

Intuitively, the Ackermann function defines generalizations of multiplication by 2 (iterated additions) and exponentiation with base 2 (iterated multiplications) to iterated exponentiation, iteration of this operation, and so on.[84]

acre
An old unit of **area**, equal to 160 square rods, 4,840 square yards, 43,560 square feet, or 4,046.856 square meters.

acute
From the Latin *acus* for "needle" (which also forms the root for *acid, acupuncture,* and *acumen*). An acute **angle** is less than 90°. An acute **triangle** is one in which all three angles are acute. Compare with **obtuse**.

adjacent
Next to. *Adjacent angles* are next to each other, and thus share one side. *Adjacent sides* of a **polygon** share a vertex.

affine geometry
The study of properties of geometric objects that remain unchanged after parallel projection from one plane to another. During such a projection, first studied by Leonhard **Euler**, each point (x, y) is mapped to a new point $(ax + cy + e, bx + dy + f)$. Circles, angles, and distances are altered by affine transformations and so are of no interest in affine geometry. Affine transformations do, however, preserve collinearity of points: if three points belong to the same straight line, their images (the points that correspond to them) under affine transformations also belong to the same line and, in addition, the middle point remains between the other two points. Similarly, under affine transformations, parallel lines remain parallel; concurrent lines remain concurrent (images of intersecting lines intersect); the ratio of lengths of line segments of a given line remains constant; the ratio of areas of two triangles remains constant; and ellipses, parabolas, and hyperbolas continue to be ellipses, parabolas, and hyperbolas.

age puzzles and tricks
Problems that ask for a person's age or, alternatively, when a person was a certain age, given several roundabout facts. They go back at least 1,500 years to the time of Metrodorus and **Diophantus's riddle**. A number of distinct types of age puzzles sprang up between the sixteenth and early twentieth centuries, in most cases best solved by a little algebra. One form asks: if X is now a years old and Y is now b years old, when will X be c times as old as Y? The single unknown, call it x, can be found from the equation $a + x = c(b + x)$. Another type of problem takes the form: if X is now a times as old as Y and after b years X will be c times as old as Y, how old are X and Y now? In this case the trick is to set up and solve two simultaneous equations: $X = aY$ and $X + b = c(Y + b)$.

PUZZLES

Around 1900, two more variants on the age puzzle became popular. Here is an example of each for the reader to try.

1. Bob is 24. He is twice as old as Alice was when Bob was as old as Alice is now. How old is Alice?
2. The combined ages of Mary and Ann are 44 years. Mary is twice as old as Ann was when Mary was half as old as Ann will be when Ann is three times as old as Mary was when Mary was three times as old as Ann. How old is Ann?"

Solutions begin on page 369.

Various mathematical sleights of hand can seem to conjure up a person's age as if by magic. For example, ask a person to multiply the first number of his or her age by 5, add 3, double this figure, add the second number of his or her age to the figure, and tell you the answer. Deduct 6 from this and you will have their age.

Alternatively, ask the person to pick a number, multiply this by 2, add 5, and multiply by 50. If the person has already had a birthday this year and it's the year 2004, she should add 1,754, otherwise she should add 1,753. Each year after 2004 these numbers need to be increased by 1. Finally, the person should subtract the year they were born. The first digits of the answer are the original number, while the last two digits are the person's age.

Here is one more trick. Take your age, multiply it by 7, then multiply again by 1,443. The result is your age repeated three times. (What you have actually done is multiplied by 10,101; if you multiply by 1,010,101, the repetition is fourfold, and so on.)

Agnesi, Maria Gaetana (1718–1799)
An Italian mathematician and scholar whose name is associated with the curve known as the *Witch of Agnesi.* Born in Milan, Maria was one of 24 children of a professor of mathematics at the University of Bologna. A child prodigy, she could speak seven languages, including Latin, Greek, and Hebrew, by the age of 11 and was solving difficult problems in geometry and ballistics by her early teens. Her father encouraged her studies and her appearance at public debates. However, Maria developed a chronic illness, marked by convulsions and headaches, and, from the age of about 20, withdrew socially and devoted herself to mathematics. Her *Instituzioni analitiche ad uso della gioventu italiana,* published in 1748, became a standard teaching manual, and in 1750, she was appointed to the chair of mathematics and natural philosophy at Bologna. Yet she never fulfilled her early promise in terms of making new

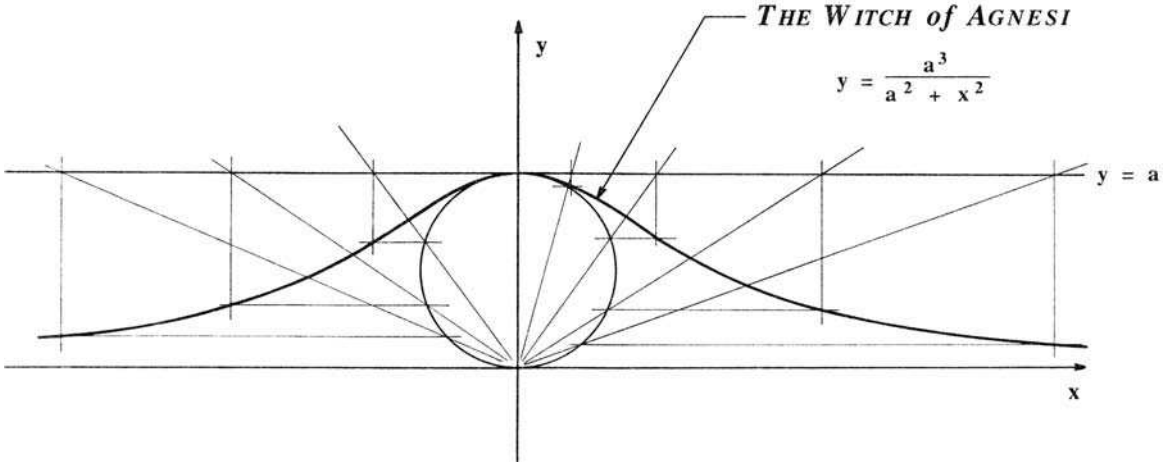

Agnesi, Maria Gaetana The Witch of Agnesi curve. *John H. Lienhard*

breakthroughs. After the death of her father in 1752, she moved into theology and, after serving for some years as the directress of the Hospice Trivulzio for Blue Nuns at Milan, joined the sisterhood herself and ended her days in this austere order.

The famous curve that bears her name had been studied earlier, in 1703, by Pierre de **Fermat** and the Italian mathematician Guido Grandi (1671–1742). Maria wrote about it in her teaching manual and referred to it as the *aversiera,* which simply means "to turn." But in translating this, the British mathematician John Colson (1680–1760), the fifth Lucasian professor of mathematics at Cambridge University, confused *aversiera* with *avversiere* which means "witch," or "wife of the devil." And so the name of the curve came down to us as the Witch of Agnesi. To draw it, start with a circle of diameter a, centered at the point $(0, a/2)$ on the y-axis. Choose a point A on the line $y = a$ and connect it to the origin with a line segment. Call the point where the segment crosses the circle B. Let P be the point where the vertical line through A crosses the horizontal line through B. The Witch is the curve traced by P as A moves along the line $y = a$. By a happy coincidence, it does look a bit like a witch's hat! In Cartesian coordinates, its equation is

$$y = a^3/(x^2 + a^2).$$

Ahmes papyrus

See **Rhind papyrus**.

Ahrens, Wilhelm Ernst Martin Georg (1872–1927)

A great German exponent of recreational mathematics whose *Mathematische Unterhaltungen und Spiele*[6] is one of the most scholarly of all books on the subject.

Alcuin (735–804)

A leading intellectual of his time and the probable compiler of *Propositiones ad Acuendos Juvenes* (Problems to sharpen the young), one of the earliest collections of recreational math problems. According to David **Singmaster** and John Hadley: "The text contains 56 problems, including 9 to 11 major types of problem which appear for the first time, 2 major types which appear in the West for the first time and 3 novel variations of known problems. . . . It has recently been realized that the **river-crossing problems** and the crossing-a-desert problem, which appear here for the first time, are probably the earliest known combinatorial problems."

Alcuin was born into a prominent family near the east coast of England. He was sent to York, where he became a pupil and, eventually, in 778, the headmaster, of Archbishop Ecgberht's School. (Ecgberht was the last person to have known the Venerable Bede.) Alcuin built up a superb library and made the school one of the chief centers of learning in Europe. Its reputation became such that, in 781, Alcuin was invited to become master of Charlemagne's Palace School at Aachen and, effectively, minister of education for Charlemagne's empire. He accepted and traveled to Aachen to a meeting of the leading scholars. Subsequently, he was made head of Charlemagne's Palace School and there developed the Carolingian minuscule, a clear, legible script that became the basis of how letters of the present Roman alphabet are written.

Before leaving Aachen, Alcuin was responsible for the most prized of the Carolingian codices, now called the

Golden Gospels: a series of illuminated masterpieces written largely in gold, on white or purple vellum. The development of Carolingian minuscule had, indirectly, a major impact on the history of mathematics. Because it was a far more easily readable script than the older unspaced capital, it led to many mathematical works being newly copied into this new style in the ninth century. Most of the works of the ancient Greek mathematicians that have survived did so because of this transcription. Alcuin lived in Aachen from 782 to 790 and again from 793 to 796. In 796, he retired from Charlemagne's Palace School and became abbot of the Abbey of St. Martin at Tours, where he and his monks continued to work with the Carolingian minuscule script.

aleph

The first letter of the Hebrew alphabet, ℵ. It was first used in mathematics by Georg **Cantor** to denote the various orders, or sizes, of **infinity**: $\aleph_0$ (aleph-null), $\aleph_1$ (aleph-one), etc. An earlier (and still used) symbol for infinity, ∞, was introduced in 1655 by John **Wallis** in his *Arithmetica infinitorum* but didn't appear in print until the *Ars conjectandi* by Jakob Bernoulli, published posthumously in 1713 by his nephew Nikolaus Bernoulli (see **Bernoulli Family**).

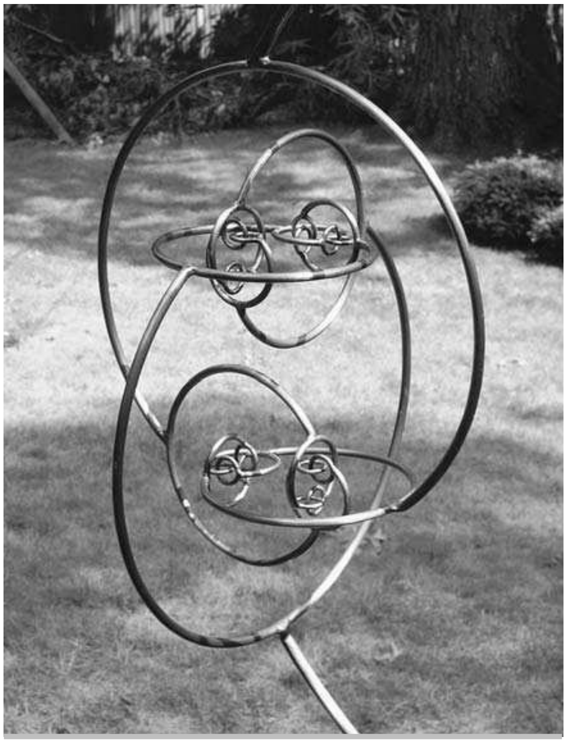

Alexander's horned sphere A sculpture of a five-level Alexander's horned sphere. *Gideon Weisz, www.gideonweisz.com*

Alexander's horned sphere

In **topology,** an example of what is called a "wild" structure; it is named after the Princeton mathematician James Waddell Alexander (1888–1971) who first described it in the early 1920s. The horned sphere is topologically equivalent to the **simply connected** surface of an ordinary hollow **sphere** but bounds a region that is not simply connected. The horns-within-horns consist of a recursive set–a **fractal**–of interlocking pairs of orthogonal rings (rings set at right angles) of decreasing radius. A rubber band around the base of any horn couldn't be removed from the structure even after infinitely many steps. The horned sphere can be embedded in the plane by reducing the interlock angle between ring pairs from 90° to 0°, then weaving the rings together in an over-under pattern. The sculptor Gideon Weisz has modeled a number of approximations to the structure, one of which is shown in the photograph.

algebra

A major branch of mathematics that, at an elementary level, involves applying the rules of arithmetic to numbers, and to letters that stand for unknown numbers, with the main aim of solving equations. Beyond the algebra learned in high school is the much vaster and more profound subject of **abstract algebra**. The word itself comes from the Arabic *al-jebr,* meaning "the reunion of broken parts." It first appeared in the title of a book, *Al-jebr w'al-mugabalah* (The science of reduction and comparison), by the ninth-century Persian scholar **al-Khowarizmi**–probably the greatest mathematician of his age, and as famous among Arabs as Euclid and Aristotle are to the Western world.

algebraic curve

A curve whose equation involves only *algebraic functions.* These are functions that, in their most general form, can be written as a sum of **polynomial**s in x multiplied by powers of y, equal to zero. Among the simplest examples are straight lines and **conic section**s.

algebraic fallacies

Misuse of **algebra** can have some surprising and absurd results. Here, for example, is a famous "proof" that 1 = 2:

$$\begin{gathered}
\text{Let } a = b.\\
\text{Then } a^2 = ab\\
a^2 + a^2 = a^2 + ab\\
2a^2 = a^2 + ab\\
2a^2 - 2ab = a^2 + ab - 2ab\\
2a^2 - 2ab = a^2 - ab\\
2(a^2 - ab) = 1(a^2 - ab).\\
\text{Dividing both sides by } a^2 - ab\\
2 = 1.
\end{gathered}$$

Where's the mistake? The problem lies with the seemingly innocuous final division. Since $a = b$, dividing by $a^2 - ab$ is the same as dividing by **zero**–the great taboo of mathematics.

Another false argument runs as follows:

$$(n + 1)^2 = n^2 + 2n + 1$$
$$(n + 1)^2 - (2n + 1) = n^2.$$

Subtracting $n(2n + 1)$ from both sides and factorizing gives

$$(n + 1)^2 - (n + 1)(2n + 1) = n^2 - n(2n + 1).$$

Adding $\frac{1}{4}(2n + 1)^2$ to both sides yields

$$(n + 1)^2 - (n + 1)(2n + 1) + \tfrac{1}{4}(2n + 1)^2$$
$$= n^2 - n(2n + 1) + \tfrac{1}{4}(2n + 1)^2.$$

This may be written:

$$[(n + 1) - \tfrac{1}{2}(2n + 1)]^2 = [(n - \tfrac{1}{2}(2n + 1)]^2.$$

Taking square roots of both sides,

$$n + 1 - \tfrac{1}{2}(2n + 1) = n - \tfrac{1}{2}(2n + 1).$$

Therefore,

$$n = n + 1.$$

The problem here is that there are *two* square roots for any positive number, one positive and one negative: the square roots of 4 are 2 and –2, which can be written as ±2. So the penultimate step should properly read:

$$\pm(n + 1 - \tfrac{1}{2}(2n + 1)) = \pm(n - \tfrac{1}{2}(2n + 1))$$

algebraic geometry

Originally, the geometry of **complex number** solutions to **polynomial** equations. Modern algebraic geometry is also concerned with algebraic varieties, which are a generalization of the solution sets found in the traditional subject, as well as solutions in fields other than complex numbers, for example *finite fields.*

algebraic number

A **real number** that is a **root** of a **polynomial** equation with integer coefficients. For example, any **rational number** a/b, where a and b are nonzero integers, is an algebraic number of degree one, because it is a root of the linear equation $bx - a = 0$. The **square root of two** is an algebraic number of degree two because it is a root of the quadratic equation $x^2 - 2 = 0$. If a real number is not algebraic, then it is a **transcendental number**. Almost all real numbers are transcendental because, whereas the set of algebraic numbers is countably infinite (see **countable set**), the set of transcendental numbers is uncountably infinite.

algebraic number theory

The branch of **number theory** that is studied without using methods such as **infinite series** and **convergence** taken from **analysis.** It contrasts with **analytical number theory**.

algebraic topology

A branch of **topology** that deals with invariants of a topological space that are algebraic structures, often **groups**.

algorithm

A systematic method for solving a problem. The word comes from the name of the Persian mathematician, **al-Khowarizmi**, and may have been first used by Gottfried **Liebniz** in the late 1600s. It remained little known in Western mathematics, however, until the Russian mathematician Andrei Markov (1903–1987) reintroduced it. The term became especially popular in the areas of math focused on computing and computation.

algorithmic complexity

A measure of **complexity** developed by Gregory **Chaitin** and others, based on Claude **Shannon**'s **information theory** and earlier work by the Russian mathematicians Andrei **Kolmogorov** and Ray Solomonoff. Algorithmic complexity quantifies how complex a system is in terms of the shortest computer program, or set of **algorithms**, needed to completely describe the system. In other words, it is the smallest model of a given system that is necessary and sufficient to capture the essential patterns of that system. Algorithmic complexity has to do with the mixture of repetition and innovation in a complex system. At one extreme, a highly regular system can be described by a very short program or algorithm. For example, the bit string 01010101010101010101 . . . follows from just three commands: print a zero, print a one, and repeat the last two commands indefinitely. The complexity of such a system is very low. At the other extreme, a totally **random** system has a very high algorithmic complexity since the random patterns can't be condensed into a smaller set of algorithms: the program is effectively as large as the system itself. See also **compressible**.

Alhambra

The former palace and citadel of the Moorish kings of Granada, and perhaps the greatest monument to Islamic mathematical art on Earth. Because the Qur'an consid-

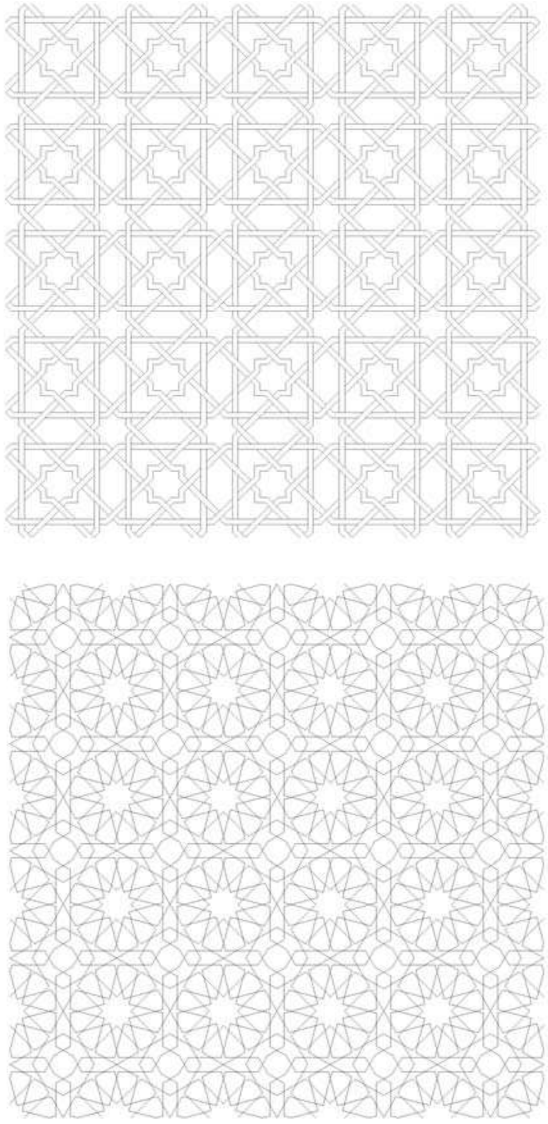

Alhambra Computer-generated tilings based on Islamic tile designs such as those found in the Alhambra. *Xah Lee, www.xahlee.org*

ers the depiction of living beings in religious settings blasphemous, Islamic artists created intricate patterns to symbolize the wonders of creation: the repetitive nature of these complex geometric designs suggests the limitless power of God. The sprawling citadel, looming high above the Andalusian plain, boasts a remarkable array of mosaics with tiles arranged in intricate patterns. The Alhambra **tiling**s are *periodic;* in other words, they consist of some basic unit that is repeated in all directions to fill up the available space. All 17 different groups of isometries–the possible ways of repeatedly tiling the plane–are used at the palace. The designs left a deep impression on Maurits **Escher**, who came here in 1936. Subsequently, Escher's art took on a much more mathematical nature, and over the next six years he produced 43 colored drawings of periodic tilings with a wide variety of symmetry types.

aliquot part

Also known as a *proper divisor,* any divisor of a number that isn't equal to the number itself. For instance, the aliquot parts of 12 are 1, 2, 3, 4, and 6. The word comes from the Latin *ali* ("other") and *quot* ("how many"). An *aliquot sequence* is formed by taking the sum of the aliquot parts of a number, adding them to form a new number, then repeating this process on the next number and so on. For example, starting with 20, we get $1 + 2 + 4 + 5 + 10 = 22$, then $1 + 2 + 11 = 14$, then $1 + 2 + 7 = 10$, then $1 + 2 + 5 = 8$, then $1 + 2 + 4 = 7$, then 1, after which the sequence doesn't change. For some numbers, the result loops back immediately to the original number; in such cases the two numbers are called **amicable numbers.** In other cases, where a sequence repeats a pattern after more than one step, the result is known as an *aliquot cycle* or a *sociable chain.* An example of this is the sequence 12496, 14288, 15472, 14536, 14264, . . . The aliquot parts of 14264 add to give 12496, so that the whole cycle begins again. Do all aliquot sequences end either in 1 or in an aliquot cycle (of which amicable numbers are a special case)? In 1888, the Belgian mathematician Eugène Catalan (1814–1894) conjectured that they do, but this remains an open question.

al-Khowarizmi (c. 780–850)

An Arabic mathematician, born in Baghdad, who is widely considered to be the founder of modern day **algebra**. He believed that any math problem, no matter how difficult, could be solved if broken down into a series of smaller steps. The word ***algorithm*** may have derived from his name.

Allais paradox

A **paradox** that stems from questions asked in 1951 by the French economist Maurice Allais (1911–).[8] Which of these would you choose: (A) an 89% chance of receiving an unknown amount and 11% chance of $1 million; or (B) an 89% chance of an unknown amount (the same amount as in A), a 10% chance of $2.5 million, and a 1% chance of nothing? Would your choice be the same if the unknown amount was $1 million? What if the unknown amount was zero?

Most people don't like risk and so prefer the better chance of winning $1 million in option A. This choice is firm when the unknown amount is $1 million, but seems to waver as the amount falls to nothing. In the latter case, the risk-averse person favors B because there isn't much difference between 10% and 11%, but there's a big difference between $1 million and $2.5 million. Thus the choice between A and B depends on the unknown amount, even though it is the same unknown amount independent of the choice. This flies in the face of the so-called *independence axiom,* that rational choice between two alternatives should depend only on how those two alternatives differ. Yet, if the amounts involved in the problem are reduced to tens of dollars instead of millions of dollars, people's behavior tends to fall back in line with the axioms of rational choice. In this case, people tend to choose option B regardless of the unknown amount. Perhaps when presented with such huge numbers, people begin to calculate qualitatively. For example, if the unknown amount is $1 million the options are essentially (A) a fortune guaranteed or (B) a fortune almost guaranteed with a small chance of a bigger fortune and a tiny chance of nothing. Choice A is then rational. However, if the unknown amount is nothing, the options are (A) a small chance of a fortune ($1 million) and a large chance of nothing, and (B) a small chance of a larger fortune ($2.5 million) and a large chance of nothing. In this case, the choice of B is rational. Thus, the Allais paradox stems from our limited ability to calculate rationally with such unusual quantities.

almost perfect number

A description sometimes applied to the powers of 2 because the **aliquot part**s (proper divisors) of 2^n sum to $2^n - 1$. So a power of 2 is a deficient number (one that is less than the sum of its proper divisors), but only just. It isn't known whether there is an odd number n whose divisors (excluding itself) sum to $n - 1$.

alphamagic square

A form of **magic square**, introduced by Lee **Sallows**,[278–280] in which the number of letters in the word for each number, in whatever language is being used, gives rise to another magic square. In English, for example, the alphamagic square:

5 (five)	22 (twenty-two)	18 (eighteen)
28 (twenty-eight)	15 (fifteen)	2 (two)
12 (twelve)	8 (eight)	25 (twenty-five)

generates the square:

4	9	8
11	7	3
6	5	10

A surprisingly large number of 3 × 3 alphamagic squares exist–in English and in other languages. French allows just one 3 × 3 alphamagic square involving numbers up to 200, but a further 255 squares if the size of the entries is increased to 300. For entries less than 100, none occurs in Danish or in Latin, but there are 6 in Dutch, 13 in Finnish, and an incredible 221 in German. Yet to be determined is whether a 3 × 3 square exists from which a magic square can be derived that, in turn, yields a third magic square–a magic triplet. Also unknown is the number of 4 × 4 and 5 × 5 language-dependent alphamagic squares. Here, for example, is a four-by-four English alphamagic square:

26	37	48	59
49	58	27	36
57	46	39	28
38	29	56	47

alphametic

A type of **cryptarithm** in which a set of words is written down in the form of a long addition sum or some other mathematical problem. The object is to replace the letters of the alphabet with decimal digits to make a valid arithmetic sum. The word *alphametic* was coined in 1955 by James **Hunter**. However, the first modern alphametic, published by Henry **Dudeney** in the July 1924 issue of *Strand Magazine,* was "Send more money," or, setting it out in the form of a long addition:

$$\begin{array}{r} \text{SEND} \\ \underline{\text{MORE}} \\ \text{MONEY} \end{array}$$

and has the (unique) solution:

$$\begin{array}{r} 9567 \\ \underline{1085} \\ 10652 \end{array}$$

PUZZLES

The reader is invited to try to solve the following elegant examples:

1. Earth, air, fire, water: nature. (Herman Nijon)
2. Saturn, Uranus, Neptune, Pluto: planets. (Peter J. Martin)
3. Martin Gardner retires. (H. Everett Moore)

Solutions begin on page 369.

Two rules are obeyed by every alphametic. First, the mapping of letters to numbers is one-to-one; that is, the same letter always stands for the same digit, and the same digit is always represented by the same letter. Second, the digit zero isn't allowed as the left-most digit in any of the numbers being added or in their sum. The best alphametics are reckoned to be those with only one correct answer.

Altekruse puzzle

A symmetrical 12-piece **burr puzzle** for which a patent was granted to William Altekruse in 1890. The Altekruse family is of Austrian-German origin and, curiously, the name means "old cross" in German, which has led some authors to incorrectly assume that it was a pseudonym. William Altekruse came to the United States as a young man in 1844 with his three brothers to escape being drafted into the German army. The Altekruse puzzle has an unusual mechanical action in the first step of disassembly by which two halves move in opposition to each other, unlike the more familiar burr types that have a key piece or pieces. Depending on how it is assembled, this action can take place along one, two, or all three axes independently but not simultaneously.

alternate

A mathematical term with several different meanings: (1) *Alternate angles* are angles on opposite sides and opposite ends of a line that cuts two parallel lines. (2) A well-known theorem called the *alternate segment theorem* involves the segment on the opposite side of a given chord of a circle. (3) An *alternate hypothesis* in statistics is the alternative offered to the **null hypothesis**. (4) To alternate is to cycle backward and forward between two different values, for example, 0, 1, 0, 1, 0, 1, ...

altitude

A perpendicular line segment from one **vertex** of a figure or solid to an edge or face opposite to that vertex. Also the length of such a line segment.

ambiguous figure

An **optical illusion** in which the subject or the perspective of a picture or shape may suddenly switch in the mind of the observer to another, equally valid possibility. Often the ambiguity stems from the fact that the figure and ground can be reversed. An example of this is the vase/profile illusion, made famous by the Danish psychologist Edgar John Rubin (1886–1951) in 1915, though earlier versions of the same illusion can be found in many eighteenth-century French prints depicting a variety of vases, usually in a naturalistic setting, and profiles of particular people. The same effect can be created in three dimensions with a suitably shaped solid vase. In some ambiguous figures, the features of a person or of an animal can suddenly be seen as different features of another individual. Classic examples include the old woman–young woman illusion and the duck–rabbit illusion. **Upside-down pictures** involve a special case of dual-purpose features in which the reversal is accomplished not mentally, by suddenly "seeing" the alternative, but physically, by turning the picture 180°. Ambiguity can also occur, particularly in some geometric drawings, when there is confusion as to which are the front and the back faces of a figure, as in the **Necker cube**, the **Thiery figure**, and **Schröder's reversible staircase**.

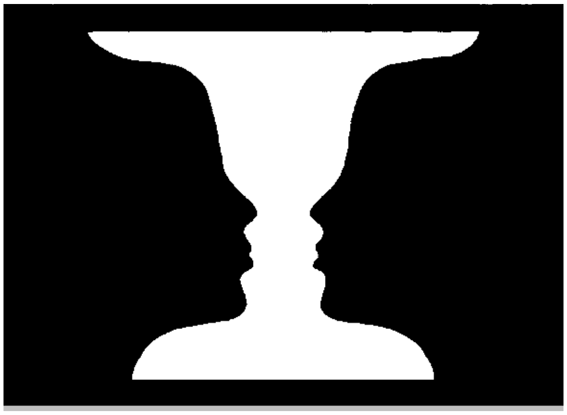

ambiguous figure The Rubin vase illusion: one moment a vase, the next two people face to face.

ambiguous connectivity

See **impossible figure**.

Ames room

The famous distorted room illusion, named after the American ophthalmologist Adelbert Ames Jr. (1880–1955), who first constructed such a room in 1946 based on a concept by the German physicist Hermann Helmholtz in the late nineteenth century. The Ames room looks cubic when seen with one eye through a specially positioned peephole; however, the room's true shape is trapezoidal. The floor, ceiling, some walls, and the far windows are trapezoidal surfaces; the floor appears level but is actually at an incline (one of the far corners being

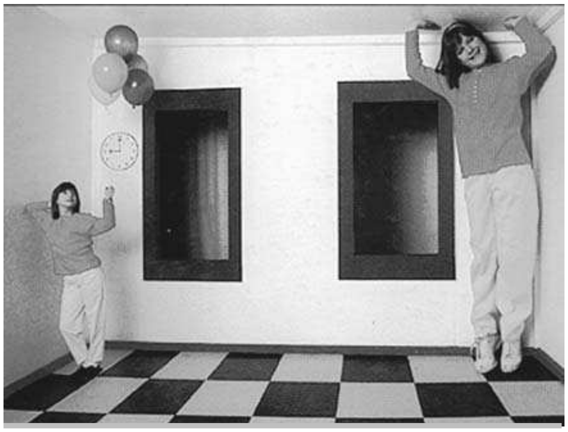

Ames room Misleading geometry makes these identical twins appear totally different in size. *Technische Universitat, Dresden*

much lower than the other); and the walls are slanted outward, though they seem perpendicular to the floor. This shape makes it look as if people or objects grow or shrink as they move from one corner of the room to another. See also **distortion illusion**.[142, 178]

amicable numbers

A pair of numbers, also known as *friendly numbers,* each of whose **aliquot part**s add to give the other number. (An aliquot part is any divisor that doesn't include the number itself.) The smallest amicable numbers are 220 (aliquot parts 1, 2, 4, 5, 10, 11, 20, 22, 44, 55, and 110, with a sum of 284) and 284 (aliquot parts 1, 2, 4, 71, and 142, with a sum of 220). This pair was known to the ancient Greeks, and the Arabs found several more. In 1636, Pierre de **Fermat** rediscovered the amicable pair 17,296 and 18,416; two years later René **Descartes** rediscovered a third pair, 9,363,584 and 9,437,056. In the eighteenth century, Leonhard **Euler** drew up a list of more than 60. Then, in 1866, B. Nicolò Paganini (not the violinist), a 16-year-old Italian, startled the mathematical world by announcing that the numbers 1,184 and 1,210 were amicable. This second-lowest pair of all had been completely overlooked! Today, the tally of known amicable numbers has grown to about 2.5 million. No amicable pair is known in which one of the two numbers is a square. An unusually high proportion of the numbers in amicable pairs ends in either 0 or 5. A *happy amicable pair* is an amicable pair in which both numbers are **happy number**s; an example is 10,572,550 and 10,854,650. See also **Harshad number**.

amplitude

Size or magnitude. The origin of the word is the same Indo-European *ple* root that gives us *plus* and *complement.* The more immediate Latin source is *amplus* for "wide." Today, **amplitude** is used to describe, among other things, the distance a **periodic** function varies from its central value, and the magnitude of a **complex number**.

anagram

The rearrangement of the letters of a word or phrase into another word or phrase, using all the letters only once. The best anagrams are meaningful and relate in some way to the original subject; for example, "stone age" and "stage one." There are also many remarkable examples of long anagrams. " 'That's one small step for a man; one giant leap for mankind.' Neil Armstrong" becomes "An 'Eagle' lands on Earth's Moon, making a first small permanent footprint."

PUZZLES

The reader is invited to untangle the following anagrams that give clues to famous people:

1. A famous German waltz god.
2. Aha! Ions made volts!
3. I'll make a wise phrase.

Solutions begin on page 369.

An *antonymous anagram,* or *antigram,* has a meaning opposite to that of the subject text; for example, "within earshot" and "I won't hear this." *Transposed couplets,* or *pairagrams,* are single word anagrams that, when placed together, create a short meaningful phrase, such as "best bets" and "lovely volley." A rare *transposed triplet,* or *trianagram,* is "discounter introduces reductions." See also **pangram**.

anallagmatic curve

A curve that is invariant under inversion (see **inverse**). Examples include the **cardioid**, **Cassinian ovals**, **limaçon of Pascal**, **strophoid**, and **Maclaurin trisectrix**.

analysis

A major branch of mathematics that has to do with *approximating* certain mathematical objects, such as numbers or **function**s, in terms of other objects that are easier to understand or to handle. A simple example of analysis is the calculation of the first few decimal places of **pi** by writing it as the **limit** of an **infinite series**. The origins of analysis go back to the seventeenth century, when people such as Isaac **Newton** began investigating how to approximate locally–in the neighborhood of a point–the behavior of quantities that vary continuously. This led to an intense study of limits, which form the basis of understanding infinite series, differentiation, and integration.

Modern analysis is subdivided into several areas: real analysis (the study of derivatives and integrals of real-valued functions); functional analysis (the study of spaces of functions); **harmonic analysis** (the study of **Fourier series** and their abstractions); **complex analysis** (the study of functions from the complex plane to the complex plane that are complex differentiable); and **nonstandard analysis** (the study of **hyperreal number**s and their functions, which leads to a rigorous treatment of infinitesimals and of infinitely large numbers).

analytical geometry

Also known as *coordinate geometry* or *Cartesian geometry,* the type of geometry that describes points, lines, and shapes in terms of **coordinate**s, and that uses **algebra** to prove things about these objects by considering their coordinates. René **Descartes** laid down the foundations for analytical geometry in 1637 in his *Discourse on the Method of Rightly Conducting the Reason in the Search for Truth in the Sciences,* commonly referred to as *Discourse on Method.* This work provided the basis for **calculus**, which was introduced later by Isaac **Newton** and Gottfried **Leibniz**.

analytical number theory

The branch of **number theory** that uses methods taken from **analysis**, especially **complex analysis**. It contrasts with **algebraic number theory**.

anamorphosis

The process of distorting the perspective of an image to such an extent that its normal appearance can only be restored by the observer completely changing the way he looks at the image. In *catoptric anamorphosis,* a curved mirror, usually of cylindrical or conical shape, is used to restore an anamorphic picture to its undistorted form. In other kinds of anamorphism, the observer has to change her viewing position–for example, by looking at the picture almost along its surface. Some anamorphic art adds deception by concealing the distorted image in an

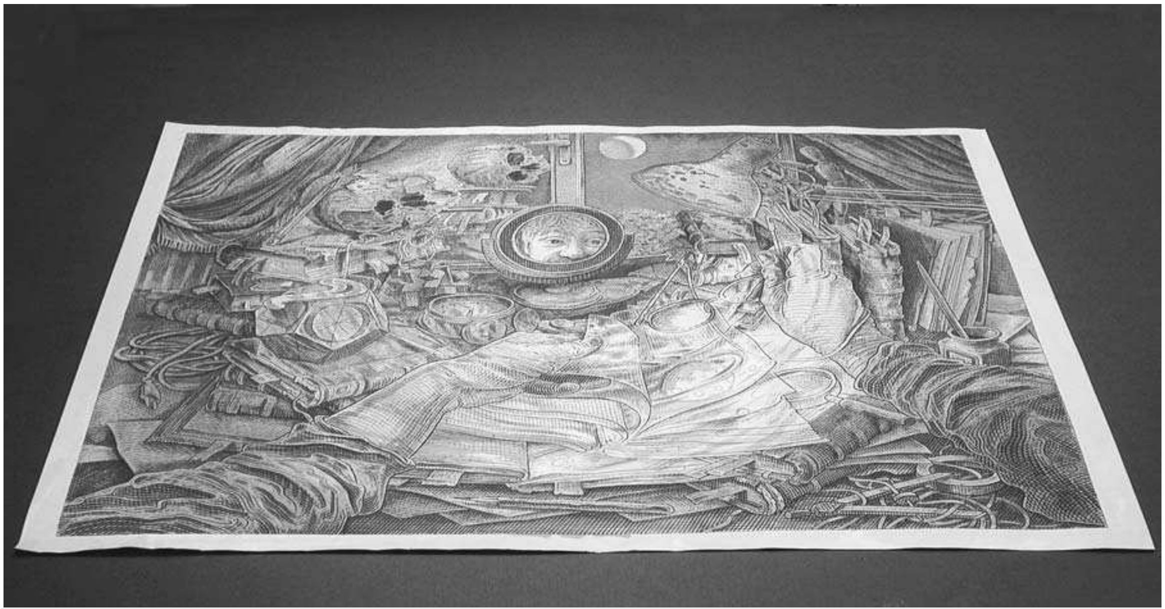

anamorphosis "Self-portrait with Albert" is a clever example of anamorphic art by the Hungarian artist Istvan Orosz. The artist's hands over his desk and a small round mirror in which the artist's face is reflected can be seen in the etching. *Istvan Orosz*

(continued)

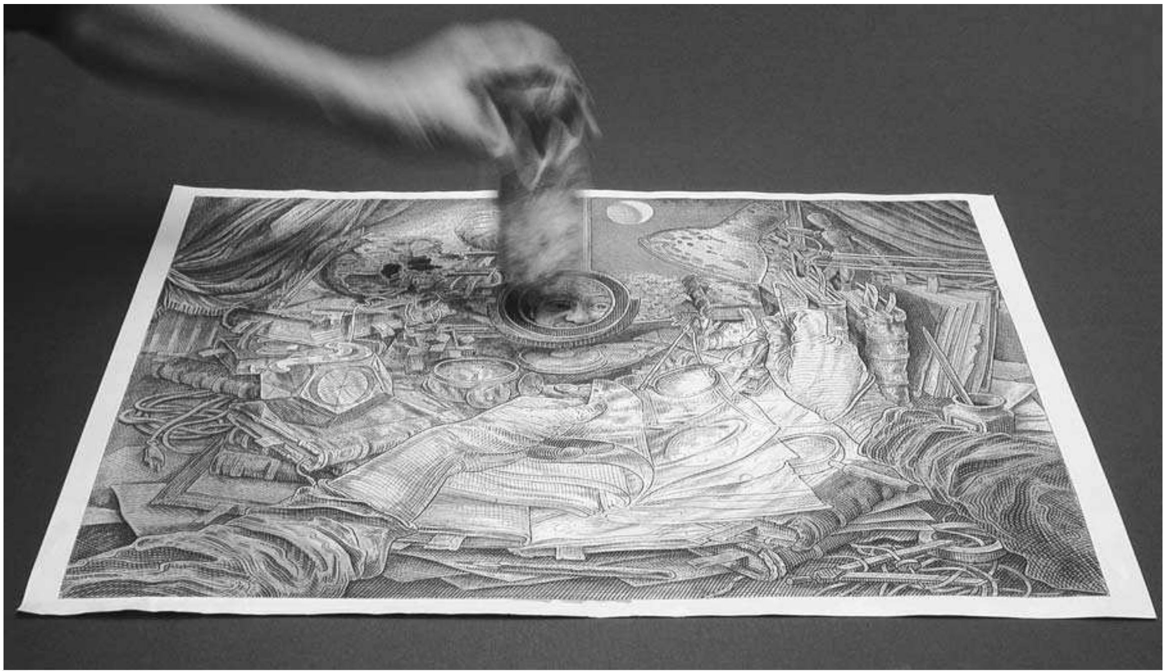

anamorphosis *(continued)* A cylindrical mirror is placed over the circle. *Istvan Orosz*

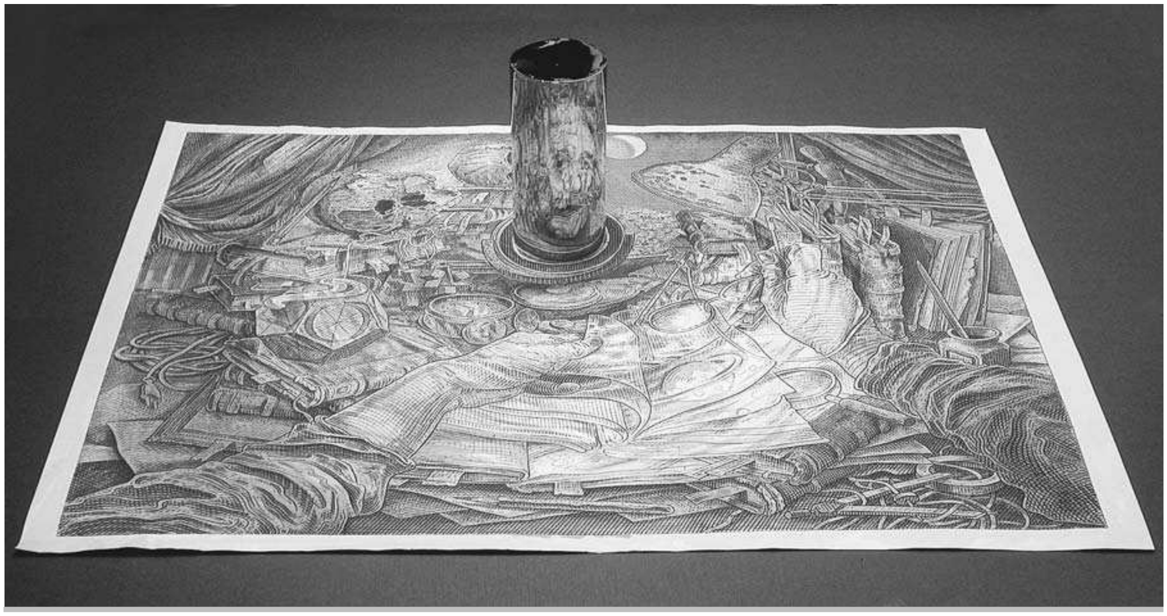

The mirror reveals a previously unsuspected aspect of the picture. The distorting effect of the curved mirror is to undistort a face hidden amid the shapes on the desk: the face of Albert Einstein. Orosz created this etching for an exhibition in Princeton, where the great scientist lived. *Istvan Orosz* otherwise normal looking picture. At one time, artists who had the mathematical knowledge to create anamorphic pictures kept their calculations and grids well-guarded secrets. Now it is relatively easy to create such images by computer.

angle

> *My geometry teacher was sometimes acute, and sometimes obtuse, but he was always right.*
>
> –Anonymous

The opening between two lines or two planes that meet; the word comes from the Latin *angulus* for "sharp bend." Angles are measured in **degree**s. A right angle has 90°, an **acute** angle less than 90°, and an **obtuse** angle has between 90° and 180°. If an angle exceeds the straight angle of 180°, it is said to be convex. *Complementary angles* add to 90°, and *supplementary angles* make a total of 180°.

angle bisection

See **bisecting an angle**.

angle trisection

See **trisecting an angle**.

animals' mathematical ability

Many different species, including rats, parrots, pigeons, raccoons, and chimpanzees, are capable of doing simple calculations. Tests on dogs have shown that they have a basic grasp of cardinality–the number of things on offer. If they're shown a pile of treats and then shown the pile again after it has been concealed and the number of treats changed slightly, they will react differently than if there's been no change. However, not all purported animal math talents stand the test of time. At the turn of the century, a horse named Clever Hans wowed audiences with his counting skills. His trainer would pose a problem, and the horse would tap out the answer. In the end, though, it was found that Hans couldn't really add or subtract but was instead responding to subtle, unintended clues from its trainer, who would visibly relax when the horse reached the correct number.

annulus

The region between the smaller and the larger of two circles that share a common center.

antigravity houses and hills

The House of Mystery in the Oregon Vortex, Gold Hill, Oregon, built during the Great Depression in the 1930s, can claim to be the first "antigravity house." It spawned many imitators around the United States and in other parts of the world. Such buildings give rise to some spectacular visual effects, which seem bewildering until the underlying cause is revealed. Of course, the visitor guides are not forthcoming about what is really going on and make fantastic claims about magnetic or gravitational anomalies, UFOs, or other weird and wonderful phenomena. The fact is that all the stunning effects stem from clever construction and concealment that make an incline seem like a horizontal in the mind of the visitor. All antigravity houses are built on hills, with a typical incline of about 25°. But unlike a normal house on the side of a hill, an antigravity house is built so that its walls are perpendicular to the (inclined) ground. In addition, the area around the house is surrounded by a tall fence that prevents the visitor from establishing a true horizontal. Thus compelled to fall back on experience, the visitor assumes that the floor of the house is horizontal and that the walls are vertical with respect to Earth's gravity. All the stunning visual shenanigans follow from this.

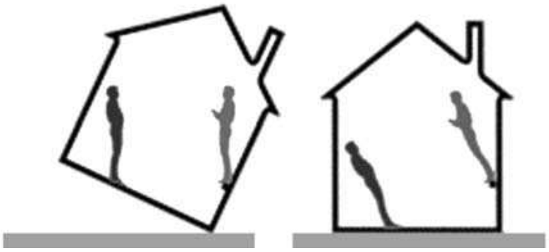

antigravity houses and hills Visitors to a "house of mystery," believing that the floor of the house is horizontal, may be astonished by the apparent gravity-defying effects (right). All these effects are easily understood, however, when it is realized that the entire house tilts at the same angle as a hill on which it is built.

In addition to man-made antigravity illusions, there are also a number of remarkable natural locations around the world where gravity seems to be out of kilter. One example is the "Electric Brae," known locally as Croy Brae, in Ayrshire, Scotland. This runs the quarter-mile from the bend overlooking the Croy railway viaduct in the west (86 meters above sea level) to the wooded Craigencroy Glen (92 meters above sea level) to the east. While there is actually a slope of 1 in 86 (a rise of 1 meter for every 86 meters horizontally) upward from the bend at the Glen, the configuration of the land on either side of the road creates the illusion that the slope runs the other way. The author is among countless folk who have parked their cars with the brakes off on this stretch of road and been amazed to see it roll apparently uphill. See also **distortion illusion**.

antimagic square

An $n \times n$ square arrangement of the numbers 1 to n^2 such that the totals of the n rows and n columns and two long diagonals form a sequence of $(2n + 2)$ consecutive integers. There are no antimagic squares of size 2×2 and 3×3 but plenty of them for larger sizes. Here is a 4×4 example:

1	13	3	12
15	9	4	10
7	2	16	8
14	6	11	5

See also **magic square**.

antiprism

A **semi-regular polyhedron** constructed from two *n-sided* **polygon**s and $2n$ triangles. An antiprism is like a **prism** in that it contains two copies of any chosen regular polygon, but is unlike a prism in that one of the copies is given a slight twist relative to the other. The polygons are connected by a band of triangles pointing alternately up and down. At each vertex, three triangles and one of the chosen polygons meet. By spacing the two polygons at the proper distance, all the triangles become equilateral. Antiprisms are named square antiprisms, pentagonal antiprisms, and so on. The simplest, the triangular antiprism, is better known as the **octahedron**.

aperiodic tiling

A **tiling** made from the same basic elements or tiles that can cover an arbitrarily large surface without ever exactly repeating itself. For a long time it was thought that whenever tiles could be used to make an aperiodic tiling, those same tiles could also be fitted together in a different way to make a **periodic tiling**. Then, in the 1960s, mathematicians began finding sets of tiles that were uniquely aperiodic. In 1966, Robert Berger produced the first set of 20,426 aperiodic tiles, and soon lowered this number to 104. Over the next few years, other mathematicians reduced the number still further.

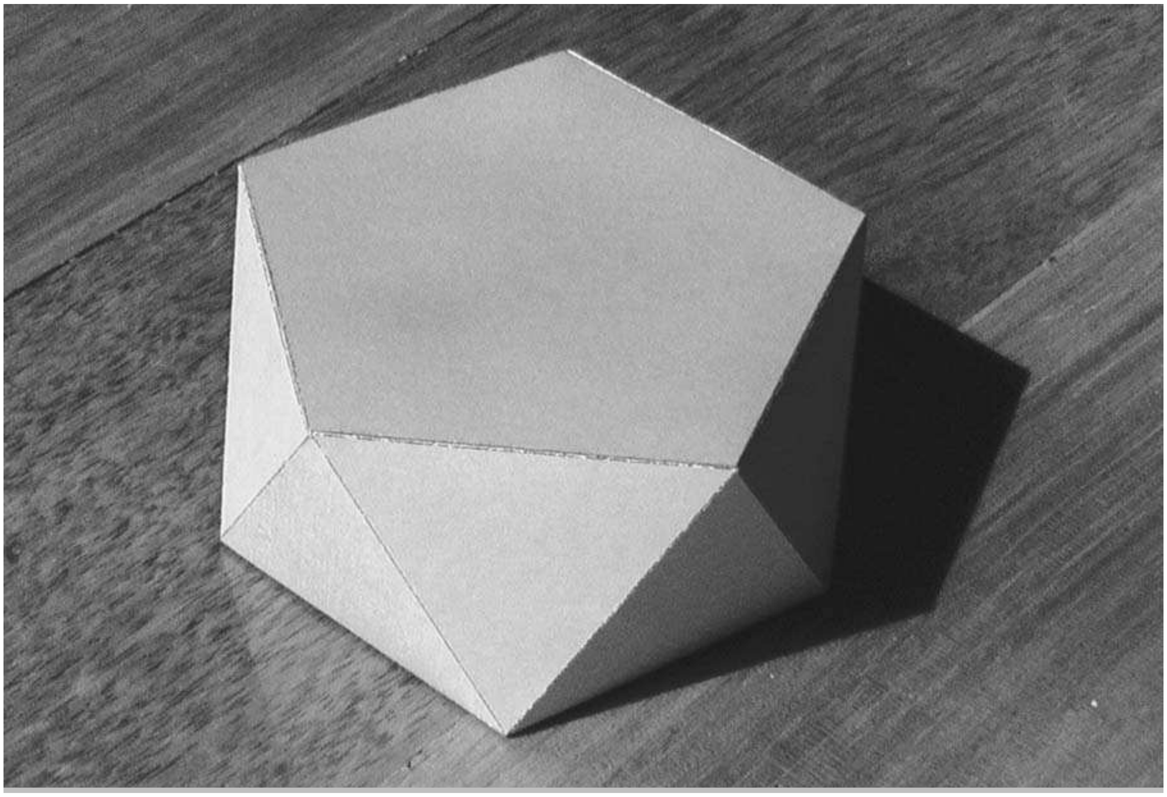

antiprism A pentagonal antiprism. *Robert Webb, www.software3d.com; created using Webb's Stella program*

In 1971, Raphael Robinson found a set of six aperiodic tiles based on notched squares; then, in 1974, Roger **Penrose** found a set of two colored aperiodic tiles (see **Penrose tilings**). The coloring can be dispensed with if the pieces are notched. There is a set of three convex (meaning no notches) aperiodic tiles, but it isn't known if there is a set of two such tiles or even a single tile (see **Einstein problem**). In three dimensions, Robert Ammann found two aperiodic polyhedra, and Ludwig Danzer found four aperiodic tetrahedra.

Apéry's constant

The number defined by the formula $\zeta/(3) = \sum_{n=1}^{\infty} 1/n^3$, where ζ is the **Riemann zeta function**: It has the value 1.202056 . . . and gives the odds (1 in 1.202056 . . .) of any three positive integers, picked at random, having no common divisor. In 1979, the French mathematician Roger Apéry (1916–1994) stunned the mathematical world with a proof that this number is irrational.[11] Whether it is a **transcendental number** remains an open question.

apex

The **vertex** of a cone or a pyramid.

apocalypse number

See **beast number**.

Apollonius of Perga (c. 255–170 B.C.)

A highly influential Greek mathematician (born in a region of what is now Turkey), known as the "Great Geometer," whose eight-part work *On Conics* introduced such terms as *ellipse, parabola,* and *hyperbola.* **Euclid** and others had written earlier about the basic properties of **conic section**s but Apollonius added many new results, particularly related to **normal**s and **tangent**s to the various conic curves. One of the most famous questions he raised is known as the **Apollonius problem**. He also wrote widely on other subjects including science, medicine, and philosophy. In *On the Burning Mirror* he showed that parallel rays of light are not brought to a focus by a spherical mirror (as had been previously thought), and he discussed the focal properties of a parabolic mirror. A few decades after his death, Emperor Hadrian collected Apollonius's works and ensured their publication throughout his realm.

Apollonius problem

A problem first recorded in *Tangencies,* written around 200 B.C. by **Apollonius of Perga**. Given three objects in the plane, each of which may be a circle *C,* a point *P* (a degenerate circle), or a line *L* (part of a circle with infinite radius), find another circle that is tangent to (just touches) each of the three. There are ten cases: *PPP, PPL, PLL, LLL, PPC, PLC, LLC, LCC, PCC, CCC.* The two easiest involve three points or three straight lines and were first solved by **Euclid**. Solutions to the eight other cases, with the exception of the three-circle problem, appeared in *Tangencies;* however, this work was lost. The most difficult case, to find a tangent circle to any three other circles, was first solved by the French mathematician François Viète (1540–1603) and involves the simultaneous solution of three **quadratic** equations, although, in principle, a solution could be found using just a compass and a straightedge. Any of the eight circles that is a solution to the general three-circle problem is called an *Apollonius circle.* If the three circles are mutually tangent then the eight solutions collapse to just two, which are known as **Soddy circle**s. A **fractal** is produced by starting with three mutually tangent circles and creating a fourth–the inner Soddy circle–that is nested between the original three. The process is repeated to yield three more circles nested between sets of three of these, and then repeated again indefinitely. The points that are never inside a circle form a fractal set called the *Apollonian gasket,* which has a fractional dimension of about 1.30568.

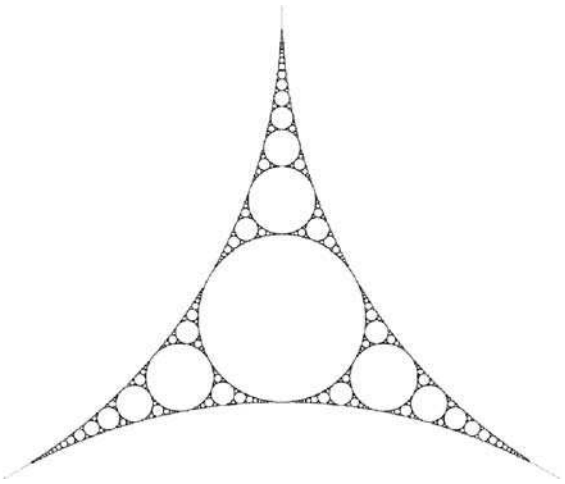

Apollonius problem The Apollonian gasket.

apothem

Also known as a *short radius,* the perpendicular distance from the center of a **regular polygon** to one of its sides. It is the same as the radius of a circle inscribed in the polygon.

apotome

One of Euclid's categories of **irrational numbers**. An apotome has the form $\sqrt{(\sqrt{A} - \sqrt{B})}$. The corresponding number with a "+" sign is called a *binomial* in Euclid's scheme.

applied mathematics

Mathematics for the sake of its use to science or society.

Arabic numeral

A **numeral** written with an Arabic digit alone: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or in combination: 10, 11, 12, . . . 594, ...

arbelos

A figure bounded by three semicircles AB, BC, and AC, where ABC is a straight line. Archimedes (about 250 B.C.) called it an *arbelos*–the Greek word for a knife of the same shape used by shoemakers to cut and trim leather–and wrote about it in his *Liber assumptorum* (Book of lemmas). Among its properties are that the sum of the two smaller arc lengths is equal to the larger; the area of the arbelos is $\pi/4$ times the product of the two smaller diameters (AB and BC); and the area of the arbelos is equal to the area of a circle whose diameter is the length of a perpendicular segment drawn from the tangent point B of the two smaller semicircles to the point D, where it meets the larger semicircle. The circles inscribed on each half of BD of the arbelos (called *Archimedes's circles*) each have a diameter of $(AB)(BC)/(AC)$. Furthermore, the smallest circumcircle of these two circles has an area equal to that of the arbelos. **Pappus of Alexandria** wrote on the relations of the chain of circles, C_1, C_2, C_3, . . . (called a *Pappus chain* or an *arbelos train*) that are mutually tangent to the two largest semicircles and to each other. The centers of these circles lie on an ellipse and the diameter of the nth circle is $(1/n)$ times the base of the perpendicular distance to the base of the semicircle.

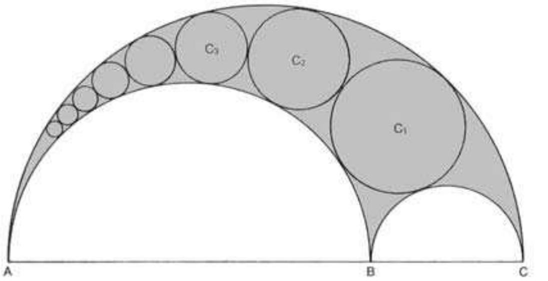

arbelos A Pappus chain of circles, C_1, C_2, C_3, . . . , inside an arbelos (shaded region).

arc

Any part of a curved line or part of the circumference of a circle; the word comes from the Latin *arcus* for a bow, which also gives rise to *arch*. *Arc length* is the distance along part of a curve.

arch

A strong, curved structure, traditionally made from wedge-shaped elements, that may take many different forms and that provides both an opening and a support for overlying material. Two common forms are the semicircular arch, first used by the Romans, and the pointed Gothic arch. The semicircular arch is the weaker of the two because it supports all the weight on the top and tends to flatten at its midpoint. It also requires massive supporting walls since all the stress on the arch acts purely downward. The pointed arch, by contrast, directs stresses both vertically and horizontally, so that the walls can be thinner, though buttressing may be required to prevent the walls from collapsing sideways. See also **Vesica Piscis**.

Archimedean dual

See **Catalan solid**.

Archimedean solid

A convex **semi-regular polyhedron**; a solid made from regular polygonal sides of two or more types that meet in a uniform pattern around each corner. (A regular polyhedron, or **Platonic solid**, has only one type of polygonal side.) There are 13 Archimedean solids (see table "Archimedian Solids"). Although they are named after their discoverer, the first surviving record of them is in the fifth book of the *Mathematical Collection* of **Pappus of Alexandria**. The **dual**s of the Archimedean solids (made by replacing each face with a vertex, and each vertex with a face) are commonly known as **Catalan solid**s. Apart from the Platonic and Archimedean solids, the only other convex uniform polyhedra with regular faces are **prism**s and **antiprism**s. This was shown by Johannes **Kepler**, who also gave the names generally used for the Archimedean solids. See also **Johnson solid**.

Archimedean spiral

A **spiral**, like that of the groove in a phonograph record, in which the distance between adjacent coils, measured radially out from the center, is constant. **Archimedes** was the first to study it and it was the main subject of his treatise *On Spirals*. The Archimedean spiral has a very simple equation in **polar coordinates** (r, θ):

$$r = a + b\theta$$

where a and b can be any real numbers. Changing the parameter a turns the spiral, while b controls the distance

Archimedean Solids

		Number of	
Name	**Vertices**	**Faces**	**Edges**
Truncated tetrahedron	8 = 4 + 4	12	18
Truncated cube	14 = 8 + 6	24	36
Truncated octahedron	14 = 6 + 8	24	36
Truncated dodecahedron	32 = 20 + 12	60	90
Truncated icosahedron	32 = 12 + 20	60	90
Cuboctahedron	14 = 8 + 6	12	24
Icosidodecahedron	32 = 20 + 12	30	60
Snub dodecahedron	92 = 80 + 12	60	150
Rhombicuboctahedron	26 = 8 + 18	24	48
Great rhombicosidodecahedron	62 = 30 + 20 + 12	120	180
Rhombicosidodecahedron	62 = 20 + 30 + 12	60	120
Great rhombicuboctahedron	26 = 8 + 12 + 6	48	72
Snub cube	38 = 32 + 6	24	60

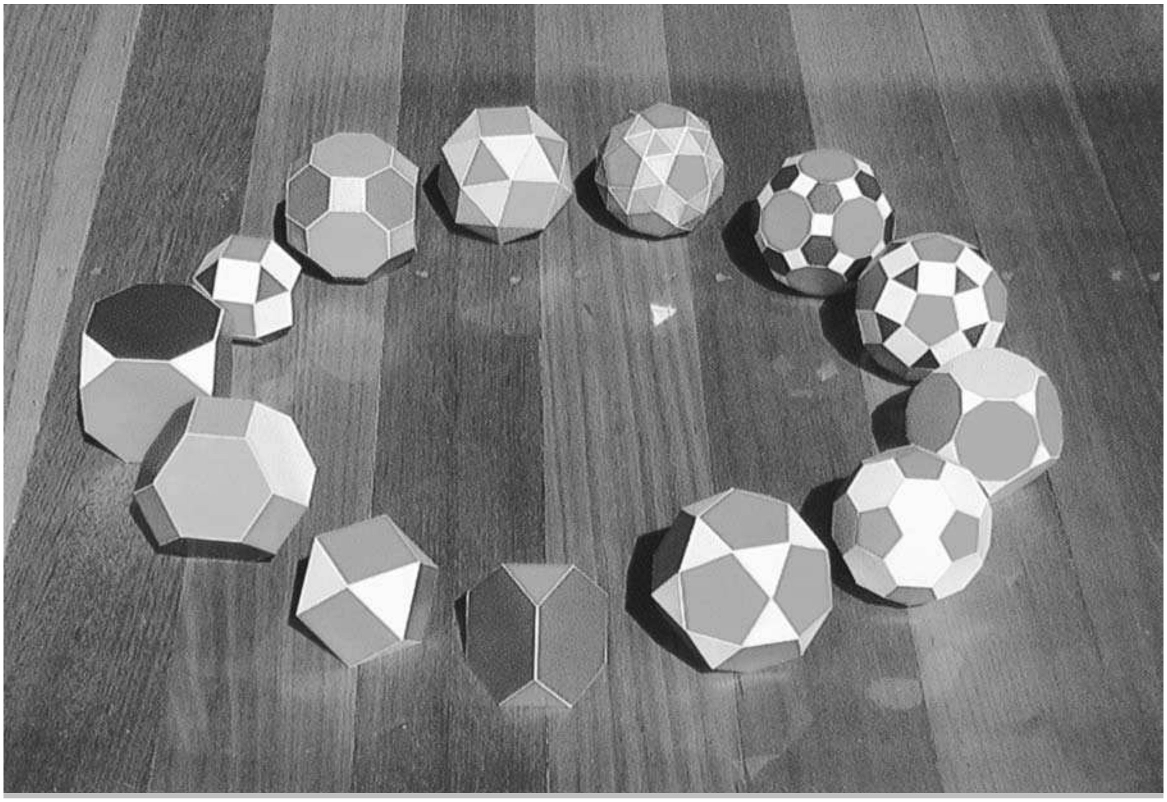

Archimedean solid The complete set of Archimedean solids, starting far left and going clockwise: truncated cube, small rhombicuboctahedron, great rhombicuboctahedron, snub cube, snub dodecahedron, great rhombicosidodecahedron, small rhombicosidodecahedron, truncated dodecahedron, truncated icosahedron (soccer ball), icosidodecahedron, truncated tetrahedron, cuboctahedron, and truncated octahedron. *Robert Webb, www.software3d.com; created using Webb's Stella program* between the arms. The Archimedean spiral is distinguished from the **logarithmic spiral** by the fact that successive arms have a fixed distance (equal to $2\pi b$ if θ is measured in radians), whereas in a logarithmic spiral these distances form a **geometric sequence**. Note that the Archimedean spiral has two possible arms that coil in opposite directions, one for $\theta > 0$ and the other for $\theta < 0$. Many examples of spirals in the man-made world, such as a watch spring or the end of a rolled carpet, are either Archimedean spirals or another curve that is very much like it, the **circle involute**.

Archimedean tessellation

Also known as a *semiregular tessellation,* a **tiling** that uses only regular **polygon**s arranged so that two or more different polygons are around each vertex and each vertex involves the same pattern of polygons. There are eight such tessellations, two involving triangles and squares, two involving triangles and hexagons, and one each involving squares and octagons; triangles and dodecagons; squares, hexagons, and dodecagons; and triangles, squares, and hexagons.

Archimedes of Syracuse (c. 287–212 B.C.)

One of the greatest mathematicians and scientists of all time. He became a popular figure because of his involvement in the defense of Syracuse against the Roman siege in the first and second Punic Wars when his war machines helped keep the Romans at bay. He also devised a scheme to move a full-size ship, complete with crew and cargo, by pulling a single rope, and invented the irrigation device known as the Archimedean screw. According to one of many legends about him, he is said to have discovered the principle of buoyancy while taking a bath and then ran into the street naked shouting "eureka" ("I found it!").

In his book *The Sand-Reckoner,* he described a positional **number system** and used it to write the equivalent of numbers up to 8×10^{64}–the number of grains of sand he thought it would take to fill the universe. He devised a rule-of-thumb method to do private calculations that closely resembles integral calculus (2,000 years before its "discovery"), but then switched to geometric proof for his results. He demonstrated that the ratio of a circle's perimeter to its diameter is the same as the ratio of the circle's area to the square of the radius. Although he didn't call this ratio "**pi**," he showed how to work it out to arbitrary accuracy and gave an approximation of it as "exceeding 3 in less than $1/7$ but more than $10/71$."

Archimedes was the first, and possibly the only, Greek mathematician to introduce mechanical curves (those traced by a moving point) as legitimate objects of study, and he used the **Archimedean spiral** to square the circle. He proved that the area and volume of the **sphere** are in the same ratio to the area and volume of a circumscribed straight cylinder, a result that pleased him so much that he made it his epitaph. Archimedes is probably also the first mathematical physicist on record, and the best before Galileo and Isaac **Newton**. He invented the field of statics, enunciated the law of the lever, the law of equilibrium of fluids, and the law of buoyancy, and was the first to identify the concept of center of gravity. He is also, perhaps erroneously, credited with the invention of a square **dissection** puzzle known as the **loculus of Archimedes**. Many of his original works were lost when the library at Alexandria burned down and they survive only in Latin or Arabic translations. Plutarch wrote of him: "Being perpetually charmed by his familiar siren, that is, by his geometry, he neglected to eat and drink and took no care of his person; that he was often carried by force to the baths, and when there he would trace geometrical figures in the ashes of the fire, and with his finger draws lines upon his body when it was anointed with oil, being in a state of great ecstasy and divinely possessed by his science."

Archimedes's cattle problem

A fiendishly hard problem involving very **large numbers** that **Archimedes** presented in a 44-line letter to **Eratosthenes,** the chief librarian at Alexandria. It ran as follows:

> If thou art diligent and wise, O stranger, compute the number of cattle of the Sun, who once upon a time grazed on the fields of the Thrinacian isle of Sicily, divided into four herds of different colors, one milk white, another a glossy black, a third yellow and the last dappled. In each herd were bulls, mighty in number according to these proportions: Understand, stranger, that the white bulls were equal to a half and a third of the black together with the whole of the yellow, while the black were equal to the fourth part of the dappled and a fifth, together with, once more, the whole of the yellow. Observe further that the remaining bulls, the dappled, were equal to a sixth part of the white and a seventh, together with all of the yellow. These were the proportions of the cows: The white were precisely equal to the third part and a fourth of the whole herd of the black; while the black were equal to the fourth part once more of the dappled and with it a fifth part, when all, including the bulls,

> went to pasture together. Now the dappled in four parts were equal in number to a fifth part and a sixth of the yellow herd. Finally the yellow were in number equal to a sixth part and a seventh of the white herd. If thou canst accurately tell, O stranger, the number of cattle of the Sun, giving separately the number of well-fed bulls and again the number of females according to each color, thou wouldst not be called unskilled or ignorant of numbers, but not yet shalt thou be numbered among the wise.
>
> But come, understand also all these conditions regarding the cattle of the Sun. When the white bulls mingled their number with the black, they stood firm, equal in depth and breadth, and the plains of Thrinacia, stretching far in all ways, were filled with their multitude. Again, when the yellow and the dappled bulls were gathered into one herd they stood in such a manner that their number, beginning from one, grew slowly greater till it completed a triangular figure, there being no bulls of other colors in their midst nor none of them lacking. If thou art able, O stranger, to find out all these things and gather them together in your mind, giving all the relations, thou shalt depart crowned with glory and knowing that thou hast been adjudged perfect in this species of wisdom.

The answer to the first part of the problem–the smallest solution for the total number of cattle–turns out to be 50,389,082. But when the extra two constraints in the second part are factored in, the solution is vastly larger. The approximate answer of 7.76×10^{202544} was found in 1880 by A. Amthor, having reduced the problem to a form called a **Pell equation**.[9] His calculations were continued by an ad hoc group called the Hillsboro Mathematical Club, of Hillsboro, Illinois, between 1889 and 1893. The club's three members (Edmund Fish, George Richards, and A. H. Bell) calculated the first 31 digits and the last 12 digits of the smallest total number of cattle to be

77602714064868182695302328332**09 . . .** 719455081800

though the two digits in bold should be 13.[31] In 1931, a correspondent to the *New York Times* wrote: "Since it has been calculated that it would take the work of a thousand men for a thousand years to determine the complete [exact] number [of cattle], it is obvious that the world will never have a complete solution." But *obvious* and *never* are words designed to make fools of prognosticators. Enter the computer. In 1965, with the help of an IBM 7040, H. C. Williams, R. A. German, and C. R. Zarnke reported a complete solution to the cattle problem, though it was 1981 before all 202,545 digits were published, by Harry Nelson, who used a Cray-1 supercomputer to generate the answer, which begins: $7.760271406486818269530232833213 \ldots \times 10^{202544}$.[341]

Archimedes's square

See **loculus of Archimedes**.

area

A measure of surface extension in two-dimensional space. *Area* is the Latin word for a vacant piece of level ground and still carries this common meaning. The French shortened form *are* denotes a square of land with a side length of 10 meters, that is, an area of 100 square meters. A *hectare* is a hundred are.

area codes

North American telephone area codes seem to have been chosen at random. But there was a method to their selection. In the mid-1950s when direct dialing of long-distance calls first became possible, it made sense to assign area codes that took the shortest time to dial to the larger cities. Almost all calls were from rotary dials. Area codes such as 212, 213, 312, and 313 took very little time for the dial to return to its starting position compared, for example, to numbers such as 809, 908, 709. The quickest-to-dial area codes were assigned to the places expected to receive the most direct-dialed calls. New York City got 212, Chicago 312, Los Angeles 213, and Washington, D.C., 202, which is a little longer to dial than 212, but much shorter than others. In order of decreasing size and estimated amount of telephone traffic, the numbers grew larger: San Francisco got 415, Miami 305, and so on. At the other end of the spectrum came places like Hawaii (the last state annexed in 1959) with 808, Puerto Rico with 809, and Newfoundland with 709. The original plan (still in use until about 1993) was that area codes had a certain construction to the numbers: the first digit is 2 through 9, the second digit is 0 or 1, and the third digit is 1 through 9. Three-digit numbers with two zeros are special codes, that is, 700, 800, or 900. Three-digit numbers with two ones are for special local codes such as 411 for local directory assistance, 611 for repairs, and so forth.

Argand diagram

A way of representing **complex number**s as points on a coordinate plane, also known as the *Argand plane* or the *complex plane*, using the x-axis as the real axis and the y-axis as the imaginary axis. It is named for the French amateur mathematician Jean Robert Argand

(1768–1822) who described it in a paper in 1806.[14] John **Wallis** suggested a similar method 120 years earlier and Casper **Wessel** extensively developed it. But Wessel's paper was published in Danish and wasn't circulated in the languages more common to mathematics at that time. In fact, it wasn't until 1895 that his paper came to the attention of the mathematical community–long after the name "Argand diagram" had stuck.

argument

(1) The input for a **function**. (2) The angle between ZO, where Z is the point representing a complex number on an **Argand diagram** and O is the origin, and the real axis. (3) A mathematical proof, possibly an informal one.

Aristotle's wheel

A **paradox** mentioned in the ancient Greek text *Mechanica,* whose author is unknown but is suspected by some to have been Aristotle. The paradox concerns two concentric circles on a wheel, as shown in the diagram. A **one-to-one** correspondence exists between points on the larger circle and those on the smaller circle. Therefore, the wheel should travel the same distance regardless of whether it is rolled from left to right on the top straight line or on the bottom one. This seems to imply that the two circumferences of the different-sized circles are equal, which is impossible. How can this apparent contradiction be resolved? The key lies in the (false) assumption that a one-to-one correspondence of points means that two curves must have the same length. In fact, the cardinalities of points in a line segment of any length (or even an infinitely long line or an infinitely large n-dimensional Euclidean space) are all the same. See also **infinity**.

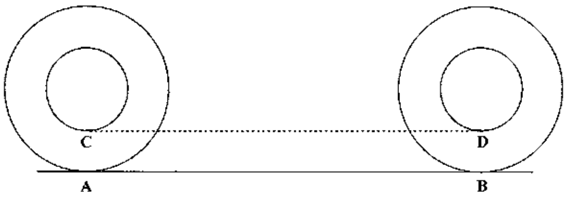

Aristotle's wheel The outer circle turns once when going from A to B, as does the inner circle when going from C to D. Yet AB is the same length as CD. How can this be, since the circles are a different size?

arithmetic

A branch of mathematics concerned with doing calculations with numbers using addition, subtraction, multiplication, and division.

arithmetic mean

The sum of n given numbers divided by n. See also **geometric mean** and **harmonic mean**.

arithmetic sequence

Also known as an *arithmetic progression,* a finite **sequence** of at least three numbers, or an infinite sequence, whose terms differ by a constant, known as the *common difference.* For example, starting with 1 and using a common difference of 4 we can get the finite arithmetic sequence: 1, 5, 9, 13, 17, 21, and also the infinite sequence 1, 5, 9, 13, 17, 21, 25, 29, . . . , $4n + 1$, ... In general, the terms of an arithmetic sequence with the first term a_0 and common difference d, have the form $a_n = dn + a_0$ ($n = 1, 2, 3, \ldots$). Does every increasing sequence of integers have to contain an arithmetic progression? Surprisingly, the answer is no. To construct a counterexample, start with 0. Then for the next term in the sequence, take the *smallest possible integer* that doesn't cause an arithmetic progression to form in the sequence constructed thus far. (There must be such an integer because there are infinitely many integers beyond the last term, and only finitely many possible progressions that the new term could complete.) This gives the nonarithmetic sequence 0, 1, 3, 4, 9, 10, 12, 13, 27, 28, ...

If the terms of an arithmetic sequence are added together the result is an *arithmetic series,* $a_0 + (a_0 + d) + \ldots + (a_0 + (n - 1)d)$, the sum of which is given by:

$$S_n = n/2\ (2a_0 + (n - 1)d) = n/2\ (a_0 + a_n).$$

See also **geometric sequence**.

around the world game

See **Icosian game**.

array

A set of numbers presented in a particular pattern, usually a grid. Matrices (see **matrix**) and **vector**s are examples of arrays.

Arrow paradox

The oldest and best-known **paradox** related to voting. The American economist Kenneth Arrow (1921–) showed that it is impossible to devise a perfect democratic voting system. In his book *Social Choice and Individual Values,*[16] Arrow identified five conditions that are universally regarded as essential for any system in which social decisions are based on individual voting preferences. The Arrow paradox is that these five conditions are logically inconsistent: under certain conditions, at least one of the essential conditions will be violated.

arrowhead

See **dart**.

artificial intelligence (AI)

The subject of "making a machine behave in ways that would be called intelligent if a human were so behaving," according John McCarthy, who coined the term in 1955. How can we tell if a computer has acquired AI at a human level? One way would be to apply the **Turing test**, though not everyone agrees that this test is foolproof (see **Chinese room**). Certainly, AI has not developed at nearly the rate many of its pioneers expected back in the 1950s and 1960s. Meanwhile, progress in fields such as **neural network**s and **fuzzy logic** continues to be made, and most computer scientists have no doubt that it is only a matter of time before computers are outperforming their biological masters in a wide variety of tasks beyond those that call for mere number-crunching ability.

artificial life

A lifelike pattern that may emerge from a **cellular automaton** and appear organic in the way it moves, grows, changes shape, reproduces, aggregates, and dies. Artificial life was pioneered by the computer scientist Chris Langton, and has been researched extensively at the Santa Fe Institute. It is being used to model various **complex system**s such as ecosystems, the economy, societies and cultures, and the immune system. The study of artificial life, though controversial, promises insights into natural processes that lead to the buildup of structure in self-organizing (see **self-organization**) systems.

associative

Three numbers, x, y, and z, are said to be *associative under addition* if

$$x + (y + z) = (x + y) + z,$$

and to be *associative under multiplication* if

$$x \times (y \times z) = (x \times y) \times z.$$

In general, three elements a, b, and c of a set S are associative under the binary operation (an operation that works on two elements at a time) $*$ if

$$a * (b * c) = (a * b) * c.$$

The word incorporates the Greek root *soci,* from which we also get *social,* and may have been first used in the modern mathematical sense by William **Hamilton** around 1850. Compare with **distributive** and **commutative**.

astroid

A **hypocycloid**–the path of a point on a circle rolling inside another circle–for which the radius of the inner circle is four times smaller than that of the larger circle; this ratio results in the astroid having four cusps. The astroid was first studied by the Danish astronomer Ole Römer in 1674, in his search for better shapes for gear teeth, and later by Johann Bernoulli (1691) (see **Bernoulli family**), Gottfried **Leibniz** (1715), and Jean **d'Alembert** (1748). Its modern name comes from the Greek *aster* for "star" and was introduced in a book by Karl Ludwig von Littrow published in Vienna in 1836; before this, the curve had a variety of names, including tetracuspid (still used), cubocycloid, and paracycle. The astroid has the Cartesian equation

$$x^{2/3} + y^{2/3} = r^{2/3}$$

where r is the radius of the fixed outer circle, and $r/4$ is the radius of the rolling circle. Its area is $3\pi r^2/8$, or $3/2$ times that of the rolling circle, and its length is $6r$. The astroid is a sextic curve and also a special form of a **Lamé curve**. It has a remarkable relationship with the quadrifolium (see **rose curve**): the radial, pedal, and orthoptic of the astroid are the quadrifolium, while the cata**caustic** of the quadrifolium is the astroid. The

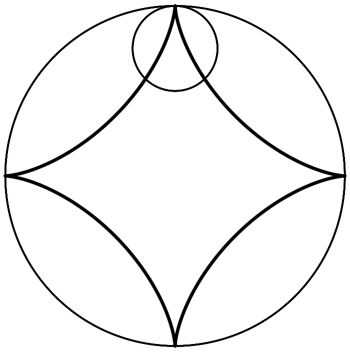

astroid As a small circle rolls around the inside of a larger one with exactly four times its circumference, a point on the rim of the small circle traces out an astroid. *© Jan Wassenaar, www.2dcurves.com* astroid is also the catacaustic of the **deltoid** and the **evolute** of the ellipse.

asymptote

A curve that gets closer and closer to a fixed straight line without ever actually touching it. Imagine facing along the direction of a great wall that is just a meter to your left. Every second, you walk forward a meter and at the same time move sideways slightly so that you halve the distance between you and the wall. The path you follow is an asymptote. The word comes from the Greek roots *a* (not), *sum* (together), and *piptein* (to fall), so that it literally means "not falling together" and was originally used in a wider sense to describe any two curves that don't intersect. **Proclus** writes about both asymptotic lines and symptotic lines (those that do cross). Nowadays, "symptotic" is almost never heard, and "asymptote" is used mainly to denote lines that serve as a limiting barrier for some curve as one of its parameters approaches plus or minus infinity. The "~" symbol is often used to show that one **function** is asymptotic to another. For example, $f(x) \sim g(x)$ indicates that the ratio of the functions $f(x)$ to $g(x)$ approaches 1 as x tends to infinity. Asymptotes are not always parallel to the x- and y-axes, as shown by the graph of $x + 1/x$, which is asymptotic to both the y-axis and the diagonal line $y = x$.

Atiyah, Michael Francis (1929–)

An English mathematician who has contributed to many topics in mathematics, notably dealing with the relationships between geometry and **analysis**. In **topology**, he developed *K-theory.* He proved the *index theorem* on the number of solutions of elliptic differential equations, linking **differential geometry**, topology, and analysis–a theorem that has been usefully applied to quantum theory. Atiyah was influential in initiating work on gauge theories and applications to nonlinear differential equations, and in making links between topology and quantum field theory. Theories of superspace and supergravity, and string theory, were all developed using ideas introduced by him.

Atomium, the

A giant steel monument in Heysel Park, Brussels, Belgium, consisting of 9 spheres that represent the body-centered cubic structure of an iron crystal magnified 150 billion times. Designed by the architect André Waterkeyn and built for the 1958 World's Fair, the 103-meter-high Atomium was originally meant to stand for only 6 months. It may be the world's largest cube. Each of its spheres have a diameter of 18 meters and are connected by escalators. Three of the upper spheres have no vertical support, and so for safety reasons are not open to the public. However, the top sphere offers a panoramic view of Brussels through its windows, and the lower spheres contain various exhibitions.

attractor

A trajectory, or set of points in **phase space**, toward which nearby orbits converge, and which is stable. Specific types of attractor include **fixed-point attractor**, **periodic attractor**, and **chaotic attractor**.

Aubel's theorem

Given a quadrilateral and a square drawn on each side of it, the two lines connecting the centers of the squares on opposite sides are perpendicular and of equal length.

autogram

See **self-enumerating sentence**.

automorphic number

Also known as an *automorph,* a number n whose square ends in n. For instance 5 is automorphic, because $5^2 = 25$, which ends in 5. A number n is called *trimorphic* if n^3 ends in n. For example $49^3 = 117{,}649$, is trimorphic. Not all trimorphic numbers are automorphic. A number n is called *tri-automorphic* if $3n^2$ ends in n; for example 6,667 is tri-automorphic because $3 \times 667^2 = 133{,}346{,}667$ ends in 7.

automorphism

An **isomorphism** from a **set** onto itself. An *automorphism group* of a **group** G is the group formed by the automorphisms of G (bijections from G to itself that preserve the multiplication). Similarly, one can consider the automorphism groups of other structures such as **ring**s and **graph**s, by looking at bijections that preserve their mathematical structure.

Avagadro constant

One of the best known examples of a **large number** in science. It is named after the Italian physicist Amedio Avagadro (1776–1856) and is defined as the number of carbon atoms in 12 grams of pure carbon, or, more generally as the number of atoms of n grams in an element with atomic weight n. It has the value $6.02214199 \times 10^{23}$.

average

A vague term that usually refers to the **arithmetic mean** but can also signify the **median**, **mode**, **geometric mean**, or weighted mean. The word stems from a commercial practice of the shipping age. The root *aver* means to declare, and the shippers of goods would declare the value of their goods. When the goods were sold, a deduction was made from each person's share, based on their declared value, for a portion of the loss or "average."

axiom

A statement that is considered to be true without need of **proof**. The term *axiom* comes from the Greek *axios* meaning "worthy" and was used by many Greek philosophers and mathematicians, including **Aristotle**. Curiously, **Euclid**, whose axioms are best known of all, seems to have favored a more general phrase meaning "common notion."

axiom of choice

An **axiom** in **set theory** that is one of the most controversial axioms in mathematics; it was formulated in 1904 by the German mathematician Ernst Zermelo (1871–1953) and, at first, seems obvious and trivial. Imagine there are many–possibly an unlimited number of–boxes in front of you, each of which has at least one thing in it. The axiom of choice (AC) says simply that you can always choose one item out of each box. More formally, if $\boldsymbol{S}$ is a collection of nonempty sets, then there exists a set that has exactly one element in common with every set S of $\boldsymbol{S}$. Put another way, there exists a function f with the property that, for each set S in the collection, $f(S)$ is a member of S. Bertrand **Russell** summed it up neatly: "To choose one sock from each of infinitely many pairs of socks requires the Axiom of Choice, but for shoes the Axiom is not needed." His point is that the two socks in a pair are identical in appearance, so, to pick one of them, we have to make an arbitrary choice. For shoes, we can use an explicit rule, such as "always choose the left shoe." Russell specifically mentions *infinitely many* pairs, because if the number is finite then AC is superfluous: we can pick one member of each pair using the definition of "nonempty" and then repeat the operation finitely many times using the rules of formal logic.

AC lies at the heart of a number of important mathematical arguments and results. For example, it is equivalent to the *well-ordering principle,* to the statement that for any two **cardinal number**s m and n, then $m < n$ or $m = n$ or $m > n$, and to *Tychonoff's theorem* (the product of any collection of compact spaces in topology is compact). Other results hinge upon it, such as the assertion that every infinite set has a denumerable subset. Yet AC was strongly attacked when it was first suggested, and still makes some mathematicians uneasy. The central issue is what it means to *choose* something from the sets in question and what it means for the choosing function to *exist.* This problem is brought into sharp focus when $\boldsymbol{S}$ happens to be the collection of all nonempty subsets of the **real number**s. No one has ever found a suitable choosing function for this collection, and there are good reasons to suspect that no one ever will. AC just mandates that there *is* such function. Because AC conjures up sets without offering workable procedures, it is said to be *nonconstructive,* as are any theorems whose proofs involve AC. Another reason that some mathematicians aren't greatly enamored with AC is that it implies the existence of some bizarre counterintuitive objects, the most famous and notorious example of which is the **Banach-Tarski paradox**. The main reason for accepting AC, as the majority of mathematicians do (albeit often reluctantly), is that it is useful. However, as a result of work by Kurt **Gödel** and, later, by Paul Cohen, it has been proven to be independent of the remaining axioms of set theory. Thus there are no contradictions in choosing to reject it; among the alternatives are to adopt a contradictory axiom or to use a completely different framework for mathematics, such as **category theory**.

axis

A line with respect to which a curve or figure is drawn, measured, rotated, and so forth. The word comes from the Greek root *aks* for a point of turning or rotation and seems to have first been used in English by Thomas Digges around 1570 in reference to the rotational axis of a right circular cone.

B

Babbage, Charles (1791–1871)

> *On two occasions I have been asked [by members of Parliament], "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.*

An English mathematician who served as Lucasian Professor of Mathematics at Cambridge (1828–1839) and became the most important figure in the prehistory of computers. Babbage noted that astronomical and other mathematical tables of the period were riddled with errors because all the calculations had to be done by hand. This gave him the idea of building a machine that would do the tedious work of computation more accurately, faster, and without ever getting tired.

In 1822, Babbage wrote a letter to one of the top British scientists of the day, Humphrey Davy, in which he talked about the design of an automatic calculator. Shortly after, he was given a grant by the British government to build this device–an elaborate symphony of rods and interlocking gear teeth–which Babbage called the Difference Engine. Construction started but never finished. Despite heroic efforts to construct a working model, the critical tolerances were beyond what engineers could provide in the first half of the eighteenth century (though the resultant gear-making skills gave Britain an edge in precision machinery for several decades and even contributed to the qualitative superiority of the British navy in World War I). The government had spent £17,000, and Babbage contributed a similar amount of his own money, on the project, when Babbage set his sights on something even more ambitious. He grasped

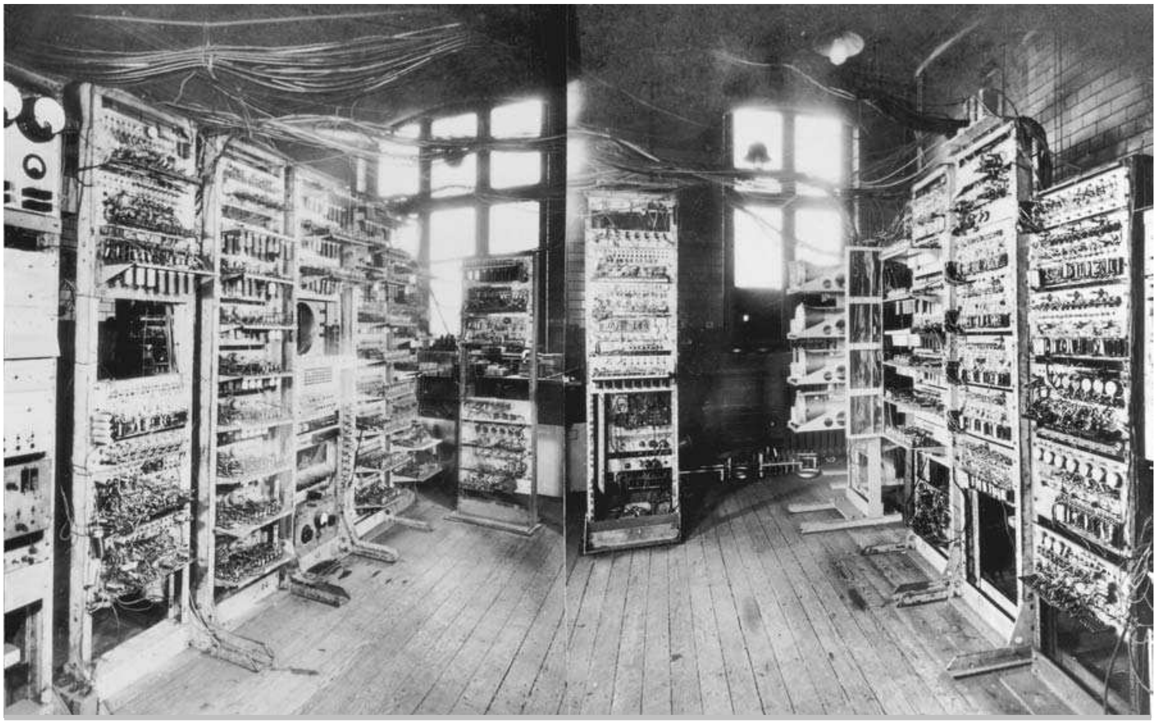

Babbage, Charles The realization of Babbage's dream: the Manchester Mark 1 computer at Manchester University, England, in 1948. This was the first computer that could store both data and programs electronically. *Ferranti Electronics Ltd.*

that the basic mechanisms of the Difference Engine could be generalized to an all-purpose calculating machine, programmable by a punched-card mechanism like that of a Jacquard loom. This vastly more powerful machine was called the Analytical Engine and would have been the world's first true computer. But it never got off the ground. "He was ill-judged enough," wrote the secretary of the Royal Astronomical Society, "to press the consideration of this new machine upon the members of Government, who were already sick of the old one." Prime Minister Robert Peel was less than enthusiastic: "I would like a little previous consideration before I move in a thin house of country gentlemen a large vote for the creation of a wooden man to calculate tables from the formula $x^2 + x + 41$." The government's eventual withdrawal of support for his schemes left Babbage a disappointed and embittered man. However, his ideas survived and proved to be the forerunner of modern computers. Parts of his uncompleted mechanisms are on display in the London Science Museum. In 1991, working from Babbage's original plans, a Difference Engine was completed–and functioned perfectly. Among Babbage's many lesser known accomplishments was his cracking of the Vigenère cipher, a discovery that helped English military campaigns but wasn't published for several years, by which time the credit had gone instead to Friedrich Kasiski, who broke the code some years after Babbage.[329] See also **Byron, Ada**.

Bachet de Méziriac, Claude-Gaspar (1581–1638)

A poet and early mathematician of the French Academy, best known for his 1621 translation of **Diophantus**'s *Arithmetica,* the book that Pierre de **Fermat** was reading when he inscribed the margin with his famous last theorem. Bachet is also remembered as a collector of mathematical puzzles, many of which he published in *Problèmes plaisans et délectables qui font par les nombres* (1612) (Pleasant and delightful problems that involve numbers), including **river-crossing problem**s, **measuring and weighing puzzle**s, number tricks, and **magic square**s. One of the puzzles is to find the least number of weights that can be used on a scale pan to weigh any integral number of pounds from 1 to 40 inclusive, if the weights can be placed in either of the scale pans. The answer is four: 1, 3, 9, and 27 pounds. On a slightly more serious note, Bachet observed that apparently every positive number can be expressed as a sum of at most four squares; for example, $5 = 2^2 + 1^2$, $6 = 2^2 + 1^2 + 1^2$, $7 = 2^2 + 1^2 + 1^2 + 1^2$, $8 = 2^2 + 2^2$, and $9 = 3^2$. The case of 7 shows that sometimes three squares wouldn't be enough. Bachet said he had checked this for more than 300 numbers but didn't know how to prove it. It wasn't until the late eighteenth century that Joseph **Lagrange** supplied a complete proof.[339]

backgammon

A gambling game for two in which each player seeks to get a set of pieces from one side of the board to the other, while trying to prevent the other player from doing the same. The distance that a piece can be moved at each turn is determined by the throw of dice.

Backgammon has roots stretching back 5,000 years. From Mesopotamia, versions of it spread to Greece and Rome as well as to India and China. The rules of the modern form of the game were largely established in England in 1743 by Edmond Hoyle but benefited from a crucial modification that emerged in American gambling clubs in the 1920s. This final innovation, which added a new level of subtlety, is known as the *doubling cube.*

Backgammon is played with two sets of 15 checkers: one player has black, the other white. The players' checkers move in opposite directions on a board with 24 spaces or *points.* Each player's goal is to be the first to bring all their own checkers "home" (into their own quarter of the board) and then "bear them off" (remove them from the board altogether). The movement of the checkers follows the outcome of a roll of two dice, the numbers on the two dice constituting separate moves.

The actual amount that changes hands at the end of the game can be more than the initial stake. For instance, in certain winning positions called *gammon* and *backgammon,* the stake is doubled or trebled, respectively. The other way the stake can change is by means of the doubling cube. If one of the players thinks that she is in a winning position, she can turn the doubling cube and *announce a double,* which means that the total stake will be doubled. If her opponent *refuses the double,* he immediately loses his (undoubled) stake and the game is finished. If he *accepts the double,* the stakes are doubled and, as a compensation, the doubling cube is handed over to him and he gets the exclusive right to announce the next double. (He is now said to *own the cube.*) If the luck of the game changes so that he later judges that he is now winning, he'll be in a position to announce a *redouble,* which means that the stake is doubled again. If the first player refuses the double, she now loses the doubled stake; if she accepts, the game goes on with a redoubled stake, four times the original value. There's no limit to how many times the stake can be doubled, but the right to announce a double switches from one player to the other every time it's exercised. (Initially either player can double–no one owns the cube.) This aspect of the game adds greatly to the variety of tactical possibilities and problems.

baker's dozen

See **thirteen**.

Bakhshali manuscript

An early mathematical manuscript, written on birch bark and found in the summer of 1881 near the village of Bakhshali in the Yusufzai subdivision of the Peshawar district (now in Pakistan). A large part of the manuscript had been destroyed and only about 70 leaves of birch bark, of which a few were mere scraps, survived to the time of its discovery. Although its date is uncertain, it is most commonly put at about the third or fourth century A.D. and appears to be a commentary on an earlier mathematical work. Among the rules and techniques it sets out for solving problems, mostly in arithmetic and algebra, but also to a lesser extent in geometry and mensuration, is this formula (stated here in modern terms) for calculating the square root of a nonsquare number Q:

$$\sqrt{Q} = \sqrt{(A^2 + b)} = A + b/2A - (b/2A)^2/[2(A + b/2A)]$$

If $Q = 41$ (so that $A = 6$ and $b = 5$) this gives $\sqrt{Q} =$ 6.403138528, which compares very favorably with the correct result of 6.403124237.

ball

Mathematicians, unlike the rest of the human race, draw a sharp distinction between a **sphere** and a ball. A sphere (in mathematics) is only a surface, whereas a ball is everything inside, and possibly including, that surface–the filling of the sphere. An *open ball* consists of all the points that are less than a given distance (the radius) away from a given point (the center); a *closed ball* consists of all the points that are less than or equal to the radius.

Mathematical balls can also exist in any number of dimensions. A one-dimensional ball of radius r is just a line segment. It consists of all the points on a line between $-r$ and r, or, in the case of a one-dimensional *unit ball* (a ball with a radius of 1), between –1 and 1. A 1 – d unit ball thus has a length, or "1 – d volume," of 2. A 2 – d unit ball, which is the filling of a unit circle, has an area, or 2 – d volume, of π. The volume of a unit ball in 3 – d is $4/3\pi$. In 4 – d it is $\pi^2/2$. Apparently, as the number of dimensions increases, so does the volume of the unit ball. What does this volume tend to do as the dimension tends to infinity? Intuitively, it might seem that in higher and higher dimensions there's more and more "room" in the unit ball, allowing its volume to become larger and larger. Does the volume become infinite, or does it approach a sufficiently large constant as the dimension increases? The answer is surprising and shows how our intuition is often misleading. Using a technique called *multivariable calculus* the volume of the unit ball in n dimensions, $V(n)$, can be shown to be $\pi^{n/2}/\ \Gamma(n/2 + 1)$, where Γ is the **gamma function** that generalizes the factorial function (i.e., $\Gamma(z + 1) = z!$). For n even, say $n = 2k$, the volume of the unit ball is thus given by $V(n) = \pi^k / k!$.

Ball, Walter William Rouse (1850–1925)

A British mathematician who lectured at Trinity College, Cambridge University, from 1878 to 1905, but is best known as a historian and as the author of the timeless classic *Mathematical Recreations and Essays*.[24] It was first published in 1892 and ran to fourteen editions, the last four with revisions by the great geometer Harold **Coxeter**.

Banach, Stefan (1892–1945)

A great Polish mathematician who founded functional **analysis** and also made important contributions to the understanding of **vector space**s, **measure theory**, and **set theory**. His name is associated with *Banach space, Banach algebra,* the *Hahn-Banach theorem,* and the remarkable **Banach-Tarski paradox**. Largely self-taught in mathematics, Banach was "discovered" by Hugo **Steinhaus** and when World War II began was president of the Polish Mathematical Society and a full professor at Lvov University. Being on good terms with Soviet mathematicians, he was allowed to hold his chair during the Soviet occupation of Lvov. The German occupation of the city in 1941 resulted in the mass murder of Polish academics. Banach survived, but the only way he could earn a living was by feeding lice with his blood in a German institute where typhoid fever research was conducted. His health declined during the occupation, and Banach died before he could be repatriated from Lvov, which was incorporated into the Soviet Union and returned to Poland after the war. *Théorie des opérations linéaires* (The theory of linear operations) is regarded as his most influential work.

Banach-Tarski paradox

> *There are more things in heaven and earth, Horatio,*
> *than are dreamt of in your philosophy.*
>
> –William Shakespeare

A seemingly bizarre and outrageous claim that it is possible to take a ball, break it into a number of pieces, and then reassemble those pieces to make two identical copies of the ball. The claim can be made even stronger: it is possible to decompose a ball the size of a marble and then reassemble the pieces to make another ball the size of Earth, or, indeed, the size of the known universe!

Before writing off Banach and Tarski as being either very bad mathematicians or very good practical jokers, it's important to understand that this is not a claim about what can actually be done with a real ball, a sharp knife, and some dabs of glue. Nor is there any chance of some entrepreneur being able to slice up a gold ingot and assemble in its place two new ones like the original. The Banach-Tarski paradox tells us nothing new about the physics of the world around us but a great deal about how volume, space, and other familiar-sounding things can assume unfamiliar guises in the strange abstract world of mathematics.

Stefan **Banach** and Alfred **Tarski** announced their startling conclusion in 1924, having built on earlier work by Felix **Hausdorff**, who proved that it's possible to chop up the unit interval (the line segment from 0 to 1) into countably many pieces, slide these bits around, and fit them together to make an interval of length 2. The Banach-Tarski paradox, which mathematicians often refer to as the *Banach-Tarski decomposition* because it's really a proof not a paradox, highlights the fact that among the infinite **set** of points that make up a mathematical **ball**, the concepts of volume and measure can't be defined for all possible subsets. What this boils down to is that quantities that can be measured in any familiar sense are not necessarily preserved when a ball is broken down into subsets and then those subsets reassembled in a different way using just translations (slides) and rotations (turns). These unmeasurable subsets are extremely complex, lacking reasonable boundaries and volume in the ordinary sense, and thus are not attainable in the real world of matter and energy. In any case, the Banach-Tarski paradox doesn't give a prescription for *how* to produce the subsets: it only proves their *existence* and that there must be at least five of them to produce a second copy of the original ball. The fact that the Banach-Tarski paradox depends on the **axiom of choice** (AC), yet is so strongly counterintuitive, has been used by some mathematics to suggest that the AC must be wrong; however, the benefits of adopting the AC are so great that such black sheep of the mathematical family as the paradox are generally tolerated.[324, 340]

Bang's theorem

If all the faces of a **tetrahedron** have the same perimeter, then the faces are all **congruent** triangles.

banker's rounding

For banking or scientific purposes it's often considered correct to round something 0.5 to the nearest *even* number (not always upward). For instance, 5.5 rounds to 6, but 12.5 rounds to 12. This method avoids introducing a bias to a large set of numbers, by rounding up more or less as often as rounding down. Unfortunately, at a lower level, it is often taught to round something 0.5 upward all the time. See also **round-off error**.

Barbaro, Daniele (1513–1570)

A Venetian geometer whose book, *La Practica della Perspectiva* (1568–9), presents the techniques of perspective, illustrated in part with a range of **polyhedra**. Partly based on the methods and writings of the great artist Piero della Francesca (1416–1492), but written in a more readable and humanistic style, it includes the earliest drawing of the *truncated icosidodecahedron* and one of the earliest representations, along with that of the German goldsmith Wenzel Jamnitzer (1508–1585), of a *rhombicosidodecahedron*. *La Practica* was one of the most respected texts on perspective in the sixteenth century, comparable to Albrecht **Dürer**'s *Painter's Manual.*

barber paradox

See **Russell's paradox**.

Barbier's theorem

See **curve of constant width**.

Barlow, Peter (1776–1862)

A self-educated English mathematician who wrote several important books on the subject but is best known for *New Mathematical Tables* (generally known as *Barlow's Tables*), a compendium of factors, squares, cubes, square roots, reciprocals, and hyperbolic logarithms of all numbers from 1 to 10,000, and his invention of a special telescope lens. "Barlows" are popularly used by amateur astronomers to this day to multiply the power of other lenses. Barlow also worked on the design of bridges and was appointed as royal commissioner for railways, conducting experiments to see if the limitation on gradients and radius of curvature proposed by George Stephenson was correct.

Barnsley's fern

A **fractal** shape, first explored by Michael F. Barnsley at the Georgia Institute of Technology in the 1980s, that has many geometric features in common with a natural fern,

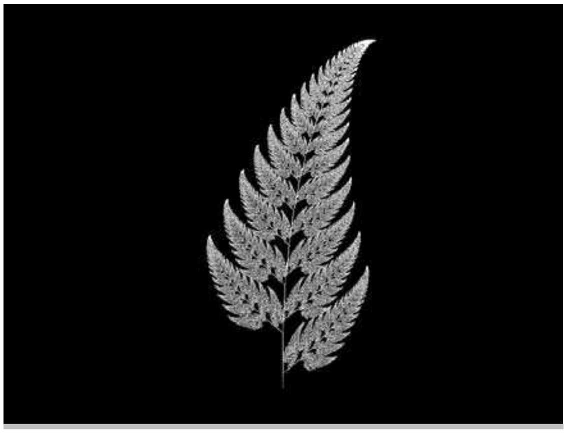

Barnsley's fern *David Nicholls*

most notably the appearance of frondlike forms at different scales. As in the case of real ferns, Barnsley's fern reveals smaller prominences along the edge of each frond that are miniature versions of the overall figure. Along these small prominences are still smaller protuberances, and so on. Barnsley's fern is created by the repetitive application of four relatively simple mathematical rules and is a type of fractal, introduced by Barnsley, known as an *iterated function system* (IFS).[26]

base

(1) The flat plane or straight line upon which a shape or a solid rests. (2) The number upon which a **number system** is based; this is also the number of different characters or figures needed by the number system. The base, or *radix,* of our familiar decimal system is 10. Thus, there are ten symbols, 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9, and a decimal number is written right to left in terms of units, tens, hundreds, and so on. Each move to the left represents a jump by a **power** of 10. The decimal number 375, for example, equals $(3 \times 10^2) + (7 \times 10) + (5 \times 1)$. This can easily be written to another base. Decimal 375_{10} becomes in octal (base 8) $567_8 = (5 \times 8^2) + (6 \times 8) + (7 \times 1)$, or in binary (base 2) 101111001_2.

basin of attraction

The set of all points in **phase space** that are under the influence of an **attractor**, or, more generally, the initial conditions of a system that evolve into the range of behavior allowed by the attractor. If one imagines a **complex system** as a sink, then the attractor can be considered the drain at the bottom, and the basin of attraction is the sink's basin.

basis

In mathematics, usually associated with linear algebra; a minimal set of **vector**s that spans a **vector space**.

Bayes, Thomas (1702–1761)

An English mathematician and theologian, remembered chiefly for the theorem named after him (see **Bayes's theorem**), and the technique of **Bayesian inference** that arises from it. Bayes wrote on probability theory, the logical basis of calculus, and asymptotic series.

Bayesian inference

Statistical inference in which probabilities are interpreted not as frequencies or proportions, but rather as degrees of belief. A prior distribution for a certain random variable is assumed; then this is modified, in the light of experimentation, using **Bayes's theorem**. Pierre **Laplace** applied Bayesian inference to estimate the mass of Saturn and in a variety of other problems.

Bayes's theorem

Also known as *Bayes's rule,* a result in **probability theory**, named after Thomas **Bayes**, who proved a special case of it. It is used in statistical inference to update estimates of the probability that different hypotheses are true, based on observations and a knowledge of how likely those observations are, given each hypothesis. In fact, it is habitually used by scientists in preference to the principle of **induction**. Bayes's theorem says that if an instance X is actually observed, then the probability of a hypothesis H must be multiplied by the following ratio:

$$\frac{\text{probability of observing } X \text{ if } H \text{ is true}}{\text{probability of observing } X}$$

In other words, the probability of a hypothesis H conditional on a given body of data X is equal to the ratio of the unconditional probability of the conjunction of the hypothesis with the data to the unconditional probability of the data alone.

Beale cipher

One of the greatest unsolved puzzles in **cryptography**–or a mere hoax. About a century ago, a fellow by the name of Thomas Beale supposedly buried two wagonloads of pots filled with silver coins in Bedford County, near Roanoke, Virginia. Local rumors claim the treasure was buried near Bedford Lake. Beale wrote three encoded letters telling what was buried, where it was buried, and to whom it belonged. He entrusted these three letters to a friend, went west, and was never heard from again. Several years later, someone examined the letters and was able to crack the code in the second one, which turned out to be based on the text from the Declaration of Independence. A number in the letter indicated which word in the document was to be used. The first letter of that word replaced the number. For example, if the first four words of the document were "We hold these truths," the number 3 in the letter would represent the letter *t.* The second letter translated as follows:

> I have deposited in the county of Bedford about four miles from Bufords in an excavation or vault six feet below the surface of the ground the following articles belonging jointly to the parties whose names are given in number three herewith. The first deposit consisted of ten hundred and fourteen pounds of gold and thirty eight hundred and twelve pounds of silver deposited Nov eighteen nineteen. The second was made Dec eighteen twenty one and consisted of nineteen hundred and seven pounds of gold and twelve hundred and eighty eight of silver, also jewels obtained in St. Louis in exchange to save transportation and valued at thirteen [t]housand dollars. The above is securely

> packed i[n] [i]ron pots with iron cov[e]rs. Th[e] vault is roughly lined with stone and the vessels rest on solid stone and are covered [w]ith others. Paper number one describes th[e] exact locality of the va[u]lt so that no difficulty will be had in finding it.

One of the remaining letters supposedly contains directions on how to find the treasure but, to date, no one has solved the code. One theory is that both the remaining letters are encoded using either the same document in a different way, or another very public document. Or, of course, this could all be an elaborate, but entertaining, wheeze. Those interested may wish to contact the Beale Cypher Association, P.O. Box 975, Beaver Falls, PA 15010.

Beal's conjecture

In 1997, the Texan financier Andrew Beal offered $75,000, later increased to $100,000, to the first person who could prove or provide a counterexample to the following conjecture:

> If $x^m + y^n = z^r$, where *x, y, z m, n,* and *r* are all positive integers, and *m, n* and *r* are greater than two, then *x, y,* and *z* have a common factor (greater than one).

Fermat's last theorem, which was proved in 1994, is a special case of Beal's conjecture. However, no one has yet been able to use this fact to prove or disprove the conjecture, nor has anyone been able to come up with a counterexample as a disproof. It is known that for any set of three exponents *m, n,* and *r,* each greater than two, there can be at most finitely many solutions. But is this finite number zero? The prize remains unclaimed.

beast number (666)

Also known as the *Apocalypse number,* the "number of the beast" mentioned in the Bible's book of Revelation, the relevant verse of which (Rev. 13:18) is often quoted as:

> Here is wisdom. Let him that hath understanding count the number of the beast: for it is the number of a man; and his number is six hundred threescore and six.

Leaving aside the thorny issue of what this actually means, the number 666 does have some interesting mathematical properties. Most notably, it is the sum of the first 36 natural numbers (all the numbers on a roulette wheel): $1 + 2 + 3 + \ldots + 36$, which makes it the thirty-sixth **triangular number**. It is also the sum of the squares of the first seven **prime numbers**, $2^2 + 3^2 + 5^2 + 7^2 + 11^2 + 13^2 + 17^2$. Other curious representations of "the beast" include:

$$1^3 + 2^3 + 3^3 + 4^3 + 5^3 + 6^3 + 5^3 + 4^3 + 3^3 + 2^3 + 1^3$$
$$3^6 - 2^6 + 1^6$$
$$6 + 6 + 6 + 6^3 + 6^3 + 6^3.$$

Furthermore, 666 is one member of a **Pythagorean triplet** (216, 630, 666), which can be written in the remarkable form:

$$(6 \times 6 \times 6)^2 + (666 - 6 \times 6)^2 = 666^2.$$

In Roman numerals, 666 represents all the numbers from 500 in descending order, namely D (500) + C (100) + L (50) + X (10) + V (5) + I (1), or DCLXVI. In fact, it's been suggested that the Roman representation of 666 may have something to do with the biblical reference. DCLXVI was often used as a generic way of referring to any unspecified or unknown large number–the Roman equivalent of our "zillion." Thus, the writer of Revelation might simply have been using "666" to mean "big but unspecified."

Beatty sequences

Suppose R is an **irrational number** greater than 1, and let S be the number satisfying the equation $1/R + 1/S = 1$. Let $[x]$ denote the *floor function* of *x,* that is, the greatest integer less than or equal to *x.* Then the sequences $[nR]$ and $[nS]$, where n ranges through the set N of positive integers, are the Beatty sequences determined by *R.* The interesting thing about them is that they partition *N;* in other words, every positive integer occurs exactly once in one sequence or the other. For example, when R is the **golden ratio** (about 1.618), the two sequences begin with

1, 3, 4, 6, 8, 9, 11, 12, 14, 16, 17, 19, 21, . . . , and
2, 5, 7, 10, 13, 15, 18, 20, 23, 26, 28, 31, 34 . . . .

Beatty sequences are named after the American mathematician Samuel Beatty (1881–1970), who introduced them in 1926 in a problem in the *American Mathematical Monthly.* Beatty was the first person to receive a Ph.D. in mathematics from a Canadian university, and later became the chairman of the mathematics department and chancellor of the University of Toronto.

beauty and mathematics

Many mathematicians and scientists have commented on the beauty they find in the structure and symmetry of the equations that underpin their work, and that beauty is often the first sign of truth. In *A Mathematician's Apology,*[151] G. H. **Hardy** wrote:

> The mathematician's patterns, like the painter's or the poet's must be beautiful; the ideas, like the colors or the words must fit together in a harmonious way. Beauty is the first test: there is no permanent place in this world for ugly mathematics.

The physicist Paul **Dirac** went even further:

> I think that there is a moral to this story, namely that it is more important to have beauty in one's

> equations than to have them fit experiment. If [Erwin] Schrödinger had been more confident of his work, he could have published it some months earlier, and he could have published a more accurate equation. It seems that if one is working from the point of view of getting beauty in one's equations, and if one has really a sound insight, one is on a sure line of progress. If there is not complete agreement between the results of one's work and experiment, one should not allow oneself to be too discouraged, because the discrepancy may well be due to minor features that are not properly taken into account and that will get cleared up with further development of the theory.[176]

The architect Richard Buckminster Fuller also saw beauty as an acid test of truth: "When I'm working on a problem, I never think about beauty. I think only how to solve the problem. But when I have finished, if the solution is not beautiful, I know it is wrong."

Bell, Eric Temple (1883–1960)

A Scottish-born mathematician and writer who, from 1903, spent most of his life in the United States, teaching at the University of Washington from 1912 until 1926, then serving as professor of mathematics at the California Institute of Technology. He did work in **number theory** but is best remembered for his books, including *Algebraic Arithmetic* (1927) and *The Development of Mathematics* (1940), which became classics, and, at a more popular level, *Men of Mathematics* (1937)[32] and *Mathematics, Queen and Servant of Science* (1951). He was also a prolific writer of science fiction under the penname John Taine.

bell curve

The characteristic shape of the graph of a normal (Gaussian) distribution.

Bell number

Named after Eric **Bell,** one of the first to analyze them in depth, the number of ways that n distinguishable objects (such as differently colored balls) can be grouped into sets (such as buckets) if no set can be empty. For example, if there are three balls, colored red (R), green (G), and blue (B), they can be grouped in five different ways: (RGB), (RG)(B), (RB)(G), (BG)(R), and (R)(G)(B), so that the third Bell number is 5. The sequence of Bell numbers, 1, 2, 5, 15, 52, 203, 877, 4,140, 21,147, . . . , can be built up in the form of a triangle, as follows. The first row has just the number one. Each successive row begins with the last number of the previous row and continues by adding the number just written down to the number immediately above and to the right of it.

1
1 2
2 3 5
5 7 10 15
15 20 27 37 52
52 . . .

The Bell numbers appear down the left-hand side of the triangle. These normal Bell numbers contrast with *ordered Bell numbers,* which count the number of ways of placing n distinguishable objects (balls) into one or more *distinguishable* sets (buckets). The ordered Bell numbers are 1, 3, 13, 75, 541, 4,683, 47,293, 545,835, ... Bell numbers are related to the **Catalan numbers**.

Benford's law

If a number is chosen at random from a large table of data or statistics, such as stock quotations, populations of towns in Germany, or half-lives of radioactive atoms, the chance that the first digit is 1 is about 30.1%, that the first digit is 2 is about 17.6%, that it is 3 is 12.4%, . . . , and that it is 9 is 4.5%. These figures fit the rule that the probability that the first digit is d is $\log_{10}(1 + 1/d)$. This rule is called Benford's law after the American physicist Frank Benford, who publicized his findings in 1938.[34] The same discovery had been made 57 years earlier by the astronomer and mathematician Simon Newcomb, who noticed that the front pages of logarithm tables tended to be more dog-eared than pages later on.[232]

Benford tested thousands of different collections of data, including the surface areas of 335 rivers, specific heats and molecular weights of thousands of chemicals, baseball statistics, and the street addresses of the first 342 people listed in the book *American Men of Science.* All these seemingly unrelated sets of numbers followed the same first-digit probability pattern as the worn pages of logarithm tables suggested. In all cases, the number 1 showed up as the first digit about 30% of the time, more often than any other, and seven times more often than the number 9.

It seems extraordinary. Why shouldn't the numbers 1 to 9 take equal turns to be the first digit? Benford's findings have been verified by other researchers. The larger and more varied the sampling of numbers from different data sets, it has been found, the more closely the distribution of numbers approaches what Benford's law predicts. Moreover these probabilities are *scale invariant* and *base invariant.* For example, it doesn't matter whether the numbers are based on the dollar prices of stocks or their prices in yen or euros, nor does it matter if the numbers are in terms of stocks per dollar; provided there are enough numbers in the sample, Benford's law will hold.[164, 263]

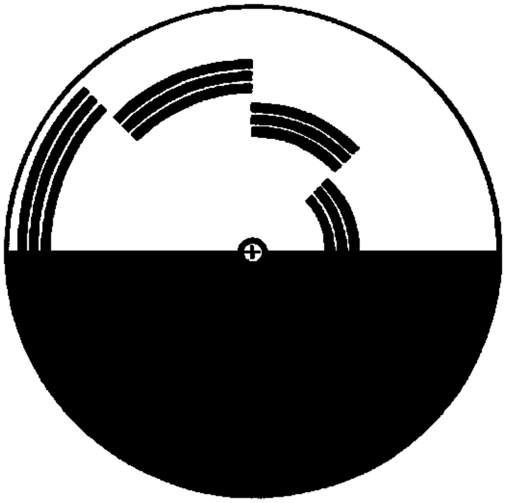

Benham's disk

Benham's disk

A disk, marked with a black and white pattern, which, when spun around, causes people to see colors. Benham's disk, also known as *Benham's wheel* and *Benham's top,* was invented in 1894 by the toy maker C. E. Benham and originally sold through Messrs. Newton and Co. as the Artificial Spectrum Top. It is one of a number of spinning disk color illusions first described by Gustav **Fechner** in 1838. For this reason, the illusory colors are known as *Fechner colors.* From the beginning, it was realized that the root of the illusion probably lay in the variation of retinal response time with wavelength. An online animated version of the disk can be seen at http://www.michaelbach.de/ot/col_benham/index.html.

Bernoulli family

An extraordinary Swiss family from Basle that produced eight outstanding mathematicians within three generations. Together with Isaac **Newton**, Gottfried **Leibniz**, Leonhard **Euler**, and Joseph **Lagrange**, the Bernoulli family dominated mathematics and physics in the seventeenth and eighteenth centuries, making important contributions to differential calculus, geometry, mechanics, ballistics, thermodynamics, hydrodynamics, optics, elasticity, magnetism, astronomy, and probability theory. Unfortunately, the Bernoullis were as conceited and arrogant as they were brilliant, and engaged in bitter rivalries and rows with one another.

The patriarchs of this mathematical dynasty were Jakob I (1654–1705) and his brother Johann I (1667–1748). (The Roman numerals are used to tell fathers, brothers, sons, and cousins apart, as the same Christian names were used repeatedly.) Next came Jakob's son, Nikolaus I, and Johann's three sons, Nikolaus II, Daniel (1700–1772), and Johann II. Finally, came Johann II's mathematical offspring, Johann III and Jakob II.

Jakob I developed a passion for science and mathematics after meeting Robert Boyle during a trip to England in 1676. He largely taught himself in these subjects and went on to lecture in experimental physics at the University of Basle. He also secretly introduced his younger brother to mathematics, against the wishes of his parents who wanted the younger brother to go into commerce. The cooperation between the two brothers soon degenerated, however, into vitriolic argument. Irked by Johann's bragging, Jakob publicly claimed that his younger brother had copied his own results. Later, having been appointed to the chair of mathematics at Basle, Jakob succeeded in blocking his brother's appointment to the same department, forcing Johann to take a teaching job at the University of Groningen instead. Johann proposed the so-called **brachistrochrone problem** and, along with Newton, Leibniz, l'Hospital, and Jakob, managed to solve it–but only after he first came up with a faulty proof and then tried to substitute one of Jakob's in its place! Eventually, Johann was offered a post at Basle as, of all things, the department head of Ancient Greek. But, en route to Basle, Johann learned that Jakob had died of tuberculosis. Upon his arrival he set about lobbying for the vacant position and, in less than two months, got his way. Jakob's most important work, his *Ars Conjectandi* (The art of conjecture), was published posthumously and formed the basis of **probability theory**.

Sadly, Johann I repeated his father's mistake and tried to force the most mathematically talented of his three sons, Daniel, into an unwanted career as a merchant. When the attempt failed, Johann let Daniel study medicine, in order to prevent his son from becoming a competitor. But all three sons followed their father's path, and Daniel, while studying medicine, was tutored in math by his older brother Nikolaus II. In 1720, Daniel went to Venice to work as a physician but won such a big reputation for his work in physics and mathematics that Peter the Great of Russia offered him a chair at the Academy of Science in St. Petersburg. Daniel went, along with Nikolaus II, who was also offered a position at the Academy. However, after just eight months, Nikolaus came down with a fever and died. Upset, Daniel wanted to return to Basle, but Johann I didn't want his son–a potential rival–back home. Instead he sent one of his pupils, none other than the great Leonhard **Euler**, to St. Petersburg to keep Daniel company. The two Swiss mathematicians became good friends, and the six years they spent together in St. Petersburg were the most productive of Daniel's life.

When Daniel finally returned to Basle, quarrels within the family flared up again after he won the prize of the Parisian Academy of Science with a paper, coauthored with his father, on astronomy. Jealous of Daniel's success, Johann threw him out of the family house. And worse was to come. In 1738 Daniel published his magnum opus, *Hydrodynamica.* Johann I read the book, hastily wrote one of his own called *Hydraulica,* back-dated it to 1732, and claimed to be the inventor of fluid dynamics! The plagiarism was soon uncovered, and Johann was ridiculed by his colleagues, but his son never recovered from the blow. See also **St. Petersburg paradox**.

Bernoulli number

A number of the type defined by Jakob **Bernoulli** in connection with evaluating sums of the form $\sum i^k$. The sequence $B_0, B_1, B_2, \ldots$ can be generated using the formula

$$x/(e^x - 1) = \sum (B_n x^n)/n!$$

though various different notations are used for them. The first few Bernoulli numbers are: $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, $B_4 = -1/30$, $B_6 = 1/42, \ldots$. They crop up in many diverse areas of mathematics including the series expansions of $\tan(x)$ and **Fermat's last theorem**.

Berry's paradox

A **paradox**, devised by G. G. Berry of the Bodleian Library at Oxford University in 1906, that involves statements of the form: "The smallest number not nameable in under ten words." At first sight, there doesn't seem anything particularly mysterious about this sentence. After all, there are only so many sentences that have less than ten words, and only a set **S** of these specify unique numbers; so there is clearly some number N that is the smallest integer not in **S**. The trouble is, the Berry sentence itself is a specification for that number *in only nine words!* Berry's paradox shows that the concept of nameability is inherently ambiguous and a dangerous concept to be used without qualification. A similar air of the paradoxical swirls around the notion of **interesting numbers.**[60]

Bertrand's box paradox

A problem, similar to the **Monty Hall problem**, that was published by the French mathematician Joseph Bertrand (1822–1900) in his 1889 text *Calcul des Probabitités.* Suppose there are three desks, each with two drawers. One desk contains a gold medal in each drawer, one contains a silver medal in each drawer, and one contains one of each, but you don't know which desk is which. If you open a drawer and find a gold medal, what are the chances that the other drawer in that desk also contains gold? This comes down, then, to figuring out the probability that you've picked the gold-gold desk instead of the gold-silver desk. Many people quickly jump to the conclusion that there are two possibilities, and since the selection was random, it must be 50-50. But this is wrong. Think of the initial selection as picking from among six drawers:

Before			After		
S	S	G		G	G
S	G	G			G
1	2	3	1	2	3

So, we have it narrowed down to three drawers, with an equal probability of each one being the one that was picked. One of the drawers is in desk 2, so there's a one-third chance that desk 2 was picked. Two of the drawers are in desk 3, so there are two one-third chances (i.e., a two-third chance) that desk 3 was picked.

Bertrand's postulate

Also known as *Betrand's conjecture,* if n is an integer greater than 3, then there is at least one **prime number** between n and $2n - 2$. This postulate (which should now be called a theorem) is named after the French mathematician Joseph Bertrand (1822–1900) who, in 1845, showed it was true for values of n up to 3 million. The Russian Pafnuty Chebyshev (1821–1894) gave the first complete proof in 1850, so that it is sometimes called *Chebyshev's theorem* (although another theorem also goes by this name). In 1932, Paul **Erdös** gave a more elegant proof, using the **binomial coefficient**s, which is the one that appears in most modern textbooks. Bertrand's postulate implies that the nth prime p_n is at most 2^n.

Bessel, Friedrich Wilhelm (1784–1846)

A German astronomer and mathematician who became director of the observatory at Königsberg (see **bridges of Königsberg**). Much of Bessel's work dealt with **perturbation**s (wobbles) in the motion of planets and stars caused by the gravitational influence of other bodies. To help analyze these perturbations he developed certain mathematical **function**s that are now known as *Bessel functions* and are used widely in physics.

beta function

The function

$$B(m, n) = \int_0^1 x^{m-1} (1 - x)^{n-1} \, dx.$$

It can be defined in terms of the **gamma function** by

$$B(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}.$$

Many integrals can be reduced to the evaluation of beta functions.

Betti number

An important topological property of a surface, named after the Italian mathematician Enrico Betti (1823–1892). The Betti number is the maximum number of cuts that can be made without dividing the surface into two separate pieces. If the surface has edges, each cut must be a "crosscut," one that goes from a point on an edge to another point on an edge. If the surface is closed, like a sphere, so that it has no edges, each cut must be a "loop cut," a cut in the form of a simple closed curve. The Betti number of a square is 0 because it is impossible to crosscut without leaving two pieces. However, if the square is folded into a tube, its topology changes–it now has two disconnected edges–and its Betti number changes to 1. A **torus**, or donut shape, has a Betti number of 2. See also **chromatic number**.

bicorn

Also known as the *cocked-hat,* any of a collection of **quartic** curves studied by James **Sylvester** in 1864 and by Arthur **Cayley** in 1867. The bicorn has the Cartesian equation

$$y^2(a^2 - x^2) = (x^2 + 2ay - a)^2.$$

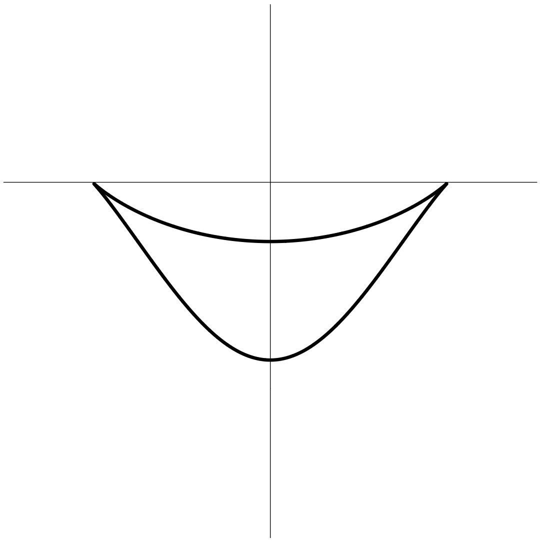

bicorn The bicorn curve © *Jan Wassenaar, www.2dcurves.com*

bicuspid curve

The **quartic** curve given by the equation

$$(x^2 - a^2)(x - a)^2 + (y^2 - a^2)^2 = 0.$$

Bieberbach conjecture

A celebrated conjecture made by the German mathematician Ludwig Bieberbach (1886–1982) in 1916, that was finally proved, after many partial results by others, by Louis de Branges of Purdue University in 1984.[54] Bierberbach is infamous in the history of mathematics because of his outspoken anti-Semitism during the Nazi era. Following the dismissal of Edmund Landau (1877–1938) from the University of Göttingen, Bierberbach wrote: "This should be seen as a prime example of the fact that representatives of overly different races do not mix as students and teachers. . . . The instincts of the Göttingen students felt that Landau was a type who handled things in an un-German manner."

Bieberbach's conjecture (BC) stemmed from the Riemann mapping theorem (RMT), which makes a claim about any region of a plane that is simply connected (in other words, any region, however complicated, that doesn't have any holes). The RMT says there must be some function, or mapping, such that every point in the arbitrary region is associated with one and only one point inside a circle with unit radius. Complex functions are best suited to plane-to-plane mappings and are often easier to work with if they can be represented as a power series. For example, given the complex number z, the function e^z can be expressed as the infinite series $1 + z + z^2/2! + z^3/3! + \ldots$. Bieberbach guessed that there is a link between the conditions imposed on a function by RMT and the numerical coefficients of the terms in a power series that represents the function. The BC says that if a function gives a one-to-one association between points in the unit circle and points in a simply connected region of the plane, the coefficients of the power series that represents the function are never larger than the corresponding power. In other words, given that $f(z) = a_0 + a_1z + a_2z^2 + a_3z^3 + \ldots$, then $|a_n| \leq n|a_1|$ for each n.

bifurcation

The value of a smoothly varying control parameter, or the point in *parameter space,* at which the behavior of a **dynamical system** undergoes a qualitative change. For example, a simple equilibrium, or a **fixed-point attractor**, might give way to a periodic oscillation as the stress on a system increases. Similarly, a **periodic attractor** might become unstable and be replaced by a **chaotic attractor**. To give a real-world example, drops fall individually at equal intervals, from a dripping faucet at low pressure. As the pressure is increased, however, the pattern of dripping abruptly changes so that two drops fall close together, with a longer interval before the next pair fall. In this case, a simple periodic process has given way to a periodic process with twice the period, a process described as "period doubling." If the flow rate of water through the faucet is increased still further, beyond the bifurcation point, often an irregular dripping is found and the behavior can become chaotic. See also **chaos**.

bilateral

Having two sides, or relating to the right and left sides of an object. *Bilateral symmetry* is a symmetrical arrangement, as of an organism or a body part, along a central axis, so that the body is divided into equivalent right and left halves by only one plane. See also **mirror reversal problem**.

bilateral diagram

A device invented by Lewis **Carroll** to represent the different logical states that two objects with two properties can take. Each cell in a four-square array represents one of the four possible object/property states and is covered by a red counter if the state is present and by a gray counter if it is absent.

billion

See **large number**s.

bimagic square

A **magic square** that becomes a new magic square when each integer is squared. If, in addition to being bimagic, the integers in the square can be cubed and the resulting square is still magic, the square is said to be *trimagic.* To date the smallest bimagic square seems to be of order 8, while the smallest trimagic square is of order 32.

binary

> *There are 10 kinds of people in the world, those who understand binary math, and those who don't.*
>
> –Anonymous

The simplest positional **number system** and the natural one to use when dealing with computers; it employs just two symbols, 0 and 1, which correspond to the possible states of an off-on switch. Each place to the left in a binary number represents the next highest power of two. The binary number 10110_2, for example, means $1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0$, or 22_{10} in the familiar decimal notation. Nonintegers can be represented by using negative powers, which are set off from the other digits by means of a *radix point* (called a *decimal point* in base 10). The binary number 11.01_2 thus means $1 \times 2^1 + 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2}$ which equals 3.25_{10}. A number that terminates in a decimal doesn't necessarily do so in binary (e.g., $0.3_{10} = 0.0100110011001\ldots_2$), and vice versa. An irrational number, however, is nonperiodic in both systems (e.g., **pi,** = $3.1415926\ldots_{10} = 11.001001000011111\ldots_2$). Binary arithmetic was first investigated by Gottfried **Leibniz** in 1672, though he didn't publish anything about it until 1701.

binary operation

An operation that involves two operands. For example, addition and subtraction are binary operations.

binomial

An expression containing two terms, joined by + or –. The **binomial theorem** gives the result of raising a binomial expression to a power; the expansion and the series it leads to are called the *binomial expansion* and the *binomial series.* A *binomial distribution* is described by a formula related to the binomial expansion. A binomial equation is a particular kind of equation that contains two terms.

binomial coefficient

A **coefficient** of x in the expansion of $(x + y)^n$. The binomial coefficient ${}_nC_m$ or $\binom{m}{n}$ gives the number of ways of picking m unordered outcomes from n possibilities, and is also known as a **combination**. It has the value $n!/(n - m)!m!$ The binomial coefficients form the rows of **Pascal's triangle**.

binomial theorem

The result that allows the expansion of a **binomial** expression:

$$(x+y)^n = x^n + a_{n-1}x^{n-1}y + a_{n-2}x^{n-2}y^2 + \ldots + y^n$$

where the coefficients a_i are called **binomial coefficient**s.

Birkhoff, George David (1884–1944)

The foremost American mathematician of the early twentieth century and the first prominent dynamicist in the New World. He is known for his work on linear **differential equation**s and **difference equation**s, and was also deeply interested in and made contributions to the analysis of **dynamical system**s, celestial mechanics, the **four-color map problem**, and function spaces. Although a geometer at heart, he discovered new symbolic solution methods. He saw beyond the theory of oscillations, created a rigorous theory of **ergodic** behavior, and foresaw dynamical models for **chaos**. In addition he wrote on the foundations of relativity and quantum mechanics and, in *Aesthetic Measure* (1933), on art and music.

birthday paradox

The fact–not really a **paradox**–that you need a group of only 23 people for there to be a better than 50/50 chance that two of these people will have the same birthday. This seems surprising because we are used to comparing our particular birthday with others and only rarely finding a perfect match. The probability of any two individuals having the same birthday is just $^1/_{365}$. Even if you were to ask 20 people, the probability of finding someone with your birthday is still less than $^1/_{20}$. But the odds improve dramatically when a group of people *ask each other* about their birthdays because then there are many more opportunities for a matchup. One way to calculate the probability of a birthday match is to count the pairs of people involved. In a room of 23 people, there are $(23 \times 22)/2$, or 253, possible pairs. Each pair has a probability of success of $^1/_{365} = 0.00274$ (0.274%), and thus a probability of failure of $(1 - 0.00274) = 0.99726$ (99.726%). The probability of no match among any of the pairs of people is 0.99726 to the 253rd power, which is 0.499 (49.9%). So the probability of a successful match is $(1 - 0.499)$, or slightly better than even (50/50). With 42 people, the probability of a birthday match climbs to 90%.

birthday surprise

Here is a way to learn someone's birthday by doing a little simple math. Ask a person to take the month number (January = 1, February = 2, and so forth) of their birthday, multiply by 5, add 6, multiply the total by 4, add 9, and multiply the new total by 5 again. Finally, have her add the number of the day on which she was born and give you the total. In your head, subtract 165 and you will have the month and day of her birth. How does this work? If M is the month number and D the day number, then after the seven steps the expression for their calculation is: $5(4(5M + 6) + 9) + D = 100M + D + 165$. Thus, if you subtract the 165, what will remain will be the month in hundreds plus the day.

bisect

To cut in half.

bisecting an angle

Splitting an angle exactly in two. The ancient Greeks knew how to easily do it using only a pair of compasses and a straightedge. Here's how: Put the point of the compass at a point O and draw a circle so that it cuts the two lines coming out from the angle. Call these intersection points A and B. Now put the point of the compass at A and draw an arc that follows within the opening of the angle. Without changing the radius at which the compass is set, move its point to B and draw another arc. Join the point where the two arcs cross, P, to O using the straight edge: angle POB is half of angle AOB. See also **trisecting an angle**.

bishops problem

To find the maximum number of bishops (chess pieces capable of moving any number of spaces along diagonals of their own color) that can be placed on an $n \times n$ chessboard in such a way that no two are attacking each other. The answer is $2n - 2$, which gives the solution 14 for a standard (8×8) chessboard. The numbers of distinct maximal arrangements for $n = 1, 2, \ldots$ bishops are 1, 4, 26, 260, 3,368, ...

bistromathics

The revolutionary new (and totally fictitious) field of mathematics in restaurants, as described by Douglas Adams in his book *Life, the Universe and Everything:*[4]

> Numbers written on restaurant bills within the confines of restaurants do not follow the same mathematical laws as numbers written on any other pieces of paper in any other parts of the Universe. This single statement took the scientific world by storm. . . . So many mathematical conferences got held in such good restaurants that many of the finest minds of a generation died of obesity and heart failure and the science of math was put back by years.

Adams explains that just as Einstein found that space and time are not an absolute but depend on the observer's movement, so numbers are not absolute, but depend on the observer's movement in restaurants:

> The first non-absolute number is the number of people for whom the table is reserved. This will vary during the course of the first three telephone calls to the restaurant, and then bear no apparent relation to the number of people who actually turn up, or to the number of people who subsequently join them after the show/match/party/gig, or to the number of people who leave when they see who else has turned up. The second non-absolute number is the given time of arrival, which is now known to be one of the most bizarre of mathematical concepts, a recipriversexcluson, a number whose existence can only be defined as being anything other than itself. In other words, the given time of arrival is the one moment of time at which it is impossible that any member of the party will arrive. . . . The third and most mysterious piece of nonabsoluteness of all lies in the relationship between the number of items on the bill, the cost of each item, the number of people at the table and what they are each prepared to pay for.

See also **large number**s.

bit

A **binary** digit, either 0 or 1. See also **byte**.

blackjack

Also known as *twenty-one,* the most popular casino game in the world and the only such game with a *fluctuating* probability: the odds of winning change with the makeup of the deck. The cards two to nine have a numerical value equal to the number printed on the card. Tens and all face cards (jack [J], queen [Q], and king [K]) have the value of 10. Aces may be counted as either 11 or one. A dealer plays against one to seven players. Every player and the dealer initially receive two cards each, dealt by the dealer. Each player's hand is played against the dealer's hand only. If a player's hand has a value closer to 21 (without going over) than the dealer's hand, the player wins. The best possible hand is known as a *blackjack* (21 in the first two cards) and consists of an ace and a ten-valued card (10, J, Q, K). The payout for a blackjack is 3-to-2: the player is paid three chips for every two chips bet. When both the player and the dealer have blackjacks, it is a normal *tie* (*push*) situation and the player retains the initial bet. The player has several choices after receiving the first two cards: (1) *Hit* or *draw:* take one or more cards to add-up to a better hand. (2) *Stand:* stop taking more cards. (3) *Double down:* double the initial amount (in cases considered more favorable). (4) *Split pairs:* if the two cards are equal in value they may be played in two separate hands. The dealer must draw until his hand adds up to 17 or more. Both the player and the dealer can go over 21, a situation known as *bust;* the player loses the bet immediately. The dealer plays his hand last, after all the players at the table. This rule creates the so-called *house edge.* John Scarne[282] was the first to calculate the house advantage at blackjack as 5.9%. However, the house edge can be cut to around 1% if the player follows certain rules. The set of rules known as *basic strategy* make blackjack one of the fairest games of any kind, almost as fair as coin tossing.

In 1962 Edward O. Thorp, an IBM computer scientist, published *Beat the Dealer,*[333] which introduced a winning method called *card counting.* This method considered the 10-valued cards and the aces as positive, and the cards 2 to 6 as negative. If the net value of the remaining deck was positive, the player must increase the bet accordingly. The method had visible results when only one deck was used and very few cards remained in the deck. Casinos responded by changing the rules dramatically. The *penetration* was introduced: not all the cards in the deck are played. Shuffling is done unexpectedly. Also, most casinos introduced *multiple-deck blackjack.*

Blanche's dissection

The simplest **dissection** of a square into rectangles of the same areas but different shapes. It is composed of seven pieces; the square is 210 units on a side, and each rectangle has area of $210^2/7 = 6{,}300$.

Bólyai, János (1802–1860)

A Hungarian mathematician who was one of the founders of **non-Euclidean geometry**, independently coming to almost the same conclusions as Nikolai **Lobachevsky**. He was initially taught by his father, Farkas, also a mathematician, then he studied at the Royal Engineering College in Vienna from 1818 to 1822. Between 1820 and 1823 he prepared his treatise on a complete system of non-Euclidean geometry, commenting, "Out of nothing I have created a strange new universe." It was published in 1832 as an appendix to an essay by his father. Carl **Gauss**, on reading the appendix, wrote to a friend saying, "I regard this young geometer Bólyai as a genius of the first order." It was not until 1848 that Bólyai learned that Lobachevsky had produced a similar piece of work in 1829. Although he never published more than the 24 pages of the appendix, Bólyai left more than 20,000 pages of manuscript of mathematical work when he died. He was an accomplished linguist, speaking nine foreign languages including Chinese and Tibetan.

book-stacking problem

How much of an overhang can be achieved by stacking books on a table before the books overbalance and fall off? Assume each book is one unit long. To balance one book on a table, the center of gravity of the book must be somewhere over the table; to achieve the maximum overhang, the center of gravity should be just over the table's edge. The maximum overhang with one book is

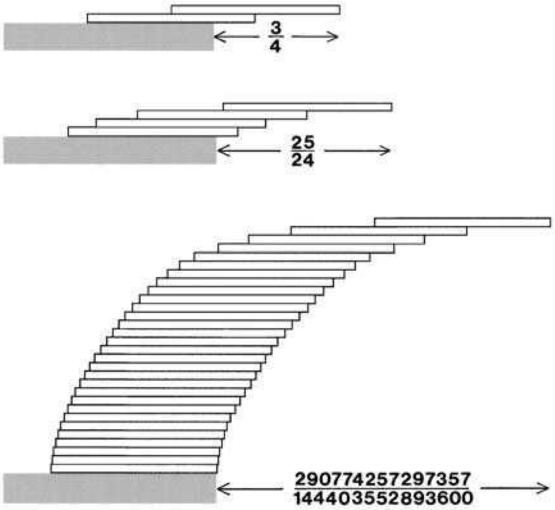

book-stacking problem The solution to the book-stacking problem.

obviously $\frac{1}{2}$ unit. For two books, the center of gravity of the first should be directly over the edge of the second, and the center of gravity of the stack of two books should be directly above the edge of the table. The center of gravity of the stack of two books is at the midpoint of the books' overlap, or $(1 + \frac{1}{2})/2$, which is $\frac{3}{4}$ unit from the far end of the top book. It turns out that the overhangs are related to the harmonic numbers H_n, (see **harmonic sequence**), which are defined as $1 + \frac{1}{2} + \frac{1}{3} + \ldots + 1/n$: the maximum overhang possible for n books is $H_n/2$. With four books, the overhang $(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4})/2$ exceeds 1, so that no part of the top book is directly over the table. With 31 books, the overhang is 2.0136 book lengths.

Boole (Stott), Alicia (1860–1940)

The third daughter of George **Boole** and an important mathematician in her own right. At the age of 18, she was introduced to a set of wooden cubes devised by her brother-in-law Charles **Hinton** as an aid to visualization of the **fourth dimension**. Despite having had no formal education, she surprised everyone by becoming adept with the cubes and developing an amazing feel for four-dimensional geometry. She introduced the word ***polytope*** to describe a four-dimensional convex solid, and went on to explore the properties of the six regular polytopes and to make 12 beautiful card models of their three-dimensional central cross sections. She sent photographs of these models to the Dutch mathematician Pieter Schoute (1846–1923), who had done similar work and with whom she subsequently published two papers. The models themselves are now housed in the Department of Pure Mathematics and Mathematical Statistics at Cambridge University.

Boole, George (1815–1864)

An English mathematician and philosopher who is regarded as one of the founders of computer science. His great contribution was to approach logic in a new way, reducing it to a simple algebra and thus incorporating logic into mathematics. He pointed out the analogy between algebraic symbols and those that represent logical forms; his algebra of logic became known as **Boolean algebra** and is now used in designing computers and analyzing logic circuits. Although he never studied for a degree, Boole was appointed to the chair of mathematics at Queens College, Cork, Ireland, in 1849. One day in 1864 he walked the two miles in pouring rain from his home to the college and then lectured in wet clothes. A fever followed but whether this alone would have caused his demise is unknown. Certainly his condition wasn't helped by his wife, Mary (a niece of Sir George Everest, after whom the mountain is named), who, following the principle that remedy should resemble cause, put Boole to bed and threw buckets of water over him. He expired shortly after. See also **Boole** (**Stott**), **Alicia**.

Boolean

Taking only 0/1, true/false, yes/no values.

Boolean algebra

An **algebra** in which the **binary** operations are chosen to model the **union** and **intersection** operations in **set theory**. For any set A, the subsets of A form a Boolean algebra under the operations of union, intersection, and complement.

Borel, Emile (1871–1956)

A French mathematician who worked on divergent series, the theory of functions, probability, and **game theory**, and was the first to define games of strategy. He also founded **measure theory**, which applies the theory of sets to the theory of functions, and thus became an originator, with Henri **Lebesgue** and René Louis Baire (1874–1932), of the modern theory of functions of a real variable.

Borges, Jorge Luis (1899–1986)

An Argentinian author, essayist, and poet, many of whose short stories explore paradoxes and other strange avenues of mathematics, logic, philosophy, and **time**. For example, the possibility of branches in time is dealt with in "The Garden of Forking Paths," while the strange notion of the **Universal Library** is the subject of "The Library of Babel." Borges was profoundly influenced by European culture, English literature, and such thinkers as George Berkeley.

Borromean rings

Three rings linked in such a way that although they can't be separated, no two rings are linked; remove any one ring, however, and the other two fall apart. Named after the Italian family of Borromeo whose family crest has borne the rings since the fifteenth century, the design has been used in many places and times as a symbol of strength in unity. A form of the Borromean link known as Odin's triangle or the walknot ("knot of the slain") was used by the Norse folk of Scandinavia in two variants: a set of Borromean triangles and a unicursal curve that makes a trefoil knot. A motif of three interlaced crescent moons, similar to the Borromean rings, can be seen at the Palace of Fontainebleau. Designed by the architect Philibert de l'Orme, it is based on the moon emblem used by Diane de Poitiers (1499–1566), mistress of King Henry II of France. A similar pattern, but with three

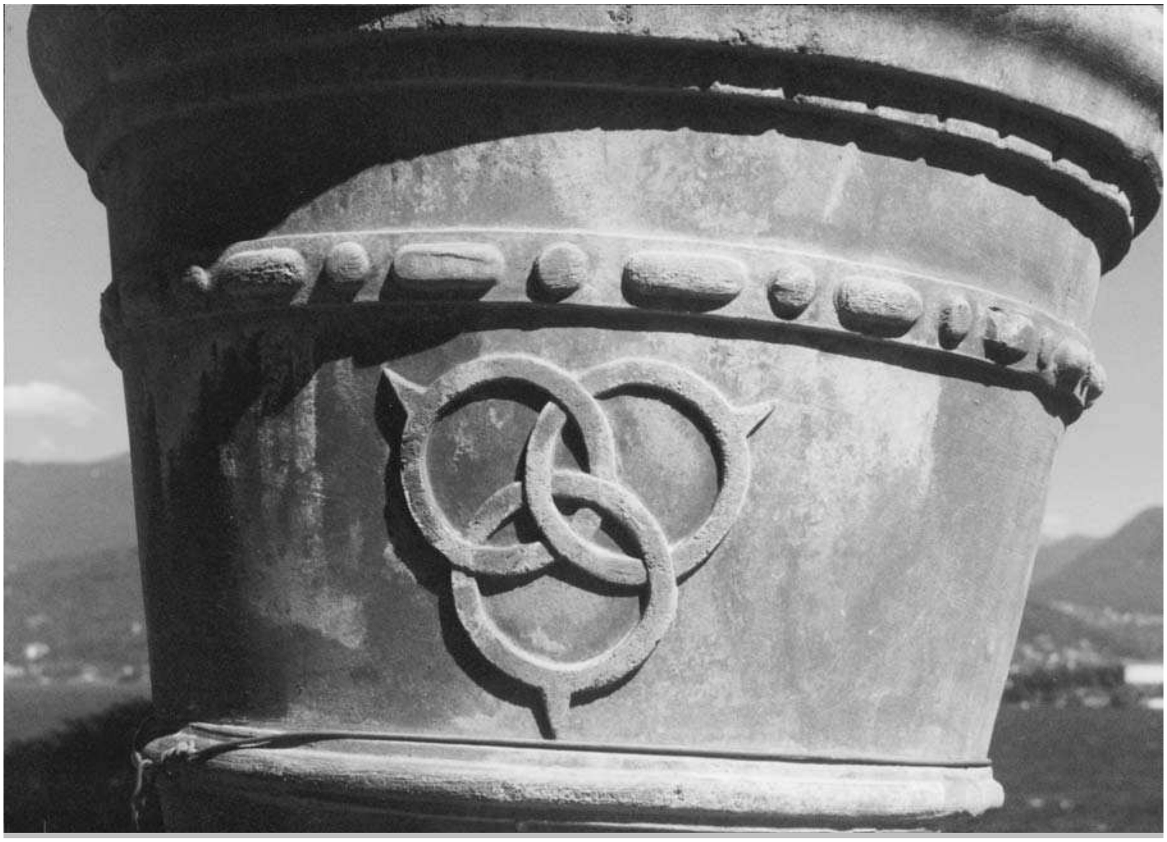

Borromean rings A flowerpot on Isola Bella, an island on Lake Maggiore, near Arona in northern Italy, that bears the Borromeo family crest. *Peter Cromwell*

interlaced snakes in place of the crescent moons, occurs at various sites in Wales, including Bangor Cathedral. The Borromean rings are commonly used to symbolize the Christianity Trinity. An early source for this was a thirteenth-century French manuscript, now lost, in which the word *unitas* appears in the center, inside all the circles, and the three syllables of "tri-ni-tas" are distributed in the outer sectors. Borromean rings can also be found on Japanese family emblems, at a Japanese Shinto shrine north of Sakurai in the province of Nara, and in the sculptures of the Australian artist John Robinson. In North America, the design is known as the Ballantine rings after the New Jersey brewing company P. Ballantine and Sons who use it as a trademark with the rings labeled Purity, Body, and Flavor.

The rings first appeared in mathematics in the earliest work on knots by Peter **Tait** in 1876. The pattern of circles can be interlaced by replacing each of the six vertices by a crossing that shows how the circles pass over and under one another. Since there are two choices for each crossing, there are $2^6 = 64$ possible interlaced patterns. However, after taking symmetry into account, these 64 reduce to only 10 geometrically distinct patterns. Two patterns are considered to be the same if one can be obtained from the other by applying one or more of the following operations: rotation by 120°, reflection, and reflection in the plane of the pattern. The last symmetry operation means that the sense of all the crossings is switched. The rings can also be analyzed from the viewpoint of **topology**, which means that the designs are thought of as links made from a flexible and elastic material. If two links can be manipulated and deformed to look like one another (without breaking and joining the rings) then they are topologically equivalent. The 10 geometrically distinct patterns boil down to only five distinct topological types.

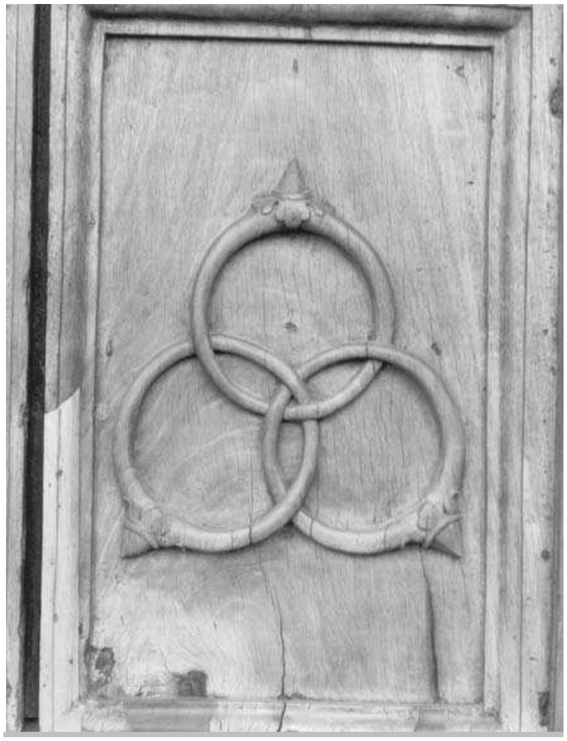

Borromean rings Three alternating rings carved in a panel in a walnut door of the church of San Sigismondo in Cremona, Italy. The emblem is one of several belonging to the Sforza family. *Peter Cromwell*

Borsuk-Ulam theorem

One of the most important and profound statements in **topology**: if there are *n* regions in *n*-dimensional space, then there is some hyperplane that cuts each region exactly in half, measured by volume. All kinds of interesting results follow from this. For example, at any given moment on Earth's surface, there must exist two antipodal points–points on exactly opposite sides of Earth–with the same temperature and barometric pressure! One way to see that this must be true is to think of two opposite points *A* and *B* on the equator. Suppose *A* starts out warmer than *B*. Now move *A* and *B* together around the equator until *A* moves into *B*'s original position, and simultaneously *B* into *A*'s original position. *A* is now cooler than *B*, so somewhere in between they must have been the same temperature. The Borsuk-Ulam theorem implies the **Brouwer fixed-point theorem** and also the **ham sandwich theorem**.

bottle sizes

Wine and champagne come in various standard bottle sizes, as shown in the table "Wine Bottle Sizes." These follow a **geometric sequence**, doubling in size with each step, up to the double-magnum, but thereafter increase in a more complicated way. There are also regional variations (for example, a Nebuchadnezzar may hold from 12 to 15 liters) and differences depending on the type of drink held.

Wine Bottle Sizes

Name of Size	Region	Capacity (liters)	Standard Bottle Equivalents
Baby/split	All	0.1875	0.25
Half-bottle	All	0.375	0.5
Bottle	All	0.75	1
Magnum	All	1.5	2
Double-magnum	All	3	4
Jeroboam	Burgundy, Champagne	5	6.67
Jeroboam	Bordeaux, Cabernet S.	4.5	6
Rehoboam	Burgundy, Champagne	4.5	6
Imperial	Bordeaux, Cabernet S.	6	8
Methuselah	Burgundy, Champagne	6	8
Salmanazar	Burgundy, Champagne	9	12
Balthazar	Burgundy, Champagne	12	16
Nebuchadnezzar	Burgundy, Champagne	15	20

boundary condition

The value of a **function** at the edge of the range of some of its variables. Recognizing the boundary conditions of an unknown function helps in its identification since other unknowns, such as variables in integrations, can then be eliminated.

boundary value problem

An **ordinary differential equation** or a **partial differential equation** given together with **boundary condition**s to ensure a unique solution.

Bourbaki, Nicholas

Not an individual but a collective mathematician. In the 1930s, the Bourbaki group, made up of some of the brightest mathematicians in France, began as a club, holding secret meetings in Strasbourg to update university lectures and texts in the wake of World War I, which had essentially wiped out a generation of young talent. In time, Bourbaki authored encyclopedic accounts of all areas of mathematics, and its influence became widespread.

Its origins can be traced to 1934 and to André Weil and Henri Cartan who were maîtres de conférences (the equivalent of assistant professors) at the University of Strasbourg. One of their duties was to teach differential and integral calculus but they found the standard text on this subject, *Traité d'Analyse* by E. Goursat, wanting. Following a suggestion by Weil to write a new "Treatise on Analysis," a group of about 10 mathematicians began to meet regularly to plan the new work. Quickly, it was decided that the work should be collective, without any acknowledgment of individual contributions, and this became a feature of Bourbaki's output. In the summer of 1935, the pen name Nicholas Bourbaki was chosen, and the initial membership consisted of Weil, Cartan, Claude Chevalley, Jean Delsarte, and Jean Dieudonné–all former students at the École Normale Supérieure in Paris. Over the years, the membership varied; some people in the first group dropped out quickly, others were added, and later there was a regular process of addition and retirement (mandatory by the age of 50).

Bourbaki adopted rules and procedures that to outsiders often seemed eccentric and even bizarre. For example, during meetings to review and revise drafts for the various books the group developed, anyone could express his opinion as loudly as he wanted at any time, so it was not uncommon for several distinguished mathematicians to be on their feet at the same time shouting monologues at the top of their voices. Somehow out of this mayhem emerged work of extreme precision, to the point of pedantry and dryness. Bourbaki would have nothing to do with geometry or any attempt at visualization, and believed that mathematics should distance itself from the sciences. However, despite its tendency to be boring and long-winded, Bourbaki did achieve its goal: to set down in writing what was no longer in doubt in modern mathematics.

brachistochrone problem

A problem with which Johann Bernoulli (see **Bernoulli Family**) challenged his contemporaries in *Acta Eruditorum* in June 1696:

> Following the example set by Pascal, Fermat, etc., I hope to gain the gratitude of the whole scientific community by placing before the finest mathematicians of our time a problem which will test their methods and the strength of their intellect. If someone communicates to me the solution of the proposed problem, I shall publicly declare him worthy of praise. . . . Given two points A and B in a vertical plane, what is the curve traced out by a point acted on only by gravity, which starts at A and reaches B in the shortest time?

Isaac **Newton** reportedly solved the problem between four in the evening and four in the morning after a hard day at the Royal Mint, later commenting: "I do not love to be dunned [pestered] and teased by foreigners about mathematical things. . . ." Other correct solutions came from Gottfried **Leibniz**, the Frenchman Guillaume **de L'Hôpital**, and Johann's brother Jakob. They, like Johann, realized that the solution to the brachistochrone problem, as it was also to the **tautochrone problem**, was a curve known as the **cycloid**.

Brahmagupta (A.D. 598–after 665)

A Hindu astronomer and mathematician who became the head of the observatory at Ujjain–the foremost mathematical center of ancient India. His main work, *Brahmasphutasiddhanta* (The opening of the universe), written in 628, contains some remarkably advanced ideas, including a good understanding of the mathematical role of **zero**, rules for manipulating both positive and **negative number**s, a method for computing square roots, methods of solving linear and some **quadratic** equations, and rules for summing series. His contributions to astronomy were equally ahead of their time. *Brahmagupta's theorem* states that in a cyclic quadrilateral (a four-sided shape whose corners lie on a circle) having perpendicular diagonals, the perpendicular to a side from the point of intersection of the diagonals always bisects the opposite side. **Brahmagupta's formula** for the area of a cyclic quadrilateral with sides of length a, b, c, and d is $\sqrt{(S-a)(S-b)(S-c)(S-d)}$, where $S = (a+b+c+d)/2$. As d goes to zero, this reduces to **Heron's formula**.

braid

A collection of lines or strings that are plaited together and whose ends are attached to two parallel straight lines.

Braid theory was pioneered by the Austrian mathematician Emil Artin (1898–1962) and is related to **knot** theory. It also has other applications: for instance, if we consider the way the roots of a **polynomial** move as one of the polynomial's coefficients changes, this motion can be thought of as a braid.

Brianchon's theorem

Given a **conic section**, if we circumscribe a hexagon about it, then the major diagonals of the hexagon are concurrent.

bridges of Königsberg

A famous routing problem that was analyzed and solved by Leonhard **Euler** in 1736, and that helped spur the development of **graph** theory. The old city of Königsberg, once the capital of East Prussia, is now called Kaliningrad. It falls within a tiny part of Russia known as the Western Russian Enclave, between Poland and Lithuania, which (to the surprise even of many modern Russians) is not connected with the rest of the country! Königsberg lay some four miles from the Baltic Sea on rising ground on both sides of the river Pregel (now the Pregolya), which flowed through the town in two branches before uniting below the Grune Brocke (Green Bridge). Seven bridges (numbered in the diagram) crossed the Pregel and connected various parts of the city (letters A to D), including Kneiphof Island (B), the site of Königsberg University and the grave of its most famous son, the great philosopher Emmanuel Kant (1724–1804).

A question arose among the town's curious citizens: Was it possible to make a journey across all seven bridges *without having to cross any bridge more than once?* No one had been able to do it, but was there a solution? Euler, who was in St. Petersburg, Russia, at the time, heard about this puzzle and looked into it. In 1736, he published a paper called *Solutio problematis ad geometriam situs pertinentis* (The solution of a problem relating to the geometry of position) in which he gave his answer. Euler reasoned that, for such a journey to be possible, each land mass would need to have an even number of bridges connected to it, or, if the journey began at one land mass and ended at another then those two land masses alone could have an odd number of connecting bridges while all the other land masses would have to have an even number of connecting bridges. Since the Königsberg bridges violated this layout, a grand tour that involved only one crossing per bridge was impossible. Euler's paper was important because it solved not just the Königsberg conundrum but the much more general case of any network of points, or *vertices,* that are connected by lines, or *arcs.* What is more, the words *geometry of position* in the title show that Euler realized that he was dealing with a different type of geometry where distance is irrelevant. So this work can be seen as a prelude to the subject of **topology**. See also **Euler path**.

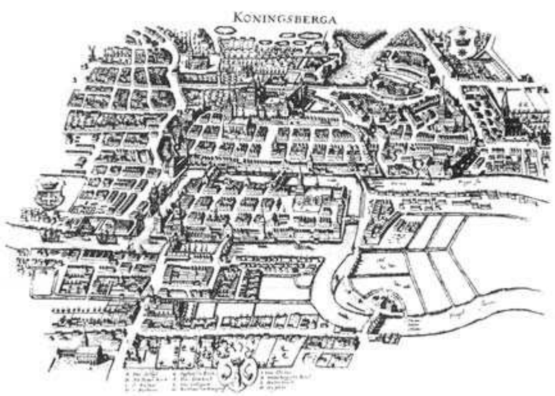

bridges of Königsberg An old map of the city showing the seven bridges.

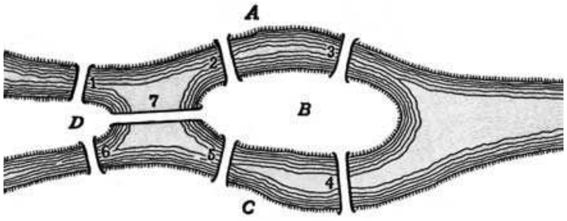

bridges of Königsberg The essential layout of the bridges. *B* represents Kneiphof Island.

Briggs, Henry (1561–1630)

An English mathematician who introduced common **logarithms** (to base 10) and was largely responsible for getting scientists to use them. Although a well-regarded mathematician in his own right, holding the Savilian chair of geometry at Oxford, Briggs was most important as a contact and as a public relations man for his field.

Brocard problem

The problem of finding the integer solutions of the equation

$$n! + 1 = m^2.$$

These solutions are called *Brown numbers,* and only three of them are known: (5, 4), (11, 5), and (71, 7). Paul **Erdös** conjectured that there are no others.

broken chessboard

See **polyomino**.

Bronowski, Jacob (1908–1974)

A Polish mathematician who worked first on operation theory and its application to military strategy, but later on the ethics of science. He is remembered for writing and narrating the television series *The Ascent of Man* in 1973.

Brouwer, Luitzen Egbertus Jan (1881–1966)

A Dutch mathematician who opposed the logicist school of Bertrand **Russell** and established the intuitionist school of mathematical thought. He was also one of the founders of **topology**, doing most of his work in this field between 1909 and 1913.

Brouwer fixed-point theorem

An amazing result in **topology** and one of the most useful theorems in mathematics. Suppose there are two sheets of paper, one lying directly on top of the other. Take the top sheet, crumple it up, and put it back on top of the other sheet. Brouwer's theorem says that there must be at least one point on the top sheet that is in exactly the same position relative to the bottom sheet as it was originally. The same idea works in three dimensions. Take a cup of coffee and stir it as much as you like. Brouwer's theorem insists that there must be some point in the coffee that is in exactly the same spot as it was before you started stirring (though it might have moved around in between). Moreover, if you stir again to move that point out of its original position, you can't help but move another point back into its original position! Not surprisingly, the formal definition of Brouwer's theorem makes no mention of sheets of paper or cups of coffee. It states that a continuous function from an *n*-**ball** into an *n*-ball (that is, any way of mapping points in one object that is topologically the same as the filling of an *n*-dimensional sphere to another such object) must have a fixed point. Continuity of the function is essential: for example, if you rip the paper in the previous example then there may not be a fixed point.

Brownian motion

The most common type of continuous **random** motion of a particle, one in which the particle's vibrations have more energy at short length and time scales. It models the motion of a particle in a fluid, fluctuation of stock prices, and many other processes. Brownian motion is named after the Scottish botanist Robert Brown (1773–1858) who first studied it.

Brun's constant

See **twin primes**.

bubbles

Whether alone or in groups joined together, bubbles get their shape by following one simple rule: soap film always tries to form a **minimal surface**. The mathematical study of bubbles and films began in earnest in the 1830s with the experiments of Joseph **Plateau**. A single bubble will always try to form a sphere because this shape, as proposed by **Archimedes** and proved by Hermann Amandus Schwarz (1843–1921) in 1884, is the minimal surface enclosing a single volume. For a long time, mathematicians believed that the minimal surface for enclosing two separate volumes of air is a double bubble separated by a third surface, which meets the other two along a circle at 120° and is flat if the bubbles enclose the same volume and, otherwise, is a spherical surface that bulges a little into the larger of the two. The *double bubble conjecture* was finally confirmed by a team of four mathematicians in 2000.

Another great bubble mystery that has recently been solved is why the bubbles in a glass of Guinness appear to sink rather than rise. Bubbles that rise, like those in a saucepan or those breathed out by a diver, are a familiar sight and easy to explain: gas bubbles are lighter than liquid and experience a buoyancy force that drives them up toward the liquid surface. But many of the bubbles in a glass of Guinness can be seen heading downward. Researchers have found that large bubbles in the center of the glass move upward relatively quickly and drag liquid with them. Since the amount of liquid in the glass stays the same (unless someone drinks it!) the liquid moving upward near the center must eventually move back down near the walls of the glass. This downward-

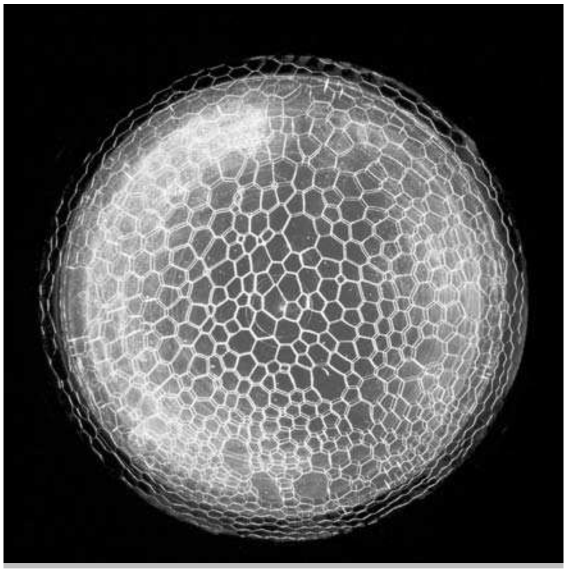

bubbles A honeycomb-like arrangement of tightly packed bubbles.
Australian National University

moving liquid has a dragging effect on the bubbles. Larger bubbles are more buoyant than smaller bubbles, and continue to move upward. Smaller bubbles (less than 0.05 mm in diameter) aren't buoyant enough to resist this drag force, and move downward with the liquid near the sides of the glass. Since Guinness is quite opaque, and these downward-moving small bubbles are close to the side of the glass, it often looks as if almost all the bubbles are moving down. See also **Plateau problem**.

buckyball

Also known as a *fullerene,* a large molecule made of carbon atoms arranged in the form of a convex polyhedral cage. Buckyballs are named after the architect Richard Buckminster Fuller because they look like the geodesic domes that he invented. The first buckyball to be discovered (by accident) was C_{60}, in which 60 carbon atoms are arranged at each of the vertices of a truncated **icosahedron**. This shape, which looks like a soccer ball, has 32 faces, of which 20 are regular hexagons and 12 are regular pentagons. Many

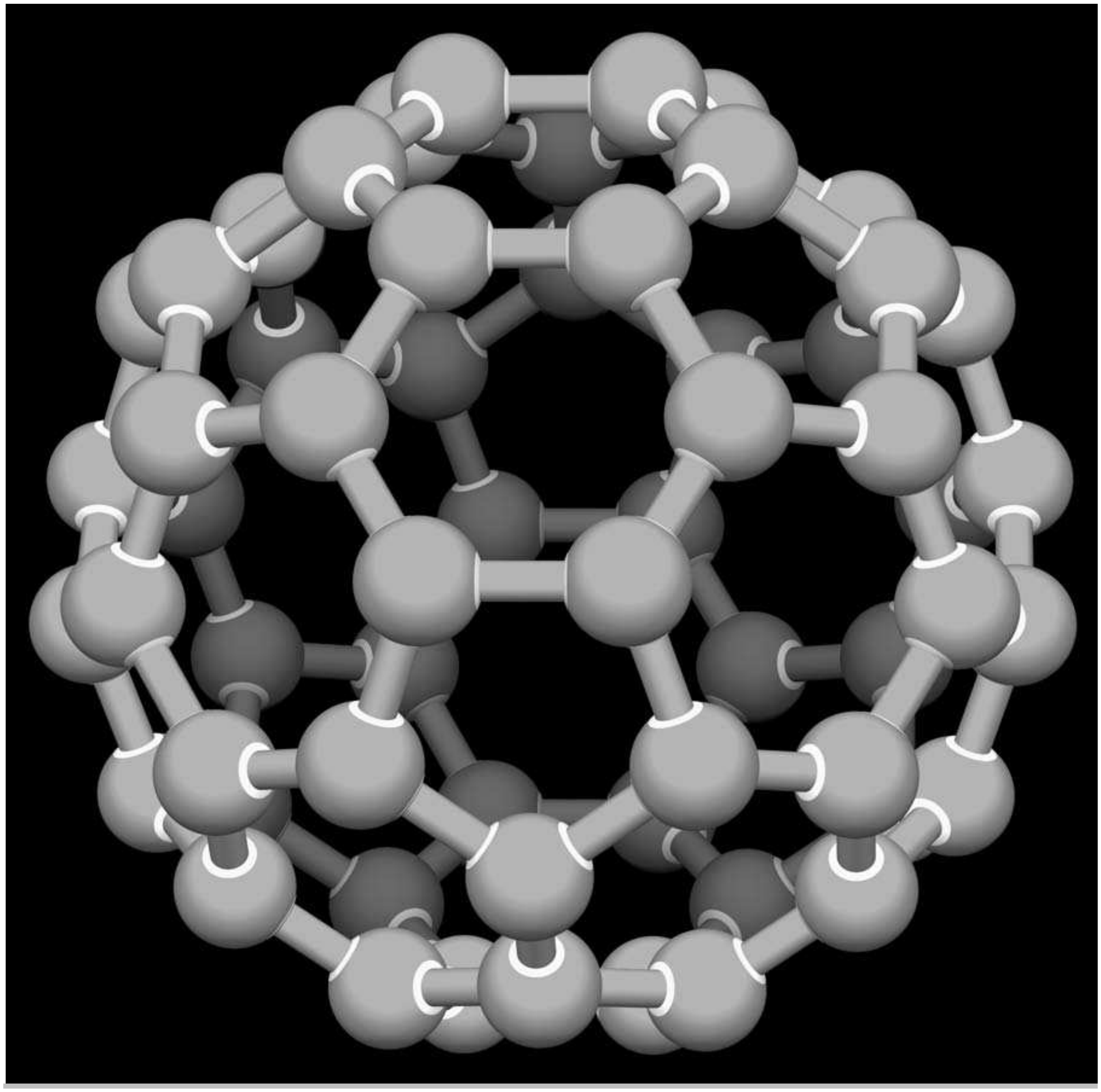

buckyball A molecule of Buckminster fullerene—a buckyball. *Nick Wilson* different types of buckyballs are known. Common ones have 70, 76, and 84 carbon atoms, but all are built up exclusively from hexagonal and pentagonal faces in arrangements that follow **Euler's formula**. This formula ensures that while the number of hexagons can vary from one type of fullerene to another, every fullerene has exactly 12 pentagons. (In fact, buckyballs with heptagonal faces have been seen, but these faces are concave and are regarded as defects.)

Buffon's needle

An early problem in geometrical probability (see **probability theory**) that was investigated experimentally in 1777 by the French naturalist and mathematician Comte Georges Louis de Buffon (1707–1788). It involves dropping a needle repeatedly onto a lined sheet of paper and finding the probability of the needle crossing one of the lines on the page. The result, surprisingly, is directly related to the value of **pi**.

Consider a simple case in which the lines are 1 cm apart and the needle is 1 cm in length. After many drops the probability of the needle lying across a line is found to be very close to $2/\pi$. Why? There are two variables: the distance from the center of the needle to the closest line, d, which can vary between 0 and 0.5 cm, and the angle, θ, at which the needle falls with respect to the lines, which can vary between 0 and 180°. The needle will hit a line if $d \leq \frac{1}{2} \sin\theta$. In a plot of d against $\frac{1}{2} \sin\theta$, the values on or below the curve represent a hit; thus, the probability of a success is the ratio of the area below the curve to the area of entire rectangle. The area below the curve is given by the integral of $\frac{1}{2} \sin\theta$ from 0 to π, which is 1. The area of the rectangle is $\pi/2$. So, the probability of a hit is $1/(\pi/2)$ or $2/\pi$ (about 0.637). Dropping a needle many times onto lined paper gives an interesting (but slow) way to find π. This kind of probabilistic means of performing calculations is the basis of a technique known as the **Monte Carlo method**.

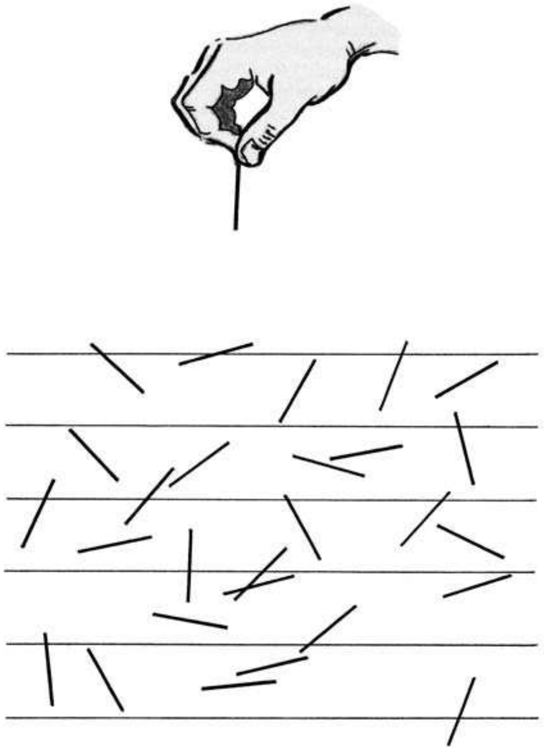

Buffon's needle Needles are dropped randomly onto a lined surface.

bundle

A map between two **topological space**s A and B, where the sets f^1 (b) for elements b of B (known as fibers), are all **homeomorphic** to a single space. The simplest example is the **Möbius band**, for which A is the Möbius band, B is a circle, and the fibers are homeomorphic to an interval on the real number line.

Burali-Forti paradox

An argument that shows that the collection of **ordinal number**s (numbers that give the position of objects) do not, unlike the **natural number**s, form a **set**. Each ordinal number can be defined as the set of all its predecessors. Thus:

0 is defined as {}, the **empty set**
1 is defined as {0} which can be written as {{}}
2 is defined as {0, 1} which can be written as {{}, {{}}}
3 is defined as {0, 1, 2} which can be written as {{}, {{}}, {{}, {{}}}} . . .
in general, n is defined as $\{0, 1, 2, \ldots n - 1\}$

If the ordinal numbers formed a set, this set would then be an ordinal number greater than any number in the set. This contradicts the assertion that the set contains all ordinal numbers. Although the ordinal numbers don't form a set, they can be regarded as a collection called a class.

Buridan's ass

A paradox of medieval logic concerning the behavior of an ass that is placed equidistantly from two piles of food of equal size and quality. Assuming that the behavior of the ass is entirely rational, it has no reason to prefer one pile to the other. Thus lacking a basis to decide which pile to eat first, it remains in its original position and starves. With one pile it would have lived; having two identical piles it dies. How can this make sense? The paradox is named after the French philosopher Jean Buridan (c. 1295–1356).

burr puzzle

An interlocking wooden puzzle that, when put together, typically looks like three rectangular blocks crossing one another at right-angles. Little is known about its early history, but it was certainly produced both in Asia and Europe in the eighteenth century. It acquired the name *Chinese puzzle,* probably because so many were produced in the Orient from the early 1900s. In 1928, Edwin Wyatt published *Puzzles in Wood,*[355] the first book devoted to the subject, and introduced the term *burr puzzle* because of the likeness of the assembled toy to a seed burr.

Burr puzzles consist of three (the smallest number), six (the most common number), twelve, or other numbers of pieces that are notched in various ways so as to pose a challenge to the would-be assembler. The earliest known reference to the popular six-piece burr appears in a Berlin catalog of 1790, but not until 1917 was a patent taken out on a particular design. In 1977, William **Cutler** proved that 25 possible notchable pieces can be used to make solid six-piece burrs and that they can be put together in 314 ways. (Pieces are considered notchable if they can be made by a sequence of notches that are produced by chiseling out the space between two saw cuts.) Cutler also proved there are 369 general pieces from which solid burrs can be made and that these can be assembled in 119,979 ways. One particular form of burr has six identical pieces, all of which move outward or inward together. Another form, with flat notched pieces, has one piece with an extra notch or an extended notch that allows it to fit in last, either by sliding or twisting, although this isn't initially obvious. This form is sometimes made with equal pieces so that it can only be assembled by force, perhaps after steaming.

butterfly effect

One of the more sensational and loudly touted claims of **chaos** theory: a butterfly beating its wings could, by an intricate chain of causes and effects, give rise to a hurricane. The gist of the argument is that minuscule disturbances can be amplified unpredictably into major phenomena. However, the overwhelming likelihood is that any effect as small as the beating of a butterfly's wing would be quickly dampened out and play no significant part in future events. See also **causality**.

butterfly theorem

Let M be the midpoint of a **chord** PQ of a circle, through which two other chords AB and CD are drawn. If AD intersects PQ at X and CB intersects PQ at Y, then M is also the midpoint of XY. The theorem gets its name from the shape of the resulting figure.

Byron, (Augusta) Ada (1815–1852)

Also known as Lady Lovelace, the daughter of Lord Byron and one of the most picturesque characters in computer prehistory. Her parents separated five weeks after her birth and she was raised by her mother (neé Annabella Milbanke), whom Lord Byron had called his "Princess of Parallelograms" because of her interest in mathematics. She was determined that Ada would become a mathematician and scientist, not a poet like her father. Ada, however, managed to combine both worlds, blending her science with poetical vision and her mathematics with metaphor.

At the age of 17 she was introduced to Mary Somerville, a remarkable woman who translated **Laplace**'s works into English, and whose texts were used at Cambridge (where a

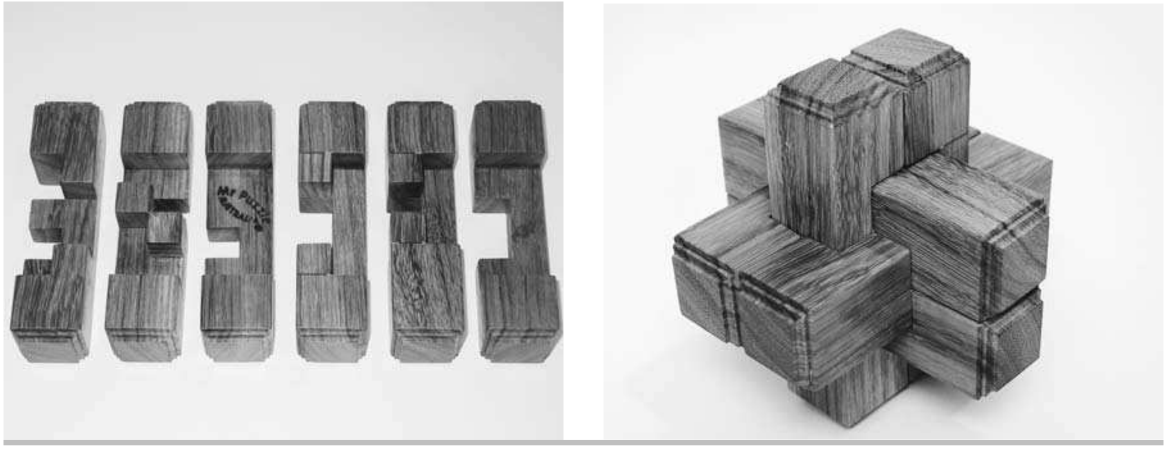

burr puzzle A difficult six-piece burr with 19 false assemblies and a final piece that requires 10 directional moves. *Mr. Puzzle Australia/William Cutler* women's college is now named after her). It was at one of Mary Somerville's dinner parties, in November 1834, that Ada first heard of Charles **Babbage**'s ideas for a new calculating machine, the Analytical Engine, and was immediately intrigued. In 1843, Ada, married and the mother of three children, translated a French article about the Engine and showed it to Babbage. He suggested that she add her own notes, which turned out to be three times the length of the original piece and included prescient comments about how such a machine might be used to compose complex music, produce graphics, and solve scientific problems. A regular correspondence ensued between Ada and Babbage, during which Ada suggested to Babbage a plan for how the engine might calculate **Bernoulli numbers**–a plan now regarded as the first computer program. In recognition of this, a software language developed by the U.S. Department of Defense was named "Ada" in 1979. Like her father, she died at the age of 36, following a series of illnesses.[335]

byte

A string of 8 **bit**s, used to represent a character.

C

caduceus

In mathematics, a pair of curves in space, each of which is a **helix** and which twist in opposite directions around one another. In mythology, the caduceus is the wing-topped staff, wound about by two snakes, carried by Hermes, the Greek messenger of the gods. The snakes became entwined after Hermes threw his staff at them to stop their fighting. A caduceus was carried by Greek officials and became a Roman symbol for truce and neutrality. Since the sixteenth century it has also served as a symbol of medicine. Before modern medicine, people infected by parasitic worms were treated by physicians using a stick and a knife. A slit would be cut in the patient's skin in front of the worm, and as the parasite crawled out of the incision, the worm would be wound around a stick until it was totally removed. The medical treatment of parasitic worm infection by knife and stick is believed to be the inspiration for the original caduceus. It was used as a promotional sign for physicians of that period.

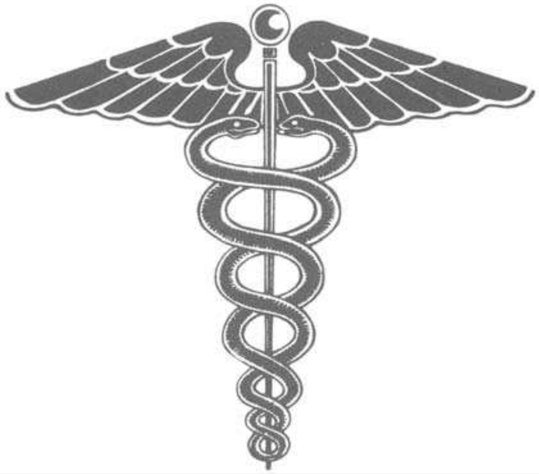

caduceus *Auckland Medical Research Foundation*

Cage, John (1912–1992)

The American avant-garde composer perhaps best known for the quietest piece of music ever written. His piano composition *4′33″* calls for the player to sit in silence for 273 seconds–this being the number of degrees below zero on the centigrade scale of **absolute zero** at which molecular motion stops. *4′33″* was inspired by Cage's visit to Harvard University's anechoic chamber about which he wrote:

> There is no such thing as empty space or empty time. There is always something to hear or something to see. In fact, try as we might to make a silence, we cannot. For certain engineering purposes, it is desirable to have as silent a situation as possible. Such a room is called an anechoic chamber, its walls made of special materials, a room without echoes. I entered one at Harvard University . . . and heard two sounds, one a high and one a low. When I described them to the engineer in charge, he informed me that the high one was my nervous system and the low one was my blood circulation.

Cage's *4′33″* breaks traditional boundaries by shifting attention from the stage to the audience and even beyond the concert hall. The listener becomes aware of all sorts of sound, from the mundane to the profound, from the expected to the surprising, from the intimate to the cosmic: shifting in seats, riffling programs, breathing, a creaking door, passing traffic, a recaptured memory. Is sitting quietly alone for 273 seconds equivalent to a private performance (and audience) of the piece? Or, in the final analysis, is it all pretentious nonsense? In his essay on "Nothing" Martin Gardner wrote: "I have not heard *4′33″* performed, but friends who have tell me it is Cage's finest composition."

Caesar cipher

The simplest and oldest known type of **substitution cipher**, attributed to Julius Caesar, who used it to send government messages. In it, each letter in the alphabet is replaced by another letter using a predefined rule that shifts the alphabet a uniform amount to the right or left. For example, a shift of three units to the right, would turn the "This is secret" into "Wklv lv vhfuhw."

cake-cutting

How can a group of people cut up a cake so that each gets what they consider to be a fair share? In its modern mathematical form, this classic problem of *fair division* dates from World War II, when Hugo **Steinhaus** tackled it using a **game theory** approach.[315] Any number of "players" are allowed. They agree on rules for dividing the cake, and then everyone follows those rules. In the end, each player must get what he or she perceives to be a fair share. In the simplest case, involving just two people, there is an easy, well-known strategy: one cuts, the other chooses. Can this method be extended to three people? Can it be extended so that each person, according to his own judgment, receives the *biggest* piece? Steinhaus was able to prove that a so-called *envy-free division,* where everyone believes they got the best deal exists in every case, for any number of players. However, it was left for others to find actual algorithms that worked for three or more players.

A three-person envy-free method was first devised by John Selfridge and John **Conway**. Suppose the players are called Alice, Bob, and Carol. The method goes like this: (1) Alice cuts the cake into what she thinks are thirds. (2) Bob trims one piece to create a two-way tie for largest, and sets the trimmings aside. (3) Carol picks a piece, then Bob, then Alice. Bob has to take a trimmed piece if Carol does not. Call the person who took the trimmed piece T, and the other (of Bob and Carol) NT. (4) To deal with the trimmings, NT cuts them into what she thinks are thirds. (5) The players pick pieces in this order: T, Alice, NT. The key to the success of the Selfridge-Conway strategy is that for the trimmings, Alice has an "irrevocable advantage" with respect to T, since Alice will never envy T even if T gets all the trimmings. Thus Alice can pick after T and allow the method to end in a finite number of steps.

For four or more cake-cutters, envy-free solutions are very complex and can take arbitrarily long to resolve. However, a general solution of the problem of fair, envy-free division was eventually found in 1992 by the Americans Steven Brams, a political scientist at New York University, and Alan Taylor, a mathematician at Union College in Schenectady, New York.[51, 52] With two players, the first player cuts the cake in half. With three players, the first player cuts the cake into thirds. With four players, Brams and Taylor showed, the first player, say Bob, cuts the cake into five equal-looking pieces. He passes them to Carol, who trims two at most to create a three-way tie for largest in her eyes. She sets the trimmings aside and gives the five pieces to Don, who trims one at most to create a two-way tie for largest in his eyes. Alice, the fourth player, now selects the piece she likes best. Choosing proceeds in the reverse order from cutting, with the proviso that anyone who trimmed one or more pieces must take one of them if any are still available when it's his or her turn to choose. The extra piece to begin with assures that no player gets second-best. If someone takes a piece she likes before it's her turn to choose, an equivalent piece or better always remains on the table. According to a formula Brams and Taylor developed, Bob must cut the cake into at least $2^{(n-2)+1}$ pieces at the start. This amounts to nine pieces for five players, 17 pieces for six, and so on. Bob has to cut all these extra pieces to make sure that, when he finally gets to choose at the end, there will be a piece left that hasn't been either trimmed or chosen by one of the many other players. With 22 players, Bob has to divide the cake into over a million pieces–small crumbs of comfort in the quest for a fairer world.[268]

Calabi-Yau space

A type of mathematical space that enters into **string theory**, where the geometry of the universe is held to consist of at least 10 dimensions–the four familiar dimensions of **space-time** and six compact dimensions of Calabi-Yau space. These extra dimensions are so tightly curled up that they aren't noticed. Although the main application of Calabi-Yau spaces is in theoretical physics, they are also interesting from a purely mathematical standpoint.

calculus

> *The calculus is the greatest aid we have to the application of physical truth in the broadest sense of the word.*
>
> –William Fogg Osgood (1864–1943)

The branch of mathematics that deals with (1) the rate of change of quantities (which can be interpreted as the slopes of curves), known as *differential calculus,* and (2) the length, area, and volume of objects, known as *integral calculus.* Calculus was one of the most important developments in mathematics and also in physics, much of which involves studying how quickly one quantity changes with respect to another. It is no coincidence that one of the founders of calculus was the brilliant English physicist Isaac **Newton**; another was Gottfried **Leibniz**. Although students nowadays learn differential calculus first, integral calculus has older roots.

calculus of variations

Calculus problems, especially **differentiation** and *maximization,* that involve **function**s on a set of functions of a real variable. An example is to find the shape of a cable suspended from both ends.

calendar curiosities

The earliest event in human history for which a definite date is known is a battle between the Lydians (allies of the Greek Spartans) and the Medes (ruled by the Persian king Cyrus) who had been locked in a war for five years. As the two sides faced each other for a crucial daytime confrontation, a solar eclipse occurred. This was taken as a sign of the gods' disapproval and the Lydians and Medes agreed to end the fighting then and there. The dates of solar eclipses can be figured out with great accuracy, and this one is known to have taken place on May 28, 586 B.C.

Much less certain is the birth date of Christ. It was not until A.D. 440 that Christmas was celebrated on December 25. This date was chosen because it coincided with the birth date of Mithras, the Persian sun-god, and was close to the pagan festival of Yule. In A.D. 534, Dionysius Exiguus (also known as Dennis the Little) created the system, still used today, of counting the years from the birth of Christ. Unfortunately, he slipped up in his calculations. No one knows exactly when Jesus was born, but it was probably around 6 B.C. and certainly before the death of Herod the Great in 4 B.C.

As for the future, there's no shortage of predictions about the end of the world. According to the Mayan "long count" linear calendar, it will happen on June 5, 2012. Other calendric curiosities: February 1865 is the only month in recorded history not to have a full moon, and months that begin on a Sunday will always have a Friday the 13th.

Caliban

A pseudonym of Hubert **Phillips**.

Caliban puzzle

A logic puzzle in which one is asked to infer one or more facts from a set of given facts.

cannonball problem

The mathematical analysis of stacks of cannonballs (or of spheres in general) that has its roots in a question posed by Sir Walter Raleigh, explorer, introducer of the potato and tobacco to Britain, and part-time pirate on the high seas. Raleigh asked his mathematical assistant, Thomas Harriot, how he could quickly figure out the number of cannonballs in a square pyramidal stack without having to count them individually. Harriot solved this problem without difficulty. If k is the number of cannonballs along the side of the bottom layer, the number of cannonballs in the pyramid n is equal to $\frac{1}{6}k(1+k)(1+2k)$. For example, if $k = 7$, $n = 420$. A more specific form of the cannonball problem asks what is the smallest number of balls that can first be laid out on the ground as an $n \times n$

cannonball problem Cannonballs stacked in Narbonne, France. *Australia National University* square, then piled into a square pyramid k balls high? In other words, what is the smallest square number that is also a square pyramidal number? This answer is the smallest solution to the **Diophantine equation**

$$\tfrac{1}{6}\, k(1 + k)(1 + 2k) = n^2$$

and turns out to be $k = 24$, $n = 70$, corresponding to 4,900 cannonballs. The ultimate form of the cannonball problem is to ask if there are any other, larger solutions. In 1875 Edouard **Lucas** conjectured that there weren't, and in 1918 G. N. Watson proved that Lucas was right.[345]

Returning to Elizabethan times, Thomas Harriot's interest in spheres extended far beyond piles of cannonballs. Harriot was an atomist, in the classical Greek sense, and believed that understanding how spheres pack together was crucial to understanding how the basic constituents of nature are arranged. Harriot also carried out numerous experiments in optics and was far ahead of his time in this field. So when, in 1909, Johannes **Kepler** wanted some advice on how to give his own theories on optics a stronger scientific underpinning who better to turn to than the Englishman? Harriot supplied Kepler with important data on the behavior of light rays passing through glass, but he also stimulated the German's interest in the sphere-packing problem. In response, Kepler published a little booklet titled *The Six-Cornered Snowflake* (1611) that would influence the science of crystallography for the next two centuries and that contained what has come to be known as **Kepler's conjecture** about the most efficient way to pack spheres.

canonical form

A form of any given **polyhedron** distorted so that every edge is **tangent** to the unit sphere and the center of gravity of the tangent points is the origin.

Cantor, Georg Ferdinand Ludwig Philipp (1845–1918)

A Russian-born German mathematician who founded **set theory** and introduced the concept of **transfinite numbers**. His shocking and counterintuitive ideas about **infinity** drew widespread criticism before being accepted as a cornerstone of modern mathematical theory.

Cantor was 11 when his family moved from St. Petersburg to Germany. Despite attempts to push him into the more lucrative field of engineering, he eventually won his father's approval to study math at the Polytechnic of Zurich. The following year, 1863, his father died and Cantor switched to the University of Berlin where he studied under some of the greats of the day, including Karl **Weierstrass** and Leopold **Kronecker**. After receiving his doctorate in 1867, he had trouble finding a good job and was forced to accept a position as an unpaid lecturer and later as an assistant professor at the backwater University of Halle. In 1872, he achieved his first breakthrough—and a promotion—by proving that if a **function** is continuous (in other words, its graph is smooth) throughout an interval, it can be represented by a unique trigonometric series. This work, suggested to him by his colleague Heinrich Heine, was crucial because it led Cantor to think about the relations between points, represented by **real numbers**, that make up an unbroken line—the so-called *continuum.* Cantor realized that **irrational numbers** can be represented as infinite sequences of **rational numbers**, so that they can be understood as geometric points on the real-number line, just as rational numbers can. He was now in uncharted territory and at odds with mathematical orthodoxy, which frowned on the idea of actual **infinity**; however, he found like-minded friends in Richard **Dedekind** and, later, Gösta **Mittag-Leffler**.

In 1873 to 1874 Cantor proved that the rational numbers could be paired off, one by one, with the natural numbers and were therefore countable, but that there was no such one-to-one correspondence with the real numbers. He then went on to show, incredibly, that there are exactly the same number of points on a short line as there are on an indefinitely long line, or on a plane, or in any mathematical space of higher dimensions. On this, he wrote to Dedekind: "I see it, but I don't believe it!"

By 1883, Cantor had abandoned his earlier reticence about dealing with irrationals only as sequences of rationals and started to think in terms of a new type of number—the **transfinite numbers**. The sets of natural numbers and of real numbers were, he reasoned, just two elements of a series of different kinds of infinity. This dramatic extension of the number system to allow for legitimate mathematics of the infinite was violently opposed. Henri **Poincaré** said that Cantor's theory of infinite sets would be regarded by future generations as "a disease from which one has recovered." Kronecker went further and did all he could to ridicule Cantor's ideas, suppress publication of his results, and block Cantor's ambition of gaining a position at the prestigious University of Berlin. In the spring of 1884 Cantor suffered the first of several attacks of depression, exacerbated if not induced by the negative reaction of his contemporaries. In between these attacks, he published further results but was increasingly troubled by his failure to prove the **continuum hypothesis**—his belief that the order of infinity of the real numbers came next after that of the natural numbers. Although his later years were spent in and out of a sanatoria, he lived long enough to see his ideas on set theory vindicated and be described by David **Hilbert** as "the finest product of mathematical genius and one of the supreme achievements of purely intellectual human activity."

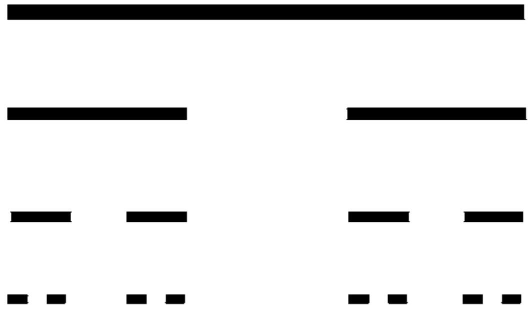

Cantor dust.

Cantor dust

Also known as the *Cantor set,* possibly the first pure **fractal** ever found. It was detected by Georg **Cantor** around 1872. To produce Cantor dust, start with a line segment, divide it in three equal smaller segments, take out the middle one, and repeat this process indefinitely. Although Cantor dust is riddled with infinitely many gaps, it still contains uncountably many points. It has a **fractal dimension** of log 2/log 3, or approximately 0.631. See also **Sierpinski carpet**.

cap

The symbol ∩ used to denote the **union** between two **set**s.

Cardano, Girolamo (1501–1570)

A celebrated Renaissance mathematician, physician, astrologer, and gambler, whose writings on the use of **negative number**s are the earliest known in Europe. As a physician, he gave one of the first clinical descriptions of typhoid fever. The illegitimate child of a mathematically gifted lawyer who was a friend of Leonardo da Vinci, he entered the University of Pavia in 1520 and later studied medicine at Padua. His eccentric and confrontational style earned him few friends, and he had trouble finding work. Eventually, he won a reputation as a physician and his services were highly valued at the courts.

Today, Cardano is remembered mostly for his achievements in algebra. He published the solutions to the **quartic** and **cubic equations** in his book *Ars magna* (1545). The solution to the cubic was communicated to him by Niccoló **Tartaglia** (who later claimed that Cardano had sworn not to reveal it, and became embroiled with Cardano in a decade-long fight), and the quartic was solved by Cardano's student Lodovico Ferrari. Both were acknowledged in the foreword of the book. Cardano was notoriously short of money and kept himself afloat by being an accomplished gambler and chess player. A book by him about games of chance, *Liber de Ludo Aleae* (Book on games of chance), written in the 1560s but published posthumously in 1663, contains the first systematic treatment of **probability theory**, as well as a section on effective cheating methods. Cardano invented several mechanical devices including the combination lock, the *Cardano suspension* (consisting of three concentric circles that allow a supported compass to rotate freely), and the *Cardan shaft,* which allows the transmission of rotary motion at various angles and is used in vehicles to this day. He made several contributions to hydrodynamics and claimed that perpetual motion is impossible, except in celestial bodies. He published two encyclopedias of natural science that contain a wide variety of inventions, facts, and occult superstitions.

Cardano led a beleaguered life. His elder and favorite son was executed in 1560 after he confessed to having poisoned his mercenary, cuckolding wife. Cardano's daughter was allegedly a prostitute who died from syphilis, prompting him to write a treatise about the disease. His younger son was a gambler who stole money from him. And Cardano himself was accused of heresy in 1570 because he computed the horoscope of Jesus Christ. Apparently, his own son contributed to the prosecution. Cardano was arrested and had to spend several months in prison, then was forced to abjure and had to give up his professorship. He moved to Rome, received a lifetime annuity from Pope Gregory XIII, and finished his not-uneventful autobiography. He died on the day he had (supposedly) astrologically predicted. See also **Chinese rings**.

Cardan's rings

See **Chinese rings**.

cardinal number

A number, often called simply a *cardinal,* that is used to count the objects or ideas in a **set** or collection: 0, 1, 2, . . . , 83, and so on. The *cardinality* of a set is just the number of elements the set contains. For finite sets this is always a **natural number**. To compare the sizes of two sets, X and Y, all that's necessary is to pair off the elements of X with those of Y and see if there are any left over. This concept is obvious in the case of finite sets but leads to some strange conclusions when dealing with infinite sets (see **infinity**). For example, it is possible to pair off all the natural numbers with all the even numbers, with none left over; thus the set of natural numbers and the set of even numbers have the same cardinality. In fact, an infinite set can be *defined* as any set that has a proper subset of the same cardinality. Every **countable set** that is infinite has a cardinality of **aleph**-null; the set of **real number**s has cardinality aleph-one. See also **ordinal number**.

cardioid

A heart-shaped curve first studied in 1674 by the Danish astronomer Ole Römer who was trying to find the

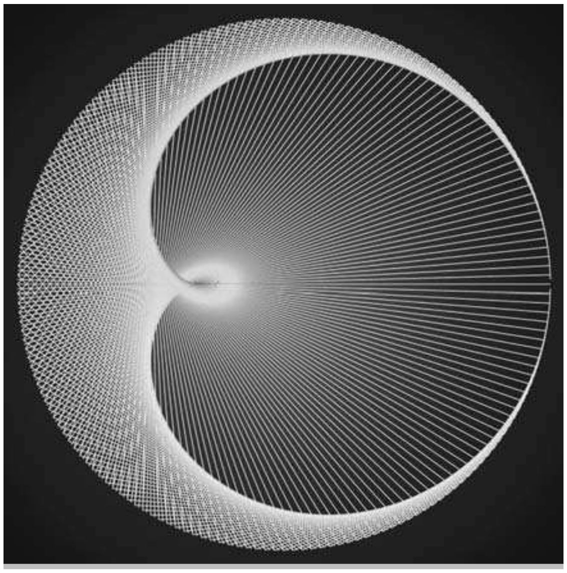

cardioid A cardioid curve spun by thread on a computer loom. *Jos Leys www.josleys.com*

best shape for gear teeth; the curve appears to have been named by Giovanni Salvemini de Castillon (1708–1791). When a circle rolls around another circle of the same size, any point on the moving circle traces out a cardioid. The Greeks used this fact when attempting to describe the motions of the planets. The cardioid is also the envelope of all circles with centers on a fixed circle, passing through one point on the fixed circle. In polar coordinates, it has the equation $r = 2a(1 - \cos\theta)$. It can also be described as an **epicycloid** with one cusp.

cards

The standard deck of 52 cards can be ordered in 52! (see **factorial**), or $8.065817517094 \times 10^{67}$ ways. There are various ways to **shuffle** cards in order to randomize them or to perform tricks with them. Each of the four kings in a deck represents a great leader from history: Charlemagne (hearts), Alexander the Great (clubs), Julius Caesar (diamonds), King David (spades). The king of hearts is the only one without a moustache. See also **blackjack**.

Carmichael number

Also known as an *absolute pseudoprime,* a number n that is a Fermat **pseudoprime** to any **base**, that is, it divides $(a^n - a)$ for any a. Another way of saying this is that a Carmichael number is actually a **composite number** even though **Fermat's little theorem** suggests it is probably a **prime number**. (Fermat's little theorem says that if P is a prime number, then for any number a, $(a^P - a)$ must be divisible by P. Carmichael numbers satisfy this condition to any base despite being composite.) There are only seven Carmichael numbers under 10,000 (they are 561, 1,105, 1,729, 2,465, 2,821, 6,601, and 8,911), and less than a quarter of a million of them under 10^{16}. Nevertheless, in 1994 it was proved that there are infinitely many of them. All Carmichael numbers are the product of at least three distinct primes, for example, $561 = 3 \times 11 \times 17$.

Carroll, Lewis (1832–1898)

The pen name of Charles Lutwidge Dodgson (obtained by anglicizing the Latin translation, "Carolus Lodovicus," of his first two names), an English mathematician, logician, and writer. Carroll was born at the Old Parsonage, Newton-by-Daresbury, Cheshire, his father being the vicar of All Saints Church, Daresbury. There is a commemorative window in the church, and a "Wonderland" weathervane, showing the Mad Hatter, the White Rabbit, and Alice, on the local primary school. Carroll was educated at Rugby School (the student mathematics society there is still called the Dodgson Society in his honor) and then at Christ Church, Oxford, at which college he was to spend the rest of his life, employed mainly as a lecturer.

Carroll's most famous book, *Alice's Adventures in Wonderland* (1865), grew out of a story he told on the

NEHSKIFD
Lewis Carroll's Chess
WORDGAME

Carroll, Lewis Lewis Carroll's Chess Wordgame, a game based on notation in Carroll's diaries with rules devised by Martin Gardner. *Kadon Enterprises, Inc., www.gamepuzzles.com* hot summer afternoon of July 4, 1862, when out rowing with the three young daughters of the Greek scholar H. G. Liddell, dean of Christ Church. Alice, named after Alice Liddell (later Hargreaves, 1852–1934), continued her adventures in *Through the Looking-Glass* (1871). Carroll wrote other books for children, including a long poem, "The Hunting of the Snark" (1876) and published several mathematical works, but was not distinguished academically. He stammered badly, never married, and seemed to find greatest pleasure in the company of little girls, with whom he lost his shyness. He was also an amateur pioneer in photography and an inventor of puzzles, games, ciphers, and mnemonics.[57–59] Carroll was a master of fantasy and his stories have their own logic. Carroll used puns and coined neologisms, including what he called "portmanteau words" like chortle (combining chuckle and snort). He played games with idioms, using such expressions as "beating time" (to music) in a literal sense. He reshaped such animals of fable or rhetoric as the Gryphon, the March Hare (said to have been inspired by a carved hare carrying a satchel located at St. Mary's Church, Beverly, Humberside, where Carroll visited), and the Cheshire Cat, and invented new ones, including the Bandersnatch and the Boojum.

PUZZLES

Here are a few examples of puzzles invented by Carroll:

1. You are given two glasses. One contains 50 tablespoons of milk, the other 50 tablespoons of water. Take one tablespoon of milk and mix it with the water. Now take one tablespoon of the water/milk mixture and mix it with the pure milk to obtain a milk/water mixture. Is there more water in the milk/water mixture or more milk in the water/milk mixture?
2. If you paint the faces of a cube with six different colors, how many ways are there to do this if each face is painted a different color and two colorings of the cube are considered equivalent if you can rotate one to get the other? What if we drop the restriction that the faces be painted different colors?
3. Make a word-ladder from FOUR to FIVE. (Every step in a word ladder differs from the previous step in exactly one letter and each step in the ladder is an English word.)
4. Why is a raven like a writing desk?

Solutions begin on page 369.

CARROLLIAN QUOTES

From *Alice's Adventures in Wonderland:*

- "The different branches of Arithmetic—Ambition, Distraction, Uglification, and Derision."
- "Then you should say what you mean," the March Hare went on.

  "I do," Alice hastily replied; "at least I mean what I say, that's the same thing, you know."

  "Not the same thing a bit!" said the Hatter. "Why, you might just as well say that 'I see what I eat' is the same thing as 'I eat what I see!' "
- "Take some more tea," the March Hare said to Alice, very earnestly.

  "I've had nothing yet," Alice replied in an offended tone, "so I can't take more."

  "You mean you can't take *less,*" said the Hatter. "It's very easy to take *more* than nothing."

From *The Hunting of the Snark:*

- "What I tell you three times is true."

From *Alice through the Looking Glass*

- "Can you do addition?" the White Queen asked.

  "What's one and one and one and one and one and one and one and one and one and one?"

  "I don't know," said Alice. "I lost count."
- "It's very good jam," said the Queen.

  "Well, I don't want any to-day, at any rate."

  "You couldn't have it if you did want it," the Queen said. "The rule is jam tomorrow and jam yesterday but never jam to-day."

  "It must come sometimes to "jam to-day," Alice objected.

  "No it can't," said the Queen. "It's jam every other day; to-day isn't any other day, you know."

  "I don't understand you," said Alice. "It's dreadfully confusing."
- "When I use a word," Humpty Dumpty said, in a rather scornful tone, "it means just what I choose it to mean—neither more nor less."

  "The question is," said Alice, "whether you can make words mean so many different things."

  "The question is," said Humpty Dumpty, "which is to be master—that's all."

Cartesian geometry

See **analytical geometry**.

Cartesian coordinates

An ordered set of **real number**s that defines the position of a point in terms of its projection onto mutually perpendicular number lines. In the plane, each point is defined by two such projections, one onto the *x*-axis and one onto the y-axis, and is written as an **ordered pair** of real numbers (x, y). The same system works equally in spaces of three or more dimensions.

Cartesian oval

A curve that actually consists of two ovals, one inside the other. It is the locus of a point whose distances s and t from two fixed points S and T satisfy the equation $s + mt = a$. When c is the distance between S and T then the curve can be expressed in the form:

$$((1 - m^2)(x^2 + y^2) + 2m^2cx + a^2 - m^2c^2)^2 = 4a^2(x^2 + y^2)$$

The curves were first studied by René **Descartes** in 1637 and are sometimes called the *ovals of Descartes;* they were also investigated by Isaac **Newton** in his classification of **cubic curves**. If $m = \pm 1$, then the Cartesian oval is a central conic. If $m = a/c$, then it becomes a **limaçon of Pascal**, in which case the inside oval touches the outside one. Cartesian ovals are **anallagmatic curves**.

Cassinian ovals

Also known as *Cassini's ovals,* a family of curves, each member of which is defined as follows: given two points A and B and a constant c^2, the locus of points P with $PA \times PB = c^2$. The locus has the equation $(x^2 + y^2)^2 - 2a^2(x^2 - y^2)^2 - a^4 + c^4 = 0$, where $a = AB$. Equivalently, Cassinian ovals can be thought of as the set of curves produced when a circular **torus** is sliced at every possible point parallel to its axis. If $c = a$, then the curve is a special case known as the **lemniscate of Bernoulli** (a figure-eight type curve). The ovals are named after the Italian-born astronomer Giovanni Cassini (1625–1712) who first investigated them in 1680 while studying the relative motions of Earth and the Sun. Cassini thought that the Sun traveled round Earth on one of these curves (rather than the **ellipse**, as correctly proposed in Kepler's heliocentric scheme), with Earth at one focus.

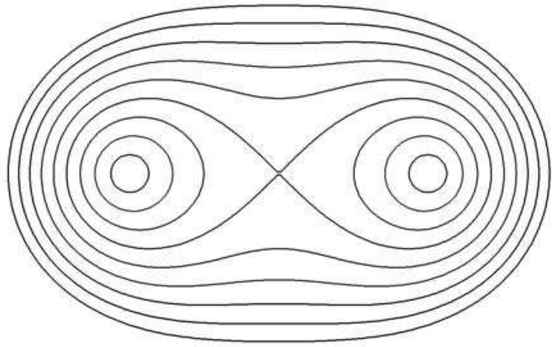

Cassinian ovals The many different ways to slice a doughnut. *Xah Lee, www.xahlee.org*

casting out nines

A method for checking arithmetic that uses the idea of the **digital root** of a number. Let the digital root of a number n be $r(n)$; for example, $r(7{,}586) = 8$. For any two numbers a and b: $r(a + b) = r(r(a) + r(b))$ and $r(a \times b) = r(r(a) \times r(b))$. These rules allow checks on addition and multiplication as the following examples show. Does $7{,}586 + 9{,}492 = 16{,}978$? $r(r(7{,}586) + r(9{,}492)) = r(8 + 6) = 5$; $r(16{,}978) = 4$; so the sum given is incorrect. Does $7{,}586 \times 9{,}492 = 72{,}006{,}312$. $r(r(7{,}586) \times r(9{,}492)) = r(8 \times 6) = r(48) = 3$; $r(72{,}006{,}312) = r(21) = 3$; which suggests that the product given is likely to be correct. The name "casting out nines" comes from the fact that nines need not be included in the calculation of the digital roots, since they have no effect on the final result. This a direct outcome of the fact that we use a **decimal** number system. If we calculated instead in octal (base eight), say, then the process would be one of "casting out sevens." This kind of checking will pick up most errors, but not all. For example, an interchange of two digits will not be detected, nor will replacing a nine by a zero or vice versa. The method appears in the work of ninth-century Arab mathematicians but may have originated earlier with the Greeks and, possibly, the Hindus.

Catalan number

Any number, u_n, from the *Catalan sequence* defined by

$$u_n = (2n)! \,/\, (n + 1)!n!.$$

It begins: 1, 2, 5, 14, 42, 132, 429, 1,430, 4,862, 16,796, 58,786, 208,012, 742,900, ... The values of u_n represent the number of ways a **polygon** with $n + 2$ sides can be cut into n triangles using straight lines joining vertices (see **vertex**). Catalan numbers are named after the Belgian mathematician Eugène Catalan (1814–1894). They also arise in other counting problems, for example in determining how many ways $2n$ beans can be divided into two containers if one container can never have less than the second.

Catalan solid

A **polyhedron** that is a dual of an **Archimedean solid**. (A dual of a polyhedron is obtained by replacing each face with a vertex, and each vertex with a face.) Catalan solids are named after the Belgian mathematician Eugène Catalan (1814–1894) who first described them in 1865. See also **Platonic solid** and **Johnson solid**. (See table, "Catalan solids.")

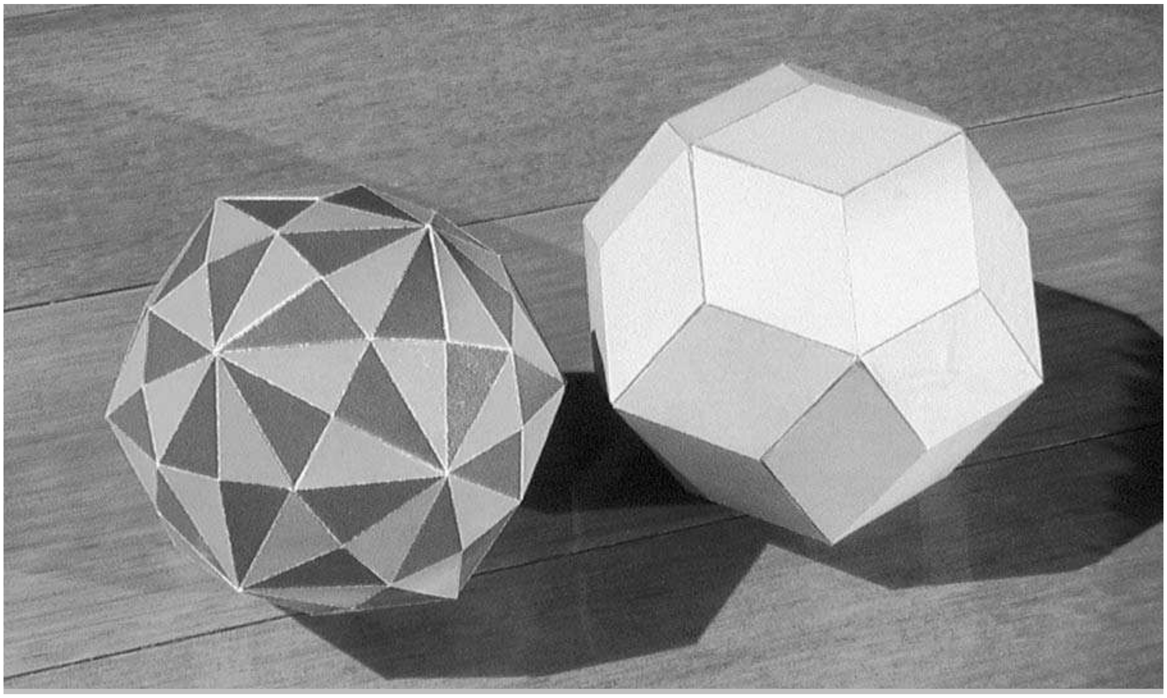

Catalan solid Two of the Catalan solids: the rhombic tricontahedron (right) and the disdyakistriacontahedron (left). *Robert Webb, www.software3d.com; created using Webb's Stella program*

Catalan Solids

Name	Corresponding Archimedean Solid
Triakis tetrahedron	Truncated tetrahedron
Rhombic dodecahedron	Cuboctahedron
Triakis octahedron	Truncated cube
Tetrakis hexahedron	Truncated octahedron
Deltoidal icositetrahedron	Small rhombicuboctahedron
Disdyakis dodecahedron	Great rhombicuboctahedron
Pentagonal icositetrahedron	Snub cube
Rhombic triacontahedron	Icosidodecahedron
Triakis icosahedron	Truncated dodecahedron
Pentakis dodecahedron	Truncated icosahedron
Deltoidal hexecontahedron	Rhombicosidodecahedron
Disdyakis triacontahedron	Great rhombicosidodecahedron
Pentagonal hexecontahedron	Snub dodecahedron

Catalan's conjecture

The hypothesis, put forward by the Belgian mathematician Eugène Catalan (1814–1894) in 1844, that 8 (= 2^3) and 9 (= 3^2) are the only pair of consecutive powers. In other words, the *Catalan equation* for **prime number**s p and q and positive integers x and y

$$x^p - y^q = 1$$

has only the one solution

$$3^2 - 2^3 = 1.$$

In 1976 R. Tijdeman took the first major step toward showing this is true by proving that for any solution, y^q is less than **e** to the power e to the power e to the power e to the power 730 (a huge number!). Since then, this bound has been reduced many times, and it is now know that the larger of p and q is at most 7.78×10^{16} and the smaller is at least 10^7. On April 18, 2002, the Romanian number theorist Preda Mihailescu sent a manuscript to several mathematicians with a proof of the entire conjecture together with an analysis by Yuri Bilu. It is expected that as soon as this work is completely reviewed by other mathematicians that Catalan's conjecture will have been proved.[266]

Solutions to Catalan's conjecture and **Fermat's last theorem** are special cases of the *Fermat-Catalan equation:*

$$x^p + y^q = z^r,$$

where x, y, and z are positive, **coprime** integers and the exponents are all primes with

$$1/p + 1/q + 1/r \leq 1.$$

The *Fermat-Catalan conjecture* is that there are only finitely many solutions to this system. These solutions include:

$1^p + 2^3 = 3^2$ ($p \geq 2$); $2^5 + 7^2 = 3^4$

$13^2 + 7^3 = 2^9$;

$33^8 + 1{,}549{,}034^2 = 15{,}613^3$

and $43^8 + 96{,}222^3 = 30{,}042{,}907^2$.

Catalan's constant

A constant that crops up regularly in combinatorial problems, especially in the evaluation of certain infinite series and integrals. For example, it is equal to

$\int_0^1 \arctan(x) / x \, dx$, and

$1 - 1/3^2 + 1/5^2 - 1/7^2 + 1/9^2 - \ldots$

It is also the solution to the following problem as n becomes arbitrarily large: If you have a $2n \times 2n$ checkerboard and a supply of $2n^2$ dominoes that are just large enough to cover two squares of the checkerboard, how many ways are there to cover the whole board with the dominoes? Catalan's constant has the value 0.915965 . . . ; it is not known if it's an **irrational number**.

catastrophe theory

A theory, developed by the French mathematician René Thom (1923–2003), that attempts to explain the behavior of complex **dynamical system**s by relating it to **topology**. The evolution of such systems consists of steady continuous change interspersed with sudden major jumps, or "catastrophes," when the topology of the set changes. Catastrophe theory has been applied, with varying degrees of success, to phenomena as diverse as earthquakes, stock market crashes, prison riots, and human conflicts, at the personal, group, and societal level. The theory was first developed by Thom in a paper published in 1968 but became well known through his book *Structural Stability and Morphogenesis* (1972).[331] Many mathematicians took up the study of catastrophe theory and it was in tremendous vogue for a while, yet it never achieved the success that its younger cousin **chaos** theory has because it failed to live up to its promise of useful predictions. Late in his career, the surrealist Salvador Dali painted *Topological Abduction of Europe: Homage to René Thom* (1983), an aerial view of a seismically fractured landscape juxtaposed with the equation that strives to explain it.

catch-22

A situation in which a person is frustrated by a paradoxical rule or set of circumstances that preclude any attempt to escape from them. The name comes from the title of a novel by Joseph Heller (1923–1999), based on his personal experiences, about an American airman's attempts to survive the madness of World War II. Heller wrote:

> There was only one catch and that was Catch-22, which specified that concern for one's own safety in the face of dangers that were real and immediate was the process of a rational mind. Orr was crazy and could be grounded. All he had to do was ask; and as soon as he did, he would no longer be crazy and would have to fly more missions. Orr would be crazy to fly more missions and sane if he didn't, but if he was sane he had to fly them. If he flew them he was crazy and didn't have to; but if he didn't want to he was sane and had to.

category theory

The study of abstracted collections of mathematical objects, such as the category of **set**s or the category of **vector space**s, together with abstracted operations sending one object to another, such as the collection of **function**s from one set to another or linear **transformation**s from one vector space to another.

catenary

The shape that a rope or telephone cable makes, under the influence of gravity, when suspended between two points. The word comes from the Latin *catena,* meaning "chain," and was first used by Christiaan **Huygens** while studying the form of suspended chains. Galileo thought the shape would be a **parabola**. In fact, near the **vertex**, a parabola and a catenary do look very similar. When x is slightly greater than three, however, the catenary begins to rapidly outgrow the value of the parabola. The two shapes are related in another way. If a parabola is rolled along a straight line, the **focus** of the parabola moves along a catenary curve. Surprisingly, too, if a bicycle with square (or any polygon-shaped) wheels is ridden along a road made of upturned catenaries the wheels will roll smoothly and the rider will stay at the same height! The St. Louis Arch, which is 192 meters wide at the base and 192 meters tall, follows the form of a catenary, the exact formula for which is displayed inside the arch: $y = 68.8 \cosh(0.01x - 1)$, where cosh is the hyperbolic cosine function.

The general equation of a catenary can be written

$$y = k \cosh(x/k),$$

where k is a constant, or, in terms of the **exponential** function,

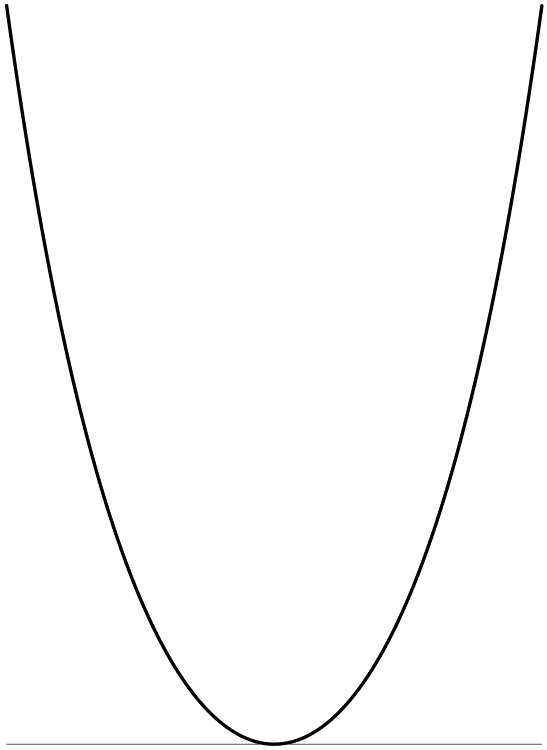

catenary The catenary curve. © *Jan Wassenaar, www.2dcurves.com*

$$y = k(e^{x/k} + e^{-x/k})/2.$$

In the special case where $k = 1$, these reduce to

$$y = \cosh(x) = (e^{x} + e^{-x})/2.$$

In terms of a polynomial series

$$\begin{aligned} y &= 0.5\ (1 + x + x^2/2! + x^3/3! + x^4/4! + \ldots + 1 - x + x^2/2! \\ &\quad - x^3/3! + x^4/4! - \ldots) \\ &= 1 + x^2/2! + x^4/4! + x^6/6! + \ldots. \end{aligned}$$

For small values of x the terms beyond $x^2/2!$ are very small, so that the equation closely approximates that of a parabola, as we have already seen.

catenoid

The **surface of revolution** produced when a **catenary** rotates about its central axis. The catenoid was first described by Leonhard **Euler** in 1740 and is the oldest known **minimal surface** (a shape of least area when bounded by a given closed space). It is the minimal surface connecting two parallel circles of unequal diameter on the same axis; soap film between two circular rings takes this form (see also **bubble**s). The catenoid is the only known minimal surface that is also a surface of revolution, and is one of only four minimal surfaces that have the topological properties of being unbounded, embedded, and non-periodic; the others are the simple plane, the **helicoid**, and *Costa's surface.*

cathetus

A line that is perpendicular to another line. Usually, it refers to one of the lines in a right triangle that is not the **hypotenuse**.

Cauchy, Augustin Louis, Baron (1789–1857)

A French mathematician who founded **complex analysis** by discovering the *Cauchy-Riemann equations* and wrote 789 papers–an output surpassed only by Leonhard **Euler**, George **Cayley**, and Paul **Erdös**. He coined the name for the **determinant** and systematized its study and gave nearly modern definitions of **limit**, **continuity**, and **convergence**.

causality

The relationship between causes and effects. An event or state of affairs *A* is the cause of an event *B* if *A* is the reason that brings about the effect *B*. For instance, one might say, "my pushing the gas pedal *caused* the car to go faster." An important question in philosophy and other fields is how (and if) causes can bring about effects. In a strict reading, if *A* causes *B*, then *A* must *always* be followed by *B*. In this sense, for example, smoking doesn't cause cancer. In everyday usage, we therefore often take "*A* causes *B*" to mean "*A* causes an increase in the probability of *B*." The establishment of cause and effect, even with this relaxed reading, is notoriously difficult. The Scottish philosopher David Hume held that causes and effects are not real, but instead are imagined by our minds to make sense of the observation that *A* often occurs together with or slightly before *B*. All we can actually observe are *correlations,* not causations. This is also expressed in the logical fallacy, "correlation implies causation." For instance, the observation that smokers have a dramatically increased lung cancer rate doesn't establish that smoking must be the *cause* of that increased cancer rate: maybe there exists a certain genetic defect which both causes cancer and a yearning for nicotine.[194]

caustic

The envelope of rays of light reflected (or refracted) by a given curve from a given point source of light; a *catacaustic* results from reflection, a *diacaustic* from refraction. Among the caustic curves of a circle are a lima, if the light source is nearby, a **nephroid**, if the source is at infinity, and a **cardioid**, if the source is on the circle.

Cavalieri's principle

If two solids have the same height and the same cross-sectional area at every level, then they have the same volume. This principle is named after the Italian mathematician Bonaventura Cavalieri (1598–1647).

Cayley, Arthur (1821–1895)

A British mathematician who made important contributions to **non-Euclidean geometry** and the algebra of matrices (see **matrix**). The former eventually found its way into the study of the **space-time** continuum, the latter into a formulation of **quantum mechanics** by the German physicist Werner Heisenberg. Cayley was also far ahead of his time in pioneering the idea of abstract **group**s.

Cayley number

See **octonion**.

Cayley's mousetrap

A **permutation** problem invented by Arthur **Cayley**. Write the numbers 1, 2, . . . , n on a set of cards and shuffle the deck. Start counting using the top card. If the card chosen does not equal the count, move it to the bottom of the deck and continue counting forward. If the card chosen *does* equal the count, discard the chosen card and begin counting again at 1. The game is won if all cards are discarded, and lost if the count reaches $n + 1$. The number of ways the cards can be arranged such that at least one card is in the proper place for $n = 1, 2, \ldots$, are 1, 1, 4, 15, 76, 455, ...

Cayley's sextic

A sinusoidal spiral curve described by the Cartesian equation

$$4(x^2 + y^2 - ax)^3 = 27a^2(x^2 + y^2)^2.$$

It was discovered by Colin **Maclaurin** but was first studied in detail by Arthur **Cayley** and named after him by R. C. Archibald in 1900.

ceiling

The largest value that something can take. The *ceiling function* of a number x is the smallest integer that is not smaller than x.

cell

(1) A three-dimensional object that is part of a higher-dimensional object, such as a **polychoron**. A cell is related to higher-dimensional objects in the way that a face, or (two-dimensional) **polygon**, is related to higher-dimensional objects. For example, a cell is to a Four-dimensional polytope, or polychoron, what a face is to a three-dimensional polytope, or **polyhedron**. Often polytopes are classified simply by how many cells they have. For example, the **tesseract** has eight cells, each one of which is a cube. (2) The fundamental spatial unit operated on by the rules of a **cellular automaton** during one generation.

cellular automaton

An array of cells that evolves according to a set of rules based on the states of surrounding cells; for example, a cell might be "on" if its four neighbor cells (east, west, north, and south) are also on. The entire array can self-organize into global patterns that may move around the screen. These patterns can be quite complex even though they emerge from just a few very simple rules governing the connections among the cells. Cellular automata are the simplest models of spatially distributed processes. They were first investigated by John **von Neumann** in about 1952. Von Neumann incorporated a cellular model into his "universal constructor" and also proved that an automaton consisting of cells with four orthogonal neighbors and 29 possible states would be capable of simulating a **Turing machine** for some configuration of about 200,000 cells. The best-known cellular automaton is Life (see **Life, Conway's game of**). Another example is **Langton's ant**. The study of cellular automata and their patterns has led to insights into the way structure is built up in biological and other complex systems, and for this reason forms part of the subject of **artificial life**.

celt

Also known as a *rattleback,* a simple ancient toy that behaves in a very counterintuitive way. When spun one way about its vertical axis, the celt spins for a long time. When spun the other way, however, a wobble quickly sets in that halts the rotation and then, incredibly, reverses it. In his 1986 paper on the subject, the British physicist Hermann Bondi wrote: "Many people, even trained scientists, find it hard to understand that the behaviour of the toy doesn't violate the principle of conservation of angular momentum."[47] The celt's remarkable antics stem from three factors: a curved base that has two different radii–one long radius for the lengthwise curve and one shorter radius for the tighter curve across the width; axes of symmetry that are skewed slightly from the principal axes of inertia; and a different distribution of mass about each of the two horizontal axes of inertia. To understand how the celt can switch direction halfway through its performance, think of the frictional force that acts at the point of contact between the celt and the surface. One component of the friction creates a torque (twisting force) that tends to rotate the celt about its vertical axis. The point of contact is moving all the time and the torque changes. If the inertial and symmetrical axes coincided, the average torque over a single oscillation would be zero. But for the celt, there's a net torque in one direction, which reverses the angular momentum. See also **Tippee Top**.

center of perspective

The point where the lines joining corresponding points of two figures that are in perspective meet.

centillion

See **large number**.

central angle

The angle subtended at the center of a **circle** by an **arc** or a **chord**; in other words, the angle between two radii.

centroid

For a **triangle**, the point of intersection of the **median**s. For any other shape, the point where coordinates are the average of the coordinates of the shape's vertices (see **vertex**). The centroid is the *center of mass* of a figure.

century

A period of 100 years. The original Latin *centuria* means simply "one hundred" and was used to describe any collection of 100 items. In the Roman army, a century was a group of 100 men, each known as a *centurion.* One of the few modern examples of "century" being used other than to denote a period of time is in the game of cricket where a batsman who scores 100 runs in an inning is said to have "made a century."

Ceva, Giovanni (1647–1734)

A Jesuit-trained Italian mathematician who specialized in **geometry**. His greatest discovery, now known as *Ceva's theorem,* can be stated as follows. Given a triangle with vertices (corners), A, B, and C and points D, E, and F on the opposite sides, the lines AD, BE, and CF will intersect at a single point if $BD \times CE \times AF = DC \times EA \times FB$. The term *cevian line* was coined by French geometers around the end of the eighteenth century to honor Ceva. It is defined as any line joining a vertex of a triangle to a point on the opposite side. The **median**, **altitude**, and angle **bisect**or are all examples of cevians. The perpendicular bisector, however, in most cases, is not a cevian because it doesn't usually pass through a vertex.

chained arrow notation

See **Conway's chained arrow notation**.

Chaitin, Gregory (1947–)

An American mathematician and computer scientist at IBM's T. J. Watson Research Center who is the chief architect of a new subject known as *algorithmic information theory,* which has profound consequences for our ideas about **random**ness. In particular, because of the limitations of computers and the programs they run, Chaitin has shown that there is an inherent uncertainty or unknowability in mathematics that is similar to the uncertainty principle in physics. Although there are an infinite number of mathematical facts, they are, for the most part, unrelated to each other and impossible to tie together with unifying theorems. His powerful message is that most of mathematics is true for no particular reason; math is true by accident. See **Chaitin's constant**.

Chaitin's constant

A **real number**, represented by capital omega (Ω) and also known as the *Halting probability,* whose digits are distributed so **random**ly that no rule can be found to predict them. Discovered by Gregory **Chaitin**, Ω is definable but not computable. It has no pattern or structure to it whatsoever, but consists instead of an infinitely long string of zeros and ones in which each digit is as unrelated to its predecessor as one coin toss is to the next. Although called a constant, it is not a constant in the sense that, for example, **pi** is, since its definition depends on the arbitrary choice of computation model and programming language. For each such model or language, Ω is the probability that a randomly produced string will represent a program that, when run, will eventually halt. To derive it, Chaitin considered all the possible programs that a hypothetical computer known as a **Turing machine** could run, and then looked for the probability that a program, chosen at random from among all the possible programs, will halt. He eventually showed that this halting probability turns Turing's question of whether a program ever stops into a real number, somewhere between zero and one. He further showed that, just as there are no computable instructions for deciding in advance whether a computer will halt, there are also no instructions for determining the digits of Ω. Omega is uncomputable and unknowable: we don't know its value for any programming language and we never will. This is extraordinary enough in itself, but Chaitin has found that Ω permeates the whole of mathematics, placing fundamental limits on what we can know.

And Ω is just the beginning. There are more disturbing numbers called Super-Omegas, whose degree of randomness is vastly greater even than that of Ω. If there were an omnipotent computer that could solve the halting problem and evaluate Ω, this mega-brain would have its own unknowable halting probability called Ω'. And if there were a still more godlike machine that could find Ω', its halting probability would be Ω''. These higher Omegas, it has been recently discovered, are not meaningless abstractions. Ω', for instance, gives the probability that an infinite computation produces only a finite amount of output. Ω″ is equivalent to the probability that, during an infinite computation, a computer will fail to produce an output–for example, get no result from a computation and move on to the next one–and that it will do this only a finite number of times. Omega and the Omega hierarchy are revealing to mathematicians an unsettling truth: the problems that we can hope ever to solve form a tiny archipelago in a vast ocean of undecidability.[130]

Champernowne's number

The first known **normal number**. It was discovered in 1933 by the English mathematician David G. Champernowne and consists of a decimal fraction in which the decimal integers are written down in increasing order: 0.12345678910111213 . . . . Champernowne's number has been proven to be a normal number in base 10 and also to be an **irrational number**. However, although its digits appear with equal frequency, the sequence of its digits are not unpredictable. An example of a number whose sequence of digits *is* unpredictable is **Chaitin's constant**.

chance

See **probability theory**.

change ringing

> *The art of change ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world. To the musical Belgian, for example, it appears that the proper thing to do with a carefully tuned ring of bells is to play a tune upon it. By the English campanologist . . . the proper use of the bells is to work out mathematical permutations and combinations.*
>
> –Dorothy L. Sayers, *The Nine Tailors*

The ringing of a set of bells in a precise relationship to one another to produce a pleasing sound. Bells are numbered 1, 2, 3, 4, 5 . . . from lightest (highest-pitched) to heaviest. After each sequence, or *round*, the order of the bells is changed slightly in a predetermined way. With 5 bells, there are 5 × 4 × 3 × 2 × 1, or 120, possible changes, which take about 4 minutes to ring. With 6, 7, or 8 bells, the number of unique changes is 720, 5,040, and 40,320, respectively. To produce pleasing variations in the sound, bells are made to change places with adjacent bells in the row, for example:

1 2 3 4 5 6 7 8
2 1 4 3 6 5 8 7

These rows are the musical notation of change ringing. No bell moves more than one place in the row at a time, although more than one pair may change in the same row. In order to ring a different row with each pull of the rope, ringers have devised methods for changing pairs in orderly ways. In ringing a method, the bells begin in rounds, ring changes according to the method, and return to rounds without repeating any row along the way. These place changes produce musical patterns, with the sounds of the bells weaving in and out. For example, a "Plain Hunt Minimus" with four bells is rung as shown in the diagram.

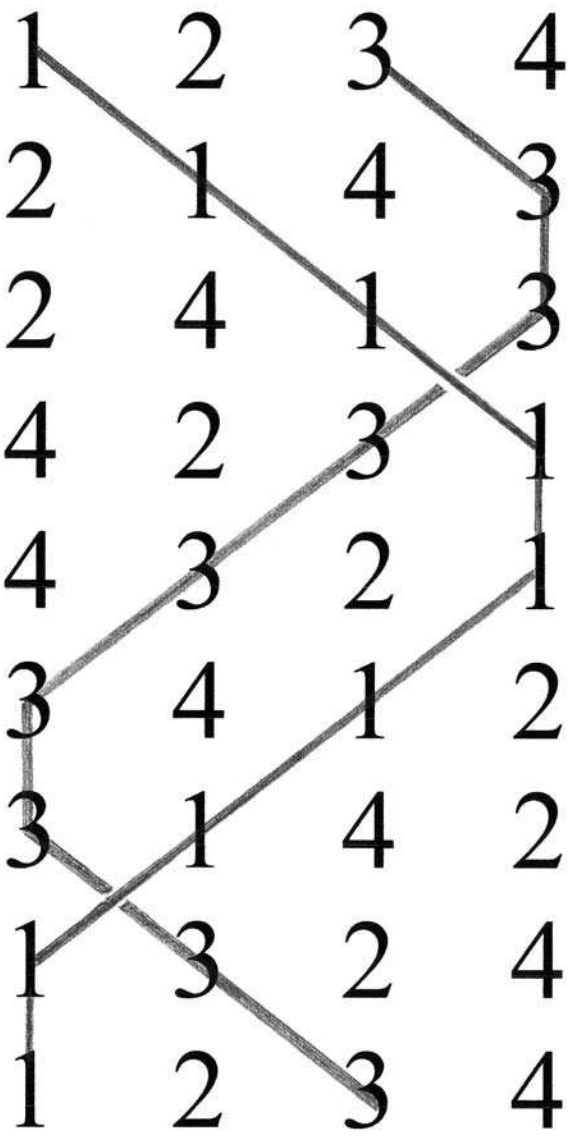

change ringing The "Plain Hunt Minimus" for four bells. The sequences for bells 1 and 3 are shown by lines.

Experienced ringers test and extend their abilities by ringing peals: 5,000 or more changes without breaks or repeating a row. Peals customarily last about three hours. The first peal was rung in England in 1715. Chiming bells (swinging them through a short arc using a rope and a lever) goes well back into the Middle Ages, but it wasn't until the seventeenth century that ringers developed the full wheel, which allowed enough control for orderly ringing. In 1668 Fabian Stedman published *Tintinnalogia* (The art of change ringing), containing all the available information on systematic ringing. The theory of change ringing set forth by Stedman has been refined in later years but remains essentially unchanged today. Bells for change ringing are hung in stout frames that allow the bells to swing through 360°. Each bell is attached to a wooden wheel with a handmade rope running around it and takes about 2 seconds to rotate. The bells are arranged in the frame so their ropes hang in a circle in the ringing chamber below. Into each rope is woven a tuft of brightly colored wool (a *sally*), which marks where the ringer must catch the rope while ringing. Bells are rung from the "mouth up" position. With a pull of the rope, the bell swings through a full circle to the up position again. With the next pull it swings back in the other direction. The plot of Dorothy Sayers's *The Nine Tailors* (1934), considered one of her best works, revolves around the art of change ringing.

chaos

> *We adore chaos because we love to produce order.*
> –M. C. Escher

A phenomenon shown by some **dynamical systems**, which consists of a curious, infinitely complex pattern of behavior that lies just beyond the edge of total order. A system is chaotic if it is predictable in principle and yet is unpredictable in practice over long periods because its behavior depends very sensitively on initial conditions. Despite this unpredictability, however, there are certain constants, such as **Feigenbaum's constant**, and certain structures, such as **chaotic attractors**, that are fixed and susceptible to analysis. The weather, the movements of a metal pendulum moving over fixed magnets, and the orbits of closely spaced moons are all examples of chaotic systems. Although the ideas behind modern chaos theory were actively studied at some level throughout most of the twentieth century, the word as a mathematical term dates only from an article in *American Mathematical Monthly* in 1975 called "Period Three Implies Chaos."

In everyday language, chaos has come to mean the exact opposite of order. But the Greek root *khaox* means "empty space" and this meaning still persists in archaic usage where it refers to a canyon or abyss. The evolution of the word to mean disorder seems to come from reference to the time before God created the universe. Empty space was formless and the creation filled the emptiness and established order. Mathematical chaos represents an unexpected third state: a deterministic system subject to simple rules that nevertheless displays infinitely complex behavior.[135]

chaos tiles

See **Penrose tiling**.

chaotic attractor

Also known as a *strange attractor,* a type of **attractor** (i.e., an attracting set of states) in a complex **dynamical system**'s **phase space** that shows **sensitivity** to initial conditions. Because of this property, once the system is on the attractor, nearby states diverge from each other exponentially fast. Consequently, small amounts of noise are amplified. Once sufficiently amplified the noise determines the system's large-scale behavior and the system is then unpredictable. Chaotic attractors themselves are markedly patterned, often having elegant, fixed geometric structures, despite the fact that the trajectories moving within them appear unpredictable. The chaotic attractor's geometric shape is the order underlying the apparent chaos. It functions in much the same way as someone kneading dough. The local separation of trajectories corresponds to stretching the dough and the global attraction property corresponds to folding the stretched dough back onto itself. One result of the stretch-and-fold aspect of chaotic attractors is that they are **fractals**; that is, some cross section of them reveals similar structure on all scales.

character theory

The study of the traces (sums of the diagonal elements) of the **matrix** representations of a **group**. The information gained is listed in *character tables,* the properties of which give insight into the group's properties.

chess

A game of strategy for two players that probably originated in India, though the earliest documentary references are in Chinese and Persian texts in about A.D. 600. Each player has 16 pieces, either black or white, consisting of eight pawns, two rooks (also known as castles), two knights, two bishops, a queen, and a king. The object is to lay siege to the opposing king in such a way that it cannot escape attack–a position known as *checkmate* (from the Persian phrase *Shah Mat,* meaning "the king is dead"). There are 400 first-move combinations–20 for white × 20 for black (though only 64 of these are regarded as strong), 318,979,564,000 ways of playing the first four moves, and 169,518,829,100,544,000 trillion ways of playing the first

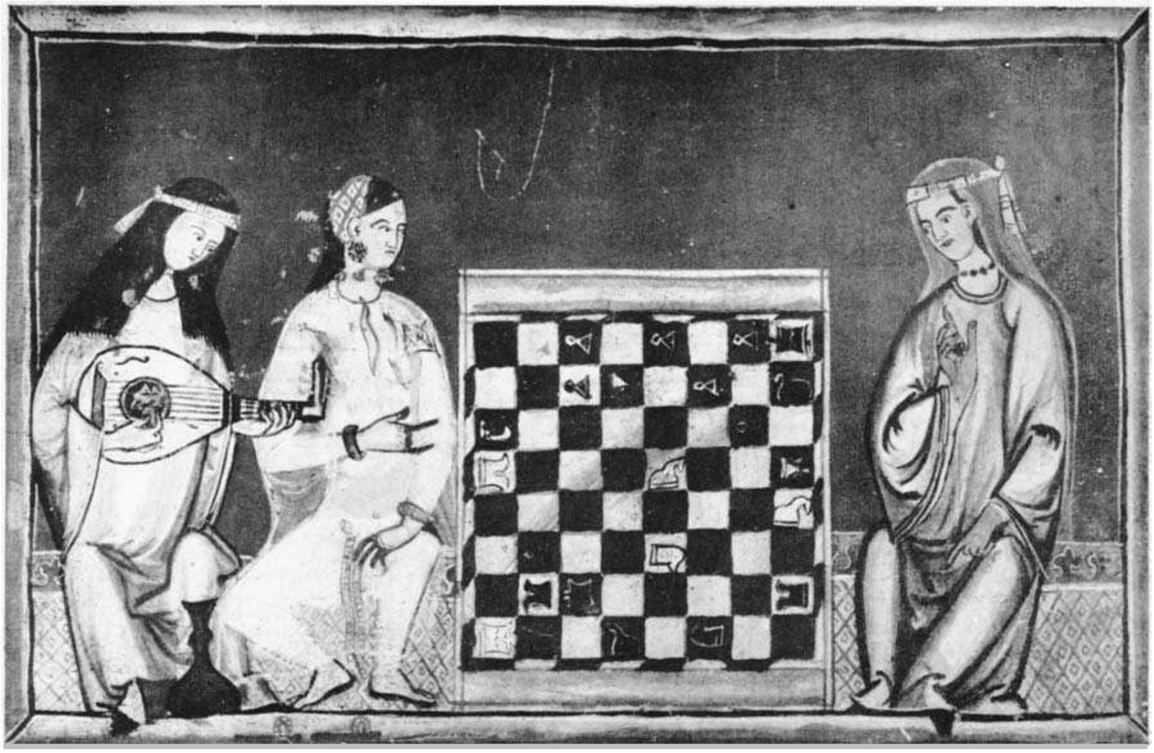

chess An illumination from the *Cantigas de Santa Maria* (a thirteenth-century collection of songs) showing a chess game in progress.

10 moves. The total number of possible board configurations is estimated at 10^{120}; for comparison, that of **Go** is generally put at about 10^{174}.

A standard chessboard is a square plane divided into 64 smaller squares by straight lines at right angles. Originally, it wasn't checkered (that is, made with its rows and columns alternately of dark and light colors), and this feature was introduced merely to help the eye in actual play. In many puzzles based on chess the utility of checkering is questionable, and the board may be generalized to any $n \times n$ size.

One of the first puzzles to use a chessboard was the **wheat and chessboard problem**, posed in 1256 by the Arabic mathematician Ibn Kallikan. Among the earliest problems to involve chess *pieces,* proposed by Guarini di Forli in 1512, asks how two white and two black knights can be interchanged, using normal knight's moves, if they are placed at the corners of a 3 × 3 board. The unusual L-shaped movement of the knight is what makes one of the best known chess puzzles, the **knight's tour**, such a challenge. Other standard puzzles, often called simply the **kings problem**, the **queens puzzle**, the **rooks problem**, the **bishops problem**, and the **knights problem**, ask for the greatest number of each of these pieces that can be placed on an 8 × 8 board or on a generalized $n \times n$ board without attacking each other, and/or the smallest number of each of these pieces that are needed to occupy or attack every square. *Fairy chess* is any variant on the standard game, which may involve a change in the form of the board, the rules of play, or the pieces used. For example, the normal rules of chess can be used but with a cylindrical or **Möbius band** connection of the edges.

Chinese cross

See **burr puzzle**.

Chinese remainder theorem

If there are n numbers, a_1 to a_n, that have no **factor**s in common (i.e., are pairwise relatively prime), then any integer greater than or equal to 0 and less than the product of all the numbers n can be uniquely represented by a series consisting of the remainders of division by the numbers n. For example, if $a_1 = 3$ and $a_2 = 5$, the Chinese remainder theorem (CRT) says that every integer from 0 to 14 will have a unique set of remainders when divided separately by (**modulo**) 3 and 5. Listing out all the possibilities shows that this is true:

0 has a remainder of 0 modulo 3 and a remainder of 0 modulo 5.

1 has a remainder of 1 modulo 3 and a remainder of 1 modulo 5.

2 has a remainder of 2 modulo 3 and a remainder of 2 modulo 5.

3 has a remainder of 0 modulo 3 and a remainder of 3 modulo 5.

4 has a remainder of 1 modulo 3 and a remainder of 4 modulo 5.

5 has a remainder of 2 modulo 3 and a remainder of 0 modulo 5.

6 has a remainder of 0 modulo 3 and a remainder of 1 modulo 5.

7 has a remainder of 1 modulo 3 and a remainder of 2 modulo 5.

8 has a remainder of 2 modulo 3 and a remainder of 3 modulo 5.

9 has a remainder of 0 modulo 3 and a remainder of 4 modulo 5.

10 has a remainder of 1 modulo 3 and a remainder of 0 modulo 5.

11 has a remainder of 2 modulo 3 and a remainder of 1 modulo 5.

12 has a remainder of 0 modulo 3 and a remainder of 2 modulo 5.

13 has a remainder of 1 modulo 3 and a remainder of 3 modulo 5.

14 has a remainder of 2 modulo 3 and a remainder of 4 modulo 5.

CRT enables problems such as the following to be solved: Find the two smallest counting numbers that will each have the remainders 2, 3, and 2 when divided by 3, 5, and 7, respectively. It is said that the ancient Chinese used a variant of this theorem to count their soldiers by having them line up in rectangles of 7 by 7, 11 by 11, and so forth. After counting only the remainders, they solved the associated system of equations for the smallest positive solution.

Chinese rings

One of the oldest known **mechanical puzzle**s, the object of which is to remove all *n* rings from a horizontal loop of stiff wire, and/or put them back on the loop. On the first move it is possible to take up to two rings off the left end of the wire. One or both of those can then be slipped through the wire loop (from top to bottom). If both are removed then the fourth ring can be slipped over the end. If just one of the first two is removed, then the next step is to slip the third ring over the end. Subsequently, rings must be put back on to the wire loop in order to remove other rings, and this procedure is repeated over and over again. In general, the minimum number of moves needed is $(2^{n+1} - 2)/3$ if n is even and $(2^{n+1} - 1)/3$ if n is odd. For example, with seven rings the solution takes 85 moves. Most of the solution is easy, as each move normally involves going forward or back to the previous state. The key to a correct solution is the first step: if *n* is even, you must remove two rings; if *n* is odd, you must remove only one. The solution is similar to that of the **Tower of Hanoi**. In fact, Edouard **Lucas**, who invented the Tower of Hanoi, gave an elegant solution to the Chinese rings that uses **binary** arithmetic.

Stewart Cullin, the noted nineteenth-century ethnologist, relates that the puzzle was invented by the famous Chinese general Chu-ko Liang (A.D. 181–234), in the second century, as a present to his wife so that, in trying to solve it, she would have something to do while he was away at the wars. However, this is anecdotal and its origins remain obscure. The earliest reference to it in Europe may be in about 1500 in the form of Problem 107 of the manuscript *De Viribus Quantitatis* by Luca **Pacioli** in which the description appears: "*Do Cavare et Mettere una Strenghetta Salda in al Quanti Anelli Saldi, Difficil Caso*" (Remove and put a little bar joined in some joined rings,

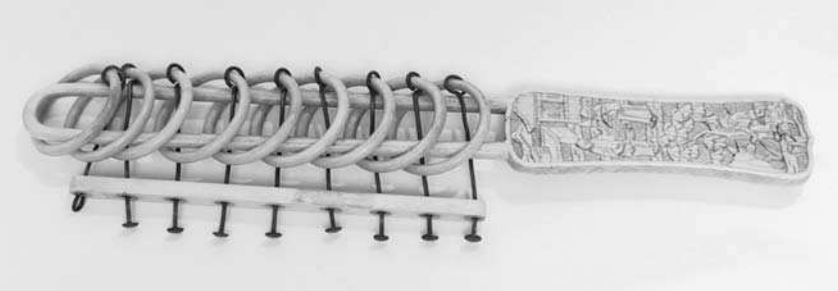

Chinese rings An unusual example of Chinese rings in ivory, dating from the mid-nineteenth century. *Sue & Brian Young/Mr. Puzzle Australia, www.mrpuzzle.com.au* difficult case). It was also mentioned by Girolamo **Cardan** in the 1550 edition of his book *De Subtililate* from which comes the name *Cardan's rings,* and was treated at length in mathematical terms by John **Wallis** in about 1685. By the end of the seventeenth century, it had become popular in many European countries. French peasants used it to lock chests and called it *baguenaudier,* or "time-waster."

Chinese room

An argument first put forward by the American philosopher John Searle (1932–) in 1980 in an attempt to show that the human mind is not a computer and that the **Turing test** is not adequate to prove that a machine can have strong **artificial intelligence** (strong AI)–in other words, can think in a humanlike way.[292] In the Chinese room scenario, a person who understands no Chinese sits in a room into which written Chinese characters are passed. The person uses a complex set of rules, established ahead of time, to manipulate these characters, and pass other characters out of the room. The idea is that a Chinese-speaking interviewer would pass questions written in Chinese into the room, and the corresponding answers would come out of the room in Chinese. Searle maintains that if such a system could indeed pass a Turing test, the person who manipulated the symbols would obviously not understand Chinese any better than he did before entering the room.

Searle proceeds systematically to refute the claims of strong AI by positioning himself as the one who manipulates the Chinese symbols. The first claim is that a system able to pass the Turing test understands the input and output. Searle replies that as the "computer" in the Chinese room, he gains no understanding of Chinese by simply manipulating the symbols according to the formal program (the complex translation rules). The operator in the room need not have any understanding of what the interviewer is asking, or of the replies that he is producing. He may not even know that there is a question-and-answer session going on outside the room.

The second claim of strong AI to which Searle objects is the claim that the system explains human understanding. Searle asserts that since the system is functioning–in this case passing the Turing Test–and yet there is no understanding on the part of the operator, then the system does not understand and therefore could not explain human understanding.

chiral

Having different left-hand and right-hand forms; not mirror symmetric. For example, the *snub cube* (one of the **Archimedean solid**s) is chiral, where as the ordinary **cube** is not.

Chladni, Ernst Florens Friedrich (1756–1827)

A German lawyer, musician (he was born in Leipzig in the same year as Mozart and died in the same year as Beethoven), and amateur scientist who founded the science of acoustics. While investigating musical tones, he had the inspired idea of making the sounds visible in a solid material. He spread fine sand over a glass or metal plate and set it into vibration with the bow of a violin by scraping the bow along one edge of the plate. The bow alternately stuck and slipped in rapid succession on the edge of the plate creating waves that moved across the plate and were reflected from the edges. These reflected waves became superimposed on the new waves coming from the bow edge, resulting in symmetrical patterns of nodal lines where the plate wasn't moving. The type of pattern produced on a Chladni plate depends on a variety of factors, including the point or points of support and their location; the point where the bow touches the plate; the frequency of the vibration, which is influenced by the speed the bow; and the shape and other properties of the plate itself.

chord

A straight line that joins two points on a curve. Most commonly, *chord* is used to mean a straight line segment joining, and included between, two points on a **circle**. In this more restricted sense it first appears in English in 1551 in Robert **Recorde**'s *The Pathwaie to Knowledge:* "*Defin.,* If the line goe crosse the circle, and passe beside the centre, then is it called a corde, or a stryngline."

Some surprising results emerge from moving chords. For example, take a chord in a circle C, and slide the chord around the circle so that the midpoint of the chord traces out a smaller concentric circle. Call the area between the two circles $A(C)$. Now do the same thing with a larger circle C' but with the same length chord. Is $A(C')$ larger or smaller than $A(C)$? Surprisingly, they are the same. In other words $A(C)$ doesn't depend on what circle you start with, only the length of the chord. An even more amazing fact is that if you slide a chord of fixed length around *any convex shape C* so that the chord's midpoint traces out another figure D, the area between C and D doesn't depend on what shape you started with.

chromatic number

(1) In **graph theory**, the minimum number of colors needed to color (the vertices of) a **connected graph** so that no two adjacent vertices are colored the same. In the case of simple graphs, this so-called *coloring problem* can be solved by inspection. In general, however, finding the chromatic number of a large graph (and, similarly, an optimal coloring) is an **NP-hard problem**. (2) In **topology**, the maximum number of regions that can be drawn

on a surface in such a way that each region has a border in common with every other region. If each region is given a different color, each color will border on every other color. The chromatic number of a square, tube, or sphere, for example, is 4; in other words, it is impossible to place more than four differently colored regions on one of these figures so that any pair has a common boundary. "Chromatic number" also indicates the least number of colors needed to color any finite map on a given surface. Again, this is 4 in the case of the plane, tube, and sphere, as was proved quite recently in the solution to the **four-color map problem**. The chromatic number, in both senses just described, is 7 for the **torus**, 6 for the **Möbius band**, and 2 for the **Klein bottle**. See also **Betti number**.

chronogram

A phrase or sentence in which certain letters represent, cryptically, a date, epoch, or, in rare cases, a non-date number. For example, the chronogram "My Day Is Closed In Immortality" commemorates the death of Queen Elizabeth the First of England: the capital letters can be rearranged to give MDCIII, or 1603, the year in which she died.

Church, Alonzo (1903–1995)

An American logician and professor at Princeton University who was an early pioneer of theoretical computer science. He is best known for his development, in 1934, of the so-called **lambda calculus**, a model of computation, and his discovery, in 1936, of an "undecidable problem" within it. This result preceded Alan **Turing**'s famous work on the **halting problem**, which also pointed out the existence of a problem unsolvable by mechanical means. Church and Turing then showed that the lambda calculus and the **Turing machine**, which is used in the halting problem, are equivalent in capability. They also demonstrated a variety of alternative "mechanical processes for computation" with equivalent computational abilities. See also **Church-Turing thesis**.

Church-Turing thesis

A logical/mathematical postulate, independently arrived at by Alan **Turing** and Alonzo **Church**, which asserts that as long as a procedure is sufficiently clear-cut and mechanical, there is some **algorithm**ic way of solving it (such as via computation on a **Turing machine**). Thus, there are some processes or problems that are computable according to some set of algorithms, and other processes or problems that are not. A strong form of the Church-Turing thesis claims that all neural and psychological processes can be simulated as computational processes on a computer.

cipher

(1) A cryptographic system (see **cryptography**) in which units of plain text of regular length, usually letters, are arbitrarily transposed (see **transposition cipher**) or substituted (see **substitution cipher**) according to a predetermined code, or a message written or transmitted in such a system. See also **Caesar cipher** and **Beal cipher**. (2) The mathematical symbol (0) for **zero**.

circle

The set of all points in a plane at a given distance, called the *radius,* from a fixed point, called the *center.* A circle is a simple **closed** curve that divides the plane into an interior and exterior. It has a perimeter, called a *circumference,* of length $2\pi r$ and encloses an area of πr^2. In coordinate geometry a circle with center (x_0, y_0) and radius r is the set of all points (x, y) such that:

$$(x - x_0)^2 + (y - y_0)^2 = r^2$$

"Circle" comes from the Latin *circus,* which refers to a large round or rounded oblong enclosure in which the famous Roman chariot races were held.

A line cutting a circle in two places is called a *secant.* The *segment* of a secant bound by the circle is called a *chord,* and the longest chord is that which passes through the center and is known as a *diameter.* The ratio of the circumference to the diameter of a circle is **pi**. The length of a circle between two radii is called an *arc;* the ratio between the length of an arc and the radius defines the *angle* between two radii in *radians.* The area bounded by two radii and an arc is known as a *sector.* A line touching a circle in one place is called a *tangent.* Tangent lines are perpendicular to radii. In **affine geometry** all circles and ellipses become **congruent**, and in **projective geometry** the other **conic section**s join them. A circle is a conic section with eccentricity zero. In **topology** all simple closed curves are **homeomorphic** to circles, and the word *circle* is often applied to them as a result. The three-dimensional analog of the circle is the **sphere,** and the four-dimensional analog is the **hypersphere**.

circle involute

The simplest kind of **spiral** to draw and understand. It is the path that a goat, tethered to a post, would follow if it walked around and around in the same direction, keeping its tether taught until it wound its way to the center. The radial distance between adjacent loops of the spiral is equal to the circumference of the central circle. Except for the innermost loop, the circle involute is hard to distinguish from the **Archimedean spiral**, though the two curves are never identical.

circular cone

A **cone** whose base is a **circle**.

circular helix

See **helix**.

circular prime

A **prime number** that remains prime on any cyclic rotation of its digits. An example (in the decimal system) is 1,193 because 1,931, 9,311, and 3,119 are also prime. Any one-digit prime is circular by default. In base ten, any circular prime with two or more digits can only contain the digits 1, 3, 7, and 9; otherwise when 0, 2, 4, 5, 6, or 8 is rotated into the units place, the result can be divided by 2 or 5. The only circular primes known, listing just the smallest representative from each cycle, are: 2, 3, 5, 7, 11, 13, 17, 37, 79, 113, 197, 199, 337, 1,193, 3,779, 11,939, 19,937, 193,939, 199,933, R_{19}, R_{23}, R_{317}, R_{1031} and possibly R_{49081}. These last five are the known **rep-unit** primes and probable primes. It's generally believed that there are infinitely many rep-unit primes, so there should be infinitely many circular primes. But it's very likely that all circular primes not on the list above are rep-units.

circumcenter

The center of a **circle** that passes through the vertices (see **vertex**) of a given **polygon**, usually a triangle. For a triangle, it is the same as the point of intersection of the perpendicular **bisect**ors of the three sides.

circumcircle

The **circle** that passes through all three vertices (see **vertex**) of a given triangle. It is said to *circumscribe* the triangle.

circumference

The distance around the outside of a **circle**. The word comes the Latin *circus* ("circle") and *ferre* ("to carry"), thus means "to carry around."

cissoid

Given a fixed point A and two curves C and D, the cissoid of the two curves with respect to A is constructed as follows: pick a point P on C, and draw a line l through P and A. This cuts D at Q. Let R be the point on l such that $AP = QR$. The **locus** of R as P moves on C is the cissoid. The name *cissoid,* meaning "ivy-shaped," first appears in the work of Geminus in the first century B.C.

A special case of this curve, now known as the *cissoid of Diocles,* was first explored by **Diocles** in his attempt to solve the classical problem of **duplicating the cube**. Later investigators of the same curve include Pierre de **Fermat**, Christiaan **Huygens**, John **Wallis**, and Isaac **Newton**. The cissoid of Diocles is traced out by the **vertex** of a **parabola** as it rolls, without slipping, on a second parabola of the same size. It has the Cartesian equation

$$y^2 = x^3/(2a - x).$$

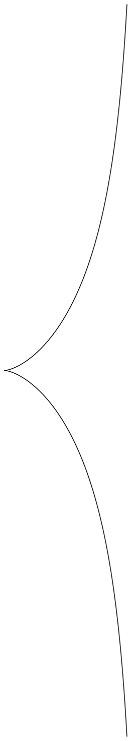

cissoid The cissoid of Diocles. *© Jan Wassenaar, www.2dcurves.com*

Interestingly, Diocles investigated the properties of the focal point of a parabola in *On Burning Mirrors* (a similar title appears in the works of Archimedes). The problem, then as now, is to find a mirror surface such that when it is placed facing the Sun, it focuses the maximum amount of heat.

classification

The goal in a branch of mathematics of providing an exhaustive list of some type of mathematical object with no repetitions. For example, the classification of 3-**manifold**s is one of the outstanding problems in **topology**. With the advent of computers, one weak but precise way to state a classification problem is to ask whether there is an **algorithm** to determine whether two given objects are equivalent.

clelia

Also known as a *clelie curve,* the **locus** of a point P that moves on the surface of a **sphere** in such a way that ϕ/θ is constant, where ϕ and θ are the longitude and colatitude (the angular distance from a pole).

Clifford, William Kingdon (1845–1879)

An English mathematician who studied **non-Euclidean geometry** and **topology**. In 1870, he wrote *On the Space Theory of Matter* in which he argued that energy and matter are simply different types of curvature of space–a remarkably advanced idea that would come to fruition in Einstein's general **relativity theory**. Although small of build Clifford was remarkably strong and able to do one-armed chin-ups. His death at an early age was the result of overwork and exhaustion.

clock puzzles

The earliest known clock problem was posed in 1694 by Jacques **Ozanam** in his *Récréations mathématiques et physiques*.

PUZZLES

Here are two clock puzzles invented by Lewis **Carroll**:

1. A clock has hour and minute hands of the same length and no numerals on its face. At what time between 6 and 7 o'clock will the time on the clock appear to be the same as the time read on the reflection of the clock in a mirror?
2. Which has a better chance of giving the right time: a clock that has stopped or one that loses a minute every day?

And here is another from Henry **Dudeney**'s *Amusements in Mathematics* called "The Club Clock":

3. One of the big clocks in the Cogitators' Club was found the other night to have stopped just when the second hand was exactly midway between the other two hands. One of the members proposed to some of his friends that they should tell him the exact time when (if the clock had not stopped) the second hand would next again have been midway between the minute hand and the hour hand. Can you find the correct time that it would happen?

Solutions begin on page 369.

closed

A *closed curve* is one that has no endpoints so that it completely encloses a certain area. A *closed interval*, which corresponds to a *closed set*, is an interval that includes its endpoints.

cochleoid

A **spiral** curve that was first studied by J. Peck in 1700 and Bernoulli in 1726. Its name, meaning "snail-form" (*kochlias* is Greek for "snail"), was coined by Benthan and Falkenburg in 1884. It can be constructed starting from a point O on the y-axis. For all circles through O (tangent to the y-axis), pace a constant distance on the circle. The collection of those points is the cochleoid. In Cartesian coordinates, it is given by the formula

$$(x^2 + y^2)\tan^{-1}(y/x) = ay$$

and in polar coordinates by

$$r = a\sin\theta / \theta.$$

The points of contact of parallel tangents to the cochleoid lie on a **strophoid**.

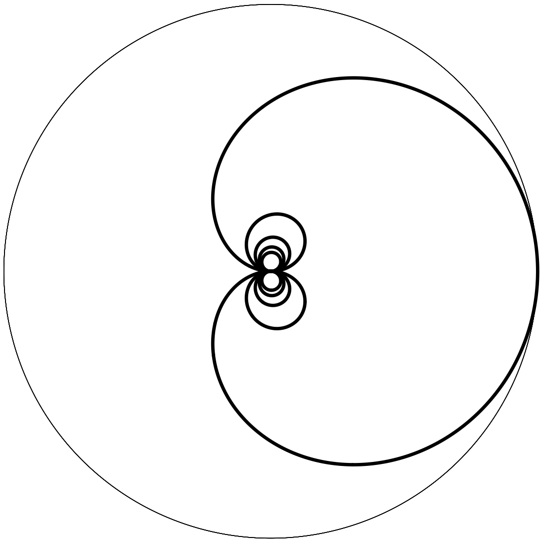

cochleoid A cochleoid inside the circle used to construct it.
© Jan Wassenaar, www.2dcurves.com

code

See **cipher**.

codimension

In general, if a mathematical object sits inside or is associated with another object of **dimension** n, then it is said to have codimension k if it has dimension $n - k$.

coding theory

The branch of mathematics concerned with sending data across noisy channels and recovering the message. Whereas **cryptography** is about making messages *hard* to read, coding theory focuses on making messages *easy* to read. The basic problem is that messages, in the form of binary digits or *bits* (strings of 0 or 1) have to be sent along a channel (such as a phone line) in which errors occur randomly, but at a predictable overall rate. To compensate for the errors, more bits have to be sent than are contained in the original message. The easiest way to detect errors in binary data is the **parity** code, which inserts an extra parity bit after every 7 bits from the source message. To correct as well as detect errors, the data has to be retransmitted. A simple way to do this is to repeat each bit a set number of times. The recipient sees which value, 0 or 1, occurs more often and assumes that to be the intended bit. This method can cope with error rates up to 1 error in every 2 bits transmitted but it means that an awful lot of extra bits have to be sent.

In 1948, Claude **Shannon** at Bell Labs began the subject of coding theory by proving the minimum number of extra bits that had to be transmitted to encode messages but without showing ways to find these optimal codes. Two years later, Richard Hamming, also at Bell Labs, gave details of error-correcting codes with information transmission rates more efficient than simple repetition. His first code, in which four data bits were followed by three check bits, allowed not only the detection but the correction of a single error.

While Shannon and Hamming were involved with information transmission in the United States, John Leech devised similar codes while working on **group** theory at Cambridge University. This research also took in the sphere **packing** problem and culminated in the amazing, 24-dimensional *Leech lattice,* the study of which proved crucial to understanding and classifying finite symmetry groups. The value of error-correcting codes for information transmission, both on Earth and from space, was immediately grasped, and a variety of codes were constructed that boosted both economy of transmission and error-correction capacity. Between 1969 and 1973 the NASA Mariner probes used a powerful *Reed-Muller code* capable of correcting 7 errors out of 32 bits transmitted. A less obvious application of error-correcting codes came with the development of the compact disk on which the signal is encoded digitally. To guard against scratches and other damage, two interleaved codes that can correct up to 4,000 consecutive errors are used. By the late 1990s the goal of finding explicit codes that reach the limits predicted by Shannon's original work had been achieved.

codomain

For a given **function** or mapping, a **set** within which the values of the function lie. This is different from the set of values, known as the **range**, that the function actually takes.

coefficient

A number or other factor that multiplies a variable. For example, in the equation $3x - 4ky = 8$, the 3 and $4k$ are coefficients of the variables x and y. The word combines three elements, the Latin *facere* ("to do"), and the prefixes *ex* ("out") and *co* ("with"), to give the overall meaning of joining two things together to bring about a result. The sixteenth-century mathematician Francois Vieta may have coined the word, but it was not commonly used until around the beginning of the eighteenth century.

Coffin, Stewart T.

A leading designer of mechanical puzzles. He is also the author of *The Puzzling World of Polyhedral Dissections,*[64] one of the most significant works produced on this subject.

cohomology

A subject that involves calculating algebraic **invariant**s of **topological space**s that are formally dual to **homology**. The invariants obtained are in general more powerful than those given by homology and usually have more algebraic structure. *Generalized cohomology theories,* both for topological spaces and for purely algebraic structures, have been developed that have some of the formal properties of cohomology but which don't have the same geometric background.

coin paradox

Consider two round coins of equal size. Imagine holding one still and then rolling the other coin around it, making sure that it doesn't slip and that the rims are touching at all times. How many times will the moving coin have rotated after it has completed one revolution of the stationary coin? Most people believe that the answer will be once and are therefore surprised to discover that the truth is in fact *twice.*

coincidence

What an amazing coincidence! Well, not really. Coincidences are bound to happen. In a world where there are a great many potential coincidences each with a small probability of happening, someone, somewhere is going to see one–and be amazed by it. The fact that there are countless numbers of noncoincidences and many people who don't see a significant coincidence in the same period of time is overlooked. Also, we tend to underestimate the probabilities of coincidences in certain situations and are therefore more surprised than we should be when coincidences happen. A classic example of this is the **birthday paradox**.

Obviously some things are extraordinarily unlikely. What are the chances, for example, of a meteorite hitting your car? Next to nothing, but not quite nothing. There are a lot of cars and there are dozens of meteorites that strike Earth every day. Sooner or later, it's bound to happen. In fact, it did happen to Michelle Knapp's Chevy Malibu parked outside her home in Peekskill, New York, on the evening of October 9, 1992. A 12-kilogram space rock smashed through the car's trunk and ended up on the driveway below.

Does coincidence completely explain away all events that might otherwise be put down to precognition? On April 15, 1912, the SS *Titanic* sunk on her maiden voyage, having been holed by an iceberg, and over 1,500 people died. Fourteen years earlier a novel had been published by Morgan Robertson that seemed to foretell the disaster. The book described a ship the same size as *Titanic* that struck an iceberg on its maiden voyage on a misty April night. The name of Robertson's fictional ship was the *Titan.* Mere happenstance or evidence of something deeper? Numerologists often spot matchups that would go unnoticed by the rest of us. Is it so strange that there are almost exactly 500 million inches in the pole-to-pole diameter of Earth? Not if you work in centimeters. And should we make such a fuss over the fact that the speed of light is within 0.1% of 300,000 kilometers per second when we give no attention to the miles-per-second value of 186,282? Yet, surely, there can be no doubt that Shakespeare wrote the Bible. The King James Version was published in 1611, when Shakespeare was 46 years old. Look up Psalm 46. Count 46 words from the beginning of the Psalm. You will find the word "Shake." Count 46 words from the end of the Psalm. You will find the word "Spear." To some, an obvious coded message. See also **thirteen**.

Collatz problem

A problem first posed by the German mathematician Lothar Collatz (1910–1990) in 1937, that is also known variously as the *$3n + 1$ problem, Kakutani's problem,* the *Syracuse problem, Thwaites' conjecture,* and *Ulam's conjecture.* It runs as follows. Let n be any integer. (1) If n is odd, put n equal to $3n + 1$; otherwise, put n equal to $n/2$. (2) If $n = 1$, stop; otherwise go back to step 1. Does this process always terminate (i.e., end in 1) for any value of n? To date, this question remains unanswered, though the process has been found to stop for all n up to 5.6×10^{13}. British mathematician Bryan Thwaites (1996) has offered a £1,000 reward for a resolution of the problem. However, John **Conway** has shown that Collatz-type problems can be formally undecidable, so it not known if a solution is even possible. The members of sequences produced by the Collatz problem are sometimes known as **hailstone sequences**.[144]

combination

A set of objects selected without reference to the order in which they are arranged. Compare with **permutation**. See also **binomial coefficient**.

combinatorics

The study of the ways of choosing and arranging objects from given collections and the study of other kinds of problems relating to counting the number of ways to do something.

commensurable

Two lines or distances are commensurable if the ratio of their lengths is a **rational number**. If the ratio is an **irrational number**, they are called *incommensurable.*

common fraction

A **fraction** that consists of the **quotient** of two **integer**s.

communication theory

See **information theory**.

commutative

Two numbers, x and z, are said to be *commutative under addition* if

$$x + y = y + x$$

and to be *commutative under multiplication* if

$$x \times y = y \times y.$$

In general, two elements a and b of a **set** S are commutative under the binary operation (an operation that works on two elements at a time) $*$ if

$$a * b = b * b.$$

Compare with **associative** and **distributive**.

complement

That which is needed to complete something. For instance, the complement of a number is what needs to be added to it to make a specified value; the complement of an angle is the angle required to turn it into a right angle. The complement of a **set** is composed of all the elements that are not members of that set.

complete

Describes a **formal system** in which all statements can be proved as being true or false. Most interesting formal systems are not complete, as demonstrated by **Gödel's incompleteness theorem**.

complete graph

A **connected graph** in which exactly one **edge** connects each pair of vertices (see **vertex**). A complete graph with n vertices, denoted K_n, has $n(n - 1)/2$ edges (i.e., the nth **triangular number**), $(n - 1)!$ **Hamilton circuit**s, and a **chromatic number** of n. Every vertex in K_n has **degree** $n - 1$; therefore K_n has an **Euler circuit** if and only if n is odd. In a *weighted complete graph,* each edge has a number called a *weight* attached to it. Each path then has a total weight, which is the sum of the weights of the edges in the path. See also **traveling salesman problem**.

complex adaptive system (CAS)

A nonlinear, interactive, **complex system** with the ability to adapt to a changing environment. CASs evolve by random mutation, **self-organization**, the transformation of their internal models of the environment, and natural selection. Examples include living organisms, the nervous system, the immune system, the economy, corporations, and societies. In a CAS, semiautonomous agents interact according to certain rules of interaction, evolving to maximize some measure like fitness. The agents are diverse in form and capability and they adapt by changing their rules and, hence, behavior, as they gain experience. CASs evolve historically–their experience determines their future trajectory. Their adaptability can either be increased or decreased by the rules shaping their interaction. Moreover, unanticipated, emergent structures can play a determining role in the evolution of such systems, which is why they are highly unpredictable. On the other hand, CASs have the potential of a great deal of creativity that was not programmed into them from the beginning.

complex analysis

The study of **function**s of a complex variable. Often, the most natural proofs for statements in real **analysis** or even **number theory** use techniques from complex analysis. Unlike real functions, which are commonly represented as two-dimensional graphs, complex functions have four-dimensional graphs and may usefully be illustrated by color-coding a three-dimensional graph to suggest four dimensions.

complex number

A **real number** plus a real number times the square root of –1; in other words, a number of the form $z = a + ib$, where a and b are real and $i = \sqrt{-1}$. The term ib is known as an **imaginary number** or the *imaginary part* of the complex number $a + ib$; a is called the real part. The names "complex," "real," and "imaginary," which came about historically, are totally misleading because complex numbers are not particularly complex and imaginary numbers are no less real than real numbers! Another way to represent a complex number is as an **ordered pair** of real numbers (a, b) together with the operations: $(a, b) + (c, d) = (a + c, b + d)$ and $(a, b) \times (c, d) = (ac - bd, bc + ad)$. Alternatively, complex numbers can be shown as points on an **Argand diagram** (a representation of the *complex plane*) in which the horizontal axis is the real number line and the vertical axis represents all possible purely imaginary numbers. Any point that appears on the complex plane off-axis has both real and imaginary parts. On an Argand diagram a complex number can also be shown as a **vector**, or directed line segment (a line of a certain length with an arrow), extending from the origin $(0 + 0i)$ to the number $(a + bi)$. The *absolute value* or *magnitude* of a complex number z, thought of as a point on a plane, is its Euclidean distance from the origin, and is denoted $|z|$; this is always a nonnegative real number. Algebraically, if $z = a + ib$, we can define $|z| = \sqrt{(a^2 + b^2)}$. If the complex number z is written in polar coordinates $z = r\,e^{i\varphi}$, then $|z| = r$.

Complex numbers are a natural extension of real numbers and form what is called an *algebraically closed field*. Because of this, mathematicians sometimes consider the complex numbers to be more "natural" than the real numbers: all **polynomial** equations have solutions among the complex numbers, which is not true for the real numbers. Complex numbers are used in electrical engineering and other branches of physics as a convenient description for periodically varying signals. In an expression $z = re^{i\varphi}$ one may think of r as the amplitude and ϕ as the phase of a sine wave of given frequency. In special and general **relativity theory**, some formulas for the metric on **space-time** become simpler if the time variable is taken to be imaginary.

complex plane

See **Argand diagram**.

complex system

A collection of many simple nonlinear units that operate in parallel and interact locally with each other so as to produce emergent (see **emergence**) behavior.

complexity

A phenomenon that has two distinct and almost opposite meanings. The first, and probably the oldest mathematically, goes back to Andrei **Kolmogorov**'s attempt to give an **algorithm**ic foundation to notions of **random**ness and probability and to Claude **Shannon**'s study of communication channels via his notion of information. In both cases, complexity is synonymous with *disorder* and a lack of structure. The more random a process, the greater its complexity. An ideal gas, for example, with its numerous molecules bouncing around in complete disarray, is complex as far as Kolmogorov and Shannon are concerned. Thus, in this sense, complexity equates to the degree of complication.

The second, and more recent notion of complexity refers instead to how structured, intricate, hierarchical, and sophisticated a natural process is. In particular, it's a property associated with **dynamical system**s in which new, unpredictable behavior arises on scales above the level of the constituent components. The distinction between these two meanings can be revealed by answering a simple question about a system: Is it complex or is it merely complicated? Measures of complexity include **algorithmic complexity**, **fractal dimension**ality; **Lyapunov fractal**s, and **logical depth**.

complexity theory

A part of the theory of computation that has to do with the resources needed to solve a given problem. The most common resources are *time* (how many steps it takes to solve a problem) and *space* (how much memory it takes to solve a problem). Complexity theory differs from **computability theory**, which deals with whether a problem can be solved at all, regardless of the resources required.

composite number

A positive integer that can be factored into smaller positive integers, neither of which is one. If a positive integer is not composite (4, 6, 8, 9, 10, 12, . . .) or one, then it is a **prime number** (2, 3, 5, 7, 11, 13, 17, . . .). As Karl **Gauss** put it in his *Disquisitiones Arithmeticae* (1801): "The problem of distinguishing prime numbers from composite numbers and of resolving the latter into their prime factors is known to be one of the most important and useful in arithmetic." One reason for its importance today is that many secret codes and much of the security of the Internet depends in part on the relative difficulty of factoring large numbers. But more basic to a mathematician is that this problem has always been central to **number theory**. Numbers that, for their size, have a lot of factors are sometimes referred to as *highly composite numbers*. Examples include 12, 24, 36, 48, 60, and 120.

compound polyhedron

An assemblage of two or more polyhedra, usually interpenetrating and having a common center. There are two types: a combination of a solid with its **dual** and an interpenetrating set of several copies of the same **polyhedron**. The simplest example of a compound polyhedron is the compound of two tetrahedra, known as the *stella octangula* and first described by Johannes **Kepler**. This shape is unique in that it falls under both of the above classes, because the **tetrahedron** is the only self-dual **uniform polyhedron**; the edges of the two tetrahedra form the diagonals of the faces of a **cube** in which the stella octangula can be inscribed.

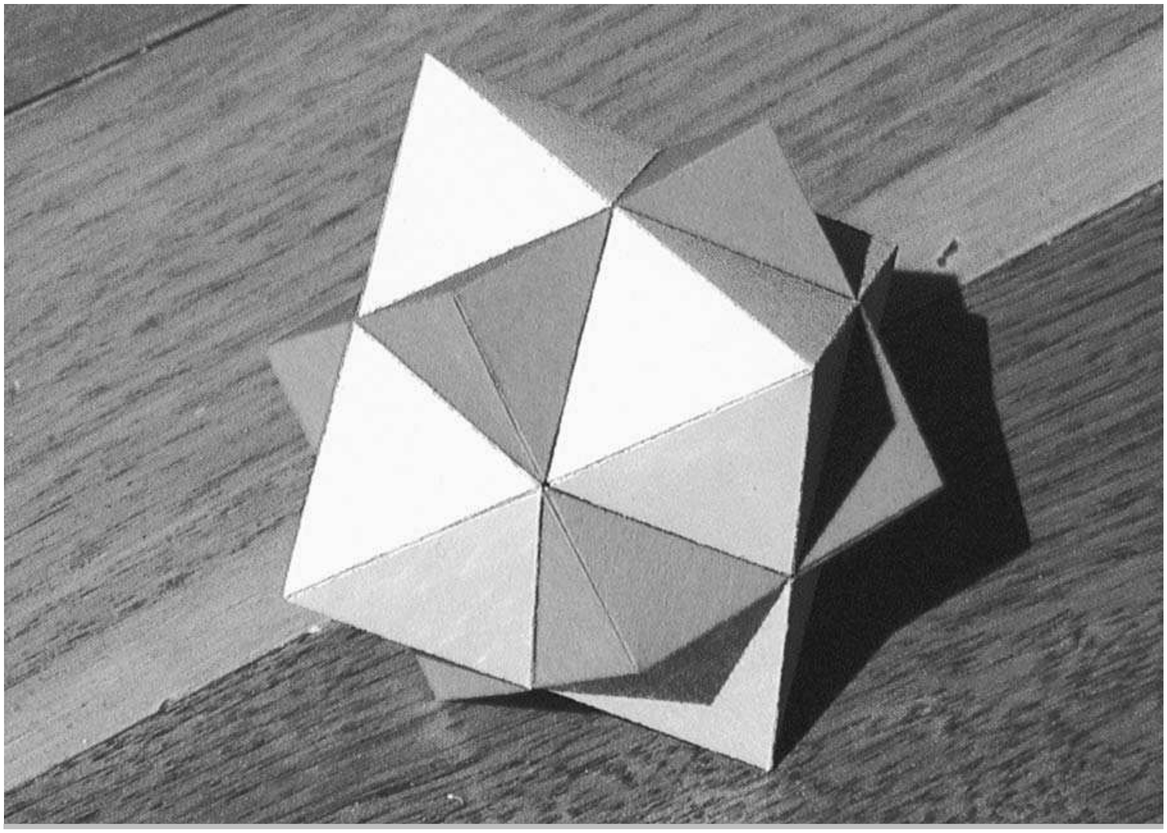

compound polyhedron A compound of duals: the cube and the octahedron. *Robert Webb, www.software3d.com; created using Webb's Stella program*

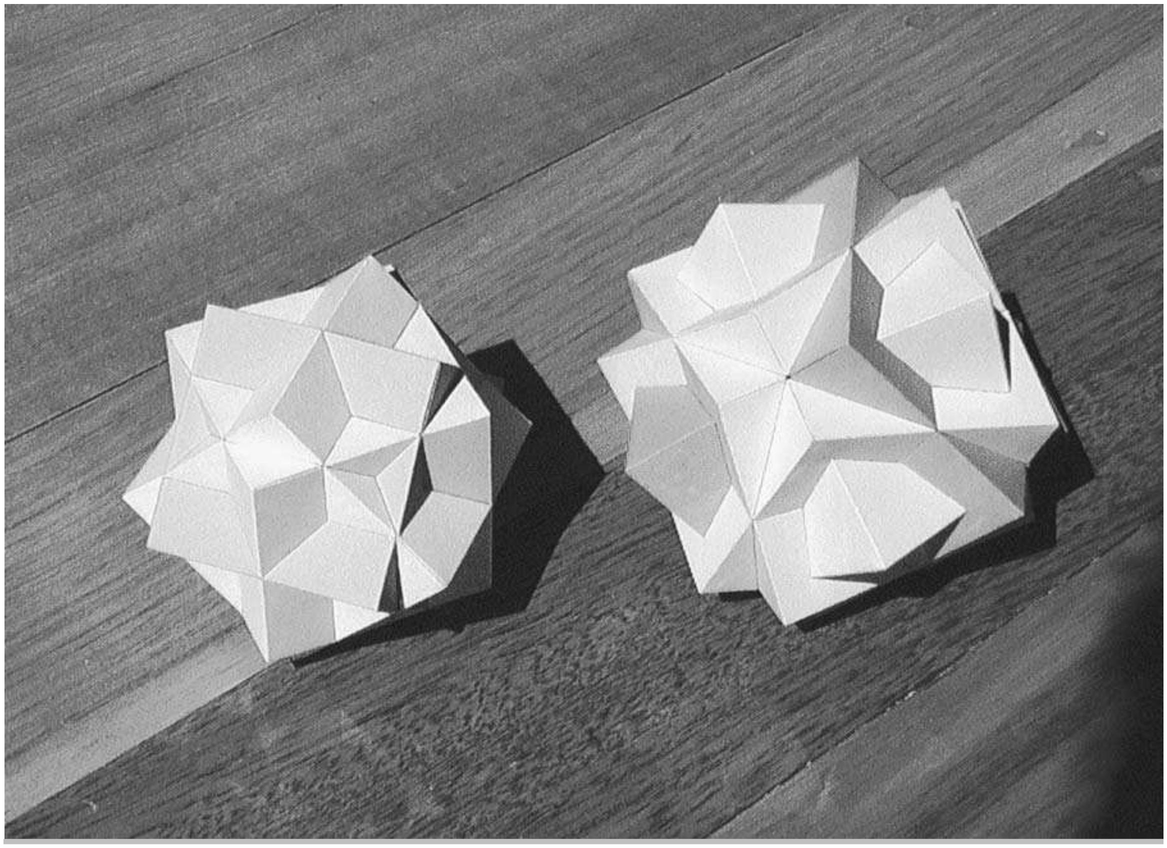

compound polyhedron A compound polyhedron of three cubes (left), such as that used in Escher's picture "Waterfall." The compound of four cubes (right) is also known as Bakos's compound. *Robert Webb, www.software3d.com; created using Webb's Stella program*

Another example of a compound follows from an important Platonic relationship: a cube can be inscribed within a **dodecahedron**. There are five different positions for a cube within a dodecahedron; superimposing all five gives the compound known as the *rhombic triacontahedron.*

compressible

Having a description that is smaller than itself; not **random**; possessing regularity.

computability theory

The part of the theory of computation that deals with problems that are solvable by **algorithms** or–what amounts to the same thing–by **Turing machine**s. Computability theory is concerned with four main questions: What problems can Turing machines solve? What other systems are equivalent to Turing machines? What problems require more powerful machines? What problems can be solved by less powerful machines? Not all problems can be solved computationally. An *undecidable problem* is one that can't be solved by any algorithm, no matter how much time, processing speed, or memory is available. Many examples are known, one of the most famous of which is the **Halting problem**. See also **cellular automaton**.

computable number

A **real number** for which there is an **algorithm** that, given *n,* calculates the *n*th digit. Alan **Turing** was the first to define a computable number and the first to prove that almost all numbers are uncomputable. An example of a number that, even though well-defined, is uncomputable is **Chaitin's constant**.

concave

Curved inward, like the inner surface of a sphere; the word comes from the Latin *concavus* for "hollow." A figure, such as a **polygon** or **polyhedron**, is said to be concave if a line segment joining any points inside the figure goes outside the figure. Similarly, a **set** is concave if it doesn't contain all the line segments connecting any pair of its points.

conchoid

A shell-shaped curve. Given a point A and a curve C, if we pick a point Q on C and draw a line L through A and Q and mark points P and P' on L at some fixed distance in either direction from Q, then the locus of P and P' as Q moves on C is a conchoid. The *conchoid of Nichomedes* is a conchoid in which the given line is a straight line; that is, given a line C and a point A we pick a point Q on C, draw a line L through A and Q, and mark P and P' on L at some fixed distance from Q. The conchoid of Nicomedes is the locus of P and P' as Q moves along C. It has the polar equation $R = a \sec\theta + k$. The *conchoid of de Sluze* is the curve with the Cartesian equation $a(x - a)(x^2 + y^2) = k^2x^2$.

cone

A shape (its name comes from the Greek *konos* for pinecone) that has a circular or elliptical base and a **vertex**, also known as an *apex,* lying outside the plane of the base and that is formed from all the line segments joining points on the edge of the base to the vertex. If the base is a circle, the shape is a *circular cone;* if the line, or *axis,* from the center of the base to the vertex is perpendicular to the base, then it is a *right cone* (an ice-cream cone is a right circular cone); otherwise it's an *oblique cone.* The curved lateral surface of the cone is called a *nappe.* If the cone is extended in both directions from the vertex, the result is a *double cone* or *bicone.* A section through a double cone that has been extended indefinitely in both directions to form a *conic surface* is known as a **conic section**. Another way to think of a cone is as a **surface of revolution** generated by a line that rotates around a fixed point, at a fixed angle from another line (the axis), both lines passing through that fixed point. The volume of a cone, of perpendicular height h and circular base of radius r, is $\frac{1}{3}\pi r^2h$.

Take a solid cylinder of radius r and height $2r$. Remove the right double cone that passes through the center of the cylinder and extends to meet the circular disks on the cylinder's top and bottom. Interestingly, the volume of the remaining object and the volume of a sphere of radius r are the same.

PUZZLE

The *Cone Puzzle* (no. 202) from Henry **Dudeney**'s *Amusements in Mathematics*[88] runs as follows: "I have a wooden cone. How am I to cut out of it the greatest possible cylinder?"

Solutions begin on page 369.

conformal mapping

A map from the plane to itself that preserves angles. Conformal mapping results in the angle between any two curves being the same as the angle between their images. The Mercator map is a conformal map of Earth's surface.

congruent

In the case of geometric figures, having exactly the same shape and size.

congruum problem

Find a **square** number x^2 such that, when a given number h is added or subtracted, new square numbers are obtained, so that $x^2 + h = a^2$ and $x^2 - h = b^2$. This problem was posed by the mathematicians Théodore and Jean de Palerma in a mathematical tournament organized by Frederick II in Pisa in 1225. The solution is $x = m^2 + n^2$ and $h = 4mn(m^2 - n^2)$, where m and n are integers.

conic section

An important, familiar, and ubiquitous family of curves obtained by slicing a right circular double **cone**, extended indefinitely in both directions, with a plane. Depending on the angle of the slice to the axis of the cone, the resulting curve may be a **circle**, an **ellipse**, a **parabola**, or a **hyperbola**. The circle is a limiting case of the ellipse, when the slice is made at right angles to the axis, while the parabola is the limiting case of both the ellipse and the hyperbola, when the slice is made parallel to the side of the cone. The name *conic sections* comes from the eight-volume work *Conics* (Κωνικα) by **Apollonius**, who also gave us the names *ellipse, parabola,* and *hyperbola.*

Another geometric way to define the conics is as the **locus** of all points in the plane whose distances, r, from a fixed point called the *focus,* and a, from a given straight line called the **directrix**, have a constant ratio. This ratio, r/a, is known as the *eccentricity, e.* The circle has an eccentricity of zero. As the eccentricity increases from near zero, corresponding to a nearly circular ellipse, the ellipse stretches until the right-hand side of it disappears to infinity, e becomes 1, and the ellipse turns into a parabola, with just one open branch. Like the circle, the parabola has only one shape, though it may look different depending on how much it is enlarged or diminished. As the eccentricity increases beyond 1, the "lost" right-hand end of the ellipse reappears from the other side of infinity, so to speak, and turns into the left-hand branch of a hyperbola.

Because a hyperbola is effectively an ellipse split in two by infinity, it comes as no surprise that these curves are related in an inverse way. An ellipse consists of all points whose distances from two foci have a constant sum, while a hyperbola is made from all points whose distances from two foci have a constant difference. These definitions also apply to the circle and the parabola, if the two foci are considered to coincide in the case of the circle and to be separated by an infinite distance in the case of the parabola.

In terms of algebra, the family of conics represents all the possible **real number** solutions to the general **quadratic** equation $ax^2 + bxy + cy^2 + dx + ey + f = 0$. In other words, the graph of any quadratic with real solutions is always a conic section. The key quantity is the difference $b^2 - 4ac$. If this is less than zero, the graph is an ellipse, a circle, a point, or no curve. If $b^2 - 4ac = 0$, the graph is a parabola, two parallel lines, one line, or no curve; if it is greater than zero, the graph is a hyperbola or two intersecting lines.

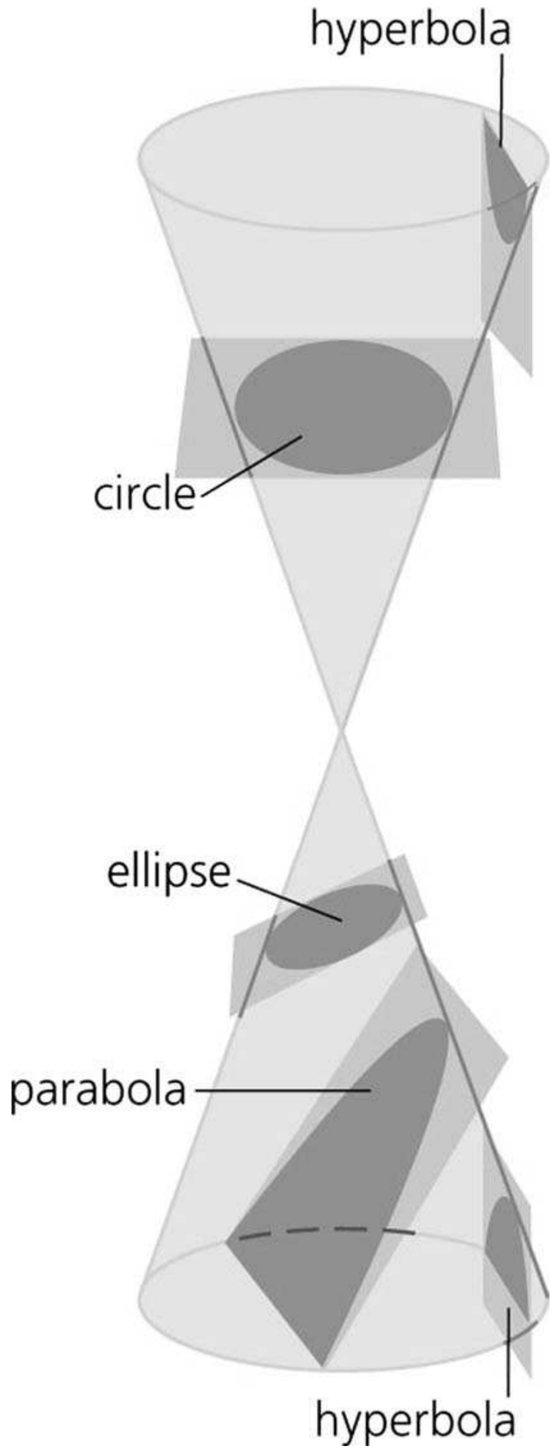

conic section The circle, ellipse, parabola, and hyperbola, obtained by slicing a right double cone in various ways.

conical helix

See **helix**.

conjecture

A mathematical statement that has been put forward as a true statement, but that no one has yet been able to prove or disprove; in mathematics, a conjecture and a hypothesis are essentially the same thing. When a conjecture has been proven to be true, it becomes known as a **theorem**. Famous conjectures include the **Riemann hypothesis**, the **Poincaré conjecture**, the **Goldbach conjecture**, and the **twin primes** conjecture. Just to show how terminology can be used inconsistently, however, the most famous of all conjectures, for centuries before its proof in 1995, was always known as **Fermat's last theorem**!

conjugate

(1) *Conjugate angles* add up to 360°. (2) The *complex conjugate* of a **complex number** $a + bi$ is $a - bi$. (3) *Conjugate lines* of a **conic section** have the property that each contains the pole point of the other, while *conjugate points* of a conic have the property that each lies on the polar line of the other. In general, conjugate indicates that there is a symmetrical relationship between two objects A and B; in other words, there is an operation that will turn A into B and B into A.

connected

A space S is said to be connected if any two points in S can be connected by a curve lying wholly within S. Two spaces can be added by what is called a *connected sum*. Roughly speaking, this involves pulling out a disk from each surface, creating **hole**s, and then sewing the two surfaces together along the boundaries of the holes. In this way, a one-holed **torus** can be added to a two-holed torus to give a three-holed torus; alternatively, a **projective**

plane can be added to a projective plane to give a **Klein bottle**. The operation is **commutative** and **associative** and there is even an **identity** element: for example, adding a sphere to any surface simply returns the same surface. See also **simply connected**.

connected graph

A **graph** in which a path exists between all pairs of vertices (see **vertex**). If the graph is also a **directed graph**, and there exists a path from each vertex to every other vertex, then it is a *strongly connected graph.* If a connected graph is such that exactly one edge connects each pair of vertices, then it is said to be a **complete graph**. See also **Euler path** and **Hamilton path**.

connectionism

A computational approach to modeling the brain that relies on the interconnection of many simple units to produce complex behavior.

connectivity

The amount of interaction in a system, the structure of the weights in a **neural network**, or the relative number of edges in a **graph**.

consistency

An **axiom**atic theory is said to be consistent if it's impossible (within the confines of the theory) to prove simultaneously a statement and its negation. **Godel's incompleteness theorem** states that any (sufficiently powerful) consistent axiomatic theory is incomplete.

constructible

In classical geometry, a figure or length that can be drawn using only an unmarked straightedge and a compass. The Greeks were adept at constructing **polygon**s, but the question of proving which **regular polygons** are constructible and which are not had to wait for the genius of Carl **Gauss**. At the age of only 19, Gauss found that a regular polygon with *n* sides is constructible if and only if *n* is a prime **Fermat number**. The only known such primes are 3, 5, 17, 257, 65,537. It is also possible to construct certain numbers, known as *constructible numbers,* that correspond to line segments, including **rational number**s and some **irrational number**s, but no **transcendental number**s. It turns out that all constructions possible with a compass and straightedge can be done with a compass alone, as long as a line is considered constructed when its two endpoints are located. The reverse is also true, since Jakob **Steiner** showed that all constructions possible with straightedge and compass can be done using only a straightedge, as long as a fixed circle and its center (or two intersecting circles without their centers, or three nonintersecting circles) have been drawn beforehand. Such a construction is known as a *Steiner construction.* The Greeks were unable to achieve certain constructions, such as **squaring the circle**, **duplicating the cube**, and **trisecting an angle**, despite numerous attempts, but it wasn't until hundreds of years later that the problems were proved to be actually impossible under the limitations imposed.

continued fraction

A representation of a **real number** in the form

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ldots}}}$$

which, mercifully for typesetters, can be written in compact notation as

$$x = [a_0;\ a_1,\ a_2,\ a_3,\ \ldots],$$

where the integers a_i are called *partial quotients.* Although rarely encountered in school and even college math courses, continued fractions (CFs) provide one of the most powerful and revealing forms of numerical expression. Numbers whose decimal expansions look unremarkable turn out, when unfolded as CFs, to have extraordinary symmetries and patterns. CFs also offer a way of constructing rational approximations to **irrational numbers** and of discovering the most irrational numbers.

CFs first appeared in the sixth century in the works of the Indian mathematician Aryabhata, who used them to solve linear equations. They surfaced in Europe in the fifteenth and sixteenth centuries and **Fibonacci** attempted to define them in a general way. The term "continued fraction" first appeared in 1653 in an edition of *Arithmetica Infinitorum* by John **Wallis**. Their properties were also studied by one of Wallis's English contemporaries, William Brouncker, who, along with Wallis, was one of the founders of the Royal Society. At about the same time, in Holland, Christiaan **Huygens** made practical use of CFs in his designs of scientific instruments. Later, in the eighteenth and early nineteenth centuries, Carl **Gauss** and Leonhard **Euler** delved into many of their deeper properties.

CFs can be finite or infinite in length. Finite CFs can be evaluated level by level (starting at the bottom) and will always reduce to a rational fraction; for example, the

CF [1; 3, 2, 4] = 40/31. By contrast, infinitely long CFs produce representations of irrational numbers. Here are the leading terms from a few notable examples of infinite CFs:

$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, \ldots]$

$\sqrt{2} = [1; 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, \ldots]$

$\sqrt{2} = [1; 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, \ldots]$

$\pi = [3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, 84, 2, \ldots]$

Each of these expansions has a simple pattern except that for π (see **pi**), which has no obvious pattern at all. There's also a preference for the quotients to be small numbers.

If an infinite CF is truncated after a finite number of steps, the result is a *rational approximation* to the original irrational. In the case of π, chopping the CF at [3; 7] gives the familiar approximation for π of 22/7 = 3.1428571 . . . . Keeping two more terms leads to [3; 7, 15, 1] = 353/113 = 3.1415929 . . . , which is an even better approximation to the true value of π (3.14159265 . . .). The more terms retained in the CF, the better the rational approximation becomes. In fact, the CF gives yields the best possible rational approximations to a general irrational number. Notice also that if a large number occurs in the expansion of quotients, then truncating the CF after that will produce an especially good rational approximation. Most CF quotients are small numbers (1 or 2), so the appearance in the CF of π of a number as large as 292 so early in the expansion is unusual. It also leads to an extremely good rational approximation to π = [3; 7, 15, 1, 292] = 103,993/33,102.

continuity

A mathematical property that has to do with how smooth or "well-behaved" a **function** or curve is. If two adjacent points on a graph, for example, are not connected or are separated by a jump, this marks a breakdown of continuity. At such a discontinuity it is impossible to obtain a **derivative**, or slope, of the curve. Usually if a curve does misbehave like this, it is only at one or two isolated places; elsewhere the curve is likely to be both continuous and differentiable. However, it is possible to construct a continuous function that has "problem points" everywhere and, therefore, is nowhere differentiable! The first example was found by Karl **Weierstrass** in 1872 and came as a total surprise. It is defined as an infinite series

$$f(x) = \sum_{n=0}^{\infty} B^n \cos(A^n \pi x),$$

where A and B can be any numbers such that B is between 0 and 1, and $A \times B$ is bigger than $1 + (3\pi/2)$.

continuum

Any **set** that can be brought into **one-to-one** correspondence with the set of **real numbers**. Examples include a finite line segment, a square, a circle, and a disk.

continuum hypothesis

In 1874 Georg **Cantor** discovered that there is more than one level of **infinity**. The lowest level is called *countable infinity;* higher levels are known as *uncountable infinities.* The **natural numbers** are an example of a countably infinite set and the **real numbers** are an example of an uncountably infinite set. The continuum hypothesis, put forward by Cantor in 1877, says that the number of real numbers is the *next* level of infinity above countable infinity. It is called the continuum hypothesis (CH) because the real numbers are used to represent a linear continuum. Let c be the cardinality of (i.e., number of points in) a continuum, **aleph**-null (χ_0) be the cardinality of any countably infinite set, and χ_1 be the next level of infinity above χ_0. CH is equivalent to saying that there is no **cardinal number** between χ_0 and c, and that $c = \chi_1$. CH has been, and continues to be, one of the most hotly pursued problems in mathematics.

convergence

A property of some **sequences**. A sequence u_i is said to be convergent if there exists a value u with the property that by choosing a large enough value of i, we can make u_i as close as we wish to u.

convex

Curved outward, like the exterior surface of a sphere; the word comes from the Latin *convexus* for "vaulted." A figure, such as a **polygon** or a **polyhedron**, is said to be convex if every line segment that joins two interior points remains inside the figure. Similarly, a **set** is convex if it contains all the line segments connecting any pair of its points.

Conway, John Horton (1937–)

A British-born (Liverpool) mathematician, who studied and taught at Cambridge University and is now a professor at Princeton University. Conway has been an extraordinarily fertile source of new ideas in mathematics and of mathematical games. His most significant contribution was the discovery of **surreal numbers**, to which he was led after watching the British **Go** champion play at Cambridge. In 1967, he found a cluster of three new sporadic **groups**, now sometimes called *Conway's constellation,* building on an earlier discovery by John Leech of an extremely dense **packing** of unit spheres in a space of 24 dimensions. He has also been active in the field of **knots** and in **coding theory**. Among amateur mathematicians, Conway is best known as the inventor of the games of **Life**, **Sprouts**, and **Phutball**,

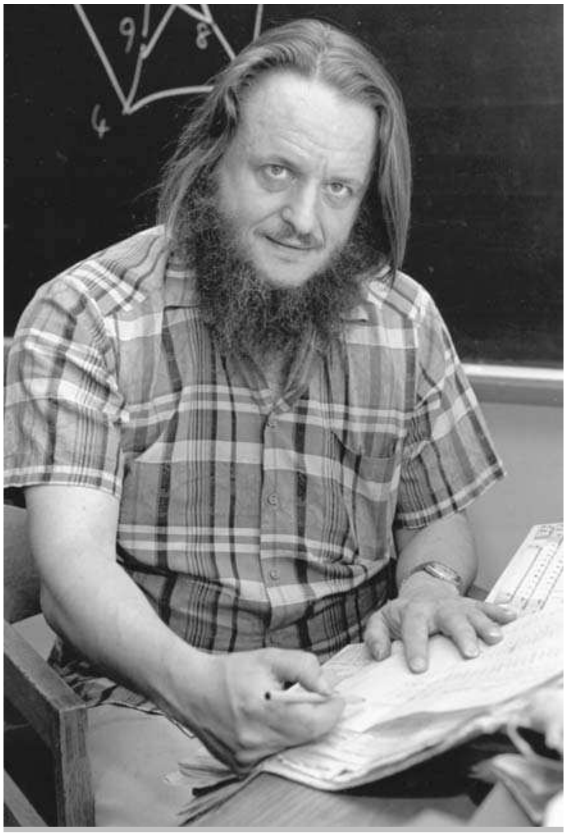

Conway, John Horton *Princeton University*

as well as for his detailed analyses of many other games and puzzles, such as the **Soma Cube**.

Conway's chained-arrow notation

One of various methods that have been devised recently for representing extremely **large number**s. Developed by John **Conway**, it is based on **Knuth's up-arrow notation** but is even more powerful. The two systems are related thus:

$$a \to b \to 1 = a \uparrow b$$
$$a \to b \to 2 = a \uparrow\uparrow b$$
$$a \to b \to 3 = a \uparrow\uparrow\uparrow b$$
$$a \to b \to c = a \uparrow\uparrow \ldots \uparrow\uparrow b \text{ (c up arrows)}$$

Longer chains are evaluated by the following general rules:

$$a \to \ldots \to b \to c \to 1 = a \to \ldots \to b \to c$$
$$a \to \ldots \to b \to 1 \to d + 1 = a \to \ldots \to b$$
$$\text{and} \quad a \to \ldots \to b \to c + 1 \to d + 1$$
$$= a \to \ldots \to b \to (a \to \ldots \to b \to c \to d) \to d$$

It's important to recognize that the Conway arrow isn't an ordinary *dyadic* operator. Where three or more numbers are joined by arrows, the arrows don't act separately but rather the whole chain has to be considered as a unit. The chain might be thought of as a **function** with a variable number of arguments, or as a function whose single argument is an ordered list or **vector**. The **Ackermann function** is equivalent to a three-element chain: $A(m, n) = (2 \to (n + 3) \to (m - 2)) - 3$. It can also be shown that **Graham's number** is bigger than $3 \to 3 \to 64 \to 2$ and smaller than $3 \to 3 \to 65 \to 2$.

coordinate

One of a set of variables that specifies the location of a point in space. If the coordinates are distances measured along perpendicular axes, they are known as **Cartesian coordinates**. See also **polar coordinates**.

coordinate geometry

See **analytical geometry**.

coprime

Two or more numbers are coprime if they have no **factor**s in common other than 1.

cosine

See **trigonometric function**.

countable set

A **set** that is either finite or *countably infinite.* A countably infinite set is one that can be put in **one-to-one** correspondence with the **natural number**s and thus has a **cardinal number** ("size") of **aleph**-null ($\aleph_0$). Examples of countable sets include the set of all people on Earth and the set of all fractions. See also **infinity**.

counterfeit coin problem

Among *n* coins, identical in size, shape, and appearance, one is a counterfeit and has a slightly different weight than the others. Using only a two-pan balance, what is the smallest number of weighings that would guarantee finding the fake coin? The problem of the counterfeit coin (or some other object), especially involving 8, 10, 12, or 13 coins, has cropped up in many guises over the years. Typically, the problem also involves finding whether the counterfeit coin is lighter or heavier than the rest. The answer depends on the specific problem and can involve quite a number of steps.

covariance

The tendency of two **random** variables to move in tandem. This is important in applications such as survey-taking and sociology, as well as in many branches of science, because if two things tend to vary together, there is a good chance they may be causally linked. See also **causality**.

Coxeter, Harold Scott MacDonald (1907–2003)

A British-born, Cambridge-educated mathematician who spent most of his career (from 1936 on) at the University of Toronto and was regarded as the greatest classical geometer of his generation. Always known as "Donald," he is best known for his work on hyperdimensional geometries and regular **polytopes**.

In 1926, at the age of 19, Coxeter discovered a new regular polyhedron, having six hexagonal faces at each vertex. He went on to study the mathematics of kaleidoscopes and, by 1933, had enumerated the *n*-dimensional kaleidoscopes. His algebraic equations expressing how many images of an object may be seen in a kaleidoscope are now known as *Coxeter groups*. His research on icosahedral symmetries played an important role in the discovery by scientists at Rice University, Texas, of the carbon-60 molecule (see **buckyball**), for which they won the 1996 Nobel Prize in Chemistry.

Coxeter was a close friend of the artist M. C. **Escher**, whom he met in 1954, and also of Buckminster Fuller, who used Coxeter's ideas in his architecture. Indeed Coxeter's work was motivated by a strong artistic temperament and a sense of what is beautiful. He had originally intended to be a composer but fascination for symmetry took him toward mathematics and a career about which he said "I am extremely fortunate for being paid for what I would have done anyway."

Several of Coxeter's books are considered classics, including *The Real Projective Plane* (1955), *Introduction to Geometry* (1961),[74] *Regular Polytopes* (1963),[75] *Non-Euclidean Geometry* (1965)[72] and, written jointly with S. L. Greitzer, *Geometry Revisited* (1967). In 1938, he revised and updated Rouse **Ball**'s *Mathematical Recreations and Essays*.[24]

cross

A shape that consists in its most basic form of an upright section and a transverse section. The *Latin cross* has the shape of an irregular **dodecahedron** with a single (vertical) line of symmetry, and can be folded up to make a cube. The *Greek cross* has the shape of a plus sign, has four lines of symmetry, and is used as the emblem of the Red Cross organization. A version of the Greek cross that has flared ends is also known as the *crux immissa* or *cross patée.* A cross of Saint Andrew is an ordinary Greek cross rotated through 45°, and is also called the *crux decussata;* it served as the basis for the multiplication sign. A cross of Saint Anthony takes the form of a capital T. The *Maltese cross* is an irregular dodecahedron whose cross pieces flange out from the center.

crunode

A point where a curve intersects itself so that two branches of the curve have distinct **tangent** lines.

cryptarithm

A number puzzle in which a group of arithmetical operations has some or all of its digits replaced by letters or symbols, and where the original digits must be found. In such a puzzle, each letter or symbol represents a unique digit. The first example appeared in *American Agriculturist* in 1864. Specific types of cryptarithm include the **alphametic**, the **digimetic**, and the **skeletal division**.

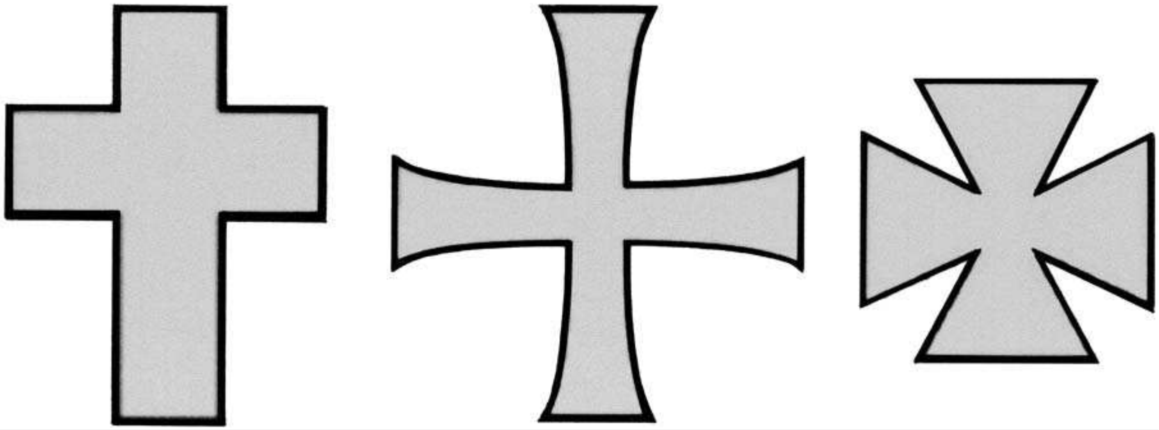

cross From left to right: a Latin cross; a *crux immissa* (a Greek cross with flared ends), also sometimes called a Latin cross; and a Maltese cross.

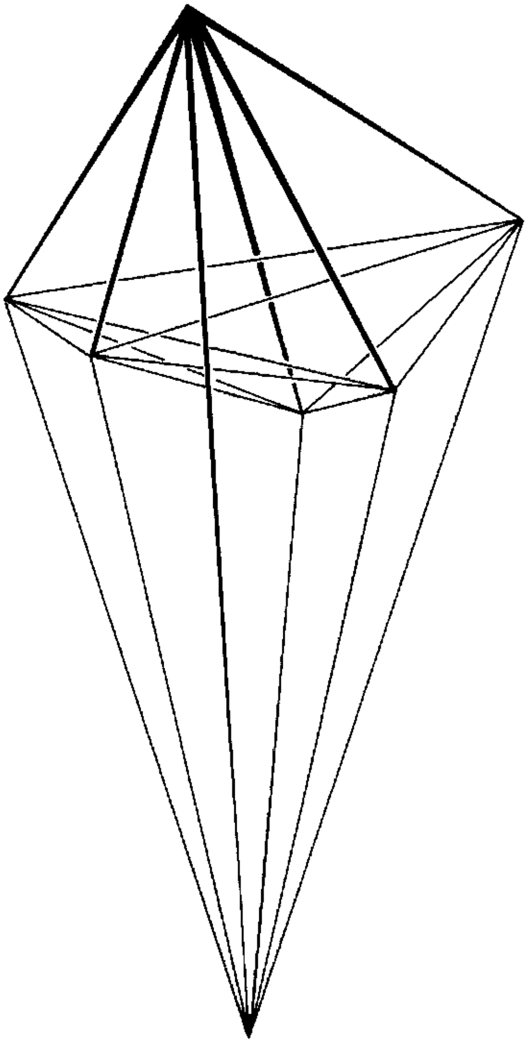

Császár polyhedron

cryptography

The science and mathematics of encoding and decoding information. See also **cipher** and **cryptarithm**.

Császár polyhedron

A **polyhedron**, first described in 1949 by the Hungarian mathematician Ákos Császár,[79] that is a solution to an interesting problem, namely: How many polyhedra exist such that every pair of vertices is joined by an edge? The first clear example is the well known **tetrahedron** (triangular pyramid). Some simple **combinatorics** specify how many vertices, edges, faces, and holes such polyhedra must have. It turns out that, other than the tetrahedron, any such polyhedron must have at least one hole. The first possible polyhedron beyond the tetrahedron has exactly one hole; this is the Császár polyhedron, which is thus topologically equivalent to a **torus** (donut). The Császár polyhedron has 7 vertices, 14 faces, and 21 edges, and is the **dual** of the **Szilassi polyhedron**. It isn't known if there are any other polyhedra in which every pair of vertices is joined by an edge. The next possible figure would have 12 faces, 66 edges, 44 vertices, and 6 holes, but this seems an unlikely configuration–as, indeed, to an even greater extent, does any more complex member of this curious family.

cube

(1) The **Platonic solid** that has a square for every one of its 6 faces; it also has 12 edges and 8 vertices (corners). The 60′ × 30′ × 30′ Double Cube Room of Wilton House (the seat of the Earl of Pembroke), near Salisbury, is considered, together with the Single Cube Room of the same domicile, among the finest surviving rooms in England from the mid-seventeenth century. A favorite with filmmakers, it has provided locations for *Barry Lyndon* by Stanley Kubrick, *The Madness of King George*, and *Sense and Sensibility*. See also **Atomium, the**. (2) To cube something is to raise it to the power of three. The result of cubing is a *cube number:* $1^3 = 1$, $2^3 = 8$, $3^3 = 27$, and so on. To take the *cube root* is the reverse process; thus, 4 cubed (4^3) is 64 and the cube root of 64 ($\sqrt[3]{64}$) is 4. For cube **dissection** problems, see **Hadwiger problem**, **Slothouber-Graatsma puzzle** and **Soma cube**. See also **tesseract** and **Prince Rupert's problem**.

cubic curve

An **algebraic curve** described by a **polynomial** equation of the general form

$$ax^3 + bx^2y + cxy^2 + dy^3 + ex^2 + fxy + gy^2 + hx + iy + j = 0,$$

where $a, b, c, d, e, f, g, h, i,$ and j are constants, such that at least one of $a, b, c,$ and d is nonzero, and x and y are variables. One of Isaac **Newton**'s many accomplishments was the classification of the cubic curves. Newton found 72 different species of curve; later investigators found six more, and it is now known that there are precisely 78 different types of cubic curves. Interesting examples include the **folium** of Decartes and the Witch of **Agnesi**.

cubic equation

A **polynomial** equation of the third degree, the general form of which is

$$ax^3 + bx^2 + cx + d = 0,$$

where $a, b, c,$ and d are constants. There was a great controversy in sixteenth-century Italy between Girolamo **Cardano** and Niccoló **Tartaglia** about who should get credit for solving the cubic. At this time symbolic algebra hadn't been developed, so all the equations were written in words instead of symbols. Early studies of cubics helped legitimize **negative number**s, give a deeper insight into equations in general, and stimulate work that eventually led to the discovery and acceptance of **complex number**s. Cardano, in his *Ars Magna,* found negative solutions to equations, but called them "fictitious." He also noted an important fact connecting solutions of a cubic equation to its **coefficient**s, namely, that the sum of the solutions is the negation of b, the coefficient of the x^2 term. At one other point, he mentions that the problem of dividing 10 into two parts so that their product is 40 would have to be $5 + v(-15)$ and $5 - v(-15)$. Cardano didn't go further than this observation of what later came to be called complex numbers, but a few years later Rafael Bombelli (1526–1672) gave several examples that involved these strange new mathematical beasts.

cubit

A measure of length used in the ancient world. It is approximately equal to the length of a person's forearm, that is, the part of the arm from the elbow to the fingers. The Romans used a cubit equal to 17.4 modern inches; the Egyptians used one of 20.64 inches.

cuboctahedron

A **polygon** obtained by cutting the corners off a cube or an **octahedron**. It has eight faces that are equilateral triangles and six faces that are squares.

cuboid

Also called a *rectangular prism,* a **hexahedron** of which all of the faces are rectangles and all of the opposite faces are identical. It is not known whether a *perfect cuboid,* whose sides, face diagonals, and space diagonals are all integers, exists. The general suspicion is that it doesn't, although several near misses have been found, including one in which $a = 240$, $b = 117$, $c = 44$, $dab = 267$, $dac = 244$, and $dbc = 125$. If there is a perfect cuboid, it has been shown that the smallest side must be at least $2^{32} = 4{,}294{,}967{,}296$.

Cullen number

A number of the form $(n \times 2^n) + 1$, denoted C_n, and named after the Reverend James Cullen (1867–1933), an Irish Jesuit priest and schoolmaster. Cullen noticed that the first, $C_1 = 3$, was a **prime number**, but with the possible exception of the fifty-third, the next 99 were all composite. Soon afterward, Cunningham discovered that 5,591 divides C_{53}, and noted that all the Cullen numbers are **composite number**s for n in the range $2 \leq n \leq 200$, with the possible exception of 141. Five decades later Robinson showed that C_{141} is a prime. Currently, the only known Cullen primes are those with n = 1, 141, 4,713, 5,795, 6,611, 18,496, 32,292, 32,469, 59,656, 90,825, 262,419, 361,275, and 481,899. Although the vast majority of Cullen numbers are composite, it has been conjectured that there are infinitely many Cullen primes. Whether n and C_n can simultaneously be prime isn't known. Sometimes, the name "Cullen number" is extended to include the Woodall numbers, $W_n = (n \times 2^n) - 1$. Finally, a few authors have defined a number of the form $(n \times b^n) + 1$, with $n + 2 > b$, to be a *generalized Cullen number.*

Cunningham chain

A sequence of **prime number**s in which each member is twice the previous one plus one. For example, {2, 5, 11, 23, 47} is the first Cunningham chain of length 5 and {89, 179, 359, 719, 1,439, 2,879} is the first of length 6. In general, a *Cunningham chain of length k of the first kind* is a sequence of k prime numbers, each of which is twice the preceding one plus one. A *Cunningham chain of length k of the second kind* is a sequence of k primes, each of which is twice the preceding one minus one. For example, {2, 3, 5} is a Cunningham chain of length 3 of the second kind and {1,531, 3,061, 6,121, 12,241, 24,481} is a Cunningham chain of length 5 of the second kind. Prime chains of both these forms are said to be complete if they can't be extended by adding either the next larger or the next smaller terms. See also **Sophie Germain prime**.

cup

The symbol $\cup$, which is used to denote the **union** of two **set**s.

curvature

A measure of the amount by which a curve, a surface, or any other **manifold** deviates from a straight line, a plane, or a hyperplane (the multidimensional equivalent of a plane). For a plane curve, the curvature at a given point has a magnitude equal to one over the radius of an *osculating* circle (a circle that "kisses," or just touches, the curve at the given point) and is a **vector** pointing in the direction of that circle's center. The smaller the radius r of the osculating circle, the greater the magnitude of the curvature $(1/r)$ will be. A straight line has zero curvature everywhere; a circle of radius r has a curvature of magnitude $1/r$ everywhere.

For a two-dimensional surface, there are two kinds of curvature: a *Gaussian* (or *scalar*) *curvature* and a *mean curvature.* To compute these at a given point, consider the intersection of the surface with a plane containing a fixed normal vector (an arrow sticking out perpendicularly) at the point. This intersection is a plane and has a curvature; if the plane is varied, this curvature also changes, and there are two extreme values–the maximal and the minimal curvature–which are known as the *main curva-*

tures, $1/R_1$ and $1/R_2$. (By convention, a curvature is taken to be positive if its vector points in the same direction as the surface's chosen normal, otherwise it is negative.) The Gaussian curvature is equal to the product $1/R_1R_2$. It is everywhere positive for a **sphere**, everywhere negative for a **hyperboloid** and **pseudosphere**, and everywhere zero for a plane. It determines whether a surface has elliptic (when it is positive) or hyperbolic (when it is negative) geometry at a point. The integral of the Gaussian curvature over the whole surface is closely related to the surface's **Euler characteristic**. The mean curvature is equal to the sum of the main curvatures, $1/R_1 + 1/R_2$.

A **minimal surface**, like that of a soap film, has a mean curvature of zero. In the case of higher-dimensional manifolds, curvature is defined in terms of a curvature **tensor**, which describes what happens to a vector that is transported around a small loop of the manifold.

curve

A continuous mapping from a one-dimensional space to an *n*-dimensional space. The most familiar mathematical curves are two- and three-dimensional **graph**s. A curve, such as a circle, that lies entirely in a plane is called a *plane curve;* by contrast, a curve that may pass through any region of three-dimensional space is called a *space curve.* See also **space-filling curve**.

curve of constant width

A curve that, when rotated in a square, makes continuous contact with all four sides. It may seem, at first sight, as if there is only one such curve–a circle. But, in fact, there are infinitely many different curves of constant width. The circle is the one with the largest area. The simplest noncircular one, and the one with the smallest area, is the **Reuleaux triangle**. Others can be constructed starting with equilateral (but not necessarily equiangular) stars. Every curve of constant width is **convex**. Moreover, *Barbier's theorem* states that every curve of constant width w has the same perimeter, πw. (The width of a convex figure is defined as the distance between parallel lines–known as *supporting lines*–that bound it.) A curve of constant width can be used in a special drill chuck to cut square holes. A generalization gives solids of constant width. These do not have the same surface area for a given width, but their shadows are curves of constant width with the *same* width.

cusp

In mathematics, a point on a curve where two branches, coming from different directions, meet and have a common **tangent**. If the two branches of the curve approach the tangent from opposite sides the cusp is called a *keratoid* (from the Greek *kera* for "horn") or *first-order cusp.* This is the case, for example, with the curve given by the equation $y^2 = x^2y + x^5$. If the two branches of the curve approach the tangent from the same side the result is a *ramphoid* or *second-order cusp.* "Cusp" derives from the Latin *cuspis* for "sharp." Outside of mathematics, the points of a crescent moon are called cusps and the sharp pointed premolar teeth of children are known as bicuspids.

cute number

A number n such that a square can be cut into n squares of, at most, two different sizes. For example, 4 and 10 are cute numbers.

Cutler, William (Bill)

An Australian puzzle maker and solver who, in 1977, became the first to completely analyze, using a computer, six-piece burrs used to make solid six-piece **burr puzzles**. Martin Gardner devoted his January 1978 "Mathematical Games" column in *Scientific American* to this and other of Cutler's discoveries. In 2003, Cutler used a computer to enumerate all solutions of the **Loculus of Archimedes**.

cybernetics

The theoretical study of communication and control processes in biological, mechanical, and electronic systems, especially the comparison of these processes in biological and artificial systems. It was pioneered by Norbert **Wiener**.

cyclic number

A number with n digits, which, when multiplied by 1, 2, 3, . . . , n produces the same digits in a different order. For example, 142,857 is a cyclic number: 142,857 × 2 = 285,714; 142,857 × 3 = 428,571; 142,857 × 4 = 571,428; 142,857 × 5 = 714,285; 142,857 × 6 = 857,142, and so on. It has been conjectured, but not yet proven, that an infinite number of cyclic numbers exist.

cyclic polygon

A **polygon** with vertices (see **vertex**) that all lie on the same circle. All triangles are cyclic (but not all of any other kind of polygon) because any set of three points, not lying on a single line, can have a circle drawn through it.

cycloid

The shape defined by a fixed point on a wheel as it rolls; more precisely, it is the **locus** of a point on the rim of a circle rolling along a perfectly straight line. The cycloid was named by Galileo in 1599. It is the solution to both the **tautochrone problem** and the **brachistochrone problem**. In 1634, the French mathematician Gilles de Roberval (1610–1675) showed that the area under a cycloid is three times the area of its generating circle. In 1658, the English architect Christopher Wren showed

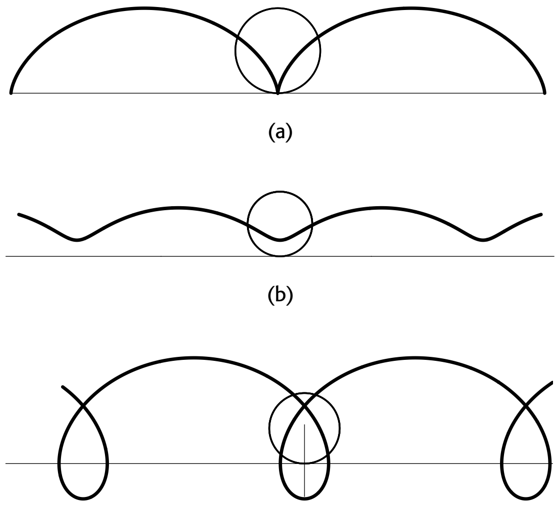

cycloid An ordinary cycloid is traced out by a point on a wheel as it rolls along a flat surface (a). A curtate cycloid is traveled by a point on the wheel that is inside the circumference (b). If the point lies outside the circumference of the wheel, the result is a prolate cycloid (c).

that the length of a cycloid is four times the diameter of its generating circle. But there was a lot of bickering and a lack of public sharing of information around this time that led to much duplication of effort, particularly over questions related to the cycloid. In fact, the confusion was so bad that the curve was nicknamed the Helen of Geometers, and Jean Montucla referred to it as "*la pomme de discorde*" (the apple of discord).

As well as the ordinary cycloid there is the *curtate cycloid,* which is the path traced out by a point on the inside of a rolling circle, and the *prolate cycloid,* which is followed by a point on the outside of the circle. A prolate cycloid is traced out, for example, by points on the flange of the wheels of a locomotive, which extends below the top of the tracks. This leads to the surprising conclusion that even as the locomotive is moving forward there are always parts of its wheels that are going backward for a moment before moving forward again. See also **epicycloid** and **hypocycloid**.

cylinder

A three-dimensional surface described by the Cartesian equation $(x/a)^2 + (y/b)^2 = 1$. If $a = b$ then the surface is a *circular cylinder,* otherwise it is an *elliptic cylinder.* The cylinder is a *degenerate quadric* because at least one of the coordinates (in this case z) doesn't appear in the equation, though by some definitions the cylinder isn't considered to be a quadric at all. In common usage, a *cylinder* is taken to mean a finite section of a right circular cylinder with its ends closed to form two circular surfaces. If the cylinder has a radius r and a length h, then its volume is $V = \pi r^2h$ and its surface area is $A = 2\pi r^2 + 2\pi rh$. For a given volume, the cylinder with the smallest surface area has $h = 2r$. For a given surface area, the cylinder with the largest volume has $h = 2r$. More unusual types of cylinder include the *imaginary elliptic cylinder:* $(x/a)^2 + (y/b)^2 = -1$, the *hyperbolic cylinder:* $(x/a)^2 - (y/b)^2 = 1$, and the *parabolic cylinder:* $x^2 + 2y = 0$.

D

d'Alembert, Jean Le Rond (1717–1783)

A French mathematician named for the church of St. Jean Baptiste de Rond upon whose steps he was abandoned as a baby, the illegitimate son of a Parisian society hostess. He clarified the concept of a **limit** in calculus, discovered the *Cauchy-Riemann equations* decades before Augustin **Cauchy** or Bernhard **Riemann**, was the first to find and solve the wave equation, and recast **Newton**'s third law in a new and powerful form through what has become known as *d'Alembert's principle.*

Dandelin spheres

If a **cone** is sliced through by a plane, the two spheres that just fit inside the cone, one on each side of the plane and both tangent to it and touching the cone, are known as Dandelin spheres. They are named after the Belgian mathematician and military engineer Germinal Pierre Dandelin (1794–1847) who gave an elegant proof that the two spheres touch the **conic section** at its foci. In 1826, Dandelin showed that the same result applies to the plane sections of a **hyperboloid** of revolution.

dart

Also known as an *arrowhead,* a special kind of quadrilateral that has one **reflex angle**. See also **Penrose tiling**.

de L'Hôpital, Guillaume François Antoine, Marquis de (1661–1704)

A French mathematician who wrote the first textbook on differential calculus, *Analyse des infiniment petits pour l'intelligence des lignes courbes* (1696). This contains the rule, now known as *L'Hôpital's rule,* for finding the limit of a rational function whose numerator and denominator tend to zero at a point. Along with Isaac **Newton**, Gottfried **Leibniz**, and Jacob Bernoulli (see **Bernoulli family**), de L'Hôpital was among the first to solve the **brachistochrone problem**.

de L'Hôpital's cubic

See **Tschirnhaus's cubic**.

de Malves's theorem

Given a **tetrahedron** in which the edges meeting at one **vertex**, X, form three right angles (i.e., the tetrahedron is the result of chopping off the corner of a **cuboid**), the square of the face opposite X is equal to the sum of the squares of the other three faces.

de Méré's problem

A question posed in the mid-seventeenth century to Blaise **Pascal** by a French nobleman and inveterate gambler, the Chevalier de Méré, which marked the birth of **probability theory**. One of de Méré's favorite bets was that at least one six would appear during a total of four rolls of a die. From past experience, he knew that this gamble paid off more often than not. Then, for a change, he started betting that he would get a double-six on 24 rolls of two dice. However, he soon realized that his old approach to the game was more profitable. He asked his friend Pascal why. Pascal showed that the probability of getting at least one six in four rolls of a die is $1 - (5/6)^4 \approx 0.5177$, which is slightly higher than the probability of at least one double-six in 24 throws of two dice, $1 - (35/36)^{24} \approx 0.4914$. This problem and others posed by de Méré are thought to have been the original inspiration for a fruitful exchange of letters on probability between Pascal and Pierre de **Fermat**. To tackle these problems, Fermat used combinatorial analysis (finding the number of possible outcomes in ideal games of chance by computing permutation and combination numbers), while Pascal reasoned by recursion (an iterative process that determines the result of the next case by the present case). Their combined work laid the foundations for probability theory as we know it today.

de Moivre, Abraham (1667–1754)

A French-British mathematician who founded analytical trigonometry and stated what has become known as **de Moivre's theorem**. He also worked on probability theory and the normal distribution, and was a good friend of Isaac **Newton**. In 1698 he wrote that the theorem had been known to Newton as early as 1676.

de Moivre's theorem

A theorem, named after Abraham **de Moivre**, that links **complex number**s and trigonometry. It states that for any real number x and any integer n,

$$(\cos x + i\sin x)^n = \cos(nx) + i\sin(nx).$$

By expanding the left-hand side and then comparing real and imaginary parts, it is possible to derive useful expressions for $\cos(nx)$ and $\sin(nx)$ in terms of $\sin(x)$ and $\cos(x)$. Furthermore, the formula can be used to find explicit expressions for the nth **root of unity**: complex numbers z such that $z^n = 1$. It can be derived from (but historically preceded) **Euler's formula** $e^{ix} = \cos x + i \sin x$ and the exponential law $(e^{ix})^n = e^{inx}$.

de Morgan, Augustus (1806–1871)

A British mathematician, born in India, who was an important innovator in the field of mathematical logic. The system he devised to express such notions as the contradictory, the converse, and the transitivity of a relation, as well as the union of two relations, laid some of the groundwork for his friend George **Boole**. De Morgan lost the sight of his right eye shortly after birth, entered Trinity College, Cambridge, at the age of 16, and received his B.A. However, he objected to a theological test required for the M.A. and returned to London to study for the bar. In 1827, he applied for the chair of mathematics in the newly founded University College, London and, despite having no mathematical publications, he was appointed. In 1831, he resigned on principle (after another professor was fired without explanation) but regained his job five years later when his replacement died in an accident. He resigned again in 1861.

His most important published work, *Formal Logic,* included the concept of the *quantification of the predicate,* an idea that solved problems that were impossible under the classic Aristotelian logic. De Morgan coined the phrase "**universe of discourse**," was the first person to define and name mathematical **induction**, and developed a set of rules to determine the **convergence** of a mathematical series. In addition, he devised a decimal coinage system, an almanac of all full moons from 2000 B.C. to A.D. 2000, and a theory on the probability of life events that is still used by insurance companies. De Morgan was also deeply interested in the history of mathematics. In *Arithmetical Books* (1847) he describes the work of over fifteen hundred mathematicians and discusses subjects such as the history of the length of a foot, while in *A Budget of Paradoxes* he gives a marvelous compendium of eccentric mathematics including the poem

> Great fleas have little fleas upon their backs to bite 'em,
> And little fleas have lesser fleas, and so ad infinitum,
> And the great fleas themselves, in turn, have greater fleas to go on,
> While these again have greater still, and greater still, and so on.

The first lines of this poem paraphrase a similar rhyme by Jonathan Swift.

PUZZLE

On one occasion, when asked his age, de Morgan replied: "I was x years old in the year x^2." How old must he have been at the time?

Solutions begin on page 369.

decagon

A **polygon** with 10 sides.

decimal

The commonly used **number system**, also known as *denary,* in which each place has a value 10 times the value of the place at its right. For example, 4,327 in the decimal (base 10) system is shorthand for $(4 \times 10^3) + (3 \times 10^2) + (2 \times 10^1) + (7 \times 10^0)$, where $10^0 = 1$. "Decimal" comes from the Latin *decimus* for "tenth." The verb *decimare,* literally "to take a tenth of," was used to describe a form of punishment applied to mutinous units in the Roman army. The men were lined up and every tenth soldier was killed as a lesson to the rest. From this custom comes our word *decimate,* which we use more loosely–in fact, incorrectly–to indicate near-total destruction. The Latin *decimare* was also used in a less ferocious sense to mean "to tax to the amount of one tenth." However, the usual word describing a one-tenth tax in English is *tithe,* which comes from the Old English *teogotha,* a form of *tenth.*

decimal fraction

A number consisting of an **integer** part, which may be zero, and a decimal part less than unity that follows the decimal marker (which may be a point or a comma). A *finite* or *terminating decimal fraction* has a sequence of decimals with a definite break-off point after which all the places are zeros. Other fractions produce endless sequences of decimals that are *periodic nonterminating.*

Dedekind, (Julius Wilhelm) Richard (1831–1916)

A German mathematician whose most important contribution was the discovery of what became known as the *Dedekind cut.* He realized that every **real number** r divides the **rational numbers** into two subsets: those greater than r and those less than r. Dedekind's brilliant idea was to represent the real numbers by such divisions of the rationals. He also provided important support for Georg **Cantor**'s set theory, which was highly controversial at the time.

Dee, John (1527–1609)

A notable English alchemist, mathematician, and astronomer, sometimes referred to as the "last magician" because of his astrological services to Queen Elizabeth I; Dee may also have influenced the writings of Shakespeare. He enrolled at St. Johns College, Cambridge, at age 15, but found the atmosphere there stifling and later went to the Continent to study and lecture. Upon his return to England, Dee cast the horoscope for Queen Mary and later visited Mary's half-sister Elizabeth in jail to determine when Mary would die. Accused of black magic, he was jailed and then released in 1555, three years before Mary's death. When Elizabeth came to the throne she consulted Dee on many matters, including the geography of newly discovered lands, and paid him well. Some of his income he spent on extensive traveling, which may have involved some spying on behalf of his sponsor.

Dee had a large library of books on witchcraft, the occult, and magic, and he wrote 79 manuscripts, only a few of which were published. He married three times and fathered eight children. He also struck up an uneasy partnership with Edward Kelly, a bad-tempered Irishman who claimed to have discovered the alchemical secret of transmuting base metal into gold but had lost his ears for forgery. In 1585 Dee and Kelly went on a four-year trek across the Continent conducting astrological readings for nobility and royalty. But Dee and Kelly had many arguments and eventually parted company. Back in England Dee found his house ransacked and many of his possessions stolen or destroyed. Elizabeth helped pay for the damage and made him warden of Christ's College in Manchester in 1595. However, Elizabeth died in 1603 and her successor James I opposed magic. Dee was forced to retire, his life ending in poverty.

deficient number

See **abundant number**.

degree

(1) The unit of measurement for angles; one degree is 1/360 of a circle. (2) The **exponent** of a variable. For example, the degree of $7x^5$ is 5. See also **degree of freedom**.

degree of freedom

A positive integer that gives the number of pieces of data that are independent.

deletable prime

See **truncatable prime**.

delta curve

A curve that can be turned inside an equilateral **triangle** while continuously making contact with all three sides. There are an infinite number of delta curves, but the simplest are the **circle** and lens-shaped delta-biangle. All the delta curves of height h have the same perimeter $2\pi h/3$. See also **Reuleaux triangle** and **rotor**.

deltahedron

A **polyhedron** whose faces consist of equilateral triangles that are all the same size. Although there are an infinite number of different deltahedra, only eight of them are convex, as O. Rausenberger first showed in 1915. Among this group of eight, faces made of coplanar equilateral triangles sharing an edge (such as the rhombic dodecahedron) aren't allowed. The eight convex deltahedra have 4, 6, 8, 10, 12, 14, 16, and 20 faces.

deltoid

A **hypocycloid** with three cusps, also known as a *tricuspoid* or *Steiner's hypocycloid* after the Swiss mathematician Jakob **Steiner** who investigated the curve in 1856. The deltoid, so-named because it looks like an uppercase Greek delta, Δ, is formed by a point on the circumference of a circle rolling inside another circle with a radius three times as large. While working on a problem in optics in 1745, Leonhard **Euler** was among the first to study its properties. The parametric equations of the cycloid with inner circle of radius r are:

$$x(\mathrm{t}) = 2r\cos t + r\cos 2t$$
$$y(\mathrm{t}) = 2r\sin t - r\sin 2t$$

The length of the path of the deltoid is $16r/3$, and the area inside the deltoid is $2\pi r^2$. If a tangent is drawn to the deltoid at some point, P, and the points where the tangent crosses the deltoids other two branches are called points A and B, then the length of AB equals $4r$. If the deltoid's tangents are drawn at points A and B, they will

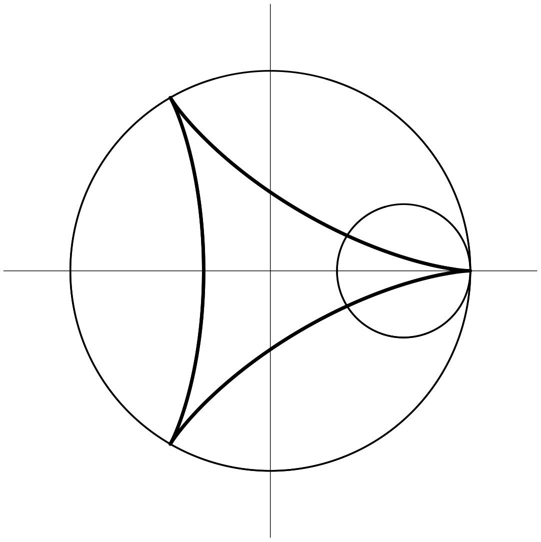

deltoid The deltoid curve. © *Jan Wassenaar, www.2dcurves.com* be perpendicular, and they will intersect at a point inside the deltoid that is the 180° rotation of point P about the center of the fixed circle.

denominator

In a **rational number**, the number below the fraction bar; it indicates into how many parts the whole is divided.

derivative

The result of differentiating a **function**; that is, the infinitesimal change in a function caused by an infinitesimal change in the variable(s) upon which it depends. The derivative gives the rate of change of a function (the slope of its curve) at a particular point. Second and third derivatives give the rate at which the rate of change is changing and the rate at which the rate of rate of change is changing, respectively. In an article in 1996, Hugo Rossi wrote: "In the fall of 1972 President Nixon announced that the rate of increase of inflation was decreasing. This was the first time a sitting president used the third derivative to advance his case for reelection."[269]

Here is a fallacious "proof" that $x = 2x$ based on derivatives. Consider the function $f(x) = x^2$, the derivative of which is $2x$. What is wrong with the following?

$$x^2 = x + x + \ldots + x \text{ (repeated } x \text{ times)}$$

Taking the derivative of both sides gives

$$(x^2)' = 1 + 1 + \ldots + 1 = x.$$

But we have already said that the derivative of x^2 is $2x$. Therefore, $x = 2x$. The error stems from taking the derivative of x different x's. Each of the terms depends not only on x, which was accounted for in taking the derivative, but also on the *number* of terms (which could be fractional) which depends on x, too, and this was not accounted for. Put another way, the derivative measures the rate of change of x^2 as x changes, but as x changes, the number of terms on the right, as well as the terms themselves, increases. For positive x, the correct answer must be larger than x–as indeed it is.

Desargues, Girard (1591–1661)

A French mathematician who is regarded as the chief founder of perspective geometry. His 12-page treatise *La perspective* (1636) consists of a single worked example in which Desargues sets out a method for constructing a perspective image without using any point lying outside the picture field. He considers the representation in the picture of a plane of lines that meet at a point and also of lines that are parallel to each another. In the last paragraph of the work he considers the problem of finding the perspective image of a conic section. Three years later, he wrote his treatise on projective geometry *Brouillon project d'une atteinte aux evenemens des rencontres du cone avec un plan* (Rough draft for an essay on the results of taking plane sections of a cone). The first part of this deals with the properties of sets of straight lines meeting at a point and of ranges of points lying on a straight line. In the second part, the properties of conics are investigated in terms of properties of ranges of points on straight lines and the modern term "point at infinity" appears for the first time. Desargues shows that he has completely grasped the connection between conics and perspective; in fact he treats the fact that any conic can be projected into any other conic as obvious. Given such innovative work it may seem surprising that the subject didn't develop rapidly in the following years. That may be partly due to mathematicians failing to recognize the power of what had been put forward. On the other hand, the algebraic approach to geometry put forward by René **Descartes** at almost exactly the same time (1637) may have diverted attention from Desargues's projective methods.

Descartes, René (1596–1650)

> *If you would be a real seeker after truth, it is necessary that at least once in your life you doubt, as far as possible, all things.*

A hugely influential philosopher and mathematician, born in La Haye (now named Descartes after its most famous son), Indre-et-Loire, France, who is often referred to as the father of modern philosophy and one of the founders of modern mathematics. He studied law at the University of Poitiers but never practiced it, served in the military for a while, and then lived in Holland for 20 years where he did the bulk of his great work. In his *Meditations on First Philosophy,* he tried to establish what can be known as true beyond doubt. His tool was methodological skepticism: the assumption that any idea that can be doubted is false. He gives the example of dreaming: in a dream, one senses things that seem to be real, but that don't actually exist. Thus, the data of the senses can't be fully trusted. Then again, he mused, perhaps there is an "evil genius"–a supremely powerful and devious being who sets out to prevent anyone from knowing the true nature of reality. Given these possibilities, what is it that one can know for certain? Descartes argues that if "I" am being deceived, then surely "I" must exist–the statement famously referred to as *cogito ergo sum* ("I think, therefore I am"), though these words don't actually appear anywhere in the *Meditations*. Descartes concludes that he can be certain that he exists. But in what form? If the senses are unreliable, Descartes reasons, all he can say for sure is that he is a *thinking thing*. He then proceeds to build a system of knowledge, discarding perception as unreliable and instead admitting only deduction as a method. Halfway through the *Meditations* he also claims to prove the existence of a benevolent **God** who has provided him with a working mind and sensory system, and who cannot desire to deceive him, and thus, finally, he establishes the possibility of acquiring knowledge about the world based on deduction *and* perception.

In mathematics, Descartes is important for his discovery of **analytical geometry**. Up to Descartes's time, **geometry**, dealing with lines and shapes, and **algebra**, dealing with numbers, were regarded as completely independent aspects of mathematics. Descartes showed how almost all problems in geometry can be translated into problems in algebra, by regarding them as questions asking for the length of a line segment, and using a coordinate system to describe the problem. Descartes's theory provided the basis for **calculus**, developed by Isaac **Newton** and Gottfried **Leibniz**, and thus for much of modern mathematics. This is particularly amazing when you consider that Descartes intended it merely as an example to support his *Discours de la méthode pour bien conduire sa raison, et chercher la verité dans les sciences* (Discourse on the method to rightly conduct the reason and search for the truth in sciences), better known under the shortened title *Discours de la méthode.*

Descartes died of pneumonia in Stockholm, where he had been invited to serve as tutor to the energetic 19-year-old Queen Christina of Sweden. Accustomed to working in a warm bed till noon, he was shocked into a rapid decline by having to teach philosophy at 5 A.M. in a freezing library. Seventeen years after his death, the Roman Catholic Church placed his works on the Index of Prohibited Books.

Descartes's circle theorem

See **Soddy's formula**.

determinant

A quantity obtained from a square ($n \times n$) array of numbers that can be useful, among other things, in solving systems of **linear** equations (equations in which the unknowns are raised to at most the first power). More generally, a determinant transforms a square **matrix** into a **scalar**–an operation that has many important properties. Two-by-two determinants were considered by Girolamo **Cardano** at the end of the sixteenth century and ones of arbitrary size by Gottfried **Leibniz** about a century later. Determinants are so named because, when applied to systems of linear equations, they "determine" if the systems are singular–that is, have multiple solutions. They also have important geometric applications, because they describe the area of a **parallelogram** and, more generally, the volume of a **parallelepiped**. A three-rowed determinant is defined by:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{array}{r} a_{11} \times a_{22} \times a_{33} + a_{12} \times a_{23} \times a_{31} \\ + a_{13} \times a_{21} \times a_{32} - a_{13} \times a_{22} \times a_{31} \\ - a_{12} \times a_{21} \times a_{33} - a_{11} \times a_{23} \times a_{32}\,. \end{array}$$

deterministic system

A system in which the later states of the system follow from, or are determined by, the earlier ones. Such a system contrasts with a *stochastic* or *random system* in which future states are not determined from previous ones. An example of a stochastic system would be the sequence of heads or tails of an unbiased coin, or radioactive decay.

If a system is deterministic, this doesn't necessarily imply that later states of the system are predictable from a knowledge of the earlier ones. In this way, **chaos** is similar to a random system. Chaos has been termed "deterministic chaos" since, although it is determined by simple rules, its property of sensitive dependence on initial conditions makes a chaotic system, in practice, largely unpredictable.

devil's curve

Also known as the *devil on two sticks,* a curve with the Cartesian equation

$$y^4 - a^2y^2 = x^4 - b^2x^2$$

and the polar equation

$$r^2\ (\sin^2\Theta - \cos^2\Theta) = a^2\sin^2\Theta - b^2\cos^2\Theta).$$

Early studies of it were carried out in 1750 by the Swiss mathematician Gabriel Cramer (1704–1752), who is

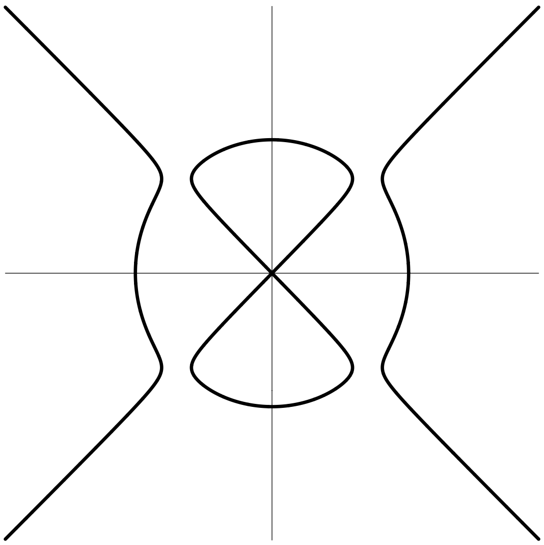

devil's curve *© Jan Wassenaar, www.2dcurves.com* most famous for his work on **determinant**s, and in 1810 by Lacroix. For $a = {}^{25}\!/_{24}$, the curve is called the *electric motor curve.*

Dewdney, Alexander Keewatin (1941–)

A Canadian computer scientist and mathematician at the University of Western Ontario, Canada, best known for his popular books and articles, most notably *The Planiverse: Computer Contact with a Two-dimensional World,* first published in 1984.[81] For several years, Dewdney wrote the "Mathematical Recreations" column for *Scientific American.*

diagonal

A line that joins any two vertices of a **polygon**, if the vertices are not next to each other; or a line that joins two vertices of a **polyhedron** that are not on the same face.

diagonal matrix

A **matrix** that has zero entries along all nondiagonal entries, that is, only the main diagonal may have nonzero values.

diameter

The distance across a **circle** through the center.

dice

Small polyhedra (see **polyhedron**), usually cubes, whose faces are numbered from one to six by patterns of dots, with opposite sides totaling seven. They are thrown, singly or in groups, from the hand or from a cup, onto a flat surface, to provide **random** numbers for gambling and other games. The face of each die that is uppermost when it comes to rest provides the value of the throw. Typical of their use today is the game of craps, in which two dice are thrown together, and bets placed on the total face-up value. Dice probably evolved from knucklebones, which are approximately tetrahedral. Even today, dice are sometimes colloquially referred to as "bones." Ivory, bone, wood, metal, and stone materials have been commonly used to make dice, though the use of plastics is now nearly universal.

Dice found in ancient tombs in the Orient point to an Asiatic origin and dicing is mentioned as an Indian game in the Rig-veda. In its primitive form, knucklebones was essentially a game of skill, played by women and children; gradually, a derivative form evolved for gambling in which four sides of the bones received different values and were counted like dice. Gambling with three, sometimes two, dice was a popular form of amusement in Greece, especially with the upper classes, and was an almost invariable accompaniment to the symposium, or drinking banquet. The Romans were passionate gamblers, and dicing was a favorite form, though it was forbidden except during the festival of Saturnalia (December 17). Throwing dice for money led to many special laws in Rome, one of which decreed that no suit could be brought by a person who allowed gambling in his house, even if he'd been cheated or assaulted! Professional gamblers were common, and some of their loaded dice are preserved in museums.

The Roman historian Tacitus states that the Germans also were passionately fond of dicing–so much so, that, having lost everything, they would even stake their personal liberty. Centuries later, in medieval times, dicing became the favorite pastime of knights, and both dicing schools and guilds of dicers flourished.

Dice are frequently used to randomize allowable moves in board games such as **backgammon**. Loaded dice can be made in many ways to cheat at such games. Weights can be added, or some edges made round while others are sharp, or some faces made slightly off-square, to make some outcomes more likely than would be predicted by pure chance. Dice with non-cubical shapes were once almost exclusively used by fortune-tellers and in other occult practices, but they have become popular lately among players of role-playing and war-games.

difference equation

An equation that describes how something changes in discrete time steps. Numerical solutions to **integral**s are usually realized as difference equations.

differential

A term such as dx used in an expression such as $y\mathrm{d}x - x\mathrm{d}y$ to denote first-order small changes in the variable. *Differentiation* is the method by which a differential is found.

differential equation

A description of how something continuously changes over time (see **continuity**). Some differential equations have an exact *analytical solution* such that all future states can be known without simulating the time evolution of the system. However, most have a *numerical solution* with only limited accuracy. A differential equation involves the first or higher derivatives of the **function** to be solved for. If the equation only involves first derivatives, it is known as an equation of *order* one, and so on. If only nth powers of the derivatives are involved, the equation is said to have *degree n.* Equations of degree one are called *linear.* Equations in only one variable are called **ordinary differential equation**s to distinguish them from **partial differential equation**s, which have two or more.

differential geometry

The study of geometry using **calculus**; it has many applications in physics, especially in **relativity theory**. The objects studied by differential geometry are known as *Riemannian **manifold**s*. These are geometrical objects, such as surfaces, that locally look like **Euclidean space** and therefore allow the definition of analytical concepts such as tangent **vector**s and tangent space, differentiability (see **differential**), and vector and **tensor** fields. Riemannian manifolds have a **metric**, which opens the door to measurement because it allows distances and angles to be evaluated locally and concepts such as **geodesic**s, **curvature**, and torsion to be defined.

differential topology

A branch of **topology** concerned with those properties of **differential geometry** that are preserved by continuous transformation.

differentiation

The method by which the **derivative** of a **function** is found.

digimetic

A **cryptarithm** in which digits are used to represent other digits.

digit

A symbol or numeral that is used to represent an **integer** in a positional **number system**. Examples of digits include the decimal characters 0 through 9, the binary characters 0 or 1, and the hexadecimal digits 0 . . . 9, A . . . F. The word comes from the Latin *digitus* for "finger" or "toe," and retains this meaning, reminding us of the origins of our base 10 number system. The earlier Indo-European root *deik* is related to many other words that hark back to the use of the hands and fingers to "point" out objects, including index, indicate, token, and teach.

digital root

Take a number, n, add its digits, then add the digits of numbers derived from it, and so on, until the remaining number has only one digit. This single digit result is called the digital root of n. For example, in the case of 5381: $5 + 3 + 8 + 1 = 19$; $1 + 9 = 10$; $1 + 0 = 1$; thus, the digital root of 5381 is 1. See also **casting out nines**.

digraph

A **graph** in which each edge has a direction associated with it.

dihedral angle

The angle defined by two given faces meeting at an edge; for example, all the dihedral angles of a cube are 90°. An almost-spherical **polyhedron** (with many faces) has small dihedral angles.

dimension

An extension in some unique direction or sense; the word comes from the Latin *dimetiri* for "measured out." The most common way to think of a dimension is as one of the three spatial dimensions (up-down, left-right, back-forth) in which we live. Mathematicians and science fiction writers alike have long imagined what it would be like in a world with a different number of spatial dimensions. Speculation has particularly focused on **two-dimensional worlds** and, to an even greater extent, on the **fourth dimension**. **Time** is also thought of as a dimension; indeed, in **relativity theory** and as a component of **space-time**, it is treated almost exactly the same as a dimension of space. The universe may have additional spatial dimensions–a total of 10, 11, or 26 are especially favored–according to some theories of the subatomic world (see **string theory** and **Kaluza-Klein theory**), though the additional ones are "curled up" incredibly small and only become important at scales far smaller than those that can be experimentally probed today.

In mathematics, the term *dimension* is used in many different ways. Some of these correspond to the everyday idea of an extension in physical space or to some of the more esoteric meanings in physics. Others are purely abstract and exist only in certain types of theoretical, mathematical space. There are, for example, *Hamel dimensions, Lebesgue covering dimensions,* and **Hilbert space**s. So-called **Hausdorff dimension**s are used to characterize **fractal**s–mathematical objects that have fractional dimensions–by giving a precise meaning to the idea of how well something, such as an extremely "wriggly" curve or surface, fills up the space in which it is embedded.

dinner party problem

See **Ramsey theory**.

Diocles (c. 240–c. 180 B.C.)

A Greek mathematician and contemporary of **Apollonius** who studied the **cissoid** as part of an attempt to duplicate the cube and also was the first to prove the focal property of a parabolic mirror.

Diophantine approximation

The approximation of a **real number** by a **rational number**.

Diophantine equation

An equation that has integer **coefficient**s and for which integer solutions are required. Such equations are named after **Diophantus**. The best known examples are those from **Pythagoras's theorem**, $a^2 = b^2 + c^2$, when a, b, and c are all required to be whole numbers–a **Pythagorean triplet**. Despite their simple appearance Diophantine equations can be fantastically difficult to solve. A notorious example comes from **Fermat's last theorem** (recently solved), $a^n = b^n + c^n$ for $n > 2$. To give a specific example, suppose we want to find integer values for x and y such that

$$x^2 = 1620y^2 + 1.$$

A trial-and-error approach using a computer would quickly find the solution: $y = 4$, $x = 161$. However, just a slight change to the equation to make it

$$x^2 = 1621y^2 + 1$$

would leave the trial-and-error method floundering, even with the resources of the most powerful computers on Earth. The smallest integer solution to this innocent looking formula involves a y-value that is on the order of a thousand trillion trillion trillion trillion trillion trillion! One of the challenges (the tenth one) that David **Hilbert** threw down to twentieth-century mathematicians in his famous list was to find a general method for solving equations of this type. In 1970, however, the Russian mathematician Yu Matijasevic showed that there is no general algorithm for determining whether a particular Diophantine equation is soluble: the problem is undecidable.[217, 218]

Diophantus of Alexandria (A.D. c. 200–c. 284)

A Greek mathematician who developed his own algebraic notation and is sometimes called "the father of algebra." His works were preserved by the Arabs and translated into Latin in the sixteenth century, when they served to inspire momentous new advances. **Diophantine equation**s are named in his honor. It was in the margin of a French translation of Diophantus's work *Aritmetike* from c. 250 that Pierre de **Fermat** scribbled his famous comment that became known as **Fermat's last theorem**.

Diophantus's riddle

One of the oldest known age problems (see **age puzzles and tricks**). It comes from the *Greek Anthology,* a collection of puzzles compiled by Metrodorus in about A.D. 500, and purports to tell how long **Diophantus** lived in the form of a riddle engraved on his tombstone:

> God vouchsafed that he should be a boy for the sixth part of his life; when a twelfth was added, his cheeks acquired a beard; He kindled for him the light of marriage after a seventh, and in the fifth year after his marriage He granted him a son. Alas! late-begotten and miserable child, when he had reached the measure of half his father's life, the chill grave took him. After consoling his grief by this science of numbers for four years, he reached the end of his life.

If d and s are the ages of Diophantus and his son when they died, then the epitaph boils down to these two equations,

$$d = (\tfrac{1}{6} + \tfrac{1}{12} + \tfrac{1}{7})d + 5 + s + 4$$
$$s = \tfrac{1}{2}\,d,$$

which can be solved simultaneously to give $s = 42$ years and $d = 84$ years.

Dirac, Paul Adrien Maurice (1902–1984)

A British theoretical physicist who played a major role in the development of **quantum mechanics** and predicted the existence of antiparticles. He made his first great breakthrough at Cambridge University in 1928, when he found a wave equation for the electron. This explained aspects of the electron that had previously been observed but not understood, and, incidentally, is the only equation to appear in Westminster Abbey, where it is engraved on Dirac's commemorative plaque. Dirac's electron equation also made the remarkable prediction that there exists a previously unseen type of matter–a particle like the electron, but with the opposite charge. This was startling at the time because only two subatomic particles, the electron and the proton, were known, and there was no suspicion that others might be waiting in the wings. The prediction was fulfilled four years later when the *positron,* as it is now called, was first seen. A central theme of Dirac's work was his belief that **beauty and mathematics** go hand in hand. When a journalist once asked him to explain the concept of mathematical beauty, Dirac asked the journalist "Do you know mathematics?" and when the journalist replied "No," Dirac said, "Then you can't understand the concept of mathematical beauty." A shy, retiring person, Dirac is not as famous as his achievements warrant.

Dirac string trick

Take a cardboard square and tie the four corners to another larger square by loose string. Rotate the small square by 360° about a vertical axis, that is, in a horizontal plane. The strings will become somewhat tangled, and it is not possible to untangle them without rotating the square. Turn the square through another 360°, for a total of 720°. Contrary to all expectations, it is now possible to untangle the string, without further rotation of the square, simply by allowing enough space for the strings to be looped over the top of the square!

Another version of the Dirac string trick has been called the Philippine wineglass trick. A glass of water held in the hand can be rotated continuously through 720° without spilling any water. Surprisingly, these geometrical demonstrations are related to the physical fact that an

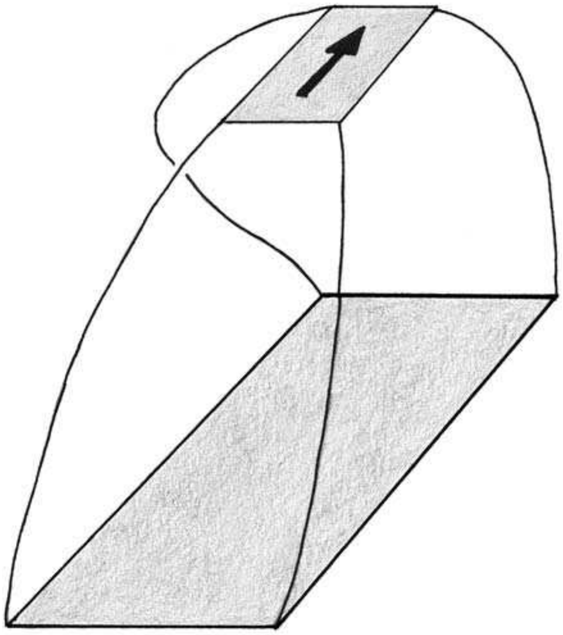

Dirac string trick The equipment needed to simulate an electron's spin property.

electron has spin ½. A particle with spin ½ is something like a ball attached to its surroundings with string. Its amplitude changes under a 360° (2π) rotation and is restored on rotation to 720° (4π).

direct proportion

The relationship two quantities have if the graph of one against the other is a straight line through the origin; so if one doubles then the other doubles, and so forth.

directed graph

Also known as a *digraph,* a **graph** in which each edge is replaced by a directed edge, indicated by an arrow. A directed graph having no multiple edges or loops is called a *simple directed graph.* A **complete graph** in which each edge is bidirected is called a complete directed graph. A directed graph having no symmetric pair of directed edges (i.e., no bidirected edges) is known as an *oriented graph.* A complete oriented graph (i.e., a directed graph in which each pair of vertices is joined by a single edge having a unique direction) is called a *tournament.*

directrix

The line that, together with a point called the **focus**, serves to define a **conic section** as the **locus** of points whose distance from the focus is proportional to the horizontal distance from the directrix. If the ratio $r = 1$, the conic is a **parabola**, if $r < 1$, it is an **ellipse**, and if $r > 1$, it is a **hyperbola**.

Dirichlet, Peter Gustav Lejeune (1805–1859)

A German mathematician who made significant contributions to **number theory**, **analysis**, and mechanics, and who is credited with the modern formal definition of a **function**. He taught at the universities of Breslau (1827) and Berlin (1828–1855) and in 1855 succeeded Carl **Gauss** at the University of Göttingen but died of a heart attack only three years later. Dirichlet continued Gauss's great work on number theory, publishing on **Diophantine equation**s of the form $x^5 + y^5 = kz^5$. His book *Lectures on Number Theory* (published posthumously in 1863) is similar in stature to Gauss's earlier *Disquisitiones* and founded modern **algebraic number theory**. In 1829 he gave the conditions sufficient for a **Fourier series** to converge (though the conditions *necessary* for it to converge are still undiscovered).

Dirichlet's theorem

For any two positive **coprime** integers, a and b, there are infinitely many **prime number**s of the form $an + b$, where $n > 0$. This theorem was first conjectured by Karl **Gauss** and proved by Peter **Dirichlet** in 1835.

discontinuity

Also called a *jump,* a point at which a **function** is not **continuous**.

discrete

Taking only non**continuous** values, for example, **Boolean** or **natural number**s.

discriminant

A quantity that gives valuable information about the solutions of an equation. In the case of the **quadratic** equation $ax^2 + bx + c = 0$, the discriminant is given by $d = b^2 - 4ac$. If $d > 0$, the roots of the equation are two different **real number**s; if $d = 0$, the roots are real and equal; if $d < 0$, the roots are **complex number**s. The concept of discriminant can also be applied in the case of **polynomial**s, **elliptic curve**s, and **metric**s.

disk

Roughly speaking, the "filling" of a **circle**. A flat (two-dimensional) disk of radius r consists of all the points that are at a distance $\leq r$ (*closed disk*) or $< r$ (*open disk*) from a fixed point in the plane. More generally, an n dimensional disk of radius r is the set of all points at a distance $\leq r$ (*closed*) or $< r$ (*open*) from a fixed point in

Euclidean *n*-space. A disk is the two-dimensional analog of a **ball**.

disme

An old word for "tenth." A notation for **decimal** fractions was introduced for the first time in 1585 in a pamphlet called *La Disme* by Simon Stevin of Holland.

dissection

Cutting apart one or more figures and rearranging the pieces to make another figure. Dissection puzzles have been around for thousands of years. The problem of dissecting two equal squares to form one larger square using four pieces dates back to at least the time of Plato (427–347 B.C.). In the tenth century, Arabian mathematicians described several dissections in their commentaries on Euclid's *Elements.* The eighteenth-century Chinese scholar Tai Chen presented an elegant dissection for approximating the value of **pi**. Others worked out dissection proofs of the **Pythagoras's theorem**. In the nineteenth century, dissection puzzles by Sam **Loyd**, Henry **Dudeney**, and others became tremendously popular in magazine and newspaper columns. A classic example is the **Haberdasher's puzzle**. Dissections can get quite elaborate: an eight-piece octahedron becomes a hexagon, a nine-piece five-pointed star becomes a pentagon, and so on. See also **tangrams** and **Loculus of Archimedes.**[201, 202]

dissipative system

A **dynamical system** that contains internal friction that deforms the structure of its **attractor**. Dissipative systems often have internal structure despite being far from **equilibrium**, like a whirlpool that preserves its basic form despite being in the midst of constant change.

distortion illusion

An illusion that distorts an image's shape and/or size. Famous examples include **Poggendorff illusion**, **Zöllner illusion**, **Titchener illusion**, **irradiation illusion**, **Fraser spiral**, **Müller-Lyer illusion**, **Orbison's illusion**, **vertical-horizontal illusion**, and **Ames room**.

distributive

Three numbers $x, y,$ and z are said to be distributive over the operation + if they obey the identity

$$x(y+z) = xy + xz.$$

Compare with **associative** and **commutative**.

diverge

If a **sequence** doesn't converge it is said to diverge (see **convergence**). This can be if it goes to infinity, or if it simply cycles between two or more values without ever staying on one of them. For example, the sequences: 1, 2, 4, 8, 16, 32, . . . and 1, 0, 1, 0, 1, 0, . . . are both divergent.

division

A counterpart to **multiplication** defined so that if

$$a \times b = c$$

where b is nonzero, then

$$a = c/b.$$

In this equation, a is the *quotient,* b is the *divisor,* and c is the *dividend.* A **skeletal division** is a long division in which most or all of the digits are replaced by a symbol (usually asterisks) to form a **cryptarithm**.

dodecagon

A **polygon** with 12 sides.

dodecahedron

A **polyhedron** with 12 faces. A *regular dodecahedron* is made from faces that are identical regular **pentagon**s and is one of the **Platonic solid**s.

dodecahedron A mechanical puzzle in the form of a dodecahedron. *Mr. Puzzle Australia, www.mrpuzzle.com.au*

Dodgson, Charles Lutwidge
See **Carroll, Lewis**.

dollar
The elaborate designs on the various denominations of American dollar bills can be used for some amusing games of the "Can you find . . . ?" variety. On a $1 bill, there is an owl in the upper left-hand corner of the "1" encased in the shield, and a spider hidden in the front upper right-hand corner. There are also at least nine occurrences of **thirteen** things: 13 steps on the pyramid, 13 letters in the Latin above the pyramid, 13 letters in "E Pluribus Unum," 13 stars above the eagle, 13 plumes of feathers on each span of the eagle's wing, 13 bars on the shield, 13 leaves on the olive branch, 13 fruits, and 13 arrows. On the back of a $5 bill the number 172 can be found in the bushes at the base of the Lincoln Memorial, while on the back of the old $10 bill are four cars and eleven light posts. On the new $100 dollar bill the time on the clock tower of Independence Hall reads 4:10.

domain
The set of numbers x for which the **function** $f(x)$ is defined. See also **codomain** and **range**.

domino
A small rectangular tile, marked with spots, that is used to play games. There are many varieties of dominoes and games based on them. Most domino tiles, however, have roughly 2-to-1 proportions and each half of each domino has spots arranged as on a six-sided die; the set generally contains tiles with all possible combinations of two numbers. A tile is identified by the number of spots on each half: for example, "1-6" or "3-3." English/American dominoes include blank sides; Chinese dominoes don't but do duplicate some whole tiles. English/American dominoes can also be bought in larger sets with numbers of spots up to nine or twelve per side. Other than games of strategy, there are many mathematical puzzles that involve dominoes. Some of these puzzles involve tiling variations on the standard 8 × 8 chessboard.

Another common pastime using domino tiles is to stand them on edge in long lines, then topple the first tile, which falls on and topples the second, and so forth, resulting in all of the tiles falling. Arrangements of thousands of tiles have been made that take several minutes to fall. By analogy, phenomena of chains of small events each causing similar events leading to eventual catastrophe are called *domino effects*. The word *domino* was first used to refer to the hooded black cape worn by priests, and later to black masks (of the Lone Ranger type) worn at masquerade balls. The domino is the simplest form of **polyomino**.

PUZZLE

A standard chessboard can easily be tiled by using four dominos in each row. But what if two squares are removed, one each from diagonally opposite corners of the chessboard? Can this reduced board be completely tiled by nonoverlapping dominoes?

Solutions begin on page 369.

domino problem
Is there an algorithm (a set of instructions) that, when given a particular shape as an input, decides if the shape can be used to tile the entire plane? The solution to this unresolved problem is tied up with **Heesch number**s. The domino problem in turn has a deep connection with the **Einstein problem**.

dozen
See **twelve**.

dragon curve
A classic example of a recursively generated **fractal** shape. Benoit **Mandelbrot** called it the "Harter-Heighway" dragon curve and it formed the subject of one of Martin

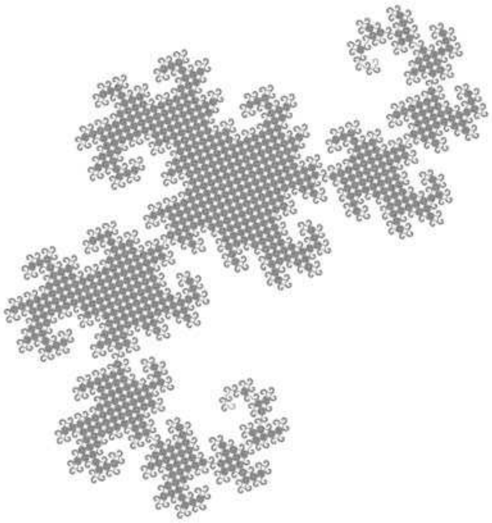

dragon curve *Jos Leys, www.josleys.com*

Gardner's *Mathematical Games* columns in *Scientific American* in 1967.[116] The dragon curve fills out an "island" of positive area with a fractal boundary.

dual

(1) The *dual of a solid* is formed by joining the centers of adjacent faces with straight lines. In the resulting dual solid, each **vertex** of the dual corresponds to a face on the original, each face on the dual to an original vertex, while the edges match, one for one. (2) The *dual of a **tessellation*** is obtained by replacing each tile with a point at its center, and each edge between tiles with an edge joining vertices. The dual of a regular tessellation is a regular tessellation; the dual of a semi-regular tessellation is not semi-regular.

Dudeney, Henry Ernest (1857–1930)

An English writer and puzzle-maker who became one of the greatest exponents of recreational mathematics of his time. **Chess** and chess problems captivated him from an early age and he was only 9 when he started contributing puzzles to a local newspaper. His education was limited and he started work as a clerk in the civil service at the age of 13. However, he kept up his interest in math and chess, wrote articles for magazines under the pseudonym "Sphinx," and joined a literary circle that included Arthur Conan Doyle. In 1893 he struck up a correspondence with the American puzzle-maker Sam **Loyd**, the other leading mathematical recreationist of the day, and the two shared many ideas. However, a rift developed after Dudeney accused Loyd of publishing many of Dudeney's puzzles under his own name. One of Dudeney's daughters "recalled her father raging and seething with anger to such an extent that she was very frightened and, thereafter, equated Sam Loyd with the devil." Dudeney was a columnist for the *Strand Magazine* for over 30 years and wrote six books. The first of these, *The Canterbury Puzzles,*[87] published in 1907, purports to include a collection of problems posed by the characters in Chaucer's *The Canterbury Tales.* The answer to the so-called **Haberdasher's puzzle** is Dudeney's best known geometrical discovery. His other books include *Amusements in Mathematics* (1917)[88] and *The World's Best Word Puzzles* (1929). See also **spider-and-fly problem** and **polyomino**.[233]

Dunsany, Lord Edward Plunkett (1878–1957)

An Irish writer who was one of the founders of the fantasy genre of literature. Edward John Moreton Drax Plunkett was born in London to a family whose roots in Ireland predate the Norman invasion. He inherited his father's title in 1899, fought in the Boer War, and returned to the ancestral home, Dunsany Castle, in 1901. Lord Dunsany was a keen marksman and hunter, a fine player of cricket (Dunsany had its own cricket ground near the village), tennis (there is a court beside the Castle), and **chess** (he was an amateur champion and once drew with Grand Master Capablanca. He also wrote chess puzzles for the *Times* over many years and invented his own variant of the game. His first of many books, *The Gods of Pegana,* was published in 1905. In writings that spanned fantasy, drama, poetry, and science fiction, he was an early explorer of such ideas as chess-playing computers (in "The Three Sailors' Gambit" from *The Last Book of Wonder* and, again, in his 1951 novel *The Last Revolution*) and paradoxes in **time travel** (e.g., in "Lost" from *The Fourth Book of Jorkens* and "The King That Was Not" from *Time and the Gods*).

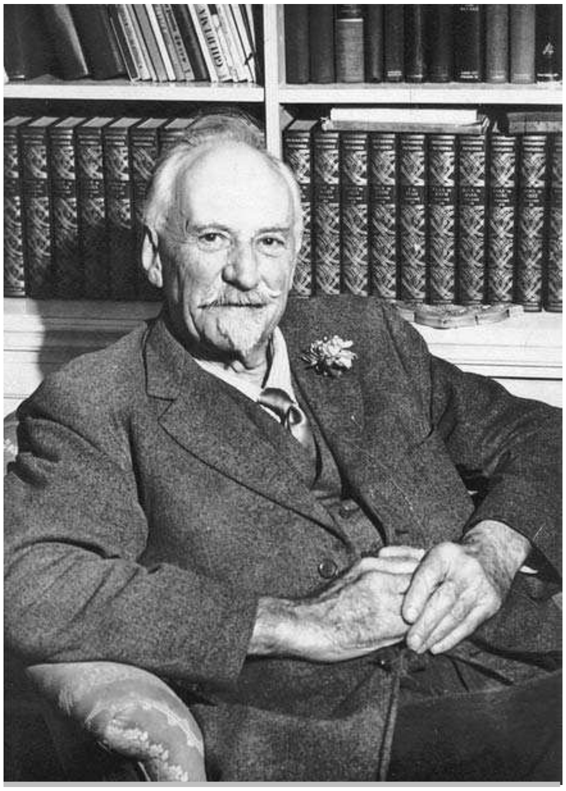

Dunsany, Lord Edward Plunkett *The Dunsany Estate*

Dupin cyclide

The envelope of all **sphere**s touching three given fixed spheres. (Each of the fixed spheres is to be touched in an assigned manner, either externally or internally.) Equivalently, the envelope of all spheres whose centers lie on a given **conic section** and which touch a given sphere. Also equivalently, the inverse of a **torus**.

duplicating the cube

A classical mathematical challenge of antiquity: using only a straightedge and compass, construct a **cube** whose volume is exactly twice that of a given cube. It is often called the "Delian" problem because of a legend that surrounds its origin. The citizens of Athens were being devastated by a plague so that, in 430 B.C., they sought advice from the oracle at Delos on how to rid their community of this pestilence. The oracle replied that the altar of Apollo, which was in the form of a cube, had to be doubled in size. Thoughtless builders merely doubled the edges of the cube, failing to appreciate that this increased the volume of the altar eightfold. The oracle said the gods had been angered; the plague grew worse. Other delegations consulted Plato. When informed of the oracle's admonition, Plato told the citizens "the god has given this oracle, not because he wanted an altar of double the size, but because he wished in setting this task before them, to reproach the Greeks for their neglect of mathematics and their contempt of geometry." Many Greek mathematicians attacked the problem. All failed, because the so-called *Delian constant,* $\sqrt[3]{2}$ (the required ratio of sides of the original cube and that to be constructed), needed for the duplication can't be constructed as prescribed. Cube duplication is possible, however, using a **Neusis construction**. See also **cissoid** of Diocles.

Dürer, Albrecht (1471–1528)

> *Sane judgment abhors nothing so much as a picture perpetrated with no technical knowledge, although with plenty of care and diligence. Now the sole reason why painters of this sort are not aware of their own error is that they have not learnt Geometry, without which no one can either be or become an absolute artist.*
>
> –from *The Art of Measurement,* 1525

A German printmaker who, through applying mathematics to art, brought important ideas to mathematics itself especially in the area of perspective geometry. Dürer was born in Nüremberg, one of 18 children, and showed an early talent for art. After a four-year apprenticeship in painting and woodcutting, he began traveling Europe, especially Italy, in search of new styles and ideas. Back in Nüremberg, he began a serious study of mathematics, absorbing *Elements* by **Euclid** and *De architectura* by the great Roman architect Vitrivius, and studying the work of Leone Alberti (1404–1472) and Luca **Pacioli** on mathematics and art, in particular their work on proportion. His mastery of perspective is clear in woodcuts *Life of a Virgin* (1502–1505). In about 1508, Dürer began to collect material for a major work on mathematics and its applications to the arts. This work was never finished but Dürer did use parts of the material in later published work. One of his most famous engravings *Melancholia,* produced in 1514, contains the first **magic square** to be seen in Europe, cleverly including the date 1514 as two entries in the middle of the bottom row. Also of mathematical interest in *Melancholia* is the **polyhedron** in the picture, the faces of which appear to consist of two equilateral triangles and six somewhat irregular pentagons. In 1825 Dürer published a four-volume treatise, *Underweysung der Messung* (available in English translation as *Painter's Manual*), which dealt with, among other things, the construction of various curves, polygons, and other solid bodies. One of the first books to teach the methods of perspective, it was highly regarded throughout the sixteenth century and presents the earliest known examples of polyhedral **net**s, that is, polyhedra unfolded to lie flat for printing.

Dürer traveled to Italy to learn about perspective and was keen to publish the methods so they weren't kept secret among a few artists. Who he learned from is not known, but Luca Pacioli is a likely possibility. Some of

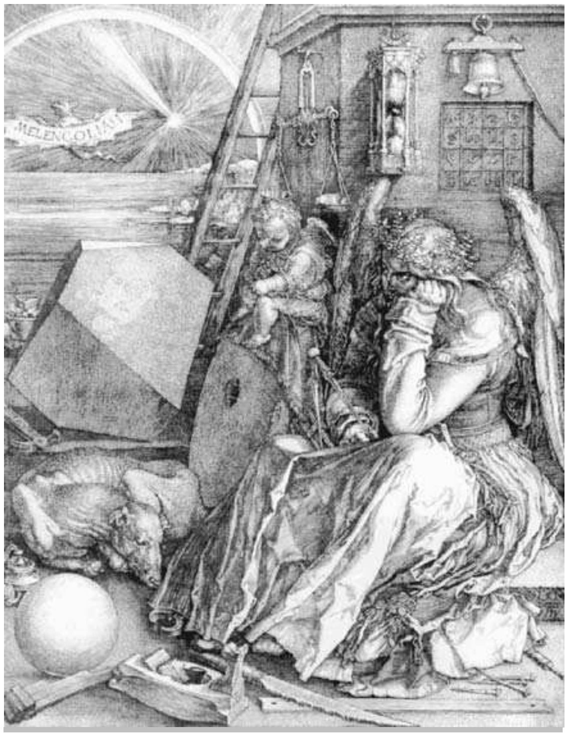

Dürer, Albrecht The famous woodcut *Melancholia,* by Albrecht Dürer, features a magic square and an unusual polyhedron.

the techniques and illustrations also follow closely the work of Piero della Francesca.

Dürer's final work, his *Treatise on Proportion,* was published posthumously and laid the groundwork for descriptive geometry and its rigorous mathematical treatment by Gaspard **Monge**.

Dürer's shell curve

Given a **parabola** and a line that is **tangent** to the parabola, the **glissette** of a point on a line sliding between the parabola and the tangent. It has the equation $(x^2 + xy + ax - b^2) = (b^2 - x^2)(x - y + a)^2$.

dynamical system

A **nonlinear**, interactive system that evolves over time, showing transformations of behavior and an increase in **complexity**. Key to this evolution is the presence and emergence of **attractor**s, most notably **chaotic attractor**s. The changes in the system's organization and behavior are known as **bifurcation**s. Dynamical systems are **deterministic system**s, although they can be influenced by random events. Times series data of dynamical systems can be graphed as phase portraits in **phase space** in order to indicate the qualitative or topological properties of the system and its attractors. For example, various physiological systems, such as the heart, can be conceptualized as dynamical systems. Seeing physiological systems as dynamical systems opens up the possibility of studying various attractor regimes. Moreover, certain diseases can be understood now as "dynamical diseases," meaning that their temporal phasing can be a key to understanding pathological conditions.

dynamics

Pertaining to the change in behavior of a system over time.

E

e

> *Pi goes on and on and on . . .*
> *And e is just as cursed.*
> *I wonder: Which is larger*
> *When their digits are reversed?*
>
> –Martin Gardner

Possibly the most important number in mathematics. Although **pi** is more familiar to the layperson, *e* is far more significant and ubiquitous in the higher reaches of the subject. One way to think of *e* is as the number of dollars you would have in the bank at the end of a year if you invested $1 at the start of the year and the bank paid an annual interest rate of 100% compounded *continuously.* Compound interest doesn't behave in quite the way intuition suggests. Because more frequent compounding causes the principal to grow faster, it might seem that continuous compounding would make the investor very rich in short order. But the effect tails off. At the end of one year, the $1 would have grown to a mere $2.72, rounded to the nearest cent. To a better approximation, *e* is 2.718281828459045 . . . , its decimal expansion stretching out forever, never repeating in any permanent pattern, because *e* is a **transcendental number**. It is the base of natural **logarithm**s, which is equivalent to the fact that the area under the curve (the integral of) $y = 1/x$ between $x = 1$ and $x = e$ is exactly equal to one unit. It also features in the **exponential** function $y = e^x$, which is unique in that its value (y) is exactly equal to its growth rate (dy/dx in calculus notation) at every point. As well as showing up in problems involving growth or decay (including compound interest) or in calculus, whenever logarithmic or exponential functions are involved, *e* is at the heart of the statistical **bell curve**; the shape of a hanging cable, known as a **catenary**; the study of the distribution of **prime number**s; and Stirling's formula for approximating **factorial**s.

Like π, *e* pops up as the limit of many **continued fraction**s and **infinite series**. Leonhard **Euler**, who was the first to study and to use the symbol *e* (in 1727), found it could be expressed as the curious fraction:

$$1+\cfrac{1}{0+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{2+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{4+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{6+\dots}}}}}}}}}}$$

No less remarkable is this infinite series of which *e* is the sum:

$$1 + 1/1! + 1/2! + 1/3! + 1/4! + \dots$$

But of all the places that *e* appears in mathematics none is more extraordinary than **Euler's formula** from which comes the most profound relationship in mathematics: $e^{i\pi} + 1 = 0$, linking *e* and π with **complex numbers**.[215]

Earthshapes

A series of 12 hypothetical Earths as conceived by American airman Joseph Portney in 1968 during a flight to the North Pole onboard a U.S. Air Force KC-135. As the North Pole was reached, Portney looked on the icy terrain below and asked himself, "What if the Earth were . . . ?" The hypothetical Earths, cylindrical, conic, donut-shaped, and so forth, were sketched and captioned by Portney and given to the Litton Guidance & Control Systems graphic arts group to create models. These models were then photographed and became the theme of a Litton publication entitled *Pilots and*

Navigators Calendar for 1969. Each month was introduced with a different one of the 12 hypothetical Earths. The result was an international sensation, attracting awards and heavy fan mail.

eccentricity

See **conic section**.

economical number

A number that has no more digits than there are digits in its prime factorization (including powers). If a number has fewer digits than are in its prime factorization it is known as a *frugal number.* The smallest frugal is 125, which has three digits, but can be written as 5^3, which has only two. The next few frugals are 128 (2^7), 243 (3^5), 256 (2^8), 343 (7^3), 512 (2^9), 625 (5^4), and 729 (3^6). An *equidigital number* is an economical number that has the same number of digits as make up its prime factorization. The smallest equidigitals are 1, 2, 3, 5, 7, and 10 ($= 2 \times 5$). All **prime number**s are equidigital. An *extravagant number* is one that has fewer digits than are in its prime factorization. The smallest extravagant number is 4 ($= 2^2$), followed by 6, 8, and 9. There are infinitely many of each of these kinds of numbers. Are there also arbitrarily long sequences of consecutive ones? Seven-member strings of consecutive economical numbers start at each of 157; 108,749; 109,997; 121,981; and 143,421. On the other hand, the longest string of consecutive frugal numbers up to 1 million is just two (for example, 4374 and 4375). Even so, it has been proved that if a certain conjecture about prime numbers known as *Dickson's conjecture* is true, then there are arbitrarily long strings of frugals.

Eddington number

"I believe there are 15,747,724,136,275,002,577,605,653, 961,181,555,468,044,717,914,527,116,709,366,231,425, 076,185,631,031,296 protons in the universe and the same number of electrons." So wrote the English astrophysicist Sir Arthur Eddington (1882–1944) in his book *Mathematical Theory of Relativity* (1923). Eddington arrived at this outrageous conclusion after a series of convoluted (and wrong!) calculations in which he first "proved" that the value of the so-called fine-structure constant was exactly $^1/_{136}$. This value appears as a factor in his prescription for the number of particles (protons + electrons; neutrons were not discovered until 1930) in the universe: $2 \times 136 \times 2^{256} = 17 \times 2^{260} \cong 3.149544 \ldots \times 10^{79}$ (double the number written out in full in the quote above). This is the Eddington number, notable for being the largest *specific* integer (as opposed to an estimate or approximation) ever thought to have a unique and tangible relationship to the physical world. Unfortunately, experimental data gave a slightly lower value for the fine-structure constant, closer to $^1/_{137}$. Unfazed, Eddington simply amended his "proof" to show that the value had to be exactly $^1/_{137}$, prompting the satirical magazine *Punch* to dub him "Sir Arthur Adding-One." See also **large number**.

edge

A line segment where two faces meet. A **cube**, for example, has 12 edges.

edge coloring theorem

See **Tait's conjecture**.

edge of chaos

The hypothesis that many natural systems tend toward dynamical behavior that borders static patterns and the chaotic regime. See also **chaos**.

egg

Specifically, a chicken's egg and its mathematical equivalent. Eggs are often described as being **oval** in shape, which is effectively tautological since "oval" comes from the Latin *ovus* for "egg." Strictly speaking, an oval is a flat two-dimensional curve, so it is more accurate to say that an egg is shaped like the **surface of revolution** of an oval. In real life, eggs, like ovals, come in a variety of forms all of which can be loosely described as "like an **ellipsoid** but with one end more pointed than the other." Because eggs vary in shape, so too do their mathematical descriptions. Having said this, there are a variety of ways to approximate the shape of a hen's egg by modifying the equation of an ellipsoid, $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$, so as to

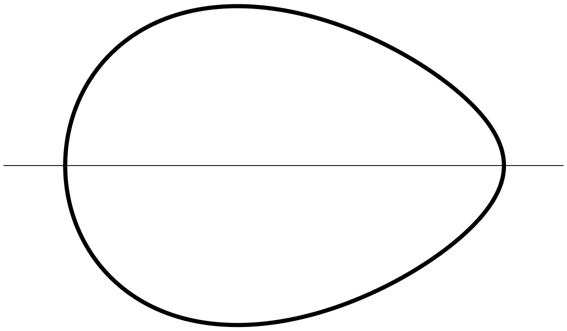

egg A good egg shape is obtained by drawing four circle arcs of different radii. *© Jan Wassenaar, www.2dcurves.com* introduce an asymmetry about the long (say, z-) axis. These involve multiplying z^2/c^2 by a suitable term, so that y becomes larger on the right side of the y-axis and smaller on the left side. For example, $x^2/a^2 + y^2/b^2 + z^2/c^2 (1 - kx) = 1$ gives a good egg. Other useful egg approximations come from surfaces of revolution of **Cartesian ovals**, **Cassinian ovals**, and sections through cones and cylinders. In France, where tennis first became popular, a zero on the scoreboard looked like an egg and was called *l'oeuf,* which is French for "egg." When tennis was introduced in the United States, Americans pronounced it "love."

Why is a hen's, or other bird's, egg shaped as it is? Because it gives strength even though the eggshell is thin enough to allow the young bird to peck its way out when ready. To demonstrate this strength, try balancing a pile of books on four half egg shells. It is even possible for a person's weight to be supported in this way.

Another trick with eggs is to distinguish between a raw egg and a hardboiled one without cracking them open to see which is which. Lay both eggs on their sides on a table, and spin them as you would a spinning top. With a bit of practice, the cooked egg will be made to rise up for a few seconds, while the raw one will remain on its side. The physics of this odd behavior was finally cracked by two mathematicians, Keith Moffat of Cambridge University and Yutaka Shimomura of Keio University, who reported their findings in 2002. They concluded that friction between the egg and the surface produces a gyroscopic effect, which causes some of the kinetic energy of the object to be translated into potential energy, raising its center of gravity (see also **Tippee Top**). As the hardboiled egg spins, its curved surface causes it to touch the tabletop at only one point. The contact point changes and traces out a little circle. If the texture of the tabletop is just right (neither too slippery nor too sticky) the egg will slide a bit as it spins. This sliding slows the spin a bit and causes a wobble. This in turn tilts the egg, lifting one end off the table more than the other, at which point the gyroscopic effect kicks in and swaps some of the kinetic energy of the spinning egg into potential energy and raises its center of gravity in a seemingly paradoxical way. This effect is heightened by the fact that as the end of the egg rises, the egg draws in closer to the axis of spin, causing it to spin more quickly—just as figure-skaters can make themselves pirouette faster by raising their arms above their heads. Why doesn't the effect occur with a raw egg? Because the inside of the egg is runny and it lags behind the shell. This lag serves as a drag, which reduces the spin rate and dissipates the egg's kinetic energy. This in turn reduces the friction between the egg and tabletop, and means that not enough energy is available to be turned into potential energy to raise the egg's center of gravity. As well as solving the mystery of the balancing egg, Moffat also found time to write a limerick to commemorate the event:

Place a hard-boiled egg on a table,
And spin it as fast as you're able;
It will stand on one end
With vectorial blend
Of precession and spin that's quite stable.

See also **superegg**.

Egyptian fraction

A unit **fraction**; in other words, a fraction in which the numerator (the number on top) is one. This type of fraction was the only kind used by the ancient Egyptians and appears extensively in the **Rhind papyrus**. Other fractions can be obtained by adding Egyptian fractions together; for example, $5/7 = 1/2 + 1/6 + 1/21$. In 1201 **Fibonacci** proved that every **rational number** can be written as a sum of Egyptian fractions.

eigenvalue

A **complex number**, λ, that satisfies the equation $Ax = \lambda x$, where A is an $n \times n$ **matrix** and x is some **vector**. In this case, x is called an *eigenvector.*

eight

The second smallest cube number (after 1^3): $8 = 2^3 = 2 \times 2 \times 2$. A queen or king in **chess** can move in eight different directions, in the same way that a compass has eight principal points: north, northeast, east, southeast, south, southwest, west, and northwest. In three dimensions, there are eight diagonal ways to move, corresponding to the eight *octants* into which three-dimensional space is divided by three mutually perpendicular planes. Add a **fourth dimension** and movement becomes possible back and forth along four directions at right angles to each other: up and down, left and right, forward and back, and to one other! The Spanish dollar was a gold coin with a value of eight reales, and was sometimes actually cut into eight wedge-shaped pieces—"pieces of eight"—to make change.

eight curve

A curve, also known as the *lemniscate of Gerono,* that has the Cartesian equation

$$x^4 = a^2(x^2 - y^2)$$

and the appearance of a figure eight lying on its side.

Einstein problem

(1) Is there a single shape that will tile a plane aperiodically (see **aperiodic tiling**)? An answer of "no" would imply the existence of a decision method for the **domino problem**. This problem, which is named because of the German translation (*ein* = "one," *stein* = "stone") and was not an invention of the famous scientist, remains unsolved. (2) A logic problem, invented by Albert Einstein, who claimed that 98% of the people in the world couldn't solve it.

PUZZLES

1. There are 5 houses (along a street) in 5 different colors: blue, green, red, white, and yellow.
2. In each house lives a person of a different nationality: Briton, Dane, German, Norwegian, and Swede.
3. These 5 owners drink a certain beverage: beer, coffee, milk, tea, or water; smoke a certain brand of cigar: Blue Master, Dunhill, Pall Mall, Prince, or blend; and keep a certain type of pet: cat, bird, dog, fish, or horse.
4. No owners have the same pet, smoke the same brand of cigar, or drink the same beverage.
5. The Briton lives in a red house. The Swede keeps dogs as pets. The Dane drinks tea. The green house is on the left of the white house (next to it). The green house owner drinks coffee. The person who smokes Pall Mall rears birds. The owner of the yellow house smokes Dunhill. The man living in the house right in the center drinks milk. The Norwegian lives in the first house. The man who smokes blend lives next to the one who keeps cats. The man who keeps horses lives next to the man who smokes Dunhill. The owner who smokes Blue Master drinks beer. The German smokes Prince. The Norwegian lives next to the blue house. The man who smokes blend has a neighbor who drinks water.

The question is: Who keeps the fish?

Solutions begin on page 369.

elementary function

Any real-value algebraic function or transcendental function (trigonometric, hyperbolic, exponential, logarithmic).

eleven

A **palindromic number**, the smallest integer that is not a **Harshad number**, a **prime number** that is a member of a **twin prime** (11 and 13), and the largest integer that is not the sum of two or more distinct primes. There are 11 players on a soccer team and on a cricket team. Strange but true: the youngest pope was 11 years old.

ellipse

A shape that looks like a squashed circle. It is one of the **conic section**s and can be defined as the **locus** of all points in a plane that have the same sum of distances from two given fixed points known as *foci.* If the two foci coincide then the ellipse is a **circle**. The line segment connecting the foci is called the *major axis* of the ellipse; half this is the *semimajor axis*, a. The line passing through the center of the ellipse (the midpoint of the foci) at right angles to the major axis is called the *minor axis*, half of which is the *semiminor axis*, b. An ellipse centered at the origin of an x-y coordinate system with its major axis along the x-axis is defined by the equation

$$x^2/a^2 + y^2/b^2 = 1.$$

The shape of an ellipse is expressed by a number called the *eccentricity*, e, which is related to a and b by the formula $b^2 = a^2(1 - e^2)$. The eccentricity is a positive number less than 1, or 0 in the case of a circle. The greater the eccentricity, the larger the ratio of a to b, and therefore the more elongated the ellipse. The distance between the foci is $2ae$. The area enclosed by an ellipse is πab. The circumference of an ellipse is $4aE(e)$, where the function E is the complete elliptical integral of the second kind.

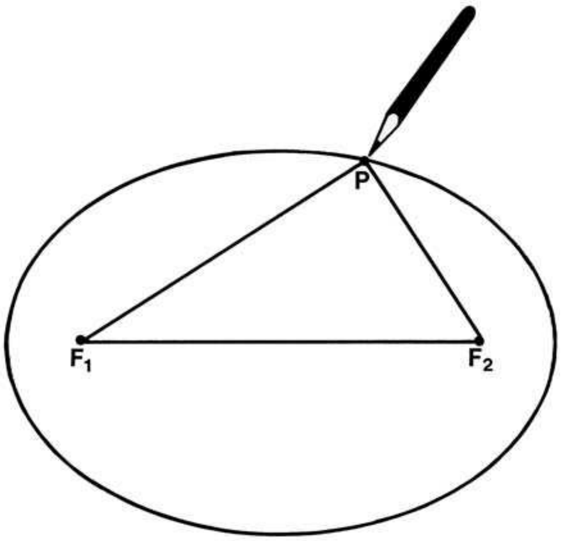

ellipse Nails mark the foci, F_1 and F_2, of an ellipse that is drawn by pencil whose moving tip, P, keeps the string threaded around the nails taut.

ellipsoid

A **quadratic** surface that is the three-dimensional analog of an **ellipse**. The general equation of an ellipsoid in Cartesian coordinates is

$$x^2/a^2 + y^2/b^2 + z^2/c^2 = 1,$$

where a, b, and c are positive real numbers determining the shape. If two of these numbers are equal, the ellipsoid is a **spheroid**; if all three are equal, it is a **sphere**. The intersection of an ellipsoid with a plane is a single point or an ellipse. Ellipsoids can also be defined in higher dimensions.

elliptic curve

The set of solutions to a type of **cubic equation** whose solutions lie on a **torus** (a donut-shaped surface). The particular type of cubic equation whose solutions lead to elliptic curves takes the form

$$y^2 + axy + by = x^3 + cx^2 + dx + e.$$

Elliptic curves, which are said to have a **genus** of 1, have an unusually rich theory and structure, and their study is linked to many other important areas of mathematics and their applications. For example, it was work done on elliptic curves by Andrew **Wiles** that finally led to a proof of **Fermat's last theorem**.

elliptic function

In **complex analysis**, a **function** defined on the complex plane that is **periodic** in two directions. The elliptic functions can be thought of as analogs of the **trigonometric functions** (which have only a single period). Leading eighteenth-century mathematicians, including Leonhard **Euler** and Joseph **Lagrange**, had studied elliptic integrals, such as the integral that gives the arc length of an ellipse; however, these cannot be expressed in terms of the elementary functions (polynomials, exponentials, and trigonometric functions). It was the insight of Karl **Jacobi**, and also of Karl **Gauss** and Niels **Abel**, that the *inverse* functions of elliptic integrals are much easier to study. They turn out to be doubly periodic functions of a complex variable. While a singly periodic function like sine has a number a (specifically $a = 2\pi$) so that $\sin(x + a) = \sin(x)$, a doubly periodic function f has the property that there are two numbers a, b, not rational multiples of each other, so that $f(x + a) = f(x + b) = f(x)$. As Jacobi proved in 1834, the ratio a/b is necessarily an imaginary number.

elliptical geometry

One of the two most important types of **non-Euclidean geometry**: the other is **hyperbolic geometry**. In elliptical geometry, Euclid's **parallel postulate** is broken because no line is parallel to any other line. The original form of elliptical geometry, known as *spherical geometry* or *Riemannian geometry*, was pioneered by Bernhard **Riemann** and Ludwig **Schläfli** and treats lines as **great circle**s on the surface of a sphere. The most familiar example of such circles, which are **geodesic**s (shortest routes) on a spherical surface, are the lines of longitude on Earth. In spherical geometry any two great circles always intersect at exactly two points. Two lines of longitude, for example, meet at the North and South Poles. Working in spherical geometry produces some surprising, nonintuitive results. For instance, it turns out that the shortest flying distance from Florida to the Philippine Islands is a path across Alaska—even though the Philippines are at a more southerly latitude than Florida! The reason is that Florida, Alaska, and the Philippines lie on the same great circle and so are collinear in spherical geometry. Another odd property of spherical geometry is that the sum of the angles of a triangle is greater than 180°. This is always the case on a surface that bulges out or, in mathematical parlance, has positive **curvature**. It was Felix **Klein** who first saw clearly how to rid spherical geometry of its one blemish: the fact that two lines have not one but two common points. He redefined the notion of a point as a **set** of antipodal points. With this definition, any two points determine a unique line so that the traditional form of Euclid's first postulate is restored. Thus modified, spherical geometry became what Klein called elliptical geometry.

embedding

Putting one mathematical object inside another, such as a subgroup within a group or one topological space inside another, while preserving all topological properties.

emergence

The arising of new, unexpected structures, patterns, or processes in a self-organizing system (see **self-organization**). These *emergents* have their own rules, laws, and possibilities, and can be understood as existing on a higher level than that of the components from which they came. The term was first used by the nineteenth-century philosopher G. H. Lewes and came into greater currency in the scientific and philosophical movement known as *emergent evolutionism* in the 1920s and 1930s.

emirp

A **prime number** that becomes a different prime number when its digits are reversed ("emirp" is "prime" spelled backward). The first twenty emirps are 13, 17, 31, 37, 71, 73, 79, 97, 107, 113, 149, 157, 167, 179, 199, 311, 337, 347, 359, and 389. Compare with a *palindromic prime* (see **palindromic number**), which gives the same prime when reversed.

empty set

The **set**, denoted by ∅ or {}, that has no members; also known as the *null set.* This is not the same as **zero**, which is the number of members of ∅. Nor is ∅ the same as **nothing** because a set with nothing in it is still a set, and a set is something. The empty set, for example, is the set of all triangles with four sides, the set of all numbers that are bigger than nine but smaller than eight, and the set of all opening moves in chess that involve a king. Applying the concept of the empty set helps distinguish between the different ways that "nothing" is used in everyday language. In his book *What Is the Name of This Book?* (1978), Raymond **Smullyan** wrote:[304] "Which is better, eternal happiness or a ham sandwich? It would appear that eternal happiness is better, but this is really not so! After all, nothing is better than eternal happiness, and a ham sandwich is certainly better than nothing. Therefore a ham sandwich is better than eternal happiness."

What is wrong with this declaration? The first statement is equivalent to "The set of things that are better than eternal happiness is ∅." The second statement is equivalent to "The set {ham sandwich} is better than the set ∅." The confusion arises because the first is comparing individual things, while the second is comparing sets of things, and ∅ plays a different role in each.

enantiomorph

The mirror image of a given **chiral** polyhedron or other figure.

enormous theorem

The largest theorem in mathematics; it concerns the classification of finite **simple group**s and encapsulates the work of hundreds of mathematicians over many years.

entropy

A measure of a system's degree of **random**ness or disorder.

envelope

A curve or a surface that touches every member of a family of lines, curves, planes, or surfaces.

epicycloid

The path traced out by a point on the circumference of a circle of radius b rolling on the outside of a **circle** of radius a. It is described by the parametric equations:

$$x = (a + b)\cos(t) - b\cos((a/b + 1)t)$$
$$y = (a + b)\sin(t) - b\sin((a/b + 1)t).$$

An epicycloid is like a **cycloid** on the circumference of a circle and is closely related to the **epitrochoid**, **hypocycloid**, and **hypotrochoid**. An epicycloid with one cusp is called a **cardioid**, one with two cusps is called a **nephroid**, and one with five cusps is called a **ranunculoid** (after the buttercup genus *Ranunculus*).

Epimenides paradox

See **liar paradox**.

epitrochoid

A curve traced out by a point that is a distance c from the center of a circle of radius b, where $c < b$, that is rolling around the outside of another circle of radius a. It is described by the parametric equations

$$x = (a + b)\cos(t) - c\cos((a/b - 1)t)$$
$$y = (a + b)\sin(t) - c\sin((a/b + 1)t).$$

Closely related to the epitrochoid are the **epicycloid**, **hypocycloid**, and the **hypotrochoid**. An example of an epitrochoid appears in Albrecht **Dürer**'s work *Instruction in Measurement with Compasses and Straightedge* (1525).

EPORN

An equal product of reversible numbers; defined by the Indian recreational mathematician Shyam Sunder Gupta as a number that can be expressed as the product of two reversible numbers (numbers whose digits are reversed) in two different ways. For example: 4,030 = 130 × 031 = 310 × 013 and 144,648 = 861 × 168 = 492 × 294. The smallest EPORN, 2,520 = 120 × 021 = 210 × 012, is also the least common multiple of all single digit natural numbers in decimal system. The **digital root**, i.e. the ultimate sum of digits, of all EPORNs is always 1, 4, 7, or 9. For example, 2,520 = 2 + 5 + 2 + 0 = 9; 4,030 = 4 + 0 + 3 + 0 = 7; 9,949,716 = 9 + 9 + 4 + 9 + 7 + 1 + 6 = 36 and 3 + 6 = 9.

equichordal point

A point inside a closed **convex** curve in the plane, all the **chord**s through which have the same length.

equilateral

Having sides of equal length, as in the case of an equilateral **polygon**. The equilateral **triangle**, with its three equal angles of 60°, is widely found in historic buildings and structures across Europe. See **Triangular Lodge**.

equilibrium

A term indicating a rest state of a system, for example, when a **dynamical system** is under the sway of a **fixed-point attractor** or **periodic attractor**. The concept originated in ancient Greece when **Archimedes** experimented with levers in balance, literally "equilibrium." The idea was elaborated through the Middle Ages, the Renaissance, and the birth of modern mathematics and physics in the seventeenth and eighteenth centuries. "Equilibrium" has come to mean pretty much the same thing as *stability*, that is, a system that is largely unaffected by internal or external changes since it easily returns to its original condition after being perturbed.

equivalent numbers

Numbers such that the sums of their **aliquot part**s (proper divisors) are the same. For example, 159, 559, and 703 are equivalent numbers because their aliquot parts all sum to 57.

Eratosthenes of Cyrene (c. 276–194 B.C.)

A Greek mathematician, astronomer, and geographer who was born in Cyrene, a Greek colony to the west of Egypt. He studied at Plato's school in Athens and eventually became the chief librarian of the great Library at Alexandria. He wrote works on geography, philosophy, history, astronomy, mathematics, and literary criticism. One of Eratosthenes' contributions to mathematics was his measurement of Earth's circumference, which he calculated to be about 252,000 stadii, or 24,700 miles (about one-tenth the actual value, but still a big improvement on earlier estimates). Eratosthenes is also known in number theory for his **sieve of Eratosthenes**, which finds all **prime number**s less than a given integer *n*.

Eratosthenes's sieve

See **sieve of Eratosthenes**.

Erdös, Paul (1913–1996)

A Hungarian mathematician (his name is pronounced "AIR-dosh"), one of the greatest mathematicians of the twentieth century and, in terms of the number of papers published (more than 1,500), the most prolific in history–beating out even Leonhard **Euler** and inspiring the term "Erdös number." A mathematician has an Erdös number of 1 if he or she has published a paper with Erdös, of 2 if he or she has published with someone who published a paper with Erdös, and so on. Erdös worked almost nonstop, 19 hours a day, 7 days a week. "A mathematician," he quipped, "is a machine for turning coffee into theorems." At age 20, Erdös discovered an elegant proof of a famous theorem in number theory, known as *Chebyshev's theorem,* which says that for each number greater than one, there is always at least one **prime number** between it and its double. **Number theory** remained one of his chief interests, though his work spread across many fields, and he became renowned for posing and solving problems that were often simple to state but notoriously difficult to solve. He did groundbreaking work in a branch of mathematics known as **Ramsey theory** long before it became fashionable in the late 1950s. Bent and slight, often wearing sandals, Erdös had no time for the material side of life. "Property is nuisance," he said. Focused totally on mathematics, Erdös traveled from meeting to meeting, carrying a half-empty suitcase and staying with mathematicians wherever he went. His colleagues took care of him, lent him money, fed him, bought him clothes, and even did his taxes. In return, he showered them with ideas and challenges–with problems to be solved and brilliant ways of attacking them. Ernst Straus, who worked with both Albert Einstein and Erdös, wrote a tribute to Erdös shortly before his own death in 1983. He said of Erdös: "In our century, in which mathematics is so strongly dominated by 'theory doctors,' he has remained the prince of problem solvers and the absolute monarch of problem posers."[170]

ergodic

The property of a **dynamical system** such that all regions of a **phase space** are visited with similar frequency and that all regions will be revisited (within a small proximity) if given enough time.

Escher, Maurits Cornelius (1898–1972)

> *My work is a game, a very serious game.*

A Dutch artist whose graphic explorations of **tiling**, figure-ground ambiguities, **impossible figure**s, and regression has attracted the interest of mathematicians and scientists. His experiences of Moorish art (see **Alhambra**) and his contact with mathematicians, most notably Harold **Coxeter**, led him to explore the way repetitive shapes can be used to tile the plane, and from this to ideas about duality and transformation. Escher's preoccupation with dualities is a constant presence in his work in the form of foreground/background, light and dark, flatness and dimensionality, representation and decoration, frame and scene, large and small, viewpoint and vanishing point, form and negative space, positive and negative, observer and observed, as well as the metaphysical aspects of good and evil. Self-referential images (see **self-referential sentence**) resonate throughout Escher's works–reflections of the artist, hands that draw themselves, the visitor in a picture gallery who looks at a print that contains him. It was for this reason that Douglas **Hofstadter** wove Escher, along with Kurt Gödel and Bach, into the "eternal golden braid" at the heart of his Pulitzer Prize–winning book.[172] Having seen some of Escher's art, Roger **Penrose** was inspired to devise impossible figures, including the **Penrose triangle**, which Escher then incorporated into several of his later works. After the artist's death Penrose regretted that Escher had not lived long enough to take advantage of the discovery of **Penrose tiling**.

escribed circle

A circle that is **tangent** to one side of a triangle and to the extensions of the other sides.

Eternity Puzzle

An enormously difficult jigsaw consisting of 209 pieces, each one different and each made from a unique configuration of equilateral triangles and half-triangles with the same total area as six triangles. The puzzle was to fit them together into an almost-regular 12-sided figure aligned to a triangular grid. The puzzle's inventor, Christopher Monckton, announced a prize of $1 million when the puzzle was released commercially in June 1999, for the first correct solution submitted, assuming there was one, when all the solutions were opened in September 2000. Monckton had run computer searches on much smaller versions of the puzzle, which had convinced him that the sheer size of Eternity would make it intractable. However, the prize was won by two British mathematicians, Alex Selby and Oliver Riordan, with the help of a couple of computers, who sent in a correct tiling on May 15, six weeks ahead of the only other puzzler known to have found a correct solution. Early on, Selby and Riordan made a surprising discovery. As the number of pieces in an Eternity-like puzzle increased, so did the difficulty–but only up to a point. The critical size is about 70 pieces, which would be almost impossible to solve. For larger puzzles, however, the number of possible correct solutions increases. In the case of Eternity itself, with its 209 pieces, there are thought to be at least 10^{95} solutions–far more than the number of subatomic particles in the universe but far, far less than the number of nonsolutions. The puzzle itself is much too large to solve by an exhaustive search but not, as it turns out, by more savvy methods that take into account what shaped regions are easiest to tile and what shaped pieces are easiest to fit. By steadily refining their search algorithm, Selby and Riordan were able to prune out the vast majority of nonsolutions and, with a bit of

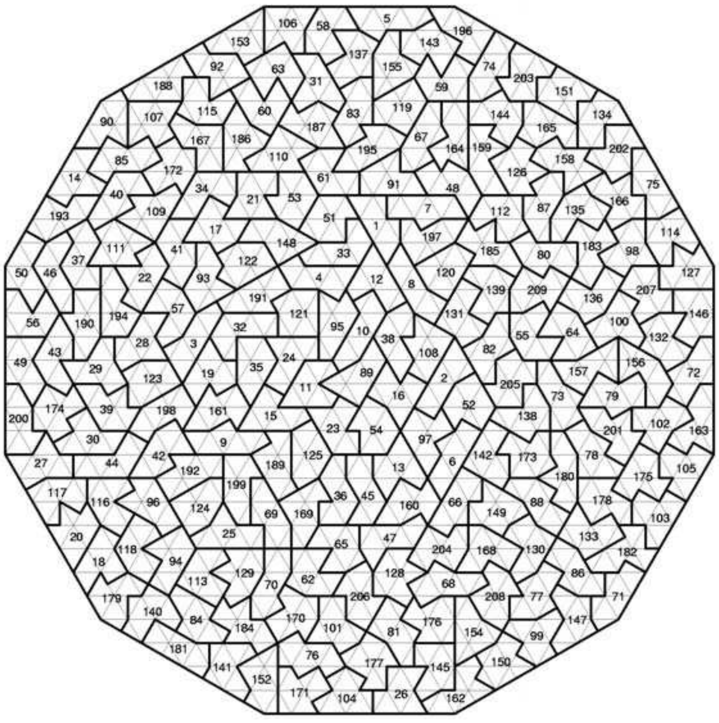
106 58 5 196
153 143 137
92 63 31 155 74
188 59 203 151
90 107 115 60 187 83 119 144 134
167 186 67 165
85 110 195 164 159 158 202
172 126
14 34 21 53 61 91 48 75
40 7 112 87 135 166
193 109 51 1
17 148 197 114
111 37 41 122 33 185 80 183 98
50 46 22 93 4 12 8 120 127
191 139 209 136 207
56 190 194 57 121 131 146
28 3 32 95 10 38 55 64 100 132
43 24 108 82 156
49 29 123 19 35 89 2 205 157 72
11 16 52 73 79
174 39 198 161 15 138
200 23 54 201 102 163
30 9 97 125
27 44 42 189 13 6 142 173 78
192 180 175 105
199
117 116 96 36 45 66 88 178
124 69 169 160 149 103
20 25 47 65 133
118 204 168 130 182
18 94 129 128 86
113 70 62 68
179 206 208 77 71
140 84 184 170 101 81 176 154 99 147
181 76 177 145 150
141 152
171 104 26 162

Eternity Puzzle The solution to the Eternity Puzzle that was awarded a $1 million prize. *Eternity pieces are copyright © 1999 by Christopher Monckton* good fortune, to hit upon a correct solution and claim the prize.

Euclid of Alexandria (c. 330–270 B.C.)

A Greek mathematician who compiled and systematically arranged the geometry and number theory of his day into the famous text *Elements.* This text, used in schools for about 2,000 years, earned him the name "the father of geometry." Even today, the geometries that don't satisfy the fifth of Euclid's "common notions" (now called axioms or postulates) are called non-Euclidean geometries. When, according to the Greek philosopher Proclus, the Egyptian ruler Ptolemy asked if there was a shorter way to the study of geometry than the *Elements,* Euclid told the pharaoh that "there is no royal road to geometry." Little is known of Euclid's life. Proclus wrote (c. A.D. 350) that Euclid lived during the reign of Ptolemy and founded the first school of mathematics in Alexandria–the site of the most impressive library of ancient times with perhaps as many as 700,000 volumes. He wrote books on other subjects such as optics and conic sections, but most of them are now lost. See also **Euclidean geometry**.

Euclidean geometry

Geometry of the type described originally by **Euclid** in his book *Elements* and based on five axioms (see **Euclid's postulates**), one of which is the controversial **parallel postulate**. Various forms of **non-Euclidean geometry** began to emerge in the nineteenth century, with enormous implications for science and philosophy. See also **Euclidean space**.

Euclidean space

Any n-dimensional mathematical space that is a generalization of the familiar two- and three-dimensional spaces described by the axioms of **Euclidean geometry**. The term "n-dimensional Euclidean space" (where n is any positive whole number) is usually abbreviated to "Euclidean n-space", or even just "n-space". Formally, Euclidean n-space is the set $\mathbf{R}^n$ (where $\mathbf{R}$ is the set of **real numbers**) together with the *distance function,* which is obtained by defining the distance between two points $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ to be the square root of $\Sigma(x_i - y_i)^2$, where the sum is over $i = 1, \ldots, n$. This distance function is based on **Pythagoras's theorem** and is called the *Euclidean metric.*

Euclid's postulates

The five postulates, which together with 23 definitions and five "common notions," form the basis of **Euclid**'s great work on geometry, *Elements.* The postulates are:

1. A straight line may be drawn from any one point to any other point.
2. A finite straight line may be produced to any length in a straight line.
3. A circle may be described with any center at any distance from that center.
4. All right angles are equal.
5. If a straight line meets two other lines, so as to make the two interior angles on one side of it together less than two right angles, the other straight lines will meet if produced on that side on which the angles are less than two right angles.

The last postulate is not as obvious as the other four, and Euclid himself was reluctant to use it. Later mathematicians, finding the fifth postulate to be complicated, thought it might be possible to derive it from the other four. However, they only succeeded in replacing it with equivalent statements. The most common of these is the **parallel postulate**.

Eudoxus of Cnidus (c. 408–c. 355 B.C.)

A Greek astronomer, mathematician, and physician whose work on ratios formed the basis for Book V of **Euclid**'s *Elements* and anticipated some aspects of algebra, such as cross multiplying, which is otherwise absent from ancient Greek mathematics. Eudoxus constructed many geometric proofs, found formulas for measuring pyramids, cones, and cylinders, and developed the **method of exhaustion**, a forerunner of integration, later extended by Archimedes. He also studied the kampyle curve, often known as the **kampyle of Eudoxus**, in connection with the classical problem of **duplicating the cube**.

Euler, Leonhard (1707–1783)

A great Swiss mathematician; the second most prolific mathematician in history, after Paul **Erdös**. His greatest contributions were to **number theory**, but Euler also did important work in **calculus**, geometry, algebra, probability, acoustics, optics, mechanics, astronomy, artillery, navigation, and finance. He had a knack for coming up with important results by intuition, he cast calculus and trigonometry in their modern forms, and he showed the importance of the number ***e***. Even the amusing puzzles he invented and, in some cases, solved have opened up new mathematical fields. The **bridges of Königsberg** problem, for example, heralded the beginning of **graph** theory and **topology**, while his **thirty-six officers problem** stimulated important work in **combinatorics**. Euler also worked on **magic squares** and the problem of the **knight's tour**. Having learned some math from his father, a Calvinist preacher, Euler studied at the University of

Basle where he became close friends with members of the **Bernoulli family**. In 1727, he moved to St. Petersburg, to the court of Catherine the Great, becoming professor of physics (1730) and of mathematics (1733). While in Russia, Euler, a devout Christian, met the encyclopedist and philosopher René Diderot, a notorious atheist. When Diderot heard that Euler had a mathematical proof of the existence of **God**, he asked for it and was quoted the equation now often referred to as **Euler's formula**.

Upon losing the use of his right eye, Euler said "Now I will have less distraction." Indeed, the quantity of his output seemed to be inversely proportional to the quality of his sight, because his rate of publication increased after he became almost totally blind in 1766. Euler died moments after calculating the orbit of Uranus on September 18, 1783.

Euler characteristic

An important kind of number, known as a *topological invariant,* that describes a closed surface. In the case of polyhedra, the Euler characteristic is the number of vertices and faces minus the number of edges (see **Euler's formula for polyhedra**).

Euler circuit

A **connected graph** such that starting at a **vertex** *a,* one can traverse every **edge** of the graph once to each of the other vertices and return to vertex *a.* In other words a Euler circuit is a **Euler path** that is a circuit. Thus, using the properties of odd and even degree vertices given in the definition of a Euler path, a Euler circuit exists if and only if every vertex of the graph has an even degree. See also **mazes**.

Euler line

A line that connects the **centroid** and the **circumcenter** of a triangle.

Euler path

A path along a **connected graph** that connects all the vertices (see **vertex**) and that traverses every **edge** of the graph only once. Note that a vertex with an odd degree allows one to travel through it and return by another path at least once, while a vertex with an even degree only allows a number of traversals through, but one cannot end a Euler path at a vertex with even degree. Thus, a connected graph has a Euler path which is a circuit (a **Euler circuit**) if all of its vertices have even degree. A connected graph has a Euler path which is non-circuitous if it has exactly two vertices with odd degree. See also **Hamilton path**.

Euler square

A square array made by combining n objects of two types such that the first and second elements form a **Latin square**. Euler squares are also known as *Graeco-Latin squares, Graeco-Roman squares,* or *Latin-Graeco squares.* For many years, Euler squares were known to exist for n = 3, 4, and for every odd n except $n = 3k$. *Euler's Graeco-Roman squares conjecture* maintained that there are no Euler squares of order $n = 4k + 2$ for $k = 1, 2, \ldots$. However, such squares were found to exist in 1959 by Bose and Shrikande, refuting the conjecture.

Euler-Mascheroni constant (γ)

Also known as *Euler's constant* or *Mascheroni's constant,* the limit (as n goes to infinity) of

$$1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \tfrac{1}{5} + \ldots + 1/n - \log n$$

It is often denoted by a lowercase gamma, γ, and is approximately 0.5772156649 . . . . Even though over one million digits of this number have been calculated, it isn't yet known if it is a rational number (the ratio of two integers a/b). If it is rational, the denominator (b) must have more than 244,663 digits. The constant γ crops up in many places in **number theory**. For example, in 1898, the French mathematician Charles de la Vallée Poussin (who proved the **prime number** theorem) proved the following: Take any positive integer n and divide it by each positive integer m less than n. Calculate the average (mean) fraction by which the quotient n/m falls short of the next integer. The larger n gets, the closer the average gets to gamma.

Euler's conjecture

It always takes n terms to sum to an nth power: two squares, three cubes, four fourth powers, and so on. This hypothesis is now known to be wrong. In 1966, L. J. Lander and T. R. Parkin found the first counterexample: four fifth powers that sum to a fifth power. They showed that $27^5 + 84^5 + 110^5 + 133^5 = 144^5$. In 1988, Noam Elkies of Harvard University found a counterexample for fourth powers: $2{,}682{,}440^4 + 15{,}365{,}639^4 + 187{,}960^4 = 20{,}615{,}673^4$. Subsequently, Roger Frye of Thinking Machines Corporation did a computer search to find the smallest example: $95{,}800^4 + 217{,}519^4 + 414{,}560^4 = 422{,}481^4$.

Euler's constant

See **Euler-Mascheroni constant**.

Euler's formula

For any **real number** x, Euler's formula is

$$e^{ix} = \cos x + i \sin x$$

where e is a fundamental constant (the base of natural logarithms) and $i = \sqrt{-1}$. If we now put $x = \pi$, we get

$$e^{i\pi} = \cos \pi + i \sin \pi,$$

and since $\cos(\pi) = -1$ and $\sin(\pi) = 0$, this reduces to

$$e^{i\pi} = -1$$

so that

$$e^{i\pi} + 1 = 0.$$

This most extraordinary equation first emerged in Leonhard **Euler**'s *Introductio,* published in 1748. It is remarkable because it links the most important mathematical constants, e and **π**, the imaginary unit i, and the basic numbers used in counting, 0 and 1. In describing the equation to students, the Harvard mathematician Benjamin Peirce said: "Gentlemen, that is surely true, it is absolutely paradoxical; we cannot understand it, and we don't know what it means, but we have proved it, and therefore, we know it must be the truth."

Euler's formula for polyhedra

The earliest known equation in **topology**. If F is the number of faces of a **polygon**, E the number of edges, and V the number of vertices, Euler's formula can be written as

$$F - E + V = 2$$

where F − E + V is known as the **Euler characteristic**. For example, the surface of a cube has six (square) faces, twelve edges, and eight vertices and, sure enough, 6 − 12 + 8 = 2.

even function

A function $f(x)$ such that $f(x) = f(-x)$ for all x.

evolute

The **locus** of the centers of curvature (the envelope) of a plane curve's normals. The original curve is then said to be the **involute** of its evolute. For example, the evolute of an **ellipse** is a **Lamé curve** and the evolute of a **tractrix** is a **catenary**.

excluded middle law

A law in (two-valued) logic which states that there is no third alternative to truth or falsehood. In other words, for any statement A, either A or not-A must be true and the other must be false. This law no longer holds in three-valued logic, in which "undecided" is a valid state, nor does it hold in **fuzzy logic**.

existence

A term that has several different meanings within **mathematics**. In the broadest sense there is the question of what it means for certain concepts, such as **pi**, to exist. Was π, for example, invented or discovered? In other words, does π exist only as an intellectual construct or was it somehow already "out there" waiting for people to find it. If it does exist independently of the human mind, when did its existence start? Does π predate the physical universe? Such ontological questions become even more difficult when applied to more complex or abstract mathematical concepts such as the **Mandelbrot set**, **surreal number**s, or **infinity**. A narrower and more technical type of "existence" in math is implied by an *existence theorem.* Such a theorem is used to prove that a number or other object with particular properties definitely exists, but does not necessarily give a specific example. Finally, there is existence in the sense of particular solutions to problems. If at least one solution can be determined for a given problem, a solution to that problem is said to exist. Something of the flavor of all three types of mathematical existence mentioned here are captured in the following anecdote:

> An engineer, a chemist, and a mathematician are staying in three adjoining cabins at an old motel. First the engineer's coffee-maker catches fire. He smells the smoke, wakes up, unplugs the coffee maker, throws it out the window, and goes back to sleep. Later that night the chemist smells smoke, too. He wakes up and sees that a cigarette butt has set the trash can on fire. He thinks to himself, "How does one put out a fire? One can reduce the temperature of the fuel below the flash point, isolate the burning material from oxygen, or both. This could be accomplished by applying water." So he picks up the trash can, puts it in the shower stall, turns on the water, and, when the fire is out, goes back to sleep. The mathematician has been watching all this out the window. So later, when he finds that his pipe ashes have set the bed sheet on fire, he's not in the least taken aback. "Aha!" he says, "A solution exists!" and goes back to sleep.

exponent

A number that gives the **power** to which a **base** is raised. For example, in 3^2 the base is 3 and the exponent is 2.

exponential

> *Who has not been amazed to learn that the function* y = e^x, *like a phoenix rising again from its own ashes, is its own derivative?*
>
> –Francois le Lionnais

Anything that grows at a rate proportional to its size is said to grow exponentially. The simplest form of the exponential function is just $y = e^x$, where ***e*** is about 2.712 . . . The exponential function to **base** *a* can be written as $f(x) = a^x$.

extrapolate

See **interpolate**.

extravagant number

See **economical number**.

F

face

A **polygon** bounding a **polyhedron**. A **cube**, for example, has six square faces. The plane angle formed by adjacent edges of a polygonal angle in space is called a *face angle.*

factor

Also known as a *divisor,* a number or variable that divides evenly into another number or algebraic expression. For example, the factors of 28 are 1, 2, 4, 7, 14, and 28. Although it is true that 28 is also divisible by the negative of each of these, "factors" is usually taken to mean only the positive divisors. *Factorization,* or *factoring,* is the decomposition of an object into a product of factors. For example, the number 15 factorizes into **prime number**s as 3 × 5; and the **polynomial** $x^2 - 4$ factorizes as $(x - 2)(x + 2)$. The aim of factoring is usually to reduce something to basic building blocks, such as numbers to prime numbers, or polynomials to linear expressions. Factoring integers is covered by the **fundamental theorem of arithmetic** and factoring polynomials by the **fundamental theorem of algebra**. *Integer factorization* for large integers appears to be a difficult problem; there are no known methods for solving it quickly, and, for this reason, it has formed the basis of some public key cryptography algorithms.

factorial

The function, denoted $n!$, that is the product of the positive integers less than or equal to n. For example, $1! = 1$; $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$; $10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 3{,}628{,}800$. 0! is defined to be 1, by working the relationship $n! = n \times (n - 1)!$ backward. An interesting equality is $1!\,10!\,22!\,1! = 11!\,0!\,2!\,21!$ in which the same digits are broken up two different ways into factorials. This may be the smallest such example. Factorials are important in **combinatorics** because there are $n!$ different ways (permutations) of arranging n distinct objects in a sequence. They also turn up in formulas in calculus, for instance in *Taylor's theorem,* because the nth derivative of the function x^n is $n!$.

factorion

A natural number that equals of the sums of the **factorial**s of its digits in a given **base**. The only known **decimal** factorions are $1 = 1!$, $2 = 2!$, $145 = 1! + 4! + 5!$, and $40{,}585 = 4! + 0! + 5! + 8! + 5!$.

Fadiman, Clifton (1904–1999)

An American essayist, literary critic, and noted intellectual who, among many other works, edited *Fantasia Mathematica*[96] and *The Mathematical Magpie.*[97] He became well known for the encyclopedic knowledge he displayed on the *Information Please* radio programs in the 1930s and '40s.

Fagnano's problem

In a given acute triangle ABC, find the inscribed triangle whose perimeter is as small as possible. The answer is the *orthic triangle* of ABC, that is, the triangle whose vertices are endpoints of the altitudes from each of the vertices of ABC. The problem was proposed and solved using calculus by Giovanni Fagnano (1715–1797) in 1775. Once the answer became known, several purely geometric solutions were also discovered.

fair division

See **cake-cutting**.

Farey sequence

A sequence of numbers named after the English geologist John Farey (1766–1826) who wrote about such sequences in an article called "On a curious property of vulgar fractions" in the *Philosophical Magazine* in 1816. Farey says that he noted the "curious property" while examining the tables of *Complete decimal quotients* produced by Henry Goodwin. To obtain the Farey sequence for a fixed number n, consider all **rational number**s between 0 and 1 which, when expressed in their lowest terms, have a denominator (the number on the bottom of a fraction) not exceeding n. Write the sequence in ascending order of magnitude beginning with the smallest. The "curious property" is that each member of the sequence is equal to the rational number whose numerator (the number on top of a fraction) is the sum of the numerators of the fractions on either side, and whose denominator is the sum of the denominators of the fractions on either side. For example, the Farey sequence for $n = 5$ is $(\frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1})$, from which it can be seen that $\frac{2}{5} = (1 + 1)/(3 + 2)$, $\frac{1}{3} = (1 + 2)/(4 + 5)$, $\frac{1}{2} = (2 + 3)/(5 + 5)$, $\frac{2}{3} = (3 + 3)/(5 + 4)$, and so forth. Farey wrote: "I am not acquainted whether this curious property of vulgar fractions has been before pointed out?; or whether it may admit of some easy or general demonstration?; which are points on which I should be glad to learn the sentiments of some of your mathematical readers."

One "mathematical reader" was Augustin **Cauchy**, who gave the necessary proof in his *Exercises de mathématique,* published in the same year as Farey's article. Farey was not the first to notice the property. C. Haros, in 1802, wrote a paper on the approximation of decimal fractions by common fractions. He explains how to construct what is in fact the Farey sequence for $n = 99$ and Farey's "curious property" is built into his construction.

Fechner, Gustav Theodor (1801–1887)

A German physicist and psychologist who studied aesthetic aspects of the **golden ratio** and published his findings in *Vorschule der Aesthetik* (Introduction to aesthetics) (1876), arguing that this ratio turns up commonly in human-made rectangular objects and is judged by people to be the most pleasing to the eye (though some later researchers have called his results into question).

Federov, E. S. (1853–1919)

A Russian geologist and crystallographer who helped lay the theoretical foundations for modern crystallography. In his famous two-part paper "Symmetry of Regular Systems of Figures" published in 1891, he proved that there are exactly 17 distinct symmetries in the **wallpaper group**.

feedback

The mutually **reciprocal** effect of one system or subsystem on another. *Negative feedback* is when two subsystems act to dampen the output of the other. For example, the relation of predators and prey can be described by a negative feedback loop since more predators lead to a decline in the population of prey, but when prey decrease too much so does the population of predators since they don't have enough food. *Positive feedback* means that two subsystems are amplifying each other's outputs, e.g., the screech heard in a public address system when the mike is too close to the speaker. The microphone amplifies the sound from the speaker which in turn amplifies the signal from the microphone, and so on. Feedback is a way of talking about the nonlinear interaction among the elements or components in a system and can be modeled by nonlinear differential or difference equations as well as by the activity of cells in a **cellular automaton** array.

Feigenbaum's constant

A universal constant, denoted δ, that governs the behavior of systems that are approaching **chaos**; it was discovered by the American mathematical physicist Mitchell Feigenbaum (1944–) in 1975 and has the value $\delta = 4.6692\ldots$. All one-dimensional chaotic systems have a behavior, as they approach instability, known as period doubling. The Feigenbaum constant gives the rate at which the period of the system doubles.

Fermat, Pierre de (1601–1665)

A French lawyer, magistrate, and gentleman scholar, often called the "Prince of Amateurs," who is best known for the **conjecture**, now proved, known as **Fermat's last theorem**. Although employed as a senior government official, Fermat somehow managed to find time to do an astonishing amount of math, for which he sought little acclaim or acknowledgment. In fact, he published only one important manuscript in his entire lifetime and even then used fake initials. When his fellow French mathematician Gilles Roberval offered to edit and publish some of his works, Fermat replied, "Whatever of my works is judged worthy of publication, I do not want my name to appear there." Most of his results are known through letters to friends, notes in book margins, and challenges to other mathematicians to find proofs for theorems he had devised.

Fermat was one of the founders, with René **Decartes**, of **analytical geometry** and, with Blaise **Pascal**, of **probability theory**. His work on the maxima and minima of curves and tangents to them was seen, by Isaac **Newton**, as a starting point for **calculus**. Yet his greatest love was for **number theory**. In 1640, while studying **perfect number**s, Fermat wrote to Mersenne that if p is prime, then $2p$ divides $2^p - 2$. Shortly after he expanded this into what is now called **Fermat's little theorem**. As usual, Fermat stated "I would send you a proof, if I did not fear its being too long." His most famous statement of this form accompanied his hasty notes on the "last theorem." See also **Fermat number**.

Fermat number

A number defined by the formula $F_n = 2^{2^n} + 1$ and named after Pierre **Fermat** who conjectured, wrongly, that all such numbers would be **prime**. The first five Fermat numbers, $F_0 = 3$, $F_1 = 5$, $F_2 = 17$, $F_3 = 257$, and $F_4 = 65{,}537$, are prime. However, in 1732, Leonhard **Euler** discovered that 641 divides F_5. It takes only two trial divisions to find this factor because Euler showed that every factor of a Fermat number F_n with n greater than 2 has the form $k \times 2^{n+2} + 1$. In the case of F_5 this is $128k + 1$, so we would try 257 and 641 (129, 385, and 513 are not prime). It is likely that there are only finitely many Fermat primes. Gauss proved that a **regular polygon** of n sides can be inscribed in a circle with Euclidean methods (e.g., by straightedge and compass) if and only if n is a power of two times a product of distinct Fermat primes.

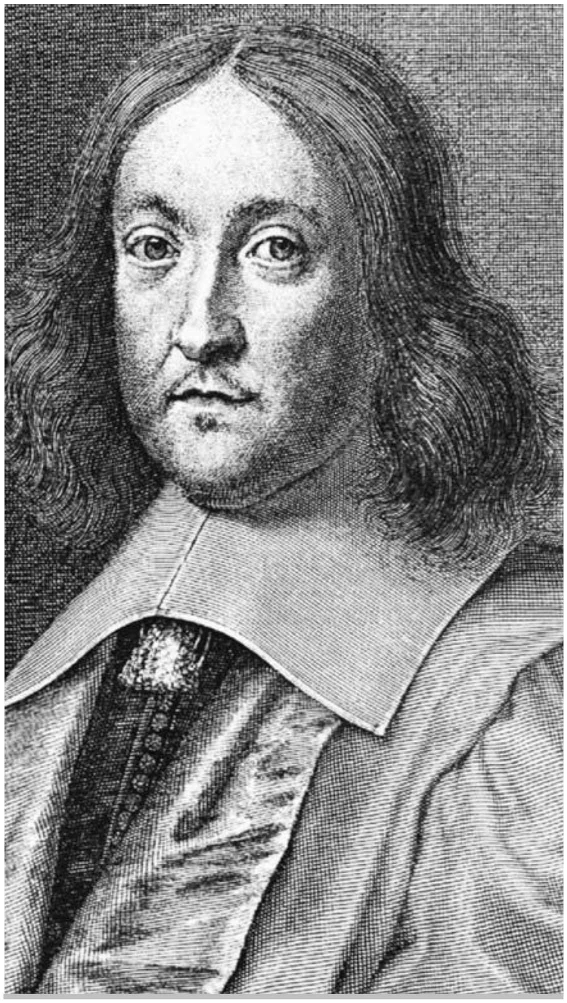

Fermat, Pierre de

Fermat's last theorem

A challenge for many long ages
Had baffled the savants and sages.
Yet at last came the light:
Seems old Fermat was right–
To the margin add 200 pages.

–Paul Chernoff

A conjecture put forward by Pierre de **Fermat** in 1637 in the form of a note scribbled in the margin of his copy of the ancient Greek text *Arithmetica* by **Diophantus**. The note was found after his death, and the original is now lost. However, a copy was included in the appendix to a book published by Fermat's son. Fermat's note read: "It is impossible to write a cube as a sum of two cubes, a fourth power as a sum of fourth powers, and, in general, any power beyond the second as a sum of two similar powers. For this, I have found a truly wonderful proof, but the margin is too small to contain it."

Fermat claimed that the **Diophantine equation** $x^n + y^n = z^n$ has no integer solutions for $n > 2$. It turns out he was right. But the proof had to wait 350 years and involved such advanced techniques, virtually none of which existed in the seventeenth century, that is seems very unlikely that Fermat really had found an elementary proof. Fermat's last theorem–now truly a theorem–was finally proved correct by Andrew **Wiles** in 1994.[353] In order to reach that dizzy height, however, Wiles had to draw on and extend several ideas at the core of modern mathematics. In particular, he tackled the *Shimura-Taniyama-Weil conjecture,* which provides links between the branches of mathematics known as **algebraic geometry** and **complex analysis**. This conjecture dates back to 1955, when it was published in Japanese as a research problem by the late Yutaka Taniyama. Goro Shimura of Princeton and Andre Weil of the Institute for Advanced Study provided key insights in formulating the conjecture, which proposes a special kind of equivalence between the mathematics of objects called elliptic curves and the mathematics of certain motions in space. Interestingly, the Wiles proof of Fermat's last theorem was a byproduct of his deep inroads into proving the Shimura-Taniyama-Weil conjecture. Now, the Wiles effort could help point the way to a general theory of three variable Diophantine equations. Historically, mathematicians have always had to state and solve such problems on a case-by-case basis. An overarching theory would represent a tremendous advance. See also **ABC conjecture**.

Fermat's little theorem

If P is a **prime number** then for any number a, $(a^P - a)$ must be divisible by P. This theorem is useful for testing if a number is *not* prime, though it can't tell if a number is prime. As usual, Pierre de **Fermat** didn't provide a proof (this time saying "I would send you the demonstration, if I did not fear its being too long"). Leonhard **Euler** first published a proof in 1736, but Gottfried **Leibniz** left virtually the same proof in an unpublished manuscript from sometime before 1683.

Fermat's spiral

A parabolic **spiral**.

Fibonacci (c. 1175–1250)

The pen name of Leonardo of Pisa, one of the greatest mathematicians of the Middle Ages. The son of a Pisan merchant who also served as a customs officer in North Africa, he traveled widely in Barbary (Algeria) and was later sent on business trips to Egypt, Syria, Greece, Sicily, and Provence. In 1200 he returned to Pisa and used the knowledge gained on his travels to write *Liber Abaci* (The book of the abacus), published in 1202, which introduced to western Europe the Hindu-Arabic numerals and **decimal** number system that remain in use today. The first chapter of Part 1 begins: "These are the nine figures of the Indians: 9 8 7 6 5 4 3 2 1. With these nine figures, and with this sign *0* which in Arabic is called zephirum, any number can be written, as will be demonstrated."

Fibonacci also showed he was capable of some amazing feats of calculation. For example, he found the positive solution of the **cubic** equation $x^3 + 2x^2 + 10x = 20$ using the Babylonian number system with base 60 (a strange choice, in view of his public advocacy of the decimal system!). He gave the result as 1, 22, 7, 42, 33, 4, 40 which is equivalent to

$$1 + \frac{22}{60} + \frac{7}{60^2} + \frac{42}{60^3} + \frac{33}{60^4} + \frac{4}{60^5} + \frac{40}{60^6}\,.$$

How on Earth he obtained this, nobody knows; it was 300 years before anyone else could obtain such accurate results. As well as serious mathematics, *Liber Abaci* contains many playful passages and it is for one of these, concerning a problem about counting the offspring of a pair of rabbits, that Fibonacci became best known after Edouard **Lucas** called the sequence of numbers discussed by the rabbit problem the **Fibonacci sequence**.

Fibonacci sequence

The sequence that arises in answer to this problem posed in **Fibonacci**'s great work *Liber Abaci:* "A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begins a new pair which from the second month on becomes productive?"

The number of pairs of rabbits in the *n*th month begins 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, . . . , where each term is the sum of the two terms preceding it. This sequence can be defined recursively as follows: $F(1) = F(2) = 1$, $F(n + 1) = F(n) + F(n - 1)$ for $n > 2$, where $F(n)$ is the *n*th Fibonacci number. Johannes **Kepler** was the first to point out that the growth rate of the Fibonacci numbers, that is, $F(n + 1) / F(n)$, converges to the **golden ratio**, ϕ (phi).

In the nineteenth century Fibonacci numbers were discovered in many natural forms. For example, many types of flower have a Fibonacci number of petals: certain types of daisies tend to have 34 or 55 petals, while sunflowers have 89 or, in some cases, 144. The seeds of sunflowers spiral outward both to the left and the right in a Fibonacci number of spirals. Similarly, the whorls on a pinecone, the numbers of rings on the trunks of palm trees, the patterns of snail shells, and the genealogy of the male bee all follow a sequence of Fibonacci numbers. The arrangement of plant leaves, or phyllotaxis, unfolds to the same pattern because this results in an optimal solution in terms of the spacing of the leaves or the amount of light that can reach them. A familiar spiral form, known as the **logarithmic spiral**, emerges when seeds on a plant grow and space themselves according the Fibonacci sequence. The logarithmic spiral is approximated by the rule: start at the origin of the Cartesian coordinate system, move $F(1)$ units to the right, move $F(2)$ units up, move $F(3)$ units to the left, move $F(4)$ units down, move $F(5)$ units to the right, and so on. By growing in this way, on structures such as sunflowers, pinecones, and pineapples, seeds are able to pack themselves together most efficiently.

Fibonacci numbers have so many interesting mathematical properties that an entire journal, *The Fibonacci Quarterly,* is devoted to them. The sequence of final digits in Fibonacci numbers repeats in cycles of 60. The last two digits repeat in 300, the last three in 1,500, the last four in 15,000, etc. The product of any four consecutive Fibonacci numbers is the area of a **Pythagorean triangle**. The shallow (least steep) diagonals of **Pascal's triangle** sum to Fibonacci numbers. Let *m* and *n* be positive integers, then

$F(n)$ divides $F(mn)$

$\gcd(F(n), F(m)) = F(\gcd(m, n))$, where "gcd" stands for "greatest common divisor."

$(F(n))^2 - F(n + 1)F(n - 1) = (-1)^{n-1}$.

$F(1) + F(3) + F(5) + \ldots + F(2n - 1) = F(2n)$.

For every *n,* there are *n* consecutive composite Fibonacci numbers.

An interesting use of the Fibonacci sequence is for converting miles to kilometers. For instance, if you want to know about how many kilometers 5 miles is, take the Fibonacci number (5) and look at the next one (8) (5 miles is about 8 kilometers). This works because it happens that the conversion factor between miles and kilometers is roughly equal to the Golden Ratio.

The first few Fibonacci numbers that are also **prime numbers** are 3; 5; 13; 89; 233; 1,597; 28,657; 514,229; . . . . It seems likely that there are infinitely many Fibonacci primes, but this has yet to be proven. However, it is relatively easy to show that for $n \geq 4$, $u_n + 1$ is never prime. The Fibonacci sequence is a special case of the **Lucas sequence**.

The *tribonacci series* is made by adding the last two digits: 1, 1, 2, 4, 7, 13, 24, 44, 81, . . . and from this the quadbonacci series, the pentbonacci series, and the

Fibonacci sequence The number of spirals of seeds on a sunflower is always a Fibonacci number—an arrangement that keeps the seeds uniformly packed no matter how large the seed head. *Thomas Stromberg*

hexbonacci series, all the way up to the *n*-bonacci series. Each ratio of successive terms forms a special constant, analogous to ϕ.

field

A number system in which addition, subtraction, multiplication, and division (except by zero) are always defined, and the **associative** and **distributive** laws are valid. For example, the set of **rational number**s is a field, whereas the set of **integer**s is not a field, because the result of dividing one integer by another is not necessarily an integer. The **real number**s also constitute a field, as do the **complex number**s. Compare with **ring**.

Fields Medal

By convention, the most prestigious award for research in mathematics. It is awarded every four years to between two and four mathematicians under the age of 40.

Fifteen Puzzle

A sliding-tile puzzle invented by Sam **Loyd** in the 1870s that became a worldwide obsession, much as **Rubik's cube** did a century later. Fifteen little tiles, numbered 1 to 15, were placed in a four by four frame in serial order except for tiles 14 and 15, which were swapped around; the lower right-hand square was left empty. The object of the puzzle was to get all the tiles in the correct order; the only allowed moves were sliding counters into the empty square. Everyone it seemed was caught up with the craze—playing the game in horse-drawn trams, during their lunch breaks, or when they were supposed to be working. The game even made its way into the solemn halls of the German parliament. "I can still visualize quite clearly the gray-haired people in the Reichstag intent on a square small box in their hands," recalled the geographer and mathematician Sigmund Gunter who was a deputy during the puzzle epidemic. "In Paris the

Fairylite
15
PUZZLE
The Puzzle of the Age –
CAN YOU SOLVE IT?

Fifteen Puzzle A version of the Fifteen Puzzle produced in England by Fairylite. *Sue & Brian Young/Mr. Puzzle Australia, www.mrpuzzle.com.au*

puzzle flourished in the open air, in the boulevards, and proliferated speedily from the capital all over the provinces," wrote a contemporary French author. "There was hardly one country cottage where this spider hadn't made its nest lying in wait for a victim to flounder in its web." Loyd offered a $1,000 reward for the first correct solution. But, although many claimed it, none were able to reproduce a winning series of moves under close scrutiny. There is a simple reason for this, which is also the reason that Loyd was unable to obtain a U.S. patent for his invention. According to regulations, Loyd had to submit a working model so that a prototype batch could be manufactured from it. Having shown the game to a patent official, he was asked if it were solvable. "No," he replied. "It's mathematically impossible." Upon which the official reasoned there could be no working model and thus no patent!

The puzzle's theory reveals that the more than 20 billion possible starting arrangements of the tiles fall into just two groups: one in which all the tiles can be maneuvered into ascending numerical order (call this group I), and one in which tiles 14 and 15 will be inverted (group II). It's impossible to combine arrangements from these two groups and impossible to turn a group I arrangement into a group II, or vice versa, using the normal rules of the game. Given a random arrangement of tiles, can we know in advance if we have the unsolvable kind? Very easily. Simply count how many instances there are of a tile numbered n appearing after the tile numbered $n + 1$. If there are an even number of such inversions, the puzzle is solvable, otherwise you are wasting your time!

figurate number

A number sequence found by creating consecutive geometrical figures from arrangements of equally spaced points. Here is an example:

```
1   3   5   7
*   *   *   *
    **  *   *
        *** *
            ****
```

The points can be arranged in one, two, three, or more dimensions. There are many different kinds of figurate numbers, such as **polygonal number**s and **tetrahedral number**s.

films and plays involving mathematics

Mathematicians and mathematics rarely make an entrance on the silver screen or the stage, unless as parodies in the form of mad professors and meaningless scrawled equations. A notable exception is *A Beautiful Mind* (2001), directed by Ron Howard and starring Russell Crowe as the brilliant but mentally troubled mathematician John **Nash**. Although a fine love story and a well-crafted film, which won four Oscars, *A Beautiful Mind* is weak on math and inaccurate in many of its details of Nash's life and his battle with schizophrenia. *Rain Man* (1988), also based on a true story, costars Dustin Hoffman as an autistic savant with a photographic memory and a genius for mental arithmetic.

Good Will Hunting (1997), written by and starring Matt Damon and Ben Affleck, and also starring Robin Williams, is about a young man who has led a troubled life but has an amazing talent for mathematics. His abilities are discovered when he comes into conflict with the law, and he soon has to decide if he should pursue his mathematical future and leave his family and friends behind. In Darren Aronofsky's disturbing independent film *Pi* (1998), the main character is a mathematician obsessed with his search for patterns within **pi**'s infinite decimal places. He believes they can be used to predict chaotic behaviors, including that of the stock market. Throughout the film he is pursued by ruthless stock market players and by rabbis trying to find a mathematical way to communicate with God. In the science fiction film *Cube* (1997), six people awake to find themselves trapped in a deadly maze, and one of the characters uses mathematical skills to solve the puzzle and find a way to escape.

Lesser known films with strong mathematical themes include Mario Martone's *Death of a Neapolitan Mathematician;* Peter Greenaway's *Drowning by Numbers;* George Csicsery's *N Is a Number;* and *Moebius,* made by students and faculty at the Universidad del Cine of Buenos Aires. Mathematics has also found its way onto the stage. The musical *Fermat's Last Tango* (2000), a fictionalized account of Andrew **Wiles**'s struggle to prove **Fermat's last theorem**, was performed in New York by the York Theatre company. It followed the Pulitzer prize–winning play *Proof* by David Aubern, about the death of a brilliant mathematician and the repercussions for his daughters and his student.

finite

Limited in extent or scope. In mathematics, a *finite set* is such that the number of elements it contains can be described by a **natural number**. For instance, the set of integers between –18 and 5 is finite, because it has a natural number (17) of elements. The set of all **prime numbers**, on the other hand, is not finite. In physics, "finite" is used to mean both "not infinite" and "nonzero."

finite-state automaton (FSA)

The simplest computing device. Although it is not nearly powerful enough to perform universal computation, it can recognize **regular expression**s. FSAs are defined by a state transition table that specifies how the FSA moves from one state to another when presented with a particular input. FSAs can be drawn as **graph**s.

Fisher, Adrian

A British professional designer and constructor of **maze**s; his company, Adrian Fisher Maze Design, has built a huge variety of mazes in Britain, continental Europe, the United States, and elsewhere. These include the formal hedge maze at Leeds Castle and the largest brick pavement maze in the world at Kentwell Hall in Long Melford. The latter is based on a Tudor rose and has 15 sepals used as locations for a board game in which live players take part in Tudor costume.[101]

Fitchneal

An Irish version of the Viking game **Hnefa-Tafl**; played on a 7 × 7 board (as was the Scottish equivalent, known as *Ard-Ri,* "High King"), it is mentioned in the *Mabinogion* and *Cormac's Glossary* of the ninth century.

five

The length of the hypotenuse of the smallest **Pythagorean triangle** (a right triangle having integral sides). Five is the only **prime number** that is a member of two pairs of **twin primes.** Every integer is the sum of five positive or negative cubes in an infinite number of ways. Five is the smallest degree of a polynomial equation for which there is no general formula for the solutions (see **quintic**).

fixed-point attractor

An **attractor** that is represented by a particular point in **phase space**, sometimes called an *equilibrium point.* As a point it corresponds to a very limited range of possible behaviors of the system. For example, in the case of a pendulum, the fixed-point attractor represents the pendulum when the bob is at rest. This state of rest attracts the system because of gravity and friction.

Flatland: A Romance of Many Dimensions

A satirical novel by Edwin A. **Abbott**,[1] first published in 1884, that portrays a **two-dimensional world**, like the surface of a map, over which its inhabitants move. Flatlanders have no concept of up and down, and appear to each other as mere points or lines. From our three-dimensional perspective we can look down on Flatland and see that its people are "really" a variety of shapes, including straight lines (females), narrow isosceles triangles (soldiers and workmen), equilateral triangles (lower middle-class men), squares and pentagons (professional men, including the pseudonymous author of the tale, A. Square), hexagons and other regular polygons with still more sides (the nobility), and circles (priests). Abbott uses these geometrical distinctions, especially the appearance of Flatland females and the working class, as a commentary on the discrimination against women, the rigid class stratification, and the lack of tolerance for "irregularity" that was prevalent in Victorian Britain.

In a dream, A. Square visits the one-dimensional world of Lineland where he tries, unsuccessfully, to persuade the king that there is such a thing as a second dimension. In turn, the narrator is told of three-dimensional space by a sphere who moves slowly through the plane of Flatland, growing and shrinking as his cross-section changes in size. (If a **hypersphere** were to move through our three-dimensional world, we would see a sphere appear, grow to a maximum size, and then shrink again before disappearing.) Abbott is aware that he cheats a little in his description of what the inhabitants of Flatland actually see. In his preface to the second edition, he gives a lengthy but not-too-convincing reply to the objection, raised by some readers, that a Flatlander, "seeing a Line, sees something that must be *thick* to the eye as well as *long* to the eye (otherwise it would not be visible . . .)." The curious and often-neglected fact is that we are just as unable to imagine what it would truly be like to see in two dimensions as we are to conceive of four dimensions! No matter how hard we try we cannot imagine being able to see a line of *zero* thickness.

flexagon

A flat model constructed from a folded strip of paper, which, when flexed, can be made to reveal a number of hidden faces. Flexagons are amusing toys but they have also caught the interest of mathematicians. They are usually square or rectangular (tetraflexagons) or hexagonal (hexaflexagons). A prefix can be added to the name to indicate the number of faces that the model can display, including the two faces (back and front) that are visible before flexing. For example, a hexaflexagon with a total of six faces is called a hexahexaflexagon. The discovery of the first flexagon, a trihexaflexagon, is credited to the British student Arthur H. Stone who was studying at

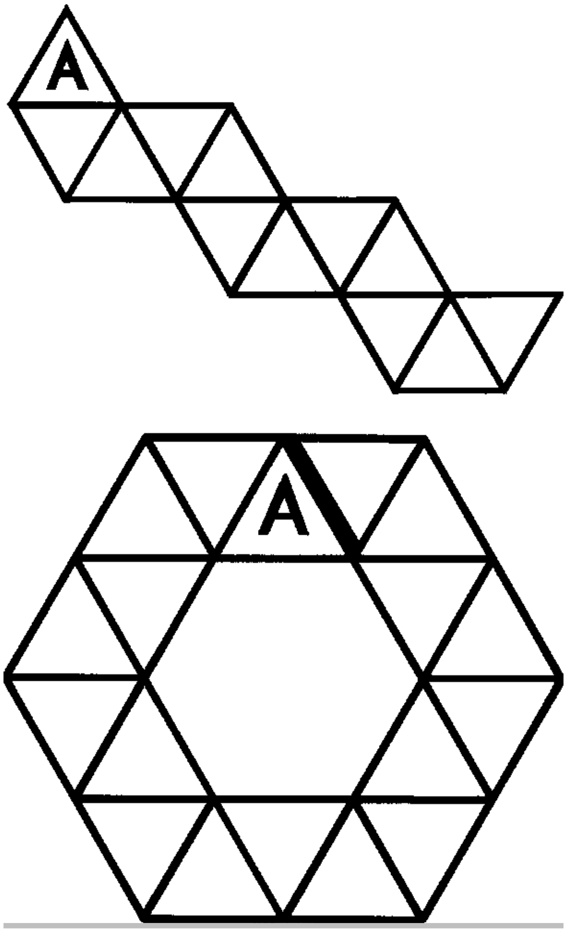

flexagon Two nets for folding into hexaflexagons: the 4-flexagon (top) and the 6-flexagon (bottom). To use the nets, photocopy and enlarge them, and label each side with these numbers:

4-flexagon $\begin{pmatrix} 2231 & 2231 & 2231 \\ 3144 & 3144 & 3144 \end{pmatrix}$,

6-flexagon $\begin{pmatrix} 662554 & 662554 & 662554 \\ 231431 & 231431 & 231431 \end{pmatrix}$.

Start at A using the top row of numbers. Number the other side of the net with the bottom row of numbers so that the top and bottom numbers appear on either side of the first triangle, and so on. *Jill Russell*

Princeton University in 1939. Stone's colleagues Bryant Tuckermann, Richard P. Feynman, and John W. Tukey became interested in the idea. Tuckerman worked out a topological method, called the *Tuckerman traverse,* for revealing all the faces of a flexagon. Tukey and Feynman developed a complete mathematical theory that has not been published. Flexagons were introduced to the general public by Martin **Gardner** writing in *Scientific American.*[110]

floor function

The greatest integer in *x,* that is, the largest integer less than or equal to *x.*

Flower of Life

One of the beautiful arrangements of circles found at the Temple of Osiris at Abydos, Egypt. The pattern also appears in Phoenician art from the ninth century B.C. The circles are placed with six-fold symmetry, forming a mesmerizing pattern of circles and lenses. A related design from the same temple, which recurs in Italian art from the thirteen century, is called the *Seed of Life.*

fly between trains problem

Two trains are approaching each another and a fly is buzzing back and forth between the two trains. Given the (constant) speed of the trains and their initial separation distance, and the (constant) speed of the fly, calculate how far the fly will travel before the trains collide. This problem appears to have been first posed by Charles Ange Laisant (1840–1921) in his *Initiation Mathématique.* There is a long-winded method of getting the answer and a much shorter way. Suppose the trains start out 200 miles apart and are each traveling at 50 miles per hour, and the fly–a speedster of its kind–is moving at 75 miles per hour. The long method involves considering the length of the back-and-forth path that the fly takes and evaluating this as the sum of an infinite series. The quick solution is to notice that the trains will collide in 2 hours and that in this time the fly will travel $2 \times 75 = 150$ miles! When this problem was put to John **von Neumann**, he immediately gave the correct answer. The poser, assuming he had spotted the shortcut, said: "It is very strange, but nearly everyone tries to sum the infinite series." Von Neumann replied: "What do you mean, strange? That's how I did it!"

focal chord

A **chord** of a **conic section** that passes through a **focus**.

focal radius

A line segment from the **focus** of an **ellipse** to a point on the **perimeter** of the ellipse.

focus

A defining point in the construction of a **conic section**. The word comes from the Latin for hearth or fireplace and appears to have been first used in mathematics, in describing an **ellipse**, by Johannes **Kepler**.

foliation

A decoration of a **manifold** in which the manifold is partitioned into sheets of some lower dimension, and the sheets are locally parallel. More technically, the foliated manifold is locally **homeomorphic** to a **vector space** decorated by cosets of a subspace.

folium

A curve, first described by Johannes **Kepler** in 1609, that corresponds to the general equation

$$(x^2 + y^2)(y^2 + x(x + b)) = 4axy^2, \text{ in Cartesian form, or}$$
$$r = -b \cos\theta + 4a \cos\theta \sin^2\theta, \text{ in polar coordinates.}$$

The Latin *folium* means "leaf-shaped." Three types, known as the simple folium, the bifolium (or double folium), and the trifolium, correspond to the cases when $b = 4a$, $b = 0$, and $b = a$, respectively. The *folium of Descartes* is given by the Cartesian equation $x^3 + y^3 = 3axy$ and was first discussed by René **Descartes** in 1638. Although he found the correct shape of the curve in the positive quadrant, he wrongly thought that this leaf shape was repeated in each quadrant like the four petals of a flower. The problem to determine the tangent to the curve was proposed to Gilles de Roberval who, having made the same incorrect assumption, called the curve *fleur de jasmin* after the four-petal jasmine bloom, a name that was later dropped. The folium of Descartes has an asymptote $x + y + a = 0$.

formal system

A mathematical formalism in which statements can be constructed and manipulated with logical rules. Some formal systems, such as **Euclidean geometry**, are built around a few basic **axiom**s and can be expanded with theorems that can be deduced through proofs.

formalism

A mathematical school of thought that was headed by the German mathematician David **Hilbert**. Formalists argue that mathematics must be developed through **axiom**atic systems. Formalists agree with Platonism on the principles of mathematical proof, but Hilbert's followers don't recognize an external world of mathematics. Formalists argue that mathematical objects don't exist until we define them. Humans create the **real number** system, for example, by establishing axioms to describe it. All that mathematics needs are inference rules to progress from one step to the next. Formalists tried to prove that within the framework of established axioms, theorems, and definitions, a mathematical system is consistent and, in the mid-twentieth century, formalism became the predominant philosophical attitude in math textbooks. However, it was undermined by **Gödel's incompleteness theorem** and also the general recognition that results can be usefully applied without having to be proved or derived axiomatically.

Fortune's conjecture

A conjecture about **prime number**s made by the New Zealander social anthropologist Reo Fortune (1903–1979), who had a reputation for unstable behavior bordering on the psychotic. He once attempted to settle an academic dispute with a colleague, Thomas McIlwraith, at the University of Ontario, by challenging him to a duel with any weapon of his choice from the collections of the Royal Ontario Museum. Fortune proposed that if q is the smallest prime greater than $P + 1$, where P is the product of the first n primes, then $q - P$ is prime. For example, if n is 3, then P is $2 \times 3 \times 5 = 30$, $q = 37$, and $q - P$ is the prime 7. These numbers, $q - P$, are now known as *Fortunate numbers.* The conjecture remains unproven but is generally thought to be true. The sequence of Fortunate numbers begins

$$3, 5, 7, 13, 23, 17, 19, 23, 37, 61, 67, \ldots$$

four

The smallest **composite number**, the second smallest **square number**, the first non-Fibonacci number (see **Fibonacci sequence**), the smallest **Smith number**, and the smallest number that can be written as the sum of two **prime number**s. Four is the number of dimensions that make up **space-time** (three of space and one of time). It is the most number of colors needed to color any map so that no two neighboring areas are the same color (see **four-color map problem**). There are four cardinal points on the compass, four Riders of the Apocalypse, and four Gospels.

four-color map problem

A long-standing problem that dates back to 1852 when Francis Guthrie, while trying to color a map of the counties of England noticed that four colors were enough to ensure that no adjacent counties were colored the same. He asked his brother Frederick if it was true that *any* map could be colored using four colors in such a way that adjacent regions (i.e., those sharing a common boundary segment, not just a point) receive different colors. Frederick Guthrie then passed on the conjecture to Augustus **de Morgan**. The first printed reference is due to Arthur **Cayley** in 1878. A year later the false "proof," by the English barrister Alfred Kempe, appeared; *its* incorrectness was pointed out by Percy Heawood 11 years later. Another failed proof is due to Peter **Tait** in 1880, a gap in *his* argument being pointed out by Julius Petersen in 1891. Both false proofs did have some value, though. Kempe discovered what became known as *Kempe chains,* and Tait found an equivalent formulation of the four-color theorem in terms of three-edge coloring.

The next major contribution came from George Birkhoff whose work allowed Philip Franklin in 1922 to prove that the four-color conjecture is true for maps with at most 25 regions. It was also used by other mathematicians to make various forms of progress on the four-color problem. In the 1970s, the German mathematician Heinrich Heesch developed the two main ingredients needed for the ultimate proof–reducibility and discharging. While the concept of reducibility was studied by other researchers as well, it seems that the idea of discharging, crucial for the unavoidability part of the proof, is due to Heesch, and that it was he who conjectured that a suitable development of this method would solve the four-color problem. This was confirmed by Kenneth Appel and Wolfgang Haken of the University of Illinois in 1977, when they published their proof of the four-color theorem.[12] Their controversial proof challenges the basic assumptions of what mathematical proof is. They used more than 1,200 hours of supercomputer time to analyze 1,478 different configurations that in turn can produce every possible map on a plane. Not everyone was happy with the method of the breakthrough, as Appel himself pointed out:

> For almost a century and a half, a Holy Grail of graph theory has been a simple incisive proof of the Four Color Theorem. It has troubled our profession that a problem that can be understood by a school child has yet to be solved in a way that better illuminates the reason that only four colors are needed for planar maps. The feelings of many mathematicians were summed up for me by Herb Wilf's response to being told that it appeared that one could prove the theorem by a long reducibility argument which used computers to test the reducibility of a large number of configurations. He simply said, "God would not allow such a beautiful theorem to have such an ugly proof."

Martin **Gardner** commented, "Whether a simple, elegant proof not requiring a computer will ever be found, is still an open question." It's interesting that such a simple, intuitive puzzle can be so difficult to settle! The four-color theorem is true for maps on a plane or on a sphere. The answer is different for geographic maps on a **torus**: in this case, seven colors are necessary and sufficient.[277]

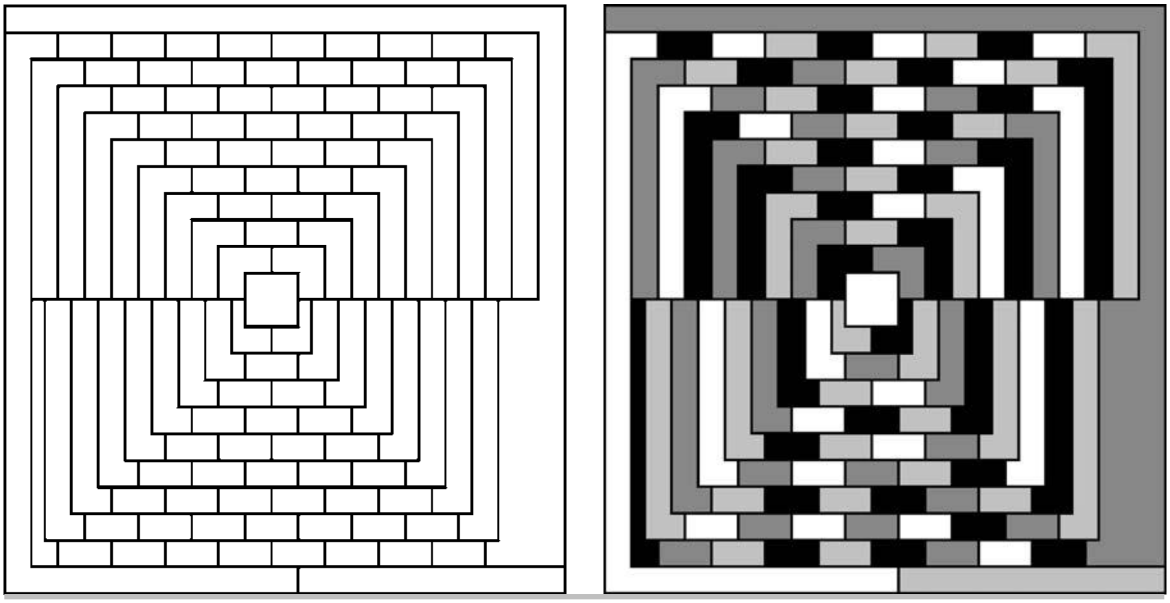

four-color map problem Martin Gardner's spoof counterexample to the four-color theorem (left) and a solution using four colors (right). The different colors are represented here in black, white, and gray.

four coins problem

Given three coins of possibly different sizes, which are arranged so that each is tangent to the other two, find the coin that is tangent to the other three coins. The solution is the inner **Soddy circle**.

four fours problem

Using arithmetic combinations of four 4's, express all the numbers from 1 to 100. For example, 1 = 44/44 and 2 = (4 × 4)/(4 + 4). The problem was first presented in *The Schoolmaster's Assistant: Being a Compendium of Arithmetic Both Practical and Theoretical* (first edition c. 1744), a popular textbook by the English schoolteacher and cleric Thomas Dilworth (d. 1780). Operations and symbols that are allowed include the four arithmetic operations (+, ×, –, /), concatenation (e.g., the use of 44), decimal points (e.g., 4.4), powers (e.g., 4^4), square roots, factorials (e.g., 4!), and overbars for repeating digits (e.g., .4 with an overbar to express 4/9). Ordinary use of parentheses are allowed. One of the trickiest numbers to represent in this way is 73, which calls for something as contorted as

$$\sqrt{(\sqrt{(\sqrt{(4^4)})})} + 4 / .4'$$

(where .4′ is shorthand for .444 . . .). Of course, the problem can be extended to represent integers greater than 100. The highest value achievable in the four four's puzzle is $10^{8.0723047260281 \times 10153} = 4^{4^{4^4}}$.

four knights puzzle

On a 3 × 3 chessboard are two white knights at the top left-hand and top right-hand squares and two black knights at the bottom left-hand and bottom right-hand squares. The problem is to exchange the black knights with the white knights in the minimum possible number of moves. One move is a normal knight's move on any vacant cell of the board, which renders the center square inaccessible.

Fourier, (Jean Baptiste) Joseph, Baron (1768–1830)

A French mathematician known chiefly for his contribution to the mathematical analysis of heat flow. Although he trained for the priesthood, Fourier didn't take his vows but instead turned toward mathematics. He first studied and later taught mathematics at the newly created Ecole Normale. In 1798 he joined Napoleon's army in its invasion of Egypt as scientific advisor, to help establish educational facilities there and to carry out archaeological explorations. After his return to France in 1801 he was appointed prefect of the department of Isere. Fourier became famous for his *Theorie analytique de la Chaleur* (1822), a mathematical treatment of how heat conducts in solid bodies. He established the **partial differential equation** governing heat flow and solved it by using an infinite series of trigonometric functions, now known as **Fourier series**. Though these series had been used before, Fourier investigated them in much greater detail and prepared the way for later work on trigonometric series and the theory of functions of a real variable. Fourier's belief that his health would be improved by wrapping himself up in blankets proved fatal: thus encumbered he tripped down the stairs of his house and died.

Fourier series

Named after Joseph **Fourier**, the expansion of a **periodic** function as an infinite sum of sines and cosines of various frequencies and amplitudes. This is similar to the approximation of an irrational number by a sum of a series of rational numbers (or a decimal expansion). Human ears effectively produce Fourier series automatically from complex sounds. Tiny hairs, known as *cilia,* vibrate at different specific frequencies. When a wave enters the ear, the cilia vibrate if the wave function contains any component of the corresponding frequency. This enables the hearer to distinguish sounds of various pitches. Fourier series are used a great deal in science and engineering to find solutions to **partial differential equations**, such as those in problems involving heat flow. They can also be used to construct some pathological functions such as ones that are continuous but nowhere differentiable. The study and computation of Fourier series is known as **harmonic analysis**.

fourth dimension

"Do you think that there are things which you cannot understand, and yet which are; that some people see things that others cannot?" said Dr. Van Helsing in Bram Stoker's *Dracula.* Instead of vampires, he may just as easily have been talking about the fourth dimension–an extension at right-angles to the three familiar directions of up-down, forward-backward, and side-to-side. In physics, especially **relativity theory**, **time** is often regarded as the fourth dimension of the **space-time** continuum in which we live. But what meaning can be attached to a fourth *spatial* dimension? The mathematics of the fourth dimension (4-d) can be approached through a simple extension of either the algebra or the geometry of one, two, and three dimensions.

Algebraically, each point in a multidimensional space can be represented by a unique sequence of **real numbers**. One-dimensional space is just the **number line** of real numbers. Two-dimensional space, the plane, corresponds to the set of all **ordered pairs** (x, y) of real numbers, and three-dimensional space to the set of all ordered triplets (x, y, z). By extrapolation, four-dimensional space corresponds to the set of all ordered quadruplets (x, y, z, w). Linked to this concept is that of **quaternions**, which can also be viewed as points in the fourth dimension.

Geometric facts about the fourth dimension are just as easy to state. The fourth dimension can be thought of as a direction perpendicular to every direction in three-dimensional space; in other words, it stretches out along an axis, say the w-axis, that is mutually perpendicular to the familiar x-, y-, and z-axes. Analogous to the cube is a hypercube or **tesseract**, and to the sphere is a 4-d **hypersphere**. Just as there are five regular polygons, known as the **Platonic solids**, so there are six four-dimensional regular **polytopes**. They are: the 4-simplex (constructed from five tetrahedra, with three tetrahedra meeting at an edge); the tesseract (made from eight cubes, meeting three per edge); the 16-cell (made from 16 tetrahedra, meeting four per edge); the 24-cell (made from 24 octahedra, meeting three per edge); the 120-cell (made from 120 dodecahedra, meeting three per edge); and the monstrous 600-cell (made from 600 tetrahedra, meeting five per edge).

Geometers have no difficulty in analyzing, describing, and cataloging the properties of all sorts of 4-d figures. The problem starts when we try to *visualize* the fourth dimension. This is a bit like trying to form a mental picture of a color different from any of those in the known rainbow from red to violet, or a "lost chord," different from any that has ever been played. The best that most of us can hope for is to understand by analogy. For example, just as a sketch of a cube is a 2-d perspective of a real cube, so a real cube can be thought of as a perspective of a tesseract. At a movie, a 2-d picture represents a 3-d world, whereas if you were to watch the action live, in three-dimensions, this would be like a screen projection in four dimensions.

Many books have been written and schemes devised to nudge our imaginations into thinking four-dimensionally. One of the oldest and best is Edwin **Abbott**'s ***Flatland***[1] written more than a century ago, around the time that mathematical discussion of **higher dimensions** was becoming popular. H. G. Wells also dabbled in the fourth dimension, most notably in *The Time Machine* (1895), but also in *The Invisible Man* (1897), in which the central character drinks a potion "involving four dimensions," and in "The Plattner Story" (1896), in which the hero of the tale, Gottfried Plattner, is hurled into a four-spatial dimension by a school chemistry experiment that goes wrong and comes back with all his internal organs switched around from right to left.[285] The most extraordinary and protracted attack on the problem, however, came from Charles **Hinton**, who believed that, through appropriate mental practice involving a complicated set of colored blocks, a higher reality would reveal itself, "bring[ing] forward a complete system of four-dimensional thought [in] mechanics, science, and art."

Victorian-age spiritualists and mystics also latched on to the idea of the fourth dimension as a home for the spirits of the departed. This would explain, they argued, how ghosts could pass through walls, disappear and reappear at will, and see what was invisible to mere three-dimensional mortals. Some distinguished scientists lent their weight to these spiritualist claims, often after being duped by clever conjuring tricks. One such unfortunate was the astronomer Karl Friedrich Zöllner who wrote about the four-dimensional spirit world in his *Transcendental Physics* (1881) after attending séances by Henry Slade, the fraudulent American medium.

Art, too, became enraptured with the fourth dimension in the early twentieth century. When the Cubist painter and theorist Albert Gleizes said, "Beyond the three dimensions of Euclid we have added another, the fourth dimension, which is to say, the figuration of space, the measure of the infinite," he united math and art and brought together two major characteristics of the fourth dimension in early Modern Art theory–the geometric orientation as a higher spatial dimension and the metaphorical association with infinity.[159] See also **Klein bottle**.[25, 156, 163, 212, 213, 231, 254, 267, 271, 273, 340]

Fox and Geese

An English board game that dates back to the Middle Ages and is unusual in that the two sides are unequal, thus making this an example of a **Tafl game**. The lone fox attempts to capture 13 (or, in later versions, 17) geese, while the geese try to hem the fox in so that it can't move. The geese start out by filling up all the points on one side of the cross-shaped grid on the board. The fox–the one counter of a different color–begins on any vacant point remaining. The fox moves first. Each side, in its turn, may move one counter. Both fox and geese can move along a line, forwards, backwards, or sideways, to the next contiguous point. The fox may move along a line or jump over a goose to an empty point, capturing the goose and removing it from the board. Two or more geese may be captured by the fox in one turn, provided that he is able to jump to an empty point after each one. The fox wins if he depletes the gaggle of geese to a number that makes it impossible for them to trap him. The geese can't jump over the fox or capture the fox but instead must try to mob him and trap him in a corner. The geese win if they make it impossible for the fox to move. A modification of this game spread with the British to India, where during the Great Mutiny the game became known as "Officers and Sepoys." In this variant, two officers in a fort attempt to hold off 24 sepoys, who must storm the fort.

fractal

A geometric shape that can be subdivided at any scale into parts that are, at least approximately, reduced-size copies of the whole. The name "fractal," from the Latin *fractus* meaning a broken surface, was coined by Benoit **Mandelbrot** in 1975. The key property of fractals is **self-similarity**, which means that zooming in or zooming out of a fractal produces no overall change in appearance.

One of several technical definitions of a fractal is "a set of points whose topological dimension is less than its **Hausdorff dimension**." The topological dimension is an object's ordinary dimensionality–one in the case of a curve, two in the case of a surface, and so forth–and is always a whole number. The Hausdorff dimension, on the other hand, measures how much space an object fills, and can take non-integer values if the object is very complex and twisty.

Some fractals show a strong regularity and rigid self-similarity and are produced by the repeated application of a set of rules that may be quite simple. Among the best known of these "iterated function" systems are the **Koch snowflake**, the **Peano curve**s, the **Sierpinski carpet**, and the **Sierpinski gasket**. Other fractals, defined by a recurrence relation at each point in space, are among the most complex, beautiful, and beguiling mathematical structures known. They include the well-known **Mandelbrot set** and **Lyapunov fractal**s. Finally, there are random fractals generated by stochastic rather than deterministic processes, for example fractal landscapes. Random fractals have the greatest practical use, and can be used to describe many highly irregular real-world objects, including clouds, mountains, coastlines, and trees. See also **fractal dimension**.

fractal dimension

A non-integer measure of the irregularity or **complexity** of a system; it is an extension of the notion of dimension found in **Euclidean geometry**. Knowing the fractal dimension helps one determine the degree of irregularity and pinpoint the number of variables that are key to determining the dynamics of the system.

fraction

A number that represents a part, or several equal parts, of a whole; examples include one-half, two-thirds, and three-fifths. The word comes from the Latin *frangere,* meaning "to break." A *simple, common,* or *vulgar fraction* is of the form a/b, where a may be any integer and b may be any integer greater than 0. If $a < b$, the fraction is said to be *proper* ("bottom heavy"); otherwise it is *improper* ("top heavy"). A **decimal fraction** has a denominator (number on the bottom) of 10, 100, 1000, and so forth. See also **continued fraction**.

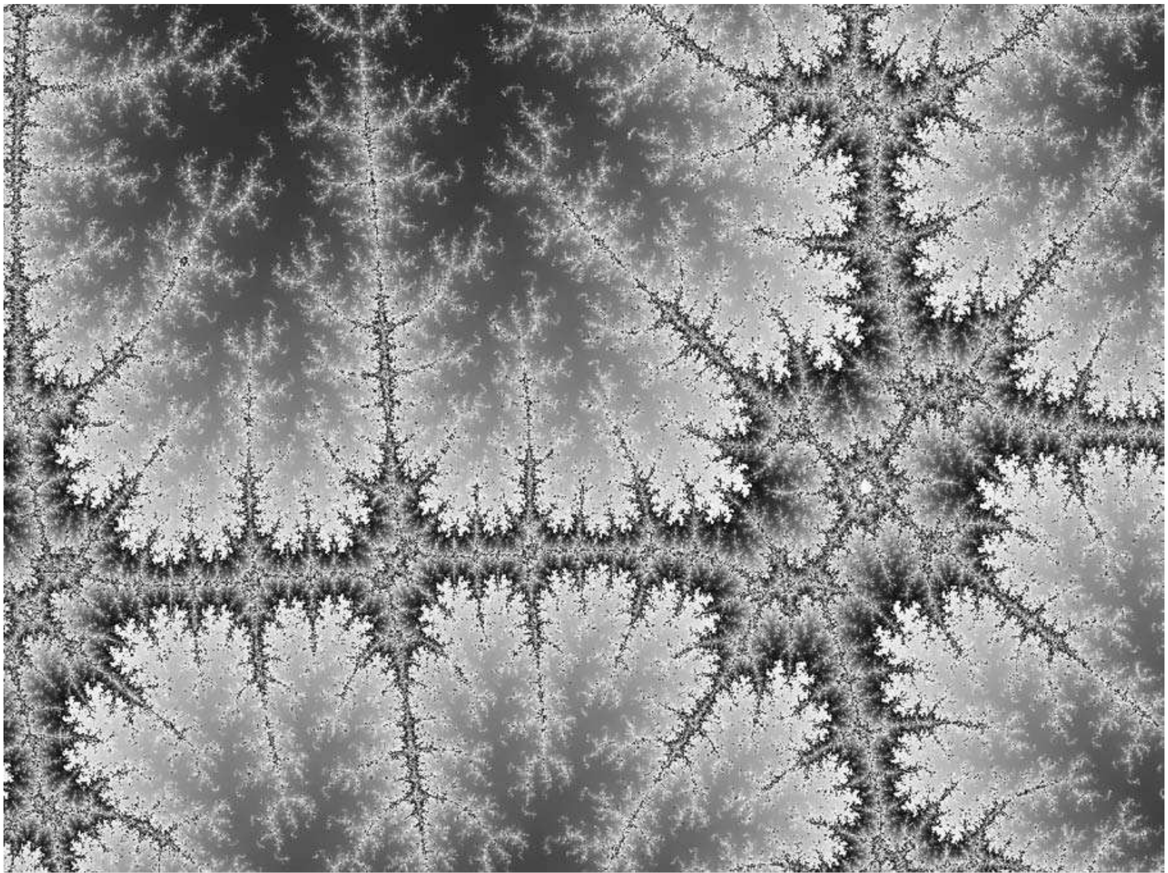

fractal A deep zoom of part of the Mandelbrot set. *Christopher Rowley*

Fraser spiral

A **distortion illusion** in which overlapping black arc segments appear to form a spiral but are in reality a series of overlapping concentric circles. This is easily demonstrated by following one of the curves with your finger. The illusion is named after the British psychologist James Fraser (1863–1936) who first published it in 1908.[104]

Fredholm, Erik Ivar (1866–1927)

A Swedish mathematician who founded the modern theory of **integral equation**s. This became a major research topic in the first quarter of the twentieth century and underpinned important theoretical developments in physics; David **Hilbert**, in particular, extended Fredholm's work to arrive at the concept of **Hilbert space**. Fredholm also devoted time to actuarial science and made a particularly important contribution by proposing an elegant formula to determine the surrender value of a life insurance policy. He earned his Ph.D. from the University of Uppsala but then spent the rest of his academic career at the University of Stockholm.

Freemish crate

See **impossible figure**.

Freeth's nephroid

See **nephroid**.

Frege, Friedrich Ludwig Gottlob (1848–1925)

A German mathematician and philosopher who virtually founded the modern discipline of mathematical logic. In *Die Grundlagen der Arithmetik* (The foundations of arithmetic, 1884), he used **set theory** to define the **cardinal number** of a given class as the class of all classes that are similar (i.e. can be placed in a **one-to-one** correspondence) to the given class. In *Grundgesetze der Arithmetik* (The basic laws of arithmetic, 2 vols., 1893 and 1903), Frege began attempting to build up mathematics from arithmetic and symbolic logic on a rigorous and contradiction-free basis. When the second volume was in the process of being printed, Bernard **Russell** pointed out a paradox in Frege's work. The paradox, which became known as **Russell's paradox**, stems from the question: "Is the class of all classes that are not members of itself a member of itself or not?" The question leads to a contradiction and cannot be resolved. Frege was thus forced to admit that the foundation of his reasoning was worthless. As he stated at the end of his work, "A scientist can hardly encounter anything more undesirable than to have the foundation collapse just as the work is finished. I was put in this position by a letter from Mr. Bertrand Russell when the work was almost through the press."

Frénicle de Bessy, Bernard (1602–1675)

An eminent French amateur mathematician who extensively researched **magic square**s; his *Des quassez ou tables magiques,* published posthumously in 1693, first identified all 880 magic squares of the fourth order. Frénicle also corresponded with **Descartes**, **Fermat**, **Huygens**, and **Mersenne**, mostly on **number theory**, the work for which he is best known.

frequency

The number of times a value occurs in some time interval.

friendly number

See **amicable numbers**.

Frogs and Toads

A puzzle in which three counters or pegs representing frogs are placed on three successive positions on the left of a string of seven squares, and three different tokens representing toads are placed on the three rightmost squares. Frogs only move to the right, toads only to the left. Every move is either a slide to the adjacent square or a jump over one position, which is allowed only if the latter is occupied by a member of the other species. No two animals are ever allowed on the same square. The goal is to move the toads into the three leftmost positions and the frogs into three rightmost positions in the fewest possible moves. Many different versions of this puzzle have appeared over the centuries and it may be Arabic in origin. The number of pieces on each side may vary, as may the number of empty starting places in the middle; other names for the puzzle have included *Sheep and Goats* and *Sphinxes and Pyramids.*

frugal number

See **economical number**.

frustum

Part of a solid cut off between two parallel planes; in particular, for a **cone** or a **pyramid**, a frustum is determined by the plane of the base and a plane parallel to the base. *Frustum* is Latin for "a piece broken off."

function

> *Old mathematicians never die; they just lose some of their functions.*
>
> –Anonymous

A way of expressing the dependence of one quantity on another quantity or quantities. Traditionally, functions were specified as explicit rules or formulas that converted some input value (or values) into an output value. If f is the name of the function and x is a name for an input value, then $f(x)$ denotes the output value corresponding to x under the rule f. An input value is also called an *argument* of the function, and an output value is called a *value* of the function. The **graph** of the function f is the collection of all pairs $(x, f(x))$, where x is an argument of f. For example, the circumference C of a circle depends on its diameter d according to the formula $C = \pi d$; therefore, one can say that the circumference is a function of the diameter, and the functional relationship is given by $C(d) = \pi d$. Equally well, the diameter can be considered a function of the circumference, with the relationship given by $d(C) = d/\pi$. In modern mathematics, the insistence on specifying an explicit effective rule has been abandoned; all that is required is that a function f associate with every element of some **set** X a unique element of some set Y. This makes it possible to prove the existence of a function without necessarily being able to calculate its values explicitly. Also, it enables general properties of functions to be proved independently of their form. The set X of all admissible arguments is called the *domain* of f; the set Y of all admissible values is called the *codomain* of f. We write $f: X \to Y$.

fundamental group

A group of a **topological space** X that is constructed by looking at how closed paths in X can be combined to get new paths. Under a suitable way of identifying paths (known as **homotopy**) one can get a group structure on the set which gives an algebraic invariant of the space X.

fundamental theorem of algebra

The result that any **polynomial** with real or complex coefficients has a root in the complex plane.

fundamental theorem of arithmetic

Every positive **integer** greater than 1 is a **prime number** or can be expressed as a unique product of primes and powers of primes.

fuzzy logic

A departure from classical two-valued logic in which something is either true or false, to allow a continuous range of truth values. Fuzzy logic was introduced by Lotfi Zadeh of the University of California at Berkeley in the 1960s as a means to model the uncertainty of natural language.

G

Gabriel's horn

The **surface of revolution** of

$$y = 1/x$$

for x greater than 1. Surprisingly this has a finite volume of **pi** cubic units but an *infinitely large* surface area! Gabriel's horn is also known as *Torricelli's trumpet* because it was investigated by the Italian Evangelista Torricelli (1608–1647). As a young man Torricelli studied in Galileo's home at Arcetri, near Florence, and then, upon Galileo's death, succeeded his teacher as mathematician and philosopher for their good friend and patron, the grand duke of Tuscany. Torricelli was amazed by the strange property of his mathematical trumpet and tried various ways to avoid the conclusion that a finite volume could be enclosed by a vessel with an infinite surface area. Unfortunately, he lived before **calculus** came along to explain the apparent paradox in terms of **infinitesimal**s.

Galois, Évariste (1811–1832)

A French mathematician who led a short, dramatic life and is often credited with founding modern **group** theory, though the Italian Paolo Ruffini (1765–1822) came up with many of the ideas first. Galois's work wasn't widely acknowledged by his contemporaries, partly because he didn't present his material very well and partly because he held unpopular political views. In fact, he was a republican revolutionary who was twice imprisoned because of his activities. During his second incarceration he fell in love with the daughter of the prison physician, Stephanie-Felice du Motel, and after being released, was killed in a duel over her with Perscheux d'Herbinville. His death started republican riots and rallies which lasted for several days. See also **Galois theory**.

Galois theory

The study of certain **group**s, known as *Galois groups,* that can be associated with **polynomial** equations. Whether

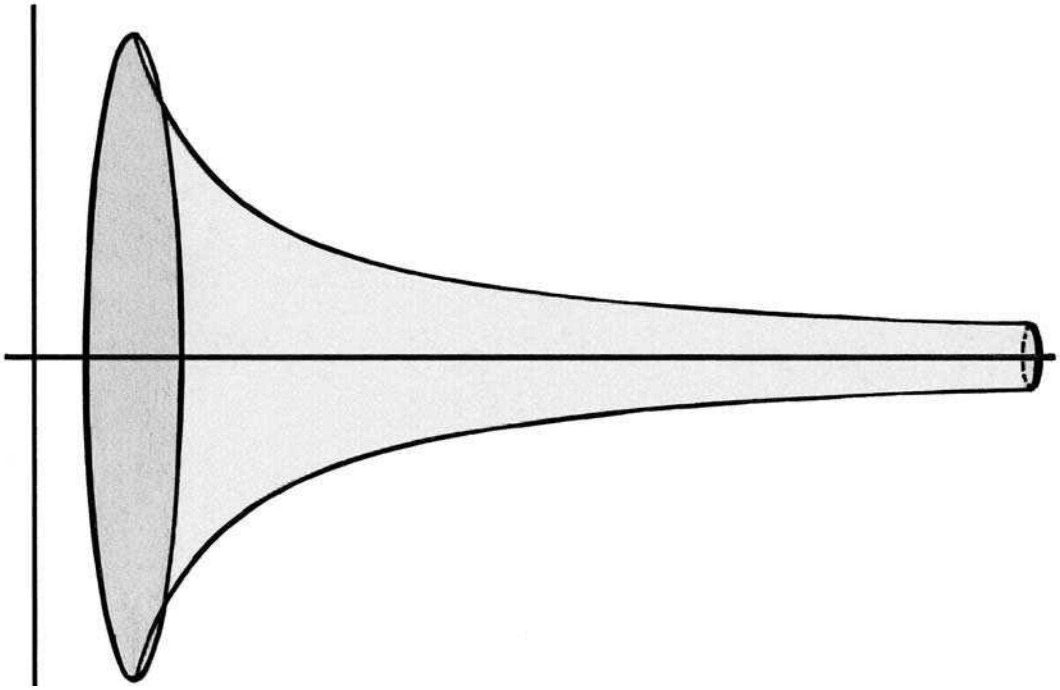

Gabriel's horn The horn for x values between 1 and 10.

or not the solutions to an equation can be written down using rational functions and square roots, cube roots, and so forth depends on certain group-theoretic properties of *Galois groups*.

game

A conflict, with formal rules and a finite number of choices of what to do at each stage, between two or more players. The study of games belongs to a branch of mathematics and logic known as **game theory**. If a game is simple enough, it can be solved for all possible outcomes. This is the case, for example, with **tic-tac-toe** and **Nim**. By harnessing the power of computers to check through vast numbers of moves, even more complicated games are succumbing to a complete analysis. In the 1990s, **nine men's morris** was shown, by searching through billions of possible endgames, to be a certain draw if both players work to an optimal strategy. Checkers may be the next to be fully determined: its roughly 500 million trillion possible positions may soon be within reach of the most powerful supercomputers. See also **blackjack**, **chess**, **Frogs and Toads**, **Ovid's game**, **TacTix**, and **Wythoff's game**.

game theory

A mathematical formalism used to study human **game**s, economics, military conflicts, and biology. The goal of game theory is to find the optimal **strategy** for one player to use when his opponent also plays optimally. A strategy may incorporate randomness, in which case it is referred to as a **mixed strategy**.

Early ideas of game theory can be found in writings throughout history as diverse as the Bible and works by René **Descartes**, Sun Tzu (author of the 2,400-year-old *The Art of War*), and Charles Darwin. The basis of modern game theory is an outgrowth of several books that deal with related subjects such as economics and probability. These include Augustin Cournot's *Researches into the Mathematical Principles of the Theory of Wealth* (1838), which gives an intuitive explanation of what would eventually be formalized by John **Nash** as *Nash equilibrium;* Francis Edgeworth's *Mathematical Psychics,* which explored the notion of competitive equilibria in a two-type (or two-person) economy; and Emile **Borel**'s *Algebre et calcul des probabilites* (1927), which gave the first insight into so-called mixed strategies.[49] Game theory finally came of age through the efforts of two European immigrants to the United States working at the Institute of Advanced Studies in Princeton. Around 1940, the idea of the **utility function** was taken up by John **von Neumann**, who had been forced to flee his native Hungary when the Nazis invaded, and the economist Oskar Morgenstern (1902–1976), who had left Austria because he loathed the National Socialists. In Princeton the two immigrants worked together on what they initially thought would be a short paper on the theory of games, but that kept growing until it finally appeared in 1944 as an opus of 600 pages with the title *Theory of Games and Economic Behavior.*[230]

GLOSSARY OF GAME THEORY

categorical game	A game in which a tie is impossible.
finite game	A game in which each player has a finite number of moves and a finite number of choices at each move.
futile game	A game that allows a tie when played properly by both players.
impartial game	A game in which the possible moves are the same for each player in any position.
mixed strategy	A collection of moves together with a corresponding set of weights which are followed probabilistically in the playing of a game.
partisan game	A game for which each player has a different set of moves in any position.
payoff matrix	An $m \times n$ matrix that gives the possible outcome of a two-person zero-sum game when player A has m possible moves and player B has n moves.
strategy	A set of moves that a player plans to follow while playing a game.
zero-sum game	A game in which players make payments only to each other. One player's loss is the other player's gain, so the total amount of "money" available is constant.

gamma

See **Euler-Mascheroni constant**.

gamma function

A generalization of the **factorial** function to the real line and to the complex plane. It is defined by:

$$\Gamma(n+1) = \int_0^\infty x^n e^{-x}\, dx$$

If n is an integer, then $\Gamma(n+1) = n!$ See also **beta function**.

Gardner, Martin (1914–)

An American recreational mathematician best known for his "Mathematical Games" column, which ran in *Scien-*

tific American for 25 years. Through this column he introduced many subjects, including **flexagon**s, **polyomino**s, Piet Hein's **Soma Cube**, and John Conway's Game of **Life**, to a wider audience. He is also an accomplished amateur magician and an active member of the skeptical movement associated with James Randi. Gardner is the author of more than 60 books, including various collections of his *Scientific American* columns, *The Ambidextrous Universe,* and *The Annotated Alice.*[108–131]

gauge theory

A force field in nature, or an analogous **vector** field in mathematics with an enormous amount of **symmetry** that expresses the redundancy or ambiguity of many parameters. The simplest example is the *electromagnetic field.*

Gauss, Carl Friedrich (1777–1855)

A German mathematician, often called the "Prince of Mathematics," whose stature and range of interests rivaled those of **Aristotle** and Isaac **Newton**. Some inkling of what was in store came when, as a 3-year-old, he corrected a mistake in one of his father's lengthy payroll calculations. In school, at age 10, when his teacher gave the class the task of adding all the integers from 1 to 100, Gauss immediately wrote down the correct answer, 5050, on his slate. He had spotted that the numbers can be paired off as (100 + 1), (99 + 2), (98 + 3), . . . , (51 + 50) so that the problem reduces to multiplying 101 by 50. At age 19, Gauss found a way to construct a heptadecagon (a **regular polygon** with 17 sides) using only a straightedge and compass–a feat that had eluded the Greeks. Then Gauss entered the mathematical stratosphere of his time by proving what is now called the *fundamental theorem of algebra,* namely, that every **polynomial** has at last one root that is a **complex number**; in fact, he gave four different proofs, the first of which appeared in his dissertation. In 1801, he proved the *fundamental theorem of arithmetic* (that every natural number can be represented as the product of **prime number**s in only one way); published a brilliant tour de force on the properties of integers in his *Disquisitiones Arithmeticae,* which systematized the study of **number theory**; and showed that every number is the sum of at most three **triangular number**s. In the same year, he also developed the method of least squares fitting and, though he didn't publish it, used it to calculate the orbit of the asteroid Ceres, that had recently been discovered by Piazzi, from only three observations. Gauss published his monumental treatise on celestial mechanics *Theoria Motus* in 1806. He became interested in the compass through surveying, and developed the magnetometer, an instrument with which, together with Wilhelm Weber, he measured the intensity of magnetic forces. With Weber, he also built the first successful telegraph.

Unfortunately for mathematics, Gauss reworked and improved papers incessantly, and, in keeping with his motto "*pauca sed matura*" (few but ripe), he published only a fraction of his work. Many of his results were subsequently repeated by and attributed to others, since his terse diary remained unpublished for years after his death. Only 19 pages long, this diary later confirmed his priority on many breakthroughs, including work on an alternative to the **parallel postulate**, which really makes him the earliest pioneer of **non-Euclidean geometry** despite the fact that János **Bólyai** and Nikolai **Lobachevsky** are normally given this accolade. Gauss did, however, publish his seminal treatment on **differential geometry** in *Disquisitiones circa superticies curvas,* and Gaussian **curvature** is named for him. Gauss wanted a heptadecagon placed on his gravestone, but the carver refused, saying it would be indistinguishable from a circle. The heptadecagon appears in the shape of a pedestal with a statue erected in his honor in his hometown of Braunschweig.

Gaussian

Normally distributed (with a bell-shaped curve) and having a **mean** at the center of the curve with tail widths proportional to the **standard deviation** of the data about the mean.

Gelfond's theorem

Also known as the *Gelfond-Schneider theorem:* a^b is a **transcendental number** if (1) a is an **algebraic number** and not equal to either 0 or 1, and (2) b is algebraic and also an **irrational number**. Gelfond's theorem enables the seventh of David **Hilbert**'s famous problems to be solved.

general relativity

See **relativity theory**.

general topology

See **point-set topology**.

genetic algorithm

A type of evolving computer program, developed by the computer scientist John Holland, whose strategy of arriving at solutions is based on principles taken from genetics. Basically, the genetic **algorithm** utilizes the mixing of genetic information in sexual reproduction, random mutations, and natural selection at arriving at solutions.

genus

In **topology**, roughly speaking, the number of **hole**s in a surface. Spheres, bowling balls (the finger holes aren't true holes because they don't go all the way through), and wine glasses have a genus of 0 and can be represented by **quadratic** equations. Bagels, inner tubes, and teacups have a genus of 1 and can be described by **cubic** equations. Humans are more difficult to specify. However, you will certainly increase your genus by one if you have your ear pierced! Different definitions of genus apply to other types of mathematical objects such as a **curve**, a **knot**, or a **set**.

geoboard

A device commonly used in elementary schools to aid in the teaching of basic geometric concepts. A simple geoboard can be made from a square piece of wood and 25 nails arranged in an evenly spaced grid of 5 vertical lines and 5 horizontal lines. These nails represent the **lattice** points in the plane. Figures are made on the geoboard by stretching rubber bands from one nail to another until the desired shape is formed.

geodesic

A path on a given **surface** that is as straight as possible; in other words, a path that doesn't deviate either to the left or to the right, and only bends when forced to do so by the **curvature** (if any) of the surface. If the surface is an ordinary **plane**, the geodesics are straight lines; on a sphere the geodesics are **great circle**s.

geometric magic square

A square array of $n \times n$ cells each occupied by a distinct geometrical figure (or piece or tile), such that the n pieces contained in every row, column, and diagonal can be fitted together to produce (i.e., tile or pack) a constant shape known as the target. The figures may be of any dimension, but are normally planar (topological disks). In tessellating the target, which may be of any shape, planar pieces are allowed to be flipped. Pieces of three or more dimensions are considered distinct from their mirror images. Geometric **magic squares** using one-dimensional entries have been known for centuries; they are the traditional magic squares in which straight lines pave a constant length, as usually represented by numbers adding to a constant total. The properties of generalized geometric magic squares were first investigated by Lee **Sallows**.

geometric mean

The geometric mean of n numbers is the nth root of the product of the numbers.

geometric sequence

Also known as a *geometric progression*, a finite **sequence** of at least three numbers, or an infinite sequence, whose terms differ by a constant multiple, known as the *common ratio*. For example, starting with 3 and using a common ratio of 2 leads to the finite geometric sequence: 3, 6, 12, 24, 48, and also to the infinite sequence 3, 6, 12, 24, 48, . . . , (3×2^n). . . . In general, the terms of a geometric sequence have the form $a_n = ar^n$ $(n = 0, 1, 2, \ldots)$ for fixed numbers a and r. If the terms of a geometric sequence are added together the result is a *geometric series*. If it is a finite series, then we add its terms to get the series sum, $S_n = a + ar + ar^2 + \ldots + ar^n = (a - ar^{n+1})/(1 - r)$. In the case of an **infinite series**, if $|r| < 1$, the sum is $a/(1 - r)$. If $|r| \geq 1$, however, the series diverges and thus has no sum. See also **arithmetic sequence**.

geometry

The study of the properties of shapes and of spaces. See also **Euclidean geometry** and **non-Euclidean geometry**.

geometry puzzles

One of the attractions of puzzles involving shapes, especially **dissection** problems, is that they appeal to the eye and very often don't call for much ability in solving equations and the like. Anyone can try to assemble the pieces of a jigsaw, whether it be of a picture or of a geometric shape, so a mathematical game such as **tangrams** or the **Soma cube** is within everyone's reach. On the other hand, some geometric puzzles call for a basic knowledge of more abstract fields such as algebra and calculus. They may also exploit our sometimes faulty intuition about how different quantities vary in one, two, and three dimensions and about how much information is needed to solve a problem.

As an example of faulty intuition, imagine that Earth, taken to be a perfect sphere with a radius r of 6,378 km, is completely covered by a thin membrane. Now suppose that 1 square meter is added to the area of this membrane to form a larger sphere. By how much does the radius and the volume of this membrane increase? This can be worked out from the formulae for the volume of a sphere $(V = (4/3)\pi r^3)$ and the area of a sphere $(A = 4\pi r^2)$, respectively. It turns out that if the area of the cover is increased by 1 square meter, then the volume it contains is increased by about 3.25 million cubic meters. This seems like a huge amount. However, the new cover wouldn't be very high above the surface of the planet–only about 6 nanometers! As an example of a problem that is both counterintuitive *and* seems to lack sufficient data for its solution see **hole-through-a-sphere problem**.

Gergonne point

In a triangle, the point at which the lines from the vertices (see **vertex**) to the points of contact of the opposite sides with the **inscribed circle** meet.

Germain, Sophie (1776–1831)

A French mathematician who made notable contributions to **number theory** and to mathematical physics despite a lack of formal training and the social prejudices of her day. She taught herself, against her parents' wishes, often at night, during the Reign of Terror following the French Revolution, with books from her father's library. When deprived of heat and light, she would wrap herself in quilts and use candles. Finally her parents acquiesced to her "incurable" passion for mathematics and let her study. Through Joseph **Lagrange**, to whom she had originally submitted work under a pseudonym, she gained access to a circle of distinguished mathematicians, including Carl **Gauss**. Among her most important work was an analysis of Ernst **Chladni**'s studies of vibrating surfaces, and her proof that if $x, y,$ and z are integers and if $x^5 + y^5 = z^5$ then at least one of $x, y,$ or z must be divisible by 5 (a result now known as *Germain's theorem*); this was an important early step towards proving **Fermat's last theorem**.

Get Off the Earth

A famous **vanishment puzzle** by Sam **Loyd**. The picture is made from a rectangular background topped with a circular card, representing the world, that can be rotated. Parts of a number of Chinese men are on each piece. With the world orientated so that the large arrow on it points to the N.E. point on the background, 13 Chinamen can be counted. But when the earth is turned slightly, so that the arrow points N.W., there are only 12 characters. Where did the thirteenth Chinaman go? The cleverness of the puzzle is that there are many bits of Chinamen–arms, legs, bodies, heads, and swords–and each has tiny slivers missing. When the earth is turned, these pieces get slightly rearranged. In particular, each of the 12 Chinamen gains a sliver of a Chinaman from his neighbor.

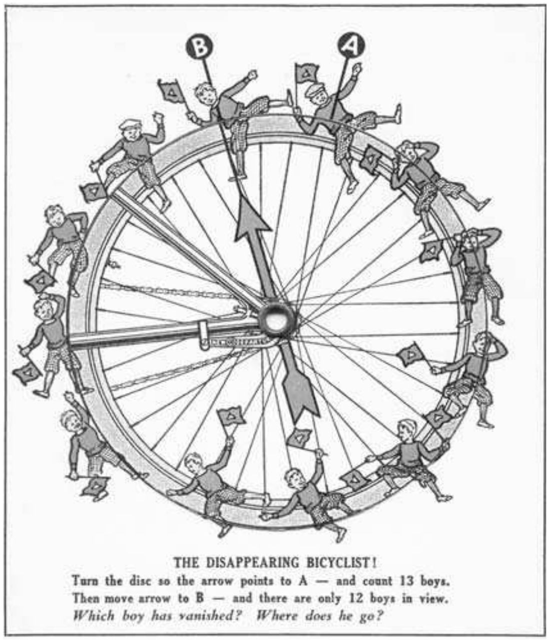
B
A
THE DISAPPEARING BICYCLIST!
Turn the disc so the arrow points to A — and count 13 boys.
Then move arrow to B — and there are only 12 boys in view.
Which boy has vanished? Where does he go?

Get Off the Earth An unusual variant of Loyd's Get Off the Earth puzzle called "The Disappearing Bicyclist." *From the collection of William Waite*

Gettier problem

A thought experiment in philosophy that throws into question the long-held supposition that to know something is equivalent to holding a belief about something that is both true and for which there is justification. Consider a case in which a lecturer has two students in her class called Mr. Havenot and Mr. Havegot. Mr. Havenot claims to own a Ferrari, drives one around, and has papers that state that the car is his. But in fact he does not actually own the car. Mr. Havegot, on the other hand, who shows no sign of Ferrari ownership, secretly has one of these rare cars. On the basis of the evidence, the teacher concludes that one of her students owns a Ferrari–and is correct in this belief. However, there is something wrong. Despite the combination of truth, justification, and belief, it seems that there is no real knowledge. The first examples of such problems were published in 1963 by the American philosopher Edmund Gettier (1927–).

Giant's Causeway

A natural structure that occurs on the coast in County Antrim, Ireland; it is one of the few places in the world where volcanic basalt has cooled in a columnar formation. The columns approximately form a **hexagon**al tessellation (see **tiling**) and tend to break off to produce a pavement with this pattern. The full length of the columns can't be seen, but it is estimated that they may be 20 feet (about 6 meters) high before merging into the underlying irregular basaltic mass. About 99% of the columns are believed to be hexagonal and only one triangular column is known. Though many of the hexagons are fairly regular, some have a side twice as long as their smallest side. Side lengths vary from 8 to 18 inches (20 to 46 cm) and the pillars break up into sections 6 to 36 inches (15 to 90 cm) long, with a concavo-convex junction rather than a plane junction. Other examples of such formations occur at Kirkjubaejarklaustri, in Iceland, and the Devil's Postpile, in California.

Gilbreath's conjecture

A strange hypothesis concerning **prime number**s that was first suggested in 1958 by the American mathematician and amateur magician Norman L. Gilbreath following

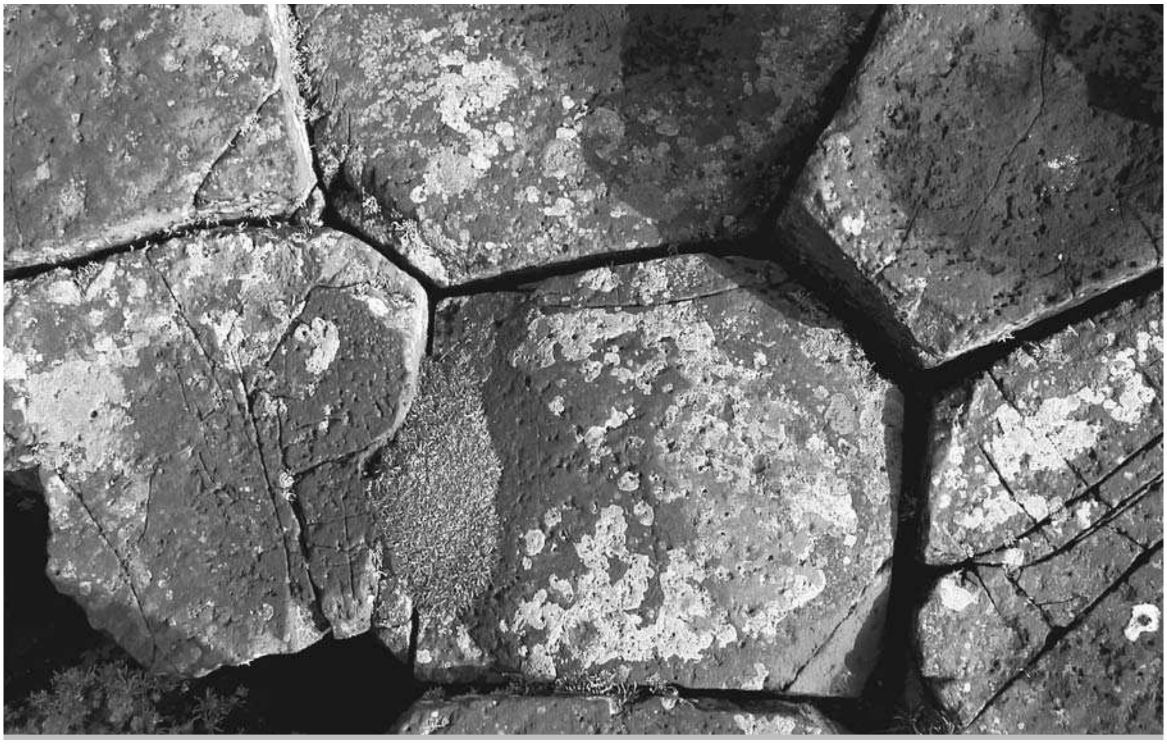

Giant's causeway Hexagonal paving on the Giant's Causeway. *Martin Melaugh, University of Ulster*

some doodlings on a napkin. Gilbreath started by writing down the first few primes:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, \ldots$$

Under these he put their differences:

$$1, 2, 2, 4, 2, 4, 2, 4, 6, 2, \ldots$$

Under these he put the unsigned difference of the differences:

$$1, 0, 2, 2, 2, 2, 2, 2, 4, \ldots$$

And he continued this process of finding iterated differences:

$$\begin{array}{l} 1, 2, 0, 0, 0, 0, 0, 2, \ldots \\ 1, 2, 0, 0, 0, 0, 2, \ldots \\ 1, 2, 0, 0, 0, 2, \ldots \\ 1, 2, 0, 0, 2, \ldots \\ 1, 2, 0, 2, \ldots \\ 1, 2, 2, \ldots \\ 1, 0, \ldots \\ 1, \ldots \end{array}$$

Gilbreath's conjecture is that, after the initial two rows, the numbers in the first column are all one. No exception has been found to date, despite searches out to several hundred billion rows, and the conjecture is generally assumed to be true. However, it may have nothing to do with primes as such. The English mathematician Hallard Croft has suggested the conjecture may apply to *any* sequence that begins with 2 and is followed by odd numbers that increase at a "reasonable" rate and with gaps of "reasonable" size. If this is the case, Gilbreath's conjecture may not be as mysterious as it first seems, though it may be very difficult to prove.

glissette

If there are two fixed curves, and a curve S of fixed shape and length that slides with its ends on the fixed curves, then the **locus** of a point moving with S is called a glissette. An example is the locus of the midpoint of a line segment sliding with its ends on two perpendicular lines; this locus is a circle.

gnomon magic square

A 3×3 array in which the elements in each 2×2 corner have the same sum. See also **magic square**.

Go

A two-player board game that originated in China in about 2000 B.C. It is often compared and contrasted with **chess**. Go is played on a board marked with 19 × 19 lines. Round, lens-shaped pieces called stones are placed, one per move, on the intersections of this grid, of which there are 361. As in chess, the pieces are colored black and white, but in Go black plays first. The board starts blank and pieces once played are not thereafter moved except to be taken off as prisoners. Pieces are captured singly or en masse by being surrounded so that they are not connected to any adjacent open intersection. The player with highest score at the end of the game, or following resignation or time expiry, wins. Go is considered to be a deeply strategic game, unlike chess, which is largely tactical. There are 32,940 opening moves, after symmetry is taken into account, 992 of which are deemed strong. Estimates of the number of possible board configurations vary but are typically on the order of 10^{174}.

God

> *I'm still an atheist, thank God.*
> –Luis Buñuel (Spanish film director, 1900–1983)

Mathematicians, logicians, and scientists have long debated the nature, existence, and dice-playing ability of a Higher Power. The pre-Renaissance French philosopher Jean Buridan (c. 1295–1358) used a version of the *liar paradox* to "prove" the existence of God. He wrote these two sentences:

God exists.

None of the sentences in this pair is true.

The only consistent way to have these two sentences be either true or false is for "God exists" to be true. (However, there is nothing to say that such consistency is necessary.) Blaise **Pascal** gave a more persuasive argument, not for the existence of God but for why we should believe in that existence: "If I believe in God and life after death and you do not, and if there is no God, we both lose when we die. However, if there is a God, you still lose and I gain everything." Pierre **Laplace**, on the other hand, replying to **Napoleon Bonaparte**, who had asked why his celestial mechanics made no mention of God, said: "Sir, I have no need of this hypothesis." The German mathematician Leopold **Kronecker** thought that "God made the Integers, all the rest is the work of man." In *The City of God,* however, Saint Augustine seems to imply that the integers were independent of God. He wrote: "**Six** is a number perfect in itself, and not because God created the world in six days; rather the contrary is true. God created the world in six days because this number is perfect, and it would remain perfect, even if the work of the six days did not exist." Augustine's statement can be taken to suggest that six would be a perfect number not only if the universe didn't exist, but even if God didn't exist. As to God's mathematical specialty, Plato said, "God ever geometrizes" while Charles Jacobi insisted that "God ever arithmetizes." James Jeans thought, "The Great Architect of the Universe now begins to appear as a pure mathematician," and Einstein ("God does not play dice") was sure He wasn't a probabilist.

Gödel, Kurt (1906–1978)

An Austrian-American mathematician and logician who, in 1931, proved that within a **formal system** questions exist that are neither provable nor disprovable on the basis of the **axiom**s that define the system. This is known as Gödel's undecidability theorem. He also showed that in a sufficiently rich formal system in which decidability of all questions is required, there will be contradictory statements. This is called **Gödel's incompleteness theorem**. In establishing these theorems Gödel showed that

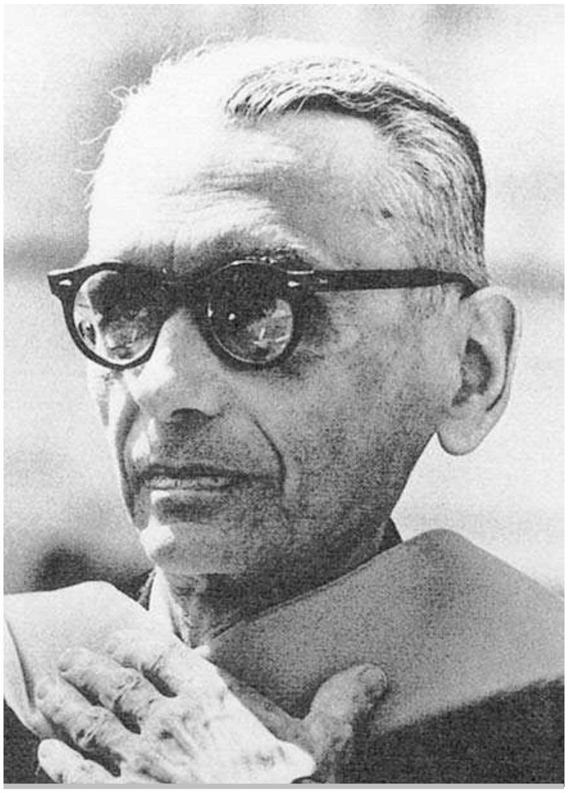

Gödel, Kurt

there are problems that can't be solved by any set of rules or procedures; instead, for these problems one must always extend the set of axioms. This disproved a common belief at the time that the different branches of mathematics could be integrated and placed on a single logical foundation. Gödel was a close friend of Albert Einstein at Princeton and contributed to his general **relativity theory** and cosmology. The so-called *Gödel universe* is a rotating model of the universe in which it is theoretically possible to travel into the past (see **time travel**).

Gödel's incompleteness theorem

In a nutshell: All consistent axiomatic systems contain undecidable propositions. What does this mean? An axiomatic system consists of some undefined terms, a number of **axiom**s that refer to those terms and partially describe their properties, and a rule or rules for deriving new propositions from already existing propositions. Axiomatic systems are powerful because they reduce large bodies of math to a simple description. Also, because they're very abstract, they allow all, and only, the results that follow from things having the formal properties specified by the axioms to be derived. An axiomatic system is *consistent* if, given the axioms and the derivation rules, it doesn't lead to any contradictions. One of the first modern axiomatic systems was a formalization of simple arithmetic (adding and multiplying whole numbers) by Giuseppe **Peano** and now known as *Peano arithmetic.* Kurt **Gödel** showed that every syntactically correct proposition in Peano arithmetic can be represented by a unique integer, called its *Gödel number.* The trick is to replace each symbol in the proposition, *including numerals,* by a different string of digits. If we represent "1" by 01, "2" by 02, "+" by 10, and "=" by 11, then the Gödel number of "1 + 1 = 2" is 0110011102. This allowed Gödel to write down, unambiguously, propositions about propositions. In particular, he was able to write down self-referential (see **self-referential sentence**) propositions–ones that include their own Gödel number. Gödel was then able to prove that, *either* the system of Peano arithmetic is inconsistent, *or* there are true propositions that can't be reached from the axioms by applying the derivation rules. The system is thus *incomplete,* and the truth of those propositions is *undecidable* (within that system). Such undecidable propositions are known as *Gödel propositions* or *Gödel sentences.* Nobody knows what the Gödel sentences for Peano arithmetic are, though people have their suspicions about the **Goldbach conjecture** (every even number is the sum of two prime numbers).

The results of an axiomatic system pertain to more than just Peano arithmetic, they apply to all kinds of things that satisfy the axioms. There are an immense number of other axiomatic systems, which either include Peano numbers among their basic entities or can be constructed from them. It follows that these systems, too, contain undecidable propositions, and are incomplete.

A common misconception is that Gödel's theorem imposes some profound limitation on knowledge, science, and mathematics. In the case of science, this ignores that Gödel's theorem applies to *deduction from axioms,* which is only one source of knowledge and not even a very common mode of reasoning in science. More generally, Gödel's incompleteness result doesn't touch directly on the most important sense of completeness and incompleteness, namely, descriptive completeness and incompleteness–the sense in which an axiom system describes a given field. In particular, the result represents no threat to the notion of truth.

Goldbach conjecture

One of the oldest and easiest-to-understand hypotheses in mathematics that remains unproven. In its original form, now known as the *weak Goldbach conjecture,* it was put forward by the Prussian amateur mathematician and historian Christian Goldbach (1690–1764) in a letter dated June 7, 1742, to Leonhard **Euler**. In this guise it says that every whole number greater than five is the sum of three **prime number**s. Euler restated this, in an equivalent form, as what is now called the *strong Goldbach conjecture* or, simply, the Goldbach conjecture: every even number greater than two is the sum of two primes. Thus, 4 = 2 + 2, 6 = 3 + 3, 8 = 3 + 5, 10 = 3 + 7, . . . , 100 = 53 + 47, ... In fact René **Descartes** knew about the two-prime version of Goldbach's conjecture before either Goldbach or Euler did. So, is it misnamed? Paul **Erdös** said, "It is better that the conjecture be named after Goldbach because, mathematically speaking, Descartes was infinitely rich and Goldbach was very poor." In any event, there is a much more important question, namely, is the conjecture true? The general assumption is that it is, but no one knows for sure. The most significant step toward a proof came in 1966 when the Chinese mathematician Chen Jing-Run showed that every sufficiently large even integer is the sum of a prime and a number that has at most two prime factors. Using powerful computers, the Goldbach conjecture has been checked out to about 400 trillion. But there is no great optimism among mathematicians that a final breakthrough is on the horizon. Even a reward of $1 million dollars for a proof offered by the publishing house Faber & Faber in 2000, to help publicize the novel *Uncle Petros and Goldbach's Conjecture* by the Greek mathematician and author Apostolos Doxiadis, went unclaimed.[85]

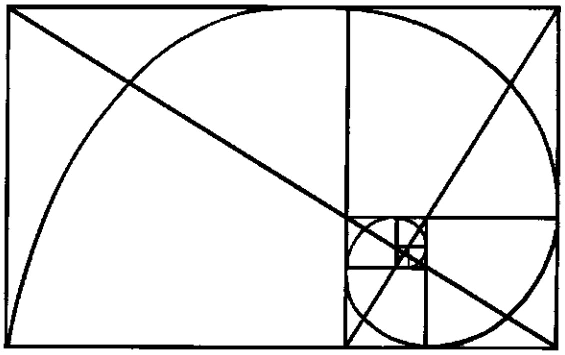

golden ratio A golden rectangle and a logarithmic spiral emerge from a whirling pattern of squares built up from two small squares of equal size at the spiral's center.

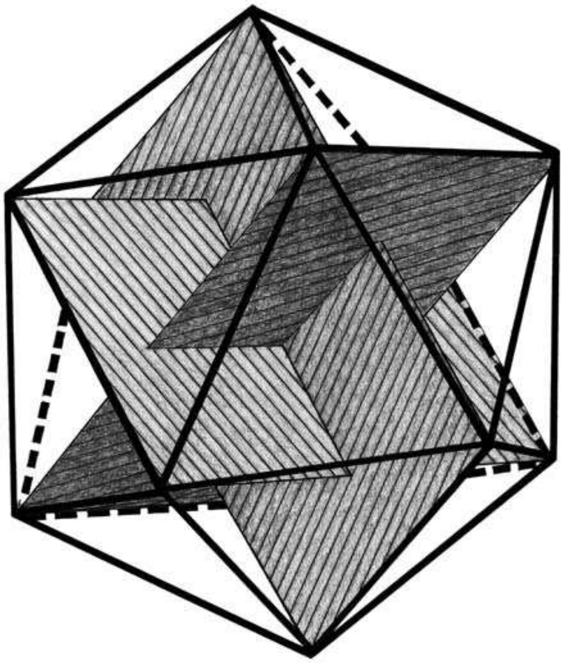

golden ratio The corners of an icosahedron meet the corners of three orthogonal golden rectangles.

golden ratio (phi, ϕ)

A remarkable number that, like **pi** and *e,* pops up all over the place in mathematics but, in some ways, has a more "human" connection, in that it seems to be linked to aesthetics. Its name, which is also given as the *golden mean,* the *golden section,* the *golden number,* and the *divine proportion,* reflects this sense of a harmonious or pleasing ideal. The golden ratio is an **irrational number** of the type known as an **algebraic number** (in contrast with π and e, which are transcendental) and is represented by the Greek letter ϕ (phi). It can be defined in various ways. For example, it is the only number equal to its own reciprocal plus 1, that is, $\phi = (1/\phi) + 1$, so that $\phi^2 = \phi + 1$. From this comes the quadratic equation $\phi^2 - \phi - 1 = 0$ of which the golden ratio is the positive solution, $(1 + \sqrt{5}) / 2 \approx 1.6180339887\ldots$. The golden ratio is also approximated by the ratio of successive terms in the **Fibonacci sequence**; in fact, $F(n + 1) / F(n)$ gets closer and closer to ϕ as n tends to infinity. Because $1/(1 - \phi) = \phi$, the **continued fraction** representation of ϕ is

$$\begin{aligned}\phi &= 1 + 1/(1 + 1/(1 + 1/(1 + 1/(1 + 1/\ldots \\ &= [1;1,1,1,\ldots].\end{aligned}$$

Two quantities are said to be in the golden ratio, if the ratio of the larger one, a, to the smaller one, b, is the same as the ratio of the smaller one to their difference, that is, $a/b = b/(a - b)$. The so-called *golden rectangle* is one whose sides a and b stand in the golden ratio. It is famously said to have great aesthetic appeal and is closely approximated by the dimensions of the front of the Parthenon in Rome. Leonardo da Vinci's masterpiece the *Mona Lisa* is said to have a face that is framed by a golden rectangle; what is certain is that Leonardo was a close personal friend of Luca **Pacioli**, who published a three-volume treatise on the golden ratio, *Divina Proportione,* in 1509. The Swiss-French architect and painter Le Corbusier designed an entire proportional system called the "Modulor," that was based on the golden ratio. The Modulor was supposed to provide a standardized system that would automatically confer harmonious proportions to everything, from door handles to high-rise buildings. Another artist who deliberately used the golden ratio is the surrealist Salvador Dali. The ratio of the dimensions of Dali's *Sacrament of the Last Supper* is equal to the golden ratio. Dali also incorporated in the painting a huge dodecahedron (a twelve-faced Platonic solid in which each side is a pentagon) engulfing the supper table. The dodecahedron, which according to Plato is the solid "which the god used for embroidering the constellations on the whole heaven," is intimately related to the golden ratio–both the surface area and the volume of a dodecahedron of unit edge length are simple functions of the golden ratio. In fact, ϕ turns up frequently in figures that have pentagonal symmetry. For instance the ratio of a regular pentagon's side and diagonal is equal to ϕ, and the vertices of a regular **icosahedron** are located on three orthogonal golden rectangles. The golden ratio is also related to **Penrose tiling** and to the **plastic number**.[205]

Golomb, Solomon W. (1932–)

A mathematician and electrical engineer at the University of Southern California who is best known for his seminal studies of **polyomino**s. His article "Checker Boards and Polyominos," published in the *American Mathematical Monthly* in 1954 when Golomb was a 22-year-old graduate student at Harvard, defined a polyomino as a simply connected set of squares (i.e., a set of squares joined along their edges). Golomb also originated the idea of **graceful graph**s.[137]

golygon

A series of straight-line segments that have lengths of one, two, three, and so on, up to some finite number of units, with the property that every segment connects at a right angle to the segment that is one unit larger, except the longest segment, which meets the shortest segment at a right angle. The name "golygon" was invented by Lee **Sallows**. Golygons have inspired some interesting puzzles as well as some intriguing research problems.

googol

A named coined in 1938 by Milton Sirotta, the 9-year-old nephew of the mathematician Edward **Kasner**, when the child was asked by his uncle to come up with a name for a very **large number**; at the same time, *googolplex* was suggested for a still larger number. Kasner defined these numbers as follows:

1 googol = 10^{100} (i.e., 1 followed by 100 zeros);
1 googolplex = $10^{\text{googol}} = 10^{10^{100}}$ (i.e., 1 followed by a googol number of zeros)

A googol is very roughly the number of years thought to be needed for all the black holes in the universe to evaporate by a process known as Hawking radiation. It is much larger than the number of protons and neutrons in the known universe (about 10^{80}), but much smaller than the number of protons and neutrons needed to pack every cubic centimeter of the known universe (about 10^{128}). The googolplex is the largest number with a proper name of which many people have heard. It is dwarfed, however, by such esoterica as **Graham's number**.

Gordian knot

The earliest reference to a **string puzzle**. In Greek mythology, a Phrygian peasant called Gordius, the father of Minos (see **maze**), became king because he was first to arrive in town after an oracle commanded the Phrygians to select as ruler the first person to drive into the public square in a wagon. In gratitude, Gordius dedicated his wagon to Zeus and placed it in the temple grove, tying the wagon pole to the yoke with a rope of bark. The knot was so intricately entwined that no one could undo it. A saying developed that whoever succeeded in untying the knot would become ruler of all Asia. Many tried, but all failed. According to legend, even Alexander the Great was unable to untie the Gordian knot, so he drew his sword and cut it through with a stroke. The expression "to cut the Gordian knot" is used to refer to a situation in which a difficult problem is solved by a quick and decisive action.

graceful graph

A **graph** of points and connecting lines that can be numbered in a certain way. Say the graph has p points and e lines ("e" for edges) connecting them. Each of the points is assigned an integer; the lowest integer (by convention) is taken to be 0, and no two integers may be alike. Each of the lines is labeled with the difference between the two integers of the points that it connects. Then, if the numbers corresponding with the lines run from 0 through e, the graph is said to be graceful. Graceful graphs were originally defined and developed by Solomon **Golomb**.

gradient

A **vector** of partial **derivative**s of a **function** that operates on vectors. Intuitively, the gradient represents the slope of a high-dimensional surface.

Graham, Ronald L. (1936–)

An American mathematician and leading combinatorialist after whom **Graham's number** is named. Graham is also one of the country's best jugglers and former president of the International Juggler's Association. In his youth, he and two friends were professional trampolinists who performed with a circus as the Bouncing Baers. His office ceiling is covered with a large net that he can lower and attach to his waist so that when he practices juggling with six or seven balls, any that are dropped will roll back to him. Graham is a professor in the department of computer science and engineering at the University of California at San Diego.

Graham's number

A stupendously **large number** that found its way in to the *Guinness Book of Records* as the biggest number ever obtained as part of a mathematical proof; it is named after its discoverer, Ronald **Graham**. Graham's number is the upper bound solution to a very exotic problem in **Ramsey theory**, namely: What is the smallest dimension n of a **hypercube** such that if the lines joining all pairs of corners are two-colored, a planar complete graph K_4 of one color will be forced? This is exactly equivalent to a problem that can be stated in plain language: Take any number of people, list every possible committee that can be formed from them, and consider every possible pair of committees. How many people must be in the original group so that no matter how the assignments are made, there will be four committees in which all the pairs fall in the same group, and all the people belong to an even number of committees. Graham's number is the *greatest* value that the answer could take. It is so large that it can only be written using special big-number notation, such as **Knuth's up-arrow notation**. Even then, it must be built up in stages. First, construct the number $G_1 = 3\uparrow\uparrow\ldots\uparrow\uparrow 3$, where there are $3\uparrow\uparrow\uparrow\uparrow 3$ up-arrows. This, in itself, is a number far, far beyond anyone's ability to even remotely comprehend. Next, construct $G_2 = 3\uparrow\uparrow\ldots\uparrow\uparrow 3$, where there are G_1 up-arrows; then construct $G_3 = 3\uparrow\uparrow\ldots\uparrow\uparrow 3$, where there are G_2 up-arrows; then continue this pattern until the number G_{62} has been made. Graham's number $G = 3\uparrow\uparrow\ldots\uparrow\uparrow 3$, where there are G_{62} up-arrows. While the unthinkably large upper boundary to the problem described earlier is given by Graham's number, nobody, including Graham himself, believes the solution is nearly so large. In fact, it is thought that the actual answer is probably 6![115]

grandfather paradox

One of the most powerful and commonly used arguments against **time travel**. It points out that if you were able to travel into the past you could (if you were so inclined) kill your grandfather when he was very young and thus render your own birth impossible. A simpler version is that you could kill a younger version of yourself so that you would not be alive in the future to travel back in time. The grandfather paradox shows how one form of time travel could violate **causality** by eliminating the cause of a phenomenon that has already taken place in the present.

The most bizarre adaptation of the grandfather paradox is found in Robert Heinlein's classic short story "All You Zombies." A baby girl is mysteriously left at an orphanage in Cleveland in 1945. "Jane" grows up lonely and dejected, not knowing who her parents are, until one day in 1963 she is strangely attracted to a drifter. She falls in love with him. But just when things are finally looking up for Jane, a series of disasters strike. First, she becomes pregnant by the drifter, who then disappears. Second, during the complicated delivery, doctors find that Jane has both sets of sex organs, and to save her life, they are forced to surgically convert "her" to a "him." Finally, a mysterious stranger kidnaps her baby from the delivery room. Reeling from these disasters, rejected by society, scorned by fate, "he" becomes a drunkard and drifter. Not only has Jane lost her parents and her lover, but he has lost his only child as well. Years later, in 1970, he stumbles into a lonely bar, called Pop's Place, and spills out his pathetic story to an elderly bartender. The bartender offers the drifter the chance to avenge the stranger who left her pregnant and abandoned, on the condition that he (Jane) join the "time travelers corps." Both of them enter a time machine, and the bartender drops off the drifter in 1963. The drifter is strangely attracted to a young orphan woman, who subsequently becomes pregnant. The bartender then goes forward nine months, kidnaps the baby girl from the hospital, and drops off the baby in an orphanage back in 1945. Then the bartender drops off the thoroughly confused drifter in 1985, to enlist in the time travelers corps. The drifter eventually gets his life together, becomes a respected and elderly member of the time travelers corps, and then disguises himself as a bartender and has his most difficult mission: a date with destiny, meeting a certain drifter at Pop's Place in 1970. The question is: who is Jane's mother, father, grandfather, grandmother, son, daughter, granddaughter, and grandson? The girl, the drifter, and the bartender, of course, are all the same person. As an exercise (on the road to insanity) try drawing Jane's family tree. You will find that not only is she her own mother and father, she is an entire family tree unto herself!

graph

> *I'll do algebra, I'll do trig, and I'll even do statistics,*
> *but graphing is where I draw the line!*
>
> –Anonymous

(1) In common usage, a plot of x values (the domain) against y values (the codomain) for a given **function**, $y = f(x)$. Such a graph is also known as a *function graph* or the *graph of a function.* (2) In strict mathematical usage, any set of dots, known as *nodes* or *vertices,* in which at least some pairs are joined by lines known as *edges* or *arcs.* What follows applies only to this second definition.

Often the lines on a graph are used to represent relationships between objects (represented by dots). Depending on the application, edges may or may not have a direction, as indicated by an arrow (see **directed graph**); edges joining a node to itself may or may not be allowed, and nodes and/or edges may be assigned weights. A *path* is a series of nodes such that each node is adjacent to both the preceding and succeeding node. A path is considered *simple* if none of the nodes in the path is repeated. The *length* of a path is the number of edges that the path uses, counting multiple edges multiple times. If it's possible to establish a path from any node to any other node of a graph, the graph is said to be a **connected graph**. A *circuit* or *cycle* is a path that begins and ends with the same node and has a length of at least two. A *tree* is a connected acyclic graph, that is, a graph without any circuits. A **complete graph** is one in which every node is adjacent to every other node. An **Euler path** in a graph is a path that uses each edge precisely once. If such a path exists, the graph is said to be *traversable.* An **Euler circuit** is a path that traverses each edge precisely once. A **Hamilton path**

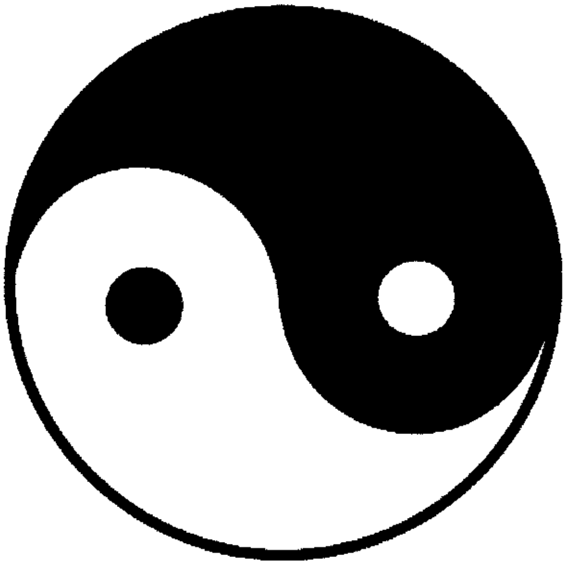

Great Monad

in a graph is a path that visits each node once and only once; a **Hamilton circuit** is a **circuit** that visits each node once and only once. Well known problems whose solution involves graphs and graph theory include the **four-color problem** and the **traveling salesman problem**.

graph theory

The study of **graph**s, either for their own sake, or as models of such diverse things as **group**s (in pure mathematics) or computer networks.

great circle

A circle that goes all the way around a **sphere** and is centered at the center of the sphere. The shortest route between two points on a sphere, such as Earth, is along the great circle that connects these points. A great circle is a **geodesic**.

Great Monad

Also known as *T'ai-Chi,* an important and ubiquitous symbol in traditional Chinese philosophy and cosmology. It represents the underlying harmony of the universe when its opposites or dualities–male and female (*yang* and *yin*), hard and soft, sun and moon, and so forth–are in balance. It occurs everywhere in Chinese art: in books, on walls, porcelain, tablets, and stitched into brocade.

greatest common divisor

The largest integer that divides each of a sequence of integers exactly. Also called the *greatest common factor.*

greatest lower bound

The largest **real number** that is smaller than each of the numbers in a **set** of real numbers.

Green, George (1793–1841)

An English mathematician who published work in the fields of hydrodynamics, electricity, and magnetism, but is best known for his theorem (see **Green's theorem**), which is the basis of **potential theory**. Green took over a bakery and adjoining windmill after the death of his father, but studied mathematics in his spare time. In 1828, he wrote his most important paper, "An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism," which, though generally overlooked at the time, is now regarded as the beginning of mathematical physics in England.

Green's theorem

A connection between path integrals over a well-connected region in the plane and the area of the region bounded in the plane. Green's theorem is a form of the *fundamental theorem of calculus,* and is used today in almost all computer codes that solve **partial differential equation**s.

Grelling's paradox

An equivalent, from the world of words and grammar, of **Russell's paradox**. Grelling's paradox involves dividing all adjectives into two sets: self-applicable and not self-applicable. Words like "English," "written," and "short" are self-applicable, while "Russian," "spoken," and "long" are not self-applicable. Now, define the adjective *heterological* to mean "not self-applicable." To which set of adjectives does "heterological" belong? This strange quandry was devised by the logician and philosopher Kurt Grelling (1886–1941/2), who was persecuted by the Nazis; it is not certain whether he died with his wife in the Auschwitz concentration camp in 1942, or whether he was killed in 1941 in the Pyrenees while trying to escape into Spain.

gross

A group of 144 items. The word comes from the Latin *grossus,* for "thick" or "large," via the Old French *gross douzaine* or "large dozen" (12 dozen), though this grouping may have started out in Germany. "Grocer" has the same origins as "gross" because a grocer is someone who deals in large quantities of food. A *great gross,* or a dozen gross, is 1728. See also **twelve**.

group

> *Wherever groups disclosed themselves, or could be introduced, simplicity crystallized out of comparative chaos.*
>
> –Eric Temple Bell

An abstract and crucially important way of representing **symmetry** and one of the most fundamental concepts in modern **algebra**. Groups were brought into mathematics in the early nineteenth century by the radical young French student Evariste **Galois** as a tool to help solve one of the outstanding problems of his day: to find a formula for solving **polynomial** equations of order five–**quintic**s–and higher. Galois showed, in notes scribbled down the night before he died in a duel, that no such formula exists. The reason for this is that the possible symmetries, or permutations, of the **root**s of fifth-degree polynomial equations are more complex than are the symmetries that can be represented by arithmetical formulas. This fact emerged from the development of the idea of a *permutation group* by Galois and, independently at about the same time, by Niels **Abel**. Half a century later, another Norwegian, Sophus **Lie**, showed how important groups are to the whole of mathematics. The theory of what became known as **Lie group**s links the discrete structure of permutations with the continuous variation of **differential equation**s. Not surprisingly, because group theory forms a common underpinning to algebra and to geometric features such as rotation, reflection, and symmetry, it crops up routinely in modern physics, from the classification of elementary particles to crystallography.

A group is a **set** whose elements are defined by a single operation. The group is called *additive* if the symbol for the operation is "+" and is called *multiplicative* if the symbol is "·" for multiplication. But any other symbol can be substituted for these. There is always a unique element (1, for multiplicative, and 0, for additive, groups) that leaves elements unchanged under the defined operation, like $a + 0 = a$. Also, for every element a there exists a unique inverse b such that, for example, in the case of the additive symbol, $a + b = 0$ and $b + a = 0$. Most often, however, the inverse is denoted as a^{-1}. Lastly, the group operation must be **associative** as in $a \cdot (b \cdot c) = (a \cdot b) \cdot c$. A group is *commutative* or *Abelian* if its operation is symmetric, as in $a + b = b + a$.

Groups come in two types: finite and infinite. The symmetry group of the roots of a polynomial equation is a finite group, because there is only a limited number of permutations possible among the roots of a given polynomial. In contrast, the Lie groups that represent symmetries of solutions of differential equations are infinite because they represent continuous transformations, and continuity carries the potential of an infinite number of changes. Finite groups can be built up from combinations of smaller groups by a process analogous to multiplication. In the same way that a whole number can be written as a product of **prime number**s, a finite group can be expressed as a combination of certain factors known as *simple groups.* Most simple groups belong to one of three families: the *cyclic groups,* the *alternating groups,* or the *groups of Lie type.* Cyclic groups consist of cyclic permutations of a prime number of objects. Alternating groups consist of even permutations–those formed by interchanging the positions of two objects an even number of times. Sixteen subfamilies make up the simple groups of Lie type, each associated with a family of infinite Lie groups. (Confusingly, a Lie group is not a group of Lie type, since the former is infinite and the latter is finite!) Altogether, there are 18 specific families of finite simple groups. There are also 26 simple groups, known as *sporadic groups,* that are highly irregular and fall outside these families. Five sporadic groups were found in the nineteenth century by Emile Mathieu. Then came a hiatus until the 1960s, when suddenly a rush of new sporadics came to light. The most remarkable of these is the so-called **monster group**, which appears to be intimately related to the structure of the universe at the subatomic level.

Grundy's game

See **Nim**.

Guy, Richard Kenneth (1916–)

A British-born mathematician who is professor emeritus of mathematics at the University of Calgary, Canada, and an expert in **combinatorics** and in **number theory**. Guy is the author of more than 250 papers and 10 books, including (as a coauthor) the game theory classic *Winning Ways.* He has been an editor of the "Problems" section of the *American Mathematical Monthly* since 1971.

H

Haberdasher's puzzle

The greatest mathematical discovery of Henry **Dudeney**, it was first published in the *Weekly Dispatch in* 1902 and then as problem no. 26 in his *The Canterbury Puzzles* (1907).[87] One must decide how to cut an equilateral triangle into four pieces that can be rearranged to make a square. The accompanying diagram shows the solution, which Dudeney describes as follows:

> Bisect AB in D and BC in E; produce the line AE to F making EF equal to EB; bisect AF in G and describe arc AHF; produce EB to H, and EH is the length of the side of the required square; from E with distance EH, describe the arc HJ, and make JK equal to BE; now from the points D and K drop perpendiculars on EJ at L and M.

A remarkable feature of the solution is that each of the pieces can be hinged at one vertex, forming a chain that can be folded into the square or the original triangle. Two of the hinges bisect sides of the triangle, while the third hinge and the corner of the large piece on the base cut the base in the approximate ratio 0.982:2:1.018. Dudeney showed just such a model of the solution, made of polished mahogany with brass hinges, at a meeting of the Royal Society on May 17, 1905.

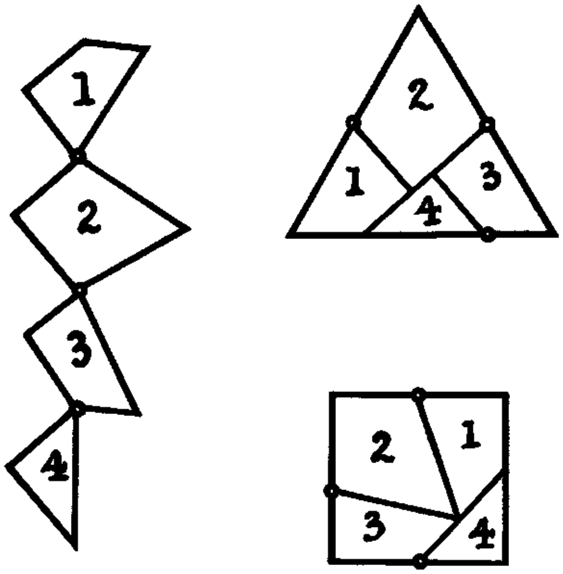

Haberdasher's puzzle The puzzle and its solution as illustrated by Henry Dudeney.

Hadwiger problem

In d dimensions, define $L(d)$ to be the largest integer n for which a cube cannot be cut into n cubes (not necessarily different). The Hadwiger problem is to find $L(d)$. Definite solutions are only known in two and three dimensions: $L(2) = 5$ and $L(3) = 47$. However, it is known that $L(4) \leq 853$ and $L(5) \leq 1,890$, and it is considered likely that $L(d)$ is odd for all values of d. See also **dissection**.

hailstone sequence

A sequence of numbers produced by the rules of the **Collatz problem**; in other words, a sequence formed in the following way: Start with any positive integer n. (1) If n is even, divide it by 2; if n is odd, multiply it by 3 and add 1. (2) If the result is not 1, repeat step (1) with the new number. For $n = 5$, this produces the sequence 5, 16, 8, 4, 2, 1, 4, 2, 1, ... For $n = 11$, the resulting sequence is 11, 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1, 4, 2, 1, ... The name "hailstone" comes from the fact that the numbers in these sequences rise and fall like hailstones in a cloud before finally falling to Earth. It seems from experiment that such a sequence will always eventually end in the repeating cycle 4, 2, 1, 4, 2, 1, . . . , but some values for n generate many values before the repeating cycle begins. An unsolved mystery is whether all such sequences eventually hit 1 (and then 4, 2, 1, 4, 2, 1, . . .) or whether there are some sequences that never settle down to a repeating cycle.

hairy ball theorem

If a sphere is covered with hair or fur, like a tennis ball, the hair cannot be brushed so that it lies flat at every point. In mathematical terms: any continuous tangent vector field on the sphere must have a point where the vector is zero. This theorem also means that somewhere on Earth's surface there has to be a point where the horizontal wind speed is zero, even if it's windy everywhere else. Does the same apply to a **torus**? Is there a hairy donut theorem? No! The number of "problem points," where the hair would stick up on a surface, is related to a quantity called the **Euler characteristic** of that surface. Basically, every point on a surface has an index that describes how many times the vector field rotates in a neighborhood of the problem point. The sum of the indices of all the vector fields is the Euler characteristic. Since the torus has Euler number 0, it is possible to have a covering of hair–a vector field–on it that lies flat at every point.

half-line

A **ray**.

half-plane

The part of a **plane** that lies on one side of a given line.

halting problem

Given a program and inputs for it, decide whether it will run forever or will eventually stop. This is not the same thing as actually running a given program and seeing what happens. The halting problem asks whether there is any *general prescription* for deciding how long to run an arbitrary program so that its halting or non-halting will be revealed. In a celebrated 1936 paper,[337] Alan **Turing** proved that the halting problem is undecidable: there's no way to construct an **algorithm** that is *always* able to determine whether another algorithm halts or not. From this it follows that there can't be an algorithm that decides whether a given statement about **natural numbers** is true or not. The undecidability of the halting problem provides an alternative proof of **Gödel's incompleteness theorem**. This is because if there were a complete and consistent axiomatization of all true statements about natural numbers, then we would be able to create a set of rules that decides whether such a statement is true or not. Another amazing consequence of the undecidability of the halting problem is *Rice's theorem,* which states that the truth of *any* nontrivial statement about the **function** that is defined by an algorithm is undecidable. So, for example, the decision problem "will this algorithm halt for the empty string" is already undecidable. Note that this theorem holds for the *function defined by the algorithm* and not the algorithm itself. It is, for example, quite possible to decide if an algorithm will halt within 100 steps, but this isn't a statement about the function that is defined by the algorithm. Many problems can be shown to be undecidable by reducing them to the halting problem. However, Gregory **Chaitin** has given an undecidable problem in algorithmic information theory that doesn't depend on the halting problem.

While Turing's proof shows that there can be no general method or algorithm to determine whether algorithms halt, individual instances of that problem may very well be susceptible to attack. Given a specific algorithm, one can often show that it must halt, and in fact computer scientists often do just that as part of a correctness proof. But every such proof requires new arguments: there is no *mechanical, general way* to determine whether algorithms halt. And there's another caveat. The undecidability of the halting problem relies on the fact that computers are assumed to have a memory of potentially infinite size. If the memory and external storage of a machine is limited, as it is for any real computer, then the halting problem for programs running on that machine can be solved with a general algorithm (albeit an extremely inefficient one).

ham sandwich theorem

Given a sandwich in which bread, ham, and cheese (three finite volumes) are mixed up, in any way at all, there is always a flat slice of a knife (a plane) that bisects each of the ham, bread, and cheese. In other words, however messed up the sandwich–even if it's been in a blender–you can always slice through it in such a way that the two halves have exactly equal amounts, by volume, of the three ingredients. This theorem generalizes to higher-dimensional ham sandwiches, when it essentially becomes the **Borsuk-Ulam theorem**: in *n*-dimensional space in which there are *n* globs of positive volume, there is always a hyperplane that cuts all the globs exactly in half.

Hamilton, William Rowan (1805–1865)

An Irish mathematician who, among other things, invented **quaternions** and a new theory of dynamics. Having excelled in Greek and mathematical physics at Trinity College, Cambridge, Hamilton was appointed Astronomer Royal of Ireland; in this position he served from 1827 to his death and, during all that time, lived in Dunsink Observatory, Dunsink Lane, to the northwest of Dublin. However, he quickly lost interest in staying up at nights to make observations–he hired three of his sisters to help run the place–and preferred instead to write poetry (badly). He was friends with Samuel Coleridge, who introduced him to the philosophy of Kant, which had a great influence on him, and with William Wordsworth, who advised him against writing any more poems.

Hamilton did early work on **caustic** curves and was led from this to his discovery of the *law of least action,* which enabled many physical problems to be expressed more elegantly. One of his greatest triumphs was his treatment of **complex numbers** as pairs of real numbers, an approach that finally exorcised long-standing suspicions about the reality of **imaginary numbers**, and helped clear the way for other algebras. From this he was led to consider ordered quartets of numbers, which he called quaternions. The idea for quaternions came to Hamilton suddenly on October 16, 1843, while he was standing on Brougham ("Broom") Bridge, where Broombridge Street crosses the Royal Canal, Dublin. A commemorative plaque under the bridge, on the towpath, was unveiled by the Taoiseach (head of the Irish parliament), Eamon De Valera, on November 13, 1958. Of his invention, Hamilton wrote:

> The quaternion was born, as a curious offspring of a quaternion of parents, say of geometry, algebra, metaphysics, and poetry. . . . I have never been able to give a clearer statement of their nature and their aim than I have done in two lines of a sonnet addressed to Sir John Herschel:
>
> "And how the One of Time, of Space the Three
> Might in the Chain of Symbols girdled be."

Hamilton's interest in complex numbers was stimulated by his friend and compatriot John Graves, who pointed Hamilton in the direction of John Warren's *A Treatise on the Geometrical Representation of the Square Root of Negative Quantities.* This book explained the concept of the complex plane, which Hamilton turned from geometry into algebra. One of Hamilton's last inventions was a curiosity called the *icosian calculus,* which was another outcome of his friendship with Graves. After a visit to the latter's house, Hamilton wrote: "Conceive me shut up and revelling for a fortnight in John Graves' Paradise of Books! of which he has really an astonishingly extensive collection, especially in the curious and mathematical kinds. Such new works from the Continent he has picked up! and such rare old ones too!" Graves posed some puzzles to Hamilton, and either Graves or his books got Hamilton to thinking about regular polyhedra. When Hamilton returned to Dublin he thought about the symmetry group of the **icosahedron**, and used it to invent an algebra he called the "icosians" and also a game called the **Icosian game**. The only complete example of this game, inscribed to Graves, is now in the keeping of the Royal Irish Academy, of which Hamilton was the president from 1837 to 1847. (In early 1996, a second example of the Icosian game came to light but only included the board.)

In some ways, Hamilton was too far ahead of his time. The operator now referred to as the *Hamiltonian* and the so-called *Hamilton-Jacobi equation* that relates waves and particles only became important when **quantum mechanics** came along, and Felix **Klein** introduced Wernher Schrödinger, the father of wave mechanics, to Hamilton's work.

Hamilton's personal life was not always happy. He fell deeply in love with a woman named Catherine Disney, who was forced by her parents to marry a wealthy man 15 years older than her. Hamilton remained hopelessly in love with her the rest of his life, though he eventually married someone else. He became an alcoholic, then foreswore drink, then relapsed. Many years after their early romance, Catherine began a secret correspondence with Hamilton. Her husband became suspicious, and she attempted suicide by taking laudanum. Five years later, she became seriously ill. Hamilton visited her and gave her a copy of his *Lectures on Quaternions.* They kissed at last, and she died two weeks later. He carried her picture with him ever afterward and talked about her to anyone who would listen.[149]

Hamilton circuit

A **Hamilton path** that starts and ends at the same vertex. See also **traveling salesman problem**.

Hamilton path

Named after William **Hamilton**, a path that traverses every vertex of a **connected graph** once and only once. The problem of the **knight's tour** is equivalent to finding a Hamilton path (or, in the case of a reentrant tour, a **Hamilton circuit**) that corresponds to the legal moves of the knight. Compare with **Euler path**.

Hankel matrix

A **matrix** in which all the elements are the same along any **diagonal** that slopes from northeast to southwest.

happy number

If you iterate the process of summing the squares of the decimal digits of a number and if this process terminates in 1, then the original number is called a happy number. For example 7 → (7^2) 49 → ($4^2 + 9^2$) 97 → ($9^2 + 7^2$) 130 → ($1^2 + 3^2$) 10 → 1. See also **amicable number**.

Hardy, Godfrey Harold (1877–1947)

One of the most prominent English mathematicians of the twentieth century; his legendary collaboration with John Littlewood lasted 35 years and produced nearly 100 papers. Hardy was a precocious child, whose tricks included factorizing hymn numbers during sermons. In 1919, he became Savilian Professor of Geometry at Oxford but returned to Cambridge in 1931 as professor of pure mathematics. His work was mainly in **analysis** and **number theory**.

Hardy had only one other passion in his life–the game of cricket. His daily routine would begin with reading *The Times* and studying the cricket scores over breakfast. Then he would do mathematical research from 9 o'clock till 1 o'clock. After a light lunch, he would walk down to the university cricket ground to watch a game. In the late afternoon he would walk slowly back to his rooms at the college, and take dinner followed by a glass of wine. Hardy was known for his eccentricities. He couldn't stand having his photo taken and only five snapshots of him are known to exist. He also hated mirrors and his first action on entering any hotel room was to cover any mirror with a towel.

Hardy's book *A Mathematician's Apology* (1940)[151] is one of the most vivid descriptions of how a mathematician thinks and the pleasure of mathematics. But the book is more, as C. P. Snow writes:

> *A Mathematician's Apology* is . . . a book of haunting sadness. Yes, it is witty and sharp with intellectual high spirits: yes, the crystalline clarity and candor are still there: yes, it is the testament of a creative artist. But it is also, in an understated stoical fashion, a passionate lament for creative powers that used to be and that will never come again. I know nothing like it in the language: partly because most people with the literary gift to express such a lament don't come to feel it: it is very rare for a writer to realise, with the finality of truth, that he is absolutely finished.

See also **Ramanujan**.

harmonic analysis

The method of expressing **periodic** functions as sums of sines and cosines.

harmonic division

The division of a line segment by two points such that it is divided externally and internally in the same ratio.

harmonic mean

The harmonic mean of two numbers a and b is $2ab/(a+b)$.

harmonic sequence

The sequence: 1, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$ . . . . Added together, these become the terms of the *harmonic series:* $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \ldots$. This series diverges (has no finite sum), though very slowly–a result first proved by the French philosopher and theologian, Nichole d' Oresme (c. 1325–1382). In fact, it still diverges if you take away every other term, and even if you take away nine out of every ten terms. However, if you take the sum of reciprocals of all natural numbers that do not contain the number nine (when written in decimal expansion) the series converges! To show this, group the terms based on the number of digits in their denominator. There are eight terms in ($\frac{1}{1} + \ldots + \frac{1}{8}$), each of which is no larger than 1. Consider the next group ($\frac{1}{10} + \ldots + \frac{1}{88}$). The number of terms is *at most* the number of ways to choose two ordered digits out of the digits 0 . . . 8, and each such term is clearly no larger than $\frac{1}{10}$. So this group's sum is no larger than $9^2/10$. Similarly, the sum of the terms in ($\frac{1}{100} + \ldots + \frac{1}{999}$) is at most $9^3/10^2$, etc. So the entire sum is no larger than

$$9 \times 1 + 9 \times (\tfrac{9}{10}) + 9 \times (9^2/10^2) + \ldots + 9 \times (9^n/10^n) + \ldots$$

This is a geometric series that converges. Thus by the comparison test, the original sum (which is smaller term-by-term) must converge.

Harshad number

A number that is divisible by the sum of its own digits; also known as a Niven number. For example, 1,729 is a Harshad number because 1 + 7 + 2 + 9 = 19 and 1,729 = 19 × 91. A *Harshad amicable pair* is an amicable pair (m, n) such that both m and n are Harshad numbers (see **amicable numbers**). For example, 2,620 and 2,924 are a Harshad amicable pair because 2,620 is divisible by 2 + 6 + 2 + 0 = 10 and 2,924 is divisible by 2 + 9 + 2 + 4 = 17 (2,924/17 = 172). There are 192 Harshad amicable pairs in the first 5,000 amicable pairs.

hat problem

A team of three contestants, Alice, Bob, and Cedric, enter a room and a hat is placed on each one's head so that he or she can't see it. The color of each hat is based on a coin toss–blue (B) for heads, red (R) for tails. After all the contestants enter the room, they look at the colors of one anothers' hats and, based on this information, they guess the color of their own hat. Each can guess red or blue, or, if she can't make up her mind, she can pass. No communication is allowed during the competition, but the players are allowed to agree on a strategy before play begins. The team wins if at least one of them guesses correctly, and none of them guesses incorrectly. What is the team's best strategy? At first sight, it may seem as if no effective strategy is possible beyond each contestant guessing his or her own hat color. In fact, this is the very worst approach since, to succeed, it requires that everyone guess correctly and the probability of this is only $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. A far better plan is for the contestants to agree that two of them will pass while the third takes a stab at the color of his own hat. Then the odds improve to one in two. Beyond this it's hard to see any way that the probability of success could be increased. Yet there is an even better strategy. The key is to realize that there are only two cases (RRR and BBB) where everyone's hat is the same color but *six* cases where two hats are the same color and the other hat is a different color (RRB, RBR, BRR, BBR, BRB, and RBB). This suggests the following strategy for members of the team: *if you see two hats of opposite colors, pass. If you see two hats of one color, guess that your hat is the other color.* If everyone's hat is the same color, all players on the team will guess wrong and the team will lose. But the chance of this happening is only $\frac{2}{8}$ (= $\frac{1}{4}$). In every other possible case, the odd person out will guess correctly and their teammates will pass, so the team will win. This strategy wins $\frac{6}{8}$ (= $\frac{3}{4}$) of the time, and can't be improved upon. Since half of each player's guesses will be wrong, it's impossible to do better than a strategy in which each player in turn guesses correctly alone three times out of four, and the fourth time all guess wrong.

What if there are more players in the team? Say the number of players is n. By the reasoning described previously, it's clear that the team can't hope to win more than $n/(n + 1)$ of the time. Yet it isn't obvious that it can do this well. Having more people, it seems, might make it harder for them to synchronize their wrong guesses. However, it turns out that, if the number of people in the team is *one less than a power of 2,* this best-possible value can be achieved. For example, with a team of 7, the team can win $^7/_8$ of the time, and with a team of 15, they can win $^{15}/_{16}$ of the time. The strategy involved is complicated but is closely linked to *Hamming codes* (see **coding theory**), which are a method of encoding and transmitting information so that even if a small number of errors occur during transmission, the original information can be entirely recovered. For teams of other sizes, such as 9, 10, or 13, mathematicians have yet to find an optimal strategy or establish what proportion of the time the team can be expected to win.

Hausdorff, Felix (1868–1942)

A German mathematician who is considered to be one of the founders of modern **topology** and who also did significant work in **set theory** and functional analysis. Among several concepts named after him is the **Hausdorff dimension**, which gives a way of assigning a fractional dimension to a curve or shape. Hausdorff also published philosophical and literary works under the pseudonym "Paul Mongré." He studied at Leipzig and taught mathematics there until 1910, when he became professor of mathematics at Bonn. When the Nazis came to power, Hausdorff, a Jew, felt that as a respected university professor he would be safe from persecution. However, his abstract mathematics was denounced as useless and "un-German" and he lost his position in 1935. He sent his daughter to Britain but stayed with his wife in Germany. In 1942, when he could no longer avoid being sent to a concentration camp, he committed suicide together with his wife and sister-in-law.

Hausdorff dimension

A way to accurately measure the **dimension** of complicated **set**s such as **fractal**s. The Hausdorff dimension, named after Felix **Hausdorff**, coincides with the more familiar notion of dimension in the case of well-behaved sets. For example, a straight line or an ordinary curve, such as a circle, has a Hausdorff dimension of 1; any **countable set** has a Hausdorff dimension of 0; and an n-dimensional **Euclidean space** has a Hausdorff dimension of n. But a Hausdorff dimension is not always a natural number. Think about a line that twists in such a complicated way that it starts to fill up the plane. Its Hausdorff dimension increases beyond 1 and takes on values that get closer and closer to 2. The same idea of ascribing a fractional dimension applies to a plane that contorts more and more in the third dimension: its Hausdorff dimension gets closer and closer to 3. As a specific example, the fractal known as the **Sierpinski carpet** has a Hausdorff dimension of just over 1.89.

Heesch number

The maximum number of times that a closed plane figure–a tile–can be completely surrounded by copies of itself. The Heesch number of a triangle, quadrilateral, regular hexagon, or any other single shape that can completely tile the plane (see **tiling**), is infinity. *Heesch's problem* is to find the largest possible *finite* Heesch number, or, more generally, what values other than zero and infinity can Heesch numbers take. In considering this problem, it's helpful to define the Heesch number more precisely. In a tiling, the *first corona* of a tile is the set of all tiles that have a common boundary point with the tile, including the original tile itself. The *second corona* is the set of tiles that share a point with anything in the first corona; and so on. The Heesch number is the maximum value of coronas (k) that can surround a shape. For a long time the record holder for the largest finite value of k was a shape found by the American computer scientist

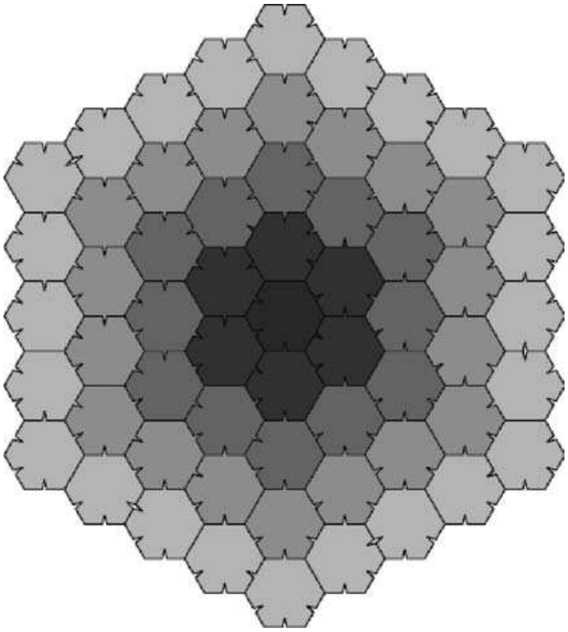

Heesch number A tiling in which copies of the same shape are used out to the fourth surrounding layer, or corona. *David Eppstein*

Robert Ammann, which consists of a regular hexagon with small bumps on two sides and matching notches on three sides. This was thought to have a Heesch number of three; however, in 2000, Alex Day argued that the Ammann hexagon actually has a Heesch number of four, though it isn't clear whether the difference has to do with a definition of tiling. In any event, it has since been shown by Casey Mann, of the University of Arkansas that there exists an infinite family of tiles (consisting of indented and outdented pentahex) with Heesch number five (or six by Day's reckoning)–the largest finite value currently known. Are there any polygons that have higher Heesch numbers? The answer is unknown but Mann thinks that more rounded **polyomino**s than the long skinny ones he's been using may have a better chance of giving unbounded Heesch numbers.

The Heesch number question is connected to two other famous unsolved tiling problems: the **domino problem** and the **Einstein problem**. Aperiodic tiling seems to act as a barrier to the existence of tiling algorithms, so it isn't expected that both of these problems have the same answer. On the other hand, if there's a maximum finite Heesch number *k*, then it seems that this could be used as the basis of an algorithm to test whether a shape tiles: simply attempt to fill out a tiling to the ($k + 1$)st corona; if successful, the shape must tile the plane, and if not, the shape will not tile. Similar questions can be asked about Heesch numbers for tilings in higher dimensions.

Hein, Piet (1905-1996)

An extraordinarily creative Danish mathematician, scientist, inventor, and poet who often wrote under the Old Norse pseudonym Kumbel, meaning "tombstone." A direct descendant of the Dutch naval hero of the sixteenth century who had the same name, Piet Hein was born in Copenhagen and studied at the Institute for Theoretical Physics of the University of Copenhagen (later the Niels Bohr Institute), the Technical University of Denmark, and the Royal Swedish Academy of Fine Art. He was later awarded an honorary doctorate by Yale University. A good friend of Albert Einstein, he is famed for his many mathematical games, including **Hex**, **Tangloids**, Polytaire, **TacTix**, and the **Soma cube**. These games were featured in numerous columns of Martin **Gardner**'s "Mathematical Recreations" column in *Scientific American* and often achieved worldwide attention in this way. As an artist and constructor, Hein gave form, in the 1950s and 1960s, to elegant pieces of furniture that helped "Scandinavian design" attract international recognition. These pieces, including a dining-room table created in cooperation with the Swedish designer Bruno Mathsson, were based on the **superellipse** curve–a shape that Hein also brought to bear in applications as varied as city planning (it's the basis for Sergel's Square in the center of Stockholm) and toy making (see **superegg**). Hein was a prolific and excellent writer of light verse, producing thousands of short, aphoristic poems known as Grooks. For him there was no unbridgeable gap between the subjectivity of fine art and the objective world of science. "Art," he said, "is a solution to problems which cannot be formulated clearly before they have been solved." His philosophy of life was summed up by his aphorism "co-existence or no existence."

helicoid

The second oldest known **minimal surface**; it was discovered by Jean-Baptiste Meusnier in 1776, thirty years after the **catenoid**. It is the only minimal surface, apart from the simple plane, that is also a **ruled surface**. The helicoid is the surface swept out by a line that always intersects a fixed axis at right angles and that rotates uniformly as its point of intersection moves uniformly along the axis. This line intersects any cylinder concentric with the axis in a **helix**. The helicoid has a wide variety of shapes and is a familiar sight in everyday life, taking the form of many things from spiraling parking ramps to screw threads.

helix

A curve in three dimensions, the tangent to which makes a constant angle with a fixed line. A *circular helix* is formed by winding a line around a **cylinder** so the radius is always the same. A *conical helix* is formed by winding a line around a **cone**, so that, consequently, its radius constantly changes. Springs often take the form of various kinds of helices. In nature, the DNA molecule is in the shape of a double helix.

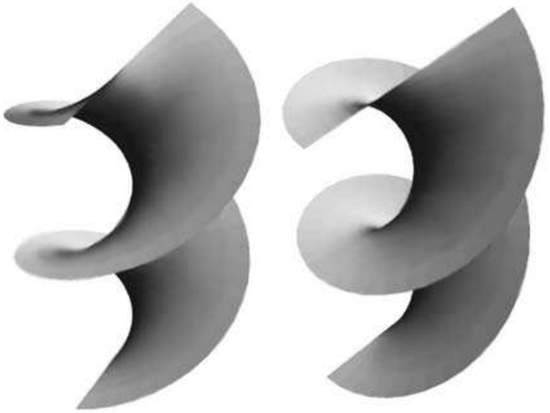

helicoid *Richard Palais*

Henon, Michele

An astronomer at the Nice Observatory in southern France. For a number of years, particularly during the 1960s, he studied the dynamics of stars moving within galaxies, using computers as a way to understand the stability of their motions. His work was very much in the spirit of Henri **Poincaré**'s approach to the classical **three-body problem**: What important geometric structures govern their behavior? The main property of these systems is that the energy of their motion is constant to a very good approximation. Consequently, their chaotic dynamics are not described by **simple attractor**s, but by objects that are markedly more difficult to analyze and visualize, existing on energy "surfaces" in three and higher dimensions. During the 1970s, Henon discovered a very simple iterated mapping that showed a **chaotic attractor**, now called *Henon's attractor,* that allowed him to make a direct connection between deterministic data and **fractal**s. The Henon attractor is self-similar (see **self-similarity**): if you zoom in on the attractor in its state space you find more and more layers, much like filo dough or a croissant.

Henstock integration

See **integration**.

heptagon

A **polygon** with 7 sides.

Hermann grid illusion

An illusion first described by the German physiologist Ludimar Hermann (1838–1914) in 1870. While reading a book on sound by the Irish physicist John Tyndall, Hermann saw gray spots in the intersections of spaces among the figures that Tyndall had arranged in a matrix. Despite the fact that the same intensity of light is reflected all the way along the white spaces in the Hermann Grid, the intersections appear gray. To explain this, consider two regions of the retina. One region views an intersection of a white horizontal and vertical band, while the other views a white band between two intersections (the region going away from the intersection). Although the two regions themselves receive the same amount of light, the situation in their neighboring regions is different. At the intersection, light comes in from all four sides, but the white band that lies between the two intersections is surrounded by two dark sides. This leads to an effect called *lateral inhibition,* which causes a bright surround to an area appear darker and, conversely, a dark surround to an area appear lighter. A similar but more powerful illusion, known as the *Lingelbach illusion* or the *Scintillating grid illusion,* was discovered in 1994 by Elke Lingelbach, the wife of a German mathematics professor, and has not

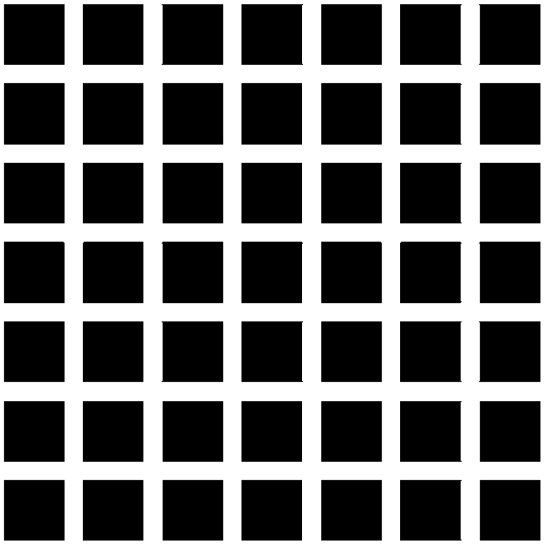

Hermann grid illusion The original illusion, in which the viewer sees gray spots at the intersections.

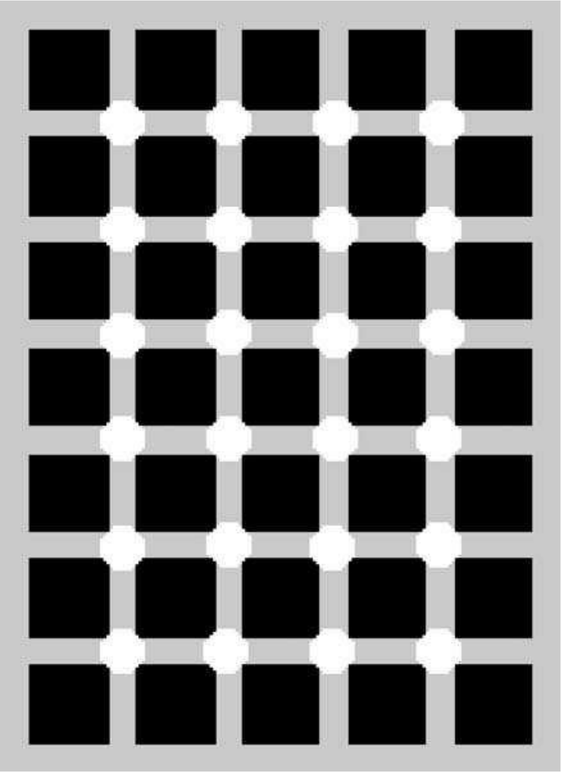

Hermann grid illusion The more striking and recently discovered "scintillating" version of the illusion.

yet been fully explained.[286] Curiously, the effect of the scintillation is lessened by tilting the head through 45°!

Hermite, Charles (1822–1901)

A French mathematician whose work in the theory of functions includes the application of **elliptic function**s to provide the first solution to the general equation of the fifth degree, the **quintic** equation. He also showed that ***e*** is a **transcendental number**, studied a class of **differential equation**s now known as *Hermite polynomials,* which later proved to be of importance in some applications of quantum mechanics, and discovered the properties of *Hermitian matrices.*

Heron of Alexandria (c. A.D. 60)

Also called *Hero,* a Greek geometer and inventor whose writings have helped preserve a knowledge of the mathematics and engineering of Babylonia, ancient Egypt, and the Greco-Roman world. His most important geometric work, *Metrica,* was lost until a fragment was discovered in 1894, followed by a complete copy in 1896. It is a compendium, in three books, of geometric rules and formulas, the best known of which is Proposition 1.8, now known as **Heron's formula**. He invented many devices operated by water, steam, or compressed air, including a fountain, a fire engine, siphons, and an engine in which the recoil of steam revolves a ball or a wheel.

Heron's formula

An important formula in plane geometry that allows the area of any triangle to be calculated without knowing the altitude (perpendicular height) of any of its sides. Let a, b, and c be the side lengths of a triangle and A its area. Heron's formula states that

$$A^2 = s(s - a)(s - b)(s - c),$$

where $s = (a + b + c)/2$. The origin of this formula is historically obscure. A medieval Arab source, for example, ascribes it to **Archimedes**. However, the first definite reference we have to it is by **Heron of Alexandria**. His proof is extremely convoluted, and it seems clear that it must have been determined by an entirely different thought process, and then dressed up in the usual synthetic form that the classical Greeks preferred for their presentations. Heron's formula contains **Pythagoras's theorem** as a degenerate case. A *Heronian triangle* is one with integer sides and integer area.

Herring illusion

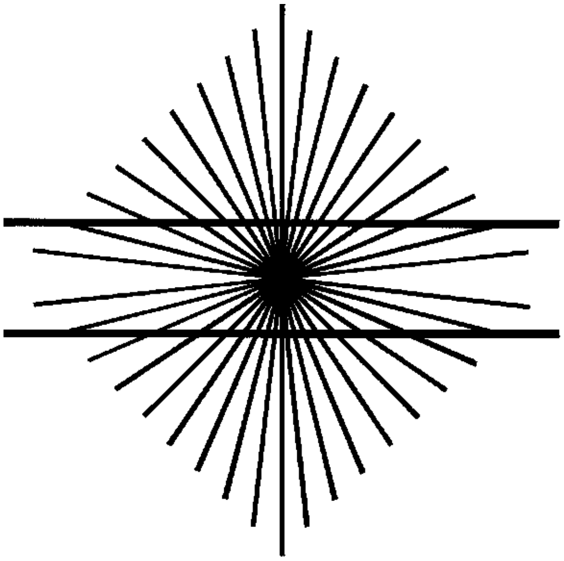

Herring illusion

A **distortion illusion** first published in 1861 by the German psychologist Ewald Herring (1834–1918), and now named after him. As in the case of the **Zöllner illusion** and others, it shows how geometrical relationships can seem to be distorted by their background (a lined background can make circles, squares, and triangles seem distorted, too). The straight horizontal lines in the illusion appear to bow out in the center. This can be explained if the brain interprets the radiating lines in terms of depth, making the central spot in the Herring diagram, and thus also the heavy black lines near the center of the diagram, appear to be farther away than the edges. Because the heavy black lines are the same thickness at the center as at the edges but are assumed to be farther away, the brain thinks they must be more widely spaced at the center.

heuristic argument

An educated guess: something that helps in finding the solution to a problem but is otherwise unjustified or incapable of justification.

Hex

A board game played by two players on a hexagonal grid, usually in the shape of an 11 × 11 rhombus. It was invented by Piet **Hein** in 1942 and independently by John **Nash** in 1948. Hein said that the game occurred to him while contemplating the **four-color problem** and it soon became popular in Denmark under the name *Polygon.* Nash's version was played by math students at Princeton and a number of other American campuses. Players use differently colored pieces–say, red and blue. They take alternate turns placing a piece of their color inside a hexagon, filling in that hexagon with their color. Red's goal is

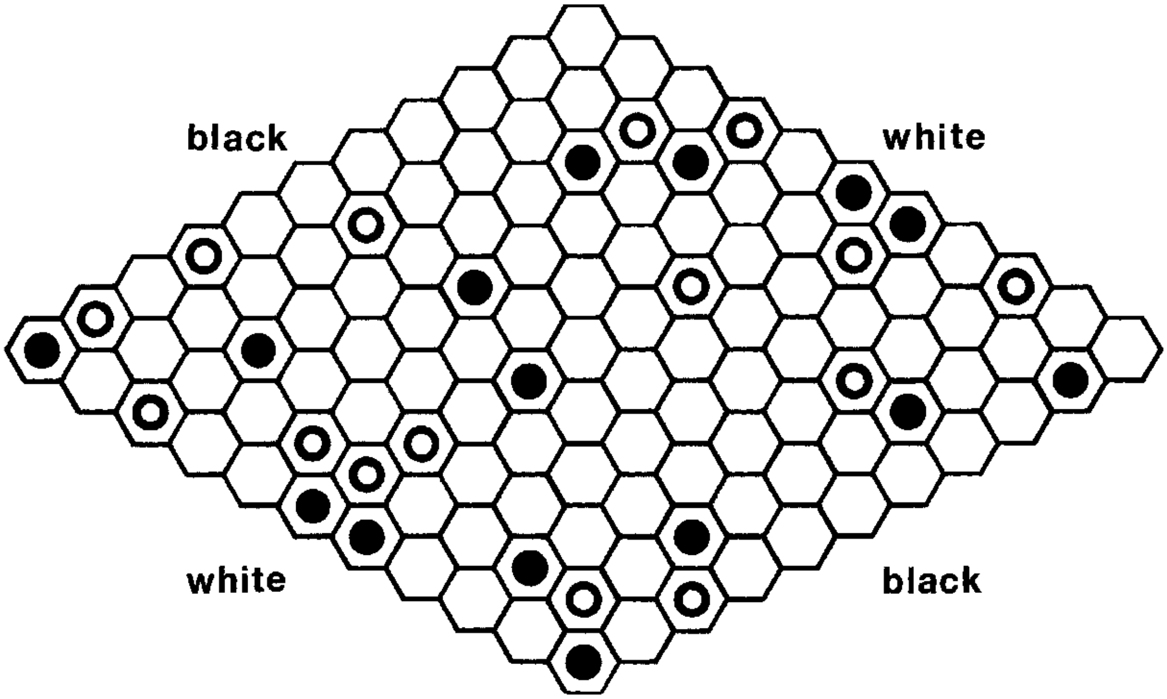

Hex A position in a game of Hex.

to form a red path connecting the top and bottom sides of the parallelogram; Blue's goal is to form a path connecting the left and right sides. The game can never end in a tie, a fact found by Nash. The only way to prevent your opponent from forming a connecting path is to form a path yourself. When the sides of the grid are equal, the game favors the first player and the first player has a winning strategy. There are two ways to make the game fairer. One is to make the second player's sides closer together, playing on a parallelogram rather than a rhombus; however, this has been proven to result in a win for the second player, so it theoretically doesn't improve matters. A better way is to allow the second player to choose his color after the first player makes the first move or to make the first three moves, which encourages the first player to intentionally even out the game.

hexa-

The Greek prefix meaning "six." A *hexagon* is a **six**-sided **polygon**. A *hexahedron* is a six-sided **polyhedron**, otherwise known as a **cube** if it is regular. *Hexadecimal* is the number system with **base** 16 (i.e., six more than the decimal system) and is used mostly in computing (because four binary digits can represent 16 different numbers). *Hexagonal numbers* are **figurate number**s (numbers that can be represented by a regular geometric arrangement of equally spaced points) of the form $n(2n - 1)$; the first few are 1, 6, 15, 28, 45, ... For *hexaflexagon* see **flexagon**. For *hexomino* see **polyomino**. See also **Giant's Causeway**.

higher dimensions

Dimensions beyond the familiar three spatial dimensions (up-down, left-right, back-forth) of which we are aware in every-day life. Intense speculation, both scientific and fictional, has naturally been directed toward the possibility of a **fourth dimension**. One way to think of points in four-dimensional space is as ordered sets of four numbers. Clearly, this algebraic representation can be extended to many arbitrary dimensions: n-dimensional space is defined as the set of the set of points $(a_1, a_2, \ldots, a_n)$ where a_1 to a_n can take any **real number** value. There has been much conjecture that the universe in which we live contains many more than three spatial dimensions. This speculation began with the **Kaluza-Klein theory** but is now firmly embedded in modern **string theory**.[25, 55, 81]

Hilbert, David (1862–1943)

> *One can measure the importance of a scientific work by the number of earlier publications rendered superfluous by it.*

A great German mathematician who was one of the colossi in the field in the twentieth century. His most important discovery was of what is now called **Hilbert space**. He was also a master of mathematical organization. During the early phase of his career, Hilbert reorganized **number theory**, crystallizing his conclusions in the classic book *Der Zahlbericht* (The theory of algebraic number fields, 1897). He then moved into geometry and performed a similar service by setting forth the first rigorous set of geometrical axioms in his *Grundlagen der Geometrie* (Foundations of geometry, 1899). He invented a simple **space-filling curve** now known as the *Hilbert curve* and also proved **Waring's conjecture**. At the Paris International Congress of 1900, Hilbert proposed 23 outstanding problems in mathematics to whose solutions he believed twentieth-century mathematicians should devote themselves. These problems have come to be known as *Hilbert's problems,* and a number still remain unsolved today. Hilbert's mathematical philosophy is partly re-vealed by a couple of remarks, one of which he made after learning that a student in his class had dropped the subject in order to become a poet. "Good," he said. "He did not have enough imagination to become a mathematician." Whether he really believed the second is open to question: "Mathematics is a game played according to certain simple rules with meaningless marks on paper."

Hilbert space

A space of infinite dimensions, named after David **Hilbert**, in which distance is preserved by making the sum of squares of coordinates a convergent sequence; it is of crucial importance in the mathematical formulation of **quantum mechanics**. See also **Fredholm, Erik Ivar**.

Hinton, Charles Howard (1853–1907)

An English-born mathematician best known for his writings and inventions aimed at helping to visualize the **fourth dimension**; he may also have coined the name ***tesseract*** for the four-dimensional analogue of a cube. Hinton matriculated at Oxford and continued to study there, earning a B.A. (1877) and an M.A. (1886), while he also taught, first at Cheltenham Ladies' School and then, from 1880 to 1886, at Uppingham School. At this time, another teacher at Uppingham was Howard Candler, who was a friend of Edwin **Abbott** and thus provides a possible link between these two explorers of other dimensions. In the early 1880s, Hinton published a series of pamphlets starting with "What Is the Fourth Dimension?" and "A Plane World" (a contemporary of Abbott's ***Flatland: A Romance of Many Dimensions***), which were reprinted in the two-volume *Scientific Romances* (1884). Hinton's descriptions owed much to the mathematical models of William **Clifford**, whose theories about four-dimensional spaces were then in vogue. But Hinton went much further in his attempts to break free of three-dimensional thought. He devised an elaborate set of small colored cubes to represent the various cross sections of a tesseract and then memorized the cubes and their many possible orientations in order to gain a window on the fourth dimension.

At the time he was teaching in England, Hinton married Mary Everest Boole, the eldest daughter of George **Boole**, the founder of mathematical logic. Regrettably, he also married a Maud Wheldon and was tried at the Old Bailey in London for bigamy. After serving a day in prison for the offence, he fled with his (first) family to Japan, where he taught for some years, before taking up a post at Princeton University. There, in 1897, he designed a species of baseball gun which, with the help of gunpowder charges, would shoot out balls at speeds of 40 to 70 miles per hour. It was used by the Princeton team for several seasons before being abandoned by the players in fear of their lives.

After a brief spell at the University of Minnesota, Hinton joined the Naval Observatory in Washington, D.C. At the same time, he more rigorously developed his ideas on the fourth dimension and presented his results before the Washington Philosophical Society in 1902. Hinton asked: What would prove the existence of a real fourth spatial dimension? He offered three possibilities, two of which involved a specific molecular structure and a particular case of electrical induction, and have since been explained by science in more mundane ways. However, Hinton's other case, pertaining to right- and left-handedness remains open because there are instances of right- and left-handedness in nature, such as the spin of elementary particles, to which his example could be applied. In any event, Hinton's final assessment that we can only regard a four-dimensional space as possible if three-dimensional mechanics fails to explain known physical phenomena still rings true today.[166, 167, 272] See also **Boole (Stott), Alicia**.

Hippias of Elis (c. 5th century B.C.)

An itinerant Greek philosopher who contributed significantly to mathematics by discovering the quadratrix, a special curve he may have used to trisect an angle (see **quadratrix of Hippias**). Hippias is one of the first mathematicians about whom a good deal is known. He came from a state in the northwest corner of Peloponnesia that was the home of the Olympic games. According to Plato, Hippias boasted, during one of his visits to the Olympics, that everything he wore–his clothing, sandals, ring, and oil flask–he'd made himself. Later, in Athens, Hippias became one of the first to teach for money, a practice forbidden by the Pythagoreans and scorned by Plato. He and other paid teachers became known as "sophists," which was a derogatory term at the time but has since come to mean "wise man."

hippopede

A **quartic** curve described by the equation

$$(x^2 + y^2)^2 + 4b(b - a)(x^2 + y^2) - 4b^2x^2 = 0,$$

where a and b are positive constants. Hippopede means "foot of a horse." It is often known as the *hippopede of Proclus,* after **Proclus** who was the first to study it (together with **Eudoxus**, who used it in his theory of how the planets move), and also the *horse fetter* and the *curve of Booth* because of work done on it by J. Booth (1810–1878). Any hippopede is the intersection of a **torus** (donut) with one of its **tangent** planes–that is, a plane parallel to its axis of rotational symmetry. The curve takes any of a variety of forms depending on where the donut is sliced. It may be a simple oval; an indented oval or *elliptical lemniscate of Booth* ($0 < b < a$); two isolated circles; a figure-eight curve or **lemniscate of Bernoulli** (the only hippopede that is also a **Cassinian oval**); or a *hyperbolic lemniscate of Booth* ($0 < a < b$).

Hi-Q

See **peg solitaire**.

Hnefa-Tafl

The Viking equivalent of chess and a particular form of a **Tafl game**. It effectively models the kind of internal conflict familiar to the Vikings and recounted in *Njal's Saga,* the best known of the Icelandic sagas. The king or chieftain sits in his hall, surrounded by his thanes. His enemies gather in secret, and in numbers sufficient to overwhelm the king's standing forces in a lightning raid. They gather around the king's hall and set it alight, forcing the defenders to fight in the open or burn in the hall. If the king can, by some desperate stratagem, break free and escape the trap, he can rally his people and strike back at his enemies. If not, he dies. Hnefa-Tafl reflects this mode of contest. The board has 19 × 19 grid lines (though sometimes has as few as 7 × 7 grid lines), and the pieces are placed at the points of the intersections (curiously analogous to the oriental game of **Go**, which also uses 19 × 19 lines and where play occurs at the intersections). The opposing forces are unequal in size, and have different objectives: the attackers attempt to trap the king in his hall, while the defenders try to open an escape route for him.

Hoffmann, Louis "Professor" (1839–1919)

The pseudonym of Angelo John Lewis, an English barrister who was the leading writer on magic, cards, and "parlor amusements" at the turn of the twentieth century. His *Puzzles Old and New*[171] (1893) is a major source of information about mathematical recreations.

Hofstadter, Douglas R. (1945–)

A physicist and philosopher best known for his 1980 Pulitzer Prize–winning book *Gödel, Escher, Bach: An Eternal Golden Braid.*[172] He is currently a professor of cognitive science and computer science at Indiana University, Bloomington, and has particular interests in themes of the mind, consciousness, self-reference, translation, and mathematical games. He is the son of the Nobel Prize–winning physicist Robert Hofstadter.

Hofstadter's law

It always takes longer than you think, even when you take Hofstadter's law into account.

Hogben, Lancelot Thomas (1895–1975)

An English zoologist and geneticist famed for his best-selling *Mathematics for the Million* (1933)[173] of which Albert Einstein said, "It makes alive the contents of the elements of mathematics" and H. G. Wells said, "A great book, a book of first class importance." Hogben was born in Southsea, Hampshire, and studied at Cambridge and London. Imprisoned in 1916 as a conscientious objector during World War I, he was released only when his health went into serious decline. He held various academic posts in Britain, Canada, and South Africa, becoming professor of social biology at London University in 1930. During World War II he was put in charge of the medical statistics records for the British Army. After the war he became professor of medical statistics at the University of Birmingham, where he remained until his retirement in 1961. Hogben first began to apply mathematical principles to the study of genetics in the 1930s, focusing on the study of generations of the fruit fly *Drosophila* and how it related to research on heredity in humans. In addition to *Mathematics for the Million,* he authored half a dozen other books, including the popular *Science for the Citizen.* Though trained as a scientist, Hogben was passionately interested in linguistics. In *The Loom of Language,* which he edited, he set out the principles of his own invented language, "Interglossa," based on Greek and Latin roots but with a syntax resembling that of Chinese.

hole

A topological structure (see **topology**) that prevents any object in which it occurs from being continuously shrunk to a point. A **sphere** has no holes; a **torus** and a teacup each have one hole. See also **genus**.

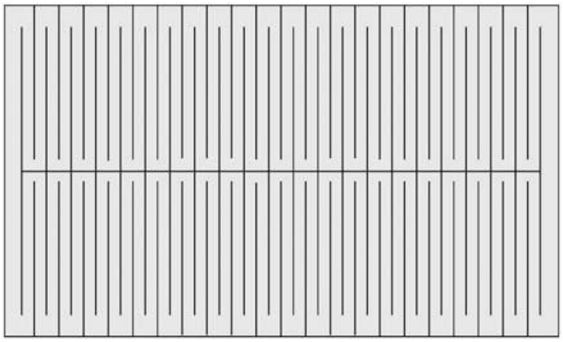

hole-in-a-postcard problem Photocopy the diagram, cut along the lines, and step through the postcard!

hole-in-a-postcard problem

With a pair of scissors, make cuts in a regular-sized postcard to create a hole large enough for a person to step through it. The solution is shown in the diagram. The number of lines determines the size of the resulting aperture. With enough cuts you could literally drive a horse and cart through the card!

hole-through-a-sphere problem

A number of **geometry puzzles** hinge on a surprising fact about spheres that have had holes bored through them. Imagine you have a round bead that is 1 inch (about 2.5 cm) in diameter and that you drill a hole exactly through the middle of this so that the remaining part of the sphere is only half an inch thick. Now imagine that an enormously large drill has been used to bore a hole though Earth so large that the part of Earth that is left behind is only half an inch thick. Amazingly, the residual volumes of these two holey spheres, the drilled bead and the drilled Earth, are exactly the same! It just happens that even though Earth is vastly larger than the bead, the drill has to take out proportionately more in order to make the thickness of the hole the same, so that the volume left doesn't depend separately on the initial size of the sphere or of the hole, but only on their *relation,* which is forced by requiring the hole to be half an inch long. This fact enables the following poem-problem to have a solution even though its seems as if not enough information has been provided:

Old Boniface he took his cheer,
Then he bored a hole through a solid sphere,
Clear through the center, straight and strong,
And the hole was just six inches long.
Now tell me, when the end was gained,
What volume in the sphere remained?
Sounds like I haven't told enough,
But I have, and the answer isn't tough.

Having already learned the secret that the volume that remains of a drilled sphere doesn't depend on the initial size of the sphere, we can cheat and give a kind of meta-argument that is much shorter than the geometric proof. The volume left behind of any sphere with a 6-inch-long hole through it must be the same as the volume left behind of a 6-inch-diameter sphere with a hole of 0 diameter drilled through it. This is equal to $\frac{4}{3}(\pi 6^3)$, or approximately 905 cubic inches.

hole-through-the-earth problem

I wonder if I shall fall right through *the earth!*
—Alice in *Alice in Wonderland*

Imagine there is a hole going from one point on Earth's surface, all the way through the center of Earth, to the exact opposite (antipodal) point on the other side. What would happen if you dropped something into this hole? The Greek historian Plutarch considered the problem some 2,000 years ago. In 1624, van Etten argued that a millstone dropped down such a hole at 1 mile per minute would take more than 2½ days to reach the center, where "it would hang in the air." The first correct answer was given by Galileo is his *Dialogue Concerning the Two Chief World Systems* (1632). Galileo realized a dropped object would accelerate until it reached the center of Earth, travel through to the other side, then oscillate back and forth.

holism

The idea that the whole is greater than the sum of the parts. Holism is credible on the basis of **emergence** alone, since **reductionism** and bottom-up descriptions of nature often fail to predict complex higher-level patterns.

homeomorphic

In **topology**, two objects are said to be homeomorphic if they can be smoothly deformed into each other. See also **homeomorphism**.

homeomorphism

A **one-to-one** continuous transformation that preserves **open** and **closed** sets.

homology

A way of attaching **Abelian group**s, or more elaborate algebraic objects, to a **topological space** so as to obtain algebraic invariants. In a sense, it detects the presence of "holes" of various dimensions in the space. The methods developed to handle this led to what is now called *homological algebra,* a subject in which homological invariants are calculated for many purely algebraic structures.

homomorphism

A **function** that preserves the operators associated with the specified structure.

homotopy

A continuous transformation from one path in a **topological space** to another, or more generally, of one **function** to another (see **continuity**). Paths connected by a homotopy are called *homotopic* and are said to be in the same *homotopy class.* Properties left unchanged by such homotopies are known as *homotopy invariants.* Homotopy classes of paths can be composed to form the **fundamental group**, or *first homotopy group.* Other maps can be used to form *higher homotopy groups.*

Hordern, L. Edward (d. 2000)

An English puzzlist and leading authority and writer on **sliding-piece puzzles**.[175]

hundred

The smallest three-digit number in the decimal system and the smallest square of a two-digit number (10). A hundred today means 100 but, over the years and in different places, it has stood for different values including 112, 120, 124, and 132. The remnants of these old measures still persist in the *hundredweight* of some countries representing 112 or 120 pounds. A hundred is also a measure of land area, frequently used in colonial America and in England to signify a division of a county or a shire having its own court. A strange custom is invoked if a member of the English Parliament's lower house, the House of Commons, wishes to resign his seat (something that is technically illegal). That member accepts stewardship of the "Chiltern Hundreds," an area of chalk hills near Oxford and Buckingham, which effects his release from Parliament. The "Hundred Years' War" between the English and the French actually lasted 116 years.

hundred fowls problem

A Chinese puzzle found in the sixth-century work of the mathematician Zhang Qiujian. Similar problems involving two constraints and three unknowns are found in early European and Arabic mathematics from about the eighth century A.D. on.

PUZZLE

If a rooster is worth five coins, a hen three coins, and three chicks together are worth one coin, how many roosters, hens, and chicks totaling 100 can be bought for 100 coins? It turns out that there are three different solutions. These can be found the long way, by trial and error, or by using algebra. Call the number of roosters R, the number of hens H, and the number of chicks C. The problem gives two constraints. First, the total number of fowl must be 100, so $R + H + C = 100$. Second, the total cost of the fowl must be 100. The cost of roosters is $5R$, the cost of hens is $3H$, and the cost of chicks is $(1/3)C$, so $5R + 3H + (1/3)C = 100$. These two equations can be used to get rid of one of the unknowns; then it's a question of guess and check.

Solutions begin on page 369.

Hunter, James Alston Hope

An American mathematician and puzzlist who has written numerous articles and several books on recreational mathematics (two in partnership with Joseph **Madachy**), and was the author of a syndicated puzzle column read throughout the United States and Canada. In 1955, he coined the name "**alphametic**" and is probably the most prolific producer of **cryptarithm**s.

Huygens, Christiaan (1629–1695)

A Dutch scientist and mathematician who solved the **tautochrone problem**, proposed a new wave theory of light, designed a new pendulum clock, discovered Saturn's largest moon (Titan), and sketched the first feature on the surface of another planet (Syrtis Major on Mars). In his final years Huygens composed one of the earliest discussions of extraterrestrial life, published after his death as the *Cosmotheoros* (1698).

Hypatia of Alexandria (c. A.D. 370–417)

The first woman known to have made a significant contribution to mathematics. Although there is no evidence that Hypatia did any original research, she assisted her father, Theon of Alexandria, in writing his 11-part commentary on Ptolemy's great work on astronomy and mathematics, the *Almagest.* It's thought that she also helped in producing a new version of **Euclid**'s *Elements,* which formed the basis for all later editions of Euclid. Hypatia became head of the Platonist school at Alexandria in about A.D. 400 and, as a pagan, represented a menace to some Christian sects who felt threatened by her learning and depth of scientific knowledge. In the end, although the exact circumstances are unclear, she was murdered by a mob. The event served as a trigger for the departure of many scholars and the beginning of the decline of Alexandria as a major academic center.

hyperbola

One of the **conic section** family of curves, which also includes the **circle**, the **ellipse**, and the **parabola**; it is obtained if a double cone is cut by a plane inclined to the axis of the cone such that it meets both branches of the cone. Of the four conic curves, the hyperbola is the one least encountered in everyday life. A rare chance to see the complete shape is when a lamp with a cylindrical or conical shade throws shadows on a nearby wall. Part of a hyperbola is produced by the liquid that climbs by capillary action between two microscope slides held vertically and almost touching.

A hyperbola is the path followed by a smaller object that is traveling fast enough to escape completely from the gravitational pull of a larger object. Some comets, for example, have hyperbolic or "open" orbits so that, after one swing around the Sun, they head off into interstellar space never to return. It can be difficult to tell, in some cases, whether a comet's orbit is hyperbolic or is highly elliptical and, therefore, closed. In fact, one way to think of a hyperbola is as a kind of ellipse that is split in half by infinity. Not surprisingly, the hyperbola and the ellipse share many inverse relationships. For example, whereas the eclipse is the **locus** of all points whose distances from two fixed points, called *foci,* have a constant sum, the hyperbola is the locus of all points whose distances, r_1 and r_2, from two fixed points, F_1 and F_2, is a constant difference, $r_2 - r_1 = k$. If a is the distance from the origin to either of the x intercepts of the hyperbola, then $k = 2a$. Also, let the distance between the foci, $F_2 - F_1 = 2c$. Then the *eccentricity,* a measure of the flatness of the hyperbola, is given by $e = c/a$. For all hyperbolas, $e > 1$; the larger the value of e, the more the hyperbola resembles two parallel lines. Just as the circle (for which $e = 0$) is the limiting case of the ellipse (for which $0 < e < 1$), so the parabola ($e = 1$) is the limiting case of both the ellipse and the hyperbola.

A hyperbola has two **asymptote**s: the never-quite-attainable limits of the curve's branches as they run away to infinity. The *transverse axis* of the hyperbola is the line on which both foci lie and that also intersects both vertices (turning points); the *conjugate axis* goes through the center and is perpendicular to the transverse axis.

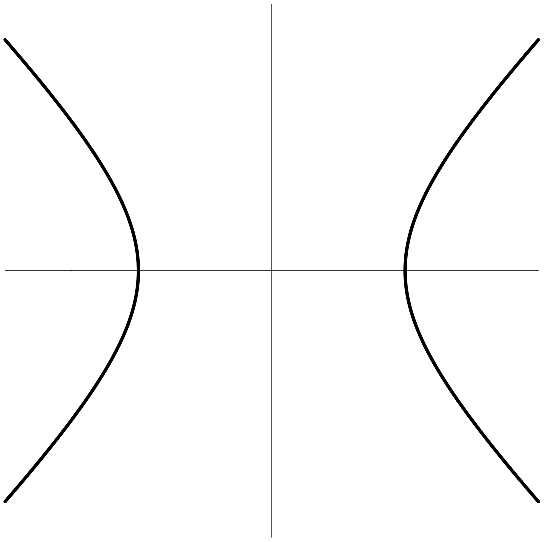

hyperbola © Jan Wassenaar, www.2dcurves.com

A *rectangular hyperbola* has an eccentricity of $\sqrt{2}$, asymptotes that are mutually perpendicular, and the property that when stretched along one or both of its asymptotes remains unchanged. This special case of the hyperbola was first studied by **Menaechmus**. **Euclid** and Aristaeus wrote about the general hyperbola but only studied one branch of it, while **Apollonius** was the first to study the two branches of the curve of the hyperbola and is generally thought to have given it its present name.

The **pedal curve** of a hyperbola with one focus as the pedal point is a circle. The pedal of a rectangular hyperbola with its center as pedal point is a **lemniscate of Bernoulli**. The **evolute** of a hyperbola is a **Lamé curve**. If the center of a rectangular hyperbola is taken as the center of inversion, the rectangular hyperbola inverts to a lemniscate. If the vertex is used as the center of inversion, the rectangular hyperbola inverts to a right **strophoid**. If the focus of a hyperbola is taken as the center of inversion, the hyperbola inverts to a limaçon (see **limaçon of Pascal**). In this last case if the asymptotes of the hyperbola make an angle of $\pi/3$ with the axis that cuts the hyperbola, then it inverts to the Maclaurin **trisectrix**. See also **hyperboloid**.

hyperbolic geometry

One of the two main types of **non-Euclidean geometry** and the first to be discovered. It is concerned with saddle-surfaces, which have negative **curvature** and on which the **geodesic**s are **hyperbolas**. In hyperbolic geometry, contrary to the **parallel postulate**, there exists a line m and a point p not on m such that at least two distinct lines parallel to m pass through p. As a result, the sum of the angles of a triangle is less than 180° and, for a right triangle, the square of the hypotenuse is greater than the sum of the squares of the other two sides. See also **elliptical geometry**.

hyperbolic spiral

The curve whose equation in **polar coordinates** is $r\theta = a$.

hyperboloid

A **quadratic** surface of which there are two basic forms: a *hyperboloid of one sheet,* generated by spinning a **hyperbola** around its conjugate axis, and a *hyperboloid of two sheets* produced by rotating a hyperbola about its transverse axis. The hyperboloid of one sheet, first described by **Archimedes**, has some particularly remarkable properties. In 1669 Christopher Wren, the architect who de-

hyperboloid The McDonnell Planetarium in St. Louis is an example of a hyperboloid. *Courtesy of the St. Louis Science Center*

signed St. Paul's Cathedral in London, showed that this kind if hyperboloid is what mathematicians now call a **ruled surface**–a surface composed of infinitely many straight lines. This fact enables a close approximation to a hyperboloid to be made in the form of a string model. Two circular disks, of the same size, are held parallel, one exactly above the other, by a framework. Strings are then run through holes near the circumference of one circle to corresponding holes in the other circle that are a fixed distance farther around the circumference. Each string is perfectly straight but the surface that emerges takes the curved form of a hyperboloid. For the same reason, a cube spun rapidly on one of its corners will appear to describe a hyperbolic curve when viewed side-on.

Prominent examples of hyperboloids can be seen in the form of cooling towers at power stations and, most strikingly, in the shape of the McDonnell Planetarium in St. Louis, Missouri. The designer of this building, Gyo Obata, chose the design because the hyperbolic paths of some comets suggest "the drama and excitement of space exploration."

hypercube

A higher dimensional analog of a cube. A four-dimensional hypercube is known as a **tesseract**.

hyperellipse

See **superellipse**.

hyperfactorial

A number such as 108, which is equal to $3^3 \times 2^2 \times 1^1$. In general, the *n*th hyperfactorial $H(n)$ is given by

$$H(n) = n^n\ (n-1)^{n-1} \ldots 3^3\ 2^2\ 1^1.$$

The first eight hyperfactorials are 1, 4, 108, 27,648, 86,400,000, 4,031,078,400,000, 3,319,766,398,771,200,000, and 55,696,437,941,726,556,979,200,000. See also **large numbers** and **superfactorials**.

hypergeometric function

The sum of the *hypergeometric series:*

$$F(a; b; c; x) = 1 + \frac{ab}{1.c}\,x + \frac{a(a+1)b(b+1)}{1.2.c(c+1)}\,x^2 + \ldots$$

Many common functions can be written as hypergeometric functions.

hyperreal number

Any of a colossal set of numbers, also known as *nonstandard reals,* that includes not only all the **real number**s but also certain classes of infinitely large (see **infinity**) and **infinitesimal** numbers as well. Hyperreals emerged in the 1960s from the work of Abraham **Robinson** who showed how infinitely large and infinitesimal numbers can be rigorously defined and developed in what is called **nonstandard analysis**. Because hyperreals represent an extension of the real numbers, R, they are usually denoted by *R.

Hyperreals include all the reals (in the technical sense that they form an ordered field containing the reals as a subfield) and they also contain infinitely many other numbers that are either infinitely large (numbers whose absolute value is greater than any positive real number) or infinitely small (numbers whose absolute value is less than any positive real number). No infinitely large number exists in the real number system and the only real infinitesimal is zero. But in the hyperreal system, it turns out that each real number is surrounded by a cloud of hyperreals that are infinitely close to it; the cloud around zero consists of the infinitesimals themselves. Conversely, every (finite) hyperreal number x is infinitely close to exactly one real number, which is called its *standard part,* st(x). In other words, there exists one and only one real number st(x) such that x – st(x) is infinitesimal.

hypersphere

A four-dimensional analog of a **sphere**; also known as a *4-sphere.* Just as the shadow cast by a sphere is a circle, the shadow cast by a hypersphere is a sphere, and just as the intersection of a sphere with a plane is a circle, the intersection of a hypersphere with a hyperplane is a sphere. These analogies are reflected in the underlying mathematics.

$x^2 + y^2 = r^2$ is the Cartesian equation of a circle of radius r;

$x^2 + y^2 + z^2 = r^2$ is the corresponding equation of a sphere;

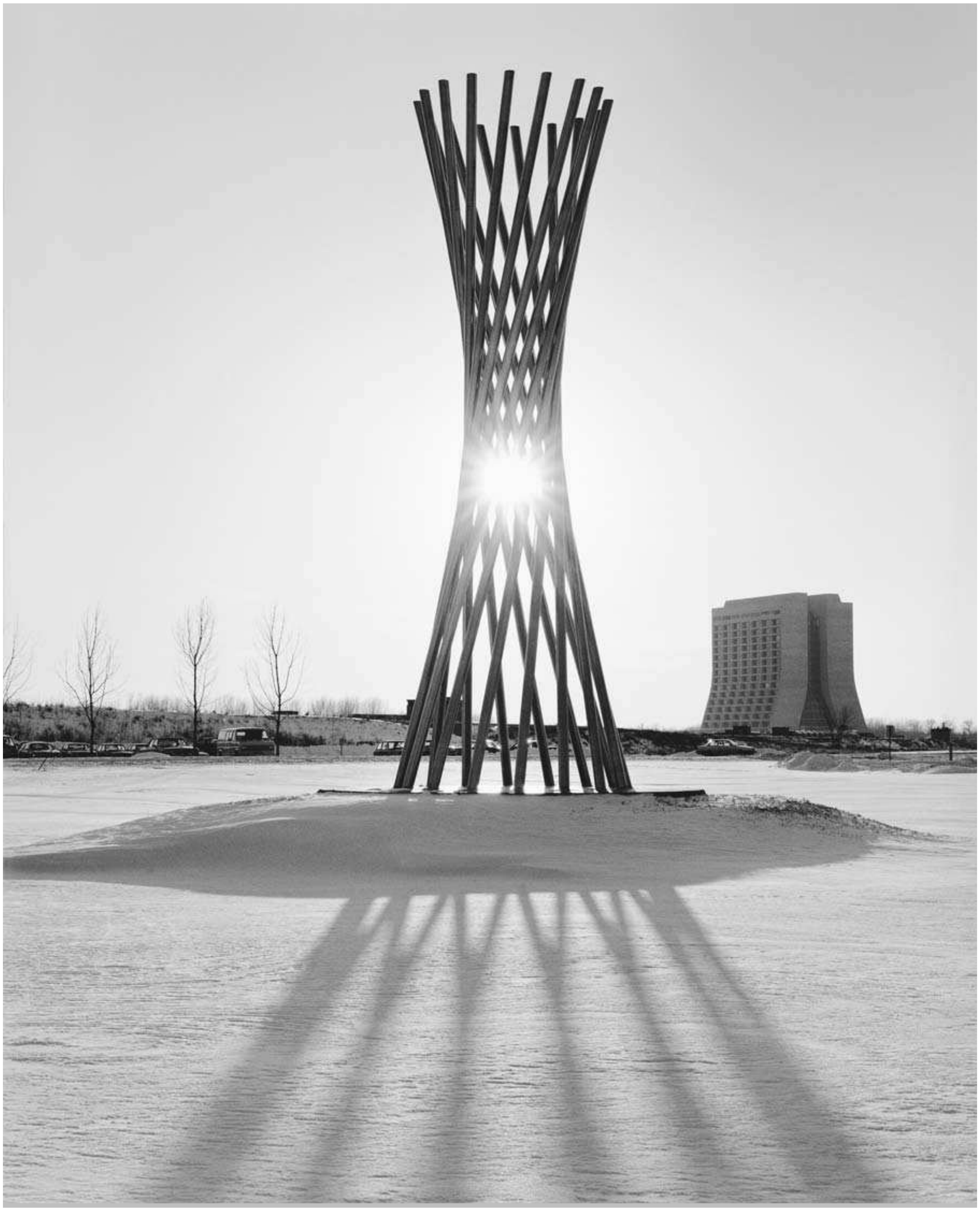

hyperboloid A hyperboloidal sculpture at the Fermi National Accelerator Laboratory. *FNAL*

$x^2 + y^2 + z^2 + w^2 = r^2$ is the equation of a hypersphere, where w is measured along a **fourth dimension** at right angles to the x-, y-, and z-axes.

The hypersphere has a *hypervolume* (analogous to the volume of a sphere) of $\pi^2 r^4/2$, and a *surface volume* (analogous to the sphere's surface area) of $2\pi^2 r^3$. A solid angle of a hypersphere is measured in *hypersteradians,* of which the hypersphere contains a total of $2\pi^2$. The apparent pattern of 2π radians in a circle and 4π steradians in a sphere does not continue with 8π hypersteradians because the *n-volume, n-area,* and number of *n-radians* of an *n-sphere* are all related to the **gamma function** and the way it can cancel out powers of π halfway between integers. In general, the term "hypersphere" may be used to refer to any n-sphere.

hypocycloid

A curve formed by the path of a point attached to a circle of radius b that rolls around the inside of a larger circle of radius a. The parametric equations for a hypocycloid can be written as:

$$x = (a - b)\cos(t) + b\cos((a/b - 1)t)$$
$$y = (a - b)\sin(t) - b\sin((a/b - 1)t)$$

The type of hypocloid depends on where the point whose path is being traced is located on the rolling circle. If it lies on the circumference of the circle, the curve generated is an ordinary hypocycloid. If it lies elsewhere the result is a **hypotrochoid**. A hypocycloid has a closed form—that is, the moving point eventually retraces its steps—when the ratio of the rolling circle and the larger circle, fixed, is equal to a rational number. When this ratio is in its simplest form, the numerator is the number of revolutions covered inside the fixed circle before the curve closes. In the same family of curves as the hypocycloid (and hypotrochoid) are the **epicycloid** and **epitrochoid**.

hypoellipse

See **superellipse**.

hypotenuse

The longest side of a **right triangle**—the one opposite the right angle. The word comes from the Greek roots *hypo* meaning "under" (which also appears, for example, in hypodermic, "under the skin") and *tein* or *ten,* for stretch. Thus hypotenuse is the line segment "stretched under" the right angle.

hypothesis

See **conjecture**.

hypotrochoid

A curve formed by the path of a point attached to a point c, which is not on the circumference, of a circle of radius b that rolls around the inside of a larger circle of radius a. The parametric equations for a hypotrochoid can be written as:

$$x = (a - b)\cos(t) + c\cos((a/b - 1)t)$$
$$y = (a - b)\sin(t) - c\sin((a/b - 1)t)$$

The hypotrochoid is a generalized **hypocycloid** and comes in two varieties: a *prolate hypocycloid* if the starting point is outside the circumference of the rolling circle and a *curtate hypocycloid* if the starting point is inside the rolling circle. In the same family of curves as the hypotrochoid (and hypocycloid) are the **epicycloid** and **epitrochoid**.

I

i

The square root of minus one; *i* is also known as the *imaginary unit number,* $\sqrt{-1}$. The product of *i* and a real number is known as an *imaginary number.* For details, see **complex number**.

I-Ching

A system of Chinese divination based on 64 hexagrams, each consisting of six horizontal lines each of which may be solid (representing the male principle, or yang) or have a break in the middle (representing the female principle, or yin). The first three lines of the hexagram, from the bottom up, constitute the lower trigram and symbolize the inner world. The fourth, fifth, and sixth lines constitute the upper trigram and symbolize the outer world. One of the first Europeans to see the hexagram structure when it was brought out of China in the early 1700s was Gottfried **Leibniz**, the first mathematician to work with base two, or **binary**, arithmetic.

icosahedron

A **polyhedron** with 20 faces. A *regular icosahedron* has faces that are all equilateral **triangle**s, and is one of the five **Platonic solid**s. The length from **vertex** to opposing vertex of a regular icosahedron is $5^{1/4} \times \phi^{1/2} \times d$ where ϕ is the **golden ratio** and d is the length of the side of one of the triangular faces. Chopping off each vertex (corner) of a regular icosahedron reveals the 12 pentagonal and 20 hexagonal faces of the *truncated icosahedron,* which is one of the 13 **Archimedean solid**s (shapes made from truncating Platonic solids in certain ways). See also **buckyball**.

Icosian game

A game devised by William **Hamilton** and first described by him in 1857 at a meeting of the British Association for the Advancement of Science in Dublin. The object of the game is to find a way around the edges of a **dodecahedron** so that every vertex (corner) is visited once and only once. A path such as this became known as a **Hamilton circuit**, though the task of finding a circuit that passes just once through every vertex of a shape seems to have arisen first in connection with Leonhard **Euler**'s study of the **knight's tour**. Two years before Hamilton introduced his game, Thomas **Kirkman** posed the problem explicitly in a paper that he submitted to the Royal Society: Given a graph of a **polyhedron**, does there exist a cycle passing through every vertex?

The Icosian game stemmed from Hamilton's invention of a curious kind of algebra that he called *icosians,* based on the symmetry properties of the **icosahedron**. Hamilton connected the mathematics of his icosians with the problem of traveling along the edges of a dodecahedron, hitting each vertex just once, and coming back to the starting point. His friend and fellow Irishman John Graves (1806–1870) suggested turning the problem into a commercial game and put Hamilton in contact with the London company of John Jacques and Sons, toy-makers and manufacturer of high quality chess sets. Jacques bought the rights to the game for £25 and marketed two versions of it, under the name Around the World. One version, for the parlor, was played on a flat board; another, for the "traveler," consisted of an actual dodecahedron. In both cases, nails at each vertex stood for a major city of the world and the player wrapped a piece of string around these nails as they went. The game was a

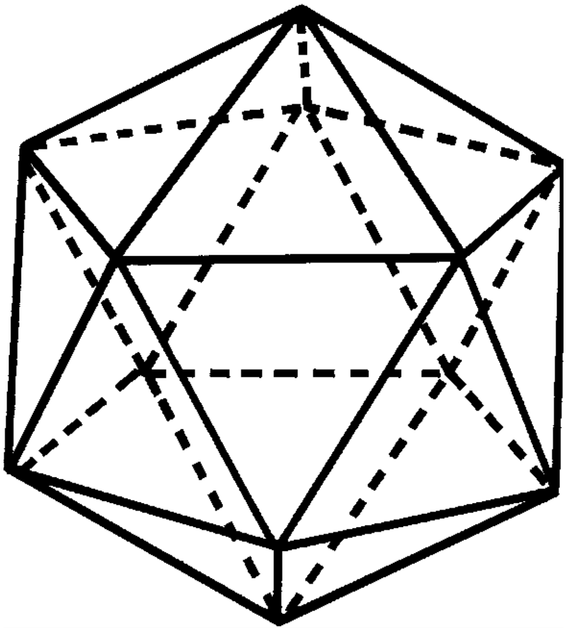

icosahedron The most multifaceted of the Platonic solids.

complete sales flop, mainly because it was too easy, even for children–but not for Hamilton himself who always used the icosian calculus to figure out his moves, instead of just trying different paths like everyone else! See also **traveling salesman problem**.

idempotent

The element x in some algebraic structures is called idempotent if $x \times x = x$.

identity

A statement that two expressions are always equal, for any values of their variables.

imaginary number

A number whose square is negative. Every imaginary number can be written in the form ib where b is a **real number** and ***i*** the *imaginary unit* with the property that $i^2 = -1$. Imaginary numbers are **complex number**s in which the real part is zero. The term "imaginary" is unfortunate because it suggests something that has less reality than a "real" number, which isn't the case.

impossibilities in mathematics

> *Alice laughed: "There's no use trying," she said;*
> *"one can't believe impossible things." "I daresay you*
> *haven't had much practice," said the Queen. "When*
> *I was younger, I always did it for half an hour a*
> *day. Why, sometimes I've believed as many as six*
> *impossible things before breakfast."*
> –Lewis Carroll, *Alice in Wonderland*

Mathematicians are used to believing things that most people would consider impossible or, at least, too outrageous to contemplate, such as the **Banach-Tarski paradox**. However, there are some genuine impossibilities, even in mathematics, including **trisecting an angle**, **duplicating the cube**, and **squaring the circle** using only a straightedge and compass; finding the center of a given circle with a straightedge alone; deriving Euclid's **parallel postulate** from the other four; and representing the **square root of 2** as a rational fraction a/b. Less well known is this little gem from Gustave Flaubert (1821–1880), who sounds as if he had seen too much of this type of problem in school:

> Since you are now studying geometry and trigonometry, I will give you a problem. A ship sails the ocean. It left Boston with a cargo of wool. It grosses 200 tons. It is bound for Le Havre. The mainmast is broken, the cabin boy is on deck, there are 12 passengers aboard, the wind is blowing East-North-East, the clock points to a quarter past three in the afternoon. It is the month of May. How old is the captain?

impossible figure

An image in two dimensions of an object that, because of spatial inconsistencies, is impossible to realize fully in three dimensions. Among the best known impossible figures are the **Penrose triangle**, the **Penrose stairway**, the **Impossible trident**, and the Freemish crate. Notable pioneers of this peculiar form of representation have been Oscar **Reutersvärd**, Roger **Penrose** (and his father), and M. C. **Escher**. See also **optical illusion**.

impossible tribar

See **Penrose triangle**.

impossible trident

One of the most notorious **impossible figure**s. It was first seen by many when it appeared on the cover of the March 1965 issue of *Mad* magazine. The two halves of the figure seem perfectly reasonable in themselves. When the top part is covered, the bottom part is taken to be three separate cylinders or tubes. With the bottom part hidden, the foreground figure is interpreted as being built of flat faces making two rectangular prongs. The trouble is that these two aspects of the figure are totally incompatible. Somewhere in the middle, the foreground and background swap places and give rise to an irreconcilable paradox. Over the years, countless adaptations of the trident have appeared with names such as the devil's fork, the three-stick clevis, the blivit, the impossible columnade, the trichotometric indicator support, and, most extravagantly, the triple encabulator tuned manifold. Swedish artist Oscar **Reutersvärd**'s mastery of such figures has led him to draw thousands of variations on the theme. When the figure is drawn long, it is easy to perceive locally as a three-dimensional object and to overlook its inherent inconsistency, because the contradictory clues are too well separated. When the figure is of medium length, the figure is easily interpreted as a three-dimensional object, and its impossibility is quickly perceived. If the prongs are very short, the two different interpretations vie for acceptance within the same local area; thus there is no consistent interpretation and the illusion breaks down. Some early writers commented that the impossible trident couldn't be built in any form in three dimensions. However, this has been shown to be false. In 1985, the Japanese artist Shigeo Fukuda made a three-dimensional model of the trident in the form of classical columns in which the illusion works–from one critical angle.

The origins of the figure are uncertain. It turns out that *Mad* magazine bought the illustration rights from a con-

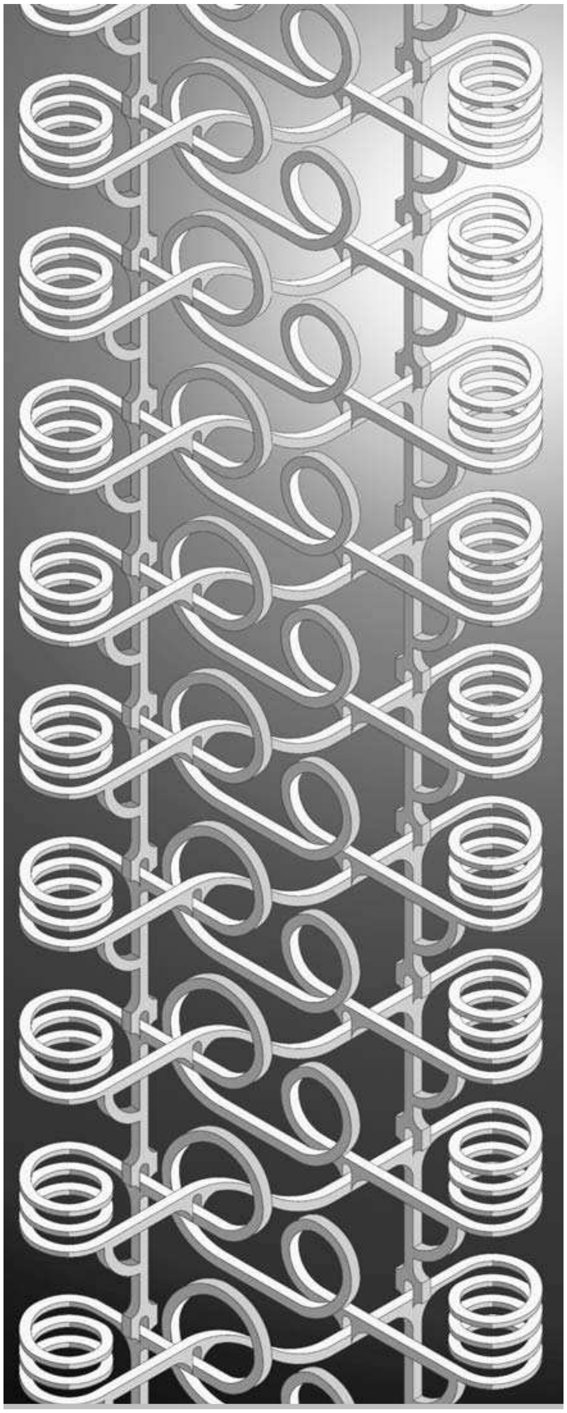

impossible figure "The Rollercoaster," a rendering in two dimensions of a structure impossible to build in three dimensions. *Jos Leys, www.josleys.com*

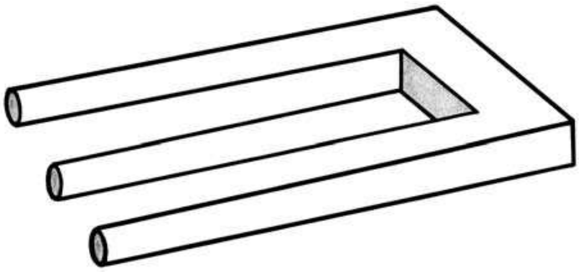

impossible trident

tributor who claimed that it was original; however, the magazine's management soon found out to their embarrassment that the figure had been previously published. It began to surface in several popular engineering, aviation, and science-fiction periodicals in May and June of 1964. D. H. Schuster published an article that same year in the *American Journal of Psychology,* which first brought the figure to the attention of the psychological community.

incircle

The **inscribed circle** of a triangle.

incomputable number

A **real number** with an infinite decimal (or binary) expansion that cannot be enumerated by any **universal computer**.

induction

A method of reasoning by which one infers a generalization from a series of instances. Say there is a hypothesis H that contains the variable n, which is a whole number. To prove by induction that H is true for every value of n is a two-step process: (1) prove that H is true for $n = 1$; (2) prove that H being true for $n = k$ implies that H is true for $n = k + 1$. This is sufficient because (1) and (2) together imply that H is true for $n = 2$, which, from (2), then implies H is true for $n = 3$, which implies H is true for $n = 4$, and so on. H is called an *inductive hypothesis.* Some philosophers don't accept this kind of proof, because it may take infinitely many steps to prove something; however, most mathematicians are happy to use it.

inequality

The statement that one quantity is less than or greater than another.

infinite dimensions

In mathematics, the concept of an infinite-dimensional space considered literally. It is a **vector space** with an infinite basis or a space with infinitely many coordinates.

infinite series

An infinite sum of the form

$$a_1 + a_2 + a_3 + \ldots = \sum_{k=1}^{\infty} a_k.$$

Such series crop up in many areas of modern mathematics. Their development began in the seventeenth century and was continued by Leonhard **Euler,** who, in the process, solved many important problems.

infinitesimal

A number that is greater than zero yet smaller than any positive **real number**. In a sense, infinitesimals are to small what **infinity** is to large. They were first introduced by Isaac **Newton** and Gottfried **Leibniz** in their early versions of **calculus**; however, the lack of a rigorous definition for them stood in the way of calculus being fully accepted. As Bertrand **Russell** later put it: "Calculus required continuity, and continuity was supposed to require the infinitely little; but nobody could discover what the infinitely little might be." In the 1800s, calculus was put on a firmer footing by Augustin **Cauchy**, Karl **Weierstrass**, and others, who clarified and redefined the notion of a **limit** without reference to infinitesimals. When a function $f(x)$ can be made as close as desired to L by taking x close enough to a, then L is the limit of $f(x)$ as x approaches a. This is the classical or *epsilon-delta* formulation of calculus, named for the common use of δ for $|x - a|$ and ε for $|f(x) - L|$. For a long time, this was the only rigorous foundation for calculus, and it is still the one taught in most calculus classes. But in 1960, Abraham **Robinson** discovered **nonstandard analysis**, which provides a rigorous formulation of infinitesimals, confers on them a new significance, and brings them closer to the vision of Newton and Leibniz.

infinity

> *Mystery has its own mysteries, and there are gods above gods. We have ours, they have theirs. That is what's known as infinity.*
>
> –Jean Cocteau (1889–1963),
> French author and filmmaker

A concept that has always fascinated philosophers and theologians, linked as it is to the notions of unending distance or space, eternity, and God, but that was avoided or met with open hostility throughout most of the history of mathematics. Only within the past century or so have mathematicians dealt with it head on and accepted infinity as a number–albeit the strangest one we know.

An early glimpse of the perils of the infinite came to Zeno of Elae through his paradoxes, the best known of which pits Achilles in a race against a tortoise (see **Zeno's paradoxes**). Confident of victory, Achilles gives the tortoise a head start. But then how can he ever overtake the sluggish reptile? First he must catch up to the point where it began, by which time the tortoise will have moved on. When he makes up the new distance that separated them, he finds his adversary has advanced again. And so it goes on, indefinitely. No matter how many times Achilles reaches the point where his competitor was, the tortoise has progressed a bit farther. So perplexed was Zeno by this problem that he decided not only was it best to avoid thinking about the infinite but also that motion was impossible!

A similar shock lay in store for **Pythagoras** and his followers who were convinced that everything in the universe could ultimately be understood in terms of whole numbers (even common fractions being just one whole number divided by another). The **square root of 2**–the length of the diagonal of a right-angled triangle whose shorter sides are both one unit long–refused to fit into this neat cosmic scheme. It was an **irrational number**, inexpressible as the ratio of two integers. Put another way, its decimal expansion goes on forever without ever settling into a recurring pattern.

These two examples highlight the basic problem in coming to grips with infinity. Our imaginations can cope with something that hasn't yet reached an end: we can always picture taking another step, adding one more to a total, or visualizing another term in a long series. But infinity, taken as a whole, boggles the mind. For mathematicians this was a particularly serious problem because mathematics deals with precise quantities and meticulously well-defined concepts. How could they work with things that clearly existed and went on indefinitely–a number like $\sqrt{2}$ or a curve that approached a line ever more closely–while avoiding a confrontation with infinity itself? **Aristotle** provided the key by arguing that there were two kinds of infinity. *Actual infinity,* or *completed infinity,* which he believed could not exist, is endlessness fully realized at some point in time. *Potential infinity,* which Aristotle insisted was manifest in nature–for example, in the unending cycle of the seasons or the indefinite divisibility of a piece of gold–is infinitude spread over unlimited time. This fundamental distinction persisted in mathematics for more than 2,000 years. In 1831, no less a figure than Karl **Gauss** expressed his "horror of the actual infinitude," saying: "I protest against the use of infinite magnitude as something completed, which is never permissible in mathematics. Infinity is merely a way of speaking, the true meaning being a limit which certain ratios approach indefinitely close, while others are permitted to increase without restriction."

By confining their attention to potential infinity, mathematicians were able to address and develop crucial concepts such as those of **infinite series**, **limit**, and **infinitesimal**s, and so arrive at **calculus**, without having to grant that infinity itself was a mathematical object. Yet as early as the Middle Ages certain paradoxes and puzzles arose, which suggested that actual infinity was not an issue to be easily dismissed. These puzzles stem from the principle that it is possible to pair off, or put in *one-to-one correspondence,* all the members of one collection of objects with all those of another of equal size. Applied to indefinitely large collections, however, this principle seemed to flout a commonsense idea first expressed by **Euclid**: the whole is always greater than any of its parts. For instance, it appeared possible to pair off *all* the positive integers with *only those that are even:* 1 with 2, 2 with 4, 3 with 6, and so on, despite the fact that positive integers also include odd numbers. Galileo, in considering such a problem, was the first to show a more enlightened attitude toward the infinite when he proposed that "infinity should obey a different arithmetic than finite numbers." Much later, David **Hilbert** offered a striking illustration of how weird the arithmetic of the endless can get.

Imagine, said Hilbert, a hotel with an infinite number of rooms. In the usual kind of hotel, with finite accommodation, no more guests can be squeezed in once all the rooms are full. But "Hilbert's Grand Hotel" is dramatically different. If the guest occupying room 1 moves to room 2, the occupant of room 2 moves to room 3, and so on, all the way down the line, a newcomer can be placed in room 1. In fact, space can be made for an infinite number of new clients by moving the occupants of rooms 1, 2, 3, etc, to rooms 2, 4, 6, etc, thus freeing up all the odd-numbered rooms. Even if an infinite number of coaches were to arrive each carrying an infinite number of passengers, no one would have to be turned away: first the odd-numbered rooms would be emptied as above, then the first coach's load would be put in rooms $3n$ for $n = 1, 2, 3, \ldots$, the second coach's load in rooms $5n$ for $n = 1, 2, \ldots$, and so on; in general, the people aboard coach number i would empty into rooms p_n where p is the $(i + 1)$th prime number.

Such is the looking-glass world that opens up once the reality of **set**s of numbers with infinitely many elements is accepted. That was a crucial issue facing mathematicians in the late nineteenth century: Were they prepared to embrace actual infinity as a number? Most were still aligned with Aristotle and Gauss in opposing the idea. But a few, including Richard **Dedekind** and, above all, Georg **Cantor**, realized that the time had come to put the concept of infinite sets on a firm logical foundation.

Cantor accepted that the well-known pairing-off principle, used to determine if two finite sets are equal, is just as applicable to infinite sets. It followed that there really are just as many even positive integers as there are positive integers altogether. This was no paradox, he realized, but the *defining property* of infinite sets: the whole is no bigger than some of its parts. He went on to show that the set of all positive integers, 1, 2, 3, . . . , contains precisely as many members–that is, has the same **cardinal number** or cardinality–as the set of all rational numbers (numbers that can be written in the form p/q, where p and q are integers). He called this infinite cardinal number **aleph**-null, $\aleph_0$ ("aleph" being the first letter of the Hebrew alphabet). He then demonstrated, using what has become known as *Cantor's theorem,* that there is a hierarchy of infinities of which $\aleph_0$ is the smallest. Essentially, he proved that the cardinal number of all the subsets–the different ways of arranging the elements–of a set of size $\aleph_0$ is a bigger form of infinity, which he called $\aleph_1$. Similarly, the cardinality of the set of subsets of $\aleph_1$ is a still bigger infinity, known as $\aleph_2$. And so on, indefinitely, leading to an infinite number of different infinities.

Cantor believed that $\aleph_1$ was identical with the total number of mathematical points on a line, which, astonishingly, he found was the same as the number of points on a plane or in any higher n-dimensional space. This infinity of spatial points, known as the *power of the continuum, c,* is the set of all **real number**s (all rational numbers plus all irrational numbers). Cantor's **continuum hypothesis** asserts that $c = \aleph_1$, which is equivalent to saying that there is no infinite set with a cardinality between that of the integers and the reals. Yet, despite much effort, Cantor was never able to prove or disprove his continuum hypothesis. We now know why–and it strikes to the very foundations of mathematics.

In the 1930s, Kurt **Gödel** showed that it is impossible to *disprove* the continuum hypothesis from the standard **axiom**s of set theory. Three decades later, Paul Cohen showed that it *cannot be proven* from those same axioms either. Such a situation had been in the cards ever since the emergence of **Gödel's incompleteness theorem**. But the independence of the continuum hypothesis was still unsettling because it was the first concrete example of an important question that could not be proven either way from the universally accepted system of axioms on which most of mathematics is built.

Currently, the preference among mathematicians is to regard the continuum hypothesis as being false, simply because of the usefulness of the results that can be derived this way. As for the nature of the various types of infinities and the very existence of infinite sets, these depend crucially on what number theory is being used. Different axioms and rules lead to different answers to the question *what lies beyond all the integers?* This can make it difficult or even meaningless to compare the various types of infinities that arise and to determine their relative size, although within any given number system the infinities can usually be put into a clear order. Certain extended number systems, such as the **surreal number**s, incorporate both the ordinary (finite) numbers and a diversity of infinite numbers. However, whatever number system is chosen, there will inevitably be inaccessible infinities–infinities that are larger than any of those the system is capable of producing.[198, 214, 275]

inflection

A point of inflection of a plane curve is a point where the curve has a stationary **tangent**, at which the tangent is changing from rotating in one direction to rotating in the opposite direction.

information theory

A mathematical theory of information born in 1948 with the publication of Claude **Shannon**'s landmark paper, "A Mathematical Theory of Communication." Its main goal is to discover the laws governing systems designed to communicate or manipulate information, and to set up quantitative measures of information and of the capacity of various systems to transmit, store, and otherwise process information. Among the problems it treats are finding the best methods of using various communication systems and the best methods for separating the wanted information, or signal, from the noise. Another of its concerns is setting upper bounds on what it is possible to achieve with a given information-carrying medium (often called an information channel). The theory overlaps heavily with communication theory but is more oriented toward the fundamental limitations on the processing and communication of information and less oriented toward the detailed operation of the devices used.

injection

A **one-to-one** mapping.

inscribed angle

The angle formed by two **chord**s of a curve that meet at the same point on the curve.

inscribed circle

A circle inside a triangle or other **polygon** whose edges are **tangent**s to the circle. Such a polygon is said to be *circumscribed* to the circle. The (unique) circle inscribed to a triangle is called the *incircle.*

instability

The condition of a system when it is easily disturbed by internal or external forces or events, in contrast to a *stable system,* which will return to its previous condition when disturbed. A pencil resting vertically on its eraser or a coin standing on edge are examples of systems that have the property of instability since they easily fall over at the slightest breeze or movement of the surface they are resting on. An unstable system is one whose **attractor**s can change; thus, instability is a characteristic of a system that is far from equilibrium or at **bifurcation**.

integer

Any positive or negative whole number or zero: . . ., –3, –2, –1, 0, 1, 2, 3, ... *Integer* is Latin for "whole" or "intact." The **set** of all integers is denoted by **Z**, which stands for *zahlen* (German for "numbers"). The integers are an extension of the **natural number**s to include **negative number**s and so make possible the solution of all equations of the form $a + x = b$, where a and b are natural numbers. Integers can be added and subtracted, multiplied, and compared. Like the natural numbers, the integers form a countably infinite set. However, the integers don't form a **field** since, for instance, there is no integer x such that $2x = 1$; the smallest field containing the integers is that of the **rational number**s. An important property of the integers is *division with remainder:* given two integers a and b with $b \neq 0$, it is always possible to find integers q and r such that $a = bq + r$, and such that $0 \leq r < |b|$. q is called the *quotient* and r is called the *remainder* resulting from division of a by b. The numbers q and r are uniquely determined by a and b. From this follows the *fundamental theorem of arithmetic,* which states that integers can be written as products of prime numbers in an essentially unique way.

integral

The area, or a generalization of area, under any section of a graph that is described by a **function**; in other words, the continuous cumulative sum of a function (see **continuity**). Not all functions have an exact formula that allows an integral to be found. In such cases, *numerical integration* has to be used, in which the area is found using approximate numerical techniques. Integrals, together with **derivative**s, are the fundamental objects of **calculus**.

integral equation

An equation that involves an **integral** of the **function** which is to be solved for.

integration

An operation that corresponds to the informal idea of finding the area under the graph of a **function**. The first theory of integration was developed by **Archimedes** with his method of *quadratures,* but this could be applied only in circumstances where there was a high degree of geometric symmetry. In the seventeenth century, Isaac **Newton** and Gottfried **Leibniz** independently discovered the idea that integration was a sort of opposite of **differentiation** (which they had just invented); this allowed mathematicians to calculate a broad class of integrals for the first time. However, unlike Archimedes's method, which was based on **Euclidean geometry**, Newton's and Leibniz's integral calculus lacked a secure foundation.

In the nineteenth century, Augustin **Cauchy** finally developed a rigorous theory of **limit**s, and Bernhard **Riemann** followed this up by formalizing what is now called the **Riemann integral**. To define this integral, one fills the area under the graph with smaller and smaller rectangles and takes the limit of the sums of the areas of the rectangles at each stage. Unfortunately, some functions don't have well-defined limits to these sums, so they have no Riemann integral. Henri **Lesbesgue** invented another method of integration to solve this problem. He first presented his ideas in *Intégrale, Longueur, Aire* (Integral, length, area) in 1902. Instead of using the areas of rectangles, a method that puts the focus on the **domain** of the function, Lebesgue turned to the **codomain** of the function for his fundamental unit of area. Lesbesgue's idea was to build first the integral for what he called simple functions–functions that take only finitely many values. Then he defined it for more complicated functions as the upper bound of all the integrals of simple functions smaller than the function in question. *Lesbesgue integration* has the beautiful property that every function with a Riemann integral also has a Lebesgue integral, and the two integrals agree. But there are many functions with a Lebesgue integral that don't have a Riemann integral. As part of the development of Lebesgue integration, Lebesgue introduced the concept of *Lebesgue measure,* which measures lengths rather than areas. Lebesgue's technique for turning a measure into an integral generalizes easily to many other situations, leading to the modern field of **measure theory**.

The Lebesgue integral was deficient in one respect. The Riemann integral had been generalized to the *improper Riemann integral* to measure functions whose domain of definition was not a **closed** interval. The Lebesgue integral integrated many of these functions (always reproducing the same answer when it did), but not all of them. The *Henstock integral* is an even more general notion of integral (based on Riemann's theory rather than Lebesgue's) that subsumes both Lebesgue integration and improper Riemann integration. However, the Henstock integral depends on specific features of the real number line and so doesn't generalize as well as the Lebesgue integral does.

interesting numbers

Clearly some numbers are of greater interest (at least to mathematicians) than are others. The number **pi**, for instance, is far more interesting than 1.283–or virtually any other number for that matter. Confining our attention to integers, can there be such a thing as an uninteresting number? It is easy to show that the answer must be "no." Suppose there were a set *U* of uninteresting integers. Then it must contain a least member, *u.* But the property of being the smallest uninteresting integer makes *u* interesting! As soon as *u* is removed from *U,* there is a new smallest uninteresting integer, which must then also be excluded. And so the argument could be continued until *U* was empty. Given that all integers are interesting, can they be ranked from least to most interesting? Again, no. To be ranked as "least interesting" is an extremely interesting property, and thus leads to another logical contradiction!

When Srinivasa **Ramanujan**, the great Indian mathematician, was ill with tuberculosis in a London hospital, his colleague G. H. **Hardy** went to visit him. Hardy opened the conversation with: "I came here in taxicab number 1729. That number seems dull to me which I hope isn't a bad omen." Ramanujan replied, without hesitation: "Nonsense, the number isn't dull at all. It's quite interesting. It's the smallest number that can be expressed as the sum of two cubes in two different ways." (1729 = $1^3 + 12^3$ and $9^3 + 10^3$.)

International Date Line

Is it possible to assign a time to every longitude on Earth, so that each longitude has a *different* time but the times at nearby longitudes are always close? The answer is no, which is mathematically equivalent to saying that there's no way to continuously map points on a **real number** line onto a circle. This explains why an International Date Line is needed. It allows most regions on Earth to have times similar to their neighbors, though, by convention, time changes are (usually) made in chunks of one hour between adjacent time zones. Then it takes care of the inevitable discontinuity by having it happen all at once, as a jump by one whole day, on a longitude that passes mostly through open water in the Pacific. The fact that there doesn't exist any continuous **one-to-one** function from the circle onto the line follows from the **Borsuk-Ulam theorem** in one dimension.

interpolate

To estimate the value of a point that lies between two known values of a **function**. This is often done by approximating a line or a smooth curve between the values, which is the literal meaning of the word. *Inter* is the

Latin prefix for "between," *polire* translates as "to adorn or polish," so together they mean "to smooth between." *Extrapolate* was created as an extension of interpolate to suggest the smoothing of a line *outside* the known points. This operation is often done in statistics when patterns are studied over time to predict future events.

intersection

A place where two or more things meet or overlap. Two lines intersect at a point, two planes can intersect in a line, and so forth. The intersection of two or more **set**s, represented by the symbol $\cap$, is the set of elements that all the sets have in common; in other words, all the elements contained in every one of the sets.

invariant

(1) Something that stays the same when a particular **transformation** is carried out. (2) A value that is unchanged when a particular **function** is applied. (3) In **topology**, a number, **polynomial**, or other quantity associated with a topological object such as a **knot** or 3-**manifold**, which depends only on the underlying object and not on its specific description or presentation.

invariant theory

The study of quantities that are associated with **polynomial** equations and that are left **invariant** under transformations of the variables. For example, the discriminant $b^2 - 4ac$ is an invariant of the **quadratic** form $ax^2 + bxy + cy^2$.

inverse

(1) The inverse of a number, or a **reciprocal**, is 1 divided by the number; for example, the inverse of 8 is $\frac{1}{8}$ and the inverse of $\frac{3}{5}$ is $\frac{5}{3}$. (2) The inverse of a **function** or a **transformation** is the function or transformation that reverses the effect of the function or transformation. For example, the inverse of addition is subtraction, and of clockwise rotation is anticlockwise rotation. (3) The inverse of an element of a **set**, or a number, with respect to a particular operation, is what has to be combined with the element or number in order to obtain that operation's *identity element.*

involute

Attach a string to a point on a curve. Extend the string so that it is **tangent** to the curve at the point of attachment. Then wind the string up, always keeping it taut. The **locus** of points traced out by the end of the string is called the involute of the original curve, and the original curve is called the **evolute** of its involute. Although a curve has a unique evolute, it has infinitely many involutes corresponding to different choices of initial point. An involute can also be thought of as any curve **orthogonal** to all the tangents to a given curve. See also **circle involute**.

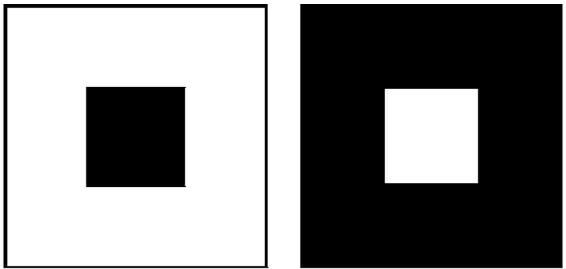

irradiation illusion Despite appearances, the inner squares are the same size.

irradiation illusion

A **distortion illusion** discovered by Hermann von Helmholtz in the nineteenth century. Despite the fact that the two figures are identical in size, the white hole looks bigger than the black one.

irrational number

A **real number** that can't be written as one whole number divided by another; in other words, a real number that isn't a **rational number**. The decimal expansion of an irrational number doesn't come to an end or repeat itself (in equal length blocks), though it may have a pattern such as 0.101001000100001000001 . . . . The vast majority of real numbers are irrational, so that if you were to pick a single point on the real **number line** at random the chances are overwhelmingly high that it would be irrational. Put another way, whereas the set of all rationals is countable, the irrationals form an **uncountable set** and therefore represent a larger kind of **infinity**. Indeed, as the Harvard logician Willard Van Orman Quine pointed out: "The irrationals exist in such variety . . . that no notation whatever is capable of providing a separate name for each of them."

There are two types of irrational number: **algebraic number**s, such as the **square root of 2**, which are the roots of algebraic equations, and the **transcendental number**s, such as **pi** and **e**, which aren't. In some cases it isn't known if a number is irrational or not; undecided cases include 2^e, π^e, π^2, and the **Euler-Mascheroni constant**, γ. An irrational number raised to a rational power can be rational; for instance, $\sqrt{2}$ to the power 2 is 2. Also, an irrational number to an irrational power can be rational. What kind of number is $\sqrt{2}^{\sqrt{2}}$? The answer is irrational. This follows from Gelfond's theorem, which says that if *a* and *b* are roots of **polynomial**s, and *a* is not 0 or 1 and *b* is irrational, then a^b must be irrational (in fact, transcendental).

irreptile

See **rep-tile**.

Ishango bone

A bone tool handle discovered around 1960 in the African area of Ishango, near Lake Edward. It has been dated to about 9000 B.C. and was at first thought to have been a **tally** stick. At one end of the bone is a piece of quartz for writing, and the bone has a series of notches carved in groups on three rows running the length of the bone. The markings on two of these rows each add to 60. The first row is consistent with a number system based on 10, since the notches are grouped as 20 + 1, 20 – 1, 10 + 1, and 10 – 1, while the second row contains the **prime number**s between 10 and 20! A third seems to show a method for multiplying by 2 that was used in later times by the Egyptians. Additional markings suggest that the bone was also used as a lunar phase counter. The Ishango Bone is kept at the Royal Institute for Natural Sciences of Belgium in Brussels. See also **Lebombo bone**.[45, 157]

isochrone

A set of points with the property that a given process or trajectory will take the same length of time to complete starting from any of the points. The curve formed by such a set of points is called an *isochronous curve*. See also **tautochrone problem**.

isogonal conjugate

Isogonal lines of a triangle are cevians (see **Ceva, Giovanni**) that are symmetric with respect to the angle bisector. Two points are isogonal conjugates if the corresponding lines to the vertices are isogonal (see **vertex**).

isometry

A symmetry operation, which may involve **translation**, **rotation**, and **reflection**, that preserves the distance of any two points. Each isometry of a **wallpaper group**, for example, can be represented by a 3 × 3 matrix.

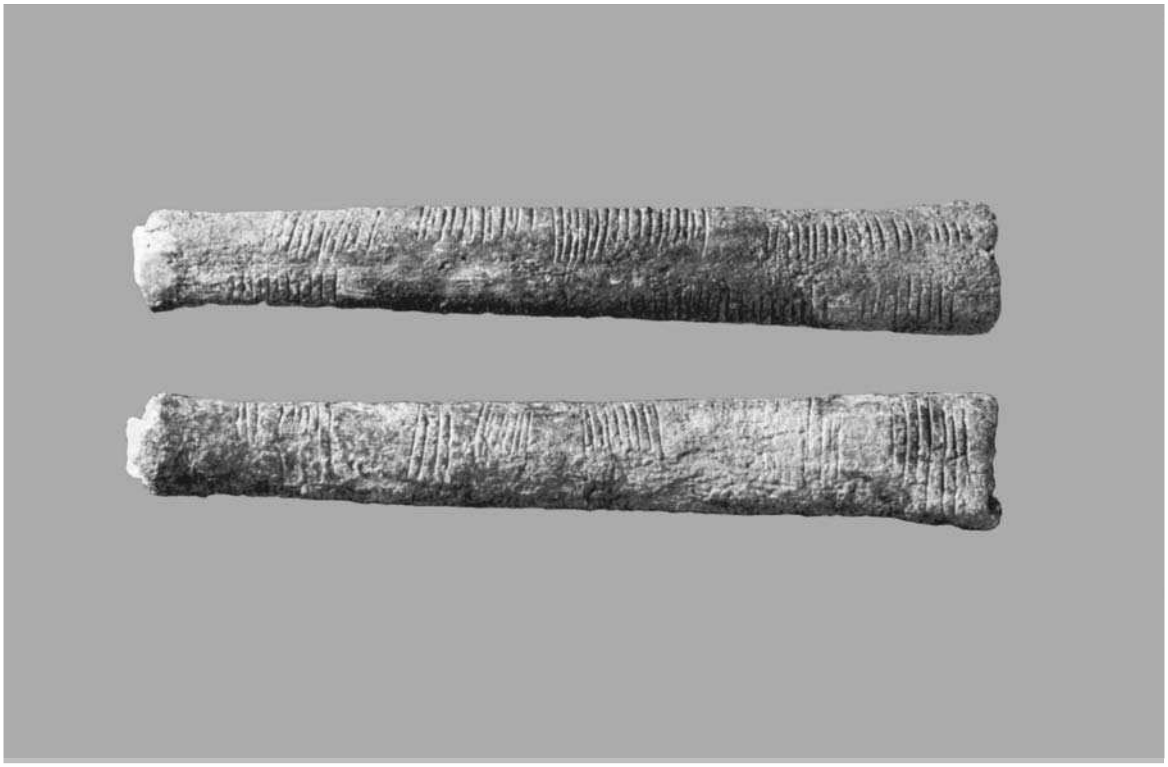

Ishango bone The marks on this 11,000-year-old bone speak of a surprisingly sophisticated mathematical knowledge. *Museum of Natural Sciences (Brussels)*

isomorphism

In geometry, a **transformation** that doesn't alter the side lengths and the angle sizes of the figure involved. Examples of such transformations include reflections, rotations, translations, or transformations by a glide. In **set theory**, an isomorphism is a **one-to-one** correspondence between the elements of two sets such that the result of an operation on elements of one set corresponds to the result of the analogous operation on their images in the other set.

isoperimetric inequality

For any closed, three-dimensional body with volume V and surface area A, the following inequality always holds:

$$36\pi V^2 \leq A^3$$

isoperimetric problem

Among all shapes of a given perimeter, which encloses the greatest area? This ancient conundrum is alluded to in the tale of Queen Dido in Virgil's *Aeneid*. Threatened by her brother, who had already murdered her father, the queen was obliged to hastily gather her valuables and flee her native city of Tyria in ancient Phoenicia. In due course, her ship landed in North Africa, where she made the following offer to a local chieftain: in return for her fortune he would give as much land as she could isolate with the skin of an ox. This was readily agreed to, whereupon the crafty queen sliced the skin into very fine threads, which she tied together and made into a giant semicircle. Combined with the natural boundary provided by the sea, this enclosed such a large area that a city–Carthage–was eventually built upon it. Two millennia later, Karl **Weierstrass** was the first to prove rigorously, using analysis and calculus, that the solution to the isoperimetric problem is a circle (something the Greeks suspected but were not able to prove geometrically). When the same question is asked in one dimension higher it becomes the **isovolume problem**.

isosceles

Having two sides of the same length, as in the case of an isosceles **triangle**. An isosceles **trapezoid** in the United States is the equivalent of a trapezium in Britain. *Isosceles* comes from the Greek *iso* ("same or equal") and *skelos* ("legs").

isotomic conjugate

Two points on the side of a triangle are *isotomic* if they are equidistant from the midpoint of that side. Two points inside a triangle are isotomic conjugates if the corresponding cevians (see **Ceva, Giovanni**) through these points meet the opposite sides in isotomic points.

isovolume problem

What surface encloses the maximum volume per unit surface? The answer is a **sphere**, for which

$$\text{volume/surface area} = (\tfrac{4}{3}\pi r^3)/(4\pi r^2) = \tfrac{1}{3}r,$$

where r is the sphere's radius. Proof of the solution to the isovolume problem came as recently as 1882 from Hermann Schwarz. See also **bubbles** and **isoperimetric problem**.

iterate

Do something repeatedly. Do something repeatedly. . . .

iteration

> *If at first you don't succeed, try, try, again. Then quit. There's no use being a damn fool about it.*
>
> –W. C. Fields

A **feedback** process that repeats n number of times. Iteration refers to the act of performing the calculation of a certain **function** and then picking the result, or output, as the starting value, or input, for the next calculation of the same function. The operation repeats on and on–even infinitely, despite Fields's comment!

J

Jacobi, Karl Gustav Jacob (1804–1851)

A German mathematician who did important work on **elliptic functions**, **partial differential equations**, and mechanics. Although he was preceded in many of his discoveries about elliptic functions by Carl **Gauss** (who didn't publish) and Niels **Abel**, Jacobi is nevertheless considered one of the founders of the subject. His name is probably best known from the *Jacobian,* an n n **determinant** formed from a set of n functions in n unknowns. He wasn't the first to use it–the "Jacobian" appears in an 1815 paper of Agustin **Cauchy**–but Jacobi did write a long memoir about it in 1841, and proved that the Jacobian of n functions vanishes if and only if the functions are related (Cauchy had proved the "if" part). He also did important work on partial differential equations and their application to physics. Along with William **Hamilton**, he developed an approach to mechanics based on generalized coordinates. In this method, the total energy of a mechanical system is represented as a function of generalized coordinates and corresponding generalized momenta; for example, in a double pendulum the two generalized coordinates could be two angles. *Hamilton-Jacobi theory* is the technique of solving the system by transforming coordinates so that the transformed coordinates and momenta are constants.

Jacobi was appointed to a position at the University of Königsberg in 1826. He gained a reputation as a gifted teacher and is credited with introducing the seminar method (giving lectures on his own ongoing research) into the university. After a collapse from overwork in 1843, Jacobi was allowed to stay in Berlin with a generous allowance from the king of Prussia. Five years later, revolution swept Europe and Jacobi was persuaded to run for parliament. This proved a disaster; not only did he loose the election, but his foray into politics annoyed his royal patron, who cut off his pension. Jacobi, with a large family to support, was faced with destitution. Only his reputation as the greatest German mathematician besides Gauss saved him; faced with the prospect of losing Jacobi to the University of Vienna, the king was prevailed upon to restore the pension. Jacobi was a notoriously hard worker (indeed, he had several breakdowns due to overwork), but his death in 1851 was the result of smallpox.

Johnson solid

Any convex **polyhedron** with regular faces that is not a **Platonic solid**, an **Archimedean solid**, or a **prism** (or an antiprism). There are 92 Johnson solids, which are named after Norman W. Johnson who was the first to catalog them in 1966. They include equilateral **deltahedra** and dipyramids (two pyramids placed symmetrically base to base) and any irregular convex solid that can be made by sticking together triangles, squares, pentagons, and so on, in a way that happens to close. Some of the simpler Johnson solids are assemblages of pyramids, prisms, and antiprisms; for example, a gyrobifastigium is two triangular prisms glued together with a twist. Others are fragments of Archimedean solids; for example, a pentagonal rotunda is half an icosidodecahedron. When stretched out with a prism this makes the marvelously named elongated pentagonal orthocupolarotunda.

Johnson's theorem

If three congruent circles all intersect in a single point, then the other three points of intersection will lie on another circle of the same radius. This simple little theorem was discovered by Roger Johnson in 1916.

Jordan, (Marie-Ennemond) Camille (1838–1922)

A French mathematician who made important contributions to **group** theory. He was the first to draw attention to the work of Evariste **Galois**, which had until then been almost entirely ignored. He built on Galois's study by grasping the intimate connection between groups of permutations and the solvability of **polynomial** equations. Jordan also introduced the idea of an infinite group. He passed on his interest in group theory to two of his most outstanding pupils, Felix **Klein** and Sophus **Lie**, both of whom went on to develop the subject in new and important ways.

Jordan curve

A simple, **closed curve**.

Jordan matrix

A **matrix** whose **diagonal** elements are all equal (and nonzero) and whose elements above the principal diagonal are equal to 1, but all other elements are 0.

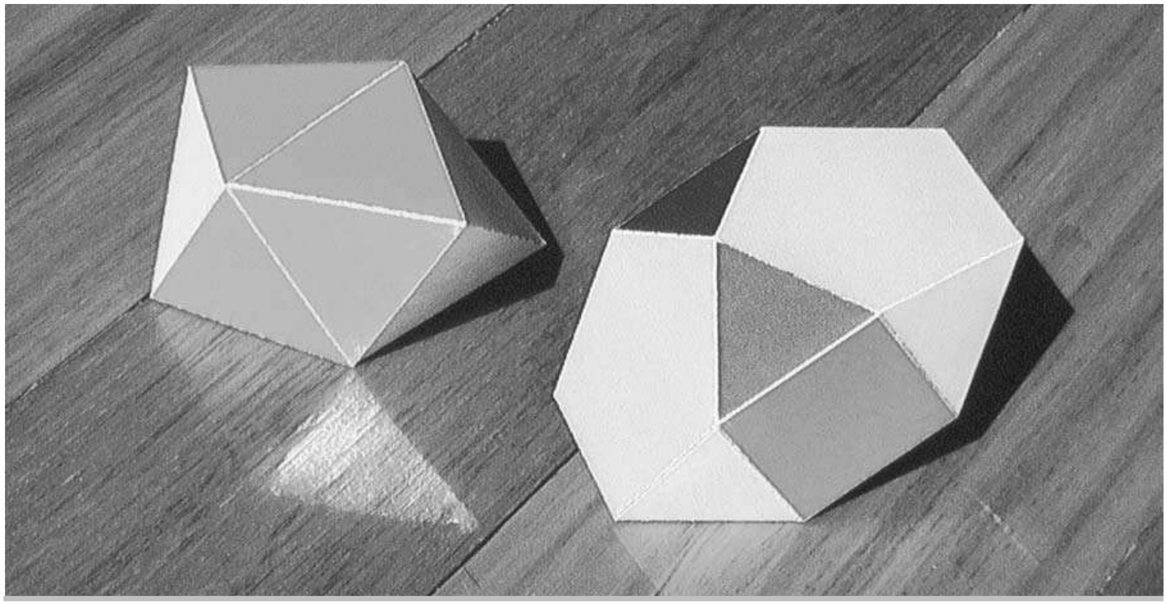

Johnson solid Two of the 92 Johnson solids: the bilunabirotunda (J91, right) and the snub disphenoid (J84, left). *Robert Webb, www.software3d.com; created using Webb's Stella program*

Journal of Recreational Mathematics

The only international journal devoted to the lighter side of mathematics. It was started in 1968 and is currently edited by Charles Ashbacher.

Julia set

Any of any infinite number of **fractal** sets of points on an **Argand diagram** (the complex plane) defined by a simple rule. Given two **complex number**s z and c, and the **recursion** $z_{n+1} = z_n^2 + c$, the Julia set for any given value of c, consists of all values of z for which z, when iterated in the equation above, does not "blow up" or tend to infinity. Julia sets are closely related to the **Mandelbrot set**, which is, in a way, an index of all Julia sets. For any point on the complex plane (which represents a value of c) a corresponding Julia set can be drawn. We can imagine a movie of a point moving about the complex plane with its corresponding Julia set. When the point lies inside the Mandelbrot set the corresponding Julia set is topologically unified or connected. As the point crosses the boundary of the Mandelbrot set, the Julia set explodes into a cloud of disconnected points called *Fatou dust.* If c is on the boundary of the Mandelbrot set, but not a waist point (where two large regions of the Mandelbrot set are connected by a narrow bridge), the Julia set of c looks like the Mandelbrot set in sufficiently small neighborhoods of c.

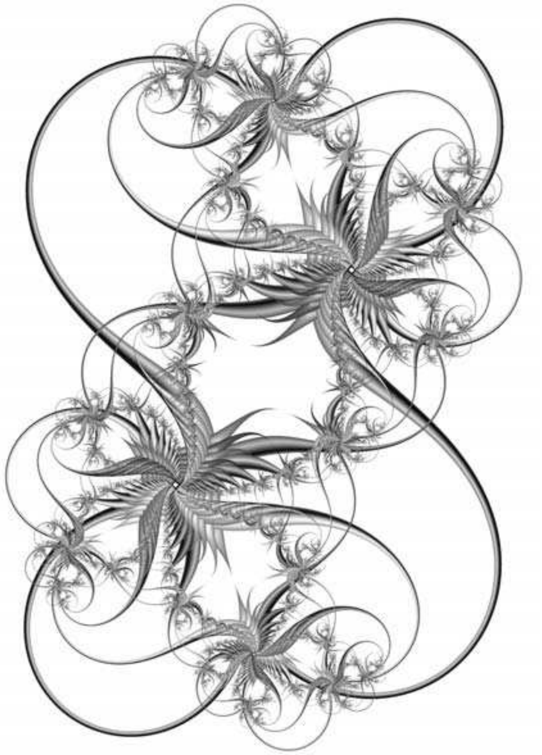

Julia set A fractal image based on a Julia set. *Jos Leys, www.josleys.com*

Julia sets are named after the French mathematician Gaston Julia (1893–1978), whose most famous work, *Memoire sur l'iteration des fonctions rationnelles,* provides the theory for Julia sets before computers were available. In 1918, at the age of 25 Julia was severely wounded in World War I and lost his nose. He wrote his greatest treatise in a hospital between the painful operations necessitated by his wounds.

jump discontinuity

A discontinuity in a **function** where the left- and right-hand limits exist but are not equal to each other.

K

Kaczynski, Theodore John (1942–)

Known as the Unabomber before his true identity was discovered, a Harvard graduate with a genius IQ who abandoned a promising career as a mathematician for a life of social isolation and intermittent terrorist attacks. Kaczynski eluded the FBI for 17 years, during which time he orchestrated 16 explosions that killed 3 people and injured 23 others. He avoided execution following his arrest in 1996 through a plea bargain and was given instead 4 life sentences plus 30 years in prison. During his mathematical career in the late 1960s, Kaczynski published a doctoral dissertation and several papers in academic journals and served as an assistant math professor at the University of California, Berkeley (1967–1969).

Kakeya needle problem

A famous problem named after the Japanese mathematician Sôichi Kakeya, who first posed it in 1917. It asks: What is the smallest-area plane figure inside which a unit straight-line segment can be rotated through 180°? For some years, the answer was thought to be a **deltoid**. However, in 1928, the Russian mathematician Abram Besicovitch shocked the mathematical world by showing that the problem *had no answer* or, to be more precise, that there was no minimum area.[39] In 1917 Besicovitch had been working on a problem in **Riemann integration**, and had reduced it to the question of existence of planar **set**s of **measure** 0, which contain a line segment in each direction. He then constructed such a set, and published his construction in a Russian journal in 1920. Due to the civil war and the blockade, there was hardly any communication between Russia and the rest of the world at the time, so Besicovitch hadn't heard about the challenge posed by Kakeya. Several years later, after he left Russia and learned about the needle problem, Besicovitch modified his original construction and was able to give the startling answer that the area in question may be made arbitrarily small.

Kaluza-Klein theory

A model that seeks to unite classical gravity and electromagnetism by resorting to **higher dimensions**. In 1919 the German mathematician Theodor Kaluza (1885–1954) pointed out that if general **relativity theory** is extended to a five-dimensional **space-time**, the equations can be separated out into ordinary four-dimensional gravitation plus an extra set (which is equivalent to Maxwell's equations for the electromagnetic field) plus an extra field known as the dilaton). Thus, electromagnetism is explained as a manifestation of curvature in a **fourth dimension** of physical space, in the same way that gravitation is explained in Einstein's theory as a manifestation of curvature in the first three. In 1926 the Swedish physicist Oskar Klein (1894–1977) proposed that the reason the extra spatial dimension goes unseen is that it is compact—curled up like a ball with a fantastically small radius. In the 1980s and 1990s, Kaluza-Klein theory experienced a big revival and can now be seen as a precursor of **string theory**.

Kampyle of Eudoxus

An hourglass-shaped curve that was first studied by **Eudoxus** in an attempt to solve the classical problem of **duplicating the cube**. It is described by the Cartesian equation

$$a^2x^4 = b^4(x^2 + y^2).$$

It is also the **radial curve** of the **catenary**.

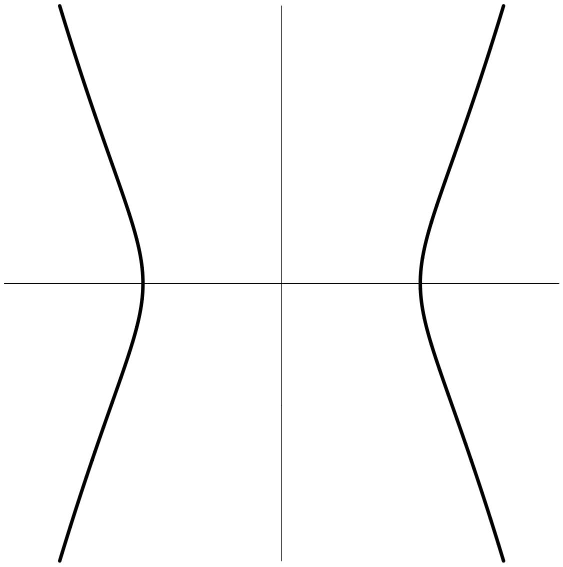

Kampyle of Eudoxus *© Jan Wassenaar, www.2dcurves.com*

kappa curve

Also known as *Gutschoven's curve,* named after Gérard van Gutschoven (1615–1668), who first studied it around 1662. The kappa curve, which resembles the Greek letter κ, was also investigated by Isaac **Newton** and, some years later, by Johann Bernoulli (see **Bernoulli family**). It is given by the Cartesian equation

$$y^2(x^2 + y^2) = a^2x^2.$$

Kaprekar constant

Take any four-digit number whose digits are not all identical. Rearrange the digits to make the largest and smallest 4-digit numbers possible. Subtract the smaller number from the larger. Use the resulting number and repeat the process. For example, starting with 4,731: 7,431 − 1,347 = 6,084; 8,640 − 468 = 8,172; 8,721 − 1,278 = 7,443; 7,443 − 3,447 = 3,996; 9,963 − 3,699 = 6,264; 6,642 − 2,466 = 4,176; 7,641 − 1,467 = 6,174. After this, the result is always 6174. Remarkably, *every* four-digit number whose digits are not all the same will eventually reach 6,174, in at most seven steps, and then stay there. This is called the Kaprekar constant for four-digit numbers, after the Indian mathematician Dattathreya Ramachandra Kaprekar who made the discovery in 1949. The Kaprekar constant for three-digit numbers is 495, which is arrived at for any three-digit number in no more than six iterations. The same process, or algorithm, can be applied to numbers of n-digits, where n is any whole number. Depending on the value of n, the algorithm will result in a nonzero constant, zero (the degenerate case), or a cycle.

Kaprekar number

Take a positive whole number n that has d number of digits. Take the square of n and separate the result into two pieces: a right-hand piece that has d digits and a left-hand piece that has either d or $d - 1$ digits. Add these two pieces together. If the result is n, then n is a Kaprekar number. Examples are 9 ($9^2 = 81$, $8 + 1 = 9$), 45 ($45^2 =$ 2,025; $20 + 25 = 45$), and 297 ($297^2 =$ 88,209; $88 + 209 = 297$). The first 10 Kaprekar numbers according to this definition are 1; 9; 45; 55; 99; 297; 703; 999; 2,223; and 2,728. Kaprekar numbers can also be defined by higher powers. For example, $45^3 =$ 91,125; and $9 + 11 + 25 = 45$. The first 10 numbers with this property are: 1; 8; 10; 45; 297; 2,322; 2,728; 4,445; 4,544; and 4,949. For fourth powers, the sequence begins 1; 7; 45; 55; 67; 100; 433; 4,950; 5,050; 38,212. Notice that 45 is a Kaprekar number for second, third, and fourth powers ($45^4 =$ 4,100,625; and $4 + 10 + 06 + 25 = 45$)–the only number in all three Kaprekar sequences, up to at least 400,000. See also **unique number**.

Kasner, Edward (1876–1955)

An American mathematician at Columbia University best remembered for introducing the words ***googol*** and *googolplex* into the popular mathematical lexicon. He is also well known as the coauthor, with James **Newman**, of *Mathematics and the Imagination,* first published in 1940.[186] In a later edition (1967), he spoke about the term *mathescope,* which was coined by the science reporter Wilson Davis after listening to one of Kasner's public lectures. In Kasner's words: "It is not a physical instrument; it is a purely intellectual instrument, the ever-increasing insight which mathematics gives into the fairyland which lies beyond intuition and beyond imagination." His main field of research was **differential geometry**, which he studied in its applications to mechanics, cartography, and stereographic projections, though he also wrote papers on circle packing and on the horn angle and studied an extension of right triangles to the complex plane.

Kepler, Johannes (1571–1630)

A German astronomer and mathematician best remembered for his three laws of planetary motion. He also did important work in optics, discovered two new regular polyhedra (1619), gave the first mathematical treatment of close **packing** of equal spheres (in 1611, leading to an explanation of the shape of the cells of a honeycomb), gave the first proof of how **logarithm**s work (1624), and devised a method of finding the volumes of **solids of revolution** that (with hindsight) can be seen as contributing to the development of **calculus** (1615–1616). Although his work eventually led to our modern view of the solar system, he himself held beliefs about the arrangement of the planets that smack of Pythagorean mysticism, though ultimately they stemmed from his belief that mathematical harmony was a reflection of God's perfection. In his first cosmological model (*Mysterium cosmographicum,* 1596) he suggested that if a sphere were drawn to touch the inside of the path of Saturn, and a cube were inscribed in the sphere, then the sphere inscribed in that cube would be the sphere circumscribing the path of Jupiter. Then if a regular tetrahedron were drawn in the sphere inscribing the path of Jupiter, the insphere of the tetrahedron would be the sphere circumscribing the path of Mars, and so inward, putting the regular dodecahedron between Mars and Earth, the regular icosahedron between Earth and Venus, and the regular octahedron between Venus and Mercury. Thus the number of (then-known) planets is explained perfectly in terms of the five convex regular solids–the **Platonic solid**s.

Kepler's second work on cosmology (*Harmonices mundi* [Harmony of the world], book V, 1619) offers a more elaborate mathematical model but still with the polyhedra in place. The mathematics in this work includes the first systematic treatment of **tessellation**s, a proof that there are only 13 convex uniform polyhedra (the **Archimedean solid**s) and the first account of two non-convex regular polyhedra (see **Kepler-Poinsot solid**s).

Kepler's mathematical work was full time and continued even during the wedding to his second wife in 1613. The dedicatory letter to a book he wrote shortly after explains that at the wedding celebration he noticed that the volumes of wine barrels were estimated by means of a rod slipped in diagonally through the bunghole, which prompted him to wonder how that could work. The result was a study of the volumes of solids of revolution, *Nova Stereometria Doliorum* (New stereometry of wine barrels, 1615).

Kepler-Poinsot solids

The four regular nonconvex polyhedra that exist in addition to the five regular convex polyhedra known as the **Platonic solid**s. As with the Platonic solids, the Kepler-Poinsot solids have identical regular **polygon**s for all their faces, and the same number of faces meet at each vertex. What is new is that we allow for a notion of "going around twice," which results in faces that intersect each other. In the *great stellated dodecahedron* and the *small stellated dodecahedron,* the faces are pentagrams (five-pointed stars). The center of each pentagram is hidden inside the polyhedron. These two polyhedra were described by Johannes **Kepler** in 1619, and he deserves credit for first understanding them mathematically, though a sixteenth-century drawing by the Nuremberg goldsmith Wentzel Jamnitzer (1508–1585) is very similar to the former and a fifteenth-century mosaic attributed to the Florentine artist Paolo Uccello (1397–1475) illustrates the latter. The *great icosahedron* and *great dodecahedron* were described by Louis **Poinsot** in 1809, though Jamnitzer made a picture of the great dodecahedron in 1568. In these two, the faces (20 triangles and 12 pentagons, respectively) that meet at each vertex "go around twice" and intersect each other, in a manner that is

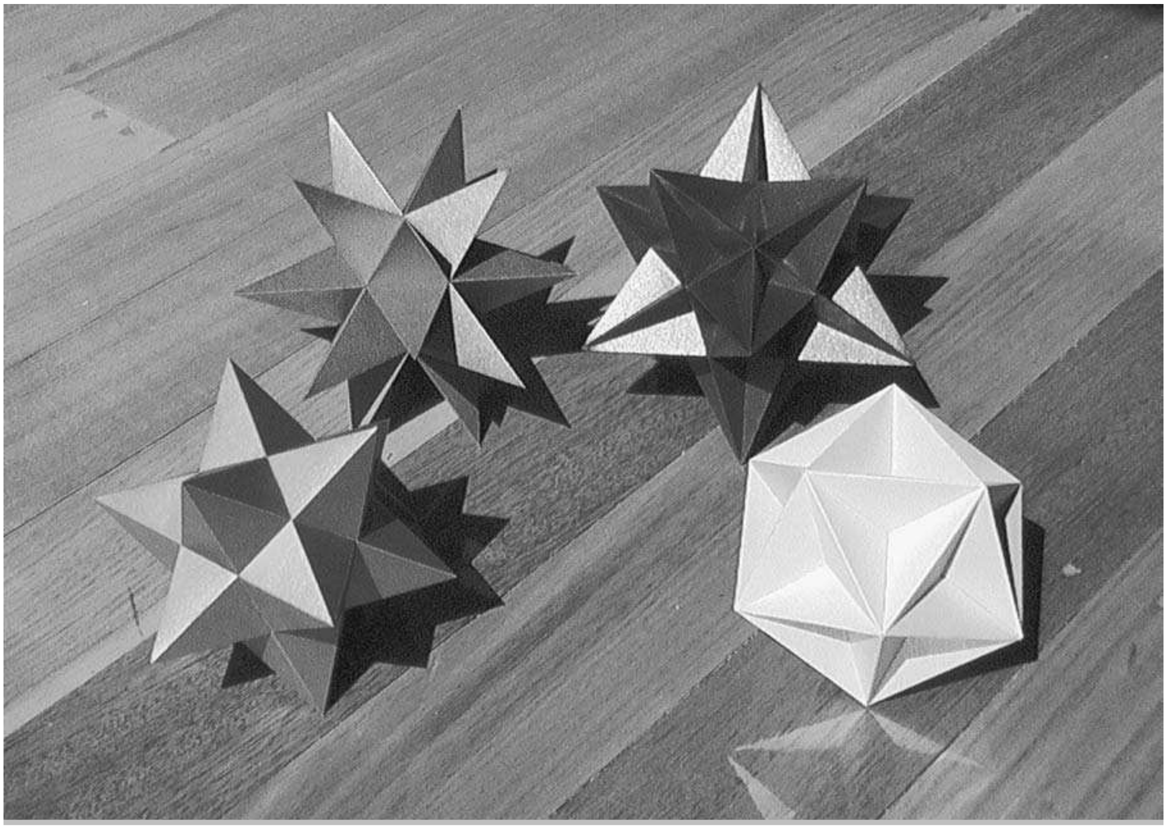

Kepler-Poinsot solids From left to right: the small stellated dodecahedron, great stellated dodecahedron, great icosahedron, and great dodecahedron. *Robert Webb, www.software3d.com; created using Webb's Stella program* the three-dimensional analog to what happens in two dimensions with a **pentagram**. Together, the Platonic solids and the Kepler-Poinsot polyhedra form the set of nine *regular* polyhedra. Augustin **Cauchy** first proved that no other polyhedra can exist with identical regular faces and identical regular vertices.

Kepler's conjecture

No **packing** of spheres of the same radius in three dimensions has a density greater than the face-centered (hexagonal) cubic packing. This claim was first published by Johannes **Kepler** in his monograph *The Six-Cornered Snowflake* (1611), a treatise inspired by his correspondence with Thomas Harriot (see **cannonball problem**). In his slender essay, Kepler asserted that face-centered cubic packing–the kind greengrocers use to stack oranges–is "the tightest possible, so that in no other arrangement could more pellets be stuffed into the same container." The question of whether Kepler's conjecture is right or not has became known, not surprisingly, as *Kepler's problem.* In the nineteenth century, Carl **Gauss** proved that face-centered cubic packing is the densest arrangement in which the centers of the spheres form a regular lattice, but he left open the question of whether an irregular stacking of spheres might be still denser. In 1953, László Tóth reduced the Kepler conjecture to an enormous calculation that involved specific cases, and later suggested that computers might be helpful for solving the problem. This was the approach taken by Thomas Hales, a mathematician at the University of Michigan at Ann Arbor, and which led him, in 1998, to claim that he had proved Kepler was right all along. Hales's proof of Kepler's conjecture remains controversial simply because of the length of the computer calculations involved and the difficulty of verifying them. The casebook on this mystery remains open. See also **monster group**.

Khayyam, Omar (1044–1123)

An outstanding Persian mathematician and astronomer, whose full name was Abu al-Fath Omar ben Ibrahim al-Khayyam, or "tent-maker"–possibly his father's profession. His work on algebra was known throughout Europe in the Middle Ages. He made several important contributions: he discovered a geometrical method to solve **cubic** equations by intersecting a **parabola** with a circle, discussed what would become known as **Pascal's triangle**, and asked if a ratio could be regarded as a number. He is best known, however, as a poet. In 1859, Edward Fitzgerald (1809–1883) translated Khayyam's *Rubaiyat*–a popular collection 600 short four-line poems. Rosehips from Khayyam's tomb were germinated at Kew Gardens, London, and planted on Fitzgerald's tomb in St. Michael's Churchyard, Boulge, Suffolk, in 1893; the original plant has died, but its descendents continue to bloom.

Khintchine's constant

One of the most remarkable, yet poorly understood, constants in mathematics, which captures, in a fascinating way, the behavior of *almost all* **real number**s. Pick a real number at random and write it down as a **continued fraction**. Almost certainly, the **geometric mean** of the terms in this continued fraction will be Khintchine's constant, which has the value 2.685452 . . . . It is important to say "almost certainly" because there are some real numbers including all **rational number**s, roots of third-order **polynomial**s, and certain other classes of number, whose continued fractions would give a different result. These exceptions, however, form a tiny minority of all real numbers.

kings problem

The problem of determining how many nonattacking kings, $K(n)$, can be placed on an $n \times n$ chessboard. For $n = 8$, the solution is 16. The general solution is $K(n) = \frac{1}{4}n^2$ if n is even and $K(n) = \frac{1}{4}(n + 1)^2$ if n is odd. The minimum number of kings needed to attack or occupy all squares on an 8×8 chessboard is nine.

kinship puzzles

Like **age puzzles and tricks**, problems to do with how family members are related go back many centuries. Some of these can be fiendishly convoluted, especially if incestual pairings are allowed (in the puzzle!). Sketching a genealogical tree is sometimes helpful. The following all involve legitimate ties.

PUZZLES

1. Brothers and sisters have I none, but that man's father is my father's son. Who is that man? (This is one of the oldest problems of the kinship variety.)
2. What is the simplest way in which two people can be an uncle to each other? (From Dudeney's *A Puzzle-Mine.*)
3. A certain family party consisted of 1 grandfather, 1 grandmother, 2 fathers, 2 mothers, 4 children, 3 grandchildren, 1 brother, 2 sisters, 2 sons, 2 daughters, 1 father-in-law, 1 mother-in-law, and 1 daughter-in-law. Twenty-three people, you will say. No; there were only seven people present. Can you show how this might be? (From *Amusements in Mathematics.*)

Solutions begin on page 369.

Kirkman, Thomas Penyngton (1806–1895)

An English rector and mathematician who did important work in **combinatorics**. His **schoolgirls problem**, which he posed in 1850, led to a more general study of certain ways of combining objects. He also explored the possibility of finding a route around the edges of a given **polyhedron** that passes through each vertex (corner) of the shape once and only once. His rather complex and ultimately unprovable ideas were later picked up and refined successfully by William **Hamilton**, but not be-fore Kirkman had managed to supply a general proof of the fact that "if a polyhedron has an odd number of vertices and each face has an even number of edges, then there is no circuit which passes through all the vertices." This result introduces the concept of a *bipartite graph*–a **graph** that can be divided into two separate sets of vertices such that every edge in one set joins a vertex.

kite

A **quadrilateral**, with two pairs of congruent adjacent sides, that is named after one of the traditional forms of toy kite. The toy itself probably draws its name from the bird commonly called a kite, or kyte, in England. The old English form of the word, cyte, may in turn derive from an early German name for an owl. A nonconvex kite is often called a *dart,* a term used by Roger **Penrose** in his proof on a nonperiodic tiling of the plane (see also **Penrose tiling**). **Proclus** referred to this shape as a "four-sided triangle" and spoke of it as a geometric paradox. John **Conway** pointed out that there is no special name for a quadrilateral that has two pairs of equal sides that, unlike in the case of a kite, are not parallel sides. He proposed the name *strombus* for such a figure, from the Greek for a spinning top.

Klein, Christian Felix (1849–1925)

A German mathematician noted for his work on **non-Euclidean geometry**, the connections between geometry and **group** theory, and the theory of functions. His Erlangen Programm (1872) for unifying the diverse forms of geometry through the study of equivalence in transformation groups was profoundly influential, especially in the United States, for over 50 years. In his *Lectures on the Icosahedron and the Solution of Equations of the Fifth Degree* (1884, tr. 1888) he showed how the rotation groups of regular solids could be applied to the solution of difficult algebraic problems. Klein was a professor of mathematics successively at the University of Erlangen, the Technical Institute in Munich, and the universities of Leipzig and Göttingen, and was a prolific writer and lecturer on the theory, history, and teaching of mathematics. See also **Klein bottle**.

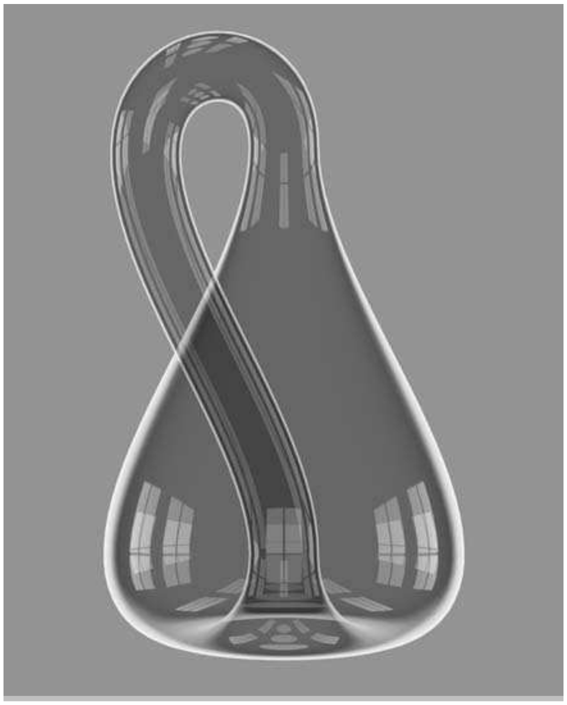

Klein bottle The famous container, with zero volume, rendered in the improbable form of a soap bubble. © *John M. Sullivan, University of Illinois and TU Berlin*

Klein bottle

> *A mathematician named Klein,*
> *Thought the Möbius band was divine,*
> *He said, "If you glue,*
> *The edges of two,*
> *You'll get a weird bottle like mine."*
>
> –Leo Moser

Take a rectangle and join one pair of opposite sides to make a cylinder. Now join the other pair with a half-twist. The result is a Klein bottle. Sound easy? It is if you have access to a **fourth dimension** because that's what is needed to carry out the second step and to allow the surface to pass through itself without a hole. A true Klein bottle is a four-dimensional object. It was discovered in 1882 by Felix **Klein** when he imagined, as in the limerick, joining two **Möbius band**s together to create a single-sided bottle with no boundary.

An ordinary (three-dimensional) bottle has a crease or fold around the opening where the inside and outside of the bottle meet. A sphere doesn't have this crease or fold, but it has no opening. A Klein bottle has an opening but no crease: like a Möbius band, it is a continuous one-sided structure. Because it has no crease or fold, there's no verifiable definition of where its inside and outside begin. Therefore, the volume of a Klein bottle is considered to be zero, and the bottle has no real contents–except itself! As the joke goes: "In topological hell the beer is packed in Klein bottles." Take a coin, slide it across the surface of a Klein bottle until it returns to its starting point, and the coin, as if by magic, will be flipped over. This is because, unlike a sphere or a regular bottle, a Klein bottle is nonorientable.

Although a Klein bottle can't be "embedded" (that is, fully realized) in three dimensions, it can be "immersed" in three dimensions. Immersion is what happens when a higher dimensional object cuts through a lower dimensional one, producing a cross section. When a sphere is immersed in a plane, for example, it produces a circle. The computer-generated soap bubble structure shown here could be created by stretching the neck of a bottle through its side and joining its end to a hole in the base. Except at the side-connection (the nexus), this properly shows the shape of a four-dimensional Klein bottle. Just as the picture here is of a three-dimensional Klein bottle immersion, so an immersion in real life is like a drawing of the true four-dimensional bottle.

knights problem

To find the maximum number of knights that can be placed on an $n \times n$ chessboard in such a way that no two pieces are attacking each other. For a standard 8×8 chessboard, the answer is 32 (all on white squares or all on black). In the general, the solution is $\frac{1}{2} n^2$ if n is even and $\frac{1}{2} (n^2 + 1)$ if n is odd, giving the sequence 1, 4, 5, 8, 13, 18, ...

knight's tour

A classic **chess** puzzle: to find a sequence of moves by which a knight can visit each square of a chessboard exactly once. If the final position is a knight's move away from the first position, the tour is said to be *closed* or *reentrant.* The earliest recorded solution for a standard 8×8 chessboard was given by Abraham **de Moivre**; the earliest known reentrant solution came from the French mathematician Adrien-Marie Legendre (1752–1833). Not to be outdone Leonhard **Euler** found a reentrant tour that visits two halves of the board in turn. The problem can be generalized to an $n \times n$ board, with some surprising results; for example, a reentrant tour is not possible on a 4×4 board.

A knight's tour is called a **magic tour** if the resulting arrangement of numbers forms a **magic square**, and a *semimagic tour* if the resulting arrangement of numbers is a **semi-magic square**. It has long been known that magic knight's tours aren't possible on $n \times n$ boards if n is odd. It was also known that such tours are possible for all boards of size $4k \times 4k$ for $k > 2$. However, while a number of semi-magic knight's tours were known on the usual 8×8 chessboard, it was *not* known if any fully magic tours existed on the 8×8 board. This longstanding open problem was finally settled in the negative by an exhaustive computer enumeration of all possibilities. The software for the computation was written by J. C. Meyrignac, and a Web site was established by Guenter Stertenbrink to distribute and collect results for all possible tours. After 61.40 CPU-days, corresponding to 138.25 days of computation at 1 GHz, the project was completed on August 5, 2003. As well as netting a total of 140 distinct semi-magic knight's tours, the computation showed for the first time that no 8×8 magic knight's tour is possible.

More magical and mysterious tours can be conducted on boards on the surfaces of cubes, cylinders, toruses, and multidimensional shapes, such as hypercubes. See also **Hamilton path**.

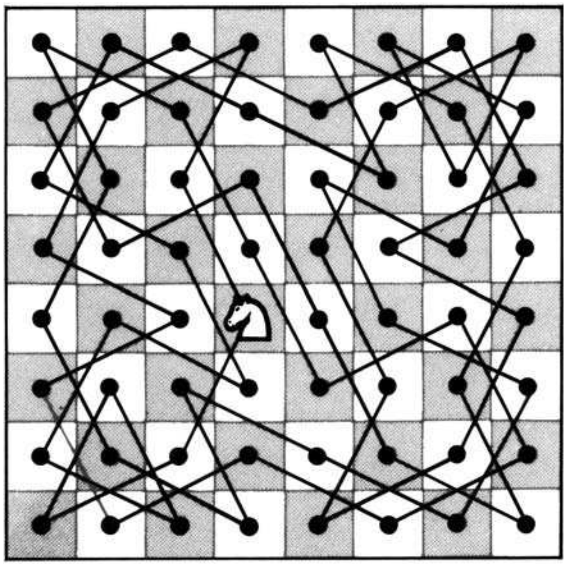

knight's tour After visiting every square of the chessboard exactly once, the knight returns to its starting point.

knot

> *Never cut what you can untie.*
>
> –Joseph Joubert

A **closed** curve in three dimensions. The two simplest nontrivial knots are the trefoil knot, whose picture has three crossings, and the figure-eight knot, whose picture has four. To date, more than 1.7 million nonequivalent knots with pictures of 16 or fewer crossings have been identified. The mathematical theory of knots was born out of attempts to model the atom. Near the end of the nineteenth century, William Thomson (Lord Kelvin) suggested that different atoms were actually different knots tied in the ether that was believed to permeate all of space. Physicists and mathematicians set to work making a table of distinct knots, believing they were making a table of the elements. A pioneer in this effort, alongside Thomson, was Peter **Tait**. By the time the theory of the ether vanished into thin air, *knot theory* was firmly tied into mainstream mathematics. It blossomed with the development of **topology** and eventually led to important applications in DNA research and molecular biology. Today it is one of the most active areas of mathematical research. See also **tie knot**, **loop**, and **braid**.

Knuth's up-arrow notation

A notation for **large number**s developed by the American mathematician Donald Knuth in 1976. A single up-arrow ($\uparrow$) is the same as exponentiation:

$$m \uparrow n = m \times m \times \ldots \times m \text{ (n terms)} = m^n.$$

Two up-arrows together represent a **power tower**:

$$m \uparrow\uparrow n = \left. m^{m^{\cdot^{\cdot^{\cdot^{m}}}}} \right\} n.$$

(a tower of height n), which is the same as the operation known as *hyper4* or *tetration.* This can very rapidly generate huge numbers. For example:

$2 \uparrow\uparrow 2 = 2 \uparrow 2 = 4$
$2 \uparrow\uparrow 3 = 2 \uparrow 2 \uparrow 2 = 2 \uparrow 4 = 16$
$2 \uparrow\uparrow 4 = 2 \uparrow 2 \uparrow 2 \uparrow 2 = 2 \uparrow 16 = 65{,}536$
$3 \uparrow\uparrow 2 = 3 \uparrow 3 = 27$
$3 \uparrow\uparrow 3 = 3 \uparrow 3 \uparrow 3 = 3 \uparrow 27 = 7{,}625{,}597{,}484{,}987$
$3 \uparrow\uparrow 4 = 3 \uparrow 3 \uparrow 3 \uparrow 3 = 3 \uparrow 3 \uparrow 27 = 3^{7625597484987}$

Three up-arrows together represent a still more vastly powerful operator, equivalent to *hyper5* or *pentation,* or a power tower of power towers:

$$m \uparrow\uparrow\uparrow n = m \uparrow\uparrow m \uparrow\uparrow \ldots \uparrow\uparrow m \text{ (n terms)}$$

For example:

$2 \uparrow\uparrow\uparrow 2 = 2 \uparrow\uparrow 2 = 4$
$2 \uparrow\uparrow\uparrow 3 = 2 \uparrow\uparrow 2 \uparrow\uparrow 2 = 2 \uparrow\uparrow 4 = 65{,}536$
$2 \uparrow\uparrow\uparrow 4 = 2 \uparrow\uparrow 2 \uparrow\uparrow 2 \uparrow\uparrow 2 = 2 \uparrow\uparrow 65{,}536 = 2 \uparrow 2 \uparrow \ldots \uparrow 2$ (65,536 terms)
$3 \uparrow\uparrow\uparrow 2 = 3 \uparrow\uparrow 3 = 7{,}625{,}597{,}484{,}987$
$3 \uparrow\uparrow\uparrow 3 = 3 \uparrow\uparrow 3 \uparrow\uparrow 3 = 3 \uparrow\uparrow 7{,}625{,}597{,}484{,}987 = 3 \uparrow 3 \uparrow \ldots \uparrow 3$ (a power tower 7,625,597,484,987 layers high)
$3 \uparrow\uparrow\uparrow 4 = 3 \uparrow\uparrow 3 \uparrow\uparrow 3 \uparrow\uparrow 3 = 3 \uparrow\uparrow 3 \uparrow\uparrow 7{,}625{,}597{,}484{,}987 = 3 \uparrow\uparrow 3 \uparrow \ldots \uparrow 3$ (a tower $3 \uparrow\uparrow 7{,}625{,}597{,}484{,}987$ layers high)

Similarly,

$$m \uparrow\uparrow\uparrow\uparrow n = m \uparrow\uparrow\uparrow m \uparrow\uparrow\uparrow \ldots \uparrow\uparrow\uparrow m \text{ (n terms)}$$

so that, for example:

$2 \uparrow\uparrow\uparrow\uparrow 2 = 2 \uparrow\uparrow\uparrow 2 = 4$
$2 \uparrow\uparrow\uparrow\uparrow 3 = 2 \uparrow\uparrow\uparrow 2 \uparrow\uparrow\uparrow 2 = 2 \uparrow\uparrow\uparrow 4 = 2 \uparrow 2 \uparrow \ldots \uparrow 2$ (65,536 terms)
$2 \uparrow\uparrow\uparrow\uparrow 4 = 2 \uparrow\uparrow\uparrow 2 \uparrow\uparrow\uparrow 2 \uparrow\uparrow\uparrow 2 = 2 \uparrow\uparrow\uparrow 2 \uparrow 2 \uparrow \ldots \uparrow 2$ (65,536 terms)
$3 \uparrow\uparrow\uparrow\uparrow 2 = 3 \uparrow\uparrow\uparrow 3 = 3 \uparrow 3 \uparrow \ldots \uparrow 3$ (7,625,597,484,987 terms)
$3 \uparrow\uparrow\uparrow\uparrow 3 = 3 \uparrow\uparrow\uparrow 3 \uparrow\uparrow\uparrow 3 = 3 \uparrow\uparrow\uparrow 3 \uparrow 3 \uparrow \ldots \uparrow 3$ (7,625,597,484,987 terms)
$= 3 \uparrow\uparrow 3 \uparrow 3 \uparrow \ldots \uparrow 3$ (3 $\uparrow 3 \uparrow \ldots \uparrow 3$ (7,625,597,484,987 terms)

Even up-arrow notation becomes cumbersome, however, when one is faced with staggeringly large numbers such as **Graham's number**. For such cases, more extensive systems such as **Conway's chained arrow notation** or *Steinhaus-Moser notation* are better suited. See also the **Ackermann function**, to which up-arrow notation is closely related.

Koch snowflake

One of the most symmetric and easy to understand **fractals**; it is named after the Swedish mathematician Helge von Koch (1870–1924), who first described it in 1906. To make the snowflake, start with a straight line and split it into three equal parts. Replace the middle part with two lines, both of the same length as the first three, creating an

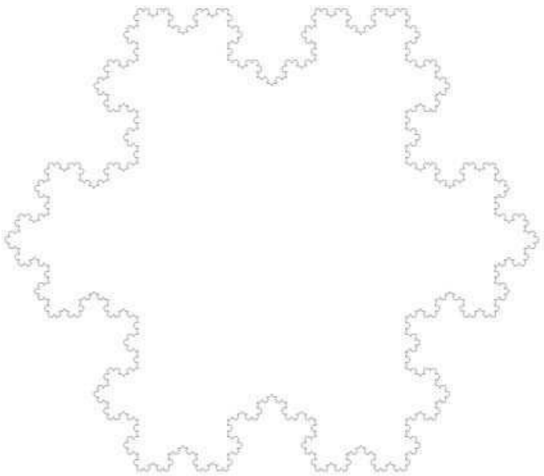

Koch snowflake This fractal shape emerges from continually sprouting equilateral triangles. *Xah Lee, www.xahlee.org* equilateral triangle missing the bottom line. The shape now consists of four straight lines with the same length. For each of these lines, repeat this process, then continue the transformation indefinitely. The Koch snowflake has infinite length because each iteration increases the length of a line segment one-third, and the iterations go on forever.

The same kind of process can be applied to a tetrahedron. Take a regular tetrahedron (all side lengths the same) and glue to each of its faces a smaller regular tetrahedron. (Each smaller tetrahedron is scaled down by a factor of ½ from the larger one, and placed on each face in an inverted fashion, so that it divides the face into four equilateral triangles and covers the center one.) Then iterate this process. Intuition suggests that the end product might be a very strange-looking jagged object. But in fact, in the limit, as the number of iterations tends to infinity, the result is a perfect cube! The cube has side length $t/\sqrt{2}$, where t is the length of one of the edges of the regular tetrahedron you started with. Variations on the flat Koch snowflake include the *exterior snowflake,* the *Koch antisnowflake,* and the *flowsnake* curves.

Kolmogorov, Andrei Nikolaievich (1903–1987)

A Russian mathematician and physicist who advanced the foundations of **probability theory**, the **algorithm**ic theory of **random**ness, and made crucial contributions to the foundations of statistical mechanics, **stochastic** processes, **information theory**, fluid mechanics, and nonlinear dynamics. All of these areas, and their interrelationships, underlie **complex system**s as they are studied today. His work on reformulating probability started with a 1933 paper in which he built up probability theory in a rigorous way from fundamental axioms, similar to Euclid's treatment of geometry. Kolmogorov went on to study the motion of the planets and turbulent fluid flows, publishing two papers in 1941 on turbulence that even today are of fundamental importance.

In 1954 he developed his work on dynamical systems in relation to planetary motion, thus demonstrating the vital role of probability theory in physics and reopening the study of apparent randomness in deterministic systems, along the lines originally conceived by Henri **Poincaré**. In 1965 he introduced the algorithmic theory of randomness via a measure of complexity, now referred to as *Kolmogorov complexity.* According to Kolmogorov, the complexity of an object is the length of the shortest computer program that can reproduce the object. Random objects, in his view, were their own shortest description, whereas periodic sequences have low Kolmogorov complexity, given by the length of the smallest repeating "template" sequence they contain. Kolmogorov's notion of complexity is a measure of randomness, one that is closely related to Claude **Shannon**'s entropy rate of an information source.

Königsberg bridge problem

See **bridges of Königsberg**.

Kronecker, Leopold (1823–1891)

A German mathematician and pioneer in the field of **algebraic number**s who formulated the relationship between the theory of numbers, the theory of equations, and **elliptic function**s. He acquired a passion for **number theory** from Ernst Kummer, his instructor at the Liegnitz Gymnasium. Kronecker, who made a fortune in business before returning to his academic studies, claimed that mathematical argumentation should be based only on integers and finite procedures. He was one of Georg **Cantor**'s sternest critics and refused to accept the validity of **Weierstrass's nondifferentiable function.**

L

labyrinth

See **maze**.

Lagrange, Joseph Louis (1736–1813)

An Italian-born French mathematician who made important contributions to **number theory** and to classical and celestial mechanics. By his mid-twenties he was recognized as one of the greatest living mathematicians because of his papers on wave propagation and the maxima and minima of curves. His prodigious output included his textbook *Mécanique Analytique* (Analytical mechanics, 1788), the basis for all later work in this field. His remarkable discoveries include the *Lagrangian,* a differential operator characterizing a system's physical state, and the *Lagrangian points,* points in space where a small body in the gravitational fields of two large ones remains relatively stable. Under Napoleon, Lagrange was made both a senator and a count; he is buried in the Panthéon.

lambda calculus

A model of computation that is capable of **universal computation**. The Lisp programming language was inspired by lambda calculus.

Lamé curve

Any of a family of curves related to the **ellipse** and that was first recognized and studied in 1818 by the French physicist and mathematician Gabriel Lamé (1795–1870). The formula for the Lamé curve family is a generalization of the equation for an ellipse ($|x/a|^2 + |y/b|^2 = 1$), namely:

$$|x/a|^n + |y/b|^n = 1,$$

where n is any real number. When $n = 0$, the curve reduces to a pair of crossed lines. As n increases, the curve changes from a curved star shape to a rectangle, with diagonals a and b, when $n = 1$. The special case when $n = 2/3$ corresponds to the **astroid**. Between $n = 1$ and $n = 2$ the curve turns from a curved rectangle to an ellipse (or a circle when both a and b are 1). For values of n greater than 2, Lamé curves are known as **superellipses**.

lamination

A decoration of a **manifold** in which some subset is partitioned into sheets of some lower dimension, and the sheets are locally parallel. It may or may not be possible to fill the gaps in a lamination to make a **foliation**.

Langley's adventitious angles

A seemingly simple problem first posed in 1922 by E. M. Langley[195] in connection with an isosceles triangle. In its original form, it is stated as follows: ABC is an isosceles triangle. B = C = 80°. CF at 30° to AC cuts AB in F. BE at 20° to AB cuts AC in E. Prove angle BEF = 30°. (No mention is made of D. Perhaps it is at the intersection of BE and CF.) A number of solutions appeared shortly after, including this one given by J. W. Mercer: Draw BG at 20° to BC, cutting CA in G. Then angle GBF = 60° and angles BGC and BCG are 80°. So BC = BG. Also angle BCF = angle BFC = 50°, so BF = BG and triangle BFG is equilateral. But angle GBE = 40° = angle BEG, so BG = GE = GF. And angle FGE = 40°, hence GEF = 70° and BEF = 30°.

Langton's ant

A type of **cellular automaton**, or a simple form of **artificial life**, named after its designer, Christopher Langton. The ant lives on an infinitely large chessboard, each square of which can be either black or white. Two pieces of information are associated with this digital insect: the

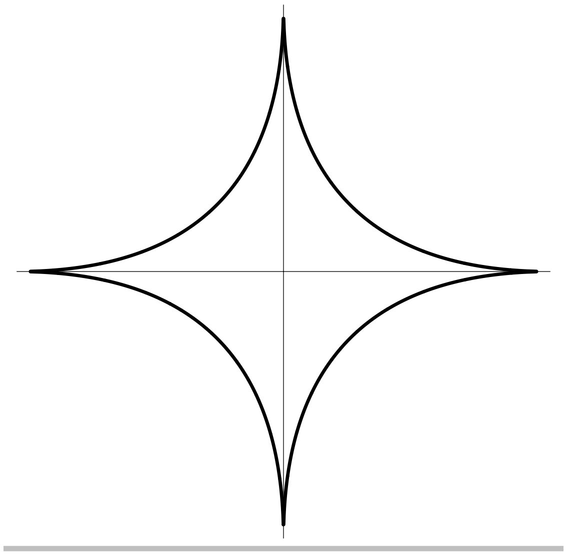

Lamé curve One example from the family of Lamé curves. © *Jan Wassenaar, www.2dcurves.com*

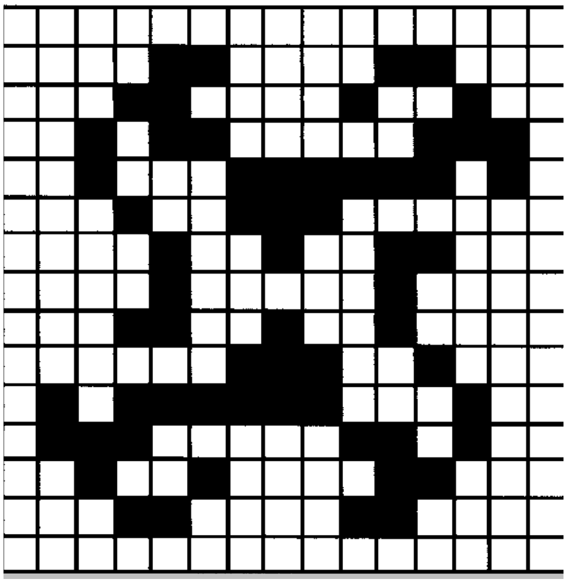

Langton's ant After 368 steps, a symmetrical pattern emerges.

direction that it's facing, and the state of the square that it currently occupies. Three simple rules govern the ant's behavior: (1) If it's on a black square, it turns left. (2) If it's on a white square, it turns right. (3) As it moves to the next square, the one it was on reverses color. Interest in Langton's ant stems from the fact that despite being a completely determined system governed by extremely simple rules, the patterns it produces are fantastically rich and complex. For the first 10,000 moves or so, the ant meanders around, building and then unbuilding structures with little pattern to them. Then, near the end of this chaotic phase, the ant begins to construct a diagonal highway off toward one edge of the board. In fact, this pattern stems from a sequence of 104 moves that, once started, will go on forever. In the language of **chaos** theory, the pattern is a *stable attractor* for the system. Remarkably, no matter what the initial arrangement of squares—even if the white and black squares are set up randomly—the ant will end up building a highway. The ant can be allowed to wrap around the edges of a finite board, thus allowing it to intersect its own path, and it will *still* end up building the highway. Are there any initial states that don't lead to the diagonal-road-building loop? No exceptions have been found from experiments—but proving it is another matter. Most mathematicians believe there is no general analytical method of predicting the position of the ant, or of any such chaotic system, after any given number of moves. Its behavior can't be reduced to the rules that govern it. In this sense, Langton's ant is a simple demonstration of the undecidability of the **halting problem**. The British mathematician Ian **Stewart** and biologist Jack Cohen, in their book *Figments of Reality*,[321] go a step further and use Langton's ant as an analog of an essential stage in the evolution of complex systems such as life: a stage in which the existence of chaotic behavior contains the potential for the spontaneous emergence of unpredictable forms of order.

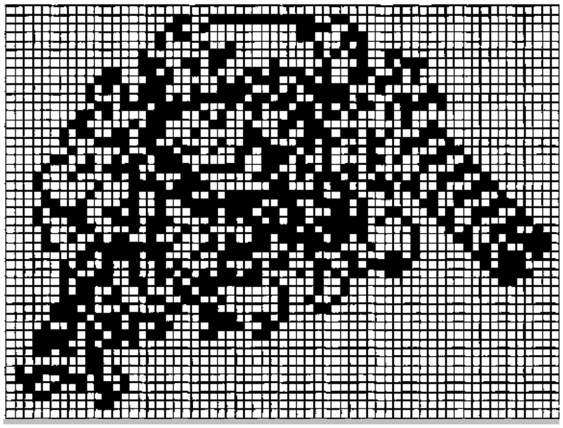

Langton's ant After 10,647 steps the "highway" is under construction.

Laplace, Pierre Simon (1749–1827)

A French mathematician and astronomer who was heavily involved with the development of celestial mechanics. He made an early impact by solving a complex problem of mutual gravitation that had eluded both Leonhard **Euler** and Joseph **Lagrange**. Laplace was among the most influential scientists of his time and was called the Newton of France for his study of and contributions to the understanding of the stability of the solar system. Laplace generalized the laws of mechanics for their application to the motion and properties of the heavenly bodies. He is also famous for his great treatises *Mécanique céleste* (Celestial mechanics, 1799–1825) and *Théorie analytique des probabilités* (Analytical theory of probabilities, 1812) which were advanced in large part by the mathematical techniques that Laplace developed early in his life.

large number

> *Bigger than the biggest thing ever and then some. Much bigger than that in fact, really amazingly immense, a totally stunning size, real "wow, that's big!" time. . . . Gigantic multiplied by colossal multiplied by staggeringly huge is the sort of concept we're trying to get across here.*
>
> —Douglas Adams, *The Restaurant at the End of the Universe*

Making, naming, and representing very large numbers is itself a big problem. A simple way to start is by adding zeros: 10; 100; 1,000; 10,000; . . . , 1,000,000; . . . . But this quickly gets tedious and exponentiation becomes a more attractive option: 10^1, 10^2, 10^3, . . . , 10^6, ... Naming the various powers of 10 follows a regular pattern of prefixes. In the United States, 10^3 is one thousand, 10^6 is one million, 10^9 is one billion, 10^{12} is one trillion, 10^{15} is one quadrillion, and so on. The "-illion" root kicks in at the sixth power of 10 prefixed by "m" for "mono"; then, every jump of three powers (factor of a thousand) comes the next prefix. Put another way, the U.S. name for 10^{3n} uses the Latinized prefix for $n - 1$. One centillion, which is the largest number with a single-word name in English, is 10^{303}. Elsewhere in the world, "billion," "trillion," and so on, can mean things other than they do in the U.S. system. A British billion, for example, is one million million, or 10^{12}, while the now largely obsolete term *milliard* was used for a thousand million. However, the American forms have become fairly standard internationally and are used without qualification in this book. It's also worth noting that, while "quadrillion," "quintillion," and so forth, are perfectly valid terms, "one thousand trillion," "one million trillion," and so forth, are generally preferred.

Exponentiating quite small numbers seems at first to be a pretty economical way of making and writing large numbers: 10^{30}, for example, is a highly effective shorthand for 1,000,000,000,000,000,000,000,000,000,000. But this method runs out of steam as the numbers get bigger and bigger. Take, for instance, the **googol** and the googolplex. One googol is the unofficial name for 10^{100}, or 1 followed by 100 zeros. This innocuous-looking number is larger than the number of atoms in the universe. What happens, therefore, if we want to represent the number that is 1 followed by *a googol number of zeros?* One way would be to write "1 followed by a googol number of zeroes"! But this is cheating because it couldn't be generalized without first giving a proper name to a fantastically large number of numbers. A better solution is to exponentiate using large numbers. Thus, 1 googolplex = $10^{\text{googol}} = 10^{10^{100}}$. This is the beginning of a **power tower**.

Are numbers as large as the googol, not to mention the googolplex, of any practical importance? Science has given us such colossi as the **Avogadro constant** (the number of molecules in a sample whose weight in grams equals its molecular weight) = 6.023×10^{23}, the **Eddington number** (astrophysicist Arthur Eddington's best estimate of the number of protons in the universe) = 1.575×10^{79}, and the Supermassive Black Hole Evaporation Time = 10^{100} years (or thereabouts)–which brings us to the level of the googol. But there's nothing known or that can be reasonably conjectured in the "real" world of physics that goes much beyond this.

Science fiction can carry us a bit further. In *The Hitchhiker's Guide to the Galaxy*[5] by Douglas Adams appears one of the largest numbers ever used in a work of fiction: 2^{260199}. These are the odds quoted against the characters Arthur Dent and Ford Prefect being rescued by a passing spaceship just after having been thrown out of an airlock. As it happens, they *are* rescued by a spaceship powered by the "infinite improbability drive." By contrast, some special numbers in mathematics make even the googolplex look tiny. **Skewes' number**, $10^{10^{10^{34}}}$, was long held up as an example of a googolplex-dwarfing number that nevertheless served a bonafide purpose in mathematics. However, even this seemingly immense integer is made to look ridiculously small by the likes of more recently described numbers, such as **Graham's number**, the *Mega*, and the *Moser*, which are so utterly vast that it takes a page or two simply to describe the various special notations used to represent them.

Just as writing out a number in full, or in place-value notation, becomes unwieldy with numbers as big as a googol, so exponentiation, in turn, endangers the world's forests if it tries to take on seriously large numbers. A more effective shorthand is *tetration–tetra* (from the Greek meaning "four") because it is the fourth dyadic operation in the series: addition, multiplication, exponentiation, tetration. Dyadic means that two numbers, or arguments, are involved in the operation. Multiplication is repeated addition (e.g., $2 \times 3 = 2 + 2 + 2$), exponentiation is repeated multiplication (e.g., $2^3 = 2 \times 2 \times 2$), and tetration is repeated exponentiation. For example, 2 tetrated to 3, represented as ${}^{3}2$, is $2^{2^2} = 2^4 = 16$; 2 tetrated to 4, or ${}^{4}2$, is $2^{2^{2^2}} = 2^{16} = 65{,}536$; and 2 tetrated to 5, or ${}^{5}2$, is $2^{2^{2^{2^2}}} = 2^{65{,}536}$ = something too big to write out in full. Tetration goes by various other names including *superpower, superdegree,* and, the one used most commonly in mathematical circles and also here, *hyper4.*

Just as the exponentiation of two numbers, a and b, is represented as a^b and defined as $a \times a \times \ldots \times a$ (b terms), the hyper4 of a and b is represented as $a^{(4)}b$ and defined as $a^{a^{a \ldots a}}$ (a power tower with b levels). Alternatively, the hyper4 operator can be represented in **Knuth's up-arrow notation** as $a \uparrow\uparrow b$. Continuing this pattern:

$$\text{hyper5 of } a \text{ and } b = a^{(5)}b = a^{(4)}a^{(4)} \ldots a^{(4)} = a \uparrow\uparrow\uparrow b$$
$$\text{hyper6 of } a \text{ and } b = a^{(6)}b = a^{(5)}a^{(5)} \ldots a^{(5)} = a \uparrow\uparrow\uparrow\uparrow b$$
$$\text{hyper7 of } a \text{ and } b = a^{(7)}b = a^{(6)}a^{(6)} \ldots a^{(6)} = a \uparrow\uparrow\uparrow\uparrow\uparrow b$$

To get some idea of the potency of this kind of representation, consider the sequence

1

10

10,000,000,000

10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000 (100 zeroes)

. . .

Notice how big even the fourth term is? The seemingly innocuous-looking number 5 ↑↑↑↑↑ 5 (or $5^{(7)}5$) is so huge that it would be around the 100,000,000,000,000,000th (one hundred thousand trillionth) term of this sequence!

Since the dyadic operators discussed previously form a pattern, they can be telescoped into a single *triadic* operator that has three arguments. This can be defined as:

$$\text{hy}(a, n, b) = \begin{cases} 1 + b \text{ for } n = 0 \\ a + b \text{ for } n = 1 \\ a \times b \text{ for } n = 2 \\ a \uparrow b \text{ for } n = 3 \\ a \uparrow \text{hy}(a, 4, b - 1) \text{ for } n = 4 \\ \text{hy}(a, n - 1, \text{hy}(a, n, b - 1)) \text{ for } n > 4 \\ a \text{ for } n > 1, b = 1 \end{cases}$$

Beyond hyper are other triadic operators capable of generating large numbers even faster. The **Ackermann function** and the *Steinhaus-Moser notation* are both equivalent to a triadic operator that is somewhat more powerful than hy(*a, n, b*). Similarly, **Conway's chained-arrow notation** marks an evolution of Knuth's symbolism. These various techniques and notations can produce immense *finite* numbers. But beyond any of these lie the many different kinds of **infinity**.

lateral inhibition illusion

See **Hermann grid illusion**.

Latin square

An $n \times n$ square grid, or matrix, whose entries consist of n symbols such that each symbol appears exactly once in each row and each column. The following are some examples:

$$\begin{array}{l} 1\,2 \\ 2\,1 \end{array} \quad \begin{array}{l} 1\,2\,3 \\ 2\,3\,1 \\ 3\,1\,2 \end{array} \quad \begin{array}{l} 1\,2\,3\,4 \\ 2\,3\,4\,1 \\ 3\,4\,1\,2 \\ 4\,1\,2\,3 \end{array} \quad \begin{array}{l} 1\,2\,3\,4 \\ 2\,1\,4\,3 \\ 3\,4\,1\,2 \\ 4\,3\,2\,1 \end{array} \quad \begin{array}{ccccc} \text{M} & \text{A} & \text{G} & \text{I} & \text{C} \\ \text{G} & \text{I} & \text{C} & \text{M} & \text{A} \\ \text{C} & \text{M} & \text{A} & \text{G} & \text{I} \\ \text{A} & \text{G} & \text{I} & \text{C} & \text{M} \\ \text{I} & \text{C} & \text{M} & \text{A} & \text{G} \end{array}$$

Latin squares have a long history, stretching back at least as far as medieval Islam (c. 1200), when they were used on amulets. Abu l'Abbas al Buni wrote about them and constructed, for example, 4 × 4 Latin squares using letters from a name of God. In his famous etching *Melancholia,* the fifteenth-century artist Albrecht **Dürer** portrays a 4 × 4 **magic square**, a relative of Latin squares, in the background. Other early references to them concern the problem of placing the 16 face cards of an ordinary deck of **cards** in the form of a square so that no row, column, or diagonal contains more than one card of each suit and each rank. Leonhard **Euler** began the systematic treatment of Latin squares in 1779 and posed a problem connected with them, known as the **thirty-six officers problem**, that wasn't solved until the beginning of the twentieth century. Arthur **Cayley** continued work on Latin squares, and in the 1930s the concept arose again in the guise of multiplication tables when the theory of quasigroups and loops began to be developed as a generalization of the **group** concept. Latin squares played an important role in the foundations of finite geometries, a subject that was also in development at that time. Also in the 1930s, a large application area for Latin squares was opened by R. A. Fisher who used them and other combinatorial structures in the design of statistical experiments.

lattice

A periodic arrangement of points such as the vertices (see **vertex**) of a **tiling** of space by cubes or the positions of atoms in a crystal. More technically, a discrete Abelian subgroup (see **Abelian group**) of an n-dimensional **vector space** that is not contained in an $(n - 1)$-dimensional vector space. Lattices play a central role in the theory of **Lie group**s, in **number theory**, in error-correcting codes, and many other areas of mathematics. See also **geoboard**.

lattice path

A sequence of points in a **lattice** such that each point differs from its predecessor by a finite list of allowed steps. **Random** lattice paths are an interesting model for the random motion of a particle and lattice paths are also important in enumerative **combinatorics**.

lattice point

A point with **integer** coordinates.

latus rectum

A **chord** of an **ellipse** that passes through a **focus** and is perpendicular to the major axis of the ellipse. Its plural is latera recta.

league

An archaic unit of traveling distance. The precise value varied, but was usually around 3 miles (4.8 kilometers).

least common multiple

The smallest **integer** that is an exact multiple of every number in a set of integers.

least upper bound

The smallest number that is larger than every member of a set of numbers.

Lebesgue, Henri Leon (1875–1941)

A French mathematician who introduced the modern definition of an **integral**. Lebesgue graduated from the École Normale Supériere and, from 1921, taught at the College de France. He and Emile **Borel** founded the modern theory of **function**s of a real variable, Lebesgue's great contribution being his new general definition of an integral (1902), which became known as the *Lebesgue integral* (see **integration**). This led to important advances in calculus, curve rectification, and trigonometric series, and, in Borel's hands, marked the start of **measure theory**. Although the Lebesgue integral was an example of the power of generalization, Lebesgue himself wasn't a fan of generalization and spent the rest of his life working on very specific problems, mostly in **analysis**.

Lebombo bone

One of the oldest mathematical artifacts known, a small piece of the fibula of a baboon, found near Border Cave in the Lebombo Mountains between South Africa and Swaziland. Discovered in the 1970s during excavations of Border Cave and dated about 35,000 B.C., the Lebombo bone is marked with 29 clearly defined notches. This suggests it may have been used as a lunar phase counter, in which case African women may have been the first mathematicians, because keeping track of menstrual cycles requires a lunar calendar. Certainly, the Lebombo bone resembles calendar sticks still used by Bushmen in Namibia. See also **Ishango bone**.

left-right reversal

See **mirror reversal problem**.

Leibniz, Gottfried Wilhelm (1646–1716)

A German philosopher, mathematician, and statesman who developed differential and integral **calculus** independently of Isaac **Newton**. He also invented a calculating machine, is considered a pioneer in mathematical logic, and proposed the metaphysical theory that we live in "the best of all possible worlds." In Leibniz's philosophical view, the universe is composed of countless conscious centers of spiritual force or energy known as *monads.* Leibniz talks about the "compossible" elements of any possible world–elements that allow a logically consistent structure. Though one of the finest minds of his age, Leibniz was not immune to blunders: he thought it just as easy to throw 12 with a pair of dice as to throw 11.

Leibniz harmonic triangle

A triangle of fractions that is related to the more famous **Pascal triangle** in a very simple way. Each row of the Leibniz harmonic triangle starts with the **reciprocal** of the row number (or the row number plus one depending on whether one starts counting from 1 or 0). Every entry is the sum of the two numbers just below it. The entries can thus be computed sequentially left to right and top to bottom using subtraction instead of addition.

1/1
1/2 1/2
1/3 1/6 1/3
1/4 1/12 1/12 1/4
1/5 1/20 1/30 1/20 1/5

lemma

A short auxiliary proposition used in the proof of a larger **theorem**.

lemniscate of Bernoulli

A curve "shaped like a figure 8, or a knot, or the bow of a ribbon" in the words of Jacob Bernoulli (see **Bernoulli family**) in an article published in 1694. Bernoulli named the curve "lemniscate" after the Greek *lemniskus* for a pendant ribbon (the type fastened to a victor's garland). It has the Cartesian equation

$$(x^2 + y^2)^2 = a^2(x^2 - y^2).$$

At the time he wrote his article, Bernoulli wasn't aware that the curve he was talking about was a special case of a **Cassinian oval**, which had been described by Cassini in 1680. The general properties of the lemniscate were established by Giovanni Fagnano (1715–1797) in 1750; Leonhard **Euler**'s investigations of the length of arc of the curve (1751) led to later work on **elliptic function**s. There is a relationship between the lemniscate and the rectangular **hyperbola**. If a **tangent** is drawn to the hyperbola and the perpendicular to the tangent is drawn through the origin, the point where the perpendicular meets the tangent is on the lemniscate. See also **hippopede.**

Leurechon, Jean (c. 1591–1670)

A French Jesuit priest and mathematician who wrote *Recreations mathématiques* (1624) under the pseudonym Henrik van Etten. Much of the mathematical content centers around Claude **Bachet**'s problems and may have been copied from it or some common source. The book also gives the earliest known description of the operation of an ear trumpet.

liar paradox

"This statement is false." What do we make of this statement (call it *S*)? If *S* is true, then *S* is false. On the other hand, if *S* is false, then it is true to say *S* is false; but, because the liar sentence is saying precisely that (namely that it is false), *S* is true. So *S* is true if and only if it is false. Since *S* is one or the other, it is both! Debate about sentences like *S* has been going on among philosophers and logicians for more than 2,000 years without any clear resolution.

The roots of the liar paradox stretch back to the philosopher Epimenides in the sixth century B.C. He said: "All Cretans are liars. . . . One of their own poets has said so." Another version of this can be found in the Bible, Titus 1:12–13, "Even one of their own prophets has said, 'Cretans are always liars, evil brutes, lazy gluttons.' This testimony is true." The poet's (or prophet's) statement is sometimes wrongly considered to be paradoxical because he himself is a Cretan. But actually there's no paradox here. A "liar," in everyday language, is someone who *on occasion* knowingly gives false answers. This isn't problematic: the poet, while lying occasionally, this time spoke the truth. However, most formulations of logic define a "liar" as an entity that always produces the negation of the true answer, that is, someone who does nothing but lie. Thus, the poet's statement can't be true: if it were, then he himself would be a liar who just spoke the truth, but liars don't do that. However, no contradiction arises if the poet's statement is taken to be false: the negation of "All Cretans are liars" is "Some Cretans aren't liars," in other words: some Cretans sometimes speak the truth. This doesn't contradict the fact that our Cretan poet just lied. Therefore, the statement "All Cretans are liars," if uttered by a Cretan, is necessarily false, but not paradoxical. Even the statement "I am a liar" is not paradoxical; depending on the definition of "liar" it may be true or false. However, the statement "I am lying now," first attributed to Eubulides of Miletus in the fourth century B.C., definitely is paradoxical. It is exactly equivalent to the sentence, we started with: "This statement is false."

Various elaborations of the basic Eubulides liar paradox have appeared over the ages. In the fourteenth century, the French philosopher Jean Buridan applied it in his argument for the existence of **God**. In 1913, the English mathematician Philip Jourdain (1879–1921) offered a version that is sometimes referred to as "Jourdain's card paradox." On one side of a card is written:

> The sentence on the other side of this card is true.

On the other side is written:

> The sentence on the other side of this card is false.

Yet another popular version of the liar paradox, guaranteed to perplex, is given by the following three sentences written on a card:

1. This sentence contains five words.
2. This sentence contains eight words.
3. Exactly one sentence on this card is true.

Lie, Marius Sophus (1842–1899)

A Norwegian mathematician who, along with his close friend Felix **Klein**, introduced **group** theory into geometry and used it to classify geometries. Lie discovered the *contact transformation,* which maps curves into surfaces (1870), and **Lie group**s, which use continuous or infinitesimal transformations. He used these groups to classify **partial differential equation**s, making the traditional methods of solution all reduce to a single principle. Lie groups also provided a basis for the growth of modern **topology**. See also **Lie algebra**.

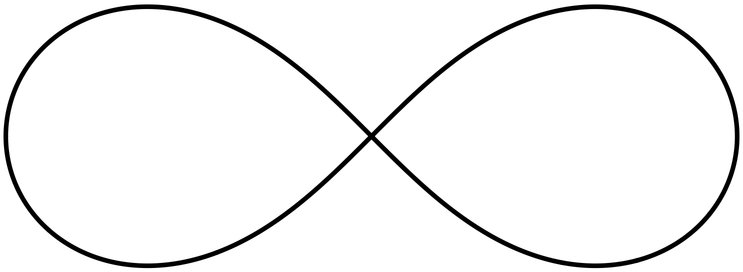

lemniscate of Bernoulli *© Jan Wassenaar, www.2dcurves.com*

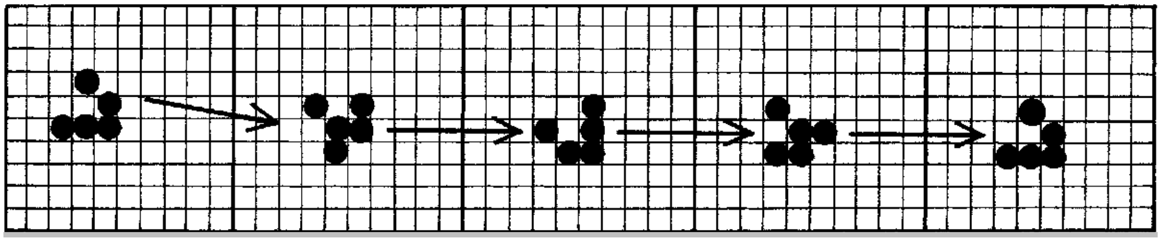

Life, Conway's game of A "glider" advances one cell down and one cell to the right.

Lie algebra

An **algebra**, named after M. Sophus **Lie**, in which multiplication satisfies properties similar to the *bracket operation* on matrices given by $[A, B] = AB - BA$, where the operation on the right-hand side involves ordinary multiplication and subtraction of matrices. The operation is not **associative**.

Lie group

A **group**, named after M. Sophus **Lie**, that is also a **manifold**. Groups of a real **matrix**, such as occur in quantum field theory, give naturally occurring examples of Lie groups. The tangent space at the identity element of a Lie group forms a **Lie algebra** in a natural way.

Life, Conway's game of

The best known example of a **cellular automaton**; it was invented by John **Conway** and first brought to public attention in Martin Gardner's *Scientific American* column in October 1970.[120] Conway's goal in creating Life was to devise a universal **Turing machine**–a sort of infinitely programmable computer. John **von Neumann** had described such a system in the 1950s, but it had very complex rules; Conway wanted to find one that was much simpler to describe and to operate. Life is played on a grid of squares on which each cell is either alive (occupied) or dead (empty). The game starts from an arbitrary initial configuration of live cells, and then progresses through generations as the life and death rules are applied. These rules are very simple: (1) A live cell survives to the next generation if it has two or three neighbors. (2) A live cell dies if it has four or more neighbors (overcrowding) or if it has only one neighbor or none (isolation). (3) A dead cell becomes a live cell in the next generation if it has exactly three neighbors (birth). The rules of Life were developed over a two-year-period during tea and coffee breaks by Conway and a group of graduate students and colleagues. Because **Go** boards and counters were used at this stage, instead of computers, it was important to have a death rule so that populations didn't tend to explode and quickly race off the board. On the other hand, to enable sufficiently interesting behavior so the game had a chance of being a universal system, it was equally important to have a birth rule that prevented populations from dying out. The rules eventually chosen provided a balance between birth and death so that the system tended to be fairly stable yet interesting enough to study. An early sign of success was the discovery of patterns, known as "gliders," that kept their shape while drifting across the plane. This was a hopeful step toward proving universality because it showed that the system had a way to transmit information from one place to another. Conway and his group went on to build nearly all the necessary configurations for arbitrary computations: AND gates, OR gates, and so on, just like the components of an ordinary computer. What they needed next was a way of producing gliders at will–a "glider gun." At this point, Conway sent a letter describing Life and the early findings to Martin Gardner, offering a prize of $50 for a configuration whose population tended toward infinity. The resulting *Scientific American* column sparked the public's imagination and very quickly a glider gun was discovered by a group at the Massachusetts Institute of Technology led by R. W. Gosper. Within two weeks of the discovery of the glider gun, both Conway's group and the group at MIT had shown that the system was indeed universal.

limaçon of Pascal

A snail-shaped curve (*limaçon* is French for "snail") named by the French mathematician Gilles Roberval after Etienne Pascal, the father of Blaise **Pascal**. It had been discovered earlier, however; Albrecht **Dürer** gave a method for drawing it as early as 1525 in his *Underweysung der Messung*. The limaçon of Pascal is a special case of an **epitrochoid** in which the rolling circle and the rolled circle have the same radius, and is also the cata**caustic** and the **pedal curve** of the **circle**. It has the **quartic** Cartesian equation

$$(x^2 + y^2 - 2rx)^2 = k^2(x^2 + y^2),$$

where r is the radius of the rolling circle or the rolled circle and k is a constant.

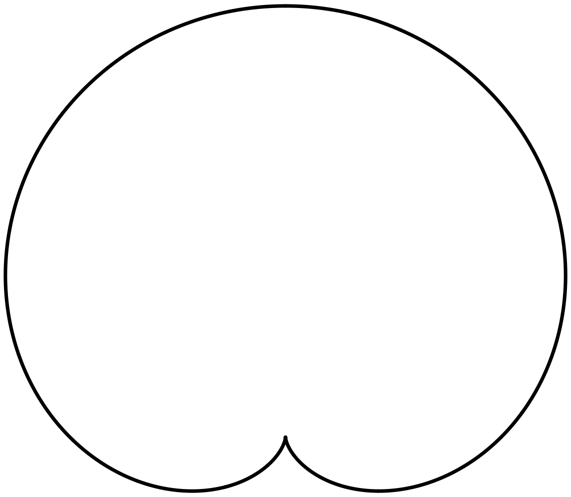

limaçon of Pascal *© Jan Wassenaar, www.2dcurves.com*

Sometimes the term *ordinary limaçon* is used to describe the curve when the value of k is greater than 0 and less than 1. When $k = 0$ the curve is a circle and when $k = 1$ the curve is a **cardioid**, so the ordinary limaçon is a transitional form between these two. The ordinary limaçon is also the **inverse** of the **ellipse**. For values of $k > 1$, a loop or noose appears in the curve. The inverse of a limaçon with a noose is a **hyperbola**. In fact, the constant k is the same as the eccentricity for a **conic section**. When $k = 2$, the limaçon is also called a **trisectrix**.

limerick

Why is this a limerick?:

$$((12 + 144 + 20) + 3\sqrt{4}/7) + 5 \times 11 = 9^2 + 0$$

Because, as its inventor, Jon Saxton, has pointed out:

A dozen, a gross, and a score,
Plus three times the square root of four,
Divided by seven,
Plus five times eleven,
Is nine squared and not a bit more.

limit

The target value that terms in a **sequence** of numbers get closer and closer to. This limit is not necessarily ever reached but can be approached arbitrarily close if the sequence is taken far enough.

limping triangle

A right **triangle** with two shorter sides (i.e., those other than the hypotenuse) that differ in length by one unit. An example is the 20-21-29 triangle ($20^2 + 21^2 = 29^2$).

line

The shortest distance between any two points in **Euclidean space**. A line is implicitly a *straight line;* the alternative is a **curve**. Mathematically, a line may be determined by the presence of any two points in an n-dimensional space (where n is two or more). A *line segment* is a piece of a line with definite endpoints.

linear

Having only a multiplicative factor. If $f(x)$ is a *linear* ***function***, then $f(a + b) = f(a) + f(b)$ and $cf(x) = f(cx)$ must both be true for all values of a, b, c, and x.

linear algebra

The study of **vector**s and **vector space**s. Linear algebra has a central place in modern mathematics, is used widely in both **abstract algebra** and functional **analysis**, and finds a concrete representation in the form of **analytical geometry**. Linear algebra began as the study of vectors in two- and three-dimensions but has now been extended to a generalized n-space.

linear group

A **group** of matrices under **matrix** multiplication.

linear programming

The problem, and associated area of mathematics, of maximizing or minimizing a linear **function** on a **convex** set, especially a **polytope**. Equivalently, maximizing a linear expression in some number of variables subject to linear equalities and inequalities.

linear system

Any system whose change of values of its variables can be represented as a series of points suggesting a straight **line** on a coordinate; hence, *linear* for "line." More generally, a linear system is one in which small changes result in small effects, and large changes in large effects. In a linear system, the components are isolated and noninteractive. Real linear systems are rare in nature since living organisms and their components are not isolated and do interact. Compare with **nonlinear system**.

Liouville number

A **transcendental number** that can be approximated very closely by a **rational number**. Normally, proving that any given number is transcendental is difficult. However, the French mathematician Joseph Liouville (1809–1882) showed the existence of a large (in fact, infinitely large) class of transcendentals whose nature is easy to ascertain. An example of a Liouville number is 0.101001000000100000000000000000000000001 . . . in which the successive groups of zeros are of length 1!, 2!, 3!, 4!, and so on.

Lissajous figure

A figure or graph of the type most often seen on an oscilloscope. The simplest Lissajous figures are circles or ellipses, but they can also take the form of **lemniscate**s and other, more complex shapes. They are named for the French scientist Jules Antoine Lissajous (1822–1880), who experimented with them in the 1850s, and are also known as *Bowditch curves,* because they had been written about earlier by the American astronomer and mathematician Nathanial Bowditch (1773–1838).

lituus

A **spiral** curve that was discovered by the English mathematician Roger Cotes (1682–1716) and named by the Scottish mathematician Colin Maclaurin (1698–1746) in 1722. The Latin *lituus* refers to a staff shaped like a bishop's crosier. The lituus is the locus of a point that moves in such a manner that the area of a circular sector remains constant; it has the polar equation

$$r^2 = a^2/\theta.$$

See **polar coordinates**.

Lobachevsky, Nikolai Ivanovich (1793–1856)

A Russian mathematician who was one of the pioneers of **non-Euclidean geometry**. He developed, independently of János **Bólyai**, the self-consistent system of **hyperbolic geometry** in which Euclid's **parallel postulate** is replaced by one allowing more than one parallel through the fixed point. Lobachevsky first announced his system in 1826 and subsequently wrote several expositions of it, including *Geometrical Researches on the Theory of Parallels* (originally published in 1840 in German). Lobachevsky studied and taught at the University of Kazan and eventually became rector of this institution in 1826. However, for some reason, despite serving his country and university well, he fell from favor and in 1846 was relieved by the government of his posts as professor and rector.

localized solution

A solution of a **differential equation**, or a similar mathematical object, that is confined to a small region even though it has the freedom to spread out.

loculus of Archimedes

A **dissection** game, similar to **tangrams**, which consists of 14 polygonal shapes that fit together to make a square. These pieces can be rearranged to make pictures of people, animals, and objects, or reassembled into their original form. There are many references to the game in ancient literature, including a description by the Roman poet and statesman Magnus Ausonius (A.D. 310–395). Only two fragmentary manuscripts, one an Arabic translation and the other a Greek manuscript dating from the tenth century discovered in Constantinople in 1899, connect the puzzle to **Archimedes** by calling it *loculus Archimedius* ("Archimedes's box"). More generally, but for unclear reasons, it is known as the *ostomachion* (Greek for "stomach"), or, in Latin texts, as the *syntemachion.* In 2003, William **Cutler** used a computer program to enumerate all 536 distinct ways (barring rotations and reflections) in which the pieces can be arranged into a square.

locus

The set of all points (usually forming a curve or surface) that satisfy some condition. For example, the locus of points in the plane equidistant from a given point is a circle. The Latin word *locus* simply means "place." (The Greek equivalent is *topos,* which crops up in "topology".)

logarithm

The logarithm of a number or variable x to base b, $\log_b x$, is the **exponent** of b needed to give x. The bases most commonly used in mathematics are ***e*** and 10. A logarithm to base e, written as $\log x$ or $\ln x$, is known as a *natural logarithm.* A logarithm to base 10 is written as $\log_{10}x$ and is known as a *common logarithm.*

logarithmic spiral

A type of **spiral**, also known as an *equiangular spiral,* that is very common in the natural world. Wonderful examples are found in the shells of some mollusks, such as that of the nautilus, and in spider webs. The angle any **tangent** to the curve makes with a tangent to a circle at the same radius, known as the *pitch angle,* is constant and results in a logarithmic spiral being *self-similar:* in other words, any part of it looks like any other part (though possibly rotated). Hawks approach their prey in the form of a logarithmic spiral and their sharpest view is at an angle to their flight direction that is the same as the spiral's pitch. On an altogether different scale, the arms of spiral galaxies are roughly logarithmic spirals. Our own galaxy, the Milky Way, is believed to have four major arms, each of which is a logarithmic spiral with pitch of about 12°. Approximate logarithmic spirals with a pitch of about 17° can be generated using the **Fibonacci sequence** or the **golden ratio**.

In **polar coordinates** (r, θ) the equation of the logarithmic spiral is

$$r = a\, b^\theta,$$

with positive real numbers a and b. Changing a rotates the spiral while b controls how tightly and in which direction it is wrapped. It can be distinguished from the

Archimedean spiral by the fact that the distance between the arms of a logarithmic spiral increase in a **geometric sequence** while in an Archimedean spiral this distance is constant. Starting at a point *P* and moving inward along the spiral, one has to circle the origin infinitely often before reaching it; yet, the total distance covered is finite. This was first realized by the Italian physicist Evangelista Torricelli (1608–1647) even before calculus had been invented. The total distance covered is $r/\sin\theta$ where θ is the pitch angle and r is the straight-line distance from *P* to the origin.

The logarithmic spiral was first described by René **Descartes** and later studied in depth by Jakob Bernoulli (see **Bernoulli family**), who called it *Spiralis mirabilis* (the wonderful spiral) and wanted one engraved on his tombstone. He did get a spiral, but unfortunately it was a rather crudely cut Archimedean type.

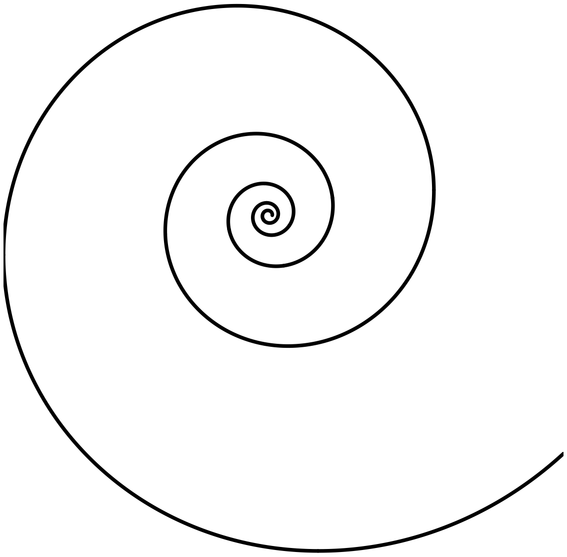

logarithmic spiral *© Jan Wassenaar, www.2dcurves.com*

logic

The branch of **mathematics** concerned with how one statement can imply others, or how sets of statements can

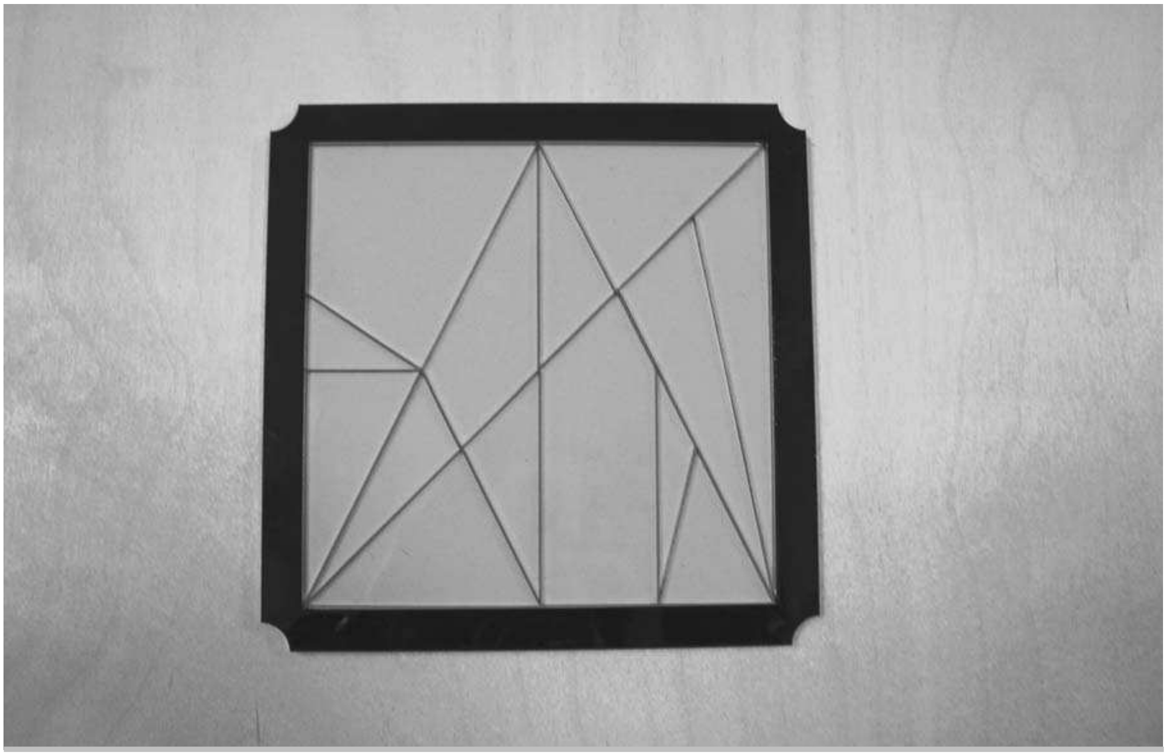

loculus of Archimedes The loculus of Archimedes is one of the most ancient dissection puzzles. *Kadon Enterprises, Inc., www.gamepuzzles.com* be connected by chains of implication. Such relationships can be written down using special symbols, and different sets of rules can be studied to see how they give rise to different sorts of structures.

logical depth

A measure of the **complexity** of a system, developed in 1988 by the American computer scientist and mathematician Charles Bennett. It contrasts with another such measure, **algorithmic complexity**, but, like it, is a measure of the algorithms needed to generate the data from a system.

loop

(1) A **knot** or hitch that holds its form. (2) A degenerate edge of a **graph** that joins a vertex to itself. (3) A sequence of instructions that is repeated either a specified number of times or until a particular condition prevails. Loops lie at the heart of most computer programs. The most common type is the *iterative loop,* signified by keywords such as *for, while, do,* and *repeat,* or their equivalent, in which a given set of instructions is repeated a specified number of times. The *recursive loop* is a more powerful construct that carries out a given set of instructions, typically including **recursive** calls with modified parameters back to the instruction set itself, until a terminating condition is met. Recursive algorithms solve problems by reducing them to smaller and smaller subproblems until a solution is found, reusing the same set of instructions as often as needed.

Lorenz, Edward Norton (1917–)

A research meteorologist at the Massachusetts Institute of Technology who, in the early 1960s, using a simple system of equations to model convection in the atmosphere, ran headlong into the phenomenon of "sensitivity to initial conditions." In the process he sketched the outlines of one of the first recognized **chaotic attractor**s. In Lorenz's meteorological computer modeling, he discovered the underlying mechanism of deterministic chaos: simply formulated systems with only a few variables can display highly complicated behavior that is unpredictable. Using his digital computer, culling through reams of printed numbers and simple strip chart plots of the variables, he saw that slight differences in one variable had profound effects on the outcome of the whole system. This was one of the first clear demonstrations of sensitive dependence on initial conditions. Equally important, Lorenz showed that this occurred in a simple, but physically relevant model. He also appreciated that in real weather situations, this sensitivity could mean the development of a front or pressure-system where there never would have been one in previous models. In his famous 1963 paper, Lorenz picturesquely explained that a butterfly flapping its wings in Beijing could affect the weather thousands of miles away some days later. This sensitivity is now called the **butterfly effect**.

Lorenz system

A system of three **differential equation**s, named after its discoverer Edward **Lorenz**, that was the first concrete example of **chaos** and a **chaotic attractor**.

Lovelace, Lady

See **Byron, Ada**.

loxodrome

A path on Earth's surface that is followed when a compass is kept pointing in the same direction. It is a straight line on a Mercator projection of the globe precisely because such a projection is designed to have the property that all paths along Earth's surface that preserve the same directional bearing appear as straight lines. The loxodrome isn't the shortest distance between two points on a sphere. The shortest distance is an arc of a **great circle**. But, in the past, it was hard for a ship's navigator to follow a great circle because this required constant changes of compass heading. The solution was to follow a loxodrome (from

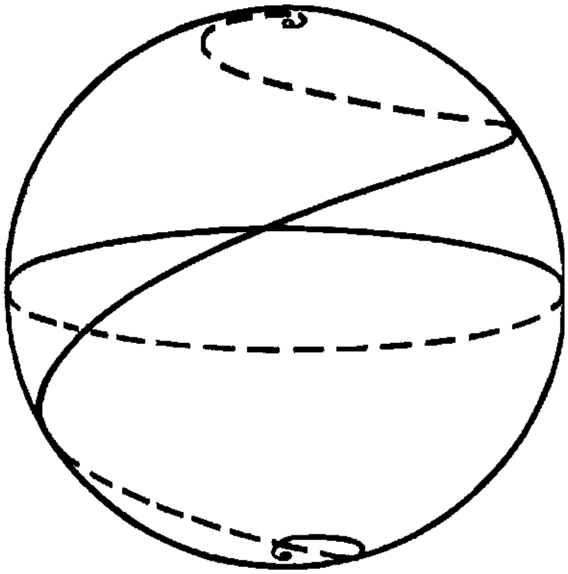

loxodrome Following a constant-compass heading between Earth's poles produces the winding path shown.

the Greek *loxos* for "slanted" and *drome* for "course"), also known as a *rhumb line,* by navigating along a constant direction. In middle latitudes, at least, this didn't lengthen the journey unduly. If a loxodrome is continued indefinitely around a sphere it will produce a spherical **spiral**, or a logarithmic spiral on a polar projection.

Loyd, Sam (1841–1911)

A great American inventor of puzzles among whose best-known creations are the Hoop-Snake Puzzle, **Get Off the Earth**, the **Pony Puzzle**, and, most famous of all, the **Fifteen Puzzle**. By age 17 he was already hailed as one of the world's leading writers of chess problems and had also just invented his deceptively simple-looking trick mules puzzle. The object of this is to cut apart three pieces that show two mules and two jockeys and then reassemble the pieces so that the jockeys are riding the mules. Loyd sold the puzzle to the showman Phineas T. Barnum (of Barnum & Bailey Circus fame) for some $10,000. This was to become his forte: devising puzzles that looked so simple to solve that people felt compelled to try them, only to find, hours later, that they were still trying to figure them out. Loyd became a full-time professional puzzlemaker only in the 1890s. He corresponded with his English counterpart Henry **Dudeney** and worked alongside his son Sam Loyd Jr. After the death of his father, Sam Loyd Jr. continued publishing puzzles that were mainly compilations of his father's work and in 1914 issued a now out-of-print and much sought-after mammoth collection of his father's puzzles called *Cyclopedia of 5,000 Puzzles, Tricks, and Conundrums.*[111, 207, 208]

lozenge

A **rhombus** with a 60° angle.

L-system

A method of constructing a **fractal** that is also a model for plant growth. L-systems use an **axiom** as a starting string and iteratively apply a set of parallel string substitution rules to yield one long string that can be used as instructions for drawing the fractal. Many fractals, including **Cantor dust**, the **Koch snowflake**, and the **Peano curve**, can be expressed as an L-system.

Lucas, (François) Edouard (Anatole) (1842–1891)

A French mathematician well known for his study of the **Fibonacci sequence** and the related **Lucas sequences** named after him. He devised methods of testing for **prime numbers**–work that was later refined by D. H. Lehmer to yield the Lucas-Lehmer test for checking **Mersenne numbers** to see if they are prime. Lucas was also interested in recreational mathematics, the **Tower of Hanoi** being his best known puzzle game. He worked at the Paris Observatory and later became a professor of mathematics in Paris.

Lucas sequences

Generalizations of the **Fibonacci sequence** first investigated by Edouard **Lucas**. One kind can be defined as follows: $L(0) = 0$, $L(1) = 1$, $L(n + 2) = PL(n + 1) + QL(n)$, where the normal Fibonacci sequence is the special case of $P = Q = 1$. Another kind of Lucas sequence begins with $L(0) = 2$, $L(1) = P$. Such sequences are used in **number theory** and in testing for **prime numbers**.

lucky number

A number in a sequence, first identified and named around 1955 by Stanislaw **Ulam**, that evades a particular type of number "sieve" (similar to the famous **sieve of Eratosthenes**), which works as follows. Start with a list of integers, including 1, and cross out every second number: 2, 4, 6, 8, ... The second surviving integer is 3. Cross out every third number not yet eliminated. This removes 5, 11, 17, 23, ... The third surviving number from the left is 7; cross out every seventh integer not yet eliminated: 19, 39, ... Repeat this process indefinitely and the numbers that survive are the "lucky" ones: 1, 3, 7, 9, 13, 15, 21, 25, 31, 33, 37, 43, 49, 51, 63, 67 . . . .

Amazingly, though produced by a sieve based solely on a number's position in an ordered list, the luckies have many properties in common with **prime numbers**. For example, there are 25 primes less than 100 and 23 luckies less than 100. In fact, primes and luckies crop up about equally often between any two given integers. Also, the gaps between successive primes and the gaps between successive luckies widen at roughly the same rate as the numbers increase, and the number of twin primes (primes that differ by 2) is close to the number of twin luckies. The luckies even have their own equivalent of the famous (still unsolved) **Goldbach conjecture**, which states that every even number greater than 2 is the sum of two primes. In the case of luckies, it's conjectured that every even number is the sum of two luckies; no exception has yet been found. Another unresolved problem is whether there are an infinite number of lucky primes. See also **Ulam spiral**.

Ludolph's number

Also known as *Ludolphine,* a name by which the number **pi** was known in Germany for many years following its evaluation to 35 digits by Ludolph von Ceulen (1540–1610).

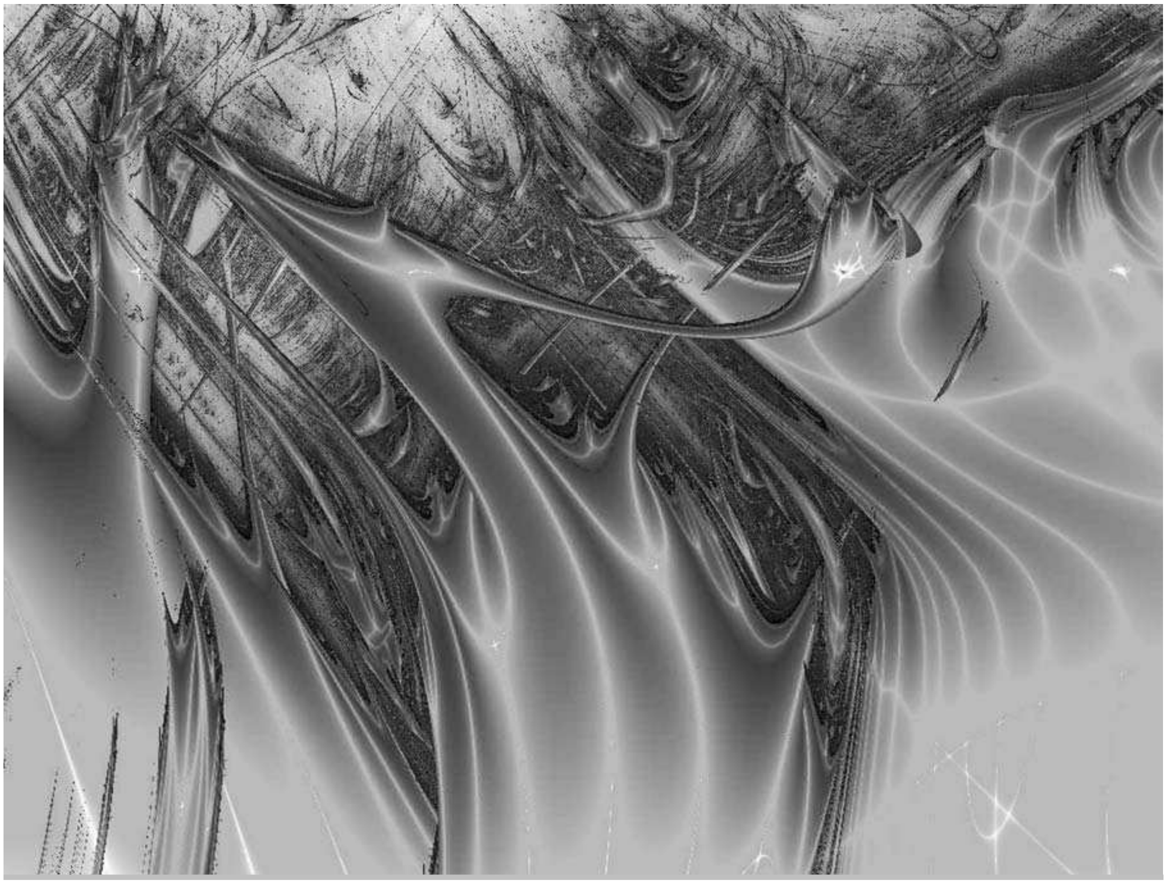

Lyapunov fractal A computer-generated image using Lyapunov exponents. *Radek Novak/Infojet*

lune

The portion of a **sphere** between two great semicircles (see **great circle**) having common endpoints (including the semicircles).

Lyapunov fractal

A particularly photogenic type of **fractal** that is popular with computer artists. It also represents a simple biological model of population growth in which the degree of the growth of the population can periodically alter between two values *a* and *b*.

M

Maclaurin, Colin (1698–1746)

A Scottish mathematician who developed and extended Isaac **Newton**'s work on **calculus** and gravitation, and did notable work on higher plane curves (see **Maclaurin trisectrix**). In his *Treatise of Fluxions* (1742), he gave the first systematic formulation of Newton's methods and set out a method for expanding functions about the origin in terms of series now known as *Maclaurin series*. Maclaurin also invented several devices, made astronomical observations, wrote on the structure of bees' honeycombs, and improved maps of the Scottish isles.

Maclaurin trisectrix

A curve first studied by Colin **Maclaurin** in 1742 with a view to solving one of the great geometric problems of antiquity: **trisecting an angle**. The Maclaurin trisectrix results from the Cartesian equation

$$y^2(a+x) = x^2(3a-x).$$

It is an **anallagmatic curve** that intersects itself at the origin.

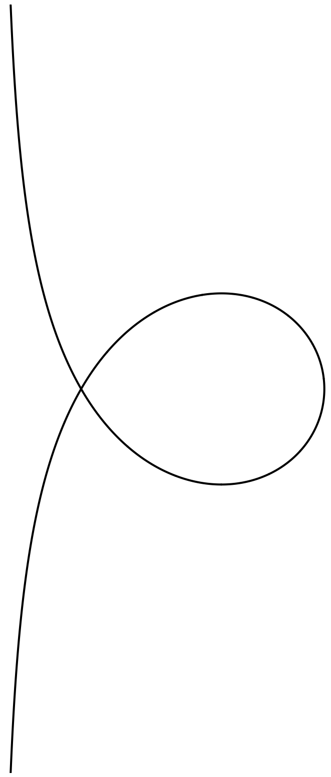

Maclaurin trisectrix *Jan Wassenaar, www.2dcurves.com*

MacMahon, Percy Alexander (1854–1929)

A British mathematician, physicist, and naval officer, born into a military family, whose leanings were evident early on when as a young child he showed a fascination with the way artillery was stacked. MacMahon later did work on missile trajectories, taking resistance into account, and on symmetric functions in the field of **combinatorics**, building on the results of James **Sylvester** and Arthur **Cayley**. His studies in symmetry led him to investigate partitions and to become a world authority on **Latin squares**. He wrote a two-volume treatise *Combinatory Analysis* (1915–1916), which became a classic, and a book on mathematical recreations called *New Mathematical Pastimes* (1921).[210] The latter shows another of the topics that intrigued MacMahon: the construction of patterns that can be repeated to fill the plane. However, much to his regret, as he wrote in the preface, "It has not been found possible to produce the book in colour." See also **MacMahon squares** and **thirty colored cubes puzzle**.

MacMahon squares

The concept of color-matching tiles based on all the permutations of colors on their edges dates from 1926, when Percy **MacMahon** invented and introduced three-color squares and four-color triangles as mathematical pastimes. MacMahon divided squares and triangles into triangles to give each edge of a piece its own color, in all possible combinations. Each set contains 24 different tiles, and MacMahon discovered that they could form a single figure with all adjacent edges matching and just one color all around the outside border. The most extensive research into these sets, over three decades, was done by the American engineer Wade Philpott (1918–1985), of Lima, Ohio, who identified all the possible symmetrical shapes that MacMahon squares and triangles could solve with both matching edge colors and uniform border color, and who calculated all the numbers of solutions for the MacMahon squares' 4 × 6 rectangle. See also **thirty colored cubes puzzle**.

Madachy, Joseph S.

An American mathematician, founder of *Recreational Mathematics Magazine,* editor of the ***Journal of Recreational Mathematics*** for nearly 30 years (now editor emeritus), the author of articles and books on the subject, and a well-known puzzlist.

magic cube

Similar to a **magic square** but having three dimensions instead of two. It contains the integers from 1 to n^3 and has $3n^2 + 4$ lines that sum correctly. All rows, columns, pillars, and the four triagonals (three-dimensional diagonals) must sum to the constant $\frac{1}{2}n(n^3 + 1)$. A cross section through a magic cube is a magic square.

magic square

An $n \times n$ square of the distinct whole numbers $1, 2, \ldots, n^2$, such that the sum of the numbers along any row, column, or main diagonal is the same. This sum is known as the *magic constant* and is equal to $\frac{1}{2}n(n^2 + 1)$. There is only one 3×3 magic square (not counting reflections and rotations), which was known to the Chinese as long ago as 650 B.C. as *lo-shu* and is bound up with a variety of myths. Associations between magic squares and the supernatural are also evident in early Indian and Arabian mathematics. The 3×3 square can be written as:

8 1 6
3 5 7
4 9 2

Each row, column, and main diagonal sums to 15. If the rows are read as three-digit numbers, forward and backward, and then squared, we find the interesting relation

$$816^2 + 357^2 + 492^2 = 618^2 + 753^2 + 294^2.$$

The reader may wish to see if the same rule holds for the columns and main diagonals.

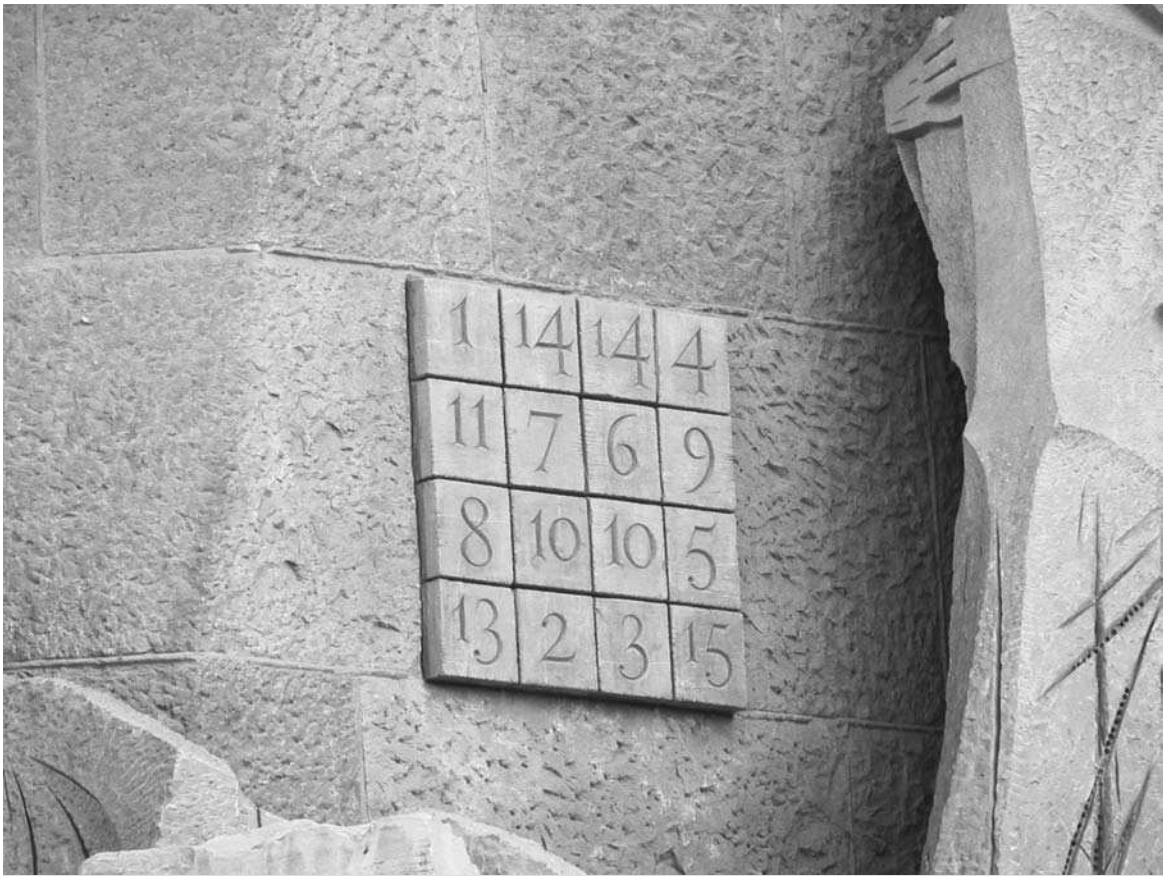

magic square A 4×4 magic square carved on the side of the Sagrada Familia cathedral in Barcelona.

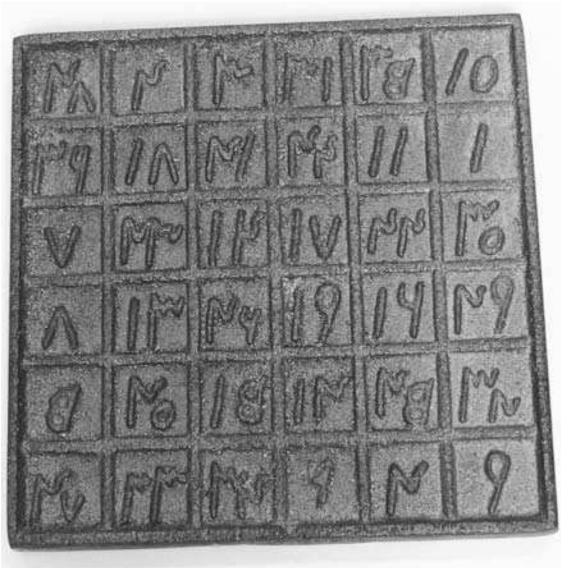

magic square A replica of a cast iron 6 × 6 magic square (vertical, horizontal, and main diagonal lines add to 111) from the Chinese Yuan dynasty (1271–1368), used as a sacred object to ward off evil spirits. *Sue & Brian Young/Mr. Puzzle Australia, www.mrpuzzle.com.au*

In the early sixteenth century Cornelius Agrippa constructed squares for $n = 3, 4, 5, 6, 7, 8$, and 9, which he associated with the seven "planets" then known (including the Sun and the Moon). Albrecht **Dürer**'s famous engraving of *Melancholia* (1514) includes a picture of an order-4 (4 × 4) magic square. There are 880 distinct squares of order-4 and 275,305,224 squares of order-5, but the number of larger squares is unknown. A square that fails to be magic only because one or both of the main diagonal sums don't equal the magic constant is called a *semi-magic square.* If *all* diagonals (including those obtained by wrapping around) of a magic square sum to the magic constant, the square is said to be a *pandiagonal square* (also known as a *panmagic* or *diabolical square*). Pandiagonal squares exist for all orders except 6, 10, 14, . . . , $2(2i + 1)$. There are 48 pandiagonal 4 × 4 squares. If replacing each number n_i by its square n_i^2 produces another magic square, the square is said to be a *bimagic* or *doubly magic square.* If a square is magic for n_i, n_i^2, and n_i^3, it is known as a *trebly magic square.*

A little trial and improvement is all it takes to construct the 3 × 3 magic square, but for building 4 × 4 squares and larger, a systematic method, or **algorithm**, is important. Interestingly, different algorithms are needed depending on whether the square is of an even order or an odd order. Odd order squares are the easier variety to make and there are several standard techniques, including the Siamese (sometimes called de la Loubere's or the Staircase), the Lozenge, and de Meziriac's methods. Here is yet another approach, known as the Pyramid or extended diagonals method: (1) Draw a pyramid of same size squares as the magic square's squares, on each side of the magic square; the number of squares on the pyramid's base should be two less than the number of squares on the side of the magic square. (2) Sequentially place the numbers 1 to n^2 of the $n \times n$ magic square in the diagonals. (3) Relocate any number not in the $n \times n$ square to the opposite hole inside the square.

An *antimagic square* is an $n \times n$ array of integers from 1 to n^2 in which each row, column, and main diagonal gives a different sum such that these sums form a sequence of consecutive integers. A 4 × 4 antimagic square is a square arrangement of the numbers 1 to 16 so that the totals of the four rows, four columns, and two main diagonals form a sequence of 10 consecutive integers, for example:

1	12	3	12
15	9	4	10
7	2	16	8
14	6	11	5

The principle of magic squares can be extrapolated from two dimensions to any number of **higher dimensions**, including **magic cube**s and *magic* ***tesseract****s,* whose cross sections consist of magic cubes, and so forth.[10, 160, 162] See also **Latin square**s and **magic tour**.

magic tour

A **tour** by a **chess** piece on an $n \times n$ chessboard, whose squares are numbered from 1 to n^2 along the path of the piece, such that the resulting arrangement of numbers is a **magic square**. The tour is a *semi-magic tour* if the resulting arrangement of numbers is a **semi-magic square**. Magic **knight's tour**s aren't possible on $n \times n$ boards if n is odd. They are possible for all boards of size $4k \times 4k$ for $k > 3$, but are believed to be impossible for $n = 8$.

main diagonal

In the $n \times n$ **matrix** $[a_{ij}]$, the elements $a_{11}, a_{22}, \ldots, a_{nn}$. See also **diagonal**.

major axis

The longest **chord** of an **ellipse**.

Malfatti circles

In 1803 the Italian mathematician Giovanni Malfatti (1731–1807) posed the following problem: Given a triangle, find three nonoverlapping circles inside it such that the sum of their areas is maximal. Malfatti and many other mathematicians thought that the solution is given by the three circles, each of which is **tangent** to the other two and also to two sides of the triangle. Malfatti computed the radii of these circles, and they are now known as Malfatti's circles. Later it became clear that Malfatti's conjecture isn't true. In particular, Goldberg proved, in 1969, that the Malfatti circles *never* give a solution of the Malfatti problem! In other words, for any triangle there are three nonintersecting circles inside it, whose areas are bigger than the area of the Malfatti circles. So far as is known, the Malfatti problem hasn't been solved yet in the general case although it seems reasonable to suppose that the solution is given by what is called the *greedy algorithm:* First inscribe a circle in the given triangle; then inscribe a circle in the smallest angle of the triangle that is tangent to the first circle. The third circle is inscribed either in the same angle or in the middle angle of the triangle, depending on which of them has the bigger area.

Mancala

What is today almost universally called Mancala is, correctly speaking, the game of *Wari.* Mancala refers to a type of counting game of which Wari is a specific example, dating back to ancient times, known in Egypt and throughout Africa and Asia, and probably brought to Europe by sailors returning from those lands. Wari is played with a board and 48 markers or playing pieces, which are usually small colored stones or shells. The board consists of a piece of wood with two rows of six hollowed-out circular spaces, called *cups,* on each side, and one larger oval space at each end called a *reservoir.* The game is set up by placing four markers in each of the twelve cups. The first player picks up the markers in one of the cups on his side, then distributes the markers by placing one (and only one) in the other cups, going counterclockwise around the board. Markers are not placed in the reservoirs, which are for holding captured pieces only. Should there be enough markers that a complete circle of the board is made, the cup just emptied is skipped over. If the last marker put down goes into a cup on the opponent's side, and ends there with a total of two or three markers in that cup, then the markers in that cup are captured, and go into the reservoir at the player's right. Also, if the next cup clockwise from the captured cup has only two or three markers in it, that cup is likewise captured and the markers taken. This continues so long as the cup is clockwise to the last captured cup, has only two or three markers in it, and is an opponent's cup. However, a player can't capture all of an opponent's remaining pieces in a move, leaving him nothing to play. In capturing cups, if there's only one opponent cup left with markers in it, it can't be captured. Also, you cannot leave your opponent with all empty cups if you have a move that would put some pieces in one or more of his cups. When one side of the board is empty, play is ended. This happens when a player has to move his last markers onto his opponent's side. At the end of play each player takes any markers left on his side of the board and adds them to his reservoir. The winner is the person with the most markers in his reservoir.

Mandelbrot, Benoît B. (1924–)

A Polish-born French mathematician, largely responsible for the present interest in **fractal** geometry. A native of Warsaw, he spent most of his early life in France. Mandelbrot was born into a family with a strong academic tradition: his mother was a doctor and his uncle Szolem Mandelbrot was a famous Parisian mathematician. His family left Poland for Paris in the 1930s to escape Hitler's regime. There, Mandelbrot was introduced to mathematics by his two uncles. Educated in France, he developed the mathematics of Gaston **Julia**, and began the (now common) graphing of equations on a computer. Mandelbrot originated what became known as fractal geometry and the object known as the **Mandelbrot set** is named after him. His work on fractals as a mathematician at IBM earned him an emeritus fellowship at the Thomas J. Watson Research Laboratories. In addition to his study of fractals in mathematics, he showed that fractals can be found in many places in nature, leading to entire new fields of exploration in **chaos theory**. He joined the faculty of Yale in 1987.

Mandelbrot set

The best known **fractal** and one of the most complex and beautiful mathematical objects known. It was discovered by Benoît **Mandelbrot** in 1980 and named after him by Adrien Douady and J. Hubbard in 1982. The set is produced by the incredibly simple **iteration** formula

$$z_{n+1} = z_n^2 + c,$$

where z and c are complex numbers and $z_0 = 0$. This can be written without complex numbers as

$$x_{n+1} = x_n^2 - y_n^2 + a, \text{ and}$$
$$y_{n+1} = 2x_n y_n + b,$$

where $z = (x, y)$ and $c = (a, b)$. The Mandelbrot set consists of all the points on the **Argand diagram** for which the function $z^2 + c$ doesn't diverge under iteration. A computer is essential for carrying out the necessary calculations and for producing pictures of this remarkable structure. For the purposes of computation, the Argand diagram is broken down into pixels (picture elements), and the coordinates of each supply the constant c in $z^2 + c$. For each pixel (value of c) the function is iterated. If the function either rapidly diverges (blows up) or rapidly

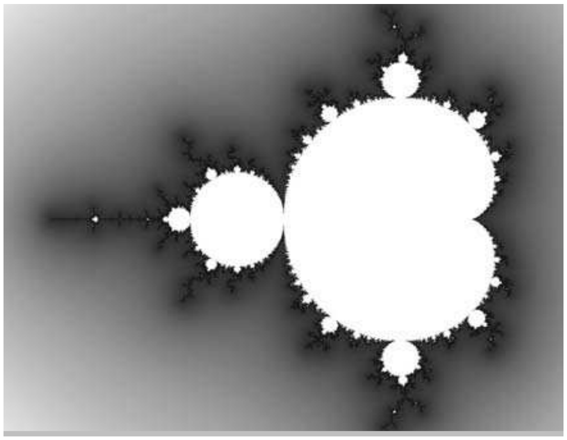

Mandelbrot set A general view of the Mandelbrot set.

converges (collapses), the pixel is left black. If the function is more indecisive about which way it is heading, it is allowed to iterate longer. In some cases the iterations could go on for a very long time before it became clear that the function would ultimately diverge, so a limit is established, known as the *depth,* beyond which iterations are stopped. If the depth is reached without divergence, the corresponding pixel is left black as though it were in the set. At locations where divergence becomes clear prior to hitting the limit, the pixel is displayed according to a scale that represents how many iterations are needed to show divergence. The whole Mandelbrot set lies within a circle of radius 2.5 centered at the origin of the Argand diagram. Although finite in area, the set has a boundary that is infinitely long and has a **Hausdorff dimension** of 2.

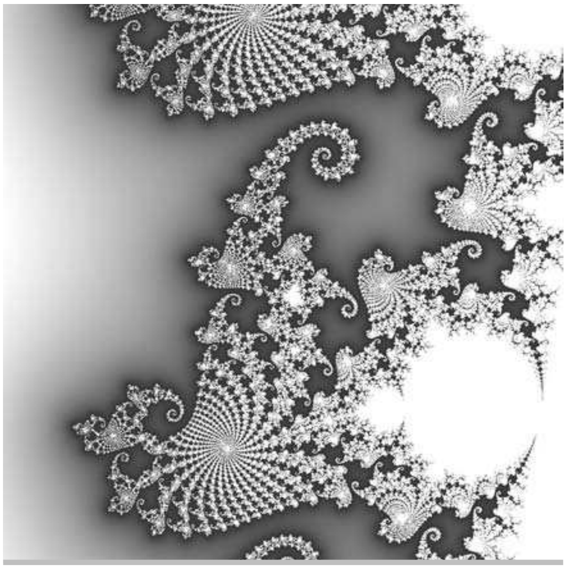

Mandelbrot set A zoom view of the Mandelbrot set reveals some of the infinite variety of remarkable patterns that inhabit it.

The overall appearance of the Mandelbrot set is that of a series of disks. These disks have irregular borders and decrease in size heading out along the negative real axis; moreover, the ratio of the diameter of one disk to the next approaches a constant. More complex shapes branch out from the disks. One region of the Mandelbrot set containing spiral shapes is known as *Seahorse Valley* because it resembles a seahorse's tail. A computer can be used, like a microscope, to zoom in on different parts of the set. This reveals that, although the shape is infinitely complex, it also displays **self-similarity** with regions that look like the outline of the entire set. The Mandelbrot set also reveals symmetry on different levels. It is identically symmetrical about the real axis, and almost symmetrical at smaller scales. This kind of "near-but-not-quite" symmetry is one of the most unexpected properties to find in an object generated from such a simple formula and process. The Mandelbrot set was created by Mandelbrot as an index to the **Julia set**s. Each point in the Argand diagram corresponds to a different Julia set, and those points within the Mandelbrot set correspond precisely to the connected Julia sets.

manifold

A mathematical object that, in geometrical terms, is nearly "flat" on a small scale (though on a larger scale it may bend and twist into exotic and intricate forms). More precisely, a manifold is a **topological space** that looks locally like ordinary **Euclidean space**. Every manifold has a *dimension,* which is the number of coordinates needed to specify it in the local coordinate system. A circle, although curved through two dimensions, is an example of a one-dimensional manifold, or one-manifold. A close-up view reveals that any small segment of the circle is practically indistinguishable from a straight line. Similarly, a sphere's two-dimensional surface, even though it curves through three dimensions, is an example of a two-manifold. Seen locally, the surface, like that of a small portion of Earth, appears flat. A manifold that is smooth enough to have locally well-defined directions is said to be *differentiable.* If it has enough structure to enable lengths and angles to be measured, then it is called a *Riemannian manifold.* Differentiable manifolds are used in mathematics to describe geometrical objects, and are also the most natural and general settings in which to study differentiability. In physics, differentiable manifolds serve as the **phase space** in classical mechanics, while four-dimensional pseudo-Riemannian manifolds are used to model **space-time** in general relativity.

mantissa

The positive fractional part of the representation of a **logarithm**. For example, in the expression log 3,300 = 3.5185 . . . , 0.5185 is the mantissa.

many worlds hypothesis

> *Do you not think it a matter worthy of lamentation*
> *that when there is such a vast multitude of them*
> *[worlds], we have not yet conquered one?*
> –Alexander the Great

An interpretation of **quantum mechanics**, first proposed by the American physicist Hugh Everett III in 1957, according to which, whenever numerous viable possibilities exist, the world splits into many worlds, one world for each different possibility (in this context, the term *worlds* refers to what most people call "universes"). In each of these worlds, everything starts out identical, except for the one initial difference; but from this point on, they develop independently. No communication is possible between the separate universes, so the people living in them (and splitting along with them) would have no idea what was really going on. Thus, according to this view, the world branches endlessly. What is "the present" to us, lies in the pasts of an uncountably huge number of different futures. Everything that *can* happen does happen, somewhere. Until the many worlds interpretation, the generally accepted interpretation of quantum mechanics was (and perhaps still is) the *Copenhagen interpretation.* The Copenhagen interpretation makes a distinction between the observer and the observed; when no one is watching, a system evolves deterministically according to a wave equation, but when someone *is* watching, the wave function of the system "collapses" to the observed state, which is why the act of observing changes the system. The Copenhagen interpretation gives the observer special status, not accorded to any other object in quantum theory, and cannot explain the observer itself, while the many worlds hypothesis models the entire observer-observee system.

map

(1) A synonym for **function**, in which context it is also known as a *mapping.* More generally, the correspondence of elements in one **set** to elements in the same set or another set. (2) A representation, usually on a plane surface, of geographical regions. See **four-color map problem**.

Markov chain

A sequence of **random** variables in which the future variable is determined by the present variable but is independent of the way in which the present state arose from its predecessors. In other words, a Markov chain describes a chance process in which the future state can be predicted from its present state as accurately as if its entire earlier history was known. Markov chains are named after the Russian mathematician Andrei Andrevich Markov (1856–1922) who first studied them in a literary context, applying the idea to an analysis of vowels and consonants in a text by Pushkin. But his work launched the theory of **stochastic** processes and has since been applied in quantum theory, particle physics, and genetics.

Martingale system

A simple, popular, and ultimately disastrous gambling system that, on the face of it, seems like a dream come true. In the short run the player has a good chance of making a few dollars using this method. But, in the long run, two things conspire to defeat him–the table betting limits and the player's available funds. The Martingale system calls for an initial bet of, say, $2. If the player loses, he doubles his bet to $4. Another loss puts the net loss at $6 and requires a doubling of the bet to $8 to recoup the losses and show a profit. Assume the player loses five hands in a row. The sixth bet requires $64. Let's say this wins–the gambler has now won $128 but has lost $124–a net win of $4. Of course, any win, however small, if repeated over and over, could produce a fortune. The trouble is that many losing streaks run longer than 6 or 8 or 10 in a row. The Martingale quickly runs into the table limits. For example, if the player is at a $2 **blackjack** table with a $500 upper limit, he has to retire after 9 losses in a row and is down over $1,000. It would take another 500 winning hands to make up this loss! Basically, one losing streak will put the Martingale gambler in a hole he is unlikely ever to climb back out of.

Mascheroni construction

A construction done using a moveable compass alone, named after the Italian geometer Lorenzo Mascheroni (1750–1800), who, in his *Geometria del compasso* (1797), astonished the mathematical world by showing how every compass-and-straightedge construction can be done in this minimalist way. (Since straight lines can't be drawn with just a compass, it's assumed that two points, obtained by arc intersections, define a straight line.) It is now known that Georg Mohr (1640–1697) proved the same results earlier in his obscure *Euclides danicus* (1672). Mascheroni, or Mohr-Mascheroni, constructions are today primarily of interest to puzzle enthusiasts who try to improve on the older solutions by finding ones with fewer steps.

matchstick puzzle

A form of **mechanical puzzle** that involves rearranging a pattern of ordinary matches, according to a given instruction, to make a new shape or solve some other mathematical problem.

mathematical lifespan

"The mathematical life of a mathematician is short. Work rarely improves after the age of twenty-five or thirty. If little has been accomplished by then, little will ever be accomplished." Thus wrote Alfred Adler in an article titled "Mathematics and Creativity" in the *New Yorker* magazine (1972), echoing a common belief that mathematicians tend to do their best work before the age of 30, physicists before the age of 40, and biologists before the age of 50 (though there are exceptions!). The mathematical physicist Freeman Dyson put it more succinctly: "Young men should prove theorems, old men should write books." On the other hand, there are compensations for early burnout, as G. H. **Hardy** pointed out (in *A Mathematician's Apology*): "Archimedes will be remembered when Aeschylus is forgotten, because languages die and mathematical ideas do not. 'Immortality' may be a silly word, but probably a mathematician has the best chance of whatever it may mean."

mathematics

> *Pure mathematics consists entirely of assertions to the effect that if such and such a proposition is true of anything, then such and such another proposition is true of that thing. It is essential not to discuss whether the first proposition is really true, and not to mention what the anything is of which it is supposed to be true. . . . Thus mathematics may be defined as the subject in which we never know what we are talking about, nor whether what we are saying is true.*
>
> –Bertrand Russell

The science of patterns, real or imagined; *mathematics* comes from the Greek *mathema* for "knowledge" or "that which is learned." Its roots lie in the practical need to carry out commercial calculations, to measure land, and to forecast astronomical events. These activities correspond roughly to the mathematics of structure, space, and change. The investigation of structure begins with numbers–initially the **natural number**s and **integer**s. The rules governing arithmetical operations are dealt with in elementary **algebra**, while the deeper properties of whole numbers are the province of **number theory**. The study of methods to solve equations leads to **abstract algebra**, which deals with structures that generalize the properties of familiar numbers. The physically important concept of **vector**, generalized to **vector space**s and studied in **linear algebra**, embraces both structure and space. The mathematics of space stems from **geometry**, first the **Euclidean geometry** and **trigonometry** of the everyday world, and later the various forms of **non-Euclidean geometry**. The modern fields of **differential geometry** and algebraic geometry generalize geometry in different ways. Differential geometry builds upon the concepts of coordinate system, smoothness, and direction, while algebraic geometry treats geometrical objects as sets of solutions to **polynomial** equations. **Group** theory places the concept of symmetry on an abstract footing and provides a bridge between space and structure. **Topology** links space and change through its emphasis on **continuity**. Analyzing and describing change in the physical world is a perennial theme of the natural sciences, and **calculus** was developed as a tool for doing this. The central concept used to describe a changing variable is that of a **function**. Many problems lead to relations between a quantity and its rate of change, and the methods to solve these are studied in the field of differential equations. The numbers used to represent continuous quantities are **real number**s, and the study of their properties and the properties of real-valued functions is known as real **analysis**. For various reasons, it's convenient to generalize to **complex number**s, which are dealt with in **complex analysis**. Functional analysis focuses attention on (typically infinite-dimensional) spaces of functions, laying the groundwork for **quantum mechanics** among many other things. In order to probe the foundations of mathematics, the fields of **set** theory, mathematical **logic**, and model theory were developed.

Since the time of the ancient Greeks, thought has been given to the ultimate nature of mathematics. What is its role and status in reality? Most crucially, is it invented or discovered? Leopold **Kronecker** was on the side of invention: "God made the integers; all the rest is the work of Man." Charles **Hermite**, by contrast, was clearly a Platonist: "There exists . . . an entire world, which is the totality of mathematical truths, to which we have access only with our mind, just as a world of physical reality exists, the one like the other independent of ourselves, both of divine creation." The German physicist Heinrich Hertz went even further: "One cannot escape the feeling that these mathematical formulas have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverers, that we get more out of them than was originally put into them." G. H. **Hardy** summed up what many mathematicians today have a tendency to believe:[151] "that mathematical reality lies outside us, that our function is to discover or observe it, and that the theorems which we prove, and which we describe grandiloquently as our 'creations,' are simply the notes of our observations." In another passage, he made this point more precisely: "317 is a prime, not because we think so, or because our minds are shaped in one way rather than another, but because it is so, because mathematical reality is built that way." We are struck by how well mathematics describes the behavior of the world in which we live. The universe, in fact, appears to have a deep mathematical infrastructure. Martin **Gardner** makes the bold claim that: "Mathematics is not only real, but it is the only reality. [The] . . . entire universe is made of matter. . . . And matter is made of particles. . . . Now what are the particles made out of? They're not made out of anything. The only thing you can say about the reality of an electron is to cite its mathematical properties. So there's a sense in which matter has completely dissolved and what is left is just a mathematical structure."

matrix

A square or rectangular array of numbers, usually written enclosed in a large pair of parentheses. Matrices, which are added and multiplied using a special set of rules, are extremely useful for representing quantities, particularly in some branches of physics. A matrix can be thought of as a **linear** operator on **vector**s. Matrix-vector multiplication can be used to carry out geometric **transformation**s such as scaling, rotation, reflection, and translation.

Matrix comes from the same Latin root that gives us *mother,* and was used to refer to the womb and to pregnant animals. It became generalized to mean any situation or substance that contributes to the origin of something. The first mathematical use of the word *matrix* was around 1850 by James **Sylvester** who saw a matrix as a way of obtaining **determinant**s, but didn't fully appreciate its potential. Within a year of his first mention of the term, he introduced the idea to Arthur **Cayley** who was the first to publish the **inverse** of a matrix and to treat matrices as purely abstract mathematical forms. The use of mathematical arrays to solve problems predates the application of the name by about 2,000 years. Around 200 B.C. in the Chinese text *Juizhang Suanshu* (Nine Chapters on the Mathematical Arts) the author solves a system of three equations in three unknowns by placing the **coefficient**s on a counting board and solving by a process that today would be called *Gaussian elimination.*

maximum

The largest of a set of values.

maze

A network of winding and interconnected passageways that a traveler must negotiate in order to reach some goal. The terms *maze* and *labyrinth* are often used interchangeably, though sometimes a distinction is made based on the layout. A labyrinth is then defined as a construction that leads from a starting point to a goal by a *single* path, with no branches or dead ends. No matter how long and twisting the route, it is predetermined by the builder: a labyrinth, according to this definition, is *unicursal.* A maze, by contrast, is *multicursal* and calls for the traveler to make a series of decisions that affect how quickly the goal is reached. In this book, *maze* and *labyrinth* mean the same thing, and refer to unicursal and multicursal mazes.

The most famous of legendary mazes was the lair (or prison) of the Minotaur at Knossos on Crete. According to Greek mythology, King Minos of Crete had his chief engineer, Daedalus, build the labyrinth in order to keep the half-human half-bovine offspring of his wife Pasiphae and a bull out of the public eye. King Aegeus of Athens was forced to pay a periodic tribute to Minos (the Athenians having earlier murdered Minos's son) in the form of seven young men and seven maidens. These unfortunates were forced to enter the maze below Minos's palace where they would get hopelessly lost and eventually be eaten by the monster. King Aegeus' son, Theseus, decided to put a stop to this and offered to take the place

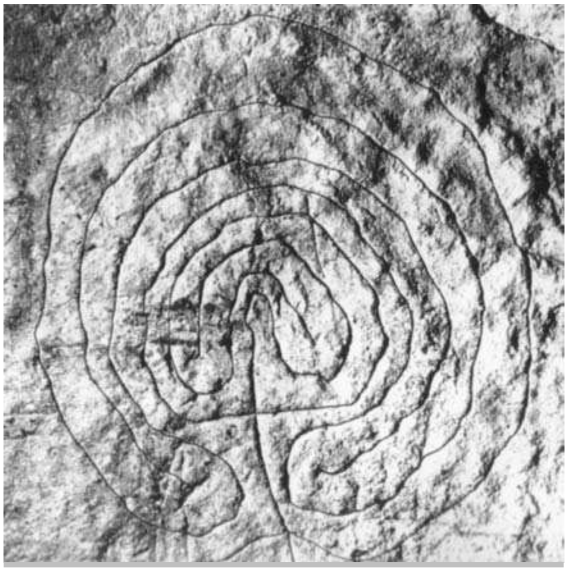

maze Possibly the world's oldest surviving labyrinth: a seven-ring labyrinth rock carving inside the Tomba del Labirinto, a Neolithic tomb at Luzzanas, Sardinia, dating somewhere between 2500 and 2000 B.C. *National University of Singapore*

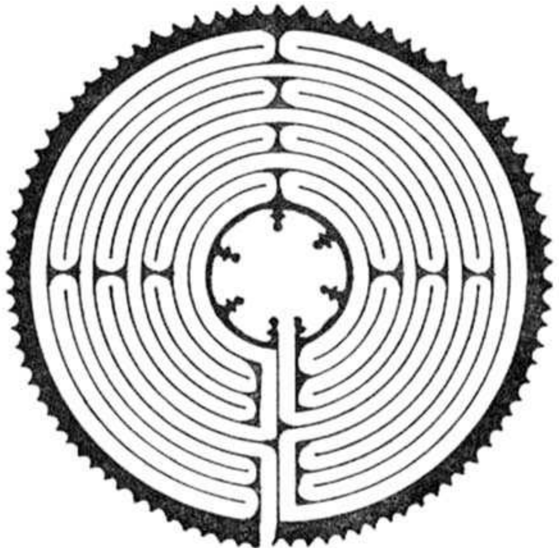

maze A plan of the maze in Chartres Cathedral as drawn by Henry Dudeney.

of one of the sacrificial virgins. He fell in love with Ariadne, one of Minos's daughters, who gave him a *clew,* or ball of yarn, to unravel as he entered the labyrinth so that he would be able to find his way back out. (Originally, *clue* and *clew* were alternative spellings of the same word. The modern sense of *clue* as a guide to solving a problem comes from the legend of the Minotaur, while *clew* retains its ancient meaning.) In true heroic style, Theseus slew the Minotaur but then spoiled the fairytale ending by abandoning Ariadne on the voyage home. Did this labyrinth really exist? In Roman times some writers suggested that a set of winding caves at Gortyan, in southern Crete, might have formed the basis for the tale. Although this complex of natural passages sounds similar to the maze in the myth, the story definitely places the labyrinth at the ancient Cretan capital. Archaeologists have found no evidence of a labyrinth structure at Knossos, but it has been suggested that the palace was so complicated, with its many levels, stairs, and rooms, that it may itself have inspired the story.

Minoan coins from about 300 B.C. bear a round, winding design, thought to be a representation of the labyrinth. A very similar geometric pattern recurs across many different cultures and times–scratched into caves in Cornwall (possibly by visiting Phoenician seafarers), on Roman coins, and in pictures drawn by native American Indians. Almost all these designs, including the one on the Minoan coins, are unicursal. Quite why Theseus would have needed a clew to navigate a maze that had only one way in and out is anybody's guess. But the underlying reason for the early unicursal design isn't hard to find. These mazes were intended not as intellectual puzzles but as symbolic representations of destiny as a matter of fate, beyond personal control. The world's oldest known maze-like designs, dating back about 3,500 years, have been found carved on rocks in northwest Spain and around the shores of the Mediterranean, and according to legend, were walked by fishermen before setting sail to ensure favorable winds and a good catch. In medieval times, mazes started to appear in churches, as art painted on walls or inlaid as "pavement mazes" on the floor. Some of the larger floor versions were traversed by people on their knees as a form of repentance or traversed as a substitute for an actual pilgrimage to the holy city, earning them the name "*Chemin de Jerusalem,*" or Road of Jerusalem. The oldest-known church labyrinth, at the Basilica of Reparatus at Orleansville, Algeria, dates from the fourth century A.D. and measures about 8 feet in diameter. One of the largest, built in 1288, formed part of the floor in the nave of Amiens Cathedral in France and spanned about 42 feet, but was destroyed in 1825. A splendid surviving example, however, is in Chartres Cathedral near Paris. Laid down in about 1200, this is an 11-circuit design (11 concentric windings) divided into quadrants, of the type often found in Gothic cathedrals. The only cathedral maze in Britain is at Ely, 16 miles north of Cambridge; when the Cathedral was restored around 1870, the architect, Gilbert Scott, installed a pavement maze of his own design under the west tower.

Mazes built primarily as puzzles or for recreation represent a different type of structure. For one thing, they are multicursal, offering those who enter a series of choices about which way to proceed. In his *Natural History,* the Roman historian Pliny comments that the classical type of labyrinth is quite distinct from "the mazes formed in the fields for the entertainment of children," suggesting that these diversions may have a long history. But it was in England that they really came of age. Church mazes never caught on in England, but turf mazes became tremendously popular. Ranging from 25 feet to over 80 feet across, they were constructed in or just outside villages across the countryside and were given names such as "Mizmaze," "Troy Town," "Shepherd's Race," and "Julian's Bower." A Welsh history book *Drych y Prif Oesoedd* (Mirror of the first age) published in 1740 notes the curious custom shepherds had of cutting the turf in the form of a labyrinth, which would seem to account for the origin of "Shepherd's Race." "Troy Town" probably refers to a legend that the city of Troy had seven exterior walls arranged as a maze to frustrate an attacking force.

From the turf maze it was no great leap to perhaps the most famous form of full-size maze–the topiary or hedge maze. While the use of hedges in gardens dates back to Roman times, the earliest references to a topiary maze appears in thirteenth-century Belgium. By the sixteenth century the hedge maze had spread to England, as a landscape painting by Tintoretto attests. In the later part of the seventeenth century, Louis XIV had a labyrinth built as a part of the gardens at Versailles, which included 39 groups of hydraulic statuary representing the fables of Aesop. The most famous surviving historic hedge maze is that on the grounds of Hampton Court Palace in England, designed by George London and Henry Wise for William of Orange, planted between 1689 and 1694, and occupying about a third of an acre. The finest turf maze in England may be that at Bridge End Gardens in Saffron Walden, Essex, which was replanted with yew in 1838–1840, abandoned and lost by 1949, then restored in 1983. A second maze, an imitation of that at Hampton Court, is on the Common.

In August 1997, Adrian **Fisher** opened the "World's Largest Maze" at Millets Farm Centre, Frilford, Oxfordshire, England, cut from about six acres of grain. However, his 1995 effort in Shippensburg, Pennsylvania, was in a 30-acre field and the 1997 maze in Reignac sur Indre covers 37 acres.[101]

When faced with a maze, what is the best way of reaching the goal, whether this is a point in the middle of the maze or an exit that forms a second opening to the structure? A unicursal maze calls for no brainwork, only footwork, since it consists of a single winding passage with no offshoots. Multicursal mazes are a different story. The easiest solution is to place a hand on one wall at the outset and follow that wall, come what may. Each blind alley will be traversed one time in and out until the whole maze is completed, or the goal is found. This simple method isn't the most efficient and it will fail altogether if the goal lies within an island in the maze–that is, a section detached from any of the exterior walls. A classic general method of "threading a maze" is: (1) never traverse a path more than twice; (2) when arriving at a new branch point or node, select either path; (3) when arriving at an old node or at a dead end by a new path, return by the same path; and (4) when arriving at an old node by an old path, select a new path, if possible; otherwise, an old path. An

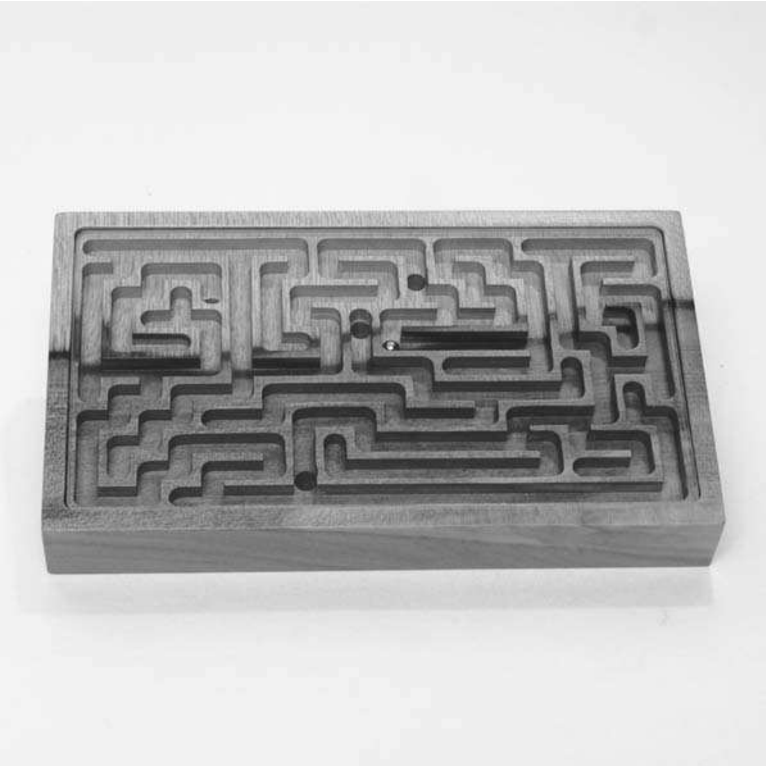

maze A handheld maze, the object of which is to get the silver ball from the starting rectangle on the left to the finishing rectangle on the right. Crafted from Tasmanian blackheart sassafrass by Kym Anderson. *Mr. Puzzle Australia, www.mrpuzzle.com.au* explorer who follows these rules, using marks on the ground, as a record, can be assured of visiting every part of the maze.[48, 219, 240, 294]

See also **Rosamund's bower**.

mean

> *When she told me I was average, she was just being mean.*
>
> –Anonymous

A precisely calculated "typical" value of a group of numbers. There are several kinds of means. The *arithmetic mean,* which is usually what is meant when people talk about an **average**, is the sum of a set of values divided by the number of values. For example, the arithmetic mean of 3, 4, 7, and 10 is (3 + 4 + 7 + 10)/4 = 6. The *geometric mean* of *n* values is the *n*th root of the product of the values. For example, the geometric mean of 3, 8, and 10 is $(3 \times 8 \times 10)^{1/3}$ or the cube root of 240. The *harmonic mean* is the reciprocal of, or one over, the mean of the reciprocals of the values. For example, the harmonic mean of 3, 8, and 10 is $1/[(1/3 + 1/8 + 1/10)/3]$. A different kind of average is *median,* which is generally the center term when a group of numbers are ordered by size. If there are an even number of values then the arithmetic mean of the center two is the median. The words *mean* and *median* both come from the Indo-European root *medhyo* meaning "middle."

measure

A way of gauging how big something is in terms of length, volume, or some other quality. One of the strangest facts in mathematics is that some objects exist that can't be measured. In the language of **set**s, the basic rules (somewhat simplified) of mathematical measures are as follows: (1) the measure of any set is a **real number**; (2) the **empty set** has measure zero; (3) if A and B are two sets with no elements in common then the measure of $A \cup B$ (the union of A and B) is equal to the measure of A plus the measure of B. The second of these rules can be very useful, for example, when integrating a **function**, since it allows us to ignore any points where the function jumps around, provided that such points are isolated. A slightly jittery function is one thing; a nonmeasurable set is a very different animal. Imagine a three-dimensional shape so fantastically intricate, so jagged and crinkled, that it is impossible to measure its volume and this gives some idea of the concept of nonmeasurability. From it flow such bizarre conclusions as the **Banach-Tarski paradox**.

measure theory

The part of **mathematics** that investigates the conditions under which **integration** can be carried out. It focuses mainly on the various ways in which the size, or **measure**, of a **set** can be estimated.

measuring and weighing puzzles

Problems that involve measuring a given quantity of liquid by pouring from one vessel into others of known capacity go back to medieval times. One of the earliest to appear in print was given by Niccoló **Tartaglia** and asks for 24 ounces of balsam to be divided into three equal portions using vessels that hold 5, 11, and 13 ounces, respectively. A similar problem was posed by a fellow traveler to the young Siméon **Poisson** while on a journey. Poisson's family had tried to steer him into careers ranging from a surgeon to a lawyer, the last on the theory that he was fit for nothing better. He seemed inept at everything he did. However, he saw the solution to the measuring problem immediately and realized his true calling. Thereafter, he threw himself into mathematics and became one of the greatest mathematicians of the nineteenth century.

The classic weighing problem was proposed by Claude-Gaspar **Bachet** and entails finding the least number of weights needed to weigh any integral number of pounds from 1 to 40 pounds inclusive, when no weights are allowed in either of the two pans. The answer is 1, 3, 9, and 27 pounds. Tartaglia had previously stated the same puzzle with the condition that the weights may only be placed in one pan, in which case the solution is 1, 2, 4, 8, 16, and 32 pounds.

PUZZLE

The following is a measuring problem from Henry **Dudeney**'s *Canterbury Puzzles,* which Dudeney claimed was the most popular of the whole collection and which the reader may like to try:

> Here be a cask of fine London ale, and in my hands do I hold two measures–one of five pints, and the other of three pints. Pray show how it is possible for me to put a true pint into each of the measures.

Solutions begin on page 369.

mechanical puzzle

A puzzle, involving several objects or a single object composed of one or more movable parts, whose solution requires moving from an initial state to a predefined final state. Mechanical puzzles were first classified by Louis **Hoffmann** in *Puzzles Old and New*[171] (1893). A modified form of his scheme is shown in the table "Mechanical Puzzles" on the following page.

medial triangle

The **triangle** whose vertices are the midpoints of the sides of a given triangle (see **vertex**).

Mechanical Puzzles

Type	Subtype	Examples
Assembly	2-dimensional assembly	**Tangrams**, **T-puzzle**, jigsaw
	3-dimensional assembly (noninterlocking)	**Soma cube**
	Matchstick puzzles	
	Miscellaneous	**Puzzle rings**
Disassembly	Trick or secret opening	**Puzzle jug**
	Miscellaneous	Trick locks, keys, etc.
Interlocking solid	**Burr puzzle**	
	3-dimensional jigsaws	
	Miscellaneous	Cubes, other objects
Disentanglement and entanglement	Wire puzzles	**Chinese rings**
	String puzzles	Cat's cradle
	Miscellaneous	
Sequential movement	**Peg solitaire** (peg removal)	
	Other counter (peg rearrangement)	
	Sliding-piece puzzles	**Fifteen Puzzle**
	Miscellaneous	**Tower of Hanoi**
Puzzle vessels	**Puzzle jug**	
	Miscellaneous	Bottom-fill teapots, pitchers
Vanishment puzzle		**Get Off the Earth**
Folding	**Origami**	
	Flexagon	
Impossible figure		**Penrose stairway**
		Penrose triangle
		Impossible trident

median

1. The line from a **vertex** of a **triangle** to the midpoint of the opposite side. 2. See **mean**.

Menaechmus (c. 380–c. 320 B.C.)

A Greek mathematician, thought to have been a pupil of **Eudoxus**, who is famed for his discovery of the **conic sections** and for being the first to show that **ellipse**s, **parabola**s, and **hyperbola**s are formed by cutting a cone in a plane that is not parallel to the base. Menaechmus made his discoveries on conic sections while attempting, unsuccessfully, to solve the problem of **duplicating the cube**. It has also been suggested that he served as a tutor to Alexander the Great.

Menger sponge

A famous **fractal** solid that is the three-dimensional equivalent of the **Sierpinski carpet** (which, in turn, is the

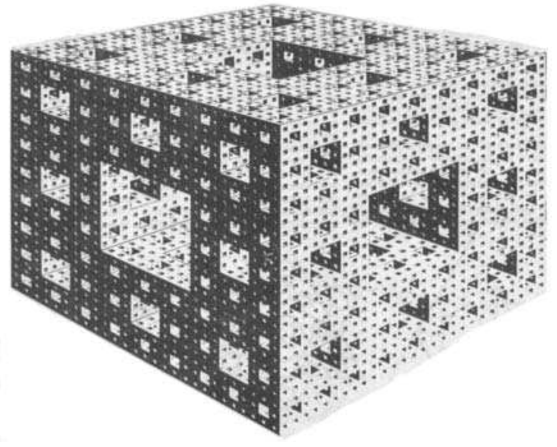

Menger sponge

two-dimensional equivalent of **Cantor dust**). To make a Menger sponge, take a cube, divide it into 27 (= 3 × 3 × 3) smaller cubes of the same size and remove the cube in the center and the six cubes that share faces with it. What's left are the eight small corner cubes and twelve small edge cubes holding them together. Now, imagine repeating this process on each of the remaining 20 cubes. Repeat it again. And again . . . ad infinitum. The Menger sponge was invented in 1926 by the Austrian mathematician Karl Menger (1902–1985).

Mersenne, Marin (1588–1648)

A French monk, philosopher, and mathematician best remembered for his work to find a formula to generate **prime number**s based on what are now known as **Mersenne number**s. However, in addition to being a mathematician, he wrote about music theory and other subjects; edited works of Euclid, Archimedes, and other Greek mathematicians; but most importantly, corresponded extensively with mathematicians and other scientists in many countries. At a time before scientific journals existed, Mersenne was at the heart of a network for information exchange.

Mersenne number

A number of the form $2^n - 1$ (one less than a power of 2), where n is a positive integer. Mersenne numbers are named after Marin **Mersenne** who wrote about them in his *Cogita Physico-Mathematica* (Physical Mathematics Knowledge 1644) and wrongly conjectured that they were prime for n = 2, 3, 5, 7, 13, 17, 19, 31, 67, and 257, and composite for $n < 257$. See also **Mersenne prime**.

Mersenne prime

A **prime number** of the form $2^p - 1$, where p is prime. A prime exponent is necessary for a **Mersenne number** to be prime but is not sufficient; for example, $2^{11} - 1 = 2{,}047 = 23 \times 89$. In fact, after an early clustering of Mersenne primes for fairly small values of p, further occurrences become increasingly rare. At the time of writing there are 40 known Mersenne primes, corresponding to values for p of 2; 3; 5; 7; 13; 17; 19; 31; 61; 89; 107; 127; 521; 607; 1,279; 2,203; 2,281; 3,217; 4,253; 4,423; 9,689; 9,941; 11,213; 19,937; 21,701; 23,209; 44,497; 86,243; 110,503; 132,049; 216,091; 756,839; 859,433; 1,257,787; 1,398,269; 2,976,221; 3,021,377; 6,972,593; 13,466,917; and 20,996,011. However, it isn't known if the current largest Mersenne prime is the fortieth in order of size because not all lower exponents have been checked. Mersenne primes rank among the largest of all known primes because they have a particularly simple test for primality, called the *Lucas-Lehmer test.*

The search for Mersenne primes has been going on for centuries. They are named after Marin **Mersenne** who, in 1644, helped the search gain wide recognition by writing to many mathematicians of his conjecture about which small exponents yield primes. Around the time that Mersenne's conjecture was finally settled, in 1947, digital computers gave a new impetus to the hunt for Mersenne primes. As time went on, larger and larger computers found many more Mersennes and, for a while the search belonged exclusively to those with the fastest computers. This changed in 1995 when the American computer scientist George Woltman began the *Great Internet Mersenne Prime Search* (*GIMPS*) by providing a database of what exponents had been checked, an efficient program based on the *Lucas-Lehmer test* that could check these numbers, and a way of reserving exponents to minimize the duplication of effort. Today GIMPS pools the combined efforts of dozens of experts and thousands of amateurs. This coordination has yielded several important results, including the discovery of the Mersennes $M_{3021377}$, $M_{2976221}$, and $M_{20996011}$ and the proof that M_{756839}, M_{859433}, and $M_{3021377}$ are the thirty-second, thirty-third, and thirty-fourth Mersennes.

meter

The basic unit of length adopted under the System International d'Unites (SI units). Over the years the definition of the meter has changed several times. Throughout all these definition changes the length of the meter hasn't changed, but the precision by which it is measured was improved. In 1793, the meter was defined to be 1/10,000,000 of the distance from the pole to the equator. During the nineteenth century, the definition was in terms of the length of standard bars of platinum kept under controlled conditions. In 1983, the current definition was adopted of the distance traveled by light in a vacuum in 1/299,792,458 of a second.

method of exhaustion

Finding an **area** by approximating it by the areas of a sequence of **polygon**s; for example, filling up the interior of a circle by inscribing polygons with more and more sides.

metric

Any **function** $d(x, y)$ that describes the *distance* between two points. Distance is formally defined as a single number with the following properties: (1) $d(x, y) = 0$ if and only if $x = y$; (2) $d(x, y) = d(y, x)$; (3) $d(x, y) + d(y, z) \geq d(x, z)$ (the triangle inequality).

metric space

A **set** that has a **metric**; in other words, a kind of space in which the concept of distance has meaning. Compare with **topological space**.

metrizable

For a **topological space**, the property that there exists a **metric** compatible with the topology. To say that a topological space is metrizable is to treat it as a **metric space**, but without distinguishing any specific or preferred distance function.

Michell, John (1724–1793)

An English natural philosopher and clergyman, educated at Queen's College, Cambridge, who discovered that the force between magnetic poles varies as $1/r^2$. In 1767, Michell became rector of Thornhill, now a suburb of Dewsbury, Yorkshire. Sometime in the early 1770s, he played music with the great astronomer William Herschel and gave Herschel his first telescope. In 1784, Michell deduced the existence of what are now called black holes from Newton's corpuscular theory of light, and suggested that some stars might have dark companions. He also devised and built the torsion balance for determining the universal gravitational constant G, but died before having the chance to use it. The balance was passed to the Cambridge physicist F. J. H. Wollaston, and from him to Henry Cavendish, who used it at his house in London and is often, mistakenly, identified as its inventor.

midpoint

The point M is the midpoint of line segment AB if $AM = MB$. That is, M is halfway between A and B.

mile

A measure of distance, the name of which is an abbreviation of the Latin *mille passes* or "one thousand paces." Since the paces (one step with each foot) of the Roman Army were supposed to be two steps, each 2.5 feet long, 1,000 paces is very close to the length now called a statute mile (5,280 feet). A *nautical mile* was developed to be a distance equal to 1 minute of arc ($^1/_{60}$ of a degree) distance along a great circle and is equal to 6,076 feet.

million

A thousand thousand, 1,000,000, or 10^6. The word comes from the Latin *mille* for "thousand" (which is also the root for ***mile*** and *millennium*) and the suffix *ion* that implies "large" or "great;" thus a million is literally a "great thousand." Although *million* seems to have come into use as early as the middle of the fourteenth century, most mathematicians would use the phrase "thousand thousands" to avoid confusion and it was not until the 1700s that "million" caught on. It appears in the King James Version of the Bible (Genesis 24:60) and in Shakespeare (*Hamlet* act II scene II)–"for the play, I remember, pleased not the million."

minimal prime

A **prime number** that is a substring of another prime when written in **base** 10. A string *a* is a substring of another string *b*, if *a* can be obtained from *b* by deleting zero or more of the characters in *b*. For example, 392 is a substring of 639,802. The minimal primes are:

2; 3; 5; 7; 11; 19; 41; 61; 89; 409; 449; 499; 881; 991; 6,469; 6,949; 9,001; 9,049; 9,649; 9,949; 60,649; 666,649; 946,669; 60,000,049; 66,000,049; 66,600,049

minimal surface

A surface that, bounded by a given closed curve or curves, has the smallest possible area. A minimal surface has a mean **curvature** of zero. Finding and classifying minimal surfaces and proving that certain surfaces are minimal have been major mathematical problems for over 200 years. If the closed curve is planar then the solution is trivial; for example, the minimal surface bounded by a circle is just a disk. But the problem becomes much more difficult if the bounding curve is nonplanar–in other words, is allowed to move up and down in the third dimension. The first nontrivial examples of minimal surfaces, the **catenoid** and the **helicoid**, were discovered by the French geometer and engineer Jean Meusnier (1754–1793) in 1776, but there was then a gap of almost 60 years before the German Heinrich Scherk found some more. In 1873 the Belgian physicist Joseph **Plateau** carried out experiments that led him to conjecture that soap **bubble**s and soap films always form minimal surfaces. Proving mathematically this was true became known as the **Plateau problem**. Most minimal surfaces are extremely hard to construct and visualize, in part because the majority of them are self-intersecting. However, the development of high-performance computer graphics has provided mathematicians with a powerful tool and the last couple of decades have seen a huge increase in the number of such surfaces that have been defined and investigated.

minimax theorem

A theorem that says there is always a rational solution to a precisely defined conflict between two people whose interests are completely opposite. It is rational in that both parties can convince themselves that they can't expect to do any better, given the nature of the conflict.

minimum

The smallest of a set of values.

minimal surface A Scherk Surface—a type of minimal surface—portrayed as a membrane. *Anders Sandberg*

Minkowski, Hermann (1864–1909)

A German mathematician, born in Lithuania, who played an important part in the early development of **relativity theory**. Minkowski was the first to realize that the work of Hendrik Lorentz and Albert **Einstein** could be best understood if space and time, formerly thought to be separate entities, were treated as part of a four-dimensional **space-time** with a **non-Euclidean geometry**. The concept of the space-time continuum, which provided a framework for all later mathematical work in relativity, appeared in Minkowski's book *Raum und Zeit* (Space and Time 1907). From 1896 to 1902, Minkowski taught at the Zurich Federal Institute of Technology when Einstein was a student. In fact, Einstein attended several of the courses he gave but didn't create a good impression at the time. Minkowski described him as a "lazy dog" who "never bothered about mathematics at all." In 1902, Minkowski accepted a chair at the University of Göttingen, where he stayed for the rest of his life. His main interest was in pure mathematics, including number theory and geometry, and it was through his understanding of the more abstract side of mathematics and geometry in more than three dimensions that he developed the idea of four-dimensional space-time.

Minkowski space

A finite-dimensional **vector space**, especially a four-dimensional one, together with an indefinite inner product with one positive or timelike direction and many negative or spacelike directions. In particular, Minkowski space is ordinary **space-time** in the special **relativity theory**.

minor axis

The smallest **chord** of an **ellipse**.

minute

(1) One sixtieth of an hour or of a **degree** of arc or angle. (2) One sixtieth of an hour. As a measure of both time and angle, it equals 60 **seconds**.

mirror reversal problem

Why does a mirror reverse right and left, but not up and down? This question crops up perennially in the letter and query columns of magazines and newspapers. It was the inspiration for Lewis **Carroll**'s *Alice through the Looking Glass*. Alice Raikes (not to be confused with Alice Liddell, after whom the fictional Alice was modeled) was another of Carroll's young friends. On one occasion, in 1868, Carroll put an orange in her right hand and then asked her to stand in front of a mirror and say which hand showed the reflection of the orange. She said the left hand and Carroll asked her to explain. She finally replied "If I was on the other side of the glass, wouldn't the orange still be in my right hand?" Carroll said this was the best answer he'd had and later said it gave him the idea for his book.

Others have struggled harder but not always convincingly to explain the phenomenon, appealing variously to gravity, the psychology of perception, and philosophy. Why *does* a mirror reverse right and left, but not up and down? A frequently given answer is that a mirror doesn't reverse right and left. It reverses *front and back*. This is certainly true: the looking glass you is facing in the opposite

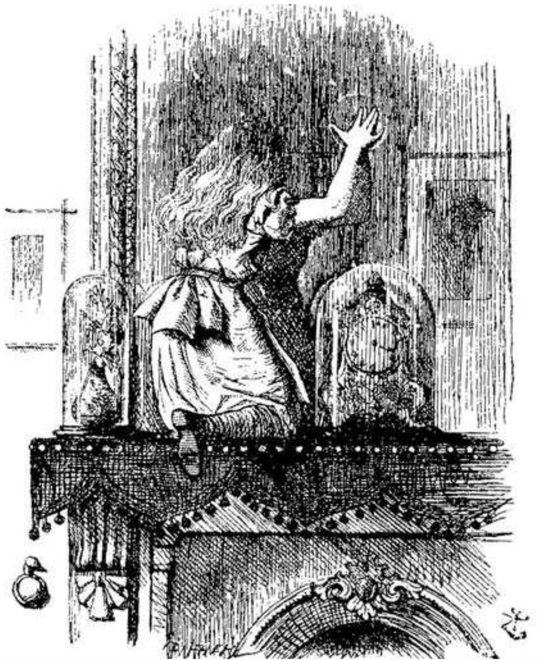

mirror reversal problem A looking-glass world beckons the curious Alice in Lewis Carroll's book *Alice through the Looking Glass.*

direction to the "real" you. But this short, crisp explanation doesn't completely dispel the mystery. If you imagine that the mirror is not there and that instead you are looking at a flesh-and-blood twin of yourself, that twin is differently handed. If you have a watch on your left wrist, the person you are facing has his/her watch on the right wrist. The mirror *has* done a left-right swap, surely! At any rate, something has happened to left and right that hasn't happened to up and down. To be more convinced of this, hold this book up to the mirror and try to read it. If no left-right swap has happened, why is the reflected writing so hard to read? First, remember that you are only looking at an image! The mirror hasn't (Carrollian fantasies aside) created some*thing* of opposite handedness. Secondly, appreciate how the writing appears in the mirror's frame of reference. This is easy to do by looking at the writing from the other side of the page (i.e., back to front, thus undoing the back to front reversal caused by the reflection). From the mirror's point of view the writing looks perfectly normal.

missing dollar problem

A version of this problem first appeared in R. M. Abraham's *Diversions and Pastimes* in 1933.[2] See also **nine nooms paradox**.

PUZZLE

Three people dine at a restaurant and receive a total bill for $30. They agree to split the amount equally and pay $10 each. The waiter hands the bill and the $30 to the manager, who realizes there has been a mistake and the correct charge should be only $25. He gives the waiter five $1 bills to return to the customers, with the restaurant's apologies. However, the waiter is dishonest. He pockets $2, and gives back only $3 to the customers. So, each of the three customers has paid $9 and the waiter has stolen $2 making a total of $29. But the original bill was for $30. Where has the missing dollar gone?

Solutions begin on page 369.

Mittag-Leffler, (Magnus) Gösta (1846–1927)

A Swedish mathematician who, in 1882, founded the international journal *Acta Mathematica,* and was its chief editor for 45 years. He studied in Paris under Charles **Hermite** and in Berlin under Karl **Weierstrass**, and made significant contributions to **analysis**. His best known work concerned the analytic representation of a one-valued **function**s and culminated in the *Mittag-Leffler theorem.* Since he took a special interest in Georg **Cantor**'s discoveries, much of Cantor's work was published in *Acta Mathematica.* Inscribed on the mantlepiece of Mittag-Leffler's home–now a research institute–in Djursholm is the epitaph: "Number is the beginning and end of thought. Thought gave birth to number but reaches not beyond. ML 1903"

mixed strategy

In **game theory**, a **strategy** that uses **random**ness by employing different actions in identical circumstances with different probabilities.

Möbius band

A simple and wonderfully entertaining two-dimensional object, also known as the Möbius strip, that has only one surface and one edge. It is named after the German mathematician and theoretical astronomer August Ferdinand Möbius (1790–1868), who discovered it in September 1858, although his compatriot and fellow mathematician Johann Benedict Listing (1808–1882) independently devised the same object in July 1858. Making a Möbius band is simple: take an ordinary sheet of typing paper, cut an 11″ × 1″ rectangle, bring the two long ends together, twist one of the ends 180°, and tape the two ends together. To prove that the band is single-sided, take a pen and start drawing a line around the band's circumference. When drawing the line, never take the pen off the paper; just

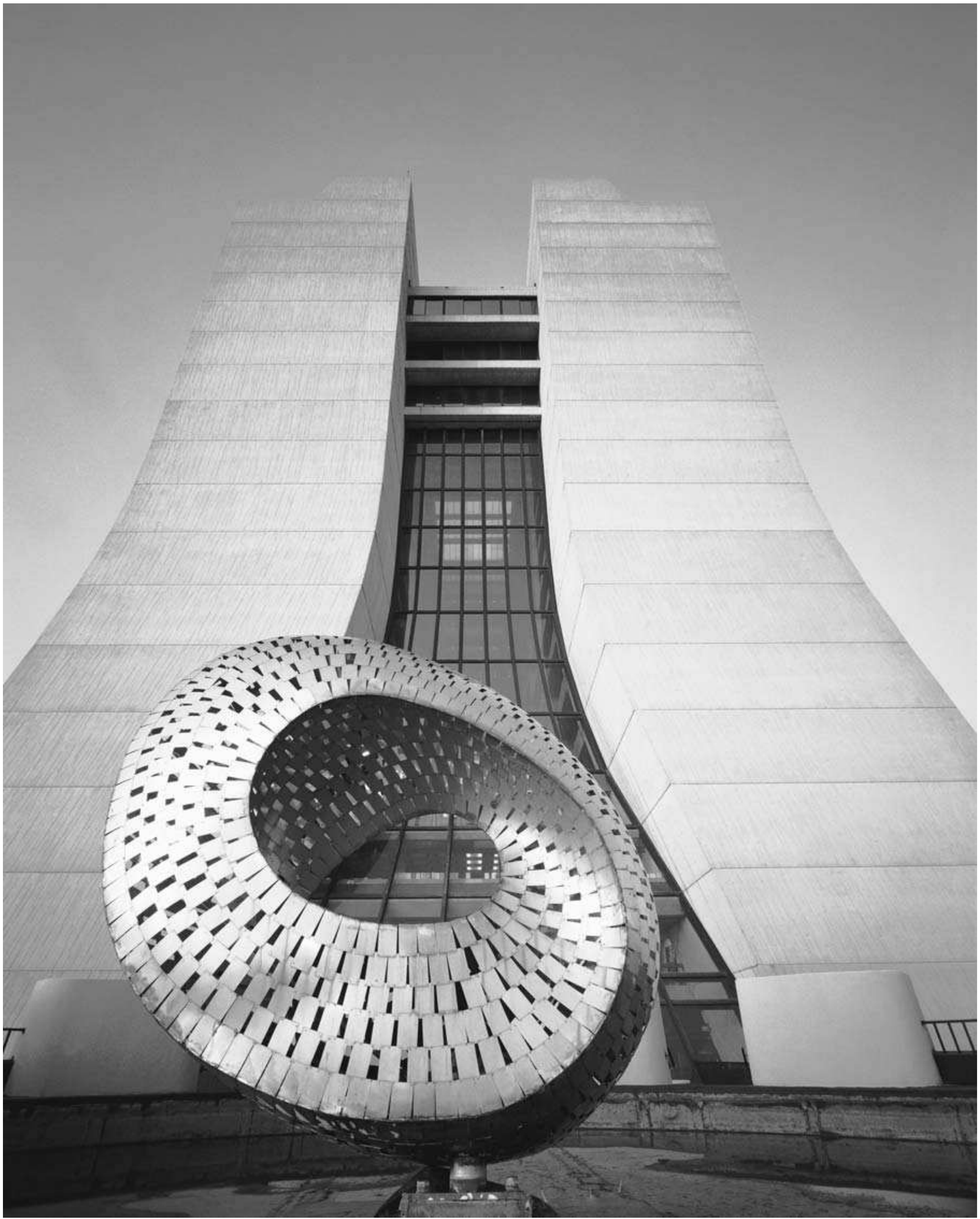

Möbius band A sculpture of a Möbius band outside the Fermi National Accelerator Laboratory. *FNAL* keep drawing the line until the starting point is reached. Once you are finished, look at both sides: there should be a line on both sides, thus proving that it is all the same side because you never took the pen off the paper.

The Möbius band has a lot of curious properties. If you cut down the middle of the band, instead of getting two separate strips, it becomes one long strip with two half-twists in it. If you cut this one down the middle, you get two strips wound around each other. Alternatively, if you cut along the band, about a third of the way in from the edge, you will get two strips; one is a thinner Möbius band, the other is a long strip with two half-twists in it. Other interesting combinations of strips can be obtained by making Möbius bands with two or more flips in them instead of one. Cutting a Möbius band, giving it extra twists, and reconnecting the ends produces unexpected figures called *paradromic rings.*

The Möbius band has provided inspiration both for sculptures and for graphical art. M. C. **Escher** was especially fond of it and based many of his lithographs on it. It is also a recurrent feature in science fiction stories, such as Arthur C. Clarke's *The Wall of Darkness.* A common fictional theme is that our universe might be some kind of generalized Möbius band. There have been technical applications; giant Möbius bands have been used as conveyor belts (to make them last longer, since "each side" gets the same amount of wear) and as continuous-loop recording tapes (to double the playing time).

A closely related strange geometrical object is the **Klein bottle**, which can be produced by gluing two Möbius bands together along their edges; however, this can't be done in ordinary three-dimensional Euclidean space without creating self-intersections.[116]

mode

The most frequently occurring value in a sequence of numbers.

model of computation

An idealized version of a computing device that usually has some simplifications such as infinite memory. A **Turing machine** and the **lambda calculus** are models of computation.

model theory

The study of mathematical structures that satisfy a particular set of **axiom**s, especially in the field of **logic**.

modulo

The integers a and b are said to be congruent modulo m if $a - b$ is divisible by m.

Moiré pattern

A radiating curved pattern created when two repetitive patterns overlap and interfere with one another. A Moiré pattern is seen, for example, when someone on TV wears a herringbone jacket. *Moiré* is the French word for "silk", and *silk moiré,* introduced from China to France in 1754, is the fabric that shows the familiar shifting patterns.

Moivre, Abraham de

See **de Moivre, Abraham**.

monad

A central concept in the Pythagorean worldview (see **Pythagoras of Samos**), in which it is regarded as the first thing that came into existence. Following the monad came, in order: the *dyad,* numbers, lines, two-dimensional entities, three-dimensional entities, bodies, the four elements (earth, air, fire, and water), and the rest of the world. The monad plays a similarly fundamental role in the metaphysics of Leibnitz as an indivisible, impenetrable unit of mental experience. It also has several different technical meanings in modern mathematics. For example, in **nonstandard analysis**, a monad consists of all those numbers infinitesimally closer to a given number. The word comes from the Latin *monas* (single) and Greek *monos* (unit).

Monge, Gaspard (1746–1818)

A French mathematician and physicist who put descriptive geometry, introduced by Albrecht **Dürer**, on a firm mathematical footing. He became professor of mathematics at Mézières (1768) and of hydraulics at the Lycée in Paris (1780), and published his groundbreaking treatise on the application of geometry to the arts of construction in 1795.

monkeys and typewriters

Six monkeys pounding away on typewriters would, by pure chance, if given enough time, be bound to write out all the works held in the British Library (or in all other libraries, for that matter). This idea was first suggested by the biologist Julian Huxley (1887–1975); it was discussed by the physicist James Jeans (1877–1946) in his *Mysterious Universe* (1930); and it has been restated in various forms over the years, in terms of chimpanzees, Shakespeare's sonnets, and the like. It was also the subject of Russell Maloney's short story "Inflexible Logic," first published in the *New Yorker* magazine (1940) and reprinted in Clifton **Fadiman**'s *Fantasia Mathematica,*[96] which tells the tragic tale of what happens when the fantastically improbable comes true. See also **Universal Library**.

monochromatic triangle

A triangle whose vertices are all colored the same (see **vertex**).

monomial

An algebraic expression that consists of just one term.

monotonic

The property of a **function** that is always strictly increasing or strictly decreasing, but never both.

Monster curve

See **Peano curve**.

Monster group

The largest, most fascinating, and most mysterious of the so-called sporadic **group**s; it was constructed by Robert Griess at Princeton in 1982, having been predicted to exist by him and Bernd Fischer in 1973, and was named the Monster by John **Conway**. Think of the Monster group as a preposterous snowflake with more than 1,050 symmetries that exists in a space of 196,883 dimensions. It contains the following number of elements:

$$2^{46} \times 3^{20} \times 5^{9} \times 7^{6} \times 11^{2} \times 13^{3} \times 17 \times 19 \times 23 \times 29 \times 31 \times 41 \times 47 \times 59 \times 71$$
$$= 808{,}017{,}424{,}794{,}512{,}875{,}886{,}459{,}904{,}961{,}710{,}757{,}005{,}754{,}368{,}000{,}000{,}000$$
$$\approx 8 \times 10^{53} \text{ (more than the number of quarks in the Sun).}$$

Despite these impressive credentials, however, it is still classified as a *simple group,* meaning that it doesn't have any normal subgroups other than the identity element and itself. All 26 simple groups have now been classified and the Monster is far and away the biggest. At first, it seemed that the Monster was just a curiosity–a Guinness Book record of pure math. Its only "useful" application seemed to be to give the best way for packing spheres in 24 dimensions! In ordinary three-dimensional space (and also four and five dimensions), the grocer's way of stacking oranges in a hexagonal lattice is thought to be the tightest possible (see **Kepler's conjecture**). But as the number of dimensions increases, the optimal packing method changes. A 24-dimensional grocer would get the most efficient arrangement of his 24-dimensional oranges by using the same symmetry as that of the Monster. This is unlikely to be immediately useful. Much more interesting, however, is the connection that has been found between the symmetry of the Monster and one of the most promising unifying theories in physics–**string theory**–which has been revealed by the **Monstrous Moonshine conjecture**.[70]

Monstrous Moonshine conjecture

An outrageous idea that stemmed from an observation made by John McKay of Concordia University in 1978. McKay was leafing through a table of abstruse mathematical data, giving possible values for coefficients of the *j*-function of certain elliptic curves, when he noticed the number 196,884 in the expression $j(q) = q - 1 + 196884q + 21493760q^2 + \ldots$. In a moment of inspiration, he recognized this number as being one more than the number of dimensions in which the **Monster group** can be most simply represented. Looking into this "coincidence" more closely, he found that it was no coincidence at all. In fact, *all* the coefficients of the *j*-function were simple combinations of the degrees of possible representations of the Monster. This pointed to some deep connection between two seemingly unrelated areas of mathematics. On the one hand were the coefficients of what is called an elliptic modular function–exactly the kind of function that would play a key role in the proof of **Fermat's last theorem**. On the other was the number of dimensions, and combinations of degrees, of a crystal lattice whose symmetry rotations and reflections formed the Monster. Subsequently, McKay and a few other mathematicians, including John **Conway** and Simon Norton, drew out the link between elliptic modular functions and the Monster in a proposition christened, because of its fantastic nature, the Monstrous Moonshine. In 1998, this conjecture was proved by Richard Borcherds (a former student of Conway's) at the University of California at Berkeley. Astonishingly, Borcherds's proof reveals a deep relationship between elliptic curves, the Monster Group, and **string theory**–the most promising theory on offer to unify our understanding of nature at the subatomic level. Borcherds showed that the Monster is the group of symmetries of 26-dimensional strings expressed in a form known as *vertex algebra.* Some people believe the connection may run even deeper and that Monstrous Moonshine may hold clues to the very existence of the reality in which we live.[69]

Monte Carlo method

A method of estimating the true value of a quantity by carrying out a lot of **random** samples. For example, suppose we want to know the probability of getting a double six when we roll two dice. We could roll a pair of dice a thousand times, and count how many times, *n,* a double six came up; the estimated probability would then be *n*/1,000. A famous example of using the Monte Carlo method is to calculate **pi**. Get a computer to generate two random numbers x and y, each in the range –1 to 1, so that the point (x, y) lies somewhere randomly inside a square of side 2 units. Do this thousands of times, and count up what proportion of the points also lie inside the circle that inscribes

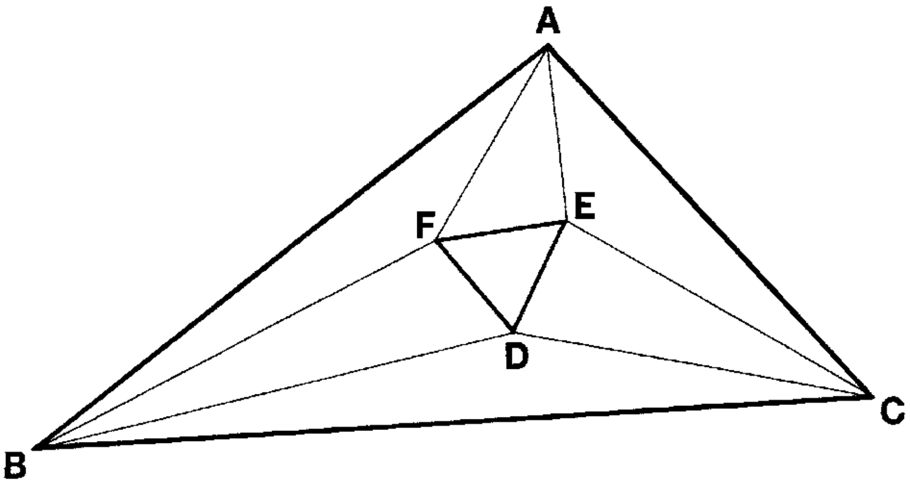

Morley's miracle The trisectors of the angles of triangle *ABC* meet the corners of a inner equilateral triangle *FED* known as Morley's triangle.

the square (you can tell whether a point does or not by working out whether $x^2 + y^2 < 1$). The proportion inside the circle is an approximation to $\pi/4$ (because the circle has area π but the square has area 4); millions of points are needed to obtain a good estimate.

This method was developed by researchers working on the Manhattan Project during World War II. To answer some of their scientific questions, they would repeatedly sample from their best estimates of the partial results, then apply the math they knew to the interactions and study the range of results. This process, which they named after the famous Monaco casino town of Monte Carlo, was created by John **von Neumann** and Stanislaw **Ulam**. The term and a description of the method seems not to have been published until some time after the war.

Montucla, Jean Etienne (1725–1799)

A French writer, mathematician, and scientist, who wrote several important early works on the history of mathematics. His *Histoire des mathématiques* (1758) was published in two volumes, the first of which covers the subject from ancient times to 1700, while the second is entirely devoted to seventeenth-century mathematics. It is considered the first attempt at a history of mathematical ideas and problems, in contrast to earlier works that were mostly lists of names, titles, and dates. Montucla had intended to produce a third volume covering the first half of the eighteenth century but the amount of new developments that had appeared during this time, and the difficulties of putting recent work into its historical context, led him to abandon this aim. A few years later he published another text for which he is famed–a new, greatly expanded and improved edition of Jacques **Ozanam**'s *Récréations mathématiques et physiques* (1778). Montucla's edition was particularly influential in popularizing geometric **dissection** problems. Charles Hutton translated it into English in 1803 and Riddle's edition was published in 1844, called *Recreations in science and natural philosophy.*

Monty Hall problem

A puzzle in probability that was inspired by the American game show *Let's Make a Deal,* hosted by Monty Hall. In its original form it goes like this: at the end of the show, you, the player, are shown three doors. Behind one of them is a new car, behind the other two are goats. Monty knows where the car is, but you don't. You choose a door. Before that door is opened however, Monty opens one of the two other doors with a goat behind it. He then gives you the option of switching to the other closed door. Should you switch or stick? At first glance, it seems as if it shouldn't make any difference. But the answer is surprising.

Suppose you stick. Your original choice made when all three doors were equally likely gives you a probability of winning the car of $\frac{1}{3}$. Now suppose you switch. In other words, you choose a door, wait for Monty to expose a goat, then switch to the other remaining door. This means that you win if the door you chose to begin with had a goat behind it. The odds that your initial choice had a goat is two thirds, so you are twice as likely to win the car if you switch! This can be hard to grasp. To make it easier, suppose there are 100 doors to choose from, but still only one car. You pick a door, Monty opens 98 that have goats behind them, then he gives you the option of switching to the other remaining closed door. Should you? Of course–it's almost certain that the car is behind the other door, and very unlikely that it's behind your original choice.

In a generalization of the original problem there are *n* doors. In the first step, you choose a door. Monty then opens some other door that's a loser. If you want, you may then switch your choice to another door. Monty will then open an as yet unopened losing door, different from your current preference. Then you may switch again, and so on. This carries on until there are only two unopened doors left: your current choice and another one. How many times should you switch, and when, if at all? The answer is: stick all the way through with your first choice but then switch at the very end.

In another variation of the problem, consider that in the actual game show there were two contestants. Both of them were allowed to pick a door but not the same one. Monty then eliminated a player with a goat behind his door (if both players had a goat, one was eliminated randomly, without letting either player know about it), opened the loser's door, and then offered the remaining player a chance to switch. Should the remaining player switch? The answer is *no.* The reason: a switcher in this game will lose if and only if either of two initial choices of the two contestants was correct. How likely is that? Two-thirds. A sticker will win in those $^2/_3$ of the cases. So stickers will win twice as often as switchers.

Morley's miracle

A remarkable theorem, discovered in 1899, by Frank Morley, then professor of Mathematics at Haverford College. Take any triangle. Mark the three points that are the intersections of adjacent angle trisectors. Then, no matter what triangle you start with, these three points will form an equilateral triangle. That such a simple and elegant result was not known to the ancient Greeks may be because it is quite hard to prove.

One of the interesting auxiliary results of some of the proofs is that the side of the equilateral triangle is equal to $8r \sin(A/3) \sin(B/3) \sin(C/3)$, where *A, B,* and *C* are the angles of the larger triangle, and *r* is the radius of the circumcircle. A surprise awaits anyone who takes the intersections of the exterior, as well as the interior, angle trisectors. In addition to the interior equilateral triangle, four exterior equilateral triangles appear, three of which have sides that are extensions of a central triangle.

Moscow papyrus

See **Rhind papyrus**.

mousetrap

See **Cayley's mousetrap**.

moving sofa problem

In Douglas Adams's book *Dirk Gently's Holistic Detective Agency,* the character Richard MacDuff says at one point, "It would be really useful to know before you buy a piece of furniture whether it's actually going to fit up the stairs or around the corner." Mathematicians call this the moving sofa problem and it has been tackled in various forms over the past few decades. One version of it, formulated by Leo Moser in 1966, asks: What is the largest sofa (in terms of area) that can be moved around a right-angled corner in a hallway of unit width? The sofa can be any shape and doesn't even have to resemble a piece of furniture! The question simply asks for the biggest, unbendable area that can be maneuvered around the corner. Several different approaches suggest that the answer is about 2.21 square units. Variations on the problem involve negotiating pianos and other items around different types of bends and passageways.[133]

Mrs. Perkins's quilt

A square **dissection** problem first posed by Henry **Dudeney** in his *Amusements in Mathematics* (1917):[88]

PUZZLE

It will be seen that in this case the square patchwork quilt is built up of 169 pieces. The puzzle is to find the smallest possible number of square portions of which the quilt could be composed and show how they might be joined together. Or, to put it the reverse way, divide the quilt into as few square portions as possible by merely cutting the stitches.

Solutions begin on page 369.

Dudeney's problem can be generalized to the dissection of a square of side *n* into a number S_n of smaller squares. Unlike a perfect **squaring the square** problem, the smaller squares needn't be all different sizes. In addition, only prime dissections are considered so that patterns that can be dissected on lower order squares aren't allowed. The smallest number of relatively prime dissections of an $n \times n$ quilt for $n = 1, 2, \ldots$, are 1, 4, 6, 7, 8, 9, 9, 10, 10, 11, 11, 11, 11, 12, ...[66]

Müller-Lyer illusion

A **distortion illusion** in which the orientation of arrowheads makes one line segment look longer than another.

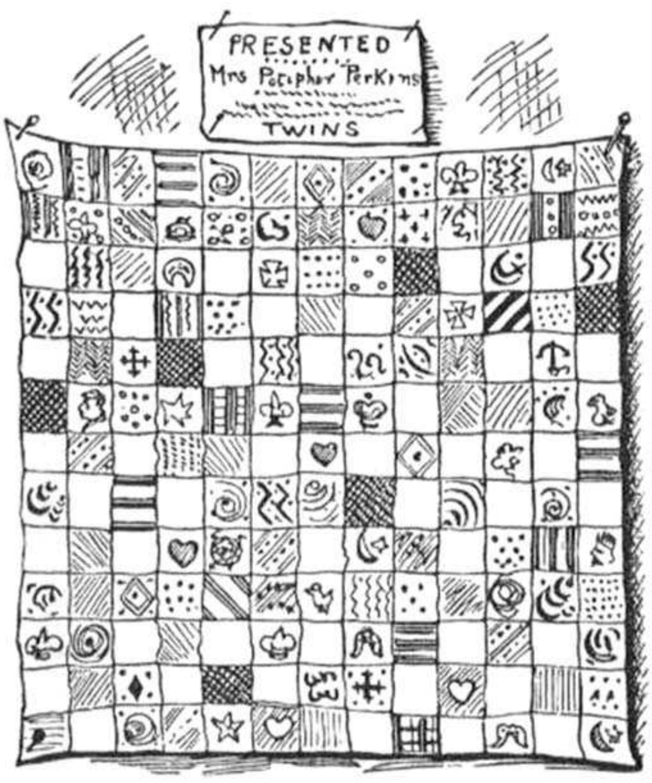

Mrs. Perkins's quilt Dudeney's drawing of one of his most famous puzzles.

multigrade

A set of equations in which the sums of powers of two different sets of numbers are the same for several different exponents. The simplest example is:

$$1 + 6 + 8 = 2 + 4 + 9$$
$$1^2 + 6^2 + 8^2 = 2^2 + 4^2 + 9^2$$

Another multigrade is:

$$1 + 8 + 10 + 17 = 36 = 2 + 5 + 13 + 16$$
$$1^2 + 8^2 + 10^2 + 17^2 = 454 = 2^2 + 5^2 + 13^2 + 16^2$$
$$1^3 + 8^3 + 10^3 + 17^3 = 6426 = 2^3 + 5^3 + 13^3 + 16^3$$

Remarkably, if any integer is added to all the terms of a multigrade it will still hold. Adding 1 to the example above, gives the multigrade (2, 9, 11, 18); (3, 6, 14, 17) (n = 1, 2, 3). Some high-order multigrades include: (1, 50, 57, 15, 22, 71); (2, 45, 61, 11, 27, 70); (5, 37, 66, 6, 35, 67) (n = 1, 2, 3, 4, 5), and (1, 9, 25, 51, 75, 79, 107, 129, 131, 157, 159, 173); (3, 15, 19, 43, 89, 93, 97, 137, 139, 141, 167, 171) (n = 1, 3, 5, 7, 9, 11, 13).

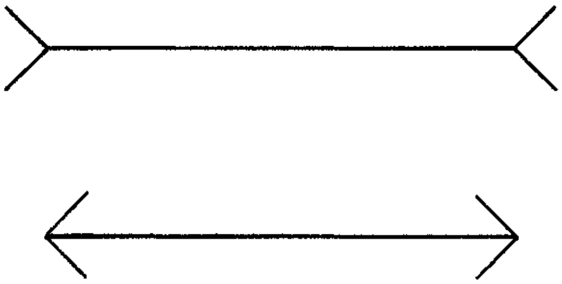

Müller–Lyer illusion Which of the horizontal lines is longer?

multiple

The integer *b* is a multiple of the integer *a* if there is an integer *d* such that $b = da$.

multiplication

A **binary operation** that is the equivalent of repeated addition and the **inverse** of **division**.

Mydorge, Claude (1585–1647)

A French mathematician who trained as a lawyer but was wealthy enough that he didn't have to work for a living. He was interested in mathematical puzzles and his book *Examen du livre des récréations mathématiques* (Study of the book of recreational mathematics, 1630) formed the basis for later works such as that by Denis Henrion (1659). Mydorge edited *Récréations mathématique* and left an unpublished manuscript of over 1,000 geometric problems and their solutions. He was also interested in optics and made a large number of instruments for his close friend René **Descartes**; the two shared a strong interest in explaining vision and the instruments and lenses were designed to help test their theories.

myriad

A term which today is normally synonymous with "very large number." Its origins go back to the Greek word *murious,* meaning "uncountable". The plural of this, *murioi,* evolved into the Latin *myriad,* which the Romans used to represent ten thousand. *Myriapod* is a general name for any many-legged anthropod, such as a millipede or a centipede

N

Nagel point

A point in a triangle where the lines from the vertices (see **vertex**) to the points of contact of the opposite sides with the excircles to those sides meet. (An excircle touches one side of a triangle and also touches the lines extended from the other two sides.)

nano-

Prefix for billionth (10^{-9}), from the Greek *nanos,* meaning "dwarf."

Napier, John (1550–1617)

A Scottish mathematician and theological writer who invented **logarithms** and wrote *Mirifici logarithmorum canonis descriptio* (The description of the wonderful canon of logarithms, 1614), which contains the first logarithmic table and the first use of the word *logarithm.* He also introduced the decimal point in writing numbers. His *Rabdologiae* (1617) describes various shortcuts for carrying out arithmetical calculations. One method of multiplication uses a system of numbered rods called *Napier's rods,* or *Napier's bones*–a major improvement on the ancient system of counters then in use. In 1619, after Napier's death, his *Mirifici logarithmorum canonis constructio,* which gave the method of construction of his logarithms, was published by his son Robert and edited by Henry **Briggs**.

Napoleon Bonaparte (1769–1821)

> *The advancement and perfection of mathematics are intimately connected with the prosperity of the state.*

Emperor of France and a very good amateur mathematician, having excelled in this subject as a student at school and at military college. Even after becoming first consul he was proud of his membership in the Institute de France (the nation's leading scientific society), and was close friends with several mathematicians and scientists, including Joseph **Fourier**, Gaspard **Monge**, Pierre Simon **Laplace**, Chaptal, and Berthollet. Indeed, in his grand expedition to Egypt in 1798 Napoleon brought along (in addition to 35,000 troops) over 150 experts in various fields, among them Monge, Fourier, and Berthollet, not to mention a complete *encyclopedie vivante* with libraries and instruments. One result of the expedition was that Fourier served for a time as the governor of lower Egypt. Likewise Laplace (who interviewed the young Napoleon for admission to the artillery) received titles and high office as a result of his friendship with Bonaparte. However, Laplace was relieved of his duties as the minister of the interior after only six weeks, and Napoleon later commented that Laplace had "sought subtleties everywhere, had only doubtful ideas, and carried the spirit of the infinitely small into administration." The most famous exchange between these two men occurred after Laplace had given Napoleon a copy of his great work, *Mecanique Celeste.* Napoleon looked it over, and remarked that in this massive volume about the universe there was not a single mention of **God**. Laplace replied "Sire, I had no need of that hypothesis." Regarding the idea that Napoleon might have discovered what is now called *Napoleon's theorem* (if equilateral triangles are constructed on the sides of any triangle (all outward or all inward), the centers of these equilateral triangles themselves form an equilateral triangle), Harold **Coxeter** and Samuel Greitzer have said that "The possibility of [Napoleon] knowing enough geometry for this feat is as questionable as the possibility of his knowing enough English to compose the famous palindrome, ABLE WAS I ERE I SAW ELBA."

nappe

Either of the two parts into which a **cone** is divided by the **vertex**. "Nappe" is the French for "tablecloth," which in turn comes from the Latin *mappa* (napkin).

narcissistic number

Also known as an *Armstrong number* or a *plus perfect number,* an *n*-digit number equal to the sum of its digits raised to the *n*th power. For instance, 371 is narcissistic because $3^3 + 7^3 + 1^3 = 371$, and 9474 is narcissistic because $9^4 + 4^4 + 7^4 + 4^4 = 9474$. The smallest narcissistic number of more than one digit is $153 = 1^3 + 5^3 + 3^3$. The largest narcissistic number (in base 10) is 115,132,219,018,763,992,565,095,597,973,971,522,401, which is the sum of the thirty-ninth powers of its digits. The reason there are no larger numbers is related to the fact that, as the number of digits increases, more and more nines are required to get a sum that has *n* digits. For example, $10^{70} - 1$ is a number consisting of 70 nines in a row, and the sum of the seventieth powers of its digits is $70 \times 9^{70} \cong 4.386051 \times 10^{68}$, which is only 69 digits long. So there is no way any 70-digit number can be equal to the sum of the seventieth powers of its digits. The reason we see the last number occur at 39 digits is because, as the limit is approached, the number of big digits like eights and nines has to increase to make sure the sum will be big enough, but this means that there are a lot fewer combinations of digits to choose from.

Nash, John Forbes Jr. (1928–)

An American mathematician not very accurately portrayed in the Oscar-winning film *A Beautiful Mind* (2001), loosely based on the biography of the same name by Sylvia Nasar (1998).[229] Nash, who worked in **game theory** and **differential geometry**, shared the 1994 Nobel prize for economics with two other game theorists, Reinhard Selten and John Harsanyi. After a promising start to his mathematical career, Nash began to suffer from schizophrenia around the age of 30 and battled with the illness for the next quarter of a century. His Ph.D. dissertation, entitled "Non-cooperative Games," contained the definition and properties of what would later be called **Nash equilibrium** and the basis of the work that, 44 years later, would make him a Nobelist. Between 1966 and 1996, Nash published nothing. However, as his mental health slowly began to improve in the mid-1990s, his ability to tackle mathematical problems returned, and he also became interested in computer programming.

Nash equilibrium

In **game theory**, a pair of strategies (see **strategy**) for a game such that neither player can improve his outcome by changing his strategy. A Nash equilibrium sometimes takes the form of a **saddle** structure. In other cases, when a strategy is at a Nash equilibrium with itself, the strategy resembles an evolutionary stable strategy.

natural logarithm

Also called a *Naperian logarith,* a **logarithm** to base **e**. For example, $\log_e 10$ (also written as 1*n* 10) is approximately 2.30258.

natural number

A number used for counting: 1, 2, 3, ... The debate about whether zero should also be included as a natural number has been going on for hundreds of years, and there's no general agreement even today. To avoid confusion, 0, 1, 2, 3, . . . , are often referred to as *nonnegative integers* or *whole numbers,* while 1, 2, 3, . . . , are called *positive integers.* Adding or multiplying natural numbers always produces other natural numbers. However, subtracting them can produce zero or negative integers, while dividing them produces **rational numbers**. An important property of the natural numbers is that they are *well-ordered,* in other words, every set of natural numbers has a smallest element. The deeper properties of the natural numbers, such as the distribution of **prime numbers**, are studied in **number theory**. Natural numbers can be used for two purposes: to describe the position of an element in an ordered sequence, which is generalized by the concept of **ordinal number**, and to specify the size of a finite set, which is generalized by the concept of **cardinal number**. In the finite world, these two concepts coincide; however, they differ when it comes to infinite sets (see **infinity**).

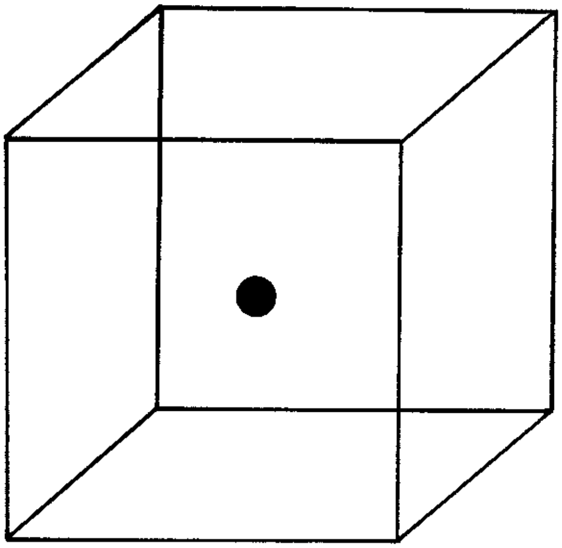

Necker cube Is the circle on the front face of the cube or the back?

Necker cube

A classic example of an **ambiguous figure**. In 1832, the Swiss crystallographer Louis Necker noticed, while examining crystals, that three-dimensional objects can fluctuate in appearance. He published pictures of an unusual **cube** that appeared to assume different orientations as one looked at it. The effect works because the drawing of the cube (an orthographic projection) carefully eliminates all depth cues. In attempting to fit the expected model of a cube to the picture, our brain must resolve the ambiguity as to which corner of the cube is closer.

negative base

The use of a negative **base** to represent numbers gives rise to some intriguing possibilities. Consider "negadecimal," for example, in which the base is *minus* 10 instead of the familiar positive 10. In this system, the number 365 is equivalent to the decimal number $5 + (6 \times -10) + (3 \times -10 \times -10)$, = 245, while 35 in negadecimal is equivalent to $5 + (3 \times -10)$, = −25, in ordinary decimal. This points to an interesting fact: the negadecimal equivalent of any positive or negative decimal number is always positive and therefore doesn't need to be accompanied by a sign. The Polish UMC-1, of which a few dozen were built in the late 1950s and earlier 1960s, is the only computer ever to use "nega**binary**" (base 2 arithmetic).

negative number

Long denied legitimacy in mathematics, negative numbers are nowhere to be found in the writings of the Babylonians, Greeks, or other ancient cultures. On the contrary, because Greek mathematics was grounded in geometry, and the concept of a negative distance is meaningless, negative numbers seemed to make no sense. They surface for the first time in bookkeeping records in seventh-century India and in a chapter of a work by the Hindu astronomer **Brahmagupta**. Their earliest documented use in Europe is in 1545 in the *Ars magna* of Girolamo **Cardano**. By the early seventeenth century, Renaissance mathematicians were explicitly using negative numbers but also meeting with heavy opposition. René **Descartes** called negative roots "false roots," and Blaise **Pascal** was convinced that numbers "less than zero" couldn't exist. Gottfried **Leibniz** admitted that they could lead to some absurd conclusions, but defended them as useful aids in calculation. By the eighteenth century, negative numbers had become an indispensable part of algebra.

Neile's parabola

Also known as the semi-cubical **parabola**, a curve discovered by the English mathematician William Neile (1637–1670) in 1657; it is the first **algebraic curve** to have its arc length calculated. (Before this, only the arc lengths of transcendental curves such as the **cycloid** and the **logarithmic spiral** had been calculated.) Neile's parabola is described in **Cartesian coordinate**s by the formula

$$y^3 = ax^2.$$

Christiaan **Huygens** showed that this curve satisfies the requirement requested by Gottfried **Leibniz** in 1687, namely, the curve along which a particle may descend under gravity so that it moves equal vertical distances in equal times. Neile's parabola is the **evolute** of a parabola.

nephroid

A type of curve often seen on the surface of a cup of coffee in the sunshine—a crescent of light formed by sunlight reflecting off the inside of the cup onto the surface of the drink. More generally, it is the shape made by parallel rays of light reflecting from the inside of any semicircle. In mathematical terms, this means that the nephroid is the cata**caustic** of a **circle** when the light source is at infinity, a fact first demonstrated by Christiaan **Huygens** in 1678 and published by him in his *Traité de la Lumière* (Treatise on light, 1690). A physical explanation wasn't forthcoming, however, until 1838 when George Airy gave a proof in terms of the wave theory of light. The name *nephroid* (from the Latin for "kidney-shaped") was introduced in 1878 by the English mathematician Richard Proctor in his book *The Geometry of Cycloids.* Prior to that it was known as a two-cusped **epicycloid**. Specifically, the nephroid is the epicycloid formed by a circle of radius a rolling around the outside on a fixed circle of radius $2a$. It has a length of $24a$, an area of $12\pi^2$, and is given by the parametric equations:

$$x = a(3\cos(t) - 3\cos(3t))$$
$$y = a(3\sin(t) - \sin(3t)).$$

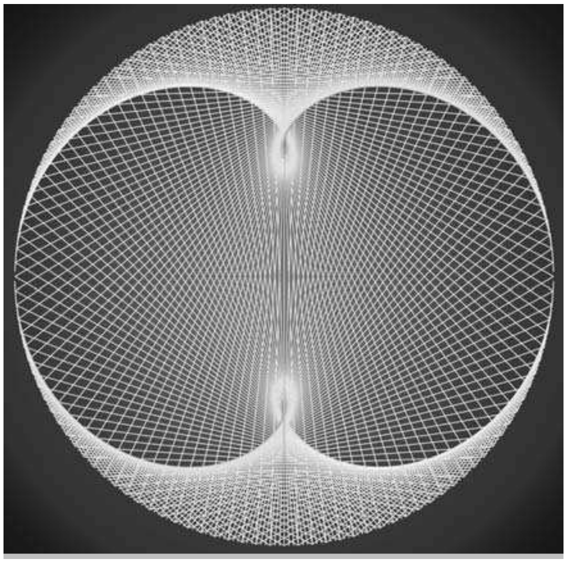

nephroid A nephroid curve spun by thread on a computer loom. *Jos Leys, www.josleys.com*

The nephroid is the **involute** of **Cayley's sextic** and is also the **envelope** of circles with their centers on a given circle, touching a given diameter of that circle. The nephroid has been described as the perfect shape for a multiseat dining table. *Freeth's nephroid,* not to be mistaken for the ordinary nephroid just described, is named after the English mathematician T. J. Freeth (1819–1904) who first wrote about it in a paper published by the London Mathematical Society in 1879. Freeth's nephroid is the **strophoid** of a circle and has the polar equation $r = a(1 + 2\sin(\theta/2))$. Freeth's nephroid is also the name of a group of mathematicians, mostly from Royal Holloway College, London, who gather weekly in a pub called the Beehive and compete in games of trivial pursuit.

net

A drawing of a **polyhedron** unfolded along its **edge**s, so as to lie flat in a plane. The earliest known examples of nets to represent polyhedra are by Albrecht **Dürer**.

neural network

An electronic automaton, similar in some ways to a **cellular automaton**, that offers a highly simplified model of a brain. As such, a neural network is a device for machine learning that is based on associative theories of human cognition. Using various algorithms and weightings of different connections between "neurons," neural networks are set up to learn how to recognize a pattern in applications such as voice recognition, visual pattern recognition, robotic control, symbol manipulation, and decision making. Generally, they consist of three layers: input neurons, output neurons, and a layer in between where information from input to output is processed. Initially the network is loaded with a random program, then the output is measured against a desired output which prompts an adjustment in the weights assigned to the connections in response to the discrepancy between the actual and desired output. This is repeated many times so that the network effectively learns as a child does: in a sense, the net discovers its own rules. Changing the rules of interaction between the "neurons" in the net can lead to interesting emergent behavior, so that neural networks have become another tool for investigating **emergence** and **self-organization**.

Neusis construction

A geometric construction that breaks the strict rules of classical Greek straightedge-and-compass construction (see **constructible**) by allowing a marked ruler to be slid into different positions. Neusis construction makes possible **duplicating the cube** and **trisecting an angle**. John **Conway** and Richard **Guy** have also shown how Neusis constructions, based on angle trisection, can be used to draw regular polygons with 7, 9, and 13 sides.

Newcomb's paradox

One of the most simply stated but astonishing of the so-called prediction paradoxes that bear on the problem of free will. It was devised in 1960 by William Newcomb, a theoretical physicist at the Lawrence Livermore Laboratory, while contemplating the **prisoner's dilemma**. A superior being, with super-predictive powers that have never been known to fail, has put $1,000 in box *A* and either nothing or $1 million in box *B*. The being presents you with a choice: (1) open box *B* only, or (2) open both box *A* and *B*. The being has put money in box *B* only if it predicted you will choose option (1). The being put nothing in box *B* if it predicted you will do anything other than choose option (1) (including choosing option (2), flipping a coin, etc.). The question is, what should you do to maximize your winnings? You might argue that since your choice now can't alter the contents of the boxes you may as well open them both and take whatever's there. This seems reasonable until you bear in mind that the being has never been known to have made an incorrect prediction. In other words, in some peculiar way, your mental state is highly correlated with the contents of the box: your choice is linked to the probability that there's money in box B. These arguments and many others have been put forward in favor of either choice. The fact is there is no known "right" answer, despite the concerted attentions of many philosophers and mathematicians over several decades.[123]

Newton, Isaac (1642–1727)

One of the great intellectual giants in human history; an English mathematician, physicist, and sometime head of the Royal Mint, Newton was one of the chief architects of **calculus** (though not in its modern form), discovered the **binomial theorem**, established the principles of universal gravitation, and saw through numerous other developments, any one of which would have brought a lesser person fame. Interestingly, his most productive period was 1665–1666, his so-called "miraculous year," when Cambridge University was closed because of the plague and Newton had to work at home. He was born in the same year that Galileo died. After publishing his *Principia* (1687), the most important and influential scientific book ever written, his interests drifted toward theology, politics, and, for a while, alchemy. His last two decades were largely spent in acrimonious debate with Gottfried **Leibniz**, over priority in the discovery of calculus, and with the astronomer John Flamsteed.

Newton's method

An iterative method for finding the zeros of a **function**.

n-gon

A **polygon** with *n* sides.

Nim

A **game**, of which there are many different versions, that involves two players alternately removing at least one item from one of two or more piles or rows. The person who picks up the last item wins. In one form of the game, five rows of matches are laid out in such a way that there is one match in the first row, two matches in the second, and so on, down to five matches in the bottom row. Players take turns to remove any nonzero number of matches from any one row. The game may have originated in China. The name "Nim" was coined by Charles Bouton, an associate professor of mathematics at Harvard at the turn of the twentieth century, who took it from an archaic English word meaning to steal or to take away. In 1901 he published a full analysis of Nim and proof of a winning strategy.[50] The first Nim-playing computer, the Nimatron, a 1-ton behemoth, was built in 1940 by the Westinghouse Electrical Corporation and was exhibited at the New York World's Fair. It played 100,000 games against spectators and attendants, and won an impressive 90% of the time; most of its losses came at the hands of attendants who were instructed to reassure incredulous onlookers that the machine could be beaten! In 1951 a Nim-playing robot, the Nimrod, was shown at the Festival of Britain, and later at the Berlin trade fair. It was so popular that spectators entirely ignored a bar at the other end of the room where free drinks were being offered. Eventually the local police had to be called in to control the crowds.[110]

nine

A number long considered to have strange, mystic properties. A phrase in a book written during the Dark Ages gave rise to the superstition that cats have nine lives. English author and satirist William Baldwin wrote in his *Beware the Cat,* "It is permitted for a witch to take her cat's body nine times." There were nine Muses, nine rivers of Hades, and nine heads on the Hydra. It took nine days for Vulcan to fall from the heavens. The phrase "nine days' wonder" comes from the proverb "a wonder lasts nine days and then the puppy's eyes are open." A cat-o'-nine-tails is a whip, usually made of nine knotted lines or cords fastened to a handle that produces scars like the scratches of a cat. Being on "cloud nine" may have its origin in Dante's ninth heaven of Paradise, whose inhabitants are blissful because they are closest to God.

The term "the whole 9 yards" came from World War II fighter pilots in the Pacific. When arming their planes on the ground, the .50-caliber machine gun ammo belts measured exactly 27 feet, before being loaded into the fuselage. If the pilots fired all their ammo at a target, it got "the whole 9 yards." Less certain–though there is no shortage of theories–is the source of the expression "dressed to the nines."

Nine is the largest single-digit number and the one that occurs least frequently in most situations; an exception is the tendency of businesses to set prices that end with one or more nines. Because nine is one less than the base of our number system, it is easy to see if a number is divisible by 9 by adding the digits (and repeating on the result if necessary: the result should be nine). This process is sometimes called **casting out nines**. Similar processes can be developed for divisibility by 99, 999, etc. or any number that divides one of these numbers. Nine has many other interesting properties. For example, write down a number containing as many digits as you like, add these digits together, and deduct the sum from the first number. The sum of the digits of this new number will always be a multiple of nine.

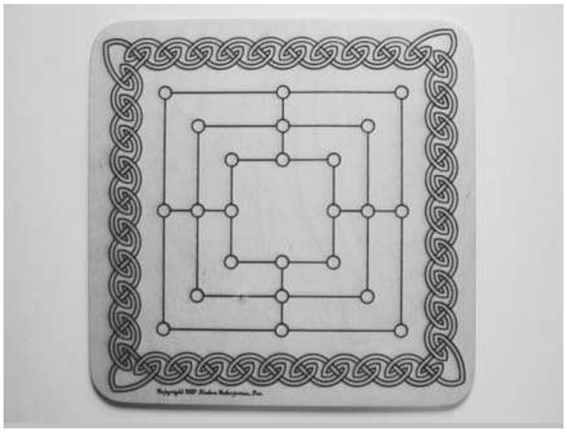
nine men's morris A modern version of the playing board.
Kadon Enterprises, Inc., www.puzzlegames.com

nine holes

See **three men's morris**.

nine men's morris

> *The nine-men's morris is filled up with mud;*
> *And the quaint mazes in the wanton green,*
> *For lack of tread, are indistinguishable.*
> –*Midsummer Night's Dream* (Act 2, scene 1),
> Shakespeare

One of the oldest of board games, known by different names and played with variations of rules in different places and periods. In France it is Marelle, in Austria it is Muhle, and in England it was known as Peg Meryll, Meg Marrylegs, and other names, all referring to a "mill" because that is the name of a run of three counters in the game. Versions of it have been found etched into the roof of the Temple of Kurna in Egypt (dated to about 1400 B.C.), cut into the oak planks that form the deck of the great Viking ship discovered at Gokstad in 1880, and carved in the choir stalls of several English cathedrals. A typical board layout is shown in the accompanying figure; diagonal lines may or may not be included. Joseph Strutt in *The Sports and Pastimes of the People of England* (1801) described the rules in this way:

> Two persons, having each of them nine pieces, or men, lay them down alternately, one by one upon the spots; and the business of either party is to prevent his antagonist from placing three of his pieces so as to form a row of three, without the intervention of an opponent piece. If a row be formed, he that made it is at liberty to take up one of his competitor's pieces from any part he thinks most to his advantage; excepting he has made a row, which must not be touched if he have another piece upon the board that is not a component part of that row. When all the pieces are laid down, they are played backwards and forwards, in any direction that the lines run, but only can move from one spot to another (next to it) at one time. He that takes off all his antagonist's pieces is the conqueror.

In 1996, the German mathematician Ralph Gasser used a computer to prove that nine men's morris is a guaranteed draw if both players make optimal moves from the outset. He programmed the computer to figure out and tabulate 10 billion positions that were known to be a win for one side or the other, then worked forward 18 moves from the beginning of the game until his opening analysis met his endgame analysis. As a result he showed that every potentially winning position could be countered by the opponent in the early stages of the game. See also **three men's morris**.[132]

nine rooms paradox

A puzzle that was first published in *Current Literature* vol. 2, April 1889. It takes the form of a poem and is similar to that of the **missing dollar**.

PUZZLE

Ten weary, footsore travellers,
All in a woeful plight,
Sought shelter at a wayside inn
One dark and stormy night.

'Nine rooms, no more,' the landlord said
'Have I to offer you.
To each of eight a single bed,
But the ninth must serve for two.'

A din arose. The troubled host
Could only scratch his head,
For of those tired men not two
Would occupy one bed.

The puzzled host was soon at ease—
He was a clever man—
And so to please his guests devised
This most ingenious plan.

In a room marked A two men were placed,
The third was lodged in B,
The fourth to C was then assigned,
The fifth retired to D.

In E the sixth he tucked away,
In F the seventh man.
The eighth and ninth in G and H,
And then to A he ran,

Wherein the host, as I have said,
Had laid two travellers by;
Then taking one—the tenth and last—
He logged him safe in I.

Nine single rooms—a room for each—
Were made to serve for ten;
And this it is that puzzles me
And many wiser men.

How has the host managed to bamboozle his patrons?

Solutions begin on page 369.

nine-point circle

Draw a triangle, any triangle (although it may be best to start with an acute triangle). Mark the midpoints of each side. Drop an altitude from each vertex to the opposite side, and mark the points where the altitudes intersect the opposite side. (If the triangle is obtuse, an altitude will be outside the triangle, so extend the opposite side until it intersects.) Notice that the altitudes intersect at a common point. Mark the midpoint between each vertex and this common point. No matter what triangle you start

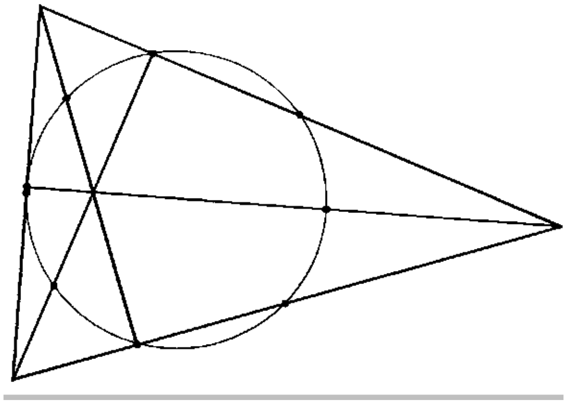

nine-point circle

with, these nine points all lie on a perfect circle! This result was known to Leonhard **Euler** in 1765, but was rediscovered by the German mathematician Karl Feuerbach (1800–1834) in 1822.

node

See **crunode**.

Noether, Emmy (Amalie) (1882–1935)

A German mathematician, one of the most talented of the early twentieth century. The crucial result now known as *Noether's theorem,* which is important in other symmetries in natural systems, is of great importance in physics. She received her doctorate in 1907 and rapidly built an international reputation, but the University of Göttingen refused to let her teach, and her colleague David **Hilbert** had to advertise her courses in the university's catalog under his own name. A long controversy ensued, with her opponents asking what the country's soldiers would think when they returned home and were expected to learn at the feet of a woman. Allowing her on the faculty would also mean letting her vote in the academic senate. Said Hilbert, "I do not see that the sex of the candidate is against her admission as a privatdozent. After all, the university senate is not a bathhouse." She was finally admitted to the faculty in 1919. A Jew, Noether was forced to flee Nazi Germany in 1933 and joined the faculty at Bryn Mawr in the United States.

non-Abelian

Non**commutative** or order-dependent. For example, the group of manipulations of **Rubik's cube** is non-Abelian because the state of the cube depends greatly on the order of the moves performed on it. See also **Abelian group**.

nonagon

A **polygon** with nine sides. A *nonagonal number* is a number of the form $n(7n - 5)/2$.

nonconvex uniform polyhedron

A **uniform polyhedron** of a type obtained by relaxing the conditions used to produce the **Archimedean solid**s (which have regular convex faces and identical convex vertices) to allow both nonconvex faces and vertex types, as in the case of the **Kepler-Poinsot solid**s. The condition that every vertex must be identical, but the faces need not be, gives rise to 53 nonconvex uniform polyhedra. An example is the *great truncated dodecahedron,* obtained by truncating the corners of the great dodecahedron at a depth which gives regular decagons.

non-Euclidean geometry

Any **geometry** in which Euclid's **parallel postulate** doesn't hold. (One way to state the parallel postulate is: given a straight line and a point A not on that line, there is only one exactly straight line through A that never intersects the original line.) The two most important types of non-Euclidean geometry are **hyperbolic geometry** and **elliptical geometry**. The different models of non-Euclidean geometry can have positive or negative **curvature**. The sign of curvature of a surface is indicated by drawing a straight line on the surface and then drawing another straight line perpendicular to it: both these lines are **geodesic**s. If the two lines curve in the same direction, the surface has a positive curvature; if they curve in opposite directions, the surface has negative curvature. Elliptical (and spherical) geometry has positive curvature whereas hyperbolic geometry has negative curvature.

The discovery of non-Euclidean geometry had immense consequences. For more than 2,000 years, people had thought that **Euclidean geometry** was the only geometric system possible. Non-Euclidean geometry showed that there are other conceivable descriptions of space–a realization that transforms mathematics into an altogether more abstract science. Thereafter, it was clear that in mathematics, one could start out with *any* set of self-consistent

Types of Geometry

	Euclidean	**Elliptical**	**Hyperbolic**
Curvature	Zero	Positive	Negative
Given a line m and a point P not on m, the number of lines passing through P and parallel to m	1	0	Many
Sum of interior angles of a triangle	180°	> 180°	< 180°
Square of hypotenuse of a right triangle with sides a and b	$a^2 + b^2$	$< a^2 + b^2$	$> a^2 + b^2$
Circumference of a circle with diameter 1	π	$< \pi$	$> \pi$

postulates and follow through their ramifications. The discovery of non-Euclidean geometry has been compared with Copernicus's theory and Einstein's **relativity theory**, the analogy being that each freed people from long-held models of thought. In fact, Einstein said about non-Euclidean geometry: "To this interpretation of geometry, I attach great importance, for should I have not been acquainted with it, I never would have been able to develop the theory of relativity."

It's important to realize that both Euclidean and non-Euclidean geometry are consistent in that the assumptions on which they rest don't involve any contradictions. In response to a question as to which geometry is true, Henri **Poincaré** said: "One geometry cannot be more true than the other; it can only be more convenient." Which geometry is valid in the physical space in which we live? On a small scale, and for all practical purposes on Earth, Euclidean geometry works just fine. But on larger scales this is no longer true. Einstein's general theory of relativity uses non-Euclidean geometry as a description of **space-time**. According to this idea, space-time has a positive curvature near gravitating matter and the geometry is non-Euclidean. When a body revolves around another body, it appears to move in a curved path due to some force exerted by the central body, but it is actually moving along a **geodesic**, without any force acting on it. Whether all of space-time contains enough matter to give itself an overall positive curvature is one of the many unanswered question in physics today, but it is generally accepted that the geometry of space-time is more non-Euclidean than Euclidean. It is proposed that if space-time does happen to have an overall positive curvature, then the universe will stop expanding after a fixed amount of time and start to shrink resulting in a "big crunch," as opposed to the "big bang" that resulted in its creation. At the moment, astronomical observations seem to favor a universe that is "open" and has a hyperbolic geometry. Another consequence of non-Euclidean geometry is the possibility of the existence of a **fourth dimension**. Just as the surface of the sphere curves in the direction of the third dimension, i.e., perpendicular to its surface, it is believed that space-time curves in the direction of the fourth dimension. Non-Euclidean geometry has applications in other areas of mathematics, including the theory of elliptic curves, which was important in the proof of **Fermat's last theorem**.[72, 140] (See table, "Types of Geometry.")

nonlinear system

Any system in which the data points coming from the measurement of the values of its variables can be represented as a curvilinear pattern on a coordinate plane; hence, "nonlinear" for "not-a-line." More generally, a system in which small changes can result in large effects, and large changes in small effects. Thus, sensitive dependence on initial conditions (see **butterfly effect**) in chaotic systems illustrates the extreme nonlinearity of these systems. In a nonlinear system the components are interactive, interdependent, and exhibit feedback effects.

nonstandard analysis

In a broad sense, the study of the infinitely small; more specifically, the study of **hyperreal number**s, their functions and properties. Nonstandard analysis, which was pioneered by Abraham **Robinson** in the 1960s, puts the concept of **infinitesimal**s on a firm mathematical footing and is, for many mathematicians, more intuitive than real **analysis**.

normal

A line that is perpendicular to a given line or plane.

normal number

A number in which digit sequences of the same length occur with the same frequency. A constant is considered normal to base 10 if any single digit in its decimal expansion appears one-tenth of the time, any two-digit combination one-hundredth of the time, any three-digit combination one-thousandth of the time, and so on. In the case of **pi**, the digit 7 is expected to appear 1 million times among the first 10 million digits of its decimal expansion. It actually occurs 1,000,207 times–very close to the expected value. Each of the other digits also turns up with approximately the same frequency, showing no significant departure from predictions. A number is said to be *absolutely normal* if its digits are normal not only to base 10 but also to every integer base greater than or equal to 2. In base 2, for example, the digits 1 and 0 would appear equally often. Émile **Borel** introduced the concept of normal numbers in 1909 as a way to characterize the resemblance between the digits of a mathematical constant such as π and a sequence of **random numbers**. He quickly established that there are lots of normal numbers, though finding a specific example of one proved to be a major challenge. The first to be found was **Champernowne's number**, which is normal to base 10. Analogous normal numbers can be created for other bases. To date, no specific "naturally occurring" **real number** has been proved to be absolutely normal, even though it is known that almost all real numbers are absolutely normal! However, in 2001, Greg Martin of the University of Toronto found some examples of the opposite extreme–real numbers that are normal to no base whatsoever. To start with, he noted that every

rational number is absolutely abnormal. For example, the fraction $1/7$ can be written in decimal form as 0.1428571428571 . . . . The digits 142857 just repeat themselves. Indeed, an expansion of a rational number to any base b or b^k eventually repeats. Martin then focused on constructing a specific *irrational* absolutely abnormal number. He nominated the following candidate, expressed in decimal form, for the honor:

$$\alpha = 0.6562499999956991\underline{999999\ldots9999998}528404201690728\ldots$$

The middle portion (underlined) of the given fragment of α consists of 23,747,291,559 nines. Martin's formulation of this number and proof of its absolute abnormality involved so-called **Liouville number**s.[216]

nothing

The absence of anything; nonexistence. Nothing is not the same as the **empty set**, which exists as the set that mathematically denotes nothing, nor is it the same as **zero**, which exists as the number that denotes how many members the empty set contains. In physics, nothing is not a **vacuum**, because a vacuum not only contains energy but exists in space and time; nor is it a singularity, which contains a great deal of concentrated matter and energy. So, can there be nothing? No. "To be" implies existence of some sort: the one thing we can be absolutely sure has never existed, or will exist, is nothing.

noughts and crosses

See **tic-tac-toe**.

NP-hard problem

A mathematical problem for which, even in theory, no shortcut or smart **algorithm** is possible that would lead to a simple or rapid solution. Instead, the only way to find an optimal solution is a computationally intensive, exhaustive analysis in which all possible outcomes are tested. Examples of NP-hard problems include the **traveling salesman problem** and the popular game **Tetris**. NP stands for "non-deterministic polynomial-time."

nucleation

A process in a physical system, or a mathematical model such as a **cellular automaton** or a statistical model, whereby a bubble or other structure appears spontaneously at a random or unpredictable spot.

null hypothesis

The **hypothesis** that is being tested in a hypothesis-testing situation.

null set

See **empty set**.

number

An abstract measure of quantity. The most familiar numbers are the **natural number**s, 0, 1, 2, . . . , used for counting. If negative numbers are included, the result is the **integer**s. Ratios of integers are called **rational number**s, which can be expressed as terminating or repeating decimals. If all infinite and nonrepeating decimal expansions are thrown in as well, the scope of numbers extends to all **real numbers**, which can be extended to the **complex numbers** in order to include all possible solutions to algebraic equations. More recent developments are the **hyperreal number**s and the **surreal number**s, which extend the real numbers by adding infinitesimal and infinitely large numbers. For measuring the size of infinite **set**s, the natural numbers have been generalized to the **ordinal number**s and to the **cardinal number**s. See also **numeral** and **number system**.

number line

A way of representing numbers by thinking of them as the positions of points on a line.

number system

A way of counting using a particular **base**. The familiar **decimal** number system is a base 10 system, because it uses 10 different digits, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and combinations of these. Other number systems, used for special purposes or by certain cultures, are the **binary** (base 2), trinary (base 3), hexadecimal (base 16), **vigesimal** (base 20), and **sexagesimal** (base 60). In a *positional number system* of base b, b basic symbols (or digits) corresponding to the first b **natural number**s, including zero, are used. To generate the rest of the numbers, the position of the symbol in the figure is used. The symbol in the last position has its own value, and as it moves to the left its value is multiplied by b. In this way, with only finitely many different symbols, every number can be expressed. This is unlike systems, such as that of **Roman numeral**s, which use different symbols for different orders of magnitude.

number theory

The study of the whole numbers and their properties and relationships. Often, a problem in number theory can be restated in terms of finding the solution, or showing that there is no solution, to a **Diophantine equation**. A Diophantine equation is one where the coefficients are whole numbers and where the solution is also constrained to be a whole number. Sometimes, what appears to be a simple Diophantine equation can lead to an **elliptic curve**.[30]

numeral

A symbol, or combination of symbols, that describes a number. See **Arabic numeral** and **Roman numeral**.

numerator

The number above the fraction bar that indicates the number of parts of the whole there are in a **rational number**.

numerical analysis

The study of methods for approximating the solutions of various classes of mathematical problems including *error analysis*.

O

obelus

The name of the symbol "÷", used as a sign for division. It comes from the Greek *obelos* meaning a pointed stick (a spit) used for cooking. This root word also gave rise to *obelisk* for a pointed stone pillar. The "÷" symbol was originally used as an editing mark in early manuscripts, sometimes only as a line without the two dots, to point out material that the editor thought needed cutting. It was also used occasionally as a symbol for subtraction. As a division symbol it was first employed by the Swiss mathematician Johann Rahn (1622–1676) in his *Teutsche Algebra* in 1659. By a misunderstanding of a credit to John Pell about other material in the book, many English writers started using the symbol and calling it "Pell's notation." Although it appears regularly in literature produced in Britain and the United States, it is virtually unknown in the rest of the world.

oblate spheroid

An **ellipsoid** produced by rotating an **ellipse** through 360° about its **minor axis**.

oblique

Slanted or not perpendicular. *Oblique coordinates* are measured on a plane coordinate system whose axes are not perpendicular. An *oblique angle* is any angle except a **right angle** and an *oblique triangle* is any triangle that doesn't contain a right angle. The first reference to "oblique angled triangles" may have been by Thomas Blundevil in his *Exercises* in 1594.

oblong

An alternate name for any **rectangle** that is not a square. The word comes from the Latin *ob* ("excessive") and *longus* ("long").

oblong number

Any positive integer that is not a **perfect square**.

obtuse

An *obtuse angle* is an **angle** greater than 90° but less than 180°. An *obtuse triangle* is a **triangle** that contains an obtuse angle. The word comes from the Latin *ob* ("against") and *tundere* ("to beat") and thus refers to things that are blunt, dull, or rounded.

octa-

The Greek prefix meaning "eight."

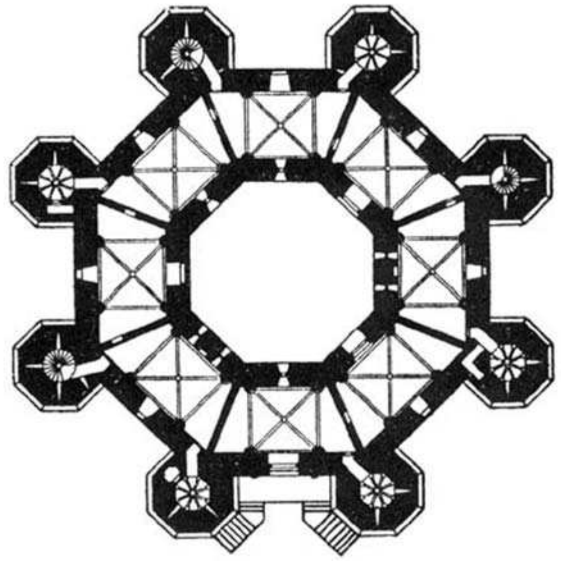

octa- An aerial view of Castel del Monte in Italy's Apulien region.

octagon

An eight-sided **polygon**. One of the most remarkable examples of octagonal design in architecture is the Castel del Monte, in southern Italy, which consists of a central octagonal core containing an inner octagonal courtyard that is surrounded by eight tall, perfectly octagonal towers. Another famous octagon is Oxford University's Radcliffe Observatory tower, built from 1772 to 1794, vaguely based on the Tower of the Winds in Athens, and considered architecturally the finest observatory in Europe.[139]

octahedron

A **polyhedron** with eight faces, each of which is an equilateral triangle. A regular octahedron, whose sides are all equilateral triangles, is one of the **Platonic solid**s; it looks like two square pyramids with their square bases stuck together. See also **octa-**.

octant

Any one of the eight portions of space determined by the three coordinate planes.

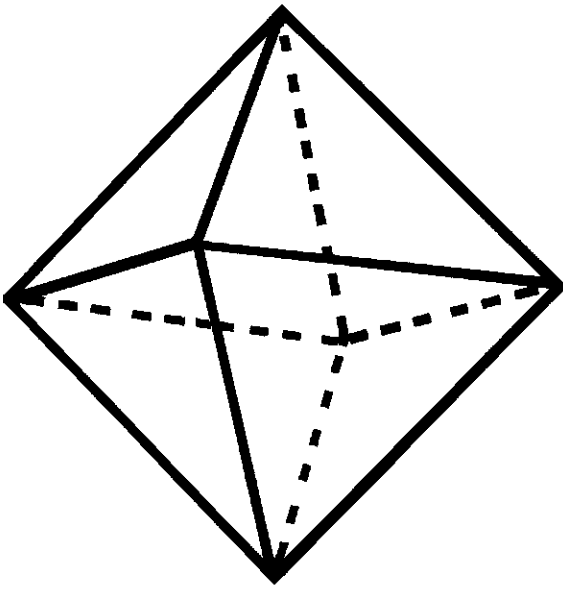

octahedron One of the Platonic solids.

octonion

Also known as a *Cayley number.* Octonions are a non**associative** generalization of the **quaternion**s and the **complex number**s involving numbers with one real **coefficient** and seven imaginary coefficients.

odd

Not divisible by two. An *odd function* is a **function** $f(x)$ with the property that $f(-x) = -f(x)$ for any value of x; an example is $\sin(x)$.

officer problem

See **thirty-six officers problem**.

omega

See **Chaitin's constant**.

one

The first positive **integer** and the first odd number; it is also known as *unity.* From the time of Euclid to the late 1500s, one wasn't generally considered to be a number but instead was thought of as the unit of which bonafide numbers were composed. The Old English (c. 550–c. 1100) *ane* served both for counting and as the indefinite article. Toward the end of the Old English period and the beginning of Middle English (c. 1100–c. 1500), *ane* developed two pronunciations, the first being used for 1 and the other for the indefinite article *an, a.* The existence of different words for the number one and the indefinite article seems to be unique to the English language. "One" can be traced back to the Latin *unus* and the Greek *oine* but probably came into English from the German *eine.*

153

A number with some very curious properties. It is the smallest number that can be expressed as the sum of the cubes of its digits: $153 = 1^3 + 5^3 + 3^3$. It is equal to the sum of the factorials of 1 to 5: $153 = 1! + 2! + 3! + 4! + 5!$ The sum of the digits of 153 is a perfect square: $1 + 5 + 3 = 9 = 3^2$. The sum of the **aliquot part**s of 153 is also a perfect square: $1 + 3 + 9 + 17 + 51 = 81 = 9^2$. On adding the number 153 to its reverse (351), the result is 504 whose square is the smallest square that can be expressed as the product of two different numbers that are the reverse of one another: $153 + 351 = 504$; $504^2 = 288 \times 882$. It can be expressed as the sum of all integers from 1 to 17. In other words, 153 is the seventeenth triangular number; its reverse is also a triangular number. In addition, 153 and its reverse are **Harshad number**s. It can be expressed as the product of two numbers formed from its own digits: $153 = 3 \times 51$. A reference to 153 occurs in the New Testament: the net that Simon Peter drew from the Sea of Tiberias held 153 fishes.

one-to-one

A **function** or map that for every possible output has only one input that yields that particular output; if $f(a) = f(b)$, then $a = b$.

open

An *open interval* is a piece of a straight line that doesn't contain its endpoints; a *half-open interval* contains one endpoint. An *open set* is one in which every point in the set has a neighborhood lying in the set.

operator

Something that acts on a **function** to give another function.

optical illusion

A picture or figure that deceives or confuses the eye and/or brain. Categorization of optical illusions is difficult because several underlying mechanisms may contribute to an effect, or the cause of the illusions may not be completely understood. For the purposes of this book illusions are grouped as shown in the table "Some Types and Examples of Optical Illusions." See individual entries for details. See also **anamorphosis**.

Some Types and Examples of Optical Illusions

Type	Examples
Distortion illusion	Fraser spiral Müller-Lyer illusion Orbison's illusion Poggendorff illusion Titchener illusion Zöllner illusion
Impossible figures and objects	Freemish crate Penrose triangle Penrose stairway Tribar illusion
Ambiguous figure	Ames room Necker cube Schröder's reversible staircase Thiery figure
Lateral inhibition illusion	Hermann grid illusion
Antigravity houses and hills	

Orbison's illusion

A **distortion illusion** in which a background of radiating lines appears to distort the shape of a superimposed figure, such as a square or circle.

orbit

(1) The path, in the form of a **conic section**, that an object takes when under the gravitational influence of another body. (2) More generally, the trajectory of a **differential equation**.

order

A word with many different meanings in mathematics. Among these are: (1) the sequence in which a set of objects or numbers is placed; (2) the number of elements in a set; (3) the number of times a shape can be fitted back onto its own outline during a complete turn (*order of symmetry*); (4) the highest order **power** in a one-variable **polynomial** (*order of a polynomial*, also known as its *degree*); (5) the type of curve described by such a polynomial (*order of a curve*).

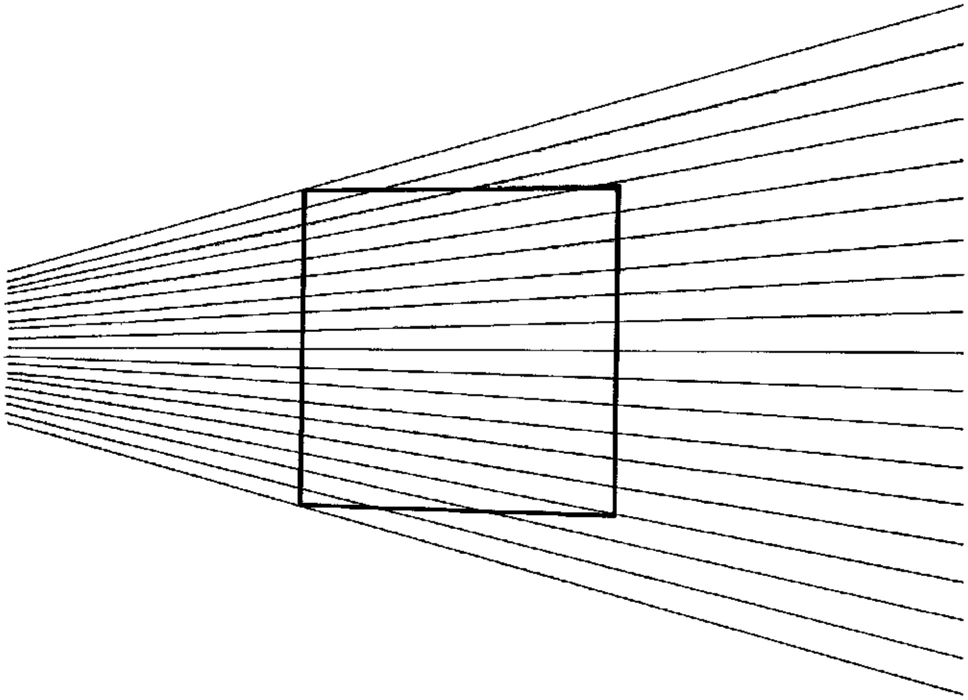

Orbison's illusion Is the square really a square?

ordered pair

A collection of two objects such that one can be distinguished as the *first element* and the other as the *second element*. An ordered pair with first element a and second element b is usually written as (a, b). Two such ordered pairs (a_1, b_1) and (a_2, b_2) are equal if and only if $a_1 = a_2$ and $b_1 = b_2$. *Ordered triples* and *ordered n-tuples* (ordered lists of n terms) are defined in the same way. An ordered triple (a, b, c) can be defined as $(a, (b, c))$, that is, as two nested pairs.

ordinal number

A number used to give the position in an ordered sequence: first, second, third, fourth, ...Ordinal numbers are distinct from **cardinal number**s (one, two, three, four, . . .), which describe the *size* of a collection. The mathematician Georg **Cantor** showed in 1897 how to extend the concept of ordinals beyond the **natural number**s to the infinite and how to do arithmetic with the resulting *transfinite ordinals* (see **infinity**).

ordinary differential equation

Any equation relating a **function** of one variable to its **derivative**s. Compare with **partial differential equation**.

ordinate

The y-coordinate, or vertical distance from the x-axis, in a system of **Cartesian coordinates**. Compare with **abscissa**.

origami

The Japanese art of the paper-folding, without cutting and joining, to form definite objects.

origin

The point (0, 0) on the coordinate axes.

orthic triangle

The triangle whose vertices (see **vertex**) are the feet of the **altitude**s of a given triangle.

orthocenter

The point of intersection of the **altitude**s of a **triangle.**

orthogonal

At right angles to; independent of. *Orthogonal curves* are two families of curves with the property that each member of one family meets members of the other family at right angles.

osculating

To share the same **tangent** and **curvature** and at a given point, as in the case of two *osculating curves*. For example the curve $y = x^3$ and the x-axis osculate at the origin.

Oughtred, William (1574–1660)

An English clergyman and mathematician who invented the **slide rule** and who, in the first English edition of his *Clavis Mathematicae* in 1647, first used the name "**pi**" for the number 3.141 . . . . He wrote it π.δ, where π stood for the English word *periphery* (what we would call circumference), the dot was his symbol for division, and δ stood for the English word *diameter.* The use of π to represent words starting with the letter *p,* like *periphery,* was not uncommon. Before π = 3.14 . . . caught on, π was variously used to indicate a point, a polygon, a positive number, a power, a proportion, the number of primes in a series, and a factorial (which is a product). Oughtred used the same notation in all later English and Latin editions of his book, but not in the earlier first Latin edition. He was a prodigious inventor of mathematical symbols, though most of them have not survived. He did, however, introduce a couple of other symbols we still use: × for multiplication (as distinct from the letter *x,* which he also used for this purpose) and ± for "plus or minus." Oughtred served as rector of Albury from 1608 or 1610 until his death. There he tutored many young mathematicians of the time, including John **Wallis**, Seth Ward, Charles Scarburgh, and Christopher Wren. Moreover, all English mathematicians for the next century, including Isaac **Newton**, learned algebra from *Clavis Mathematicae* (first published in 1631).

oval

A curve that looks like a squashed **circle** but, in contrast with the **ellipse**, doesn't have a precise mathematical definition. The word *oval* comes from the Latin *ovus* for "egg." Unlike ellipses, ovals sometimes have only a single axis of reflection symmetry (instead of two).

Ovid's game

A board game, described by the Roman poet Ovid in Book III of his *Art of Love,* that was popular in ancient Greece, Rome, and China. Each player has three counters, which they place alternately on to nine points laid out in a 3 × 3 grid, with the object of getting three in a line and so winning. This may sound like **tic-tac-toe**, and there is no doubt that both Ovid's game and **nine men's morris** are antecedents of the familiar noughts and crosses. But in Ovid's game play continues after the six counters are down, if no one has yet won, by moving a single counter on each turn, though not diagonally, to any adjacent square. As in the case of tic-tac-toe, two experts (i.e., "rational" players) will always draw. Because the first player is ensured a win by covering the center square, this move is usually not allowed. Ovid advised women to master the game in order to gain the attention of men!

Ozanam, Jacques (1640–1717)

A French mathematician, scientist, and writer best remembered for his book on mathematical and scientific puzzles *Récréations Mathématiques et Physiques* (four volumes, 1694), which later went through 10 editions. Based on earlier works by Claude **Bachet**, Claude **Mydorge**, Jean **Leurechon**, and Daniel Schwenter (1585–1636), it was later revised and enlarged by Jean **Montucla**, then translated into English by Charles Hutton (1803, 1814). Edward Riddle edited a new edition, which was published in 1844, removing some old material and adding new material so that "[T]he work might continue to be to the present generation a useful manual of scientific recreation, as its predecessors have been to the generation which has passed." Ozanam's original edition contained an early example of a problem about orthogonal **Latin squares**: "Arrange the 16 court cards so that each row and each column contains one of each suit and one of each value."

P

Pacioli, Luca (1445–1517)

An Italian mathematician and Franciscan monk who wrote several influential books. His encyclopedic *Summa de Arithmetica, Geometria, Proportioni et Proportionalita* (Summary of arithmetic, geometry, proportion and proportionality, 1494) summarized what was known about contemporary arithmetic, algebra, geometry, and trigonometry and gave a basis for the major progress in mathematics which subsequently took place in Europe. *Divina proportione* (1509), with drawings by none other than Leonardo da Vinci (surely no mathematical text was more impressively illustrated!), deals with the **golden ratio**, a subject that Pacioli treats from an architectural standpoint in a second volume. At his death, he left a major book unpublished, *De Viribus Amanuensis,* on recreational problems, geometrical problems, and proverbs. It makes frequent reference to Leonardo, who assisted him with the project: many of the problems in this treatise are also in Leonardo's notebooks. Again it is a work for which Pacioli claimed no originality, describing it as a compendium.

packing

A way to place objects of the same kind so that they touch in some specified way, often inside a container with specified properties. The objects to be packed may be polyhedra, polygons, spheres, ellipsoids, hyperspheres, or any other type of shape, and the number of dimensions involved may range upward from two. The fraction of a space filled by a given collection of objects is called the *packing density.* The densest packing of circles in the plane is the hexagonal lattice of the bee's honeycomb, which has a packing density of 0.9069 . . . . In 1611, Johannes **Kepler** proposed that hexagonal, or face-centered cubic, packing is also the densest possible way to arrange spheres in three dimensions–an assertion known as **Kepler's conjecture**. Currently, the worst known **convex** packer in two-dimensions is the smoothed octagon, with a packing density of about 0.902. Stanislaw **Ulam** conjectured that the sphere was the worst packing object in three-dimensional space.

Padovan sequence

See **plastic number**.

palindrome

A word, verse, sentence, or passage that reads the same forward or backward; the term comes from the Greek *palindromos* for "running back again." Well-known examples include: "Madam, I'm Adam;" "A man, a plan, a canal–Panama!;" and "Able was I ere I saw Elba." A slightly longer one, devised by Peter Hilton, a code-breaker on the British team that cracked the German Enigma, is "Doc, note. I dissent. A fast never prevents a fatness. I diet on cod." Credit for inventing the palindrome is often given to Sotades the Obscene of Maronea (third century B.C.). Though only eleven lines of his work have survived, he is thought to have recast the entire Iliad in palindromic verse. Sotades also wrote lines (now sometimes called Sotadic verses) that, when read backward, had the opposite meaning. His acid tongue eventually landed him in jail by order of Ptolemy II, though worse was to follow. Sotades escaped but was captured by Ptolemy's admiral Patroclus, who sealed him in a leaden chest and tossed him into the sea. A musical palindrome is formed by Haydn's Symphony No. 47 in G, sometimes referred to as *The Palindrome,* because in both the minuet and the trio the orchestra plays the music twice forward and then twice backward to arrive at the beginning. See also **palindromic number**.

palindromic number

A number such as 1,234,321 that reads the same forward and backward; more generally, a symmetrical number written in some **base** a as $a_1\, a_2\, a_3 \ldots | \ldots a_3\, a_2\, a_1$. In the familiar base 10 system, there are nine two-digit palindromic numbers: 11, 22, 33, 44, 55, 66, 77, 88, 99; there are 90 palindromics with three digits: 101, 111, 121, 131, 141, . . . , 959, 969, 979, 989, 999; and there are 90 palindromics with four digits: 1,001; 1,111; 1,221; 1,331; 1,441; . . . , 9,559; 9,669; 9,779; 9,889; 9,999, giving a total of 199 palindromic numbers below 10^4. Below 10^5 there are 1,099 palindromics and for other exponents of 10^n there are 1,999; 10,999; 19,999; 109,999; 199,999; 1,099,999; . . . . It is conjectured, but has not been proven, that there are an infinite number of palindromic **prime numbers**. With the exception of 11, palindromic primes must have an odd number of digits. A normally quick way to produce a palindromic number is to pick a positive integer of two or more digits, reverse the digits, and add to the original, then repeat this process with the new number, and so on. For example, 3,462 gives the sequence 3,462; 6,105; 11,121; 23,232. Does the series formed by adding a number to its reverse always end in a palindrome? It used to be thought so. However, this conjecture has been proven false for bases 2, 4, 8, and other powers of 2, and seems to be false for base 10 as well. Among the first 100,000 numbers, 5,996 numbers are known that have not produced palindromic numbers by the add-and-reverse method in calculations carried out to date. The first few of these are 196; 887; 1,675; 7,436; 13,783; 52,514; . . . . A proof that these numbers never produce palindromes, however, has yet to be found. The largest known palindromic prime, containing 30,913 digits, was found by David Broadhurst in 2003.

pandiagonal magic square

A **magic square** in which all the broken diagonals as well as the main diagonals add up to the magic constant.

pandigital number

An integer that contains each of the digits from zero to nine exactly once and whose leading digit is nonzero. The smallest pandigital numbers are 1,023,456,789; 1,023,456,798; 1,023,456,879; 1,023,456,897; and 1,023,456,978. Like all pandigitals these are divisible by nine. The first few "zeroless" pandigitals are 123,456,789; 123,456,798; 123,456,879; 123,456,897; 123,456,978; and 123,456,987; and the first few zeroless pandigital primes are 1,123,465,789; 1,123,465,879; 1,123,468,597; 1,123,469,587; and 1,123,478,659. The sum of the first 32,423 (a **palindromic number**) consecutive primes is 5,897,230,146, which is pandigital. No other palindromic number shares this property. Examples of palindromic numbers that are the product of pandigital numbers are 2,970,408,257,528,040,792 (= 1,023,687,954 × 2,901,673,548) and 5,550,518,471,748,150,555 (= 1,023,746,895 × 5,421,768,309), both found in 2001. See also **pandigital product**.

pandigital product

A product in which the digits of the multiplicand, multiplier, and product, taken together, form a **pandigital number**.

pangram

A phrase or sentence that contains every letter of the alphabet at least once. The best known example in English and a familiar typing test is "The quick brown fox jumps over the lazy dog." One of the shortest known that still makes some kind of sense is "The five boxing wizards jump quickly."

Pappus of Alexandria (c. A.D. 300)

The last of the great Greek geometers whose eight-volume *Mathematical Collection* summarized the bulk of mathematics known at that time. In this compendium, Pappus added a considerable number of his own explanations and amplifications of the earlier work of **Euclid**, **Archimedes**, **Apollonius**, and others.

parabola

One of the **conic section**s and one of the most studied curves in the history of mathematics. A parabola is the outline of the figure obtained if a right circular **cone** is cut by a plane that is exactly parallel to the cone's side. Just as the circle is a limiting case of the **ellipse** when the two foci coincide, the parabola is a limiting case of the ellipse when one of the foci is moved to infinity. As the French mathematician Henri Fabre eloquently put it, the parabola is an ellipse that "seeks in vain for its second, lost center." Johannes **Kepler** drew this further connection between the parabola and other conic sections: "Because of its intermediate nature the parabola occupies a middle position [between the ellipse and the **hyperbola**]. As it is produced it does not spread out its arms like the hyperbola but contracts them and brings them nearer to parallel, always encompassing more, yet always striving for less–whereas the hyperbola, the more it encompasses, the more it tries to obtain."

A parabola is the **locus** of all points in a plane that are equidistant from a given line, known as the *directric,* and

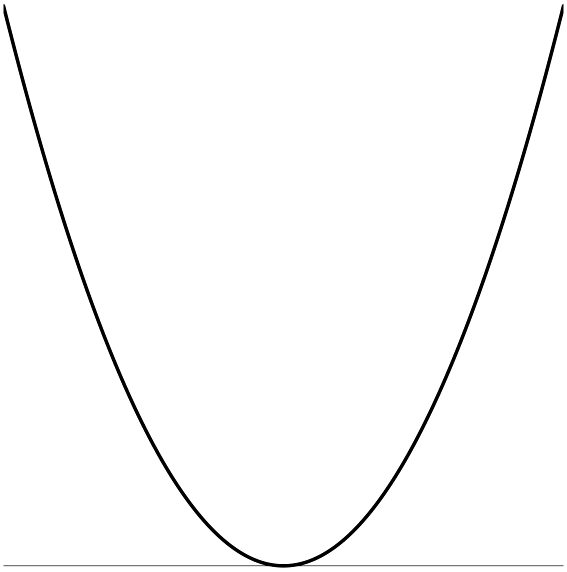

parabola *Jan Wassenaar, www.2dcurves.com* a given point not on the line, known as the focus. The Cartesian equation of a parabola that opens upward and has its vertex (turning point) at the origin is $y = 4ax^2$, where a is the distance from the vertex to the focus, and the quantity $4a$ is known as the *latus rectum.* More generally, any **quadratic** equation of the form $y = ax^2 + bx + c$, where a is not zero, graphs a parabola. The simplest form of this, when $a = 1$, and both b and c are zero, is $y = x^2$. Like the circle, but unlike the ellipse and the hyperbola, the parabola has only one distinct shape. In other words, any parabola can be superimposed exactly on any other parabola simply by rotating, translating (sliding), and/or enlarging or shrinking it.

Euclid dealt with the parabola in his *Conic Sections,* and, although this treatise was lost, it provided a foundation for the first four books by **Apollonius** of the same name. Galileo (1564–1642) discovered that a cannonball, or any other projectile launched at an angle to ground, follows a parabolic path: a result that immediately grabbed the attention of not only scientists but of monarchs and military leaders. René **Descartes**, in writing *La Géométrie* (1637), chose the parabola to illustrate his innovative **analytical geometry**. In 1992, Rudolph Marcus of the California Institute of Technology won the Nobel Prize in Chemistry for his work showing that parabolic reaction surfaces can be used to calculate how fast electrons travel in molecules. His most famous theoretical result, an inverted rate-energy parabola, predicts electron transfer will slow down at very high-reaction free energies.

paraboloid

The surface of revolution of the **parabola**. It is a **quadratic** surface described by the equation $z = a(x^2 + y^2)$.

paradox

> *Please accept my resignation. I don't want to belong to any club that will accept me as a member.*
>
> –Groucho Marx (1895–1977)

A statement that seems to lead to a logical self-contradiction, or to a situation that contradicts common intuition. The word *paradox* comes from the Greek *para* ("beyond") and *doxa* ("opinion" or "belief"). The identification of a paradox based on seemingly simple and reasonable concepts has often led to significant advances in science, philosophy, and mathematics. See **Allais paradox**, **Arrow paradox**, **Banach-Tarski paradox**, **Berry's Paradox**, **birthday paradox**, **Burali-Forti paradox**, **coin paradox**, **grandfather paradox**, **Grelling's paradox**, **liar paradox**, **Newcomb's paradox**, **nine rooms paradox**, **Parrondo's paradox**, **raven paradox**, **Russell's paradox**, **St. Petersburg paradox**, **Siegel's paradox**, **unexpected hanging**, and **Zeno's paradoxes**.

parallel

Said of two or more things, such as lines or planes, that are equally distant from one another at all points.

parallel postulate

The fifth and most controversial of **Euclid's postulates** set forth in the Greek geometer's great work, *Elements.* To later mathematicians, the parallel postulate seemed less obvious than the other four and many attempts were made to derive it from them, but without success. In 1823, János **Bólyai** and Nikolai **Lobachevsky** independently realized that entirely self-consistent types of **non-Euclidean geometry** could be created in which the parallel postulate doesn't hold. Carl **Gauss** had made the same discovery earlier but kept the fact secret.

parallelepiped

A **polyhedron** with six faces bounded by three pairs of parallel planes, so that all its faces are **parallelograms**. It is also a **prism** with a parallelogram for a base. A *rectangular parallelepiped* has the shape of a shoebox. The word comes from the Greek *parallelepipedon* for the same shape, which in turn comes from the roots *para* ("beside"), *allel* ("other"), *epi* ("on"), and *pedon* ("ground"). *Parallelepipedon* may have been first used in English by Billingsley in his 1570 translation of Euclid. It seems to have then given way to *parallelepiped* in the last quarter of the nineteenth century.

parallelogram

A **quadrilateral** (four-sided figure) whose opposite sides are parallel, and whose opposite angles, therefore, are equal. The diagonals of a parallelogram bisect each other. A parallelogram of base b and height h has an area

$$bh = ab \sin A = ab\sin B.$$

The height of a parallelogram is $h = a \sin A = a \sin B$. The sides a, b, c, and d and diagonals p and q of a parallelogram satisfy the equality

$$p^2 + q^2 = a^2 + b^2 + c^2 + d^2.$$

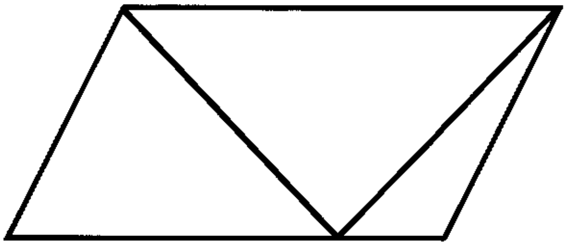

parallelogram The two lines drawn inside this parallelogram have the same length.

Special cases of a parallelogram are a **rhombus**, which has sides of equal length; a **rectangle**, which has two sets of parallel sides that are perpendicular to each other; and a **square**, which meets the conditions of both a rectangle and a rhombus.

In the *parallelogram illusion* shown, despite appearances to the contrary, the interior diagonal lines are of equal length.

parameter

An independent variable; one of the inputs for a **function**. *Parametric equations* are equations in which the variables of interest are given in terms of another variable.

parity

(1) The sum of the digits in a **binary** representation of a number. (2) An indication of whether two numbers are both even or both odd (same parity), or are such that one is even and one is odd (opposite parity). This concept, of same or opposite parity, can be applied to any things that can have two different states–the sense in which a knot is tied, the color of a chessboard square, and so on.

Parrondo's paradox

Two losing gambling games can be set up so that when they are played one after the other, they become winning. This paradox is named after the Spanish physicist Juan Parrondo who, in the late 1990s, discovered how to construct such a scenario. The simplest way is to use three biased coins. Imagine you are standing on stair zero, in the middle of a long staircase with 1001 stairs numbered from –500 to 500. You win if you can get to the top of the staircase, and the way you move depends on the outcome of flipping one of two coins. Heads you move up a stair, tails you move down a stair. In game 1, you use coin *A,* which is slightly biased and comes up heads 49.5% of the time and tails 50.5%. Obviously, these are losing odds. In game 2, you use two coins, *B* and *C.* Coin *B* comes up heads only 9.5% of the time, tails 90.5%. Coin *C* comes up heads 74.5% of the time, tails 25.5%. In game 2 if the number of the stair you are on at the time is a multiple of 3 (that is, . . . , –9, –6, –3, 0, 3, 6, 9, 12, . . .), then you flip coin *B;* otherwise you flip coin *C.* Game 2, it turns out, is also a losing game and would eventually take you to the bottom of the stairs. What Parrondo found, however, is that if you play these two games in succession in random order, keeping your place on the staircase as you switch between games, you will steadily rise to the top of the staircase! This kind of process has been called a *Brownian motion*.[153]

partial differential equation

An equation that involves **derivative**s with respect to more than one variable. Many of the equations used to model the physics of the real world are partial differential equations.

partition number

A number that gives the number of ways of placing *n* indistinguishable balls into *n* indistinguishable urns. For example:

1: (*)
2: (**) (*)(*)
3: (***) (**)(*) (*)(*)(*)
5: (****) (***)(*) (**)(**) (**)(*)(*) (*)(*)(*)(*)
7: (*****) (****)(*) (***)(**) (***)(*)(*) (**)(**)(*)
(**)(*)(*)(*)(*)(*)(*)(*)(*)
11: (******) (*****)(*) (****)(**) (****)(*)(*)
(***)(***) (***)(**)(*)(***)(*)(*)(*)
(**)(**)(**) (**)(**)(*)(*) (**)(*)(*)(*)(*)
(*)(*)(*)(*)(*)(*)

The sequence runs: 1, 2, 3, 5, 7, 11, 15, 22, 30, 42, 56, 77, 101, 135, 176, 231, 297, 385, ... If the urns are distinguishable, the number of ways is 2^n. If the balls are distinguishable, the number of ways is given by the *n*th **Bell number**.

Pascal, Blaise (1623–1662)

A French mathematician, philosopher, and pioneer of **probability theory** whose short life was rich with mathematical invention but whose name, ironically, is most familiar from its association with an array of numbers known as **Pascal's triangle** (which he didn't discover, though he did important work on it). Educated by his father (after whom the **limaçon of Pascal** is named), Pascal showed his intellectual prowess early on by proving one of the most important theorems in **projective geometry** at the age of 16. Three years later he devised the world's second mechanical calculating machine (the first was made by Wilhelm Schickard in 1623) to help with his father's business; he sold about 50 of these "Pascalines," several of which survive. In 1654, he and Pierre de **Fermat**, in an exchange of correspondence, laid the foundation for **probability theory**. They considered the dice problem, already studied by Girolamo **Cardano**, and the problem of points also considered by Cardano and, around the same time, by Luca **Pacioli** and Niccoló **Tartaglia**. The dice problem asks how many times one can expect to throw a pair of dice before getting a double six; the problem of points asks how to divide the stakes if a game of dice is incomplete. Pascal and Fermat solved the problem of points for a two-player game but didn't develop powerful enough methods to solve it for three or more players. In the same year, Pascal was almost killed in an incident in which the horses pulling his carriage bolted and the carriage was left hanging over a bridge above the river Seine. Though rescued unharmed, he shortly after converted to the rigorous Jansenist sect of the Catholic Church. His philosophical work *Pensées,* written between 1656 and 1658 contains his famous argument, often called Pascal's wager, for belief in **God**. Having suffered poor health for most of his adult life, he died in great pain of cancer at the age of 39.

Pascal's mystic hexagon

If a **hexagon** *ADBFCE* (not necessarily convex) is inscribed in a **conic section** (in particular a circle), then the points of intersections of opposite sides (*AD* with *FC*, *DB* with *CE* and *BF* with *EA*) are collinear. This line is called the *Pascal line* of the hexagon. A special case is when the conic degenerates into two lines; the theorem still holds but is then usually called *Pappus's theorem.*

Pascal's triangle

A triangular pattern of numbers in which each number is equal to the sum of the two numbers immediately above it:

$$\begin{array}{c}
1\\
1\quad 1\\
1\quad 2\quad 1\\
1\quad 3\quad 3\quad 1\\
1\quad 4\quad 6\quad 4\quad 1\\
1\quad 5\quad 10\quad 10\quad 5\quad 1\\
1\quad 6\quad 15\quad 20\quad 15\quad 6\quad 1\\
1\quad 7\quad 21\quad 35\quad 35\quad 21\quad 7\quad 1\\
1\quad 8\quad 28\quad 56\quad 70\quad 56\quad 28\quad 8\quad 1\\
\vdots
\end{array}$$

Although named after Blaise **Pascal**, who studied it, this arithmetic triangle has been recognized since the twelfth century and has a variety of other names. In Italy it is called *Tartaglia's triangle* and in many parts of Asia it is referred to as *Yang Hui's triangle.* Yang Hui was a minor Chinese official who wrote two books, dated 1261 and 1275, that use decimal fractions (long before they appeared in the West) and contain one of the earliest accounts of the triangle. At about the same time, Omar **Khayyam** also wrote about it. The Chinese triangle appears again in 1303 on the front of Chu Shi-Chieh's *Ssu Yuan Yü Chien* (Precious mirror of the four elements), a book in which Chu says the triangle was known in China more than two centuries before his time.

The numbers in Pascal's triangle give the number of ways of picking r unordered outcomes from n possibilities. This is equivalent to saying that the numbers in each row are the **binomial coefficient**s in the expansion of $(x+y)^n$:

$$\begin{aligned}
(x+y)^0 &= 1\\
(x+y)^1 &= 1x + 1y\\
(x+y)^2 &= 1x^2 + 2xy + 1y^2\\
(x+y)^3 &= 1x^3 + 3x^2y + 3xy^2 + 1y^3\\
(x+y)^4 &= 1x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + 1y^4
\end{aligned}$$

and so on. In addition, the shallow diagonals of the triangle sum to give the **Fibonacci sequence**.

path

See **trajectory**.

pathological curve

A curve often specifically devised to show the falseness of certain intuitive concepts. In particular, the image of **continuity** as a smooth curve in our mind's eye severely misrepresents the situation and is the result of a bias stemming from an overexposure to the much smaller class of differentiable functions (see **differentiation**). A chief lesson of pathological curves is that continuity is a weaker notion than differentiability. Many pathological curves are **fractal**s, such as **Cantor dust**, including **space-filling curve**s, such as the **Peano curve**. The earliest known example is the graph of **Weierstrass' nondifferentiable function**.

payoff

In **game theory**, the amount that a player wins, given the player's and his opponent's actions.

Peano, Guiseppe (1858–1932)

An Italian pioneer of mathematical **logic** and the axiomatization of mathematics. In 1889, he published his first version of a system of mathematical logic in *Arithmetics Principia,* which included his famous **axiom**s of natural numbers, now known as *Peano's axioms.* Two years later, he established a journal, the *Rivista,* in which he proposed the symbolizing of all mathematical propositions into his system. The project, which became known as the *Formulario,* was his chief focus for the next fifteen years.

Peano curve

The first known example of a **space-filling curve**. Discovered by Guiseppe **Peano** in 1890, its effect was like that of an earthquake on the traditional structure of mathematics. Commenting in 1965 on the impact of the curve in Peano's day, N. Ya Vilenkin said: "Everything has come unstrung! It's difficult to put into words the effect that

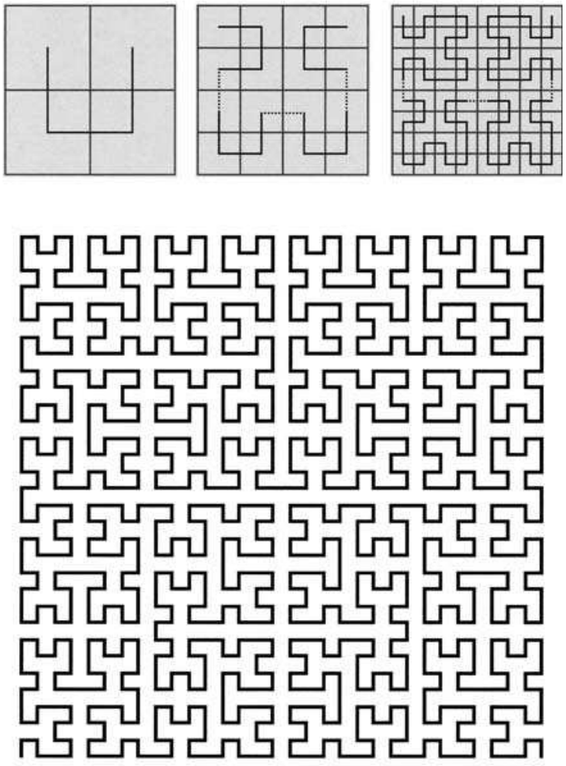

Peano curve A simple iterative process produces a remarkable space-filling shape.

Peano's result had on the mathematical world. It seemed that everything was in ruins, that all the basic mathematical concepts had lost their meaning." Today, the Peano curve is recognized as just one of an infinite class of familiar objects known as **fractal**s. But at the end of the nineteenth century it was an extravagant, completely counterintuitive thing; indeed, it was something that had been believed impossible. Writing of Peano's result in *Grundzüge der Mengenlehre* (Basic features of set theory) in 1914, Felix **Hausdorff** said: "This is one of the most remarkable facts of set theory." Originally, the Peano curve was derived purely analytically, without any kind of drawing or attempt at visualization. But the first few steps in drawing it, as shown in the diagrams, are easy enough, even though the finished product is unattainable in this way–and totally unimaginable. To fill the unit square, as the Peano curve does, without leaving any holes, the curve has to be both continuous and self-intersecting.

pearls of Sluze

Curves that are generated by the Cartesian equation

$$y^n = k(a - x)^p\, x^m,$$

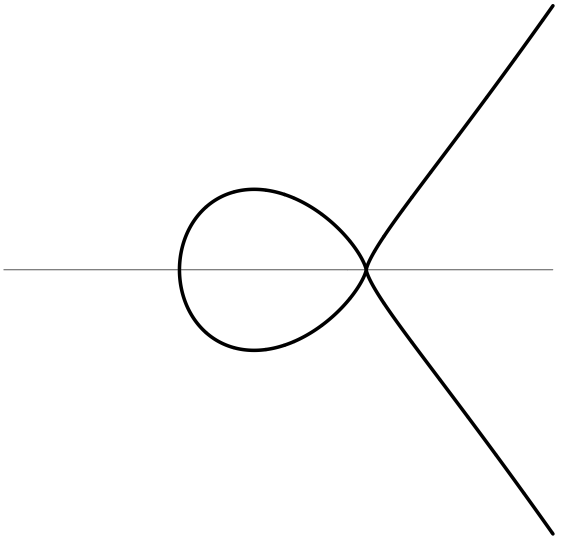

pearls of Sluze *Jan Wassenaar, www.2dcurves.com*

where n, m, and p are integers. They were first studied by the French mathematician René de Sluze (1622–1685) and named the pearls of Sluze by Blaise **Pascal**.

pedal curve

The locus of the feet of the perpendiculars from a given point to the **tangent**s to the given figure. The "given point" is known as the *pedal point.*

pedal triangle

The pedal triangle of a point P with respect to a triangle ABC is the triangle whose vertices (see **vertex**) are the feet of the perpendiculars dropped from P to the sides of triangle ABC.

peg solitaire

Also known as Hi-Q, a game that, in its commonest form, is played with 32 pegs or marbles on a rectangular grid, the middle position of which starts out empty. A peg may jump horizontally or vertically, but not diagonally, over a peg in an adjacent hole into a vacant hole immediately beyond. The peg that was jumped over is then removed. The object is to be left with a single peg in the center position.

The quickest solution of peg solitaire was found by Ernest Bergholt in 1912 and was proved to be minimal by John Beasley in 1964. Suppose the rows and columns of the board are each labeled 1 to 7 so that, for example, a peg on the fourth row and the third column is in position 43. The quickest solution is: (1) 46→44. (2) 65→45.

5 LEFT - Good
4 LEFT - Better
3 LEFT - Really Clever
2 LEFT - A Pat on the back
1 LEFT - Take a deep bow
PERFECT SCORE
1 LEFT in centre hole
RULES

peg solitaire A version of peg solitaire made of tin with marbles by the British firm MAR Toys in the 1950s. *Sue & Brian Young/Mr. Puzzle Australia, www.mrpuzzle.com.au*

(3) 57→55. (4) 54→56. (5) 52→54. (6) 73→53. (7) 43→63. (8) 75→73→53. (9) 35→55. (10) 15→35. (11) 23→43→63→65→45→25. (12) 37→57→55→53. (13) 31→33. (14) 34→32. (15) 51→31→33. (16) 13→15→35. (17) 36→34→32→52→54→34. (18) 24→44.[28]

Pell equation

An equation of the form $y^2 = ax^2 + 1$, where a is any positive whole number except a square number. The name comes from the English mathematician John Pell (1611–1685); however, he was wrongly credited. In writing about some of the work done on this type of equation, Leonhard **Euler** gave priority to Pell whereas, in fact, Pell had done no more than copy it in his papers from some of Pierre de **Fermat**'s letters. Fermat had been the first to state that an equation of this form always has an unlimited number of integer solutions. For example, the equation $y^2 = 92x^2 + 1$, has the solutions $x = 0, y = 1$; $x = 120, y = 1{,}151$; $x = 276{,}240, y = 2{,}649{,}601$; and so on. Each successive solution is about 2,300 times the previous solution. In fact, the solutions are every eighth partial fraction (where x is the numerator and y is the denominator) of the **continued fraction** for $\sqrt{92}$. A Pell equation was used in finding the solution to **Archimedes's cattle problem**.

Pell numbers

Numbers, named after John Pell (1611–1685), that are similar to the **Fibonacci sequence** and are generated by the formula $A_n = 2\,A_{n-1} + A_{n-2}$. The sequence runs: 1; 2; 5; 12; 29; 70; 169; 408; 985; 2,378; 5,741; 13,860; 33,461; 80,782; 195,025; . . . . The ratio of successive terms approaches 1 plus the **square root of 2**.

Penrose, Roger (1931–)

An English mathematical physicist famous for his important contributions to cosmology and the physics of black holes, for his controversial views on the nature of human consciousness and its relationship to **quantum mechanics**, and for his work in the field of recreational mathematics. The **Penrose tiling** and the **Penrose triangle** are named after him (but not the **Penrose stairway**, which is named after his father). In his book *The Emperor's New Mind* he argues that there must be errors in the known laws of physics, notably in quantum mechanics, and that true **artificial intelligence** is impossible.[242] The latter claim is based on his assertion that humans can do things outside the power of formal logic systems, such as knowing the truth of unprovable statements, or solving the **halting problem** (claims that were originally made by the philosopher John Lucas of Merton College, Oxford). These are controversial views, with which most of the mathematical and computer science communities disagree.

Penrose stairway

An **impossible figure** named after the British geneticist Lionel Penrose (1898–1972), father of Roger **Penrose**. It served as an inspiration for the staircase in M. C. **Escher**'s famous print "Ascending and Descending." Although the Penrose stairway cannot be realized in

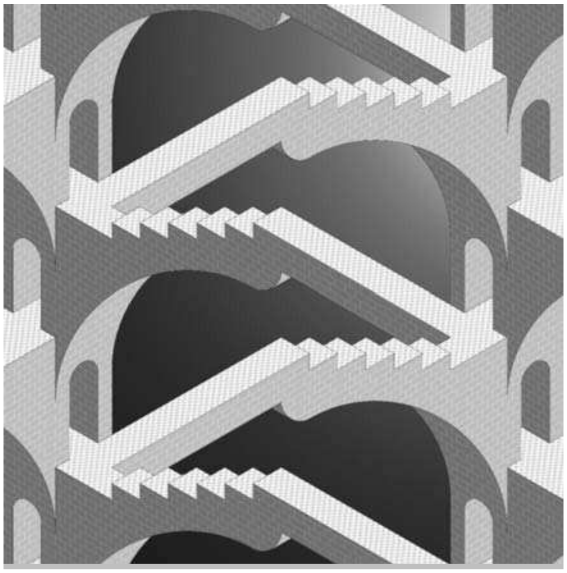

Penrose stairway Steps that ascend and descend simultaneously. *Jos Leys, www.josleys.com*

three dimensions, this impossibility is not immediately perceived and, in fact, the paradox is not even apparent to many people at a quick glance. Although Escher and the Penroses made the Stairway famous, it was, unbeknownst to them, independently discovered and refined years before by the Swedish artist Oscar **Reutersvärd**. In the 1960s the Stanford psychologist Roger Shepard created an auditory analogue of the Stairway. See also **Penrose triangle**.

Penrose tiling

A kind of **aperiodic tiling** discovered by Roger **Penrose**. In 1973 he announced a tiling made from a set of six tiles and then, by slicing and re-gluing, was able to reduce the number of tiles to just two. The most elegant of Penrose tilings use two rhombi (see **rhombus**), a thick one called a "kite" and a thin one called a "dart," which are fitted together so that no two tiles are aligned to form a single parallelogram (otherwise, a single rhombus could be used to make a periodic tiling). All angles are multiples of π/5 radians (36°, or one-tenth of a circle). Each tile has four sides with a length of one unit. One tile has four corners with the angles 72°, 72°, 108°, and 108° (2, 2, 3, and 3 multiples of 36°); the other has angles of 36°, 36°, 144°, and 144° (1, 1, 4, and 4 multiples of 36°). On each tile one of the vertices (corners) is colored black and two of the sides are marked with arrows. The only rules for assembling the tiles to ensure an aperiodic tiling are that two adjacent vertices must be of the same color, and two adjacent edges must have arrows pointing in the same direction or have no arrows at all. These rules ensure that, taken over a large enough area of the plane, the pattern of tiles doesn't repeat. In a correct Penrose tiling, the ratio of kites to darts is always the same and is equal to the **golden ratio**. Although Penrose tilings started out as nothing more than an interesting mathematical diver-

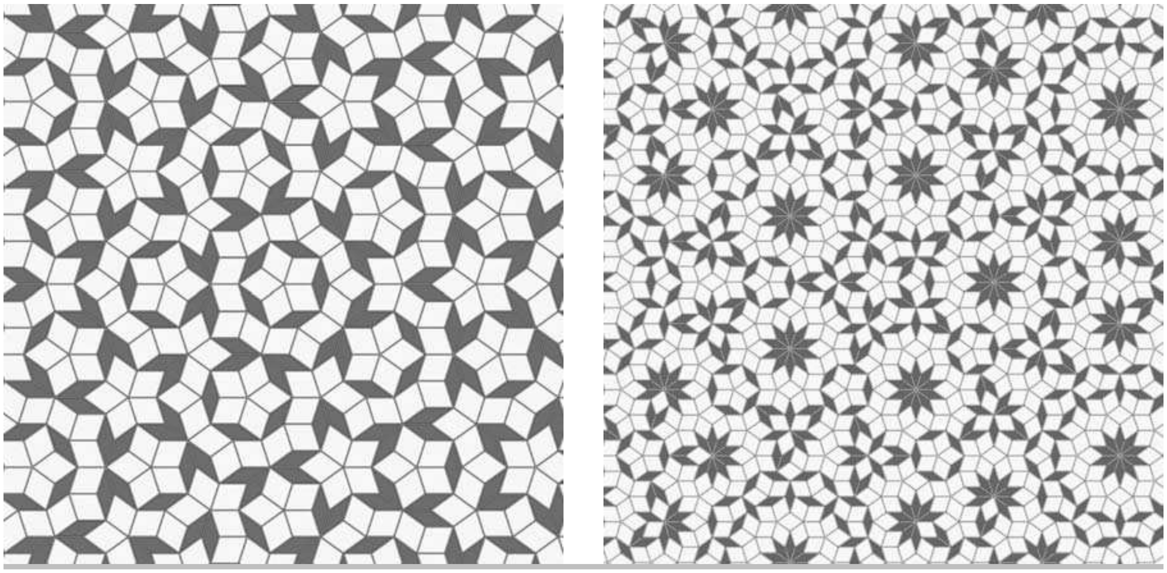

Penrose tiling Two of an infinite variety of nonperiodic patterns possible using Penrose tiles. *Jos Leys, www.josleys.com*

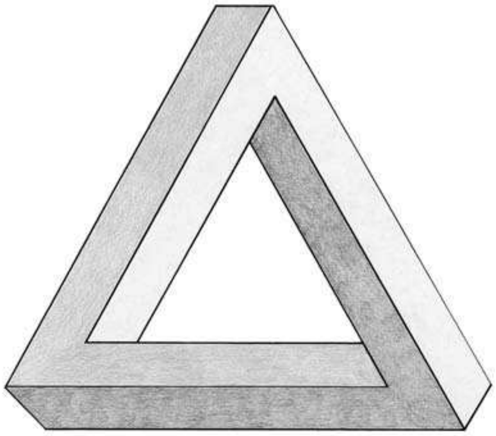

Penrose triangle

sion, they have turned out to mimic the arrangements of atoms in some newly discovered materials, known as **quasicrystal**s.[127]

Penrose triangle

The most famous and one of the simplest **impossible figure**s. Its roots go back to 1934 when Oscar **Reutersvärd** made the first recognizable impossible triangle out of a strange two-dimensional representation of cubes; this artwork appeared on a Swedish postage stamp issued in 1982. In 1954, Roger **Penrose**, after attending a lecture by the artist M. C. **Escher**, rediscovered the impossible triangle and drew it in its most familiar form, which he published in a 1958 article in the *British Journal of Psychology,* coauthored with his father Lionel.[241] Penrose was unfamiliar with the work of Reutersvärd, Giovanni Piranesi, and others who had previously created impossible figures. Penrose's impossible triangle, unlike Reutersvärd's earlier version, was drawn in perspective, which added an additional size paradox to the object. In 1961, Escher, inspired by Penrose's version of the impossible triangle (he was sent a copy of the article by the Penroses), incorporated it into his famous lithograph *Waterfall.*

penta-

The Greek prefix meaning "five." A *pentagon* is a five-sided **polygon** and a *pentahedron,* such as the Great Pyramid of Gisa, is a five-sided **polyhedron**. A *pentagram,* or *pentacle,* is a five-pointed star formed by the pairwise extension of the sides of a regular polygon so that each side meets not with the next side but with the one after that. The Pythagoreans (see **Pythagoras of Samos**) used the pentagram as a secret identification emblem; later it became a trademark of alchemists and, perhaps because of its repeating properties, a sign of the occult. According to legend, a pentagram was used by Doctor Faustus to exorcise Mephistopheles. *Pentagonal numbers* are **figurate number**s (numbers that can be represented by a regular geometric arrangement of equally spaced points) of the form $n(3n - 1)/2$; the first few are 1, 5, 12, 22, ...

pentomino

A five-square **polyomino**.

Perelman, Yakov Isidorovitch (1882–1942)

A Russian scientist and exponent of recreational mathematics whose stature in his own country was, and remains, similar to that of Martin **Gardner** in the United States. Between 1913 and his death, Perelman authored a dozen books and scores of articles covering many different aspects of popular mathematics, physics, and astronomy. His books have been republished so often in Russia that Perelman is universally known among both amateur and professional mathematicians across several generations. He was also a leading proponent of the ideas of the spaceflight visionary Konstantin Tsiolkovsky.

perfect cube

An integer of the form n^3 where n is an integer.

perfect cuboid problem

See **cuboid**.

perfect number

A whole number that is equal to the sum of all its factors except itself. For example, 6 is a perfect number because its factors, 1, 2, and 3 add to give 6. The next smallest is 28 (the sum of 1 + 2 + 4 + 7 + 14). Augustine (354–430) argued that God could have created the world in an instant but chose to do it in a perfect number of days, 6. Early Jewish commentators saw the perfection of the universe in the Moon's period of 28 days. The next in line are 496, 8,128, and 33,550,336. As René **Descartes** pointed out: "Perfect numbers like perfect men are very rare." All end in six or eight, though what seems to be an alternating pattern of sixes and eights for the first few perfect numbers doesn't continue. All are of the form $2^{n-1}(2^n - 1)$, where $2^n - 1$ is a **Mersenne prime**, so that the search for perfect numbers is the search for Mersenne primes. The largest one found is $2^{3021376}(2^{3021377} - 1)$. It isn't known if there are infinitely many perfect numbers or if there are any odd perfect numbers. A *pseudoperfect number* or *semi-perfect number* is a number equal to the sum of some of its divisors, for example, 12 = 2 + 4 + 6, 20 = 1 + 4 + 5 + 10. An *irreducible semi-perfect number* is a semi-perfect number, none of whose factors is semi-perfect, for example, 104. A *quasi-perfect number* would be a number n whose divisors (excluding itself) sum to $n + 1$, but it isn't known if such a number exists. A *multiply perfect number* is a number n whose divisors sum to a multiple of n. An example is 120, whose divisors (including itself) sum to $360 = 3 \times 120$. If the divisors sum to $3n$, n is called multiply perfect of order 3, or *tri-perfect.* Ordinary perfect numbers are multiply perfect of order 2. Multiply perfect numbers are known of order up to 8. See also **abundant number**.

perfect power

An integer of the form m^n where m and n are integers and $n > 1$.

perfect square

A number that is the product of two equal whole numbers; for example, $1 = 1 \times 1$, $4 = 2 \times 2$, $9 = 3 \times 3$, $16 = 4 \times 4$.

Perigal, Henry (1801–1898)

An English amateur mathematician best known for his elegant **dissection** proof of **Pythagoras's theorem**, a diagram of which is carved on his gravestone. Perigal also discovered a number of other interesting geometrical dissections and, though employed modestly for much of his life as a stockbroker's clerk, was well known in British scientific society.

perimeter

The distance around a two-dimensional shape.

period, of a decimal expansion

The length of the smallest block of repeating digits in the decimal expansion of a **rational number** that does not terminate. For example:

$1/3 = 0.333333333333\ldots$
repeating block = 3, period = 1
$5/7 = 0.71428571428571\ldots$
repeating block = 714285, period = 6
$89/26 = 3.4230769230769\ldots$
repeating block = 230769, period = 6

periodic

Refers to motion or to an entity that goes through a finite number of regions, returns to a previous state, and repeats the same fixed pattern forever.

periodic attractor

Also called a *limit cycle attractor,* an **attractor** that consists of a periodic movement back and forth between two or more values. The periodic attractor represents more possibilities for system behavior than does the **fixed-point attractor**. An example of a period two attractor is the oscillating movement of a metronome or, in psychiatry, a bipolar disorder that causes a person's mood to shift back and forth from elation to depression.

periodic tiling

A **tiling** in which a region can be outlined that tiles the plane by translation, that is, by shifting the position of the region without rotating or reflecting it. M. C. **Escher** is famous for his many pictures of periodic tilings with shapes that resemble living things. An infinity of shapes–for instance, the regular hexagon–tile only periodically, though all these fall into 17 distinct **wallpaper group**s. An infinity of other shapes tile both periodically and aperiodically. But it was only quite recently that the first **aperiodic tiling**s were discovered.

periphery

A curved **perimeter**.

permutable prime

Also known as an *absolute prime,* a **prime number** with at least two distinct digits, which remains prime on every rearrangement (permutation) of the digits. For example, 337 is a permutable prime because 337, 373, and 733 are all prime. Most likely, in base 10, the only permutable primes are 13, 17, 37, 79, 113, 199, 337, and their permutations. Permutable primes can't have any of the digits 2, 4, 6, 8 or 5, nor can they have all four of the digits 1, 3, 7, and 9 simultaneously.

permutation

A particular ordering of a collection of objects. For example, if an athlete has won three medals, a bronze one (B), a silver one (S), and a gold one (G), there are six ways they can be permuted or lined up: BSG, BGS, SBG, SGB, GBS, and GSB. If six people want to sit on the same park bench, there are 720 ways in which they can organize themselves. In general, n things can be permuted in $n \times (n - 1) \times (n - 2) \times \ldots \times 2 \times 1 = n!$ ways (where "!" is the symbol for **factorial**). How about if there are n distinct objects but we want to permute them in groups of k (where $k \leq n$): how many ways can that be done? The first member of the group can be picked in n ways because there are n objects to pick from. The second member can be filled in $(n - 1)$ ways since one of the n elements has already been taken. The third member can be filled in $(n - 2)$ ways since 2 elements have already been used, and so on. This pattern continues until k things have been chosen. This means that the last member can be filled in $(n - k + 1)$ ways. Therefore a total of $n(n - 1)(n - 2)\ldots(n - k + 1)$ different permutations of k

objects, taken from a pool of *n* objects, exist. If we denote this number by $P(n, k)$, we can write $P(n, k) = n! / (n - k)!$

perpendicular

At right angles. Two lines, planes, etc., are said to be perpendicular if they are 90° apart.

Perrin sequence

The sequence of integers defined by the recurrence relation $P(n) = P(n - 2) + P(n - 3)$, with the initial conditions $P(0) = 3$, $P(1) = 0$, $P(2) = 2$. Although the sequence is named after R. Perrin who studied it in 1899, it had been explored earlier, in 1876, by Édouard **Lucas**. As in the case of the similar *Padovan sequence,* the ratio of consecutive Perrin numbers tends toward the **plastic number**. More importantly, it appears that *n* divides $P(n)$ exactly if and only if *n* is a **prime number**. For example, 19 is prime, $P(19) = 209$ and $^{209}/_{19} = 11$; on the other hand, 18 is composite, $P(18) = 158$ and $^{158}/_{18} = 8.777$, which is not a whole number. Lucas conjectured that this is true for all values of *n* so that the Perrin sequence can be used as a test for non-primality: any number *n* that does not divide $P(n)$ is composite. Whether this conjecture is correct remains an open question. No one has ever found a composite *n* that divides $A(n)$, nor has anyone been able to prove that such numbers, known as *Perrin pseudoprimes,* don't exist. In 1991 Steven Arno of the Supercomputing Research Center in Bowie, Maryland, proved that Perrin pseudoprimes must have at least 15 digits.[15] The conjecture that no Perrin pseudoprimes exist is important, because the remainder on dividing $P(n)$ by *n* can be calculated very rapidly.

perturbation

A slight nudge that, temporarily or permanently, displaces an object or system out of **equilibrium**.

peta-

Prefix for 10^{15}, from the Greek *pentakis,* meaning "five times."

phase space

The mathematical space of all possibilities in a given situation. A motion is then described by a *path, trajectory,* or *orbit* in this space. This is not the usual kind of path laid out on the ground, but a series of locations in phase space, describing motion or change over a period of time. The terms do, however, recall the origins of qualitative dynamics in Henri **Poincaré**'s study of planetary motion. The *dimension* of the phase space is the number of initial conditions needed to uniquely specify a path and is equal to the number of variables in the dynamical system. The temporal behavior of the system is viewed as the succession of states in the system's state space. In the case of a simple pendulum, for example, the instantaneous configuration is given by just two numbers–the position of the pendulum bob and its velocity–which completely describe the system's state. For more complex systems, such as a chain of *n* pendulums coupled together, the state of the system is much larger. It requires, in this case, 2*n* numbers to specify the state of the entire system. This collection of all possible configurations is the phase space.

phase transition

In physics, a change from one state of matter to another. In **dynamical system**s theory, a change from one mode of behavior to another.

phi

See **golden ratio**.

Phillips, Hubert (1891–1964)

An English compiler of crosswords and word puzzles who wrote under the pseudonyms "Dogberry" in the *News Chronicle* magazine and "Caliban" in the *New Statesman.* He was also a prolific writer of epigrams, parodies, and satirical verse, and appeared on many radio quizzes. Phillips earned a first class degree in history at Oxford, served in the army throughout World War I, taught economics at Bristol University, and was active in the British Liberal Party. He was an accomplished player of contract bridge, captaining England in 1937 and 1938. Among his many publications are *Caliban's Problem Book, The Complete Book of Card Games, Brush Up Your Wits, My Best Puzzles in Mathematics,* over a hundred crime-problem stories, and a novel, *Charteris Royal.*[249, 250, 251]

Phutball

Also known as *philosopher's football,* a two-player board **game** first described by Elwyn Berlekamp, John **Conway**, and Richard **Guy**, in their book *Winning Ways for your Mathematical Plays.* Phutball is played on the intersections of a 19 × 15 grid using one white piece and as many black pieces as necessary. The objective is to score goals by using the men (black pieces) to move the football (white piece) onto or over the opponent's goal line, that is, either rows 1 or 0 or rows 19 and 20 (rows 0 and 20 being off the board). At the start of the game the football is placed on the central point, although a handicapping scheme exists where the ball can start nearer to the stronger player's goal line. Players alternate making moves, which consist either of adding a man to any vacant point on the board or of moving the ball. There is no difference between men played by the two opponents. The football is moved by a series of jumps over adjacent men. One jump is from the football's current point to the first vacant point in a straight line orthogonally or diagonally over one or more men. The jumped men are then removed from the board (before any subsequent jump occurs). Jumping is optional and can be repeated for as long as there are men available to be jumped and the player desires to do so. In contrast to checkers, multiple men in a row are jumped and removed as a group. If the football ends the move on or over the opponent's goal line then a goal has been scored. If the football passes through a goal line, but ends up elsewhere due to further jumps, the game continues. The game is sufficiently complex that, although theoretically one of the players has a winning strategy, this strategy is not known.

pi (π)

The ratio of the circumference to the diameter of a **circle**. Pi, sometimes known as *Archimedes' constant,* is one of the most important and ubiquitous numbers in mathematics, popping up in all kinds of seemingly unrelated areas. Its approximate value is 3.1419 but, being an **irrational number**, it can't be written as a terminating or recurring decimal. The fact that π can't be expressed as the ratio of two integers was proved by Johann Lambert in 1761. Then, in 1881, Ferdinand Lindemann showed that it is a **transcendental number**, meaning that there is no polynomial with integer or rational coefficients of which π is a root. Because of this, it's impossible to write π in terms of *any* finite number of integers or fractions or their roots. From the outset, this dooms any attempt at **squaring the circle**, that is, constructing a square whose area is equal to the area of a given circle, using straightedge and compass alone.

The association of the Greek letter π with the number 3.141 . . . started with William **Oughtred** in 1647 who used π.δ as a stand-in for "periphery-diameter" in his *Clavis mathematicae.* The Welsh mathematician William Jones was the first to use π as a symbol on its own in his 1706 *Synopsis palmariorum matheseos.* But it fell to the great Leonhard **Euler** to popularize the notation 30 years later.

Calculating π with greater and greater precision became a popular pursuit. Around 1600, Ludolph van Ceulen computed the first 35 decimals and was so proud of his accomplishment that he had the number inscribed on his tombstone. In 1789, the Slovene mathematician Jurij Vega extended known π to 140 places (of which only 137 were correct), while, in 1873, William Shanks pushed back the frontier to 707 places–a record that stood until 1949 when the first electronic computers appeared on the scene. In September 2002, Yasumasa Kanada and his colleagues at Tokyo University used 400 hours of supercomputer time to calculate π to 1.24 trillion places, beating their previous best, set in 1999, of 206 billion places. Although such an exercise may seem pointless, it serves as a benchmark for new high-speed computers and algorithms, and also to test the long-standing, but still unproven assertion that the distribution of digits in π is completely random.

For the record, the first 100 digits of π are: 3.1415926535 8979323846 2643383279 5028841971 6939937510 5820974944 5923078164 0628620899 8628034825 3421170679. The wonderfully pointless pastime of "piphilology" is concerned with devising mnemonics for remembering some of these: the number of letters in each word of the mnemonic gives the corresponding digit. A famous example is Isaac Asimov's "How I want a drink, alcoholic of course, after the heavy lectures involving quantum mechanics!" For the more ambitious, there is Michael Keith's 1995 poem that supplies the first 42 digits:

Poe, E. Near A Raven
Midnights so dreary, tired and weary,
Silently pondering volumes extolling all by-now
obsolete lore.
During my rather long nap, the weirdest tap!
An ominous vibrating sound disturbing my chamber's antedoor.
"This," I whispered quietly, "I ignore."

There are some reasonably good approximations to π in the form of ordinary (rational) fractions. The best known is $^{22}/_{7}$, but this is only accurate to two decimal places. A fraction with a larger denominator offers a better chance of getting a more refined estimate. There is also $^{333}/_{106}$, which is good to 5 places. But an outstanding approximation is $^{355}/_{113}$ which is accurate to 6 places; in fact, there is no better approximation among all fractions with denominators less than 30,000. See also **continued fraction**.

In 1897 the General Assembly of the state of Indiana tried to pass legislation to the effect that the exact value of π is 3.2. The bill was referred, for some bizarre reason, to the House Committee on Canals. But then the true motive for the attempted change to the law became clear. By chance a professor of mathematics happened to be present during a debate and heard an ex-teacher saying "The case is perfectly simple. If we pass this bill, which establishes a new and correct value for π, the author offers to our state without cost the use of his discovery and its free publication in our school textbooks, while everyone else must pay him a royalty." Fortunately, the professor was able to teach the senators a little math and the bill was stopped in its tracks.

Any formulas to do with circles or spheres, not surprisingly, have π in them. For example, the circumference of a circle of radius r is $2\pi r$, the area of a circle is πr^2, the volume of a sphere is $(4/3)\pi r^3$, and the surface area of a sphere is $4\pi r^2$. But π also has a strange habit of appearing in the most unexpected places. In the eighteenth century, the French naturalist Comte de Buffon showed that you could estimate π by experiment using the unlikely apparatus of a repeatedly dropped needle and a shove-halfpenny board (see **Buffon's needle**). Two of the most important, but (as far as is known) unrelated, equations in modern physics, Heisenberg's uncertainty principle ($\Delta x\ \Delta p = h/4\pi$) in **quantum mechanics** and Einstein's field equation ($R_{ik} - \frac{1}{2} g_{ik} R + \Lambda g_{ik} = 8\pi G/c^4\ T_{ik}$) in general **relativity theory** include π. The ever-present constant also emerges as a result of various remarkable **infinite series**. These include:

$1/1^2 + 1/2^2 + 1/3^2 + 1/4^2 + \ldots = \pi^2/6$
(found by Leonhard **Euler**)

$1/1 - 1/3 + 1/5 - 1/7 + 1/9 - \ldots = \pi/4$
(found by Gottfried **Leibniz**)

$2/1 \times 2/3 \times 4/3 \times 4/5 \times 6/5 \times 6/7 \times 8/7 \times 8/9 \times \ldots = \pi/2$
(found by John **Wallis**)

Among many other formulas that give rise to or contain π are

the integral

$$\int_{-\infty}^{\infty} e^{-x^2}\, \mathrm{d}x = \sqrt{\pi},$$

Stirling's formula, $n! \sim \sqrt{2\ \pi\ n}\ \ (n/e)^n$ and, the most intriguing equation in mathematics, $e^{i\pi} + 1 = 0$.

In **number theory**, the probability that two randomly chosen integers have no common divisors (in other words, that they're *relatively prime*) is $6/\pi^2$ or 1 in 1.644934 . . . , and the average number of ways to write a positive integer as the sum of two **perfect square**s is π/4. Both of these facts are astonishing because it is hard to see how a basically geometric constant to do with circles has any bearing at all on how various types of numbers are distributed. Deep truths are buried here!

In 1995, David Bailey, Peter Borwein, and Simon Plouffe of the University of Quebec at Montreal discovered a new formula for π as an infinite series:[21]

$$\pi = \sum_{k=0}^{8} \frac{1}{16^k}\left[\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6}\right]$$

What is so remarkable about this is that it enables the calculation of *isolated digits* of π–say, the trillionth digit–without computing and keeping track of all the preceding digits. How such a formula could possibly arise constitutes a mystery in itself. The only catch is that the formula works for base 2 (**binary**) and 16 (**hexadecimal**) but not base 10. So, it's possible to use the formula to determine, for example, that the five-trillionth binary digit of π is 0, but there's no way to convert the result into its decimal equivalent without knowing all the binary digits that come before the one of interest. The new formula allows the calculation of the *n*th base 2 or base 16 digit of π in a time that is essentially linear in *n,* with memory requirements that grow logarithmically (very slowly) in *n.* One possible use of the Bailey-Borwein-Plouffe formula is to help shed light on whether the distribution of π's digits are truly random, as most mathematicians suppose.[29, 42, 68, 114] See also **pyramid**.

Pick's theorem

First published in 1899, a theorem that was brought to broad attention as recently as 1969 through Hugo **Steinhaus**'s popular book *Mathematical Snapshots.*[316] Pick's theorem gives an elegant formula for the area of *lattice polygons*–**polygon**s that have vertices located at the integral nodes of a square grid or lattice that are spaced a unit distance from their immediate neighbors. Pick's theorem says that the area of such a polygon can be found simply by counting the lattice points on the interior and boundary of the polygon. The area is given by

$$i + (b/2) - 1,$$

where i is the number of interior lattice points and b is the number of boundary lattice points. The Austrian mathematician Georg Pick (1859–1942) after whom the result is named, was born in Vienna and perished during World War II in the Theresienstadt concentration camp. Over the past few decades, beginning with a paper by J. E. Reeve in 1957, various generalizations of Pick's theorem have been made to more general polygons, to higher-dimensional polyhedra, and to lattices other than square lattices. Most recently, mathematicians have become interested in the theorem because it provides a link between traditional Euclidean geometry and the modern subject of digital (discrete) geometry.

pico-

Prefix meaning a trillionth (10^{-12}), from the Italian *piccolo,* meaning "small."

pitch drop experiment

The world's longest running experiment. It started in 1927 when Thomas Parnell, the first professor of physics at the University of Queensland in Brisbane, Australia, heated a sample of pitch (a derivative of tar) and poured it into a glass funnel with a sealed stem. Three years were allowed for the pitch to settle, then, in 1930, the sealed stem was cut. From that date on the pitch has slowly dripped out of the funnel–so slowly that, up to the present time, only eight drops have fallen! The experiment

pitch drop experiment Professor John Mainstone alongside the world's longest-running experiment. *John Mainstone, University of Queensland*

stands in a display cabinet in the foyer of the Department of Physics at the University of Queensland demonstrating for all to see the fact that pitch, though it feels like a solid and is brittle enough to smash with a hammer, is really a fluid of very high viscosity (about 100 billion times as viscous as water).[91]

pizza

Order a pizza, pick an arbitrary point in it, and cut the pizza into eight slices by slicing at 45° angles through the chosen point. Color the alternate pieces red, with ketchup, and yellow, with mustard. Measure the area of the red slices and the yellow slices. Surprisingly, it will be found that these areas are the same! This fact holds true for any multiple of four slices cut by using equal angles through a fixed arbitrary point in the pizza. See also **tautology**.

place-value system

A **number system** in which the value of a number symbol depends not only on the symbol itself but also on the position where it occurs.

Planck time

In **quantum mechanics**, the shortest meaningful period of **time**; any two events that are separated by less than this amount of time can be considered simultaneous. It has the value 5.390×10^{-44} second. Related to this is the *Planck length* of 6.160×10^{-35} meters, which is the distance that light can travel in the Planck time.

plane

A flat surface such that a straight **line** joining any two points on it will also lie entirely on the surface.

plane partition

A stack of unit cubes in a rectangular box or in the positive **octant** in space such that to the left, behind, and below every cube lies either another cube or a wall. A plane partition in a box is equivalent to a **lozenge** tiling of a hexagon in the plane.

plastic number

A little-known number that has much in common with the **golden ratio** in that it is closely linked to architecture and to aesthetics. The concept of the plastic number was first described by the Dutchman Hans van der Laan (1904–1991) in 1928, shortly after he had abandoned his

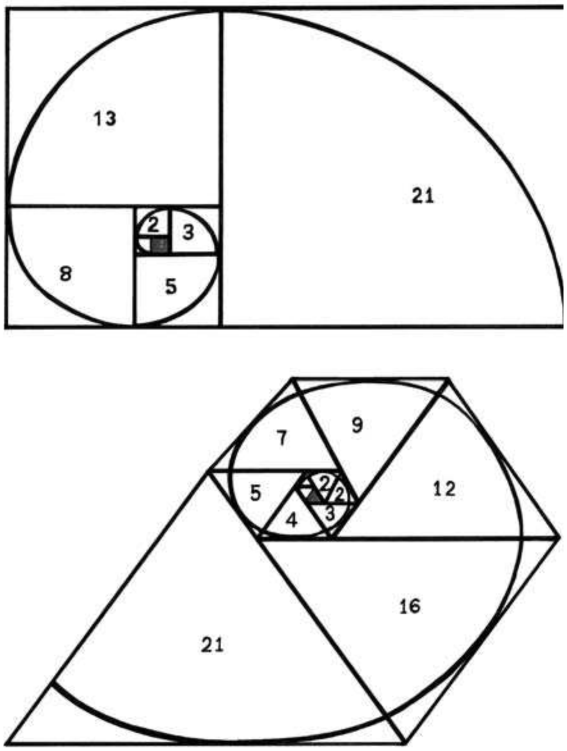

plastic number Spiraling systems illustrate the Fibonacci numbers (top) and the Padovan sequence.

architectural studies and become a novice monk, and has subsequently been explored by the English architect Richard Padovan (1935–). It is derived from a **cubic** equation, rather than a **quadratic** in the case of the golden ratio, and is intimately linked to two ratios, approximately 3:4 and 1:7, which van der Laan considered fundamental in the relationship between human perception and shape and form. These ratios, he believed, express the lower and upper limits of our normal ability to perceive differences of size among three-dimensional objects. The lower limit is that at which things differ just enough to be of distinct types of size. The upper limit is that beyond which they differ too much to relate to each other; they then belong to different orders of size. According to van der Laan, these limits are precisely definable. The mutual proportion of three-dimensional things first becomes perceptible when the largest dimension of one thing equals the sum of the two smaller dimensions of the other. This initial proportion determines, in turn, the limit beyond which things cease to have any perceptible mutual relation.

In mathematical terms, the plastic number is the unique **real number** solution to the equation $x^3 - x - 1 = 0$ and has the approximate value 1.324718 . . . . Just as the golden ratio is approximated better and better by successive terms of the **Fibonacci sequence**, $F(n + 1) = F(n) + F(n - 1)$, where $F(0) = F(1) = 1$, so the plastic number arises as the limit of the ratio of successive numbers in the sequence $P(n + 1) = P(n - 1) + P(n - 2)$ where $P(0) = P(1) = P(2) = 1$. This sequence has been called the *Padovan sequence,* and its members the *Padovan numbers.* The Padovan sequence increases much more slowly than does the Fibonacci sequence. Some numbers, such as 3, 5, and 21, are common to both sequences; however, it is not known if there are finitely many or infinitely many such pairs. Some Padovan numbers, such as 9, 16, and 49, are perfect squares; the square roots of these (3, 4, and 7) are also Padovan numbers, but it is not known if this is a coincidence or a general rule. Another way to generate the Padovan numbers is to mimic the use of squares for Fibonacci numbers, but with cuboid structures–boxes with rectangular faces. A kind of three-dimensional spiral of boxes emerges. Start with a cube of side 1 and place another adjacent to it: the result is a 1 × 1 × 2 cuboid. On the 1 × 2 face, add another 1 × 1 × 2 box to produce a 1 × 2 × 2 cuboid. Then on a 2 × 2 face, add a 2 × 2 × 2 cube to form a 2 × 2 × 3 cuboid overall. To a 2 × 3 face, add a 2 × 2 × 3 to get a 2 × 3 × 4 box overall, and so on. Continue this process, always adding cuboids in the sequence east, south, down, west, north, and up. At each stage the new cuboid formed will have three consecutive Padovan numbers as its sides. Moreover, if successive square faces of the added cuboids are connected by straight lines, the result is a spiral that lies in a plane. The Padovan sequence is very similar to the **Perrin sequence**.[319]

Plateau, Joseph Antoine Ferdinand (1801–1883)

A Belgian physicist who did pioneering experimental work on soap bubbles and soap films, which stimulated the mathematical study of bubbles as **minimal surface**s. In 1829 Plateau carried out an optical experiment that involved looking at the Sun for 25 seconds: this damaged his eyes, and he eventually became blind.

Plateau curves

A set of curves, named after Joseph **Plateau**, that are described by the parametric equations:

$$x = a \sin(m + n)t / \sin(m - n)t$$
$$y = 2a \sin(mt) \sin(nt) / \sin(m - n)t$$

If $m = 2n$ the Plateau curves become a circle with a center at (1, 0) and a radius of 2.

Plateau problem

The general problem of determining the shape of the **minimal surface** constrained by a given boundary. It is named after Joseph **Plateau** who noticed that a handful of simple patterns seemed to completely describe the geometry of how soap **bubble**s fit together. Plateau claimed that soap bubble surfaces always make contact in one of two ways: either three surfaces meet at 120° angles along a curve; or six surfaces meet at a vertex, forming angles of about 109°. For instance, in a cluster of bubbles, two intersecting bubbles (of possibly different sizes) will have a common dividing wall (the third surface) that meets the outer surfaces of the bubbles in 120° angles. On the other hand, the edges of the six soap-film faces that emerge within a tetrahedral wire frame, when dipped in a soapy solution, form angles of roughly 109° at a central vertex. Until the American mathematician Jean Taylor came along in the mid-1970s, Plateau's patterns were just a set of empirical rules. However, as a follow-up to her doctoral thesis, Taylor was able to prove that Plateau's rules were a necessary consequence of the energy-minimizing principle–no other yet unobserved configurations were possible–thus settling a question that had been open for more than a century. The forces acting along the surface of a soap bubble all have the same magnitude in all directions. In crystals this is not the case (magnitudes of surface forces differ in different directions, though they may exhibit a grain, analogous to that in a piece of wood), but they still require the least energy to enclose a given volume. Minimal surfaces that model these conditions, like a cube with its corners chopped off or the bottom half of a cone

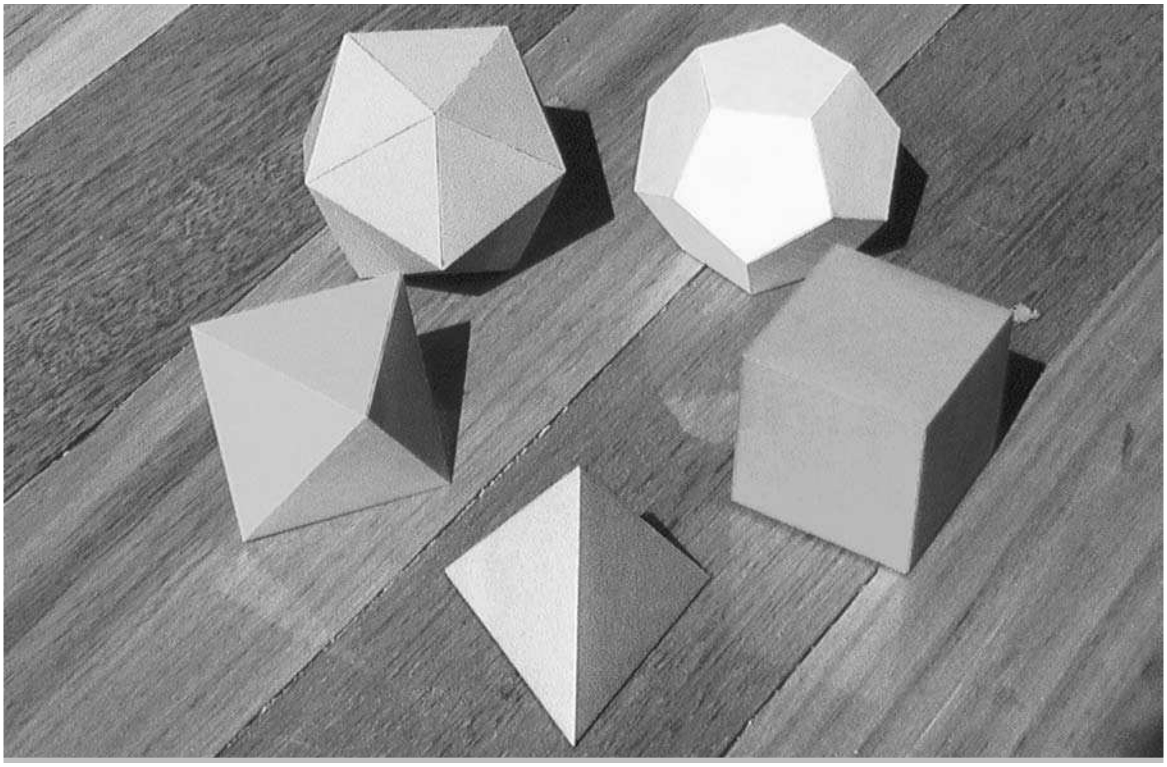

Platonic solid Clockwise from the extreme right: the cube, tetrahedron, octahedron, icosahedron, and dodecahedron. *Robert Webb, www.software3d.com; created using Webb's Stella program*

mounted on a cylinder, are known as *Wulff Shapes,* and provide fertile ground for mathematical study today.[328]

Platonic solid

Any of the five regular **polyhedron**s–solids with regular **polygon** faces and the same number of faces meeting at each corner–that are possible in three dimensions. They are the **tetrahedron** (a pyramid with triangular faces), the **octahedron** (an eight-sided figure with triangular faces), the **dodecahedron** (a 12-sided figure with pentagonal faces), the **icosahedron** (a 20-sided figure with triangular faces), and the **cube** (see table, "The Platonic Solids"). They are named after **Plato** who described them in one of his books, though it was **Euclid** who proved that there are only five regular polyhedra. A regular solid with hexagonal faces cannot exist because if it did, the sum of the angles of any three hexagonal corners that meet would already equal 360°, so such an object would be

The Platonic Solids

	Number of			
Name	**Faces**	**Edges**	**Vertices**	**Schläfi Symbol**
Tetrahedron	4	6	4	{3,3}
Cube	6	12	8	{4,3}
Octagon	8	12	6	{3,4}
Dodecahedron	12	30	20	{5,3}
Icosahedron	20	30	12	{3,5}

planar. See also **Archimedean solid**, **Catalan solid**, and **Johnson solid**.

Platonism

The belief that mathematical objects exist independent of physical models. It is, at the very least, a useful pretense in mathematics, especially in **geometry**.

pleated surface

A surface in **Euclidean space** or hyperbolic space (see **hyperbolic geometry**) that resembles a **polyhedron** in the sense that it has flat faces that meet along edges. Unlike a polyhedron, a pleated surface has no corners, but it may have infinitely many edges that form a **lamination**.

Poggendorff illusion

A **distortion illusion** in which the two ends of a straight line passing behind a rectangle appear offset when, in fact, they are aligned. It was discovered in 1860 by the physicist J. C. Poggendorff, editor of *Annalen der Physik und Chemie* (Annals of physics and chemistry), after receiving a letter from the astronomer Friedrich Zöllner. In this letter, Zöllner described an illusion (see **Zöllner illusion**) he had noticed on a fabric design in which parallel lines intersected by a pattern of short diagonal lines appear to diverge.[134]

Poincaré, (Jules) Henri (1854–1912)

A French theoretical physicist and mathematician who almost discovered **relativity theory** before Einstein and who is most famous in mathematics because of his hypothesis known as the **Poincaré conjecture**. Poincaré, who taught for most of his life at the University of Paris, was sometimes referred to as "the last universalist," because of his huge published output on a wide variety of mathematics and mathematical physics. Strangely, he was also clumsy, absent-minded, and inept at simple arithmetic. In order to translate problems in **topology** into questions in **algebra**, he devised (or discovered) ***homotopy** groups*–quantities that capture the essence of multidimensional spaces in algebraic terms, and have the power to reveal similarities between them.

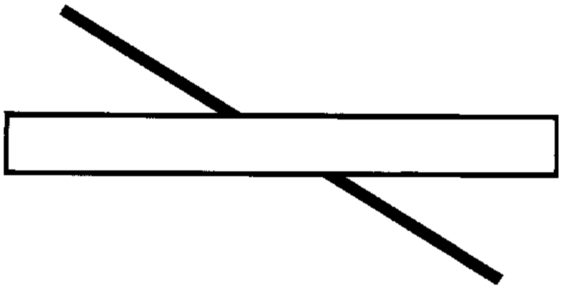

Poggendorff illusion

Poincaré conjecture

A proposition in **topology** put forward by Henri **Poincaré** in 1904. Poincaré was led to make his conjecture during his pioneering work in topology, the mathematical study of the properties of objects that stay unchanged when the objects are stretched or bent. In loose terms, the conjecture is that every three-dimensional object that has a set of spherelike properties (i.e., is topologically equivalent to a sphere) can be stretched or squeezed until it is a three-dimensional sphere (a 3-sphere) without tearing (i.e., making a hole in) it. Strictly speaking, the conjecture says that every **closed**, **simply connected** three-**manifold** is **homeomorphic** to the 3-sphere.

Poincaré proved the two-dimensional case and he guessed that the principle would hold in three dimensions. Determining if the Poincaré conjecture is correct has been widely judged the most important outstanding problem in topology–so important that, in 2000, the Clay Mathematics Institute in Boston named it as one of seven Millennium Prize Problems and offered a $1 million prize for its solution. Since the 1960s, mathematicians have shown by various means that the generalized conjecture is true for all dimensions higher than three–the four-dimensional case finally falling in 1982. But none of these strategies work in three dimensions. On April 7, 2002, came reports that the Poincaré conjecture might have been proved by Martin Dunwoody of Southampton University, but within a few days a fatal flaw was found in his proof. Then, in April 2003, what appears to be a genuine breakthrough emerged during a series of lectures delivered at the Massachusetts Institute of Technology by the Russian mathematician Grigori Perelman of the Steklov Institute of Mathematics (part of the Russian Academy of Sciences in St. Petersburg). His lectures, entitled "Ricci Flow and Geometrization of Three-Manifolds," constituted Perelman's first public discussion of important results contained in two earlier preprints. Mathematicians will now scrutinize the validity of Perelman's work (which does not actually mention the Poincaré conjecture by name). In any event, the Clay Institute requires a two-year cooling-off period before the prize can be awarded.

Poincaré disk

The region inside (but not including) a bounding circle, where straight lines are defined to be either diameters of the bounding circle, or arcs of circles that are perpendicular to the bounding circle. The Poincaré disk is a model

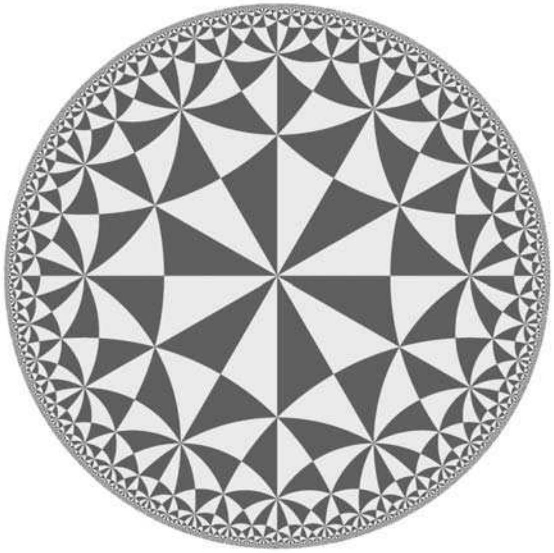

Poincaré disk A pattern on a hyperbolic surface. *Jos Leys, www.josleys.com*

of **hyperbolic geometry**. The great flexibility it allows in specifying the angles of a triangle leads to an infinite variety of ways in which the disk can be tiled.

Poinsot, Louis (1777–1859)

A French mathematician who invented *geometrical mechanics* while investigating how a system of forces acting on a rigid body can be resolved into a single force and a couple (a pair of equal and oppositely directed forces, as when you try to unscrew a bottle cap). Together with Gaspard **Monge**, he helped geometry regain its leading role in mathematical research in France in the eighteenth century. He wrote an important work on polyhedra in 1809, discovering four new regular polyhedra, which are now known as the **Kepler-Poinsot solids**. Two of these had been found by Johannes **Kepler** in 1619 but Poinsot was unaware of this; the two additional ones that Poinsot discovered were the *great dodecahedron* and the *great icosahedron.* In 1810 Augustin **Cauchy** proved that, with this definition of regular, the enumeration of regular polyhedra is complete (although a mistake was discovered in Poinsot's, and hence Cauchy's, definition in 1990 when an internal inconsistency became apparent). Poinsot also worked in number theory, studying **Diophantine equation**s with a view to expressing numbers as the difference of two squares and primitive roots.

point

A dimensionless geometric object having no properties other than location or place. More generally, an element in a geometrically described **set**.

point-set topology

Also known as *general topology,* a branch of **topology** concerned with how to put a structure on a **set** in such a way as to generalize the idea of **continuity** for maps from the real numbers to itself. A topology on a set X is a certain set of *open subsets* of the set X which satisfy various axioms. The set X together with this topology is called a **topological space**.

Poisson, Siméon Denis (1781–1840)

A French mathematician whose main interest lay in the application of mathematics to physics, especially in electrostatics and magnetism. He developed a two-fluid theory of electricity and provided theoretical support for the experimental results of others, notably Charles de Coulomb. Poisson also made important contributions to mechanics, especially the theory of elasticity; optics; calculus, especially definite integrals; differential geometry; and probability theory. In all, he wrote more than 300 papers on mathematics, physics, and astronomy, and his *Traité de Mécanique* (1811) was long a standard work.

polar coordinates

A **coordinate** system in which distances are measured from a fixed reference point (the pole) and angles from a fixed reference line. A *polar equation* is one that uses polar coordinates.

pole

(1) The **origin** of a system of **polar coordinates**. (2) In **complex analysis**, a function of a certain simple type of **singularity**. (3) One of the two points, where the axis of rotation of a rotating body, such as Earth, passes through the surface. (4) An old unit of length, also called a *rod,* equal to 5.5 yards.

Pólya's conjecture

A hypothesis put forward by the Hungarian mathematician George Pólya (1887–1985) in 1919. A positive integer is said to be of *even type* if it factorizes into an even number of **prime number**s; otherwise it is said to be of *odd type.* For example, 4, $= 2 \times 2$, is of even type, whereas 18, $= 2 \times 3 \times 3$, is of odd type. Let $O(n)$ be the number of odd type and $E(n)$ be the number of even type integers in the first n integers. Pólya's conjecture says that $O(n) \geq E(n)$ for all $n \geq 2$. After the conjecture had been checked for all values of n up to 1 million, many mathematicians assumed it was probably true. However, in 1942 A. E. Ingham came up with an ingenious method to show how a counterexample could be constructed, even though there wasn't enough computing power around at the time to do the necessary calculations.[177] Twenty years later, R. S. Lehman ran Ingham's method on a computer to find a counterexample to Pólya's conjecture at $n = 906180359$.[199]

polychoron

An unofficial name (from the Greek *poly* meaning "many" and *choros* meaning "room" or "space") for a four-dimensional **polytope**; it was first advocated by George Olshevsky.

polycube

A **polyhedron** formed by joining unit cubes by their faces. Examples of puzzles that involve polycubes are the **Soma cube** and **Rubik's cube**.

polygon

A plane closed figure whose sides are straight lines. The term *polygon* (from the Greek *poly* for "many" and *gwnos* for "angle") sometimes also refers to the *interior* of the polygon (the open area that this path encloses) or to the **union** of both. A polygon is *simple* if it is described by a single, nonintersecting boundary; otherwise it is said to be *complex.* A simple polygon is called *convex* if it has no internal angles greater than 180°; otherwise it is called *concave.* A polygon is called *regular* if all its sides are of equal length and all its angles are equal (see table, "Regular Polygons"). Any polygon, regular or irregular, has as many angles as it has sides. See **constructible** (for constructible polygons).

Regular Polygons

Name	Sides	Angle (= 180°–360°/sides)
Equilateral **triangle**	3	60°
Square	4	90°
Regular **penta**gon	5	108°
Regular **hexa**gon	6	120°
Regular heptagon	7	128.57° (approx.)
Regular **octa**gon	8	135°
Regular nonagon	9	140°
Regular **decagon**	10	144°
Regular hectagon	100	176.4°
Regular megagon	10^6	179.99964°
Regular **googol**gon	10^{100}	180° (approx.)

polygonal number

The number of equally spaced dots needed to draw a **polygon**. Polygonal numbers, which are a type of **figurate number**, include square numbers, triangular numbers, and hexagonal numbers.

polyhedron

A three-dimensional object whose faces are all **polygon**s and whose edges are shared by exactly two polygons. *Polyhedron* comes from the Greek *poly* for "many" and -*hedron* meaning "base," "seat," or "face." Every polyhedron in three-dimensional space consists of (two-dimensional) faces, (one-dimensional) edges, and (zero-dimensional) vertices. Sometimes the term *polyhedron* is used to apply to figures in more than three dimensions; however, analogs of polyhedra in the **fourth dimension** or higher are also referred to as **polytope**s. Polyhedrons, like polygons, may be *convex* or *nonconvex.* If a line that connects any two points on the surface of a polyhedron is completely inside or on the polyhedron, the figure is convex. Otherwise, it is nonconvex or *concave.* A polygon is *regular* if all of its faces are exactly the same size and shape and if the same number of faces meet at each vertex. There are only five regular convex polyhedra–the **Platonic solid**s. Another four regular nonconvex polyhedra are called **Kepler-Poinsot solid**s. However, the term *regular polyhedra* is sometimes used to describe only the Platonic solids. A convex polyhedron is said to be *semi-regular* if its faces have a similar arrangement of nonintersecting regular plane convex polygons of two or more different types about each vertex. These solids, of which there are 13 different kinds, are commonly called the **Archimedean solids**. A *dual* of a polyhedron is another polyhedron in which faces and vertices occupy complementary locations. The duals of the Archimedean solids are known as the **Catalan solid**s. A **quasi-regular polyhedron** is the solid region interior to two dual regular polyhedra; only two exist: the cuboctahedron and the icoidodecahedron. There are also infinite families of **prism**s and **antiprism**s. In total there are 92 convex polyhedra with regular polygonal faces (and not necessarily equivalent vertices); these are the **Johnson solids**.

The oldest known examples of man-made polyhedra were found on the islands of northeastern Scotland and date back to Neolithic times, between 2000 and 3000 B.C. These stone figures are about 2 inches in diameter and many are carved into rounded forms of regular polyhedra. Examples including cubical, tetrahedral, octahedral, and dodecahedral forms, one which is the dual of the pentagonal prism, are on display in the Museum of Scotland and in Oxford's Ashmolean Museum.

polyiamond

A shape made from identical equilateral **triangle**s that have been joined at their edges.

polynomial

An expression in which whole-numbered powers of a variable multiplied by numerical coefficients are added together. The powers of the variable must be positive integers, or zero. An example of a polynomial is $3x^3 + 7x - 2$. A polynomial *equation* has a polynomial expression, or zero, on each side of the equal sign: for example, $4a^2 - 5.6a + 1.7 = 0$ or $5x^2 - 1 = 8x + 2x^4$. The type of polynomial expression or equation is determined by the highest power present of the variable. A **quadratic**, for example, has nothing higher than a squared term. **Cubic**s, **quartic**s, and **quintic**s have maximum powers of three, four, and five, respectively. Mathematicians who have done pioneering work on each of these higher types of polynomial equations tended, for some reason, to have had colorful and star-crossed lives. Niccolo **Tartaglia**, who first solved the cubic, failed miserably as a mathematician for the rest of his life, largely because he spent it trying to discredit Girolamo **Cardano**. Tartaglia told Cardano his method of solution and swore him to secrecy but Cardano went ahead and published the solution anyway. Cardono himself lived a long unhappy life and his only son was executed for murder. Lodovico Ferrara, Cardano's student, who solved the general quartic, was poisoned, probably by his sister, over an inheritance dispute. Finally, Evariste **Galois**, who showed that the general quintic was unsolvable, died in a duel at the age of 20.

polyomino

A two-dimensional shape made by connecting *n* squares of the same size along their edges. The polyomino is a generalization of the **domino**, which can be thought of as a polyomino with $n = 2$. Several puzzles involving polyominoes made from three or more squares stuck together became popular around the beginning of the twentieth century. The best-known of these is the Broken Chessboard, presented as problem no. 74 in Henry **Dudeney**'s book *The Canterbury Puzzles* (1907).[87] Dudeney gives an amusing introduction to the problem, first quoting from Hayward's *Life of William the Conqueror* (1613) about an incident in which Prince Henry,

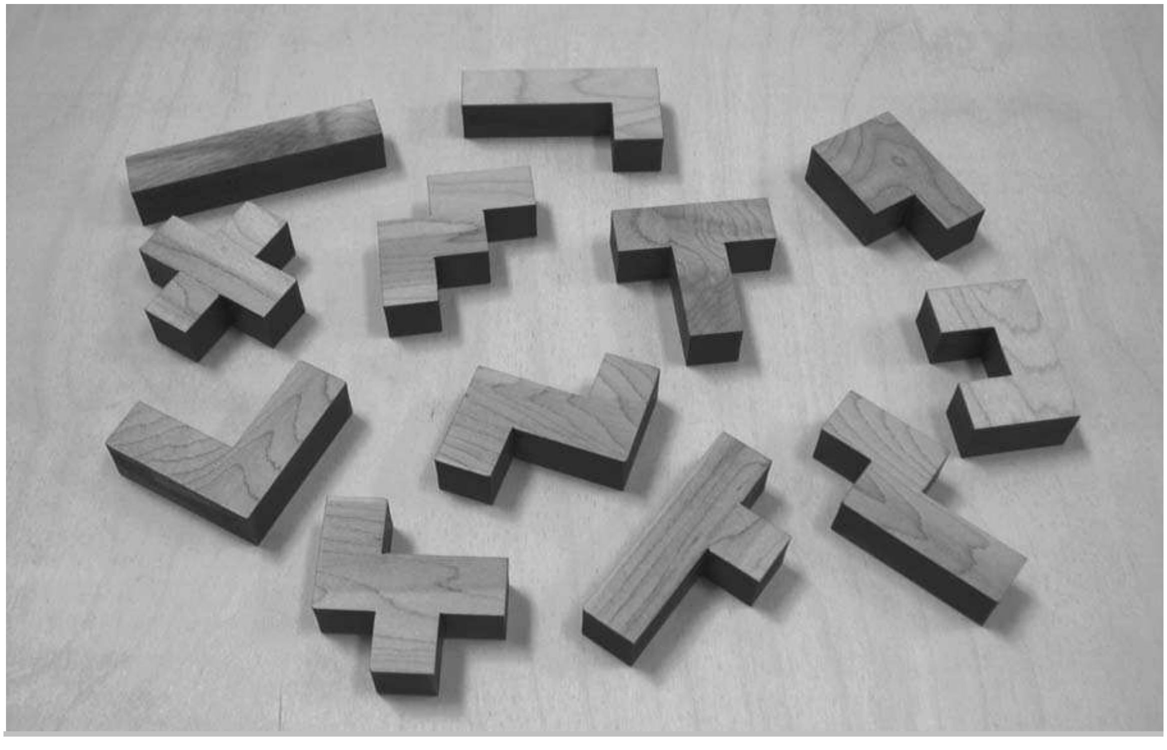

polyomino A complete set of 12 solid pentominos, also known as planar pentacubes. *Kadon Enterprises, Inc., www.gamepuzzles.com* one of William's sons, smashes a chessboard over the head of his brother, Prince Robert, and then adding his own mathematical appendix. By a curious quirk, Dudeney reveals, the board breaks into 13 pieces, 12 of which are the different possible ways in which five squares can be arranged and one square 2 × 2 piece. The puzzle is to reconstitute the board from these fragments. This is the earliest example of a mathematical recreation that involves pentominoes, a name that wasn't coined until 1953 when Solomon **Golomb** first used it in a talk to the Harvard Math Club. Golomb not only invented the nomenclature for polyominoes but did much of the pioneering research on them; his work was brought to a wide audience and popularized by Martin **Gardner** in his *Scientific American* column, beginning in 1957.[114, 136, 137]

Golomb was particularly interested in the pentomino, for a reason that becomes clear (see table, "Types of Polyomino"). The familiar domino comes in just one configuration as, obviously, does the trivial monomino. A triomino can have two different shapes–three squares in a line or in an L shape. The tetromino (or tetramino) has five distinct arrangements and is used in the popular video game **Tetris**. Thus for $n \leq 4$ the number of distinct pieces is ≤ 5, which restricts the variety of combinations possible. On the other hand, for $n > 5$, the number of different pieces is large, making analysis of problems difficult and games based on such polyominoes difficult and unwieldy. However, pentominoes, for which $n = 5$, come in 12 unique configurations, which is just about an ideal balance between tractability and combinatorial richness, hence the high level of interest in this particular type of polyomino. The problem of tiling an 8 × 8 chessboard with a square hole in the center using pentominoes was first solved in 1935 and found, by computer in 1958, to have exactly 65 solutions. Another standard pentomino puzzle is to arrange the set of 12 possible shapes into rectangles without holes: 3 × 20, 4 × 15, 5 × 12, and 6 × 10. Pentominoes are prominently featured in a subplot of the novel *Imperial Earth* by Arthur C. Clarke.

Hexominoes and heptominoes come in 35 and 108 unique arrangements, respectively. One of the 108 heptomino configurations, however, has a hole–a region that is not tiled with squares but that is unconnected to the exterior of the polyomino–and may or may not be counted as a valid piece depending on the rules of a particular game. All polyominoes made of seven or more squares may contain holes. There is no known algorithm or formula for calculating how many distinct polyominoes of each order there are.

Related to polyominoes are **polyiamonds** (formed from equilateral triangles) and polyhexes (formed from regular hexagons). The three-dimensional analog of polyominoes uses cubes instead of squares; an example is the **Soma cube**.

Types of Polyomino

Name	Number of Unit Squares	Number of Distinct Configurations
Monomino	1	1
Domino	2	1
Triomino	3	2
Tetromino	4	5
Pentomino	5	12
Hexomino	6	35
Heptomino	7	108

polytope

A higher-dimensional analogue of a **polygon** or **polyhedron**. The number of possible regular polytopes depends on the number of dimensions. In two dimensions there

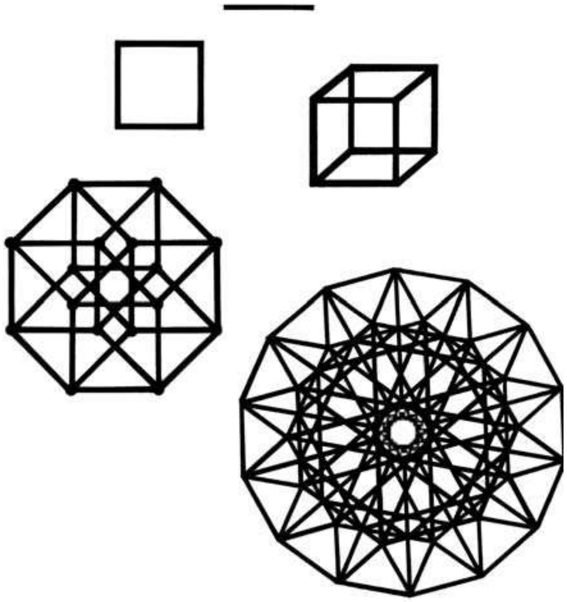

polytope In descending order: a straight line, a square, a cube, a tesseract (four-dimensional cube), and a regular, multidimensional polytope.

are infinitely many possible regular polygons; in three dimensions there are five possible regular polyhedra; in four dimensions there are six possible regular polytopes; and in each number of dimensions higher than four, there are just three possible regular polytopes, analogous to the three-dimensional tetrahedron, cube, and octahedron. A four-dimensional polytope is also sometimes called a *polyhedroid* or a *polychoron.* Just as a polygon has vertices and edges, and a polyhedron has vertices, edges, and faces, a four-dimensional polytope has vertices, edges, faces, and cells, where a cell is a three-dimensional figure.[75] See also **Boole (Stott), Alicia**.

Poncelet, Jean Victor (1788–1867)

A French mathematician who substantially advanced **projective geometry**. With Brianchon, he proved *Feuerbach's theorem* on the **nine-point circle** in 1820–1821, and also suggested the theorem proved by Jakob **Steiner** and now called the *Poncelot-Steiner theorem* that Euclidean constructions (see **constructible**) can be done with a straightedge alone. As a soldier in Napoleon's army, he was captured and imprisoned in Russia. While in prison from 1813–1814, he organized and wrote down his discoveries, and the result was published as *Traité des Propriétés Projectives des Figures* (Treatise on the properties of the projections of shapes, 1822). To serve as an introduction to this work, he also wrote *Applications D'analyse et de Géométrie* (Applications of analysis and geometry, 2 vols., 1862–1864).

Poncelet's theorem

Given an **ellipse**, and a smaller ellipse entirely inside it, start at a point on the outer ellipse, and, moving clockwise, follow a line that is **tangent** to the inner ellipse until you hit the outer ellipse again. Repeat this over and over again. It may be that this path will never hit the same points on the outer ellipse twice. However, if it *does* close up in a certain number of steps, then something amazing is true: *all* such paths, starting at *any point* on the outer ellipse, close up in the same number of steps. This fact is Poncelet's theorem, also known as *Poncelet's closure theorem,* and is named after the Jean **Poncelet**.

Pony Puzzle

One of Sam **Loyd**'s best known and commercially successful puzzles. It consists of just six pieces of the silhouette of a horse that have to be assembled in the most sensible way. Loyd got the idea for the puzzle from the governor of Philadelphia, Andrew Curtin, when the two were on a steamer returning from Europe and talking about the White Horse of Uffington–a famous chalk figure carved on a hillside in Berkshire, England.

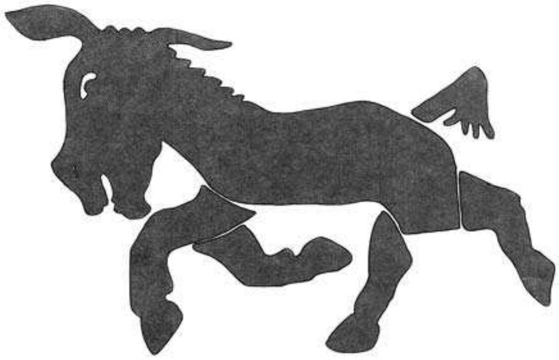

Pony Puzzle One of Sam Loyd's best known brain-teasers.

postage stamp problems

Mathematical puzzles that involve postage stamps have been around almost as long as postage stamps themselves, the first of which, the Penny Black, was issued by Great Britain on May 6, 1840. Some such puzzles ask what postage amounts can or can't be made with stamps of certain values. Others are based on the ways that a block of stamps can be folded or torn along the perforations.

PUZZLE

One of this type appears as problem no. 285 in Henry **Dudeney**'s *Amusements in Mathematics* (1917).[88] It starts by saying you have just bought 12 stamps in a rectangular block of three rows with four stamps in each. (In Dudeney's diagram, they are labeled 1, 2, 3, 4 across the top row, and so on.) He goes on: "[A] friend asks you to oblige him with four stamps, all joined together–no stamp hanging on by a mere corner. In how many different ways is it possible for you to tear off those four stamps? You see, you can give him 1, 2, 3, 4, or 2, 3, 6, 7, . . . , and so on. Can you count the number of different ways in which those four stamps might be delivered?" This can be thought of as a problem involving tetrominoes, which are a type of **polyomino**.

Solutions begin on page 369.

The postage stamp problem, also known as the *Frobenius problem,* is a long-standing challenge in **number theory** and in **computer science**. Suppose a country issues n different denominations of stamps but allows no more than m stamps to be put on a single letter. The postage stamp problem is to write and implement an algorithm (a stepwise set of rules) that, for any given values of m and *n,* computes the greatest consecutive range of postage values, from one on up, and all possible sets of denominations that realize that range. For example, for $n = 4$ and $m = 5$, the stamps with values (1, 4, 12, 21) allow the postage values 1 through 71. If the values of the stamps are constant and not part of the input, algorithms can be, and have been, devised that give a short-cut solution. However, in the general case where the number of stamp values *is* part of the input, the postage stamp problem has been shown to be an **NP-hard problem**, and thus not susceptible to an efficient algorithmic approach.

potato paradox

Fred brings home 100 pounds of potatoes, which (being purely mathematical potatoes) consist of 99 percent water. He then leaves them outside overnight so that they consist of 98 percent water. What is their new weight? The surprising answer is 50 pounds!

potential theory

The study of *harmonic functions.* These functions satisfy **Laplace**'s equation, a certain type of **partial differential equation** that commonly arises in physics in problems to do with gravity and electromagnetism.

power

A word that is almost never used in its correct, original sense any more. Strictly speaking, if we write $8 = 2^3$, then 2 is the **base**, 3 is the **exponent**, and 8 is the power. But almost everyone, including most mathematicians, would say that 3 is the power, and that "power" and "exponent" mean the same thing. The misuse has probably come about from a misunderstanding of statements such "eight is the third power of two."

power law

A type of mathematical pattern in which the frequency of an occurrence of a given size is inversely **proportional** to some **power** (or exponent) of its size. For example, in the case of avalanches or earthquakes, large ones are fairly rare, smaller ones are much more frequent, and in between are cascades of different sizes and frequencies. Power laws define the distribution of catastrophic events in self-organized critical systems (see **self-organization**).

power series

An infinite sum of the form

$$a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots,$$

where a is a number of any type and x is the variable. Power series are commonly used to define **function**s. For example, the **sine** function can be written

$$\sin x = x - x^3/3! + x^5/5! - x^7/7! + \ldots,$$

where "!" stands for **factorial**. Although the series has infinitely many terms, these get small so quickly that only the first few terms make much of a contribution.

power set

The **set** of all **subsets** of a given set, including the **empty set** and the original set. For example, if the original set is $\{a, b, c\}$ then the power set is $\{\{\varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$.

power tower

A way of representing very large numbers in terms of stacks of **exponent**s. For example:

$$10^{3^{3^{3^3}}} = 10^{3^{3^{27}}} = 10^{3^{2541865828329}} = \text{enormous}$$

In general, and especially if the number at the top of the tower is fairly large, adding another exponent to the bottom of a tower will make the value of the tower much larger than will increasing the size of the bottom exponent. This leads to the (somewhat counterintuitive) result that to know which of two towers is the larger, you can look at how many exponents are in the tower and know right away which is larger. For example:

$$1.1^{1.1^{1.1^{1,000}}} \text{ is much larger than } 1{,}000^{1{,}000^{1{,}000}}$$

In the case of an infinite power tower of the form

$$x^{x^{x^{\cdots}}},$$

the maximum value that x can take and still cause the tower to converge to a finite value is $e^{(1/e)} = 1.444667\ldots.$ The minimum value of x that will produce convergence is $1/e^e = 0.065988\ldots.$

powerful number

Also known as a *squarefull number,* a positive whole number n such that for every **prime number** p dividing n, p^2 also divides n. Every powerful number can be written as a^2b^3, where a and b are positive integers. The first few powerful numbers are 1, 4, 8, 9, 16, 25, 27, 32, 36, 49, 64, 72, 81, 100, and 108. Pairs of consecutive powerful numbers exist, such as (8,9), (288,289), and (675,676). However, no three consecutive powerful numbers are known and, in 1978, Paul **Erdös** conjectured that none exist.

practical number

A number n such that every positive integer less than n is either a divisor or a sum of distinct divisors of n. The first few practical numbers are 1, 2, 4, 6, 8, 12, 16, 18, 20, and 24. All **perfect number**s are practical.

prime number

[U]pon looking at prime numbers one has the feeling of being in the presence of one of the inexplicable secrets of creation.

–Don Zagier

An **integer** greater than 1 that is divisible only by 1 and itself. Prime numbers have fascinated mathematicians for centuries, in large part because of how they are distributed. At first sight, their occurrence looks random; yet, on closer inspection, they reveal a subtle order or pattern, that seems to hold deep truths about the nature of mathematics and of the world in which we live. The German-born American mathematician Don Zagier (1951–), in his inaugural lecture at Bonn University, put it this way:

> There are two facts about the distribution of prime numbers which I hope to convince you. . . . The first is that despite their simple definition and role as the building blocks of the natural numbers, the prime numbers . . . grow like weeds among the natural numbers, seeming to obey no other law than that of chance, and nobody can predict where the next one will sprout. The second fact is even more astonishing, for it states just the opposite: that the prime numbers exhibit stunning regularity, that there are laws governing their behavior, and that they obey these laws with almost military precision.

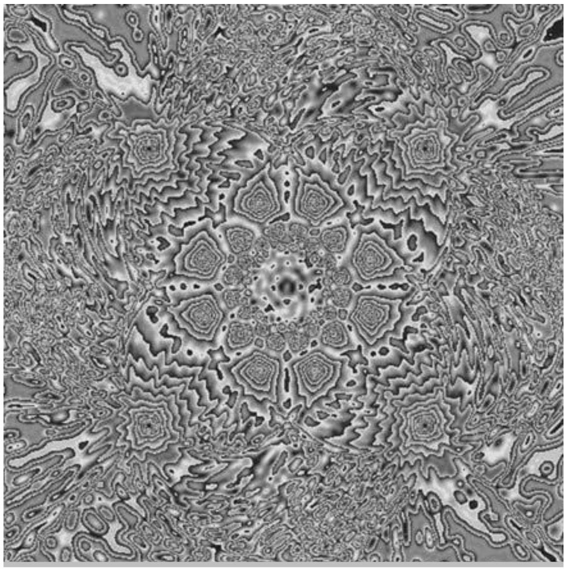

prime number A remarkable pattern generated by the distribution of the first few billion primes. *Jean-Francois Colonna/Ecole Polytechnique*

The prime numbers are 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, ... The *fundamental theorem of arithmetic* declares that the primes are the building blocks of the positive integers: every positive integer is a product of prime numbers in one and only one way, except for the order of the factors. This is the key to their importance: the prime factors of an integer determine its properties. The ancient Greeks proved (c. 300 B.C.) that there are infinitely many primes and that they are irregularly spaced; in fact, there can be arbitrarily large gaps between successive primes. On the other hand, in the nineteenth century it was shown that the number of primes less than or equal to *n* approaches *n*/log *n*, as *n* gets very large (a result known as the *prime number theorem*), so that a rough estimate for the *n*th prime is *n* log *n*. In his *Disquisitiones Arithmeticae* (1801), Carl **Gauss** wrote: "The problem of distinguishing prime numbers from composite numbers and of resolving the latter into their prime factors is known to be one of the most important and useful in arithmetic. It has engaged the industry and wisdom of ancient and modern geometers to such an extent that it would be superfluous to discuss the problem at length. . . . Further, the dignity of the science itself seems to require that every possible means be explored for the solution of a problem so elegant and so celebrated." The earliest known primality test is the **sieve of Eratosthenes**, which dates from around 240 B.C. However, high-speed computers and fast algorithms are needed to identify large primes. New record-breaking primes tend to be of the variety known as **Mersenne prime**s, since these are the easiest to find. About 6,000 prime numbers are known of which the largest is $2^{20996011} - 1$.

Much remains unknown about the primes. As Martin **Gardner** said:[113] "No branch of **number theory** is more saturated with mystery. . . . Some problems concerning primes are so simple that a child can understand them and yet so deep and far from solved that many mathematicians now suspect they have no solution. Perhaps they are 'undecideable.' Perhaps number theory, like quantum mechanics, has its own uncertainty principle that makes it necessary, in certain areas, to abandon exactness for probabilistic formulations." One of the greatest unsolved problems in mathematics is the **Riemann hypothesis** concerning the distribution of prime numbers. See also **Goldbach's conjecture**, **twin primes**, **Ulam spiral**, and **Ishango bone**.

primitive root

A primitive root for a **prime number** *p* is one whose powers generate all the nonzero integers **modulo** (or mod) *p*. For example, 3 is a primitive root modulo 7 since $3 = 3^1$, $2 = 3^2 \bmod 7$, $6 = 3^3 \bmod 7$, $4 = 3^4 \bmod 7$, $5 = 3^5 \bmod 7$, $1 = 3^6 \bmod 7$.

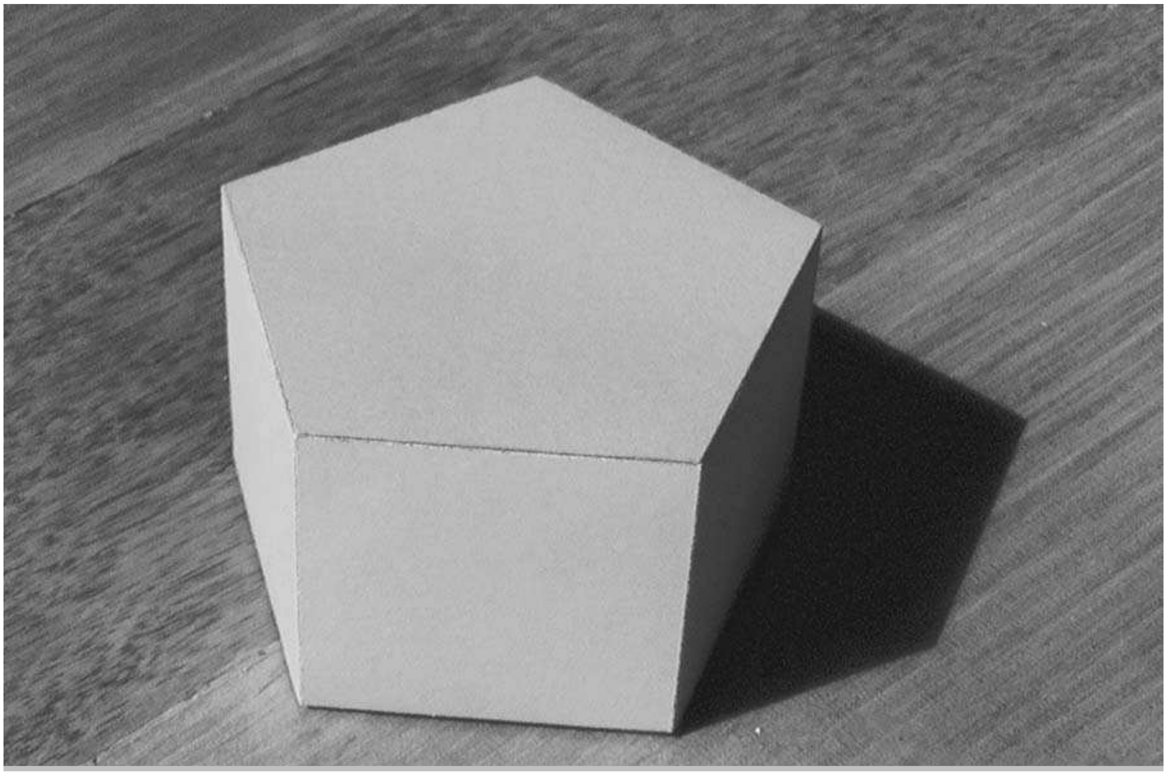

prism A pentagonal prism. *Robert Webb, www.software3d.com; created using Webb's Stella program*

primitive root of unity

The **complex number** z such that $z^n = 1$ but z^k is not equal to 1 for any positive integer k less than n.

primorial

Also known as a *prime factorial,* the product of all **prime numbers** that are less than or equal to a given prime p; it is denoted $p\#$. For example $3\# = 2 \times 3 = 6$, $5\# = 2 \times 3 \times 5 = 30$, and $13\# = 2 \times 3 \times 5 \times 7 \times 11 \times 13 = 30{,}030$.

Prince Rupert's problem

The problem of pushing a **cube** through a hole in another cube of equal or less size; it is named after Prince Rupert (1619–1682), a nephew of England's King Charles I, who won a wager that a hole could be made in one of two equal cubes large enough for the other cube to slide through. The mathematics of cubes passing through cubes was considered by John **Wallis**. Later, in 1816, a solution was published posthumously by the Dutch mathematician Pieter Nieuwland (1764–1794) to the question: What is the largest cube that can be passed through a unit cube (a cube of which each side is one unit long)? Nieuwland answered this by finding the largest square that fits inside a unit cube. When viewed from directly above one apex, a unit cube has the outline of a regular hexagon of side $\sqrt{3}/\sqrt{2}$. The largest square that will go into a cube has a face that can be inscribed within this hexagon; the length of its edge is $\sqrt{6} - \sqrt{2} = 1.03527618$.

prism

A **semi-regular polyhedron** constructed from two **congruent** *n-sided* **polygons** and *n* **parallelogram**s. The word comes from the Greek *prizma,* which relates to cutting or sawing. A *prismoid* resembles a prism but has bases that are **similar** rather than congruent, and sides that are **trapezoids** rather than parallelograms. An example of a prismoid is the **frustum** of a **pyramid**. A *prismatoid* is a polyhedron with all its vertices lying in two parallel planes.

prisoner's dilemma

A problem in **game theory** first described by the Canadian-born Princeton mathematician Albert Tucker (1905–1995) in 1950, while addressing an audience of psychologists at Stanford University, where he was a visiting professor. It runs along these lines: Al and Bob have been arrested for holding up the Anyapolis State Bank and have been put in separate cells. Each cares a lot more about his personal freedom than he does about his accomplice's welfare. A clever prosecutor makes the following offer to each. "You may choose to confess or remain silent. If you confess and your accomplice remains silent, I'll drop all charges against you and use your testimony to ensure that your accomplice does serious time. Likewise, if your accomplice confesses while you remain silent, he'll go free while you do the time. If you both confess I get two convictions, but I'll see to it that you both get early parole. If you both remain silent, I'll have to settle for token sentences on firearms possession charges. If you wish to confess, you must leave a note with the jailer before I come back tomorrow morning." The dilemma faced by the prisoners is that, whatever the other does, each is better off confessing than remaining silent. But the outcome when both confess is worse for each than the outcome if both stay silent!

Tucker's paradox was based on puzzles with a similar structure that had been devised in 1950 by Merrill Flood and Melvin Dresher as part of the Rand Corporation's investigations into game theory (which Rand pursued because of possible applications to global nuclear strategy). Flood and Dresher hadn't published much about their work, but the prisoner's dilemma attracted an enormous amount of attention in subjects as diverse philosophy, biology, sociology, political science, and economics, as well as game theory itself. A common view is that the puzzle illustrates a conflict between individual and group rationality. A group whose members pursue rational self-interest may all end up worse off than a group whose members act contrary to rational self-interest. More generally, if the payoffs aren't assumed to represent self-interest, a group whose members rationally pursue any goals may all meet less success than if they hadn't rationally pursued their goals individually.

probability

A measure of how likely it is that some event will occur, given as a number between 0 (impossible) and 1 (certain). Usually, probability is expressed as a ratio: the number of experimental results that would produce the event divided by the number of experimental results considered possible. For example, the probability of drawing the five of hearts from an ordinary deck of cards is one in fifty-two (1:52).

probability theory

The branch of **mathematics** that deals with the possible outcomes of events and their relative likelihoods. While mathematicians agree on how to calculate the probability of certain events and how to use those calculations in certain ways, there's plenty of disagreement as to what the numbers actually mean. Probability divides into two main concepts: *aleatory probability,* which represents the likelihood of future events whose occurrence is governed by some random physical phenomenon like tossing dice or spinning a wheel; and *epistemic probability,* which represents our uncertainty of belief about past events that either did or didn't occur, or uncertainty about the causes of future events. An example of the latter is when we say that it's "probable" that a certain suspect committed a crime based on the available evidence. It is an open question whether aleatory probability is reducible to epistemic probability based on our inability to precisely predict every force that might affect the roll of a die, or whether such uncertainties exist in the nature of reality itself, particularly at the level of **quantum mechanics**. One of the earliest mathematical studies on probability was written by Girolamo **Cardano**. Among other important contributions to the development of the subject were those by Blaise **Pascal**, Pierre de **Fermat**, Jakob Bernouilli (see **Bernouilli family**), Joseph **Lagrange**, Pierre **Laplace**, Carl **Gauss**, Siméon **Poisson**, Abraham **de Moivre**, Pafnuty Chebyshev, Andrei Markov (see **Markov chain**), and Andrei **Kolmogorov**.

PUZZLES

The following are two problems in probability. Both are easier to solve if the various possible outcomes are written down in the form of a table.

1. You meet a stranger on the street, and ask how many children he has. He truthfully says two. You ask "Is the older one a girl?" He truthfully says yes. What is the probability that both children are girls? What would the probability be if your second question had been "Is at least one of them a girl?", with the other conditions unchanged?
2. You are in a game of Russian roulette, but this time the gun (a six-shooter revolver) has three bullets in a row in three of the chambers. The barrel is spun only once. Each player then points the gun at his head and pulls the trigger. If he is still alive, the gun is passed to the other player who then points it at his own head and pulls the trigger. The game stops when one player dies. Would you stand a better chance of surviving if you shoot first or second, or does it make any difference?

Solutions begin on page 369.

See also **Buffon's needle**, **birthday problem**, **Monty Hall problem**, and **St. Petersburg paradox**.

Proclus Diadochus (c. A.D. 410–485)

The last major Greek philosopher; his *Commentary on Euclid* is the main original source we have on the early history of Greek geometry. He also wrote *Hypotyposis,* which gives a detailed account of the Earth-centered astronomical theories of Hipparchus and Ptolemy.

product

The result of one or more multiplications.

projectile

An object, such as a baseball, spear, or cannonball, that is thrown, fired, or otherwise propelled but that can't propel itself. For centuries, philosophers and mathematicians debated what path a projectile follows under gravity. Galilei Galileo was the first to establish that (in the absence of air resistance) this path is a **parabola**.

projective geometry

The branch of **geometry** that deals with properties of geometric figures that remain unchanged under projection. A mathematical theory of perspective grew out of the studies of Renaissance architects and painters who asked themselves how to best represent a three-dimensional object on a two-dimensional surface. The Greeks had done some early work on perspective, and the great geometer **Pappus of Alexandria** is credited with the first theorem in projective geometry. However, the subject reached mathematical maturity through the efforts first of Girard **Desargues**, and then, much later through the work of Jean **Poncelet** and by Karl von Staudt (1798–1867).

The basic elements of projective geometry are points, lines, and planes. These elements retain their characteristics under projection; for example, the projection of a line is another line, and the point of intersection of two lines is projected into another point that is the intersection of the projections of the two original lines. However, lengths and ratios of lengths are not invariant under projection, nor are angles or the shapes of figures. The concept of parallelism doesn't appear at all in projective geometry; any pair of distinct lines intersects in a point, and if these lines are parallel in the sense of **Euclidean geometry**, then their point of intersection is at infinity. The plane that includes the ideal line, or line at infinity, consisting of all such ideal points, is called the **projective plane**. Two properties that are invariant under projection are the order of three or more points on a line and the harmonic relationship, or cross-ratio, among four points, A, B, C, D, that is, $AC/BC : AD/BD$. The most remarkable concept in projective geometry is that of duality. In the plane, the terms *point* and *line* are dual and can be interchanged in any valid statement to yield another valid statement; in space, the terms *plane, line,* and *point* are interchanged with *point, line,* and *plane,* respectively, to yield dual statements. Entire theorems also occur in dual pairs, so that one can be instantly transformed into the other. For example, *Pascal's theorem* (given a hexagon inscribed in a **conic section**, the three pairs of the continuations of opposite sides meet on a straight line) is the dual of *Brianchon's theorem* (given a hexagon circumscribed on a conic section, the lines joining opposite diagonals meet in a single point). In fact, *all* the propositions in projective geometry occur in dual pairs.

projective plane

The surface you would get if you glued the edge of a disk to the edge of a **Möbius band**. This sounds easy to do, since a disk and a Möbius band each have one edge. But the process becomes hopelessly tangled and is, in fact, impossible. The projective plane needs a **fourth dimension**, in addition to the three we live in (up-down, left-right, and back-forth), to be fully realized. The idea of the projective plane arose from the study of perspective by mathematicians and painters in the Renaissance. In trying to represent parallel lines in space on the two-dimensional surface of a painting, it was found useful to introduce the notion of a *line at infinity* on which parallel lines met. The study of the geometry that adds this extra line of ideal points to the ordinary familiar plane came to be known as **projective geometry**, because of its use in studying projections of figures onto different lines. This idea was even more important in three dimensions, since projections are used for representing three-dimensional figures on planes. An interesting property of the projective plane is that any "straight" line on it, followed far enough, comes back to the starting point. (The old arcade version of the game of Asteroids was played on a virtual projective plane: the screen was a disk, and when an asteroid went off one edge of the screen it emerged on the opposite side.) The projective plane is also *nonorientable,* as a result of which any two-dimensional object pushed along a path back to its starting point would be reversed.

prolate

(1) Rounded like an **egg**. (2) Having a polar diameter greater than the equatorial diameter. See also **spheroid**.

pronic number

Also known as a *rectangular* or *oblong number,* a number that is the product of two consecutive integers: 2 (1 × 2), 6 (2 × 3), 12 (3 × 4), 20 (4 × 5), ... The pronic numbers are twice the **triangular number**s, and represent the lengths that produce the musical intervals: octave (1:2), fifth (2:3), fourth (3:4), major third (4:5) . . . . Pronic seems to be a misspelling of *promic,* from the Greek *promekes,* for "rectangular" or "oblong"; however, the "n" form goes back at least as far as Leonhard **Euler** who used it in series one, volume fifteen of his *Opera.*

proof

A sequence of statements in which each subsequent statement is derivable from one of the previous statements or from an **axiom** of a **formal system**. The final statement of a proof is usually the **theorem** that one has set out to prove.

proper divisor

See **aliquot part**.

proportional

A variable a is said to be *directly proportional* to b if a/b is a constant. The relationship is written $a \propto b$, which implies that $a = kb$, where k is a constant. If a is *inversely proportional* to b, this is written $a \propto 1/b$.

pseudoprime

A number that passes the test of **Fermat's little theorem** (FLT) for **prime numbers** but actually isn't a prime. FLT says that if p is prime and a is **coprime** to p, then $a^{p-1} - 1$ is divisible by p. If a number x is not prime, a is coprime to x, and x divides $a^{x-1} - 1$, then x is called a *pseudoprime to base a.* A number x that is a pseudoprime for all values of a that are coprime to x is called a **Carmichael number**. The smallest pseudoprime in base 2 is 341. This isn't prime because $341 = 11 \times 31$; however, it satisfies FLT: $2^{340} - 1$ is divisible by 341.

pseudosphere

A saddle-shaped surface that is produced by rotating a **tractrix** about its **asymptote**. The name *pseudosphere,* which means "false sphere," is misleading because it suggests something that is spherelike; however, a pseudosphere is almost exactly the opposite of a sphere. Whereas a sphere has a constant positive **curvature** (equal to $+l/r$, where r is the radius) at every point on its surface, a pseudosphere has a constant negative curvature (equal to $-l/r$) everywhere. As a result, a sphere has a closed surface and a finite area, while a pseudosphere has an open surface and an infinite area. In fact, although both the two-dimensional plane and a pseudosphere are infinite, the pseudosphere manages to have more room! One way to think of this is that a pseudosphere is more intensely infinite than the plane. Another result of the pseudosphere's negative curvature is that the angles of a triangle drawn on its surface add up to less than 180°. The geometry on the surfaces of both the sphere and the pseudosphere is a two-dimensional **non-Euclidean geometry**–spherical (or elliptical) geometry in the case of the sphere and **hyperbolic geometry** in the case of the pseudosphere. Astronomers currently suspect that the universe we live in may have a hyperbolic geometry and thus have properties analogous to those of a pseudosphere.

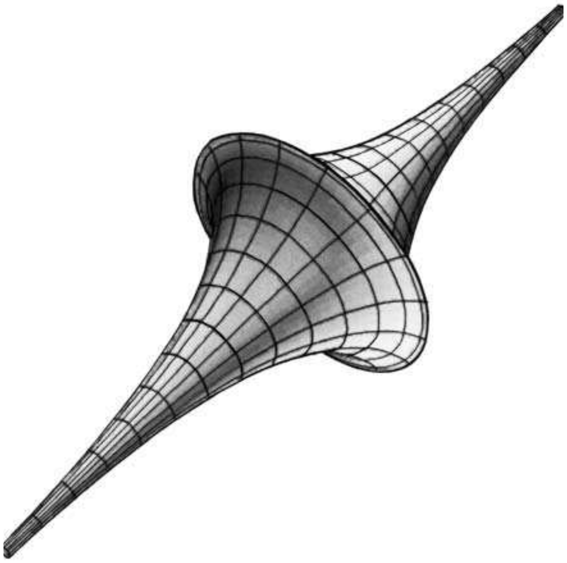

pseudosphere

Ptolemy's theorem

The sum of the products of the two pairs of opposite sides of a convex cyclic **quadrilateral** (see **cyclic polygon**) is equal to the product of the lengths of the diagonals. The theorem is named after the mathematician, astronomer, and geographer Ptolemy of Alexandria.

pure mathematics

Mathematics for the sake of its internal beauty or logical strength. Compare with **applied mathematics**.

pursuit curve

The path an object takes when chasing after another object in the most effective way. Pursuit curves can arise in a variety of situations, for example when a lion is chasing after a gazelle or a heat-seeking missile is homing in on a moving target. Suppose that four ants are at the corners of a square. They start to crawl clockwise at a constant rate, each moving toward its neighbor. At any instant, they mark the corners of a square. As the ants get closer to

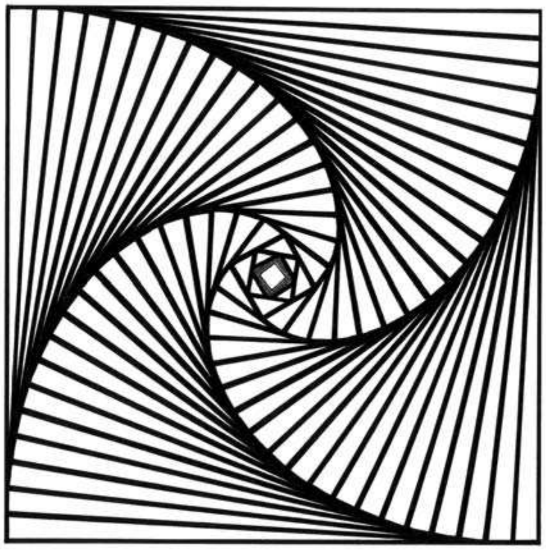

pursuit curve A set of superimposed snapshots shows the lines of sight at regular intervals of four bugs chasing one another, all moving at the same speed after starting at the corners of a square. *John Sharp*

the original square's center, the new square they define rotates and diminishes in size. In reaching the center, each ant travels on a **logarithmic spiral** with a length equal to the side of the original square. Superimposed snapshots of the ants' progress give rise to an intriguing pattern.

puzzle jug

A practical joke in the form of a drinking vessel that, unless the trick is known, spills all over the user when he tries to take a sip. As the jug is tilted, liquid pours out from the pierced decoration surrounding the pot. The puzzle is to learn the secret of how to successfully drink without getting wet. Puzzle jugs are the oldest known **mechanical puzzle**s. Several Phoenician examples are in New York's Metropolitan Museum of Art, and in ninth-century Turkey they were especially popular. Puzzle jugs have been manufactured in Germany, Holland, France, and several other European countries since the end of the thirteenth century. The Exeter puzzle jug, on display in the Royal Albert Museum in Exeter, Devon, was probably made in the Saintonge region of western France, around 1300, and is one of the finest examples of medieval pottery imported to England. Puzzle jugs came back into vogue in the eighteenth and nineteenth centuries and are still manufactured by ceramicists in France, Germany, and Britain. And the secret? That is really quite open, and can be handled by anyone prepared to get to the bottom of it.

puzzle rings

A **mechanical puzzle** consisting of a group of interlocking rings that fit together to form an intricate design; when apart, the rings remain interconnected and pose a challenging puzzle. The puzzle itself has ancient origins in Egypt, though many stories have been told of its use throughout the Middle East. According to one, a person would give the ring to their lover telling them of the magical qualities of the ring. If the ring was ever worn while being amorous to another, it would tell the giver of the ring about the unfaithfulness of the ring wearer. The wearer, wishing to avoid the watchfulness of the ring, might remove it from their finger so that it wouldn't witness an unfaithful act. But once removed, the ring would fall into its separate pieces and so the giver would know their lover had been cheating.

pyramid

A **polyhedron** whose base is a **polygon** and whose other faces are triangles that meet at a common **vertex**, sometimes called the *apex*. A right pyramid has its apex directly above the center of its base. A square-based pyramid, like the pyramids in Egypt, has a square base and four triangular sides. A triangular pyramid, or **tetrahedron**, is made from four triangular sides; if regular, it is one of the **Platonic solid**s. The volume of a pyramid is $\frac{1}{3}A_bh$, where A_b is the area of the base and h is the perpendicular height of the apex above the base; the surface area is ps, where p is the base perimeter and s is the slant height.

Pyramidology, which (in its original form) claims links between biblical prophesy and the construction of the Egyptian pyramids, stemmed from the publication in 1859 of *The Great Pyramid: Why Was It Built? And Who Built It?* by

pyramid The Great Pyramid of Cheops at Giza.

John Taylor, a British amateur mathematician and astronomer who was editor of the *London Observer.* Taylor became convinced that the Great Pyramid of Cheops, in its architectural proportions, embodied various remarkable and deeply meaningful geometric and mathematical properties. Chief among these, Taylor noted, was that the ratio of the perimeter of the pyramid's base to twice its height closely approximates the universal constant **pi**. The said ratio of the Great Pyramid gives a better value for pi than any found in written Egyptian records. The **Rhind papyrus** contains several problems that involve a multistep method for finding the area of a circle from its diameter. This method implies a value of π equal to 256/81, or 3.1605, which is less than 1% larger than the true value of 3.14159 . . . . The values Taylor used for the Great Pyramid's base and height gave him a value of π correct to two decimal places. Further analysis of the pyramid's dimensions and various manipulations of numbers led him to conclude that the builders of the edifice had used a unit that he called "pyramid inch" (equal to about 1.01 times a standard inch). Twenty-five pyramid inches made a "pyramid cubit," and 10 million pyramid cubits, Taylor pointed out, approximate the length of Earth's polar radius. These and a series of similar calculations provided what Taylor considered to be adequate evidence that the Great Pyramid had been built as a model of Earth. Taylor's fantastic claims may never have become popular had not Charles Piazzi Smyth, the Astronomer Royal of Scotland taken up the cause of pyramidology. He popularized it in Britain, the rest of Europe, and the United States, through a number of books including *Our Inheritance in the Great Pyramid* (1864) and *Life and Work at the Great Pyramid* (1867). (In the churchyard in Sharow, North Yorkshire, Smyth has a pyramidal grave marker.) Others who espoused pyramid numerology have included Helena Blavatsky, the theosophist; Charles Taze Russell, founder of the Watchtower Bible and Tract Society; and Edgar Cayce, the American psychic; not to mention a variety of more recent pseudohistorians who have linked the Great Pyramid to everything from lost civilizations to UFOs. Yet Taylor himself was well aware of other, less dramatic explanations for the pyramid's dimensions, including the possibility that the pyramid had been constructed so that the area of one of its faces would equal the square of its height. The mathematical sophistication required to achieve this is not great and would have resulted in a ratio of the perimeter of the base to twice the height of 3.145–about as close as the approximation of pi used by Smyth in his studies. So, the ratio could have occurred as a completely coincidental byproduct of a design that had nothing to do with the ratio of the circumference of a circle to its diameter. The pyramid is a rich source of the kind of data that Taylor and Smyth worked with, and it would be surprising if they'd been unable to come up with some interesting number combinations given the extent of their efforts.

pyramidal number

The number of dots that may be arranged in a **pyramid** with a regular **polygon** as a base.

Pythagoras of Samos (c. 580–500 B.C.)

A Greek philosopher and mathematician, a native of the Aegean island of Samos, and the founder of a secretive pseudoreligious community in Croton, southern Italy. Pythagoras left no writings and virtually nothing is known about him as an individual, so it is almost impossible to disentangle the beliefs and discoveries of the "Pythagoreans" from those of their leader. To the Pythagoreans, "everything is number" and every number was supposed to be a quantity that could be expressed as the ratio of two integers. We now call such a number a **rational number**. The Pythagoreans used music as an example of the perfection and harmony of numbers that can be expressed as ratios. They showed that pitch could be represented as a simple ratio that came from the length of equally tight strings that could be plucked. Perhaps the most famous of the Pythagoreans' mathematical results is **Pythagoras's theorem**. Then the sky fell in on the Pythagoreans' worldview. Using their very own theorems they showed that not all numbers are rational. Their discovery that the **square root of 2** (the length of the hypotenuse of a triangle with sides 1 and 1) can't be expressed as ratio of two whole numbers was to have been kept a closely guarded a secret, but was later revealed by one of the cult's members.

Pythagoras's lute

The **kite**-shaped figure that forms the enclosing shape for a progression of diminishing **penta**gons and pentagrams, linking the vertices together. The resulting diagram is replete with lines in the **golden ratio**.

Pythagoras's theorem

The square of the length of the **hypotenuse** of a right **triangle** is the sum of the squares of the lengths of the two sides. This is usually expressed as $a^2 + b^2 = c^2$. See also **Pythagorean triplet**.

Pythagorean square puzzle

A deceptively hard assembly puzzle in which a small square piece must be combined with four pieces forming a larger square to make an even larger square.

PUZZLE

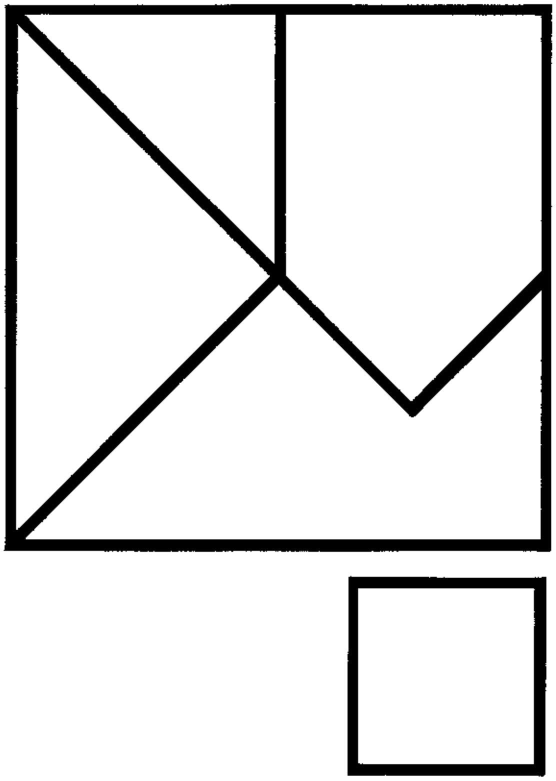

Rearrange the pieces in the Pythagorean square puzzle, including the small square, to make a single larger square.

Solutions begin on page 369.

Pythagorean triangle

A right triangle whose sides are integers. A *primitive Pythagorean triangle* is one whose sides are **coprime**.

Pythagorean triplet

Also called a *Pythagorean triple,* a set of three whole numbers that satisfies **Pythagoras's theorem**, that is, the squares of two of the numbers add up to the square of the third number. Examples include (3, 4, 5), (5, 12, 13), and (7, 24, 25). These are called *primitive triplets* because they have no common divisors. If the members of a primitive triplet are multiplied by the same integer, the result is a new (but not primitive) triplet. In any primitive Pythagorean triplet, one, and only one, of the three numbers must be even (but can't equal 2); the other two numbers are **coprime**. There are infinitely many such triplets, and they are easy to generate using a classic formula, known since ancient times. If the numbers in the triplet are a, b, and c, then: $a = n^2 - m^2$, $b = 2mn$, $c = m^2 + n^2$, where m and n are two integers and m is less than n. Because the **square root of 2** is irrational, there can't be any Pythagorean triplets (a, a, c). However, there are an infinite number of triplets $(a, a + 1, c)$, the first three of which (apart from the trivial (0, 1, 1)) are (3, 4, 5), (20, 21, 29), and (119, 120, 169).

There are also an infinite number of Pythagorean quartets (a, b, c, d) such that $a^2 + b^2 + c^2 = d^2$. This is simply the three-dimensional form of Pythagoras's theorem and can be interpreted as the fact that the point in three-dimensions with Cartesian coordinates (a, b, c) lies an integer distance d from the origin. A formula that generates Pythagorean quartets is: $a = m^2$, $b = 2mn$, $c = 2n^2$, $d = (m^2 + 2n^2) = a + c$. Also note that $b^2 = 2ac$. When $m = 1$ and $n = 1$, we get the quartet (1, 2, 2, 3)–the simplest example.

Although there are an infinite number of Pythagorean triplets, **Fermat's last theorem**, which is now know to be true, ensures that there are no triplets for higher powers. See also **Euler's conjecture** and **multigrade**.

Q

QED

Abbreviation for *quod erat demonstrandum* ("that which was to be shown"), used to denote the end of a proof.

quadrangle

A plane figure consisting of four points, each of which is joined to two other points by a line segment. A quadrangle may be **concave** or **convex** depending on whether the line segments do or don't intersect. A convex quadrangle is a **quadrilateral**. The word is from the Latin *quadrangulum* for "four-cornered" and is also used to describe a rectangular area surrounded on all four sides by buildings, or to such buildings themselves.

quadrant

Any one of the four portions of the plane into which the plane is divided by the Cartesian coordinate axes.

quadratic

An expression or an equation that contains the variable squared, but not raised to any higher power. For instance a *quadratic equation* in x contains x^2 but not x^3. Similarly a *quadratic expression,* or a *quadratic form,* contains its variable(s) squared but not raised to any higher power. If there is more than one variable (say, x and y), quadratic can mean that they are multiplied together in pairs (xy) but not in threes (such as x^2y). The graph of a quadratic equation is known as a *quadratic curve;* the curve of the general quadratic equation $y = ax^2 + bx + c$ is a **parabola**.

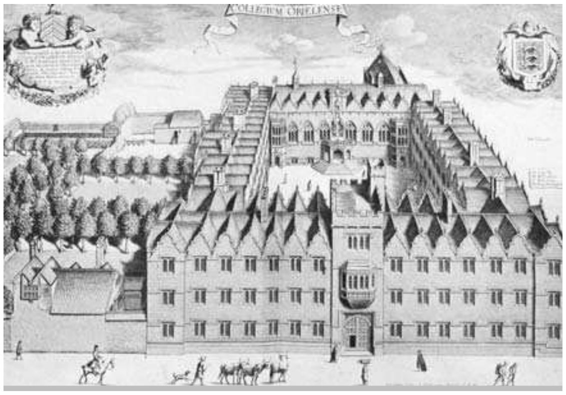

quadrangle The sixteenth-century quadrangle of Oxford University's Oriel College. *Oriel College, Oxford*

quadratrix of Hippias

The first curve in recorded history that was not part of a line or a circle, and the first curve known that is not **constructible** in the classical sense; in other words, it can't be drawn using a straightedge and a compass alone, but instead has to be plotted point by point. The quadratrix can be thought of as the intersection of two lines moving with constant velocity: the first line rotates (e.g., counterclockwise) while the second line moves along (say, in the direction of the positive y-axis). It has the Cartesian equation $y = x \cot(\pi x/2a)$. The quadratrix was discovered by **Hippias of Elis** in about 430 B.C. and was used by him in his work on **trisecting an angle** and **squaring the circle**. In fact, its name refers to its use in turning curvilinear space into a rectangular area.

quadrature

The determination of the area of a geometric figure.

quadric

A surface in three dimensions that is described by equations containing the squares of x, y, and z, but no higher powers of them. Examples of such surfaces include the **sphere**, **ellipsoid**, **cone**, and **cylinder**.

quadrifolium

See **rose curve**.

quadrilateral

A **polygon** that has four sides and four vertices (corners). Quadrilaterals, and polygons in general, may be **convex** or **concave**. A convex quadrilateral may be further classified as a **trapezoid** or a British **trapezium** (one pair of opposite sides are parallel), a trapezium (no sides parallel); an isosceles trapezoid (United States) or an isosceles trapezium (United Kingdom) (two of the opposite sides parallel, the two other sides equal, and the two ends of each parallel side of equal angles); a **parallelogram** (opposite sides are parallel); a **kite** (two adjacent sides of equal length, the other two sides of equal length); a **rhombus** (four sides of equal length); a **rectangle** (each angle is a right angle); or a **square** (four sides of equal length, each angle a right angle). *Quadrangular prisms* and *quadrangular pyramids* are ones whose bases are quadrilateral.

quantifier

In symbolic logic, the *universal quantifier* ∀ indicates "for every" or "for all." For example, $\forall\ x\ A, p(x)$ means for all x belonging to A, the **proposition** $p(x)$ is valid. The *existential quantifier* ∃ indicates "there exists." So, for instance, $\exists\ x\ A,\ p(x)$ means there exists at least one x, belonging in A, for which the proposition $p(x)$ is valid.

quantum field theory

The study of force fields such as the electromagnetic field in the context of **quantum mechanics**, and often special **relativity theory**. The mainstay of modern high-energy physics.

quantum mechanics

The science and mathematics that describe the behavior of nature at the atomic and subatomic level. At the heart of quantum mechanics are two basic concepts: (1) that every small bit of matter or energy can behave as if it were either a particle or a wave; and (2) that certain combinations of properties such as position and velocity, and energy and time, can't be known with arbitrary precision. The latter idea is encapsulated in *Heisenberg's uncertainty principle.* See also **many worlds hypothesis**.

quartic

A **polynomial** or polynomial equation that contains the fourth power of the variable, but no higher power. Many famous curves are described by such equations, including the **bicorn**, **Cartesian oval**, **conchoid**, **deltoid**, **devil's curve**, **folium**, **kampyle of Eudoxus**, and **limacon of Pascal**.

quartile

The first quartile of a **sequence** of numbers is the number such that one quarter of the numbers in the sequence are less than this number.

quasicrystal

A strange type of solid whose atomic structure is very regular but never quite repeats. Quasicrystalline structures don't have a simple unit cell that can be repeated periodically in all directions to fill space, although they do have local patterns that repeat almost periodically. They also have local rotational symmetries, such as those of a **pentagon**, that can't exist in ordinary crystals. Prior to the discovery of quasicrystals, it was thought that five-fold crystal symmetry was impossible, because there are no space-filling periodic tilings of this kind. The best known examples of quasicrystals resemble **Penrose tiling**s, which use repeated copies of two different rhombi to cover an infinite plane in intricate, interlocking patterns. In fact, some quasicrystals can be sliced in such a way that the atoms on the surface follow the exact pattern of the Penrose tiling.[293]

quasiperiodic

Refers to a form of motion that is regular but never exactly repeating. Quasiperiodic motion is always composed of multiple but simpler **periodic** motions. In the general case for motion that is the sum of simpler periodic motions, if there exists a length of time that evenly divides the frequencies of the underlying motions, then the composite motion will also be periodic; however, if no such length of time exists, then the motion will be quasiperiodic.

quasiregular polyhedron

A **polyhedron** that consists of two sets of regular **polygon**s, m-sided and n-sided respectively, and is constructed so that each polygon in one set is surrounded by members of the other set. There are three convex quasiregular solids: the *cuboctahedron* ($m = 3$, $n = 4$), the *icosidodecahedron* ($m = 3$, $n = 5$), and the **octahedron** ($m = n = 3$). In each case four faces meet at each vertex in the cyclic order (m, n, m, n). Because of this, these polyhedra have some special properties, one of which is that their edges form a system of **great circle**s. The edges of the octahedron form three squares; the edges of the cuboctahedron form four hexagons, and the edges of the icosidodecahedron form six decagons. Among the nonconvex polyhedra are two examples of type (m, n, m, n): the *dodecadodecahedron* ($m = 5$, $n = 5/2$) and the *great icosidodecahedron* ($m = 3$, $n = 5/2$), which can be made by truncating the **Kepler-Poinsot polyhedra** at their edge midpoints. There are also three nonconvex examples of type (m, n, m, n, m, n): the *small triambic icosidodecahedron* ($m = 3$, $n = 5/2$), the *triambic dodecadodecahedron* ($m = 5/3$, $n = 5$), and the *great triambic icosidodecahedron* ($m = 3$, $n = 5$). Finally, there is a group of nine *hemihedra,* in which some faces pass through the polyhedron's center. These hemifaces each cut a sphere into two hemispheres.

quaternion

An ordered **set** of four numbers. Quaternions, first introduced by William **Hamilton**, can also be written in the

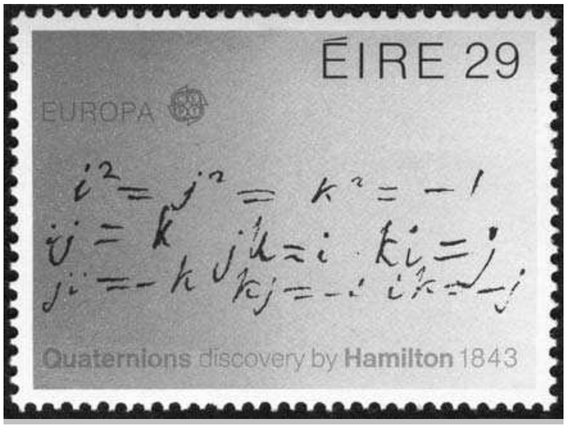
ÉIRE 29
EUROPA
$i^2 = j^2 = k^2 = -1$
$ij = k \quad jk = i \quad ki = j$
$ji = -k \quad kj = -i \quad ik = -j$
Quaternions discovery by Hamilton 1843

quaternion An Irish stamp showing the quaternion equations in Hamilton's own hand.

form $a + bi + cj + dk$, where a, b, c, and d are real numbers and i, j, and k are imaginary numbers, which is similar to that of **complex numbers**. Whereas complex numbers can be represented by points of a two-dimensional plane, quaternions can be viewed as points in the **fourth dimension**. For a while, quaternions were very influential: they were taught in many mathematics departments in the United States in the late 1800s, and were a mandatory topic of study at Dublin, where Hamilton ran the observatory. But then they were driven out by the **vector** notation of William Gibbs and Oliver Heaviside. Had quaternions come along later, when theoretical physicists were trying to understand patterns among subatomic particles, they may have found a place in modern science; after all, the unit quaternions form the group SU(2), which is perfect for studying spin-$\frac{1}{2}$ particles. But the way things turned out, quaternions had fallen from favor by the twentieth century and Wolfgang Pauli used 2×2 complex matrices instead to describe the generators of SU(2).

queens puzzle

A famous **chess** problem that asks in how many ways eight queens can be placed on a chessboard so that no two attack each other. The generalized problem, to find how many ways n queens can be placed on an $n \times n$ board so that no two attack each other, was first posed by Franz Nauck in 1850. In 1874 Günther and Glaisher described methods for solving this problem based on **determinants**. The number of distinct solutions, not counting rotations and reflections, for board sizes ranging from 1×1 to 10×10 is 1, 0, 0, 1, 2, 1, 6, 12, 46, and 92, respectively. The 6×6 puzzle, for which there is a solitary unique solution, was sold for one penny in Victorian London in the form of a wooden board with 36 holes into which pins were placed.

quine

> *Before we put the motion "that the motion be now put," should we not first put the motion "that the motion 'that the motion be now put' be now put?"*
>
> –Chairman of the meeting of the Society of Logicians

A term named by Douglas **Hofstadter** after the Harvard logician Willard van Orman Quine. It can be used either as a noun or a verb. (1) Quine (noun). A computer program that produces an exact copy of itself (or, alternatively, that prints its own listing.) This means that when the program is run, it must duplicate (or print out) precisely those instructions that the programmer wrote as part of the program, including the instructions that do the copying (or printing) and the data used in the copying (or printing.) A respectable quine–one that doesn't cheat–is not allowed to do anything as underhand or trivial as seeking the source file on the disk, opening it, and copying (or printing) its contents. Although writing a quine is not always easy, and in fact may seem impossible, it can always be done in any programming language that is Turing complete (see **Turing machine**), which includes every programming language actually in use. (2) Quine (verb). To write a sentence fragment a first time, and then to write it a second time, but with quotation marks around it. For example, if we quine "say," we get "say 'say' "). Thus, if we quine "quine", we get "quine 'quine,' " so that the sentence "quine 'quine' " is a quine. In this linguistic analogy, the verb "to quine," plays the role of the code, and "quine" in quotation marks plays the role of the data.

quintic

A **polynomial** or polynomial equation that contains the fifth power of the variable, but no higher power. Niels **Abel** and Evariste **Galois** independently proved that although there exist formulas for the general solution of **quadratic**, **cubic**, and **quartic** equations, no such formula exists for quintic equations.

quipu

A recording device generally associated with the Incas, who ruled Peru before the Spanish conquest. Quipu consisted of a number of color-coded cords; knots of various kinds were tied on these cords to represent a variety of information. An important use of quipu was to record numbers for use in trade, keeping accounts, and calendars, but the knotted strings might also have served as mnemonics for important historical events, astronomical data, and mythology (the Incas had no written language). Some evidence supports the idea that the knots and cords followed in a decimal system. Similar devices were used by several other Indian tribes and also described in Chinese and Persian documents from the fifth and sixth centuries B.C., and are still used by shepherds in the Andes for keeping accounts of their herds.

quotient

The number of times that one number can be divided exactly into another.

R

radian

A unit of angular measurement such that there are 2π radians in a complete circle. One radian = $180/\pi$ degrees. One radian is approximately 57.3°.

radical

The symbol that indicates a **root**, $\sqrt[n]{\ }$. It seems to have been first used in 1525 by Christoff Rudolff (1499–1545) in his *Die Coss*.

radical axis

The **locus** of points of equal power with respect to two **circle**s. The *radical center* of three circles is the common point of intersection of the radical axes of each pair of circles.

radius

The distance from the center of a **circle** to its circumference, or from the center of a regular **polygon** to any one of its vertices. The *radius of curvature, r,* at any point of a curve is $r = 1/\kappa$, where κ is the **curvature**.

radix

See **base**.

railroad problems

See **shunting puzzles**.

Ramanujan, Srinivasa Aaiyangar (1887–1920)

An extraordinary, largely self-taught, Indian mathematician who, in the most unorthodox way, made significant contributions to **number theory**, including the subject of **elliptic function**s, **continued fraction**s, and **infinite series**. During much of his early work, he was unaware that he was rediscovering results that had taken other mathematicians centuries to achieve. But even in cases where he arrived at conclusions already known, he'd often travel an original route, and, in many cases, almost purely by intuition. Ramanujan was employed in a lowly clerk's position in Madras, when, in 1913, he wrote letters to three eminent mathematicians in England describing some of his results. Two of the three letters were returned unopened. However, G. H. **Hardy** recognized Ramanujan's abilities and arranged for him to come to Cambridge. Because of his lack of formal training, Ramanujan sometimes failed to distinguish between formal proof and apparent truth based on intuition or numerical evidence. His extraordinary innate familiarity with numbers was revealed by an incident recalled by Hardy:[150] "I remember once going to see him when he was lying ill at Putney. I had ridden in taxi cab number 1729 and remarked that the number seemed to me rather a dull one, and that I hoped it was not an unfavorable omen. 'No,' he replied, 'it is a very interesting number; it is the smallest number expressible as the sum of two cubes in two different ways [$1729 = 1^3 + 12^3 = 9^3 + 10^3$].' "

Unfortunately, Ramanujan's health deteriorated rapidly in England, perhaps due to the unfamiliar climate and food, and to the isolation which Ramanujan felt as the sole Indian and a devout Hindu in a culture which was alien to him. Ramanujan was sent home to recuperate in 1919, but tragically died the following year at the age of only 32. Although he published some of his results in journals, much of his work and conclusions have only

SRINIVASA RAMANUJAN
1887–1920
न.पै.
15 nP
INDIA POSTAGE
INDIA SECURITY PRESS

Ramanujan, Srinivasa Aaiyangar The enigmatic mathematician on an Indian commemorative stamp.

come to light more recently, from a scrutiny of his disorganized but fascinating notebooks.[184]

Ramsey theory

A branch of mathematics that asks questions such as: Can order always be found in what appears to be disorder? If so, how much can be found and how big a chunk of disorder is needed to find a particular amount of order in it? Ramsey theory is named after the English mathematician Frank P. Ramsey (1904–1930) who started the field in 1928 while wrestling with a problem in logic. (Frank's one-year-younger brother, Arthur, served as Archbishop of Canterbury from 1961 to 1974.) His life was cut short at the age of 26, following a bout of jaundice. Ramsey suspected that if a system was big enough, even if it seemed to be disorderly to an arbitrary degree, it was bound to contain pockets of order from which information about the system could be gleaned.

random

Without cause; not **compressible**; obeying the statistics of a fair coin toss.

random number

A number generated by a process that is fundamentally nondeterministic and unpredictable. Computer-generated "random numbers," which are calculated through a deterministic process, cannot, by definition, be **random**. Given knowledge of the **algorithm** used to create the numbers and its internal state, it's possible to predict all of the numbers returned by subsequent calls to the algorithm. For this reason the numbers produced by computer-based "random number generators" are often referred to as *pseudorandom numbers.* In the case of genuinely random numbers, knowledge of one number or an arbitrarily long sequence of numbers offers no clue of the next number to be generated. Humans are among the worst random number gen-

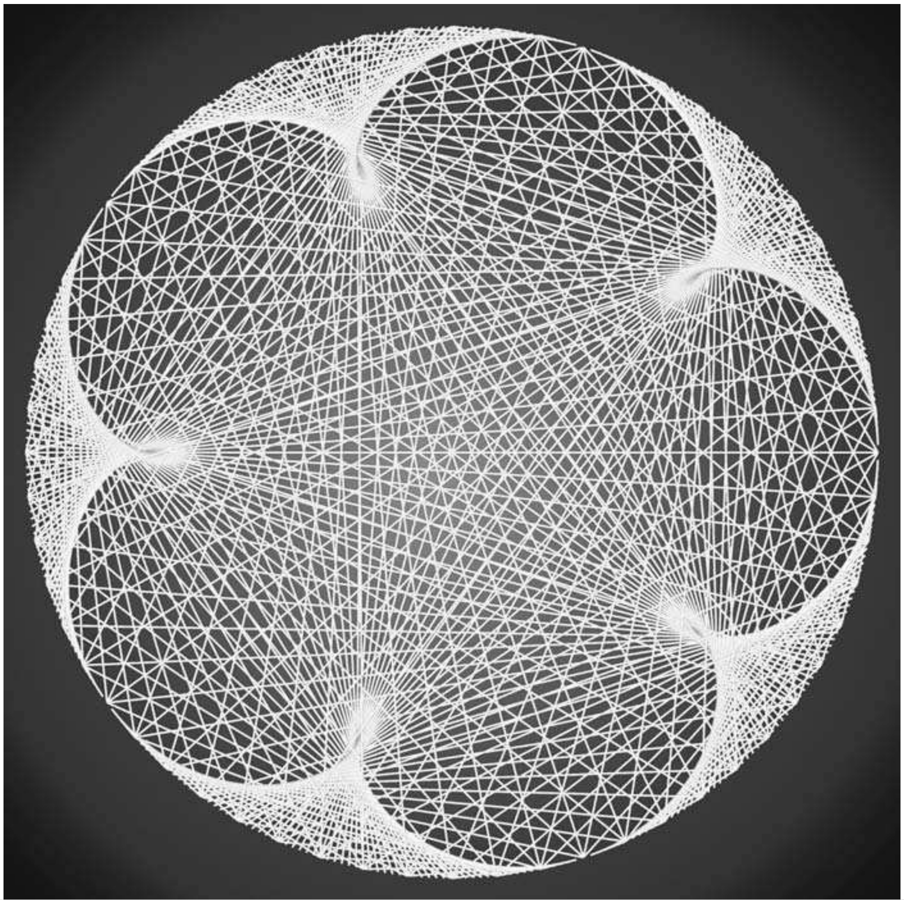

ranunculoid A ranunculoid curve spun by thread on a computer loom. *Jos Leys, www.josleys.com* erators. Ask someone to pick a number "at random" between 1 and 20, and the number they're most likely to choose is 17. *Psychologically random numbers,* such as 17, are usually odd and don't end in 5, so that they frequently tend to be **prime numbers**. See also **Chaitin's constant**.

random walk

A process in which the position of a particle changes by discrete steps of fixed length, and the direction of each step is chosen randomly. Random walks have interesting mathematical properties that vary greatly depending on the number of dimensions in which the walk takes place and whether it is confined to a **lattice**. For a random walk in one dimension there are only two directions to choose from. Imagine a drunken person wandering on the **number line** who starts at 0, and then moves left or right (+/−1) with probability ½. The probability that the walker will eventually return to his starting point is 1; in other words, it is certain to happen. The same is true for a random walk in the plane, moving on the integer lattice points, with probability ¼ in each of the coordinate directions: the probability of ending up back at the starting point is 1. However, the situation changes in three dimensions. Suppose a drunken fly moves randomly from one point to another in a three-dimensional lattice with a probability of 1 in 6 of arriving at any of the six adjacent lattice points on each move. No matter how long the fly roams, it has only a 0.34054 . . . probability of ever getting back to where it started. Probabilists say that random walks on the line and plane are *recurrent,* whereas random walks in three dimensions or more are *transient.* Effectively, this is because there is so much more "space" in three or more dimensions. The numbers giving the probability of eventually returning to the starting point are known as *random walk constants.* The random thermal perturbations in a liquid are responsible for a random walk phenomenon known as *Brownian motion,* and the collisions of molecules in a gas are a random walk responsible for *diffusion.*

range

(1) The set of possible values in which a **function**'s output can be. See also **codomain**. (2) The set of all points on a **line segment**.

rank

(1) Any of the rows of squares running crosswise to the files on a playing board in **chess** or checkers. (2) The rank of a **matrix** is equal to the dimension of the largest submatrix that can be obtained by deleting rows and columns of the parent matrix and that has a nonzero determinant. See also **tensor**.

ranunculoid

An **epicycloid** with five cusps ($n = 5$), named after the buttercup genus *Ranunculus.*

ratio

A **rational number** of the form *a/b* where *a* is called the *numerator* and *b* is called the *denominator.* It may be written with a colon (:), as a fraction, or with the word *to.*

rational number

A number that can be written as an ordinary **fraction**—a ratio, *a/b,* of two integers, *a* and *b,* where *b* isn't zero—or as a decimal expansion that either stops (like 4.58) or is periodic (like 1.315315 . . .). Other examples include 1, 1.2, 385.66, and ⅓. Rational numbers are countable, which means that, although there are infinitely many of them, they can always be put in a definite order, from smallest to largest, and can thus be counted. They also form what's called a *densely ordered set;* in other words, between any two rationals there always sits another one—in fact infinitely many others. The rational numbers are a subset of the **real numbers**; real numbers that aren't rational are called, rationally enough, **irrational numbers**. Although rationals are dense on the real **number line**, in the sense that any **open** set contains a rational, they're pretty sparse in comparison with the irrationals. One way to think of this is that the **infinity** of rationals (which, strangely enough, is exactly the same size as the infinity of whole numbers) is smaller than the infinity of irrational numbers. Another way to grasp the scarcity versus density issue, is to realize that the rationals can be covered by a set whose "length" is arbitrarily small. In other words, given a string of any positive length, no matter how short, it will still be long enough to cover all the rationals. In mathematical parlance, the rationals are a **measure** *zero* set. The irrationals, by contrast, are a measure one set. This difference in measure means that the rationals and irrationals are quite different even though a rational can always be found between any two irrationals, and an irrational exists between any two rationals.

raven paradox

A **paradox** put forward by the German logician Carl Hempel (1905–1997) in the 1940s to highlight a situation where the logic of **induction** seems to fly in the face of intuition. According to the *principle of induction,* the more that a theory is supported by observation, the greater the probability that the theory is true. Consider, said Hempel, the theory that all ravens are black. After each observation of a black raven, our belief in the theory "all ravens are black" increases. But here's the rub. The statement "all ravens are black" is logically equivalent to the statement "all nonblack things are nonravens." The observation of a white swan is consistent with this statement. A white swan is a nonblack thing, and when we examine it, we observe that it is a nonraven. So by the principle of induction, observing a white swan should increase our belief that all ravens are black!

Many solutions have been offered to this enigma. The American logician Nelson Goodman (1906–1998) suggested imposing restrictions to our reasoning, such as never considering an instance as support for "All P are Q" if it would also support "No P are Q." Others have questioned the *principle of equivalence.* Perhaps seeing a white swan should strengthen our belief in the theory "all nonblack things are nonravens," without increasing our conviction that "all ravens are black." Yet others have argued that our intuition is flawed. Observing a white swan really does increase the probability that all ravens are black! After all, if you were shown all the nonblack things in existence, and you noticed that none was a raven, then you could properly conclude that all ravens were black. The example only seems counterintuitive because the set of nonblack things is vastly larger than the set of ravens. Thus observing one more nonblack thing that is not a raven should make a tiny difference to our degree of belief in the proposition compared to the difference made by observing one more raven that is black.

A way to sidestep the paradox is by using **Bayes's theorem**. According to this the probability of a hypothesis H must be multiplied by the ratio:

$$\frac{\text{probability of observing } X \text{ if } H \text{ is true}}{\text{probability of observing } X}$$

If a swan is picked at random, the probability of it being white is independent of the colors of ravens. The numerator in the above ratio will equal the denominator, the ratio will equal one, and the probability will remain unchanged. Seeing a white swan doesn't affect our belief about whether all ravens are black. If a nonblack thing is chosen at random, and a white swan is shown, then the numerator will be bigger than the denominator by a tiny amount. Seeing the white swan will only slightly increase our belief that all ravens are black. We'd have to see almost every nonblack thing in the universe (and see that they're all nonravens) before our belief in "all ravens are black" would increase appreciably. In both cases, these results are in line with intuition.

ray

A straight path of points that begins at one point and continues in one direction.

real number

Any number that can be represented as a decimal, possibly infinitely long and nonrepeating. Real numbers stand in one-to-one correspondence with the points on a continuous line, known as the *real number line,* that stretches from zero to infinity in both directions. The set of real numbers contains the set of all **rational number**s and the set of all **irrational number**s. The name "real number" is a retronym, coined by René **Descartes** in response to the concept of **imaginary number**s. Number systems that are even more general than the real numbers include the **complex numbers** and, of much more recent discovery, **hyperreal number**s and **surreal number**s.

realm

A term advocated for a three-dimensional version of the two-dimensional **plane**.

reciprocal

One over a given number; for example, the reciprocal of 4 is $^1/_4$.

Recorde, Robert (c. 1510–1558)

A Welsh physician and mathematician, born in Tenby, Pembrokeshire, and trained at Oxford and Cambridge, who held various positions, including master of the mint in Bristol and later in Ireland, and wrote a number of influential math textbooks. These books formed a complete course and were written in English, rather than the usual Latin or Greek, so that they could be read by anyone. In one of them Recorde introduces the "=" sign for "equals."

rectangle

A **quadrilateral** whose interior angles are all 90°. If all of its sides are the same length, it is a **square**. The smallest square that can be cut into $m \times n$ rectangles, such that all m and n are different integers, is the 11 × 11 square, and the **tiling** uses five rectangles. The smallest rectangle that can be cut into $m \times n$ rectangles, such that all m and n are different integers, is the 9 × 13 rectangle; this tiling also uses five rectangles.

rectangular coordinates

See **Cartesian coordinates**.

rectangular hyperbola

See **hyperbola**.

recursion

See **recursion**.

No, seriously. Recursion is a process that wraps back on itself and feeds the output of a process or function back in as the input. Using some sort of *recurrence relation,* an entire class of objects can be built up from a few initial values and a small number of rules. The **Fibonacci sequence**, for example, is defined recursively, as are many **fractal** figures. *Self-recursion,* of which the first two lines of this entry are an example, leads to an endless feedback loop. One, somewhat disturbing, notion of reality is that we live in a *recursive universe* in which nature resembles an infinite nest of Russian dolls. One day, we will simulate the big bang in one of our supercomputers. Inside this artificial universe (rather like an immense *Star Trek* "holodeck" simulation) will evolve new star systems and new life forms. They too will evolve intelligence one day and invent computers. And they too may one day simulate the big bang inside one of their supercomputers. This chain of existence will continue as long as we continue to run our simulation. Our universe will continue to exist as long as our parent universes' continue to run their simulations. And therein lies madness.

recursive function

Strictly speaking, a **function** that is *computable;* however, in the usual sense of the word, a function is said to be recursive if its definition makes reference to itself (see **recursion**). For example, **factorial** can be defined as $x! = x(x - 1)!$ with the base case of 1! equal to 1. See also **self-referential sentence**.

recursively enumerable set

A potentially infinite **set** whose members can be enumerated by a **universal computer**; however, a universal computer may not be able to determine that something is not a member of a recursively enumerable set. The *halting set,* a concept related to the **halting problem**, is recursively enumerable but not recursive.

reductio ad absurdum

"Reduction to the absurd"; the process of demonstrating that an idea is probably false by first assuming its truth, and then showing how that truth leads to conclusions that can't possibly be true. In *A Mathematician's Apology* (1941),[151] G. H. **Hardy** said: "Reductio ad absurdum, which Euclid loved so much, is one of a mathematician's finest weapons. It is a far finer gambit than any chess play: a chess player may offer the sacrifice of a pawn or even a piece, but a mathematician offers the game."

reductionism

The idea that nature can be understood by taking it apart. In other words, knowing the lowest-level details of how things work (at, say, the level of subatomic physics) reveals how higher-level phenomena come about. This is a bottom-up way of looking at the universe, and is the exact opposite of **holism**.

redundancy

The existence of repetitive patterns or structures. In an important sense, redundancy refers to *order* in a **complex system** since order is defined as the existence of structures that maintain themselves over time. In **information theory**, redundancy refers to repetition in patterns of messages in a communication channel. If the message contains these redundancies, they can be compressed further; for example, a message containing a series of 250 ones, could be compressed into a command that effectively says "and then repeat one 250 times," instead of writing out all 250 ones.

reentrant angle

An inward-pointing angle of a **concave** polygon.

reflection

A way of transforming a shape in the same way that a mirror does. The reflection of a shape in a mirror line is an identical shape that has been flipped over. When an object is placed a certain distance in front of a mirror, its image in the mirror appears the same distance away from the edge of the mirror. Likewise, all the points on a shape and all the points on its image are the same distance away from the mirror line.

reflex angle

An angle between 180° and 360°.

reflexible

Having a plane of mirror symmetry. Compare with **chiral**.

regular polygon

A **polygon** in which all the sides are equal and all the angles are equal.

regular polyhedron

A polyhedron in which every face and vertex figure is regular. There are nine regular polyhedra: the five **Platonic solid**s and the four **Kepler-Poinsot solid**s. However, others are sometimes allowed, depending on the definition of polyhedron.

relativity theory

The physical and mathematical theory due to Albert Einstein (1879–1955) that revolutionized our understanding of **space**, **time**, and gravity. In it, space and time are seen as a unified and inseparable whole–the four-dimensional continuum of **space-time**. The curvature of space-time due to the presence of matter becomes the extraordinary new explanation and interpretation of gravity. Einstein's theory of relativity was published by him in two great parts. Special relativity, published in 1905, deals exclusively with *inertial* frames of reference, that is, reference frames that don't accelerate with respect to one another. Its two central premises are that the laws of physics are the same in all reference frames and that the speed of light (in a vacuum) is constant in all reference frames. General relativity, published in 1915, centers on the *equivalence principle,* the idea that acceleration and gravity are equivalent. See also **non-Euclidean space**.

renormalization

A mathematical technique for looking at a physical system at different levels of magnification.

repdigit

A number composed of repetition of a single digit in a given base, generally taken as base 10 unless otherwise specified. For example, the **beast number** 666 is a (base-10) repdigit.

representation theory

A theory that seeks to understand an abstract algebraic system such as a **group** by obtaining it in a more concrete way as a *permutation group* or as a group of matrixes (see **matrix**).

rep-tile

A repetitive **tiling**: a shape with the property that it tiles a larger version of itself, using identical copies of itself. A simple example is a square because four copies of any square tile a larger square. Any triangle also is a rep-tile, because four copies of it tile a larger version of this triangle. Rep-tiles that require *n* tiles to build a larger version of themselves are said to be rep-*n;* thus a square is rep-4. Since any of these larger replicas can be combined to give an even larger, second-generation copy, a rep-*n* tile is also rep-n^2, rep-n^3, and so on. Often tiles have several rep-numbers. If a tile is rep-*n* and rep-*m,* it is also rep-*mn,* since replicas can be built with *n* tiles, then combined, *m* at a time, to give a yet larger version.

The set of rep-tiles is a subset of the set of *irreptiles.* An irreptile is any shape that tiles a larger version of itself using either differently sized or identical copies of itself. The problem to find all irreptiles in the Euclidean plane has been studied but not yet completely solved. A

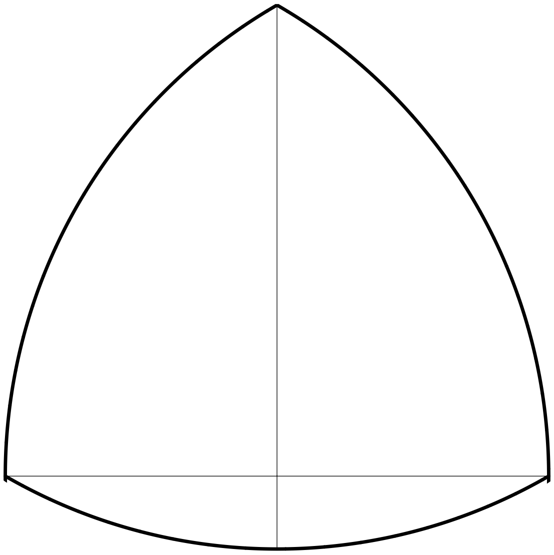

Reuleaux triangle *© Jan Wassenaar, www.2dcurves.com*

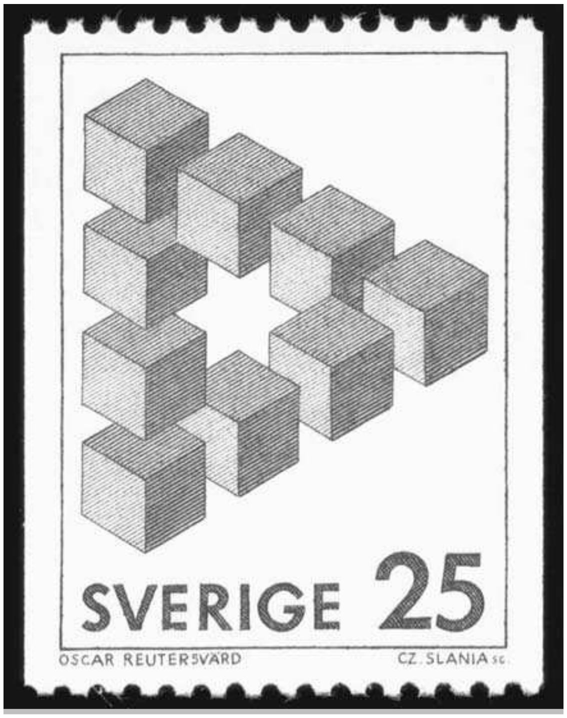

Reutersvärd, Oscar A Swedish stamp depicting one of Reutersvärd's impossible figures.

re-lated set of problems is to find for each irreptile the minimum number of smaller copies needed to tile the original shape; in many cases it is difficult to prove such a minimality. The name "rep-tile" was coined by Simon **Golomb**.

rep-unit

A number whose digits are all units; the rep-unit (repeated unit) with n digits is denoted R_n. For example, $R_1 = 1$, $R_2 = 11$, $R_3 = 111$, and $R_n = (10^n - 1)/9$. R_n divides R_m whenever n divides m. No rep-unit can be a square, but it is not known if one can be a cube. *Rep-unit primes* are rep-units that are **prime number**s. The only known rep-unit primes are R_2 (11), R_{19}, R_{23}, R_{317}, and R_{1031}, though R_{49081} and R_{86453} are suspected primes.

resultant

A **vector** that is the sum of a given set of vectors.

Reuleaux polytope

A **convex** body in the plane or in higher dimensions that, like the **Reuleaux triangle**, consists of pieces of round spheres, each centered at one of the corners of the convex body.

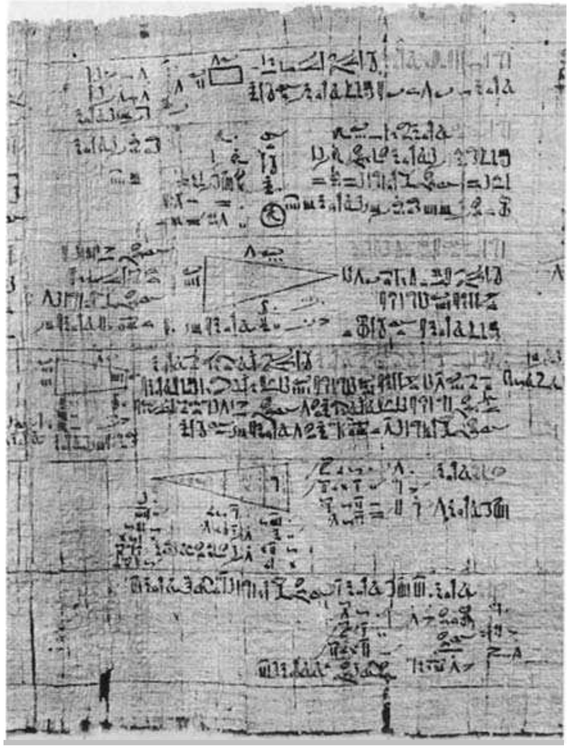

Rhind papyrus *British Museum*

Reuleaux triangle

The simplest noncircular **curve of constant width**; also known as the *Reuleaux wheel,* it is named after the German engineer and mathematician Franz Reuleaux (1829–1905). Although it was known to earlier mathematicians, Reuleaux was the first to show its constant-width properties. To form a Reuleaux triangle, take the three points at the corners of an equilateral triangle and connect each pair of points by a circular arc centered at the remaining point. The ratio of the circumference to the width of the triangle is, remarkably, **pi**. By rotating the centroid of a Reuleaux triangle appropriately, the figure can be made to trace out a square, perfect except for slightly rounded corners. This idea has formed the basis of a drill that will carve out squares, first patented by Harry Watts in 1914. Bits for square, pentagonal, hexagonal, and octagonal holes are still sold by the Watts Brothers Tool Works in Wilmerding, Pennsylvania. The actual drill bit for the square is a Reuleaux triangle made concave in three spots to allow for unobstructed corner-cutting and the discharge of shavings. The Reuleaux triangle may also form the shape of the piston in a rotary, or Wankel, engine, in which gasoline burns in crescent-shaped chambers, turning a rotating piston that drives an axle through its center.

Reutersvärd, Oscar (1915–)

A Swedish artist who pioneered the creation and design of **impossible figure**s. His work in this area goes back to one day in 1934 when, as a young student in Stockholm, he started doodling in the margins of a textbook during a long lecture. Reutersvärd's doodle began with an outline of a perfect six-pointed star. Once the star was complete he added cubes around the star, nestled into the spaces between the points. He soon realized that what he'd drawn was paradoxical: something that couldn't be built in the real world. A different version of this figure, independently created by Roger **Penrose**, would later be called the **Penrose triangle**. Thus began a lifelong fascination with such objects, which later included work on the impossible staircase (a design he sketched in 1950 while on a cross-country train ride), which is known as the **Penrose stairway**, and the discovery of the **tribar illusion**. In the early 1980s the Swedish government honored Reutersvärd's achievements by issuing a set of three stamps depicting impossible figures, including a version of his 1934 weird cubes design.

Rhind papyrus

A papyrus scroll, 33 cm high and 565 cm wide, found in a tomb in Thebes, that is the most valuable source of information we have about Egyptian mathematics. The scroll was bought at a market in Luxor in 1858 by a 25-year-old Scotsman, Henry Rhind, who went to Egypt for health reasons and became interested in archeology. After his early death at the age of 30, the scroll found its way to the British Museum in London in 1864 and has remained there ever since. It is often referred to as the *Rhind mathematical papyrus,* or RMP for short. The hieroglyphs on the papyrus were deciphered in 1842, while the Babylonian clay-tablet cuneiform writing was deciphered later in the nineteenth century. The text begins by stating that the scribe "Ahmes" is writing it (in about 1600 B.C., and he is thus the earliest named individual in the history of mathematics) but that he has copied it from "ancient writings," which probably go back to at least 2000 B.C. Although there is some strictly practical mathematics on the papyrus, including calculations needed for surveying, building, and accounting, some of which involve **Egyptian fractions**, many of the problems in the RMP take the form of arithmetic puzzles. One of these is: Seven houses contain seven cats. Each cat kills seven mice. Each mouse had eaten seven ears of grain. Each ear of grain would have produced seven hekats of wheat. What is the total of all of these? This is very similar to the **St. Ives problem**.

There are also four lesser documents preserving Egyptian arithmetic: the Moscow papyrus and the Berlin papyrus (named for the places they are kept), the Kahun papyrus (named for where it was found), and the Leather Roll (named for its composition). The Moscow papyrus is sometimes called the Golenischev papyrus after the Russian V. S. Golenischev, who purchased it in 1893 from two Egyptian brothers who found it in a tomb at Deir el-Bahri. It measures 8 cm high and 540 cm wide and contains 25 problems and their solutions. The most unusual are the tenth, which seems to give the area of the surface of a hemisphere or perhaps a cylinder, and the fourteenth, which gives the formula for the volume of the frustum of a **pyramid**.

rhombus

A **quadrilateral** in which both pairs of opposite sides are parallel and all sides are the same length, that is, an equilateral **parallelogram**. A rhombus is also sometimes called a *rhomb* or a diamond. A rhombus whose acute angles are 45° is called a *lozenge.* The diagonals p and q of a rhombus are perpendicular and satisfy the relationship $p^2 + q^2 = 4a^2$. The area of a rhombus is given by $A = \frac{1}{2} pq$.

Richard's paradox

See **Berry's paradox**.

Riemann, (Georg Friedrich) Bernhard (1826–1866)

A German mathematician who was the first person to provide a thorough treatment of **non-Euclidean geometry** and to see how it might be applied in physics; he thus helped pave the way for the general **relativity theory**. Among several profound aspects of mathematics now named after him are the **Riemann hypothesis** and the related **Riemann zeta function**. His father, a Lutheran pastor, encouraged him to study theology at Göttingen. But even as a child Riemann had shown a tremendous aptitude for mathematics and, in 1847, he persuaded his father to let him go to Berlin to learn mathematics from the likes of Karl **Jacobi**, Peter **Dirichlet**, and Jakob **Steiner**. Two years later, he returned to Göttingen to study for his Ph.D. and begin his climb up the professorial ladder. In 1854, his inaugural lecture, "Concerning the hypotheses which underlie geometry," covered a breathtaking array of topics, including a workable definition of the curvature of space and how it could be measured, the first description of **elliptical geometry**, and, most important of all, the extension of geometry into more than three dimensions with the aid of algebra.

Riemann hypothesis

> *If I were to awaken after having slept for a thousand years, my first question would be: Has the Riemann Hypothesis been proven?*
>
> –David Hilbert

The most important open question in **number theory** and, possibly, in the whole of mathematics. A $1 million prize has been offered by the Clay Mathematics Institute for a proof. The hypothesis was first formulated by Bernhard **Riemann** in 1859, was included in David **Hilbert**'s list of challenging problems for twentieth-century mathematicians, and is widely believed to be true. Yet a proof remains tantalizingly out of reach. What the Riemann hypothesis says is that the nontrivial zeros of the **Riemann zeta function** all have real parts equal to $\frac{1}{2}$. In plain language, the hypothesis asserts that there is an underlying order, akin to musical harmonics, in the way **prime numbers** are distributed. It's known that for any given number n there are approximately $n/\log n$ prime numbers that are less than n. The formula is not exact: sometimes it is a little high and sometimes it is a little low. Riemann looked at these deviations and found that they contain periodicities. His hypothesis quantifies and formalizes this discovery, positing that the zeros of the zeta function can be regarded as the harmonic frequencies in the distribution of primes. If the Riemann hypothesis turns out to be true, what do these harmonics in the "music" of the primes mean? Remarkably, it is been found by the English physicist Michael Berry and his colleagues that there is a deep connection between the harmonics–the Riemann zeros–and the allowable energy states of physical systems that are on the border between the quantum world (see **quantum**

mechanics) and the everyday world of classical physics. The Riemann harmonics, or "magic numbers," behave exactly like the energy levels in quantum systems that classically would be chaotic. This deep connection between number theory and the physics of the real universe, if upheld, is utterly astonishing. If the Riemann hypothesis is proved true, it could open an entirely new window on the nature of reality and the relationship between the abstract world of mathematics and the behavior of matter and energy. On the other hand, if it is disproven, there will be an even deeper mystery to explore: How can the Riemann zeta function so convincingly mimic a quantum system without actually being one?[40]

Riemann integral

The kind of **integral** familiar from **calculus** texts and normally used by scientists and engineers, in which the process of **integration** of function on an interval amounts to finding the area under the curve.

Riemann sphere

A topological sphere consisting of the **complex plane** and the point at infinity; an example of a **Riemann surface**.

Riemann surface

Also known as a *complex curve,* a complex **manifold** with one complex dimension. The Riemann surface is a conformal structure (see **conformal mapping**).

Riemann zeta function $\zeta(s)$

> *We may–paraphrasing the famous sentence of George Orwell–say that "all mathematics is beautiful, yet some is more beautiful than the other." But the most beautiful in all mathematics is the zeta function. There is no doubt about it.*
>
> –Krzysztof Maslanka, Polish cosmologist

One of the most profound and mysterious objects in modern mathematics; from it has sprung the **Riemann hypothesis** and all that this conjecture seems to imply. The Riemann zeta function is closely tied to the distribution of **prime numbers**. It is an extension of the *Euler zeta function,* first studied by Leonhard **Euler**, which is the sum

$$\zeta(s) = 1 + 1/2^s + 1/3^s + 1/4^s + \ldots = \sum_{n=1}^{\infty} 1/n^s.$$

Euler found that this function is linked to the occurrence of prime numbers by the following fundamental relationship:

$$\begin{aligned}&\zeta(s) = 1 + 1/2^s + 1/3^s + 1/4^s + \ldots = 2^n/(2^n - 1)\\ &\times 3^n/(3^n - 1) \times 5^n/(5^n - 1) \times 7^n/(7^n - 1) \times \ldots\\ &= \prod_p 1/(1 - p^{-s})\end{aligned}$$

The Riemann zeta function extends the definition of Euler's zeta function to all **complex numbers**.

Riemannian geometry

See **elliptical geometry**.

right

A *right angle* is an angle of 90°. A *right triangle* is a triangle that contains a right angle.

ring

(1) Another name for an **annulus**. (2) A number system in which addition, subtraction, and multiplication are always defined and the **associative** and **distributive** laws are valid. Compare with **field**.

Rithmomachia

A medieval, chesslike board game for two players that is based on the number theories of Pythagoras and Boethius. Rithmomachia or "battle of numbers" (*rithmo,* "arithmetic, numbers"; *machia* "battle") dates back to about A.D. 1150 although the earliest publication of the rules was by Jean de Boissiere in the sixteenth century. Used as an educational tool (the only game allowed in medieval schools and universities) and as an intellectual exercise, it enjoyed a last wave of popularity during the Renaissance before the early Scientific Revolution led to its disappearance.

The game is played on an 8 × 16 board. Each player starts with either 24 black or white pieces: 8 circles, 8 triangles, 7 squares, and 1 pyramid. Each of these has a *number,* which is how many places it can move. A square can move 4 spaces, a triangle 3, a circle 1, and the pyramid as many spaces as the player chooses. Opponent pieces can be captured in a variety of ways: *siege capture* (surrounding the opponent piece on all four sides; *meeting capture* (attacking a piece with the same type of piece); *assault capture* (achieved if the piece's number times the number of spaces it moved lands it next to an opposing piece that equals the product); *ambuscade* (achieved if two pieces of a player that are on either sides of a piece sum to equal the opponent's piece). There are also a number of different ways to win, including: *de corpore* (players agree on a number of pieces to be captured); *de bonis* (players agree on a number value target); *de lite* (the winner is determined by the sum of the pieces as well as the number of digits on all those pieces); *victoria magna* (if there is a common difference between the pieces a player has captured, or the squares of three consecutive integers are captured; or there is a difference of 2, 4, and 6 in even pieces, or 3, 5, and 7 in odd pieces).

river-crossing problem

A puzzle in which a variety of objects and living things, some of them mutually incompatible, must be conveyed in small groups from one side of a river to another without any loss along the way. The earliest known examples are in *Propositiones ad Acuendos Juvenes* (Propositions for sharpening youths), which is generally attributed to Abbott **Alcuin**. They are: the problem of three jealous husbands (each of whom won't let another man be alone with his wife), the problem of the two adults and two children where the children weigh half as much as the adults (and the boat has a limited weight capacity), and the problem of the wolf, the goat, and the cabbage. In the last case, the difficulty is that only one item can be ferried across at once but, if left unattended, the sheep will eat the cabbage and the wolf will eat the sheep. The solution, which involves a stratagem common to all these types of problems, is to bring back to the starting bank of the river an item that has already been taken across. In this case, the sheep must be taken across first, followed by either the cabbage or the wolf, but then the sheep must be brought back before the next item is taken across to avoid the sheep becoming either a diner or a dinner.

These medieval puzzles were considered and elaborated on by Niccoló **Tartaglia**, Luca **Pacioli**, and Claude-Gaspar **Bachet**, and even more so by later mathematicians such as Edouard **Lucas** and Gaston **Tarry**. Ways of complicating river-crossing problems include adding more people and objects, using a bigger boat, and inserting an island in the river. The reader may care to try the problem of the missionaries and the cannibals.

PUZZLE

Three missionaries and three cannibals must cross a river in a boat that holds a maximum of two people. If the cannibals outnumber the missionaries, on either side of the river, the missionaries are in trouble. Each missionary and each cannibal is capable of rowing the boat. How can all six get across the river safely?

Solutions begin on page 369.

Robinson, Abraham (1918–1974)

A German-born mathematician who founded **nonstandard analysis** and did important work in a wide diversity of fields from aerodynamics to mathematical logic. A modern counterpart to Gottfried **Leibniz** in his range of interests and the significance of his research on **infinitesimal**s, Robinson (he changed his name from Robinsohn) taught at various universities in Israel, England, Canada, and the United States.

Rolle's theorem

Suppose a continuous function (see **continuity**) crosses the x-axis at two points a and b and is differentiable at all points between a and b; that is, it has a **tangent** at all points on the curve between a and b. Then there's at least one point between a and b where the **derivative** is 0, and the tangent is parallel to the x-axis.

Roman numerals

A **number system** in which each symbol represents a fixed value regardless of its position; this differs from the place-value system of **Arabic numeral**s. The earliest form of the Roman system was, however, decimal. In this primitive version a series of I's represented any number from 1 to 9, and a new symbol was introduced for each higher power of 10: X for 10, C for 100, and M for 1,000. The symbols V, L, and D, which stand for 5, 50, and 500, are thought to have been introduced by the Etruscans. A common remark is that multiplication and division using Roman numerals is so awkward that it is totally impractical. However, an article by James G. Kennedy in *The American Mathematical Monthly* in 1980 gives algorithms for these operations that are actually more straightforward in the Roman system than in the Arabic. In multiplication the first step is to rewrite the numbers in a simple place-value notation. Seven columns are set up, headed by the symbols M, D, C, L, X, V, and I, and tallies are marked in each column corresponding to the number of times that symbol appears in the multiplicand. For example, if the multiplicand is XIII (13), one tally is marked in the X column and three tallies are marked in the I column. The multiplier is written in the same way. The multiplication itself is done by forming partial products according to two simple rules. In most cases the partial product given by any one tally in the multiplier is simply the set of tallies that represents the multiplicand, shifted to the left an appropriate number of columns. If the multiplier digit is I, the multiplicand is not shifted at all; the multiplicand is shifted one place to the left for V, two places for X, three places for L, and so on. The second rule is applied only when one Etruscan character is multiplied by another. In such cases the tallies representing the multiplicand digit are written twice in the appropriately shifted column and an additional tally is written one column to the right. Once a partial product has been formed for every tally in the multiplier, the tallies in each column are accumulated and replaced by the Roman symbol at the head of the column, giving the final answer. Only a slight change in the method is needed for Roman numerals in "subtractive notation," where 10 is written as IX, and so on. If all this sounds not quite so simple, the method for multiplying Arabic numbers is just as

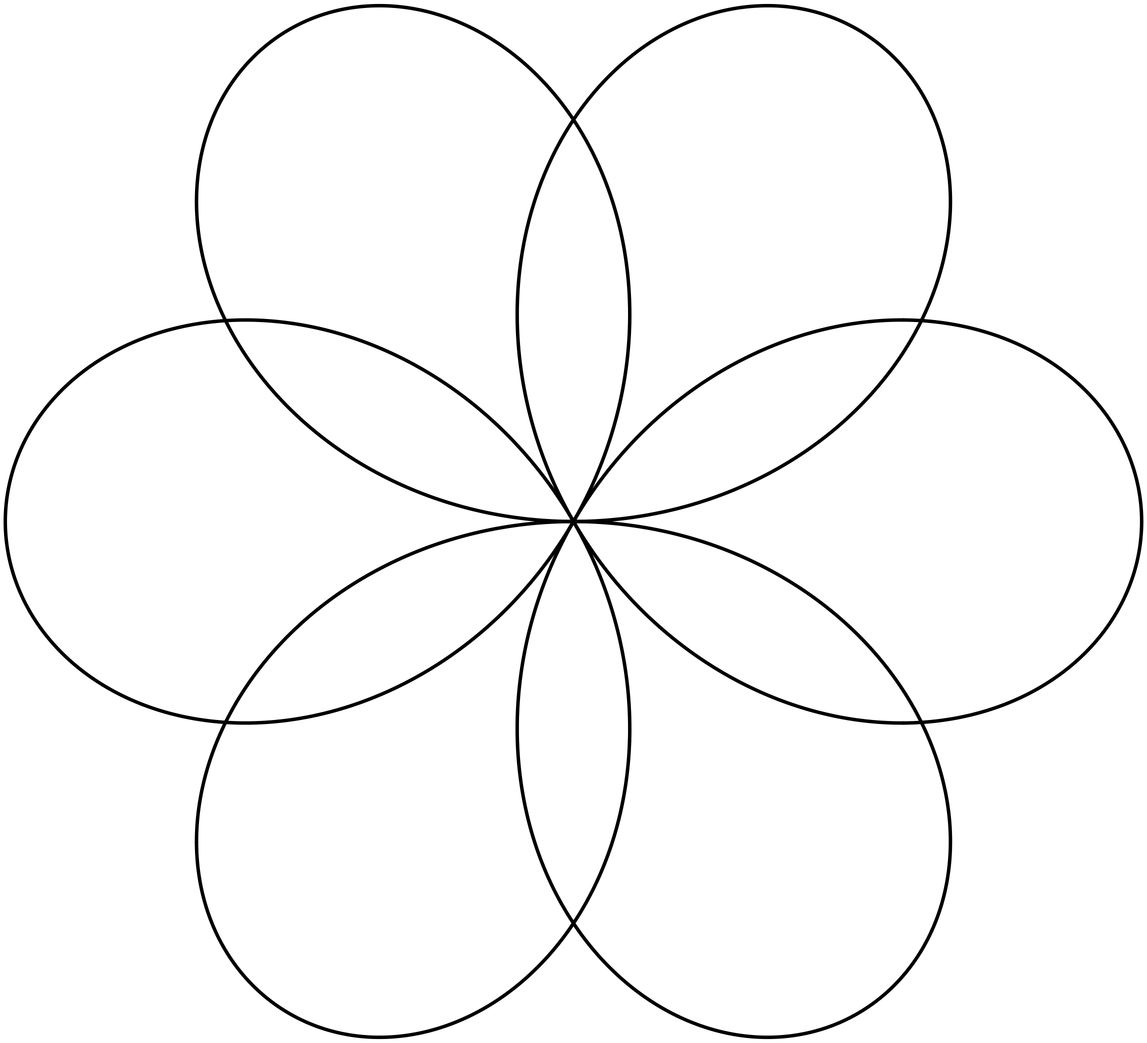

rose curve *© Jan Wassenaar, www.2dcurves.com*

involved if they are written in explicit form. Furthermore, Arabic operations require a multiplication table giving the 100 products of all the possible pairs of Arabic digits. No comparable table is needed with Roman numerals, where all arithmetical operations can be defined in terms of shifting rules, addition, and subtraction.

rooks problem

To find the maximum number of rooks that can be placed on an $n \times n$ **chess**board such that no rook attacks another. Since each rook attacks all squares in the rank and file upon which it rests, this number is *n:* the rooks may be placed along the main diagonal. The total number of ways of placing *n* nonattacking rooks is *n* factorial (*n*!).

root

(1) A number used to build up another number by repeated multiplication. For example, since $2 \times 2 \times 2 = 8$, two is said to be the third root or cube root of eight. (2) A solution of an equation. For example, 3 is a root of the

Rubik's cube *Peter Knoppers, www.buttonius.com*

equation $x^2 = 9$. A root is also called a *zero of a function* because it is a value that will make the function zero ($x = 3$ will make the function $f(x) = x^2 - 9$ zero. The word comes from the Indo-European *werad,* which originally meant the roots of a plant but was later generalized to mean the origins or beginnings of something, whether it was physical or mental.

root of unity

A solution of the equation $x^n = 1$, where n is a positive integer.

rope around the earth puzzle

Imagine a rope that fits snugly all the way around Earth like a ring on a person's finger. Now imagine the rope is made just 1 meter longer and lifted uniformly off the surface until it is once again taught. What will its height be above the surface? This puzzle, or one very like it, appeared in a students' book on Euclid written in 1702 by the English clergyman, mathematician, and natural philosopher William Whiston (1667–1752). The answer in the form just given is remarkable: about 16 cm. It comes simply from the formula for the circumference of a circle. If the extra radius of the rope is r and Earth's radius is R, then

$$2\pi(R + r) = 2\pi R + 100$$
$$\text{so that } r = 100/2\pi \cong 15.9.$$

Rosamund's bower

A legendary **maze**, located in Woodstock Park, Oxfordshire, whose purported site is marked today by a well and fountain. It was supposedly intended to conceal Rosamund Clifford, the mistress of King Henry II (1133–1189), from the queen, Eleanor of Aquitaine. Legend has it that about 1176, Eleanor managed to solve the maze and confronted Rosamund with the choice of a dagger or poison; she drank the poison and Henry never smiled again. Historically, Henry had imprisoned Eleanor for fomenting rebellion by her sons, and Rosamund was his acknowledged mistress. Rosamund probably spent her last days at a nunnery in Godstow, near Oxford. The legend of the bower dates from the fourteenth century and her murder is a later addition. In the nineteenth century, many puzzle collections had a maze called Rosamund's bower.

rose curve

A curve that has the shape of a flower with petals; it was named *rhodonea* ("rose") by the Italian mathematician Guido Grandi in the 1720s. It is given by the polar equation

$$r = a \sin (n\theta).$$

If n is odd the rose has n petals, if n is even the rose has $2n$ petals, and if $n = 2$ the rose becomes the **quadrifolium**. If n is an *irrational* **number**, then there are an infinite number of petals.

rotation

A **transformation** in which a figure turns through a specific angle about a fixed point, called the center of rotation. The center of rotation may be inside or outside the figure that is being transformed. If the figure is turned anticlockwise, the rotation is considered positive, while a negative rotation turns the figure clockwise.

rotor

A **convex** figure that can be rotated inside a **polygon** (or **polyhedron**) while always touching every side (or face). The least area rotor in a square is the **Reuleaux triangle**. The least area rotor in an equilateral triangle is a lens with two 60° arcs of circles and a radius equal to the triangle's altitude. There exist nonspherical rotors for the tetrahedron, octahedron, and cube, but not for the dodecahedron and icosahedron. See also **curve of constant width**.

roulette

(1) The curve traced by a fixed point on a closed **convex** curve as that curve rolls without slipping along a second curve. (2) A gambling game in which players bet on which slot of a rotating disk a small ball will come to rest in. On August 18, 1913, on an unbiased roulette wheel at Monte Carlo, evens came up 26 times in a row. The probability of this occurring is 1 in 136,823,184.

round

In **topology**, the terms ***circle*** and ***sphere*** refer to topological objects and not geometric ones, so that the surface of an egg shape is a sphere. A *round sphere* is, topologically speaking, not a tautology, but a sphere with constant **curvature**; that is, a sphere in the sense of geometry.

rounding

Replacing a number by another number having fewer **significant digits** or, for integer numbers, fewer value-carrying (nonzero digits). For example, 386.804 may be rounded successively to 386.80, 386.8, 387, 390, and 400. Rounding may be carried out in two ways: by *rounding down,* which is equivalent to truncation, and by *rounding up* the last digit to be retained by one unit. See also **banker's rounding**.

round-off error

The error accumulated during a calculation due to rounding intermediate results. See also **banker's rounding**.

Rubik's cube

A 3 × 3 × 3 cube in which the 26 subcubes on the outside are internally hinged in such a way that rotation (by a quarter turn in either direction or a half turn) is possible in any plane of cubes. Each of the six sides is painted a distinct color, and the goal of the puzzle is to return the cube to a state in which each side has a single color after it has been randomized by repeated rotations. Invented in 1974 by the Hungarian Ernö Rubik, patented in 1975, and put on the market in Hungary in 1977, it went on to sell some 100 million copies worldwide over the next decade. Since there are over 43 million trillion different arrangements of the small cubes, only one of which corresponds to the desired goal, to solve Rubik's cube in a time significantly less than the current age of the universe (let alone the world record, which stands at around 20 seconds) calls for some kind of methodical approach. Algorithms exist for solving a cube from an arbitrary initial position, but they are not necessarily optimal (i.e., requiring a minimum number of turns).[270, 297]

Rucker, Rudy (Rudolf von Bitter) (1946–)

An American mathematician best known for his entertaining popular mathematics, science, and science fiction books, including *Infinity and the Mind,*[275] *The Fourth Dimension,*[273] and *Mind Tools.*[274] Rucker has a doctorate in mathematical logic from Rutgers and teaches at San José State University. His great-great-great-grandfather was the famous German philosopher Georg Hegel.

ruled surface

A surface that is built up from an infinite number of perfectly straight lines. A **cylinder**, for example, is a ruled surface of parallel straight lines. A **cone** is a ruled surface of straight lines that meet at the apex of the cone. Also known as *scrolls*, ruled surfaces have been studied for centuries by geometers such as the Jesuits Roger Boscovich and Andre Tacquet as well as by their famous students, including Gaspar **Monge** and Phillippe de Lahire. Examples that stand out because of both their striking shape and their relative ease of construction include **hyperboloid**s, **helicoid**s and **Möbius band**s. Most ruled surfaces, however, are so complicated that, before the computer age, they were almost impossible to construct.

ruler-and-compass construction

See **Mascheroni construction**.

Russell, Bertrand Arthur William (1872–1970)

A British philosopher, mathematician, and logician who rose to prominence with his first major work, *The Principles of Mathematics* (1902), in which he attempted to remove mathematics from the realm of abstract philosophical notions and to give it a precise scientific framework. Russell then collaborated for eight years with the British philosopher and mathematician Alfred North Whitehead (1861–1947) to produce the monumental work *Principia Mathematica* (3 volumes, 1910–1913). This work showed that mathematics can be stated in terms of the concepts of general **logic**, such as class and membership in a class. It became a masterpiece of rational thought. Russell and Whitehead proved that numbers can be defined as classes of a certain type, and in the process they developed logic concepts and a logic notation that established symbolic logic as an important specialization within the field of philosophy. In his next major work, *The Problems of Philosophy* (1912), Russell borrowed from the fields of sociology, psychology, physics, and mathematics to refute the tenets of idealism, the dominant philosophical school of the period, which held that all objects and experiences are the product of the intellect. Russell, a realist, believed that objects perceived by the senses have an inherent reality independent of the mind.

Russell condemned both sides in World War I, and for his uncompromising stand he was fined, imprisoned, and deprived of his teaching post at Cambridge. In prison he wrote *Introduction to Mathematical Philosophy* (1919). After the war he visited the Soviet Union, and in his book *Practice and Theory of Bolshevism* (1920) he expressed his disappointment with the form of socialism practiced there. He felt that the methods used to achieve a Communist system were intolerable and that the results obtained were not worth the price paid. Russell taught at Beijing University during 1921 and 1922, and in the United States from 1938 to 1944, though he was barred from teaching at the College of the City of New York (now City College of the City University of New York) by the state supreme court because of his attacks on religion and his advocacy of sexual freedom. Russell returned to England in 1944 and was reinstated as a fellow of Trinity College. Although he abandoned pacifism to support the Allied cause in World War II, he became an ardent opponent of nuclear weapons. Russell received the 1950 Nobel Prize for Literature and was cited as "the champion of humanity and freedom of thought." He led a movement in the late 1950s advocating unilateral nuclear disarmament by Britain, and at the age of 89 he was imprisoned after an antinuclear demonstration.[276]

Russell's paradox

A **paradox** uncovered by Bertrand **Russell** in 1901 that forced a reformulation of **set theory**. One version of Russell's paradox, known as the barber paradox, considers a town with a male barber who, every day, shaves every man who doesn't shave himself, and no one else. Does the barber shave himself? The scenario as described requires that the barber shave himself if and only if he does not! Russell's paradox, in its original form considers the set of all sets that aren't members of themselves. Most sets, it would seem, aren't members of themselves–for example, the set of elephants is not an elephant–and so could be said to be "run-of-the-mill." However, some "self-swallowing" sets *do* contain themselves as members, such as the set of all sets, or the set of all things except Julius Caesar, and so on. Clearly, every set is either run-of-the-mill or self-swallowing, and no set can be both. But then, asked Russell, what about the set *S* of all sets that aren't members of themselves? Somehow, *S* is neither a member of itself nor not a member of itself. Russell discovered this strange situation while studying a foundational work in symbolic logic by Gottlob **Frege**. After he described it, set theory had to be reformulated axiomatically in a way that avoided such problems. Russell himself, together with Alfred North Whitehead (1861–1947), developed a comprehensive system of types in *Principia Mathematica.* Although this system does avoid troublesome paradoxes and allows for the construction of all of mathematics, it never became widely accepted. Instead, the most common version of axiomatic set theory in use today is the *Zermelo-Fraenkel set theory,* which avoids the notion of types and restricts the universe of sets to those that can be built up from given sets using certain axioms. Russell's paradox underlies the proof of **Gödel's incompleteness theorem** as well as Alan **Turing**'s proof of the undecidability of the **halting problem**.

Russian multiplication

Multiplication by repeated doubling, also known as *peasant multiplication*. For example to multiply 17 by 13, double the 17 and halve the 13 rounding down to the next whole number where necessary; then add the doubles that correspond to an odd number in the other column.

So	17×13
doubled and halved	34×6
doubled and halved	68×3
doubled and halved	136×1

adding up the numbers in the first column that correspond to an odd number in the second $(17 + 68 + 136) = 221 = 17 \times 13$.

S

Saccheri, Giovanni Girolamo (1667–1733)

A Jesuit priest, philosopher, and mathematician who did early work on **non-Euclidean geometry**, although he didn't see it as such. His *Euclides ab Omni Naevo Vindicatus* (1733) was actually an attempt to prove Euclid's **parallel postulate** but ended up laying the groundwork for both **hyperbolic geometry** and **elliptical geometry**.

saddle

A type of surface that is neither a peak nor a valley but still has a zero gradient. Saddle points are situated such that moving in one direction takes one uphill, while moving in another direction would be downhill. A *saddle function* is a **function** $f(x, y)$ of two vectors x and y (which typically lie in different **vector spaces**) that is **concave** up in x and concave down in y. See also **pseudosphere**.

St. Ives problem

A well-known and simple puzzle in arithmetic set by Mother Goose in the rhyme "As I Was Going to St. Ives":

> As I was going to St. Ives, I met a man with seven wives. Every wife had seven sacks, and every sack had seven cats, every cat had seven kittens. Kittens, cats, sacks, and wives, how many were going to St. Ives?

The various individuals and items make up a **geometric sequence**: 7 wives + 7^2 sacks + 7^3 cats + 7^4 kittens = 7 + 49 + 343 + 2,401. This gives a total of 2,800 (or 2,801 if the narrator also happens to be a wife). It's been said that the real answer to this rhyme is 1, since "I *met* a man with seven wives" means the 2,800 were going in the *opposite* direction!

A similar problem, and solution, is contained in the **Rhind papyrus** by Ahmose, written about 1650 B.C. (some of which is copied from an older document of about 1800 B.C.). Here the geometric series has one more power of 7: 7 houses + 49 cats + 343 mice + 2,401 ears of grain + 16,807 hekats (a measure) of grains, giving a total of 19,607.

St. Petersburg paradox

A strange state of affairs that arises from a game proposed by Nikolaus (I) Bernoulli (see **Bernoulli family**) in 1713. It is named after the fact that a treatise on the paradox was written by Nikolaus' cousin, Daniel, and published (1738) in the *Commentaries of the Imperial Academy of Science of St. Petersburg.* The game goes as follows. You toss a coin. If it shows heads, you win whatever is in the pot and the game is over. If it shows tails, the pot is doubled and you get to toss the coin again. If the coin shows heads on the first toss you win \$2; if it shows tails, you toss again. If the coin now shows heads you win \$4, and so on. After n tosses you get $\$2^n$ if heads appear for the first time. The only catch is you have to pay to play the game. How much should you be willing to pay? Classical decision theory says that you should be willing to pay any amount up to the expected prize, the value of which is obtained by multiplying all the possible prizes by the probability that they are obtained and adding the resulting numbers. The chance of winning \$2 is ½ (heads on the first toss); the chance of winning \$4 is ¼ (tails followed by heads); the chance of winning \$8 is ⅛ (tails followed by tails followed by heads); and so on. Since the expected payoff of each possible consequence is \$1 (\$2 × ½, \$4 × ¼, etc.) and there are an infinite number of them, the total expected payoff is an infinite sum of money. A rational gambler would enter a game if and only if the price of entry was less than the expected value. In the St. Petersburg game, any finite price of entry is smaller than the expected value of the game. Thus, the rational gambler would play no matter how large the entry price was! But there's clearly something wrong with this. Most people would offer between \$5 and \$20 on the grounds that the chance of winning more than \$4 is only 25% and the odds of winning a fortune are very small. And therein lies the paradox: If the expected payoff is infinite, why is no one willing to pay a huge amount to play?

The classical solution to this mystery, provided by Daniel Bernoulli and another Swiss mathematician, Gabriel Cremer, goes beyond probability theory to touch areas of psychology and economics. Bernoulli and Cremer pointed out that a given amount of money isn't always of the same use to its owner. For example, to a millionaire \$1 is nothing, whereas to a beggar it can mean not going hungry. In a similar way, the utility of \$2 million is not twice the utility of \$1 million. Thus, the important quantity in the St. Petersburg game is the *expected utility* of the game (the utility of the prize multiplied by its probability) which is far less than the *expected prize.* This explanation forms the theoretical basis of the insurance business. The existence of a **utility function** means that most people prefer, for example, having $98 in cash to gambling in a lottery where they could win $70 or $130 each with a chance of 50%, even though the lottery has the higher expected prize of $100. The difference of $2 is the premium most of us would be willing to pay for insurance. That many people pay for insurance to avoid any risk, yet at the same time spend money on lottery tickets in order to take a risk of a different kind, is another paradox, which is still waiting to be explained.

salient

At a *salient point,* two branches of a curve meet and stop, and have different **tangent**s. A *salient angle* is an outward-pointing angle of a **polygon**; compare with **reentrant angle**.

salinon

A figure formed from four connected semicircles. The word *salinon* is Greek for "salt cellar," which the figure resembles. In his *Book of Lemmas,* **Archimedes** proved that the salinon has an area equal to the circle having the line segment joining the top and bottom points as its diameter. See also **arbelos**.

Sallows, Lee C. F. (1944–)

A British electronics engineer and puzzle enthusiast at the University of Nijmegen in the Netherlands. Among his many accomplishments in recreational mathematics, he devised the first **self-enumerating sentence** (published in 1982), introduced **alphamagic square**s in 1986, coined the word ***golygon*** in 1990, invented reflexicons (minimal self-enumerating phrases) in 1992, demonstrated the parallelogram theorem for 3 × 3 **magic square**s in 1997, and discovered **geometric magic squares** in 2001.

scalar

A quantity specified by a single number or value (as opposed to a vector, matrix, or array) that contains multiple values. Examples of scalars include mass, volume, and temperature. A *scalar field* is an arrangement of scalar values distributed in a space.

scalene triangle

A **triangle** whose sides are all unequal.

schizophrenic number

An informal name for an **irrational number** that displays such persistent patterns in its decimal expansion, that it

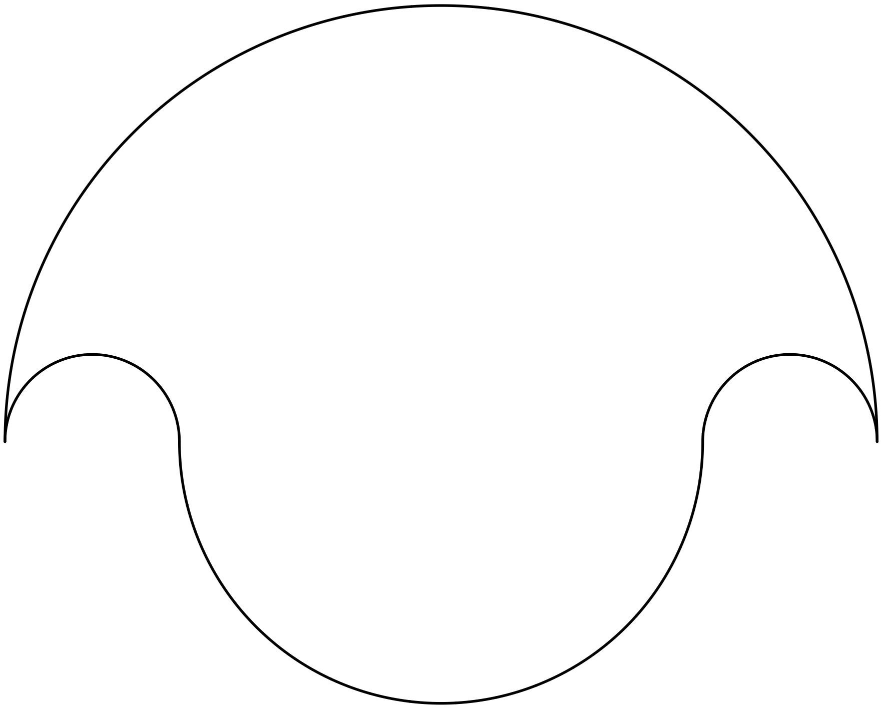

salinon *© Jan Wassenaar, www.2dcurves.com* has the appearance of a **rational number**. A schizophrenic number can be obtained as follows. For any positive integer n let $f(n)$ denote the integer given by the recurrence $f(n) = 10f(n-1) + n$ with the initial value $f(0) = 0$. Thus, $f(1) = 1$, $f(2) = 12$, $f(3) = 123$, and so on. The square roots of $f(n)$ for odd integers n give rise to a curious mixture appearing to be rational for periods, and then disintegrating into irrationality. This is illustrated by the first 500 digits of $\sqrt{f(49)}$:

1111111111111111111111111.1111111111111111111111 0860
555555555555555555555555555555555555555555555 2730541
66666666666666666666666666666666666666666 0296260347
2222222222222222222222222222222222222 0426563940928819
4444444444444444444444444444444 3877555125040117187 4
9999999999999999999999999999 8082496877114863053385 41
66666666666666666666666 59871857386214406386555989 58
33333333333333333333 0843460407627608206940277099609374
99999999999999 06422275875559830666394303215874565 97
222222222 18634920167911808330818 44 . . . .

The repeating strings become progressively shorter and the scrabbled strings become larger until eventually the repeating strings disappear. However, by increasing n we can forestall the disappearance of the repeating strings as long as we like. The repeating digits are always 1, 5, 6, 2, 4, 9, 6, 3, 9, 2, ...

Schläfli, Ludwig (1814–1895)

A German mathematician whose work centered on geometry, arithmetic, and the theory of functions. He made an important contribution to **non-Euclidean geometry** when he proposed that spherical three-dimensional space could be thought of as the surface of a **hypersphere** in Euclidean four-dimensional space. Schläfli started out as a schoolteacher and amateur mathematician. He was also an expert linguist and spoke many languages, including Sanskrit. In 1843 he served as a translator for the great mathematicians Jakob **Steiner**, Karl **Jacobi**, and Peter **Dirichlet** during their visit to Rome and learned a great deal from them. Ten years later he became professor of mathematics at Bern. However, his true importance was only appreciated following the publication of his magnum opus *Theory of Continuous Manifolds* in 1901, several years after his death.

Schläfli symbol

A notation, devised by Ludwig **Schläfli**, which describes the number of edges of each **polygon** meeting at a **vertex** of a regular or semi-regular **tessellation** or solid. For a **Platonic solid**, it is written $\{p, q\}$, where p is the number of sides each face has, and q is the number of faces that touch at each vertex.

schoolgirls problem

A problem in **combinatorics** posed by the Rev. Thomas **Kirkman** in a letter in 1850 following a paper he wrote on the same subject in 1847:[188] A school mistress has fifteen girl pupils and she wishes to take them on a daily walk. The girls are to walk in five rows of three girls each. It is required that no two girls should walk in the same row more than once per week. Can this be done?

In fact, provided n is divisible by 3, we can ask the more general question about n schoolgirls walking for $(n-1)/2$ days so that no girl walks with any other girl in the same triplet more than once. Solutions for $n = 9$, 15, and 27 were given in 1850 and much work was done on the problem thereafter. The general problem of how many triads can be made out of **n** symbols, so that no pair of symbols is comprised more than once among them gave rise to the study of Steiner triple systems. However, Jakob **Steiner** had little to do with them and they should rightfully be named after Kirkman. They are important in the modern theory of **combinatorics**.

Schröder's reversible staircase

A classic example of an **ambiguous figure**, first drawn by Schröder in 1858. Not to be confused with the **Penrose stairway**.

Schubert, Hermann Cäsar Hannibal (1848–1911)

A German mathematician who worked mainly in enumerative geometry–the parts of algebraic geometry that

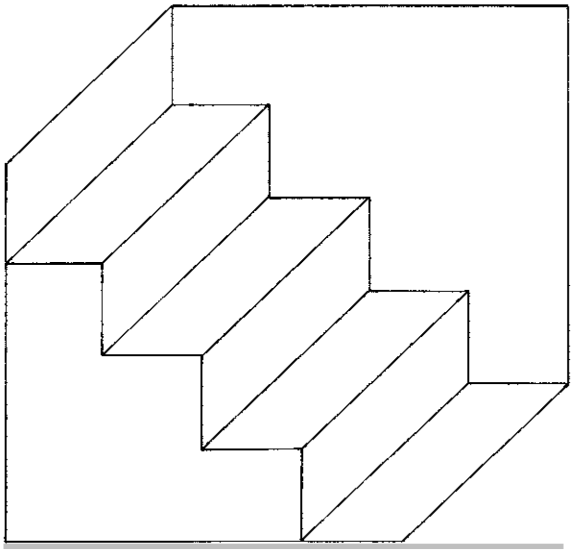

Schröder's reversible staircase

involve a finite number of solutions. He also wrote extensively on recreational math.

Schuh, Frederick (1875–1966)

A Dutch mathematician who wrote many textbooks and a number of books on recreational mathematics, including *The Master Book of Mathematical Recreations* (1943), the English edition of which appeared in 1968 (Dover Publications).[291] Schuh was professor of mathematics at the Technische Hoogeschool at Delft (1907–1909 and 1916–1945) and professor of mathematics at Groningen (1909–1916).

scientific notation

A number of the form $a \times 10^n$, where n is an integer, positive or negative, and a is a **real number** larger than or equal to 1, but less than 10. Scientific notation provides a compact way of writing **large number**s.

scintillating grid illusion

See **Hermann grid illusion**.

score

A group of 20 items. The word comes from the Old Norse *skor* for a heavy mark used to indicate a string of 20 smaller marks; *skor,* in turn, is descended from the Indo-European *sker,* for "cutting" or "slicing." From about 1400, *score* was also the word for a record or an amount due–the total of the score marks on a tally. It became a common word for the total of a tradesman's or innkeeper's account. So, *to settle the score* originally meant just to pay one's bill. But it acquired the figurative sense of taking revenge on somebody, and that's usually what is meant by the expression now. The more general meaning of score, as a tally, is used daily when the results of sports competitions are reported.

search space

A characterization of every possible solution to a problem instance.

secant

A straight line that meets a curve in two or more points.

second

One-sixtieth of a **minute** in both time and angle. Literally, the second division of the hour or the circle, the minute being the first; from the Latin *secundus.* In the System International d'Unites (SI units) one second is defined as the duration of 9,192,631,770 periods of radiation corresponding to the transition between two hyperfine levels of cesium-133 in a ground state at a temperature of 0°K (Kelvin).

secretary problem

See **sultan's dowry**.

sector

Part of a **circle** bounded by two radii and the included **arc**.

segment

Part of a **circle** bounded by a **chord** and the **arc** subtending the chord.

self-enumerating sentence

Also known as an *autogram,* a **self-referential sentence** whose text consists solely of the enumeration of its letter content. The answer to the question whether such a sentence exists in English was given by Lee **Sallows**, and was first published in *Scientific American* in January 1982:

> Only the fool would take trouble to verify that his sentence was composed of ten a's, three b's, four c's, four d's, forty-six e's, sixteen f's, four g's, thirteen h's, fifteen i's, two k's, nine l's, four m's, twenty-five n's, twenty-four o's, five p's, sixteen r's, forty-one s's, thirty-seven t's, ten u's, eight v's, eight w's, four x's, eleven y's, twenty-seven commas, twenty-three apostrophes, seven hyphens and, last but not least, a single !

This remarkable sentence counts not only its own letters, but also its punctuation marks, although it fails to enumerate three letters of the alphabet (j, q, and z). Sallows went on to devise a "pangram machine"–a computer purposely built to search for sentences of this type. Among its many successes is:

> This pangram lists four a's, one b, one c, two d's, twenty-nine e's, eight f's, three g's, five h's, eleven i's, one j, one k, three l's, two m's, twenty-two n's, fifteen o's, two p's, one q, seven r's, twenty-six s's, nineteen t's, four u's, five v's, nine w's, two x's, four y's, and one z.

self-intersecting

A *self-intersecting polygon* is a **polygon** with edges that cross other edges. A *self-intersecting polyhedron* is a **polyhedron** with faces that cross other faces.

self-organization

A process in a **complex system** whereby new emergent structures, patterns, and properties arise without being externally imposed on the system. Not controlled by a centralized, hierarchical "command and control" center, self-organization is usually distributed throughout a system. It requires a complex, **nonlinear system** under appropriate conditions, variously described as "far-from-equilibrium," critical values of control parameters leading to "bifurcation," or the "edge of chaos." First investigated in the 1960s in physical systems by Ilya Prigogine and his followers, as well as the Synergetics School founded by Hermann Haken, self-organization is now studied mainly through computer simulations (see **cellular automaton**), Boolean networks, and other phenomena of **artificial life**. However, self-organization is now recognized as a crucial way of understanding emergent, collective behavior in a large variety of systems including the economy, the brain and nervous system, the immune system, and ecosystems. The buildup of system order via self-organization is now conceived as a primary tendency of complex systems in contrast to the past emphasis on the degrading of order in association with the principle of entropy (Second Law of Thermodynamics). However, rather than denying entropy, self-organization can be understood as a way that entropy increases in complex, nonlinear systems.

self-organized criticality

A mathematical theory that describes how systems composed of many interacting parts can tune themselves toward dynamical behavior that is critical in the sense that it is neither stable nor unstable but at a region near a **phase transition**. See also **edge of chaos** and **self-organization**.

self-referential sentence

A sentence that refers to itself and nothing else. Here are some examples:

> This statement is short.
>
> This sentence has five words.
>
> The last word of this sentence is "wrong."
>
> "Pentasyllabic" is pentasyllabic.
>
> How long is the answer to this question? Ten letters.

Some self-referential statements take the form of jokes. For example:

> The two rules for success are: Never tell them everything you know.
>
> There are three kinds of people in the world: those who can count and those who can't.

Finally, some take the form of maxims, as in this case by Thomas Macaulay (1800–1859):

> Nothing is so useless as a general maxim.

See also **Hofstadter's law**.

self-similarity

The property an object has when a part of itself looks the same or similar to the whole. Many objects in the real world, such as coastlines, are statistically self-similar: parts of them show the same statistical properties at many scales. Self-similarity is a defining characteristic of **fractal**s.

semigroup

A **set** together with a method of combining elements, such as addition or multiplication, to get new ones, which satisfies only some of the properties required to get a **group**. In particular, a semigroup need not have an **identity** element and elements need not have **inverse**s.

semi-magic square

A square array of n numbers such that sum of the n numbers in any row or column is a constant (known as the magic sum). See also **magic square**.

semi-regular polyhedron

A **polyhedron** that consists of two or more types of **regular polygon**s, all of whose vertices are identical. This category includes the **Archimedean solid**s, **prism**s and **antiprism**s, and the nonconvex uniform polyhedra (see **nonconvex uniform polyhedron**).

Senet

A popular two-player board game in ancient Egypt, enjoyed by both commoners and nobility, that may be an ancestor of modern **backgammon**. The rules are not known, though about 40 sets have been found in tombs, some in very good condition, together with paintings of games on tomb walls, dating back to the reign of Hesy (c. 2686–2613 B.C.). Senet, or the "game of passing," was played on a rectangular board consisting of three rows of 10 squares called "houses" that represented good or bad fortune. The board could be a grid drawn on a smooth surface or an elaborate box of wood and other precious materials. A perfectly preserved traveling version of Senet was found in Tutankhamen's tomb. The pieces, called *ibau* ("dancers" in Egyptian), varied in number from five to ten per player–five and seven being commonest. Cone-shaped pieces were pitted against reel-shaped pieces. The object was to get one's pieces on the board, then around the board in an S-shaped pattern, and finally off again at the far end. Strategy was mixed with chance (as it is in backgammon), introduced by the throw of four, two-sided sticks (as depicted in the Hesy painting) or, in later times, of knucklebones. Later depictions of the game, in the New Kingdom period, often showed just one player in competition–the opponent being a spirit from the afterlife. This has been interpreted as a change

Senet A modern version of the ancient Egyptian board game. *Fundex Games Ltd.*

in the significance of Senet, from a simple amusement to a symbolic representation of the deceased's journey through the underworld. See also **nine men's morris**.

sensitivity

The tendency of a system, which may be chaotic (see **chaos**) to change dramatically with only small **perturbation**s.

sequence

An ordered list of values that may be finite or infinite in length. Among the different types of sequence are **arithmetic sequence**s, **geometric sequence**s, and the **harmonic sequence**.

series

A sum of all or some of the terms of a **sequence**. A series may or may not *converge* to a particular value as more and more terms are included. A series is said to be *absolutely convergent* if the sum of the absolute values of the terms converges; in this case the series converges no matter how the terms of the sum are arranged. Series that are *conditionally convergent* only converge for some arrangements of the terms and, even then, converge to different values for different arrangements.

serpentine

A curve named and studied by Isaac **Newton** in 1701 and contained in his classification of **cubic curve**s. It had been studied earlier by **de L'Hôpital** and Christiaan **Huygens** in 1692. The curve is given by the Cartesian equation

$$y(x) = abx/(x^2 - a^2).$$

set

A finite or infinite collection of objects known as *elements*. Sets are one of the most basic and important concepts in mathematics. An example of a finite set is the set of whole numbers from 1 to 58; an example of an infinite set is the set of all the **rational number**s. Two sets are equal if, and only if, they contain the same objects. Standard notation uses braces around the list of elements, as in: {red, green, blue}. If A and B are two sets and every x in A is also contained in B, then A is said to be a *subset* of B. Every set has as subsets itself, known as the *improper subset*, and the **empty set**. The *union* of a collection of sets $S = \{S_1, S_2, S_3, \ldots\}$ is the set of all elements contained in at least one of the sets $S_1, S_2, S_3, \ldots$. The *intersection* of a collection of sets $T = \{T_1, T_2, T_3, \ldots\}$ is the set of all elements contained in all of the sets. The union and intersection of sets, say A_1 and A_2, is denoted $A_1 \cup A_2$ and $A_1 \cap A_2$, respectively. The set of all subsets of X is called its *power set* and is denoted 2^X or $P(X)$. See also **set theory**, **Venn diagram**, and **Russell's paradox**.

set of all sets

See **Russell's paradox**.

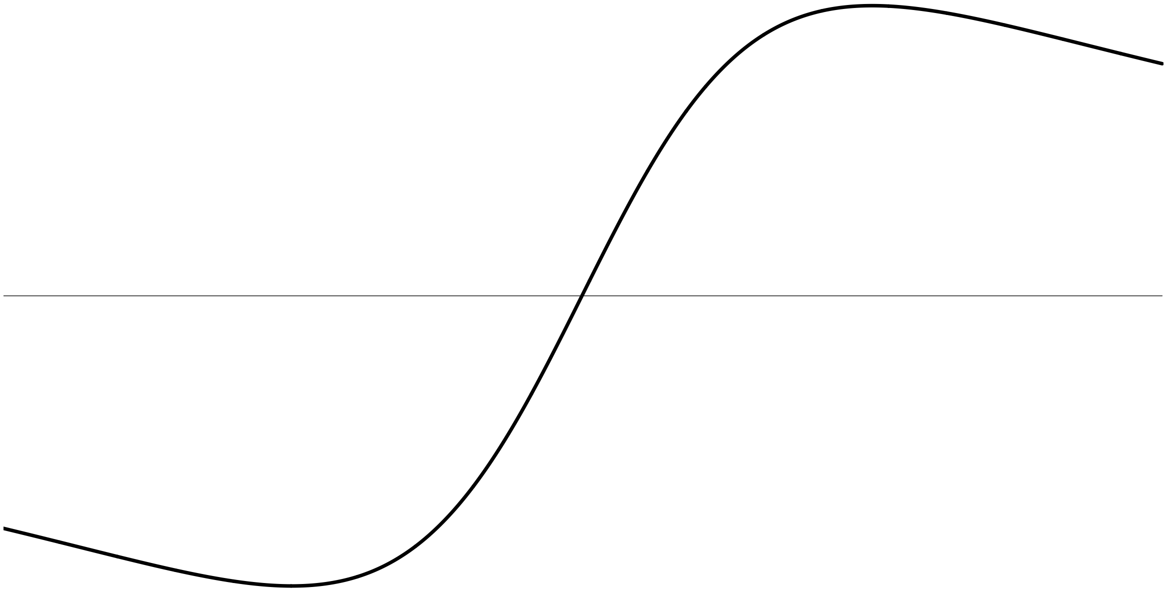

serpentine The serpentine curve. *© Jan Wassenaar, www.2dcurves.com*

set theory

A branch of **mathematics** created by Georg **Cantor** at the end of the nineteenth century. Initially controversial, set theory has come to play a foundational role in modern mathematics, in that it is used to justify assumptions made concerning the existence of mathematical objects (such as numbers or functions) and their properties. Formal versions of set theory also figure centrally in specifying a theoretical ideal of mathematical rigor in proofs. Cantor's basic discovery was that if we define two sets A and B to have the same number of members (the same *cardinality*), then there is a way of pairing off members of A exhaustively with members of B. The appearance around the turn of the century of set-theoretical paradoxes, such as **Russell's paradox**, prompted the formulation in 1908 by Ernst Zermelo of an axiomatic theory of sets. The axioms for set theory now most often studied and used are those called the *Zermelo-Fraenkel axioms,* usually together with the **axiom of choice**. The Zermelo-Fraenkel axioms are commonly abbreviated to ZF, or ZFC if the axiom of choice is included. An important feature of ZFC is that every object that it deals with is a set. In particular, every element of a set is itself a set. Other familiar mathematical objects, such as numbers, must be subsequently defined in terms of sets.

seven

A lucky number in the eyes of many people and one that has been given much spiritual significance. The early religious and cultural use of the seven-day week almost certainly stems from the fact that the Moon goes through its four phases in a bit over 28 days, which divides nicely into seven days per phase. There are seven moving objects in the sky visible to the naked eye (the Sun, Moon, Mercury, Venus, Mars, Jupiter, and Saturn), seven seas, seven orders of architecture, seven deadly sins, seven liberal arts and sciences, and seven dwarfs. The seventh son of a seventh son is supposed to be born gifted (Donny Osmond was such a person). In the Bible, there were seven years of famine and seven years of plenty, and seven years were taken to construct King Solomon's Temple. The Pythagoreans were especially intrigued by the number as it is the sum of three and four, which are the number of sides of a triangle and a square–shapes of enormous importance to the sect. These links with Solomon's Temple and the Pythagoreans help explain the importance of seven in freemasonry. Seven is the smallest positive integer whose reciprocal has a pattern of more than one repeating digit: $1/7 = 0.142857142857\ldots$ and is the smallest number for which the digit sequence of $1/n$ is of length $n - 1$ (the longest such a sequence can be). The next such numbers are 17, 19, 23, 29, 47, 59, 61, 97, 109, 113, ... Other curios: the citrus soda 7-UP, created in 1929, was so called because the original containers were 7 ounces and "up" was the direction of the bubbles, and seven is the maximum number of times you can fold any sheet of paper (try it!).

seventeen

The number most often picked in response to the request "Pick a random number from 1 to 20." Seventeen is a *Fermat prime* (a **prime number** of the form $2^{2^n} + 1$, where n is a positive integer), the exponent of a *Mersenne prime* (a prime p for which $2^p - 1$ is prime), and the only prime that is the sum of four consecutive primes $(2 + 3 + 5 + 7)$. Seventeen is also the smallest number for which the sum of the digits of its cube is equal to the number: $17^3 = 4913$, $4 + 9 + 1 + 3 = 17$, and the smallest number that can be written as $a^2 + b^3$ in two different ways: $17 = 3^2 + 2^3 = 4^2 + 1^3$. The pair (8, 9), whose sum is 17, is the only pair of consecutive numbers where one is a square and the other is a cube (a result proved by Leonhard **Euler**.) There are 17 planar crystallographic groups, called **wallpaper groups**. The minimum number of faces on a convex **polyhedron** that has only one stable face is 17. (A stable face is one that the figure can rest on without falling over; most polygons have more than one such face.) Seventeen is also the answer to the following problem: At a party where any two people have previously met each other in one of three other places, what is the least number of people who must be at the party to guarantee that there is at least one group of three people who have met each other before in the same place?

sexagesimal

Of, relating to, or based on the number 60. Sexagesimal refers especially to the **number system** with base 60. The Babylonians began using such a scheme around the beginning of the second millennium B.C. in what was the first example of a **place-value system**. Our degree of 60 minutes, minute of 60 seconds (in both time and angle measure), and hour of 60 minutes hark back to this ancient method of numeration. Why the Babylonians counted using sexagesimal isn't known, but 60 certainly has more factors than any other number of comparable size.

Shannon, Claude Elwood (1916–2001)

An American mathematician and a pioneer of **information theory**. Shannon was the first to realize that any sort of message can be transmitted as a series of 0's and 1's, regardless of whether it consists of words, numbers, pictures, or sound. In his master's thesis, he explained how electrical switches could represent **binary** digits–a 1 when the switch is on and 0 when it is off. He also used **Boolean algebra** to show that complex operations could be carried out automatically on these electrical circuits, thus manipulating the data they were storing. It was in one of his papers, "A Mathematical Theory of Communication" published in 1948, that the word "bit" (short for binary digit) was used for the first time. In fact the framework and terminology for information theory he developed remains standard today. Shannon was driven by curiosity, and in his own words he "just wondered how things were put together." Among his inventions were rocket-powered Frisbees, motorized Pogo sticks, a device that could solve the **Rubik's cube** puzzle, and a juggling machine. (He could ride a unicycle while juggling three balls.) He was involved in pioneering **artificial intelligence** research, which included building the electromechanical mouse called "Theseus" that could navigate a metal **maze** using magnetic signals. Shannon also built a **chess**-playing computer, many years before IBM's Deep Blue, that played well against the world champion of the time, Mikhail Botvinnik (the computer lost only after 42 moves).

shell curve

See **Dürer's shell curve**.

shuffle

How many shuffles does it take to **randomize** a deck of cards–in other words, to mix up the cards about as thoroughly as dropping them all on the table and stirring them around for several minutes (the author's usual method). The answer depends on the kind of shuffle used. The beginner's *overhand shuffle*, for example, is a really bad way to mix cards: about 2,500 such shuffles are need to randomize a deck of 52 cards. A magician's *perfect shuffle*, on the other hand, in which the cards are cut exactly in half and then perfectly interlaced, *never* produces randomization (see below). One of the most effective ways to get a random deck is the *riffle shuffle* in which the deck is cut in half and imperfectly interlaced by dropping cards one by one from either half of the deck with a probability proportional to the current sizes of the deck splits. In 1992 Persi Diaconis (then at Harvard) and David Bayer demonstrated that, starting with a completely ordered deck, it takes *seven* riffle shuffles to produce randomization.[27] Any more than this and there's no significant increase in the randomness; any less and the shuffle is far from random. In fact, not only are five or six riffles not enough to randomize, there are some configurations of cards that are impossible to reach in this number of shuffles! To understand this, suppose the starting order of the cards is marked 1 to 52, top to bottom. After one shuffle, only configurations with two or fewer rising sequences are possible. A rising sequence is a maximal increasing sequential ordering of cards that appear in the deck (with other cards possibly interspersed) as it is run through from top to bottom. For instance, in an eight-card deck, 12345678 is the ordered deck and it has one rising sequence. After one shuffle, 16237845 is a possible configuration, and there are two rising sequences (the underlined numerals form one, the nonunderlined numerals form the other). Clearly the rising sequences are formed when the deck is cut before the cards are interleaved in the shuffle. After two shuffles, there can be at most four rising sequences, since each of the two rising sequences from the first shuffle has a chance of being cut in the second. This pattern continues: the number of rising sequences can at most double during each shuffle. After five shuffles, there are at most 32 rising sequences. But the *reversed* deck, numbered 52 down to 1, has 52 rising sequences. Thus, this is one (of many) arrangements that are unattainable in five riffle shuffles. Interestingly, Diaconis and other researchers have also found that decks can undergo sudden changes in their degree of randomness; after six riffle shuffles, a deck is still visibly ordered, but this order vanishes one shuffle later.

Perfect shuffles do the exact opposite of randomizing: they preserve order at every stage. There are two kinds of perfect shuffles. The *out-shuffle* is one in which the top card stays on top; the *in-shuffle* is one in which the top card moves to the second position of the deck. Amazingly, eight perfect out-shuffles restore the deck to its original order! Magicians use combinations of out and in shuffles to perform a variety of baffling tricks and to control the position of any given card in a deck. How could you make the top card (call it position 0) go to position *n*? Easy: write *n* in **binary** (base 2), read the 0's and 1's from left to right, perform an out-shuffle for a 0 and an in-shuffle for a 1, and, as if by magic, the top card will have materialized at position *n*.

shunting puzzles

Railroad modelers, especially those with limited space available for their layouts, often enjoy setting up track that allows interesting shunting problems to be tried out and solved. The most famous mathematical puzzle of this type, called the *railroad shunting puzzle,* comes in a number of variations, but basically the problem is that there are two trains (*A* and *B* in the diagram) facing each other on a single line with just one short siding, which will only hold one item of rolling stock at a time. In order to enable the two trains to pass each other and to continue their journey, a series of movements using the siding is required. First time around it's quite a brain-teaser, which probably explains why railroad companies all over the world took the more costly but easier way out and built passing sidings!

Siegel's paradox

If a fixed fraction *x* of a given amount of money *P* is lost, and then the same fraction *x* of the remaining amount is gained, the result is less than the original and equal to the final amount if a fraction *x* is first gained, then lost.

Sierpinski, Waclaw Franciszek (1882–1969)

A Polish mathematician who made outstanding contributions to **set theory**, which included research on the **axiom of choice** and the **continuum hypothesis**, **number theory**, and **topology**. Two well-known fractals, the **Sierpinski carpet** and the **Sierpinski gasket**, are named after him.

Sierpinski carpet

A fractal, named after Waclaw **Sierpinski**, that is derived from a square by cutting it into nine equal squares with a 3 × 3 grid, removing the central piece, and then applying the same procedure ad infinitum to the remaining eight squares. It is one of two generalizations of the Cantor set to two dimensions; the other is the **Cantor dust**. The carpet's **Hausdorff dimension** is log 8/log 3 = 1.8928. . . .

Sierpinski gasket

A fractal, also known as the *Sierpinski triangle* or *Sierpinski sieve* after its inventor Waclaw **Sierpinski**. It is produced by the following set of rules: (1) start with any triangle in a plane; (2) shrink the triangle by ½, make three copies, and translate them so that each triangle touches the two other triangles at a corner; (3) repeat step 2 ad infinitum. The gasket can also be made by starting with **Pascal's**

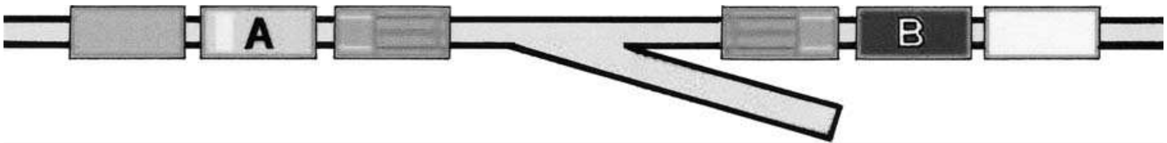

shunting puzzles How can two trains, traveling in opposite directions on a single track, get past each other by using a siding that can only accommodate a single item of rolling stock at a time?

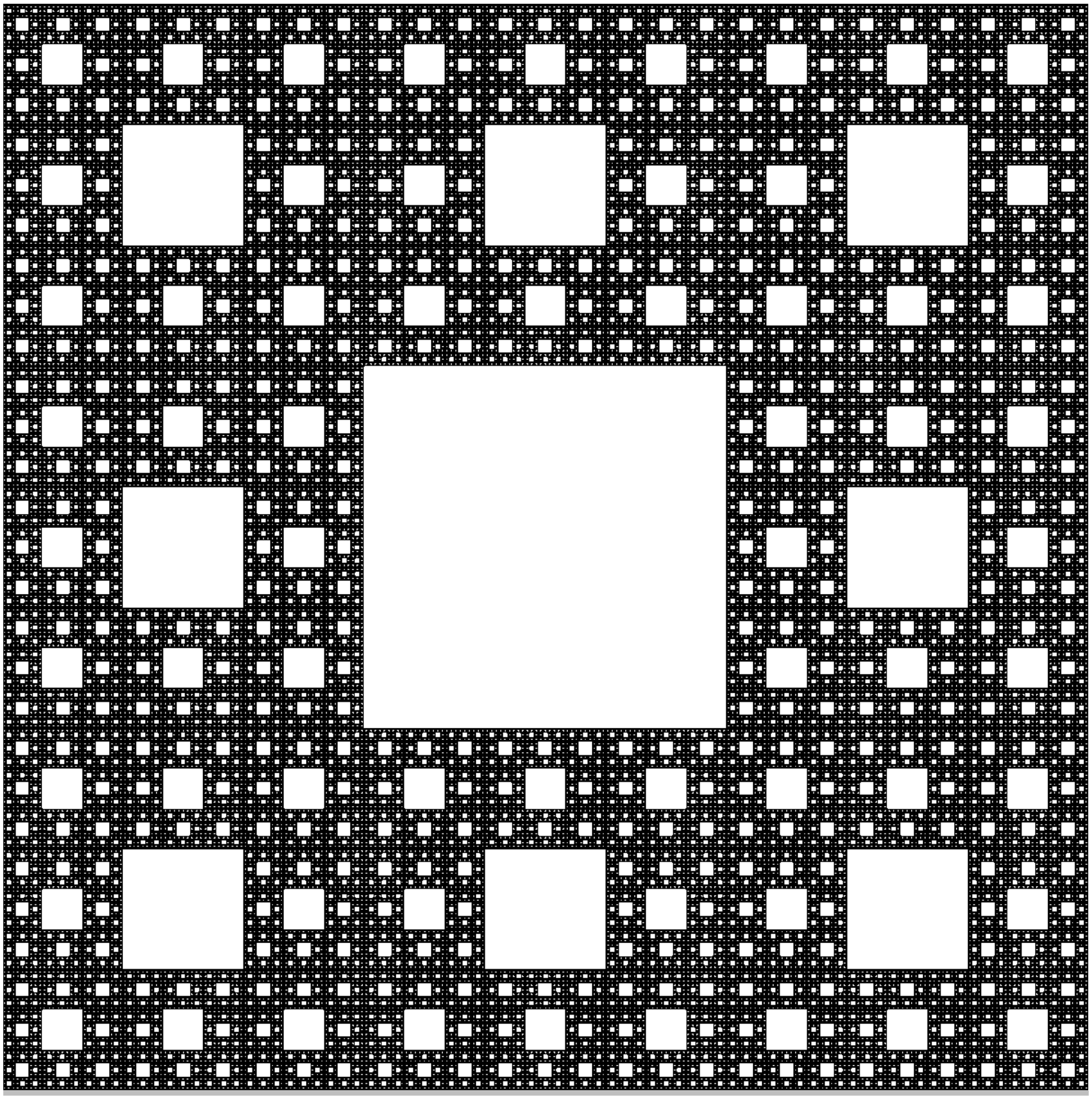

Sierpinski carpet

triangle, then coloring the even numbers white and the odd numbers black. Most curiously, it can be generated by a game of chance. Begin with three points, labeled 1, 2, and 3, and any starting point, *S*. Then select randomly 1, 2, or 3, using a die or some other method. Each random number defines a new point halfway between the latest point and the labeled point that the random number indicates. When the game has gone on long enough the pattern produced is the Sierpinski gasket. The gasket has a **Hausdorff dimension** of log 3/log 2 = 1.585 . . . , which follows from the fact that it is a union of three copies of itself, each scaled by a factor of ½. Adding rounded corners to the defining curve gives a nonintersecting curve that traverses the gasket from one corner to another and which Benoit **Mandelbrot** called the *Sierpinski arrowhead.*

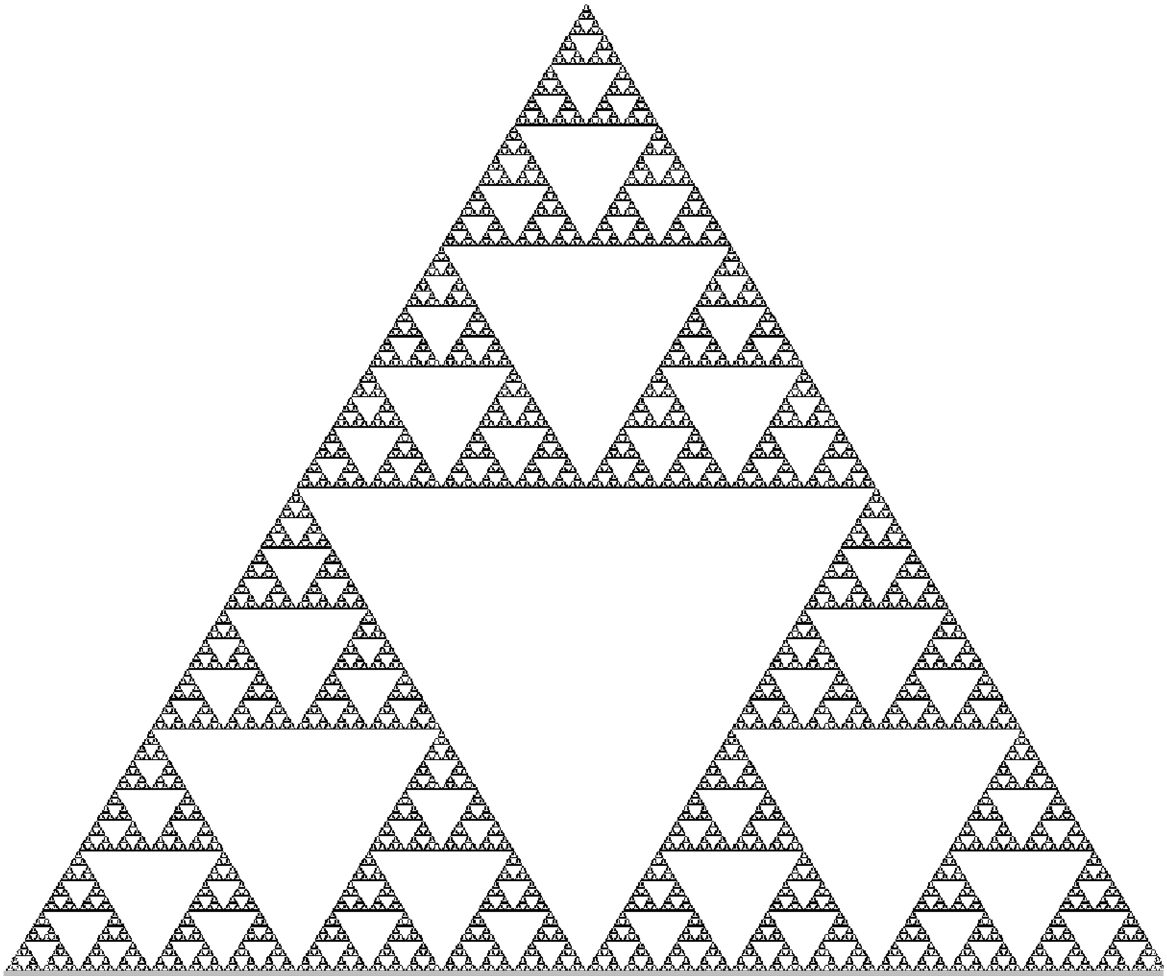

Sierpinski gasket

Sierpinski number

A positive, odd integer *k* such that *k* times $2^n + 1$ is *never* a **prime number** for any value of *n*. In 1960 Waclaw **Sierpinski** showed that there were infinitely many such numbers (though he didn't give a specific example). This is a strange result. Why should it be that while the vast majority of expressions of the form *m* times $2^n + 1$ eventually produce a prime, some don't? For now, mathematicians are focused on a more manageable problem posed by Sierpinski: What is the smallest Sierpinski number? In 1962, John Selfridge discovered the smallest known Sierpinski number, $k = 78{,}557$. The next largest is 271,129. Is there a smaller Sierpinski number? No one yet knows. However, to establish that 78,557 is really the smallest, it would be sufficient to find a prime of the form $k(2^n + 1)$ for every value of *k* less than 78,557. In early 2001, there were only 17 candidate values of *k* left to check: 4,847; 5,359; 10,223; 19,249; 21,181; 22,699; 24,737; 27,653; 28,433; 33,661; 44,131; 46,157; 54,767; 55,459; 65,567; 67,607; and 69,109. In March 2002, Louis Helm of the University of Michigan and David Norris of the University of Illinois started a project called "Seventeen or Bust," the goal of which is to harness the computing power of a worldwide network of hundreds of personal computers to check for primes among the remaining candidates. The team's effort have so far eliminated six candidates–5,359; 44,131; 46,157; 54,767; 65,567; and 69,109. Despite this encouraging start, it may take as long as a decade, with many additional participants, to check the eleven remaining candidates.

sieve of Eratosthenes

The most efficient way to find all of the smallest **prime numbers**. First described by **Eratosthenes of Cyrene**, it involves making a list of all the integers less than or equal to *n* (and greater than one), then striking out the multiples of all primes less than or equal to the square root of *n*. The numbers that are left are the primes. For example, to find all the primes less than or equal to 30, we list the numbers from 2 to 30:

2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30

The first number, 2, is prime, so we keep it and strike out all of its multiples, leaving

2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29

The next number left, 3, is prime, so again we retain it and delete all of its multiples, leaving

2, 3, 5, 7, 11, 13, 17, 19, 23, 25, 29

Now we do the same thing for 5, another prime, to give

2, 3, 5, 7, 11, 13, 17, 19, 23, 29

The next number, 7, is larger than the square root of 30, so all of the numbers left are primes.

significant digits

The digits that define a numerical value. The significant digits of a given number begin with the first nonzero integer digit or, if this number is less than unity, with the first (zero or nonzero) decimal digit. They end with the final (zero or nonzero) decimal digit; the final zero or zeros of an integer may or may not be significant.

similar

Having the same shape but not necessarily the same size. Two triangles are similar if they have equal angles and their corresponding sides, say, a_1, b_1, c_1 and a_2, b_2, c_2, have a common ratio, r: $a_1/a_2 = b_1/b_2 = c_1/c_2$. In general, a *similarity* is a transformation under which the distance between any corresponding pair of points changes by the same factor.

simple group

A **group** that has no nontrivial proper normal **subgroup**s. Simple groups are important because they can be thought of as the blocks out of which other groups can be built. Much activity has been expended in the classification of all finite simple groups.

simplex

The *n*-dimensional generalization of the **triangle** and the **tetrahedron**; in other words, a **polytope** in *n* dimensions with $n + 1$ vertices.

simply connected

The condition of a geometrical object if it consists of one piece and doesn't have any holes or "handles." For example, a line, disk, and sphere are simply connected, but a torus (doughnut) and teapot are not. See also **connected**.

simulation

Experimentation in the space of theories, or a combination of experimentation and theorization. Some numerical simulations are programs that represent a model for how nature works. Usually, the outcome of a simulation is as much a surprise as the outcome of a natural event, due to the richness and uncertainty of computation.

sine

The **trigonometric function** of an angle of a right-angled triangle other than the right angle that is equal to the length of the side adjacent to the angle divided by the length of the hypotenuse. The curve of $y = \sin x$ is called the sine curve, or *sinusoid.*

Singmaster, David

A professor at the school of computing, information systems, and mathematics at South Bank University, London, who is one of the world's leading compilers and historians of mathematical puzzles.

singularity

(1) A point at which the **derivative** does not exist for a given **function** but every neighborhood of which contains points for which the derivative exists. (2) A point in **space-time** at which gravitational forces cause matter to have infinite density and infinitesimal volume, and space and time to become infinitely distorted.

six

The smallest **perfect number**, the number of faces of a **cube**, and the number of sides of a **hexagon**. There are six players on a volleyball team, six kinds of chessmen, and six types of quark (not including antiquarks). A touchdown in American football earns six points and a hit across the boundary rope of a cricket field without bouncing scores six runs. Long ago people indicated a number by pointing to a part of their body; this is echoed in the New Guinea word for six, which is the same as that for "wrist."

six circles theorem

Given a triangle and a circle inside it that touches two of the triangle's sides, draw a second circle touching another two sides and touching the first circle. Draw a third circle touching two sides and the second circle, and so on. This chain ends with the sixth circle, which will touch the first.

sixty

See **sexagesimal**.

skeletal division

A long division in which most or all of the digits are replaced by symbols (usually asterisks) to form a **cryptarithm**.

skew lines

Also known as *crossing lines,* lines that lie in different planes and do not intersect one another.

Skewes' number

A famous **large number**, commonly given as $10^{10^{10^{34}}}$, that was first derived in 1933 by the South African mathematician Samuel Skewes in a proof involving **prime number**s.[299] G. H. **Hardy** once described Skewes' number as "the largest number which has ever served any definite purpose in mathematics," though it has long since lost that distinction. Skewes' numbers—there are actually two of them—came about from a study of the frequency with which **prime number**s occur. Gauss's well-known estimate of the number of prime numbers less than or equal to n, pi(n), is the integral from $u = 0$ to $u = n$ of $1/(\log u)$; this integral is called Li(n). In 1914 the English mathematician John Littlewood proved that pi(x) – Li(x) assumes both positive and negative values infinitely often. For all values of n up to 10^{22}, which is as far as computations have gone so far, Li(n) has turned out to be an overestimate. But Littlewood's result showed that above some value of n it becomes an underestimate, then at an even higher value of n it becomes an overestimate again, and so on. This is where Skewes' number comes in. Skewes showed that, if the **Riemann hypothesis** is true, the first crossing can't be greater than $e^{e^{e^{79}}}$. This is called the first or *Riemann true Skewes' number.* Converted to base 10, the value can be approximated as $10^{10^{10^{34}}}$, or more accurately as $10^{10^{8.852142 \times 10^{33}}}$ or $10^{10^{8852142197543270606106100452735038.55}}$. In 1987, the Dutch mathematician Herman te Riele[330] reduced dramatically the upper bound of the first crossing to $e^{e^{27/4}}$, or approximately 8.185×10^{370}, while John **Conway** and Richard Guy[68] have made the contradictory claim that the lower bound is 10^{1167}. In any event, Skewes' number is now only of historical interest. Skewes also defined the limit if the Riemann hypothesis is false: $10^{10^{10^{1000}}}$. This is known as the second Skewes' number.

slide rule

A calculating device consisting of two sliding logarithmic scales.

sliding-piece puzzle

A type of *sequential-movement puzzle,* within the larger category of **mechanical puzzle**s, that involves sliding one piece at a time into a single vacant opening in order to advance toward the solution—a certain orderly arrangement of the pieces. The best known is Loyd's **Fifteen Puzzle**.

Slocum, Jerry

An American historian of and writer on **mechanical puzzles**.

slope

"Rise over run." For a straight line in the plane, the *slope* is the **tangent** of the angle it forms with the positive x-axis. For a curve, the slope is, by definition, the slope of the *tangent* line. Therefore, if the slope is constant, a line is straight.

Slothouber-Graatsma puzzle

A **packing** puzzle in which six $1 \times 2 \times 2$ blocks and three $1 \times 1 \times 1$ blocks must be fitted together to make a $3 \times 3 \times 3$ cube. There is only one solution. A similar but much more difficult puzzle, named after its inventor, John **Conway**, calls for packing three $1 \times 1 \times 3$ blocks, one $1 \times 2 \times 2$ block, one $2 \times 2 \times 2$ block, and thirteen $1 \times 2 \times 4$ blocks into a $5 \times 5 \times 5$ box.

Smith number

A **composite number**, the sum of whose digits equals the sum of the digits of its prime factors (its **factor**s that are **prime number**s). The name stems from a phone call in 1984 by the mathematician Albert Wilansky to his brother-in-law Smith, during which Wilansky noticed that the phone number, 493-7775, obeyed the condition just mentioned. Specifically:

$$4{,}937{,}775 = 3 \times 5 \times 5 \times 65{,}837$$
$$4 + 9 + 3 + 7 + 7 + 7 + 5 = 3 + 5 + 5 + 6 + 5 + 8 + 3 + 7$$

Trivially, all **prime numbers** have this property, so they are excluded. The first few Smith numbers are: 4, 22, 27, 58, 85, 94, 121, 166, 202, 265, 274, 319, 346, ... In 1987, Wayne McDaniel proved that there are infinitely many Smiths.

smooth

(1) Infinitely differentiable; possessing infinitely many **derivative**s. For example, sin(x) is a smooth function, while $|x|^3$ is not. More complicated mathematical objects such as **manifold**s are called smooth if they are defined or described by smooth functions. (2) Continuously differentiable (see **continuity**); possessing a continuous **tangent** or derivative.

Smullyan, Raymond (1919–)

An American mathematical logician, puzzle-maker, and magician, who has taught in various colleges but is best known for his books on recreational mathematics. Among these are *What Is the Name of This Book?, The Lady or the Tiger, The Tao is Silent,* and *This Book Needs No Title: A Budget of Living Paradoxes.*[304–311] Smullyan is an inventive maker of logic paradoxes and a pioneer of chess problems that involve "retrograde analysis," in which the object is to deduce the past history of a game from some given present position.

snow

It is often said that no two snowflakes are alike. While this is hard to prove, individual samples can be captured on a chilled glass microscope slide and preserved with artist's spray fixative. All are six-sided and the more ornate kind, called *dendritic snowflakes,* form when the air temperature is between −12°C and −16°C (10°F and 3°F). Typical snowflakes fall at a rate of a meter or two per second; assuming 1.5 m/s and a cloud base of 3,000 m (roughly the height of nimbostratus clouds) gives a descent time of 20 minutes. One of the great urban legends is that the Inuit have *n* words for "snow," where *n* is a large number. This story may have started in 1911 when anthropologist Franz Boaz casually mentioned that the Inuit–he called them "Eskimos," using the derogatory term of a tribe to the south of them for eaters of raw meat–had four different words for snow. With each succeeding reference in textbooks and the popular press the number grew to as many as 400 words. A problem with trying to pin down exactly how many Inuit words there are for snow and/or ice, or for anything else, is that the various dialects of Inuit are polysynthetic, which means that words can effectively be made up on the spot by concatenating various particles to the root word. For example, the suffix *-tluk,* for "bad," might be added to *kaniktshaq,* for "snow," to give *kaniktshartluk,* "bad snow." This can give rise to any number of snow terms, from *akelrorak* ("newly drifting snow") to *mitailak* ("soft snow over an opening in an ice floe").

snowball prime

Also known as a *right-truncatable prime,* a **prime number** whose digits can be chopped off, one by one, from the right-hand side, yet still leave a prime number. This means that even if you stop writing before you finish the number, you will still have written a prime. The largest snowball prime is 73,939,133 (7, 73, 739, . . . , 73,939,133 are all prime).

snowflake curve

See **Koch snowflake**.

soap film

See **bubbles**.

Soddy circle

A solution to the three-circle form of the **Apollonius problem** in which each of the given circles is tangent to (just touches) the other two.[312] There are two Soddy circles: the *outer Soddy circle,* which surrounds the three given circles, and the *inner Soddy circle,* which is interior to them. The inner Soddy circle is the solution to the **four coins problem**.

Soddy's formula

If four circles *A, B, C,* and *D,* of radii r_1, r_2, r_3, and r_4, are drawn so that they do not overlap but each touches the other three, and if we let $b_1 = 1/r_1$, etc., then

$$(b_1 + b_2 + b_3 + b_4)^2 = 2(b_1^2 + b_2^2 + b_3^2 + b_4^2).$$

solid

Of or relating to three-dimensional geometric figures or bodies.

solid angle

An angle formed by three or more planes intersecting at a common point. Solid angles are measured in **steradian**s.

solid geometry

The geometry of three-dimensional space.

solidus

The slanted line in a fraction such as *a/b* dividing the **numerator** from the **denominator**.

solitaire

See **peg solitaire**.

solitary number

A number that is not one of a pair of **amicable numbers**. Examples include all **prime number**s, all integer powers or primes, and other numbers such as 9, 16, 18, 52, and 160.

soliton

A solitary wave that can travel for long distances without changing its shape or losing energy. Mathematically, a soliton is a solution to a **partial differential equation** that is localized in some directions but not localized in time, and which does not change its shape. The first soliton to be described was a water wave seen by the engineer and shipbuilder John Scott Russell (1808–1882) in 1834 when he was riding by the Grand Union Canal at Hermiston, Glasgow. He observed that when a canal boat stopped, its bow wave continued onward as a well-defined elevation of the water at constant speed. (Another account says Scott Russell observed it on the Glasgow and Ardrossan Canal when a horse bolted with a light canal boat in tow.) The phenomenon was largely forgotten until the 1960s, when the American physicist Martin Kruskal rediscovered it and called it a soliton wave. On July 12, 1995, a viaduct at Hermiston was renamed the John Scott Russell viaduct, with Kruskal unveiling plaques and attempting to re-create a soliton wave. A classic soliton occurs on the River Severn in England, which starts at the head of the triangular Bristol Channel and narrows rapidly upstream. When the tide comes in, it is greatly compressed and produces the Severn Bore, a tidal wave up to 6 feet (2 m) high which rushes about 20 miles (32 km) up the river to Gloucester at a speed of up to 10 mph (16 km/hr). It is strongest at the spring tides, i.e. at full and new moons, and is now popular with surfers and canoeists.

solution

A value that satisfies the requirements of an equation. See also **lost solutions**.

Soma cube

A mathematical puzzle devised by Piet **Hein** in 1936 during a lecture on quantum mechanics by Werner Heisenberg in which the great German physicist was describing a space sliced into cubes. In a moment of genius, Hein grasped that the result of combining all seven of the irregular shapes that can be made from no more than four unit cubes joined at their faces is a single larger (3 × 3 × 3) cube. The Soma cube was first brought to popular attention by Martin **Gardner** in his "Mathematical Games" column in *Scientific American* in 1958. All 240 possible solutions were first identified by John **Conway** and Mickael Guy in 1961. The pieces can also be used to make a variety of other interesting three-dimensional

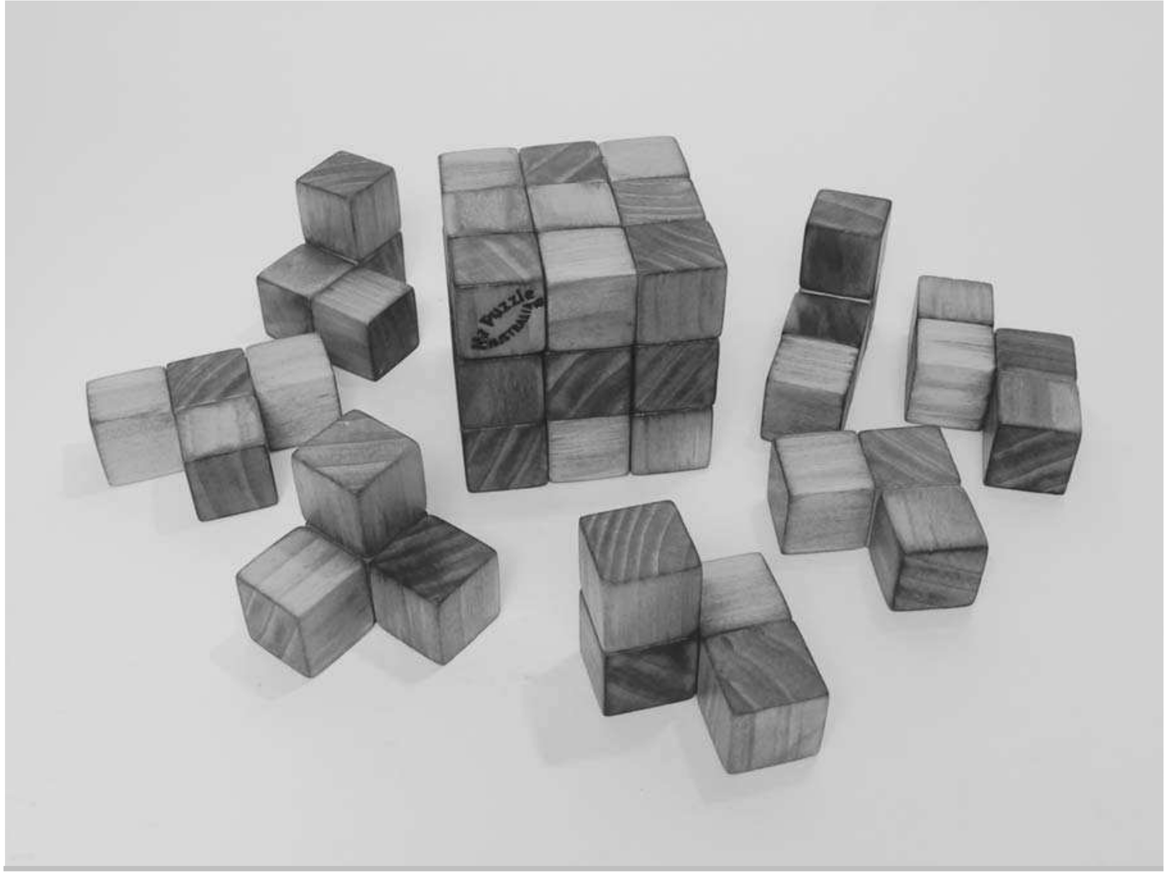

Soma cube How the pieces of the cube fit together. *Mr. Puzzle Australia, www.mrpuzzle.com.au*

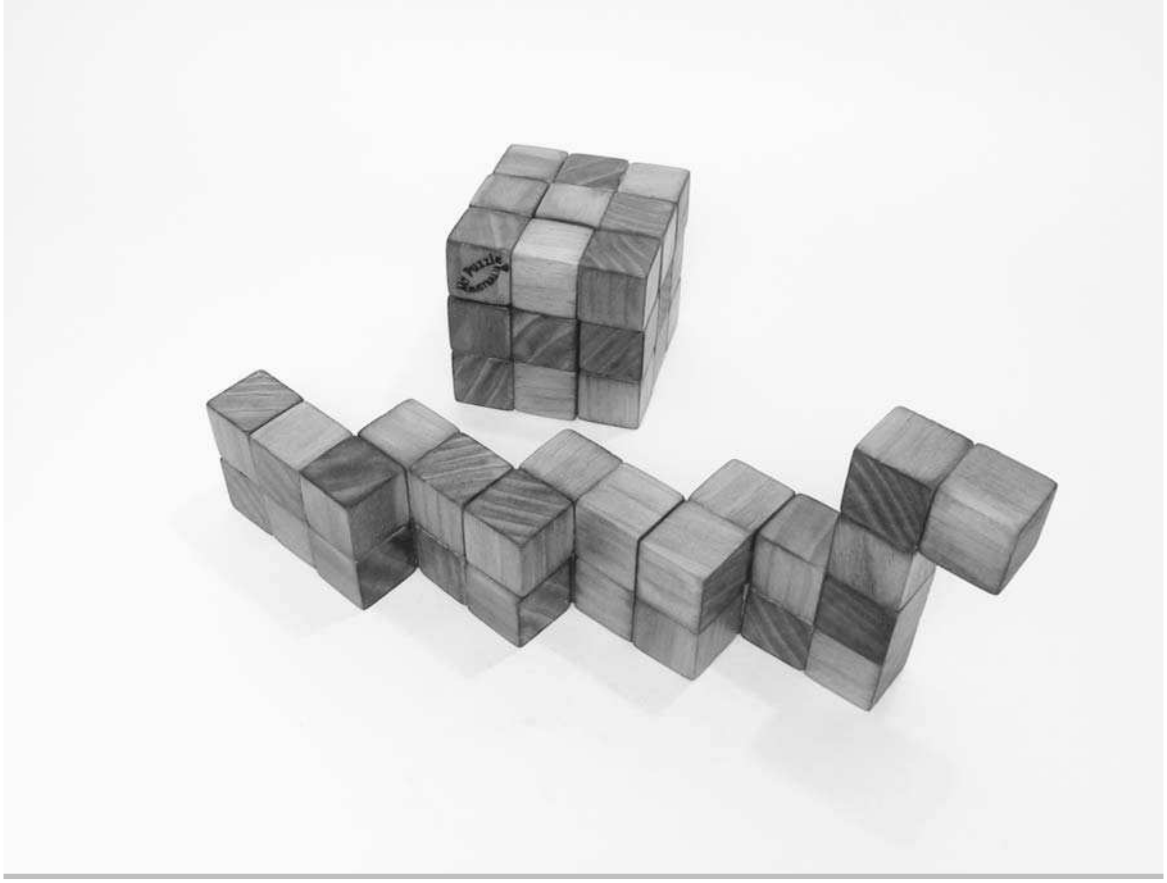

Soma cube One of many animals into which the cube pieces can be arranged. *Mr. Puzzle Australia, www.mrpuzzle.com.au*

shapes so that the Soma cube is often regarded as a three-dimensional analog of **tangrams**. It is possible that the puzzle is named after the fictitious drug "soma" in Aldous Huxley's novel *Brave New World.* See also **polyomino**.[112]

Sophie Germain prime

Any **prime number** p such that $2p + 1$ is also prime; the smallest examples are 2, 3, 5, 11, 23, 29, 41, 53, 83, 89, 113, and 131. Around 1825 Sophie **Germain** proved that the first case of **Fermat's last theorem** (FLT) is true for such primes. Soon after, Adrien-Marie Legendre began to generalize this by showing the first case of FLT also holds for odd primes p such that $kp + 1$ is prime, $k = 4, 8, 10, 14,$ and 16. In 1991 Fee and Granville extended this to $k < 100$, where k is not a multiple of three. Many similar results were also shown, but now that FLT has been proven correct, they are of less interest.

soroban

See **abacus**.

space

(1) The three-dimensional theater in which things as we know them can exist or in which events can take place. In the Einsteinian worldview, space and **time** are united inextricably in a **space-time** continuum and there is also the possibility of **higher dimensions**. (See also **fourth dimension**.) (2) In mathematics, there are additionally many other types of space, most of them too abstract to imagine or to describe accurately in a few sentences. Generally, a mathematical space is a **set** of points with additional features. In a *topological space* (see **topology**) every point has a collection of neighborhoods to which it belongs. In an *affine space,* which is a generalization of the familiar concepts of a straight line, a plane, and ordinary three-dimensional space, a defining feature is the ability

to fix a point and a set of coordinate axes through it so that every point in the space can be represented as a "tuple," or ordered set, of coordinates. Other examples of mathematical spaces include *vector spaces, measure spaces,* and *metric spaces.*

space-filling curve

A curve that passes through every point of a finite region (such as a unit square or unit cube) of an n-dimensional space, where $n \geq 2$. A well-known example is the **Peano curve**.

space-time

The inseparable four-dimensional **manifold**, or combination, which **space** and **time** are considered to form in the special and general theories of relativity (see **relativity theory**). A point in space-time is known as an *event.* Each event has four coordinates (x, y, z, t). Just as the x, y, z coordinates of a point depend on the axes being used, so distances and time intervals, (which are invariant in Newtonian physics) may depend (in relativistic physics) on the reference frame of an observer; this can lead to bizarre effects such as *length contraction* and *time dilation.* A *space-time interval* between two events is the invariant quantity analogous to distance in **Euclidean space**. The space-time interval s along a curve is defined by the quantity

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2dt^2,$$

where c is the speed of light. A basic assumption of relativity theory is that coordinate transformations leave intervals invariant. However, note that whereas distances are always positive, intervals may be positive, zero, or negative. Events with a space-time interval of zero are separated by the propagation of a light signal. Events with a positive space-time interval are in each other's future or past, and the value of the interval defines the proper time measured by an observer traveling between them.

special function

A **function**, often named after the person who introduced it, that has a particular use in physics or some branch of mathematics. Examples include *Bessel functions, Lagrange polynomials,* **beta functions**, **gamma functions**, and **hypergeometric functions**.

special relativity

See **relativity theory**.

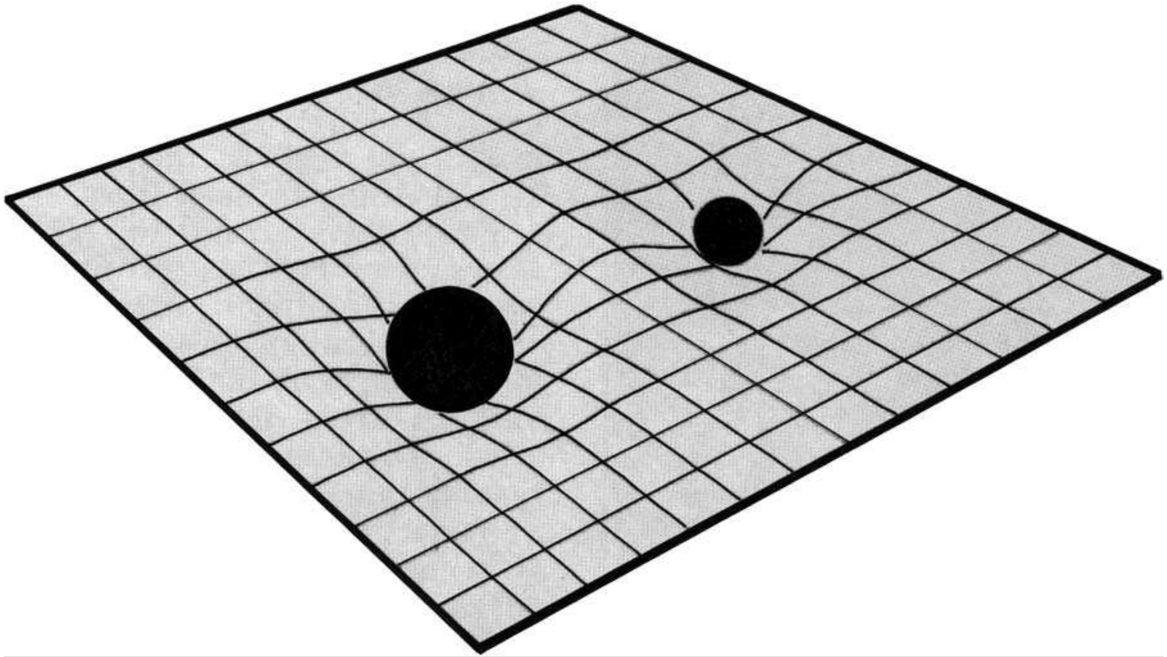

space-time Gravity seen as a curvature of the fabric of space and time.

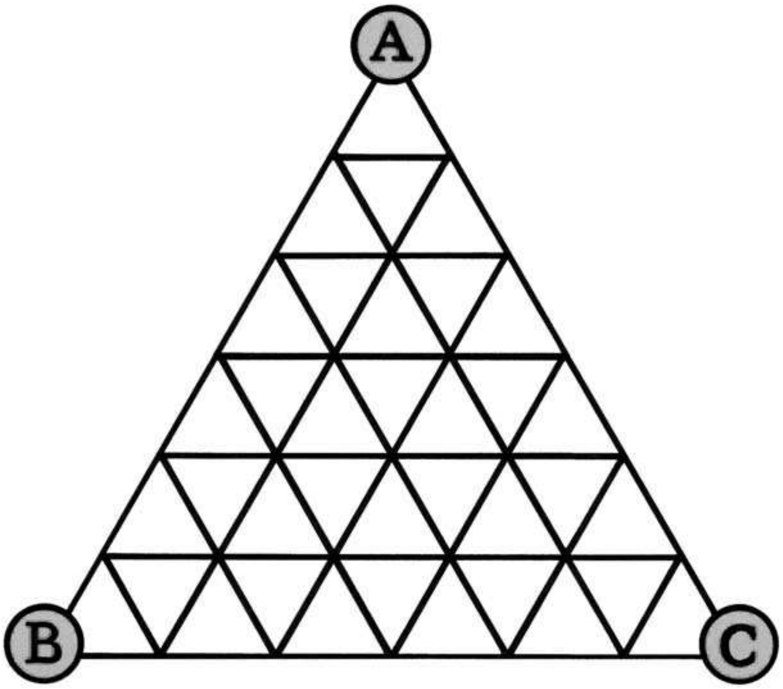

Sperner's lemma A triangular lattice that can be used for demonstrating Sperner's lemma. To understand how, see the accompanying entry.

spectrum

(1) In **quantum mechanics**, the set of allowable energy levels of a particle or system. It is directly related to bright or dark lines in a spectrum of light produced by a prism. (2) In mathematics, the set of **eigenvalue**s of a linear transformation. By historical coincidence, it is equivalent to the notion of a spectrum in quantum mechanics.

Sperner's lemma

Take a triangle *ABC,* labeled counterclockwise, and subdivide it into lots of smaller triangles in any arbitrary way. Then label all the new vertices as follows: (1) vertices along *AB* may be labeled either *A* or *B,* but not *C;* (2) vertices along *BC* may be labeled either *B* or *C,* but not *A;* (3) vertices along *CA* may be labeled either *C* or *A,* but not *B;* (4) vertices inside triangle *ABC* may be labeled *A* or *B* or *C.* Now shade in every small triangle that has three different labels. Use two different shadings to distinguish the triangles that have been labeled counterclockwise (i.e., in the same sense as triangle *ABC*) from the triangles that have been labeled clockwise (i.e., in the sense opposite to that of triangle *ABC*). Then there will be exactly one more counterclockwise triangle than clockwise triangles. In particular, the number of shaded triangles will be odd. This is Sperner's lemma, named after its discoverer, the German mathematician Emanuel Sperner (1905–1980). Sperner's lemma is equivalent to the **Brouwer fixed-point theorem**; a version of it holds in all dimensions.

sphere

Roughly speaking, a ball-shaped object. In everyday usage a sphere is often considered to be solid; mathematicians call this the *interior* of the sphere. In mathematics, a sphere is a **quadric** consisting only of a surface and is therefore hollow. More precisely, a sphere is the set of all points in three-dimensional **Euclidean space** that lie at distance *r,* the radius, from a fixed point. In **analytical geometry** a sphere with center (x_0, y_0, z_0) and radius r is the set of all points (x, y, z) such that

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = r^2.$$

A sphere can also be defined as the **surface of revolution** formed by rotating a circle about its diameter. If the circle is replaced by an **ellipse**, the shape becomes a **spheroid**. The surface area of a sphere is $4\pi r^2$ and its volume is $4\pi r^3/3$. The sphere has the smallest surface area among all surfaces enclosing a given volume and it encloses the largest volume among all closed surfaces with a given surface area. In nature, **bubbles** and water drops tend to form spheres because surface tension always tries to minimize surface area. The circumscribed cylinder for a given sphere has a volume which is 3/2 times the volume of the sphere. If a spherical egg were cut up by an egg-slicer with evenly spaced wires, the bands between the cuts (on the surface of the sphere) would have exactly the same area. Spheres can be generalized to other dimensions. For any natural number n, an n-sphere is the set of points in $(n + 1)$-dimensional **Euclidean space** that lie at distance r from a fixed point of that space. A 2-sphere is therefore an ordinary sphere, while a 1-sphere is a **circle** and a 0-sphere is a pair of points. An n-sphere for which $n = 3$ or more is often called a **hypersphere**.

sphere packing

See **packing**, **Kepler's conjecture**, and **cannonball problem**.

spherical geometry

See **elliptical geometry**.

sphericon

A curious and mathematically delightful three-dimensional object made from a right double-**cone**–two identical, 90° cones joined base to base–and an added twist. To create a sphericon, a right double-cone is sliced along a plane that includes both vertices. The resulting cross section is a square, which enables one of the halves to be rotated through a right angle and the two halves to be glued back together without any overlap. This final twist enables the sphericon to roll in an unusual way. An ordinary cone placed on a flat surface rolls around in circles. A double-cone can roll in a clockwise circle or a counterclockwise one. A sphericon, in contrast, performs a controlled wiggle, with first one conical sector in contact with the flat surface, then the other. Two sphericons placed next to each other can roll on each other's surfaces. Four sphericons arranged in a square block can all roll around one another simultaneously. And eight sphericons can fit on the surface of one sphericon so that any one of the outer solids can roll on the surface of the central one. The sphericon was first found by the En-glishman Colin Roberts in 1969, while he was still in school. In 1999 he brought his discovery to the attention of Ian **Stewart** who subsequently wrote about the new object in his "Mathematical Recreations" column in *Scientific American.*[322]

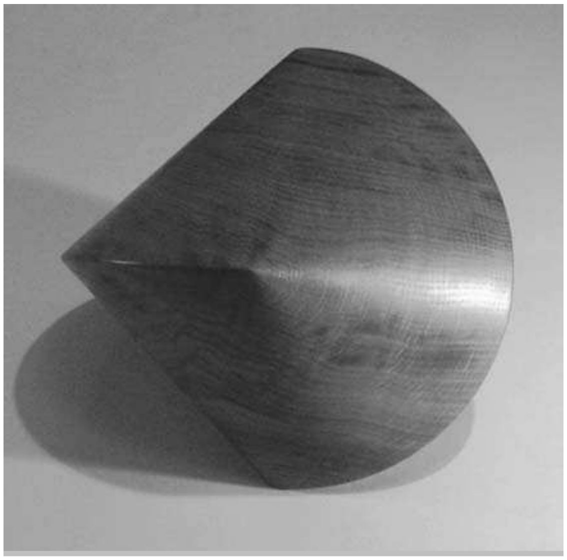

sphericon A model in oak of this remarkable shape. *Paul and Colin Roberts*

spheroid

A surface in three dimensions obtained by rotating an **ellipse** about one of its principal axes. If the ellipse is rotated about its major axis, the surface is called a *prolate spheroid* (similar to the shape of a rugby ball). If the minor axis is chosen, the surface is called an *oblate spheroid* (similar to the shape of Earth). The **sphere** is a special case of the spheroid in which the generating ellipse is a **circle**. A spheroid is a special case of an **ellipsoid** where two of the three major axes are equal.

Sphinx riddle

In Greek mythology, the Sphinx sat outside Thebes and asked this riddle of all travelers who passed by. If the traveler failed to solve the riddle, then the Sphinx killed him/her. And if the traveler answered the riddle correctly, then the Sphinx would destroy herself. The riddle: what goes on four legs in the morning, on two legs at noon, and on three legs in the evening? Oedipus solved the riddle, and the Sphinx destroyed herself. The solution: a man, who crawls on all fours as a baby, walks on two legs as an adult, and walks with a cane in old age. Of course morning, noon, and night are metaphors for the times in a person's life. Such metaphors are common in riddles. There were two Thebes; apparently the Thebes in this myth was the one in Greece, and the Sphinx was different from the one that stands at Giza, in Egypt.

spider-and-fly problem

A puzzle that was originally posed in the *Weekly Dispatch,* an English newspaper, on June 14, 1903, by Henry **Dudeney**, and that appears as one of the problems in *The Canterbury Puzzles* (1907).[87] A simple but elegant exercise in **geodesic**s, it is Dudeney's best-known brain-teaser. In a cuboidal (shoebox-shaped) room measuring 30′ × 12′ × 12′, a spider is in the middle of one 12′ × 12′ wall, 1 foot away from the ceiling. A fly is in the middle of the opposite wall 1 foot away from the floor. If the fly remains stationary, what is the shortest total distance (the geodesic) the spider must crawl along the walls, ceiling, and floor in order to get to the fly? The answer, 40′, can be obtained by flattening out the walls. Note that this distance is shorter than the 42′ the spider would have to travel if it first crawled along the wall to the floor, then across the floor, then up 1′ to get to the fly. A twist to the problem can be obtained by a spider that suspends himself from a strand of cobweb and thus takes a shortcut by not being forced to remain glued to a surface of the room. If the spider attaches a strand of cobweb to the wall at his starting position and lowers himself down to the floor (thus not crawling a single inch), he can then walk across the length of the room (30′) and ascend a single foot, thus reaching his prey after a total crawl of 31′ (although the total distance traveled is of course 42′). If the spider is not proficient with fastening strands to vertical walls, he must first ascend 1′ to the ceiling, from where it can lower himself to the floor, traverse the length of the room, and climb one foot to get to the fly, for a total distance crawled of 32′.

Dudeney and Sam **Loyd** offered several versions of the problem in a rectangular room. In 1926 Dudeney gave a version on a cylindrical glass with the source and the target on opposite sides.

spinor

A mathematical object similar to a **vector**, but which changes sign when rotated through 360°. Spinors were invented by Wolfgang Pauli and Paul **Dirac** (they were named by Paul Ehrenfest) to represent the spin of a subatomic particle. In the early 1930s, Dirac, Piet **Hein**, and others at the Niels Bohr Institute created games such as **Tangloids** to teach and model the calculus of spinors.

spiral

A curve that turns around some central point, getting progressively closer to it or progressively farther from it, depending on which way the curve is followed. Among the best known types are the **Archimedean spiral**, the **logarithmic spiral**, the **circle involute**, and the **lituus**. Like their space-curve cousin the **helix**, all spirals are asymmetric and each come in two forms that are mirror reflections of one another.

spirograph curve

See **roulette**.

Sprague-Grundy theory

A theory of certain classes of games called *impartial games,* discovered independently by Roland Percival Sprague (in 1936) and Patrick Michael Grundy (in 1939)

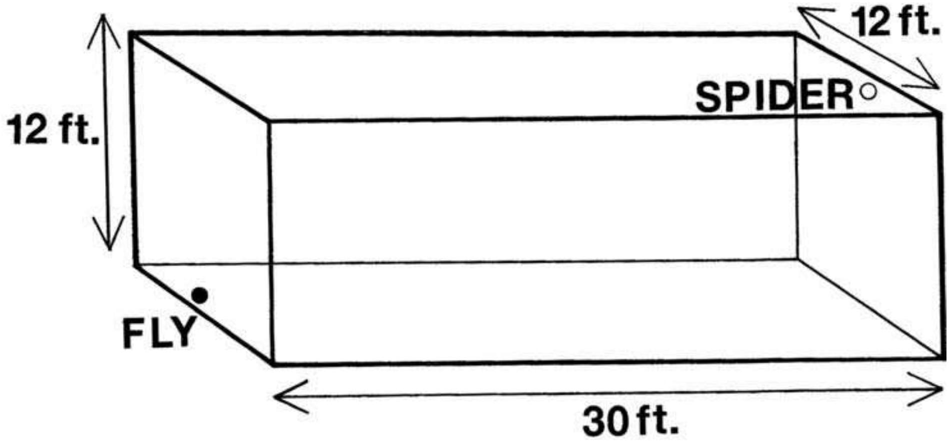

Spider-and-fly problem What is the shortest route for the spider to catch its prey?

and originally applied to **Nim**. In simple terms, they showed that one could take any impartial game and analyze it in terms of Nim heaps that could grow or decrease in size. The theory was developed further by E. R. Berlekamp, John **Conway**, and others, and presented comprehensively in the books *Winning Ways for Your Mathematical Plays* and *On Numbers*. Sprague-Grundy theory has been applied to other combinatorial games, including kayles.

Sprouts

A pencil-and-paper game invented by John **Conway** and Michael S. Paterson at Cambridge University in 1967. Sprouts is played by two players, starting with a few dots (called *spots*) drawn on a sheet of paper. To make a move, a player draws a curve between two spots or a loop from a spot to itself; the curve may not cross any other curve. The player marks a new spot on the curve, dividing it in two. Each spot can have at most three curves connected to it. The player who makes the last move wins. Sprouts has been studied from the perspectives of **graph** theory and of **topology**. It can be proven that a game started with n spots will last at least $2n$ moves and at most $3n - 1$ moves. By enumerating all possible moves, one can show that the first player is guaranteed a win in games involving three, four, or five spots, while the second player can always win a game that starts with one, two, or six spots. Following a 1990 computer analysis of the game at Bell Labs out to 11 spots, David Applegate, Guy Jacobsen, and Daniel Sleator conjectured that the first player has a winning strategy when the number of spots divided by six leaves a remainder of three, four, or five. Sprouts was featured in the plot of the first part of the science fiction novel *Macroscope* by Piers Anthony.

square

(1) A **quadrilateral** with four equal sides that meet at right angles. (2) To square something is to multiply it by itself. To take the *square root* is the reverse process. A *square number* is the square of a whole number (see also **figurate number**). A well-known formula for the *difference of squares* is $a^2 - b^2 = (a + b)(a - b)$. This formula enables some otherwise difficult calculations to be done easily in the head. For example, $43 \times 37 = (40 + 3)(40 - 3) = 40^2 - 3^2 = 1600 - 9 = 1591$.

square free

An integer that is not divisible by a perfect square, n^2, for $n > 1$.

square pyramid problem

See **cannonball problem**.

square root of 2 ($\sqrt{2}$)

The first number shown to be what is now known as an **irrational number** (a number that can't be written in the form a/b, where both a and b are integers). This discovery was made by **Pythagoras** or, at any rate, by the Pythagorean group that he founded. The square root of 2 is the length of the hypotenuse (longest side) of a right **triangle** whose other two sides are each one unit long. A ***reductio ad absurdum*** proof that $\sqrt{2}$ is irrational is straightforward. Suppose that $\sqrt{2}$ is rational, in other words that $\sqrt{2} = a/b$, where a and b are coprime integers (that is, they have no common factors other than 1) and $b > 0$. It follows that $a^2/b^2 = 2$, so that $a^2 = 2b^2$. Since a^2 is even (because it has a factor of 2), a must be even, so that $a = 2c$, say. Therefore, $(2c)^2 = 2b^2$, or $2c^2 = b^2$, so b must also be even. Thus, in a/b, both a and b are even. But we started out by assuming that we'd reduced the fraction to its lowest terms. So there is a contradiction and therefore $\sqrt{2}$ must not be irrational. This type of proof can be generalized to show that any root of any natural number is either a natural number or irrational.

As a **continued fraction**, $\sqrt{2}$ can be written $1 + 1/(2 + 1/(2 + 1/(2 + \ldots)))$, which yields the series of rational approximations: 1/1, 3/2, 7/5, 17/12, 41/29, 99/70, 239/169, ... Multiplying each numerator (number on the top) by its denominator (number on the bottom) gives the series 1; 6; 35; 204; 1,189; 6,930; 40,391; 235,416; . . . which follows the pattern: $A_n = 6A_{n-1} - A_{n-2}$. Squaring each of these numbers gives 1; 36; 1,225; 41,616; 1,413,721; 48,024,900; 1,631,432,881; . . . each of which is also a **triangular number**. The numbers in this sequence are the only numbers that are both square and triangular.

squarefull number

See **powerful number**.

squaring the circle

The problem to find a construction (see **constructible**), using only a straightedge and compass, that would give a square of the same area as a given circle. It is now known that this is impossible because, as the German mathematician Ferdinand von Lindemann showed in 1882, it would amount to finding a **polynomial** expression for **pi**, which can't be done as π is a **transcendental number**. (All the coordinates of all points that can be constructed with ruler and compass are algebraic numbers.) This minor inconvenience, however, hasn't prevented some amateur mathematicians from continuing to claim they have found proofs that the circle can be squared. Some of the most bizarre attempts have involved proposing a different and rational value for π. In 1897, the Indiana

State Legislature came within a hair's breadth of introducing a bill to set the value of π equal to 3.2! In fact, long before π was finally proved to be transcendental, most learned societies had conjectured that a proof was impossible and had stopped considering circle-squaring arguments sent to them.

squaring the square

The problem of how to tile a square with *integral squares* (squares of integral side-length). Of course squaring the square is a trivial task unless additional conditions are set. The most studied restriction is the *perfect squared square:* a square such that each of the smaller squares has a different size. The name was coined in humorous analogy with **squaring the circle** and is first recorded as being studied by R. L. Brooks, C. A. B. Smith, A. H. Stone, and W. T. Tutte at Cambridge University. The first perfect squared square was found by Roland Sprague in 1939. If such a **tiling** is enlarged so that the formerly smallest tile becomes as big as the original square, it becomes clear that the whole plane can be tiled with integral squares, each having a different size. It is still an unsolved problem, however, whether the plane can be tiled with a set of integral square tiles such that each natural number is used exactly once as the size of a tile. A *simple squared square* is one where no subset of the squares forms a rectangle. The smallest simple perfect squared square was discovered by A. J. W. Duijvestin using a computer search. His tiling uses 21 squares, and has been proved to be minimal. Other possible conditions that lead to interesting results are nowhere-neat squared squares and no-touch squared squares. Developments leading to squaring the square can be traced back to 1902 and the first appearance of Henry **Dudeney**'s Lady Isabel's casket, later published as problem #40 in *The Canterbury Puzzles.*[87]

standard deviation

A measure of the spread of a **set** of data. For a **Gaussian** distribution, the standard deviation hints at the width of the tails of the distribution **function**.

Stanhope, Earl

A line of English earls, several of whom were notable mathematicians and polymaths. The third earl, Charles Stanhope (1753–1816), invented the process of stereotyping and, in 1777, devised the first mechanical logical calculator. The fourth earl, Philip Henry Stanhope (1781–1855) was a mathematician who applied his knowledge to mapmaking and **maze** design. From 1818 to 1830 he planted a maze according to a basic plan laid down by the second earl (1714–1786) that was probably the first to incorporate islands that defeat the familiar "hand-on-wall" method. (There are islands in the famous Hampton Court maze, yet the hand-on-wall method solves it.) His designs often had a number of islands, or just isolated lengths of hedge.

Star of David

Also known as a *hexagram,* the six-pointed star obtained by extending the sides of a regular **hexa**gon to the points of intersection.

star of Lakshmi

An eight-pointed star design often used in architecture, particularly as a tiling or other decoration on the floor of a room that has four- or eight-fold symmetry. A notable example is in the **octagonal** central lobby in the Houses of Parliament in London. It was used by Hindus to symbolize *Ashtalakshmi,* the eight forms of wealth.

state space

See **phase space**.

stationary point

A point on the graph of a **function** where the **tangent** to the graph is parallel to the x-axis or, equivalently, where the **derivative** of the function is 0. There are four kinds of stationary points: (1) a *local minimum,* where the derivative of the function changes from negative to positive; (2) a *local maximum,* where the derivative changes from positive to negative; (3) a *rising point of inflection,* where the derivative is positive on both sides of the stationary point; and (4) a *falling point of inflection,* where the derivative is negative on both sides of the stationary point.

statistical mechanics

The study of statistical and thermal properties of physical materials and their idealized mathematical models.

statistics

The study of ways that lots of data can be represented using a few numbers and the study of how such numbers can be chosen and used to draw reasonable conclusions about the data. The word *statistics* comes from the Latin *statis* for "political state"; one of the main tasks of the subject involves analyzing facts and figures about governments, resources, and populations. Although a powerful tool, statistics is open to abuse, both intentional and unintentional. Benjamin Disraeli (1804–1881) may have gone a little over the top when he said, "There are lies, damned lies, and statistics," but Scottish author Andrew Lang (1844–1912) could have been describing many a politician when he remarked, "He uses statistics as a drunken man uses lamp posts–for support rather than

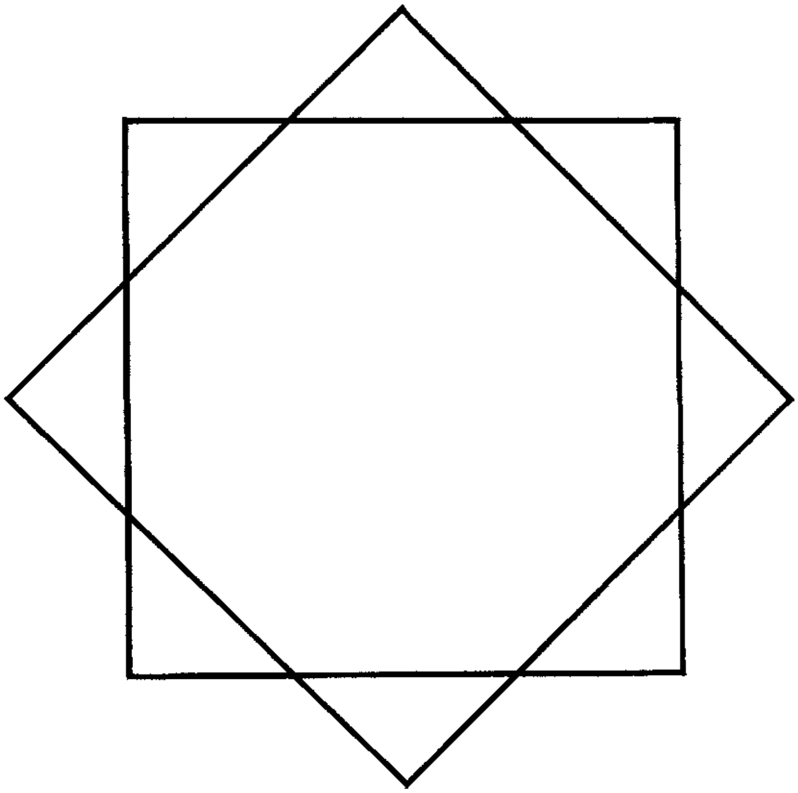

star of Lakshmi

illumination." The branch of statistics most commonly used in recreational mathematics is **probability theory.**

Steiner, Jakob (1796–1863)

A Swiss mathematician, considered by many to be the greatest geometer since **Apollonius** of Perga. Largely self-taught, he became a professor at the University of Berlin, and was a pioneer in the field of **projective geometry**. He had an important influence on his students, including Bernhard **Riemann**.

Steiner-Lehmus theorem

Any triangle that has two equal angle bisectors (each measured from a vertex to the opposite sides) is an **isosceles** triangle. In 1840, a Berlin professor Ludolph Lehmus wondered if this statement is true, given that it is the inverse of the already proven rule: If a triangle is isosceles then two of its internal bisectors are equal. He put the problem to Jakob **Steiner** who was quickly able to show its validity. Shortly after, Lehmus himself found a neater proof and it is has since become a favorite pastime of geometry hobbyists to search for still simpler proofs of the theorem.

Steinhaus, Hugo Dyonizy (1887–1972)

A Polish mathematician who was an influential member of the *Lvov school,* based at the Jan Kazimierz University in Lvov, which also included Stefan **Banach** and which focused on problems in functional **analysis**, real functions, and probability in the 1920s and '30s. Early on, Steinhaus's work revolved around applications of the **Lebesque** measure and integral. In 1923 he published the first rigorous account of the theory of tossing coins based on **measure theory**, and in 1925 was the first to define and discuss the concept of strategy in **game theory**. During World War II, as a Jew he was compelled to hide from persecution by the Nazis, yet continued his mathematical work despite great hardship. In 1944, Steinhaus proposed the problem of dividing a cake into *n* pieces so that it is proportional and envy free (see **cake-cutting**). He is also well known as the author of the widely read *Mathematical Snapshots.*[316]

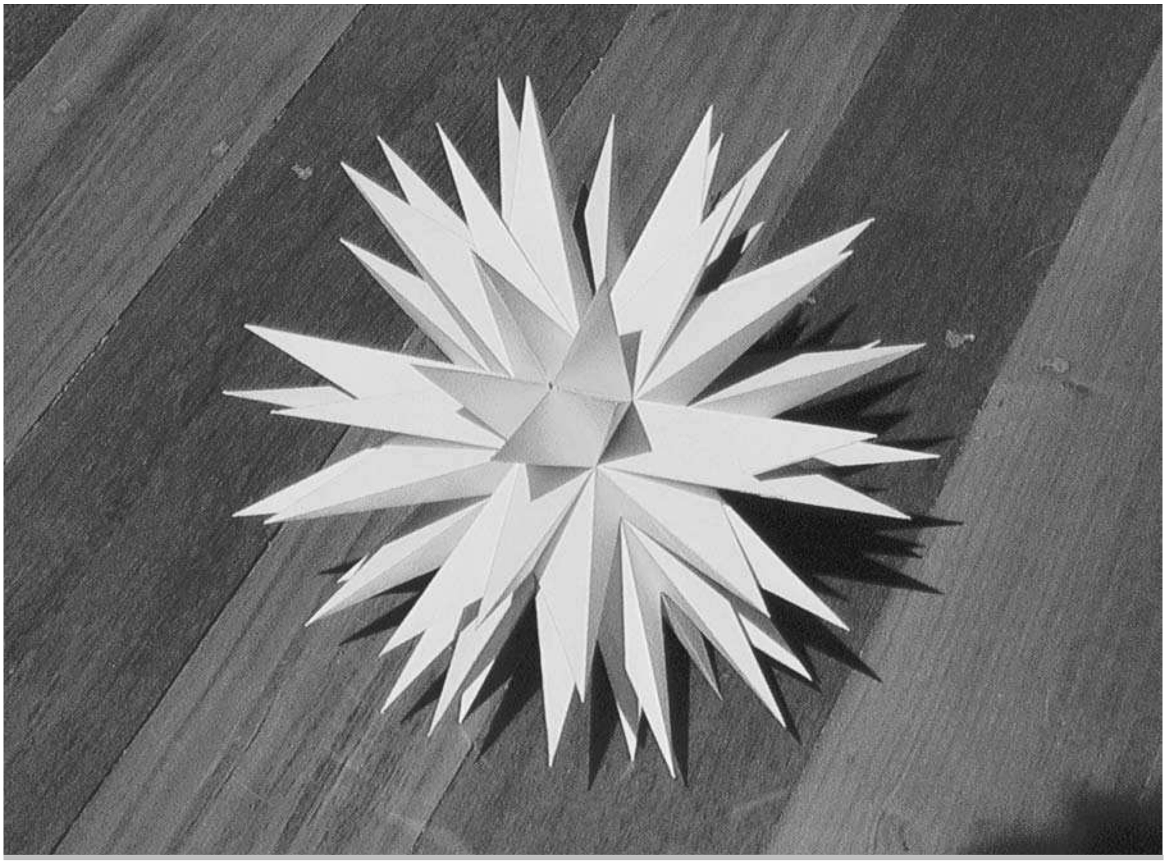

stellation The final stellation of the icosahedron. *Robert Webb, www.software3d.com; created using Webb's Stella program*

stellation

(1) The process of constructing a new **polyhedron** by extending the face planes of a given polyhedron past their edges. (2) The new polyhedron thus obtained. Starting with the **icosahedron**, for example, there are 59 possible stellations, including the *great icosahedron,* which is one of the **Kepler-Poinsot solids**.

steradian

The international (SI) unit in which **solid angle**s are measured. The steradian (sr) is a solid angle with its **vertex** in the center of a **sphere** and which cuts off from the spherical surface an area equal to the square of the radius of the sphere. The full sphere represents a spherical angle of 4π steradians.

Stewart, Ian (1945–)

A British mathematician at Warwick University who has written many books and articles on recreational math and popular science. From 1990 to 2001 he wrote the "Mathematical Recreations" column in *Scientific American.* His books include *Flatterland: Like Flatland Only More So* (a modern version of Edwin Abbott's classic ***Flatland: A Romance of Many Dimensions***); *Does God Play Dice: The New Mathematics of Chaos,* and *Evolving the Alien* (with Jack Cohen).[317, 318, 321, 323]

Stewart toroid

A **polygon** with a hole through it that meets some or all of the criteria specified by Bonnie Madison Stewart in his 1980 book *Adventures among the Toroids.* These criteria are as follows: all faces must be regular; faces that meet mustn't lie in the same plane; the polygon must be quasi-convex; the hole through the polygon must change its **genus**; and, the faces aren't allowed to intersect with themselves or each other. Stewart toroids, pierced relatives of the familiar Platonic solids, combine dazzling complexity with attractive symmetry.

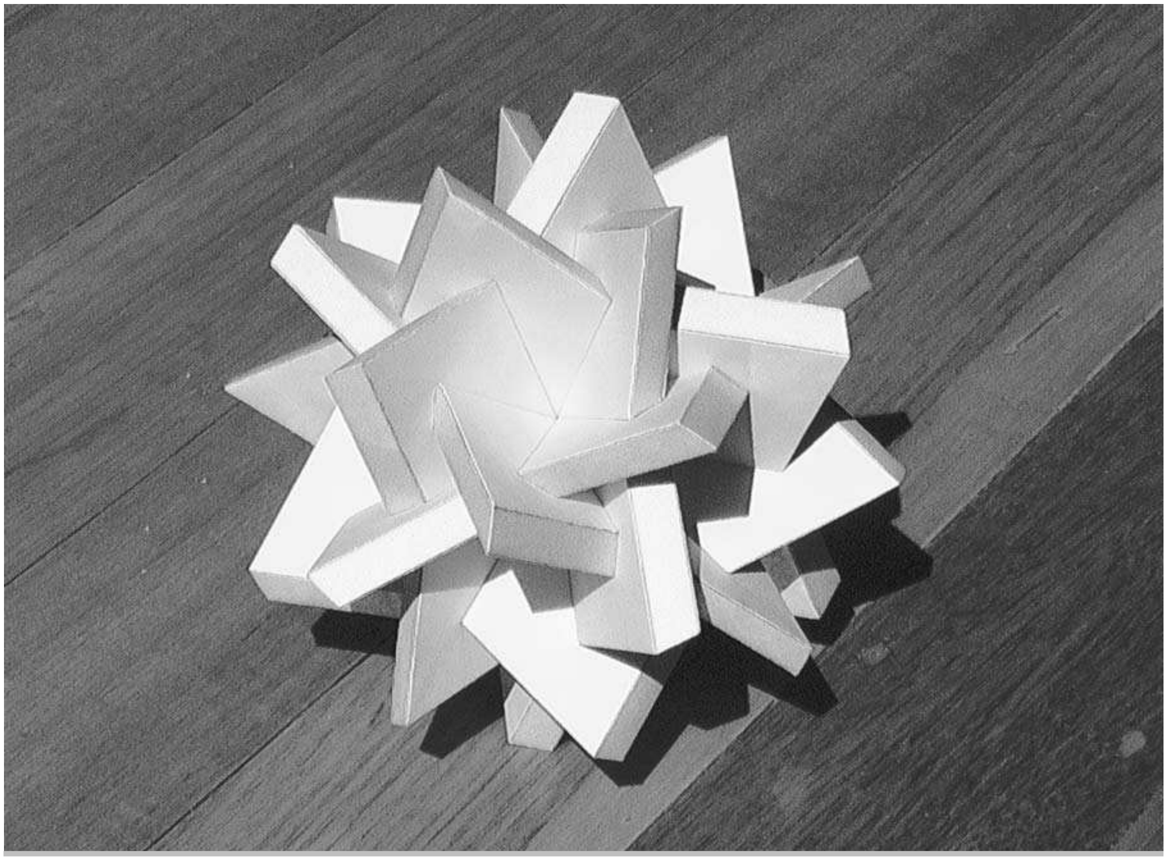

stellation A stellation of the small stellated truncated dodecahedron, a uniform polyhedron. *Robert Webb, www.software3d.com; created using Webb's Stella program*

stochastic
Something that is **random**.

stochastic process
A **dynamical system** with **random** fluctuations at each iteration or that is influenced by random noise. A random variable that, at each stage in time, depends on its previous values and on further random choices. For example, the price of a stock is often modeled as a stochastic process.

Stomachion
See **Loculus of Archimedes**.

straight
Having no deviations. A *straight line* is usually simply called a **line**. A *straight angle,* or *flat angle,* is exactly 180°.

strange attractor
See **chaotic attractor**.

strange loop
A phenomenon in which, whenever movement is made upward or downward through the levels of some hierarchical system, the system unexpectedly arrives back where it started. Douglas **Hofstadter** has used the strange loop as a paradigm in which to interpret paradoxes in logic, such as **Grelling's paradox** and **Russell's paradox**, and has called a system in which a strange loop appears a *tangled hierarchy.*

strategy
In **game theory**, a policy for playing a game. A strategy is a complete recipe for how a player should act in a game under all circumstances. A policy may employ **randomness**, in which case it is referred to as a **mixed strategy**.

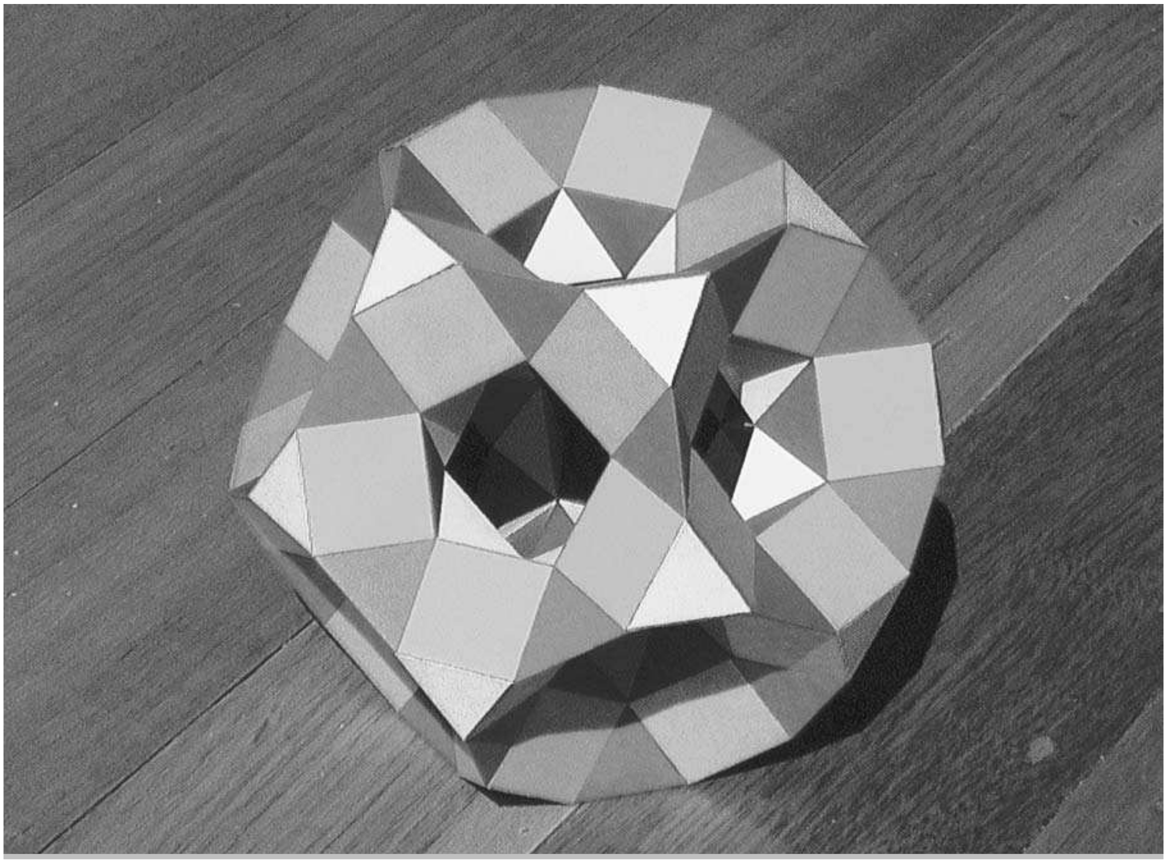

Stewart toroid A drilled truncated dodecahedron. *Robert Webb, www.software3d.com; created using Webb's Stella program*

string

Any sequence of letters, numbers, digits, **bit**s, or symbols.

string puzzle

The origins of puzzles using string, a form of *disentanglement puzzle,* are lost in antiquity but one of the earliest stories involving this type of problem is that of the **Gordian knot**. Today many kinds of string puzzles are available, most of which involve removing a ring from loops of string that are threaded through a solid (typically wooden) body. See also **mechanical puzzle**s and **knot**.

string theory

An important theory in modern physics in which the fundamental particles in nature are thought of as the "musical notes" or excitation modes of vibrating elementary strings. These strings have the shortest meaningful physical length, the so-called *Planck length* (equal to about 10^{-33} cm), and no thickness. Even more bizarre, for the theory to make sense, the universe must have nine space dimensions and one time dimension, for a total of 10 dimensions. This idea of a 10-dimensional universe first appeared in the **Kaluza-Klein theory**. We're familiar with time and three of the space dimensions: the other six together are known as **Calabi-Yau space**s. In string theory, as in the case of a stringed instrument, the string must be stretched under tension in order to vibrate. This tension is fantastically high–equivalent to a weight of about 10^{39} tons. String theories are classified according to whether or not the strings are required to be closed loops, and whether or not the particle spectrum includes *fermions*. In order to include fermions in string theory, there must be a special kind of symmetry called *supersymmetry,* which means that for every *boson* (a particle, of integral spin, that transmits a force) there is a corresponding fermion (a particle, of half-integral spin, that makes up matter). So supersymmetry relates the particles that transmit forces to the particles that make up matter. Supersymmetric partners to currently known particles have not been observed in particle experiments, but theorists believe this is because supersymmetric particles are too massive to be produced by present-day high-energy accelerators. Particle accelerators could be on the verge of finding evidence for high energy supersymmetry in the next decade. Evidence for supersymmetry at high energy would be compelling evidence that string theory was a good mathematical model for nature at the smallest distance scales. In string theory, all of the properties of elementary particles–charge, mass, spin, and so forth–come from the vibration of the string. The easiest of these to understand is mass. The more frantic the vibration, the more energy; and since mass and energy are the same thing, higher mass comes from faster vibration.

strobogrammatic prime

A **prime number** that remains unchanged when rotated through 180°. An example is 619, which looks the same when read upside-down. To be strobogrammatic, a prime cannot contain digits other than 0, 1, and 8, which have a horizontal line of symmetry (ignoring font variations), and 6 and 9, which are vertical reflections of each other. An *invertible prime* is one that yields a different prime when the digits are inverted. Of course, these definitions are not taken seriously by mathematicians!

strophoid

A looping curve, first studied in 1670 by Isaac Barrow (1630–1677), the first Lucasian Professor of Mathematics at Cambridge and the immediate predecessor of Isaac **Newton** in this job. The strophoid, which is a special case of the general **cissoid**, was named by Enrico Montucci in 1846 from the Latin for "twisted belt shape." The elaborate rules for drawing a strophoid are as follows. Let C be a curve, let O be a fixed point called the pole, and let O' be a second fixed point. Let P and P' be points on a line through O meeting C at Q such that $P'Q = QP = QO'$. The locus of P and P' is the strophoid of C with respect to the pole O and fixed point O'. If the curve C is a straight line, the pole P is not on C, and the second point O' is on C, the resulting strophoid is called an *oblique strophoid.* If these same conditions apply except that O' is the point where the perpendicular from O to C cuts C, then the strophoid produced is called a *right strophoid.* On the other hand if C is a circle, O is the center of the circle, and O' a point on its circumference, then the strophoid that results is known as *Freeth's nephroid.* The French mathematician Gilles Roberval (1602–1675) found the strophoid in a different way–as the result of planes cutting a **cone**. When the plane rotates (about the tangent at its vertex) the collection of foci of the obtained conics gives the strophoid.

subgroup

A subset of a **group** that is a group under the same operation.

sublime number

A number such that both the *sum* of its divisors and the *number* of its divisors are **perfect numbers**. The smallest sublime number is 12. There are 6 divisors of 12–1, 2, 3, 4, 6, and 12–the sum of which is 28. Both 6 and 28 are perfect. The second sublime number begins 60,865 . . . , ends . . . 91,264, and has a total of 76 digits! It is not known if there are larger even sublime numbers, nor if there are any odd sublime numbers.

subset

A **set** whose members are members of another set; a set contained within another set.

substitution cipher

A **cipher** that replaces each plaintext (original message) symbol with a ciphertext (coded text) symbol. The receiver decodes using the inverse substitution. A simple example is the **Caesar cipher**.

subtraction

The **binary operation** of finding the difference between two quantities or numbers.

sultan's dowry

A sticky problem in **probability** that first came to light in Martin **Gardner**'s "Mathematical Recreations" column in the February 1960 issue of *Scientific American.* Gardner's original version has become known as the secretary problem. In the exactly equivalent form called the sultan's dowry problem, a sultan has granted a commoner the chance to marry one of his hundred daughters. The commoner will be shown the daughters one at a time and will be told each daughter's dowry. The commoner has only one chance to accept or reject each daughter; he can't go back and choose one that he has previously rejected. The sultan's catch is that the commoner may only marry the daughter with the highest dowry. What is the commoner's best strategy, assuming that he knows nothing in advance about the way the dowries are distributed?

Many mathematicians have tackled this question and numerous papers have been written on the subject. It has even spawned its own area of study within the field of management science. The consensus among those who have worked on the problem is that the commoner's best strategy is to let a certain fraction of the daughters pass and then choose the next one who has a dowry higher than any of the ones seen up to that point. The exact number to skip is determined by the condition that the odds that the highest dowry has already been seen is just greater than the odds that it remains to be seen *and that if it is seen it will be picked.* This amounts to finding the smallest x such that:

$$x/n > x/n \times [1/(x+1) + \ldots + 1/(n-1)]$$

Substituting $n = 100$ leads to the conclusion that the commoner should wait until he has seen 37 of the daughters, then pick the first daughter with a dowry that is bigger than any that have already been revealed. With this strategy, his odds of choosing the daughter with the highest dowry are surprisingly high: about 37%.[224]

supercomputer

A computer that, in its day, is faster than any contemporary conventional computer. Supercomputers are typically used for enormous number-crunching tasks such as weather-forecasting, simulations of fusion experiments or galaxy evolution, design of cars and planes, and code-cracking. At the start of 2004, the world's most powerful computer, by a long way, is the Earth Simulator in Japan, which is capable of a maximum of 40,960 gigaflops, that is, 40.96 trillion "floating point operations" per second. The top 500 supercomputers are listed at the following Web site: http://www.top500.org.

superegg

The **surface of revolution** of a **superellipse** given by the formula $|x/a|^{2.5} + |y/b|^{2.5} = 1$, where $a/b = 4/3$. The superegg was named by Piet **Hein** and singled out by him because of an unusual property: stood on either end it has a peculiar and surprising stability. Supereggs, made of metal, wood, and other materials, were sold as novelties in the 1960s; small, solid-steel ones were marketed as an "executive toy." The world's largest superegg, made of steel and aluminum and weighing 1 ton, was placed outside Kelvin Hall in Glasgow in 1971 to honor Hein's appearance there as a speaker.

superellipse

A **Lamé curve**, described by the formula $|x/a|^n + |y/b|^n = 1$, for which $n > 2$. Superellipses have a form partway between an **ellipse** and a rounded rectangle (or, if $a = b$, partway between a circle and a rounded square). The Danish poet and architect Piet **Hein** decided that the superellipse with $n = 5/2$ and $a/b = 6/5$ is the most pleasing to the eye. This so-called *Piet Hein ellipse* was quickly adopted as the basic motif for planning an open space at the center of Stockholm and was also incorporated into Scandinavian designs for office tables, desks, beds, and even roundabouts in roads. The **surface of revolution** of a superellipse is a *superellipsoid,* one special form of which has been nicknamed the **superegg**.

superegg

superfactorial

A **function** based on the **factorial** that produces **large numbers** very quickly. It is a recent creation that has not yet entered the mathematical mainstream, and has been defined in a couple of different ways. In 1995, in his book *Keys to Infinity,* Clifford Pickover gave the superfactorial as

$$n\$ = \underbrace{n!^{n!^{\cdot^{\cdot^{\cdot n!}}}}}_{n!\text{ terms}}$$

In the same year, Sloane and Plouffe offered an alternative definition:

$$n\$ = \prod_{i}^{n} i!$$

Pickover's superfactorial grows with extraordinary speed. The first two terms are 1 and 4, but the third term 3$, already has more digits than could be written down on this page. Sloan and Plouffe's factorial grows sedately by comparison–the first few values are 1, 1, 2, 12, 288, and 345,690–and is related to **Bell numbers**.

supersymmetry

A theory in physics that postulates a counterintuitive symmetric relationship between fermions, which are particles such as electrons that obey the Pauli exclusion principle and thus cannot occupy the same quantum state, and bosons, which are particles such as photons that can coexist in the same state.

supertetrahedral number

A type of **figurate number** in four dimensions (see **fourth dimension**). Supertetrahedral numbers are obtained by piling up the **tetrahedral number**s 1, 4, 10, 20, 35, etc., as in:

$1 = 1$

$1 + 4 = 5$

$1 + 4 + 10 = 15$

$1 + 4 + 10 + 20 = 35$

$1 + 4 + 10 + 20 + 35 = 70$

etc.

supplementary angles

Two angles that add up to 180°.

surd

A **radical** that expresses an **irrational number**. Surds may be *quadratic* (e.g., $\sqrt{2}$), *cubic* (e.g., $\sqrt[3]{2}$), *quartic* (e.g., $\sqrt[4]{2}$), and so on. (The term is sometimes used as a synonym for irrational number.) A *pure surd,* or entire surd, contains no **rational number**; that, is, all its factors or terms are surds (e.g., $\sqrt{2} + \sqrt{3}$). A *mixed surd* contains at least one rational number (e.g., $2 + \sqrt{3}$ or $3\sqrt{2}$).

surface

In mathematics, any object that locally (if you zoom in close enough to it) looks like a piece of a flat **plane**. A **sphere**, a **torus**, a **pseudosphere**, and a **Klein bottle** are examples of different types of surfaces.

surface of revolution

A surface produced by rotating a line or a curve about some axis. For example, a **sphere** is the surface of revolution generated when a circle spins about its diameter.

surreal number

A member of a mind-boggling vast class of numbers that includes all of the **real number**s, all of Georg Cantor's infinite **ordinal number**s (different kinds of infinity), a set of **infinitesimal**s (infinitely small numbers) produced from these ordinals, and strange numbers that previously lived outside the known realm of mathematics. Each real number, it turns out, is surrounded by a "cloud" of surreals that lie closer to it than do any other real numbers. One of these surreal clouds occupies the curious space between zero and the smallest real number greater than zero and is made up of the infinitesimals.

Surreal numbers were invented or discovered (depending on your philosophy) by John **Conway** to help with his analysis of certain kinds of games. The idea came to him after watching the British **Go** champion playing in the mathematics department at Cambridge. Conway noticed that endgames in Go tend to break up into a sum of games, and that some positions behave like numbers. He then found that, in the case of infinite games, some positions behaved like a new kind of number–the surreals. The name "surreal" was introduced by Donald **Knuth** in his 1974 book *Surreal Numbers: How Two Ex-Students Turned on to Pure Mathematics and Found Total Happiness.*[191] This novelette is notable as being the only instance where a major mathematical idea has been first presented in a work of fiction. Conway went on to describe the surreal numbers and their use in analyzing games in his 1976 book *On Numbers and Games.*[67] The surreals are similar to the **hyperreal number**s, but they are constructed in a very different way and the class of surreals is larger and contains the hyperreals as a subset.

swastika

An ancient ideogram, signs of which have been found in the Euphrates-Tigris Valley, and in some areas of the Indus Valley, dating back 3,000 years; it became commonly used

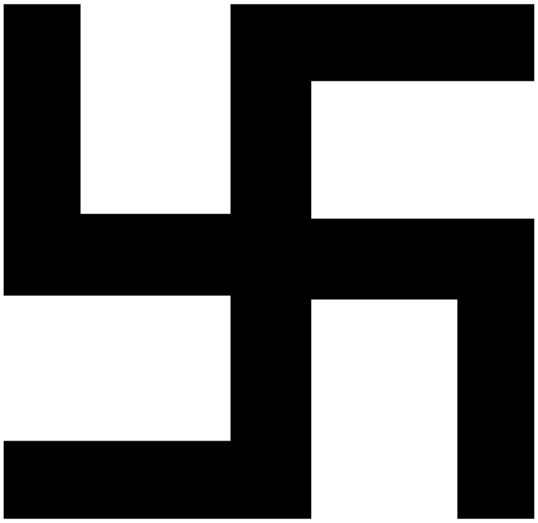

swastika

around 1000 B.C., possibly first in ancient Troy, in the northwest of modern Turkey. The swastika is an irregular icosagon (a 20-sided **polygon**), which in Arabic and Indian culture originally represented good luck. Of course, in more recent times, it was adopted as the symbol of the Nazi Party in Hitler's Germany and thus came to stand for anti-Semitism.

syllogism

An argument composed of three parts–a major premise, a minor premise, and a conclusion. For example: All men are mortal (major premise). Socrates is a man (minor premise). Therefore, Socrates is mortal (conclusion). The syllogism forms the basis of Aristotle's system of logic, which went unchallenged for over 2,000 years. Aristotle believed that by setting out any argument in syllogistic form, it should be possible to avoid fallacies. However, Bertrand **Russell** discovered several formal errors in the doctrine of syllogism.

Sylvester, James Joseph (1814–1897)

An English mathematician and lawyer who, in 1850, first used the term ***matrix*** in mathematics and gave it its present meaning of a rectangular array of numbers from which **determinant**s may be formed. Together with Arthur **Cayley**, he founded the *theory of invariants.* Hotheaded and vociferous in opposition of anti-Semitism, he was thrown out of the University of London for threatening another student with a table knife. Later he studied at Cambridge, emerging as Second Wrangler but without a degree (though in 1871 he earned an MA) because, as a Jew, he refused to accept the articles of the Church of England. After a spell working as an actuary and a barrister, and also taking private pupils (one of whom was Flor-ence Nightingale), he met Cayley, with whom he forged a lifelong friendship and collaboration, and returned to mathematics. He became professor of mathematics at the Royal Military Academy at Woolwich (1855–1870) and at the newly established Johns Hopkins University at Baltimore (1877–1883), founding the *American Journal of Mathematics,* before accepting the Savilian chair at Oxford (1883–1894). Remarkably, especially in the field of mathematics, he produced an extraordinary flood of ideas well into his old age. At 82, he worked out the theory of compound partitions. He also published on the roots of **quintic** equations and on **number theory**. His partnership with Cayley worked perfectly, since Cayley supplied the rigor which the brilliantly creative Sylvester lacked.

Sylvester's problem of collinear points

A problem posed in 1893 by James **Sylvester** who wrote: "Prove that it is not possible to arrange any finite number of real points so that a right line through every two of them shall pass through a third, unless they all lie in the same right line." No correct proof was forthcoming at the time, but the problem was revived by Paul **Erdös** in 1943 and correctly solved by T. Grünwald in 1944.

symmedian

Reflection of a **median** of a triangle about the corresponding angle **bisector**.

symmetric group

The **group** of all **permutation**s of a finite set. The symmetric group of a **set** of size n is denoted S_n and has $n!$ elements.

symmetry

An intrinsic property of a mathematical object that allows it to remain unchanged under certain types of **transformation**, such as rotation, reflection, or more abstract operations. The mathematical study of symmetry is systematized and formalized in the extremely powerful subject known as **group** theory. Symmetries and apparent symmetries in the laws of nature have played a part in the construction of physical theories since the time of Galileo and Newton. The most familiar symmetries are spatial or geometric ones. In a **snow**flake, for example, the presence of a symmetrical pattern can be detected at a glance. One of the most remarkable developments of the past half century has been the emergence of symmetry as a central theme of subatomic physics. This came about through a series of subtle evolutions in the concept of symmetry itself. Many researchers believe that this evolutionary process has not come to an end, and that further meaning of the concept of symmetry, with perhaps new mathematical structures, will develop in the coming years.

symmetry group

The **group** formed by the set of all rigid motions (translations, rotations, reflections, etc.) of Euclidean space that map all points of a subset F into F.

system

Something that can be studied as a whole. Systems may consist of subsystems that are interesting in their own right. Or they may exist in an environment that consists of other similar systems. Systems are generally understood to have an internal state, inputs from an environment, and methods for manipulating the environment or themselves. Since cause and effect can flow in both directions of a system and environment, interesting systems often possess **feedback**, which is self-referential in the strongest case.

Szilassi polyhedron

A **toroid**al heptahedron (seven-sided **polyhedron**) first described in 1977 by the Hungarian mathematician Lajos Szilassi. It has 7 faces, 14 vertices, 21 edges, and 1 hole. The Szilassi polyhedron is the **dual** of the **Császár polyhedron** and, like it, shares with the **tetrahedron** the property that each of its faces touches all the other faces. Whereas a tetrahedron demonstrates that four colors are necessary for a map on a surface topologically equivalent to a sphere, the Szilassi and Császár polyhedra show that seven colors are necessary for a map on a surface topologically equivalent to a **torus**.

T

tachyon

A hypothetical particle that travels faster than the speed of light. Tachyons were first proposed in prerelativistic times by the physicist Arnold Sommerfeld and named in the 1960s by Gerald Feinberg from the Greek *tachys* meaning "swift." By extension of this terminology, particles that travel slower than light are called *tardyons* (or *bradyons* in more modern usage) and particles, such as photons, that travel exactly at the speed of light are called *luxons*. The existence of tachyons is allowed by the mathematics of special **relativity theory**, one of the basic equations of which is

$$E = m/\sqrt{(1 - v^2/c^2)}$$

where E is the mass-energy of a particle, m its rest mass, v its velocity, and c the speed of light. This shows that for tardyons (particles of ordinary matter), E increases as v increases and becomes infinite when $v = c$, thus preventing an initially slower-than-light particle from being accelerated up to the speed of light and beyond. What about a particle for which v is always greater than c? In this case, $v^2/c^2 > 1$, so that the denominator in the equation above is an **imaginary number**–the square root of a negative real number. If m has a real value, E is imaginary, which is hard for physicists to swallow because E is a measurable quantity. If m takes an imaginary value, however (because one imaginary number divided by another is real), then E is real. Tachyons are allowed, therefore, providing (a) they never cross from one side of the light-barrier to the other, and (b) they have an imaginary rest mass (which is physically more acceptable, since the rest mass of an object that never stops isn't directly measurable). Tachyons would slow down if they lost energy, and accelerate if they gained energy. This leads to a problem in the case of charged tachyons because charged particles that move faster than the speed of light in the surrounding medium give off energy in the form of Cherenkov radiation. Charged tachyons would continuously lose energy, even in a vacuum, through Cherenkov emission. This would cause them to gain speed, thus lose energy at an even greater rate, thus accelerate even more, and so on, leading to a runaway reaction and the release of an arbitrarily large amount of energy.

More worrisome, as the physicist Gregory Benford and his colleagues first pointed out in their 1970 paper "The Tachyonic Antitelephone," tachyons seem to lead to a **time travel** paradox because of their ability to send messages into the past. Suppose Alice on Earth and Boole on a planet circling around Sirius can communicate using what has been called a tachyon "antitelephone." They agree in advance that when Boole receives a message from Alice, he will reply immediately. Alice promises to send a message to Boole at noon her time, if and only if she has not received a message from Boole by 10 A.M. The snag is that both messages, being superluminal, travel back in time. If Alice sends her message at noon, Boole's reply could reach her before 10 A.M. "Then," as Benford and colleagues wrote, "the exchange of messages will take place if and only if it does not take place. . . ." Perhaps not surprisingly, despite numerous searches, no tachyon detection has so far been confirmed. The same is true of another hypothetical faster-than-light particle called a dybbuk (Hebrew for a "roving spirit"), which would have imaginary mass, energy, and momentum. Dybbuks, proposed by Raymond Fox of the Israel Institute of Technology, have properties even stranger than those of tachyons yet, interestingly, they avoid the **causality** problem that affects their superluminal cousins.

TacTix

A two-player game of strategy, devised by Piet **Hein**, that is essentially a two-dimensional version of **Nim**. Though the game is nontrivial, the first player can always win, at least in the 5 × 5 matrix version by choosing the center piece and symmetrically mirroring the second player's moves. On a 5 × 5 grid, players alternate taking away as many contiguous pieces as desired from a single row or column.

Tafl game

A type of board game in which the contest is between two forces of unequal number or strength. The earliest form of Tafl (old Norse for "table"), known as **Hnefa-Tafl** ("king's table"), originated in Scandinavia before A.D. 400 and was then exported by the Vikings to Greenland, Iceland (where it is mentioned in the *Grettis Saga* dating back to A.D. 1300), Ireland, England, Wales, and as far east as the Ukraine. Several boards unearthed in both Viking and Anglo-Saxon contexts, including the board found at the Gokstad ship burial, have had

Hnefa-Tafl on one side and **nine men's morris** on the other. Later variants of Tafl include Tabula (the medieval ancestor of **backgammon**, introduced from the French as *Quatre* and thus *Kvatru-Tafl*), **fox and geese** (*Ref-Skak,* "fox chess", *Hala-Tafl* or *Freys-Tafl*), **three men's morris** (*Hrœ-Tafl,* "Quick-Tafl"), and **nine men's morris.**

Tait, Peter Guthrie (1831–1901)

A Scottish scientist and mathematician who carried out the world's first systematic investigation of **knot** theory. Early in his career he formed a friendship with William **Hamilton** and became fascinated in the application of Hamilton's **quaternion**s to problems in physics. In 1857, he also took an interest in Hermann Helmholtz's theories on the behavior of vortex rings, and began experimenting with smoke rings and their interactions. These experiments greatly impressed William Thomson (Lord Kelvin) who saw in them a possible way (wrong, as we now know) to explain atomic structure and the buildup of different elements. This idea, in turn, led Tait, Thomson, and James Maxwell to do seminal work on knot theory, since the basic building blocks in Thomson's vortex atom model were rings knotted in three dimensions. Without any rigorous theory, which would have been well beyond nineteenth-century mathematics, Tait began to classify knots using his geometric intuition. By 1877 he had classified all knots with seven crossings. He then went on to consider the coloring of **graph**s and put forward a hypothesis (see **Tait's conjecture**) that, if true (which it wasn't), would have proved the **four-color map problem**. Among his many other accomplishments, Tait wrote a classic paper on the trajectory of golf balls (1896). This was a subject close to his heart because the third of his four sons was Frederick Gutherie Tait, the leading amateur golfer in 1893 and winner of the Open Golf Championship in 1896 and 1898.

Tait's conjecture

A hypothesis put forward by Peter **Tait** in 1884, which says that every **polyhedron** has a **Hamilton circuit** through its vertices. In other words, it is possible to travel around all the edges of a polyhedron, passing through each vertex (corner) exactly once and arriving back at the starting point. If true, Tait's conjecture would have provided an immediate proof of the **four color map problem**. However, in 1946, the British mathematician William Tutte (1917–2002), whose work at Bletchley Park on cracking the German FISH cipher played an important role in World War II, found a counterexample to the conjecture in the form of a polygon with 25 faces, 69 edges, and 46 vertices.

tally

To count or keep score. In the past this was often done by making marks on a stick; the word comes from the Latin *talea* meaning "one who cuts," which is also the root of *tailor.* The oldest known tally stick is thought to be the **Lebombo bone** dating back about 37,000 years. Until around 1828 British tax records were kept on wooden tally sticks. When the system was finally abandoned, the government was left with a mountain of wood which, in 1834, it decided to dispose of by having a giant bonfire. So successful was that blaze that it also burned down the parliament buildings. What Guy Fawkes had failed to do with dynamite, the Exchequer did with tally sticks!

tangent

(1) A straight line that touches a given curve exactly once, at a given point. (2) In a right triangle, if one of the angles is θ, then the tangent of θ is the ratio of lengths of the side opposite θ to the side next to θ. See also **trigonometric function**.

tangle

A system in which a **strange loop** appears.

tangled graph

A **graph** in 3-dimensional space; equivalently, a graph drawn in the plane so that when edges cross, one edge goes over the other.

Tangloids

A mathematical game for two players devised by Piet **Hein** to model the calculus of **spinor**s. Two flat blocks of wood each pierced with three tiny holes are joined with three parallel strings. Each player holds one of the blocks of wood. The first player holds one block of wood still, while the other player rotates the other block of wood around any axis for 4π radians (two full revolutions). Then the first player tries to untangle the strings without rotating either piece of wood. Only translations (sliding the pieces) are allowed. Afterward, the players reverse roles; whoever can untangle the strings fastest is the winner.

tangrams

A puzzle of Chinese origin, the objective of which is to form given shapes using a set of seven pieces (five triangles of various sizes, one square, and one parallelogram) that come from slicing up a square. The produced shape has to contain all the pieces, which mustn't overlap. Tangrams became popular in England around the middle of the nineteenth century, having been brought back by sailors from Hong Kong. It received a further boost when Lewis **Carroll** used the pieces to create illustrations of the characters in the Alice books. The origin of the name

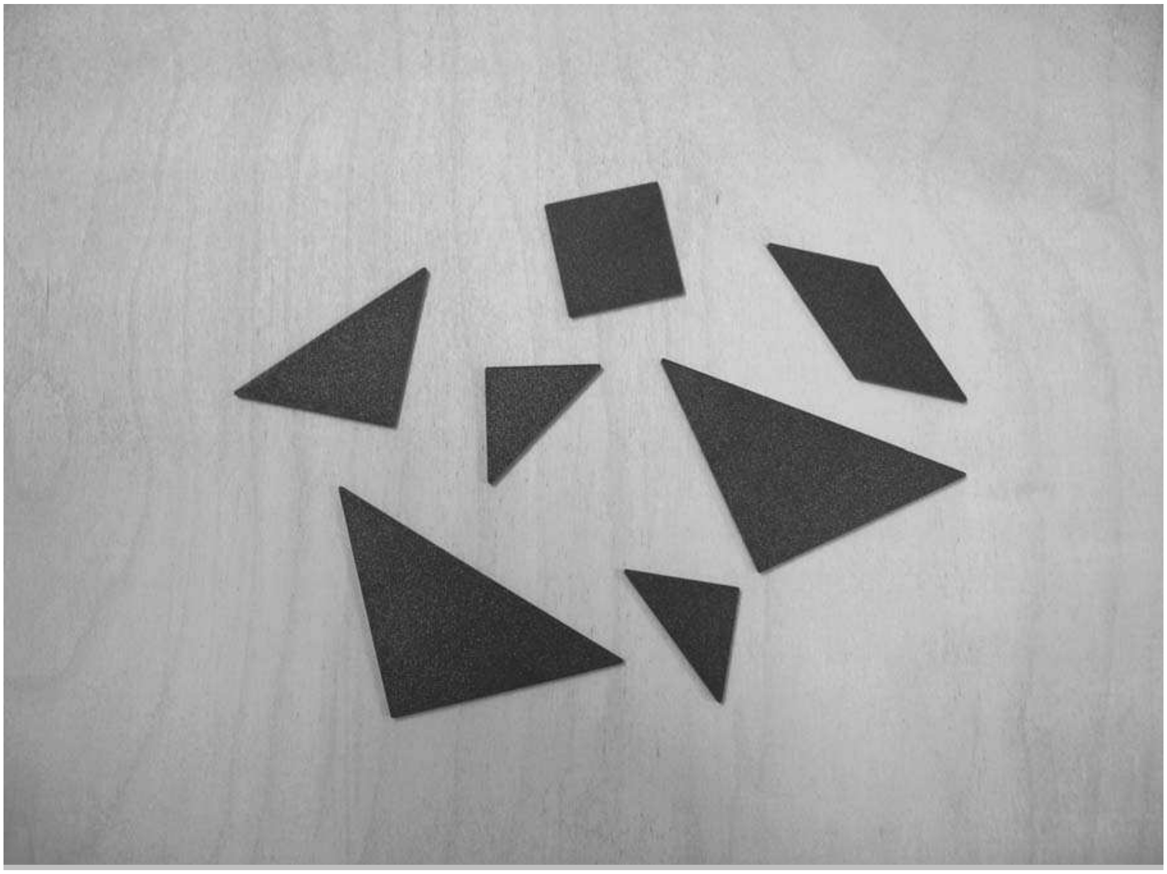

tangrams The seven tangrams. *Kadon Enterprises, Inc., www.gamepuzzles.com*

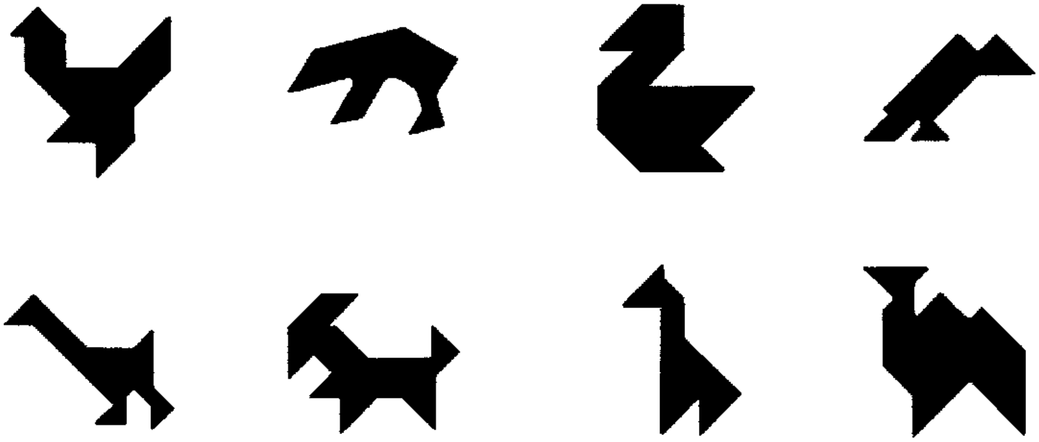

tangrams A tangram menagerie built up from the seven simple shapes.

isn't certain. One theory is that it comes from the Cantonese word for chin. A second is that it stems from a mispronunciation of a Chinese term that the sailors used for the ladies of the night from whom they learned the game! A third suggestion is that it is from the archaic Chinese root for seven, which still persists in the Tanabata festival held on July 7 in Japan. The first definitive appearance of tangrams in the Far East seems to be in Japan in the early eighteenth century; by around 1805 it had become a fad in both China and Europe. The **loculus of Archimedes** is a similar game and it has been suggested, though without evidence, that this was a direct precursor of tangrams, having been transmitted to the East via Arab sources.[93, 265]

Tarry, Gaston (1843–1913)

A French civil servant and amateur mathematician who spent the whole of his working career in Algeria. He published numerous articles on geometry, **number theory**, and **magic square**s from 1882 until his death, and is best known for his contribution to Euler's **thirty-six officers problem**. He also published an algorithm for exploring **maze**s that is named after him.

Tarski, Alfred (1902–1983)

A Polish-American mathematician who, along with Aristotle, Friedrich **Frege**, and Kurt **Gödel** is considered one of the greatest logicians of all time and certainly the most prolific: his collected works, excluding books, runs to 2,500 pages. Tarski made important contributions in many areas of mathematics, including **set theory**, **topology**, algebraic **logic**, and metamathematics. His most important contribution to logic is the *semantic method*–a technique that allows a more exacting study of formal scientific languages.

Tartaglia, Niccoló Fontana (1499–1557)

An Italian mathematician who, along with Girolamo **Cardano**, discovered the algebraic solution of the **cubic**. He was also a well-known inventor of mathematical recreations. He devised many arithmetical problems, and contributed especially to **measuring and weighing puzzles** and to **river-crossing problem**s. Although his real name was Niccoló Fontana, he is always referred to as Tartaglia, "the stammerer."

tautochrone problem

Find the curve down which an object can slide from any point to the bottom (accelerated by gravity and ignoring friction), always in the same length of time. *Tautochrone* comes from the Greek *tauto* for "the same" (which also gives us ***tautology***) and *chronos* for "time." The solution, first found by Christiaan **Huygens** and published in his *Horologium oscillatorium* (1673), is a **cycloid**. Thus, if you were to upturn a cycloid, in the manner of an inverted arch, and then release a marble from any point on it, it would take exactly the same time to reach the bottom, no matter where on the curve you started. Huygens used his discovery to design a more accurate pendulum–one with curved jaws from the point of support that forced the string to follow the right curve no matter how large or small the swing. The cycloid's unique property is mentioned in the following passage from Herman Melville's *Moby Dick:* "[The try-pot] is also a place for profound mathematical meditation. It was in the left-hand try-pot of the *Pequod,* with the soapstone diligently circling round me, that I was first indirectly struck by the remarkable fact, that in geometry all bodies gliding along a cycloid, my soapstone, for example, will descend from any point in precisely the same time." The cycloid is also the curve that answers the **brachistochrone problem**.

tautology

In mathematics, a logical statement in which the conclusion is equivalent to the premise. According to the school of thought known as *logicism,* all of mathematics is derived from **logic** and is thus inherently tautological. Tautology is also the needless, pointless, meaningless, and unwarranted repetition of words and phrases that mean the same thing. Examples are to be found in the previous sentence and the next one. Have a slice of pizza (*pizza* is Italian for "slice") and spin the **roulette** wheel (*roulette* is French for "little wheel").

ten

The **base** of our familiar **number system**, which stems directly from the fact that we have 10 fingers on which to count. Ten is the only **triangular number** that is a sum of consecutive odd squares ($10 = 1^2 + 3^2$) and the only composite integer such that all of its positive integer divisors other than 1 are of the form $x^2 + 1$ ($2 = 1^2 + 1$, $5 = 2^2 + 1$, $10 = 3^2 + 1$). Strange but true: the life span of a taste bud is 10 days.

tensor

A generalization of the concept of a **vector**. A *scalar* is a tensor of rank zero and a vector is a tensor of rank one. There are tensors of rank two, three, and so on, used mainly in manipulating and transforming sets of equations within and between different coordinate systems. A tensor of order n has n^2 components and is often best thought of as an array of values that can be manipulated like a **matrix**. To illustrate a tensor of rank two, imagine a plane surface area with a force acting on it. The total effect depends on two things: the magnitude and direction of the force, and the size of the area and its orientation. In fact this latter property can be represented uniquely by a vector of magnitude proportional to the size of the area, and in a direction normal to the area. So the effect of the force upon the surface depends on two vectors and is equivalent to a tensor of rank two. Tensors were used by Einstein in deriving his law of gravitation in general **relativity theory**.

tera-

Prefix for 10^{15}, from the Greek *pentakis* for "five times."

terminating decimal

A decimal fraction that comes to an end, as in 0.3 or 0.7194. All terminating decimals are **rational number**s, but not all rational numbers have terminating decimal expansions. For example, ⅓ is rational, but its decimal expansion goes on forever.

ternary

(1) Having the **base** three. (2) Involving three **variable**s. From the Latin *ternarius* ("three each").

tessellation

See **tiling**.

tesseract

The four-dimensional analogue of a cube, also known as a 4-space **hypercube** or an 8-cell. The name, possibly coined by Charles **Hinton**, comes from *tesser,* meaning "four," and *aktis,* meaning "ray," thus "four rays." Just as a cube is obtained by "thickening" a square in the third dimension, which can be imagined as stacking infinitely many infinitely thin sheets of paper, a tesseract is a cube thickened in the **fourth dimension**. We can't imagine this because we can't think four dimensionally but it is possible to appreciate that just as perspectives of cubes can be drawn on a 2-d surface, so real cubes can serve as perspectives of tesseracts. A square drawn inside a larger square with the vertices connected by lines is one way to provide a perspective of a cube. Similarly, a hypercube is sometimes portrayed as a small cube within a larger cube with lines drawn from the vertices of the smaller cube to the vertices of the larger cube. This kind of representation is a bit misleading, however, and reveals very little of the nature of a tesseract. It doesn't show, for instance, how a tesseract can be subdivided into smaller 4-d blocks in the same way that a cube can be divided into smaller cubes, or a square into smaller squares. A more useful way to think of a tesseract is as a folding, in the fourth dimension, of a 3-d net of eight cubes, just as a cube is a folding in the third dimension of a 2-d net of six squares. Start with a stack of four cubes, with four more cubes arranged in a cross around the second cube from the top. A tesseract is made by folding (in the fourth dimension) so that the top face of the cube at the top of the stack merges with the bottom face of the bottom cube, and so that the adjacent edges of the cubes in the cross join. (See table, "A Comparison of the Square, Cube, and Tesseract.")

A Comparison of the Square, Cube, and Tesseract

	Vertices	Edges	Squares	Cubes
Square	4	4	1	–
Cube	8	12	6	1
Tesseract	16	32	24	8

A tesseract is bounded by eight hyperplanes, each of which intersects it to form a cube. Two cubes, and so three squares, intersect at each edge. There are three cubes meeting at every vertex, the vertex polyhedron of which is a regular tetrahedron, leading to the **Schläfli symbol** {4,3,3}. The distance between opposite corners of a hypercube is twice the length of a side–much tidier than the corresponding values of $\sqrt{2}$ for a square and $\sqrt{3}$ for a cube.

If a cube is hung from one of its vertices and sliced horizontally through its center, the result is a hexagon. What if the same is done to a tesseract? The slice will yield a 3-dimensional object–but what kind? The answer is an octahedron. By analogy with the slice of the 3-cube, the slice of the 4-cube must cut every "face." The number of faces of a 4-cube is eight and the only regular eight-sided solid is an octahedron.

Tesseracts turn up in both art and literature. Salvador Dali's *Christus Hypercubus* shows Christ being crucified on a tesseract. In Robert Heinlein's short story "And He Built a Crooked House" (1940), a house built as a three-dimensional projection of a tesseract collapses, or folds up, to become a real tesseract–with unusual consequences for the person trapped inside. The tesseract is also mentioned in Madeleine l'Engles children's fantasy *A Wrinkle in Time* as a way of introducing the concept of higher dimensions.[161, 338]

tetragon

The less familiar name for what's normally called a **quadrilateral**.

tetrahedral number

A number that can be made by considering a tetrahedral pattern of beads in three dimensions. For example, if a tri-

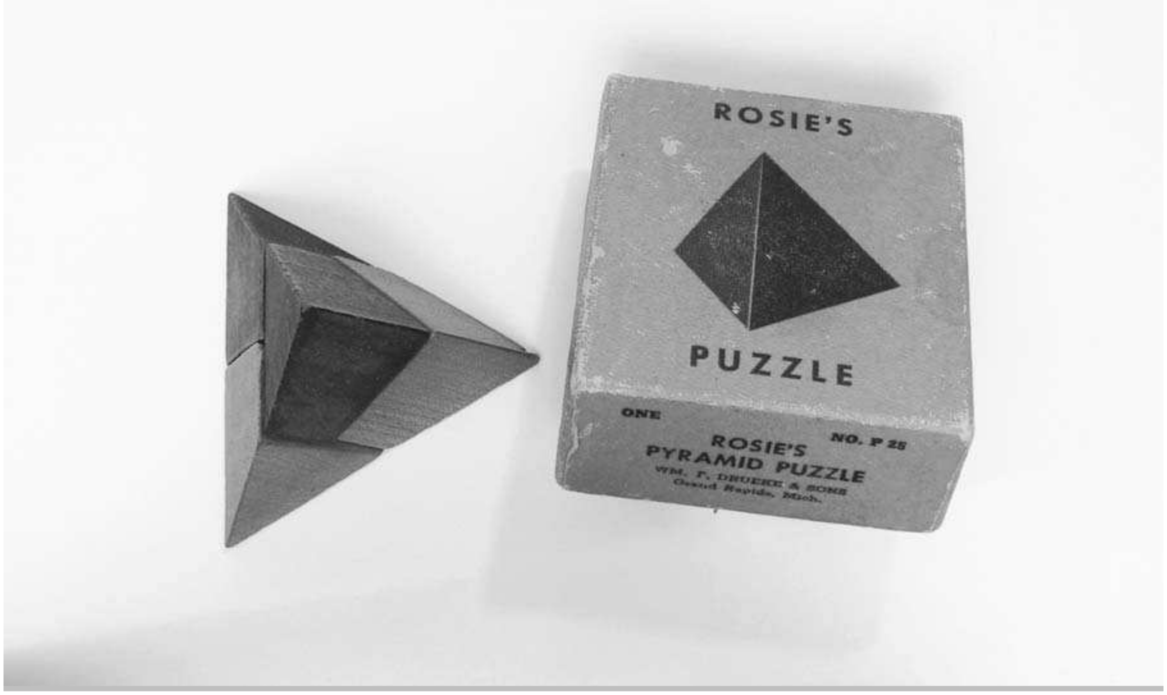

tetrahedron Rosie's pyramid puzzle, a vintage game that calls for four pieces to be assembled into a tetrahedron. *Sue & Brian Young/Mr. Puzzle Australia, www.mrpuzzle.com.au*

angle of beads is made with three beads to a side, and on top of this is placed a triangle with two beads to a side, and on top of that a triangle with one bead to a side, the result is a tetrahedron of beads. In this case the total number of beads is (third triangular number) + (second triangular number) + (first triangular number) = 6 + 3 + 1 = 10. In general the *n*th tetrahedral number is equal to the sum of the first *n* triangular numbers. This is the same as the fourth number from the left in the $(n + 3)$th row of **Pascal's triangle**. We can use the **binomial** formula for numbers in Pascal's triangle to show that the *n*th tetrahedral number is ${}^{n+2}C_3$, or $(n + 2)(n + 1)n/6$. The only numbers that are both tetrahedral and square are 4 ($= 2^2 = T_2$) and 19,600 ($= 140^2 = T_{48}$).

tetrahedron

A four-sided **polyhedron**. A regular tetrahedron, one of the **Platonic solid**s, is a regular three-sided pyramid in which the base-edges and side-edges are of equal length. The projection of a regular tetrahedron can be an equilateral triangle or a square. The centers of the faces of a tetrahedron form another tetrahedron.

tetraktys

The sum $1 + 2 + 3 + 4 = 10$ (the fourth **triangular number**), held in reverence by the Pythagoreans (see **Pythagoras of Samos**). The tetraktus, or "holy fourfoldness," was taken to represent the four elements: fire, water, air, and earth.

Tetris

A video and computer game, invented in 1985 by the Russian Alexey Pajitnov, that has become one of the most widely played games of all time. In 2002, computer scientists Erik Demaine, Susan Hohenberger, and David Liben-Nowell of the Massachusetts Institute of Technology (MIT) analyzed the game to determine its computational complexity and found it to be an **NP-hard problem** (one that is immune to simple solution and instead demands exhaustive analysis to work out the best way to be completed). Many people first played Tetris on the Nintendo Gameboy handheld console but it has since become available for virtually every personal computer–based device. The game gives the player the task of creating complete lines from a series of regularly shaped blocks–tetrominos, which are a type of **polyomino**–that advance steadily

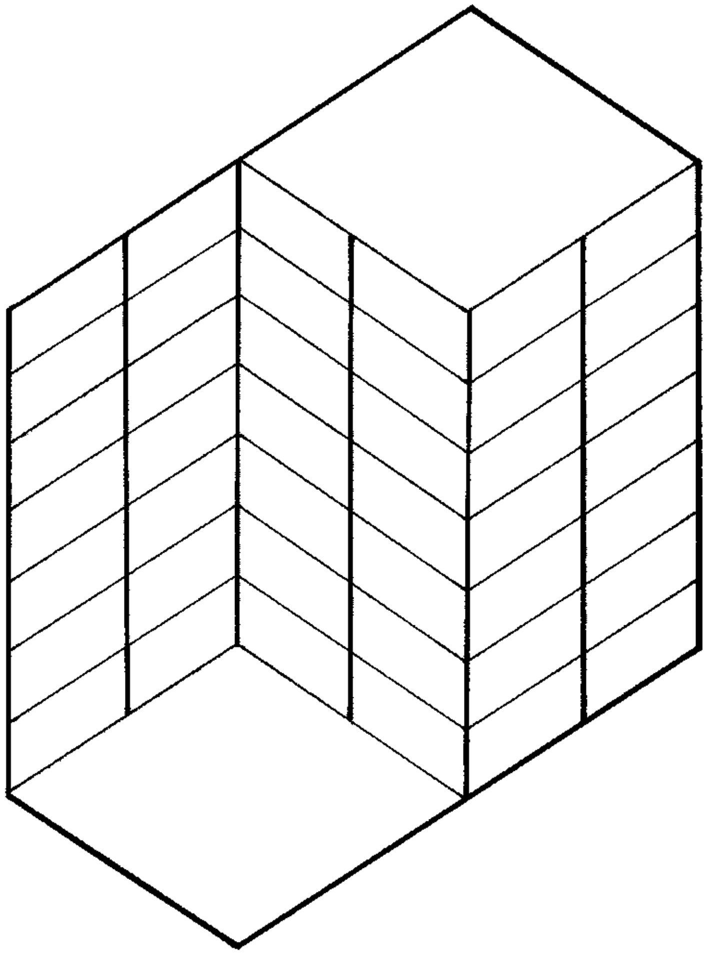

Thiery figure

down a narrow grid. The blocks can be spun to make them fit together better and complete lines. The game gets faster as levels are completed, making it harder to spin and fit blocks together fast enough to form lines. The MIT team found that, subject to certain conditions, Tetris has much in common with some of the knottiest mathematical conundrums, including the **traveling salesman problem**. Because Tetris is NP-hard, there is no easy way to maximize a score for the game, even when the sequence of blocks is known in advance.

tetromino

A four-square **polyomino**.

theorem

A major mathematical proposition that has been proved correct. More precisely, a statement in a **formal system** for which there exists a **proof**. See also **conjecture** and **lemma**.

theory of everything

A unified theory of all fundamental forces and interactions in nature; a grand unified theory that also includes gravity or general relativity (see **relativity theory**).

Thiery figure

A classic **ambiguous figure** devised by the psychologist A. Thiery in the late nineteenth century.

thirteen

The unluckiest of numbers if you happen to be superstitious. This belief has a couple of historical roots. According to biblical tradition, there were 13 people at Christ's Last Supper, and Christ was crucified on a Friday the thirteenth. Further back in time, Alexander the Great decided he wanted to be the thirteenth god alongside the 12 that already stood for each month of the year, so he had a thirteenth statue built on the place of his capital. His death shortly after gave the number a bad name. Many buildings don't have a floor labeled 13 and many hotels will have a room numbered 12A instead of 13. There is even a name for a morbid fear of 13: triskaidekaphobia.

Fresh disasters involving the number hardly help triskaidekaphobics overcome their affliction. The most notorious of these involved the Apollo 13 Moon mission, which was launched on April 11, 1970 (the sum of 4, 11, and 70 equals 85, the digital sum of which is 13), from Pad 39 (3 × 13) at 13:13 local time, and suffered an explosion on April 13. (The astronauts did, however, make it home safely, which could be considered good luck.) There is always at least one Friday the thirteenth in each year; in some years there are two, and rarely three (e.g., 1998 and 2009). There were 13 original U.S. colonies (hence the 13 stripes on the American flag) and 13 signers of the Declaration of Independence. In Japan, the numbers 4 and 9 are considered unlucky, not because 13 can be represented as sum of these two perfect squares but because of their pronunciation. In Japanese, four is *shi,* which is pronounced the same as the word for death; nine is *ku,* which sounds like the word for torture. And speaking of torture, it was not unusual in times past for bakers to come in for stiff punishment if they short-changed their customers. In ancient Egypt, someone found selling light loaves might end up with his ear nailed to a doorpost, while in medieval Britain the punishment was likely to be a spell in the pillory. This led to the custom of adding a thirteenth loaf to every batch of 12 to be on the safe side, and hence the expression "a baker's dozen."

Mathematically, the reverse of the square of 13 is the same as the square of the reverse of 13: $13^2 = 169$; the reverse of 169 is 961 and the reverse of 13 is 31; $31^2 = 961$. Thirteen is the smallest **prime number** that can be expressed as the sum of the squares of two prime numbers: $13 = 2^2 + 3^2$. Also the sum of all prime numbers up to 13 ($2 + 3 + 5 + 7 + 11 + 13$) is equal to the thirteenth prime number (41), and this is the largest such number. On the anagrammatical front is this nice equation:

ELEVEN + TWO = TWELVE + ONE

See also **dollar**.

thirty colored cubes puzzle

A game devised in 1921 by Percy **MacMahon**, which was marketed under the name "Mayblox." It is played with 30 cubes that have all possible permutations of six different colors on their faces. A number of different games can be played with the blocks. One is to choose a cube at random and then choose seven other cubes to make a 2 × 2 × 2 cube with the same arrangement of colors for its faces as the first chosen cube. Each face of the 2 × 2 × 2 cube has to be a single color and the interior faces have to match in color. After building one such cube, only one other with the same properties can by made from the remaining 22 cubes–the mirror image of the first.

thirty-six officers problem

Arrange 36 officers in a 6 × 6 square so that one officer from each of six regiments appears in each row and one from each of six ranks appears in each column. This problem, first posed by Leonhard **Euler** in 1779, is equivalent to finding two mutually orthogonal **Latin square**s of order six. Euler correctly conjectured that there was no solution; the search for a proof led to important developments in **combinatorics**.

Thompson, D'Arcy Wentworth (1860–1948)

A Scottish naturalist and polymath. Thompson held a professorial chair at St. Andrews and Dundee, Scotland, for the amazing period of 64 years, a record for tenure unlikely ever to be broken. Although he wrote more than 300 scientific articles and books, his reputation is based primarily on his efforts to reduce biological phenomena to mathematics in his magnum opus *On Growth and Form* (1917).[332] In this book, full of marvelous sketches of such things as nautilus shells and honeycombs, Thompson claimed that much about animals and plants could be understood by the laws of physics, as mirrored in the

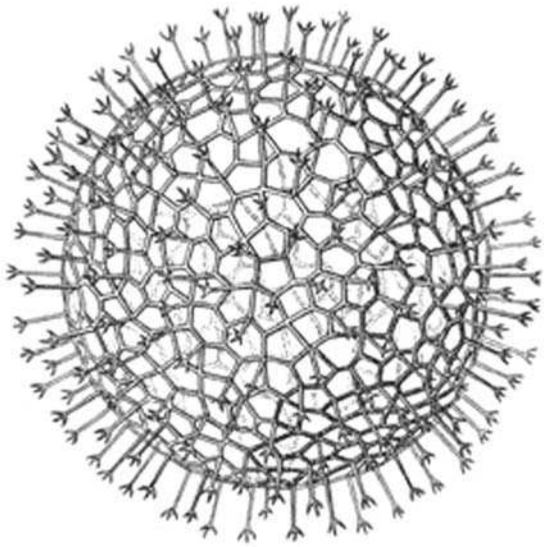

Thompson, D'Arcy Wentworth An illustration from Thompson's *On Growth and Form:* the shell of the radiolarian Aulastrum triceros.

structures and patterns of mathematics. His most novel idea was to show how mathematical functions could be applied to the shape of one organism to continuously transform it into other, physically similar organisms. One memorable example is the squeezing and stretching of a rectangular Cartesian grid that transforms the fish species *Scarus sp.* to the species *Pomacanthus.* Thompson used the same principle to transform skulls of baboons into those of other primates, and to show how corresponding bones like the shoulder blade are related in different species. Doubtless, if he were alive today, he would be heavily into "morphing"–the digital technique that allows a computer to do exactly the same kind of transformation of one object to another, but vastly more efficiently. Thompson acquired a local reputation as a mild eccentric; indeed, older inhabitants of St. Andrews still recall seeing him strolling about town with a parrot on his shoulder.

thousand

A number that, linguistically, comes from an extension of "hundred" and has roots in the Germanic *teue* and *hundt. Teue* refers to a thickening or swelling, and *hundt* is the source of our present-day hundred. A thousand, then, literally means a swollen or large hundred. The root *teue* is the basis of such common words today as thigh, thumb, tumor, and tuber. One thousand is the answer to the question: If you were to spell out numbers, how far would you have to go until you would find the letter "A"? While Americans may say "Thanks a million" to express gratitude, Norwegians offer "Thanks a thousand" ("*tusen takk*").

three

> *1 + 1 = 3, for large values of 1.*
>
> –Anonymous

The number of dimensions of space in which we live; three is also the smallest odd **prime number**, the second **triangular number**, and member of the **Fibonacci sequence**. Three is often the number of repetitions in jokes and children's stories (for example, the tale of the *Three Little Pigs* and of the number of chairs, beds, and porridge-eating bears in *Goldilocks*), because it is the minimum number needed to establish a pattern (such as a regular tempo) or to convey the impression of an ongoing sequence or succession. In the Christian tradition, three plays a crucial role: Christ represents one third of the Trinity (the Father, Son, and Holy Spirit), was visited by the three wise men and, 33 years later, when Peter disowned him three times, rose on the third day after the crucifixion, having died at 3 P.M. Other things that come in threes: musketeers, primary colors, wishes, blind mice, bad luck, and London buses. See also **Triangular Lodge** and **tinner's rabbits**.

three-hat problem

See **hat problem**.

three men's morris

An old game played on a 3 × 3 board that is thought to be a direct ancestor of **tic-tac-toe**; it is known by many other names, including nine holes, and is related to six men's morris and **nine men's morris**. The game involves two differently colored sets of four pieces (one set for each player). Players take turns placing pieces on intersection points, and the first person to place three along a line wins the game. The earliest known board for three men's morris was found on the roof of the temple in Kurna, Egypt, dating back almost three and a half thousand years. Its earliest known appearance in literature is in Ovid's *Ars Amatoria.* The Chinese are believed to have played it under the name *Luk tsut K'i* during the time of Confucius (c. 500 B.C.). Boards for three men's morris dating back to the thirteenth century can be found carved into the cloister seats at the cathedrals at Canterbury, Gloucester, Norwich, and Salisbury, and at Westminster Abbey.

three-body problem

The problem of determining the future positions and velocities of three gravitational bodies. The problem was proved unsolvable in the general case by Henri **Poincaré**, which foreshadowed the importance of **chaos**. Although no analytical solutions are possible in the worst case, a numerical solution is sometimes sufficient for many tasks.

Thue-Morse constant

Also known as the *parity constant,* the number defined as follows: Take a string of 1's and 0's and follow it by its complement (the same string with 1's switched to 0's and vice versa) to give a string twice as long. Repeat this process forever (starting with 0 as the initial string) to get the sequence 01101001100101101001011 0 . . . . Make this a binary fraction, $0.0110100110010110\ldots_2$, and rewrite it in base 10. The resulting **transcendental number**, 0.41245403364 . . . , is the Thue-Morse constant.

tic-tac-toe

Also known as noughts and crosses, and spelled in a variety of ways (such as ticktacktoe), this well-known pastime is not as ancient as popularly believed, though it certainly has roots in older games. The earliest clear description of the rules, but without a name, comes from Charles **Babbage** around 1820. Later on, Babbage started to call the game tit tat to and by slight variants on this, gave the first detailed analysis of it, and designed the first robot to play it! He speculated that "[T]he machine might consist of the figures of two children playing against each other, accompanied by a lamb and a cock. That the child who won the game might clap his hands whilst the cock was crowing, after which, that the child who was beaten might cry and wring his hands whilst the lamb was bleating."

His plan was to exhibit the machine in London to raise money for his more serious projects, including the fabulous Analytical Engine. However, on hearing that similar devices, such as a mechanical composer of Latin verse, had flopped financially he abandoned the scheme.

Two expert players of tic-tac-toe will always draw. In other words, there is no winning strategy against an opponent who knows the game well. Nor is it difficult to become unbeatable. Despite the fact that there are 9 × 8 × 7 × 6 × 5 (= 15,120) possible layouts of noughts and crosses for the first five moves alone, these reduce to just a few basic patterns of play and counterplay that ensure the canny contestant never loses. The game is only interesting when at least one novice is involved. The strongest opening play is in a corner because an unwary opponent will be trapped unless her countermove is in the center. By the same token, a center opening must be countered by a corner response or the first player has an easy win. Of course, a master player will not only never lose but will learn his opponent's weaknesses and exploit them in the most devastating way.

The game also becomes more interesting when played on a larger board and/or in more dimensions. An exception to this is the 3 × 3 × 3 cube which gives an easy win to the first player; in fact, it cannot end in a draw because the first player has 14 plays and it is impossible to make all of these without scoring. The 4 × 4 × 4 cube is much more interesting and is sold commercially as the well-known Score Four game. For the ambitious, tic-tac-toe can be played in four dimensions on a **tesseract** by sectioning it into two-dimensional squares. Older, similar, and mathematically more interesting relatives of tic-tac-toe include **nine men's morris** and **Ovid's game**.

tie knots

How many ways can you tie a tie? For many years there were just three styles of **knot**: the Four-in-Hand, the Windsor, and the Half-Windsor. Then the Pratt was introduced to the world on the cover of the *New York Times* in 1989. Intrigued that only one new knot had been added to the tie-tying repertoire in more than half a century, two researchers from the University of Cambridge's physics department, Thomas Fink and Yong Mao, decided to see how many tie knots were actually possible. To this end, they applied **random walk** theory–a technique useful for describing movement which, although unpredictable in detail, reveals large-scale patterns. Such patterns, the researchers realized, were essential to a successfully accomplished tie-knot. For example, if the end of the tie is moved to the right, its next move can't be to the right again–it has to be either to the left or to the center. This means that each move made in tying a tie can only be followed by one of two alternatives. Fink and Mao found that the simplest possible knot involves just three moves. They went on to discover 85 possible tie-knots, including the four popular knots, six new knots that they consider aesthetically pleasing, and two complicated nine-move knots.[100]

tiling

Also called a *tesselation,* a collection of smaller shapes that precisely covers a larger shape, without any gaps or overlaps. Usually, the shape to be tiled is a flat plane but other shapes and three-dimensional objects can be tiled, too. In a game that involves tiling, certain conditions are applied; for example, all the tiles may have to be identical, or they may all have to be squares but every one of a different size. It's been known for some time that all simple regular tilings in the plane belong to one of the 17 plane symmetry groups known as **wallpaper group**s. All 17 of these patterns are known to exist in the **Alhambra** palace. This doesn't exhaust the apparently simple problem of tiling the plane: adding extra constraints or removing the requirement for regularity leads to a large number of interesting problems. These include *alternating tilings,* for examples of squares or dominoes, such that two tiles have a side or a part of a side in common, or *colored tilings,* in which no two adjacent tiles have the same color. Colored tilings are also called *colored maps.* The most famous problem relating to colored tilings is the **four-color map problem**, which has been solved. Other problems involve *n*-tesselations, in which each tile has an integral area and for each natural number *n* there is exactly one tile with area *n.* See also **Penrose tiling**, **rep-tile**, **rectangle**, **squaring the square**, **dissection**, and **packing**.

time

> *Time is a great teacher. Unfortunately, it kills all its pupils.*
>
> –Hector Berlioz (1803–1869)

One of the most familiar and yet mysterious properties of the universe. The "flow" of time is one of the strongest impressions we have, yet it may simply be

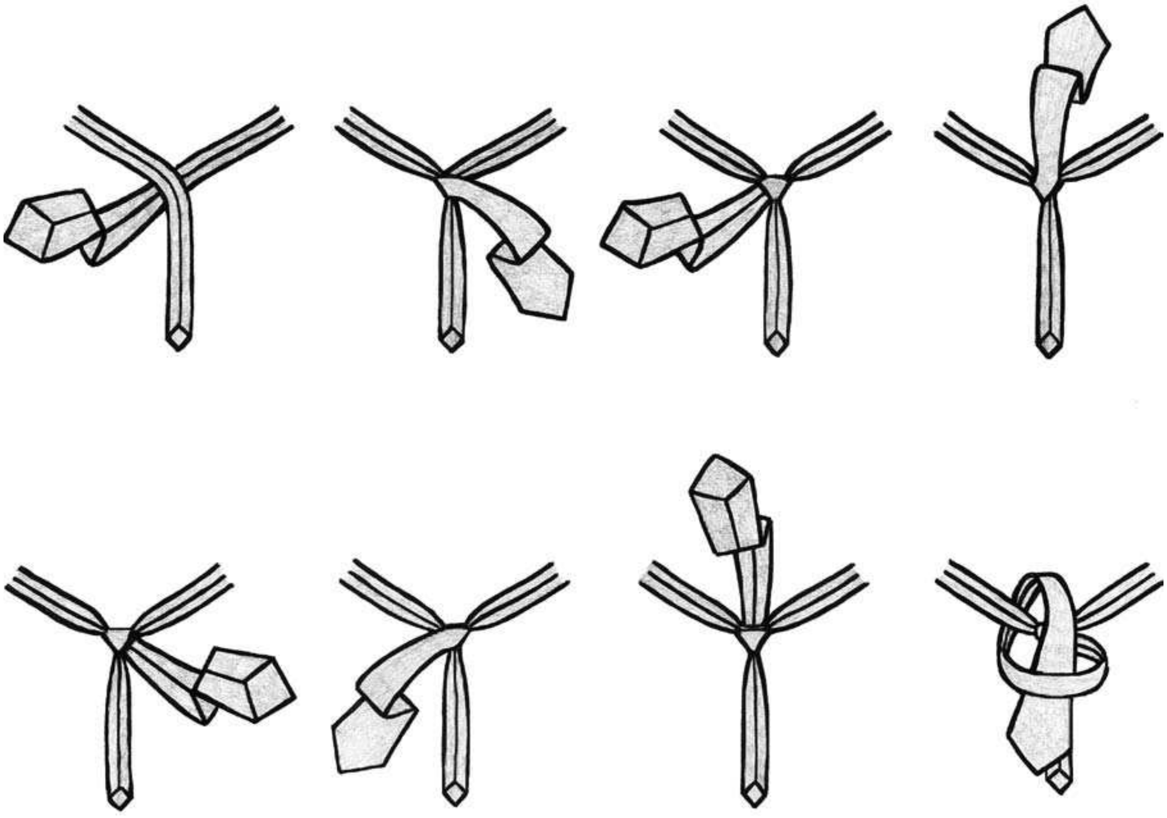

tie knots The "St. Andrew": one of several new tie knots recently discovered. *Thomas Fink*

an illusion or a product of the conscious mind. The very notion that time somehow moves leads to a logical paradox because, as the Australian philosopher J. J. C. Smart asked: "In what units is the rate of time flow to be measured? Seconds per ____?" John Dunne in his classic book *An Experiment with Time,* argued that the human mind has the ability to rove back and forth along the time line so that precognition is a physical possibility. However his theory involves an infinite regress of time and of the observer that is philosophically hard to swallow. All the same, it may be that the apparent movement from past to present to future has less to do with the universe at large than it has to do with our individual subjective experience. In some way, still to be fathomed, time, consciousness, free will, and the individual are intimately entwined. In physics, by contrast, time is treated no differently (with one important exception, noted below) than space. It is simply another **dimension**–another axis, or extension, of physical reality. Just as the various spatial dimensions prevent everything from happening at a single point, so time prevents everything from happening all at once. As one wag put it, "Time is just one damned thing after another!" In Einstein's **relativity theory**, time is effectively "spatialized" so that, instead of speaking of an absolute three-dimensional space and a separate one-dimensional time, there is a four-dimensional **space-time** continuum. So closely related are time and space in relativity theory that time can be converted into space and vice versa. In particular, different observers may not agree on the distance or the duration between any two events in space-time, but they will always agree on the *space-time interval.* If the two points events occur at (t, x, y, z) and $(t + dt, x + dx, y + dy, z + dz)$, then the (constant) space-time interval between them is given by

$$s^2 = c^2(t_2^2 - t_1^2) - (x_2^2 - x_1^2) - (y_2^2 - y_1^2) - (z_2^2 - z_1^2).$$

But the time of relativity, like that of classical physics, remains reversible.

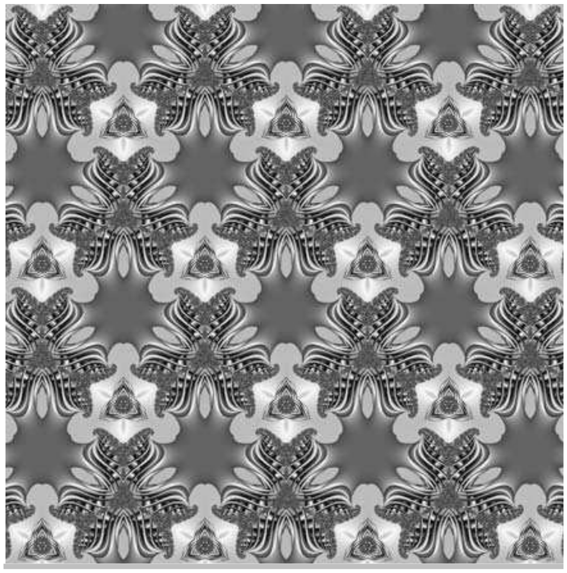

tiling A computer-generated periodic tiling. *Jos Leys, www.josleys.com*

Chrononauts
Chrononauts Official Rules
MISSION
ID
1912
Titanic Avoids Iceberg
Linchpin

time travel The game of Chrononauts, in which players must resolve a series of paradoxes that threaten the time line. *Looney Labs, www.LooneyLabs.com*

time complexity

A **function** that describes the amount of time required for a program to run on a computer to perform a particular task. The function is parameterized by the length of the program's input.

time dilation

See **relativity theory**.

time-reversible

A property of **dynamical system**s that can be run unambiguously both forward and backward in time. The **Lorenz system**, for example, is time-reversible.

time travel

Can we travel through **time**? Of course, we do it all the time! But can we do it at a different rate than normal? Again, the answer is "yes" because of the phenomenon known as *time dilation* in Einstein's **relativity theory**. However, time dilation enables, even in principle, only a limited kind of leap into the future–one from which we cannot return to the present. Genuine time travel is the ability to jump forward or backward through time at a rate other than that of the ordinary progression of events or that enabled by the relativistic time dilation effect.

The possibility of traveling through time poses such a threat to **causality** and opens the door to so many disturbing paradoxes that many scientists feel inclined to dismiss it out of hand. However, it has been a favorite theme of science fiction since the 1880s. In *The Time Machine* (1895), H. G. Wells gives a pleasant preamble about the nature of the **fourth dimension** before whisking his hero 802,000 years into the future. Says the Time Traveller (we never learn his real name), "[A]ny real body must have extension in four directions: it must have Length, Breadth, Thickness, and–Duration. There are really four dimensions, three of which we call the three planes of Space, and a fourth, Time. There is, however, a tendency to draw an unreal distinction between the former three dimensions and the latter, because it happens that our consciousness moves intermittently in one direction along the latter from the beginning to the end of our lives."

Unfortunately for would-be chrononauts (an early version of *The Time Machine* was called *The Chronic Astronauts*), Wells is not specific about how his time traveling device works, though we know that "Parts were of nickel, parts of ivory, parts had certainly been filed or sawn out of rock crystal." In more recent times, physicists, speculating on some of the more esoteric byways of relativity and **quantum mechanics**, have been a little more forthcoming about how time travel might be achieved in practice. These speculations have variously involved wormholes (shortcuts outside of normal space and time), faster-than-light particles known as **tachyon**s, and unusual cosmological models, such as the ***Gödel** universe,* which allow movement to any point in the future or the past. Let us leave aside the practical aspects, however, and focus on the logic of breaking the time barrier.

The various time travel possibilities dealt with in science fiction fall into two broad categories. In the first the timeline, from deepest past to darkest future, is frozen and immutable, like a filmstrip. Any time traveling that takes place is constrained by this preordained structure–effectively, already written into the narrative of the world (the "block universe" of Einsteinian physics)–and is thus prevented from leading to paradoxes. In one variant of this scenario, the *Novikov self-consistency principle* applies. Named after Igor Novikov, an astrophysicist at Copenhagen University, this asserts that any attempt at time travel that would lead to a paradox, such as the **grandfather paradox**, is bound to fail even if the cause of failure is an extremely improbable event. In other words, try as you might to introduce a contradiction into the timeline, like killing yourself or one of your ancestors in the past, circumstances will always conspire to prevent you. An excellent example of this type of universe is found in Robert L. Forward's novel *Timemaster.* Another variant on the fixed timeline concept is that any event that appears to have caused a paradox has, in fact, created a new timeline. The old timeline remains unaltered, and the time traveler becomes part of a new temporal branch line. One difficulty with this arrangement is that it might violate the principle of conservation of mass-energy, unless the mechanics of time travel demand that mass-energy be exchanged in precise balance between past and future at the moment of travel. However, the concept of branching universes and alternative histories is not outrageous in physics where the **many worlds hypothesis** and of Feynmann's sum-over-histories are routinely debated.

The second main type of time travel entertained in science fiction assumes that the timeline is flexible and changeable. This can lead to all sorts of mind-boggling difficulties and contradictions. A way to offset some of these problems is to stipulate that the timeline is very resistant to change. In the extreme case, as writer Larry Niven has argued, it may be a fundamental rule that in any universe where time travel is allowed, no actual time machine is ever invented. The English physicist and mathematician Stephen Hawking put this idea on a more formal footing with his *chronology protection conjecture.* On the other hand if the timeline is presumed to be easily changed, paradoxes threaten to spring up at every turn. One of the most remarkable of these is the *closed causal curve paradox* in which, it seems, something can be gotten for nothing. Samuel Mines summarized the plot of his 1946 short story as follows:

> A scientist builds a time machine, goes 500 years into the future. He finds a statue of himself commemorating the first time traveler. He brings it back to his own time and it is subsequently set up in his honor. You see the catch here? It had to be set up in his own time so that it would be there waiting for him when he went into the future to find it. He had to go into the future to bring it back so it could be set up in his own time. Somewhere a piece of the cycle is missing. When was the statue made?

Closed loops in time can also conjure knowledge out of thin air. A man builds a time machine and travels into the past to give the plans for the device to his younger self who then builds the machine, travels into the past, and so on. Where did the plans originate? A curious thing about time loops is that they have no easily discernible future or past because all the events taking place in them affect one another in a circular way. Time loops also put a question mark over free will. What happens if the younger man, given the time machine plans by his older self, decides not to build the device? *Can* he make that choice given that, in some sense, he has already built it? Perhaps the apparent absence of time travelers and time machines in the real world is a sign that we do not have to worry about such issues–at least, for the present.

tinner's rabbits

A name that has emerged recently to describe a pattern of three rabbits or hares that has been found in many parts of the world, including England and Wales, mainland Europe, China, and Russia. It occurs, for example, on the medieval roof bosses of some churches in Devon and Cornwall and is thought to be connected with the local tin-mining industry. One theory suggests the following link: Tin is alloyed with copper to make bronze, copper came from Cyprus (the words *Cyprus* and *copper* have the same root), Cyprus is the island of the goddess Venus or Aphrodite (she was born there), rabbits are symbols of Venus. Three intertwined fishes are a common Christian symbol, so the three rabbits may also have stood for the Trinity.

Tippee Top

Also known as a *Tippy Top,* a type of top, patented in Britain in 1953, that consists of a peg with a ball-shaped body. If the top is spun quickly on the rounded body, with the peg pointing upward, it flips itself over and spins on its peg. Crucial to this counterintuitive behavior is the shape of the body, which is smooth and spheroidal with no sharp point. After release, the top, like tops of every description, begins to show precession, that is, its axis of rotation moves in a small circle. After a while, the contact point of the top with the table no longer coincides with the rotation axis but instead moves to other points of the top's head. Due to frictional forces and precession, the top seeks a more stable position, which it finds by flipping on to its stalk. See also **celt**.

Titchener illusion

Also known as the *Ebbinghaus size illusion,* a well-known **distortion illusion**. Two circles are surrounded by either six big circles or six small circles. Despite appearances, the two center circles are exactly the same size.

Toeplitz matrix

A **matrix** in which all the elements are the same along any **diagonal** that slopes from northwest to southeast.

topological group

Also called a *continuous group,* a **set** that has both the structure of a **group** and of a **topological space** in such a way that the operations defining the group structure give continuous maps in the topological structure. Many groups of matrices (see **matrix**) give topological groups.

topological dimension

An integer that defines the number of coordinates needed to specify a given point of an object of set *X*. A single point, therefore, has a topological dimension equal to zero; a curve has dimension one, a surface has dimension two, and so on.

topological space

A type of generalized mathematical space in which the idea of closeness, or limits, is described in terms of relationships between **set**s rather than in terms of distance. Every topological space consists of: (1) a set of points; (2) a class of subsets defined axiomatically as open sets; and (3) the set operations of union and intersection.

topology

The study of those properties of mathematical objects that remain unaffected by smooth deformations, such as stretching and squeezing, but that don't involve tearing.

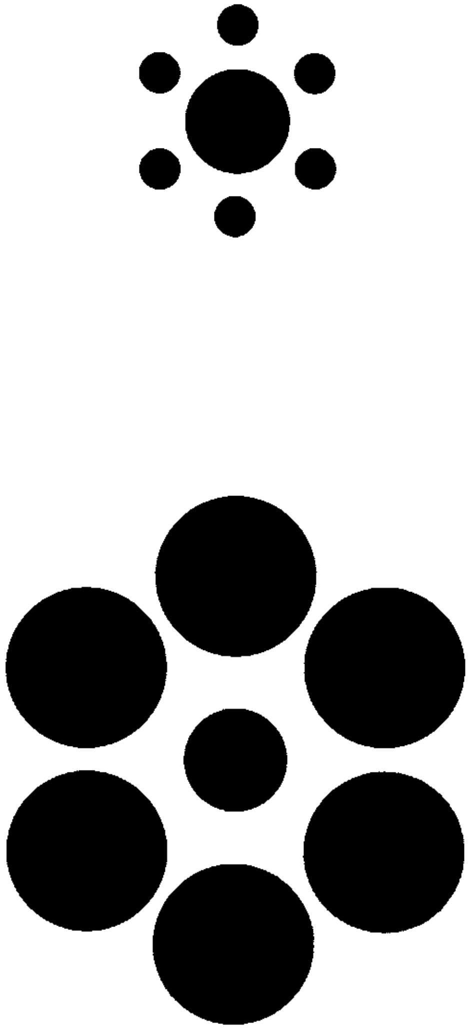

Titchener illusion The center circles are the same size.

The word comes from the Greek *topos* for "place," and was introduced into English by Solomon Lefschetz in the late 1920s. A topologist has been described as someone who doesn't know the difference between a doughnut and a coffee cup. Substitute "care about" for "know" and this becomes more accurate. Imagine a donut made of soft clay. A potter can easily shape this into a cup with a handle without removing or creating any new holes. Both shapes, in topology, are said to be genus 1–objects with a single hole. A sphere, by contrast, is genus 0 (no holes), while an eyeglass frame, with the lenses removed, is genus 2. For more on topologically intriguing structures, see **Möbius band** and **Klein bottle**.

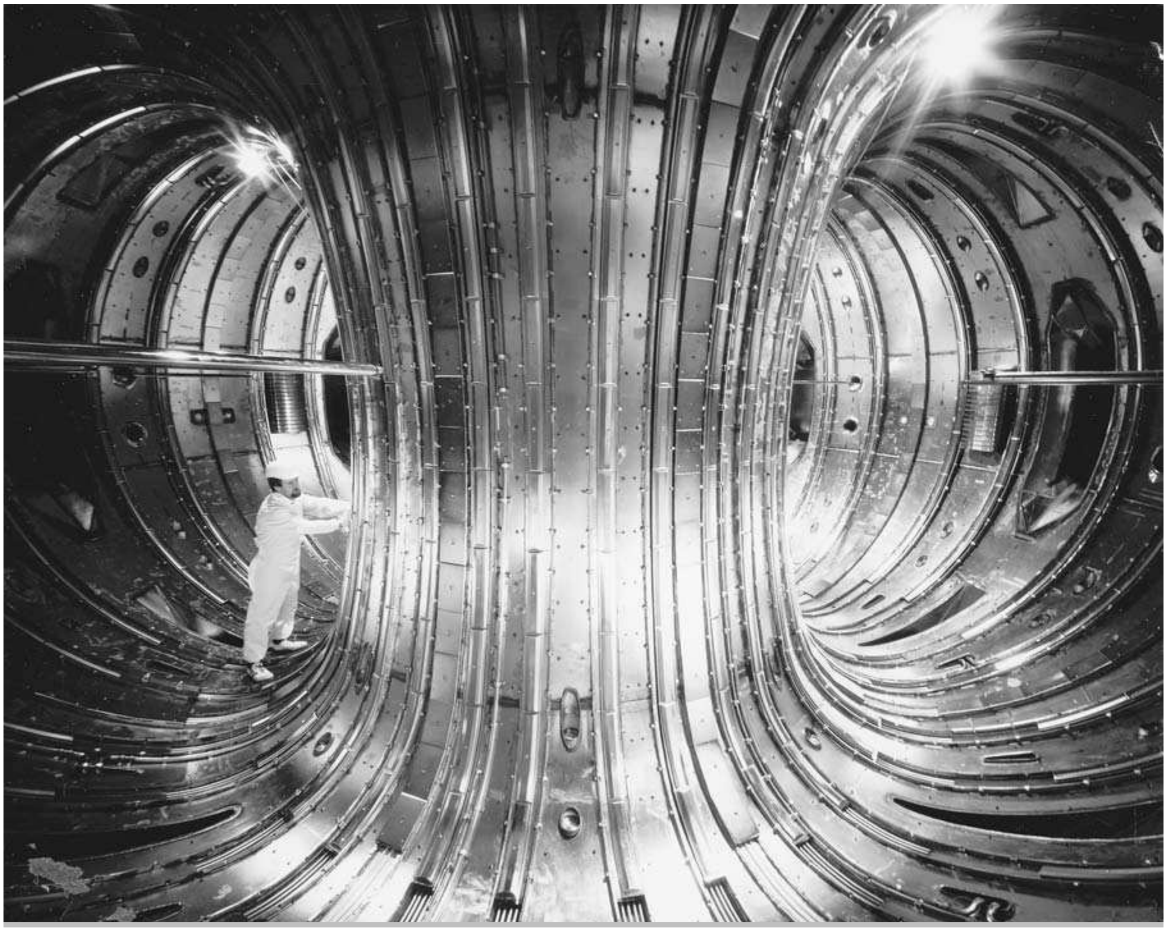

torus Inside the torus of an experimental nuclear fusion reactor. *Joint European Torus*

torus

A doughnut, bagel, or inner-tube shape; the word comes from the Latin for "bulge" and was first used to describe the molding around the base of a column. One way to think of a torus is as a **surface of revolution** obtained by rotating a circle around an axis that lies in the plane of the circle but doesn't intersect the circle. In the general case, where the shape being so rotated is any closed plane curve, the resulting surface is called a *toroid*. Although the usual torus in three-dimensional space is shaped like a doughnut, the concept of the torus is extremely useful in higher dimensional space as well.

tour

A sequence of moves by a **chess** piece on a chessboard in which each square of the board is visited exactly once. See also **knight's tour** and **magic tour**.

Tower of Brahma

A romantic legend manufactured by Edouard **Lucas** as an accompaniment to the popular game he invented, the **Tower of Hanoi**. According to the tale of the Tower of Brahma, in the Indian city of Benares, beneath a dome that marked the center of the world, is to be found a brass plate in which are set three diamond needles, "each a cubit high and as thick as the body of a bee." Brahma placed 64 disks of pure gold on one of these needles at the time of Creation. Each disk is a different size, and each is placed so that it rests on top of another disk of greater size, with the largest resting on the brass plate at the bottom and the smallest at the top. Within the temple are priests whose job it is to transfer all the gold disks from their original needle to one of the others, without ever moving more than one disk at a time. No priest can ever place any disk on top of a smaller one, or anywhere

AND BRAHMA SAID UNTO THE PRIESTS TRANSFER THESE DISKS FROM THE FIRST PYRAMID TO THE THIRD TRANSPOSING ONE DISK AT A TIME ONLY
AND PUTTING IT EITHER ON A VACANT PYRAMID OR ON A LARGER DISK. BY THE TIME YOU HAVE EXECUTED THIS TASK THE END OF THE WORLD WILL BE NEAR.
PYRAMIDS
A PROBLEM FROM THE ANCIENTS
TRADE MARK
A MASTER THINKER CREATION
THIS MODERN VERSION MAY BE SOLVED IN A SHORT TIME..
PYRAMIDS
A GAME
A PUZZLE

Tower of Hanoi A vintage version of the game called Pyramids, manufactured by Knapp Electric, Inc., of New York. *Sue & Brian Young/Mr. Puzzle Australia, www.mrpuzzle.com.au*

else except on one of the needles. When the task is done, and all 64 disks have been successfully transferred to another needle, "tower, temple, and Brahmins alike will crumble into dust, and with a thunder-clap the world will vanish." The prediction (thunder-clap aside) seems fairly safe given that the number of steps required to transfer all the disks is $2^{64} - 1$, which is approximately 1.8447×10^{19}. Assuming one second per move, this would take about five times longer than the current age of the universe! Interestingly, $2^{64} - 1$ is also the answer to the **wheat and chessboard problem**.

Tower of Hanoi

A game invented by Edouard **Lucas** and sold as a toy in 1883. Early versions of it carried the name "Prof. Claus" of the College of "Li-Sou-Stain," but these were quickly discovered to be anagrams for "Prof. Lucas" of the College of "Saint Louis." The game, in its usual form, consists of three pegs on one of which are eight disks, stacked from largest to smallest. The problem is to transfer the tower to either of the vacant pegs in the fewest possible moves, by moving one disk at a time and never placing any disk on top of a smaller one. The minimum number of moves turns out to be $2^n - 1$, where n is the number of disks; this equals 255 in the case of eight disks. The original toy came with a description saying that it was a small version of the great **Tower of Brahma**.[257]

T-puzzle

A surprisingly difficult puzzle, given that there are only four pieces; it dates back to the start of the twentieth century. Photocopy and cut out the four pieces shown in the figure on the following page, and then try to arrange them to make the symmetric capital T. You are allowed to rotate the pieces as you wish and even turn them over, but they mustn't overlap in the final letter. In fact *two* different symmetric capital T letters can be made from the pieces. Also, two other symmetric shapes can be formed from the set, including an isosceles trapezoid. Can you find all of these?

trace

The sum of the terms along the **main diagonal** of a **matrix**.

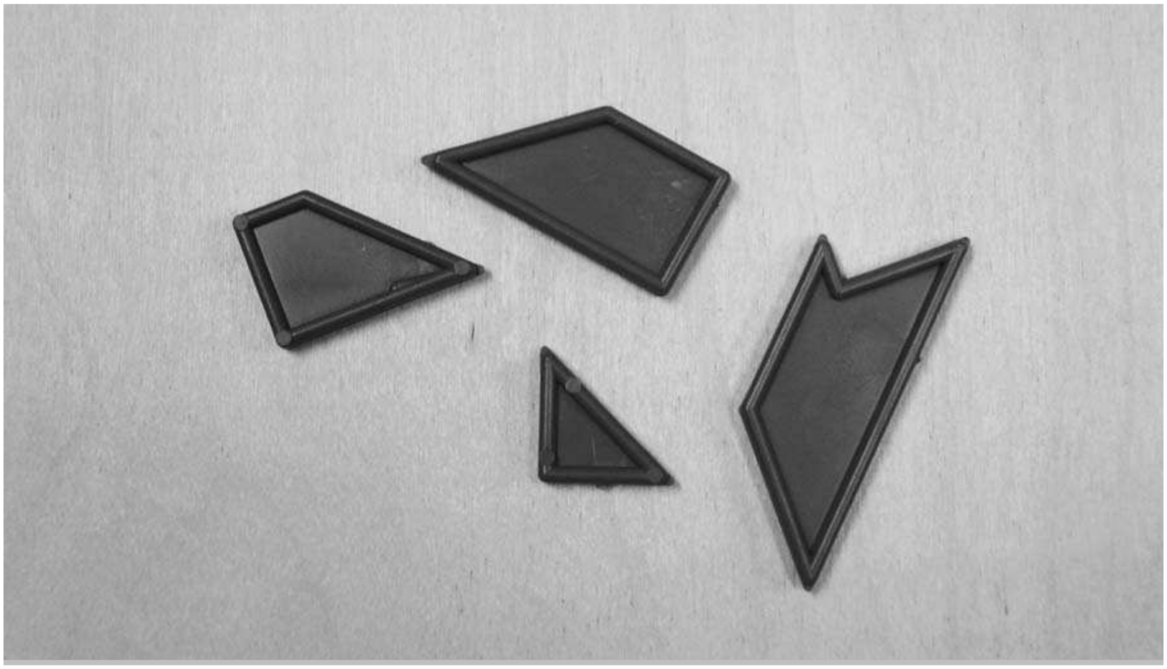

T-puzzle The pieces of the T-puzzle. *Kadon Enterprises, Inc., www.puzzlegames.com*

tractrix

A curve, sometimes called the *trajectory curve* or *equitangential curve,* that is the answer to a question asked by the Frenchman Claude Perrault (1613–1688). Perrault is not a giant in the annals of mathematics; in fact, he trained as a doctor and gained a minor reputation as an architect and an anatomist before dying in unusual style as a result of an infection he caught while dissecting a camel. His greatest claim to fame, aside from his connection with the tractrix, is that he was the brother of the author of "Cinderella" and "Puss-in-Boots."

In 1676, at about the time Gottfried **Leibniz** was doing groundbreaking work on the calculus, Perrault placed his pocket watch on the middle of a table, pulled the end of its chain along the edge of the table, and asked: What is the shape of the curve traced by the watch? The first known solution was given in a letter to a friend in 1693 by Christiaan **Huygens**, who also coined the name "tractrix" from the Latin *tractus* for something that is pulled along. (The corresponding German name is *hundkurve,* or "hound curve," which makes sense if you imagine the path a dog might follow on its leash as its master walks away.) The tractrix can also be found by taking the **involute** of a **catenary**. (Imagine a horizontal bar held at the vertex of the catenary and the point of contact marked as *P.* When the bar is rolled against the catenary without slipping, the path of *P* is a tractrix.) It is described by the parametric equations: $x = 1/\cosh(t)$, $y = t - \tanh(t)$. The **surface of revolution** of the tractrix is the **pseudosphere**, which is the classic model for **hyperbolic geometry** and one possible three-dimensional analog for the shape of the four-dimensional space-time in which we live.

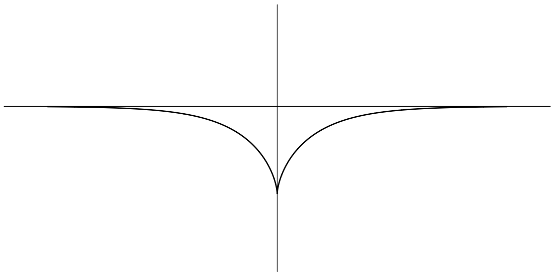

tractrix © *Jan Wassenaar, www.2dcurves.com*

trajectory

(1) The path of a **projectile** or other moving body through space. (2) A curve that intersects all curves of a given family at the same angles; if the intersection is at a right angle, this is an *orthogonal trajectory.* (3) The path through **phase space** taken by a system.

transcendental number

A number that can't be expressed as the root of a **polynomial** equation with integer coefficients. Transcendental numbers are one of the two types of **irrational number**, the other being **algebraic numbers**. Their existence was proved in 1844 by the French mathematician Joseph Liouville (1809–1882). Although transcendentals make up the vast majority of real numbers, it is often surprisingly hard, and may even be impossible, to tell whether a certain number is transcendental or algebraic. For example it is known that both **pi** (π) and ***e*** are transcendental and also that at least one of $\pi + e$ and $\pi \times e$ must be transcendental, but it is not known which. It is also known that e^π is transcendental. This follows from the *Gelfond-Schneider theorem,* which says that if a and b are algebraic, a is not 0 or 1, and b is not rational, then a^b is transcendental. Using Euler's formula, $e^{i\pi} = -1$, and taking both sides to the power $-i$ gives $(-1)^{-i} = (e^{i\pi})^{-i} = e^\pi$. Since the theorem tells us that the left-hand side is transcendental, it follows that the right-hand side is too. (It also follows that $e \times \pi$ and $e + \pi$ are not both algebraic, because if they were, then the equation $x^2 + x(e + \pi) + e\pi = 0$ would have roots e and π, making both numbers algebraic.) But although it is known that e^π is transcendental, the status of e^e, π^e, and π^π remains uncertain.

transfinite number

Any of the infinite **ordinal numbers** first described by Georg **Cantor**.

transformation

In geometry, a change to an object due to a process such as **rotation**, **reflection**, enlargement, or **translation**. In algebra, a transformation is the action of a **function**; in other words, what happens when there is a **one-to-one** mapping between sets of objects.

translation

Any **transformation** that takes the form of a constant offset with no rotation or distortion.

transpose

An operation that flips a **matrix** about the **main diagonal**.

transposition cipher

A **cipher** that encodes a message by reordering the plaintext. The receiver decodes the message using the inverse transposition. A simple kind of transposition cipher writes the message in a rectangle by rows, for example:

Asimplekin
doftranspo
sitionciph
erwritesth
emessagein
toarectang
lebyrowsan
dreadsitou
tbycolumns

and reads it by columns:

Adsee tldts oirmo erbif tweab eymti rsrya cproi serdo lanta cosle ncegt wiuks iseas tmipp tinao nnohh ngnus.

This type of cipher can be made more difficult to crack by permuting the rows and columns. See also **substitution cipher**.

transversal

A line that cuts across parallel lines, intersecting each of them.

trapezoid

A **quadrilateral** with one pair of parallel sides; in Britain this shape is known as a **trapezium**. If the parallel sides are of length a and b, and h is the perpendicular distance between them, then the area of the trapezoid is given by $A = \frac{1}{2}(a + b)h$.

trapezium

The American definition of a trapezium is a **quadrilateral** with no parallel sides. The British definition is equivalent to that of a **trapezoid**.

traveling salesman problem

Given a number n of cities, along with the cost of travel between each pair of them, find the cheapest way of visiting all the cities and returning to the starting point. This is equivalent to finding the **Hamilton circuit** of minimum weight in a weighted **complete graph**. Mathematical problems related to the traveling salesman problem (TSP) were treated in the nineteenth century by William **Hamil-**

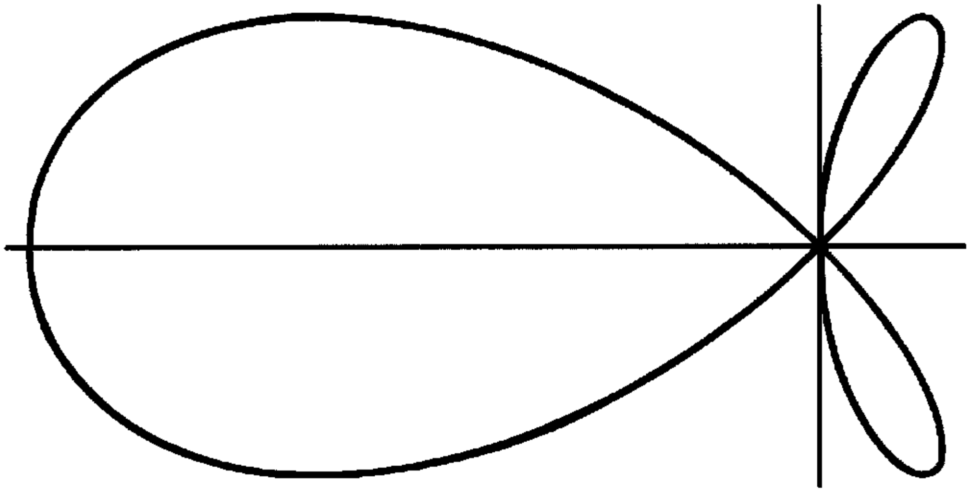

trefoil curve

ton, for example in his **Icosian game**, and by Thomas **Kirkman**. The general form of the TSP appears to be have been first studied by mathematicians in the 1930s, notably by Karl Menger in Vienna and Harvard, and later promoted by Hassler Whitney and Merrill Flood at Princeton. It has become a classic challenge to computer scientists seeking fast **algorithm**s to complex problems. An approximate solution to the TSP for 15,112 cities, towns, and villages in Germany was found in 2001 by Princeton researchers using 110 computer processors and the equivalent of more than five years computer time on a 2 GHz machine.

tree

A **graph** with the property that there is a unique path from any **vertex** to any other vertex traveling along the **edge**s.

trefoil curve

The plane curve given by the equation

$$x^4 + x^2y^2 + y^4 = x(x^2 - y^2).$$

triangle

A three-sided **polygon**. The sum of the interior angles of a triangle is always 180°, unless the triangle is drawn in a **non-Euclidean geometry**. Triangles can be classified either by their angles, as **acute**, **obtuse**, or **right**; or by their sides, as scalene (all different), **isosceles** (two the same), or **equilateral** (all equal).

Triangular Lodge

One of the few triangular buildings in England; it was built by Sir Thomas Tresham in about 1595 at Rushton, Northamptonshire. Tresham was a Catholic (spending some 15 years in prison because of this) and also a mystic numerologist. The whole design of the Lodge is based on the number **three**, which Tresham saw connected with his own surname and as an expression of his faith in the Christian Trinity. The Lodge's ground plan is a perfect **equilateral** triangle, each side 33 feet long–by tradition, the age of Christ at his death. The building has three floors, each floor has three windows, and each window is a three-fold **trefoil**. There are three gables and three gargoyles on each side. Even the central chimney is three-sided. The inscriptions all have 33 letters. Other buildings with a three-sided equilateral theme in Europe include a triangular castle at Gripsholm in Sweden and part of the Chateau de Chantilly in France, which is based on an equilateral plan of gigantic scale.

triangular number

Any number that can be represented by a triangular array of dots: 1, 3, 6, 10, ... The nth triangular number is $n(n + 1)/2$. Every integer is the sum of at most three triangular numbers. Every triangular number is a **perfect number**. If T is a triangular number, $8T + 1$ is a perfect square and $9T + 1$ is another triangular number. The square of the nth triangular number is equal to the sum of the first n cubes. Certain triangular numbers are also squares, but no triangular number can be a third, fourth, or fifth power, nor can one end in 2, 4, 7, or 9.

triangulation

A **tiling** of some object such as a **manifold** by simplices (see **simplex**).

tribar illusion

See **Penrose triangle**.

tricuspoid

See **deltoid**.

trident of Newton

A curve investigated and named "trident" by Isaac **Newton** as part of his systematic study of **cubic** equations. René **Descartes** also studied it and it is sometimes called the parabola of Descartes, although it isn't a parabola. It has the Cartesian equation

$$xy = cx^3 + dx^2 + ex + f.$$

trifolium

See **rose curve**.

trigonometric curve

Any of the curves produced when a **trigonometric function** is graphed.

trigonometric function

Any of the **function**s *sine* (sin), *cosine* (cos), *tangent* (tan), *secant* (sec), *cosecant* (cosec), or *cotangent* (cot), or their inverses, sin^{-1}, etc, which deal with certain proportions in right **triangle**s. For example, the sine of an angle θ, sin θ, in a right triangle, is equal to the side opposite the angle divided by the hypotenuse (the longest side). Similarly cos is adjacent over hypotenuse and tan is opposite over adjacent. Sec, cosec, and cot are the multiplicative inverses of cos, sin, and tan, respectively: $\sec\theta = 1/\cos\theta$, and so forth. These are not the same as the inverse functions $\cos^{-1}$, $\sin^{-1}$, and $\tan^{-1}$, which are also known as arccos, arcsin, and arctan. Graphs of trigonometric functions produce **trigonometric curves**.

trigonometry

The branch of **mathematics** that deals with the relationships between the sides and the angles of **triangle**s and the calculations based on them, particularly the **trigonometric function**s. Sherlock Holmes relies on a little trigonometry to solve a 250-year-old mystery known as the Musgrave Ritual (in a short story of the same name)–an enigmatic series of clues that refers to the shadow of an elm tree when the sun is just visible at the top of a nearby oak to point toward buried treasure. The great detective recalls to Watson his conversation with Reginald Musgrave:

> "Have you any old elms?" . . .
>
> "There used to be a very old one over yonder, but it was struck by lightening ten years ago, and we cut down the stump."
>
> "You can see where it used to be?"
>
> "Oh, yes." . . .
>
> "I suppose it is impossible to find out how high the elm was?"
>
> "I can give it to you at once. It was sixty-four feet. . . . When my old tutor used to give me an exercise in trigonometry, it always took the shape of measuring heights." . . .
>
> I went with Musgrave to his study and whittled myself this peg, to which I tied this long string with a knot at each yard. Then I took two lengths of a fishing-rod, which came to just six feet. . . . The sun was just grazing the top of the oak. I fastened the rod on end, marked out the direction of the shadow. . . . It was nine feet in length. Of course, the calculation was now a simple one. If a rod of six feet threw a shadow of nine, a tree of sixty-four feet would throw one of ninety-six. . . . I measured out the distance . . . and I thrust a peg into the spot.

trillion

In American and general usage, a million million–1,000,000,000,000 or 10^{12}. A European trillion is a million times larger than this, or 10^{18}. Counting one number every second, 24 hours a day, it would take 31,688 years to reach one (American) trillion. The first trillion-dollar lawsuit ($116 trillion) was filed in August 2002 by 600 family members against a company run by Osama bin Laden's family, Saudi Arabian princes, and Sudan. See also **large number**.

trimorphic number

See **automorphic number**.

trinomial

An algebraic expression consisting of three terms.

triomino

Also called a *tromino,* a three-square **polyomino**.

triple

A multiple of **three**. A *triple integral* is one in which the integrand is integrated three times. See also **Pythagorean triplet**.

trisecting an angle

Whereas **bisecting an angle** could hardly be simpler, splitting an angle in three equal parts with compass and straightedge alone is impossible, except in a few special cases such as when the angle happens to be 90°. Trisecting an arbitrary angle can be done if you cheat by using a measuring ruler instead of a plain straightedge, or even if you draw just two marks on the straightedge, but not if you play by the rules and the straightedge is completely blank. The Greeks put a huge effort into the problem but couldn't crack it. In fact, the question of whether trisection could ever be done in the general case remained open until 1837, when it was finally shown to be impossible by Pierre Wantzel, a 23-year-old French mathematician. Why is it impossible? Wantzel showed that the two problems of trisecting an angle and of solving a cubic equation are equivalent. Moreover, he showed that only a very few cubic equations can be solved using the straightedge-and-compass method. He thus deduced that most angles cannot be trisected.

trisector theorem

See **Morley's miracle**.

trisectrix

A general name for curves that can be used in **trisecting an angle**. The name "trisectrix," on its own, is often applied specifically to the **limaçon of Pascal**. Other famous trisectrix curves include the **Maclaurin trisectrix** and the **conchoid** of Nicomedes.

triskaidekaphobia

See **thirteen**.

trochoid

The curve formed by the path of a point on the extension of a radius of a circle as it rolls along a curve or line. It is also the curve formed by the path of a point on a perpendicular to a straight line as the straight line rolls along the convex side of a base curve. By the first definition, the trochoid is derived from the **cycloid**; by the second definition it is derived from the **involute**. See also **roulette**.

truel

A three-cornered gunfight or its logical equivalent. Imagine a truel between Arnie, Bullseye, and Clint, who are standing at the corners of an equilateral triangle. All know that Arnie's chance of hitting a target is 0.3 and Clint's is 0.5, while Bullseye never misses. They have to fire at their choice of target in the order Arnie then Bullseye then Clint until only one man is left. A man who's been hit is out of the fight and can no longer be shot at. What should Arnie's strategy be?

Truels, like this one, have become a significant topic in **game theory** because they're analogs of various real-life situations, from rivalry among animals to competition between television networks. Small changes in the rules can lead to strikingly different, sometimes counterintuitive outcomes. Different firing rules are possible: sequential in fixed order (players fire one at a time in a predetermined, repeating sequence), sequential in random order (the first player to fire and each subsequent player is chosen at random from among the survivors), or simultaneous (all surviving players fire at the same time in every round). In certain truels, a participant is allowed to shoot at the ground rather than try to eliminate an opponent (an optimal strategy if the firing order is fixed and each player has only one bullet and is a perfect shot). If the first shooter misses on purpose, he eliminates himself as a threat, and the other two fight it out, leaving two survivors in the end. Any other course of action would lead to the first shooter's own demise, with only one survivor. Even if the players have an unlimited supply of bullets, the truel may still end with more than one survivor because no player wants to be the first to shoot. Indeed, under the fixed firing order rule, no player has an incentive to eliminate another player. Only in the case of simultaneous firing is there a chance that nobody will survive.

Most of the mathematical research on truels concerns the relationship between a player's marksmanship (probability of hitting a target) and his or her survival probability. It's possible to show, for example, that better marksmanship can hurt in many situations. In a sequential truel in which contendors aren't allowed to shoot in the air, a player maximizes his probability of survival by firing at the opponent against whom he'd less prefer to fight in a duel–regardless of what the other players do. If his shot misses, it makes no difference who the target was. If the shot hits the target, the shooter is better off because his opponent in the next duel is weaker. Thus, the first shooter fires at the opponent who's the better marksman. In general, depending on the marksmanship values, the survival probabilities of the truelists could end up in any order, including one that is the reverse order of shooting skill. Optimal play can be very sensitive to slight changes in the rules, such as the number of rounds of play allowed. On the other hand, some factors are fairly constant: the disadvantage of being the best marksman, the weakness of pacts, the possibility that an endless supply of ammunition may stabilize rather than undermine cooperation, and the deterrent effect of an indefinite number of rounds of play (which can prevent players from trying to get the last shot). Some of these findings are counterintuitive, even paradoxical. An understanding of them might well dampen the desire of aggressive players to score quick but temporary wins, rendering them more cautious. In particular, contemplating the consequences of a long, drawn-out conflict, truelists may come to realize that their own actions, while immediately beneficial, may trigger forces that ultimately lead to their own destruction.

truncate

To slice off a corner of a **polyhedron** around a **vertex**.

truncatable prime

A **prime number** *n* that remains a prime when digits are deleted from it one at a time. For example 410,256,793 is a truncatable prime because each number created by the removal of the digit underlined produces a new prime: 410,256,793; 41,256,793; 4,125,673; 415,673; 45,673; 4,567; 467; 67; 7. It is conjectured that there are infinitely many of these primes. If the digits from a prime can be deleted only from the right, to leave a prime, then *n* is called a *right truncatable prime.* If they can be deleted only from the left, to leave a prime, then *n* is a *left truncatable prime.* The list of primes from which *any* digit can be deleted at each step to leave a prime is very short indeed, because it demands that each digit be a prime and also that no digit occurs twice. Only these numbers satisfy this requirement: 2, 3, 5, 7, 23, 37, 53, and 73.

Tschirnhaus's cubic

A curve with the Cartesian equation $3ay^2 = x(x - a)^2$.

Turing, Alan Mathison (1912–1954)

An English mathematician considered to be one of the fathers of modern digital computing. At an early age, Turing showed signs of the genius and eccentricity that became hallmarks of his adult personality. He taught himself to read in three weeks and made a habit of stopping at street corners to read the serial numbers of traffic lights. Later, he became a near-Olympic-class runner and ran long distances with an alarm clock tied to his waist to time himself.

At Cambridge, Turing studied under G. H. **Hardy** and got involved with problems that David **Hilbert** and Kurt **Gödel** had proposed to do with completeness and decidability in mathematics. In 1936, he introduced the idea of what became known as **Turing machine**s–formal devices capable of solving any conceivable mathematical problem that could be represented by an **algorithm**. However, the Turing machine was only a theoretical possibility at that time and not a working implementation. It would remain for later researchers to solve the various practical difficulties required to make the computer a reality. Turing also showed that there were mathematical problems that a Turing machine could never solve. One of these is the **halting problem**. While his proof was published after that of Alonzo **Church**, Turing's work is more accessible and intuitive. During World War II, Turing was a major player at Bletchley Park, near present-day Milton-Keynes (a town built after the war), in the successful efforts to crack the Nazi Enigma ciphers. While serving at Bletchley Park (1939–1944), he stayed at the Crown Inn, Shenley Brook End, and somewhere near here he buried two silver bars, carefully recording the site with respect to local landmarks. When he returned to recover them, the area had been rebuilt and all his landmarks were gone. Despite several attempts with metal detectors, he never recovered them and no one else is known to have found them. The Crown is now a private house and the area where he buried the bars is a housing estate.

Turing, Alan Mathison

Turing's interest in computing continued after the war, when he worked at the National Physical Laboratory on the development of a stored-program computer (the ACE or Automatic Computing Engine). In 1948 he moved to the University of Manchester, where the first stored-program digital computer ran later that year (see

Babbage, Charles for a photo of this machine). In 1950, in the article "Computing Machinery and Intelligence," Turing tackled the problem of **artificial intelligence**, and proposed an experiment now known as the **Turing test**.

In 1952 his lover helped a compatriot to break into Turing's house and commit larceny. Turing went to the police to report the crime. As a result of the police investigation, he was charged with homosexuality (formerly a crime), offered no defense, and was convicted. Following the well-publicized trial, he was given a choice between incarceration and libido-reducing hormone injections. He chose the latter, which lasted for a year and had side effects including the development of breasts during that period. In 1954 he died of poisoning after eating a cyanide-laced apple. Most (though not his mother) believed that his death was intentional, and the death was ruled a suicide. According to one urban legend the Apple company's logo is symbolic of this event: an apple with two bites (or possibly bytes) out of it and rainbow colors that code for homosexuality. See also **Church-Turing thesis**.[168]

Turing machine

An abstract model of computer execution and storage introduced in 1936 by Alan **Turing** to give a mathematically precise definition of **algorithm**. A Turing machine can be thought of as a black box that carries out a calculation of some kind on an input number. If the calculation reaches a conclusion, or halts, then an output number is returned. Otherwise, the machine theoretically carries on forever (see **halting problem**). There are an infinite number of Turing machines, as there are an infinite number of calculations that can be done with a finite list of rules. A Turing machine that can simulate any other Turing machine is called a *universal Turing machine* or a **universal computer**. The concept of Turing machines is still widely used in theoretical computer science, especially in **complexity theory** and the theory of computation.

Turing test

A proposed way of deciding if a machine has human-level intelligence. First described by Alan **Turing** in 1950, it goes like this: A human judge engages in a natural language conversation with other parties; if the judge can't reliably tell whether the other party is human or machine, then the machine is said to pass the test. It is assumed that both the human and the machine try to appear human. The origin of the test is a party game in which guests try to guess the gender of a person in another room by writing a series of questions on notes and reading the answers sent back. In Turing's original proposal, the human participants had to pretend to be the other gender, and the test was limited to a 5-minute conversation. These features are nowadays not considered to be essential and are generally not included in the specification of the Turing test. Turing proposed the test in order to replace the emotionally charged and, for him, meaningless question "Can machines think?" with a more well-defined one. Turing predicted that machines would eventually be able to pass the test. In fact, he estimated that by the year 2000, machines with 10^9 bits (about 119MB) of memory would be able to fool 30% of human judges during a 5-minute test. He also predicted that people would then no longer consider the phrase "thinking machine" contradictory.

It has been argued that the Turing test can't serve as a valid definition of **artificial intelligence** for at least two reasons: (1) A machine passing the Turing test might be able to simulate human conversational behavior, but this could be much weaker than true intelligence. The machine might just follow some cleverly devised rules. (2) A machine might well be intelligent without being able to chat like a human. Simple conversational programs, such as ELIZA, have fooled people into believing they are talking to another human being; however, such limited successes don't amount to passing the Turing test. Most obviously, the human party in the conversation has no reason to suspect he is talking to anything other than a human, whereas in a real Turing test the questioner is actively trying to determine the nature of the entity he is chatting with. The Loebner Prize is an annual competition to determine the best Turing test competitors. See also **Chinese room**.

twelve

A number heavily used for grouping things (inches, hours, 12-packs), partly because it can be divided evenly in several different ways (by 2, 3, 4, and 6) and partly because there are roughly 12 cycles of the Moon for every one of the Sun. The Latin *duodecim* (two + ten) for 12 forms the root of dodecagon (originally duodecagon), meaning a 12-sided shape, and duodenum, the first part of the intestine that is about 12 inches long. Contracted and modified over the years, *duodecim* became "dozen." Multiples of 12 have also been used by many cultures for various units and measures. A "shock" was 60 or five dozen (a dozen for each finger on one hand), and many cultures had a "great hundred" of 120 or 10 dozen (a dozen for each finger on both hands). The Romans used a fraction system based on 12 and the smallest part, an *uncil,* became our word for "ounce." The French emperor Charlemagne established a monetary system that had a base of 12 and 20, the remnants of which persist. Until 1970, the English pound sterling consisted of 20 shillings, and each shilling contained 12 pence. In 1944, The Duodecimal Society was formed in New York with the purpose of proposing a switch to base 12 for all scientific work. There are 12 signs of the zodiac and there were 12 apostles of Christ. Twelve is the smallest **abundant number**, a **Harshad number**, and a semiperfect number (because 12 = 1 + 2 + 3 + 6; see **perfect number**). See also **gross**.

twelve-color map problem

If on a plane or sphere each country has at the most one colony, requiring the same color as its parent country, at most 12 different colors are needed to distinguish the political regions on a map. The problem of determining if this is true or not remains open. See also **four-color map problem**.

twenty

Many early cultures, including some in Europe; the Mayans of Central America; and the Ainu, the indigenous people of the Japanese islands, used a **base** of 20 for counting. The base 20 system was retained until about 1970 in the British monetary system, in which there were 20 shillings to the pound. A group of 20 is often called a **score**. Two of the five Platonic solids involve 20: the **icosahedron** has 20 triangular faces and the **dodecahedron** has 20 vertices. Twenty is a **Harshad number**, a semiperfect number (see **perfect number**), and a **practical number**. It is also a tetratrahedral number–the sum of consecutive triangular numbers (1 + 3 + 6 + 10).

twin primes

Pairs of **prime numbers** that differ by two, the first of which are 3 and 5, 5 and 7, 11 and 13, and 17 and 19. The largest example known, as of February 2003, is a pair of 51,090-digit primes discovered by Yves Gallot and Daniel Papp, with the value $33{,}218{,}925 \times 2^{169690} \pm 1$. Other than the first, all twin primes have the form $\{6n - 1, 6n + 1\}$; also, the integers n and $n + 2$ form twin primes if and only if $4[(n - 1)! + 1] = -n \pmod{n(n + 2)}$. In 1919 the Norwegian mathematician Viggo Brun (1885–1978) showed that the sum of the reciprocals of the twin primes converges to a sum now known as *Brun's constant:*

$$(1/3 + 1/5) + (1/5 + 1/7) + (1/11 + 1/13) + (1/17 + 1/19) + \ldots$$

In 1994, by calculating the twin primes up to 10^{14} (and discovering the infamous Pentium bug in the process), Thomas Nicely of Lynchburg College estimated Brun's constant to be 1.902160578. According to the (unsolved) *twin-prime conjecture* there are infinitely many twin primes. The twin-prime conjecture generalizes to prime pairs that differ by any even number n, and generalizes even further to certain finite patterns of numbers separated by specified even differences. For example, the following triplets of primes all fit the pattern k, $k + 2$, and $k + 6$: 5, 7, and 11; 11, 13, and 17; 17, 19, and 23; 41, 43, and 47. It is believed that for any such pattern not outlawed by divisibility considerations there are infinitely many examples. (The pattern k, $k + 2$, and $k + 4$ has only one solution in primes, 3, 5, and 7, because any larger such triplet would contain a number divisible by 3.) Quartets of the form k, $k + 2$, $k + 6$, and $k + 8$ (the smallest example is 5, 7, 11, and 13) are thought to be infinite. For some patterns no example is known, or only one.

twins paradox

See **relativity theory**.

twisted cubic

A curve in three-dimensional space or projective space whose points are given by $(x(t), y(t), z(t))$ for a parameter t and where x, y, z are **polynomials** of at most degree 3.

two

The first even number and the only even **prime number**. The word comes from the Greek *dyo* and the Latin *duo* through the Old English *twa.* Early languages often had both feminine and masculine forms for two and so there are a lot of diverse roots related to "two-ness." Many "two" words use the Greek root *bi* such as biannual, binary, biscuit, and biceps. Others come from the Old English *twa,* such as between, twilight, twist, and twin. From *duo* we get dual, duet, dubious (of two minds), duplex (two layers), and double. The Latin *di* gives us diploma (two papers) and dihedral. The earlier Greek *dyo* produces dyad, composed of two parts. Two is the only positive real number that gives the same result when added to itself as when multiplied by itself. It is conjectured that 2 is the only even integer that cannot be written as the sum of two primes (see **Goldbach conjecture**) and it has recently been proven that 2 is the largest value of n for which the equation $x^n + y^n = z^n$ has nonzero integer solutions (see **Fermat's last theorem**). Two is the base of the **binary** number system.

two-dimensional world

Life in three dimensions is familiar and there is a huge body of literature on the **fourth dimension**. But what would a universe of just two dimensions be like? The first and the most charming book on the subject is Edwin Abbott's ***Flatland: A Romance of Many Dimensions***

(1884).[1] This was followed by Charles **Hinton**'s lengthier *An Episode of Flatland* (1907)[272] in which the 2-d world is not a plane, as in Abbott's yarn, but the rim of a large circular world called Astria. Hinton was the first to explore in some depth what science and technology might be like in two dimensions; in fact, an earlier pamphlet of his called "A Plane World," (reprinted in *Scientific Romances* in 1884) may have helped inspire Abbott's novel. Hinton's speculations were taken much further by Alexander **Dewdney** in *Planiverse* (1984).[81]

U

Ulam, Stanislaw Marcin (1909–1984)

A Polish-born American mathematician and physicist who solved the problem of how to initiate fusion in the hydrogen bomb and also devised the **Monte Carlo method** of solving mathematical problems using statistical sampling. He first came to the United States in 1935 following an invitation from John **von Neumann**. One morning in 1946 an event happened that changed Ulam's life, as colleague Gian-Carlo Rota recalled:

> Ulam, a newly appointed professor at the University of Southern California, awoke to find himself unable to speak. A few hours later, he underwent a dangerous surgical operation after the diagnosis of encephalitis. . . . In time, however, some changes in his personality became obvious to those who knew him . . . [H]is ideas, which he spouted out at odd intervals, were fascinating beyond anything I have witnessed before or since. However, he seemed to studiously avoid going into details . . . . [H]e came to lean on his unimpaired imagination for his ideas, and on . . . others for technical support . . . . A crippling technical weakness coupled with an extraordinarily creative imagination is the drama of Stan Ulam.

Ulam spiral

A remarkable geometric pattern accidentally found among the **prime number**s by Stanislaw **Ulam**;[314] it is also known as the *Prime Spiral.* During a boring meeting one day in 1963, Ulam drew a square, marked the number 1 at the center, and then wrote the increasing whole numbers as a spiral that wound its way out to the edge of the paper. He then circled all the prime numbers and was immediately struck by how they tended to fall on diagonal lines radiating from the central 1. In Ulam's words the arrangement of primes "appears to exhibit a strongly nonrandom appearance." Ulam rushed home and

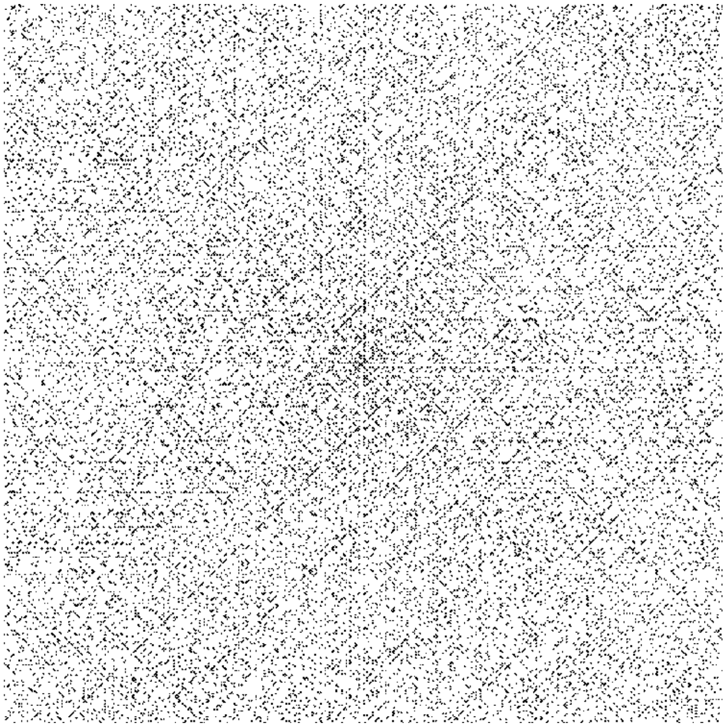

Ulam spiral Patterns amid the primes.

expanded the spiral to cover a much larger portion of the number sequence. The strange pattern persisted. Primes had a tendency to occur in clusters and all clusters tended to make a beautiful image that couldn't be predicted. With the help of computers this pattern can now be explored almost indefinitely and it reveals a wonderfully rich combination of symmetry and surprise–very reminiscent of some **fractal**s.

The Ulam spiral should perhaps be known as the "Clarke spiral" in view of the fact that Arthur C. Clarke described the phenomenon in his novel *The City and the Stars* (1956, ch. 6, p. 54),[63] predating Ulam's discovery by several years. Clarke wrote:

> Jeserac sat motionless within a whirlpool of numbers. The first thousand primes. . . . Jeserac was no mathematician, though sometimes he liked to believe he was. All he could do was to search among the infinite array of primes for special relationships and rules which more talented men might incorporate in general laws. He could find how numbers behaved, but he could not explain why. It was his pleasure to hack his way through the arithmetical jungle, and sometimes he discovered wonders that more skillful explorers had missed. He set up the matrix of all possible integers, and started his computer stringing the primes across its surface as beads might be arranged at the intersections of a mesh.

Ulam's conjecture

See **Collatz problem**.

uncountable set

A **set** of numbers that can't be put in a definite order from smallest to largest and so can't be counted. All uncountable sets are infinite, but not all infinite sets are uncountable. The best known uncountable set is the set of all **real number**s. By contrast the set of all **natural number**s, which represents the "smallest" type of **infinity**, is countable.

undulating number

An integer whose digits, in a given **base**, alternate–that is, one written in the form *ababab . . .*, where *a* and *b* are digits. For example, 434,343 and 101,010,101 are undulating numbers.

unduloid

A member of a family curves that is formed by films or liquid drops suspended between certain boundaries. Examples of unduloids are seen on a spider web when viewed through a microscope. They consist of blobs of viscous liquid that make up the sticky part of the web and are mostly gathered into a lemon shape. The family of unduloids includes shapes ranging from very thin to almost spherical, depending on the diameter of the thread and the volume of liquid in the blob. The shape of the curve is a result of the equality of pressure throughout the blob, which means that the total **curvature** at all points on the surface must be the same. The total curvature is the sum as the curvatures in two planes at right angles, and so varies from one blob to the next. A common property of all unduloids, however, is that they have a constant nonzero main curvature.

unexpected hanging

A remarkable logical **paradox** that appears to have begun circulating by word of mouth in the 1940s, often in the form of a puzzle about a man condemned to be hanged. A judge, with a reputation for reliability, tells a prisoner on Saturday that he will be hung on one of the next seven days but that he will not know which day until he is informed on the morning of the execution. Back in his cell, the prisoner reasons that the judge must be wrong. The hanging cannot be left until Saturday, because the prisoner would certainly know, if this day dawned, that it was his last. But if Saturday is eliminated, the hanging cannot take place on Friday either, because if the prisoner survived Thursday he would know that the hanging was scheduled for the next day. By the same argument, Thursday can be crossed off, then Wednesday, and so forth, all the way back to Sunday. But with every other day ruled out for a possible unexpected hanging, the hangman cannot arrive on Sunday without the prisoner knowing in advance. Thus, the condemned man reasons, the sentence can't be carried out as the judged decreed. But then Wednesday morning comes around and, with it, the hangman–unexpectedly! The judge was right after all and something was awry with the prisoner's seemingly impeccable logic. But what? More than half a century of attack by numerous logicians and mathematicians has failed to produce a resolution that is universally accepted. The paradox seems to stem from the fact that whereas the judge knows beyond doubt that his words are true (the hanging will occur on a day unknown in advance to the prisoner), the prisoner does not have this same degree of certainty. Even if the prisoner is alive on Saturday morning, can he be certain that the hangman will arrive?[61, 260]

uniform polyhedron

A **polyhedron** in which each face is regular and each **vertex** is equivalently arranged. Uniform polyhedra include the **Platonic solid**s, the **Archimedean solid**s, the **prism**s and **antiprism**s, and the nonconvex uniform polyhedra (see **nonconvex uniform polyhedron**).

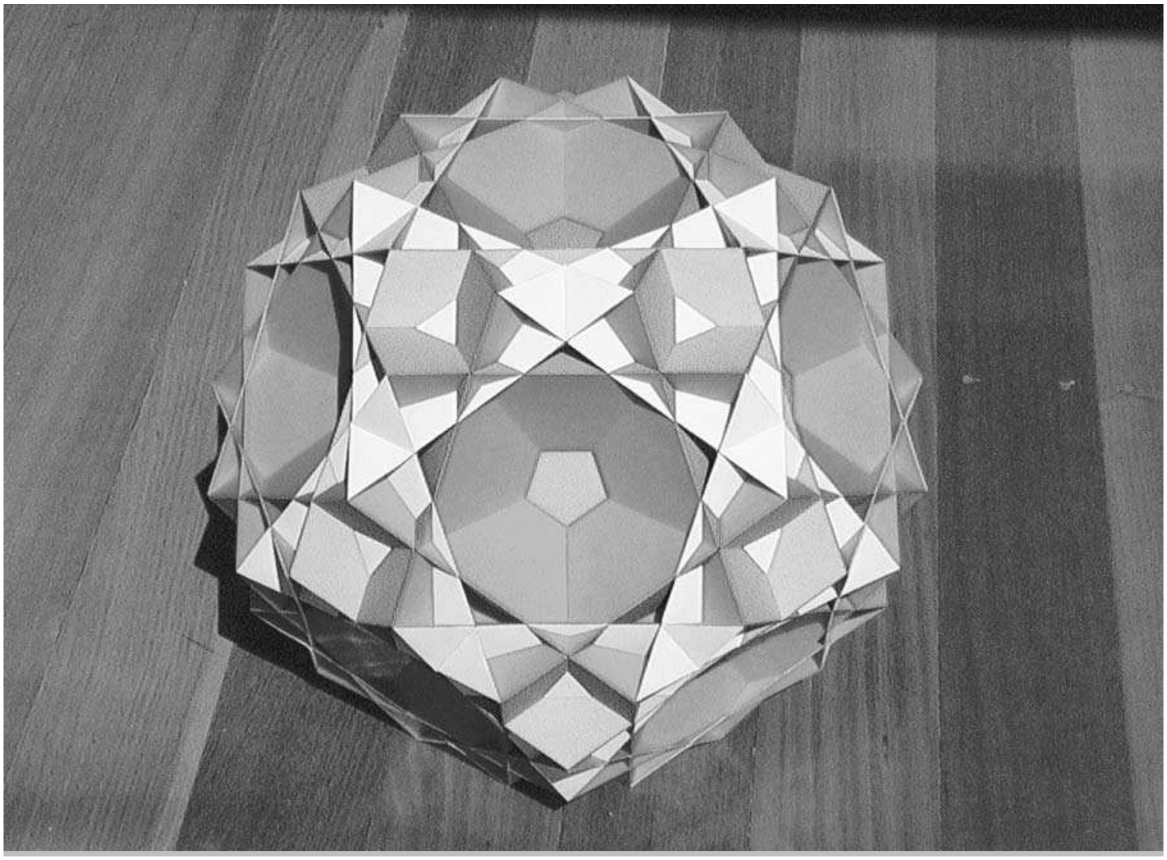

uniform polyhedron The great icosicosidodecahedron, a uniform polyhedron. *Robert Webb, www.software3d.com; created using Webb's Stella program*

unilluminable room problem

Imagine an L-shaped room in which Amy is standing near one corner holding a match. If Bob stands round the corner, he can see light from the match because a light ray can bounce off the two opposite walls. This is true wherever Bob stands in the room: the whole room is illuminated by the one match. Would this be true for a room of *any* shape, or is there at least one room that is *so* complicated that there's somewhere inside it that light from the match never reaches. This problem was first asked by Ernst Strauss in the 1950s. Nobody knew the answer until 1995, when George Tokarsky of the University of Alberta showed that the answer is "yes," there *is* a room that is not completely illuminable. His published floor plan showed a room with 26 sides–the smallest such room currently known. But a mystery remains. The room Tokarsky found contains one particular place where the match can be held which leaves part of the room dark. But if the match is moved slightly, the whole room is lit up again. Is there a room so fiendishly complicated that *wherever* the match is held there are some places that its light can never reach? For the moment we remain in the dark.[320, 334]

unique number

The constant U_n that results if a number A_n consisting of n consecutive digits, in ascending order, is subtracted from the number A_n' obtained by reversing the digits of A_n. For example, a three-digit number 345, if subtracted from its reverse 543, yields a difference of 198. Any other three-digit number subtracted from its reverse gives the same difference. Thus $U_3 = 198$. Similarly for a number with four consecutive digits, the unique number $U_4 = 3{,}087$. The first ten unique numbers are: $U_1 = 0$, $U_2 = 9$, $U_3 = 198$, $U_4 = 3{,}087$, $U_5 = 41{,}976$, $U_6 = 530{,}865$, $U_7 = 6{,}419{,}754$, $U_8 = 75{,}308{,}643$, $U_9 = 864{,}197{,}532$, and $U_{10} = 9{,}753{,}086{,}421$. Unique numbers are related to **Kaprekar numbers**, K_n, by the formula

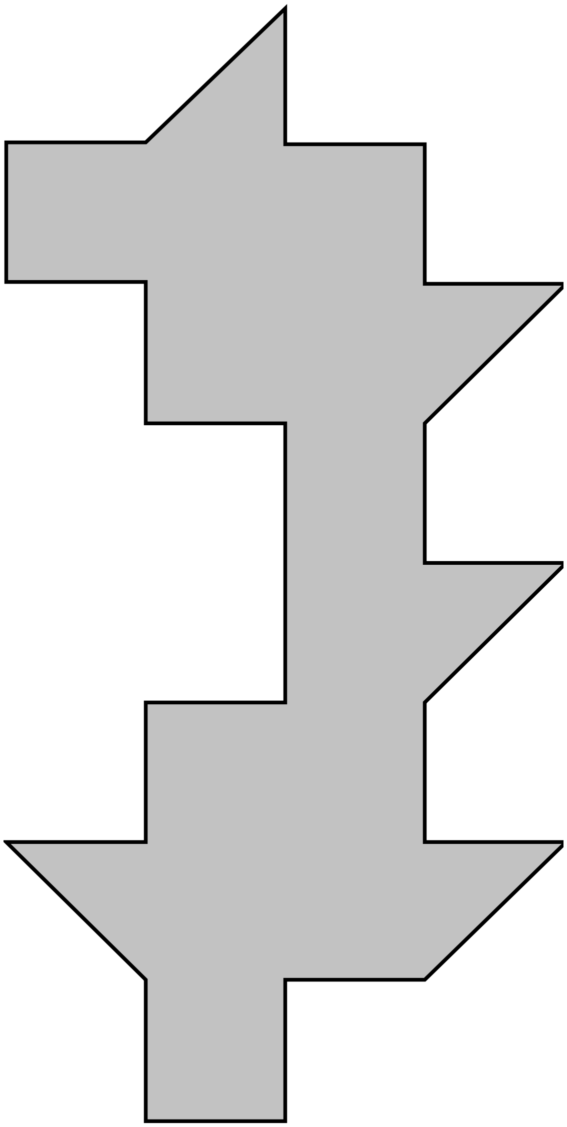

unilluminable room problem No single light source can light up every corner of a room with this shape.

$$U_n + U'_n = K_n + K'_n.$$

For example, when $n = 4$, $K_4 = 6{,}174$, $K_4{}' = 4{,}716$, $U_4 = 3{,}087$, $U_4{}' = 7{,}803$, and

$$3{,}087 + 7{,}803 = 10{,}890 = 6{,}174 + 4{,}716.$$

unimodal sequence

A **sequence** that first increases and then decreases.

unimodular matrix

A square **matrix** whose **determinant** is 1.

union

The **set** of all elements that belong to at least one of two or more given sets. It is denoted by the symbol ∪.

unit circle

A circle with radius 1.

unit cube

A cube with edge length 1. A *unit square* has a side length of 1.

unit fraction

A fraction whose **numerator** (number on top) is 1.

universal approximation

Having the ability to approximate any **function** to an arbitrary degree of accuracy. **Neural network**s are universal approximators.

universal computation

Capable of computing anything that can in principle be computed; being equivalent in computing power to a **Turing machine** or the **lambda calculus**.

universal computer

A computer that is capable of **universal computation**, which means that given a description of any other computer or program and some data, it can perfectly emulate this second computer or program. Strictly speaking, home PCs are not universal computers because they have only a finite amount of memory. However, in practice, this is usually ignored.

Universal Library

A library that contains not just one copy of every book that has ever been printed but one copy of every book that it is possible to print. A version of such a fantastic place is described by Jorge Luis **Borges** in his melancholic short story "Library of Babel" from *The Garden of Forking Paths* (1941). It begins: "The universe (which others call the Library) is composed of an indefinite, perhaps infinite number of hexagonal galleries." Each gallery is identical to all the others and contains 800 books identical in format. "[E]ach book contains four hundred ten pages; each page, forty lines, each line, approximately eighty black letters. . . ." There are 25 symbols–22 letters, the comma, the period, and the space. Because the library contains every possible combination of these

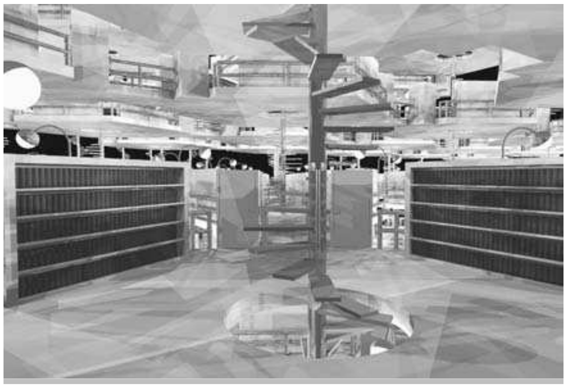

Universal Library Inside the library that goes on forever.
Joseph Formoso

symbols it contains, in addition to vast tracts of gibberish, every truth, falsehood, idea, novel, thought, and description of events, past and future, that are possible. It contains, writes Borges,

> *All*–the detailed history of the future, the autobiographies of the archangels, the faithful catalog of the Library, thousands and thousands of false catalogs, the proof of the falsity of those catalogs, a proof of the falsity of the *true* catalog, the gnostic gospel of Basilides, the commentary upon that gospel, the commentary on the commentary of that gospel, the true story of your death, the translation of every book into every language, the interpolations of every book into all books, the treatise Bede could have written (but did not) on the mythology of the Saxon people, the lost books of Tacitus.

The name "Universal Library" was first used by the German philosopher and science fiction writer Kurt Lasswitz (1848–1910) as the title of a short story published in 1901. He, in turn, borrowed the concept from the German psychologist **Theodor Fechner**. But to get to the root of speculation about listing all possible combinations of words and meanings, we have to go back much further, to Ramon Lully (1235–1315), a Spaniard who was a missionary and a mystic philosopher and who had an idea that later became known as Lully's Great Art. His idea was simply this: if one property of a thing is chosen, say the color of blood, and all the possibilities for that property are listed–blood is green, blood is yellow, etc.–then one of them must be true. One list alone might not be enough to point out the truth. However, other lists could be made that would eliminate some of the possible colors. Done in the right way, the one and only true answer should emerge. Lully even tried to build a device that used a series of concentric rings to bring different combinations of words into alignment. Eventually, the idea came down to Fechner who ruminated on the idea of permuting all combinations of letters to express all possible statements and concepts. There are, however, two barriers to this dream of ultimate truth. The first is that there isn't enough matter or space in the universe to represent all the different ways that book-length sequences of letters can expressed. Second, even if there were, it would take an all-seeing, all-knowing intelligence to sort the rare grain of meaningful wheat from the vast quantities of vapid chaff. See also **monkeys and typewriters**.

universal set
The **set** that contains all elements capable of being accepted to the problem. Also known as the *universe,* as in the **universe of discourse**, it is usually denoted *U.*

universe of discourse
The part of the world under discussion; more precisely, the **set** of all objects presumed or hypothesized to exist for some specific purpose. Objects may be concrete (e.g., a specific carbon atom, Confucius, the Sun) or abstract (e.g., the number 2, the set of all integers, the concept of justice). Objects may be primitive or composite (e.g., a circuit that consists of many subcircuits). Objects may even be fictional (e.g., a unicorn, Sherlock Holmes). The universe of discourse is a familiar concept in **logic**, linguistics, and mathematics.

unknown
A quantity, denoted by a letter, that is to be found by solving one or more equations.

untouchable number
A number that is not the sum of the **aliquot part**s of any other number. The first few untouchable numbers are 2, 5, 52, 88.

up-arrow notation
See **Knuth's up-arrow notation**.

upside-down picture
A picture (or figure) that, when inverted, looks the same or changes into the picture of a different subject. Possibly the most remarkable examples of upside-down art were the cartoons drawn by Gustave Verbeek for the *Sunday New York Herald* in the early 1900s. The first part of the cartoon is read normally; then the newspaper is turned

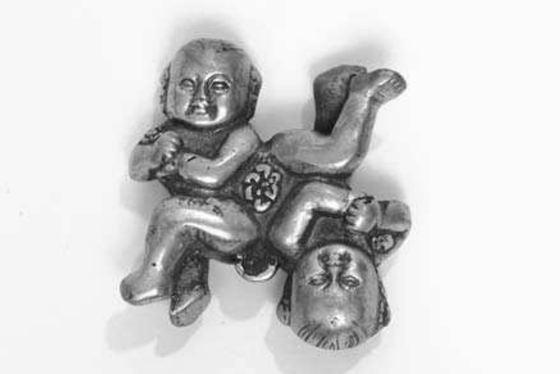

upside-down picture An upside-down figure from China.
Sue & Brian Young/Mr. Puzzle Australia, www.mrpuzzle.com.au

through 180° and the second part read from the same boxes in reverse order. As if by magic, Little Lady Lovekins transforms into Old Man Muffaroo, a giant fish becomes a giant bird, and a pouncing tiger turns into a tiger buried under a pile of stones.

Ussher, James (1581–1656)

An Irish clergyman, born in Fishamble Street, Dublin, who studied at Trinity College, Dublin, and later became a fellow there. He entered the Church and eventually became Archbishop of Armagh. In 1650 he published his famous assertion that the Creation had taken place "upon the entrance of the night preceding" Sunday, October 23, 4004 B.C.

utility function

The lesson of the **St. Petersburg paradox** is that people do not play games as if they are maximizing the expected monetary value they receive. However, certain rationality assumptions about the way people behave, known as the *von Neumann and Morganstern axioms,* imply that people *do* act as though they are maximizing *something.* This "something" is often referred to as a utility function.

V

vampire number

A natural number x that can be factorized as $y \times z$ in such a way that the number of occurrences of a particular digit in the representation of x in a given **base** (say 10) appears the same number of times in the representations in that same base of y and z together. For example, 2,187 is a vampire number since 2,187 = 21 × 87; similarly 136,948 is a vampire because 136,948 = 146 × 938. Vampire numbers are a whimsical idea that was introduced by Clifford Pickover in 1995.[253]

van der Pol, Balthazar (1889–1959)

A Dutch electrical engineer who began the modern experimental study of **dynamical system**s in the 1920s and '30s. Van der Pol discovered that electrical circuits employing vacuum tubes display stable oscillations, now called limit cycles, but that when these circuits are driven with a signal whose frequency is near that of the limit cycle, the periodic response shifts its frequency to that of the driving signal. The resulting waveform, however, can be quite complicated and contain a rich structure of harmonics and subharmonics. In 1927, van der Pol and his colleague van der Mark reported that an "irregular noise" was heard at certain driving frequencies between the natural entrainment frequencies. It's now clear that, without realizing it, they had described one of the first experimental instances of **chaos**.

vanishment puzzle

A **mechanical puzzle** in which the total area of a collection of pieces, or the number of items in a picture, appear to change following some manipulation. A well-known puzzle, first seen in 1868, involves an 8 × 8 square that is divided up into two triangles and two trapezoids. The pieces are reassembled into an oblong shape that measures 5 × 13. Where did the extra bit of area come from? The answer is that final shape is not perfectly rectangular but instead has a narrow gap that runs the length of one of the diagonals. Most famous of all vanishment puzzles is **Get off the Earth**.

variable

An **unknown** that has no fixed quantitative value.

vector

A quantity that is specified by a number, indicating size or "magnitude," and a direction; for example, "80 kilometers per hour, heading due south." More generally, a vector is any element of a **vector space** and is also a type of **tensor**. A vector is usually shown in a graph or other diagram as an arrow whose length and direction represent the vector's magnitude and direction, respectively. In n-dimensional space, it is easy to deal with vectors algebraically in the form of n-tuples, which are ordered lists (one-dimensional arrays) of n components.

vector space

Also known as a *linear space,* the most fundamental concept in **linear algebra**. It is a generalization of the set of all geometric **vector**s and is used throughout modern mathematics. Like the concepts of **group**, **ring**, and **field**, that of a vector space is entirely abstract.

Venn diagram

A simple way of representing **set**s and subsets, which makes use of overlapping circles. Venn diagrams are named after the Englishman John Venn (1834–1923), a fellow of Cambridge University. Venn was a cleric in the Anglican Church, an authority on what was then called "moral science," the compiler of a massive index of all Cambridge alumni, and a rather mundane mathematician who worked in logic and probability theory. The diagrams he used for representing **syllogism**s appear to have been first called "Venn diagrams" by Clarence Irving in his book *A Survey of Symbolic Logic* in 1918. However, Venn was lucky to be so immortalized. Both Gottfried **Leibniz** and Leonhard **Euler** used very similar forms of representation many years earlier.

vertex

The point where two sides of a closed figure, or two sides of an angle, meet; otherwise known as a *corner.* A cube, for example, has eight vertices.

vertex figure

The **polygon** that appears if a **polyhedron** is **truncate**d at a **vertex**. The vertex figure of a cube, for example, is an equilateral triangle. To ensure consistency, the truncation may be done at the midpoints of the edges.

vertical-horizontal illusion

Vertical lines looks considerably longer than horizontal ones of the same length. For centuries, it has been known

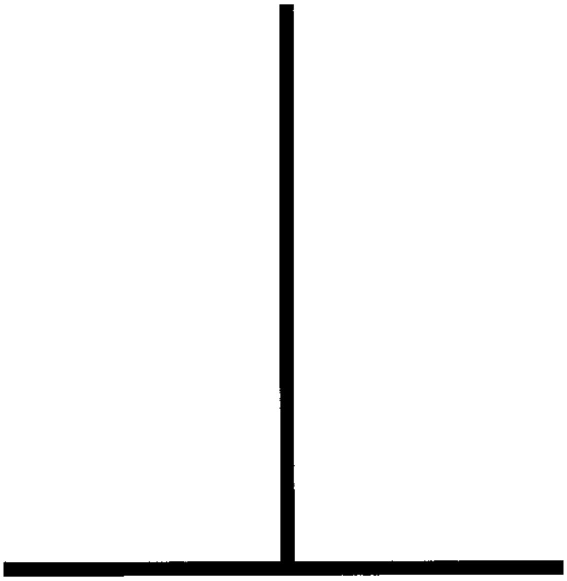

vertical-horizontal illusion

that the height of trees and buildings is perceived to be greater than the horizontal distance between them. This may be because we live in gravity. Rotate the book and you will see that the vertical line in the diagram appears to shrink in relation to what was originally the horizontal line. See also **distortion illusion**.

Vesica Piscis

Literally "fish's bladder," the almond shape formed when two identical circles overlap so that the outside edge of each just touches the center of the other. The Vesica Piscis appears as the Christian fish symbol and also frequently in medieval art and architecture. Most significantly, it provides the template for the pointed Gothic **arch**.

Vickrey auction

A silent auction in which the object being sold goes to the *second-highest bidder.* In a normal silent auction prospective buyers place their bids in sealed envelopes and the highest bid wins. This makes it risky for the seller because if the object is highly valuable and all buyers think they are the only ones who recognize this fact, they may offer much less than they think the object is actually worth. However, a Vickrey auction induces people to bid truthfully. Why? Because when all other bids are fixed and unknown to a given bidder, that bidder's optimal strategy is to bid what she thinks the object is worth. Suppose Alice places a bid for an antique vase. Let V be the amount that Alice thinks the vase is actually worth, and B the bid that she actually makes. Let M be the maximum of all other bids. If M is more than V, then Alice should set her bid B less than or equal to V, so that she does not get the vase for more than she thinks it is worth. If M is less than V, then Alice should set $B = V$, because if she bids any less, she will not get the vase any cheaper, and she may lose it altogether.

vigesimal

Of, relating to, or based on the number 20; the term comes from the Latin *vigesimus* for "twentieth." Mayan arithmetic, which took account of all the toes as well as the fingers, used a vigesimal system. In place of the multiples of 10 used in the decimal system: 1; 10; 100; 1,000; 10,000, . . . , the Mayans dealt in multiples of 20: 1; 20; 400; 8,000; 160,000; . . . .

vinculum

The bar that is placed over repeating decimal fractions to indicate the portion of the pattern that repeats. In the original Latin, *vinculum* referred to a small cord for binding the hands or feet. The symbol was once used in the same way that parentheses and brackets are now used to bind together a group of numbers or symbols. Originally the line was placed under the items to be grouped. What today might be written $7(3x + 4)$ the early users of the vinculum would write $\underline{3x + 4}\ 7$. Sometimes the horizontal fraction bar is called a vinculum as it binds the numerator and denominator into a single value.

Vinogradov's theorem

Every sufficiently large odd number can be expressed as the sum of three **prime numbers**. The theorem was named after the Russian mathematician Ivan Vinogradov (1891–1983) who proved it in 1937. This is a partial solution of the **Goldbach conjecture** and is related to **Waring's conjecture**.

Viviani's curve

The space curve that marks the intersection of the cylinder $(x - a)^2 + y^2 = a^2$ and the sphere $x^2 + y^2 + z^2 = a^2$. It is given by the parametric equations:

$$x = a(1 + \cos t)$$
$$y = a \sin t$$
$$z = 2a \sin(\tfrac{1}{2} t)$$

Viviani's theorem

For a given point inside an equilateral **triangle**, the sum of the perpendicular distances from the point to the sides is equal to the height of the triangle. If the point is outside the triangle, the relationship still holds if one or more of the perpendiculars is treated as a negative value.

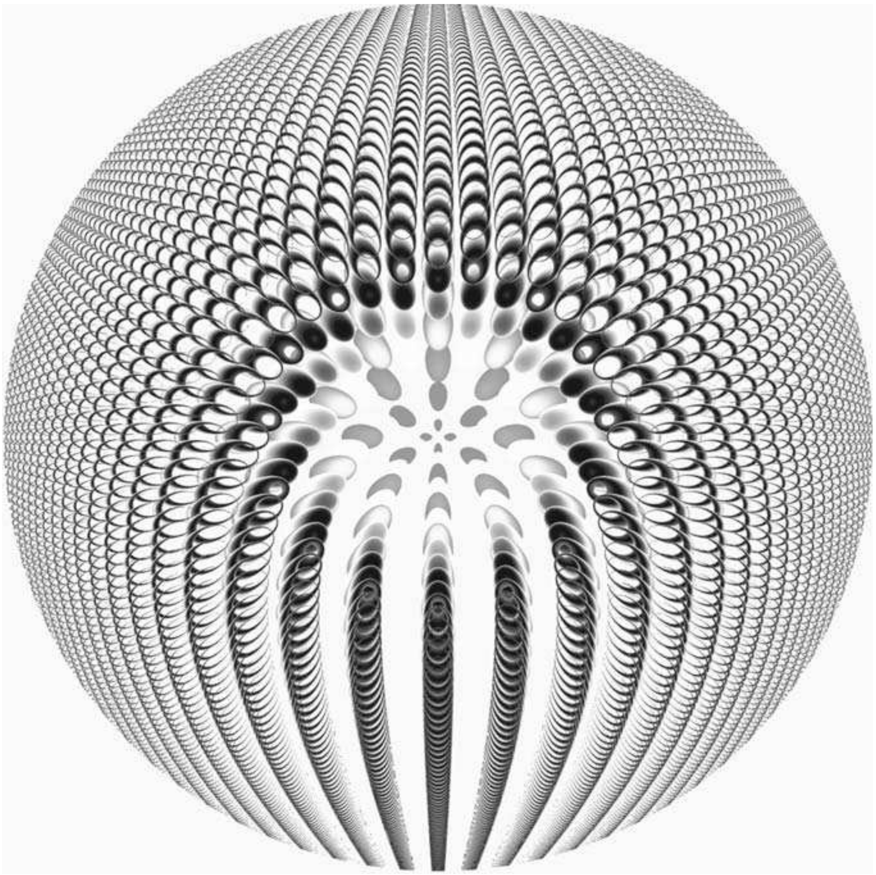

Volvox fractal *Jos Leys, www.josleys.com*

Viviani's theorem generalizes to a regular *n*-sided **polygon**: the sum of the perpendicular distances from an interior point to the *n* sides being *n* times the **apothem** of the figure. The theorem is named for Vincenzo Viviani (1622–1703), a pupil of Galileo and Torricelli, who is also remembered for a reconstruction of a book on the **conic section**s of **Apollonius** and for finding a way of **trisecting an angle** through the use of an equilateral hyperbola.

volume

The measure of space occupied by a solid body.

Volvox fractal

A **fractal** that is similar in appearance to *Volvox*–a unicellular life form that lives in spherical colonies with thousands of members.

von Neumann, John (1903–1957)

> *Young man, in mathematics you don't understand things, you just get used to them.*

A Hungarian-American mathematician who made important contributions to **set theory**, computer science, economics, and **quantum mechanics**. He received a Ph.D. in mathematics from the University of Budapest and later he worked at the Institute for Advanced Study in Princeton. *Theory of Games and Economic Behavior,*[230] which he coauthored with Oskar Morgenstern in 1944, is considered a seminal work in the field of **game theory**. Von Neumann devised the von Neumann architecture used in all modern computers and studied cellular automata (see **cellular automaton**) in order to construct the first examples of self-replicating automata, now known as **von Neumann machine**s. Von Neumann had a mind of great

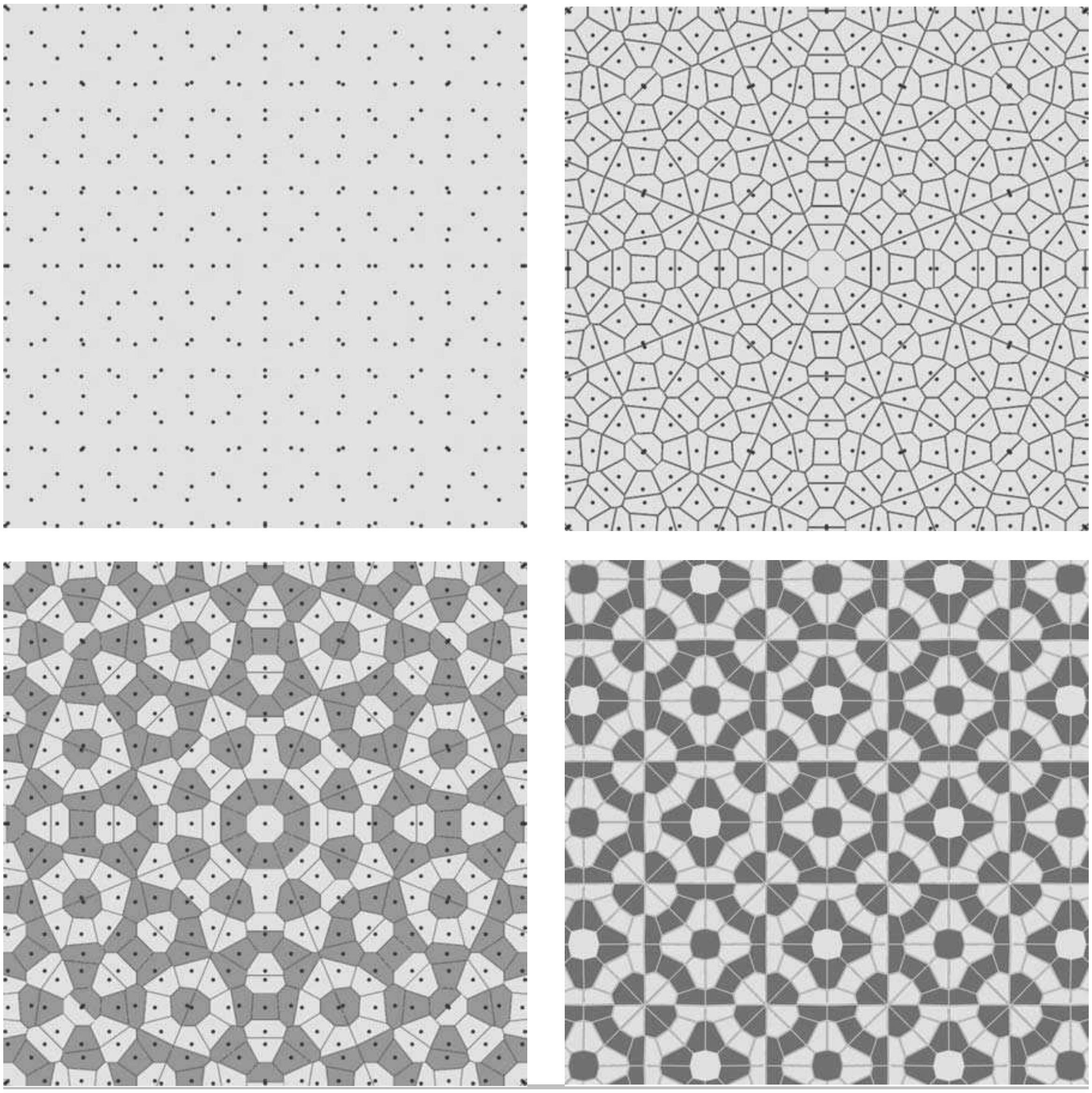

Voronoi diagram Top (left to right): Start from a grid of points. Draw lines that limit areas closest to the points of the grid. Bottom (left to right): Color the areas in some fashion. An infinite variety of patterns is possible. *Jos Leys, www.josleys.com*

ingenuity, nearly total recall of what he'd learned, immense arrogance, and a great love of jokes and humor.[259]

von Neumann machine

(1) A model for a computing machine that uses a single storage structure to hold both the set of instructions on how to perform the computation and the data required or generated by the computation. John **von Neumann** helped to create the model as an example of a general-purpose computing machine. By treating the instructions in the same way as the data, the machine could easily change the instructions. In other words the machine was reprogrammable. (2) A self-replicating machine. In principle, if a machine (e.g., an industrial

robot) could be given enough capability, raw material, and instructions, then that robot could make an exact physical copy of itself. The copy would need to be programmed in order to do anything. If both robots were reprogrammable, then the original robot could be instructed to copy its program to the new robot. Both robots would now have the capability of building copies of themselves. Since such a machine is capable of reproduction, it could arguably qualify as a simple form of life.

Voronoi diagram

Also known as a *Dirichlet tesselation,* a partitioning of space into cells, each of which consists of the points closer to one particular object than to any others. More specifically, in two dimensions a Voronoi diagram consists of breaking up a plane containing *n* points into *n* convex **polygon**s in such a way that each polygon contains exactly one point and every point in a given polygon is closer to its central point than to any other. Voronoi diagrams, their boundaries (known as *medial axes*), and their **dual**s (called *Delaunay triangulations*) have been reinvented, given different names, generalized, studied, and applied many times over in many different fields. Voronoi diagrams tend to be involved in situations where a space should be partitioned into "spheres of influence"; examples include models of crystal and cell growth and protein molecule volume analysis.

vulgar fraction

See **common fraction**.

W

walking in the rain problem

The question is whether you get wetter by walking or running a given distance through rain that is falling at a constant rate. An early appearance of this problem was in Bagley's *Paradox Pie* (1944). A simple answer is that moving faster is better. If the rain falls vertically and the density of water in the air is assumed constant, then, no matter what your speed, you'll sweep out the same volume and will always get the same amount of water hitting your front. However, running rather than walking will reduce the amount of water landing on your head.

Wallis, John (1616–1703)

The most influential English mathematician before Isaac **Newton** and an important contributor to the origins of **calculus**. He was a skilled linguist, was one of the first to proclaim in public Harvey's discovery of the circulation of the blood, and had an extraordinary memory for figures. His *Arithmetica Infinitorum* was described as "the most stimulating mathematical work so far published in England" and introduced the symbol ∞ for infinity (see also **aleph**). It contained the germs of the differential calculus, and it suggested to Newton, who was delighted by it, the **binomial theorem**.

Wallis formula

See **pi**.

wallpaper group

Also known as a *crystallographic group*, a distinct way to tile the plane that repeats indefinitely in two dimensions; that is, a collection of two-dimensional symmetric patterns on a plane surface, containing two nonparallel **translation**s (see **periodic tiling**). There are only 17 kinds of these patterns, known as isometries (see **isometry**), each uniquely identified by its translation and rotation symmetries, as discovered in the late nineteenth century by E. S. **Fedorov** and, independently, by the German A. M. Schoenflies and the Englishman William Barlow. Thirteen of the isometries include some kind of rotational symmetry, while four do not; twelve show rectangular symmetries, while five involve hexagonal symmetries. Every two-dimensional repetitive pattern in wallpaper, textiles, brickwork, or the arrangement of atoms in a plane of a crystal is just a minor variation on one of these 17 patterns.

Wari

See **Mancala**.

Waring's conjecture

A hypothesis given, without proof, by the English mathematician Edward Waring (1734–1798) in his *Meditationes algebraicae* (1770). It states that for every number k, there is another number s such that every natural number can be represented as the sum of s kth powers. For example, every natural number can be written as a sum of 4 squares, 9 cubes and so on. Waring's conjecture was first proven in full by David **Hilbert** in 1909.

weak inequality

An **inequality** that permits the equality case. For example, a is less than or equal to ($\leq$) b.

Weierstrass, Karl Wilhelm Theodor (1815–1897)

A German mathematician who is considered the father of modern **analysis**. Compelled by his father to study law, he instead spent four years at the University of Bonn, fencing, drinking, and reading math. He left under a cloud and

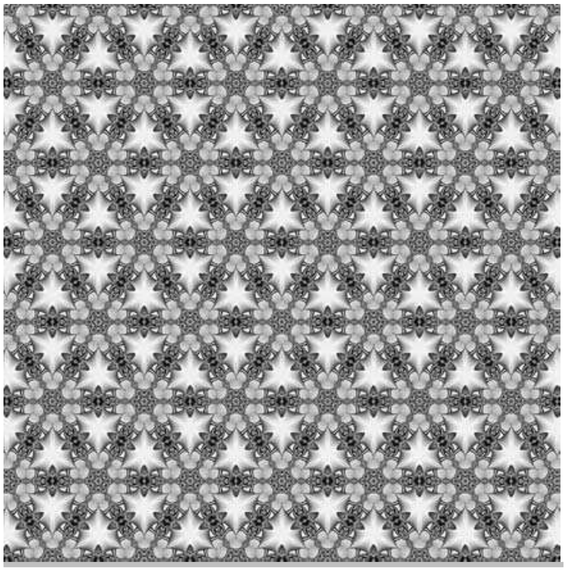

wallpaper group A pattern made from one of the wallpaper groups. *Jos Leys, www.josleys.com* ended up teaching in secondary schools for many years. In 1854 he published a paper, written 14 years earlier when he was fresh out of college, in *Crelle's Journal,* on *Abelian functions* which completed work that Niels **Abel** and Karl **Jacobi** had begun. Its importance was immediately recognized and Weierstrass was appointed a professor at the Royal Polytechnic School and a lecturer at the University of Berlin. He went on to give the first rigorous definitions of **limit**, **derivative**, differentiability, and **convergence**, and investigated under what conditions a power series will converge.

Weierstrass's nondifferentiable function

The earliest known example of a *pathological function*–a **function** that gives rise to a **pathological curve**. It was investigated by Karl **Weierstrass**, but had been first discovered by Bernhard **Riemann**, and is defined as:

$$f(x) = \sum_{k=1}^{\infty} \frac{\sin(\pi k^2 x)}{\pi k^2}$$

The Weierstrass function is everywhere continuous (see **continuity**) but nowhere differentiable; in other words, no **tangent** exists to its curve at any point. Constructed from an infinite sum of trigonometric functions, it is the densely nested oscillating structure that makes the definition of a tangent line impossible.

weighing puzzles

See **measuring and weighing puzzles**.

weird number

See **abundant number**.

Wessel, Caspar (1745–1818)

A Norwegian surveyor whose mathematical fame rests on a single paper, published in 1799, that gave the first geometrical interpretation of **complex number**s. His priority in this discovery, however, went unrecognized for many years. Thus what should really be called a *Wessel diagram* is known instead as an **Argand diagram** after the man whose work on the same subject, published in 1806, first came to the attention of the mathematical world. Wessel's paper, by contrast, wasn't noticed by the mathematical community until 1895 when the Danish mathematician Sophus Juel drew attention to it and, in the same year, Sophus **Lie** republished Wessel's paper. Astonishingly, Wessel's remarkable work was not translated into English until 1999–its bicentenary!

Weyl, Hermann Klaus Hugo (1885–1955)

A German mathematician (known as "Peter" to his close friends) whose work involved symmetry theory, **topology**, and **non-Euclidean geometry**. Weyl studied under David **Hilbert** at Göttingen. Then, as a colleague of Albert Einstein at Zurich 1913, he got involved with **relativity theory** and came to believe (erroneously) that he had found a way to unite gravity and electromagnetism. From 1923 to 1938 he concentrated on **group** theory and made some important contributions to **quantum mechanics**. As the Nazi tide swept over Europe, Weyl came to the United States and spent the rest of his career at the Institute for Advanced Studies at Princeton. He said: "My work always tried to unite the truth with the beautiful, but when I had to choose one or the other, I usually chose the beautiful." See also **beauty and mathematics**.

wff

A well-formed formula.

what color was the bear?

A hunter walks one mile due south, then one mile due east, then one mile north and arrives back at his starting point. He shoots a bear. What color was it? Such a trip is possible if the hunter starts from one of the geographical poles, then circumnavigates the sides of a spherical triangle (see **elliptical geometry**). Since there are no bears in Antarctica, the trip is assumed to have taken place in the Arctic where there are polar bears and thus the answer is "white." Versions of this problem began to appear in the 1940s. A closer examination reveals that there are many more points on the globe, other than an exact pole, from which the hunter could have begun his trek. One example is any point (of which there are an infinite number) on a circle drawn at a distance of slightly more than 1 + $\frac{1}{2}\pi$ mile (about 1.16 miles) from a pole–"slightly more" because of Earth's curvature. But this is not all. The hunter could also satisfy the conditions by starting at points closer to the pole, so that the walk east would carry him exactly twice around the pole, or three times, and so on. Of course, the bear would still be white (except that polar bears don't live that far north!).

wheat and chessboard problem

According to one myth, **chess** was invented by Grand Vizier Sissa Ben Dahir and then given to King Shirham of India. The king was so pleased that he offered his subject a great reward in gold, but the wily vizier said that he would be happy merely to have some wheat: one grain for the first square of the chessboard, two grains for the second square, four for the third, and so on, doubling each time. The king thought this was a very modest request, granted it, and asked for a bag of wheat to be brought in. However the bag was emptied by the twentieth square. The king asked for another bag, but then realized that this entire bag was needed for the next square. In 20 more squares, as many bags would have been exhausted as there

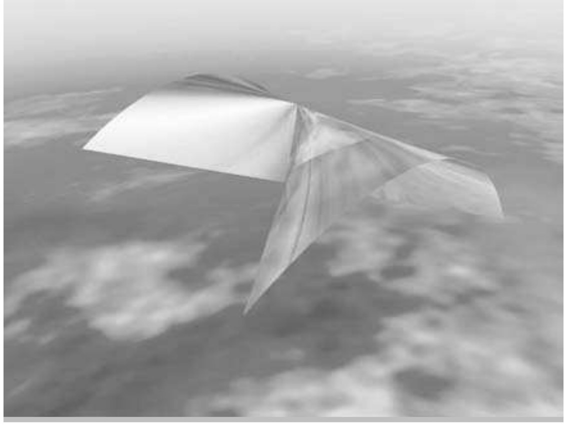

Whitney's umbrella *Takashi Nishimura*

were grains in the first bag! The number of grains on the sixty-fourth square would have been 2^{63}, and the total for the whole board $2^{64} - 1 = 18,446,744,073,709,551,615$. This is more wheat than in the entire whole world; in fact, it would fill a building 40 km long, 40 km wide, and 300 meters tall. See also **Tower of Brahma**.

Whitney's umbrella

A strange-looking, whimsically named geometrical object first studied by Hassler Whitney in the 1940s. It can be pictured as a self-intersecting rectangle in three dimensions. A *pinch point,* also known as a *Whitney singularity* or a *branch point* occurs at the top endpoint of the segment of self-intersection: every neighborhood of the pinch point intersects itself.

Wiener, Norbert (1894–1964)

An American mathematician who established the subject of **cybernetics.** As a precocious youngster, Wiener hopped from subject to subject at college, finally earning a Ph.D. in mathematics from Harvard at 19, before embarking on an even more erratic early career that took him into a variety of activities, including journalism. Having settled upon mathematical research, obtaining a post at the Massachusetts Institute of Technology in 1919, he nevertheless continued to range across fields, from random processes, including ergodic theory (concerned with the onset of chaos in a system), to integral equations, quantum mechanics, and potential theory. Wartime work that involved applying statistical methods to control and communication engineering, led to him extending these studies into control and communication in complex electronic systems and in animals, especially humans–the science of cybernetics.

Wiles, Andrew (1953–)

The English mathematician who, in 1994, finally proved **Fermat's last theorem**. Wiles studied at Oxford (B.A. 1974) and Cambridge (Ph.D. 1977) and has held posts at Cambridge, Oxford, and Princeton. From the mid-1980s his work was focused on proving a proposition known as the *Shimura-Taniyama conjecture,* since from this, it had been shown, Fermat's last theorem would follow. In 1993 he gave a series of lectures at Cambridge University ending on June 23, 1993. At the end of the final lecture he announced he had a proof of Fermat's last theorem. However, when the results were written up for publication, a subtle error was found. Wiles worked hard for about a year, helped in particular by a colleague, R. Taylor, and by September 19, 1994, having almost given up, he decided to have one last try. As he recalled, "suddenly, totally unexpectedly, I had this incredible revelation. It was the most important moment of my working life. . . . [I]t was so indescribably beautiful, it was so simple and so elegant . . . [that] I just stared in disbelief for twenty minutes, then during the day I walked round the department. I'd keep coming back to my desk to see it was still there–it was still there."

In 1994 Wiles was appointed Eugene Higgins Professor of Mathematics at Princeton. His paper that proves Fermat's last theorem is called "Modular elliptic curves and Fermat's Last Theorem" and appeared in the *Annals of Mathematics* in 1995.[296]

Wilson's theorem

Any number p is a **prime number** if, and only if, $(p - 1)! + 1$ is divisible by p. We can easily check this for some small numbers: $(2 - 1)! + 1 = 2$, which is divisible by 2; $(5 - 1)! + 1 = 25$, which is divisible by 5; $(9 - 1)! + 1 = 40,321$, which is not divisible by 9. The theorem is named for Sir John Wilson (1741–1793), who came across it (but left no formal proof) while he was a student at Peterhouse College, Cambridge. Wilson went on to become a judge and seems to have done little else in mathematics. The theorem was first published and named after Wilson by Edward Waring (1734–1798) around 1770. However, it is now clear that the result was known to Gottfried **Leibniz** and perhaps, much earlier to Ibn al-Haytham (965–1040). The first known proof was provided by Joseph **Lagrange**.

winding number

The number of times a **closed** curve in the plane passes around a given point in the counterclockwise direction.

wine

The creation of a fine wine may be a complex business, but it is not mathematically intractable according to a Californian winemaker and businessman, Leo McCloskey, who runs a consulting firm. McCloskey's formula takes into account color, fragrance, and flavor and attempts to predict wine quality two to three years in advance. The formula stems from a comparison of key chemical characteristics for a range of wines with what experts made of the final product. Putting this information into a database, McCloskey has claimed, enables the math of wine flavor to be derived. More than 50,000 wines were analyzed for the database but, in each case, only a handful of the 400 to 500 constituent chemicals were considered crucial flags to future color, taste, and bouquet. The proportion of tannins and phenols is particularly crucial in determining a wine's future, thus reducing the number of essential flags for an accurate mathematical model to just 10 or 20. With this system, McCloskey believes, he can predict what a wine will be like before it is bottled, giving winemakers the opportunity to modify the flavor if necessary. However, whether any wine formula can effectively take into account the subjective impressions of the end user remains to be seen. See also **bottle sizes**.

Witch of Agnesi curve

See **Agnesi**.

word puzzles

1. What positive integer, when spelled out, has a Scrabble score equal to that integer?
2. What is the only English word that ends in "mt?"
3. Only four words in the English language end in "-dous." What are they?
4. What is the shortest complete sentence in the English language?
5. Name the only word (not including proper names) in the English language that has two "i" 's together.
6. Name a one-syllable word that becomes a three-syllable word by adding one letter to the end of it.
7. What word begins and ends with "und?"

Solutions begin on page 369.

word trivia

- The hardest word to define briefly is thought to be the word *mamihlapinatopai* from the Fuegian language spoken by the natives of the Andaman Islands. Its simplest definition is "two people looking at each other without speaking hoping that the other will offer to do something which both parties desire but neither are willing to do."
- No word in the English language rhymes with orange, silver, purple, or month.
- The only 15-letter word that can be written without repeating a letter is *uncopyrightable.*
- The combination "ough" can be pronounced in nine different ways. The following sentence contains them all: "A rough-coated, dough-faced, thoughtful ploughman strode through the streets of Scarborough; after falling into a slough, he coughed and hiccoughed."
- *Facetious* and *abstemious* contain all the vowels in the correct order, as does *arsenious,* meaning "containing arsenic."
- The longest word in the English language, according to the Oxford English Dictionary, is *pneumonoultramicroscopicsilicovolcanoconiosis.* The only other word with the same amount of letters is its plural: *pneumonoultramicroscopicsilicovolcanoconioses.* It means an infection of the lungs.
- The longest English word with one vowel is *strengths.*
- The "sixth sick sheik's sixth sheep's sick" is said to be the toughest tongue twister in the English language.
- Richard Millhouse Nixon was the first U.S. president whose name contains all the letters from the word *criminal.*
- In Scotland, a new game was invented. It was called Gentlemen Only Ladies Forbidden and thus the word GOLF entered the vocabulary.
- The longest English words that don't use any of vowels a, e, i, o, or u are "rhythm" and "syzygy."
- The verb *cleave* is one of many English words with two synonyms which are antonyms of each other: adhere and separate.
- There is a seven letter word in the English language that contains ten words without rearranging any of its letters, *therein:* the, there, he, in, rein, her, here, ere, therein, herein.
- The longest English word that consists entirely of consonants is *crwth,* which is from the fourteenth century and means crowd.
- The word *trivia* comes from the Latin *tri-* + *via,* and means "three streets." This is because in ancient times, at an intersection of three streets in Rome, they would have a type of kiosk where ancillary information was listed. You might be interested in it, you might not, hence they were bits of "trivia."

worldline

The path of an object through **space-time**. On a *Minkowski diagram,* in which the three dimensions of space are represented by the horizontal axis and time is represented by the vertical axis, worldlines appear as wiggly curves extending from the past into the future. *My*

Worldline is the title of the autobiography of George Gamow, the Ukrainian-American physicist.

Wundt illusion

A **distortion illusion** devised by the German "father of experimental psychology," and one-time assistant to the physicist Hermann von Helmholtz at Heidelberg, Wilhelm Wundt (1832–1920). In the figure, two horizontal lines are both straight, although they look as if they bow in at the middle. The distortion is induced by crooked lines on the background, as in **Orbison's illusion**. The simplest of all distortion illusions–the **vertical-horizontal illusion**–was also discovered by Wundt.

Wythoff's game

A variation on the game of **Nim** suggested by W. A. Wythoff in 1907. It is played with two heaps of counters in which a player may take any number from either heap or the same number from both. The player who takes the last counter wins.

X

x

The most frequently used symbol for an unknown in an expression or equation.

x-axis

The horizontal axis of a two-dimensional plot in **Cartesian coordinates**.

X-pentomino

A pentomino, a five-square **polyomino**, in the shape of the letter X.

Y

yard
A unit of distance equal to 3 feet (36 inches) or 0.9144 meters. The name originates with the old German word *gazdaz* for a staff or stick, which could be used for measurement. This changed to *gierd* in Old English and eventually to *yard*. The yard-arm of a sailing ship–a tapered spar used to support a square sail–harkens back to the earlier meaning. In France, the equivalent of the yard measure is called a "verge," from the Latin *virga* for a stick or rod.

***y*-axis**
The vertical axis of a two-dimensional plot in **Cartesian coordinates**.

yin-yang symbol
See **Great Monad**.

yocto-/yotta-
The prefix yocto-, for 10^{-24}, derives from the Greek *oktakis* ("eight times"). The prefix yotta-, for 10^{24}, comes from the same source.

Z

Zeller's formula

A formula invented by the German clergyman Christian Zeller (1824–1899) for figuring out the day of the week of a given date, without the need of tables. Let

J be the century-number,
K the last two digits in the year,
e the residue, which remains when J is divided by 4,
m the number of the month,
q the day of the month,
h the number of the day of the week.

Then h is the remainder that results when

$$h = q + 26(m + 1)/10 + K + K/4 - 2e$$

is divided by 7. For the formula to work, January and February have to be taken as months 13 and 14 of the preceding year. For example, Frederick the Great was born on 24 January 1712, so $J = 17$, $e = 1$, $K = 11$ (not 12 because of the special way of numbering the months), $m = 13$, and $q = 24$. Plugging these values into the formula, we get

$$\begin{aligned} &24 + 26(13 + 1)/10 + 11 + 11/4 - 2 \times 1 \\ &\quad = 71 = 70 \times 1 + 1. \end{aligned}$$

The remainder h is 1, so Frederick the Great was born on the first day of the week, Sunday.

Zenodorus (c. 180 B.C.)

The Greek mathematician and philosopher who proposed in *On Isometric Figures* that the circle has the maximum area of all isoperimetric figures in a plane, and that the sphere has the maximum volume of all bodies with equal surface.

Zeno's paradoxes

A series of **paradox**es posed by the philosopher Zeno of Elea (c. 490–c. 425 B.C.). Little is known about Zeno's life. He was born in Elea (now Lucania) in southern Italy and was a friend and student of Parmenides. None of his writings survive but he is known to have written a book, which **Proclus** says contained 40 paradoxes. Four of these, which all concern motion, have had a profound influence on the development of mathematics. They are described in Aristotle's great work *Physics* and are called the Dichotomy, Achilles (and the Tortoise), the Arrow, and the Stadium.

The Dichotomy argues that "there is no motion because that which is moved must arrive at the middle of its course before it arrives at the end." In order to traverse a line segment it's necessary to reach the halfway point, but this requires first reaching the quarter-way point, which first requires reaching the eighth-way point, and so on without end. Hence motion can never begin. This problem isn't alleviated by the well-known infinite sum $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 1$ because Zeno is effectively insisting that the sum be tackled in the reverse direction. What is the first term in such a series?

Zeno's paradox of Achilles is told by Aristotle in this way: "The slower when running will never be overtaken by the quicker; for that which is pursuing must first reach the point from which that which is fleeing started, so that the slower must necessarily always be some distance ahead." Thus, Achilles, however fast he runs, will never catch the plodding Tortoise who started first. And yet, of course, in the real world, faster things do overtake slower ones. So how is the paradox to be solved? The German set theorist Adolf Frankel (1891–1965) is one of many modern mathematicians (Bertrand Russell is another) who have pointed out that 2,000 years of attempted explanations have not cleared away the mysteries of Zeno's paradoxes: "Although they have often been dismissed as logical nonsense, many attempts have also been made to dispose of them by means of mathematical theorems, such as the theory of convergent series or the theory of sets. In the end, however, the difficulties inherent in his arguments have always come back with a vengeance, for the human mind is so constructed that it can look at a continuum in two ways that are not quite reconcilable."

zepto-/zetta-

The prefix zepto-, for 10^{-21}, derives from the Greek *heptakis* ("seven times"). The prefix zetta-, for 10^{21}, comes from the same source.

zero

The **integer**, denoted 0, which, when used as a counting number, indicates that no objects are present. It is the only integer that is neither negative nor positive. Zero is both a **number** and a **numeral**. The number zero is the size of the **empty set** but it is not the empty set itself, nor is it the same thing as **nothing**. The numeral or digit zero is used in positional **number system**s, where the position of a digit signifies its value, with successive positions having higher values, and the digit zero is used to skip a position. The earliest roots of the numeral zero stretch back 5,000 years to the Sumerians in Mesopotamia, who inserted a slanted double wedge between cuneiform characters for numbers, written positionally, to indicate a number's absence. The symbol changed over time as positional notation made its way to India, via the Greeks (in whose own culture zero made a late and only occasional appearance). Our word *zero* derives from the Hindi *sunya* for "void" or "emptiness," through the Arabic *sifr* (which also gives us *cipher*), and the Italian *zevero*. As a number in its own right, aside from its use as a position marker, zero took a much longer time to become established, and even now is not equal in status to other numbers: division by zero is not allowed.

zero divisors

Nonzero elements of a **ring** whose product is 0.

zero of a function

See **root**.

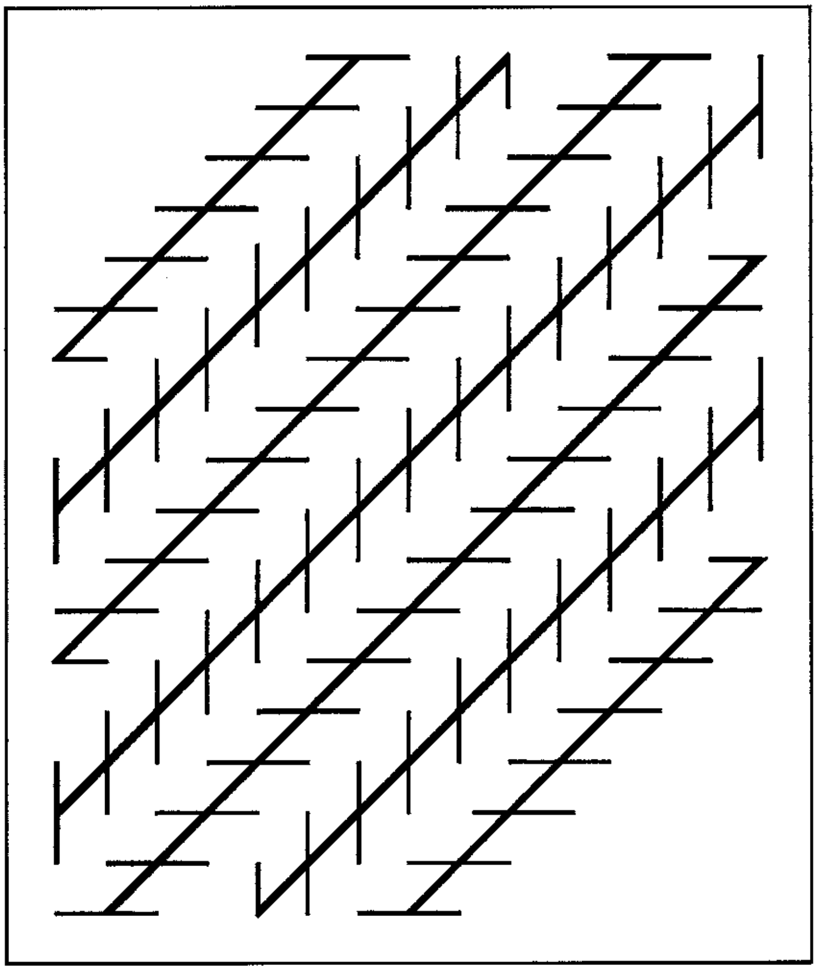

Zöllner illusion

zero-sum game

In **game theory**, a game in which a win for one player results in an equal but opposite loss for the other players.

zeta function

A **function** that has certain properties and is calculated as an infinite sum of negative powers. The most commonly encountered zeta function is the **Riemann zeta function**.

zigzag

A general word for a type of curve that consists of several straight lines joined at points. A zigzag usually goes alternately from side to side.

Zöllner illusion

A line **distortion illusion** first published by the astronomer Johann Zöllner in 1860. The diagonal lines, although parallel, appear not to be. The illusion was one of a series specifically designed to cause errors in optical equipment of that time. They did cause errors, and also great concern among scientists over the validity of all human observations. See also **Poggendorff illusion**.

zombie

A hypothetical being that behaves like us and may share our functional organization and even, perhaps, our neurophysiological makeup, but lacks consciousness or any form of subjective awareness. The concept is used in discussions of **artificial intelligence**.

zone

The portion of a **sphere** between two parallel planes.

zonohedron

A **polyhedron** in which the faces are all **parallelogram**s or parallel-sided. The faces of a zonohedron can be grouped into *zones*–encircling bands of faces which share a common edge direction (and length).

References

1. Abbott, Edwin A. *Flatland: A Romance of Many Dimensions.* London: Seely and Co., 1884. Reprint, Mineola, N.Y.: Dover, 1992.

2. Abraham, R. M. *Diversions and Pastimes.* London: Constable & Co., 1933. Reprinted, 1964.

3. Ackermann, A. S. E. *Scientific Paradoxes and Problems.* London, 1925.

4. Adams, Douglas. *Life, the Universe and Everything.* New York: Harmony Books, 1982.

5. Adams, Douglas. *The Hitchhiker's Guide to the Galaxy.* New York: Ballantine, 1995.

6. Ahrens, W. *Mathematische Unterhaltungen und Spiele.* Leipzig, Germany: Teubner, 1910.

7. Ainley, Stephen. *Mathematical Puzzles.* London: Bell, 1977.

8. Allais, M. "Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine." *Econometrica,* 21: 503–546 (1953).

9. Amthor, A., and B. Krumbiegel. "Das Problema bovinum des Archimedes." *Zeitschrift für Mathematik und Physik,* 25: 121–171 (1880).

10. Andrews, William S. *Magic Squares and Cubes,* 2nd ed. Mineola, N.Y.: Dover, 1960.

11. Apéry, R. "Irrationalité de ζ(2) et ζ(3)." *Astérisque,* 61: 11–13 (1979).

12. Appel, K., and W. Haken. "The Solution of the Four-Color Map Problem." *Scientific American,* 237: 108–121 (1977).

13. ApSimon, Hugh. *Mathematical Byways in Ayling, Beeling, and Ceiling.* New York: Oxford University Press, 1984.

14. Argand, R. *Essai sur une manière de représenter les quantités imaginaires dans les constructions géométriques.* Paris: Albert Blanchard, 1971. Reprint of the 2nd ed., published by G. J. Hoel in 1874. First published in 1806 in Paris.

15. Arno, Steven. "A Note on Perrin Pseudoprimes." *Mathematics of Computation,* 56(193): 371–376 (1991).

16. Arrow, Kenneth. *Social Choice and Individual Values.* New York: John Wiley & Sons, 1951.

17. Ascher, Marcia. *Ethnomathematics: A Multicultural View of Mathematical Ideas.* Pacific Grove, Calif.: Brooks/Cole Publishing Co., 1991.

18. Avedon, Elliot M., and Brian Sutton-Smith. *The Study of Games.* New York: John Wiley & Sons, 1971.

19. Averbach, Bonnie, and Orin Chein. *Mathematics: Problem Solving through Recreational Mathematics.* New York: W. H. Freeman, 1980.

20. Bagley, William A. *Puzzle Pie: A Unique Collection of Scientific Paradoxes, Posers and Oddities.* London: Vawser and Wiles, 1944.

21. Bailey, D. H., P. B. Borwein, and S. Plouffe. "On the Rapid Computation of Various Polylogarithmic Constants." *Mathematics of Computation,* 66(218): 903–913 (1997).

22. Bakst, Aaron. *Mathematical Puzzles and Pastimes.* New York: Van Nostrand, 1954.

23. Bakst, Aaron. *Mathematics: Its Magic and Mystery,* 3rd ed. Princeton, N.J.: Van Nostrand, 1967.

24. Ball, Walter William Rouse. *Mathematical Recreations and Problems.* London: Macmillan, 1892. (Ball, W. W. Rouse and H. S. M. Coxeter. *Mathematical Recreations and Essays,* 13th ed. Mineola, N.Y.: Dover, 1987.)

25. Banchoff, Thomas F. *Beyond the Third Dimension: Geometry, Computer Graphics, and Higher Dimensions.* New York: W. H. Freeman, 1990.

26. Barnsley, Michael. *Fractals Everywhere,* 2nd edition. San Francisco: Morgan Kaufmann, 1993.

27. Bayer, D., and P. Diaconis. "Trailing the Dovetail Shuffle to Its Lair." *Annals of Applied Probability,* 2(2): 294–313 (1992).

28. Beasley, John D. *The Ins and Outs of Peg Solitaire.* New York: Oxford University Press, 1985.

29. Beckmann, P. *A History of Pi,* 3rd ed. New York: Dorset Press, 1989.

30. Beiler, Albert H. *Recreations in the Theory of Numbers.* Mineola, N.Y.: Dover, 1964.

31. Bell, A. H. " 'Cattle Problem.' By Archimedes 251 B.C." *American Mathematical Monthly,* 2: 140 (1895).

32. Bell, E. T. *Men of Mathematics.* New York: Simon & Schuster, 1937.

33. Bell, Robbie, and Michael Cornelius. *Board Games Round the World: A Resource Book for Mathematical Investigations.* New York: Cambridge University Press, 1990.

34. Benford, F. "The Law of Anomalous Numbers."

Proceedings of the American Philosophical Society, 78: 551–572 (1938).

35. Bergholt, E. *Complete Handbook to the Game of Solitaire on the English Board of Thirty-Three Holes.* Routledge: London, 1920.

36. Berlekamp, Elwyn R., John Horton Conway, and Richard K. Guy. *Winning Ways for Your Mathematical Plays,* 2 vols. New York: Academic Press, 1982.

37. Berloquin, Pierre. *Le Jardin du Sphinx.* Paris: Dunod, 1981.

38. Berrondo, Marie. *Mathematical Games.* Englewood Cliffs, N.J.: Prentice Hall, 1983.

39. Besicovitch, A. S. "On Kakeya's Problem and a Similar One." *Mathematische Zeitschrift,* 27: 312–320, (1928).

40. Berry, Michael. "Quantum Physics on the Edge of Chaos." *New Scientist,* 116: 44–47 (1987).

41. Birtwistle, Claude. *The Calculator Puzzle Book.* New York: Bell Publishing Co., 1978.

42. Blatner, David. *The Joy of Pi.* New York: Walker, 1997.

43. Blocksma, Mary. *Reading the Numbers: A Survival Guide to the Measurements, Numbers, and Sizes Encountered in Everyday Life.* New York: Viking Penguin, 1989.

44. Blyth, Will. *Match-Stick Magic, Puzzles, Games and Conjuring Tricks.* London: C. A. Pearson Ltd., 1921.

45. Bogoshi, J., Naidoo, K., and Webb, J. "The Oldest Mathematical Artifact." *Mathematical Gazette,* 71: 458 (1987).

46. Bond, Raymond T. *Famous Stories of Code and Cipher.* New York: Rinehart & Co., 1947.

47. Bondi, Hermann. "The Rigid Body Dynamics of Unidirectional Spin." *Proceedings of the Royal Society,* A405: 265–274, 1986.

48. Bord, Janet. *Mazes and Labyrinths of the World.* London: Latimer, 1976.

49. Borel, Emile. *Algebre et Calcul des Probabilites,* vol. 184. Paris: Comptes Rendus Academie des Sciences, 1927.

50. Bouton, Charles L. "Nim, a Game with a Complete Mathematical Theory." *Annals of Mathematics,* Series 2. 3: 35–39, (1901–1902).

51. Brams, S. J., and A. D. Taylor. "An Envy-Free Cake Division Protocol." *American Mathematical Monthly,* 102: 9–19 (1995).

52. Brams, S. J., and A. D. Taylor. *Fair Division: From Cake-Cutting to Dispute Resolution.* New York: Cambridge University Press, 1996.

53. Brandreth, Gyles. *Numberplay.* New York: Rawson, 1984.

54. Branges, L., de, "A Proof of the Bieberbach Conjecture." *Acta Mathematica,* 154: 137–152 (1985).

55. Burger, Dionys. *Sphereland: A Fantasy about Curved Spaces and an Expanding Universe.* New York: Barnes & Noble, 1983.

56. Cadwell, J. H. *Topics in Recreational Mathematics.* New York: Cambridge University Press, 1966.

57. Carroll, Lewis. *Mathematical Recreations of Lewis Carroll.* 2 vols. Mineola, N.Y.: Dover, 1958.

58. Carroll, Lewis. *Symbolic Logic and the Game of Logic.* 2 vols. Mineola, N.Y.: Dover, 1958.

59. Carroll, Lewis. *Pillow Problems and A Tangles Tale.* New York: Dover, 1958.

60. Chaitin, G. J. "The Berry Paradox." *Complexity,* 1: 26–30 (1995).

61. Chow, T. Y. "The Surprise Examination or Unexpected Hanging Paradox." *American Mathematical Monthly,* 105: 41–51 (1998).

62. Church, Alonzo. *Introduction to Mathematical Logic.* Princeton, N.J.: Princeton University Press, 1956.

63. Clarke, Arthur C. *The City and the Stars.* New York: Harcourt, Brace, 1956.

64. Coffin, Stewart T. *The Puzzling World of Polyhedral Dissections.* New York: Oxford University Press, 1990.

65. Collins, A. Frederick. *Fun with Figures.* New York: D. Appleton & Co., 1928.

66. Conway, J. H. "Mrs. Perkins's Quilt." *Proceedings of the Cambridge Philosophical Society,* 60: 363–368, 1964.

67. Conway, John Horton. *On Numbers and Games.* New York: Academic Press, 1976.

68. Conway, John Horton, and Richard K. Guy. *The Book of Numbers.* New York: Springer-Verlag, 1996.

69. Conway, J. H., and S. P. Norton. "Monstrous Moonshine." *Bulletin London Mathematical Society,* 11: 308–339 (1979).

70. Conway, J. H., and N. J. A. Sloane. "The Monster Group and its 196884-Dimensional Space" and "A Monster Lie Algebra?" Chs. 29–30 in *Sphere Packings Lattices, and Groups,* 2nd ed. New York: Springer-Verlag, pp. 554–571, 1993.

71. Cooper, Necia Grant (ed.). *From Cardinals to Chaos: Reflections on the Life and Legacy of Stanislaw Ulam.* Cambridge: Cambridge University Press, 1989.

72. Coxeter, H. S. M. *Non-Euclidean Geometry.* Toronto: The University of Toronto Press, 1942.

73. Coxeter, H. S. M., M. S. Longuet-Higgins, and J. C. P. Miller. "Uniform Polyhedra." *Philosophical Transactions of the Royal Society of London,* Series, *A* 246: 401–450, 1954.

74. Coxeter, H. S. M. *Introduction to Geometry,* 2nd ed. New York: John Wiley & Sons, 1969.

75. Coxeter, H. S. M. *Regular Polytopes,* 3rd ed. New York: Dover, 1973.

76. Coxeter, H. S. M., P. DuVal, H. T. Flather, and J. F. Petrie. *The Fifty-Nine Icosahedra.* Toronto: University of

Toronto Press, 1938. Reprint, New York: Springer-Verlag, 1982.

77. Coxeter, H. S. M., M. Emmer, R. Penrose, and M. L. Teuber, (eds). *M. C. Escher: Art and Science.* New York: North-Holland, 1986.

78. Cromwell, Peter R. *Polyhedra.* Cambridge: Cambridge University Press, 1997.

79. Császár, Ákos. "A polyhedron without diagonals." *Acta Univ Szegendiensis, Acta Scientia Math,* 13: 140–142, 1949.

80. Cundy, H. Martyn, and A. P. Rollett. *Mathematical Models,* 3rd ed. Oxford: Clarendon Press, 1961; Diss, England: Tarquin Publications, 1981.

81. Dewdney, A. K. *The Planiverse: Computer Contact with a Two-dimensional World.* New York: Simon & Schuster, 1984.

82. Dickson, Leonard E. *History of the Theory of Numbers.* 3 vol. London: Chelsea Pub. Co., 1919–23. Reprint, Providence, R.I.: American Mathematical Society, 1999.

83. Domoryad, A. P. *Mathematical Games and Pastimes.* Elmsford, N.Y.: Pergamon Press, 1963.

84. Dötzel, Gunter. "A Function to End All Functions." *Algorithm: Recreational Programming 2.4,* 16–17 (1991).

85. Doxiadis, Apostolos K. *Uncle Petros and Goldbach's Conjecture.* New York: Bloomsbury, 2000.

86. Dresner, Simon (ed.). *Science World Book of Brain Teasers.* New York: Scholastic Book Services, 1962.

87. Dudeney, H. E. *The Canterbury Puzzles.* London: Nelson, 1907. Reprinted Mineola, N.Y.: Dover, 1958.

88. Dudeney, H. E. *Amusements in Mathematics.* New York: Dover, 1917. Reprinted Mineola, N.Y.: Dover, 1958.

89. Dudeney, Henry E. *The World's Best Word Puzzles.* London: Daily News, 1925.

90. Dudeney, H. E. *536 Puzzles and Curious Problems,* ed. by Martin Gardner. New York: Charles Scribner, 1967.

91. Edgeworth, R., Dalton, B. J., and Parnell, T. "The Pitch Drop Experiment." *European Journal of Physics,* 5: 198–200 (1984).

92. Eiss, Harry Edwin. *Dictionary of Mathematical Games, Puzzles, and Amusements.* Westport, Conn.: Greenwood Press, 1988.

93. Elffers, J. *Tangram—The Ancient Chinese Puzzle.* New York: Harry N. Abrams, 1979.

94. Escher, M. C. *The Graphic Work of M. C. Escher,* New York: Ballantine, 1971.

95. Ewing, John, and Kosniowski, Czes. *Puzzle It Out: Cubes, Groups and Puzzles.* New York: Cambridge University Press, 1982.

96. Fadiman, Clifton. *Fantasia Mathematica: Being a set of stories, together with a group of oddments and diversions, all drawn from the universe of mathematics.* New York: Simon and Schuster, 1958.

97. Fadiman, Clifton. *The Mathematical Magpie: Being more stories, mainly transcendental, plus subsets of essays, rhymes, music, anecdotes, epigrams and other prime oddments and diversions, rational or irrational, all derived from the infinite domain of mathematics.* New York: Simon and Schuster, 1962.

Federico, P. J. "The Melancholy Octahedron." *Mathematics Magazine,* 45: 30–36, 1972.

98. Filipiak, Anthony S. *100 Puzzles: How to Make and Solve Them.* New York: A. S. Barnes, 1942.

99. Filipiak, Anthony S. *Mathematical Puzzles and Other Brain Twisters.* New York: A. S. Barnes, 1964.

100. Fink, Thomas M., and Yong Mao Mao. "Designing Tie Knots Using Random Walks." *Nature,* 398: 31 (1999).

101. Fisher, Adrian. Arian Fisher Maze Design Web site: http://www.mazemaker.com.

102. Fisher, John (ed.). *The Magic of Lewis Carroll.* New York: Simon and Schuster, 1973.

103. Foster, James E. *Mathematics as Diversion.* Astoria, IL: Fulton County Press, 1978.

104. Fraser, James. "A new visual illusion of direction." *British Journal of Psychology,* 2: 307–337 (1908).

105. Frederickson, Greg N. *Dissections: Plane and Fancy.* Cambridge: Cambridge University Press, 1997.

106. Frey, Alexander H., Jr., and David Singmaster. *Handbook of Cubik Math.* Hillside, N.J.: Enslow, 1982.

107. Gamow, G., and M. Stern. *Puzzle-Math.* New York: Viking, 1958.

108. Gardiner, M. *Mathematical Puzzling.* New York: Oxford University Press, 1987.

109. Gardner, Martin. *Mathematics, Magic and Mystery.* Mineola, N.Y.: Dover, 1956.

110. Gardner, M. *Mathematical Puzzles and Diversions.* New York: Simon and Schuster, 1959.

111. Gardner, Martin, ed. *The Mathematical Puzzles of Sam Loyd,* 2 vols. Mineola, N.Y.: Dover, 1959.

112. Gardner, M. *The Second Scientific American Book of Mathematical Puzzles & Diversions.* New York: Simon and Schuster, 1961.

113. Gardner, Martin. "The Remarkable Lore of the Prime Numbers," *Scientific American,* March 1964.

114. Gardner, M. *New Mathematical Diversions from Scientific American.* New York: Simon and Schuster, 1966.

115. Gardner, M. "Mathematical Games." *Scientific American,* 237, 18–28, Nov. 1977.

116. Gardner, Martin. *Mathematical Magic Show: More Puzzles, Games, Diversions, Illusions and Other Mathematical Sleights-of-Mind from Scientific American.* New York: Vintage, 1978.

117. Gardner, Martin. *The Ambidextrous Universe.* New York: Charles Scribner's Sons, 1978.

118. Gardner, Martin. *Mathematical Circus: More Games, Puzzles, Paradoxes, and Other Mathematical*

Entertainments from Scientific American. New York: Alfred A. Knopf, 1979.

119. Gardner, Martin. *Science Fiction Puzzle Tales.* New York: Clarkson N. Potter, 1981.

120. Gardner, Martin. *Wheels, Life, and other Mathematical Amusements.* New York: W. H. Freeman, 1983.

121. Gardner, Martin. *Martin Gardner's Sixth Book of Mathematical Diversions from Scientific American.* Chicago: University of Chicago Press, 1983.

122. Gardner, Martin. *The Magic Numbers of Dr. Matrix.* Buffalo, N.Y.: Prometheus Books, 1985.

123. Gardner, M. *Knotted Doughnuts and Other Mathematical Entertainments.* New York: W. H. Freeman, 1986.

124. Gardner, Martin. *Riddles of the Sphinx And Other Mathematical Puzzle Tales.* Washington, DC: Mathematical Association of America, 1987.

125. Gardner, Martin. *Time Travel and Other Mathematical Bewilderments.* New York: W. H. Freeman, 1987.

126. Gardner, Martin. *Hexaflexagons and Other Mathematical Diversions.* Chicago: University of Chicago Press, 1988.

127. Gardner, Martin. *Penrose Tiles and Trapdoor Ciphers . . . and the Return of Dr. Matrix.* New York: W. H. Freeman, 1989.

128. Gardner, Martin. *Mathematical Carnival.* Washington, D.C.: Mathematical Association of America, 1989.

129. Gardner, Martin. *Mathematical Magic Show.* Washington, D.C.: Mathematical Association of America, 1990.

130. Gardner, M. *Fractal Music, Hypercards, and More Mathematical Recreations from Scientific American Magazine.* New York: W. H. Freeman, 1992.

131. Gardner, Martin (introduction and notes). *The Annotated Alice: Alice's Adventures in Wonderland and Through the Looking Glass.* New York: Random House, 1998.

132. Gasser, Ralph. "Solving Nine Men's Morris." *Computational Intelligence,* 12: 24–41 (1996).

133. Gerver, J. L. "On Moving a Sofa Around a Corner." *Geometriae Dedicata,* 42: 267–283 (1992).

134. Gillam, B. "Geometrical Illusions." *Scientific American,* 242: 102–111 (Jan. 1980).

135. Gleick, James. *Chaos: Making a New Science.* New York: Penguin, 1988.

136. Golomb, S. W. "Checker Boards and Polyominoes." *American Mathematical Monthly,* 61: 675–682 (1954).

137. Golomb, Solomon W. *Polyominoes: Puzzles, Patterns, Problems, and Packings.* New York: Charles Scribner's, 1965. Reprint, Princeton, N.J.: Princeton University Press, 1995.

138. Gomme, Alice Bertha. *Traditional Games of England, Scotland and Ireland.* 2 vols. London: Thames and Hudson, 1894–1898.

139. Gotze, Heinz. *Castel del Monte, Geometric Marvel of the Middle Ages.* New York: Prestel, 1998.

140. Greenberg, Marvin J. *Euclidean and non-Euclidean Geometries: Development and History.* San Francisco: W. H. Freeman, 1974.

141. Greenblatt, M. H. *Mathematical Entertainments.* London: George Allen and Unwin, 1968.

142. Gregory, R. L. "Analogue Transactions with Adelbert Ames." *Perception,* 16: 277–282 (1987).

143. Grunbaum, B., and G. C. Shephard. *Tilings and Patterns.* New York: W. H. Freeman, 1987.

144. Guy, Richard K. *Unsolved Problems in Number Theory,* 2nd ed. New York: Springer-Verlag, 1994.

145. Guy, Richard K., and Robert E. Woodrow (eds). *The Lighter Side of Mathematics: Proceedings of the Eugène Strens Memorial Conference on Recreational Mathematics and its History.* Washington, D.C.: Mathematical Association of America, 1994.

146. Haeckel, Ernst. Kunstformen der Natur (Art Forms in Nature). Leipzig, Wien: Bibliographisches Institut 1899–1904. Reprinted Mineola, N.Y.: Dover, 1974.

147. Haldeman-Julius, E. *Problems, Puzzles and Brain-Teasers.* Girard, KS: Appeal to Reason, 1937.

148. Hall, Trevor H. *Old Conjuring Books.* London: Duckworth, 1972.

149. Hankins, Thomas L. *Sir William Rowan Hamilton.* Baltimore: Johns Hopkins University Press, 1980.

150. Hardy, G. H. *Ramanujan: Twelve Lectures on Subjects Suggested by His Life and Work.* London: Cambridge University Press, 1940.

151. Hardy, G. H. *A Mathematician's Apology.* London: Cambridge University Press, 1940.

152. Hargrave, Catherine Perry. *A History of Playing Cards and a Bibliography of Cards and Gaming.* New York: Houghton Mifflin, 1930. Reprint, New York: Dover, 1966.

153. Harmer, G. P., and D. Abbott. "Losing Strategies Can Win by Parrondo's Paradox." *Nature,* 402: 864 (1999).

154. Hartley, Miles C. *Patterns of Polyhedra.* Ann Arbor, MI: Edwards Brothers, 1957.

155. Heafford, Philip. *The Math Entertainer.* New York: Vintage Press, 1983.

156. Heinlein, Robert. "And He Built a Crooked House," in Isaac Asimov (ed.) *Where Do We Go From Here?* New York: Doubleday, 1971.

157. Heinzelin, J., de, "Ishango." *Scientific American,* 206(6): 105–116 (Jun. 1962).

158. Held, Richard (ed.). *Image, Object and Illusion.* San Francisco: W. H. Freeman, 1974.

159. Henderson, Linda Dalrymple. *The Fourth Dimension and Non-Euclidean Geometry in Modern Art.* Princeton, N.J.: Princeton University Press, 1983.

160. Hendricks, John R. "Magic Tesseracts and N-Dimensional Magic Hypercubes." *Journal of Recreational Mathematics.* 6(3) (Summer 1973).

161. Hendricks, John R. "Black and White Vertices of a Hypercube." *Journal of Recreational Mathematics,* 11(4), 1978–1979.

162. Hendricks, John R. "Magic Hypercubes." Winnipeg, Canada: self-published pamphlet, 1988.

163. Hendricks, John R. "Images from the Fourth Dimension." Winnipeg, Canada: self-published pamphlet, 1993.

164. Hill, T. P. "The First Digit Phenomenon." *American Scientist,* 86: 358–363 (1998).

165. Hilton, Peter, and Jean Pedersen. *Build Your Own Polyhedra.* New York: Addison Wesley, 1988.

166. Hinton, Charles Howard. *A New Era of Thought.* London: Swan Sonnenschein, 1888.

167. Hinton, Charles Howard. *The Fourth Dimension.* London: Allen & Unwin, 1904.

168. Hodges, Andrew, and Douglas Hofstadter. *Alan Turing: The Enigma.* New York: Walker & Co., 2000.

169. Hoffman, Paul. *Archimedes' Revenge: The Joys and Perils of Mathematics.* New York: W. W. Norton, 1988.

170. Hoffman, Paul. *The Man Who Loved Only Numbers: The Story of Paul Erdos and the Search for Mathematical Truth.* New York: Hyperion, 1998.

171. Hoffmann, Professor (Angelo Louis). *Puzzles Old and New.* London: Frederick Warne, 1893. Reprint, L. Edward Hordern, 1993.

172. Hofstadter, Douglas R. *Gödel. Escher, Bach: An Eternal Golden Braid.* New York: Basic Books, 1999.

173. Hogben, Lancelot. *Mathematics for the Million.* New York: W. W. Norton, 1933.

174. Holden, Alan. *Shapes, Spaces and Symmetry.* New York: Columbia University Press, 1971. Reprint, Mineola, N.Y.: Dover, 1991.

175. Hordern, Edward. *Sliding Piece Puzzles.* New York: Oxford University Press, 1986.

176. Hovis, R. Corby, and Helge Kragh. "P. A. M. Dirac and the Beauty of Physics." *Scientific American,* 268(5): 104–109, May 1993.

177. Ingham, A. E. "On Two Conjectures in the Theory of Numbers." *American Journal of Mathematics,* 64: 313–319 (1942).

178. Ittelson, W. H., and F. P. Kilpatrick. "Experiments in Perception." *Scientific American,* 185(2), 50–55, August 1951.

179. Jackson, John. *Rational Amusement for Winter Evenings.* London: Longmans & Co., 1821.

180. Jayne, Caroline Furness. *String Figures and How to Make Them: A Study of Cat's-Cradle in Many Lands.* New York: Dover, 1975.

181. Jones, John Winter. *Riddles, Charades, and Conundrums.* London, 1821.

182. Jones, Samuel Isaac. *Mathematical Nuts for Lovers of Mathematics.* Nashville, 1932.

183. Kadesch, Robert R. *Math Menagerie.* New York: Harper & Row, 1970.

184. Kanigal, Robert. *The Man Who Knew Infinity: A Life of the Genius Ramanujan.* New York: Scribner, 1991.

185. Kapur, J. N. *The Fascinating World of Mathematics.* 3 vols. New Delhi, India: Mathematical Sciences Trust Society, 1989.

186. Kasner, Edward, and James R. Newman, *Mathematics and the Imagination.* New York: Simon and Schuster, 1940.

187. King, T. *The Best 100 Puzzles Solved and Answered.* Slough, England: W. Foulsham & Co., 1927.

188. Kirkman, T. P. "On a Problem in Combinatorics." *Cambridge and Dublin Math. J.,* 2: 191–204 (1847).

189. Klarner, David A., ed. *The Mathematical Gardner.* Boston: Prindle, Weber and Schmidt, 1981.

190. Klarner, David A., ed. *Mathematical Recreations: A Collection in Honor of Martin Gardner.* New York: Dover, 1998.

191. Knuth, Donald. *Surreal Numbers: How Two Ex-Students Turned On to Pure Mathematics and Found Total Happiness.* Reading, Mass.: Addison-Wesley, 1974.

192. Körner, T. W. *The Pleasures of Counting.* Cambridge: Cambridge University Press, 1997.

193. Kraitchik, Maurice. *Mathematical Recreations.* New York: Norton, 1942.

194. Krasnikov, S. V. "Causality Violations and Paradoxes." *Physical Review D,* 55(6): 3427–30 (1997).

195. Langley, E. M. "Problem 644." *Mathematical Gazette,* 11: 173, 1922.

196. Langman, H. *Play Mathematics.* New York: Hafner, 1962.

197. Lausmann, Raymond F. *Fun With Figures.* New York: McGraw-Hill, 1965.

198. Lavine, S. *Understanding the Infinite.* Cambridge, Mass.: Harvard University Press, 1994.

199. Lehman, R. S. "On Liouville's Function." *Mathematics of Computation,* 14: 311–320, 1960.

200. Licks, H. E. *Recreations in Mathematics.* New York: D. Van Nostrand, 1921.

201. Lindgren, Harry. *Geometric Dissections.* New York: Van Nostrand Reinhold, 1964.

202. Lindgren, Harry. *Recreational Problems in Geometrical Dissections and How to Solve Them.* New York: Dover Publications, 1972.

203. Lines, Malcolm. *Think of a Number.* New York: Adam Hilger, 1990.

204. Littlewood, J. E. *A Mathematician's Miscellany.* London: Methuen and Co., 1953.

205. Livio, Mario. *The Golden Ratio: The Story of Phi, the World's Most Astonishing Number.* New York: Broadway Books, 2002.

206. Locher J. L., ed. *M. C. Esher: His Life and Complete Graphic Work.* New York: Abradale Press, 1982.

207. Loyd, Sam. *Sam Loyd's Cyclopedia of 5000 Puzzles, Tricks, and Conundrums.* 1914. Reprint, New York: Pinnacle, 1976.

208. Loyd, Sam. *Sam Loyd's Picture Puzzles with Answers.* Brooklyn: S. Loyd, 1924. Reprint, Ontario, Canada: Algrove, 2000.

209. Lucas, Edouard. *Récréations Mathématiques,* 4 vols. Paris: Gauthier-Villars et fils, 1881–1894. Reprinted by Paris: A. Blanchard, 1960–1974.

210. MacMahon, Alexander: *New Mathematical Pastimes.* Cambridge: Cambridge University Press, 1921.

211. Madachy, Joseph S. *Mathematics on Vacation.* New York: Charles Scribner's, 1966. Republished as *Madachy's Mathematical Recreations.* New York: Dover, 1979.

212. Manning, H. P. *Geometry of Four Dimensions.* New York: Dover, 1956.

213. Manning, H. P. *The Fourth Dimension Simply Explained.* New York: Dover, 1960.

214. Maor, E. *To Infinity and Beyond: A Cultural History of the Infinite.* Boston, Mass.: Birkhäuser, 1987.

215. Maor, E. *e: The Story of a Number.* Princeton, N.J.: Princeton University Press, 1994.

216. Martin, G. "Absolutely abnormal numbers." *American Mathematical Monthly,* 108: 746–754, 2001.

217. Matijasevic, Yu. V. "Solution of the Tenth Problem of Hilbert." *Matematikai Lapok,* 21: 83–87 (1970).

218. Matijasevic, Yu. V. *Hilbert's Tenth Problem.* Cambridge, Mass.: MIT Press, 1993.

219. Matthews, W. H. *Mazes & Labyrinths: A General Account of Their History and Development.* London: Longmans, Green and Co., 1922. Reprint, New York: Dover, 1970.

220. Maxwell, E. A. *Fallacies in Mathematics.* New York: Cambridge University Press, 1959.

221. Meeus, J., and P. J. Torbijn. *Polycubes.* Paris: Cedic, 1977.

222. Melzak, Z. A. *Companion to Concrete Mathematics.* 2 vols. New York: John Wiley, 1973, 1976.

223. Mercer, J. W., et al. "Solutions to Langley's Adventitious Angles Problem." *Mathematical Gazette,* 11: 321–323, 1923.

224. Mosteller, F. *Fifty Challenging Problems in Probability with Solutions,* Reading, Mass.: Addison-Wesley, 1965, #47; "Mathematical Plums," edited by Ross Honsberger, pp. 104–110.

225. Mott-Smith, Geoffrey. *Mathematical Puzzles for Beginners and Enthusiasts,* 2nd rev. ed. New York: Dover, 1954.

226. Moyer, Ann E. *The Philosophers' Game: Rithmomachia in Medieval and Renaissance Europe.* Ann Arbor, Mich.: University of Michigan Press, 2001.

227. Murray, H. J. R. *History of Chess.* Oxford: Clarendon Press, 1913.

228. Murray, H. J. R. *A History of Board-Games Other than Chess.* Oxford: Clarendon Press, 1952.

229. Nasar, Sylvia. *A Beautiful Mind: A Biography of John Forbes Nash, Jr.* New York: Simon & Schuster, 1998.

230. Neumann, J. von, and Morgenstern, O. *Theory of Games and Economic Behavior.* New York: Wiley, 1964.

231. Neville, E. H. *The Fourth Dimension.* Cambridge: Cambridge University Press, 1921.

232. Newcomb, S. "Note on the Frequency of the Use of Digits in Natural Numbers." *American Journal of Mathematics,* 4: 39–40, 1881.

233. Newing, Angela. "The Life and Work of H. E. Dudeney." *Mathematical Spectrum,* 21: 37–44 (1988–1989).

234. Northrop, Eugene P. *Riddles in Mathematics. A Book of Paradoxes.* Princeton, N.J.: Van Nostrand, 1944.

235. O'Beirne, T. H. *Puzzles and Paradoxes: Fascinating Excursions in Recreational Mathematics.* Mineola, N.Y.: Dover, 1984.

236. Ogilvy, C. Stanley. *Tomorrow's Math: Unsolved Problems for the Amateur,* 2nd ed. New York: Oxford University Press, 1972.

237. Olivastro, Dominic. *Ancient Puzzles.* New York: Bantam, 1993.

238. Panofsky, Erwin. *The Life and Art of Albrecht Durer,* Princeton, N.J.: Princeton University Press, 1955.

239. Pardon, G. F. *Parlour Pastimes: A Repertoire of Acting Charades, Fire-Side Games, Enigmas, Riddles, Etc.* London, 1868.

240. Pennick, Nigel. *Mazes and Labyrinths.* London: Robert Hale, 1990.

241. Penrose, L. S., and R. Penrose. "Impossible Objects: A Special Type of Illusion," *British Journal of Psychology,* 49: 31, 1958.

242. Penrose Roger. *The Emperor's New Mind.* New York: Oxford University Press, 1989.

243. Péter, Rózsa. *Playing with Infinity: Mathematical Explorations and Excursions.* Mineola, N.Y.: Dover, 1976.

244. Peterson, Ivars. *The Mathematical Tourist: Snapshots of Modern Mathematics.* New York: W. H. Freeman, 1988.

245. Peterson, Ivars. *Islands of Truth: A Mathematical Mystery Cruise.* New York: W. H. Freeman, 1990.

246. Peterson, Ivars. *The Jungles of Randomness: A Mathematical Safari.* New York: Wiley, 1997.

247. Peterson, Ivars. *The Mathematical Tourist: New and Updated Snapshots of Modern Mathematics.* New York: W. H. Freeman, 1998.

248. Peterson, Ivars. *Mathematical Treks: From Surreal Numbers to Magic Circles.* Washington, D.C.: Mathematical Association of America, 2001.

249. Phillips, Hubert. *Journey to Nowhere: A Discursive Autobiography.* London: Macgibbon & Kee, 1960.

250. Phillips, Hubert. *My Best Puzzles in Mathematics.* Mineola, N.Y.: Dover, 1961.

251. Phillips, Hubert. *My Best Puzzles in Logic and Reasoning.* Mineola, N.Y.: Dover, 1961.

252. Picciotto, Henri. *Pentomino Activities.* Mountain View, Calif.: Creative Publications, 1986.

253. Pickover, Clifford. *Keys to Infinity.* New York: W. H. Freeman, 1995.

254. Pickover, Clifford. *Surfing Through Hyperspace.* New York: Oxford University Press, 1999.

255. Pickover, Clifford A. *Wonders of Numbers: Adventures in Mathematics, Mind, and Meaning.* Oxford: Oxford University Press, 2001.

256. Pickover, C. A. *The Zen of Magic Squares, Circles, and Stars: An Exhibition of Surprising Structures across Dimensions.* Princeton, N.J.: Princeton University Press, 2002.

257. Poole, D. G. "The Towers and Triangles of Professor Claus (or, Pascal Knows Hanoi)." *Mathematics Magazine,* 67: 323–344 (1994).

258. Poundstone, William. *The Recursive Universe.* New York: William Morrow, 1984.

259. Poundstone, William. *Prisoner's Dilemma: John Von Neumann, Game Theory and the Puzzle of the Bomb.* New York: Doubleday, 1992.

260. Quine, W. V. O. "On a So-Called Paradox." *Mind,* 62: 65–67 (1953).

261. Rabinowitz, Stanley. *Index to Mathematical Problems 1980–1984.* Westford, Mass.: Mathpro Press, 1992.

262. Rademacher, Hans, and Otto Toeplitz. *The Enjoyment of Mathematics: Selections from Mathematics for the Amateur.* Princeton, N.J.: Princeton University Press, 1957.

263. Raimi, R. A. "The Peculiar Distribution of First Digits." *Scientific American,* 221: 109–119 (Dec. 1969).

264. Ransom, William R. *One Hundred Mathematical Curiosities.* Portland, Maine: Weston Walch, 1955.

265. Read, Ronald C. *Tangrams: 330 Puzzles.* Mineola, N.Y.: Dover, 1965.

266. Ribenboim, P. *Catalan's Conjecture.* Boston, Mass.: Academic Press, 1994.

267. Robbin, Tony. *Fourfield: Computers, Art, and the Fourth Dimension.* Boston: Little, Brown, 1992.

268. Robertson, J., and Webb, W. *Cake-cutting Algorithms.* Natick, Mass.: A. K. Peters, 1998.

269. Rossi, Hugo. "Mathematics Is an Edifice, Not a Toolbox." *Notices of the AMS,* 43(10) (Oct. 1996).

270. Rubik, Erno, et al. *Rubik's Cubic Compendium.* New York: Oxford University Press, 1987.

271. Rucker, Rudolf v.B. *Geometry, Relativity, and the Fourth Dimension.* New York: Dover, 1977.

272. Rucker, R. (ed). *Speculations on the Fourth Dimension: Selected Writings of C. H. Hinton.* New York: Dover, 1980.

273. Rucker, Rudy. *The Fourth Dimension.* Boston: Houghton Mifflin, 1984.

274. Rucker, Rudy. *Mind Tools.* Boston: Houghton Mifflin, 1987.

275. Rucker, Rudy. *Infinity and the Mind.* Princeton, N.J.: Princeton University Press, 1995.

276. Russell, Bertrand. *Bertrand Russell Autobiography.* New York: Routledge, 1992.

277. Saaty, T. L., and P. C. Kainen. *The Four-Color Problem: Assaults and Conquest.* New York: Dover, 1986.

278. Sallows, L. C. F. "Alphamagic squares." *Abacus* 4 (No. 1): 28–45, 1986.

279. Sallows, L. C. F. "Alphamagic squares, part II". *Abacus* 4 (No. 2): 20–29, 43, 1987.

280. Sallows, L. C. F. "Alphamagic squares." In *The Lighter Side of Mathematics: Proceedings of the Eugène Strens Memorial Conference on Recreational Mathematics and Its History,* R. K. Guy and R. E. Woodrow, eds. Washington, D.C.: Mathematical Association of America, 1994.

281. Sanford, Vera. *A Short History of Mathematics.* Boston, Mass.: Houghton Mifflin, 1930.

282. Scarne, John. *Scarne's New Complete Guide to Gambling.* New York: Simon and Schuster, 1986.

283. Schaaf, William Leonard. *A Bibliography of Recreational Mathematics.* 4 vols. Washington, D.C.: National Council of Teachers of Mathematics, 1955–78.

284. Schattschneider, Doris. *M. C. Escher: Visions of Symmetry.* New York: W. H. Freeman, 1990.

285. Scheick, William J. "The Fourth Dimension in Wells's Novels of the 1920s." *Criticism,* 20: 167–190, 1978.

286. Schrauf, M., B. Lingelbach, E. Lingelbach, and E. R. Wist. "The Hermann Grid and the Scintillation Effect." *Perception,* 24, suppl. A: 88–89, 1995.

287. Schreiber, P. "A New Hypothesis on Durer's Enigmatic Polyhedron in His Copper Engraving 'Melencholia I'," *Historia Mathematica,* 26: 369–377, 1999.

288. Schubert, Hermann. *Mathematical Essays and Recreations.* Thomas J. McCormack, trans. Chicago: Open Court, 1899.

289. Schubert, Hermann. *Mathematische Mussestunden.* Walter de Gruyter, 1940. First published 1900? (German).

290. Schubert, Hermann. *Mathematische Mussestunden II.* Berlin: G. J. Goshen'sche, 1909.

291. Schuh, Fred. *The Master Book of Mathematical Recreations.* Mineola, N.Y.: Dover, 1968.

292. Searle, John R. "Minds, Brains and Programs" in *The Brain and Behavioral Sciences,* vol. 3. Cambridge: Cambridge University Press, 1980.

293. Senechal, M. *Quasicrystals and Geometry.* New York: Cambridge University Press, 1995.

294. Shepherd, Walter. *For Amazement Only.* New York: Penguin, 1942. Reprinted as *Mazes and Labyrinths: A Book of Puzzles.* New York: Dover, 1961.

295. Simon, William. *Mathematical Magic.* New York: Charles Scribner's, 1964.

296. Singh, Simon. *Fermat's Enigma: The Epic Quest to Solve the World's Greatest Mathematical Problem.* New York: Walker & Co., 1997.

297. Singmaster, David. *Notes on Rubik's 'Magic Cube.* Hillside, N.J.: Enslow, 1981.

298. Singmaster, David. *Sources in Recreational Mathematics: An Annotated Bibliography,* 7th prelim, ed. Unpublished.

299. Skewes, S. "On the difference $\pi(x) - li(x)$." *Journal of the London Mathematical Society,* 8: 277–283 (1933).

300. Slocum, Jerry. *Compendium of Mechanical Puzzles.* 3rd ed. Slocum and Haubrich, 1977.

301. Slocum, Jerry. *The Tangram Book: The Story of the Chinese Puzzle with Over 2000 Puzzles to Solve.* New York: Sterling, 2003.

302. Smith, David Eugene. *Number Stories of Long Ago.* Washington, D.C.: National Council of Teachers of Mathematics, 1919.

303. Smith, David Eugene. *A Source Book in Mathematics.* 2 vols. New York: McGraw-Hill, 1929.

304. Smullyan, Raymond M. *What Is the Name of This Book? The Riddle of Dracula and Other Logical Puzzles.* Englewood Cliffs, N.J.: Prentice Hall, 1978.

305. Smullyan, Raymond M. *The Chess Mysteries of Sherlock Holmes.* New York: Alfred A. Knopf, 1979.

306. Smullyan, Raymond. *This Book Needs No Title: A Budget of Living Paradoxes.* Englewood Cliffs, N.J.: Prentice-Hall, 1980.

307. Smullyan, Raymond M. *The Lady or the Tiger? And Other Logic Puzzles.* New York: Alfred A. Knopf, 1982.

308. Smullyan, Raymond M. *Alice in Puzzleland.* New York: William Morrow, 1982.

309. Smullyan, Raymond M. *To Mock a Mockingbird.* New York: Alfred A. Knopf, 1985.

310. Smullyan, Raymond M. *Forever Undecided: A Puzzle Guide to Godel.* New York: Alfred A. Knopf, 1987.

311. Smullyan, Raymond. *The Tao is Silent.* New York: HarperCollins, 1992.

312. Soddy, F. "The Kiss Precise." *Nature,* 137: 1021, 1936.

313. Sprague, Roland. *Recreation in Mathematics.* Mineola, N.Y.: Dover, 1963.

314. Stein, M. L., S. M. Ulam, and M. B. Wells. "A Visual Display of Some Properties of the Distribution of Primes." *American Mathematical Monthly,* 71: 516–520, 1964.

315. Steinhaus, H. "Sur la division pragmatique." *Ekonometrika* (*Supp.*), 17: 315–319, 1949.

316. Steinhaus, Hugo. *Mathematical Snapshots.* 3rd rev. ed. New York: Dover, 1983.

317. Stewart, Ian. *Does God Play Dice: The New Mathematics of Chaos.* Cambridge, Mass.: Basil Blackwell, 1989.

318. Stewart, Ian. *Game, Set, and Math: Enigmas and Conundrums.* Cambridge, Mass.: Basil Blackwell, 1989.

319. Stewart, Ian. "Tales of a Neglected Number." *Scientific American,* 274: 102–103 (Jun. 1996).

320. Stewart, Ian. "Unilluminable Rooms." *Scientific American,* 275(2), 100–103 (Aug. 1996).

321. Stewart, Ian and Cohen, Jack Cohen. *Figments of Reality.* Cambridge: Cambridge University Press, 1997.

322. Stewart, Ian. "Cone with a Twist." *Scientific American,* 281: 116–117 (Oct. 1999).

323. Stewart, Ian. *Flatterland: Like Flatland Only More So.* New York: Perseus, 2001.

324. Stromberg, K. "The Banach-Tarski Paradox." *American Mathematical Monthly,* 86: 3, 1979.

325. Stubbs, A. Duncan. *Miscellaneous Puzzles.* London: Frederick Warne, 1931.

326. Struik, Dirk Jan, ed. *A Source Book in Mathematics 1200–1800.* Princeton, N.J.: Princeton University Press, 1969.

327. Strutt, Joseph. *The Sports and Pastimes of the People of England.* London: J. White, 1801.

328. Stuwe, M. *Plateau's Problem and the Calculus of Variations.* Princeton, N.J.: Princeton University Press, 1989.

329. Swade, Doron. *The Difference Engine: Charles Babbage and the Quest to Build the First Computer.* New York: Viking, 2000.

330. te Riele, H. J. J. "On the Sign of the Difference $\pi(x) - \mathrm{Li}(x)$." *Mathematics of Computation,* 48: 323–328 (1987).

331. Thom, R. *Structural Stability and Morphogenesis: An Outline of a General Theory of Models.* Reading, Mass.: Addison-Wesley, 1972.

332. Thompson, d'Arcy W. *On Growth and Form.* Cambridge: Cambridge University Press, 1917. (Reprint, New York: Dover, 1992.)

333. Thorpe, Edward O. *Beat the Dealer,* rev. ed. New York: Random House, 1966.

334. Tokarsky, George W. "Polygonal Rooms Not Illuminable from Every Point." *American Mathematical Monthly,* 102(10): 867–879, 1995.

335. Toole, Betty A. (ed.). *Ada, the Enchantress of Numbers: A Selection from the Letters of Lord Byron's Daughter and Her Description of the First Computer.* Mill Valley, Calif.: Strawberry Press, 1998.

336. Tormey, Alan and Tormey, Judith Farr. "Renaissance Intarsia: The Art of Geometry." *Scientific American,* 247: 136–143, July 1982.

337. Turing, A. M. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings*

of the London Mathematical Society, Series 2, 42: 230–265, 1937.

338. Turney, P, “Unfolding the Tesseract.” *Journal of Recreational Mathematics,* 17(1): 1–16, 1984–1985.

339. Underwood, Dudley. “The First Recreational Mathematics Book.” *Journal of Recreational Mathematics,* 3, 164–169, 1970.

340. Van Cleve, James. “Right, Left, and the Fourth Dimension.” *The Philosophical Review,* 96: 33–68 (1987).

341. Vardi, I. “Archimedes’ Cattle Problem.” *American Mathematical Monthly,* 105: 305–319 (1998).

342. Vardi, Ilan. *Computational Recreations in Mathematica.* Redwood City, Calif.: Addison-Wesley, 1991.

343. Wagon, S. *The Banach-Tarski Paradox.* New York: Cambridge University Press, 1993.

344. Waller, Mary D. *Chladni Figures: A Study in Symmetry.* London: G. Bell & Sons, 1961.

345. Watson, G. N. “The Problem of the Square Pyramid.” *Messenger of Mathematics,* 48: 1–22 (1918–1919).

346. Weisstein, Eric W. *The CRC Concise Encyclopedia of Mathematics.* Boca Raton, Fla.: CRC Press, 1998.

347. Wells, David Graham. *Hidden Connections, Double Meanings: A Mathematical Exploration.* Cambridge: Cambridge University Press, 1988.

348. Wells, David. *The Penguin Dictionary of Curious and Interesting Geometry.* London: Penguin Books, 1991.

349. Wells, David. *The Penguin Book of Curious and Interesting Mathematics.* New York: Penguin, 1997.

350. Wells, Kenneth. *Wooden Puzzles and Games.* New York: Sterling Publications, 1983.

351. Weyl, Hermann. *Symmetry.* Princeton, N.J.: Princeton University Press, 1952.

352. White, Alain C. *Sam Loyd and His Chess Problems.* Leeds, England: Whitehead & Miller, 1913.

353. Wiles, A. “Modular Elliptic-Curves and Fermat’s Last Theorem.” *Annals of Mathematics,* 141: 443–551, 1995.

354. Williams, W. T., and Savage, G. H. *The First Penguin Problems Book.* New York: Penguin Books, 1940.

355. Wyatt, Edwin Mather. *Puzzles in Wood.* Milwaukee, Wis.: Bruce Publishing Company, 1928.

356. Yan, Li, and Du, Shiran. *Chinese Mathematics: A Concise History.* Oxford: Oxford University Press, 1988.

357. Zaslavsky, Claudia. *Africa Counts.* Brooklyn, N.Y.: Lawrence Hill Books, 1973.

Solutions to Puzzles

abracadabra

To solve this problem, redraw the diamond and replace every letter on the outside top half of the diamond with a 1 and every other point in the diamond with a dot. This gives

```
          1
        1   1
      1   .   1
    1   .   .   1
  1   .   .   .   1
1   .   .   .   .   1
  .   .   .   .   .
    .   .   .   .
      .   .   .
        .   .
          .
```

Next, replace each of the dots, starting at the top, with the sum of the two numbers in the northwest and northeast positions, so that the pyramid now starts

```
      1
    1   1
  1   2   1
1   3   3   1
```

After you have worked your way down to the very bottom, the bottom number in the diamond is the answer to Polya's problem: 252.

age puzzles and tricks

1. 18.
2. $16\frac{1}{2}$ (Mary is $27\frac{1}{2}$).

alphametic

1.
67432 (EARTH)
704 (AIR)
8046 (FIRE)
+ 97364 (WATER)
———————
173546 (NATURE)

A = 7, E = 6, F = 8, H = 2, I = 0, R = 4, T = 3, W = 9

2.
127503 (SATURN)
502351 (URANUS)
3947539 (NEPTUNE)
+ 46578 (PLUTO)
———————
4623971 (PLANETS)

A = 2, E = 9, L = 6, N = 3, O = 8, P = 4, R = 0, S = 1, T = 7, U = 5

3.
862903 (MARTIN)
+ 1627342 (GARDNER)
———————
2490245 (RETIRES)

A = 6, D = 7, E = 4, G = 1, I = 0, M = 8, N = 3, R = 2, S = 5, T = 9

anagram

1. Wolfgang Amadeus Mozart.
2. Thomas Alva Edison.
3. William Shakespeare.

Carroll, Lewis

1. There is as much water in the milk/water mixture as milk in the water/milk mixture.
2. There are 30 ways of painting the cube. If the restriction that each face be painted a different color is dropped, there are 2,226 ways of painting the cube.
3. FOUR → FOUL → FOOL → FOOT → FORT → FORE → FIRE → FIVE

Alternatively:
FOUR → POUR → POUT → ROUT → ROUE → ROVE → DOVE → DIVE → FIVE
(Thanks to my editor, Stephen Power, for this one.)

4. Carroll didn't have an answer in mind when he wrote the riddle, though he later came up with: "Because it can produce a few notes, though they are *very* flat; and it is nevar put wrong end in front!" (Note the variant spelling of "never.") Other authors have come up with: "Because Poe wrote on both"

(Sam Loyd); "Because it slopes with a flap" (Cyril Pearson); and "Because both have quills dipped in ink" (David Jodrey).

clock puzzles

1. Approximately 27 minutes and 42 seconds (exactly 360/13 minutes) after 6.
2. The stopped clock, because it will give the right time twice a day, whereas the other is only correct approximately every two years.
3. The positions of the hands shown in the illustration could only indicate that the clock stopped at 44 min. 51 1,143/1,427 sec. after eleven o'clock. The second hand would next be "exactly midway between the other two hands" at 45 min. 52 496/1,427 sec. after eleven o'clock. If we had been dealing with the points on the circle to which the three hands are directed, the answer would be 45 min. 22 106/1,427 sec. after eleven; but the question applied to the hands, and the second hand would not be between the others at that time, but outside them.

cone

The simple rule is that the cone must be cut at one-third of its height.

de Morgan, Augustus

43 (the only number that, when squared, gives a number between the years of de Morgan's birth and death).

domino

No! Suppose it *were* possible to totally cover the modified chessboard with nonoverlapping dominos. In any complete tiling, every domino must cover exactly one white square and one black square. Thus the modified board must have exactly the same number of black and white squares. But the two removed squares, from diagonally opposite corners of a chessboard, must be same color. Since there can't be the same number of white squares and black squares on the modified board it must be impossible to tile the modified board with nonoverlapping dominos.

hundred fowls problem

Using the two equations to eliminate C, gives the equation $7R + 4H = 100$. R must be less than 15 (since $7 \times 15 = 105$). Trial-and-error shows that values of R that allow whole number values of H are 4, 8, and 12, from which it follows that the three possible solutions to the problem are 4 roosters, 18 hens, and 78 chicks; 8 roosters, 11 hens, and 81 chicks; and 12 roosters, 4 hens, and 84 chicks.

kinship puzzles

1. Your son.
2. Sons of two men who married each other's mothers.
3. The party consisted of two girls and a boy, their father and mother, and their father's father and mother.

measuring and weighing puzzles

The innkeeper first filled the 5-pint and 3-pint measures, then turned the tap on the barrel and allowed the rest of its contents to run to waste. He closed the tap and emptied the 3-pint into the barrel; filled the 3-pint from the 5-pint; emptied the 3-pint into the barrel; transferred the 2 pints from the 5-pint to the 3-pint; filled the 5-pint from the barrel, leaving 1 pint now in the barrel; filled the 3-pint from the 5-pint; allowed the company to drink the contents of the 3-pint; filled the 3-pint from the 5-pint, leaving 1 pint now in the 5-pint; drank the contents of the 3-pint; and finally drew off 1 pint from the barrel into the 3-pint. He now had 1 pint of ale in each measure!

missing dollar problem

There is no missing dollar (of course!). Adding $27 and $2 (to get $29) is a bogus operation. They paid $27, $2 went to the dishonest waiter, and $25 went to the restaurant. You have to subtract $27 minus $2 to get $25. There never was a $29; it has nothing to do with anything.

Mrs. Perkins's quilt

The following diagram shows how the quilt should be made.

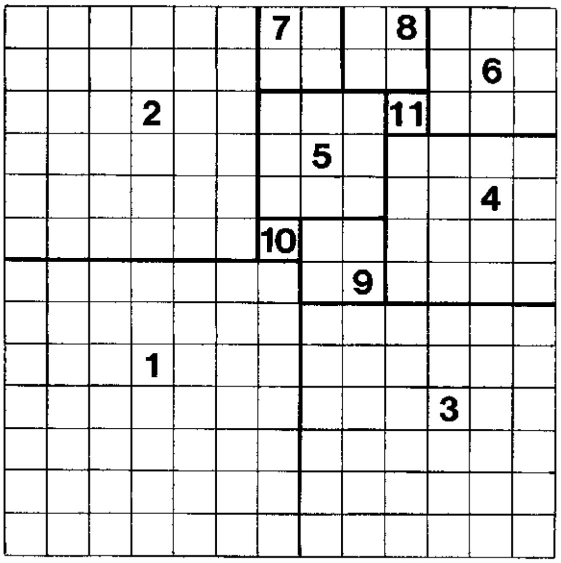

The solution of Mrs. Perkins's quilt.

In *Amusements in Mathematics,* Dudeney writes: "There is, I believe, only one solution to this puzzle. The fewest separate squares must be eleven. The portions must be of the sizes given, the three largest pieces must be arranged as shown, and the remaining group of eight squares may be 'reflected,' but cannot be differently arranged."

nine rooms paradox

One of the two customers who is initially placed in room A, whom we refer to as the "first" customer, is later transferred to room I and treated as if he were also the tenth customer!

postage stamp problems

Dudeney gives the solution to his "The Four Postage Stamps" problem as follows: "[T]he four stamps may be given in the shape 1, 2, 3, 4, in three ways; in the shape 1, 2, 5, 6, in six ways; in the shape 1, 2, 3, 5, or 1, 2, 3, 7, or 1, 5, 6, 7, or 3, 5, 6, 7, in twenty-eight ways; in shape 1, 2, 3, 6, or 2, 5, 6, 7, in fourteen ways; in shape 1, 2, 6, 7, or 2, 3, 5, 6, or 1, 5, 6, 10, or 2, 5, 6, 9, in fourteen ways. Thus there are sixty-five ways in all."

probability theory

1. There are four possibilities:

Oldest Child	Youngest Child
1. Girl	Girl
2. Girl	Boy
3. Boy	Girl
4. Boy	Boy

If your friend says "My oldest child is a girl," he has eliminated cases 3 and 4, and in the remaining cases both are girls ½ of the time. If your friend says "At least one of my children is a girl," he has eliminated case 4 only, and in the remaining cases both are girls ⅓ of the time.

2. There are six possible bullet configurations (B = bullet, E = empty):

B B B E E E → player 1 dies
E B B B E E → player 2 dies
E E B B B E → player 1 dies
E E E B B B → player 2 dies
B E E E B B → player 1 dies
B B E E E B → player 1 dies

One therefore has a ⅔ probability of winning (and a ⅓ probability of dying) by shooting second.

Pythagorean square puzzle

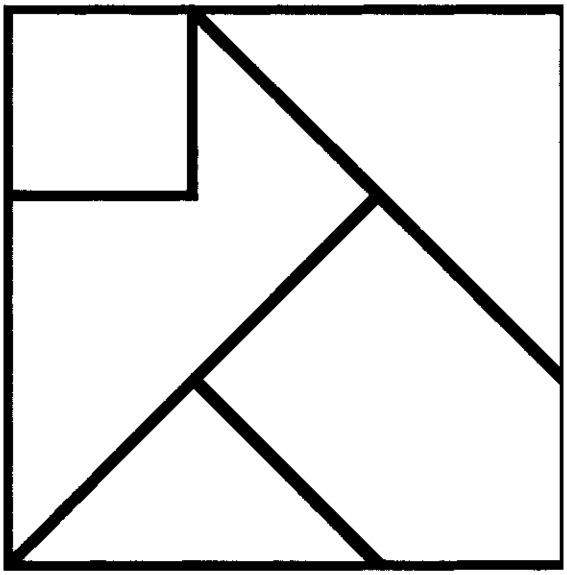

Pythagorean puzzle solution.

river-crossing problem

The missionaries (M) can avoid being eaten by the cannibals (C) if the crossings of the river and the returns are arranged as follows. Crossing #1: 1M + 1C; return #1: 1M; crossing #2: 2C; return #2: 1C; crossing #3: 2M; return #3: 1M + 1C; crossing #4: 2M; return #4: 1C; crossing #5: 2C; return #5: 1C; crossing #6: 2C.

word puzzles

(1) $T_1W_4E_1L_1V_4E_1$.

(2) Dreamt.

(3) Tremendous, horrendous, stupendous, and hazardous.

(4) I am.

(5) skiing

(6) "Are" by an adding an "a" becomes "area."

(7) underground.

$9.95 in USA

Essential Dictionary of Music Notation

A Practical and Concise Source for Students to Professionals

Take the mystery out of music notation. Whether you're using a pencil to sketch a melody or a computer to score a symphony, this multi-purpose quick reference is for you.

Every aspect of notation is covered in a versatile, easy-to-use format. From beginning music students to seasoned engraving professionals, everyone notating music will love this dictionary!

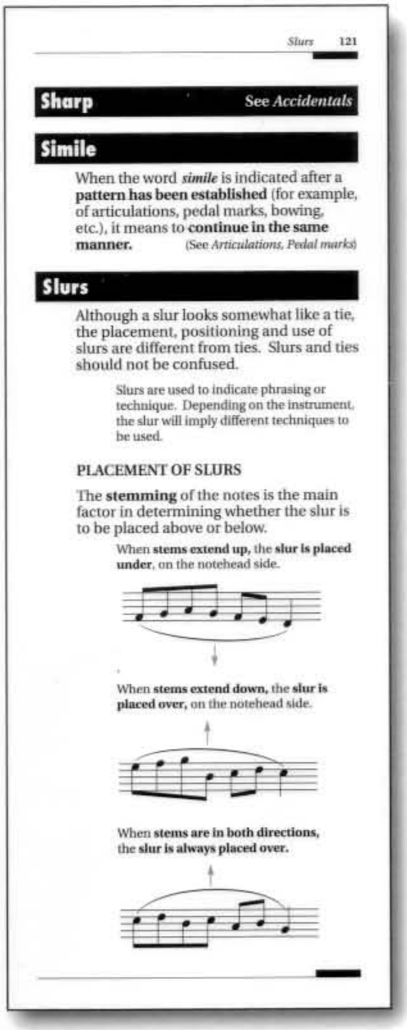
Slurs 121

Sharp See *Accidentals*

Simile

When the word ***simile*** is indicated after a **pattern has been established** (for example, of articulations, pedal marks, bowing, etc.), it means to **continue in the same manner.** (See *Articulations, Pedal marks*)

Slurs

Although a slur looks somewhat like a tie, the placement, positioning and use of slurs are different from ties. Slurs and ties should not be confused.

Slurs are used to indicate phrasing or technique. Depending on the instrument, the slur will imply different techniques to be used.

PLACEMENT OF SLURS

The **stemming** of the notes is the main factor in determining whether the slur is to be placed above or below.

When **stems extend up,** the **slur is placed under**, on the notehead side.

When **stems extend down,** the **slur is placed over,** on the notehead side.

When **stems are in both directions,** the **slur is always placed over.**

- Easy–to–use dictionary format
- Concise text is clearly illustrated
- Key words and concepts are highlighted
- Over 75 topics spanning 158 pages

Alfred Publishing Co
16380 Roscoe Blvd.
P.O. Box 10003
Van Nuys, CA 9141

17241

AN ALFRED HANDY GUIDE

Essential Dictionary of Music Notation

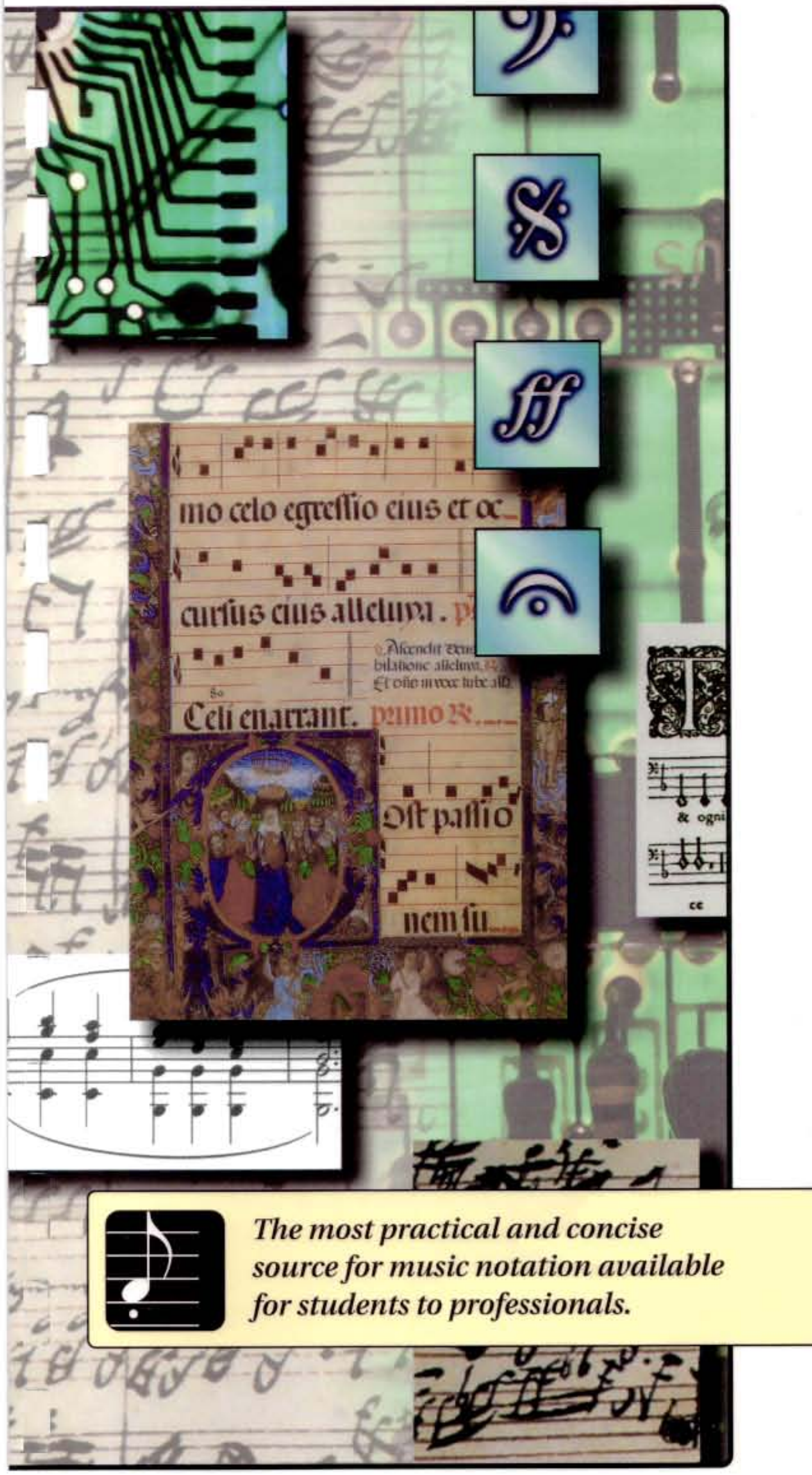

The most practical and concise source for music notation available for students to professionals.

TOM GEROU • LINDA LUSK

Essential Dictionary of Music Notation

The most practical and concise source for music notation

TOM GEROU
LINDA LUSK

- Easy-to-use format, including cross-referencing
- For any musician, whether using pencil or computer
- Emphasis on correct and current notation practices
- Numerous examples, kept simple and to-the-point

Thanks to Susy Christiansen, Mitchell Newman, Greg Plumblee and Patrick Wilson

Cover design: Carol Kascsak

Cover art, center photo: From an Antiphoner,
or book of music for the choir services.
Copyright © Sotheby's

Copyright © MCMXCVI by Alfred Publishing Co., Inc.
All rights reserved. Printed in USA.

ISBN 0-88284-768-6

APPROACHING MUSIC NOTATION

Music notation is a language and, as such, it must be communicable. The closer the notation adheres to common rules and guidelines, the more successfully the music will be performed. The rules and guidelines presented here will make the notation easier to read and will help to make the musical intent clear. Although these rules are not inflexible, they should not be ignored.

For music notation, the first priority is always clarity. Choices need to be made that are solely dependent on the situation; a rule may not provide the most desirable answer to a problem. (It would be an endless task to list every musical exception to a rule.) Problems are usually solved by deciding which rule is more flexible (least important to clarity).

Although meticulous adjustment can only clarify and beautify the notation, sometimes such attention to detail is not practical. All decisions must depend on a wide variety of situations. An understanding of the rules allows these choices to be made wisely—the notation can only be as good as the knowledge at hand.

ABOUT THIS BOOK

The *Essential Dictionary of Music Notation* is in an easy-to-use format, which includes cross referencing. At the top of each page, running heads show exactly which topic is being treated on that page. Bold treatments of text highlight important points for clarity and quick reference.

Text is accompanied by numerous examples directly related to the topic being discussed. These examples are kept as simple and complete as possible, with an emphasis on presenting correct and current notation practices. Obsolete rules are mentioned only when necessary.

The focus of this book has been narrowed, avoiding the discussion of orchestration or music theory, in order to present as much information as possible that is directly related to the actual *notation* of music. The serious musician is encouraged to seek out other sources that more thoroughly address the subjects of orchestration and music theory.

Whether you are using a pencil or a computer, all principles within this book are applicable. No computer software or specific tool is referenced, in order to make this a useful guide for any musician.

An *Index of Topics* is provided on the last page.

A cappella

Vocal music that is to be sung **without accompaniment** (a cappella) often has a piano part for rehearsal only.

The piano part may be full size or cue size (see *Cue notes*). If cue size, the staff should be cue size also. (Cue-size notes on a full-size staff are not as easy to read.)

The indication ***for rehearsal only*** should be clearly placed above the piano part.

A due (a2)

See *Sharing a staff*

A tempo

Place above the staff (sometimes within the grand staff) to indicate a return to the normal tempo. Used after tempo alterations such as ***ritardando, accelerando, più lento*** or ***ad libitum.*** (See *Tempo marks*)

Accents

See *Articulations*

Accidentals

Temporary alterations to the pitch of a note.

THE 5 SYMBOLS

If appearing in the music, apart from use in the key signature, these symbols are called *accidentals.*

♯ **Sharp**
Raises a note one half step.

𝄪 **Double-sharp**
Raises a note two half steps.

♭ **Flat**
Lowers a note one half step.

𝄫 **Double-flat**
Lowers a note two half steps. Notice that the flats touch, forming one symbol.

♮ **Natural**
Cancels any of the above, whether indicated in the key signature or as an accidental.

ONE ACCIDENTAL PER NOTE

Do not use multiple accidentals on one note. (A double-flat is a single accidental.

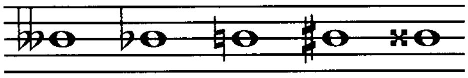

Accidentals shown altering one note from lowest to highest.

The lowest a note can be lowered is a 𝄫.

The highest a note can be raised is a 𝄪.

WITHIN A MEASURE

Any accidental is **cancelled when a new accidental appears** on the like pitch.

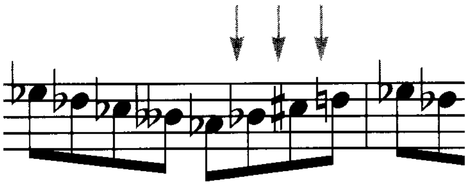

Notice that only the flat (not ♮♭) is needed to change from a double-flat to a flat.

An accidental affects all **subsequent notes of the same pitch** within the measure.

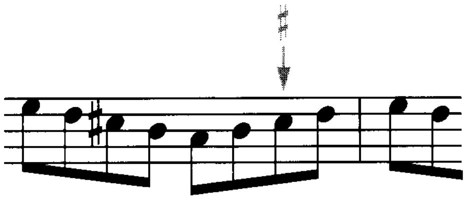

An accidental **must be added again** in the following measure.

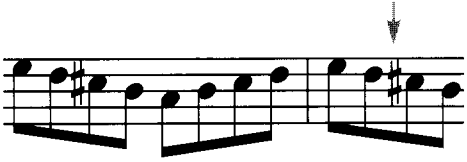

An accidental **carries over to a tied note** extending from one measure to the next.

The accidental will reappear only when the note is re-struck.

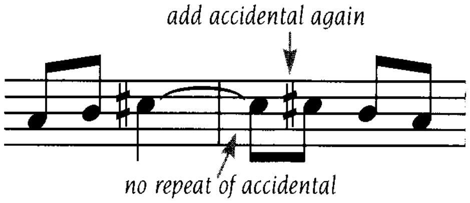

OCTAVE ALTERATIONS

Accidentals **do not apply if an *8va*** sign is added; repeat the accidentals.

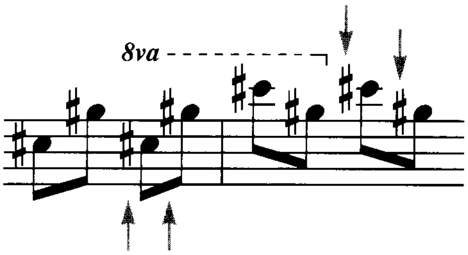

COURTESY ACCIDENTALS

Used as **reminders of the key signature** when an accidental has been used in the previous measure, or to **avoid ambiguity** in certain situations.

Use a courtesy accidental as a **reminder** to return to the original pitch.

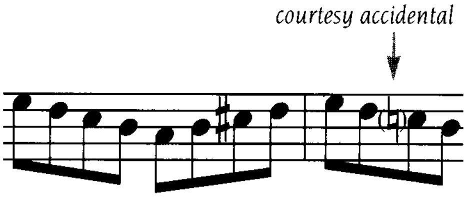

Courtesy accidentals may or may not be enclosed in **parentheses.** The parentheses show that it *is* a courtesy accidental, and help avoid confusion about the key signature. However, if many courtesy accidentals are used, and the piece is also crowded, the parentheses could make it look cluttered and more complicated. For this reason some editors choose not to use parentheses at all.

If a chord has more than one courtesy accidental, **each accidental gets its own pair of parentheses.**

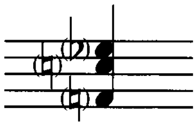

It is best to **add a courtesy accidental for octave shifts** although the notes of the same pitch in different octaves are considered separately.

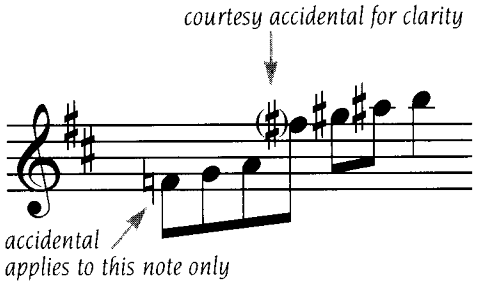
courtesy accidental for clarity
accidental applies to this note only

GRACE NOTES WITH ACCIDENTALS

The accidental placed on a grace note will be smaller (cue size), like the grace note.

The accidental **affects all subsequent notes of the same pitch** within the measure.

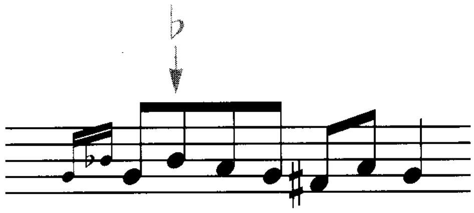

Depending on the clarity of the situation, **courtesy accidentals are often needed** to avoid confusion.

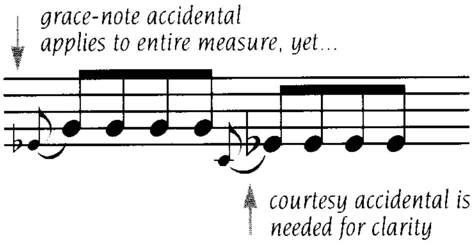
grace-note accidental applies to entire measure, yet...
courtesy accidental is needed for clarity

ALIGNMENT FOR INTERVALS (2 Notes)

For intervals of a **2nd through a 6th,** place the upper accidental closest to the note, the lower accidental to the left.

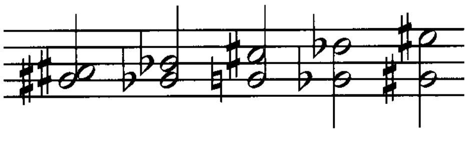

When intervals are **greater than a 6th,** accidentals align vertically.

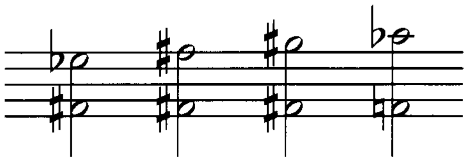

For the **interval of a 6th,** if the two accidentals don't collide they may be aligned vertically.

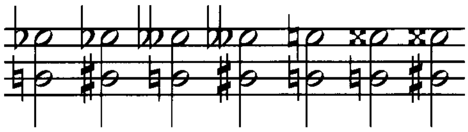

ALIGNMENT FOR CHORDS (3 Notes)

When outer notes are a **6th or less,** the upper accidental is closest to the note, the lower accidental is placed left, and the middle accidental is placed farthest left.

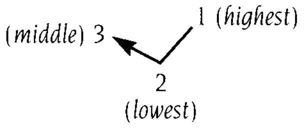
(middle) 3
1 (highest)
2
(lowest)

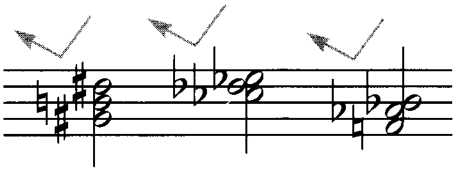

When outer notes are **greater than a 6th,** upper and lower accidentals are aligned closest to the notes—the middle accidental is placed to the left.

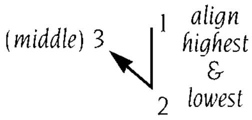
(middle) 3
1
2
align highest & lowest

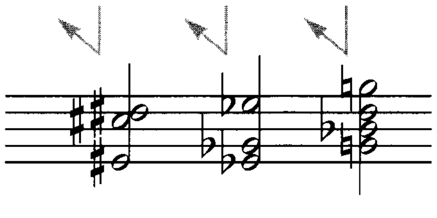

ALIGNMENT FOR CHORDS (More than 3 Notes)

When dealing with accidentals for complex chords, rules are treated more as guidelines or suggestions. Keep the arrangement of the accidentals as compact as possible.

Make the arrangement of accidentals as **easy to read as possible** for the situation.

Align the highest and lowest accidentals whenever possible.

The **center accidentals are usually arranged diagonally** from highest to lowest.

Align accidentals for octaves whenever possible.

Accidentals for the 2nd should usually be **shaped like the 2nd** (if the outer notes are greater than a 6th).

The upper accidental is placed closest to the note, the lower accidental to the left.

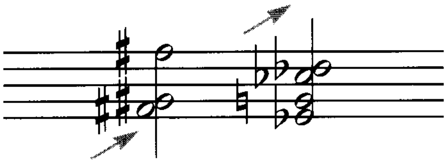

Accidentals are always placed before the entire note structure. (Do not place an accidental between notes that are played together, even if they are stemmed in opposite direction.)

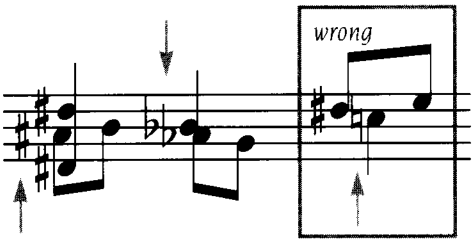

TWO PARTS ON A STAFF

For multiple notes played by **one instrument,** an accidental carries through the measure. Cancellations must be applied.

one instrument—
the accidentals carry through the measure

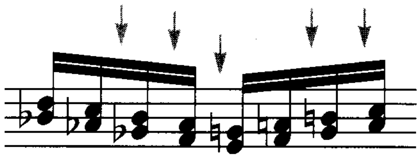

Parts on the same staff that are played by **more than one instrument** (including vocal parts) are treated individually. Repeat accidentals in the same measure if the voices cross.

two instruments sharing a staff—
the accidentals are considered separately

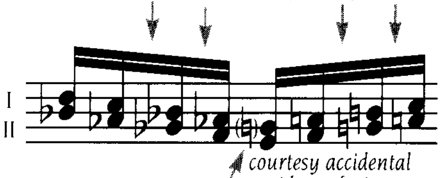

courtesy accidental
avoids confusion

CHROMATIC PASSAGES

In general, use **sharps when ascending, flats when descending.**

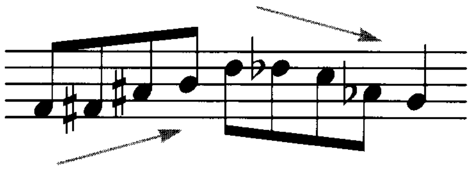

When choosing accidentals, consider the context of the music, such as key signature, harmony and voice leading.

In the chord below, both a sharp and a flat are required (C♯, not D♭; E♭ not D♯).

augmented 6th chord

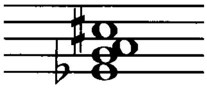

ACCIDENTALS WITH TRILLS (See *Trill*)

ACCIDENTALS IN HARP MUSIC

(See *Pedal marks*)

Alternating instruments

CONCERT ORCHESTRA / BAND

1. To alternate instruments, enough **time must be given** to allow the change.
2. During the measures of rest the indication ***"Muta in [second instrument]"*** is indicated **below** the staff as soon as the first instrument stops.
3. Any change to the **key signature** occurs after the barline following the rest, before the new instrument begins.
4. The **new instrument is indicated** above the entrance of the instrument.

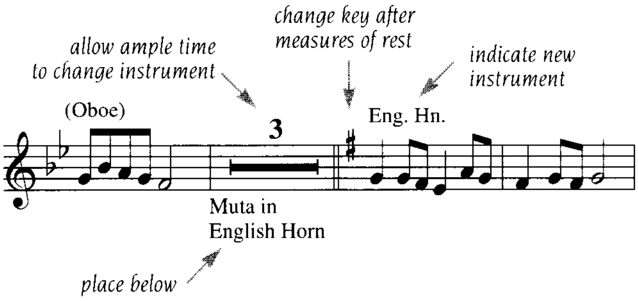
change key after measures of rest
allow ample time to change instrument
indicate new instrument
(Oboe)
3
Eng. Hn.
Muta in English Horn
place below

INCORPORATING CUES

1. Cue notes should be **added after many measures of rest,** to help the performer make a correct entrance.
2. The cued instrument is **indicated above the passage** in a smaller type size.
3. The **key signature changes at the beginning** of the cue.
4. **Full-size rests** are indicated below (or above, depending on stem direction).

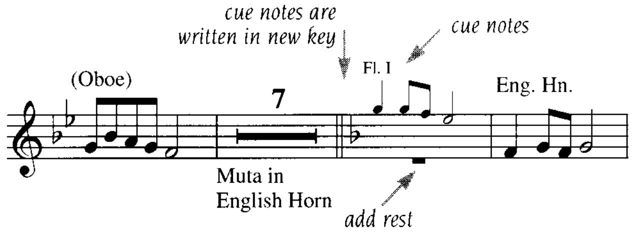
cue notes are written in new key
cue notes
Fl. I
(Oboe)
7
Eng. Hn.
Muta in English Horn
add rest

(See *Cue notes*)

JAZZ ENSEMBLE

All rules above apply except that "muta in [...]" is written **"Change to [new instrument]"** and is placed **above** the staff.

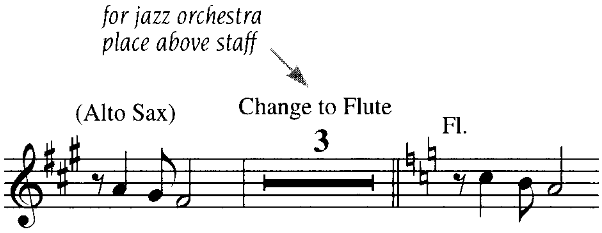
for jazz orchestra place above staff
(Alto Sax)
Change to Flute
3
Fl.

COMMONLY ALTERNATING INSTRUMENTS

The following is a listing of commonly alternating instruments in a standard orchestra:

	Alternate
Flute III	Piccolo
Oboe III	English horn
Clarinet III	Bass clarinet
Bassoon III	Contrabassoon

In smaller orchestras, the second player alternates in all cases above.

Arpeggio signs

A vertical wavy line before a chord or interval indicates an arpeggio (rolled chord).

ARPEGGIO DIRECTION

The chord is to be played quickly from **bottom to top** (no direction arrow).

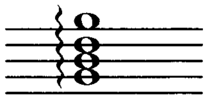

played upward (no direction arrow)

If the arpeggio is to be played from **top to bottom, an arrowhead is added** to the bottom of the wavy line.

played downward

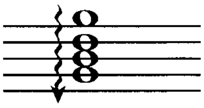

When the arpeggio returns to normal (bottom to top) or alternates with an arpeggio played downward, an arrowhead is added to the top of the wavy line.

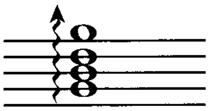

played upward

The arpeggio sign **should not extend much past the chord** affected.

PLACEMENT WITH ACCIDENTALS

Arpeggio signs are always **placed before accidentals**.

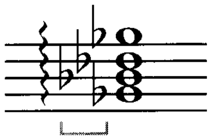

allow space
for accidentals

FOR KEYBOARD INSTRUMENTS

The arpeggio sign **can extend across both staves.** The right hand begins after the left hand has completed the chord.

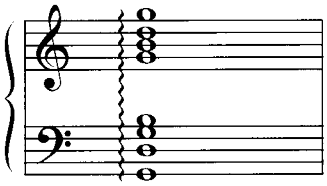

played upward,
from bottom to top

If the wavy line is broken between staves, both hands begin the arpeggio simultaneously.

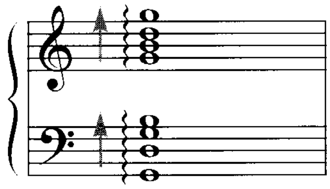

hands start
simultaneously

Articulations

Articulations are symbols used to indicate *how* a note or chord is played, alone or in relation to other notes and chords.

There are five main articulations: ***staccatissimo, staccato, tenuto, accent, marcato.***

An articulation's **influence on a note or chord varies** depending on its context. A staccato note in a slower tempo is not played as short as it would be played in a faster tempo. An accent is softer in the dynamic ***p*** than in the dynamic ***f***.

One articulation applies to all notes of an interval or chord (on one stem).

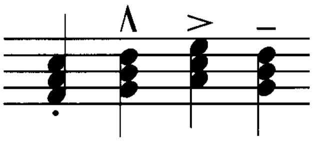

DURATIONAL ARTICULATIONS

There are three articulations that **affect the duration** of a note or chord: *staccatissimo* (▾), *staccato* (·) and *tenuto* (–).

The following illustrates durational articulations compared to an unmarked note, from shortest to longest.

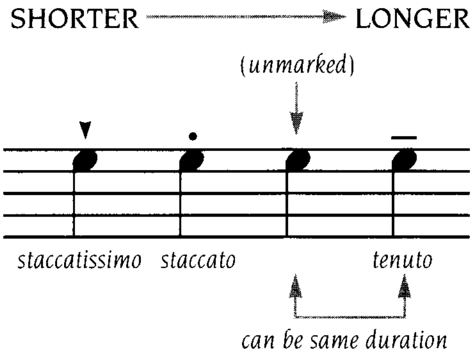

Staccatissimo

The shape varies slightly among engravers, but is always wedge-shaped. It indicates that a note or chord is to be played as **short as possible.**

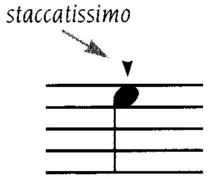

Used only with note values of a quarter note or less.

The *staccatissimo* is **centered on the notehead**. Placement is in the next space from the notehead, whether the note is on a line or space.

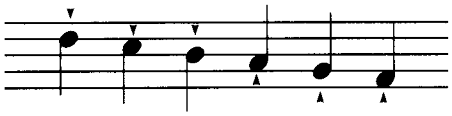

Center on the stem for opposite stem direction.

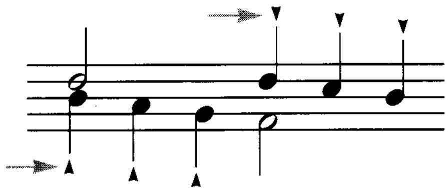

Staccato

A dot above or below a note or chord indicates that the note or chord is to be played **short.**

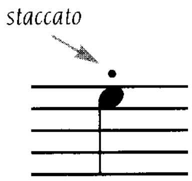

Used with note values of a quarter note or less.

The *staccato* is **centered on the notehead.** Placement is in the next space from the notehead, whether the note is on a line or space.

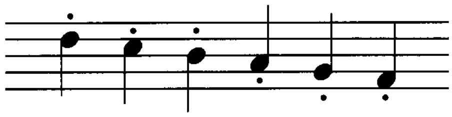

Center on the stem for opposite stem direction.

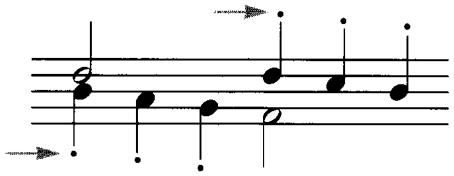

Tenuto

A short, heavy line above or below a note or chord, **thicker than the staff lines.**

A tenuto indicates that a note or chord is to be **held for its full value** (durational articulation), or the intent may be to apply a **slight stress** (articulation of force).

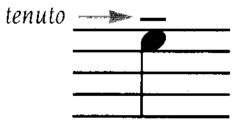

However it is interpreted, the tenuto is meant to ensure that a note or chord will be given a **certain amount of attention.**

This emphasis allows the tenuto to be **flexible in its application,** especially when combined with other articulations.

The tenuto is sometimes referred to as a *sostenuto* or *stress.*

Used with notes of any rhythmic value.

The *tenuto* is **centered on the notehead.** Placement is in the next space from the notehead, whether the note is on a line or space.

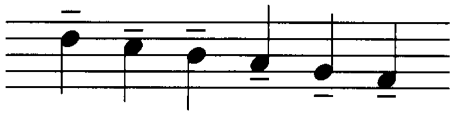

Also **center to the notehead,** not the stem, for **opposite stem direction.**

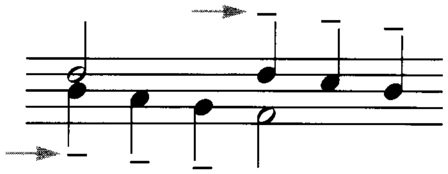

ARTICULATIONS OF FORCE

There are three articulations that affect the **force of attack** of a note or chord: the *tenuto* [–] discussed earlier, *accent* [>] and *marcato* [Λ].*

The following illustrates articulations of force compared to an unmarked note, from lesser to greater force of attack.

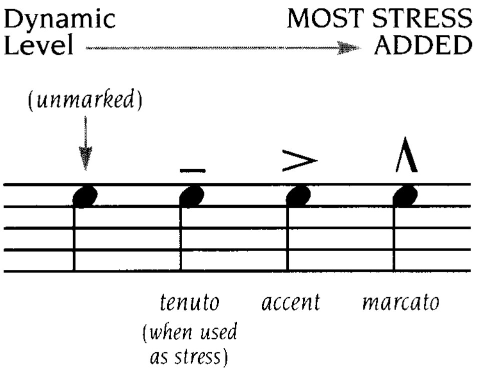
Dynamic Level
MOST STRESS ADDED
(unmarked)
tenuto (when used as stress)
accent
marcato

*Terminology varies for the symbols [>] and [Λ]. For the sake of this discussion, the words *accent* and *marcato* have been attached to the symbols.

Accent

When an *accent* is placed over or under a note or chord, the note or chord is to be played with **more attack, more marked.**

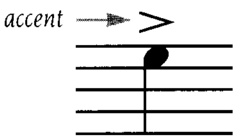

Used with notes of any rhythmic value.

The *accent* is **centered on the notehead.** Preferred placement is **outside the staff.**

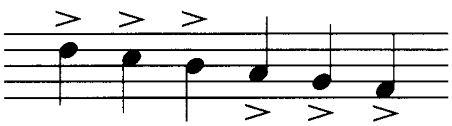

Although it is to be avoided, sometimes placing an accent **within the staff** is acceptable. A decision needs to be made whether it is preferred to have the accent closer to the notehead, or outside the staff.

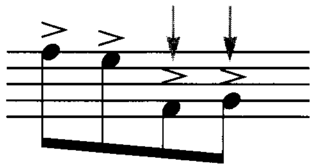

Center to the notehead, not the stem, for **opposite stem direction.**

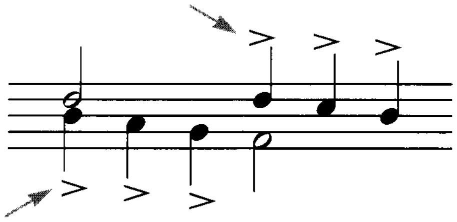

Marcato

When the *marcato* is placed over or under a note or chord, the note or chord is to be played with **even more attack, and more marked,** than an *accent.*

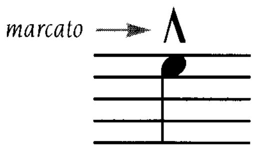

Used with notes of any rhythmic value.

The *marcato* is **centered on the notehead.** Preferred placement is **outside the staff** and ***above* the staff.**

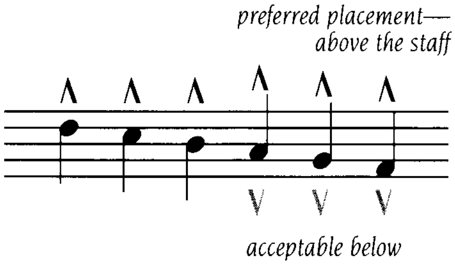
preferred placement—
above the staff
acceptable below

Also **center to the notehead** (not the stem) for **opposite stem direction.**

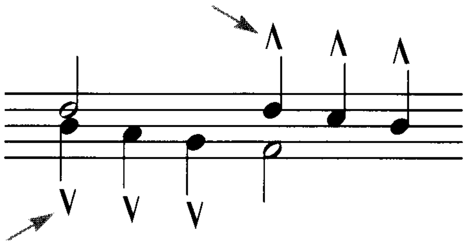

OTHER ARTICULATIONS OF FORCE

Sforzando (sf), forzando/forzato (fz) and *sforzato (sfz)* are also included in this articulation group and are interchangeable with accents and marcato accents, depending on the dynamic range.

These three articulations have the benefit of a built-in **indicator of dynamics.**

For instance, a sf should be indicated only in the dynamic ranges from ppp to f. With each increase of dynamic level, the sforzando reflects the higher dynamic level: sff in ff, sfff in fff.

The following chart illustrates the relationships between *accent* [>] and *marcato* [Λ] to *sforzando [sf], forzando [fz]*, and *sforzato [sfz]* within various dynamic levels.

Dynamic Range

+	*ppp* to *f*	*ff*	*fff*
>	*sf*	*sff*	*sfff*
Λ	*fz*	*ffz*	*fffz*
Λ̲	*sfz*	*sffz*	*sfffz*

A [>] in *f* is comparable to a [*sf*].
A [Λ̲] in *fff* is comparable to a [*sfffz*].

COMBINED ARTICULATIONS

Durational articulations can be combined with articulations of force to create variations of the basic articulations.

The following examples are the most used combinations.

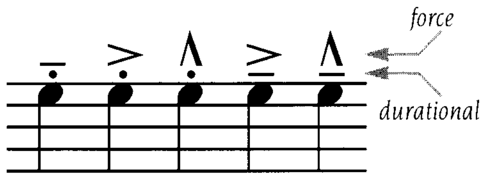

Staccatissimo and **staccato** are never combined (they basically have the same function in varying degrees).

Accent and **marcato** are never combined for the same reason.

Staccatissimo and **tenuto** are not commonly combined.

Staccatissimo and **marcato** are not commonly combined.

Combining articulations with tenuto

The **dual function** of the tenuto—the idea of *more importance, more attention,* whether of duration or force (or both)—allows the tenuto to be combined with either articulation group.

When a tenuto is **combined with an articulation of force,** the tenuto has more of a *durational function,* but may also retain its idea of force.

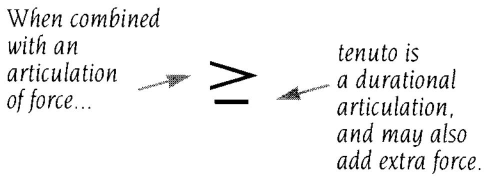

When combined with an articulation of force...

tenuto is a durational articulation, and may also add extra force.

When a tenuto is **combined with a durational articulation,** the tenuto has the function of an *articulation of force,* while retaining its durational function.

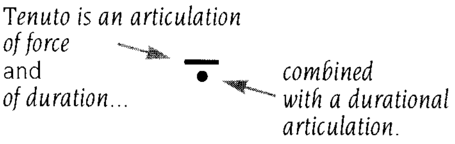

Tenuto is an articulation of force and of duration...

combined with a durational articulation.

Usual interpretation: play detached and with a slight stress.

PLACEMENT WHEN COMBINED

The following can be used as guides for the placement of combined articulations:

tenuto & staccato

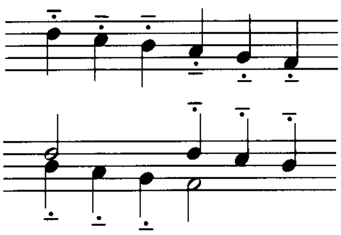

accent & staccato

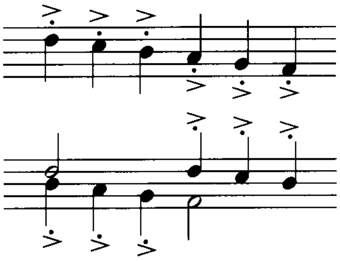

marcato & staccato

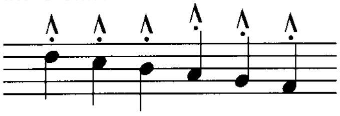

Keep the two combined articulations together (on the same side of the note) whenever possible.

Notice that when a **staccato** is combined with another articulation (and is on the same side of the note), it is **centered on the notehead,** even when it is on the stem side.

SIMILE

When an articulation (or pattern of articulations) is applied in a repetitious and predictable manner, the word ***simile*** may be used.

Simile is indicated **after the pattern has been clearly established** (usually two or three measures).

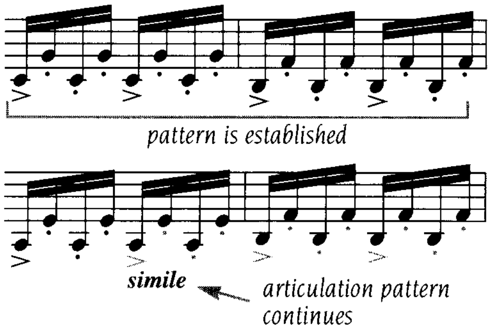

Notice that the ***simile*** was **not used at the beginning of the line,** even though the pattern had already been established.

The ***simile*** is ended when the pattern ends or when different articulations are added.

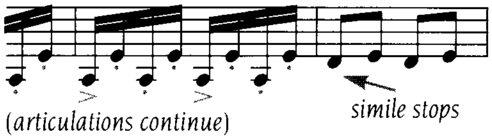

SLURS WITH ARTICULATIONS (See *Slurs*)

Augmentation dot

PLACEMENT

Always place **one dot per space** and **only in a space**—never on a staff line or on the same level as a leger line.

For a line note—place to the right of the notehead, **in the space above.**

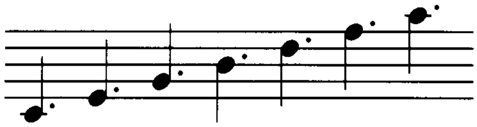

For a space note—place to the right of the notehead **in the same space** as the note.

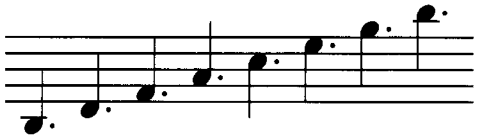

For notes with a flag—the dot is placed further right, altogether **avoiding the flag.** Never place a dot between the notehead and its flag.

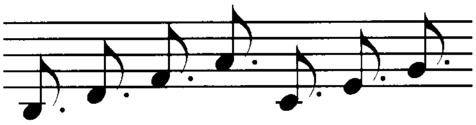

Double and triple dots are placed directly to the right of the first dot. Horizontal spacing is equal to that of the first dot. *Multiple dots should be used only in situations where they will be easily understood.*

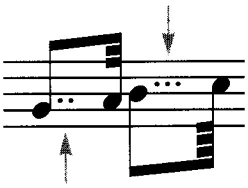

Never separate a dot from its notehead. Stem direction and creative positioning of a dotted note must be considered.

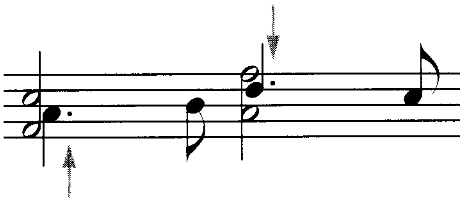

Ties always avoid the dot—place the tie clearly to the right of the dot.

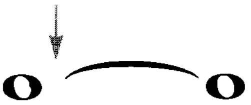

For two or more notes in a chord with the same stem direction, **always align the dots vertically.**

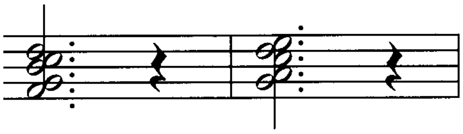

Two dots never share the same space.

FOR INTERVALS OF A SECOND

The dot for the space note is always placed in the space to the right of the notehead.

If the **line note is above,** place the dot in the space above. If the **line note is below,** place the dot in the space below.

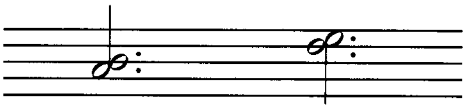

FOR TWO OR MORE PARTS

For stems up, place dots for line notes in the space above. **For stems down,** place in the space below.

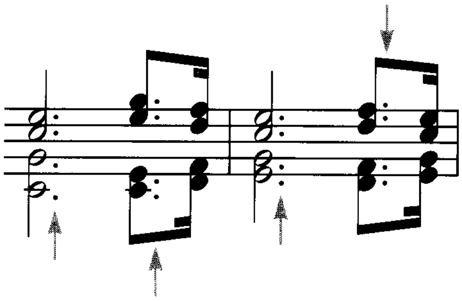

DOTTED RESTS

Rests are dotted in the same way notes are dotted, but with some restrictions on their use.

Place the dots as follows:

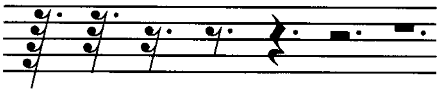

Dotted rests are used mainly to **clarify the subdivision of a measure.** (See *Meter*)

Dotted whole rests are the least used of all the dotted rests. They are never used in simple meter.

Possible use of dotted whole rest:

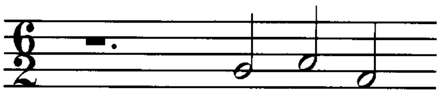

Likewise, **a dotted half rest** should not be used in simple meter.

Possible uses of dotted half rest:

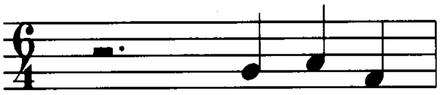

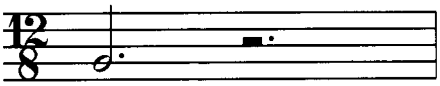

Dotted quarter rests are used in compound meters. Avoid using them in simple meters.

In 6/8 time, dotted quarter rests can be used on beats 1 and 4 only.

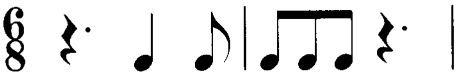

In 9/8 time, dotted quarter rests can be used on beats 1, 4 and 7 only (the pulses).

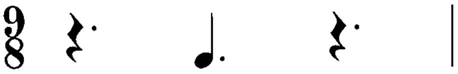

In 12/8 time, do NOT use a dotted half rest across the middle of the measure. Use dotted quarter rests.

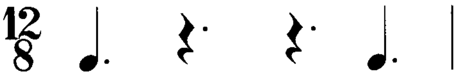

It is acceptable to use the **dotted eighth rest, dotted sixteenth rest, etc.,** for a cleaner look.

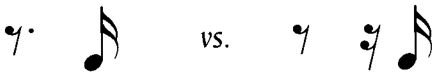

Barlines

Barlines are vertical lines that divide a staff into measures.

SINGLE BARLINE

The **thickness of a barline** is equal to, or greater than, the staff line thickness (staff lines being thicker than stems).

To help the performer follow the music from staff to staff, **avoid the alignment of barlines** from staff to staff (or from system to system).

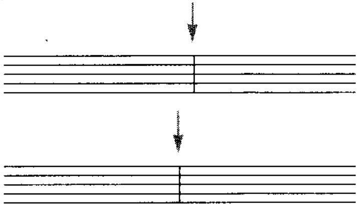

Allow a small amount of space after a barline—depending upon the density of notes within the measure.

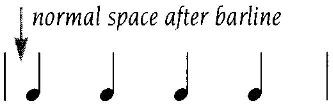

Less space may be allowed **after** the barline if the **density of notes is high.**

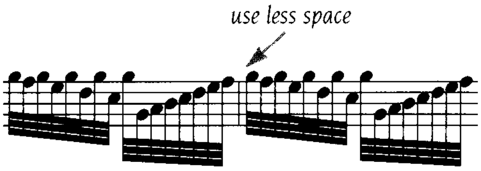

More space may be allowed **after** the barline if the **density of notes is low.**

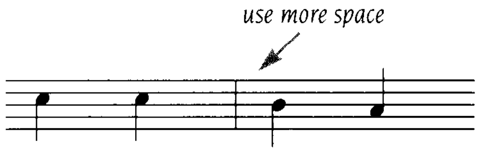

Less space may be allowed **before** the barline if **density is high.** Avoid *adding* space before the barline.

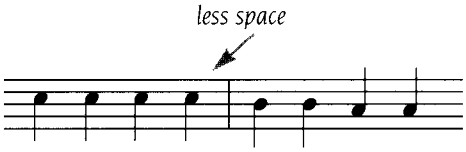

DOUBLE BARLINE
(Thin/thin)

Used to clarify the end of a section of music, such as in a *Da Capo* layout.

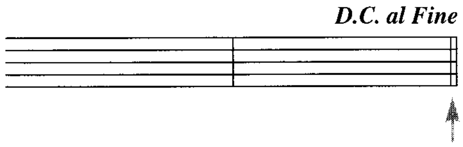

FINAL DOUBLE BARLINE
(Thin/thick)

Marks the **end of a composition, or the end of a movement** within a larger work.

Consists of a thin barline followed by a thick barline.

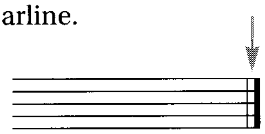

BROKEN BARLINES

Broken (dashed) barlines are sometimes used when a situation calls for the completion of a section **mid-measure,** although the **final double barline** is preferred.

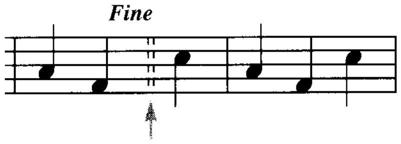

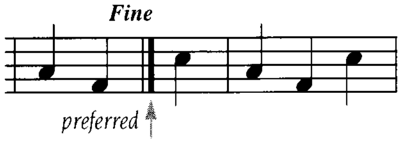
Fine
preferred

For composite meters, a broken barline is sometimes used to **clarify the meter division** within a measure.

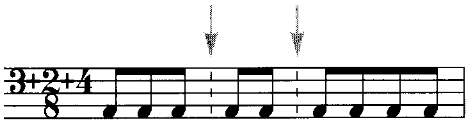

THE LEFT BARLINE

A single staff should ***never*** **begin** with a barline.

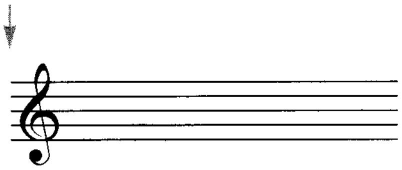

SYSTEMIC BARLINE

A systemic barline connects two or more staves together, as in the grand staff.

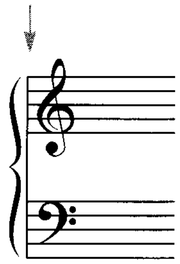

Barlines are **used to group staves of like instruments** in a score layout.

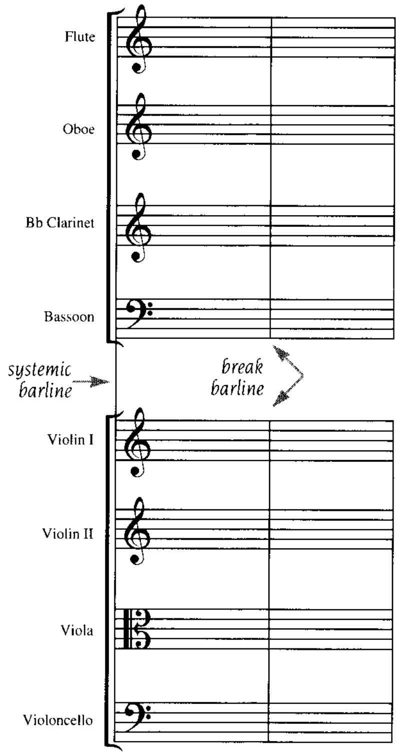
Flute
Oboe
Bb Clarinet
Bassoon
systemic barline
break barline
Violin I
Violin II
Viola
Violoncello

WITH CHANGING CLEFS

The new clef (cue size) is placed **before the barline** if the clef changes mid-staff.

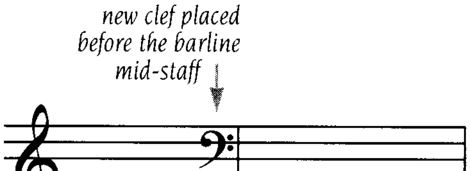
new clef placed before the barline mid-staff

A **courtesy clef is placed before the last barline** of the staff if the new clef begins immediately on the next staff.

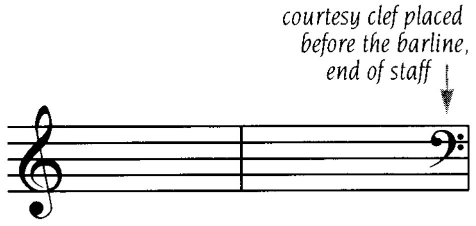
courtesy clef placed
before the barline,
end of staff

(See *Clef signs*)

WITH CHANGING KEY SIGNATURES

The new key signature is **placed after** an added **double barline** if the key signature changes mid-staff.

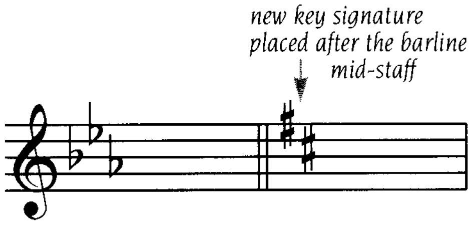
new key signature
placed after the barline
mid-staff

A **courtesy** key signature is **placed after** the last barline (add double bar) if the new key signature begins on the following staff.

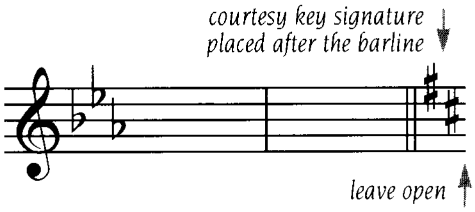
courtesy key signature
placed after the barline
leave open

Note: *A single barline may be used in both the above instances, but a double bar calls more attention to the change.*

WITH CHANGING TIME SIGNATURES

The time signature is **placed after** the barline if the time signature changes mid-staff.

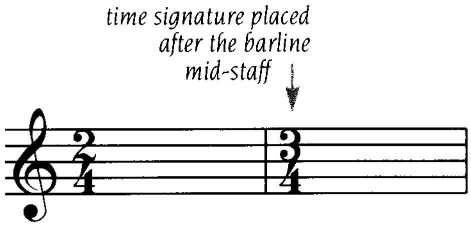
time signature placed
after the barline
mid-staff

A **courtesy** time signature is **placed after** the last barline if the new time signature begins on the following staff.

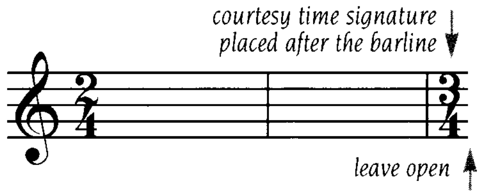
courtesy time signature
placed after the barline
leave open

WITH MULTIPLE COURTESY SIGNS

If a courtesy clef, key signature and time signature all appear together, they follow the rules for each individual courtesy change.

The order is the same as the order presented at the beginning of the piece (clef, key signature, time signature).

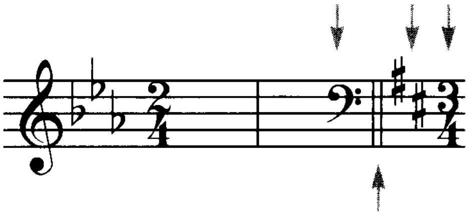

INCOMPLETE MEASURES

If a piece begins with a **pick-up,** the barline is positioned **according to the time value of the beat(s).**

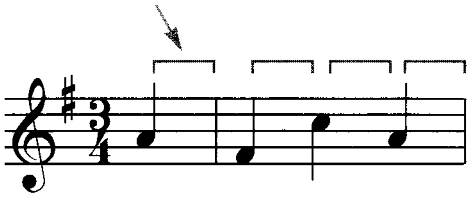

Beams

Although sometimes referred to as "balkans" or "ligatures," **"beam"** is the preferred term for the connection between two or more consecutive notes, replacing the flag.

Beams greatly simplify the reading of music, **substituting for individual flags** in groupings of **notes smaller than a quarter note.**

Because of the **ease in reading of beams,** the **use of flags in vocal music**—in relation to the lyric—has become **obsolete.**

CONNECTION TO THE STEM

Beams are connected to the side of the stem. They must be **flush to the stem,** both vertically and horizontally.

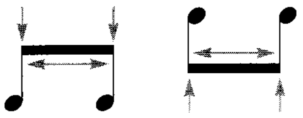

PRIMARY AND SECONDARY BEAMS

The beam that is furthest from the noteheads and connects a group of notes is called a *primary beam.* This beam **remains unbroken** throughout the grouping.

Any beam other than the primary beam is a *secondary beam.* The secondary beam **may be broken,** to divide the grouping into smaller units for easier reading.

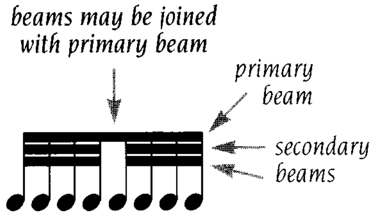

Stems extend through secondary beams and connect to the primary beam.

DOUBLE BEAMS

Normal stem length will accommodate a double beam—the stem need not be adjusted.

TRIPLE AND QUADRUPLE BEAMS

When a third or fourth beam is added below the secondary beam, **the stems must be extended** by approximately one staff space per beam.

The extension of the stems allows the space between the noteheads and the lowest beam to remain normal.

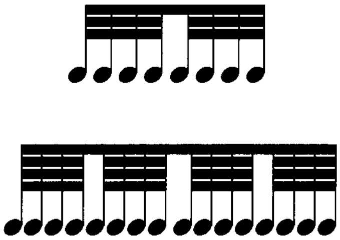

If 32nd- and 64th-note values occupy the majority of the rhythmic values, it may be better to double the rhythmic values of the entire piece and alterthe tempo.

FRACTIONAL BEAMS

A fractional beam is also a **secondary beam.**

Fractional beams are **associated with only one note.**

The **length** of a fractional beam is the same as the width of a notehead.

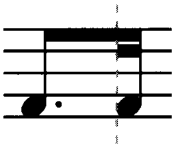

Place exactly the same as a full-length secondary beam would be placed. The beam is always *inside* the grouping, and it usually follows or precedes a beamed dotted note.

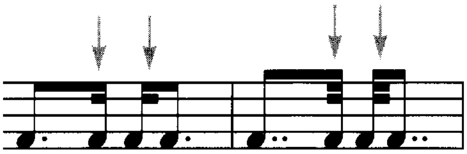

BEAMING AND METER

The basic purpose of beaming is to **connect two or more notes within the same beat (or pulse).** Meter must be considered when grouping notes with a beam. (See *Meter*)

In **any simple meter,** each beat is divisible by two; a beam may connect the two notes. Simple duple, simple triple and simple quadruple meters each have slightly different considerations.

Simple duple meter ($\frac{2}{2}$, $\frac{2}{4}$, $\frac{2}{8}$)

In some cases, the entire measure may be beamed.

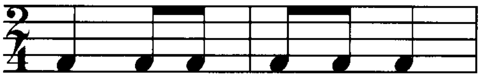

or

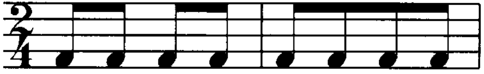

Some combinations of beams, flags and rests in simple duple meter:

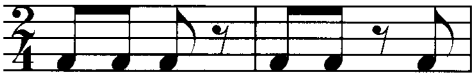

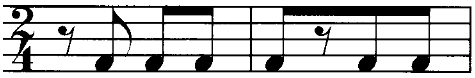

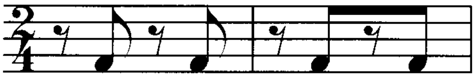

Although some contemporary engravers beam over a rest at the beginning or ending of a beamed grouping 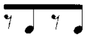, it is not the preferred notation.

Simple triple meter ($\frac{3}{2}$, $\frac{3}{4}$, $\frac{3}{8}$)

Consideration is given to **visually maintain** the three beats within the measure. In some cases the entire measure may be beamed.

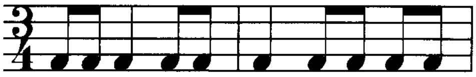

or

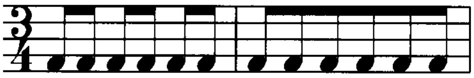

It is better to **avoid arbitrary groupings** of 4+2 or 2+4. However, they are sometimes used to emphasize phrasing.

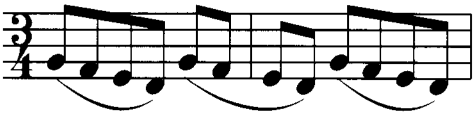

The combined use of **flags and beams** clarify the meter.

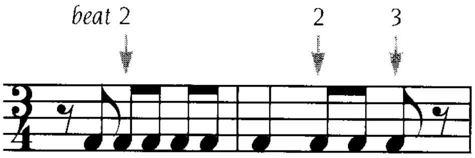

Avoid beaming in a manner that would suggest $\frac{6}{8}$ time. Use flagged notes to differentiate $\frac{3}{4}$ meter from $\frac{6}{8}$ meter.

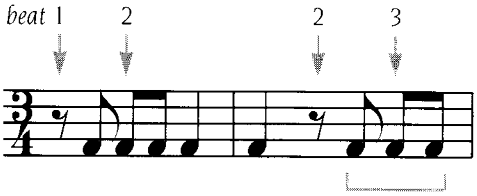

use flagged notes to avoid groupings of three

Beaming to include rests inside the grouping is acceptable.

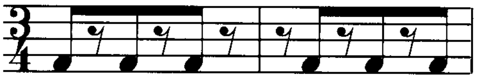

Simple quadruple meter ($\frac{4}{4}$, $\frac{4}{8}$)

In $\frac{4}{4}$ time, beats 1 & 2 and beats 3 & 4 may be beamed together to form a unit.

Never connect the notes of beat 2 with beat 3.

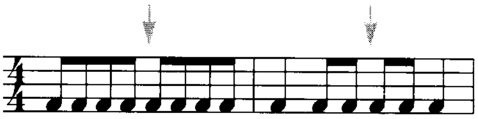

The basic rule in $\frac{4}{4}$ time is that the **two halves** of the measure (beats 1 & 3) **must be immediately recognizable.** The only exception to this would be a very simple syncopation (such as 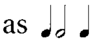).

Because of this rule, all the notes in the measure would *not* be beamed together, as is possible in the other simple meters.

or

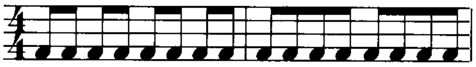

THE SAME RHYTHM THREE WAYS

Each could be the best for a particular situation.

Notice that **each beat** is recognizable in the following example.

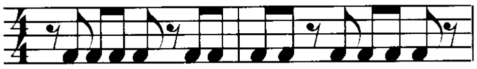

Notated with beams only, the **two halves** of the measure are easily recognizable:

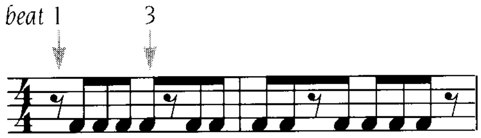

If the above example were in a piece that contained triplets, it could at first glance be **mistaken for triplets.** In that case, a better notation would be:

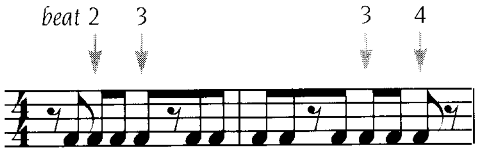

For any simple meter, **when a beat is uneven or contains more than two notes,** the two adjacent beats should not be beamed together as a unit.

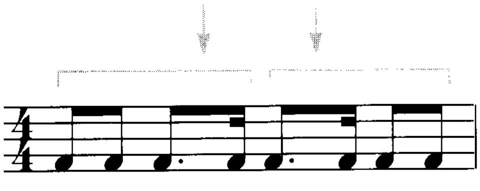

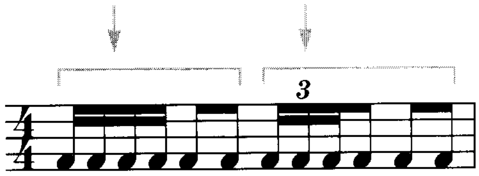

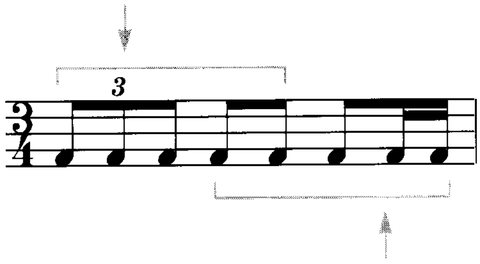

Compound meter

When each **pulse** of the measure is **divided into three beats,** the meter is called "compound." A beam may connect the three beats.

compound duple (2 pulses, 6 beats)

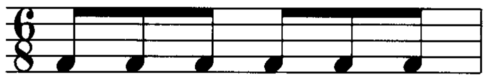

compound triple (3 pulses, 9 beats)

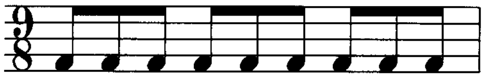

compound quadruple (4 pulses, 12 beats)

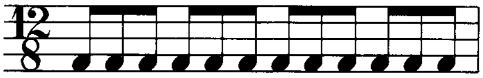

The **units** of three beats **for each pulse** in compound meter **must always be maintained.** Even when rests are included in the measure, as in the example below, beats 1, 4 & 7 must be clearly visible.

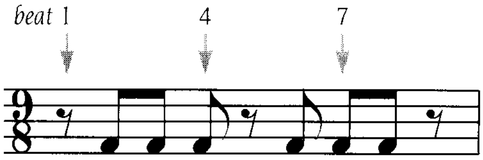

The three units in the above example could be made even more obvious by including the rest on beat 5 in a beamed unit.

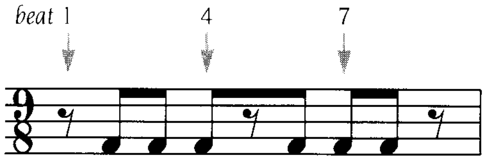

It is **possible to group all notes within a pulse.** All the notes in the *measure,* however, cannot be beamed together—this could cause the measure to appear to be in some other meter.

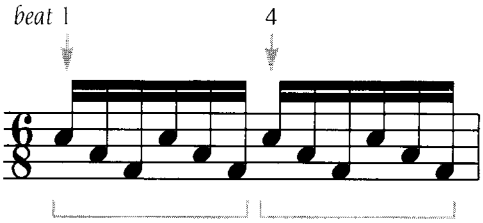

Although different groupings are sometimes used for phrasing or interpretation, the above should be generally adhered to.

TIES AND BEAMS

Avoid using tied notes *within* a beamed grouping. The **unit should be broken** where the tie is placed.

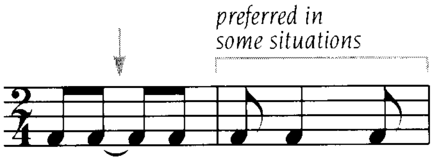
preferred in
some situations

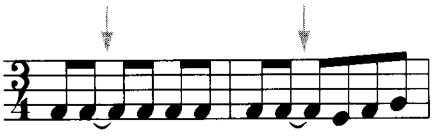

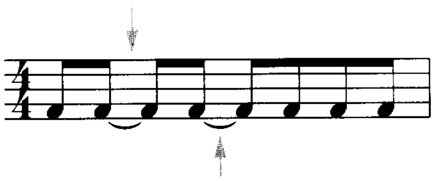

In this case, there would not be a continuous beam, since beats 2 and 3 should not be beamed together in this meter.

STEM DIRECTION OF BEAMED GROUPS

Simple rules for stemming of single notes apply to beamed groups when possible (for example, if all notes in a beamed grouping are on or above the middle line of the staff, stems are down), but some additional guidelines are needed.

(See *Stems*)

For two beamed notes

Stem direction is **determined by the note farther from the middle staff line.**

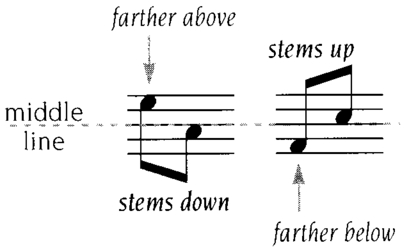
farther above
stems up
middle
line
stems down
farther below

If both notes are **the same distance** from the middle line, the preference is for **stems down.**

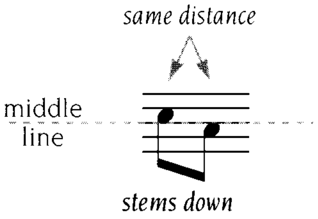
same distance
middle
line
stems down

If all other stems in the surrounding measures are up-stem, that direction is sometimes continued if the two notes are the same distance from the middle line. (This is a very subtle point, and somewhat obsolete.)

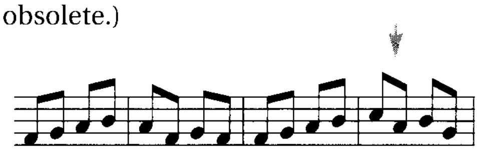

For three or more beamed notes

If the majority of the notes are **on or above** the middle line, the stems go down.

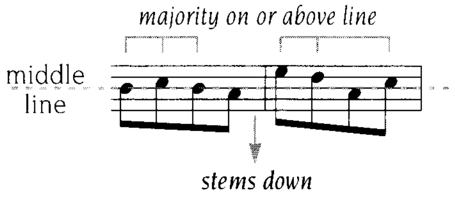
majority on or above line
middle
line
stems down

If the majority of the notes are **below** the middle line, the stems are up.

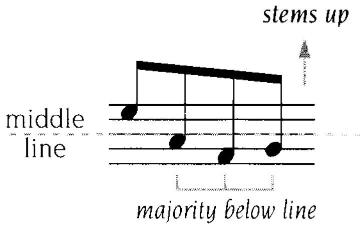
stems up
middle
line
majority below line

If the number of notes **above and below the middle line are equal,** the note farthest from the middle line dictates the stem direction. (If all notes are the same distance, the preference is down.)

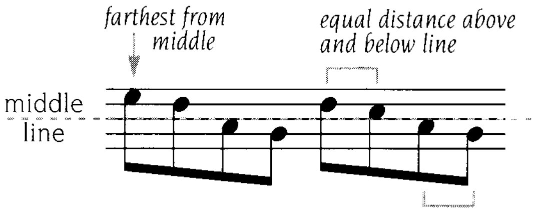
farthest from
middle
equal distance above
and below line
middle
line

BEAM SLANT & PLACEMENT

Concerning the computer

With most computer engraving today, the direction, slant and placement of the beam are automatic.

However, the defaults—or even the settings allowable—may not conform to traditional practices. Even though some of these practices are in the process of change, there are some guidelines to keep in mind for a more professional look.

Beginning placement of the beam

When **stems are up,** the stem length of the **highest note** determines the beam placement.

When **stems are down,** the stem length of the **lowest note** determines the beam placement.

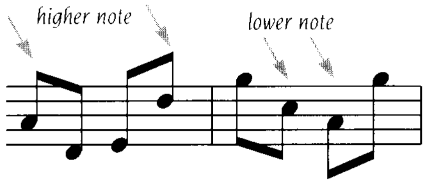

Direction of slant

Beams *usually* **follow the contour of the notes,** ascending or descending.

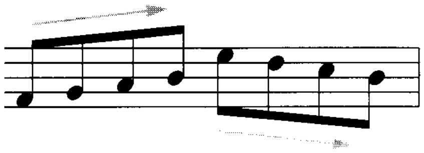

For mixed direction of notes, the beam *usually* **follows the direction between the first and last note** of a grouping.

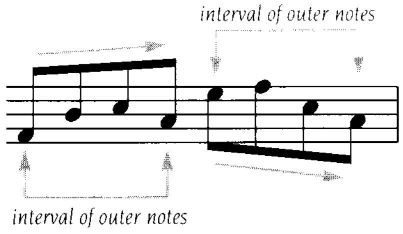

When two or more notes share a stem, the note **closest to the beam** dictates the beam direction.

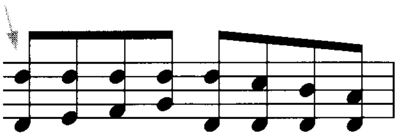

Horizontal beaming

A beam should be horizontal if the first and last note of the beamed group is on the **same pitch level.**

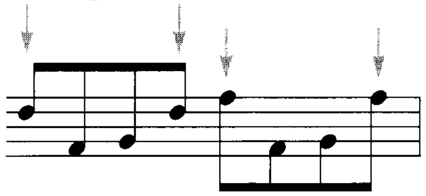

If the grouping consists of **repeated intervals,** the beam should be horizontal.

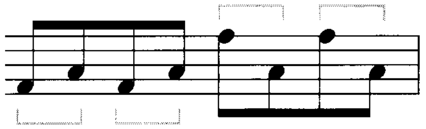

Horizontal beams may be used if inner notes **do not follow the interval direction** of the outer notes.

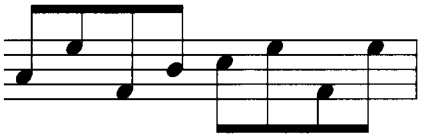

Some other examples of horizontal beaming:

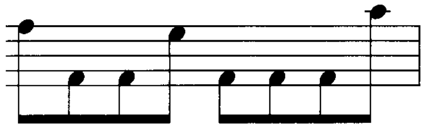

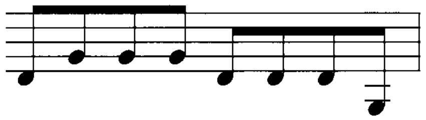

Interval-specific beam slant

Determining the placement and slant of beams in traditional plate engraving was complicated and detailed.

The following are *approximate* guidelines, **modified from the strict traditional** rules for beaming of intervals. There are two main reasons for this modification:

1. The consistency of the computer and the quality of modern printing make it no longer necessary to avoid the small "wedges" of white space that in the past often filled in with ink.
2. Adjusting many beams individually on the computer is very time-consuming.

A pair of **notes of different pitch should not have a horizontal beam.** (Exception: two notes on the same line or space, one or both altered with an accidental)

If a **beam slants, it is usually from 1/2 to 2 staff spaces**—very seldom more.

Intervals of a **2nd** slant 1/2 space.

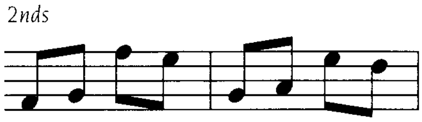

Intervals of a **3rd** slant 1 staff space.

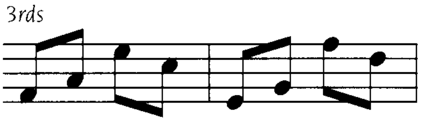

Intervals of a **4th** slant from 1 to 1-1/2 staff spaces.

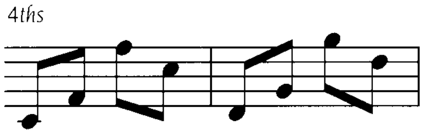

Intervals of a **5th, 6th & 7th** each slant approximately 1 to 1-1/2 staff spaces.

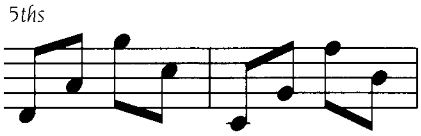

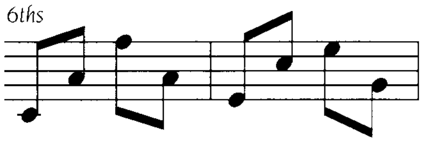

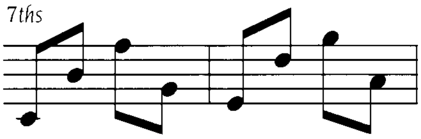

Intervals of an **octave and greater** slant 1-1/2 to 2 staff spaces.

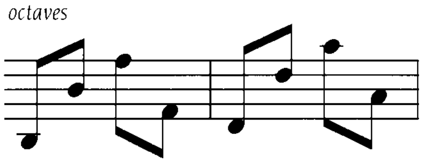

Beamed notes with **leger lines** slant 1/2 staff space. (Remember that stems must extend to the middle line.)

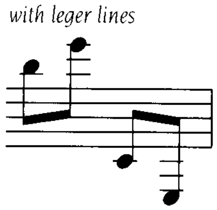

Depending upon the **amount of space** allowed between the beamed notes, the slant of the beam naturally *decreases* with an *increase* in the amount of space allotted, if stem lengths remain the same.

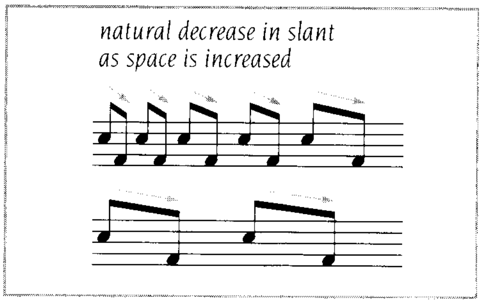

For tightly spaced beamed notes, **adjust the slant** (by lengthening the stem of the farthest note) so that it is not so extreme.

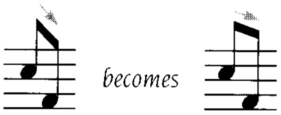

Alternative beaming

For grand staff notation, **cross-staff beaming** is acceptable in some situations. The slant of the beam is determined by the amount of space between the two staves, the length of the stems, and by avoiding exaggerated angles.

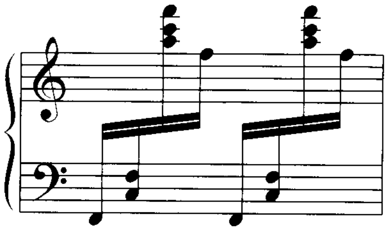

For single-staff music, alternative beaming should be used only in **exceptional circumstances.** Stems must not be too short.

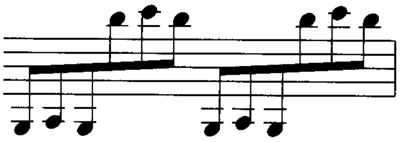

Bowing

Down-bow and **up-bow** markings are centered to the notehead and are placed above the staff, except when two parts share a staff.

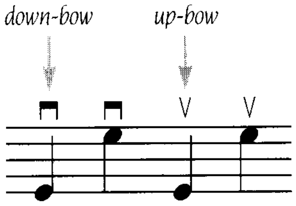

String techniques such as ***spiccato, legato, portato*** and ***jeté*** are notated conventionally, with slurs and/or articulations.

The study of orchestration is necessary to properly specify string indications.

Slurs can indicate either phrasing or bowing, or both.

Brackets to the left of the chord may be used for **multiple stops** if the chord is to be divided a particular way. Usually no bracket is used.

Multiple stops may be *double, triple* or *quadruple* stops.

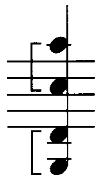

PLACEMENT OF TEXT

All text for bowing indications is **placed above** the staff.

When a particular string is to be played exclusively, ***sul*** [string] is indicated.

sul G

Pizzicato is cancelled with the ***arco*** indication.

pizz. ***arco***

With any of the following indications, ***ord.*** or ***ordinario*** is commonly used to indicate the return to normal playing.

non vib. ***sul pont.*** ***sul tasto*** ***col legno***

Brace

The brace is used to connect the two staves in scores for instruments such as piano, organ, harp and marimba.

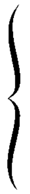

The brace connects two staves to form a **grand staff.**

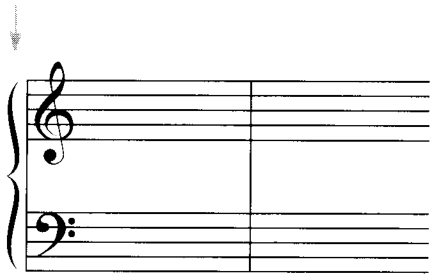

Bracket

The bracket is used **to connect two or more staves** to form a system or to **group certain instruments** within a system.

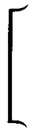

The bracket is different from the brace in that it connects the staves for two or more **separate instruments,** while the brace connects staves for one instrument (such as keyboard).

The grouping may be two or more **unlike** instruments, **identical** instruments, or a **family** of instruments, such as strings.

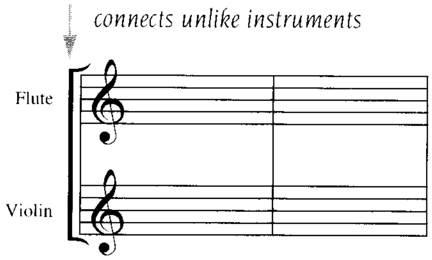

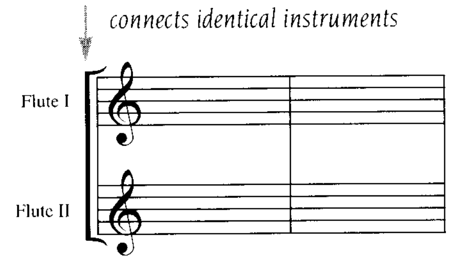

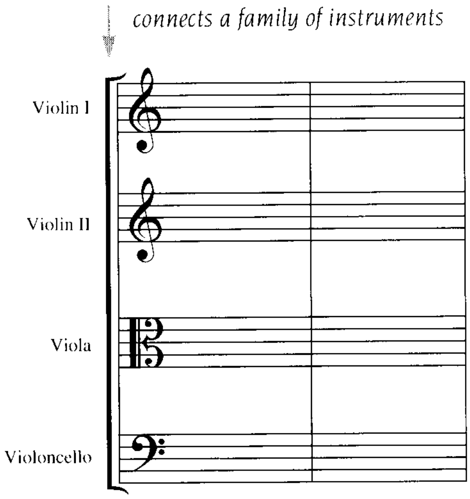

(See *Barlines, Scores*)

Breath mark

The breath mark (,) has two similar uses:

- **an actual breath indication** for instruments (including voice) requiring the breath to sound.
- **a small pause or break** (as if taking a breath) for instruments *not* requiring breath to sound.

A breath mark may or may not interrupt the tempo. If it does interrupt the tempo, the break or pause is shorter than that of a caesura or fermata.

PLACEMENT

The breath mark is indicated just above the staff, at the point **where a breath or break is desired.**

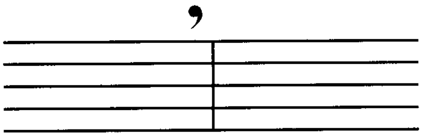

If the breath mark occurs between two notes, the placement is toward the **end of the space for the full value** of the note.

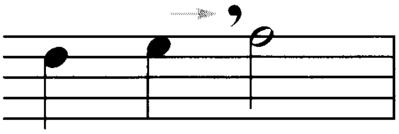

If the breath mark occurs at the end of a measure, the placement is just **before the barline.**

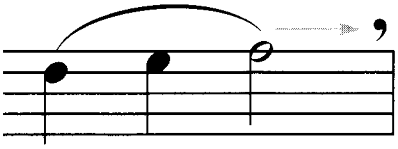

Caesura

The caesura (//) is a small pause or break in the tempo (implying a break in sound).

The pause of the caesura is **slightly longer than a breath mark,** but shorter than a fermata.

PLACEMENT

The caesura is placed on the **fourth line from the bottom** of the staff (extending to the first leger line), at the point **where a break is desired.**

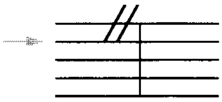

The caesura is placed toward the **end of the full value** of the note.

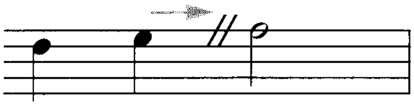

If space is tight, the caesura **can be centered** between two notes.

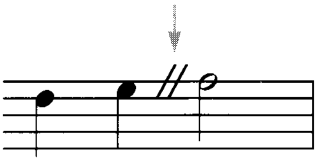

If the caesura occurs at the end of a measure, the placement is just **before the barline.**

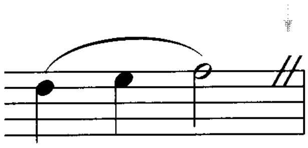

If a longer pause is desired, **a fermata may be placed above** the caesura.

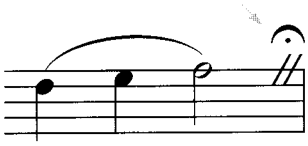

Chord frames & symbols

Chord frames are diagrams that contain all the information necessary to play a particular chord on a fretted instrument.

The photograph below shows the finger numbers corresponding to the fingerings used for guitar.

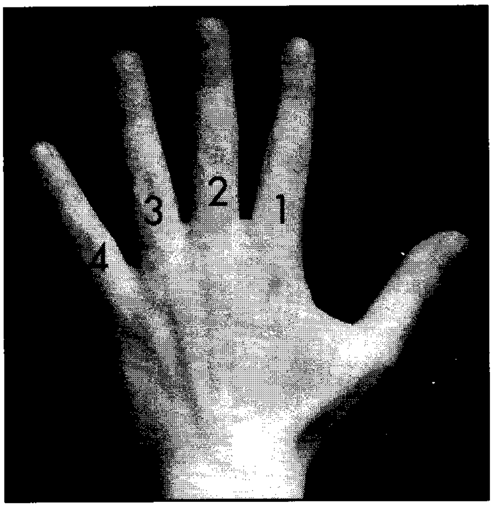

The number of strings and tunings vary between string instruments. The **chord frame is modified** (number of strings, tunings) according to the differences between each instrument.

The following examples explain the various **chord frame elements** for guitar.

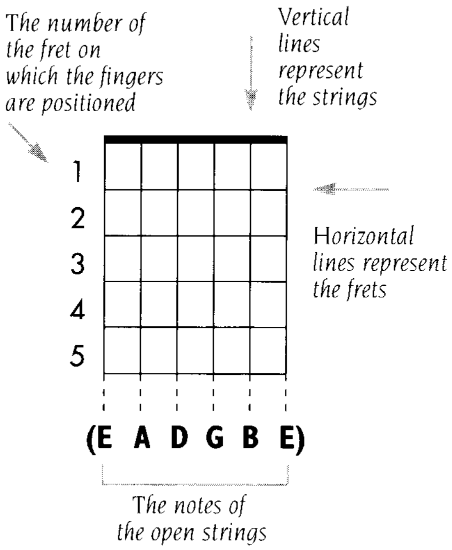
The number of the fret on which the fingers are positioned
Vertical lines represent the strings
1
2
3
4
5
Horizontal lines represent the frets
(E A D G B E)
The notes of the open strings

The **fingerings, note names** and **position** of the chord on the neck are all provided on the chord frame.

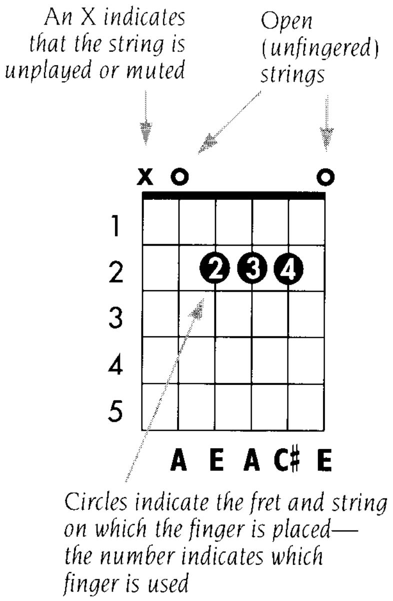
An X indicates that the string is unplayed or muted
Open (unfingered) strings
X O O
1
2
3
4
5
2 3 4
A E A C♯ E
Circles indicate the fret and string on which the finger is placed— the number indicates which finger is used

A "slur" indicates that the finger is placed flat, covering the marked notes

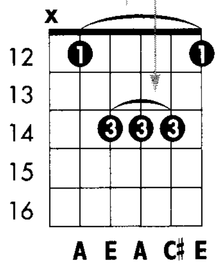
x
12
13
14
15
16
A E A C♯ E

Resulting note names

CHORD SYMBOLS

The **chord symbol** is always placed above the frame to identify the chord.

Chord symbol

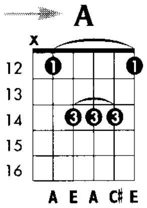
A
x
12
13
14
15
16
A E A C♯ E

Careful attention must be given to the **consistent identification** of chords, in the **least ambiguous way** possible.

Chord symbols often appear without a frame. Whether or not there is a frame, space must be allowed so that the symbol is **aligned with the proper beat.**

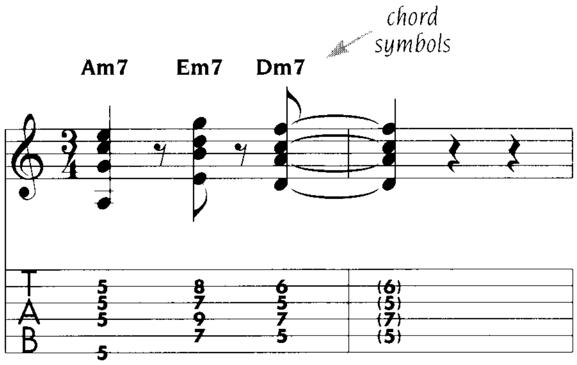
chord symbols
Am7
Em7
Dm7
T
A
B
5
5
5
5
8
7
9
7
6
5
7
5
(6)
(5)
(7)
(5)

(See *Tablature*)

A complete listing of guitar chords, along with chord theory, is provided in Alfred's Guitar Chord Encyclopedia *(Item #4432). It also includes a listing of chord symbols and a listing of ambiguously named symbols.*

Circle of fifths See *Key signatures*

Clef signs

The following are commonly used clefs.

Middle C is shown for comparison in all examples except neutral clefs.

TREBLE AND BASS CLEFS

Treble and bass clefs are most frequently used. Keyboard, harp and organ music is written using only these clefs.

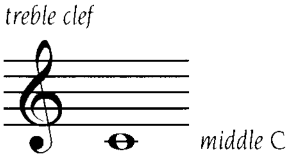
treble clef
middle C

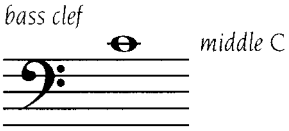
bass clef
middle C

The **octave treble clef** is sometimes used for tenor vocal parts, to indicate that the part is one octave lower than written.

The **double treble clef** serves the same function as the octave treble clef but is much less frequently seen.

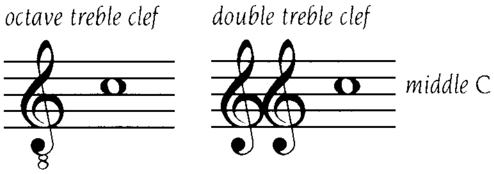
octave treble clef
double treble clef
middle C

The regular **treble clef** is also commonly used for tenor vocal parts; the octave is assumed.

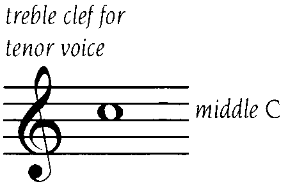
treble clef for tenor voice
middle C

C CLEFS

Alto clef is commonly used for viola, to avoid excessive leger lines. (Viola also uses treble clef.)

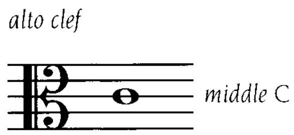
alto clef
middle C

Tenor clef is primarily used for bassoon, cello and trombone when playing in higher registers.

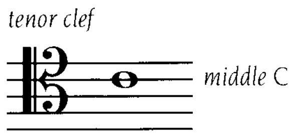
tenor clef
middle C

The following C clefs are **obsolete:**

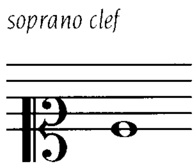
soprano clef

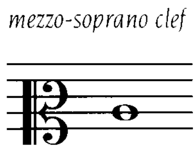
mezzo-soprano clef

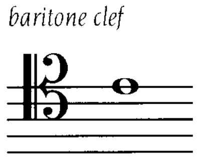
baritone clef

NEUTRAL CLEFS

Neutral clefs are used by percussion instruments of indefinite pitch. The following is most commonly used.

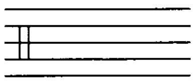

Neutral clef for single-line percussion:

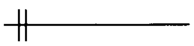

POSITIONING

Clefs are **slightly indented on the staff,** to the right of a systemic barline or the open single staff. Be sure to leave the ***single* staff open—do not put a barline before the clef!**

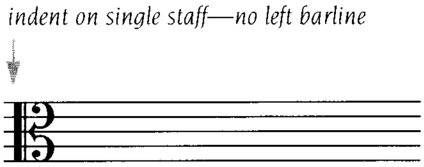
indent on single staff—no left barline

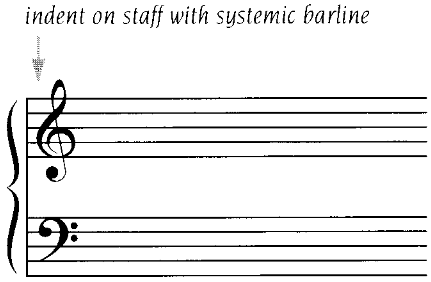
indent on staff with systemic barline

The clef is always **placed before the key signature and time signature.**

Notice the distance from the end of the staff to the clef, from the clef to the key signature, and from the key signature to the time signature.

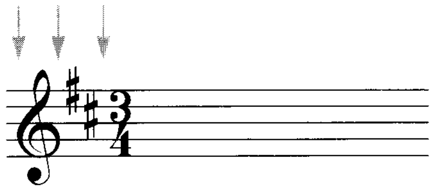

CLEF CHANGES

When the clef changes **within a staff,** a cue-size clef is used (usually 75% of the original clef size). Courtesy clefs are also cue size.

If a clef change affects a complete measure, it is always placed before the barline.

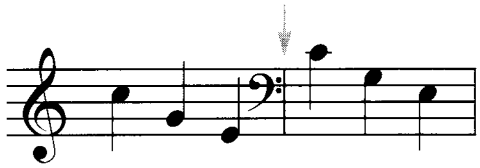

Only a *systemic* barline can precede a clef sign. (See *Systems*)

If a clef change is within a measure, the clef is placed directly before the first note involved.

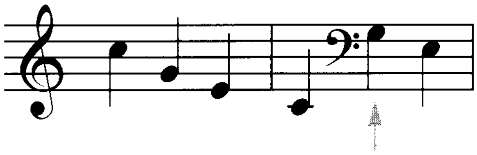

If a clef change **begins after a rest and is on a beat,** the clef is placed directly before the first note affected.

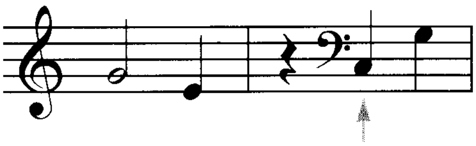

If a clef change **begins after a rest but is not on a beat,** the clef precedes the rest.

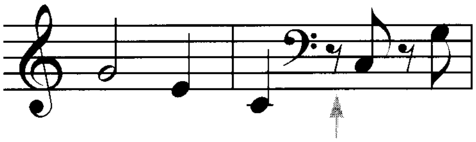

If a clef change **occurs at the beginning of a staff,** a courtesy clef is placed before the barline at the end of the previous staff, followed by the new clef at full size in the new staff.

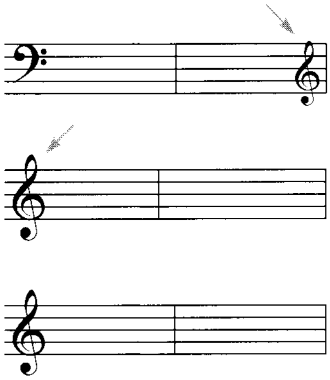

The rules for clef changes are not affected by the **addition of a key change and/or time signature change.** The clef sign precedes the barline; the key signature and time signature follow the barline.

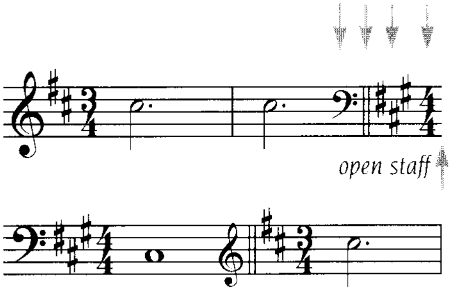

Notice that the **order** of courtesy clef, key signature and time signature is the same as the order at the beginning of a piece.

The **courtesy clef** is cue size; the key signature and time signature are normal size.

Also notice the **open staff** after the *courtesy* key signature and time signature.

Coda sign See *D.C. al Coda, D.S. al Coda*

Courtesy accidentals See *Accidentals*

Credits

Credits may consist of the **composer, lyricist** and **arranger,** or a word such as *traditional, anonymous, folk song,* etc.

1. The **composer** credit is placed flush right with the page margin. It is not horizontally aligned with the tempo marking.
2. The **arranger** credit is placed below the composer credit (also not horizontally aligned with the tempo marking).
3. The **lyric** credit is placed flush left with the page margin, horizontally aligned with the composer credit.

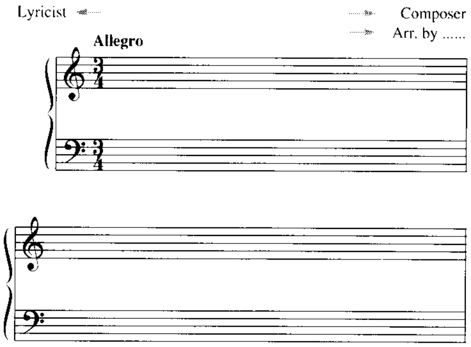

The typeface for all three credits is usually Roman.

For a less formal style, such as pop piano/vocal music, the lyric credit would go under the composer credit, with the arranger under that. **All credits would be flush right** with the page margin.

Some examples:

Words and Music by...................

Words and Music by
....

Music by...................
Words by...................
Arranged by...................

The word "arranged" may be abbreviated (Arr.).

Cue notes

Cue notes may be given as guidance only, **to assist a performer's entrance** after numerous measures of rest.

Cue notes may also be given for **possible performance** if the cue instrument is weak or missing. (See *Alternating instruments*)

The size of cue notes is somewhat **smaller than normal note size,** but still large enough to be legible (65–75% of normal note size).

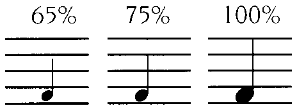

All musical elements associated with the cue notes will also be at **cue size.**

There are **two ways to notate** cue passages.

With **reverse stem direction** and added rests:

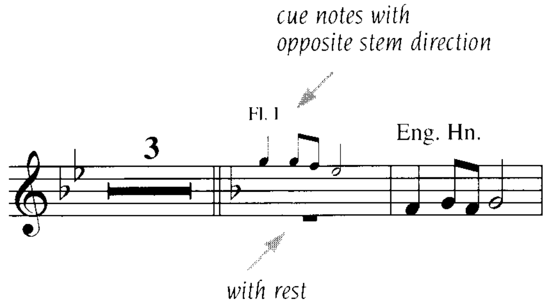
cue notes with
opposite stem direction
Fl. I
Eng. Hn.
3
with rest

With **normal stem direction** and no added rests:

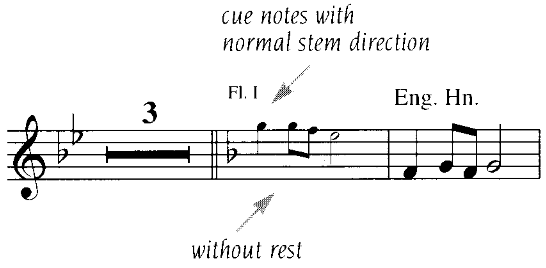
cue notes with
normal stem direction
Fl. I
Eng. Hn.
3
without rest

Piano accompaniment parts may include a cue-size solo part above the grand staff.

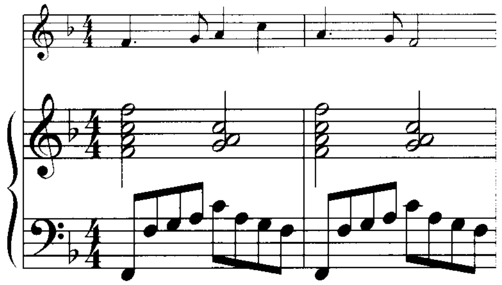

Not all notes that are cue *size* are used as a cue.

Ossia passages and **grace notes** are typically notated at cue size. (See *Ossia, Grace notes*)

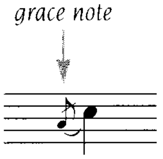

D.C. al Coda (da capo al coda)

Means go back to the beginning, play to the "to coda" sign (𝄌), then jump to the coda to finish the piece.

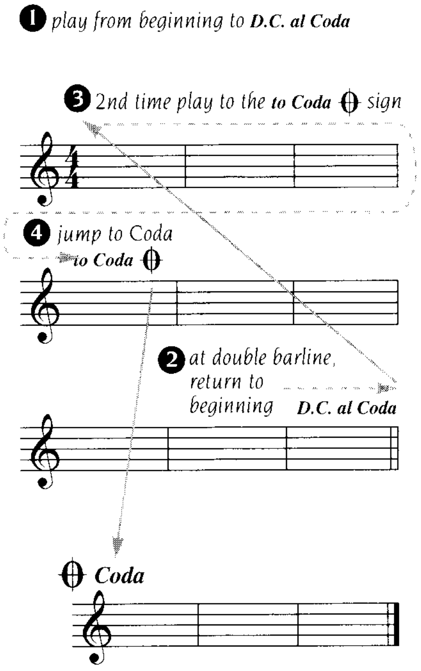
❶ play from beginning to D.C. al Coda
❸ 2nd time play to the to Coda 𝄌 sign
❹ jump to Coda
to Coda 𝄌
❷ at double barline, return to beginning
D.C. al Coda
𝄌 Coda

Align the edge of the "to coda" sign **flush right** with the barline.

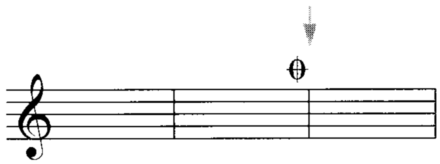

The words *to Coda* are often omitted from the "to coda" sign.

Place a **double barline** (thin/thin) at the ***D.C. al Coda.*** (***D.C. al Coda*** can also be written ***D.C. al Coda* 𝄌**)

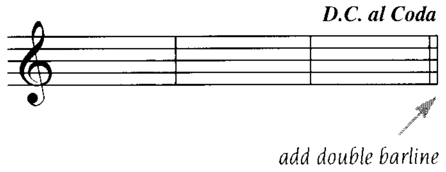
D.C. al Coda
add double barline

Separate the coda from the staff above (optional) and **indent.**

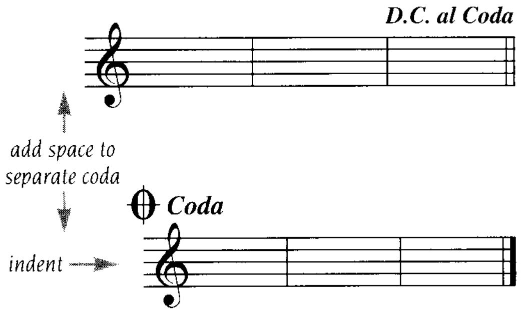
D.C. al Coda
add space to separate coda
𝄌 Coda
indent

Align the coda sign with the beginning of the staff. It should be somewhat **larger** than the "to coda" sign.

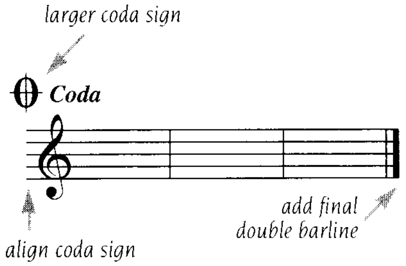
larger coda sign
𝄌 Coda
add final double barline
align coda sign

The coda will end with a **final double barline** (thin/thick).

It is sometimes necessary to begin the coda on the same staff as the ***D.C. al Coda.*** If so, **break the staff** and **add a clef and key signature.** (Do not add a time signature unless it changes at the coda.)

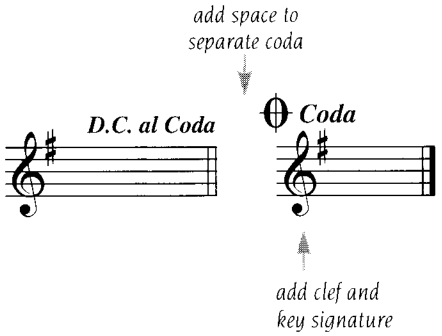
add space to separate coda
D.C. al Coda
𝄌 Coda
add clef and key signature

D.C. al Fine (da capo al fine)

Means go back to the beginning and play to the ***Fine.***

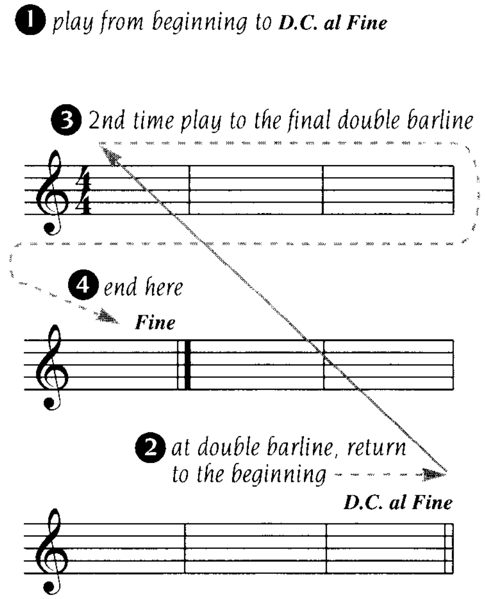
❶ play from beginning to D.C. al Fine
❸ 2nd time play to the final double barline
❹ end here
Fine
❷ at double barline, return to the beginning
D.C. al Fine

Place a **double barline** (thin/thin) at the ***D.C. al Fine.***

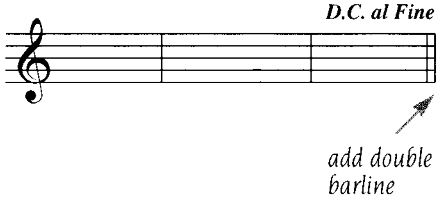
D.C. al Fine
add double barline

A **final double barline** (thin/thick) is added at the ***Fine.***

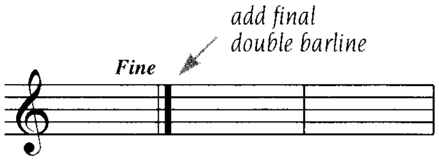
add final double barline
Fine

D.S. al Coda (dal segno al coda)

Means go back to the sign (𝄋), play to the "to coda" sign (𝄌), then jump to the coda to finish the piece.

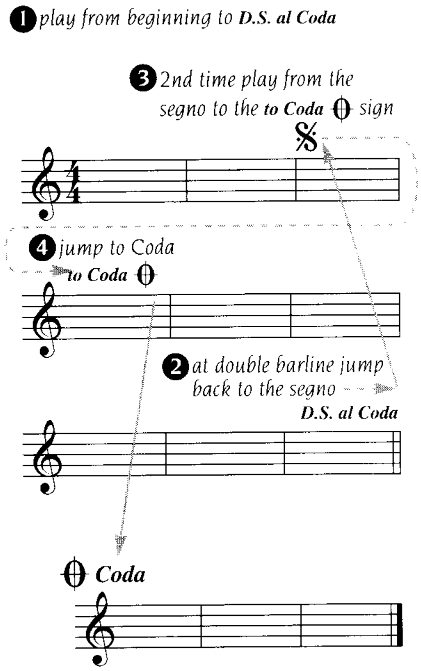
❶ play from beginning to D.S. al Coda
❸ 2nd time play from the segno to the to Coda 𝄌 sign
❹ jump to Coda
to Coda 𝄌
❷ at double barline jump back to the segno
D.S. al Coda
𝄌 Coda

The segno 𝄋 should be **aligned flush left** with the barline.

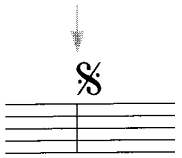

Align the edge of the "to coda" sign **flush right** with the barline.

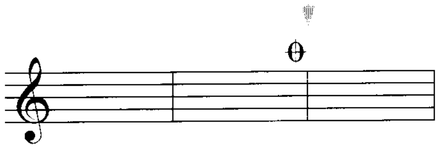

The words *to Coda* are often omitted from the "to Coda" sign.

Place a **double barline** (thin/thin) at the *D.S. al Coda.* (*D.S. al Coda* can also be written *D.S. 𝄋 al Coda 𝄌.*)

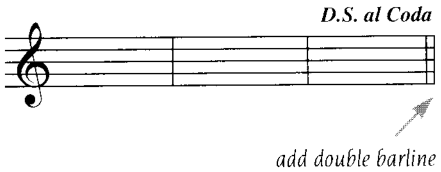
D.S. al Coda
add double barline

Separate the coda from the staff above (optional) and **indent.**

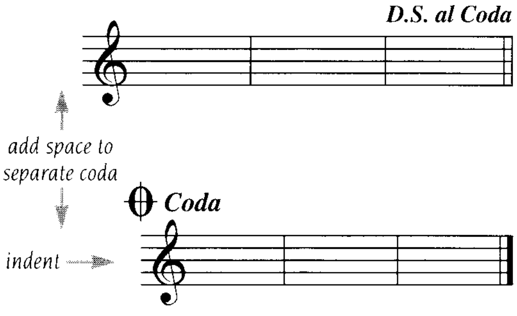
D.S. al Coda
add space to separate coda
𝄌 Coda
indent

Align the coda sign with the beginning of the staff. It should be somewhat **larger** than the "to coda" sign.

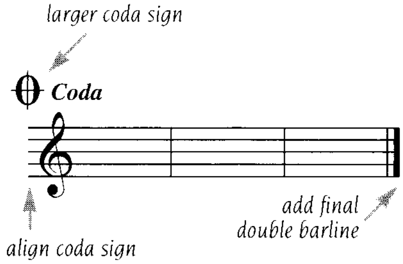
larger coda sign
𝄌 Coda
add final double barline
align coda sign

The coda will end with a **final double barline** (thin/thick).

D.S. al Fine (dal segno al fine)

Means go back to the sign and play to the *Fine.*

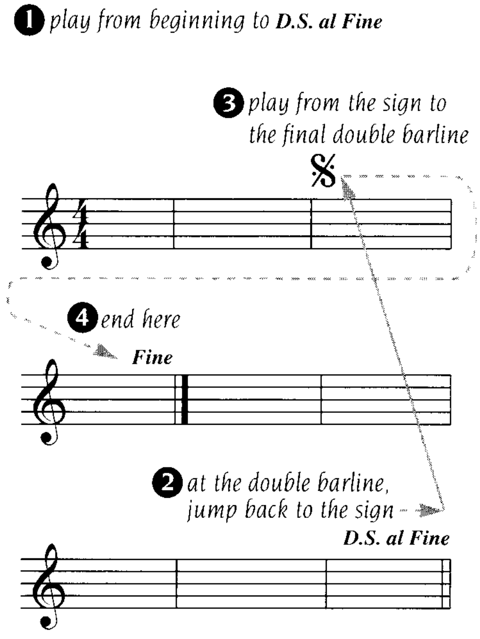
❶ play from beginning to D.S. al Fine
❸ play from the sign to the final double barline
❹ end here
Fine
❷ at the double barline, jump back to the sign
D.S. al Fine

The segno 𝄋 should be **aligned flush left** with the barline.

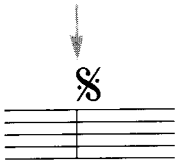

Add a **double barline** (thin/thin) at the ***D.S. al Fine.***

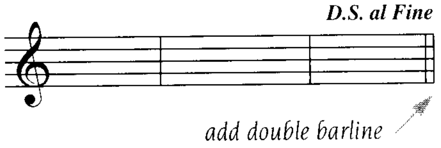

A **final double barline** (thin/thick) is added at the ***Fine.***

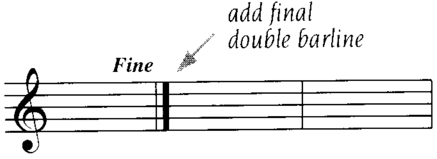

Divisi

Divisi is used for instruments (such as strings) that can produce **more than one note at a time,** indicating that the notes should be divided among the *group* of instruments (i.e., instead of multiple stops). (See *Sharing a staff*)

Divisi dots See *Tremolo*

Dotted rhythms See *Augmentation dot*

Dynamics

Dynamics indicate the varying degree of volume or intensity of a note, phrase or section of music.

The following lists the most common dynamics from the softest to the loudest:

softest → *loudest*

ppp pp p mp mf f ff fff

Dynamics **softer than *ppp* or louder than *fff* are impractical,** although theoretically any dynamic is possible.

COMBINED DYNAMICS

Dynamics such as ***mf***, ***f***, ***sf*** can be combined with ***p*** or ***pp***. The resulting dynamics function as a somewhat broader form of articulation.

fp sfp sfpp

***fp** would mean to play loudly, then immediately play softly.*

(See *Articulations*)

CRESCENDO & DIMINUENDO

The **crescendo sign** is wedge-shaped and **opens toward the right.**

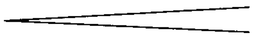

The **diminuendo sign** is wedge-shaped and **opens toward the left.** (The words *diminuendo* and *decrescendo* are interchangeable, but *diminuendo* is preferred.)

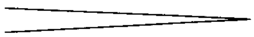

A crescendo or diminuendo sign is used when a **gradual increase or decrease** is wanted from one dynamic to another.

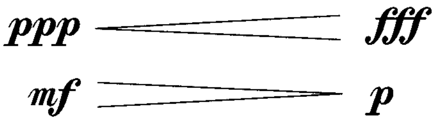

Musical elements should not be placed within the opening of the signs.

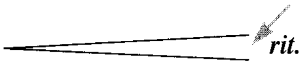

The word ***crescendo*** or ***diminuendo (decrescendo)*** may be written or abbreviated ***cresc., dim. (decresc.).***

The **effect is identical** whether using the sign, word or abbreviation.

If space doesn't allow the use of the sign, or the passage is a long one, it is best to use the word.

A **dashed line** may be used if it helps clarify the length of the passage.

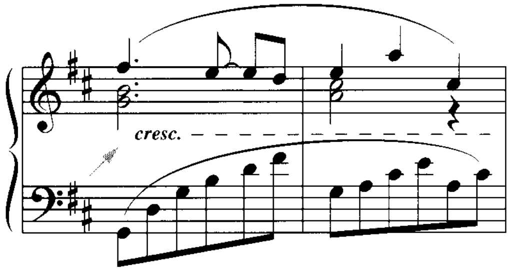

The word is sometimes **hyphenated** to extend throughout a section of music.

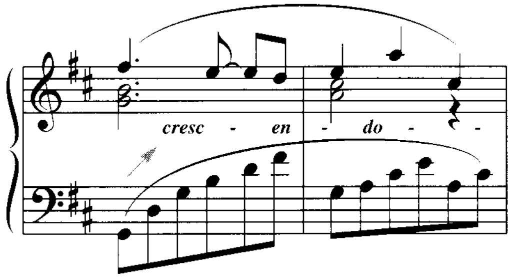

GENERAL PLACEMENT

Place the dynamic sign *slightly* before the notehead whenever possible.

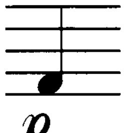

For single-staff instruments

Dynamics are placed **below the staff.**

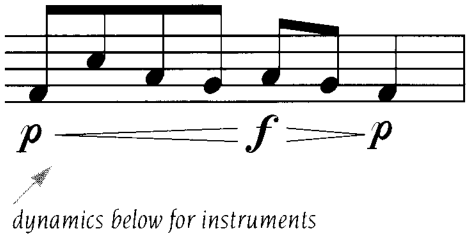
p
f
p
dynamics below for instruments

When two parts share a staff and the dynamics vary between parts, the dynamics will be placed above and below the staff.

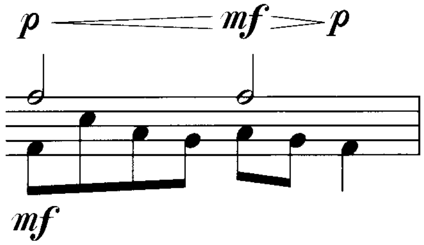

For vocal music

Dynamics are placed **above the staff** to avoid conflicting with the lyric.

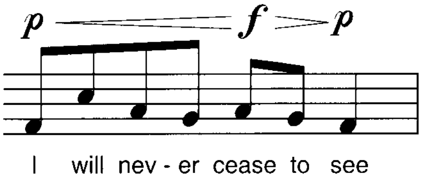
p
f
p
I will nev - er cease to see

For the grand staff

Dynamics are placed **between the two staves.**

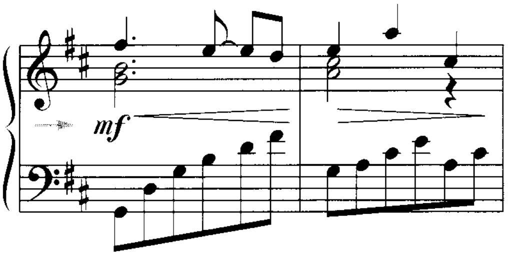

If two different dynamics are needed for each staff, the dynamics may be placed above for the top staff and below for the bottom, or in positions that clarify the situation best.

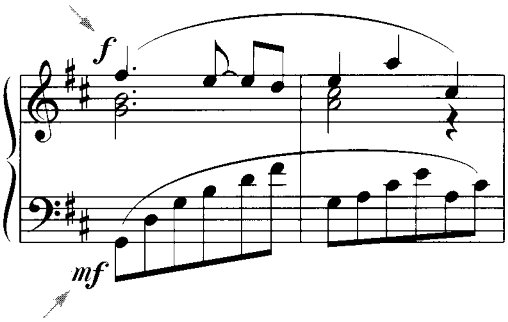

Horizontal placement is preferred for crescendo and diminuendo signs, although an angled placement is acceptable if necessary.

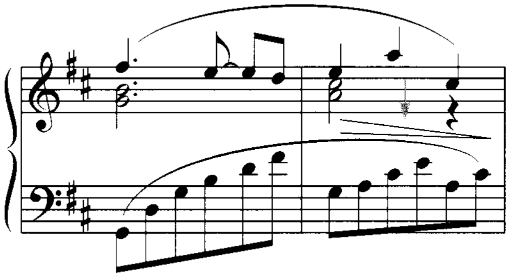

The signs may also be **placed outside the staff** if space is tight within the grand staff.

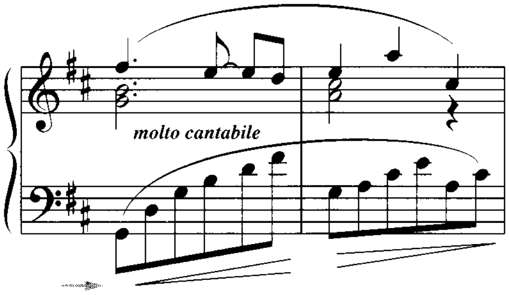

If a dynamic mark must be placed on a barline, the barline should be **broken.**

For crescendo and diminuendo signs, the **barline is left intact.**

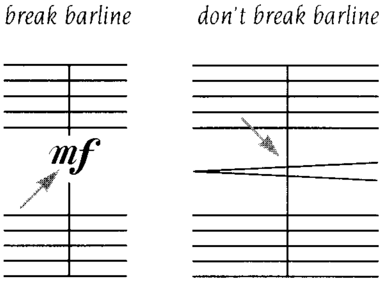

Dynamics with crescendo/ diminuendo signs

The intensity of the increase/decrease in volume is **relative to the dynamic level** of the section of music.

If a dynamic is not indicated at the end of a crescendo sign, and the music is consistently gentle, the increase in volume would probably be subtle.

The **level at the beginning of the sign** is the same as **the dynamic last indicated** in the music.

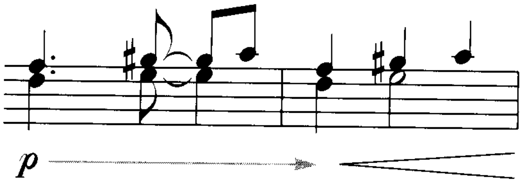

It is advised to **carefully begin and end the sign** so that it makes sense.

While the example below would make perfect sense in a clarinet part, it would make no sense in a piano part.

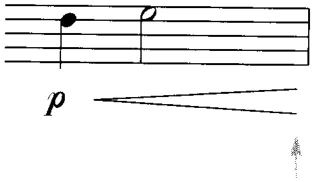

Endings

See *Repeat signs*

Fermata

A hold or pause sign (𝄐). It is used with any rhythmic value—single notes, intervals, chords or rests.

PLACEMENT

Preferred placement is above, regardless of stem direction, except when parts share a staff.

The fermata is **centered on the notehead** (or rest), whether it is on the notehead side or the stem side.

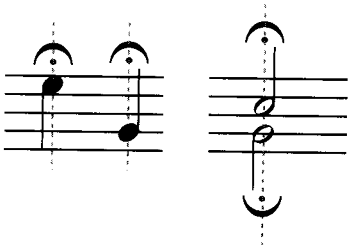

One fermata applies to all notes of an interval or chord.

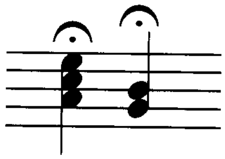

Place outside the staff if at all possible.

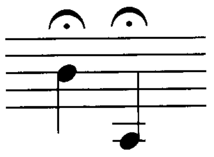

Place outside articulations and slurs.

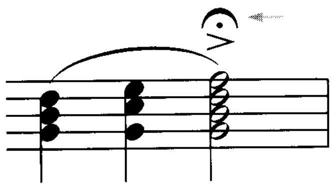

Place the fermata on the exact beat that is to be held, so that there can be no confusion. In the first example below, it might be wondered whether the fermata over the whole notes would apply to the eighth rest in the lower voice.

Confusing:

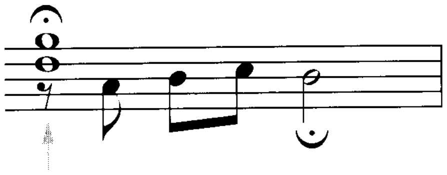

Same example, more clearly written:

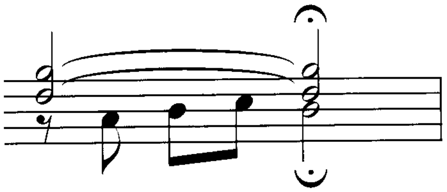

TIME VALUE

A fermata over a note or chord not only indicates that the tempo is interrupted but that the note or chord is **sustained.**

The length of a fermata is **determined by the performer** and is relative to the musical situation.

In a fast tempo, placed on a sixteenth note, the fermata would be a shorter value than if placed on a whole note, in a slow tempo.

USING WITH RESTS

In the following example, the fermata **prolongs silence rather than sound.**

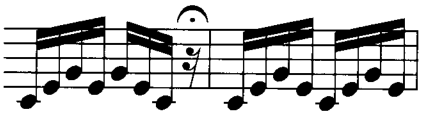

ABOVE A CAESURA

The following example is similar to the last one. The caesura indicates a small break in the sound; the addition of a *fermata* **lengthens the break.**

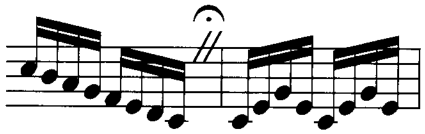

Occasionally a fermata is seen over a final double barline before another movement. However, since a barline has no time value, **avoid using a fermata over a single barline.**

ACCOMPANYING TERMS

Lunga *(long)* or **poco** *(little* or *short)* may be used with a *fermata* for a more specific intent.

These should be **centered above** the *fermata.*

Fingerings (for keyboard)

GENERAL PLACEMENT

Fingerings for the **right hand** should be placed **above** the staff, away from all musical elements.

Fingerings for the **left hand** should be placed **below** the staff, away from all musical elements.

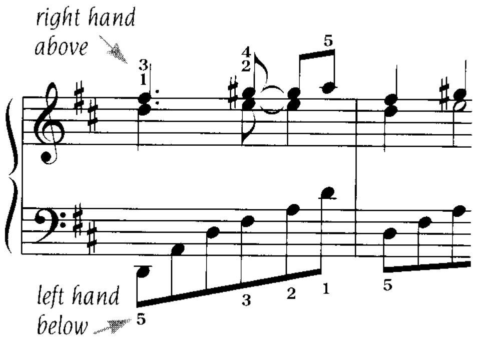

Fingerings are **centered** above or below the notehead, when possible.

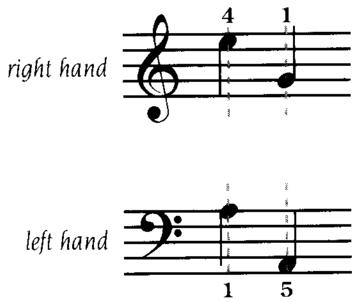

A finger number is preferably placed **outside a beam.**

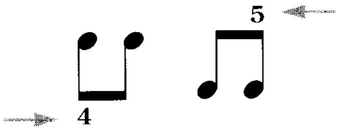

CHORDS

Fingerings are **stacked above or below** a chord; all notes should have a fingering.

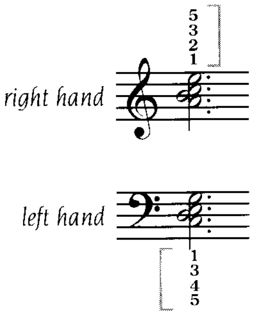

If a chord includes **notes that are tied** and a fingering is indicated for a note that is *not tied,* only that fingering is necessary.

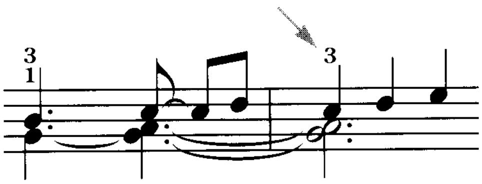

ALTERNATING FINGERINGS

If a fingering changes on a note that is held, a hyphen (-) or an "en-dash" (–) **separates the two fingerings.** The first fingering is centered on the note; the second is to the right.

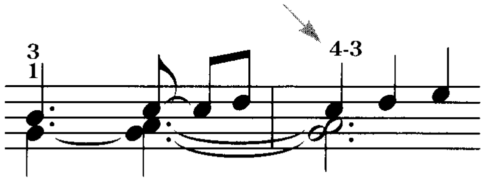

An alternate method uses a small slur instead of the hyphen or en-dash.

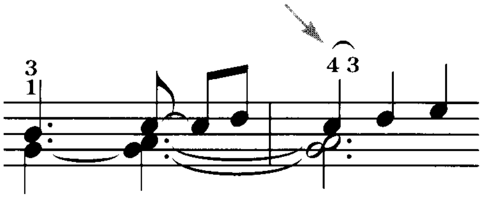

USE WITH TRIPLETS

The triplet ***3*** is sometimes mistaken for a fingering when the fonts are similar.

The **fingering font should be smaller** and of a **different typeface** than the font used for triplets.

Avoid positioning the triplet *3* on the same side of the staff as a fingering.

Instead of moving the fingering, **move the triplet *3*** to the opposite side from the fingering. The triplet ***3*** would normally be on the *beam* side; **if on the notehead side** it is helpful to **add a bracket.** (See *Tuplets*)

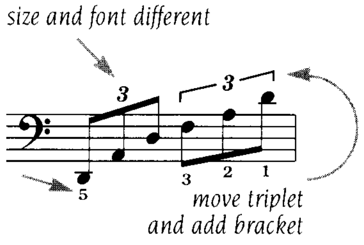
size and font different

move triplet and add bracket

TWO HANDS ON ONE STAFF

If both hands share the same staff, the fingering for the **right hand is above,** the **left hand below.**

RH (right hand) or **LH** (left hand) may be added for **clarification.**

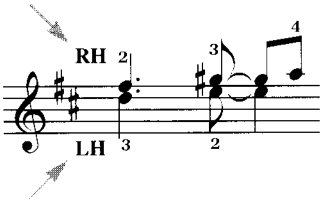

AMOUNT OF FINGERING

Fingering should be **used sparingly** and only to establish a pattern or hand position.

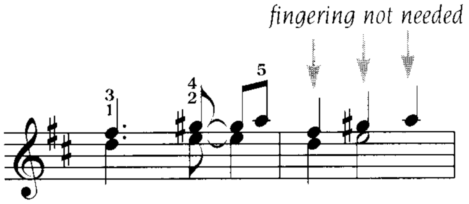

OVERLAPPING ELEMENTS

Fingerings **may overlap a slur.** The preference is to break the slur at the point of intersection, if possible.

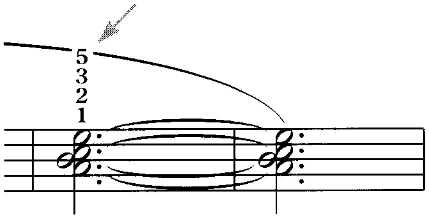

The vertical space sometimes dictates that fingerings be placed within the staff. The **staff lines should be broken** to accommodate the overlapping fingerings.

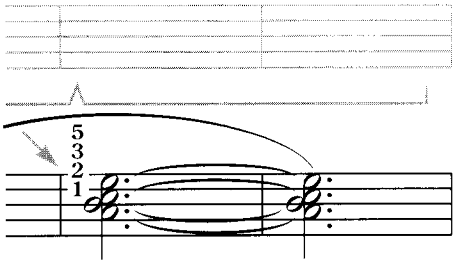

Flat

See *Accidentals, Key signatures*

Glissando sign

A glissando is indicated by a wavy or straight line placed at an angle, ascending or descending.

Indicate the beginning and ending notes of the glissando.

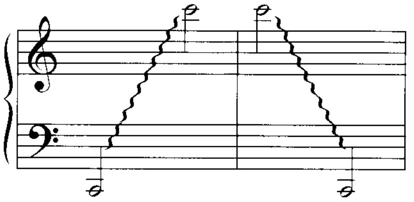

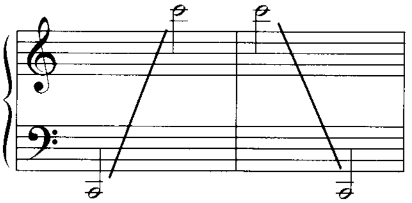

ADD THE ABBREVIATION

It is helpful to **add the abbreviation** (*gliss.*)**,** placing it at the same angle as the symbol.

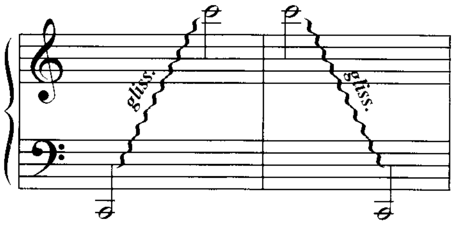

TWO OR MORE NOTES

If the glissando affects more than one note, add a **glissando symbol for each note.** The word (*gliss.*) is used only once regardless of the number of notes.

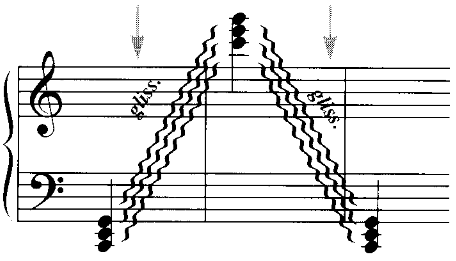

Grace notes

Grace notes are small notes **without a rhythmic value** of their own, taking their value from the previous or following beat.

Grace notes are notated at **cue size** or slightly smaller (65% of normal size works well).

SINGLE GRACE NOTES

Single grace notes are **considered eighth notes.**

Place the **stem up** regardless of the main note's stem direction (except when two parts share a staff).

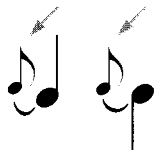

Usually a **small slur connects** the grace note to the main note; the slur usually begins below the grace note.

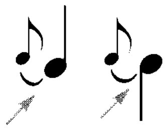

The slur **end does not need to be centered** on the main notehead like slurs normally do. It may end **slightly to the left** of the notehead. (See *Slurs*)

For single grace notes with a flag, a **small line** can intersect the flag, always **slanting upward, from left to right,** regardless of stem direction.

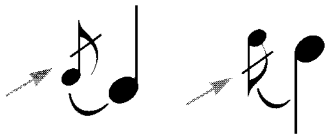

(for a lower voice sharing a staff)

PLACEMENT ON THE STAFF

The grace note is centered in a space or placed directly on a line.

Grace-note **accidentals** are also cue size. (See *Accidentals*)

Stems for grace notes on leger lines do not have to be lengthened to meet the middle line.

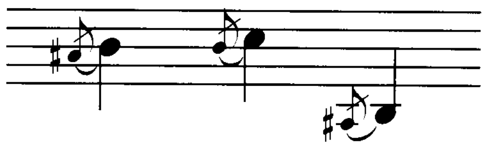

SHARING A STAFF

When two parts share a staff, the grace note stems follow the stem direction of the parts.

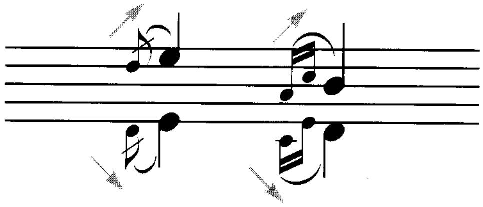

MULTIPLE GRACE NOTES

Multiple grace notes should **never have a slanted line** through the stems or beams. The slanted line is only for single, flagged grace notes.

If a slur is used, the **slur connects** all the grace notes **to the main note.**

Two grace notes are **considered sixteenth notes** and should be beamed accordingly. Three and four can be either sixteenth or 32nd notes.

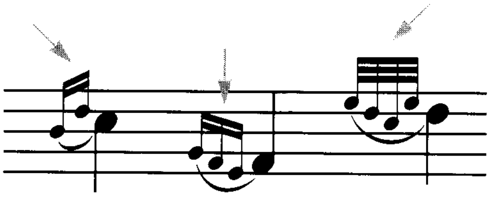

Larger groupings, such as five and six grace notes, are usually beamed as 32nd notes.

DIRECTION OF GRACE-NOTE SLURS

If a grace note precedes an interval or chord, the **slur will follow the direction of resolution** (from the grace note to the appropriate main note).

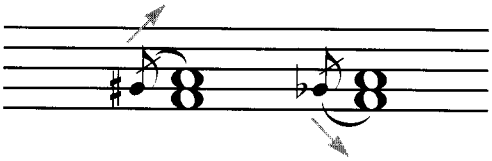

WITH LEGER LINES

If leger lines are used, the slur may be **placed above to avoid the leger lines.**

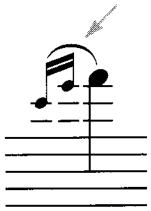

The grace-note leger lines can be the same thickness as staff lines, but shorter.

Grand pause

The grand pause or general pause (G.P.) over a measure of rest indicates that the **entire orchestra is silent,** often unexpectedly.

Grand staff

The grand staff is used to notate scores for instruments such as piano, organ, harp and marimba.

The grand staff consists of **two staves joined by a brace** and a systemic barline. (See *Brace*)

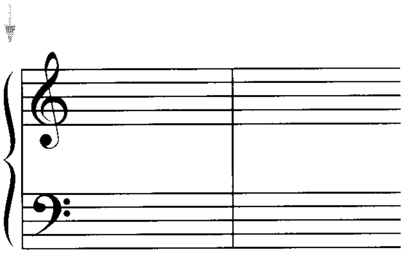

Although the "normal" clefs for the grand staff are treble above and bass below, **either staff may have either clef,** and may change clefs at any time. (See *Clef signs*)

FOR ORGAN

Organ scores use the grand staff connected to a staff used to notate music for the organ pedals.

Except for the systemic barline, **barlines are separated** between the manuals and pedal.

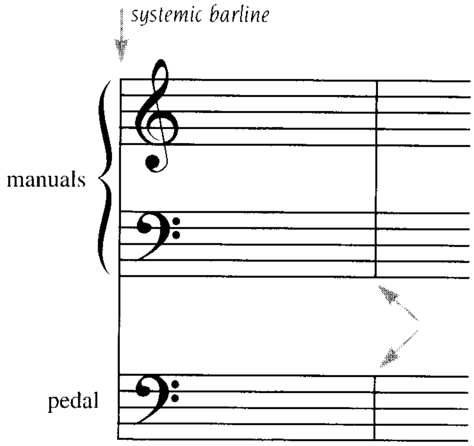

FOR INSTRUMENT & PIANO

Scores for an instrument (including voice) and piano use a **grand staff, with the instrument staff above.**

The staff above the grand staff is sometimes **cue size,** if the instrument has its own full-size, separate part.

Except for the systemic barline, **barlines are broken** between the instrument staff and the grand staff.

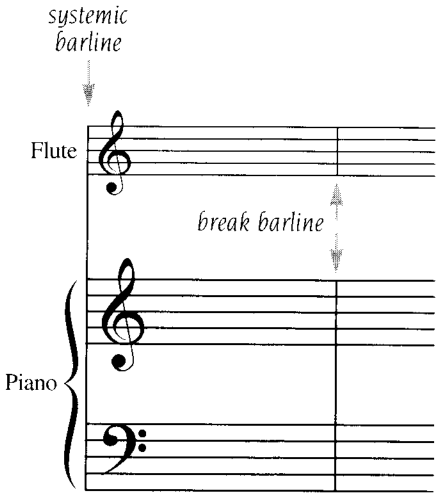

TWO-PIANO MUSIC

Two-piano music requires two grand staves joined together with a **systemic barline.**

The measure **barlines connect staves for each grand staff only.**

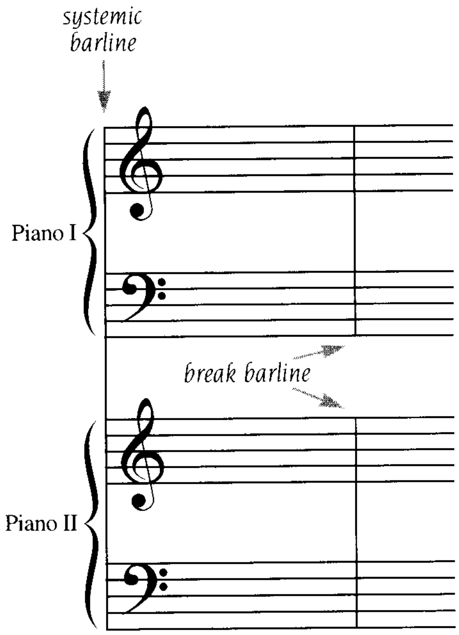
systemic barline
Piano I
break barline
Piano II

ONE PIANO—FOUR HANDS

Facing pages are required.

The **Secundo** part is notated on the left-hand page; usually both staves are in **bass clef.**

The **Primo** part is notated on the right-hand page; usually both staves are in **treble clef.**

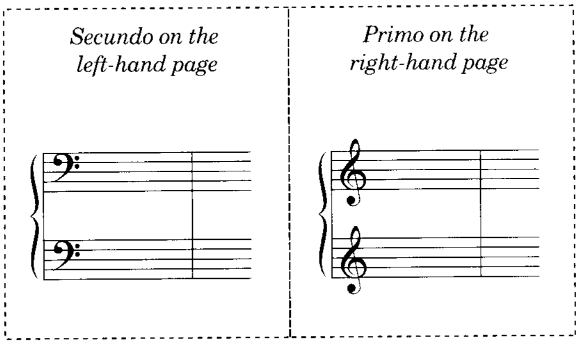
Secundo on the left-hand page
Primo on the right-hand page

The **page layout** for primo and secondo **is preferably identical**, system-for-system; if they are not identical, the facing pages *must* begin and end with the same measures.

Harmonics

NATURAL HARMONICS

Natural harmonics are indicated by a **small circle** centered above the notehead.

The indicated note is the SOUNDING NOTE.

Such notation indicates a **natural harmonic.**

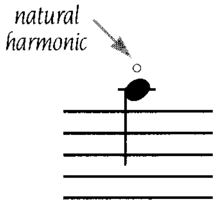
natural harmonic

Careful consideration of an instrument's capabilities is necessary. Only some instruments are capable of producing harmonics.

Another way to notate a natural harmonic (for strings) is to use a **diamond-shaped notehead.**

The position on the staff **indicates where the finger is placed**—NOT the sounding note.

The sounding note may be indicated in cue size and in parentheses.

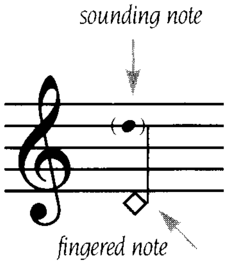
sounding note
fingered note

ARTIFICIAL HARMONICS (STRINGS)

A diamond-shaped notehead with a standard notehead comprises an artificial harmonic. The **standard notehead indicates the placed finger**; the diamond indicates where the second finger lightly touches the node to produce the harmonic.

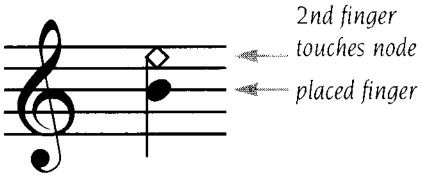
2nd finger touches node
placed finger

Often the sounding note is indicated in cue size and in parentheses.

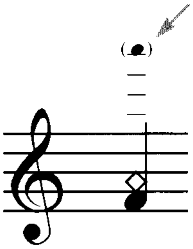

Key signatures

Key signatures appear **after the clef but before the time signature.**

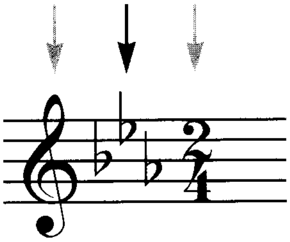

CHANGING KEYS

The preference is to make key changes at the beginning of a staff or system, with a **courtesy key signature** at the end of the previous one.

1. A key change is traditionally preceded by a **double barline.**
2. If the key change is at the beginning of a staff or system, a **courtesy key signature** is placed at the end of the previous one.
3. The **staff is left open** after the courtesy key signature.
4. The **new key signature is indicated** on the next staff.

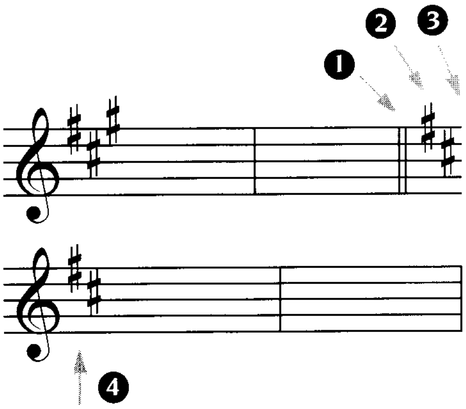

If a key **changes mid-staff:**

1. a double barline is added,
2. followed by the new key signature.
3. The new key signature continues on the next staff.

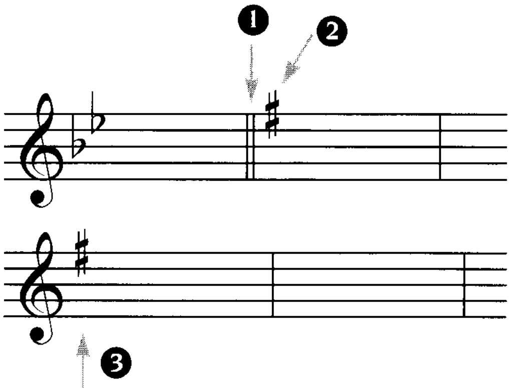

CANCELLATIONS

Cancellations are no longer considered necessary, unless the new key is C major or A minor (no sharps or flats). In that case, cancel the old sharps or flats with natural signs, in the same order as the old key signature .

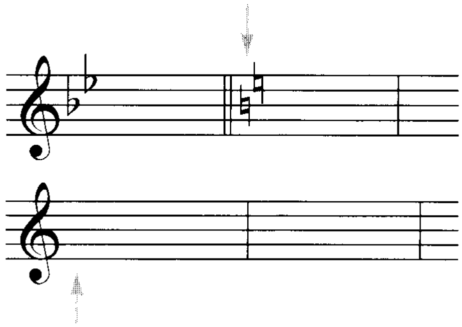

The following illustrates the positioning of the clef, key signature and time signature if all three change simultaneously.

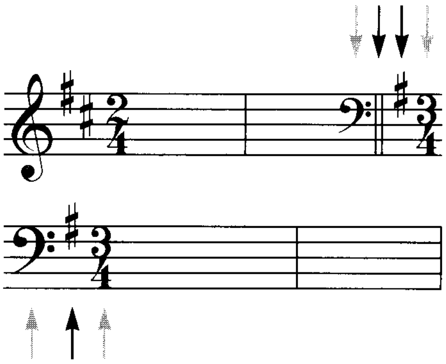

Key signatures in the **TREBLE CLEF**

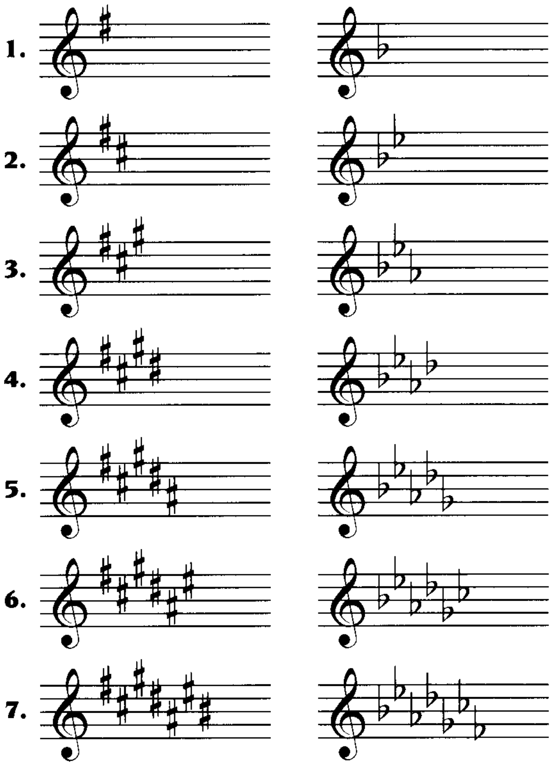

Key signatures in the **TENOR CLEF**

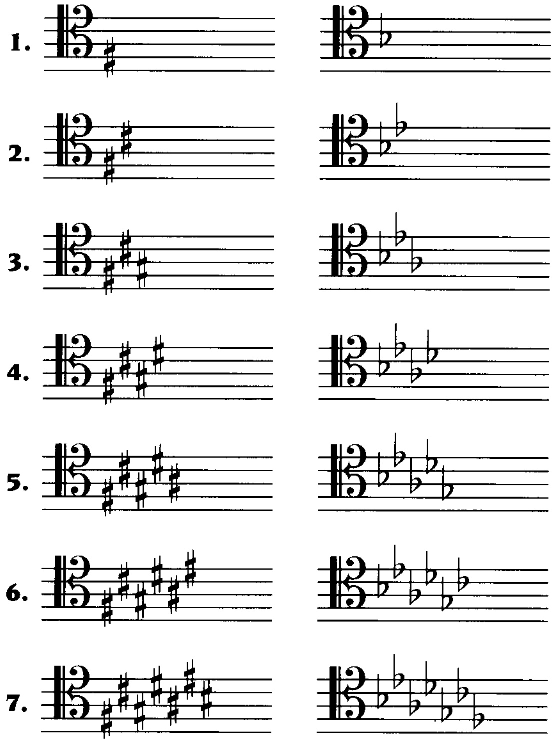

Key signatures in the **BASS CLEF**

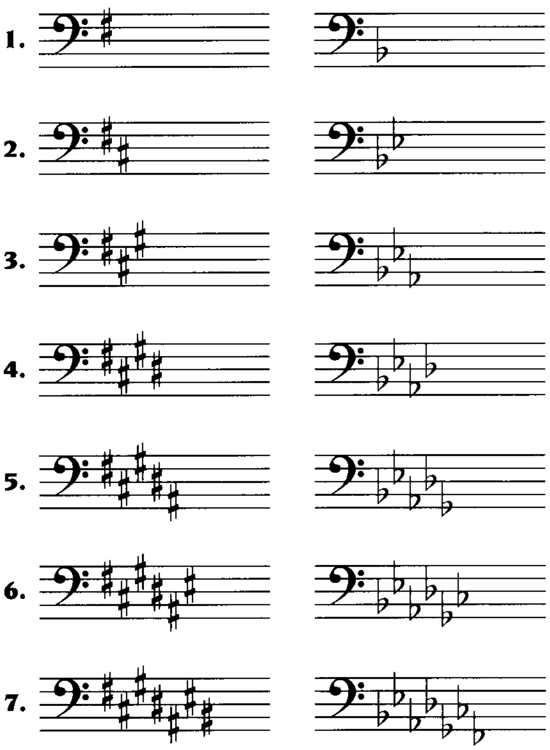

Key signatures in the **ALTO CLEF**

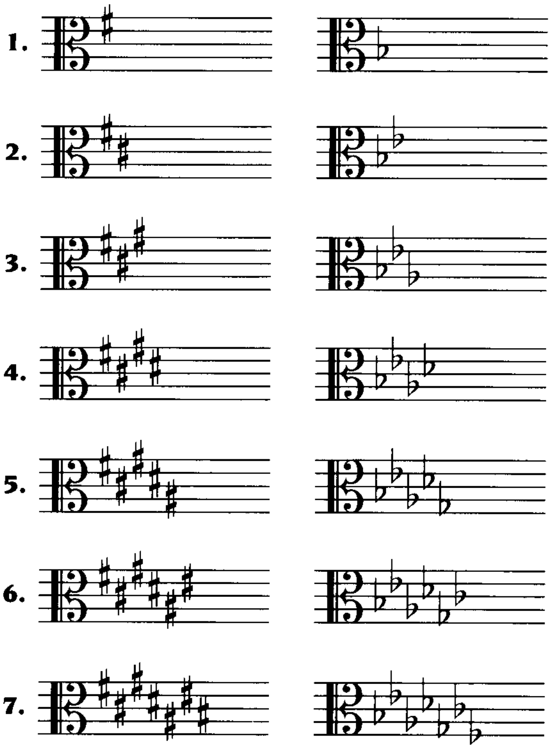

THE CIRCLE OF FIFTHS

The Circle of Fifths serves as a quick reference guide to the relationship of the keys and key signatures, and how they can be figured out in a logical manner.

The Circle of Fifths will also help to clarify which keys are enharmonic equivalents. (*Enharmonic* means notes that sound the same but are written differently.)

Clockwise movement *(up a perfect 5th) provides all of the sharp keys by adding one sharp to the key signature progressively.*

Counter-clockwise *(down a perfect 5th) provides the flat keys by adding one flat similarly.*

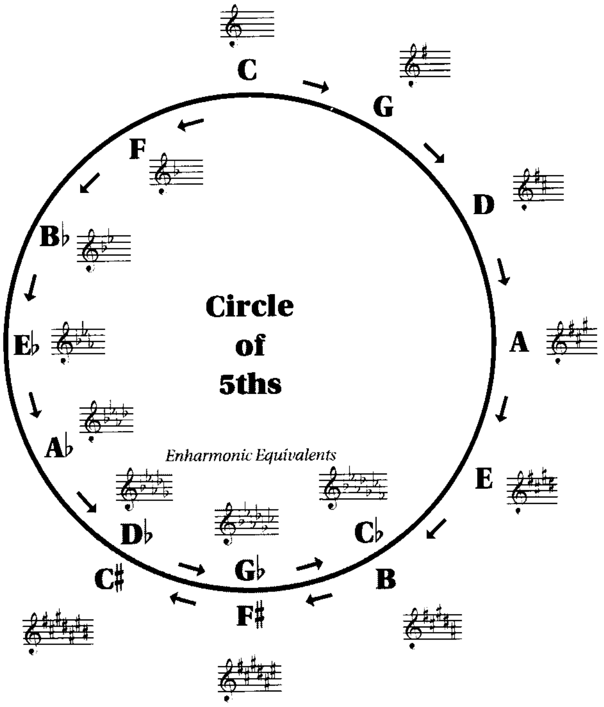

Layout

See *Spacing*

Leger lines

Leger lines are used to extend the range of the five-line staff.

Leger lines are the **same line weight as staff lines,** or slightly heavier.

The **vertical spacing** of leger lines must be identical to the vertical spacing between staff lines.

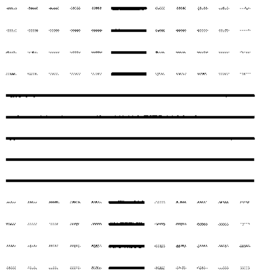

Leger lines **extend slightly past the notehead.** They will need to extend sufficiently enough to be seen.

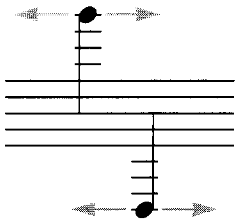

STEM LENGTH

Stems of all notes ABOVE or BELOW the first leger line past the staff **must extend at least to the middle staff line.**

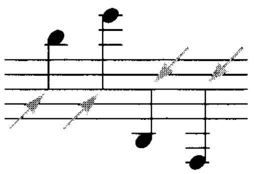

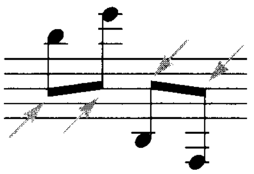

OVERCROWDING

To avoid leger lines that touch, adjustments must be made either in the **horizontal spacing** of the music or by **shortening the leger line length.** The first is preferred, but is not always possible.

In the first example below, the leger lines are overcrowded and appear to be touching.

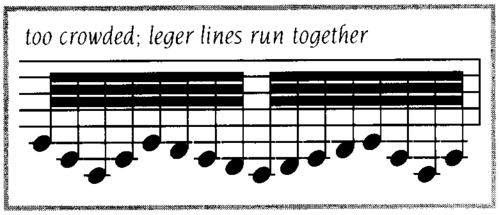

Corrected by shortening the leger lines to create a slight amount of space between the lines.

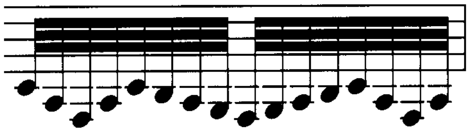

CLEF CONSIDERATIONS

Choose the appropriate clef for the situation. The viola, for instance, may be written in alto clef to avoid excessive leger lines. (See *Clefs*)

The following examples illustrate the same musical example in three different clefs.

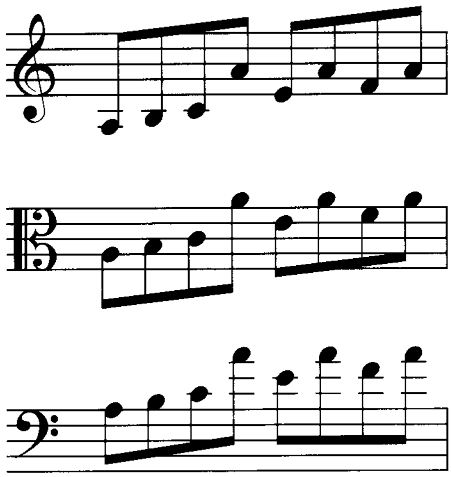

Lyric

Beaming of notes associated with a lyric now follows standard notational practice.

Traditional practice, now obsolete, was to use flags for eighth notes, sixteenth notes, etc. Beams were used only for melismas.

However, in several other ways the lyric makes the notation of a vocal part different from that of other parts.

PLACEMENT OF STAFF INDICATIONS

Dynamics and tempo indications go **above the staff** to avoid conflicting with the lyric.

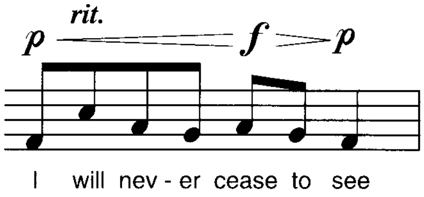

USE OF SLURS

Slurs are **seldom used in a vocal part for anything other than a *melisma*** (a term used here simply to mean a syllable or word that is sung to more than one note).

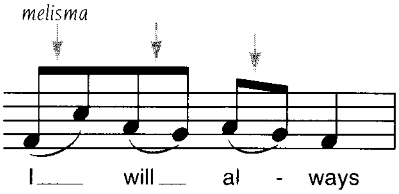

ALIGNMENT OF LYRIC FOR MELISMA

Another help in reading a melisma is the proper **alignment of the lyric.** The word or syllable, instead of being centered under the note, should be aligned **flush left with the left edge of the notehead** (see above).

EXTENDER LINES

In addition to the slur, **extender lines at the baseline of the lyric** help to **show the length of a melisma.** However, extender lines are only used for **one-syllable words, or for the last syllable of a word.** A hyphen is used in the middle of a word.

The **thickness of extender lines** should be less than that of the staff lines.

The **end of the extender line** should be **aligned flush right** with the last note of the melisma. *The most common mistake in the use of lyric extender lines is to extend the line for the full value of the note.*

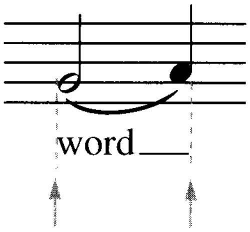

When an extender line is called for, there should always be one, even if the words are crowded and it has to be very short. (In this case, the line can extend slightly past the note.)

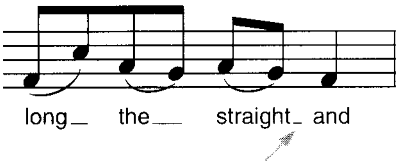

When an extender line wraps to the next staff, begin the line at the end of the key signature.

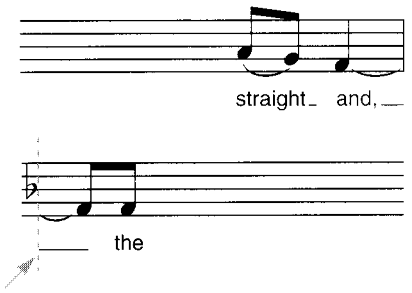

Notice the comma in the previous example. Any **punctuation always precedes an extender line.**

TIES

Ties are treated exactly like melismas, including the use of extender lines and alignment of lyric.

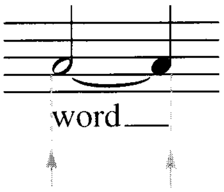

WORD PLACEMENT (NORMAL)

Words or syllables for notes other than melismas or tied notes are **centered to the notehead.** If the lack of space does not allow a word or syllable to be exactly centered, at least some part of the word or syllable should be under the note.

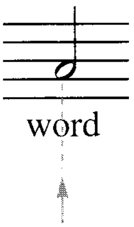

HYPHENATED WORDS

Center the *hyphen* (not a dash or extender line) between syllables.

If the melisma (or tie) is extremely long, **more than one hyphen** may be used.

If the hyphenated word **wraps to the next line, another hyphen *may* be used** at the beginning of the line.

Hyphenate words **according to the dictionary.** Do not hyphenate according to how the word *sounds* like it should divide. If there is any question at all, check the dictionary.

PUNCTUATION

Use **normal punctuation and capitalization.** (Sometimes capitalization is done as if the words were poetry, with each "line" capped, but this is becoming less common.) *A common mistake is to use little or no punctuation.*

MULTIPLE VERSES & REPEATS

Broken slurs

A **broken slur** is used when one verse has a melisma, especially if it is not the first verse.

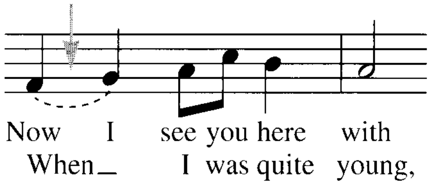

Treat the lyric for each verse separately. Notice in the above example that the melisma is aligned *as a melisma,* while the previous verse is aligned normally.

Omission lines

An omission line is sometimes used to clarify that there is **no word or syllable sung.**

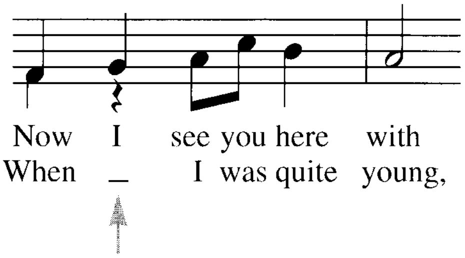

In many cases, however, there is a better way to treat the situation. In the previous example, the omission line could be mistaken for an extender line, so it would be better without it, letting the cue-size rest serve the same purpose.

Cue-size notes, rests

Use cue-size notes and rests **when a rhythm is different from the first verse** (see previous example). Stems and beams should also be cue size.

If there are **more than two verses** with different rhythms, the notation can become quite complicated. It may be better to write out the music.

Numbering

Multiple verses are numbered. There are several ways to treat these when there are pick-ups, repeats, first and second endings, etc. Be clear and consistent.

→ 1. Verse, verse, verse
→ 2. Verse, verse, verse
→ 3. Verse, verse, verse

Using a brace

When multiple lines converge into one line, such as a *chorus* or *refrain*, a **brace** is used.

→ 1. Verse, verse, verse
→ 2. Verse, verse, verse } chorus or refrain →
→ 3. Verse, verse, verse

Meter

The understanding of meter is essential to the placement of ties, rests, syncopated rhythms and proper beaming.

Time signatures identify the meter.

There are two overall classifications of meter: **perfect** and **imperfect** (odd) meter.

PERFECT METER

In perfect meter the measure can be divided into **equal halves or thirds**.

There are two classifications of perfect meter: **simple meter** and **compound meter.**

In **simple** meters each **BEAT** can be subdivided by 2.

In **compound** meters each **PULSE** can be subdivided by 3.

Metric subdivisions should not be confused with tuplets.

SIMPLE METER

The classification of **duple, triple** or **quadruple** reflects the number of **BEATS** in the measure.

Simple duple meter

Simple duple time signatures have a 2 as the upper number of the time signature.

$\frac{2}{2}$ $\frac{2}{4}$ $\frac{2}{8}$

Each measure contains **2 BEATS.**

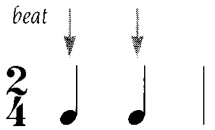

Each BEAT can be **subdivided by 2**.

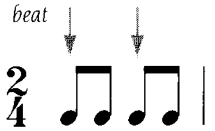

Only **rests of equal value to the beat** or less may be used (except in a complete **measure of rest,** which is indicated by a whole rest).

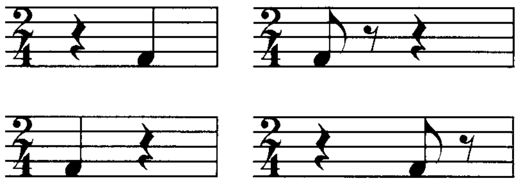

Rests of equal value to the beat are used ONLY on the beat.

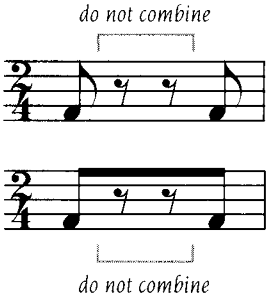

The following are all acceptable examples in simple duple meter.

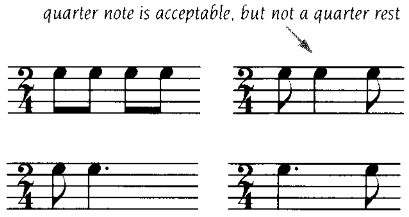

Simple triple meter

Simple triple time signatures have a 3 as the upper number of the time signature.

$\frac{3}{2}\quad\frac{3}{4}\quad\frac{3}{8}$

Each measure contains **3 BEATS.**

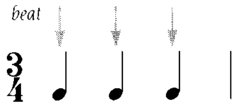

Each beat can be **subdivided by 2.**

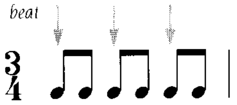

Only **rests of equal value to the beat** or less may be used (except in a complete **measure of rest,** which is indicated by a whole rest).

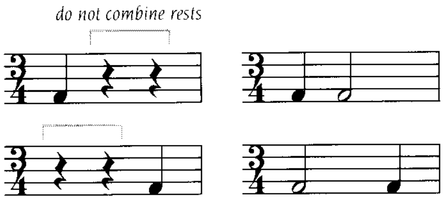

Beaming should reflect the simple triple meter, NOT compound duple meter.

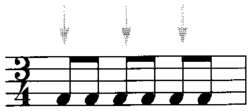

NOT

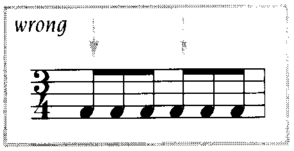

Simple quadruple meter

Simple quadruple time signatures have a 4 as the upper number of the time signature.

$\frac{4}{2}\quad\frac{4}{4}\quad\frac{4}{8}$

Each measure contains **4 BEATS.**
Each beat can be **subdivided by 2.**

To successfully notate *simple quadruple* meter, it is best to consider the meter as a combination of **2 units of duple meter.**

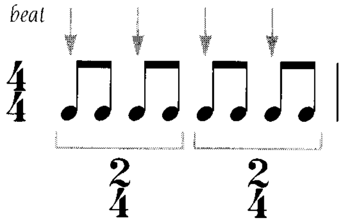

*In simple quadruple meter, these **2 units** must always be clearly distinguished. The only exception to this would be a very simple syncopation (such as ♩𝅗𝅥 ♩).*

The following examples illustrate the **subdivision of the 2 units.** The notation of each unit, separately, is identical to that of simple duple meter.

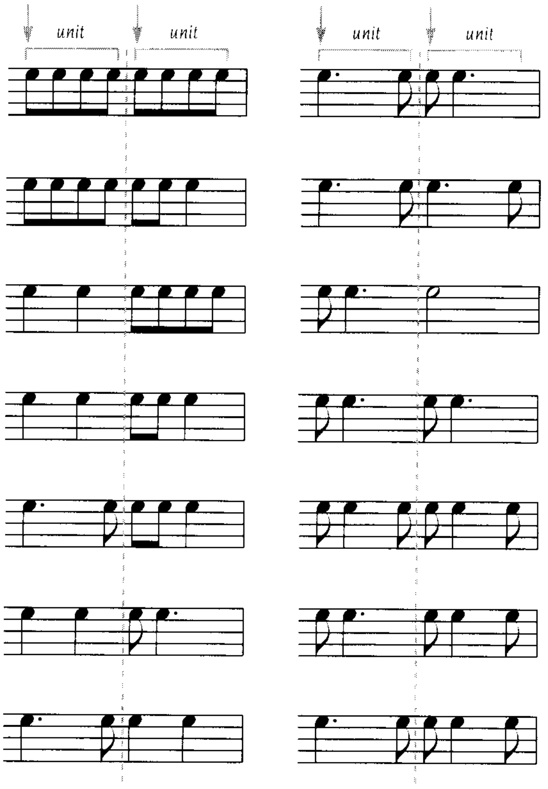

Whole rests are used to indicate a complete measure of rest.

Rests twice the value of the beat are used ONLY on beats 1 & 3 (the first beat of each half of the measure).

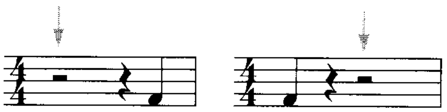

Rests of equal value to the beat can ONLY appear on the beat.

Rests less than the value of the beat are used freely as long as the two units are apparent.

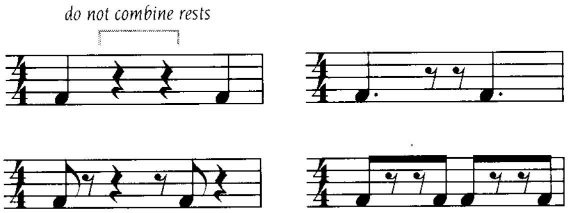

COMPOUND METER

The classification of **duple, triple** or **quadruple** reflects the number of **PULSES** in the measure.

Compound duple meter

Compound duple time signatures have a 6 as the upper number of the time signature.

$\genfrac{}{}{0pt}{}{6}{4}$ $\genfrac{}{}{0pt}{}{6}{8}$

Each measure contains **2 PULSES.**

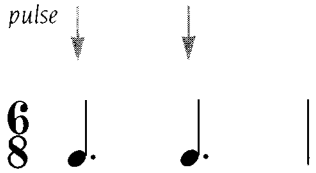

Each PULSE can be **divided into 3 BEATS.**

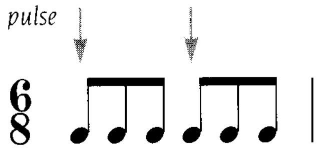

Only **rests of equal value to the pulse** or less may be used (except in a complete **measure of rest** which is indicated by a whole rest).

Rests of equal value to the pulse can ONLY appear on the pulse.

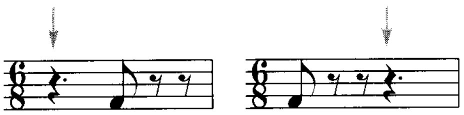

Rests of $2/3$ the value of a pulse may be used ONLY on the pulse.

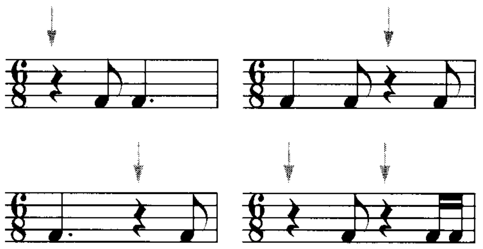

Rests of equal value to the beat, **not appearing on the pulse,** CANNOT be combined.

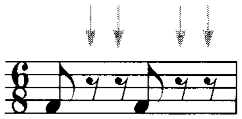

Beaming should reflect the compound duple meter, NOT simple triple meter.

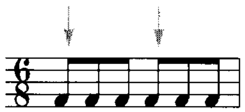

NOT

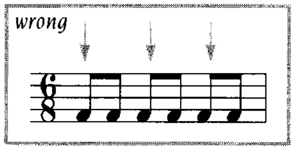

Compound triple meter

Compound triple time signatures have a 9 as the upper number of the time signature. An example is $\genfrac{}{}{0pt}{}{9}{8}$.

Each measure contains **3 PULSES.**

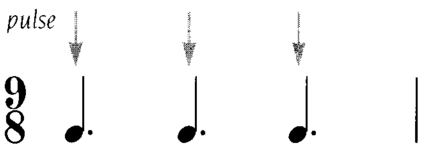

Each PULSE can be **divided into 3 BEATS.**

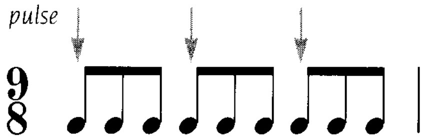

Rules governing **rests and beaming** are identical to compound duple meter.

Compound quadruple meter

Compound quadruple time signatures have a 12 as the upper number of the time signature. An example is $\frac{12}{8}$.

Each measure contains **4 PULSES.**

Each PULSE can be **divided into 3 BEATS.**

To successfully notate *compound quadruple* meter it is best to consider the meter as a combination of **2 units of compound duple meter.**

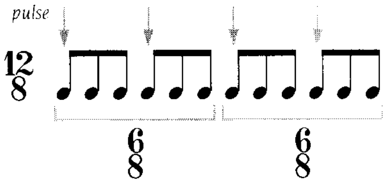
pulse
12/8
6/8
6/8

Rules governing **rests and beaming** are identical to compound duple meter.

IMPERFECT METER (ODD METER)

In imperfect meter (odd meter) the measure **CANNOT be divided into equal halves or thirds.**

Imperfect meter can be considered as any combination of **duple or quadruple meter** AND **triple meter.**

$\frac{7}{8}$ time can be considered a combination of $\frac{2}{4} + \frac{3}{8}$ or $\frac{3}{8} + \frac{2}{4}$.

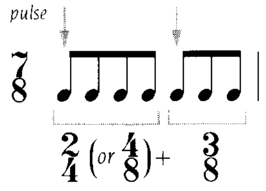
pulse
7/8
2/4 (or 4/8)+
3/8

Notice that 7 is not divisible by 2 or 3.

MODERN APPROACHES TO METER

A **broken barline** is sometimes added to clarify imperfect meters.

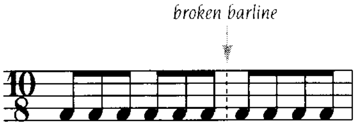
broken barline
10/8

A **composite meter** is another option to clarify the subdivision of imperfect meters.

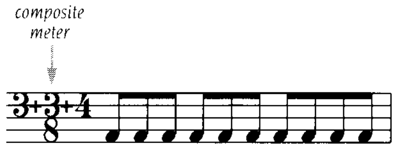
composite meter
3+3+4
8

When two or more measures alternate meters in a manner that repeats predictably, a time signature reflecting the **alternating meter** may be used.

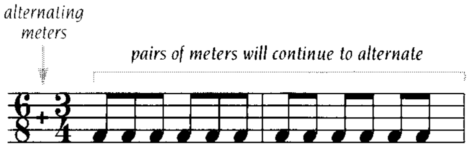
alternating meters
pairs of meters will continue to alternate
6/8 + 3/4

Metronome marks See *Tempo marks*

Natural sign See *Accidentals*

Notes

Traditionally there are five main shapes representing noteheads:

Pitched	Non-pitched
𝅝 *whole note*	◇ *whole note, half note*
𝅗 *half note*	
● *quarter, eighth, sixteenth, 32nd, etc.*	X *quarter, eighth, sixteenth, 32nd, etc.*

PLACEMENT ON THE STAFF

On a **five-line staff,** notes are either in a space or on a line.

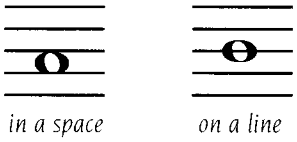
in a space
on a line

On a **single-line staff,** the notehead usually intersects the line, with the stem always up.

When two instruments share a staff, one part sits on the line; the other hangs from the line.

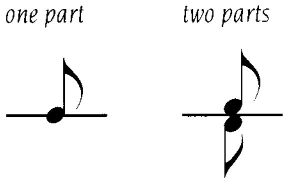

STEMS AND NOTEHEADS

Double whole notes (or *breves*) and **whole notes** are without stems.

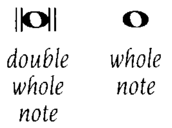

Half notes and **quarter notes** are always stemmed.

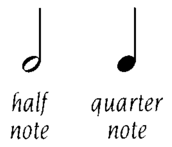

Single **eighth, sixteenth, 32nd and 64th notes** are always stemmed and with a flag(s).

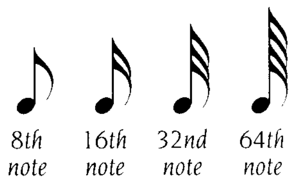

Two or more eighth, sixteenth, 32nd and 64th notes are usually **beamed into groups.**

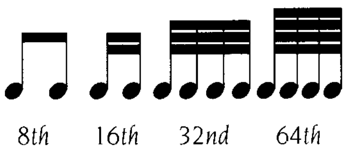

An **X** notehead is used primarily in **non-pitched percussion** music. It is also seen in vocal music to indicate approximate-pitched singing *(Sprechstimme),* rhythmically spoken parts, clapping, etc.

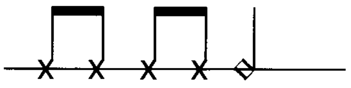

The diamond-shaped notehead serves as a half note (with stem) or whole note (without stem).

When used as a **harmonic,** the diamond-shaped **notehead is always open,** whatever the rhythmic value.

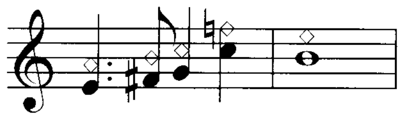

In some software programs the characters for the diamond-shaped notehead and the harmonic are slightly different.

Octave signs

The use of octave signs is primarily to **avoid multiple leger lines.** While frequently used in piano music, octave signs are only selectively used in instrumental music.

8va or *all'ottava* means "at the octave."

8va (or sometimes ***8***) is indicated above a **treble clef staff only.** Do not use above any other clef.

Usually accompanied by an extender line—**a broken line with a downstroke** at the end.

The alignment of the ***8va*** is centered above the notehead.

The end of the ***8va*** bracket **extends slightly past the last note** affected.

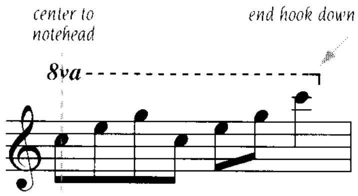

8va bassa or *ottava bassa* means "at the octave below."

__8va__ (or 8) below the staff, especially in keyboard music, is always understood to mean __8va bassa.__

8va bassa is indicated below a **bass clef staff only.** Do not use for treble, alto or tenor clefs.

Usually accompanied by an extender line—**a broken line with an upstroke** at the end.

The beginning of the ***8va bassa*** is placed below the notehead.

The end of the ***8va bassa*** line **extends slightly past the last note** affected.

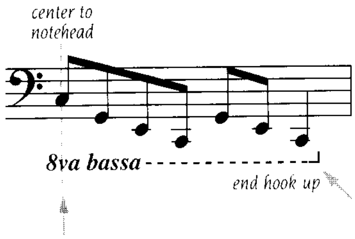

The abbreviation ***8vb*** is *only* a copyist's shorthand and **should not be used in engraved music.**

PLACEMENT CONSIDERATIONS

The ***8va*** or ***8va bassa*** with its extender line should be **clearly placed,** to avoid conflict with as many musical elements as possible and yet be as close to the notes affected as possible.

In the example below, the 8va is placed entirely above the slur.

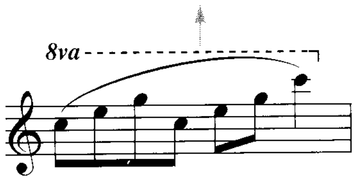

In this example, the 8va intersects the slur due to a possible limitation of space.

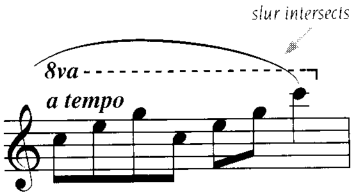

WITH ACCIDENTALS

Accidentals must be repeated for the notes included within ***8va*** or ***8va bassa*** octave alterations.

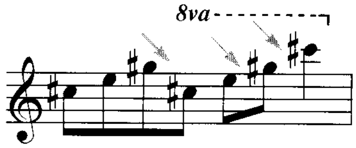

WRAPPING FROM STAFF TO STAFF

The ***8va*** or ***8va bassa*** **breaks at the end of the first staff,** aligned with the right barline.

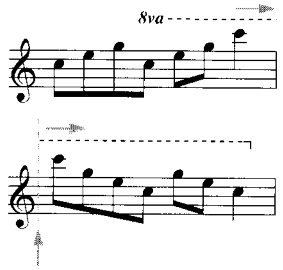

The extender line **begins again just past the clef** or key signature.

A **courtesy *8va*** or ***8va bassa*** may also be indicated at the beginning of the second staff when wrapping from staff to staff.

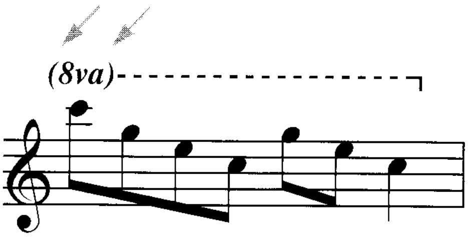

LOCO

Although the ending of the extender line should suffice, the word ***loco*** (meaning "at place") may be added as a reminder that the ***8va*** or ***8va bassa*** no longer applies.

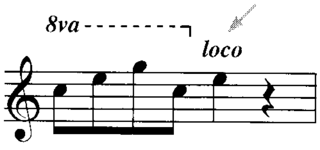

ON AN ENDING NOTE

The *8va* or *8va bassa* **is centered to the notehead** and, since it is the last note of the piece, it may or may not be accompanied with a downstroke (or upstroke).

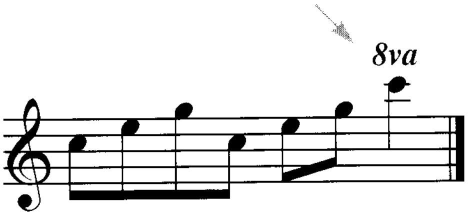

If a single note is altered but is **not the last note** of the piece, an **upstroke (or downstroke) *must* be added.**

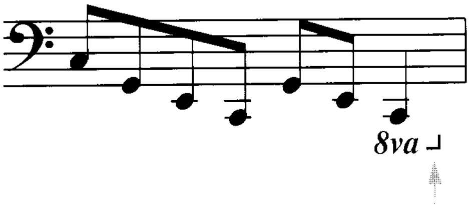

SPECIAL SITUATIONS

The following examples are incorrectly notated. The *8va* or *8va bassa* is **never placed between the two staves** of a grand staff.

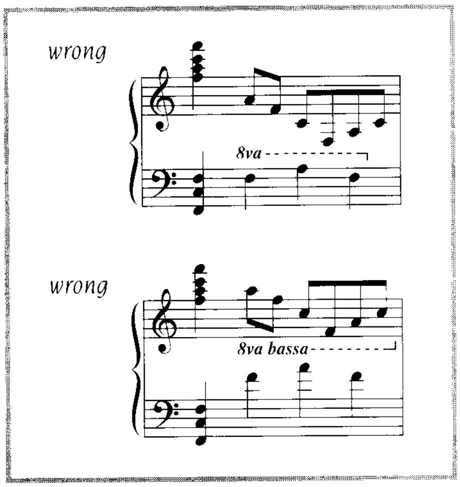

These two incorrect examples would need to be rewritten—perhaps in one of the following ways:

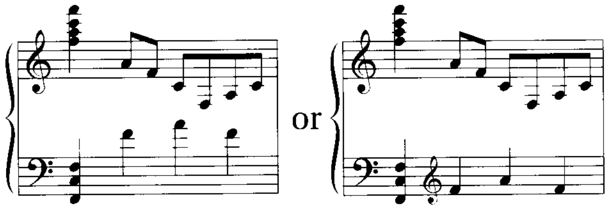

WITH RESTS

If the *8va* or *8va bassa* is interrupted with short rests, the **extender line may continue over** them.

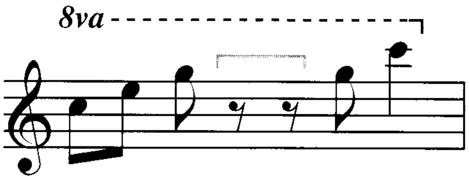

If the rests are of a **larger value, the line should be broken** and reintroduced.

When **rests precede** the *8va* or *8va bassa,* begin on the first note after the rests.

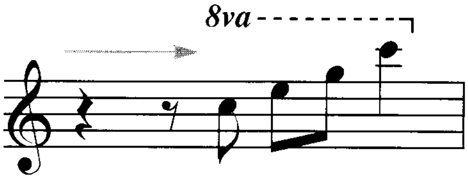

When **rests follow** the *8va* or *8va bassa,* the octave alteration ends with the last note in the grouping—the rests should not be included in the grouping.

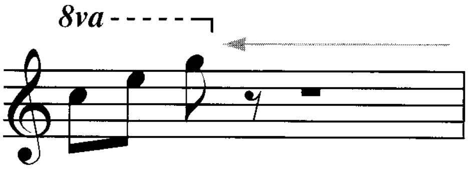

CUE SIZE

For any passage at **cue size** the *8va* will likewise be at cue size.

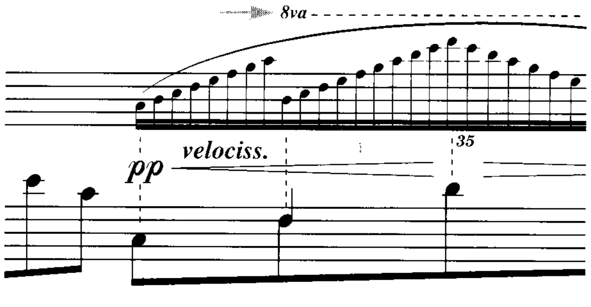

15MA

15ma or *quindicesima* means "at the fifteenth" (two octaves). The situations to use ***15ma bassa*** are very rare.

All rules for ***8va*** apply to ***15ma.***

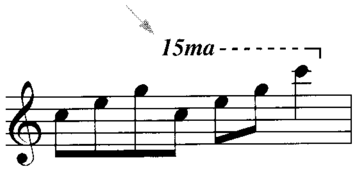

COLL' 8VA

Coll' 8va (or ***coll' 8***) is an abbreviation for *coll' ottava* meaning **"with the octave."** It is a shorthand system of writing octaves, **discouraged in traditional engraving.**

The function of ***coll' 8va*** differs from ***8va*** in that the written notes are **doubled at the octave.**

For the extender line, a longer dash length or a solid line is recommended, to make the ***coll' 8va*** distinguishable from ***8va.***

The ***coll' 8va*** doubles at the **octave above** when notated in the **treble clef.**

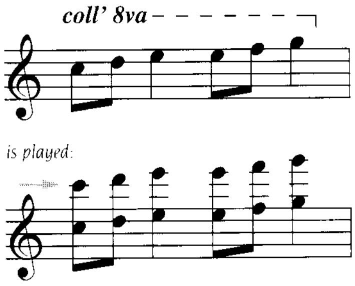

The ***coll' 8va*** doubles at the **octave below** when notated in the **bass clef.**

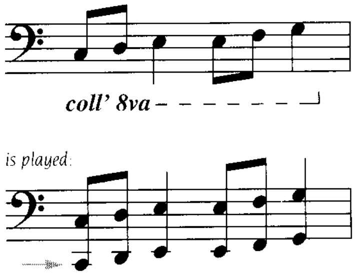

INSTRUMENT CONSIDERATIONS

ORCHESTRA / BAND SCORES

Often a score will have **limited space,** causing leger lines to become too crowded and illegible. Octave signs help to **eliminate vertical overcrowding.**

WOODWINDS

Wind players prefer to read leger lines.

STRINGS

Octave signs should be used very sparingly for viola and cello. Violins are accustomed to reading either leger lines or ***8va*** (but not ***15ma***).

KEYBOARD INSTRUMENTS

Celesta uses ***8va*** but not ***8va bassa*** (the range does not extend low enough).

Harpsichord rarely uses ***8va*** or ***8va bassa.*** Leger lines are more commonly used.

Piano uses ***8va*** and ***15ma*** in the treble clef and ***8va bassa*** in the bass clef.

Organ uses ***8va*** but not ***8va bassa*** (the range does not extend low enough).

Ornaments

Ornaments are primarily used in Baroque, Classical or Romantic period music. The interpretation of an ornament **varies according to the musical period** or composer. It is advised to consult detailed references on ornament usage and interpretation.

Ornament symbols are **placed above the staff** regardless of stem direction.

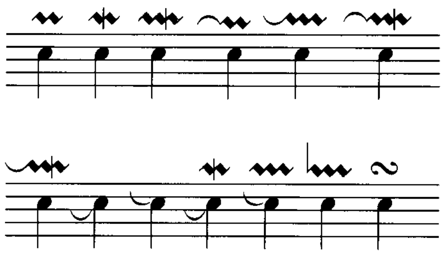

When two parts share a staff, an ornament symbol applying to the lower part is placed below the staff.

Ornaments are usually **centered** to the notehead. **Depending upon the position** of the ornament, the realization of the ornament will vary.

In the example below, the turn symbol is presented two ways, centered above the note and positioned after the note.

Notice the different realizations of the ornament.

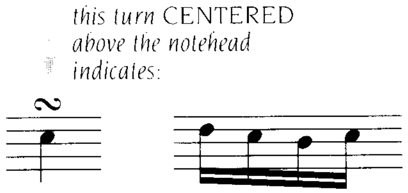
this turn CENTERED
above the notehead
indicates:

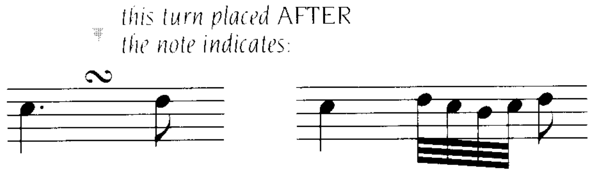
this turn placed AFTER
the note indicates:

Ossia

Ossia is an indication for an **alternate version,** which may be easier or more difficult. It is treated the same as a realized ornament.

The ossia passage is **placed above** the full-size staff. The passage is cue size (65–75%).

Only the **passage concerned is notated**—the staff begins directly before and ends directly after.

Align all beats.

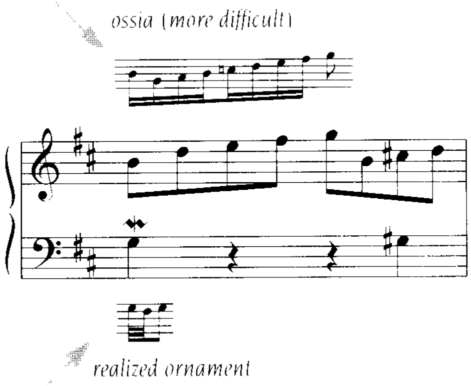
ossia (more difficult)
realized ornament

Pedal marks

FOR THE PIANO

There are three pedals on a grand piano: the **damper pedal** (right), the **sostenuto pedal** (center) and the **una corda pedal** (left). Some upright pianos have only two pedals—the damper and una corda.

Damper pedal

Markings for the damper pedal (sustain pedal) are **always placed below the grand staff** and are usually placed below all other musical elements.

The most commonly used pedal marks consist of three elements: *pedal down, pedal change and pedal up.*

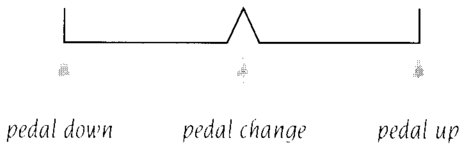
pedal down
pedal change
pedal up

The **pedal-down marking** begins vertically aligned with the left edge of the notehead.*

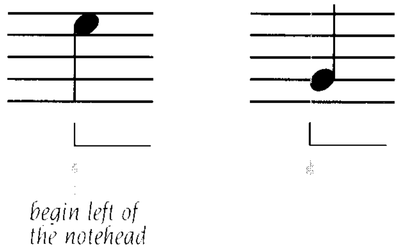
begin left of
the notehead

The *point* of the **pedal-change marking** is vertically aligned with the center of the notehead.

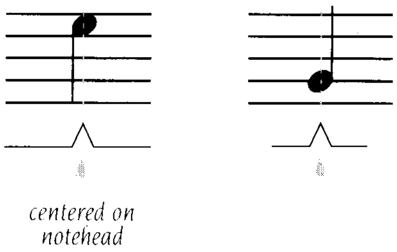
centered on
notehead

---

*Some publishers center or align flush right. In any case, a style decision should be made.

The **pedal-up marking** is vertically aligned to the left of the notehead.

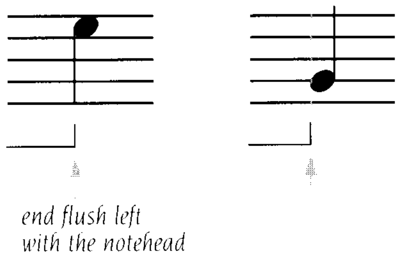
end flush left
with the notehead

The traditional method of pedal marking consisted of **pedal down** and **pedal up.**

This method is somewhat obsolete.

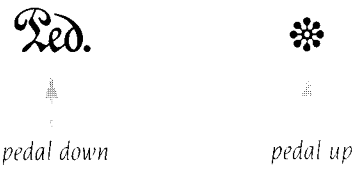
pedal down
pedal up

Drawbacks of this method of pedal indication include the potential for inaccuracy and a cluttered look.

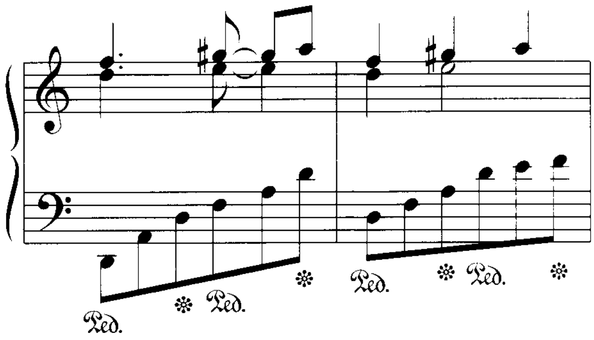

Flutter Pedal

To notate the rapid fluctuating of the pedal, a **zigzag pattern** interrupts the usual pedal marking.

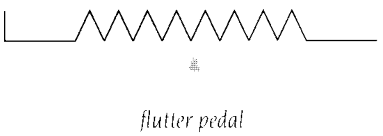
flutter pedal

Wrapping from staff to staff

The pedal mark is interrupted at the end of the staff. It continues on the next staff, **beginning immediately after the key signature.**

If the pedal mark continues to the end of the piece, it is vertically aligned with the thin line of the final double barline.

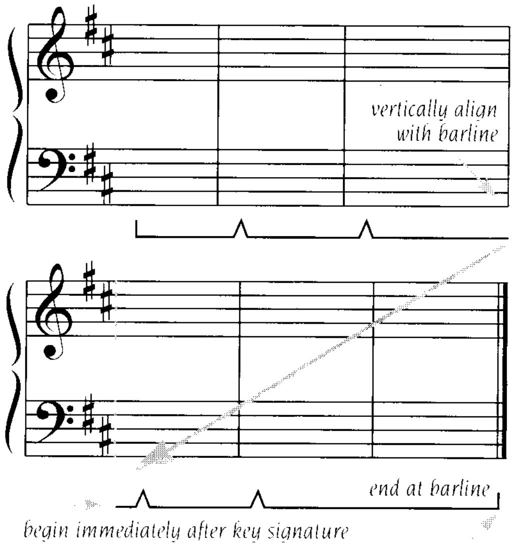
vertically align
with barline
end at barline
begin immediately after key signature

Simile

When the pattern **has been clearly established,** ***simile*** may be used instead of continuing the pedal marks. (See *Simile*)

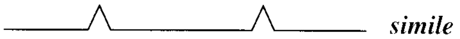

If the pattern changes, simply begin the pedal marks again.

Half pedal

The clearest indication is to simply **add "1/2" to the pedal marking,** whether at the down pedal or at the pedal change.

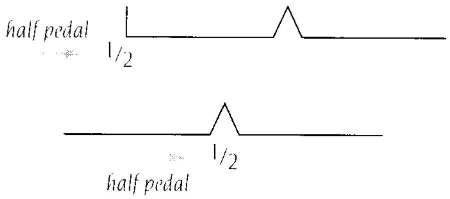
half pedal
1/2
1/2
half pedal

Una corda pedal

The indication ***una corda*** (one string) or ***u.c.*** is placed where the pedal is pressed; ***tre corde*** (three strings) or ***t.c.*** indicates the return to normal.

Sostenuto pedal

This is indicated by the abbreviation ***sos.*** (sometimes in combination with a bracket).

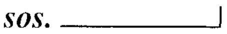

FOR THE ORGAN

Two symbols are used, one for the *heel*, one for the *toe.*

U Λ

heel *toe*

The symbols are **NOT inverted** when moved from one side of the staff to the other.

The pedal marks for the right foot are always **placed above the staff and centered on the notehead.**

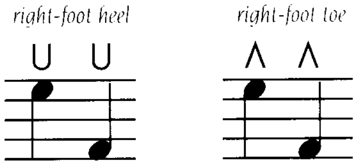
right-foot heel
right-foot toe
U U
Λ Λ

The pedal marks for the left foot are always **placed below the staff and centered on the notehead.**

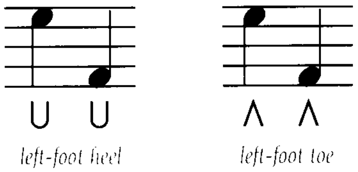
U U
Λ Λ
left-foot heel
left-foot toe

HARP PEDALS

Pedal indications for harp are **settings for tuning.**

The left and right foot have **separate sets** of pedals.

Notation for the harp is closely linked to the position of each of these pedals. Therefore, **enharmonic spellings** are frequent in harp music.

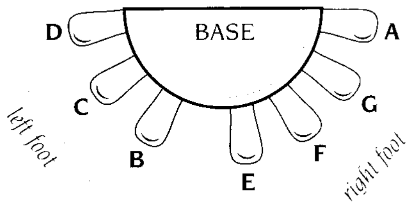
D
BASE
A
C
G
left foot
B
E
F
right foot

The tuning of the harp is based on the three **notch positions** of the pedals.

When a pedal is in its resting position, **center,** the string is tuned to the note's letter name (D, C, B, E, F, G & A).

When a pedal is **raised,** the note is lowered a half step (loosens the string).

When a pedal is **lowered,** the note is raised a half step (tightens the string).

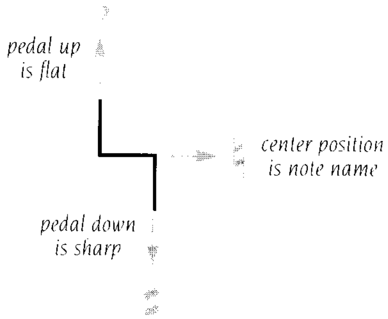
pedal up is flat
center position is note name
pedal down is sharp

The following three examples illustrate possible ways to indicate harp pedal settings in a score.

They may be listed **horizontally, stacked or using the Salzedo diagram.**

D♭ C♯ B♭ E F♭ G A♯

or

D♭ C♯ B♭
E F♭ G A♯

or

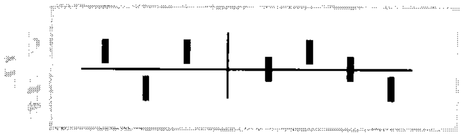

the Salzedo diagram

Picking

Picking is notated with *down-stroke* and *up-stroke* symbols. The same symbols may be used as those that are used for down-bow and up-bow.

Down-stroke and **up-stroke** indications are centered to the notehead and are always placed above the staff.

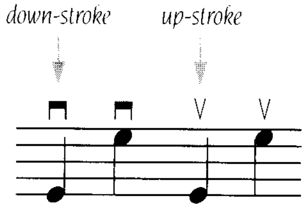
down-stroke
up-stroke

Repeat signs

Indications for repeats occur on three levels: **beat repetitions, measure repetitions and section repetitions.** The first two are used almost exclusively in manuscript or in popular music.

BEAT REPETITIONS

Angled slashes following a single beat indicate that whatever is written on the **first beat is to be repeated.** These are commonly **used in manuscript.** Do not use in engraved music.

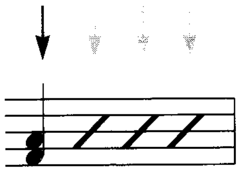

Another use of beat repeats is in **guitar or rhythm parts,** where they help pinpoint where chord changes occur. These are used in engraved music. Stems, beams and ties may be used when necessary.

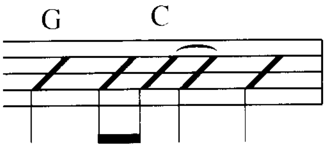

ONE-MEASURE REPEAT

This sign (𝄎), meaning **repeat the previous measure,** consists of a slash with two dots and is centered in an otherwise empty measure. In engraved music, its use should be limited to rhythm parts.

When several consecutive measure repeats are used, numbering is helpful. **Place a small number** above the staff and **center it on the measure repeat sign.**

The **notated measure is considered the first measure,** with the actual numbering beginning on the first repeated measure.

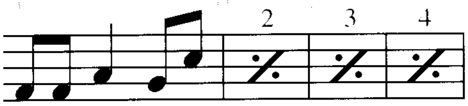

Write out the music **at key points** such as rehearsal numbers, page turns, etc.

TWO-MEASURE REPEAT

When **two measures are to be repeated,** a sign consisting of *two* slashes with dots is placed **on the center bar line** of two otherwise empty measures, **with the number 2 above it.** The number is in the same typeface and size as a time signature number.

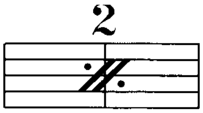

This sign should *not* be used when one measure is to be repeated twice. It should only be used for **two consecutive measures that are *different*,** and are to be repeated. In engraved music, it should only be used in rhythm parts.

If more than two measures are to be repeated, the **repeat sign** should be used.

SECTION REPETITIONS

The **repeat sign** is used to indicate the repeat of one or more measures, or an entire section.

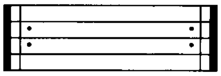

The repeat sign looks like a final double barline, with two dots positioned in the spaces above and below the center line of the staff.

The right-facing repeat sign is **placed at the beginning** of the first measure of the section to be repeated. The left-facing sign is **placed at the end** of the last measure of the same section.

Thus, the section that is to be repeated is **framed** by the inward-facing repeat signs.

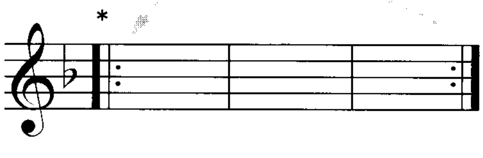

*If the repeat begins **at the beginning of a piece, the right-facing sign is not needed.**

For **two adjoining repeated sections,** repeat signs are placed, as before, framing each section. The **adjoining repeat signs may share the thick bar.**

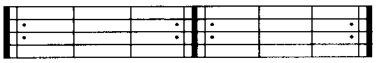

With key changes

When a key signature change occurs at the beginning of a repeated section, a double bar is added and the **key signature is placed before the first repeat sign**.

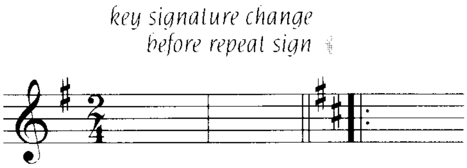
key signature change
before repeat sign

With multiple endings

A section that is repeated may have a different ending each time it is played.

The repeat sign is placed at the beginning of the section and at the last bar line of the 1st ending.

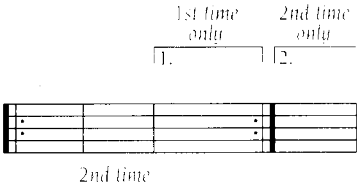
1st time only
2nd time only
1.
2.
2nd time

The **brackets do not intersect or touch the staff** but are placed above the staff, avoiding all musical elements. They are aligned vertically with the barlines.

1st ending brackets are **always closed.**

If the **2nd ending** is **not the end of the piece,** the bracket will be **open-ended.**

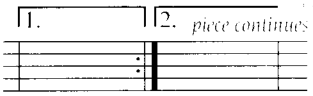
1.
2. piece continues

If it is the **end of the piece** (or movement), the bracket will be **closed.**

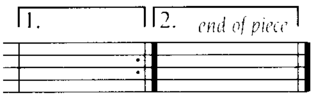
1.
2. end of piece

If a section is played three or more times but there are only two endings, number the brackets accordingly. This occurs often when there are lyrics.

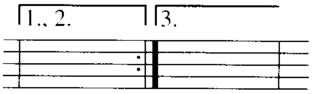

Rests

This section deals mainly with the positioning of rests. For more on the use of rests, see *Meter* and *Augmentation dot.*

TABLE OF RESTS

In the listing of rests below, each rest is half the value of the previous rest, starting with the double whole rest.

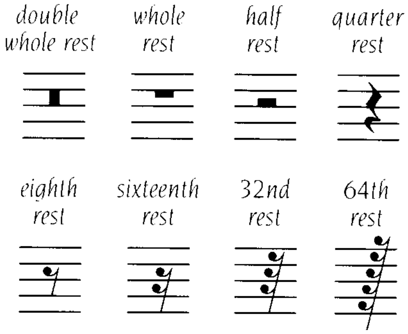
double whole rest
whole rest
half rest
quarter rest
eighth rest
sixteenth rest
32nd rest
64th rest

DOUBLE WHOLE REST (BREVE)

The double whole rest is placed between the third and fourth staff lines, touching both lines.

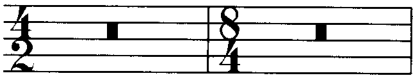

WHOLE REST

With the exception of $\frac{4}{2}$ and $\frac{8}{4}$, the whole rest is used to signify a **complete measure of rest** in any time signature.

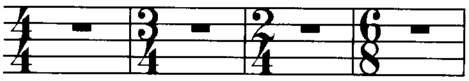

It **hangs** from the fourth staff line.

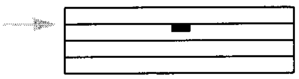

Whole rests are **centered in the measure.**

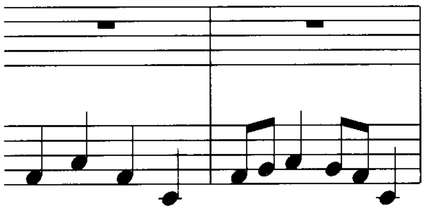

When two parts share a staff, whole rests are positioned on the **fifth line for the top** part and the **first line for the bottom.**

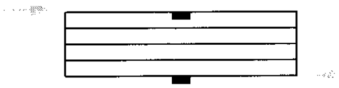

If either part conflicts with the placement of the rest, the rest may be **positioned outside the staff.**

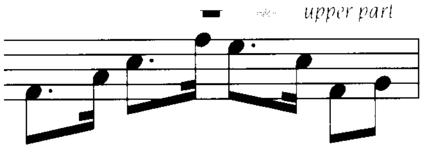

HALF REST

The half rest is positioned to **sit on the third staff line.**

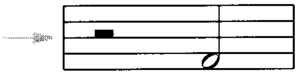

When two parts share a staff, the half rest will sit on **either the fifth line or the first line.**

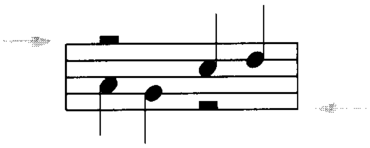

If either part conflicts with this placement, the rest may be positioned **outside the staff.**

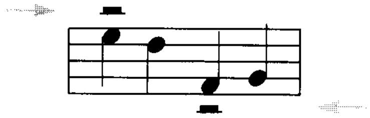

Half rests are **positioned at the beginning of beat 1 or 3**—never centered between two beats.

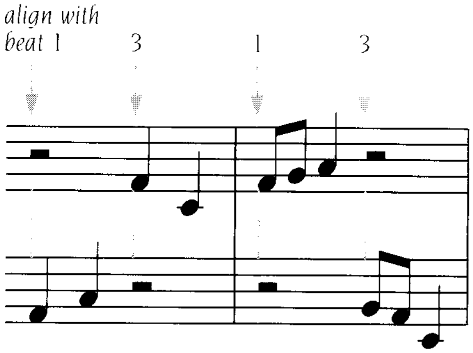

In the above example the use of half rests instead of two quarter rests is correct. However, meter must be considered—do not automatically substitute a half rest for two quarter rests.

(See *Meter)*

QUARTER REST

The quarter rest is positioned as follows. Notice the position of the bottom hook.

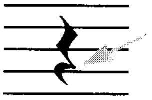

When two parts share a staff, the quarter rest is moved **up or down** from its normal position. The **relationship** of the rest **to lines and spaces** must be the same as in normal position.

The rest may be completely outside the staff when necessary.

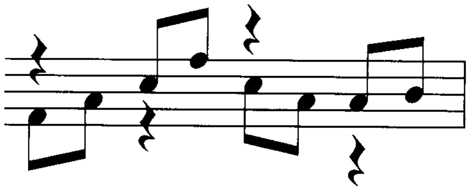

In some situations the above rests could be left out entirely. In other situations, if *both parts* were resting, one rest in normal position could serve for both.

EIGHTH, SIXTEENTH & 32ND RESTS

The guidelines for positioning quarter rests in the example above also apply for eighth, sixteenth and 32nd rests.

The ***hook* of these rests is always placed in a space** (or between imaginary leger lines if placed outside the staff).

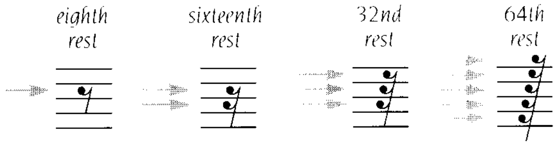

If one of the above rests **interferes with a beam, the rest moves up or down** so that stem lengths may be more normal. Other rests in the measure remain in normal position.

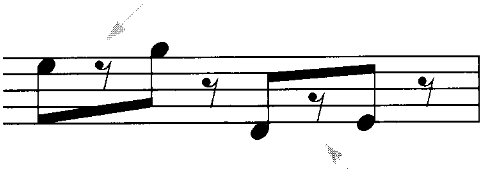

MULTIPLE-MEASURE RESTS

For a rest of more than one measure, a **horizontal line** (the thickness of a beam) is placed on the middle staff line, with a thinner vertical line at each end extending from the second to the fourth staff line.

A number the size and font of a time-signature number is placed above the staff and centered in the measure.

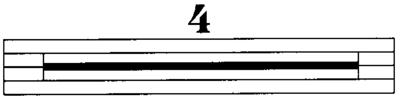

Break the multiple-measure rest for rehearsal numbers, tempo changes, etc.

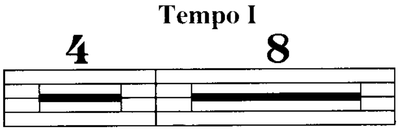

RESTS ON A SINGLE-LINE STAFF

When there is only one part on a staff, position the rests as follows:

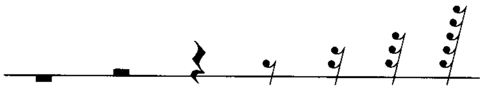

When **two parts share a single-line staff,** place the rests above and below the line, not touching the line. Use a leger line for whole and half rests.

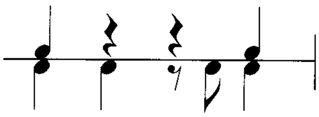

Scores

On a complete score all parts are included on a system and are arranged in a certain order.

All information contained on the parts and score must be **identical.**

On the first page of the score the system (or first system) is **indented.** Brackets connect instrument *choirs.* (See *Bracket, Systems*)

The **full name of each instrument** is given to the left of the first system.

Abbreviations are substituted for the full instrument names on subsequent systems.

Rehearsal numbers (indicated at the top of the score) serve as key reference points.

They may be **letters** or **measure numbers** enclosed in a circle or square.

All **extracted parts** will have **identical rehearsal** numbers.

SCORE ORDER

Instruments should be arranged in what is known as *score order.*

Woodwinds
Brasses
Percussion
Other instruments
Strings

Each **family** of instruments is predominantly **grouped together** from highest-pitched instruments to lowest-pitched instruments, with some traditional exceptions.

The French horns, although not the highest-pitched, are traditionally placed above the other brass instruments.

Sharing a staff

In band or orchestra music some parts of instrument families share the same staff. Combine parts 1 & 2 or 2 & 3; avoid combining three instruments on one staff.

BEGINNING A PASSAGE

If parts are divisi, no indication is necessary unless one or both instruments (such as strings) can play more than one note at a time. (See *Divisi*)

Indicate ***a2*** above staff if both parts play in unison. Use only one stem per note.

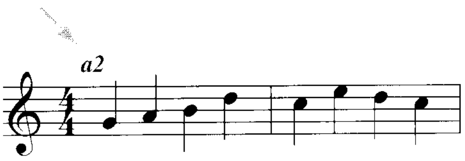

If the unison continues after a page turn, repeat ***a2*** in parentheses.

If like instruments are sharing a staff and only one part plays, indicate which part: I. or II.

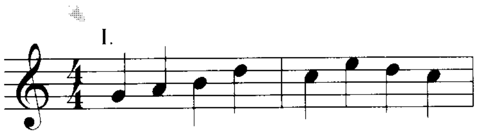

If unlike instruments are sharing the staff and one part plays, indicate which instrument is playing.

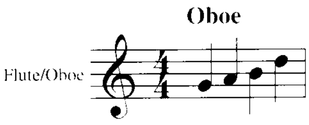

STEMMING OF TWO PARTS

If two parts are playing **different pitches but the same rhythm,** they share **one stem.**

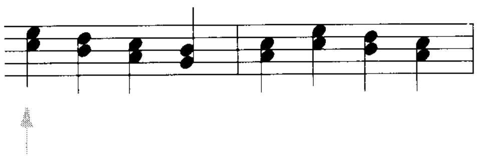

If the parts are **alternating between unison and divisi, and share the same rhythm,** opposite stem direction makes the unisons clear.

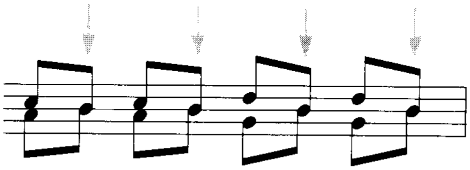

If both parts have **different rhythms,** opposite stem direction is necessary.

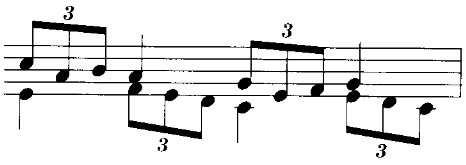

(See *Accidentals, Rests*)

FROM TWO PARTS TO UNISON

Opposite stem direction is used for clarity on **unison notes in short passages.**

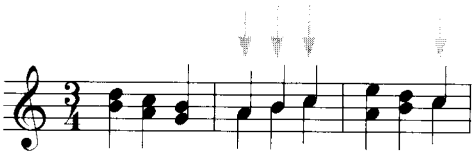

Indicate ***a2*** above staff and use one stem for longer passages.

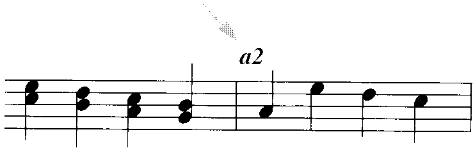

FROM TWO PARTS TO ONE:

Use rests in shorter passages.

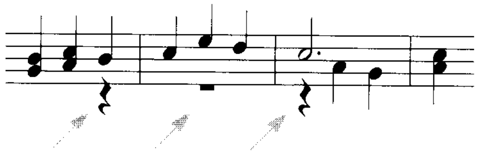

Place I. or II. above the staff and use one stem (for two *like* instruments in longer passages or after long rests).

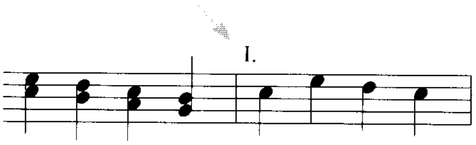

Mark passage with instrument name (for two *unlike* instruments in longer passages or after long rests).

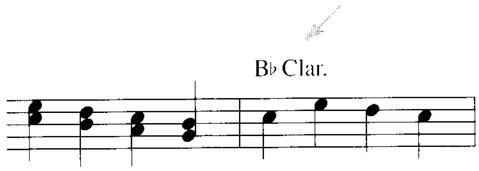

STAFF INDICATIONS

Staff indications (dynamic, slurs, tempi) are **placed below** the staff.

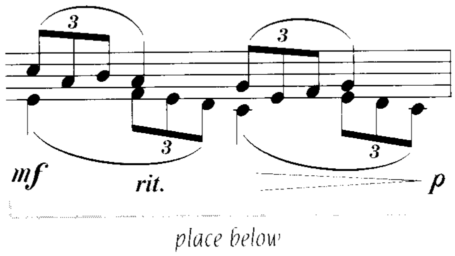

If the two parts move **independently,** a **separate set of directions** is needed for each part, above and below the staff.

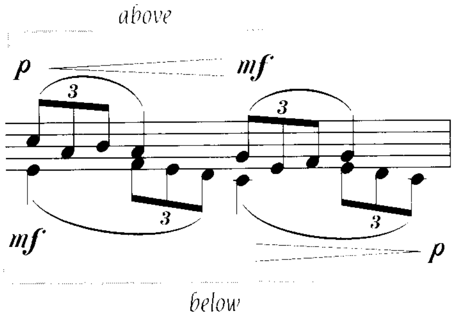

WHEN NOT TO SHARE A STAFF

1. If the range of one part tends to cross the range of the other, the parts should have **separate staves.**
2. If the parts require **different clefs** or **clef changes** the parts should be separated.
3. If the two parts are **too rhythmically diverse,** they should be separated.

In all cases **a bracket should be used,** not a brace.

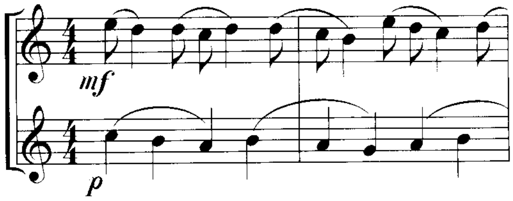

Sharp

See *Accidentals*

Simile

When the word ***simile*** is indicated after a **pattern has been established** (for example, of articulations, pedal marks, bowing, etc.), it means to **continue in the same manner.** (See *Articulations, Pedal marks*)

Slurs

Although a slur looks somewhat like a tie, the placement, positioning and use of slurs are different from ties. Slurs and ties should not be confused.

Slurs are used to indicate phrasing or technique. Depending on the instrument, the slur will imply different techniques to be used.

PLACEMENT OF SLURS

The **stemming** of the notes is the main factor in determining whether the slur is to be placed above or below.

When **stems extend up,** the **slur is placed under**, on the notehead side.

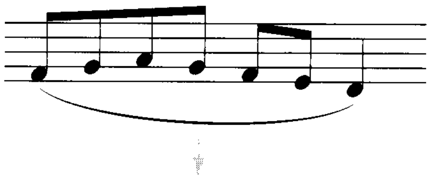

When **stems extend down,** the **slur is placed over,** on the notehead side.

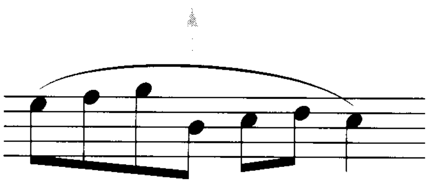

When **stems are in both directions,** the **slur is always placed over.**

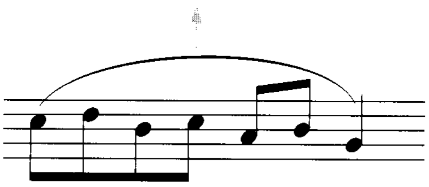

If the notes are whole notes, *imagine* the stems!

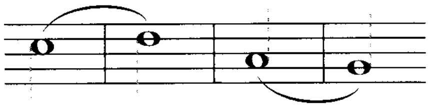

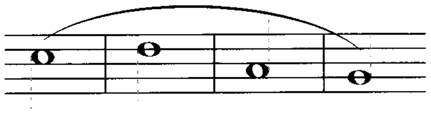

If the **phrase is substantially long,** the slur may be placed over, regardless of stem direction.

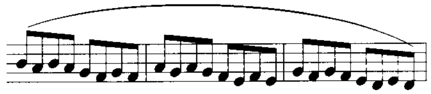

PLACEMENT ON THE NOTEHEAD

Slurs have a **definite beginning and ending;** they should not be "hanging in space" so that the beginning and ending notes must be guessed at.

If the slur is on the notehead side, the beginning and ending of the slur should be **centered on the notehead.** It should not touch the notehead.

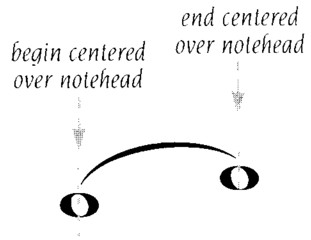
end centered over notehead
begin centered over notehead

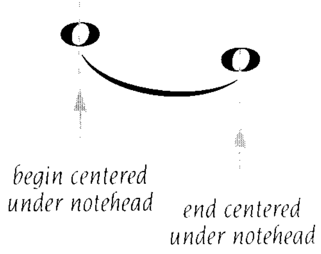
begin centered under notehead
end centered under notehead

BEGINNING FROM THE STEM

Place the beginning of the slur so it **starts mid-stem** (to the **right** of the stem—never begin to the left, crossing the stem).

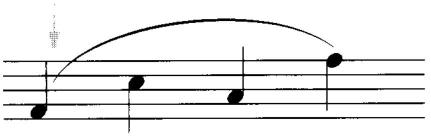

In some cases the slur may begin **at the end of the stem.**

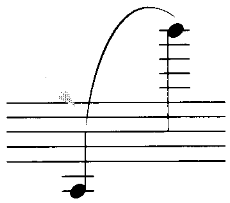

Begin at the end of the stem **if beaming is involved.** Never cross the beam with a slur.

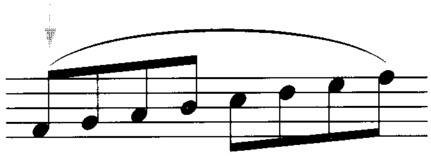

ENDING ON THE STEM

Place the end of the slur so it **slopes towards the notehead,** away from the end of the stem. **Center the slur** on the notehead.

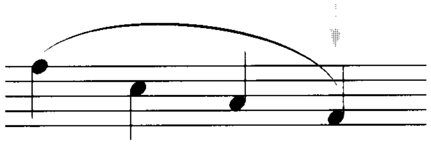

The slur may also end **at the end of the stem** if it would make a more pleasing shape.

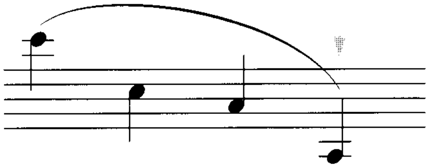

End at the end of the stem **if beaming is involved.** Never cross the beam with a slur.

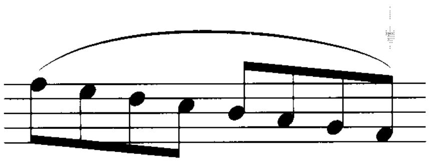

SLUR DIRECTION

If the direction of the **phrase ascends,** the slur will **slope upward.** If the direction of the **phrase descends,** the slur will **slope downward.**

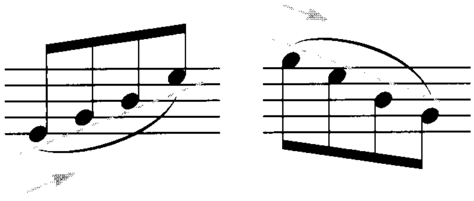

When a slur **begins and ends on the same note**, the slur can remain on the same horizontal level.

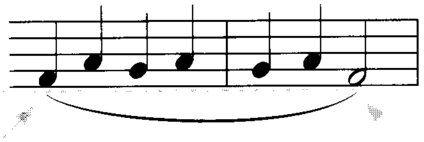

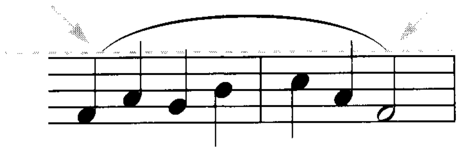

SLURS WITH CHORDS

Only one slur is needed when there is only one stem per chord.

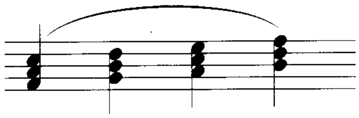

SHARING THE SAME STAFF

For two parts, place slurs **above for the upper part, below for the lower part.**

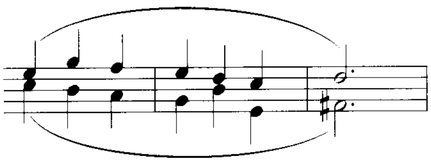

For two *sets* of parts, place slurs **above for the upper parts, below for the lower parts.**

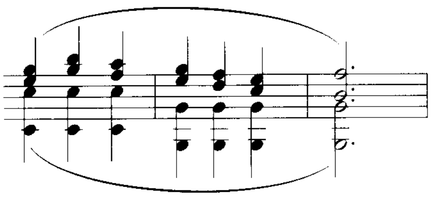

WRAPPING FROM STAFF TO STAFF

Align the end of the slur with the end of the staff. The slur ends at an **angle** (ascending or descending) at the end of the staff.

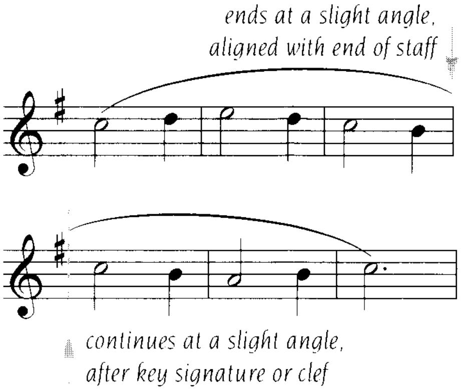
ends at a slight angle,
aligned with end of staff
continues at a slight angle,
after key signature or clef

Continue the slur in the lower staff immediately after the key signature (or clef), also at an **angle.** Be sure the first note is far enough to the right so that it is very clear that the slur does not begin on the note.

If the slur is over the staff it **remains over** when continuing on the next staff. If it is under, it **remains under** on the next staff.

The **direction of the slur** (ascending or descending) continues logically in the following staff.

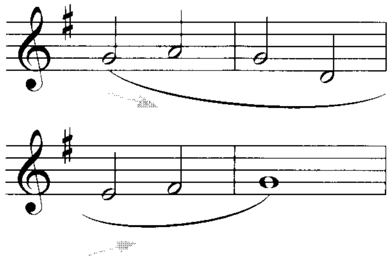

slur ascends to follow direction of notes

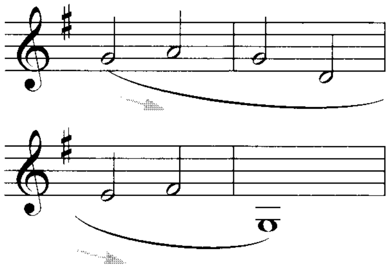

slur descends to follow direction of notes

DIFFERENTIATING SLURS FROM TIES

When slurs and ties are both involved in a chord progression such as the following, **position the slur opposite to the tie** and at a noticeable **angle.**

slur MUST be moved opposite to the tie

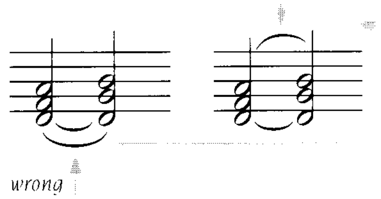

wrong

An even better way to write the above:

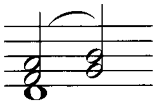

If a slur **begins and ends on the same pitch** and could possibly be mistaken for a tie, position the slur on the stem side.

slur is moved above
to avoid being mistaken for a tie

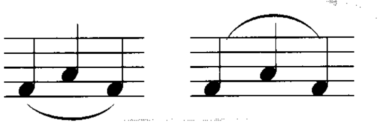

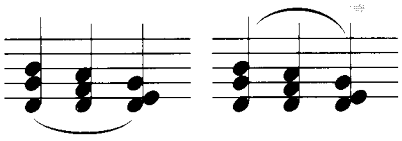

When a slur connects chords with **tied notes within**, combine the tied notes into a single note with opposite stem direction.

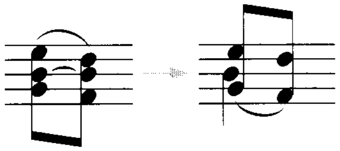

BEGINNING OR ENDING WITH A TIE

A slur **must include both notes of the tie,** whether the tie is at the beginning or at the ending of the slur. (For an exception, see "Elisions," on the next page.)

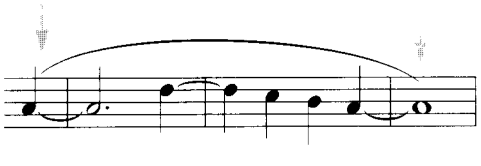

ELISIONS

In keyboard music, one slur might end at the same point where another slur begins.

The two slurs **share the center of the notehead** (elide) without touching each other.

end of
one phrase

beginning of
another phrase

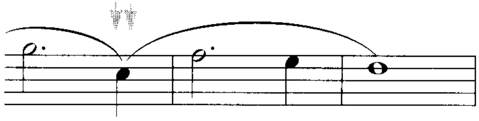

When the **elision occurs on tied notes**, the first phrase may end on the first tied note. The second phrase can then begin on the second tied note.

end of
one phrase

beginning of
another phrase

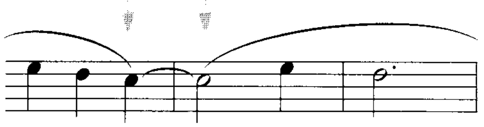

GRACE-NOTE SLURS

Grace notes **may or may not have slurs.** The slurs are usually somewhat smaller to match the grace notes.

Since grace notes are always stemmed up (unless there are two voices), the slur is usually **placed below.**

Grace-note slurs may be placed more to the inside edges of the noteheads.

grace-note slur

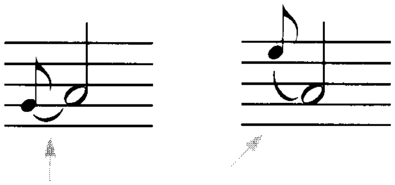

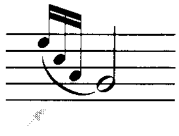

slur for beamed
grace notes

Grace-note slurs elide with slurs.

grace-note slurs and slurs

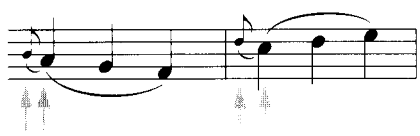

SLURS AND ARTICULATIONS

Beginning or ending on notes with an articulation:

The beginning or ending of a slur is placed **outside staccato and tenuto** marks.

Center over or under the articulation.

slur is outside notehead and articulation

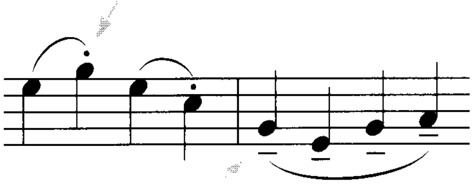

The beginning or ending of a slur is usually placed **between accents** and the notehead.

slur is between notehead and articulation

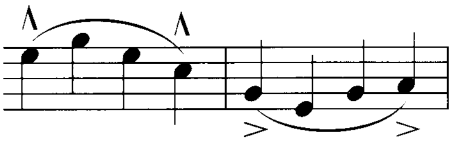

here the accent is placed inside the slur

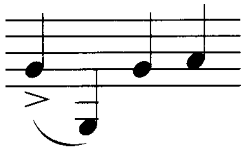

Articulations between the beginning and ending notes remain inside the slur.

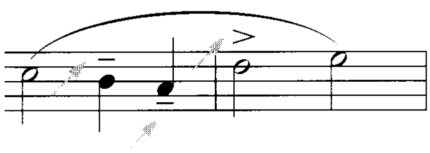

Fermatas are placed **outside** the end of a slur.

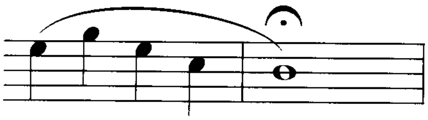

"LET RING" SLURS

The use of a short slur to indicate continued vibration should be placed on notes of longer duration only. Angle the slur slightly.

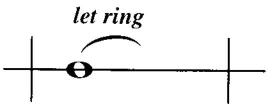

INSTRUMENT CONSIDERATIONS

Woodwinds and brass

Legato technique is indicated by the use of a slur. The slur might also indicate breathing and phrasing.

Slurs in combination with staccato or tenuto marks show varying degrees of legato.

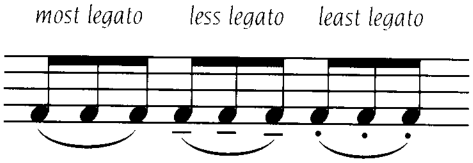

For keyboards

If the space between the two staves is limited, slurs may be placed outside the staff.

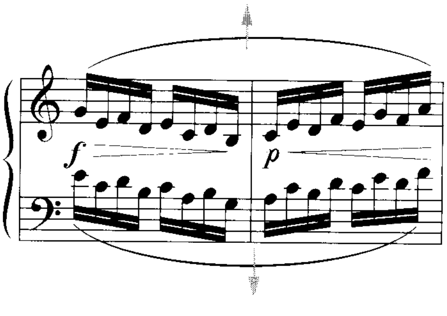

An S-shaped slur is sometimes used in music notated on the grand staff.

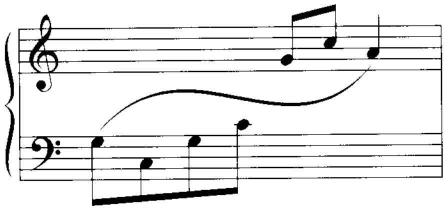

For strings

Slurs can only partially indicate phrasing. Slurs are also used to indicate various bowing techniques. (See *Bowing*)

For vocal music

A **slur is used for a *melisma*** (two or more notes that are to be sung on a single syllable).

The word or syllable should be placed flush left with the notehead.

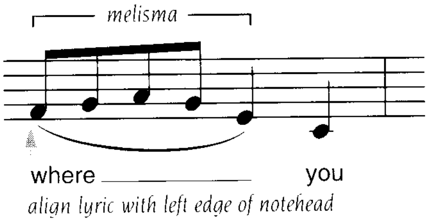
melisma
where
you
align lyric with left edge of notehead

Slurs to indicate phrasing in vocal music are seldom used. Since breathing is a major factor in phrasing, rests and breath marks can be used as a partial indicator of phrases.

Slurs can be drawn using a **broken line** instead of a solid one. The broken slur is most often **used to accommodate lyrics for multiple verses.**

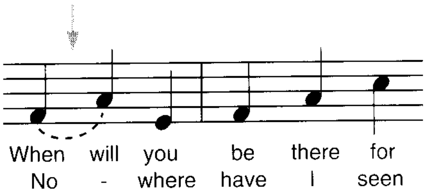
When will you be there for
No - where have I seen

Similarly, **broken slurs** may also be used for syllable differences of translated text.

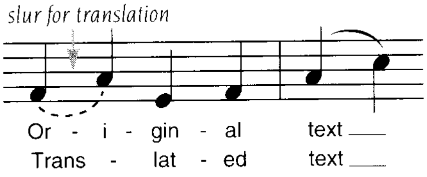
slur for translation
Or - i - gin - al text
Trans - lat - ed text

Spacing

Spacing is of primary importance to the **visual quality** of any musical notation. While the attention to *musical* detail may be excellent, poor judgment in spacing may make the score look ill-conceived.

An awareness of spacing occurs on many levels, from the **smallest element** (the beat) **to the largest** (the score or book). The desired effect is that the notation *and* layout will immediately convey the musical intent. Ultimately, any score should look **accessible** regardless of the level of difficulty.

PROPORTIONAL SPACING

The concept of proportional spacing takes into consideration the **mathematical spacing** of a beat, measure and system, in relation to the **density of activity** (the number of notes per beat, measure and system).

With proportional spacing, a half note, for example, must get *more* space than a quarter note, but not *twice* as much.

SPACING WITHIN THE MEASURE

Notes are spaced within the measure according to a **compromise between** two extremes:

1) **Space given according to note-value** (for example, a half note getting twice as much space as a quarter note).

2) **Equal spacing given to each note.**

Compare the normal (proportional) spacing with the two extreme examples.

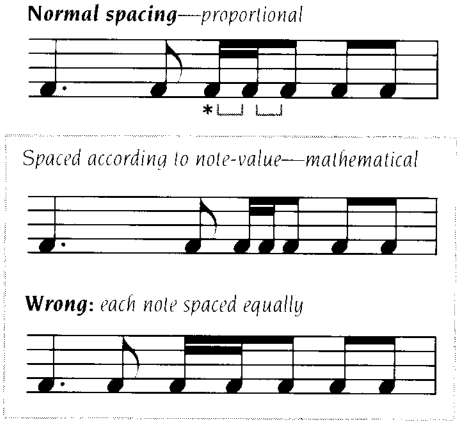
Normal spacing—*proportional*
*
Spaced according to note-value—mathematical
Wrong: *each note spaced equally*

*Make sure that consecutive notes of equal value have equal space *following* each note.

MEASURES WITHIN THE STAFF

The size of the measure within the staff is determined according to a **compromise between** two extremes:

1) **Equal space** for each measure.

2) Measure size **based on note density only.**

Normal spacing of measures

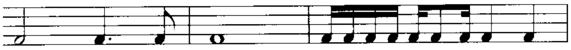

Wrong: *spaced according to equal measure size*

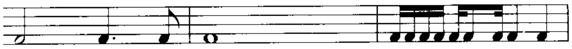

Wrong: *spaced according to measure density only*

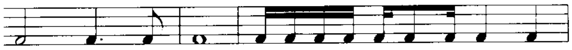

VERTICAL SPACING

Between staves

Musical elements (notes on leger lines, articulations, slurs, pedal marks) **projecting beyond a staff should never conflict** with the elements of another staff or system. There should be enough space allowed for ease of reading.

Inside systems

On the other hand, the **staves of a system should not be too far apart.** This is especially important for a system that is to be read by one performer—a keyboard player, for example. The eye should not have to jump too far from one staff to the other, in order to read notes for the right and left hand simultaneously.

> *The most common misjudgment is to place staves (or systems) exactly the same distance from one another without adjusting to compensate for the projecting musical elements.*

Between systems

More space should be be provided between systems than between the *staves* of the systems.

The **spacing between systems should be balanced,** with the page margins and the number of systems on the page in mind. If the **page is dense,** the systems will be closer to each other and to the minimum vertical margins, while allowing enough distance between systems to distinguish one from another. If the **page is sparse,** the upper and lower page margins could be larger.

GENERAL PAGE LAYOUT

Page numbers

The position of the page number in relation to the page edges should *never* change, regardless of page spacing. Printers can use the page number as a guide for positioning the image area on the page.

If page numbers are placed to the left and right, as is usual, **even numbers** are always to the **left, odd numbers** always to the **right.** Page numbers may be centered, but this is not the usual practice. Centered page numbers interfere more with the music and text.

Page numbers should be positioned **no closer than 1/4 inch from** the top (or bottom) of the **page edge.**

Title page

The title, credits, copyright and any accompanying text can be **vertically adjusted to accommodate** a loosely spaced page or a tightly spaced page. Such type should always be well placed, considering the space allowed.

Facing pages

Facing pages **should look good together.** As much as possible, they should have a **balanced look** (notes on a system, number of systems, white space, etc.). The vertical page margins should be the same if at all possible.

Page turns

It is very important to lay out the measures and systems in a manner that will allow a good page turn. Having **rests in the last measure** before the page turn is the most desirable. Second to rests would be perhaps a whole note (that could be held with the pedal) or a sparse measure—anything that would allow a hand to be free to turn the page.

Try to **minimize the number of page turns** by adjusting the amount of music on the page or by changing the sequence of pieces.

Book layout

Calculating the number of pages needed or what will work in the allowed number of pages is crucial to an appealing layout.

If a book contains many short pieces and a particular order is not important (or can be adjusted slightly), there are several things that can be done to make a good layout.

Often a piece can be **condensed or expanded** to offer choice page turns and good facing pages.

Consider reordering the pieces to **minimize the number of page turns. Group pieces** with an odd number of pages together so that the ones with an even number of pages can have facing pages.

A poorly placed piece at the beginning of a book can throw off the facing pages for the rest of the book.

Backing up

Expect to go back and adjust **various elements** (page turns, systems, staves, measures or notes within a measure) when planning a layout for a book. A change in the order of pieces and the number of pages might create density problems on other pages.

The part must be considered with the whole and the whole considered with the part.

White space

With experience comes the **appreciation of white space** on a page or within a book. There is nothing more daunting than overwhelming density. The eye should be guided from note to note, measure to measure, staff to staff and page to page with a minimum of surprises. If a performer must stop playing in order to understand the notation, or loses his place because of poor placement, the notation and layout are *not* successful.

Staccato

See *Articulations*

Staff

The staff commonly consists of five lines.

The **staff line weight** should be thick enough to be clearly legible but thin enough for the notes of the staff to be easily read.

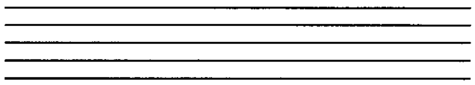

A staff line will be **thicker than stem lines** and **equal to** (or thinner than) **barlines.**

If a note is on the line, the line will always run through the **center of the notehead,** regardless of the note value.

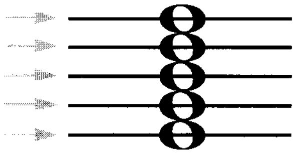

A notehead in a space is always clearly positioned between the two staff lines.

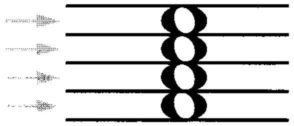

A **one-line staff** is used for non-pitched percussion.

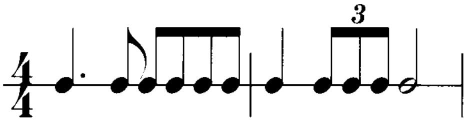

or

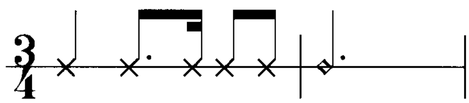

Leger lines extend the range of the staff.

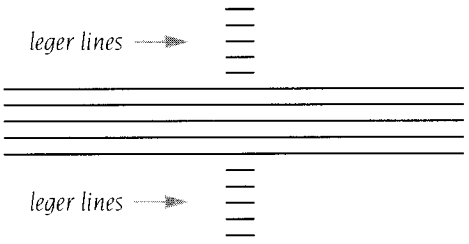
leger lines
leger lines

Barlines divide the staff into measures.

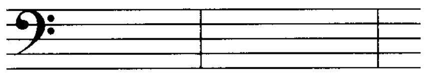

A **clef** placed on a five-line staff determines the pitches for each line and space.

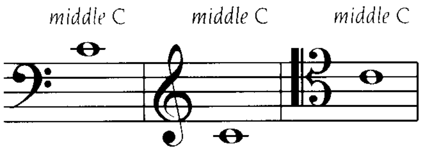
middle C
middle C
middle C

A **key signature** determines the key of the music notated on the staff.

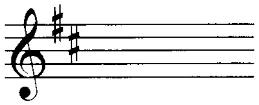

Two or more staves grouped together (with a systemic barline) form a **system.**

A **bracket** groups staves into systems or parts of systems. (See *Bracket*)

A **brace** groups staves into a **grand staff.** (See *Brace*)

Some examples of systems:

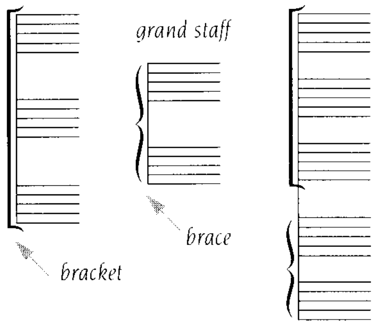
grand staff
brace
bracket

The **first system** (or staff) of music is usually **indented** about 1/2 inch.

Stems

Stems are **thinner than staff lines and barlines.**

STEM LENGTH

The **normal stem length** is 3^{1}/2 spaces (one octave).

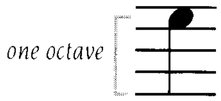

With leger lines

When a note extends **beyond one leger line,** the stem **must touch the middle staff line.**

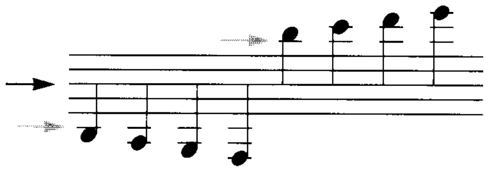

When parts share a staff

Up-stem notes **above the middle line are shorter** than normal. Traditionally, their stem length is relative to surrounding notes, getting progressively shorter as the notes go higher. The shortest stem length is 2^{1}/2 spaces (interval of a 6th).

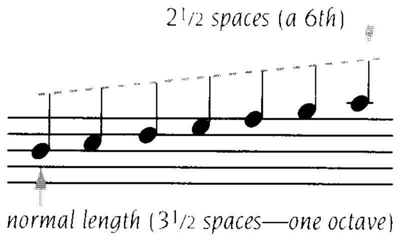
2 1/2 spaces (a 6th)
normal length (3 1/2 spaces—one octave)

For the computer, a setting should be chosen that works well for most situations.

Down-stem notes **below the middle line** follow the same guidelines as the above.

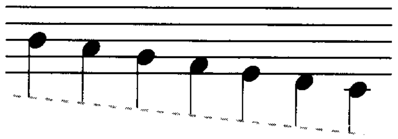

With the interval of a 2nd

The stem is always placed **between the two notes** of an interval of a 2nd, with the **upper note always to the right**, the **lower note always to the left.**

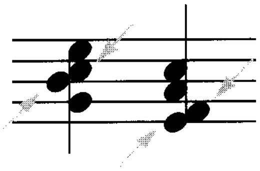

STEM DIRECTION

For notes on the **middle line and above,** the stem is down.

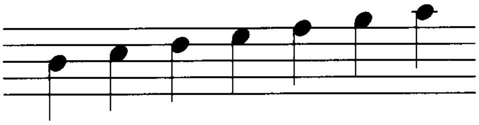

For notes **below the middle line,** the stem is up.

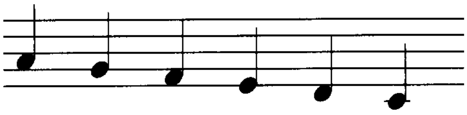

When sharing a staff

Notes have **opposite stem direction** when sharing a staff. (Stems in the "wrong direction" will have shorter stems, as previously mentioned.)

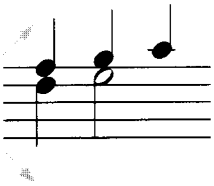

When notes share a stem

If the **note above is farther** from the middle line than the note below, the **stem goes down.**

If the note **below is farther** from the middle line, the **stem goes up.**

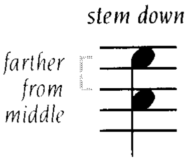

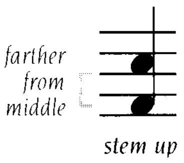

When the two notes are an **equal distance from the middle line,** the preferred direction is down.

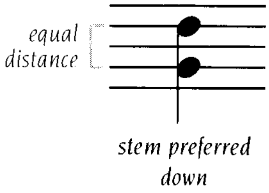

If a **pattern of stem direction** is established, stem direction is *sometimes* maintained when the notes are an equal distance from the middle. (This practice is all but obsolete.)

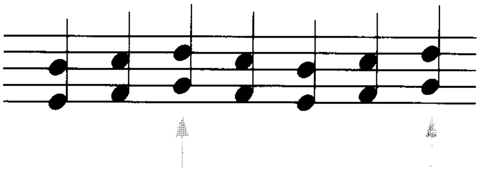

The same rules apply to *more than two notes* sharing a stem. The **distance of the OUTER NOTES from the middle line** determines stem direction.

When the **outer notes** are an **equal distance from the middle line** and the majority of notes are above the middle line, the stem goes down.

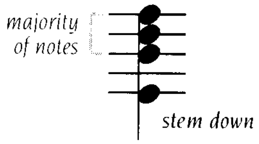

When the **outer notes** are an **equal distance from the middle line** and the majority of notes are below the middle line, the stem goes up.

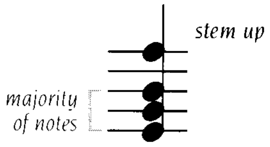

For beamed groups

Simple rules for stemming of single notes and chords apply to beamed groups when possible (for example, if all notes in a beamed grouping are on or above the middle line of the staff, stems are down).

(See *Beams*)

Systems

Two or more staves grouped together (with a systemic barline) form a system.

A **bracket** groups staves into systems or parts of systems. (See *Bracket*)

A **brace** groups staves into a **grand staff.** (See *Brace* and *Grand staff*)

Some examples of systems:

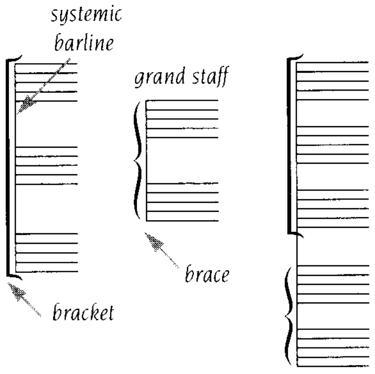
systemic
barline
grand staff
brace
bracket

(See *Barlines*)

The **first system** of music is usually **indented** about 1/2 inch.

Tablature

Tablature is a system of notation that graphically represents the strings and frets of a stringed instrument (such as guitar).

The letters **"TAB"** are placed where a clef would normally be. The only function of these letters is to identify the staff as being tablature.

Each note is indicated by **placing the fret number on the appropriate string.**

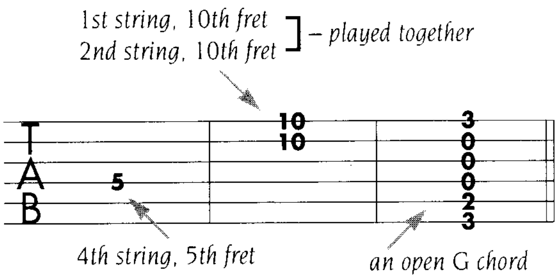
1st string, 10th fret
2nd string, 10th fret
– played together
T
A
B
5
10
10
3
0
0
0
2
3
4th string, 5th fret
an open G chord

Tablature does NOT indicate rhythmic values; therefore, a five-line staff using **standard notation** is placed above the tablature staff.

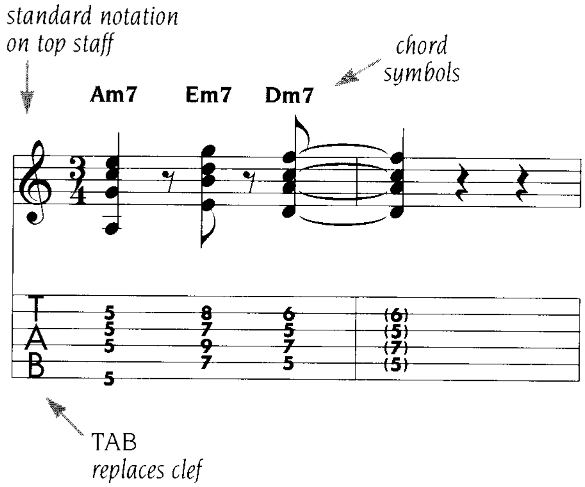
standard notation
on top staff
chord
symbols
Am7
Em7
Dm7
T
A
B
5
5
5
5
8
7
9
7
6
5
7
5
(6)
(5)
(7)
(5)
TAB
replaces clef

A **systemic barline connects** the two staves together. Measure **barlines** are broken.

Chord symbols (when appropriate) are often added above the top staff.

Picking or fingering techniques notated on the top staff are repeated on the tablature staff.

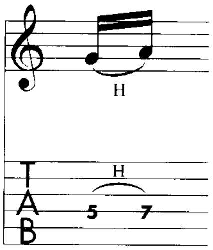

Every note on the top staff must have a **corresponding finger number** on the tablature staff—**tied notes** are in parentheses.

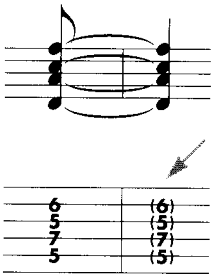

Tablature may consist of **four lines** (for electric bass and mandolin) or **five lines** (five-string banjo).

The **notation** for guitar tablature outlined on the previous pages applies to all tablature, regardless of the number of strings.

Electric bass

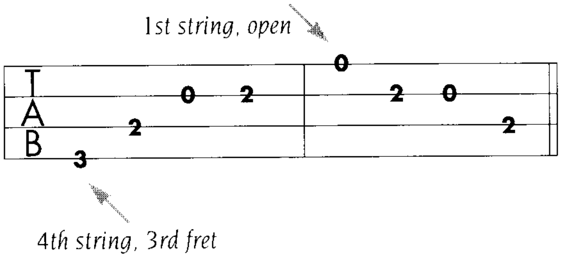

Five-string banjo

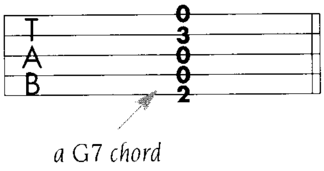

Tempo marks

Tempo marks indicating the rate of speed may be followed by an expression mark and/or metronome mark.

The left edge of the tempo mark is **vertically aligned to the left edge of the time signature.**

The type is bold Roman.

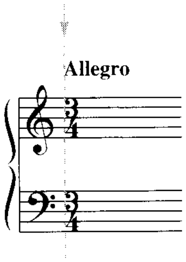

WITH EXPRESSION MARKS

Tempo marks are often accompanied by an indication of **expression;** however, expression marks should **never *replace*** tempo marks.

Presto con brio

Slowly, with much emotion

WITH METRONOME MARKS

The tempo mark may be more precisely indicated with the addition of a metronome mark.

The metronome mark may be ***specific,*** a tempo ***range*** or ***general.*** The type is slightly smaller than that of the tempo and is usually enclosed in parentheses. The note is cue size or smaller.

Allegro con moto (♩ = 144)

Allegro con moto (♩ = 132–144)

Allegro con moto (♩ = ca. 144)

TEMPO VARIATIONS

If the tempo is varied by the use of markings such as ***ritardando, rallentando*** or ***accelerando,*** the term ***a tempo*** is placed at the point where the tempo is reinstated. *The type will be bold italic.*

If the tempo changes, and then returns to the original tempo later in the piece, the term **Tempo I** is indicated **above the staff.** *The type will be bold Roman.*

Tenuto

See *Articulations*

Ties

A tie connects **two consecutive notes of the same pitch,** extending the duration of the first note to include the second. (The second note does not have its own attack.)

The shape of a tie is somewhat similar to a slur, but the **placement, positioning and use of ties and slurs are different;** the two should not be confused.

BEGINNING AND ENDING

Begin and end on the same horizontal level.

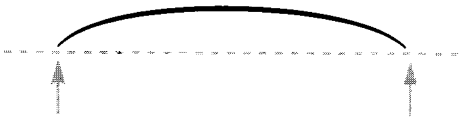

Begin to the right of the notehead.
End to the left of the next notehead.
The tie **does not touch the noteheads.**

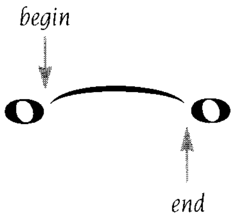

AVOID AUGMENTATION DOTS

Ties should never collide with augmentation dots. **Begin the tie to the right of an augmentation dot.**

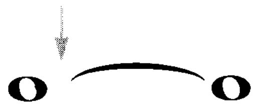

PLACEMENT ON THE STAFF

Avoid touching staff lines.
Place the **center** of the curve in a "space."

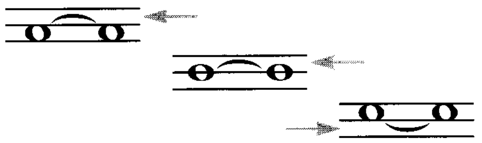

Ends of ties may cross staff lines—it is better to cross even slightly, than to just touch a staff line.

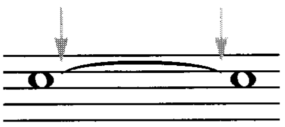

Adjust curve for longer ties.
Raise or lower the center of the curve to the next space.

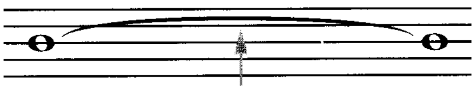

DIRECTION ON THE STAFF

Curve ties opposite of stem direction.
Stem down—curve above notehead.
Stem up—curve below notehead.

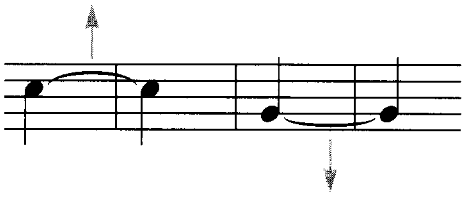

Ties on whole notes are treated as if the whole notes had stems.

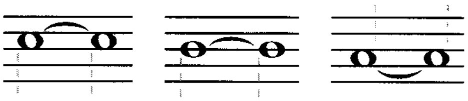

For mixed stem direction, the tie is always placed **above.**

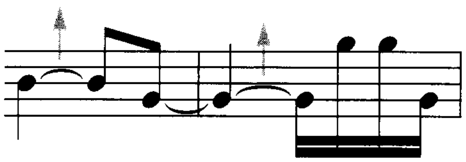

SHARING THE SAME STAFF

For two parts, place ties **above for the upper part, below for the lower part.**

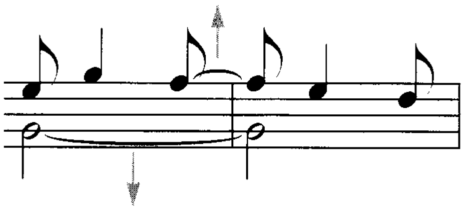

For two *sets* of parts, place ties **above for the upper two parts, below for the lower two parts.**

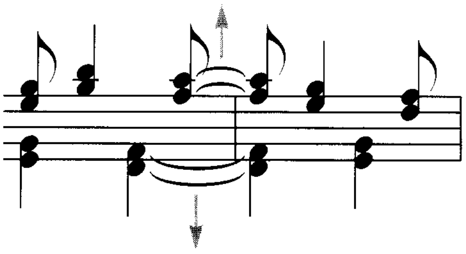

MULTIPLE TIED NOTES

When more than two notes are tied, ties must be added from note to note.

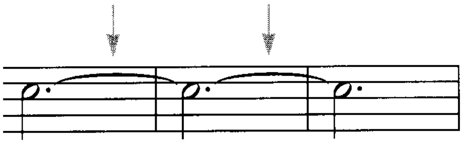

WRAPPING FROM STAFF TO STAFF

Align end of tie with end of staff.

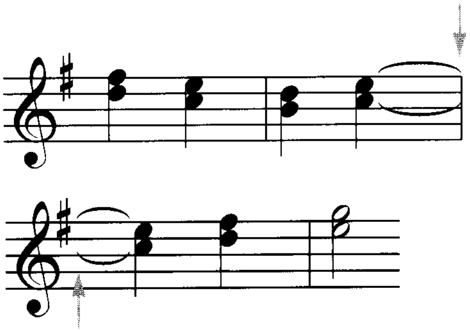

Continue tie in the lower staff immediately after the key signature (or clef).

Beginning and ending points always remain on the **same horizontal level.**

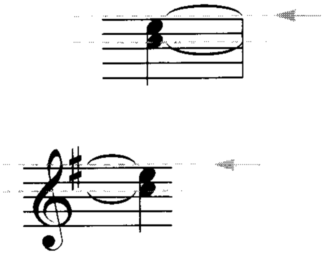

DIRECTION OF TIES FOR CHORDS

Highest note—tie is **above.**

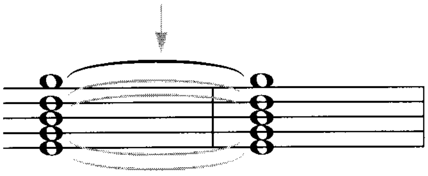

Lowest note—tie is **below.**

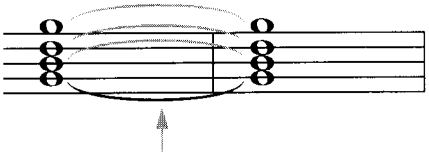

Notes between highest and lowest—on the **middle line and above,** ties are **above.**

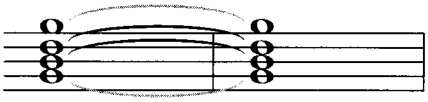

Below the **middle line,** ties are **below.**

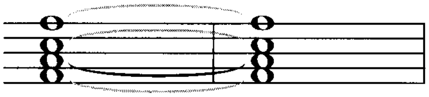

For **intervals of a 2nd**, ties will be in **opposite directions** whenever possible.

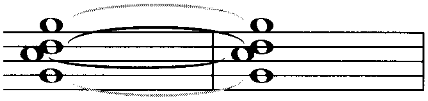

If some notes are not tied within the chords, the uppermost tied note has the tie above; the lowermost tied note has the tie below.

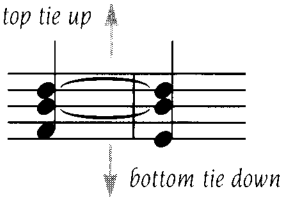

ALIGNMENT OF TIE ENDS

Align ends of ties within a chord.

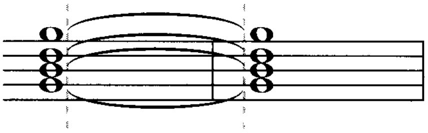

Tie ends for intervals of a 2nd should be adjusted if possible.

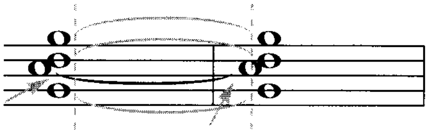

ADJUSTMENT OF TIES

Whenever possible, **maintain the curve** of a tie in uncommon situations. It should look like a tie, not a slur.

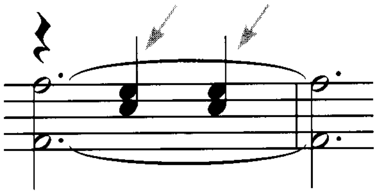

In this instance it is better to have the top tie curved up even though it is a part of the lower voice.

If a time signature interrupts a tie, break the tie to avoid colliding with the time signature.

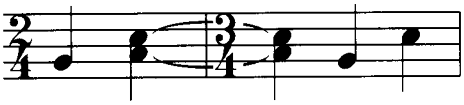

BROKEN (DASHED) TIES

Ties using a broken line instead of a solid one are often **used to accommodate lyrics for multiple verses.**

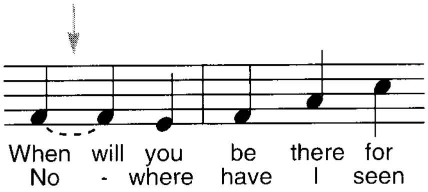

TIES IN MEASURED ARPEGGIOS

Although traditionally considered incorrect, ties are sometimes **extended to the unbroken chord** to avoid a cluttered look.

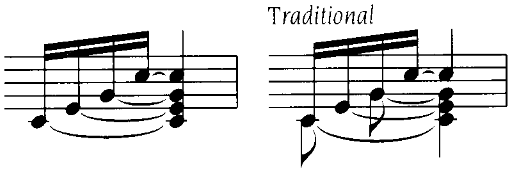

Time signatures

Time signatures indicate the meter for a measure(s) or an entire piece.

Since meter is so closely connected with the proper use of ties, rests, beams and syncopated rhythms, the **correct time signature** must be chosen.

The **upper number** indicates the number of beats per measure.

The **lower number** indicates the note that gets one beat (2 = half note, 4 = quarter note, 8 = eighth note, 16 = sixteenth note, etc.).

All measures after a time signature will be in the same meter until a new time signature is indicated.

PLACEMENT

The **upper number** is always placed between the top line and the middle line of the staff.

The **lower number** is always placed between the middle line and the bottom line of the staff.

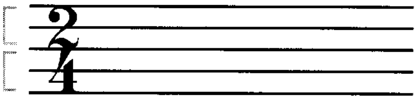

For a **single-line staff,** the line intersects the two numbers.

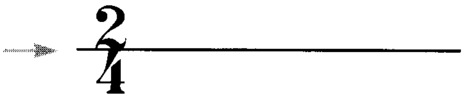

The time signature is **indicated after the clef and key signature.**

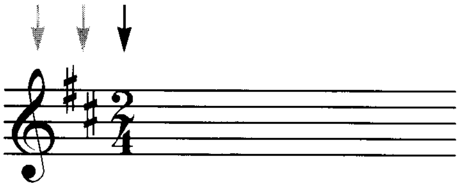

CHANGING TIME SIGNATURES

Time signatures may change from measure to measure, applying to all measures until the next time signature change. **A change in time signature always appears AFTER a barline.**

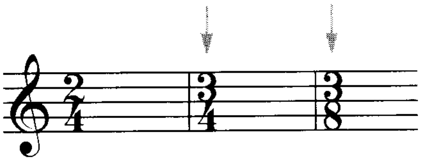

If the time signature changes at the beginning of a line, a **courtesy time signature** is added at the end of the previous staff, *after* the last barline. The new time signature appears after the clef (and key signature, if any) on the next staff.

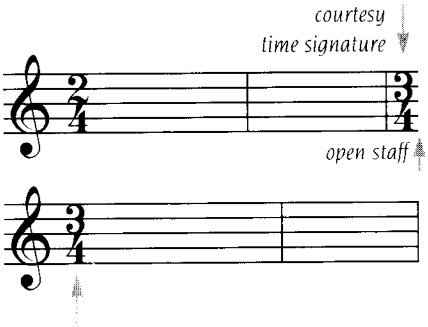

The following illustrates the positioning of the clef, key signature and time signature if all three change simultaneously.

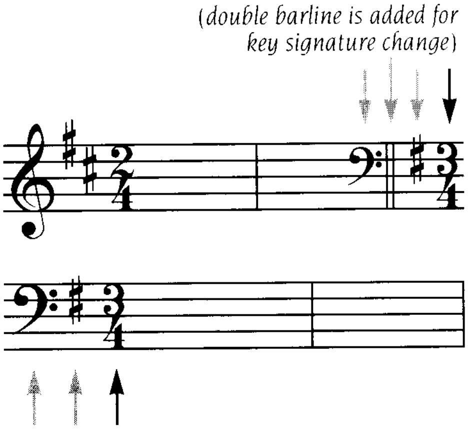

COMPOSITE METER (See *Meter*)

ALTERNATING METER (See *Meter*)

Tremolo

A tremolo may be **on one note,** or **between two or more notes.** It may be **measured** (a clear subdivision) or **unmeasured** (as fast as possible).

ON ONE NOTE

Measured tremolo

A **short slanted line,** ascending from left to right, through the stem of a note indicates that a note is to be subdivided.

Each line placed on the stem indicates the **rhythmic value** of the tremolo.

For instance, one line will indicate eighth notes, two lines—sixteenth notes.

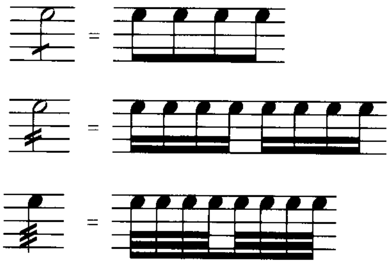

In string music, a measured tremolo is sometimes indicated by placing **divisi dots** above the note, in addition to the slashes through the stem.

The divisi dots precisely indicate the divisions per note. The slashes through the stem indicate the rhythmic value.

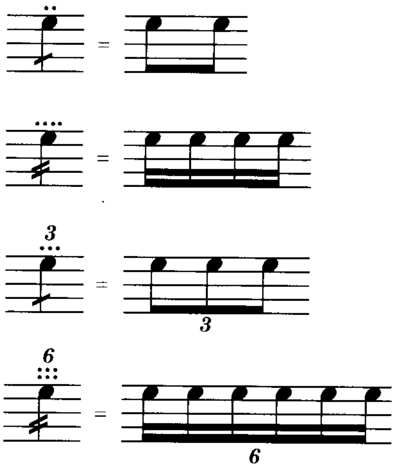

Unmeasured tremolo

An **unmeasured tremolo** (on one note) is indicated by placing three lines through the stem and adding the abbreviation ***trem.*** above the note. (If placed on a beamed note, the beam counts as one of the lines.) Most often seen in notation for strings, percussion and fretted instruments.

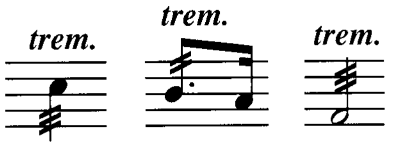

Tremolo lines on whole notes are traditionally **indicated above or below the note** (positioned as if there were a stem, but centered on the notehead).

traditional style

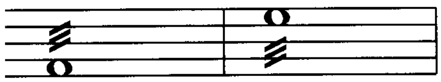

Modern style breaks the whole note into two half notes and places the slanted lines through the stems as usual.

modern style

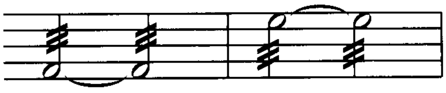

ON MORE THAN ONE NOTE

When a tremolo is indicated between two (or more) different pitches, the notes are connected with a beam, with incomplete beams placed between to indicate the rhythmic value.

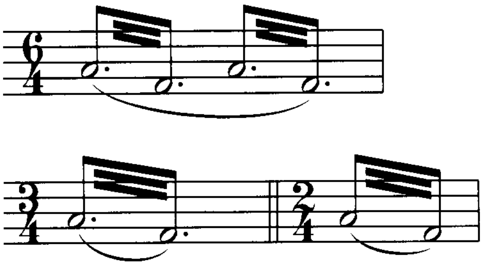

Stems are lengthened to accommodate added beams.

With such a subdivision of the tremolo, **each note equals the total rhythmic value of the tremolo.**

unmeasured whole-note tremolo (traditional style)

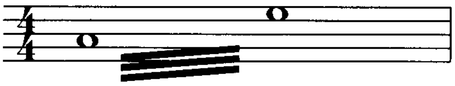

modern style

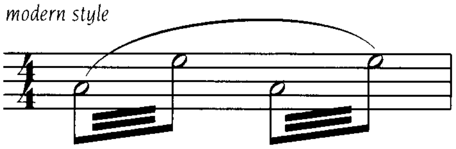

In the previous example, a complete measure of tremolo in $\frac{4}{4}$ time consists of two sets of half notes beamed together.

INTERVAL / CHORD TREMOLO

For some instruments, a tremolo is possible between two, three or more notes.

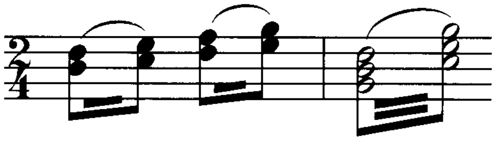

Trill

The abbreviation ***tr*** indicates a trill.

The ***tr*** is always **placed above the note,** regardless of stem direction (unless two parts share a staff).

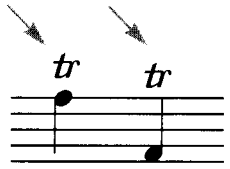

Each trilled note must have its own trill symbol.

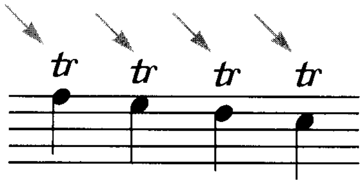

A **wavy line** may be indicated after the ***tr***, especially if the trill is extended.

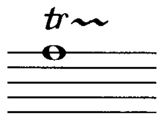

If the trill **extends over two or more notes,** the wavy line continues to the end of the notehead of the last note affected by the trill.

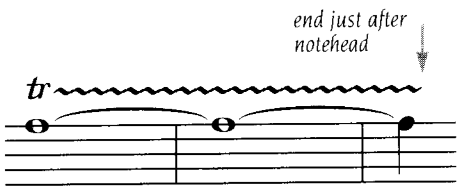

If the **wavy line extends over consecutive notes of the same pitch** but without ties, the notes are to be clearly articulated.

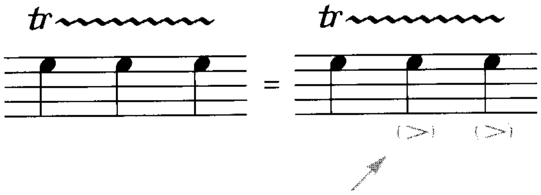

When two parts share a staff, the trill for the lower part is placed below the note.

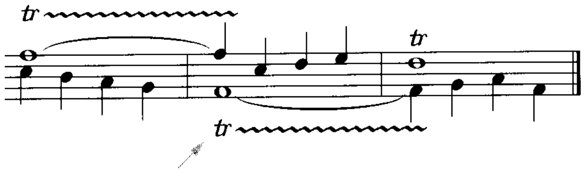

When the upper note of a trill is **altered from the diatonic scale,** a small **accidental** is used with the trill sign. The accidental is placed after the trill sign but before the wavy line (if used.)

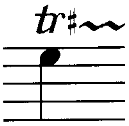

The accidental **may also appear above** the trill sign.

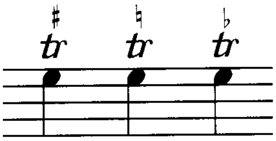

A **cue-size notehead** is sometimes used.

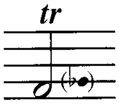

Terminations can be indicated by cue-size or full-size notes.

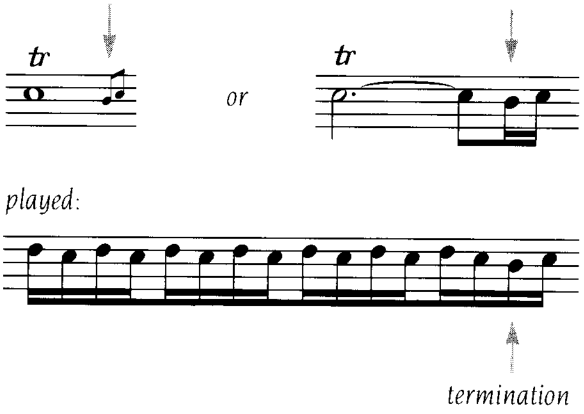

Triplet See *Tuplets*

Tuplets

Rhythmic groupings of notes that are NOT metric groupings are known as **tuplets.**

Duplets, triplets, quadruplets, quintuplets, sextuplets and septuplets are all examples of **tuplets.**

TUPLET NUMERALS

The numeral is larger than a finger number and is **distinctly italic, preferably bold italic.** (See *Fingerings [for keyboard]*)

For beamed tuplets, the numeral is preferred **on the beam side, centered.**

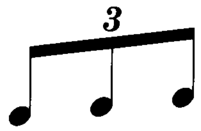

Do not use a tuplet slur with the numeral. This is an obsolete practice and causes confusion between normal slurs and tuplet slurs.

Place the tuplet numeral **outside the staff** whenever possible.

If placed within the staff, the **numeral must be clearly positioned,** avoiding staff lines as much as possible.

The numeral **remains placed with the beam,** whatever the stem direction.

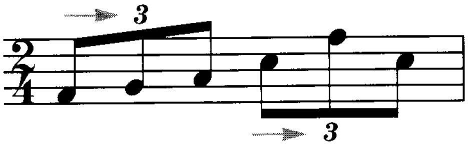

When parts share a staff, the numeral is placed normally (beam/stem side).

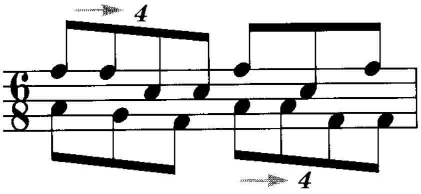

After a **pattern of tuplets** has been established, the numeral may be omitted. ***Simile*** may also be used for clarity.

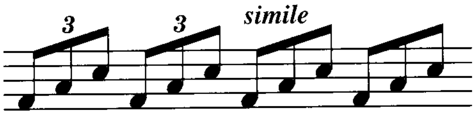

TUPLET BRACKETS

If the tuplet is not beamed, a **bracket is added** and the **numeral is centered** within the bracket.

The bracket always **begins flush left** with the notehead and **ends flush right** with the notehead, whether on the notehead or stem side.

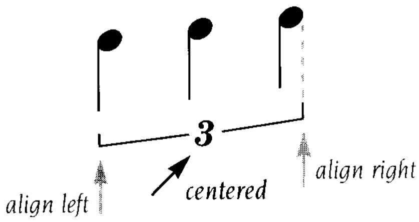

The **bracket is broken** to accommodate the number; the **ends** are always vertical.

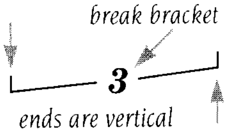

If the tuplet numeral is **moved to the notehead side** of a beamed tuplet (for example, to avoid fingerings), add a bracket for clarity.

The **bracket is placed** at the same angle as the beam.

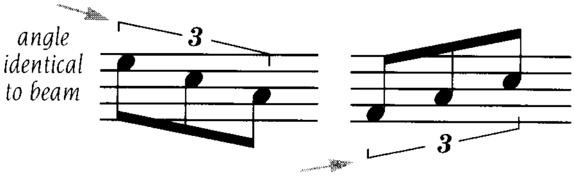

When placing a **bracket with unbeamed notes,** the bracket angle can vary.

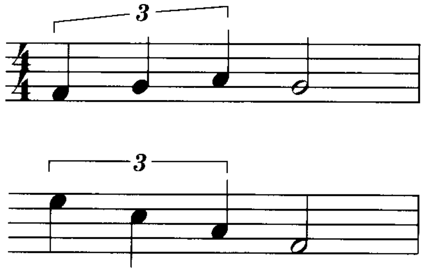

BRACKETS AND UNEQUAL VALUES

For **unequal rhythmic values** within the tuplet group, the bracket must **include all notes** of the group.

The **numeral remains centered** within the bracket.

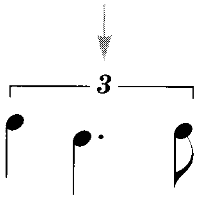

The tuplet unit is clearly seen if the bracket extends to **include the entire allotted space** for the tuplet (as in the first tuplet below).

The bracket does not extend to include the flag (as shown in the second tuplet).

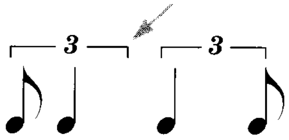

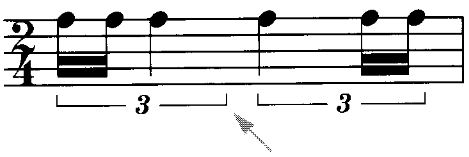

WITH RESTS

The **bracket extends** to include rests.

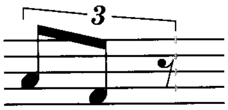

Although the bracket is preferred, an option is to **extend the beam** to include the rest within the tuplet.

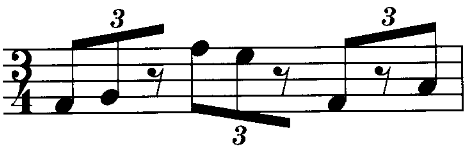

WITH FINGERINGS

It is advised to **move the tuplet figure to avoid confusion with finger numbers.**

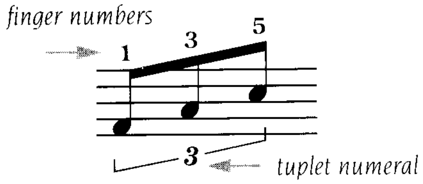
finger numbers
1
3
5
3
tuplet numeral

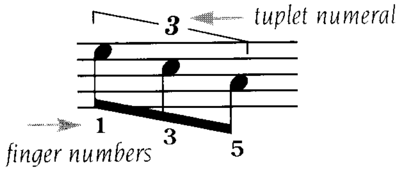
3
tuplet numeral
1
3
5
finger numbers

WITH ARTICULATIONS

If the bracket must be placed on the same side as articulations, the bracket is **placed outside articulations.**

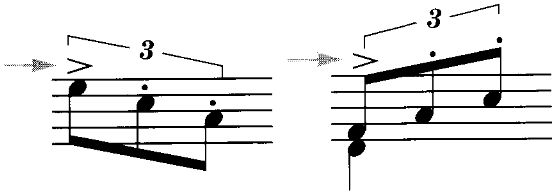

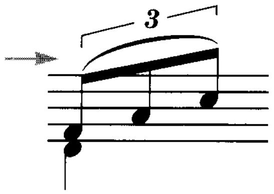

Short slurs may be considered articulations.

DETERMINING NOTE VALUE

A tuplet may occupy a beat, a pulse, a *portion* of a beat or pulse, or an entire measure.

To determine the rhythmic value of a tuplet that occupies **one beat,** subdivide the beat metrically as many times as you can without exceeding the number of notes in the tuplet. The note value of the tuplet will be equal to the note value of that metric subdivision.

In the example below, if each beat were subdivided further, it would have more notes than the tuplet; therefore the tuplet notes are sixteenth notes.

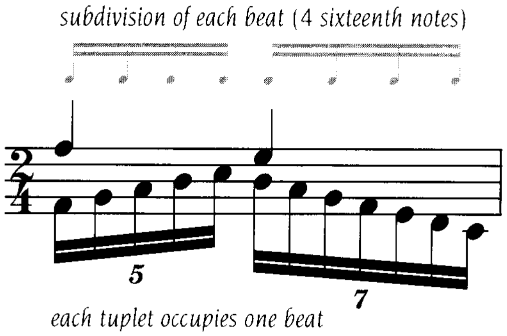
subdivision of each beat (4 sixteenth notes)
5
7

each tuplet occupies one beat

Use the same principle to determine the note value of tuplets that occupy an **entire measure** (or another unit of a measure).

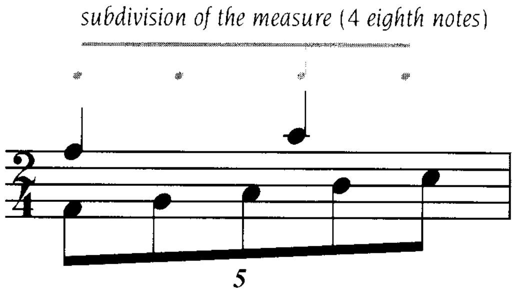
subdivision of the measure (4 eighth notes)
5

tuplets occupying an entire measure

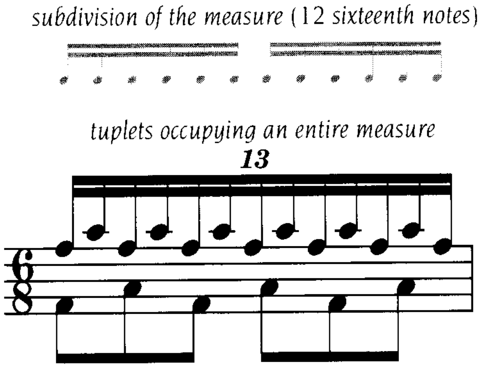
subdivision of the measure (12 sixteenth notes)
tuplets occupying an entire measure
13

Duplets are the only exception to the previous rule deciding tuplet rhythmic value.

Duplets are **equal in rhythmic value to the beat** (not to be confused with pulse).

Duplets only occur in **compound meter.**

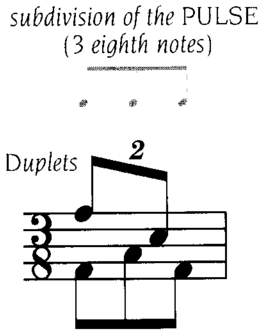
subdivision of the PULSE
(3 eighth notes)
Duplets
2

HORIZONTAL SPACING

A tuplet is spaced mathematically, as proper spacing dictates—not aligned with the metric rhythm.

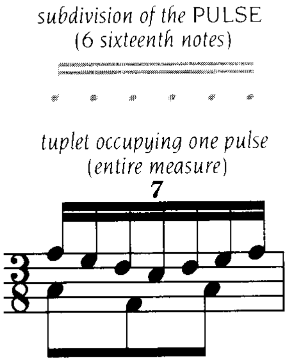
subdivision of the PULSE
(6 sixteenth notes)
tuplet occupying one pulse
(entire measure)
7

Words

See *Lyric*

X notehead

See *Notes*

INDEX OF TOPICS

A cappella 3
A due (a2) 3
A tempo 3
Accents 3
Accidentals 3
Alternating instruments 10
Arpeggio signs 11
Articulations 13
Augmentation dot 21
Barlines 25
Beams 29
Bowing 42
Brace 43
Bracket 43
Breath mark 44
Caesura 45
Chord frames & symbols 46
Circle of fifths 49
Clef signs 49
Coda sign 53
Courtesy accidentals 53
Credits 53
Cue notes 54
D.C. al Coda 55
D.C. al Fine 57
D.S. al Coda 58
D.S. al Fine 59
Divisi 60
Divisi dots 60
Dotted rhythms 60
Dynamics 60
Endings 65
Fermata 65
Fingerings (keyboard) 67
Flat 71
Glissando sign 71
Grace notes 72
Grand pause 74
Grand staff 74
Harmonics 76
Key signatures 78
Layout 82
Leger lines 83
Lyric 84
Meter 88
Metronome marks 95
Natural sign 95
Notes 95
Octave signs 97
Ornaments 103
Ossia 104
Pedal marks 105
Picking 109
Repeat signs 110
Rests 113
Scores 117
Sharing a staff 118
Sharp 121
Simile 121
Slurs 121
Spacing 131
Staccato 134
Staff 135
Stems 137
Systems 140
Tablature 140
Tempo marks 142
Tenuto 143
Ties 143
Time signatures 149
Tremolo 151
Trill 153
Triplet 155
Tuplets 155
Words 160
X notehead 160

Michel Marie Deza
Elena Deza

Encyclopedia of Distances

Fourth Edition

Springer

[redacted]

Preface to *Dictionary of Distances*, 2006

The concept of *distance* is a basic one in the whole human experience. In everyday life, it usually means some degree of closeness of two physical objects or ideas, i.e., length, time interval, gap, rank difference, coolness, or remoteness, while the term *metric* is often used as a standard for a measurement.

But here we consider, except for the last two chapters, the mathematical meaning of those terms which is an abstraction of measurement. The mathematical notions of *distance metric* (i.e., a function $d(x, y)$ from $X \times X$ to the set of real numbers satisfying to $d(x, y) \geq 0$ with equality only for $x = y$, $d(x, y) = d(y, x)$, and $d(x, y) \leq d(x, z) + d(z, y)$) and of *metric space* (X, d) were originated a century ago by M. Fréchet (1906) and F. Hausdorff (1914) as a special case of an infinite topological space. The *triangle inequality* above appears already in Euclid. The infinite metric spaces are usually seen as a generalization of the metric $|x - y|$ on the real numbers. Their main classes are the measurable spaces (add measure) and Banach spaces (add norm and completeness).

However, starting from K. Menger (who, in 1928, introduced metric spaces in geometry) and L.M. Blumenthal (1953), an explosion of interest in both finite and infinite metric spaces occurred. Another trend is that many mathematical theories, in the process of their generalization, settled on the level of metric space. It is an ongoing process, for example, for Riemannian geometry, real analysis, and approximation theory.

Distance metrics and distances have become now an essential tool in many areas of mathematics and its applications including geometry, probability, statistics, coding/graph theory, clustering, data analysis, pattern recognition, networks, engineering, computer graphics/vision, astronomy, cosmology, molecular biology, and many other areas of science. Devising the most suitable distance metrics and similarities, in order to quantify the proximity between objects, has become a standard task for many researchers. Especially intense ongoing search for such distances occurs, for example, in computational biology, image analysis, speech recognition, and information retrieval.

Often the same distance metric appears independently in several different areas, for example, the edit distance between words, the evolutionary distance in biology, the Levenshtein metric in coding theory, and the Hamming+Gap or shuffle-Hamming distance.

This body of knowledge has become too big and disparate to operate within. The number of worldwide web entries offered by Google on the topics “distance,” “metric space,” and “distance metric” is about 216, 3, and 9 million, respectively, not to mention all the printed information outside the web, or the vast “invisible web” of searchable databases. About 15,000 books on Amazon.com contain “distance” in their titles. However, this huge information on distances is too scattered: the works evaluating distance from some list usually treat very specific areas and are hardly accessible for nonexperts.

Therefore many researchers, including us, keep and cherish a collection of distances for use in their areas of science. In view of the growing general need for an accessible interdisciplinary source for a vast multitude of researchers, we have expanded our private collection into this dictionary. Some additional material was reworked from various encyclopedias, especially Encyclopedia of Mathematics ([EM98]), MathWorld ([Weis99]), PlanetMath ([PM]), and Wikipedia ([WFE]). However, the majority of distances are extracted directly from specialist literature.

Besides distances themselves, we collected here many distance-related notions (especially in Chap. 1) and paradigms, enabling people from applications to get those (arcane for nonspecialists) research tools, in ready-to-use fashion. This and the appearance of some distances in different contexts can be a source of new research.

In the time when overspecialization and terminology fences isolate researchers, this dictionary tries to be “centripetal” and “ecumenical,” providing some access and altitude of vision but without taking the route of scientific vulgarization. This attempted balance defined the structure and style of the dictionary.

This reference book is a specialized encyclopedic dictionary organized by subject area. It is divided into 29 chapters grouped into 7 parts of about the same length. The titles of the parts are purposely approximative: they just allow a reader to figure out her/his area of interest and competence. For example, Parts II, III, IV, and V require some culture in, respectively, pure and applied mathematics. Part VII can be read by a layman.

The chapters are thematic lists, by areas of mathematics or applications, which can be read independently. When necessary, a chapter or a section starts with a short introduction: a field trip with the main concepts. Besides these introductions, the main properties and uses of distances are given, within items, only exceptionally. We also tried, when it was easy, to trace distances to their originator(s), but the proposed extensive bibliography has a less general ambition: just to provide convenient sources for a quick search.

Each chapter consists of items ordered in a way that hints of connections between them. All item titles and (with majuscules only for proper nouns) selected key terms can be traced in the large subject index; they are boldfaced unless the meaning is clear from the context. So, the definitions are easy to locate, by subject, in chapters and/or, by alphabetic order, in the subject index.

[redacted]

The introductions and definitions are reader-friendly and maximally independent each from another; still they are interconnected, in the 3D HTML manner, by hyperlink-like boldfaced references to similar definitions.

Many nice curiosities appear in this "Who Is Who" of distances. Examples of such sundry terms are ubiquitous Euclidean distance ("as-the-crow-flies"), flower-shop metric (shortest way between two points, visiting a "flower-shop" point first), knight-move metric on a chessboard, Gordian distance of knots, Earth mover's distance, biotope distance, Procrustes distance, lift metric, post office metric, Internet hop metric, WWW hyperlink quasi-metric, Moscow metric, and dogkeeper distance.

Besides abstract distances, the distances having physical meaning appear also (especially in Part VI); they range from 1.6×10^{-35} m (Planck length) to 8.8×10^{26} m (the estimated size of the observable Universe, about 5.4×10^{61} Planck lengths).

The number of distance metrics is infinite, and therefore our dictionary cannot enumerate all of them. But we were inspired by several successful thematic dictionaries on other infinite lists, for example, on numbers, integer sequences, inequalities, and random processes, and by atlases of functions, groups, fullerenes, etc. On the other hand, the large scope often forced us to switch to the mode of laconic tutorial.

The target audience consists of all researchers working on some measuring schemes and, to a certain degree, students and a part of the general public interested in science.

We tried to address all scientific uses of the notion of distance. But some distances did not made it to this dictionary due to space limitations (being too specific and/or complex) or our oversight. In general, the size/interdisciplinarity cut-off, i.e., decision where to stop, was our main headache. We would be grateful to the readers who will send us their favorite distances missed here.

Contents

Part I Mathematics of Distances

1 General Definitions ... 3
1.1 Basic Definitions ... 3
1.2 Main Distance-Related Notions ... 12
1.3 Metric Numerical Invariants ... 23
1.4 Main Mappings of Metric Spaces ... 36
1.5 General Distances ... 46

2 Topological Spaces ... 63

3 Generalizations of Metric Spaces ... 71
3.1 m-Tuple Generalizations of Metrics ... 71
3.2 Indefinite Metrics ... 73
3.3 Topological Generalizations ... 75
3.4 Beyond Numbers ... 78

4 Metric Transforms ... 87
4.1 Metrics on the Same Set ... 87
4.2 Metrics on Set Extensions ... 91
4.3 Metrics on Other Sets ... 94

5 Metrics on Normed Structures ... 97

Part II Geometry and Distances

6 Distances in Geometry ... 111
6.1 Geodesic Geometry ... 111
6.2 Projective Geometry ... 117
6.3 Affine Geometry ... 123
6.4 Non-Euclidean Geometry ... 125

7 Riemannian and Hermitian Metrics 135
- 7.1 Riemannian Metrics and Generalizations 136
- 7.2 Riemannian Metrics in Information Theory 152
- 7.3 Hermitian Metrics and Generalizations 156

8 Distances on Surfaces and Knots 169
- 8.1 General Surface Metrics 169
- 8.2 Intrinsic Metrics on Surfaces 175
- 8.3 Distances on Knots 180

9 Distances on Convex Bodies, Cones, and Simplicial Complexes 183
- 9.1 Distances on Convex Bodies 183
- 9.2 Distances on Cones 189
- 9.3 Distances on Simplicial Complexes 192

Part III Distances in Classical Mathematics

10 Distances in Algebra 199
- 10.1 Group Metrics 199
- 10.2 Metrics on Binary Relations 209
- 10.3 Metrics on Semilattices 210

11 Distances on Strings and Permutations 215
- 11.1 Distances on General Strings 216
- 11.2 Distances on Permutations 225

12 Distances on Numbers, Polynomials, and Matrices 229
- 12.1 Metrics on Numbers 229
- 12.2 Metrics on Polynomials 235
- 12.3 Metrics on Matrices 237

13 Distances in Functional Analysis 247
- 13.1 Metrics on Function Spaces 247
- 13.2 Metrics on Linear Operators 254

14 Distances in Probability Theory 259
- 14.1 Distances on Random Variables 260
- 14.2 Distances on Distribution Laws 261

Part IV Distances in Applied Mathematics

15 Distances in Graph Theory 277
- 15.1 Distances on the Vertices of a Graph 278
- 15.2 Distance-Defined Graphs 286
- 15.3 Distances on Graphs 297
- 15.4 Distances on Trees 305

[redacted]

16 Distances in Coding Theory 313
16.1 Minimum Distance and Relatives 315
16.2 Main Coding Distances 317

17 Distances and Similarities in Data Analysis 327
17.1 Similarities and Distances for Numerical Data 329
17.2 Relatives of Euclidean Distance 333
17.3 Similarities and Distances for Binary Data 337
17.4 Correlation Similarities and Distances 340

18 Distances in Systems and Mathematical Engineering 347
18.1 Distances in State Transition and Dynamical Systems 348
18.2 Distances in Control Theory 354
18.3 Motion Planning Distances 356
18.4 MOEA Distances 361

Part V Computer-Related Distances

19 Distances on Real and Digital Planes 365
19.1 Metrics on Real Plane 365
19.2 Digital Metrics 376

20 Voronoi Diagram Distances 383
20.1 Classical Voronoi Generation Distances 384
20.2 Plane Voronoi Generation Distances 386
20.3 Other Voronoi Generation Distances 390

21 Image and Audio Distances 393
21.1 Image Distances 393
21.2 Audio Distances 409

22 Distances in Networks 419
22.1 Scale-Free Networks 419
22.2 Network-Based Semantic Distances 424
22.3 Distances in Internet and Web 428

Part VI Distances in Natural Sciences

23 Distances in Biology 439
23.1 Genetic Distances 443
23.2 Distances for DNA/RNA and Protein Data 455
23.3 Distances in Ecology, Biogeography, Ethology 467
23.4 Other Biological Distances 484

24 Distances in Physics and Chemistry 499
24.1 Distances in Physics 499
24.2 Distances in Chemistry and Crystallography 524

25 Distances in Earth Science and Astronomy 533
25.1 Distances in Geography 533
25.2 Distances in Geophysics 546
25.3 Distances in Astronomy 558

26 Distances in Cosmology and Theory of Relativity 577
26.1 Distances in Cosmology 577
26.2 Distances in Theory of Relativity 587

Part VII Real-World Distances

27 Length Measures and Scales 613
27.1 Length Scales 613
27.2 Orders of Magnitude for Length 623

28 Distances in Applied Social Sciences 629
28.1 Distances in Perception and Psychology 629
28.2 Distances in Economics and Human Geography 640
28.3 Distances in Sociology and Language 653
28.4 Distances in Philosophy, Religion and Art 664

29 Other Distances 683
29.1 Distances in Medicine, Anthropometry and Sport 683
29.2 Equipment Distances 701
29.3 Miscellany 716

References 723

Index 731

[redacted]

Part I
Mathematics of Distances

Chapter 1
General Definitions

1.1 Basic Definitions

- **Distance**

  A **distance space** (X, d) is a set X (*carrier*) equipped with a **distance** d.

  A function $d : X \times X \to \mathbb{R}$ is called a **distance** (or **dissimilarity**) on X if, for all $x, y \in X$, it holds:

  1. $d(x, y) \geq 0$ (*nonnegativity*);
  2. $d(x, y) = d(y, x)$ (*symmetry*);
  3. $d(x, x) = 0$ (*reflexivity*).

  In Topology, a distance with $d(x, y) = 0$ implying $x = y$ is called a **symmetric**.

  For any distance d, the function D_1 defined for $x \neq y$ by $D_1(x, y) = d(x, y) + c$, where $c = \max_{x,y,z \in X}(d(x, y) - d(x, z) - d(y, z))$, and $D(x, x) = 0$, is a **metric**. Also, $D_2(x, y) = d(x, y)^c$ is a metric for sufficiently small $c \geq 0$.

  The function $D_3(x, y) = \inf \sum_i d(z_i, z_{i+1})$, where the infimum is taken over all sequences $x = z_0, \dots, z_{n+1} = y$, is the **path semimetric** of the complete weighted graph on X, where, for any $x, y \in X$, the weight of edge xy is $d(x, y)$.
- **Similarity**

  Let X be a set. A function $s : X \times X \to \mathbb{R}$ is called a **similarity** on X if s is nonnegative, symmetric and the inequality

  $$s(x, y) \leq s(x, x)$$

  holds for all $x, y \in X$, with equality if and only if $x = y$.

  The main transforms used to obtain a distance (dissimilarity) d from a similarity s bounded by 1 from above are: $d = 1 - s$, $d = \frac{1-s}{s}$, $d = \sqrt{1-s}$, $d = \sqrt{2(1-s^2)}$, $d = \arccos s$, $d = -\ln s$ (cf. Chap. 4).

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_1

[redacted]

- **Semimetric**

  Let X be a set. A function $d : X \times X \to \mathbb{R}$ is called a **semimetric** on X if d is nonnegative, symmetric, *reflexive* ($d(x, x) = 0$ for $x \in X$) and it holds

  $$d(x, y) \leq d(x, z) + d(z, y)$$

  for all $x, y, z \in X$ (**triangle inequality** or, sometimes, *triangular inequality*).

  In Topology, it is called a **pseudo-metric** (or, rarely, **semidistance**, *gauge*), while the term *semimetric* is sometimes used for a **symmetric** (a distance $d(x, y)$ with $d(x, y) = 0$ only if $x = y$); cf. **symmetrizable space** in Chap. 2.

  For a semimetric d, the triangle inequality is equivalent, for each fixed $n \geq 4$ and all $x, y, z_1, \dots, z_{n-2} \in X$, to the following *n-gon inequality*

  $$d(x, y) \leq d(x, z_1) + d(z_1, z_2) + \cdots + d(z_{n-2}, y).$$

  Equivalent *rectangle inequality* is $|d(x, y) - d(z_1, z_2)| \leq d(x, z_1) + +d(y, z_2)$.

  For a semimetric d on X, define an equivalence relation, called **metric identification**, by $x \sim y$ if $d(x, y) = 0$; equivalent points are equidistant from all other points. Let $[x]$ denote the equivalence class containing x; then $D([x], [y]) = d(x, y)$ is a **metric** on the set $\{[x] : x \in X\}$ of equivalence classes.

- **Metric**

  Let X be a set. A function $d : X \times X \to \mathbb{R}$ is called a **metric** on X if, for all $x, y, z \in X$, it holds:

  1. $d(x, y) \geq 0$ (*nonnegativity*);
  2. $d(x, y) = 0$ if and only if $x = y$ (*identity of indiscernibles*);
  3. $d(x, y) = d(y, x)$ (*symmetry*);
  4. $d(x, y) \leq d(x, z) + d(z, y)$ (**triangle inequality**).

  In fact, the above condition 1. follows from above 2., 3. and 4.

  If 2. is dropped, then d is called (Bukatin, 2002) **relaxed semimetric**. If 2. is weakened to "$d(x, x) = d(x, y) = d(y, y)$ implies $x = y$", then d is called **relaxed metric**. A **partial metric** is a **partial semimetric**, which is a relaxed metric.

  If above 2. is weakened to "$d(x, y) = 0$ implies $x = y$", then d is called (Amini-Harandi, 2012) **metric-like function**. Any **partial metric** is metric-like.

- **Metric space**

  A **metric space** (X, d) is a set X equipped with a metric d.

  It is called a **metric frame** (or *metric scheme*, *integral*) if d is integer-valued.

  A **pointed metric space** (or *rooted metric space*) (X, d, x_0) is a metric space (X, d) with a selected base point $x_0 \in X$.

- **Extended metric**

  An **extended metric** is a generalization of the notion of metric: the value ∞ is allowed for a metric d.

- **Quasi-distance**

  Let X be a set. A function $d : X \times X \to \mathbb{R}$ is called a **quasi-distance** on X if d is nonnegative, and $d(x, x) = 0$ holds for all $x \in X$. It is also called a **premetric** or **prametric** in Topology and a *divergence* in Probability.

If a quasi-distance d satisfies the **strong triangle inequality** $d(x, y) \leq d(x, z) + d(y, z)$, then (Lindenbaum, 1926) it is symmetric and so, a semimetric. A **quasi-semimetric** d is a semimetric if and only if (Weiss, 2012) it satisfies the **full triangle inequality** $|d(x, z) - d(z, y)| \leq d(x, z) \leq d(x, z) + d(z, y)$.

The distance/metric notions are usually named as weakenings or modifications of the fundamental notion of **metric**, using various prefixes and modifiers. But, perhaps, extended (i.e., the value ∞ is allowed) semimetric and quasi-semimetric should be (as suggested in Lawvere, 2002) used as the basic terms, since, together with their **short mappings**, they are best behaved of the **metric space categories**.

- **Quasi-semimetric**

A function $d : X \times X \to \mathbb{R}$ is called a **quasi-semimetric** (or **hemimetric**, *ostensible metric*) on X if $d(x, x) = 0$, $d(x, y) \geq 0$ and the **oriented triangle inequality**

$$d(x, y) \leq d(x, z) + d(z, y)$$

holds for all $x, y, z \in X$. The set X can be partially ordered by the *specialization order*: $x \preceq y$ if and only if $d(x, y) = 0$.

A **weak quasi-metric** is a quasi-semimetric d on X with *weak symmetry*, i.e., for all $x, y \in X$ the equality $d(x, y) = 0$ implies $d(y, x) = 0$.

An **Albert quasi-metric** is a quasi-semimetric d on X with *weak definiteness*, i.e., for all $x, y \in X$ the equality $d(x, y) = d(y, x) = 0$ implies $x = y$.

Both, weak and Albert, quasi-metric, is a usual **quasi-metric**.

Any *pre-order* $(X, \prec)$ (satisfying for all $x, y, z \in X$, $x \prec x$ and if $x \prec y$ and $y \prec z$ then $x \prec z$) can be viewed as a **pre-order extended quasi-semimetric** (X, d) by defining $d(x, y) = 0$ if $x \prec y$ and $d(x, y) = \infty$, otherwise.

A **weightable quasi-semimetric** is a quasi-semimetric d on X with *relaxed symmetry*, i.e., for all $x, y, z \in X$

$$d(x, y) + d(y, z) + d(z, x) = d(x, z) + d(z, y) + d(y, x),$$

holds or, equivalently, there exists a weight function $w(x) \in \mathbb{R}$ on X with $d(x, y) - d(y, x) = w(y) - w(x)$ for all $x, y \in X$ (i.e., $d(x, y) + \frac{1}{2}(w(x) - w(y))$ is a semimetric). If d is a weightable quasi-semimetric, then $d(x, y) + w(x)$ is a **partial semimetric** (moreover, a **partial metric** if d is an Albert quasi-metric).

- **Partial metric**

Let X be a set. A nonnegative symmetric function $p : X \times X \to \mathbb{R}$ is called a **partial metric** ([Matt92]) if, for all $x, y, z \in X$, it holds:

1. $p(x, x) \leq p(x, y)$, i.e., every **self-distance** (or *extent*) $p(x, x)$ is *small*;
2. $x = y$ if $p(x, x) = p(x, y) = p(y, y) = 0$ (T_0 *separation axiom*);
3. $p(x, y) \leq p(x, z) + p(z, y) - p(z, z)$ (**sharp triangle inequality**).

The 1-st above condition means that p is a **forward resemblance**, cf. Chap. 3.

[redacted]

If the 2-nd above condition is dropped, the function p is called a **partial semimetric**. The nonnegative function p is a partial semimetric if and only if $p(x,y) - p(x,x)$ is a **weightable quasi-semimetric** with $w(x) = p(x,x)$.

If the 1-st above condition is also dropped, the function p is called (Heckmann, 1999) a **weak partial semimetric**. The nonnegative function p is a weak partial semimetric if and only if $2p(x,y) - p(x,x) - p(y,y)$ is a semimetric.

Sometimes, the term *partial metric* is used when a metric $d(x,y)$ is defined only on a subset of the set of all pairs x, y of points.

- **Protometric**

  A function $p : X \times X \to \mathbb{R}$ is called a **protometric** if, for all (equivalently, for all different) $x, y, z \in X$, the **sharp triangle inequality** holds:

$$p(x,y) \le p(x,z) + p(z,y) - p(z,z).$$

For finite X, the matrix $((p(x,y)))$ is (Burkard et al., 1996) *weak Monge array*.

A **strong protometric** is a protometric p with $p(x,x) = 0$ for all $x \in X$. Such a protometric is exactly a quasi-semimetric, but with the condition $p(x,y) \ge 0$ (for any $x, y \in X$) being relaxed to $p(x,y) + p(y,x) \ge 0$.

A **partial semimetric** is a **symmetric protometric** (i.e., $p(x,y) = p(y,x)$ with $p(x,y) \ge p(x,x) \ge 0$ for all $x, y \in X$.) An example of a nonpositive symmetric protometric is given by $p(x,y) = -(x.y)_{x_0} = \frac{1}{2}(d(x,y) - d(x,x_0) - d(y,y_0))$, where (X,d) is a metric space with a fixed base point $x_0 \in X$; see **Gromov product similarity** $(x.y)_{x_0}$ and, in Chap. 4, **Farris transform metric** $C - (x.y)_{x_0}$.

A 0**-protometric** is a protometric p for which all sharp triangle inequalities (equivalently, all inequalities $p(x,y) + p(y,x) \ge p(x,x) + p(y,y)$ implied by them) hold as equalities. For any $u \in X$, denote by A'_u, A''_u the 0-protometrics p with $p(x,y) = 1_{x=u}, 1_{y=u}$, respectively. The protometrics on X form a flat convex cone in which the 0-protometrics form the largest linear space. For finite X, a basis of this space is given by all but one A'_u, A''_u (since $\sum_u A'_u = \sum_u A''_u$) and, for the flat subcone of all symmetric 0-protometrics on X, by all $A'_u + A''_u$.

A **weighted protometric** on X is a protometric with a point-weight function $w : X \to \mathbb{R}$. The mappings $p(x,y) = \frac{1}{2}(d(x,y) + w(x) + w(y))$ and $d(x,y) = 2p(x,y) - p(x,x) - p(y,y)$, $w(x) = p(x,x)$ establish a bijection between the weighted strong protometrics (d, w) and the protometrics p on X, as well as between the weighted semimetrics and the symmetric protometrics. For example, a weighted semimetric (d, w) with $w(x) = -d(x, x_0)$ corresponds to a protometric $-(x.y)_{x_0}$. For finite $|X|$, the above mappings amount to the representation

$$2p = d + \sum_{u \in X} p(u,u)(A'_u + A''_u).$$

- **Quasi-metric**

  A function $d : X \times X \to \mathbb{R}$ is called a **quasi-metric** (or *asymmetric metric*, *directed metric*) on X if $d(x,y) \ge 0$ holds for all $x, y \in X$ with equality if and only if $x = y$, and for all $x, y, z \in X$ the **oriented triangle inequality**

$$d(x, y) \leq d(x, z) + d(z, y)$$

holds. A *quasi-metric space* (X, d) is a set X equipped with a quasi-metric d.

For any quasi-metric d, the functions $\max\{d(x, y), d(y, x)\}$ (called sometimes *bi-distance*), $\min\{d(x, y), d(y, x)\}$, $\frac{1}{2}(d^p(x, y) + d^p(y, x))^{\frac{1}{p}}$ with given $p \geq 1$ are **metric generating**; cf. Chap. 4.

A **non-Archimedean quasi-metric** d is a quasi-distance on X which, for all $x, y, z \in X$, satisfies the following strengthened oriented triangle inequality:

$$d(x, y) \leq \max\{d(x, z), d(z, y)\}.$$

- **Directed-metric**

  Let X be a set. A function $d : X \times X \to \mathbb{R}$ is called (Jegede, 2005) a **directed-metric** on X if, for all $x, y, z \in X$, it holds $d(x, y) = -d(y, x)$ and

  $$|d(x, y)| \leq |d(x, z)| + |d(z, y)|.$$

  Cf. **displacement** in Chap. 24 and **rigid motion of metric space**.
- **Coarse-path metric**

  Let X be a set. A metric d on X is called a **coarse-path metric** if, for a fixed $C \geq 0$ and for every pair of points $x, y \in X$, there exists a sequence $x = x_0, x_1, \dots, x_t = y$ for which $d(x_{i-1}, x_i) \leq C$ for $i = 1, \dots, t$, and it holds

  $$d(x, y) \geq d(x_0, x_1) + d(x_1, x_2) + \cdots + d(x_{t-1}, x_t) - C.$$

- **Near-metric**

  Let X be a set. A distance d on X is called a **near-metric** (or *C-near-metric*) if $d(x, y) > 0$ for $x \neq y$ and the *C-relaxed triangle inequality*

  $$d(x, y) \leq C(d(x, z) + d(z, y))$$

  holds for all $x, y, z \in X$ and some constant $C \geq 1$.

  If $d(x, y) > 0$ for $x \neq y$ and the *C-asymmetric triangle inequality* $d(x, y) \leq d(x, z) + Cd(z, y)$ holds, d is a $\frac{C+1}{2}$-near-metric.

  A **C-inframetric** is a C-near-metric, while a C-near-metric is a $2C$-inframetric.

  Some recent papers use the term *quasi-triangle inequality* for the above inequality and so, *quasi-metric* for the notion of near-metric.

  The **power transform** (Chap. 4) $(d(x, y))^\alpha$ of any near-metric is a near-metric for any $\alpha > 0$. Also, any near-metric d admits a **bi-Lipschitz mapping** on $(D(x, y))^\alpha$ for some semimetric D on the same set and a positive number α.

[redacted]

A near-metric d on X is called a **Hölder near-metric** if the inequality

$$|d(x,y) - d(x,z)| \leq \beta d^{\alpha}(y,z)(d(x,y) + d(x,z))^{1-\alpha}$$

holds for some $\beta > 0$, $0 < \alpha \leq 1$ and all $x, y, z \in X$. Cf. **Hölder mapping**.

A distance d on set X is said (Greenhoe, 2015) to satisfy (C, p) **power triangle inequality** if, for given positive C, p and any $x, y, z \in X$, it holds

$$d(x,y) \leq 2C|\frac{1}{2}d^{p}(x,z) + \frac{1}{2}d^{p}(z,y)|^{\frac{1}{p}}.$$

- ***f*-quasi-metric**

  Let $f(t,t') : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a function with $\lim_{(t,t') \to 0}, f(t,t') = f(0,0) = 0$.

  Let X be a set. A function $d : X \times X \to \mathbb{R}$ is called (Arutyunov et al., 2016) a ***f*-quasi-metric** on X if $d(x,y) \geq 0$ holds for all $x, y \in X$ with equality if and only if $x = y$, and for all $x, y, z \in X$ holds the ***f*-triangle inequality**

  $$d(x,y) \leq f(d(x,z), d(z,y)).$$

  The *f-quasi-metric space* (X, d) with symmetric d and $f(t,t') = \max(t,t')$ is exactly the **Fréchet *V*-space** (1906); cf. the **partially ordered distance** in Sect. 3.4.

  The case $f(t,t') = t + t'$ of a f-quasi-metric corresponds to a **quasi-metric**. Given $q, q' \geq 1$, the f-quasi-metric with $f(t,t') = qt + q't'$ is called (q,q')*-quasi-metric*.

  The inequality $d(x,y) \leq F(d(x,z), d(y,z))$ implies $d(x,y) \leq f(d(x,z), d(z,y))$ for the function $f(t,t') = F(t, F(0,t'))$.

- **Weak ultrametric**

  A **weak ultrametric** (or ***C*-inframetric**, *C-pseudo-distance*) d is a distance on X such that $d(x,y) > 0$ for $x \neq y$ and the *C-inframetric inequality*

  $$d(x,y) \leq C \max\{d(x,z), d(z,y)\}$$

  holds for all $x, y, z \in X$ and some constant $C \geq 1$.

  The term **pseudo-distance** is also used, in some applications, for any of a **pseudo-metric**, a **quasi-distance**, a **near-metric**, a distance which can be infinite, a distance with an error, etc. Another unsettled term is **weak metric**: it is used for both a **near-metric** and a **quasi-semimetric**.

- **Ultrametric**

  An **ultrametric** (or *non-Archimedean metric*) is (Krasner, 1944) a metric d on X which satisfies, for all $x, y, z \in X$, the following strengthened version of the triangle inequality (Hausdorff, 1934), called the **ultrametric inequality**:

  $$d(x,y) \leq \max\{d(x,z), d(z,y)\}$$

An ultrametric space is also called an *isosceles space* since at least two of $d(x, y)$, $d(z, y)$, $d(x, z)$ are equal. An ultrametric on a set V has at most $|V|$ values.

A metric d is an ultrametric if and only if its **power transform** (see Chap. 4) d^{α} is a metric for any real positive number α. Any ultrametric satisfies the **four-point inequality**. A metric d is an ultrametric if and only if it is a **Farris transform metric** (Chap. 4) of a **four-point inequality metric**.

- **Robinsonian distance**

  A distance d on X is called a **Robinsonian distance** (or *monotone distance*) if there exists a total order $\preceq$ on X *compatible* with it, i.e., for $x, y, w, z \in X$,

  $$x \preceq y \preceq w \preceq z \text{ implies } d(y, w) \leq d(x, z),$$

  or, equivalently, for $x, y, z \in X$, it holds

  $$x \preceq y \preceq z \text{ implies } d(x, y) \leq \max\{d(x, z), d(z, y)\}.$$

  Any **ultrametric** is a Robinsonian distance.

- **Four-point inequality metric**

  A metric d on X is a **four-point inequality metric** (or **additive metric**) if it satisfies the following strengthened version of the triangle inequality called the **four-point inequality** (Buneman, 1974): for all $x, y, z, u \in X$

  $$d(x, y) + d(z, u) \leq \max\{d(x, z) + d(y, u), d(x, u) + d(y, z)\}$$

  holds. Equivalently, among the three sums $d(x, y) + d(z, u)$, $d(x, z) + d(y, u)$, $d(x, u) + d(y, z)$ the two largest sums are equal.

  A metric satisfies the four-point inequality if and only if it is a **tree-like metric**.

  Any metric, satisfying the four-point inequality, is a **Ptolemaic metric** and an L_1*-metric*. Cf. L_p**-metric** in Chap. 5.

  A **bush metric** is a metric for which all four-point inequalities are equalities, i.e., $d(x, y) + d(u, z) = d(x, u) + d(y, z)$ holds for any $u, x, y, z \in X$.

- **Relaxed four-point inequality metric**

  A metric d on X satisfies the **relaxed four-point inequality** if, for all $x, y, z, u \in X$, among the three sums

  $$d(x, y) + d(z, u), d(x, z) + d(y, u), d(x, u) + d(y, z)$$

  at least two (not necessarily the two largest) are equal. A metric satisfies this inequality if and only if it is a **relaxed tree-like metric**.

- **Ptolemaic metric**

  A **Ptolemaic metric** d is a metric on X which satisfies the **Ptolemaic inequality**

  $$d(x, y)d(u, z) \leq d(x, u)d(y, z) + d(x, z)d(y, u)$$

[redacted]

for all $x, y, u, z \in X$. A classical result, attributed to Ptolemy, says that this inequality holds in the Euclidean plane, with equality if and only if the points x, y, u, z lie on a circle in that order.

A *Ptolemaic space* is a *normed vector space* $(V, ||.||)$ such that its norm metric $||x - y||$ is a Ptolemaic metric. A normed vector space is a Ptolemaic space if and only if it is an **inner product space** (Chap. 5); so, a **Minkowskian metric** (Chap. 6) is Euclidean if and only if it is Ptolemaic.

For any metric d, the metric $\sqrt{d}$ is Ptolemaic ([FoSc06]).

- **δ-hyperbolic metric**

  Given a number $\delta \geq 0$, a metric d on a set X is called **δ-hyperbolic** if it satisfies the following *Gromov δ-hyperbolic inequality* (another weakening of the **four-point inequality**): for all $x, y, z, u \in X$, it holds that

  $$d(x,y) + d(z,u) \leq 2\delta + \max\{d(x,z) + d(y,u), d(x,u) + d(y,z)\}.$$

  A metric space (X, d) is δ-hyperbolic if and only if for all $x_0, x, y, z \in X$ it holds

  $$(x.y)_{x_0} \geq \min\{(x.z)_{x_0}, (y.z)_{x_0}\} - \delta,$$

  where $(x.y)_{x_0} = \frac{1}{2}(d(x_0, x) + d(x_0, y) - d(x, y))$ is the **Gromov product** of the points x and y of X with respect to the base point $x_0 \in X$.

  A metric space (X, d) is 0-hyperbolic exactly when d satisfies the **four-point inequality**. Every bounded metric space of diameter D is D-hyperbolic. The n-dimensional *hyperbolic space* is $\ln 3$-hyperbolic.

  Every δ-hyperbolic metric space is isometrically embeddable into a **geodesic metric space** (Bonk and Schramm, 2000).

- **Gromov product similarity**

  Given a metric space (X, d) with a fixed point $x_0 \in X$, the **Gromov product similarity** (or *Gromov product*, *covariance*, *overlap function*) $(.)_{x_0}$ is a similarity on X defined by

  $$(x.y)_{x_0} = \frac{1}{2}(d(x, x_0) + d(y, x_0) - d(x, y)).$$

  The triangle inequality for d implies $(x.y)_{x_0} \geq (x.z)_{x_0} + (y.z)_{x_0} - (z.z)_{x_0}$ (**covariance triangle inequality**), i.e., **sharp triangle inequality** for **protometric** $-(x.y)_{x_0}$.

  If (X, d) is a tree, then $(x.y)_{x_0} = d(x_0, [x, y])$. If (X, d) is a **measure semimetric space**, i.e., $d(x, y) = \mu(x \triangle y)$ for a Borel measure μ on X, then $(x.y)_{\emptyset} = \mu(x \cap y)$. If d is a **distance of negative type**, i.e., $d(x, y) = d_E^2(x, y)$ for a subset X of a Euclidean space $\mathbb{E}^n$, then $(x.y)_0$ is the usual *inner product* on $\mathbb{E}^n$.

  Cf. **Farris transform metric** $d_{x_0}(x, y) = C - (x.y)_{x_0}$ in Chap. 4.

- **Cross-difference**

  Given a metric space (X, d) and quadruple (x, y, z, w) of its points, the **cross-difference** is the real number cd defined by

  $$cd(x, y, z, w) = d(x, y) + d(z, w) - d(x, z) - d(y, w).$$

In terms of the **Gromov product similarity**, for all $x, y, z, w, p \in X$, it holds

$$\frac{1}{2} cd(x, y, z, w) = -(x.y)_p - (z.w)_p + (x.z)_p + (y.w)_p;$$

in particular, it becomes $(x.y)_p$ if $y = w = p$.

If $x \neq z$ and $y \neq w$, the **cross-ratio** is the positive number defined by

$$cr((x, y, z, w), d) = \frac{d(x, y)d(z, w)}{d(x, z)d(y, w)}.$$

- **$2k$-gonal distance**

  A **$2k$-gonal distance** d is a distance on X which satisfies, for all distinct elements $x_1, \dots, x_n \in X$, the **$2k$-gonal inequality**

  $$\sum_{1 \leq i < j \leq n} b_i b_j d(x_i, x_j) \leq 0$$

  for all $b \in \mathbb{Z}^n$ with $\sum_{i=1}^n b_i = 0$ and $\sum_{i=1}^n |b_i| = 2k$.
- **Distance of negative type**

  A **distance of negative type** d is a distance on X which is **$2k$-gonal** for any $k \geq 1$, i.e., satisfies the **negative type inequality**

  $$\sum_{1 \leq i < j \leq n} b_i b_j d(x_i, x_j) \leq 0$$

  for all $b \in \mathbb{Z}^n$ with $\sum_{i=1}^n b_i = 0$, and for all distinct elements $x_1, \dots, x_n \in X$.

  A distance can be of negative type without being a semimetric. Cayley proved that a metric d is an **L_2-metric** if and only if d^2 is a distance of negative type.
- **$(2k+1)$-gonal distance**

  A **$(2k+1)$-gonal distance** d is a distance on X which satisfies, for all distinct elements $x_1, \dots, x_n \in X$, the **$(2k+1)$-gonal inequality**

  $$\sum_{1 \leq i < j \leq n} b_i b_j d(x_i, x_j) \leq 0$$

  for all $b \in \mathbb{Z}^n$ with $\sum_{i=1}^n b_i = 1$ and $\sum_{i=1}^n |b_i| = 2k + 1$.

  The $(2k+1)$-gonal inequality with $k = 1$ is the usual triangle inequality. The $(2k+1)$-gonal inequality implies the **$2k$-gonal inequality**.
- **Hypermetric**

  A **hypermetric** d is a distance on X which is **$(2k+1)$-gonal** for any $k \geq 1$, i.e., satisfies the **hypermetric inequality** (Deza, 1960)

  $$\sum_{1 \leq i < j \leq n} b_i b_j d(x_i, x_j) \leq 0$$

  for all $b \in \mathbb{Z}^n$ with $\sum_{i=1}^n b_i = 1$, and for all distinct elements $x_1, \dots, x_n \in X$.

[redacted]

Any hypermetric is a semimetric, a **distance of negative type** and, moreover, it can be isometrically embedded into some n-sphere $\mathbb{S}^n$ with squared Euclidean distance. Any L_1-*metric* (cf. L_p**-metric** in Chap. 5) is a hypermetric.

- ***P*-metric**

  A P**-metric** d is a metric on X with values in $[0, 1]$ which satisfies the **correlation triangle inequality**

  $$d(x, y) \leq d(x, z) + d(z, y) - d(x, z)d(z, y).$$

  The equivalent inequality $1 - d(x, y) \geq (1 - d(x, z))(1 - d(z, y))$ expresses that the probability, say, to reach x from y via z is either equal to $(1 - d(x, z))(1 - d(z, y))$ (independence of reaching z from x and y from z), or greater than it (positive correlation). A metric is a P-metric if and only if it is a **Schoenberg transform metric** (Chap. 4).

1.2 Main Distance-Related Notions

- **Metric ball**

  Given a metric space (X, d), the **metric ball** (or *closed metric ball*) with center $x_0 \in X$ and radius $r > 0$ is defined by $\overline{B}(x_0, r) = \{x \in X : d(x_0, x) \leq r\}$, and the **open metric ball** with center $x_0 \in X$ and radius $r > 0$ is defined by $B(x_0, r) = \{x \in X : d(x_0, x) < r\}$. The closed ball is a subset of the closure of the open ball; it is a proper subset for, say, the **discrete metric** on X.

  The **metric sphere** with center $x_0 \in X$ and radius $r > 0$ is defined by $S(x_0, r) = \{x \in X : d(x_0, x) = r\}$.

  For the **norm metric** on an n-dimensional *normed vector space* $(V, ||.||)$, the metric ball $\overline{B}^n = \{x \in V : ||x|| \leq 1\}$ is called the *unit ball*, and the set $S^{n-1} = \{x \in V : ||x|| = 1\}$ is called the *unit sphere*. In a two-dimensional vector space, a metric ball (closed or open) is called a **metric disk** (closed or open, respectively).

- **Metric hull**

  Given a metric space (X, d), let M be a **bounded** subset of X.

  The **metric hull** $H(M)$ of M is the intersection of all metric balls containing M.

  The set of *surface points* $S(M)$ of M is the set of all $x \in H(M)$ such that x lies on the sphere of one of the metric balls containing M.

- **Distance-invariant metric space**

  A metric space (X, d) is **distance-invariant** if all **metric balls** $\overline{B}(x_0, r) = \{x \in X : d(x_0, x) \leq r\}$ of the same radius have the same number of elements.

  Then the **growth rate of a metric space** (X, d) is the function $f(n) = |\overline{B}(x, n)|$.

(X, d) is a *metric space of polynomial growth* if there are some positive constants k, C such that $f(n) \leq Cn^k$ for all $n \geq 0$. Cf. **graph of polynomial growth**, including the group case, in Chap. 15.

For a **metrically discrete metric space** (X, d) (i.e., with $a = \inf_{x,y \in X, x \neq y} d(x, y) > 0$), its *growth rate* was defined also (Gordon–Linial–Rabinovich, 1998) by

$$\max_{x \in X, r \geq 2} \frac{\log |\overline{B}(x, ar)|}{\log r}.$$

- **Ahlfors q-regular metric space**

  A metric space (X, d) endowed with a Borel measure μ is called **Ahlfors q-regular** if there exists a constant $C \geq 1$ such that for every ball in (X, d) with radius $r < diam(X, d)$ it holds

  $$C^{-1} r^q \leq \mu(\overline{B}(x_0, r)) \leq Cr^q.$$

  If such an (X, d) is locally compact, then the **Hausdorff q-measure** can be taken as μ and q is the **Hausdorff dimension**. For two disjoint **continua** (nonempty **connected compact metric subspaces**) C_1, C_2 of such space (X, d), let Γ be the set of rectifiable curves connecting C_1 to C_2. The *q-modulus* between C_1 and C_2 is $M_q(C_1, C_2) = \inf\{\int_X \rho^q : \inf_{\gamma \in \Gamma} \int_\gamma \rho \geq 1\}$, where $\rho : X \to \mathbb{R}_{>0}$ is any density function on X; cf. the **modulus metric** in Chap. 6.

  The *relative distance* between C_1 and C_2 is $\delta(C_1, C_2) = \frac{\inf\{d(p_1, p_2) : p_1 \in C_1, p_2 \in C_2\}}{\min\{diam(C_1), diam(C_2)\}}$. (X, d) is a **q-Loewner space** if there are increasing functions $f, g : [0, \infty) \to [0, \infty)$ such that for all C_1, C_2 it holds $f(\delta(C_1, C_2)) \leq M_q(C_1, C_2) \leq g(\delta(C_1, C_2))$.
- **Connected metric space**

  A metric space (X, d) is called **connected** if it cannot be partitioned into two nonempty **open** sets. Cf. **connected space** in Chap. 2.

  The maximal connected subspaces of a metric space are called its *connected components*. A **totally disconnected metric space** is a space in which all connected subsets are $\emptyset$ and one-point sets.

  A **path-connected metric space** is a connected metric space such that any two its points can be joined by an **arc** (cf. **metric curve**).
- **Cantor connected metric space**

  A metric space (X, d) is called **Cantor** (or *pre-*) *connected* if, for any two its points x, y and any $\epsilon > 0$, there exists an *ϵ-chain* joining them, i.e., a sequence of points $x = z_0, z_1, \ldots, z_{n-1}, z_n = y$ such that $d(z_k, z_{k+1}) \leq \epsilon$ for every $0 \leq k \leq n$. A metric space (X, d) is Cantor connected if and only if it cannot be partitioned into two *remote parts* A and B, i.e., such that $\inf\{d(x, y) : x \in A, y \in B\} > 0$.

  The maximal Cantor connected subspaces of a metric space are called its *Cantor connected components*. A **totally Cantor disconnected metric** is the metric of a metric space in which all Cantor connected components are one-point sets.

[redacted]

- **Indivisible metric space**

  A metric space (X, d) is called **indivisible** if it cannot be partitioned into two parts, neither of which contains an isometric copy of (X, d). Any indivisible metric space with $|X| \geq 2$ is infinite, bounded and **totally Cantor disconnected** (Delhomme–Laflamme–Pouzet–Sauer, 2007).

  A metric space (X, d) is called an **oscillation stable metric space** (Nguyen Van Thé, 2006) if, given any $\epsilon > 0$ and any partition of X into finitely many pieces, the **ϵ-neighborhood** of one of the pieces includes an isometric copy of (X, d).

- **Closed subset of metric space**

  Given a subset M of a metric space (X, d), a point $x \in X$ is called a *limit* (or *accumulation*) *point of* M if any **open metric ball** $B(x, r) = \{y \in X : d(x, y) < r\}$ contains a point $x' \in M$ with $x' \neq x$. The *boundary* $\vartheta(M)$ of M is the set of all its limit points. The *closure of* M, denoted by $cl(M)$, is $M \cup \vartheta(M)$, and M is called **closed subset**, if $M = cl(M)$, and **dense subset**, if $X = cl(M)$.

  Every point of M which is not its limit point, is called an *isolated point*. The *interior* $int(M)$ of M is the set of all its isolated points, and the *exterior* $ext(M)$ of M is $int(X \backslash M)$. A subset M is called *nowhere dense* if $int(cl(M)) = \emptyset$.

  A subset M is called *topologically discrete* (cf. **metrically discrete metric space**) if $int(M) = M$ and *dense-in-itself* if $int(M) = \emptyset$. A dense-in-itself subset is called *perfect* (cf. **perfect metric space**) if it is closed. The subsets *Irr* (irrational numbers) and $\mathbb{Q}$ (rational numbers) of $\mathbb{R}$ are dense, dense-in-itself but not perfect. The set $\mathbb{Q} \cap [0, 1]$ is dense-in-itself but not dense in $\mathbb{R}$.

- **Open subset of metric space**

  A subset M of a metric space (X, d) is called *open* if, given any point $x \in M$, the **open metric ball** $B(x, r) = \{y \in X : d(x, y) < r\}$ is contained in M for some number $r > 0$. The family of open subsets of a metric space forms a natural topology on it. A **closed subset** is the complement of an open subset.

  An open subset is called *clopen*, if it is closed, and a *domain* if it is **connected**.

  A *door space* is a metric (in general, topological) space in which every subset is either open or closed.

- **Metric topology**

  A **metric topology** is a *topology* induced by a metric; cf. **equivalent metrics**. More exactly, the **metric topology** on a metric space (X, d) is the set of all *open sets* of X, i.e., arbitrary unions of (finitely or infinitely many) open metric balls $B(x, r) = \{y \in X : d(x, y) < r\}, x \in X, r \in \mathbb{R}, r > 0$.

  A topological space which can arise in this way from a metric space is called a **metrizable space** (Chap. 2). **Metrization theorems** are theorems which give sufficient conditions for a topological space to be metrizable.

  On the other hand, the adjective *metric* in several important mathematical terms indicates connection to a measure, rather than distance, for example, *metric Number Theory*, *metric Theory of Functions*, *metric transitivity*.

- **Equivalent metrics**

  Two metrics d_1 and d_2 on a set X are called **equivalent** if they define the same *topology* on X, i.e., if, for every point $x_0 \in X$, every open metric ball with center at x_0 defined with respect to d_1, contains an open metric ball with the same center but defined with respect to d_2, and conversely.

Two metrics d_1 and d_2 are equivalent if and only if, for every $\epsilon > 0$ and every $x \in X$, there exists $\delta > 0$ such that $d_1(x, y) \leq \delta$ implies $d_2(x, y) \leq \epsilon$ and, conversely, $d_2(x, y) \leq \delta$ implies $d_1(x, y) \leq \epsilon$.

All metrics on a finite set are equivalent; they generate the *discrete topology*.

- **Metric betweenness**

  The **metric betweenness** of a metric space (X, d) is (Menger, 1928) the set of all ordered triples (x, y, z) such that x, y, z are (not necessarily distinct) points of X for which the **triangle equality** $d(x, y) + d(y, z) = d(x, z)$ holds.

- **Monometric**

  A ternary relation R on a set X is called a *betweenness relation* if $(x, y, z) \in R$ if and only if $(z, y, x) \in R$ and $(x, y, z), (x, z, y) \in R$ if and only if $y = z$.

  Given a such relation R, a **monometric** is (Perez-Fernández et al., 2016) a function $d : X \times X \to \mathbb{R}_{\geq 0}$ with $d(x, y) = 0$ if and only if $x = y$ and (x, y, z) implying $d(x, y) \leq d(x, z)$. Clearly, any metric is a monometric.

  Cf. a **distance-rationalizable voting rule** in Sect. 11.2.

- **Closed metric interval**

  Given two different points $x, y \in X$ of a metric space (X, d), the **closed metric interval** between them (or *line induced by*) them is the set of the points z, for which the **triangle equality** (or **metric betweenness** (x, z, y)) holds:

  $$I(x, y) = \{z \in X : d(x, y) = d(x, z) + d(z, y)\}.$$

  Cf. **inner product space** (Chap. 5) and **cutpoint additive metric** (Chap. 15).

  Let $Ext(x, y) = \{z : y \in I(x, z) \setminus \{x, z\}\}$. A *CC-line* $CC(x, y)$ is $I(x, y) \cup Ext(x, y) \cup Ext(y, x)$. Chen–Chvátal, 2008, conjectured that every metric space on $n, n \geq 2$, points, either has at least n distinct CC-lines or consists of a unique CC-line.

- **Underlying graph of a metric space**

  The **underlying graph** (or *neighborhood graph*) of a metric space (X, d) is a graph with the vertex-set X and xy being an edge if $I(x, y) = \{x, y\}$, i.e., there is no third point $z \in X$, for which $d(x, y) = d(x, z) + d(z, y)$.

- **Distance monotone metric space**

  A metric space (X, d) is called **distance monotone** if for any its **closed metric interval** $I(x, y)$ and $u \in X \setminus I(x, y)$, there exists $z \in I(x, xy)$ with $d(u, z) > d(x, y)$.

- **Metric triangle**

  Three distinct points $x, y, z \in X$ of a metric space (X, d) form a **metric triangle** if the **closed metric intervals** $I(x, y), I(y, z)$ and $I(z, x)$ intersect only in the common endpoints.

- **Metric space having collinearity**

  A metric space (X, d) has **collinearity** if for any $\epsilon > 0$ each of its infinite subsets contains distinct *ϵ-collinear* (i.e., with $d(x, y) + d(y, z) - d(x, z) \leq \epsilon$) points x, y, z.

[redacted]

- **Modular metric space**

  A metric space (X, d) is called **modular** if, for any three different points $x, y, z \in X$, there exists a point $u \in I(x, y) \cap I(y, z) \cap I(z, x)$. This should not be confused with **modular distance** in Chap. 10 and **modulus metric** in Chap. 6.
- **Median metric space**

  A metric space (X, d) is called a **median metric space** if, for any three points $x, y, z \in X$, there exists a unique point $u \in I(x, y) \cap I(y, z) \cap I(z, x)$, or, equivalently,

  $$d(x, u) + d(y, u) + d(z, u) = \frac{1}{2}((x, y) + d(y, z) + d(z, x)).$$

  The point u is called *median for* $\{x, y, z\}$, since it minimises the sum of distances to them. Any median metric space is an L_1*-metric*; cf. L_p**-metric** in Chap. 5 and **median graph** in Chap. 15.

  A metric space (X, d) is called an **antimedian metric space** if, for any three points $x, y, z \in X$, there exists a unique point $u \in X$ maximizing $d(x, u)+d(y, u)+d(z, u)$.
- **Metric quadrangle**

  Four different points $x, y, z, u \in X$ of a metric space (X, d) form a **metric quadrangle** if $x, z \in I(y, u)$ and $y, u \in I(x, z)$; then $d(x, y) = d(z, u)$ and $d(x, u) = d(y, z)$.

  A metric space (X, d) is called *weakly spherical* if any three different points $x, y, z \in X$ with $y \in I(x, z)$, form a metric quadrangle with some point $u \in X$.
- **Metric curve**

  A **metric curve** (or, simply, *curve*) γ in a metric space (X, d) is a continuous mapping $\gamma : I \to X$ from an interval I of $\mathbb{R}$ into X. A curve is called an **arc** (or **path**, *simple curve*) if it is injective. A curve $\gamma : [a, b] \to X$ is called a *Jordan curve* (or *simple closed curve*) if it does not cross itself, and $\gamma(a) = \gamma(b)$.

  The **length of a curve** $\gamma : [a, b] \to X$ is the number $l(\gamma)$ defined by

  $$l(\gamma) = \sup\{\sum_{1 \leq i \leq n} d(\gamma(t_i), \gamma(t_{i-1})) : n \in \mathbb{N}, a = t_0 < t_1 < \cdots < t_n = b\}.$$

  A *rectifiable curve* is a curve with a finite length. A metric space (X, d), where every two points can be joined by a rectifiable curve, is called a **quasi-convex metric space** (or, specifically, C**-quasi-convex metric space**) if there exists a constant $C \geq 1$ such that every pair $x, y \in X$ can be joined by a rectifiable curve of length at most $Cd(x, y)$. If $C = 1$, then this length is equal to $d(x, y)$, i.e., (X, d) is a **geodesic metric space** (Chap. 6).

  In a quasi-convex metric space (X, d), the infimum of the lengths of all rectifiable curves, connecting $x, y \in X$ is called the **internal metric**.

  The metric d on X is called the **intrinsic metric** (and then (X, d) is called a **length space**) if it coincides with the internal metric of (X, d).

  If, moreover, any pair x, y of points can be joined by a curve of length $d(x, y)$, the metric d is called **strictly intrinsic**, and the length space (X, d) is a geodesic metric space. Hopf–Rinow, 1931, showed that any complete locally compact length space is geodesic and **proper**. The **punctured plane** $(\mathbb{R}^2 \setminus \{0\}, ||x-y||_2)$ is **locally compact** and **path-connected** but not geodesic: the distance between $(-1, 0)$ and $(1, 0)$ is 2 but there is no geodesic realizing this distance.

The **metric derivative** of a metric curve $\gamma : [a, b] \to X$ at a limit point t is

$$\lim_{s \to 0} \frac{d(\gamma(t+s), \gamma(t))}{|s|},$$

if it exists. It is the rate of change, with respect to t, of the length of the curve at almost every point, i.e., a generalization of the notion of *speed* to metric spaces.

- **Geodesic**

  Given a metric space (X, d), a **geodesic** is a locally shortest **metric curve**, i.e., it is a locally isometric embedding of $\mathbb{R}$ into X; cf. Chap. 6.

  A subset S of X is called a **geodesic segment** (or **metric segment**, *shortest* path, *minimizing geodesic*) between two distinct points x and y in X, if there exists a *segment* (closed interval) [a,b] on the real line $\mathbb{R}$ and an isometric embedding $\gamma : [a, b] \to X$, such that $\gamma[a, b] = S$, $\gamma(a) = x$ and $\gamma(b) = y$.

  A **metric straight line** is a geodesic which is minimal between any two of its points; it is an isometric embedding of the whole of $\mathbb{R}$ into X. A **metric ray** and **metric great circle** are isometric embeddings of, respectively, the half-line $\mathbb{R}_{\geq 0}$ and a circle $S^1(0, r)$ into X.

  A **geodesic metric space** (Chap. 6) is a metric space in which any two points are joined by a geodesic segment. If, moreover, the geodesic is unique, the space is called *totally geodesic* (or *uniquely geodesic*).

  A geodesic metric space (X, d) is called *geodesically complete* if every geodesic is a subarc of a metric straight line. If (X, d) is **complete**, then it is geodesically complete. The **punctured plane** $(\mathbb{R}^2 \setminus \{0\}, ||x-y||_2)$ is not geodesically complete: any geodesic going to 0 is not a subarc of a metric straight line.

- **Length spectrum**

  Given a metric space (X, d), a *closed geodesic* is a map $\gamma : \mathbb{S}^1 \to X$ which is locally minimizing around every point of $\mathbb{S}^1$.

  If (X, d) is a compact **length space**, its **length spectrum** is the collection of lengths of closed geodesics. Each length is counted with *multiplicity* equal to the number of distinct *free homotopy* classes that contain a closed geodesic of such length. The **minimal length spectrum** is the set of lengths of closed geodesics which are the shortest in their free homotopy class. Cf. the **distance list**.

- **Systole of metric space**

  Given a compact metric space (X, d), its **systole** $sys(X, d)$ is the length of the shortest noncontractible loop in X; such a loop is a closed geodesic. So, $sys(X, d) = 0$ exactly if (X, d) is **simply connected**. Cf. **connected space** in Chap. 2.

  If (X, d) is a graph with path metric, then its systole is referred to as the *girth*. If (X, d) is a closed surface, then its *systolic ratio* is the ratio $SR = \frac{sys^2(X,d)}{area(X,d)}$.

[redacted]

Some tight upper bounds of SR for every metric on a surface are: $\frac{2}{\sqrt{3}} = \gamma_2$ (*Hermite constant* in 2D) for 2-torus (Loewner, 1949), $\frac{\pi}{2}$ for the real projective plane (Pu, 1952) and $\frac{\pi}{\sqrt{8}}$ for the Klein bottle (Bavard, 1986). Tight asymptotic bounds for a surface S of large genus g are $\frac{4}{9} \cdot \frac{\log^2 g}{\pi g} \leq SR(S) \leq \frac{\log^2 g}{\pi g}$ (Katz et al., 2007).

- **Shankar–Sormani radii**

  Given a **geodesic metric space** (X, d), Shankar and Sormani, 2009, defined its **unique injectivity radius** $Uirad(X)$ as the supremum over all $r \geq 0$ such that any two points at distance at most r are joined by a unique geodesic, and its **minimal radius** $Mrad(X)$ as $\inf_{p \in X} d(p, MinCut(p))$.

  Here the *minimal cut locus of p* $MinCut(p)$ is the set of points $q \in X$ for which there is a geodesic γ running from p to q such that γ extends past q but is not minimizing from p to any point past q. If (X, d) is a Riemannian space, then the distance function from p is a smooth function except at p itself and the cut locus. Cf. **medial axis and skeleton** in Chap. 21.

  It holds $Uirad(X) \leq Mrad(X)$ with equality if (X, d) is a Riemannian space in which case it is the **injectivity radius**. It holds $Uirad(X) = \infty$ for a flat disk but $Mrad(X) < \infty$ if (X, d) is compact and at least one geodesic is extendible.

- **Geodesic convexity**

  Given a **geodesic metric space** (X, d) and a subset $M \subset X$, the set M is called **geodesically convex** (or *convex*) if, for any two points of M, there exists a geodesic segment connecting them which lies entirely in M; the space is **strongly convex** if such a segment is unique and no other geodesic connecting those points lies entirely in M. The space is called **locally convex** if such a segment exists for any two sufficiently close points in M.

  For a given point $x \in M$, the **radius of convexity** is $r_x = \sup\{r \geq 0 : B(x, r) \subset M\}$, where the **metric ball** $B(x, r)$ is convex. The point x is called the *center of mass* of points $y_1, \dots, y_k \in M$ if it minimizes the function $\sum_i d(x, y_i)^2$ (cf. **Fréchet mean**); such point is unique if $d(y_i, y_j) < r_x$ for all $1 \leq i < j \leq k$.

  The **injectivity radius** of the set M is the supremum over all $r \geq 0$ such that any two points in M at distance $\leq r$ are joined by unique geodesic segment which lies in M. The **Hawaiian Earring** is a compact complete metric space consisting of a set of circles of radius $\frac{1}{i}$ for each $i \in \mathbb{N}$ all joined at a common point; its injectivity radius is 0. It is **path-connected** but not **simply connected**.

  The set $M \subset X$ is called a **totally convex metric subspace** of (X, d) if, for any two points of M, any geodesic segment connecting them lies entirely in M.

- **Busemann convexity**

  A **geodesic metric space** (X, d) is called **Busemann convex** (or *Busemann space*, *nonpositively curved in the sense of Busemann*) if, for any three points $x, y, z \in X$ and *midpoints* $m(x, z)$ and $m(y, z)$ (i.e., $d(x, m(x, z)) = d(m(x, z), z) = \frac{1}{2}d(x, z)$ and $d(y, m(y, z)) = d(m(y, z), z) = \frac{1}{2}d(y, z)$), there holds

$$d(m(x, z), m(y, z)) \leq \frac{1}{2}d(x, y).$$

The *flat Euclidean strip* $\{(x, y) \in \mathbb{R}^2 : 0 < x < 1\}$ is **Gromov hyperbolic metric space** (Chap. 6) but not Busemann convex one. In a complete Busemann convex metric space any two points are joined by a unique geodesic segment.

A locally geodesic metric space (X, d) is called **Busemann locally convex** if the above inequality holds locally. Any **locally CAT**(0) metric space is Busemann locally convex.

- **Menger convexity**

A metric space (X, d) is called **Menger convex** if, for any different points $x, y \in X$, there exists a third point $z \in X$ for which $d(x, y) = d(x, z) + d(z, y)$, i.e., $|I(x, y)| > 2$ holds for the **closed metric interval** $I(x, y) = \{z \in X : (x, y) = d(x, z) + d(z, y)\}$. It is called **strictly Menger convex** if such a z is unique for all $x, y \in X$.

Geodesic convexity implies Menger convexity. The converse holds for **complete** metric spaces.

A subset $M \subset X$ is called (Menger, 1928) a *d-convex set* (or *interval-convex set*) if $I(x, y) \subset M$ for any different points $x, y \in M$. A function $f : M \to \mathbb{R}$ defined on a d-convex set $M \subset X$ is a d**-convex function** if for any $z \in I(x, y) \subset M$

$$f(z) \le \frac{d(y, z)}{d(x, y)} f(x) + \frac{d(x, z)}{d(x, y)} f(y).$$

A subset $M \subset X$ is a *gated set* if for every $x \in X$ there exists a unique $x' \in M$, the *gate*, such that $d(x, y) = d(x, x') + d(x', y)$ for $y \in M$. Any such set is d-convex.

- **Midpoint convexity**

A metric space (X, d) is called **midpoint convex** (or **having midpoints**, *admitting a midpoint map*) if, for any different points $x, y \in X$, there exists a third point $m(x, y) \in X$ for which $d(x, m(x, y)) = d(m(x, y), y) = \frac{1}{2} d(x, y)$. Such a point $m(x, y)$ is called a *midpoint* and the map $m : X \times X \to X$ is called a *midpoint map* (cf. **midset**); this map is unique if $m(x, y)$ is unique for all $x, y \in X$.

For example, the geometric mean $\sqrt{xy}$ is the midpoint map for the metric space $(\mathbb{R}_{>0}, d(x, y) = |\log x - \log y|)$.

A **complete** metric space is **geodesic** if and only if it is midpoint convex.

A metric space (X, d) is said to have **approximate midpoints** if, for any points $x, y \in X$ and any $\epsilon > 0$, there exists an *ϵ-midpoint*, i.e., a point $z \in X$ such that $d(x, z) \le \frac{1}{2} d(x, y) + \epsilon \ge d(z, y)$.

- **Ball convexity**

A **midpoint convex** metric space (X, d) is called **ball convex** if

$$d(m(x, y), z) \le \max\{d(x, z), d(y, z)\}$$

for all $x, y, z \in X$ and any midpoint map $m(x, y)$.

[redacted]

Ball convexity implies that all metric balls are **totally convex** and, in the case of a **geodesic** metric space, vice versa. Ball convexity implies also the uniqueness of a midpoint map (geodesics in the case of **complete** metric space).

The metric space $(\mathbb{R}^2, d(x,y) = \sum_{i=1}^{2} \sqrt{|x_i - y_i|})$ is not ball convex.

- **Distance convexity**

  A **midpoint convex** metric space (X, d) is called **distance convex** if

  $$d(m(x,y),z) \leq \frac{1}{2}(d(x,z) + d(y,z)).$$

  A **geodesic metric space** is distance convex if and only if the restriction of the distance function $d(x,\cdot)$, $x \in X$, to every geodesic segment is a convex function.

  Distance convexity implies **ball convexity** and, in the case of **Busemann convex** metric space, vice versa.

- **Metric convexity**

  A metric space (X, d) is called **metrically convex** if, for any different points $x, y \in X$ and any $\lambda \in (0,1)$, there exists a third point $z = z(x,y,\lambda) \in X$ for which $d(x,y) = d(x,z) + d(z,y)$ and $d(x,z) = \lambda d(x,y)$.

  The space is called **strictly metrically convex** if such a point $z(x,y,\lambda)$ is unique for all $x, y \in X$ and any $\lambda \in (0,1)$.

  A metric space (X, d) is called **strongly metrically convex** if, for any different points $x, y \in X$ and any $\lambda_1, \lambda_2 \in (0,1)$, there exists a third point $z = z(x,y,\lambda) \in X$ for which $d(z(x,y,\lambda_1), z(x,y,\lambda_2)) = |\lambda_1 - \lambda_2| d(x,y)$.

  Metric convexity implies **Menger convexity**, and every Menger convex **complete** metric space is strongly metrically convex.

  A metric space (X, d) is called **nearly convex** (Mandelkern, 1983) if, for any different points $x, y \in X$ and any $\lambda, \mu > 0$ such that $d(x,y) < \lambda + \mu$, there exists a third point $z \in X$ for which $d(x,z) < \lambda$ and $d(z,y) < \mu$, i.e., $z \in B(x,\lambda) \cap B(y,\mu)$. Metric convexity implies near convexity.

- **Takahashi convexity**

  A metric space (X, d) is called **Takahashi convex** if, for any different points $x, y \in X$ and any $\lambda \in (0,1)$, there exists a third point $z = z(x,y,\lambda) \in X$ such that $d(z(x,y,\lambda), u) \leq \lambda d(x,u) + (1-\lambda) d(y,u)$ for all $u \in X$. Any convex subset of a normed space is a Takahashi convex metric space with $z(x,y,\lambda) = \lambda x + (1-\lambda)y$.

  A set $M \subset X$ is *Takahashi convex* if $z(x,y,\lambda) \in M$ for all $x, y \in X$ and any $\lambda \in [0,1]$. In a Takahashi convex metric space, all metric balls, open metric balls, and arbitrary intersections of Takahashi convex subsets are all Takahashi convex.

- **Hyperconvexity**

  A metric space (X, d) is called **hyperconvex** (Aronszajn–Panitchpakdi, 1956) if it is **metrically convex** and its metric balls have the *infinite Helly property*, i.e., any family of mutually intersecting closed balls in X has nonempty intersection. A metric space (X, d) is hyperconvex if and only if it is an **injective metric space**.

  The spaces l_∞^n, l_∞^∞ and l_1^2 are hyperconvex but l_2^∞ is not.

- **Distance matrix**

  Given a finite metric space $(X = \{x_1, \cdots, x_n\}, d)$, its **distance matrix** is the symmetric $n \times n$ matrix $((d_{ij}))$, where $d_{ij} = d(x_i, x_j)$ for any $1 \leq i, j \leq n$.

The probability that a symmetric $n \times n$ matrix, whose diagonal elements are zeros and all other elements are uniformly random real numbers, is a distance matrix is (Mascioni, 2005) $\frac{1}{2}$, $\frac{17}{120}$ for $n = 3, 4$, respectively.

- **Magnitude of a finite metric space**

  Let $(X = \{x_1, \dots, x_n\}, d)$ be a finite metric space, such that there exists a vector $w = \{w_1, \dots, w_n\}$ with $((e^{-d(x_i,x_j)}))w = (1, \dots, 1)^T$.

  Then the **magnitude** of (X, d) is (Leinster–Meckes, 2016) the sum $\sum_{i=1}^{n} w_i$. In fact, the definition of *Euler characteristic of a category* was generalized to *enriched categories*, renamed *magnitude*, then re-specialized to metric spaces.

- **Distance product of matrices**

  Given $n \times n$ matrices $A = ((a_{ij}))$ and $B = ((b_{ij}))$, their **distance** (or *min-plus*) **product** is the $n \times n$ matrix $C = ((c_{ij}))$ with $c_{ij} = \min_{k=1}^{n}(a_{ik} + b_{kj})$.

  It is the usual matrix multiplication in the *tropical semiring* $(\mathbb{R} \cup \{\infty\}, \min, +)$ (Chap. 18). If A is the matrix of weights of an edge-weighted complete graph K_n, then its *direct power* A^n is the (shortest path) **distance matrix** of this graph.

- **Distance list**

  Given a metric space (X, d), its **distance set** and **distance list** are the set and the *multiset* (i.e., multiplicities are counted) and of all pairwise distances.

  Two subsets $A, B \subset X$ are said to be *homometric sets* if they have the same distance list. Cf. **homometric structures** in Chap. 24.

  A finite metric space is called *tie-breaking* if all pairwise distances are distinct.

- **Degree of distance near-equality**

  Given a finite metric space (X, d) with $|X| = n \geq 3$, let $f = \min |\frac{d(x,y)}{d(a,b)} - 1|$ (**degree of distance near-equality**) and $f' = \min |\frac{d(x,y)}{d(x,b)} - 1|$, where the minimum is over different 2-subsets $\{x, y\}, \{a, b\}$ of X and, respectively, over different $x, y, b \in X$. [OpPi14] proved $f \leq \frac{9 \log n}{n^2}$ and $f' \leq \frac{3}{n}$, while $f \geq \frac{\log n}{20n^2}$ and $f' \geq \frac{1}{2n}$ for some (X, d).

- **Semimetric cone**

  The **semimetric cone** MET_n is the polyhedral cone in $\mathbb{R}^{\binom{n}{2}}$ of all **distance matrices** of semimetrics on the set $V_n = \{1, \dots, n\}$. Vershik, 2004, considers MET_∞, i.e., the weakly closed convex cone of infinite distance matrices of semimetrics on $\mathbb{N}$.

  The cone of n-point **weightable quasi-semimetrics** is a projection along an extreme ray of the semimetric cone Met_{n+1} (Grishukhin–Deza–Deza, 2011).

  The **metric fan** is a canonical decomposition MF_n of MET_n into subcones whose faces belong to the fan, and the intersection of any two of them is their common boundary. Two semimetrics $d, d' \in MET_n$ lie in the same cone of the metric fan if the subdivisions $\delta_d, \delta_{d'}$ of the polyhedron $\delta(n, 2) = conv\{e_i + e_j : 1 \leq i < j \leq n\} \subset \mathbb{R}^n$ are equal. Here a subpolytope P of $\delta(n, 2)$ is a cell of the subdivision δ_d if there exists $y \in \mathbb{R}^n$ satisfying $y_i + y_j = d_{ij}$ if $e_i + e_j$ is a vertex of P, and $y_i + y_j > d_{ij}$, otherwise. The complex of bounded faces of the polyhedron dual to δ_d is the **tight span** of the semimetric d.

[redacted]

- **Cayley–Menger matrix**

  Given a finite metric space $(X = \{x_1, \cdots, x_n\}, d)$, its **Cayley–Menger matrix** is the symmetric $(n+1) \times (n+1)$ matrix

  $$CM(X,d) = \begin{pmatrix} 0 & e \\ e^T & D \end{pmatrix},$$

  where $D = ((d^2(x_i, x_j)))$ and e is the n-vector all components of which are 1.

  The determinant of $CM(X,d)$ is called the *Cayley–Menger determinant*. If (X,d) is a metric subspace of the Euclidean space $\mathbb{E}^{n-1}$, then $CM(X,d)$ is $(-1)^n 2^{n-1}((n-1)!)^2$ times the squared $(n-1)$*-dimensional volume* of the convex hull of X in $\mathbb{R}^{n-1}$.
- **Gram matrix**

  Given elements $v_1, \dots, v_k$ of a Euclidean space, their **Gram matrix** is the symmetric $k \times k$ matrix VV^T, where $V = ((v_{ij}))$, of pairwise *inner products* of $v_1, \dots, v_k$:

  $$G(v_1, \dots, v_k) = ((\langle v_i, v_j \rangle)).$$

  It holds $G(v_1, \dots, v_k) = \frac{1}{2}((d_E^2(v_0, v_i) + d_E^2(v_0, v_j) - d_E^2(v_i, v_j)))$, i.e., the inner product $\langle \cdot, \cdot \rangle$ is the **Gromov product similarity** of the **squared Euclidean distance** d_E^2. A $k \times k$ matrix $((d_E^2(v_i, v_j)))$ is called *Euclidean distance matrix* (or *EDM*). It defines a **distance of negative type** on $\{1, \dots, k\}$; all such matrices form the (nonpolyhedral) closed convex cone of all such distances.

  The determinant of a Gram matrix is called the *Gram determinant*; it is equal to the square of the *k-dimensional volume* of the *parallelotope* constructed on $v_1, \dots v_k$.

  A symmetric $k \times k$ real matrix M is said to be *positive-semidefinite* (PSD) if $xMx^T \geq 0$ for any nonzero $x \in \mathbb{R}^k$ and *positive-definite* (PD) if $xMx^T > 0$. A matrix is PSD if and only if it is a Gram matrix; it is PD if and only the vectors $v_1, \dots, v_k$ are linearly independent. In Statistics, the *covariance matrices* and *correlation matrices* are exactly PSD and PD ones, respectively.
- **Midset**

  Given a metric space (X,d) and distinct $y, z \in X$, the **midset** (or *bisector*) of points y and z is the set $M = \{x \in X : d(x,y) = d(x,z)\}$ of *midpoints* x.

  A metric space is said to have the *n-point midset property* if, for every pair of its points, the midset has exactly n points. The one-point midset property means uniqueness of the *midpoint map*. Cf. **midpoint convexity**.
- **Distance k-sector**

  Given a metric space (X,d) and disjoint subsets $Y, Z \subset X$, the *bisector* of Y and Z is the set $M = \{x \in X : \inf_{y \in Y} d(x,y) = \inf_{z \in Z} d(x,z)\}$.

  The **distance k-sector** of Y and Z is the sequence $M_1, \dots, M_{k-1}$ of subsets of X such that M_i, for any $1 \leq i \leq k-1$, is the bisector of sets M_{i-1} and M_{i+1}, where $Y = M_0$ and $Z = M_k$. Asano–Matousek–Tokuyama, 2006, considered the distance k-sector on the Euclidean plane $(\mathbb{R}^2, l_2)$; for compact sets Y and Z, the sets $M_1, \dots, M_{k-1}$ are curves partitioning the plane into k parts.

- **Metric basis**

  Given a metric space (X, d) and a subset $M \subset X$, for any point $x \in X$, its *metric M-representation* is the set $\{(m, d(x, m)) : m \in M\}$ of its *metric M-coordinates* $(m, d(x, m))$. The set M is called (Blumenthal, 1953) a **metric basis** (or *resolving set*, *locating set*, *set of uniqueness*, *set of landmarks*) if distinct points $x \in X$ have distinct M-representations. A vertex-subset M of a connected graph is (Okamoto et al., 2009) a *local metric basis* if adjacent vertices have distinct M-representations.

  The *resolving number* of a finite (X, d) is (Chartrand–Poisson–Zhang, 2000) minimum k such that any k-subset of X is a metric basis.

  The vertices of a non degenerate simplex form a metric basis of $\mathbb{E}^n$, but l_1- and l_∞-metrics on $\mathbb{R}^n$, $n > 1$, have no finite metric basis.

  The **distance similarity** is (Saenpholphat–Zhang, 2003) an equivalence relation on X defined by $x \sim y$ if $d(z, x) = d(z, y)$ for any $z \in X \setminus \{x, y\}$. Any metric basis contains all or all but one elements from each equivalence class.

1.3 Metric Numerical Invariants

- **Resolving dimension**

  Given a metric space (X, d), its **resolving dimension** (or **location number** (Slater, 1975), *metric dimension* (Harary–Melter, 1976)) is the minimum cardinality of its **metric basis**. The **upper resolving dimension** of (X, d) is the maximum cardinality of its metric basis not containing another metric basis as a proper subset. *Adjacency dimension* of (X, d) is the metric dimension of $(X, \min(2, d))$.

  A **metric independence number** of (X, d) is (Currie–Oellermann, 2001) the maximum cardinality I of a collection of pairs of points of X, such that for any two, (say, (x, y) and (x', y')) of them there is no point $z \in X$ with $d(z, x) \neq d(z, y)$ and $d(z, x') \neq d(z, y')$. A function $f : X \to [0, 1]$ is a *resolving function* of (X, d) if $\sum_{z \in X : d(x,z) \neq d(y,z)} f(z) \geq 1$ for any distinct $x, y \in X$. The *fractional resolving dimension* of (X, d) is $F = \min \sum_{x \in X} g(x)$, where the minimum is taken over resolving functions f such that any function f' with f', f is not resolving.

  The *partition dimension* of (X, d) is (Chartrand–Salevi–Zhang, 1998) the minimum cardinality P of its *resolving partition*, i.e., a partition $X = \cup_{1 \leq i \leq k} S_i$ such that no two points have, for $1 \leq i \leq k$, the same minimal distances to the set S_i.

  Related *locating a robber* game on a graph $G = (V, E)$ was considered in 2012 by Seager and by Carraher et al.: *cop win* on G if every sequence $r = r_1, \ldots, r_n$ of robber's steps ($r_i \in V$ and $d_{\text{path}}(r_i, r_{i+1}) \leq 1$) is uniquely identified by a sequence $d(r_1, c_1), \ldots, d(r_n, c_n)$ of cop's distance queries for some $c_1, \ldots, c_n \in V$.

[redacted]

- **Metric dimension**

  For a metric space (X, d) and a number $\epsilon > 0$, let C_ϵ be the minimal size of an ϵ-**net** of (X, d), i.e., a subset $M \subset X$ with $\cup_{x \in M} B(x, \epsilon) = X$. The number

$$dim(X, d) = \lim_{\epsilon \to 0} \frac{\ln C_\epsilon}{-\ln \epsilon}$$

(if it exists) is called the **metric dimension** (or **Minkowski–Bouligand dimension**, **box-counting dimension**) of X. If the limit above does not exist, then the following notions of dimension are considered:

1. $\underline{dim}(X, d) = \underline{\lim}_{\epsilon \to 0} \frac{\ln C_\epsilon}{-\ln \epsilon}$ called the **lower Minkowski dimension** (or *lower dimension*, *lower box dimension*, *Pontryagin–Snirelman dimension*);
2. $\overline{dim}(X, d) = \overline{\lim}_{\epsilon \to 0} \frac{\ln C_\epsilon}{-\ln \epsilon}$ called the **Kolmogorov–Tikhomirov dimension** (or *upper dimension*, *entropy dimension*, *upper box dimension*).

See below examples of other, less prominent, notions of *metric dimension*.

1. The (equilateral) *metric dimension* of a metric space is the maximum cardinality of its *equidistant* subset, i.e., such that any two of its distinct points are at the same distance. For a normed space, this dimension is equal to the maximum number of translates of its unit ball that touch pairwise.
2. For any $c > 1$, the (normed space) *metric dimension* $dim_c(X)$ of a finite metric space (X, d) is the least dimension of a real *normed space* $(V, ||.||)$ such that there is an embedding $f : X \to V$ with $\frac{1}{c} d(x, y) \leq ||f(x) - f(y)|| \leq d(x, y)$.
3. The (Euclidean) *metric dimension* of a finite metric space (X, d) is the least dimension n of a Euclidean space $\mathbb{E}^n$ such that $(X, f(d))$ is its metric subspace, where the minimum is taken over all continuous monotone increasing functions $f(t)$ of $t \geq 0$.
4. The *dimensionality* of a metric space is $\frac{\mu^2}{2\sigma^2}$, where μ and σ^2 are the mean and variance of its histogram of distance values; this notion is used in Information Retrieval for proximity searching.

   The term *dimensionality* is also used for the minimal dimension, if it is finite, of Euclidean space in which a given metric space embeds isometrically.

- **Hausdorff dimension**

  Given a metric space (X, d) and $p, q > 0$, let $H_p^q = \inf \sum_{i=1}^{\infty} (diam(A_i))^p$, where the infimum is taken over all countable coverings $\{A_i\}$ with diameter of A_i less than q. The **Hausdorff** q**-measure** of X is the **metric outer measure** defined by

$$H^p = \lim_{q \to 0} H_p^q.$$

The **Hausdorff dimension** (or **fractal dimension**) of (X, d) is defined by

$$dim_{Haus}(X, d) = \inf\{p \geq 0 : H^p(X) = 0\}.$$

Any countable metric space has $dim_{Haus} = 0$, $dim_{Haus}(\mathbb{E}^n) = n$, and any $X \subset \mathbb{E}^n$ with $Int\ X \neq \emptyset$ has $dim_{Haus} = \overline{dim}$. For any **totally bounded** (X, d), it holds

$$dim_{top} \leq dim_{Haus} \leq \underline{dim} \leq dim \leq \overline{dim}.$$

- **Rough dimension**

  Given a metric space (X, d), its *rough n-volume* $Vol_n X$ is $\overline{\lim}_{\epsilon \to 0} \epsilon^n \beta_X(\epsilon)$, where $\epsilon > 0$ and $\beta_X(\epsilon) = \max |Y|$ for $Y \subseteq X$ with $d(a, b) \geq \epsilon$ if $a \in Y, b \in Y \setminus \{a\}$; $\beta_X(\epsilon) = \infty$ is permitted. The **rough dimension** is defined ([BBI01]) by

$$dim_{rough}(X, d) = \sup\{n : Vol_n X = \infty\} \text{ or, equivalently, } = \inf\{n : Vol_n X = 0\}.$$

  The space (X, d) can be not locally compact. It holds $dim_{Haus} \leq dim_{rough}$.

- **Packing dimension**

  Given a metric space (X, d) and $p, q > 0$, let $P^q_p = \sup \sum_{i=1}^{\infty} (diam(B_i))^p$, where the supremum is taken over all countable packings (by disjoint balls) $\{B_i\}$ with the diameter of B_i less than q.

  The **packing** q**-pre-measure** is $P^p_0 = \lim_{q \to 0} P^q_p$. The **packing** q**-measure** is the **metric outer measure** which is the infimum of packing q-pre-measures of countable coverings of X. The **packing dimension** of (X, d) is defined by

$$dim_{pack}(X, d) = \inf\{p \geq 0 : P^p(X) = 0\}.$$

- **Topological dimension**

  For any compact metric space (X, d) its **topological dimension** (or **Lebesgue covering dimension**) is defined by

$$\dim_{top}(X, d) = \inf_{d'}\{\dim_{Haus}(X, d')\},$$

  where d' is any metric on X **equivalent** to d. So, it holds $\dim_{top} \leq \dim_{Haus}$. A **fractal** (Chap. 18) is a metric space for which this inequality is strict.

  This dimension does not exceed also the **Assouad–Nagata dimension** of (X, d).

  In general, the **topological dimension** of a topological space X is the smallest integer n such that, for any finite open covering of X, there exists a finite open refinement of it with no point of X belonging to more than $n + 1$ elements.

  The *geometric dimension* is (Kleiner, 1999; [BBI01]) $\sup dim_{top}(Y, d)$ over compact $Y \subset X$.

- **Doubling dimension**

  The **doubling dimension** ($dim_{doubl}(X, d)$) of a metric space (X, d) is the smallest integer n (or ∞ if such an n does not exist) such that every metric ball (or, say, a set of finite diameter) can be covered by a family of at most 2^n metric balls (respectively, sets) of half the diameter.

[redacted]

If (X, d) has finite doubling dimension, then d is called a **doubling metric** and the smallest integer m such that every metric ball can be covered by a family of at most m metric balls of half the diameter is called *doubling constant*.

- **Assouad–Nagata dimension**

  The **Assouad–Nagata dimension** $dim_{AN}(X, d)$ of a metric space (X, d) is the smallest integer n (or ∞ if such an n does not exist) for which there exists a constant $C > 0$ such that, for all $s > 0$, there exists a covering of X by its subsets of diameter $\leq Cs$ with every subset of X of diameter $\leq s$ meeting $\leq n+1$ elements of covering. It holds (LeDonne–Rajala, 2014) $dim_{AN} \leq dim_{doubl}$; but $dim_{AN} = 1$, while $dim_{doubl} = \infty$, holds (Lang–Schlichenmaier, 2014) for some **real trees** (X, d).

  Replacing "for all $s > 0$" in the above definition by "for $s > 0$ sufficiently large" or by "for $s > 0$ sufficiently small", gives the *microscopic mi-dim*$_{AN}(X, d)$ and *macroscopic ma-dim*$_{AN}(X, d)$ Assouad–Nagata dimensions, respectively. Then (Brodskiy et al., 2006) $mi\text{-}dim_{AN}(X, d) = dim_{AN}(X, \min\{d, 1\})$ and

  $ma\text{-}dim_{AN}(X, d) = dim_{AN}(X, \max\{d, 1\})$ (here $\max\{d(x, y), 1\}$ means 0 for $x = y$).

  The Assouad–Nagata dimension is preserved (Lang–Schlichenmaier, 2004) under **quasi-symmetric mapping** but, in general, not under **quasi-isometry**.

- **Vol'berg–Konyagin dimension**

  The **Vol'berg–Konyagin dimension** of a metric space (X, d) is the smallest constant $C > 1$ (or ∞ if such a C does not exist) for which X carries a *doubling measure*, i.e., a Borel measure μ such that, for all $x \in X$ and $r > 0$, it holds that

$$\mu(\overline{B}(x, 2r)) \leq C\mu(\overline{B}(x, r)).$$

  A metric space (X, d) carries a doubling measure if and only if d is a **doubling metric**, and any complete doubling metric carries a doubling measure.

  The **Karger–Ruhl constant** of a metric space (X, d) is the smallest $c > 1$ (or ∞ if such a c does not exist) such that for all $x \in X$ and $r > 0$ it holds

$$|\overline{B}(x, 2r)| \leq c|\overline{B}(x, r)|.$$

  If c is finite, then the **doubling dimension** of (X, d) is at most $4c$.

- **Hyperbolic dimension**

  A metric space (X, d) is called an (R, N)-*large-scale doubling* if there exists a number $R > 0$ and integer $N > 0$ such that every ball of radius $r \geq R$ in (X, d) can be covered by N balls of radius $\frac{r}{2}$.

  The **hyperbolic dimension** $hypdim(X, d)$ of a metric space (X, d) (Buyalo–Schroeder, 2004) is the smallest integer n such that for every $r > 0$ there are $R > 0$, an integer $N > 0$ and a covering of X with the following properties:

  1. Every ball of radius r meets at most $n + 1$ elements of the covering;
  2. The covering is an (R, N)-large-scale doubling, and any finite union of its elements is an (R', N)-large-scale doubling for some $R' > 0$.

The hyperbolic dimension is 0 if (X, d) is a large-scale doubling, and it is n if (X, d) is n-dimensional hyperbolic space.

Also, $hypdim(X, d) \leq asdim(X, d)$ since the **asymptotic dimension** $asdim(X, d)$ corresponds to the case $N = 1$ in the definition of $hypdim(X, d)$.

The hyperbolic dimension is preserved under a **quasi-isometry**.

- **Asymptotic dimension**

The **asymptotic dimension** $asdim(X, d)$ of a metric space (X, d) (Gromov, 1993) is the smallest integer n such that, for every $r > 0$, there exists a constant $D = D(r)$ and a covering of X by its subsets of diameter at most D such that every ball of radius r meets at most $n + 1$ elements of the covering.

The asymptotic dimension is preserved under a **quasi-isometry**.

- **Width dimension**

Let (X, d) be a compact metric space. For a given number $\epsilon > 0$, the **width dimension** $Widim_\epsilon(X, d)$ of (X, d) is (Gromov, 1999) the minimum integer n such that there exists an n-dimensional polyhedron P and a continuous map $f : X \rightarrow P$ (called an *ϵ-embedding*) with $diam(f^{-1}(y)) \leq \epsilon$ for all $y \in P$.

The width dimension is a *macroscopic dimension at the scale* $\geq \epsilon$ of (X, d), because its limit for $\epsilon \rightarrow 0$ is the **topological dimension** of (X, d).

- **Godsil–McKay dimension**

We say that a metric space (X, d) has **Godsil–McKay dimension** $n \geq 0$ if there exists an element $x_0 \in X$ and two positive constants c and C such that the inequality $ck^n \leq |\{x \in X : d(x, x_0) \leq k\}| \leq Ck^n$ holds for every integer $k \geq 0$.

This notion was introduced in [GoMc80] for the **path metric** of a countable locally finite graph. They proved that, if the group $\mathbb{Z}^n$ acts faithfully and with a finite number of orbits on the vertices of the graph, then this dimension is n.

- **Metric outer measure**

A *σ-algebra over X* is any nonempty collection Σ of subsets of X, including X itself, that is closed under complementation and countable unions of its members.

Given a σ-algebra Σ over X, a *measure* on (X, Σ) is a function $\mu : \Sigma \rightarrow [0, \infty]$ with the following properties:

1. $\mu(\emptyset) = 0$;
2. For any sequence $\{A_i\}$ of pairwise disjoint subsets of X, $\mu(\sum_i A_i) = \sum_i \mu(A_i)$ (*countable σ-additivity*).

The triple (X, Σ, μ) is called a *measure space*. If $M \subset A \in \Sigma$ and $\mu(A) = 0$ implies $M \in \Sigma$, then (X, Σ, μ) is called a *complete measure space*. A measure μ with $\mu(X) = 1$ is called a *probability measure*.

If X is a *topological space* (see Chap. 2), then the σ-algebra over X, consisting of all *open* and *closed sets* of X, is called the *Borel σ-algebra* of X, (X, Σ) is called a *Borel space*, and a measure on Σ is called a *Borel measure*. So, any metric space (X, d) admits a Borel measure coming from its **metric topology**, where the *open set* is an arbitrary union of open **metric d-balls**.

An *outer measure* on X is a function $\nu : P(X) \rightarrow [0, \infty]$ (where $P(X)$ is the set of all subsets of X) with the following properties:

1. $\nu(\emptyset) = 0$;
2. For any subsets $A, B \subset X$, $A \subset B$ implies $\nu(A) \leq \nu(B)$ (*monotonicity*);

[redacted]

3. For any sequence $\{A_i\}$ of subsets of X, $\nu(\sum_i A_i) \leq \sum_i \nu(A_i)$ (*countable subadditivity*).

A subset $M \subset X$ is called *ν-measurable* if $\nu(A) = \nu(A \cup M) + \nu(A \setminus M)$ for any $A \subset X$. The set Σ' of all ν-measurable sets forms a σ-algebra over X, and (X, Σ', ν) is a complete measure space.

A **metric outer measure** is an outer measure ν defined on the subsets of a given metric space (X, d) such that $\nu(A \cup B) = \nu(A) + \nu(B)$ for every pair of nonempty subsets $A, B \subset X$ with positive **set-set distance** $\inf_{a \in A, b \in B} d(a, b)$. An example is **Hausdorff q-measure**; cf. **Hausdorff dimension**.

- **Length of metric space**

The **Fremlin length** of a metric space (X, d) is its **Hausdorff 1-measure** $H^1(X)$.

The **Hejcman length** $lng(M)$ of a subset $M \subset X$ of a metric space (X, d) is $\sup\{lng(M') : M' \subset M, |M'| < \infty\}$. Here $lng(\emptyset) = 0$ and, for a finite subset $M' \subset X$, $lng(M') = \min \sum_{i=1}^{n} d(x_{i-1}, x_i)$ over all sequences $x_0, \dots, x_n$ such that $\{x_i : i = 0, 1, \dots, n\} = M'$.

The **Schechtman length** of a finite metric space (X, d) is $\inf \sqrt{\sum_{i=1}^{n} a_i^2}$ over all sequences $a_1, \dots, a_n$ of positive numbers such that there exists a sequence $X_0, \dots, X_n$ of partitions of X with following properties:

1. $X_0 = \{X\}$ and $X_n = \{\{x\} : x \in X\}$;
2. X_i refines X_{i-1} for $i = 1, \dots, n$;
3. For $i = 1, \dots, n$ and $B, C \subset A \in X_{i-1}$ with $B, C \in X_i$, there exists a one-to-one map f from B onto C such that $d(x, f(x)) \leq a_i$ for all $x \in B$.

- **Volume of finite metric space**

Given a metric space (X, d) with $|X| = k < \infty$, its **volume** (Feige, 2000) is the maximal $(k-1)$-dimensional volume of the simplex with vertices $\{f(x) : x \in X\}$ over all **metric mappings** $f : (X, d) \to (\mathbb{R}^{k-1}, l_2)$. The volume coincides with the metric for $k = 2$. It is monotonically increasing and continuous in the metric d.

- **Rank of metric space**

The **Minkowski rank of a metric space** (X, d) is the maximal dimension of a normed vector space $(V, ||.||)$ such that there is an isometry $(V, ||.||) \to (X, d)$.

The **Euclidean rank of a metric space** (X, d) is the maximal dimension of a *flat* in it, that is of a Euclidean space $\mathbb{E}^n$ such that there is an isometric embedding $\mathbb{E}^n \to (X, d)$.

The **quasi-Euclidean rank of a metric space** (X, d) is the maximal dimension of a *quasi-flat* in it, i.e., of an Euclidean space $\mathbb{E}^n$ admitting a **quasi-isometry** $\mathbb{E}^n \to (X, d)$. Every **Gromov hyperbolic metric space** has this rank 1.

- **Roundness of metric space**

The **roundness of a metric space** (X, d) is the supremum of all q such that

$$d(x_1, x_2)^q + d(y_1, y_2)^q \leq d(x_1, y_1)^q + d(x_1, y_2)^q + d(x_2, y_1)^q + d(x_2, y_2)^q$$

for any four points $x_1, x_2, y_1, y_2 \in X$.

Every metric space has roundness ≥ 1; it is ≤ 2 if the space has **approximate midpoints**. The roundness of L_p**-space** is p if $1 \leq p \leq 2$.

The *generalized roundness of a metric space* (X, d) is (Enflo, 1969) the supremum of all q such that, for any $2k \geq 4$ points $x_i, y_i \in X$ with $1 \leq i \leq k$,

$$\sum_{1\leq i<j\leq k} d^q(x_i, x_j) + d^q(y_i, y_j) \leq \sum_{1\leq i,j\leq k} d^q(x_i, y_j).$$

Lennard–Tonge–Weston, 1997, have shown that the generalized roundness is the supremum of q such that d is of q**-negative type**, i.e., d^q is of **negative type**.

Every **CAT(0) space** (Chap. 6) has roundness 2, but some of them have generalized roundness 0 (Lafont–Prassidis, 2006).

- **Type of metric space**

The **Enflo type** of a metric space (X, d) is p if there exists a constant $1 \leq C < \infty$ such that, for every $n \in \mathbb{N}$ and every function $f : \{-1, 1\}^n \to X$, $\sum_{\epsilon\in\{-1,1\}^n} d^p(f(\epsilon), f(-\epsilon))$ is at most

$C^p \sum_{j=1}^n \sum_{\epsilon\in\{-1,1\}^n} d^p(f(\epsilon_1, \dots, \epsilon_{j-1}, \epsilon_j, \epsilon_{j+1}, \dots, \epsilon_n), f(\epsilon_1, \dots, \epsilon_{j-1}, -\epsilon_j, \epsilon_{j+1}, \dots, \epsilon_n))$.

A Banach space $(V, ||.||)$ of Enflo type p has *Rademacher type* p, i.e., for every $x_1, \dots, x_n \in V$, it holds

$$\sum_{\epsilon\in\{-1,1\}^n} ||\sum_{j=1}^n \epsilon_j x_j||^p \leq C^p \sum_{j=1}^n ||x_j||^p.$$

Given a metric space (X, d), a *symmetric Markov chain on* X is a Markov chain $\{Z_l\}_{l=0}^\infty$ on a state space $\{x_1, \dots, x_m\} \subset X$ with a symmetrical transition $m \times m$ matrix $((a_{ij}))$, such that $P(Z_{l+1} = x_j : Z_l = x_i) = a_{ij}$ and $P(Z_0 = x_i) = \frac{1}{m}$ for all integers $1 \leq i, j \leq m$ and $l \geq 0$. A metric space (X, d) has **Markov type** p (Ball, 1992) if $\sup_T M_p(X, T) < \infty$ where $M_p(X, T)$ is the smallest constant $C > 0$ such that the inequality

$$\mathbb{E}d^p(Z_T, Z_0) \leq TC^p\mathbb{E}d^p(Z_1, Z_0)$$

holds for every symmetric Markov chain $\{Z_l\}_{l=0}^\infty$ on X holds, in terms of expected value (mean) $\mathbb{E}[X] = \sum_x xp(x)$ of the discrete random variable X.

A metric space of Markov type p has Enflo type p.

- **Strength of metric space**

Given a finite metric space (X, d) with s different nonzero values of $d_{ij} = d(i, j)$, its **strength** is the largest number t such that, for any integers $p, q \geq 0$ with $p + q \leq t$, there is a polynomial $f_{pq}(s)$ of degree at most $\min\{p, q\}$ such that $((d_{ij}^{2p}))((d_{ij}^{2q})) = ((f_{pq}(d_{ij}^2)))$.

- **Rendez-vous number**

Given a metric space (X, d), its **rendez-vous number** (or *Gross number, magic number*) is a positive real number $r(X, d)$ (if it exists) defined by the

[redacted]

property that for each integer n and all (not necessarily distinct) $x_1, \dots, x_n \in X$ there exists a point $x \in X$ such that

$$r(X, d) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, x).$$

If the number $r(X, d)$ exists, then it is said that (X, d) has the **average distance property**. Every compact connected metric space has this property. The *unit ball* $\{x \in V : ||x|| \leq 1\}$ of a **Banach space** $(V, ||.||)$ has the rendez-vous number 1.

- **Wiener-like distance indices**

  Given a finite subset M of a metric space (X, d) and a parameter q, the **Wiener polynomial** of M (as defined by Hosoya, 1988, for the graphic metric d_{path}) is

$$W(M; q) = \frac{1}{2} \sum_{x,y \in M:\, x \neq y} q^{d(x,y)}.$$

It is a *generating function* for the **distance distribution** (Chap. 16) of M, i.e., the coefficient of q^i in $W(M; q)$ is the number $|\{\{x, y\} \in M \times M : d(x, y) = i\}|$.

In the main case when M is the vertex-set V of a connected graph $G = (V, E)$ and d is the **path metric** of G, the number $W(M; 1) = \frac{1}{2} \sum_{x,y \in M} d(x, y)$ is called the **Wiener index** of G. This notion is originated (Wiener, 1947) and applied, together with its many analogs, in Chemistry; cf. **chemical distance** in Chap. 24.

The *hyper-Wiener index* is $\sum_{x,y \in M}(d(x, y) + d(x, y)^2)$. The *reverse-Wiener index* is $\frac{1}{2} \sum_{x,y \in M}(D - d(x, y))$, where D is the diameter of M. The *complementary reciprocal Wiener index* is $\frac{1}{2} \sum_{x,y \in M}(1 + D - d(x, y))^{-1}$. The *Harary index* is $\sum_{x,y \in M}(d(x, y))^{-1}$. The *Szeged index* and the *vertex PI index* are $\sum_{e \in E} n_x(e) n_y(e)$ and $\sum_{e \in E}(n_x(e) + n_y(e))$, where $e = (xy)$ and $n_x(e) = |\{z \in V : d(x, z) < d(y, z)\}|$.

Two studied *edge-Wiener indices* of G are the Wiener index of its *line graph* and $\sum_{(xy),(x'y') \in E} \max\{d(x, x'), d(x, y'), d(y, x'), d(y, y')\}$.

The *Gutman–Schultz index*, **degree distance** (Dobrynin–Kochetova, 1994), *reciprocal degree distance* and *terminal Wiener index* are:

$$\sum_{x,y \in M} r_x r_y d(x, y),\ \sum_{x,y \in M} d(x, y)(r_x + r_y),\ \sum_{x,y \in M} \frac{1}{d(x, y)}(r_x + r_y),\ \sum_{x,y \in \{z \in M : r_z = 1\}} d(x, y),$$

where r_z is the degree of the vertex $z \in M$. The **eccentric distance sum** (Gupta et al., 2002) is $\sum_{y \in M}(\max\{d(x, y) : x \in M\} d_y)$, where d_y is $\sum_{x \in M} d(x, y)$. The *Balaban index* is $\frac{|E|}{c+1} \sum_{(yz) \in E}(\sqrt{d_y d_z})^{-1}$, where c is the number of primitive cycles. The *multiplicative Wiener index* is (Das–Gutman, 2016) $\prod_{x,y \in M, x \neq y} d(x, y)$.

Given a partition $P = \{V_1, \dots, V_k\}$ of the vertex-set V, set $f_P(x) = i$ for $x \in V_i$. The *colored distance* (Dankelman et al., 2001) and the *partition distance* (Klavžar, 2016) of G are $\sum_{f_P(x) \neq f_P(y)} d(x, y)$ and $\sum_{f_P(x) = f_P(y)} d(x, y)$, respectively.

Above indices are called (corresponding) *Kirchhoff indices* if d the **resistance metric** (Chap. 15) of G.

The *average distance* of M is the number $\frac{1}{|M|(|M|-1)}\sum_{x,y\in M} d(x,y)$. In general, for a quasi-metric space (X,d), the numbers $\sum_{x,y\in M} d(x,y)$ and $\frac{1}{|M|(|M|-1)}\sum_{x,y\in M, x\neq y}\frac{1}{d(x,y)}$ are called, respectively, the *transmission* and *global efficiency* of M.

- **Distance polynomial**

  Given an ordered finite subset M of a metric space (X,d), let D be the **distance matrix** of M. The **distance polynomial** of M is the *characteristic polynomial* of D, i.e., the determinant $det(D-\lambda I)$.

  Usually, D is the distance matrix of the **path metric** of a graph. Sometimes, the distance polynomial is defined as $det(\lambda I - D)$ or $(-1)^n det(D-\lambda I)$.

  The roots of the distance polynomial constitute the **distance spectrum** (or *D-spectrum* of *D-eigenvalues*) of M. Let $\rho_{\max}$ and $\rho_{\min}$ be the largest and the smallest roots; then $\rho_{\max}$ and $\rho_{\max}-\rho_{\min}$ are called (distance spectral) *radius* and *spread* of M. The **distance degree** of $x \in M$ is $\sum_{y\in M} d(x,y)$. The **distance energy** of M is the sum of the absolute values of its D-eigenvalues. It is $2\rho_{\max}$ if (as, for example, for the path metric of a tree) exactly one D-eigenvalue is positive.

- ***s*-energy**

  Given a finite subset M of a metric space (X,d) and a number $s>0$, the ***s*-energy** and **0-energy** of M are, respectively, the numbers

$$\sum_{x,y\in M, x\neq y}\frac{1}{d^s(x,y)} \quad \text{and} \quad \sum_{x,y\in M, x\neq y}\log\frac{1}{d(x,y)} = -\log\prod_{x,y\in M, x\neq y} d(x,y).$$

  The (unnormalized) ***s*-moment** of M is the number $\sum_{x,y\in M} d^s(x,y)$.

  The *discrete Riesz s-energy* is the s-energy for Euclidean distance d. In general, let μ be a finite Borel probability measure on (X,d). Then $U_s^{\mu}(x) = \int\frac{\mu(dy)}{d(x,y)^s}$ is the (abstract) *s-potential* at a point $x\in X$. The *Newton gravitational potential* is the case $(X,d)=(\mathbb{R}^3,|x-y|)$, $s=1$, for the mass distribution μ.

  The *s-energy* of μ is $E_s^{\mu}=\int U_s^{\mu}(x)\mu(dx)=\int\int\frac{\mu(dx)\mu(dy)}{d(x,y)^s}$, and the *s-capacity* of (X,d) is $(\inf_{\mu}E_s^{\mu})^{-1}$. Cf. the **metric capacity**.

- **Fréchet mean**

  Given a metric space (X,d) and a number $s>0$, the *Fréchet function* is $F_s(x)=\mathbb{E}[d^s(x,y)]$. For a finite subset M of X, this expected value is the mean $F_s(x)=\sum_{y\in M} w(y)d^s(x,y)$, where $w(y)$ is a weight function on M.

  The points, minimizing $F_1(x)$ and $F_2(x)$, are called the **Fréchet median** (or *weighted geometric median*) and **Fréchet mean** (or *Karcher mean*), respectively.

  If $(X,d)=(\mathbb{R}^n,||x-y||_2)$ and the weights are equal, these points are called the *geometric median* (or *Fermat–Weber point*, *1-median*) and the *centroid* (or *geometric center*, *barycenter*), respectively.

  The *k-median* and *k-mean* of M are the k-sets C minimizing, respectively, the sums $\sum_{y\in M}\min_{c\in C} d(y,c)=\sum_{y\in M} d(y,C)$ and $\sum_{y\in M} d^2(y,C)$.

[redacted]

Let (X, d) be the metric space $(\mathbb{R}_{>0}, |f(x) - f(y)|)$, where $f : \mathbb{R}_{>0} \to \mathbb{R}$ is a given injective and continuous function. Then the Fréchet mean of $M \subset \mathbb{R}_{>0}$ is the *f-mean* (or *Kolmogorov mean*, *quasi-arithmetic mean*) $f^{-1}(\frac{\sum_{x \in M} f(x)}{|M|})$. It is the arithmetic, geometric, harmonic, and power mean if $f = x, \log(x), \frac{1}{x}$, and $f = x^p$ (for a given $p \neq 0$), respectively. The cases $p \to +\infty, p \to -\infty$ correspond to maximum and minimum, while $p = 2, = 1, \to 0, \to -1$ correspond to the quadratic, arithmetic, geometric and harmonic mean.

Given a *completely monotonic* (i.e., $(-1)^k f^{(k)} \geq 0$ for any k) function $f \in \mathbb{C}^{\infty}$, the **$f$-potential energy** of a finite subset M of (X, d) is $\sum_{x,y \in M, x \neq y} f(d^2(x, y))$. The set M is called (Cohn–Kumar, 2007) *universally optimal* if it minimizes, among sets $M' \subset X$ with $|M'| = |M|$, the f-potential energy for any such f. Among universally optimal subsets of $(\mathbb{S}^{n-1}, ||x - y||_2)$, there are the vertex-sets of a polygon, simplex, cross-polytope, icosahedron, 600-cell, E_8 root system.

- **Distance-weighted mean**

  In Statistics, the **distance-weighted mean** between given data points $x_1, \dots, x_n$ is defined (Dodonov–Dodonova, 2011) by

$$\frac{\sum_{1 \leq i \leq n} w_i x_i}{\sum_{1 \leq i \leq n} w_i} \quad \text{with} \quad w_i = \frac{n-1}{\sum_{1 \leq j \leq n} |x_i - x_j|}.$$

  The case $w_i = 1$ for all i corresponds to the arithmetic mean.

- **Inverse distance weighting**

  In Numerical Analysis, *multivariate* (or *spatial*) interpolation is interpolation on functions of more than one variable. **Inverse distance weighting** is a method (Shepard, 1968) for multivariate interpolation. Let $x_1, \dots, x_n$ be interpolating points (i.e., samples $u_i = u(x_i)$ are known), x be an interpolated (unknown) point and $d(x, x_i)$ be a given distance. A general form of interpolated value $u(x)$ is

$$u(x) = \frac{\sum_{1 \leq i \leq n} w_i(x) u_i}{\sum_{1 \leq i \leq n} w_i(x)}, \text{ with } w_i(x) = \frac{1}{(d(x, x_i))^p},$$

  where $p > 0$ (usually $p = 2$) is a fixed *power parameter*.

- **Transfinite diameter**

  The *n-th diameter $D_n(M)$* and the *n-th Chebyshev constant $C_n(M)$* of a set $M \subseteq X$ in a metric space (X, d) are defined (Fekete, 1923, for the complex plane $\mathbb{C}$) as

$$D_n(M) = \sup_{x_1, \dots, x_n \in M} \prod_{i \neq j} d(x_i, x_j)^{\frac{1}{n(n-1)}} \text{ and } C_n(M) = \inf_{x \in X} \sup_{x_1, \dots, x_n \in M} \prod_{j=1}^{n} d(x, x_j)^{\frac{1}{n}}.$$

  The number $\log D_n(M)$ (the supremum of the average distance) is called the *n-extent* of M. The numbers $D_n(M), C_n(M)$ come from the geometric mean averaging; they also come as the limit case $s \to 0$ of the *s-moment* $\sum_{i \neq j} d(x_i, x_j)^s$ averaging.

The **transfinite diameter** (or ∞-*th diameter*) and the ∞-*th Chebyshev constant* $C_{\infty}(M)$ of M are defined as

$$D_{\infty}(M) = \lim_{n\to\infty} D_n(M) \text{ and } C_{\infty}(M) = \lim_{n\to\infty} C_n(M);$$

these limits existing since $\{D_n(M)\}$ and $\{C_n(M)\}$ are nonincreasing sequences of nonnegative real numbers. Define $D_{\infty}(\emptyset) = 0$.

The transfinite diameter of a compact subset of $\mathbb{C}$ is its **conformal radius** at infinity (cf. Chap. 6); for a segment in $\mathbb{C}$, it is $\frac{1}{4}$ of its length.

- **Metric diameter**

The **metric diameter** (or **diameter**, *width*) *diam*(M) of a set $M \subseteq X$ in a metric space (X, d) is defined by

$$\sup_{x,y\in M} d(x, y).$$

The *diameter graph* of M has, as vertices, all points $x \in M$ with $d(x, y) = diam(M)$ for some $y \in M$; it has, as edges, all pairs of its vertices at distance $diam(M)$ in (X, d). (X, d) is called a **diametrical metric space** if any $x \in X$ has the *antipode*, i.e., a unique $x' \in X$ such that the **closed metric interval** $I(x, x')$ is X.

The *furthest neighbor digraph* of M is a directed graph on M, where xy is an arc (called a *furthest neighbor pair*) whenever y is at maximal distance from x.

In a metric space endowed with a measure, one says that the *isodiametric inequality* holds if the metric balls maximize the measure among all sets with given diameter. It holds for the volume in Euclidean space but not, for example, for the **Heisenberg metric** on the *Heisenberg group* (Chap. 10).

The k**-ameter** (Grove–Markvorsen, 1992) is $\sup_{K\subseteq X:\,|K|=k} \frac{1}{2}\sum_{x,y\in K} d(x, y)$, and the k**-diameter** (Chung–Delorme–Sole, 1999) is $\sup_{K\subseteq X:\,|K|=k} \inf_{x,y\in K:\,x\neq y}$ $d(x, y)$.

Given a property $P \subseteq X \times X$ of a pair (K, K') of subsets of a finite metric space (X, d), the **conditional diameter** (called *P-diameter* in Balbuena et al., 1996) is $\max_{(K,K')\in P} \min_{(x,y)\in K\times K'} d(x, y)$. It is $diam(X, d)$ if $P = \{(K, K') \in X \times X : |K| = |K'| = 1\}$. When (X, d) models an interconnection network, the P-diameter corresponds to the maximum delay of the messages interchanged between any pair of clusters of nodes, K and K', satisfying a given property P of interest.

- **Metric spread**

A subset M of a metric space (X, d) is called **Delone set** (or *separated* ϵ**-net**, (A, a)-*Delone set*) if it is **bounded** (with a finite **diameter** $A = \sup_{x,y\in M} d(x, y)$) and **metrically discrete** (with a *separation* $a = \inf_{x,y\in M, x\neq y} d(x, y) > 0$).

The **metric spread** (or *distance ratio*, *normalized diameter*) of M is the ratio $\frac{A}{a}$.

[redacted]

The **aspect ratio** (or *axial ratio*) of a shape is the ratio of its longer and shorter dimensions, say, the length and diameter of a rod, major and minor axes of a torus or width and height of a rectangle (image, display, pixel, etc.).

For a mesh M with separation a and **covering radius** (or *mesh norm*) $c = \sup_{y \in X} \inf_{x \in M} d(x,y)$, the *mesh ratio* is $\frac{c}{a}$.

In Physics, the *aspect ratio* is the ratio of height-to-length scale characteristics. Cf. the *wing's aspect ratio* among **aircraft distances** in Chap. 29.

Dynamic range DNR is the ratio between the largest and smallest possible values of a quantity, such as in sound or light signals; cf. **SNR distance** in Chap. 21.

- **Eccentricity**

  Given a bounded metric space (X,d), the **eccentricity** (or *Koenig number*) of a point $x \in X$ is the number $e(x) = \max_{y \in X} d(x,y)$.

  The numbers $D = \max_{x \in X} e(x)$ and $r = \min_{x \in X} e(x)$ are called the **diameter** and the **radius** of (X,d), respectively. The point $z \in X$ is called *central* if $e(z) = r$, *peripheral* if $e(z) = D$, and *pseudo-peripheral* if for each point x with $d(z,x) = e(z)$ it holds that $e(z) = e(x)$. For finite $|X|$, the *average eccentricity* is $\frac{1}{|X|}\sum_{x \in X} e(x)$, and the *contour* of (X,d) is the set of points $x \in X$ such that no *neighbor* (closest point) of x has an eccentricity greater than x.

  The **eccentric digraph** (Buckley, 2001) of (X,d) has, as vertices, all points $x \in X$ and, as arcs, all ordered pairs (x,y) of points with $d(x,y) = e(y)$. The **eccentric graph** (Akyiama–Ando–Avis, 1976) of (X,d) has, as vertices, all points $x \in X$ and, as edges, all pairs (x,y) of points at distance $\min\{e(x), e(y)\}$.

  The **super-eccentric graph** (Iqbalunnisa–Janairaman–Srinivasan, 1989) of (X,d) has, as vertices, all points $x \in X$ and, as edges, all pairs (x,y) of points at distance no less than the radius of (X,d). The **radial graph** (Kathiresan–Marimuthu, 2009) of (X,d) has, as vertices, all points $x \in X$ and, as edges, all pairs (x,y) of points at distance equal to the radius of (X,d).

  The sets $\{x \in X : e(x) \leq e(z)$ for any $z \in X\}$, $\{x \in X : e(x) \geq e(z)$ for any $z \in X\}$ and $\{x \in X : \sum_{y \in X} d(x,y) \leq \sum_{y \in X} d(z,y)$ for any $z \in X\}$ are called, respectively, the **metric center** (or *eccentricity center*, *center*), **metric antimedian** (or *periphery*) and the **metric median** (or *distance center*) of (X,d).

- **Radii of metric space**

  Given a bounded metric space (X,d) and a set $M \subseteq X$ of **diameter** D, its **metric radius** (or **radius**) Mr, **covering radius** (or **directed Hausdorff distance** from X to M) Cr and **remoteness** (or **Chebyshev radius**) Re are the numbers $\inf_{x \in M} \sup_{y \in M} d(x,y)$, $\sup_{x \in X} \inf_{y \in M} d(x,y)$ and $\inf_{x \in X} \sup_{y \in M} d(x,y)$, respectively. It holds that $\frac{D}{2} \leq Re \leq Mr \leq D$ with $Mr = \frac{D}{2}$ in any **injective metric space**. Somemimes, $\frac{D}{2}$ is called the *radius*.

  For $m > 0$, a **minimax distance design of size** m is an m-subset of X having smallest covering radius. This radius is called the *m-point mesh norm* of (X,d).

  The **packing radius** Pr of M is the number $\sup\{r : \inf_{x,y \in M, x \neq y} d(x,y) > 2r\}$. For $m > 0$, a **maximum distance design of size** m is an m-subset of X having largest packing radius. This radius is the *m-point best packing distance* on (X,d).

- **ϵ-net**

  Given a metric space (X, d), a subset $M \subset X$, and a number $\epsilon > 0$, the **ϵ-neighborhood** of M is the set $M^{\epsilon} = \cup_{x \in M} B(x, \epsilon)$.

  The set M is called an **ϵ-net** (or *ϵ-covering*, *ϵ-approximation*) of (X, d) if $M^{\epsilon} = X$, i.e., the **covering radius** of M is at most ϵ.

  Let C_{ϵ} denote the *ϵ-covering number*, i.e., the smallest size of an ϵ-net in (X, d). The number $\lg_2 C_{\epsilon}$ is called (Kolmogorov–Tikhomirov, 1959) the **metric entropy** (or *ϵ-entropy*) of (X, d). It holds $P_{\epsilon} \leq C_{\epsilon} \leq P_{\frac{\epsilon}{2}}$, where P_{ϵ} denote the *ϵ-packing number* of (X, d), i.e., $\sup\{|M| : M \subset X, \overline{B}(x, \epsilon) \cap \overline{B}(y, \epsilon) = \emptyset$ for any $x, y \in M, x \neq y\}$. The number $\lg_2 P_{\epsilon}$ is called the **metric capacity** (or *ϵ-capacity*) of (X, d).
- **Steiner ratio**

  Given a metric space (X, d) and a finite subset $V \subset X$, let $G = (V, E)$ be the complete weighted graph on V with edge-weights $d(x, y)$ for all $x, y \in V$.

  Given a tree T, its *weight* is the sum $d(T)$ of its edge-weights. A *spanning tree* of V is a subset of $|V| - 1$ edges forming a tree on V. Let $MSpT_V$ be a *minimum spanning tree* of V, i.e., a spanning tree with the minimal weight $d(MSpT_V)$.

  A *Steiner tree* of V is a tree on Y, $V \subset Y \subset X$, connecting vertices from V; elements of $Y \setminus V$ are called *Steiner points*. Let $StMT_V$ be a *minimum Steiner tree* of V, i.e., a Steiner tree with the minimal weight $d(StMT_V) = \inf_{Y \subset X : V \subset Y} d(MSpT_Y)$. This weight is called the **Steiner diversity** of V; cf. **diversity** in Chap. 3. It is the **Steiner distance of set** V (Chap. 15) if (X, d) is graphic metric space.

  The **Steiner ratio** $St(X, d)$ of the metric space (X, d) is defined by

$$\inf_{V \subset X} \frac{d(StMT_V)}{d(MSpT_V)}.$$

  Cf. **arc routing problems** in Chap. 15.
- **Chromatic numbers of metric space**

  Given a metric space (X, d) and a set D of positive real numbers, the **D-chromatic number** of (X, d) is the standard *chromatic number* of its **D-distance graph**, i.e., the graph (X, E) with the vertex-set X and the edge-set $E = \{xy : d(x, y) \in D\}$ (Chap. 15). Usually, (X, d) is an **l_p-space** and $D = \{1\}$ (**Benda–Perles chromatic number**) or $D = [1 - \epsilon, 1 + \epsilon]$.

  For a metric space (X, d), the **polychromatic number** is the minimum number of colors needed to color all the points $x \in X$ so that, for each color class C_i, there is a distance d_i such that no two points of C_i are at distance d_i.

  For a metric space (X, d), the **packing chromatic number** is the minimum number of colors needed to color all the points $x \in X$ so that, for each color class C_i, no two distinct points of C_i are at distance at most i.

  For any integer $t > 0$, the **t-distance chromatic number** of a metric space (X, d) is the minimum number of colors needed to color all the points $x \in X$ so that any two points whose distance is $\leq t$ have distinct colors. Cf. **k-distance chromatic number** in Chap. 15.

[redacted]

For any integer $t > 0$, the **t-th Babai number** of a metric space (X, d) is the minimum number of colors needed to color all the points in X so that, for any set D of positive distances with $|D| \leq t$, any two points $x, y \in X$ with $d(x, y) \in D$ have distinct colors.

- **Congruence order of metric space**

  A metric space (X, d) has **congruence order** n if every finite metric space which is not **isometrically embeddable** in (X, d) has a subspace with at most n points which is not isometrically embeddable in (X, d). For example, the congruence order of l_2^n is $n + 3$ (Menger, 1928); it is 4 for the **path metric** of a tree.

1.4 Main Mappings of Metric Spaces

- **Distance function**

  In Topology, the term *distance function* is often used for **distance**. But, in general, a **distance function** (or *ray function*) is a continuous function on a metric space (X, d) (usually, on a Euclidean space $\mathbb{E}^n$) $f : X \to \mathbb{R}_{\geq 0}$ which is *homogeneous*, i.e., $f(tx) = tf(x)$ for all $t \geq 0$ and all $x \in X$.

  Such function f is called *positive* if $f(x) > 0$ for all $x \neq 0$, *symmetric* if $f(x) = f(-x)$, *convex* if $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ for any $0 < t < 1$ and $x \neq y$, and *strictly convex* if this inequality is strict.

  If $X = \mathbb{E}^n$, the set $S_f = \{x \in \mathbb{R}^n : f(x) < 1\}$ is *star body*, i.e., $x \in S_f$ implies $[0, x] \subset S_f$. Any star body S corresponds to a unique distance function $g(x) = \inf_{tx \in S, t>0} \frac{1}{t}$, and $S = S_g$. The star body is bounded if f is positive, symmetric about the origin if f is symmetric, convex if f is convex, and *strictly convex* (i.e., the boundary ∂B does not contain a segment) if f is strictly convex.

  For a quadratic distance function of the form $f_A = xAx^T$, where A is a real matrix and $x \in \mathbb{R}^n$, the matrix A is *positive-definite* (i.e., the **Gram matrix** $VV^T = ((\langle v_i, v_j \rangle))$ of n linearly independent vectors $v_i = (v_{i1}, \dots, v_{in})$) if and only if f_A is symmetric and strictly convex function. The *homogeneous minimum* of f_A is

$$\min(f_A) = \inf_{x \in \mathbb{Z}^n \setminus \{0\}} f_A(x) = \inf_{x \in L \setminus \{0\}} \sum_{1 \leq i \leq n} x_i^2,$$

  where $L = \{\sum x_i v_i : x_i \in \mathbb{Z}\}$ is a *lattice*, i.e., a discrete subgroup of $\mathbb{R}^n$ spanning it. The *Hermite constant* γ_n, a central notion in Geometry of Numbers, is the supremum, over all positive-definite $(n \times n)$-matrices, of $\min(f_A) \det(A)^{\frac{1}{n}}$. It is known only for $2 \leq n \leq 8$ and $n = 24$; cf. **systole of metric space**.

- **Convex distance function**

  Given a compact convex region $B \subset \mathbb{R}^n$ containing the origin O in its interior, the **convex distance function** (or **Minkowski distance function**, *Minkowski* *seminorm*, *gauge*) is the function $||P||_B$ whose value at a point $P \in \mathbb{R}^n$ is the *distance ratio* $\frac{OP}{OQ}$, where $Q \in B$ is the furthest from O point on the ray OP.

Then $d_B(x,y) = ||x-y||_B$ is the quasi-metric on $\mathbb{R}^n$ defined, for $x \neq y$, by

$$\inf\{\alpha > 0 : y - x \in \alpha B\},$$

and $B = \{x \in \mathbb{R}^n : d_B(0,x) \leq 1\}$ with equality only for $x \in \partial B$.

The function $||P||_B$ is called a **polyhedral distance function** if B is a n-polytope, *simplicial distance function* if it is a n-simplex, and so on.

If B is centrally-symmetric with respect to the origin, then d_B is a **Minkowskian metric** (Chap. 6) whose unit ball is B. This is the l_1-metric if B is the n-cross-polytope and the l_∞-metric if B is the n-cube.

- **Funk distance**

  Let B be an nonempty open convex subset of $\mathbb{R}^n$. For any $x, y \in B$, denote by $R(x,y)$ the ray from x through y. The **Funk distance** (Funk, 1929) on B is the quasi-semimetric defined, for any $x, y \in B$, as 0 if the boundary $\partial(B)$ and $R(x,y)$ are disjoint, and, otherwise, i.e., if $R(x,y) \cap \partial B = \{z\}$, by

  $$\ln \frac{||x-z||_2}{||y-z||_2}.$$

  The **Hilbert projective metric** in Chap. 6 is a symmetrization of this distance.

- **Metric projection**

  Given a metric space (X,d) and a subset $M \subset X$, an element $u_0 \in M$ is called an **element of best approximation** (or **nearest point**) to a given element $x \in X$ if $d(x,u_0) = \inf_{u \in M} d(x,u)$, i.e., if $d(x,u_0)$ is the **point-set distance** $d(x,M)$.

  A **metric projection** (or *operator of best approximation*, *nearest point map*) is a multivalued mapping associating to each element $x \in X$ the set of elements of best approximation from the set M (cf. **distance map**).

  A **Chebyshev set** in a metric space (X,d) is a subset $M \subset X$ containing a unique element of best approximation for every $x \in X$.

  A subset $M \subset X$ is called a **semi-Chebyshev set** if the number of such elements is at most one, and a **proximinal set** if this number is at least one.

  While the **Chebyshev radius** (or **remoteness**; cf. **radii of metric space**) of the set M is $\inf_{x \in X} \sup_{y \in M} d(x,y)$, a **Chebyshev center** of M is an element $x_0 \in X$ realizing this infimum. Sometimes (say, for a finite graphic metric space), $\frac{1}{|M|} \inf_{x \in X} \sum_{y \in M} d(x,y)$ and $\frac{1}{|M|} \sup_{x \in X} \sum_{y \in M} d(x,y)$ are called *proximity* and *remoteness* of M.

- **Distance map**

  Given a metric space (X,d) and a subset $M \subset X$, the **distance map** is a function $f_M : X \to \mathbb{R}_{\geq 0}$, where $f_M(x) = \inf_{u \in M} d(x,u)$ is the **point-set distance** $d(x,M)$ (cf. **metric projection**).

[redacted]

If the boundary $B(M)$ of the set M is defined, then the **signed distance function** g_M is defined by $g_M(x) = -\inf_{u \in B(M)} d(x,u)$ for $x \in M$, and $g_M(x) = \inf_{u \in B(M)} d(x,u)$, otherwise. If M is a (closed orientable) ***n*-manifold** (Chap. 2), then g_M is the solution of the *eikonal equation* $|\nabla g| = 1$ for its *gradient* ∇.

If $X = \mathbb{R}^n$ and, for every $x \in X$, there is unique element $u(x)$ with $d(x,M) = d(x,u(x))$ (i.e., M is a **Chebyshev set**), then $||x - u(x)||$ is called a **vector distance function**.

Distance maps are used in Robot Motion (M being the set of obstacle points) and, especially, in Image Processing (M being the set of all or only boundary pixels of the image). For $X = \mathbb{R}^2$, the graph $\{(x, f_M(x)) : x \in X\}$ of $d(x,M)$ is called the *Voronoi surface* of M.

- **Isometry**

Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \rightarrow Y$ is called an **isometric embedding** of X into Y if it is injective and the equality $d_Y(f(x), f(y)) = d_X(x,y)$ holds for all $x, y \in X$.

An **isometry** (or *congruence mapping*) is a bijective isometric embedding. Two metric spaces are called **isometric** (or *isometrically isomorphic*) if there exists an isometry between them.

A property of metric spaces which is invariant with respect to isometries (completeness, boundedness, etc.) is called a **metric property** (or *metric invariant*).

A **path isometry** (or *arcwise isometry*) is a mapping from X into Y (not necessarily bijective) preserving lengths of curves.

- **Rigid motion of metric space**

A **rigid motion** (or, simply, **motion**) of a metric space (X,d) is an **isometry** of (X,d) onto itself.

For a motion f, the **displacement function** $d_f(x)$ is $d(x, f(x))$. The motion f is called *semisimple* if $\inf_{x \in X} d_f(x) = d(x_0, f(x_0))$ for some $x_0 \in X$, and *parabolic*, otherwise. A semisimple motion is called *elliptic* if $\inf_{x \in X} d_f(x) = 0$, and *axial* (or *hyperbolic*), otherwise. A motion is called a *Clifford translation* if the displacement function $d_f(x)$ is a constant for all $x \in X$.

- **Symmetric metric space**

A metric space (X,d) is called **symmetric** if, for any point $p \in X$, there exists a *symmetry* relative to that point, i.e., a **motion** f_p of this metric space such that $f_p(f_p(x)) = x$ for all $x \in X$, and p is an isolated fixed point of f_p.

- **Homogeneous metric space**

A metric space is called **homogeneous** (or *point-homogeneous*) if, for any two points of it, there exists a motion mapping one of the points to the other.

In general, a *homogeneous space* is a set together with a given transitive group of *symmetries*. Moss, 1992, defined similar *distance-homogeneous distanced graph*.

A metric space is called **ultrahomogeneous space** (or *highly transitive*) if any isometry between two of its finite subspaces extends to the whole space.

A metric space (X,d) is called (Grünbaum–Kelly) a **metrically homogeneous metric space** if $\{d(x,z) : z \in X\} = \{d(y,z) : z \in X\}$ for any $x, y \in X$.

- **Flat space**

  A **flat space** is any metric space with **local isometry** to some $\mathbb{E}^n$, i.e., each point has a neighborhood isometric to an open set in $\mathbb{E}^n$. A space is *locally Euclidean* if every point has a neighborhood homeomorphic to an open subset in $\mathbb{E}^n$.

- **Dilation of metric space**

  Given a metric space (X, d), its **dilation** (or r**-dilation**) is a mapping $f : X \rightarrow X$ with $d(f(x), f(y)) = rd(x, y)$ for some $r > 0$ and any $x \in X$.

- **Wobbling of metric space**

  Given a metric space (X, d), its **wobbling** (or r**-wobbling**) is a mapping $f : X \rightarrow X$ with $d(x, f(x)) < r$ for some $r > 0$ and any $x \in X$.

- **Paradoxical metric space**

  Given a metric space (X, d) and an equivalence relation on the subsets of X, the space (X, d) is called **paradoxical** if X can be decomposed into two disjoint sets M_1, M_2 so that M_1, M_2 and X are pairwise equivalent.

  Deuber, Simonovitz and Sós, 1995, introduced this idea for *wobbling equivalent* subsets $M_1, M_2 \subset X$, i.e., there is a bijective r**-wobbling** $f : M_1 \rightarrow M_2$. For example, $(\mathbb{R}^2, l_2)$ is paradoxical for wobbling but not for isometry equivalence.

- **Metric cone**

  A **pointed metric space** (X, d, x_0) is called a **metric cone**, if it is isometric to $(\lambda X, d, x_0)$ for all $\lambda > 0$. A **metric cone structure** on (X, d, x_0) is a (pointwise) continuous family f_t $(t \in \mathbb{R}_{>0})$ of **dilations** of X, leaving the point x_0 invariant, such that $d(f_t(x), f_t(y)) = td(x, y)$ for all x, y and $f_t \circ f_s = f_{ts}$. A Banach space has such a structure for the dilations $f_t(x) = tx$ $(t \in \mathbb{R}_{>0})$. The *Euclidean cone over a metric space* (cf. **cone over metric space** in Chap. 9) is another example.

  The **tangent metric cone** over a metric space (X, d) at a point x_0 is (for all dilations $tX = (X, td)$) the closure of $\cup_{t>0} tX$, i.e., of $\lim_{t\rightarrow\infty} tX$ taken in the pointed Gromov–Hausdorff topology (cf. **Gromov–Hausdorff metric**).

  The **asymptotic metric cone** over (X, d) is its tangent metric cone "at infinity", i.e., $\cap_{t>0} tX = \lim_{t\rightarrow 0} tX$. Cf. **boundary of metric space** in Chap. 6.

  The term *metric cone* was also used by Bronshtein, 1998, for a convex cone C equipped with a complete metric compatible with its operations of addition (continuous on $C \times C$) and multiplication (continuous on $C \times \mathbb{R}_{\geq 0}$). by all $\lambda \geq 0$.

- **Metric fibration**

  Given a **complete** metric space (X, d), two subsets M_1 and M_2 of X are called *equidistant* if for each $x \in M_1$ there exists $y \in M_2$ with $d(x, y)$ being equal to the **Hausdorff metric** between the sets M_1 and M_2. A **metric fibration** of (X, d) is a partition $\mathcal{F}$ of X into isometric mutually equidistant closed sets.

  The quotient metric space $X/\mathcal{F}$ inherits a natural metric for which the **distance map** is a **submetry**.

- **Homeomorphic metric spaces**

  Two metric spaces (X, d_X) and (Y, d_Y) are called **homeomorphic** (or *topologically isomorphic*) if there exists a *homeomorphism* from X to Y, i.e., a bijective function $f : X \rightarrow Y$ such that f and f^{-1} are *continuous* (the preimage of every open set in Y is open in X).

[redacted]

Two metric spaces (X, d_X) and (Y, d_Y) are called *uniformly isomorphic* if there exists a bijective function $f : X \to Y$ such that f and f^{-1} are *uniformly continuous*. A function g is *uniformly continuous* if, for any $\epsilon > 0$, there exists $\delta > 0$ such that, for any $x, y \in X$, the inequality $d_X(x, y) < \delta$ implies that $d_Y(g(x), f(y)) < \epsilon$; a continuous function is uniformly continuous if X is compact.

- **Möbius mapping**

  Given distinct points x, y, z, w of a metric space (X, d), their **cross-ratio** is

$$cr((x, y, z, w), d) = \frac{d(x, y)d(z, w)}{d(x, z)d(y, w)} > 0.$$

Given metric spaces (X, d_X) and (Y, d_Y), a **homeomorphism** $f : X \to Y$ is called a **Möbius mapping** if, for every distinct points $x, y, z, w \in X$, it holds

$$cr((x, y, z, w), d_X) = cr((f(x), f(y), f(z), f(w)), d_Y).$$

A homeomorphism $f : X \to Y$ is called a **quasi-Möbius mapping** (Väisälä, 1984) if there exists a homeomorphism $\tau : [0, \infty) \to [0, \infty)$ such that, for every quadruple x, y, z, w of distinct points of X, it holds

$$cr((f(x), f(y), f(z), f(w)), d_Y) \leq \tau(cr((x, y, z, w), d_X)).$$

A metric space (X, d) is called *metrically dense* (or *μ-dense* for given $\mu > 1$, Aseev–Trotsenko, 1987) if for any $x, y \in X$, there exists a sequence $\{z_i, i \in \mathbb{Z}\}$ with $z_i \to x$ as $i \to -\infty$, $z_i \to y$ as $i \to \infty$, and $\log cr((x, z_i, z_{i+1}, y), d) \leq \log \mu$ for all $i \in \mathbb{Z}$. The space (X, d) is μ-dense if and only if (Tukia-Väisälä, 1980), for any $x, y \in X$, there exists $z \in X$ with $\frac{d(x,y)}{6\mu} \leq d(x, z) \leq \frac{d(x,y)}{4}$.

- **Quasi-symmetric mapping**

  Given metric spaces (X, d_X) and (Y, d_Y), a **homeomorphism** $f : X \to Y$ is called a **quasi-symmetric mapping** (Tukia–Väisälä, 1980) if there is a homeomorphism $\tau : [0, \infty) \to [0, \infty)$ such that, for every triple (x, y, z) of distinct points of X,

$$\frac{d_Y(f(x), f(y))}{d_Y(f(x), f(z))} \leq \tau \frac{d_X(x, y)}{d_X(x, z)}.$$

Quasi-symmetric mappings are **quasi-Möbius**, and quasi-Möbius mappings between bounded metric spaces are quasi-symmetric. In the case $f : \mathbb{R}^n \to \mathbb{R}^n$, quasi-symmetric mappings are exactly the same as **quasi-conformal mappings**.

- **Conformal metric mapping**

  Given metric spaces (X, d_X) and (Y, d_Y) which are domains in $\mathbb{R}^n$, a **homeomorphism** $f : X \to Y$ is called a **conformal metric mapping** if, for any nonisolated point $x \in X$, the limit $\lim_{y \to x} \frac{d_Y(f(x), f(y))}{d(x,y)}$ exists, is finite and positive.

A homeomorphism $f : X \to Y$ is called a **quasi-conformal mapping** (or, specifically, *C-quasi-conformal mapping*) if there exists a constant C such that

$$\limsup_{r \to 0} \frac{\max\{d_Y(f(x), f(y)) : d_X(x,y) \le r\}}{\min\{d_Y(f(x), f(y)) : d_X(x,y) \ge r\}} \le C$$

for each $x \in X$. The smallest such constant C is called the **conformal dilation**.

The **conformal dimension** of a metric space (X, d) (Pansu, 1989) is the infimum of the **Hausdorff dimension** over all quasi-conformal mappings of (X, d) into some metric space. For the middle-third Cantor set on $[0, 1]$, it is 0 but, for any of its quasi-conformal images, it is positive.

- **Hölder mapping**

  Let $c, \alpha \ge 0$ be constants. Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called the **Hölder mapping** (or *α-Hölder mapping* if the constant α should be mentioned) if for all $x, y \in X$

  $$d_Y(f(x), f(y)) \le c(d_X(x,y))^{\alpha}.$$

  A 1-Hölder mapping is a **Lipschitz mapping**; 0-Hölder mapping means that the metric d_Y is bounded.

- **Lipschitz mapping**

  Let c be a positive constant. Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called a **Lipschitz** (or **Lipschitz continuous**, *c-Lipschitz* if the constant c should be mentioned) **mapping** if for all $x, y \in X$ it holds

  $$d_Y(f(x), f(y)) \le c d_X(x,y).$$

  A c-Lipschitz mapping is called a **metric mapping** if $c = 1$, and is called a **contraction** if $c < 1$.

- **Bi-Lipschitz mapping**

  Given metric spaces (X, d_X), (Y, d_Y) and a constant $c > 1$, a function $f : X \to Y$ is called a **bi-Lipschitz mapping** (or *c-bi-Lipschitz mapping*, **c-embedding**) if there exists a number $r > 0$ such that for any $x, y \in X$ it holds

  $$r d_X(x,y) \le d_Y(f(x), f(y)) \le c r d_X(x,y).$$

  Every bi-Lipschitz mapping is a **quasi-symmetric mapping**.

  The smallest c for which f is a c-bi-Lipschitz mapping is called the **distortion** of f. Bourgain, 1985, proved that every k-point metric space c-embeds into a Euclidean space with distortion $O(\ln k)$. Gromov's *distortion for curves* is the maximum ratio of arc length to chord length.

  Two metrics d_1 and d_2 on X are called **bi-Lipschitz equivalent metrics** if there are positive constants c and C such that $c d_1(x,y) \le d_2(x,y) \le C d_1(x,y)$ for all $x, y \in X$, i.e., the identity mapping is a bi-Lipschitz mapping from (X, d_1) into (X, d_2). Bi-Lipschitz equivalent metrics are **equivalent**, i.e., generate the same

[redacted]

topology but, for example, equivalent L_1-metric and L_2-metric (cf. **L_p-metric** in Chap. 5) on $\mathbb{R}$ are not bi-Lipschitz equivalent.

A bi-Lipschitz mapping $f : X \to Y$ is a **c-isomorphism** $f : X \to f(X)$.

- **c-isomorphism of metric spaces**

  Given two metric spaces (X, d_X) and (Y, d_Y), the *Lipschitz norm* $||.||_{Lip}$ on the set of all injective mappings $f : X \to Y$ is defined by

  $$||f||_{Lip} = \sup_{x,y \in X, x \neq y} \frac{d_Y(f(x), f(y))}{d_X(x, y)}.$$

  Two metric spaces X and Y are called **c-isomorphic** if there exists an injective mapping $f : X \to Y$ such that $||f||_{Lip}||f^{-1}||_{Lip} \leq c$.

- **Metric Ramsey number**

  For a given class $\mathcal{M}$ of metric spaces (usually, *l_p-spaces*), an integer $n \geq 1$, and a real number $c \geq 1$, the **metric Ramsey number** (or *c-metric Ramsey number*) $R_{\mathcal{M}}(c, n)$ is the largest integer m such that every n-point metric space has a subspace of cardinality m that c-embeds into a member of $\mathcal{M}$ (see [BLMN05]).

  The *Ramsey number* R_n is the minimal number of vertices of a complete graph such that any edge-coloring with n colors produces a monochromatic triangle. The following metric analog of R_n was considered in [Masc04]: the least number of points a finite metric space must contain in order to contain an equilateral triangle, i.e., to have **equilateral metric dimension** greater than two.

- **Uniform metric mapping**

  Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called a **uniform metric mapping** if there are two nondecreasing functions g_1 and g_2 from $\mathbb{R}_{\geq 0}$ to itself with $\lim_{r \to \infty} g_i(r) = \infty$ for $i = 1, 2$, such that the inequality

  $$g_1(d_X(x, y)) \leq d_Y(f(x), f(y)) \leq g_2(d_X(x, y))$$

  holds for all $x, y \in X$. A **bi-Lipschitz mapping** is a uniform metric mapping with linear functions g_1, g_2.

- **Metric compression**

  Given metric spaces (X, d_X) (unbounded) and (Y, d_Y), a function $f : X \to Y$ is a *large scale Lipschitz mapping* if, for some $c > 0, D \geq 0$ and all $x, y \in X$,

  $$d_Y(f(x), f(y)) \leq c d_X(x, y) + D.$$

  The *compression* of such a mapping f is $\rho_f(r) = \inf_{d_X(x,y) \geq r} d_Y(f(x), f(y))$.

  The **metric compression** of (X, d_X) in (Y, d_Y) is defined by

  $$R(X, Y) = \sup_f \{\underline{\lim}_{r \to \infty} \frac{\log \max\{\rho_f(r), 1\}}{\log r}\},$$

  where the supremum is over all large scale Lipschitz mappings f.

In the main interesting case—when (Y, d_Y) is a Hilbert space and (X, d_X) is a (finitely generated discrete) group with **word metric**—$R(X, Y) = 0$ if there is no (Guentner–Kaminker, 2004) **uniform metric mapping** $(X, d_X) \to (Y, d_Y)$, and $R(X, Y) = 1$ for free groups, even if there is no **quasi-isometry**. Arzhantzeva–Guba–Sapir, 2006, found groups with $\frac{1}{2} \leq R(X, Y) \leq \frac{3}{4}$.

- **Quasi-isometry**

  Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called a **quasi-isometry** (or (C, c)**-quasi-isometry**) if it holds

$$C^{-1} d_X(x, y) - c \leq d_Y(f(x), f(y)) \leq C d_X(x, y) + c,$$

  for some $C \geq 1, c \geq 0$, and $Y = \cup_{x \in X} B_{d_Y}(f(x), c)$, i.e., for every point $y \in Y$, there exists $x \in X$ such that $d_Y(y, f(x)) < \frac{c}{2}$. Quasi-isometry is an equivalence relation on metric spaces; it is a bi-Lipschitz equivalence up to small distances. Quasi-isometry means that metric spaces contain bi-Lipschitz equivalent **Delone sets**.

  A quasi-isometry with $C = 1$ is called a **coarse isometry** (or *rough isometry*, *almost isometry*). Cf. **quasi-Euclidean rank of a metric space**.

- **Coarse embedding**

  Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called a **coarse embedding** if there exist nondecreasing functions $\rho_1, \rho_2 : [0, \infty) \to [0, \infty)$ with $\rho_1(d_X(x, x')) \leq d_Y(f(x), f(x')) \leq \rho_2(d_X(x, x'))$ if $x, x' \in X$ and $\lim_{t \to \infty} \rho_1(t) = +\infty$.

  Metrics d_1, d_2 on X are called **coarsely equivalent metrics** if there exist nondecreasing functions $f, g : [0, \infty) \to [0, \infty)$ such that $d_1 \leq f(d_2), d_2 \leq g(d_1)$.

- **Metrically regular mapping**

  Let (X, d_X) and (Y, d_Y) be metric spaces, and let F be a set-valued mapping from X to Y, having *inverse* F^{-1}, i.e., with $x \in F^{-1}(y)$ if and only if $y \in F(x)$.

  The mapping F is said to be **metrically regular at** $\overline{x}$ **for** $\overline{y}$ (Dontchev–Lewis–Rockafeller, 2002) if there exists $c > 0$ such that it holds

$$d_X(x, F^{-1}(y)) \leq c d_Y(y, F(x))$$

  for all (x, y) close to $(\overline{x}, \overline{y})$. Here $d(z, A) = \inf_{a \in A} d(z, a)$ and $d(z, \emptyset) = +\infty$.

- **Contraction**

  Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called a **contraction** if the inequality

$$d_Y(f(x), f(y)) \leq c d_X(x, y)$$

  holds for all $x, y \in X$ and some real number c, $0 \leq c < 1$.

  Every contraction is a **contractive mapping**, and it is uniformly continuous. *Banach fixed point theorem* (or *contraction principle*): every contraction from a **complete** metric space into itself has a unique fixed point.

[redacted]

- **Contractive mapping**

  Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called a **contractive** (or *strictly short*, *distance-decreasing*) **mapping** if

  $$d_Y(f(x), f(y)) < d_X(x, y)$$

  holds for all different $x, y \in X$. A function $f : X \to Y$ is called a **noncontractive mapping** (or *dominating mapping*) if for all $x, y \in X$ it holds

  $$d_Y(f(x), f(y)) \geq d_X(x, y).$$

  Every noncontractive bijection from a **totally bounded** metric space onto itself is an **isometry**.

- **Short mapping**

  Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \to Y$ is called a **short** (or *1-Lipschitz*, *nonexpansive*, *distance-noninreasing*, *metric*) **mapping** (or *semicontraction*) if for all $x, y \in X$ it holds

  $$d_Y(f(x), f(y)) \leq d_X(x, y).$$

  A **submetry** is a short mapping such that the image of any metric ball is a metric ball of the same radius.

  The set of short mappings $f : X \to Y$ for bounded metric spaces (X, d_X) and (Y, d_Y) is a metric space under the **uniform metric** $\sup\{d_Y(f(x), g(x)) : x \in X\}$.

  Two subsets A and B of a metric space (X, d) are called (Gowers, 2000) **similar** if there exist short mappings $f : A \to X$, $g : B \to X$ and a small $\epsilon > 0$ such that every point of A is within ϵ of some point of B, every point of B is within ϵ of some point of A, and $|d(x, g(f(x))) - d(y, f(g(y)))| \leq \epsilon$ for any $x \in A, y \in B$.

- **Category of metric spaces**

  A *category* Ψ consists of a class $Ob(\Psi)$ of *objects* and a class $Mor(\Psi)$ of *morphisms* (or *arrows*) satisfying the following conditions:

  1. To each ordered pair of objects A, B is associated a set $\Psi(A, B)$ of morphisms, and each morphism belongs to only one set $\Psi(A, B)$;
  2. The composition $f \cdot g$ of two morphisms $f : A \to B$, $g : C \to D$ is defined if $B = C$ in which case it belongs to $\Psi(A, D)$, and it is associative;
  3. Each set $\Psi(A, A)$ contains, as an *identity*, a morphism id_A such that $f \cdot id_A = f$ and $id_A \cdot g = g$ for any morphisms $f : X \to A$ and $g : A \to Y$.

  The **category of metric spaces**, denoted by *Met* (see [Isbe64]), is a category which has metric spaces as objects and **short mappings** as morphisms. A unique **injective envelope** exists in this category for every one of its objects; it can be identified with its **tight span**. In *Met*, the *monomorphisms* are injective short mappings, and *isomorphisms* are **isometries**. *Met* is a subcategory of the category which has metric spaces as objects and **Lipschitz mappings** as morphisms.

Cf. **metric** 1**-space** on the objects of a category in Chap. 3.

- **Injective metric space**

  A metric space (X, d) is called **injective** if, for every isometric embedding $f : X \to X'$ of (X, d) into another metric space (X', d'), there exists a **short mapping** f' from X' into X with $f' \cdot f = id_X$, i.e., X is a *retract* of X'.

  Equivalently, X is an *absolute retract*, i.e., a retract of every metric space into which it embeds isometrically. A metric space (X, d) is injective if and only if it is **hyperconvex**. Examples of such metric spaces are l_1^2-space, l_∞^n-space, any **real tree** and the **tight span** of a metric space.

- **Injective envelope**

  The **injective envelope** (introduced first in [Isbe64] as *injective hull*) is a generalization of **Cauchy completion**. Given a metric space (X, d), it can be embedded isometrically into an **injective metric space** $(\hat{X}, \hat{d})$; given any such isometric embedding $f : X \to \hat{X}$, there exists a unique smallest injective subspace $(\overline{X}, \overline{d})$ of $(\hat{X}, \hat{d})$ containing $f(X)$ which is called the **injective envelope** of X. It is isometrically identified with the **tight span** of (X, d).

  A metric space coincides with its injective envelope if and only if it is injective.

- **Tight extension**

  An extension (X', d') of a metric space (X, d) is called a **tight extension** if, for every semimetric d'' on X' satisfying the conditions $d''(x_1, x_2) = d(x_1, x_2)$ for all $x_1, x_2 \in X$, and $d''(y_1, y_2) \leq d'(y_1, y_2)$ for any $y_1, y_2 \in X'$, one has $d''(y_1, y_2) = d'(y_1, y_2)$ for all $y_1, y_2 \in X'$.

  The **tight span** is the *universal tight extension* of X, i.e., it contains, up to isometries, every tight extension of X, and it has no proper tight extension itself.

- **Tight span**

  Given a metric space (X, d) of finite diameter, consider the set $\mathbb{R}^X = \{f : X \to \mathbb{R}\}$. The **tight span** $T(X, d)$ of (X, d) is defined as the set $T(X, d) = \{f \in \mathbb{R}^X : f(x) = \sup_{y \in X}(d(x, y) - f(y))$ for all $x \in X\}$, endowed with the metric induced on $T(X, d)$ by the *sup norm* $||f|| = \sup_{x \in X} |f(x)|$.

  The set X can be identified with the set $\{h_x \in T(X, d) : h_x(y) = d(y, x)\}$ or, equivalently, with the set $T^0(X, d) = \{f \in T(X, d) : 0 \in f(X)\}$. The **injective envelope** $(\overline{X}, \overline{d})$ of X is isometrically identified with the tight span $T(X, d)$ by

  $$\overline{X} \to T(X, d),\ \overline{x} \to h_{\overline{x}} \in T(X, d) : h_{\overline{x}}(y) = \overline{d}(f(y), \overline{x}).$$

  The tight span $T(X, d)$ of a finite metric space is the metric space $(T(X), D(f, g) = \max |f(x) - g(x)|)$, where $T(X)$ is the set of functions $f : X \to \mathbb{R}$ such that for any $x, y \in X$, $f(x) + f(y) \geq d(x, y)$ and, for each $x \in X$, there exists $y \in X$ with $f(x) + f(y) = d(x, y)$. The mapping of any x into the function $f_x(y) = d(x, y)$ gives an isometric embedding of (X, d) into $T(X, d)$. For example, if $X = \{x_1, x_2\}$, then $T(X, d)$ is the interval of length $d(x_1, x_2)$.

[redacted]

The tight span of a metric space (X, d) of finite diameter can be considered as a polytopal complex of bounded faces of the polyhedron

$$\{y \in \mathbb{R}^n_{\geq 0} : y_i + y_j \geq d(x_i, x_j) \text{ for } 1 \leq i < j \leq n\}$$

if, for example, $X = \{x_1, \dots, x_n\}$. The dimension of this complex is called (Dress, 1984) the **combinatorial dimension** of (X, d).

- **Real tree**

  A metric space (X, d) is called (Tits, 1977) a **real tree** (or **$\mathbb{R}$-tree**) if, for all $x, y \in X$, there exists a unique **arc** from x to y, and this arc is a **geodesic segment**. So, an $\mathbb{R}$-tree is a (uniquely) arcwise connected metric space in which each arc is isometric to a subarc of $\mathbb{R}$. $\mathbb{R}$-tree is not related to a **metric tree** in Chap. 17.

  A metric space (X, d) is a real tree if and only if it is **path-connected** and Gromov 0-**hyperbolic** (i.e., satisfies the **four-point inequality**). The plane $\mathbb{R}^2$ with the **Paris metric** or **lift metric** (Chap. 19) are examples of an $\mathbb{R}$-tree.

  Real trees are exactly **tree-like** metric spaces which are **geodesic**; they are **injective** metric spaces among tree-like spaces. Tree-like metric spaces are by definition metric subspaces of real trees.

  If (X, d) is a finite metric space, then the **tight span** $T(X, d)$ is a real tree and can be viewed as an edge-weighted graph-theoretical tree.

  A metric space is a complete real tree if and only if it is **hyperconvex** and any two points are joined by a **metric segment**.

  A geodesic metric space (X, d) is called (Druţu–Sapir, 2005) *tree-graded with respect to* a collection $\mathcal{P}$ of connected proper subsets with $|P \cap P'| \leq 1$ for any distinct $P, P' \in \mathcal{P}$, if every its simple loop composed of three geodesics is contained in one $P \in \mathcal{P}$. $\mathbb{R}$-trees are tree-graded with respect to the empty set.

1.5 General Distances

- **Discrete metric**

  Given a set X, the **discrete metric** (or **trivial metric**, **sorting distance**, **drastic distance**, **Dirac distance**, *overlap*) is a metric on X, defined by $d(x, y) = 1$ for all distinct $x, y \in X$ and $d(x, x) = 0$. Cf. the much more general notion of a (metrically or topologically) **discrete metric space**.

- **Indiscrete semimetric**

  Given a set X, the **indiscrete semimetric** d is a semimetric on X defined by $d(x, y) = 0$ for all $x, y \in X$.

- **Equidistant metric**

  Given a set X and a positive real number t, the **equidistant metric** d is a metric on X defined by $d(x, y) = t$ for all distinct $x, y \in X$ (and $d(x, x) = 0$).

- **$(1, 2) - B$-metric**

  Given a set X, the $(1, 2) - B$**-metric** d is a metric on X such that, for any $x \in X$, the number of points $y \in X$ with $d(x, y) = 1$ is at most B, and all other distances are equal to 2. The $(1,2)-B$-**metric** is the **truncated metric** of a graph with maximal vertex degree B.

- **Permutation metric**

  Given a finite set X, a metric d on it is called a **permutation metric** (or *linear arrangement metric*) if there exists a bijection $\omega : X \to \{1, \dots, |X|\}$ such that

$$d(x, y) = |\omega(x) - \omega(y)|$$

  holds for all $x, y \in X$. Even–Naor–Rao–Schieber, 2000, defined a more general **spreading metric**, i.e., any metric d on $\{1, \dots, n\}$ such that $\sum_{y \in M} d(x, y) \geq \frac{|M|(|M|+2)}{4}$ for any $1 \leq x \leq n$ and $M \subseteq \{1, \dots, n\} \setminus \{x\}$ with $|M| \geq 2$.

- **Induced metric**

  Given a metric space (X, d) and a subset $X' \subset X$, an **induced metric** (or **submetric**) is the restriction d' of d to X'. A metric space (X', d') is called a **metric subspace** of (X, d), and (X, d) is called a **metric extension** of (X', d').

- **Katětov mapping**

  Given a metric space (X, d), the mapping $f : X \to \mathbb{R}$ is a **Katětov mapping** if

$$|f(x) - f(y)| \leq d(x, y) \leq f(x) + f(y)$$

  for any $x, y \in X$, i.e., setting $d(x, z) = f(x)$ defines a one-point **metric extension** $(X \cup \{z\}, d)$ of (X, d).

  The set $E(X)$ of Katětov mappings on X is a complete metric space with metric $D(f, g) = \sup_{x \in X} |f(x) - g(x)|$; (X, d) embeds isometrically in it via the *Kuratowski mapping* $x \to d(x, .)$, with unique extension of each isometry of X to one of $E(X)$.

- **Dominating metric**

  Given metrics d and d_1 on a set X, d_1 **dominates** d if $d_1(x, y) \geq d(x, y)$ for all $x, y \in X$. Cf. **noncontractive mapping** (or *dominating mapping*).

- **Barbilian semimetric**

  Given sets X and P, the function $f : P \times X \to \mathbb{R}_{>0}$ is called an *influence* (of P over X) if for any $x, y \in X$ the ratio $g_{xy}(p) = \frac{f(p,x)}{f(p,y)}$ has a maximum when $p \in P$.

  The **Barbilian semimetric** is defined on the set X by

$$\ln \frac{\max_{p \in P} g_{xy}(p)}{\min_{p \in P} g_{xy}(p)}$$

  for any $x, y \in X$. Barbilian, 1959, proved that the above function is well defined (moreover, $\min_{p \in P} g_{xy}(p) = \frac{1}{\max_{p \in P} g_{yx}(p)}$) and is a semimetric. Also, it is a metric if the influence f is *effective*, i.e., there is no pair $x, y \in X$ such that $g_{xy}(p)$ is constant for all $p \in P$. Cf. a special case **Barbilian metric** in Chap. 6.

- **Metric transform**

  A **metric transform** is a distance obtained as a function of a given metric (cf. Chap. 4).

[redacted]

- **Complete metric**

  Given a metric space (X, d), a sequence $\{x_n\}$, $x_n \in X$, is said to have *convergence to* $x^* \in X$ if $\lim_{n\to\infty} d(x_n, x^*) = 0$, i.e., for any $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $d(x_n, x^*) < \epsilon$ for any $n > n_0$. Any sequence converges to at most one limit in X; it is not so, in general, if d is a semimetric.

  A sequence $\{x_n\}_n$, $x_n \in X$, is called a *Cauchy sequence* if, for any $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $d(x_n, x_m) < \epsilon$ for any $m, n > n_0$.

  A metric space (X, d) is called a **complete metric space** if every *Cauchy sequence* in it converges. In this case the metric d is called a **complete metric**. An example of an incomplete metric space is $(\mathbb{N}, d(m, n) = \frac{|m-n|}{mn})$.
- **Cauchy completion**

  Given a metric space (X, d), its **Cauchy completion** is a metric space (X^*, d^*) on the set X^* of all equivalence classes of *Cauchy sequences*, where the sequence $\{x_n\}_n$ is called *equivalent to* $\{y_n\}_n$ if $\lim_{n\to\infty} d(x_n, y_n) = 0$. The metric d^* is defined by

  $$d^*(x^*, y^*) = \lim_{n\to\infty} d(x_n, y_n),$$

  for any $x^*, y^* \in X^*$, where $\{z_n\}_n$ is any element in the equivalence class z^*.

  The Cauchy completion (X^*, d^*) is a unique, up to isometry, **complete** metric space, into which the metric space (X, d) embeds as a dense metric subspace.

  The Cauchy completion of the metric space $(\mathbb{Q}, |x - y|)$ of rational numbers is the *real line* $(\mathbb{R}, |x - y|)$. A **Banach space** is the Cauchy completion of a *normed vector space* $(V, ||.||)$ with the **norm metric** $||x - y||$. A **Hilbert space** corresponds to the case an *inner product norm* $||x|| = \sqrt{\langle x, x\rangle}$.
- **Perfect metric space**

  A complete metric space (X, d) is called **perfect** if every point $x \in X$ is a *limit point*, i.e., $|B(x, r) = \{y \in X : d(x, y) < r\}| > 1$ holds for any $r > 0$.

  A topological space is a **Cantor space** (i.e., *homeomorphic* to the *Cantor set* with the natural metric $|x - y|$) if and only if it is nonempty, perfect, **totally disconnected**, compact and metrizable. The totally disconnected countable metric space $(\mathbb{Q}, |x-y|)$ of rational numbers also consists only of limit points but it is not complete and not **locally compact**.

  Every proper metric ball of radius r in a metric space has diameter at most $2r$. Given a number $0 < c \leq 1$, a metric space is called a c**-uniformly perfect metric space** if this diameter is at least $2cr$. Cf. the **radii of metric space**.
- **Metrically discrete metric space**

  A metric space (X, d) is called **metrically** (or *uniformly*) **discrete** if there exists a number $r > 0$ such that $B(x, r) = \{y \in X : d(x, y) < r\} = \{x\}$ for every $x \in X$.

  (X, d) is a **topologically discrete metric space** (or a *discrete metric space*) if the underlying topological space is **discrete**, i.e., each point $x \in X$ is an *isolated point*: there exists a number $r(x) > 0$ such that $B(x, r(x)) = \{x\}$. For $X = \{\frac{1}{n} :$

$n = 1, 2, 3, \dots\}$, the metric space $(X, |x - y|)$ is topologically but not metrically discrete. Cf. **translation discrete metric** in Chap. 10.

Alternatively, a metric space (X, d) is called *discrete* if any of the following holds:

1. (Burdyuk–Burdyuk 1991) it has a proper *isolated subset*, i.e., $M \subset X$ with $\inf\{d(x, y) : x \in M, y \notin M\} > 0$ (any such space admits a unique decomposition into *continuous*, i.e., nondiscrete, components);
2. (Lebedeva–Sergienko–Soltan, 1984) for any distinct points $x, y \in X$, there exists a point z of the **closed metric interval** $I(x, y)$ with $I(x, z) = \{x, z\}$;
3. a stronger property holds: for any two distinct points $x, y \in X$, every sequence of points $z_1, z_2, \dots$ with $z_k \in I(x, y)$ but $z_{k+1} \in I(x, z_k) \setminus \{z_k\}$ for $k = 1, 2, \dots$ is a finite sequence.

- **Locally finite metric space**

  Let (X, d) be a **metrically discrete metric space**. Then it is called **locally finite** if for every $x \in X$ and every $r \geq 0$, the ball $|B(x, r)|$ is finite.

  If, moreover, $|B(x, r)| \leq C(r)$ for some number $C(r)$ depending only on r, then (X, d) is said to have *bounded geometry*.
- **Bounded metric space**

  A metric (moreover, a distance) d on a set X is called **bounded** if there exists a constant $C > 0$ such that $d(x, y) \leq C$ for any $x, y \in X$.

  For example, given a metric d on X, the metric D on X, defined by $D(x, y) = \frac{d(x,y)}{1+d(x,y)}$, is bounded with $C = 1$.

  A metric space (X, d) with a bounded metric d is called a **bounded metric space**.
- **Totally bounded metric space**

  A metric space (X, d) is called **totally bounded** if, for every $\epsilon > 0$, there exists a finite **ϵ-net**, i.e., a finite subset $M \subset X$ with the **point-set distance** $d(x, M) < \epsilon$ for any $x \in X$ (cf. **totally bounded space** in Chap. 2).

  Every totally bounded metric space is **bounded** and **separable**. A metric space is totally bounded if and only if its **Cauchy completion** is **compact**.
- **Separable metric space**

  A metric space (X, d) is called **separable** if it contains a countable **dense subset** M, i.e., a subset with which all its elements can be approached: X is the *closure* $cl(M)$ (M together with all its limit points).

  A metric space is separable if and only if it is **second-countable** (cf. Chap. 2).
- **Compact metric space**

  A **compact metric space** (or **metric compactum**) is a metric space in which every sequence has a *Cauchy subsequence*, and those subsequences are convergent. A metric space is compact if and only if it is **totally bounded** and **complete**.

  Every bounded and closed subset of a Euclidean space is compact. Every finite metric space is compact. Every compact metric space is **second-countable**.

  A **continuum** is a nonempty **connected** metric compactum.

[redacted]

- **Proper metric space**

  A metric space is called **proper** (or *finitely compact*, *having the Heine–Borel property*) if every its closed metric ball is compact. Any such space is **complete**.
- **UC metric space**

  A metric space is called a **UC metric space** (or *Atsuji space*) if any continuous function from it into an arbitrary metric space is *uniformly continuous*.

  Every such space is **complete**. Every **metric compactum** is a UC metric space.
- **Metric measure space**

  A **metric measure space** (or *mm-space*, *metric triple*) is a triple (X, d, μ), where (X, d) is a *Polish* (i.e., complete separable; cf. Chap. 2) *metric space* and (X, Σ, μ) is a *probability measure space* ($\mu(X) = 1$) with Σ being a *Borel* σ-*algebra* of all open and closed sets of the **metric topology** (Chap. 2) induced by the metric d on X. Cf. **metric outer measure**.
- **Norm metric**

  Given a *normed vector space* $(V, ||.||)$, the **norm metric** on V is defined by

$$||x - y||.$$

  The metric space $(V, ||x - y||)$ is called a **Banach space** if it is **complete**. Examples of norm metrics are l_p**- and** L_p**-metrics**, in particular, the **Euclidean metric**.

  Any metric space (X, d) admits an isometric embedding into a Banach space B such that its convex hull is dense in B (cf. **Monge–Kantorovich metric** in Chap. 14); (X, d) is a **linearly rigid metric space** if such an embedding is unique up to isometry. A metric space isometrically embeds into the unit sphere of a Banach space if and only if its diameter is at most 2.
- **Path metric**

  Given a connected graph $G = (V, E)$, its **path metric** (or *graphic metric*) d_{path} is a metric on V defined as the length (i.e., the number of edges) of a shortest path connecting two given vertices x and y from V (cf. Chap. 15).
- **Editing metric**

  Given a finite set X and a finite set $\mathcal{O}$ of (unary) *editing operations* on X, the **editing metric** on X is the **path metric** of the graph with the vertex-set X and xy being an edge if y can be obtained from x by one of the operations from $\mathcal{O}$.
- **Gallery metric**

  A *chamber system* is a set X (its elements are called *chambers*) equipped with n equivalence relations $\sim_i$, $1 \leq i \leq n$. A *gallery* is a sequence of chambers $x_1, \dots, x_m$ such that $x_i \sim_j x_{i+1}$ for every i and some j depending on i.

  The **gallery metric** is an **extended metric** on X which is the length of the shortest gallery connecting x and $y \in X$ (and is equal to ∞ if there is no connecting gallery). The gallery metric is the (extended) **path metric** of the graph with the vertex-set X and xy being an edge if $x \sim_i y$ for some $1 \leq i \leq n$.

- **Metric on incidence structure**

  An *incidence structure* (P, L, I) consists of 3 sets: points P, lines L and flags $I \subset P \times L$, where a point $p \in P$ is said to be *incident* with a line $l \in L$ if $(p, l) \in I$.

  If, moreover, for any pair of distinct points, there is at most one line incident with both of them, then the collinearity graph is a graph whose vertices are the points with two vertices being adjacent if they determine a line.

  The **metric on incidence structure** is the **path metric** of this graph.
- **Riemannian metric**

  Given a connected n-dimensional smooth *manifold* M^n (cf. Chaps. 2 and 7), its **Riemannian metric** is a collection of positive-definite symmetric bilinear forms $((g_{ij}))$ on the tangent spaces of M^n which varies smoothly from point to point.

  The length of a curve γ on M^n is expressed as $\int_\gamma \sqrt{\sum_{i,j} g_{ij} dx_i dx_j}$, and the **intrinsic metric** on M^n, also called the **Riemannian distance**, is the infimum of lengths of curves connecting any two given points $x, y \in M^n$. Cf. Chap. 7.
- **Linearly additive metric**

  A **linearly additive** (or *additive on lines*) **metric** is a continuous metric d on $\mathbb{R}^n$ which, for any points x, y, z lying in that order on a common line, satisfies

  $$d(x, z) = d(x, y) + d(y, z).$$

  Hilbert's 4-th problem asked in 1900 to classify such metrics; it is solved only for dimension $n = 2$ ([Amba76]). Cf. **projective metric** in Chap. 6.

  Every **norm metric** on $\mathbb{R}^n$ is linearly additive. Every linearly additive metric on $\mathbb{R}^2$ is a **hypermetric**.
- **Hamming metric**

  The **Hamming metric** d_H (called sometimes *Dalal distance* in Semantics) is a metric on $\mathbb{R}^n$ defined (Hamming, 1950) by

  $$|\{i : 1 \leq i \leq n, x_i \neq y_i\}|.$$

  On binary vectors $x, y \in \{0, 1\}^n$ the Hamming metric and the l_1*-metric* (cf. **L_p-metric** in Chap. 5) coincide; they are equal to $|I(x) \Delta I(y)| = |I(x) \setminus I(y)| + |I(y) \setminus I(x)|$, where $I(z) = \{1 \leq t \leq n : z_i = 1\}$.

  In fact, $\max\{|I(x) \setminus I(y)|, |I(y) \setminus I(x)|\}$ is also a metric.
- **Lee metric**

  Given $m, n \in \mathbb{N}$, $m \geq 2$, the **Lee metric** d_{Lee} is a metric on $\mathbb{Z}_m^n = \{0, 1, \ldots, m-1\}^n$ defined (Lee, 1958) by

  $$\sum_{1 \leq i \leq n} \min\{|x_i - y_i|, m - |x_i - y_i|\}.$$

  The metric space (Z_m^n, d_{Lee}) is a discrete analog of the *elliptic space*.

[redacted]

The Lee metric coincides with the Hamming metric d_H if $m = 2$ or $m = 3$. The metric spaces (Z_4^n, d_{Lee}) and $Z_2^{2n}, d_H)$ are isometric. Lee and Hamming metrics are applied for phase and orthogonal modulation, respectively.

Cf. **absolute summation distance** and **generalized Lee metric** in Chap. 16.

- **Enomoto–Katona metric**

Given a finite set X and an integer k, $2k \leq |X|$, the **Enomoto–Katona metric** (2001) is the distance between unordered pairs (X_1, X_2) and (Y_1, Y_2) of disjoint k-subsets of X defined by

$$\min\{|X_1 \setminus Y_1| + |X_2 \setminus Y_2|, |X_1 \setminus Y_2| + |X_2 \setminus Y_1|\}.$$

Cf. **Earth Mover's distance**, **transportation distance** in Chaps. 21 and 14.

- **Symmetric difference metric**

Given a *measure space* $(\Omega, \mathcal{A}, \mu)$, the **symmetric difference** (or *measure*) *semimetric* on the set $\mathcal{A}_\mu = \{A \in \mathcal{A} : \mu(A) < \infty\}$ is defined by

$$od_\triangle(A, B) = \mu(A \triangle B),$$

where $A \triangle B = (A \cup B) \backslash (A \cap B)$ is the *symmetric difference* of A and $B \in \mathcal{A}_\mu$.

The value $d_\triangle(A, B) = 0$ if and only if $\mu(A \triangle B) = 0$, i.e., A and B are equal *almost everywhere*. Identifying two sets $A, B \in \mathcal{A}_\mu$ if $\mu(A \triangle B) = 0$, we obtain the **symmetric difference metric** (or **Fréchet–Nikodym–Aronszyan distance**, **measure metric**).

If μ is the *cardinality measure*, i.e., $\mu(A) = |A|$, then $d_\triangle(A, B) = |A \triangle B| = |A \setminus B| + |B \setminus A|$. In this case $|A \triangle B| = 0$ if and only if $A = B$.

The metrics $d_{\max}(A, B) = \max(|A \setminus B|, |B \setminus A|)$ and $1 - \frac{|A \cap B|}{\max(|A|, |B|)}$ (its normalised version) are special cases of **Zelinka distance** and **Bunke–Shearer metric** in Chap. 15. For each $p \geq 1$, the **p-difference metric** (Noradam–Nyblom, 2014) is $d_p(A, B) = (|A \setminus B|^p + |B \setminus A|^p)^{\frac{1}{p}}$; so, $d_1 = d_\triangle$ and $\lim_{p \to \infty} d_p = d_{\max}$.

The **Johnson distance** between k-sets A and B is $\frac{|A \triangle B|}{2} = k - |A \cap B|$.

The *symmetric difference metric between ordered q-partitions* $A = (A_1, \ldots, A_q)$ and $B = (B_1, \ldots, B_q)$ is $\sum_{i=1}^{q} |A_i \Delta B_i|$. Cf. **metrics between partitions** in Chap. 10.

- **Steinhaus distance**

Given a *measure space* $(\Omega, \mathcal{A}, \mu)$, the **Steinhaus distance** d_{St} is a semimetric on the set $\mathcal{A}_\mu = \{A \in \mathcal{A} : \mu(A) < \infty\}$ defined as 0 if $\mu(A) = \mu(B) = 0$, and by

$$\frac{\mu(A \triangle B)}{\mu(A \cup B)} = 1 - \frac{\mu(A \cap B)}{\mu(A \cup B)}$$

if $\mu(A \cup B) > 0$. It becomes a metric on the set of equivalence classes of elements from $\mathcal{A}_\mu$; here $A, B \in \mathcal{A}_\mu$ are called *equivalent* if $\mu(A \triangle B) = 0$.

The **biotope** (or **Tanimoto**) **distance** $\frac{|A \triangle B|}{|A \cup B|}$ is the special case of Steinhaus distance obtained for the *cardinality measure* $\mu(A) = |A|$ for finite sets.

Cf. also the **generalized biotope transform metric** in Chap. 4.

- **Fréchet metric**

  Let (X, d) be a metric space. Consider a set $\mathcal{F}$ of all continuous mappings $f : A \to X, g : B \to X, \dots$, where $A, B, \dots$ are subsets of $\mathbb{R}^n$, homeomorphic to $[0, 1]^n$ for a fixed dimension $n \in \mathbb{N}$.

  The *Fréchet semimetric* d_F is a semimetric on $\mathcal{F}$ defined by

$$\inf_{\sigma} \sup_{x \in A} d(f(x), g(\sigma(x))),$$

  where the infimum is taken over all orientation preserving homeomorphisms $\sigma : A \to B$. It becomes the **Fréchet metric** on the set of equivalence classes $f^* = \{g : d_F(g,f) = 0\}$. Cf. the **Fréchet surface metric** in Chap. 8.

- **Hausdorff metric**

  Given a metric space (X, d), the **Hausdorff metric** (or *two-sided Hausdorff distance*) is a metric on the family $\mathcal{F}$ of nonempty compact subsets of X defined by

$$d_{Haus} = \max\{d_{dHaus}(A, B), d_{dHaus}(B, A)\},$$

  where $d_{dHaus}(A, B) = \max_{x \in A} \min_{y \in B} d(x, y)$ is the **directed Hausdorff distance** (or *one-sided Hausdorff distance*) from A to B. The metric space $(\mathcal{F}, d_{Haus})$ is called **hyperspace of metric space** (X, d); cf. **hyperspace** in Chap. 2.

  In other words, $d_{Haus}(A, B)$ is the minimal number ϵ (called also the **Blaschke distance**) such that a closed **ϵ-neighborhood** of A contains B and a closed ϵ-neighborhood of B contains A. Then $d_{Haus}(A, B)$ is equal to

$$\sup_{x \in X} |d(x, A) - d(x, B)|,$$

  where $d(x, A) = \min_{y \in A} d(x, y)$ is the **point-set distance**.

  If the above definition is extended for noncompact closed subsets A and B of X, then $d_{Haus}(A, B)$ can be infinite, i.e., it becomes an **extended metric**.

  For not necessarily closed subsets A and B of X, the **Hausdorff semimetric** between them is defined as the Hausdorff metric between their closures. If X is finite, d_{Haus} is a metric on the class of all subsets of X.

- **L_p-Hausdorff distance**

  Given a finite metric space (X, d), the **L_p-Hausdorff distance** ([Badd92]) between two subsets A and B of X is defined by

$$(\sum_{x \in X} |d(x, A) - d(x, B)|^p)^{\frac{1}{p}},$$

[redacted]

where $d(x, A)$ is the **point-set distance**. The usual **Hausdorff metric** corresponds to the case $p = \infty$.

- **Generalized G-Hausdorff metric**

  Given a group $(G, \cdot, e)$ acting on a metric space (X, d), the **generalized G-Hausdorff metric** between two closed bounded subsets A and B of X is

$$\min_{g_1, g_2 \in G} d_{Haus}(g_1(A), g_2(B)),$$

  where d_{Haus} is the **Hausdorff metric**. If $d(g(x), g(y)) = d(x, y)$ for any $g \in G$ (i.e., if the metric d is *left-invariant* with respect of G), then above metric is equal to $\min_{g \in G} d_{Haus}(A, g(B))$.

- **Gromov–Hausdorff metric**

  The **Gromov–Hausdorff metric** is a metric on the set of all *isometry classes* of compact metric spaces defined by

$$\inf d_{Haus}(f(X), g(Y))$$

  for any two classes X^* and Y^* with the representatives X and Y, respectively, where d_{Haus} is the **Hausdorff metric**, and the minimum is taken over all metric spaces M and all isometric embeddings $f : X \to M$, $g : Y \to M$. The corresponding metric space is called the *Gromov–Hausdorff space*.

  The **Hausdorff–Lipschitz distance** between isometry classes of compact metric spaces X and Y is defined by

$$\inf\{d_{GH}(X, X_1) + d_L(X_1, Y_1) + d_{GH}(Y, Y_1)\},$$

  where d_{GH} is the Gromov–Hausdorff metric, d_L is the **Lipschitz metric**, and the minimum is taken over all (isometry classes of compact) metric spaces X_1, Y_1.

- **Kadets distance**

  The *gap* (or *opening*) between two closed subspaces X and Y of a Banach space $(V, ||.||)$ is defined by

$$gap(X, Y) = \max\{\delta(X, Y), \delta(Y, X)\},$$

  where $\delta(X, Y) = \sup\{\inf_{y \in Y} ||x - y|| : x \in X, ||x|| = 1\}$ (cf. **gap distance** in Chap. 12 and **gap metric** in Chap. 18).

  The **Kadets distance** between two Banach spaces V and W is a semimetric defined (Kadets, 1975) by

$$\inf_{Z, f, g} gap(\overline{B}_{f(V)}, \overline{B}_{g(W)}),$$

  where the infimum is taken over all Banach spaces Z and all linear isometric embeddings $f : V \to Z$ and $g : W \to Z$; here $\overline{B}_{f(V)}$ and $\overline{B}_{g(W)}$ are the closed unit balls of Banach spaces $f(V)$ and $g(W)$, respectively.

The nonlinear analog of the Kadets distance is the following **Gromov–Hausdorff distance between Banach spaces** U and W:

$$\inf_{Z,f,g} d_{Haus}(f(\overline{B}_V), g(\overline{B}_W)),$$

where the infimum is taken over all metric spaces Z and all isometric embeddings $f : V \to Z$ and $g : W \to Z$; here d_{Haus} is the **Hausdorff metric**.

The **Kadets path distance** between Banach spaces V and W is defined (Ostrovskii, 2000) as the infimum of the lengths (with respect to the Kadets distance) of all curves joining V and W (and is equal to ∞ if there is no such curve).

- **Banach–Mazur distance**

The **Banach–Mazur distance** d_{BM} between two Banach spaces V and W is

$$\ln \inf_T ||T|| \cdot ||T^{-1}||,$$

where the infimum is taken over all isomorphisms $T : V \to W$.

It can also be written as $\ln d(V, W)$, where the number $d(V, W)$ is the smallest positive $d \geq 1$ such that $\overline{B}^n_W \subset T(\overline{B}^n_V) \subset d\overline{B}^n_W$ for some linear invertible transformation $T : V \to W$. Here $\overline{B}^n_V = \{x \in V : ||x||_V \leq 1\}$ and $\overline{B}^n_W = \{x \in W; ||x||_W \leq 1\}$ are the *unit balls* of the normed spaces $(V, ||.||_V)$ and $(W, ||.||_W)$, respectively.

One has $d_{BM}(V, W) = 0$ if and only if V and W are *isometric*, and d_{BM} becomes a metric on the set X^n of all equivalence classes of n-dimensional normed spaces, where $V \sim W$ if they are isometric. The pair (X^n, d_{BM}) is a compact metric space which is called the **Banach–Mazur compactum**.

The **modified Banach–Mazur distance** (Glushkin, 1963, and Khrabrov, 2001) is

$$\inf\{||T||_{X \to Y} : |detT| = 1\} \cdot \inf\{||T||_{Y \to X} : |detT| = 1\}.$$

The **weak Banach–Mazur distance** (Tomczak–Jaegermann, 1984) is

$$\max\{\overline{\gamma}_Y(id_X), \overline{\gamma}_X(id_Y)\},$$

where *id* is the identity map and, for an operator $U : X \to Y$, $\overline{\gamma}_Z(U)$ denotes $\inf \sum ||W_k|| \cdot ||V_k||$. Here the infimum is taken over all representations $U = \sum W_k V_k$ for $W_k : X \to Z$ and $V_k : Z \to Y$. This distance never exceeds the corresponding Banach–Mazur distance.

[redacted]

- **Lipschitz distance**

  Given $\alpha \geq 0$ and two metric spaces (X, d_X), (Y, d_Y), the *α-Hölder norm* $||.||_{Hol}$ on the set of all injective functions $f : X \to Y$ is defined by

  $$||f||_{Hol} = \sup_{x,y \in X, x \neq y} \frac{d_Y(f(x), f(y))}{d_X(x,y)^{\alpha}}.$$

  The *Lipschitz norm* $||.||_{Lip}$ is the case $\alpha = 1$ of $||.||_{Hol}$.

  The **Lipschitz distance** between metric spaces (X, d_X) and (Y, d_Y) is defined by

  $$\ln \inf_f ||f||_{Lip} \cdot ||f^{-1}||_{Lip},$$

  where the infimum is taken over all bijective functions $f : X \to Y$. Equivalently, it is the infimum of numbers $\ln a$ such that there exists a bijective **bi-Lipschitz mapping** between (X, d_X) and (Y, d_Y) with constants $\exp(-a)$, $\exp(a)$.

  It becomes a metric (**Lipschitz metric**) on the set of all isometry classes of compact metric spaces. Cf. **Hausdorff–Lipschitz distance**.

  This distance is an analog to the **Banach–Mazur distance** and, in the case of finite-dimensional real Banach spaces, coincides with it.

  It also coincides with the **Hilbert projective metric** on *nonnegative* projective spaces, obtained by starting with $\mathbb{R}^n_{>0}$ and identifying any point x with cx, $c > 0$.

- **Lipschitz distance between measures**

  Given a compact metric space (X, d), the *Lipschitz seminorm* $||.||_{Lip}$ on the set of all functions $f : X \to \mathbb{R}$ is defined by $||f||_{Lip} = \sup_{x,y \in X, x \neq y} \frac{|f(x)-f(y)|}{d(x,y)}$.

  The **Lipschitz distance between measures** μ and ν on X is defined by

  $$\sup_{||f||_{Lip} \leq 1} \int f d(\mu - \nu).$$

  It is the **transportation distance** (Chap. 14) if μ, ν are probability measures. Let a such measure $m_x(.)$ be attached to any $x \in X$; for distinct x, y the *coarse Ricci curvature along* (xy) is defined (Ollivier, 2009) as $\kappa(x,y) = 1 - \frac{W_1(m_x, m_y)}{d(x,y)}$. *Ollivier's curvature* generalizes the *Ricci curvature* in Riemannian space (cf. Chap. 7).

- **Barycentric metric space**

  Given a metric space (X, d), let $(B(X), ||\mu - \nu||_{TV})$ be the metric space, where $B(X)$ is the set of all regular Borel probability measures on X with bounded support, and $||\mu - \nu||_{TV}$ is the **variational distance** $\int_X |p(\mu) - p(\nu)| d\lambda$ (cf. Chap. 14). Here $p(\mu)$ and $p(\nu)$ are the density functions of measures μ and ν, respectively, with respect to the σ-finite measure $\frac{\mu+\nu}{2}$.

A metric space (X, d) is **barycentric** if there exists a constant $\beta > 0$ and a surjection $f : B(X) \rightarrow X$ such that for any measures $\mu, \nu \in B(X)$ it holds the inequality

$$d(f(\mu), f(\nu)) \leq \beta diam(supp(\mu + \nu))||\mu - \nu||_{TV}.$$

Any Banach space $(X, d = ||x - y||)$ is a barycentric metric space with the smallest β being 1 and the map $f(\mu)$ being the usual *center of mass* $\int_X x d\mu(x)$.

Any **Hadamard** (i.e., a complete **CAT**(0) **space**, cf. Chap. 6, is barycentric with the smallest β being 1 and the map $f(\mu)$ being the unique minimizer of the function $g(y) = \int_X d^2(x, y) d\mu(x)$ on X.

- **Point-set distance**

  Given a metric space (X, d), the **point-set distance** $d(x, A)$ between a point $x \in X$ and a subset A of X is defined as

  $$\inf_{y \in A} d(x, y).$$

  For any $x, y \in X$ and for any nonempty subset A of X, we have the following version of the triangle inequality: $d(x, A) \leq d(x, y) + d(y, A)$ (cf. **distance map**).

  For a given point-measure $\mu(x)$ on X and a *penalty function* p, an **optimal quantizer** is a set $B \subset X$ such that $\int p(d(x, B)) d\mu(x)$ is as small as possible.

- **Set-set distance**

  Given a metric space (X, d), the **set-set distance** between two subsets A and B of X is defined by

  $$d_{ss}(A, B) = \inf_{x \in A, y \in B} d(x, y).$$

  This distance can be 0 even for disjoint sets, for example, for the intervals $(1, 2)$, $(2, 3)$ on $\mathbb{R}$. The sets A and B are *positively separated* if $d_{ss}(A, B) > 0$. A constructive *appartness space* is a generalization of this relation on subsets of X.

  The **spanning distance** between A and B is $\sup_{x \in A, y \in B} d(x, y)$.

  In Data Analysis, (cf. Chap. 17) the set-set and spanning distances between clusters are called the **single** and **complete linkage**, respectively.

- **Matching distance**

  Given a metric space (X, d), the **matching distance** (or *multiset-multiset distance*) between two multisets A and B in X is defined by

  $$\inf_{\phi} \max_{x \in A} d(x, \phi(x)),$$

  where ϕ runs over all bijections between A and B, as multisets.

[redacted]

The matching distance is not related to the **perfect matching distance** in Chap. 15 and to the **nonlinear elastic matching distance** in Chap. 21. But the **bottleneck distance** in Chap. 21 is a special case of it.

- **Metrics between multisets**

  A *multiset* (or *bag*) drawn from a set S is a mapping $m : S \to \mathbb{Z}_{\geq 0}$, where $m(x)$ represents the "multiplicity" of $x \in S$. The *dimensionality*, *cardinality* and *height* of multiset m is $|S|$, $|m| = \sum_{x \in S} m(x)$ and $\max_{x \in S} m(x)$, respectively.

  Multisets are good models for multi-attribute objects such as, say, all symbols in a string, all words in a document, etc.

  A multiset m is finite if S and all $m(x)$ are finite; the *complement* of a finite multiset m is the multiset $\overline{m} : S \to \mathbb{Z}_{\geq 0}$, where $\overline{m}(x) = \max_{y \in S} m(y) - m(x)$. Given two multisets m_1 and m_2, denote by $m_1 \cup m_2$, $m_1 \cap m_2$, $m_1 \backslash m_2$ and $m_1 \Delta m_2$ the multisets on S defined , for any $x \in S$, by $m_1 \cup m_2(x) = \max\{m_1(x), m_2(x)\}$, $m_1 \cap m_2(x) = \min\{m_1(x), m_2(x)\}$, $m_1 \backslash m_2(x) = \max\{0, m_1(x) - m_2(x)\}$ and $m_1 \Delta m_2(x) = |m_1(x) - m_2(x)|$, respectively. Also, $m_1 \subseteq m_2$ denotes that $m_1(x) \leq m_2(x)$ for all $x \in S$.

  The *measure* $\mu(m)$ of a multiset m is a linear combination $\mu(m) = \sum_{x \in S} \lambda(x) m(x)$ with $\lambda(x) \geq 0$. In particular, $|m|$ is the *counting measure*.

  For any measure $\mu(m) \in \mathbb{R}_{\geq 0}$, Miyamoto, 1990, and Petrovsky, 2003, proposed several **semimetrics between multisets** m_1 and m_2 including $d_1(m_1, m_2) = \mu(m_1 \Delta m_2)$ and $d_2(m_1, m_2) = \frac{\mu(m_1 \Delta m_2)}{\mu(m_1 \cup m_2)}$ (with $d_2(\emptyset, \emptyset) = 0$ by definition). Cf. **symmetric difference metric** and **Steinhaus distance**.

  Among examples of other metrics between multisets are **matching distance**, **metric space of roots** in Chap. 12, **μ-metric** in Chap. 15 and, in Chap. 11, **bag distance** $\max\{|m_1 \backslash m_2|, |m_2 \backslash m_1|\}$ and **q-gram similarity**.

  See also **Vitanyi multiset metric** in Chap. 3.

- **Metrics between fuzzy sets**

  A *fuzzy subset* of a set S is a mapping $\mu : S \to [0, 1]$, where $\mu(x)$ represents the "degree of membership" of $x \in S$. It is an ordinary (*crisp*) if all $\mu(x)$ are 0 or 1. Fuzzy sets are good models for *gray scale images* (cf. **gray scale images distances** in Chap. 21), random objects and objects with nonsharp boundaries.

  Bhutani–Rosenfeld, 2003, introduced the following two metrics between two fuzzy subsets μ and ν of a finite set S. The **diff-dissimilarity** is a metric (a fuzzy generalization of **Hamming metric**), defined by

$$d(\mu, \nu) = \sum_{x \in S} |\mu(x) - \nu(x)|.$$

  The **perm-dissimilarity** is a semimetric defined by

$$\min\{d(\mu, p(\nu))\},$$

  where the minimum is taken over all permutations p of S.

The **Chaudhuri–Rosenfeld metric** (1996) between two fuzzy sets μ and ν *with crisp points* (i.e., the sets $\{x \in S : \mu(x) = 1\}$ and $\{x \in S : \nu(x) = 1\}$ are nonempty) is an **extended metric**, defined the **Hausdorff metric** d_{Haus} by

$$\int_0^1 2t d_{Haus}(\{x \in S : \mu(x) \geq t\}, \{x \in S : \nu(x) \geq t\}) dt.$$

A *fuzzy number* is a fuzzy subset μ of the real line $\mathbb{R}$, such that the *level set* (or *t-cut*) $A_\mu(t) = \{x \in \mathbb{R} : \mu(x) \geq t\}$ is convex for every $t \in [0, 1]$. The *sendograph* of a fuzzy set μ is the set $send(\mu) = \{(x, t) \in S \times [0, 1] : \mu(x) > 0,\ \mu(x) \geq t\}$. The **sendograph metric** (Kloeden, 1980) between two fuzzy numbers μ, ν with crisp points and compact sendographs is the **Hausdorff metric**

$$\max\{\sup_{a=(x,t)\in send(\mu)} d(a, send(\nu)), \sup_{b=(x',t')\in send(\nu)} d(b, send(\mu))\},$$

where $d(a, b) = d((x, t), (x', t'))$ is a **box metric** (Chap. 4) $\max\{|x - x'|, |t - t'|\}$.

The **Klement–Puri–Ralesku metric** (1988) between fuzzy numbers μ, ν is

$$\int_0^1 d_{Haus}(A_\mu(t), A_\nu(t)) dt,$$

where $d_{Haus}(A_\mu(t), A_\nu(t))$ is the **Hausdorff metric**

$$\max\{\sup_{x\in A_\mu(t)} \inf_{y\in A_\nu(t)} |x - y|, \sup_{x\in A_\nu(t)} \inf_{x\in A_\mu(t)} |x - y|\}.$$

Several other Hausdorff-like metrics on some families of fuzzy sets were proposed by Boxer in 1997, Fan in 1998 and Brass in 2002; Brass also argued the nonexistence of a "good" such metric.

If q is a quasi-metric on $[0, 1]$ and S is a finite set, then $Q(\mu, \nu) = \sup_{x\in S} q(\mu(x), \nu(x))$ is a quasi-metric on fuzzy subsets of S.

Cf. **fuzzy Hamming distance** in Chap. 11 and, in Chap. 23, **fuzzy set distance** and **fuzzy polynucleotide metric**. Cf. **fuzzy metric spaces** in Chap. 3 for fuzzy-valued generalizations of metrics and, for example, [Bloc99] for a survey.

- **Metrics between intuitionistic fuzzy sets**

An *intuitionistic fuzzy subset* of a set S is (Atanassov, 1999) an ordered pair of mappings $\mu, \nu :\to [0, 1]$ with $0 \leq \mu(x) + \nu(x) \leq 1$ for all $x \in S$, representing the "degree of membership" and the "degree of nonmembership" of $x \in S$, respectively. It is an ordinary *fuzzy subset* if $\mu(x) + \nu(x) = 1$ for all $x \in S$.

Given two intuitionistic fuzzy subsets $(\mu(x), \nu(x))$ and $(\mu'(x), \nu'(x))$ of a finite set $S = \{x_1, \dots, x_n\}$, their **Atanassov distances** (1999) are:

$$\frac{1}{2}\sum_{i=1}^{n}(|\mu(x_i) - \mu'(x_i)| + |\nu(x_i) - \nu'(x_i)|) \text{ (Hamming distance)}$$

[redacted]

and, in general, for any given numbers $p \geq 1$ and $0 \leq q \leq 1$, the distance

$$(\sum_{i=1}^{n}(1-q)(\mu(x_i)-\mu'(x_i))^p + q(\nu(x_i)-\nu'(x_i)^p)^{\frac{1}{p}}.$$

Their **Grzegorzewski distances** (2004) are:

$$\sum_{i=1}^{n}\max\{|\mu(x_i)-\mu'(x_i)|, |\nu(x_i)-\nu'(x_i)|\} \text{ (Hamming distance);}$$

$$\sqrt{\sum_{i=1}^{n}\max\{(\mu(x_i)-\mu'(x_i))^2, (\nu(x_i)-\nu'(x_i))^2\}} \text{ (Euclidean distance).}$$

The normalized versions (dividing the above sums by n) were also proposed.

Szmidt–Kacprzyk, 1997, proposed a modification of the above, adding $\pi(x) - \pi'(x)$, where $\pi(x)$ is the third mapping $1-\mu(x)-\nu(x)$.

An *interval-valued fuzzy subset* of a set S is a mapping $\mu :\rightarrow [I]$, where $[I]$ is the set of closed intervals $[a^-, a^+] \subseteq [0,1]$. Let $\mu(x) = [\mu^-(x), \mu^+(x)]$, where $0 \leq \mu^-(x) \leq \mu^+(x) \leq 1$ and an interval-valued fuzzy subset is an ordered pair of mappings μ^- and μ^+. This notion is close to the above intuitionistic one; so, above distance can easily be adapted. For example, $\sum_{i=1}^{n}\max\{|\mu^-(x_i) - \mu'^-(x_i)|, |\mu^+(x_i) - \mu'^+(x_i)|\}$ is a Hamming distance between interval-valued fuzzy subsets (μ^-, μ^+) and (μ'^-, μ'^+).

- **Polynomial metric space**

  Let (X,d) be a metric space with a finite diameter D and a finite normalized measure μ_X. Let the Hilbert space $L_2(X,d)$ of complex-valued functions decompose into a countable (when X is infinite) or a finite (with $D+1$ members when X is finite) direct sum of mutually orthogonal subspaces $L_2(X,d) = V_0 \oplus V_1 \oplus \dots$.

  Then (X,d) is a **polynomial metric space** if there exists an ordering of the spaces $V_0, V_1, \dots$ such that, for $i = 0, 1, \dots,$ there exist *zonal spherical functions*, i.e., real polynomials $Q_i(t)$ of degree i such that

$$Q_i(t(d(x,y))) = \frac{1}{r_i}\sum_{j=1}^{r_i} v_{ij}(x)\overline{v_{ij}(y)}$$

  for all $x, y \in X$, where r_i is the dimension of V_i, $\{v_{ii}(x) : 1 \leq j \leq r_i\}$ is an orthonormal basis of V_i, and $t(d)$ is a continuous decreasing real function such that $t(0) = 1$ and $t(D) = -1$. The zonal spherical functions constitute an orthogonal system of polynomials with respect to some weight $w(t)$.

  The finite polynomial metric spaces are also called *(P and Q)-polynomial association schemes*; cf. **distance-regular graph** in Chap. 15. The infinite polynomial metric spaces are the *compact connected two-point homogeneous* *spaces*. Wang, 1952, classified them as the Euclidean unit spheres, the real, complex, quaternionic projective spaces or the Cayley projective line and plane.

- **Universal metric space**

  A metric space (U, d) is called **universal** for a collection $\mathcal{M}$ of metric spaces if any metric space (M, d_M) from $\mathcal{M}$ is *isometrically embeddable* in (U, d), i.e., there exists a mapping $f : M \to U$ which satisfies $d_M(x, y) = d(f(x), f(y))$ for any $x, y \in M$. Some examples follow.

  Every separable metric space (X, d) isometrically embeds (Fréchet, 1909) in (a nonseparable) **Banach space** l^{∞}_{∞}. In fact, $d(x, y) = \sup_i |d(x, a_i) - d(y, a_i)|$, where $(a_1, \dots, a_i, \dots)$ is a dense countable subset of X.

  Every metric space isometrically embeds (Kuratowski, 1935) in the **Banach space** $L^{\infty}(X)$ of bounded functions $f : X \to \mathbb{R}$ with the norm $\sup_{x \in X} |f(x)|$.

  The **Urysohn space** is a **homogeneous** complete separable space which is the universal metric space for all separable metric spaces. The **Hilbert cube** (Chap. 4) is the universal space for the class of metric spaces with a countable base.

  The **graphic** metric space of the **random graph** (Rado, 1964; the vertex-set consists of all prime numbers $p \equiv 1 \pmod 4$ with pq being an edge if p is a quadratic residue modulo q) is the universal metric space for any finite or countable metric space with distances 0, 1 and 2 only. It is a discrete analog of the Urysohn space.

  There exists a metric d on $\mathbb{R}$, inducing the usual (interval) topology, such that $(\mathbb{R}, d)$ is a universal metric space for all finite metric spaces (Holsztynski, 1978). The Banach space l^n_{∞} is a universal metric space for all metric spaces (X, d) with $|X| \leq n + 2$ (Wolfe, 1967). The Euclidean space $\mathbb{E}^n$ is a universal metric space for all ultrametric spaces (X, d) with $|X| \leq n + 1$; the space of all finite functions $f(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}$ equipped with the metric $d(f, g) = \sup\{t : f(t) \neq g(t)\}$ is a universal metric space for all ultrametric spaces (Lemin–Lemin, 1996).

  The universality can be defined also for mappings, other than isometric embeddings, of metric spaces, say, a bi-Lipschitz embedding, etc. For example, any compact metric space is a continuous image of the **Cantor set** with the natural metric $|x - y|$ inherited from $\mathbb{R}$, and any complete separable metric space is a continuous image of the space of irrational numbers.

- **Constructive metric space**

  A **constructive metric space** is a pair (X, d), where X is a set of constructive objects (say, words over an alphabet), and d is an algorithm converting any pair of elements of X into a constructive real number $d(x, y)$ such that d is a metric on X.

- **Computable metric space**

  Let $\{x_n\}_{n \in \mathbb{N}}$ be a sequence of elements from a given *Polish* (i.e., complete separable) *metric space* (X, d) such that the set $\{x_n : n \in \mathbb{N}\}$ is dense in (X, d). Let $\mathcal{N}(m, n, k)$ be the *Cantor tuple function* of a triple $(n, m, k) \in \mathbb{N}^3$, and let $\{q_k\}_{k \in \mathbb{N}}$ be a fixed total standard numbering of the set $\mathbb{Q}$ of rational numbers.

[redacted]

The triple $(X, d, \{x_n\}_{n\in\mathbb{N}})$ is called an *effective* (or *semicomputable*) *metric space* ([Hemm02]) if the set $\{\mathcal{N}(n, m, k) : d(x_m, x_n) < q_k\}$ is *recursively enumerable*, i.e., there is an algorithm that enumerates the members of this set. If, moreover, the set $\{\mathcal{N}(n, m, k) : d(s_m, s_m) > q_k\}$ is recursively enumerable, then this triple is called (Lacombe, 1951) **computable metric space**, (or **recursive metric space**). In other words, the map $d \circ (q \times q) : \mathbb{N}^2 \to \mathbb{R}$ is a computable (double) sequence of real numbers, i.e., is recursive over $\mathbb{R}$.

Chapter 2
Topological Spaces

A *topological space* (X, τ) is a set X with a *topology* τ, i.e., a collection of subsets of X with the following properties:

1. $X \in \tau, \emptyset \in \tau$;
2. If $A, B \in \tau$, then $A \cap B \in \tau$;
3. For any collection $\{A_\alpha\}_\alpha$, if all $A_\alpha \in \tau$, then $\cup_\alpha A_\alpha \in \tau$.

The sets in τ are called *open sets*, and their complements are called *closed sets*. A *base* of the topology τ is a collection of open sets such that every open set is a union of sets in the base. The coarsest topology has two open sets, the empty set and X, and is called the *trivial topology* (or *indiscrete topology*). The finest topology contains all subsets as open sets, and is called the *discrete topology*.

In a metric space (X, d) define the *open ball* as the set $B(x, r) = \{y \in X : d(x, y) < r\}$, where $x \in X$ (the *center* of the ball), and $r \in \mathbb{R}, r > 0$ (the *radius* of the ball). A subset of X which is the union of (finitely or infinitely many) open balls, is called an *open set*. Equivalently, a subset U of X is called *open* if, given any point $x \in U$, there exists a real number $\epsilon > 0$ such that, for any point $y \in X$ with $d(x, y) < \epsilon$, $y \in U$.

Any metric space is a topological space, the topology (**metric topology**, *topology induced by the metric d*) being the set of all open sets. The metric topology is always T_4 (see below a list of topological spaces). A topological space which can arise in this way from a metric space, is called a **metrizable space**.

A *quasi-pseudo-metric topology* is a topology on X induced similarly by a quasi-semimetric d on X, using the set of open d-balls $B(x, r)$ as the base. In particular, *quasi-metric topology* and *pseudo-metric topology* are the terms used for the case of, respectively, quasi-metric and semimetric d. In general, those topologies are not T_0.

Given a topological space (X, τ), a *neighborhood* of a point $x \in X$ is a set containing an open set which in turn contains x. The *closure* of a subset of a topological space is the smallest closed set which contains it. An *open cover* of X is a collection $\mathcal{L}$ of open sets, the union of which is X; its *subcover* is a cover $\mathcal{K}$

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_2

[redacted]

such that every member of $\mathcal{K}$ is a member of $\mathcal{L}$; its *refinement* is a cover $\mathcal{K}$, where every member of $\mathcal{K}$ is a subset of some member of $\mathcal{L}$. A collection of subsets of X is called *locally finite* if every point of X has a neighborhood which meets only finitely many of these subsets.

A subset $A \subset X$ is called *dense* if $X = cl(A)$, i.e., it consists of A and its *limit points*; cf. **closed subset of metric space** in Chap. 1. The *density* of a topological space is the least cardinality of its dense subset. A *local base* of a point $x \in X$ is a collection $\mathcal{U}$ of neighborhoods of x such that every neighborhood of x contains some member of $\mathcal{U}$.

A function from one topological space to another is called *continuous* if the preimage of every open set is open. Roughly, given $x \in X$, all points close to x map to points close to $f(x)$. A function f from one metric space (X, d_X) to another metric space (Y, d_Y) is *continuous* at the point $c \in X$ if, for any positive real number ϵ, there exists a positive real number δ such that all $x \in X$ satisfying $d_X(x, c) < \delta$ will also satisfy $d_Y(f(x), f(c)) < \epsilon$; the function is continuous on an interval I if it is continuous at any point of I.

The following classes of topological spaces (up to T_4) include any metric space.

- **T_0-space**

  A **T_0-space** (or *Kolmogorov space*) is a topological space in which every two distinct points are *topologically distinguishable*, i.e., have different neighborhoods.
- **T_1-space**

  A **T_1-space** (or *accessible space*) is a topological space in which every two distinct points are *separated*, i.e., each does not belong to other's closure. T_1-spaces are always T_0.
- **T_2-space**

  A **T_2-space** (or **Hausdorff space**) is a topological space in which every two distinct points are *separated by neighborhoods*, i.e., have disjoint neighborhoods. T_2-spaces are always T_1.

  A space is T_2 if and only if it is both T_0 and *pre-regular*, i.e., any two *topologically distinguishable* points are separated by neighborhoods.
- **Regular space**

  A **regular space** is a topological space in which every neighborhood of a point contains a closed neighborhood of the same point. A **T_3-space** (or *Vietoris space*, *regular Hausdorff space*) is a topological space which is T_1 and regular.

  Bing, Nagata, Smirnov showed in 1950–1951 that a topological space is metrizable if and only if it is regular, T_0 and has a countably locally finite base.

  A **completely regular space** (or *Tychonoff space*) is a **Hausdorff space** (X, τ) in which any closed set A and any $x \notin A$ are *functionally separated*, i.e., there is a continuous function $f : X \to [0, 1]$ such that $f(A) = 0$ and $f(B) = 1$.
- **Normal space**

  A **normal space** is a topological space in which, for any two disjoint closed sets A and B, there exist two disjoint open sets U and V such that $A \subset U$, and

$B \subset V$. A T_4**-space** (or *Tietze space*, *normal Hausdorff space*) is a topological space which is T_1 and normal. Any metric space is a perfectly normal T_4-space.

A **completely** (or *hereditarily*) **normal space** is a topological space in which any two *separated* (i.e., disjoint from the other's closure) sets have disjoint neighborhoods. A T_5**-space** (or *completely normal Hausdorff space*) is a topological space which is completely normal and T_1. T_5-spaces are always T_4.

A **monotonically normal space** is a **completely normal space** in which any two *separated* subsets A and B are *strongly separated*, i.e., there exist open sets U and V with $A \subset U, B \subset V$ and $Cl(U) \cap Cl(V) = \emptyset$.

A **perfectly normal space** is a topological space (X, τ) in which any two disjoint closed subsets of X are *functionally separated*. A T_6**-space** (or *perfectly normal Hausdorff space*) is a topological space which is T_1 and perfectly normal. T_6-spaces are always T_5.

- **Moore space**

A **Moore space** is a **regular space** with a *development*.

A *development* is a sequence $\{\mathcal{U}_n\}_n$ of open covers such that, for every $x \in X$ and every open set A containing x, there exists n such that $St(x, \mathcal{U}_n) = \cup\{U \in \mathcal{U}_n : x \in U\} \subset A$, i.e., $\{St(x, \mathcal{U}_n)\}_n$ is a *neighborhood base* at x.

- **Polish space**

A **separable space** is a topological space which has a countable dense subset.

A **Polish space** is a separable space which can be equipped with a complete metric. A *Lusin space* is a topological space such that some weaker topology makes it into a Polish space; every Polish space is Lusin. A *Souslin space* is a continuous image of a Polish space; every Lusin space is Suslin.

- **Lindelöf space**

A **Lindelöf space** is a topological space in which every open cover has a countable subcover.

- **First-countable space**

A topological space is called **first-countable** if every point has a countable local base. Every metric space is first-countable.

- **Second-countable space**

A topological space is called **second-countable** if its topology has a countable base. Such space is **quasi-metrizable** and, if and only if it is a T_3-space, **metrizable**.

Second-countable spaces are **first-countable**, **separable** and **Lindelöf**. The properties **second-countable**, **separable** and **Lindelöf** are equivalent for metric spaces.

The Euclidean space $\mathbb{E}^n$ with its usual topology is second-countable.

- **Baire space**

A **Baire space** is a topological space in which every intersection of countably many dense open sets is dense. Every complete metric space is a Baire space. Every locally compact T_2-space (hence, every n-**manifold**) is a Baire space.

- **Alexandrov space**

An **Alexandrov space** is a topological space in which every intersection of arbitrarily many open sets is open.

[redacted]

A topological space is called a **P-space** if every *G_δ-set* (i.e., the intersection of countably many open sets) is open.

A topological space (X, τ) is called a **Q-space** if every subset $A \subset X$ is a G_δ-set.

- **Connected space**

  A topological space (X, τ) is called **connected** if it is not the union of a pair of disjoint nonempty open sets. In this case the set X is called a *connected set*.

  A connected topological space (X, τ) is called *unicoherent* if the intersection $A \cap B$ is connected for any closed connected sets A, B with $A \cup B = X$.

  A topological space (X, τ) is called **locally connected** if every point $x \in X$ has a local base consisting of connected sets.

  A topological space (X, τ) is called **path-connected** (or 0-*connected*) if for every points $x, y \in X$ there is a path γ from x to y, i.e., a continuous function $\gamma : [0, 1] \to X$ with $\gamma(x) = 0, \gamma(y) = 1$.

  A topological space (X, τ) is called **simply connected** (or 1-*connected*) if it consists of one piece, and has no circle-shaped "holes" or "handles" or, equivalently, if every continuous curve of X is *contractible*, i.e., can be reduced to one of its points by a *continuous deformation*.

  A topological space (X, τ) is called **hyperconnected** (or *irreducible*) if X cannot be written as the union of two proper closed sets.

- **Sober space**

  A topological space (X, τ) is called **sober** if every **hyperconnected** closed subset of X is the closure of exactly one point of X. Any sober space is a T_0-space.

  Any T_2-space is a sober T_1-space but some sober T_1-spaces are not T_2.

- **Paracompact space**

  A topological space is called **paracompact** if every open cover of it has an open locally finite refinement. Every **metrizable** space is paracompact.

- **Totally bounded space**

  A topological space (X, τ) is called **totally bounded** (or *pre-compact*) if it can be covered by finitely many subsets of any fixed cardinality.

  A metric space (X, d) is a **totally bounded metric space** if, for every real number $r > 0$, there exist finitely many open balls of radius r, whose union is equal to X.

- **Compact space**

  A topological space (X, τ) is called **compact** if every open cover of X has a finite subcover.

  Compact spaces are always **Lindelöf**, **totally bounded**, and **paracompact**. A metric space is compact if and only if it is **complete** and **totally bounded**. A subset of a Euclidean space $\mathbb{E}^n$ is compact if and only if it is closed and bounded.

  There exist a number of topological properties which are equivalent to compactness in metric spaces, but are nonequivalent in general topological spaces. Thus, a metric space is compact if and only if it is a *sequentially compact space* (every sequence has a convergent subsequence), or a *countably compact space* (every countable open cover has a finite subcover), or a *pseudo-compact* *space* (every real-valued continuous function on the space is bounded), or a *weakly countably compact space* (i.e., every infinite subset has an accumulation point).

Sometimes, a compact **connected** T_2**-space** is called *continuum*; cf. **continuum** in Chap. 1.

- **Locally compact space**

  A topological space is called **locally compact** if every point has a local base consisting of compact neighborhoods. The Euclidean spaces $\mathbb{E}^n$ and the spaces $\mathbb{Q}_p$ of *p-adic numbers* are locally compact.

  A topological space (X, τ) is called a *k-space* if, for any compact set $Y \subset X$ and $A \subset X$, the set A is closed whenever $A \cap Y$ is closed. The k-spaces are precisely quotient images of locally compact spaces.

- **Locally convex space**

  A **topological vector space** is a real (complex) vector space V which is a T_2**-space** with continuous vector addition and scalar multiplication. It is a **uniform space** (cf. Chap. 3).

  A **locally convex space** is a topological vector space whose topology has a base, where each member is a *convex balanced absorbent* set. A subset A of V is called *convex* if, for all $x, y \in A$ and all $t \in [0, 1]$, the point $tx + (1-t)y \in A$, i.e., every point on the *line segment* connecting x and y belongs to A. A subset A is *balanced* if it contains the line segment between x and $-x$ for every $x \in A$; A is *absorbent* if, for every $x \in V$, there exist $t > 0$ such that $tx \in A$.

  The locally convex spaces are precisely vector spaces with topology induced by a family $\{||.||_\alpha\}$ of seminorms such that $x = 0$ if $||x||_\alpha = 0$ for every α.

  Any metric space $(V, ||x - y||)$ on a real (complex) vector space V with a **norm metric** $||x - y||$ is a locally convex space; each point of V has a local base consisting of convex sets. Every L_p with $0 < p < 1$ is an example of a vector space which is not locally convex.

- n**-manifold**

  Broadly, a *manifold* is a topological space locally homeomorphic to a **topological vector space** over the reals.

  But usually, a **topological manifold** is a **second-countable** T_2**-space** that is locally homeomorphic to Euclidean space. An n**-manifold** is a topological manifold such that every point has a neighborhood homeomorphic to $\mathbb{E}^n$.

- **Fréchet space**

  A **Fréchet space** is a **locally convex space** (V, τ) which is complete as a **uniform space** and whose topology is defined using a countable set of seminorms $||.||_1, \dots, ||.||_n, \dots$, i.e., a subset $U \subset V$ is *open in* (V, τ) if, for every $u \in U$, there exist $\epsilon > 0$ and $N \geq 1$ with $\{v \in V : ||u - v||_i < \epsilon \text{ if } i \leq N\} \subset U$.

  A Fréchet space is precisely a locally convex **F-space** (cf. Chap. 5). Its topology can be induced by a **translation invariant metric** (Chap. 5) and it is a complete and **metrizable space** with respect to this topology. But this topology may be induced by many such metrics. Every **Banach space** is a Fréchet space.

[redacted]

- **Countably-normed space**

  A **countably-normed space** is a **locally convex space** (V, τ) whose topology is defined using a countable set of *compatible norms* $||.||_1, \dots, ||.||_n, \dots$. It means that, if a sequence $\{x_n\}_n$ of elements of V that is fundamental in the norms $||.||_i$ and $||.||_j$ converges to zero in one of these norms, then it also converges in the other. A countably-normed space is a **metrizable space**, and its metric can be defined by

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \cdot \frac{||x-y||_n}{1+||x-y||_n}.$$

- **Metrizable space**

  A topological space (T, τ) is called **metrizable** if it is homeomorphic to a metric space, i.e., X admits a metric d such that the set of open d-balls $\{B(x, r) : r > 0\}$ forms a neighborhood base at each point $x \in X$. If, moreover, (X, d) is a complete metric space for one of such metrics d, then (X, d) is a *completely metrizable* (or *topologically complete*) space.

  Metrizable spaces are always **paracompact** T_2**-spaces** (hence, **normal** and **completely regular**), and **first-countable**.

  A topological space is called **locally metrizable** if every point in it has a metrizable neighborhood.

  A topological space (X, τ) is called **submetrizable** if there exists a metrizable topology τ' on X which is coarser than τ.

  A topological space (X, τ) is called **proto-metrizable** if it is paracompact and has an *orthobase*, i.e., a base $\mathcal{B}$ such that, for $\mathcal{B}' \subset \mathcal{B}$, either $\cap\mathcal{B}'$ is open, or $\mathcal{B}'$ is a local base at the unique point in $\cap\mathcal{B}'$. It is not related to the **protometric** in Chap. 1.

  Some examples of other direct generalizations of metrizable spaces follow.

  A **sequential space** is a quotient image of a metrizable space.

  Morita's M**-space** is a topological space (X, τ) from which there exists a continuous map f onto a metrizable topological space (Y, τ') such that f is closed and $f^{-1}(y)$ is countably compact for each $y \in Y$.

  Ceder's M_1**-space** is a topological space (X, τ) having a σ-closure-preserving base (metrizable spaces have σ-locally finite bases).

  Okuyama's σ**-space** is a topological space (X, τ) having a σ-locally finite *net*, i.e., a collection $\mathcal{U}$ of subsets of X such that, given of a point $x \in U$ with U open, there exists $U' \in \mathcal{U}$ with $x \in U' \subset U$ (a base is a net consisting of open sets). Every compact subset of a σ-space is metrizable.

  Michael's **cosmic space** is a topological space (X, τ) having a countable net (equivalently, a Lindelöf σ-space). It is exactly a continuous image of a separable metric space. A T_2**-space** is called **analytic** if it is a continuous image of a complete separable metric space; it is called a **Lusin space** if, moreover, the image is one-to-one.

- **Quasi-metrizable space**

  A topological space (X, τ) is called a **quasi-metrizable space** if X admits a quasi-metric d such that the set of open d-balls $\{B(x,r) : r > 0\}$ forms a neighborhood base at each point $x \in X$.

  A more general γ**-space** is a topological space admitting a γ**-metric** d (i.e., a function $d : X \times X \to \mathbb{R}_{\geq 0}$ with $d(x, z_n) \to 0$ whenever $d(x, y_n) \to 0$ and $d(y_n, z_n) \to 0$) such that the set of open *forward* d-balls $\{B(x,r) : r > 0\}$ forms a neighborhood base at each point $x \in X$.

  The *Sorgenfrey line* is the topological space $(\mathbb{R}, \tau)$ defined by the base $\{[a,b) : a, b \in \mathbb{R}, a < b\}$. It is not metrizable but it is a first-countable separable and paracompact T_5**-space**; neither it is second-countable, nor locally compact or locally connected. However, the Sorgenfrey line is quasi-metrizable by the **Sorgenfrey quasi-metric** (cf. Chap. 12) defined as $y - x$ if $y \geq x$, and 1, otherwise.
- **Symmetrizable space**

  A topological space (X, τ) is called **symmetrizable** (and τ is called the **distance topology**) if there is a **symmetric** d on X (i.e., a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ with $d(x,y) = 0$ implying $x = y$) such that a subset $U \subset X$ is open if and only if, for each $x \in U$, there exists $\epsilon > 0$ with $B(x, \epsilon) = \{y \in X : d(x,y) < \epsilon\} \subset U$.

  In other words, a subset $H \subset X$ is closed if and only if $d(x, H) = \inf_y \{d(x,y) : y \in H\} > 0$ for each $x \in X \backslash U$. A symmetrizable space is **metrizable** if and only if it is a Morita's M**-space**.

  In Topology, the term **semimetrizable space** refers to a topological space (X, τ) admitting a symmetric d such that, for each $x \in X$, the family $\{B(x, \epsilon) : \epsilon > 0\}$ of balls forms a (not necessarily open) neighborhood base at x. In other words, a point x is in the closure of a set H if and only if $d(x, H) = 0$.

  A topological space is semimetrizable if and only if it is symmetrizable and **first-countable**. Also, a symmetrizable space is semimetrizable if and only if it is a *Fréchet–Urysohn space* (or *E-space*), i.e., for any subset A and for any point x of its closure, there is a sequence in A converging to x.
- **Hyperspace**

  A **hyperspace** of a topological space (X, τ) is a topological space on the set $CL(X)$ of all nonempty closed (or, moreover, compact) subsets of X. The topology of a hyperspace of X is called a *hypertopology*. Examples of such a *hit-and-miss topology* are the *Vietoris topology*, and the *Fell topology*. Examples of such a *weak hyperspace topology* are the *Hausdorff metric topology*, and the *Wijsman topology*.
- **Discrete topological space**

  A topological space (X, τ) is **discrete** if τ is the *discrete topology* (the finest topology on X), i.e., containing all subsets of X as open sets. Equivalently, it does not contain any *limit point*, i.e., it consists only of *isolated points*.
- **Indiscrete topological space**

  A topological space (X, τ) is **indiscrete** if τ is the *indiscrete topology* (the coarsest topology on X), i.e., having only two open sets, $\emptyset$ and X.

[redacted]

It can be considered as the semimetric space (X, d) with the **indiscrete semimetric**: $d(x, y) = 0$ for any $x, y \in X$.

- **Extended topology**

  Consider a set X and a map $cl : P(X) \to P(X)$, where $P(X)$ is the set of all subsets of X. The set $cl(A)$ (for $A \subset X$), its dual set $int(A) = X \backslash cl(X \backslash A)$ and the map $N : X \to P(X)$ with $N(x) = \{A \subset X : x \in int(A)\}$ are called the *closure*, *interior* and *neighborhood* map, respectively.

  So, $x \in cl(A)$ is equivalent to $X \backslash A \in P(X) \backslash N(x)$. A subset $A \subset X$ is *closed* if $A = cl(A)$ and *open* if $A = int(A)$. Consider the following possible properties of cl; they are meant to hold for all $A, B \in P(X)$.

  1. $cl(\emptyset) = \emptyset$;
  2. $A \subseteq B$ implies $cl(A) \subseteq cl(B)$ (*isotony*);
  3. $A \subseteq cl(A)$(*enlarging*);
  4. $cl(A \cup B) = cl(A) \cup cl(B)$ (*linearity*, and, in fact, 4 implies 2);
  5. $cl(cl(A)) = cl(A)$ (*idempotency*).

  The pair (X, cl) satisfying 1 is called an **extended topology** if 2 holds, a **Brissaud space** (Brissaud, 1974) if 3 holds, a **neighborhood space** (Hammer, 1964) if 2 and 3 hold, a **Smyth space** (Smyth, 1995) if 4 holds, a **pre-topology** (Čech, 1966) if 3 and 4 hold, and a **closure space** (Soltan, 1984) if 2, 3 and 5 hold.

  (X, cl) is the usual topology, in closure terms, if 1, 3, 4 and 5 hold.

Chapter 3
Generalizations of Metric Spaces

Some immediate generalizations of the notion of metric, for example, **quasi-metric**, **near-metric**, **extended metric**, were defined in Chap. 1. Here we give some generalizations in the direction of Topology, Probability, Algebra, etc.

3.1 *m*-Tuple Generalizations of Metrics

In the definition of a metric, for every *two points* there is a *unique associated number*. Here we group some generalizations of metrics in which *several points* or *several numbers* are considered instead.

- ***m*-hemimetric**

  Let X be a nonempty set. A function $d : X^{m+1} \to \mathbb{R}_{\geq 0}$ is called a *m*-**hemimetric** (Deza–Rosenberg, 2000) if it have the following properties:

  1. d is *totally symmetric*, i.e., satisfies $d(x_1, \dots, x_{m+1}) = d(x_{\pi(1)}, \dots, x_{\pi(m+1)})$ for all $x_1, \dots, x_{m+1} \in X$ and for any permutation π of $\{1, \dots, m+1\}$;
  2. $d(x_1, \dots, x_{m+1}) = 0$ if $x_1, \dots, x_{m+1}$ are not pairwise distinct;
  3. for all $x_1, \dots, x_{m+2} \in X$, d satisfies the ***m*-simplex inequality**

  $$d(x_1, \dots, x_{m+1}) \leq \sum_{i=1}^{m+1} d(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{m+2}).$$

  Cf. unrelated **hemimetric** (i.e., a quasi-semimetric) in Chap. 1.

  If in above 3. $d(x_1, \dots, x_{m+1})$ is replaced by $sd(x_1, \dots, x_{m+1})$ for some $s, 0 < s \leq 1$, then d is called (m, s)**-super-metric** ([DeDu03]). $(m, 1)$- and $(1, s)$-super-metrics are exactly m-hemimetric and $\frac{1}{s}$-*near-semimetric*; cf. **near-metric** in Chap. 1.

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_3

[redacted]

If above 3. is dropped, d is called m**-dissimilarity**. 1-dissimilarity and 1-hemimetric are exactly a distance and a semimetric.

- **2-metric**

  An m**-hemimetric** with $m = 2$ satisfies **2-simplex** (or *tetrahedron*) **inequality**

  $$d(x_1, x_2, x_3) \leq d(x_4, x_2, x_3) + d(x_1, x_4, x_3) + d(x_1, x_2, x_4).$$

  A **2-metric** (Gähler, 1963 and 1966) is a **2-hemimetric** d in which, for any distinct points x_1, x_2, there is a point x_3 with $d(x_1, x_2, x_3) > 0$. The area of the triangle spanned by x_1, x_2, x_3 on $\mathbb{R}^2$ or $\mathbb{S}^2$ is a 2-metric.

  A *D-space* (Dhage, 1992) is an **2-hemimetric space** (X, d) in which the condition "$d(x_1, x_2, x_3) = 0$ if two of x_1, x_2, x_3 are equal" is replaced by "$d(x_1, x_2, x_3) = 0$ if and only if $x_1 = x_2 = x_3$." Mustafa and Sims, 2003, showed that D-spaces are not suitable for topological constructions. In 2006, they defined instead a function, let us call it $MS -$ **2-metric**, $D : X^3 \to \mathbb{R}_{\geq 0}$ which satisfies

  1. $D(x_1, x_2, x_3) = 0$ if $x_1 = x_2 = x_3$;
  2. $D(x_1, x_1, x_2) > 0$ whenever $x_1 \neq x_2$;
  3. $D(x_1, x_2, x_3) \geq D(x_1, x_1, x_2)$ whenever $x_3 \neq x_2$;
  4. D is a *totally symmetric* function of its three variables, and
  5. $D(x_1, x_2, x_3) \leq D(x_1, x_4, x_4) + D(x_4, x_2, x_3)$ for all $x_1, x_2, x_3, x_4 \in X$.

  The perimeter of the triangle spanned by x_1, x_2, x_3 on $\mathbb{R}^2$ is a $MS -$ 2-metric. If d is a metric, then $\frac{1}{2}(d(x_1, x_2) + d(x_2, x_3) + d(x_1, x_3))$ and $\max(d(x_1, x_2), d(x_2, x_3), d(x_1, x_3))$ are $MS -$ 2-metrics. If D is a $MS -$ 2-metric, then $D(x_1, x_2, x_2) + D(x_1, x_1, x_2)$ is a metric. If (X, D) is a $MS -$ 2-metric space, the open *D-ball with center* x_0 *and radius* r is $B_D(x_0, r) = \{x_1 \in X : D(x_0, x_1, x_1) < r\}$.

- **Multidistance**

  Given a set X, a function $D : \cup_{m>1} X^m \to \mathbb{R}_{\geq 0}$ is called a **multidistance** (Martin–Major, 2009) if, for all m and all $x_1, \dots, x_m, y \in X$, it satisfies:

  1. $D(x_1, \dots, x_m) = 0$ if $x_1 = \cdots = x_m$;
  2. $D(x_1, \dots, x_m) = D(x_{\pi(1)}, \dots, x_{\pi(m)})$ for any permutation π of $\{1, \dots, m\}$;
  3. $D(x_1, \dots, x_m) \leq \sum_{i=1}^{m} D(x_i, y)$.

  Clearly, the restriction of a multidistance on X^2 is a semimetric.

  A multidistance D is called *regular*, if all $D(x_1, \dots, x_m) \leq D(x_1, \dots, x_m, y)$ hold, and *stable*, if all $D(x_1, \dots, x_m) = D(x_1, \dots, x_m, x_i)$ hold. Given a metric space (X, d), the *Fermat multidistance* is $\min_{x \in X} \sum_{i=1}^{m} d(x_i, x)$; it is regular, but not stable.

  The regular multidistances on X form a convex cone.

- **Multimetric**

  In Mao, 2006, a **multimetric space** is the union of some metric spaces $(X_i, d_i), i \in J$. In the case $X_i = X, i \in J$, the **multimetric** is defined as the sequence-valued map $d(x, y) = (d_i), i \in J$, from $X \times X$ to $R^{|J|}_{\geq 0}$.

Cf. **bimetric theory of gravity** in Chap. 24 and (in the item **meter-related terms**) *multimetric crystallography* in Chap. 27.

Also, Jörnsten, 2007, consider *clustering* (Chap. 17) under several distance metrics simultaneously. In Rintanen, 2004, a *linear multimetric* is defined as $d = w_1 d_1 + \cdots + w_m d_m$, where d_i are metrics and $w_i \in [0, 1]$ are weights.

- **Diversity**

  Given a set X, a function f from its finite subsets to $\mathbb{R}_{\geq 0}$ is called (Bryant–Tupper, 2012) *diversity on* X if $f(A) = 0$ for all $A \subset X$ with $|A| \leq 1$ and

  $$f(A \cup B) + f(B \cup C) \geq f(A \cup C) \text{ for all } A, B, C \subset X \text{ with } B \neq \emptyset.$$

  The **induced diversity metric** $d(x, y)$ is $f(\{x, y\})$.

  For any diversity $f(A)$ with induced metric space (X, d), it holds $f_{diam}(A) \leq f(A) \leq f_S(A) \leq (|A| - 1)f_{diam}(A)$, where the **diameter diversity** $f_{diam}(A)$ is $\max_{x,y \in A} d(x, y) = diam(A)$ and the **Steiner diversity** $f_S(A)$ is the minimum weight of a Steiner tree connecting elements of A.

  l_1**-diversity** is defined by $f_{m1}(A) = \max |a_i - b_i| : a, b \in A$ for all finite $A \subset \mathbb{R}^m$.

  Any diversity is a **Vitanyi multiset metric**, restricted to subsets. But much of Bryant–Tupper's theory of diversities does not extend on multisets.

- **Vitanyi multiset metric**

  Given two multisets m and m', define $n = mm'$ if n is the multiset consisting of the elements of the multisets m and m', that is, if x occurs once in m and once in m', then it occurs twice in n. A function d on the set of nonempty finite multisets is (Vitanyi, 2011) a **multiset metric** if

  1. $d(m) = 0$ if all elements of m are equal and $d(m) > 0$ otherwise.
  2. $d(X)$ is invariant under all permutations of m.
  3. $d(mm') \leq d(mm'') + d(m''m')$ (*multiset triangle inequality*).

  The usual metric between two elements results if the multiset m has two elements in 1. and 2. and the multisets m, m', m'' have one element each in 3.

  An example is the set of all nonempty finite multisets m of integers with $d(m) = \max\{x : x \in m\} - \min\{x : x \in m\}$. Cohen–Vitanyi, 2012, defined another multiset metric, generalising **normalised web distance** (Chap. 22).

3.2 Indefinite Metrics

- **Indefinite metric**

  An **indefinite metric** (or *G-metric*) on a real (complex) vector space V is a *bilinear* (in the complex case, *sesquilinear*) *form* G on V, i.e., a function $G : V \times V \to \mathbb{R}$ ($\mathbb{C}$), such that, for any $x, y, z \in V$ and for any scalars α, β, we have the following properties: $G(\alpha x + \beta y, z) = \alpha G(x, z) + \beta G(y, z)$, and $G(x, \alpha y + \beta z) = \overline{\alpha} G(x, y) + \overline{\beta} G(x, z)$, where $\overline{\alpha} = \overline{a + bi} = a - bi$ denotes the *complex conjugation*.

[redacted]

If a positive-definite form G is symmetric, then it is an *inner product* on V, and one can use it to canonically introduce a *norm* and the corresponding **norm metric** on V. In the case of a general form G, there is neither a norm, nor a metric canonically related to G, and the term **indefinite metric** only recalls the close relation of such forms with certain metrics in vector spaces (cf. Chaps. 7 and 26).

The pair (V, G) is called a *space with an indefinite metric*. A finite-dimensional space with an indefinite metric is called a *bilinear metric space*. A **Hilbert space** H, endowed with a continuous G-metric, is called a *Hilbert space with an indefinite metric*. The most important example of such space is a *J-space*; cf. **J-metric**.

A subspace L in a space (V, G) with an indefinite metric is called a *positive subspace*, *negative subspace*, or *neutral subspace*, depending on whether $G(x,x) > 0$, $G(x,x) < 0$, or $G(x,x) = 0$ for all $x \in L$.

- **Hermitian G-metric**

  A **Hermitian G-metric** is an **indefinite metric** G^H on a complex vector space V such that, for all $x, y \in V$, we have the equality

$$G^H(x,y) = \overline{G^H(y,x)},$$

where $\overline{\alpha} = \overline{a + bi} = a - bi$ denotes the *complex conjugation*.

- **Regular G-metric**

  A **regular G-metric** is a continuous **indefinite metric** G on a **Hilbert space** H over $\mathbb{C}$, generated by an invertible *Hermitian operator* T by the formula

$$G(x,y) = \langle T(x), y \rangle,$$

where $\langle , \rangle$ is the *inner product* on H.

A *Hermitian operator* on a Hilbert space H is a *linear operator* T on H defined on a *domain* $D(T)$ of H such that $\langle T(x), y \rangle = \langle x, T(y) \rangle$ for any $x, y \in D(T)$. A bounded Hermitian operator is either defined on the whole of H, or can be so extended by continuity, and then $T = T^*$. On a finite-dimensional space a Hermitian operator can be described by a *Hermitian matrix* $((a_{ij})) = ((\overline{a}_{ji}))$.

- **J-metric**

  A **J-metric** is a continuous **indefinite metric** G on a **Hilbert space** H over $\mathbb{C}$ defined by a certain *Hermitian involution* J on H by the formula

$$G(x,y) = \langle J(x), y \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the *inner product* on H.

An *involution* is a mapping H onto H whose square is the *identity mapping*. The involution J may be represented as $J = P_+ - P_-$, where P_+ and P_- are orthogonal projections in H, and $P_+ + P_- = H$. The rank *of indefiniteness* of the J-metric is defined as $\min\{\dim P_+, \dim P_-\}$.

The space (H, G) is called a *J-space*. A J-space with finite rank of indefiniteness is called a *Pontryagin space*.

3.3 Topological Generalizations

- **Metametric space**

  A **metametric space** (Väisälä, 2003) is a pair (X, d), where X is a set, and d is a nonnegative symmetric function $d : X \times X \to \mathbb{R}$ such that $d(x, y) = 0$ implies $x = y$ and triangle inequality $d(x, y) \leq d(x, z) + d(z, y)$ holds for all $x, y, z \in X$.

  A metametric space is metrizable: the metametric d defines the same topology as the metric d' defined by $d'(x, x) = 0$ and $d'(x, y) = d(x, y)$ if $x \neq y$. A metametric d induces a Hausdorff topology with the usual definition of a *ball* $B(x_0, r) = \{x \in X : d(x_0, x) < r\}$. Any **partial metric** (Chap. 1) is a metametric.

- **Resemblance**

  Let X be a set. A function $d : X \times X \to \mathbb{R}$ is called (Batagelj-Bren, 1993) a **resemblance** on X if d is *symmetric* and if, for all $x, y \in X$, either $d(x, x) \leq d(x, y)$ (in which case d is called a **forward resemblance**), or $d(x, x) \geq d(x, y)$ (in which case d is called a **backward resemblance**).

  Every resemblance d induces a *strict partial order* $\prec$ on the set of all unordered pairs of elements of X by defining $\{x, y\} \prec \{u, v\}$ if and only if $d(x, y) < d(u, v)$.

- **w-distance**

  Given a metric space (X, d), a **w-distance** on X (Kada–Suzuki–Takahashi, 1996) is a nonnegative function $p : X \times X \to \mathbb{R}$ which satisfies the following conditions:

  1. $p(x, z) \leq p(x, y) + p(y, z)$ for all $x, y, z \in X$;
  2. for any $x \in X$, the function $p(x, .) : X \to \mathbb{R}$ is *lower semicontinuous*, i.e., if a sequence $\{y_n\}_n$ in X converges to $y \in X$, then $p(x, y) \leq \underline{\lim}_{n \to \infty} p(x, y_n)$;
  3. for any $\epsilon > 0$, there exists $\delta > 0$ such that $p(z, x) \leq \delta$ and $p(z, y) \leq \delta$ imply $d(x, y) \leq \epsilon$, for each $x, y, z \in X$.

- **τ-distance space**

  A **τ-distance space** is a pair (X, f), where X is a topological space and f is an Aamri-Moutawakil's *τ-distance* on X, i.e., a nonnegative function $f : X \times X \to \mathbb{R}$ such that, for any $x \in X$ and any neighborhood U of x, there exists $\epsilon > 0$ with $\{y \in X : f(x, y) < \epsilon\} \subset U$.

  Any distance space (X, d) is a τ-distance space for the topology τ_f defined as follows: $A \in \tau_f$ if, for any $x \in X$, there exists $\epsilon > 0$ with $\{y \in X : f(x, y) < \epsilon\} \subset A$. However, there exist nonmetrizable τ-distance spaces. A τ-distance $f(x, y)$ need be neither symmetric, nor vanishing for $x = y$; for example, $e^{|x-y|}$ is a τ-distance on $X = \mathbb{R}$ with usual topology.

[redacted]

- **Proximity space**

  A **proximity space** (Efremovich, 1936) is a set X with a binary relation δ on the *power set* $P(X)$ of all of its subsets which satisfies the following conditions:

  1. $A\delta B$ if and only if $B\delta A$ (*symmetry*);
  2. $A\delta(B \cup C)$ if and only if $A\delta B$ or $A\delta C$ (*additivity*);
  3. $A\delta A$ if and only if $A \neq \emptyset$ (*reflexivity*).

  The relation δ defines a **proximity** (or *proximity structure*) on X. If $A\delta B$ fails, the sets A and B are called *remote sets*.

  Every metric space (X, d) is a proximity space: define $A\delta B$ if and only if $d(A, B) = \inf_{x \in A, y \in B} d(x, y) = 0$.

  Every proximity on X induces a (**completely regular**) topology on X by defining the *closure operator* $cl : P(X) \rightarrow P(X)$ on the set of all subsets of X as $cl(A) = \{x \in X : \{x\}\delta A\}$.

- **Uniform space**

  A **uniform space** is a topological space (with additional structure) providing a generalization of metric space, based on **set-set distance**.

  A **uniform space** (Weil, 1937) is a set X with an *uniformity* (or *uniform structure*) $\mathcal{U}$, i.e., a nonempty collection of subsets of $X \times X$, called *entourages*, with the following properties:

  1. Every subset of $X \times X$ which contains a set of $\mathcal{U}$ belongs to $\mathcal{U}$;
  2. Every finite intersection of sets of $\mathcal{U}$ belongs to $\mathcal{U}$;
  3. Every set $V \in \mathcal{U}$ contains the *diagonal*, i.e., the set $\{(x, x) : x \in X\} \subset X \times X$;
  4. If V belongs to $\mathcal{U}$, then the set $\{(y, x) : (x, y) \in V\}$ belongs to $\mathcal{U}$;
  5. If V belongs to $\mathcal{U}$, then there exists $V' \in \mathcal{U}$ such that $(x, z) \in V$ whenever $(x, y), (y, z) \in V'$.

  Every metric space (X, d) is a uniform space. An entourage in (X, d) is a subset of $X \times X$ which contains the set $V_\epsilon = \{(x, y) \in X \times X : d(x, y) < \epsilon\}$ for some positive real number ϵ. Other basic example of uniform space are *topological groups*.

  Every uniform space $(X, \mathcal{U})$ generates a topology consisting of all sets $A \subset X$ such that, for any $x \in A$, there is a set $V \in \mathcal{U}$ with $\{y : (x, y) \in V\} \subset A$.

  Every uniformity induces a **proximity** σ where $A\sigma B$ if and only if $A \times B$ has nonempty intersection with any entourage.

  A topological space admits a uniform structure inducing its topology if only if the topology is **completely regular** (Chap. 2) and, also, if only if it is a *gauge space*, i.e., the topology is defined by a *$\geq$-filter* of semimetrics.

- **Nearness space**

  A **nearness space** (Herrich, 1974) is a set X with a *nearness structure*, i.e., a nonempty collection $\mathcal{U}$ of families of subsets of X, called *near families*, with the following properties:

  1. Each family refining a near family is near;
  2. Every family with nonempty intersection is near;

3. V is near if $\{cl(A) : A \in V\}$ is near, where $cl(A)$ is $\{x \in X : \{\{x\}, A\} \in \mathcal{U}\}$;
4. $\emptyset$ is near, while the set of all subsets of X is not;
5. If $\{A \cup B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$ is near family, then so is $\mathcal{F}_1$ or $\mathcal{F}_2$.

The **uniform spaces** are precisely **paracompact** nearness spaces.

- **Approach space**

An **approach space** is a topological space providing a generalization of metric space, based on **point-set distance**.

An **approach space** (Lowen, 1989) is a pair (X, D), where X is a set and D is a **point-set distance**, i.e., a function $X \times P(X) \to [0, \infty]$ (where $P(X)$ is the set of all subsets of X) satisfying, for all $x \in X$ and all $A, B \in P(X)$, the following conditions:

1. $D(x, \{x\}) = 0$;
2. $D(x, \{\emptyset\}) = \infty$;
3. $D(x, A \cup B) = \min\{D(x, A), D(x, B)\}$;
4. $D(x, A) \leq D(x, A^{\epsilon}) + \epsilon$ for any $\epsilon \in [0, \infty]$, where $A^{\epsilon} = \{x : D(x, A) \leq \epsilon\}$ is the "ϵ-ball" with center x.

Every metric space (X, d) (moreover, any extended quasi-semimetric space) is an approach space with $D(x, A)$ being the usual point-set distance $\min_{y \in A} d(x, y)$.

Given a **locally compact separable** metric space (X, d) and the family $\mathcal{F}$ of its nonempty closed subsets, the **Baddeley–Molchanov distance function** gives a tool for another generalization. It is a function $D : X \times \mathcal{F} \to \mathbb{R}$ which is lower semicontinuous with respect to its first argument, measurable with respect to the second, and satisfies the following two conditions: $F = \{x \in X : D(x, F) \leq 0\}$ for $F \in \mathcal{F}$, and $D(x, F_1) \geq D(x, F_2)$ for $x \in X$, whenever $F_1, F_2 \in \mathcal{F}$ and $F_1 \subset F_2$.

The additional conditions $D(x, \{y\}) = D(y, \{x\})$, and $D(x, F) \leq D(x, \{y\}) + D(y, F)$ for all $x, y \in X$ and every $F \in \mathcal{F}$, provide analogs of symmetry and the triangle inequality. The case $D(x, F) = d(x, F)$ corresponds to the usual point-set distance for the metric space (X, d); the case $D(x, F) = d(x, F)$ for $x \in X \backslash F$ and $D(x, F) = -d(x, X \backslash F)$ for $x \in X$ corresponds to the **signed distance function** in Chap. 1.

- **Metric bornology**

Given a topological space X, a *bornology* of X is any family $\mathcal{A}$ of proper subsets A of X such that the following conditions hold:

1. $\cup_{A \in \mathcal{A}} A = X$;
2. $\mathcal{A}$ is an *ideal*, i.e., contains all subsets and finite unions of its members.

   The family $\mathcal{A}$ is a **metric bornology** ([Beer99]) if, moreover

3. $\mathcal{A}$ contains a countable base;
4. For any $A \in \mathcal{A}$ there exists $A' \in \mathcal{A}$ such that the closure of A coincides with the interior of A'.

The metric bornology is called *trivial* if $\mathcal{A}$ is the set $P(X)$ of all subsets of X; such a metric bornology corresponds to the family of bounded sets of some bounded

[redacted]

metric. For any noncompact **metrizable** topological space X, there exists an unbounded metric compatible with this topology. A nontrivial metric bornology on such a space X corresponds to the family of bounded subsets with respect to some such unbounded metric. A noncompact metrizable topological space X admits uncountably many nontrivial metric bornologies.

3.4 Beyond Numbers

- **Metric 1-space**

  A *category* Ψ consists (Eilenberg and MacLane, 1945) of a set $Ob(\Psi)$ of *objects*, a set $Mor(\Psi)$ of *morphisms* (or *arrows*)) and a set-valued map associating a set $\Psi(x, y)$ of arrows to each ordered pair of objects x, y, so that each arrow belongs to only one set $\Psi(x, y)$. An element of $\Psi(x, y)$ is also denoted by $f : x \rightarrow y$.

  Moreover, the composition $f \cdot g \in \Psi(x, z)$ of two arrows $f : x \rightarrow y, g : y \rightarrow z$ is defined, and it is associative. Finally, each set $\Psi(x, x)$ contains an *identity arrow* id_x such that $f \cdot id_x = f$ and $id_x \cdot g = g$ for any arrows $f : y \rightarrow x$ and $g : x \rightarrow z$. Cf. **category of metric spaces** in Chap. 1.

  Weiss defined in [Weis12] a **metric 1-space** as a category Ψ together with a weight-function $w : \Psi(x, y) \rightarrow \mathbb{R}_{\geq 0} \cup \{\infty\}$ on arrows, which satisfies

  1. $w(id_x) = 0$ holds for each object $x \in Ob(\Psi)$ (*reflexivity*).
  2. $|w(g) - w(f)| \leq w(g \cdot f) \leq w(g) + w(f)$ holds for any objects x, y, z and arrows $f : x \rightarrow y, g : y \rightarrow z$ (*full triangle inequality*).

  Any set X produces an *indiscrete category* I_X, in which $Ob(I_X) = X$ and $|I_X(x, y)| = 1$ for all $x, y \in X$. Any metric space (X, d) produces a metric 1-space on I_X by defining $w(f) = d(x, y)$, and it is unique metric 1-space on I_X. But, in general, the function w on arrows can be seen as a multivalued function on $Ob \times Ob$.

  [Weis12] also outlined a **metric** m**-space** as a kind of an m**-hemimetric** on an m*-category* consisting of i-dimensional cells, $0 \leq i \leq m$ (objects, arrows, ...) and a associative-like composition rule for the cells with matching boundaries.

- V**-continuity space**

  Let $(V, \wedge, \vee)$ be a *complete* (having $\wedge S := \wedge_{x \in S} x$ and $\vee S := \vee_{x \in S}$ for all $S \subseteq V$) lattice with bottom element 0. For $a, b \in V$, a is said to be *well above* b, denoted by $b \prec a$, if given any $S \subseteq V$ such that $\wedge S \prec b$, there exists $s \in S$ with $s \prec a$.

  A *value quantale* is a pair $(V, +)$, where V is a complete lattice and $+$ is an associative and commutative operation o such that for all $a, b \in V$ and $S \subseteq V$,

  1. $a + \wedge S = \wedge(a + S)$,
  2. $a + 0 = a$,
  3. $a = \wedge\{b \in Va \prec b\}$,
  4. $0 \prec a \wedge b$ if $0 \prec a, b$.

A V**-continuity space** is (Flagg–Koperman, 1997) a triple (X, d, V), where V is a value quantale, X is a set, and $d : X \times X \to V$ is a function satisfying

$$d(x,x) = 0 \quad \text{and} \quad d(x,z) \leq d(x,y) + d(y,z).$$

Any extended quasi-semimetric space is a V-continuity space, where V is the value quantale $[0, \infty]$, seen as a complete lattice, with ordinary addition.

Weiss, 2013, showed that taken with continuous functions, the categories of all V-continuity spaces and of all topological spaces are equivalent. In particular, every topological space (X, τ) is "metrizable" in the sense that there exists a V-continuity space (X, d, V) such that τ is the topology generated by *open balls* $\{y \in X : \prec \epsilon\}$.

- **Probabilistic metric space**

A notion of **probabilistic metric space** is a generalization of the notion of metric space (see, for example, [ScSk83]) in two ways: distances become probability distributions, and the sum in the triangle inequality becomes a **triangle operation**.

Formally, let A be the set of all *probability distribution functions*, whose support lies in $[0, \infty]$. For any $a \in [0, \infty]$ define *step functions* $\epsilon_a \in A$ by $\epsilon_a(x) = 1$ if $x > a$ or $x = \infty$, and $\epsilon_a(x) = 0$, otherwise. The functions in A are ordered by defining $F \leq G$ to mean $F(x) \leq G(x)$ for all $x \geq 0$; the minimal element is ϵ_0.

A commutative and associative operation τ on A is called a **triangle function** if $\tau(F, \epsilon_0) = F$ for any $F \in A$ and $\tau(E, F) \leq \tau(G, H)$ whenever $E \leq G, F \leq H$. The semigroup (A, τ) generalizes the group $(\mathbb{R}, +)$.

A **probabilistic metric space** is a triple (X, D, τ), where X is a set, D is a function $X \times X \to A$, and τ is a triangle function, such that for any $p, q, r \in X$

1. $D(p,q) = \epsilon_0$ if and only if $p = q$;
2. $D(p,q) = D(q,p)$;
3. $D(p,r) \geq \tau(D(p,q), D(q,r))$.

For any metric space (X, d) and any triangle function τ, such that $\tau(\epsilon_a, \epsilon_b) \geq \epsilon_{a+b}$ for all $a, b \geq 0$, the triple $(X, D = \epsilon_{d(x,y)}, \tau)$ is a probabilistic metric space.

For any $x \geq 0$, the value $D(p,q)$ at x can be interpreted as "the probability that the distance between p and q is less than x"; this was approach of Menger, who proposed in 1942 the original version, *statistical metric space*, of this notion.

A probabilistic metric space is called a *Wald space* if the triangle function is a convolution, i.e., of the form $\tau_x(E, F) = \int_{\mathbb{R}} E(x-t)dF(t)$.

A probabilistic metric space is called a **generalized Menger space** if the triangle function has form $\tau_x(E, F) = \sup_{u+v=x} T(E(u), F(v))$ for a *t-norm* T, i.e., such a commutative and associative operation on $[0, 1]$ that $T(a, 1) = a$, $T(0, 0) = 0$ and $T(c, d) \geq T(a, b)$ whenever $c \geq a, d \geq b$.

- **Fuzzy metric spaces**

A *fuzzy subset* of a set S is a mapping $\mu : S \to [0, 1]$, where $\mu(x)$ represents the "degree of membership" of $x \in S$.

[redacted]

A *continuous t-norm* is a binary commutative and associative continuous operation T on $[0, 1]$, such that $T(a, 1) = a$ and $T(c, d) \geq T(a, b)$ whenever $c \geq a, d \geq b$.

A **KM fuzzy metric space** (Kramosil–Michalek, 1975) is a pair $(X, (\mu, T))$, where X is a nonempty set and a *fuzzy metric* (μ, T) is a pair comprising a continuous t-norm T and a fuzzy set $\mu : X^2 \times \mathbb{R}_{\geq 0} \to [0, 1]$, such that, for $x, y, z \in X$ and $s, t \geq 0$, the following conditions hold:

1. $\mu(x, y, 0) = 0$;
2. $\mu(x, y, t) = 1$ if and only if $x = y, t > 0$;
3. $\mu(x, y, t) = \mu(y, x, t)$;
4. $T(\mu(x, y, t), \mu(y, z, s)) \leq \mu(x, z, t + s)$;
5. the function $\mu(x, y, \cdot) : \mathbb{R}_{\geq 0} \to [0, 1]$ is left continuous.

A KM fuzzy metric space is called also a **fuzzy Menger space** since by defining $D_t(p, q) = \mu(p, q, t)$ one gets a **generalized Menger space**. The following modification of the above notion, using a stronger form of metric fuzziness, it a generalized Menger space with $D_t(p, q)$ positive and continuous on $\mathbb{R}_{>0}$ for all p, q.

A **GV fuzzy metric space** (George–Veeramani, 1994) is a pair $(X, (\mu, T))$, where X is a nonempty set, and a *fuzzy metric* (μ, T) is a pair comprising a continuous t-norm T and a fuzzy set $\mu : X^2 \times \mathbb{R}_{>0} \to [0, 1]$, such that for $x, y, z \in X$ and $s, t > 0$

1. $\mu(x, y, t) > 0$;
2. $\mu(x, y, t) = 1$ if and only if $x = y$;
3. $\mu(x, y, t) = \mu(y, x, t)$;
4. $T(\mu(x, y, t), \mu(y, z, s)) \leq \mu(x, z, t + s)$;
5. the function $\mu(x, y, \cdot) : \mathbb{R}_{>0} \to [0, 1]$ is continuous.

An example of a GV fuzzy metric space comes from any metric space (X, d) by defining $T(a, b) = b - ab$ and $\mu(x, y, t) = \frac{t}{t+d(x,y)}$. Conversely, any GV fuzzy metric space (and also any KM fuzzy metric space) generates a metrizable topology. Most GV fuzzy metrics are *strong*, i.e., $T(\mu(x, y, t), \mu(y, z, t)) \leq \mu(x, z, t)$ holds.

A *fuzzy number* is a fuzzy set $\mu : \mathbb{R} \to [0, 1]$ which is *normal* ($\{x \in \mathbb{R} : \mu(x) = 1\} \neq \emptyset$), *convex* ($\mu(tx + (1 - t)y) \geq \min\{\mu(x), \mu(y)\}$ for every $x, y \in \mathbb{R}$ and $t \in [0, 1]$) and *upper semicontinuous* (at each point x_0, the values $\mu(x)$ for x near x_0 are either close to $\mu(x_0)$ or less than $\mu(x_0)$). Denote the set of all fuzzy numbers which are *nonnegative*, i.e., $\mu(x) = 0$ for all $x < 0$, by G. The additive and multiplicative identities of fuzzy numbers are denoted by $\tilde{0}$ and $\tilde{1}$, respectively. The *level set* $[\mu]_t = \{x : \mu(x) \geq t\}$ of a fuzzy number μ is a closed interval.

Given a nonempty set X and a mapping $d : X^2 \to G$, let the mappings $L, R : [0, 1]^2 \to [0, 1]$ be symmetric and nondecreasing in both arguments and satisfy $L(0, 0) = 0$, $R(1, 1) = 1$. For all $x, y \in X$ and $t \in (0, 1]$, let $[d(x, y)]_t = [\lambda_t(x, y), \rho_t(x, y)]$.

A **KS fuzzy metric space** (Kaleva–Seikkala, 1984) is a quadruple (X, d, L, R) with *fuzzy metric* d, if for all $x, y, z \in X$

1. $d(x, y) = \tilde{0}$ if and only if $x = y$;
2. $d(x, y) = d(y, x)$;
3. $d(x, y)(s + t) \geq L(d(x, z)(s), d(z, y)(t))$ whenever $s \leq \lambda_1(x, z)$, $t \leq \lambda_1(z, y)$, and $s + t \leq \lambda_1(x, y)$;
4. $d(x, y)(s + t) \leq R(d(x, z)(s), d(z, y)(t))$ whenever $s \geq \lambda_1(x, z)$, $t \geq \lambda_1(z, y)$, and $s + t \geq \lambda_1(x, y)$.

The following functions are some frequently used choices for L and R:

$$\max\{a + b - 1, 0\}, ab, \min\{a, b\}, \max\{a, b\}, a + b - ab, \min\{a + b, 1\}.$$

Several other notions of **fuzzy metric space** were proposed, including those by Erceg, 1979, Deng, 1982, and Voxman, 1998, Xu–Li, 2001, Tran–Duckstein, 2002, Chakraborty–Chakraborty, 2006. Cf. also **metrics between fuzzy sets**, **fuzzy Hamming distance**, **gray-scale image distances** and **fuzzy polynucleotide metric** in Chaps. 1, 11, 21 and 23, respectively.

- **Interval-valued metric space**

Let $I(\mathbb{R}_{\geq 0})$ denote the set of closed intervals of $\mathbb{R}_{\geq 0}$.

An **interval-valued metric space** (Coppola–Pacelli, 2006) is a pair $((X, \leq), \Delta)$, where $(X, \leq)$ is a partially ordered set and Δ is an interval-valued mapping $\Delta : X \times X \to I(\mathbb{R}_{\geq 0})$, such that for every $x, y, z \in X$

1. $\Delta(x, x) \star [0, 1] = \Delta(x, x)$;
2. $\Delta(x, y) = \Delta(y, x)$;
3. $\Delta(x, y) - \Delta(z, z) \preceq \Delta(x, z) + \Delta(z, y)$;
4. $\Delta(x, y) - \Delta(x, y) \preceq \Delta(x, x) + \Delta(y, y)$;
5. $x \leq x'$ and $y \leq y'$ imply $\Delta(x, y) \subseteq \Delta(x', y')$;
6. $\Delta(x, y) = 0$ if and only if $x = y$ and x, y are *atoms* (minimal elements of $(X, \leq)$).

Here the following *interval arithmetic* rules hold: $[u, v] \preceq [u', v']$ if and only if $u \leq u'$,

$[u, v] + [u', v'] = [u + u', v + v']$, $[u, v] - [u', v'] = [u - u', v - v']$,
$[u, v] \star [u', v'] = [\min\{uu', uv', vu', vv'\}, \max\{uu', uv', vu', vv'\}]$ and
$\frac{[u,v]}{[u',v']} = [\min\{\frac{u}{u'}, \frac{u}{v'}, \frac{v}{u'}, \frac{v}{v'}\}, \max\{\frac{u}{u'}, \frac{u}{v'}, \frac{v}{u'}, \frac{v}{v'}\}]$ when $0 \notin [u', v']$.

The addition and multiplication operations are commutative, associative and *subdistributive*: it holds $X \star (Y + Z) \subset (X \star Y + X \star Z)$.

Cf. **metric between intervals** in Chap. 10.

The usual metric spaces coincide with above spaces in which all $x \in X$ are atoms.

- **Direction distance**

Given a normed real vector space $(V, ||.||)$, for any $x \in V \setminus \{0\}$, denote by $[x]$ the *direction* (ray) $\{\lambda x : \lambda > 0\}$ and by x_0 the point $\frac{x}{||x||}$. An *oriented angle* is an

[redacted]

ordered pair $([x],[y])$ of directions. The **direction distance** from x to y is defined (Busch–Ruch, 1992) as the family of distances $||\alpha x_0 - \beta y_0||$ with $\alpha, \beta \in \mathbb{R}_{>0}$.

The **mixing distance** is defined as the restriction of the direction distance to pairs of directions in the cone $\{\lambda v : v \in V, \lambda > 0\}$. In fact, authors introduced these distances on some special normed spaces used in Quantum Mechanics.

- **Generalized metric**

  Let X be a set. Let $(V,+,\leq)$ be an *ordered semigroup* (not necessarily commutative) with a least element θ and with $x \leq y, x_1 \leq y_1$ implying $x+x_1 \leq y+y_1$. Let $(V,+)$ be also endowed with an order-preserving *involution* x^* (i.e., $(x^*)^* = x$), which is operation-reversing, i.e., $(x+y)^* = y^* + x^*$.

  A function $d : X \times X \to G$ is called (Li–Wang–Pouzet, 1987) a **generalized metric** over $(V,+,\leq)$ if the following conditions hold:

  1. $d(x,y) = \theta$ if and only if $x = y$;
  2. $d(x,y) \leq d(x,z) + d(z,y)$ for all $x,y \in X$;
  3. $d^*(x,y) = d(y,x)$.

- **Cone metric**

  Let C be a *proper cone* in a real Banach space W, i.e., C is closed, $C \neq \emptyset$, the interior of C is not equal to $\{\theta\}$ (where θ is the zero vector in W) and

  1. if $x,y \in C$ and $a,b \in \mathbb{R}_{\geq 0}$, then $ax+by \in C$;
  2. if $x \in C$ and $-x \in C$, then $x = 0$.

  Define a partial ordering $(W,\leq)$ on W by letting $x \leq y$ if $y-x \in C$. The following variation of **generalized metric** and **partially ordered distance** was defined in Huang–Zhang, 2007, and, partially, in Rzepecki, 1980. Given a set X, a **cone metric** is a mapping $d : X \times X \to (W,\leq)$ such that

  1. $\theta \leq d(x,y)$ with equality if and only if $x = y$;
  2. $d(x,y) = d(y,x)$ for all $x,y \in X$;
  3. $d(x,y) \leq d(x,z) + d(z,y)$ for all $x,y \in X$;

  The pair (X,d) is called a **cone metric space**.

- ***W*-distance on building**

  Let X be a set, and let $(W,\cdot,1)$ be a group. A *W-distance* on X is a W-valued map $\sigma : X \times X \to W$ having the following properties:

  1. $\sigma(x,y) = 1$ if and only if $x = y$;
  2. $\sigma(y,x) = (\sigma(x,y))^{-1}$.

  A natural W-distance on W is $\sigma(x,y) = x^{-1}y$.

  A *Coxeter group* is a group $(W,\cdot,1)$ generated by the elements

$$\{w_1,\dots,w_n : (w_i w_j)^{m_{ij}} = 1, 1 \leq i,j \leq n\}.$$

  Here $M = ((m_{ij}))$ is a *Coxeter matrix*, i.e., an arbitrary symmetric $n \times n$ matrix with $m_{ii} = 1$, and the other values are positive integers or ∞. The *length* $l(x)$ of $x \in W$ is the smallest number of generators $w_1,\dots,w_n$ needed to represent x.

Let X be a set, let $(W, \cdot, 1)$ be a Coxeter group and let $\sigma(x, y)$ be a W-distance on X. The pair (X, σ) is called (Tits, 1981) a *building* over $(W, \cdot, 1)$ if it holds

1. the relation $\sim_i$ defined by $x \sim_i y$ if $\sigma(x, y) = 1$ or w_i, is an equivalence relation;
2. given $x \in X$ and an equivalence class C of $\sim_i$, there exists a unique $y \in C$ such that $\sigma(x, y)$ is *shortest* (i.e., of smallest length), and $\sigma(x, y') = \sigma(x, y)w_i$ for any $y' \in C, y' \neq y$.

The **gallery distance on building** d is a usual metric on X defined by $l(d(x, y))$. The distance d is the **path metric** in the graph with the vertex-set X and xy being an edge if $\sigma(x, y) = w_i$ for some $1 \leq i \leq n$. The gallery distance on building is a special case of a **gallery metric** (of *chamber system* X).

- **Boolean metric space**

  A *Boolean algebra* (or *Boolean lattice*) is a *distributive lattice* $(B, \vee, \wedge)$ admitting a least element 0 and greatest element 1 such that every $x \in B$ has a *complement* $\bar{x}$ with $x \vee \bar{x} = 1$ and $x \wedge \bar{x} = 0$.

  Let X be a set, and let $(B, \vee, \wedge)$ be a Boolean algebra. The pair (X, d) is called (Blumenthal, 1953) a **Boolean metric space** over B if the function $d : X \times X \to B$ has the following properties:

  1. $d(x, y) = 0$ if and only if $x = y$;
  2. $d(x, y) \leq d(x, z) \vee d(z, y)$ for all $x, y, z \in X$.

- **Space over algebra**

  A **space over algebra** is a metric space with a differential-geometric structure, whose points can be provided with coordinates from some *algebra* (usually, an associative algebra with identity).

  A *module* over an algebra is a generalization of a vector space over a field, and its definition can be obtained from the definition of a vector space by replacing the field by an associative algebra with identity. An *affine space over an algebra* is a similar generalization of an *affine space* over a field. In affine spaces over algebras one can specify a Hermitian metric, while in the case of commutative algebras even a quadratic metric can be given. To do this one defines in a unital module a *scalar product* $\langle x, y\rangle$, in the first case with the property $\langle x, y\rangle = J(\langle y, x\rangle)$, where J is an *involution* of the algebra, and in the second case with the property $\langle y, x\rangle = \langle x, y\rangle$.

  The n-dimensional *projective space over an algebra* is defined as the variety of one-dimensional submodules of an $(n + 1)$-dimensional unital module over this algebra. The introduction of a *scalar product* $\langle x, y\rangle$ in a unital module makes it possible to define a Hermitian metric in a projective space constructed by means of this module or, in the case of a commutative algebra, quadratic elliptic and hyperbolic metrics. The metric invariant of the points of these spaces is the **cross-ratio** $W = \langle x, x\rangle^{-1}\langle x, y\rangle\langle y, y\rangle^{-1}\langle y, x\rangle$. If W is a real number, then $w = \arccos\sqrt{W}$ is called the **distance** between x and y **in the space over algebra**.

[redacted]

- **Partially ordered distance**

  Let X be a set. Let $(G, \leq)$ be a *partially ordered set* with a least element g_0. A **partially ordered distance** is a function $d : X \times X \to G$ such that, for any $x, y \in X$, $d(x, y) = g_0$ if and only if $x = y$.

  A **generalized ultrametric** (Priess-Crampe and Ribenboim, 1993) is a symmetric (i.e., $d(x, y) = d(y, x)$) partially ordered distance, such that $d(z, x) \leq g$ and $d(z, y) \leq g$ imply $d(x, y) \leq g$ for any $x, y, z \in X$ and $g \in G$.

  Suppose that $G' = G \backslash \{g_0\} \neq \emptyset$ and, for any $g_1, g_2 \in G'$, there exists $g_3 \in G'$ such that $g_3 \leq g_1$ and $g_3 \leq g_2$. Consider the following possible properties:

  1. For any $g_1 \in G'$, there exists $g_2 \in G'$ such that, for any $x, y \in X$, from $d(x, y) \leq g_2$ it follows that $d(y, x) \leq g_1$;
  2. For any $g_1 \in G'$, there exist $g_2, g_3 \in G'$ such that, for any $x, y, z \in X$, from $d(x, y) \leq g_2$ and $d(y, z) \leq g_3$ it follows that $d(x, z) \leq g_1$;
  3. For any $g_1 \in G'$, there exists $g_2 \in G'$ such that, for any $x, y, z \in X$, from $d(x, y) \leq g_2$ and $d(y, z) \leq g_2$ it follows that $d(y, x) \leq g_1$;
  4. G' has no first element;
  5. $d(x, y) = d(y, x)$ for any $x, y \in X$;
  6. For any $g_1 \in G'$, there exists $g_2 \in G'$ such that, for any $x, y, z \in X$, from $d(x, y) <^* g_2$ and $d(y, z) <^* g_2$ it follows that $d(x, z) <^* g_1$; here $p <^* q$ means that either $p < q$, or p is not comparable to q;
  7. The order relation $<$ is a total ordering of G.

  In terms of above properties, d is called: the **Appert partially ordered distance** if 1 and 2 hold; the **Golmez partially ordered distance of first type** if 4, 5, and 6 hold; the **Golmez partially ordered distance of second type** if 3, 4, and 5 hold; the **Kurepa–Fréchet distance** if 3, 4, 5, and 7 hold.

  The case $G = \mathbb{R}_{\geq 0}$ of the Kurepa–Fréchet distance corresponds to the **Fréchet V-space**; cf. the **f-quasi-metric** in Sect. 1.1. The general case was considered in Kurepa, 1934, and rediscovered in Fréchet, 1946.

- **Distance from measurement**

  Distance from measurement is an analog of distance on domains in Computer Science; it was developed in [Mart00].

  A *po* (partially ordered set) $(D, \preceq)$ is called *dcpo* (directed-complete po) if every *directed subset* $S \subset D$ (i.e., $S \neq \emptyset$ and any pair $x, y \in S$ is *bounded*: there is $z \in S$ with $x, y \preceq z$) has a *supremum* $\sqcup S$, i.e., the least of such upper bounds z.

  For $x, y \in D$, y is an *approximation* of x if, for all directed subsets $S \subset D$, $x \preceq \sqcup S$ implies $y \preceq s$ for some $s \in S$. A dcpo $(D, \preceq)$ is *continuous* if for all $x \in D$ the set of all approximations of x is directed and x is its supremum. A *domain* is a continuous dcpo $(D, \preceq)$ such that for all $x, y \in D$ there is $z \in D$ with $z \preceq x, y$. A *Scott domain* is a domain with least element, in which any bounded pair has a supremum.

  A subset U of a dcpo $(D, \preceq)$ is *Alexandrov open* if, for any $x \in U$ and $y \in D$, $x \preceq y$ implies $y \in U$; it is *Scott open* if also, for any directed subset $S \subset D$, $\sqcup S \in U$ implies $S \cap U \neq \emptyset$. The set of Scott open sets form the *Scott topology*; it is a **T_0-space** (Chap. 2) with generalized metrization by a **partial metric** (Chap. 1).

A *measurement* is a mapping $\mu : D \to \mathbb{R}_{\geq 0}$ between dcpo $(D, \preceq)$ and dcpo $(\mathbb{R}_{\geq 0}, \preceq)$, where $\mathbb{R}_{\geq 0}$ is ordered as $x \preceq y$ if $y \leq x$, such that

1. $x \preceq y$ implies $\mu(x) \preceq \mu(y)$;
2. $\mu(\sqcup S) = \sqcup(\{\mu(s) : s \in S\})$ for every directed subset $S \subset D$;
3. For all $x \in D$ with $\mu(x) = 0$ and all sequences $(x_n), n \to \infty$, of approximations of x with $\lim_{n \to \infty} \mu(x_n) = \mu(x)$, one has $\sqcup(\cup_{n=1}^{\infty}\{x_n\}) = x$.

Given a measurement μ, the **distance from measurement** is a mapping $d : D \times D \to \mathbb{R}_{\geq 0}$ given by

$$d(x, y) = \inf\{\mu(z) : z \text{ approximates } x, y\} = \inf\{\mu(z) : z \preceq x, y\}.$$

One has $d(x, x) \preceq \mu(x)$. The function $d(x, y)$ is a metric on the set $\{x \in D : \mu(x) = 0\}$ if μ satisfies the following **measurement triangle inequality**: for all bounded pairs $x, y \in D$, there is an element $z \preceq x, y$ such that $\mu(z) \leq \mu(x) + \mu(y)$.

Waszkiewicz, 2001, found topological connections between topologies coming from a distance from measurement and from a **partial metric** defined in Chap. 1.

[redacted]

Chapter 4
Metric Transforms

There are many ways to obtain new distances (metrics) from given distances (metrics). Metric transforms give new distances as a functions of given metrics (or given distances) on the same set X. A metric so obtained is called a **transform metric**. We give some important examples of transform metrics in Sect. 4.1.

Given a metric space (X, d), one can construct a new metric on an extension of X; similarly, given a collection of metrics on sets $X_1, \dots, X_n$, one can obtain a new metric on an extension of $X_1, \dots, X_n$. Examples of such operations are given in Sect. 4.2. There are many distances on other structures connected with X, say, on the set of all subsets of X. The main distances of this kind are considered in Sect. 4.3.

4.1 Metrics on the Same Set

- **Metric transform**

  A **metric transform** is a distance on a set X, obtained as a function of given metrics (or given distances) on X.

  In particular, given a continuous monotone increasing function $f(x)$ of $x \geq 0$ with $f(0) = 0$, called the *scale*, and a distance space (X, d), one obtains another distance space (X, d_f), called a **scale metric transform** of X, defining $d_f(x, y) = f(d(x, y))$. For every finite distance space (X, d), there exists a scale f, such that (X, d_f) is a metric subspace of a Euclidean space $\mathbb{R}^n$.

  If (X, d) is a metric space and f is a continuous differentiable strictly increasing scale with $f(0) = 0$ and nonincreasing f', then (X, d_f) is a metric space (cf. **functional transform metric**).

  The metric d is an **ultrametric** if and only if $f(d)$ is a metric for every nondecreasing function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

- **Transform metric**

  A **transform metric** is a metric on a set X which is a **metric transform**, i.e., is obtained as a function of a given metric (or given metrics) on X. In particular,

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_4

transform metrics can be obtained from a given metric d (or given metrics d_1 and d_2) on X by any of the following operations (here $t > 0$):

1. $td(x, y)$ (**t-scaled metric**, or **dilated metric**, **similar metric**);
2. $\min\{t, d(x, y)\}$ (**t-truncated metric**);
3. $\max\{t, d(x, y)\}$ for $x \neq y$ (**t-uniformly discrete metric**);
4. $d(x, y) + t$ for $x \neq y$ (**t-translated metric**);
5. $\frac{kd(x,y)}{1+d(x,y)}$ (this metric has diameter less than k);
6. $d^p(x, y) = \frac{2d(x,y)}{d(x,p)+d(y,p)+d(x,y)}$, where p is an fixed element of X (**biotope transform metric**, or **p-smoothing distance** on $X \setminus \{p\}$);
7. $\max\{d_1(x, y), d_2(x, y)\}$;
8. $\alpha d_1(x, y) + \beta d_2(x, y)$, where $\alpha, \beta > 0$ (cf. **semimetric cone** in Chap. 1).

- **Generalized biotope transform metric**

  For a given metric d on a set X and a closed set $M \subset X$, the **generalized biotope transform metric** d^M on X is defined by

$$d^M(x, y) = \frac{2d(x, y)}{d(x, y) + \inf_{z \in M}(d(x, z) + d(y, z))}.$$

  In fact, $d^M(x, y)$ and its **1-truncation** $\min\{1, d^M(x, y)\}$ are both metrics.

  The **biotope transform metric** is $d^M(x, y)$ with $|M| = 1$. The **Steinhaus distance** from Chap. 1 is the case $d(x, y) = \mu(x \triangle y)$ with $p \neq \emptyset$ and the **biotope distance** from Chap. 23 is its subcase $d(x, y) = \mu(x \triangle y) = |x \triangle y|$.

- **Metric-preserving function**

  A function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $f^{-1}(0) = \{0\}$ is a **metric-preserving function** if, for each metric space (X, d), the **metric transform**

$$d_f(x, y) = f(d(x, y))$$

  is a metric on X; cf. [Cora99]. In this case d_f is called a **functional transform metric**. For example, αd ($\alpha > 0$), $d^\alpha (0 < \alpha \leq 1)$, $\ln(1 + d)$, $\operatorname{arcsinh} d$, $\operatorname{arccosh}(1 + d)$, and $\frac{d}{1+d}$ are functional transform metrics.

  The superposition, sum and maximum of two metric-preserving functions are metric-preserving. If f is *subadditive*, i.e. $f(x + y) \leq f(x) + f(y)$ for all $x, y \geq 0$, and nondecreasing, then it is metric-preserving. But, for example, the function $f(x) = \frac{x+2}{x+1}$, for $x > 0$, and $f(0) = 0$, is decreasing and metric-preserving. If f is metric-preserving, then it is subadditive.

  If f is *concave*, i.e., $f(\frac{x+y}{2}) \geq \frac{f(x)+f(y)}{2}$ for all $x, y \geq 0$, then it is metric-preserving. In particular, a twice differentiable function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $f(0) = 0, f'(x) > 0$ for all $x \geq 0$, and $f''(x) \leq 0$ for all $x \geq 0$, is metric-preserving.

  The function f is **strongly metric-preserving function** if d and $f(d(x, y))$ are **equivalent metrics** on X, for each metric space (X, d). A metric-preserving function is strongly metric-preserving if and only if it is continuous at 0.

[redacted]

- **Metric aggregating function**

  A function $f : \mathbb{R}^2_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $f(a, b) = 0$ if and only if $a = b = 0$, is said to be **metric** (respectively, **quasi-metric**) **aggregating function** if the function $d_f : X \times X \to \mathbb{R}_{\geq 0}$ is a metric for every pair of metric spaces (respectively, a quasi-metric for every pair of quasi-metric spaces) (X_1, d_1) and (X_2, d_2), where $X = X_1 \times X_2$ and, for all $(x, z), (y, w) \in X$, it holds

  $$d_f((x,z),(y,w)) = f(d_1(x,z), d_2(y,w)).$$

  Borsík–Doboš, 1981, proved that a function f is metric aggregating if and only if, for all $a, b, c, a', b', c' \geq 0$ with $|a-b| \leq c \leq a+b$ and $|a'-b'| \leq c' \leq a'+b'$, it holds

  $$|f(a,a') - f(b,b')| \leq f(c,c') \leq f(a,a') + f(b,b').$$

  Cf. **spin triangle inequality** in Chap. 15.

  Major–Valero, 2008, proved that a function f is quasi-metric agregating if and only if it holds $f(a,a') \leq f(b,c') + f(c,b')$ for all $a, b, c, a', b', c' \geq 0$ such that $a \leq b + c$ and $a' \leq b' + c'$; so, any quasi-metric agregating function is metric agregating.
- **Metric generating function**

  A symmetric function $f : \mathbb{R}^2_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $f(a, b) = 0$ if and only if $a = b = 0$, is said to be **metric generating** if the function defined by

  $$d_f(x,y) = f(d(x,y), d(y,x))$$

  for all $x, y \in X$ is a metric on X for every quasi-metric space (X, d).

  Martin–Major–Valero, 2013, proved that a function f is metric generating if and only if it holds $f(a,a') \leq f(b,c') + f(c,b')$ for all a, b, c, a', b', c' such that $a \leq b + c, b \leq a + b', c \leq c' + a$ and $a' \leq b' + c', b' \leq a' + b, c' \leq c + a'$.
- **Power transform metric**

  Let $0 < \alpha \leq 1$. Given a metric space (X, d), the **power** (or *α-snowflake*) **transform metric** is a **functional transform metric** on X defined by

  $$(d(x,y))^{\alpha}.$$

  The distance $d(x,y) = (\sum_1^n |x_i - y_i|^p)^{\frac{1}{p}}$ with $0 < p = \alpha < 1$ is not a metric on $\mathbb{R}^n$, but its power transform $(d(x,y)^{\alpha})$ is a metric.

  For a given metric d on X and any $\alpha > 1$, the function d^{α} is, in general, only a distance on X. It is a metric, for any positive α, if and only if d is an **ultrametric**.

  A metric d is a **doubling metric** if and only if (Assouad, 1983) the power transform metric d^{α} admits a **bi-Lipschitz embedding** in some Euclidean space for every $0 < \alpha < 1$ (cf. Chap. 1 for definitions).

- **Quadrance**

  A distance which is a squared distance d^2 is called a **quadrance**.

  Rational trigonometry is the proposal (Wildberger, 2007) to use as its fundamental units, quadrance and *spread* (square of sine of angle), instead of distance and angle.

  It makes some problems easier to computers: solvable with only addition, subtraction, multiplication, and division, while avoiding square roots, sine, and cosine functions. Also, such trigonometry can be done over any field.
- **Schoenberg transform metric**

  Let $\lambda > 0$. Given a metric space (X, d), the **Schoenberg transform metric** is a **functional transform metric** on X defined by

  $$1 - e^{-\lambda d(x,y)}.$$

  The Schoenberg transform metrics are exactly P**-metrics** (cf. Chap. 1).
- **Pullback metric**

  Given two metric spaces (X, d_X), (Y, d_Y) and an injective mapping $g : X \to Y$, the **pullback metric** (of (Y, d_Y) by g) on X is defined by

  $$d_Y(g(x), g(y)).$$

  If $(X, d_X) = (Y, d_Y)$, then the pullback metric is called a g**-transform metric**.
- **Internal metric**

  Given a metric space (X, d) in which every pair of points x, y is joined by a *rectifiable curve*, the **internal metric** (or **inner metric**, induced **intrinsic metric**, **interior metric**) D is a **transform metric** on X, obtained from d as the infimum of the lengths of all rectifiable curves connecting two given points x and $y \in X$.

  The metric d is called an **intrinsic metric** (or **length metric** if it coincides with its internal metric. Cf. Chap. 6 and **metric curve** in Chap. 1.
- **Farris transform metric**

  Given a metric space (X, d) and a point $z \in X$, the **Farris transform** is a metric transform D_z on $X \backslash \{z\}$ defined by $D_z(x, x) = 0$ and, for different $x, y \in X \backslash \{z\}$, by

  $$D_z(x, y) = C - (x.y)_z,$$

  where C is a positive constant, and $(x.y)_z = \frac{1}{2}(d(x, z) + d(y, z) - d(x, y))$ is the **Gromov product** (cf. Chap. 1). It is a metric if $C \geq \max_{x \in X \backslash \{z\}} d(x, z)$; in fact, there exists a number $C_0 \in (\max_{x,y \in X \backslash \{z\}, x \neq y} (x.y)_z, \max_{x \in X \backslash \{z\}} d(x, z)]$ such that it is a metric if and only if $C \geq C_0$. The Farris transform is an **ultrametric** if and only if d satisfies the **four-point inequality**. In Phylogenetics, where it was applied first, the term *Farris transform* is used for the function $d(x, y) - d(x, z) - d(y, z)$.

[redacted]

- **Involution transform metric**

  Given a metric space (X, d) and a point $z \in X$, the **involution transform metric** is a metric transform d_z on $X \backslash \{z\}$ defined by

$$d_z(x, y) = \frac{d(x, y)}{d(x, z)d(y, z)}.$$

  It is a metric for any $z \in X$, if and only if d is a **Ptolemaic metric** ([FoSc06]).

4.2 Metrics on Set Extensions

- **Extension distances**

  If d is a metric on $V_n = \{1, \dots, n\}$, and $\alpha \in \mathbb{R}, \alpha > 0$, then the following extension distances (see, for example, [DeLa97]) are used.

  The **gate extension distance** $gat = gat_\alpha^d$ is a metric on $V_{n+1} = \{1, \dots, n+1\}$ defined by the following conditions:

  1. $gat(1, n+1) = \alpha$;
  2. $gat(i, n+1) = \alpha + d(1, i)$ if $2 \leq i \leq n$;
  3. $gat(i,j) = d(i,j)$ if $1 \leq i < j \leq n$.

  The distance gat_0^d is called the **gate** 0**-extension** or, simply, 0**-extension** of d.

  If $\alpha \geq \max_{2 \leq i \leq n} d(1, i)$, then the **antipodal extension distance** $ant = ant_\alpha^d$ is a distance on V_{n+1} defined by the following conditions:

  1. $ant(1, n+1) = \alpha$;
  2. $ant(i, n+1) = \alpha - d(1, i)$ if $2 \leq i \leq n$;
  3. $ant(i,j) = d(i,j)$ if $1 \leq i < j \leq n$.

  If $\alpha \geq \max_{1 \leq i,j \leq n} d(i,j)$, then the **full antipodal extension distance** $Ant = Ant_\alpha^d$ is a distance on $V_{2n} = \{1, \dots, 2n\}$ defined by the following conditions:

  1. $Ant(i, n+i) = \alpha$ if $1 \leq i \leq n$;
  2. $Ant(i, n+j) = \alpha - d(i,j)$ if $1 \leq i \neq j \leq n$;
  3. $Ant(i,j) = d(i,j)$ if $1 \leq i \neq j \leq n$;
  4. $Ant(n+i, n+j) = d(i,j)$ if $1 \leq i \neq j \leq n$.

  It is obtained by applying the antipodal extension operation iteratively n times, starting from d.

  The **spherical extension distance** $sph = sph_\alpha^d$ is a metric on V_{n+1} defined by the following conditions:

  1. $sph(i, n+1) = \alpha$ if $1 \leq i \leq n$;
  2. $sph(i,j) = d(i,j)$ if $1 \leq i < j \leq n$.

- **1-sum distance**

  Let d_1 be a distance on a set X_1, let d_2 be a distance on a set X_2, and suppose that $X_1 \cap X_2 = \{x_0\}$. The **1-sum distance** of d_1 and d_2 is the distance d on $X_1 \cup X_2$ defined by the following conditions:

$$d(x,y) = \begin{cases} d_1(x,y), & \text{if} \quad x,y \in X_1, \\ d_2(x,y), & \text{if} \quad x,y \in X_2, \\ d(x,x_0) + d(x_0,y), & \text{if} \ x \in X_1, y \in X_2. \end{cases}$$

  In Graph Theory, the 1-sum distance is a **path metric**, corresponding to the clique 1-sum operation for graphs.
- **Disjoint union metric**

  Given a family (X_t, d_t), $t \in T$, of metric spaces, the **disjoint union metric** is an **extended metric** on the set $\bigcup_t X_t \times \{t\}$ defined by

$$d((x,t_1),(y,t_2)) = d_t(x,y)$$

  for $t_1 = t_2$, and $d((x,t_1),(y,t_2)) = \infty$, otherwise.
- **Metric bouquet**

  Given a family (X_t, d_t), $t \in T$, of metric spaces with marked points x_t, the **metric bouquet** is obtained from their **disjoint union** by gluing all points x_t together.
- **Product metric**

  Given finite or countable number n of metric spaces (X_1,d_1), (X_2,d_2), $\dots$, (X_n,d_n), the **product metric** is a metric on the *Cartesian product* $X_1 \times X_2 \times \cdots \times X_n = \{x = (x_1,x_2,\dots,x_n) : x_1 \in X_1,\dots,x_n \in X_n\}$ defined as a function of $d_1,\dots,d_n$. The simplest finite product metrics are defined by

  1. $(\sum_{i=1}^{n} d_i^p(x_i,y_i))^{\frac{1}{p}}$, $1 \le p < \infty$;
  2. $\max_{1 \le i \le n} d_i(x_i,y_i)$;
  3. $\sum_{i=1}^{n} \frac{1}{2^i} \frac{d_i(x_i,y_i)}{1+d_i(x_i,y_i)}$.

  The last metric is **bounded** and can be extended to the product of countably many metric spaces.

  If $X_1 = \cdots = X_n = \mathbb{R}$, and $d_1 = \cdots = d_n = d$, where $d(x,y) = |x-y|$ is the **natural metric** on $\mathbb{R}$, all product metrics above induce the Euclidean topology on the n-dimensional space $\mathbb{R}^n$. They do not coincide with the Euclidean metric on $\mathbb{R}^n$, but they are equivalent to it. In particular, the set $\mathbb{R}^n$ with the Euclidean metric can be considered as the Cartesian product $\mathbb{R} \times \cdots \times \mathbb{R}$ of n copies of the *real line* $(\mathbb{R}, d)$ with the product metric defined by $\sqrt{\sum_{i=1}^{n} d^2(x_i,y_i)}$.
- **Box metric**

  Let (X,d) be a metric space and I the unit interval of $\mathbb{R}$. The **box metric** is the **product metric** d' on the Cartesian product $X \times I$ defined by

$$d'((x_1,t_1),(x_2,t_2)) = \max(d(x_1,x_2), |t_1 - t_2|).$$

[redacted]

Cf. unrelated **bounded box metric** in Chap. 18.

- **Fréchet product metric**

  Let (X, d) be a metric space with a **bounded** metric d. Let $X^{\infty} = X \times \cdots \times X \cdots = \{x = (x_1, \dots, x_n, \dots) : x_1 \in X_1, \dots, x_n \in X_n, \dots\}$ be the *countable Cartesian product space* of X.

  The **Fréchet product metric** is a **product metric** on X^{∞} defined by

$$\sum_{n=1}^{\infty} A_n d(x_n, y_n),$$

  where $\sum_{n=1}^{\infty} A_n$ is any convergent series of positive terms. Usually, $A_n = \frac{1}{2^n}$ is used.

  A metric (sometimes called the *Fréchet metric*) on the set of all sequences $\{x_n\}_n$ of real (complex) numbers, defined by

$$\sum_{n=1}^{\infty} A_n \frac{|x_n - y_n|}{1 + |x_n - y_n|},$$

  where $\sum_{n=1}^{\infty} A_n$ is any convergent series of positive terms, is a Fréchet product metric of countably many copies of $\mathbb{R}$ ($\mathbb{C}$). Usually, $A_n = \frac{1}{n!}$ or $A_n = \frac{1}{2^n}$ are used.

- **Hilbert cube metric**

  The *Hilbert cube* $I^{\aleph_0}$ is the *Cartesian product* of countable many copies of the interval $[0, 1]$, equipped with the metric

$$\sum_{i=1}^{\infty} 2^{-i} |x_i - y_i|$$

  (cf. **Fréchet infinite metric product**). It also can be identified up to homeomorphisms with the compact metric space formed by all sequences $\{x_n\}_n$ of real numbers such that $0 \leq x_n \leq \frac{1}{n}$, where the metric is defined as $\sqrt{\sum_{n=1}^{\infty} (x_n - y_n)^2}$.

  The Cartesian products $[0, 1]^{\tau}$ and $\{0, 1\}^{\tau}$, where τ is an arbitrary cardinal number, are called a *Tikhonov cube* and *Cantor cube*, respectively.

- **Hamming cube**

  Given integers $n \geq 1$ and $q \geq 2$, the *Hamming space* $H(n, q)$ is the set of all n-tuples over an alphabet of size q (say, the *Cartesian product* of n copies of the set $\{0, 1, \dots, q - 1\}$), equipped with the **Hamming metric** (cf. Chap. 1), i.e., the distance between two n-tuples is the number of coordinates where they differ. The **Hamming cube** is the Hamming space $H(n, 2)$.

  The *infinite Hamming cube* $H(\infty)$ is the set of all infinite strings over the alphabet $\{0, 1\}$ containing only finitely many 1's, equipped with the Hamming metric.

  The *half-cube* $\frac{1}{2}H(n)$ is the set of all n-tuples over $\{0, 1\}$, containing even number of 1's, with two of them being adjacent if they differ exactly in two coordinates. For an integer k, $1 \leq k < n$, the *Johnson graph* $J(n, k)$ is the set of all n-tuples over $\{0, 1\}$, containing k 1's, with the same notion of adjacence.

The *Fibonacci cube* $F(n)$ is the set of all n-tuples over $\{0, 1\}$ that contain no two consecutive 1's, equipped with the Hamming metric; it is a **partial cube** (Chap. 15), i.e., an isometric subgraph of $H(n, 2)$. The *Lucas cube* $L(n)$ is obtained from $F(n)$ by removing n-tuples that start and end with 1.

- **Cameron–Tarzi cube**

Given integers $n \geq 1$ and $q \geq 2$, the *normalized Hamming space* $H_n(q)$ is the set of all n-tuples over an alphabet of size q, equipped with the **Hamming metric** divided by n. Clearly, there are isometric embeddings

$$H_1(q) \to H_2(q) \to H_4(q) \to H_8(q) \to \dots$$

Let $H(q)$ denote the **Cauchy completion** (Chap. 1) of the union (denote it by $H_\omega(q)$) of all metric spaces $H_n(q)$ with $n \geq 1$. This metric space was introduced in [CaTa08]. Call $H(2)$ the **Cameron–Tarzi cube**.

It is shown in [CaTa08] that $H_\omega(2)$ is the **word metric** space (Chap. 20) of the *countable Nim group*, i.e., the elementary Abelian 2-group of all natural numbers under bitwise addition modulo 2 of the number expressions in base 2. The Cameron-Tarzi cube is also the word metric space of an Abelian group.

- **Rubik cube**

There is a bijection between legal positions of the *Rubik* $3 \times 3 \times 3$ *cube* and elements of the subgroup G of the group Sym_{48} (of all permutations of $6(9 - 1)$ movable facets) generated by the 6 face rotations. The number of possible positions attainable by the cube is $|G| \approx 43 \times 10^{18}$.

The maximum number of face turns needed to solve any instance of the Rubik cube is the diameter (maximal **word metric**), 20, of the *Cayley graph* of G.

- **Warped product metric**

Let (X, d_X) and (Y, d_Y) be two complete **length spaces** (cf. Chap. 6), and let $f : X \to \mathbb{R}$ be a positive continuous function. Given a curve $\gamma : [a, b] \to X \times Y$, consider its projections $\gamma_1 : [a, b] \to X$ and $\gamma_2 : [a, b] \to Y$ to X and Y, and define the length of γ by the formula $\int_a^b \sqrt{|\gamma_1'|^2(t) + f^2(\gamma_1(t))|\gamma_2'|^2(t)}dt$.

The **warped product metric** is a metric on $X \times Y$, defined as the infimum of lengths of all rectifiable curves connecting two given points in $X \times Y$ (see [BBI01]).

4.3 Metrics on Other Sets

Given a metric space (X, d), one can construct several distances between some subsets of X. The main such distances are: the **point-set distance** $d(x, A) = \inf_{y \in A} d(x, y)$ between a point $x \in X$ and a subset $A \subset X$, the **set-set distance** $\inf_{x \in A, y \in B} d(x, y)$ between two subsets A and B of X, and the **Hausdorff metric**

[redacted]

between compact subsets of X which are considered in Chap. 1. In this section we list some other distances of this kind.

- **Line-line distance**

  The **line-line distance** (or *vertical distance between lines*) is the **set-set distance** in $\mathbb{R}^3$ between two *skew* lines, i.e., two straight lines that do not lie in a plane.

  It is the length of the segment of their common perpendicular whose endpoints lie on the lines. For l_1 and l_2 with equations $l_1\colon x = p+qt$, $t \in \mathbb{R}$, and $l_2\colon x = r+st$, $t \in \mathbb{R}$, the distance is given by

  $$\frac{|\langle r-p, q\times s\rangle|}{||q\times s||_2},$$

  where $\times$ is the *cross product* on $\mathbb{R}^3$, $\langle,\rangle$ is the *inner product* on $\mathbb{R}^3$, and $||.||_2$ is the Euclidean norm. For $x = (q_1, q_2, q_3)$, $s = (s_1, s_2, s_3)$, one has $q \times s = (q_2s_3 - q_3s_2, q_3s_1 - q_1s_3, q_1s_2 - q_2s_1)$.
- **Point-line distance**

  The **point-line distance** is the **point-set distance** between a point and a line.

  In $\mathbb{R}^2$, the distance between a point $P = (x_1, y_1)$ and a line l: $ax + by + c = 0$ (in Cartesian coordinates) is the **perpendicular distance** given by

  $$\frac{|ax_1 + by_1 + c|}{\sqrt{a^2+b^2}}.$$

  In $\mathbb{R}^2$, the *directed distance* between a point P and a line l is given by

  $$\frac{ax_1 + by_1 + c}{\pm\sqrt{a^2+b^2}},$$

  where the denominator is given the sign of b. It is negative if P is below the line.

  In $\mathbb{R}^3$, the distance between a point P and a line l: $x = p + qt$, $t \in \mathbb{R}$ (in vector formulation) is given by

  $$\frac{||q\times(p-P)||_2}{||q||_2},$$

  where $\times$ is the *cross product* on $\mathbb{R}^3$, and $||.||_2$ is the Euclidean norm.
- **Point-plane distance**

  The **point-plane distance** is the **point-set distance** in $\mathbb{R}^3$ between a point $P = (x_1, y_1, z_1)$ and a plane α: $ax + by + cz + d = 0$ given by

  $$\frac{|ax_1 + by_1 + cz_1 + d|}{\sqrt{a^2+b^2+c^2}}.$$

- **Algebraic point-conic distance**

  The **algebraic distance** of a point y to the curve given by the quadratic equation x^TAx (such as a conic in homogeneous coordinates) is defined as y^TAy.
- **Prime number distance**

  The **prime number distance** is the **point-set distance** in $(\mathbb{N}, |n-m|)$ between a number $n \in \mathbb{N}$ and the set of prime numbers $P \subset \mathbb{N}$. It is the absolute difference between n and the nearest prime number.
- **Distance up to nearest integer**

  The **distance up to nearest integer** is the **point-set distance** in $(\mathbb{R}, |x-y|)$ between a number $x \in \mathbb{R}$ and the set of integers $\mathbb{Z} \subset \mathbb{R}$, i.e., $\min_{n\in\mathbb{Z}} |x-n|$.
- **Busemann metric of sets**

  Given a metric space (X, d), the **Busemann metric of sets** (see [Buse55]) is a metric on the set of all nonempty closed subsets of X defined by

  $$\sup_{x\in X} |d(x,A) - d(x,B)| e^{-d(p,x)},$$

  where $p \in X$ is fixed, and $d(x,A) = \min_{y\in A} d(x,y)$ is the **point-set distance**.

  Instead of the weighting factor $e^{-d(p,x)}$, one can take any distance transform function which decreases fast enough (cf. L_p**-Hausdorff distance** in Chap. 1, and the list of variations of the **Hausdorff metric** in Chap. 21).
- **Quotient semimetric**

  Given an **extended metric space** (X, d) (i.e., a possibly infinite metric) and an equivalence relation $\sim$ on X, the **quotient semimetric** is a semimetric on the set $\overline{X} = X/\sim$ of equivalence classes defined, for any $\overline{x}, \overline{y} \in \overline{X}$, by

  $$\overline{d}(\overline{x}, \overline{y}) = \inf_{m\in\mathbb{N}} \sum_{i=1}^{m} d(x_i, y_i),$$

  where the infimum is taken over all sequences $x_1, y_1, x_2, y_2, \dots, x_m, y_m$ with $x_1 \in \overline{x}$, $y_m \in \overline{y}$, and $y_i \sim x_{i+1}$ for $i = 1, 2, \dots, m-1$. One has $\overline{d}(\overline{x}, \overline{y}) \le d(x,y)$ for all $x, y \in X$, and $\overline{d}$ is the biggest semimetric on $\overline{X}$ with this property.

[redacted]

Chapter 5
Metrics on Normed Structures

In this chapter we consider a special class of metrics defined on some *normed structures*, as the norm of the difference between two given elements. This structure can be a group (with a *group norm*), a vector space (with a *vector norm* or, simply, a *norm*), a vector lattice (with a *Riesz norm*), a field (with a *valuation*), etc.

Any norm is *subadditive*, i.e., triangle inequality $||x + y|| \leq ||x|| + ||y||$ holds. A norm is *submultiplicative* if **multiplicative triangle inequality** $||xy|| \leq ||x||||y||$ holds.

- **Group norm metric**

  A **group norm metric** is a metric on a *group* $(G, +, 0)$ defined by

  $$||x + (-y)|| = ||x - y||,$$

  where $||.||$ is a *group norm* on G, i.e., a function $||.|| : G \to \mathbb{R}$ such that, for all $x, y \in G$, we have the following properties:

  1. $||x|| \geq 0$, with $||x|| = 0$ if and only if $x = 0$;
  2. $||x|| = || - x||$;
  3. $||x + y|| \leq ||x|| + ||y||$ (triangle inequality).

  Any group norm metric d is **right-invariant**, i.e., $d(x, y) = d(x + z, y + z)$ for any $x, y, z \in G$. Conversely, any right-invariant (as well as any **left-invariant**, and, in particular, any **bi-invariant**) metric d on G is a group norm metric, since one can define a group norm on G by $||x|| = d(x, 0)$.

- ***F*-norm metric**

  A *vector space* (or *linear space*) over a *field* $\mathbb{F}$ is a set V equipped with operations of *vector addition* $+ : V \times V \to V$ and *scalar multiplication* $\cdot : \mathbb{F} \times V \to V$ such that $(V, +, 0)$ forms an *Abelian group* (where $0 \in V$ is the *zero vector*), and, for all *vectors* $x, y \in V$ and any *scalars* $a, b \in \mathbb{F}$, we have the following properties: $1 \cdot x = x$ (where 1 is the multiplicative unit of $\mathbb{F}$), $(ab) \cdot x = a \cdot (b \cdot x)$, $(a + b) \cdot x = a \cdot x + b \cdot x$, and $a \cdot (x + y) = a \cdot x + a \cdot y$.

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_5

A vector space over the field $\mathbb{R}$ of real numbers is called a *real vector space*. A vector space over the field $\mathbb{C}$ of complex numbers is called *complex vector space*.

A **F-norm metric** is a metric on a real (complex) vector space V defined by

$$||x-y||_F,$$

where $||.||_F$ is an *F-norm* on V, i.e., a function $||.||_F : V \to \mathbb{R}$ such that, for all $x, y \in V$ and for any scalar a with $|a| = 1$, we have the following properties:

1. $||x||_F \geq 0$, with $||x||_F = 0$ if and only if $x = 0$;
2. $||ax||_F \leq ||x||_F$ if $|a| \leq 1$;
3. $\lim_{a \to 0} ||ax||_F = 0$;
4. $||x+y||_F \leq ||x||_F + ||y||_F$ (triangle inequality).

An F-norm is called *p-homogeneous* if $||ax||_F = |a|^p ||x||_F$ for any scalar a.

Any F-norm metric d is a **translation invariant metric**, i.e., $d(x, y) = d(x + z, y + z)$ for all $x, y, z \in V$. Conversely, if d is a translation invariant metric on V, then $||x||_F = d(x, 0)$ is an F-norm on V.

- **F^*-metric**

  An **F^*-metric** is an **F-norm metric** $||x-y||_F$ on a real (complex) vector space V such that the operations of scalar multiplication and vector addition are continuous with respect to $||.||_F$. Thus $||.||_F$ is a function $||.||_F : V \to \mathbb{R}$ such that, for all $x, y, x_n \in V$ and for all scalars a, a_n, we have the following properties:

  1. $||x||_F \geq 0$, with $||x||_F = 0$ if and only if $x = 0$;
  2. $||ax||_F = ||x||_F$ for all a with $|a| = 1$;
  3. $||x+y||_F \leq ||x||_F + ||y||_F$;
  4. $||a_n x||_F \to 0$ if $a_n \to 0$;
  5. $||a x_n||_F \to 0$ if $x_n \to 0$;
  6. $||a_n x_n||_F \to 0$ if $a_n \to 0$, $x_n \to 0$.

  The metric space $(V, ||x-y||_F)$ with an F^*-metric is called a n**F^*-space**. Equivalently, an F^*-space is a metric space (V, d) with a **translation invariant metric** d such that the operation of scalar multiplication and vector addition are continuous with respect to this metric.

  A **complete** F^*-space is called an **F-space**. A **locally convex** F-space is known as a **Fréchet space** (Chap. 2) in Functional Analysis.

  A **modular space** is an F^*-space $(V, ||.||_F)$ in which the F-norm $||.||_F$ is defined by

$$||x||_F = \inf\{\lambda > 0 : \rho\left(\frac{x}{\lambda}\right) < \lambda\},$$

  and ρ is a *metrizing modular* on V, i.e., a function $\rho : V \to [0, \infty]$ such that, for all $x, y, x_n \in V$ and for all scalars a, a_n, we have the following properties:

  1. $\rho(x) = 0$ if and only if $x = 0$;
  2. $\rho(ax) = \rho(x)$ implies $|a| = 1$;

[redacted]

3. $\rho(ax + by) \leq \rho(x) + \rho(y)$ implies $a, b \geq 0, a + b = 1$;
4. $\rho(a_n x) \to 0$ if $a_n \to 0$ and $\rho(x) < \infty$;
5. $\rho(ax_n) \to 0$ if $\rho(x_n) \to 0$ (*metrizing property*);
6. For any $x \in V$, there exists $k > 0$ such that $\rho(kx) < \infty$.

- **Norm metric**

  A **norm metric** is a metric on a real (complex) vector space V defined by

$$||x - y||,$$

where $||.||$ is a *norm* on V, i.e., a function $||.|| : V \to \mathbb{R}$ such that, for all $x, y \in V$ and for any scalar a, we have the following properties:

1. $||x|| \geq 0$, with $||x|| = 0$ if and only if $x = 0$;
2. $||ax|| = |a|||x||$;
3. $||x + y|| \leq ||x|| + ||y||$ (triangle inequality).

Therefore, a norm $||.||$ is a 1-*homogeneous F-norm*. The vector space $(V, ||.||)$ is called a *normed vector space* or, simply, *normed space*.

Any metric space can be embedded isometrically in some normed vector space as a closed linearly independent subset. Every finite-dimensional normed space is **complete**, and all norms on it are equivalent.

In general, the norm $||.||$ is equivalent (Maligranda, 2008) to the norm

$$||x||_{u,p} = (||x + ||x|| \cdot u||^p + ||x - ||x|| \cdot u||^p)^{\frac{1}{p}},$$

introduced, for any $u \in V$ and $p \geq 1$, by Odell and Schlumprecht, 1998.

The **norm-angular distance** between x and y is defined (Clarkson, 1936) by

$$d(x, y) = ||\frac{x}{||x||} - \frac{y}{||y||}||.$$

The following sharpening of the triangle inequality (Maligranda, 2003) holds:

$$\frac{||x - y|| - |||x|| - ||y|||}{\min\{||x||, ||y||\}} \leq d(x, y) \leq \frac{||x - y|| + |||x|| - ||y|||}{\max\{||x||, ||y||\}}, \text{ i.e.,}$$

$$(2 - d(x, -y)) \min\{||x||, ||y||\} \leq ||x|| + ||y|| - ||x + y||$$
$$\leq (2 - d(x, -y)) \max\{||x||, ||y||\}.$$

Dragomir, 2004, call $|\int_a^b f(x)dx| \leq \int_a^b |f(x)|dx$ *continuous triangle inequality*.

- **Reverse triangle inequality**

  The triangle inequality $||x+y|| \leq ||x|| + ||y||$ in a normed space $(V, ||.||)$ is equivalent to the following inequality, for any $x_1, \dots, x_n \in V$ with $n \geq 2$:

$$||\sum_{i=1}^{n} x_i|| \leq \sum_{i=1}^{n} ||x_i||.$$

  If in the normed space $(V, ||.||)$, for some $C \geq 1$ one has

$$C||\sum_{i=1}^{n} x_i|| \geq \sum_{i=1}^{n} ||x_i||,$$

  then this inequality is called the **reverse triangle inequality**.

  This term is used, sometimes, also for the **inverse triangle inequality** (cf. **kinematic metric** in Chap. 26) and for the eventual inequality $Cd(x,z) \geq d(x,y) + d(y,z)$ with $C \geq 1$ in a metric space (X, d).

  The triangle inequality $||x+y|| \leq ||x|| + ||y||$, for any $x, y \in V$, in a normed space $(V, ||.||)$ is, for any number $q > 1$, equivalent (Belbachir, Mirzavaziri and Moslenian, 2005) to the following inequality:

$$||x+y||^q \leq 2^{q-1}(||x||^q + ||y||^q).$$

  The *parallelogram inequality* $||x+y||^2 \leq 2(||x||^2 + ||y||^2)$ is the case $q = 2$ of above.

  Given a number q, $0 < q \leq 1$, the norm is called *q-subadditive* if $||x+y||^q \leq ||x||^q + ||y||^q$ holds for $x, y \in V$.
- **Seminorm semimetric**

  A **seminorm semimetric** on a real (complex) vector space V is defined by

$$||x-y||,$$

  where $||.||$ is a *seminorm* (or *pseudo-norm*) on V, i.e., a function $||.|| : V \rightarrow \mathbb{R}$ such that, for all $x, y \in V$ and for any scalar a, we have the following properties:

  1. $||x|| \geq 0$, with $||0|| = 0$;
  2. $||ax|| = |a|||x||$;
  3. $||x+y|| \leq ||x|| + ||y||$ (triangle inequality).

  The vector space $(V, ||.||)$ is called a *seminormed vector space*. Many *normed vector spaces*, in particular, **Banach spaces**, are defined as the quotient space by the subspace of elements of seminorm zero.

  A *quasi-normed space* is a vector space V, on which a *quasi-norm* is given. A *quasi-norm* on V is a nonnegative function $||.|| : V \rightarrow \mathbb{R}$ which satisfies the same axioms as a norm, except for the triangle inequality which is replaced by the weaker requirement: there exists a constant $C > 0$ such that, for all $x, y \in V$,

[redacted]

the following ***C*-triangle inequality** (cf. **near-metric** in Chap. 1) holds:

$$||x+y|| \leq C(||x||+||y||)$$

An example of a quasi-normed space, that is not normed, is the *Lebesgue space* $L_p(\Omega)$ with $0 < p < 1$ in which a quasi-norm is defined by

$$||f|| = (\int_{\Omega} |f(x)|^p dx)^{1/p}, f \in L_p(\Omega).$$

- **Banach space**

  A **Banach space** (or *B-space*) is a **complete** metric space $(V, ||x-y||)$ on a vector space V with a norm metric $||x-y||$. Equivalently, it is the complete *normed space* $(V, ||.||)$. In this case, the norm $||.||$ on V is called the *Banach norm*. Some examples of Banach spaces are:

  1. l_p^n*-spaces*, l_p^∞*-spaces*, $1 \leq p \leq \infty$, $n \in \mathbb{N}$;
  2. The space C of convergent numerical sequences with the norm $||x|| = \sup_n |x_n|$;
  3. The space C_0 of numerical sequences which converge to zero with the norm $||x|| = \max_n |x_n|$;
  4. The space $C^p_{[a,b]}$, $1 \leq p \leq \infty$, of continuous functions on $[a,b]$ with the L_p*-norm* $||f||_p = (\int_a^b |f(t)|^p dt)^{\frac{1}{p}}$;
  5. The space C_K of continuous functions on a compactum K with the norm $||f|| = \max_{t \in K} |f(t)|$;
  6. The space $(C_{[a,b]})^n$ of functions on $[a,b]$ with continuous derivatives up to and including the order n with the norm $||f||_n = \sum_{k=0}^n \max_{a \leq t \leq b} |f^{(k)}(t)|$;
  7. The space $C^n[I^m]$ of all functions defined in an m-dimensional cube that are continuously differentiable up to and including the order n with the norm of uniform boundedness in all derivatives of order at most n;
  8. The space $M_{[a,b]}$ of bounded measurable functions on $[a,b]$ with the norm

  $$||f|| = ess \sup_{a \leq t \leq b} |f(t)| = \inf_{e, \mu(e)=0} \sup_{t \in [a,b] \setminus e} |f(t)|;$$

  9. The space $A(\Delta)$ of functions analytic in the open *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ and continuous in the closed disk $\overline{\Delta}$ with the norm $||f|| = \max_{z \in \overline{\Delta}} |f(z)|$;
  10. The **Lebesgue spaces** $L_p(\Omega)$, $1 \leq p \leq \infty$;
  11. The *Sobolev spaces* $W^{k,p}(\Omega)$, $\Omega \subset \mathbb{R}^n$, $1 \leq p \leq \infty$, of functions f on Ω such that f and its derivatives, up to some order k, have a finite L_p*-norm*, with the norm $||f||_{k,p} = \sum_{i=0}^k ||f^{(i)}||_p$;

12. The *Bohr space AP* of almost periodic functions with the norm

$$||f|| = \sup_{-\infty<t<+\infty} |f(t)|.$$

A finite-dimensional real Banach space is called a *Minkowskian space*. A norm metric of a Minkowskian space is called a **Minkowskian metric** (Chap. 6). In particular, any l_p**-metric** is a Minkowskian metric.

All n-dimensional Banach spaces are pairwise isomorphic; the set of such spaces becomes compact if one introduces the **Banach–Mazur distance** by $d_{BM}(V, W) = \ln \inf_T ||T|| \cdot ||T^{-1}||$, where the infimum is taken over all operators which realize an isomorphism $T : V \to W$.

- l_p**-metric**

The l_p**-metric** d_{l_p}, $1 \le p \le \infty$, is a norm metric on $\mathbb{R}^n$ (or on $\mathbb{C}^n$), defined by

$$||x - y||_p,$$

where the l_p*-norm* $||.||_p$ is defined by

$$||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}.$$

For $p = \infty$, we obtain $||x||_\infty = \lim_{p\to\infty} \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} = \max_{1\le i\le n} |x_i|$. The metric space $(\mathbb{R}^n, d_{l_p})$ is abbreviated as l_p^n and is called l_p^n*-space*.

The l_p**-metric**, $1 \le p \le \infty$, on the set of all sequences $x = \{x_n\}_{n=1}^{\infty}$ of real (complex) numbers, for which the sum $\sum_{i=1}^{\infty} |x_i|^p$ (for $p = \infty$, the sum $\sum_{i=1}^{\infty} |x_i|$) is finite, is

$$(\sum_{i=1}^{\infty} |x_i - y_i|^p)^{\frac{1}{p}}.$$

For $p = \infty$, we obtain $\max_{i\ge 1} |x_i - y_i|$. This metric space is abbreviated as l_p^∞ and is called l_p^∞*-space*.

Most important are l_1-, l_2- and l_∞-metrics. Among l_p-metrics, only l_1- and l_∞-metrics are **crystalline metrics**, i.e., metrics having polygonal unit balls. On $\mathbb{R}$ all l_p-metrics coincide with the **natural metric** (Chap. 12) $|x - y|$.

The l_2-norm $||(x_1, x_2)||_2 = \sqrt{x_1^2 + x_2^2}$ on $\mathbb{R}^2$ is also called *Pythagorean addition* of the numbers x_1 and x_2. Under this commutative operation, $\mathbb{R}$ form a semigroup, and $\mathbb{R}_{\ge 0}$ form a *monoid* (semigroup with identity, 0).

[redacted]

- **Euclidean metric**

  The **Euclidean metric** (or **Pythagorean distance**, **as-the-crow-flies distance**, **beeline distance**) d_E is the metric on $\mathbb{R}^n$ defined by

$$||x-y||_2 = \sqrt{(x_1-y_1)^2 + \cdots + (x_n-y_n)^2}.$$

  It is the ordinary l_2**-metric** on $\mathbb{R}^n$. The metric space $(\mathbb{R}^n, d_E)$ is abbreviated as $\mathbb{E}^n$ and is called **Euclidean space** "Euclidean space" stands for the case $n = 3$, as opposed, for $n = 2$, to *Euclidean plane* and, for $n = 1$, *Euclidean* (or *real*) *line*.

  In fact, $\mathbb{E}^n$ is an **inner product space** (and even a **Hilbert space**), i.e., $d_E(x,y) = ||x-y||_2 = \sqrt{\langle x-y, x-y\rangle}$, where $\langle x,y\rangle$ is the *inner product* on $\mathbb{R}^n$ which is given in the Cartesian coordinate system by $\langle x,y\rangle = \sum_{i=1}^{n} x_i y_i$. In a standard coordinate system one has $\langle x,y\rangle = \sum_{i,j} g_{ij}x_i y_j$, where $g_{ij} = \langle e_i, e_j\rangle$, and the **metric tensor** $((g_{ij}))$ (cf. Chap. 7) is a positive-definite symmetric $n \times n$ matrix.

  In general, a Euclidean space is defined as a space, the properties of which are described by the axioms of *Euclidean Geometry*.

- **Norm transform metric**

  A **norm transform metric** is a metric $d(x,y)$ on a vector space $(V, ||.||)$, which is a function of $||x||$ and $||y|$. Usually, $V = \mathbb{R}^n$ and, moreover, $\mathbb{E}^n = (\mathbb{R}^n, ||.||_2)$.

  Some examples are (p,q)**-relative metric**, M**-relative metric** and, from Chap. 19, the **British Rail metric** $||x|| + ||y||$ for $x \neq y$, (and equal to 0, otherwise), the **radar screen metric** $\min\{1, ||x-y||\}$ and $\max\{1, ||x-y||\}$ for $x \neq y$. Cf. t**-truncated** and t**-uniformly discrete** metrics in Chap. 4.

- (p,q)**-relative distance**

  Let $(V, ||.||) \neq \emptyset, \{0\}$ be a *Ptolemaic space*, i.e., the norm metric $||x-y||$ is a **Ptolemaic metric** (Chap. 1). Let $p, q > 0$.

  The (p,q)**-relative distance** on $(V, ||.||)$ is defined, for x or $y \neq 0$, by

$$\rho_{p,q}(x,y) + \frac{||x-y||}{(\frac{1}{2}(||x||^p + ||y||^p))^{\frac{q}{p}}}$$

  (and equal to 0, otherwise). In the case of $p = \infty$, it has the form

$$\frac{||x-y||}{(\max\{||x||, ||y||\})^q}.$$

  This distance is a metric (Hästö, 2002) if and only if $0 < q \leq 1, p \geq \max\{1-q, \frac{2-q}{3}\}$.

  $(p,1)$-, $(\infty,1)$- and the original $(1,1)$-relative metric on $\mathbb{E}^n$ are called p**-relative** (or **Klamkin–Meir metric**), **relative metric** and **Schattschneider metric**.

- ***M*-relative distance**

  Let $(V, ||.||) \neq \emptyset, \{0\}$ be a *Ptolemaic space*, i.e., $||x - y||$ is a **Ptolemaic metric**. Let $M : [0, \infty) \to (0, \infty)$ be a symmetric function.

  The ***M*-relative distance** on $(V, ||.||)$ is defined by

  $$\rho_M(x, y) = \frac{||x - y||}{M(||x||, ||y||)}.$$

  So, it is the **(p, q)-relative distance** if $M(x, y) = (x^p + y^p)^{\frac{q}{p}}$ for $p, q > 0$.

  Call a function $f : [0, \infty) \to (0, \infty)$ *moderately increasing* (MI) if $f(x)$ is increasing but $\frac{f(x)}{x}$ is decreasing for $x > 0$. Hästö, 2002, showed that

  (i) If $M = f(x)f(y)$, then $\rho_M(x, y)$ is a metric if and only if f is MI and convex;
  (ii) If both, $M(x, \cdot)$ and $M(\cdot, x)$, are MI for each fixed $x > 0$, then $\rho_M(x, y)$ is a metric if and only if it is a metric on $\mathbb{R}$; it holds for any Ptolemeaic metric.

- **Unitary metric**

  The **unitary** (or *complex Euclidean*) **metric** is the **l_2-metric** on $\mathbb{C}^n$ defined by

  $$||x - y||_2 = \sqrt{|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2}.$$

  For $n = 1$, it is the **complex modulus metric** $|x - y| = \sqrt{(x - y)\overline{(x - y)}}$ on the *Wessel–Argand plane* (Chap. 12).

- **L_p-metric**

  An **L_p-metric** d_{L_p}, $1 \leq p \leq \infty$, is a norm metric on $L_p(\Omega, \mathcal{A}, \mu)$ defined by

  $$||f - g||_p$$

  for any $f, g \in L_p(\Omega, \mathcal{A}, \mu)$. The metric space $(L_p(\Omega, \mathcal{A}, \mu), d_{L_p})$ is called the **L_p-space** (or **Lebesgue space**).

  Here Ω is a set, and $\mathcal{A}$ is n *σ-algebra* of subsets of Ω, i.e., a collection of subsets of Ω satisfying the following properties:

  1. $\Omega \in \mathcal{A}$;
  2. If $A \in \mathcal{A}$, then $\Omega \backslash A \in \mathcal{A}$;
  3. If $A = \cup_{i=1}^{\infty} A_i$ with $A_i \in \mathcal{A}$, then $A \in \mathcal{A}$.

  A function $\mu : \mathcal{A} \to \mathbb{R}_{\geq 0}$ is called a *measure* on $\mathcal{A}$ if it is *additive*, i.e., $\mu(\cup_{i \geq 1} A_i) = \sum_{i \geq 1} \mu(A_i)$ for all pairwise disjoint sets $A_i \in \mathcal{A}$, and satisfies $\mu(\emptyset) = 0$. A *measure space* is a triple $(\Omega, \mathcal{A}, \mu)$.

  Given a function $f : \Omega \to \mathbb{R}(\mathbb{C})$, its *$L_p$-norm* is defined by

  $$||f||_p = \left(\int_{\Omega} |f(\omega)|^p \mu(d\omega) \right)^{\frac{1}{p}}.$$

[redacted]

Let $L_p(\Omega, \mathcal{A}, \mu) = L_p(\Omega)$ denote the set of all functions $f : \Omega \to \mathbb{R}$ ($\mathbb{C}$) such that $||f||_p < \infty$. Strictly speaking, $L_p(\Omega, \mathcal{A}, \mu)$ consists of equivalence classes of functions, where two functions are *equivalent* if they are equal *almost everywhere*, i.e., the set on which they differ has measure zero. The set $L_\infty(\Omega, \mathcal{A}, \mu)$ is the set of equivalence classes of measurable functions $f : \Omega \to \mathbb{R}$ ($\mathbb{C}$) whose absolute values are bounded almost everywhere.

The most classical example of an L_p-metric is d_{L_p} on the set $L_p(\Omega, \mathcal{A}, \mu)$, where Ω is the open interval $(0, 1)$, $\mathcal{A}$ is the *Borel σ-algebra* on $(0, 1)$, and μ is the *Lebesgue measure*. This metric space is abbreviated by $L_p(0, 1)$ and is called *$L_p(0, 1)$-space*.

In the same way, one can define the L_p-metric on the set $C_{[a,b]}$ of all real (complex) continuous functions on $[a, b]$: $d_{L_p}(f, g) = (\int_a^b |f(x) - g(x)|^p dx)^{\frac{1}{p}}$. For $p = \infty$, $d_{L_\infty}(f, g) = \max_{a \le x \le b} |f(x) - g(x)|$. This metric space is abbreviated by $C^p_{[a,b]}$ and is called *$C^p_{[a,b]}$-space*.

If $\Omega = \mathbb{N}$, $\mathcal{A} = 2^\Omega$ is the collection of all subsets of Ω, and μ is the *cardinality measure* (i.e., $\mu(A) = |A|$ if A is a finite subset of Ω, and $\mu(A) = \infty$, otherwise), then the metric space $(L_p(\Omega, 2^\Omega, |.|), d_{L_p})$ coincides with the space l_p^∞.

If $\Omega = V_n$ is a set of cardinality n, $\mathcal{A} = 2^{V_n}$, and μ is the cardinality measure, then the metric space $(L_p(V_n, 2^{V_n}, |.|), d_{L_p})$ coincides with the space l_p^n.

- **Dual metrics**

  The **l_p-metric** and the **l_q-metric**, $1 < p, q < \infty$, are called **dual** if $1/p + 1/q = 1$.

  In general, when dealing with a *normed vector space* $(V, ||.||_V)$, one is interested in the *continuous* linear functionals from V into the base field ($\mathbb{R}$ or $\mathbb{C}$). These functionals form a **Banach space** $(V', ||.||_{V'})$, called the *continuous dual* of V. The norm $||.||_{V'}$ on V' is defined by $||T||_{V'} = \sup_{||x||_V \le 1} |T(x)|$.

  The continuous dual for the metric space l_p^n (l_p^∞) is l_q^n (l_q^∞, respectively). The continuous dual of l_1^n (l_1^∞) is l_∞^n (l_∞^∞, respectively). The continuous duals of the Banach spaces C (consisting of all convergent sequences, with **l_∞-metric**) and C_0 (consisting of the sequences converging to zero, with **l_∞-metric**) are both naturally identified with l_1^∞.

- **Inner product space**

  An **inner product space** (or *pre-Hilbert space*) is a metric space $(V, ||x - y||)$ on a real (complex) vector space V with an *inner product* $\langle x, y \rangle$ such that the norm metric $||x - y||$ is constructed using the *inner product norm* $||x|| = \sqrt{\langle x, x \rangle}$.

  An *inner product* $\langle, \rangle$ on a real (complex) vector space V is a *symmetric bilinear* (in the complex case, *sesquilinear*) form on V, i.e., a function $\langle, \rangle : V \times V \longrightarrow \mathbb{R}$ ($\mathbb{C}$) such that, for all $x, y, z \in V$ and for all scalars α, β, we have the following properties:

1. $\langle x, x \rangle \ge 0$, with $\langle x, x \rangle = 0$ if and only if $x = 0$;
2. $\langle x, y \rangle = \overline{\langle y, x \rangle}$, where the bar denotes *complex conjugation*;
3. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$.

For a complex vector space, an inner product is called also a *Hermitian inner product*, and the corresponding metric space is called a *Hermitian inner product space*.

A norm $||.||$ in a *normed space* $(V, ||.||)$ is generated by an inner product if and only if, for all $x, y \in V$, we have: $||x + y||^2 + ||x - y||^2 = 2(||x||^2 + ||y||^2)$.

In an inner product space, the **triangle equality** (Chap. 1) $||x - y|| = ||x|| + ||y||$, for $x, y \neq 0$, holds if and only if $\frac{x}{||x||} = \frac{y}{||y||}$, i.e., $x - y \in [x, y]$.

- **Hilbert space**

  A **Hilbert space** is an **inner product space** which, as a metric space, is **complete**. More precisely, a Hilbert space is a complete metric space $(H, ||x-y||)$ on a real (complex) vector space H with an *inner product* $\langle , \rangle$ such that the norm metric $||x - y||$ is constructed using the *inner product norm* $||x|| = \sqrt{\langle x, x \rangle}$. Any Hilbert space is a **Banach space**.

  An example of a Hilbert space is the set of all sequences $x = \{x_n\}_n$ of real (complex) numbers such that $\sum_{i=1}^{\infty} |x_i|^2$ converges, with the **Hilbert metric** defined by

  $$(\sum_{i=1}^{\infty}(x_i - y_i)^2)^{\frac{1}{2}}.$$

  Other examples of Hilbert spaces are any L_2**-space**, and any finite-dimensional inner product space. In particular, any Euclidean space is a Hilbert space.

  A direct product of two **Hilbert spaces** is called a *Liouville space* (or *line space*, *extended Hilbert space*).

  Given an infinite cardinal number τ and a set A of the cardinality τ, let $\mathbb{R}_a$, $a \in A$, be the copies of $\mathbb{R}$. Let $H(A) = \{\{x_a\} \in \prod_{a \in A} \mathbb{R}_a : \sum_a x_a^2 < \infty\}$; then $H(A)$ with the metric defined for $\{x_a\}, \{y_a\} \in H(A)$ as

  $$(\sum_{a \in A}(x_a - y_a)^2)^{\frac{1}{2}},$$

  is called the **generalized Hilbert space** of weight τ.

- **Erdös space**

  The **Erdös space** (or *rational Hilbert space*) is the metric subspace of l_2 consisting of all vectors in l_2 with only rational coordinates. It has topological dimension 1 and is not complete. Erdös space is **homeomorphic** to its countable infinite power, and every nonempty open subset of it is homeomorphic to whole space.

  The **complete Erdös space** (or *irrational Hilbert space*) is the complete metric subspace of l_2 consisting of all vectors in l_2 the coordinates of which are all irrational.

[redacted]

- **Riesz norm metric**

  A *Riesz space* (or *vector lattice*) is a partially ordered vector space $(V_{Ri}, \preceq)$ in which the following conditions hold:

  1. The vector space structure and the partial order structure are compatible, i.e., from $x \preceq y$ it follows that $x + z \preceq y + z$, and from $x \succ 0$, $a \in \mathbb{R}$, $a > 0$ it follows that $ax \succ 0$;
  2. For any two elements $x, y \in V_{Ri}$, there exist the *join* $x \vee y \in V_{Ri}$ and *meet* $x \wedge y \in V_{Ri}$ (cf. Chap. 10).

  The **Riesz norm metric** is a norm metric on V_{Ri} defined by

  $$||x - y||_{Ri},$$

  where $||.||_{Ri}$ is a *Riesz norm* on V_{Ri}, i.e., a norm such that, for any $x, y \in V_{Ri}$, the inequality $|x| \preceq |y|$, where $|x| = (-x) \vee (x)$, implies $||x||_{Ri} \leq ||y||_{Ri}$.

  The space $(V_{Ri}, ||.||_{Ri})$ is called a *normed Riesz space*. In the case of completeness, it is called a *Banach lattice*.

- **Banach–Mazur compactum**

  The **Banach–Mazur distance** d_{BM} between two n-dimensional *normed spaces* $(V, ||.||_V)$ and $(W, ||.||_W)$ is defined by

  $$\ln \inf_T ||T|| \cdot ||T^{-1}||,$$

  where the infimum is taken over all isomorphisms $T : V \to W$. It is a metric on the set X^n of all equivalence classes of n-dimensional normed spaces, where $V \sim W$ if and only if they are *isometric*. Then the pair (X^n, d_{BM}) is a compact metric space which is called the **Banach–Mazur compactum**.

- **Quotient metric**

  Given a *normed space* $(V, ||.||_V)$ with a norm $||.||_V$ and a closed subspace W of V, let $(V/W, ||.||_{V/W})$ be the normed space of cosets $x+W = \{x+w : w \in W\}$, $x \in V$, with the *quotient norm* $||x + W||_{V/W} = \inf_{w \in W} ||x + w||_V$.

  The **quotient metric** is a norm metric on V/W defined by

  $$||(x + W) - (y + W)||_{V/W}.$$

- **Tensor norm metric**

  Given *normed spaces* $(V, ||.||_V)$ and $(W, ||.||_W)$, a norm $||.||_\otimes$ on the *tensor product* $V \otimes W$ is called *tensor norm* (or *cross norm*) if $||x \otimes y||_\otimes = ||x||_V ||y||_W$ for all *decomposable* tensors $x \otimes y$.

  The **tensor product metric** is a norm metric on $V \otimes W$ defined by

  $$||z - t||_\otimes.$$

For any $z \in V \otimes W$, $z = \sum_j x_j \otimes y_j$, $x_j \in V$, $y_j \in W$, the *projective norm* (or *π-norm*) of z is defined by $||z||_{pr} = \inf \sum_j ||x_j||_V ||y_j||_W$, where the infimum is taken over all representations of z as a sum of decomposable vectors. It is the largest tensor norm on $V \otimes W$.

- **Valuation metric**

  A **valuation metric** is a metric on a *field* $\mathbb{F}$ defined by

  $$||x - y||,$$

  where $||.||$ is a *valuation* on $\mathbb{F}$, i.e., a function $||.|| : \mathbb{F} \to \mathbb{R}$ such that, for all $x, y \in \mathbb{F}$, we have the following properties:

  1. $||x|| \geq 0$, with $||x|| = 0$ if and only if $x = 0$;
  2. $||xy|| = ||x|| \, ||y||$,
  3. $||x + y|| \leq ||x|| + ||y||$ (triangle inequality).

  If $||x+y|| \leq \max\{||x||, ||y||\}$, the valuation $||.||$ is called *non-Archimedean*. In this case, the valuation metric is an **ultrametric**. The simplest valuation is the *trivial valuation* $||.||_{tr}$: $||0||_{tr} = 0$, and $||x||_{tr} = 1$ for $x \in \mathbb{F} \backslash \{0\}$. It is non-Archimedean.

  There are different definitions of valuation in Mathematics. Thus, the function $\nu : \mathbb{F} \to \mathbb{R} \cup \{\infty\}$ is called a *valuation* if $\nu(x) \geq 0$, $\nu(0) = \infty$, $\nu(xy) = \nu(x) + \nu(y)$, and $\nu(x + y) \geq \min\{\nu(x), \nu(y)\}$ for all $x, y \in \mathbb{F}$. The valuation $||.||$ can be obtained from the function ν by the formula $||x|| = \alpha^{\nu(x)}$ for some fixed $0 < \alpha < 1$ (cf. ***p*-adic metric** in Chap. 12).

  The *Kürschäk valuation* $|.|_{Krs}$ is a function $|.|_{Krs} : \mathbb{F} \to \mathbb{R}$ such that $|x|_{Krs} \geq 0$, $|x|_{Krs} = 0$ if and only if $x = 0$, $|xy|_{Krs} = |x|_{Krs}|y|_{Krs}$, and $|x + y|_{Krs} \leq C \max\{|x|_{Krs}, |y|_{Krs}\}$ for all $x, y \in \mathbb{F}$ and for some positive constant C, called the *constant of valuation*. If $C \leq 2$, one obtains the ordinary valuation $||.||$ which is non-Archimedean if $C \leq 1$. In general, any $|.|_{Krs}$ is *equivalent* to some $||.||$, i.e., $|.|^p_{Krs} = ||.||$ for some $p > 0$.

  Finally, given an *ordered group* $(G, \cdot, e, \leq)$ equipped with zero, the *Krull valuation* is a function $|.| : \mathbb{F} \to G$ such that $|x| = 0$ if and only if $x = 0$, $|xy| = |x||y|$, and $|x + y| \leq \max\{|x|, |y|\}$ for any $x, y \in \mathbb{F}$. It is a generalization of the definition of non-Archimedean valuation $||.||$ (cf. **generalized metric** in Chap. 3).

- **Power series metric**

  Let $\mathbb{F}$ be an arbitrary algebraic field, and let $\mathbb{F}\langle x^{-1}\rangle$ be the field of power series of the form $w = \alpha_{-m}x^m + \dots + \alpha_0 + \alpha_1 x^{-1} + \dots$, $\alpha_i \in \mathbb{F}$. Given $l > 1$, a *non-Archimedean valuation* $||.||$ on $\mathbb{F}\langle x^{-1}\rangle$ is defined by

  $$||w|| = \begin{cases} l^m, & \text{if } w \neq 0, \\ 0, & \text{if } w = 0. \end{cases}$$

  The **power series metric** is the **valuation metric** $||w - v||$ on $\mathbb{F}\langle x^{-1}\rangle$.

[redacted]

Part II
Geometry and Distances

Chapter 6
Distances in Geometry

Geometry arose as the field of knowledge dealing with spatial relationships. It was one of the two fields of pre-modern Mathematics, the other being the study of numbers.

Earliest known evidence of abstract representation—ochre rocks marked with cross hatches and lines to create a consistent complex geometric motif, dated about 75,000 BC—were found in Blombos Cave, South Africa. In modern times, geometric concepts have been generalized to a high level of abstraction and complexity.

6.1 Geodesic Geometry

In Mathematics, the notion of “geodesic” is a generalization of the notion of “straight line” to curved spaces. This term is taken from *Geodesy*, the science of measuring the size and shape of the Earth.

Given a metric space (X, d), a **metric curve** γ is a continuous function $\gamma : I \to X$, where I is an *interval* (i.e., nonempty connected subset) of $\mathbb{R}$. If γ is r times continuously differentiable, it is called a *regular curve* of class C^r; if $r = \infty$, γ is called a *smooth curve*.

In general, a curve may cross itself. A curve is called a *simple curve* (or *arc*, path) if it does not cross itself, i.e., if it is injective. A curve $\gamma : [a, b] \to X$ is called a *Jordan curve* (or *simple closed curve*) if it does not cross itself, and $\gamma(a) = \gamma(b)$.

The *length* (which may be equal to ∞) $l(\gamma)$ of a curve $\gamma : [a, b] \to X$ is defined by $\sup \sum_{i=1}^{n} d(\gamma(t_{i-1}), \gamma(t_i))$, where the supremum is taken over all finite decompositions $a = t_0 < t_1 < \ldots < t_n = b$, $n \in \mathbb{N}$, of $[a, b]$.

A curve with finite length is called *rectifiable*. For each regular curve $\gamma : [a, b] \to X$ define the *natural parameter* s of γ by $s = s(t) = l(\gamma|_{[a,t]})$, where $l(\gamma|_{[a,t]})$ is the length of the part of γ corresponding to the interval $[a, t]$. A curve with this *natural*

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_6

[redacted]

parametrization $\gamma = \gamma(s)$ is called **of unit speed**, (or *parametrized by arc length*, *normalized*); in this parametrization, for any $t_1, t_2 \in I$, one has $l(\gamma|_{[t_1,t_2]}) = |t_2 - t_1|$, and $l(\gamma) = |b - a|$.

The length of any curve $\gamma : [a, b] \to X$ is at least the distance between its endpoints: $l(\gamma) \geq d(\gamma(a), \gamma(b))$. The curve γ, for which $l(\gamma) = d(\gamma(a), \gamma(b))$, is called the **geodesic segment** (or *shortest* path) from $x = \gamma(a)$ to $y = \gamma(b)$, and denoted by $[x, y]$.

Thus, a geodesic segment is a shortest join of its endpoints; it is an isometric embedding of $[a, b]$ in X. In general, geodesic segments need not exist, unless the segment consists of one point only. A geodesic segment joining two points need not be unique.

A **geodesic** (Chap. 1) is a curve which extends indefinitely in both directions and behaves locally like a segment, i.e., is everywhere locally a distance minimizer.

More exactly, a curve $\gamma : \mathbb{R} \to X$, given in the natural parametrization, is called a *geodesic* if, for any $t \in \mathbb{R}$, there exists a *neighborhood* U of t such that, for any $t_1, t_2 \in U$, we have $d(\gamma(t_1), \gamma(t_2)) = |t_1 - t_2|$. Thus, any geodesic is a locally isometric embedding of the whole of $\mathbb{R}$ in X.

A geodesic is called a **metric straight line** if the equality $d(\gamma(t_1), \gamma(t_2)) = |t_1 - t_2|$ holds for all $t_1, t_2 \in \mathbb{R}$. Such a geodesic is an isometric embedding of the whole real line $\mathbb{R}$ in X. A geodesic is called a **metric great circle** if it is an isometric embedding of a circle $S^1(0, r)$ in X. In general, geodesics need not exist.

- **Geodesic metric space**

  A metric space (X, d) is called **geodesic** if any two points in X can be joined by a **geodesic segment**, i.e., for any two points $x, y \in X$, there is an isometry from the segment $[0, d(x, y)]$ into X. Examples of geodesic spaces are complete *Riemannian spaces*, **Banach spaces**, **metric graphs** from Chap. 15 and (Ivanov–Nikolaeva–Tuzhilin, 2015) *Gromov–Hausdorff space*.

  A metric space (X, d) is called a **locally geodesic metric space** if any two sufficiently close points in X can be joined by a geodesic segment; it is called **D-geodesic** if any two points at distance $< D$ can be joined by a geodesic segment.
- **Geodesic distance**

  The **geodesic distance** (or **shortest path distance**) is the length of a **geodesic segment** (i.e., a *shortest* path) between two points.
- **Intrinsic metric**

  Given a metric space (X, d) in which every two points are joined by a rectifiable curve, the **internal metric** (Chap. 4) D on X is defined as the infimum of the lengths of all rectifiable curves, connecting two given points $x, y \in X$.

  The metric d on X is called the **intrinsic metric** (or **length metric**) if it coincides with its internal metric D. A metric space with the intrinsic metric is called a **length space** (or **path metric space**, *inner metric space*, *intrinsic space*).

  If, moreover, any pair x, y of points can be joined by a curve of length $d(x, y)$, the intrinsic metric d is called *strictly intrinsic*, and the length space (X, d) is a **geodesic metric space** (or *shortest* path *metric space*).

A complete metric space (X, d) is a length space if and only if it is having **approximate midpoints**, i.e., for any points $x, y \in X$ and for any $\epsilon > 0$, there exists a third point $z \in X$ with $d(x, z), d(y, z) \leq \frac{1}{2}d(x, y) + \epsilon$. A complete metric space (X, d) is a **geodesic metric space** if and only if it is having **midpoints**.

Any complete locally compact length space is a **proper** geodesic metric space.

- **G-space**

  A **G-space** (or **space of geodesics**) is a metric space (X, d) with the geometry characterized by the fact that extensions of geodesics, defined as locally shortest lines, are unique. Such geometry is a generalization of *Hilbert Geometry* (see [Buse55]).

  More exactly, a G-space (X, d) is defined by the following conditions:

  1. It is **proper** (or *finitely compact*), i.e., all metric balls are compact;
  2. It is **Menger-convex**, i.e., for any different $x, y \in X$, there exists a third point $z \in X$, $z \neq x, y$, such that $d(x, z) + d(z, y) = d(x, y)$;
  3. It is *locally extendable*, i.e., for any $a \in X$, there exists $r > 0$ such that, for any distinct points x, y in the ball $B(a, r)$, there exists z distinct from x and y such that $d(x, y) + d(y, z) = d(x, z)$;
  4. It is *uniquely extendable*, i.e., if in 3 above two points z_1 and z_2 were found, so that $d(y, z_1) = d(y, z_2)$, then $z_1 = z_2$.

  The existence of geodesic segments is ensured by finite compactness and Menger-convexity: any two points of a finitely compact Menger-convex set X can be joined by a geodesic segment in X. The existence of geodesics is ensured by the axiom of local prolongation: if a finitely compact Menger-convex set X is locally extendable, then there exists a geodesic containing a given segment. Finally, the uniqueness of prolongation ensures the assumption of Differential Geometry that a *line element* determines a geodesic uniquely.

  All *Riemannian* and *Finsler spaces* are G-spaces. A 1D G-space is a metric straight line or metric great circle. Any 2D G-space is a **topological manifold** (Chap. 2).

  Every G-space is a **chord metric space**, i.e., a metric space with a set distinguished geodesic segments such that any two points are joined by a unique such segment (see [BuPh87]).

- **Desarguesian space**

  A **Desarguesian space** is a **G-space** (X, d) in which the role of geodesics is played by ordinary straight lines. Thus, X may be topologically mapped into a *projective space* $\mathbb{R}P^n$ so that each geodesic of X is mapped into a straight line of $\mathbb{R}P^n$.

  Any X mapped into $\mathbb{R}P^n$ must either cover all of $\mathbb{R}P^n$ and, in such a case, the geodesics of X are all metric great circles of the same length, or X may be considered as an open convex subset of an affine space A^n.

  A space (X, d) of geodesics is a Desarguesian space if and only if the following conditions hold:

[redacted]

1. The geodesic passing through two different points is unique;
2. For dimension $n = 2$, both the direct and the converse *Desargues theorems* are valid and, for dimension $n > 2$, any three points in X lie in one plane.

Among *Riemannian spaces*, the only Desarguesian spaces are Euclidean, *hyperbolic*, and *elliptic* spaces. An example of the non-Riemannian Desarguesian space is the *Minkowskian space* which can be regarded as the prototype of all non-Riemannian spaces, including *Finsler spaces*.

- ***G*-space of elliptic type**

  A ***G*-space of elliptic type** is a ***G*-space** in which the geodesic through two points is unique, and all geodesics are the metric great circles of the same length.

  Every G-space such that there is unique geodesic through each given pair of points is either a G-space of elliptic type, or a **straight *G*-space**.

- **Straight *G*-space**

  A **straight *G*-space** is a ***G*-space** in which extension of a geodesic is possible globally, so that any segment of the geodesic remains a shortest path. In other words, for any two points $x, y \in X$, there is a unique geodesic segment joining x to y, and a unique metric straight line containing x and y.

  Any geodesic in a straight G-space is a metric straight line, and is uniquely determined by any two of its points. Any such 2D space is homeomorphic to the plane.

  All simply connected *Riemannian spaces* of nonpositive curvature (including Euclidean and *hyperbolic* spaces), *Hilbert geometries*, and Teichmüller spaces of compact Riemann surfaces of genus $g > 1$ (when metrized by the **Teichmüller metric**) are straight G-spaces.

- **Gromov hyperbolic metric space**

  A metric space (X, d) is called **Gromov hyperbolic** if it is **geodesic** and **δ-hyperbolic** for some $\delta \geq 0$.

  An important class of such spaces are the *hyperbolic groups*, i.e., finitely generated groups whose **word metric** is Gromov hyperbolic. A metric space is a **real tree** exactly when it is 0-hyperbolic.

  Every bounded metric space X is (diam(X,d))-hyperbolic. A normed vector space is Gromov hyperbolic if and only it has dimension 1. Any complete simply connected *Riemannian space* of *sectional curvature* $k \leq -a^2 < 0$ is $\frac{\ln 3}{a}$-hyperbolic. Every **CAT(κ)** space with $\kappa < 0$ is Gromov hyperbolic.

- **CAT(κ) space**

  Let (X, d) be a metric space. Let M^2 be a simply connected 2D *Riemannian manifold* (Chap. 7) of *constant curvature* κ, i.e., the 2-sphere S^2_κ with $\kappa > 0$, the Euclidean plane $\mathbb{E}^2$ with $\kappa = 0$, or the hyperbolic plane H^2_κ with $\kappa < 0$. Let D_κ denote the *diameter* of M^2, i.e., $D_\kappa = \frac{\pi}{\sqrt{\kappa}}$ if $\kappa > 0$, and $D_\kappa = \infty$ if $\kappa \leq 0$.

  A *triangle* T in X consists of three points in X together with three *geodesic segments* joining them pairwise; the segments are called the *sides of the triangle*. For a triangle $T \subset X$, a *comparison triangle* for T in M^2 is a triangle $T' \subset M^2$ together with a map f_T which sends each side of T isometrically onto a side of T'. A triangle T is said (Gromov, 1987) to satisfy the **CAT(κ) inequality** (for

Cartan, Alexandrov and Toponogov) if, for every $x, y \in T$, we have

$$d(x, y) \leq d_{M^2}(f_T(x), f_T(y)),$$

where f_T is the map associated to a comparison triangle for T in M^2. So, the geodesic triangle T is at least as "thin" as its comparison triangle in M^2.

The metric space (X, d) is a **CAT(κ) space** if it is **D_κ-geodesic** (i.e., any two points at distance $< D_\kappa$ can be joined by a geodesic segment), and all triangles T with perimeter $< 2D_\kappa$ satisfy the CAT(κ) inequality.

Every CAT(κ_1) space is a CAT(κ_2) space if $\kappa_1 < \kappa_2$. Every **real tree** is a CAT($-\infty$) *space*, i.e., is a CAT(κ_1) space for all $\kappa \in \mathbb{R}$.

A **locally CAT(κ) space** (called **metric space with curvature** $\leq \kappa$ in Alexandrov, 1951) is a metric space (X, d) in which every point $p \in X$ has a neighborhood U such that any two points $x, y \in U$ are connected by a geodesic segment, and the CAT(κ) inequality holds for any $x, y, z \in U$. A Riemannian manifold is locally CAT(κ) if and only if its *sectional curvature* is at most κ.

A **metric space with curvature** $\geq \kappa$ is (Alexandrov, 1951) a metric space (X, d) in which every $p \in X$ has a neighborhood U such that any $x, y \in U$ are connected by a geodesic segment, and the *reverse CAT(κ) inequality*

$$d(x, y) \geq d_{M^2}(f_T(x), f_T(y))$$

holds for any $x, y, z \in U$, where f_T is the map associated to a comparison triangle for T in M^2. It is a **generalized Riemannian space** (Chap. 7).

Above two definitions differ only by the sign of $d(x, y) - d_{M^2}(f_T(x), f_T(y))$. In the case $\kappa = 0$, the above spaces are called **nonpositively curved** and **nonnegatively curved** metric spaces, respectively. For complete metric spaces, they differ also (Bruhat–Tits, 1972) by the sign (≤ 0 or ≥ 0, respectively) of

$$F(x, y, z) = 4d^2(z, m(x, y)) - (d^2(z, x) + d^2(z, y) - d^2(x, y)),$$

where x, y, z are any three points and $m(x, y)$ is the midpoint of the **metric interval** $I(x, y)$. A complete CAT(0) space is called **Hadamard space**.

The inequality $F(x, y, z) \leq 0$ for all $x, y, z \in X$, characterizing Hadamard spaces, is called *semiparallelogram inequality*, because the usual vector parallelogram law $||u - v||^2 + ||u + v||^2 = 2||u||^2 + 2||v||^2$, characterizing norms induced by inner products, is equivalent to the equality $F(x, y, z) = 0$. A normed space is an Hadamard space if and only if it is a **Hilbert space**.

Every two points in an Hadamard space are connected by a unique geodesic (and hence unique shortest path), while in a general CAT(0) space, they are connected by a unique geodesic segment, and the distance is a convex function.

Foertsch–Lytchack–Schroeder, 2007, proved that a metric space is CAT(0) if and only if it is **Busemann convex** and **Ptolemaic**; cf. Chap. 1. Euclidean spaces, hyperbolic spaces, and trees are CAT(0) spaces.

[redacted]

- **δ-bolic metric space**

  Given a number $\delta > 0$, a metric space (X, d) is called **δ-bolic** (Kasparov–Skandalis, 1994, simplified by Bucher–Karlsson, 2002) if for any $x, y, z \in X$ and some function $m : X \times X \to X$, it holds

  $$2d(z, m(x, y)) \le \sqrt{2d^2(z, x) + 2d^2(z, y) - d^2(x, y)} + \frac{4}{3}\delta.$$

  A **δ-hyperbolic space** with **approximate δ-midpoints** (Chap. 1) is $\frac{3\delta}{2}$-bolic.

  Every CAT(0)-space is δ-bolic for any $\delta > 0$; for complete spaces the converse holds as well. An l_p-metric space of dimension > 1 is δ-bolic for any $\delta > 0$ only if $p = 2$.

- **Boundary of metric space**

  There are many notions of the **boundary ∂X of a metric space** (X, d). We give below some of the most general among them. Usually, if (X, d) is locally compact, $X \cup \partial X$ is its *compactification*.

1. **Ideal boundary** (or *boundary at* ∞). Given a geodesic metric space (X, d), let γ^1 and γ^2 be two **metric rays**, i.e., geodesics with isometry of $\mathbb{R}_{\ge 0}$ into X. These rays are called *equivalent* if the **Hausdorff distance** between them (associated with the metric d) is finite, i.e., if $\sup_{t \ge 0} d(\gamma^1(t), \gamma^2(t)) < \infty$.

   The **ideal boundary** of (X, d) is the set $\partial_\infty X$ of equivalence classes γ_∞ of all metric rays. Cf. **asymptotic metric cone** (Chap. 1).

   If (X, d) is a complete CAT(0) space, then the **Tits metric** (or *asymptotic angle of divergence*) on $\partial_\infty X$ is defined by $2 \arcsin\left(\frac{\rho}{2}\right)$ for all $\gamma^1_\infty, \gamma^2_\infty \in \partial_\infty X$, where $\rho = \lim_{t \to \infty} \frac{1}{t} d(\gamma^1(t), \gamma^2(t))$. The set $\partial_\infty X$ equipped with the Tits metric is called the **Tits boundary** of X.

   If (X, d, x_0) is a pointed complete CAT(-1) space, then the **Bourdon metric** (or **visual distance**) on $\partial_\infty X$ is defined, for any distinct $x, y \in \partial_\infty X$, by $e^{-(x.y)}$, where $(x.y)$ denotes the **Gromov product** $(x.y)_{x_0}$.

   The **visual sphere of** (X, d) **at a point** $x_0 \in X$ is the set of equivalence classes of all metric rays emanating from x_0.
2. **Gromov boundary**. Given a **pointed metric space** (X, d, x_0) (i.e., one with a selected base point $x_0 \in X$), the **Gromov boundary** of it (as generalized by Buckley and Kokkendorff, 2005, from the case of the Gromov hyperbolic space) is the set $\partial_G X$ of equivalence classes of *Gromov sequences*.

   A sequence $x = \{x_n\}_n$ in X is a *Gromov sequence* if the Gromov product $(x_i.x_j)_{x_0} \to \infty$ as $i, j \to \infty$. Two Gromov sequences x and y are *equivalent* if there is a finite chain of Gromov sequences x^k, $0 \le k \le k'$, such that $x = x^0, y = x^{k'}$, and $\lim_{i,j \to \infty} \inf(x_i^{k-1}.x_j^k) = \infty$ for $0 \le k \le k'$.

   In a **proper** geodesic Gromov hyperbolic space (X, d), the visual sphere does not depends on the base point x_0 and is naturally isomorphic to its *Gromov boundary* $\partial_G X$ which can be identified with $\partial_\infty X$.
3. **g-boundary**. Denote by $\overline{X_d}$ the metric completion of (X, d) and, viewing X as a subset of $\overline{X_d}$, denote by ∂X_d the difference $\overline{X_d} \backslash X$. Let (X, l, x_0) be a pointed

unbounded **length space**, i.e., its metric coincides with the **internal metric** l of (X, d). Given a measurable function $g : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$, the g**-boundary** of (X, d, x_0) (as generalized by Buckley–Kokkendorff, 2005, from *spherical* and *Floyd boundaries*) is $\partial_g X = \partial X_\sigma \backslash \partial X_l$, where $\sigma(x, y) = \inf \int_\gamma g(z)dl(z)$ for all $x, y \in X$ (here the infimum is taken over all metric segments $\gamma = [x, y]$).

4. **Hotchkiss boundary**. Given a pointed proper **Busemann convex** metric space (X, d, x_0), the **Hotchkiss boundary** of it is the set $\partial_H(X, x_0)$ of isometries $f : \mathbb{R}_{\geq 0} \rightarrow X$ with $f(0) = x_0$. The boundaries $\partial_H^{x_0} X$ and $\partial_H^{x_1} X$ are homeomorphic for distinct $x_0, x_1 \in X$. In a Gromov hyperbolic space, $\partial_H^{x_0} X$ is homeomorphic to the Gromov boundary $\partial_G X$.
5. **Metric boundary**. Given a pointed metric space (X, d, x_0) and an unbounded subset S of $\mathbb{R}_{\geq 0}$, a ray $\gamma : S \rightarrow X$ is called a *weakly geodesic ray* if, for every $x \in X$ and every $\epsilon > 0$, there is an integer N such that $|d(\gamma(t), \gamma(0)) - t| < \epsilon$, and $|d(\gamma(t), x) - d(\gamma(s), x) - (t - s)| < \epsilon$ for all $s, t \in T$ with $s, t \geq N$.

   Let $\mathcal{G}(X, d)$ be the *commutative unital* C^**-algebra* with the norm $||.||_\infty$, generated by the (bounded, continuous) functions which vanish at infinity, the constant functions, and the functions of the form $g_y(x) = d(x, x_0) - d(x, y)$; cf. **Rieffel metric space** in Chap. 7 for definitions.

   The Rieffel's **metric boundary** $\partial_R X$ of (X, d) is the difference $\overline{X}^d \backslash X$, where $\overline{X}^d$ is the *metric compactification* of (X, d), i.e., the maximum ideal space (the set of *pure states*) of this C^*-algebra.

   For a **proper metric space** (X, d) (Chap. 1) with a countable base, the boundary $\partial_R X$ consists of the limits $\lim_{t \rightarrow \infty} f(\gamma(t))$ for every weakly geodesic ray γ and every function f from the above C^*-algebra (Rieffel, 2002).

- **Projectively flat metric space**

  A metric space, in which geodesics are defined, is called **projectively flat** if it locally admits a *geodesic mapping* (or *projective mapping*), i.e., a mapping preserving geodesics into an Euclidean space. Cf. Euclidean **rank of metric space** in Chap. 1; similar terms are: *affinely flat*, *conformally flat*, etc.

  A Riemannian space is projectively flat if and only if it has constant (sectional) curvature. Cf. **flat metric** in Chap. 8.

6.2 Projective Geometry

Projective Geometry is a branch of Geometry dealing with the properties and invariants of geometric figures under *projection*. Affine Geometry, Metric Geometry and Euclidean Geometry are subsets of Projective Geometry of increasing complexity. The main invariants of Projective, Affine, Metric, Euclidean Geometry are, respectively, **cross-ratio**, parallelism (and relative distances), angles (and relative distances), absolute distances.

[redacted]

An n-dimensional *projective space* $\mathbb{F}P^n$ is the space of one-dimensional vector subspaces of a given $(n+1)$-dimensional vector space V over a field $\mathbb{F}$. The basic construction is to form the set of equivalence classes of nonzero vectors in V under the relation of scalar proportionality. This idea goes back to mathematical descriptions of *perspective*.

The use of a basis of V allows the introduction of *homogeneous coordinates* of a point in $\mathbb{F}P^n$ which are usually written as $(x_1 : x_2 : \ldots : x_n : x_{n+1})$—a vector of length $n+1$, other than $(0 : 0 : 0 : \ldots : 0)$. Two sets of coordinates that are proportional denote the same point of the projective space. Any point of projective space which can be represented as $(x_1 : x_2 : \ldots : x_n : 0)$ is called a *point at infinity*. The part of a projective space $\mathbb{F}P^n$ not "at infinity" , i.e., the set of points of the projective space which can be represented as $(x_1 : x_2 : \ldots : x_n : 1)$, is an n-dimensional *affine space* A^n.

The notation $\mathbb{R}P^n$ denotes the *real projective space* of dimension n, i.e., the space of 1D vector subspaces of $\mathbb{R}^{n+1}$. The notation $\mathbb{C}P^n$ denotes the *complex projective space* of dimension n. The projective space $\mathbb{R}P^n$ carries a natural structure of a compact smooth n-**manifold**. It can be viewed as the space of lines through the zero element 0 of $\mathbb{R}^{n+1}$ (i.e., as a *ray space*). It can be viewed also as the set $\mathbb{R}^n$, considered as an *affine space*, together with its points at infinity. Also it can be seen as the set of points of an n-sphere in $\mathbb{R}^{n+1}$ with identified diametrically-opposite points.

The projective points, projective straight lines, projective planes, ..., projective hyperplanes of $\mathbb{F}P^n$ are one-, two-, three-, ..., n-dimensional subspaces of V, respectively. Any two projective straight lines in a projective plane have one and only one common point. A *projective transformation* (or *collineation*, *projectivity*) is a bijection of a projective space onto itself, preserving collinearity (the property of points to be on one line) in both directions. Any projective transformation is a composition of a pair of *perspective projections*. Projective transformations do not preserve sizes or angles but do preserve *type* (that is, points remain points, and lines remain lines), *incidence* (that is, whether a point lies on a line), and **cross-ratio** (Chap. 1).

Here, given four collinear points $x, y, z, t \in \mathbb{F}P^n$, their **cross-ratio** (x, y, z, t) is $\frac{(x-z)(y-t)}{(y-z)(x-t)}$, where $\frac{x-z}{x-t}$ denotes the ratio $\frac{f(x)-f(z)}{f(x)-f(t)}$ for some affine bijection f from the straight line $l_{x,y}$ through the points x and y onto $\mathbb{F}$.

Given four projective straight lines l_x, l_y, l_z, l_t, containing points x, y, z, t, respectively, and passing through a given point, their **cross-ratio** (l_x, l_y, l_z, l_t) is $\frac{\sin(l_x,l_z)\sin(l_y,l_t)}{\sin(l_y,l_z)\sin(l_x,l_t)}$, coincides with (x, y, z, t). The cross-ratio (x, y, z, t) of four complex numbers x, y, z, t is $\frac{(x-z)(y-t)}{(y-z)(x-t)}$. It is real if and only if the four numbers are either collinear or concyclic.

- **Projective metric**

  Given a convex subset D of a projective space $\mathbb{R}P^n$, the **projective metric** d is a metric on D such that shortest paths with respect to this metric are parts of or entire projective straight lines. It is assumed that the following conditions hold:

1. D does not belong to a hyperplane;
2. For any three noncollinear points $x, y, z \in D$, the triangle inequality holds in the strict sense: $d(x, y) < d(x, z) + d(z, y)$;
3. If x, y are different points in D, then the intersection of the straight line $l_{x,y}$ through x and y with D is either all of $l_{x,y}$, and forms a **metric great circle**, or is obtained from $l_{x,y}$ by discarding some segment (which can be reduced to a point), and forms a **metric straight line**.

The metric space (D, d) is called a **projective metric space**. The problem of determining all projective metrics on $\mathbb{R}^n$ (called **linearly additive metrics** in Chap. 1) is the 4-*th problem of Hilbert*; it has been solved only for $n = 2$. In fact, given a smooth measure on the space of hyperplanes in $\mathbb{R}P^n$, define the distance between any two points $x, y \in \mathbb{R}P^n$ as one-half the measure of all hyperplanes intersecting the line segment joining x and y. The obtained metric is projective; it is the *Busemann's construction* of projective metrics. [Amba76] proved that all projective metrics on $\mathbb{R}^2$ can be obtained by this construction.

In a projective metric space there cannot simultaneously be both types of straight lines: they are either all metric straight lines, or they are all metric great circles of the same length (*Hamel's theorem*). Spaces of the first kind are called *open*. They coincide with subspaces of an affine space; the geometry of open projective metric spaces is a *Hilbert Geometry*. *Hyperbolic Geometry* is a Hilbert Geometry in which there exist reflections at all straight lines.

Thus, the set D has Hyperbolic Geometry if and only if it is the interior of an ellipsoid. The geometry of open projective metric spaces whose subsets coincide with all of affine space, is a *Minkowski Geometry*. *Euclidean Geometry* is a Hilbert Geometry and a Minkowski Geometry, simultaneously. Spaces of the second kind are called *closed*; they coincide with the whole of $\mathbb{R}P^n$. *Elliptic Geometry* is the geometry of a projective metric space of the second kind.

- **Strip projective metric**

  The **strip projective metric** ([BuKe53]) is a **projective metric** on the strip $St = \{x \in \mathbb{R}^2 : -\pi/2 < x_2 < \pi/2\}$ defined by

  $$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} + |\tan x_2 - \tan y_2|.$$

  The Euclidean metric $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$ is not a projective metric on St.
- **Half-plane projective metric**

  The **half-plane projective metric** ([BuKe53]) is a **projective metric** on $\mathbb{R}^2_+ = \{x \in \mathbb{R}^2 : x_2 > 0\}$ defined by

  $$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} + \left|\frac{1}{x_2} - \frac{1}{y_2}\right|.$$

- **Hilbert projective metric**

  Given a set H, the **Hilbert projective metric** h is a **complete projective metric** on H. It means that H contains, together with two arbitrary distinct

[redacted]

points x and y, also the points z and t for which $h(x,z) + h(z,y) = h(x,y)$, $h(x,y) + h(y,t) = h(x,t)$, and that H is homeomorphic to a convex set in an n-dimensional affine space A^n, the geodesics in H being mapped to straight lines of A^n.

The metric space (H,h) is called the *Hilbert projective space*, and the geometry of a Hilbert projective space is called *Hilbert Geometry*.

Formally, let D be a nonempty convex open set in A^n with the boundary ∂D not containing two proper coplanar but noncollinear segments (ordinarily the boundary of D is a strictly convex closed curve, and D is its interior). Let $x, y \in D$ be located on a straight line which intersects ∂D at z and t, z is on the side of y, and t is on the side of x. Then the **Hilbert projective metric** h on D is the symmetrization of the **Funk distance** (Chap. 1):

$$h(x,y) = \frac{1}{2}\left(\ln\frac{x-z}{y-z} + \ln\frac{x-t}{y-t}\right) = \frac{1}{2}\ln(x,y,z,t),$$

where (x,y,z,t) is the **cross-ratio** of x, y, z, t.

The metric space (D,h) is a **straight *G*-space**. If D is an ellipsoid, then h is the **hyperbolic metric**, and defines *Hyperbolic Geometry* on D. On the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ the metric h coincides with the **Cayley–Klein–Hilbert metric**. If $n = 1$, the metric h makes D isometric to the Euclidean line.

If ∂D contains coplanar but noncollinear segments, a projective metric on D can be given by $h(x,y) + d(x,y)$, where d is any **Minkowskian metric**.

- **Minkowskian metric**

The **Minkowskian metric** (or **Minkowski–Hölder distance**) is the **norm metric** of a finite-dimensional real **Banach space**.

Formally, let $\mathbb{R}^n$ be an n-dimensional real vector space, let K be a *symmetric convex body* in $\mathbb{R}^n$, i.e., an open neighborhood of the origin which is bounded, convex, and *symmetric* ($x \in K$ if and only if $-x \in K$). Then the **Minkowski distance function** $||x||_K : \mathbb{R}^n \to [0,\infty)$, defined as $\inf\{\alpha > 0 : \frac{x}{\alpha} \in \partial K\}$ (cf. Chap. 1), is a *norm* on $\mathbb{R}^n$, and the Minkowskian metric m_K on $\mathbb{R}^n$ is defined by

$$m_K(x,y) = ||x-y||_K.$$

The metric space $(\mathbb{R}^n, m)$ is called *Minkowskian space*; its geometry is *Minkowski Geometry*. It can be seen as an affine space A^n with a metric m in which the *unit ball* is the body K. For a strictly convex symmetric body the Minkowskian metric is a **projective metric**, and $(\mathbb{R}^n, m)$ is a ***G*-straight space**. A Minkowski Geometry is Euclidean if and only if its *unit sphere* is an ellipsoid.

The Minkowskian metric m is proportional to the Euclidean metric d_E on every given line l, i.e., $m(x,y) = \phi(l)d_E(x,y)$. Thus, the Minkowskian metric can be considered as a metric which is defined in the whole affine space A^n and has the property that the *affine ratio* $\frac{ac}{ab}$ of any three collinear points a, b, c (cf. Sect. 6.3) is equal to their *distance ratio* $\frac{m(a,c)}{m(a,b)}$.

Given a convex body C in a Minkowskian space with unit ball K, the *Minkowskian thickness* and *Minkowskian diameter* of C are (Averkov, 2003):

$$\sup\{\alpha > 0 : \alpha K \subset C - C\} \text{ and } \inf\{\alpha > 0 : C - C \subset \alpha K\}.$$

- ***C*-distance**

  Given a convex body $C \subset \mathbb{E}^n$, the ***C*-distance** (or *relative distance*; Lassak, 1991) is a distance on $\mathbb{E}^n$ defined, for any $x, y \in \mathbb{E}^n$, by

$$d_C(x, y) = 2\frac{d_E(x, y)}{d_E(x', y')},$$

  where $x'y'$ is the longest chord of C parallel to the segment xy. C-distance is not related to C-metric in Chap. 10 and to **rotating *C*-metric** in Chap. 26.

  The unit ball of the normed space with the norm $||x|| = d_C(x, 0)$ is $\frac{1}{2}(C - C)$. For every $r \in [-1, 1]$, it holds $d_C(x, y) = d_{rC+(1-r)(-C)}(x, y)$.

- **Busemann metric**

  The **Busemann metric** ([Buse55]) is a metric on the real n-dimensional projective space $\mathbb{R}P^n$ defined by

$$\min\left\{\sum_{i=1}^{n+1}\left|\frac{x_i}{||x||} - \frac{y_i}{||y||}\right|, \sum_{i=1}^{n+1}\left|\frac{x_i}{||x||} + \frac{y_i}{||y||}\right|\right\}$$

  for any $x = (x_1 : \ldots : x_{n+1}), y = (y_1 : \ldots : y_{n+1}) \in \mathbb{R}P^n$, where $||x|| = \sqrt{\sum_{i=1}^{n+1} x_1^2}$.

- **Flag metric**

  Given an n-dimensional *projective space* $\mathbb{F}P^n$, the **flag metric** d is a metric on $\mathbb{F}P^n$ defined by a *flag*, i.e., an *absolute* consisting of a collection of m-planes α_m, $m = 0, \ldots, n-1$, with α_{i-1} belonging to α_i for all $i \in \{1, \ldots, n-1\}$. The metric space $(\mathbb{F}P^n, d)$ is abbreviated by F^n and is called a *flag space*.

  If one chooses an affine coordinate system $(x_i)_i$ in a space F^n, so that the vectors of the lines passing through the $(n-m-1)$-plane α_{n-m-1} are defined by the condition $x_1 = \ldots x_m = 0$, then the flag metric $d(x, y)$ between the points $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ is defined by

$$d(x, y) = |x_1 - y_1|, \text{ if } x_1 \neq y_1, \; d(x, y) = |x_2 - y_2|, \text{ if } x_1 = y_1, x_2 \neq y_2, \ldots$$
$$\ldots, d(x, y) = |x_k - y_k|, \text{ if } x_1 = y_1, \ldots, x_{k-1} = y_{k-1}, x_k \neq y_k, \ldots .$$

- **Projective determination of a metric**

  The **projective determination of a metric** is an introduction, in subsets of a projective space, of a metric such that these subsets become isomorphic to a Euclidean, *hyperbolic*, or *elliptic space*.

[redacted]

To obtain a *Euclidean determination of a metric* in $\mathbb{R}P^n$, one should distinguish in this space an $(n-1)$-dimensional hyperplane π, called the *hyperplane at infinity*, and define $\mathbb{E}^n$ as the subset of the projective space obtained by removing from it this hyperplane π. In terms of homogeneous coordinates, π consists of all points $(x_1 : \ldots : x_n : 0)$, and $\mathbb{E}^n$ consists of all points $(x_1 : \ldots : x_n : x_{n+1})$ with $x_{n+1} \neq 0$. Hence, it can be written as $\mathbb{E}^n = \{x \in \mathbb{R}P^n : x = (x_1 : \ldots : x_n : 1)\}$. The Euclidean metric d_E on $\mathbb{E}^n$ is defined by

$$\sqrt{\langle x-y, x-y\rangle},$$

where, for any $x = (x_1 : \ldots : x_n : 1), y = (y_1 : \ldots : y_n : 1) \in \mathbb{E}^n$, one has $\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$.

To obtain a *hyperbolic determination of a metric* in $\mathbb{R}P^n$, a set D of interior points of a real oval hypersurface Ω of order two in $\mathbb{R}P^n$ is considered. The **hyperbolic metric** d_{hyp} on D is defined by

$$\frac{r}{2}|\ln(x, y, z, t)|,$$

where z and t are the points of intersection of the straight line $l_{x,y}$ through the points x and y with Ω, (x, y, z, t) is the **cross-ratio** of the points x, y, z, t, and $r > 0$ is a fixed constant. If, for any $x = (x_1 : \ldots : x_{n+1}), y = (y_1 : \ldots : y_{n+1}) \in \mathbb{R}P^n$, the *scalar product* $\langle x, y\rangle = -x_1 y_1 + \sum_{i=2}^{n+1} x_i y_i$ is defined, the hyperbolic metric on the set $D = \{x \in \mathbb{R}P^n : \langle x, x\rangle < 0\}$ can be written, for a fixed constant $r > 0$, as

$$r \operatorname{arccosh} \frac{|\langle x, y\rangle|}{\sqrt{\langle x, x\rangle \langle y, y\rangle}},$$

where arccosh denotes the nonnegative values of the inverse hyperbolic cosine.

To obtain an *elliptic determination of a metric* in $\mathbb{R}P^n$, one should consider, for any $x = (x_1 : \ldots : x_{n+1}), y = (y_1 : \ldots : y_{n+1}) \in \mathbb{R}P^n$, the *inner product* $\langle x, y\rangle = \sum_{i=1}^{n+1} x_i y_i$. The **elliptic metric** d_{ell} on $\mathbb{R}P^n$ is defined now by

$$r \arccos \frac{|\langle x, y\rangle|}{\sqrt{\langle x, x\rangle \langle y, y\rangle}},$$

where $r > 0$ is a fixed constant, and arccos is the inverse cosine in $[0, \pi]$.

In all the considered cases, some hypersurfaces of the second-order remain invariant under given **motions**, i.e., projective transformations preserving a given metric. These hypersurfaces are called *absolutes*. In the case of a Euclidean determination of a metric, the absolute is an imaginary $(n-2)$-dimensional oval surface of order two, in fact, the degenerate absolute $x_1^2 + \cdots + x_n^2 = 0, x_{n+1} = 0$.

In the case of a hyperbolic determination of a metric, the absolute is a real $(n-1)$-dimensional oval hypersurface of order two, in the simplest case, the absolute $-x_1^2 + x_2^2 + \cdots + x_{n+1}^2 = 0$. In the case of an elliptic determination of a metric, the absolute is an imaginary $(n-1)$-dimensional oval hypersurface of order two, in fact, the absolute $x_1^2 + \cdots + x_{n+1}^2 = 0$.

6.3 Affine Geometry

An n-dimensional *affine space* over a field $\mathbb{F}$ is a set A^n (the elements of which are called *points* of the affine space) to which corresponds an n-dimensional vector space V over $\mathbb{F}$ (called the *space associated to* A^n) such that, for any $a \in A^n$, $A = a + V = \{a + v : v \in V\}$. In the other words, if $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n) \in A^n$, then the vector $\overrightarrow{ab} = (b_1 - a_1, \dots, b_n - a_n)$ belongs to V.

In an affine space, one can add a vector to a point to get another point, and subtract points to get vectors, but one cannot add points, since there is no origin. Given points $a, b, c, d \in A^n$ such that $c \neq d$, and the vectors $\overrightarrow{ab}$ and $\overrightarrow{cd}$ are collinear, the scalar λ, defined by $\overrightarrow{ab} = \lambda\overrightarrow{cd}$, is called the *affine ratio* of ab and cd, and is denoted by $\frac{ab}{cd}$.

An *affine transformation* (or *affinity*) is a bijection of A^n onto itself which preserves *collinearity* and *ratios of distances* In this sense, *affine* indicates a special class of *projective transformations* that do not move any objects from the affine space to the plane at infinity or conversely. Any affine transformation is a composition of *rotations*, *translations*, *dilations*, and *shears*. The set of all affine transformations of A^n forms a group $Aff(A^n)$, called the *general affine group* of A^n. Each element $f \in Aff(A)$ can be given by a formula $f(a) = b$, $b_i = \sum_{j=1}^n p_{ij}a_j + c_j$, where $((p_{ij}))$ is an invertible matrix.

The subgroup of $Aff(A^n)$, consisting of affine transformations with $det((p_{ij})) = 1$, is called the *equi-affine group* of A^n. An *equi-affine space* is an affine space with the equi-affine group of transformations. The fundamental invariants of an equi-affine space are volumes of parallelepipeds. In an *equi-affine plane* A^2, any two vectors v_1, v_2 have an invariant $|v_1 \times v_2|$ (the modulus of their cross product)—the surface area of the parallelogram constructed on v_1 and v_2.

Given a nonrectilinear curve $\gamma = \gamma(t)$, its *affine parameter* (or *equi-affine arc length*) is an invariant $s = \int_{t_0}^t |\gamma' \times \gamma''|^{1/3} dt$. The invariant $k = \frac{d^2\gamma}{ds^2} \times \frac{d^3\gamma}{ds^3}$ is called the *equi-affine curvature* of γ. Passing to the general affine group, two more invariants of the curve are considered: the *affine arc length* $\sigma = \int k^{1/2} ds$, and the *affine curvature* $\mathrm{k} = \frac{1}{k^{3/2}} \frac{dk}{ds}$.

For A^n, $n > 2$, the *affine parameter* (or *equi-affine arc length*) of a curve $\gamma = \gamma(t)$ is defined by $s = \int_{t_0}^t |(\gamma', \gamma'', \dots, \gamma^{(n)})|^{\frac{2}{n(n+1)}} dt$, where the invariant $(v_1, \dots, v_n)$ is the (oriented) volume spanned by the vectors $v_1, \dots, v_n$ which is equal to the determinant of the $n \times n$ matrix whose i-th column is the vector v_i.

[redacted]

- **Affine distance**

  Given an *affine plane* A^2, a *line element* (a, l_a) of A^2 consists of a point $a \in A^2$ together with a straight line $l_a \subset A^2$ passing through a.

  The **affine distance** is a distance on the set of all line elements of A^2 defined by

$$2f^{1/3},$$

  where, for a given line elements (a, l_a) and (b, l_b), f is the surface area of the triangle abc if c is the point of intersection of the straight lines l_a and l_b. The affine distance between (a, l_a) and (b, l_b) can be interpreted as the affine length of the arc ab of a parabola such that l_a and l_b are tangent to the parabola at a and b, respectively.

- **Affine pseudo-distance**

  Let A^2 be an *equi-affine plane*, and let $\gamma = \gamma(s)$ be a curve in A^2 defined as a function of the *affine parameter s*. The **affine pseudo-distance** dp_{aff} for A^2 is

$$dp_{aff}(a, b) = \left| \overrightarrow{ab} \times \frac{d\gamma}{ds} \right|,$$

  i.e., it is equal to the surface area of the parallelogram constructed on the vectors $\overrightarrow{ab}$ and $\frac{d\gamma}{ds}$, where b is an arbitrary point in A^2, a is a point on γ, and $\frac{d\gamma}{ds}$ is the tangent vector to the curve γ at the point a.

  Similarly, the **affine pseudo-distance** for an *equi-affine space* A^3 is defined as

$$\left| (\overrightarrow{ab}, \frac{d\gamma}{ds}, \frac{d^2\gamma}{ds^2}) \right|,$$

  where $\gamma = \gamma(s)$ is a curve in A^3, defined as a function of the *affine parameter s*, $b \in A^3$, a is a point of γ, and the vectors $\frac{d\gamma}{ds}, \frac{d^2\gamma}{ds^2}$ are obtained at the point a.

  For A^n, $n > 3$, we have $dp_{aff}(a, b) = |(\overrightarrow{ab}, \frac{d\gamma}{ds}, \dots, \frac{d^{n-1}\gamma}{ds^{n-1}})|$. For an arbitrary parametrization $\gamma = \gamma(t)$, one obtains $dp_{aff}(a, b) = |(\overrightarrow{ab}, \gamma', \dots, \gamma^{(n-1)})||(\gamma', \dots, \gamma^{(n-1)})|^{\frac{1-n}{1+n}}$.

- **Affine metric**

  The **affine metric** is a metric on a *nondevelopable surface* $r = r(u_1, u_2)$ in an *equi-affine space* A^3, given by its **metric tensor** $((g_{ij}))$:

$$g_{ij} = \frac{a_{ij}}{|det((a_{ij}))|^{1/4}},$$

  where $a_{ij} = (\partial_1 r, \partial_2 r, \partial_{ij} r)$, $i, j \in \{1, 2\}$.

6.4 Non-Euclidean Geometry

The term *non-Euclidean Geometry* describes both *Hyperbolic Geometry* (or *Lobachevsky-Bolyai-Gauss Geometry*) and *Elliptic Geometry* which are contrasted with *Euclidean Geometry* (or *Parabolic Geometry*). The essential difference between Euclidean and non-Euclidean Geometry is the nature of parallel lines. In Euclidean Geometry, if we start with a line l and a point a, which is not on l, then there is only one line through a that is parallel to l. In Hyperbolic Geometry there are infinitely many lines through a parallel to l. In Elliptic Geometry, parallel lines do not exist. The *Spherical Geometry* is also "non-Euclidean", but it fails the axiom that any two points determine exactly one line.

- **Spherical metric**

  Let $S^n(0,r) = \{x \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = r^2\}$ be the sphere in $\mathbb{R}^{n+1}$ with the center 0 and the radius $r > 0$.

  The **spherical metric** (or *great circle metric*) is a metric on $S^n(0,r)$ defined by

  $$d_{sph} = r \arccos\left(\frac{|\sum_{i=1}^{n+1} x_i y_i|}{r^2}\right),$$

  where arccos is the inverse cosine in $[0,\pi]$. It is the length of the *great circle* arc, passing through x and y. In terms of the standard *inner product* $\langle x,y\rangle = \sum_{i=1}^{n+1} x_i y_i$ on $\mathbb{R}^{n+1}$, the spherical metric can be written as $r \arccos \frac{|\langle x,y\rangle|}{\sqrt{\langle x,x\rangle \langle y,y\rangle}}$.

  The metric space $(S^n(0,r), d_{sph})$ is called *n-dimensional spherical space*. It is a space of curvature $1/r^2$, and r is the radius of curvature. It is a model of n-dimensional *Spherical Geometry*. The great circles of the sphere are its geodesics and all geodesics are closed and of the same length. See, for example, [Blum70].

- **Elliptic metric**

  Let $\mathbb{R}P^n$ be the real n-dimensional projective space. The **elliptic metric** d_{ell} is a metric on $\mathbb{R}P^n$ defined by

  $$r \arccos \frac{|\langle x,y\rangle|}{\sqrt{\langle x,x\rangle \langle y,y\rangle}},$$

  where, for any $x = (x_1 : \ldots : x_{n+1})$ and $y = (y_1 : \ldots : y_{n+1}) \in \mathbb{R}P^n$, one has $\langle x,y\rangle = \sum_{i=1}^{n+1} x_i y_i$, $r > 0$ is a constant and arccos is the inverse cosine in $[0,\pi]$.

  The metric space $(\mathbb{R}P^n, d_{ell})$ is called *n-dimensional elliptic space*. It is a model of n-dimensional *Elliptic Geometry*. It is the space of curvature $1/r^2$, and r is the radius of curvature. As $r \to \infty$, the metric formulas of Elliptic Geometry yield formulas of Euclidean Geometry (or become meaningless).

  If $\mathbb{R}P^n$ is viewed as the set $E^n(0,r)$, obtained from the sphere $S^n(0,r) = \{x \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = r^2\}$ in $\mathbb{R}^{n+1}$ with center 0 and radius r by identifying diametrically-opposite points, then the elliptic metric on $E^n(0,r)$ can be written

[redacted]

as $d_{sph}(x, y)$ if $d_{sph}(x, y) \leq \frac{\pi}{2}r$, and as $\pi r - d_{sph}(x, y)$ if $d_{sph}(x, y) > \frac{\pi}{2}r$, where d_{sph} is the **spherical metric** on $S^n(0, r)$. Thus, no two points of $E^n(0, r)$ have distance exceeding $\frac{\pi}{2}r$. The elliptic space $(E^2(0, r), d_{ell})$ is called the *Poincaré sphere*.

If $\mathbb{R}P^n$ is viewed as the set E^n of lines through the zero element 0 in $\mathbb{R}^{n+1}$, then the elliptic metric on E^n is defined as the angle between the corresponding subspaces.

An n-dimensional elliptic space is a *Riemannian space* of constant positive curvature. It is the only such space which is topologically equivalent to a projective space. See, for example, [Blum70] and [Buse55].

- **Hermitian elliptic metric**

Let $\mathbb{C}P^n$ be the n-dimensional complex projective space. The **Hermitian elliptic metric** d^H_{ell} (see, for example, [Buse55]) is a metric on $\mathbb{C}P^n$ defined by

$$r \arccos \frac{|\langle x, y\rangle|}{\sqrt{\langle x, x\rangle\langle y, y\rangle}},$$

where, for any $x = (x_1 : \ldots : x_{n+1})$ and $y = (y_1 : \ldots : y_{n+1}) \in \mathbb{C}P^n$, one has $\langle x, y\rangle = \sum_{i=1}^{n+1} \overline{x}_i y_i$, $r > 0$ is a constant and arccos is the inverse cosine in $[0, \pi]$.

The metric space $(\mathbb{C}P^n, d^H_{ell})$ is called n-dimensional *Hermitian elliptic space* (cf. **Fubini–Study metric** in Chap. 7).

- **Elliptic plane metric**

The **elliptic plane metric** is the **elliptic metric** on the *elliptic plane* $\mathbb{R}P^2$.

If $\mathbb{R}P^2$ is viewed as the *Poincaré sphere* (i.e., a sphere in $\mathbb{R}^3$ with identified diametrically-opposite points) of diameter 1 tangent to the extended complex plane $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ at the point $z = 0$, then, under the stereographic projection from the "north pole" $(0, 0, 1)$, $\overline{\mathbb{C}}$ with identified points z and $-\frac{1}{\overline{z}}$ is a model of the elliptic plane.

The elliptic plane metric d_{ell} on it is defined by its *line element* $ds^2 = \frac{|dz|^2}{(1+|z|^2)^2}$.

- **Pseudo-elliptic distance**

The **pseudo-elliptic distance** (or *elliptic pseudo-distance*) dp_{ell} is defined, on the extended complex plane $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ with identified points z and $-\frac{1}{\overline{z}}$, by

$$\left|\frac{z - u}{1 + \overline{z}u}\right|.$$

In fact, $dp_{ell}(z, u) = \tan d_{ell}(z, u)$, where d_{ell} is the **elliptic plane metric**.

- **Hyperbolic metric**

Let $\mathbb{R}P^n$ be the n-dimensional real projective space. Let, for any $x = (x_1 : \ldots : x_{n+1})$, $y = (y_1 : \ldots : y_{n+1}) \in \mathbb{R}P^n$, their *scalar product* $\langle x, y\rangle$ be $-x_1y_1 + \sum_{i=2}^{n+1} x_i y_i$.

The **hyperbolic metric** d_{hyp} is a metric on the set $H^n = \{x \in \mathbb{R}P^n : \langle x, x\rangle < 0\}$ defined, for a fixed constant $r > 0$, by

$$r \operatorname{arccosh} \frac{|\langle x, y\rangle|}{\sqrt{\langle x, x\rangle \langle y, y\rangle}},$$

where arccosh denotes the nonnegative values of the inverse hyperbolic cosine.

In this construction, the points of H^n can be viewed as the one-spaces of the *pseudo-Euclidean space* $\mathbb{R}^{n,1}$ inside the cone $C = \{x \in \mathbb{R}^{n,1} : \langle x, x\rangle = 0\}$.

The metric space (H^n, d_{hyp}) is called *n-dimensional hyperbolic space*. It is a model of n-dimensional *Hyperbolic Geometry*. It is the space of curvature $-1/r^2$, and r is the radius of curvature. Replacement of r by ir transforms all metric formulas of Hyperbolic Geometry into the corresponding formulas of Elliptic Geometry. As $r \to \infty$, both systems yield formulas of Euclidean Geometry (or become meaningless).

If H^n is viewed as the set $\{x \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 < K\}$, where $K > 1$ is any fixed constant, the hyperbolic metric can be written as

$$\frac{r}{2} \ln \frac{1 + \sqrt{1 - \gamma(x, y)}}{1 - \sqrt{1 - \gamma(x, y)}},$$

where $\gamma(x, y) = \frac{(K - \sum_{i=1}^n x_i^2)(K - \sum_{i=1}^n y_i^2)}{(K - \sum_{i=1}^n x_i y_i)^2}$, and $r > 0$ is a number with $\tanh \frac{1}{r} = \frac{1}{\sqrt{K}}$.

If H^n is viewed as a submanifold of the $(n+1)$-dimensional *pseudo-Euclidean space* $\mathbb{R}^{n,1}$ with the scalar product $\langle x, y\rangle = -x_1 y_1 + \sum_{i=2}^{n+1} x_i y_i$ (in fact, as the top sheet $\{x \in \mathbb{R}^{n,1} : \langle x, x\rangle = -1, x_1 > 0\}$ of the two-sheeted *hyperboloid of revolution*), then the hyperbolic metric on H^n is induced from the **pseudo-Riemannian metric** on $R^{n,1}$ (cf. **Lorentz metric** in Chap. 26).

An n-dimensional hyperbolic space is a *Riemannian space* of constant negative curvature. It is the only such space which is **complete** and topologically equivalent to an Euclidean space. (See, for example, [Blum70, Buse55].)

- **Hermitian hyperbolic metric**

  Let $\mathbb{C}P^n$ be the n-dimensional complex projective space. Let, for any $x = (x_1 : \ldots : x_{n+1})$, $y = (y_1 : \ldots : y_{n+1}) \in \mathbb{C}P^n$, their *scalar product* $\langle x, y\rangle$ be $-\overline{x}_1 y_1 + \sum_{i=2}^{n+1} \overline{x}_i y_i$.

  The **Hermitian hyperbolic metric** d_{hyp}^H (see, for example, [Buse55]) is a metric on the set $\mathbb{C}H^n = \{x \in \mathbb{C}P^n : \langle x, x\rangle < 0\}$ defined, for a fixed constant $r > 0$, by

$$\operatorname{arccosh} \frac{|\langle x, y\rangle|}{\sqrt{\langle x, x\rangle \langle y, y\rangle}},$$

  where arccosh denotes the nonnegative values of the inverse hyperbolic cosine.

  The metric space $(\mathbb{C}H^n, d_{hyp}^H)$ is called n-dimensional *Hermitian hyperbolic space*.

[redacted]

- **Poincaré metric**

The **Poincaré metric** d_P is the **hyperbolic metric** for the *Poincaré disk model* of Hyperbolic Geometry. In this model the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ is called the *hyperbolic plane*, every point of Δ is called a *hyperbolic point*, circular arcs (and diameters) in Δ which are orthogonal to the *absolute* $\Omega = \{z \in \mathbb{C} : |z| = 1\}$ are called *hyperbolic straight lines*. Every point of Ω is called an *ideal point*.

The angular measurements in this model are the same as in Hyperbolic Geometry, i.e., it is a *conformal model*. There is a one-to-one correspondence between segments and acute angles. The Poincaré metric on Δ is defined by its *line element*

$$ds^2 = \frac{|dz|^2}{(1-|z|^2)^2} = \frac{dz_1^2 + dz_2^2}{(1-z_1^2-z_2^2)^2}.$$

The distance d_P between two points z and u of Δ can be written as

$$\frac{1}{2}\ln\frac{|1-z\overline{u}| + |z-u|}{|1-z\overline{u}| - |z-u|} = \operatorname{arctanh}\frac{|z-u|}{|1-z\overline{u}|}.$$

In terms of **cross-ratio**, it is equal to

$$\frac{1}{2}\ln(z,u,z^*,u^*) = \frac{1}{2}\ln\frac{(z^*-z)(u^*-u)}{(z^*-u)(u^*-z)},$$

where z^* and u^* are the points of intersection of the hyperbolic straight line passing through z and u with Ω, z^* on the side of u, and u^* on the side of z.

The **multiplicative distance function** on the segments zu of Δ is defined (Hartshorne, 2003) by $\mu(zu) = (z,u,z^*,u^*)^{-1}$; it allows the definition of trigonometric functions in the absence of continuity.

In the conformal *Poincaré half-plane model* of Hyperbolic Geometry the *hyperbolic plane* is the upper half-plane $H^2 = \{z \in \mathbb{C} : z_2 > 0\}$, and the *hyperbolic lines* are semicircles and half-lines which are orthogonal to the real axis. The *absolute* (i.e., the set of *ideal points*) is the real axis together with the point at infinity.

The *line element* of the **Poincaré metric** on H^2 is given by

$$ds^2 = \frac{|dz|^2}{(\Im z)^2} = \frac{dz_1^2 + dz_2^2}{z_2^2}.$$

The distance between two points z, u can be written as

$$\frac{1}{2}\ln\frac{|z-\overline{u}| + |z-u|}{|z-\overline{u}| - |z-u|} = \operatorname{arctanh}\frac{|z-u|}{|z-\overline{u}|} = \frac{1}{2}\ln(z,u,z^*,u^*) = \frac{1}{2}\ln\frac{(z^*-z)(u^*-u)}{(z^*-u)(u^*-z)},$$

where z^* is the ideal point of the half-line emanating from z and passing through u, and u^* is the ideal point of the half-line emanating from u and passing through z.

In general, the **hyperbolic metric** in any *domain* $D \subset \mathbb{C}$ with at least three boundary points is defined as the preimage of the Poincaré metric in Δ under a *conformal mapping* $f : D \to \Delta$. Its *line element* has the form

$$ds^2 = \frac{|f'(z)|^2|dz|^2}{(1-|f(z)|^2)^2}.$$

The distance between two points z and u in D can be written as

$$\frac{1}{2}\ln\frac{|1-f(z)\overline{f(u)}|+|f(z)-f(u)|}{|1-f(z)\overline{f(u)}|-|f(z)-f(u)|}.$$

- **Pseudo-hyperbolic distance**

  The **pseudo-hyperbolic distance** (or **Gleason distance**, *hyperbolic pseudo-distance*) dp_{hyp} is a metric on the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$, defined by

  $$\left|\frac{z-u}{1-\overline{z}u}\right|.$$

  In fact, $dp_{hyp}(z,u) = \tanh d_P(z,u)$, where d_P is the **Poincaré metric** on Δ.
- **Cayley–Klein–Hilbert metric**

  The **Cayley–Klein–Hilbert metric** d_{CKH} is the **hyperbolic metric** for the *Klein model* (or *projective disk model*, for Hyperbolic Geometry. In this model the *hyperbolic plane* is realized as the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$, and the *hyperbolic straight lines* are realized as the chords of Δ. Every point of the *absolute* $\Omega = \{z \in \mathbb{C} : |z| = 1\}$ is called an *ideal point*. This model is not conformal: the angular measurements are distorted. The **Cayley–Klein–Hilbert metric** on Δ is given by its **metric tensor** $((g_{ij}))$, $i,j = 1,2$:

  $$g_{11} = \frac{r^2(1-z_2^2)}{(1-z_1^2-z_2^2)^2},\ g_{12} = \frac{r^2 z_1 z_2}{(1-z_1^2-z_2^2)^2},\ g_{22} = \frac{r^2(1-z_1^2)}{(1-z_1^2-z_2^2)^2},$$

  where r is any positive constant. The distance between points z and u in Δ is

  $$r\operatorname{arccosh}\left(\frac{1-z_1u_1-z_2u_2}{\sqrt{1-z_1^2-z_2^2}\sqrt{1-u_1^2-u_2^2}}\right),$$

  where arccosh denotes the nonnegative values of the inverse hyperbolic cosine.

[redacted]

- **Weierstrass metric**

  Given a real n-dimensional **inner product space** $(\mathbb{R}^n, \langle , \rangle)$, $n \geq 2$, the **Weierstrass metric** d_W is a metric on $\mathbb{R}^n$ defined by

  $$\operatorname{arccosh}(\sqrt{1 + \langle x, x\rangle}\sqrt{1 + \langle y, y\rangle} - \langle x, y\rangle),$$

  where arccosh denotes the nonnegative values of the inverse hyperbolic cosine.

  Here, $(x, \sqrt{1 + \langle x, x\rangle}) \in \mathbb{R}^n \oplus \mathbb{R}$ are the *Weierstrass coordinates* of $x \in \mathbb{R}^n$, and the metric space $(\mathbb{R}^n, d_W)$ can be seen as the *Weierstrass model* of Hyperbolic Geometry.

  The **Cayley–Klein–Hilbert metric** $d_{CKH}(x, y) = \operatorname{arccosh}\frac{1-\langle x,y\rangle}{\sqrt{1-\langle x,x\rangle}\sqrt{1-\langle y,y\rangle}}$ on the open ball $B^n = \{x \in \mathbb{R}^n : \langle x, x\rangle < 1\}$ can be obtained from d_W by $d_{GKH}(x, y) = d_W(\mu(x), \mu(y))$, where $\mu : \mathbb{R}^n \to B^n$ is the *Weierstrass mapping*: $\mu(x) = \frac{x}{\sqrt{1-\langle x,x\rangle}}$.

- **Harnack metric**

  Given a *domain* $D \subset \mathbb{R}^n$, $n \geq 2$, the **Harnack metric** is a metric on D defined by

  $$\sup_f |\log\frac{f(x)}{f(y)}|,$$

  where the supremum is taken over all positive functions which are harmonic on D.

- **Quasi-hyperbolic metric**

  Given a *domain* $D \subset \mathbb{R}^n$, $n \geq 2$, the **quasi-hyperbolic metric** on D is defined by

  $$\inf_{\gamma\in\Gamma}\int_\gamma \frac{|dz|}{\rho(z)},$$

  where the infimum is taken over the set Γ of all rectifiable curves connecting x and y in D, $\rho(z) = \inf_{u\in\partial D} ||z - u||_2$ is the distance between z and the boundary ∂D of D, and $||.||_2$ is the Euclidean norm on $\mathbb{R}^n$.

  This metric is **Gromov hyperbolic** if the domain D is *uniform*, i.e., there exist constants C, C' such that each pair of points $x, y \in D$ can be joined by a rectifiable curve $\gamma = \gamma(x, y) \in D$ of length $l(\gamma)$ at most $C|x - y|$, and $\min\{l(\gamma(x, z)), l(\gamma(z, y))\} \leq C'd(z, \partial D)$ holds for all $z \in \gamma$. Also, the quasi-hyperbolic metric is the **inner metric** (Chap. 4) of the **distance ratio metric**.

  For $n = 2$, one can define the **hyperbolic metric** on D by

  $$\inf_{\gamma\in\Gamma}\int_\gamma \frac{2|f'(z)|}{1 - |f(z)|^2}|dz|,$$

where $f : D \to \Delta$ is any conformal mapping of D onto the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$. For $n \geq 3$, it is defined only for the half-hyperplane H^n and for the *open unit ball* B^n as the infimum over all $\gamma \in \Gamma$ of the integrals $\int_\gamma \frac{|dz|}{z_n}$ and $\int_\gamma \frac{2|dz|}{1-||z||_2^2}$.

- **Apollonian metric**

  Let $D \subset \mathbb{R}^n$, be a *domain* such that its complement is not contained in a hyperplane or a sphere. The **Apollonian metric** (or **Barbilian metric**, [Barb35]) on D is defined (denoting the boundary of D by ∂D) by the **cross-ratio** as

$$\sup_{a,b \in \partial D} \ln \frac{||a-x||_2 ||b-y||_2}{||a-y||_2 ||b-x||_2}.$$

  This metric is **Gromov hyperbolic**.

- **Half-Apollonian metric**

  Given a *domain* $D \subset \mathbb{R}^n$, the **half-Apollonian metric** η_D (Hăsto and Lindén, 2004) on D is defined (denoting the boundary of D by ∂D) by

$$\sup_{a \in \partial D} \left| \ln \frac{||a-y||_2}{||a-x||_2} \right|.$$

  This metric is **Gromov hyperbolic** only if the domain is $\mathbb{R}^n \backslash \{x\}$.

- **Gehring metric**

  Given a *domain* $D \subset \mathbb{R}^n$, the **Gehring metric** $\tilde{j}_D$ (Gehring, 1982) is a metric on D, defined by

$$\frac{1}{2} \ln \left(\left(1 + \frac{||x-y||_2}{\rho(x)} \right) \left(1 + \frac{||x-y||_2}{\rho(y)} \right) \right),$$

  where $\rho(x) = \inf_{u \in \partial D} ||x-u||_2$ is the distance between x and the boundary of D.

  This metric is **Gromov hyperbolic**.

- **Distance ratio metric**

  Given a *domain* $D \subset \mathbb{R}^n$, the **distance ratio metric** (or *j_D-metric*) is (Gehring–Palka, 1976, and Vuorinen, 1985) a metric on D defined by

$$\ln \left(1 + \frac{||x-y||_2}{\min\{\rho(x), \rho(y)\}} \right),$$

  where $\rho(x) = \inf_{u \in \partial D} ||x-u||_2$ is the distance between x and the boundary of D.

  This metric is **Gromov hyperbolic** only if the domain is $\mathbb{R}^n \backslash \{x\}$.

- **Triangular ratio metric**

  Given a *domain* $D \subset \mathbb{R}^n$, the **triangular ratio metric** on D is defined (denoting the boundary of D by ∂D) by

$$\sup_{z \in \partial D} \frac{||x-y||_2}{(||x-z||_2 + ||y-z||_2|}.$$

[redacted]

- **Visual angle metric**

  Given a *domain* $D \subset \mathbb{R}^n$, the **visual angle metric** on D is defined (denoting the boundary of D by ∂D) by Klein et al., 2014, as

$$\sup_{z \in \partial D} \angle(x, z, y).$$

- **Dovgoshev–Hariri–Vuorinen metric**

  Given a metric space (X, d) and an open set D in it with nonempty boundary ∂D, Dovgoshev–Hariri–Vuorinen, 2015, showed that

$$\ln\left(1 + \frac{cd(x, y)}{\sqrt{d(x, \partial D)d(y, \partial D)}}\right)$$

  is a metric for every $c \geq 2$ with 2 being the best possible constant.

- **Ferrand metric**

  Given a *domain* $D \subset \mathbb{R}^n$, the **Ferrand metric** σ_D (Ferrand, 1987) is a metric on D defined by

$$\inf_{\gamma \in \Gamma} \int_\gamma \sup_{a,b \in \partial D} \frac{||a - b||_2}{||z - a||_2||z - b||_2} |dz|,$$

  where the infimum is taken over the set Γ of all rectifiable curves connecting x and y in D, ∂D is the boundary of D, and $||.||_2$ is the Euclidean norm on $\mathbb{R}^n$.

  This metric is the **inner metric** (Chap. 4) of the **Möbius metric**.

- **Möbius metric**

  Given a *domain* $D \subset \mathbb{R}^n$, the **Möbius** (or *absolute ratio*, δ_D-) **metric**; Siettenranta, 1999) is a metric on D defined by

$$\sup_{a,b \in \partial D} \ln\left(1 + \frac{||a - x||_2||b - y||_2}{||a - b||_2||x - y||_2}\right).$$

  This metric is **Gromov hyperbolic**.

- **Modulus metric**

  Let $D \subset \mathbb{R}^n$, be a *domain*. The *conformal modulus* of a family Γ of locally rectifiable curves in D is $M(\Gamma) = \inf_\rho \int_{\mathbb{R}^n} \rho^n dm$, where m is the n-dimensional Lebesgue measure, and ρ is any Borel-measurable function with $\int_\gamma \rho ds \geq 1$ and $\rho \geq 0$ for each $\gamma \in \Gamma$. Cf. general *modulus* in **extremal metric**, Chap. 8.

  Let $\Delta(E, F; D)$ denote the family of all closed nonconstant curves in D joining E and F. The **modulus metric** μ_D (Gál, 1960) is a metric on D, defined by

$$\inf_{C_{xy}} M(\Delta(C_{xy}, \partial D; D)),$$

where C_{xy} is a compact connected set such that for some $\gamma : [0, 1] \to D$, it holds $C_{xy} = \gamma([0, 1])$ and $\gamma(0) = x$, $\gamma(1) = y$.

The **Ferrand second metric** λ^*_D (Ferrand, 1997) is a metric on D, defined by

$$\left(\inf_{C_x, C_y} M(\Delta(C_x, C_y; D))\right)^{\frac{1}{1-n}},$$

where C_z $(z = x, y)$ is a compact connected set such that, for some $\gamma_z : [0, 1] \to D$, it holds $C_z = \gamma([0, 1))$, $z \in |\gamma_z|$ and $\gamma_z(t) \to \partial D$ as $t \to 1$.

Above two metrics are **Gromov hyperbolic** if D is the open ball $B^n = \{x \in \mathbb{R}^n : \langle x, x\rangle < 1\}$ or a simply connected domain in $\mathbb{R}^2$.

- **Conformal radius**

Let $D \subset \mathbb{C}$, be a simply connected *domain* and let $z \in D$, $z \neq \infty$.

The **conformal** (or *harmonic*) **radius** is defined by

$$rad(z, D) = (f'(z))^{-1},$$

where $f : D \to \Delta$ is the *uniformizing map*, i.e., the unique conformal mapping onto the unit disk with $f(z) = 0 \in \Delta$ and $f'(z) > 0$.

The Euclidean distance between z and the boundary ∂D of D (i.e., the radius of the largest disk inscribed in D) lies in the segment $[\frac{rad(z,D)}{4}, rad(z, D)]$.

If D is compact, define $rad(\infty, D)$ as $\lim_{z\to\infty} \frac{f(z)}{z}$, where $f : (\mathbb{C}\backslash\Delta) \to (\mathbb{C}\backslash D)$ is the unique conformal mapping with $f(\infty) = \infty$ and positive above limit. This radius is the **transfinite diameter** from Chap. 1.

- **Parabolic distance**

The **parabolic distance** is a metric on $\mathbb{R}^{n+1}$, considered as $\mathbb{R}^n \times \mathbb{R}$ defined by

$$\sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2 + |t_x - t_y|^{1/m}}, m \in \mathbb{N},$$

for any $x = (x_1, \ldots, x_n, t_x), y = (y_1, \ldots, y_n, t_y) \in \mathbb{R}^n \times \mathbb{R}$.

The space $\mathbb{R}^n \times \mathbb{R}$ can be interpreted as multidimensional *space-time*.

Usually, the value $m = 2$ is applied. There exist some variants of the parabolic distance, for example, the parabolic distance

$$\sup\{|x_1 - y_1|, |x_2 - y_2|^{1/2}\}$$

on $\mathbb{R}^2$ (cf. also **Rickman's rug metric** in Chap. 19), or the **half-space parabolic distance** on $\mathbb{R}^3_+ = \{x \in \mathbb{R}^3 : x_1 \geq 0\}$ defined by

$$\frac{|x_1 - y_1| + |x_2 - y_2|}{\sqrt{x_1} + \sqrt{x_2} + \sqrt{|x_2 - y_2|}} + \sqrt{|x_3 - y_3|}.$$

[redacted]

Chapter 7
Riemannian and Hermitian Metrics

Riemannian Geometry is a multidimensional generalization of the intrinsic geometry of 2D surfaces in the Euclidean space $\mathbb{E}^3$. It studies *real smooth manifolds* equipped with **Riemannian metrics**, i.e., collections of positive-definite symmetric bilinear forms $((g_{ij}))$ on their tangent spaces which vary smoothly from point to point. The geometry of such (*Riemannian*) manifolds is based on the *line element* $ds^2 = \sum_{i,j} g_{ij}dx_idx_j$. This gives, in particular, local notions of angle, length of curve, and volume.

From these notions some other global quantities can be derived, by integrating local contributions. Thus, the value ds is interpreted as the length of the vector $(dx_1, \dots, dx_n)$, and it is called the **infinitesimal distance**. The arc length of a curve γ is expressed by $\int_\gamma \sqrt{\sum_{i,j} g_{ij}dx_idx_j}$, and then the **intrinsic metric** on a Riemannian manifold is defined as the infimum of lengths of curves joining two given points of the manifold.

Therefore, a Riemannian metric is not an ordinary metric, but it induces an ordinary metric, in fact, the intrinsic metric, called **Riemannian distance**, on any connected Riemannian manifold. A Riemannian metric is an infinitesimal form of the corresponding Riemannian distance.

As particular special cases of Riemannian Geometry, there occur *Euclidean Geometry* as well as the two standard types, *Elliptic Geometry* and *Hyperbolic Geometry*, of *non-Euclidean Geometry*. If the bilinear forms $((g_{ij}))$ are nondegenerate but indefinite, one obtains *pseudo-Riemannian Geometry*. In the case of dimension four (and *signature* $(1, 3)$) it is the main object of the General Theory of Relativity.

If $ds = F(x_1, \dots, x_n, dx_1, \dots, dx_n)$, where F is a real positive-definite convex function which cannot be given as the square root of a symmetric bilinear form (as in the Riemannian case), one obtains the *Finsler Geometry* generalizing Riemannian Geometry.

Hermitian Geometry studies *complex manifolds* equipped with **Hermitian metrics**, i.e., collections of positive-definite symmetric *sesquilinear forms* (or $\frac{3}{2}$*-linear*

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_7

forms) since they are linear in one argument and *antilinear* in the other) on their tangent spaces, which vary smoothly from point to point. It is a complex analog of Riemannian Geometry.

A special class of Hermitian metrics form **Kähler metrics** which have a closed fundamental form w. A generalization of Hermitian metrics give **complex Finsler metrics** which cannot be written as a bilinear symmetric positive-definite sesqulinear form.

7.1 Riemannian Metrics and Generalizations

A *real n-manifold* M^n *with boundary* is (cf. Chap. 2) a **Hausdorff space** in which every point has an open neighborhood homeomorphic to either an open subset of $\mathbb{E}^n$, or an open subset of the closed half of $\mathbb{E}^n$. The set of points which have an open neighborhood homeomorphic to $\mathbb{E}^n$ is called the *interior* (of the manifold); it is always nonempty.

The complement of the interior is called the *boundary* (of the manifold); it is an $(n-1)$-dimensional manifold. If it is empty, one obtains a *real n-manifold without boundary*. Such manifold is called *closed* if it is compact, and *open*, otherwise.

An open set of M^n together with a homeomorphism between the open set and an open set of $\mathbb{E}^n$ is called a *coordinate chart*. A collection of charts which cover M^n is an *atlas* on M^n. The homeomorphisms of two overlapping charts provide a transition mapping from a subset of $\mathbb{E}^n$ to some other subset of $\mathbb{E}^n$.

If all these mappings are continuously differentiable, then M^n is a *differentiable manifold*. If they are k times (infinitely often) continuously differentiable, then the manifold is a C^k *manifold* (respectively, a *smooth manifold*, or C^∞ *manifold*).

An atlas of a manifold is called *oriented* if the Jacobians of the coordinate transformations between any two charts are positive at every point. An *orientable manifold* is a manifold admitting an oriented atlas.

Manifolds inherit many local properties of the Euclidean space: they are locally path-connected, locally compact, and locally metrizable. Every smooth Riemannian manifold embeds isometrically (Nash, 1956) in some finite-dimensional Euclidean space.

Associated with every point on a differentiable manifold is a *tangent space* and its dual, a *cotangent space*. Formally, let M^n be a C^k manifold, $k \geq 1$, and p a point of M^n. Fix a chart $\varphi : U \to \mathbb{E}^n$, where U is an open subset of M^n containing p. Suppose that two curves $\gamma^1 : (-1,1) \to M^n$ and $\gamma^2 : (-1,1) \to M^n$ with $\gamma^1(0) = \gamma^2(0) = p$ are given such that $\varphi \cdot \gamma^1$ and $\varphi \cdot \gamma^2$ are both differentiable at 0.

Then γ^1 and γ^2 are called *tangent at* 0 if $(\varphi \cdot \gamma^1)'(0)= (\varphi \cdot \gamma^2)'(0)$. If the functions $\varphi \cdot \gamma^i : (-1,1) \to \mathbb{E}^n$, $i = 1,2$, are given by n real-valued component functions $(\varphi \cdot \gamma^i)_1(t), \dots, (\varphi \cdot \gamma^i)_n(t)$, the condition above means that their Jacobians $\left(\frac{d(\varphi\cdot\gamma^i)_1(t)}{dt}, \dots, \frac{d(\varphi\cdot\gamma^i)_n(t)}{dt}\right)$ coincide at 0. This is an equivalence relation, and the equivalence class $\gamma'(0)$ of the curve γ is called a *tangent vector* of M^n at p.

[redacted]

The *tangent space* $T_p(M^n)$ of M^n at p is defined as the set of all tangent vectors at p. The function $(d\varphi)_p : T_p(M^n) \to \mathbb{E}^n$ defined by $(d\varphi)_p(\gamma'(0)) = (\varphi \cdot \gamma)'(0)$, is bijective and can be used to transfer the vector space operations from $\mathbb{E}^n$ over to $T_p(M^n)$.

All the tangent spaces $T_p(M^n), p \in M^n$, when "glued together", form the *tangent bundle* $T(M^n)$ of M^n. Any element of $T(M^n)$ is a pair (p, v), where $v \in T_p(M^n)$.

If for an open neighborhood U of p the function $\varphi : U \to \mathbb{R}^n$ is a coordinate chart, then the preimage V of U in $T(M^n)$ admits a mapping $\psi : V \to \mathbb{R}^n \times \mathbb{R}^n$ defined by $\psi(p, v) = (\varphi(p), d\varphi(p))$. It defines the structure of a smooth $2n$-dimensional manifold on $T(M^n)$. The *cotangent bundle* $T^*(M^n)$ of M^n is obtained in similar manner using cotangent spaces $T_p^*(M^n), p \in M^n$.

A *vector field* on a manifold M^n is a *section* of its tangent bundle $T(M^n)$, i.e., a smooth function $f : M^n \to T(M^n)$ which assigns to every point $p \in M^n$ a vector $v \in T_p(M^n)$.

A *connection* (or *covariant derivative*) is a way of specifying a derivative of a vector field along another vector field on a manifold.

Formally, the covariant derivative ∇ of a vector u (defined at a point $p \in M^n$) in the direction of the vector v (defined at the same point p) is a rule that defines a third vector at p, called $\nabla_v u$ which has the properties of a derivative. A Riemannian metric uniquely defines a special covariant derivative called the *Levi-Civita connection*. It is the torsion-free connection ∇ of the tangent bundle, preserving the given Riemannian metric.

The *Riemann curvature tensor* R is the standard way to express the *curvature* of *Riemannian manifolds*. The Riemann curvature tensor can be given in terms of the Levi-Civita connection ∇ by the following formula:

$$R(u, v)w = \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w,$$

where $R(u, v)$ is a linear transformation of the tangent space of the manifold M^n; it is linear in each argument. If $u = \frac{\partial}{\partial x_i}$ and $v = \frac{\partial}{\partial x_j}$ are coordinate vector fields, then $[u, v] = 0$, and the formula simplifies to $R(u, v)w = \nabla_u \nabla_v w - \nabla_v \nabla_u w$, i.e., the curvature tensor measures anti-commutativity of the covariant derivative. The linear transformation $w \to R(u, v)w$ is also called the *curvature transformation*.

The *Ricci curvature tensor* (or *Ricci curvature*) *Ric* is obtained as the trace of the full curvature tensor R. It can be thought of as a Laplacian of the Riemannian metric tensor in the case of Riemannian manifolds. Ricci curvature is a linear operator on the tangent space at a point. Given an orthonormal basis $(e_i)_i$ in the tangent space $T_p(M^n)$, we have

$$Ric(u) = \sum_i R(u, e_i)e_i.$$

The value of $Ric(u)$ does not depend on the choice of an orthonormal basis. Starting with dimension four, the Ricci curvature does not describe the curvature tensor completely.

The *Ricci scalar* (or *scalar curvature*) Sc of a Riemannian manifold M^n is the full trace of the curvature tensor; given an orthonormal basis $(e_i)_i$ at $p \in M^n$, we have

$$Sc = \sum_{i,j} \langle R(e_i, e_j)e_j, e_i \rangle = \sum_i \langle Ric(e_i), e_i \rangle.$$

The *sectional curvature* $K(\sigma)$ of a Riemannian manifold M^n is defined as the *Gauss curvature* of an *σ-section* at a point $p \in M^n$, where a *σ-section* is a locally-defined piece of surface which has the 2-plane σ as a tangent plane at p, obtained from geodesics which start at p in the directions of the image of σ under the exponential mapping.

- **Metric tensor**

  The **metric** (or *basic*, *fundamental*) **tensor** is a symmetric tensor of rank 2, that is used to measure distances and angles in a real n-dimensional differentiable manifold M^n. Once a local coordinate system $(x_i)_i$ is chosen, the metric tensor appears as a real symmetric $n \times n$ matrix $((g_{ij}))$.

  The assignment of a metric tensor on M^n introduces a *scalar product* (i.e., symmetric bilinear, but in general not positive-definite, form) $\langle , \rangle_p$ on the tangent space $T_p(M^n)$ at any $p \in M^n$ defined by

  $$\langle x, y \rangle_p = g_p(x, y) = \sum_{i,j} g_{ij}(p) x_i y_j,$$

  where $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n) \in T_p(M^n)$. The collection of all these scalar products is called the **metric** g with the metric tensor $((g_{ij}))$. The length ds of the vector $(dx_1, \dots, dx_n)$ is expressed by the quadratic differential form

  $$ds^2 = \sum_{i,j} g_{ij} dx_i dx_j,$$

  which is called the *line element* (or *first fundamental form*) of the metric g.

  The *length* of a curve γ is expressed by the formula $\int_\gamma \sqrt{\sum_{i,j} g_{ij} dx_i dx_j}$. In general it may be real, purely imaginary, or zero (an *isotropic curve*).

  Let p, q and r be the numbers of positive, negative and zero *eigenvalues* of the matrix $((g_{ij}))$ of the metric g; so, $p+q+r = n$. The **metric signature** (or, simply, *signature*) of g is the pair (p, q). A **nondegenerated metric** (i.e., one with $r = 0$) is Riemannian or pseudo-Riemannian if its signature is *positive-definite* ($q = 0$) or *indefinite* ($pq > 0$), respectively.

  The **nonmetricity tensor** is the *covariant derivative* of a metric tensor. It is 0 for **Riemannian metrics** but can be $\neq 0$ for **pseudo-Riemannian** ones.

[redacted]

- **Nondegenerate metric**

  A **nondegenerate metric** is a metric g with the metric tensor $((g_{ij}))$, for which the *metric discriminant* $\det((g_{ij})) \neq 0$. All Riemannian and pseudo-Riemannian metrics are nondegenerate.

  A **degenerate metric** is a metric g with $\det((g_{ij})) = 0$ (cf. **semi-Riemannian metric** and **semi-pseudo-Riemannian metric**). A manifold with a degenerate metric is called an *isotropic manifold*.

- **Diagonal metric**

  A **diagonal metric** is a metric g with a metric tensor $((g_{ij}))$ which is zero for $i \neq j$. The Euclidean metric is a diagonal metric, as its metric tensor has the form $g_{ii} = 1, g_{ij} = 0$ for $i \neq j$.

- **Riemannian metric**

  Consider a real n-dimensional differentiable manifold M^n in which each tangent space is equipped with an *inner product* (i.e., a symmetric positive-definite bilinear form) which varies smoothly from point to point.

  A **Riemannian metric** on M^n is a collection of inner products $\langle , \rangle_p$ on the tangent spaces $T_p(M^n)$, one for each $p \in M^n$.

  Every inner product $\langle , \rangle_p$ is completely defined by inner products $\langle e_i, e_j \rangle_p = g_{ij}(p)$ of elements $e_1, \dots, e_n$ of a standard basis in $\mathbb{E}^n$, i.e., by the real symmetric and positive-definite $n \times n$ matrix $((g_{ij})) = ((g_{ij}(p)))$, called a **metric tensor**. In fact, $\langle x, y \rangle_p = \sum_{i,j} g_{ij}(p) x_i y_j$, where $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n) \in T_p(M^n)$. The smooth function g completely determines the Riemannian metric.

  A Riemannian metric on M^n is not an ordinary metric on M^n. However, for a connected manifold M^n, every Riemannian metric on M^n induces an ordinary metric on M^n, in fact, the **intrinsic metric** of M^n,

  For any points $p, q \in M^n$ the **Riemannian distance** between them is defined as

$$\inf_{\gamma} \int_0^1 \langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \rangle^{\frac{1}{2}} dt = \inf_{\gamma} \int_0^1 \sqrt{\sum_{i,j} g_{ij} \frac{dx_i}{dt} \frac{dx_j}{dt}} dt,$$

  where the infimum is over all rectifiable curves $\gamma : [0, 1] \to M^n$, connecting p and q.

  A *Riemannian manifold* (or *Riemannian space*) is a real n-dimensional differentiable manifold M^n equipped with a Riemannian metric. The theory of Riemannian spaces is called *Riemannian Geometry*. The simplest examples of Riemannian spaces are Euclidean spaces, *hyperbolic spaces*, and *elliptic spaces*.

- **Conformal metric**

  A *conformal structure on a vector space V* is a class of pairwise-homothetic Euclidean metrics on V. Any Euclidean metric d_E on V defines a conformal structure $\{\lambda d_E : \lambda > 0\}$.

  A *conformal structure on a manifold* is a field of conformal structures on the tangent spaces or, equivalently, a class of *conformally equivalent Riemannian metrics*. Two Riemannian metrics g and h on a smooth manifold M^n are called *conformally equivalent* if $g = f \cdot h$ for some positive function f on M^n, called a *conformal factor*.

A **conformal metric** is a Riemannian metric that represents the conformal structure. Cf. **conformally invariant metric** in Chap. 8.

- **Conformal space**

The **conformal space** (or *inversive space*) is the Euclidean space $\mathbb{E}^n$ extended by an ideal point (at infinity). Under *conformal* transformations, i.e., continuous transformations preserving local angles, the ideal point can be taken to be an ordinary point. Therefore, in a conformal space a sphere is indistinguishable from a plane: a plane is a sphere passing through the ideal point.

Conformal spaces are considered in *Conformal* (or *angle-preserving*, *Möbius*) *Geometry* in which properties of figures are studied that are invariant under conformal transformations. It is the set of transformations that map spheres into spheres, i.e., generated by the Euclidean transformations together with *inversions* which in coordinate form are conjugate to $x_i \to \frac{r^2 x_i}{\sum_j x_j^2}$, where r is the radius of the inversion. An inversion in a sphere becomes an everywhere well defined automorphism of period two. Any angle inverts into an equal angle.

The 2D conformal space is the *Riemann sphere*, on which the conformal transformations are given by the *Möbius transformations* $z \to \frac{az+b}{cz+d}$, $ad - bc \neq 0$.

In general, a **conformal mapping** between two Riemannian manifolds is a diffeomorphism between them such that the pulled back metric is *conformally equivalent* to the original one. A *conformal Euclidean space* is a *Riemannian space* admitting a conformal mapping onto an Euclidean space.

In the General Theory of Relativity, conformal transformations are considered on the *Minkowski space* $\mathbb{R}^{1,3}$ extended by two ideal points.

- **Space of constant curvature**

A **space of constant curvature** is a *Riemannian space* M^n for which the sectional curvature $K(\sigma)$ is constant in all 2D directions σ.

A *space form* is a connected complete space of constant curvature k. Examples of a *flat space form*, i.e., with $k = 0$, are the Euclidean space and flat torus. The sphere and hyperbolic space are space forms with $k > 0$ and $k < 0$, respectively.

- **Generalized Riemannian space**

A **generalized Riemannian space** is a metric space with the **intrinsic metric**, subject to certain restrictions on the curvature. Such spaces include *spaces of bounded curvature*, *Riemannian spaces*, etc. They are defined and investigated on the basis of their metric alone, without coordinates.

A *space of bounded curvature* ($\leq k$ and $\geq k'$) is defined by the condition: for any sequence of *geodesic triangles* T_n contracting to a point, we have

$$k \geq \overline{\lim} \frac{\overline{\delta}(T_n)}{\sigma(T_n^0)} \geq \underline{\lim} \frac{\overline{\delta}(T_n)}{\sigma(T_n^0)} \geq k',$$

where a *geodesic triangle* $T = xyz$ is the triplet of geodesic segments $[x, y]$, $[y, z]$, $[z, x]$ (the sides of T) connecting in pairs three different points x, y, z, $\overline{\delta}(T) =$

[redacted]

$\alpha + \beta + \gamma - \pi$ is the *excess* of the geodesic triangle T, and $\sigma(T^0)$ is the area of a Euclidean triangle T^0 with the sides of the same lengths. The **intrinsic metric** on the space of bounded curvature is called a **metric of bounded curvature**.

Such a space turns out to be Riemannian under two additional conditions: local compactness of the space (this ensures the condition of local existence of geodesics), and local extendability of geodesics. If in this case $k = k'$, it is a Riemannian space of constant curvature k (cf. **space of geodesics** in Chap. 6).

A space of curvature $\leq k$ is defined by the condition $\overline{\lim}\frac{\overline{\delta}(T_n)}{\sigma(T_n^0)} \leq k$. In such a space any point has a neighborhood in which the sum $\alpha + \beta + \gamma$ of the angles of a geodesic triangle T does not exceed the sum $\alpha_k + \beta_k + \gamma_k$ of the angles of a triangle T^k with sides of the same lengths in a space of constant curvature k. The intrinsic metric of such space is called a k**-concave metric**.

A space of curvature $\geq k$ is defined by the condition $\underline{\lim}\frac{\overline{\delta}(T_n)}{\sigma(T_n^0)} \geq k$. In such a space any point has a neighborhood in which $\alpha + \beta + \gamma \geq \alpha_k + \beta_k + \gamma_k$ for triangles T and T^k. The intrinsic metric of such space is called a K**-concave metric**.

An *Alexandrov metric space* is a generalized Riemannian space with upper, lower or integral curvature bounds. Cf. a **CAT(**κ_1**) space** in Chap. 6.

- **Complete Riemannian metric**

  A Riemannian metric g on a manifold M^n is called **complete** if M^n forms a complete metric space with respect to g.

  Any Riemannian metric on a compact manifold is complete.

- **Ricci-flat metric**

  A **Ricci-flat metric** is a Riemannian metric with vanished Ricci curvature tensor.

  A *Ricci-flat manifold* is a Riemannian manifold equipped with a Ricci-flat metric. Ricci-flat manifolds represent vacuum solutions to the *Einstein field equation*, and are special cases of *Kähler–Einstein manifolds*. Important Ricci-flat manifolds are *Calabi–Yau manifolds*, and *hyper-Kähler manifolds*.

- **Osserman metric**

  An **Osserman metric** is a Riemannian metric for which the Riemannian curvature tensor R is *Osserman*, i.e., the eigenvalues of the *Jacobi operator* $\mathcal{J}(x) : y \rightarrow R(y, x)x$ are constant on the *unit sphere* S^{n-1} in $\mathbb{E}^n$ (they are independent of the unit vectors x).

- G**-invariant Riemannian metric**

  Given a *Lie group* $(G, \cdot, id)$ of transformations, a Riemannian metric g on a differentiable manifold M^n is called G**-invariant**, if it does not change under any $x \in G$. The group $(G, \cdot, id)$ is called the *group of motions* (or *group of isometries*) of the Riemannian space (M^n, g). Cf. G**-invariant metric** in Chap. 10.

- **Ivanov–Petrova metric**

  Let R be the Riemannian curvature tensor of a Riemannian manifold M^n, and let $\{x, y\}$ be an orthogonal basis for an oriented 2-plane π in the tangent space $T_p(M^n)$ at a point p of M^n.

The **Ivanov–Petrova metric** is a Riemannian metric on M^n for which the eigenvalues of the antisymmetric curvature operator $\mathcal{R}(\pi) = R(x, y)$ ([IvSt95]) depend only on the point p of a Riemannian manifold M^n, but not upon the plane π.

- **Zoll metric**

  A **Zoll metric** is a Riemannian metric on a smooth manifold M^n whose geodesics are all simple closed curves of an equal length. A 2D sphere S^2 admits many such metrics, besides the obvious metrics of constant curvature. In terms of cylindrical coordinates (z, θ) ($z \in [-1, 1]$, $\theta \in [0, 2\pi]$), the *line element*

$$ds^2 = \frac{(1 + f(z))^2}{1 - z^2} dz^2 + (1 - z^2) d\theta^2$$

  defines a Zoll metric on S^2 for any smooth odd function $f : [-1, 1] \to (-1, 1)$ which vanishes at the endpoints of the interval.

- **Berger metric**

  The **Berger metric** is a Riemannian metric on the *Berger sphere* (i.e., the three-sphere S^3 squashed in one direction) defined by the *line element*

$$ds^2 = d\theta^2 + \sin^2\theta d\phi^2 + \cos^2\alpha(d\psi + \cos\theta d\phi)^2,$$

  where α is a constant, and θ, ϕ, ψ are *Euler angles*.

- **Cycloidal metric**

  The **cycloidal metric** is a Riemannian metric on the half-plane $\mathbb{R}^2_+ = \{x \in \mathbb{R}^2 : x_2 > 0\}$ defined by the *line element*

$$ds^2 = \frac{dx_1^2 + dx_2^2}{2x_2}.$$

  It is called *cycloidal* because its geodesics are cycloid curves. The corresponding distance $d(x, y)$ between two points $x, y \in \mathbb{R}^2_+$ is equivalent to the distance

$$\rho(x, y) = \frac{|x_1 - y_1| + |x_2 - y_2|}{\sqrt{x_1} + \sqrt{x_2} + \sqrt{|x_2 - y_2|}}$$

  in the sense that $d \le C\rho$, and $\rho \le Cd$ for some positive constant C.

- **Klein metric**

  The **Klein metric** is a Riemannian metric on the *open unit ball* $B^n = \{x \in \mathbb{R}^n : ||x||_2 < 1\}$ in $\mathbb{R}^n$ defined by

$$\frac{\sqrt{||y||_2^2 - (||x||_2^2||y||_2^2 - \langle x, y\rangle^2)}}{1 - ||x||_2^2}$$

[redacted]

for any $x \in B^n$ and $y \in T_x(B^n)$, where $||.||_2$ is the Euclidean norm on $\mathbb{R}^n$, and $\langle , \rangle$ is the ordinary *inner product* on $\mathbb{R}^n$.

The Klein metric is the hyperbolic case $a = -1$ of the general form

$$\frac{\sqrt{(1 + a||x||^2)||y||^2 - a\langle x, y\rangle^2}}{1 + a||x||^2},$$

while $a = 0, 1$ correspond to the Euclidean and spherical cases.

- **Carnot–Carathéodory metric**

  A *distribution* (or *polarization*) on a manifold M^n is a subbundle of the tangent bundle $T(M^n)$ of M^n. Given a distribution $H(M^n)$, a vector field in $H(M^n)$ is called *horizontal*. A curve γ on M^n is called *horizontal* (or *distinguished*, *admissible*) with respect to $H(M^n)$ if $\gamma'(t) \in H_{\gamma(t)}(M^n)$ for any t.

  A distribution $H(M^n)$ is called *completely nonintegrable* if the Lie brackets of $H(M^n)$, i.e., $[\cdots, [H(M^n), H(M^n)]]$, span the tangent bundle $T(M^n)$, i.e., for all $p \in M^n$ any tangent vector v from $T_p(M^n)$ can be presented as a linear combination of vectors of the following types: u, $[u, w]$, $[u, [w, t]]$, $[u, [w, [t, s]]], \cdots \in T_p(M^n)$, where all vector fields $u, w, t, s, \ldots$ are horizontal.

  The **Carnot–Carathéodory metric** (or **CC metric**, **sub-Riemannian metric**, *control metric*) is a metric on a manifold M^n with a completely nonintegrable horizontal distribution $H(M^n)$ defined as the section g_C of positive-definite *scalar products* on $H(M^n)$. The distance $d_C(p, q)$ between any points $p, q \in M^n$ is defined as the infimum of the g_C-lengths of the horizontal curves joining p and q.

  A *sub-Riemannian manifold* (or *polarized manifold*) is a manifold M^n equipped with a Carnot–Carathéodory metric. It is a generalization of a Riemannian manifold. Roughly, in order to measure distances in a sub-Riemannian manifold, one is allowed to go only along curves tangent to horizontal spaces.

- **Pseudo-Riemannian metric**

  Consider a real n-dimensional differentiable manifold M^n in which every tangent space $T_p(M^n)$, $p \in M^n$, is equipped with a *scalar product* which varies smoothly from point to point and is nondegenerate, but indefinite.

  A **pseudo-Riemannian metric** on M^n is a collection of scalar products $\langle , \rangle_p$ on the tangent spaces $T_p(M^n)$, $p \in M^n$, one for each $p \in M^n$.

  Every scalar product $\langle , \rangle_p$ is completely defined by scalar products $\langle e_i, e_j\rangle_p = g_{ij}(p)$ of elements $e_1, \ldots, e_n$ of a standard basis in $\mathbb{E}^n$, i.e., by the real symmetric indefinite $n \times n$ matrix $((g_{ij})) = ((g_{ij}(p)))$, called a **metric tensor** (cf. **Riemannian metric** in which case this tensor is not only nondegenerate but, moreover, positive-definite).

  In fact, $\langle x, y\rangle_p = \sum_{i,j} g_{ij}(p)x_i y_j$, where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n) \in T_p(M^n)$. The smooth function g determines the pseudo-Riemannian metric.

The length ds of the vector $(dx_1, \dots, dx_n)$ is given by the quadratic differential form

$$ds^2 = \sum_{i,j} g_{ij} dx_i dx_j.$$

The length of a curve $\gamma : [0, 1] \to M^n$ is expressed by the formula

$$\int_\gamma \sqrt{\sum_{i,j} g_{ij} dx_i dx_j} = \int_0^1 \sqrt{\sum_{i,j} g_{ij} \frac{dx_i}{dt} \frac{dx_j}{dt}} dt.$$

In general it may be real, purely imaginary or zero (an *isotropic curve*).

A pseudo-Riemannian metric on M^n is a metric with a fixed, but indefinite signature (p, q), $p + q = n$. A pseudo-Riemannian metric is nondegenerate, i.e., its metric discriminant $\det((g_{ij})) \neq 0$. Therefore, it is a **nondegenerate indefinite metric**.

A *pseudo-Riemannian manifold* (or *pseudo-Riemannian space*) is a real n-dimensional differentiable manifold M^n equipped with a pseudo-Riemannian metric. The theory of pseudo-Riemannian spaces is called *Pseudo-Riemannian Geometry*.

- **Pseudo-Euclidean distance**

The model space of a **pseudo-Riemannian space** of signature (p, q) is the *pseudo-Euclidean space* $\mathbb{R}^{p,q}$, $p + q = n$ which is a real n-dimensional vector space $\mathbb{R}^n$ equipped with the metric tensor $((g_{ij}))$ of signature (p, q) defined, for $i \neq j$, by $g_{11} = \cdots = g_{pp} = 1$, $g_{p+1,p+1} = \cdots = g_{nn} = -1$, $g_{ij} = 0$.

The *line element* of the corresponding metric is given by

$$ds^2 = dx_1^2 + \cdots + dx_p^2 - dx_{p+1}^2 - \cdots - dx_n^2.$$

The **pseudo-Euclidean distance** of signature $(p, q = n - p)$ on $\mathbb{R}^n$ is defined by

$$d_{pE}^2(x, y) = D(x, y) = \sum_{i=1}^{p} (x_i - y_i)^2 - \sum_{i=p+1}^{n} (x_i - y_i)^2.$$

Such a pseudo-Euclidean space can be seen as $\mathbb{R}^p \times i\mathbb{R}^q$, where $i = \sqrt{-1}$.

The pseudo-Euclidean space with $(p, q) = (1, 3)$ is used as flat space-time model of Special Relativity; cf. **Minkowski metric** in Chap. 26.

The points correspond to *events*; the line spanned by x and y is *space-like* if $D(x, y) > 0$ and *time-like* if $D(x, y) < 0$. If $D(x, y) > 0$, then $\sqrt{D(x, y)}$ is Euclidean distance and if $D(x, y) < 0$, then $\sqrt{|D(x, y)|}$ is the lifetime of a particle (from x to y).

[redacted]

The pseudo-Euclidean distance of signature $(p, q = n - p)$ is the case $A = \mathrm{diag}(a_i)$ with $a_i = 1$ for $1 \le i \le p$ and $a_i = -1$ for $p + 1 \le i \le n$, of the **weighted Euclidean distance** $\sqrt{\sum_{1 \le i \le n} a_i(x_i - y_i)^2}$ in Chap. 17.

- **Blaschke metric**

  The **Blaschke metric** on a nondegenerate hypersurface is a pseudo-Riemannian metric, associated to the affine normal of the immersion $\phi : M^n \to \mathbb{R}^{n+1}$, where M^n is an n-dimensional manifold, and $\mathbb{R}^{n+1}$ is considered as an affine space.

- **Semi-Riemannian metric**

  A **semi-Riemannian metric** on a real n-dimensional differentiable manifold M^n is a degenerate Riemannian metric, i.e., a collection of positive-semidefinite *scalar products* $\langle x, y \rangle_p = \sum_{i,j} g_{ij}(p) x_i y_j$ on the tangent spaces $T_p(M^n)$, $p \in M^n$; the metric discriminant $det((g_{ij})) = 0$.

  A *semi-Riemannian manifold* (or *semi-Riemannian space*) is a real n-dimensional differentiable manifold M^n equipped with a semi-Riemannian metric.

  The model space of a semi-Riemannian manifold is the *semi-Euclidean space* R^n_d, $d \ge 1$ (sometimes denoted also by $\mathbb{R}^n_{n-d}$), i.e., a real n-dimensional vector space $\mathbb{R}^n$ equipped with a semi-Riemannian metric.

  It means that there exists a scalar product of vectors such that, relative to a suitably chosen basis, the scalar product $\langle x, x \rangle$ has the form $\langle x, x \rangle = \sum_{i=1}^{n-d} x_i^2$. The number $d \ge 1$ is called the *defect* (or *deficiency*) of the space.

- **Grushin metric**

  The **Grushin metric** is a semi-Riemannian metric on $\mathbb{R}^2$ defined by the *line element*

$$ds^2 = dx_1^2 + \frac{dx_2^2}{x_1^2}.$$

- **Agmon distance**

  The **Agmon metric** attached to an energy E and a potential V is defined as

$$ds^2 = \max\{0, V(x) - E_0(h)\} dx^2,$$

  where dx^2 is the standard metric on $\mathbb{R}^d$. Then the **Agmon distance** on $\mathbb{R}^d$ is the corresponding Riemannian distance defined, for any $x, y \in \mathbb{R}^d$, by

$$\inf_{\gamma}\{\int_0^1 \sqrt{\max\{V(\gamma(s)) - E_0(h), 0\}} \cdot |\gamma'(s)| ds : \gamma(0) = x, \gamma(1) = y, \gamma \in C^1\}.$$

- **Semi-pseudo-Riemannian metric**

  A **semi-pseudo-Riemannian metric** on a real n-dimensional differentiable manifold M^n is a degenerate pseudo-Riemannian metric, i.e., a collection of degenerate indefinite *scalar products* $\langle x, y \rangle_p = \sum_{i,j} g_{ij}(p) x_i y_j$ on the tangent spaces $T_p(M^n)$, $p \in M^n$; the metric discriminant $det((g_{ij})) = 0$. In fact, a semi-pseudo-Riemannian metric is a **degenerate indefinite metric**.

A *semi-pseudo-Riemannian manifold* (or *semi-pseudo-Riemannian space*) is a real n-dimensional differentiable manifold M^n equipped with a semi-pseudo-Riemannian metric. The model space of such manifold is the *semi-pseudo-Euclidean space* $\underset{m_1,\dots,m_{r-1}}{\mathbb{R}^n_{l_1,\dots,l_r}}$, i.e., a vector space $\mathbb{R}^n$ equipped with a semi-pseudo-Riemannian metric.

It means that there exist r scalar products $\langle x, y\rangle_a = \sum \epsilon_{i_a} x_{i_a} y_{i_a}$, where $a = 1, \dots r$, $0 = m_0 < m_1 < \cdots < m_r = n$, $i_a = m_{a-1} + 1, \dots m_a$, $\epsilon_{i_a} = \pm 1$, and -1 occurs l_a times among the numbers ϵ_{i_a}. The product $\langle x, y\rangle_a$ is defined for those vectors for which all coordinates x_i, $i \leq m_{a-1}$ or $i > m_a + 1$ are zero.

The first scalar square of an arbitrary vector x is a degenerate quadratic form $\langle x, x\rangle_1 = -\sum_{i=1}^{l_1} x_i^2 + \sum_{j=l_1+1}^{n-d} x_j^2$. The number $l_1 \geq 0$ is called the *index*, and the number $d = n - m_1$ is called the *defect* of the space. If $l_1 = \cdots = l_r = 0$, we obtain a *semi-Euclidean space*. The spaces $\underset{m}{\mathbb{R}^n}$ and $\underset{m}{\mathbb{R}^n_{k,l}}$ are called *quasi-Euclidean spaces*.

The *semi-pseudo-non-Euclidean space* $\underset{m_1,\dots,m_{r-1}}{\mathbb{S}^n_{l_1,\dots,l_r}}$ is a hypersphere in $\underset{m_1,\dots,m_{r-1}}{\mathbb{R}^{n+1}_{l_1,\dots,l_r}}$ with identified antipodal points. It is called *semielliptic* (or *semi-non-Euclidean*) *space* if $l_1 = \cdots = l_r = 0$ and a *semihyperbolic space* if there exist $l_i \neq 0$.

- **Finsler metric**

Consider a real n-dimensional differentiable manifold M^n in which every tangent space $T_p(M^n)$, $p \in M^n$, is equipped with a *Banach norm* $||.||$ such that the Banach norm as a function of position is smooth, and the matrix $((g_{ij}))$,

$$g_{ij} = g_{ij}(p, x) = \frac{1}{2}\frac{\partial^2 ||x||^2}{\partial x_i \partial x_j},$$

is positive-definite for any $p \in M^n$ and any $x \in T_p(M^n)$.

A **Finsler metric** on M^n is a collection of Banach norms $||.||$ on the tangent spaces $T_p(M^n)$, one for each $p \in M^n$. Its *line element* has the form

$$ds^2 = \sum_{i,j} g_{ij} dx_i dx_j.$$

The Finsler metric can be given by *fundamental function*, i.e., a real positive-definite convex function $F(p, x)$ of $p \in M^n$ and $x \in T_p(M^n)$ acting at the point p. $F(p, x)$ is positively homogeneous of degree one in x: $F(p, \lambda x) = \lambda F(p, x)$ for every $\lambda > 0$. Then $F(p, x)$ is the length of the vector x.

The *Finsler metric tensor* has the form $((g_{ij})) = ((\frac{1}{2}\frac{\partial^2 F^2(p,x)}{\partial x_i \partial x_j}))$. The length of a curve $\gamma : [0, 1] \to M^n$ is given by $\int_0^1 F(p, \frac{dp}{dt})dt$. For each fixed p the Finsler metric tensor is Riemannian in the variables x.

[redacted]

The Finsler metric is a generalization of the Riemannian metric, where the general definition of the length $||x||$ of a vector $x \in T_p(M^n)$ is not necessarily given in the form of the square root of a symmetric bilinear form as in the Riemannian case.

A *Finsler manifold* (or *Finsler space*) is a real differentiable n-manifold M^n equipped with a Finsler metric. The theory of such spaces is *Finsler Geometry*.

The difference between a Riemannian space and a Finsler space is that the former behaves locally like a Euclidean space, and the latter locally like a *Minkowskian space* or, analytically, the difference is that to an ellipsoid in the Riemannian case there corresponds an arbitrary convex surface which has the origin as the center.

A **pseudo-Finsler metric** F is defined by weakening the definition of a Finsler metric): $((g_{ij}))$ should be nondegenerate and of constant signature (not necessarily positive-definite) and hence F could be negative. The pseudo-Finsler metric is a generalization of the pseudo-Riemannian metric.

- **(α, β)-metric**

  Let $\alpha(x, y) = \sqrt{\alpha_{ij}(x)y^i y^j}$ be a Riemannian metric and $\beta(x, y) = b_i(x)y^i$ be a 1-form on a n-dimensional manifold M^n. Let $s = \frac{\beta}{\alpha}$ and $\phi(s)$ is an C^∞-positive function on some symmetric interval $(-r, r)$ with $r > \frac{\beta}{\alpha}$ for all (x, y) in the tangent bundle $TM = \cup_{x \in M} T_x(M^n)$ of the tangent spaces $T_x(M^n)$. Then $F = \alpha\phi(s)$ is a Finsler metric (Matsumoto, 1972) called an **(α, β)-metric**. The main examples of (α, β)-metrics follow.

  The **Kropina metric** is the case $\phi(s) = \frac{1}{s}$, i.e., $F = \frac{\alpha^2}{\beta}$.
  The **generalized Kropina metric** is the case $\phi(s) = s^m$, i.e., $F = \beta^m \alpha^{1-m}$.
  The **Randers metric** (1941) is the case $\phi(s) = 1 + s$, i.e., $F = \alpha + \beta$.
  The **Matsumoto slope metric** is the case $\phi(s) = \frac{1}{1-s}$, i.e., $F = \frac{\alpha^2}{\alpha - \beta}$.
  The **Riemann-type (α, β)-metric** is the case $\phi(s) = \sqrt{1 + s^2}$, i.e., $F = \alpha^2 + \beta^2$.
  Park and Lee, 1998, considered the case $\phi(s) = 1 + s^2$, i.e., $F = \alpha + \frac{\beta^2}{\alpha}$.

- **Shen metric**

  Given a vector $a \in \mathbb{R}^n$, $||a||_2 < 1$, the **Shen metric** (2003) is a Finsler metric on the *open unit ball* $B^n = \{x \in \mathbb{R}^n : ||x||_2 < 1\}$ in $\mathbb{R}^n$ defined by

$$\frac{\sqrt{||y||_2^2 - (||x||_2^2 ||y||_2^2 - \langle x, y \rangle^2)} + \langle x, y \rangle}{1 - ||x||_2^2} + \frac{\langle a, y \rangle}{1 + \langle a, x \rangle}$$

  for any $x \in B^n$ and $y \in T_x(B^n)$, where $||.||_2$ is the Euclidean norm on $\mathbb{R}^n$, and $\langle , \rangle$ is the ordinary *inner product* on $\mathbb{R}^n$. It is a **Randers metric** and a **projective metric**. Cf. **Klein metric** and **Berwald metric**.

- **Berwald metric**

  The **Berwald metric** (1929) is a Finsler metric F_{Be} on the *open unit ball* $B^n = \{x \in \mathbb{R}^n : ||x||_2 < 1\}$ in $\mathbb{R}^n$ defined, for any $x \in B^n$ and $y \in T_x(B^n)$, by

$$\frac{\left(\sqrt{||y||_2^2 - (||x||_2^2||y||_2^2 - \langle x,y\rangle^2)} + \langle x,y\rangle\right)^2}{(1-||x||_2^2)^2\sqrt{||y||_2^2 - (||x||_2^2||y||_2^2 - \langle x,y\rangle^2)}},$$

  where $||.||_2$ is the Euclidean norm on $\mathbb{R}^n$, and $\langle,\rangle$ is the *inner product* on $\mathbb{R}^n$. It is a **projective metric** and an (α,β)**-metric** with $\phi(s) = (1+s)^2$, i.e., $F = \frac{(\alpha+\beta)^2}{\alpha}$.

  The Riemannian metrics are special Berwald metrics. Every Berwald metric is affinely equivalent to a Riemannian metric.

  In general, every Finsler metric on a manifold M^n induces a *spray* (second-order homogeneous ordinary differential equation) $y_i\frac{\partial}{\partial x_i} - 2G^i\frac{\partial}{\partial y_i}$ which determines the geodesics. A Finsler metric is a Berwald metric if the spray coefficients $G^i = G^i(x,y)$ are quadratic in $y \in T_x(M^n)$ at any point $x \in M^n$, i.e., $G^i = \frac{1}{2}\Gamma^i_{jk}(x)y^jy^k$.

  A Finsler metric is a more general **Landsberg metric** $\Gamma^i_{jk} = \frac{1}{2}\partial_{y^j}\partial_{y^k}(\Gamma^i_{jk}(x)y^jy^k)$. The Landsberg metric is the one for which the *Landsberg curvature* (the covariant derivative of the *Cartan torsion along a geodesic*) is zero.

- **Douglas metric**

  A **Douglas metric** a Finsler metric for which the *spray coefficients* $G^i = G^i(x,y)$ have the following form:

$$G^i = \frac{1}{2}\Gamma^i_{jk}(x)y_iy_k + P(x,y)y_i.$$

  Every Finsler metric which is projectively equivalent to a **Berwald metric** is a Douglas metric. Every **Berwald metric** is a Douglas metric. Every known Douglas metric is (locally) projectively equivalent to a Berwald metric.

- **Bryant metric**

  Let α be an angle with $|\alpha| < \frac{\pi}{2}$. Let, for any $x, y \in \mathbb{R}^n$, $A = ||y||_2^4\sin^2 2\alpha + \left(||y||_2^2\cos 2\alpha + ||x||_2^2||y||_2^2 - \langle x,y\rangle^2\right)^2$, $B = ||y||_2^2\cos 2\alpha + ||x||_2^2||y||_2^2 - \langle x,y\rangle^2$, $C = \langle x,y\rangle\sin 2\alpha$, $D = ||x||_2^4 + 2||x||_2^2\cos 2\alpha + 1$. Then we get a Finsler metric

$$\sqrt{\frac{\sqrt{A}+B}{2D} + \left(\frac{C}{D}\right)^2} + \frac{C}{D}.$$

  On the 2D *unit sphere* S^2, it is the **Bryant metric** (1996).

[redacted]

- ***m*-th root pseudo-Finsler metric**

  An **m-th root pseudo-Finsler metric** is (Shimada, 1979) a **pseudo-Finsler metric** defined (with $a_{i_1 \dots i_m}$ symmetric in all its indices) by

  $$F(x,y) = (a_{i_1 \dots i_m}(x) y^{i_1 \dots i_m})^{\frac{1}{m}}.$$

  For $m = 2$, it is a pseudo-Riemannian metric. The 3-rd and 4-th root pseudo-Finsler metrics are called *cubic metric* and *quartic metric*, respectively.

- **Antonelli–Shimada metric**

  The **Antonelli–Shimada metric** (or *ecological Finsler metric*) is an **m-th root pseudo-Finsler metric** defined, via linearly independent 1-forms a^i, by

  $$F(x,y) = (\sum_{i=1}^{n} (a^i)^m)^{\frac{1}{m}}.$$

  The **Uchijo metric** is defined, for a constant k, by

  $$F(x,y) = (\sum_{i=1}^{n} (a^i)^2)^{\frac{1}{2}} + ka^1.$$

- **Berwald–Moör metric**

  The **Berwald–Moör metric** is an **m-th root pseudo-Finsler metric**, defined by

  $$F(x,y) = (y^1 \dots y^n)^{\frac{1}{n}}.$$

  More general **Asanov metric** is defined, via linearly independent 1-forms a^i, by

  $$F(x,y) = (a^1 \dots a^n)^{\frac{1}{n}}.$$

  The Berwald–Moör metrics with $n = 4$ and $n = 6$ are applied in Relativity Theory and Diffusion Imaging, respectively. The pseudo-Finsler spaces which are locally isomorphic to the 4-th root Berwald–Moör metric, are expected to be more general and productive space-time models than usual pseudo-Riemannian spaces, which are locally isomorphic to the Minkowski metric.

- **Kawaguchi metric**

  The **Kawaguchi metric** is a metric on a smooth n-dimensional manifold M^n, given by the arc element ds of a regular curve $x = x(t)$, $t \in [t_0, t_1]$ via the formula

  $$ds = F(x, \frac{dx}{dt}, \dots, \frac{d^k x}{dt^k}) dt,$$

where the *metric function* F satisfies Zermelo's conditions: $\sum_{s=1}^{k} sx^{(s)}F_{(s)i} = F$, $\sum_{s=r}^{k}\binom{s}{k}x^{(s-r+1)i}F_{(s)i} = 0$, $x^{(s)i} = \frac{d^s x^i}{dt^s}$, $F_{(s)i} = \frac{\partial F}{\partial x^{(s)i}}$, and $r = 2, \dots, k$.

These conditions ensure that the arc element ds is independent of the parametrization of the curve $x = x(t)$.

A *Kawaguchi manifold* (or *Kawaguchi space*) is a smooth manifold equipped with a Kawaguchi metric. It is a generalization of a *Finsler manifold*.

- **Lagrange metric**

  Consider a real n-dimensional manifold M^n. A set of symmetric nondegenerated matrices $((g_{ij}(p,x)))$ define a **generalized Lagrange metric** on M^n if a change of coordinates $(p,x) \to (q,y)$, such that $q_i = q_i(p_1, \dots, p_n)$, $y_i = (\partial_j q_i)x_j$ and rank $(\partial_j q_i) = n$, implies $g_{ij}(p,x) = (\partial_i q_i)(\partial_j q_j)g_{ij}(q,y)$.

  A generalized Lagrange metric is called a **Lagrange metric** if there exists a *Lagrangian*, i.e., a smooth function $L(p,x)$ such that it holds

  $$g_{ij}(p,x) = \frac{1}{2} \cdot \frac{\partial^2 L(p,x)}{\partial x_i \partial x_j}.$$

  Every Finsler metric is a Lagrange metric with $L = F^2$.

- **DeWitt supermetric**

  The **DeWitt supermetric** (or **Wheeler–DeWitt supermetric**) $G = ((G_{ijkl}))$ calculates distances between metrics on a given manifold, and it is a generalization of a Riemannian (or pseudo-Riemannian) metric $g = ((g_{ij}))$.

  For example, for a given connected smooth 3-dimensional manifold M^3, consider the space $\mathcal{M}(M^3)$ of all Riemannian (or pseudo-Riemannian) metrics on M^3. Identifying points of $\mathcal{M}(M^3)$ that are related by a diffeomorphism of M^3, one obtains the space $Geom(M^3)$ of 3-*geometries* (of fixed topology), points of which are the classes of diffeomorphically equivalent metrics. The space $Geom(M^3)$ is called a *superspace*. It plays an important role in several formulations of Quantum Gravity.

  A **supermetric**, i.e., a "metric on metrics", is a metric on $\mathcal{M}(M^3)$ (or on $Geom(M^3)$) which is used for measuring distances between metrics on M^3 (or between their equivalence classes). Given $g = ((g_{ij})) \in \mathcal{M}(M^3)$, we obtain

  $$||\delta g||^2 = \int_{M^3} d^3x G^{ijkl}(x)\delta g_{ij}(x)\delta g_{kl}(x),$$

  where G^{ijkl} is the inverse of the **DeWitt supermetric**

  $$G_{ijkl} = \frac{1}{2\sqrt{det((g_{ij}))}}(g_{ik}g_{jl} + g_{il}g_{jk} - \lambda g_{ij}g_{kl}).$$

  The value λ parametrizes the distance between metrics in $\mathcal{M}(M^3)$, and may take any real value except $\lambda = \frac{2}{3}$, for which the supermetric is *singular*.

[redacted]

- **Lund–Regge supermetric**

  The **Lund–Regge supermetric** (or **simplicial supermetric**) is an analog of the **DeWitt supermetric**, used to measure the distances between *simplicial* 3-*geometries* in a *simplicial configuration space*.

  More exactly, given a closed *simplicial* 3D manifold M^3 consisting of several *tetrahedra* (i.e., 3-*simplices*), a *simplicial geometry* on M^3 is fixed by an assignment of values to the squared edge lengths of M^3, and a flat Riemannian Geometry to the interior of each tetrahedron consistent with those values.

  The squared edge lengths should be positive and constrained by the triangle inequalities and their analogs for the tetrahedra, i.e., all squared measures (lengths, areas, volumes) must be nonnegative (cf. **tetrahedron inequality** in Chap. 3).

  The set $\mathcal{T}(M^3)$ of all simplicial geometries on M^3 is called a *simplicial configuration space*. The Lund–Regge supermetric $((G_{mn}))$ on $\mathcal{T}(M^3)$ is induced from the DeWitt supermetric on $\mathcal{M}(M^3)$, using for representations of points in $\mathcal{T}(M^3)$ such metrics in $\mathcal{M}(M^3)$ which are piecewise flat in the tetrahedra.

- **Space of Lorentz metrics**

  Let M^n be an n-dimensional compact manifold, and $\mathcal{L}(M^n)$ the set of all **Lorentz metrics** (i.e., the pseudo-Riemannian metrics of signature $(n-1,1)$) on M^n.

  Given a Riemannian metric g on M^n, one can identify the vector space $S^2(M^n)$ of all symmetric 2-tensors with the vector space of endomorphisms of the tangent to M^n which are symmetric with respect to g. In fact, if $\tilde{h}$ is the endomorphism associated to a tensor h, then the distance on $S^2(M^n)$ is given by

  $$d_g(h,t) = \sup_{x \in M^n} \sqrt{tr(\tilde{h}_x - \tilde{t}_x)^2}.$$

  The set $\mathcal{L}(M^n)$ taken with the distance d_g is an open subset of $S^2(M^n)$ called the **space of Lorentz metrics**. Cf. **manifold triangulation metric** in Chap. 9.

- **Perelman supermetric proof**

  The *Thurston's Geometrization Conjecture* is that, after two well-known splittings, any 3D manifold admits, as remaining components, only one of eight *Thurston model geometries*. If true, this conjecture implies the validity of the famous *Poincaré Conjecture* of 1904, that any 3-manifold, in which every simple closed curve can be deformed continuously to a point, is homeomorphic to the 3-sphere.

  In 2002, Perelman gave a gapless "sketch of an eclectic proof" of Thurston's conjecture using a kind of supermetric approach to the space of all Riemannian metrics on a given smooth 3-manifold. In a *Ricci flow* the distances decrease in directions of positive curvature since the metric is time-dependent. Perelman's modification of the standard Ricci flow permitted systematic elimination of arising singularities.

7.2 Riemannian Metrics in Information Theory

Some special Riemannian metrics are commonly used in Information Theory. A list of such metrics is given below.

- **Thermodynamic metrics**

  Given the space of all *extensive* (additive in magnitude, mechanically conserved) thermodynamic variables of a system (energy, entropy, amounts of materials), a **thermodynamic metric** is a Riemannian metric on the manifold of equilibrium states defined as the 2-nd derivative of one extensive quantity, usually entropy or energy, with respect to the other extensive quantities. This information geometric approach provides a geometric description of thermodynamic systems in equilibrium.

  The **Ruppeiner metric** (Ruppeiner, 1979) is defined by the *line element* $ds_R^2 = g_{ij}^R dx^i dx^j$, where the matrix $((g_{ij}^R))$ of the symmetric metric tensor is a negative *Hessian* (the matrix of 2-nd order partial derivatives) of the entropy function S:

  $$g_{ij}^R = -\partial_i \partial_j S(M, N^a).$$

  Here M is the internal energy (which is the mass in black hole applications) of the system and N^a refer to other extensive parameters such as charge, angular momentum, volume, etc. This metric is flat if and only if the statistical mechanical system is noninteracting, while curvature singularities are a signal of critical behavior, or, more precisely, of divergent **correlation lengths** (Chap. 24).

  The **Weinhold metric** (Weinhold, 1975) is defined by $g_{ij}^W = \partial_i \partial_j M(S, N^a)$.

  The Ruppeiner and Weinhold metrics are *conformally equivalent* (cf. **conformal metric**) via $ds^2 = g_{ij}^R dM^i dM^j = \frac{1}{T} g_{ij}^W dS^i dS^j$, where T is the temperature.

  The **thermodynamic length** in Chap. 24 is a path function that measures the distance along a path in the state space.

- **Fisher information metric**

  In Statistics, Probability, and Information Geometry, the **Fisher information metric** is a Riemannian metric for a statistical differential manifold (see, for example, [Amar85, Frie98]). Formally, let $p_\theta = p(x, \theta)$ be a family of densities, indexed by n parameters $\theta = (\theta_1, \dots, \theta_n)$ which form the *parameter manifold* P.

  The **Fisher information metric** $g = g_\theta$ on P is a Riemannian metric, defined by the *Fisher information matrix* $((I(\theta)_{ij}))$, where

  $$I(\theta)_{ij} = \mathbb{E}_\theta \left[\frac{\partial \ln p_\theta}{\partial \theta_i} \cdot \frac{\partial \ln p_\theta}{\partial \theta_j} \right] = \int \frac{\partial \ln p(x, \theta)}{\partial \theta_i} \frac{\partial \ln p(x, \theta)}{\partial \theta_j} p(x, \theta) dx.$$

  It is a symmetric bilinear form which gives a classical measure (*Rao measure*) for the statistical distinguishability of distribution parameters.

[redacted]

Putting $i(x, \theta) = -\ln p(x, \theta)$, one obtains an equivalent formula

$$I(\theta)_{ij} = \mathbb{E}_\theta \left[\frac{\partial^2 i(x, \theta)}{\partial\theta_i \partial\theta_j} \right] = \int \frac{\partial^2 i(x, \theta)}{\partial\theta_i \partial\theta_j} p(x, \theta) dx.$$

In a coordinate-free language, we get

$$I(\theta)(u, v) = \mathbb{E}_\theta \left[u(\ln p_\theta) \cdot v(\ln p_\theta) \right],$$

where u and v are vectors tangent to the parameter manifold P, and $u(\ln p_\theta) = \frac{d}{dt} \ln p_{\theta+tu|t=0}$ is the derivative of $\ln p_\theta$ along the direction u.

A *manifold of densities M* is the image of the parameter manifold P under the mapping $\theta \to p_\theta$ with certain regularity conditions. A vector u tangent to this manifold is of the form $u = \frac{d}{dt} p_{\theta+tu|t=0}$, and the Fisher information metric $g = g_p$ on M, obtained from the metric g_θ on P, can be written as

$$g_p(u, v) = \mathbb{E}_p \left[\frac{u}{p} \cdot \frac{v}{p} \right].$$

- **Fisher–Rao metric**

  Let $\mathcal{P}_n = \{p \in \mathbb{R}^n : \sum_{i=1}^n p_i = 1, p_i > 0\}$ be the simplex of strictly positive probability vectors. An element $p \in \mathcal{P}_n$ is a density of the n-point set $\{1, \dots, n\}$ with $p(i) = p_i$. An element u of the tangent space $T_p(\mathcal{P}_n) = \{u \in \mathbb{R}^n : \sum_{i=1}^n u_i = 0\}$ at a point $p \in \mathcal{P}_n$ is a function on $\{1, \dots, n\}$ with $u(i) = u_i$.

  The **Fisher–Rao metric** g_p on $\mathcal{P}_n$ is a Riemannian metric defined by

$$g_p(u, v) = \sum_{i=1}^n \frac{u_i v_i}{p_i}$$

  for any $u, v \in T_p(\mathcal{P}_n)$, i.e., it is the **Fisher information metric** on $\mathcal{P}_n$.

  The Fisher–Rao metric is the unique (up to a constant factor) Riemannian metric on $\mathcal{P}_n$, contracting under stochastic maps ([Chen72]).

  This metric is isometric, by $p \to 2(\sqrt{p_1}, \dots, \sqrt{p_n})$, with the standard metric on an open subset of the sphere of radius two in $\mathbb{R}^n$. This identification allows one to obtain on $\mathcal{P}_n$ the **geodesic distance**, called the **Rao distance**, by

$$2 \arccos(\sum_i p_i^{1/2} q_i^{1/2}).$$

  The Fisher–Rao metric can be extended to the set $\mathcal{M}_n = \{p \in \mathbb{R}^n, p_i > 0\}$ of all finite strictly positive measures on the set $\{1, \dots, n\}$. In this case, the geodesic distance on $\mathcal{M}_n$ can be written as

$$2(\sum_i (\sqrt{p_i} - \sqrt{q_i})^2)^{1/2}$$

for any $p, q \in \mathcal{M}_n$ (cf. **Hellinger metric** in Chap. 14).

- **Monotone metrics**

  Let M_n be the set of all complex $n \times n$ matrices. Let $\mathcal{M} \subset M_n$ be the manifold of all such positive-definite matrices. Let $\mathcal{D} \subset \mathcal{M}$, $\mathcal{D} = \{\rho \in \mathcal{M} : \mathrm{Tr}\rho = 1\}$, be the submanifold of all *density matrices*. It is the space of faithful states of an n-level quantum system; cf. **distances between quantum states** in Chap. 24.

  The tangent space of $\mathcal{M}$ at $\rho \in \mathcal{M}$ is $T_\rho(\mathcal{M}) = \{x \in M_n : x = x^*\}$, i.e., the set of all $n \times n$ *Hermitian matrices*. The tangent space $T_\rho(\mathcal{D})$ at $\rho \in \mathcal{D}$ is the subspace of *traceless* (i.e., with trace 0) matrices in $T_\rho(\mathcal{M})$.

  A Riemannian metric λ on $\mathcal{M}$ is called **monotone metric** if the inequality

$$\lambda_{h(\rho)}(h(u), h(u)) \leq \lambda_\rho(u, u)$$

  holds for any $\rho \in \mathcal{M}$, any $u \in T_\rho(\mathcal{M})$, and any *stochastic*, i.e., completely positive trace preserving mapping h.

  It was proved in [Petz96] that λ is monotone if and only if it can be written as

$$\lambda_\rho(u, v) = \mathrm{Tr}\ uJ_\rho(v),$$

  where J_ρ is an operator of the form $J_\rho = \frac{1}{f(L_\rho/R_\rho)R_\rho}$. Here L_ρ and R_ρ are the left and the right multiplication operators, and $f : (0, \infty) \to \mathbb{R}$ is an operator monotone function which is *symmetric*, i.e., $f(t) = tf(t^{-1})$, and *normalized*, i.e., $f(1) = 1$. Then $J_\rho(v) = \rho^{-1}v$ if v and ρ are commute, i.e., any monotone metric is equal to the **Fisher information metric** on commutative submanifolds.

  The **Bures metric** (or *Helstrom metric*) is the smallest monotone metric, $\frac{1}{2}\mathrm{Tr}(d\rho g)$, obtained for $f(t) = \frac{1+t}{2}$. In this case $J_\rho(v) = g$, $\rho g + g\rho = 2v$.

  For any $\rho_1, \rho_2 \in \mathcal{M}$ the **geodesic distance** defined by the Bures metric, (cf. **Bures distance** in Chap. 24) can be written as

$$\sqrt{\mathrm{Tr}(\rho_1) + \mathrm{Tr}(\rho_2) - 2\mathrm{Tr}(\sqrt{\sqrt{\rho_1}\rho_2\sqrt{\rho_1}})}.$$

  On the submanifold $\mathcal{D} = \{\rho \in \mathcal{M} : \mathrm{Tr}\rho = 1\}$ of density matrices it has the form

$$\arccos \mathrm{Tr}(\sqrt{\sqrt{\rho_1}\rho_2\sqrt{\rho_1}}).$$

  The **right logarithmic derivative metric** (or *RLD-metric*) is the greatest monotone metric, corresponding to $f(t) = \frac{2t}{1+t}$, i.e., $J_\rho(v) = \frac{1}{2}(\rho^{-1}v + v\rho^{-1})$.

[redacted]

The **Bogolubov–Kubo–Mori metric** (or *BKM-metric*) is obtained for $f(x) = \frac{x-1}{\ln x}$. It can be written as $\lambda_\rho(u,v) = \frac{\partial^2}{\partial s \partial t}\mathrm{Tr}(\rho + su)\ln(\rho + tv)|_{s,t=0}$.

- **Wigner–Yanase–Dyson metrics**

  The **Wigner–Yanase–Dyson metrics** (or *WYD-metrics*) form a family of metrics on the manifold $\mathcal{M}$ of all complex positive-definite $n \times n$ matrices defined by

$$\lambda_\rho^\alpha(u,v) = \frac{\partial^2}{\partial t \partial s}\mathrm{Tr} f_\alpha(\rho + tu) f_{-\alpha}(\rho + sv)|_{s,t=0},$$

  where $f_\alpha(x) = \frac{2}{1-\alpha}x^{\frac{1-\alpha}{2}}$, if $\alpha \neq 1$, and is $\ln x$, if $\alpha = 1$. These metrics are monotone for $\alpha \in [-3,3]$. For $\alpha = \pm 1$ one obtains the **Bogolubov–Kubo–Mori metric**; for $\alpha = \pm 3$ one obtains the **right logarithmic derivative metric**.

  The **Wigner–Yanase metric** (or *WY-metric*) is λ_ρ^0, the smallest metric in the family. It can be written as $\lambda_\rho(u,v) = 4\mathrm{Tr}\ u(\sqrt{L_\rho} + \sqrt{R_\rho})^2(v)$.

- **Connes metric**

  Roughly, the **Connes metric** is a generalization (from the space of all probability measures of a set X, to the *state space* of any *unital C^*-algebra*) of the **transportation distance** (Chap. 14) defined via *Lipschitz seminorm*.

  Let M^n be a smooth n-dimensional manifold. Let $A = C^\infty(M^n)$ be the (commutative) algebra of smooth complex-valued functions on M^n, represented as multiplication operators on the Hilbert space $H = L^2(M^n, S)$ of square integrable sections of the spinor bundle on M^n by $(f\xi)(p) = f(p)\xi(p)$ for all $f \in A$ and for all $\xi \in H$.

  Let D be the *Dirac operator*. Let the commutator $[D,f]$ for $f \in A$ be the *Clifford multiplication* by the gradient ∇f, so that its operator norm $||.||$ in H is given by $||[D,f]|| = \sup_{p \in M^n} ||\nabla f||$.

  The **Connes metric** is the **intrinsic metric** on M^n, defined by

$$\sup_{f \in A, ||\,[D,f]\,|| \leq 1} |f(p) - f(q)|.$$

  This definition can also be applied to discrete spaces, and even generalized to *C^*-algebras*; cf. **Rieffel metric space**. In particular, for a labeled connected *locally finite* graph $G = (V,E)$ with the vertex-set $V = \{v_1, \dots, v_n, \dots\}$, the Connes metric on V is defined, for any $v_i, v_j \in V$, by $\sup_{||\,[D,f]\,|| = ||df|| \leq 1} |f_{v_i} - f_{v_j}|$, where $\{f = \sum f_{v_i} v_i : \sum |f_{v_i}|^2 < \infty\}$ is the set of formal sums f, forming a Hilbert space, and $||\,[D,f]\,||$ is $\sup_i (\sum_{k=1}^{deg(v_i)} (f_{v_k} - f_{v_i})^2)^{\frac{1}{2}}$.

- **Rieffel metric space**

  Let V be a *normed space* (or, more generally, a **locally convex topological vector space**, cf. Chap. 2), and let V' be its **continuous dual space**, i.e., the set of all continuous linear functionals f on V. The *weak-* topology* on V' is defined as the weakest (i.e., with the fewest open sets) topology on V' such that, for every

$x \in V$, the map $F_x : V' \to \mathbb{R}$ defined by $F_x(f) = f(x)$ for all $f \in V'$, remains continuous.

An *order-unit space* is a *partially ordered* real (complex) vector space $(A, \preceq)$ with a special distinguished element e (*order unit*) satisfying the following properties:

1. For any $a \in A$, there exists $r \in \mathbb{R}$ with $a \preceq re$;
2. If $a \in A$ and $a \preceq re$ for all positive $r \in \mathbb{R}$, then $a \preceq 0$ (*Archimedean property*).

The main example of an order-unit space is the vector space of all self-adjoint elements in a *unital C^*-algebra* with the identity element being the order unit. Here a *C^*-algebra* is a *Banach algebra* over $\mathbb{C}$ equipped with a special *involution*. It is called *unital* if it has a *unit* (multiplicative identity element); such C^*-algebras are also called, roughly, *compact noncommutative topological spaces*.

Main example of a unital C^*-algebra is the complex algebra of linear operators on a complex **Hilbert space** which is topologically closed in the norm topology of operators, and is closed under the operation of taking adjoints of operators.

The *state space* of an order-unit space $(A, \preceq, e)$ is the set $S(A) = \{f \in A'_+ : ||f|| = 1\}$ of *states*, i.e., continuous linear functionals f with $||f|| = f(e) = 1$.

A **Rieffel** (or *compact quantum* as in Rieffel, 1999) **metric space** is a pair $(A, ||.||_{Lip})$, where $(A, \preceq, e)$ is an order-unit space, and $||.||_{Lip}$ is a $[0, +\infty]$-valued seminorm on A (generalizing the *Lipschitz seminorm*) for which it holds:

1. For $a \in A$, $||a||_{Lip} = 0$ holds if and only if $a \in \mathbb{R}e$;
2. the metric $d_{Lip}(f, g) = \sup_{a \in A:||a||_{Lip} \leq 1} |f(a) - g(a)|$ generates on the state space $S(A)$ its weak-* topology.

So, $(S(A), d_{Lip})$ is a usual metric space. If the order-unit space $(A, \preceq, e)$ is a C^*-algebra, then d_{Lip} is the **Connes metric**, and if, moreover, the C^*-algebra is noncommutative, $(S(A), d_{Lip})$ is called a **noncommutative metric space**.

The term *quantum* is due to the belief that the Planck-scale geometry of *space-time* comes from such C^*-algebras; cf. **quantum space-time** in Chap. 24.

Kuperberg and Weaver, 2010, proposed a new definition of *quantum metric space*, in the setting of *von Neumann algebras*.

7.3 Hermitian Metrics and Generalizations

A *vector bundle* is a geometrical construct where to every point of a *topological space M* we attach a vector space so that all those vector spaces "glued together" form another topological space E. A continuous mapping $\pi : E \to M$ is called a *projection E* on M. For every $p \in M$, the vector space $\pi^{-1}(p)$ is called a *fiber* of the vector bundle.

A *real* (*complex*) *vector bundle* is a vector bundle $\pi : E \to M$ whose fibers $\pi^{-1}(p), p \in M$, are real (complex) vector spaces.

[redacted]

In a real vector bundle, for every $p \in M$, the fiber $\pi^{-1}(p)$ locally looks like the vector space $\mathbb{R}^n$, i.e., there is an *open neighborhood* U of p, a natural number n, and a homeomorphism $\varphi : U \times \mathbb{R}^n \to \pi^{-1}(U)$ such that, for all $x \in U$ and $v \in \mathbb{R}^n$, one has $\pi(\varphi(x, v)) = v$, and the mapping $v \to \varphi(x, v)$ yields an isomorphism between $\mathbb{R}^n$ and $\pi^{-1}(x)$. The set U, together with φ, is called a *local trivialization* of the bundle.

If there exists a "global trivialization", then a real vector bundle $\pi : M \times \mathbb{R}^n \to M$ is called *trivial*. Similarly, in a complex vector bundle, for every $p \in M$, the fiber $\pi^{-1}(p)$ locally looks like the vector space $\mathbb{C}^n$. The basic example of such bundle is the trivial bundle $\pi : U \times \mathbb{C}^n \to U$, where U is an open subset of $\mathbb{R}^k$.

Important special cases of a real vector bundle are the *tangent bundle* $T(M^n)$ and the *cotangent bundle* $T^*(M^n)$ of a *real n-dimensional manifold* $M^n_{\mathbb{R}} = M^n$. Important special cases of a complex vector bundle are the tangent bundle and the cotangent bundle of a *complex n-dimensional manifold.*

Namely, a *complex n-dimensional manifold* $M^n_{\mathbb{C}}$ is a *topological space* in which every point has an open neighborhood homeomorphic to an open set of the n-dimensional complex vector space $\mathbb{C}^n$, and there is an atlas of charts such that the change of coordinates between charts is analytic. The (complex) tangent bundle $T_{\mathbb{C}}(M^n_{\mathbb{C}})$ of a complex manifold $M^n_{\mathbb{C}}$ is a vector bundle of all (complex) *tangent spaces* of $M^n_{\mathbb{C}}$ at every point $p \in M^n_{\mathbb{C}}$. It can be obtained as a *complexification* $T_{\mathbb{R}}(M^n_{\mathbb{R}}) \otimes \mathbb{C} = T(M^n) \otimes \mathbb{C}$ of the corresponding real tangent bundle, and is called the *complexified tangent bundle* of $M^n_{\mathbb{C}}$.

The *complexified cotangent bundle* of $M^n_{\mathbb{C}}$ is obtained similarly as $T^*(M^n) \otimes \mathbb{C}$. Any complex n-dimensional manifold $M^n_{\mathbb{C}} = M^n$ can be regarded as a real $2n$-dimensional manifold equipped with a *complex structure* on each tangent space.

A *complex structure* on a real vector space V is the structure of a complex vector space on V that is compatible with the original real structure. It is completely determined by the operator of multiplication by the number i, the role of which can be taken by an arbitrary linear transformation $J : V \to V$, $J^2 = -id$, where id is the *identity mapping*.

A *connection* (or *covariant derivative*) is a way of specifying a derivative of a *vector field* along another vector field in a vector bundle. A **metric connection** is a linear connection in a vector bundle $\pi : E \to M$, equipped with a bilinear form in the fibers, for which parallel displacement along an arbitrary piecewise-smooth curve in M preserves the form, that is, the *scalar product* of two vectors remains constant under parallel displacement.

In the case of a nondegenerate symmetric bilinear form, the metric connection is called the *Euclidean connection*. In the case of nondegenerate antisymmetric bilinear form, the metric connection is called the *symplectic connection*.

- **Bundle metric**

  A **bundle metric** is a metric on a vector bundle.
- **Hermitian metric**

  A **Hermitian metric** on a complex vector bundle $\pi : E \to M$ is a collection of *Hermitian inner products* (i.e., positive-definite symmetric sesquilinear forms) on every fiber $E_p = \pi^{-1}(p), p \in M$, that varies smoothly with the point p in M. Any complex vector bundle has a Hermitian metric.

The basic example of a vector bundle is the trivial bundle $\pi : U \times \mathbb{C}^n \to U$, where U is an open set in $\mathbb{R}^k$. In this case a Hermitian inner product on $\mathbb{C}^n$, and hence, a Hermitian metric on the bundle $\pi : U \times \mathbb{C}^n \to U$, is defined by

$$\langle u, v \rangle = u^T H \overline{v},$$

where H is a *positive-definite Hermitian matrix*, i.e., a complex $n \times n$ matrix such that $H^* = \overline{H}^T = H$, and $\overline{v}^T H v > 0$ for all $v \in \mathbb{C}^n \backslash \{0\}$. In the simplest case, one has $\langle u, v \rangle = \sum_{i=1}^{n} u_i \overline{v}_i$.

An important special case is a Hermitian metric h on a complex manifold M^n, i.e., on the complexified tangent bundle $T(M^n) \otimes \mathbb{C}$ of M^n. This is the Hermitian analog of a Riemannian metric. In this case $h = g + iw$, and its real part g is a Riemannian metric, while its imaginary part w is a nondegenerate antisymmetric bilinear form, called a *fundamental form*. Here $g(J(x), J(y)) = g(x, y)$, $w(J(x), J(y)) = w(x, y)$, and $w(x, y) = g(x, J(y))$, where the operator J is an operator of complex structure on M^n; as a rule, $J(x) = ix$. Any of the forms g, w determines h uniquely.

The term *Hermitian metric* can also refer to the corresponding Riemannian metric g, which gives M^n a Hermitian structure.

On a complex manifold, a Hermitian metric h can be expressed in local coordinates by a *Hermitian symmetric tensor* $((h_{ij}))$:

$$h = \sum_{i,j} h_{ij} dz_i \otimes d\overline{z}_j,$$

where $((h_{ij}))$ is a positive-definite Hermitian matrix. The associated fundamental form w is then written as $w = \frac{i}{2} \sum_{i,j} h_{ij} dz_i \wedge d\overline{z}_j$. A *Hermitian manifold* (or *Hermitian space*) is a complex manifold equipped with a Hermitian metric.

- **Kähler metric**

A **Kähler metric** (or *Kählerian metric*) is a Hermitian metric $h = g + iw$ on a complex manifold M^n whose fundamental form w is *closed*, i.e., $dw = 0$ holds. A *Kähler manifold* is a complex manifold equipped with a Kähler metric.

If h is expressed in local coordinates, i.e., $h = \sum_{i,j} h_{ij} dz_i \otimes d\overline{z}_j$, then the associated fundamental form w can be written as $w = \frac{i}{2} \sum_{i,j} h_{ij} dz_i \wedge d\overline{z}_j$, where $\wedge$ is the *wedge product* which is antisymmetric, i.e., $dx \wedge dy = -dy \wedge dx$ (hence, $dx \wedge dx = 0$).

In fact, w is a *differential* 2-*form* on M^n, i.e., a tensor of rank 2 that is antisymmetric under exchange of any pair of indices: $w = \sum_{i,j} f_{ij} dx^i \wedge dx^j$, where f_{ij} is a function on M^n. The *exterior derivative* dw of w is defined by $dw = \sum_{i,j} \sum_k \frac{\partial f_{ij}}{\partial x_k} dx_k \wedge dx_i \wedge dx_k$. If $dw = 0$, then w is a *symplectic* (i.e., closed nondegenerate) differential 2-form. Such differential 2-forms are called *Kähler forms*.

[redacted]

The metric on a Kähler manifold locally satisfies $h_{ij} = \frac{\partial^2 K}{\partial z_i \partial \overline{z}_j}$. for some function K, called the *Kähler potential*. The term *Kähler metric* can also refer to the corresponding Riemannian metric g, which gives M^n a Kähler structure. Then a Kähler manifold is defined as a complex manifold which carries a Riemannian metric and a Kähler form on the underlying real manifold.

- **Hessian metric**

  Given a smooth f on an open subset of a real vector space, the associated **Hessian metric** is defined by

$$g_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

  A Hessian metric is also called an **affine Kähler metric** since a Kähler metric on a complex manifold has an analogous description as $\frac{\partial^2 f}{\partial z_i \partial \overline{z}_j}$.

- **Calabi–Yau metric**

  The **Calabi–Yau metric** is a **Kähler metric** which is **Ricci-flat**.

  A *Calabi–Yau manifold* (or *Calabi–Yau space*) is a simply connected complex manifold equipped with a Calabi–Yau metric. It can be considered as a $2n$-dimensional (6D being particularly interesting) smooth manifold with holonomy group (i.e., the set of linear transformations of tangent vectors arising from parallel transport along closed loops) in the special unitary group.

- **Kähler–Einstein metric**

  A **Kähler–Einstein metric** is a **Kähler metric** on a complex manifold M^n whose *Ricci curvature tensor* is proportional to the metric tensor. This proportionality is an analog of the *Einstein field equation* in the General Theory of Relativity.

  A *Kähler–Einstein manifold* (or *Einstein manifold*) is a complex manifold equipped with a Kähler–Einstein metric. In this case the Ricci curvature tensor, seen as an operator on the tangent space, is just multiplication by a constant.

  Such a metric exists on any *domain* $D \subset \mathbb{C}^n$ that is bounded and *pseudo-convex*. It can be given by the *line element*

$$ds^2 = \sum_{i,j} \frac{\partial^2 u(z)}{\partial z_i \partial \overline{z}_j} dz_i d\overline{z}_j,$$

  where u is a solution to the *boundary value problem*: $det(\frac{\partial^2 u}{\partial z_i \partial \overline{z}_j}) = e^{2u}$ on D, and $u = \infty$ on ∂D. The Kähler–Einstein metric is a complete metric. On the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ it is coincides with the **Poincaré metric**.

  Let h be the Einstein metric on a smooth compact manifold M^{n-1} without boundary, having scalar curvature $(n-1)(n-2)$. A **generalized Delaunay metric** on $\mathbb{R} \times M^{n-1}$ is (Delay, 2010) of the form $g = u^{\frac{4}{n-2}}(dy^2 + h)$, where $u = u(y) > 0$ is a periodic solution of $u'' - \frac{(n-2)^2}{4} u + \frac{n(n-2)}{4} u^{\frac{n+2}{n-2}} = 0$.

There is one parameter family of constant positive curvature conformal metrics on $\mathbb{R} \times \mathbb{S}^{n-1}$, referred to as **Delaunay metric**; cf. **Kottler metric** in Chap. 26.

- **Hodge metric**

The **Hodge metric** is a **Kähler metric** whose *fundamental form w* defines an integral cohomology class or, equivalently, has integral periods.

A *Hodge manifold* (or *Hodge variety*) is a complex manifold equipped with a Hodge metric. A compact complex manifold is a Hodge manifold if and only if it is isomorphic to a smooth algebraic subvariety of some complex projective space.

- **Fubini–Study metric**

The **Fubini–Study metric** (or *Cayley–Fubini–Study metric*) is a **Kähler metric** on a *complex projective space* $\mathbb{C}P^n$ defined by a *Hermitian inner product* $\langle , \rangle$ in $\mathbb{C}^{n+1}$. It is given by the *line element*

$$ds^2 = \frac{\langle x, x\rangle\langle dx, dx\rangle - \langle x, d\overline{x}\rangle\langle \overline{x}, dx\rangle}{\langle x, x\rangle^2}.$$

The **Fubini–Study distance** between points $(x_1 : \ldots : x_{n+1})$ and $(y_1 : \ldots : y_{n+1}) \in \mathbb{C}P^n$, where $x = (x_1, \ldots, x_{n+1})$ and $y = (y_1, \ldots, y_{n+1}) \in \mathbb{C}^{n+1} \backslash \{0\}$, is equal to

$$\arccos \frac{|\langle x, y\rangle|}{\sqrt{\langle x, x\rangle\langle y, y\rangle}}.$$

The Fubini–Study metric is a **Hodge metric**. The space $\mathbb{C}P^n$ endowed with this metric is called a *Hermitian elliptic space* (cf. **Hermitian elliptic metric**).

- **Bergman metric**

The **Bergman metric** is a **Kähler metric** on a bounded *domain* $D \subset \mathbb{C}^n$ defined, for the *Bergman kernel* $K(z, u)$, by the *line element*

$$ds^2 = \sum_{i,j} \frac{\partial^2 \ln K(z, z)}{\partial z_i \partial \overline{z}_j} dz_i d\overline{z}_j.$$

It is a **biholomorhically invariant metric** on D, and it is complete if D is homogeneous. For the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ the Bergman metric coincides with the **Poincaré metric**; cf. also **Bergman p-metric** in Chap. 13.

The set of all analytic functions $f \neq 0$ of class $L_2(D)$ with respect to the Lebesgue measure, forms the **Hilbert space** $L_{2,a}(D) \subset L_2(D)$ with an orthonormal basis $(\phi_i)_i$. The *Bergman kernel* is a function in the domain $D \times D \subset \mathbb{C}^{2n}$, defined by $K_D(z, u) = K(z, u) = \sum_{i=1}^{\infty} \phi_i(z)\overline{\phi_i(u)}$.

[redacted]

The **Skwarczynski distance** is defined by

$$(1 \frac{|K(z,u)|}{\sqrt{K(z,z)}\sqrt{K(u,u)}})^{\frac{1}{2}}$$

.

- **Hyper-Kähler metric**

  A **hyper-Kähler metric** is a Riemannian metric g on a $4n$-dimensional *Riemannian manifold* which is compatible with a quaternionic structure on the tangent bundle of the manifold.

  Thus, the metric g is Kählerian with respect to 3 Kähler structures (I, w_I, g), (J, w_J, g), and (K, w_K, g), corresponding to the complex structures, as endomorphisms of the tangent bundle, which satisfy the quaternionic relationship

$$I^2 = J^2 = K^2 = IJK = -JIK = -1.$$

  A *hyper-Kähler manifold* is a Riemannian manifold equipped with a hyper-Kähler metric. manifolds are Ricci-flat. Compact 4D hyper-Kähler manifolds are called K_3-*surfaces*; they are studied in Algebraic Geometry.

- **Calabi metric**

  The **Calabi metric** is a **hyper-Kähler metric** on the cotangent bundle $T^*(\mathbb{C}P^{n+1})$ of a *complex projective space* $\mathbb{C}P^{n+1}$.

  For $n = 4k + 4$, this metric can be given by the *line element*

$$ds^2 = \frac{dr^2}{1-r^{-4}} + \frac{1}{4}r^2(1-r^{-4})\lambda^2 + r^2(\nu_1^2+\nu_2^2) + \frac{1}{2}(r^2-1)(\sigma_{1\alpha}^2+\sigma_{2\alpha}^2) + \frac{1}{2}(r^2+1)(\Sigma_{1\alpha}^2+\Sigma_{2\alpha}^2),$$

  where $(\lambda, \nu_1, \nu_2, \sigma_{1\alpha}, \sigma_{2\alpha}, \Sigma_{1\alpha}, \Sigma_{2\alpha})$, with α running over k values, are left-invariant *one-forms* (i.e., linear real-valued functions) on the coset $SU(k+2)/U(k)$. Here $U(k)$ is the *unitary group* consisting of complex $k \times k$ *unitary matrices*, and $SU(k)$ is the *special unitary group* of complex $k \times k$ unitary matrices with determinant 1.

  For $k = 0$, the Calabi metric coincides with the **Eguchi–Hanson metric**.

- **Stenzel metric**

  The **Stenzel metric** is a **hyper-Kähler metric** on the cotangent bundle $T^*(S^{n+1})$ of a sphere S^{n+1}.

- $SO(3)$**-invariant metric**

  An $SO(3)$**-invariant metric** is a 4D 4-dimensional hyper-Kähler metric with the *line element* given, in the Bianchi type *IX* formalism (cf. **Bianchi metrics** in Chap. 26) by

$$ds^2 = f^2(t)dt^2 + a^2(t)\sigma_1^2 + b^2(t)\sigma_2^2 + c^2(t)\sigma_3^2,$$

  where the invariant *one-forms* $\sigma_1, \sigma_2, \sigma_3$ of $SO(3)$ are expressed in terms of *Euler angles* θ, ψ, ϕ as $\sigma_1 = \frac{1}{2}(\sin\psi d\theta - \sin\theta\cos\psi d\phi)$, $\sigma_2 = -\frac{1}{2}(\cos\psi d\theta +$

$\sin\theta \sin\psi d\phi)$, $\sigma_3 = \frac{1}{2}(d\psi + \cos\theta d\phi)$, and the normalization has been chosen so that $\sigma_i \wedge \sigma_j = \frac{1}{2}\epsilon_{ijk} d\sigma_k$. The coordinate t of the metric can always be chosen so that $f(t) = \frac{1}{2}abc$, using a suitable reparametrization.

- **Atiyah–Hitchin metric**

  The **Atiyah–Hitchin metric** is a **complete regular** $SO(3)$**-invariant metric** with the *line element*

$$ds^2 = \frac{1}{4}a^2b^2c^2\left(\frac{dk}{k(1-k^2)K^2}\right)^2 + a^2(k)\sigma_1^2 + b^2(k)\sigma_2^2 + c^2(k)\sigma_3^2,$$

  where a, b, c are functions of k, $ab = -K(k)(E(k) - K(k))$, $bc = -K(k)(E(k) - (1-k^2)K(k))$, $ac = -K(k)E(k)$, and $K(k)$, $E(k)$ are the complete elliptic integrals, respectively, of the first and second kind, with $0 < k < 1$. The coordinate t is given by the change of variables $t = -\frac{2K(1-k^2)}{\pi K(k)}$ up to an additive constant.

- **Taub–NUT metric**

  The **Taub–NUT metric** (cf. also Chap. 26) is a **complete regular** $SO(3)$**-invariant metric** with the *line element*

$$ds^2 = \frac{1}{4}\frac{r+m}{r-m}dr^2 + (r^2 - m^2)(\sigma_1^2 + \sigma_2^2) + 4m^2\frac{r-m}{r+m}\sigma_3^2,$$

  where m is the relevant moduli parameter, and the coordinate r is related to t by $r = m + \frac{1}{2mt}$. NUT manifold was discovered in Ehlers, 1957, and rediscovered in Newman–Tamburino–Unti, 1963; it is closely related to the metric in Taub, 1951.

- **Eguchi–Hanson metric**

  The **Eguchi–Hanson metric** is a **complete regular** $SO(3)$**-invariant metric** with the *line element*

$$ds^2 = \frac{dr^2}{1 - \left(\frac{a}{r}\right)^4} + r^2\left(\sigma_1^2 + \sigma_2^2 + \left(1 - \left(\frac{a}{r}\right)^4\right)\sigma_3^2\right),$$

  where a is the moduli parameter, and the coordinate r is $a\sqrt{\coth(a^2t)}$.

  The Eguchi–Hanson metric coincides with the 4D **Calabi metric**.

- **Complex Finsler metric**

  A **complex Finsler metric** is an upper semicontinuous function $F : T(M^n) \to \mathbb{R}_+$ on a complex manifold M^n with the analytic tangent bundle $T(M^n)$ satisfying the following conditions:

  1. F^2 is smooth on $\check{M}^n$, where $\check{M}^n$ is the complement in $T(M^n)$ of the zero section;
  2. $F(p, x) > 0$ for all $p \in M^n$ and $x \in \check{M}^n_p$;
  3. $F(p, \lambda x) = |\lambda| F(p, x)$ for all $p \in M^n$, $x \in T_p(M^n)$, and $\lambda \in \mathbb{C}$.

[redacted]

The function $G = F^2$ can be locally expressed in terms of the coordinates $(p_1, \dots, p_n,\ x_1, \dots, x_n)$; the *Finsler metric tensor* of the complex Finsler metric is given by the matrix$((G_{ij})) = ((\frac{1}{2}\frac{\partial^2 F^2}{\partial x_i \partial \overline{x}_j}))$, called the *Levi matrix*. If the matrix $((G_{ij}))$ is positive-definite, the complex Finsler metric F is called *strongly pseudo-convex*.

- **Distance-decreasing semimetric**

  Let d be a semimetric which can be defined on some class $\mathcal{M}$ of complex manifolds containing the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$. It is called **distance-decreasing** if, for any analytic mapping$f : M_1 \to M_2$ with $M_1, M_2 \in \mathcal{M}$, the inequality $d(f(p), f(q)) \leq d(p, q)$ holds for all $p, q \in M_1$.

  The **Carathéodory semimetric** F_C, **Sibony semimetric** F_S, **Azukawa semimetric** F_A and **Kobayashi semimetric** F_K are distance-decreasing with F_C and F_K being the smallest and the greatest distance-decreasing semimetrics. They are generalizations of the **Poincaré metric** to higher-dimensional domains, since $F_C = F_K$ is the **Poincaré metric** on the unit disk Δ, and $F_C = F_K \equiv 0$ on $\mathbb{C}^n$.

  It holds $F_C(z, u) \leq F_S(z, u) \leq F_A(z, u) \leq F_K(z, u)$ for all $z \in D$ and $u \in \mathbb{C}^n$. If D is convex, then all these metrics coincide.

- **Biholomorphically invariant semimetric**

  A *biholomorphism* is a bijective *holomorphic* (complex differentiable in a neighborhood of every point in its domain) function whose inverse is also holomorphic.

  A semimetric $F(z, u) : D \times \mathbb{C}^n \to [0, \infty]$ on a domain D in $\mathbb{C}^n$ is called **biholomorphically invariant** if $F(z, u) = |\lambda| F(z, u)$ for all $\lambda \in \mathbb{C}$, and $F(z, u) = F(f(z), f'(z)u)$ for any biholomorphism$f : D \to D'$.

  Invariant metrics, including the **Carathéodory**, **Kobayashi**, **Sibony**, **Azukawa**, **Bergman**, and **Kähler–Einstein** metrics, play an important role in Complex Function Theory, Complex Dynamics and Convex Geometry. The first four metrics are used mostly because they are **distance-decreasing**. But they are almost never Hermitian. On the other hand, the Bergman metric and the Kähler–Einstein metric are Hermitian (in fact, Kählerian), but, in general, not distance-decreasing.

  The **Wu metric** (Cheung and Kim, 1996) is an invariant non-Kähler Hermitian metric on a complex manifold M^n which factor, for any holomorphic mapping between two complex manifolds.

- **Kobayashi metric**

  Let D be a *domain* in $\mathbb{C}^n$. Let $\mathcal{O}(\Delta, D)$ be the set of all analytic mappings $f : \Delta \to D$, where $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ is the *unit disk*.

  The **Kobayashi metric** (or **Kobayashi–Royden metric**) F_K is a **complex Finsler metric** defined, for all $z \in D$ and $u \in \mathbb{C}^n$, by

$$F_K(z, u) = \inf\{\alpha > 0 : \exists f \in \mathcal{O}(\Delta, D), f(0) = z, \alpha f'(0) = u\}.$$

Given a complex manifold M^n, the **Kobayashi semimetric** F_K is defined by

$$F_K(p,u) = \inf\{\alpha > 0 : \exists f \in \mathcal{O}(\Delta, M^n), f(0) = p, \alpha f'(0) = u\}$$

for all $p \in M^n$ and $u \in T_p(M^n)$.

$F_K(p,u)$ is a seminorm of the tangent vector u, called the *Kobayashi seminorm*. F_K is a metric if M^n is *taut*, i.e., $\mathcal{O}(\Delta, M^n)$ is a *normal family* (every sequence has a subsequence which either converge or diverge compactly).

The Kobayashi semimetric is an infinitesimal form of the **Kobayashi semidistance** (or *Kobayashi pseudo-distance*, 1967) K_{M^n} on M^n, defined as follows. Given $p, q \in M^n$, a *chain of disks* α from p to q is a collection of points $p = p^0, p^1, \dots, p^k = q$ of M^n, pairs of points $a^1, b^1; \dots; a^k, b^k$ of the unit disk Δ, and analytic mappings $f_1, \dots f_k$ from Δ into M^n, such that $f_j(a^j) = p^{j-1}$ and $f_j(b^j) = p^j$ for all j.

The length $l(\alpha)$ of a chain α is the sum $d_P(a^1, b^1) + \cdots + d_P(a^k, b^k)$, where d_P is the Poincaré metric. The Kobayashi semimetric K_{M^n} on M^n is defined by

$$K_{M^n}(p,q) = \inf_{\alpha} l(\alpha),$$

where the infimum is taken over all lengths $l(\alpha)$ of chains of disks α from p to q.

Given a complex manifold M^n, the **Kobayashi–Busemann semimetric** on M^n is the double dual of the **Kobayashi semimetric**. It is a metric if M^n is taut.

- **Carathéodory metric**

  Let D be a *domain* in $\mathbb{C}^n$. Let $\mathcal{O}(D, \Delta)$ be the set of all analytic mappings $f : D \to \Delta$, where $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ is the *unit disk*.

  The **Carathéodory metric** F_C is a **complex Finsler metric** defined by

$$F_C(z,u) = \sup\{|f'(z)u| : f \in \mathcal{O}(D, \Delta)\}$$

for any $z \in D$ and $u \in \mathbb{C}^n$.

Given a complex manifold M^n, the **Carathéodory semimetric** F_C is defined by

$$F_C(p,u) = \sup\{|f'(p)u| : f \in \mathcal{O}(M^n, \Delta)\}$$

for all $p \in M^n$ and $u \in T_p(M^n)$. F_C is a metric if M^n is *taut*, i.e., every sequence in $\mathcal{O}(\Delta, M^n)$ has a subsequence which either converge or diverge compactly.

The **Carathéodory semidistance** (or *Carathéodory pseudo-distance*, 1926) C_{M^n} is a semimetric on a complex manifold M^n, defined by

$$C_{M^n}(p,q) = \sup\{d_P(f(p), f(q)) : f \in \mathcal{O}(M^n, \Delta)\},$$

where d_P is the Poincaré metric.

[redacted]

In general, the integrated semimetric of the infinitesimal Carathéodory semimetric is **internal** for the Carathéodory semidistance, but does not equal to it.

- **Azukawa semimetric**

  Let D be a *domain* in $\mathbb{C}^n$. Let $K_D(z)$ be the set of all *logarithmically plurisubharmonic* functions $f : D \to [0,1)$ such that there exist $M, r > 0$ with $f(u) \leq M||u-z||_2$ for all $u \in B(z,r) \subset D$; here $||.||_2$ is the l_2*-norm* on $\mathbb{C}^n$, and $B(z,r) = \{x \in \mathbb{C}^n : ||z-x||_2 < r\}$. Let $g_D(z,u)$ be $\sup\{f(u) : f \in K_D(z)\}$.

  The **Azukawa semimetric** F_A is a **complex Finsler metric** defined by

  $$F_A(z,u) = \overline{\lim}_{\lambda \to 0} \frac{1}{|\lambda|} g_D(z, z+\lambda u)$$

  for all $z \in D$ and $u \in \mathbb{C}^n$.

  The Azukawa metric is an infinitesimal form of the **Azukawa semidistance**.

- **Sibony semimetric**

  Let D be a *domain* in $\mathbb{C}^n$. Let $K_D(z)$ be the set of all *logarithmically plurisubharmonic* functions $f : D \to [0,1)$ such that there exist $M, r > 0$ with $f(u) \leq M||u-z||_2$ for all $u \in B(z,r) = \{x \in \mathbb{C}^n : ||z-x||_2 < r\} \subset D$. Let $C^2_{loc}(z)$ be the set of all functions of class C^2 on some open neighborhood of z.

  The **Sibony semimetric** F_S is a **complex Finsler semimetric** defined by

  $$F_S(z,u) = \sup_{f \in K_D(z) \cap C^2_{loc}(z)} \sqrt{\sum_{i,j} \frac{\partial^2 f}{\partial z_i \partial \overline{z}_j}(z) u_i \overline{u}_j}$$

  for all $z \in D$ and $u \in \mathbb{C}^n$.

  The Sibony semimetric is an infinitesimal form of the **Sibony semidistance**.

- **Teichmüller metric**

  A *Riemann surface* R is a one-dimensional complex manifold. Two Riemann surfaces R_1 and R_2 are called *conformally equivalent* if there exists a bijective analytic function (i.e., a conformal homeomorphism) from R_1 into R_2. More precisely, consider a fixed closed Riemann surface R_0 of a given *genus* $g \geq 2$.

  For a closed Riemann surface R of genus g, one can construct a pair (R,f), where $f : R_0 \to R$ is a homeomorphism. Two pairs (R,f) and (R_1,f_1) are called *conformally equivalent* if there exists a conformal homeomorphism $h : R \to R_1$ such that the mapping $(f_1)^{-1} \cdot h \cdot f : R_0 \to R_0$ is homotopic to the identity.

  An *abstract Riemann surface* $R^* = (R,f)^*$ is the equivalence class of all Riemann surfaces, conformally equivalent to R. The set of all equivalence classes is called the *Teichmüller space* $T(R_0)$ of the surface R_0.

  For closed surfaces R_0 of given genus g, the spaces $T(R_0)$ are isometrically isomorphic, and one can speak of the *Teichmüller space* T_g *of surfaces of genus* g. T_g is a complex manifold. If R_0 is obtained from a compact surface of genus $g \geq 2$ by removing n points, then the complex dimension of T_g is $3g-3+n$.

The **Teichmüller metric** is a metric on T_g defined by

$$\frac{1}{2}\inf_h \ln K(h)$$

for any $R_1^*, R_2^* \in T_g$, where $h : R_1 \to R_2$ is a quasi-conformal homeomorphism, homotopic to the identity, and $K(h)$ is the *maximal dilation* of h. In fact, there exists a unique extremal mapping, called the *Teichmüller mapping* which minimizes the maximal dilation of all such h, and the distance between R_1^* and R_2^* is equal to $\frac{1}{2}\ln K$, where the constant K is the dilation of the Teichmüller mapping.

In terms of the *extremal length* $ext_{R^*}(\gamma)$, the distance between R_1^* and R_2^* is

$$\frac{1}{2}\ln \sup_\gamma \frac{ext_{R_1^*}(\gamma)}{ext_{R_2^*}(\gamma)},$$

where the supremum is taken over all simple closed curves on R_0.

The Teichmüller space T_g, with the Teichmüller metric on it, is a **geodesic** metric space (moreover, a **straight** G**-space**) but it is neither **Gromov hyperbolic**, nor a **Busemann convex** metric space.

The **Thurston quasi-metric** on the *Teichmüller space* T_g is defined by

$$\frac{1}{2}\inf_h \ln ||h||_{Lip}$$

for any $R_1^*, R_2^* \in T_g$, where $h : R_1 \to R_2$ is a quasi-conformal homeomorphism, homotopic to the identity, and $||.||_{Lip}$ is the *Lipschitz norm* on the set of all injective functions $f : X \to Y$ defined by $||f||_{Lip} = \sup_{x,y \in X, x \neq y} \frac{d_Y(f(x),f(y))}{d_X(x,y)}$.

The *moduli space* R_g of conformal classes of Riemann surfaces of genus g is obtained by factorization of T_g by some countable group of automorphisms of it, called the *modular group*. The **Zamolodchikov metric**, defined (1986) in terms of *exactly marginal operators*, is a natural metric on the conformal moduli spaces.

Liu, Sun and Yau, 2005, showed that all known complete metrics on the Teichmüller space and moduli space (including **Teichmüller metric**, **Bergman metric**, *Cheng–Yau–Mok* **Kähler–Einstein metric**, **Carathéodory metric**, *McMullen metric*) are equivalent since they are **quasi-isometric** (Chap. 1) to the *Ricci metric* and the *perturbed Ricci metric* introduced by them.

- **Weil–Petersson metric**

The **Weil–Petersson metric** is a **Kähler metric** on the Teichmüller space $T_{g,n}$ of abstract Riemann surfaces of genus g with n punctures and negative Euler characteristic. This metric has negative Ricci curvature; it is **geodesically convex** (Chap. 1) and not complete.

The **Weil–Peterson metric** is **Gromov hyperbolic** if and only if (Brock and Farb, 2006) the complex dimension $3g - 3 + n$ of $T_{g,n}$ is at most two.

[redacted]

- **Gibbons–Manton metric**

  The **Gibbons–Manton metric** is a $4n$-dimensional **hyper-Kähler metric** on the moduli space of *n-monopoles* which admits an isometric action of the n-dimensional torus T^n. It is a hyper-Kähler quotient of a flat quaternionic vector space.

- **Metrics on determinant lines**

  Let M^n be an n-dimensional compact smooth manifold, and let F be a flat vector bundle over M^n. Let $H^{\bullet}(M^n, F) = \oplus_{i=0}^{n} H^i(M^n, F)$ be the *de Rham cohomology* of M^n with coefficients in F. Given an n-dimensional vector space V, the *determinant line det V* of V is defined as the top exterior power of V, i.e., $det\ V = \wedge^n V$. Given a finite-dimensional graded vector space $V = \oplus_{i=0}^{n} V_i$, the determinant line of V is defined as the tensor product $det\ V = \otimes_{i=0}^{n}(detV_i)^{(-1)^i}$.

  Thus, the determinant line $detH^{\bullet}(M^n, F)$ of the cogomology $H^{\bullet}(M^n, F)$ can be written as $detH^{\bullet}(M^n, F) = \otimes_{i=0}^{n}(detH^i(M^n, F))^{(-1)^i}$.

  The **Reidemeister metric** is a metric on $detH^{\bullet}(M^n, F)$, defined by a given smooth triangulation of M^n, and the classical *Reidemeister–Franz torsion.*

  Let g^F and $g^{T(M^n)}$ be smooth metrics on the vector bundle F and tangent bundle $T(M^n)$, respectively. These metrics induce a canonical **L_2-metric** $h^{H^{\bullet}(M^n,F)}$ on $H^{\bullet}(M^n, F)$. The **Ray–Singler metric** on $detH^{\bullet}(M^n, F)$ is defined as the product of the metric induced on $detH^{\bullet}(M^n, F)$ by $h^{H^{\bullet}(M^n,F)}$ with the *Ray–Singler analytic torsion.* The **Milnor metric** on $detH^{\bullet}(M^n, F)$ can be defined in a similar manner using the *Milnor analytic torsion.* If g^F is flat, the above two metrics coincide with the Reidemeister metric. Using a co-Euler structure, one can define a *modified Ray–Singler metric* on $detH^{\bullet}(M^n, F)$.

  The **Poincaré–Reidemeister metric** is a metric on the cohomological determinant line $detH^{\bullet}(M^n, F)$ of a closed connected oriented odd-dimensional manifold M^n. It can be constructed using a combination of the Reidemeister torsion with the Poincaré duality. Equivalently, one can define the *Poincaré–Reidemeister scalar product* on $detH^{\bullet}(M^n, F)$ which completely determines the Poincaré–Reidemeister metric but contains an additional sign or phase information.

  The **Quillen metric** is a metric on the inverse of the cohomological determinant line of a compact Hermitian one-dimensional complex manifold. It can be defined as the product of the L_2-metric with the Ray-Singler analytic torsion.

- **Kähler supermetric**

  The **Kähler supermetric** is a generalization of the **Kähler metric** for the case of a *supermanifold.* A *supermanifold* is a generalization of the usual manifold with *fermonic* as well as *bosonic* coordinates. The bosonic coordinates are ordinary numbers, whereas the fermonic coordinates are *Grassmann numbers.*

  Here the term *supermetric* differs from the one used in this chapter.

- **Hofer metric**

  A *symplectic manifold* (M^n, w), $n = 2k$, is a smooth even-dimensional manifold M^n equipped with a *symplectic form*, i.e, a closed nondegenerate 2-form, w.

A *Lagrangian manifold* is a k-dimensional smooth submanifold L^k of a symplectic manifold (M^n, w), $n = 2k$, such that the form w vanishes identically on L^k, i.e., for any $p \in L^k$ and any $x, y \in T_p(L^k)$, one has $w(x, y) = 0$.

Let $L(M^n, \Delta)$ be the set of all Lagrangian submanifolds of a closed symplectic manifold (M^n, w), diffeomorphic to a given Lagrangian submanifold Δ. A smooth family $\alpha = \{L_t\}_t$, $t \in [0, 1]$, of Lagrangian submanifolds $L_t \in L(M^n, \Delta)$ is called an *exact* path connecting L_0 and L_1, if there exists a smooth mapping $\Psi : \Delta \times [0, 1] \to M^n$ such that, for every $t \in [0, 1]$, one has $\Psi(\Delta \times \{t\}) = L_t$, and $\Psi * w = dH_t \wedge dt$ for some smooth function $H : \Delta \times [0, 1] \to \mathbb{R}$. The *Hofer length* $l(\alpha)$ of an exact path α is defined by $l(\alpha) = \int_0^1 \{\max_{p \in \Delta} H(p, t) - \min_{p \in \Delta} H(p, t)\} dt$.

The **Hofer metric** on the set $L(M^n, \Delta)$ is defined by

$$\inf_{\alpha} l(\alpha)$$

for any $L_0, L_1 \in L(M^n, \Delta)$, where the infimum is taken over all exact paths on $L(M^n, \Delta)$, that connect L_0 and L_1.

The Hofer metric can be defined similarly on the group $Ham(M^n, w)$ of *Hamiltonian diffeomorphisms* of a closed symplectic manifold (M^n, w), whose elements are *time-one mappings* of *Hamiltonian flows* ϕ_t^H: it is $\inf_\alpha l(\alpha)$, where the infimum is taken over all smooth paths $\alpha = \{\phi_t^H\}$, $t \in [0, 1]$, connecting ϕ and ψ.

- **Sasakian metric**

A **Sasakian metric** is a metric on a *contact manifold*, naturally adapted to the *contact structure*.

A contact manifold equipped with a Sasakian metric is called a *Sasakian space*, and it is an odd-dimensional analog of a *Kähler manifold*. The scalar curvature of a Sasakian metric which is also **Einstein metric**, is positive.

- **Cartan metric**

A *Killing form* (or *Cartan–Killing form*) on a finite-dimensional *Lie algebra* Ω over a field $\mathbb{F}$ is a symmetric bilinear form

$$B(x, y) = \mathrm{Tr}(ad_x \cdot ad_y),$$

where Tr denotes the trace of a linear operator, and ad_x is the image of x under *the adjoint representation* of Ω, i.e., the linear operator on the vector space Ω defined by the rule $z \to [x, z]$, where $[,]$ is the Lie bracket.

Let $e_1, \dots e_n$ be a basis for the Lie algebra Ω, and $[e_i, e_j] = \sum_{k=1}^n \gamma_{ij}^k e_k$, where γ_{ij}^k are corresponding *structure constants*. Then the Killing form is given by

$$B(x_i, x_j) = g_{ij} = \sum_{k,l=1}^{n} \gamma_{il}^k \gamma_{ik}^l.$$

In Theoretical Physics, the **metric tensor** $((g_{ij}))$ is called a **Cartan metric**.

[redacted]

Chapter 8
Distances on Surfaces and Knots

8.1 General Surface Metrics

A *surface* is a real 2D (two-dimensional) *manifold* M^2, i.e., a **Hausdorff space**, each point of which has a neighborhood which is homeomorphic to a plane $\mathbb{E}^2$, or a closed half-plane (cf. Chap. 7).

A compact orientable surface is called *closed* if it has no boundary, and it is called a *surface with boundary*, otherwise. There are compact nonorientable surfaces (closed or with boundary); the simplest such surface is the *Möbius strip*. Noncompact surfaces without boundary are called *open*.

Any closed connected surface is homeomorphic to either a sphere with, say, g (cylindric) handles, or a sphere with, say, g *cross-caps* (i.e., caps with a twist like Möbius strip in them). In both cases the number g is called the *genus* of the surface. In the case of handles, the surface is orientable; it is called a *torus* (doughnut), *double torus*, and *triple torus* for $g = 1, 2$ and 3, respectively. In the case of cross-caps, the surface is nonorientable; it is called the *real projective plane*, *Klein bottle*, and *Dyck's surface* for $g = 1, 2$ and 3, respectively. The genus is the maximal number of disjoint simple closed curves which can be cut from a surface without disconnecting it (the *Jordan curve theorem* for surfaces).

The *Euler–Poincaré characteristic* of a surface is (the same for all polyhedral decompositions of a given surface) the number $\chi = v - e + f$, where v, e and f are, respectively, the number of vertices, edges and faces of the decomposition. Then $\chi = 2 - 2g$ if the surface is orientable, and $\chi = 2 - g$ if not. Every surface with boundary is homeomorphic to a sphere with an appropriate number of (disjoint) *holes* (i.e., what remains if an open disk is removed) and handles or cross-caps. If h is the number of holes, then $\chi = 2 - 2g - h$ holds if the surface is orientable, and $\chi = 2 - g - h$ if not.

The *connectivity number* of a surface is the largest number of closed cuts that can be made on the surface without separating it into two or more parts. This number is equal to $3 - \chi$ for closed surfaces, and $2 - \chi$ for surfaces with boundaries. A surface

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_8

with connectivity number 1, 2 and 3 is called, respectively, *simply, doubly* and *triply connected*. A sphere is simply connected, while a torus is triply connected.

A surface can be considered as a metric space with its own **intrinsic metric**, or as a figure in space. A surface in $\mathbb{E}^3$ is called *complete* if it is a **complete** metric space with respect to its intrinsic metric.

Useful *shape-aware* (preserved by isomorphic deformations of the surface) distances on the interior of a surface mesh can be defined by isometric embedding of the surface into a suitable high-dimensional Euclidean space; for example, **diffusion metric** (cf. Chap. 15 and **histogram diffusion distance** from Chap. 21) and Rustamov et al., 2009.

A surface is called *differentiable*, *regular*, or *analytic*, respectively, if in a neighborhood of each of its points it can be given by an expression

$$r = r(u, v) = r(x_1(u, v), x_2(u, v), x_3(u, v)),$$

where the *position vector* $r = r(u, v)$ is a differentiable, *regular* (i.e., a sufficient number of times differentiable), or *real analytic*, respectively, vector function satisfying the condition $r_u \times r_v \neq 0$.

Any regular surface has the intrinsic metric with the *line element* (or *first fundamental form*)

$$ds^2 = dr^2 = E(u, v)du^2 + 2F(u, v)dudv + G(u, v)dv^2,$$

where $E(u, v) = \langle r_u, r_u \rangle$, $F(u, v) = \langle r_u, r_v \rangle$, $G(u, v) = \langle r_v, r_v \rangle$. The length of a curve defined on the surface by the equations $u = u(t)$, $v = v(t)$, $t \in [0, 1]$, is computed by

$$\int_0^1 \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}dt,$$

and the distance between any points $p, q \in M^2$ is defined as the infimum of the lengths of all curves on M^2, connecting p and q. A **Riemannian metric** is a generalization of the first fundamental form of a surface.

For surfaces, two kinds of *curvature* are considered: *Gaussian curvature*, and *mean curvature*. To compute these curvatures at a given point of the surface, consider the intersection of the surface with a plane, containing a fixed *normal vector*, i.e., a vector which is perpendicular to the surface at this point. This intersection is a plane curve. The *curvature* k of this plane curve is called the *normal curvature* of the surface at the given point. If we vary the plane, the normal curvature k will change, and there are two extremal values, the *maximal curvature* k_1, and the *minimal curvature* k_2, called the *principal curvatures* of the surface. A curvature is taken to be *positive* if the curve turns in the same direction as the surface's chosen normal, otherwise it is taken to be *negative*.

[redacted]

The *Gaussian curvature* is $K = k_1k_2$ (it can be given entirely in terms of the first fundamental form). The *mean curvature* is $H = \frac{1}{2}(k_1 + k_2)$.

A *minimal surface* is a surface with mean curvature zero or, equivalently, a surface of minimum area subject to constraints on the location of its boundary.

A *Riemann surface* is a one-dimensional *complex manifold*, or a 2D real manifold with a complex structure, i.e., in which the local coordinates in neighborhoods of points are related by complex analytic functions. It can be thought of as a deformed version of the complex plane. All Riemann surfaces are orientable. Closed Riemann surfaces are geometrical models of *complex algebraic curves*. Every connected Riemann surface can be turned into a *complete* 2D *Riemannian manifold* with constant curvature $-1, 0$, or 1. The Riemann surfaces with curvature -1 are called *hyperbolic*, and the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ is the canonical example. The Riemann surfaces with curvature 0 are called *parabolic*, and $\mathbb{C}$ is a typical example. The Riemann surfaces with curvature 1 are called *elliptic*, and the *Riemann sphere* $\mathbb{C} \cup \{\infty\}$ is a typical example.

- **Regular metric**

  The intrinsic metric of a surface is **regular** if it can be specified by the *line element*

  $$ds^2 = Edu^2 + 2Fdudv + Gdv^2,$$

  where the coefficients of the form ds^2 are regular functions.

  Any regular surface, given by an expression $r = r(u, v)$, has a regular metric with the *line element* ds^2, where $E(u, v) = \langle r_u, r_u\rangle$, $F(u, v) = \langle r_u, r_v\rangle$, $G(u, v) = \langle r_v, r_v\rangle$.

- **Analytic metric**

  The intrinsic metric on a surface is **analytic** if it can be specified by the *line element*

  $$ds^2 = Edu^2 + 2Fdudv + Gdv^2,$$

  where the coefficients of the form ds^2 are real analytic functions.

  Any analytic surface, given by an expression $r = r(u, v)$, has an analytic metric with the *line element* ds^2, where $E(u, v) = \langle r_u, r_u\rangle$, $F(u, v) = \langle r_u, r_v\rangle$, $G(u, v) = \langle r_v, r_v\rangle$.

- **Metric of nonpositive curvature**

  A **metric of nonpositive curvature** is the intrinsic metric on a *saddle-like surface*. A *saddle-like surface* is a generalization of a surface of negative curvature: a twice continuously-differentiable surface is a saddle-like surface if and only if at each point of the surface its Gaussian curvature is nonpositive.

  These surfaces can be seen as antipodes of *convex surfaces*, but they do not form such a natural class of surfaces as do convex surfaces.

  A **metric of negative curvature** is the intrinsic metric on a *surface of negative curvature*, i.e., a surface in $\mathbb{E}^3$ that has negative Gaussian curvature at every point.

A surface of negative curvature locally has a saddle-like structure. The intrinsic geometry of a surface of constant negative curvature (in particular, of a *pseudo-sphere*) locally coincides with the geometry of the *Lobachevsky plane*. There exists no surface in $\mathbb{E}^3$ whose intrinsic geometry coincides completely with the geometry of the Lobachevsky plane (i.e., a complete regular surface of constant negative curvature).

- **Metric of nonnegative curvature**

A **metric of nonnegative curvature** is the intrinsic metric on a *convex surface*.

A *convex surface* is a *domain* (i.e., a connected open set) on the boundary of a *convex body* in $\mathbb{E}^3$ (in some sense, it is an antipode of a saddle-like surface).

The entire boundary of a convex body is called a *complete convex surface*. If the body is finite (bounded), the complete convex surface is called *closed*. Otherwise, it is called *infinite* (an infinite convex surface is homeomorphic to a plane or to a circular cylinder).

Any convex surface M^2 in $\mathbb{E}^3$ is a *surface of bounded curvature*. The *total Gaussian curvature* $w(A) = \int\int_A K(x)d\sigma(x)$ of a set $A \subset M^2$ is always nonnegative (here $\sigma(.)$ is the *area*, and $K(x)$ is the *Gaussian curvature* of M^2 at a point x), i.e., a convex surface can be seen as a *surface of nonnegative curvature*.

The intrinsic metric of a convex surface is a **convex metric** (not to be confused with **metric convexity** from Chap. 1) in the sense of Surface Theory, i.e., it displays the *convexity condition*: the sum of the angles of any triangle whose sides are shortest curves is not less that π.

A **metric of positive curvature** is the intrinsic metric on a *surface of positive curvature*, i.e., a surface in $\mathbb{E}^3$ that has positive Gaussian curvature at every point.

- **Metric with alternating curvature**

A **metric with alternating curvature** is the intrinsic metric on a surface with alternating (positive or negative) Gaussian curvature.

- **Flat metric**

A **flat metric** is the intrinsic metric on a *developable surface*, i.e., a surface, on which the Gaussian curvature is everywhere zero. Cf. **flat space** in Chap. 1.

In general, a Riemannian metric on a surface is locally Euclidean up to a third order error (distortion of metric) measured by the Gaussian curvature.

- **Metric of bounded curvature**

A **metric of bounded curvature** is the intrinsic metric ρ on a *surface of bounded curvature*.

A surface M^2 with an intrinsic metric ρ is called a *surface of bounded curvature* if there exists a sequence of **Riemannian metrics** ρ_n defined on M^2, such that $\rho_n \to \rho$ uniformly for any compact set $A \subset M^2$, and the sequence $|w_n|(A)$ is bounded, where $|w|_n(A) = \int\int_A |K(x)|d\sigma(x)$ is the *total absolute curvature* of the metric ρ_n (here $K(x)$ is the Gaussian curvature of M^2 at a point x, and $\sigma(.)$ is the *area*).

- **Λ-metric**

A **Λ-metric** (or *metric of type* Λ) is a **complete** metric on a surface with curvature bounded from above by a negative constant.

[redacted]

A Λ-metric does not have embeddings into $\mathbb{E}^3$. It is a generalization of the result in Hilbert, 1901: no complete regular surface of constant negative curvature (i.e., a surface whose intrinsic geometry is the geometry of the Lobachevsky plane) exists in $\mathbb{E}^3$.

- **(h, Δ)-metric**

  A **(h, Δ)-metric** is a metric on a surface with a slowly-changing negative curvature.

  A **complete** (h, Δ)-metric does not permit a regular *isometric embedding* in three-dimensional Euclidean space (cf. **Λ-metric**).

- **G-distance**

  A connected set G of points on a surface M^2 is called a *geodesic region* if, for each point $x \in G$, there exists a *disk* $B(x, r)$ with center at x, such that $B_G = G \cap B(x, r)$ has one of the following forms: $B_G = B(x, r)$ (x is a *regular interior point* of G); B_G is a semidisk of $B(x, r)$ (x is a *regular boundary point* of G); B_G is a sector of $B(x, r)$ other than a semidisk (x is an *angular point* of G); B_G consists of a finite number of sectors of $B(x, r)$ with no common points except x (a *nodal point* of G).

  The **G-distance** between any x and $y \in G$ is the greatest lower bound of the lengths of all rectifiable curves connecting x and $y \in G$ and completely contained in G.

- **Conformally invariant metric**

  Let R be a Riemann surface. A *local parameter* (or *local uniformizing parameter*, *local uniformizer*) is a complex variable z considered as a continuous function $z_{p_0} = \phi_{p_0}(p)$ of a point $p \in R$ which is defined everywhere in some neighborhood (*parametric neighborhood*) $V(p_0)$ of a point $p_0 \in R$ and which realizes a homeomorphic mapping (*parametric mapping*) of $V(p_0)$ onto the disk (*parametric disk*) $\Delta(p_0) = \{z \in \mathbb{C} : |z| < r(p_0)\}$, where $\phi_{p_0}(p_0) = 0$. Under a parametric mapping, any point function $g(p)$ defined in the parametric neighborhood $V(p_0)$, goes into a function of the local parameter z: $g(p) = g(\phi_{p_0}^{-1}(z)) = G(z)$.

  A **conformally invariant metric** is a differential $\rho(z)|dz|$ on the Riemann surface R which is invariant with respect to the choice of the local parameter z. Thus, to each local parameter z ($z : U \to \overline{\mathbb{C}}$) a function $\rho_z : z(U) \to [0, \infty]$ is associated such that, for any local parameters z_1 and z_2, we have

$$\frac{\rho_{z_2}(z_2(p))}{\rho_{z_1}(z_1(p))} = \left|\frac{dz_1(p)}{dz_2(p)}\right| \text{ for any } p \in U_1 \cap U_2.$$

  Every linear differential $\lambda(z)dz$ and every *quadratic differential* $Q(z)dz^2$ induce conformally invariant metrics $|\lambda(z)||dz|$ and $|Q(z)|^{1/2}|dz|$, respectively (cf. **Q-metric**).

- **Q-metric**

  An **Q-metric** is a **conformally invariant metric** $\rho(z)|dz| = |Q(z)|^{1/2}|dz|$ on a Riemann surface R defined by a *quadratic differential* $Q(z)dz^2$.

A *quadratic differential* $Q(z)dz^2$ is a nonlinear differential on a Riemann surface R which is invariant with respect to the choice of the local parameter z. Thus, to each local parameter z ($z : U \to \overline{\mathbb{C}}$) a function $Q_z : z(U) \to \overline{\mathbb{C}}$ is associated such that, for any local parameters z_1 and z_2, we have

$$\frac{Q_{z_2}(z_2(p))}{Q_{z_1}(z_1(p))} = \left(\frac{dz_1(p)}{dz_2(p)}\right)^2 \text{ for any } p \in U_1 \cap U_2.$$

- **Extremal metric**

  Let Γ be a family of locally rectifiable curves on a Riemann surface R and let P be a class of **conformally invariant metrics** $\rho(z)|dz|$ on R such that $\rho(z)$ is square-integrable in the z-plane for every local parameter z, and the following Lebesgue integrals are not simultaneously equal to 0 or ∞:

  $$A_\rho(R) = \int\int_R \rho^2(z)dxdy \text{ and } L_\rho(\Gamma) = \inf_{\gamma\in\Gamma} \int_y \rho(z)|dz|.$$

  The *modulus of the family of curves* Γ is defined by

  $$M(\Gamma) = \inf_{\rho\in P} \frac{A_\rho(R)}{(L_\rho(\Gamma))^2}.$$

  The *extremal length of the family of curves* Γ is the reciprocal of $M(\Gamma)$.

  Let P_L be the subclass of P such that, for any $\rho(z)|dz| \in P_L$ and any $\gamma \in \Gamma$, one has $\int_\gamma \rho(z)|dz| \geq 1$. If $P_L \neq \emptyset$, then $M(\Gamma) = \inf_{\rho\in P_L} A_\rho(R)$. Every metric from P_L is called an *admissible metric* for the modulus on Γ. If there exists ρ^* for which

  $$M(\Gamma) = \inf_{\rho\in P_L} A_\rho(R) = A_{\rho^*}(R),$$

  the metric $\rho^*|dz|$ is called an **extremal metric** for the modulus on Γ. It is a **conformally invariant metric**.
- **Fréchet surface metric**

  Let (X, d) be a metric space, M^2 a compact 2D manifold, f a continuous mapping $f : M^2 \to X$, called a *parametrized surface*, and $\sigma : M^2 \to M^2$ a homeomorphism of M^2 onto itself. Two parametrized surfaces f_1 and f_2 are called *equivalent* if $\inf_\sigma \max_{p\in M^2} d(f_1(p), f_2(\sigma(p))) = 0$, where the infimum is taken over all possible homeomorphisms σ. A class f^* of parametrized surfaces, equivalent to f, is called a *Fréchet surface*. It is a generalization of the notion of a surface in Euclidean space to the case of an arbitrary metric space (X, d).

  The **Fréchet surface metric** on the set of all Fréchet surfaces is defined by

  $$\inf_\sigma \max_{p\in M^2} d(f_1(p), f_2(\sigma(p)))$$

[redacted]

for any Fréchet surfaces f_1^* and f_2^*, where the infimum is taken over all possible homeomorphisms σ. Cf. the **Fréchet metric** in Chap. 1.

- **Hempel metric**

  A *handlebody* of genus g is the boundary sum of g copies of a solid torus; it is homeomorphic to the closure of a regular neighborhood of some finite graph in $\mathbb{R}^3$. Given a closed orientable 3-manifold M, its *Heegaard splitting* (of genus g) is $M = A \cup_P B$ where A, B are genus g handlebodies in M such that $M = A \cup B$ and $A \cap B = \partial A = \partial B = P$. Then P is called a (genus g) *Heegaard surface* of M. In knot applications, Heegaard splitting of the *exterior of* a knot K (the complement of an open solid torus knotted like K) are considered.

  Two embedded curves are *isotopic* if there exists a continuous deformation of one embedding to another through a path of embeddings. Given a closed connected orientable surface S of genus at least two, let $C(S) = (V, E)$ denotes the graph whose vertices are isotopy classes of *essential* (not bounding disk on the surface) simple closed curves and whose edges are drawn between vertices with disjoint representative curves. This graph is connected. For any subsets of vertices $X, Y \subset V$, denote by $d_S(X, Y)$ their **set-to-set distance** $\min d_S(x, y) : x \in X, y \in Y$, where $d_S(x, y)$ is the **path metric** of $C(S)$.

  If S is the boundary of a handlebody H, let $M(H)$ denotes the set of vertices with representatives bounding *meridian disks* D of H, i.e., such that ∂D are essential simple closed curves in ∂H. The **Hempel distance** of a (genus $g \geq 2$) Heegaard splitting $M = A \cup_P B$ is defined (Hempel, 2001) to be $d_P(M(A), M(B))$.

  A Heegaard splitting $M = A \cup_P B$ is *stabilized*, if there are meridian disks D_A, D_B of A, B respectively such that ∂D_A and ∂D_B intersects transversely in a single point. The **Reidemeister–Singer distance** between two Heegaard surfaces/splittings is the minimal number of *stabilizations* (roughly, additions of a "trivial" handle) and *destabilizations* (inverse operation) relating them.

8.2 Intrinsic Metrics on Surfaces

In this section we list intrinsic metrics, given by their *line elements* (which, in fact, are 2D **Riemannian metrics**), for some selected surfaces.

- **Quadric metric**

  A *quadric* (or *quadratic surface*, *surface of second-order*) is a set of points in $\mathbb{E}^3$, whose coordinates in a Cartesian coordinate system satisfy an algebraic equation of degree two. There are 17 classes of such surfaces. Among them are: *ellipsoids*, *one-sheet* and *two-sheet hyperboloids*, *elliptic paraboloids*, *hyperbolic paraboloids*, *elliptic*, *hyperbolic* and *parabolic cylinders*, and *conical surfaces*.

  For example, a *cylinder* can be given by the following parametric equations:

$$x_1(u, v) = a \cos v, \ \ x_2(u, v) = a \sin v, \ \ x_3(u, v) = u.$$

The intrinsic metric on it is given by the *line element*

$$ds^2 = du^2 + a^2dv^2.$$

An *elliptic cone* (i.e., a cone with elliptical cross-section) has the following equations:

$$x_1(u,v) = a\frac{h-u}{h}\cos v,\ x_2(u,v) = b\frac{h-u}{h}\sin v,\ x_3(u,v) = u,$$

where h is the *height*, a is the *semimajor axis*, and b is the *semiminor axis* of the cone. The intrinsic metric on it is given by the *line element*

$$ds^2 = \frac{h^2 + a^2\cos^2 v + b^2\sin^2 v}{h^2}du^2 + 2\frac{(a^2-b^2)(h-u)\cos v\sin v}{h^2}dudv +$$
$$+\frac{(h-u)^2(a^2\sin^2 v + b^2\cos^2 v)}{h^2}dv^2.$$

- **Sphere metric**

  A *sphere* is a *quadric*, given by the Cartesian equation $(x_1-a)^2 + (x_2-b)^2 + (x_3-c)^2 = r^2$, where the point (a,b,c) is the *center* of the sphere, and $r > 0$ is the *radius* of the sphere. The sphere of radius r, centered at the origin, can be given by the following parametric equations:

  $$x_1(\theta,\phi) = r\sin\theta\cos\phi,\ x_2(\theta,\phi) = r\sin\theta\sin\phi,\ x_3(\theta,\phi) = r\cos\theta,$$

  where the *azimuthal angle* $\phi \in [0, 2\pi)$, and the *polar angle* $\theta \in [0,\pi]$.

  The intrinsic metric on it (in fact, the 2D **spherical metric**) is given by the *line element*

  $$ds^2 = r^2d\theta^2 + r^2\sin^2\theta d\phi^2.$$

  A sphere of radius r has constant positive Gaussian curvature equal to r.
- **Ellipsoid metric**

  An *ellipsoid* is a *quadric* given by the Cartesian equation $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2} = 1$, or by the following parametric equations:

  $$x_1(\theta,\phi) = a\cos\phi\sin\theta,\ x_2(\theta,\phi) = b\sin\phi\sin\theta,\ x_3(\theta,\phi) = c\cos\theta,$$

  where the *azimuthal angle* $\phi \in [0, 2\pi)$, and the *polar angle* $\theta \in [0,\pi]$.

  The intrinsic metric on it is given by the *line element*

  $$ds^2 = (b^2\cos^2\phi + a^2\sin^2\phi)\sin^2\theta d\phi^2 + (b^2-a^2)\cos\phi\sin\phi\cos\theta\sin\theta d\theta d\phi+$$
  $$+((a^2\cos^2\phi + b^2\sin^2\phi)\cos^2\theta + c^2\sin^2\theta)d\theta^2.$$

[redacted]

- **Spheroid metric**

  A *spheroid* is an *ellipsoid* having two axes of equal length. It is also a *rotation surface*, given by the following parametric equations:

  $$x_1(u,v) = a\sin v\cos u,\ x_2(u,v) = a\sin v\sin u,\ x_3(u,v) = c\cos v,$$

  where $0 \le u < 2\pi$, and $0 \le v \le \pi$.

  The intrinsic metric on it is given by the *line element*

  $$ds^2 = a^2\sin^2 v du^2 + \frac{1}{2}(a^2 + c^2 + (a^2 - c^2)\cos(2v))dv^2.$$

- **Hyperboloid metric**

  A *hyperboloid* is a *quadric* which may be one- or two-sheeted.

  The one-sheeted hyperboloid is a *surface of revolution* obtained by rotating a hyperbola about the perpendicular bisector to the line between the foci, while the two-sheeted hyperboloid is a surface of revolution obtained by rotating a hyperbola about the line joining the foci.

  The one-sheeted circular hyperboloid, oriented along the x_3 axis, is given by the Cartesian equation $\frac{x_1^2}{a^2} + \frac{x_2^2}{a^2} - \frac{x_3^2}{c^2} = 1$, or by the following parametric equations:

  $$x_1(u,v) = a\sqrt{1+u^2}\cos v,\ x_2(u,v) = a\sqrt{1+u^2}\sin v,\ x_3(u,v) = cu,$$

  where $v \in [0, 2\pi)$. The intrinsic metric on it is given by the *line element*

  $$ds^2 = \left(c^2 + \frac{a^2u^2}{u^2+1}\right)du^2 + a^2(u^2+1)dv^2.$$

- **Rotation surface metric**

  A *rotation surface* (or *surface of revolution*) is a surface generated by rotating a 2D curve about an axis. It is given by the following parametric equations:

  $$x_1(u,v) = \phi(v)\cos u,\ x_2(u,v) = \phi(v)\sin u,\ x_3(u,v) = \psi(v).$$

  The intrinsic metric on it is given by the *line element*

  $$ds^2 = \phi^2 du^2 + (\phi'^2 + \psi'^2)dv^2.$$

- **Pseudo-sphere metric**

  A *pseudo-sphere* is a half of the *rotation surface* generated by rotating a *tractrix* about its asymptote. It is given by the following parametric equations:

  $$x_1(u,v) = \mathrm{sech} u\cos v,\ x_2(u,v) = \mathrm{sech} u\sin v,\ x_3(u,v) = u - \tanh u,$$

where $u \geq 0$, and $0 \leq v < 2\pi$. The intrinsic metric on it is given by the *line element*

$$ds^2 = \tanh^2 u du^2 + \text{sech}^2 u dv^2.$$

The pseudo-sphere has constant negative Gaussian curvature equal to -1, and in this sense is an analog of the sphere which has constant positive Gaussian curvature.

- **Torus metric**

  A *torus* is a surface having genus one. A torus azimuthally symmetric about the x_3 axis is given by the Cartesian equation $(c - \sqrt{x_1^2 + x_2^2})^2 + x_3^2 = a^2$, or by the following parametric equations:

  $$x_1(u, v) = (c + a\cos v)\cos u,\ x_2(u, v) = (c + a\cos v)\sin u,\ x_3(u, v) = a\sin v,$$

  where $c > a$, and $u, v \in [0, 2\pi)$.

  The intrinsic metric on it is given by the *line element*

  $$ds^2 = (c + a\cos v)^2 du^2 + a^2 dv^2.$$

  For toroidally confined plasma, such as in magnetic confinement fusion, the coordinates u, v and a correspond to the directions called , respectively, *toroidal* (long, as lines of latitude, way around the torus), *poloidal* (short way around the torus) and *radial*. The **poloidal distance**, used in plasma context, is the distance in the poloidal direction.

- **Helical surface metric**

  A *helical surface* (or *surface of screw motion*) is a surface described by a plane curve γ which, while rotating around an axis at a uniform rate, also advances along that axis at a uniform rate. If γ is located in the plane of the axis of rotation x_3 and is defined by the equation $x_3 = f(u)$, the position vector of the helical surface is

  $$r = (u\cos v,\ u\sin v,\ f(u) = hv),\ h = \text{const},$$

  and the intrinsic metric on it is given by the *line element*

  $$ds^2 = (1 + f'^2)du^2 + 2hf' dudv + (u^2 + h^2)dv^2.$$

  If $f = const$, one has a *helicoid*; if $h = 0$, one has a *rotation surface*.

- **Catalan surface metric**

  The *Catalan surface* is a *minimal surface*, given by the following equations:

  $$x_1(u, v) = u - \sin u \cosh v,\ x_2(u, v) = 1 - \cos u \cosh v,\ x_3(u, v) = 4\sin\left(\frac{u}{2}\right)\sinh\left(\frac{v}{2}\right).$$

[redacted]

The intrinsic metric on it is given by the *line element*

$$ds^2 = 2\cosh^2\left(\frac{v}{2}\right)(\cosh v - \cos u)du^2 + 2\cosh^2(\frac{v}{2})\,(\cosh v - \cos u)\,dv^2.$$

- **Monkey saddle metric**

  The *monkey saddle* is a surface, given by the Cartesian equation $x_3 = x_1(x_1^2 - 3x_2^2)$, or by the following parametric equations:

  $$x_1(u, v) = u,\ x_2(u, v) = v,\ x_3(u, v) = u^3 - 3uv^2.$$

  This is a surface which a monkey can straddle with both legs and his tail. The intrinsic metric on it is given by the *line element*

  $$ds^2 = (1 + (su^2 - 3v^2)^2)du^2 - 2(18uv(u^2 - v^2))dudv + (1 + 36u^2v^2)dv^2).$$

- **Distance-defined surfaces and curves**

  We give below examples of plane curves and surfaces which are the loci of points with given value of some function of their Euclidean distances to the given objects.

  A *parabola* is the locus of all points in $\mathbb{R}^2$ that are equidistant from the given point (*focus*) and given line (*directrix*) on the plane.

  A *hyperbola* is the locus of all points in $\mathbb{R}^2$ such that the ratio of their distances to the given point and line is a constant (*eccentricity*) greater than 1. It is also the locus of all points in $\mathbb{R}^2$ such that the absolute value of the difference of their distances to the two given foci is constant.

  An *ellipse* is the locus of all points in $\mathbb{R}^2$ such that the sum of their distances to the two given points (*foci*) is constant; cf. **elliptic orbit distance** in Chap. 25. A *circle* is an ellipse in which the two foci are coincident.

  A *Cassini oval* is the locus of all points in $\mathbb{R}^2$ such that the product of their distances to two given points is a constant k. If the distance between two points is $2\sqrt{k}$, then such oval is called a *lemniscate of Bernoulli*.

  A *circle of Appolonius* is the locus of points in $\mathbb{R}^2$ such that the ratio of their distances to the first and second given points is constant.

  A *Cartesian oval* is the locus of points in $\mathbb{R}^2$ such that their distances r_1, r_2 to the foci $(-1, 0), (1, 0)$ are related linearly by $ar_1 + br_2 = 1$. The cases $a = b, a = -b$ and $a = \frac{1}{2}$ (or $b = \frac{1}{2}$) correspond to the ellipse, hyperbola and *limaçon of Pascal*, respectively.

  A *Cassinian curve* is the locus of all points in $\mathbb{R}^2$ such that the product of their distances to n given points (*poles*) is constant. If the poles form a regular n-gon, then this (algebraic of degree $2n$) curve is a *sinusoidal spiral* given also by polar equation $r^n = 2\cos(n\theta)$, and the case $n = 3$ corresponds to the *Kiepert curve*.

  Farouki and Moon, 2000, considered many other multipolar generalizations of above curves. For example, their *trifocal ellipse* is the locus of all points in

$\mathbb{R}^2$ (seen as the complex plane) such that the sum of their distances to the 3 cube roots of unity is a constant k. If $k = 2\sqrt{3}$, the curve pass through (and is singular at) the 3 poles.

In $\mathbb{R}^3$, a surface, rotationally symmetric about an axis, is a locus defined via Euclidean distances of its points to the two given poles belonging to this axis. For example, a *spheroid* (or *ellipsoid of revolution*) is a quadric obtained by rotating an ellipse about one of its principal axes.

It is a sphere, if this ellipse is a circle. If the ellipse is rotated about its major axis, the result is an elongated (as the rugby ball) spheroid which is the locus of all points in $\mathbb{R}^3$ such that the sum of their distances to the two given points is constant. The rotation about its minor axis results in a flattened spheroid (as the Earth) which is the locus of all points in $\mathbb{R}^3$ such that the sum of the distances to the closest and the farthest points of given circle is constant.

A *hyperboloid of revolution of two sheets* is a quadric obtained by revolving a hyperbola about its semimajor (real) axis. Such hyperboloid with axis AB is the locus of all points in $\mathbb{R}^3$ such that the absolute value of the difference of their distances to the points A and B is constant.

Any point in $\mathbb{R}^n$ is uniquely defined by its Euclidean distances to the vertices of a nondegenerated n-simplex. If a surface which is not rotationally symmetric about an axis, is a locus in $\mathbb{R}^3$ defined via distances of its points to the given poles, then three noncollinear poles is needed, and the surface is symmetric with respect to reflection in the plane defined by the three poles.

8.3 Distances on Knots

A *knot* is a closed, self-nonintersecting curve that is embedded in S^3. The *trivial knot* (or *unknot*) O is a closed loop that is not knotted. A knot can be generalized to a link which is a set of disjoint knots. Every link has its *Seifert surface*, i.e., a compact oriented surface with the given link as boundary.

Two knots (links) are called *equivalent* if one can be smoothly deformed into another. Formally, a link is defined as a smooth one-dimensional *submanifold* of the 3-sphere S^3; a knot is a link consisting of one component; two links L_1 and L_2 are called *equivalent* if there exists an orientation-preserving homeomorphism $f : S^3 \to S^3$ such that $f(L_1) = L_2$.

All the information about a knot can be described using a *knot diagram*. It is a projection of a knot onto a plane such that no more than two points of the knot are projected to the same point on the plane, and at each such point it is indicated which strand is closest to the plane, usually by erasing part of the lower strand. Two different knot diagrams may both represent the same knot. Much of Knot Theory is devoted to telling when two knot diagrams represent the same knot.

[redacted]

An *unknotting operation* is an operation which changes the overcrossing and the undercrossing at a double point of a given knot diagram. The *unknotting number* of a knot K is the minimum number of unknotting operations needed to deform a diagram of K into that of the trivial knot, where the minimum is taken over all diagrams of K. Roughly, the unknotting number is the smallest number of times a knot K must be passed through itself to untie it.

An $\sharp$-*unknotting operation* in a diagram of a knot K is an analog of the unknotting operation for a $\sharp$-*part* of the diagram consisting of two pairs of parallel strands with one of the pair overcrossing another. Thus, an $\sharp$-unknotting operation changes the overcrossing and the undercrossing at each vertex of obtained quadrangle.

- **Gordian distance**

  The **Gordian distance** is a metric on the set of all knots defined, for given knots K and K', as the minimum number of unknotting operations needed to deform a diagram of K into that of K', where the minimum is taken over all diagrams of K from which one can obtain diagrams of K'. The unknotting number of K is equal to the Gordian distance between K and the trivial knot O.

  Let rK be the knot obtained from K by taking its mirror image, and let $-K$ be the knot with the reversed orientation. The **positive reflection distance** $Ref_+(K)$ is the Gordian distance between K and rK. The **negative reflection distance** $Ref_-(K)$ is the Gordian distance between K and $-rK$. The **inversion distance** $Inv(K)$ is the Gordian distance between K and $-K$.

  The Gordian distance is the case $k = 1$ of the C_k**-distance** which is the minimum number of C_k-*moves* needed to transform K into K'; Habiro, 1994 and Goussarov, 1995, independently proved that, for $k > 1$, it is finite if and only if both knots have the same *Vassiliev invariants of order less than k*. A C_1-move is a single crossing change, a C_2-move (or *delta-move*) is a simultaneous crossing change for 3 arcs forming a triangle. C_2- and C_3-distances are called **delta distance** and **clasp-pass distance**, respectively.
- **$\sharp$-Gordian distance**

  The **$\sharp$-Gordian distance** (see, for example, [Mura85]) is a metric on the set of all knots defined, for given knots K and K', as the minimum number of $\sharp$-unknotting operations needed to deform a diagram of K into that of K', where the minimum is taken over all diagrams of K from which one can obtain diagrams of K'.

  Let rK be the knot obtained from K by taking its mirror image, and let $-K$ be the knot with the reversed orientation. The **positive $\sharp$-reflection distance** $Ref^{\sharp}_{+}(K)$ is the $\sharp$-Gordian distance between K and rK. The **negative $\sharp$-reflection distance** $Ref^{\sharp}_{-}(K)$ is the $\sharp$-Gordian distance between K and $-rK$. The **$\sharp$-inversion distance** $Inv^{\sharp}(K)$ is the $\sharp$-Gordian distance between K and $-K$.
- **Knot complement hyperbolic metric**

  The *complement* of a knot K (or a link L) is $S^3\backslash K$ (or $S^3\backslash L$, respectively).

A knot (or, in general, a link) is called *hyperbolic* if its complement supports a complete Riemannian metric of constant curvature -1. In this case, the metric is called a **knot** (or **link**) **complement hyperbolic metric**, and it is unique.

A knot is hyperbolic if and only if (Thurston, 1978) it is not a *satellite knot* (then it supports a complete locally homogeneous Riemannian metric) and not a *torus knot* (does not lie on a trivially embedded torus in S^3). The complement of any nontrivial knot supports a complete nonpositively curved Riemannian metric.

[redacted]

Chapter 9
Distances on Convex Bodies, Cones, and Simplicial Complexes

9.1 Distances on Convex Bodies

A *convex body* in the n-dimensional Euclidean space $\mathbb{E}^n$ is a convex *compact connected* subset of $\mathbb{E}^n$. It is called *solid* (or *proper*) if it has nonempty interior. Let K denote the space of all convex bodies in $\mathbb{E}^n$, and let K_p be the subspace of all proper convex bodies. Given a set $X \subset \mathbb{E}^n$, its *convex hull* $conv(X)$ is the minimal *convex* set containing X.

Any metric space (K, d) on K is called a *metric space of convex bodies*. Such spaces, in particular the metrization by the **Hausdorff metric**, or by the **symmetric difference metric**, play a basic role in Convex Geometry (see, for example, [Grub93]).

For $C, D \in K \backslash \{\emptyset\}$, the *Minkowski addition* and the *Minkowski nonnegative scalar multiplication* are defined by $C + D = \{x + y : x \in C, y \in D\}$, and $\alpha C = \{\alpha x : x \in C\}$, $\alpha \geq 0$, respectively. The Abelian semigroup $(K, +)$ equipped with nonnegative scalar multiplication operators can be considered as a *convex cone*.

The *support function* $h_C : S^{n-1} \to \mathbb{R}$ of $C \in K$ is defined by $h_C(u) = \sup\{\langle u, x\rangle : x \in C\}$ for any $u \in S^{n-1}$, where S^{n-1} is the $(n-1)$-dimensional *unit sphere* in $\mathbb{E}^n$, and $\langle , \rangle$ is the *inner product* in $\mathbb{E}^n$. The *width* $w_C(u)$ is $h_C(u) + h_C(-u) = h_{C-C}(u)$. It is the perpendicular distance between the parallel supporting hyperplanes perpendicular to given direction. The *mean width* is the average of width over all directions in S^{n-1}.

- **Area deviation**

  The **area deviation** (or **template metric**) is a metric on the set K_p in $\mathbb{E}^2$ (i.e., on the set of plane convex disks) defined by

$$A(C \triangle D),$$

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_9

where $A(.)$ is the *area*, and $\triangle$ is the *symmetric difference*. If $C \subset D$, then it is equal to $A(D) - A(C)$.

- **Perimeter deviation**

  The **perimeter deviation** is a metric on K_p in $\mathbb{E}^2$ defined by

$$2p(conv(C \cup D)) - p(C) - p(D),$$

where $p(.)$ is the *perimeter*. In the case $C \subset D$, it is equal to $p(D) - p(C)$.

- **Mean width metric**

  The **mean width metric** is a metric on K_p in $\mathbb{E}^2$ defined by

$$v2W(conv(C \cup D)) - W(C) - W(D),$$

where $W(.)$ is the *mean width*: $W(C) = p(C)/\pi$, and $p(.)$ is the *perimeter*.

- **Florian metric**

  The **Florian metric** is a metric on K defined by

$$\int_{S^{n-1}} |h_C(u) - h_D(u)| d\sigma(u) = ||h_C - h_D||_1.$$

It can be expressed in the form $2S(conv(C \cup D)) - S(C) - S(D)$ for $n = 2$ (cf. **perimeter deviation**); it can be expressed also in the form $nk_n(2W(conv(C \cup D)) - W(C) - W(D))$ for $n \geq 2$ (cf. **mean width metric**).

Here $S(.)$ is the *surface area*, k_n is the *volume* of the *unit ball* $\overline{B}^n$ of $\mathbb{E}^n$, and $W(.)$ is the *mean width*: $W(C) = \frac{1}{nk_n} \int_{S^{n-1}} (h_C(u) + h_C(-u)) d\sigma(u)$.

- **McClure–Vitale metric**

  Given $1 \leq p \leq \infty$, the **McClure–Vitale metric** is a metric on K, defined by

$$\left(\int_{S^{n-1}} |h_C(u) - h_D(u)|^p d\sigma(u) \right)^{\frac{1}{p}} = ||h_C - h_D||_p.$$

- **Pompeiu–Hausdorff–Blaschke metric**

  The **Pompeiu–Hausdorff–Blaschke metric** is a metric on K defined by

$$\max\{\sup_{x \in C} \inf_{y \in D} ||x - y||_2, \sup_{y \in D} \inf_{x \in C} ||x - y||_2\},$$

where $||.||_2$ is the Euclidean norm on $\mathbb{E}^n$.

In terms of support functions and using Minkowski addition, this metric is

$$\sup_{u \in S^{n-1}} |h_C(u) - h_D(u)| = ||h_C - h_D||_\infty = \inf\{\lambda \geq 0 : C \subset D + \lambda \overline{B}^n, D \subset C + \lambda \overline{B}^n\},$$

where $\overline{B}^n$ is the *unit ball* of $\mathbb{E}^n$. This metric can be defined using any norm on $\mathbb{R}^n$ and for the space of bounded closed subsets of any metric space.

[redacted]

- **Pompeiu–Eggleston metric**

  The **Pompeiu–Eggleston metric** is a metric on K defined by

$$\sup_{x\in C}\inf_{y\in D}||x-y||_2+\sup_{y\in D}\inf_{x\in C}||x-y||_2,$$

  where $||.||_2$ is the Euclidean norm on $\mathbb{E}^n$.

  In terms of support functions and using Minkowski addition, this metric is

$$\max\{0,\sup_{u\in S^{n-1}}(h_C(u)-h_D(u))\}+\max\{0,\sup_{u\in S^{n-1}}(h_D(u)-h_C(u))\}=$$
$$=\inf\{\lambda\geq 0:C\subset D+\lambda\overline{B}^n\}+\inf\{\lambda\geq 0:D\subset C+\lambda\overline{B}^n\},$$

  where $\overline{B}^n$ is the *unit ball* of $\mathbb{E}^n$. This metric can be defined using any norm on $\mathbb{R}^n$ and for the space of bounded closed subsets of any metric space.
- **Sobolev distance**

  The **Sobolev distance** is a metric on K defined by

$$||h_C-h_D||_w,$$

  where $||.||_w$ is the *Sobolev* 1*-norm* on the set $G_{S^{n-1}}$ of all real continuous functions on the *unit sphere* S^{n-1} of $\mathbb{E}^n$.

  The *Sobolev* 1*-norm* is defined by $||f||_w=\langle f,f\rangle_w^{1/2}$, where $\langle,\rangle_w$ is an *inner product* on $G_{S^{n-1}}$, given by

$$\langle f,g\rangle_w=\int_{S^{n-1}}(fg+\nabla_s(f,g))dw_0,\ \ w_0=\frac{1}{n\cdot k_n}w,$$

  where $\nabla_s(f,g)=\langle grad_s f,\ grad_s g\rangle$, $\langle,\rangle$ is the *inner product* in $\mathbb{E}^n$, and $grad_s$ is the *gradient* on S^{n-1} (see [ArWe92]).
- **Shephard metric**

  The **Shephard metric** is a metric on K_p defined by

$$\ln(1+2\inf\{\lambda\geq 0:C\subset D+\lambda(D-D),D\subset C+\lambda(C-C)\}).$$

- **Nikodym metric**

  The **Nikodym metric** (or **volume of symmetric difference**, **Dinghas distance**) is a metric on K_p defined by

$$V(C\triangle D)=\int(1_{x\in C}-1_{x\in D})^2dx,$$

  where $V(.)$ is the *volume* (i.e., the Lebesgue n-dimensional measure), and $\triangle$ is the *symmetric difference*. For $n=2$, one obtains the **area deviation**.

Normalized volume of symmetric difference is a variant of **Steinhaus distance** defined by

$$\frac{V(C\triangle D)}{V(C\cup D)}.$$

- **Eggleston distance**

The **Eggleston distance** (or **symmetric surface area deviation**) is a distance on K_p defined by

$$S(C\cup D)-S(C\cap D),$$

where $S(.)$ is the *surface area*. It is not a metric.

- **Asplund metric**

The **Asplund metric** is a metric on the space $K_p/\approx$ of affine-equivalence classes in K_p defined by

$$\ln\inf\{\lambda\geq 1:\exists T:\mathbb{E}^n\to\mathbb{E}^n \text{ affine}, x\in\mathbb{E}^n, C\subset T(D)\subset\lambda C+x\}$$

for any equivalence classes C^* and D^* with the representatives C and D, respectively.

- **Macbeath metric**

The **Macbeath metric** is a metric on the space $K_p/\approx$ of affine-equivalence classes in K_p defined by

$$\ln\inf\{|\det T\cdot P|:\exists T,P:\mathbb{E}^n\to\mathbb{E}^n \text{ regular affine}, C\subset T(D), D\subset P(C)\}$$

for any equivalence classes C^* and D^* with the representatives C and D, respectively.

Equivalently, it can be written as $\ln\delta(C,D)+\ ln\delta(D,C)$, where $\delta(C,D)=\inf_T\{\frac{V(T(D))}{V(C)};C\subset T(D)\}$, and T is a regular affine mapping of $\mathbb{E}^n$ onto itself.

- **Banach–Mazur metric**

The **Banach–Mazur metric** is a metric on the space $K_{po}/\sim$ of the equivalence classes of proper 0-symmetric convex bodies with respect to linear transformations defined by

$$\ln\inf\{\lambda\geq 1:\exists T:\mathbb{E}^n\to\mathbb{E}^n \text{ linear}, C\subset T(D)\subset\lambda C\}$$

for any equivalence classes C^* and D^* with the representatives C and D, respectively.

It is a special case of the **Banach–Mazur distance** (Chap. 1).

- **Separation distance**

The **separation distance** between two disjoint convex bodies C and D in $\mathbb{E}^n$ (in general, between any two disjoint subsets) $\mathbb{E}^n$) is (Buckley, 1985) their

[redacted]

Euclidean **set-set distance** $\inf\{||x-y||_2 : x \in C, y \in D\}$, while $\sup\{||x-y||_2 : x \in C, y \in D\}$ is their **spanning distance**.

- **Penetration depth distance**

  The **penetration depth distance** between two interpenetrating convex bodies C and D in $\mathbb{E}^n$ (in general, between any two interpenetrating subsets of $\mathbb{E}^n$) is (Cameron–Culley, 1986) defined as the minimum *translation distance* that one body undergoes to make the interiors of C and D disjoint:

  $$\min\{||t||_2 : interior(C+t) \cap D = \emptyset\}.$$

  Keerthi–Sridharan, 1991, considered $||t||_1$- and $||t||_\infty$-analogs of this distance.

  Cf. **penetration distance** in Chap. 23 and **penetration depth** in Chap. 24.

- **Growth distances**

  Let $C, D \in K_p$ be two compact convex proper bodies. Fix their *seed points* $p_C \in int\,C$ and $p_D \in int\,D$; usually, they are the centroids of C and D. The *growth function* $g(C, D)$ is the minimal number $\lambda > 0$, such that

  $$(\{p_C\} + \lambda(C\backslash\{p_C\})) \cap (\{p_D\} + \lambda(D\backslash\{p_D\})) \neq \emptyset.$$

  It is the amount objects must be grown if $g(C, D) > 1$ (i.e., $C \cap D = \emptyset$), or contracted if $g(C, D) > 1$ (i.e., $int\,C \cap int\,D \neq \emptyset$) from their internal seed points until their surfaces just touch. The **growth separation distance** $d_S(C, D)$ and the **growth penetration distance** $d_P(C, D)$ ([OnGi96]) are defined as

  $$d_S(C, D) = \max\{0, r_{CD}(g(C, D)-1)\} \text{ and } d_P(C, D) = \max\{0, r_{CD}(1-g(C, D))\},$$

  where r_{CD} is the scaling coefficient (usually, the sum of radii of circumscribing spheres for the sets $C\backslash\{p_C\}$ and $D\backslash\{p_D\}$).

  The *one-sided growth distance* between disjoint C and D (Leven–Sharir, 1987) is

  $$-1 + \min \lambda > 0 : (\{p_C\} + \lambda\{(C\backslash\{p_C\})) \cap D \neq \emptyset\}.$$

- **Minkowski difference**

  The **Minkowski difference** on the set of all compact subsets, in particular, on the set of all *sculptured objects* (or *free form objects*), of $\mathbb{R}^3$ is defined by

  $$A - B = \{x - y : x \in A, y \in B\}.$$

  If we consider object B to be free to move with fixed orientation, the Minkowski difference is a set containing all the translations that bring B to intersect with A. The closest point from the Minkowski difference boundary, $\partial(A-B)$, to the origin gives the **separation distance** between A and B.

If both objects intersect, the origin is inside of their Minkowski difference, and the obtained distance can be interpreted as a **penetration depth distance**.

- **Demyanov distance**

  Given $C \in K_p$ and $u \in S^{n-1}$, denote, if $|\{c \in C : \langle u, c\rangle = h_C(u)\}| = 1$, this unique point by $y(u, C)$ (*exposed point of C in direction u*).

  The *Demyanov difference* $A \ominus B$ of two subsets $A, B \in K_p$ is the closure of

$$conv(\cup_{T(A)\cap T(B)}\{y(u,A) - y(u,B)\}),$$

  where $T(C) = \{u \in S^{n-1} : |\{c \in C : \langle u, c\rangle = h_C(u)\}| = 1\}$.

  The **Demyanov distance** between two subsets $A, B \in K_p$ is defined by

$$||A \ominus B|| = \max_{c \in A \ominus B} ||c||_2.$$

  It is shown in [BaFa07] that $||A \ominus B|| = \sup_\alpha ||St_\alpha(A) - St_\alpha(M)||_2$, where $St_\alpha(C)$ is a *generalized Steiner point* and the supremum is over all "sufficiently smooth" probabilistic measures α.

- **Maximum polygon distance**

  The **maximum polygon distance** is a distance between two convex polygons $P = (p_1, \dots, p_n)$ and $Q = (q_1, \dots, q_m)$ defined by

$$\max_{i,j} ||p_i - q_j||_2, \;\; i \in \{1, \dots, n\}, \; j \in \{1, \dots, m\}.$$

- **Grenander distance**

  Let $P = (p_1, \dots, p_n)$ and $Q = (q_1, \dots, q_m)$ be two disjoint convex polygons, and let $L(p_i, q_j), L(p_l, q_m)$ be two intersecting *critical support lines* for P and Q. Then the **Grenander distance** between P and Q is defined by

$$||p_i - q_j||_2 + ||p_l - q_m||_2 - \Sigma(p_i, p_l) - \Sigma(g_j, q_m),$$

  where $||.||_2$ is the Euclidean norm, and $\Sigma(p_i, p_l)$ is the sum of the edges lengths of the polynomial chain $p_i, \dots, p_l$.

  Here $P = (p_1, \dots, p_n)$ is a convex polygon with the vertices in standard form, i.e., the vertices are specified according to Cartesian coordinates in a clockwise order, and no three consecutive vertices are collinear. A line L is a *line of support* of P if the interior of P lies completely to one side of L.

  Given two disjoint polygons P and Q, the line $L(p_i, q_j)$ is a *critical support line* if it is a line of support for P at p_i, a line of support for Q at q_j, and P and Q lie on opposite sides of $L(p_i, q_j)$. In general, a chord $[a, b]$ of a convex body C is called its **affine diameter** if there is a pair of different hyperplanes each containing one of the endpoints a, b and supporting C.

[redacted]

9.2 Distances on Cones

A *convex cone* C in a real vector space V is a subset C of V such that $C + C \subset C$, $\lambda C \subset C$ for any $\lambda \geq 0$. A cone C induces a *partial order* on V by

$$x \preceq y \text{ if and only if } y - x \in C.$$

The order $\preceq$ respects the vector structure of V, i.e., if $x \preceq y$ and $z \preceq u$, then $x + z \preceq y + u$, and if $x \preceq y$, then $\lambda x \preceq \lambda y$, $\lambda \in \mathbb{R}$, $\lambda \geq 0$. Elements $x, y \in V$ are called *comparable* and denoted by $x \sim y$ if there exist positive real numbers α and β such that $\alpha y \preceq x \preceq \beta y$. Comparability is an equivalence relation; its equivalence classes (which belong to C or to $-C$) are called *parts* (or *components*, *constituents*).

Given a convex cone C, a subset $S = \{x \in C : T(x) = 1\}$, where $T : V \to \mathbb{R}$ is a positive linear functional, is called a *cross-section* of C. A convex cone C is called *almost Archimedean* if the closure of its restriction to any 2D subspace is also a cone.

A convex cone C is called *pointed* if $C \cup (-C) = \{0\}$ and *solid* if $int\ C \neq \emptyset$.

- **Koszul–Vinberg metric**

  Given an open pointed convex cone C, let C^* be its dual cone.

  The **Koszul–Vinberg metric** on C (Vinberg, 1963, and Koszul, 1965) is an affine invariant Riemannian metric defined as the Hessian $g = d^2\psi_C$, where $\psi_C(x) = -\log \int_{C^*} e^{-\langle \epsilon, x\rangle} d\epsilon$ for any $x \in C$.

  The Hessian of the *entropy* (Legendre transform of $\psi_C(x)$) defines a metric on C^*, which ([Barb14]) is equivalent to the **Fisher–Rao metric** (Sect. 7.2). [Barb14] also observed that *Fisher–Souriau metric* ([Sour70]) generalises Fisher–Rao metric for Lie group thermodynamics and interpreted it as a geometric *heat capacity*.
- **Invariant distances on symmetric cones**

  An open convex cone C in an Euclidean space V is said to be *homogeneous* if its group of linear automorphisms $G = \{g \in GL(V) : g(C) = C\}$ act transitively on C. If, moreover, $\overline{C}$ is pointed and C is self-dual with respect to the given inner product $\langle , \rangle$, then it is called a *symmetric cone*. Any symmetric cone is a Cartesian product of such cones of only 5 types: the cones $Sym(n, \mathbb{R})^+$, $Her(n, \mathbb{C})^+$ (cf. Chap. 12), $Her(n, \mathbb{H})^+$ of positive-definite Hermitian matrices with real, complex or quaternion entries, the *Lorentz cone* (or *forward light cone*) $\{(t, x_1, \dots, x_n) \in \mathbb{R}^{n+1} : t^2 > x_1^2 + \cdots + x_n^2\}$ and 27-dimensional exceptional cone of 3×3 positive-definite matrices over the octonions $\mathbb{O}$. An $n \times n$ quaternion matrix A can be seen as a $2n \times 2n$ complex matrix A'; so, $A \in Her(n, \mathbb{H})^+$ means $A' \in Her(2n, \mathbb{C})^+$.

  Let V be an *Euclidean Jordan algebra*, i.e., a finite-dimensional *Jordan algebra* (commutative algebra satisfying $x^2(xy) = x(x^2y)$ and having a multiplicative identity e) equipped with an *associative* ($\langle xy, z\rangle = \langle y, xz\rangle$) inner product $\langle , \rangle$. Then the set of square elements of V is a symmetric cone, and every symmetric cone arises in this way. Denote $P(x)y = 2x(xy) - x^2y$ for any $x, y \in C$.

For example, for $C = PD_n(\mathbb{R})$, the group G is $GL(n, \mathbb{R})$, the inner product is $\langle X, Y\rangle = \mathrm{Tr}(XY)$, the Jordan product is $\frac{1}{2}(XY + YX)$, and $P(X)Y = XYX$, where the multiplication on the right-hand side is the usual matrix multiplication.

If r is the rank of V, then for any $x \in V$ there is a complete set of orthogonal primitive idempotents $c_1, \dots, c_r \neq 0$ (i.e., $c_i^2 = c_i$, c_i indecomposable, $c_ic_j = 0$ if $i \neq j$, $\sum_{i=1}^{r} c_i = e$) and real numbers $\lambda_1, \dots, \lambda_r$, called *eigenvalues* of x, such that $x = \sum_{i=1}^{r} \lambda_i c_i$. Let $x, y \in C$ and $\lambda_1, \dots, \lambda_r$ be the eigenvalues of $P(x^{-\frac{1}{2}})y$. Lim, 2001, defined following three G-invariant distances on any symmetric cone C:

$$d_R = (\sum_{1\leq i\leq r} \ln^2 \lambda_i)^{\frac{1}{2}}, \; d_F = \max_{1\leq i\leq r} \ln|\lambda_i|, \; d_H = \ln(\max_{1\leq i\leq r} \lambda_i(\min_{1\leq i\leq r} \lambda_i)^{-1}).$$

For above distances, the geometric mean $P(x^{\frac{1}{2}})(P(x^{-\frac{1}{2}}y))^{\frac{1}{2}}$ is the midpoint of x and y. The distances $d_R(x, y)$, $d_F(x, y)$ are the intrinsic metrics of G-invariant Riemannian and Finsler metrics on C. The Riemannian geodesic curve $\alpha(t) = P(x^{\frac{1}{2}})(P(x^{-\frac{1}{2}}y))^t$ is one of infinitely many shortest Finsler curves passing through x and y. The space $(C, d_R(x, y))$ is an **Hadamard space** (Chap. 6), while $(C, d_F(x, y))$ is not. The distance $d_F(x, y)$ is the **Thompson's part metric** on C, and the distance $d_H(x, y)$ is the **Hilbert projective semimetric** on C which is a complete metric on the unit sphere on C.

- **Thompson's part metric**

  Given a convex cone C in a real Banach space V, the **Thompson's part metric** on a *part* $K \subset C\backslash\{0\}$ is defined (Thompson, 1963) by

$$\log\max\{m(x, y), m(y, x)\}$$

  for any $x, y \in K$, where $m(x, y) = \inf\{\lambda \in \mathbb{R} : y \preceq \lambda x\}$.

  If C is *almost Archimedean*, then K equipped with this metric is a **complete** metric space. If C is finite-dimensional, then one obtains a **chord space** (Chap. 6). The *positive cone* $\mathbb{R}^n_+ = \{(x_1, \dots, x_n) : x_i \geq 0 \text{ for } 1 \leq i \leq n\}$ equipped with this metric is isometric to a *normed space* which can be seen as being flat. The same holds for the **Hilbert projective semimetric** on $\mathbb{R}^n_+$.

  If C is a closed solid cone in $\mathbb{R}^n$, then *int* C can be seen as an n-dimensional manifold M^n. If for any tangent vector $v \in T_p(M^n)$, $p \in M^n$, we define a norm $||v||_p^T = \inf\{\alpha > 0 : -\alpha p \preceq v \preceq \alpha p\}$, then the length of any piecewise differentiable curve $\gamma : [0, 1] \to M^n$ is $l(\gamma) = \int_0^1 ||\gamma'(t)||^T_{\gamma(t)}dt$, and the distance between x and y is $\inf_\gamma l(\gamma)$, where the infimum is taken over all such curves γ with $\gamma(0) = x$, $\gamma(1) = y$.

- **Hilbert projective semimetric**

  Given a pointed closed convex cone C in a real Banach space V, the **Hilbert projective semimetric** on $C\backslash\{0\}$ is defined (Bushell, 1973), for $x, y \in C\backslash\{0\}$, by

$$h(x, y) = \log(m(x, y)m(y, x)),$$

[redacted]

where $m(x, y) = \inf\{\lambda \in \mathbb{R} : y \preceq \lambda x\}$; it holds $\frac{1}{m(y,x)} = \sup\{\lambda \in \mathbb{R} : \lambda y \preceq x\}$. This semimetric is finite on the interior of C and $h(\lambda x, \lambda' y) = h(x, y)$ for $\lambda, \lambda' > 0$. So, $h(x, y)$ is a metric on the *projectivization* of C, i.e., the space of rays of this cone.

If C is finite-dimensional, and S is a *cross-section* of C (in particular, $S = \{x \in C : ||x|| = 1\}$, where $||.||$ is a norm on V), then, for any distinct points $x, y \in S$, it holds $h(x, y) = |\ln(x, y, z, t)|$, where z, t are the points of the intersection of the line $l_{x,y}$ with the boundary of S, and (x, y, z, t) is the **cross-ratio** of x, y, z, t. Cf. the **Hilbert projective metric** in Chap. 6.

If C is finite-dimensional and *almost Archimedean*, then each part of C is a **chord space** (Chap. 6) under the Hilbert projective semimetric. On the *Lorentz cone* $L = \{x = (t, x_1, \dots, x_n) \in \mathbb{R}^{n+1} : t^2 > x_1^2 + \dots + x_n^2\}$, this semimetric is isometric to the n-dimensional *hyperbolic space*. On the hyperbolic subspace $H = \{x \in L : \det(x) = 1\}$, it holds $h(x, y) = 2d(x, y)$, where $d(x, y)$ is the **Thompson's part metric** which is (on H) the usual hyperbolic distance $\operatorname{arccosh}\langle x, y\rangle$.

If C is a closed solid cone in $\mathbb{R}^n$, then *int* C can be seen as an n**-manifold** M^n (Chap. 2). If for any tangent vector $v \in T_p(M^n)$, $p \in M^n$, we define a seminorm $||v||_p^H = m(p, v) - m(v, p)$, then the length of any piecewise differentiable curve $\gamma : [0, 1] \to M^n$ is $l(\gamma) = \int_0^1 ||\gamma'(t)||_{\gamma(t)}^H dt$, and $h(x, y) = \inf_\gamma l(\gamma)$, where the infimum is taken over all such curves γ with $\gamma(0) = x$ and $\gamma(1) = y$.

- **Bushell metric**

  Given a convex cone C in a real Banach space V, the **Bushell metric** on the set $S = \{x \in C : \sum_{i=1}^n |x_i| = 1\}$ (in general, on any *cross-section* of C) is defined by

$$\frac{1 - m(x, y) \cdot m(y, x)}{1 + m(x, y) \cdot m(y, x)}$$

  for any $x, y \in S$, where $m(x, y) = \inf\{\lambda \in \mathbb{R} : y \preceq \lambda x\}$. In fact, it is equal to $\tanh(\frac{1}{2} h(x, y))$, where h is the **Hilbert projective semimetric**.

- ***k*-oriented distance**

  A *simplicial cone* C in $\mathbb{R}^n$ is defined as the intersection of n (open or closed) half-spaces, each of whose supporting planes contain the origin 0. For any set M of n points on the *unit sphere*, there is a unique simplicial cone C that contains these points. The *axes* of the cone C can be constructed as the set of the n rays, where each ray originates at the origin, and contains one of the points from M.

  Given a *partition* $\{C_1, \dots, C_k\}$ of $\mathbb{R}^n$ into a set of simplicial cones $C_1, \dots, C_k$, the k**-oriented distance** is a metric on $\mathbb{R}^n$ defined by

$$d_k(x - y)$$

  for all $x, y \in \mathbb{R}^n$, where, for any $x \in C_i$, the value $d_k(x)$ is the length of the shortest path from the origin 0 to x traveling only in directions parallel to the axes of C_i.

- **Cones over metric space**

  A **cone over a metric space** (X, d) is the quotient space $Con(X, d)$=$(X \times [0, 1])/(X \times \{0\})$ obtained from the product $X \times \mathbb{R}_{\geq 0}$ by collapsing the *fiber* (subspace $X \times \{0\}$) to a point (the apex of the cone). Cf. **metric cone** in Chap. 1.

  The *Euclidean cone over the metric space* (X, d) is the cone $Con(X, d)$ with a metric d defined, for any $(x, t), (y, s) \in Con(X, d)$, by

$$\sqrt{t^2 + s^2 - 2ts\cos(\min\{d(x, y), \pi\})}.$$

  If (X, d) is a compact metric space with diameter < 2, the **Krakus metric** is a metric on $Con(X, d)$ defined, for any $(x, t), (y, s) \in Con(X, d)$, by

$$\min\{s, t\}d(x, y) + |t - s|.$$

  The cone $Con(X, d)$ with the Krakus metric admits a unique *midpoint* for each pair of its points if (X, d) has this property.

  If M^n is a manifold with a pseudo-Riemannian metric g, one can consider a metric $dr^2 + r^2g$ (in general, a metric $\frac{1}{k}dr^2 + r^2g, k \neq 0$) on $Con(M^n) = M^n \times \mathbb{R}_{>0}$. For example, $Con(M^n) = \mathbb{R}^n \setminus \{0\}$ if (M^n, g) is the unit sphere in $\mathbb{R}^n$.

  A *spherical cone* (or *suspension*) $\Sigma(X)$ over a metric space (X, d) is the quotient of the product $X \times [0, a]$ obtained by identifying all points in the fibers $X \times \{0\}$ and $X \times \{a\}$. If (X, d) is a **length space** (Chap. 6) with $diam(X) \leq \pi$, and $a = \pi$, the **suspension metric** on $\Sigma(X)$ is defined, for any $(x, t), (y, s) \in \Sigma(X)$, by

$$\arccos(\cos t \cos s + \sin t \sin s \cos d(x, y)).$$

9.3 Distances on Simplicial Complexes

An r-dimensional *simplex* (or *geometrical simplex*, *hypertetrahedron*) is the *convex hull* of $r + 1$ points of $\mathbb{E}^n$ which do not lie in any $(r - 1)$-plane. The boundary of an r-simplex has $r + 1$ 0-*faces* (polytope vertices), $\frac{r(r+1)}{2}$ 1-*faces* (polytope edges), and $\binom{r+1}{i+1}$ *i-faces*, where $\binom{r}{i}$ is the binomial coefficient. The *content* (i.e., the *hypervolume*) of a simplex can be computed using the *Cayley–Menger determinant*. The regular simplex of dimension r is denoted by α_r. *Simplicial depth of a point* $p \in \mathbb{E}^n$ *relative to a set* $P \subset \mathbb{E}^n$ is the number of simplices S, generated by $(n + 1)$-subsets of P and containing p.

Roughly, a *geometrical simplicial complex* is a space with a *triangulation*, i.e., a decomposition of it into closed simplices such that any two simplices either do not intersect or intersect only along a common face.

An *abstract simplicial complex* S is a set, whose elements are called *vertices*, in which a family of finite nonempty subsets, called *simplices*, is distinguished, such

[redacted]

that every nonempty subset of a simplex s is a simplex, called a *face* of s, and every one-element subset is a simplex. A simplex is called i-dimensional if it consists of $i + 1$ vertices. The *dimension* of S is the maximal dimension of its simplices. For every simplicial complex S there exists a triangulation of a polyhedron whose simplicial complex is S. This geometric simplicial complex, denoted by GS, is called the *geometric realization* of S.

- **Vietoris–Rips complex**

  Given a metric space (X, d) and distance δ, their **Vietoris–Rips complex** is an abstract simplicial complex, the simplexes of which are the finite subsets M of (X, d) having diameter at most δ; the dimension of a simplex defined by M is $|M| - 1$.
- **Simplicial metric**

  Given an abstract simplicial complex S, the points of geometric simplicial complex GS, realizing S, can be identified with the functions $\alpha : S \to [0, 1]$ for which the set $\{x \in S : \alpha(x) \neq 0\}$ is a simplex in S, and $\sum_{x \in S} \alpha(x) = 1$. The number $\alpha(x)$ is called the *x-th barycentric coordinate* of α.

  The **simplicial metric** on GS (Lefschetz, 1939) is the Euclidean metric on it:

  $$\sqrt{\sum_{x \in S} (\alpha(x) - \beta(x))^2}.$$

  Tukey, 1939, found another metric on GS, topologically equivalent to a simplicial one. His **polyhedral metric** is the **intrinsic metric**, defined as the infimum of the lengths of the polygonal lines joining the points α and β such that each link is within one of the simplices. An example of a polyhedral metric is the intrinsic metric on the surface of a convex polyhedron in $\mathbb{E}^3$.
- **Polyhedral space**

  A Euclidean **polyhedral space** is a simplicial complex with a **polyhedral metric**. Every simplex is a **flat space** (a metric space locally isometric to some $\mathbb{E}^n$; cf. Chap. 1), and the metrics of any two simplices coincide on their intersection. The metric is the maximal metric not exceeding the metrics of simplices.

  If such a space is an n-**manifold** (Chap. 2), a point in it is a **metric singularity** if it has no neighborhood isometric to an open subset of $\mathbb{E}^n$.

  A polyhedral metric on a simplicial complex in a space of constant (positive or negative) curvature results in *spherical* and *hyperbolic polyhedral spaces*.

  The *dimension* of a polyhedral space is the maximal dimension of simplices used to glue it. **Metric graphs** (Chap. 15) are just one-dimensional polyhedral spaces.

  The surface of a convex polyhedron is a 2D polyhedral space. A polyhedral metric d on a triangulated surface is a **circle-packing metric** (Thurston, 1985) if there exists a vertex-weighting $w(x) > 0$ with $d(x, y) = w(x) + w(y)$ for any edge xy.

- **Manifold edge-distance**

  A (boundaryless) *combinatorial n-manifold* is an abstract n-dimensional simplicial complex M^n in which the *link* of each r-simplex is an $(n-r-1)$-sphere. The category of such spaces is equivalent to the category of piecewise-linear (PL) manifolds.

  The *link* of a simplex S is $Cl(Star_S) - Star_S$, where $Star_S$ is the set of all simplices in M^n having a face S, and $Cl(Star_S)$ is the smallest simplicial subcomplex of M^n containing $Star_S$.

  The **edge-distance** between vertices $u, v \in M^n$ is the minimum number of edges needed to connect them.

- **Manifold triangulation metric**

  Let M^n be a compact PL (piecewise-linear) n-dimensional manifold. A *triangulation* of M^n is a simplicial complex such that its corresponding polyhedron is PL-homeomorphic to M^n. Let T_{M^n} be the set of all *combinatorial types* of triangulations, where two triangulations are equivalent if they are simplicially isomorphic.

  Every such triangulation can be seen as a metric on the smooth manifold M if one assigns the unit length for any of its 1-dimensional simplices; so, T_{M^n} can be seen as a discrete analog of the space of Riemannian structures, i.e., isometry classes of Riemannian metrics on M^n.

  A **manifold triangulation metric** between two triangulations x and y is (Nabutovsky and Ben-Av, 1993) an **editing metric** on T_{M^n}, i.e., the minimal number of elementary moves, from a given finite list of operations, needed to obtain y from x.

  For example, the *bistellar move* consists of replacing a subcomplex of a given triangulation, which is simplicially isomorphic to a subcomplex of the boundary of the standard $(n+1)$-simplex, by the complementary subcomplex of the boundary of an $(n+1)$-simplex, containing all remaining n-simplices and their faces. Every triangulation can be obtained from any other one by a finite sequence of bistellar moves.

- **Polyhedral chain metric**

  An r-dimensional *polyhedral chain A* in $\mathbb{E}^n$ is a linear expression $\sum_{i=1}^{m} d_i t_i^r$, where, for any i, the value t_i^r is an r-dimensional simplex of $\mathbb{E}^n$. The *boundary* ∂A of a chain AD is the linear combination of boundaries of the simplices in the chain. The boundary of an r-dimensional chain is an $(r-1)$-dimensional chain.

  A **polyhedral chain metric** is a **norm metric** $||A - B||$ on the set $C_r(\mathbb{E}^n)$ of all r-dimensional polyhedral chains. As a norm $||.||$ on $C_r(\mathbb{E}^n)$ one can take:

  1. The *mass* of a polyhedral chain, i.e., $|A| = \sum_{i=1}^{m} |d_i||t_i^r|$, where $|t^r|$ is the volume of the cell t_i^r;
  2. The *flat norm* of a polyhedral chain, i.e., $|A|^\flat = \inf_D\{|A - \partial D| + |D|\}$, where the infimum is taken over all $(r+1)$-dimensional polyhedral chains;

[redacted]

3. The *sharp norm* of a polyhedral chain, i.e.,

$$|A|^{\sharp} = \inf\left(\frac{\sum_{i=1}^{m} |d_i||t_i^r||v_i|}{r+1} + |\sum_{i=1}^{m} d_i T_{v_i} t_i^r|^{\flat}\right),$$

   where the infimum is taken over all *shifts* v (here $T_v t^r$ is the cell obtained by shifting t^r by a vector v of length $|v|$). A flat chain of finite mass is a sharp chain. If $r = 0$, than $|A|^{\flat} = |A|^{\sharp}$.

The metric space of *polyhedral co-chains* (i.e., linear functions of polyhedral chains) can be defined similarly. As a norm of a polyhedral co-chain X one can take:

1. The *co-mass* of a polyhedral co-chain, i.e., $|X| = \sup_{|A|=1} |X(A)|$, where $X(A)$ is the value of the co-chain X on a chain A;
2. The *flat co-norm* of a polyhedral co-chain, i.e., $|X|^{\flat} = \sup_{|A|^{\flat}=1} |X(A)|$;
3. The *sharp co-norm* of a polyhedral co-chain, i.e., $|X|^{\sharp} = \sup_{|A|^{\sharp}=1} |X(A)|$.

Part III
Distances in Classical Mathematics

[redacted]

Chapter 10
Distances in Algebra

10.1 Group Metrics

A *group* $(G, \cdot, e)$ is a set G of elements with a binary operation $\cdot$, called the *group operation*, that together satisfy the four fundamental properties of *closure* ($x \cdot y \in G$ for any $x, y \in G$), *associativity* ($x \cdot (y \cdot z) = (x \cdot y) \cdot z$ for any $x, y, z \in G$), the *identity property* ($x \cdot e = e \cdot x = x$ for any $x \in G$), and the *inverse property* (for any $x \in G$, there exists an element $x^{-1} \in G$ such that $x \cdot x^{-1} = x^{-1} \cdot x = e$).

In additive notation, a group $(G, +, 0)$ is a set G with a binary operation $+$ such that the following properties hold: $x+y \in G$ for any $x, y \in G$, $x+(y+z) = (x+y)+z$ for any $x, y, z \in G$, $x + 0 = 0 + x = x$ for any $x \in G$, and, for any $x \in G$, there exists an element $-x \in G$ such that $x + (-x) = (-x) + x = 0$.

A group $(G, \cdot, e)$ is called *finite* if the set G is finite. A group $(G, \cdot, e)$ is called *Abelian* if it is *commutative*, i.e., $x \cdot y = y \cdot x$ for any $x, y \in G$.

Most metrics considered in this section are **group norm metrics** on a group $(G, \cdot, e)$, defined by

$$||x \cdot y^{-1}||$$

(or, sometimes, by $||y^{-1} \cdot x||$), where $||.||$ is a *group norm*, i.e., a function $||.|| : G \to \mathbb{R}$ such that, for any $x, y \in G$, we have the following properties:

1. $||x|| \geq 0$, with $||x|| = 0$ if and only if $x = e$;
2. $||x|| = ||x^{-1}||$;
3. $||x \cdot y|| \leq ||x|| + ||y||$ (*triangle inequality*).

In additive notation, a group norm metric on a group $(G, +, 0)$ is defined by $||x + (-y)|| = ||x - y||$, or, sometimes, by $||(-y) + x||$.

The simplest example of a group norm metric is the **bi-invariant ultrametric** (sometimes called the *Hamming metric*) $||x \cdot y^{-1}||_H$, where $||x||_H = 1$ for $x \neq e$, and $||e||_H = 0$.

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_10

- **Bi-invariant metric**

  A metric (in general, a semimetric) d on a group $(G, \cdot, e)$ is called **bi-invariant** if

$$d(x, y) = d(x \cdot z, y \cdot z) = d(z \cdot x, z \cdot y)$$

  for any $x, y, z \in G$ (cf. **translation invariant metric** in Chap. 5). Any **group norm metric** on an Abelian group is bi-invariant.

  A metric (in general, a semimetric) d on a group $(G, \cdot, e)$ is called a **right-invariant metric** if $d(x, y) = d(x \cdot z, y \cdot z)$ for any $x, y, z \in G$, i.e., the operation of right multiplication by an element z is a **motion** of the metric space (G, d). Any group norm metric defined by $||x \cdot y^{-1}||$, is right-invariant.

  A metric (in general, a semimetric) d on a group $(G, \cdot, e)$ is called a **left-invariant metric** if $d(x, y) = d(z \cdot x, z \cdot y)$ holds for any $x, y, z \in G$, i.e., the operation of left multiplication by an element z is a motion of the metric space (G, d). Any group norm metric defined by $||y^{-1} \cdot x||$, is left-invariant.

  Any right-invariant or left-invariant (in particular, bi-invariant) metric d on G is a group norm metric, since one can define a group norm on G by $||x|| = d(x, 0)$.
- ***G*-invariant metric**

  Given a metric space (X, d) and an action $g(x)$ of a group G on it, the metric d is called ***G*-invariant** (under this action) if for all $x, y \in X, g \in G$ it holds

$$d(g(x), g(y)) = d(x, y).$$

  For every G-invariant metric d_X on X and every point $x \in X$, the function

$$d_G(g_1, g_2) = d_X(g_1(x), g_2(x))$$

  is a **left-invariant metric** on G. This metric is called **orbit metric** in [BBI01], since it is the restriction of d on the orbit Gx, which can be identified with G.
- **Positively homogeneous distance**

  A distance d on an Abelian group $(G, +, 0)$ is called **positively homogeneous** if

$$d(mx, my) = md(x, y)$$

  for all $x, y \in G$ and all $m \in \mathbb{N}$, where mx is the sum of m terms all equal to x.
- **Translation discrete metric**

  A **group norm metric** (in general, a group norm semimetric) on a group $(G, \cdot, e)$ is called **translation discrete** if the *translation distances* (or *translation numbers*)

$$\tau_G(x) = \lim_{n \to \infty} \frac{||x^n||}{n}$$

[redacted]

of the *nontorsion elements* x (i.e., such that $x^n \neq e$ for any $n \in \mathbb{N}$) of the group with respect to that metric are bounded away from zero.

If the numbers $\tau_G(x)$ are just nonzero, such a group norm metric is called a **translation proper metric**.

- **Word metric**

  Let $(G, \cdot, e)$ be a finitely-generated group with a set A of generators (i.e., A is finite, and every element of G can be expressed as a product of finitely many elements A and their inverses). The *word length* $w_W^A(x)$ of an element $x \in G \backslash \{e\}$ is defined by

  $$w_W^A(x) = \inf\{r : x = a_1^{\epsilon_1} \dots a_r^{\epsilon_r}, a_i \in A, \epsilon_i \in \{\pm 1\}\} \text{ and } w_W^A(e) = 0.$$

  The **word metric** d_W^A *associated with* A is a **group norm metric** on G defined by

  $$w_W^A(x \cdot y^{-1}).$$

  As the word length w_W^A is a *group norm* on G, d_W^A is **right-invariant**. Sometimes it is defined as $w_W^A(y^{-1} \cdot x)$, and then it is **left-invariant**. In fact, d_W^A is the maximal metric on G that is right-invariant, and such that the distance from any element of A or A^{-1} to the identity element e is equal to one.

  If A and B are two finite sets of generators of the group $(G, \cdot, e)$, then the identity mapping between the metric spaces (G, d_W^A) and (G, d_W^B) is a **quasi-isometry**, i.e., the word metric is unique up to quasi-isometry.

  The word metric is the **path metric** of the *Cayley graph* Γ of $(G, \cdot, e)$, constructed with respect to A. Namely, Γ is a graph with the vertex-set G in which two vertices x and $y \in G$ are connected by an edge if and only if $y = a^\epsilon x$, $\epsilon = \pm 1, a \in A$.

- **Weighted word metric**

  Let $(G, \cdot, e)$ be a finitely-generated group with a set A of generators. Given a bounded *weight function* $w : A \to (0, \infty)$, the *weighted word length* $w_{WW}^A(x)$ of an element $x \in G \backslash \{e\}$ is defined by $w_{WW}^A(e) = 0$ and

  $$w_{WW}^A(x) = \inf\left\{\sum_{i=1}^{t} w(a_i), t \in \mathbb{N} : x = a_1^{\epsilon_1} \dots a_t^{\epsilon_t}, a_i \in A, \epsilon_i \in \{\pm 1\}\right\}.$$

  The **weighted word metric** d_{WW}^A *associated with* A is a **group norm metric** on G defined by

  $$w_{WW}^A(x \cdot y^{-1}).$$

  As the weighted word length w_{WW}^A is a *group norm* on G, d_{WW}^A is **right-invariant**. Sometimes it is defined as $w_{WW}^A(y^{-1} \cdot x)$, and then it is **left-invariant**.

The metric d^A_{WW} is the supremum of semimetrics d on G with the property that $d(e,a) \leq w(a)$ for any $a \in A$.

The metric d^A_{WW} is a **coarse-path metric**, and every right-invariant coarse path metric is a weighted word metric up to **coarse isometry**.

The metric d^A_{WW} is the **path metric** of the *weighted Cayley graph* Γ_W of $(G,\cdot,e)$ constructed with respect to A. Namely, Γ_W is a weighted graph with the vertex-set G in which two vertices x and $y \in G$ are connected by an edge with the weight $w(a)$ if and only if $y = a^{\epsilon}x$, $\epsilon = \pm 1$, $a \in A$.

- **Interval norm metric**

  An **interval norm metric** is a **group norm metric** on a finite group $(G,\cdot,e)$ defined by

$$||x \cdot y^{-1}||_{int},$$

where $||.||_{int}$ is an *interval norm* on G, i.e., a *group norm* such that the values of $||.||_{int}$ form a set of consecutive integers starting with 0.

To each interval norm $||.||_{int}$ corresponds an ordered *partition* $\{B_0,\dots,B_m\}$ of G with $B_i = \{x \in G : ||x||_{int} = i\}$; cf. **Sharma–Kaushik distance** in Chap. 16. The *Hamming* and *Lee* norms are special cases of interval norm. A *generalized Lee norm* is an interval norm for which each class has a form $B_i = \{a, a^{-1}\}$.

- ***C*-metric**

  A ***C*-metric** d is a metric on a group $(G,\cdot,e)$ satisfying the following conditions:

1. The values of d form a set of consecutive integers starting with 0;
2. The cardinality of the sphere $B(x,r) = \{y \in G : d(x,y) = r\}$ is independent of the particular choice of $x \in G$.

The **word metric**, the **Hamming metric**, and the **Lee metric** are C-metrics. Any **interval norm metric** is a C-metric.

- **Order norm metric**

  Let $(G,\cdot,e)$ be a finite Abelian group. Let $ord(x)$ be the *order* of an element $x \in G$, i.e., the smallest positive integer n such that $x^n = e$. Then the function $||.||_{ord} : G \to \mathbb{R}$ defined by $||x||_{ord} = \ln ord(x)$, is a *group norm* on G, called the *order norm*.

  The **order norm metric** is a **group norm metric** on G, defined by

$$||x \cdot y^{-1}||_{ord}.$$

- **Tărnăuceanu metric**

  Let $o(a)$ denote the order of the element a of a group. Let C be the class of finite groups G in which $o(ab) < o(a) + o(b)$ for every $a, b \in G$. Tărnăuceanu,

[redacted]

2015, noted that the function $d : G \times G \to \mathbb{N}$ defined by

$$d(x, y) = o(xy^{-1}) - 1$$

for all $x, y \in G$ is a metric on G if and only if $G \in C$.

He found that C contains all Abelian p-groups, Q_8, and A_4, but not nonabelian finite simple groups, alternating groups $A(n)$ with $n \geq 5$, and, for $n \geq 4$, $Sym(n)$, quaternion groups Q_{2^n}, dihedral groups D_{2n}. C is closed under subgroups, but not under direct products or extensions. The centralizers of nontrivial elements of such groups contain only elements of prime power order.

- **Monomorphism norm metric**

  Let $(G, +, 0)$ be a group. Let $(H, \cdot, e)$ be a group with a *group norm* $||.||_H$. Let $f : G \to H$ be a *monomorphism* of groups G and H, i.e., an injective function such that $f(x+y) = f(x) \cdot f(y)$ for any $x, y \in G$. Then the function $||.||^f_G : G \to \mathbb{R}$ defined by $||x||^f_G = ||f(x)||_H$, is a *group norm* on G, called the *monomorphism norm.*

  The **monomorphism norm metric** is a **group norm metric** on G defined by

$$||x - y||^f_G.$$

- **Product norm metric**

  Let $(G, +, 0)$ be a group with a *group norm* $||.||_G$. Let $(H, \cdot, e)$ be a group with a group norm $||.||_H$. Let $G \times H = \{\alpha = (x, y) : x \in G, y \in H\}$ be the Cartesian product of G and H, and $(x, y) \cdot (z, t) = (x + z, y \cdot t)$.

  Then the function $||.||_{G\times H} : G\times H \to \mathbb{R}$ defined by $||\alpha||_{G\times H} = ||(x, y)||_{G\times H} = ||x||_G + ||y||_H$, is a group norm on $G \times H$, called the *product norm.*

  The **product norm metric** is a **group norm metric** on $G \times H$ defined by

$$||\alpha \cdot \beta^{-1}||_{G\times F}.$$

  On the Cartesian product $G \times H$ of two finite groups with the *interval norms* $||.||^{int}_G$ and $||.||^{int}_H$, an interval norm $||.||^{int}_{G\times H}$ can be defined. In fact, $||\alpha||^{int}_{G\times H} = ||(x, y)||^{int}_{G\times H} = ||x||_G + (m + 1)||y||_H$, where $m = \max_{a\in G} ||a||^{int}_G$.

- **Quotient norm metric**

  Let $(G, \cdot, e)$ be a group with a *group norm* $||.||_G$. Let $(N, \cdot, e)$ be a *normal subgroup* of $(G, \cdot, e)$, i.e., $xN = Nx$ for any $x \in G$. Let $(G/N, \cdot, eN)$ be the *quotient group* of G, i.e., $G/N = \{xN : x \in G\}$ with $xN = \{x \cdot a : a \in N\}$, and $xN \cdot yN = xyN$. Then the function $||.||_{G/N} : G/N \to \mathbb{R}$ defined by $||xN||_{G/N} = \min_{a\in N} ||xa||_X$, is a group norm on G/N, called the *quotient norm.*

  A **quotient norm metric** is a **group norm metric** on G/N defined by

$$||xN \cdot (yN)^{-1}||_{G/N} = ||xy^{-1}N||_{G/N}.$$

If $G = \mathbb{Z}$ with the norm being the absolute value, and $N = m\mathbb{Z}$, $m \in \mathbb{N}$, then the quotient norm on $\mathbb{Z}/m\mathbb{Z} = \mathbb{Z}_m$ coincides with the *Lee norm*.

If a metric d on a group $(G, \cdot, e)$ is **right-invariant**, then for any normal subgroup $(N, \cdot, e)$ of $(G, \cdot, e)$ the metric d induces a right-invariant metric (in fact, the **Hausdorff metric**) d^* on G/N by

$$d^*(xN, yN) = \max\{\max_{b\in yN}\min_{a\in xN} d(a, b), \max_{a\in xN}\min_{b\in yN} d(a, b)\}.$$

- **Commutation distance**

  Let $(G, \cdot, e)$ be a finite nonabelian group. Let $Z(G) = \{c \in G : x \cdot c = c \cdot x \text{ for any } x \in G\}$ be the *center* of G.

  The *commutation graph* of G is defined as a graph with the vertex-set G in which distinct elements $x, y \in G$ are connected by an edge whenever they *commute*, i.e., $x \cdot y = y \cdot x$. (Darafsheh, 2009, consider noncommuting graph on $G \setminus Z(G)$.)

  Any two noncommuting elements $x, y \in G$ are connected in this graph by the path x, c, y, where c is any element of $Z(G)$ (for example, e). A path $x = x^1, x^2, \dots, x^k = y$ in the commutation graph is called an $(x - y)$ N-path if $x^i \notin Z(G)$ for any $i \in \{1, \dots, k\}$. In this case the elements $x, y \in G\backslash Z(G)$ are called *N-connected*.

  The **commutation distance** (see [DeHu98]) d is an extended distance on G defined by the following conditions:

  1. $d(x, x) = 0$;
  2. $d(x, y) = 1$ if $x \neq y$, and $x \cdot y = y \cdot x$;
  3. $d(x, y)$ is the minimum length of an $(x - y)$ N-path for any N-connected elements x and $y \in G\backslash Z(G)$;
  4. $d(x, y) = \infty$ if $x, y \in G\backslash Z(G)$ are not connected by any N-path.

  Given a group G and a G-conjugacy class X in it, Bates–Bundy–Perkins–Rowley in 2003, 2004, 2007, 2008 considered *commuting graph* (X, E) whose vertex set is X and distinct vertices $x, y \in X$ are joined by an edge $e \in E$ whenever they commute.

- **Modular distance**

  Let $(\mathbb{Z}_m, +, 0)$, $m \geq 2$, be a finite *cyclic group*. Let $r \in \mathbb{N}$, $r \geq 2$. The *modular r-weight* $w_r(x)$ of an element $x \in \mathbb{Z}_m = \{0, 1, \dots, m\}$ is defined as $w_r(x) = \min\{w_r(x), w_r(m - x)\}$, where $w_r(x)$ is the *arithmetic r-weight* of the integer x.

  The value $w_r(x)$ can be obtained as the number of nonzero coefficients in the *generalized nonadjacent form* $x = e_n r^n + \dots e_1 r + e_0$ with $e_i \in \mathbb{Z}$, $|e_i| < r$, $|e_i + e_{i+1}| < r$, and $|e_i| < |e_{i+1}|$ if $e_i e_{i+1} < 0$. Cf. **arithmetic *r*-norm metric** in Chap. 12.

  The **modular distance** is a distance on $\mathbb{Z}_m$, defined by

  $$w_r(x - y).$$

[redacted]

The modular distance is a metric for $w_r(m) = 1$, $w_r(m) = 2$, and for several special cases with $w_r(m) = 3$ or 4. In particular, it is a metric for $m = r^n$ or $m = r^n - 1$; if $r = 2$, it is a metric also for $m = 2^n + 1$ (see, for example, [Ernv85]).

The most popular metric on $\mathbb{Z}_m$ is the **Lee metric** defined by $||x-y||_{Lee}$, where $||x||_{Lee} = \min\{x, m-x\}$ is the *Lee norm* of an element $x \in \mathbb{Z}_m$.

- ***G*-norm metric**

  Consider a finite field $\mathbb{F}_{p^n}$ for a prime p and a natural number n. Given a compact convex centrally-symmetric body G in $\mathbb{R}^n$, define the *G-norm* of an element $x \in \mathbb{F}_{p^n}$ by $||x||_G = \inf\{\mu \geq 0 : x \in p\mathbb{Z}^n + \mu G\}$.

  The ***G*-norm metric** is a **group norm metric** on $\mathbb{F}_{p^n}$ defined by

$$||x \cdot y^{-1}||_G.$$

- **Permutation norm metric**

  Given a finite metric space (X, d), the **permutation norm metric** is a **group norm metric** on the group $(Sym_X, \cdot, id)$ of all permutations of X (*id* is the *identity mapping*) defined by

$$||f \cdot g^{-1}||_{Sym},$$

  where the *group norm* $||.||_{Sym}$ on Sym_X is given by $||f||_{Sym} = \max_{x \in X} d(x, f(x))$.

- **Metric of motions**

  Let (X, d) be a metric space, and let $p \in X$ be a fixed element of X.

  The **metric of motions** (see [Buse55]) is a metric on the group $(\Omega, \cdot, id)$ of all **motions** of (X, d) (*id* is the *identity mapping*) defined by

$$\sup_{x \in X} d(f(x), g(x)) \cdot e^{-d(p,x)}$$

  for any $f, g \in \Omega$ (cf. **Busemann metric of sets** in Chap. 3). If the space (X, d) is bounded, a similar metric on Ω can be defined as

$$\sup_{x \in X} d(f(x), g(x)).$$

  Given a semimetric space (X, d), the **semimetric of motions** on $(\Omega, \cdot, id)$ is

$$d(f(p), g(p)).$$

- **General linear group semimetric**

  Let $\mathbb{F}$ be a locally compact nondiscrete *topological field*. Let $(\mathbb{F}^n, ||.||_{\mathbb{F}^n})$, $n \geq 2$, be a *normed vector space* over $\mathbb{F}$. Let $||.||$ be the *operator norm* associated with the normed vector space $(\mathbb{F}^n, ||.||_{\mathbb{F}^n})$. Let $GL(n, \mathbb{F})$ be the *general linear group* over $\mathbb{F}$. Then the function $|.|_{op} : GL(n, \mathbb{F}) \to \mathbb{R}$ defined by $|g|_{op} = \sup\{|\ln ||g|| |, |\ln ||g^{-1}|| |\}$, is a seminorm on $GL(n, \mathbb{F})$.

The **general linear group semimetric** on the group $GL(n, \mathbb{F})$ is defined by

$$|g \cdot h^{-1}|_{op}.$$

It is a **right-invariant** semimetric which is unique, up to **coarse isometry**, since any two norms on $\mathbb{F}^n$ are **bi-Lipschitz equivalent**.

- **Generalized torus semimetric**

  Let $(T, \cdot, e)$ be a *generalized torus*, i.e., a *topological group* which is isomorphic to a direct product of n multiplicative groups $\mathbb{F}_i^*$ of locally compact nondiscrete *topological fields* $\mathbb{F}_i$. Then there is a proper continuous homomorphism $v : T \to \mathbb{R}^n$, namely, $v(x_1, \dots, x_n) = (v_1(x_1), \dots, v_n(x_n))$, where $v_i : \mathbb{F}_i^* \to \mathbb{R}$ are proper continuous homomorphisms from the $\mathbb{F}_i^*$ to the additive group $\mathbb{R}$, given by the logarithm of the *valuation*. Every other proper continuous homomorphism $v' : T \to \mathbb{R}^n$ is of the form $v' = \alpha \cdot v$ with $\alpha \in GL(n, \mathbb{R})$. If $||.||$ is a norm on $\mathbb{R}^n$, one obtains the corresponding seminorm $||x||_T = ||v(x)||$ on T.

  The **generalized torus semimetric** is defined on the group $(T, \cdot, e)$ by

$$||xy^{-1}||_T = ||v(xy^{-1})|| = ||v(x) - v(y)||.$$

- **Stable norm metric**

  Given a Riemannian manifold (M, g), the **stable norm metric** is a **group norm metric** on its *real homology group* $H_k(M, \mathbb{R})$ defined by the following *stable norm* $||h||_s$: the infimum of the Riemannian k-volumes of real cycles representing h.

  The Riemannian manifold $(\mathbb{R}^n, g)$ is within finite **Gromov–Hausdorff distance** (cf. Chap. 1) from an n-dimensional normed vector space $(\mathbb{R}^n, ||.||_s)$.

  If (M, g) is a compact connected oriented Riemannian manifold, then the manifold $H_1(M, \mathbb{R})/H_1(M, \mathbb{R})$ with metric induced by $||.||_s$ is called the *Albanese torus* (or *Jacobi torus*) of (M, g). This **Albanese metric** is a **flat metric** (Chap. 8).
- **Heisenberg metric**

  Let $(H, \cdot, e)$ be the (real) *Heisenberg group* $\mathcal{H}^n$, i.e., a group on the set $H = \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ with the group law $h \cdot h' = (x, y, t) \cdot (x', y', t') = (x + x', y + y', t + t' + 2\sum_{i=1}^{n}(x_i' y_i - x_i y_i'))$, and the identity $e = (0, 0, 0)$. Let $|.|_{Heis}$ be the *Heisenberg gauge* (Cygan, 1978) on $\mathcal{H}^n$ defined by $|h|_{Heis} = |(x, y, t)|_{Heis} = ((\sum_{i=1}^{n}(x_i^2 + y_i^2))^2 + t^2)^{1/4}$.

  The **Heisenberg metric** (or **Korányi metric**, **Cygan metric**, **gauge metric**) d_{Heis} is a **group norm metric** on $\mathcal{H}^n$ defined by

$$|x^{-1} \cdot y|_{Heis}.$$

  One can identify the Heisenberg group $\mathcal{H}^{n-1} = \mathbb{C}^{n-1} \times \mathbb{R}$ with $\partial\mathbb{H}_{\mathbb{C}}^n \setminus \{\infty\}$, where $\mathbb{H}_{\mathbb{C}}^n$ is the Hermitian (i.e., complex) hyperbolic n-space, and ∞ is any point of its boundary $\partial\mathbb{H}_{\mathbb{C}}^n$. So, the usual hyperbolic metric of $\mathbb{H}_{\mathbb{C}}^{n+1}$ induces a metric

[redacted]

on $\mathcal{H}^n$. The **Hamenstädt distance** on $\partial \mathbb{H}^n_{\mathbb{C}} \setminus \{\infty\}$ (Hersonsky–Paulin, 2004) is $\frac{1}{\sqrt{2}} d_{Heis}$.

Sometimes, the term *Cygan metric* is reserved for the extension of the metric d_{Heis} on whole $\mathbb{H}^n_{\mathbb{C}}$ and (Apanasov, 2004) for its generalization (via the *Carnot group* $\mathbb{F}^{n-1} \times Im\mathbb{F}$) on $\mathbb{F}$-hyperbolic spaces $\mathbb{H}^n_{\mathbb{F}}$ over numbers $\mathbb{F}$ that can be complex numbers, or quaternions or, for $n = 2$, octonions. Also, the generalization of d_{Heis} on Carnot groups *of Heisenberg type* is called the *Cygan metric*.

The second natural metric on $\mathcal{H}^n$ is the **Carnot–Carathéodory metric** (or **CC metric**, **sub-Riemannian metric**; cf. Chap. 7) d_C defined as the **length metric** (Chap. 6) using *horizontal vector fields* on $\mathcal{H}^n$. This metric is the **internal metric** (Chap. 4) corresponding to d_{Heis}.

The metric d_{Heis} is **bi-Lipschitz equivalent** with d_C but not with any Riemannian distance and, in particular, not with any Euclidean metric. For both metrics, the Heisenberg group $\mathcal{H}^n$ is a **fractal** since its **Hausdorff dimension**, $2n + 2$, is strictly greater than its **topological dimension**, $2n + 1$.

- **Metric between intervals**

  Let G be the set of all intervals $[a, b]$ of $\mathbb{R}$. The set G forms semigroups $(G, +)$ and $(G, \cdot)$ under addition $I + J = \{x + y : x \in I, y \in J\}$ and under multiplication $I \cdot J = \{x \cdot y : x \in I, y \in J\}$, respectively.

  The **metric between intervals** is a metric on G, defined by

$$\max\{|I|, |J|\}$$

  for all $I, J \in G$, where, for $K = [a, b]$, one has $|K| = |a - b|$.

- **Metric between games**

  Consider *positional games*, i.e., two-player nonrandom games of perfect information with real-valued outcomes. Play is alternating with a nonterminated game having move options for both players. Real-world examples include Chess, Go and Tic-Tac-Toe. Formally, let $F_{\mathbb{R}}$ be the universe of games defined inductively as follows:

  1. Every real number $r \in \mathbb{R}$ belongs to $F_{\mathbb{R}}$ and is called an *atomic game*.
  2. If $A, B \subset F_{\mathbb{R}}$ with $1 \leq |A|, |B| < \infty$, then $\{A|B\} \in F_{\mathbb{R}}$ (*nonatomic game*).

  Write any game $G = \{A|B\}$ as $\{G^L|G^R\}$, where $G^L = A$ and $G^R = B$ are the set of left and right moves of G, respectively.

  $F_{\mathbb{R}}$ becomes a commutative semigroup under the following addition operation:

  1. If p and q are atomic games, then $p + q$ is the usual addition in $\mathbb{R}$.
  2. $p + \{g_{l_1}, \dots |g_{r_1}, \dots\} = \{g_{l_1} + p, \dots |g_{r_1} + p, \dots\}$.
  3. If G and H are both nonatomic, then $\{G^L|G^R\} + \{H^L|H^R\} = \{I^L|I^R\}$, where $I^L = \{g_l + H, G + h_l : g_l \in G^L, h_l \in H^L\}$ and $I^R = \{g_r + H, G + h_r : g_r \in G^R, h_r \in H^R\}$.

For any game $G \in F_{\mathbb{R}}$, define the optimal outcomes $\overline{L}(G)$ and $\overline{R}(G)$ (if both players play optimally with Left and Right starting, respectively) as follows:

$\overline{L}(p) = \overline{R}(p) = p$ and $\overline{L}(G) = \max\{\overline{R}(g_l) : g_l \in G^L\}$, $\overline{R}(G) = \max\{\overline{L}(g_r) : g_r \in G^R\}$.

The **metric between games** G and H defined by Ettinger, 2000, is the following **extended metric** on $F_{\mathbb{R}}$:

$$\sup_X |\overline{L}(G+X) - \overline{L}(H+X)| = \sup_X |\overline{R}(G+X) - \overline{R}(H+X)|.$$

- **Helly semimetric**

  Consider a game $(\mathcal{A}, \mathcal{B}, H)$ between players A and B with *strategy sets* $\mathcal{A}$ and $\mathcal{B}$, respectively. Here $H = H(\cdot, \cdot)$ is the *payoff function*, i.e., if player A plays $a \in \mathcal{A}$ and player B plays $b \in \mathcal{B}$, then A pays H(a,b) to B. A player's *strategy set* is the set of available to him *pure strategies*, i.e., complete algorithms for playing the game, indicating the move for every possible situation throughout it.

  The **Helly semimetric** between strategies $a_1 \in \mathcal{A}$ and $a_2 \in \mathcal{A}$ of A is defined by

$$\sup_{b \in \mathcal{B}} |H(a_1, b) - H(a_2, b)|.$$

- **Factorial ring semimetric**

  Let $(A, +, \cdot)$ be a *factorial ring*, i.e., an *integral domain* (nonzero commutative ring with no nonzero zero divisors), in which every nonzero nonunit element can be written as a product of (nonunit) irreducible elements, and such factorization is unique up to permutation.

  The **factorial ring semimetric** is a semimetric on the set $A \backslash \{0\}$, defined by

$$\ln \frac{lcm(x, y)}{gcd(x, y)},$$

  where $lcm(x, y)$ is the *least common multiple*, and $gcd(x, y)$ is the *greatest common divisor* of elements $x, y \in A \backslash \{0\}$.
- **Frankild–Sather–Wagstaff metric**

  Let $\mathcal{G}(R)$ be the set of isomorphism classes, up to a shift, of semidualizing complexes over a local Noetherian commutative ring R. An *R-complex* is a particular sequence of R-module homomorphisms; see [FrSa07]) for exact definitions.

  The **Frankild–Sather–Wagstaff metric** ([FrSa07]) is a metric on $\mathcal{G}(R)$ defined, for any classes $[K], [L] \in \mathcal{G}(R)$, as the infimum of the *lengths* of chains of pairwise comparable elements starting with $[K]$ and ending with $[L]$.

[redacted]

10.2 Metrics on Binary Relations

A *binary relation* R on a set X is a subset of $X \times X$; it is the arc-set of the directed graph (X, R) with the vertex-set X.

A binary relation R which is *symmetric* ($(x, y) \in R$ implies $(y, x) \in R$), *reflexive* (all $(x, x) \in R$), and *transitive* ($(x, y), (y, z) \in R$ imply $(x, z) \in R$) is called an *equivalence relation* or a *partition* (of X into equivalence classes). Any *q-ary sequence* $x = (x_1, \dots, x_n)$, $q \geq 2$ (i.e., with $0 \leq x_i \leq q - 1$ for $1 \leq i \leq n$), corresponds to the partition $\{B_0, \dots, B_{q-1}\}$ of $V_n = \{1, \dots, n\}$, where $B_j = \{1 \leq i \leq n : x_i = j\}$ are the equivalence classes.

A binary relation R which is *antisymmetric* ($(x, y), (y, x) \in R$ imply $x = y$), reflexive, and transitive is called a *partial order*, and the pair (X, R) is called a *poset* (partially ordered set). A partial order R on X is denoted also by $\preceq$ with $x \preceq y$ if and only if $(x, y) \in R$. The order $\preceq$ is called *linear* if any elements $x, y \in X$ are *compatible*, i.e., $x \preceq y$ or $y \preceq x$.

A poset $(L, \preceq)$ is called a *lattice* if every two elements $x, y \in L$ have the *join* $x \vee y$ and the *meet* $x \wedge y$. All partitions of X form a lattice $\mathbb{P}_X$ by refinement; it is a sublattice of the lattice (by set-inclusion) of all binary relations.

- **Kemeny distance**

  The **Kemeny distance** between binary relations R_1 and R_2 on a set X is the **Hamming metric** $|R_1 \triangle R_2|$. It is twice the minimal number of inversions of pairs of adjacent elements of X which is necessary to obtain R_2 from R_1.

  If R_1, R_2 are *partitions*, then the Kemeny distance coincides with the **Mirkin–Tcherny distance**, and $1 - \frac{|R_1 \triangle R_2|}{n(n-1)}$ is the *Rand index*.

  If binary relations R_1, R_2 are *linear orders* (or *permutations*) on the set X, then the Kemeny distance coincides with the **Kendall τ distance** (Chap. 11).

- **Drápal–Kepka distance**

  The **Drápal–Kepka distance** between distinct *quasigroups* (differing from groups in that they need not be associative) $(X, +)$ and $(X, \cdot)$ is the **Hamming metric** $|\{(x, y) : x + y \neq x \cdot y\}|$ between their *Cayley tables*.

  For finite nonisomorphic groups, this distance is (Ivanyos, Le Gall and Yoshida, 2012) at least $2(\frac{|X|}{3})^2$ with equality (Drápal, 2003) for some 3-groups.

- **Editing metrics between partitions**

  Let X be a finite set, $|X| = n$, and let A, B be nonempty subsets of X. Let $\mathcal{P}_X$ be the set of partitions of X, and $P, Q \in \mathcal{P}_X$. Let $P_1, \dots, P_q$ be *blocks* in the partition P, i.e., the pairwise disjoint sets such that $X = P_1 \cup \dots \cup P_q$, $q \geq 1$. Let $P \vee Q$ and $P \wedge Q$ be the *join* and *meet* of P and Q in the *lattice* $\mathbb{P}_X$ *of partitions* of X.

  Consider the following *editing operations* on partitions (clusterings):

  - An *augmentation* transforms a partition P of $A \backslash \{B\}$ into a partition of A by either including the objects of B in a block, or including B as a new block;
  - An *removal* transforms a partition P of A into a partition of $A \backslash \{B\}$ by deleting the objects in B from each block that contains them;

- A *division* transforms one partition P into another by the simultaneous removal of B from P_i (where $B \subset P_i$, $B \neq P_i$), and augmentation of B as a new block;
- A *merging* transforms one partition P into another by the simultaneous removal of B from P_i (where $B = P_i$), and augmentation of B to P_j (where $j \neq i$);
- A *transfer* transforms one partition P into another by the simultaneous removal of B from P_i (where $B \subset P_i$), and augmentation of B to P_j (where $j \neq i$).

Define (see, say, [Day81]), using above operations, the following metrics on $\mathcal{P}_X$:

1. The minimum number of augmentations and removals of single objects needed to transform P into Q;
2. The minimum number of divisions, mergings, and transfers of single objects needed to transform P into Q;
3. The minimum number of divisions, mergings, and transfers needed to transform P into Q;
4. The minimum number of divisions and mergings needed to transform P into Q; in fact, it is equal to $|P| + |Q| - 2|P \vee Q|$;
5. $\sigma(P) + \sigma(Q) - 2\sigma(P \wedge Q)$, where $\sigma(P) = \sum_{P_i \in P} |P_i|(|P_i| - 1)$;
6. $e(P) + e(Q) - 2e(P \wedge Q)$, where $e(P) = \log_2 n + \sum_{P_i \in P} \frac{|P_i|}{n} \log_2 \frac{|P_i|}{n}$;
7. $2n - \sum_{P_i \in P} \max_{Q_j \in Q} |P_i \cap Q_j| - \sum_{Q_j \in Q} \max_{P_i \in P} |P_i \cap Q_j|$ (van Dongen, 2000).

The **maximum matching distance** (or *partition-distance* as defined in Gusfield, 2002) is (Réignier, 1965) the minimum number of elements that must be moved between the blocks of partition P in order to transform it into Q.

- **Rossi–Hamming metric**

  Given a partition $P = (P_1, \dots, P_q)$ of a finite set X, its *size* is defined as $s(P) = \frac{1}{2} \sum_{1 \le i \le q} |P_i|(|P_i| - 1)$. We call the **Rossi–Hamming metric** the metric between partitions P and Q, defined in Rossi, 2014, as

$$d_{RH}(P, Q) = s(P) + s(Q) - 2s(P \wedge Q).$$

  One has $d_{RH}(P, Q) \le s(P \vee Q) - s(P \wedge Q)$, where the right-hand side is the *size-based distance* (Rossi, 2011). The inequality is strict only for some noncomparable P, Q.

10.3 Metrics on Semilattices

Consider a poset $(L, \preceq)$. The *meet* (or *infimum*) $x \wedge y$ (if it exists) of two elements x and y is the unique element satisfying $x \wedge y \preceq x, y$, and $z \preceq x \wedge y$ if $z \preceq x, y$. The *join* (or *supremum*) $x \vee y$ (if it exists) is the unique element such that $x, y \preceq x \vee y$,

[redacted]

and $x \vee y \preceq z$ if $x, y \preceq z$. A poset $(L, \preceq)$ is called a *lattice* if every its elements x, y have the join $x \vee y$ and the meet $x \wedge y$. A poset is a *meet* (or *lower*) *semilattice* if only the meet-operation is defined. A poset is a *join* (or *upper*) *semilattice* if only the join-operation is defined.

A lattice $\mathbb{L} = (L, \preceq, \vee, \wedge)$ is called a *semimodular lattice* if the *modularity relation xMy* is symmetric: *xMy* implies *yMx* for any $x, y \in L$. Here two elements x and y are said to constitute a *modular pair*, in symbols *xMy*, if $x \wedge (y \vee z) = (x \wedge y) \vee z$ for any $z \preceq x$. A lattice $\mathbb{L}$ in which every pair of elements is modular, is called a *modular lattice*.

Given a lattice $\mathbb{L}$, a function $v : L \to \mathbb{R}_{\geq 0}$, satisfying $v(x \vee y) + v(x \wedge y) \leq v(x) + v(y)$ for all $x, y \in L$, is called a *subvaluation* on $\mathbb{L}$. A subvaluation v is *isotone* if $v(x) \leq v(y)$ whenever $x \preceq y$, and it is *positive* if $v(x) < v(y)$ whenever $x \preceq y$, $x \neq y$. A subvaluation v is called a *valuation* if it is isotone and $v(x \vee y) + v(x \wedge y) = v(x) + v(y)$ for all $x, y \in L$.

- **Lattice valuation metric**

  Let $\mathbb{L} = (L, \preceq, \vee, \wedge)$ be a lattice, and let v be an isotone subvaluation on $\mathbb{L}$. The *lattice subvaluation semimetric* d_v on L is defined by

  $$2v(x \vee y) - v(x) - v(y).$$

  (It can be defined also on some semilattices.) If v is a positive subvaluation on $\mathbb{L}$, one obtains a metric, called the **lattice subvaluation metric**. If v is a valuation, d_v is called the *valuation semimetric* and can be written as

  $$v(x \vee y) - v(x \wedge y) = v(x) + v(y) - 2v(x \wedge y).$$

  If v is a positive valuation on $\mathbb{L}$, one obtains a metric, called the **lattice valuation metric**, and the lattice is called a **metric lattice**.

  An example is the **Hamming distance** $d_H(A, B) = |A \cup B| - |A \cap B|$ on the lattice $(P(X), \cup, \cap)$ of all subsets of the set X. Cf. also the **Shannon distance** (Chap. 14), which can be seen as a distance on partitions.

  If $L = \mathbb{N}$ (the set of positive integers), $x \vee y = lcm(x, y)$ (least common multiple), $x \wedge y = gcd(x, y)$ (greatest common divisor), and the positive valuation $v(x) = \ln x$, then $d_v(x, y) = \ln \frac{lcm(x,y)}{gcd(x,y)}$.

  This metric can be generalized on any *factorial ring* equipped with a positive valuation v such that $v(x) \geq 0$ with equality only for the multiplicative unit of the ring, and $v(xy) = v(x) + v(y)$. Cf. **factorial ring semimetric**.
- **Finite subgroup metric**

  Let $(G, \cdot, e)$ be a group. Let $\mathbb{L} = (L, \subset, \cap)$ be the meet semilattice of all finite subgroups of the group $(G, \cdot, e)$ with the meet $X \cap Y$ and the valuation $v(X) = \ln |X|$.

  The **finite subgroup metric** is a **valuation metric** on L defined by

  $$v(X) + v(Y) - 2v(X \wedge Y) = \ln \frac{|X||Y|}{(|X \cap Y|)^2}.$$

- **Join semilattice distances**

  Let $\mathbb{L} = (L, \preceq, \vee)$ be a join semilattice, finite or infinite, such that every maximal chain in every interval $[x, y]$ is finite. For $x \preceq y$, the *height* $h(x, y)$ *of* y *above* x is the least cardinality of a finite maximal (by inclusion) chain of $[x, y]$ minus 1. Call the join semilattice $\mathbb{L}$ *semimodular* if for all $x, y \in L$, whenever there exists an element z covered by both x and y, the join $x \vee y$ covers both x and y, or, in other words, whenever elements x, y have a common lower bound z, it holds $h(x, x \vee y) \leq h(z, y)$. Any *tree* (i.e., all intervals $[x, z]$ are finite, each pair x, y of uncomparable elements have a least common upper bound $x \vee y$ but they never have a common lower bound) is semimodular. Consider the following distances on L:

  $d_{\text{path}}(x, y)$ is the path metric of the *Hasse diagram* of $(L, \preceq)$, i.e., a graph with vertex-set L and an edge between two elements if they are comparable.

  $d_{a.\text{path}}(x, y)$ is the smallest number of the form $h(x, z) + h(y, z)$, where z is a common upper bound of x and y, i.e., it is the **ancestral path distance**; cf. **pedigree-based distances** in Chap. 23. This and next distance reflect the way how Roman civil law and medieval canon law, respectively, measured degree of kinship.

  $d_{\max}(x, y)$ is defined by $\max(h(x, x \vee y), h(y, x \vee y))$.

  It holds $d_{a.\text{path}}(x, y) \geq d_{\text{path}}(x, y) \geq d_{\max}(x, y)$. Foldes, 2013, proved that $d_{\max}(x, y)$ is a metric if $\mathbb{L}$ is semimodular and that $d_{a.\text{path}}(x, y)$ is a metric if and only if $\mathbb{L}$ is semimodular, in which case $d_{a.\text{path}}(x, y) = d_{\text{path}}(x, y)$.

- **Gallery distance of flags**

  Let $\mathbb{L}$ be a lattice. A *chain* C in $\mathbb{L}$ is a subset of L which is *linearly ordered*, i.e., any two elements of C are compatible. A *flag* is a chain in $\mathbb{L}$ which is maximal with respect to inclusion. If $\mathbb{L}$ is a semimodular lattice, containing a finite flag, then $\mathbb{L}$ has a unique minimal and a unique maximal element, and any two flags C, D in $\mathbb{L}$ have the same cardinality, $n + 1$. Then n is the *height* of the lattice $\mathbb{L}$.

  Two flags C, D are called *adjacent* if either they are equal or D contains exactly one element not in C. A *gallery* from C to D of length m is a sequence of flags $C = C_0, C_1, \dots, C_m = D$ such that C_{i-1} and C_i are adjacent for $i = 1, \dots, m$.

  A **gallery distance of flags** (see [Abel91]) is a distance on the set of all flags of a semimodular lattice $\mathbb{L}$ with finite height defined as the minimum of lengths of galleries from C to D. It can be written as

  $$|C \vee D| - |C| = |C \vee D| - |D|,$$

  where $C \vee D = \{c \vee d : c \in C, d \in D\}$ is the subsemilattice generated by C and D. This distance is the **gallery metric** of the *chamber system* consisting of flags.

- **Scalar and vectorial metrics**

  Let $\mathbb{L} = (L, \leq, \max, \min)$ be a lattice with the join $\max\{x, y\}$, and the meet $\min\{x, y\}$ on a set $L \subset [0, \infty)$ which has a fixed number a as the greatest element and is closed under *negation*, i.e., for any $x \in L$, one has $\overline{x} = a - x \in L$.

[redacted]

The **scalar metric** d on L is defined, for $x \neq y$, by

$$d(x, y) = \max\{\min\{x, \overline{y}\}, \min\{\overline{x}, y\}\}.$$

The **scalar metric** d^* on $L^* = L \cup \{*\}$, $* \notin L$, is defined, for $x \neq y$, by

$$d^*(x, y) = \begin{cases} d(x, y), & \text{if} \quad x, y \in L, \\ \max\{x, \overline{x}\}, & \text{if} \quad y = *, x \neq *, \\ \max\{y, \overline{y}\}, & \text{if} \quad x = *, y \neq *. \end{cases}$$

Given a norm $||.||$ on $\mathbb{R}^n$, $n \geq 2$, the **vectorial metric** on L^n is defined by

$$||(d(x_1, y_1), \dots, d(x_n, y_n))||,$$

and the **vectorial metric** on $(L^*)^n$ is defined by

$$||(d^*(x_1, y_1), \dots, d^*(x_n, y_n))||.$$

The vectorial metric on $L_2^n = \{0, 1\}^n$ with l_1-*norm* on $\mathbb{R}^n$ is the **Fréchet–Nikodym–Aronszyan distance**. The vectorial metric on $L_m^n = \{0, \frac{1}{m-1}, \dots, \frac{m-2}{m-1}, 1\}^n$ with l_1-*norm* on $\mathbb{R}^n$ is the **Sgarro m-valued metric**. The vectorial metric on $[0, 1]^n$ with l_1-*norm* on $\mathbb{R}^n$ is the **Sgarro fuzzy metric**.

If L is L_m or $[0, 1]$, and $x = (x_1, \dots, x_n, x_{n+1}, \dots, x_{n+r})$, $y = (y_1, \dots, y_n, *, \dots, *)$, where $*$ stands in r places, then the vectorial metric between x and y is the **Sgarro metric** (see, for example, [CSY01]).

- **Metrics on Riesz space**

  A *Riesz space* (or *vector lattice*) is a partially ordered vector space $(V_{Ri}, \preceq)$ in which the following conditions hold:

  1. The vector space structure and the partial order structure are compatible: $x \preceq y$ implies $x + z \preceq y + z$, and $x \succ 0$, $\lambda \in \mathbb{R}, \lambda > 0$ implies $\lambda x \succ 0$;
  2. For any two elements $x, y \in V_{Ri}$ there exists the join $x \vee y \in V_{Ri}$ (in particular, the join and the meet of any finite set of elements from V_{Ri} exist).

  The **Riesz norm metric** is a **norm metric** on V_{Ri} defined by

$$||x - y||_{Ri},$$

  where $||.||_{Ri}$ is a *Riesz norm*, i.e., a *norm* on V_{Ri} such that, for any $x, y \in V_{Ri}$, the inequality $|x| \leq |y|$, where $|x| = (-x) \vee (x)$, implies $||x||_{Ri} \leq ||y||_{Ri}$.

  The space $(V_{Ri}, ||.||_{Ri})$ is called a *normed Riesz space*. In the case of completeness it is called a *Banach lattice*. All Riesz norms on a Banach lattice are equivalent.

  An element $e \in V_{Ri}^+ = \{x \in V_{Ri} : x \succ 0\}$ is called a *strong unit* of V_{Ri} if for each $x \in V_{Ri}$ there exists $\lambda \in \mathbb{R}$ such that $|x| \preceq \lambda e$. If a Riesz space V_{Ri} has a strong unit e, then $||x|| = \inf\{\lambda \in \mathbb{R} : |x| \preceq \lambda e\}$ is a Riesz norm, and one obtains on V_{Ri} a Riesz norm metric

$$\inf\{\lambda \in \mathbb{R} : |x - y| \preceq \lambda e\}.$$

A *weak unit* of V_{Ri} is an element e of V_{Ri}^{+} such that $e \wedge |x| = 0$ implies $x = 0$. A Riesz space V_{Ri} is called *Archimedean* if, for any two $x, y \in V_{Ri}^{+}$, there exists a natural number n, such that $nx \preceq y$. The **uniform metric** on an Archimedean Riesz space with a weak unit e is defined by

$$\inf\{\lambda \in \mathbb{R} : |x - y| \wedge e \preceq \lambda e\}.$$

- **Machida metric**

For a fixed integer $k \geq 2$ and the set $V_k = \{0, 1, \ldots, k - 1\}$, let $O_k^{(n)}$ be the set of all n-ary functions from $(V_k)^n$ into V_k and $O_k = \cup_{n=1}^{\infty} O_k^{(n)}$. Let Pr_k be the set of all *projections* pr_i^n over V_k, where $pr_i^n(x_1, \ldots, x_i, \ldots, x_n) = x_i$ for any $x_1, \ldots, x_n \in V_k$.

A *clone over* V_k is a subset C of O_k containing Pr_k and closed under (functional) composition. The set L_k of all clones over V_k is a lattice. The *Post lattice* L_2 defined over Boolean functions, is countable but any L_k with $k \geq 3$ is not. For $n \geq 1$ and a clone $C \in L_k$, let $C^{(n)}$ denote *n-slice* $C \cap O_k^{(n)}$.

For any two clones $C_1, C_2 \in L_k$, Machida, 1998, defined the distance to be 0 if $C_1 = C_2$ and $(\min\{n : C_1^{(n)} \neq C_2^{(n)}\})^{-1}$, otherwise. The lattice L_k of clones with this distance is a compact ultrametric space. Cf. **Baire metric** in Chap. 11.

[redacted]

Chapter 11
Distances on Strings and Permutations

An *alphabet* is a finite set $\mathcal{A}$, $|\mathcal{A}| \geq 2$, elements of which are called *characters* (or *symbols*). A *string* (or *word*) is a sequence of characters over a given finite alphabet $\mathcal{A}$. The set of all finite strings over the alphabet $\mathcal{A}$ is denoted by $W(\mathcal{A})$. Examples of real world applications, using distances and similarities of string pairs, are Speech Recognition, Bioinformatics, Information Retrieval, Machine Translation, Lexicography, Dialectology.

A *substring* (or *factor*, *chain*, *block*) of the string $x = x_1 \dots x_n$ is any contiguous subsequence $x_i x_{i+1} \dots x_k$ with $1 \leq i \leq k \leq n$. A *prefix* of a string x is any its substring starting with x_1; a *suffix* is any its substring finishing with x_n. If a string is a part of a text, then the *delimiters* (a space, a dot, a comma, etc.) are added to $\mathcal{A}$.

A *vector* is any finite sequence consisting of real numbers, i.e., a finite string over the *infinite alphabet* $\mathbb{R}$. A *frequency vector* (or *discrete probability distribution*) is any string $x_1 \dots x_n$ with all $x_i \geq 0$ and $\sum_{i=1}^{n} x_i = 1$. A *permutation* (or *ranking*) is any string $x_1 \dots x_n$ with all x_i being different numbers from $\{1, \dots, n\}$.

An *editing operation* is an operation on strings, i.e., a *symmetric binary relation* on the set of all considered strings. Given a set of editing operations $\mathcal{O} = \{O_1, \dots, O_m\}$, the corresponding **editing metric** (or *unit cost edit distance*) between strings x and y is the minimum number of editing operations from $\mathcal{O}$ needed to obtain y from x. It is the **path metric** of a graph with the vertex-set $W(\mathcal{A})$ and xy being an edge if y can be obtained from x by one of the operations from $\mathcal{O}$.

In some applications, a *cost function* is assigned to each type of editing operation; then the editing distance is the minimal total cost of transforming x into y. Given a set of editing operations $\mathcal{O}$ on strings, the corresponding **necklace editing metric** between cyclic strings x and y is the minimum number of editing operations from $\mathcal{O}$ needed to obtain y from x, minimized over all rotations of x.

The main editing operations on strings are:

- *Character indel*, i.e., insertion or deletion of a character;
- *Character replacement*;
- *Character swap*, i.e., an interchange of adjacent characters;

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_11

- *Substring move*, i.e., transforming, say, the string $x = x_1 \dots x_n$ into the string $x_1 \dots x_{i-1}\mathbf{x_j} \dots \mathbf{x_{k-1}}x_i \dots x_{j-1}x_k \dots x_n$;
- *Substring copy*, i.e., transforming, say, $x = x_1 \dots x_n$ into $x_1 \dots x_{i-1}\mathbf{x_j} \dots \mathbf{x_{k-1}}$ $x_i \dots x_n$;
- *Substring uncopy*, i.e., the removal of a substring provided that a copy of it remains in the string.

We list below the main distances on strings. However, some string distances will appear in Chaps. 15, 21 and 23, where they fit better, with respect to the needed level of generalization or specification.

11.1 Distances on General Strings

- **Levenstein metric**

  The **Levenstein metric** (or **edit distance**, *Hamming+Gap metric*, *shuffle-Hamming distance*) is (Levenstein, 1965) an editing metric on $W(\mathcal{A})$, obtained for $\mathcal{O}$ consisting of only character replacements and indels.

  The Levenstein metric between strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_n$ is

  $$d_L(x, y) = \min\{d_H(x^*, y^*)\},$$

  where x^*, y^* are strings of length k, $k \geq \max\{m, n\}$, over the alphabet $\mathcal{A}^* = \mathcal{A} \cup \{*\}$ so that, after deleting all new characters $*$, strings x^* and y^* shrink to x and y, respectively. Here, the *gap* is the new symbol $*$, and x^*, y^* are *shuffles* of strings x and y with strings consisting of only $*$.

  The *Levenstein similarity* is $1 - \frac{d_L(x,y)}{\max\{m,n\}}$.

  The **Damerau–Levenstein metric** (Damerau, 1964) is an editing metric on $W(\mathcal{A})$, obtained for $\mathcal{O}$ consisting only of character replacements, indels and transpositions. In the Levenstein metric, a transposition corresponds to two editing operations: one insertion and one deletion.

  The **constrained edit distance** (Oomen, 1986) is the Levenstein metric, but the ranges for the number of replacements, insertions and deletions are specified.
- **Editing metric with moves**

  The **editing metric with moves** is an editing metric on $W(\mathcal{A})$ ([Corm03]), obtained for $\mathcal{O}$ consisting of only substring moves and indels.
- **Editing compression metric**

  The **editing compression metric** is an editing metric on $W(\mathcal{A})$ ([Corm03]), obtained for $\mathcal{O}$ consisting of only indels, copy and uncopy operations.
- **Swap metric**

  The **swap metric** (or *interchange distance*, *Dodson distance*) is an editing metric on $W(\mathcal{A})$, obtained for $\mathcal{O}$ consisting only of character swaps, i.e., it is the minimum number of interchanges of adjacent pairs of symbols, converting x into y.

[redacted]

- **Indel metric**

  The **indel metric** is an editing metric on $W(\mathcal{A})$, obtained for $\mathcal{O}$ consisting of only indels. It is an analog of the **Hamming metric** $|X \Delta Y|$ between sets X and Y. For strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_n$ it is $m + n - 2LCS(x, y)$, where the similarity $LCS(x, y)$ is the length of the longest common subsequence of x and y.

  The **factor distance** is $m + n - 2LCF(x, y)$, where $LCF(x, y)$ is the length of the longest common substring (factor) of x and y.

  The *LCS ratio* and the *LCF ratio* are the similarities $\frac{LCS(x,y)}{\min\{m,n\}}$ and $\frac{LCF(x,y)}{\min\{m,n\}}$, respectively; sometimes, the denominator is $\max\{m, n\}$ or $\frac{m+n}{2}$.

- **Prefix, suffix, and substring distances**

  Given strings x and y, their **prefix distance**, **suffix distance**, and **substring distance** are the numbers of symbol occurrences in the strings that do not belong to their longest common prefix, suffix or substring, respectively. Cf. **Jaro–Winkler similarity**, **factor distance.**

- **Antidistance**

  There are $(n - 1)!$ *circular permutations*, i.e., cyclic orders, of a set X of size n. The **antidistance** between circular permutations x and y is the **swap metric** between x and the reversal of y.

  Also, given complex $n \times n$ matrices A and B, the *unitary similarity orbit* through B is $\sup_{U \in \mathbb{U}_n} ||U^* BU||_\infty$, where $U \in \mathbb{U}_n$ is the group of unitary matrices. Ando, 1996, define *anti-distance* between A and this orbit as $\sup_{U \in \mathbb{U}_n} ||A - U^* BU||_\infty$.

  Also, given a simple connected graph (V, E), we assign directions to edges and the weight of each edge (either 1 or -1) depending on the direction of the traverse. Iravanian, 2012, define *anti-distance* $d(u, v) = -d(v, u)$ between vertices as the weighted average length of all simple paths from u to v.

- **Edit distance with costs**

  Given a set of editing operations $\mathcal{O} = \{O_1, \dots, O_m\}$ and a *weight* (or *cost function*) $w_i \geq 0$, assigned to each type O_i of operation, the **edit distance with costs** between strings x and y is the minimal total cost of an *editing path* between them, i.e., the minimal sum of weights for a sequence of operations transforming x into y.

  The **normalized edit distance** between strings x and y (Marzal–Vidal, 1993) is the minimum, over all editing paths P between them, of $\frac{W(P)}{L(P)}$, where $W(P)$ and $L(P)$ are the total cost and the length of the editing path P.

- **Transduction edit distances**

  The **Levenstein metric** with costs between strings x and y is modeled in [RiYi98] as a memoryless stochastic transduction between x and y.

  Each step of transduction generates either a character replacement pair (a, b), a deletion pair $(a, \emptyset)$, an insertion pair $(\emptyset, b)$, or the specific termination symbol t according to a probability function $\delta : E \cup \{t\} \to [0, 1]$, where E is the set of all possible above pairs. Such a transducer induces a probability function on the set of all sequences of operations.

The **transduction edit distances** between strings x and y are ([RiYi98]) $\ln p$ of the following probabilities p:

– for the **Viterbi edit distance**, the probability of the most likely sequence of editing operations transforming x into y;
– for the **stochastic edit distance**, the probability of the string pair (x, y).

This model allows one to learn, in order to reduce error rate, the edit costs for the Levenstein metric from a corpus of examples (training set of string pairs). This learning is automatic; it reduces to estimating the parameters of above transducer.

- **Bag distance**

  The **bag distance** (or *multiset metric*, *counting filter*) is a metric on $W(\mathcal{A})$ defined (Navarro, 1997) by

  $$\max\{|X\backslash Y|, |Y\backslash X|\}$$

  for any strings x and y, where X and Y are the *bags of symbols* (multisets of characters) in strings x and y, respectively, and, say, $|X\backslash Y|$ counts the number of elements in the multiset $X\backslash Y$. It is a (computationally) cheap approximation of the **Levenstein metric**. Cf. **metrics between multisets** in Chap. 1.

- **Marking metric**

  The **marking metric** is a metric on $W(\mathcal{A})$ ([EhHa88]) defined by

  $$\ln_2((\mathit{diff}(x, y) + 1)(\mathit{diff}(y, x) + 1))$$

  for any strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_n$, where $\mathit{diff}(x, y)$ is the minimal cardinality $|M|$ of a subset $M \subset \{1, \dots, m\}$ such that any substring of x, not containing any x_i with $i \in M$, is a substring of y.

  Another metric defined in [EhHa88], is $\ln_2(\mathit{diff}(x, y) + \mathit{diff}(y, x) + 1)$.

- **Transformation distance**

  The **transformation distance** is an **editing distance with costs** on $W(\mathcal{A})$ (Varre–Delahaye–Rivals, 1999) obtained for $\mathcal{O}$ consisting only of substring copy, uncopy and substring indels. The distance between strings x and y is the minimal cost of transformation x into y using these operations, where the cost of each operation is the length of its description.

  For example, the description of the copy requires a binary code specifying the type of operation, an offset between the substring locations in x and in y, and the length of the substring. A code for insertion specifies the type of operation, the length of the substring and the sequence of the substring.

- **L_1-rearrangement distance**

  The **L_1-rearrangement distance** (Amir et al., 2007) between strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_m$ is defined by

  $$\min_{\pi} \sum_{i=1}^{m} |i - \pi(i)|,$$

[redacted]

where $\pi : \{1, \dots, m\} \rightarrow \{1, \dots, m\}$ is a permutation transforming x into y; if there are no such permutations, the distance is equal to ∞.

The **L_∞-rearrangement distance** (Amir et al., 2007) between x and y is $\min_\pi \max_{1 \leq i \leq m} |i - \pi(i)|$ and it is ∞ if such a permutation does not exist.

Cf. **genome rearrangement distances** in Chap. 23.

- **Normalized information distance**

  The **normalized information distance** d between two binary strings x and y is a symmetric function on $W(\{0,1\})$ ([LCLMV04]) defined by

$$\frac{\max\{K(x|y^*), K(y|x^*)\}}{\max\{K(x), K(y)\}}.$$

Here, for binary strings u and v, u^* is a shortest binary program to compute u on an appropriate (i.e., using a *Turing-complete* language) universal computer, the *Kolmogorov complexity* (or *algorithmic entropy*) $K(u)$ is the length of u^* (the ultimate compressed version of u), and $K(u|v)$ is the length of the shortest program to compute u if v is provided as an auxiliary input.

The function $d(x, y)$ is a metric up to small error term: $d(x, x) = O((K(x))^{-1})$, and $d(x, z) - d(x, y) - d(y, z) = O((\max\{K(x), K(y), K(z)\})^{-1})$. Cf. in Chap. 15 the **shared information distance** $H(X|Y) + H(Y|X)$ between sources X and Y.

The Kolmogorov complexity is uncomputable and depends on the chosen computer language; so, instead of $K(u)$, were proposed the *minimum message length* (shortest overall message) by Wallace, 1968, and the *minimum description length* (largest compression of data) by Rissanen, 1978.

The **normalized compression distance** is a metric on $W(\{0,1\})$ (derived by Cilibrasi and Vitányi, 2005, from [LCLMV04, BGLVZ98]) defined by

$$\frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

for any binary strings x and y, where $C(x), C(y)$, and $C(xy)$ denote the size of the compression (by fixed compressor C, such as gzip, bzip2, or PPMZ) of strings x, y, and their *concatenation* xy. It is an approximation of the normalized information distance. A similar distance is defined by $\frac{C(xy)}{C(x)+C(y)} - \frac{1}{2}$.

- **Lempel–Ziv distance**

  The **Lempel–Ziv distance** between two binary strings x and y of length n is

$$\max\{\frac{LZ(x|y)}{LZ(x)}, \frac{LZ(y|x)}{LZ(y)}\},$$

where $LZ(x) = \frac{|P(x)| \log |P(x)|}{n}$ is the *Lempel–Ziv complexity* of x, approximating its *Kolmogorov complexity* $K(x)$. Here $P(x)$ is the set of nonoverlapping substrings into which x is parsed sequentially, so that the new substring is not yet contained in the set of substrings generated so far. For example, such a *Lempel–Ziv* *parsing* for $x = 001100101010011$ is $0|01|1|00|10|101|001|11$. Now, $LZ(x|y) = \frac{|P(x) \backslash P(y)| \log |P(x) \backslash P(y)|}{n}$.

- **Anthony–Hammer similarity**

  The **Anthony–Hammer similarity** between a binary string $x = x_1 \dots x_n$ and the set Y of binary strings $y = y_1 \dots y_n$ is the maximal number m such that, for every m-subset $M \subset \{1, \dots, n\}$, the substring of x, containing only x_i with $i \in M$, is a substring of some $y \in Y$ containing only y_i with $i \in M$.

- **Jaro similarity**

  Given strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_n$, call a character x_i *common with* y if $x_i = y_j$, where $|i - j| \le \frac{\min\{m,n\}}{2}$. Let $x' = x'_1 \dots x'_{m'}$ be all the characters of x which are common with y (in the same order as they appear in x), and let $y' = y'_1 \dots y'_{n'}$ be the analogic string for y.

  The **Jaro similarity** $Jaro(x, y)$ between strings x and y is defined by

$$\frac{1}{3}\left(\frac{m'}{m} + \frac{n'}{n} + \frac{|\{1 \le i \le \min\{m', n'\} : x'_i = y'_i\}|}{\min\{m', n'\}}\right).$$

  This and following two similarities are used in Record Linkage.

- **Jaro–Winkler similarity**

  The **Jaro–Winkler similarity** between strings x and y is defined by

$$Jaro(x, y) + \frac{\max\{4, LCP(x, y)\}}{10}(1 - Jaro(x, y)),$$

  where $Jaro(x, y)$ is the **Jaro similarity**, and $LCP(x, y)$ is the length of the longest common prefix of x and y.

- **q-gram similarity**

  Given an integer $q \ge 1$ (usually, q is 2 or 3), the **q-gram similarity** between strings x and y is defined by

$$\frac{2q(x, y)}{q(x) + q(y)},$$

  where $q(x)$, $q(y)$ and $q(x, y)$ are the sizes of multisets of all *q-grams* (substrings of length q) occurring in x, y and both of them, respectively.

  Sometimes, $q(x, y)$ is divided not by the average of $q(x)$ and $q(y)$, as above, but by their minimum, maximum or *harmonic mean* $\frac{2q(x)q(y)}{q(x)+q(y)}$. Cf. **metrics between multisets** in Chap. 1 and, in Chap. 17, **Dice similarity**, **Simpson similarity**, **Braun–Blanquet similarity** and **Anderberg similarity**.

  The q-gram similarity is an example of **token-based similarities**, i.e., ones defined in terms of *tokens* (selected substrings or words). Here tokens are q-grams. A generic **dictionary-based metric** between strings x and y is $|D(x) \Delta D(y)|$, where $D(z)$ denotes the full *dictionary* of z, i.e., the set of all of its substrings.

[redacted]

- **Prefix-Hamming metric**

  The **prefix-Hamming metric** between strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_n$ is

  $$(\max\{m, n\} - \min\{m, n\}) + |\{1 \le i \le \min\{m, n\} : x_i \neq y_i\}|.$$

- **Weighted Hamming metric**

  If $(\mathcal{A}, d)$ is a metric space, then the **weighted Hamming metric** between strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_m$ is defined by

  $$\sum_{i=1}^{m} d(x_i, y_i).$$

  The term *weighted Hamming metric* (or *weighted Hamming distance*) is also used for $\sum_{1 \le i \le m, x_i \neq y_i} w_i$, where, for any $1 \le i \le m$, $w(i) > 0$ is its *weight*.
- **Fuzzy Hamming distance**

  If $(\mathcal{A}, d)$ is a metric space, the **fuzzy Hamming distance** between strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_m$ is an **editing distance with costs** on $W(\mathcal{A})$ obtained for $\mathcal{O}$ consisting of only indels, each of fixed cost $q > 0$, and *character shifts* (i.e., moves of 1-character substrings), where the cost of replacement of i by j is a function $f(|i{-}j|)$. This distance is the minimal total cost of transforming x into y by these operations. Bookstein–Klein–Raita, 2001, introduced this distance for Information Retrieval and proved that it is a metric if f is a monotonically increasing concave function on integers vanishing only at 0.

  The case $f(|i - j|) = C|i - j|$, where $C > 0$ is a constant and $|i - j|$ is a time shift, corresponds to the Victor–Purpura **spike train distance** in Chap. 23.

  Ralescu, 2003, introduced, for Image Retrieval, another **fuzzy Hamming distance** on $\mathcal{R}^m$. The **Ralescu distance** between two strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_m$ is the fuzzy cardinality of the difference fuzzy set $D_\alpha(x, y)$ (where α is a parameter) with membership function

  $$\mu_i = 1 - e^{-\alpha(x_i - y_i)^2}, 1 \le i \le m.$$

  The *nonfuzzy cardinality of the fuzzy set* $D_\alpha(x, y)$ approximating its fuzzy cardinality is $|\{1 \le i \le m : \mu_i > \frac{1}{2}\}|$.
- **Needleman–Wunsch–Sellers metric**

  If $(\mathcal{A}, d)$ is a metric space, the **Needleman–Wunsch–Sellers metric** (or *global alignment metric*) is an **editing distance with costs** on $W(\mathcal{A})$ ([NeWu70]), obtained for $\mathcal{O}$ consisting of only indels, each of fixed cost $q > 0$, and character replacements, where the cost of replacement of i by j is $d(i, j)$. This metric is the minimal total cost of transforming x into y by these operations. It is

  $$\min\{d_{wH}(x^*, y^*)\},$$

where x^*, y^* are strings of length k, $k \geq \max\{m, n\}$, over the alphabet $\mathcal{A}^* = \mathcal{A} \cup \{*\}$, so that, after deleting all new characters $*$, strings x^* and y^* shrink to x and y, respectively. Here $d_{wH}(x^*, y^*)$ is the **weighted Hamming metric** between x^* and y^* with weight $d(x_i^*, y_i^*) = q$ (i.e., the editing operation is an indel) if one of x_i^*, y_i^* is $*$, and $d(x_i^*, y_i^*) = d(i, j)$, otherwise.

The **Gotoh–Smith–Waterman distance** (or *string distance with affine gaps*) is a more specialized editing metric with costs (see [Goto82]). It discounts mismatching parts at the beginning and end of the strings x, y, and introduces two indel costs: one for starting an *affine gap* (contiguous block of indels), and another one (lower) for extending a gap.

- **Duncan metric**

Consider the set X of all strictly increasing infinite sequences $x = \{x_n\}_n$ of positive integers. Define $N(n, x)$ as the number of elements in $x = \{x_n\}_n$ which are less than n, and $\delta(x)$ as the *density* of x, i.e., $\delta(x) = \lim_{n \to \infty} \frac{N(n,x)}{n}$. Let Y be the subset of X consisting of all sequences $x = \{x_n\}_n$ for which $\delta(x) < \infty$.

The **Duncan metric** is a metric on Y defined, for $x \neq y$, by

$$\frac{1}{1 + LCP(x, y)} + |\delta(x) - \delta(y)|,$$

where $LCP(x, y)$ is the length of the longest common prefix of x and y.

- **Martin metric**

The **Martin metric** d^a between strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_n$ is

$$|2^{-m} - 2^{-n}| + \sum_{t=1}^{\max\{m,n\}} \frac{a_t}{|\mathcal{A}|^t} \sup_z |k(z, x) - k(z, y)|,$$

where z is any string of length t, $k(z, x)$ is the *Martin kernel* of a *Markov chain* $M = \{M_t\}_{t=0}^{\infty}$, and the sequence $a \in \{a = \{a_t\}_{t=0}^{\infty} : a_t > 0, \sum_{t=1}^{\infty} a_t < \infty\}$ is a parameter.

- **Baire metric**

The **Baire metric** is an ultrametric between strings x and y defined, for $x \neq y$, by

$$\frac{1}{1 + LCP(x, y)},$$

where $LCP(x, y)$ is the length of the longest common prefix of strings (finite or infinite) x and y. Cf. **Baire space** in Chap. 2.

Given an infinite *cardinal number* κ and a set A of cardinality κ, the Cartesian product of countably many copies of A endowed with above ultrametric $\frac{1}{1+LCP(x,y)}$ is called the **Baire space of weight** κ and denoted by $B(\kappa)$. In particular, $B(\aleph_0)$ (called the *Baire* 0*-dimensional space*) is homeomorphic to the space *Irr* of irrationals with **continued fraction metric** (Chap. 12).

[redacted]

- **Generalized Cantor metric**

  The **generalized Cantor metric** (or, sometimes, *Baire distance*) is an ultrametric between infinite strings x and y defined, for $x \neq y$, by

  $$a^{1+LCP(x,y)},$$

  where a is a fixed number from the interval $(0, 1)$, and $LCP(x, y)$ is the length of the longest common prefix of x and y.

  This ultrametric space is **compact**. In the case $a = \frac{1}{2}$, this metric was considered on a remarkable **fractal**, the *Cantor set*; cf. **Cantor metric** in Chap. 18. Another important case is $a = \frac{1}{e} \approx 0.367879441$.

  Comyn–Dauchet, 1985, and Kwiatkowska, 1990, introduced some analogs of generalized Cantor metric for *traces*, i.e., equivalence classes of strings with respect to a congruence relation identifying strings x, y that are identical up to permutation of concurrent actions ($xy = yx$).
- **Parikh distance**

  Given an ordered alphabet $\mathcal{A} = \{a_1, \dots, a_k\}$, the **Parikh distance** between words x and y over it is the **Manhattan metric** $\sum_{i=1}^{k} |x_i - y_i|$ between their *Parikh maps* (or *commutative images*) $P(x)$ and $P(y)$, where, for a word w, w_i denotes the number of occurrences of a_i in w and $P(w)$ is $(w_1, \dots, w_k)$.
- **Parentheses string metrics**

  Let P_n be the set of all strings on the alphabet $\{(,)\}$ generated by a grammar and having n open and n closed parentheses. A **parentheses string metric** is an editing metric on P_n corresponding to a given set of editing operations.

  For example, the **Monjardet metric** (Monjardet, 1981) between two strings $x, y \in P_n$ is the minimum number of adjacent parentheses interchanges ("()" to ")(" or ")(" to "()") needed to obtain y from x. It is the **Manhattan** metric between their *representations* p_x and p_x, where $p_z = (p_z(1), \dots, p_z(n))$ and $p_z(i)$ is the number of open parentheses written before the i-th closed parentheses of $z \in P_n$.

  There is a bijection between parentheses strings and binary trees; cf. the **tree rotation distance** in Chap. 15.
- **Dehornoy–Autord distance**

  The **Dehornoy–Autord distance** (2010) between two shortest expressions x and y of a permutation as a product of transpositions t_i, is the minimal, needed to get x from y, number of *braid relations*: $t_i t_j t_i = t_j t_i t_j$ with $|i-j| = 1$ and $t_i t_j = t_j t_i$ with $|i-j| \geq 2$.

  This distance can be extended to the decompositions of any given *positive braid* in terms of *Artin's generators*. The permutations corresponds to the *simple braids* which are the divisors of *Garside's fundamental braid* in the *braid monoid*.
- **Schellenkens complexity quasi-metric**

  The **Schellenkens complexity quasi-metric** between infinite strings $x = (x_i)$ and $y = (y_i)$ $(i = 0, 1, \dots)$ over $\mathbb{R}_{\geq 0}$ with $\sum_{i=0}^{\infty} 2^{-i} \frac{1}{x_i} < \infty$ (seen as complexity functions) is defined (Schellenkens, 1995) by

$$\sum_{i=0}^{\infty} 2^{-i} \max\{0, \frac{1}{x_i} - \frac{1}{y_i}\}.$$

- **Graev metrics**

  Let (X, d) be a metric space. Let $\overline{X} = X \cup X' \cup \{e\}$, where $X' = \{x' : x \in X\}$ is a disjoint copy of X, and $e \notin X \cup X'$. We use the notation $(e')' = e$ and $(x')' = x$ for any $x \in X$; also, the letters x, y, x_i, y_i will denote elements of $\overline{X}$. Let $(\overline{X}, D)$ be a metric space such that $D(x, y) = D(x', y') = d(x, y)$, $D(x, e) = D(x', e)$ and $D(x, y') = D(x', y)$ for all $x, y \in X$.

  Denote by $W(X)$ the set of all words over $\overline{X}$ and, for each word $w \in W(X)$, denote by $l(w)$ its length. A word $w \in W(X)$ is called *irreducible* if $w = e$ or $w = x_0 \dots x_n$, where $x_i \neq e$ and $x_{i+1} \neq x_i'$ for $0 \leq i < n$.

  For each word w over $\overline{X}$, denote by $\hat{w}$ the unique irreducible word obtained from w by successively replacing any occurrence of xx' in w by e and eliminating e from any occurrence of the form $w_1 e w_2$, where $w_1 = w_2 - \emptyset$ is excluded.

  Denote by $F(X)$ the set of all irreducible words over $\overline{X}$ and, for $u, v \in F(X)$, define $u \cdot v = w'$, where w is the concatenation of words u and v. Then $F(X)$ becomes a group; its identity element is the (nonempty) word e.

  For any two words $v = x_0 \dots x_n$ and $u = y_0 \dots y_n$ over $\overline{X}$ of the same length, let $\rho(v, u) = \sum_{i=0}^{n} D(x_i, y_i)$. The **Graev metric** between two irreducible words $u = u, v \in F(X)$ is defined ([DiGa07]) by

  $$\inf\{\rho(u^*, v^*) : u^*, v^* \in W(X),\ l(u^*) = l(v^*),\ \widehat{u^*} = u, \widehat{v^*} = v\}.$$

  Graev proved that this metric is **bi-invariant metric** on $F(X)$ and that $F(X)$ is a topological group in the topology induced by it.

- **String-induced alphabet distance**

  Let $a = (a_1, \dots, a_m)$ be a finite string over alphabet X, $|X| = n \geq 2$. Let $A(x) = \{1 \leq i \leq m : a_i = x\} \neq \emptyset$ for any $x \in X$.

  The **string-induced distance** between symbols $x, y \in X$ is the **set-set distance** (Chap. 1) defined by

  $$d_a(x, y) = \min\{|i - j| : i \in A(x), j \in A(y)\}.$$

  A *k*-**radius sequence** (Jaromczyk and Lonc, 2004) is a string a over X with $\max_{x,y \in X} d_a(x, y) \leq k$, i.e., any two symbols (say, large digital images) occur in some window (say, memory cache) of length $k+1$. Minimal length m corresponds to most efficient pipelining of images when no more than $k + 1$ of them can be placed in main memory in any given time.

[redacted]

11.2 Distances on Permutations

A *permutation* (or *ranking*) is any string $x_1 \dots x_n$ with all x_i being different numbers from $\{1, \dots, n\}$; a *signed permutation* is any string $x_1 \dots x_n$ with all $|x_i|$ being different numbers from $\{1, \dots, n\}$. Denote by $(Sym_n, \cdot, id)$ the group of all permutations of the set $\{1, \dots, n\}$, where *id* is the *identity mapping*.

The restriction, on the set Sym_n of all n-permutation vectors, of any metric on $\mathbb{R}^n$ is a metric on Sym_n; the main example is the l_p**-metric** $(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}, p \geq 1$.

The main editing operations on permutations are:

- *Block transposition*, i.e., a substring move;
- *Character move*, i.e., a transposition of a block consisting of only one character;
- *Character swap*, i.e., interchanging of any two adjacent characters;
- *Character exchange*, i.e., interchanging of any two characters (in Group Theory, it is called *transposition*);
- *One-level character exchange*, i.e., exchange of characters x_i and x_j, $i < j$, such that, for any k with $i < k < j$, either $\min\{x_i, x_j\} > x_k$, or $x_k > \max\{x_i, x_j\}$;
- *Block reversal*, i.e., transforming, say, the permutation $x = x_1 \dots x_n$ into the permutation $x_1 \dots x_{i-1} \mathbf{x_j x_{j-1} \dots x_{i+1} x_i} x_{j+1} \dots x_n$ (so, a swap is a reversal of a block consisting only of two characters);
- *Signed reversal*, i.e., a reversal in signed permutation, followed by multiplication on -1 of all characters of the reversed block.

Below we list the most used editing and other metrics on Sym_n.

- **Hamming metric on permutations**

  The **Hamming metric on permutations** d_H is an editing metric on Sym_n, obtained for $\mathcal{O}$ consisting of only character replacements. It is a **bi-invariant** metric. Also, $n - d_H(x, y)$ is the number of fixed points of xy^{-1}.
- **Spearman ρ distance**

  The **Spearman ρ distance** is the Euclidean metric on Sym_n:

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

  Its square is a 2**-near-metric**. Cf. **Spearman ρ rank correlation** in Chap. 17.
- **Spearman footrule distance**

  The **Spearman footrule distance** is the l_1*-metric* on Sym_n:

$$\sum_{i=1}^{n} |x_i - y_i|.$$

  Cf. **Spearman footrule similarity** in Chap. 17.

  Both above Spearman distances are **bi-invariant**.

- **Kendall τ distance**

  The **Kendall τ distance** (or *inversion metric*, permutation **swap metric**, *bubble-sort distance*) I is an editing metric on Sym_n, obtained for $\mathcal{O}$ consisting only of character swaps.

  In terms of Group Theory, $I(x, y)$ is the number of adjacent transpositions needed to obtain x from y. Also, $I(x, y)$ is the number of *relative inversions* of x and y, i.e., pairs (i, j), $1 \leq i < j \leq n$, with $(x_i - x_j)(y_i - y_j) < 0$. Cf. **Kendall τ rank correlation** in Chap. 17.

  In [BCFS97] the following metrics, associated with $I(x, y)$, were given:

  1. $\min_{z \in Sym_n}(I(x, z) + I(z^{-1}, y^{-1}))$;
  2. $\max_{z \in Sym_n} I(zx, zy)$;
  3. $\min_{z \in Sym_n} I(zx, zy) = T(x, y)$, where T is the **Cayley metric**;
  4. editing metric with $\mathcal{O}$ consisting only of one-level character exchanges.

- **Daniels–Guilbaud semimetric**

  The **Daniels–Guilbaud semimetric** (see [Monj98]) is defined, for any $x, y \in Sym_n$, as the number of triples (i, j, k), $1 \leq i < j < k \leq n$, such that (x_i, x_j, x_k) is not a cyclic shift of (y_i, y_j, y_k). So, it is 0 if and only if x is a cyclic shift of y.
- **Cayley metric**

  The **Cayley metric** (or **transposition distance**) T is an editing metric on Sym_n, obtained for $\mathcal{O}$ consisting only of character exchanges. In terms of Group Theory, $T(x, y)$ is the minimum number of transpositions needed to obtain x from y.

  The metric T is **bi-invariant**. Also, $n - T(x, y)$ is the number of cycles in xy^{-1}, and, for the **Hamming metric on permutations**, $d_H(x, y) - T(x, y)$ is the number of cycles with length at least 2 in xy^{-1}.
- **Ulam metric**

  The **Ulam metric** (or **permutation editing metric**) U is an editing metric on Sym_n, obtained for $\mathcal{O}$ consisting only of character moves. It is the half of the **indel metric** on Sym_n.

  Also, $n - U(x, y) = LCS(x, y) = LIS(xy^{-1})$, where $LCS(x, y)$ is the length of the longest common subsequence (not necessarily a substring) of x and y, while $LIS(z)$ is the length of the longest increasing subsequence of $z \in Sym_n$.

  This and the preceding six metrics are **right-invariant**.
- **Reversal metric**

  The **reversal metric** is an editing metric on Sym_n, obtained for $\mathcal{O}$ consisting only of block reversals.
- **Signed reversal metric**

  The **signed reversal metric** (Sankoff, 1989) is an editing metric on the set of all $2^n n!$ signed permutations of the set $\{1, \ldots, n\}$, obtained for $\mathcal{O}$ consisting only of signed reversals.

  This metric is used in Biology, where a signed permutation represents a single-chromosome genome, seen as a permutation of genes (along the chromosome) each having a direction (so, a sign $+$ or $-$).

[redacted]

- **Chain metric**

  The **chain metric** (or *rearrangement metric*) is a metric on Sym_n ([Page65]) defined, for any $x, y \in Sym_n$, as the minimum number, minus 1, of chains (substrings) $y'_1, \dots, y'_t$ of y, such that x can be *parsed* (concatenated) into, i.e., $x = y'_1 \dots y'_t$.
- **Lexicographic metric**

  The **lexicographic metric** (Golenko–Ginzburg, 1973) is a metric on Sym_n:

  $$|N(x) - N(y)|,$$

  where $N(x)$ is the ordinal number of the position (among $1, \dots, n!$) occupied by the permutation x in the *lexicographic ordering* of the set Sym_n.

  In the *lexicographic ordering* of Sym_n, $x = x_1 \dots x_n \prec y = y_1 \dots y_n$ if there exists $1 \leq i \leq n$ such that $x_1 = x_1, \dots, x_{i-1} = y_{i-1}$, but $x_i < y_i$.
- **Fréchet permutation metric**

  The **Fréchet permutation metric** is the **Fréchet product metric** (Chap. 4) on the set Sym_∞ of permutations of positive integers defined by

  $$\sum_{i=1}^{\infty} \frac{1}{2^i} \frac{|x_i - y_i|}{1 + |x_i - y_i|}.$$
- **Distance-rationalizable voting rule**

  Let $e = (\pi_1, \dots, \pi_m)$ be a finite string over alphabet Sym_n; it can be seen as an election in which, for each i, $1 \leq i \leq m$, the *voter* v_i give the ranking $\pi_i = (\pi_i(c_1), \dots, \pi_i(c_n))$ on the set $C = \{c_1, \dots, c_n)$ of *candidates*. Let $X = Sym_n^m$ be the set of all possible elections with m voters in each.

  A *voting rule* is any map $R : X \to P(C)$ assigning to each election e a set $R(e) \subset C$ of its *R-winners*. For example, the winners of *plurality rule* are candidates with the largest number of first-place votes. A candidate is a *unanimity winner* if all voters rank him first. A candidate c_i is a *Condorcet winner* if for each $c_j \in C \setminus \{c_i\}$, a strict majority of voters prefer c_i to c_j. A candidate is a *Dodson winner* if the number of swaps of adjacent candidates in the rankings by voters after which he became a Condorcet winner, is minimal. So, $|R(e)| \leq 1$ for elections with unanimity or Condorcet rule, and $|R(e)| \geq 1$ for plurality or Dodson rule.

  A *consensus class* is a pair (Y, W), where $Y \subset X$ is a set of elections and W is a voting rule with unique (Y, W)-*winner* (i.e., $|W(e)| = 1$) for all $e \in Y$. Let $\mathcal{U}$ and $\mathcal{C}$ denote the consensus classes of all elections having the Condorcet winner and the unanimity winner, respectively.

  Given a distance d on X and consensus class (Y, W), the voting rule R is called (Meskanen–Nurmi, 2008, and Elkind–Faliszewski–Slinko, 2009) **$(d; (Y, W))$-distance-rationalizable** if, for each election e, a candidate c_i is its R-winner if and only if he is the (Y, W)-winner in a d-closest election in Y.

The plurality rule is $(d_H; \mathcal{U})$-rationalizable, where $d_H(e, e')$ is the **Hamming distance** $|\{i \leq i \leq m : \pi_i \neq \pi'_i\}|$. The Dodson rule is $(d_{sw}; \mathcal{C})$-rationalizable, where $d_{sw}(e, e') = \sum_{1 \leq i \leq m} d_{sw}(\pi_i, \pi'_i)$ and d_{sw} on rankings is the **swap metric**.

Similar framework (minimization of an *aggregation function* of distances between a collective opinion and the individual judgements) is used in *distance-based jugement aggregation* and in general *distance-based semantics* for decision or choice.

[redacted]

Chapter 12
Distances on Numbers, Polynomials, and Matrices

12.1 Metrics on Numbers

Here we consider the most important metrics on the classical number systems: the semiring $\mathbb{N}$ of natural numbers, the ring $\mathbb{Z}$ of integers, and the fields $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ of rational, real, complex numbers, respectively. We consider also the algebra $\mathcal{Q}$ of quaternions.

- **Metrics on natural numbers**

  There are several well-known metrics on the set $\mathbb{N}$ of natural numbers:

  1. $|n - m|$; the restriction of the **natural metric** (from $\mathbb{R}$) on $\mathbb{N}$;
  2. $p^{-\alpha}$, where α is the highest power of a given prime number p dividing $m - n$, for $m \neq n$ (and equal to 0 for $m = n$); the restriction of the **p-adic metric** (from $\mathbb{Q}$) on $\mathbb{N}$;
  3. $\ln \frac{lcm(m,n)}{gcd(m,n)}$; an example of the **lattice valuation metric**;
  4. $w_r(n - m)$, where $w_r(n)$ is the *arithmetic r-weight* of n; the restriction of the **arithmetic r-norm metric** (from $\mathbb{Z}$) on $\mathbb{N}$;
  5. $\frac{|n-m|}{mn}$ (cf. **M-relative metric** in Chap. 5);
  6. $1 + \frac{1}{m+n}$ for $m \neq n$ (and equal to 0 for $m = n$); the **Sierpinski metric**.

  Most of these metrics on $\mathbb{N}$ can be extended on $\mathbb{Z}$. Moreover, any one of the above metrics can be used in the case of an arbitrary countable set X. For example, the **Sierpinski metric** is defined, in general, on a countable set $X = \{x_n : n \in \mathbb{N}\}$ by $1 + \frac{1}{m+n}$ for all $x_m, x_n \in X$ with $m \neq n$ (and is equal to 0, otherwise).

- **Arithmetic r-norm metric**

  Let $r \in \mathbb{N}$, $r \geq 2$. The *modified r-ary form* of an integer x is a representation

$$x = e_n r^n + \cdots + e_1 r + e_0,$$

  where $e_i \in \mathbb{Z}$, and $|e_i| < r$ for all $i = 0, \ldots, n$.

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_12

An r-ary form is called *minimal* if the number of nonzero coefficients is minimal. The minimal form is not unique, in general. But if the coefficients e_i, $0 \le i \le n-1$, satisfy the conditions $|e_i + e_{i+1}| < r$, and $|e_i| < |e_{i+1}|$ if $e_i e_{i+1} < 0$, then the above form is unique and minimal; it is called the *generalized nonadjacent form*.

The *arithmetic r-weight* $w_r(x)$ of an integer x is the number of nonzero coefficients in a *minimal r-ary form* of x, in particular, in the generalized nonadjacent form. The **arithmetic r-norm metric** on $\mathbb{Z}$ (see, for example, [Ernv85]) is defined by

$$w_r(x-y).$$

- **Distance between consecutive primes**

  The **distance between consecutive primes** (or *prime gap*, *prime difference function*) is the difference $g_n = p_{n+1} - p_n$ between two successive prime numbers.

  It holds $g_n \le p_n$, $\overline{\lim}_{n\to\infty} g_n = \infty$ and (Zhang, 2013) $\underline{\lim}_{n\to\infty} g_n < 7 \times 10^7$, improved to ≤ 246 (conjecturally, to ≤ 6) by Polymath8, 2014. There is no $\lim_{n\to\infty} g_n$ but $g_n \approx \ln p_n$ for the average g_n.

  Open *Polignac's conjecture*: for any $k \ge 1$, there are infinitely many n with $g_n = 2k$; the case $k = 1$ (i.e., that $\underline{\lim}_{n\to\infty} g_n = 2$ holds) is the *twin prime conjecture*.
- **Distance Fibonacci numbers**

  Fibonacci numbers are defined by the recurrence $F_n = F_{n-1} + F_{n-2}$ for $n \ge 2$ with initial terms $F_0 = 0$ and $F_1 = 1$. **Distance Fibonacci numbers** are three following generalizations of them in the distance sense, considered by Wloch et al..

  Kwaśnik–Wloch, 2000: $F(k,n) = F(k,n-1) + F(k,n-k)$ for $n > k$ and $F(k,n) = n+1$ for $n \le k$.

  Bednarz et al., 2012: $Fd(k,n) = Fd(k,n-k+1) + Fd(k,n-k)$ for $n \ge k > 1$ and $Fd(k,n) = 1$ for $0 \le n < k$.

  Wloch et al., 2013: $F_2(k,n) = F_2(k,n-2) + F_2(k,n-k)$ for $n \ge k \ge 1$ and $F_2(k,n) = 1$ for $0 \le n < k$.
- **p-adic metric**

  Let p be a prime number. Any nonzero rational number x can be represented as $x = p^{\alpha}\frac{c}{d}$, where c and d are integers not divisible by p, and α is a unique integer. The *p-adic norm* of x is defined by $|x|_p = p^{-\alpha}$. Moreover, $|0|_p = 0$ is defined.

  The **p-adic metric** is a **norm metric** on the set $\mathbb{Q}$ of rational numbers defined by

$$|x-y|_p.$$

This metric forms the basis for the algebra of p-adic numbers. The **Cauchy completions** of the metric spaces ($\mathbb{Q}$, $|x-y|_p$) and ($\mathbb{Q}$, $|x-y|$) with the **natural metric** $|x-y|$ give the fields $\mathbb{Q}_p$ of *p-adic numbers* and $\mathbb{R}$ of real numbers, respectively.

[redacted]

The **Gajić metric** is an **ultrametric** on the set $\mathbb{Q}$ of rational numbers defined, for $x \neq y$ (via the integer part $\lfloor z \rfloor$ of a real number z), by

$$\inf\{2^{-n} : n \in \mathbb{Z},\ \lfloor 2^n(x-e) \rfloor = \lfloor 2^n(y-e) \rfloor\},$$

where e is any fixed irrational number. This metric is **equivalent** to the **natural metric** $|x-y|$ on $\mathbb{Q}$.

- **Continued fraction metric on irrationals**

The **continued fraction metric on irrationals** is a complete metric on the set *Irr* of irrational numbers defined, for $x \neq y$, by

$$\frac{1}{n},$$

where n is the first index for which the continued fraction expansions of x and y differ. This metric is **equivalent** to the **natural metric** $|x-y|$ on *Irr* which is noncomplete and disconnected. Also, the *Baire* 0*-dimensional space* $B(\aleph_0)$ (cf. **Baire metric** in Chap. 11) is homeomorphic to *Irr* endowed with this metric.

- **Natural metric**

The **natural metric** (or **absolute value metric**, **line metric**, *the distance between numbers*) is a metric on $\mathbb{R}$ defined by

$$|x-y| = \begin{cases} y-x, \text{ if } x-y < 0, \\ x-y, \text{ if } x-y \geq 0. \end{cases}$$

On $\mathbb{R}$ all l_p**-metrics** coincide with the natural metric. The metric space $(\mathbb{R}, |x-y|)$ is called the *real line* (or *Euclidean line*).

There exist many other metrics on $\mathbb{R}$ coming from $|x-y|$ by some **metric transform** (Chap. 4). For example: $\min\{1, |x-y|\}$, $\frac{|x-y|}{1+|x-y|}$, $|x|+|x-y|+|y|$ (for $x \neq y$) and, for a given $0 < \alpha < 1$, the **generalized absolute value metric** $|x-y|^{\alpha}$.

Some authors use $|x-y|$ as the *Polish notation* (parentheses-free and computer-friendly) of the distance function in any metric space.

- **Zero bias metric**

The **zero bias metric** is a metric on $\mathbb{R}$ defined by

$$1+|x-y|$$

if one and only one of x and y is strictly positive, and by

$$|x-y|,$$

otherwise, where $|x-y|$ is the **natural metric** (see, for example, [Gile87]).

- **Sorgenfrey quasi-metric**

  The **Sorgenfrey quasi-metric** is a quasi-metric d on $\mathbb{R}$ defined by

$$y - x$$

if $y \geq x$, and equal to 1, otherwise. Some similar quasi-metrics on $\mathbb{R}$ are:

1. $d_1(x, y) = \max\{y - x, 0\}$ (in general, $\max\{f(y) - f(x), 0\}$ is a quasi-metric on a set X if $f : X \to \mathbb{R}_{\geq 0}$ is an injective function);
2. $d_2(x, y) = \min\{y - x, 1\}$ if $y \geq x$, and equal to 1, otherwise;
3. $d_3(x, y) = y - x$ if $y \geq x$, and equal to $a(x - y)$ (for fixed $a > 0$), otherwise;
4. $d_4(x, y) = e^y - e^x$ if $y \geq x$, and equal to $e^{-y} - e^{-x}$ otherwise.

- **Real half-line quasi-semimetric**

  The **real half-line quasi-semimetric** is defined on the half-line $\mathbb{R}_{>0}$ by

$$\max\{0, \ln \frac{y}{x}\}.$$

- **Janous–Hametner metric**

  The **Janous–Hametner metric** is defined on the half-line $\mathbb{R}_{>0}$ by

$$\frac{|x - y|}{(x + y)^t},$$

where $t = -1$ or $0 \leq t \leq 1$, and $|x - y|$ is the **natural metric**.

- **Extended real line metric**

  An **extended real line metric** is a metric on $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$. The main example (see, for example, [Cops68]) of such metric is given by

$$|f(x) - f(y)|,$$

where $f(x) = \frac{x}{1+|x|}$ for $x \in \mathbb{R}$, $f(+\infty) = 1$, and $f(-\infty) = -1$.

  Another metric, commonly used on $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$, is defined by

$$|\arctan x - \arctan y|,$$

where $-\frac{1}{2}\pi < \arctan x < \frac{1}{2}\pi$ for $-\infty < x < \infty$, and $\arctan(\pm\infty) = \pm\frac{1}{2}\pi$.

- **Complex modulus metric**

  The **complex modulus metric** on the set $\mathbb{C}$ of complex numbers is defined by

$$|z - u|,$$

where, for any $z = z_1 + z_2 i \in \mathbb{C}$, the number $|z| = \sqrt{z\overline{z}} = \sqrt{z_1^2 + z_2^2}$ is the *complex modulus*. The *complex argument* θ is defined by $z = |z|(\cos(\theta) + i \sin(\theta))$.

[redacted]

The metric space $(\mathbb{C}, |z-u|)$ is called the *complex* (or *Wessel–Argand*) *plane*. It is isometric to the Euclidean plane $(\mathbb{R}^2, ||x-y||_2)$. So, the metrics on $\mathbb{R}^2$, given in Chaps. 19 and 5, can be seen as metrics on $\mathbb{C}$. For example, the **British Rail metric** on $\mathbb{C}$ is $|z| + |u|$ for $z \neq u$. The **p-relative** (if $1 \leq p < \infty$) and **relative metric** (if $p = \infty$) on $\mathbb{C}$ are defined for $|z| + |u| \neq 0$ respectively, by

$$\frac{|z-u|}{\sqrt[p]{|z|^p + |u|^p}} \text{ and } \frac{|z-u|}{\max\{|z|, |u|\}}.$$

- **$\mathbb{Z}(\eta_m)$-related norm metrics**

A *Kummer* (or *cyclotomic*) *ring* $\mathbb{Z}(\eta_m)$ is a subring of the ring $\mathbb{C}$ (and an extension of the ring $\mathbb{Z}$), such that each of its elements has the form $\sum_{j=0}^{m-1} a_j \eta_m^j$, where η_m is a primitive m-th root $\exp(\frac{2\pi i}{m})$ of unity, and all a_j are integers.

The *complex modulus* $|z|$ of $z = a + b\eta_m \in \mathbb{C}$ is defined by

$$|z|^2 = z\overline{z} = a^2 + (\eta_m + \overline{\eta_m})ab + b^2 = a^2 + 2ab\cos(\frac{2\pi i}{m}) + b^2.$$

Then $(a+b)^2 = q^2$ for $m = 2$ (or 1), $a^2 + b^2$ for $m = 4$, and $a^2 + ab + b^2$ for $m = 6$ (or 3), i.e., for the ring $\mathbb{Z}$ of usual integers, $\mathbb{Z}(i)$ of *Gaussian integers* and $\mathbb{Z}(\rho)$ of *Eisenstein–Jacobi* (or *EJ*) *integers*.

The set of units of $\mathbb{Z}(\eta_m)$ contain $\eta_m^j, 0 \leq j \leq m-1$; for $m = 5$ and $m \geq 6$, units of infinite order appear also, since $\cos(\frac{2\pi i}{m})$ is irrational. For $m = 2, 4, 6$, the set of units is $\{\pm 1\}$, $\{\pm 1, \pm i\}$, $\{\pm 1, \pm \rho, \pm \rho^2\}$, where $i = \eta_4$ and $\rho = \eta_6 = \frac{1+i\sqrt{3}}{2}$.

The norms $|z| = \sqrt{a^2 + b^2}$ and $||z||_i = |a| + |b|$ for $z = a + bi \in \mathbb{C}$ give rise to the **complex modulus** and **i-Manhattan** metrics on $\mathbb{C}$. They coincide with the Euclidean (l_2-) and Manhattan (l_1-) metrics, respectively, on $\mathbb{R}^2$ seen as the complex plane. The restriction of the i-Manhattan metric on $\mathbb{Z}(i)$ is the path metric of the square grid $\mathbb{Z}^2$ of $\mathbb{R}^2$; cf. **grid metric** in Chap. 19.

The **ρ-Manhattan metric** on $\mathbb{C}$ is defined by the norm $||z||_\rho$, i.e.,

$$\min\{|a|+|b|+|c| : z = a+b\rho+c\rho^2\} = \min\{|a|+|b|, |a+b|+|b|, |a+b|+|a| : z = a+b\rho\}.$$

The restriction of the ρ-Manhattan metric on $\mathbb{Z}(\rho)$ is the path metric of the triangular grid of $\mathbb{R}^2$ (seen as the *hexagonal lattice* $A_2 = \{(a,b,c) \in \mathbb{Z}^3 : a + b + c = 0\}$), i.e., the **hexagonal metric** (Chap. 19).

Let f denote either i or $\rho = \frac{1+i\sqrt{3}}{2}$. Given a $\pi \in \mathbb{Z}(f) \setminus \{0\}$ and $z, z' \in \mathbb{Z}(f)$, we write $z \equiv z' \,(\mathrm{mod}\,\pi)$ if $z - z' = \delta\pi$ for some $\delta \in \mathbb{Z}(f)$. For the quotient ring $\mathbb{Z}_\pi(f) = \{z\,(\mathrm{mod}\,\pi) : z \in \mathbb{Z}(f)\}$, it holds $|\mathbb{Z}_\pi(f)| = ||\pi||_f^2$.

Call two congruence classes $z\,(\mathrm{mod}\,\pi)$ and $z'\,(\mathrm{mod}\,\pi)$ *adjacent* if $z - z' \equiv f^j\,(\mathrm{mod}\,\pi)$ for some j. The resulting graph on $\mathbb{Z}_\pi(f)$ called a *Gaussian network* or *EJ network* if, respectively, $f = i$ or $f = \rho$. The path metrics of these networks coincide with their norm metrics, defined (Fan–Gao, 2004) for $z \,(\mathrm{mod}\, \pi)$ and $z' \,(\mathrm{mod}\, \pi)$, by

$$\min ||u||_f : u \in z - z' \,(\mathrm{mod}\, \pi).$$

These metrics are different from the previously defined ([Hube94a, Hube94b]) distance on $\mathbb{Z}_\pi(f)$: $||v||_f$, where $v \in z - z' \,(\mathrm{mod}\, \pi)$ is selected by minimizing the complex modulus. For $f = i$, this is the **Mannheim distance** (Chap. 16), which is not a metric.

- **Chordal metric**

  The **chordal metric** d_χ is a metric on the set $\overline{\mathbb{C}}$=$\mathbb{C} \cup \{\infty\}$ defined by

$$d_\chi(z, u) = \frac{2|z - u|}{\sqrt{1 + |z|^2}\sqrt{1 + |u|^2}} \text{ and } d_\chi(z, \infty) = \frac{2}{\sqrt{1 + |z|^2}}$$

for all $u, z \in \mathbb{C}$ (cf. ***M*-relative metric** in Chap. 5).

The metric space $(\overline{\mathbb{C}}, d_\chi)$ is called the *extended complex plane*. It is homeomorphic and conformally equivalent to the *Riemann sphere*, i.e., the *unit sphere* $S^2 = \{(x_1, x_2, x_3) \in \mathbb{E}^3 : x_1^2 + x_2^2 + x_3^2 = 1\}$ (considered as a metric subspace of $\mathbb{E}^3$), onto which $(\overline{\mathbb{C}}, d_\chi)$ is one-to-one mapped under stereographic projection.

The plane $\overline{\mathbb{C}}$ can be identified with the plane $x_3 = 0$ such that the and imaginary axes coincide with the x_1 and x_2 axes. Under stereographic projection, each point $z \in \mathbb{C}$ corresponds to the point $(x_1, x_2, x_3) \in S^2$, where the ray drawn from the "north pole" $(0, 0, 1)$ to the point z meets the sphere S^2; the "north pole" corresponds to the point at ∞. The chordal (spherical) metric between two points $p, q \in S^2$ is taken to be the distance between their preimages $z, u \in \overline{\mathbb{C}}$.

The chordal metric is defined equivalently on $\overline{\mathbb{R}}^n = \mathbb{R}^n \cup \{\infty\}$:

$$d_\chi(x, y) = \frac{2||x - y||_2}{\sqrt{1 + ||x||_2^2}\sqrt{1 + ||y||_2^2}} \text{ and } d_\chi(x, \infty) = \frac{2}{\sqrt{1 + ||x||_2^2}}.$$

The restriction of the metric d_χ on $\mathbb{R}^n$ is a **Ptolemaic metric**; cf. Chap. 1.

Given $\alpha > 0, \beta \geq 0, p \geq 1$, the **generalized chordal metric** is a metric on $\mathbb{C}$ (in general, on $(\mathbb{R}^n, ||.||_2)$ and even on any *Ptolemaic space* $(V, ||.||)$), defined by

$$\frac{|z - u|}{\sqrt[p]{\alpha + \beta|z|^p} \cdot \sqrt[p]{\alpha + \beta|u|^p}}.$$

- **Metrics on quaternions**

  Quaternions are members of a noncommutative division algebra $\mathcal{Q}$ over the field $\mathbb{R}$, geometrically realizable in $\mathbb{R}^4$ ([Hami66]). Formally,

$$\mathcal{Q} = \{q = q_1 + q_2 i + q_3 j + q_4 k : q_i \in \mathbb{R}\},$$

where the *basic units* $1, i, j, k \in \mathcal{Q}$ satisfy $i^2 = j^2 = k^2 = -1$ and $ij = -ji = k$.

[redacted]

The *quaternion norm* is defined by $||q|| = \sqrt{q\overline{q}} = \sqrt{q_1^2 + q_2^2 + q_3^2 + q_4^2}$, where $\overline{q} = q_1 - q_2 i - q_3 j - q_4 k$. The **quaternion metric** is the norm metric $||q - q'||$ on $\mathcal{Q}$.

The set of all *Lipschitz integers* and *Hurwitz integers* are defined, respectively, by

$$L = \{q_1 + q_2 i + q_3 j + q_4 k : q_i \in \mathbb{Z}\} \text{ and}$$

$$H = \{q_1 + q_2 i + q_3 j + q_4 k : \text{ all } q_i \in \mathbb{Z} \text{ or all } q_i + \frac{1}{2} \in \mathbb{Z}\}.$$

A quaternion $q \in L$ is *irreducible* (i.e., $q = q'q''$ implies $\{q', q''\} \cap \{\pm 1, \pm i, \pm j, \pm k\} \neq \emptyset$) if and only if $||q||$ is a prime. Given an irreducible $\pi \in L$ and $q, q' \in H$, we write $q \equiv q' \pmod{\pi}$ if $q - q' = \delta\pi$ for some $\delta \in L$.

For the rings $L_\pi = \{q \pmod{\pi} : q \in L\}$ and $H_\pi = \{q \pmod{\pi} : q \in H\}$ it holds $|L_\pi| = ||\pi||^2$ and $|H_\pi| = 2||\pi||^2 - 1$.

The **quaternion Lipschitz metric** on L_π is defined (Martinez et al., 2009) by

$$d_L(\alpha, \beta) = \min \sum_{1 \le s \le 4} |q_s| : \alpha - \beta \equiv q_1 + q_2 i + q_3 j + q_4 k \pmod{\pi}.$$

The ring H is additively generated by its subring L and $w = \frac{1}{2}(1 + i + j + k)$. The **Hurwitz metric** on the ring H_π is defined (Guzëltepe, 2013) by

$$d_H(\alpha, \beta) = \min \sum_{1 \le s \le 5} |q_s| : \alpha - \beta \equiv q_1 + q_2 i + q_3 j + q_4 k + q_5 w \pmod{\pi}.$$

Cf. the **hyper-Kähler** and **Gibbons–Manton** metrics in Sect. 7.3 and the **unit quaternions** and **joint angle** metrics in Sect. 18.3.

12.2 Metrics on Polynomials

A *polynomial* is a sum of powers in one or more variables multiplied by coefficients. A *polynomial in one variable* (or *monic polynomial*) with constant real (complex) coefficients is given by $P = P(z) = \sum_{k=0}^{n} a_k z^k$, $a_k \in \mathbb{R}$ ($a_k \in \mathbb{C}$). The set $\mathcal{P}$ of all real (complex) polynomials forms a ring $(\mathcal{P}, +, \cdot, 0)$. It is also a vector space over $\mathbb{R}$ (over $\mathbb{C}$).

- **Polynomial norm metric**

  A **polynomial norm metric** is a **norm metric** on the vector space $\mathcal{P}$ of all real (complex) polynomials defined by

$$||P - Q||,$$

where $||.||$ is a *polynomial norm*, i.e., a function $||.|| : \mathcal{P} \to \mathbb{R}$ such that, for all $P, Q \in \mathcal{P}$ and for any scalar k, we have the following properties:

1. $||P|| \geq 0$, with $||P|| = 0$ if and only if $P \equiv 0$;
2. $||kP|| = |k| ||P||$;
3. $||P + Q|| \leq ||P|| + ||Q||$ (triangle inequality).

The l_p*-norm* and L_p*-norm* of a polynomial $P(z) = \sum_{k=0}^{n} a_k z^k$ are defined by

$$||P||_p = (\sum_{k=0}^{n} |a_k|^p)^{1/p} \text{ and } ||P||_{L_p} = (\int_0^{2\pi} |P(e^{i\theta})|^p \frac{d\theta}{2\pi})^{\frac{1}{p}} \text{ for } 1 \leq p < \infty,$$

$$||P||_\infty = \max_{0 \leq k \leq n} |a_k| \text{ and } ||P||_{L_\infty} = \sup_{|z|=1} |P(z)| \text{ for } p = \infty.$$

The values $||P||_1$ and $||P||_\infty$ are called the *length* and *height* of polynomial P.

- **Distance from irreducible polynomials**

  For any field $\mathbb{F}$, a polynomial with coefficients in $\mathbb{F}$ is said to be *irreducible over* $\mathbb{F}$ if it cannot be factored into the product of two nonconstant polynomials with coefficients in $\mathbb{F}$. Given a metric d on the polynomials over $\mathbb{F}$, the **distance** (of a given polynomial $P(z)$) **from irreducible polynomials** is $d_{ir}(P) = \inf d(P, Q)$, where $Q(z)$ is any irreducible polynomial of the same degree over $\mathbb{F}$.

  Polynomial conjecture of Turán, 1967, is that there exists a constant C with $d_{ir}(P) \leq C$ for every polynomial P over $\mathbb{Z}$, where $d(P, Q)$ is the *length* $||P - Q||_1$ of $P - Q$.

  Lee–Ruskey–Williams, 2007, conjectured that there exists a constant C with $d_{ir}(P) \leq C$ for every polynomial P over the Galois field $\mathbb{F}_2$, where $d(P, Q)$ is the **Hamming distance** between the $(0, 1)$-sequences of coefficients of P and Q.

- **Bombieri metric**

  The **Bombieri metric** (or **polynomial bracket metric**) is a **polynomial norm metric** on the set $\mathcal{P}$ of all real (complex) polynomials defined by

$$[P - Q]_p,$$

  where $[.]_p$, $0 \leq p \leq \infty$, is the *Bombieri* p*-norm*.

  For a polynomial $P(z) = \sum_{k=0}^{n} a_k z^k$ it is defined by

$$[P]_p = (\sum_{k=0}^{n} \binom{n}{k}^{1-p} |a_k|^p)^{\frac{1}{p}}.$$

- **Metric space of roots**

  The **metric space of roots** is (Ćurgus–Mascioni, 2006) the space (X, d) where X is the family of all multisets of complex numbers with n elements and the distance between multisets $U = \{u_1, \ldots, u_n\}$ and $V = \{v_1, \ldots, v_n\}$ is defined by

[redacted]

the following analog of the **Fréchet metric**:

$$\min_{\tau \in Sym_n} \max_{1 \leq j \leq n} |u_j - v_{\tau(j)}|,$$

where τ is any permutation of $\{1, \dots, n\}$. Here the set of roots of some monic complex polynomial of degree n is considered as a multiset with n elements. Cf. **metrics between multisets** in Chap. 1.

The function assigning to each polynomial the multiset of its roots is a **homeomorphism** between the metric space of all monic complex polynomials of degree n with the **polynomial norm metric** l_∞ and the metric space of roots.

12.3 Metrics on Matrices

An $m \times n$ matrix $A = ((a_{ij}))$ over a field $\mathbb{F}$ is a table consisting of m rows and n columns with the entries a_{ij} from $\mathbb{F}$. The set of all $m \times n$ matrices with real (complex) entries is denoted by $M_{m,n}$ or $\mathbb{R}^{m \times n}$ ($\mathbb{C}^{m \times n}$). It forms a *group* $(M_{m,n}, +, 0_{m,n})$, where $((a_{ij})) + ((b_{ij})) = ((a_{ij} + b_{ij}))$, and the matrix $0_{m,n} \equiv 0$. It is also an mn-dimensional vector space over $\mathbb{R}$ ($\mathbb{C}$).

The *transpose* of a matrix $A = ((a_{ij})) \in M_{m,n}$ is the matrix $A^T = ((a_{ji})) \in M_{n,m}$. A $m \times n$ matrix A is called a *square matrix* if $m = n$, and a *symmetric matrix* if $A = A^T$. The *conjugate transpose* (or *adjoint*) of a matrix $A = ((a_{ij})) \in M_{m,n}$ is the matrix $A^* = ((\overline{a}_{ji})) \in M_{n,m}$. An *Hermitian matrix* is a complex square matrix A with $A = A^*$.

The set of all square $n \times n$ matrices with real (complex) entries is denoted by M_n. It forms a *ring* $(M_n, +, \cdot, 0_n)$, where $+$ and 0_n are defined as above, and $((a_{ij})) \cdot ((b_{ij})) = ((\sum_{k=1}^{n} a_{ik} b_{kj}))$. It is also an n^2-dimensional vector space over $\mathbb{R}$ (over $\mathbb{C}$). The *trace* of a square $n \times n$ matrix $A = ((a_{ij}))$ is defined by $\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{ii}$.

The *identity matrix* is $1_n = ((c_{ij}))$ with $c_{ii} = 1$, and $c_{ij} = 0$, $i \neq j$. An *unitary matrix* $U = ((u_{ij}))$ is a square matrix defined by $U^{-1} = U^*$, where U^{-1} is the *inverse matrix* of U, i.e., $UU^{-1} = 1_n$. A matrix $A \in M_{m,n}$ is *orthonormal* if $A^* A = 1_n$. A matrix $A \in \mathbb{R}^{n \times n}$ is *orthogonal* if $A^T = A^{-1}$, *normal* if $A^T A = AA^T$ and *singular* if its determinant is 0.

If for a matrix $A \in M_n$ there is a vector x such that $Ax = \lambda x$ for some scalar λ, then λ is called an *eigenvalue* of A with corresponding *eigenvector* x. Given a matrix $A \in \mathbb{C}^{m \times n}$, its *singular values* $s_i(A)$ are defined as $\sqrt{\lambda(A^* A)}$. A real matrix A is *positive-definite* if $v^T A v > 0$ for all nonzero real vectors v; it holds if and only if all eigenvalues of $A_H = \frac{1}{2}(A + A^T)$ are positive. An Hermitian matrix A is *positive-definite* if $v^* A v > 0$ for all nonzero complex vectors v; it holds if and only if all $\lambda(A)$ are positive.

The *mixed states* of a n-dimensional *quantum system* are described by their *density matrices*, i.e., positive-semidefinite Hermitian $n \times n$ matrices of trace 1. The set of such matrices is convex, and its extremal points describe the *pure states*. Cf. **monotone metrics** in Chap. 7 and **distances between quantum states** in Chap. 24.

- **Matrix norm metric**

  A **matrix norm metric** is a **norm metric** on the set $M_{m,n}$ of all real (complex) $m \times n$ matrices defined by

$$||A - B||,$$

  where $||.||$ is a *matrix norm*, i.e., a function $||.|| : M_{m,n} \to \mathbb{R}$ such that, for all $A, B \in M_{m,n}$, and for any scalar k, we have the following properties:

  1. $||A|| \geq 0$, with $||A|| = 0$ if and only if $A = 0_{m,n}$;
  2. $||kA|| = |k| ||A||$;
  3. $||A + B|| \leq ||A|| + ||B||$ (triangle inequality).
  4. $||AB|| \leq ||A|| \cdot ||B||$ (*submultiplicativity*).

  All matrix norm metrics on $M_{m,n}$ are equivalent. The simplest example of such metric is the **Hamming metric** on $M_{m,n}$ (in general, on the set $M_{m,n}(\mathbb{F})$ of all $m \times n$ matrices with entries from a field $\mathbb{F}$) defined by $||A - B||_H$, where $||A||_H$ is the *Hamming norm* of $A \in M_{m,n}$, i.e., the number of nonzero entries in A. Example of a *generalized* (i.e., not submultiplicative one) *matrix norm* is the *max element norm* $||A = ((a_{ij}))||\mathrm{max} = \max_{i,j} |a_{ij}|$; but $\sqrt{mn}||A||_{\max}$ is a matrix norm.

- **Natural norm metric**

  A **natural** (or *operator*, *induced*) **norm metric** is a **matrix norm metric** on the set M_n defined by

$$||A - B||_{\mathrm{nat}},$$

  where $||.||_{\mathrm{nat}}$ is a *natural* (or *operator*, *induced*) *norm* on M_n, induced by the vector norm $||x||$, $x \in \mathbb{R}^n$ ($x \in \mathbb{C}^n$), is a matrix norm defined by

$$||A||_{\mathrm{nat}} = \sup_{||x|| \neq 0} \frac{||Ax||}{||x||} = \sup_{||x||=1} ||Ax|| = \sup_{||x|| \leq 1} ||Ax||.$$

  The natural norm metric can be defined in similar way on the set $M_{m,n}$ of all $m \times n$ real (complex) matrices: given vector norms $||.||_{\mathbb{R}^m}$ on $\mathbb{R}^m$ and $||.||_{\mathbb{R}^n}$ on $\mathbb{R}^n$, the *natural norm* $||A||_{\mathrm{nat}}$ of a matrix $A \in M_{m,n}$, induced by $||.||_{\mathbb{R}^n}$ and $||.||_{\mathbb{R}^m}$, is a matrix norm defined by $||A||_{\mathrm{nat}} = \sup_{||x||_{\mathbb{R}^n}=1} ||Ax||_{\mathbb{R}^m}$.

- **Matrix p-norm metric**

  A **matrix p-norm metric** is a **natural norm metric** on M_n defined by

$$||A - B||^p_{\mathrm{nat}},$$

  where $||.||^p_{\mathrm{nat}}$ is the *matrix* (or *operator*) *p-norm*, i.e., a *natural norm*, induced by the vector *l_p-norm*, $1 \leq p \leq \infty$:

$$||A||^p_{\mathrm{nat}} = \max_{||x||_p=1} ||Ax||_p, \quad \text{where} \quad ||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}.$$

[redacted]

The **maximum absolute column** and **maximum absolute row metric** are the **matrix** 1**-norm** and **matrix** ∞**-norm metric** on M_n. For a matrix $A = ((a_{ij})) \in M_n$, the *maximum absolute column* and *maximum absolute row sum norm* are

$$||A||^1_{\text{nat}} = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}| \text{ and } |A||^{\infty}_{\text{nat}} = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|.$$

The **spectral norm metric** is the **matrix** 2**-norm metric** $||A - B||^2_{\text{nat}}$ on M_n. The matrix 2-norm $||.||^2_{\text{nat}}$, induced by the vector l_2*-norm*, is also called the *spectral norm* and denoted by $||.||_{sp}$. For a symmetric matrix $A = ((a_{ij})) \in M_n$, it is

$$||A||_{sp} = s_{\max}(A) = \sqrt{\lambda_{\max}(A^*A)},$$

where $A^* = ((\overline{a}_{ji}))$, while $s_{\max}$ and $\lambda_{\max}$ are largest singular value and eigenvalue.

- **Frobenius norm metric**

The **Frobenius norm metric** is a **matrix norm metric** on $M_{m,n}$ defined by

$$||A - B||_{Fr},$$

where $||.||_{Fr}$ is the *Frobenius* (or *Hilbert–Schmidt*) *norm*. For $A = ((a_{ij}))$, it is

$$||A||_{Fr} = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\text{Tr}(A^*A)} = \sqrt{\sum_{1 \le i \le \text{rank}(A)} \lambda_i} = \sqrt{\sum_{1 \le i \le \text{rank}(A)} s_i^2},$$

where λ_i, s_i are the eigenvalues and singular values of A.

This norm is strictly convex, is a differentiable function of its elements a_{ij} and is the only unitarily invariant norm among $||A||_p = (\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^p)^{\frac{1}{p}}, p \ge 1$.

The **trace norm metric** is a matrix norm metric on $M_{m,n}$ defined by

$$||A - B||_{tr},$$

where $||.||_{tr}$ is the *trace norm* (or *nuclear norm*) on $M_{m,n}$ defined by

$$||A||_{tr} = \sum_{i=1}^{\min\{m,n\}} s_i(A) = \text{Tr}(\sqrt{A^*A}).$$

- **Schatten norm metric**

Given $1 \le p < \infty$, the **Schatten norm metric** is a **matrix norm metric** on $M_{m,n}$ defined by

$$||A - B||^p_{Sch},$$

where $||.||^p_{Sch}$ is the *Schatten p-norm* on $M_{m,n}$. For a matrix $A \in M_{m,n}$, it is defined as the p-th root of the sum of the p-th powers of all its *singular values*:

$$||A||^p_{Sch} = (\sum_{i=1}^{\min\{m,n\}} s_i^p(A))^{\frac{1}{p}}.$$

For $p = \infty$, 2 and 1, one obtains the **spectral norm metric**, **Frobenius norm metric** and **trace norm metric**, respectively.

- **(c, p)-norm metric**

  Let $k \in \mathbb{N}$, $k \leq \min\{m, n\}$, $c \in \mathbb{R}^k$, $c_1 \geq c_2 \geq \cdots \geq c_k > 0$, and $1 \leq p < \infty$. The (c, p)**-norm metric** is a **matrix norm metric** on $M_{m,n}$ defined by

$$||A - B||^k_{(c,p)},$$

where $||.||^k_{(c,p)}$ is the (c, p)*-norm* on $M_{m,n}$. For a matrix $A \in M_{m,n}$, it is defined by

$$||A||^k_{(c,p)} = (\sum_{i=1}^{k} c_i s_i^p(A))^{\frac{1}{p}},$$

where $s_1(A) \geq s_2(A) \geq \cdots \geq s_k(A)$ are the first k *singular values* of A.

If $p = 1$, it is the *c-norm*. If, moreover, $c_1 = \cdots = c_k = 1$, it is the *Ky Fan k-norm*.

- **Ky Fan k-norm metric**

  Given $k \in \mathbb{N}$, $k \leq \min\{m, n\}$, the **Ky Fan k-norm metric** is a **matrix norm metric** on $M_{m,n}$ defined by

$$||A - B||^k_{KF},$$

where $||.||^k_{KF}$ is the *Ky Fan k-norm* on $M_{m,n}$. For a matrix $A \in M_{m,n}$, it is defined as the sum of its first k *singular values*:

$$||A||^k_{KF} = \sum_{i=1}^{k} s_i(A).$$

For $k = 1$ and $k = \min\{m, n\}$, one obtains the **spectral** and **trace** norm metrics.

- **Cut norm metric**

  The **cut norm metric** is a **matrix norm metric** on $M_{m,n}$ defined by

$$||A - B||_{cut},$$

[redacted]

where $||.||_{cut}$ is the *cut norm* on $M_{m,n}$ defined, for a matrix $A = ((a_{ij})) \in M_{m,n}$, as:

$$||A||_{cut} = \max_{I \subset \{1,\dots,m\}, J \subset \{1,\dots,n\}} |\sum_{i \in I, j \in J} a_{ij}|.$$

Cf. in Chap. 15 the **rectangle distance on weighted graphs** and the **cut semimetric**, but the **weighted cut metric** in Chap. 19 is not related.

- **Matrix nearness problems**

  A norm $||.||$ is *unitarily invariant* on $M_{m,n}$ if $||B|| = ||UBV||$ for all $B \in M_{m,n}$ and all unitary matrices U, V. All *Schatten p-norms* are unitarily invariant.

  Given a unitarily invariant norm $||.||$ on $M_{m,n}$, a matrix property $\mathcal{P}$ defining a subspace or compact subset of $M_{m,n}$ (so that $d_{||.||}(A, \mathcal{P})$ below is well defined) and a matrix $A \in M_{m,n}$, then the *distance to* $\mathcal{P}$ is the **point-set distance** on $M_{m,n}$

  $$d(A) = d_{||.||}(A, \mathcal{P}) = \min\{||E|| : A + E \text{ has property } \mathcal{P}\}.$$

  A **matrix nearness problem** is ([High89]) to find an explicit formula for $d(A)$, the *$\mathcal{P}$-closest matrix* (or matrices) $X_{||.||}(A) = A + E$, satisfying the above minimum, and efficient algorithms for computing $d(A)$ and $X_{||.||}(A)$. The *componentwise nearness problem* is to find $d'(A) = \min\{\epsilon : |E| \leq \epsilon|A|, A + E \text{ has property } \mathcal{P}\}$, where $|B| = ((|b_{ij}|))$ and the matrix inequality is interpreted componentwise.

  The most used norms for $B = ((b_{ij}))$ are the *Schatten* 2- and ∞-*norms* (cf. **Schatten norm metric**): the *Frobenius norm* $||B||_{Fr} = \sqrt{\mathrm{Tr}(B^*B)} = \sqrt{\sum_{1 \leq i \leq \mathrm{rank}(B)} s_i^2}$ and the *spectral norm* $||B||_{sp} = \sqrt{\lambda_{\max}(B^*B)} = s_1(B)$.

  Examples of closest matrices $X = X_{||.||}(A, \mathcal{P})$ follow.

  Let $A \in \mathbb{C}^{n \times n}$. Then $A = A_H + A_S$, where $A_H = \frac{1}{2}(A + A^*)$ is Hermitian and $A_H = \frac{1}{2}(A - A^*)$ is *skew-Hermitian* (i.e., $A_H^* = -A_H$). Let $A = U\Sigma V^*$ be a *singular value decomposition* (SVD) of A, i.e., $U \in M_m$ and $V^* \in M_n$ are unitary, while $\Sigma = \mathrm{diag}(s_1, s_2, \dots, s_{\min\{m,n\}})$ is an $m \times n$ diagonal matrix with $s_1 \geq s_2 \geq \dots \geq s_{\mathrm{rank}(A)} > 0 = \dots = 0$. Fan and Hoffman, 1955, showed that, for any unitarily invariant norm, A_H, A_S, UV^* are closest Hermitian (symmetric), skew-Hermitian (skew-symmetric) and unitary (orthogonal) matrices, respectively. Such matrix $X_{Fr}(A)$ is a unique minimizer in all three cases.

  Let $A \in \mathbb{R}^{n \times n}$. Gabriel, 1979, found the closest normal matrix $X_{Fr}(A)$. Higham found in 1988 a unique closest symmetric positive-semidefinite matrix $X_{Fr}(A)$ and, in 2001, the closest matrix of this type with unit diagonal (i.e., ab correlation matrix).

  Given a SVD $A = U\Sigma V^*$ of A, let A_k denote $U\Sigma_k V^*$, where Σ_k is a diagonal matrix $\mathrm{diag}(s_1, s_2, \dots, s_k, 0, \dots, 0)$ containing the largest k singular values of A. Then (Mirsky, 1960) A_k achieves $\min_{\mathrm{rank}(A+E) \leq k} ||E||$ for any unitarily invariant norm. So, $||A - A_k||_{Fr} = \sqrt{\sum_{i=k+1}^{\mathrm{rank}(A)} s_i^2}$ (Eckart–Young, 1936) and $||A - A_k||_{sp} = s_{max}(A - A_k) = s_{k+1}(A)$. A_k is a unique minimizer $X_{Fr}(A)$ if $s_k > s_{k+1}$.

Let $A \in \mathbb{R}^{n \times n}$ be nonsingular. Then its **distance to singularity** $d(A, Sing) = \min\{||E|| : A + E \text{ is singular}\}$ is, for both above norms, $s_n(A) = \frac{1}{s_1(A^{-1})} = \frac{1}{||A^{-1}||_{sp}} = \sup\{\delta : \delta\mathbb{B}_{\mathbb{R}^n} \subseteq A\mathbb{B}_{\mathbb{R}^n}\}$; here $\mathbb{B}_{\mathbb{R}^n} = \{x \in \mathbb{R}^n : ||x|| \leq 1\}$.

Given a closed convex cone $C \subseteq \mathbb{R}^n$, call a matrix $A \in \mathbb{R}^{m \times n}$ *feasible* if $\{Ax : x \in C\} = \mathbb{R}^m$; so, for $m = n$ and $C = \mathbb{R}^n$, feasibly means nonsingularity. Renegar, 1995, showed that, for feasible matrix A, its **distance to infeasibility** $\min\{||E||_{\text{nat}} : A + E \text{ is not feasible}\}$ is $\sup\{\delta : \delta\mathbb{B}_{\mathbb{R}^m} \subseteq A(\mathbb{B}_{\mathbb{R}^n} \cap C)\}$.

Lewis, 2003, generalized this by showing that, given two real normed spaces X, Y and a surjective *convex process* (or *set valued sublinear mapping*) F from X to Y, i.e., a multifunction for which $\{(x, y) : y \in F(x)\}$ is a closed convex cone, it holds

$$\min\{||E||_{\text{nat}} : E \text{ is any linear map } X \to Y, F + E \text{ is not surjective}\} = \frac{1}{||F^{-1}||_{\text{nat}}}.$$

Donchev et al. 2002, extended this, computing **distance to irregularity**; cf. **metric regularity** (Chap. 1). Cf. the above four *distances to ill-posedness* with **distance to uncontrollability** (Chap. 18) and **distances from symmetry** (Chap. 21).

- **$Sym(n, \mathbb{R})^+$ and $Her(n, \mathbb{C})^+$ metrics**

  Let $Sym(n, \mathbb{R})^+$ and $Her(n, \mathbb{C})^+$ be the cones of $n \times n$ symmetric real and Hermitian complex positive-definite $n \times n$ matrices. The $Sym(n, \mathbb{R})^+$ **metric** is defined, for any $A, B \in Sym(n, \mathbb{R})^+$, as

$$(\sum_{i=1}^{n} \log^2 \lambda_i)^{\frac{1}{2}},$$

where λ_1, c, λ_n are the *eigenvalues* of the matrix $A^{-1}B$ (the same as those of $A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$). It is the **Riemannian distance**, arising from the Riemannian metric $ds^2 = \text{Tr}((A^{-1}(dA))^2)$. This metric was rediscovered in Förstner–Moonen, 1999, and Pennec et al., 2004, via *generalized eigenvalue problem*: $det(\lambda A - B) = 0$.

The $Her(n, \mathbb{C})^+$ **metric** is defined, for any $A, B \in Her(n, \mathbb{C})^+$, by

$$d_R(A, B) = || \log(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})||_{Fr},$$

where $||H||_{Fr} = (\sum_{i,j} |h_{ij}|^2)^{\frac{1}{2}}$ is the *Frobenius norm* of the matrix $H = ((h_{ij}))$. It is the **Riemannian distance** arising from the Riemannian metric of nonpositive curvature, defined locally (at H) by $ds = ||H^{-\frac{1}{2}}\, dH\, H^{-\frac{1}{2}}||_{Fr}$. In other words, this distance is the **geodesic distance**

$$\inf\{L(\gamma) : \gamma \text{ is a (differentiable) path from A to B}\},$$

[redacted]

where $L(\gamma) = \int_A^B ||\gamma^{-\frac{1}{2}}(t)\gamma'(t)\gamma^{-\frac{1}{2}}(t)||_{Fr}dt$ and the geodesic $[A, B]$ is parametrized by $\gamma(t) = A^{\frac{1}{2}}(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})^t A^{\frac{1}{2}}$ in the sense that $d_R(A, \gamma(t)) = td_R(A, B)$ for each $t \in [0, 1]$. In particular, the geodesic midpoint $\gamma(\frac{1}{2})$ of $[A, B]$ can be seen as the *geometric mean* of two positive-definite matrices A and B.

The space $(Her(n, \mathbb{C})^+, d_R))$ is an **Hadamard** (i.e., complete and CAT(0)) **space**, cf. Chap. 6. But $Her(n, \mathbb{C})^+$ is not complete with respect to matrix norms; it has a boundary consisting of the singular positive-semidefinite matrices.

Above $Sym(n, \mathbb{R})^+$ and $Her(n, \mathbb{C})^+$ metrics are the special cases of the distance $d_R(x, y)$ among **invariant distances on symmetric cones** in Chap. 9.

Cf. also, in Chap. 24, the **trace distance** on all Hermitian of trace 1 positive-definite $n \times n$ matrices and in Chap. 7, the **Wigner–Yanase–Dyson metrics** on all complex positive-definite $n \times n$ matrices.

The **Bartlett distance** between two matrices $A, B \in Her(n, \mathbb{C})^+$, is defined (Conradsen et al., 2003, for radar applications) by

$$\ln\left(\frac{(det(A+B))^2}{4det(A)det(B)}\right).$$

- **Siegel distance**

The *Siegel half-plane* is the set SH_n of $n \times n$ matrices $Z = X + iY$, where X, Y are symmetric or Hermitian and Y is positive-definite. The **Siegel–Hua metric** (Siegel, 1943, and independently, Hua, 1944) on SH_n is defined by

$$ds^2 = \mathrm{Tr}(Y^{-1}(dZ)Y^{-1}(d\overline{Z})).$$

It is unique metric preserved by any automorphism of SH_n. The Siegel–Hua metric on the *Siegel disk* $SD_n = \{W = (Z - iI)(Z + iI)^{-1} : Z \in SH_n\}$ is defined by

$$ds^2 = \mathrm{Tr}((I - WW^*)^{-1}dW(I - W^*W)^{-1}dW^*).$$

For n=1, the Siegel–Hua metric is the **Poincaré metric** (cf. Chap. 6) on the *Poincaré half-plane* SH_1 and the *Poincaré disk* SD_1, respectively.

Let $A_n = \{Z = iY : Y > 0\}$ be the imaginary axe on the Siegel half-plane. The Siegel–Hua metric on A_n is (cf. [Barb12]) the Riemannian **trace metric** $ds^2 = \mathrm{Tr}((P^1 dP)^2)$. The corresponding distances are $Sym(n, \mathbb{R})^+$ **metric** or $Her(n, \mathbb{C})^+$ **metric**. The **Siegel distance** on $SH_n \setminus A_n$ is defined by

$$d^2_{Siegel}(Z_1, Z_2) = \sum_{i=1}^{n} \log^2\left(\frac{1 + \sqrt{\lambda_i}}{1 - \sqrt{\lambda_i}}\right);$$

$\lambda_1, \ldots, \lambda_n$ are the *eigenvalues* of the matrix $(Z_1 - Z_2)(Z_1 - \overline{Z_2}) - 1(\overline{Z_1} - \overline{Z_2})(\overline{Z_1} - Z_2)^{-1}$.

- **Barbaresco metrics**

  Let $z(k)$ be a complex temporal (discrete time) *stationary* signal, i.e., its mean value is constant and its *covariance function* $\mathbb{E}[z(k_1)z^*(k_2)]$ is only a function of $k_1 - k_2$. Such signal can be represented by its covariance $n \times n$ matrix $R = ((r_{ij}))$, where $r_{ij} = \mathbb{E}[z(i), z*(j)] = \mathbb{E}[z(n)z*(n-i+j)]$. It is a positive-definite *Toeplitz* (i.e. diagonal-constant) Hermitian matrix. In radar applications, such matrices represent the Doppler spectra of the signal. Matrices R admit a parametrization (complex ARM, i.e., m-th order autoregressive model) by *partial autocorrelation coefficients* defined recursively as the complex correlation between the forward and backward prediction errors of the $(m-1)$-th order complex ARM.

  Barbaresco ([Barb12]) defined, via this parametrization, a **Bergman metric** (Chap. 7) on the bounded domain $\mathbb{R} + xD_n \subset \mathbb{C}^n$ of above matrices R; here D is a *Poincaré disk*. He also defined a related **Kähler metric** on $M \times S_n$, where M is the set of positive-definite Hermitian matrices and SD_n is the *Siegel disk* (cf. **Siegel distance**). Such matrices represent spatiotemporal stationary signals, i.e., in radar applications, the Doppler spectra and spatial directions of the signal.

  Ben Jeuris, 2015, extended above metrics on *block Toeplitz matrices*, i.e., those having blocks that are repeated (as elements of a Toeplitz matrix) down the diagonals of the matrix.

  Cf. **Ruppeiner metric** (Chap. 7) and **Martin cepstrum distance** (Chap. 21).

- **Distances between graphs of matrices**

  The *graph $G(A)$ of a complex $m \times n$ matrix A* is the *range* (i.e., the span of columns) of the matrix $R(A) = ([IA^T])^T$. So, $G(A)$ is a subspace of $\mathbb{C}^{m+n}$ of all vectors v, for which the equation $R(A)x = v$ has a solution.

  A **distance between graphs of matrices** A and B is a distance between the subspaces $G(A)$ and $G(B)$. It can be an **angle distance between subspaces** or, for example, the following distance (cf. also the **Kadets distance** in Chap. 1 and the **gap metric** in Chap. 18).

  The **spherical gap distance** between subspaces A and B is defined by

  $$\max\{\max_{x \in S(A)} d_E(x, S(B)), \max_{y \in S(B)} d_E(y, S(A))\},$$

  where $S(A), S(B)$ are the unit spheres of the subspaces A, B, $d(z, C)$ is the **point-set distance** $\inf_{y \in C} d(z, y)$ and $d_E(z, y)$ is the Euclidean distance.

- **Angle distances between subspaces**

  Consider the *Grassmannian space $G(m, n)$* of all n-dimensional subspaces of Euclidean space $\mathbb{E}^m$; it is a compact *Riemannian manifold* of dimension $n(m-n)$.

  Given two subspaces $A, B \in G(m, n)$, the *principal angles* $\frac{\pi}{2} \geq \theta_1 \geq \cdots \geq \theta_n \geq 0$ between them are defined, for $k = 1, \dots, n$, inductively by

  $$\cos\theta_k = \max_{x \in A} \max_{y \in B} x^T y = (x^k)^T y^k$$

  subject to the conditions $||x||_2 = ||y||_2 = 1, x^T x^i = 0, y^T y^i = 0$, for $1 \leq i \leq k-1$, where $||.||_2$ is the Euclidean norm.

[redacted]

The principal angles can also be defined in terms of orthonormal matrices Q_A and Q_B spanning subspaces A and B, respectively: in fact, n ordered singular values of the matrix $Q_A Q_B \in M_n$ can be expressed as cosines $\cos\theta_1, \dots, \cos\theta_n$.

The **Grassmann distance** between subspaces A and B of the same dimension is their geodesic distance defined by

$$\sqrt{\sum_{i=1}^{n} \theta_i^2}.$$

The **Martin distance** between subspaces A and B is defined by

$$\sqrt{\ln \prod_{i=1}^{n} \frac{1}{\cos^2\theta_i}}.$$

In the case when the subspaces represent ARMs (*autoregressive models*), the Martin distance can be expressed in terms of the *cepstrum* of the autocorrelation functions of the models. Cf. the **Martin cepstrum distance** in Chap. 21.

The **Asimov distance** between subspaces A and B is defined by θ_1. The **spectral distance** (or *chordal* 2*-norm distance*) is defined by $2\sin(\frac{\theta_1}{2})$.

The **containment gap distance** (or *projection distance*) is $\sin\theta_1$. It is the l_2-*norm* of the difference of the *orthogonal projectors* onto A and B. Many versions of this distance are used in Control Theory (cf. **gap metric** in Chap. 18).

The **Frobenius distance** and **chordal distance** between subspaces A and B are

$$\sqrt{2\sum_{i=1}^{n}\sin^2\theta_i} \quad\text{and}\quad \sqrt{\sum_{i=1}^{n}\sin^2\theta_i}, \quad \text{respectively.}$$

It is the *Frobenius norm* of the difference of above projectors onto A and B.

Similar distances $\sqrt{1-\prod_{i=1}^{n}\cos^2\theta_i}$ and $\arccos(\prod_{i=1}^{n}\cos\theta_i)$ are called the **Binet–Cauchy distance** and (cf. Chap. 7) **Fubini–Study distance**, respectively.

- **Larsson–Villani metric**

  Let A and B be two arbitrary orthonormal $m \times n$ matrices of full rank, and let θ_{ij} be the angle between the i-th column of A and the j-th column of B.

  We call **Larsson–Villani metric** the distance between A and B (used by Larsson and Villani, 2000, for multivariate models) the square of which is defined by

$$n - \sum_{i=1}^{n}\sum_{j=1}^{n}\cos^2\theta_{ij}.$$

The square of usual Euclidean distance between A and B is $2(1-\sum_{i=1}^{n}\cos\theta_{ii})$. For $n = 1$, above two distances are $\sin\theta$ and $\sqrt{2(1-\cos\theta)}$, respectively.

- **Lerman metric**

  Given a finite set X and real symmetric $|X| \times |X|$ matrices $((d_1(x, y)))$, $((d_2(x, y)))$ with $x, y \in X$, their **Lerman semimetric** (cf. **Kendall τ distance** on permutations in Chap. 11) is defined by

$$|\{(\{x,y\},\{u,v\}) : (d_1(x,y) - d_1(u,v))(d_2(x,y) - d_2(u,v)) < 0\}| \binom{|X|+1}{2}^{-2},$$

  where $(\{x, y\}, \{u, v\})$ is any pair of unordered pairs of elements x, y, u, v from X.

  Similar **Kaufman semimetric** between $((d_1(x, y)))$ and $((d_2(x, y)))$ is

$$\frac{|\{(\{x,y\},\{u,v\}) : (d_1(x,y) - d_1(u,v))(d_2(x,y) - d_2(u,v)) < 0\}|}{|\{(\{x,y\},\{u,v\}) : (d_1(x,y) - d_1(u,v))(d_2(x,y) - d_2(u,v)) \neq 0\}|}.$$

[redacted]

Chapter 13
Distances in Functional Analysis

Functional Analysis is the branch of Mathematics concerned with the study of spaces of functions. This usage of the word *functional* goes back to the calculus of variations which studies functions whose argument is a function. In the modern view, Functional Analysis is seen as the study of complete *normed vector spaces*, i.e., **Banach spaces**.

For any real number $p \geq 1$, an example of a Banach space is given by **L_p-space** of all Lebesgue-measurable functions whose absolute value's p-th power has finite integral.

A **Hilbert space** is a Banach space in which the norm arises from an *inner product.* Also, in Functional Analysis are considered *continuous linear operators* defined on Banach and Hilbert spaces.

13.1 Metrics on Function Spaces

Let $I \subset \mathbb{R}$ be an *open interval* (i.e., a nonempty connected open set) in $\mathbb{R}$. A real function $f : I \to \mathbb{R}$ is called *real analytic* on I if it agrees with its *Taylor series* in an *open neighborhood* U_{x_0} of every point $x_0 \in I$: $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$ for any $x \in U_{x_0}$. Let $D \subset \mathbb{C}$ be a *domain* (i.e., a *convex* open set) in $\mathbb{C}$.

A complex function $f : D \to \mathbb{C}$ is called *complex analytic* (or, simply, *analytic*) on D if it agrees with its Taylor series in an open neighborhood of every point $z_0 \in D$. A complex function f is analytic on D if and only if it is *holomorphic* on D, i.e., if it has a complex derivative $f'(z_0) = \lim_{z \to z_0} \frac{f(z)-f(z_0)}{z-z_0}$ at every point $z_0 \in D$.

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_13

- **Integral metric**

  The **integral metric** is the L_1*-metric* on the set $C_{[a,b]}$ of all continuous real (complex) functions on a given segment $[a,b]$ defined by

  $$\int_a^b |f(x) - g(x)|dx.$$

  The corresponding metric space is abbreviated by $C^1_{[a,b]}$. It is a Banach space.

  In general, for any **compact** topological space X, the integral metric is defined on the set of all continuous functions $f : X \to \mathbb{R}$ $(\mathbb{C})$ by $\int_X |f(x) - g(x)|dx$.

- **Uniform metric**

  The **uniform metric** (or **sup metric**) is the L_∞**-metric** on the set $C_{[a,b]}$ of all real (complex) continuous functions on a given segment $[a,b]$ defined by

  $$\sup_{x\in[a,b]} |f(x) - g(x)|.$$

  The corresponding metric space is abbreviated by $C^\infty_{[a,b]}$. It is a Banach space.

  A generalization of $C^\infty_{[a,b]}$ is the *space of continuous functions* $C(X)$, i.e., a metric space on the set of all continuous (more generally, bounded) functions $f : X \to \mathbb{C}$ of a topological space X with the L_∞-metric $\sup_{x\in X} |f(x) - g(x)|$.

  In the case of the metric space $C(X,Y)$ of continuous (more generally, bounded) functions $f : X \to Y$ from one **metric compactum** (X, d_X) to another (Y, d_Y), the sup metric between two functions $f, g \in C(X,Y)$ is defined by $\sup_{x\in X} d_Y(f(x), g(x))$.

  The metric space $C^\infty_{[a,b]}$, as well as the metric space $C^1_{[a,b]}$, are two of the most important cases of the metric space $C^p_{[a,b]}$, $1 \leq p \leq \infty$, on the set $C_{[a,b]}$ with the L_p-metric $(\int_a^b |f(x) - g(x)|^p dx)^{\frac{1}{p}}$. The space $C^p_{[a,b]}$ is an example of an L_p*-space*.

- **Dogkeeper distance**

  Given a metric space (X, d), the **dogkeeper distance** is a metric on the set of all functions $f : [0,1] \to X$, defined by

  $$\inf_\sigma \sup_{t\in[0,1]} d(f(t), g(\sigma(t))),$$

  where $\sigma : [0,1] \to [0,1]$ is a continuous, monotone increasing function such that $\sigma(0) = 0, \sigma(1) = 1$. This metric is a special case of the **Fréchet metric**.

  For the case, when (X, d) is Euclidean space $\mathbb{R}^n$, this metric is the original (1906) **Fréchet distance** between parametric curves $f, g : [0,1] \to \mathbb{R}^n$. This distance can be seen as the length of the shortest leash that is sufficient for the man and the dog to walk their paths f and g from start to end. For example, the Fréchet distance between two concentric circles of radius r_1 and r_2 is $|r_1 - r_2|$.

[redacted]

The **discrete Fréchet distance** (or *coupling distance*, Eiter and Mannila, 1994) is an approximation of the Fréchet metric for polygonal curves f and g. It considers only positions of the leash where its endpoints are located at vertices of f and g. So, this distance is the minimum, over all order-preserving pairings of vertices in f and g, of the maximal Euclidean distance between paired vertices.

If the two curves are embedded in a more general metric space (say, a polyhedral terrain or some Euclidean space with obstacles), the distance between two points on them is most naturally defined as the length of the shortest path. The resulting **geodesic Fréchet distance** allows the leash to switch discontinuously.

- **Bohr metric**

  Let $\mathbb{R}$ be a metric space with a metric ρ. A continuous function $f : \mathbb{R} \to \mathbb{R}$ is called *almost periodic* if, for every $\epsilon > 0$, there exists $l = l(\epsilon) > 0$ such that every interval $[t_0, t_0 + l(\epsilon)]$ contains at least one number τ for which $\rho(f(t), f(t+\tau)) < \epsilon$ for $-\infty < t < +\infty$.

  The **Bohr metric** is the **norm metric** $||f - g||$ on the set AP of all almost periodic functions defined by the norm

$$||f|| = \sup_{-\infty < t < +\infty} |f(t)|.$$

  It makes AP a Banach space. Some generalizations of almost periodic functions were obtained using other norms; cf. **Stepanov distance**, **Weyl distance**, **Besicovitch distance** and **Bochner metric**.

- **Stepanov distance**

  The **Stepanov distance** is a distance on the set of all measurable functions $f : \mathbb{R} \to \mathbb{C}$ with summable p-th power on each bounded integral, defined by

$$\sup_{x \in \mathbb{R}} \left(\frac{1}{l} \int_{x}^{x+l} |f(x) - g(x)|^p dx \right)^{1/p}.$$

  The **Weyl distance** is a distance on the same set defined by

$$\lim_{l \to \infty} \sup_{x \in \mathbb{R}} \left(\frac{1}{l} \int_{x}^{x+l} |f(x) - g(x)|^p dx \right)^{1/p}.$$

- **Besicovitch distance**

  The **Besicovitch distance** is a distance on the set of all measurable functions $f : \mathbb{R} \to \mathbb{C}$ with summable p-th power on each bounded integral defined by

$$\left(\overline{\lim}_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |f(x) - g(x)|^p dx \right)^{1/p}.$$

  The *generalized Besicovitch almost periodic functions* correspond to this distance.

- **Bochner metric**
  Given a measure space $(\Omega, \mathcal{A}, \mu)$, a Banach space $(V, ||.||_V)$, and $1 \le p \le \infty$, the *Bochner space* (or *Lebesgue–Bochner space*) $L^p(\Omega, V)$ is the set of all measurable functions $f : \Omega \to V$ such that $||f||_{L^p(\Omega,V)} \le \infty$.

  Here the *Bochner norm* $||f||_{L^p(\Omega,V)}$ is defined by $(\int_\Omega ||f(\omega)||_V^p d\mu(\omega))^{\frac{1}{p}}$ for $1 \le p < \infty$, and, for $p = \infty$, by $\operatorname{ess\,sup}_{\omega \in \Omega} ||f(\omega)||_V$.
- **Bergman p-metric**
  Given $1 \le p < \infty$, let $L_p(\Delta)$ be the L_p-space of Lebesgue measurable functions f on the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ with $||f||_p = \left(\int_\Delta |f(z)|^p \mu(dz)\right)^{\frac{1}{p}} < \infty$.

  The *Bergman space* $L_p^a(\Delta)$ is the subspace of $L_p(\Delta)$ consisting of analytic functions, and the **Bergman p-metric** is the **L_p-metric** on $L_p^a(\Delta)$ (cf. **Bergman metric** in Chap. 7). Any Bergman space is a Banach space.
- **Bloch metric**
  The *Bloch space* B on the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ is the set of all analytic functions f on Δ such that $||f||_B = \sup_{z \in \Delta}(1 - |z|^2)|f'(z)| < \infty$. Using the complete *seminorm* $||.||_B$, a norm on B is defined by

$$||f|| = |f(0)| + ||f||_B.$$

  The **Bloch metric** is the **norm metric** $||f - g||$ on B. It makes B a Banach space.
- **Besov metric**
  Given $1 < p < \infty$, the *Besov space* B_p on the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$ is the set of all analytic functions f in Δ such that $||f||_{B_p} = \left(\int_\Delta (1 - |z|^2)^p |f'(z)|^p d\lambda(z)\right)^{\frac{1}{p}} < \infty$, where $d\lambda(z) = \frac{\mu(dz)}{(1-|z|^2)^2}$ is the Möbius invariant measure on Δ. Using the complete *seminorm* $||.||_{B_p}$, the *Besov norm* on B_p is defined by

$$||f|| = |f(0)| + ||f||_{B_p}.$$

  The **Besov metric** is the **norm metric** $||f - g||$ on B_p.

  It makes B_p a Banach space. The set B_2 is the classical *Dirichlet space* of functions analytic on Δ with square integrable derivative, equipped with the **Dirichlet metric**. The *Bloch space B* can be considered as B_∞.
- **Hardy metric**
  Given $1 \le p < \infty$, the *Hardy space* $H^p(\Delta)$ is the class of functions, analytic on the *unit disk* $\Delta = \{z \in \mathbb{C} : |z| < 1\}$, and satisfying the following growth condition for the *Hardy norm* $||.||_{H^p}$:

$$||f||_{H^p(\Delta)} = \sup_{0<r<1} \left(\frac{1}{2\pi} \int_0^{2\pi} |f(re^{i\theta})|^p d\theta \right)^{\frac{1}{p}} < \infty.$$

[redacted]

The **Hardy metric** is the **norm metric** $||f-g||_{H^p(\Delta)}$ on $H^p(\Delta)$. It makes $H^p(\Delta)$ a Banach space.

In Complex Analysis, the Hardy spaces are analogs of the L_p-spaces of Functional Analysis. Such spaces are applied in Mathematical Analysis itself, and also in Scattering Theory and Control Theory (cf. Chap. 18).

- **Part metric**

  The **part metric** is a metric on a *domain* D of $\mathbb{R}^2$ defined for any $x, y \in \mathbb{R}^2$ by

$$\sup_{f \in H^+} \left| \ln \left(\frac{f(x)}{f(y)} \right) \right|,$$

where H^+ is the set of all positive *harmonic functions* on the domain D.

A twice-differentiable real function $f : D \to \mathbb{R}$ is called *harmonic* on D if its *Laplacian* $\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2}$ vanishes on D.

- **Orlicz metric**

  Let $M(u)$ be an even convex function of a real variable which is increasing for u positive, and $\lim_{u \to 0} u^{-1}M(u) = \lim_{u \to \infty} u(M(u))^{-1} = 0$. In this case the function $p(v) = M^{'}(v)$ does not decrease on $[0, \infty)$, $p(0) = \lim_{v \to 0} p(v) = 0$, and $p(v) > 0$ when $v > 0$. Writing $M(u) = \int_0^{|u|} p(v)dv$, and defining $N(u) = \int_0^{|u|} p^{-1}(v)dv$, one obtains a pair $(M(u), N(u))$ of *complementary functions*.

Let $(M(u), N(u))$ be a pair of complementary functions, and let G be a bounded closed set in $\mathbb{R}^n$. The *Orlicz space* $L_M^*(G)$ is the set of Lebesgue-measurable functions f on G satisfying the following growth condition for the *Orlicz norm* $||f||_M$:

$$||f||_M = \sup \left\{ \int_G f(t)g(t)dt : \int_G N(g(t))dt \le 1 \right\} < \infty.$$

The **Orlicz metric** is the norm metric $||f-g||_M$ on $L_M^*(G)$. It makes $L_M^*(G)$ a Banach space ([Orli32]).

When $M(u) = u^p, 1 < p < \infty$, $L_M^*(G)$ coincides with the space $L_p(G)$, and, up to scalar factor, the *L_p-norm* $||f||_p$ coincides with $||f||_M$.

The Orlicz norm is equivalent to the *Luxemburg norm* $||f||_{(M)} = \inf\{\lambda > 0 : \int_G M(\lambda^{-1}f(t))dt \le 1\}$; in fact, $||f||_{(M)} \le ||f||_M \le 2||f||_{(M)}$.

- **Orlicz–Lorentz metric**

  Let $w : (0, \infty) \to (0, \infty)$ be a nonincreasing function. Let $M : [0, \infty) \to [0, \infty)$ be a nondecreasing and convex function with $M(0) = 0$. Let G be a bounded closed set in $\mathbb{R}^n$.

  The *Orlicz–Lorentz space* $L_{w,M}(G)$ is the set of all Lebesgue-measurable functions f on G satisfying the following growth condition for the *Orlicz–Lorentz* *norm* $||f||_{w,M}$:

$$||f||_{w,M} = \inf\left\{\lambda > 0 : \int_0^\infty w(x)M\left(\frac{f^*(x)}{\lambda}\right)dx \leq 1\right\} < \infty,$$

where $f^*(x) = \sup\{t : \mu(|f| \geq t) \geq x\}$ is the *nonincreasing rearrangement* of f.

The **Orlicz–Lorentz metric** is the **norm metric** $||f - g||_{w,M}$ on $L_{w,M}(G)$. It makes $L_{w,M}(G)$ a Banach space.

The Orlicz–Lorentz space is a generalization of the *Orlicz space* $L_M^*(G)$ (cf. **Orlicz metric**), and the *Lorentz space* $L_{w,q}(G)$, $1 \leq q < \infty$, of all Lebesgue-measurable functions f on G satisfying the following growth condition for the *Lorentz norm*:

$$||f||_{w,q} = \left(\int_0^\infty w(x)(f^*(x))^q\right)^{\frac{1}{q}} < \infty.$$

- **Hölder metric**

  Let $L^\alpha(G)$ be the set of all bounded continuous functions f defined on a subset G of $\mathbb{R}^n$, and satisfying the *Hölder condition* on G. Here, a function f satisfies the *Hölder condition* at a point $y \in G$ with *index* (or *order*) α, $0 < \alpha \leq 1$, and with coefficient $A(y)$, if $|f(x) - f(y)| \leq A(y)|x - y|^\alpha$ for all $x \in G$ sufficiently close to y.

  If $A = \sup_{y \in G}(A(y)) < \infty$, the Hölder condition is called *uniform* on G, and A is called the *Hölder coefficient* of G. The quantity $|f|_\alpha = \sup_{x,y \in G}\frac{|f(x)-f(y)|}{|x-y|^\alpha}$, $0 \leq \alpha \leq 1$, is called the *Hölder α-seminorm* of f, and the *Hölder norm* of f is defined by

$$||f||_{L^\alpha(G)} = \sup_{x \in G}|f(x)| + |f|_\alpha.$$

  The **Hölder metric** is the **norm metric** $||f - g||_{L^\alpha(G)}$ on $L^\alpha(G)$. It makes $L^\alpha(G)$ a Banach space.

- **Sobolev metric**

  The *Sobolev space* $W^{k,p}$ is a subset of an L_p-space such that f and its derivatives up to order k have a finite L_p*-norm*. Formally, given a subset G of $\mathbb{R}^n$, define

$$W^{k,p} = W^{k,p}(G) = \{f \in L_p(G) : f^{(i)} \in L_p(G), 1 \leq i \leq k\},$$

  where $f^{(i)} = \partial_{x_1}^{\alpha_1} \ldots \partial_{x_n}^{\alpha_n} f$, $\alpha_1 + \cdots + \alpha_n = i$, and the derivatives are taken in a weak sense. The *Sobolev norm* on $W^{k,p}$ is defined by

$$||f||_{k,p} = \sum_{i=0}^{k}||f^{(i)}||_p.$$

[redacted]

In fact, it is enough to take only the first and last in the sequence, i.e., the norm defined by $||f||_{k,p} = ||f||_p + ||f^{(k)}||_p$ is equivalent to the norm above.

For $p = \infty$, the Sobolev norm is equal to the *essential supremum* of $|f|$: $||f||_{k,\infty} = ess\sup_{x\in G}|f(x)|$, i.e., it is the infimum of all numbers $a \in \mathbb{R}$ for which $|f(x)| > a$ on a set of measure zero.

The **Sobolev metric** is the **norm metric** $||f - g||_{k,p}$ on $W^{k,p}$. It makes $W^{k,p}$ a Banach space.

The Sobolev space $W^{k,2}$ is denoted by H^k. It is a Hilbert space for the *inner product* $\langle f, g\rangle_k = \sum_{i=1}^{k}\langle f^{(i)}, g^{(i)}\rangle_{L_2} = \sum_{i=1}^{k}\int_G f^{(i)}\overline{g}^{(i)}\mu(d\omega)$.

- **Variable exponent space metrics**

  Let G be a nonempty open subset of $\mathbb{R}^n$, and let $p : G \to [1, \infty)$ be a measurable bounded function, called a *variable exponent*. The *variable exponent Lebesgue space* $L_{p(.)}(G)$ is the set of all measurable functions $f : G \to \mathbb{R}$ for which the *modular* $\varrho_{p(.)}(f) = \int_G |f(x)|^{p(x)}dx$ is finite. The *Luxemburg norm* on this space is defined by

$$||f||_{p(.)} = \inf\{\lambda > 0 : \varrho_{p(.)}(f/\lambda) \leq 1\}.$$

  The **variable exponent Lebesgue space metric** is the **norm metric** $||f - g||_{p(.)}$ on $L_{p(.)}(G)$.

  A *variable exponent Sobolev space* $W^{1,p(.)}(G)$ is a subspace of $L_{p(.)}(G)$ consisting of functions f whose distributional gradient exists almost everywhere and satisfies the condition $|\nabla f| \in L_{p(.)}(G)$. The norm

$$||f||_{1,p(.)} = ||f||_{p(.)} + ||\nabla f||_{p(.)}$$

  makes $W^{1,p(.)}(G)$ a Banach space. The **variable exponent Sobolev space metric** is the norm metric $||f - g||_{1,p(.)}$ on $W^{1,p(.)}$.

- **Schwartz metric**

  The *Schwartz space* (or *space of rapidly decreasing functions*) $S(\mathbb{R}^n)$ is the class of all *Schwartz functions*, i.e., infinitely-differentiable functions $f : \mathbb{R}^n \to \mathbb{C}$ that decrease at infinity, as do all their derivatives, faster than any inverse power of x. More precisely, f is a Schwartz function if we have the following growth condition:

$$||f||_{\alpha\beta} = \sup_{x\in\mathbb{R}^n} |x_1^{\beta_1} \dots x_n^{\beta_n} \frac{\partial^{\alpha_1+\dots+\alpha_n} f(x_1, \dots, x_n)}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}| < \infty$$

  for any nonnegative integer vectors α and β. The family of *seminorms* $||.||_{\alpha\beta}$ defines a **locally convex** topology of $S(\mathbb{R}^n)$ which is **metrizable** and complete. The **Schwartz metric** is a metric on $S(\mathbb{R}^n)$ which can be obtained using this topology (cf. **countably normed space** in Chap. 2).

  The corresponding metric space on $S(\mathbb{R}^n)$ is a *Fréchet space* in the sense of Functional Analysis, i.e., a locally convex *F-space*.

- **Bregman quasi-distance**
  Let $G \subset \mathbb{R}^n$ be a closed set with the nonempty interior G^0. Let f be a *Bregman function with zone G*.
  The **Bregman quasi-distance** $D_f : G \times G^0 \to \mathbb{R}_{\geq 0}$ is defined by
  $$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle,$$
  where $\nabla f = (\frac{\partial f}{\partial x_1}, \dots \frac{\partial f}{\partial x_n})$. $D_f(x, y) = 0$ if and only if $x = y$. Also $D_f(x, y) + D_f(y, z) - D_f(x, z) = \langle \nabla f(z) - \nabla f(y), x - y \rangle$ but, in general, D_f does not satisfy the triangle inequality, and is not symmetric.
  A real-valued function f whose effective domain contains G is called a *Bregman function with zone G* if the following conditions hold:
  1. f is continuously differentiable on G^0;
  2. f is strictly convex and continuous on G;
  3. For all $\delta \in \mathbb{R}$ the *partial level sets* $\Gamma(x, \delta) = \{y \in G^0 : D_f(x, y) \leq \delta\}$ are bounded for all $x \in G$;
  4. If $\{y_n\}_n \subset G^0$ converges to y^*, then $D_f(y^*, y_n)$ converges to 0;
  5. If $\{x_n\}_n \subset G$ and $\{y_n\}_n \subset G^0$ are sequences such that $\{x_n\}_n$ is bounded, $\lim_{n\to\infty} y_n = y^*$, and $\lim_{n\to\infty} D_f(x_n, y_n) = 0$, then $\lim_{n\to\infty} x_n = y^*$.

  When $G = \mathbb{R}^n$, a sufficient condition for a strictly convex function to be a Bregman function has the form: $\lim_{||x||\to\infty} \frac{f(x)}{||x||} = \infty$.

13.2 Metrics on Linear Operators

A *linear operator* is a function $T : V \to W$ between two vector spaces V, W over a field $\mathbb{F}$, that is compatible with their linear structures, i.e., for any $x, y \in V$ and any scalar $k \in \mathbb{F}$, we have the following properties: $T(x + y) = T(x) + T(y)$, and $T(kx) = kT(x)$.

- **Operator norm metric**
  Consider the set of all linear operators from a *normed space* $(V, ||.||_V)$ into a normed space $(W, ||.||_W)$. The *operator norm* $||T||$ of a *linear operator* $T : V \to W$ is defined as the largest value by which T stretches an element of V, i.e.,
  $$||T|| = \sup_{||v||_V \neq 0} \frac{||T(v)||_W}{||v||_V} = \sup_{||v||_V = 1} ||T(v)||_W = \sup_{||v||_V \leq 1} ||T(v)||_W.$$
  A linear operator $T : V \to W$ from a normed space V into a normed space W is called *bounded* if its operator norm is finite. For normed spaces, a linear operator is bounded if and only if it is *continuous*.

[redacted]

The **operator norm metric** is a **norm metric** on the set $B(V, W)$ of all bounded linear operators from V into W, defined by

$$||T - P||.$$

The space $(B(V, W), ||.||)$ is called the *space of bounded linear operators*. This metric space is **complete** if W is. If $V = W$ is complete, the space $B(V, V)$ is a *Banach algebra*, as the operator norm is a *submultiplicative norm*.

A linear operator $T : V \rightarrow W$ from a Banach space V into another Banach space W is called *compact* if the image of any bounded subset of V is a relatively compact subset of W. Any compact operator is bounded (and, hence, continuous). The space $(K(V, W), ||.||)$ on the set $K(V, W)$ of all compact operators from V into W with the operator norm $||.||$ is called the *space of compact operators*.

- **Nuclear norm metric**

  Let $B(V, W)$ be the space of all bounded linear operators mapping a Banach space $(V, ||.||_V)$ into another Banach space $(W, ||.||_W)$. Let the *Banach dual* of V be denoted by $V^{'}$, and the value of a functional $x^{'} \in V^{'}$ at a vector $x \in V$ by $\langle x, x^{'} \rangle$.

  A linear operator $T \in B(V, W)$ is called a *nuclear operator* if it can be represented in the form $x \mapsto T(x) = \sum_{i=1}^{\infty} \langle x, x_i^{'} \rangle y_i$, where $\{x_i^{'}\}_i$ and $\{y_i\}_i$ are sequences in $V^{'}$ and W, respectively, such that $\sum_{i=1}^{\infty} ||x_i^{'}||_{V^{'}} ||y_i||_W < \infty$. This representation is called *nuclear*, and can be regarded as an expansion of T as a sum of operators of rank 1 (i.e., with one-dimensional range). The *nuclear norm* of T is defined as

$$||T||_{nuc} = \inf \sum_{i=1}^{\infty} ||x_i^{'}||_{V^{'}} ||y_i||_W,$$

  where the infimum is taken over all possible nuclear representations of T.

  The **nuclear norm metric** is the **norm metric** $||T - P||_{nuc}$ on the set $N(V, W)$ of all nuclear operators mapping V into W. The space $(N(V, W), ||.||_{nuc})$, called the *space of nuclear operators*, is a Banach space.

  A *nuclear space* is defined as a **locally convex** space for which all continuous linear functions into an arbitrary Banach space are nuclear operators. A nuclear space is constructed as a projective limit of Hilbert spaces H_α with the property that, for each $\alpha \in I$, one can find $\beta \in I$ such that $H_\beta \subset H_\alpha$, and the embedding operator $H_\beta \ni x \rightarrow x \in H_\alpha$ is a *Hilbert–Schmidt operator*. A *normed space* is nuclear if and only if it is finite-dimensional.

- **Finite nuclear norm metric**

  Let $F(V, W)$ be the space of all *linear operators of finite rank* (i.e., with finite-dimensional range) mapping a Banach space $(V, ||.||_V)$ into another Banach space $(W, ||.||_W)$. A linear operator $T \in F(V, W)$ can be represented in the form $x \mapsto T(x) = \sum_{i=1}^{n} \langle x, x_i^{'} \rangle y_i$, where $\{x_i^{'}\}_i$ and $\{y_i\}_i$ are sequences in $V^{'}$ (*Banach dual* of V) and W, respectively, and $\langle x, x^{'} \rangle$ is the value of a functional $x^{'} \in V^{'}$ at a vector

$x \in V$. The *finite nuclear norm* of T is defined as

$$||T||_{fnuc} = \inf \sum_{i=1}^{n} ||x_i'||_{V'} ||y_i||_W,$$

where the infimum is taken over all possible finite representations of T.

The **finite nuclear norm metric** is the **norm metric** $||T-P||_{fnuc}$ on $F(V, W)$. The space $(F(V, W), ||.||_{fnuc})$ is called the *space of operators of finite rank*. It is a dense linear subspace of the *space of nuclear operators* $N(V, W)$.

- **Hilbert–Schmidt norm metric**

Consider the set of all linear operators from a Hilbert space $(H_1, ||.||_{H_1})$ into a Hilbert space $(H_2, ||.||_{H_2})$. The *Hilbert–Schmidt norm* $||T||_{HS}$ of a linear operator $T : H_1 \rightarrow H_2$ is defined by

$$||T||_{HS} = (\sum_{\alpha \in I} ||T(e_\alpha)||^2_{H_2})^{1/2},$$

where $(e_\alpha)_{\alpha \in I}$ is an orthonormal basis in H_1. A linear operator $T : H_1 \rightarrow H_2$ is called a *Hilbert–Schmidt operator* if $||T||^2_{HS} < \infty$.

The **Hilbert–Schmidt norm metric** is the **norm metric** $||T-P||_{HS}$ on the set $S(H_1, H_2)$ of all Hilbert–Schmidt operators from H_1 into H_2. In Euclidean space $||.||_{HS}$ is also called *Frobenius norm*; cf. **Frobenius norm metric** in Chap. 12.

For $H_1 = H_2 = H$, the algebra $S(H, H) = S(H)$ with the Hilbert–Schmidt norm is a *Banach algebra*. It contains operators of finite rank as a dense subset, and is contained in the space $K(H)$ of compact operators. An *inner product* $\langle , \rangle_{HS}$ on $S(H)$ is defined by $\langle T, P\rangle_{HS} = \sum_{\alpha \in I} \langle T(e_\alpha), P(e_\alpha)\rangle$, and $||T||_{HS} = \langle T, T\rangle^{1/2}_{HS}$. So, $S(H)$ is a Hilbert space (independent of the chosen basis $(e_\alpha)_{\alpha \in I}$).

- **Trace-class norm metric**

Given a Hilbert space H, the *trace-class norm* of a linear operator $T : H \rightarrow H$ is

$$||T||_{tc} = \sum_{\alpha \in I} \langle |T|(e_\alpha), e_\alpha\rangle,$$

where $|T|$ is the *absolute value* of T in the *Banach algebra* $B(H)$ of all bounded operators from H into itself, and $(e_\alpha)_{\alpha \in I}$ is an orthonormal basis of H.

An operator $T : H \rightarrow H$ is called a *trace-class operator* if $||T||_{tc} < \infty$. Any such operator is the product of two *Hilbert–Schmidt operators*.

The **trace-class norm metric** is the **norm metric** $||T - P||_{tc}$ on the set $L(H)$ of all trace-class operators from H into itself.

The set $L(H)$ with the norm $||.||_{tc}$ forms a Banach algebra which is contained in the algebra $K(H)$ (of all compact operators from H into itself), and contains the algebra $S(H)$ of all Hilbert–Schmidt operators from H into itself.

[redacted]

- **Schatten p-class norm metric**

  Let $1 \leq p < \infty$. Given a separable Hilbert space H, the *Schatten p-class norm* of a compact linear operator $T : H \to H$ is defined by

$$||T||_{Sch}^{p} = \left(\sum_{n} |s_n|^p \right)^{\frac{1}{p}},$$

  where $\{s_n\}_n$ is the sequence of *singular values* of T. A compact operator $T : H \to H$ is called a *Schatten p-class operator* if $||T||_{Sch}^{p} < \infty$.

  The **Schatten p-class norm metric** is the **norm metric** $||T - P||_{Sch}^{p}$ on the set $S_p(H)$ of all Schatten p-class operators from H onto itself. The set $S_p(H)$ with the norm $||.||_{Sch}^{p}$ forms a Banach space. $S_1(H)$ is the *trace-class* of H, and $S_2(H)$ is the *Hilbert–Schmidt class* of H. Cf. **Schatten norm metric** (in Chap. 12) for which **trace** and **Frobenius** norm metrics are cases $p = 1$ and $p = 2$, respectively.

- **Continuous dual space**

  For any vector space V over some field, its *algebraic dual space* is the set of all linear functionals on V.

  Let $(V, ||.||)$ be a *normed vector space*. Let V' be the set of all *continuous* linear functionals T from V into the base field ($\mathbb{R}$ or $\mathbb{C}$). Let $||.||'$ be the *operator norm* on V' defined by

$$||T||' = \sup_{||x|| \leq 1} |T(x)|.$$

  The space $(V', ||.||')$ is a Banach space which is called the **continuous dual** (or *Banach dual*) of $(V, ||.||)$.

  The continuous dual of the metric space l_p^n (l_p^∞) is l_q^n (l_q^∞, respectively), where q is defined by $\frac{1}{p} + \frac{1}{q} = 1$. The continuous dual of l_1^n (l_1^∞) is l_∞^n (l_∞^∞, respectively).

- **Distance constant of operator algebra**

  Let $\mathcal{A}$ be an subalgebra of $B(H)$, the algebra of all bounded operators on a Hilbert space H. For any operator $T \in B(H)$, let P be a projection, $P^\perp$ be its orthogonal complement and $\beta(T, \mathcal{A}) = \sup\{||P^\perp TP|| : P^\perp \mathcal{A} P = (0)\}$.

  Let $dist(T, \mathcal{A}) = \inf_{A \in \mathcal{A}} ||T - A||$ be the *distance of* T *to algebra* $\mathcal{A}$; cf. **matrix nearness problems** in Chap. 12. It holds $dist(T, \mathcal{A}) \geq \beta(T, \mathcal{A})$.

  The algebra $\mathcal{A}$ is *reflexive* if $\beta(T, \mathcal{A}) = 0$ implies $T \in \mathcal{A}$; it is *hyperreflexive* if there exists a constant $C \geq 1$ such that, for any operator $T \in B(H)$, it holds

$$dist(T, \mathcal{A}) \leq C\beta(T, \mathcal{A}).$$

  The smallest such C is called the **distance constant** of the algebra $\mathcal{A}$.

  In the case of a reflexive algebra of matrices with nonzero entries specified by a given pattern, the problem of finding the distance constant can be formulated as a matrix-filling problem: given a partially completed matrix, fill in the remaining entries so that the operator norm of the resulting complete matrix is minimized.

Chapter 14
Distances in Probability Theory

A *probability space* is a *measurable space* $(\Omega, \mathcal{A}, P)$, where $\mathcal{A}$ is the set of all measurable subsets of Ω, and P is a measure on $\mathcal{A}$ with $P(\Omega) = 1$. The set Ω is called a *sample space*. An element $a \in \mathcal{A}$ is called an *event*. $P(a)$ is called the *probability* of the event a. The measure P on $\mathcal{A}$ is called a *probability measure*, or *(probability) distribution law*, or simply *(probability) distribution*.

A *random variable* X is a measurable function from a probability space $(\Omega, \mathcal{A}, P)$ into a measurable space, called a *state space* of possible values of the variable; it is usually taken to be $\mathbb{R}$ with the *Borel σ-algebra*, so $X : \Omega \to \mathbb{R}$. The range $\mathcal{X}$ of the variable X is called the *support* of the distribution P; an element $x \in \mathcal{X}$ is called a *state*.

A distribution law can be uniquely described via a *cumulative distribution* (or simply, *distribution*) *function* CDF, which describes the probability that a random value X takes on a value at most x: $F(x) = P(X \leq x) = P(\omega \in \Omega : X(\omega) \leq x)$.

So, any random variable X gives rise to a *probability distribution* which assigns to the interval $[a, b]$ the probability $P(a \leq X \leq b) = P(\omega \in \Omega : a \leq X(\omega) \leq b)$, i.e., the probability that the variable X will take a value in the interval $[a, b]$.

A distribution is called *discrete* if $F(x)$ consists of a sequence of finite jumps at x_i; a distribution is called *continuous* if $F(x)$ is continuous. We consider (as in the majority of applications) only discrete or *absolutely continuous* distributions, i.e., the CDF function $F : \mathbb{R} \to \mathbb{R}$ is *absolutely continuous*. It means that, for every number $\epsilon > 0$, there is a number $\delta > 0$ such that, for any sequence of pairwise disjoint intervals $[x_k, y_k]$, $1 \leq k \leq n$, the inequality $\sum_{1 \leq k \leq n}(y_k - x_k) < \delta$ implies the inequality $\sum_{1 \leq k \leq n} |F(y_k) - F(x_k)| < \epsilon$.

A distribution law also can be uniquely defined via a *probability density* (or *density*, *probability*) *function* PDF of the underlying random variable. For an absolutely continuous distribution, the CDF is almost everywhere differentiable, and the PDF is defined as the derivative $p(x) = F^{'}(x)$ of the CDF; so, $F(x) = P(X \leq x) = \int_{-\infty}^{x} p(t)dt$, and $\int_{a}^{b} p(t)dt = P(a \leq X \leq b)$. In the discrete case, the PDF is is

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_14

[redacted]

$\sum_{x_i \leq x} p(x_i)$, where $p(x) = P(X = x)$ is the *probability mass function*. But $p(x) = 0$ for each fixed x in any continuous case.

The random variable X is used to "push-forward" the measure P on Ω to a measure dF on $\mathbb{R}$. The underlying probability space is a technical device used to guarantee the existence of random variables and sometimes to construct them.

We usually present the discrete version of probability metrics, but many of them are defined on any measurable space; see [Bass89, Bass13, Cha08]. For a probability distance d on random quantities, the conditions $P(X = Y) = 1$ or equality of distributions imply (and characterize) $d(X, Y) = 0$; such distances are called ([Rach91]) *compound* or *simple* distances, respectively. Often, some *ground* distance d is given on the state space $\mathcal{X}$ and the presented distance is a lifting of it to a distance on distributions. A quasi-distance between distributions is also called **divergence** or *distance statistic*.

Below we denote $p_X = p(x) = P(X = x)$, $F_X = F(x) = P(X \leq x)$, $p(x, y) = P(X = x, Y = y)$. We denote by $\mathbb{E}[X]$ the *expected value* (or *mean*) of the random variable X: in the discrete case $\mathbb{E}[X] = \sum_x xp(x)$, in the continuous case $\mathbb{E}[X] = \int xp(x)dx$.

The *covariance* between the random variables X and Y is $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. The *variance* and *standard deviation* of X are $Var(X) = Cov(X, X)$ and $\sigma(X) = \sqrt{Var(X)}$, respectively. The *correlation* between X and Y is $Corr(X, Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$; cf. Chap. 17.

14.1 Distances on Random Variables

All distances in this section are defined on the set $\mathbf{Z}$ of all random variables with the same support $\mathcal{X}$; here $X, Y \in \mathbf{Z}$.

- ***p*-average compound metric**

  Given $p \geq 1$, the ***p*-average compound metric** (or *L_p-metric between variables*) is a metric on $\mathbf{Z}$ with $\mathcal{X} \subset \mathbb{R}$ and $\mathbb{E}[|Z|^p] < \infty$ for all $Z \in \mathbf{Z}$ defined by

$$(\mathbb{E}[|X - Y|^p])^{1/p} = \Big(\sum_{(x,y) \in \mathcal{X} \times \mathcal{X}} |x - y|^p p(x, y)\Big)^{1/p}.$$

  For $p = 2$ and ∞, it is called, respectively, the *mean-square distance* and *essential supremum distance* between variables.

- **Lukaszyk–Karmovski metric**

  The **Lukaszyk–Karmovski metric** (2001) on $\mathbb{Z}$ with $\mathcal{X} \subset \mathbb{R}$ is defined by

$$\sum_{(x,y) \in \mathcal{X} \times \mathcal{X}} |x - y| p(x)p(y).$$

For continuous random variables, it is defined by $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|x-y|F(x)F(y)dxdy$. This function can be positive for $X = Y$. Such possibility is excluded, and so, it will be a distance metric, if and only if it holds

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|x-y|\delta(x-\mathbb{E}[X])\delta(y-\mathbb{E}[Y])dxdy = |\mathbb{E}[X]-\mathbb{E}[Y]|.$$

- **Absolute moment metric**

  Given $p \geq 1$, the **absolute moment metric** is a metric on **Z** with $\mathcal{X} \subset \mathbb{R}$ and $\mathbb{E}[|Z|^p] < \infty$ for all $Z \in \mathbf{Z}$ defined by

$$|(\mathbb{E}[|X|^p])^{1/p} - (\mathbb{E}[|Y|^p])^{1/p}|.$$

  For $p = 1$ it is called the *engineer metric*.
- **Indicator metric**

  The **indicator metric** is a metric on **Z** defined by

$$\mathbb{E}[1_{X\neq Y}] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} 1_{x\neq y}p(x,y) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{X},x\neq y} p(x,y).$$

  (Cf. **Hamming metric** in Chap. 1.)
- **Ky Fan metric** K

  The **Ky Fan metric** K is a metric K on **Z**, defined by

$$\inf\{\epsilon > 0 : P(|X-Y| > \epsilon) < \epsilon\}.$$

  It is the case $d(x,y) = |X-Y|$ of the **probability distance**.
- **Ky Fan metric** K^*

  The **Ky Fan metric** K^* is a metric on **Z** defined by

$$\mathbb{E}\left[\frac{|X-Y|}{1+|X-Y|}\right] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \frac{|x-y|}{1+|x-y|}p(x,y).$$

- **Probability distance**

  Given a metric space $(\mathcal{X}, d)$, the **probability distance** on **Z** is defined by

$$\inf\{\epsilon > 0 : P(d(X,Y) > \epsilon) < \epsilon\}.$$

14.2 Distances on Distribution Laws

All distances in this section are defined on the set $\mathcal{P}$ of all distribution laws such that corresponding random variables have the same range $\mathcal{X}$; here $P_1, P_2 \in \mathcal{P}$.

[redacted]

- **L_p-metric between densities**

  The **L_p-metric between densities** is a metric on $\mathcal{P}$ (for a countable $\mathcal{X}$) defined, for any $p \geq 1$, by

  $$(\sum_x |p_1(x) - p_2(x)|^p)^{\frac{1}{p}}.$$

  For $p = 1$, one half of it is called the **variational distance** (or *total variation distance*, *Kolmogorov distance*). For $p = 2$, it is the **Patrick–Fisher distance**. The *point metric* $\sup_x |p_1(x) - p_2(x)|$ corresponds to $p = \infty$.

  The **Lissak–Fu distance** with parameter $\alpha > 0$ is defined as $\sum_x |p_1(x) - p_2(x)|^\alpha$.
- **Bayesian distance**

  The *error probability in classification* is the following error probability of the optimal Bayes rule for the classification into two classes with a priori probabilities $\phi, 1 - \phi$ and corresponding densities p_1, p_2 of the observations:

  $$P_e = \sum_x \min(\phi p_1(x), (1 - \phi)p_2(x)).$$

  The **Bayesian distance** on $\mathcal{P}$ is defined by $1 - P_e$.

  For the classification into m classes with *a priori* probabilities ϕ_i, $1 \leq i \leq m$, and corresponding densities p_i of the observations, the error probability becomes

  $$P_e = 1 - \sum_x p(x) \max_i P(C_i|x),$$

  where $P(C_i|x)$ is the *a posteriori* probability of the class C_i given the observation x and $p(x) = \sum_{i=1}^m \phi_i P(x|C_i)$. The *general mean distance between m classes C_i* (cf. m-hemimetric in Chap. 3) is defined (Van der Lubbe, 1979) for $\alpha > 0, \beta > 1$ by

  $$\sum_x p(x)(\sum_i P(C_i|x)^\beta)^\alpha.$$

  The case $\alpha = 1, \beta = 2$ corresponds to the *Bayesian distance* in Devijver, 1974; the case $\beta = \frac{1}{\alpha}$ was considered in Trouborst et al., 1974.
- **Mahalanobis semimetric**

  The **Mahalanobis semimetric** is a semimetric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}^n$) defined by

  $$\sqrt{(\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X])^T A(\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X])}$$

  for a given positive-semidefinite matrix A; its square is a **Bregman quasi-distance** (Chap. 13). Cf. also the **Mahalanobis distance** in Chap. 17.

- **Engineer semimetric**

  The **engineer semimetric** is a semimetric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$) defined by

$$|\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X]| = |\sum_x x(p_1(x) - p_2(x))|.$$

- **Stop-loss metric of order** m

  The **stop-loss metric of order** m is a metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$) defined by

$$\sup_{t\in\mathbb{R}} \sum_{x\geq t} \frac{(x-t)^m}{m!}(p_1(x) - p_2(x)).$$

- **Kolmogorov–Smirnov metric**

  The **Kolmogorov–Smirnov metric** (or *Kolmogorov metric*, *uniform metric*) is a metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$) defined (1948) by

$$\sup_{x\in\mathbb{R}} |P_1(X \leq x) - P_2(X \leq x)|.$$

  This metric is used, for example, in Biology as a measure of sexual dimorphism.

  The **Kuiper distance** on $\mathcal{P}$ is defined by

$$\sup_{x\in\mathbb{R}}(P_1(X \leq x) - P_2(X \leq x)) + \sup_{x\in\mathbb{R}}(P_2(X \leq x) - P_1(X \leq x)).$$

  (Cf. **Pompeiu–Eggleston metric** in Chap. 9.)

  The **Crnkovic–Drachma distance** is defined by

$$\sup_{x\in\mathbb{R}}(P_1(X \leq x) - P_2(X \leq x)) \ln \frac{1}{\sqrt{(P_1(X \leq x)(1 - P_1(X \leq x))}} +$$

$$+ \sup_{x\in\mathbb{R}}(P_2(X \leq x) - P_1(X \leq x)) \ln \frac{1}{\sqrt{(P_1(X \leq x)(1 - P_1(X \leq x))}}.$$

- **Cramér–von Mises distance**

  The **Cramér–von Mises distance** (1928) is defined on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$) by

$$\omega^2 = \int_{-\infty}^{+\infty} (P_1(X \leq x) - P_2(X \leq x))^2 dP_2(x).$$

  The **Anderson–Darling distance** (1954) on $\mathcal{P}$ is defined by

$$\int_{-\infty}^{+\infty} \frac{(P_1(X \leq x) - P_2)(X \leq x))^2}{(P_2(X \leq x)(1 - P_2(X \leq x))} dP_2(x).$$

[redacted]

In Statistics, above distances of Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling and, below, χ^2**-distance** are the main measures of *goodness of fit* between estimated, P_2, and theoretical, P_1, distributions.

They and other distances were generalized (for example by Kiefer, 1955, and Glick, 1969) on *K-sample setting*, i.e., some convenient generalized distances $d(P_1, \dots, P_K)$ were defined. Cf. **m-hemimetric** in Chap. 3.

- **Energy distance**

  The **energy distance** (Széely, 1985) between cumulative density functions $F(X)$, $F(Y)$ of two independent random vectors $X, Y \in \mathbb{R}^n$ is defined by

  $$d(F(X), F(Y)) = 2\mathbb{E}[||(X - Y||] - \mathbb{E}[||X - X'||] - \mathbb{E}[||(Y - Y'||],$$

  where X, X' are *iid* (independent and identically distributed), Y, Y' are iid and $||.||$ is the length of a vector. For real-valued random variables this distance is exactly twice **Cramér–von Mises distance**. Cf. **distance covariance** in Chap. 17.

  It holds $d(F(X), F(Y)) = 0$ if and only if X, Y are iid.

- **Prokhorov metric**

  Given a metric space $(\mathcal{X}, d)$, the **Prokhorov metric** on $\mathcal{P}$ is defined (1956) by

  $$\inf\{\epsilon > 0 : P_1(X \in B) \leq P_2(X \in B^\epsilon) + \epsilon \text{ and } P_2(X \in B) \leq P_1(X \in B^\epsilon) + \epsilon\},$$

  where B is any Borel subset of $\mathcal{X}$, and $B^\epsilon = \{x : d(x, y) < \epsilon, y \in B\}$.

  It is the smallest (over all joint distributions of pairs (X, Y) of random variables X, Y such that the marginal distributions of X and Y are P_1 and P_2, respectively) **probability distance** between random variables X and Y.

- **Levy–Sibley metric**

  The **Levy–Sibley metric** is a metric on $\mathcal{P}$ (for $\mathcal{X} \subset \mathbb{R}$ only) defined by

  $$\inf\{\epsilon > 0 : P_1(X \leq x - \epsilon) - \epsilon \leq P_2(X \leq x) \leq P_1(X \leq x + \epsilon) + \epsilon \text{ for any } x \in \mathbb{R}\}.$$

  It is a special case of the **Prokhorov metric** for $(\mathcal{X}, d) = (\mathbb{R}, |x - y|)$.

- **Dudley metric**

  Given a metric space $(\mathcal{X}, d)$, the **Dudley metric** on $\mathcal{P}$ is defined by

  $$\sup_{f \in F} |\mathbb{E}_{P_1}[f(X)] - \mathbb{E}_{P_2}[f(X)]| = \sup_{f \in F} |\sum_{x \in \mathcal{X}} f(x)(p_1(x) - p_2(x))|,$$

  where $F = \{f : \mathcal{X} \to \mathbb{R}, ||f||_\infty + Lip_d(f) \leq 1\}$, and $Lip_d(f) = \sup_{x,y \in \mathcal{X}, x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}$.

- **Szulga metric**
  Given a metric space $(\mathcal{X}, d)$, the **Szulga metric** (1982) on $\mathcal{P}$ is defined by

$$\sup_{f \in F} |(\sum_{x \in \mathcal{X}} |f(x)|^p p_1(x))^{1/p} - (\sum_{x \in \mathcal{X}} |f(x)|^p p_2(x))^{1/p}|,$$

  where $F = \{f : X \to \mathbb{R},\ Lip_d(f) \leq 1\}$, and $Lip_d(f) = \sup_{x,y \in \mathcal{X}, x \neq y} \frac{|f(x)-f(y)|}{d(x,y)}$.
- **Zolotarev semimetric**
  The **Zolotarev semimetric** is a semimetric on $\mathcal{P}$, defined (1976) by

$$\sup_{f \in F} | \sum_{x \in \mathcal{X}} f(x)(p_1(x) - p_2(x))|,$$

  where F is any set of functions $f : \mathcal{X} \to \mathbb{R}$ (in the continuous case, F is any set of such bounded continuous functions); cf. **Szulga metric**, **Dudley metric**.
- **Convolution metric**
  Let G be a separable locally compact Abelian group, and let $C(G)$ be the set of all real bounded continuous functions on G vanishing at infinity. Fix a function $g \in C(G)$ such that $|g|$ is integrable with respect to the Haar measure on G, and $\{\beta \in G^* : \hat{g}(\beta) = 0\}$ has empty interior; here G^* is the dual group of G, and $\hat{g}$ is the Fourier transform of g.

  The **convolution metric** (or *smoothing metric*) is defined (Yukich, 1985), for any two finite signed Baire measures P_1 and P_2 on G, by

$$\sup_{x \in G} | \int_{y \in G} g(xy^{-1})(dP_1 - dP_2)(y)|.$$

  It can also be seen as the difference $T_{P_1}(g) - T_{P_2}(g)$ of *convolution operators* on $C(G)$ where, for any $f \in C(G)$, the operator $T_P f(x)$ is $\int_{y \in G} f(xy^{-1}) dP(y)$.

  In particular, this metric can be defined on the space of probability measures on $\mathbb{R}^n$, where g is a PDF satisfying above conditions.
- **Discrepancy metric**
  Given a metric space $(\mathcal{X}, d)$, the **discrepancy metric** on $\mathcal{P}$ is defined by

$$\sup\{|P_1(X \in B) - P_2(X \in B)| : B \text{ is any closed ball}\}.$$

- **Bi-discrepancy semimetric**
  The **bi-discrepancy semimetric** (evaluating the proximity of distributions P_1, P_2 over different collections $\mathcal{A}_1, \mathcal{A}_2$ of measurable sets) is defined by

$$D(P_1, P_2) + D(P_2, P_1),$$

  where $D(P_1, P_2) = \sup\{\inf\{P_2(C) : B \subset C \in \mathcal{A}_2\} - P_1(B) : B \in \mathcal{A}_1\}$ (*discrepancy*).

[redacted]

- **Le Cam distance**

  The **Le Cam distance** (1974) is a semimetric, evaluating the proximity of probability distributions P_1, P_2 (on different spaces $\mathcal{X}_1, \mathcal{X}_2$) and defined as follows:

  $$\max\{\delta(P_1, P_2), \delta(P_2, P_1)\},$$

  where $\delta(P_1, P_2) = \inf_B \sum_{x_2 \in \mathcal{X}_2} |BP_1(X_2 = x_2) - BP_2(X_2 = x_2)|$ is the *Le Cam deficiency*. Here $BP_1(X_2 = x_2) = \sum_{x_1 \in \mathcal{X}_1} p_1(x_1) b(x_2|x_1)$, where B is a probability distribution over $\mathcal{X}_1 \times \mathcal{X}_2$, and

  $$b(x_2|x_1) = \frac{B(X_1 = x_1, X_2 = x_2)}{B(X_1 = x_1)} = \frac{B(X_1 = x_1, X_2 = x_2)}{\sum_{x \in \mathcal{X}_2} B(X_1 = x_1, X_2 = x)}.$$

  So, $BP_2(X_2 = x_2)$ is a probability distribution over $\mathcal{X}_2$, since $\sum_{x_2 \in \mathcal{X}_2} b(x_2|x_1) = 1$.

  Le Cam distance is not a probabilistic distance, since P_1 and P_2 are defined over different spaces; it is a distance between statistical experiments (models).

- **Skorokhod–Billingsley metric**

  The **Skorokhod–Billingsley metric** is a metric on $\mathcal{P}$, defined by

  $$\inf_f \max\left\{\sup_x |P_1(X \le x) - P_2(X \le f(x))|, \sup_x |f(x) - x|, \sup_{x \ne y} \left|\ln \frac{f(y) - f(x)}{y - x}\right|\right\},$$

  where $f : \mathbb{R} \to \mathbb{R}$ is any strictly increasing continuous function.

- **Skorokhod metric**

  The **Skorokhod metric** is a metric on $\mathcal{P}$ defined (1956) by

  $$\inf\{\epsilon > 0 : \max\{\sup_x |P_1(X < x) - P_2(X \le f(x))|, \sup_x |f(x) - x|\} < \epsilon\},$$

  where $f : \mathbb{R} \to \mathbb{R}$ is a strictly increasing continuous function.

- **Birnbaum–Orlicz distance**

  The **Birnbaum–Orlicz distance** (1931) is a distance on $\mathcal{P}$ defined by

  $$\sup_{x \in \mathbb{R}} f(|P_1(X \le x) - P_2(X \le x)|),$$

  where $f : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ is any nondecreasing continuous function with $f(0) = 0$, and $f(2t) \le Cf(t)$ for any $t > 0$ and some fixed $C \ge 1$. It is a **near-metric**, since the **C-triangle inequality** $d(P_1, P_2) \le C(d(P_1, P_3) + d(P_3, P_2))$ holds.

  Birnbaum–Orlicz distance is also used, in Functional Analysis, on the set of all integrable functions on the segment $[0, 1]$, where it is defined by $\int_0^1 H(|f(x) - g(x)|)dx$, where H is a nondecreasing continuous function from $[0, \infty)$ onto $[0, \infty)$ which vanishes at the origin and satisfies the *Orlicz condition*: $\sup_{t>0} \frac{H(2t)}{H(t)} < \infty$.

- **Kruglov distance**

  The **Kruglov distance** (1973) is a distance on $\mathcal{P}$, defined by

  $$\int f(P_1(X \leq x) - P_2(X \leq x))dx,$$

  where $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is any even strictly increasing function with $f(0) = 0$, and $f(s + t) \leq C(f(s) + f(t))$ for any $s, t \geq 0$ and some fixed $C \geq 1$. It is a **near-metric**, since the **C-triangle inequality** $d(P_1, P_2) \leq C(d(P_1, P_3) + d(P_3, P_2))$ holds.
- **Bregman divergence**

  Given a differentiable strictly convex function $\phi(p) : \mathbb{R}^n \to \mathbb{R}$ and $\beta \in (0, 1)$, the **skew Jensen** (or *skew Burbea–Rao*) **divergence** on $\mathcal{P}$ is (Basseville–Cardoso, 1995)

  $$J_{\phi}^{(\beta)}(P_1, P_2) = \beta\phi(p_1) + (1 - \beta)\phi(p_2) - \phi(\beta p_1 + (1 - \beta)p_2).$$

  The **Burbea–Rao distance** (1982) is the case $\beta = \frac{1}{2}$ of it, i.e., it is

  $$\sum_x \left(\frac{\phi(p_1(x)) + \phi(p_2(x))}{2} - \phi(\frac{p_1(x) + (p_2(x)}{2}) \right).$$

  The **Bregman divergence** (1967) is a quasi-distance on $\mathcal{P}$ defined by

  $$\sum_x (\phi(p_1(x)) - \phi(p_2(x)) - (p_1(x) - p_2(x))\phi'(p_2(x))) = \lim_{\beta \to 1} \frac{1}{\beta} J_{\phi}^{(\beta)}(P_1, P_2).$$

  The **generalised Kullback–Leibler distance** $\sum_x p_1(x) \ln \frac{p_1(x)}{p_2(x)} - \sum_x (p_1(x) - p_2(x))$ and **Itakura–Saito distance** (Chap. 21) $\sum_x \frac{p_1(x)}{p_2(x)} - \ln \frac{p_1(x)}{p_2(x)} - 1$ are the cases $\phi(p) = \sum_x p(x) \ln p(x) - \sum_x p(x)$ and $\phi(p) = -\sum_x \ln p(x)$ of the Bregman divergence. Cf. **Bregman quasi-distance** in Chap. 13.

  Csizár, 1991, proved that the **Kullback–Leibler distance** is the only **Bregman divergence** which is an **f-divergence**.
- **f-divergence**

  Given a convex function $f(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with $f(1) = 0, f'(1) = 0, f''(1) = 1$, the **$f$-divergence** (independently, Csizár, 1963, Morimoto, 1963, Ali–Silvey, 1966, Ziv–Zakai, 1973, and Akaike, 1974) on $\mathcal{P}$ is defined by

  $$\sum_x p_2(x) f\left(\frac{p_1(x)}{p_2(x)}\right).$$

  The cases $f(t) = t \ln t$ and $f(t) = (t - 1)^2$ correspond to the **Kullback–Leibler distance** and to the **χ^2-distance** below, respectively. The case $f(t) = |t - 1|$

[redacted]

corresponds to the **variational distance**, and the case $f(t) = 4(1 - \sqrt{t})$ (as well as $f(t) = 2(t + 1) - 4\sqrt{t}$) corresponds to the squared **Hellinger metric**.

Semimetrics can also be obtained, as the square root of the f-divergence, in the cases $f(t) = (t-1)^2/(t+1)$ (the **Vajda–Kus semimetric**), $f(t) = |t^a - 1|^{1/a}$ with $0 < a \leq 1$ (the **generalized Matusita distance**), and $f(t) = \frac{(t^a+1)^{1/a} - 2^{(1-a)/a}(t+1)}{1-1/\alpha}$ (the **Osterreicher semimetric**).

- **α-divergence**

  Given $\alpha \in \mathbb{R}$, the **α-divergence** (independently, Csizár, 1967, Havrda–Charvát, 1967, Cressie–Read, 1984, and Amari, 1985) is defined as $KL(P_1, P_2)$, $KL(P_2, P_1)$ for $\alpha = 1, 0$ and for $\alpha \neq 0, 1$, it is

$$\frac{1}{\alpha(1-\alpha)}(1 - \sum_x p_2(x)\left(\frac{p_1(x)}{p_2(x)}\right)^{\alpha}).$$

  The **Amari divergence** come from the above by the transformation $\alpha = \frac{1+t}{2}$.

- **Harmonic mean similarity**

  The **harmonic mean similarity** is a similarity on $\mathcal{P}$ defined by

$$2\sum_x \frac{p_1(x)p_2(x)}{p_1(x) + p_2(x)}.$$

- **Fidelity similarity**

  The **fidelity similarity** (or *Bhattacharya coefficient*, *Hellinger affinity*) on $\mathcal{P}$ is

$$\rho(P_1, P_2) = \sum_x \sqrt{p_1(x)p_2(x)}.$$

  Cf. more general **quantum fidelity similarity** in Chap. 24.

- **Hellinger metric**

  In terms of the **fidelity similarity** ρ, the **Hellinger metric** (or **Matusita distance**, *Hellinger–Kakutani metric*) on $\mathcal{P}$ is defined by

$$(\sum_x (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2)^{\frac{1}{2}} = \sqrt{2(1 - \rho(P_1, P_2))}.$$

- **Bhattacharya distance** 1

  In terms of the **fidelity similarity** ρ, the **Bhattacharya distance** 1 (1946) is

$$(\arccos \rho(P_1, P_2))^2$$

  for $P_1, P_2 \in \mathcal{P}$. Twice this distance is the **Rao distance** from Chap. 7. It is used also in Statistics and Machine Learning, where it is called the *Fisher distance*.

The **Bhattacharya distance** 2 (1943) on $\mathcal{P}$ is defined by

$$-\ln \rho(P_1, P_2).$$

- **χ^2-distance**

  The **χ^2-distance** (or **Pearson χ^2-distance**) is a quasi-distance on $\mathcal{P}$, defined by

  $$\sum_x \frac{(p_1(x) - p_2(x))^2}{p_2(x)}.$$

  The **Neyman χ^2-distance** is a quasi-distance on $\mathcal{P}$, defined by

  $$\sum_x \frac{(p_1(x) - p_2(x))^2}{p_1(x)}.$$

  The half of χ^2-distance is also called *Kagan's divergence*.

  The probabilistic **symmetric χ^2-measure** is a distance on $\mathcal{P}$, defined by

  $$2\sum_x \frac{(p_1(x) - p_2(x))^2}{p_1(x) + p_2(x)}.$$

- **Separation quasi-distance**

  The **separation distance** is a quasi-distance on $\mathcal{P}$ (for a countable $\mathcal{X}$) defined by

  $$\max_x \left(1 - \frac{p_1(x)}{p_2(x)}\right).$$

  (Not to be confused with **separation distance** in Chap. 9.)

- **Kullback–Leibler distance**

  The **Kullback–Leibler distance** (or *relative entropy*, *information deviation*, *information gain*, *KL-distance*) is a quasi-distance on $\mathcal{P}$, defined (1951) by

  $$KL(P_1, P_2) = \mathbb{E}_{P_1}[\ln L] = \sum_x p_1(x) \ln \frac{p_1(x)}{p_2(x)},$$

  where $L = \frac{p_1(x)}{p_2(x)}$ is the *likelihood ratio*. Therefore,

  $$KL(P_1, P_2) = -\sum_x p_1(x) \ln p_2(x) + \sum_x p_1(x) \ln p_1(x) = H(P_1, P_2) - H(P_1),$$

  where $H(P_1)$ is the *entropy* of P_1, and $H(P_1, P_2)$ is the *cross-entropy* of P_1 and P_2.

[redacted]

If P_2 is the product of marginals of P_1 (say, $p_2(x,y) = p_1(x)p_1(y)$), the KL-distance $KL(P_1, P_2)$ is called the *Shannon information quantity* and (cf. **Shannon distance**) is equal to $\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} p_1(x,y)\ln\frac{p_1(x,y)}{p_1(x)p_1(y)}$.

The **exponential divergence** is defined by $\sum_x p_1(x)(\ln\frac{p_1(x)}{p_2(x)})^2$.

- **Distance to normality**

For a continuous distribution P on $\mathbb{R}$, the *differential entropy* is defined by

$$h(P) = -\int_{-\infty}^{\infty} p(x)\ln p(x)dx.$$

It is $\ln(\delta\sqrt{2\pi e})$ for a *normal* (or *Gaussian*) *distribution* $g_{\delta,\mu}(x) = \frac{1}{\sqrt{2\pi\delta^2}}\exp\left(-\frac{(x-\mu)^2}{2\delta^2}\right)$ with variance δ^2 and mean μ.

The **distance to normality** (or *negentropy*) of P is the **Kullback–Leibler distance** $KL(P, g) = \int_{-\infty}^{\infty} p(x)\ln\left(\frac{p(x)}{g(x)}\right)dx = h(g) - h(P)$, where q is a normal distribution with the same variance as P. So, it is nonnegative and equal to 0 if and only if $P = g$ almost everywhere. Cf. **Shannon distance**.

Also, $h(u_{a,b}) = \ln(b-a)$ for an *uniform distribution* with minimum a and maximum $b > a$, i.e., $u_{a,b}(x) = \frac{1}{b-a}$, if $x \in [a,b]$, and it is 0, otherwise. It holds $h(u_{a,b}) \geq h(P)$ for any distribution P with support contained in $[a,b]$; so, $h(u_{a,b}) - h(P)$ can be called the *distance to uniformity*. Tononi, 2008, used it in his model of consciousness.

- **Jeffrey distance**

The **Jeffrey distance** (or *J-divergence*, *KL2-distance*) is a symmetric version of the **Kullback–Leibler distance** defined (1946) on $\mathcal{P}$ by

$$KL(P_1,P_2) + KL(P_2,P_1) = \sum_x((p_1(x) - p_2(x))\ln\frac{p_1(x)}{p_2(x)}.$$

The **Aitchison distance** (1986) is defined by $\sqrt{\sum_x(\ln\frac{p_1(x)g(p_1)}{p_2(x)g(p_2)})^2}$, where $g(p) = (\prod_x p(x))^{\frac{1}{n}}$ is the geometric mean of components $p(x)$ of p.

- **Resistor-average distance**

The **resistor-average distance** is (Johnson–Simanović, 2000) a symmetric version of the **Kullback–Leibler distance** on $\mathcal{P}$ which is defined by the harmonic sum

$$\left(\frac{1}{KL(P_1,P_2)} + \frac{1}{KL(P_2,P_1)}\right)^{-1}.$$

- **Jensen–Shannon divergence**

Given a number $\beta \in [0,1]$ and $P_1, P_2 \in \mathcal{P}$, let P_3 denote $\beta P_1 + (1-\beta)P_2$.

The **skew divergence** and the **Jensen–Shannon divergence** between P_1 and P_2 are defined on $\mathcal{P}$ as $KL(P_1, P_3)$ and $\beta KL(P_1, P_3) + (1 - \beta)KL(P_2, P_3)$, respectively. Here KL is the **Kullback–Leibler distance**; cf. **clarity similarity**.

In terms of *entropy* $H(P) = -\sum_x p(x) \ln p(x)$, the Jensen–Shannon divergence is $H(\beta P_1 + (1-\beta)P_2) - \beta H(P_1) - (1-\beta)H(P_2)$, i.e., the **Jensen divergence** (cf. **Bregman divergence**).

Let $P_3 = \frac{1}{2}(P_1 + P_2)$, i.e., $\beta = \frac{1}{2}$. Then the skew divergence and twice the Jensen–Shannon divergence are called **K-divergence** and **Topsøe distance** (or *information statistics*), respectively. The Topsøe distance is a symmetric version of $KL(P_1, P_2)$. It is not a metric, but its square root is a metric.

Related symmetric divergencies between P_1 and P_2—**Taneja distance** (1995) and **Kumar–Johnson distance** (2005)—are defined, respectively, by

$$\sum_x \left(\frac{p_1(x) + p_2(x)}{2} \ln \frac{p_1(x) + p_2(x)}{2\sqrt{p_1(x)p_2(x)}} \right) \text{ and } \sum_x \left(\frac{(p_1^2(x) - p_2^2(x))^2}{2\sqrt{p_1(x)p_2(x)}} \right).$$

- **Clarity similarity**

  The **clarity similarity** is a similarity on $\mathcal{P}$, defined by

$$(KL(P_1, P_3) + KL(P_2, P_3)) - (KL(P_1, P_2) + KL(P_2, P_1)) =$$
$$= \sum_x \left(p_1(x) \ln \frac{p_2(x)}{p_3(x)} + p_2(x) \ln \frac{p_1(x)}{p_3(x)} \right),$$

  where KL is the **Kullback–Leibler distance**, and P_3 is a fixed probability law.

  It was introduced in [CCL01] with P_3 being the probability distribution of English.

- **Ali–Silvey distance**

  The **Ali–Silvey distance** is a quasi-distance on $\mathcal{P}$ defined by the functional

$$f(\mathbb{E}_{P_1}[g(L)]),$$

  where $L = \frac{p_1(x)}{p_2(x)}$ is the *likelihood ratio*, f is a nondecreasing function on $\mathbb{R}$, and g is a continuous convex function on $\mathbb{R}_{\geq 0}$ (cf. **f-divergence**).

  The case $f(x) = x$, $g(x) = x \ln x$ corresponds to the **Kullback–Leibler distance**; the case $f(x) = -\ln x$, $g(x) = x^t$ corresponds to the **Chernoff distance**.

- **Chernoff distance**

  The **Chernoff distance** (or *Rényi cross-entropy*) on $\mathcal{P}$ is defined (1954) by

$$\max_{t \in (0,1)} D_t(P_1, P_2),$$

  where $0 < t < 1$ and $D_t(P_1, P_2) = -\ln \sum_x (p_1(x))^t (p_2(x))^{1-t}$ (called the *Chernoff coefficient*) which is proportional to the **Rényi distance**.

[redacted]

- **Rényi distance**

  Given $t \in \mathbb{R}$, the **Rényi distance** (or *order t Rényi entropy*, 1961) is a quasi-distance on $\mathcal{P}$ defined as the **Kullback–Leibler distance** $KL(P_1, P_2)$ if $t = 1$, and, otherwise, by

$$\frac{1}{1-t} \ln \sum_x p_2(x) \left(\frac{p_1(x)}{p_2(x)} \right)^t .$$

  For $t = \frac{1}{2}$, one half of the Rényi distance is the **Bhattacharya distance** 2. Cf. f**-divergence** and **Chernoff distance**.

- **Shannon distance**

  Given a *measure space* $(\Omega, \mathcal{A}, P)$, where the set Ω is finite and P is a probability measure, the *entropy* (or *Shannon information entropy*) of a function $f : \Omega \to X$, where X is a finite set, is defined by

$$H(f) = -\sum_{x \in X} P(f = x) \log_a (P(f = x)).$$

  Here $a = 2$, e, or 10 and the unit of entropy is called a *bit*, *nat*, or *dit* (digit), respectively. The function f can be seen as a partition of the measure space.

  For any two such partitions $f : \Omega \to X$ and $g : \Omega \to Y$, denote by $H(f, g)$ the entropy of the partition $(f, g) : \Omega \to X \times Y$ (*joint entropy*), and by $H(f|g)$ the *conditional entropy* (or *equivocation*). Then the **Shannon distance** between f and g is a metric defined by

$$H(f|g) + H(g|f) = 2H(f, g) - H(f) - H(g) = H(f) + H(g) - 2I(f; g),$$

  where $I(f; g) = H(f) + H(g) - H(f, g)$ is the *Shannon mutual information.*

  If P is the uniform probability law, then Goppa showed that the Shannon distance can be obtained as a limiting case of the **finite subgroup metric**.

  In general, the **shared information distance** (or **variation of information**, **entropy metric**) between random variables (information sources) X and Y is

$$H(X|Y) + H(Y|X) = H(X, Y) - I(X; Y),$$

  where the *conditional entropy* $H(X|Y)$ is $\sum_{x \in X} \sum_{y \in Y} p(x, y) \ln p(x|y)$, and $p(x|y) = P(X = x | Y = y)$ is the conditional probability.

  The **Rajski distance** (or *normalized information metric*) is defined (Rajski, 1961, for discrete probability distributions X, Y) by

$$\frac{H(X|Y) + H(Y|X)}{H(X, Y)} = 1 - \frac{I(X; Y)}{H(X, Y)}.$$

It is equal to 1 if X and Y are independent. Cf. **Tanimoto distance** in Chap. 17 and the **normalized information distance** in Chap. 11.

Now, $\max\{H(X), H(Y)\} - I(X;Y)$ is also a metric.

- **Transportation distance**

  Given a metric space $(\mathcal{X}, d)$, the **transportation distance** (and/or, according to Villani, 2009, **Monge–Kantorovich–Wasserstein–Rubinstein–Ornstein–Gini–Dall'Aglio–Mallows–Tanaka distance**) is the metric defined by

  $$W_1(P_1, P_2) = \inf \mathbb{E}_S[d(X,Y)] = \inf_S \int_{(X,Y)\in\mathcal{X}\times\mathcal{X}} d(X,Y)dS(X,Y),$$

  where the infimum is taken over all joint distributions S of pairs (X, Y) of random variables X, Y such that marginal distributions of X and Y are P_1 and P_2.

  For any **separable** metric space $(\mathcal{X}, d)$, this is equivalent to the **Lipschitz distance between measures** $\sup_f \int f d(P_1 - P_2)$, where the supremum is taken over all functions f with $|f(x) - f(y)| \leq d(x,y)$ for any $x, y \in \mathcal{X}$. Cf. **Dudley metric**.

  In general, for a Borel function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, the **$c$-transportation distance** $T_c(P_1, P_2)$ is $\inf \mathbb{E}_S[c(X,Y)]$. It is the minimal total transportation cost if $c(X, Y)$ is the cost of transporting a unit of mass from the location X to the location Y. Cf. the **Earth Mover's distance** (Chap. 21), which is a discrete form of it.

  The **L_p-Wasserstein distance** is $W_p = (T_{d^p})^{1/p} = (\inf \mathbb{E}_S[d^p(X,Y)])^{1/p}$. For $(\mathcal{X}, d) = (\mathbb{R}, |x-y|)$, it is also called the *L_p-metric between distribution functions* (CDF) F_i with $F_i^{-1}(x) = \sup_u (P_i(X \leq x) < u)$, and can be written as

  $$(\inf \mathbb{E}[|X-Y|^p])^{1/p} = \left(\int_{\mathbb{R}} |F_1(x) - F_2(x)|^p dx\right)^{1/p} = \left(\int_0^1 |F_1^{-1}(x) - F_2^{-1}(x)|^p dx\right)^{1/p}.$$

  For $p = 1$, this metric is called **Monge–Kantorovich metric** (or **Wasserstein metric**, **Fortet–Mourier metric**, **Hutchinson metric**, **Kantorovich–Rubinstein metric**). For $p = 2$, it is the **Levy–Fréchet metric** (Fréchet, 1957).

- **Ornstein $\bar{d}$-metric**

  The **Ornstein $\bar{d}$-metric** is a metric on $\mathcal{P}$ (for $\mathcal{X} = \mathbb{R}^n$) defined (1974) by

  $$\frac{1}{n} \inf \int_{x,y} \left(\sum_{i=1}^{n} 1_{x_i \neq y_i}\right) dS,$$

  where the infimum is taken over all joint distributions S of pairs (X, Y) of random variables X, Y such that marginal distributions of X and Y are P_1 and P_2, respectively.

- **Distances between belief assignments**

  In *Bayesian* (or *subjective*, *evidential*) interpretation, a probability can be assigned to any statement, even if no random process is involved, as a way

[redacted]

to represent its subjective plausibility, or the degree to which it is supported by the available evidence, or, mainly, degree of belief. Within this approach, *imprecise probability* generalizes Probability Theory to deal with scarce, vague, or conflicting information. The main model is *Dempster–Shafer theory*, which allows evidence to be combined.

Given a set X, a (basic) **belief assignment** is a function $m : P(X) \to [0, 1]$ (where $P(X)$ is the set of all subsets of X) with $m(\emptyset) = 0$ and $\sum_{A \subset P(X)} m(A) = 1$. Probability measures are a special case in which $m(A) > 0$ only for singletons.

For the classic probability $P(A)$, it holds then $\mathrm{Bel}(A) \leq P(A) \leq \mathrm{Pl}(A)$, where the *belief function* and *plausibility function* are defined, respectively, by

$$\mathrm{Bel}(A) = \sum_{B:B \subset A} m(B) \text{ and } \mathrm{Pl}(A) = \sum_{B:B \cap A \neq \emptyset} m(B) = 1 - \mathrm{Bel}(\overline{A}).$$

The original (Dempster, 1967) *conflict factor* between two belief assignments m_1 and m_2 was defined as $c(m_1, m_2) = \sum_{A \cap B = \emptyset} m_1(A) m_2(B)$. This is not a distance since $c(m, m) > 0$. The combination of m_1 and m_2, seen as independent sources of evidence, is defined by $m_1 \oplus m_2(A) = \frac{1}{1 - c(m_1, m_2)} \sum_{B \cap C = A} m_1(B) m_2(C)$.

Usually, a distance between m_1 and m_2 estimates the difference between these sources in the form $d_U = |U(m_1) - U(m_2)|$, where U is an uncertainty measure; see Sarabi-Jamab et al., 2013, for a comparison of their performance. In particular, this distance is called:

confusion (Hoehle, 1981) if $U(m) = \sum_A m(A) \log_2 \mathrm{Bel}(A)$;
dissonance (Yager, 1983) if $U(m) = E(m) = -\sum_A m(A) \log_2 \mathrm{Pl}(A)$;
Yager's factor (Eager, 1983) if $U(m) = 1 - \sum_{A \neq \emptyset} \frac{m(A)}{|A|}$;
possibility-based (Smets, 1983) if $U(m) = -\sum_A \log_2 \sum_{B:A \subset B} m(B)$;
U-uncertainty (Dubois–Prade, 1985) if $U(m) = I(m) = -\sum_A m(A) \log_2 |A|$;
Lamata–Moral's (1988) if $U(m) = \log_2(\sum_A m(A)|A|)$ and $U(m) = E(m) + I(m)$;
discord (Klir–Ramer, 1990) if $U(m) = D(m) = -\sum_A m(A) \log_2(1 - \sum_B m(B) \frac{|B \backslash A|}{|B|})$ and a variant: $U(m) = D(m) + I(m)$;
strife (Klir–Parviz, 1992) if $U(m) = -\sum_A m(A) \log_2(\sum_B m(B) \frac{|A \cap B|}{|A|})$;
Pal et al.'s (1993) if $U(m) = G(m) = -\sum_A \log_2 m(A)$ and $U(m) = G(m) + I(m)$;
total conflict (George–Pal, 1996) if $U(m) = \sum_A m(A) \sum_B (m(B)(1 - \frac{|A \cap B|}{|A \cup B|}))$.

Among other distances used are the **cosine distance** $1 - \frac{m_1^T m_2}{||m_1|| ||m_2||}$, the **Mahalanobis distance** $\sqrt{(m_1 - m_2)^T A (m_1 - m_2)}$ for some matrices A, and *pignistic-based* one (Tessem, 1993) $\max_A \{ | \sum_{B \neg \emptyset} (m_1(B) - m_2(B) \frac{|A \cap B|}{|B|} | \}$.

Part IV
Distances in Applied Mathematics

[redacted]

Chapter 15
Distances in Graph Theory

A *graph* is a pair $G = (V, E)$, where V is a set, called the set of *vertices* of the graph G, and E is a set of unordered pairs of vertices, called the *edges* of the graph G. A *directed graph* (or *digraph*) is a pair $D = (V, E)$, where V is a set, called the set of *vertices* of the digraph D, and E is a set of ordered pairs of vertices, called *arcs* of the digraph D.

A graph in which at most one edge may connect any two vertices, is called a *simple graph*. If multiple edges are allowed between vertices, the graph is called a *multigraph*. A graph, together with a function which assigns a positive weight to each edge, is called a *weighted graph* or *network*.

The graph is called *finite* (*infinite*) if the set V of its vertices is finite (infinite, respectively). The *order* and *size* of a finite graph (V, E) are $|V|$ and $|E|$, respectively.

A *subgraph* of a graph $G = (V, E)$ is a graph $G' = (V', E')$ with $V' \subset V$ and $E' \subset E$. If G' is a subgraph of G, then G is called a *supergraph* of G'. A subgraph (V', E') of (V, E) is its *induced subgraph* if $E' = \{e = uv \in E : u, v \in V'\}$.

A graph $G = (V, E)$ is called *connected* if, for any $u, v \in V$, there exists a $(u - v)$ *walk*, i.e., a sequence of edges $uw_1 = w_0w_1, w_1w_2, \dots, w_{n-1}w_n = w_{n-1}v$ from E. A $(u - v)$ *path* is a $(u - v)$ walk with distinct edges. A graph is called *m-connected* if there is no set of $m - 1$ edges whose removal disconnects the graph; a connected graph is 1-connected. A digraph $D = (V, E)$ is called *strongly connected* if, for any $u, v \in V$, the *directed* $(u - v)$ and $(v - u)$ paths both exist. A maximal connected subgraph of a graph G is called its *connected component*.

Vertices connected by an edge are called *adjacent*. The *degree* $deg(v)$ of a vertex $v \in V$ of a graph $G = (V, E)$ is the number of its vertices adjacent to v.

A *complete graph* is a graph in which each pair of vertices is connected by an edge. A *bipartite graph* is a graph in which the set V of vertices is decomposed into two disjoint subsets so that no two vertices within the same subset are adjacent. A *simple path* is a simple connected graph in which two vertices have degree 1, and

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_15

other vertices (if they exist) have degree 2; the *length* of a path is the number of its edges.

A *cycle* is a *closed simple path*, i.e., a simple connected graph in which every vertex has degree 2. The *circumference* of a graph is the length of the longest cycle in it. A *tree* is a simple connected graph without cycles. A tree having a path from which every vertex has distance ≤ 1 or ≤ 2, is called a *caterpillar* or *lobster*, respectively.

Two graphs which contain the same number of vertices connected in the same way are called *isomorphic*. Formally, two graphs $G = (V(G), E(G))$ and $H = (V(H), E(H))$ are called *isomorphic* if there is a bijection $f : V(G) \to V(H)$ such that, for any $u, v \in V(G)$, $uv \in E(G)$ if and only if $f(u)f(v) \in E(H)$.

We will consider mainly simple finite graphs and digraphs; more exactly, the equivalence classes of such isomorphic graphs.

15.1 Distances on the Vertices of a Graph

- **Path metric**

  The **path metric** (or **graphic metric**, *shortest path metric*) d_{path} is a metric on the vertex-set V of a connected graph $G = (V, E)$ defined, for any $u, v \in V$, as the length of a shortest $(u - v)$ path in G, i.e., a *geodesic*. Examples follow.

  Given an integer $n \geq 1$, the **line metric on** $\{1, \dots, n\}$ in Chap. 1 is the path metric of the path $P_n = \{1, \dots, n\}$. The path metric of the *Cayley graph* Γ of a finitely generated group $(G, \cdot, e)$ is called a **word metric**.

  The **hypercube metric** is the path metric of a *hypercube graph* $H(m, 2)$ with the vertex-set $V = \{0, 1\}^m$, and whose edges are the pairs of vectors $x, y \in \{0, 1\}^m$ such that $|\{i \in \{1, \dots, n\} : x_i \neq y_i\}| = 1$; it is equal to $|\{i \in \{1, \dots, n\} : x_i = 1\} \triangle \{i \in \{1, \dots, n\} : y_i = 1\}|$. The graphic metric space associated with a hypercube graph coincides with a **Hamming cube**, i.e., the metric space $(\{0, 1\}^m, d_{l_1})$.

  The **belt distance** (Garber–Dolbilin, 2010) is the path metric of a *belt graph* $B(P)$ of a polytope P with centrally symmetric facets. The vertices of $B(P)$ are the facets of P and two vertices are connected by an edge if the corresponding facets lie in the same *belt* (the set of all facets of P parallel to a given face of codimension 2).

  The reciprocal path metric is called **geodesic similarity**.

- **Weighted path metric**

  The **weighted path metric** d_{wpath} is a metric on the vertex-set V of a connected weighted graph $G = (V, E)$ with positive edge-weights $(w(e))_{e \in E}$ defined by

$$\min_P \sum_{e \in P} w(e),$$

  where the minimum is taken over all $(u - v)$ paths P in G.

[redacted]

Sometimes, $\frac{1}{w(e)}$ is called the *length* of the edge e. In the theory of electrical networks, the edge-length $\frac{1}{w(e)}$ is identified with the *resistance* of the edge e. The **inverse weighted path metric** is $\min_P \sum_{e \in P} \frac{1}{w(e)}$.

- **Filling of metric space**

Given a finite metric space (X, d), a connected weighted graph $G = (V, E; \omega)$ with $X \subset V$ and edge-weights $\omega(e) \geq 0$ is called (Ivanov–Tuzhilin, 2012) **filling** of (X, d) if for all $x, y \in X$, $d(x, y)$ is bounded from above by the **weighted path metric** $d_G(x, y)$. Ivanov–Tuzhilin showed the existence of filings with minimal $\sum_{e \in E} \omega(e)$ and some relations with Steiner problem; (cf. **Steiner ratio** in Sect. 1.3.

For the case when X is Riemannian manifold, a similar notion was defined by Gromov (1983), and related to Systolic Geometry (cf. **systole of metric space** in Sect. 1.2). Here a *minimal filing* is any compact manifold of minimal volume, having boundary X and a distance function that bounds d on X from above.

- **Metric graph**

A **metric** (or *metrized*) **graph** is a connected graph $G = (V, E)$, where edges e are identified with line segments $[0, l(e)]$ of length $l(e)$. Let x_e be the coordinate on the segment $[0, l(e)]$ with vertices corresponding to $x_e = 0, l(e)$; the ends of distinct segments are identified if they correspond to the same vertex of G. A *function* f on G is the $|E|$-tuple of functions $f_e(x_e)$ on the segments.

A metric graph can be seen as an infinite metric space (X, d), where X is the set of all points on above segments, and the distance between two points is the length of the shortest, along the line segments traversed, path connecting them. Also, it can be seen as one-dimensional Riemannian manifold with singularities.

There is a bijection between the metric graphs, the equivalence classes of finite connected edge-weighted graphs and the resistive electrical networks: if an edge e of a metric graph has length $l(e)$, then $\frac{1}{l(e)}$ is the weight of e in the corresponding edge-weighted graph and $l(e)$ is the resistance along e in the corresponding resistive electric circuit. Cf. the **resistance metric**.

A **quantum graph** is a metric graph equipped with a self-adjoint differential operator (such as a *Laplacian*) acting on functions on the graph. The *Hilbert space* of the graph is $\oplus_{e \in E} L^2([0, w(e)])$, where the inner product of functions is $\langle f, g \rangle = \sum_{e \in E} \int_0^{w(e)} f_e^*(x_e) g_e(x_e) dx_e$.

- **Spin network**

A **spin network** is (Penrose, 1971) a connected graph (V, E) with edge-weights $(w(e))_{e \in E}$ (*spins*), $w(e) \in \mathbb{N}$, such that for any distinct edges e_1, e_2, e_3 with a common vertex, it holds **spin triangle inequality** $|w(e_1) - w(e_2)| \leq w(e_3) \leq w(e_1) + w(e_2)$ and *fermion conservation*: $w(e_1) + w(e_2) + w(e_3)$ is an even number.

The **quantum space-time** (Chap. 24) in *Loop Quantum Gravity* is a network of loops at Planck scale. Loops are represented by adapted spin networks: directed graphs whose arcs are labeled by irreducible representations of a compact Lie group and vertices are labeled by *interwinning operators* from the tensor product of labels on incoming arcs to the tensor product of labels on outgoing arcs. Such networks represent "quantum states" of the gravitational field on a 3D hypersurface.

- **Detour distance**

  Given a connected graph $G = (V, E)$, the **detour distance** is (Chartrand and Zhang, 2004) a metric on the vertex-set V defined, for $u \neq v$, as the length of the longest $(u - v)$ path in G. So, this distance is 1 or $|V| - 1$ if and only if uv is a *bridge* of G or, respectively, G contains a Hamiltonian $(u - v)$ path.

  The **monophonic distance** is (Santhakumaran and Titus, 2011) a distance (in general, not a metric) on the V defined, for $u \neq v$, as the length of a longest *monophonic* (or *induced*, *minimal*), i.e., chordless $(u - v)$ path in G.

  The *height* of a DAG (acyclic digraph) is the number of vertices in a longest directed path.

- **Rainbow distance**

  In an edge-colored graph, the **rainbow distance** is (Chartrand and Zhang, 2005) the length of a shortest *rainbow* (i.e., containing no color twice) path.

  In a vertex-colored graph, the **colored distance** is (Dankelmann et al., 2001) the sum of distances between all unordered pairs of vertices having different colors.

- **Cutpoint additive metric**

  Given a graph $G = (V, E)$, Klein–Zhu, 1998, call a metric d on V **graph-geodetic metric** if, for $u, w, v \in V$, the **triangle equality** $d(u, w) + d(w, v) = d(u, v)$ holds if w is a (u, v)-*gatekeeper*, i.e., w lies on any path connecting u and v. Cf. **metric interval** in Chap. 1. Any gatekeeper is a *cutpoint*, i.e., removing it disconnects G and a *pivotal point*, i.e., it lies on any shortest path between u and v.

  Chebotarev, 2010, call a metric d on the vertices of a multigraph without loops **cutpoint additive** if $d(u, w) + d(w, v) = d(u, v)$ holds if and only if w lies on any path connecting u and v. The **resistance metric** is cutpoint additive (Gvishiani and Gurvich, 1992), while the **path metric** is graph-geodetic only (in the weaker Klein–Zhu sense). See also **Chebotarev–Shamis metric**.

- **Graph boundary**

  Given a connected graph $G = (V, E)$, a vertex $v \in V$ is (Chartrand et al., 2003) a *boundary vertex* if there exists a *witness*, i.e., a vertex $u \in V$ such that $d(u, v) \geq d(u, w)$ for all neighbors w of v. So, the end-vertices of a longest path are boundary vertices. The **boundary** of G is the set of all boundary vertices.

  The *boundary of a subset* $M \subset V$ is the set $\partial M \subset E$ of edges having precisely one endpoint in M. The **isoperimetric number** of G is (Buser, 1978) $\inf \frac{\partial M}{|M|}$, where the infimum is taken over all $M \subset V$ with $2|M| \leq |V|$.

- **Graph diameter**

  Given a connected graph $G = (V, E)$, its **graph diameter** is the largest value of the **path metric** between vertices of G.

  A connected graph is *vertex-critical* (*edge-critical*) if deleting any vertex (edge) increases its diameter. A graph G of diameter k is *goal-minimal* if for every edge uv, the inequality $d_{G-uv}(x, y) > k$ holds if and only if $\{u, v\} = \{x, y\}$.

[redacted]

The **mixed fault diameter** $D_{(a,b)}(G)$ is the maximum diameter among all subgraphs obtained from G by deleting a vertices and b edges.

The **a-wide distance** $d_a(u, v)$ between vertices u and v is the minimum integer l, for which there are at least a internally disjoint $(u - v)$ paths of length at most l in G: cf. **Hsu–Lyuu–Flandrin–Li distance**. The *a-wide diameter* of G is $\max_{u,v \in V} d_a(u, v)$; it is at least $D_{(a-1,0)}(G)$.

Given a *strong orientation O* of a connected graph $G = (V, E)$, i.e., a strongly connected digraph $D = (V, E')$ with arcs $e' \in E'$ obtained from edges $e \in E$ by orientation O, the *diameter* of D is the maximal length of shortest directed $(u-v)$ path in it. The **oriented diameter** of a graph G is the smallest diameter among strong orientations of G. If it is equal to the diameter of G, then any orientation realizing this equality is called *tight*. For example, a *hypercube graph $H(m, 2)$* admits a tight orientation if $m \geq 4$ (McCanna, 1988).

- **Path quasi-metric in digraphs**

  The **path quasi-metric in digraphs** d_{dpath} is a quasi-metric on the vertex-set V of a strongly connected digraph $D = (V, E)$ defined, for any $u, v \in V$, as the length of a shortest directed $(u - v)$ path in D.

  The **circular metric in digraphs** is a metric on the vertex-set V of a strongly connected digraph $D = (V, E)$, defined by $d_{\text{dpath}}(u, v) + d_{\text{dpath}}(v, u)$.

- **Strong distance in digraphs**

  The **strong distance in digraphs** is a metric between vertices v and v of a strongly connected digraph $D = (V, E)$ defined (Chartrand–Erwin–Raines–Zhang, 1999) as the minimum *size* (the number of edges) of a strongly connected subdigraph of D containing v and v. Cf. **Steiner distance of a set**.

- **Υ-metric**

  Given a class Υ of connected graphs, the metric d of a metric space (X, d) is called a **Υ-metric** if (X, d) is isometric to a subspace of a metric space (V, d_{wpath}), where $G = (V, E) \in \Upsilon$, and d_{wpath} is the **weighted path metric** on V with positive edge-weight function w; cf. **tree-like metric**.

- **Tree-like metric**

  A **tree-like metric** (or **weighted tree metric**) d on a set X is a **Υ-metric** for the class Υ of all trees, i.e., the metric space (X, d) is isometric to a subspace of a metric space (V, d_{wpath}), where $T = (V, E)$ is a tree, and d_{wpath} is the **weighted path metric** on the vertex-set V of T with a positive weight function w. A metric is a tree-like metric if and only if it satisfies the **four-point inequality**. Any such metric has (Hendy, 1992) a unique tree representation.

  A metric d on a set X is called a **relaxed tree-like metric** if the set X can be embedded in some (not necessary positively) edge-weighted tree such that, for any $x, y \in X$, $d(x, y)$ is equal to the sum of all edge weights along the (unique) path between corresponding vertices x and y in the tree. A metric is a relaxed tree-like metric if and only if it satisfies the **relaxed four-point inequality**.

- **Katz similarity**

  Given a connected graph $G = (V, E)$ with positive edge-weight function $w = (w(e))_{e \in E}$, let $V = \{v_1, \dots, v_n\}$. Denote by A the $(n \times n)$-matrix $((a_{ij}))$, where $a_{ij} = a_{ji} = w(ij)$ if ij is an edge, and $a_{ij} = 0$, otherwise. Let I be the identity

$(n \times n)$-matrix, and let $t, 0 < t < \frac{1}{\lambda}$, be a parameter, where $\lambda = \max_i |\lambda_i|$ is the *spectral radius* of A and λ_i are the eigenvalues of A. Define the $(n \times n)$-matrix

$$K = ((k_{ij})) = \sum_{i=1}^{\infty} t^i A^i = (I - tA)^{-1} - I.$$

The number k_{ij} is called the **Katz similarity** between v_i and v_j. Katz, 1953, proposed it for evaluating social status.

Chebotarev, 2011, defined, for a similar $(n \times n)$-matrix $((c_{ij})) = \sum_{i=0}^{\infty} t^i A^i = (I - tA)^{-1}$ and connected edge-weighted multigraphs allowing loops, the **walk distance** between vertices v_i and v_j as any positive multiple of $d_t(i,j) = -\ln \frac{c_{ij}}{\sqrt{c_{ii}\, c_{jj}}}$ (cf. the **Nei standard genetic distance** in Chap. 23). He proved that d_t is a **cutpoint additive metric** and the **path metric** in G coincides with the *short walk distance* $\lim_{t \to 0^+} \frac{d_t}{-\ln t}$ in G, while the **resistance metric** in G coincides with the *long walk distance* $\lim_{t \to \frac{1}{\lambda}^-} \frac{2d_t}{n(t^{-1} - \lambda)}$ in the graph G' obtained from G by attaching weighted loops that provide G' with uniform weighted degrees.

If G is a simple unweighted graph, then A is its adjacency matrix. Let J be the $(n \times n)$-matrix of all ones and let $\mu = \min_i \lambda_i$. Let $N = ((n_{ij})) = \mu(I - J) - A$. Neumaier, 1980, remarked that $((\sqrt{n_{ij}}))$ is a semimetric on the vertices of G.

- **Resistance metric**

Given a connected graph $G = (V, E)$ with positive edge-weight function $w = (w(e))_{e \in E}$, let us interpret the edge-weights as electrical conductances and their inverses as resistances. For any two different vertices u and v, suppose that a battery is connected across them, so that one unit of a current flows in at u and out in v. The voltage (potential) difference, required for this, is, by Ohm's law, the effective resistance between u and v in an electrical network; it is called the **resistance** (or *electric*) **metric** $\Omega(u, v)$ between them (Sharpe, 1967, Gvishiani–Gurvich, 1987, and Klein–Randic, 1993). So, if a potential of one volt is applied across vertices u and v, a current of $\frac{1}{\Omega(u,v)}$ will flow. The number $\frac{1}{\Omega(u,v)}$ is a measure of the *connectivity* between u and v.

Let $r(u, v) = \frac{1}{w(e)}$ if uv is an edge, and $r(u, v) = 0$, otherwise. Formally,

$$\Omega(u, v) = \Big(\sum_{w \in V} f(w) r(w, v)\Big)^{-1},$$

where $f : V \to [0, 1]$ is the unique function with $f(u) = 1, f(v) = 0$ and $\sum_{z \in V}(f(w) - f(z)) r(w, z) = 0$ for any $w \neq u, v$.

The resistance metric is a weighted average of the lengths of all $(u - v)$ paths. It is applied when the number of $(u - v)$ paths, for any $u, v \in V$, matters.

A probabilistic interpretation (Gobel–Jagers, 1974) is: $\Omega(u, v) = (deg(u) Pr(u \to v))^{-1}$, where $deg(u)$ is the degree of the vertex u, and $Pr(u \to v)$ is the probability for a random walk leaving u to arrive at v before returning to u. The expected commuting time between u and v is $2 \sum_{e \in E} w(e) \Omega(u, v)$.

[redacted]

Then $\Omega(u, v) \le \min_P \sum_{e \in P} \frac{1}{w(e)}$, where P is any $(u - v)$ path (cf. **inverse weighted path metric**), with equality if and only if such a path P is unique. So, if $w(e) = 1$ for all edges, the equality means that G is a **geodetic graph**, and hence the path and resistance metrics coincide. Also, it holds that $\Omega(u, v) = \frac{|\{t : uv \in t \in T\}|}{|T|}$ if uv is an edge, and $\Omega(u, v) = \frac{|T' - T|}{|T|}$, otherwise, where T, T' are the sets of spanning trees for $G = (V, E)$ and $G' = (V, E \cup \{uv\})$.

If $w(e) = 1$ for all edges, then $\Omega(u, v) = (g_{uu} + g_{vv}) - (g_{uv} + g_{vu})$, where $((g_{ij}))$ is the Moore–Penrose *generalized inverse* of the *Laplacian matrix* $((l_{ij}))$ of the graph G: here l_{ii} is the degree of vertex i, while, for $i \neq j$, $l_{ij} = -1$ if the vertices i and j are adjacent, and $l_{ij} = 0$, otherwise. A symmetric (for an undirected graph) and positive-semidefinite matrix $((g_{ij}))$ admits a representation KK^T. So, $\Omega(u, v)$ is the squared Euclidean distance between the u-th and v-th rows of K.

The distance $\sqrt{\Omega(u, v)}$ is a **Mahalanobis distance** (Chap. 17) with a weighting matrix $((g_{ij}))$. So, $\Omega_{u,v} = a_{uv}|((g_{ij}))|a_{uv}$, where a_{uv} are the vectors of zeros except for $+1$ and -1 in the u-th and v-th positions. This distance is called a *diffusion metric* in [CLMNWZ05] because it depends on a random walk.

The number $\frac{1}{2} \sum_{u,v \in V} \Omega(u, v)$ is called the *total resistance* (or *Kirchhoff index*) of G.

- **Hitting time quasi-metric**

Let $G = (V, E)$ be a connected graph. Consider random walks on G, where at each step the walk moves to a vertex randomly with uniform probability from the neighbors of the current vertex. The **hitting** (or *first-passage*) **time quasi-metric** $H(u, v)$ from $u \in V$ to $v \in V$ is the expected number of steps (edges) for a random walk on G beginning at u to reach v for the first time; it is 0 for $u = v$. This quasi-metric is a **weightable quasi-semimetric** (Chap. 1).

The **commuting time metric** is $C(u, v) = H(u, v) + H(v, u)$.

Then $C(u, v) = 2|E|\Omega(u, v)$, where $\Omega(u, v)$ is the **resistance metric** (or *effective resistance*), i.e., 0 if $u = v$ and, otherwise, $\frac{1}{\Omega(u,v)}$ is the current flowing into v, when grounding v and applying a 1 volt potential to u (each edge is seen as a resistor of 1 ohm). Also, $\Omega(u, v) = \sup_{f : V \to \mathbb{R}, D(f) > 0} \frac{(f(u) - f(v))^2}{DE(f)}$, where $DE(f)$ is the *Dirichlet energy* of f, i.e., $\sum_{st \in E} (f(s) - f(t))^2$.

The above setting can be generalized to weighted digraphs $D = (V, E)$ with arc-weights c_{ij} for $ij \in E$ and the *cost* of a directed $(u - v)$ path being the sum of the weights of its arcs. Consider the random walk on D, where at each step the walk moves by arc ij with *reference probability* p_{ij} proportional to $\frac{1}{c_{ij}}$; set $p_{ij} = 0$ if $ij \notin E$. Saerens et al., 2008, defined the *randomized* et al. *shortest path quasi-distance* $d(u, v)$ on vertices of D as the minimum expected cost of a directed $(u - v)$ path in the probability distribution minimizing the expected cost among all distributions having a fixed **Kullback–Leibler distance** (Chap. 14) with reference probability distribution. In fact, their biased random walk model depends on a parameter $\theta \ge 0$. For $\theta = 0$ and large θ, the distance $d(u, v) + d(v, u)$ become a metric; it is proportional to the commuting time and the usual path metric, respectively.

- **Chebotarev–Shamis metric**

  Given $\alpha > 0$ and a connected weighted *multigraph* $G = (V, E; w)$ with positive edge-weight function $w = (w(e))_{e \in E}$, denote by $L = ((l_{ij}))$ the *Laplacian* (or *Kirchhoff*) matrix of G, i.e., $l_{ij} = -w(ij)$ for $i \neq j$ and $l_{ii} = \sum_{j \neq i} w(ij)$. The **Chebotarev–Shamis metric** $d_{\alpha}(u, v)$ (Chebotarev and Shamis, 2000, called $\frac{1}{2} d_{\alpha}(u, v)$ **α-forest metric**) between vertices u and v is defined by

  $$2q_{uv} - q_{uu} - q_{vv}$$

  for the **protometric** $((g_{ij})) = -(I + \alpha L)^{-1}$, where I is the identity matrix.

  Chebotarev and Shamis showed that their metric of $G = (V, E; w)$ is the **resistance metric** of another weighted multigraph, $G' = (V', E'; w')$, where $V' = V \cup \{0\}$, $E' = E \cup \{u0 : u \in V\}$, while $w'(e) = \alpha w(e)$ for all $e \in E$ and $w'(u0) = 1$ for all $u \in V$. In fact, there is a bijection between the forests of G and trees of G'. This metric becomes the resistance metric of $G = (V, E; w)$ as $\alpha \to \infty$.

  Their **forest metric** (1997) is the case $\alpha = 1$ of the α-forest metric.

  Chebotarev, 2010, remarked that $2 \ln q_{uv} - \ln q_{uu} - \ln q_{vv}$ is a **cutpoint additive metric** $d''_{\alpha}(u, v)$, i.e., $d''_{\alpha}(u, w) + d''_{\alpha}(w, v) = d''_{\alpha}(u, v)$ holds if and only if w lies on any path connecting u and v. The metric d''_{α} is the **path metric** if $\alpha \to 0^{+}$ and the **resistance metric** if $\alpha \to \infty$.
- **Truncated metric**

  The **truncated metric** is a metric on the vertex-set of a graph, which is equal to 1 for any two adjacent vertices, and is equal to 2 for any nonadjacent different vertices. It is the 2**-truncated metric** for the path metric of the graph. It is the $(1, 2) - B$**-metric** if the degree of any vertex is at most B.
- **Hsu–Lyuu–Flandrin–Li distance**

  Given an *m-connected* graph $G = (V, E)$ and two vertices $u, v \in V$, a *container* $C(u, v)$ *of width* m is a set of m $(u - v)$ paths with any two of them intersecting only in u and v. The *length of a container* is the length of the longest path in it.

  The **Hsu–Lyuu–Flandrin–Li distance** between vertices u and v (Hsu–Lyuu, 1991, and Flandrin–Li, 1994) is the minimum of container lengths taken over all containers $C(u, v)$ of width m. This generalization of the path metric is used in parallel architectures for interconnection networks.
- **Multiply-sure distance**

  The **multiply-sure distance** is a distance on the vertex-set V of an *m-connected* weighted graph $G = (V, E)$, defined, for any $u, v \in V$, as the minimum weighted sum of lengths of m disjoint $(u - v)$ paths. This generalization of the path metric helps when several disjoint paths between two points are needed, for example, in communication networks, where $m - 1$ of $(u - v)$ paths are used to code the message sent by the remaining $(u - v)$ path (see [McCa97]).

[redacted]

- **Cut semimetric**

  A *cut* is a *partition* of a set into two parts. Given a subset S of $V_n = \{1, \dots, n\}$, we obtain the partition $\{S, V_n \backslash S\}$ of V_n. The **cut semimetric** (or **split semimetric**) δ_S defined by this partition, is a semimetric on V_n defined by

  $$\delta_S(i,j) = \begin{cases} 1, & \text{if } i \neq j, |S \cap \{i,j\}| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

  Usually, it is considered as a vector in $\mathbb{R}^{|E_n|}$, $E(n) = \{\{i,j\} : 1 \leq i < j \leq n\}$.

  A *circular cut* of V_n is defined by a subset $S_{[k+1,l]} = \{k+1, \dots, l\} (\mod n) \subset V_n$: if we consider the points $\{1, \dots, n\}$ as being ordered along a circle in that circular order, then $S_{[k+1,l]}$ is the set of its consecutive vertices from $k+1$ to l. For a circular cut, the corresponding cut semimetric is called a **circular cut semimetric**.

  An **even cut semimetric** (**odd cut semimetric**) is δ_S on V_n with even (odd, respectively) $|S|$. A **k-uniform cut semimetric** is δ_S on V_n with $|S| \in \{k, n-k\}$. An **equicut semimetric** (**inequicut semimetric**) is δ_S on V_n with $|S| \in \{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil\}$ ($|S| \notin \{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil\}$, respectively); see, for example, [DeLa97].
- **Decomposable semimetric**

  A **decomposable semimetric** is a semimetric on $V_n = \{1, \dots, n\}$ which can be represented as a nonnegative linear combination of **cut semimetrics**. The set of all decomposable semimetrics on V_n is a *convex cone*, called the *cut cone* CUT_n.

  A semimetric on V_n is decomposable if and only if it is a **finite l_1-semimetric**.

  A **circular decomposable semimetric** is a semimetric on $V_n = \{1, \dots, n\}$ which can be represented as a nonnegative linear combination of **circular cut semimetrics**. A semimetric on V_n is circular decomposable if and only if it is a **Kalmanson semimetric** with respect to the same ordering (see [ChFi98]).
- **Finite l_p-semimetric**

  A **finite l_p-semimetric** d is a semimetric on $V_n = \{1, \dots, n\}$ such that (V_n, d) is a semimetric subspace of the l_p^m*-space* $(\mathbb{R}^m, d_{l_p})$ for some $m \in \mathbb{N}$.

  If, instead of V_n, is taken $X = \{0,1\}^n$, the metric space (X, d) is called the l_p^n*-cube*. The l_1^n-cube is called a **Hamming cube**; cf. Chap. 4. It is the graphic metric space associated with a hypercube graph $H(n,2)$, and any subspace of it is called a **partial cube**.
- **Kalmanson semimetric**

  A **Kalmanson semimetric** d *with respect to the ordering* $1, \dots, n$ is a semimetric on $V_n = \{1, \dots, n\}$ which satisfies the condition

  $$\max\{d(i,j) + d(r,s), d(i,s) + d(j,r)\} \leq d(i,r) + d(j,s)$$

  for all $1 \leq i \leq j \leq r \leq s \leq n$.

  Equivalently, if the points $\{1, \dots, n\}$ are ordered along a circle C_n in that circular order, then the distance d on V_n is a Kalmanson semimetric if the inequality

$$d(i,r) + d(j,s) \leq d(i,j) + d(r,s)$$

holds for $i, j, r, s \in V_n$ whenever the segments $[i,j]$, $[r,s]$ are crossing chords of C_n.

A **tree-like metric** is a Kalmanson metric for some ordering of the vertices of the tree. The Euclidean metric, restricted to the points that form a convex polygon in the plane, is a Kalmanson metric.

- **Multicut semimetric**

  Let $\{S_1, \dots, S_q\}$, $q \geq 2$, be a *partition* of the set $V_n = \{1, \dots, n\}$, i.e., a collection $S_1, \dots, S_q$ of pairwise disjoint subsets of V_n such that $S_1 \cup \cdots \cup S_q = V_n$.

  The **multicut semimetric** $\delta_{S_1,\dots,S_q}$ is a semimetric on V_n defined by

$$\delta_{S_1,\dots,S_q}(i,j) = \begin{cases} 0, & \text{if } i,j \in S_h \text{ for some } h, 1 \leq h \leq q, \\ 1, & \text{otherwise.} \end{cases}$$

- **Oriented cut quasi-semimetric**

  Given a subset S of $V_n = \{1, \dots, n\}$, the **oriented cut quasi-semimetric** δ'_S is a quasi-semimetric on V_n defined by

$$\delta'_S(i,j) = \begin{cases} 1, & \text{if } i \in S, j \notin S, \\ 0, & \text{otherwise.} \end{cases}$$

  Usually, it is considered as the vector of $\mathbb{R}^{|I_n|}$, $I(n) = \{(i,j) : 1 \leq i \neq j \leq n\}$. The **cut semimetric** δ_S is $\delta'_S + \delta'_{V_n \setminus S}$.

- **Oriented multicut quasi-semimetric**

  Given a *partition* $\{S_1, \dots, S_q\}$, $q \geq 2$, of V_n, the **oriented multicut quasi-semimetric** $\delta'_{S_1,\dots,S_q}$ is a quasi-semimetric on V_n defined by

$$\delta'_{S_1,\dots,S_n}(i,j) = \begin{cases} 1, & \text{if} \quad i \in S_h, j \in S_m, h < m, \\ 0, & \text{otherwise} \quad . \end{cases}$$

15.2 Distance-Defined Graphs

Below we first give some graphs defined in terms of distances between their vertices. Then some graphs associated with metric spaces are presented.

A graph (V, E) is, say, *distance-invariant* or *distance monotone* if its metric space (V, d_{path}) is **distance invariant** or **distance monotone**, respectively (cf. Chap. 1). The definitions of such graphs, being straightforward subcases of corresponding metric spaces, will be not given below.

[redacted]

- ***k*-power of a graph**

  The ***k*-power** of a graph $G = (V, E)$ is the supergraph $G^k = (V, E')$ of G with edges between all pairs of vertices having path distance at most k.

- **Distance-residual subgraph**

  For a connected finite graph $G = (V, E)$ and a set $M \subset V$ of its vertices, a **distance-residual subgraph** is (Luksic and Pisanski, 2010) a subgraph induced on the set of vertices u of G at the maximal **point-set distance** $\min_{v \in M} d_{\text{path}}(u, v)$ from M. Such a subgraph is called *vertex-residual* if M consists of a vertex, and *edge-residual* if M consists of two adjacent vertices.

- **Isometric subgraph**

  A subgraph H of a graph $G = (V, E)$ is called an **isometric subgraph** if the path metric between any two points of H is the same as their path metric in G.

  A subgraph H is called a *convex subgraph* if it is isometric, and for any $u, v \in H$ every vertex on a shortest $(u - v)$ path belonging to H also belongs to H. Taking, instead of shortest, other types of paths (say, any, longest, induced) or walks, one gets corresponding notions of convexity.

  A subset $M \subset V$ is called *gated* if for every $u \in V \setminus M$ there exists a unique vertex $g \in M$ (called a *gate*) lying on a shortest $(u - v)$ path for every $v \in M$. The subgraph induced by a gated set is a convex subgraph.

- **Retract subgraph**

  A subgraph H of G is called a **retract subgraph** if it is induced by an idempotent **metric mapping** of G into itself, i.e., $f^2 = f : V \to V$ with $d_{\text{path}}(f(u), f(v)) \leq d_{\text{path}}(u, v)$ for $u, v \in V$. Any retract subgraph is isometric.

- **Partial cube**

  A **partial cube** is an **isometric subgraph** of a **Hamming cube**, i.e., of a hypercube $H(m, 2)$. Similar topological notion was introduced by Acharya, 1983: any graph (V, E) admits a *set-indexing* $f : V \cup E \to 2^X$ with injective $f|_V, f|_R$ and $f(uv) = f(u) \Delta f(v)$ for any $(uv) \in E$. The *set-indexing number* is $\min |X|$.

- **Median graph**

  A connected graph $G = (V, E)$ is called a **median graph** if, for any three vertices $u, v, w \in V$, there exists a unique vertex that lies simultaneously on a shortest $(u - v)$, $(u - w)$ and $(w - v)$ paths, i.e., (V, d_{path}) is a **median metric space**.

  The median graphs are exactly **retract subgraphs** of hypercubes and exactly 1-skeletons of a **CAT(0)** cube complexes. Also, they are exactly **partial cubes** such that the vertex-set of any *convex subgraph* is *gated* (cf. **isometric subgraph**).

- **Geodetic graph**

  A graph is called **geodetic** if there exists at most one shortest path between any two of its vertices. A graph is called *strongly geodetic* if there exists at most one path of length less than or equal to the diameter between any two of its vertices.

  A *uniformly geodetic graph* is a connected graph such that the number of shortest paths between any two vertices u and v depends only on $d(u, v)$.

  A graph is a *forest* (disjoint union of trees) if and only if there exists at most one path between any two of its vertices.

The *geodetic number* of a finite connected graph (V, E) ([BuHa90]) is min $|M|$ over sets $M \subset V$ such that any $x \in V$ lies on a shortest $(u-v)$ path with $u, v \in M$.

- ***k*-geodetically connected graph**

  A k-connected graph is called (Entringer–Jackson–Slater, 1977) ***k*-geodetically connected** $(k - GC)$ if the removal of less than k vertices (or, equivalently, edges) does not affect the **path metric** between any pair of the remaining vertices.

  $2 - GC$ graphs are called *self-repairing*. Cf. **Hsu–Lyuu–Flandrin–Li distance**.

- **Interval distance monotone graph**

  A connected graph $G = (V, E)$ is called **interval distance monotone** if any of its intervals $I_G(u, v)$ induces a *distance monotone graph*, i.e., its path metric is **distance monotone**, cf. Chap. 1.

  A graph is interval distance monotone if and only if (Zhang–Wang, 2007) each of its intervals is isomorphic to either a path, a cycle or a hypercube.

- **Distance-regular graph**

  A connected *regular* (i.e., every vertex has the same degree) graph $G = (V, E)$ of diameter T is called **distance-regular** (or *drg*) if, for every two its vertices u, v and any integers $0 \leq i, j \leq T$, the number $|\{w \in V : d_{\text{path}}(u, w) = i, d_{\text{path}}(v, w) = j\}|$ depends only on i, j and $k = d_{\text{path}}(u, v)$, but not on the choice of u and v.

  A special case of it is a **distance-transitive graph**, i.e., such that its group of automorphisms is transitive, for any $0 \leq i \leq T$, on the pairs of vertices (u, v) with $d_{\text{path}}(u, v) = i$. Given a number s, $1 \leq s \leq T$, a graph is said to be *s-geodesic-transitive* or *s-distance transitive* if its automorphism group acts transitively on the sets of *i-geodesics* (shortest paths of length i), for $0 \leq i \leq s$, or, respectively, on the sets of ordered vertex pairs at distance i, for $0 \leq i \leq s$.

  Any drg is a **distance-balanced graph** (or *dbg*), i.e., $|W_{u,v}| = |W_{v,u}|$, where $W_{u,v} = \{x \in V : d(x, u) < d(x, v)\}$. Such graph is also called *self-median* since it is exactly one, **metric median** (cf. **eccentricity** in Chap. 1) of which is V. A gbg is called *nicely distance-balanced* if $|W_{u,v}|$ is the same for all edges uv.

  Any drg is a **distance degree-regular graph** (i.e., $|\{x \in V : d(x, u) = i\}|$ depends only on i; such graph is also called *strongly distance-balanced*), and a **walk-regular graph** (i.e., the number of closed walks of length i starting at u depends only on i). van Dam–Omidi, 2013, call a graph *strongly walk-regular* if there is an $l \geq 2$ such that the number of walks of length l from u to v depends only on whether the $d(u, v)$ is $0, 1$, or ≥ 2; for $l = 2$, it is a *strongly regular graph*, i.e., a drg of diameter 2. A *d-Deza graph* (Gu, 2013) is a regular graph (V, E) in which there are exactly d different values of $|\{w \in V : d(u, w) = d(v, w) = 1\}|$ for distinct $u, v \in V$.

  Any drg is a **sphere-regular graph** (Dehmer, 2008), i.e., its path metric is **distance-invariant** (all closed balls of the same radius have the same size; cf. Chap. 1).

  A graph G is a **distance-regularized graph** if for each $u \in V$, if admits an *intersection array at vertex u*, i.e., the numbers $a_i(u) = |G_i(u) \cap G_1(v)|$, $b_i(u) = |G_{i+1}(u) \cap G_1(v)|$ and $c_i(u) = |G_{i-1}(v) \cap G_1(v)|$ depend only on the

[redacted]

distance $d(u, v) = i$ and are independent of the choice of the vertex $v \in G_i(u)$. Here, for any i, $G_i(w)$ is the set of all vertices at the distance i from w. Godsil–Shawe-Taylor, 1987, defined such graph and proved that it is either drg or *distance-biregular* (a bipartite one with vertices in the same class having the same intersection array).

A drg is also called a **metric association scheme** or *P-polynomial association scheme*. A finite **polynomial metric space** (Chap. 1) is a special case of it, also called a *(P and Q)-polynomial association scheme*.

- **Distance-regular digraph**

  A strongly connected digraph $D = (V, E)$ is called **distance-regular** (Damerell, 1981) if, for any its vertices u, v with $d_{\text{path}}(u, v) = k$ and for any integer $0 \leq i \leq k + 1$, the number of vertices w, such that $d_{\text{path}}(u, w) = i$ and $d_{\text{path}}(v, w) = 1$, depends only on k and i, but not on the choice of u and v. In order to find interesting classes of distance-regular digraphs with unbounded diameter, the above definition was weakened by two teams in different directions.

  Call $\overline{d(x,y)} = (d(x,y), d(y,x))$ the **two-way distance in digraph** D. A strongly connected digraph $D = (V, E)$ is called **weakly distance-regular** (Wang and Suzuku, 2003) if, for any its vertices u, v with $\overline{d(u, v)} = (k_1, k_2)$, the number of vertices w, such that $\overline{d(w, u)} = (i_1, i_2)$ and $\overline{d(w, v)} = (j_1, j_2)$, depends only on the values $k_1, k_2, i_1, i_2, j_1, j_2$. Comellas et al., 2004, defined a **weakly distance-regular digraph** as one in which, for any vertices u and v, the number of $u \rightarrow v$ walks of every given length only depends on the distance $d(u, v)$.

- **Metrically almost transitive graph**

  An *automorphism* of a graph $G = (V, E)$ is a map $g : V \rightarrow V$ such that u is adjacent to v if and only if $g(u)$ is adjacent to $g(v)$, for any $u, v \in V$. The set $Aut(G)$ of automorphisms of G is a group with respect to the composition of functions.

  A graph G is **metrically almost transitive** (Krön–Möller, 2008) if there is an integer r such that, for any vertex $u \in V$ it holds

$$\cup_{g \in Aut(G)} \{g(\overline{B}(u, r) = \{v \in V : d_{\text{path}}(u, v) \leq r\})\} = V.$$

- **Metric end**

  Given an infinite graph $G = (V, E)$, a *ray* is a sequence $(x_0, x_1, \dots)$ of distinct vertices such that x_i and x_{i+1} are adjacent for $i \geq 0$.

  Two rays R_1 and R_2 are equivalent whenever it is impossible to find a bounded—set of vertices F such that any path from R_1 to R_2 contains an element of F.

  Metric ends are defined as equivalence classes of *metric rays* which are rays without infinite, bounded subsets.

- **Graph of polynomial growth**

  Let $G = (V, E)$ be a transitive locally finite graph. For a vertex $v \in V$, the *growth function* is defined by

  $$f(n) = |\{u \in V : d(u, v) \leq n\}|,$$

  and it does not depend on v. Cf. **growth rate of metric space** in Chap. 1.

  The graph G is a **graph of polynomial growth** if there are some positive constants k, C such that $f(n) \leq Cn^k$ for all $n \geq 0$. It is a **graph of exponential growth** if there is a constant $C > 1$ such that $f(n) > C^n$ for all $n \geq 0$.

  A group with a finite symmetric set of generators has *polynomial growth rate* if the corresponding *Cayley graph* has polynomial growth. Here the metric ball consists of all elements of the group which can be expressed as products of at most n generators, i.e., it is a closed ball centered in the identity in the **word metric**, cf. Chap. 10.
- **Distance-polynomial graph**

  Given a connected graph $G = (V, E)$ of diameter T, for any $2 \leq i \leq T$ denote by G_i the graph $(V, E^{'})$ with $E^{'} = \{e = uv \in E : d_{\text{path}}(u, v) = i\}$. The graph G is called a **distance-polynomial** if the adjacency matrix of any G_i, $2 \leq i \leq T$, is a polynomial in terms of the adjacency matrix of G.

  Any **distance-regular** graph is a distance-polynomial.
- **Distance-hereditary graph**

  A connected graph is called **distance-hereditary** (Howorka, 1977) if each of its connected induced subgraphs is isometric.

  A graph is distance-hereditary if each of its induced paths is isometric. A graph is distance-hereditary, bipartite distance-hereditary, **block graph**, tree if and only if its path metric is a **relaxed tree-like metric** for edge-weights being, respectively, nonzero half-integers, nonzero integers, positive half-integers, positive integers.

  A graph is called a **parity graph** if, for any $u, v \in V$, the lengths of all induced $(u - v)$ paths have the same parity. A graph is a parity graph (moreover, distance-hereditary) if and only if every induced subgraph of odd (moreover, any) order of at least five has an even number of Hamiltonian cycles (McKee, 2008).
- **Distance magic graph**

  A graph $G = (V, E)$ is called (Vilfred, 1994) a **distance magic graph** if it admits a *distance magic labeling*, i.e., a *magic constant* $k > 0$ and a bijection $f : V \to \{1, 2, \dots, |V|\}$ with $w(v) = \sum_{uv \in E} f(v) = k$ for every $u \in V$.

  Cf. **rendez-vous number** in Chap. 1. These graphs generalize *magic squares* (such complete n-partite graphs with parts of size n). Among such trees, cycles and K_n, only P_1, P_3, C_4 are distance magic. The *hypercube graph* $H(m, 2)$ is distance magic if $m = 2, 6$ but not if $m \equiv 0, 1, 3 \pmod 4$.

  The graph $G = (V, E)$ is said (Kamatchi–Arumugam, 2013) to be *distance anti-magic* if $w(u) \neq w(v)$ for all $u, v \in V$. Cichacz et al., 2015, extended both above notions on vertex-labeling by elements of an Abelian group.

[redacted]

The graph $G = (V, E)$ is said (O'Neil–Slater, 2011) to be *D-distance magic* if it admits a D**-distance magic labeling**, i.e., a *magic constant* $k > 0$ and a bijection $f : V \to \{1, 2, \ldots, |V|\}$ with $w(v) = \sum_{u \in V : d(u,v) \in D} f(u) = k$ for every $u \in V$.

- **Block graph**

  A graph is called a **block graph** if each of its *blocks* (i.e., a maximal 2-connected induced subgraph) is a complete graph. Any tree is a block graph.

  A graph is a block graph if and only if its path metric is a **tree-like metric** or, equivalently, satisfies the **four-point inequality**.

- **Ptolemaic graph**

  A graph is called **Ptolemaic** if its path metric satisfies the **Ptolemaic inequality**

$$d(x, y)d(u, z) \leq d(x, u)d(y, z) + d(x, z)d(y, u).$$

  A graph is Ptolemaic if and only if it is distance-hereditary and *chordal*, i.e., every cycle of length greater than 3 has a chord. So, any **block graph** is Ptolemaic.

- k**-cocomparability graph**

  A graph $G = (V, E)$ is called (Chang–Ho–Ko, 2003) k**-cocomparability graph** if its vertex-set admits a linear ordering $<$ such that for any three vertices $u < v < w$, $d(u, w) \leq k$ implies $d(u, v) \leq k$ or $d(v, w) \leq k$.

- **Distance-perfect graph**

  Cvetković et al., 2007, observed that any graph of diameter T has at most $k + T^k$ vertices, where k is its **location number** (Chap. 1), i.e., the minimal cardinality of a set of vertices, the path distances from which uniquely determines any vertex. They called a graph **distance-perfect** if it meets this upper bound and proved that such a graph has $T \neq 2$.

- t**-irredundant set**

  A set $S \subset V$ of vertices in a connected graph $G = (V, E)$ is called t**-irredundant** (Hattingh–Henning, 1994) if for any $u \in S$ there exists a vertex $v \in V$ such that, for the path metric d_{path} of G, it holds

$$d_{\text{path}}(v, x) \leq t < d_{\text{path}}(v, V \backslash S) = \min_{u \notin S} d_{\text{path}}(v, u).$$

  The t**-irredundance number** ir_t of G is the smallest cardinality $|S|$ such that S is t-irredundant but $S \cup \{v\}$ is not, for every $v \in V \backslash S$.

  The *t-independent number* α_t and **distance** t**-domination number** γ_t of G are, respectively, the cardinality of the largest $\lceil \frac{t}{2} \rceil$*-packing* and smallest $(t + 1)$-*covering* (by the open balls of the radius $t+1$) of the metric space $(V, d_{\text{path}}(u, v))$; cf. the **radii of metric space** in Chap. 1. Then it holds $\frac{\gamma_t + 1}{2} \leq ir_t \leq \gamma_t \leq \alpha_t$.

  Let C_r be the largest constant such that $\gamma_1 \geq C_r \sum_{1 \leq i < j \leq r} d(x_i, x_j)$ for r vertices of any connected graph; Kang, 2015, showed that $C_2 = \frac{1}{3}$ and $C_r = \frac{1}{r(r-1)}$ for $r \geq 3$, implying that $\gamma_1 \geq ad(G)$, where ad denotes the average distance.

The *distance t-vertex cover number* of G is (Canales et al., 2015) the minimum cardinality of the set $S \subset V$, such that for each edge e of G, there is a path of length at most t that contains e and a vertex from S.

If G is plane graph, then its *distance t-guarding number* is (Canales et al., 2015) the minimum cardinality of the set $S \subset V$, such that for every face f of G, there is a vertex u in the boundary of f, and a vertex $v \in S$ with $d_{\text{path}}(u, v) < t$.

- ***r*-locating-dominating set**

  Let D=(V,E) be a digraph and $C \subset V$, and let $B_r^-(v)$ denote the set of all vertices x such that there exists a directed $(x - v)$ path with at most r arcs.

  If $B_r^-(v) \cap C$, $v \in V \setminus C$ (respectively, $v \in V$), are nonempty distinct sets, C is called (Slater, 1984) an ***r*-locating-dominating set** (respectively, an ***r*-identifying code**; cf. Chap. 16) of D. Such sets of smallest cardinality are called *optimal*.

- **Locating chromatic number**

  The **locating chromatic number** of a graph $G = (V, E)$ is the minimum number of color classes $C_1, \dots, C_t$ needed to color vertices of G so that any two adjacent vertices have distinct colors and each vertex $u \in V$ has distinct *color code* $(\min_{v \in C_1} d(u, v), \dots, \min_{v \in C_k} d(u, v))$.

- ***k*-distant chromatic number**

  The ***k*-distant chromatic number** of a graph $G = (V, E)$ is the minimum number of colors needed to color vertices of G so that any two vertices at distance at most k have distinct colors, i.e., it is the chromatic number of the ***k*-power of** G.

- **Distance between edges**

  The **distance between edges** in a connected graph $G = (X, E)$ is the number of vertices in a shortest path between them. So, adjacent edges have distance 1.

  A **distance-*k* matching** of G is a set of edges no two of which are within distance k. For $k = 1$, it is the usual matching. For $k = 2$, it is also *induced* (or *strong*) matching. A distance-k matching of G is equivalent to an independent set in the ***k*-power** of the line graph of G. A **distance-*k* edge-coloring** of G is an edge-coloring such that each color class induces a distance-k matching.

  The **distance-*k* chromatic index** $\mu_k(G)$ is the least integer t such that there exists a distance-t edge-coloring of G. The **distance-*k* matching number** $\nu_k(G)$ is the largest integer t such that there exists a distance-t matching in G with t edges. It holds that $\mu_k(G)\nu_k(G) \geq |E|$.

  The **distance between faces** of a plane graph is the number of vertices in a shortest path between them. A **distance-*k* face-coloring** is a face-coloring such that any two faces at distance at most k have different colors. The **distance-*k* face chromatic index** is the least integer t such that such coloring by t colors exists. The dual notion is *k-distance coloring* of a graph $G = (V, E)$, i.e., a vertex-coloring such that any two vertices at distance at most k receives different colors.

- **Distance integral graph**

  Given a finite graph G, its **distance polynomial** (cf. Chap. 1) is the determinant $det(D - \lambda I)$, where D is the distance matrix of the path metric of G.

  A graph is called **distance integral** if all roots of this polynomial are integers.

[redacted]

- ***D*-distance graph**

  Given a set D of positive numbers containing 1 and a metric space (X, d), the **D-distance graph** is a graph $G = (V = X, E)$ with the edge-set $E = \{uv : d(u, v) \in D\}$ (cf. **D-chromatic number** in Chap. 1). If (X, d) is path metric of a graph H, then G is called the **distance power** H^D of H.

  Alon–Kupavsky, 2014, call G (in the case $(X, d) = \mathbb{E}^n$, $d = \{1\}$) the *faithful unit-distance graph*, using term *unit-distance graph* for $E \subseteq \{(u, v) : ||u-v||_2 = 1\}$.

  For a positive number t, the *signed distance graph* is (Fiedler, 1969) a signed graph with the vertex-set X in which vertices x, y are joined by a positive edge if $t > d(x, y)$, by a negative edge if $d(x, y) > t$, and not joined if $d(x, y) = t$.

  A D-distance graph is called a **distance graph** (or *unit-distance graph*) if $D = \{1\}$, an *ϵ-unit graph* if $D = [1 - \epsilon, 1 + \epsilon]$, a *unit-neighborhood graph* if $D = (0, 1]$, an *integral-distance graph* if $D = \mathbb{Z}_+$, a *rational-distance graph* if $D = \mathbb{Q}_+$, and a *prime-distance graph* if D is the set of prime numbers (with 1).

  Every finite graph can be represented by a D-distance graph in some $\mathbb{E}^n$. The minimum dimension of such a Euclidean space is called the *D-dimension* of G. A *matchstick graph* is a crossingless unit-distance graph in $\mathbb{E}^2$.

- **Distance-number of a graph**

  Given a graph $G = (V, E)$, its *degenerate drawing* is a mapping $f : V \to \mathbb{R}^2$ such that $|f(V)| = |V|$ and $f(uv)$ is an open straight-line segment joining the vertices $f(u)$ and $f(v)$ for any edge $uv \in E$; it is a *drawing* if, moreover, $f(w) \notin f(uv)$ for any $uv \in E$ and $w \in V$.

  The **distance-number** $dn(G)$ **of a graph** G is (Carmi et al. 2008) the minimum number of distinct edge-lengths in a drawing of G.

  The *degenerate distance-number* of G, denoted by $ddn(G)$, is the minimum number of distinct edge-lengths in a degenerated drawing of G. The first of the **Erdös-type distance problems** in Chap. 19 is equivalent to determining $ddn(K_n)$.

- **Dimension of a graph**

  The **dimension** $dim(G)$ **of a graph** G is (Erdös–Harary–Tutte, 1965) the minimum k such that G has a *unit-distance representation* in $\mathbb{R}^k$, i.e., every edge is of length 1. The vertices are mapped to distinct points of $\mathbb{R}^k$, but edges may cross.

  For example, $dim(G) = n - 1, 4, 2$ for $G = K_n, K_{m,n}, C_n$ $(m \geq n \geq 3)$.

- **Bar-and-joint framework**

  A n-dimensional **bar-and-joint framework** is a pair (G,f), where $G = (V, E)$ is a finite graph (no loops and multiple edges) and $f : V \to \mathbb{R}^n$ is a map with $f(u) \neq f(v)$ whenever $uv \in E$. The **framework** is a straight line realization of G in $\mathbb{R}^n$ in which the length of an edge $uv \in E$ is given by $||f(u) - f(v)||_2$.

  The vertices and edges are called *joints* and *bars*, respectively, in terms of Structural Engineering. A **tensegrity structure** (Fuller, 1948) is a mechanically stable bar framework in which bars are either *cables* (tension elements which cannot get further apart), or *struts* (compression elements which cannot get closer together).

A framework (G,f) is *globally rigid* if every framework (G,f'), satisfying $||f(u) - f(v)||_2 = ||f'(u) - f'(v)||_2$ for all $uv \in E$, also satisfy it for all $u, v \in V$. A framework (G,f) is *rigid* if every continuous motion of its vertices which preserves the lengths of all edges, also preserves the distances between all pairs of vertices. The framework (G,f) is *generic* if the set containing the coordinates of all the points $f(v)$ is algebraically independent over the rationals. The graph G is *n-rigid* if every its n-dimensional generic realization is rigid. For generic frameworks, rigidity is equivalent to the stronger property of infinitesimal rigidity.

An *infinitesimal motion* of (G,f) is a map $m : V \to \mathbb{R}^n$ with $(m(u) - m(v))(f(u) - f(v)) = 0$ whenever $uv \in E$. A motion is *trivial* if it can be extended to an isometry of $\mathbb{R}^n$. A framework is an *infinitesimally rigid* if every motion of it is trivial, and it is *isostatic* if, moreover, the deletion of any its edge will cause loss of rigidity. (G,f) is an *elastic framework* if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that for every edge-weighting $w : E \to \mathbb{R}_{>0}$ with $\max_{uv \in E} |w(uv) - ||f(u) - f(v)||_2| \leq \delta$, there exist a framework (G,f') with $\max_{v \in V} ||f(u) - f'(v)||_2 < \epsilon$.

A framework (G,f) with $||f(u) - f(v)||_2 > r$ if $u, v \in V, u \neq c$ and $||f(u),f(v)||_2 \leq R$ if $uv \in E$, for some $0 < r < R$, is called (Doyle–Snell, 1984) a *civilized drawing of a graph*. The random walks on such graphs are recurrent if $n = 1, 2$.

- **Distance constrained labeling**

  Given a sequence $\alpha = (\alpha_1, \dots, \alpha_k)$ of **distance constraints** $\alpha_1 \geq \cdots \geq \alpha_k > 0$, a λ_α*-labeling* of a graph $G = (V, E)$ is an assignment of labels $f(v)$ from the set $\{0, 1, \dots, \lambda\}$ of integers to the vertices $v \in V$ such that, for any t with $0 \leq t \leq k$, $|f(v) - f(u)| \geq \alpha_t$ whenever the path distance between u and v is t.

  The *radio frequency assignment problem*, where vertices are transmitters (available channels) and labels represent frequencies of not-interfering channels, consists of minimizing λ. **Distance-two labeling** is the main interesting case $\alpha = (2, 1)$; its *span* is the difference between the largest and smallest labels used.

- **Distance-related graph embedding**

  An *embedding* of the guest graph $G = (V_1, E_1)$ into the host graph $H = (V_2, E_2)$ with $|V_1| \leq |V_2|$, is an injective map from V_1 into V_2.

  The **wire length**, *dilation* and *antidilation* of G in H are

$$\min_f \sum_{(uv) \in E_1} d_H(f(u),f(v)), \ \min_f \max_{(uv) \in E_1} d_H(f(u),f(v)), \ \max_f \min_{(uv) \in E_1} d_H(f(u),f(v)),$$

  respectively, where f is any embedding of G into H. The main **distance-related graph embedding** problems consist of finding or estimating these three parameters.

  The *bandwidth* and *antibandwidth* of G is the dilation and antidilation, respectively, of G in a path H with V_1 vertices.

[redacted]

- **Bandwidth of a graph**

  Given a graph $G = (V, E)$ with $|V| = n$, its *ordering* is a bijective mapping $f : V \to \{1, \dots, n\}$. Given a number $b > 0$, the *bandwidth problem* for (G, b) is the existence of ordering f with the *stretch* $\max_{uv \in E} |f(u) - f(v)|$ at most b.

  The **bandwidth** of G, denoted by $bw(G)$, is the minimum stretch over all f.

  The *antibandwidth problem* for G is to find ordering f with maximal $\min_{uv \in E} |f(u) - f(v)|$ (*antibandwidth*).

- **Path distance width of a graph**

  Given a connected graph $G = (V, E)$, an ordered partition $V = \cup_{i=1}^{t} L_i$ of its vertices is called a *distance structure* on G if $L_i = \{v \in V : \min_{u \in L_1} d_{\text{path}}(u, v) = i - 1\}$ for $1 \le i \le t$. The structure is *rooted* if $|L_1| = 1$.

  The **path distance width** $pwd(G)$ of G is defined (Yamazaki et al., 1999) as $\min \max_{1 \le i \le t} |L_i|$ over all distance structures on G.

  An ordered partition $V = \cup_{i=1}^{t} L_i$ is called a *level structure* on G if for each edge uv with $u \in L_i$ and $v \in L_j$, it holds that $|i - j| \le 1$. The *level width* (or *strong pathwidth*) $lw(G)$ is $\min \max_{1 \le i \le t} |L_i|$ over all level structures.

  Clearly, $lw(G) \le pdw(G)$. Yamazaki et al., 1999, proved that $pdw(G)$ can be arbitrarily larger than the **bandwidth** $bw(G)$ and $lw(G) \le bw(G) < 2lw(G)$.

- **Tree-length of a graph**

  A *tree decomposition* of a graph $G = (V, E)$ is a pair of a tree T with vertex-set W and a family of subsets $\{X_i : i \in W\}$ of V with $\cup_{i \in W} X_i = V$ such that

  1. for every edge $(uv) \in E$, there is a subset X_i containing u, v, and
  2. for every $v \in V$, the set $\{i \in W : v \in X_i\}$ induces a connected subtree of T.

  The *chordal graphs* (i.e., ones without induced cycles of length at least 4) are exactly those admitting a tree decomposition where every X_i is a clique.

  For tree decomposition, the *tree-length* is $\max_{i \in W} diam(X_i)$ ($diam(X_i)$ is the diameter of the subgraph of G induced by X_i) and *tree-width* is $\max_{i \in W} |X_i| - 1$. The **tree-length** of G (Dourisboure–Gavoille, 2004) and its **tree-width** (Robertson–Seymour, 1986) are the minima, over all tree decompositions, of above tree-length and tree-width. The *path-length* G is defined taking as trees only paths.

  Given a linear ordering $e_1, \dots, e_{|E|}$ of the edges of G, let, for $1 \le i < |E|$, denote by $G_{\le i}$ and $G_{i<}$ the graphs induced by the edges $\{e_1, \dots, e_i\}$ and $\{e_{i+1}, \dots, e_{|E|}\}$, respectively. The *linear-length* is $\max_{1 \le i < |E|} diam(V(G_{\le i}) \cap V(G_{i<}))$. The **linear-length** of G (Umezawa–Yamazaki, 2009) is the minimum of the above linear-length taken over all the linear orderings of its edges.

- **Spatial graph**

  A **spatial graph** (or *spatial network*) is a graph $G = (V, E)$, where each vertex v has a spatial position $(v_1, \dots, v_n) \in \mathbb{R}^n$. ($G$ is called a *geometric graph* if it is drawn on $\mathbb{R}^2$ and its edges are straight-line segments.)

  The *graph-theoretic dilation* and *geometric dilation* of G are, respectively:

$$\max_{v,u \in V} \frac{d(v,u)}{||v-u||_2} \text{ and } \max_{(vu) \in E} \frac{d(v,u)}{||v-u||_2}.$$

- **Distance Geometry problem**

  Given a weighted finite graph $G = (V, E; w)$, the **Distance Geometry problem** (DGP) is the problem of realizing it as a **spatial graph** $G = (V', E')$, where $x : V \to V'$ is a bijection with $x(v) = (v_1, \dots, v_n) \in \mathbb{R}^n$ for every $v \in V$ and $E' = \{(x(u)x(v)) : (uv) \in E\}$, so that for every edge $(uv) \in E$ it holds that

  $$||x(u) - x(u)||_2 = w(uv).$$

  The main application of DGP is the *molecular DGP*: to find the coordinates of the atoms of a given molecular conformation are by exploiting only some of the distances between pairs of atoms found experimentally; cf. [MLLM13].

- **Arc routing problems**

  Given a finite set X, a quasi-distance $d(x, y)$ on it and a set $A \subseteq \{(x, y) : x, y \in X\}$, consider the weighted digraph $D = (X, A)$ with the vertex-set X and arc-weights $d(x, y)$ for all arcs $(x, y) \in A$. For given sets V of vertices and E of arcs, the **arc routing problem** consists of finding a *shortest* (i.e., with minimal sum of weights of its arcs) (V, E)*-tour*, i.e., a circuit in $D = (X, A)$, visiting each vertex in V and each arc in E exactly once or, in a variation, at least once.

  The *Asymmetric Traveling Salesman problem* corresponds to the case $V = X$, $E = \emptyset$; the *Traveling Salesman problem* is the symmetric version of it (usually, each vertex should be visited exactly once). The *Bottleneck Traveling Salesman problem* consists of finding a (V, E)-tour T with smallest $\max_{(x,y) \in T} d(x, y)$.

  The *Windy Postman problem* corresponds to the case $V = \emptyset$, $E = A$, while the Chinese Postman problem is the symmetric version of it.

  The above problems are also considered for general arc- or edge-weights; then, for example, the term *Metric TSP* is used when edge-weights in the Traveling Salesman problem satisfy the triangle inequality, i.e., d is a quasi-semimetric.

- **Steiner distance of a set**

  The **Steiner distance of a set** $S \subset V$ of vertices in a connected graph $G = (V, E)$ is (Chartrand et al., 1989) the minimum *size* (number of edges) of a connected subgraph of G, containing S. Such a subgraph is a tree, and is called a *Steiner tree* for S. Cf. general **Steiner diversity** in **Steiner ratio** (Chap. 1).

  The Steiner distance of the set $S = \{u, v\}$ is the path metric between u and v. The *Steiner k-diameter* of G is the maximum Steiner distance of any k-subset of V.

- ***t*-spanner**

  A *factor*, i.e., a spanning subgraph, $H = (V, E(H))$ of a connected graph $G = (V, E)$ is called a ***t*-spanner** (or *t-multiplicative spanner*) of G if, for every $u, v \in V$, the inequality $d_{\text{path}}^{H}(u, v)/d_{\text{path}}^{G}(u, v) \leq t$ holds. The value t is called the *stretch factor* (or *dilation*) of H. Cf. **distance-related graph embedding** and **spatial graph**.

  The graph $H = (V, E(H))$ is called a *k-additive spanner* of G if, for every $u, v \in V$, the inequality $d_{\text{path}}^{H}(u, v) \leq d_{\text{path}}^{G}(u, v) + k$ holds.

[redacted]

Mulder and Nebeský, 2012, defined, for connected H, the *guide* of (H, G) as the ternary relation $R \subset V \times V \times V$ consisting of ordered triples (u, w, v) such that $uw \in E$ and $d^H_{\text{path}}(u, w) + d^H_{\text{path}}(w, v) = d^H_{\text{path}}(u, v)$. The guide of (G, G) is called the *step* ternary relation; cf. **metric betweenness** in Chap. 1.

- **Optimal realization of metric space**

  Given a finite metric space (X, d), a *realization* of it is a weighted graph $G = (V, E; w)$ with $X \subset V$ such that $d(x, y) = d_G(x, y)$ holds for all $x, y \in X$.

  The realization is **optimal** if it has minimal $\sum_{(uv) \in E} w(uv)$.

- **Proximity graph**

  Given a finite subset V of a metric space (X, d), its **proximity graph** is a graph representing neighbor relationships between points of V. Such graphs are used in Computational Geometry and many real-world problems. The main examples are presented below. Cf. **underlying graph of a metric space** in Chap. 1.

  A *spanning tree* of V is a set T of $|V| - 1$ unordered pairs (x, y) of different points of V forming a tree on V; the *weight* of T is $\sum_{(x,y) \in T} d(x, y)$. A **minimum spanning tree** $MST(V)$ of V is a spanning tree with the minimal weight. Such a tree is unique if the edge-weights are distinct.

  A **nearest neighbor graph** is the digraph $NNG(V) = (V, E)$ with vertex-set $V = v_1, \dots, v_{|V|}$ and, for $x, y \in V$, $xy \in E$ if y is the *nearest neighbor* of x, i.e., $d(x, y) = \min_{v_i \in V \setminus \{x\}} d(x, v_i)$ and only v_i with maximal index i is picked. The *k-nearest neighbor graph* arises if k such v_i with maximal indices are picked. The undirect version of $NNG(V)$ is a subgraph of $MST(V)$.

  A **relative neighborhood graph** is (Toussaint, 1980) the graph $RNG(V) = (V, E)$ with vertex-set V and, for $x, y \in V$, $xy \in E$ if there is no point $z \in V$ with $\max\{d(x, z), d(y, z)\} < d(x, y)$. Also considered, for $(X, d) = (\mathbb{R}^2, ||x - y||_2)$, the related *Gabriel graph* $GG(V)$ (in general, *β-skeleton*) and *Delaunay triangulation* $DT(V)$; then $NNG(V) \subseteq MST(V) \subseteq RNG(V) \subseteq GG(V) \subseteq DT(V)$.

  For any $x \in V$, its *sphere of influence* is the open metric ball $B(x, r_x) = \{z \in X : d(x, z) < r\}$ in (X, d) centered at x with radius $r_x = \min_{z \in V \setminus \{x\}} d(x, z)$.

  Sphere of influence graph is the graph $SIG(V) = (V, E)$ with vertex-set V and, for $x, y \in V$, $xy \in E$ if $B(x, r_x) \cap B(y, r_y) \neq \emptyset$; so, it is a proximity graph and an *intersection graph*. The *closed sphere of influence graph* is the graph $CSIG(V) = (V, E)$ with $xy \in E$ if $\overline{B(x, r_x)} \cap \overline{B(y, r_y)} \neq \emptyset$.

15.3 Distances on Graphs

- **Chartrand–Kubicki–Schultz distance**

  The **Chartrand–Kubicki–Schultz distance** (or *ϕ-distance*, 1998) between two connected graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ with $|V_1| = |V_2| = n$ is

$$\min\{\sum |d_{G_1}(u, v) - d_{G_2}(\phi(u), \phi(v))|\},$$

where d_{G_1}, d_{G_2} are the path metrics of graphs G_1, G_2, the sum is taken over all unordered pairs u, v of vertices of G_1, and the minimum is taken over all bijections $\phi : V_1 \to V_2$.

- **Subgraph metric**

  Let $\mathbb{F} = \{F_1 = (V_1, E_1), F_2 = (V_2, E_2), \dots,\}$ be the set of isomorphism classes of finite graphs. Given a finite graph $G = (V, E)$, denote by $s_i(G)$ the number of *injective homomorphisms* from F_i into G, i.e., the number of injections $\phi : V_i \to V$ with $\phi(x)\phi(y) \in E$ if $xy \in E_i$ divided by the number $\frac{|V|!}{(|V|-|V_i|)!}$ of such injections from F_i with $|V_i| \leq |V|$ into $K_{|V|}$. Set $s(G) = (s_i(G))_{i=1}^{\infty} \in [0, 1]^{\infty}$.

  Let d be the **Cantor metric** (Chap. 18) $d(x, y) = \sum_{i=1}^{\infty} 2^{-i}|x_i - y_i|$ on $[0, 1]^{\infty}$ or any metric on $[0, 1]^{\infty}$ inducing the *product topology*. Then Bollobás–Riordan, 2007, defined the **subgraph metric** between the graphs G_1 and G_2 as

$$d(s(G_1), s(G_2))$$

  and generalized it on *kernels* (or *graphons*), i.e., symmetric measurable functions $k : [0, 1] \times [0, 1] \to \mathbb{R}_{\geq 0}$, replacing G by k and the above $s_i(G)$ by

$$s_i(k) = \int_{[0,1]^{|V_i|}} \prod_{st \in E_i} k(x_s x_t) \prod_{s=1}^{|V_i|} dx_s.$$

- **Benjamini–Schramm metric**

  The rooted graphs (G, o) and (G', o') (where $G = (V, E), G' = (V', E')$ and $o \in V, o' \in V'$) are *isomorphic* is there is a graph-isomorphism of G onto G' taking o to o'. Let X be the set of isomorphism classes of rooted connected locally finite graphs and let $(G, o), (G', o')$ be representatives of two classes.

  Let k be the supremum of all radii r, for which rooted **metric balls** $(\overline{B}_G(o, r), o)$ and $(\overline{B}_{G'}(o', r), o')$ (in the usual **path metric**) are isomorphic as rooted graphs. Benjamini and Schramm, 2001, defined the metric 2^{-k} between classes represented by (G, o) and (G', o'). Here $2^{-\infty}$ means 0. Benjamini and Curien, 2011, defined the similar distance $\frac{1}{1+k}$.

- **Rectangle distance on weighted graphs**

  Let $G = G(\alpha, \beta)$ be a complete weighted graph on $\{1, \dots, n\}$ with vertex-weights $\alpha_i > 0$, $1 \leq i \leq n$, and edge-weights $\beta_{ij} \in \mathbb{R}$, $1 \leq i < j \leq n$. Denote by $A(G)$ the $n \times n$ matrix $((a_{ij}))$, where $a_{ij} = \frac{\alpha_i \alpha_j \beta_{ij}}{(\sum_{1 \leq i \leq n} \alpha_i)^2}$.

  The **rectangle distance** (or *cut distance*) between two weighted graphs $G = G(\alpha, \beta)$ and $G' = G(\alpha', \beta')$ (with vertex-weights (α'_i) and edge-weights (β'_{ij})) is defined (Borgs–Chayes–Lovász–Sós–Vesztergombi, 2007) by

$$\max_{I,J \subset \{1,\dots,n\}} |\sum_{i \in I, j \in J} (a_{ij} - a'_{ij})| + \sum_{i=1}^{n} |\frac{\alpha_i}{\sum_{1 \leq j \leq n} \alpha_j} - \frac{\alpha'_i}{\sum_{1 \leq j \leq n} \alpha'_j}|,$$

  where $A(G) = ((a_{ij}))$ and $A(G') = ((a'_{ij}))$.

[redacted]

In the case $(\alpha'_i) = (\alpha_i)$, the rectangle distance is $||A(G) - A(G')||_{cut}$, i.e., the **cut norm metric** (cf. Chap. 12) between matrices $A(G)$ and $A(G')$ and the *rectangle distance* from Frieze–Kannan, 1999. In this case, the l_1- and l_2-metrics between two weighted graphs G and G' are defined as $||A(G) - A(G')||_1$ and $||A(G) - A(G')||_2$, respectively. The subcase $\alpha_i = 1$ for all $1 \leq i \leq n$ corresponds to unweighted vertices. Cf. the **Robinson–Foulds weighted metric**.

Authors generalized the rectangle distance on *kernels* (or *graphons*), i.e., symmetric measurable functions $k : [0,1] \times [0,1] \to \mathbb{R}_{\geq 0}$, using the *cut norm* $||k||_{cut} = \sup_{S,T \subset [0,1]} |\int_{S \times T} k(x,y)dxdy|$.

A map $\phi : [0,1] \to [0,1]$ is *measure-preserving* if, for any measurable subset $A \subset [0,1]$, the measures of A and $\phi^{-1}(A)$ are equal. For a kernel k, define the kernel k^{ϕ} by $k^{\phi}(x,y) = k(\phi(x), \phi(y))$. The **Lovász–Szegedy semimetric** (2007) between kernels k_1 and k_1 is defined by

$$\inf_{\phi} ||k_1^{\phi} - k_2||_{cut},$$

where ϕ ranges over all measure-preserving bijections $[0,1] \to [0,1]$. Cf. **Chartrand–Kubicki–Schultz distance**.

- **Spectral semimetric between graphs**

  Given a finite weighted graph $G = (V, E, w)$, its *normalised Laplacian* matrix is $\Delta_G = I - D^{-1}A$, where D is the diagonal matrix $((l_{ij}))$ with l_{ii} being the degree of vertex i, while A is the adjacency matrix of G; cf. the **resistance metric**.

  The *spectrum* of G consists of all eigenvalues λ_i, $1 \leq i \leq |V|$, of Δ_G. The spectrum, organised as $\lambda_G = (\lambda_1, \dots, \lambda_{|V|})$, where the λ_i do not decrease, is called the *spectral vector*.

  The **spectral semimetric** between two finite weighted graphs $G = (V, E, w)$ and $G' = (V', E', w')$ with $|V| = |V'| = n$ is defined (Gu–Hua–Liu, 2015) as

$$d_p(G, G') = \frac{1}{n} \|\lambda_G - \lambda_{G'}\|_p,$$

  i.e., the l_p**-distance** between the spectral vectors, for any $1 \leq p < \infty$.

  In fact, by assigning a probability measure via the spectrum of the normalised Laplacian matrix to each graph, Gu, Hua, and Liu defined their *spectral distance* d_p as the L_p**-Wasserstein distance** (cf. Chap. 14) between probability measures on the set of all, including infinite and random, weighted graphs.

- **Subgraph–supergraph distances**

  A *common subgraph* of graphs G_1 and G_2 is a graph which is isomorphic to induced subgraphs of both G_1 and G_2. A *common supergraph* of graphs G_1 and G_2 is a graph which contains induced subgraphs isomorphic to G_1 and G_2.

  The **Zelinka distance** d_Z ([Zeli75]) on the set **G** of all graphs (more exactly, on the set of all equivalence classes of isomorphic graphs) is defined by

$$d_Z = \max\{n(G_1), n(G_2)\} - n(G_1, G_2)$$

for any $G_1, G_2 \in \mathbf{G}$, where $n(G_i)$ is the number of vertices in G_i, $i = 1, 2$, and $n(G_1, G_2)$ is the maximum number of vertices of their common subgraph.

The **Bunke–Shearer metric** (1998) on the set of nonempty graphs is defined by

$$1 - \frac{n(G_1, G_2)}{\max\{n(G_1), n(G_2)\}}.$$

Given any set $\mathbf{M}$ of graphs, the **common subgraph distance** d_M on $\mathbf{M}$ is

$$\max\{n(G_1), n(G_2)\} - n(G_1, G_2),$$

and the **common supergraph distance** d_M^* is defined, for any $G_1, G_2 \in \mathbf{M}$, by

$$N(G_1, G_2) - \min\{n(G_1), n(G_2)\},$$

where $n(G_i)$ is the number of vertices in G_i, $i = 1, 2$, while $n(G_1, G_2)$ and $N(G_1, G_2)$ are the maximal order of a common subgraph $G \in \mathbf{M}$ and the minimal order of a common supergraph $H \in \mathbf{M}$, respectively, of G_1 and G_2.

d_M is a metric on $\mathbf{M}$ if the following condition (i) holds:

(i) if $H \in \mathbf{M}$ is a common supergraph of $G_1, G_2 \in \mathbf{M}$, then there exists a common subgraph $G \in \mathbf{M}$ of G_1 and G_2 with $n(G) \geq n(G_1) + n(G_2) - n(H)$.
d_M^* is a metric on $\mathbf{M}$ if the following condition (ii) holds:

(ii) if $G \in \mathbf{M}$ is a common subgraph of $G_1, G_2 \in \mathbf{M}$, then there exists a common supergraph $H \in \mathbf{M}$ of G_1 and G_2 with $n(H) \leq n(G_1) + n(G_2) - n(G)$.

One has $d_M \leq d_M^*$ if the condition (i) holds, and $d_M \geq d_M^*$ if (ii) holds.

The distance d_M is a metric on the set $\mathbf{G}$ of all graphs, the set of all cycle-free graphs, the set of all bipartite graphs, and the set of all trees. The distance d_M^* is a metric on the set $\mathbf{G}$ of all graphs, the set of all connected graphs, the set of all connected bipartite graphs, and the set of all trees. The Zelinka distance d_Z coincides with d_M and d_M^* on the set $\mathbf{G}$ of all graphs. On the set $\mathbf{T}$ of all trees the distances d_M and d_M^* are identical, but different from the Zelinka distance.

The Zelinka distance d_Z is a metric on the set $\mathbf{G}(n)$ of all graphs with n vertices, and is equal to $n - k$ or to $K - n$ for all G_1, $G_2 \in \mathbf{G}(n)$, where k is the maximum number of vertices of a common subgraph of G_1 and G_2, and K is the minimum number of vertices of a common supergraph of G_1 and G_2.

On the set $\mathbf{T}(n)$ of all trees with n vertices the distance d_Z is called the **Zelinka tree distance** (see, for example, [Zeli75]).

- **Fernández–Valiente metric**

  Given graphs G and H, let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be their *maximum common subgraph* and *minimum common supergraph*; cf. **subgraph–supergraph distances**. The **Fernández–Valiente metric** (2001) between G and H is

$$(|V_2| + |E_2|) - (|V_1| + |E_1|).$$

[redacted]

- **Graph edit distance**

  The **graph edit distance** (Axenovich–Kézdy–Martin, 2008, and Alon–Stav, 2008) between graphs G and G' on the same labeled vertex-set is defined by

$$d_{ed}(G,G') = |E(G) \Delta E(G')|.$$

It is the minimum number of edge deletions or additions needed to transform G into G', and half of the Hamming distance between their adjacency matrices.

Given a *graph property* (i.e., a family $\mathcal{H}$ of graphs), let $d_{ed}(G,\mathcal{H})$ be $\min\{d_{ed}(G,G') : V(G') = V(G), G' \in \mathcal{H}\}$. Given a number $p \in (0,1]$, the **edit distance function of a property** $\mathcal{H}$ is (if this limit exists) defined by

$$ed_{\mathcal{H}}(p) = \lim_{n\to\infty} \max\{d_{ed}(G,\mathcal{H}) : |V(G)| = n, |E(G)| = \lfloor p\binom{n}{2}\rfloor\}(\binom{n}{2})^{-1}.$$

If $\mathcal{H}$ is *hereditary* (closed under the taking induced subgraphs) and *nontrivial* (contains arbitrarily large graphs), then (Balogh–Martin, 2008) it holds

$$ed_{\mathcal{H}}(p) = \lim_{n\to\infty} \mathbb{E}[d_{ed}(G(n,p),\mathcal{H})](\binom{n}{2})^{-1};$$

$G(n,p)$ is the **random graph** (Chap. 1) on n vertices with edge probability p.

Bunke, 1997, defined the *graph edit distance* between vertex- and edge-labeled graphs G_1 and G_2 as the minimal total cost of matching G_1 and G_2, using deletions, additions and substitutions of vertices and edges. Cf. also **tree**, **top-down**, **unit cost** and **restricted edit distance** between rooted trees.

The **Bayesian graph edit distance** between two *relational graphs* (i.e., triples (V,E,A), where V,E,A are the sets of vertices, edges, *vertex-attributes*) is (Myers–Wilson–Hancock, 2000) their graph edit distance with costs defined by probabilities of operations along an editing path seen as a memoryless error process. Cf. **transduction edit distances** (Chap. 11) and **Bayesian distance** (Chap. 14).

The **structural Hamming distance** between two digraphs $G = (X,E)$ and $G' = (X,E')$ is defined (Acid–Campos, 2003) as $SHD(G,G') = |E \Delta E'|$. The *ring sum* (or *symmetric difference*) of G and G' is defined (Deo, 1974) as $(X \cup X', E \Delta E')$.

- **Edge distance**

  The **edge distance** on the set of all graphs is defined (Baláž et al., 1986) by

$$|E_1| + |E_2| - 2|E_{12}| + ||V_1| - |V_2||$$

for any graphs $G_1 = (V_1,E_1)$ and $G_2 = (V_2,E_2)$, where $G_{12} = (V_{12},E_{12})$ is a common subgraph of G_1 and G_2 with maximal number of edges. This distance has many applications in Organic and Medical Chemistry.

- **Contraction distance**

  The **contraction distance** is a distance on the set $\mathbf{G}(n)$ of all graphs with n vertices defined by

  $$n - k$$

  for any $G_1, G_2 \in \mathbf{G}(n)$, where k is the maximum number of vertices of a graph which is isomorphic simultaneously to a graph, obtained from each of G_1 and G_2 by a finite number of *edge contractions*. To perform the *contraction* of the edge $uv \in E$ of a graph $G = (V, E)$ means to replace u and v by one vertex that is adjacent to all vertices of $V \backslash \{u, v\}$ which were adjacent to u or to v.
- **Edge move distance**

  The **edge move distance** (Balâž et al., 1986) is a metric on the set $\mathbf{G}(n, m)$ of all graphs with n vertices and m edges, defined, for any $G_1, G_2 \in \mathbf{G}(m, n)$, as the minimum number of *edge moves* necessary for transforming the graph G_1 into the graph G_2. It is equal to $m - k$, where k is the maximum size of a common subgraph of G_1 and G_2.

  An *edge move* is one of the *edge transformations*, defined as follows: H can be obtained from G by an edge move if there exist (not necessarily distinct) vertices u, v, w, and x in G such that $uv \in E(G)$, $wx \notin E(G)$, and $H = G - uv + wx$.
- **Edge jump distance**

  The **edge jump distance** is an extended metric (which in general can take the value ∞) on the set $\mathbf{G}(n, m)$ of all graphs with n vertices and m edges defined, for any $G_1, G_2 \in \mathbf{G}(m, n)$, as the minimum number of *edge jumps* necessary for transforming G_1 into G_2.

  An *edge jump* is one of the *edge transformations*, defined as follows: H can be obtained from G by an edge jump if there exist four distinct vertices u, v, w, and x in G, such that $uv \in E(G)$, $wx \notin E(G)$, and $H = G - av + wx$.
- **Edge flipping distance**

  Let $P = \{v_1, \dots, v_n\}$ be a collection of points on the plane. A *triangulation* T of P is a partition of the convex hull of P into a set of triangles such that each triangle has a disjoint interior and the vertices of each triangle are points of P.

  The **edge flipping distance** is a distance on the set of all triangulations of P defined, for any triangulations T and T_1, as the minimum number of edge flippings necessary for transforming T into T_1.

  An edge e of T is called *flippable* if it is the boundary of two triangles t and t' of T, and $C = t \cup t'$ is a convex quadrilateral. The *flipping* e is one of the *edge transformations*, which consists of removing e and replacing it by the other diagonal of C. Edge flipping is an special case of *edge jump*.

  The edge flipping distance can be extended on *pseudo-triangulations*, i.e., partitions of the convex hull of P into a set of disjoint interior *pseudo-triangles* (simply connected subsets of the plane that lie between any three mutually tangent convex sets) whose vertices are given points.

[redacted]

- **Edge rotation distance**

  The **edge rotation distance** (Chartand–Saba–Zou, 1985) is a metric on the set $\mathbf{G}(n,m)$ of graphs with n vertices and m edges, defined, for any G_1, G_2, as the minimum number of *edge rotations* needed for transforming G_1 into G_2. Cf. **Pospichal–Kvasnička chemical distance**, called also *reaction metric*, in Chap. 24.

  An *edge rotation* is one of the *edge transformations*, defined as follows: H can be obtained from G by an edge rotation if there exist distinct vertices u, v, and w in G, such that $uv \in E(G)$, $uw \notin E(G)$, and $H = G - uv + uw$.
- **Tree edge rotation distance**

  The **tree edge rotation distance** is a metric on the set $\mathbf{T}(n)$ of all trees with n vertices defined, for all T_1, $T_2 \in \mathbf{T}(n)$, as the minimum number of *tree edge rotations* necessary for transforming T_1 into T_2. A *tree edge rotation* is an *edge rotation* performed on a tree, and resulting in a tree.

  For $\mathbf{T}(n)$ the tree edge rotation and the edge rotation distances may differ.
- **Edge shift distance**

  The **edge shift distance** (or **edge slide distance**) is a metric (Johnson, 1985) on the set $\mathbf{G}_c(n,m)$ of all connected graphs with n vertices and m edges defined, for any $G_1, G_2 \in \mathbf{G}_c(m,n)$, as the minimum number of *edge shifts* necessary for transforming G_1 into G_2.

  An *edge shift* is one of the *edge transformations*, defined as follows: H can be obtained from G by an edge shift if there exist distinct vertices u, v, and w in G such that $uv, vw \in E(G)$, $uw \notin E(G)$, and $H = G - uv + uw$. Edge shift is a special kind of *edge rotation* in the case when the vertices v, w are adjacent in G.

  The edge shift distance can be defined between any graphs G and H with components $G_i (1 \leq i \leq k)$ and $H_i (1 \leq i \leq k)$, respectively, such that G_i and H_i have the same order and the same size.
- **F-rotation distance**

  The **F-rotation distance** is a distance on the set $\mathbf{G}_F(n,m)$ of all graphs with n vertices and m edges, containing a subgraph isomorphic to a given graph F of order at least 2 defined, for all G_1, $G_2 \in \mathbf{G}_F(m,n)$, as the minimum number of *F-rotations* necessary for transforming G_1 into G_2.

  An *F-rotation* is one of the *edge transformations*, defined as follows: let F' be a subgraph of a graph G, isomorphic to F, let u, v, w be three distinct vertices of the graph G such that $u \notin V(F')$, $v, w \in V(F')$, $uv \in E(G)$, and $uw \notin E(G)$; H can be obtained from G by the F-rotation of the edge uv into the position uw if $H = G - uv + uw$.
- **Binary relation distance**

  Let R be a nonreflexive *binary relation* between graphs, i.e., $R \subset \mathbf{G} \times \mathbf{G}$, and there exists $G \in \mathbf{G}$ such that $(G,G) \notin R$.

  The **binary relation distance** is a metric (which can take the value ∞) on the set $\mathbf{G}$ of all graphs defined, for any graphs G_1 and G_2, as the minimum number of *R-transformations* necessary for transforming G_1 into G_2. We say that a graph H can be obtained from a graph G by an *R-transformation* if $(H,G) \in R$.

An example is the distance between two *triangular embeddings of a complete graph* (i.e., its cellular embeddings in a surface with only 3-gonal faces) defined as the minimal number t such that, up to replacing t faces, the embeddings are isomorphic.

- **Crossing-free transformation metrics**

  Given a subset S of $\mathbb{R}^2$, a *noncrossing spanning tree* of S is a tree whose vertices are points of S, and edges are pairwise noncrossing straight line segments.

  The **crossing-free edge move metric** (see [AAH00]) on the set $\mathbf{T}_S$ of all noncrossing spanning trees of a set S, is defined, for any $T_1, T_2 \in \mathbf{T}_S$, as the minimum number of *crossing-free edge moves* needed to transform T_1 into T_2. Such move is an edge transformation which consists of adding some edge e in $T \in \mathbf{T}_S$ and removing some edge f from the induced cycle so that e and f do not cross.

  The **crossing-free edge slide metric** is a metric on the set $\mathbf{T}_S$ of all *noncrossing spanning trees* of a set S defined, for any $T_1, T_2 \in \mathbf{T}_S$, as the minimum number of *crossing-free edge slides* necessary for transforming T_1 into T_2. Such slide is one of the edge transformations which consists of taking some edge e in $T \in \mathbf{T}_S$ and moving one of its endpoints along some edge adjacent to e in T, without introducing edge crossings and without sweeping across points in S (that gives a new edge f instead of e). The edge slide is a special kind of crossing-free edge move: the new tree is obtained by closing with f a cycle C of length 3 in T, and removing e from C, in such a way that f avoids the interior of the triangle C.

- **Traveling salesman tours distances**

  The *Traveling Salesman problem* is the problem of finding the shortest tour that visits a set of cities. We will consider only Traveling Salesman problem with undirected links. For an n-city traveling salesman problem, the space $\mathcal{T}_n$ of tours is the set of $\frac{(n-1)!}{2}$ cyclic permutations of the cities $1, 2, \dots, n$.

  The metric D on $\mathcal{T}_n$ is defined in terms of the difference in form: if tours $T, T' \in \mathcal{T}_n$ differ in m links, then $D(T, T') = m$.

  A *k-OPT transformation* of a tour T is obtained by deleting k links from T, and reconnecting. A tour T', obtained from T by a k-OPT transformation, is called a *k-OPT* of T. The distance d on the set $\mathcal{T}_N$ is defined in terms of the 2-OPT transformations: $d(T, T')$ is the minimal i, for which there exists a sequence of i 2-OPT transformations which transforms T to T'. In fact, $d(T, T') \leq D(T, T')$ for any $T, T' \in \mathcal{T}_N$ (see, for example, [MaMo95]). Cf. **arc routing problems**.

- **Orientation distance**

  The **orientation distance** (Chartrand–Erwin–Raines–Zhang, 2001) between two orientations D and D' of a finite graph is the minimum number of arcs of D whose directions must be reversed to produce an orientation isomorphic to D'.

[redacted]

- **Subgraphs distances**

  The standard distance on the set of all subgraphs of a connected graph $G = (V, E)$ is defined by

$$\min\{d_{\text{path}}(u, v) : u \in V(F), v \in V(H)\}$$

  for any subgraphs F, H of G. For any subgraphs F, H of a strongly connected digraph $D = (V, E)$, the standard quasi-distance is defined by

$$\min\{d_{\text{dpath}}(u, v) : u \in V(F), v \in V(H)\}.$$

  Using standard operations (rotation, shift, etc.) on the edge-set of a graph, one gets corresponding distances between its edge-induced subgraphs of given size which are subcases of similar distances on the set of all graphs of a given size and order.

  The **edge rotation distance** on the set $\mathbf{S}^k(G)$ of all edge-induced subgraphs with k edges in a connected graph G is defined as the minimum number of *edge rotations* required to transform $F \in \mathbf{S}^k(G)$ into $H \in \mathbf{S}^k(G)$. We say that H can be obtained from F by an *edge rotation* if there exist distinct vertices u, v, and w in G such that $uv \in E(F)$, $uw \in E(G) \backslash E(F)$, and $H = F - uv + uw$.

  The **edge shift distance** on the set $\mathbf{S}^k(G)$ of all edge-induced subgraphs with k edges in a connected graph G is defined as the minimum number of *edge shifts* required to transform $F \in \mathbf{S}^k(G)$ into $H \in \mathbf{S}^k(G)$. We say that H can be obtained from F by an *edge shift* if there exist distinct vertices u, v and w in G such that $uv, vw \in E(F)$, $uw \in E(G) \backslash E(F)$, and $H = F - uv + uw$.

  The **edge move distance** on the set $\mathbf{S}^k(G)$ of all edge-induced subgraphs with k edges of a graph G (not necessary connected) is defined as the minimum number of *edge moves* required to transform $F \in \mathbf{S}^k(G)$ into $H \in \mathbf{S}^k(G)$. We say that H can be obtained from F by an *edge move* if there exist (not necessarily distinct) vertices u, v, w, and x in G such that $uv \in E(F)$, $wx \in E(G) \backslash E(F)$, and $H = F - uv + wx$. The edge move distance is a metric on $\mathbf{S}^k(G)$. If F and H have s edges in common, then it is equal to $k - s$.

  The **edge jump distance** (which in general can take the value ∞) on the set $\mathbf{S}^k(G)$ of all edge-induced subgraphs with k edges of a graph G (not necessary connected) is defined as the minimum number of *edge jumps* required to transform $F \in \mathbf{S}^k(G)$ into $H \in \mathbf{S}^k(G)$. We say that H can be obtained from F by an *edge jump* if there exist four distinct vertices u, v, w, and x in G such that $uv \in E(F)$, $wx \in E(G) \backslash E(F)$, and $H = F - uv + wx$.

15.4 Distances on Trees

Let T be a *rooted tree*, i.e., a tree with one of its vertices being chosen as the *root*. The *depth* of a vertex v, $depth(v)$, is the number of edges on the path from v to the root. A vertex v is called a *parent* of a vertex u, $v = par(u)$, if they are adjacent, and *depth*$(u) =$ *depth*$(v) + 1$; in this case u is called a *child* of v. A *leaf* is a vertex without child. Two vertices are *siblings* if they have the same parent.

The *in-degree* of a vertex is the number of its children. $T(v)$ is the subtree of T, rooted at a node $v \in V(T)$. If $w \in V(T(v))$, then v is an *ancestor* of w, and w is a *descendant* of v; $nca(u, v)$ is the *nearest common ancestor* of the vertices u and v.

T is called a *labeled tree* if a symbol from a fixed finite alphabet $\mathcal{A}$ is assigned to each node. T is called an *ordered tree* if a left-to-right order among siblings in T is given. On the set $\mathbb{T}_{rlo}$ of all rooted labeled ordered trees there are three *editing operations*:

- *Relabel*—change the label of a vertex v;
- *Deletion*—delete a nonrooted vertex v with parent $v^{\prime}$, making the children of v become the children of $v^{\prime}$; the children are inserted in the place of v as a subsequence in the left-to-right order of the children of $v^{\prime}$;
- *Insertion*—the complement of deletion; insert a vertex v as a child of a $v^{\prime}$ making v the parent of a consecutive subsequence of the children of $v^{\prime}$.

For unordered trees above operations (and so, distances) are defined similarly, but the insert and delete operations work on a subset instead of a subsequence.

We assume that there is a *cost function* defined on each editing operation, and the *cost* of a sequence of editing operations is the sum of the costs of these operations.

The *ordered edit distance mapping* is a representation of the editing operations. Formally, the triple (M, T_1, T_2) is an *ordered edit distance mapping* from T_1 to T_2, $T_1, T_2 \in \mathbb{T}_{rlo}$, if $M \subset V(T_1) \times V(T_2)$ and, for any (v_1, w_1), $(v_2, w_2) \in M$, the following conditions hold: $v_1 = v_2$ if and only if $w_1 = w_2$ (*one-to-one condition*), v_1 is an ancestor of v_2 if and only if w_1 is an ancestor of w_2 (*ancestor condition*), v_1 is to the left of v_2 if and only if w_1 is to the left of w_2 (*sibling condition*).

We say that a vertex v in T_1 and T_2 is *touched by a line* in M if v occurs in some pair in M. Let N_1 and N_2 be the set of vertices in T_1 and T_2, respectively, not touched by any line in M. The *cost* of M is given by $\gamma(M) = \sum_{(v,w)\in M} \gamma(v \rightarrow w) + \sum_{v\in N_1} \gamma(v \rightarrow \lambda) + \sum_{w\in N_2} \gamma(\lambda \rightarrow w)$, where $\gamma(a \rightarrow b) = \gamma(a, b)$ is the *cost* of an editing operation $a \rightarrow b$ which is a relabel if $a, b \in \mathcal{A}$, a deletion if $b = \lambda$, and an insertion if $a = \lambda$. Here $\lambda \notin \mathcal{A}$ is a special *blank symbol*, and γ is a metric on the set $\mathcal{A} \cup \lambda$ (excepting the value $\gamma(\lambda, \lambda)$).

- **Tree edit distance**

  The **tree edit distance** (see [Tai79]) on the set $\mathbb{T}_{rlo}$ of all rooted labeled ordered trees is defined, for any $T_1, T_2 \in \mathbb{T}_{rlo}$, as the minimum cost of a sequence of editing operations (relabels, insertions, and deletions) turning T_1 into T_2.

  In terms of ordered edit distance mappings, it is equal to $\min_{(M,T_1,T_2)} \gamma(M)$, where the minimum is taken over all such mappings (M, T_1, T_2).

  The **unit cost edit distance** between T_1 and T_2 is the minimum number of three above editing operations turning T_1 into T_2, i.e., it is the tree edit distance with cost 1 of any operation.

[redacted]

- **Selkow distance**

  The **Selkow distance** (or *top-down edit distance*, *degree*-1 *edit distance*) is a distance on the set $\mathbb{T}_{rlo}$ of all rooted labeled ordered trees defined, for any $T_1, T_2 \in \mathbb{T}_{rlo}$, as the minimum cost of a sequence of editing operations (relabels, insertions, and deletions) turning T_1 into T_2 if insertions and deletions are restricted to leaves of the trees (see [Selk77]).

  The root of T_1 must be mapped to the root of T_2, and if a node v is to be deleted (inserted), then any subtree rooted at v is to be deleted (inserted).

  In terms of ordered edit distance mappings, it is equal to $\min_{(M,T_1,T_2)} \gamma(M)$, where the minimum is taken over all such mappings (M, T_1, T_2) such that $(par(v), par(w)) \in M$ if $(v, w) \in M$, where neither v nor w is the root.

- **Restricted edit distance**

  The **restricted edit distance** is a distance on the set $\mathbb{T}_{rlo}$ of all rooted labeled ordered trees defined, for any $T_1, T_2 \in \mathbb{T}_{rlo}$, as the minimum cost of a sequence of editing operations (relabels, insertions, and deletions) turning T_1 into T_2 with the restriction that disjoint subtrees should be mapped to disjoint subtrees.

  In terms of ordered edit distance mappings, it is equal to $\min_{(M,T_1,T_2)} \gamma(M)$, where the minimum is taken over all such mappings (M, T_1, T_2) satisfying the following condition: for all (v_1, w_1), (v_2, w_2), $(v_3, w_3) \in M$, $nca(v_1, v_2)$ is a proper ancestor of v_3 if and only if $nca(w_1, w_2)$ is a proper ancestor of w_3.

  This distance is equivalent to the *structure respecting edit distance* which is defined by $\min_{(M,T_1,T_2)} \gamma(M)$. Here the minimum is taken over all ordered edit distance mappings (M, T_1, T_2), satisfying the following condition: for all (v_1, w_1), (v_2, w_2), $(v_3, w_3) \in M$, such that none of v_1, v_2, and v_3 is an ancestor of the others, $nca(v_1, v_2) = nca(v_1, v_3)$ if and only if $nca(w_1, w_2) = nca(w_1, w_3)$.

  Cf. **constrained edit distance** in Chap. 11.

- **Alignment distance**

  The **alignment distance** (see [JWZ94]) is a distance on the set $\mathbb{T}_{rlo}$ of all rooted labeled ordered trees defined, for any $T_1, T_2 \in \mathbb{T}_{rlo}$, as the minimum *cost* of an *alignment* of T_1 and T_2. It corresponds to a restricted edit distance, where all insertions must be performed before any deletions.

  Thus, one inserts *spaces*, i.e., vertices labeled with a *blank symbol* λ, into T_1 and T_2 so that they become isomorphic when labels are ignored; the resulting trees are overlaid on top of each other giving the *alignment* $T_{\mathcal{A}}$ which is a tree, where each vertex is labeled by a pair of labels. The *cost* of $T_{\mathcal{A}}$ is the sum of the costs of all pairs of opposite labels in $T_{\mathcal{A}}$.

- **Splitting-merging distance**

  The **splitting-merging distance** (see [ChLu85]) is a distance on the set $\mathbb{T}_{rlo}$ of all rooted labeled ordered trees defined, for any $T_1, T_2 \in \mathbb{T}_{rlo}$, as the minimum number of vertex splittings and mergings needed to transform T_1 into T_2.

- **Degree-2 distance**

  The **degree-2 distance** is a metric on the set $\mathbb{T}_l$ of all labeled trees (*labeled free trees*), defined, for any $T_1, T_2 \in \mathbb{T}_l$, as the minimum number of editing operations (relabels, insertions, and deletions) turning T_1 into T_2 if any vertex to be inserted (deleted) has no more than two neighbors. This metric is a natural extension of the **tree edit distance** and the **Selkow distance**.

A *phylogenetic X-tree* is an unordered unrooted tree with the labeled leaf set X and no vertices of degree two. If every interior vertex has degree three, the tree is called *binary*. Let $\mathbb{T}(X)$ denote the set of all phylogenetic X-trees.

- **Robinson–Foulds metric**

  A *cut* $A|B$ of X is a *partition* of X into two subsets A and B (see **cut semimetric**). Removing an edge e from a phylogenetic X-tree induces a cut of the leaf set X which is called the *cut associated with* e.

  The **Robinson–Foulds metric** (or *Bourque metric*, *bipartition distance*) is a metric on the set $\mathbb{T}(X)$, defined, for any phylogenetic X-trees $T_1, T_2 \in \mathbb{T}(X)$, by

$$\frac{1}{2}|\Sigma(T_1)\triangle\Sigma(T_2)| = \frac{1}{2}|\Sigma(T_1) \setminus \Sigma(T_2)| + \frac{1}{2}|\Sigma(T_2) \setminus \Sigma(T_1)|,$$

  where $\Sigma(T)$ is the collection of all cuts of X associated with edges of T.

  The **Robinson–Foulds weighted metric** is a metric on the set $\mathbb{T}(X)$ of all phylogenetic X-trees defined by

$$\sum_{A|B\in\Sigma(T_1)\cup\Sigma(T_2)} |w_1(A|B) - w_2(A|B)|$$

  for all $T_1, T_2 \in \mathbb{T}(X)$, where $w_i = (w(e))_{e\in E(T_i)}$ is the collection of positive weights, associated with the edges of the X-tree T_i, $\Sigma(T_i)$ is the collection of all cuts of X, associated with edges of T_i, and $w_i(A|B)$ is the weight of the edge, corresponding to the cut $A|B$ of X, $i = 1, 2$. Cf. more general **cut norm metric** in Chap. 12 and **rectangle distance on weighted graphs**.

- **μ-metric**

  Given a phylogenetic X-tree T with n leaves and a vertex v in it, let $\mu(v) = (\mu_1(v), \dots, \mu_n(v))$, where $\mu_i(v)$ is the number of different paths from the vertex v to the i-th leaf. Let $\mu(T)$ denote the multiset on the vertex-set of T with $\mu(v)$ being the multiplicity of the vertex v.

  The **μ-metric** (Cardona–Roselló–Valiente, 2008) is a metric on the set $\mathbb{T}(X)$ of all phylogenetic X-trees defined, for all $T_1, T_2 \in \mathbb{T}(X)$, by

$$\frac{1}{2}|\mu(T_1)\Delta\mu(T_2)|,$$

  where Δ denotes the symmetric difference of multisets.

  Cf. the **metrics between multisets** in Chap. 1 and the **Dodge–Shiode WebX quasi-distance** in Chap. 22.

- **Nearest neighbor interchange metric**

  The **nearest neighbor interchange metric** (or **crossover metric**) on the set $\mathbb{T}(X)$ of all phylogenetic X-trees, is defined, for all $T_1, T_2 \in \mathbb{T}(X)$, as the

[redacted]

minimum number of *nearest neighbor interchanges* required to transform T_1 into T_2.

A *nearest neighbor interchange* consists of swapping two subtrees in a tree that are adjacent to the same internal edge; the remainder of the tree is unchanged.

- **Subtree prune and regraft distance**

  The **subtree prune and regraft distance** is a metric on the set $\mathbb{T}(X)$ of all phylogenetic X-trees defined, for all $T_1, T_2 \in \mathbb{T}(X)$, as the minimum number of *subtree prune and regraft transformations* required to transform T_1 into T_2.

  A *subtree prune and regraft transformation* proceeds in three steps: one selects and removes an edge uv of the tree, thereby dividing the tree into two subtrees T_u (containing u) and T_v (containing v); then one selects and subdivides an edge of T_v, giving a new vertex w; finally, one connects u and w by an edge, and removes all vertices of degree two.

- **Tree bisection-reconnection metric**

  The **tree bisection-reconnection metric** (or **TBR-metric**) on the set $\mathbb{T}(X)$ of all phylogenetic X-trees is defined, for all $T_1, T_2 \in \mathbb{T}(X)$, as the minimum number of *tree bisection and reconnection* transformations required to transform T_1 into T_2.

  A *tree bisection and reconnection transformation* proceeds in three steps: one selects and removes an edge uv of the tree, thereby dividing the tree into two subtrees T_u (containing u) and T_v (containing v); then one selects and subdivides an edge of T_v, giving a new vertex w, and an edge of T_u, giving a new vertex z; finally one connects w and z by an edge, and removes all vertices of degree two.

- **Quartet distance**

  The **quartet distance** (see [EMM85]) is a distance of the set $\mathbb{T}_b(X)$ of all binary phylogenetic X-trees defined, for all $T_1, T_2 \in \mathbb{T}_b(X)$, as the number of mismatched *quartets* (from the total number $\binom{n}{4}$ possible quartets) for T_1 and T_2.

  This distance is based on the fact that, given four leaves $\{1, 2, 3, 4\}$ of a tree, they can only be combined in a binary subtree in three ways: $(12|34)$, $(13|24)$, or $(14|23)$: the notation $(12|34)$ refers to the binary tree with the leaf set $\{1, 2, 3, 4\}$ in which removing the inner edge yields the trees with the leaf sets $\{1, 2\}$ and $\{3, 4\}$.

- **Triples distance**

  The **triples distance** (see [CPQ96]) is a distance of the set $\mathbb{T}_b(X)$ of all binary phylogenetic X-trees defined, for all $T_1, T_2 \in \mathbb{T}_b(X)$, as the number of triples (from the total number $\binom{n}{3}$ possible triples) that differ (for example, by which leaf is the outlier) for T_1 and T_2.

- **Perfect matching distance**

  The **perfect matching distance** is a distance on the set $\mathbb{T}_{br}(X)$ of all rooted binary phylogenetic X-trees with the set X of n labeled leaves defined, for any $T_1, T_2 \in \mathbb{T}_{br}(X)$, as the minimum number of interchanges necessary to bring the perfect matching of T_1 to the perfect matching of T_2.

  Given a set $A = \{1, \dots, 2k\}$ of $2k$ points, a *perfect matching* of A is a *partition* of A into k pairs. A rooted binary phylogenetic tree with n labeled leaves has a root and $n - 2$ internal vertices distinct from the root. It can be identified with a perfect matching on $2n-2$, different from the root, vertices by following construction: label the internal vertices with numbers $n+1, \dots, 2n-2$ by putting the smallest available label as the parent of the pair of labeled children of which one has the smallest label among pairs of labeled children; now a matching is formed by peeling off the children, or sibling pairs, two by two.

- **Tree rotation distance**

  The **tree rotation distance** is a distance on the set $\mathbf{T}_n$ of all rooted ordered binary trees with n interior vertices defined, for all $T_1, T_2 \in \mathbf{T}_n$, as the minimum number of *rotations*, required to transform T_1 into T_2.

  Given interior edges uv, vv', vv'' and uw of a binary tree, the *rotation* is replacing them by edges uv, uv'', vv' and vw.

  There is a bijection between edge flipping operations in triangulations of convex polygons with $n+2$ vertices and rotations in binary trees with n interior vertices.

- **Attributed tree metrics**

  An *attributed tree* is a triple (V, E, α), where $T = (V, E)$ is the underlying tree, and α is a function which assigns an *attribute vector* $\alpha(v)$ to every vertex $v \in V$. Given two attributed trees (V_1, E_1, α) and (V_2, E_2, β), consider the set of all *subtree isomorphisms* between them, i.e., the set of all isomorphisms $f : H_1 \to H_2, H_1 \subset V_1, H_2 \subset V_2$, between their *induced subtrees*.

  Given a similarity s on the set of attributes, the similarity between isomorphic induced subtrees is defined as $W_s(f) = \sum_{v \in H_1} s(\alpha(v), \beta(f(v)))$. Let ϕ be the isomorphism with maximal similarity $W_s(\phi) = W(\phi)$.

  The following four semimetrics on the set $\mathbf{T}_{att}$ of all attributed trees are used:

$$\max\{|V_1|, |V_2|\} - W(\phi), \quad |V_1| + |V_2| - 2W(\phi) \quad \text{and}$$

$$1 - \frac{W(\phi)}{\max\{|V_1|, |V_2|\}}, \quad 1 - \frac{W(\phi)}{|V_1| + |V_2| - W(\phi)}.$$

  They become metrics on the set of equivalences classes of attributed trees: two such trees (V_1, E_1, α) and (V_2, E_2, β) are called *equivalent* if they are *attribute-isomorphic*, i.e., if there exists an isomorphism $g : V_1 \to V_2$ between the trees such that, for any $v \in V_1$, we have $\alpha(v) = \beta(g(v))$. Then $|V_1| = |V_2| = W(g)$.

- **Maximal agreement subtree distance**

  The **maximal agreement subtree distance** (MAST) is (Finden–Gordon, 1985) a distance of the set $\mathbf{T}$ of all trees defined, for all $T_1, T_2 \in \mathbf{T}$, as the minimum number of leaves removed to obtain a *(greatest) agreement subtree*.

  An *agreement subtree* (or *common pruned tree*) of two trees is an identical subtree that can be obtained from both trees by pruning leaves with the same label.

[redacted]

- **l_p-distances between trees**

  Given $p \geq 1$, the **l_p-distance between trees** T, T' on the same vertex-set is the l_p-metric between their (shortest path) distance matrices D, D' seen as vectors.

  For $p = 1$, it is **edge difference distance** (Williams–Clifford, 1971). For $p = 2$, it is **path difference distance** (Steel–Penny, 1993). For $p = \infty$, it is **path interval** (or **k-interval cospeciation**, as in original Hugguns et al., 2012) **distance**.

Chapter 16
Distances in Coding Theory

Coding Theory deals with the design and properties of *error-correcting codes* for the reliable transmission of information across noisy channels in transmission lines and storage devices. The aim of Coding Theory is to find codes which transmit and decode fast, contain many valid code words, and can correct, or at least detect, many errors. These aims are mutually exclusive, however; so, each application has its own good code.

In communications, a *code* is a rule for converting a piece of information (for example, a letter, word, or phrase) into another form or representation, not necessarily of the same sort. *Encoding* is the process by which a source (object) performs this conversion of information into data, which is then sent to a receiver (observer), such as a data processing system. *Decoding* is the reverse process of converting data which has been sent by a source, into information understandable by a receiver.

An *error-correcting code* is a code in which every data signal conforms to specific rules of construction so that departures from this construction in the received signal can generally be automatically detected and corrected. It is used in computer data storage, for example in dynamic RAM, and in data transmission. Error detection is much simpler than error correction, and one or more "check" digits are commonly embedded in credit card numbers in order to detect mistakes. The two main classes of error-correcting codes are *block codes*, and *convolutional codes*.

A *block code* (or *uniform code*) of length n over an alphabet $\mathcal{A}$, usually, over a finite field $\mathbb{F}_q = \{0, \dots, q-1\}$, is a subset $C \subset \mathcal{A}^n$; every vector $x \in C$ is called a *codeword*, and $M = |C|$ is called *size* of the code. Given a metric d on $\mathbb{F}_q^n$ (for example, the **Hamming metric**, **Lee metric**, **Levenstein metric**), the value $d^* = d^*(C) = \min_{x,y \in C, x \neq y} d(x,y)$ is called the **minimum distance** of the code C. The *weight* $w(x)$ of a codeword $x \in C$ is defined as $w(x) = d(x, 0)$. An (n, M, d^*)-*code* is a q-ary block code of length n, size M, and minimum distance d^*. A *binary code* is a code over $\mathbb{F}_2$.

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_16

[redacted]

When codewords are chosen such that the distance between them is maximized, the code is called *error-correcting*, since slightly garbled vectors can be recovered by choosing the nearest codeword. A code C is a *t-error-correcting code* (and a *$2t$-error-detecting code*) if $d^*(C) \geq 2t + 1$. In this case each *neighborhood* $U_t(x) = \{y \in C : d(x, y) \leq t\}$ of $x \in C$ is disjoint with $U_t(y)$ for any $y \in C, y \neq x$.

A *perfect code* is a q-ary $(n, M, 2t + 1)$-code for which the M spheres $U_t(x)$ of radius t centered on the codewords fill the whole space $\mathbb{F}_q^n$ completely, without overlapping.

A block code $C \subset \mathbb{F}_q^n$ is called *linear* if C is a vector subspace of $\mathbb{F}_q^n$. An $[n, k]$-code is a k-dimensional linear code $C \subset \mathbb{F}_q^n$ (with the minimum distance d^*); it has size q^k, i.e., it is an (n, q^k, d^*)-code. The *Hamming code* is the linear perfect one-error correcting $(\frac{q^r-1}{q-1}, \frac{q^r-1}{q-1} - r, 3)$-code.

A $k \times n$ matrix G with rows that are basis vectors for a linear $[n, k]$-code C is called a *generator matrix* of C. In *standard form* it can be written as $(1_k|A)$, where 1_k is the $k \times k$ identity matrix. Each *message* (or *information symbol*, *source symbol*) $u = (u_1, \dots, u_k) \in \mathbb{F}_q^k$ can be encoded by multiplying it (on the right) by the generator matrix: $uG \in C$.

The matrix $H = (-A^T|1_{n-k})$ is called the *parity-check matrix* of C. The number $r = n - k$ corresponds to the number of parity check digits in the code, and is called the *redundancy* of the code C. The *information rate* (or *code rate*) of a code C is the number $R = \frac{\log_2 M}{n}$. For a q-ary $[n, k]$-code, $R = \frac{k}{n} \log_2 q$; for a binary $[n, k]$-code, $R = \frac{k}{n}$.

A *convolutional code* is a type of error-correction code in which each k-bit information symbol to be encoded is transformed into an n-bit codeword, where $R = \frac{k}{n}$ is the code rate ($n \geq k$), and the transformation is a function of the last m information symbols, where m is the *constraint length* of the code. Convolutional codes are often used to improve the performance of radio and satellite links.

A *variable length code* is a code with codewords of different lengths.

In contrast to error-correcting codes which are designed only to increase the reliability of data communications, *cryptographic codes* are designed to increase their security. In Cryptography, the sender uses a *key* to encrypt a message before it is sent through an insecure channel, and an authorized receiver at the other end then uses a key to decrypt the received data to a message.

Often, data compression algorithms and error-correcting codes are used in tandem with cryptographic codes to yield communications that are efficient, robust to data transmission errors, and secure to eavesdropping and tampering. Encrypted messages which are, moreover, hidden in text, image, etc., are called *steganographic messages*.

The *encryption/assortment theory of humor* (Flamson–Barrett, 2008) proposes that people signal similarity in locally variable personal features through humor. In a successful joke, both the producer and the receiver share common background information—the key—and the joke is engineered in such a way (via devices such as incongruity) that there is a nonrandom fit between the surface utterance and this information that would only be apparent to a person with access to it. The function of encrypted humor is not secrecy per se, but rather, honestly indexing the presence of shared keys.

16.1 Minimum Distance and Relatives

- **Minimum distance**

  Given a code $C \subset V$, where V is an n-dimensional vector space equipped with a metric d, the **minimum distance** $d^* = d^*(C)$ of the code C is defined by

$$\min_{x,y \in C, x \neq y} d(x, y).$$

  The metric d depends on the nature of the errors for the correction of which the code is intended. For a prescribed correcting capacity it is necessary to use codes with a maximum number of codewords. Such most widely investigated codes are the *q-ary block codes* in the **Hamming metric** $d_H(x, y) = |\{i : x_i \neq y_i, i = 1, \dots, n\}|$.

  For a *linear code* C the minimum distance $d^*(C) = w(C)$, where $w(C) = \min\{w(x) : x \in C\}$ is a *minimum weight* of the code C. As there are $\mathrm{rank}(H) \leq n - k$ independent columns in the parity check matrix H of an $[n, k]$-code C, then $d^*(C) \leq n - k + 1$ (*Singleton upper bound*).
- **Dual distance**

  The **dual distance** $d^{\perp}$ of a *linear* $[n, k]$-code $C \subset \mathbb{F}_q^n$ is the **minimum distance** of the *dual code* $C^{\perp}$ of C defined by $C^{\perp} = \{v \in \mathbb{F}_q^n : \langle v, u \rangle = 0$ for any $u \in C\}$.

  The code $C^{\perp}$ is a linear $[n, n-k]$-code, and its $(n - k) \times n$ generator matrix is the parity-check matrix of C.
- **Bar product distance**

  Given linear codes C_1 and C_2 of length n with $C_2 \subset C_1$, their *bar product* $C_1|C_2$ is a linear code of length $2n$ defined by $C_1|C_2 = \{x|x+y : x \in C_1, y \in C_2\}$.

  The **bar product distance** between C_1 and C_2 is the minimum distance $d^*(C_1|C_2)$ of their bar product $C_1|C_2$.
- **Design distance**

  A *linear code* is called a *cyclic code* if all cyclic shifts of a codeword also belong to C, i.e., if for any $(a_0, \dots, a_{n-1}) \in C$ the vector $(a_{n-1}, a_0, \dots, a_{n-2}) \in C$. It is convenient to identify a codeword $(a_0, \dots, a_{n-1})$ with the polynomial $c(x) = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1}$; then every cyclic $[n, k]$-code can be represented as the principal ideal $\langle g(x) \rangle = \{r(x)g(x) : r(x) \in R_n\}$ of the ring $R_n = \mathbb{F}_q(x)/(x^n - 1)$, generated by the *generator polynomial* $g(x) = g_0 + g_1 x + \cdots + x^{n-k}$ of C.

  Given an element α of order n in a finite field $\mathbb{F}_{q^s}$, a *Bose–Chaudhuri–Hocquenghem* $[n, k]$*-code* of **design distance** d is a cyclic code of length n, generated by a polynomial $g(x)$ in $\mathbb{F}_q(x)$ of degree $n - k$, that has roots at α,

[redacted]

$\alpha^2, \dots, \alpha^{d-1}$. The minimum distance d^* of such a code of odd design distance d is at least d.

A *Reed–Solomon code* is a Bose–Chaudhuri–Hocquenghem code with $s = 1$. The generator polynomial of a Reed–Solomon code of design distance d is $g(x) = (x - \alpha)\dots(x - \alpha^{d-1})$ with degree $n - k = d - 1$; that is, for a Reed–Solomon code the design distance $d = n-k+1$, and the minimum distance $d^* \geq d$. Since, for a linear $[n, k]$-code, the minimum distance $d^* \leq n-k+1$ (*Singleton upper bound*), a Reed–Solomon code achieves this bound. Compact disc players use a double-error correcting $(255, 251, 5)$ Reed–Solomon code over $\mathbb{F}_{256}$.

- **Goppa designed minimum distance**

  The **Goppa designed minimum distance** ([Gopp71]) is a lower bound $d^\star(m)$ for the minimum distance of *one-point geometric Goppa codes* (or *algebraic geometry codes*) $G(m)$. For $G(m)$, associated to the divisors D and mP, $m \in \mathbb{N}$, of a smooth projective absolutely irreducible algebraic curve of genus $g > 0$ over a finite field $\mathbb{F}_q$, one has $d^\star(m) = m + 2 - 2g$ if $2g - 2 < m < n$.

  In fact, for a Goppa code $C(m)$ the structure of the gap sequence at P may allow one to give a better lower bound of the minimum distance (cf. **Feng–Rao distance**).

- **Feng–Rao distance**

  The **Feng–Rao distance** $\delta_{FR}(m)$ is a lower bound for the minimum distance of *one-point geometric Goppa codes* $G(m)$ which is better than the **Goppa designed minimum distance**. The method of Feng and Rao for encoding the code $C(m)$ decodes errors up to half the Feng–Rao distance $\delta_{FR}(m)$, and gives an improvement of the number of errors that one can correct for one-point geometric Goppa codes.

  Formally, the Feng–Rao distance is defined as follows. Let S be a subsemigroup S of $\mathbb{N} \cup \{0\}$ such that the *genus* $g = |\mathbb{N} \cup \{0\} \backslash S|$ of S is finite, and $0 \in S$. The **Feng–Rao distance** on S is a function $\delta_{FR} : S \to \mathbb{N} \cup \{0\}$ such that $\delta_{FR}(m) = \min\{\nu(r) : r \geq m, r \in S\}$, where $\nu(r) = |\{(a, b) \in S^2 : a + b = r\}|$.

  The generalized ***r*-th Feng–Rao distance** on S is $\delta^r_{FR}(m) = \min\{\nu[m_1, \dots, m_r] : m \leq m_1 < \dots < m_r, m_i \in S\}$, where $\nu[m_1, \dots, m_r] = |\{a \in S : m_i - a \in S \text{ for some } i = 1, \dots, r\}|$. Then $\delta_{FR}(m) = \delta^1_{FR}(m)$. See, for example, [FaMu03].

- **Free distance**

  The **free distance** is the minimum nonzero *Hamming weight* of any codeword in a *convolutional code* or a *variable length code*.

  Formally, the ***k*-th minimum distance** d^*_k of such code is the smallest Hamming distance between any two initial codeword segments which are k frame long and disagree in the initial frame. The sequence $d^*_1, d^*_2, d^*_3, \dots$ $(d^*_1 \leq d^*_2 \leq d^*_3 \leq \dots)$ is called the *distance profile* of the code. The **free distance** of a convolutional code or a variable length code is $\max_l d^*_l = \lim_{l\to\infty} d^*_l = d^*_\infty$.

- **Effective free distance**

  A *turbo code* is a long *block code* in which there are L input bits, and each of these bits is encoded q times. In the j-th encoding, the L bits are sent through a permutation box P_j, and then encoded via an $[N_j, L]$ block encoder (*code fragment* *encoder*) which can be thought of as an $L \times N_j$ matrix. The overall turbo code is then a *linear* $[N_1 + \cdots + N_q, L]$-code (see, for example, [BGT93]).

The *weight-i input minimum distance* $d^i(C)$ of a turbo code C is the minimum weight among codewords corresponding to input words of weight i. The **effective free distance** of C is its *weight-2 input minimum distance* $d^2(C)$, i.e., the minimum *weight* among codewords corresponding to input words of weight 2.

Turbo codes were the first practical codes to closely approach the *Shannon limit* (or *channel capacity*), the theoretical limit of maximum information transfer rate over a symmetric memory-less noisy channel. These codes are used in 3G mobile and satellite communications. Another capacity-approaching codes with similar performance are linear LDPC (*low-density parity-check*) codes.

- **Distance distribution**

  Given a code C over a finite metric space (X, d) with the diameter $diam(X, d) = D$, the **distance distribution** of C is a $(D+1)$-vector $(A_0, \dots, A_D)$, where $A_i = \frac{1}{|C|}|\{(c, c') \in C^2 : d(c, c') = i\}|$. That is, one considers $A_i(c)$ as the number of code words at distance i from the codeword c, and takes A_i as the average of $A_i(c)$ over all $c \in C$. $A_0 = 1$ and, if $d^* = d^*(C)$ is the minimum distance of C, then $A_1 = \cdots = A_{d^*-1} = 0$.

  The distance distribution of a code with given parameters is important, in particular, for bounding the probability of decoding error under different decoding procedures from *maximum likelihood* decoding to error detection. It can also be helpful in revealing structural properties of codes and establishing nonexistence of some codes.

- **Unicity distance**

  The **unicity distance** of a cryptosystem (Shannon, 1949) is the minimal length of a cyphertext that is required in order to expect that there exists only one meaningful decryption for it. For classic cryptosystems with fixed key space, the unicity distance is approximated by the formula $H(K)/D$, where $H(K)$ is the *key space entropy* (roughly $\log_2 N$, where N is the number of keys), and D measures the *redundancy* of the plaintext source language in bits per letter.

  A cryptosystem offers perfect secrecy if its unicity distance is infinite. For example, the *one-time pads* offer perfect secrecy; they were used for the "red telephone" between the Kremlin and the White House.

  More generally, **Pe-security distance** of a cryptosystem (Tilburg–Boekee, 1987) is the minimal expected length of cyphertext that is required in order to break the cryptogram with an average error probability of at most Pe.

16.2 Main Coding Distances

- **Arithmetic codes distance**

  An *arithmetic code* (or *code with correction of arithmetic errors*) is a finite subset of the set $\mathbb{Z}$ of integers (usually, nonnegative integers). It is intended for the control of the functioning of an *adder* (a module performing addition). When

[redacted]

adding numbers represented in the binary number system, a single slip in the functioning of the adder leads to a change in the result by some power of 2, thus, to a single *arithmetic error*. Formally, a single *arithmetic error* on $\mathbb{Z}$ is defined as a transformation of a number $n \in \mathbb{Z}$ to a number $n' = n \pm 2^i$, $i = 1, 2, \dots$.

The **arithmetic codes distance** is a metric on $\mathbb{Z}$ defined, for any $n_1, n_2 \in \mathbb{Z}$, as the minimum number of *arithmetic errors* taking n_1 to n_2. It is $w_2(n_1 - n_2)$, where $w_2(n)$ is the *arithmetic* 2-*weight* of n, i.e., the smallest possible number of nonzero coefficients in representations $n = \sum_{i=0}^{k} e_i 2^i$, where $e_i = 0, \pm 1$, and k is some nonnegative integer. For each n there is a unique such representation with $e_k \neq 0$, $e_i e_{i+1} = 0$ for all $i = 0, \dots, k-1$, which has the smallest number of nonzero coefficients (cf. **arithmetic r-norm metric** in Chap. 12).

- ***b*-burst metric**

  Given the number $b > 1$ and the set $\mathbb{Z}_m^n = \{0, 1, \dots, m-1\}^n$, each its element $x = (x_1, \dots, x_n)$ can be uniquely represented as

$$(0^{k_1} u_1 v_1^{b-1} 0^{k_2} u_2 v_2^{b-1} \dots),$$

where $u_i \neq 0$, 0^k is the string of $k \geq 0$ zeroes and v^{b-1} is any string of length $b-1$.

The ***b*-burst metric** between elements x and y of $\mathbb{Z}_m^n$ is (Bridewell and Wolf, 1979) the number of b-tuples uv^{b-1} in $x - y$. It describes the burst errors.

- **Sharma–Kaushik metrics**

  Let $q \geq 2$, $m \geq 2$. A *partition* $\{B_0, B_1, \dots B_{q-1}\}$ of $\mathbb{Z}_m$ is called a *Sharma–Kaushik partition* if the following conditions hold:

1. $B_0 = \{0\}$;
2. For any $i \in Z_m$, $i \in B_s$ if and only if $m - i \in B_s$, $s = 1, 2, \dots, q-1$;
3. If $i \in B_s, j \in B_t$, and $s > t$, then $\min\{i, m-i\} > \min\{j, m-j\}$;
4. If $s \geq t$, $s, t = 0, 1, \dots, q-1$, then $|B_s| \geq |B_t|$ except for $s = q-1$ in which case $|B_{q-1}| \geq \frac{1}{2}|B_{q-2}|$.

Given a Sharma–Kaushik partition of $\mathbb{Z}_m$, the *Sharma-Kaushik weight* $w_{SK}(x)$ of any element $x \in \mathbb{Z}_m$ is defined by $w_{SK}(x) = i$ if $x \in B_i$, $i \in \{0, 1, \dots, q-1\}$.

The **Sharma–Kaushik metric** ([ShKa79]) is a metric on $\mathbb{Z}_m$ defined by

$$w_{SK}(x - y).$$

The Sharma–Kaushik metric on Z_m^n is defined by $w_{SK}^n(x - y)$ where, for $x = (x_1, \dots x_n) \in \mathbb{Z}_m^n$, one has $w_{SK}^n(x) = \sum_{i=1}^{n} w_{SK}(x_i)$.

The **Hamming metric** and the **Lee metric** arise from two specific partitions of the above type: $P_H = \{B_0, B_1\}$, where $B_1 = \{1, 2, \dots, q-1\}$, and $P_L = \{B_0, B_1, \dots, B_{\lfloor q/2 \rfloor}\}$, where $B_i = \{i, m-i\}$, $i = 1, \dots, \lfloor \frac{q}{2} \rfloor$.

- **Varshamov metric**

  The **Varshamov metric** between two binary n-vectors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ from $\mathbb{Z}_2^n = \{0, 1\}^n$ is defined by

$$\max(\sum_{i=1}^{n} I_{x_i=1-y_i=0}, \sum_{i=1}^{n} I_{x_i=1-y_i=1}).$$

  This metric was introduced by Varshamov, 1965, to describe asymmetric errors.

- **Absolute summation distance**

  The **absolute summation distance** (or *Lee distance*) is the **Lee metric** on the set $\mathbb{Z}_m^n = \{0, 1, \dots, m-1\}^n$ defined by

$$w_{Lee}(x-y),$$

  where $w_{Lee}(x) = \sum_{i=1}^{n} \min\{x_i, m-x_i\}$ is the *Lee weight* of $x = (x_1, \dots, x_n) \in \mathbb{Z}_m^n$.

  If $\mathbb{Z}_m^n$ is equipped with the absolute summation distance, then a subset C of $\mathbb{Z}_m^n$ is called a *Lee distance code*. The most important such codes are *negacyclic codes*.

- **Mannheim distance**

  The **Mannheim distance** is a 2D generalization of the **Lee metric**.

  Let $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$ be the set of *Gaussian integers*. Let $\pi = a + bi$ $(a > b > 0)$ be a *Gaussian prime*, i.e., either

  (i) $(a + bi)(a - bi) = a^2 + b^2 = p$, where $p \equiv 1 \pmod 4$ is a prime number, or
  (ii) up to an integer, $\pi = p + 0 \cdot i$, where $p \equiv 3 \pmod 4$ is a prime number.

  The **Mannheim distance** is not a metric; it is defined ([Hube94a]), for any $x, y \in \mathbb{Z}[i]$, as $|x'| + |y'|$, where $x' + y'i = x - y \pmod \pi$, which is defined as $(x - y) - [\frac{(x-y)\pi}{\pi\overline{\pi}}]\pi$ in the case (i). Here [.] denotes rounding to the *closest Gaussian integer*, i.e., $[c + di] = [c] + [d]i$ with $[c]$ denoting the rounding to the closest integer.

  In general, the elements of the finite field $\mathbb{F}_p = \{0, 1, \dots, p-1\}$ for $p \equiv 1 \pmod 4$, $p = a^2 + b^2$, and of the finite field $\mathbb{F}_{p^2}$ for $p \equiv 3 \pmod 4$, $p = a$, can be mapped on a subset of $\mathbb{Z}[i]$ using the complex modulo function $\mu(k) = k - [\frac{k(a-bi)}{p}](a + bi)$, $k = 0, \dots, p-1$. The set of the selected Gaussian integers $a+bi$ with the minimal *complex modulus* norms $\sqrt{(a + bi)(a - bi)} = \sqrt{a^2 + b^2}$ is called a *constellation*.

  The Mannheim distance between two vectors over $\mathbb{Z}[i]$ is the sum of the Mannheim distances of corresponding components. It was introduced to make 2D QAM-like signals more susceptible to algebraic decoding methods.

  For codes over hexagonal signal constellations, a similar metric was introduced over $\mathbb{Z}(\frac{i\sqrt{3}+1}{2})$ in [Hube94b]. Cf. $\mathbb{Z}(\eta_m)$**-related norm metrics** in Chap. 12.

[redacted]

- **Generalized Lee metric**

  Let $\mathbb{F}_{p^m}$ denote the finite field with p^m elements, where p is prime number and $m \geq 1$ is an integer. Let $e_i = (0, \dots, 0, 1, 0, \dots, 0)$, $1 \leq i \leq k$, be the standard basis of $\mathbb{Z}^k$. Choose elements $a_i \in \mathbb{F}_{p^m}$, $1 \leq i \leq k$, and the mapping $\phi : \mathbb{Z}^k \to \mathbb{F}_{p^m}$, sending any $x = \sum_{i=1}^{k} x_i e_i$, $x_i \in \mathbb{Z}^k$, to $\phi(x) = \sum_{i=1}^{k} a_i x_i (\text{mod } p)$, so that ϕ is surjective. So, for each $a \in \mathbb{F}_{p^m}$, there exists $x \in \mathbb{Z}^k$ such that $a = \phi(x)$. For each $a \in \mathbb{F}_{p^m}$, its *k-dimensional Lee weight* is $w_{kL}(a) = \min\{\sum_{i=1}^{k} |x_i| : x = (x_i) \in \mathbb{Z}, a = \phi(x)\}$.

  The **generalized Lee metric** between vectors (a_j) and (b_j) of $\mathbb{F}_{p^m}^n$ is defined (Nishimura–Hiramatsu, 2008) by

$$\sum_{j=1}^{n} w_{kL}(a_j - b_j).$$

  It is the **Lee metric** (or **absolute summation distance**) if $\phi(e_1) = 1$ while $\phi(e_i) = 0$ for $2 \leq i \leq k$. It is the **Mannheim distance** if $k = 2, p \equiv 1 \pmod 4$, $\phi(e_1) = 1$ while $\phi(e_2) = a$ is a solution in $\mathbb{F}_p$ of the quadratic congruence $x^2 \equiv -1 \pmod p$.

- ***p*-Lee metric**

  The ***p*-Lee metric** on $\mathbb{Z}_q^n$ is given by $d_p^p(x, y) = \sum_{i=1}^{n} d(x_i, y_i)^p$ if $p \in [1, \infty)$, and $d(x, y) = \max_{i=1}^{n} d(x_i, y_i)$ if $p = \infty$, where $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n) \in \mathbb{Z}_q^n$, and $d(x_i, y_i)$ is the **Lee metric** $\min\{|x_i - y_i|, q - |x_i - y_i|\}$ on $\mathbb{Z}$.

- **Poset metric**

  Let $(V_n, \preceq)$ be a *poset* on $V_n = \{1, \dots, n\}$. A subset I of V_n is called *ideal* if $x \in I$ and $y \preceq x$ imply that $y \in I$. If $J \subset V_n$, then $\langle J \rangle$ denotes the smallest ideal of V_n which contains J. Consider the vector space $\mathbb{F}_q^n$ over a finite field $\mathbb{F}_q$. The *P-weight* of an element $x = (x_1, \dots, x_n) \in \mathbb{F}_q^n$ is defined as the cardinality of the smallest ideal of V_n containing the *support* of x: $w_P(x) = |\langle supp(x) \rangle|$, where $supp(x) = \{i : x_i \neq 0\}$.

  The **poset metric** (see [BGL95]) is a metric on $\mathbb{F}_q^n$ defined by

$$w_P(x - y).$$

  If $\mathbb{F}_q^n$ is equipped with a poset metric, then a subset C of $\mathbb{F}_q^n$ is called a *poset code*. If V_n forms the chain $1 \leq 2 \leq \cdots \leq n$, then the linear code C of dimension k consisting of all vectors $(0, \dots, 0, a_{n-k+1}, \dots, a_n) \in \mathbb{F}_q^n$ is a perfect poset code with the minimum (poset) metric $d_P^*(C) = n - k + 1$.

  If V_n forms an antichain, then the poset distance coincides with the **Hamming metric**. If V_n consists of finite disjoint union of chains of equal lengths, then the poset distance coincides with the **NRT metric**.

- **Rank metric**

  Let $\mathbb{F}_q$ be a finite field, $\mathbb{K} = \mathbb{F}_{q^m}$ an extension of degree m of $\mathbb{F}_q$, and $\mathbb{K}^n$ a vector space of dimension n over $\mathbb{K}$. For any $a = (a_1, \dots a_n) \in \mathbb{K}^n$ define its *rank*, $\operatorname{rank}(a)$, as the dimension of the vector space over $\mathbb{F}_q$, generated by $\{a_1, \dots, a_n\}$.

  The **rank metric** (Delsarte, 1978) is a metric on $\mathbb{K}^n$ defined by

  $$\operatorname{rank}(a - b).$$

  A *constant rank-distance k set* is (Gow et al., 2014) a set U of $n \times n$ matrices over a field $\mathbb{F}$ such that $\operatorname{rank}(A - B) = k$ for all $A, B \in U, A \neq B$ and $\operatorname{rank}(A) = k$ for all $A \in U, A \neq 0$. Such set is called a *partial spread set* if $k = n$; it defines a partial spread in the $(2n - 1)$-dimensional *projective*, *hermitian polar* or *symplectic polar space*, if U consists of arbitrary, hermitian or symmetric matrices, respectively.
- **Gabidulin–Simonis metrics**

  Let $\mathbb{F}_q^n$ be the vector space over a finite field $\mathbb{F}_q$ and let $F = \{F_i : i \in I\}$ be a finite family of its subsets such that the minimal linear subspace of $\mathbb{F}_q^n$ containing $\cup_{i \in I} F_i$ is $\mathbb{F}_q^n$. Without loss of generality, F can be an antichain of linear subspaces of $\mathbb{F}_q^n$.

  The *F-weight* w_F of a vector $x = (x_1, \dots, x_n) \in \mathbb{F}_q^n$ is the smallest $|J|$ over such subsets $J \subset I$ that x belongs to the minimal linear subspace of $\mathbb{F}_q^n$ containing $\cup_{i \in J} F_i$. A **Gabidulin–Simonis metric** (or **F-distance**, see [GaSi98]) on $\mathbb{F}_q^n$ is defined by

  $$w_F(x - y).$$

  The **Hamming metric** corresponds to the case of $F_i, i \in I$, forming the standard basis. The **Vandermonde metric** is F-distance with $F_i, i \in I$, being the columns of a generalized Vandermonde matrix. Among other examples are: the **rank metric** and the *combinatorial metrics* (by Gabidulin, 1984), including the **b-burst metric**.
- **Subspace metric**

  Let $\mathbb{F}_q^n$ be the vector space over a finite field $\mathbb{F}_q$ and let $\mathcal{P}_{n,q}$ be the set of all subspaces of $\mathbb{F}_q^n$. For any subspace $U \in \mathcal{P}_{n,q}$, let $dim(U)$ denote its dimension and let $U^\perp = \{v \in \mathbb{F}_q^n : \langle u, v \rangle = 0 \text{ for all } u \in U\}$ be its orthogonal space.

  Let $U + V = \{u + v : u \in U, v \in V\}$, i.e., $U + V$ is the smallest subspace of $\mathbb{F}_q^n$ containing both V and V. Then $dim(U + V) = dim(U) + dim(V) - dim(U \cap V)$. If $U \cap V = \emptyset$, then $U + V$ is a direct sum $U \oplus V$.

  The **subspace metric** between two subspaces U and V from $\mathcal{P}_{n,q}$ is defined by

  $$d(U, V) = dim(U + V) - dim(U \cap V) = dim(U) + dim(V) - 2dim(U \cap V).$$

  This metric was introduced by Koetter and Kschischang, 2007, for network coding. It holds $d(U, V) = d(U^\perp, V^\perp)$. Cf. the **lattice valuation metric** in Chap. 10 and distances between subspaces in Chap. 12.

[redacted]

- **NRT metric**

  Let $M_{m,n}(\mathbb{F}_q)$ be the set of all $m \times n$ matrices with entries from a finite field $\mathbb{F}_q$ (in general, from any finite alphabet $\mathcal{A} = \{a_1, \dots, a_q\}$). The *NRT norm* $||.||_{RT}$ on $M_{m,n}(\mathbb{F}_q)$ is defined as follows: if $m = 1$ and $a = (\xi_1, \xi_2, \dots, \xi_n) \in M_{1,n}(\mathbb{F}_q)$, then $||0_{1,n}||_{RT} = 0$, and $||a||_{RT} = \max\{i : \xi_i \neq 0\}$ for $a \neq 0_{1,n}$; if $A = (a_1, \dots, a_m)^T \in M_{m,n}(\mathbb{F}_q)$, $a_j \in M_{1,n}(\mathbb{F}_q)$, $1 \leq j \leq m$, then $||A||_{RT} = \sum_{j=1}^{m} ||a_j||_{RT}$.

  The **NRT metric** (or *Niederreiter–Rosenbloom–Tsfasman metric*, since introduced by Niederreiter, 1991, and Rosenbloom–Tsfasman, 1997; or *ordered Hamming distance*, in ([MaSt99]) is a **matrix norm metric** (in fact, an **ultrametric**) on $M_{m,n}(\mathbb{F}_q)$, defined by

  $$||A - B||_{RT}.$$

  For every matrix code $C \subset M_{m,n}(F_q)$ with q^k elements the minimum NRT distance $d^*_{RT}(C) \leq mn - k + 1$. Codes meeting this bound are called *maximum distance separable codes*.

  The most used distance between codewords of a matrix code $C \subset M_{m,n}(F_q)$ is the **Hamming metric** on $M_{m,n}(F_q)$ defined by $||A - B||_H$, where $||A||_H$ is the *Hamming weight* of a matrix $A \in M_{m,n}(\mathbb{F}_q)$, i.e., the number of its nonzero entries.

  The **LRTJ-metric** (introduced as *Generalized Lee–Rosenbloom–Tsfasman pseudo-metric* by Jain, 2008) is the **norm metric** for the following generalization of the above norm $||a||_{RT}$ in the case $a \neq 0_{1,n}$:

  $$||a||_{LRTJ} = \max_{1 \leq i \leq n} \min\{\xi_i, q - \xi_i\} + \max\{i - 1 : \xi_i \neq 0\}.$$

  It is the **Lee metric** for $m = 1$ and the NRT metric for $q = 2, 3$.

- **ACME distance**

  The **ACME distance** on a code $C \subset \mathcal{A}^n$ over an alphabet $\mathcal{A}$ is defined by

  $$\min\{d_H(x, y), d_I(x, y)\},$$

  where d_H is the **Hamming metric**, and d_I is the **swap metric** (Chap. 11), i.e., the minimum number of interchanges of adjacent pairs of symbols, converting x into y.

- **Indel distance**

  Let W be the set of all words over an alphabet $\mathcal{A}$. A *deletion* of a letter in a word $\beta = b_1 \dots b_n$ of the length n is a transformation of β into a word $\beta' = b_1 \dots b_{i-1} b_{i+1} \dots b_n$ of the length $n - 1$. An *insertion* of a letter in a word $\beta = b_1 \dots b_n$ of the length n is a transformation of β into a word $\beta'' = b_1 \dots b_i b b_{i+1} \dots b_n$, of the length $n + 1$.

The **indel distance** is a metric on W, defined as the minimum number of deletions and insertions of letters converting α into β. Cf. **indel metric** in Chap. 11.

A *code C with correction of deletions and insertions* is an arbitrary finite subset of W. An example of such a code is the set of words $\beta = b_1 \dots b_n$ of length n over the alphabet $\mathcal{A} = \{0, 1\}$ for which $\sum_{i=1}^{n} ib_i \equiv 0 \pmod{n+1}$. The number of words in this code is equal to $\frac{1}{2(n+1)} \sum_k \phi(k) 2^{(n+1)/k}$, where the sum is taken over all odd divisors k of $n+1$, and ϕ is the *Euler function*.

- **Interval distance**

  The **interval distance** (see, for example, [Bata95]) is a metric on a finite group $(G, +, 0)$ defined by

$$w_{int}(x - y),$$

  where $w_{int}(x)$ is an *interval weight* on G, i.e., a *group norm* whose values are consecutive nonnegative integers $0, \dots, m$. This distance is used for *group codes* $C \subset G$.

- **Fano metric**

  The **Fano metric** is a *decoding metric* with the goal to find the best sequence estimate used for the *Fano algorithm* of *sequential decoding* of *convolutional codes*. In a *convolutional code* each k-bit information symbol to be encoded is transformed into an n-bit codeword, where $R = \frac{k}{n}$ is the code rate ($n \geq k$), and the transformation is a function of the last m information symbols.

  The linear time-invariant decoder (*fixed convolutional decoder*) maps an information symbol $u_i \in \{u_1, \dots, u_N\}$, $u_i = (u_{i1}, \dots u_{ik})$, $u_{ij} \in \mathbb{F}_2$, into a codeword $x_i \in \{x_1, \dots, x_N\}$, $x_i = (x_{i1}, \dots, x_{in})$, $x_{ij} \in \mathbb{F}_2$, so one has a code $\{x_1, \dots, x_N\}$ with N codewords which occur with probabilities $\{p(x_1), \dots, p(x_N)\}$. A sequence of l codewords forms a *path* $x = x_{[1,l]} = \{x_1, \dots, x_l\}$ which is transmitted through a *discrete memoryless channel*, resulting in the received sequence $y = y_{[1,l]}$.

  The task of a decoder minimizing the sequence error probability is to find a sequence maximizing the joint probability of input and output channel sequences $p(y, x) = p(y|x) \cdot p(x)$. Usually it is sufficient to find a procedure that maximizes $p(y|x)$, and a decoder that always chooses as its estimate one of the sequences that maximizes it or, equivalently, the **Fano metric**, is called a *max-likelihood decoder*.

  Roughly, we consider each code as a tree, where each branch represents one codeword. The decoder begins at the first vertex in the tree, and computes the branch metric for each possible branch, determining the best branch to be the one corresponding to the codeword x_j resulting in the largest branch metric, $\mu_F(x_j)$.

  This branch is added to the path, and the algorithm continues from the new node which represents the sum of the previous node and the number of bits in the current best codeword. Through iterating until a terminal node of the tree is reached, the algorithm traces the most likely path.

[redacted]

In this construction, the **bit Fano metric** is defined by

$$\log_2 \frac{p(y_i|x_i)}{p(y_i)} - R,$$

the **branch Fano metric** is defined by

$$\mu_F(x_j) = \sum_{i=1}^{n} (\log_2 \frac{p(y_i|x_{ji})}{p(y_i)} - R),$$

and the **path Fano metric** is defined by

$$\mu_F(x_{[1,l]}) = \sum_{j=1}^{l} \mu_F(x_j),$$

where $p(y_i|x_{ji})$ are the channel transition probabilities, $p(y_i) = \sum_{x_m} p(x_m)p(y_i|x_m)$ is the probability distribution of the output given the input symbols averaged over all input symbols, and $R = \frac{k}{n}$ is the code rate.

For a hard-decision decoder $p(y_j = 0|x_j = 1) = p(y_j = 1|x_j = 0) = p$, $0 < p < \frac{1}{2}$, the Fano metric for a path $x_{[1,l]}$ can be written as

$$\mu_F(x_{[1,l]}) = -\alpha d_H(y_{[1,l]}, x_{[1,l]}) + \beta \cdot l \cdot n,$$

where $\alpha = -\log_2 \frac{p}{1-p} > 0$, $\beta = 1 - R + \log_2(1-p)$, and d_H is the **Hamming metric**.

The **generalized Fano metric** is defined, for $0 \leq w \leq 1$, by

$$\mu_F^w(x_{[1,l]}) = \sum_{j=1}^{ln} \left(\log_2 \frac{p(y_j|x_j)^w}{p(y_j)^{1-w}} - wR \right).$$

For $w = 1/2$, it is the Fano metric with a multiplicative constant $1/2$.

- **Channel metrization**

  A *square channel* over $[n] = \{1, \ldots, n\}$ is an $n \times n$ probability matrix $P = (()P_{ij})$ such that P_{ij} is the probability $P(j|i) = P(j$ received if i sent).

  For a code $C \subseteq [n]$, the *maximum likelihood decoder* (MLD) decodes j as $c \in C$ maximizing $P(j|c)$. If a metric d is defined on $[n]$, then the *minimum distance decoder* (MDD) decodes j as $c \in C$ minimizing $d(j, c)$.

  Channel metrization is (D'Oliveira–Firer, 2015), for a given channel P over $[n]$, to find a metric d on $[n]$ with coinciding decoders MLD and MDD, i.e., for every $C \subseteq [n]$ and $j \in [n]$, $\arg\min\{d(j, c) : c \in C\} = \arg\max\{P(j|c) : c \in C\}$. Main example: metrization of the binary symmetric channel by the Hamming metric.

- **Metric recursion of a MAP decoding**

  Maximum a posteriori sequence estimation, or *MAP decoding* for *variable length codes*, used the *Viterbi algorithm*, and is based on the **metric recursion**

$$\Lambda_k^{(m)} = \Lambda_{k-1}^{(m)} + \sum_{n=1}^{l_k^{(m)}} x_{k,n}^{(m)} \log_2 \frac{p(y_{k,n}|x_{k,n}^{(m)} = +1)}{p(y_{k,n}|x_{k,n}^{(m)} = -1)} + 2\log_2 p(u_k^{(m)}),$$

  where $\Lambda_k^{(m)}$ is the **branch metric** of branch m at time (level) k, $x_{k,n}$ is the n-th bit of the codeword having $l_k^{(m)}$ bits labeled at each branch, $y_{k,n}$ is the respective received soft-bit, u_k^m is the source symbol of branch m at time k and, assuming statistical independence of the source symbols, the probability $p(u_k^{(m)})$ is equivalent to the probability of the source symbol labeled at branch m, that may be known or estimated. The metric increment is computed for each branch, and the largest value, when using *log-likelihood* values, of each state is used for further recursion. The decoder first computes the metric of all branches, and then the branch sequence with largest metric starting from the final state backward is selected.
- **Distance decoder**

  A graph family A is said (Peleg, 2000) to have an $l(n)$ **distance labeling scheme** if there is a function L_G labeling the vertices of each n-vertex graph $G \in A$ with distinct labels up to $l(n)$ bits, and there exists an algorithm, called a **distance decoder**, that decides the distance $d(u, v)$ between any two vertices $u, v \in X$ in a graph $G \in A$, i.e.,$d(u, v) = f(L_G(u), L_G(v))$, polynomial in time in the length of their labels $L(u), L(v)$.

  Cf. **distance constrained labeling** in Chap. 15.
- **Identifying code**

  Let $G = (X, E)$ be a digraph and $C \subset V$, and let $B(v)$ denote the set consisting of v and all of its incoming neighbors in G. If the sets $B(v) \cap C$ are nonempty and distinct, C is called **identifying code** of G. Such sets of smallest cardinality are called (Karpovsky–Chakrabarty–Levitin, 1998) *minimum identifying codes*; denote this cardinality by $M(G)$. An **r-locating-dominating set** (Chap. 15) with $r = 1$ differs from an identifying code only in that $B(v) \cap C$ are not required to be unique *identifying sets* for $v \in C$.

  A *minimum identifying code graph* of order n is a graph $G = (X, E)$ with $X = n$ and $M(G) = \lceil \log 2_2(n + 1)\rceil$ having the minimum number of edges $|E|$.

[redacted]

Chapter 17
Distances and Similarities in Data Analysis

A *data set* is a finite set comprising m sequences $(x_1^j, \dots, x_n^j)$, $j \in \{1, \dots, m\}$, of length n. The values $x_i^1, \dots, x_i^m$ represent an *attribute* S_i.

Among numerical data, **metric data** is any reading at an *interval scale*, measuring the degree of difference between items, or at a *ratio scale* measuring the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind; with them one have a meter permitting define distances between scale values. **Nonmetric** (or *categorial*, *qualitative*) **data** are collected from *binary* (presence/absence expressed by $1/0$), *ordinal* (numbers expressing rank only), or *nominal* (items are not ordered) *scale*.

Geometric data analysis refer to geometric aspects of image, pattern and shape analysis that treats arbitrary data sets as clouds of points in $\mathbb{R}^n$.

Often data are organized in a **metric database** (especially, **metric tree**), i.e., a database indexed in a metric space. The term *metric indexing* is also used.

Cluster Analysis (or *Classification*, *Taxonomy*, *Pattern Recognition*) consists mainly of partition of data A into a relatively small number of *clusters*, i.e., such sets of objects that (with respect to a selected measure of distance) are at best possible degree, “close” if they belong to the same cluster, “far” if they belong to different clusters, and further subdivision into clusters will impair the above two conditions.

We give three typical examples. In *Information Retrieval* applications, nodes of peer-to-peer database network export data (collection of text documents); each document is characterized by a vector from $\mathbb{R}^n$. An user needs to retrieve all documents in the database which are *relevant* to a query object (say, a vector $x \in \mathbb{R}^n$), i.e., belong to the *ball* in $\mathbb{R}^n$, center x, of fixed radius and with a convenient distance function. Such similarity query is called a **metric range query**. In *Record Linkage*, each document (database record) is represented by a term-frequency vector $x \in \mathbb{R}^n$ or a string, and one wants to measure semantic relevancy of syntactically different records. In *Ecology*, let x, y be *species abundance distributions*, obtained by two sample methods (i.e., x_j, y_j are the numbers of individuals of species j,

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_17

observed in a corresponding sample); one needs a measure of the distance between x and y, in order to compare two methods.

Once a distance d between objects is selected, it is **intra-distance** or **inter-distance** if the objects are within the same cluster or in two different clusters, respectively.

The **linkage metric**, i.e., a distance between clusters $A = \{a_1, \dots, a_m\}$ and $B = \{b_1, \dots, b_n\}$ is usually one of the following:

average linkage: the average of the distances between the all members of the clusters, i.e., $\frac{\sum_i \sum_j d(a_i, b_j)}{mn}$;

single linkage: the distance $\min_{i,j} d(a_{i,j})$ between the nearest members of the clusters, i.e., the **set-set distance** (Chap. 1);

complete linkage: the distance $\max_{i,j} d(a_i, b_j)$ between the furthest members of the clusters, i.e., the **spanning distance** (Chap. 1);

centroid linkage: the distance between the *centroids* of the clusters, i.e, $||\tilde{a} - \tilde{b}||_2$, where $\tilde{a} = \frac{\sum_i a_i}{m}$, and $\tilde{b} = \frac{\sum_j b_j}{n}$;

Ward linkage: the distance $\sqrt{\frac{mn}{m+n}}||\tilde{a} - \tilde{b}||_2$.

Multidimensional Scaling is a technique developed in the behavioral and Social Sciences for studying the structure of objects or people. Together with Cluster Analysis, it is based on distance methods. But in Multidimensional Scaling, as opposed to Cluster Analysis, one starts only with some $m \times m$ matrix D of distances of the objects and (iteratively) looks for a representation of objects in $\mathbb{R}^n$ with low n, so that their Euclidean distance matrix has minimal square deviation from the original matrix D.

The related *Metric Nearness Problem* (Dhillon–Sra–Tropp, 2003) is to approximate a given finite distance space (X, d) by a metric space (X, d'). Other examples of distance methods in Data Analysis are *distance-based outlier detection* (in Data Mining) and *distance-based redundancy analysis* (in Multivariate Statistics).

There are many **similarities** used in Data Analysis; the choice depends on the nature of data and is not an exact science. We list below the main such similarities and distances.

Given two objects, represented by nonzero vectors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ from $\mathbb{R}^n$, the following notation is used in this chapter.

$\sum x_i$ means $\sum_{i=1}^{n} x_i$.

1_F is the *characteristic function* of event F: $1_F = 1$ if F happens, and $1_F = 0$, otherwise.

$||x||_2 = \sqrt{\sum x_i^2}$ is the ordinary Euclidean norm on $\mathbb{R}^n$.

$\overline{x}$ denotes $\frac{\sum x_i}{n}$, i.e., the *mean value* of components of x. So, $\overline{x} = \frac{1}{n}$ if x is a *frequency vector* (*discrete probability distribution*), i.e., all $x_i \geq 0$, and $\sum x_i = 1$; and $\overline{x} = \frac{n+1}{2}$ if x is a *ranking* (*permutation*), i.e., all x_i are different numbers from $\{1, \dots, n\}$.

The *k-th moment* is $\frac{\sum (x_i - \overline{x})^k}{n}$; it is called *variance*, *skewness*, *kurtosis* if $k = 2, 3, 4$.

[redacted]

In the binary case $x \in \{0,1\}^n$ (i.e., when x is a binary n-sequence), let $X = \{1 \leq i \leq n : x_i = 1\}$ and $\overline{X} = \{1 \leq i \leq n : x_i = 0\}$. Let $|X \cap Y|$, $|X \cup Y|$, $|X \backslash Y|$ and $|X \triangle Y|$ denote the cardinality of the intersection, union, difference and symmetric difference $(X \backslash Y) \cup (Y \backslash X)$ of the sets X and Y, respectively.

17.1 Similarities and Distances for Numerical Data

- **Ruzicka similarity**

  The **Ruzicka similarity** is a similarity on $\mathbb{R}^n$, defined by

$$\frac{\sum \min\{x_i, y_i\}}{\sum \max\{x_i, y_i\}}.$$

The corresponding **Soergel distance**

$$1 - \frac{\sum \min\{x_i, y_i\}}{\sum \max\{x_i, y_i\}} = \frac{\sum |x_i - y_i|}{\sum \max\{x_i, y_i\}}$$

coincides on $\mathbb{R}^n_{\geq 0}$ with the **fuzzy polynucleotide metric** (Chap. 23).

The **Wave–Edges distance** is defined by

$$\sum (1 - \frac{\min\{x_i, y_i\}}{\max\{x_i, y_i\}}) = \sum \frac{|x_i - y_i|}{\max\{x_i, y_i\}}.$$

- **Roberts similarity**

  The **Roberts similarity** is a similarity on $\mathbb{R}^n$, defined by

$$\frac{\sum (x_i + y_i) \frac{\min\{x_i, y_i\}}{\max\{x_i, y_i\}}}{\sum (x_i + y_i)}.$$

- **Ellenberg similarity**

  The **Ellenberg similarity** is a similarity on $\mathbb{R}^n$ defined by

$$\frac{\sum (x_i + y_i) 1_{x_i \cdot y_i \neq 0}}{\sum (x_i + y_i)(1 + 1_{x_i y_i = 0})}.$$

- **Gleason similarity**

  The **Gleason similarity** is a similarity on $\mathbb{R}^n$, defined by

$$\frac{\sum (x_i + y_i) 1_{x_i \cdot y_i \neq 0}}{\sum (x_i + y_i)}.$$

The **Czekanowsky–Dice distance** (or *nonmetric coefficient*, Bray–Curtis, 1957) is a **near-metric** on $\{0,1\}^n$ defined by

$$1-\frac{2|X\cap Y|}{|X|+|Y|}=\frac{|X\Delta Y|}{|X|+|Y|}.$$

- **Warrens inequalities on similarities**

  Denote by S_k, $1\le k\le 7$ above **intersection**, **Kulczynski 2**, **Bray–Curtis**, **Roberts**, **Ruzicka**, **Eilenberg** and **Gleason similarities**, respectively.

  Warrens (2016), showed that

$$1\ge S_1\ge S_2\ge S_3\ge S_4\ge S_5\ge 0 \quad \text{and} \quad S_5\le S_6\le S_7\ge S_3.$$

- **Jaccard similarity**

  The **Jaccard similarity of community**, Jaccard, 1908, is a similarity on $\mathbb{R}^n$ defined by

$$\frac{\sum x_iy_i}{\sum x_i^2+\sum y_i^2-\sum x_iy_i}.$$

The corresponding **Jaccard distance** is defined by

$$1-\frac{\sum x_iy_i}{\sum x_i^2+\sum y_i^2-\sum x_iy_i}=\frac{(\sum x_i-y_i)^2}{\sum x_i^2+\sum y_i^2-\sum x_iy_i}.$$

The binary cases of Jaccard, **Ellenberg** and **Ruzicka similarities** coincide; it is called **Tanimoto similarity**:

$$\frac{|X\cap Y|}{|X\cup Y|}.$$

The **Tanimoto distance** (or **biotope distance** from Chap. 23) is a distance on $\{0,1\}^n$ defined by

$$1-\frac{|X\cap Y|}{|X\cup Y|}=\frac{|X\Delta Y|}{|X\cup Y|}.$$

- **Czekanowsky similarity**

  The **Czekanowsky similarity** is a similarity on $\mathbb{R}^n$, defined by

$$\frac{\sum \min\{x_i,y_i\}}{\sum(x_i+y_i)}.$$

The corresponding **Czekanowsky distance** is defined by

$$1-\frac{\sum \min\{x_i,y_i\}}{\sum(x_i+y_i)}=\frac{\sum |x_i-y_i|}{\sum(x_i+y_i)}.$$

[redacted]

- **Dice similarity**

  The **Dice similarity** is a similarity on $\mathbb{R}^n$, defined by

$$\frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2}.$$

  The corresponding **Dice distance** is defined by

$$1 - \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2} = \frac{\sum |x_i - y_i|}{\sum x_i^2 + \sum y_i^2}.$$

- **Maryland Bridge similarity**

  The **Maryland Bridge similarity** is (Mirkin and Koonin, 2003) a similarity on $\mathbb{R}^n$ defined by

$$\frac{1}{2}\left(\frac{\sum x_i y_i}{\sum x_i^2} + \frac{\sum x_i y_i}{\sum y_i^2}\right).$$

  The corresponding **Maryland Bridge distance** is defined by

$$1 - \frac{1}{2}\left(\frac{\sum x_i y_i}{\sum x_i^2} + \frac{\sum x_i y_i}{\sum y_i^2}\right).$$

- **Simpson similarity**

  The **Simpson** (or *overlap*) **similarity** is a similarity on $\mathbb{R}^n$ defined by

$$\frac{\sum x_i y_i}{\min\{\sum x_i, \sum y_i\}}.$$

- **Intersection distance**

  The **intersection distance** is a distance on $\mathbb{R}^n$, defined by

$$1 - \frac{\sum \min\{x_i, y_i\}}{\min\{\sum x_i, \sum y_i\}}.$$

  It becomes $\frac{1}{2} \sum |x_i - y_i|$ if x, y are *frequency vectors.*
- **Kulczynski similarity** 1

  The **Kulczynski similarity** 1 is a similarity on $\mathbb{R}^n$ defined by

$$\frac{\sum \min\{x_i, y_i\}}{\sum |x_i - y_i|}.$$

  The corresponding **Kulczynski distance** is

$$\frac{\sum |x_i - y_i|}{\sum \min\{x_i, y_i\}}.$$

- **Kulczynski similarity** 2

  The **Kulczynski similarity** 2 is a similarity on $\mathbb{R}^n$ defined by

$$\frac{n}{2}\left(\frac{1}{\overline{x}}+\frac{1}{\overline{y}}\right)\sum \min\{x_i, y_i\}.$$

  In the binary case it coincides with **Maryland bridge similarity**; its form is

$$\frac{|X\cap Y|\cdot(|X|+|Y|)}{2|X|\cdot|Y|}=\frac{|X\cap Y|}{2|X|}+\frac{|X\cap Y|}{2|Y|}.$$

- **Motyka similarity**

  The **Motyka similarity** is a similarity on $\mathbb{R}^n$, defined by

$$\frac{\sum \min\{x_i, y_i\}}{\sum(x_i+y_i)}=n\frac{\sum \min\{x_i, y_i\}}{\overline{x}+\overline{y}}.$$

  The corresponding **Motyka distance** is

$$1-\frac{\sum \min\{x_i, y_i\}}{\sum(x_i+y_i)}=\frac{\sum \max\{x_i, y_i\}}{\sum(x_i+y_i)}.$$

- **Bray–Curtis similarity**

  The **Bray–Curtis similarity**, 1957, is a similarity on $\mathbb{R}^n$ defined by

$$\frac{2}{n(\overline{x}+\overline{y})}\sum \min\{x_i, y_j\}.$$

  It is called *Renkonen percentage similarity* if x, y are *frequency vectors*.

- **Sørensen distance**

  The **Sørensen** (or *Bray–Curtis*) **distance** on $\mathbb{R}^n$ is defined (Sørensen, 1948) by

$$1-\frac{2}{n(\overline{x}+\overline{y})}\sum \min\{x_i, y_j\}=\frac{\sum |x_i-y_i|}{\sum(x_i+y_i)}.$$

  The binary cases of **Bray–Curtis**, **Cleason**, **Czekanowsky** and **Dice similarities** coincide; it is called **Sørensen similarity**:

$$\frac{2|X\cap Y|}{|X\cup Y|+|X\cap Y|}=\frac{2|X\cap Y|}{|X|+|Y|}.$$

- **Canberra distance**

  The **Canberra distance** (Lance–Williams, 1967) is a distance on $\mathbb{R}^n$, defined by

$$\sum \frac{|x_i-y_i|}{|x_i|+|y_i|}.$$

[redacted]

- **Baroni–Urbani–Buser similarity**

  The **Baroni–Urbani–Buser similarity** is a similarity on $\mathbb{R}^n$ defined by

$$\frac{\sum \min\{x_i, y_i\} + \sqrt{\sum \min\{x_i, y_i\} \sum(\max_{1 \le j \le n} x_j - \max\{x_i, y_i\})}}{\sum \max\{x_i, y_i\} + \sqrt{\sum \min\{x_i, y_i\} \sum(\max_{1 \le j \le n} x_j - \max\{x_i, y_i\})}}.$$

  In the binary case it takes the form

$$\frac{|X \cap Y| + \sqrt{|X \cap Y| \cdot |\overline{X \cup Y}|}}{|X \cup Y| + \sqrt{|X \cap Y| \cdot |\overline{X \cup Y}|}}.$$

17.2 Relatives of Euclidean Distance

- **Power (p, r)-distance**

  The **power (p, r)-distance** is a distance on $\mathbb{R}^n$ defined, for $x, y \in \mathbb{R}^n$, by

$$(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{r}}.$$

  For $p = r \ge 1$, it is the l_p**-metric**, including the **Euclidean**, **Manhattan** (or *magnitude*) and **Chebyshev** (or *maximum-value*, *dominance*, *template*) **metrics** for $p = 2, 1$ and ∞, respectively.

  The case $(p, r) = (2, 1)$ corresponds to the **squared Euclidean distance**.

  The power (p, r)-distance with $0 < p = r < 1$ is called the **fractional** l_p**-distance** (not a metric since the unit balls are not convex). It is used for "dimensionality-cursed" data, i.e., when there are few observations and the number n of variables is large. The case $0 < p < r = 1$, i.e., of the p-th power of the fractional l_p-distance, corresponds to a metric on $\mathbb{R}^n$.

  The weighted versions $(\sum w_i |x_i - y_i|^p)^{\frac{1}{p}}$ (with nonnegative weights w_i) are also used, for $p = 1, 2$, in applications. Given weights $w_i \ge 0$, the **weighted Manhattan quasi-metric** for $x, y \in \mathbb{R}^n$ is $\sum_{i=1}^{n} d_i$, where every d_i is the quasi-metric defined by $d_i = w_i(x_i - y_i)$ if $x_i > y_i$ and $d_i = W_i(y_i - x_i)$, otherwise.

  The **ordinal distance** on $\mathbb{R}^n$ is defined (Bahari and Van hamme, 2014) by

$$(\sum_{i=1}^{n} | \sum_{1 \le j \le i} (x_j - y_j)|^p)^{\frac{1}{p}}.$$

- **YJHHR metrics**

  We call **YJHHR metrics** the following metrics, introduced by Yang, Jiang, Hahn, Housworth, and Radivojac, 2016.

For any $p \geq 1$ and two finite sets X, Y, define

$$d(X, Y) = (|X \setminus Y|^p + |Y \setminus X|^p)^{\frac{1}{p}}$$

$$\text{and } \; d'(X, Y) = \frac{d(X, Y)}{|X \cup Y|} \; \text{ if } \; |X \cup Y| > 0, \; d'(X, Y) = 0, \; \text{otherwise.}$$

For any $p \geq 1$ and two bounded integrable functions f, g on $\mathbb{R}$, define

$$D(f, g) = \left(\left(\int (\max(f - g, 0)dx\right)^p + \left(\int (\max(g - f, 0))x\right)^p\right)^{\frac{1}{p}}$$

$$\text{and } \; D'(f, g) = \frac{D(f, g)}{\int \max(|f|, |f|, |f - g|)dx}.$$

- **Multiplicative distance**

  In order to offset instability of the norm distances for high-dimensional data (i.e., in $\mathbb{R}^n$ with large n), Mansouri and Khademi, 2014, introduced the following **multiplicative distance** for any $x, y \in \mathbb{R}^n$:

  $$d_{MK}(x, y) = -1 + \prod_{i=1}^{n} (1 + |x_i - y_i|)^{c_i},$$

  where $c_1, \dots, c_n$ are given positive numbers.

  Qureshi, 2015, introduced another *multiplicative distance* for any $x, y \in \mathbb{N}^n$:

  $$d_Q(x, y) = \sum_{i=1}^{n} ||l(x_i) - l(y_i)||_1,$$

  where $l(a) = (\alpha_1, \alpha_2, \dots)$ for any number $a \in \mathbb{N}$ represented as $p_1^{\alpha_1} p_2^{\alpha_2} \dots$; here $p_1, p_2, \dots$ is the sequence of prime numbers.
- **Penrose size and shape distances**

  The **Penrose size distance** and **Penrose shape distance** are the distances on $\mathbb{R}^n$ defined, respectively, by

  $$\sqrt{n} \sum |x_i - y_i| \; \text{ and } \; \sqrt{\sum ((x_i - \overline{x}) - (y_i - \overline{y}))^2}.$$

  The sum of their squares is the **squared Euclidean distance**.

  The **mean character distance** (Czekanowsky, 1909) is defined by $\frac{\sum |x_i - y_i|}{n}$.
- **Lorentzian distance**

  The **Lorentzian distance** is a distance on $\mathbb{R}^n$, defined by

  $$\sum \ln(1 + |x_i - y_i|).$$

[redacted]

- **Effect size**

  Let $\overline{x}, \overline{y}$ be the means of samples x, y and let s^2 be the pooled variance of both samples. The **effect size** (a term used mainly in social sciences) is defined by

  $$\frac{\overline{x}-\overline{y}}{s}.$$

  Its symmetric version $\frac{|\overline{x}-\overline{y}|}{s}$ is called *statistical distance* by Johnson–Wichern, 1982, and *standard distance* by Flury–Riedwyl, 1986.

  Cf. the **engineer semimetric** in Chap. 14 and the **ward linkage**.

- **Binary Euclidean distance**

  The **binary Euclidean distance** is a distance on $\mathbb{R}^n$ defined by

  $$\sqrt{\sum (1_{x_i>0} - 1_{y_i>0})^2}.$$

- **Mean censored Euclidean distance**

  The **mean censored Euclidean distance** is a distance on $\mathbb{R}^n$ defined by

  $$\sqrt{\frac{\sum (x_i - y_i)^2}{\sum 1_{x_i^2+y_i^2 \neq 0}}}.$$

- **Normalized l_p-distance**

  The **normalized l_p-distance**, $1 \leq p \leq \infty$, is a distance on $\mathbb{R}^n$ defined by

  $$\frac{||x-y||_p}{||x||_p + ||y||_p}.$$

  The only integer value p for which the normalized l_p-distance is a metric, is p=2. Moreover, the distance $\frac{||x-y||_2}{a+b(||x||_2+||y||_2)}$ is a metric for any $a, b > 0$ ([Yian91]).

- **Clark distance**

  The **Clark distance** (Clark, 1952) is a distance on $\mathbb{R}^n$, defined by

  $$\left(\frac{1}{n} \sum \left(\frac{x_i - y_i}{|x_i| + |y_i|} \right)^2 \right)^{\frac{1}{2}}.$$

- **Meehl distance**

  The **Meehl distance** (or *Meehl index*) is a distance on $\mathbb{R}^n$ defined by

  $$\sum_{1 \leq i \leq n-1} (x_i - y_i - x_{i+1} + y_{i+1})^2.$$

- **Hellinger distance**

  The **Hellinger distance** is a distance on $\mathbb{R}^n_+$ defined by

$$\sqrt{2\sum\left(\sqrt{\frac{x_i}{\overline{x}}}-\sqrt{\frac{y_i}{\overline{y}}}\right)^2}.$$

  Cf. **Hellinger metric** in Chap. 14.

  The *Whittaker index of association* is defined by $\frac{1}{2}\sum|\frac{x_i}{\overline{x}}-\frac{y_i}{\overline{y}}|$.

- **Symmetric χ^2-measure**

  The **symmetric χ^2-measure** is a distance on $\mathbb{R}^n$ defined by

$$\sum\frac{2}{\overline{x}\cdot\overline{y}}\cdot\frac{(x_i\overline{y}-y_i\overline{x})^2}{x_i+y_i}.$$

- **Symmetric χ^2-distance**

  The **symmetric χ^2-distance** (or *chi-distance*) is a distance on $\mathbb{R}^n$ defined by

$$\sqrt{\sum\frac{\overline{x}+\overline{y}}{n(x_i+y_i)}(\frac{x_i}{\overline{x}}-\frac{y_i}{\overline{y}})^2}=\sqrt{\sum\frac{\overline{x}+\overline{y}}{n(\overline{x}\cdot\overline{y})^2}\cdot\frac{(x_i\overline{y}-y_i\overline{x})^2}{x_i+y_i}}.$$

  It is a **weighted Euclidean distance**.

- **Weighted Euclidean distance**

  The general **quadratic-form distance** on $\mathbb{R}^n$ is defined by

$$\sqrt{(x-y)^TA(x-y)},$$

  where A is a real nonsingular symmetric $n\times n$ matrix; cf. **Mahalanobis distance**.

  The **weighted Euclidean distance** is the case $A=\operatorname{diag}(a_i), a_i\neq 0$, i.e., it is

$$\sqrt{\sum a_i(x_i-y_i)^2}.$$

  Some examples are: **pseudo-Euclidean distance** (Chap. 7), **standardized Euclidean distance** and first two metrics (Euclidean $\mathbb{R}^6$-distances) in Sect. 18.3.

- **Mahalanobis distance**

  The **Mahalanobis distance** (or **quadratic distance**, or *directionally weighted distance*) is a semimetric on $\mathbb{R}^n$ defined (Mahalanobis, 1936) by

$$||x-y||_A=\sqrt{(x-y)A(x-y)^T},$$

  where A is a positive-semidefinite matrix. It is a metric if A is positive-definite. Cf. **Mahalanobis semimetric** in Chap. 14. The square $||x-y||^2_A$ is called *generalized ellipsoid* (or *generalized squared interpoint*) *distance*.

[redacted]

Usually, $A = C^{-1}$, where C is a *covariance matrix* $((Cov(x_i, x_j)))$ of some data points $x, y \in \mathbb{R}^n$ (say, random vectors with the same distribution), or $A = (det(C))^{\frac{1}{n}} C^{-1}$ so that $det(A) = 1$.

Clearly, $||x-y||_I$ is the Euclidean distance. If $C = ((c_{ij}))$ is a diagonal matrix, then $c_{ii} = Var(x_i) = Var(y_i) = \sigma_i^2$ and it holds

$$||x-y||_{C^{-1}} = \sqrt{\sum_i \frac{(x_i - y_i)^2}{\sigma_i^2}}.$$

Such diagonal Mahalanobis distance is called the **standardized Euclidean distance** (or **normalized Euclidean distance**, *scaled Euclidean distance*).

The **maximum scaled difference** (Maxwell–Buddemeier, 2002) is defined by

$$\max_i \frac{(x_i - y_i)^2}{\sigma_i^2}.$$

17.3 Similarities and Distances for Binary Data

Usually, such similarities s range from 0 to 1 or from -1 to 1; the corresponding distances are usually $1 - s$ or $\frac{1-s}{2}$, respectively.

- **Hamann similarity**

  The **Hamann similarity**, 1961, is a similarity on $\{0, 1\}^n$, defined by

$$\frac{2|\overline{X \Delta Y}|}{n} - 1 = \frac{n - 2|X \Delta Y|}{n}.$$

- **Rand similarity**

  The **Rand similarity** (or Sokal–Michener's *simple matching*) is a similarity on $\{0, 1\}^n$ defined by

$$\frac{|\overline{X \Delta Y}|}{n} = 1 - \frac{|X \Delta Y|}{n}.$$

  Its square root is called the *Euclidean similarity*. The corresponding metric $\frac{|X \Delta Y|}{n}$ is called the *variance* or *Manhattan similarity*; cf. **Penrose size distance**.

- **Sokal–Sneath similarities**

  The **Sokal-Sneath similarities** $1, 2, 3$ are the similarity on $\{0, 1\}^n$ defined by

$$\frac{2|\overline{X \Delta Y}|}{n + |\overline{X \Delta Y}|}, \quad \frac{|X \cap Y|}{|X \cup Y| + |X \Delta Y|}, \quad \frac{|\overline{X \Delta Y}|}{|\overline{X \Delta Y}|}.$$

- **Russel–Rao similarity**
  The **Russel–Rao similarity** is a similarity on $\{0,1\}^n$, defined by
$$\frac{|X \cap Y|}{n}.$$
- **Forbes–Mozley similarity**
  The **Forbes–Mozley similarity** is a similarity on $\{0,1\}^n$ defined by
$$\frac{n|X \cap Y|}{|X||Y|}.$$
- **Braun–Blanquet similarity**
  The **Braun–Blanquet similarity** is a similarity on $\{0,1\}^n$ defined by
$$\frac{|X \cap Y|}{\max\{|X|,|Y|\}}.$$
- **Roger–Tanimoto similarity**
  The **Roger–Tanimoto similarity**, 1960, is a similarity on $\{0,1\}^n$ defined by
$$\frac{|\overline{X \Delta Y}|}{n + |X \Delta Y|}.$$
- **Faith similarity**
  The **Faith similarity** is a similarity on $\{0,1\}^n$, defined by
$$\frac{|X \cap Y| + |\overline{X \Delta Y}|}{2n}.$$
- **Tversky similarity**
  The **Tversky similarity** is a similarity on $\{0,1\}^n$, defined by
$$\frac{|X \cap Y|}{a|X \Delta Y| + b|X \cap Y|}.$$
  It becomes the **Tanimoto**, **Sørensen** and (the binary case of) **Kulczynsky** 1 **similarities** for $(a,b) = (1,1)$, $(\frac{1}{2},1)$ and $(1,0)$, respectively.
- **Mountford similarity**
  The **Mountford similarity**, 1962, is a similarity on $\{0,1\}^n$, defined by
$$\frac{2|X \cap Y|}{|X||Y \backslash X| + |Y||X \backslash Y|}.$$

[redacted]

- **Gower–Legendre similarity**
  The **Gower–Legendre similarity** is a similarity on $\{0,1\}^n$ defined by

$$\frac{|\overline{X \Delta Y}|}{a|X \Delta Y| + |\overline{X \Delta Y}|} = \frac{|\overline{X \Delta Y}|}{n + (a-1)|X \Delta Y|}.$$

- **Anderberg similarity**
  The **Anderberg** (or *Sokal–Sneath* 4 **similarity**) on $\{0,1\}^n$ is defined by

$$\frac{|X \cap Y|}{4}\left(\frac{1}{|X|} + \frac{1}{|Y|}\right) + \frac{|\overline{X \cup Y}|}{4}\left(\frac{1}{|\overline{X}|} + \frac{1}{|\overline{Y}|}\right).$$

- **Yule similarities**
  The **Yule** Q **similarity** (Yule, 1900) is a similarity on $\{0,1\}^n$, defined by

$$\frac{|X \cap Y| \cdot |\overline{X \cup Y}| - |X \backslash Y| \cdot |Y \backslash X|}{|X \cap Y| \cdot |\overline{X \cup Y}| + |X \backslash Y| \cdot |Y \backslash X|}.$$

  The **Yule** Y **similarity of colligation** (1912) is a similarity on $\{0,1\}^n$ defined by

$$\frac{\sqrt{|X \cap Y| \cdot |\overline{X \cup Y}|} - \sqrt{|X \backslash Y| \cdot |Y \backslash X|}}{\sqrt{|X \cap Y| \cdot |\overline{X \cup Y}|} + \sqrt{|X \backslash Y| \cdot |Y \backslash X|}}.$$

- **Dispersion similarity**
  The **dispersion similarity** is a similarity on $\{0,1\}^n$, defined by

$$\frac{|X \cap Y| \cdot |\overline{X \cup Y}| - |X \backslash Y| \cdot |Y \backslash X|}{n^2}.$$

- **Pearson ϕ similarity**
  The **Pearson ϕ similarity** is a similarity on $\{0,1\}^n$ defined by

$$\frac{|X \cap Y| \cdot |\overline{X \cup Y}| - |X \backslash Y| \cdot |Y \backslash X|}{\sqrt{|X| \cdot |\overline{X}| \cdot |Y| \cdot |\overline{Y}|}}.$$

- **Gower similarity** 2
  The **Gower** 2 (or *Sokal–Sneath* 5) **similarity** on $\{0,1\}^n$ is defined by

$$\frac{|X \cap Y| \cdot |\overline{X \cup Y}|}{\sqrt{|X| \cdot |\overline{X}| \cdot |Y| \cdot |\overline{Y}|}}.$$

- **Pattern difference**

  The **pattern difference** is a distance on $\{0,1\}^n$, defined by

$$\frac{4|X \backslash Y| \cdot |Y \backslash X|}{n^2}.$$

- **Q_0-difference**

  The **Q_0-difference** is a distance on $\{0,1\}^n$, defined by

$$\frac{|X \backslash Y| \cdot |Y \backslash X|}{|X \cap Y| \cdot |\overline{X \cup Y}|}.$$

- **Model distance**

  Let X, Y be two data sets, and let λ_j be the eigenvalues of the symmetrized cross-correlation matrix $C_{X \backslash Y\, Y \backslash X} \times C_{Y \backslash X\, X \backslash Y}$.

  The **model distance** (Todeschini, 2004) is a distance on $\{0,1\}^n$ defined by

$$\sqrt{|X \backslash Y| + |Y \backslash X| - 2\sum_j \sqrt{\lambda_j}}.$$

  The **CMD-distance** (or, *canonical measure of distance*, Todeschini et al., 2009) is

$$\sqrt{|X| + |Y| - 2\sum_j \sqrt{\lambda_j}},$$

  where λ_j are the nonzero eigenvalues of the cross-correlation matrix $C_{XY} \times C_{YX}$.

17.4 Correlation Similarities and Distances

The *covariance* between two real-valued random variables X and Y is $Cov(x,y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. The *variance* of X is $Var(X) = Cov(X,X)$ and the **Pearson correlation** of X and Y is $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$; cf. Chap. 14.

Let $(X,Y), (X',Y'), (X'',Y'')$ be independent and identically distributed. The *distance covariance* (Székely, 2005) is the square root of $dCov^2(X,Y) = \mathbb{E}[|X-X'||Y-Y'|] + \mathbb{E}[|X-X'|]\mathbb{E}[|Y-Y'|] - \mathbb{E}[|X-X'||Y-Y''|] - \mathbb{E}[|X-X''||Y-Y'|] = \mathbb{E}[|X-X'||Y-Y'|] + \mathbb{E}[|X-X'|]\mathbb{E}[|Y-Y'|] - 2\mathbb{E}[|X-X'||Y-Y''|]$. It is 0 if and only if X and Y are independent. The **distance correlation** $dCor(X,Y)$ is $\frac{dCov(X,Y)}{\sqrt{dCov(X,X)dCov(Y,Y)}}$.

The vectors x, y below can be seen as *samples* (series of n measurements) of X, Y.

[redacted]

- **Covariance similarity**

  The **covariance similarity** is a similarity on $\mathbb{R}^n$ defined by

$$\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{n}=\frac{\sum x_i y_i}{n}-\overline{x}\cdot\overline{y}.$$

- **Pearson correlation similarity**

  The **Pearson correlation similarity**, or, by its full name, *Pearson product-moment correlation coefficient*) is a similarity on $\mathbb{R}^n$ defined by

$$s=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_j-\overline{x})^2\sum(y_j-\overline{y})^2}}.$$

  The **Pearson distance** (or **correlation distance**) is defined by

$$1-s=\frac{1}{2}\sum\left(\frac{x_i-\overline{x}}{\sqrt{\sum(x_j-\overline{x})^2}}-\frac{y_i-\overline{y}}{\sqrt{\sum(y_j-\overline{y})^2}}\right)$$

  A multivariate generalization of the Pearson correlation similarity is the *RV coefficient* (Escoufier, 1973) $RV(X,Y)=\frac{Covv(X,Y)}{\sqrt{Covv(X,X)Covv(Y,Y)}}$, where X,Y are matrices of centered random (column) vectors with covariance matrix $C(X,Y)=\mathbb{E}[X^TY]$, and $Covv(X,Y)$ is the trace of the matrix $C(X,Y)C(Y,X)$.
- **Cosine similarity**

  The **cosine similarity** (or *Orchini similarity*, *angular similarity*, *normalized dot product*) is the case $\overline{x}=\overline{y}=0$ of the **Pearson correlation similarity**, i.e., it is

$$\frac{\langle x,y\rangle}{||x||_2\cdot||y||_2}=\cos\phi,$$

  where ϕ is the angle between vectors x and y. In the binary case, it becomes

$$\frac{|X\cap Y|}{\sqrt{|X|\cdot|Y|}}$$

  and is called the **Ochiai–Otsuka similarity**.

  In Record Linkage, cosine similarity is called **TF-IDF** similarity; it (or *tf-idf*, *TFIDF*) are used as an abbreviation of *Frequency-Inverse Document Frequency*.

  The **angular semimetric** on $\mathbb{R}^n$ is defined by $\arccos\phi$. The **cosine distance** is $1-\cos\phi$, and the **Orloci distance** (or *chord distance*) is

$$\sqrt{2(1-\cos\phi)}=\sqrt{\sum(\frac{x_i}{||x||_2}-\frac{y_i}{||y||_2})^2}.$$

- **Similarity ratio**

  The **similarity ratio** (or *Kohonen similarity*, *Kumar–Hassebrook similarity*) is a similarity on $\mathbb{R}^n$ defined by

  $$\frac{\langle x, y\rangle}{\langle x, y\rangle + ||x-y||_2^2}.$$

  Its binary case is the **Tanimoto similarity**.
- **Morisita–Horn similarity**

  The **Morisita–Horn similarity** (Morisita, 1959) is a similarity on $\mathbb{R}^n$ defined by

  $$\frac{2\langle x, y\rangle}{||x||_2^2 \cdot \frac{\overline{y}}{\overline{x}} + ||y||_2^2 \cdot \frac{\overline{x}}{\overline{y}}}.$$

- **Spearman rank correlation**

  If the sequences $x, y \in \mathbb{R}^n$ are ranked separately, then the **Pearson correlation similarity** is approximated by the following **Spearman ρ rank correlation**:

  $$\frac{\sum(a_i-\overline{a})(b_i-\overline{b})}{\sqrt{\sum(a_j-\overline{a})^2\sum(b_j-\overline{b})^2}} = 1 - \frac{6}{n(n^2-1)}\sum(a_i-b_i)^2,$$

  where $n > 1$ and $a_i = \text{rank}(x_i), b_i = \text{rank}(y_i), a = (a_1, \dots, a_n), b = (b_1, \dots, b_n)$. This approximation is good for such *ordinal* data when it holds $\overline{x} = \overline{y} = \frac{n+1}{2}$.

  The **Spearman footrule** is defined by

  $$1 - \frac{3}{n^2-1}\sum|x_i - y_i|.$$

  Cf. the **Spearman ρ distance** and **Spearman footrule distance** in Chap. 11.

  Another correlation similarity for rankings is the **Kendall τ rank correlation**:

  $$\frac{2\sum_{1\le i<j\le n} sign(x_i-x_j)sign(y_i-y_j)}{n(n-1)}.$$

  Cf. the **Kendall τ distance** on permutations in Chap. 11.
- **Global correlation distance**

  Let $x \in \mathbb{R}^n$ and (A, d) be a metric space with n points $a_1, \dots, a_n$. For any $d > 0$, the *Moran autocorrelation coefficient* is defined by

  $$I(d) = \frac{n\sum_{1\le i\ne j\le n} w_{ij}(d)(x_i-\overline{x})(x_j-\overline{x})}{\sum_{1\le i\ne j\le n} w_{ij}(d)\sum_{1\le i\le n}(x_i-\overline{x})^2},$$

[redacted]

where the weight $w_{ij}(d)$ is 1 if $d(a_i, a_j) = d$ and 0, otherwise. In **spatial analysis**, eventual clustering of (A, d) implies that $I(d)$ decreases with increasing d. $I(d)$ is a global indicator of the presumed **spatial dependence** that evaluate the existence/size of clusters in the spatial arrangement (A, d) of a given variable.

The **global correlation distance** is the least value d' for which $I(d) = 0$.

- **Log-likelihood distance**

  Given two clusters A and B, their **log-likelihood distance** is the decrease in *log-likelihood* (cf. the **Kullback–Leibler distance** in Chap. 14 and the **log-likelihood ratio quasi-distance** in Chap. 21) as they are combined into one cluster. Simplifying (taking $A, B \subset \mathbb{R}_{>0}$), it is defined by

$$\sum_{x \in A} x \log \frac{x}{|A|} + \sum_{x \in B} x \log \frac{x}{|B|} - \sum_{x \in A \cup B} x \log \frac{x}{|A \cup B|}.$$

- **Spatial analysis**

  In Statistics, **spatial analysis** (or *spatial statistics*) includes the formal techniques for studying entities using their topological, geometric, or geographic properties. More restrictively, it refers to *Geostatistics* and *Human Geography*. It considers spatially distributed data as *a priori* dependent one on another.

  Spatial dependence is a measure for the degree of associative dependence between independently measured values in an ordered set, determined in samples selected at different positions in a sample space. Cf. **spatial correlation** in Chap. 24. An example of such space-time dynamics: Gould, 1997, showed that $\approx$ 80 % of the diffusion of HIV in the US is highly correlated with the air passenger traffic (origin-destination) matrix for 102 major urban centers.

  SADIE (Spatial Analysis by Distance IndicEs) is a methodology (Perry, 1998) to measure the degree of nonrandomness in 2D spatial patterns of populations. Given n sample units $x_i \in \mathbb{R}^2$ with associated counts N_i, the **distance to regularity** is the minimal total Euclidean distance that the individuals in the sample would have to move, from unit to unit, so that all units contained an identical number of individuals. The **distance to crowding** is the minimal total distance that individuals in the sample must move so that all are congregated in one unit. The *indices of aggregation* are defined by dividing above distances by their mean values. Cf. **Earth Mover's distance** in Chap. 21.

- **Distance sampling**

  Distance sampling is a widely-used group of methods for estimating the density and abundance of biological populations. It is an extension of plot- (or quadrate-based) sampling, where the number of objects at given distance from a point or a segment is counted. Also, *Distance* is the name of a Windows-based computer package that allows to design and analyze distance sampling surveys.

  A standardized survey along a series of lines or points is performed, searching for objects of interest (say, animals, plants or their clusters). *Detection distances* r (perpendicular ones from given lines and radial ones from given points) are measured to each detected object. The *detection function* $g(r)$ (the probability that an object at distance r is detected) is fit then to the observed distances, and this fitted function is used to estimate the proportion of objects missed by the survey. It gives estimates for the density and abundance of objects in the survey area.

- **Cook distance**

  The **Cook distance** is a distance on $\mathbb{R}^n$ giving a statistical measure of deciding if some i-th observation alone affects much regression estimates. It is a normalized **squared Euclidean distance** between estimated parameters from regression models constructed from all data and from data without i-th observation.

  The main similar distances, used in Regression Analysis for detecting influential observations, are *DFITS distance*, *Welsch distance*, and *Hadi distance*.

- **Periodicity p-self-distance**

  Ergun–Muthukrishnan–Sahinalp, 2004, call a data stream $x = (x_1, \dots, x_n)$ *p-periodic approximatively*, for given $1 \leq p \leq \frac{n}{2}$ and distance function d between p-blocks of x, if the **periodicity p-self-distance** $\sum_{i \neq j} d((x_{jp+1}, \dots, x_{jp+p}), (x_{ip+1}, \dots, x_{ip+p}))$ is below some threshold.

  Above notion of self-distance is different from ones given in Chaps. 1 and 28. Also, the term *self-distance* is used for *round-off error* (or *rounding error*), i.e., the difference between the calculated approximation of a number and its exact value.

- **Distance metric learning**

  Let $x_1, \dots, x_n$ denote the samples in the training set $X \subset \mathbb{R}^m$; here m is the number of features. **Distance metric learning** is an approach for the problem of clustering with side information, when algorithm learns a distance function d prior to clustering and then tries to satisfy some *positive* (or *equivalence*) constraints P and *negative* constraints D. Here S and D are the sets of *similar* (belonging to the same class) and *dissimilar* pairs (x_i, x_j), respectively.

  Usually d is a **Mahalanobis metric** $||x_i - x_j||_A = \sqrt{(x_i - x_j)^T A (x_i - x_j)}$, where A is a positive-semidefinite matrix, i.e., $A = W^T W$ for a matrix W with m columns and $||x_i - x_j||_A^2 = ||Wx_i - Wx_j||^2$. Then, for example, one look for (Xing et al., 2003) A minimizing $\sum_{(x_i,x_j) \in S} ||x_i - x_j||_A^2$ while $\sum_{(x_i,x_j) \in D} ||x_i - x_j||_A^2 \geq 1$.

- **Heterogeneous distance**

  The following IBL (instance-based learning) setting is used for many real-world applications (neural networks, etc.), where data are incomplete and have both continuous and nominal attributes. Given an $m \times (n+1)$ matrix $((x_{ij}))$, its row $(x_{i0}, x_{i1}, \dots, x_{in})$ denotes an *instance input vector* $x_i = (x_{i1}, \dots, x_{in})$ with output class x_{i0}; the set of m instances represents a training set during learning. For any new input vector $y = (y_1, \dots, y_n)$, the closest (in terms of a selected distance d) instance x_i is sought, in order to *classify* y, i.e., predict its output class as x_{i0}.

  A **heterogeneous distance** $d(x_i, y)$ is defined ([WiMa97]) by

$$\sqrt{\sum_{j=1}^{n} d_j^2(x_{ij}, y_j)}$$

[redacted]

with $d_j(x_{ij}, y_j) = 1$ if x_{ij} or y_j is unknown. If the *attribute* (input variable) j is nominal, then $d_j(x_{ij}, y_j)$ is defined, for example, as $1_{x_{ij} \neq y_j}$, or as

$$\sum_a \left| \frac{|\{1 \le t \le m : x_{t0} = a, x_{ij} = x_{ij}\}|}{|\{1 \le t \le m : x_{tj} = x_{ij}\}|} - \frac{|\{1 \le t \le m : x_{t0} = a, x_{ij} = y_j\}|}{|\{1 \le t \le m : x_{tj} = y_j\}|} \right|^q$$

for $q = 1$ or 2; the sum is taken over all output classes, i.e., values a from $\{x_{t0} : 1 \le t \le m\}$. For continuous attributes j, the number d_j is taken to be $|x_{ij} - y_j|$ divided by $\max_t x_{tj} - \min_t x_{tj}$, or by $\frac{1}{4}$ of the standard deviation of values x_{tj}, $1 \le t \le m$.

Chapter 18
Distances in Systems and Mathematical Engineering

In this chapter we group the main distances used in Systems Theory (such as *Transition Systems*, *Dynamical Systems*, *Cellular Automata*, *Feedback Systems*) and other interdisciplinary branches of Mathematics, Engineering and Theoretical Computer Science (such as, say, *Robot Motion* and *Multi-objective Optimization*).

A *labeled transition system* (LTS) is a triple (S, T, F) where S is a set of *states*, T is a set of *labels* (or *actions*) and $F \subseteq S \times T \times S$ is a ternary relation. Any $(x, t, y) \in F$ represents a *t-labeled transition* from state x to state y. A LTS with $|T| = 1$ corresponds to an *unlabeled transition system*.

A *path* is a sequence $((x_1, t_1, x_2), \dots, (x_i, t_i, x_{i+1}), \dots)$ of transitions; it gives rise to a *trace* $(t_1, \dots, t_i, \dots)$. Two paths are *trace-equivalent* if they have the same traces. The term *trace*, in Computer Science, refers in general to the equivalence classes of strings of a *trace monoid*, wherein certain letters in the string are allowed to commute. It is not related to the *trace* in Linear Algebra.

A LTS is called *deterministic* if for any $x \in S$ and $t \in T$ it holds that $|\{y \in S : (x, t, y) \in F\}| = 1$. Such LTS without output is called a *semiautomaton* (S, T, f) where S is a set of *states*, T is an *input alphabet* and $f : X \times T \to S$ is a *transition function*.

A *deterministic finite-state machine* is a tuple (S, s_0, T, f, S') with S, T, f as above but $0 < |S|, |T| < \infty$, while $s_o \in S$ is an *initial state*, and $S' \subset S$ is the set of *final states*.

The *free monoid* on a set T is a *monoid* (algebraic structure with an associative binary operation and an identity element) T^* whose elements are all the finite sequences $x = x_0, \dots, x_m$ of elements from T. The identity element is the empty string λ, and the operation is string concatenation. The free semigroup on T is $T^+ = T^* \setminus \{\lambda\}$. Let T^ω denote the set of all infinite sequences $x = (x_0, x_1, \dots)$ in T, and let T^∞ denote $T^* \cup T^\omega$.

A finite-state machine is *nondeterministic* if the next possible state is not uniquely determined. A *weighted automaton* is a such machine, say, M equipped

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_18

[redacted]

with a cost function $c \geq 0$, over some semiring $(S, \oplus, \otimes)$, on transitions. For a *probabilistic automaton*, the semiring is $(\mathbb{R}_{\geq 0}, +, \times)$ and $0 \leq c \leq 1$.

A **distance automaton** is (Hashiguchi, 1982) a weighted automaton over the *tropical semiring TROP* $= (\mathbb{N} \cup \{\infty\}, \min, +)$. A *run* over a word (string in the *language* of M) $(a_1, \dots, a_k)$ is a sequence $(s_0, \dots, s_k)$ of states. The *run's distance* is the sum $\sum_{i=1}^{k} c(a_i)_{p_{i-1}p_i}$ of costs of involved transitions. The run is *accepting* if s_0 is initial and s_k is a final state. The *distance of a word recognized by M* is the minimum of the distances over the all accepting runs. The *distance* of M is the supremum over the distances of all recognized words. Distance automata are equivalent to finitely generated monoids of matrices over *TROP*: nondeterministic automata recognize the same language as some deterministic ones but with transitions acting on the sets of original states.

18.1 Distances in State Transition and Dynamical Systems

- **Distances on formal languages**

  A *formal language* over an alphabet T is a set of words over T.

  The similarity between two words of a language are measured usually by the **Hamming metric** (cf. Sect. 1.5) or (cf. Sect. 11.1) by the **Levenstein metric** (or **edit distance**) and **prefix, suffix and substring distances**.

  Given a language L and a distance d on it, the **distance between a word** u **and** L is the **point-set distance** (Sect. 1.5) $\min_{v \in L} d(u, v)$.

  The **inner distance** (or *self distance*) d **of** L is (Mihri, 2007) $\min_{u,v \in L, u \neq v} d(u, v)$; cf. *separation* in the item **metric spread** in Sect. 1.3.

  The similarity between two languages L and L' are measured usually by the following extensions: $\min_{u \in L, v \in L', u \neq v} d(u, v)$ and (called by Choffrut–Pighizzini, 2002, the **relative distance d**) $\max_{u \in L} \min_{v \in L'} d(u, v)$. In terms of Sect. 1.5, they are, respectively, the **set-set distance** and the **directed Hausdorff distance**.

  See also in Sect. 28.3 the **language distance from English** and other notions of distance between words of a natural language and between languages.

- **Fahrenberg–Legay–Thrane distances**

  Given a *labeled transition system* (LTS) (S, T, F) Fahrenberg–Legay–Thrane, 2011, call T^{∞} the set of *traces* and define a *trace distance* as an extended **hemimetric** (or **quasi-semimetric**) $h : T^{\infty} \times T^{\infty} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that $h(x, y) = \infty$ for any sequences $x, y \in T^{\infty}$ of different length.

  For a given distance d on the set T of labels and a *discount factor* q $(0 < q \leq 1)$, they defined the *pointwise*, *accumulating* and *limit-average trace distance* as, respectively, $PW_{d,q}(x, y) = \sup_i q^i d(x_i, y_i)$, $ACC_{d,q}(x, y) = \sum_i q^i d(x_i, y_i)$ and $AVG_d = \underline{\lim}_{i \to \infty} \frac{1}{i+1} \sum_{j=0}^{i} d(x_j, y_j)$.

  If d is a **discrete metric**, i.e., $d(t, t') = 1$ whenever $t \neq t'$, then $ACC_{d,1}$ is the **Hamming metric** for finite traces of the same length, and $ACC_{d,q}$ with $q < 1$ and AVG_d are analogs of the Hamming metric for infinite traces.

Other examples of trace distances are a Cantor-like distance $(1 + \inf\{i : x_i \neq y_i\})^{-1}$ and the *maximum-lead distance*, defined, for $T \subseteq \Sigma \times \mathbb{R}$, by Henzinger–Majumdar–Prabhu, 2005, as $\sup_i |\sum_{j=0}^{i} x''_j - \sum_{j=0}^{i} y''_j|$ if $x'_i = y'_i$ for all i and ∞, otherwise. Here any $z \in T$ is denoted by (z', z''), where $z' \in \Sigma$ and $z'' \in \mathbb{R}$.

Fahrenberg–Legay–Thrane, 2011, also define the two following extended simulation hemimetrics between states $x, y \in S$.

The *accumulating simulation distance* $h_{ac}(x, y)$ and the *pointwise simulation distance* $h_{po}(x, y)$ are the least fixed points, respectively, to the set of equations

$$h_{ac}(x, y) = \max_{t \in T : (x,t,x') \in F} \ \min_{t' \in T : (y,t',y') \in F} (d(t, t') + q h_{ac}(x', y')) \text{ and}$$

$$h_{po}(x, y) = \max_{t \in T : (x,t,x') \in F} \ \min_{t' \in T : (y,t',y') \in F} \max\{d(t, t'), h_{po}(x', y')\}.$$

The above hemimetrics generalize the lifting by Alfaro–Faella–Stoelinga, 2004, of the quasi-metric $\max\{x'' - y'', 0\}$ between labels $x, y \in T = \Sigma \times \mathbb{R}$ on an accumulating trace distance and then the lifting of it on the **directed Hausdorff distance** (Chap. 1) between the sets of traces from two given states.

The case $h_{ac}(x, y) = h_{po}(x, y) = 0$ corresponds to the *simulation* of x by y, written $x \leq y$, i.e., to the existence of a *weighted simulation relation* $R \subseteq S \times S$, i.e., whenever $(x, y) \in R$ and $(x, t, x') \in F$, then $(y, t, y') \in F$ with $(x', y') \in R$.

The case $h_{ac}(x, y) < \infty$ or $h_{po}(x, y) < \infty$ corresponds to the existence of an *unweighted simulation relation* $R \subseteq S \times S$, i.e., whenever $(x, y) \in R$ and $(x, t, x') \in F$, then $(y, t', y') \in F$ with $(x', y') \in R$ and $d(t, t') < \infty$.

The relation $\leq$ is a pre-order on S. Two states x and y are *similar* if $x \leq y$ and $y \leq x$; they are *bisimilar* if, moreover, the simulation R of x by y is the inverse of the simulation of y by x. Similarity is an equivalence relation on S which is coarser than the bisimilarity congruence.

The above trace and similarity system hemimetrics are quantitative generalizations of system relations: trace-equivalence and simulation pre-order, respectively.

- **Cellular automata distances**

Let $S, |S| \geq 2$, be a finite set (*alphabet*), and let S^∞ be the set of $\mathbb{Z}$-indexed bi-infinite sequences $\{x_i\}_{i=-\infty}^{\infty}$ (*configurations*) of elements of S. A (one-dimensional) *cellular automaton* is a continuous self-map $f : S^\infty \to S^\infty$ that commutes with all *shift* (or *translation*) maps $g : S^\infty \to S^\infty$ defined by $g(x_i) = x_{i+1}$.

Such cellular automaton form a discrete **dynamical system** with the time set $T = \mathbb{Z}$ (of *cells*, positions of a finite-state machine) on the finite-state space S. The main distances between configurations $\{x_i\}_i$ and $\{y_i\}_i$ (see [BFK99]) follow.

The **Cantor metric** is a metric on S^∞ defined, for $x \neq y$, by

$$2^{-\min\{i \geq 0 : |x_i - y_i| + |x_{-i} - y_{-i}| \neq 0\}}.$$

It corresponds to the case $a = \frac{1}{2}$ of the **generalized Cantor metric** in Chap. 11. The corresponding metric space is compact.

[redacted]

The **Besicovitch semimetric** is a semimetric on S^{∞} defined, for $x \neq y$, by

$$\overline{\lim}_{l \to \infty} \frac{|-l \leq i \leq l : x_i \neq y_i|}{2l+1}.$$

Cf. **Besicovitch distance** on measurable functions in Chap. 13. The corresponding semimetric space is complete.

The **Weyl semimetric** is a semimetric on S^{∞}, defined by

$$\overline{\lim}_{l \to \infty} \max_{k \in \mathbb{Z}} \frac{|k+1 \leq i \leq k+l : x_i \neq y_i|}{l}.$$

This and the above semimetric are **translation invariant**, but are neither separable nor locally compact. Cf. **Weyl distance** in Chap. 13.

- **Dynamical system**

  A (deterministic) **dynamical system** is a tuple (T, X, f) consisting of a metric space (X, d), called the *state space*, a *time set* T and an *evolution function* $f : T \times X \to X$. Usually, T is a monoid, (X, d) is a manifold locally diffeomorphic to a Banach space, and f is a continuous function.

  The system is *discrete* if $T = \mathbb{Z}$ (*cascade*) or if $T = \{0, 1, 2 \dots\}$. It is *real* (or *flow*) if T is an open interval in $\mathbb{R}$, and it is a *cellular automaton* if X is finite and $T = \mathbb{Z}^n$. Dynamical systems are studied in Control Theory in the context of stability; Chaos Theory considers the systems with maximal possible instability.

  A discrete dynamical system with $T = \{0, 1, 2 \dots\}$ is defined by a self-map $f : X \to X$. For any $x \in X$, its *orbit* is the sequence $\{f^n(x)\}_n$; here $f^n(x) = f(f^{n-1}(x))$ with $f^0(x) = x$. The orbit of $x \in X$ is called *periodic* if $f^n(x) = x$ for some $n > 0$.

  A pair $(x, y) \in X \times X$ is called *proximal* if $\underline{\lim}_{n \to \infty} d(f^n(x), f^n(y)) = 0$, and *distal*, otherwise. The system is called *distal* if any pair (x, y) of distinct points is distal.

  The dynamical system is called *expansive* if there exists a constant $D > 0$ such that the inequality $d(f^n(x), f^n(y)) \geq D$ holds for any distinct $x, y \in X$ and some n.

  An *attractor* is a closed subset A of X such that there exists an *open neighborhood* U of A with the property that $\lim_{n \to \infty} d(f^n(b), A) = 0$ for every $b \in U$, i.e., A *attracts* all nearby orbits. Here $d(x, A) = \inf_{y \in A} d(x, y)$ is the **point-set distance**.

  If for large n and small r there exists a number α such that

$$C(X, n, r) = \frac{|\{(i,j) : d(f^i(x), f^j(x)) \leq r, 1 \leq i, j \leq n\}|}{n^2} \sim r^{\alpha},$$

  then α is called (Grassberger–Hentschel–Procaccia, 1983) the *correlation dimension*.

  The *Lyapunov time* is the characteristic timescale on which a dynamical system is *chaotic*, i.e., small differences in initial conditions yield widely diverging outcomes, rendering long-term prediction impossible in general.

- **Melnikov distance**

  The evolution of a planar **dynamical system** can be represented in a 3D state space with orthogonal coordinate axes Ox, Ox', Ot. A *homoclinic orbit* (nongeneric orbit that joins a saddle point) can be seen in that space as the intersection with a plane of section $t = const$ of the *stable manifold* (the surface consisting of all trajectories that approach $\gamma_0 = Ot$ asymptotically in forward time) and the *unstable manifold* (the surface consisting of all trajectories that approach Ot in reverse time).

  Under a sufficiently small perturbation ϵ which is bounded and smooth enough, Ot persists as a smooth curve $\gamma_\epsilon = \gamma_0 + O(\epsilon)$, and the perturbed system has (not coinciding since $\epsilon > 0$) stable and unstable manifolds contained in an $O(\epsilon)$ neighborhood of the unperturbed manifolds.

  The **Melnikov distance** is the distance between stable and unstable manifolds measured along a line normal to the unperturbed manifolds, i.e., a direction that is perpendicular to the unperturbed homoclinic orbit. Cf. Sect. 18.2.
- **Fractal**

  For a metric space, its **topological dimension** does not exceed its **Hausdorff dimension**; cf. Chap. 1. A **fractal** is a metric space for which this inequality is strict. (Originally, Mandelbrot defined a fractal as a point set with noninteger Hausdorff dimension.) For example, the *Cantor set*, seen as a compact metric subspace of $(\mathbb{R}, d(x, y) = |x - y|)$ has the Hausdorff dimension $\frac{\ln 2}{\ln 3}$; cf. another **Cantor metric** in Chap. 11. Another classical fractal, the *Sierpinski carpet* of $[0, 1] \times [0, 1]$, is a **complete geodesic** metric subspace of $(\mathbb{R}^2, d(x, y) = ||x - y||_1)$.

  The term *fractal* is used also, more generally, for a *self-similar* (i.e., roughly, looking similar at any scale) object (usually, a subset of $\mathbb{R}^n$). Cf. **scale invariance**.
- **Scale invariance**

  Scale invariance is a feature of laws or objects which do not change if length scales are multiplied by a common factor.

  Examples of scale invariant phenomena are **fractals** and *power laws*; cf. **scale-free network** in Chap. 22 and *self-similarity* in **long range dependence**. Scale invariance arising from a power law $y = Cx^k$, for a constant C and scale exponent k, amounts to linearity $\log y = \log C + k \log x$ for logarithms. Much of scale invariant behavior (and complexity in nature) is explained (Bak–Tang–Wiesenfeld, 1987) by *self-organized cruciality (SOC)* of many **dynamical systems**, i.e., the property to have the critical point of a phase transition as an attractor which can be attained spontaneously without any fine-tuning of control parameters.

  Two moving systems are *dynamically similar* if the motion of one can be made identical to the motion of the other by multiplying all lengths by one scale factor, all forces by another one and all time periods by a third scale factor. Dynamic similarity can be formulated in terms of dimensionless parameters as, for example, the **Reynolds number** in Chap. 24.
- **Long range dependence**

  A (second-order stationary) stochastic process $X_k, k \in \mathbb{Z}$, is called **long range dependent** (or *long memory*) if there exist numbers $\alpha, 0 < \alpha < 1$, and $c_\rho > 0$

[redacted]

such that $\lim_{k\to\infty} c_\rho k^\alpha \rho(k) = 1$, where $\rho(k)$ is the autocorrelation function. So, correlations decay very slowly (asymptotically hyperbolic) to zero implying that $\sum_{k\in\mathbb{Z}} |\rho(k)| = \infty$, and that events so far apart are correlated (long memory). If the above sum is finite and the decay is exponential, then the process is *short range*.

Examples of such processes are the exponential, normal and Poisson processes which are memoryless, and, in physical terms, systems in thermodynamic equilibrium. The above power law decay for correlations as a function of time translates into a power law decay of the Fourier spectrum as a function of frequency f and is called $\frac{1}{f}$ *noise*.

A process has a *self-similarity exponent* (or *Hust parameter*) H if X_k and $t^{-H}X_{tk}$ have the same finite-dimensional distributions for any positive t. The cases $H = \frac{1}{2}$ and $H = 1$ correspond, respectively, to purely random process and to exact self-similarity: the same behavior on all scales. Cf. **fractal**, **scale invariance** and, in Chap. 22, **scale-free network**. The processes with $\frac{1}{2} < H < 1$ are long range dependent with $\alpha = 2(1 - H)$.

Long range dependence corresponds to *heavy-tailed* (or *power law*) distributions. The *distribution function* and *tail* of a nonnegative random variable X are $F(x) = P(X \le x)$ and $\overline{F(x)} = P(X > x)$. A distribution $F(X)$ is *heavy-tailed* if there exists a number $\alpha, 0 < \alpha < 1$, such that $\lim_{x\to\infty} x^\alpha \overline{F(x)} = 1$.

Many such distributions occur in the real world (for example, in Physics, Economics, the Internet) in both space (distances) and time (durations). A standard example is the Pareto distribution $\overline{F(x)} = x^{-k}$, $x \ge 1$, where $k > 0$ is a parameter. Cf. Sect. 18.4 and, in Chap. 29, **distance decay**.

Also, the random-copying model (the cultural analog of genetic drift) of the frequency distributions of various cultural traits (such as of scientific papers citations, first names, dog breeds, pottery decorations) results (Bentley–Hahn–Shennan, 2004) in a power law distribution $y = Cx^{-k}$, where y is the proportion of cultural traits that occur with frequency x in the population, and C and k are parameters.

A general *Lévy flight* is a random walk in which the increments have a power law probability distribution.

- **Lévy walks in human mobility**

  A *jump* is a longest straight line trip from one location to another done without a directional change or pause. Consider a 2D *random walk* (taking successive jumps, each in a random direction) model that involves two distributions: a uniform one for the *turning angle* θ_i and a power law $P(l_i) \sim l_i^{-\alpha}$ for the *jump length* l_i.

  Brownian motion has $\alpha \ge 3$ and *normal diffusion*, i.e., the MSD (mean squared displacement) grows linearly with time t: $MSD \sim t^\gamma, \gamma = 1$.

  A *Lévy walk* has $1 < \alpha < 3$. Its jump length is *scale-free*, i.e., lacks an average scale $\overline{l_i}$, and it is *superdiffusive*: $MSD \sim t^\gamma, \gamma > 1$. Intuitively, Lévy walks consist of many short jumps and, exceptionally, long jumps eliminating the effect of short ones in average jump lengths.

Lévy-like walk dispersal was observed in our Web browsing and image scanning, as well as in foraging animals. It and two-mode Brownian search strategies might be optimal for finding patches of randomly dispersed abundant resources: to cluster, in order to save time and effort, closely located activities and then make many short jumps within the clustered areas and a few long jumps among areas.

Human mobility occurs on many length scales, ranging from walking to air travel. On average, humans spend 1.1 h of their daily time budget traveling. Schafer and Victor, 2000, estimated the average travel distance, per person per year, as 1814, 4382 and 6787 km for 1960, 1990 and 2020, respectively.

Brockmann–Hafnagel–Geisel, 2006, studied long range human traffic via the geographic circulation of money. To track a bill, a user stamps it and enters data (serial number, series and local ZIP code) in a computer. The site www.wheresgeorge.com reports the time and distance between the bill's consecutive sightings. 57 % of all $\approx$ 465,000 considered bills traveled 50–800 km over 9 months in US. The probability of a bill traversing a distance r (an estimate of the probability of humans moving such a distance) followed, over 10–3500 km, a power law $P(r) = r^{-1.6}$. Banknote dispersal was fractal, and the bill trajectories resembled Lévy walks.

González–Hidalgo–Barabási, 2008, studied the trajectory of 100,000 anonymized mobile phone users (a random sample of 6 million) over 6 months. The probability of finding a user at a location of *rank* k (by the number of times a user was recorded in the vicinity) was $P(k) \sim \frac{1}{k}$. 40 % of the time users were found at their first two preferred locations (home, work), while spending remaining time in 5–50 places. Phithakkitnukoon et al., 2011, found that $\approx$ 80 % of places visited by mobile phone users are within of their *geo-social radius* (nearest social ties' locations) 20 km.

Jiang–Yin–Zhao, 2009, analyzed people's moving trajectories, obtained from GPS data of 50 taxicabs over 6 months in a large street network. They found a Lévy behavior in walks (both origin-destination and between streets) and attributed it to the fractal property of the underlying street network, not to the goal-directed nature of human movement. Rhee et al., 2009, analyzed $\approx$1000 h of GPS traces of walks of 44 participants. They also got Lévy walks.

- **Total distance between trajectories**

  A *trajectory* in $\mathbb{R}^k$ is a sequence $X = ((x_1, t_1), \dots, (x_n, t_n))$ of state-time points with $x_i \in \mathbb{R}^k$ and increasing "time" $t_i \in \mathbb{N}$. In $\mathbb{R}^2$, a trajectory can be seen as a list of consecutive GPS points for a moving object; sometimes, a semantic tag (say, "Shop, Restaurant") is attached to each point.

  Let $c(X)$ be the center of mass of $\{x_1, \dots, x_n\}$, $l(X)$ the length of X, and $\mathbf{s}(X)$ the **displacement** of X (cf. Sect. 24.1), i.e., the vector from x_1 to x_n.

  The **total distance between trajectories** $X = ((x_1, t_1), \dots, (x_n, t_n))$ and $X' = ((x'_1, t'_1), \dots, (x'_n, t'_{n'}))$ is (Liu–Schneider, 2012) $\frac{Geo(X,X')}{1+\alpha Sem(X.X')}$, where $\alpha \in [0, 1]$ is a parameter, the *semantic similarity* $Sem(X, X')$ is the length of their longest

[redacted]

common subsequence divided by $\min(n, n')$, while

$$d_E(c(X), c(X'))(1 + \frac{|l(X) - l(X')|}{\max(l(X), l(X'))}) - \frac{||\mathbf{s}(X)|| + ||\mathbf{s}(X')||}{2} \frac{\langle \mathbf{s}(X), \mathbf{s}(X') \rangle}{||\mathbf{s}(X)|| \cdot ||\mathbf{s}(X')||}$$

is their *geographic distance* $Geo(X, X')$.

The *Lyapunov time*, reflecting the limits of system's predictability, is the time for the distance between nearby system's trajectories to increase by a factor of e.

18.2 Distances in Control Theory

Control Theory deals with influencing the behavior of *dynamical systems*. It considers the feedback loop of a *plant* P (a function representing the object to be controlled, a system) and a *controller* C (a function to design). The output y, measured by a sensor, is fed back to the reference value r.

Then the controller takes the *error* $e = r - y$ to make inputs $u = Ce$. Subject to zero initial conditions, the input and output signals to the plant are related by $y = Pu$, where r, u, v and P, C are functions of the frequency variable s. So, $y = \frac{PC}{1+PC} r$ and $y \approx r$ (i.e., one controls the output by simply setting the reference) if PC is large for any value of s.

If the system is modeled by a system of linear differential equations, then its *transfer function* $\frac{PC}{1+PC}$, relating the output with the input, is a rational function. The plant P is *stable* if it has no poles in the closed right half-plane $\mathbb{C}_+ = \{z \in \mathbb{C} : Re(z) \geq 0\}$.

The *robust stabilization problem* is: given a *nominal* plant (a model) P_0 and some metric d on plants, find the open ball of maximal radius which is centered in P_0, such that some controller (rational function) C stabilizes every element of this ball.

The *graph* $G(P)$ of the plant P is the set of all bounded input-output pairs $(u, y = Pu)$. Both u and y belong to the *Hardy space* $H^2(\mathbb{C}_+)$ of the right half-plane; the graph is a closed subspace of $H^2(\mathbb{C}_+) + H^2(\mathbb{C}_+)$. In fact, $G(P)$ is a closed subspace of $H^2(\mathbb{C}^2)$, and $G(P) = f(P) \cdot H^2(\mathbb{C}^2)$ for some function $f(P)$, called the *graph symbol*.

Cf. a **dynamical system** and the **Melnikov distance** in Sect. 18.1.

- **Gap metric**

  The **gap metric** between plants P_1 and P_2 (Zames–El-Sakkary, 1980) is defined by

$$gap(P_1, P_2) = ||\Pi(P_1) - \Pi(P_2)||_2,$$

where $\Pi(P_i)$, $i = 1, 2$, is the orthogonal projection of the graph $G(P_i)$ of P_i seen as a closed subspace of $H^2(\mathbb{C}^2)$. We have

$$gap(P_1, P_2) = \max\{\delta_1(P_1, P_2), \delta_1(P_2, P_1)\},$$

where $\delta_1(P_1, P_2) = \inf_{Q \in H^\infty} ||f(P_1) - f(P_2)Q||_{H^\infty}$, and $f(P)$ is a graph symbol.

Here H^∞ is the space of matrix-valued functions that are analytic and bounded in the open right half-plane $\{s \in \mathbb{C} : \Re s > 0\}$; the H^∞-norm is the maximum singular value of the function over this space.

If A is an $m \times n$ matrix with $m < n$, then its n columns span an n-dimensional subspace, and the matrix B of the orthogonal projection onto the column space of A is $A(A^TA)^{-1}A^T$. If the basis is orthonormal, then $B = AA^T$.

In general, the **gap metric** between two subspaces of the same dimension is the *l_2-norm* of the difference of their orthogonal projections; see also the definition of this distance as an **angle distance between subspaces**.

In applications, when subspaces correspond to autoregressive models, the *Frobenius norm* is used instead of the l_2-norm. Cf. **Frobenius distance** in Chap. 12.

- **Vidyasagar metric**

  The **Vidyasagar metric** (or *graph metric*) between plants P_1 and P_2 is defined by

  $$\max\{\delta_2(P_1, P_2), \delta_2(P_2, P_1)\},$$

  where $\delta_2(P_1, P_2) = \inf_{||Q|| \leq 1} ||f(P_1) - f(P_2)Q||_{H^\infty}$.

  The **behavioral distance** is the gap between *extended* graphs of P_1 and P_2; a term is added to the graph $G(P)$, in order to reflect all possible initial conditions (instead of the usual setup with the initial conditions being zero).

- **Vinnicombe metric**

  The **Vinnicombe metric** (*ν-gap metric*) between plants P_1 and P_2 is defined by

  $$\delta_\nu(P_1, P_2) = ||(1 + P_2P_2^*)^{-\frac{1}{2}}(P_2 - P_1)(1 + P_1^*P_1)^{-\frac{1}{2}}||_\infty$$

  if $wno(f^*(P_2)f(P_1)) = 0$, and it is equal to 1, otherwise.

  Here $f(P)$ is the graph symbol function of plant P. See [Youn98] for the definition of the *winding number* $wno(f)$ of a rational function f and for a good introduction to Feedback Stabilization.

- **Lanzon–Papageorgiou quasi-distance**

  Given a plant P, a *perturbed plant* $\hat{P}$ and an uncertainty structure expressed via a *generalized plant* H, let Δ be the set of all possible perturbations that explain the discrepancy between P and $\hat{P}$. Then **Lanzon–Papageorgiou quasi-distance** (2009) between P and $\hat{P}$ is defined as ∞ if $\Delta = \emptyset$ and $\inf_{\delta \in \Delta} ||\delta||_\infty$, otherwise.

  This quasi-distance corresponds to the worst-case degradation of the stability margin due to a plant perturbation. For standard uncertainity structures H, it is a metric, but it is only a quasi-metric for multiplicative uncertainity.

- **Distance to uncontrollability**

  Linear Control Theory concerns a system of the form $\bar{x} = Ax(t) + Bu(t)$, where, at each time t, $x(t) \in \mathbb{C}^n$ is the *state vector*, $u(t) \in \mathbb{C}^m$ is the *control input vector*, and $A \in \mathbb{C}^{n \times n}$, $B \in \mathbb{C}^{n \times m}$ are the given matrices. The system (matrix pair (A, B)) is called *controllable* if, for any initial and final states $x(0)$ and $x(T)$, there

[redacted]

exists $u(t), 0 \leq t \leq T$, that drive the state from $x(0)$ to $x(T)$ within finite time, or, equivalently (Kalman, 1963) the matrix $A - \lambda IB$ has full row rank for all $\lambda \in \mathbb{C}$. The **distance to uncontrollability** (Paige, 1981, and Eising, 1984) is defined as

$$\min\{||[E, F]|| : (A + E, B + F) \text{ is uncontrollable}\} = \min_{\lambda \in \mathbb{C}} \sigma_n(A - \lambda IB),$$

where $||.||$ is the *spectral* or *Frobenius norm* (cf. Sect. 12.3) and $\sigma_n(A - \lambda IB)$ denotes the n-th largest singular value of the $(n \times (n + m))$-matrix $A - \lambda IB$.

A matrix $A \in \mathbb{C}^{n \times n}$ is *stable* if any its eigenvalue λ has real part $Re(\lambda) < 0$. The **distance to instability** is (Van Loan, 1985) $\min\{||E||\} : A + E$ is unstable$\}$, where $||.||$ is one of two above norms. Cf. **nearness matrix problems** in Chap. 12.

18.3 Motion Planning Distances

Automatic motion planning methods are applied in *Robotics*, *Virtual Reality Systems* and *Computer Aided Design*. A **motion planning metric** is a metric used in automatic motion planning methods.

Let a *robot* be a finite collection of rigid links organized in a kinematic hierarchy. If the robot has n degrees of freedom, this leads to an n-dimensional *manifold* C, called the *configuration space* (or *C-space*) of the robot. The *workspace* W of the robot is the space (usually, $\mathbb{E}^3$) in which the robot moves. Usually, it is modeled as the Euclidean space $\mathbb{E}^3$. A **workspace metric** is a motion planning metric in the workspace $\mathbb{R}^3$.

The *obstacle region* CB is the set of all configurations $q \in C$ that either cause the robot to collide with obstacles B, or cause different links of the robot to collide among themselves. The closure $cl(C_{free})$ of $C_{free} = C \backslash \{CB\}$ is called the *space of collision-free configurations*. A *motion planning algorithm* must find a collision-free path from an initial configuration to a goal configuration.

A **configuration metric** is a motion planning metric on the configuration space C of a robot. Usually, the configuration space C consists of six-tuples $(x, y, z, \alpha, \beta, \gamma)$, where the first three coordinates define the position, and the last three the orientation. The orientation coordinates are the angles in radians.

Intuitively, a good measure of the distance between two configurations is a measure of the workspace region swept by the robot as it moves between them (the **swept volume distance**). However, the computation of such a metric is prohibitively expensive.

The simplest approach has been to consider the C-space as a Cartesian space and to use Euclidean distance or its generalizations. For such **configuration metrics**, one normalizes the orientation coordinates so that they get the same magnitude as the position coordinates. Roughly, one multiplies the orientation coordinates by the maximum x, y or z range of the workspace bounding box. Examples of such metrics are given below.

More generally, the configuration space of a 3D rigid body can be identified with the Lie group $ISO(3)$: $C \cong \mathbb{R}^3 \times \mathbb{R}P^3$. The general form of a matrix in $ISO(3)$ is given by

$$\begin{pmatrix} R & X \\ 0 & 1 \end{pmatrix},$$

where $R \in SO(3) \cong \mathbb{R}P^3$, and $X \in \mathbb{R}^3$.

If X_q and R_q represent the translation and rotation components of the configuration $q = (X_q, R_q) \in ISO(3)$, then a configuration metric between configurations q and r is given by $w_{tr}||X_q - X_r|| + w_{rot}f(R_q, R_r)$, where the **translation distance** $||X_q - X_r||$ is obtained using some norm $||.||$ on $\mathbb{R}^3$, and the **rotation distance** $f(R_q, R_r)$ is a positive scalar function which gives the distance between the rotations $R_q, R_r \in SO(3)$. The rotation distance is scaled relative to the translation distance via the weights w_{tr}, w_{rot}.

There are many other types of metrics used in motion planning methods, in particular, the **Riemannian metrics**, the **Hausdorff metric** and, in Chap. 9, the **separation distance**, the **penetration depth distance** and the **growth distances**.

- **Weighted Euclidean $\mathbb{R}^6$-distance**

  The **weighted Euclidean $\mathbb{R}^6$-distance** is a **configuration metric** on $\mathbb{R}^6$ defined, for any $x, y \in \mathbb{R}^6$, by

  $$\left(\sum_{i=1}^{3} |x_i - y_i|^2 + \sum_{i=4}^{6}(w_i|x_i - y_i|)^2\right)^{\frac{1}{2}},$$

  where $x = (x_1, \dots, x_6)$, x_1, x_2, x_3 are the position coordinates, x_4, x_5, x_6 are the orientation coordinates, and w_i is the normalization factor. Cf. the general, i.e., in $\mathbb{R}^n$, **weighted Euclidean distance** in Chap. 17.

  The *scaled weighted Euclidean $\mathbb{R}^6$-distance* is defined, for any $x, y \in \mathbb{R}^6$, by

  $$\left(s\sum_{i=1}^{3} |x_i - y_i|^2 + (1-s)\sum_{i=4}^{6}(w_i|x_i - y_i|)^2\right)^{\frac{1}{2}}.$$

  This distance changes the relative importance of the position and orientation components through the scale parameter s.

- **Weighted Minkowskian distance**

  The **weighted Minkowskian distance** is a **configuration metric** on $\mathbb{R}^6$ defined, for any $x, y \in \mathbb{R}^6$, by

  $$\left(\sum_{i=1}^{3} |x_i - y_i|^p + \sum_{i=4}^{6}(w_i|x_i - y_i|)^p\right)^{\frac{1}{p}}.$$

[redacted]

It gives the same importance to both position and orientation.

- **Modified Minkowskian distance**

  The **modified Minkowskian distance** is a **configuration metric** on $\mathbb{R}^6$ defined, for any $x, y \in \mathbb{R}^6$, by

$$\left(\sum_{i=1}^{3} |x_i - y_i|^{p_1} + \sum_{i=4}^{6} (w_i |x_i - y_i|)^{p_2}\right)^{\frac{1}{p_3}}.$$

  It distinguishes between position and orientation coordinates using the parameters $p_1 \geq 1$ (for the position) and $p_2 \geq 1$ (for the orientation).

- **Weighted Manhattan distance**

  The **weighted Manhattan distance** is a **configuration metric** on $\mathbb{R}^6$ defined, for any $x, y \in \mathbb{R}^6$, by

$$\sum_{i=1}^{3} |x_i - y_i| + \sum_{i=4}^{6} w_i |x_i - y_i|.$$

- **Robot displacement metric**

  The **robot displacement metric** (or *DISP distance*, Latombe, 1991, and LaValle, 2006) is a **configuration metric** on a configuration space C of a robot defined by

$$\max_{a \in A} ||a(q) - a(r)||$$

  for any two configurations $q, r \in C$, where $a(q)$ is the position of the point a in the workspace $\mathbb{R}^3$ when the robot is at configuration q, and $||.||$ is one of the norms on $\mathbb{R}^3$, usually the Euclidean norm. Intuitively, this metric yields the maximum amount in workspace that any part of the robot is displaced when moving from one configuration to another (cf. **bounded box metric**).

- **Euler angle metric**

  The **Euler angle metric** is a **rotation metric** on the group $SO(3)$ (for the case of using three—Heading-Elevation-Bank—*Euler angles* to describe the orientation of a rigid body) defined by

$$w_{rot}\sqrt{\Delta(\theta_1, \theta_2)^2 + \Delta(\phi_1, \phi_2)^2 + \Delta(\eta_1, \eta_2)^2}$$

  for all $R_1, R_2 \in SO(3)$, given by Euler angles $(\theta_1, \phi_1, \eta_1)$, $(\theta_2, \phi_2, \eta_2)$, respectively, where $\Delta(\theta_1, \theta_2) = \min\{|\theta_1 - \theta_2|, 2\pi - |\theta_1 - \theta_2|\}$, $\theta_i \in [0, 2\pi]$, is the **metric between angles**, and w_{rot} is a scaling factor.

- **Unit quaternions metric**

  The **unit quaternions metric** is a **rotation metric** on the *unit quaternion representation* of $SO(3)$, i.e., a representation of $SO(3)$ as the set of points (*unit* *quaternions*) on the *unit sphere* S^3 in $\mathbb{R}^4$ with identified antipodal points ($q \sim -q$).

This representation of $SO(3)$ suggested a number of possible metrics on it, for example, the following ones:

1. $\min\{||q - r||, ||q + r||\}$,
2. $|| \ln(q^{-1}r)||$,
3. $w_{rot}(1 - |\lambda|)$,
4. $\arccos |\lambda|$,

where $q = q_1 + q_2 i + q_3 j + q_4 k$, $\sum_{i=1}^{4} q_i^2 = 1$, $||.||$ is a norm on $\mathbb{R}^4$, $\lambda = \langle q, r\rangle = \sum_{i=1}^{4} q_i r_i$, and w_{rot} is a scaling factor.

- **Center of mass metric**

  The **center of mass metric** is a **workspace metric**, defined as the Euclidean distance between the *centers of mass* of the robot in the two configurations. The center of mass is approximated by averaging all object vertices.

- **Bounded box metric**

  The **bounded box metric** is a **workspace metric** defined as the maximum Euclidean distance between any vertex of the *bounding box* of the robot in one configuration and its corresponding vertex in the other configuration.

  The **box metric** in Chap. 4 is unrelated.

- **Pose distance**

  A **pose distance** provides a measure of dissimilarity between actions of *agents* (including robots and humans) for Learning by Imitation in Robotics.

  In this context, agents are considered as *kinematic chains*, and are represented in the form of a *kinematic tree*, such that every link in the kinematic chain is represented by a unique edge in the corresponding tree.

  The configuration of the chain is represented by the *pose* of the corresponding tree which is obtained by an assignment of the pair (n_i, l_i) to every edge e_i. Here n_i is the unit normal, representing the orientation of the corresponding link in the chain, and l_i is the length of the link.

  The *pose class* consists of all poses of a given kinematic tree. One of the possible pose distances is a distance on a given pose class which is the sum of measures of dissimilarity for every pair of compatible segments in the two given poses.

  Another way is to view a *pose* $D(m)$ in the context of the a precedent and a subsequent frames as a *3D point cloud* $\{D^j(i) : m - a \leq i \leq m + a, j \in J\}$, where J is the joint set. The set $D(m)$ contains $k = |J|(2a + 1)$ points (joint positions) $p_i = (x_i, y_i, z_i)$, $1 \leq i \leq k$. Let $T_{\theta,x,z}$ denote the linear transformation which simultaneously rotates all points of a point cloud about the y axis by an angle $\theta \in [0.2\pi]$ and then shifts the resulting points in the xz plane by a vector $(x, 0, z) \in \mathbb{R}^3$. Then the **3D point cloud distance** (Kover and Gleicher, 2002)

[redacted]

between the poses $D(m) = (p_i)_{i\in[1,k]}$ and $D(n) = (q_i)_{i\in[1,k]}$ is defined as

$$\min_{\theta,x,z}\{\sum_{i=1}^{k} ||p_i - T_{\theta,x,z}(q_i)||_2^2\}.$$

Cf. **Procrustes distance** in Chap. 21.

- **Joint angle metric**

  For a given frame (or pose) i in an animation, let us define $p_i \in \mathbb{R}^3$ as the global (root) position and $q_{i,k} \in S^3$ as the unit quaternion describing the orientation of a joint k from the joint set J. Cf. **unit quaternions metric** and **3D point cloud distance**. The **joint angle metric** between frames x and y is defined as follows:

$$|p_x - p_y|^2 + \sum_{k\in J} w_k|\log(q_{y,k}^{-1}q_{x,k})|^2.$$

  The second term describes the weighted sum of the orientation differences; cf. **weighted Euclidean $\mathbb{R}^6$-distance**. Sometimes, the terms expressing differences in derivatives, such as joint velocity and acceleration, are added.

- **Millibot train metrics**

  In *Microbotics* (the field of miniature mobile robots), *nanorobot*, *microrobot*, *millirobot*, *minirobot*, and *small robot* are terms for robots with characteristic dimensions at most one micrometer, mm, cm, dm, and m, respectively.

  A *millibot train* is a team of heterogeneous, resource-limited millirobots which can collectively share information. They are able to fuse range information from a variety of different platforms to build a global occupancy map that represents a single collective view of the environment.

  In the construction of a **motion planning metric** of millibot trains, one casts a series of random points about a robot and poses each point as a candidate position for movement. The point with the highest overall utility is then selected, and the robot is directed to that point. Thus:

  the **free space metric**, determined by free space contours, only allows candidate points that do not drive the robot through obstructions;
  the **obstacle avoidance metric** penalizes for moves that get too close to obstacles;
  the **frontier metric** rewards for moves that take the robot towards open space;
  the **formation metric** rewards for moves that maintain formation;
  the **localization metric**, based on the separation angle between one or more localization pairs, rewards for moves that maximize localization (see [GKC04]).

  Cf. **collision avoidance distance** and **piano movers distance** in Chap. 19.

  A *swarm-bot* can form more complex (more sensors and actuators) and flexible (interconnecting at several angles and with less accuracy) configurations.

The wingspan range of flying robots includes 2.8 cm (quadcopter Lisa/S) and 40 m (Global Hawk). During 2012, a robot Papa Mau (PacX Wave Glider), piloted remotely, swam 16,668 km from San Francisco to Australia.

18.4 MOEA Distances

Most optimization problems have several objectives but, for simplicity, only one of them is optimized, and the others are handled as constraints. *Multi-objective optimization* considers (besides some inequality constraints) an objective vector function $f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}^k$ from the *search* (or *genotype*, *decision variables*) space X to the *objective* (or *phenotype*, *decision vectors*) space $f(X) = \{f(x) : x \in X\} \subset \mathbb{R}^k$.

A point $x^* \in X$ is a *Pareto-optimal solution* if, for every other $x \in X$, the decision vector $f(x)$ does not *Pareto-dominate* $f(x^*)$, i.e., $f(x) \leq f(x^*)$. The *Pareto-optimal front* is the set $PF^* = \{f(x) : x \in X^*\}$, where X^* is the set of all Pareto-optimal solutions.

Multi-objective evolutionary algorithms (MOEA) produce, at each generation, an *approximation set* (the found Pareto front PF_{known} approximating the desired Pareto front PF^*) in objective space in which no element Pareto-dominates another element. Examples of **MOEA metrics**, i.e., measures evaluating how close PF_{known} is to PF^*, follow.

- **Generational distance**

  The **generational distance** is defined by

$$\frac{(\sum_{j=1}^{m} d_j^2)^{\frac{1}{2}}}{m},$$

  where $m = |PF_{known}|$, and d_j is the Euclidean distance (in the objective space) between $f^j(x)$ (i.e., j-th member of PF_{known}) and the nearest member of PF^*. This distance is zero if and only if $PF_{known} = PF^*$.

  The term **generational distance** (or *rate of turnover*) is also used for the minimal number of branches between two positions in any system of ranked descent represented by a hierarchical tree. Examples are: **phylogenetic distance** on a phylogenetic tree (cf. Chap. 23), the number of generations separating a photocopy from the original block print, and the number of generations separating the audience at a memorial from the commemorated event.
- **Spacing**

  The **spacing** is defined by

$$\left(\frac{\sum_{j=1}^{m}(\overline{d} - d_j)^2}{m-1}\right)^{\frac{1}{2}},$$

[redacted]

where $m = |PF_{known}|$, d_j is the L_1-*metric* (in the objective space) between $f^j(x)$ (i.e., j-th member of PF_{known}) and the nearest other member of PF_{known}, while $\overline{d}$ is the mean of all d_j.

- **Overall nondominated vector ratio**
  The **overall nondominated vector ratio** is defined by

$$\frac{|PF_{known}|}{|PF^*|}.$$

- **Crowding distance**
  The **crowding distance** (Deb et al., 2002) is a diversity metric assigned to each Pareto-optimal solution. It is the sum, for all objectives, of the absolute difference of the objective values of two nearest solutions on each side, if they exist.

  The *boundary solutions*, i.e., those with the smallest or the highest such value, are assigned an infinite crowding distance.

Part V
Computer-Related Distances

[redacted]

Chapter 19
Distances on Real and Digital Planes

19.1 Metrics on Real Plane

Any L_p**-metric** (as well as any **norm metric** for a given norm $||.||$ on $\mathbb{R}^2$) can be used on the plane $\mathbb{R}^2$, and the most natural is the L_2-metric, i.e., the Euclidean metric $d_E(x,y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$ which gives the length of the straight line segment $[x,y]$, and is the **intrinsic metric** of the plane.

However, there are other, often "exotic", metrics on $\mathbb{R}^2$. Many of them are used for the construction of *generalized Voronoi diagrams* on $\mathbb{R}^2$ (see, for example, **Moscow metric**, **network metric**, **nice metric**). Some of them are used in Digital Geometry.

- **Erdős-type distance problems**

  Those distance problems were given by Erdős and his collaborators, usually, for the Euclidean metric on $\mathbb{R}^2$, but they are of interest for $\mathbb{R}^n$ and for other metrics on $\mathbb{R}^2$. Examples of such problems are to find out:

  the least number of different distances (or largest occurrence of a given distance) in an m-subset of $\mathbb{R}^2$;
  the largest cardinality of a subset of $\mathbb{R}^2$ determining at most m distances;
  the minimum diameter of an m-subset of $\mathbb{R}^2$ with only integral distances (or, say, without a pair (d_1, d_2) of distances with $0 < |d_1 - d_2| < 1$);
  the *Erdős-diameter* of a given set S, i.e., the minimum diameter of a rescaled set rS, $r > 0$, in which any two different positive distances differ at least by one;
  the largest cardinality of an *isosceles set* in $\mathbb{R}^2$, i.e., a set of points, any three of which form an isosceles triangle;
  existence of an m-subset of $\mathbb{R}^2$ with, for each $1 \leq i \leq m$, a distance occurring exactly i times (examples are known for $m \leq 8$);

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_19

existence of a dense subset of $\mathbb{R}^2$ with rational distances (Ulam problem);
existence of $m, m > 7$, noncollinear points of $\mathbb{R}^2$ with integral distances;
forbidden (not occurring within each part) *distances* of a partition of $\mathbb{R}^2$.

The general **Erdős distinct distances problem**, still open for $n > 2$, is to prove that if $A \subset \mathbb{R}^n$, $|A| = m$ and $d(A)$ denotes the set $\{\sum_{i=1}^{n}(x_i - y_i)^2 : x, y \in A\}$, then $|d(A)| \geq Cm^{\frac{n}{2}}$ for some constant $C > 0$. This problem was generalized for distinct "distances" (cf. Chap. 3) over a finite field. Also, its continuous analog, open **Falconer distance problem** is to prove that if the **Hausdorff dimension** of $A \subset \mathbb{R}^n$ is $> \frac{n}{2}$, then 1-dimensional *Lebesque measure* of $d(A)$ is positive. Related result by Quas, 2009: if the *upper density* of $A \subset \mathbb{R}^n$ is positive, then there is $r_0 > 0$ such that for any $r > r_0$, there are $x, y \in A$ with $d_E(x, y) = r$.

The *three-distance theorem* (Sós, 1957): given $a \in (0, 1)$ and $n \in \mathbb{N}$, the points $\{0\}, \{a\}, \{2a\}, \dots, \{na\} \pmod 1$ on the circle of perimeter 1, partition it into $n + 1$ intervals having at most three lengths, one being the sum of the other two.

The **problem of dispersion** (or *finding System of Distant Representatives*) in a family of n subsets of a metric space is (Fiala et al., 2005) that of selecting n points, one in each subset, such that the minimum inter-point distance is maximized.

- **Distance inequalities in a triangle**

  The multitude of inequalities, involving Euclidean distances between points of $\mathbb{R}^n$, is represented below by some **distance inequalities in a triangle**.

  Let $\triangle ABC$ be a triangle on $\mathbb{R}^2$ with side-lengths $a = d(B, C), b = d(C, A), c = d(A, B)$ and area $\mathcal{A} = \frac{1}{4}\sqrt{(a^2 + b^2 + c^2)^2 - 2(a^4 + b^4 + c^4)}$.

  Let P, P' be two arbitrary interior points in $\triangle ABC$. Denote by D_A, D_B, D_C the distances $d(P, A), d(P, B), d(P, C)$ and by d_A, d_B, d_C the **point-line distances** (Chap. 4) from P to the sides BC, CA, AB opposite to A, B, C. For the point P' define D'_A, D'_B, D'_C and d'_A, d'_B, d'_C similarly.

  The point P is *circumcenter* if $D_A = D_B = D_C$; this distance, $R = \frac{abc}{4\mathcal{A}}$, is circumcircle's radius. The point P is *incenter* if $d_A = d_B = d_C$; this distance, $r = \frac{2\mathcal{A}}{a+b+c}$, is incircle's radius. The *centroid* (the center of mass) is the point G of concurrency of three triangle's medians m_a, m_b, m_c; it holds $d(A, G) = \frac{2}{3}m_a, d(B, G) = \frac{2}{3}m_b, d(C, G) = \frac{2}{3}m_c$. The *symmedian point* is the point of concurrency of three triangle's *symmedians* (reflections of medians at corresponding angle bisectors).

  The *orthocenter* is the point of concurrency of three triangle's altitudes. The centroid is situated on the *Euler line* through the circumcenter and the orthocenter, at $\frac{1}{3}$ of their distance. At $\frac{1}{2}$ of their distance lies the center of the circle going through the midpoints of three sides and the feet of three altitudes.

  – If P and P' are the circumcenter and incenter of $\triangle ABC$, then (Euler, 1765)

$$d^2(P, P') \geq R(R - 2r)$$

[redacted]

holds implying $R \geq 2r$ with equality if and only if triangle is equilateral. In fact, the general *Euler's inequality* $R \geq nr$ holds (Klamkin–Tsintsifas, 1979) for the radii R, r of circumscribed and inscribed spheres of an n-simplex.

- For any P, P', the *Erdős–Mordell* inequality (Mordell–Barrow, 1937) is

$$D_A + D_B + D_C \geq 2(d_A + d_B + d_C).$$

Liu, 2008, generalized above as follows: for all $x, y, z \geq 0$ it holds

$$\sqrt{D_A D'_A}x^2 + \sqrt{D_B D'_B}y^2 + \sqrt{D_C D'_C}z^2 \geq 2(\sqrt{d_A d'_A}yz + \sqrt{d_B d'_B}xz + \sqrt{d_C d'_C}xy).$$

- Lemoine, 1873, proved that

$$\frac{4\mathcal{A}^2}{a^2 + b^2 + c^2} \leq d_A^2 + d_B^2 + d_C^2$$

with equality if and only if P is the symmedian point.

- Posamentier and Salkind, 1996, showed

$$\frac{3}{4}(a+b+c) < m_a+m_b+m_c < a+b+c, \text{ while } \frac{3}{4}(a^2+b^2+c^2) = m_a^2+m_b^2+m_c^2.$$

- Kimberling, 2010, proved that

$$d_A d_B d_C \leq \frac{8\mathcal{A}^3}{27abc}$$

with equality if and only if P is the centroid.

He also gave (together with unique point realizing equality) inequality

$$\frac{(2\mathcal{A})^q}{(a^{\frac{2}{q-1}} + b^{\frac{2}{q-1}} + c^{\frac{2}{q-1}})^{q-1}} \leq d_A^q + d_B^q + d_C^q$$

for any $q < 0$ or $q > 1$. For $0 < q < 1$, the reverse inequality holds.

The side-lengths $d(A, B), d(B, C), d(C, A)$ of a *right triangle* are in arithmetic progression only if their ratio is 3:4:5. They are in geometric progression only if their ratio is $1{:}\sqrt{\varphi}{:}\varphi$, where φ is the *golden section* $\frac{1+\sqrt{5}}{2}$.

- **City-block metric**

The **city-block metric** is the L_1*-metric* on $\mathbb{R}^2$ defined by

$$||x - y||_1 = |x_1 - y_1| + |x_2 - y_2|.$$

It is also called the **taxicab metric**, **Manhattan metric**, **rectilinear metric**, **right-angle metric**, 4-**metric** and, on $\mathbb{Z}^n$, **grid metric**. The *von Neumann neighborhood* of a point is the set of points at a Manhattan distance of 1 from it.

- **Chebyshev metric**

  The **Chebyshev metric** (or **chessboard metric**, **king-move metric**, 8-**metric**) is the L_∞-*metric* on $\mathbb{R}^2$ defined by

  $$||x-y||_\infty = \max\{|x_1-y_1|, |x_2-y_2|\}.$$

  On $\mathbb{Z}^n$, this metric is called also the **lattice** (or **uniform**, **sup**) **metric**. A point's *Moore neighborhood* is the set of points at a Chebyshev distance of 1.

- α-**metric**

  Given $\alpha \in [0, \frac{\pi}{4}]$, the α-**metric** for $x, y \in \mathbb{R}^2$ is defined (Tian, 2005) by

  $$d_\alpha(x,y) = \max\{|x_1-y_1|, |x_2-y_2|\} - (\sec\alpha - \tan\alpha)\min\{|x_1-y_1|, |x_2-y_2|.$$

  It is the **city-block metric** if $\alpha = 0$. For $\alpha = \frac{\pi}{4}$, i.e., $\sec\alpha - \tan\alpha = \sqrt{2}-1$, it is the **Chinese checkers metric** (Chen, 1992). *Chinese checkers* (as well as *Hexagonal chess*, *Masonic chess*, *Sannin shogi*, *Hexshogi*) is a strategy board game with hexagonal cells, while *Tiangular chess*, *Tri-chess*, *Trishogi* have triangular cells. Cf. **hexagonal metric**. Gelişgen and Kaya, 2006, generalized α-metric on $\mathbb{R}^n$.

- **Relative metrics on** $\mathbb{R}^2$

  The (p,q)-**relative** and M-**relative** distances are defined in Chap. 5 on any *Ptolemaic space*. The (p,q)-**relative distance** on $\mathbb{R}^2$ (in general, on $\mathbb{R}^n$) is defined (for x or $y \neq 0$) in the cases $1 \leq p < \infty$ and $p = \infty$, respectively, by

  $$\frac{||x-y||_2}{(\frac{1}{2}(||x||_2^p + ||y||_2^p))^{\frac{q}{p}}} \text{ and } \frac{||x-y||_2}{(\max\{||x||_2, ||y||_2\})^q}.$$

  Let $f : [0,\infty) \to (0,\infty)$ be a convex increasing function such that $\frac{f(x)}{x}$ is decreasing for $x > 0$. The M-**relative distance** on $\mathbb{R}^2$ (in general, on $\mathbb{R}^n$), is defined by

  $$\frac{||x-y||_2}{f(||x||_2) \cdot f(||y||_2)}.$$

  In particular, the distance below is a metric if and only if $p \geq 1$:

  $$\frac{||x-y||_2}{\sqrt[p]{1+||x||_2^p}\sqrt[p]{1+||y||_2^p}}.$$

  A similar metric on $\mathbb{R}^2 \backslash \{0\}$ (in general, on $\mathbb{R}^n \backslash \{0\}$) is defined by $\frac{||x-y||_2}{||x||_2 \cdot ||y||_2}$.

[redacted]

- **MBR metric**

  The **MBR metric** (Schönemann, 1982, for bounded response scales in Psychology) is a metric $d((x_1, x_2), (y_1, y_2))$ on $\mathbb{R}^2$, defined by

$$\frac{|x_1 - y_1| + |x_2 - y_2|}{1 + |x_1 - y_1||x_2 - y_2|} = \tanh(\operatorname{arctanh}(|x_1 - y_1|) + \operatorname{arctanh}(|x_2 - y_2|)).$$

- **Moscow metric**

  The **Moscow** (or *Karlsruhe*, *Amsterdam*) *metric* is a metric on $\mathbb{R}^2$, defined as the minimum Euclidean length of all *admissible* connecting curves between x and $y \in \mathbb{R}^2$, where a curve is called *admissible* if it consists only of *radial streets* (segments of straight lines passing through the origin) and *circular avenues* (segments of circles centered at the origin); see, for example, [Klei88]).

  If the polar coordinates for points $x, y \in \mathbb{R}^2$ are (r_x, θ_x), (r_y, θ_y), respectively, then the distance between them is equal to $\min\{r_x, r_y\}\Delta(\theta_x - \theta_y) + |r_x - r_y|$ if $0 \leq \Delta(\theta_x, \theta_y) < 2$, and is equal to $r_x + r_y$ if $2 \leq \Delta(\theta_x, \theta_y) < \pi$, where $\Delta(\theta_x, \theta_y) = \min\{|\theta_x - \theta_y|, 2\pi - |\theta_x - \theta_y|\}$, $\theta_x, \theta_y \in [0, 2\pi)$, is the **metric between angles**.
- **French Metro metric**

  Given a norm $||.||$ on $\mathbb{R}^2$, the **French Metro metric** is a metric on $\mathbb{R}^2$ defined by

$$||x - y|| \text{ if } x = cy \text{ for some } 0 \neq c \in \mathbb{R} \text{ (i.e., } x_1y_2 = x_2y_1),$$

  and by

$$||x|| + ||y||, \text{ otherwise.}$$

  For the Euclidean norm $||.||_2$, it is called the **Paris metric**, **radial metric**, **hedgehog metric**, or **French railroad metric**, **enhanced SNCF metric**.

  In this case it can be defined as the minimum Euclidean length of all *admissible* connecting curves between two given points x and y, where a curve is called *admissible* if it consists only of segments of straight lines passing through the origin.

  In graph terms, this metric is similar to the **path metric** of the tree consisting of a point from which radiate several disjoint paths. In the case when only one line radiates from the point, this metric is called the **train metric**.

  The Paris metric is an example of an $\mathbb{R}$**-tree** T which is *simplicial*, i.e., its set of points x with $T\backslash\{x\}$ not having exactly two components, is discrete and closed.
- **Lift metric**

  The **lift metric** (or **jungle river metric**, *raspberry picker metric*, *barbed wire metric*) is a metric $d((x_1, x_2), (y_1, y_2))$ on $\mathbb{R}^2$ defined (see, for example, [Brya85]) by

$$|x_1 - y_1| \text{ if } x_2 = y_2,$$

and by

$$|x_1| + |x_2 - y_2| + |y_1| \text{ if } x_2 \neq y_2.$$

It is the minimum Euclidean length of all *admissible* (consisting only of segments of straight lines parallel to the x_1 axis and segments of the x_2 axis) connecting curves between points (x_1, x_2) and (y_1, y_2).

The lift metric is an *nonsimplicial* (cf. **French Metro metric**) **$\mathbb{R}$-tree**.

- **Radar screen metric**

  Given a norm $||.||$ on $\mathbb{R}^2$ (in general, on $\mathbb{R}^n$), the **radar screen metric** is a special case of the ***t*-truncated metric** (Chap. 4) defined by

  $$\min\{1, ||x - y||\}.$$

- **British Rail metric**

  Given a norm $||.||$ on $\mathbb{R}^2$ (in general, on $\mathbb{R}^n$), the **British Rail metric** is a metric defined as 0 for $x = y$ and, otherwise, by

  $$||x|| + ||y||.$$

  It is also called the **Post Office metric**, **caterpillar metric** and **shuttle metric**.

- **Flower-shop metric**

  Let d be a metric on $\mathbb{R}^2$, and let f be a fixed point (a *flower-shop*) in the plane.

  The **flower-shop metric** (sometimes called **SNCF metric**) is a metric on $\mathbb{R}^2$ (in general, on any metric space) defined by

  $$d(x, f) + d(f, y)$$

  for $x \neq y$ (and is equal to 0, otherwise). So, a person living at point x, who wants to visit someone else living at point y, first goes to f, to buy some flowers. In the case $d(x, y) = ||x - y||$ and the point f being the origin, it is the **British Rail metric**.

  If $k > 1$ flower-shops $f_1, \dots, f_k$ are available, one buys the flowers, where the detour is a minimum, i.e., the distance between distinct points x, y is equal to $\min_{1 \leq i \leq k}\{d(x, f_i) + d(f_i, y)\}$.

- **Rickman's rug metric**

  Given a number $\alpha \in (0, 1)$, the **Rickman's rug metric** on $\mathbb{R}^2$ is a 2D case of the **parabolic distance** (Chap. 6) defined by

  $$d((x_1, x_2), (y_1, y_2)) = |x_1 - y_1| + |x_2 - y_2|^{\alpha}.$$

- **Burago–Burago–Ivanov metric**

  The **Burago–Burago–Ivanov metric** ([BBI01]) is a metric on $\mathbb{R}^2$ defined by

  $$|\,||x||_2 - ||y||_2| + \min\{||x||_2, ||y||_2\} \cdot \sqrt{\angle(x, y)},$$

[redacted]

where $\angle(x, y)$ is the angle between vectors x and y, and $||.||_2$ is the Euclidean norm on $\mathbb{R}^2$. The corresponding **internal metric** on $\mathbb{R}^2$ is equal to $|\,||x||_2 - ||y||_2|$ if $\angle(x, y) = 0$, and is equal to $||x||_2 + ||y||_2$, otherwise.

- **2*n*-gon metric**

  Given a centrally symmetric regular $2n$-gon K on the plane, the **2*n*-gon metric** is a metric on $\mathbb{R}^2$ defined, for any $x, y \in \mathbb{R}^2$, as the shortest Euclidean length of a polygonal line from x to y with each of its sides parallel to some edge of K.

  If K is a square with the vertices $\{(\pm 1, \pm 1)\}$, one obtains the **Manhattan metric**. The Manhattan metric arises also as the **Minkowskian metric** with the unit ball being the *diamond*, i.e., a square with the vertices $\{(1, 0), (0, 1), (-1, 0), (0, -1)\}$.

- **A-distance**

  Given a set A, $|A| \geq 2$, of distinct *orientations* (i.e., angles with fixed x axis) on the plane $\mathbb{R}^2$, the ***A*-distance** (or **fixed orientation metric**) is (Widmayer–Wu–Wong, 1987) Euclidean length of the shortest (zig-zag) path of line segments with orientations from A. Any A-distance is a metric.

  A-distance with $A = \{\frac{i\pi}{n} : 1 \leq i \leq n\}$ for fixed $n \in [2, \infty]$, is called a *uniform orientation metric*; cf. **2*n*-gon metric**. It is the l_1-metric, **hexagonal metric**, l_2-metric for $n = 2, 3, \infty$, respectively.

- **Central Park metric**

  The **Central Park metric** is a metric on $\mathbb{R}^2$, defined as the length of a shortest L_1-path (*Manhattan path*) between two points $x, y \in \mathbb{R}^2$ in the presence of a given set of areas which are traversed by a shortest Euclidean path (for example, Central Park in Manhattan).

- **Collision avoidance distance**

  Let $\mathcal{O} = \{O_1, \dots, O_m\}$ be a collection of pairwise disjoint polygons on the Euclidean plane representing a set of obstacles which are neither transparent nor traversable.

  The **collision avoidance distance** (or **piano movers distance**, **shortest path metric with obstacles**) is a metric on the set $\mathbb{R}^2 \backslash \{\mathcal{O}\}$, defined, for any $x, y \in \mathbb{R}^2 \backslash \{\mathcal{O}\}$, as the length of the shortest path among all possible continuous paths, connecting x and y, that do not intersect obstacles $O_i \backslash \partial O_i$ (a path can pass through points on the boundary ∂O_i of O_i), $i = 1, \dots m$.

- **Rectilinear distance with barriers**

  Let $\mathcal{O} = \{O_1, \dots, O_m\}$ be a set of pairwise disjoint open polygonal barriers on $\mathbb{R}^2$. A *rectilinear path* (or *Manhattan path*) P_{xy} from x to y is a collection of horizontal and vertical segments in the plane, joining x and y. The path P_{xy} is called *feasible* if $P_{xy} \cap (\cup_{i=1}^{m} B_i) = \emptyset$.

  The **rectilinear distance with barriers** (or *rectilinear distance in the presence of barriers*) is a metric on $\mathbb{R}^2 \backslash \{\mathcal{O}\}$, defined, for any $x, y \in \mathbb{R}^2 \backslash \{\mathcal{O}\}$, as the length of the shortest *feasible rectilinear path* from x to y.

  The rectilinear distance in the presence of barriers is a restriction of the **Manhattan metric**, and usually it is considered on the set $\{q_1, \dots, q_n\} \subset \mathbb{R}^2$ of *n origin-destination points*: the problem to find such a path arises, for example, in Urban Transportation, or in Plant and Facility Layout (see, for example, [LaLi81]).

- **Link distance**

  Let $P \subset \mathbb{R}^2$. The **polygonal distance** (or **link distance** as defined by Suri, 1986) between any two points of P is the smallest number of edges of a polygonal path in P connecting them if such path exists and ∞, otherwise.

  If the path is restricted to be rectilinear, one obtains the *rectilinear link distance*. If each line segment of the path is parallel to one from a set A of fixed orientations, one obtains the *A-oriented link distance*; cf. **fixed orientation metric** above.

  If the turning points of the path are constrained to lie on the boundary of P, then the path is called *drp* (diffuse reflection path). The *drp-diameter* of P is the minimum number of *diffuse reflections* (segments in a drp) needed to illuminate any target point from any point light source inside P.

- **Facility layout distances**

  A *layout* is a partition of a rectangular plane region into smaller rectangles, called *departments*, by lines parallel to the sides of original rectangle. All interior vertices should be of degree 3, and some of them, at least one on the boundary of each department, are *doors*, i.e., input-output locations.

  The problem is to design a convenient notion of distance $d(x, y)$ between departments x and y which minimizes the *cost function* $\sum_{x,y} F(x, y)d(x, y)$, where $F(x, y)$ is some *material flow* between x and y. The main distances used are:

  the **centroid distance**, i.e., the shortest Euclidean or **Manhattan distance** between *centroids* (the intersections of the diagonals) of x and y;

  the **perimeter distance**, i.e., the shortest rectilinear distance between doors of x and y, but going only along the *walls* (department perimeters).

- **Quickest path metric**

  A **quickest path metric** (or **network metric**, **time metric**) is a metric on $\mathbb{R}^2$ (or on a subset of $\mathbb{R}^2$) in the presence of a given *transportation network*, i.e., a finite graph $G = (V, E)$ with $V \subset \mathbb{R}^2$ and edge-weight function $w(e) > 1$: the vertices and edges are *stations* and *roads*. For any $x, y \in \mathbb{R}^2$, it is the time needed for a *quickest path* (i.e., a path minimizing the travel duration) between them when using, eventually, the network.

  Movement takes place, either off the network with unit speed, or along its roads $e \in E$ with fixed speeds $w(e) >> 1$, with respect to a given (usually, Euclidean or **Manhattan**) metric d on the plane. The network G can be accessed or exited only at stations (usual discrete model) or at any point of roads (the continuous model).

  The **heavy luggage metric** (Abellanas–Hurtado–Palop, 2005) is a quickest path metric on $\mathbb{R}^2$ in the presence of a network with speed 1 outside of the network and speed ∞ (so, travel time 0) inside of it.

  The **airlift metric** is a quickest path metric on $\mathbb{R}^2$ in the presence of an *airports network*, i.e., a planar graph $G = (V, E)$ on n vertices (*airports*) with positive edge weights $(w_e)_{e \in E}$ (*flight durations*). The graph may be entered and

[redacted]

exited only at the airports. Movement off the network takes place with unit speed with respect to the Euclidean metric. We assume that going by car takes time equal to the Euclidean distance d, whereas the flight along an edge $e = uv$ of G takes time $w(e) < d(u, v)$. In the simplest case, when there is an airlift between two points $a, b \in \mathbb{R}^2$, the distance between x and y is equal to

$$\min\{d(x,y), d(x,a) + w + d(b,y), d(x,b) + w + d(a,y)\},$$

where w is the flight duration from a to b.

The **city metric** is a quickest path metric on $\mathbb{R}^2$ in the presence of a *city public transportation network*, i.e., a planar straight line graph G with horizontal or vertical edges. G may be composed of many connected components, and may contain cycles.

One can enter/exit G at any point, be it at a vertex or on an edge (it is possible to postulate fixed entry points, too). Once having accessed G, one travels at fixed speed $v > 1$ in one of the available directions. Movement off the network takes place with unit speed with respect to the **Manhattan metric**, as in a large modern-style city with streets arranged in north–south and east–west directions.

A variant of such semimetric is the **subway semimetric** defined ([O'Bri03]), for $x, y \in \mathbb{R}^2$, as $\min(d(x,y), d(x,L) + d(y,L))$, where d is the Manhattan metric and L is a (*subway*) line.

- **Shantaram metric**

  For any numbers a, b with $0 < b \leq 2a \leq 2b$, the **Shantaram metric** between two points $x, y \in \mathbb{R}^2$ is 0, a or b if x and y coincide in exactly 2, 1 or 0 coordinates, respectively.

- **Periodic metric**

  A metric d on $\mathbb{R}^2$ is called **periodic** if there exist two linearly independent vectors v and u such that the *translation* by any vector $w = mv + nu$, $m, n \in \mathbb{Z}$, preserves distances, i.e., $d(x,y) = d(x+w, y+w)$ for any $x, y \in \mathbb{R}^2$.

  Cf. **translation invariant metric** in Chap. 5.

- **Nice metric**

  A metric d on $\mathbb{R}^2$ with the following properties is called **nice** (Klein–Wood, 1989):

1. d induces the Euclidean topology;
2. The d-circles are bounded with respect to the Euclidean metric;
3. If $x, y \in \mathbb{R}^2$ and $x \neq y$, then there exists a point $z, z \neq x, z \neq y$, such that $d(x,y) = d(x,z) + d(z,y)$;
4. If $x, y \in \mathbb{R}^2$, $x \prec y$ (where $\prec$ is a fixed order on $\mathbb{R}^2$, the lexicographic order, for example), $C(x,y) = \{z \in \mathbb{R}^2 : d(x,z) \leq d(y,z)\}$, $D(x,y) = \{z \in \mathbb{R}^2 : d(x,z) < d(y,z)\}$, and $\overline{D(x,y)}$ is the closure of $D(x,y)$, then $J(x,y) = C(x,y) \cap \overline{D(x,y)}$ is a curve homeomorphic to $(0,1)$. The intersection of two such curves consists of finitely many connected components.

Every **norm metric** fulfills 1, 2, and 3 Property 2 means that the metric d is continuous at infinity with respect to the Euclidean metric. Property 4 is to ensure that the boundaries of the correspondent *Voronoi diagrams* are curves, and that not too many intersections exist in a neighborhood of a point, or at infinity.

A nice metric d has a nice Voronoi diagram: in the Voronoi diagram $V(P, d, \mathbb{R}^2)$ (where $P = \{p_1, \dots, p_k\}$, $k \geq 2$, is the set of *generator points*) each *Voronoi region* $V(p_i)$ is a path-connected set with a nonempty interior, and the system $\{V(p_1), \dots, V(p_k)\}$ forms a *partition* of the plane.

- **Contact quasi-distances**

  The **contact quasi-distances** are the following variations of the **distance convex function** (Chap. 1) defined on $\mathbb{R}^2$ (in general, on $\mathbb{R}^n$) for any $x, y \in \mathbb{R}^2$.

  Given a set $B \subset \mathbb{R}^2$, the **first contact quasi-distance** d_B is defined by

  $$\inf\{\alpha > 0 : y - x \in \alpha B\}$$

  (cf. **sensor network distances** in Chap. 29).

  Given, moreover, a point $b \in B$ and a set $A \subset \mathbb{R}^2$, the **linear contact quasi-distance** is a **point-set distance** defined by $d_b(x, A) = \inf\{\alpha \geq 0 : \alpha b + x \in A\}$.

  The **intercept quasi-distance** is, for a finite set B, defined by $\frac{\sum_{b \in B} d_b(x,y)}{|B|}$.

- **Radar discrimination distance**

  The **radar discrimination distance** is a distance on $\mathbb{R}^2$ defined by

  $$|\rho_x - \rho_y + \theta_{xy}|$$

  if $x, y \in \mathbb{R}^2 \backslash \{0\}$, and by

  $$|\rho_x - \rho_y|$$

  if $x = 0$ or $y = 0$, where, for each $x \in \mathbb{R}^2$, ρ_x denotes the radial distance of x from $\{0\}$ and, for any $x, y \in R^2 \backslash \{0\}$, θ_{xy} denotes the radian angle between them.

- **Ehrenfeucht–Haussler semimetric**

  Let S be a subset of $\mathbb{R}^2$ such that $x_1 \geq x_2 - 1 \geq 0$ for any $x = (x_1, x_2) \in S$.

  The **Ehrenfeucht–Haussler semimetric** (see [EhHa88]) on S is defined by

  $$\log_2 \left(\left(\frac{x_1}{y_2} + 1 \right) \left(\frac{y_1}{x_2} + 1 \right) \right).$$

- **Toroidal metric**

  The **toroidal metric** is a metric on $T = [0, 1) \times [0, 1) = \{x = (x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_1, x_2 < 1\}$ defined for any $x, y \in \mathbb{R}^2$ by

  $$\sqrt{t_1^2 + t_2^2},$$

  where $t_i = \min\{|x_i - y_i|, |x_i - y_i + 1|\}$ for $i = 1, 2$ (cf. **torus metric**).

- **Circle metric**

  The **circle metric** is the **intrinsic metric** on the *unit circle* S^1 in the plane.

[redacted]

As $S^1 = \{(x,y) : x^2 + y^2 = 1\} = \{e^{i\theta} : 0 \leq \theta < 2\pi\}$, it is the length of the shorter of the two arcs joining the points $e^{i\theta}, e^{i\vartheta} \in S^1$, and can be written as

$$\min\{|\theta - \vartheta|, 2\pi - |\theta - \vartheta|\} = \begin{cases} |\vartheta - \theta|, & \text{if } 0 \leq |\vartheta - \theta| \leq \pi, \\ 2\pi - |\vartheta - \theta|, & \text{if } \quad |\vartheta - \theta| > \pi. \end{cases}$$

- **Metric between angles**

  The **metric between angles** Δ is a metric on the set of all angles in the plane $\mathbb{R}^2$ defined for any $\theta, \vartheta \in [0, 2\pi)$ (cf. **circle metric**) by

$$\min\{|\theta - \vartheta|, 2\pi - |\theta - \vartheta|\} = \begin{cases} |\vartheta - \theta|, & \text{if } 0 \leq |\vartheta - \theta| \leq \pi, \\ 2\pi - |\vartheta - \theta|, & \text{if } \quad |\vartheta - \theta| > \pi. \end{cases}$$

- **Metric between directions**

  On $\mathbb{R}^2$, a *direction* $\hat{l}$ is a class of all straight lines which are parallel to a given straight line $l \subset \mathbb{R}^2$. The **metric between directions** is a metric on the set $\mathcal{L}$ of all directions on the plane defined, for any directions $\hat{l}, \hat{m} \in \mathcal{L}$, as the angle between any two representatives.
- **Angular distance**

  The **angular distance** traveled around a circle is the number $\theta = \frac{l}{r}$ of radians the path subtends, $\theta = \frac{l}{r}$, where l is the path length, and r is the circle's radius.
- **Circular distance**

  The **circular distance** is the distance traveled by a wheel. Each revolution of a wheel with radius r is equivalent to the distance of $2\pi r$ radians.
- **Circular-railroad quasi-metric**

  The **circular-railroad quasi-metric** on the *unit circle* $S^1 \subset \mathbb{R}^2$ is defined, for any $x, y \in S^1$, as the length of the counterclockwise circular arc from x to y in S^1.
- **Inversive distance**

  The **inversive distance** between two nonintersecting circles in the plane $\mathbb{R}^2$ is defined as the natural logarithm of the ratio of the radii (the larger to the smaller) of two concentric circles into which the given circles can be inverted.

  Let c be the distance between the centers of two nonintersecting circles of radii a and $b < a$. Then their inversive distance is given by

$$\cosh^{-1}\left|\frac{a^2 + b^2 - c^2}{2ab}\right|.$$

The *circumcircle* and *incircle* of a triangle with *circumradius* R and *inradius* r are at the inversive distance $2\sinh^{-1}(\frac{1}{2}\sqrt{\frac{r}{R}})$.

Given three noncollinear points, construct three tangent circles such that one is centered at each point and the circles are pairwise tangent to one another. Then there exist exactly two nonintersecting circles, called the *Soddy circles*, that are tangent to all three circles. Their inversive distance is $2\cosh^{-1} 2$.

19.2 Digital Metrics

Here we list special metrics which are used in *Computer Vision* (or *Pattern Recognition*, *Robot Vision*, *Digital Geometry*).

A *computer picture* (or *computer image*) is a subset of $\mathbb{Z}^n$ which is called a *digital nD space*. Usually, pictures are represented in the *digital plane* (or *image plane*) $\mathbb{Z}^2$, or in the *digital space* (or *image space*) $\mathbb{Z}^3$. The points of $\mathbb{Z}^2$ and $\mathbb{Z}^3$ are called *pixels* and *voxels*, respectively. An *nD m-quantized space* is a scaling $\frac{1}{m}\mathbb{Z}^n$.

A **digital metric** (see, for example, [RoPf68]) is any metric on a digital *nD* space. Usually, it should take integer values.

The metrics on $\mathbb{Z}^n$ that are mainly used are the L_1- and L_∞*-metrics*, as well as the L_2*-metric* after rounding to the nearest greater (or lesser) integer. In general, a given list of *neighbors* of a pixel can be seen as a list of permitted *one-step moves* on $\mathbb{Z}^2$. Let us associate a **prime distance**, i.e., a positive weight, to each type of such move.

Many digital metrics can be obtained now as the minimum, over all admissible paths (i.e., sequences of permitted moves), of the sum of corresponding prime distances.

In practice, the subset $(\mathbb{Z}_m)^n = \{0, 1, \dots, m-1\}^n$ is considered instead of the full space $\mathbb{Z}^n$. $(\mathbb{Z}_m)^2$ and $(\mathbb{Z}_m)^3$ are called the *m-grill* and *m-framework*, respectively. The most used metrics on $(\mathbb{Z}_m)^n$ are the **Hamming metric** and the **Lee metric**.

- **Grid metric**

  The **grid metric** is the L_1*-metric* on $\mathbb{Z}^n$. It can be seen as the path metric of an infinite graph: two points of $\mathbb{Z}^n$ are adjacent if their L_1-distance is 1.

  For $n = 2$, this metric is the restriction on $\mathbb{Z}^2$ of the **city-block metric** which is also called the **taxicab** (or **rectilinear**, **Manhattan**, 4-) **metric**.
- **Lattice metric**

  The **lattice metric** is the L_∞*-metric* on $\mathbb{Z}^n$. It can be seen as the path metric of an infinite graph: two points of $\mathbb{Z}^n$ are adjacent if their L_∞-distance is 1. For $\mathbb{Z}^2$, the adjacency corresponds to the king move in chessboard terms, and this graph is called the L_∞*-grid*, while this metric is also called the **chessboard metric**, **king-move metric**, 8**-metric**, or **checking distance**.

  This metric is the restriction on $\mathbb{Z}^n$ of the **Chebyshev metric** which is also called the **sup metric**, or **uniform metric**.
- **Hexagonal metric**

  The **hexagonal metric** (or 6**-metric**) is a metric on $\mathbb{Z}^2$ with a *unit sphere* (centered at $x \in \mathbb{Z}^2$) defined by $S^1(x) = S^1_{L_1}(x) \cup \{(x_1-1, x_2-1), (x_1-1, x_2+1)\}$ for even x_2, and $S^1(x) = S^1_{L_1}(x) \cup \{(x_1+1, x_2-1), (x_1+1, x_2+1)\}$ for odd x_2. For any $x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{Z}^2$, this metric $d_6(x, y)$ can be written as

  $$\max\left\{|x_2 - y_2|, \frac{|x_2 - y_2|}{2} \pm \left(\frac{x_2 - y_2}{2} + \left\lfloor \frac{x_2 + 1}{2} \right\rfloor - \left\lfloor \frac{y_2 + 1}{2} \right\rfloor - (x_1 - y_1)\right)\right\}.$$

  It is the path metric of the *triangular grid* (or, dually, the minimum number of cell moves of the *hexagonal grid*) on the plane. In *hexagonal coordinates* (h_1, h_2)

[redacted]

(in which the h_1- and h_2-axes are parallel to the grid's edges) the hexagonal distance between points $h = (h_1, h_2)$ and $i = (i_1, i_2)$ is

$$d_6(h, i) = \frac{1}{2}(|h_1 - i_1| + |h_2 - i_2| + |(h_1 - i_1) - (h_2 - i_2)|),$$

i.e., $|h_1 - i_1| + |h_2 - i_2|$, if $(h_1 - i_1)(h_2 - i_2) \leq 0$, and $\max\{|h_1 - i_1|, |h_2 - i_2|\}$, if otherwise; cf. [LuRo76]. The coordinates (h_1, h_2) of a point x are related to its Cartesian coordinates (x_1, x_2) by $h_1 = x_1 - \lfloor \frac{x_2}{2} \rfloor$, $h_2 = x_2$.

This metric approximates the Euclidean metric better than L_1- or L_∞*-metric*.

The **hexagonal Hausdorff metric** is a metric on the set of all bounded subsets (*pictures*, or *images*) of the hexagonal grid on the plane defined by

$$\inf\{p, q : A \subset B + qH,\ B \subset A + pH\}$$

for any pictures A and B, where pH is the *regular hexagon of size p* (i.e., with $p+1$ pixels on each edge), centered at the origin and including its interior, and $+$ is the *Minkowski addition*: $A + B = \{x + y : x \in A, y \in B\}$ (cf. **Pompeiu–Hausdorff–Blaschke metric** in Chap. 9). If A is a pixel x, then the distance between x and B is equal to $\sup_{y \in B} d_6(x, y)$, where d_6 is the hexagonal metric.

- **Digital volume metric**

  The **digital volume metric** is a metric on the set K of all bounded subsets (*pictures*, or *images*) of $\mathbb{Z}^2$ (in general, of $\mathbb{Z}^n$) defined by

  $$vol(A \triangle B),$$

  where $vol(A) = |A|$, i.e., the number of pixels contained in A, and $A \triangle B$ is the *symmetric difference* between sets A and B.

  This metric is a digital analog of the **Nikodym metric** in Chap. 9.

- **Neighborhood sequence metric**

  On the digital plane $\mathbb{Z}^2$, consider two types of motions: the *city-block motion*, restricting movements only to the horizontal or vertical directions, and the *chessboard motion*, also allowing diagonal movements.

  The use of both these motions is determined by a *neighborhood sequence* $B = \{b(1), b(2), \dots, b(l)\}$, where $b(i) \in \{1, 2\}$ is a particular type of neighborhood, with $b(i) = 1$ signifying unit change in 1 coordinate (*city-block neighborhood*), and $b(i) = 2$ meaning unit change also in 2 coordinates (*chessboard neighborhood*). The sequence B defines the type of motion to be used at every step (see [Das90]).

  The **neighborhood sequence metric** is a metric on $\mathbb{Z}^2$ defined as the length of a shortest path between x and $y \in \mathbb{Z}^2$, determined by a given neighborhood sequence B. It can be written as

  $$\max\{d_B^1(u), d_B^2(u)\},$$

where $u_1 = x_1 - y_1$, $u_2 = x_2 - y_2$, $d_B^1(u) = \max\{|u_1|, |u_2|\}$, $d_B^2(u) = \sum_{j=1}^{l} \lfloor \frac{|u_1|+|u_2|+g(j)}{f(l)} \rfloor$, $f(0) = 0$, $f(i) = \sum_{j=1}^{i} b(j)$, $1 \le i \le l$, $g(j) = f(l) - f(j-1) - 1$, $1 \le j \le l$.

For $B = \{1\}$ one obtains the **city-block metric**, for $B = \{2\}$ one obtains the **chessboard metric**. The case $B = \{1, 2\}$, i.e., the alternative use of these motions, results in the **octagonal metric**, introduced in [RoPf68].

A proper selection of the B-sequence can make the corresponding metric very close to the Euclidean metric. It is always greater than the chessboard metric, but smaller than the city-block metric.

- ***nD*-neighborhood sequence metric**

  The ***nD*-neighborhood sequence metric** is a metric on $\mathbb{Z}^n$, defined as the length of a shortest path between x and $y \in \mathbb{Z}^n$, determined by a given *nD-neighborhood sequence B* (see [Faze99]).

  Formally, two points $x, y \in \mathbb{Z}^n$ are called *m-neighbors*, $0 \le m \le n$, if $0 \le |x_i - y_i| \le 1$, $1 \le i \le n$, and $\sum_{i=1}^{n} |x_i - y_i| \le m$. A finite sequence $B = \{b(1), \dots, b(l)\}$, $b(i) \in \{1, 2, \dots, n\}$, is called an *nD-neighborhood sequence with period l*. For any $x, y \in \mathbb{Z}^n$, a point sequence $x = x^0, x^1, \dots, x^k = y$, where x^i and x^{i+1}, $0 \le i \le k-1$, are r-neighbors, $r = b((i \bmod l) + 1)$, is called a *path from x to y determined by B* with length k. The distance between x and y can be written as

$$\max_{1 \le i \le n} d_i(u) \quad \text{with} \quad d_i(x, y) = \sum_{j=1}^{l} \left\lfloor \frac{a_i + g_i(j)}{f_i(l)} \right\rfloor ,$$

  where $u = (|u_1|, |u_2|, \dots, |u_n|)$ is the nonincreasing ordering of $|u_m|$, $u_m = x_m - y_m$, $m = 1, \dots, n$, that is, $|u_i| \le |u_j|$ if $i < j$; $a_i = \sum_{j=1}^{n-i+1} u_j$; $b_i(j) = b(j)$ if $b(j) < n - i + 2$, and is $n - i + 1$, otherwise; $f_i(j) = \sum_{k=1}^{j} b_i(k)$ if $1 \le j \le l$, and is 0 if $j = 0$; $g_i(j) = f_i(l) - f_i(j-1) - 1$, $1 \le j \le l$.

- **Strand–Nagy distances**

  The *face-centered cubic lattice* is $A_3 = \{(a_1, a_2, a_3) \in \mathbb{Z}^3 : a_1 + a_2 + a_3 \equiv 0 (\bmod 2)\}$, and the *body-centered cubic lattice* is its dual

$$A_3^* = \{(a_1, a_2, a_3) \in \mathbb{Z}^3 : a_1 \equiv a_2 \equiv a_3 (\bmod 2)\}.$$

  Let $L \in \{A_3, A_3^*\}$. For any points $x, y \in L$, let $d_1(x, y) = \sum_{j=1}^{3} |x_j - y_j|$ denote the L_1-metric and $d_\infty(x, y) = \max_{j \in \{1,2,3\}} |x_j - y_j|$ denote the L_∞-metric between them. Two points $x, y \in L$ are called *1-neighbors* if $d_1(x, y) \le 3$ and $0 < d_\infty(x, y) \le 1$; they are called *2-neighbors* if $d_1(x, y) \le 3$ and $1 < d_\infty(x, y) \le 2$.

  Given a sequence $B = \{b(i)\}_{i=1}^{\infty}$ over the alphabet $\{1, 2\}$, a *B-path* in L is a point sequence $x = x^0, x^1, \dots, x^k = y$, where x^i and x^{i+1}, $0 \le i \le k-1$, are 1-neighbors if $b(i) = 1$ and 2-neighbors if $b(i) = 2$.

  The **Strand–Nagy distance** between two points $x, y \in L$ (or *B-distance* in Strand and Nagy, 2007) is the length of a shortest B-path between them. For

[redacted]

$L = A_3$, it is

$$\min\{k : k \geq \max\{\frac{d_1(x,y)}{2}, d_\infty(x,y) - |\{1 \leq i \leq k : b(i) = 2\}|\}.$$

The Strand–Nagy distance is a metric, for example, for the periodic sequence $B = (1,2,1,2,1,2,\dots)$ but not for the periodic sequence $B = (2,1,2,1,2,1,\dots)$.

- **Path-generated metric**

  Consider the l_∞-*grid*, i.e., the graph with the vertex-set $\mathbb{Z}^2$, and two vertices being *neighbors* if their l_∞-distance is 1. Let $\mathcal{P}$ be a collection of paths in the l_∞-grid such that, for any $x, y \in \mathbb{Z}^2$, there exists at least one path from $\mathcal{P}$ between x and y, and if $\mathcal{P}$ contains a path Q, then it also contains every path contained in Q.

  Let $d_{\mathcal{P}}(x,y)$ be the length of the shortest path from $\mathcal{P}$ between x and $y \in \mathbb{Z}^2$. If $d_{\mathcal{P}}$ is a metric on $\mathbb{Z}^2$, then it is called a **path-generated metric** (see [Melt91]).

  Let G be one of the sets: $G_1 = \{\uparrow, \rightarrow\}$, $G_{2A} = \{\uparrow, \nearrow\}$, $G_{2B} = \{\uparrow, \nwarrow\}$, $G_{2C} = \{\nearrow, \nwarrow\}$, $G_{2D} = \{\rightarrow, \nwarrow\}$, $G_{3A} = \{\rightarrow, \uparrow, \nearrow\}$, $G_{3B}\{\rightarrow, \uparrow, \nwarrow\}$, $G_{4A} = \{\rightarrow, \nearrow, \nwarrow\}$, $G_{4B}\{\uparrow, \nearrow, \nwarrow, \}$, $G_5 = \{\rightarrow, \uparrow, \nearrow, \nwarrow\}$. Let $\mathcal{P}(G)$ be the set of paths which are obtained by concatenation of paths in G and the corresponding paths in the opposite directions. Any path-generated metric coincides with one of the metrics $d_{\mathcal{P}(G)}$. Moreover, one can obtain the following formulas:

  1. $d_{\mathcal{P}(G_1)}(x,y) = |u_1| + |u_2|$;
  2. $d_{\mathcal{P}(G_{2A})}(x,y) = \max\{|2u_1 - u_2|, |u_2|\}$;
  3. $d_{\mathcal{P}(G_{2B})}(x,y) = \max\{|2u_1 + u_2|, |u_2|\}$;
  4. $d_{\mathcal{P}(G_{2C})}(x,y) = \max\{|2u_2 + u_1|, |u_1|\}$;
  5. $d_{\mathcal{P}(G_{2D})}(x,y) = \max\{|2u_2 - u_1|, |u_1|\}$;
  6. $d_{\mathcal{P}(G_{3A})}(x,y) = \max\{|u_1|, |u_2|, |u_1 - u_2|\}$;
  7. $d_{\mathcal{P}(G_{3B})}(x,y) = \max\{|u_1|, |u_2|, |u_1 + u_2|\}$;
  8. $d_{\mathcal{P}(G_{4A})}(x,y) = \max\{2\lceil(|u_1| - |u_2|)/2\rceil, 0\} + |u_2|$;
  9. $d_{\mathcal{P}(G_{4B})}(x,y) = \max\{2\lceil(|u_2| - |u_1|)/2\rceil, 0\} + |u_1|$;
  10. $d_{\mathcal{P}(G_5)}(x,y) = \max\{|u_1|, |u_2|\}$,

  where $u_1 = x_1 - y_1$, $u_2 = x_2 - y_2$, and $\lceil . \rceil$ is the *ceiling function*: for any real x the number $\lceil x \rceil$ is the least integer greater than or equal to x.

  The metric spaces obtained from G-sets with the same numerical index are isometric. $d_{\mathcal{P}(G_1)}$ is the **city-block metric**, and $d_{\mathcal{P}(G_5)}$ is the **chessboard metric**.

- **Chamfer metric**

  Given numbers α, β with $0 < \alpha \leq \beta < 2\alpha$, the (α, β)-*weighted* l_∞-*grid* is the graph with the vertex-set $\mathbb{Z}^2$, two vertices being adjacent if their l_∞-distance is one, while horizontal/vertical and diagonal edges have *weights* α and β, respectively.

  A **chamfer metric** (or (α, β)-*chamfer metric*, [Borg86]) is the weighted path metric in this graph. For any $x, y \in \mathbb{Z}^2$ it can be written as

  $$\beta m + \alpha(M - m),$$

  where $M = \max\{|u_1|, |u_2|\}$, $m = \min\{|u_1|, |u_2|\}$, $u_1 = x_1 - y_1$, $u_2 = x_2 - y_2$.

If the weights α and β are equal to the Euclidean lengths 1, $\sqrt{2}$ of horizontal/vertical and diagonal edges, respectively, then one obtains the Euclidean length of the shortest chessboard path between x and y. If $\alpha = \beta = 1$, one obtains the **chessboard metric**. The $(3, 4)$-chamfer metric is the most used one for digital images.

A 3*D*-**chamfer metric** is the weighted path metric of the graph with the vertex-set $\mathbb{Z}^3$ of *voxels*, two voxels being adjacent if their l_∞-distance is one, while weights α, β, and γ are associated, respectively, to the distance from 6 face neighbors, 12 edge neighbors, and 8 corner neighbors.

- **Weighted cut metric**

  Consider the *weighted* l_∞*-grid*, i.e., the graph with the vertex-set $\mathbb{Z}^2$, two vertices being adjacent if their l_∞-distance is one, and each edge having some positive *weight* (or *cost*). The usual **weighted path metric** between two pixels is the minimal cost of a path connecting them. The **weighted cut metric** between two pixels is the minimal cost (defined now as the sum of costs of crossed edges) of a *cut*, i.e., a plane curve connecting them while avoiding pixels.

- **Knight metric**

  The **knight** (or *octogonal*) **metric** on $\mathbb{Z}^2$ is the minimum number of moves a chess knight would take to travel from x to $y \in \mathbb{Z}^2$. Its *unit sphere* S^1_{knight}, centered at the origin, contains exactly 8 integral points $\{(\pm 2, \pm 1), (\pm 1, \pm 2)\}$, and can be written as $S^1_{knight} = S^3_{L_1} \cap S^2_{l_\infty}$, where $S^3_{L_1}$ is the L_1-sphere of radius 3, and $S^2_{L_\infty}$ is the L_∞-sphere of radius 2, both centered at the origin (see [DaCh88]).

  The distance between x and y is 3 if $(M, m) = (1, 0)$, is 4 if $(M, m) = (2, 2)$ and is equal to $\max\{\lceil \frac{M}{2} \rceil, \lceil \frac{M+m}{3} \rceil\} + (M + m) - \max\{\lceil \frac{M}{2} \rceil, \lceil \frac{M+m}{3} \rceil\} \pmod 2$, otherwise, where $M = \max\{|u_1|, |u_2|\}$, $m = \min\{|u_1|, |u_2|\}$, $u_1 = x_1 - y_1$, $u_2 = x_2 - y_2$.

- **Super-knight metric**

  Let $p, q \in \mathbb{N}$. A (p, q)*-super-knight* (or (p, q)*-leaper*, (p, q)*-spider*) is a (variant) chess piece whose move consists of a leap p squares in one orthogonal direction followed by a 90° direction change, and q squares leap to the destination square. Rook, bishop and queen have $q = 0$, $q = p$ and $q = 0, p$, respectively.

  Chess-variant terms exist for a $(p, 1)$-leaper with $p = 0, 1, 2, 3, 4$ (*Wazir*, *Ferz*, usual *Knight*, *Camel*, *Giraffe*), and for a $(p, 2)$-leaper with $p = 0, 1, 2, 3$ (*Dabbaba*, usual *Knight*, *Alfil*, *Zebra*).

  A (p, q)**-super-knight metric** (or (p, q)*-leaper metric*) is a metric on $\mathbb{Z}^2$ defined as the minimum number of moves a chess (p, q)-super-knight would take to travel from x to $y \in \mathbb{Z}^2$. Thus, its *unit sphere* $S^1_{p,q}$, centered at the origin, contains exactly 8 integral points $\{(\pm p, \pm q), (\pm q, \pm p)\}$. (See [DaMu90].)

  The **knight metric** is the $(1, 2)$-super-knight metric. The **city-block metric** can be considered as the *Wazir metric*, i.e., (0, 1)-super-knight metric.

- **Rook metric**

  The **rook metric** is a metric on $\mathbb{Z}^2$ defined as the minimum number of moves a chess rook would take to travel from x to $y \in \mathbb{Z}^2$. This metric can take only the values $\{0, 1, 2\}$, and coincides with the **Hamming metric** on $\mathbb{Z}^2$.

[redacted]

- **Chess programming distances**

  On a chessboard $\mathbb{Z}_8^2$, *files* are 8 columns labeled from *a* to *h* and *ranks* are 8 rows labeled from 1 to 8. Given two squares, their **file-distance** and **rank-distance** are the absolute differences between the 0–7 indices of their files or, respectively, ranks. The **Chebyshev distance** and **Manhattan distance** are the maximum or, respectively, the sum of their file-distance and rank-distance.

  The *center distance* and *corner distance* of a square are its (Chebyshev or Manhattan) distance to closest square among $\{d4, d5, e4, e5\}$ or, respectively, closest corner. For example, the program *Chess 4.x* uses in endgame evaluation $4.7d + 1.6(14 - d')$, where d is the center Manhattan distance of losing king and d' is the Manhattan distance between kings.

  Two kings at rank- and file-distances d_r, d_f, are in *opposition*, which is *direct*, or *diagonal*, or *distant* if $(d_r, d_f) \in \{(0,2), (2,0)\}$, or $= (2,2)$, or their Manhattan distance is even ≥ 6 and no pawns interfere between them.

  Unrelated *cavalry file distance* is the number of files in which it rides.

Chapter 20
Voronoi Diagram Distances

Given a finite set A of objects A_i in a space S, computing the *Voronoi diagram* of A means partitioning the space S into *Voronoi regions* $V(A_i)$ in such a way that $V(A_i)$ contains all points of S that are "closer" to A_i than to any other object A_j in A.

Given a *generator set* $P = \{p_1, \dots p_k\}$, $k \geq 2$, of distinct points (*generators*) from $\mathbb{R}^n$, $n \geq 2$, the ordinary **Voronoi polytope** $V(p_i)$ associated with a generator p_i is defined by

$$V(p_i) = \{x \in \mathbb{R}^n : d_E(x, p_i) \leq d_E(x, p_j) \text{ for any } j \neq i\},$$

where d_E is the Euclidean distance on $\mathbb{R}^n$. The set

$$V(P, d_E, \mathbb{R}^n) = \{V(p_1), \dots, V(p_k)\}$$

is called the *n-dimensional ordinary Voronoi diagram, generated by P*.

The boundaries of (n-dimensional) Voronoi polytopes are called (($n-1$)-dimensional) *Voronoi facets*, the boundaries of Voronoi facets are called ($n-2$)*-dimensional Voronoi faces*, …, the boundaries of 2D Voronoi faces are called *Voronoi edges*, and the boundaries of Voronoi edges are called *Voronoi vertices*.

The ordinary Voronoi diagram can be generalized in the following three ways:

1. The generalization with respect to the generator set $A = \{A_1, \dots, A_k\}$ which can be a set of lines, a set of areas, etc.;
2. The generalization with respect to the space S which can be a sphere (*spherical Voronoi diagram*), a cylinder (*cylindrical Voronoi diagram*), a cone (*conic Voronoi diagram*), a polyhedral surface (*polyhedral Voronoi diagram*), etc.;
3. The generalization with respect to the function d, where $d(x, A_i)$ measures the "distance" from a point $x \in S$ to a generator $A_i \in A$.

This generalized distance function d is called the **Voronoi generation distance** (or *Voronoi distance*, *V-distance*), and allows many more functions than the

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_20

[redacted]

Euclidean metric on S. If F is a strictly increasing function of a V-distance d, i.e., $F(d(x,A_i)) \leq F(d(x,A_j))$ if and only if $d(x,A_i) \leq d(x,A_j)$, then the generalized Voronoi diagrams $V(A,F(d),S)$ and $V(A,d,S)$ coincide, and one says that the V-distance $F(d)$ is *transformable* to the V-distance d, and that the generalized Voronoi diagram $V(A,F(d),S)$ is a *trivial generalization* of the generalized Voronoi diagram $V(A,d,S)$.

In applications, one often uses for trivial generalizations of the ordinary Voronoi diagram $V(P,d,\mathbb{R}^n)$ the **exponential distance**, the **logarithmic distance**, and the **power distance**. There are generalized Voronoi diagrams $V(P,D,\mathbb{R}^n)$, defined by V-distances, that are not transformable to the Euclidean distance d_E: the **multiplicatively weighted Voronoi distance**, the **additively weighted Voronoi distance**, etc.

The theory of generalized Voronoi diagrams $V(P,D,\mathbb{R}^n)$, where D is a norm metric $||x-p||$ collapses even for the case, when P is a lattice in $\mathbb{R}^n$. But [DeDu13] adapted it for *polyhedral*, i.e., with a polytopal unit ball, norms; $||.||_1$ and $||.||_\infty$ are among them.

For additional information see, for example, [OBS92, Klei89].

20.1 Classical Voronoi Generation Distances

- **Exponential distance**

  The **exponential distance** is the Voronoi generation distance

  $$D_{exp}(x,p_i) = e^{d_E(x,p_i)}$$

  for the trivial generalization $V(P,D_{exp},\mathbb{R}^n)$ of the ordinary Voronoi diagram $V(P,d_E,\mathbb{R}^n)$, where d_E is the Euclidean distance.
- **Logarithmic distance**

  The **logarithmic distance** is the Voronoi generation distance

  $$D_{\ln}(x,p_i) = \ln d_E(x,p_i)$$

  for the trivial generalization $V(P,D_{\ln},\mathbb{R}^n)$ of the ordinary Voronoi diagram $V(P,d_E,\mathbb{R}^n)$, where d_E is the Euclidean distance.
- **Power distance**

  The **power distance** is the Voronoi generation distance

  $$D_\alpha(x,p_i) = d_E(x,p_i)^\alpha, \ \alpha > 0,$$

  for the trivial generalization $V(P,D_\alpha,\mathbb{R}^n)$ of the ordinary Voronoi diagram $V(P,d_E,\mathbb{R}^n)$, where d_E is the Euclidean distance.

- **Multiplicatively weighted distance**

  The **multiplicatively weighted distance** d_{MW} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{MW}, \mathbb{R}^n)$ (*multiplicatively weighted Voronoi diagram*) defined by

$$d_{MW}(x, p_i) = \frac{1}{w_i} d_E(x, p_i)$$

  for any point $x \in \mathbb{R}^n$ and any *generator point* $p_i \in P = \{p_1, \dots, p_k\}$, $k \geq 2$, where $w_i \in w = \{w_i, \dots, w_k\}$ is a given positive *multiplicative weight* of the generator p_i, and d_E is the Euclidean distance.

  A *Möbius diagram* (Boissonnat–Karavelas, 2003) is a diagram the **midsets** (bisectors) of which are hyperspheres. It generalizes the Euclidean Voronoi and power diagrams, and it is equivalent to power diagrams in $\mathbb{R}^{n+1}$.

  For $\mathbb{R}^2$, the multiplicatively weighted Voronoi diagram is called a *circular Dirichlet tessellation*. An edge in this diagram is a circular arc or a straight line.

  In the plane $\mathbb{R}^2$, there exists a generalization of the multiplicatively weighted Voronoi diagram, the *crystal Voronoi diagram*, with the same definition of the distance (where w_i is the speed of growth of the crystal p_i), but a different partition of the plane, as the crystals can grow only in an empty area.

- **Additively weighted distance**

  The **additively weighted distance** d_{AW} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{AW}, \mathbb{R}^n)$ (*additively weighted Voronoi diagram*) defined by

$$d_{AW}(x, p_i) = d_E(x, p_i) - w_i$$

  for any point $x \in \mathbb{R}^n$ and any generator point $p_i \in P = \{p_1, \dots, p_k\}$, $k \geq 2$, where $w_i \in w = \{w_i, \dots, w_k\}$ is a given *additive weight* of the generator p_i, and d_E is the Euclidean distance.

  For $\mathbb{R}^2$, the additively weighted Voronoi diagram is called a *hyperbolic Dirichlet tessellation*. An edge in this diagram is a hyperbolic arc or a straight line segment.

- **Additively weighted power distance**

  The **additively weighted power distance** d_{PW} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{PW}, \mathbb{R}^n)$ (*additively weighted power Voronoi diagram*) defined by

$$d_{PW}(x, p_i) = d_E^2(x, p_i) - w_i$$

  for any point $x \in \mathbb{R}^n$ and any *generator point* $p_i \in P = \{p_1, \dots, p_k\}$, $k \geq 2$, where $w_i \in w = \{w_i, \dots, w_k\}$ is a given *additive weight* of the generator p_i, and d_E is the Euclidean distance.

  This diagram can be seen as a Voronoi diagram of circles or as a Voronoi diagram with the *Laguerre geometry*.

[redacted]

The **multiplicatively weighted power distance** $d_{MPW}(x,p_i) = \frac{1}{w_i}d_E^2(x,p_i)$, $w_i > 0$, is transformable to the **multiplicatively weighted distance**, and gives a trivial extension of the multiplicatively weighted Voronoi diagram.

- **Compoundly weighted distance**

  The **compoundly weighted distance** d_{CW} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{CW}, \mathbb{R}^n)$ (*compoundly weighted Voronoi diagram*) defined by

  $$d_{CW}(x,p_i) = \frac{1}{w_i}d_E(x,p_i) - v_i$$

  for any point $x \in \mathbb{R}^n$ and any generator point $p_i \in P = \{p_1,\dots,p_k\}$, $k \geq 2$, where $w_i \in w = \{w_i,\dots,w_k\}$ is a given positive *multiplicative weight* of the generator p_i, $v_i \in v = \{v_1,\dots,v_k\}$ is a given *additive weight* of the generator p_i, and d_E is the Euclidean distance.

  An edge in the 2D compoundly weighted Voronoi diagram is a part of a fourth-order polynomial curve, a hyperbolic arc, a circular arc, or a straight line.

20.2 Plane Voronoi Generation Distances

- **Shortest path distance with obstacles**

  Let $\mathcal{O} = \{O_1,\dots,O_m\}$ be a collection of pairwise disjoint polygons on the Euclidean plane, representing a set of obstacles which are neither transparent nor traversable.

  The **shortest path distance with obstacles** d_{sp} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{sp}, \mathbb{R}^2\backslash\{\mathcal{O}\})$ (*shortest path Voronoi diagram with obstacles*) defined, for any $x, y \in \mathbb{R}^2\backslash\{\mathcal{O}\}$, as the length of the shortest path among all possible continuous $(x-y)$-paths that do not intersect obstacles $O_i\backslash\partial O_i$ (a path can pass through points on the boundary ∂O_i of O_i), $i = 1,\dots m$.

  The shortest path is constructed with the aid of the *visibility polygon* and the *visibility graph* of $V(P, d_{sp}, \mathbb{R}^2\backslash\{\mathcal{O}\})$.

- **Visibility shortest path distance**

  Let $\mathcal{O} = \{O_1,\dots,O_m\}$ be a collection of pairwise disjoint line segments $O_l = [a_l, b_l]$ in the Euclidean plane, with $P = \{p_1,\dots,p_k\}$, $k \geq 2$, the set of generator points,

  $$VIS(p_i) = \{x \in \mathbb{R}^2 : [x,p_i] \cap]a_l, b_l[= \emptyset \text{ for all } l = 1,\dots,m\}$$

  the *visibility polygon* of the generator p_i, and d_E the Euclidean distance.

  The **visibility shortest path distance** d_{vsp} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{vsp}, \mathbb{R}^2\backslash\{\mathcal{O}\})$ (*visibility shortest path*

Voronoi diagram with line obstacles), defined by

$$d_{vsp}(x,p_i) = \begin{cases} d_E(x,p_i), & \text{if } x \in VIS(p_i), \\ \infty, & \text{otherwise} . \end{cases}$$

- **Network distances**

  A *network* on $\mathbb{R}^2$ is a connected planar geometrical graph $G = (V,E)$ with the set V of vertices and the set E of edges (links).

  Let the generator set $P = \{p_1,\ldots,p_k\}$ be a subset of the set $V = \{p_1,\ldots,p_l\}$ of vertices of G, and let the set L be given by points of links of G.

  The **network distance** d_{netv} on the set V is the Voronoi generation distance of the *network Voronoi node diagram* $V(P, d_{netv}, V)$ defined as the shortest path along the links of G from $p_i \in V$ to $p_j \in V$. It is the weighted path metric of the graph G, where w_e is the Euclidean length of the link $e \in E$.

  The **network distance** d_{netl} on the set L is the Voronoi generation distance of the *network Voronoi link diagram* $V(P, d_{netl}, L)$ defined as the shortest path along the links from $x \in L$ to $y \in L$.

  The **access network distance** d_{accnet} on $\mathbb{R}^2$ is the Voronoi generation distance of the *network Voronoi area diagram* $V(P, d_{accnet}, \mathbb{R}^2)$ defined by

$$d_{accnet}(x,y) = d_{netl}(l(x), l(y)) + d_{acc}(x) + d_{acc}(y),$$

  where $d_{acc}(x) = \min_{l \in L} d(x,l) = d_E(x, l(x))$ is the *access distance* of a point x. In fact, $d_{acc}(x)$ is the Euclidean distance from x to the *access point* $l(x) \in L$ of x which is the nearest to x point on the links of G. The **access distance**, among **memory distances** in Chap. 22, is not related.
- **Airlift distance**

  An *airports network* is an arbitrary planar graph G on n vertices (*airports*) with positive edge weights (*flight durations*). This graph may be entered and exited only at the airports. Once having accessed G, one travels at fixed speed $v > 1$ within the network. Movement off the network takes place with the unit speed with respect to the Euclidean distance.

  The **airlift distance** d_{al} is the Voronoi generation distance of the *airlift Voronoi diagram* $V(P, d_{al}, \mathbb{R}^2)$, defined as the time needed for a *quickest*, i.e., minimizing the travel time, path between x and y in the presence of the airports network G.
- **City distance**

  A *city public transportation network*, like a subway or a bus transportation system, is a planar straight line graph G with horizontal or vertical edges. G may be composed of many connected components, and may contain cycles. One is free to enter G at any point, be it at a vertex or on an edge (it is possible to postulate fixed entry points, too). Once having accessed G, one travels at a fixed speed $v > 1$ in one of the available directions. Movement off the network takes place with the unit speed with respect to the **Manhattan metric**.

[redacted]

The **city distance** d_{city} is the Voronoi generation distance of the *city Voronoi diagram* $V(P, d_{city}, \mathbb{R}^2)$, defined as the time needed for the *quickest path*, i.e., the one minimizing the travel time, between x and y in the presence of the network G.

The set $P = \{p_1, \dots, p_k\}$, $k \geq 2$, can be seen as a set of some city facilities (say, post offices or hospitals): for some people several facilities of the same kind are equally attractive, and they want to find out which facility is reachable first.

- **Distance in a river**

  The **distance in a river** d_{riv} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{riv}, \mathbb{R}^2)$ (*Voronoi diagram in a river*), defined by

  $$d_{riv}(x, y) = \frac{-\alpha(x_1 - y_1) + \sqrt{(x_1 - y_1)^2 + (1 - \alpha^2)(x_2 - y_2)^2}}{v(1 - \alpha^2)},$$

  where v is the speed of the boat on still water, $w > 0$ is the speed of constant flow in the positive direction of the x_1 axis, and $\alpha = \frac{w}{v}$ $(0 < \alpha < 1)$ is the *relative flow speed.*

- **Boat-sail distance**

  Let $\Omega \subset \mathbb{R}^2$ be a *domain* in the plane (*water surface*), let $f : \Omega \to \mathbb{R}^2$ be a continuous vector field on Ω, representing the velocity of the water flow (*flow field*); let $P = \{p_1, \dots, p_k\} \subset \Omega$, $k \geq 2$, be a set of k points in Ω (*harbors*).

  The **boat-sail distance** ([NiSu03]) d_{bs} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{bs}, \Omega)$ (*boat-sail Voronoi diagram*) defined by

  $$d_{bs}(x, y) = \inf_{\gamma} \delta(\gamma, x, y)$$

  for all $x, y \in \Omega$, where $\delta(\gamma, x, y) = \int_0^1 \left| F \frac{\gamma'(s)}{|\gamma'(s)|} + f(\gamma(s)) \right|^{-1} ds$ is the time necessary for the boat with the maximum speed F on still water to move from x to y along the curve $\gamma : [0, 1] \to \Omega$, $\gamma(0) = x$, $\gamma(1) = y$, and the infimum is taken over all possible curves γ.

- **Peeper distance**

  Let $S = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > 0\}$ be the half-plane in $\mathbb{R}^2$, let $P = \{p_1, \dots, p_k\}$, $k \geq 2$, be a set of points contained in the half-plane $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 < 0\}$, and let the *window* be the interval (a, b) with $a = (0, 1)$ and $b = (0, -1)$.

  The **peeper distance** d_{pee} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{pee}, S)$ (*peeper's Voronoi diagram*) defined by

  $$d_{pee}(x, p_i) = \begin{cases} d_E(x, p_i), & \text{if } [x, p] \cap]a, b[\neq \emptyset, \\ \infty, & \text{otherwise}, \end{cases}$$

  where d_E is the Euclidean distance.

- **Snowmobile distance**

  Let $\Omega \subset \mathbb{R}^2$ be a *domain* in the x_1x_2-plane of the space $\mathbb{R}^3$ (a *2D mapping*), and let $\Omega^* = \{(q, h(q)) : q = (x_1(q), x_2(q)) \in \Omega, h(q) \in \mathbb{R}\}$ be the three-dimensional *land surface* associated with the mapping Ω. Let $P = \{p_1, \dots, p_k\} \subset \Omega$, $k \geq 2$, be a set of k points in Ω (*snowmobile stations*).

  The **snowmobile distance** d_{sm} is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{sm}, \Omega)$ (*snowmobile Voronoi diagram*) defined by

  $$d_{sm}(q, r) = \inf_{\gamma} \int_{\gamma} \frac{1}{F\left(1 - \alpha \frac{dh(\gamma(s))}{ds}\right)} ds$$

  for any $q, r \in \Omega$, and calculating the minimum time necessary for the snowmobile with the speed F on flat land to move from $(q, h(q))$ to $(r, h(r))$ along the *land path* $\gamma^* : \gamma^*(s) = (\gamma(s), h(\gamma(s)))$ associated with the *domain path* $\gamma : [0, 1] \to \Omega$, $\gamma(0) = q$, $\gamma(1) = r$. The infimum is taken over all possible paths γ, and α is a positive constant, since a snowmobile goes uphill more slowly than downhill.

  The situation is opposite for a forest fire, and it can be modeled using a negative value of α. The resulting distance and Voronoi diagram are called **forest-fire distance** and *forest-fire Voronoi diagram*.

- **Skew distance**

  Let T be a *tilted plane* in $\mathbb{R}^3$, obtained by rotating the x_1x_2 plane around the x_1 axis through the angle α, $0 < \alpha < \frac{\pi}{2}$, with the coordinate system obtained by taking the coordinate system of the x_1x_2 plane, accordingly rotated. For a point $q \in T$, $q = (x_1(q), x_2(q))$, define the *height* $h(q)$ as its x_3 coordinate in $\mathbb{R}^3$. Thus, $h(q) = x_2(q) \cdot \sin \alpha$. Let $P = \{p_1, \dots, p_k\} \subset T$, $k \geq 2$.

  The **skew distance** is the Voronoi generation distance of the generalized Voronoi diagram $V(P, d_{skew}, T)$ (*skew Voronoi diagram*) defined ([AACLMP98]) by

  $$d_{skew}(q, r) = d_E(q, r) + (h(r) - h(q)) = d_E(q, r) + \sin \alpha (x_2(r) - x_2(q))$$

  or, more generally, by

  $$d_{skew}(q, r) = d_E(q, r) + k(x_2(r) - x_2(q))$$

  for all $q, r \in T$, where d_E is the Euclidean distance, and $k \geq 0$ is a constant.

[redacted]

20.3 Other Voronoi Generation Distances

- **Voronoi distance for line segments**

  The **Voronoi distance for** (a set of) **line segments** d_{sl} is the Voronoi generation distance of the generalized Voronoi diagram $V(A, d_{sl}, \mathbb{R}^2)$ (*line Voronoi diagram generated by straight line segments*) defined by

$$d_{sl}(x, A_i) = \inf_{y \in A_i} d_E(x, y),$$

  where the *generator set* $A = \{A_1, \dots, A_k\}$, $k \geq 2$, is a set of pairwise disjoint straight line segments $A_i = [a_i, b_i]$, and d_E is the ordinary Euclidean distance. In fact,

$$d_{sl}(x, A_i) = \begin{cases} d_E(x, a_i), & \text{if} \quad x \in R_{a_i}, \\ d_E(x, b_i), & \text{if} \quad x \in R_{b_i}, \\ d_E(x - a_i, \frac{(x-a_i)^T(b_i-a_i)}{d_E^2(a_i,b_i)}(b_i - a_i)), & \text{if} \ x \in \mathbb{R}^2 \backslash \{R_{a_i} \cup R_{b_i}\}, \end{cases}$$

  where $R_{a_i} = \{x \in \mathbb{R}^2 : (b_i - a_i)^T(x - a_i) < 0\}$, $R_{b_i} = \{x \in \mathbb{R}^2 : (a_i - b_i)^T(x - b_i) < 0\}$.
- **Voronoi distance for arcs**

  The **Voronoi distance for** (a set of circle) **arcs** d_{ca} is the Voronoi generation distance of the generalized Voronoi diagram $V(A, d_{ca}, \mathbb{R}^2)$ (*line Voronoi diagram generated by circle arcs*) defined by

$$d_{ca}(x, A_i) = \inf_{y \in A_i} d_E(x, y),$$

  where the *generator set* $A = \{A_1, \dots, A_k\}$, $k \geq 2$, is a set of pairwise disjoint circle arcs A_i (less than or equal to a semicircle) with radius r_i centered at x_{c_i}, and d_E is the Euclidean distance. In fact,

$$d_{ca}(x, A_i) = \min\{d_E(x, a_i), d_E(x, b_i), |d_E(x, x_{c_i}) - r_i|\},$$

  where a_i and b_i are the endpoints of A_i.
- **Voronoi distance for circles**

  The **Voronoi distance for** (a set of) **circles** d_{cl} is the Voronoi generation distance of a generalized Voronoi diagram $V(A, d_{cl}, \mathbb{R}^2)$ (*line Voronoi diagram generated by circles*) defined by

$$d_{cl}(x, A_i) = \inf_{y \in A_i} d_E(x, y),$$

where the *generator set* $A = \{A_1, \dots, A_k\}$, $k \geq 2$, is a set of pairwise disjoint circles A_i with radius r_i centered at x_{c_i}, and d_E is the Euclidean distance. In fact,

$$d_{cl}(x, A_i) = |d_E(x, x_{c_i}) - r_i|.$$

Examples of above Voronoi distances are $d^{\star}_{cl}(x, A_i) = d_E(x, x_{c_i}) - r_i$ and $d^{*}_{cl}(x, A_i) = d^2_E(x, x_{c_i}) - r_i^2$ (the *Laguerre Voronoi diagram*).

- **Voronoi distance for areas**

  The **Voronoi distance for areas** d_{ar} is the Voronoi generation distance of the generalized Voronoi diagram $V(A, d_{ar}, \mathbb{R}^2)$ (*area Voronoi diagram*) defined by

  $$d_{ar}(x, A_i) = \inf_{y \in A_i} d_E(x, y),$$

  where $A = \{A_1, \dots, A_k\}$, $k \geq 2$, is a collection of pairwise disjoint connected closed sets (*areas*), and d_E is the ordinary Euclidean distance.

  For any generalized generator set $A = \{A_1, \dots, A_k\}$, $k \geq 2$, one can use as the Voronoi generation distance the **Hausdorff distance** from a point x to a set A_i: $d_{Haus}(x, A_i) = \sup_{y \in A_i} d_E(x, y)$, where d_E is the Euclidean distance.

- **Cylindrical distance**

  The **cylindrical distance** d_{cyl} is the **intrinsic metric** on the surface of a cylinder S which is used as the Voronoi generation distance in the *cylindrical Voronoi diagram $V(P, d_{cyl}, S)$*. If the axis of a cylinder with unit radius is placed at the x_3 axis in $\mathbb{R}^3$, the cylindrical distance between any points $x, y \in S$ with the cylindrical coordinates $(1, \theta_x, z_x)$ and $(1, \theta_y, z_y)$ is given by

  $$d_{cyl}(x, y) = \begin{cases} \sqrt{(\theta_x - \theta_y)^2 + (z_x - z_y)^2}, & \text{if } \theta_y - \theta_x \leq \pi, \\ \sqrt{(\theta_x + 2\pi - \theta_y)^2 + (z_x - z_y)^2}, & \text{if } \theta_y - \theta_x > \pi. \end{cases}$$

- **Cone distance**

  The **cone distance** d_{con} is the **intrinsic metric** on the surface of a cone S which is used as the Voronoi generation distance in the *conic Voronoi diagram* $V(P, d_{con}, S)$. If the axis of the cone S is placed at the x_3 axis in $\mathbb{R}^3$, and the radius of the circle made by the intersection of the cone S with the $x_1 x_2$ plane is equal to one, then the cone distance between any points $x, y \in S$ is given by

  $$d_{con}(x, y) = \begin{cases} \sqrt{r_x^2 + r_y^2 - 2 r_x r_y \cos(\theta'_y - \theta'_x)}, \\ \qquad \text{if } \theta'_y \leq \theta'_x + \pi \sin(\alpha/2), \\ \sqrt{r_x^2 + r_y^2 - 2 r_x r_y \cos(\theta'_x + 2\pi \sin(\alpha/2) - \theta'_y)}, \\ \qquad \text{if } \theta'_y > \theta'_x + \pi \sin(\alpha/2), \end{cases}$$

  where (x_1, x_2, x_3) are the Cartesian coordinates of a point x on S, α is the top angle of the cone, θ_x is the counterclockwise angle from the x_1 axis to the ray from the origin to the point $(x_1, x_2, 0)$, $\theta'_x = \theta_x \sin(\alpha/2)$, $r_x =$

[redacted]

$\sqrt{x_1^2 + x_2^2 + (x_3 - \coth(\alpha/2))^2}$ is the straight line distance from the top of the cone to the point (x_1, x_2, x_3).

- **Voronoi distances of order** m

  Given a finite set A of objects in a metric space (S, d), and an integer $m \geq 1$, consider the set of all *m-subsets* M_i of A (i.e., $M_i \subset A$, and $|M_i| = m$). The *Voronoi diagram of order m* of A is a partition of S into *Voronoi regions* $V(M_i)$ *of m-subsets* of A in such a way that $V(M_i)$ contains all points $s \in S$ which are "closer" to M_i than to any other m-set M_j: $d(s, x) < d(s, y)$ for any $x \in M_i$ and $y \in S \backslash M_i$. This diagram provides first, second, …, m-th closest neighbors of a point in S.

  Such diagrams can be defined in terms of some "distance function" $D(s, M_i)$, in particular, some m**-hemimetric** (Chap. 3) on S. For $M_i = \{a_i, b_i\}$, there were considered the functions $|d(s, a_i) - d(s, b_i)|$, $d(s, a_i) + d(s, b_i)$, $d(s, a_i) \cdot d(s, b_i)$, as well as 2**-metrics** $d(s, a_i) + d(s, b_i) + d(a_i, b_i)$ and the area of triangle (s, a_i, b_i).

Chapter 21
Image and Audio Distances

21.1 Image Distances

Image Processing treats signals such as photographs, video, or tomographic output. In particular, *Computer Graphics* consists of image synthesis from some abstract models, while *Computer Vision* extracts some abstract information: say, the 3D description of a scene from video footage of it. From about 2000, analog image processing (by optical devices) gave way to digital processing, and, in particular, digital image editing (for example, processing of images taken by popular digital cameras).

Computer graphics (and our brains) deals with *vector graphics images*, i.e., those represented geometrically by curves, polygons, etc. A *raster graphics image* (or *digital image*, *bitmap*) in $2D$ is a representation of a $2D$ image as a finite set of digital values, called *pixels* (short for picture elements) placed on a square grid $\mathbb{Z}^2$ or a hexagonal grid. Typically, the image raster is a square $2^k \times 2^k$ grid with $k = 8, 9$ or 10.

Video images and *tomographic* or magnetic resonance (obtained by cross-sectional slices) images are $3D$ ($2D$ plus time); their digital values are called *voxels* (volume elements). The spacing between two pixels in one slice is referred to as the *interpixel distance*, while the spacing between slices is the *interslice distance*.

A *digital binary image* corresponds to only two values 0,1 with 1 being interpreted as logical “true” and displayed as black; so, such image is identified with the set of black pixels. A *continuous binary image* is a (usually, compact) subset of a **locally compact** metric space (usually, Euclidean space $\mathbb{E}^n$ with $n = 2, 3$).

The *gray-scale images* can be seen as point-weighted binary images. In general, a *fuzzy set* is a point-weighted set with weights (*membership values*); cf. **metrics between fuzzy sets** in Chap. 1. For the gray-scale images, xyi-representation is used, where plane coordinates (x, y) indicate shape, while the weight i (short for intensity, i.e., brightness) indicates *texture*. Sometimes, the matrix $((i_{xy}))$ of gray-levels is used.

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_21

[redacted]

The *brightness histogram* of a gray-scale image provides the frequency of each brightness value found in that image. If an image has m brightness levels (bins of gray-scale), then there are 2^m different possible intensities. Usually, $m = 8$ and numbers $0, 1, \dots, 255$ represent the intensity range from black to white; other typical values are $m = 10, 12, 14, 16$. Humans can differ between around 10 million different colors but between only 30 different gray-levels; so, color has much higher discriminatory power.

For color images, (RGB)-representation is the better known, where space coordinates R, G, B indicate red, green and blue levels; a $3D$ histogram provides brightness at each point. Among many other $3D$ color models (spaces) are: (CMY) cube (Cyan, Magenta, Yellow colors), (HSL) cone (Hue-color type given as an angle, Saturation in %, Luminosity in %), and (YUV), (YIQ) used, respectively, in PAL, NTSC television. CIE-approved conversion of (RGB) into luminance (luminosity) of gray-level is $0.299R + 0.587G + 0.114B$. The *color histogram* is a feature vector with components representing either the total number of pixels, or the percentage of pixels of a given color in the image.

Images are often represented by *feature vectors*, including color histograms, color moments, textures, shape descriptors, etc. Examples of feature spaces are: *raw intensity* (pixel values), *edges* (boundaries, contours, surfaces), *salient features* (corners, line intersections, points of high curvature), and *statistical features* (moment invariants, centroids). Typical video features are in terms of overlapping frames and motions.

Image Retrieval (similarity search) consists of (as for other data: audio recordings, DNA sequences, text documents, time-series, etc.) finding images whose features have values either mutual similarity, or similarity to a given query or in a given range.

There are two methods to compare images directly: intensity-based (color and texture histograms), and geometry-based (shape representations by *medial axis*, *skeletons*, etc.). The imprecise term *shape* is used for the extent (silhouette) of the object, for its local geometry or geometrical pattern (conspicuous geometric details, points, curves, etc.), or for that pattern modulo a similarity transformation group (translations, rotations, and scalings). The imprecise term *texture* means all that is left after color and shape have been considered, or it is defined via structure and randomness.

The similarity between vector representations of images is measured by the usual practical distances: **l_p-metrics**, **weighted editing metrics**, **Tanimoto distance**, **cosine distance**, **Mahalanobis distance** and its extension, **distance**.

Among probabilistic distances, the following ones are most used: **Bhattacharya** 2, **Hellinger**, **Kullback–Leibler**, **Kolmogorov–Smirnov**, **Jeffrey** and (especially, for histograms) **χ^2-**, **Kuiper distances**.

The main distances applied for compact subsets X and Y of $\mathbb{R}^n$ (usually, $n = 2, 3$) or their digital versions are: **Asplund metric**, **Shephard metric**, **symmetric difference semimetric** $Vol(X \Delta Y)$ (cf. **Nykodym metric**, **area deviation**, **digital volume metric** and their normalizations) and variations of the **Hausdorff distance** (see below).

For Image Processing, the distances below are between “true” and approximated digital images, in order to assess the performance of algorithms. For Image Retrieval, distances are between feature vectors of a query and reference.

- **Color distances**

  The visible spectrum of a typical human eye is about 380–760 nm. It matches the range of wavelengths sustaining photosynthesis; also, at those wavelengths opacity often coincides with impenetrability. A light-adapted eye has its maximum sensitivity at $\approx$ 555 nm (540 THz), in the green region of the optical spectrum.

  A *color space* is a 3-parameter description of colors. The need for exactly three parameters comes from the existence of three kinds of receptors (cells on the retina) in the human eye: for short, middle and long wavelengths, corresponding to blue, green, and red. Their respective sensitivity peaks are situated around 570 nm, 543 nm and 442 nm. About 1 of 10 women has a 4-th type of color receptor. Color blindness is 10 times more common in males. People with absent or removed lens of the eye, can see UV (ultraviolet) wavelengths (400–10 nm).

  The mantis shrimp has 12 types of color receptors including 4 for UV; its species *Gonodactylus smithii* is the only organism known to have optimal polarization vision. Some dragonflies have 30 different receptors; they can see UV and polarised light. Some spiders, jellyfish and scallops have, respectively, $8, 24$, over 110 eyes.

  The CIE (International Commission on Illumination) derived (XYZ) color space in 1931 from the (RGB)-model and measurements of the human eye. In the CIE (XYZ) color space, the values X, Y and Z are also roughly red, green and blue.

  In Colorimetry, *chromaticity* is intensity of a colour (hue), defined as the distance in the particular colour space of a colour from the neutral grey colour with the same value. The basic assumption of Colorimetry (Indow, 1991), is that the perceptual color space admits a metric, the true **color distance**. This metric is expected to be almost locally Euclidean, i.e., a **Riemannian metric**. A continuous mapping from the metric space of light stimuli to this metric space is also expected.

  Such a *uniform color scale*, where equal distances in the color space correspond to equal differences in color, is not obtained yet and existing **color distances** are various approximations of it. A first step in this direction was given by *MacAdam ellipses*, i.e., regions on a chromaticity (x, y) diagram which contains all colors looking indistinguishable to the average human eye; cf. JND (just-noticeable difference) **video quality metric**. For any $\epsilon > 0$, the **MacAdam metric** in a color space is the metric for which those 25 ellipses are circles of radius ϵ. Here $x = \frac{X}{X+Y+Z}$ and $y = \frac{Y}{X+Y+Z}$ are projective coordinates, and the colors of the chromaticity diagram occupy a region of the real projective plane $\mathbb{R}P^2$.

[redacted]

The CIE ($L^*a^*b^*$) (CIELAB) is an adaptation of CIE 1931 (XYZ) color space; it gives a partial linearization of the MacAdam color metric. The L^*, a^*, b^* of the most complete model are derived from L, a, b which are: the luminance L of the color from black $L = 0$ to white $L = 100$, its position a between green $a < 0$ and red $a > 0$, and its position b between green $b < 0$ and yellow $b > 0$.

Examples of non-Euclidean metrics, fitting well to observed color differences, are CIE94 (McDonald–Smith, 1995) and CIEDE2000 (Luo–Cui–Rigg, 2001). The best performing Euclidean metrics are versions of DIN99 (Cui et al., 2002) and the log-compressed OSA-UCS (Oleari et al., 2009).

Farup, 2014, proposed to transform a chromatic plane $\{(x, y)\}$ (say, $\{(a^*, b^*)\}$ of CIELAB) with polar coordinates (r, θ) and an existing Euclidean metric, into the R-scaled **Poincaré disk** by $\tilde{r} = \tanh(\frac{r}{2R})$ and observed that the resulting **hyperbolic metrics** (Chap. 6)) with optimised R perform better.

- **Average color distance**

  For a given $3D$ color space OSA-UC a list of n colors, let (c_{i1}, c_{i2}, c_{i3}) be the representation of the i-th color of the list in this space. For a color histogram $x = (x_1, \dots, x_n)$, its *average color* is the vector $(x_{(1)}, x_{(2)}, x_{(3)})$, where $x_{(j)} = \sum_{i=1}^{n} x_i c_{ij}$ (for example, the average red, blue and green values in (RGB)).

  The **average color distance** between two color histograms ([HSEFN95]) is the Euclidean distance of their average colors.

- **Color component distance**

  Given an image (as a subset of $\mathbb{R}^2$), let p_i denote the area percentage of this image occupied by the color c_i. A *color component* of the image is a pair (c_i, p_i).

  The **color component distance** (Ma–Deng–Manjunath, 1997) between color components (c_i, p_i) and (c_j, p_j) is defined by

$$|p_i - p_j| \cdot d(c_i, c_j),$$

  where $d(c_i, c_j)$ is the distance between c_i and c_j in a given color space. Mojsilović–Hu–Soljanin, 2002, did an **Earth Mover's distance**-like modification of it.

- **Riemannian color space**

  The proposal to measure perceptual dissimilarity of colors by a **Riemannian metric** (Chap. 7) on a strictly convex cone $C \subset \mathbb{R}^3$ comes from von Helmholtz, 1891, and Schrödinger, 1920. The challenge here is to choose a "good" one.

  Roughly, it was shown in [Resn74] that the only such *GL-homogeneous* cones C (i.e., the group of all orientation preserving linear transformations of $\mathbb{R}^3$, carrying C into itself, acts transitively on C) are either $C_1 = \mathbb{R}_{>0} \times (\mathbb{R}_{>0} \times \mathbb{R}_{>0})$, or $C_2 = \mathbb{R}_{>0} \times C'$, where C' is the set $SL(2, \mathbb{R})/SO(2)$ of 2×2 real symmetric matrices with determinant 1. The first factor $\mathbb{R}_{>0}$ can be identified with variation of brightness and the other with the set of lights of a fixed brightness.

Given some $\alpha_i > 0$, the **Stiles color metric** (1946) is the GL-invariant Riemannian metric on $C_1 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_i > 0\}$ given by the *line element*

$$ds^2 = \sum_{i=1}^{3} \alpha_i (\frac{dx_i}{x_i})^2.$$

The **Resnikoff color metric** (1974) is the GL-invariant Riemannian metric on $C_2 = \{(x, u) : x \in \mathbb{R}_{>0}, u \in C'\}$ given by the *line element*

$$ds^2 = \alpha_1 (\frac{dx}{x})^2 + \alpha_2 ds^2_{C'},$$

where $ds^2_{C'}$, is the **Poincaré metric** (Chap. 6) on C'.

- **Histogram intersection quasi-distance**

  Given two color histograms $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ (with x_i, y_i representing the number of pixels in the bin i), the **histogram intersection quasi-distance** between them (cf. **intersection distance** in Chap. 17) is (Swain–Ballard, 1991) defined by

  $$1 - \frac{\sum_{i=1}^{n} \min\{x_i, y_i\}}{\sum_{i=1}^{n} x_i}.$$

  For normalized histograms (total sum is 1) the above quasi-distance becomes the usual l_1-metric $\sum_{i=1}^{n} |x_i - y_i|$. The *normalized cross-correlation* (Rosenfeld–Kak, 1982) between x and y is a similarity defined by $\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$.

- **Histogram quadratic distance**

  Given two color histograms $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ (usually, $n = 256$ or $n = 64$) representing the color percentages of two images, their **histogram quadratic distance** (used in IBM's Query By Image Content system) is their **Mahalanobis distance** defined in Chap. !7 by

  $$\sqrt{(x-y)^T A (x-y)},$$

  where $A = ((a_{ij}))$ is a symmetric positive-definite matrix, and the weight a_{ij} is some, perceptually justified, similarity between colors i and j.

  For example (cf. [HSEFN95]), $a_{ij} = 1 - \frac{d_{ij}}{\max_{1 \le p,q \le n} d_{pq}}$, where d_{ij} is the Euclidean distance between 3-vectors representing i and j in some color space.

  If (h_i, s_i, v_i) and (h_j, s_j, v_j) are the representations of the colors i and j in the color space (HSV), then $a_{ij} = 1 - \frac{1}{\sqrt{5}}((v_i - v_j)^2 + (s_i \cos h_i - s_j \cos h_j)^2 + (s_i \sin h_i - s_j \sin h_j)^2)^{\frac{1}{2}}$ is used.

[redacted]

- **Histogram diffusion distance**

  Given two histogram-based descriptors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, their **histogram diffusion distance** (Ling–Okada, 2006) is defined by

$$\int_0^T ||u(t)||_1 dt,$$

  where T is a constant, and $u(t)$ is a heat diffusion process with initial condition $u(0) = x - y$. In order to approximate the diffusion, the initial condition is convoluted with a Gaussian window; then the sums of l_1-norms after each convolution approximate the integral.
- **Gray-scale image distances**

  Let $f(x)$ and $g(x)$ denote the brightness values of two digital gray-scale images f and g at the pixel $x \in X$, where X is a raster of pixels. Any distance between point-weighted sets (X, f) and (X, g) (for example, the **Earth Mover's distance**) can be applied for measuring distances between f and g. However, the main used distances (called also *errors*) between the images f and g are:

  1. The *root-mean-square error* $RMS(f, g) = \left(\frac{1}{|X|}\sum_{x \in X}(f(x) - g(x))^2\right)^{\frac{1}{2}}$ (a variant is to use the l_1-*norm* $|f(x) - g(x)|$ instead of the l_2-*norm*);
  2. The **signal-to-noise ratio** $SNR(f, g) = \left(\frac{\sum_{x \in X} g(x)^2}{\sum_{x \in X}(f(x)-g(x))^2}\right)^{\frac{1}{2}}$ (cf. **SNR distance** between sonograms);
  3. The *pixel misclassification error rate* $\frac{1}{|X|}|\{x \in X : f(x) \neq g(x)\}|$ (normalized **Hamming distance**);
  4. The *frequency root-mean-square error* $\left(\frac{1}{|U|^2}\sum_{u \in U}(F(u) - G(u))^2\right)^{\frac{1}{2}}$, where F and G are the discrete Fourier transforms of f and g, respectively, and U is the frequency domain;
  5. The *Sobolev norm of order* δ *error* $\left(\frac{1}{|U|^2}\sum_{u \in U}(1 + |\eta_u|^2)^{\delta}(F(u) - G(u))^2\right)^{\frac{1}{2}}$, where $0 < \delta < 1$ is fixed (usually, $\delta = \frac{1}{2}$), and η_u is the $2D$ frequency vector associated with position u in the frequency domain U.

  Cf. **metrics between fuzzy sets** in Chap. 1.
- **Image compression L_p-metric**

  Given a number r, $0 \leq r < 1$, the **image compression L_p-metric** is the usual **L_p-metric** on $\mathbb{R}^{n^2}_{\geq 0}$ (the set of gray-scale images seen as $n \times n$ matrices) with p being a solution of the equation $r = \frac{p-1}{2p-1} \cdot e^{\frac{p}{2p-1}}$. So, $p = 1, 2$, or ∞ for, respectively, $r = 0, r = \frac{1}{3}e^{\frac{2}{3}} \approx 0.65$, or $r \geq \frac{\sqrt{e}}{2} \approx 0.82$. Here r estimates the *informative* (i.e., filled with nonzeros) part of the image. According to [KKN02], it is the best quality metric to select a lossy compression scheme.
- **Chamfering distances**

  The **chamfering distances** are distances approximating Euclidean distance by a weighted path distance on the graph $G = (\mathbb{Z}^2, E)$, where two pixels are neighbors if one can be obtained from another by an *one-step move* on $\mathbb{Z}^2$. The list of permitted moves is given, and a **prime distance**, i.e., a positive weight (cf. Chap. 19), is associated to each type of such move.

An (α, β)**-chamfer metric** corresponds to two permitted moves—with l_1-distance 1 and with l_∞-distance 1 (diagonal moves only)—weighted α and β, respectively.

The main applied cases are $(\alpha, \beta) = (1, 0)$ (**city-block metric**, or 4**-metric**), (1,1) (**chessboard metric**, or 8**-metric**), $(1, \sqrt{2})$ (**Montanari metric**), $(3, 4)$ ($(3, 4)$**-metric**), $(2, 3)$ (**Hilditch–Rutovitz metric**), $(5, 7)$ (**Verwer metric**).

The **Borgefors metric** corresponds to three permitted moves—with l_1-distance 1, with l_∞-distance 1 (diagonal moves only) and knight moves—weighted 5, 7 and 11.

An 3*D***-chamfer metric** (or (α, β, γ)*-chamfer metric*) is the weighted path metric of the infinite graph with the vertex-set $\mathbb{Z}^3$ of voxels, two vertices being adjacent if their l_∞-distance is one, while weights α, β and γ are associated to 6 face, 12 edge and 8 corner neighbors, respectively. If $\alpha = \beta = \gamma = 1$, we obtain l_∞-metric. The $(3, 4, 5)$- and $(1, 2, 3)$-chamfer metrics are the most used ones.

The **Chaudhuri–Murthy–Chaudhuri metric** between sequences $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_n)$ is defined by

$$|x_{i(x,y)} - y_{i(x,y)}| + \frac{1}{1 + \lceil \frac{n}{2} \rceil} \sum_{1 \le i \le n, i \ne i(x,y)} |x_i - y_i|,$$

where $\max_i |x_i - y_i| = |x_{i(x,y)} - y_{i(x,y)}|$. For $n = 2$ it is the $(1, \frac{3}{2})$-chamfer metric.

- **Earth Mover's distance**

  The **Earth Mover's distance** is a discrete form of the **transportation distance** (Chap. 14). Roughly, it is the minimal amount of work needed to transport earth or mass from one position (properly spread in space) to the other (a collection of holes). For any two finite sequences $(x_1, \dots, x_m)$ and $(y_1, \dots, y_n)$ over a metric space (X, d), consider *signatures*, i.e., point-weighted sets $P_1 = (p_1(x_1), \dots, p_1(x_m))$ and $P_2 = (p_2(y_1), \dots, p_2(y_n))$.

  For example, in [RTG00]) signatures represent clustered color or texture content of images: elements of X are centroids of clusters, and $p_1(x_i), p_2(y_j)$ are cardinalities of corresponding clusters. The ground distance d is a **color distance**, say, the Euclidean distance in 3*D* CIE $(L^* a^* b^*)$ color space.

  Let $W_1 = \sum_i p_1(x_i)$ and $W_2 = \sum_j p_2(y_j)$ be the *total weights* of P_1 and P_2, respectively. Then the **Earth Mover's distance** between P_1 and P_2 is defined as

$$\frac{\sum_{i,j} f_{ij}^* d(x_i, y_j)}{\sum_{i,j} f_{ij}^*},$$

[redacted]

where the $m \times n$ matrix $S^* = ((f^*_{ij}))$ is an *optimal*, i.e., minimizing $\sum_{i,j} f_{ij} d(x_i, y_j)$, *flow*. A *flow* is an $m \times n$ matrix $S = ((f_{ij}))$ with following constraints:

1. all $f_{ij} \geq 0$;
2. $\sum_{i,j} f_{ij} = \min\{W_1, W_2\}$;
3. $\sum_i f_{ij} \leq p_2(y_j)$ and $\sum_j f_{ij} \leq p_1(x_i)$.

So, this distance is the average ground distance d that weights travel during an optimal flow. It is not a *bin-to-bin* (component-wise, as L_p-, Kullback–Leibler, χ^2-distances), but a **cross-bin histogram distance**.

In the case $W_1 = W_2$, the above two inequalities 3. become equalities. Normalizing signatures to $W_1 = W_2 = 1$ (which not changes the distance) allows us to see P_1 and P_2 as probability distributions of random variables, say, X and Y. Then $\sum_{i,j} f_{ij} d(x_i, y_j)$ is $\mathbb{E}_S[d(X, Y)]$, i.e., the Earth Mover's distance coincides, in this case, with the **transportation distance** (Chap. 14).

For $W_1 \neq W_2$, it is not a metric in general. However, replacing the inequalities 3 in the above definition by equalities

$3'$. $\sum_i f_{ij} = p_2(y_j)$ and $\sum_j f_{ij} = \frac{p_1(x_i)W_1}{W_2}$

produces the Giannopoulos–Veltkamp's **proportional transport semimetric**.

- **Parameterized curves distance**

  The shape can be represented by a parametrized curve on the plane. Usually, such a curve is *simple*, i.e., it has no self-intersections. Let $X = X(x(t))$ and $Y = Y(y(t))$ be two parametrized curves, where the (continuous) parametrization functions $x(t)$ and $y(t)$ on $[0, 1]$ satisfy $x(0) = y(0) = 0$ and $x(1) = y(1) = 1$.

  The most used **parametrized curves distance** is the minimum, over all monotone increasing parametrizations $x(t)$ and $y(t)$, of the maximal Euclidean distance $d_E(X(x(t)), Y(y(t)))$. It is the Euclidean special case of the **dogkeeper distance** (cf. Sect. 13.1), i.e., the **Fréchet metric** for the case of curves.

  Among variations of this distance are dropping the monotonicity condition of the parametrization (*weak Fréchet distance*), or finding the part of one curve to which the other has the smallest such distance ([VeHa01]).

  In a general Riemannian manifold M, the *shape* $[C]$ of a curve C can be seen as the class of all curves that differ from C only by a parametration, i.e., they pass through the same points of M but at different speeds. The geodesic distances d on curves and D on shapes are linked by $D([C], [C']) = \inf d(C, C' \circ \alpha)$, where $\alpha : [0, 1] \to [0, 1]$ is any *reparameterization* (increasing diffeomorphism(.

  A reparametrization-invariant metric on the space of curves induces a Riemannian structure on the *shape space*. Two such metrics are **pullback metrics** (cf. Chap. 4) on the tangent bundle via *square root velocity function*: from L_2-metric by Srivastava et al., 2011, and from a more adapted metric by Le Brigant, 2016.

- **Homotopic Fréchet distance**

  Given a metric space (X, d), a *curve* is a continuous function $C : [0, 1] \to X$. The *length* of C is defined by the metric d. A *reparameterization* of C is any curve $C \circ \alpha$, where $\alpha : [0, 1] \to [0, 1]$ is a continuous nondecreasing surjection.

An (s, t)-*leash* between two curves C and C' is another curve $\lambda : [0, 1] \to X$ such that $\lambda(0) = C(s)$ and $\lambda(1) = C'(t)$. A *leash map* is a continuous function $l : [0, 1]^2 \to X$ such that $l(\cdot, 0)$ and $l(\cdot, 1)$ are reparameterizations of C and C', respectively. It describes the continuous motion of a leash between a dog walking along C and its owner walking along C'; the curve $l(t, \cdot)$ is the leash at time t. The *length* of l is the maximum length of any leash $l(t, \cdot)$.

The **homotopic Fréchet distance** between C and C' is ([CCELLT10]) the infimum, over all leash maps l between C and C', of the length of l. This distance can be thought of as the minimal amount of deformation needed to transform C into C', identifying points which are joined by a leash during its motion.

- **Nonlinear elastic matching distance**

Consider a digital representation of curves. Let $r \geq 1$ be a constant, and let $A = \{a_1, \dots, a_m\}$, $B = \{b_1, \dots, b_n\}$ be finite ordered sets of consecutive points on two closed curves. For any order-preserving correspondence f between all points of A and all points of B, the *stretch* $s(a_i, b_j)$ *of* $(a_i, f(a_i) = b_j)$ is r if either $f(a_{i-1}) = b_j$ or $f(a_i) = b_{j-1}$, or zero otherwise.

The **relaxed nonlinear elastic matching distance** is $\min_f \sum(s(a_i, b_j) + d(a_i, b_j))$, where $d(a_i, b_j)$ is the difference between the tangent angles of a_i and b_j. It is a **near-metric** for some r. For $r = 1$, it is called the **nonlinear elastic matching distance**. In general, Younes, 1998, and Mio–Srivastava–Joshi, 2005, introduced *elastic* Riemannian distances between (seen as elastic) plane curves (or enclosed shapes) measuring the minimal cost of elastic reshaping of a curve into another.

- **Turning function distance**

For a plane polygon P, its *turning function* $T_P(s)$ is the angle between the counterclockwise tangent and the x axis as a function of the arc length s. This function increases with each left hand turn and decreases with right-hand turns.

Given two polygons of equal perimeters, their **turning function distance** is the L_p**-metric** between their turning functions.

- **Size function distance**

For a shape, seen as a plane graph $G = (V, E)$, and a *measuring function* f on its vertex-set V (for example, the distance from $v \in V$ to the center of mass of V), the *size function* $S_G(x, y)$ is defined, on the points $(x, y) \in \mathbb{R}^2$, as the number of connected components of the restriction of G on vertices $\{v \in V : f(v) \leq y\}$ which contain a point v' with $f(v') \leq x$.

Given two plane graphs with vertex-sets belonging to a raster $R \subset \mathbb{Z}^2$, their Uras–Verri's **size function distance** is the normalized l_1-distance between their size functions over raster pixels. The **matching distance** (Chap. 1) between the *cornerpoints/cornerlines* multisets of two size functions is also used.

- **Reflection distance**

For a finite union A of plane curves and each point $x \in \mathbb{R}^2$, let V_A^x denote the union of intervals (x, a), $a \in A$ which are *visible from* x, i.e., $(x, a) \cap A = \emptyset$. Denote by ρ_A^x the area of the set $\{x + v \in V_A^x : x - v \in V_A^x\}$.

The Hagedoorn–Veltkamp's **reflection distance** between finite unions A and B of plane curves is the normalized l_1-distance between the corresponding

[redacted]

functions ρ_A^x and ρ_B^x defined by

$$\frac{\int_{\mathbb{R}^2} |\rho_A^x - \rho_B^x| dx}{\int_{\mathbb{R}^2} \max\{\rho_A^x, \rho_B^x\} dx}.$$

- **Distance transform**

  Given a metric space $(X = \mathbb{Z}^2, d)$ and a binary digital image $M \subset X$, the **distance transform** is a function $f_M : X \to \mathbb{R}_{\geq 0}$, where $f_M(x) = \inf_{u \in M} d(x, u)$ is the **point-set distance** $d(x, M)$. So, a distance transform can be seen as a gray-scale digital image where each pixel is given a label (a gray-level) which corresponds to the distance to the nearest pixel of the background. Distance transforms, in Image Processing, are also called *distance fields* and **distance maps**; but we reserve the last term only for this notion in any metric space as in Chap. 1.

  A *distance transform of a shape* is the distance transform with M being the boundary of the image. For $X = \mathbb{R}^2$, the graph $\{(x, f(x)) : x \in X\}$ of $d(x, M)$ is called the *Voronoi surface* of M.

- **Medial axis and skeleton**

  Let (X, d) be a metric space, and let M be a subset of X. The **medial axis** of X is the set $MA(X) = \{x \in X : |\{m \in M : d(x, m) = d(x, M)\}| \geq 2\}$, i.e., all points of X which have in M at least two **elements of best approximation**; cf. **metric projection** in Chap. 1. $MA(X)$ consists of all points of boundaries of *Voronoi regions* of points of M. The *reach* of M is the **set-set distance** (Chap. 1) between M and $MA(X)$.

  The *cut locus* of X is the closure $\overline{MA(X)}$ of the medial axis. Cf. **Shankar–Sormani radii** in Chap. 1. The *medial axis transform* $MAT(X)$ is the point-weighted set $MA(X)$ (the restriction of the **distance transform** on $MA(X)$) with $d(x, M)$ being the weight of $x \in X$.

  If (as usual in applications) $X \subset \mathbb{R}^n$ and M is the boundary of X, then the **skeleton** $Skel(X)$ of X is the set of the centers of all d-balls inscribed in X and not belonging to any other such ball; so, $Skel(X) = MA(X)$. The skeleton with M being continuous boundary is a limit of *Voronoi diagrams* as the number of the generating points becomes infinite. For $2D$ binary images X, the skeleton is a curve, a single-pixel thin one, in the digital case. The *exoskeleton* of X is the skeleton of the complement of X, i.e., of the background of the image for which X is the foreground.

- **Procrustes distance**

  The *shape* of a *form* (configuration of points in $\mathbb{R}^2$), seen as expression of translation-, rotation- and scale-invariant properties of form, can be represented by a sequence of *landmarks*, i.e., specific points on the form, selected accordingly to some rule. Each landmark point a can be seen as an element $(a', a'') \in \mathbb{R}^2$ or an element $a' + a''i \in \mathbb{C}$.

  Consider two shapes x and y, represented by their landmark vectors $(x_1, \dots, x_n)$ and $(y_1, \dots, y_n)$ from $\mathbb{C}^n$. Suppose that x and y are corrected for translation by setting $\sum_t x_t = \sum_t y_t = 0$. Then their **Procrustes distance** is defined by

$$\sqrt{\sum_{1 \le t \le n} |x_t - y_t|^2},$$

where two forms are, first, optimally (by least squares criterion) aligned to correct for scale, and their **Kendall shape distance** is defined by

$$\arccos \sqrt{\frac{(\sum_t x_t \bar{y}_t)(\sum_t y_t \bar{x}_t)}{(\sum_t x_t \bar{x}_t)(\sum_t y_t \bar{y}_t)}},$$

where $\overline{\alpha} = a' - a''i$ is the *complex conjugate* of $\alpha = a' + a''i$.

Petitjean, 2002, extended the L_2-Wasserstein distance (cf. Chap. 14) to *colored mixtures*, i.e., ordinary mixtures of random vectors, for which an new axis (the space of colors) has been added. He remarked that the Procrustes distance is an instance of this *colored Wasserstein distance*, when this latter is minimized for a class of affine transformations (rotations and translations).

- **Shape parameters**

Let X be a figure in $\mathbb{R}^2$ with area $A(X)$, perimeter $P(X)$ and convex hull *conv* X. The main **shape parameters** of X are given below.

$D_A(X) = 2\sqrt{\frac{A(X)}{\pi}}$ and $D_P(X) = \frac{P(X)}{\pi}$ are the diameters of circles with area $A(X)$ and with perimeter $P(X)$, respectively.

Feret's diameters $F_x(X), F_y(X), F_{min}(X), F_{\max}(X)$ are the orthogonal projections of X on the x and y axes and such minimal and maximal projections on a line.

Martin's diameter $M(X)$ is the distance between opposite sides of X measured crosswise of it on a line bisecting the figure's area. $M_x(X)$ and $M_y(X)$ are Martin's diameters for horizontal and vertical directions, respectively.

$R_{in}(X)$ and $R_{out}(X)$ are the radii of the largest disc in X and the smallest disc including X. $a(X)$ and $b(X)$ are the lengths of the major and minor semiaxes of the ellipse with area $A(X)$ and perimeter $P(X)$.

Examples of the ratios, describing some shape properties in above terms, follow.

The *area-perimeter ratio* (or *projection sphericity*) and *Petland's projection sphericity ratio* are $ArPe = \frac{4\pi A(X)}{(P(X))^2}$ and $\frac{4A(X)}{\pi(F_{\max}(X))^2}$.

The *circularity shape factor* and *Horton's compactness factor* are $\frac{1}{ArPe}$ and $\frac{1}{\sqrt{ArPe}}$.

Wadell's circularity shape and *drainage-basin circularity shape* ratios are $\frac{D_A(X)}{F_{\max}(X)}$ and $\frac{A(X)}{D_P(X)}$. Both ratios and $ArPe$ are at most 1 with equality only for a disc.

[redacted]

Tickell's ratio is $(\frac{D_A(X)}{D_{out}(X)})^2$. *Cailleux's roundness ratio* is $\frac{2r(X)}{F_{\max}(X)}$, where $r(X)$ is the radius of curvature at a most convex part of the contour of X.

The *rugosity coefficient* and *convexity ratio* (or *solidity*) are $\frac{P(X)}{P(conv\,X)}$ and $\frac{A(X)}{A(conv\,X)}$. Both the solidity and $\frac{P(conv\,X)}{P(X)}$ are at most 1 with equality only for convex sets.

The *diameters ratios* are $\frac{MD_x(X)}{F_x(X)}$ and $\frac{MD_y(X)}{F_y(X)}$. The *radii ratio* and *ellipse ratio* are $\frac{R_{in}(X)}{R_{out}(X)}$ and $\frac{a(X)}{b(X)}$. The *Feret's ratio* and *modification ratio* are $\frac{F_{min}(X)}{F_{\max}(X)}$ and $\frac{R_{in}(X)}{F_{\max}(X)}$. The **aspect ratio** in Chap. 1 is the reciprocal of the Feret's ratio.

The *symmetry factor of Blaschke* is $1-\frac{A(X)}{A(S(X))}$, where $S(X) = \frac{1}{2}(X \oplus \{x : -x \in X\})$.

- **Distances from symmetry**

  Many measures of chirality and, in general, given symmetry G of a given set $A \in \mathbb{R}^n$, were proposed. Several examples follow.

  Let A' be the *enantiomorph* (mirror image) of A. Gilat, 1985, proposed to measure *distance from achirality* of A by $\frac{V(A\Delta A')}{V(A)}$; cf. **normalized volume of symmetric difference** in Chap. 9.

  Let shape A be represented by a sequence $(a_1, \dots, a_m)$ of points. Then the **symmetry distance** of A is defined by Zabrodsky–Peleg–Avnir, 1992, as the point-set distance $\inf_b \frac{1}{m}\sum_{i=1}^{m} ||a_i - b_i||_2^2$, where $b = (b_1, \dots, b_n)$ is the L_2-nearest to a representation of a *symmetric* (i.e., invariant to rotation and translation) shape. The *symmetry distance of a function f* with respect to any transformation G is the L_2-distance between f and the nearest function invariant to G.

  If A is a $2D$ object, and it is represented by its *radial function* $R(r)$, then the *distance of A from symmetry G* can be measured (Köhler, 1993) by $\int_0^{2\pi} |G(R(r)) - R(r)|dr$. For a sequence $(a_1, \dots, a_m)$ of points, similar distance is (Köhler, 1999) $\min_p \sum_{i=1}^{m} d_E(a_i, G(p(a_i)))$, where p is any of $m!$ permutations of $(a_1, \dots, a_m)$ and d_E is the Euclidean distance.

- **Tangent distance**

  For any $x \in \mathbb{R}^n$ and a family of *transformations* $t(x, \alpha)$, where $\alpha \in \mathbb{R}^k$ is the vector of k parameters (for example, the scaling factor and rotation angle), the set $M_x = \{t(x, \alpha) : \alpha \in \mathbb{R}^k\} \subset \mathbb{R}^n$ is a manifold of dimension at most k. It is a curve if $k = 1$. The minimum Euclidean distance between manifolds M_x and M_y would be a useful distance since it is invariant with respect to transformations $t(x, \alpha)$.

  However, the computation of such a distance is too difficult in general; so, M_x is approximated by its *tangent subspace at x*: $\{x + \sum_{i=1}^{k} \alpha_k x^i : \alpha \in \mathbb{R}^k\} \subset \mathbb{R}^n$, where the *tangent vectors* x^i, $1 \le i \le k$, spanning it are the partial derivatives of $t(x, \alpha)$ with respect to α. The **one-sided** (or *directed*) **tangent distance** between elements x and y of $\mathbb{R}^n$ is a quasi-distance d defined by

$$\sqrt{\min_{\alpha} ||x + \sum_{i=1}^{k} \alpha_k x^i - y||^2}.$$

The Simard–Le Cun–Denker's **tangent distance** is defined by $\min\{d(x,y), d(y,x)\}$.

- **Pixel distance**

  Consider two digital images, seen as binary $m \times n$ matrices $x = ((x_{ij}))$ and $y = ((y_{ij}))$, where a pixel x_{ij} is black or white if it is equal to 1 or 0, respectively.

  For each pixel x_{ij}, the *fringe distance map to the nearest pixel of opposite color* $D_{BW}(x_{ij})$ is the number of *fringes* expanded from (i,j) (where each fringe is composed by the pixels that are at the same distance from (i,j)) until the first fringe holding a pixel of opposite color is reached.

  The **pixel distance** (Smith–Bourgoin–Sims–Voorhees, 1994) is defined by

$$\sum_{1\le i\le m}\sum_{1\le j\le n}|x_{ij}-y_{ij}|(D_{BW}(x_{ij})+D_{BW}(y_{ij})).$$

  In a pixel-based device (computer monitor, printer, scanner), the **pixel pitch** (or *dot pitch*) is the spacing between subpixels (dots) of the same color on the inside of a display screen. Closer spacing usually produce a sharper image.

- **Pratt's figure of merit**

  In general, a *figure of merit* is a quantity used to characterize the performance of a device, system or method, relative to its alternatives. Given two binary images, seen as nonempty subsets, A and B, of a finite metric space (X,d), their **Pratt's figure of merit** (or *FOM*, Abdou–Pratt, 1979) is a quasi-distance defined by

$$\left(\max\{|A|,|B|\}\sum_{x\in B}\frac{1}{1+\alpha d(x,A)^2}\right)^{-1},$$

  where α is a scaling constant (usually, $\frac{1}{9}$), and $d(x,A) = \min_{y\in A} d(x,y)$ is the **point-set distance**.

  Similar quasi-distances are Peli–Malah's *mean error distance* $\frac{1}{|B|}\sum_{x\in B} d(x,A)$, and the *mean square error distance* $\frac{1}{|B|}\sum_{x\in B} d(x,A)^2$.

- **p-th order mean Hausdorff distance**

  Given $p \ge 1$ and two binary images, seen as nonempty subsets A and B of a finite metric space (say, a raster of pixels) (X,d), their **p-th order mean Hausdorff distance** is ([Badd92]) a normalized **L_p-Hausdorff distance**, defined by

$$\left(\frac{1}{|X|}\sum_{x\in X}|d(x,A)-d(x,B)|^p\right)^{\frac{1}{p}},$$

  where $d(x,A) = \min_{y\in A} d(x,y)$ is the **point-set distance**. The usual Hausdorff metric is proportional to the ∞-th order mean Hausdorff distance.

[redacted]

Venkatasubraminian's **Σ-Hausdorff distance** $d_{dHaus}(A,B) + d_{dHaus}(B,A)$ is equal to $\sum_{x \in A \cup B} |d(x,A) - d(x,B)|$, i.e., it is a version of L_1-Hausdorff distance.

Another version of the 1-st order mean Hausdorff distance is Lindstrom–Turk's *mean geometric error* (1998) between two images (seen as surfaces A and B) defined by

$$\frac{1}{Area(A) + Area(B)} \left(\int_{x \in A} d(x,B)dS + \int_{x \in B} d(x,A)dS \right),$$

where $Area(A)$ denotes the area of A. If the images are seen as finite sets A and B, their *mean geometric error* is defined by

$$\frac{1}{|A| + |B|} \left(\sum_{x \in A} d(x,B) + \sum_{x \in B} d(x,A) \right).$$

- **Modified Hausdorff distance**

  Given two binary images, seen as nonempty subsets A and B of a finite metric space (X,d), their Dubuisson–Jain's **modified Hausdorff distance** (1994) is defined as the maximum of **point-set distances** averaged over A and B:

  $$\max \left\{ \frac{1}{|A|} \sum_{x \in A} d(x,B), \frac{1}{|B|} \sum_{x \in B} d(x,A) \right\},$$

  while their Eiter–Mannila's **sum of minimal distances** (1997) is defined as

  $$\frac{1}{2}(\sum_{x \in A} d(x,B) + \sum_{x \in B} d(x,A)).$$

- **Partial Hausdorff quasi-distance**

  Given two binary images, seen as subsets $A, B \neq \emptyset$ of a finite metric space (X,d), and integers k, l with $1 \leq k \leq |A|$, $1 \leq l \leq |B|$, their Huttenlocher–Rucklidge's **partial (k,l)-Hausdorff quasi-distance** (1992) is defined by

  $$\max\{k^{th}_{x \in A} d(x,B), l^{th}_{x \in B} d(x,A)\},$$

  where $k^{th}_{x \in A} d(x,B)$ means the k-th (rather than the largest $|A|$-th ranked one) among $|A|$ distances $d(x,B)$ ranked in increasing order. The case $k = \lfloor \frac{|A|}{2} \rfloor$, $l = \lfloor \frac{|B|}{2} \rfloor$ corresponds to the *modified median Hausdorff quasi-distance*.
- **Bottleneck distance**

  Given two binary images, seen as subsets $A, B \neq \emptyset$ with $|A| = |B| = m$, of a metric space (X,d), their **bottleneck distance** is defined by

  $$\min_{f} \max_{x \in A} d(x, f(x)),$$

where f is any bijective mapping between A and B. Cf. more general **matching distance** in Chap. 1.

Variations of the above distance are:

1. The **minimum weight matching**: $\min_f \sum_{x \in A} d(x, f(x))$;
2. The **uniform matching**: $\min_f \{\max_{x \in A} d(x, f(x)) - \min_{x \in A} d(x, f(x)\}$;
3. The **minimum deviation matching**: $\min_f \{\max_{x \in A} d(x, f(x)) - \frac{1}{|A|} \sum_{x \in A} d(x, f(x)\}$.

Given an integer t with $1 \le t \le |A|$, the t**-bottleneck distance** between A and B ([InVe00]) is the above minimum but with f being any mapping from A to B such that $|\{x \in A : f(x) = y\}| \le t$.

The cases $t = 1$ and $t = |A|$ correspond, respectively, to the bottleneck distance and **directed Hausdorff distance** $d_{dHaus}(A, B) = \max_{x \in A} \min_{y \in B} d(x, y)$ (Chap. 1).

- **Hausdorff distance up to** G

  Given a group $(G, \cdot, id)$ acting on the Euclidean space $\mathbb{E}^n$, the **Hausdorff distance up to** G between two compact subsets A and B (used in Image Processing) is their **generalized** G**-Hausdorff distance** (Chap. 1), i.e., the minimum of $d_{Haus}(A, g(B))$ over all $g \in G$. Usually, G is the group of all isometries or all translations of $\mathbb{E}^n$.
- **Hyperbolic Hausdorff distance**

  For any compact subset A of $\mathbb{R}^n$, denote by $MAT(A)$ its *Blum's medial axis transform*, i.e., the subset of $X = \mathbb{R}^n \times \mathbb{R}_{\ge 0}$, whose elements are all pairs $x = (x', r_x)$ of the centers x' and the radii r_x of the maximal inscribed (in A) balls, in terms of the Euclidean distance d_E in $\mathbb{R}^n$. (Cf. **medial axis and skeleton** transforms for the general case.)

  The **hyperbolic Hausdorff distance** ([ChSe00]) is the **Hausdorff metric** on nonempty compact subsets $MAT(A)$ of the metric space (X, d), where the *hyperbolic distance* d on X is defined, for its elements $x = (x', r_x)$ and $y = (y', r_y)$, by

$$\max\{0, d_E(x', y') - (r_y - r_x)\}.$$

- **Nonlinear Hausdorff metric**

  Given two compact subsets A and B of a metric space (X, d), their **nonlinear Hausdorff metric** (or *Szatmári–Rekeczky–Roska wave distance*) is the **Hausdorff distance** $d_{Haus}(A \cap B, (A \cup B)^*)$, where $(A \cup B)^*$ is the subset of $A \cup B$ which forms a closed contiguous region with $A \cap B$, and the distances between points are allowed to be measured only along paths wholly in $A \cup B$.
- **Handwriting spatial gap distances**

  Automatic recognition of unconstrained handwritten texts (for example, legal amounts on bank checks or pre-hospital care reports) require measuring the spatial gaps between connected components in order to extract words.

[redacted]

Three most used ones, among **handwriting spatial gap distances** between two adjacent connected components x and y of text line, are:

- Seni–Cohen, 1994: the *run-length* (minimum horizontal Euclidean distance) between x and y, and the horizontal distance between their bounding boxes;
- Mahadevan–Nagabushnam, 1995: Euclidean distance between the convex hulls of x and y, on the line linking hull centroids.

- **Video quality metrics**

  These metrics are between test and reference color video sequences, usually aimed at optimization of encoding/compression/decoding algorithms. Each of them is based on some perceptual model of the human vision system, the simplest ones being RMSE (root-mean-square error) and PSNR (peak signal-to-noise ratio) error measures. The *threshold metrics* estimate the probability of detecting an *artifact* (i.e., a visible distortion that gets added to a video signal during digital encoding).

  Examples are: Sarnoff's JND (just-noticeable difference), Winkler's PDM (perceptual distortion), and Watson's DVQ (digital video quality) metrics. DVQ is an **l_p-metric** between feature vectors representing two video sequences. Some metrics measure special artifacts in the video: the appearance of block structure, blurriness, added "mosquito" noise (ambiguity in the edge direction), texture distortion, etc.
- **Time series video distances**

  The **time series video distances** are objective wavelet-based spatial-temporal **video quality metrics**. A video stream x is processed into a time series $x(t)$ (seen as a curve on coordinate plane) which is then (piecewise linearly) approximated by a set of n contiguous line segments that can be defined by $n + 1$ endpoints (x_i, x'_i), $0 \le i \le n$, in the coordinate plane. In [WoPi99] are given the following (cf. **Meehl distance**) distances between video streams x and y:

  - $Shape(x,y) = \sum_{i=0}^{n-1} |(x'_{i+1} - x'_i) - (y'_{i+1} - y'_i)|$;
  - $Offset(x,y) = \sum_{i=0}^{n-1} |\frac{x'_{i+1}+x'_i}{2} - \frac{y'_{i+1}+y'_i}{2}|$.

- **Dynamic time wrapping distance**

  A *time series* is a temporal sequence, varied in time or speed, of numerical data points, collected usually at regular intervals over a period of time; see above **time series video distances** and, in Chap. 23, **spike train distances**.

  Dynamic time warping (DTW) is a sequence alignment method allowing a nonlinear mapping of one sequence to another by minimizing the (total cumulative) distance between them. Used originally in speech recognition, DTW is applied now to temporal sequences of video, audio, and graphics data.

  Consider two sequences (say, time series) $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_n$ over a distance space (S, d). The matrix $((DTW(i,j)))_{0\le i\le m}^{0\le j\le n}$ of *cumulative distances* is defined by $DTW(0,0) = 0$, all $DTW(0,i)$ and $DTW(i,0)$

are ∞, while

$$DTW(i,j) = d(x_i, y_j) + \min(DTW(i-1, j, DTW(i, j-1, DTW(i-1, j-1)),$$

for any $i,j \geq 1$, minimizing insertion, deletion, match, respectively. A *warping path* $W = (w_1, \dots, w_k)$ is a contiguous set of matrix elements w_i that represent a mapping between x and y. The *optimal* path is the one minimizing the *warping cost*; this minimum $DTW(x, y) = \min \sqrt{\sum_{1 \le i \le k} w_k^2}$ is called the **DTW-distance**.

The time series can be seen as discretizations of differentiable functions of time. Optimal nonlinear time warping functions are computed by minimizing a measure of distance of the set of functions to their warped average. Related *elastic matching* is an optimization problem of 2D warping specifying corresponding pixels between subjected images; cf. **nonlinear elastic matching distance** between curves.

21.2 Audio Distances

Sound is the vibration of gas or air particles that causes pressure variations within our eardrums. *Audio* (speech, music, etc.) *Signal Processing* is the processing of analog (continuous) or, mainly, digital representation of the air pressure waveform of the sound. A *sound spectrogram* (or *sonogram*) is a visual $3D$ representation of acoustic signal. It is obtained either by a series of bandpass filters (an analog processing), or by application of the *short-time Fourier transform* to the electronic analog of an acoustic wave. Three axes represent time, frequency and *intensity* (acoustic energy). Often this $3D$ curve is reduced to two dimensions by indicating the intensity with more thick lines or more intense gray or color values.

Sound is called *tone* if it is periodic (the lowest *fundamental* frequency plus its multiples, *harmonics* or *overtones*) and *noise*, otherwise. The frequency is measured in *cps* (the number of complete cycles per second) or Hz (Hertz). The range of audible sound frequencies to humans is typically 20 Hz to 20 kHz. A moth *Galleria mellonella* can hear up to 300 kHz, in order to locate predatory bats using ultrasound.

The *power* $P(f)$ of a signal is energy per unit of time; it is proportional to the square of signal's amplitude $A(f)$. *Decibel dB* is the unit used to express the relative strength of two signals. One tenth of 1 dB is *bel*, the original outdated unit.

The amplitude of an audio signal in *dB* is $20 \log_{10} \frac{A(f)}{A(f')} = 10 \log_{10} \frac{P(f)}{P(f')}$, where f' is a reference signal selected to correspond to 0 dB (usually, the threshold of human hearing). The threshold of pain is about 120–140 dB.

Pitch and *loudness* are auditory subjective terms for frequency and amplitude.

The *mel scale* is a perceptual frequency scale, corresponding to the auditory sensation of tone height and based on *mel*, a unit of pitch. It is connected to

[redacted]

the acoustic frequency f hertz scale by $Mel(f) = 1127 \ln(1 + \frac{f}{700})$ or, simply, $Mel(f) = 1000 \log_2(1 + \frac{f}{1000}$.

The *Bark scale* (named after Barkhausen) is a psycho-acoustic scale of frequency: it ranges from 0 to 24 *Bark* corresponding to the first 24 critical bands of hearing:

$0, 100, 200, \ldots, 1270, 1480, 1720, \ldots, 9500, 12{,}000, 15{,}500$ Hz.

Those bands correspond to spatial regions of the basilar membrane (of the inner ear), where oscillations, produced by the sound of given frequency, activate the hair cells and neurons. Our ears are most sensitive in 2000–5000 Hz. The Bark scale is connected to the acoustic frequency f kilohertz scale by $Bark(f) = 13 \arctan(0.76f) + 3.5 \arctan(\frac{f}{0.75})^2$.

Terrestrial vertebrates perceive frequency on a logarithmic scale, i.e., pitch perception is better described by frequency ratios than by differences on a linear scale. It is matched by the distribution of cells sensitive to different frequencies in their ears.

Power spectral density $PSD(f)$ of a wave is the power per Hz. It is the Fourier transform of the autocorrelation sequence. So, the power of the signal in the band $(-W, W)$ is given by $\int_{-W}^{W} PSD(f)df$. A *power law noise* has $PSD(f) \sim f^{\alpha}$. The noise is called *violet*, *blue*, *white*, *pink* (or $\frac{1}{f}$), *red* (or *brown(ian)*), *black* (or *silent*) if $\alpha = 2, 1, 0, -1, -2, < -2$. PSD changes by 3α dB per *octave* (i.e., with frequency doubling); it decreases for $\alpha < 0$.

Pink noise occurs in many physical, biological and economic systems; cf. **long range dependence** in Chap. 18. It has equal power in proportionally wide frequency ranges. Humans also process frequencies in a such logarithmic space (approximated by the Bark scale). So, every octave contains the same amount of energy. Thus pink noise is used as a reference signal in Audio Engineering. Steady pink noise (including light music) reduces brain wave complexity and improve sleep quality.

Intensity of speech signal goes up/down within a 3–8 Hz frequency which resonates with the theta rhythm of neocortex. The speakers produce 3–8 syllables per second.

The main way that humans control their *phonation* (speech, song, laughter) is by control over the *vocal tract* (the throat and mouth) shape. This shape, i.e., the cross-sectional profile of the tube from the closure in the *glottis* (the space between the vocal cords) to the opening (lips), is represented by the cross-sectional area function $Area(x)$, where x is the distance to the glottis. The vocal tract acts as a resonator during vowel phonation, because it is kept relatively open. These resonances reinforce the source sound (ongoing flow of lung air) at particular *resonant frequencies* (or *formants*) of the vocal tract, producing peaks in the *spectrum* of the sound.

Each vowel has two characteristic formants, depending on the vertical and horizontal position of the tongue in the mouth. The source sound function is modified by the frequency response function for a given area function. If the vocal tract is approximated as a sequence of concatenated tubes of constant cross-sectional area, then the *area ratio coefficients* are the ratios $\frac{Area(x_{i+1})}{Area(x_i)}$ for consecutive tubes; those coefficients can be computed by *LPC* (linear predictive coding).

The *spectrum* of a sound is the distribution of magnitude (dB) (and sometimes the phases) in frequency (kHz) of the components of the wave. The *spectral envelope* is a smooth contour that connects the spectral peaks. Its estimation is based on either LPC, or FFT (fast Fourier transform) using *real cepstrum*, i.e., the log amplitude spectrum.

FT (Fourier transform) maps time-domain functions into frequency-domain representations. The *complex cepstrum* of the signal $f(t)$ is $FT(\ln(FT(f(t) + 2\pi mi)))$, where m is the integer needed to unwrap the angle or imaginary part of the complex logarithm function. The FFT performs the Fourier transform on the signal and samples the discrete transform output at the desired frequencies usually in the *mel* scale.

Parameter-based distances used in recognition and processing of speech data are usually derived by LPC, modeling the speech spectrum as a linear combination of the previous samples (as in autoregressive processes). Roughly, LPC processes each word of the speech signal in the following 6 steps: filtering, energy normalization, partition into frames, *windowing* (to minimize discontinuities at the borders of frames), obtaining LPC parameters by the autocorrelation method and conversion to the *LPC-derived cepstral coefficients*. LPC assumes that speech is produced by a buzzer at the glottis (with occasionally added hissing and popping sounds), and it removes the formants by filtering.

The majority of distortion measures between sonograms are variations of **squared Euclidean distance** (including a covariance-weighted one, i.e., **Mahalanobis**, distance) and probabilistic distances belonging to following general types: generalized **variational distance**, **f-divergence** and **Chernoff distance**; cf. Chap. 14.

The distances for sound processing below are between vectors x and y representing two signals to compare. For recognition, they are a template reference and input signal, while for noise reduction they are the original (reference) and distorted signal (see, for example, [OASM03]). Often distances are calculated for small segments, between vectors representing short-time spectra, and then averaged.

- **SNR distance**

  Given a sound, let P and A_s denote its average power and RMS (root-mean-square) amplitude. The **signal-to-noise ratio in decibels** is defined by

  $$SNR_{dB} = 10\log_{10}(\frac{P_{signal}}{P_{noise}}) = P_{signal,dB} - P_{noise,dB} = 10\log_{10}(\frac{A_{signal}}{A_{noise}})^2.$$

  The *dynamic range* is such ratio between the strongest undistorted and minimum discernable signals. It is roughly 140 dB for human hearing, 40 dB for human speech and 80 dB for a music in a concert hall.

[redacted]

The *Shannon–Hartley theorem* express the *capacity* (maximal possible information rate) of a channel with additive *colored* (frequency-dependent) Gaussian noise, on the bandwidth B in Hz as $\int_0^B \log_2(1+\frac{P_{signal}(f)}{P_{noise}(f)})df$.

The **SNR distance** between signals $x=(x_i)$ and $y=(y_i)$ with n frames is

$$10\log_{10}\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i-y_i)^2}.$$

If M is the number of segments, the *segmented SNR* between x and y is defined by

$$\frac{10}{m}\sum_{m=0}^{M-1}\left(\log_{10}\sum_{i=nm+1}^{nm+n}\frac{x_i^2}{(x_i-y_i)^2}\right).$$

- **Spectral magnitude-phase distortion**

  The **spectral magnitude-phase distortion** between signals $x=x(\omega)$ and $y=y(\omega)$ is defined by

  $$\frac{1}{n}\left(\lambda\sum_{i=1}^n(|x(w)|-|y(w)|)^2+(1-\lambda)\sum_{i=1}^n(\angle x(w)-\angle y(w))^2\right),$$

  where $|x(w)|$, $|y(w)|$ are magnitude spectra, and $\angle x(w)$, $\angle y(w)$ are phase spectra of x and y, respectively, while the parameter $\lambda, 0\le\lambda\le 1$, is chosen in order to attach commensurate weights to the magnitude and phase terms. The case $\lambda=0$ corresponds to the **spectral phase distance**.

  Given a signal $f(t)=ae^{-bt}u(t)$, $a,b>0$ which has Fourier transform $x(w)=\frac{a}{b+iw}$, its *magnitude* (or *amplitude*) spectrum is $|x|=\frac{a}{\sqrt{b^2+w^2}}$, and its *phase* spectrum (in radians) is $\alpha(x)=\tan^{-1}\frac{w}{b}$, i.e., $x(w)=|x|e^{i\alpha}=|x|(\cos\alpha+i\sin\alpha)$.

  The **Fourier distance** and *Fourier phase distance* are $||FFT(x)-FFT(y)||_2$ and $||arg(FFT(x))-arg(FFT(y))||_2$, where the sums only contain the lower frequency terms of fast Fourier transform in order to reduce noise. The similar *wavelet distance* is based on the *discrete wavelet transform* separating low and high frequencies.

- **Spectral distances**

  Given two discrete spectra $x=(x_i)$ and $y=(y_i)$ with n channel filters, their **Euclidean metric** *EM*, **slope metric** *SM* (Klatt, 1982) and 2**-nd differential metric** 2*DM* (Assmann and Summerfield, 1989) are defined, respectively, by

  $$\sqrt{\sum_{i=1}^n(x_i-y_i)^2},\ \sqrt{\sum_{i=1}^n(x_i'-y_i')^2}\ \text{and}\ \sqrt{\sum_{i=1}^n(x_i''-y_i'')^2},$$

  where $z_i'=z_{i+1}-z_i$ and $z_i''=\max(2z_i-z_{i+1}-z_{i-1},0)$. Comparing, say, the auditory excitation patterns of vowels, *EM* gives equal weight to peaks and troughs although spectral peaks have more perceptual weight. *SM* emphasizes the formant frequencies, while 2*DM* sets to zero the spectral properties other than the formants.

The **RMS log spectral distance** (or *root-mean-square distance*, *quadratic mean distance*) $LSD(x,y)$ is defined by

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \ln y_i)^2}.$$

The corresponding l_1- and l_∞-distances are called *mean absolute distance* and *maximum deviation*. These three distances are related to decibel variations in the log spectral domain by the multiple $\frac{10}{\log 10}$. In fact, the *log spectral distance between power spectra* $x(\omega)$ and $y(\omega)$ is defined as (expressed in dB):

$$\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}(10\log_{10}\frac{x(\omega)}{y(\omega)})^2 d\omega}.$$

The square of $LSD(x,y)$, via the cepstrum representation $\ln x(\omega) = \sum_{j=-\infty}^{\infty} c_j e^{-j\omega i}$ (where $x(\omega)$ is the *power cepstrum* $|FT(\ln(|FT(f(t)))|^2)|^2)$ becomes, in the complex cepstral space, the **cepstral distance**.

The **log area ratio distance** $LAR(x,y)$ between x and y is defined by

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} 10(\log_{10} Area(x_i) - \log_{10} Area(y_i))^2},$$

where $Area(z_i)$ is the cross-sectional area of the i-th segment of the vocal tract.

- **Bark spectral distance**

  Let (x_i) and (y_i) be the *Bark spectra* of x and y, where the i-th component corresponds to the i-th auditory critical band in the Bark scale. The **Bark spectral distance** (Wang–Sekey–Gersho, 1992) is a perceptual distance, defined by

  $$BSD(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2,$$

  i.e., it is the **squared Euclidean distance** between the Bark spectra.

  A modification of the Bark spectral distance excludes critical bands i on which the loudness distortion $|x_i - y_i|$ is less than the noise masking threshold.

[redacted]

- **Itakura–Saito quasi-distance**

  The **Itakura–Saito** (or *maximum likelihood*) **quasi-distance** between LPC-derived spectral envelopes $x = x(\omega)$ and $y = y(\omega)$ is defined (1968) by

$$IS(x,y) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{x(w)}{y(w)} - \ln\frac{x(w)}{y(w)} - 1\right)dw.$$

  The **cosh distance** is defined by $IS(x,y) + IS(y,x)$, i.e., is equal to

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{x(w)}{y(w)} + \frac{y(w)}{x(w)} - 2\right)dw = \frac{1}{2\pi}\int_{-\pi}^{\pi} 2\cosh\left(\ln\frac{x(w)}{y(w)} - 1\right)dw,$$

  where $\cosh(t) = \frac{e^t+e^{-t}}{2}$ is the hyperbolic cosine function.

- **Log-likelihood ratio quasi-distance**

  The **log-likelihood ratio quasi-distance** between LPC-derived spectral envelopes $x = x(\omega), y = y(\omega)$ is defined (cf. **Kullback–Leibler distance** in Chap. 14) by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} x(w)\ln\frac{x(w)}{y(w)}dw.$$

  The **weighted likelihood ratio distance** between $x(\omega)$ and $y(\omega)$ is defined by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{\left(\ln\left(\frac{x(w)}{y(w)}\right) + \frac{y(w)}{x(w)} - 1\right)x(w)}{p_x} + \frac{\left(\ln\left(\frac{y(w)}{x(w)}\right) + \frac{x(w)}{y(w)} - 1\right)y(w)}{p_y}\right)dw,$$

  where $P(x)$ and $P(y)$ denote the power of the spectra $x(w)$ and $y(w)$.

- **Cepstral distance**

  The **cepstral distance** (or *squared Euclidean cepstrum metric*) $CEP(x,y)$ between the LPC-derived spectral envelopes $x = x(\omega)$ and $y = y(\omega)$ is defined by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\ln\frac{x(w)}{y(w)}\right)^2 dw = \frac{1}{2\pi}\int_{-\pi}^{\pi}(\ln x(w) - \ln y(w))^2\,dw = \sum_{j=-\infty}^{\infty}(c_j(x) - c_j(y))^2,$$

  where $c_j(z) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{jwi}\ln|z(w)|dw$ is j-th cepstral (real) coefficient of z derived from the Fourier transform or LPC.

  The **quefrency-weighted cepstral distance** (or **Yegnanarayana distance**, *weighted slope distance*) between x and y is defined by

$$\sum_{i=-\infty}^{\infty} i^2(c_i(x) - c_i(y))^2.$$

"Quefrency" and "cepstrum" are anagrams of "frequency" and "spectrum".

The **Martin cepstrum distance** between two ARMs (autoregressive models) is defined, in terms of their cepstra, by

$$\sqrt{\sum_{i=0}^{\infty} i(c_i(x) - c_i(y))^2}.$$

Cf. general **Martin distance** in Chap. 12 and **Martin metric** in Chap. 11.

- **Distances in Musicology**

  Pitch is a subjective correlate of the fundamental frequency. A *note* (or *tone*) is a named pitch. Pitch, seen as extending along a 1D continuum from high to low, is called *pitch height*. But it also varies circularly: a *pitch class* is a set of all pitches that are a whole number of *octaves* (intervals between a frequency and its double) apart. About 10 octaves cover the range of human hearing. In Western music, the most used octave division is the *chromatic scale*: 12 notes $C, C\#, D, D\#, E, F, F\#, G, G\#, A, A\#, B$ drawn usually as *pitch class space*: a circle of *equal temperament*, i.e., divided into 12 equal *semitones* (or *half steps*). The **distance between notes** whose frequencies are f_1, f_2 is $12\log_2(\frac{f_1}{f_2})$ semitones.

  An *interval* is the difference between two pitches. Its *width* is the ratio $\frac{a}{b}$ (with $\text{g.c.d}(a, b) = 1$) between their frequencies. The *Benedetti height* of this ratio is ab; *Tenney height* (or **Tenney harmonic distance**) is $\log_2 ab$ and *Kees height* is $\max(a', b')$, where a', b' come from a, b by removing factors of 2. The width of a semitone is $\sqrt[12]{2}$ or 100 cents. The width of octave is 2 or 1200 cents.

  A **pitch distance** (or *melodic distance*) is the size of the section of the pitch-continuum bounded by those two pitches, such as modeled in a given scale. A MIDI (Musical Instrument Digital Interface) number of fundamental frequency f is defined by $p(f) = 69 + 12\log_2 \frac{f}{440}$. The distance between notes, in terms of this *linear pitch space*, becomes the **natural metric** $|p(f_1) - p(f_2)|$ on $\mathbb{R}$. This pitch distance corresponds to physical distance on keyboard instruments, and psychological distance as measured by experiments and musicians.

  Using integer notation $0, 1, \dots, 11$ of pitches, a *pitch interval* $PI(x, y)$ between the pitches x and y is the number of semitones $|x-y|$ that separates them linearly, while a *pitch-interval class* $PIC(x, y)$ is $|x - y| \mod 12$ and an *interval class* $ic(x, y)$ is their **Lee distance** $\min |x - y|, 12 - |x - y|$ on the circle.

  In integer notation, the *circle of fifths* is $\{7i \mod 12\}_{0,\dots,11}$, and its reversal, the *circle of fourths*, is $\{5i \mod 12\}_{0,\dots,11}$. Neighboring pitches are separated by a *perfect fifth* (interval of 5 staff positions or 7 semitones).

  7 letters of a musical alphabet, C–D–E–F–G–A–B, are called the *natural tones*; they are the names of the white keys on a piano/keyboard, forming an octave. Above sequence and any of its translations is a *major diatonic scale*. A *diatonic scale* is a scale of 7 notes most used in Western music. Its structure is 1–1–0.5–1–1–1–0.5, in terms of interval succession of steps. A **distance model** (in Music) is the alternation of two different intervals to create a nondiatonic musical

[redacted]

mode/scale, such as the 1:3 distance model (alternation of semitones and minor thirds).

In tonal music, composition written in home key; it *modulates* (move to other keys) and usually returns. The distance between keys approximates the ease of modulation. Every key is associated with a scale of pitches, usually, major or *minor* diatonic. The **interkey distance** of two keys is 7 minus the number of tones shared by their scales. It is also their distance around the circle of fifths, i.e., the difference in the number of *sharps* (or *flats*) in their signatures. The *relative* (having the same signatures) major and minor key share all 7 notes.

A *chord* in music is any set of at least 3 pitch classes in the same octave that is heard as if sounding simultaneously. Music can be seen as a sequence of chords. *Interval vector* of a given chord c is $V(c) = (c_1, \dots, c_6)$, where c_i is the number of times i-th interval class (having i or $12 - i$ semitones) appears in it. **Intervalic distance** and **Estrada distance** between chord c and c' are (Mathieu, 2002):

$$\sum_{i=1}^{6} |c_i - c'_i| \text{ and } \max |c|, |c'| - |V(c) \cap V(c')| - 1.$$

The **root distance** is the number of fifths between the *roots* (pitches upon which a chord may be built, often by stacking thirds) of the chords. In [RRHD10], a survey of 8 **distances between chords** is given: above 3 and those by Chew (2000), Costère (1962), Lerdahl (2001), Paiement et al. (2005) and Yoo et al. (2006).

Alternatively to equal-temperement, *just intonation* is a tuning in which the frequencies of notes are related by ratios of small whole numbers, say, $\frac{3}{2}$ for perfect fifth (G) and $\frac{4}{3}$ for perfect forth (F). The pitches can be arranged in a $2D$ diagram. For an odd number $n > 0$, the *n-limit diagram* contains all rational numbers such that any odd divisor of the numerator or denominator is at most n. Such 5- and 7-limit can be seen as the *hexagonal lattice* $A_2 = \{(a, b, c) \in \mathbb{Z}^3 : a + b + c = 0\}$ and *face-centered cubic lattice* $A_3 = \{(a, b, c) \in \mathbb{Z}^3 : a + b + c \equiv 0(\mod 2)\}$, respectively, with vector space norms $\sqrt{a^2 + ab + b^2}$ and $\sqrt{a^2 + ab + b^2 + c(a + b + c)}$.

- **Distances between rhythms**

  A rhythm timeline (music pattern) is represented, besides the standard music notation, in the following ways, used in computational music analysis.

1. By a binary vector $x = (x_1, \dots, x_m)$ of m time intervals (equal in a metric timeline), where $x_i = 1$ denotes a beat, while $x_i = 0$ denotes a rest interval (silence). For example, the five $12/8$ metric timelines of Flamenco music are represented by five binary sequences of length 12.
2. By a *pitch vector* $q = (q_1, \dots, q_n)$ of absolute pitch values q_i and a *pitch difference vector* $p = (p_1, \dots, p_{n-1})$ where $p_i = q_{i+1} - q_i$ represents the number of semitones (positive or negative) from q_i to q_{i+1}.

3. By an *interonset interval vector* $t = (t_1, \dots, t_n)$ of n time intervals between consecutive onsets.
4. By a *chronotonic representation* which is a histogram visualizing t as a sequence of squares of sides $t_1, \dots, t_n$; it can be seen as a piecewise linear function.
5. By a *rhythm difference vector* $r = (r_1, \dots, r_{n-1})$, where $r_i = \frac{t_{i+1}}{t_i}$.

Examples of general **distances between rhythms** are the Hamming distance, **swap metric** (Chap. 11) and **Earth Mover's distance** between their given vector representations.

The **Euclidean interval vector distance** is the Euclidean distance between two interonset interval vectors. The Gustafson **chronotonic distance** is a variation of l_1-distance between these vectors using the chronotonic representation.

Coyle–Shmulevich **interval-ratio distance** is defined by

$$1 - n + \sum_{i=1}^{n-1} \frac{\max\{r_i, r'_i\}}{\min\{r_i, r'_i\}},$$

where r and r' are rhythm difference vectors of two rhythms (cf. the reciprocal of **Ruzicka similarity** in Chap. 17).

- **Long-distance drumming**

Long-distance drumming (or *drum telegraphy*) is an early form of long-distance communication which was used by cultures in Africa, New Guinea and the tropical America living in deforested areas. A rhythm could represent an signal, repeat the profile of a spoken utterance or simply be subject to musical laws.

The *message drums* (or *slit gongs*) were developed from hollow tree trunks. The sound could be understood at $\leq$ 8 km but usually it was relayed to a next village. Another oldest tools of audio telecommunication were *horns* (tapered sound guides providing an acoustic impedance match between a sound source and free air). Any mode of communication (as by means of drums or horns) for use beyond the range of the articulate voice, is called **distance language**.

Soldier termites of some species drum their heads (11 times per second) on the ground to signal danger. The initial vibrations travel 40 cm, but a chain of soldiers relay the resulting wave, moving 1.3 m/s, over much greater distances.

- **Sonority distance effect**

People in warm-climate cultures spend more time outdoors and engage, on average, in more distal oral communication. So, such populations have greater *sonority* (audibility) of their phoneme inventory. Munroe et al., 1996 and 2009, observed that speakers in such languages use more simple consonant-vowel syllables, vowels and *sonorant* (say, nasal "n", "m" rather than obstruents as "t", "g") consonants.

Ember and Ember, 2007, found that number of cold months, as well as the combination of cold climate and sparse vegetation, predicts less sonority. Larger

[redacted]

average distance of the baby from its caregivers, as well as higher frequency of premarital and extramarital sex predicts more sonority.

- **Vocal deviation**

  Vocal deviation is (Podos, 2001) the distance of birdsong performance to the upper performance limit. Performance is measured by a variable combining frequency bandwidth and *note repetition rate* (the number of notes per second).

  The **vocal deviation** of a bird is the minimal **point-line distance** (Chap. 4) of data points of its recordings from the (upper-bound regression) line representing performance limit.

  Fitch et al., 2014, found in hermit thrush's songs the same intervals of pitch as the common major chords (octave, perfect fifth, perfect third) in human music.

- **Acoustics distances**

  The *wavelength* of a wave is the distance it travels to complete one cycle. This distance is measured perpendicular to the wavefront in the direction of propagation between one peak of a *sine wave* (sinusoid) and the next corresponding peak. The wavelength of any frequency sound may be found by dividing the speed of sound (331.4 m/s at sea level) in the medium by the fundamental frequency.

  The **near field** (cf. Chap. 24) is the part of a sound field (usually within about two wavelengths from the source) where there is no simple relationship between sound level and distance. The **far field** (cf. Chap. 24) is the area beyond the near field boundary. It is comprised of the *reverberant field* and *free field*, where sound intensity decreases as $\frac{1}{d^2}$ with the distance d from the source. This law corresponds to a reduction of $\approx$ 6 dB in the sound level for each doubling of distance and to halving of loudness (subjective response) for each reduction of $\approx$ 10 dB.

  The **critical distance** (or *room radius*) is the distance from the source at which the direct sound and reverberant sound (reflected echo produced by the direct sound bouncing off, say, walls, floor, etc.) are equal in amplitude.

  The **pickup distance** of a microphone is the effective distance that it can be used at. For an electric guitar, it is the distance from *pickup* (transducer that captures mechanical vibrations) to strings.

  A directional microphone may be placed farther away from a desired sound source than an omnidirectional one of equal quality; the ratio of distances is called the **distance factor**.

  The *proximity effect (audio)* is the anomaly of low frequencies being enhanced when a directional microphone is very close to the source.

  Auditory **distance cues** (Chap. 28) are based on differences in loudness, spectrum, direct-to-reverb ratio and binaural ones. The closer sound object is louder, has more bass, high-frequencies, transient detail, dynamic contrast. Also, it appear wider, has more direct sound level over its reflected sound and has greater time delay between the direct sound and its reflections.

  The *acoustic metric* is the term used occasionally for some distances between vowels; for example, the Euclidean distance between vectors of formant frequencies of pronounced and intended vowel. Cf. **acoustic metric** in Physics (Chap. 24).

Chapter 22
Distances in Networks

22.1 Scale-Free Networks

A **network** is a graph, directed or undirected, with a positive number (weight) assigned to each of its arcs or edges. Real-world complex networks usually have a gigantic number N of vertices and are sparse, i.e., with relatively few edges.

Interaction networks (Internet, Web, social networks, etc.) tend to be **small-world** ([Watt99]), i.e., interpolate between regular geometric lattices and random graphs in the following sense. They have a large *clustering coefficient* (the probability that two distinct neighbors of a vertex are neighbors), as lattices in a local neighborhood, while the average path distance between two vertices is small, about $\ln N$, as in a random graph.

A **scale-free network** ([Bara01]) ia a network with probability distribution for a vertex to have degree k being similar to $k^{-\gamma}$, for some constant $\gamma > 0$ which usually belongs to the segment $[2, 3]$. This *power law* implies that very few vertices, called *hubs* (connectors, gateways, super-spreaders), are far more connected than other vertices. The power law (or **long range dependent**, *heavy-tail*) distributions, in space or time, has been observed in many natural phenomena (both physical and sociological).

- **Collaboration distance**

  The **collaboration distance** is the path metric of the *Collaboration graph*, having authors in Mathematical Reviews database as vertices with xy being an edge if authors x and y have a joint publication among the papers from this database.

  The vertex of largest degree (1416) corresponds to Paul Erdős; the *Erdős number* of a mathematician is his collaboration distance to Paul Erdős. An example of a 3-path: Michel Deza–Paul Erdős–Ernst Gabor Straus–Albert Einstein.

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_22

[redacted]

- **Co-starring distance**

  The **co-starring distance** is the path metric of the *Hollywood graph*, having about 250,000 vertices (actors in the Internet Movie database) with *xy* being an edge if the actors *x* and *y* appeared in a feature film together. The vertices of largest degree are Christopher Lee and Kevin Bacon; the trivia game *Six degrees of Kevin Bacon* uses the *Bacon number*, i.e., the co-starring distance to this actor.

  The *Morphy* and *Shusaku numbers* are the similar measures of a chess or Go player's connection to Paul Morphy and Honinbo Shusaku by way of playing games. *Kasparov number* of a chess-player is the length of a shortest directed path, if any, from him/her to Garry Kasparov; here arc *uv* means victory of *u* over v.

  Similar popular examples of such social scale-free networks are graphs of musicians (who played in the same rock band), baseball players (as team-mates), scientific publications (who cite each other), mail exchanges, acquaintances among classmates in a college, business board membership.

  Among other such studied networks are air travel connections, word co-occurrences in human language, US power grid, sensor networks, worm neuronal network, gene co-expression networks, protein interaction networks and metabolic networks (with two substrates forming an edge if a reaction occurs between them via enzymes).

- **WikiDistance**

  In 2015, Wikipedia had about 38 million articles in 281 languages and 74,000 active editors. English Wikipedia alone had 5 million articles (4 % of estimated number of notable articles needed to cover all human knowledge) and 814 billion of edits.

  The **WikiDistance** is the directed path quasi-metric of the *Wikipedia digraph*, having English Wikipedia articles as vertices, with *xy* being an arc if the article *x* contains an hyperlink to the article *y*; cf. http://software.tanos.co.uk/wikidistance and the **Web hyperlink quasi-metric**.

  Gabrilovich–Markovich, 2007, proposed to measure semantic relatedness of two texts by the **cosine distance** (cf. **Web similarity metrics**) between weighted vectors, interpreting texts in terms of affinity with a host of Wikipedia concepts.

  Crandall et al., 2008, considered the social network of Wikipedia editors: two editors are assumed to be connected if one of them posted to the other's discussion page. Brandes et al., 2009, considered the *edit network* of a Wikipedia page, where nodes are the authors of this page and edges correspond to undoing each other edits.

  The *editing depth* of Wikipedia is an indicator of its collaborativeness defined as $D = \frac{e}{a+n} \times (\frac{n}{a})^2$, where e, a, n are the numbers of page edits, articles and *nonarticles* (redirects, talk, user pages). At February 2016, English Wikipedia had $D = 916$.

- **Virtual community distance**

  Largest, in millions of active user accounts, *virtual communities* (online social networking services) are, as of December 2014, *Facebook* (1590), *WhatsApp* (1000), *Tencent QQ* (860), *Facebook Messenger* (800), *Tencent Ozone* (653), *WeChat* (650), *Google+* (540), *Istagram* (400), *Twitter* (305), *Skype* (300), all of origin in the US or China. 3.3 billion has used the services of the Internet within 2015.

In 2012, about 30 billion documents were uploaded on Facebook, 300 million tweets sent on Twitter and 24 petabytes of data processed by Google *per day*, while mankind published only $\approx$5000 petabytes for the 20,000 years before 2003.

A **virtual community distance** is the path metric of the graph of active users, two of them forming an edge if they are "friends". In Twitter it means that both "follow" each other. In particular, for the **Facebook hop distance** in November 2011, 99.6 % of all pairs of users were connected by paths of length at most 5. The mean distance was 4.74, down from 5.28 in 2008.

The **Twitter friendship distance** in Avril 2010 was $4, 5, 6$ among 37 %, 41 %, 13 % of 5.2 billion friendships. The average distance was 4.67 steps. Cf. mean distance 5.2 in Milgram's (1967) theory of *six degrees of separation* on a planetary scale.

An example of application: analysing linguistically emotional content of tweets and comments, one can obtain an interaction graph of the targeted region mood.

- **Distance effect in large e-mail networks**

  Takhteyev–Gruzd–Wellman, 2012, considered a sample representing Twitter *tie* (i.e., "follow" relation in both directions) network. They found that distance constrains ties, despite the seeming ease with which they can be formed: 39 % of the ties are shorter than 100 km (within the same regional cluster), ties up to 1000 km are more frequent than random ones, and ties longer than 5000 km are rare. Cf. **distance decay** in Chap. 28. But the nonlocal ties are predicted better by the frequency of airline connections than by physical proximity.

  State et al., 2013, started with a graph of a sample about 10 million users of Yahoo! email with an edge between two users whenever they exchanged at least one email message in both directions, during the observation period in 2012. A weighted complete graph of 141 countries was derived, with edge-weight being the rescaled logarithm of the communication density between countries. For each doubling of distance (between each country centroids) and doubling of the number of direct flights, the density decreased by 66 % and increased by 33 %, respectively.

  But the main (besides colonial link and common language) cultural factor, nearly doubling the density, happens to be the common membership in the same civilization from the list produced by Huntigton in a 1993 article *The Clash of Civilizations*: Latin American, Islamic, Orthodox, Sinic, Buddist, Western, African, Hindu, Japonic. For Latin American, Islamic, and Orthodox civilization, this factor increases the density by the factor of 5.4, 3.1 and 2.4, respectively.

- **Network's hidden metric**

  Many social, biologic, and communication networks, including the Internet and Web, are scale-free and *strongly clustered* (many triangular subgraphs).

[redacted]

Greedy routing is a navigation strategy to do always the locally optimal step with the hope of finding a globally shortest path. Krioukov et al., 2009, found that successful greedy paths are shortest, mostly and asymptotically, in the large complex networks.

They explain such efficiency by the existence of a **hidden metric space** (V, d) on the set V of nodes, so that a node passes information to the neighbor that is closest in (V, d) to the final destination. Moreover, they suggest that (V, d) is **hyperbolic**, because the nodes are *heterogeneous* (can be classified into groups, subgroups, and so on) implying a tree-like structure of such network.

- **Sexual distance**

  Given a group of people, its *sexual network* is the graph of members two of them forming an edge if they had a sexual contact. The **sexual distance** is the path metric of a sexual network. Such networks of heterosexual individuals are usually scale-free but not small-world since they have no 3-cycles and very few 4-cycles.

  Several sexual networks were mapped in order to trace the spread of sexually infectious diseases. The sexual network of all adults aged 18–35 in Licoma (almost isolated island 18 km^2 on lake Malawi) have a giant connected component containing half of nonisolated vertices, and more than one quarter were connected *robustly*, i.e., by multiple disjoint paths. Also, in the sexual network of students of an Midwestern US high school, 52 % of nonisolated vertices belong to a giant connected component. But this graph contains very few cycles and have large diameter (37).

  A study of persons at risk for HIV (Colorado Springs, 1988–1992) compared their sexual and geographical distance, measured as the actual distance between their residences. The closest (at mean 2.9 km) pairs were HIV-positive persons and their contacts. The most distant (at mean 6.1 km) pairs were prostitutes and their paying partners. The mean distance between all persons in Colorado Springs was 12.4 km compared with 5.4 km between all dyads the study.

  Moslonka-Lefebre et al., 2012, consider weighted sexual networks, where the weight of an edge is the number of sex acts that are actually realized between two individuals per, say, a week. Such model is more consistent with epidemiological data.

  The sexual network for the human race have a giant connected component containing many vertices of degree 1 and almost all vertices of larger degree.

  A dating application Tinder, linked to Facebook, has an estimated 50 million users since 2012; it generates 15 million mutual matches a day.

- **Subway network core**

  Roth et al., 2012, observed that the world's largest subway networks converge to a similar shape: a core (ring-shaped set of central stations) with quasi-1D/linear branches radiating from it. The average degree of core stations is 2.5; among them $\approx 20\,\%$ are transfer stations and $> 60\,\%$ have degree 2.

  The average *radial* (from the geographical centroid of all stations) distance (in km) to branched stations is about double of such distance to core stations, while the number of branches scales roughly as the square root of the number of stations.

Cf. **Moscow metric**, **Paris metric**, and **subway semimetric** in Chap. 19.

- **Normalized web distance**

  The **normalized web distance** (or *normalized Google distance*, as it was derived by Cilibrasi and Vitanyi, 2007, from [BGLVZ98]) is a semantic distance, not a metric, between search terms induced by a large data base (say, the Internet) and a search engine (say, Google). It is applicable to any data-base search engine pair. This distance between two search terms x and y is defined by

  $$NWD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}},$$

  where $f(x)$ is the number of pages containing x, the frequency $f(x,y)$ is the number of pages containing both x and y, and N is the total number of indexed pages.

  Cf. **normalized information distance** in Chap. 11.

- **Drift distance**

  The **drift distance** is the absolute value of the difference between observed and actual coordinates of a node in a NVE (Networked Virtual Environment).

  In models of such large-scale peer-to-peer NVE (for example, Massively Multiplayer Online Games), the users are represented as coordinate points on the plane (*nodes*) which can move at discrete *time-steps*, and each has a visibility range called the *Area of Interest*. NVE creates a synthetic $3D$ world where each user assumes *avatar* (a virtual identity) to interact with other users or computer AI.

  The primary metric tool in MMOG and Virtual Worlds is the proximity sensor recording when an avatar is within its specified range.

  The term **drift distance** is also used for the current going through a material, in tire production, etc.

- **Betweenness centrality**

  For a **geodesic** metric space (X,d) (in particular, for the path metric of a graph), the **stress centrality** of a point $x \in X$ is defined (Shimbel, 1953) by

  $$\sum_{y,z \in X, y \neq x \neq z} \text{Number of shortest } (y-z) \text{ paths through } x,$$

  the **betweenness centrality** of a point $x \in X$ is defined (Freeman, 1977) by

  $$g(x) = \sum_{y,z \in X, y \neq x \neq z} \frac{\text{Number of shortest } (y-z) \text{ paths through } x}{\text{Number of shortest } (y-z) \text{ paths}},$$

[redacted]

and the **distance-mass function** is a function $M : \mathbb{R}_{\geq 0} \to \mathbb{Q}$ defined by

$$M(a) = \frac{|\{y \in X : d(x,y) + d(y,z) = a \text{ for some } x, y \in X\}|}{|\{(x,z) \in X \times X : d(x,z) = a\}|}.$$

[GOJKK02] estimated that $\frac{M(a)}{a} \approx 4.5$ for the **Internet AS metric**, and ≈ 1 for the **Web hyperlink quasi-metric** for which the shortest paths are almost unique.

- **Distance centrality**

  Given a finite metric space (X, d) (usually, the **path metric** on the graph of a network) and a point $x \in X$, we give here examples of metric functionals used to measure **distance centrality**, i.e., the amount of centrality of the point x in X expressed in terms of its distances $d(x, y)$ to other points.

  1. The **eccentricity** (or *Koenig number*) $\max_{y \in X} d(x,y)$ was given in Chap. 1; Hage-Harary, 1995, considered $\frac{1}{\max_{y \in X} d(x,y)}$.
  2. The **closeness** (Sabidussi, 1966) is the inverse $\frac{1}{\sum_{y \in X} d(x,y)}$ of the *farness*.
  3. Dangalchev, 2006, introduced $\sum_{y \in X, y \neq x} 2^{-d(x,y)}$ which allows the case $d(x,y) = \infty$ (disconnected graphs).
  4. The functions $f_1 = \sum_{y \in X} d(x,y)$ and $f_2 = \sum_{y \in X} d^2(x,y)$; cf. **Fréchet mean** in Chap. 1.

  In *Location Theory* applications, $X' \subset X$ is a set of positions of "clients" and one seeks points $x \in X$ of acceptable facility positions. The appropriate objective function is, say, $\min \max_{y \in X'} d(x,y)$ to locate an emergency service, $\min \sum_{y \in X'} d(x,y)$ for a goods delivering facility and $\max \sum_{y \in X'} d(x,y)$ for a hazardous facility.

22.2 Network-Based Semantic Distances

Among the main lexical networks (such as WordNet, Framenet, Medical Search Headings, Roget's Thesaurus) a semantic lexicon WordNet is the most popular lexical resource used in Natural Language Processing and Computational Linguistics.

WordNet (see http://wordnet.princeton.edu) is an online lexical database in which English nouns, verbs, adjectives and adverbs are organized into *synsets* (synonym sets), each representing one underlying lexical concept.

Two synsets can be linked semantically by one of the following links: upwards x (*hyponym*) *IS-A* y (*hypernym*) link, downwards x (*meronym*) *CONTAINS* y (*holonym*) link, or a horizontal link expressing frequent co-occurrence (*antonymy*), etc. *IS-A* links induce a partial order, called *IS-A taxonomy*. The version 2.0 of WordNet has 80,000 noun concepts and 13,500 verb concepts, organized into 9 and 554 separate *IS-A* hierarchies.

In the resulting DAG (*directed acyclic graph*) of concepts, for any two synsets (or concepts) x and y, let $l(x, y)$ denote the length of the shortest path between them, using only *IS-A* links, and let $LPS(x, y)$ denote their *least common subsumer* (ancestor) by *IS-A* taxonomy. Let $d(x)$ denote the *depth* of x (i.e., its distance from the root in *IS-A* taxonomy) and let $D = \max_x d(x)$.

The semantic relatedness of two nouns can be estimated by their **ancestral path distance** (cf. Chaps. 10 and 23), i.e., the length of the shortest *ancestral path* (concatenation of two directed paths from a common ancestor) to them). A list of the other main semantic similarities and distances follows. See also [HRJM13].

- **Length similarities**

  The **path similarity** and **Leacock–Chodorow similarity** between synsets x and y are defined by

  $$path(x, y) = (l(x, y))^{-1} \text{ and } lch(x, y) = -\ln \frac{l(x, y)}{2D}.$$

  The **conceptual distance** between x and y is defined by $\frac{l(x,y)}{D}$.
- **Wu–Palmer similarity**

  The **Wu–Palmer similarity** between synsets x and y is defined by

  $$wup(x, y) = \frac{2d(LPS(x, y))}{d(x) + d(y)}.$$

- **Resnik similarity**

  The **Resnik similarity** between synsets x and y is defined by

  $$res(x, y) = -\ln p(LPS(x, y)),$$

  where $p(z)$ is the probability of encountering an instance of concept z in a large corpus, and $-\ln p(z)$ is called the *information content* of z.
- **Lin similarity**

  The **Lin similarity** between synsets x and y is defined by

  $$lin(x, y) = \frac{2 \ln p(LPS(x, y))}{\ln p(x) + \ln p(y)}.$$

- **Jiang–Conrath distance**

  The **Jiang–Conrath distance** between synsets x and y is defined by

  $$jcn(x, y) = 2 \ln p(LPS(x, y)) - (\ln p(x) + \ln p(y)).$$

- **Lesk similarities**

  A *gloss* of a synonym set z is the member of this set giving a definition or explanation of an underlying concept. The **Lesk similarities** are those defined by a function of the overlap of glosses of corresponding concepts; for example,

[redacted]

the **gloss overlap** is

$$\frac{2t(x,y)}{t(x)+t(y)},$$

where $t(z)$ is the number of words in the synset z, and $t(x,y)$ is the number of common words in x and y.

- **Hirst–St-Onge similarity**

  The **Hirst–St-Onge similarity** between synsets x and y is defined by

  $$hso(x,y) = C - L(x,y) - ck,$$

  where $L(x,y)$ is the length of a shortest path between x and y using all links, k is the number of changes of direction in that path, and C, c are constants.

  The **Hirst–St-Onge distance** is defined by $\frac{L(x,y)}{k}$.

- **Semantic biomedical distances**

  The **semantic biomedical distances** are the distances used in biomedical lexical networks. The main clinical terminologies are UMLS (United Medical Language System) and SNOMED (Systematized Nomenclature of Medicine) CT.

  An example of such distances used in SNOMED and presented in Melton et al., 2006, is given by the *interpatient distance* between two *medical cases* (sets X and Y of patient data). It is their **Tanimoto distance** (Chap. 1) $\frac{|X \Delta Y|}{|X \cup Y|}$.

  The *conceptual distance* between two biomedical concepts in UMLS is (Caviedes and Cimino, 2004) the minimum number of *IS-A* parent links between them in the directed acyclic graph of *IS-A* taxonomy of concepts.

- **Semantic proximity**

  For the words in a document, there are short range syntactic relations and long range *semantic correlations*, i.e., meaning correlations between concepts.

  The main document networks are Web and bibliographic databases (digital libraries, scientific databases, etc.); the documents in them are related by, respectively, hyperlinks and citation or collaboration.

  Also, some semantic tags (keywords) can be attached to the documents in order to index (classify) them: terms selected by author, title words, journal titles, etc.

  The **semantic proximity** between two keywords x and y is their **Tanimoto similarity** $\frac{|X \cap Y|}{|X \cup Y|}$, where X and Y are the sets of documents indexed by x and y, respectively. Their **keyword distance** is defined by $\frac{|X \Delta Y|}{|X \cap Y|}$; it is not a metric.

- **Dictionary digraph**

  Dictionary digraph (V,E) have the words of a given dictionary as vertices, and arcs $uv \in E$ whenever word u is used to define word v. The *kernel* (V',E') is its subdigraph induced by the vertices with out-degree $\neq 0$. MF (*minimum feedback vertex set*) is a smallest set of vertices, from which any $v \in V$ can be reached.

Picard et al., 2013, found that $|V| \approx 10|V'| \approx 20|MF| \approx 20|V''|$ in such digraphs for four English dictionaries; here (V'', E'') is the *core* (largest strongly connected component) of the kernel. They observed that the words in the kernel V' are learned at a much younger age, and are more concrete, imageable and frequent than the words in $V \setminus V'$. The same is true, but more so, comparing V'' with $V \setminus V''$ and any MF with $V \setminus MF$. Cf. **Swadesh similarity** (Chap. 28).

- **SimRank similarity**

Let D be a directed multigraph representing a cross-referred document corpus (say, a set of citation-related scientific papers, hyperlink-related web pages, etc.) and $I(v)$ be the set of in-neighbors of a vertex v.

SimRank similarity $s(x, y)$ between vertices x and y of D is defined (Jeh and Widom, 2002) as 1 if $x = y$, 0 if $|I(x)||I(y)| = 0$ and, otherwise, as

$$\frac{C}{|I(x)||I(y)|} \sum_{a \in I(x), b \in I(y)} s(a, b),$$

where C is a constant, $0 < C < 1$ (usually, $C = 0.8$ or 0.6 is used).

- ***D*-separation in Bayesian network**

A *Bayesian network* is a DAG (digraph with no directed cycles) (V, E) whose vertices represent random variables and arcs represent conditional dependencies; so, the likelihood of each vertex can be calculated from the likelihood of its ancestors. Bayesian networks, including *causal networks*, are used for modeling knowledge.

A vertex $v \in V$ is called a *collider* of a *trail* (undirected path) t if there are two consecutive arcs $uv, vu' \in E$ on t. A trail t is *active by* a set $Z \subset V$ of vertices if every its collider is or has a descendent in Z, while every other vertex along t is outside of Z. If $X, Y, Z \subset V$ are disjoint sets of vertices, then Z is said (Pearl, 1988) to ***d*-separate** X from Y if there is no active trail by Z between a vertex in X and a vertex in Y. Such d-separation means that the variable sets, represented by X and Y, are independent conditional on variables, represented by Z, in all probability distributions the DAG (V, E) can represent.

The minimal set which d-separates vertex v from all other vertices is v's *Markov blanket*; it consists of v's parents, its children, and its children's parents. A *moral graph* of the DAG (V, E), used to find its equivalent undirect form, is the graph (V, E'), where E' consists all arcs from E made undirected plus all missing *marriages* (edges between vertices having a common child).

Cf. the **Bayesian graph edit distance** in Chap. 15.

- **Forward quasi-distance**

In a directed network, where edge-weights correspond to a point in time, the **forward quasi-distance** (**backward quasi-distance**) is the length of the shortest directed path, but only among paths on which consecutive edge-weights are increasing (decreasing, respectively).

The forward quasi-distance is useful in epidemiological networks (disease spreading by contact, or, say, heresy spreading within a church), while the

[redacted]

backward quasi-distance is appropriated in P2P (i.e., peer-to-peer) file-sharing networks.

Berman, 1996, introduced *scheduled network*: a directed network (of, say, airports), in which each edge (say, flight) is labeled by departure and arrival times. Kempe–Kleinberg–Kumar, 2002, defined more general *temporal network*: an edge-weighted graph, in which the weight of an edge is the time at which its endpoints communicated. A path is *time-respecting* if the weights of its edges are nondecreasing. Besides Scheduling and Epidemiology, such networks occur in Distributed Systems (say, dissemination of information using node-to-node communication).

In order to handle large temporal data on human behavior, Kostakos, 2009, introduced *temporal graph*: an arc-weighted directed graph, where the vertices are instances $a_i t_k$ (person a_i in point t_k of time), and the arcs are $(t_{k+1} - t_k)$-weighted ones $(a_i t_k, a_i t_{k+1})$ linking time-consecutive pairs and unweighted ones $(a_i t_k, a_j t_k)$ representing a communication (say, e-mail) from a_i to a_j at time t_k.

In order to handle *temporally disconnected* (not connected by a time-respecting path) nodes, Tang et al., 2009, defined *time-varying network*: an ordered set $\{D_t\}_{t=1,\dots,T}$ of directed (or not) graphs $D_t = (X, E_t)$, where the arc-sets E_t may change in time and the arcs have temporal duration. As real-world examples, they considered brain cortical and social interaction networks.

22.3 Distances in Internet and Web

Let us consider in detail the graphs of the Web and of its hardware substrate, Internet which are small-world and scale-free.

The *Internet* is the largest WAN (wide area network), spanning the Earth. This publicly available worldwide computer network came from 13-node ARPANET (started in 1969 by US Department of Defense), NSFNet, Usenet, Bitnet, and other networks. In 1995, the National Science Foundation in the US gave up the stewardship of the Internet, and in 2009, US Department of Commerce accepted privatization/internationalization of ICANN, the body responsible for domain names in the Internet.

Its nodes are *routers*, i.e., devices that forward packets of data along networks from one computer to another, using IP (Internet Protocol relating names and numbers), TCP and UDP (for sending data), and (built on top of them) HTTP, Telnet, FTP and many other *protocols* (i.e., technical specifications of data transfer). Routers are located at *gateways*, i.e., places where at least two networks connect.

The links that join the nodes together are various physical connectors, such as telephone wires, optical cables and satellite networks. The Internet uses *packet switching*, i.e., data (fragmented if needed) are forwarded not along a previously established path, but so as to optimize the use of available *bandwidth* (bit rate, in million bits per second) and minimize the *latency* (the time, in milliseconds, needed for a request to arrive).

Each computer linked to the Internet is usually given a unique "address", called its *IP address*. The new Internet Protocol IPv6 has address space $2^{128} \approx 4.4 \times 10^{38}$. The most popular applications supported by the Internet are e-mail, file transfer, Web, and some multimedia as YouTube and Internet TV. In 2012, 144 billions e-mails (68.8 % of which was spam) were sent daily by 2.2 billions users worldwide. In 2015, global IP traffic will reach 1.0 zettabytes (1000^7 bytes) per year.

The *Internet IP graph* has, as the vertex-set, the IP addresses of all computers linked to the Internet; two vertices are adjacent if a router connects them directly, i.e., the passing datagram makes only one *hop*. The Internet also can be partitioned into ASs (administratively Autonomous Systems). Within each AS the intradomain routing is done by IGP (Interior Gateway Protocol), while interdomain routing is done by BGP (Border Gateway Protocol) which assigns an ASN (16-bit number) to each AS. The *Internet AS graph* has ASs (about 42,000 in 2012) as vertices and edges represent the existence of a BGP peer connection between corresponding ASs.

The *World Wide Web* (*WWW* or *Web*, for short) is a major part of Internet content consisting of interconnected documents (resources). It corresponds to HTTP (Hyper Text Transfer Protocol) between browser and server, HTML (Hyper Text Markup Language) of encoding information for a display, and URLs (Uniform Resource Locators), giving unique "addresses" to web pages. The Web was started in 1989 in CERN which gave it for public use in 1993. The *Web digraph* is a virtual network, the nodes of which are *documents* (i.e., static HTML pages or their URLs) which are connected by incoming or outcoming HTML *hyperlinks*, i.e., hypertext links. It was at least 4.64 billion nodes (pages) in the Indexed Web digraph in May 2014.

The number of operating *web sites* (collections of related web pages found at a single address) reached 634 million in 2012 from 18,957 in 1995. In 2012, 54.7 % of websites were in English, followed by 5.9 %, 5.7 % in Russian and German. Along with the Web lies the *Deep* (or *Invisible*) *Web*, i.e., content, which is not indexed by standard search engines. This content (say, unlinked, or having dynamic URL, non-HTML/text, technically limited access, or scripted, requiring registration/login) has (Bergman, 2001) about 3000 times more pages than Surface Web, where Internet searchers are searching.

There are several hundred thousand *cyber-communities*, i.e., clusters of nodes of the Web digraph, where the link density is greater among members than between members and the rest. The cyber-communities (a customer group, a social network, a concept in a technical paper, etc.) are usually focused around a definite topic and contain a bipartite *hubs-authorities* subgraph, where all hubs (guides and resource lists) point to all authorities (useful and relevant pages on the topic).

Examples of new media, created by the Web are *(we)blogs* (digital diaries posted on the Web), Skype (telephone calls), social sites (as Facebook, Twitter, Linkedin) and Wikipedia (the collaborative encyclopedia). Original Web-as-information-source is often referred as *Web* 1.0, while *Web* 2.0 means present Web-as-participation-platform as, for example, web-based communities, blogs, social-networking (and video-sharing) sites, wikis, hosted services and web applications. For example, with *cloud servers* one can access his data and applications from the Internet rather than having them housed on-site.

[redacted]

Web 3.0 is the third generation of WWW conjectured to include semantic tagging of content. The project Semantic Web by W3C (WWW Consortium) aims at linking to metadata, merging social data and (making all things addressable by the existing naming protocols) transformation of WWW into GGG (Giant Global Graph) of users.

The *Internet of Things* refers to uniquely identifiable objects (things) and their virtual representations in an Internet-like structure. It would encode geographic location and dimensions of 50–100 trillion objects, and be able to follow their movement and send data between them. Every human being is surrounded by 1000–5000 objects.

On average, nodes of the Web digraph are of size 10 kilobytes, out-degree 7.2, and probability k^{-2} to have out-degree or in-degree k. A study in [BKMR00] of over 200 million web pages gave, approximately, the largest connected component "core" of 56 million pages, with another 44 million of pages connected to the core (newcomers?), 44 million to which the giant core is connected (corporations?) and 44 million connected to the core only by undirected paths or disconnected from it. For randomly chosen nodes x and y, the probability of the existence of a directed path from x to y was 0.25 and the average length of such a shortest path (if it exists) was 16, while maximal length of a shortest path was over 28 in the core and over 500 in the whole digraph.

A study in [CHKSS07] of Internet AS graphs revealed the following *Medusa structure* of the Internet: "nucleus" (diameter 2 cluster of $\approx$ 100 nodes), "fractal" ($\approx$15,000 nodes around it), and "tentacles" ($\approx$5000 nodes in isolated subnetworks communicating with the outside world only via the nucleus).

The distances below are examples of host-to-host **routing metrics**, i.e., values used by routing algorithms in the Internet, in order to compare possible routes. Examples of other such measures are: bandwidth consumption, communication cost, reliability (probability of packet loss). Also, the main computer-related *quality metrics* are mentioned.

- **Distance-vector routing protocol**

  A **distance-vector routing protocol** (DVRP) requires that a router informs its neighbors of topology changes periodically and, in some cases, when a change is detected in the topology of a network. Routers are advertised as vectors of a distance (say, **Internet IP metric**) and direction, given by next hop address and exit interface. Cf. **displacement** in Chap. 24.

  Ad hoc on-demand distance-vector routing is a (both unicast and multicast) routing protocol for mobile and other wireless ad hoc networks. It establishes a route to a destination only on demand and avoids the counting-to-infinity problem of other distance-vector protocols by using sequence numbers on route updates.

  Between nodes of an ad hoc network with end-to-end delay constraints, head-of-line packets compete for access to the shared medium. Each packet with remaining lifetime T and remaining **Internet IP metric** H to its destination, is associated with a ranking function $\gamma(H, T) = \frac{T^{\alpha}}{H}$, denoting its transmission priority. The number $\alpha \geq 0$ is called **lifetime-distance factor**; it should be optimized in order to minimize the probability of packet loss due to excessive delay.

- **Internet IP metric**

  The **Internet IP metric** (or **hop count**, *RIP metric*, *IP path length*) is the path metric in the *Internet IP graph*, i.e., the minimal number of hops (or, equivalently, routers, represented by their IP addresses) needed to forward a packet of data.

  RIP (a **distance-vector routing protocol** first defined in 1988) imposes a maximum distance of 15 and advertises by 16 nonreachable routes.

- **Internet AS metric**

  The **Internet AS metric** (or *BGP-metric*) is the **path metric** in the *Internet AS graph*, i.e., the minimal number of ISPs (Independent Service Providers), represented by their ASs, needed to forward a packet of data.

- **Geographic distance**

  The **geographic distance** is the **great circle distance** (Chap. 25) on the Earth from the client x (destination) to the server y (source).

  However, for economical reasons, the data often do not follow such geodesics; for example, most data from Japan to Europe transits via US.

- **RTT-distance**

  The **RTT-distance** (or *ping time*) is the round-trip time (to send a packet and receive an acknowledgment back) of transmission between x and y, measured in milliseconds (usually, by the *ping* command).

  See [HFPMC02] for variations of this distance and connections with the above three metrics. Fraigniaud–Lebbar–Viennot, 2008, found that RTT is a **C-inframetric** (Chap. 1) with $C \approx 7$.

- **Synchcronization distance**

  In the Network Time Protocol (NTP), the **synchcronization distance** is the *root dispersion* (maximum error relative to the primary reference source at the root of the synchronization subnet) plus one half the *root delay* (total round-trip delay to the primary reference source at the root of the synchronization subnet).

- **Administrative cost distance**

  The **administrative cost distance** is the nominal number (rating the trustworthiness of a routing information), assigned by the network to the route between x and y. For example, Cisco Systems assigns values $0, 1, \dots, 200, 255$ for the Connected Interface, Static Route, …, Internal BGP, Unknown, respectively.

- **DRP-metrics**

  The DD (Distributed Director) system of Cisco uses (with priorities and weights) the **administrative cost distance**, the **random metric** (selecting a random number for each IP address) and the **DRP** (Direct Response Protocol) metrics. DRP-metrics ask from all DRP-associated routers one of the following distances:

  1. The **DRP-external metric**: the number of BGP (Border Gateway Protocol) hops between the client requesting service and the DRP server agent;
  2. The **DRP-internal metric**: the number of IGP hops between the DRP server agent and the closest border router at the edge of the autonomous system;

[redacted]

3. The **DRP-server metric**: the number of IGP hops between the DRP server agent and the associated server.

- **Reported distance**

  In a Cisco Systems routing protocol EIGRP, **reported distance** (or *RD*, *advertised distance*) is the total metric along a path to a destination network as advertised by an upstream neighbor. RD is equal to the current lowest total distance through a successor for a neighboring router.

  A **feasible distance** is the lowest known distance from a router to a particular destination. This is RD plus the cost to reach the neighboring router from which the RD was sent; so, it is a historically lowest known distance to a particular destination.

- **Network tomography metrics**

  Consider a network with fixed routing protocol, i.e., a *strongly connected* digraph $D = (V, E)$ with a unique directed path $T(u, v)$ selected for any pair (u, v) of vertices. The routing protocol is described by a binary *routing matrix* $A = ((a_{ij}))$, where $a_{ij} = 1$ if the arc $e \in E$, indexed i, belongs to the directed path $T(u, v)$, indexed j. The **Hamming distance** between two rows (columns) of A is called the **distance between corresponding arcs** (directed paths) of the network.

  Consider two networks with the same digraph, but different routing protocols with routing matrices A and A', respectively. Then a **routing protocol semi-metric** ([Vard04]) is the smallest Hamming distance between A and a matrix B, obtained from A' by permutations of rows and columns (both matrices are seen as strings).

- **Web hyperlink quasi-metric**

  The **Web hyperlink quasi-metric** (or *click count*) is the length of the shortest directed path (if it exists) between two web pages (vertices in the Web digraph), i.e., the minimal number of necessary mouse-clicks in this digraph.

- **Average-clicks Web quasi-distance**

  The **average-clicks Web quasi-distance** between two web pages x and y in the Web digraph ([YOI03]) is the minimum $\sum_{i=1}^{m} \ln p \frac{z_i^+}{\alpha}$ over all directed paths $x = z_0, z_1, \dots, z_m = y$ connecting x and y, where z_i^+ is the out-degree of the page z_i. The parameter α is 1 or 0.85, while p (the average out-degree) is 7 or 6.

- **Dodge–Shiode WebX quasi-distance**

  The **Dodge–Shiode WebX quasi-distance** between two web pages x and y of the Web digraph is the number $\frac{1}{h(x,y)}$, where $h(x, y)$ is the number of shortest directed paths connecting x and y.

- **Web similarity metrics**

  Web similarity metrics form a family of indicators used to quantify the extent of relatedness (in content, links or/and usage) between two web pages x and y.

  Some examples are: topical resemblance in overlap terms, *co-citation* (the number of pages, where both are given as hyperlinks), *bibliographical coupling* (the number of hyperlinks in common) and *co-occurrence frequency*

$\min\{P(x|y), P(y|x)\}$, where $P(x|y)$ is the probability that a visitor of the page y will visit the page x.

In particular, **search-centric change metrics** are metrics used by search engines on the Web, in order to measure the degree of change between two versions x and y of a web page. If X and Y are the set of all words (excluding HTML markup) in x and y, respectively, then the **word page distance** is the **Czekanowsky–Dice distance** (cf. Chap. 17)

$$\frac{|X \triangle Y|}{|X| + |Y|} = 1 - \frac{2|X \cup Y|}{|X| + |Y|}.$$

If v_x and v_y are weighted vector representations of x and y, then their **cosine page distance** is given (cf. **TF-IDF similarity** in Chap. 17) by

$$1 - \frac{\langle v_x, v_y \rangle}{||v_x||_2 \cdot ||v_y||_2}.$$

- **Web quality control distance function**

  Let P be a query quality parameter and X its domain. For example, P can be query *response time*, or accuracy, relevancy, size of result.

  The **Web quality control distance function** (Chen–Zhu–Wang, 1998) for evaluating the relative goodness of two values, x and y, of parameter P is a function $\rho : X \times X \to \mathbb{R}$ (not a **distance**) such that, for all $x, y, z \in X$:

  1. $\rho(x, y) = 0$ if and only if $x = y$,
  2. $\rho(x, y) > 0$ if and only if $\rho(y, x) < 0$,
  3. if $\rho(x, y) > 0$ and $\rho(y, z) > 0$, then $\rho(x, z) > 0$.

  The inequality $\rho(x, y) > 0$ means that x is better than y; so, it defines a *partial order* (reflexive, antisymmetric and transitive binary relation) on X.
- **Lostness metric**

  Users navigating within hypertext systems often experience *disorientation* (the tendency to lose sense of location and direction in a nonlinear document) and *cognitive overhead* (the additional effort and concentration needed to maintain several tasks/trails at the same time). Smith's **lostness metric** measures it by

  $$(\frac{d}{t} - 1)^2 + (\frac{r}{d} - 1)^2,$$

  where t is the total number of nodes visited, d is the number of different nodes among them, and r is the number of remaining nodes needed to complete a task.
- **Trust metrics**

  A **trust metric** is, in Computer Security, a measure to evaluate a set of peer certificates resulting in a set of accounts accepted and, in Sociology, a measure of how a member of the group is trusted by the others in the group.

[redacted]

For example, the UNIX access metric is a combination of only *read*, *write* and *execute* kinds of access to a resource. The much finer *Advogato* trust metric (used in the community of open source developers to rank them) is based on bonds of trust formed when a person issues a certificate about someone else. Other examples are: *Technorati*, *TrustFlow*, Richardson et al., Mui et al., *eBay* trust metrics.

- **Software metrics**

  A **software metric** is a measure of software quality which indicates the complexity, understandability, description, testability and intricacy of code. Managers use mainly **process metrics** which help in monitoring the processes that produce the software (say, the number of times the program failed to rebuild overnight).

  An **architectural metric** is a measure of software architecture (development of large software systems) quality which indicates the coupling (interconnectivity of composites), cohesion (intraconnectivity), abstractness, instability, etc.

- **Locality metric**

  The **locality metric** is a physical metric measuring globally the locations of the program components, their calls, and the depth of nested calls by

$$\frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}},$$

  where d_{ij} is a distance between calling components i and j, while f_{ij} is the frequency of calls from i to j. If the program components are of about same size, $d_{ij} = |i - j|$ is taken. In the general case, Zhang–Gorla, 2000, proposed to distinguish *forward* calls which are placed before the called component, and *backward* (other) calls. Define $d_{ij} = d'_i + d''_{ij}$, where d'_i is the number of lines of code between the calling statement and the end of i if call is forward, and between the beginning of i and the call, otherwise, while $d''_{ij} = \sum_{k=i+1}^{j-1} L_k$ if the call is forward, and $d''_{ij} = \sum_{k=j+1}^{i-1} L_k$ otherwise. Here L_k is the number of lines in component k.

- **Memory distances**

  In a computer, the *microprocessor* (or *processor*) is the chip doing all the computations, and the *memory* usually refers to *RAM* (random access memory). A (processor) *cache* stores small amounts of recently used information right next to the processor, where it can be accessed much faster than memory. The following distances estimate the cache behavior of programs.

  The **reuse distance** (Mattson et al., 1970, and Ding–Zhong, 2003) of a memory location x is the number of distinct memory references between two accesses of x. Each memory reference is counted only once, because after access it is moved in the cache. The reuse distance from the current access to the previous one or to the next one is called the *backward* or *forward* reuse distance, respectively.

In general, a **memory distance** is (Fang et al., 2006) any dynamic quantifiable distance in terms of memory references between two accesses to the same memory location. In particular, the **access distance** of a memory reference is the number of memory instructions between a store to and a load from the same address. The *instruction distance* is the number of instructions between two references to the same address; it typically correlates well to access distance.

The *value distance* of a reference is the access distance of a load to the first store in a sequence of stores of the same value.

- **Tethering distance**

  Tethering is connecting one device to another. For mobile phones and tablet computers, it allows sharing their Internet connection with, say, laptops. Such connection can be done over wireless LAN (Wi-Fi), over Bluetooth, or by a cable, say USB.

  If tethering is done over WLAN, the feature becomes a mobile *hotspot* (Wi-Fi network access point), and the smartphone serves as a portable router. **Tethering distance** is the maximum phone–computer distance for Wi-Fi tethering.

- **Action at a distance (in Computing)**

  In Computing, the **action at a distance** is a class of programming problems in which the state in one part of a program's data structure varies wildly because of difficult-to-identify operations in another part of the program.

  In Software Engineering, Holland's *Law of Demeter* is a style guideline: an unit should "talk only to immediate friends" (closely related units) and have limited knowledge about other units; cf. **principle of locality** in Chap. 24.

[redacted]

Part VI
Distances in Natural Sciences

Chapter 23

Distances in Biology

Distances are mainly used in *Biology* to pursue basic classification tasks, for instance, for reconstructing the evolutionary history of organisms in the form of phylogenetic trees. In the classical approach those distances were based on comparative morphology, physiology, mating studies, paleaontology and immunodiffusion. The progress of modern *Molecular Biology* also allowed the use of nuclear- and amino-acid sequences to estimate distances between genes, proteins, genomes, organisms, species, etc.

DNA is a sequence of *nucleotides* (or *nuclei acids*) A, T, G, C, and it can be seen as a word over this alphabet of four letters. The (single ring) nucleotides A, G (short for adenine and guanine) are called *purines*, while (double ring) T, C (short for thymine and cytosine) are called *pyrimidines* (in RNA, uracil U replaces T). Over 25 base modifications are known, say, variants 5mC of cytosine and 6mA of adenine.

Two strands of DNA are held together and in the opposite orientation (forming a double helix) by weak hydrogen bonds between corresponding *base pair* of nucleotides (necessarily, a purine and a pyrimidine) in the strands alignment.

A *transition mutation* is a substitution of a base pair, so that a purine/pyrimidine is replaced by another purine/pyrimidine; say, GC is replaced by AT. A *transversion mutation* is a substitution of a base pair, so that a purine/pyrimidine is replaced by a pyrimidine/purine base pair, or vice versa; say, GC is replaced by TA.

DNA molecules occur (in the nuclei of eukaryote cells) in the form of long chains called *chromosomes*. DNA from one human cell has length/width $\approx$ 1.8 m/2.4 nm.

Most human cells contain 46 chromosomes (23 pairs, one set of 23 from each parent); the human *gamete* (sperm or egg) is a *haploid*, i.e., contains only one set of 23 chromosomes. The (normal) males and females differ only in the 23-rd pair: *XY* for males, and *XX* for females. But a male ant *Mirmecia pilosula* has only 1 chromosome, while a plant *Ophioglossum* has 1260. A protozoan *Tetrahymena thermophila* occurs in seven different variants (*sexes*) that can reproduce in $\binom{7}{2} = 21$ combinations. A fungus *Cryptococcus neoformans* has two sexes but their ratio

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_23

[redacted]

is 99.9 %. More than 99 % of multicellular eukaryotes reproduce sexually, while bacteria only reproduce asexually.

A *gene* is a segment of DNA encoding (via *transcription*, information flow to RNA, and then *translation*, information flow from RNA to enzymes) for a protein or an RNA chain. The location of a gene on its chromosome is called the *gene locus*. Different versions (states) of a gene are called its *alleles*.

A *protein* is a large molecule which is a chain of *amino acids*; among them are hormones, catalysts (enzymes), antibodies, etc. The **protein length** is the number of amino acids in the chain; average protein length is around 300.

The *genetic code* is a map, universal to all organisms, of $4^3 = 64$ *codons* (ordered triples of nucleotides) onto 21 messages: 20 *standard* amino acids and stop-signal. It express the *genotype* (information contained in genes, i.e., in DNA) as the *phenotype* (proteins). Some codons have two meanings, one related to protein sequence, and one related to gene control. Slight variations of the code were found for some mitochondria, ciliates, yeasts, etc. The code also was expanded by encoding new amino acids.

Besides genetic and *epigenetic* (not modifying the sequence) changes of DNA, *evolution* (heritable changes) can happen by "protein mutations" (prions) or culturally (via behavior and symbolic communication). Holliger et al., 2012, synthesized (replacing the natural sugar in DNA) a new polymer (*HNA*) capable of replication and evolution.

A *genome* is the entire genetic constitution of a species or of an organism. The human genome is the set of 23 chromosomes consisting of $\approx$ 3.2 billion *bp* (base pairs) of DNA. It organized into $\approx$21,000 genes, coding for 250,000–1,000,000 proteins. An apple have $\approx$57,000 genes and a lungfish genome contains $\approx$ 133 billion bp.

A *hologenome* is the collection of genomes in a *holobiont* (host plus all its symbionts), a possible unit of selection in evolution. The human microbiota consists of $\approx 10^{14}$ (mainly, bacterial and fungal) cells of $\approx$ 500 species with 3 million distinct genes.

First known evidence of life, photosynthesis, multicellular organisms and of animals (bilaterians) are dated about 4100, 3850, 2100 and 560 Ma (million years ago), respectively. During the *Cambrian Explosion* 540–520 Ma, the rate of evolution was 4–5 times faster than in any other era. But there is no evolution without changes in environment: some deep sea bacteria remained virtually unchanged for 2.3 billion years.

1.9 million extant nonvirus species are known: 1,200,000 invertebrates, 400,000 beetles, 290,000 plants, 250,000 bacteria/protists, 70,000 fungi and 60,000 vertebrates, including 5416 mammals. About 10 million species are living today. The numbers of living nonhuman mammals, birds, trees, fishes, ants, insects, nematodes, viruses are about 5×10^{11}, 2×10^{11}, 3×10^{12}, 4×10^{12}, 5×10^{17}, 5×10^{18}, 10^{22}, 10^{31}, respectively. At least 100 billion of microbial species are expected, but genomes of only 100,000 have been sequenced and about 10,000 have been grown in a laboratory.

About 80 % of the species are parasites of others, parasites included; over 100 are human-specific ones. About 75 % of the links in food webs involve a parasitic species.

About 30 % of the vertebrates and 60 % of invertebrates are nocturnal.

The global nonbacterial biomass is 560 GtC (billion tonnes of organically bound carbon); only 5–10 GtC found in the oceans. Humans and their main symbionts, domesticated animals and cultivated plants, contribute $0.1, 0.7, 2$ GtC. The animal species with largest biomass ($\approx$ 0.5 GtC) is Antarctic krill, the largest species ($\sim 10^{27}$ individuals) is the smallest photosynthetic organism, cyanobacteria *Prochlorococcus*.

99 % of over 5 billion species that have ever existed on Earth became extinct. But sponges originated at least 760 Ma (million years ago) and are still living. Mammalian species' longevity is $\approx$ 1 Ma; our ancestor, *Homo erectus*, survived from 1.8 to 0.55 Ma. Our subspecies is relatively young (0.25–0.6 Ma), 6–7 % of all humans that have ever been born are living today, and their median age is 29.7 years.

The world population was about 1 million 0.05 Ma and 5 million 0.01 Ma ago, after the last glaciation. It grew continuously since the end of the Black Death in 1350, when it was $\approx$ 370 million, and reached 3 billion in 1960, 7.4 billion in 2015.

Gott, 2007: with a 95 % chance, the human race will last anywhere from another 5000 to 7,800,000 years; the same *doomsday argument* by Carter, 1983, gave only 10,000 years for us. Earth's life was only unicellular 3.8–1.3 Ga (billion years) ago and will be so again in $\approx$ 0.8 Ga. But Earth will support some prokaryotes in refuges until mean surface temperature reach 146 °C in 1.6–2.8 Ga. Another Ga life can stay on Mars.

IAM (infinite-alleles model of evolution) assumes that an allele can change from any given state into any other given state. It corresponds to a primary role for *genetic drift* (i.e., random variation in gene frequencies from one generation to another), especially in small populations, over *natural selection* (stepwise mutations). IAM corresponds to low-rate and short-term evolution, while SMM corresponds to high-rate evolution.

SMM (stepwise mutation model of evolution) is more convenient for (recently, most popular) microsatellite data. A *repeat* is a stretch of base pairs that is repeated with a high degree of similarity in the same sequence. *Microsatellites* are highly variable repeating short sequences of DNA; their mutation rate is 1 per 1000–10,000 replication events, while it is 1 per 1,000,000 for *allozymes* used by IAM. Microsatellite data (for example, for DNA fingerprinting) consist of numbers of repeats of microsatellites for each allele.

Evolution, without design and purpose, has increased the life's size, diversity and maximal complexity. But organisms can evolve to become simpler and thus multiply faster; for the *Black Queen model*, such evolution pushes microorganisms to lose functions which are performed by another species around. Evolution has, perhaps, a direction: *convergent gene evolution* (say, bats/dolphins echolocation, primates/crows cognition), increase of energy flow per gram per second (Chaisson, 2003), etc.

[redacted]

Evolution and learning share many common underlying principles. Natural selection can favor increased evolvability under environmental pressure. Besides natural selection, some species alter their environment through *niche construction*. In general, selection can act at genic, cellular, individual, holobiont and group level. Selection of species and even phyla could happen during rare abrupt extreme events.

Locally and over short time spans, macroevolution, is dominated by biotic factors (competition, predation, etc.) as in the *Red Queen model*. But larger-scale (geographic and temporal) patterns and species diversity are driven largely by extrinsic abiotic factors (climate, landscape, food supply, tectonic events, etc.), as in the *Court Jester model*. The organisms evolve rapidly (sometimes, by macromutations), but most changes cancel each other out. So, in the longer term, the evolution appears slow. It is not simple accumulation of microevolutionary adaptations, but rather nonlinear (or chaotic).

Besides *vertical gene transfer* (reproduction within species), the evolution is affected by *HGT* (horizontal gene transfer), when an organism incorporates genetic material from another one without being its offspring, and *hybridization* (extra-species sexual reproduction). HGT is common among unicellular life and viruses, even across large **taxonomic distance**. It accounts for $\approx 85\,\%$ of the prokaryotic protein evolution. HGT happens also in plants and animals, usually, by viruses. 40–50 % of the human genome consists of DNA imported horizontally by viruses. The most taxonomically distant fertile hybrids are (very rare) interfamilial ones, for instance, blue-winged parrot × cocktatiel, chicken × guineafowl in birds and (under UV irradiation) carrot with tobacco, rice or barley. In 2012, an RNA-DNA virus hybrid and a *virophage* of a (giant) virus were found.

The *life* is not well defined (say, for *viroids*), DNA could be only its recent attribute. Neither life can be “anything undergoing evolution”, since the unit is this evolution (gene, cell, organism, group, species?) is not clear. Lineweaver, 2012, defined life as a *far-from-equilibrium dissipative system*. For Eigen, life is a type of behavior of matter. An essential feature of life is *autopoiesis* (self-making); for example, the human body replaces 98 % of its atoms every year while maintaining its unique pattern.

Neither *complexity* is well defined. Chaisson, 2014, measures complexity of any Nature’s system by the amount of energy passing through it per second and per gram. Earth’s biosphere use energy at $\approx 10^{21}$ erg/s ($\approx 0.1\,\%$ of obtained solar power).

Examples of distances, representing general schemes of measurement in Biology, follow.

The term **taxonomic distance** is used for every distance between two *taxa*, i.e., entities or groups which are arranged into a hierarchy (in the form of a tree designed to indicate degrees of relationship).

The *Linnaean taxonomic hierarchy* is arranged in ascending series of ranks: Zoology (Kingdom, Phylum, Class, Order, Family Genus, Species) and Botany (12 ranks). A *phenogram* is a hierarchy expressing *phenetic relationship*, i.e., unweighted overall similarity. A *cladogram* is a strictly genealogical (by ancestry) hierarchy in which no attempt is made to estimate/depict rates or amount of genetic divergence between taxa.

A *phylogenetic tree* is a hierarchy representing a hypothesis of *phylogeny*, i.e., evolutionary relationships within and between taxonomic levels, especially the patterns of lines of descent. The **phenetic distance** is a measure of the difference in phenotype between any two nodes on a phylogenetic tree; see, for example the **biodistances** in Chap. 29.

The **phylogenetic distance** (or **cladistic distance**, **genealogical distance**) between two taxa is the *branch length*, i.e., the minimum number of edges, separating them in a phylogenetic tree. In such edge-weighted tree, the **additive distance** between two taxa is the minimal sum of edge-weights in a path connecting them. The *phylogenetic diversity* is (Faith, 1992) the minimum total length of all the phylogenetic branches required to span a given set of taxa on the phylogenetic tree.

The **evolutionary distance** (or **patristic distance**) between two taxa is a measure of genetic divergence estimating the **temporal remoteness** of their most recent co-ancestor. Their general **immunological distance** is a measure of the strength of antigen-antibody reactions, indicating their evolutionary distance.

23.1 Genetic Distances

The general **genetic distance** between two taxa is a distance between the sets of DNA-related data chosen to represent them. Among the three most popular genetic distances below, the **Nei standard genetic distance** assumes that differences arise due to mutation and genetic drift, while the **Cavalli-Sforza–Edwards chord distance** and the **Reynolds–Weir–Cockerham distance** assume genetic drift only.

A *population* is represented by a vector $x = (x_{ij})$ with $\sum_{j=1}^{n} m_j$ components, where x_{ij} is the frequency of the i-th *allele* (the label for a state of a gene) at the j-th gene locus (the position of a gene on a chromosome), m_j is the number of alleles at the j-th locus, and n is the number of considered loci. Since x_{ij} is the frequency, we have $x_{ij} \geq 0$ and $\sum_{i=1}^{m_j} x_{ij} = 1$. Denote by $\sum$ summation over all i and j.

- **Shared allele distance**

  The **shared allele distance** D_{SA} (Stephens et al., 1992, corrected by Chakraborty–Jin, 1993) between individuals a, b is $1 - SA(a, b)$; for populations x, y it is

$$1 - \frac{\overline{SA(x,y)}}{\overline{SA(x)} + \overline{SA(y)}},$$

  where $SA(a, b)$ denotes the number of shared alleles summed over all n loci and divided by $2n$, while $\overline{SA(x)}$, $\overline{SA(y)}$, and $\overline{SA(x,y)}$ are $SA(a, b)$ averaged over all pairs (a, b) with individuals a, b being in populations x, y, respectively.

[redacted]

- **MHC genetic dissimilarity**

  The **MHC genetic dissimilarity** of two individuals is defined as the number of shared alleles in their MHC (*major histocompatibility complex*).

  MHC is the most gene-dense and fast-evolving region of the mammalian genome. In humans, it is a 3.6 Mb region containing 140 genes on chromosome 6 and called HLA (*human leukocyte antigen system*). HLA has the largest *polymorphism* (allelic diversity) found in the population. Three most diverse loci (HLA-A, HLA-B,HLA-DRB1) have roughly 1000, 1600, 870 known alleles. This diversity is essential for immune function since it broadens the range of *antigens* (proteins bound by MHC and presented to T-cells for destruction); cf. **immunological distance**.

  MHC (and related gut microbiota) diversity allows the marking of each individual of a species with a unique body odor permitting kin recognition and mate selection. *MHC-negative assortative mating* (the tendency to select MHC-dissimilar mates) increases MHC variation and so provides progeny with an enhanced immunological surveillance and reduced disease levels.

  While about 6 % of the non-African modern human genome is common with other hominins (Neanderthals and Denisovans), the share of such HLA-A alleles is 50 %, 72 %, 90 % for people in Europe, China, Papua New Guinea.

- ***Dps* distance**

  The **Thorpe similarity** (proportion of shared alleles) between populations x and y is defined by $\sum \min\{x_{ij}, y_{ij}\}$. The *Dps* **distance** between x and y is defined by

$$-\ln \frac{\sum \min\{x_{ij}, y_{ij}\}}{\sum_{j=1}^{n} m_j}.$$

- **Prevosti–Ocana–Alonso distance**

  The **Prevosti–Ocana–Alonso distance** (1975) between populations x and y is defined (cf. **Manhattan metric** in Chap. 19) by

$$\frac{\sum |x_{ij} - y_{ij}|}{2n}.$$

- **Roger distance**

  The **Roger distance** D_R (1972) between populations x and y is defined by

$$\frac{1}{\sqrt{2}n} \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m_j} (x_{ij} - y_{ij})^2}.$$

- **Cavalli-Sforza–Edwards chord distance**

  The **Cavalli-Sforza–Edwards chord distance** D_{CH} (1967) between populations x and y (cf. **Hellinger distance** in Chap. 17) is defined by

$$\frac{2\sqrt{2}}{\pi n}\sum_{j=1}^{n}\sqrt{1-\sum_{i=1}^{m_j}\sqrt{x_{ij}y_{ij}}}.$$

  The **Cavalli-Sforza arc distance** between populations x and y is defined by

$$\frac{2}{\pi}\arccos\left(\sum\sqrt{x_{ij}y_{ij}}\right).$$

  Cf. **Bhattacharya distance 1** in Chap. 14.

- **Nei–Tajima–Tateno distance**

  The **Nei–Tajima–Tateno distance** D_A (1983) between populations x and y is

$$1-\frac{1}{n}\sum\sqrt{x_{ij}y_{ij}}.$$

  The **Tomiuk–Loeschcke distance** (1998) is $-\ln\frac{1}{n}\sqrt{\sum x_{ij}\sum y_{ij}}$.

  The **Nei standard genetic distance** D_s (1972) between x and y is defined by

$$-\ln\frac{\langle x,y\rangle}{||x||_2\cdot||y||_2}.$$

  Cf. **Bhattacharya distances** in Chap. 14 and **angular semimetric** in Chap. 17.

  Under IAM, D_s increases linearly with time; cf. **temporal remoteness**.

  The **kinship distance** is defined by $-\ln\langle x,y\rangle$. Caballero and Toro, 2002, defined the *molecular kinship coefficient* between x and y as the probability that two randomly sampled alleles from the same locus in them are *identical by state*. Computing it as $\langle x,y\rangle$ and using the analogy with the **coefficient of kinship** defined via *identity by descent*, they proposed several distances adapted to molecular markers (polymorphisms). Cf. **co-ancestry coefficient**.

  The **Nei minimum genetic distance** D_m (1973) between x and y is defined by

$$\frac{1}{2n}\sum(x_{ij}-y_{ij})^2.$$

- **Sangvi χ^2 distance**

  The **Sangvi χ^2 (1953) distance** between populations x and y is defined by

$$\frac{2}{n}\sum\frac{(x_{ij}-y_{ij})^2}{x_{ij}+y_{ij}}.$$

[redacted]

- **Fuzzy set distance**

  The **fuzzy set distance** D_{fs} between populations x and y (Dubois–Prade, 1983; cf. **Tanimoto distance** in Chap. 1) is defined by

$$\frac{\sum 1_{x_{ij} \neq y_{ij}}}{\sum_{j=1}^{n} m_j}.$$

- **Goldstein et al. distance**

  The **Goldstein et al. distance** (1995) between populations x and y is

$$(\delta\mu)^2 = \frac{1}{n}\sum (ix_{ij} - iy_{ij})^2.$$

It is the loci-averaged value $(\delta\mu)^2 = (\mu(x)_j - \mu(y)_j)^2$, where $\mu(z)_j = \sum_i iz_{ij}$ is the mean number of repeats of allele at the j-th (microsatellite) locus in population z.

The **Feldman et al. distance** (1997) is $\log(1 - \frac{\sum_i (\delta\mu)_i^2}{M})$, where the summation is over loci and M is the average value of the distance at maximal divergence.

The above two and the next two distances assume high-rate SMM.

- **Average square distance**

  The **average square distance** between populations x and y is defined by

$$\frac{1}{n}\sum_{k=1}^{n}\left(\sum_{1 \le i < j \le m_j} (i-j)^2 x_{ik} y_{jk}\right).$$

- **Shriver et al. stepwise distance**

  The **Shriver et al. stepwise distance** (1995) between populations x and y is

$$D_{SW} = \frac{1}{n}\sum_{k=1}^{n}\sum_{1 \le i,j \le m_k} |i-j|(2x_{ik}y_{jk} - x_{ik}x_{jk} - y_{ik}y_{jk}).$$

- **Latter F-statistics distance**

  The **Latter F-statistics distance** (1972) between populations x and y is defined by the following F_{ST}-*estimator*:

$$\theta^* = \frac{\sum (x_{ij} - y_{ij})^2}{2(n - \langle x, y\rangle)}.$$

The **Latter distance** D_L (1973) is $-\ln(1 - \theta^*)$.

- **Reynolds–Weir–Cockerham distance**

  The **Reynolds–Weir–Cockerham distance** (1983) between populations x and y is defined by

  $$D_W = -\ln(1-\theta),$$

  where θ is their **co-ancestry coefficient** estimated as $\frac{a}{a+b}$.

  Here a is the variance between populations x and y, and b is the variance within them. If the sample size is large, then D_W is close to the **Latter F-statistics distance**. For short-term evolution (i.e., $\frac{t}{N}$ small), $D_W \approx \frac{t}{2N}$, where N is the population size, and t is the number of generations; cf. **temporal remoteness**.
- **Co-ancestry coefficient**

  The **co-ancestry coefficient** (or *coefficient of kinship*) of two populations (or individuals) x and y is defined (Wright, 1922, and Malécot, 1948) as the probability $\theta(x,y)$ that two alleles, sampled at random from x and y, are *IBD* (or *identical by descent*), i.e., descending from the same ancestral allele.

  Two genes can be *IBS* (or *identical by state*), i.e., similar due to random chance. Cf. **Nei standard genetic distance** and **coefficient of relationship**.

  An DNA segment, found consistently to be identical in two related people (or populations) is IBD if it is so due to their common ancestry. The total and mean **IBD segment length** of two people g generations since the founding event (i.e., with g meioses on the path of descent) are $\approx \frac{1}{2^g}$-th of total genome length and $\approx \frac{50}{g}$ centiMorgans, respectively; cf. the **map distance**. For example, two people are cryptic relatives if those lengths are at least 1500 cM and 25 cM.
- **F_{ST}-based distances**

  Given a population T of size $|T|$ partitioned into subpopulations $S_1,\dots,S_k$, the *F-statistics* (or *fixation indices*) are the measures

  $$F_{IS} = 1-\frac{H_I}{H_S},\ F_{ST} = 1-\frac{H_S}{H_T},\ F_{IT} = 1-\frac{H_I}{H_T}$$

  of the correlation between genes drawn within subpopulations S_i, among them and within the entire T, respectively.

  Here H_I, H_S and H_T are the *heterozygosity indices* over (i)ndividuals, (s)ubpopulations and (t)otal T used to compare observed variation in gene frequencies (partitioned into within and between group components) with the expected one in HWE (*Hardy–Weinberg equilibrium*, i.e., an ideal state when allele and genotype frequencies in population remain constant from generation to generation). $H_I = \frac{\sum_{1\le j\le k}|S_j|H_{obs\,j}}{|T|}$ (where $H_{obs\,j}$ is the observed *heterozygosity*, i.e., proportion of heterozygotes, in subpopulation S_j) is the mean actual heterozygosity in individuals within subpopulations. $H_S = \frac{\sum_{1\le j\le k}|S_j|H_{exp\,j}}{|T|}$ (where $H_{exp\,j} = 1-\sum_i p_i^2$ is the expected, assuming HWE, heterozygosity in S_j and p_i is the frequency of the i-th allele of the locus) is the mean expected

[redacted]

heterozygosity within subpopulations. $H_T = 1 - \sum_i \bar{p}_i^2$ (where $\bar{p}_i$ is the frequency of the i-th allele averaged over all subpopulations) is the expected, assuming HWE, heterozygosity in T.

The above Nei's (1973) F_{ST} generalizes Wright's (1951) F_{ST}, when there are only two alleles at a locus. This measure is equivalent to the **co-ancestry coefficient** if all the alleles in the population are different. Nei, 1987, generalized F_{ST} to multi-loci as $G_{ST} = 1 - \frac{\bar{H}_S}{\bar{H}_T}$, where H_S and H_T are averaged across all loci.

The above relative measures underestimate the between-population difference if the within-population diversity is high, such as, say, for microsatellites. Slatkin's R_{ST} is an analog of Wright's F_{ST}, adapted for microsatellite loci by assuming SMM. It is defined by $R_{ST} = \frac{\bar{S} - S_W}{\bar{S}}$, where S_W is the sum over all loci of twice the weighted mean of the within-population variances $var(A)$ and $var(B)$, and $\bar{S}$ is the sum over all loci of twice the variance $var(A \cup B)$ of the combined population.

In fact, S_W and $\bar{S}$ are the **average square distance** within a subpopulation and the entire population. Slatkin (1995) developed R_{ST} using his (1991) SMM-based F-statistics $F_{ST} = \frac{\bar{t} - t_0}{\bar{t}}$, where $\bar{t}$ and t_0 are the average **temporal remoteness** to the closest co-ancestor of any two randomly chosen alleles from the entire population and from the same subpopulation, respectively.

Jost's D_{est} (2008) is an estimator $\frac{k}{k-1}\frac{H_T - H_S}{1 - H_S}$ of the actual differentiation based on H's estimated from allele identities rather than ratios of heterozygosity.

The Weir–Cockerham θ_{ST} (1984) is an estimation of F_{ST}, seen as the correlation of pairs of alleles between individuals within a subpopulation and based on partition of variance rather than heterozygosity. The total variance of allele frequency within a population is the sum $a+b+c$ of variances between subpopulations, between individuals within a subpopulation, and between gametes within individuals. Then θ_{ST} is defined, generalizing the **Reynolds–Weir–Cockerham distance**, as $\frac{\sum a}{\sum (a+b+c)}$, where the sum is taken over all alleles and loci.

The **genetic F_{ST}-distance** is the pairwise F_{ST} taking account only of the data for the two subpopulations concerned, not all the data simultaneously. Such a measure is valid only if the breeding system is similar for both populations.

Cavalli-Sforza–Menozzi–Piazza, 1994, evaluated, using 120 blood polymorphisms, the doubled genetic F_{ST}-distance between 42 native human populations and between 9 resulting clusters. The largest such distances between two continents were Africa–Oceania (0.247) and Africa–Americas (0.226), while the shortest distances were Americas–Asia (0.089) and Americas–Europe (0.095).

The largest distance in Europe, $F_{ST} = 0.02$–0.023, was between Finland and Southern Italy; cf. 0.11 (Europeans–Chinese) and 0.153 (Europeans–Africans (Yoruba)). Mbuti Pygmies (the least "Neanderthal") and Papuans (the most "Denisovan") are the two most divergent living humans with $F_{ST} = 0.377$.

A similar analysis by Atzmon et al., 2010, of seven Jewish groups indicated a common origin and, 100–150 generations ago, the split into Middle Eastern and European clusters. The most distant and differentiated are among Mizrahim: F_{ST} of Iranian Jews to other Jews is 0.016. Ashkenazi Jews have the highest admixture with non-Jews but they are not descendants of converted Khazars or Slavs. The closest by F_{ST} to them are Northern Italians, French, Palestinians, and Druze.

Genetic variation in alleles of genes occurs both within (due to mutations and gene exchange during meiosis) and among (due to natural selection and *genetic drift*, i.e., random gene changes) populations. Human total genetic variation is 0.5 % consisting of 0.1 % in SNPs (single nucleotide polymorphisms), $\approx$ 0.4 % in *CNVs* (copy number variants—deletion, duplication or more—in a DNA segment, instead of exactly two copies of DNA per cell) and a small variation in repetitive DNA. Sudmant et al., 2015, found that CNVs (especially, duplications) are a source of seven times greater diversity compared to SNPs.

Each human is born with about 50 new mutations, rarely noticeable. The main other mechanisms of our genetic diversity are migrations and hybridization.

The genetic similarity of humans is 99.9 % among them, while it is 99.85 % with Neanderthals and 96–98 % with (having SNP diversity 0.2 %) chimpanzees. After initial division, there was interbreeding with chimpanzees and, later, with Neanderthals, Denisovans and archaic Africans. There are 2 % of archaic genes in sub-Saharan Africans, 2 % of Neanderthal genes in Central Asians and 4 % of them in Europeans and Americans. There are 2.5 % of Neanderthal and 6–8 % Denisovan genes in South Asians and Australo-Melanesians.

75–85 % of human SNP variation 0.1 % is among individuals within any population, 5–10 % between local populations of the same continent, and 6–10 % between large groups from different continents. So, differentiation between continental groups is $F_{ST} \leq 0.1$, less than the threshold 0.25 used to define a *subspecies* (*race*).

- **Temporal remoteness**

The **temporal remoteness** of most recent common ancestor (or *TMRCA*, *divergence time*, *time to coalescence*) of two taxa is the time (or the number of generations) that has passed since those populations existed as a single one. The *molecular clock hypothesis* estimates that one unit of **Nei standard genetic distance** between two taxa corresponds to 18–20 Ma of their TMRCA.

A human phylogenetic tree is derived from matrilineal mitochondrial DNA, or patrilineal nonrecombinant part of the Y-chromosome of (usually blood) protein sequences by measuring accumulated mutations. TMRCA is 0.2–0.19 Ma ago along all-female ancestry lines for the *Mitochondrial Eve* and 0.24–0.58 Ma ago along all-male lines for the *Y-chromosomal Adam*.

The resulting phylogenetic tree is rooted in the common ancestor of chimpanzees and humans, which originated in Africa 8–6 Ma ago. The corresponding **genetic F_{ST}-distance** between humans and chimpanzee is about 0.02, i.e., at least 30 million point mutations affecting 80 % of genes.

Our primate ancestors evolved, probably, 40 Ma ago in Asia and sailed across a narrow sea to Africa. Our genus *Homo* had diverged, as a carnivorous scavenger, from the *Australopithecines* (bipedal ape-like using rudimentary stone tools) $\approx$ 2.5 Ma ago in East (*Homo habilis*) or South Africa. Then *Homo erectus*, the first global and using fire human species, moved to Eurasia 1.8 Ma ago,

[redacted]

followed by Denisovans and, later, Neanderthals, the common ancestor of which split from our line 0.8 Ma ago.

At least ten extinct Homo species are known. *Archaic Homo sapiens* originated 0.7–0.6 Ma ago. They evolved to anatomically modern humans *Homo sapiens sapiens* $\approx$ 0.2 Ma ago, as shown by the temporal remoteness of their mitochondrial most recent common ancestor. Then their mitochondrial lineage L3 (among L0, L1, L2, L3) migrated out of (east or southern) Africa 0.125 and 0.065 Ma ago.

Humans passed via population bottleneck $\approx$ 0.074 Ma ago (when Toba supervolcano erupted), followed by a rapid expansion. They arrived about 0.075, 0.046, 0.044, 0.016 Ma ago in West Africa, Australia, Europe and the Americas, respectively. The last place on Earth (besides the Antarctica, occupied only in 1899, and tiny atolls) humans colonised was New Zealand where they arrived $\approx$1300 years ago.

Savanna living, use of fire, speech and sophisticated hand axes appeared about 1.7, 1.6, 0.6, 0.5 Ma ago. Modern human behavior (language, symbolic thought, *cultural universals*) emerged 0.07–0.05 Ma, i.e., $\approx$3000 generations, ago.

The main known gene mutations leading to us: improving blood supply to the primate brain, weakening jaw muscle (so skull/brains could expand), speeding up the neuron migration (crucial to intelligence), increasing the production of the salivary enzyme (helping to the emergence of agriculture). Also, noncoding sequence HACNS1 had 16 variations during last 6 Ma; it led to more fine muscle control allowing bipedality and tool use. The gene miR-941, unique to humans, emerged 6–1 Ma ago; it could initiate our advanced brain functions.

- **Pedigree-based distances**

A *cousin* (or *blood relative*) is a relative with whom one shares a common ancestor. A *cousin chart* (or *table of consanguinity*, *family tree*, *pedigree digraph*) is a directed tree, where vertices represent relatives (usually humans, show dogs, race horses or cultivars), and the arc uv means that v is a child of u. So, the in-degree of each vertex is at most two (known parents). Moreover, unoriented edges are added with edge uv meaning *reproductive affinity*, i.e., that u and v are mated.

The **genealogical quasi-distance** (or, in Anthropology, *genealogical distance*, *degree of relative consanguinity*) from the individual x to its relative y is defined (Schneider, 1968) as the number of generations one must go before a common ancestor is found, i.e., it is the length $q(x,y)$ of the shortest directed path to x from a common ancestor of x and y in the family tree. Recently, the value $\min\{q(x,y), q(y,x)\}$ is preferred in English pedigree documents.

An *ancestral path* between the vertices x and y in a family tree (or any acyclic digraph) is a concatenation of two directed paths from a common ancestor to them. The **ancestral path distance** is the length of a shortest ancestral path, i.e., it is $q(x,y)+q(y,x)$. Cf. **genealogical distance** between the vertices x and y (of a phylogenetic tree representing taxa) which is the length of a shortest $(x-y)$-path in the undirected family tree, i.e., also $q(x,y)+q(y,x)$. Cf. ancestral path distance in Chap. 22 and **join semilattice distances** in Chap. 10.

The *ancestral distance of an extant taxon* (Hearn and Huber, 2006) is the time (or the number of speciation events) separating it from its most recent ancestor with at least one extant descendant having an independent trait.

Mycielski and Ulam, 1969, called *genealogical distances* between x and y the value $|S(x)\Delta S(y)|$, where $S(z)$ is the set of ancestors of z in a given family tree, and the **Manhattan metric** between some vector representations of x and y.

Two cousins are *a-removed of degree b* if they are separated by a generations and the minimum number of generations between either cousin and their common ancestor is b. The *direct relatives* are spouses and cousins with $(a,b) = (1,0),(2,0),(1,1),(0,2)$ and $(0,1)$, i.e., parents/children, grandparents/grandchildren, uncles (aunts)/nieces (nephews), first degree cousins and siblings. Clearly, $a = |q(x,y)-q(y,x)|$ and $b = \min\{q(x,y), q(y,x)\}$. Worldwide, $\approx 10\,\%$ of marriages are between closer than third degree cousins; the case of third degree cousins results in progeny only slightly more homozygous than the general population.

The above pedigree notions are important also in some family, inheritance and nationality rules. For example, the Roman Catholic Church prohibits marriage of x with a relative y if $q(x,y) + q(y,x) \leq 4$. The closest legally permissible unions are between *double-first cousins*, i.e., those sharing four grandparents (in Muslim populations), or uncle-niece (in South India).

Another example: a *Jew* in Halakha's (Jewish Law) sense is a child born to a Jewish mother or an converted adult. Israel's Law of Return permits independent repatriation to anyone with a nonapostate Jewish grandparent and/or his spouse. In Nazi Germany, a *full Jew* was anyone with three Jewish grandparents, while *part-Jews* of first/second degree were those (not practicing Judaism and not having a Jewish spouse) who had two/one Jewish grandparents.

The *inbreeding coefficient* $F(z)$ of an individual z is the probability of *autozygosity*, i.e., that z received the same ancestral gene from both its parents; so, $F(z)$ is the **co-ancestry coefficient** $\theta(z_1, z_2)$ of its parents z_1, z_2. When pedigree data are available, $\theta(x,y)$ is estimated as $\sum_{z\in Z(x,y)} 0.5^{|P(z)|}(1 + F(z))$, where $Z(x,y)$ is the set of least common ancestors of x and y in the pedigree digraph, and $|P(z)|$ is the number of vertices in the shortest ancestral path between x and y through z. In practice, ancestors z are counted only up to a given number of generations and not all of them are known.

The **coefficient of relationship** between two relatives x and y is the fraction of genome inherited from common ancestors. It is almost 1 for identical twins (they differ due to mutations during development and gene copy number variation) and $\approx \frac{3}{4}$ for *semi-identical twins* inheriting the same genes from only one parent. Otherwise, it is $2\theta(x,y)$, since any progeny have a risk $\frac{1}{2}$ of inheriting identical alleles from both parents. It is $\frac{1}{2}$ for siblings and for parent-offspring.

The **coefficient of relatedness** (or *genetic similarity*) between social partners x, y relative to the population is defined (Hamilton, 1970) by

$$r(x,y) = \frac{cov(g,g')}{cov(g,g)} = \frac{E[(g-E[g])(g'-E[g'])]}{E[(g-E[g])(g-E[g])]},$$

[redacted]

where g, g' are genetic (i.e., heritable) components of the phenotype (for the character of interest) of x, y, respectively, and *cov* denotes a statistical covariance taken over all individuals in the population. This coefficient quantifies the *indirect fitness*, i.e., the component of fitness gained from aiding related individuals.

Fitness is an individual's ability to propagate its genes, i.e., to both survive and reproduce. A measure of it is the average contribution to the gene pool of the next generation that is made by an average individual of the specified genotype or phenotype. The relative *reproductive value* of an individual is the probability that it is the ancestor of a randomly chosen individual in a distant future generation.

Fowler–Christakis, 2013: pairs of nonkin friends are, on average, as genetically similar to one another as fourth cousins,

- **Mating distances**

  Individual migration distances are the distances between birthplaces of paired individuals. If the pairs are spouses (gametes) or siblings, we have **marital distance** or **sib distance**, respectively. Also, the **parent-offspring distance** is used to describe gene migration per generation.

  For humans, those distances are measured either in km, or, say, as the number of municipalities crossed by a straight line between municipality midpoints of each pair's birthplaces. The term *marital migration distance* is also used for the distance between premarital town of a person and town of marriage. Cf. **migration distances (in Economics)** in Chap. 28.

  Until the twentieth century, men usually went courting no more than about 8 km from home (the distance they could walk out and back on their day off from work). 80 % of all marriages in history could be between second cousins or closer. Also, young birds, leaving the nest, usually move 4–5 home ranges away; so, they stay within breeding distance of their cousins.

  For a population, **critical mating distances**, are maximum spatial (physical) and genetic (number of genes bearing different alleles) distances allowed for mating; cf. **isolation by distance**. For honey bees, the *mating distance* is the range of queen's nuptial flight from her hive to the drone congregation areas over their hives; it is typically within 7.5 km but can reach 17 km.

- **Lasker distance**

  The **Lasker distance** between two human populations x and y, characterized by surname frequency vectors (x_i) and (y_i), is the number $-\ln 2R_{x,y}$, where $R_{x,y} = \frac{1}{2}\sum_i x_i y_i$ is Lasker's (1977) *coefficient of relationship by isonymy*.

  Surname structure is related to inbreeding and (in patrilineal societies) to random genetic drift, mutation and migration. Surnames can be considered as alleles of one locus, and their distribution can be analyzed by Kimura's theory of neutral mutations; an isonymy points to a common ancestry.

- **Isolation by distance**

  Isolation by distance (or *ibd*, Wright, 1943) is the tendency for most individuals to migrate and find mates between neighbors; so, populations that are a geographically closer are more similar than those that are further apart. It results in a smooth increase in a *cline*, i.e., the gradual change in a character (say, allele frequency, within- or between-population genetic differentiation) or feature (phenotype) with increasing geographic distance. The above distance can be Euclidean or along a great circle, river, or topographic isocline.

Ibd for humans was studied, for example, via migration patterns and the distribution of surnames (cf. **Lasker distance**). At both continental and global scales, the **genetic F_{ST}-distance** and differentiation in cranial morphology between populations increases with **great circle distance** (Chap. 25).

The geographic distance explains $> 75\,\%$ of the variation between human populations, and this distance from East Africa explains 85 % of the smooth decrease in genetic diversity. Atkinson, 2011, claims that phoneme diversity also declines with distance from Africa. The occurrence of alleles $7R$ and $2R$, linked to risk-taking, of the dopamine-related gene *DRD*4 increases with distance from Africa.

A strong Europe-wide (except Basques, Finns and Sardinians isolates) correlation, based on >300,000 single nucleotide polymorphisms, between geographic and genetic distance was found. South-to-North was the main smooth gradient.

The ibd model explains the emergence of regional differences (races) and new species by restricted gene flow and adaptive variations. *Speciation* (branching of new species from an ancestral population) occurs when subpopulations become reproductively isolated. The dominating mode of speciation is *allopatry* when habitat splits into discontinuous parts by the formation of a physical barrier to gene flow or dispersal. Examples of natural barriers are the Himalayas, Wallace Line, Grand Canyon. All modes, in a continuum from complete (allopatric) to zero (*sympatric*) spatial segregation of diverging groups, occur, mainly, in marine ecosystems.

In spatially extended population, another mode—*topopatric* (or *distance-forced*) speciation can occur via ibd only, without geographic isolation and selection. de Aguiar and Bat-Yam, 2011, gave the conditions for speciation in such population as a function of its density, mutation rate, genome size and **critical mating distances**. They see such speciation as a case of breakdown of unstable uniform distribution, leading to the self-organization of its members into clusters.

Absolute distances between diverged groups can be, say, tens of meters for pathogen resistance to hundreds of kilometers in marine invertebrates.

- **Wright**

  Dispersal neighborhood (DN) is the geographic area within which individuals and genes regularly move and interact. It is estimated as the area within a radius extending 2 standard deviations from the mean of species's dispersal distribution.

  Richardson et al., 2014, proposed to measure microgeographic adaptive evolution by the **wright** indicating the phenotypic difference between populations relative to the number of species-specific DN's separating them. It is

$$w = \frac{|x_1 - x_2|}{ds_p},$$

[redacted]

where x_1, x_2 are the means of the genetically determined traits of populations, s_p is the pooled standard deviation of those trait values across populations, and d is the distance in number of DN's separating the two populations.

The wright is a spatial analog of the *haldane*, a metric for rate of microevolution defined (Gingerich, 1993, and Lynch, 1990) as $\frac{|x_1-x_2|}{gs_p}$, where g the number of generations separating the populations (or samples of the same populations). Haldan, 1949, defined the *darwin* as $|\ln x_1 - \ln x_2|$ (or, respectively, $\frac{|\ln x_1 - \ln x_2|}{t}$), where t is the time in Ma separating samples of the same populations).

- **Malécot's distance model**

  Genealogy, migration and surname isonymy are used to predict kinship (usually estimated from blood samples). But because of incomplete knowledge on ancestors, pedigree-independent methods for kinship assays utilize the distance-dependent correlations of any parameter influenced by identity in descent: phenotype, gene frequency, or, say, isonymy.

  Malécot's distance model (1948, 1959) is expressed by the following *kinship-distance formula* for the mean **coefficient of kinship** between two populations **isolated by distance** d:

$$\theta_d = ae^{-bd}d^c,$$

  where $c = 0, \frac{1}{2}$ correspond to one-, two-dimensional migration, b is a function of the *systematic pressure* (joint effect of co-ancestry, selection, mutations and long range migration), and a is the *local kinship* (the correlation between random gametes from the same locality). In fact, the results in 2D for small and moderate distances agree closely with $c = 0$. The model is most successful when the systematic pressure is dominated by migration.

  Malécot's model was adapted for the dependency ρ_d of alleles at two loci at distance d (*allelic association*, *linkage disequilibrium*, *polymorphism distance*):

$$\rho_d = (1-L)Me^{-\epsilon d} + L,$$

  where d is the distance (say, from a disease gene) between loci along the chromosome (either **genome distance** on the physical scale in kilobases, or **map distance** on the genetic scale in centiMorgans), ϵ is a constant for a specified region, $M \leq 1$ is a parameter expressing mutation rate and L is the parameter predicting association between unlinked loci.

  Selection generates long *blocks of linkage disequilibrium* (places in the genome where genetic variations are occurring more often than by chance, as in the genetic drift) across hundreds of kilobases. Using it, Hawks et al., 2007, found that selection in humans much accelerated during the last 40,000 years, driven by exponential population growth and cultural adaptations.

  Examples of accelerated (perhaps, under diet and diseases pressures) human evolution and variation include disease resistance, lactose tolerance, skin color, skeletal gracility. A mutation in *microcephalin* (gene MCPH1) appeared

14,000–62,000 years ago and is now carried by 70 % of people but not in sub-Saharan Africa. Distinctive traits of East Asians (about 93 % of Han Chinese and 70 % of Japanese and Tai)—thicker hair shafts, more sweat glands, smaller breasts and specific teeth—are the result of a mutation in gene EDAR that occurred $\approx$35,000 years ago.

The fastest genetic change ever observed in humans is that the ethnic Tibetans split off from the Han Chinese less than 3000 years ago and since then evolved a unique ability to thrive at high (4000 m above sea level) altitudes and low oxygen levels. It also come from their gene *EPAS1* found only in Denisovans. An example of quick nongenetic evolution: the average height of a European male at age 21 rose from 167 cm in early 1870s to 178 cm in 1980. Crabtree, 2012, argues that our intellectual and emotional abilities diminish gradually (after a peak 2000–6000 years ago) because of weakened control of genetic mutations by natural selection.

Over the past 20,000 years, the average volume of the human brain has decreased by 10 %. Possible reason: our dwindling intelligence (Geary, 2011), or improved brain efficiency (Hawks, 2011), or social self-domestication (Hood, 2014).

23.2 Distances for DNA/RNA and Protein Data

The main way to estimate the genetic distance between DNA, RNA or proteins is to compare their nucleotide or amino acid, sequences, respectively. Besides sequencing, the main techniques used are immunological ones, *annealing* (cf. **hybridization metric**) and *gel electrophoresis* (cf. **read length**).

Distances between nucleotide (DNA/RNA) or protein sequences are usually measured in terms of substitutions, i.e., mutations, between them.

A *DNA sequence* is a sequence $x = (x_1, \dots, x_n)$ over the four-letter alphabet of four nucleotides A, T, C, G (or two-letter alphabet purine/pyrimidine, or 16-letter *dinucleotide* alphabet of ordered nucleotide pairs, etc.). Let $\sum$ denote $\sum_{i=1}^{n}$.

A *protein sequence* is a sequence $x = (x_1, \dots, x_n)$ over a 20-letter alphabet of 20 standard amino acids; $\sum$ again denotes $\sum_{i=1}^{n}$.

A short sequence is called *nullomer* if it do not occur in a given species and *prime* if it has not been found in nature. Hampikian–Andersen, 2007, lists 80 human DNA nullomers of length 11 and many primes: DNA of length 15 and protein of length 5.

For a macromolecule, a *primary structure* is its atomic composition and the chemical bonds connecting atoms. For DNA, RNA or protein, it is specified by its sequence. The *secondary structure* is the 3D form of local segments defined by the hydrogen bonds. The *tertiary structure* is the 3D structure, as defined by atomic positions. The *quaternary structure* describes the arrangement of multiple molecules into larger complexes.

[redacted]

- **Number of DNA differences**

  The **number of DNA differences** between DNA sequences $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ is the number of mutations, i.e., their **Hamming metric**:

$$\sum 1_{x_i \neq y_i}.$$

- ***p*-distance**

  The ***p*-distance** d_p between DNA sequences $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ is defined by

$$\frac{\sum 1_{x_i \neq y_i}}{n}.$$

- **Jukes–Cantor nucleotide distance**

  The **Jukes–Cantor nucleotide distance** between DNA sequences x and y is defined, using the ***p*-distance** d_p with $d_p \leq \frac{3}{4}$, by

$$-\frac{3}{4} \ln \left(1 - \frac{4}{3} d_p(x, y) \right).$$

  If the rate of substitution varies with the gamma distribution, and a is the parameter describing the shape of this distribution, then the **gamma distance for the Jukes–Cantor model** is defined by

$$\frac{3a}{4} \left(\left(1 - \frac{4}{3} d_p(x, y) \right)^{-1/a} - 1 \right).$$

- **Tajima–Nei distance**

  The **Tajima–Nei distance** between DNA sequences x and y is defined by

$$-b \ln \left(1 - \frac{d_p(x, y)}{b} \right), \quad \text{where}$$

$$b = \frac{1}{2} \left(1 - \sum_{j=A,T,C,G} \left(\frac{1_{x_i = y_i = j}}{n} \right)^2 + \frac{1}{c} \sum \left(\frac{1_{x_i \neq y_i}}{n} \right)^2 \right), \quad \text{and}$$

$$c = \frac{1}{2} \sum_{i,k \in \{A,T,G,C\}, j \neq k} \frac{\left(\sum 1_{(x_i, y_i) = (j,k)} \right)^2}{\left(\sum 1_{x_i = y_i = j} \right) \left(\sum 1_{x_i = y_i = k} \right)}.$$

Let $P = \frac{1}{n} |\{1 \leq i \leq n : \{x_i, y_i\} = \{A, G\} \text{ or } \{T, C\}\}|$, and $Q = \frac{1}{n} |\{1 \leq i \leq n : \{x_i, y_i\} = \{A, T\} \text{ or } \{G, C\}\}|$, i.e., P and Q are the frequencies of, respectively, transition and transversion mutations between DNA sequences x and y.

The following four distances are given in terms of P and Q.

- **Jin–Nei gamma distance**

  The **Jin–Nei gamma distance** between DNA sequences is defined by

$$\frac{a}{2}\left((1-2P-Q)^{-1/a}+\frac{1}{2}(1-2Q)^{-1/a}-\frac{3}{2}\right),$$

  where the rate of substitution varies with the gamma distribution, and a is the parameter describing the shape of this distribution.
- **Kimura 2-parameter distance**

  The **Kimura 2-parameter distance** $K2P$ (Kimura, 1980) between DNA sequences is defined by

$$-\frac{1}{2}\ln(1-2P-Q)-\frac{1}{4}\ln\sqrt{1-2Q}.$$

- **Tamura 3-parameter distance**

  The **Tamura 3-parameter distance** between DNA sequences is defined by

$$-b\ln\left(1-\frac{P}{b}-Q\right)-\frac{1}{2}(1-b)\ln(1-2Q),$$

  where $f_x=\frac{1}{n}|\{1\leq i\leq n: x_i=G \text{ or } C\}|, f_y=\frac{1}{n}|\{1\leq i\leq n: y_i=G \text{ or } C\}|$, and $b=f_x+f_y-2f_xf_y$. If $b=\frac{1}{2}$, it is the **Kimura 2-parameter distance**.
- **Tamura–Nei distance**

  The **Tamura–Nei distance** between DNA sequences is defined by

$$-\frac{2f_Af_G}{f_R}\ln\left(1-\frac{f_R}{2f_Af_G}P_{AG}-\frac{1}{2f_R}P_{RY}\right)-\frac{2f_Tf_C}{f_Y}\ln\left(1-\frac{f_Y}{2f_Tf_C}P_{TC}-\frac{1}{2f_Y}P_{RY}\right)-$$
$$-2\left(f_Rf_Y-\frac{f_Af_Gf_Y}{f_R}-\frac{f_Tf_Cf_R}{f_Y}\right)\ln\left(1-\frac{1}{2f_Rf_Y}P_{RY}\right),$$

  where $f_j=\frac{1}{2n}\sum(1_{x_i=j}+1_{y_i=j})$ for $j=A,G,T,C$, and $f_R=f_A+f_G, f_Y=f_T+f_C$, while $P_{RY}=\frac{1}{n}|\{1\leq i\leq n: |\{x_i,y_i\}\cap\{A,G\}|=|\{x_i,y_i\}\cap\{T,C\}|=1\}|$ (the proportion of transversion differences), $P_{AG}=\frac{1}{n}|\{1\leq i\leq n:\{x_i,y_i\}=\{A,G\}\}|$ (the proportion of transitions within purines), and $P_{TC}=\frac{1}{n}|\{1\leq i\leq n:\{x_i,y_i\}=\{T,C\}\}|$ (the proportion of transitions within pyrimidines).
- **Lake paralinear distance**

  Given two DNA sequences $x=(x_1,\dots,x_n)$ and $y=(y_1,\dots,y_n)$, denote by $det(J)$ the determinant of the 4×4 matrix $J=((J_{ij}))$, where $J_{ij}=\frac{1}{n}|\{1\leq t\leq n: x_t=i, y_t=j\}|$ (joint probability) and indices $i,j=1,2,3,4$ represent nucleotides A, T, C, G, respectively. Let $f_i(x)$ denote the frequency of the i-th nucleotide in the sequence x (marginal probability), and let $f(x)=f_1(x)f_2(x)f_3(x)f_4(x)$.

[redacted]

The **Lake paralinear distance** (1994) between sequences x and y is defined by

$$-\frac{1}{4}\ln\frac{det(J)}{\sqrt{f(x)f(y)}}.$$

It is a **four-point inequality metric**, and it generalizes trivially for sequences over any alphabet. Related are the **LogDet distance** (Lockhart et al., 1994) $-\frac{1}{4}\ln\det(J)$ and the symmetrization $\frac{1}{2}(d(x,y)+d(y,x))$ of the **Barry–Hartigan quasi-metric** (1987) $d(x,y)=-\frac{1}{4}\ln\frac{det(J)}{\sqrt{f(x)}}$.

- **Eigen–McCaskill–Schuster distance**

  The **Eigen–McCaskill–Schuster distance** between DNA sequences $x=(x_1,\dots,x_n)$ and $y=(y_1,\dots,y_n)$ is defined by

  $$|\{1\le i\le n:\{x_i,y_i\}\ne\{A,G\},\{T,C\}\}|.$$

  It is the number of *transversions*, i.e., positions i with one of x_i, y_i denoting a purine and another one denoting a pyrimidine.

- **Watson–Crick distance**

  The **Watson–Crick distance** between DNA sequences $x=(x_1,\dots,x_n)$ and $y=(y_1,\dots,y_n)$ is defined, for $x\ne y$, by

  $$|\{1\le i\le n:\{x_i,y_i\}\ne\{A,T\},\{G,C\}\}|$$

  It is the **Hamming metric** (**number of DNA differences**) $\sum 1_{x_i\ne\bar{y}_i}$ between x and the *Watson–Crick complement* $\bar{y}=(\bar{y}_1,\dots,\bar{y}_n)$ of y, where $\bar{y}_i=A,T,G,C$ if $y_i=T,A,C,G$, respectively. Let y^* be the reversal $(\bar{y}_n,\dots,\bar{y}_1)$ of $\bar{y}$.

  Hybridization is the process of combining complementary single-stranded nucleic acids into a single molecule. *Annealing* is the binding of two strands by the Watson–Crick complementation. *Denaturation* is the reverse process.

  A *DNA cube* is any maximal set of DNA n-sequences, such that, for any two x, y of them, it holds that $H(x,y)=\min_{-n\le k\le n}\sum 1_{x_i\ne y^*_{i+k \pmod n}}=0$. The **hybridization metric** (Garzon et al., 1997) between DNA cubes A and B is

  $$\min_{x\in A,y\in B}H(x,y).$$

- **RNA structural distances**

  An *RNA sequence* is a string over the alphabet $\{A,C,G,U\}$ of nucleotides (bases). Inside a cell, such a string folds in 3D space, because of pairing of nucleotide bases (usually, by bonds $A-U$, $G-C$ and $G-U$). The *secondary structure* of an RNA is, roughly, the set of helices (or the list of paired bases) making up the RNA. Such structure can be represented as a planar graph and further, as a rooted tree. The *tertiary structure* is the geometric form the RNA takes in space; the secondary structure is its simplified/localized model.

An **RNA structural distance** between two RNA sequences is a distance between their secondary structures. These distances are given in terms of their selected representation. For example, the **tree edit distance** (and other distances on rooted trees given in Chap. 15) are based on the rooted tree representation.

Let an RNA secondary structure be represented by a simple graph (V, E) with vertex-set $V = \{1, \dots, n\}$ such that, for every $1 \le i \le n$, $(i, i+1) \notin E$ and $(i,j), (i,k) \in E$ imply $j = k$. Let $E = \{(i_1, j_1), \dots, (i_k, j_k)\}$, and let (ij) denote the transposition of i and j. Then $\pi(G) = \prod_{t=1}^{k}(i_t j_t)$ is an involution.

Let $G = (V, E)$ and $G' = (V, E')$ be such planar graph representations of two RNA secondary structures. The **base pair distance** between G and G' is the number $|E \Delta E'|$, i.e., the **symmetric difference metric** between secondary structures seen as sets of paired bases.

The **Zuker distance** between G and G' is the smallest number k such that, for every edge $(i,j) \in E$, there is an edge $(i',j') \in E'$ with $\max\{|i - i'|, |j - j'|\} \le k$ and, for every $(k', l') \in E'$, there is an $(k, l) \in E$ with $\max\{|k - k'|, |l - l'|\} \le k$.

The **Reidys–Stadler–Roselló metric** between G and G' is defined by

$$|E \Delta E'| - 2T,$$

where T is the number of cyclic orbits of length greater than 2 induced by the action on V of the subgroup $\langle \pi(G), \pi(G') \rangle$ of the group Sym_n of permutations on V. It is the number of transpositions needed to represent $\pi(G)\pi(G')$.

Let $I_G = \langle x_i x_j : (x_i, x_j) \in E \rangle$ be the monomial ideal (in the ring of polynomials in the variables $x_1, \dots, x_n$ with coefficients 0, 1), and let $M(I_G)_m$ denote the set of all monomials of total degree $\le m$ that belong to I_G. For every $m \ge 3$, a Liabrés-Roselló **monomial metric** between $G = (V, E)$ and $G' = (V', E')$ is

$$|M(I_G)_{m-1} \Delta M(I_{G'})_{m-1}|.$$

Chen–Li–Chen, 2010, proposed the following variation of the **directed Hausdorff distance** (Chap. 1) between two intervals $x = [x_1, x_2]$ and $y = [y_1, y_2]$, representing two RNA secondary structures:

$$\max_{a \in x} \min_{b \in y} |a - b| \left(1 - \frac{O(x, y)}{x_2 - x_1 + 1}\right),$$

where $O(x, y) = \min\{x_2, y_2\} - \max\{x_1, y_1\}$, represents the overlap of intervals x and y; it is seen as a negative gap between x and y, if they are disjoint.

- **Fuzzy polynucleotide metric**

  The **fuzzy polynucleotide metric** (or **NTV-metric**) is the metric introduced by Nieto, Torres and Valques-Trasande, 2003, on the 12-dimensional unit cube I^{12}. Four nucleotides U, C, A and G of the RNA alphabet being coded as $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$ and $(0, 0, 0, 1)$, respectively, 64 possible triplet codons of the genetic code can be seen as vertices of I^{12}.

[redacted]

So, any point $(x_1, \dots, x_{12}) \in I^{12}$ can be seen as a *fuzzy polynucleotide codon* with each x_i expressing the grade of membership of element i, $1 \leq i \leq 12$, in the fuzzy set x. The vertices of the cube are called the *crisp sets*.

The **NTV-metric** between different points $x, y \in I^{12}$ is defined by

$$\frac{\sum_{1 \leq i \leq 12} |x_i - y_i|}{\sum_{1 \leq i \leq 12} \max\{x_i, y_i\}}.$$

Dress and Lokot showed that $\frac{\sum_{1 \leq i \leq n} |x_i - y_i|}{\sum_{1 \leq i \leq n} \max\{|x_i|, |y_i|\}}$ is a metric on the whole of $\mathbb{R}^n$. On $\mathbb{R}^n_{\geq 0}$ this metric is equal to $1 - s(x, y)$, where $s(x, y) = \frac{\sum_{1 \leq i \leq n} \min\{x_i, y_i\}}{\sum_{1 \leq i \leq n} \max\{x_i, y_i\}}$ is the **Ruzicka similarity** (Chap. 17).

- **Genome rearrangement distances**

  The *genomes* of related unichromosomal species or single chromosome organelles (such as small viruses and mitochondria) are represented by the order of genes along chromosomes, i.e., as *permutations* (or *rankings*) of a given set of n homologous genes. If one takes into account the directionality of the genes, a chromosome is described by a *signed permutation*, i.e., by a vector $x = (x_1, \dots, x_n)$, where $|x_i|$ are different numbers $1, \dots, n$, and any x_i can be positive or negative.

  The circular genomes are represented by circular (signed) permutations $(x_1, \dots, x_n)$, where $x_{n+1} = x_1$ and so on.

  Given a set of considered mutation moves, the corresponding *genomic distance* between two such genomes is the **editing metric** (Chap. 11) with the editing operations being these moves, i.e., the minimal number of moves needed to transform one (signed) permutation into another.

  In addition to (and, usually, instead of) local mutation events, such as character indels or replacements in the DNA sequence, the *large* (i.e., happening on a large portion of the chromosome) mutations are considered, and the corresponding genomic editing metrics are called **genome rearrangement distances**. In fact, such rearrangement mutations being rarer, these distances estimate better the true genomic evolutionary distance.

  The genome (chromosomal) rearrangements are *inversions* (block reversals), *transpositions* (exchanges of two adjacent blocks), *inverted transpositions* (inversions combined with transpositions) in a permutation, and, for signed permutations, *signed reversals* (sign reversal combined with inversion). The main genome rearrangement distances between two unichromosomal genomes are:

  Cayley, **reversal** and **signed reversal** metrics (cf. Chap. 11);

  ITT-distance: the minimal number of inversions, transpositions and inverted transpositions needed to transform one of them into another.

  Given two circular signed permutations $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ (so, $x_{n+1} = x_1$, etc.), a *breakpoint* is a number i, $1 \leq i \leq n$, such that $y_{i+1} \neq x_{j(i)+1}$, where the number $j(i)$, $1 \leq j(i) \leq n$, is defined by the equality $y_i = x_{j(i)}$.

The **breakpoint distance** (Watterson et al., 1982) between genomes, represented by *x* and *y*, is the number of breakpoints.

This distance and the **permutation editing metric** (the **Ulam metric** from Chap. 11: the minimal needed number of one-character transpositions) are used for the approximation of genome rearrangement distances.

- **Syntenic distance**

  This is a *genomic distance* between multi-chromosomal genomes, seen as unordered collections of *synteny sets* of genes, where two genes are *syntenic* if they appear in the same chromosome. The **syntenic distance** (Ferretti–Nadeau–Sankoff, 1996) between two such genomes is the minimal number of mutation moves—*translocations* (exchanges of genes between two chromosomes), *fusions* (merging of two chromosomes into one) and *fissions* (splitting of one chromosome into two)—needed to transfer one genome into another. All (input and output) chromosomes of these mutations should be nonempty and not duplicated.

  The above three mutation moves correspond to interchromosomal genome rearrangements which are rarer than intrachromosomal ones; so, they give information about deeper evolutionary history.

- **Genome distance**

  The **genome distance** between two loci on a chromosome is a physical distance: the number of base pairs (bp) separating them on the chromosome.

  In particular, the **intragenic distance** of two neighboring genes is the smallest distance in bp separating them on the chromosome. Sometimes, it is defined as the genome distance between the transcription start sites of those genes.

  Nelson, Hersh and Carrol, 2004, defined the *intergenic distance* of a gene as the amount of noncoding DNA between the gene and its nearest neighbors, i.e., the sum of upstream and downstream distances, where *upstream distance* is the genome distance between the start of a gene's first exon and the boundary of the closest upstream neighboring exon and *downstream distance* is the distance between the end of a gene's last exon and the boundary of the closest downstream neighboring exon. If exons overlap, the intergenic distance is 0.

- **Strand length**

  A single strand of nucleic acid (DNA or RNA sequence) is oriented *downstream*, i.e., from the $5'$ *end* toward the $3'$ *end* (sites terminating at the 5-th and 3-rd carbon in the sugar-ring; $5'$-phosphate binds covalently to the $3'$-hydroxyl of another nucleotide). So, the structures along it (genes, transcription factors, polymerases) are either downstream or upstream. The **strand length** is the distance from its 5' to 3' end. Cf. **end-to-end distance** (in Chap. 24) for a general polymer.

  For a molecule of *messenger RNA* (mRNA), the **gene length** is the distance from the *cap site* $5'$, where post-translational stability is ensured, to the *polyadenylation site* $3'$, where a poly(A) tail of 50–250 adenines is attached after translation.

- **Map distance**

  The **map distance** between two loci on a genetic map is the recombination frequency expressed as a percentage; it is measured in *centiMorgans* cM (or *map*

[redacted]

units), where 1 cM corresponds to a 1 % ($\frac{1}{100}$) chance that a segment of DNA will crossover or recombine within one generation. Genes at map distance 50 cM are unlinked.

For humans, 1.3 cM corresponds to a **genome distance** of 1 Mb (million bp). In the female this recombination rate (and so map distances) are twice that of the male. In males, the total length of intervals between linked genes is 2500 cM.

During meiosis in humans, there is an average of 2 to 3 crossovers for each pair of homologous chromosomes. The *intermarker meiotic recombination distance* (Dib et al., 1992) counts only meiotic crossovers. Mitotic crossover is rare.

- **tRNA interspecies distance**

  Transfer RNA (tRNA) molecules are necessary to translate *codons* (nucleotide triplets) into amino acids; eukaryotes have up to 80 different tRNAs. Two tRNA molecules are called *isoacceptor tRNAs* if they bind the same amino acid.

  The **tRNA interspecies distance** between species m and n is (Xue et al., 2003), averaged for all 20 amino acids, the *tRNA distance for a given amino acid* aa_i which is, averaged for all pairs, the **Jukes–Cantor protein distance** between each isoacceptor tRNAs of aa_i from species m and each isoacceptor tRNAs of the same amino acid from species n.

- **PAM distance**

  There are many notions of similarity/distance (20 × 20 *scoring matrices*) on the set of 20 standard amino acids, based on genetic codes, physico-chemical properties, secondary structural matching, structural properties (hydrophilicity, polarity, charge, shape, etc.) and observed frequency of mutations. The most frequently used one is the **Dayhoff distance**, based on the 20 × 20 *Dayhoff PAM250* matrix which expresses the relative mutability of amino acids.

  The **PAM distance** (or **Dayhoff–Eck distance**, *PAM value*) between protein sequences is defined as the minimal number of accepted (i.e., fixed) point mutations per 100 amino acids needed to transform one protein into another.

  1 PAM is a unit of evolution: it corresponds to 1 point mutation per 100 amino acids. PAM values 80, 100, 200, 250 correspond to the distance (in %) 50, 60, 75, 92 between proteins.

- **Genetic code distance**

  The **genetic code distance** (Fitch and Margoliash, 1967) between amino acids x and y is the minimum number of nucleotides that must be changed to obtain x from y. In fact, it is $1, 2$ or 3, since each amino acid corresponds to three bases.

- **Miyata–Miyazawa–Yasanaga distance**

  The **Miyata–Masada–Yasanaga distance** (or *Miyata's biochemical distance*, 1979) between amino acids x, y with polarities p_x, p_y and volumes v_x, v_y is defined by

$$\sqrt{(\frac{|p_x - p_y|}{\sigma_p})^2 + (\frac{|v_x - v_y|}{\sigma_v})^2},$$

  where σ_p and σ_v are the standard deviations of $|p_x - p_y|$ and $|v_x - v_y|$.

This distance is derived from the similar *Grantam's chemical distance* (Grantam, 1974) based on polarity, volume and carbon-composition of amino acids.

- **Polar distance (in Biology)**

  The following three physico-chemical distances between amino acids x and y were defined in Hughes–Ota–Nei, 1990.

  Dividing amino acids into two groups—*polar* (C, D, E, H, K, N, Q, R, S, T, W, Y) and nonpolar (the rest)—the **polar distance** is 1, if x, y belong to different groups, and 0, otherwise. The second polarity distance is the absolute difference between the polarity indices of x and y. Dividing amino acids into three groups—*positive* (H, K, R), *negative* (D, E) and *neutral* (the rest)—the **charge distance** is 1, if x, y belong to different groups, and 0, otherwise.

- **Feng–Wang distance**

  20 amino acids can be ordered linearly by their *rank-scaled* functions CI, NI of pK_a values for the terminal amino acid groups COOH and NH_3^+, respectively. 17 CI is 1,2,3,4,5,6,7,7,8,9,10,11,12,13,14,14,15,15,16,17 for C,H,F,P,N,D,R,Q,K,E, Y,S,M,V,G,A,L,I,W,T, while 18 NI is 1,2,3,4,5,5,6,7,8,9, 10,10,11,12,13,14,15,16,17,18 for N,K,R,Y,F,Q,S,H,M,W,G,L,V,E,I,A,D,T,P,C.

  Given a protein sequence $x = (x_1, \dots, x_m)$, define $x_i < x_j$ if $i < j$, $CI(x_i) < CI(x_i)$ and $NI(x_i) < NI(x_i)$ hold. Represent the sequence x by the augmented $m \times m$ *Hasse matrix* $((a_{ij}(x)))$, where $a_{ii}(x) = \frac{CI(x_i)+NI(x_i)}{2}$ and, for $i \neq j$, $a_{ij}(x) = -1$ or 1 if $x_i < x_j$ or $x_i \geq x_j$, respectively.

  The **Feng–Wang distance** ([FeWa08]) between protein sequences $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_n)$ is defined by

$$||\frac{\lambda(x)}{\sqrt{m}} - \frac{\lambda(y)}{\sqrt{n}}||_2,$$

  where $\lambda(z)$ denotes the largest eigenvalue of the matrix $((a_{ij}(z)))$.

- **Number of protein differences**

  The **number of protein differences** between protein sequences $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_n)$ is just the **Hamming metric** between protein sequences:

$$\sum 1_{x_i \neq y_i}.$$

- **Amino p-distance**

  The **amino p-distance** (or *uncorrected distance*) d_p between protein sequences $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_n)$ is defined by

$$\frac{\sum 1_{x_i \neq y_i}}{n}.$$

[redacted]

- **Amino Poisson correction distance**

  The **amino Poisson correction distance** between protein sequences x and y is defined, via the **amino p-distance** d_p, by

$$-\ln(1 - d_p(x, y)).$$

- **Amino gamma distance**

  The **amino gamma distance** (or *Poisson correction gamma distance*) between protein sequences x and y is defined, via the **amino p-distance** d_p, by

$$a((1 - d_p(x, y))^{-1/a} - 1),$$

  where the substitution rate varies with $i = 1, \dots, n$ according to the gamma distribution with the shape described by the parameter a. For $a = 2.25$ and $a = 0.65$, it estimates the **Dayhoff distance** and **Grishin distances**, respectively. In some applications, this distance with $a = 2.25$ is called simply the **Dayhoff distance**.
- **Jukes–Cantor protein distance**

  The **Jukes–Cantor protein distance** between protein sequences x and y is defined, via the **amino p-distance** d_p, by

$$-\frac{19}{20} \ln\left(1 - \frac{20}{19} d_p(x, y)\right).$$

- **Kimura protein distance**

  The **Kimura protein distance** between protein sequences x and y is defined, via the **amino p-distance** d_p, by

$$-\ln\left(1 - d_p(x, y) - \frac{d_p^2(x, y)}{5}\right).$$

- **Grishin distance**

  The **Grishin distance** d between protein sequences x and y can be obtained, via the **amino p-distance** d_p, from the formula

$$\frac{\ln(1 + 2d(x, y))}{2d(x, y)} = 1 - d_p(x, y).$$

- **k-mer distance**

  The **k-mer distance** (Edgar, 2004) between sequences $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_n)$ over a compressed amino acid alphabet is defined by

$$\ln\left(\frac{1}{10} + \frac{\sum_a \min\{x(a), y(a)\}}{\min\{m, n\} - k + 1}\right),$$

where a is any *k-mer* (a word of length k over the alphabet), while $x(a)$ and $y(a)$ are the number of times a occurs in x and y, respectively, as a *block* (contiguous subsequence). Cf. **q-gram similarity** in Chap. 11.

- **Santoni–Felici–Vergni proximity**

  Given a protein sequence $v = (v_1, \dots, v_N)$ and two amino acids a and b in it, let $(x_1, \dots, x_n)$ and $(y_1, \dots y_m)$ be arrays of occurrence positions in v of a and b, respectively. The **proximity in v between a and b** is (Santoni–Felici–Vergni, 2016)

$$P_v(a,b) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,m} \{|x_i - y_j|\}.$$

  Using the above function and machine learning, the authors found a way to discriminate between natural and random protein sequences. It shows how natural selection dominates the randomising effect of casual mutations in protein evolution.

- **Whole genome composition distance**

  Let A_k be the set of all $\sum_{i=1}^{k} 4^i$ nonempty words of length at most k over the alphabet of four RNA nucleotides. For an RNA sequence $x = (x_1, \dots, x_n)$ and any $a \in A_k$, let $g_a(x)$ be the number of occurrences of a as a *block* (contiguous subsequence) in x and $f_a(x)$ be $g_a(x)$ divided by the number of blocks of the same length in x.

  The **whole genome composition distance** (Wu et al., 2006) between RNA sequences x and y (of two strains of HIV-1 virus) is the Euclidean distance

$$\sqrt{\sum_{a \in A_k} (f_a(x) - f_a(y))^2}.$$

  The D^2 *distance* (Torney et al., 1990) is $\sum_{a \in A_k \setminus A_l} (g_a(x) - g_a(y))^2$ for some $l \le k$. The D_2 *statistic* (Lippert et al., 2002) is the number of k-word matches of x, y.

- **Huang ensemble distance**

  Consider a genome, represented by a nucleotide sequence $s = (s_1, \dots, s_l)$, of length l, where $s_i \in \{AC, G, T\}$. For $t = A, C, G, T$, define $w_t(s_i) = 1$ if $s_i = t$ and $w_t(s_1) = 0$, otherwise. The *natural vector* of s is defined to be the concatenation

$$(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_A, D_C, D_G, D_T)$$

  of three following vectors: the *counting vector* (n_A, n_C, n_G, n_T), where $n_t = \sum_{i=1}^{l} w_t(s_i)$, the *first moment of positions* $(\mu_A, \mu_C, \mu_G, \mu_T)$, where $\mu_t = \sum_{i=1}^{l} i \frac{w_t(s_i)}{n_t}$ and the *second moment of positions* (D_A, D_C, D_G, D_T), where $D_t = \sum_{i=1}^{l} (i - \mu_t)^2 \frac{w_t(s_i)}{l n_t}$.

[redacted]

The *k-mer vector* is $(m(1),\dots,m(4^k))$ where $m(j)$ is the number of times the j-th of 4^k possible strings of k nucleotides occurs as k consecutive elements in s.

A *multipartite*, i.e., consisting of several nucleic acid molecules, virus can be represented by a set of natural and k-mer counting vectors of its multiple genome segments. The **Huang ensemble distance** between such viruses X and Y is

$$\sqrt{\frac{N_{XY}^2}{\delta_N^2}+\frac{K_{XY}^2}{\delta_K^2}}$$

(Huang, 2016), where N_{XY}, K_{XY} are the **Hausdorff distances** between the sets of their natural vectors and between the sets of their k-mer counting vectors, and δ_N^2, δ_K^2 are the variances of the upper triangular matrix with elements N_{IJ} and K_{IJ} (for all pairs IJ of considered viruses), respectively.

N_{XY}, N_{XY} are $\max\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(x,y)\}$, where $d(x,y)$ is the Euclidean distance of two natural or k-mer counting vectors $x \in X$, $y \in Y$.

- **Additive stem w-distance**

  Given an alphabet $\mathcal{A}$, let $w = w(a,b) > 0$ for $a, b \in \mathcal{A}$, be a weight function on it. The **additive stem w-distance** between two n-sequences $x, y \in \mathcal{A}^n$ is defined (D'yachkov and Voronina, 2008) by

  $$D_w(x,y) = \sum_{i=1}^{n-1}(s_i^w(x,x) - s_i^w(x,y)),$$

  where $s_i^w(x,y) = w(a,b)$ if $x_i = y_i = a, x_{i+1} = y_{i+1} = b$ and $s_i^w(x,y) = 0$, otherwise. If all $w(a,b) = 1$, then $\sum_{i=1}^{n-1} s_i(x,y)$ is the number of common 2-blocks containing adjacent symbols in the longest common subsequence of x and y; then $D_w(x,y)$ is called a **stem Hamming distance**.
- **ACS-distance**

  Given an alphabet $\mathcal{A}$, the *average common substring length* between sequences $x = (x_1,\dots,x_m)$ and $y = (y_1,\dots,y_n)$ over $\mathcal{A}$ is (Ulitsky et al., 2006) $L(x,y) = \frac{1}{m}\sum_{i=1}^{m} l_i$, where l_i is the length of the longest substring $(x_i,\dots,x_{i-1+l_i})$ which matches a substring of y. So, $L(x,x) = \frac{m+1}{2}$.

  The **ACS-distance** is defined by

  $$\frac{1}{2}\left(\frac{\log(n)}{L(x,y)} - \frac{\log(m)}{L(x,x)} + \frac{\log(m)}{L(y,x)} - \frac{\log(n)}{L(y,y)}\right).$$

  A similar distance was considered (Haubold et al., 2009) replacing the longest common substring by the shortest absent one.

23.3 Distances in Ecology, Biogeography, Ethology

Main distance-related notions in Ecology, Biogeography and Animal Behavior follow.

- **Niche overlap similarities**

  A *niche* is environmental space, while a *biotope* is a geographic space.

  Let $p(x) = (p_1(x), \dots, p_n(x))$ be a *frequency vector* (i.e., all $p_i(x) \geq 0$ and $\sum_i p_i(x) = 1$) representing an ecological niche of species x, for instance, the proportion of resource i, $i \in \{1, \dots, n\}$, used by species x.

  Four main **niche overlap similarities** of species x and y are:

  Schoener's D, introduced by Renkonen in 1938:

$$D(x,y) = 1 - \frac{1}{2} \sum_{i=1}^{n} |p_i(x) - p_i(y)|;$$

  cosine similarity (Chap. 17), called in Ecology (from 1973) *Pianka's O*:

$$O(x,y) = \frac{\langle p(x), p(y) \rangle}{||p(x)||_2 \cdot ||p(y)||_2};$$

  Hellinger I (i.e., **fidelity similarity**, cf. Chap. 14) and **Bray–Curtis** (or, since $p(x), p(y)$ are frequency vectors, *Renkonen percentage*) **similarity** (Chap. 17).

- **Ecological distance**

  Let a given species be distributed in subpopulations over a given *landscape*, i.e., a textured mosaic of *patches* (homogeneous areas of land use, such as fields, lakes, forest) and linear *frontiers* (river shores, hedges and road sides). The individuals move across the landscape, preferentially by frontiers, until they reach a different subpopulation or they exceed a maximum **dispersal distance**.

  The **ecological distance** between two subpopulations (patches) x and y is defined (Vuilleumier–Fontanillas, 2007) by

$$\frac{D(x,y) + D(y,x)}{2},$$

  where $D(x,y)$ is the distance an individual covers to reach patch y from patch x, averaged over all successful dispersers from x to y. If no such dispersers exist, $D(x,y)$ is defined as $\min_z(D(x,z) + D(z,x))$.

  Ecotopes are the smallest ecologically distinct features in a landscape mapping.

  Effective ecological distance (or *cost-distance*) is the Euclidean distance modified for the effect of landscape and behavior on the dispersal of an organism between locations in the landscape. Such functional distance can be computed as *least-cost path* using either *cost surface* (matrix between patches assigning degree of permeability depending on hostile habitat or physical barriers), or

[redacted]

resistance surface accounting for costs (resistance per unit distance) of passing through different landscape elements. Cf. **resistance metric** in Chap. 15. Pinto–Kein, 2009, proposed *least-cost corridors* formed by multiple paths with similar costs, since animals, even birds, rarely move along straight-line paths in a landscape.

- **Biotope distance**

  The *biotopes* here are represented as binary sequences $x = (x_1, \dots, x_n)$, where $x_i = 1$ means the presence of the species i. The **biotope** (or **Tanimoto**, cf. Chap. 17) **distance** between biotopes x and y is defined by

  $$\frac{|\{1 \leq i \leq n : x_i \neq y_i\}|}{|\{1 \leq i \leq n : x_i + y_i > 0\}|} = \frac{|X \Delta Y|}{|X \cup Y|},$$

  where $X = \{1 \leq i \leq n : x_i = 1\}$ and $Y = \{1 \leq i \leq n : y_i = 1\}$.

- **Prototype distance**

  Given a finite metric space (X, d) (usually, a Euclidean space) and a selected, as typical by some criterion, vertex $x_0 \in X$, called the *prototype*, the **prototype distance** of every $x \in X$ is the number $d(x, x_0)$.

  Usually, the elements of X represent phenotypes or morphological traits. The average of $d(x, x_0)$ over $x \in X$ estimates the corresponding *variability*.

- **Critical domain size**

  In Spatial Ecology, the **critical domain size** is (Kierstead and Slobodkin, 1953) the minimal amount of habitat, surrounded by a hostile matrix, required for a population to persist. For example, in the invasion and persistence of algal and insect populations in rivers, such a size is the minimal length of a river (with a given, less than the threshold, flow speed) that supports a population.

- **Island distance effect**

  An *island*, in Biogeography, is any area of habitat surrounded by areas unsuitable for the species on the island: true islands surrounded by ocean, mountains isolated by surrounding lowlands, lakes surrounded by dry land, isolated springs in the desert, grassland or forest fragments surrounded by human-altered landscapes.

  The **island distance effect** is that the number of species found on an island is smaller when the degree of isolation (distance to nearest neighbor and mainland) is larger. Also, organisms with high dispersal, such as plants and birds, are much more common on islands than are poorly dispersing taxa like mammals.

- **Dispersal distance**

  In Biology, the **dispersal distance** is a **range distance** to which a species maintains or expands the distribution of a population. It refers, say, to seed dispersal by pollination and to natal, breeding and migration dispersal. For animals, *natal dispersal* is permanent emigration from the natal range to a disjoint adult range, and **dispersal distance** is the distance between their barycenters.

  The dispersal distance is usually described by a *dispersal kernel* which gives the probability distribution of the distance traveled by any individual.

When *outcrossing* (gene flow) is used to increase genetic diversity of a plant species, the *optimal outcrossing distance* is the dispersal distance at which seed production is maximized. It is less than the mean pollen dispersal distance.

Plant height matters more than seed mass for its dispersal distance. Unusual way of wind dispersal include tumbleweeds.

Pollen from *Pinus sylvestris* can fly 100 km, but oceanic larvae dispersal is at least one order of magnitude greater than that of pollen-dispersing terrestrial biotas.

- **Long-distance dispersal**

Long-distance dispersal (or *LDD*) refers to the rare events of biological dispersal on distances an order of magnitude greater than the median **dispersal distance**. For the regional survival of some plants, LDD is more important than local (median-distance) dispersal. The longest recorded distance traveled by a drift seed is 28,000 km by a Mary's bean from the Marshall Islands to Norway.

LDD emerged in Biogeography as greater factor of biodiversity and species migration patterns than original *vicarience theory* (dispersal via land bridges) based on continental drift. Such relatively recent chance dispersal explain the fast spread of organisms in new habitats, for example, plant pathogens, invasive species and in paleocolonization events, such as the joining of North and South America 3 Ma ago, or Africa and India with Eurasia 30 and 50 Ma ago.

Human colonization of Madagascar (isolated for 88 Ma) $\approx$2000 years ago may have resulted from an accidental transoceanic crossing; other animals arrived by rafting from Africa 60–70 Ma ago. LDD followed traders and explorers, especially, in Columbian Exchange after 1492.

Transoceanic LDD by wind/water currents can explain strong floristic similarities among landmasses in the southern hemisphere. Monkey, rodents, and crocodiles dispersed 50–30 Ma ago to the Americas from Africa via the Atlantic.

Free-living microbes occupy every niche but their biodiversity is low, because they are carried by wind thousands of km on dust particles protecting them from UV. Extreme example of such (or via underground rivers, before the continents split) LDD: sunlight-independent bacterium *Desulforudis audaxviator*, living 1–3.3 km deep in South Africa (the only species known to be alone in its ecosystem and radiation-relying), reached deep boreholes in eastern California.

Some other LDD vehicles are: rafting by water (corals can traverse 40,000 km during their lifetime), migrating birds, human transport, ship ballast water, and extreme climatic events. Snails can travel hundreds of km inside bird guts: 1–10 % of eaten snails survive up to 5 h until being ejected in bird feces.

Also, cancer invasion (spread from primary tumors invading new tissues) can be thought as an invasive species spread via LDD, followed by localized dispersal.

The most invasive mammal species (besides humans) are: rabbits, black rats, gray squirrels, goats, pigs, deers, mice, cats, red foxes, mongooses. Invasive *Argentine ants* form the largest global insect mega-colony: they do not attack each other.

[redacted]

- **Migration distance (in Biogeography)**

  Migration distance is the distance between regular breeding and foraging areas within seasonal large-scale return movement of birds, fish, insects, turtles, seals, etc.

  The longest such recorded round-trip is 71,000 km pole-to-pole traveled each year by the Arctic tern. The highest migration altitude is 9 km by bar-headed goose. Longest each way migration for a mammal is $\approx$9800 km, traveled by a humpback whale from the Brazilian coast to Madagascar, and, for an insect, $\approx$4500 km by desert locust and Monarch butterfly.

  One of unsolved problems in Biology is: how do the descendants of Monarch butterfly, migrating from Canada to central Mexico for several generations, manage to return to a few small overwintering spots? Also, why the ancient murrelets fly almost 8000 km across entire North Pacific and back—for no obvious benefit?

  Migration differs from *ranging*, i.e., the movement of an animal beyond its home range which ceases when a suitable new home range (a resource: food, mates, shelter) is found. It differs also from foraging/commuting as occurs, say, for albatrosses or plankton. Wandering albatrosses, having the largest (3.63 m) wingspan, make several-days foraging round trips of up to 3000 km. Krill, 1–2 cm long, move up to 500 m vertically each night, to feed in the sunlit waters, where plants are abundant, while avoiding being seen by predators. *Mesopelagic* (living 0.2–1 km deep) fish also travel to upper layers at night.

  At the population level, migration involves displacement, while ranging/foraging result only in mixing. Entire species migrate slowly by shifting, because of rapid climate change, their geographical or elevation ranges. Chen et al., 2011, found that the mean travel poleward of over 1300 species, considered by them, was 17 km a decade; vertically, the mean had a species moving 11 m uphill.

  During ice ages species move to hotspots, say, volcanoes. In early Holocene, forests migrated northward, up to 5 km per decade, using LDD (by birds) and low-density founder populations. Root et al., 2003, claim that butterflies, birds and plants move towards the poles by 6.1 km per decade over the past 100 years.

  Lyons et al., 2015: during last 300 Mya, more pairs of land species with correlated co-occurrence were positively correlated than negatively correlated ones, and 6000 years ago this pattern was reversed. They explains it by barriers to dispersal introduced by increasing human population and the spread of agriculture.

- **Daily distance traveled**

  Daily distance traveled D (m/day) is an important parameter of the energy budget of ranging/foraging mammals.

  The *foraging efficiency* is the ratio $\frac{B}{C}$, where C, B (J/m) are the energy costs of travel and of acquiring energy. Over a day, the expected total net energy return is $D(B - C)$. The *locomotor cost* is the distance traveled per unit energy spent on locomotion. The limb length determines this cost in terrestrial animals but no link with D has been observed. Pontzer, 2011, explains this paradox by high $\frac{B}{C}$ in most taxa: only for $\frac{B}{C} < 10$, would selection for limb length be needed.

Within species, over a lifetime, increased D is associated with decreased $B - C$, reproductive effort and maintenance. But among species, over evolutionary time, it is associated with a greater number of offspring and their total mass per lifetime.

The mean D traveled by carnivores is four times such distance by herbivores. Also, D and feeding/grooming time are much greater in larger groups of primates. Foraging radius, D and annual travel distance of Neanderthal was $\approx 75\,\%$ of that of humans.

- **Collective motion of organisms**

Organisms aggregate to procure resources (pack-hunting), to find mates (plankton, plants) and to lower predation risk (meerkats, schools of sardines, flocks of starlings). Animals moving in large groups at the same speed and in the same direction, tend to have similar size and to be regularly spaced.

The near-constant distance which an animal maintains from its immediate neighbors is called the **nearest-neighbor distance** (NDD). When NDD decreases, the mode of movement can change: marching locusts align, ants build bridges, etc.

Moving in file when perception is limited to one individual (ants, caterpillars in processions up to 300, spiny lobsters in parallel chains of 3–30), animals use tactile cues or just perceive and follow the choice of the preceding individual, such as sheep in mountain path or cows in cattle-handling facilities. Penguins in the huddle move ("traveling wave", like the stop-and-go of cars in a traffic jam) trigger movements in their neighbors as soon as the threshold distance (≈ 2 cm, i.e., twice the thickness of their compressive feather layer) is formed between two penguins.

The greatest recorded group of moving animals was a swarm in US, 1875, by 12.5 trillion insects (Rocky Mountain locust, extinct by now) covering 510,000 km^2. A swarm by extant desert locusts in Kenya, 1954, covered 200 km^2.

Flights of migratory pest insects occur usually at altitudes up to 1 km, and are downwind; they last for a few hours with displacement up tp 400 km. Flocks of *red-billed Quelea* (the most abundant wild bird species) take up to 5 h to fly past. Herring schools occupy up to 4.8 km^3 with density 0.5–1.0 fish per m^3. Jellyfish *Rhopilema nomadica* form swarms 100 km long.

Schools of sardines, anchovy and krill, despite being variable in size, share a ratio $\frac{S}{V} = 3.3$ m^{-1} of surface area to volume; it has been interpreted as the optimal trade-off between predator avoidance and resource acquisition.

The spatiotemporal movement patterns, emerging from such groups, result from interactions between individuals. This local mechanism can be *allelomimesis* ("do what your neighbor does"), social attraction (say, to the center of mass of neighbors), or the threat of cannibalism from behind (in running Mormon crickets and desert locusts), mass mate-searching (in burrow-dwelling crabs). Vicsek, 1995, modelled a swarm as a collection of particles moving with a constant

[redacted]

speed but adopting, if perturbation, at each time increment the average motion's direction of the neighbors.

Migrating birds tend to fly in a V, J, or W shaped formation. In energy-saving V-formation (or *skein*), they sync their flapping to hook the former bird's updraft. The birds flying at the front and the tips are rotated. The changes in distance between members and in direction propagate like sound waves: for distance better in bigger flocks, while for direction better in smaller flocks.

Most spectacular are aerial displays of flocks of starlings highly variable in shape. Scale-free behavioral correlation was observed: regardless of flock size, the correlations of a bird's orientation and velocity with the other birds did not vary and was near-instantaneous. Cf. *SOC* in **scale invariance** (Chap. 18).

Silverberg et al., 2013, discovered self-organized emergent behavior in *moshing* (when 100–100,000 fans at heavy metal concert form circles and then run together with abandon, bouncing off one another). In fact, the speed distribution of people closely matches that of molecules in a 2D gas at equilibrium and moshing corresponds to domination of the model's parameters by noise.

Such emerging, when their number increases, collective behavior can be seen as a *critical phase transition*; it was observed also for simple automatons.

Besides animals, collective directed motion occurs also in cellular populations. Some aggregated bacterial populations (say, foraging swarms of billions of *Paenibacillus vortex*) can migrate rapidly and coordinately over a surface. A *grex* is an slug-like aggregate 2–4 mm long of up to 100,000 amoebas formed when they are under stress. It moves as a unit, only forward, 1 mm/h. Other example of small-scale moving super-organism: oceanic *pyrosomes* (clonal colonies of zooids-tunicates).

In a multicellular organism, collective cell migration occurs (usually by *chemotaxis*: response to chemical concentration) throughout embryonic development, wound healing (fibroplasts and epithelial cells), immune response (leukocytes), and cancerous tumor invasion. Similarly to migration of songbirds, cancerous cells prepare for metastatic travel by gathering proteins near their leading edges.

During development, some cells migrate to very long distances. For example, newborn neurons in the adult brain can traverse $\frac{2}{3}$ of its length.

- **Distances in Animal Behavior**

  The first such distance was derived by Hediger for zoos; his *interanimal distance* is the maximum species-specific distance at which conspecifics approach each other. In 1955, he defined *flight distance* (run boundary), *critical distance* (attack boundary), *personal distance* (at which members of noncontact species feel comfortable) and *social distance* (at which within-species groups tolerate each other).

  The exact such distances are highly context dependent. An example: a tamer manipulate a semi-tamed lion moving in and out of its critical zone.

  For humans, flight and critical distances have been, with few exceptions, eliminated. So, Hall adapted above space boundaries; cf. his **distances between people** in Chap. 28. The main **distances in Animal Behavior** follow.

The **individual distance**: the distance which an animal attempts to maintain between itself and other animals. It ranges between "proximity" and "far apart" (for example, $\leq$ 8 m and $\geq$ 61 m in elephant social calls). Bell et al., 2012, found that gaining and maintaining a preferred interanimal distance, accounts for much of the variability in dodging by rats and field crickets.

The **group distance**: the distance which a group of animals attempts to maintain between it and other groups. Cf. the **nearest-neighbor distance**.

The **alert distance**: the distance from the disturbance source (say, a predator or a dominating conspecific) when the animal changes its behavior (say, turns towards as *perception advertisement*) in response to an approaching threat.

The **flight initiation distance** (or FID, *escape distance*): the distance from the disturbing stimulus when escape begins. FID, corrected for the *distance to refuge*, is a measure of animal's boldness.

The **reaction distance**: the distance at which the animal reacts to the appearance of prey; *catching distance*: the distance at which the predator can strike a prey.

The **detection distance**: the maximal distance from the observer at which the individual or cluster of them is seen, heard, or detected by some other remote way. For example, it is 2000 m for an eagle searching for displaying sage-grouse and 1450 m for a sage-grouse scanning for a flying eagle.

The **social recognition distance**: the distance over which a contact call can be identified as belonging to a family.

In the main nonresource-based mating system, *lek mating*, females in estrous visit a congregation of displaying males, the *lek* and mate preferentially with males of higher **lekking distance rank**, i.e., relative distance from male territory (the median of his positions) to the center of the lek. High-ranking individuals have smaller, centrally located (so, less far to travel and more secure) home ranges.

The **distance-to-shore**: the distance to the coastline used to study clustering of whale strandings (by distorted echo-location, anomalies of magnetic field, etc.).

The **sleeping distance** of a mating pair: for example, it is no further than 0.5 m in Arctic blue fox, but more than 2–2.5 m one month after copulation.

- **Snout-vent length**

  The **snout-vent length** (SVL) or *nose–cloaca distance* (NCD) is an animal's length taken from the tip of the nose (*snout*) to the cloaca's opening (*vent*) at the tail base.

  The **gape distance** is the width of the widely opened mouth of a vertebrate.

- **Animal proximity network**

  An *animal proximity network* is a digraph of individuals with a pair of them forming an edge or arc in the case of spatiotemporal co-occurrence using some range criteria: a mutual distance threshold or some distance rule (say, nearest neighbor) within a given distance around a focal individual.

  In general, the amount of time spent in proximity to another individual does not correlate with the proportion of social interactions between pairs of individuals. But proximity is a good proxy for interaction rates in mixed-species

[redacted]

bird flocks, for social information or disease transmission, and for some social behaviors, such as foraging, grooming, sleeping, mate guarding, and parent–offspring relationships.

- **Animal depth/distance perception**

  Many animals, including humans, have two eyes with overlapping visual fields that use parallax (cf. **parallax distance** in Chap. 26) for depth perception and distance estimation. Some animals (for example, pigeons) use *motion parallax* in which they move head to gain different viewpoints. Anoter example: the velocity of the mantis's head movement is kept constant during peering. So, the distance to the target (prey) is inversely proportional to the velocity of the retinal image.

  All animals have a binocular region (growing as eyes become more forward-facing) which allows for vision through the clutter, as long as the width of the objects causing clutter is less than the **interpupillary distance** (cf. Chap. 29) d.

  Changizi–Shimojo, 2008, suggested that the degree of binocular convergence is selected to maximize how much the animal can see. Most animals exist in noncluttered environments or surroundings where the cluttering objects are bigger in size than d. They tend to have sideways-facing eyes allowing panoramic vision. But humans and other large mammals evolved in leafy environments like forests and their forward-facing eyes (and smaller distance d) maximize ability to see.

  Still, the human eye sockets are much wider relative to their height and the outer margin is recessed much further back in the human skull, than those of the apes. So, unlike "forest view" of other apes, our "savannah eyes" can have a lateral view that is unimpeded by the bones of the skull.

  Larvae of the sunburst diving beetle (*Thermonectus marmoratus*) have 6 pairs of eyes. Four eyes of two frontal pairs (used to scan potential prey) have bifocal lenses and at least two 1D-retinas: distant and close-up. The two focused images produced by the lens sit at different distances and vertically separated.

- **Distance-related animal settings**

  Spatial fidelity zones specific to individuals (say, at a given distance from a colony center, or within a particular zone of the total foraging area) have been observed for some social insect species, molluscan communities, birds, etc.

  Home range is the area where an animal (or a group) lives and travels within. Within it, the area of intensive and exclusive use by resident animals is the *core area*. The *distance between range centroids* of two individuals (or groups) is a parameter used in studies of spatially based animal social systems. Cf. **dispersal distance**.

  An animal is **territorial** if it consistently occupies, marks and defends a *territory*, i.e., an area with a focused resource (say, a nest, den, mating site or sufficient food resources). Territories are held by an individual, a mated pair, or a group. An *extraterritorial foray* is the movement of a territorial animal into a conspecific's territory.*Dear enemy recognition* is the stronger response of a territorial animal to strangers than to its neighbors from adjacent territories.

  Lizards *Varanus panoptes* dig, in a spiral, the deepest known, $\leq$ 3.6 m, nests.

A linear relationship between the logarithm of faecal egg (of internal parasites) counts and the mean distance moved by a sheep per time step was found.

The **defense region** is the region that a male must defend in a mating competition to monopolize a female. It can be 1D (burrow, tunnel), 2D (dry land), bounded 3D (tree, coral reef), or open 3D (air, water). Puts, 2010, claims that 1D and 2D (as for humans) mating environments favor the evolution of contests.

The reliability of a threat display in animal contests is maintained by the *proximity risk*, i.e., the display is credible only within a certain distance of the opponent. This threshold distance is related to weaponry and the species-specific fighting technique. Here, greater formidability and dominance can be reached solely behaviorally; for example, an elephant's musth status overrides its body size and tusks.

The **landscape of fear** of a foraging animal is defined by the spatial variation of presumed predation risk. Its horizontal and vertical components correspond to terrestrial and aerial predators. It include clearness of sightlines (to spot predators), shrubs/trees/edge cover and the interplay of the distances to food and shelter. For example, small fish stay close to the coral reef when grazing seaweed; this creates "grazing halos" of bare sand, visible from space, around all reefs. Similar natural features are "fairy rings" of green eelgrass (up to 1500 m in width, off Denmark's coast), of mushrooms (10–600 m) and of barren sand (2–15 m, in Africa).

The **domain of danger** (DOD, or **Voronoi polygon**, cf. Chap. 20) of an animal, risking predation, in aggregation is the area closer to it than to any other group member. *Selfish herd theory* (Hamilton, 1971) posits that a cover-seeking dominant animal tends to minimize its DOD by occupying the center, thus diluting its risk by placing another individual between itself and a predator or parasite. Moreover, some fish bite a group member, when exposed to a searching predator.

During traveling, dominant animals are closer to the front of the herd. During foraging, their trajectories are shorter, more direct and more aligned both with their nearest neighbors and with the whole herd.

Romey et al., 2015: for beetles, *flash expansion of repulsive herd*, i.e., a rapid disassembly of a prey group detecting a predator, is away from the centroid rather than from density maximum or starter. The starter tend to be a female at the group's edge; she moved more quickly than others and favoured the group's centroid.

Distance senses include sight, hearing, and smell (they can be in stereo), while *contact senses* include taste, the senses of pressure, thermoception, and *internal senses* include the sense of balance and muscle stretch. The buzzard can see small rodents from a height of 4.6 km. The spotted hyaena hears noises from predators feeding on carcasses over distances of up to 10 km. The silkmoth detects pheromones up to 11 km distant. The grizzly bear smells food from up to 29 km away.

An example of unexplained **distance prediction** by animals is given (Vannini et al., 2008) by snails *Cerithidea decollata* migrating up and down mangrove

[redacted]

shores in synchrony with tidal phases. In the absence of visual cues and chemical marks, snails cluster just above the high water line, and the distance from the ground correlates better with the incoming tide level than with previous ones.

Ants initially wander randomly and upon finding food return to their colony while laying down pheromone trails. So, when one ant finds a **shortest path** to a food source, other (and eventually all) ants are likely to follow it. Inspired by this idea, the *ant colony optimization algorithm* (ACO) is a probabilistic technique for finding shortest paths through graphs; cf. **arc routing problems** in Chap. 15. Also, ants routinely find the maximal distance from all entrances to dispose of dead bodies.

The **distance effect avoidance** is the observed selection of some good distant source of interest over a poor but nearer one in the same direction. For example, females at a chorusing lek of anurians or arthropods may use the lower pitch of a bigger or better distant male's call to select it over a weaker but louder call nearby. High-quality males help them by timing their calls to precede or follow those of inferior males. Ant colonies can select a good distant nest over a poorer one in the way, even when it is 9 times closer. Ants compensate for the distance effect by increasing recruitment latencies and quorum thresholds at nearby poor nests. Another example: foraging vervet monkeys optimize travel distance when alone but prioritize high-reward food sites when in competition.

In land locomotion, animals crawl, walk, run, hop, climb or crawl, slither, undulate. In fluids (water, air) animals swim and fly by beating flagella, tails, wings, undulating their bodies, or actuating pumps. Some animals can switch the medium. Fish *Exocoetidae* can spend 45 s in flight gliding up to 200 m at altitudes of up to 6 m; using waves, it can span distances up to 400 m. Some squids fly in shoals covering up to 50 m at 6 m above the water. Squirrels *Petauristinae*, snakes *Chrysopelea* and lemurs *Dermoptera* can glide with small loss of height up to 200, 100 and 70 m, respectively. The deepest dive for a flying bird is 210 m by a thick-billed murre. The deepest (3 km) and longest (2.3 h) mammal's dive was recorded for a Cuvier's beaked whale in 2014. The human free-diving record is 253.2 m by Nitsch in 2012.

Flying and swimming animals can move through volumes with six degrees of freedom: 3 translational (left/right, forwards/backwards, up/down) and 3 rotational (pitch, roll, yaw). Surface-constrained animals have only three degrees: left/right, forwards/backwards and yaw; moving in 3D, they have higher place field resolution in the horizontal plane and showed a preference for movement in the horizontal.

Maximal relative speed (in body lengths per second) is 6.2, 16 and 322 for human (Usain Bolt, 2009), cheetah and a mite *Paratarsotomus macropalpis*. But Meyer-Vernet and Rospars, 2015, showed that it is (except flying and very large organisms) $\sim b_m \rho \sigma^{-1} \sim 10$ across 20 orders of magnitude in body mass. Here $\rho \approx 10^3$ kg m^{-3} is the density of liquid water, while $\sigma \approx 2 \times 10^5$ N m^{-2} and $b_m \approx 2 \times 10^3$ W kg^{-1}, roughly constant across life forms, are the applied force per unit cross-sectional area of tissue and maximum metabolic rate per unit of working tissue.

Navigating animals use an egocentric orientation mechanism and simple panoramic views, within which proximal objects dominate because their image on the retina change significantly with displacement. Animals rely on the spatial arrangement of the objects/landmarks across the scene rather than on their individual identification and geometric cues. Humans and, perhaps, chimpanzees and capuchin monkeys, possess, in addition, an allocentric reference system, centered on objects/features of the environment, and a more flexible geometric representation of space, with true distance and direction, i.e., closer to an abstract mental map.

Gaze monitoring and *pointing*: four great apes, canids and ravens follow another's head and eye orientation into distant space, even behind an obstacle. Moreover, bonobos and chimpanzees take barrier opacity into consideration. African elephants can use communicative intent of human pointing as a cue to find food. Horses use their facial expressions (direction of eyes and ears) to "talk" to other horses.

Great apes, dolphins, elephants, magpies recognise themselves in mirrors. *Metacognition* (cognitive self-awareness) was found in great apes, dolphins and rhesus monkeys. A basic *Theory of Mind* (ability to attribute mental states), mental time travel, meta-tool use and empathy are expected in primates and corvides. Ravens are as clever as chimpanzees, despite having much smaller brains.

Mammals, birds and octopuses possess neurological substrates generating consciousness. Chimpanzees have a system of intentional communication. But shared intentionality and cumulative culture seems to be uniquely human.

- **Animal communication**

Only humans, songbirds, hummingbirds, parrots, cetaceans and bats have complex, learned vocalisation. Conceptual generalizations (bottlenose dolphins can transmit up to 9 km identity information independent of the caller's voice/location), syntax (calls of some monkeys, Bengal finches, Japanese great tits are built as "word sequences") and meta-communication ("play face" and tail-to-the-right signals in dogs that the subsequent aggressive signal is a play) have been observed.

Matters of relevance *at a distance* (a distant food source or shelter) can be communicated by body language. For example, honeybees dancing convey (by duration of the waggle phase and orientation of the dance relative to the vertical) the polar coordinates (distance D to the goal and angle between the direction towards it and to that of the Sun's azimuth) of locations of interest. The mean number of waggings of bee's *waggle phases* increases with D. Also, wolves, before a hunt, howl to rally the pack, become tense and have their tails pointing straight. Dogs express their spatial needs by body language and vocalizations. Stiffness, pilorection, aggressive barking/lunging are *distance-increasing*, while play bow, tail wagging to the right, "positive" barking/lunging are *distance-decreasing signals*.

A **distance pheromone** is a soluble (for example, in the urine) and/or evaporable substance emitted by an animal, as a chemosensory cue, in order to

[redacted]

send a message (on alarm, sex, food trail, recognition, etc.) to other members of the same species. In contrast, a *contact pheromone* is such an insoluble nonevaporable substance; it coats the animal's body and is a contact cue. The *action radius* of a distance pheromone is its attraction (or repulsion) range, the maximum distance over which animals can be shown to direct their movement to (or from) a source.

In species, such as carnivores occurring at low densities or having large home ranges, individuals are widely spaced and communicate via chemical broadcast signaling at *latrines*, i.e., collections of scent marks (feces, urine or glandular secretions), or via visually conspicuous landmarks of the boundary such as scratches and middens. Herrings communicate by farting. Shelter-dwelling caterpillars ballistically eject faecal pellets great distances (7–39 times their body length) at great speeds, in order to remove olfactory chemical cues for natural enemies.

The **communication distance** is the maximal distance at which the receiver can still get the signal. For example, a typical bird can detect a sound, coming from 300 m away amidst a background racket of human speech (roughly, 60 dB).

Animals can vary the signal amplitude and visual display with *receiver distance* in order to ensure signal transmission. For example, baleen whales have been observed calling more loudly, in order to compensate for human-generated noise.

Another example of *distance-dependent communication* is the protective coloration of some aposematic animals: it switches from *conspicuousness* (signaling nonedibility) to *crypsis* (camouflage) with increasing distance from a predator. Examples of interspecies communication of nonhuman animals, other than predator-prey signaling, are: eavesdropping, heterospecific alarm calls and cooperative hunting.

The main modes of animal communication are infrasound (< 20 Hz), sound, ultrasound (> 20 kHz), vision (light), chemical (odor), tactile, seismic and electrical. Infrasound, low-pitched sound (as territorial calls) and light in air can be long-distance. Some frogs, spiders, insects, small mammals have vibrotactile sense.

In a diverse ecosystem's soundscape, mammals tend to monopolize the low frequencies; birds call at higher pitches; and sound of insects is higher still.

A blue whale infrasound could (prior to noise pollution caused by ships) travel over 6000 km through the ocean water using the **SOFAR channel** (Chap. 25).

Most elephant communication is in the form of infrasonic rumbles which may be heard by conspecifics 5–10 km away. Also, they drum their soles on the ground, and resulting seismic waves can be detected as far as 16–32 km.

Many animals hear infrasound generated by earthquakes, tsunami and hurricanes before they strike. Elephants can hear storms 160–240 km away.

High-frequency sounds attenuate more rapidly with distance, more directional and vulnerable to scattering. But ultrasounds are used by bats (echo-location) and arthropods. Rodents use them to communicate to nearby receivers without alerting predators and competitors. Some anurans shift to ultrasound signals in the presence of continuous background noise (such as waterfall, human traffic). Animals, including frogs, insects, birds and whales, increase the minimum frequency, amplitude or **signal-to-noise ratio** (Chap. 21) in the presence of antropogenic noise.

- **Plant long-distance communication**

  Long-distance signaling was observed from roots and mature leaves, exposed to an environmental stress, to newly developing leaves of a higher plant.

  This communication is done cell-to-cell through the plant vascular transpiration system. In this system, macromolecules (except for water, ions and hormones) carry nutrients and signals, via *phloem* and *xylem* tissues, only in one direction: from lower mature regions to shoots. The identity of long-distance signals in plants is still unknown but the existence of information macromolecules is expected.

  Besides the above vascular signaling, plants communicate chemically with each other or with mutualistic animals (pollinators, bodyguards, etc.). For example, plants respond to attack by herbivores or pathogens with the release of volatile compounds, informing neighboring plants and attracting predators of attackers.

  Some 80 % of plants are colonized by ectosymbiotic fungi that form a network of fine white threads, *mycorrhizae*, which take in water and minerals from the soil, and hand some over to the plant in exchange for nutrients. A mycorrhizal network can take over an entire forest and tie together plants of different species. Plants use this network as a signaling and kin (or host) detection system too. They assist neighbors or kin in deterring pests, attracting pollinators and nutrient uptake.

- **Internodal distance**

  A *node* on a plant stem is a joint where a leaf is attached. The **internodal distance** (or *internode length*) is the distance between two consecutive nodes.

  A *ramet* is an independent member of a clone. The **interramet distance** (or *propagule dispersal distance*) is the internodal distance in plant clonal species.

- **Leaf size**

  According to Boland et al., 2006, the leaf is *mesophyll, notophyll, microphyll, nanophyll*, if its length is > 12.7, 7.6–12.7, 2.5–7.6, < 2.5 cm, respectively.

- **Body size rules**

  Body size, measured as mass or length, is one of the most important traits of an organism. Food webs, describing "who eat whom" (cf. **trophic distance**), are *nearly interval*, i.e., the species can be ordered so that almost all the resources of each consumer are adjacent in the order. Zook et al., 2011, found that ordering by body size is the best proxy to produce this near-interval ordering.

  The lower limits (10 kg and 2 g) to body size is set by the size of offspring for marine and by energetic limitations for terrestrial mammals. The largest known sizes for them are 190 t and 16 t, but the upper limit is still unclear. The length ranges for mammals, fish, insects, vascular plants, algae, bacteria, virions are $[10^{-1}, 10^{2}]$, $[10^{-2}, 10^{1}]$, $[10^{-4}, 10^{-1}]$, $[10^{-2}, 10^{2}]$, $[10^{-5}, 10^{0}]$, $[10^{-7}, 10^{-3}]$, $[10^{-8}, 10^{-6}]$ m.

[redacted]

According to Payne et al., 2008, the maximum size of the Earth's organisms increased by 16 orders of magnitude over the last 3.5 billion years. 75 % of the increase happened in two great leaps (about 1900 and 600–400 Ma ago: the appearance of eukaryotic cells and multi-cellularity) due to leaps in the oxygen level, and each time it increased about million times. The average animal in the oceans today is 150 times larger in mass than the average one half a billion years ago.

Smith et al., 2010: the maximum size of mammals increased (from 2 g to 190 t) near-exponentially after the C-T (Cretaceous-Paleogene) extinction of the nonavian dinosaurs 65.5 Ma ago; with little food, only small, no larger than a cat, species survived on land. After this C-T event, homeothermic animals (mammals and birds) radiated globally from northern Asia and became dominant.

The maximum size of insects also followed O_2 level 350–150 Ma ago, reaching 71 cm. Then it dropped (while O_2 went up) with evolution of birds and 65 Ma ago with their specialization and evolution of bats. Larsson–Dececchi, 2013, explain the origin of birds by a change of body-to-limb length ratio in Maniraptoran dinosaurs: the hind legs shrank, while forelimbs got long enough to work as an airfoil. From 230–220 to 163 Ma ago, theropods shrinked ($\approx$ 0.5 % of mass) to first birds.

Evans et al., 2012, claim that an increase in size (100, 1000, 5000 times) of land and marine mammals took 1.6, 5.1, 10 and 1, 1.3, 5 million generations, respectively. Mouse-sized mammals evolved into elephant-sized ones during 24 million generations, but decreasing in size occurred about 30 times faster.

Clauset and Erwin, 2008: 60 Ma of mammalian body size evolution can be explained by simple diffusion model of a trade-off between the short-term selective advantages (**Cope's rule**, common among mammals: a slight within-lineage drift toward larger masses) and long-term selective risks of increased size.

The size has costs as well as benefits; for example, reversals to unicellularity occurred at least 5 times in cyanobacteria. It favors the individual but renders the clade more susceptible to extinction via, for example, dietary specialization.

Large size enhances reproductive success, the ability to avoid predators and capture prey, and improves thermal efficiency. In large carnivores, bigger species dominate better over smaller competitors. Predator-prey mass ratio is typically around 10. But, for example, cookiecutter shark, only 0.5–1 m in length, preys on all larger animals in ocean, and the larvae of beetle *Epomis* preys on amphibians.

By mean body size (67 kg now and 50 kg in the Stone Age) humans are a small *megafauna* ($\geq$ 44 kg) species. A rapid average decline of $\approx$ 20 % in size-related traits was observed in human-harvested species.

Benthic (living on/in the bottom of a water body) animals are generally classified according to size: *microbenthos* $<$ 0.063 mm, *meiobenthos* 0.063–1 (or 0.5) mm, *macrobenthos* $>$ 1 (or 0.5) mm and, sometimes, *megabenthos* $>$ 10 mm.

Given below are the other main rules of large-scale Ecology involving body size.

Foster's (or *island*) **rule** is a principle that size of members of a species changes depending on the resources available in the environment. Damuth, 1993: there is an optimum mammal body size $\approx$ 1 kg for energy acquisition, and so island species should, in the absence of the usual competitors and predators, evolve to it.

Insular dwarfism is an evolutionary trend of the reduction in size of large mammals when their gene pool is limited to a very small environment (say, islands). One explanation is that smaller animals need fewer resources and reproduce faster.

Island gigantism is a form of natural selection where the size of animals isolated on an island increases dramatically over generations due the removal of constraints. Also, island bird species evolve smaller flight muscles and longer legs.

Abyssal gigantism is a tendency of deep-sea species to be larger than their shallow-water counterparts. For example, the colossal squid and the king-of-herrings (giant oarfish) can reach 14 and 17 m in length. It can be adaptation for scarcer food resources, greater pressure and lower temperature.

The **lilliput effect** is a reduction in the biota's size after mass extinctions.

1 °C of warming reduces the adult body mass of cold-blooded organisms by 2.5 % on average. For warm-blooded animals, **Allen's rule** holds: those from colder climates have shorter limbs than the equivalent ones from warmer climates.

Rensch's rule is that males are the larger sex in big-bodied species (such as humans) and the smaller sex in small-bodied species (such as spiders). It holds for plants also. Often, natural selection on females to maximize fecundity results in female-biased sexual size dimorphism, whereas sexual selection for large males promotes male-biased dimorphism. The males in some cichlid fish are up to 60 times larger than that of the females, while tremoctopus females may reach 2 m versus the males, at most a few cm long.

Size-assortative mating (positive correlation between male and female size among couples) has been found in crustaceans, insects, birds, reptiles, fishes and humans, for which it is a part of *homophily* (tendency to associate and bond with similar others). Humans have the largest, among apes, penises and breasts.

Cognitive and behavioral capacities do not correlate either with body or brain size, nor with their ratio, which is, say, $\frac{1}{7}, \frac{1}{40}, \frac{1}{2496}$ for small (0.06 mg) ant, human and shark. The *encephalization quotient* is the ratio of actual to predicted brain mass for a given size animal; it is the record 7.4–7.8 for humans. The number of neurons is $302, 85 \times 10^9, 2 \times 10^{11}$ in a nematode, human and elephant. Fish with smaller brain have more offspring. Echinoderms (say, starfish) lack a brain entirely.

Bromage et al., 2012, found a correlation between body mass and *RI* (*repeat interval*), i.e., the number of days between adjacent *striae of Retzius* (incremental growth lines seen in tooth) in primate's enamel. RI is also represented by the increments in bone. RI is an integer within $[1, 11]$; the mean RI is 8–9 in humans.

[redacted]

The *surface-to-volume ratio* $\frac{A}{V}$ is the main compactness measure for 3D shapes in Biology. Higher $\frac{A}{V}$ means increased exposure to the environment. It permits smaller cells to gather nutrients and reproduce very rapidly. Also, smaller animals in hot and dry climates lose heat better through the skin and cool the body. But lower $\frac{A}{V}$ (and so, larger size) improves temperature control: slower heat loss or gain. **Bergmann's rule** is a principle that, within a species, the body size increases with colder climate. For example, Northern Europeans on average are taller than Southern ones.

The *Hutchinson's ratio* is the ratio of the size differences between competing species when they were living together as compared to when they were isolated.

- **Allometry**

Allometry concerns relative change in structure or function of organisms (within individuals, among individuals, and across groups/species) with change in body size (length l, or mass M). The dependence of a biological variable Y on body size is usually given by an empirical *allometric scaling law*, i.e., a power function $Y = Y_0 M^b$ or $Y = Y_0 l^b$, where b is the *scaling exponent* and Y_0 a constant.

Geometrically similar (i.e., of the same basic shape) objects have proportional relationships between surface area A, volume V, and length l, i.e., $A' = A(\frac{l'}{l})^2$ and $V' = V(\frac{l'}{l})^3$, in terms of the *square–cube law*. Also, $M' = M(\frac{l'}{l})^3$ holds for the mass M if a scaled-up object maintains the same density.

Isometric scaling occurs when the above proportions are preserved as size changes. Frogs grow isometrically, as well as Ediacaran biota.

An isometrically scaling organism will have all V- (or M-), A-, and l-based properties change to the power $1, \frac{2}{3}, \frac{1}{3}$ of M, respectively.

Allometric scaling is any change that deviates from isometry. Such scaling is *positive* or *negative* if the value is larger or smaller than predicted by isometry. Positive scaling occurs for some fishes, whose weight grows as about $l^{3.325}$.

An example of negative allometric scaling is *Kleiber's law*: metabolic rate (V-based property) is proportional to $M^{0.75}$ over 27 orders of magnitude, from the molecular level up to the largest organisms. According to Garland, 1983, the speed of mammals tend to increase as $M^{0.17}$, and 119 kg is the optimal weight for the fastest runner.

A unicellular organism of size l has metabolic activity roughly proportional to cell volume (so, to l^3) and flux of nutrient and energy dissipation proportional to cell envelope area (so, to l^2). Hence, l is close to their ratio. The size of viral particles is roughly proportional to the third root of the genome size.

Let d, D denote mean size and distance between organisms. A proxy $\frac{D}{d}$ of animal abundance is very similar over a size span of 7 orders of magnitude. In fact, $\frac{D}{d} = \frac{44\mu}{3\mu}, \frac{390\mu}{40\mu}, \frac{5.8mm}{0.5mm}, \frac{35mm}{5mm}, \frac{50cm}{5cm}, \frac{300cm}{10cm}, \frac{100m}{1.5m}$, respectively, for bacteria, amoeba, nematodes, antropodes, shrews, birds and large mammals.

- **Size spectrum**

The term **size spectrum** is used generally when comparing objects of a given class, say, shoes or phones. But mainly, it is (Sheldon–Parsons, 1967) the relationship between body size of individuals and their abundance or biomass, regardless of their species, in a given (aquatic or soil) **size-based food web**.

For a population, the main considered sizes (lengths or masses) are: maximal, *asymptotic* (which individuals would reach if they were to grow indefnitely), of *maximal yield* (with highest biomass) and average in maturity. Example of corresponding *size-spectrum models*: Andersen and Beyer, 2006, derived proportionality of the number of individuals of given species and size to their asymptotic size raised to the power -2.05.

- **Trophic distance**

  Given an ecosystem, its *ecological network* is a digraph in which species are (biomass- or abundance-weighted) vertices with two of them being connected by arc or edge if there is a trophic or, respectively, symbiotic interaction. A *community food web* (or *ecological pyramid*) is a such digraph with only trophic arcs.

  The **trophic distance** from resource u to consumer v is the length of a shortest *food chain* (directed $(u - v)$ path) if it exists,

  The *trophic level* of a vertex v is 1 if it is a primary resource (usually, *producer* as plants, algae, phytoplankton) and 1 plus the trophic level of its principal diet, otherwise. The *fractional tropic level* of v is (Pauly–Palomares, 2005) 1 plus the weighted average (using stomach contents) trophic level of all its food items.

  The *mean trophic level* for fishery overall catch should be preserved to avoid *fishing down the food web*, when fisheries in a given ecosystem deplete the large predatory fish and end up with small fish and invertebrates.

  In a **size-based food web**, the layers are defined by body-size class rather than by trophic level. Community-based *predator-prey body mass ratios* (PPMR) and *energy transfer efficiency* (TE) are key parameters in such webs. In marine food webs, typically, PPMR $\in [100, 3000]$ and TE $\in [0.1, 0.13]$, i.e., 10–13 % of prey biomass is converted into predator production.

  An *energy* and *functional food webs* are weighted digraphs where arcs correspond to energy flow and interaction strength. Consumers at each level convert to tissue about 10 % of their food's chemical energy.

- **Insecticide distance effect**

  The main means of pest (termites, ants, etc.) control are chemical liquid insecticides and repellents. The efficiency of an insecticide can be measured by its **all dead distance**, i.e., the maximum distance from the standard toxicant source within which no targeted insects are found alive after a fixed period.

  The **insecticide distance effect** is that the toxicant is spread through the colony because insects groom and feed each other. The toxicant should act slowly in order to maximize this effect and minimize secondary repellency created by the presence of dying, dead and decaying insects. Nearly all animals, when they die, emit the same stench of fatty acids which acts as repellent and it is universal.

[redacted]

23.4 Other Biological Distances

Here we collect the main examples of other notions of distance and distance-related models used in Biology.

- **Immunologic distance**

  An *antigen* (or immunogen, pathogen) is any molecule eliciting an immune response. Once it gets into the body, the immune system either neutralizes its pathogenic effect or destroys the infected cells. The most important cells in this response are white blood cells: *T-cells* and *B-cells* responsible for the production and secretion of *antibodies* (specific proteins that bind to the antigen).

  When an antibody strongly matches an antigen, the corresponding B-cell is stimulated to divide, produce clones of itself that then produce more antibodies, and then differentiate into a plasma or memory cell. A secreted antibody binds to an antigen, and antigen-antibody complexes are removed.

  A mammal (usually a rabbit) when injected with an antigen will produce immunoglobulins (antibodies) specific for this antigen. Then *antiserum* (blood serum containing antibodies) is purified from the mammal's serum. The produced antiserum is used to pass on passive immunity to many diseases.

  Immunological distance procedures (*immunodiffusion* and, the mainly used now, *micro-complement fixation*) measure the relative strengths of the immunological responses to antigens from different taxa. This strength is dependent upon the similarity of the proteins, and the dissimilarity of the proteins is related to the evolutionary distance between the taxa concerned.

  The *index of dissimilarity* $id(x, y)$ between two taxa x and y is the factor $\frac{r(x,x)}{r(x,y)}$ by which the *heterologous* (reacting with an antibody not induced by it) antigen concentration must be raised to produce a reaction as strong as that to the *homologous* (reacting with its specific antibody) antigen.

  The **immunological distance** between two taxa is given by

  $$100(\log id(x, y) + \log id(y, x)).$$

  It can be 0 for two closely related species. It is not symmetric in general.

  Earlier immunodiffusion procedures compared the amount of precipitate when heterologous bloods were added in similar amounts as homologous ones, or compared with the highest dilution giving a positive reaction.

  The name of the applied antigen (target protein) can be used to specify immunological distance, say, albumin, transferring lysozyme distances. Proponents of the *molecular clock hypothesis* estimate that one unit of albumin distance between two taxa corresponds to $\approx$ 0.54 Ma of their divergence time, and that one unit of **Nei standard genetic distance** corresponds to 18–20 Ma.

  Adams and Boots, 2006, call the *immunological distance* between two immunologically similar pathogen strains (actually, serotypes of dengue virus) their *cross-immunity*, i.e., 1 minus the probability that primary infection with one strain prevents secondary infection with the other. Lee and Chen, 2004, define the *antigenic distance* between two influenza viruses to be the reciprocal of their *antigenic relatedness* which is (presented as a percentage) the geometric mean $\sqrt{\frac{r(x,y)}{r(x,x)}\frac{r(y,x)}{r(y,y)}}$ of two ratios between the heterologous and homologous antibody *titers*.

An antiserum *titer* is a measurement of concentration of antibodies found in a serum. Titers are expressed in their highest positive dilution.

- **Metabolic distance**

  Enzymes are proteins that *catalyze* (increase the rates of) chemical reactions.

  The **metabolic distance** (or *pathway distance*) between enzymes is the minimum number of metabolic steps separating two enzymes in the metabolic pathways.

- **Pharmacological distance**

  The *protein kinases* are enzymes which transmit signals and control cells using transfer of *phosphate groups* from high-energy donor molecules to specific target proteins. So, many drug molecules (against cancer, inflammation, etc.) are kinase inhibitors (blockers). Designed drugs should be *specific* (say, not to bind to $\geq$ 95 % of other proteins), in order to avoid toxic side-effects.

  Given a set $\{a_1, \dots, a_n\}$ of drugs in use, the *affinity vector* of kinase x is defined as $(-\ln B_1(x), \dots, -\ln B_n(x))$, where $B_i(x)$ is the *binding constant* for the reaction of x with drug a_i, and $B_i(x) = 1$ if no interaction was observed. The binding constants are the average of several experiments where the concentration of binding kinase is measured at equilibrium. The **pharmacological distance** (Fabian et al., 2005) between kinases x and y is the Euclidean distance $(\sum_{i=1}^{n}(\ln B_i(x) - \ln B_i(y))^2)^{\frac{1}{2}}$ between their affinity vectors.

  The *secondary structure* of a protein is given by the hydrogen bonds between its residues. A *dehydron* in a solvable protein is a hydrogen bond which is solvent-accessible. The *dehydron matrix* of kinase x with residue-set $\{R_1, \dots, R_m\}$ is the $m \times m$ matrix $((D_{ij}(x)))$, where $D_{ij}(x)$ is 1 if residues R_i and R_j are paired by a dehydron, and is 0, otherwise. The **packing distance** (Maddipati–Fernándes, 2006) between kinases x and y is the Hamming distance $\sum_{1 \leq i,j \leq m} |D_{ij}(x) - D_{ij}(y)|$ between their dehydron matrices; cf. **base pair distance** among **RNA structural distances**. The *environmental distance* (Chen, Zhang and Fernándes, 2007) between kinases is a normalized variation of their packing distance.

  Besides hydrogen bonding, residues in protein helices adopt backbone dihedral angles. So, the secondary structure of a protein much depends on its sequence of dihedral angles defining the backbone. Wang and Zheng, 2007, presented a variation of **Lempel–Ziv distance** between two such sequences.

- **Global distance test**

  The *secondary structures* of proteins are mainly composed of the α-helices, β-sheets and loops. Protein *tertiary structure* refers to the 3D structure of a single protein molecule. The α and β structures are folded into a compact globule.

  The **global distance test** (GDT) is a measure of similarity between two (model and experimental) proteins x and y with identical *primary structures*

[redacted]

(amino acid sequences) but different tertiary structures. GDT is calculated as the largest set of amino acid residues' α carbon atoms in x falling within a defined **cutoff distance** (Chap. 29) d_0 of their position y.

For proteins, in order for this set to define all intermolecular stabilizing (relevant short range) interactions, $d_0 = 0.5$ nm is usually sufficient. Sometimes, $d_0 = 0.6$ nm, in order to include contacts acting through another atom.

- **Migration distance (in Biomotility)**

  The **migration** (or *penetration*) **distance**, in cattle reproduction and human infertility diagnosis, is the distance in mm traveled by the vanguard spermatozoon during sperm displacement *in vitro* through a capillary tube filled with homologous cervical mucus or a gel mimicking it. Sperm swim 1–4 mm/min. 90 % of human sperm swim forward with small side-to-side movements, while $\approx$ 5 % swim in a faster-paced helical pattern and the remaining $\approx$ 5 % swim in a hyper-helical manner, where the sperm are more active but less directional.

  Such measurements, under different specifications (duration, temperature, etc.) of incubation, estimate the ability of spermatozoa to colonize the oviduct *in vivo*.

  In general, the term **migration distance** is used in biological measurements of directional motility using controlled migration; for example, determining the molecular weight of an unknown protein via its migration distance through a gel.

- **Penetration distance**

  The **penetration distance** is a general term used in (especially, biological) measurements for the distance from the given surface to the point where the concentration of the penetrating substance (say, a drug) in the medium (say, a tissue) had dropped to the given level. Several examples follow.

  During penetration of a macromolecular drug into the tumor interstitium, *tumor interstitial penetration* is the distance that the drug carrier moved away from the source at a vascular surface; it is measured in 3D to the nearest vascular surface.

  During the intraperitoneal delivery of cisplatin and heat to tumor metastases in tissues adjacent to the peritoneal cavity, the *penetration distance* is the depth to which the drug diffuses directly from the cavity into tissues. Specifically, it is the distance beyond which such delivery is not preferable to intravenous delivery.

  It can be the distance from the cavity surface into the tissues within which drug concentration is, for example, (a) greater, at a given time point, than that in control cells distant from the cavity, or (b) is much higher than in equivalent intravenous delivery, or (c) has a first peak approaching its plateau value within 1 % deviation.

  The *penetration distance* of a drug in the brain is the distance from the probe surface to the point where the concentration is roughly half its **far field** value.

  The *penetration distance* of chemicals into wood is the distance between the point of application and the 5 mm cut section in which the contaminant concentration is at least 3 % of the total.

  The *forest edge-effect penetration distance* is the distance to the point where invertebrate abundance ceased to differ from forest interior abundance.

Cf. **penetration depth distance** in Chap. 9, **penetration depth** in Chap. 24 and **distance sampling** in Chap. 17.

- **Capillary diffusion distance**

  One of the diffusion processes is *osmosis*, i.e., the net movement of water through a permeable membrane to a region of lower solvent potential. In the respiratory system (the alveoli of mammalian lungs), oxygen O_2 diffuses into the blood and carbon dioxide CO_2 diffuses out.

  The **capillary diffusion distance** is, similarly to **penetration distance**, a general term used in biological measurements for the distance, from the capillary blood through the tissues to the mitochondria, to the point where the concentration of oxygen has dropped to the given low level.

  This distance is measured as the average distance from the capillary wall to the mitochondria, or the distance between the closest capillary endothelial cell to the epidermis, or in percentage terms, say, the distance where a given percentage (95 % for maximal, 50 % for average) of the fiber area is served by a capillary.

  Another practical example: the *effective diffusion distance* of nitric oxide NO in microcirculation *in vivo* is the distance within which N concentration is greater than the equilibrium dissociation constant of the target enzyme for oxide action.

  Cf. the **immunological distance** for immunodiffusion and, in Chap. 29, the **diffusion tensor distance** among **distances in Medicine**.

- **Főrster distance**

  FRET (fluorescence resonance energy transfer; Főrster, 1948) is a distance-dependent quantum mechanical property of a *fluorophore* (molecule component causing its fluorescence) resulting in direct nonradiative energy transfer between the electronic excited states of two dye molecules, the donor fluorophore and a suitable acceptor fluorophore, via a dipole. In FRET microscopy, fluorescent proteins are used as noninvasive probes in living cells since they fuse genetically to proteins of interest.

  The efficiency of FRET transfer depends on the square of the donor electric field magnitude, and this field decays as the inverse sixth power of the intermolecular separation (the physical donor-acceptor distance). The distance at which this energy transfer is 50 % efficient, i.e., 50 % of excited donors are deactivated by FRET, is called the **Főrster distance** of these two fluorophores.

  Measurable FRET occurs only if the donor-acceptor distance is less than ≈ 10 nm, the mutual orientation of the molecules is favorable, and the spectral overlap of the donor emission with acceptor absorption is sufficient.

- **Gendron–Lemieux–Major distance**

  The **Gendron–Lemieux–Major distance** (2001) between two base-base interactions, represented by 4×4 *homogeneous transformation matrices* X, Y, is

$$\frac{S(XY^{-1}) + S(X^{-1}Y)}{2},$$

  where $S(M) = \sqrt{l^2 + (\theta/\alpha)^2}$, l is the translation length, θ is the rotation angle, and α is a scaling factor between the translation and rotation contributions.

[redacted]

- **Spike train distances**

  A human brain has 85×10^9 *neurons* (nerve cells) each communicating with an average 1000 other neurons dozens of times per second. Most neurons are capable of making 10^4–10^6 individual microconnections. One human brain, using $\approx 10^{15}$ synapses, produces $\approx 6.4 \times 10^{18}$ nerve impulses per second.

  The neuronal response to a stimulus is a continuous time series. It can be reduced, by a threshold criterion, to a simpler discrete series of *spikes* (short electrical pulses). A *spike train* is a sequence $x = (t_1, \dots, t_s)$ of s events (neuronal spikes, or heart beats, etc.) listing absolute spike times or interspike time intervals. The main **distances between spike trains** $x = x_1, \dots, x_m$ and $y = y_1, \dots, y_n$ follow.

1. The **spike count distance** is defined by
$$\frac{|n-m|}{\max\{m,n\}}.$$
2. The **firing rate distance** is defined by
$$\sum_{1 \le i \le s} (x'_i - y'_i)^2,$$
where $x' = x'_1, \dots, x'_s$ is the sequence of local firing rates of train $x = x_1, \dots, x_m$ partitioned in s time intervals of length T_{rate}.
3. Let $\tau_{ij} = \frac{1}{2}\min\{x_{i+1} - x_i, x_i - x_{i-1}, y_{i+1} - y_i, y_i - y_{i-1}\}$ and $c(x|y) = \sum_{i=1}^{m}\sum_{j=1}^{n} J_{ij}$, where $J_{ij} = 1$ if $0 < x_i - y_i \le \tau_{ij}$, $= \frac{1}{2}$ if $x_i = y_i$ and $= 0$, otherwise. The **event synchronization distance** (Quiroga et al., 2002) is defined by
$$1 - \frac{c(x|y) + c(y|x)}{\sqrt{mn}}.$$
4. Let $x_{isi}(t) = \min\{x_i : x_i > t\} - \max\{x_i : x_i < t\}$ for $x_1 < t < x_m$, let $I(t) = \frac{x_{isi}(t)}{y_{isi}(t)} - 1$ if $x_{isi}(t) \le x_{isi}(t)$ and $I(t) = 1 - \frac{y_{isi}(t)}{x_{isi}(t)}$, otherwise. The time-weighted and spike-weighted **ISI distances** (Kreuz et al., 2007) are
$$\int_0^T |I(t)|dt \text{ and } \sum_{i=1}^{m} |I(x_i)|.$$
5. Various *information distances* were applied to spike trains: the **Kullback–Leibler distance**, and the **Chernoff distance** (Chap. 14). Also, if x and y are mapped into binary sequences, the **Lempel–Ziv distance** and a version of the **normalized information distance** (Chap. 11) are used.
6. The **Victor–Purpura distance** (1996) is a cost-based **editing metric** (i.e., the minimal cost of transforming x into y) defined by the following operations

with their associated costs: insert a spike (cost 1), delete a spike (cost 1), shift a spike by time t (cost qt); here $q > 0$ is a parameter. The **fuzzy Hamming distance** (Chap. 11), introduced in 2001, identifies cost functions of shift preserving the triangle inequality.

7. The **van Rossum distance**, 2001, is defined by

$$\sqrt{\int_0^{\infty} (f_t(x) - f_t(y))^2 dt},$$

where x is convoluted with $h(t) = \frac{1}{\tau}e^{-t/\tau}$ and $\tau \approx 12$ ms (best); $f_t(x) = \sum_0^m h(t - x_i)$. This and above distances are the most commonly used metrics.
8. Given two sets of spike trains labeled by neurons firing them, the **Aronov et al. distance** (2003) between them is a cost-based **editing metric** (i.e., the minimal cost of transforming one into the other) defined by the following operations: insert or delete a spike (cost 1), shift a spike by time t (cost qt), relabel a spike (cost k), where $q, k > 0$ are parameters.

- **Bursting distances**

  Bursts refers to the periods in a spike train when the spike frequency is relatively high, separated by periods when it is relatively low or spikes are absent.

  Given neurons $x_1, \dots, x_n$ and SBEs (synchronized bursting events) $Y_1, \dots, Y_m$ with similar patterns of neuronal activity, let C^{ij} denote the cross-correlation between the activity of a neuron in Y_i and Y_j maximized over neurons, and let C_{ij} denote the correlation between neurons x_i and x_j averaged over SBEs.

  Baruchi and Ben-Jacob, 2004, defined the **interSBE distance** between Y_i and Y_j and the **interneuron distance** between x_i and x_j by $\frac{1}{m}(\sum_{s=1}^{m}(C^{is} - C^{js})^2)^{\frac{1}{2}}$ and $\frac{1}{n}(\sum_{s=1}^{n}(C_{is} - C_{js})^2)^{\frac{1}{2}}$, respectively.
- **Long-distance neural connection**

  Unlike Computing, neural systems are not exclusively optimized for minimal global wiring, but for a variety of factors including the minimization of processing steps. Kaiser and Hilgetag, 2006, showed that, due to the existence of long-distance projections, the total wiring among 95 primate (Macaque) cortical areas could be decreased by 32 %, and the wiring of neuronal networks in the nematode *C. elegans* could be reduced by 48 % on the global level. For example, $>$ 10 % of the primate cortical projections connect components separated by $>$ 40 mm, while 69 mm is the maximal possible distance. For the global *C. elegans* network, some connections are almost as long as the entire organism.

  Long-distance and long-lasting neural connections may be the architecture linking the separate regions/processes together during a single global conscious state.

  In autism there are more local connections and more local processing, while the psychosis/schizophrenia spectrum is marked by more long-distance connections.

[redacted]

About 5 %, 10 %, 6.7 % of variation in individual intelligence is predicted by activity level in LPFC (lateral prefrontal cortex), by the strength of neural pathways connecting left LPFC to the rest of the brain and by overall brain size.

- **Long-distance cell communication**

  Human cell size is within [4–135] μm; typically, 10 μm. In *gap junctions*, the intercellular spacing is reduced from 25–450 nm to a gap of 1–3 nm, bridged by hollow tubes. Animal cells may communicate locally, either directly through gap junctions, or by cell-cell recognition (in immune cells), or (*paracrine signaling*) using messenger molecules that travel, by diffusion, only short distances. Mammal's, astrocytes form, via gap junctions, a network of neurons and vasculature. Neurons may use interferon signals transmitted over great distances to fend off viral infection.

  In *synaptic signaling*, the electrical signal along a neuron's axon triggers the release of a neurotransmitter to diffuse across the synapse through a gap junction. Signal transmission through the nervous system is a long-distance signaling. Slower long-distance signaling is done by hormones transported in the blood. A hormone reaches all parts of the body, but only target cells have receptors for it.

  Another means of long-distance cell communication, via TNTs (*tunneling nanotubes*), was found in 1999. TNTs are membrane tubes, 50–200 nm thick with length up to several cell diameters. Cells can send out several TNTs, creating a network lasting hours. TNTs can carry cellular components and pathogens (HIV and prions). Also, electrical signals can spread bidirectionally between TNT-connected cells (over distances 10–70 μm) through interposed gap junctions.

  Some bacteria gain energy by oxidizing H_2S via electron transfer, hundreds of cell-lengths away. Thousands of *Desulfobulbus* form cm-long conductive chains, transporting electrons from H_2S-rich marine sediment to the upper O_2-rich one.

- **Length constant**

  In an excitable cell (nerve or muscle), the **length constant** is the distance over which a nonpropagating, passively conducted electrical signal decays to $\frac{1}{e}$ (36.8 %) of its maximum.

  During a measurement, the **conduction distance** between two positions on a cell is the distance between the first recording electrode for each position.

- **Ontogenetic depth**

  The **ontogenetic depth** (or *egg-adult distance*) is (Nelson, 2003) the number of cell divisions, from the unicellular state (fertilized egg) to the adult metazoan capable of reproduction (production of viable gametes).

  The **mitotic length** is the number of intervening mitoses, from the normal (neither immortal nor malignant) cells in the immature precursor stage to their progeny in a state of *mitotic death* (terminal differentiation) and phenotypic maturity.

  Adult humans produce daily $\approx$ 200 billion red cells, 100 billion neutrophils and 50 billion other cells. About 7×10^{15} mature cells are produced in a human lifetime and these could be produced in 53 cell generations ($2^{53} \approx 9 \times 10^{15}$).

- **Interspot distance**.

  A *DNA microarray* is a technology consisting of an arrayed series of thousands of *features* (microscopic spots of DNA oligonucleotides, each containing picomoles of a specific DNA sequence) that are used as probes to hybridize a *target* (cRNA sample) under high-stringency conditions. Probe-target hybridization is quantified by fluorescence-based detection of fluorophore-labeled targets to determine the relative abundance of nucleic acid sequences in the target.

  The **interspot distance** is the **spacing distance** (Chap. 29) between features. Typical values are 375, 750, 1500 micrometers ($1\ \mu m = 10^{-6}$ m).

- **Read length**

  In gene sequencing, automated sequencers transform electropherograms (obtained by *electrophoresis* using fluorescent dyes) into a four-color chromatogram where peaks represent each of the DNA bases A, T, C, G. Chromosomes stained by some dyes show a 2D pattern of traverse bands of light and heavy staining.

  The **read length** is the length, in the number of bases, of the sequence obtained from an individual clone chosen. Computers then assemble those short blocks into long continuous stretches which are analyzed for errors, gene-coding regions, etc.

- **Action at a distance along DNA/RNA**

  An **action at a distance along DNA/RNA** happens when an event at one location on a molecule affects an event at a distant (say, more than 2500 base pairs) location on the same molecule.

  Many genes are regulated by distant (up to a million bp away and, possibly, located on another chromosome) or short (30–200 bp) regions of DNA, *enhancers*. Enhancers increase the probability of such a gene to be transcribed in a manner independent of distance and position (the same or opposite strand of DNA) relative to the transcription initiation site (the promoter).

  DNA supercoiling is the twisting of a DNA double helix around its axis, once every 10.4 bp of sequence (forming circles and figures of eight) because it has been bent, overwound or underwound. Such folding puts a long range enhancer, which is far from a regulated gene in **genome distance**, geometrically closer to the promoter.

  The *genomic radius of regulatory activity* of a genome is the genome distance of the most distant known enhancer from the corresponding promoter; in the human genome it is $\approx 10^6$ bp (for the enhancer of SSH, *Sonic Hedgehog* gene).

  There is evidence that genomes are organized into enhancer-promoter loops. But the long range enhancer function is not fully understood yet.

  Similarly, some viral RNA elements interact across thousands of intervening nucleotides to control translation, genomic RNA synthesis and mRNA transcription.

  Genes are controlled either locally (from the same molecule) by specialized *cis* regulators, or at a distance by *trans* regulators. Comparing genes in key brain regions of human and primates, the most drastic changes were found in *trans*-controlled genes.

[redacted]

- **Length variation in 5-HTTLPR**

  5-HTTLPR is a repeat polymorphic region in *SLC6A4*, the gene (on chromosome 17) coding for SERT (serotonin transporter) protein. This polymorphism has *short* (14 repeats) and *long* (16 repeats) variations. So, an individual can have short/short, short/long, or long/long genotypes at this location in the DNA.

  A short/short allele leads to less transcription for SLC6A4, and its carriers are more attuned and responsive to their environment; so, social support is more important for their well-being. They have less gray matter, more neurons and a larger thalamus. Whereas $\frac{2}{3}$ of East Asians have the short/short variant, only $\frac{1}{5}$ of Americans and Western Europeans have it.

  Other gene variants of central neurotransmitter systems—dopamine receptor (*DRD4 7R*), dopamine/serotonin breaking enzyme (*MAOA VNTR*) and μ-opioid receptor (*OPRM1 A118G*)—are also associated with novelty-seeking, plasticity and social sensitivity. They appeared < 0.08 Ma ago and spread into 20–50 % of the population. They generate anxiety and aggression, but could be selected for extending behavioral range and boosting resilience at the group level.
- **ER-mitochondria distance**

  The *endoplasmic reticulum* (ER) is a large membrane-bound compartment spread throughout the cytoplasm of eukaryotic cells. The interface between mitochondria and ER is a crucial hub for bidirectional communication regulating mitochondrial energy, lipid biosynthesis, calcium Ca2+ signalling, and cell death.

  Mitochondria are surrounded by tubules of the ER that lie preferentially within 200 nm apart, whereas 20 % of the mitochondrial surface is in direct contact with the ER. Mitochondria and ER are physically coupled by protein tethers. The size of these bridges between the two organelles are 10–25 nm in length. Altering this physical **ER-mitochondria distance** leads inevitably to cell dysfunction.

  When this distance is decreased, Ca(2+) overload occurs both in the cytosol and mitochondria. Qi–Shuai, 2016, proposed that the resultant death of neurons, which are responsible for memory and cognition, contributes to Alzheimer's disease.
- **Telomere length**

  The *telomeres* are the caps of repetitive DNA sequences ($(TTAGGG)_n$ in vertebrates cells) at both ends of each linear chromosome in the cell nucleus. They are long stretches of noncoding DNA protecting coding DNA. The number n of *TTAGGG* repeats is called the **telomere length** (TL); it is $\approx$2000 in humans. TL is a robust indicator of biological age and a prognostic marker of disease risk. A limit of life—about 120 years—can be defined by TL in blood stem cells.

  Every time a normal cell divides, its telomeres shorten (due to incomplete lagging strand DNA synthesis) and eventually they are so short that cell stops dividing, self-destructs, or tries to self-replicate and creates cancer. The *Hayflick limit* is the maximal number of divisions beneath which a normal cell will stop dividing, because of shortened telomeres or DNA damage, and die; for humans it is about 52.

  Human telomeres are 15–20 kb in length at birth and shorten gradually throughout life in dividing cells: they lose $\approx$ 100 bp, i.e., 16 repeats, at each mitosis (i.e., every 20–180 min). But telomere length can increase: by transfer of repeats between telomers or by synthesizing new telomeric DNA from either an RNA template using the enzyme *telomerase*, or a DNA template using *alternative lengthening of telomeres* (ALT). In humans, telomerase acts only in germ, stem or actively dividing tumor cells; ALT occurs only in cancer cells.

Hydras, lobsters, planarian flatworms, trees maintain telomere lengths. Also, bacterial colonies and *Turritopsis dohrnii*, whose medusa form can revert to the polyp stage, are *biologically immortal*, i.e., there is no aging (sustained increase of mortality rate with age) since the Hayflick limit does not apply. Animals with negligible aging die mainly because of growth: they lose agility to get food. The oldest living animals are some sponges and black corals: 2000–10,000 years. The oldest known cell line is 11,000 years-old canine transmissible venereal tumor.

Phenoptsis is genetically programmed death of organism. It acts quickly in *semelparous* (capable of only single reproduction) species, say, Pacific salmon, cicada, mayflies, annual plants and some bamboo, arachnids, squids. Extreme examples: the male praying mantis ejaculates only after being decapitated by the female, and the *Adactyllidium* tick larvae kill their mother eating her from the inside out.

Aging is slow phenoptsis in other species. The telomere shortening is one of its main mechanisms. Vascular disease, osteoarthritis, cancer and menopause are other means of human phenoptsis. Human mortality doubles every 8 years. Mortality rate of people with cancer behave as if the cancer had aged them by 15 years.

- **Gerontologic distance**

The **gerontologic distance** between individuals of ages x and y from a population with survival fraction distributions $S_1(t)$ and $S_2(t)$, respectively, is defined by

$$|\ln \frac{S_2(y)}{S_1(x)}|.$$

A function $S(t)$ can be either an empirical distribution, or a parametric one based on modeling. The main survival functions $S(t)$ are: $\frac{N(t)}{N(0)}$ (where $N(t)$ is the number of survivors, from an initial population $N(0)$, at time t), e^{kt} (exponential model), $e^{\frac{a}{b}(1-e^{bt})}$ (Gompertz model), and $e^{-\frac{at^{b+1}}{b+1}}$ (Weibull model); here a and b are, respectively, age independent and age dependent mortality rate coefficients.

But *late-life mortality deceleration* was observed for humans and fruit flies: the probability that organism's somatic cells become senescent tends to be independent of its age in the long-time limit. The one-year probability of death at advanced age asymptotically approaches 44 % for women and 54 % for men. Such a plateau is typical for many Markov processes. Human *species-specific life span* (age at which death rates of different populations converge) is close to 95 years.

[redacted]

Since 1840, life expectancy has risen linearly: almost three months per year for women and more slowly for men. But the age of super-long livers is linked to their genes rather than their lifestyle; at least 100 genes are linked to longevity.

Distances are used in Human Gerontology also to model the link between geographical distance and contact between adult children and their elderly parents.

Aging/death are adaptive species-specific trade-offs with reproduction. But the *Akela effect* (long post-reproductory period with intergenerational transfers) was observed, besides humans, in two species of toothed whales, Asian elephants, social aphids and, in zoo, Japanese macaques and chimpanzees.

- **Distance to death**

  80 % of the persons, who die in any one year, are age 65 or older. Elderly persons think and talk readily about death, but perceived temporal nearness of it is not quantified by $\approx$ 50 % of them. Still this proximity determines one's attitude on it.

  Gerstorf *et al., 2008*: relative to age-related decline, mortality-related one (i.e., **distance to death**) in reported life satisfaction account for more variance in the change of subjective well-being. At a point about 4 years before death of an old, i.e., 70+ years, person, this decline showed a 2-fold increase (3-fold for the *oldest old*, i.e., 85+ years) in steepness relative to the preterminal phase.

  Bosworth *et al., 1999*: distance to death explains much of the variance in intellectual performance (verbal meaning, psychomotor speed, spatial and reasoning abilities) associated with age. Higher baseline intelligence test scores are associated with reduced risk of mortality and reduced effects of impending death on cognition.

  The *terminal drop hypothesis* (Riegel–Riegel, 1972) states that death is preceded by a decrease in cognitive (especially, verbal) functioning over an $\approx$ 5 years period.

  The *cascade model* (Birren–Cunningham, 1985) posits primary (normal), secondary (disease-related) and tertiary (distance to death) aging, which influence 3-rd, 2–3-rd and 1–3-rd, respectively, classes of intellectual function: *cristallilized abilities* (to think logically and solve problems knowledge-independently), *fluid abilities* (to use skills, knowledge and experience) and perceptual speed.

  Borjigin et al., 2013, observed neural correlates of heightened conscious processing at near-death: a surge 30 s of coherence and connectivity in the dying rat's brain.

  Micromort and *microlife* are the units of risk: 10^{-6} probability of death and half an hour ($\approx 10^{-6}$-th of 57 years) change of life expectancy, respectively.

  The farthest from Earth that any human has died is about 167 km, when three cosmonauts on Soyuz 11 suffered a depressurization accident while returning to Earth. They were moving at about 7755 m/sec at the time, which is also the highest forward speed at which any human has ever died.

- **Distance running model**

  Bipedality is a key behavior of hominins which appeared 6–4.2 Ma ago. It allowed australopithecines to see approaching danger further off, to walk long distances and to use hands for gathering food. Our genus Homo emerged $\approx$ 2.5 Ma ago.

  The distance running model anthropogenesis, proposed in [BrLi04], claims that our capacity to run long distances in the savanna arised, prior to the invention of the spear, as adaptation for persistence hunting (by running prey to exhaustion) and scavenging (allowing to compete for widely dispersed carcasses).

  This model specifies how endurance running defined the human body form, producing balanced head, low/wide shoulders, narrow chest, short forearms and heels, large hip, etc. Even now, a good athlete can run at 20 km/h for several hours which is comparable to endurance specialists as, say, zebras and antelopes. By sweating we can dissipate body heat faster than any other large mammal and reach large sustainable distance. The capacity of humans to travel vast distances using little energy contributed also to the evolution of their complex social networks.
- **Distance coercion model**

  The **distance coercion model** ([OkBi08]) of the origin of uniquely human kinship-independent cooperation see all complex symbolic speech, cognitive virtuosity, transmission of fitness-relevant information, etc. as elements and effects of this cooperation catalyzed by advances in lethal projectile weapons.

  The model argues that such cooperation can arise only as a result of the pursuit of individual self-interest by animals who can project "death from a distance".

  Among rare organisms able to project coercive threat remotely, humans are the most efficient on long distances, say, to kill adult conspecifics up to 18–27 m by throwing a spear and up to 91 m by a bow. The chimpanzee and Neanderthal also could throw objects but not with human's precision.

  The model posits that this capacity, permitting to repel predators and scavenge their kills in the African savanna, briefly preceded the emergence of brain expansion and social support. Comparing with Neanderthals, evidence of a huge number of injuries suggests that their hunting involved dangerously close contact with large prey animals; they used conventional spears rather than true projectile weapons.

  Throwing and language capacities enabled humans to survive rapid climatic and environmental changes, to spread and to become the dominant large-scale species on the planet. Historical increases in social cooperation could be associated with prior acquisition of a new coercive technology; for instance, the bow and agricultural civilizations, gunpowder weaponry and the modern state.

[redacted]

Humans are most efficient enforcers of cooperation (even relying mainly on indirect cues): our cognitive abilities expanded the range of situations in which cooperation can be favored. Also, while the *strong reciprocity* (generous third-party enforcement) is prevalent in large societies, Marlowe et al., 2012, claim that motivated by the basic emotion of anger, humans-special tendency to retaliate on their own behalf, even at a cost, is sufficient to explain the origin of human cooperation.

- **Distance model of altruism**

  In Evolutionary Ecology, altruism is explained by kin selection, reciprocity, sexual selection, etc. The cooperation between nonrelatives was a driving force in some major transitions (say, from symbiotic bacteria to mitochondria, eukaryotes or multicellular organisms). Individual selection, including *social selection* in which fitness is influenced by the behaviors of others, interacts with group selection.

  The **distance model of altruism** ([Koel00]) claims that altruists spread locally, i.e., with small interaction distance and offspring dispersal distance, while the egoists invest in increasing of those distances. The intermediate behaviors are not maintained, and evolution will lead to a stable bimodal spatial pattern.

- **Distance grooming model of language**

  In primates, being groomed produces mildly narcotic effects, because it stimulates the production of the body's natural opiates, the *endogenous opioid peptides*. Kulahci et al., 2015, observed in lemurs vocal grooming-at-a-distance.

  Language, according to Dunbar, 1993, evolved in archaic *Homo sapiens* as more distance/time efficient replacement of social grooming. Their brain size expanded 0.5 Ma ago from 900 cm^3 in *Homo erectus* to 1300 cm^3, and they lived in large groups (over 120 individuals) requiring cohesion. Language allowed them to produce the reinforcing, social-bonding effects of grooming-at-a-distance and to use more efficiently the time available for social interaction.

  Language achieves this through information transfer, gossip and emotional means (say, laughter, facial expression, Duchenne smile). Many primate species extensively use contact calls such as the long-distance *pant-hoot* call of chimpanzees. Dunbar interprets such calls as a grooming-at-a-distance from which language evolved. But gestures are far more likely precursor of language than vocalizations.

  He deduced (from the link between group and brain sizes in primates) that human social networks tend to be structured in layers: 5 intimates (support clique), 15 best friends (sympathy group), active network of "persons" (50 good friends and 150 friends), 500 acquaintances, 1500 "people I recognize". One need to be in contact every week, month, half-year, year with groups 1–4, respectively. A natural group size (*Dunbar's number*) is 150 for humans and 50 for chimpanzees.

Dunbar explain above sizes by cognitive and time constraints on the number of relationships ego can maintain at a given level of intensity. The clique size correlates with the highest achievable order of *intentionality recursion*, in which mind states are reflexively attributed to others. 0-order means responses to stimuli (as bacteria and computers); 1-st order: belief about the real or imagined world (as most organisms with brains); 2-nd order: belief about the mental state of others; i-th order: as, for $i = 5$, in the sentence "I *think* that you *believe* that I *suppose* that we *understand* that Jane *wants*...". We operate usually at 3-rd and sometimes at 4-th, 5-th, 6-th or 7-th order. Language is essential for 4-th order recursion.

[redacted]

Chapter 24
Distances in Physics and Chemistry

24.1 Distances in Physics

Physics studies the behavior and properties of matter in a wide variety of contexts, ranging from the submicroscopic particles from which all ordinary matter is made (*Particle Physics*) to the behavior of the material Universe as a whole (*Cosmology*).

Physical forces which act at a distance (i.e., a push or pull which acts without "physical contact") are nuclear and molecular attraction and, beyond the atomic level, gravity (completed, perhaps, by anti-gravity), static electricity, and magnetism. Last two forces can be both push and pull, depending on the charges of involved bodies. The nucleon-nucleon interaction (or *residual strong force*) is attractive but becomes repulsive at very small distances keeping the nucleons apart. Dark matter is attractive while dark energy is repulsive (if they exist).

Distances on a relatively small scale are treated in this chapter, while large distances (as in Astronomy and Cosmology) are the subject of Chaps. 25 and 26.

The distances having physical meaning range from 1.6×10^{-35} m (**Planck length**) to 8.8×10^{26} m (estimated size of the observable Universe). We can see things of about $10^{-4} - 10^{21}$ m and measure them within $[10^{-18}, 10^{26}]$ m. The smallest measurable distance, time and weight are 10^{-18} m (by LHC), 10^{-17} sec and 10^{-24} g.

The Theory of Relativity, Quantum Theory and Newtonian laws permit us to describe and predict the behavior of physical systems in the range $10^{-15} - 10^{12}$ m, i.e., from proton to Solar System. Weakened description is still possible up to 10^{25} m.

The world appears Euclidean at distances less than about 10^{25} m (if gravitational fields are not too strong). Relativity and Quantum Theory effects, governing Physics on very large and small scales, are already accounted for in technology, say, of GPS satellites and nanocrystals of solar cells.

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_24

- **Moment**

  In Physics and Engineering, **moment** is the product of a quantity (usually, force) and a distance (or a power of it) to some point associated with that quantity.
- **Momentum**

  In classical mechanics, **momentum** $\mathbf{p} = (p_x, p_y, p_z)$ is the product $m\mathbf{v}$ of the mass m and velocity vector $\mathbf{v} = (v_x, v_y, v_z)$ of an object.

  In relativistic 4D mechanics, *momentum-energy* $(\frac{E}{c}, p_x, p_y, p_z)$, where c is the speed of light and $E = mc^2$ is energy, is compared with space-time (ct, x, y, z).
- **Displacement**

  In Mechanics, a *displacement* (or *relative position*) *vector* of a moving particle from its initial position P_i to the final position P_f, is the vector $\overrightarrow{P_iP_f} = \overrightarrow{0(P_f - P_i)}$, where O is a reference point (usually the origin of a coordinate system).

  A **displacement** is the length $||P_f - P_i||_2$ of this vector, i.e., the Euclidean distance from P_i to P_f. It is never greater than the distance traveled by a particle.
- **Acceleration distance**

  The **acceleration distance** is the minimum distance at which an object (or, say, flow, flame), accelerating in given conditions, reaches a given speed.
- **Mechanic distance**

  The **mechanic distance** is the position of a particle as a function of time t.

  For a particle, moving linearly with initial position x_0 and initial speed v_0, which is acted upon by a constant acceleration a, it and the speed are given by

  $$x(t) = x_0 + v_0t + \frac{1}{2}at^2 \quad \text{and} \quad v(t) = v_0 + at.$$

  So, the **acceleration distance** fallen under uniform acceleration a, in order to reach a speed v, is $\frac{v^2}{2a}$. A body is *free falling* if it is falling subject only to acceleration g by gravity; the **free fall distance** (distance fallen by it) is $y(t) = \frac{1}{2}gt^2$.
- **Terminal distance**

  The **terminal distance** is the distance of an object, moving linearly in a resistive medium, from an initial position to a stop.

  If object's initial position and speed are x_0, v_0, and the drag per unit mass in the medium is proportional to speed with constant of proportionality β, then the position and speed of a body are given by

  $$x(t) = x_0 + \frac{v_0}{\beta}(1 - e^{-\beta t}) \quad \text{and} \quad v(t) = x^{'}(t) = v_0e^{-\beta t}.$$

  The speed decreases to 0, and the body reaches a *maximum terminal distance*

  $$x_{terminal} = \lim_{t \to \infty} x(t) = x_0 + \frac{v_0}{\beta}.$$

[redacted]

For a body, moving from initial position (x_0, y_0) and speed (v_{x_0}, v_{y_0}), the position $(x(t), y(t))$ is $x(t) = x_0 + \frac{v_{x0}}{\beta}(1 - e^{-\beta t})$, $y(t) = (y_0 + \frac{v_{y0}}{\beta} - \frac{g}{\beta^2}) + \frac{v_{y0}\beta - g}{\beta^2} e^{-\beta t}$. The horizontal motion ceases at a maximum terminal distance $x_{terminal} = x_0 + \frac{v_{x0}}{\beta}$.

- **Ballistics distances**

  Ballistics is the study of the motion of *projectiles*, i.e., bodies which are propelled (or thrown) with some initial velocity, and then allowed to be acted upon by the forces of gravity and possible drag.

  The **trajectory**, **range** and **height** of a projectile are its parabolic path, total horizontal distance traveled and maximum upward distance reached. If projectile is launched on flat ground at velocity v and angle θ to the horizontal, then at the time t of motion, its horizontal and vertical positions are

$$x(t) = vt\cos\theta \quad \text{and} \quad y(t) = vt\cos\theta - \frac{1}{2}gt^2.$$

  So, the range, realized by the *time of flight* $t_{of} = \frac{2v\sin\theta}{g}$, and height are

$$x_{\max} = x(t_{of}) = \frac{v^2 \sin 2\theta}{g} \quad \text{and} \quad y_{\max} = y(\frac{1}{2}t_{of}) = \frac{v\sin^2\theta}{2g},$$

  which are maximized when $\theta = \pi/4$ and $\theta = \pi/2$, respectively.

  The *bullet drop* is the height it loses, because of gravity, between leaving the rifle and reaching the target. In order to ensure that the "zero" (point at which the bullet's path intersects with the *LOS*, line of sight, to the target) will be at a specific range, the shooter should set (using a *sight*, device mounted on the rifle) the *bore angle* between the rifle bore and the LOS. A properly adjusted rifle barrel and sight are said to be *zeroed* (or *sighted-in*). The shooter zeroes rifle at a standard *zero range* and then adjustments are made for other ranges.

  The **point-blank range** is the distance at which the bullet is expected to strike a target of a given size without adjusting the elevation of the firearm.

- **Interaction distance**

  The **impact parameter** is the perpendicular distance between the velocity vector of a projectile and the center of the object it is approaching.

  The **interaction distance** between two particles is the farthest distance of their approach at which it is discernible that they will not pass at the impact parameter, i.e., their distance of closest approach if they had continued to move in their original direction at their original speed.

  The *coefficient of restitution* (COR) of colliding objects A, B is the ratio of speeds after and before an impact, taken along its line. The collision is *inelastic* if COR< 1. COR^2 is the ratio of *rebound* and *drop* distances if A bounces off stationary B.

- **Mean free path (length)**

  The **mean free path (length)** of a particle (photon, atom or molecule) in a medium measures its probability to undergo a situation of a given kind K; it is the average of an exponential distribution of distances until the situation K occurs. In particular, this average distance d is called:

nuclear collusion length if K is a nuclear reaction;
interaction length if K is an interaction which is neither elastic, nor quasi-elastic;
scattering length if K is a scattering event;
attenuation length (or *absorption length*) if K means that the probability $P(d)$, that a particle has not been absorbed, drops to $\frac{1}{e} \approx 0.368$, cf. **Beer-Lambert law**;
radiation length (or *cascade unit*) if K means that the energy of (high energy electromagnetic-interacting) relativistic charged particles drops by the factor $\frac{1}{e}$;
free streaming length if K means that particles become nonrelativistic.

In Gamma-ray Radiography, the *mean free path* of a beam of photons is the average distance a photon travels between collisions with atoms of the target material. It is $\frac{1}{\alpha\rho}$, where α is the material *opacity* and ρ is its density.

- **Neutron scattering length**

  In Physics, *scattering* is the random deviation or reflection of a beam of radiation or a stream of particles by the particles in the medium.

  In Neutron Interferometry, the **scattering length** a is the zero-energy limit of the scattering amplitude $f = -\frac{\sin\delta}{k}$. Since the *total scattering cross-section* (the likelihood of particle interactions) is $4\pi|f|^2$, it can be seen as the radius of a hard sphere from which a point neutron is scattered.

  The spin-independent part of the scattering length is the *coherent scattering length*. In order to expand the scattering formalism to absorption, the scattering length is made complex $a = a' - ia''$.

  Thomson scattering length is the *classical electron radius* $\approx 2.818 \times 10^{-15}$ m.

- **Inelastic mean free path**

  In Electron Microscopy, the **inelastic mean free path** (or IMFP) is the average total distance that an electron traverses between events of inelastic scattering, while the **effective attenuation length** (or EAL) is an experimental parameter reflecting the average net distance traveled.

  The EAL is the thickness in the material through which electron can pass with probability $\frac{1}{e}$ that it survives without inelastic scattering. It is about 20 % less than the IMFP due to the elastic scatterings which deflect the electron trajectories.

  Both are smaller than the total electron range which may be 10–100 times greater.

- **Sampling distance**

  In Electron Spectroscopy for chemical analysis, the **sampling distance** is the lateral distance between areas to be measured for characterizing a *surface*, i.e., the volume from which the photo-electrons can escape.

[redacted]

- **Debye screening distance**

  The **Debye screening distance** (or *Debye length*, *Debye–Hückel length*) is the distance over which a local electric field affects the distribution of mobile charge carriers (for example, electrons) present in the material (plasmas and other conductors).

  Its order increases with decreasing concentration of free charge carriers, from 10^{-4} m in gas discharge to 10^{5} m in intergalactic medium.
- **Range of a charged particle**

  The **range of a charged particle**, passing through a medium and ionizing, is the distance to the point where its energy drops to almost zero.
- **Gyroradius**

  The **gyroradius** (or *cyclotron radius*, *Larmor radius*) is the radius of the circular orbit of a charged particle in the presence of a uniform magnetic field.
- **Radius of gyration**

  The **radius of gyration** of a body about a given axis is the distance from this axis to the centre of gyration. It is the RMS (square root of the mean of the squares) of the distances from the axis of rotation to all the points in the body.
- **Inverse-square laws**

  An **inverse-square law** is any law stating that a quantity or intensity is inversely proportional to the square of the distance d from its source.

  Such law applies when some conserved quantity is evenly radiated outward from a point source in 3D space. The intensity of radiation passing through any unit area (directly facing the point source) is inversely proportional to d^2, since radiation is spread out over the surface area $4\pi d^2$ of a sphere.

  Newton's **law of universal gravitation** (checked above 6×10^{-5} m): the gravitational attraction between two point-like masses m_1, m_2 at distance d is

  $$G\frac{m_1 m_2}{d^2},$$

  where $G = 6.67384(80) \times 10^{-11}$ m^{3}kg^{-1}s^{-2} is the Newton *gravitational constant*.

  The existence of extra dimensions, postulated by M-theory, will be checked by LHC (Large Hadron Collider at CERN, near Geneva) based on the inverse proportionality of the gravitational attraction in nD space to the $(n-1)$-th degree of the distance between objects; if the Universe has a 4-th dimension, LHC will find out the inverse proportionality to the cube of the small interparticle distance.

  The effects of electric, magnetic, light, sound, and radiation phenomena also follows, an inverse-square law for large d. **Coulomb law**: the force of attraction or repulsion between two point-like objects with charges e_1, e_2 at distance d is given by

  $$k_e\frac{e_1 e_2}{d^2},$$

where k_e is the *Coulomb constant* depending upon the medium that the charged objects are immersed in. The gravitational and electrostatic forces of two bodies with *Planck mass* m_P and $e_1 = e_2 = 1$ have equal strength.

The *irradiance* (power per unit area in the direction of propagation) of a spherical wavefront (light, sound, etc.) radiating from a point source decreases (assuming that there are no losses caused by absorption or scattering) inversely proportional to the square d^2 of the distance from the source (cf. **distance decay** in Chap. 29). However, for a radio wave, it decrease like $\frac{1}{d}$.

The **far field** is the region, where the emitter can be considered as a point source (so, inverse-square law holds), usually at a distance of more than 5–10 times the source diameter; the **near field** is the region closer to source.

For light, the distance, from which far field starts, is called **photometric distance**. Cf. **Fraunhofer distance** and, in Chap. 21, **acoustics distances**.

- **Fraunhofer distance**

The Maxwell equations, governing the field strength decay, can be approximated as d^{-3}, d^{-2} and d^{-1} for three regions surrounding an radiating antenna: the *reactive near field*, the *radiating near field* (or *Fresnel region*) and the **far field** (or *Fraunhofer region*). Cf. **inverse-square law**. The Fresnel region begins about at $0.62(\frac{D^3}{\lambda})^{\frac{1}{2}}$, where D is the largest dimension of the antenna and λ is the wavelength. The propagating waves start to dominate here, but only in the far field the distribution of the diffracted energy no longer change with distance.

The **Fraunhofer** (or *far field*, *Rayleigh*) **distance** is $\frac{2D^2}{\lambda}$, the distance where the far field begin. Cf. **acoustic distances** in Chap. 21.

In Optics, beam divergence is defined by its *radius*, i.e., for a Gaussian beam, the distance from the beam propagation axis where intensity drops to $\frac{1}{e^2} \approx 13.5\,\%$ of the maximum. The beam's *waist* (or *focus*) is the position on its axis where the radius is at its minimum. The *imaginary-distance BPM* (Jevick–Hermansson, 1989) refers to beam propagating the (complex electric) field along the imaginary axis.

The beam's **Rayleigh length** (or *Rayleigh range*) Rl is the distance from the waist (in the propagation direction) where the radius increases from w_0 to $\sqrt{2}w_0$, i.e., the beam propagates without diverging much. For Gaussian beams, $Rl = \frac{\pi w_0^2}{\lambda}$, where λ is the vacuum wavelength divided by the refractive index of the material. The Rayleigh length divides the near-field and *mid-field*; it is the distance from the waist at which the wavefront curvature is at a maximum. The divergence really starts in the far field where the beam radius is at least $10Rl$. The *confocal parameter* (or *depth of focus*) of the beam is $2Rl$. Cf. **lens distances** in Chap. 28.

- **Half-value layer**

Ionizing radiation consists of highly-energetic particles or waves (especially, X-rays, gamma rays and far ultraviolet light) which are progressively absorbed during propagation through the surrounding medium, via *ionization*, i.e., removing an electron from some of its atoms or molecules. The **half-value layer** is the depth within a material where half of the incident radiation is absorbed.

[redacted]

A basic rule of protection against ionizing radiation exposure: multiplying the distance from its source by a *distance factor* r decreases this exposure to $\frac{1}{r^2}$ of it.

In Maxwell Render light simulation software, the *attenuation distance* (or *transparency*) is the thickness of object that absorbs 50 % of light energy.

- **Range of fundamental forces**

The fundamental forces (or interactions) are gravity and electromagnetic, weak nuclear and strong nuclear forces. The **range** of a force is considered *short* if it decays (approaches 0) exponentially as the distance d increases.

Both electromagnetic force and gravity are forces of infinite range which obey **inverse-square laws**. The shorter the range, the higher the energy. Both weak and strong forces are very short range (about 10^{-17} m and 10^{-15} m, respectively) which is limited by the uncertainty principle.

At subatomic distances, Quantum Field Theory describes electromagnetic, weak and strong interactions with the same formalism but different constants. Quantum Electrodynamics describes electromagnetism via photon exchanges between charged particles and Quantum Chromodynamics describe strong interactions via gluon exchanges between quarks. Strong interaction force grows stronger with the distance. Three forces almost coincide at very large energy, but at large distances they are irrelevant compared with gravity. The number of fundamental particles increases on smaller distance scales. But at macroscopic scales, those particles can collectively create emerging phenomena, say, superconductivity.

General Relativity has been probed from submillimeter up to Solar System scales but at cosmological scale it require the presence of *dark matter* and *dark energy*. Maxwell's electromagnetism has been probed from atomic distances up to 1.3 AU (order of the **coherence lengths** of the magnetic fields dragged by the solar wind) but it does not explain magnetic fields found in galaxies, clusters and voids.

At cosmological scale, the repulsive force of putative *dark energy*, due to *vacuum energy* (or *cosmological constant*) overtakes gravity; cf. **metric expansion of space** in Chap. 26. Dark energy is the only substance known to act both on subatomic and cosmological scale. Its effect is measured only on a scale larger than superclusters. Khoury–Weltman, 2004, in order to explain dark energy, conjectured *fifth force* with range depending on density of matter in its environment, say, 1 mm in Earth's vicinity and 10^7 ly in cosmos. An alternative to dark energy: possible, in String Theory, modifications of gravity at ultra large distances (i.e., small curvatures) due to some specific compactification of extra dimensions.

Alexander-Katz et al., 2016, found a new kind of long-range interaction between particles, in a liquid medium, that is based entirely on their motions. A small number of magnetic metal microparticles a few μm across, were interspersed with a much larger quantity of inert particles of comparable size, all suspended in water. When a rotating magnetic field was applied, the metal particles would begin to spin and, even when separated by distances tens of times their size, would migrate toward each other, through a slow series of motions. The attraction depends on the speed of spinning and amount of inert matter.

- **EM radiation wavelength range**

  The *wavelength* is the distance $\lambda = \frac{c}{f}$ the wave travels to complete one cycle.

  Electromagnetic (EM) **radiation wavelength range** is infinite and continuous in principle. The limits of short and long waves are the vicinity of the **Planck length** and the size of Universe, respectively.

  The wavelengths are: < 0.01 nm for gamma rays, 0.01–10 nm for X-rays, 100–400 nm for ultraviolet, 400–780 nm for visible light, 0.78–1000 μm for infrared (in lasers), 1–330 mm for microwave, 0.33–3000 m for radio frequency radiation, > 3 km for low frequency, and ∞ for static field.

  Besides gamma rays, X-rays and far ultraviolet, the EM radiation is *nonionizing*, i.e., passing through matter, it only *excites* electrons: moves them to a higher energy state, instead of removing them completely from an atom or molecule.

- **Compton wavelength**

  Compton scattering is the scattering of (X-ray or gamma ray) photons by electrons. It results in a decrease in energy (so, increase in wavelength) of the photon.

  Compton wavelength $\lambda_C(m)$ and **reduced Compton wavelength** $\overline{\lambda}_C(m)$ of a particle with rest mass m (where c is the speed of light, $\hbar$ is the *reduced Planck constant* $\frac{h}{2\pi}$ and l_P, m_P are Planck length and mass) are defined by

$$\frac{1}{2\pi}\lambda_C(m) = \overline{\lambda}_C(m) = \frac{\hbar}{mc} = \frac{m_P}{m} l_P.$$

- **Radiation attenuation with distance**

  Radiation is the process by which energy is emitted from a source and propagated through the surrounding medium. Radiant energy described in wave terms includes sound and electromagnetic radiation, such as light, X-rays and gamma rays. The incident radiation partially changes its direction, gets absorbed, and the remainder transmitted. The change of direction is *reflection*, *diffraction*, or *scattering* if the direction of the outgoing radiation is reversed, split into separate rays, or randomized (diffused), respectively. Scattering occurs in nonhomogeneous media.

  In Physics, *attenuation* is any process in which the flux density, power amplitude or intensity of a wave, beam or signal decreases with increasing distance from the energy source, as a result of absorption of energy and scattering out of the beam by the transmitting medium. It comes in addition to the divergence of flux caused by distance alone as described by the **inverse-square laws**.

  Attenuation of light is caused mainly by scattering and absorption of photons. The primary causes of attenuation in matter are the *photoelectric effect* (emission of electrons), *Compton scattering* (cf. **Compton wavelength**) and *pair production* (creation of an elementary particle and its antiparticle from a high-energy photon).

[redacted]

In Physics, *absorption* is a process in which atoms, molecules, or ions enter some bulk phase—gas, liquid or solid material; in *adsorption*, the molecules are taken up by the surface, not by the volume. *Absorption of EM radiation* is the process by which the energy of a photon is taken up (and destroyed) by, for example, an atom whose valence electrons make the transition between two electronic energy levels. The absorbed energy may be re-emitted or transformed into heat.

Attenuation is measured in units of decibels (dB) or *nepers* ($\approx$ 8.7 dB) per length unit of the medium and is represented by the medium *attenuation coefficient* α. When possible, specific absorption or scattering coefficient is used instead.

Attenuation (or *loss*) *of signal* is the reduction of its strength during transmission. In Signal Propagation, attenuation of a propagating EM wave is called the *path loss*. Path loss may be due to free-space loss, refraction, diffraction, reflection, absorption, aperture-medium coupling loss, etc. of antennas. Path loss in decibels is $L = 10n \lg d + C$, where n is the path loss exponent, d is the transmitter-receiver distance in m, and C is a constant accounting for system losses.

The *free-space path loss* (FSPL) is the loss in signal strength of an EM wave that would result from a line-of-sight path through free space, with no obstacles to cause reflection or diffraction. FSPL is $(\frac{4\pi d}{\lambda})^2$, where d is the distance from the transmitter and λ is the signal wavelength (both in m), i.e., in dB it is $10 \lg (FSPL) = 20 \lg d + 20 \lg f - 147.56$, where f is the frequency in Hz.

- **Beer–Lambert law**

  The **Beer–Lambert law** is an empirical relationship for the *absorbance Ab* of a substance when a radiation beam of given frequency goes through it:

$$Ab = \alpha d = -\log_a T,$$

where $a = e$ or (for liquids) 10, d is the *path length* (distance the beam travels through the medium), $T = \frac{I_d}{I_0}$ is the *transmittance* (I_d and I_0 are the intensity of the transmitted and incident radiation), and α is the medium *opacity* (or *linear attenuation coefficient*, *absorption coefficient*); α is the fraction of radiation lost to absorption and/or scattering per unit length of the medium.

The *extinction coefficient* is $\frac{\lambda_w}{4\pi}\alpha$, where λ_w is the same frequency wavelength in a vacuum. In Chemistry, α is given as ϵC, where C is the absorber *concentration*, and ϵ is the *molar extinction coefficient*.

The **optical depth** is $\tau = -\ln \frac{I_d}{I_0}$, measured along the true (slant) optical path.

The **penetration depth** (or **attenuation length**, *mean free path*, *optical extinction length*) is the thickness d in the medium where the intensity I_d has decreased to $\frac{1}{e}$ of I_0; so, it is $\frac{1}{\alpha}$. Cf. **half-value layer**.

Also, in Helioseismology, the (meridional flow) *penetration depth* is the distance from the base of the solar convection zone to the location of the first reversal of the meridional velocity. In an information network, the *message* *penetration distance* is the maximum distance from the event message traverses in the valid routing region.

The **skin depth** is the thickness d where the amplitude A_d of a propagating wave (say, alternating current in a conductor) has decreased to $\frac{1}{e}$ of its initial value A_0; it is twice the penetration depth. The *propagation constant* is $\gamma = -\ln\frac{A_d}{A_0}$.

The Beer–Lambert law can describe also the attenuation of solar or stellar radiation. The main components of the atmospheric light attenuation are: absorption and scattering by aerosols, Rayleigh scattering (from molecular oxygen O_2 and nitrogen Ni_2) and (only absorption) by carbon dioxide CO_2, O_2, nitrogen dioxide NiO_2, water vapor, ozone O_3. Cf. **atmospheric visibility distances** in Chap. 25.

The sea is nearly opaque to light: less than 1 % penetrates 100 m deep. Cf. **distances in Oceanography** in Chap. 25. In Oceanography, attenuation of light is the decrease in its intensity with depth due to absorption (by water molecules) and scattering (by suspended fine particles). The transparency of the water in oceans and lakes is measured by the *Secchi depth* d_S at which the reflectance equals the intensity of light backscattered from the water. Then $\alpha = \frac{10d_S}{17}$ is used as the *light attenuation coefficient* in the Beer–Lambert law $\alpha d = -\ln\frac{I_d}{I_0}$, in order to estimate I_d, the intensity of light at depth d, from I_0, its intensity at the surface.

In Astronomy, attenuation of EM radiation is called *extinction* (or *reddening*). It arises from the absorption and scattering by the interstellar medium, the Earth's atmosphere and dust around an observed object.

The *photosphere* of a star is the surface where its optical depth is $\frac{2}{3}$. energy emitted. The *optical depth of a planetary ring* is the proportion of light blocked by the ring when it lies between the source and the observer.

- **Atmospheric depth**

  The total absorption along a slanted path through matter of density ρ at the angle θ from the vertical is measured by the **interaction depth**, i.e., $\frac{1}{\cos\theta}\int\rho(x)d(x)$.

  In particular, the **atmospheric depth** (or *effective thickness of the atmosphere*) at altitude h is defined as

$$A(h) = \frac{1}{\cos\theta}\int_h^\infty \rho(x)d(x).$$

  It is $\approx$ 11 times larger when the Sun is just above the horizon than at the shortest path (solar zenith). Vertical $A(h)$ is approximatively equal to the pressure divided by $g = 9.8$ m/s^2.

- **Arago distance**

  The *Arago point* is a *neutral point* (where the degree of polarization of skylight goes to zero) located $\approx 20°$ directly above the *antisolar point* (the point on the celestial sphere that lies directly opposite the sun from the observer) in relatively clear air and at higher elevations in turbid air.

[redacted]

So, the **Arago distance**, i.e., the angular distance from the antisolar point to the Arago point, is a measure of *atmospheric turbidity* (effect of aerosols in reducing the transmission of direct solar radiation).

Another useful measure of turbidity is *aerosol optical depth*, i.e., the **optical depth** due to extinction by the aerosol component of the atmosphere.

- **Sound attenuation with distance**

Vibrations propagate through elastic solids and liquids, including the Earth, and consist of *elastic* (or *seismic*, *body*) *waves* and *surface* (occurring since it acts as an solid-gas interface) *waves*. Elastic waves are: primary (P) wave moving in the propagation direction of the wave and *secondary* (S) wave moving in this direction and perpendicular to it. Surface waves are: the *Love* wave moving perpendicular to the direction of the wave and the *Rayleigh* (R) wave moving in the direction of the wave and circularly within the vertical surface perpendicular to it. The attenuation of P- and S-waves is proportional to $\frac{1}{d^2}$ or $\frac{1}{d}$, when propagated by the surface or inside of an infinite elastic body. For the R-wave, it is proportional to $\frac{1}{\sqrt{d}}$.

Sound propagates through gas (say, air) as a P-wave and attenuates over a distance, at a rate of $\frac{1}{d^2}$. The *far field* (cf. **Fraunhofer distance**) is the part of a sound field in which sound pressure (if it is the same in all directions) decreases according to the *inverse distance law* $\frac{1}{d}$; but sound intensity decreases as $\frac{1}{d^2}$.

In natural media, further weakening occurs from *attenuation*, i.e., *scattering* (reflection of the sound in other directions) and *absorption* (conversion of the sound energy to heat). Cf. **critical distance** among **acoustics distances** in Chap. 21.

The **sound extinction distance** is the distance over which its intensity falls to $\frac{1}{e}$ of its original value. For sonic boom intensities (say, supersonic flights), the lateral *extinction distance* is the distance where in 99 % of cases the sound intensity is lower than 0.1–0.2 mbar (10–20 pascals) of atmospheric pressure. Cf. earthquake *extinction length* in **distances in Seismology** (Chap. 25).

Water is transparent to sound. Sound energy is absorbed (due to viscosity) and $\approx$ 6 % of it is scattered (due to water inhomogeneities). Absorbed less, low frequency sounds can propagate over large distances along lines of minimum sound speed. High frequency waves attenuate more rapidly. So, low frequency waves are dominant further from the source (say, a musical band or earthquake).

Attenuation of ultrasound waves with frequency f MHz at a given distance r cm is αfr decibels, where α is the *attenuation coefficient* of the medium. It is used in Ultrasound Biomicroscopy; in a homogeneous medium (so, without scattering) α is 0.0022, 0.18, 0.85, 20, 41 for water, blood, brain, bone, lung, respectively.

- **Lighting distance**

Sound travels through air at 330–350 m/s (depending on altitude, relative humidity, pressure, etc.), while the speed of light is $c \approx 300 \times 10^6$ m/s. In fact, structure of light pulses reduces their speed, even in a vacuum.

So, the **lighting distance** (of a lightning bolt from an observer) in km is $\approx \frac{1}{3}$ of the delay, in seconds, between observer's seeing it and hearing thunder.

- **Optical distance**

  The **optical distance** (or *optical path length*) is a distance dn traveled by light, where d is the physical distance in a medium and $n = \frac{c}{v}$ is the medium's *refractive index* (c, v are the speeds of an EM wave in a vacuum and in the medium). By *Fermat's principle* light follows the shortest optical path. Cf. **optical depth**.

  The **light extinction distance** is the distance where light propagating through a given medium reaches its *steady-state speed*, i.e., a characteristic speed that it can maintain indefinitely. It is proportional to $\frac{1}{\rho\lambda}$, where ρ is the density of the medium and λ is the wavelength, and it is very small for most common media.

- **Edge perimeter distance**

  In semiconductor technology, the **edge perimeter distance** is the distance from the edge of a *wafer* (thin slice with parallel faces cut from a semiconductor crystal) in a wafer carrier to the top face of the wafer carrier.

- **Proximity effects**

  In Electronic Engineering, an alternating current flowing through an electric conductor induces (via the associated magnetic field) eddy currents within the conductor. The *electromagnetic proximity effect* is the "current crowding" which occurs when such currents are flowing through several nearby conductors such as within a wire. It increases the alternating current resistance (so, electrical losses) and generates undesirable heating.

  In Nanotechnology, the *quantum* $\frac{1}{f}$ *proximity effect* is that the $\frac{1}{f}$ fundamental noise in a semiconductor sample is increased by the presence of another similar current-carrying sample placed in the close vicinity.

  The *superconducting proximity effect* is the propagation of superconductivity through a NS (normal-superconductor) interface, i.e., a very thin layer of "normal" metal behaves like a superconductor (that is, with no resistance) when placed between two thicker superconductor slices.

  In Lithography, if a material is exposed to an electronic beam, some molecular chains break and many electron scattering events occur. Any pattern written by the beam on the material can be distorted by this *E-beam proximity effect*.

  In LECD (localized electrochemical deposition) technique for fabrication of miniature devices, the electrode (anode) is placed close to the tip of a fabricated structure (cathode). Voltage is applied and the structure is grown by deposition. The *LECD proximity effect*: at small cathode-anode distances, migration overcomes diffusion, the deposition rate increases greatly and the products are porous.

  In Atomic Physics, the **proximity effect** refers to the intramolecular interaction between two (or more) functional groups (in terms of group contributions models of a molecule) that affects their properties and those of the groups located nearby.

  Cf. also *proximity effect (audio)* among **acoustics distances** in Chap. 21.

[redacted]

The term *proximity effect* is also used more abstractly, to describe some undesirable proximity phenomena. For example, the *proximity effect in the production of chromosome aberrations* (when ionizing radiation breaks double-stranded DNA) is that DNA strands can misrejoin if separated by less than $\frac{1}{3}$ of the diameter of a cell nucleus. The *proximity effect in innovation process* is the tendency to the geographic agglomeration of innovation activity.

- **Hopping distance**

  Hopping is atomic-scale long range dynamics that controls diffusivity and conductivity. For example, oxidation of DNA (loss of an electron) generates a radical cation which can migrate a long (more that 20 nm) distance, called the **hopping distance**, from site to site before it is trapped by reaction with water.

- **Atomic jump distance**

  In the solid state, the atoms are about closely packed on a rigid lattice. The atoms of some elements (carbon, hydrogen, nitrogen), being too small to replace the atoms of metallic elements on the lattice, are located in the interstices between metal atoms and they diffuse by squeezing between the host atoms.

  Interstitial diffusion is the only mechanism by which atoms can be transported through a solid substance while, in a gas or liquid, mass transport is possible by both diffusion and the flow of fluid (for example, convection currents).

  The **jump distance** is the distance an atom is moved through the lattice in a given direction by one exchange of its position with an adjacent lattice site.

  Some crystals can jump 1000 times their own length under light, since light energy rearranges atoms and builds strain, which is then explosively released.

  The **mean square diffusion distance** d_t from the starting point which a molecule will have diffused in time t, satisfies $d_t^2 = r^2 N = r^2 \nu t = 2nDt$, where r is the jump distance, N is the number of jumps (equal to νt assuming a fixed jump rate ν), $n = 1, 2, 3$ for $1, 2, 3$-dimensional diffusion, and $D = \frac{\nu r^2}{2n}$ is the *diffusivity* in cm^2/s.

  In diffusion alloy bonding, a **characteristic diffusion distance** is the distance between the joint interface and the site wherein the concentration of the diffusing substance (say, aluminum in high carbon-steel) falls to zero up to a given error.

- **Diffusion length**

  Diffusion is a process of spontaneous spreading of matter, heat, momentum, or light: particles move to lower chemical potential implying concentration change.

  In Microfluidics, the **diffusion length** is the distance from the point of initial mixing to the complete mixing point where the equilibrium composition is reached.

  In semiconductors, electron-hole pairs are generated and recombine. The (*minority carrier*) **diffusion length** of a material is the average distance a minority carrier can move from the point of generation until it recombines with majority carriers. Also, in electron transport by diffusion, the *diffusion length* is the distance over which concentration of free charge carriers injected into semiconductor falls to $\frac{1}{e}$ of its original value.

  Cf. **jump distance** and, in Chap. 23, **capillary diffusion distance**.

- **Thermal diffusion length**

  The heat propagation into material is represented by the **thermal diffusion length**, i.e., the propagation distance of the thermal wave producing an attenuation of the peak temperature to about 0.1 of the maximum surface value.

  The propagation of the laser-generated shock wave is approximated as blast wave (instantaneous, massless point explosion). The **expansion distance** is the distance between the surface of the target and the position of a blast wave; it depends on the energy converted into the plasma state.
- **Thermal entrance length**

  In heat transfer at a boundary (surface) within a fluid, the **thermal entrance length** is the distance required for the *Nusselt number* (ratio of convective to conductive heat transfer across normal to the boundary) associated with the pipe flow to decrease to within 5 % of its value for a fully developed heat flow.
- **Distance-to-spot ratio**

  The **distance-to-spot ratio** of an infrared temperature sensing device is the ratio of the distance to the object and the diameter of the t° measurement area.
- **Bjerrum length**

  The **Bjerrum length** is the separation at which the electrostatic interaction between two elementary charges is comparable in magnitude to the thermal energy scale, k_B, T, where k_B is the Boltzmann constant and T is the temperature in kelvin.
- **Lagrangian radius**

  The **Lagrangian radius** of the particle is the distance from the explosion center to a particle at the moment the shock front passes through it.

  Cf. unrelated *Lagrangian radii* among the **radii of a star system** (Chap. 25).
- **Reynolds number**

  For an object of a characteristic length (Chap. 29) l, flowing with mean relative velocity v in a *fluid* (liquid or gas) of the density ρ and dynamic viscosity μ, the **Reynolds number** is the ratio $Re = \frac{\rho v l}{\mu}$ of inertial forces to viscous forces. The *flow* is *smooth* (or *laminar*) if Re is low (viscous forces dominate), *rough* (or *turbulent*) if Re is high (usually $Re \geq 10^5$) and *transitional* in between. In a *Stokes flow* (laminar flow with very low Re), the inertial forces are negligible.

  The *law of the wall* (von Kármán, 1930) states that the average velocity of a turbulent flow close to the *wall* (boundary of the fluid region) is proportional to $\ln y^+$, where the **wall distance** $y^+ = \frac{u'}{\nu} y$ is the distance y to the wall, made dimensionless with the friction velocity u' at the wall and fluid's kinematic viscosity ν.

  In swimming, Re is 10^{-5}, 4×10^{-3}, $10^{-1}-10$, 5×10^4 and 3×10^8 for bacterium, spermatozoa, small zooplankton, large fish and whale, respectively. In flying, Re is $30–4 \times 10^4$ for insects, $10^3–10^5$ for birds, 1.6×10^6 for a glider and 2×10^9 for Boeing 747. Blood flow has $Re = 10^{-3}$, 140, 500 and 3.4×10^3 in capillary, vein, artery and aorta, respectively. Re is a *dimensionless parameter*, i.e., the units of measurement in it cancel out. Examples of other such flow parameters follow.

  The *Mach number Ma* is a ratio of the speed of flow to the speed of sound in a fluid. Ma is ratio of inertia to *compressibility* (volume change as a response to

[redacted]

a pressure). The flow is *subsonic*, *supersonic*, *transonic* or *hypersonic* if $Ma < 1, 0.8 \leq Ma \leq 1.5$ or $Ma \geq 5$, respectively. Ma governs *compressible* (i.e., those with $Ma > 0.3$) flows.

The *Froude number* $Fr = \frac{v}{\sqrt{gl}}$, where v, d, g are characteristic flow velocity, depth and external field (usually, gravity), is the ratio of the flow inertia to the external field; it governs open-channel flows. Alexander, 1984: legged animals moving with the same Froude number have similar gait patterns and *duty factors* (percentages of a stride when a foot touch the ground); cf. **gait distances** in Chap. 29.

The *lift* L and *drag* D are perpendicular and, respectively, parallel (to the oncoming flow direction) components of the force fluid flowing past the surface of a body exerts on it. In a flight without wind, the lift-to-drag ratio $\frac{L}{D}$ is the horizontal distance traveled divided by the altitude lost. $\frac{L}{D}$ is $4, 17, 20, 37$ for cruising house sparrow, Boeing 747, albatross, Lockheed U-2, respectively. Küchemann, 1978, found that range-maximizing L for high Ma is $\approx 4 + \frac{12}{Ma}$.

- **Turbulence length scales**

  Turbulence is the time dependent chaotic behavior of fluid flows. The turbulent field consists of the superposition of interacting *eddies* (coherent patterns of velocity, vorticity and pressure) of different length scales. The kinetic energy cascades from the eddies of largest scales down to the smallest ones generated from the larger ones through the nonlinear process of vortex stretching.

  The **turbulence length scales** are measures of the eddy scale sizes in turbulent flow. Such standard length scales for largest, smallest and intermediate eddy sizes are called **integral length scale**, **Kolmogorov microscale** and **Taylor microscale**, respectively. The corresponding ranges are called *energy-containing*, *dissipation* and *inertial* range.

  Integral length scale measures the largest separation distance over which components of the eddy velocities at two distinct points are correlated; it depends usually of the flow geometry. For example, the largest integral scale of pipe flow is the pipe diameter. For atmospheric turbulence, this length can reach several hundreds km. On intermediate Taylor microscale, turbulence kinetic energy is neither generated nor destroyed but is transferred from larger to smaller scales.

  At the smallest scale, the dynamics of the small eddies become independent of the large-scale eddies, and the rate at which energy is supplied is equal to the rate at which it is dissipated into heat by viscosity. The Kolmogorov length microscale is given by $\tau = (\frac{\nu^3}{\epsilon})^{\frac{1}{4}}$, where ϵ is the average rate of energy dissipation per unit mass and ν is the kinematic viscosity of the fluid. This microscale describe the smallest scales of turbulence before viscosity dominates.

  Similarly, the *Batchelor scale* (usually, smaller) describes the smallest length of fluctuations in scalar concentration before molecular diffusion dominates.

  Quantum turbulence is the chaotic motion of quantum fluids (say, *superfluids*) at high flow rates and close to absolute zero.

Turbulence is well described by the *Navier–Sokes equations*. Clay Mathematics Institute list the investigation, whether those equations in 3D always have a nonsingular solution, among the six US 1,000,000-valued open problems.

- **Meter of water equivalent**

  The **meter of water equivalent** (mwe) of a material is the thickness of that material that provides the equivalent shielding of 1 m of water.

  Also, the *mass balance* of a glacier is reported in mwe as the ratio of the volume (of water that would be obtained from melting the snow/ice) and area; it gives the change of thickness in water depth.

  Unrelated *centimeter of water* (cmH_2O) is the pressure $\approx$ 98.1 Pa (pascals) exerted by a column of water of 1 cm in height at 4° at acceleration g. A similar manometric unit of pressure is the *millimeter of mercury* (mmHg) $\approx$ 1 Torr $\approx$ 133.3 Pa.

- **Hydraulic diameter**

  For flow in a (in general, noncircular) pipe or tube, the **hydraulic diameter** is $\frac{4A}{P}$, where A is the cross-sectional area and P is the *wetted perimeter*, i.e., the perimeter of all channel walls that are in direct contact with the flow.

  So, in open liquid flow, the length exposed to air is not included in P. The hydraulic diameter of a circular tube is equal to its inside diameter. The **hydraulic radius** is (nonstandardly) defined as $\frac{1}{4}$ of the hydraulic diameter.

- **Hydrodynamic radius**

  The **hydrodynamic radius** (or *Stokes radius*, *Stokes–Einstein radius*) of a molecule, undergoing diffusion in a solution (homogeneous mixture of two or more substances), is the hypothetical radius of a hard sphere which diffuses with the same rate as the molecule. Cf. the **characteristic diameters** in Chap. 29.

- **Wigner–Seitz radius**

  The **Wigner–Seitz radius** is the radius of a sphere whose volume is equal to the mean volume $\frac{V}{n} = \frac{1}{n}$ per particle in a solid; n is the particle density.

  So, it is $(\frac{4}{3\pi n})^{\frac{1}{3}}$ an estimation of the mean interparticle distance.

- **Chromatographic migration distances**

  In thin layer Chromatography, the **solvent migration distance** is the distance d_{sl} traveled by the front line of the liquid or gas entering a chromatographic bed for *elution* (the process of using a solvent to extract an absorbed substance from a solid medium). The **migration distance of substance** is distance d_{sb} traveled by the center of a spot. The *retardation* and *retention factors* are $\frac{d_{sl}}{d_{sb}}$ and $\frac{d_{sl}}{d_{sb}} - 1$.

  The **retention distance** is a measure of equal-spreading of the spots on the chromatographic plate, defined via retention factors of sorted compounds.

- **Droplet radii**

  Let A be a small liquid droplet in equilibrium with a *supersaturated vapor*, i.e., a vapor which will begin to condense in the presence of nucleation centers.

  Let ρ_l, ρ_v be the liquid and vapor density, respectively, and let p_l, p_v be the liquid and vapor pressure. Let γ and γ_0 be the actual value at the surface of tension and planar limit value of surface tension. The **capillarity radius** R_c of A is defined by the *Young–Laplace equation* $\frac{\gamma_0}{R_e} = \frac{1}{2}(p_l - p_v)$.

[redacted]

The **surface of tension radius** (or **Kelvin–Laplace radius**, *equilibrium radius of curvature*) R_s is defined by $\frac{\gamma}{R_s} = \frac{1}{2}(p_l - p_v)$. The reciprocal of R_s is the mean curvature $H = \frac{1}{2}(k_1 + k_2)$ (cf. Chap. 8) of the *Gibbs surface of tension*, for which the Young–Laplace equation holds exactly for all droplet radii.

The **equimolar radius** (or *Gibbs adsorption radius*) R_e of A is the radius of a ball of *equimolar* (i.e., with the same *molar concentration*) volume. Roughly, this ball has uniform density ρ_l in the cubic cell of density ρ_v.

The **Tolman length** and the **excess equimolar radius** of the droplet A are $\delta = R_e - R_s$ and $\tau = R_e - R_c$, respectively.

On the other hand, the **cloud drop effective radius** is a weighted mean of the size distribution of cloud droplets.

- **Dephasing length**

  Intense laser pulses traveling through plasma can generate, for example, a wake (the region of turbulence around a solid body moving relative to a liquid, caused by its flow around the body) or X-rays. The **dephasing length** is the distance after which the electrons outrun the wake, or (for a given mismatch in speed of pulses and X-rays) laser and X-rays slip out of phase.

- **Healing length**

  A *Bose–Einstein condensate* (BEC) is a state of dilute gas of weakly interacting bosons confined in an external potential, and cooled to temperatures near absolute zero (0 K, i.e., −273.15 °C), so that a large fraction of them occupy the lowest quantum state of the potential, and quantum effects become apparent on a macroscopic scale. Examples of BEC are superconductors (materials loosing all electrical resistance if cooled below critical temperature), *superfluids* liquid states with no viscosity) and *supersolids* (spatially ordered materials with superfluid properties).

  The **healing length** of BEC is the width of the bounding region over which the probability density of the condensate drops to zero. For a superfluid, say, it is a length over which the wave function can vary while still minimizing energy.

- **Coupling length**

  In optical fiber devices mode coupling occurs during transmission by multimode fibers (mainly because of random bending of the fiber axis). Between two modes, a and b, the **coupling length** l_c is the length for which the complete power transfer cycle (from a to b and back) take place, and the **beating length** z is the length along which the modes accumulate a 2π phase difference. The resonant coupling effect is *adiabatic* (no heat is transferred) if and only if $l_c > z$. Furuya–Suematsu–Tokiwa, 1978, define the *coupling length* of modes a and b as the length of transmission at which the ratio $\frac{I_a}{I_b}$ of mode intensities reach e^2.

- **Localization length**

  Generally, the **localization length** is the average distance between two obstacles in a given scale.

  The localization scaling theory of metal-insulator transitions predicts that, in zero magnetic field, electronic wave functions are always localized in disordered 2D systems over a length scale called the **localization length**.

- **Thermodynamic length**

  Thermodynamic length (Weinhold, 1975) is a Riemannian metric defined on a manifold of equilibrium states of a thermodynamic system.

  It is a path function that measures the distance along a path in the state space. Cf. the **thermodynamic metrics** in Chap. 7.

- **Magnetic length**

  The **magnetic length** (or *effective magnetic length*) is the distance between the effective magnetic poles of a magnet.

  The *magnetic correlation length* is a magnetic-field dependent **correlation length**.

- **Correlation length**

  The **correlation length** (or *correlation radius*) is the distance from a point beyond which there is no further correlation of a physical property associated with that point. It is used mainly in statistical mechanics as a measure of the order in a system for phase transitions (fluid, ferromagnetic, nematic).

  For example, in a spin system at high temperature, the correlation length is $-\frac{\ln d \cdot C(d)}{d}$ where d is the distance between spins and $C(d)$ is the correlation function.

  In particular, the *percolation correlation length* is an average distance between two sites belonging to the same cluster, while the *thermal correlation length* is an average diameter of spin clusters in thermal equilibrium at a given temperature. In second-order phase transitions, the correlation length diverges at the *critical point*.

  In wireless communication systems with multiple antennas, **spatial correlation** is a correlation between a signal's direction and the average received signal gain.

- **Long range order**

  A physical system has **long range order** if remote portions of the same sample exhibit correlated behavior. For example, in crystals and some liquids, the positions of an atom and its neighbors define the positions of all other atoms.

  Examples of long range ordered states are: superfluidity and, in solids, magnetism, charge density wave, superconductivity. Most strongly correlated systems develop long range order in their ground state.

  Short range refers to the finite **correlation length**, say, to the first- or second-nearest neighbors of an atom.

  The system has **long range order**, *quasi-long range order* or is *disordered* if the corresponding correlation function decays at large distances to a constant, to 0 polynomially, or to 0 exponentially. Cf. **long range dependency** in Chap. 18.

- **Spatial coherence length**

  The **spatial coherence length** is the propagation distance from a coherent source to the farthest point where an electromagnetic wave still maintains a specific degree of coherence. This notion is used in Telecommunication Engineering (usually, for the optical regime) and in synchrotron X-ray Optics (the advanced characteristics of synchrotron sources provide highly coherent X-rays).

[redacted]

The spatial coherence length is about 20 cm, 100 m and 100 km for helium-neon, semiconductor and fiber lasers, respectively. Cf. *temporal coherence length* which describes the correlation between signals observed at different moments of time.

For vortex-loop phase transitions (superconductors, superfluid, etc.), **coherence length** is the diameter of the largest thermally excited loop. Besides coherence length, the second **characteristic length** (Chap. 29) in a superconductor is its **penetration depth**. If the ratio of these values (the *Ginzburg–Landau parameter*) is $< \sqrt{2}$, then the phase transition to superconductivity is of second-order.

- **Decoherence length**

  In disordered media, the **decoherence length** is the propagation distance of a wave from a coherent source to the point beyond which the phase is irreversibly destroyed (for example, by a coupling with noisy environment).

- **Critical radius**

  Critical radius is the minimum size that must be formed by atoms or molecules clustering together (in a gas, liquid or solid) before a new-phase inclusion (a bubble, a droplet, or a solid particle) is stable and begins to grow. Say, when atoms, under changes in temperature, pressure or composition (*doping*), get closer together in a solid, their electronic state levels broaden, overlap and hybridize. At a certain *critical distance*, a *Mott insulator-metal transition* occurs.

- **Binding energy**

  The **binding energy** of a system is the mechanical energy required to separate its parts so that their relative distances become infinite. For example, the binding energy of an electron or proton is the energy needed to remove it from the atom or the nucleus, respectively, to an infinite distance.

  In Astrophysics, *gravitational binding energy* of a celestial body is the energy required to disassemble it into dust and gas, while the lower *gravitational potential energy* is needed to separate two bodies to infinite distance, keeping each intact.

- **Metric theory of gravity**

  A **metric theory of gravity** assumes the existence of a symmetric metric (seen as a property of space-time itself) to which matter and nongravitational fields respond. Such theories differ by the types of additional gravitational fields, say, by dependency or not on the location and/or velocity of the local systems. General Relativity is one such theory; it contains only one gravitational field, the space-time metric itself, and it is governed by Einstein's partial differential equations. It has been found empirically that, besides Nordström's 1913 *conformally-flat scalar theory*, every other metric theory of gravity introduces auxiliary gravitational fields.

  A **bimetric theory of gravity** is (Rosen, 1973) a metric theory of gravity in which two, instead of one, metric tensors are used for, say, effective Riemannian and background Minkowski space-times. But usually, rather two frames (not two metric tensors) are considered. Cf. **multimetric** in Chap. 3.

The *Brans–Dicke theory* is a metric theory of gravity, in which $\frac{1}{G}$, where G is the gravitational constant, is replaced by a scalar field. Another direct competitor of General Relativity is affine *Einstein–Cartan–Sciama–Kibble theory* relaxing the assumption that the metric be torsion-free and interpreting spin as affine torsion. It supposes (Sakharov, 1967) *Induced Gravity* with space-time background emerging as a mean field approximation of underlying microscopic degrees of freedom. Such quantum gravity is implied by a *World Crystal model* of **quantum space-time**.

Østvang, 2001, proposed a quasi-metric framework for relativistic gravity.

Classical physics adequately describes gravity only for masses of $10^{-23} - 10^{30}$ kg.

- **Schwarzschild radius**

The **Schwarzschild radius** of a mass m is the radius $r_s(m) = \frac{2Gm}{c^2} = \frac{2m}{m_P} l_P$ of a sphere S such that, if m is compressed within S, it will become a *Schwarzschild* (i.e., uncharged and with angular momentum zero) *black hole*, and so, the escape speed from the surface of S would be the speed c of light.

For such hole, the radii of *photon sphere* (where photons are forced by gravity to travel in circular orbits), of *marginally bound orbit* (where a test particle starts to be gravitationally bound) and of *marginally stable orbit* (smallest circular orbit for material, usually the inner edge of the accretion cloud) are $\frac{3}{2} r_g$, $2r_g$ and $3r_g$.

A typical (stellar) black hole has mass $\approx 6\,M_\odot$ (where $M_\odot \approx 1.98 \times 10^{30}$ kg denotes the solar mass), diameter $\approx$ 18 km, temperature $\approx 10^{-8}$ K and lifetime $\approx 2 \times 10^{68}$ years. The mass of Sgr A^*, the black hole in the center of Milky Way is $4 \times 10^6 M_\odot$; its radius is at most 12.5 light-hours (45 AU). The masses of known stellar and supermassive black holes are within [3.8–15.7] and [5×10^4–3×10^{10}] suns, respectively. Most black holes do not exceed 0.1 % of the mass of their host galaxies, but the one in NGC 1277 reached $17 \times 10^9 M_\odot$ (14 % of this galaxy's mass).

The Schwarzschild radius of H1821+643 (black hole with largest precisely measured mass, $3 \times 10^{10} M_\odot$) is about 575 AU, 28 times the radius of Pluto's orbit.

The transition point separating neutron stars and black holes is expected within 1.7–2.7 $\mathrm{M}_\odot$. *Neutron stars* are composed of the densest known form of matter. The radius of J0348-0432, the largest known ($2.04\,M_\odot$) neutron star is $\approx$ 10 km, i.e., only about twice its Schwarzschild radius.

The "mini" black hole would be a hypothetical *Planck particle* with mass $\sqrt{\pi} m_P$, for which $r_s(m) = \lambda_C(m)$ (cf. **Compton wavelength**), and radius $r_s(\sqrt{\pi} m_P) = 2\sqrt{\pi} l_P$. Cf. *planckeon* in **quantum space-time**; it should have radius $l_P = \bar{\lambda}_C(m_P)$ and mass m_P, for which $\bar{\lambda}_C(m) = \frac{1}{2} r_s(m)$.

A *quasar* (quasi-stellar radio source) is a compact region in the center of a massive galaxy surrounding its central supermassive black hole; its size is 10–10,000 times the Schwarzschild radius of the black hole.

[redacted]

The Schwarzschild radius of observable Universe is $\approx$ 10 Gly. So, Pathria–Good, 1972, then Poplawski, from 2010, proposed that the observable Universe is the interior of a black hole existing inside a larger universe, or multiverse.

- **Jeans length**

  The **Jeans length** (or *acoustic instability scale*) is (Jeans, 1902) the length scale $L_J = v_s t_g = \frac{v_s}{\sqrt{\rho G}}$ of a cloud (usually, of interstellar dust) where thermal energy causing the cloud to expand, is counteracted by self-gravity causing it to collapse. Here v_s, t_g, ρ are the speed of sound, gravitational free fall time and enclosed mass density. So, L_j is also the distance a sound wave would travel in the collapse time.

  The *Jeans mass* is the mass contained in a sphere of Jeans length diameter.

- **Acoustic metric**

  In Acoustics and Fluid Dynamics, the **acoustic metric** (or **sonic metric**) is a characteristic of sound-carrying properties of a given medium: air, water, etc.

  In General Relativity and Quantum Gravity, it is a characteristic of signal-carrying in a given *analog model* (with respect to Condensed Matter Physics) where, for example, the propagation of scalar fields in curved *space-time* is modeled (see, for example, [BLV05]) as the propagation of sound in a moving fluid, or slow light in a moving fluid dielectric, or *superfluid* (quasi-particles in quantum fluid).

  The passage of a signal through an acoustic metric modifies the metric; for example, the motion of sound in air moves air and modifies the local speed of the sound. Such "effective" (i.e., recognized by its "effects") **Lorentz metric** (Chap. 26) governs, instead of the background metric, the propagation of fluctuations: the particles associated to the perturbations follow geodesics of that metric.

  In fact, if a fluid is barotropic and inviscid, and the flow is irrotational, then the propagation of sound is described by an **acoustic metric** which depends on the density ρ of flow, velocity $\mathbf{v}$ of flow and local speed s of sound in the fluid. It can be given by the *acoustic tensor*

$$g = g(t, \mathbf{x}) = \frac{\rho}{s} \begin{pmatrix} -(s^2 - v^2) & \vdots & -\mathbf{v}^T \\ \cdots & & \cdots \\ -\mathbf{v} & \vdots & 1_3 \end{pmatrix},$$

  where 1_3 is the 3×3 identity matrix, and $v = ||\mathbf{v}||$. The *acoustic line element* is

$$ds^2 = \frac{\rho}{s}\left(-(s^2 - v^2)dt^2 - 2\mathbf{v}d\mathbf{x}dt + (d\mathbf{x})^2\right) = \frac{\rho}{s}\left(-s^2dt^2 + (d\mathbf{x} - \mathbf{v}dt)^2\right).$$

  The signature of this metric is (3, 1), i.e., it is a **Lorentz metric**. If the speed of the fluid becomes supersonic, then the sound waves will be unable to come back, i.e., there exists a *mute hole*, the acoustic analog of a *black hole*.

The **optical metrics** are also used in analog gravity and effective metric techniques; they correspond to the representation of a gravitational field by an equivalent optical medium with magnetic permittivity equal to electric one.

- **Aichelburg–Sexl metric**

In Quantum Gravity, the **Aichelburg–Sexl metric** (Aichelburg and Sexl, 1971) is a 4D metric created by a relativistic particle (having an energy of the order of the Planck mass) of momentum p along the x axis, described by its *line element*

$$ds^2 = dudv - d\rho^2 - \rho^2 d\phi^2 + 8p \ln \frac{\rho}{\rho_0} \delta(u) du^2,$$

where $u = t - x, v = t + x$ are null coordinates, ρ and ϕ are standard polar coordinates, $\rho = \sqrt{y^2 + z^2}$, and ρ_0 is an arbitrary scale constant.

This metric admits an nD generalization (de Vega and Sánchez, 1989), given by

$$ds^2 = dudv - (dX^i)^2 + f_n(\rho)\delta(u)du^2,$$

where X^i are the traverse coordinates, $\rho = \sqrt{\sum_{1 \leq i \leq n-2} (X^i)^2}, f_n(\rho) = K(\frac{\rho}{\rho_0})^{4-n}$, $k = \frac{8\pi^{2-0.5n}}{n-4}\Gamma(0.5n-1)GP, n > 4, f_4 = 8GP \ln \frac{\rho}{\rho_0}$, P is the particle's momentum.

- **Quantum space-time**

Quantum space-time is a generalization of the usual space-time in which some variables that ordinarily commute are assumed not to commute, form a different Lie algebra, and, as a result, some variables may become discrete. For example, noncommutative field theory supposes that, on sufficiently small (quantum) distances, the spatial coordinates do not commute, i.e., it is impossible to measure exactly the position of a particle with respect to more than one axis. Any noncommutative algebra with $\geq$ 4 generators could be interpreted as a quantum space-time.

At Planck scale $l_P \approx 1.6 \times 10^{-35}$ m, "quantum foam" (Wheeler,1950) is expected: violent warping and turbulence of space-time, which loses the smooth continuous structure (apparent macroscopically) of a *Riemannian manifold*, to become discrete, fractal, nondifferentiable.

Many models of *granular space* were proposed. *Planckeon* is (Markov, 1965) a hypothetical "grain of space" of size l_P and Planck rest mass m_P. In the *World Crystal model*, quantum space-time is a lattice with spacing of the order l_P, and matter creates defects generating curvature and all effects of General Relativity.

A **quantum metric** is a general term used for a metric expected to describe the space-time at quantum scales. Cf. **Rieffel metric space**, **Fubini–Study distance** (Chaps. 7, 12), **quantum graph** (Chap. 15), statistical geometry of fuzzy lumps ([ReRo01]), quantization of the **semimetric cone** (Chap. 1) in [IKP90].

Loop Quantum Gravity (LQG), *String Theory*, *Causal Sets* and *Black Hole Thermodynamics*, predict a quantum space-time at Planck scale. LQG predict,

[redacted]

moreover, that its geometry (area, volume) is quantized via **spin networks** (Chap. 15). Analyses of gamma ray bursts rule out quantum graininess at $> 10^{-48}$ m.

- **Distances between quantum states**

  A **distance between quantum states** is a metric which is preferably preserved by unitary operations, monotone under quantum operations, stable under addition of systems and having clear operational interpretation.

  The pure states correspond to the rays in the Hilbert space of wave functions. Every mixed state can be purified in a larger Hilbert space. The mixed quantum states are represented by *density operators* (i.e., positive operators of unit trace) in the complex projective space over the infinite-dimensional Hilbert space. Let X denote the set of all density operators in this Hilbert space. For two given quantum states, represented by $x, y \in X$, we mention the following main distances on X.

  The **trace distance** is a metric on density matrices defined by

$$T(x,y) = \frac{1}{2}||x-y||_{tr} = \frac{1}{2}\mathrm{Tr}\sqrt{(x-y)^*(x-y)} = \frac{1}{2}\mathrm{Tr}\sqrt{(x-y)^2} = \frac{1}{2}\sum_i |\lambda_i|,$$

  where λ_i are eigenvalues of the Hermitian matrix $x - y$. It is the maximum probability that a quantum measurement will distinguish x from y. Cf. the **trace norm metric** $||x - y||_{tr}$ in Chap. 12. When matrices x and y commute, i.e., are diagonal in the same basis, $T(x,y)$ coincides with **variational distance** in Chap. 14.

  The **quantum fidelity similarity** is defined (Jorza, 1994) by

$$F(x,y) = (\mathrm{Tr}(\sqrt{x}y\sqrt{x}))^2 = (||\sqrt{x}\sqrt{y}||_{tr})^2.$$

  When the states x and y are *classical*, i.e., they commute, $\sqrt{F(x,y)}$ is the classical **fidelity similarity** $\rho(P_1, P_2) = \sum_z \sqrt{p_1(z)p_2(z)}$ from Chap. 14.

  When x and y are pure states, $F(x,y)$ is called *transition probability* and $\sqrt{F(x,y)} = |\langle x', y'\rangle|$ (where x', y' are the unit vectors representing x, y) is called *overlap*. In general, $F(x,y)$ is the maximum overlap between purifications of x and y. Useful lower and upper bounds for $F(x,y)$ are

$$\mathrm{Tr}(xy) + \sqrt{2((\mathrm{Tr}(xy))^2 - \mathrm{Tr}(xyxy))} \quad (\textit{subfidelity}) \text{ and}$$

$$\mathrm{Tr}(xy) + \sqrt{((\mathrm{Tr}(x))^2 - \mathrm{Tr}(x^2))((\mathrm{Tr}(y))^2 - \mathrm{Tr}(y^2))} \quad (\textit{super-fidelity}).$$

  The **Bures distance** is $\sqrt{2(1-\sqrt{F(x,y)})}$. The **Bures length** (or *Bures angle*) is $\arccos\sqrt{F(x,y)}$; it is the minimal such distance between purifications of x and y. Cf. the **Bures metric** and **Fubini–Study metric** in Chap. 7. In general, the Riemannian **monotone metrics** in Chap. 7 generalize the **Fisher information**

metric on the class of probability densities (*classical* or commutative case) to the class of density matrices (*quantum* or noncommutative case). Fujiwara–Nagaoka, 1995, derived the *quantum Fisher metric*, coinciding, up to a scalar, on pure states with **Fubini–Study metric** on the associated projective Hilbert space.

The distances based on the *Shannon entropy* $H(p) = -\sum_i p_i \log p_i$ are generalized on quantum setting via the *von Neumann entropy* $S(x) = -\mathrm{Tr}(x \log x)$. Balian–Alhassid–Reinhardt, 1986 (see [Bali14]) defined a such metric by the Hessian of von Neumann entropy; cf. **Hessian metric** in Sect. 7.3.

The **sine distance** (Rastegin, 2006) is a metric defined by

$$\sin \min_{x',y'}(\arccos(|\langle x', y' \rangle|)) = \sqrt{1 - F(x,y)},$$

where x', y' are purifications of x, y. It holds $1 - \sqrt{F(x,y)} \leq T(x,y) \leq \sqrt{1 - F(x,y)}$.

Examples of other known metrics generalized to the class of density matrices are the **Hilbert–Schmidt norm metric**, **Sobolev metric** (Chap. 13) and **Monge–Kantorovich metric** (Chap. 21).

- **Action at a distance (in Physics)**

  An **action at a distance** is the interaction, without known mediator, of two objects separated in space. Einstein used the term *spooky action at a distance* for quantum mechanical interaction (like *entanglement* and *quantum nonlocality*) which is instantaneous, regardless of distance. His **principle of locality** is: distant objects cannot have direct influence on one another, an object is influenced directly only by its immediate surroundings.

  Alice–Bob distance is the distance between two entangled particles, "Alice" and "Bob". Quantum Theory predicts that the correlations based on quantum entanglement should be maintained over arbitrary Alice–Bob distances. But a *strong nonlocality*, i.e., a measurable action at a distance (a superluminal propagation of real, physical information) never was observed and is not expected. Salart et al., 2008, estimated that such signal should be 10,000 times faster than light.

  At 2012, some quantum information—the polarization property of a photon—to its mate in an entangled pair of photons, was teleported over 143 km. But such photon systems transport only half of the information. Used with continuous variable systems, they are 100 % effective but limited to short distances.

  Two-particle entanglement occurs in any temperature.

  "Mental action at a distance" (say, telepathy, clairvoyance, distant anticipation, psychokinesis) is controversial because it challenge classical concepts of time/causality as well as space/distance.

  The term *short range interaction* is used for the transmission of action at a distance by a material medium from point to point with a certain velocity dependent on properties of this medium. In Information Storage, the term *near-field interaction* is used for very short distance interaction using scanning probe techniques. *Near-field communication* is a set of standards-based technologies

[redacted]

enabling short range ($\leq$ 4 cm) wireless communication between electronic devices.

- **Macroscale entanglement/superposition**

  Quantum superposition is the addition of the amplitudes of wave-functions, occurring when an object simultaneously "possesses" two or more values for an observable quantity, say, the position or energy of a particle. If the system interacts with its environment in a *thermodynamically irreversible way* (say, the quantity is measured), then *quantum decoherence* occurs: the state randomly collapses onto one of those values. But it can happen also without any influence from the outside world.

  Superposition and *entanglement* (nonlocal correlation which cannot be described by classical communication or common causes) were observed at atomic scale. Entangling in time (a pair of photons that never existed at the same time) was observed as well. With increasing duration and size/complexity of objects, these quantum effects are lost: decoherence, due to many interactions at the molecular level, occurs. To find out this threshold, if any, is a hot research topic. Kasevich et al., 2015, got superposition on human scales: 54 cm apart for about 1 sec. Kaiser *et al.*, 2016, got it, for the oscillations of neutrinos, at 735 km.

  Nimmrichter and Hornberger, 2013 assign the *macroscopicity* μ to a quantum state, if the *equivalent* (in terms of ruling out even a minimal modification of Quantum Mechanics, which would predict a failure of the superposition principle on the macroscale) superposition state of a single electron last for 10^{τ} seconds. The record score so far is $\mu \approx 12$ and 24 looks reachable. But the Schrödinger's cat (seen as a 4-kg sphere of water) in a superposition, where it sits in two positions spaced 10 cm apart for 1 s, would score unconceivable 57.

  Szarek–Aubrun–Ye, 2013, found a threshold $k_0 \approx \frac{N}{5}$ such that two subsystems of k particles each of a system of N identical particles in a random pure state, typically share entanglement if $k > k_0$, and typically do not share it if $k < k_0$.

  "Warm" quantum coherence was observed in plant photosynthesis, animal magnetoreception, our sense of smell and microtubules inside brain neurons. Hameroff and Penrose, 2014: EEG rhythms (brain waves) and consciousness derive from quantum vibrations in microtubules, i.e., on the quantum-realm scale ($\sim$ 100 nm) rather than, or in addition to, the larger scale of neurons (4–100 μm).

  Schmied et al., 2016, got a special case of entanglement—*Bell correlation* (correlations between the spins)—of about 480 atoms in a Bose-Einstein condensate.

- **Entanglement distance**

  The **entanglement distance** is the maximal distance between two entangled electrons in a *degenerate* electron gas beyond which all entanglement is observed to vanish. *Degenerate matter* (say, a white dwarf star) is matter having so high density that the main contribution to its pressure arises from the *Pauli exclusion principle*: no two identical fermions may occupy the same quantum state together.

- **Tunneling distance**

  Quantum Tunneling is the quantum mechanical phenomenon where a particle tunnels through a barrier that it classically could not surmount.

  For example, in *STM* (Scanning Tunneling Microscope), electron tunneling current and a net electric current from a metal tip of STM to a conducting surface result from overlap of electron wavefunctions of tip and sample, if they are brought close enough together and an electric voltage is applied between them.

  The tip-sample current depends exponentially (about $exp(-d^{0.5})$) on their distance d, called **tunneling distance**. Formally, d is the sum of the radii of the electron delocalization regions in the donor and the acceptor atoms.

  By keeping the current constant while scanning the tip over the surface and measuring its height, the contours of the surface can be mapped out. The tunneling distance is longer ($<$ 1 nm) in aqueous solution than in vacuum ($<$ 0.3 nm).

24.2 Distances in Chemistry and Crystallography

Main chemical substances are ionic (held together by ionic bonds), metallic (giant close packed structures held together by metallic bonds), giant covalent (as diamond and graphite), or molecular (small covalent). Molecules are made of a fixed number of atoms joined together by covalent bonds; they range from small (single-atom molecules in the noble gases) to very large ones (as in polymers, proteins or DNA).

The largest known (55 tons and $12, 4$ m in diameter) crystal is a selenite found in Naica Mine, Mexico. The largest stable synthetic molecule is PG_5 with a diameter of 10 nm and a mass equal to 2×10^8 hydrogen atoms.

The **interatomic distance** of two atoms is the distance (in angstroms or picometers, where $1\,\text{Å} = 10^{-10}$ m $=$ 10 pm) between their nuclei. The bond between helium atoms in molecules He_2 is the longest (54.6 Å) and weakest known; it is 0.75 Å in H_2. The *$C - C$ distance* (distance between two carbon atoms) in graphene is 1.42 Å. The *C-length* (number of carbon atoms) is *carbyne* (chain of carbon atoms) can reach 6400.

- **Atomic radius**

  Quantum Mechanics implies that an atom is not a ball having an exactly defined boundary. Hence, **atomic radius** is defined as the distance from the atomic nucleus to the outermost stable electron orbital in a atom that is at equilibrium. Atomic radii represent the sizes of isolated, electrically neutral atoms, unaffected by bonding.

  Atomic radii are estimated from **bond distances** if the atoms of the element form bonds; otherwise (like the noble gases), only **van der Waals radii** are used.

  The atomic radii of elements increase as one moves down the column (or to the left) in the Periodic Table of Elements. internuclear distance, Re is the equilibrium internuclear distance (bond length)

[redacted]

- **Bond distance**

  The **bond distance** (or *bond length*) is the equilibrium *internuclear distance* of two bonded atoms. For example, typical bond distances for carbon-carbon bonds in an organic molecule are 0.15, 0.13 and 0.12 pm (picometers 10^{-9} m) for single, double and triple bonds, respectively. The atomic nuclei repel each other; the **equilibrium distance** between two atoms in a molecule is the internuclear distance at the minimum of the electronic (or potential) energy surface.

  Depending on the type of bonding of the element, its atomic radius is called *covalent* or *metallic*. The *metallic radius* is one half of the **metallic distance**, i.e., the closest internuclear distance in a *metallic crystal* (lattice of metallic element).

  Covalent radii of atoms of elements that form covalent bonds are inferred from bond distances between pairs of covalently-bonded atoms. If the two atoms are of the same kind, then their covalent radius is one half of their bond distance. Covalent radii for other elements is inferred by combining the radii of those that bond with bond distances between pairs of atoms of different kind.

- **van der Waals contact distance**

  Intermolecular distance data are interpreted by viewing atoms as hard spheres. The spheres of two neighboring nonbonded atoms (in touching molecules or atoms) are supposed to just touch. So, their interatomic distance, called the **van der Waals contact distance**, is the sum of radii, called **van der Waals radii** (of *effective sizes*), of their hard spheres.

  The van der Waals contact distance corresponds to a "weak bond", when repulsion forces of electronic shells exceed London (attractive electrostatic) forces.

- **Molecular RMS radius**

  The **molecular RMS radius** (cf. **radius of gyration** in Sect. 24.1) is the root-mean-square distance of a molecule's atoms from their common center of gravity:

  $$\sqrt{\frac{\sum_{1\le i\le n} d_{0i}^2}{n+1}} = \sqrt{\frac{\sum_i \sum_j d_{ij}^2}{(n+1)^2}},$$

  where n is the number of atoms in the molecule, d_{0i} is the Euclidean distance of the i-th atom from the center of gravity of the molecule (in a specified conformation), and d_{ij} is the Euclidean distance between the i-th and j-th atoms.

  The **mean molecular radius** is the number $\frac{\sum_i r_i}{n}$, where n is the number of atoms, and r_i is the Euclidean distance of the i-th atom from the centroid $\frac{\sum_j x_{ij}}{n}$ of the molecule (here x_{ij} is the i-th Cartesian coordinate of the j-th atom).

- **Molecular sizes**

  There are various descriptions of the **molecular sizes**; examples as follows.

  The *kinetic diameter* of a molecule (most applicable to transport phenomena) is its smallest effective dimension.

  The **effective diameter** of a molecule is the general extent of the electron cloud surrounding it as calculated in any of several ways.

Sometimes, it is defined as diameter of the sphere containing 98 % of the total electron density; then its half is close to the experimental **van der Waals radius**.

The **effective molecular radius** is the size a molecule displays in solution. For liquids and solids it is usually defined via packing density.

For a gas, molecular sizes can be estimated from the intermolecular separation, speed, mean free path and collision rate of gas molecules.

For example, in the model of kinetic theory of gases, assuming that molecules interact like hard spheres, the **molecular diameter** d is $\sqrt{\frac{m}{\pi\sqrt{2}\rho}}$, where m is the mass of molecule, l is mean free path and ρ is density.

- **Range of molecular forces**

  Molecular forces (or interactions) are the following electromagnetic forces: ionic bonds (charges), hydrogen bonds (dipolar), dipole-dipole interactions, London forces (the attraction part of van der Waals forces) and steric repulsion (the repulsion part of van der Waals forces). If the distance (between two molecules or atoms) is d, then (experimental observation) the potential energy function P relates inversely to d^n with $n = 1, 3, 3, 6, 12$ for the above five forces, respectively.

  The **range** (or the *radius*) of an interaction is considered *short* if P approaches 0 rapidly as d increases. It is also called *short* if it is at most 3 Å; so, only the range of steric repulsion is short (cf. **range of fundamental forces**).

  An example: for polyelectrolyte solutions, the long range ionic solvent-water force competes with the shorter range water-water (hydrogen bonding) force.

  In protein molecule, the range of London van der Waals force is $\approx$ 5 Å, and the range of hydrophobic effect is up to 12 Å, while the length of hydrogen bond is $\approx$ 3 Å, and the length of *peptide bond* is $\approx$ 1.5 Å.

- **Chemical distance**

  Various chemical systems (single molecules, their fragments, crystals, polymers, clusters) are well represented by graphs where vertices (say, atoms, molecules acting as monomers, molecular fragments) are linked by, say, chemical bonding, van der Waals interactions, hydrogen bonding, reactions path.

  In Organic Chemistry, a *molecular graph* $G = (V, E)$ is a graph representing a given molecule, so that the vertices $v \in V$ are atoms and the edges $e \in E$ correspond to electron pair bonds. The *Wiener number* of a molecule is one half of the sum of all pairwise distances between vertices of its molecular graph.The *Wiener polarity index* is the number of unordered vertex pairs at distance 3 in this graph. Cf. **Wiener-like distance indices** in Chap. 1.

  The (bonds and electrons) *BE-matrix* of a molecule is the $|V| \times |V|$ matrix $((e_{ij}))$, where e_{ii} is the number of free unshared valence electrons of the atom A_i and, for $i \neq j$, $e_{ij} = e_{ji} = 1$ if there is a bond between atoms A_i and A_j, and $= 0$, otherwise.

  Given two *stoichiometric* (i.e., with the same number of atoms) molecules x and y, their **Dugundji–Ugi chemical distance** is the **Hamming metric**

$$\sum_{1 \leq i,j \leq |V|} |e_{ij}(x) - e_{ij}(y)|,$$

[redacted]

and their **Pospichal–Kvasnička chemical distance** is

$$\min_{\pi} \sum_{1 \le i,j \le |V|} |e_{ij}(x) - e_{\pi(i)\pi(j)}(y)|,$$

where π is any permutation of the atoms. The above distance is equal to $|E(x)| + |E(y)| - 2|E(x,y)|$, where $E(x,y)$ is the edge-set of the maximum common subgraph of the molecular graphs $G(x)$ and $G(y)$. Cf. **Zelinka distance** in Chap. 15.

The **Pospichal–Kvasnička reaction distance**, assigned to a molecular transformation $x \to y$, is the minimum number of *elementary transformations* needed to transform $G(x)$ onto $G(y)$.

- **Molecular similarities**

  Given two 3D molecules x and y characterized by some structural (shape or electronic) property P, their similarities are called **molecular similarities**.

  The main electronic similarities correspond to some correlation similarities from Chap. 17. For example, the *Carbó similarity* (Carbó–Leyda–Arnau, 1980) is the **cosine similarity** (Chap. 17) defined by

$$\frac{\langle f(x), f(y)\rangle}{||f(x)||_2 \cdot ||f(y)||_2},$$

  where the *electron density function* $f(z)$ of a molecule z is the volumic integral $\int P(z)dv$ over the whole space.

  The *Hodgkin–Richards similarity* (1991) is defined (cf. the **Morisita–Horn similarity** in Chap. 17) by

$$\frac{2\langle f(x), f(y)\rangle}{||f(x)||_2^2 + ||f(y)||_2^2},$$

  where $f(z)$ is the electrostatic potential or electrostatic field of a molecule z.

  Petitjean, 1995, proposed to use the distance $V(x \cup y) - V(x \cap y)$, where the *volume* $V(z)$ of a molecule z is the union of *van der Waals spheres* of its atoms. Cf. **van der Waals contact distance** and, in Chap. 9, **Nikodym metric** $V(x \Delta y)$.

- **End-to-end distance**

  A *polymer* is a large macromolecule composed of repeating structural units connected by covalent chemical bonds.

  For a coiled polymer, the **end-to-end distance** (or *displacement length*) is the distance between the ends of the polymer chain. The maximal possible such distance (i.e., when the polymer is stretched out) is called *contour length*.

  The root-mean-square end-to-end distance of ideal linear or randomly branched polymer scales as $n^{0.5}$ or, respectively, $n^{0.25}$ if n is the number of monomers. For a polymer chain following a random walk in 3D, it is also 6 times **molecular RMS radius**. The **strand length** in Chap. 23 is the end-to-end distance for a special linear polymer, single-stranded RNA or DNA.

- **Persistence length**

  The **persistence length** of a polymer chain is the length over which correlations in the direction of the tangent are lost.

  The molecule behaves as a flexible elastic rod for shorter segments, while for much longer ones it can only be described statistically, like a 3D random walk. Cf. **correlation length**.

  Twice the persistence length is the *Kuhn length*, i.e., the length of hypothetical segments which can be thought of as if they are freely jointed with each other in order to form given polymer chain.

- **Bend radius**

  In Polymer Tubing, the **bend radius** of a tube is the distance from the center of an imaginary circle on which the arc of the bent tube falls to a point on that arc.

- **Intermicellar distance**

  Micelle is an electrically charged particle built up from polymeric molecules or ions and occurring in certain colloidal electrolytic solutions like soaps and detergents. This term is also used for a submicroscopic aggregation of molecules, such as a droplet in a colloidal system, and for a coherent strand or structure in a fiber.

  The **intermicellar distance** is the average distance between micelles.

- **Interionic distance**

  An *ion* is an atom that has a positive or negative electrical charge. The **interionic distance** is the distance between the centers of two adjacent (bonded) ions. **Ionic radii** are inferred from ionic bond distances in real molecules and crystals.

  The ion radii of *cations* (positive ions, for example, sodium Na^+) are smaller than the atomic radii of the atoms they come from, while *anions* (negative ions, for example, chlorine Cl^-) are larger than their atoms.

- **Repeat distance**

  Given a periodic layered structure, its **repeat distance** is the period, i.e., the spacing between layers (say, lattice planes, bilayers in a liquid-crystal system, or graphite sheets along the unit cell's hexagonal axis).

  A crystal lattice, the unit cell in it and cell spacing are called also a *repeat pattern*, *basic repeat unit* and *cell repeat distance* (or *lattice spacing*, *interplaner distance*).

  The *repeat distance* in a polymer is the ratio of the unit cell length along its axis of propagation to the number of monomeric units this length covers.

- **Metric symmetry**

  The full crystal symmetry is given by its *space group*.

  The **metric symmetry** of the crystal lattice is its symmetry without taking into account the arrangement of the atoms in the unit cell.

  In between lies the *Laue group* giving equivalence of different reflections, i.e., the symmetry of the crystal diffraction pattern. In other words, it is the symmetry in the *reciprocal space* (taking into account the reflex intensities).

[redacted]

The Laue symmetry can be lower than the metric symmetry (for example, an orthorhombic unit cell with $a = b$ is metrically tetragonal) but never higher.

There are seven *crystal systems*—*triclinic*, *monoclinic*, *orthorhombic*, *tetragonal*, trigonal, *hexagonal*, and *cubic* (or *isometric*). Taken together with possible *lattice centerings*, there are 14 *Bravais lattices*.

- **Homometric structures**

  Two structures of identical atoms are **homometric** if they are characterized by the same multiset of interatomic distances; cf. **distance list** in Chap. 1.

  Homometric crystal structures produce identical X-ray diffraction patterns.

  In Music, two rhythms with the same multiset of intervals are called **homometric**.

- **Dislocation distances**

  In Crystallography, a *dislocation* is a defect extending through a crystal for some distance (**dislocation path length**) along a *dislocation line*. It either forms a complete loop within the crystal or ends at a surface or other dislocation.

  The **mean free path** of a dislocation is (Gao et al., 2007), in 2D, the average distance between its origin and the nearest particle or, in 3D, the maximum radius of a dislocation loop before it reaches a particle in the slip plane.

  The **pinning distance** is the distance between two endpoints of a mobile dislocation, where one of the endpoints has to be within the volume. It is a characteristic length for the dislocation microstructure.

  The *Burgers vector* of a dislocation is a crystal vector denoting the direction and magnitude of the atomic displacement that occurs within a crystal when a dislocation moves through the lattice. A dislocation is called *edge*, *screw* or *mixed* if the angle between its line vector and the Burgers vector is 90°, 0° or otherwise, respectively. The **edge dislocation width** is the distance over which the magnitude of the displacement of the atoms from their perfect crystal position is greater than $\frac{1}{4}$ of the magnitude of the Burgers vector.

  The *dislocation density* ρ is the total length of dislocation lines per unit volume; typically, it is 10 km per cm^3 but can reach 10^6 km per cm^3 in a heavily deformed metal. The **average distance** between dislocations depends on their arrangement; it is $\rho^{-\frac{1}{2}}$ for a quadratic array of parallel dislocations. If the average distance decreases, dislocations start to cancel each other's motion.

  The **spacing dislocation distance** is the minimum distance between two dislocations which can coexist on separate planes without recombining spontaneously.

- **Dynamical diffraction distances**

  Diffraction is the apparent bending of propagating waves around obstacles of about the wavelength size. Diffraction from a 3D periodic structure such as an atomic crystal is called *Bragg diffraction*. It is a convolution of the simultaneous scattering of the probe beam (light as X-rays, or matter waves such as electrons or neutrons) by the sample and interference (superposition of reflections from crystal planes). Some materials, lacking 3D translational periodicity, still have aspects of **long range order**, that give rise to sharp Bragg reflections in their

X-ray diffraction patterns. Such distance doubling between layers of a material is called *c-axis periodicity*.

The *Bragg Law*, modeling diffraction as reflection from crystal planes of atoms, states that waves (with wavelength λ scattered under angle θ from planes at spacing d) interfere only if they remain in phase, i.e., $\frac{2d \sin \theta}{\lambda}$ is an integer.

The decay of intensity with depth traversed in the crystal occurs by *dynamical extinction*, redistributing energy within the wave field, and by *photoelectric absorption* (a loss of energy from the wave field to the atoms of the crystal).

The former *kinematic* theory works for imperfect crystals and estimates absorption. The *dynamical* (multiple diffraction) theory is used to model the *perfect* (no disruptions in the periodicity) crystals. It considers the incident and diffracted wave fronts as coupled/interacting parts of a wave field and the periodically varying electrical susceptibility of the medium so as to satisfy the Maxwell equations.

Dynamical theory distinguishes two cases: Laue (or *transmission*) and Bragg (or *reflection*) case, when the reflected wave is directed toward the inside and outside of the crystal. The wave field is represented by its *dispersion surface*. The inverse of the diameter of this surface is called (Autier, 2001) the **Pendellösung distance** Λ_L in the Laue case and the **extinction distance** Λ_B in the Bragg case.

At the exit face of the crystal, the wave splits into two single waves with different directions: incident 0-beam and diffracted H-beam. With increasing thickness of the crystal, the wave leaving it will first appear mainly in the 0-beam, then entirely in the H-beam at thickness $\frac{\Lambda_L}{2}$, and then it will oscillate between these beams with a period Λ_L, called the **Pendellösung length**; cf. similar **coupling length**.

The wave amplitude (and the intensity of the diffracted beam) is transferred back-and-forth once, i.e., the physical distance acquires a phase change of 2π. Pendellösung oscillations happen also in Bragg case, but with very rapidly decaying amplitudes, and *Pendellösung fringes* are visible only for θ close to 0^0 or 45^0.

Diffraction that involves multiple scattering events is called *extinction* since it reduces the observed integrated diffracted intensity. Extinction is very significant for perfect crystals and is then called *primary extinction*. In the Bragg case, the **primary extinction length** (James, 1964) is the inverse of the *extinction factor* (maximum extinction coefficient for the middle of the range of total reflection):

$$\frac{\pi V \cos \theta}{\lambda r_e |F| C},$$

where F, C (valued 1 or $\cos 2\theta$) are the structure and polarization factors, V is the volume of unit cell, $r_e \approx 2.81794 \times 10^{-15}$m is the *classical electron radius* and λ is the X-ray beam wavelength. The diffracted intensity with sufficiently large thickness no longer increases significantly with increased thickness.

The **extinction length** of an electron or neutron diffraction is $\frac{\pi V \cos \theta}{\lambda |F|}$. Half of it gives the number of atom planes needed to reduce the beam to 0 intensity.

[redacted]

The **X-ray penetration depth** (or *attenuation length*, *mean free path*, *extinction distance*) is (Wolfstieg, 1976) the depth into the material where the intensity of the diffracted beam has decreased e-fold. Cf. **penetration depth**.

In Gullity, 1956, *X-ray penetration depth* is the depth z such that $\frac{I_z}{I_\infty} = 1 - \frac{1}{e}$, where I_∞, I_z are the total diffraction intensities given from the whole specimen and, respectively, the range between the surface and the depth, z, from it.

- **X-ray absorption length**

The *absorption edge* is a sharp discontinuity in the absorption spectrum of X-rays by an element that occurs when the energy of the photon is just above the **binding energy** of an electron in a specific shell of the atom.

The **X-ray absorption length** of a crystal is the thickness s of the sample such that the intensity of the X-rays incident upon it at an energy 50 eV above the absorption edge is attenuated e-fold.

For an X-ray laser, the *extinction length* is the thickness needed to fully reflect the beam; usually, it is a few microns while the absorption length is much larger.

In Segmüller, 1968, the *absorption length* is $\frac{\sin\theta}{\mu}$, where μ is the linear absorption coefficient, and the beam enters the crystal at an angle θ.

- **Diamond-cutting distances**

Diamond is the hardest natural gem and the only gemstone composed of a single element—carbon. Diamond takes a fine polish, which makes its surfaces highly reflective. Color in diamond (the rarest being red) is caused by structural irregularities, or trace elements. Diamonds are graded according to carat weight, clarity, color and cut. Diamonds are cut to maximize the play of light within the stone. Their beauty comes from a combination of *fire* (rainbow flash from within) and *brilliance* (burst of sparkling light). Both are a direct result of the cut.

A faceted stone can be divided into an upper (*crown*) and lower (*pavilion*) section. The perimeter, where both parts meet, is referred to as the *girdle*. The **depth of a gemstone** is measured from the *table* (highest crown facet) to the *culet* (tip of the pavilion). On a round brilliant diamond, the **depth percentage** represents the ratio of the table-culet distance to the average girdle diameter.

Normally, the table is the largest surface on a gemstone. On a round brilliant-cut diamond it forms an octagon, but some cutting styles do not have a table. The **table percentage** of a diamond represents the ratio of table width to overall stone width. A beautiful, well-cut stone will normally have a table percentage $53-$ $-64\,\%$. A stone's *luster* (appearance of the surface dependent upon its reflecting qualities) is directly affected by its depth and table percentages.

Chapter 25
Distances in Earth Science and Astronomy

25.1 Distances in Geography

- **Spatial scale**

  In Geography, **spatial scales** are shorthand terms for distances, sizes and areas. For example, micro, meso, macro, mega may refer to local (0.001–1), regional (1–100), continental (100–10,000), global (>10,000) km, respectively.
- **Earth radii**

  The Earth's maximal and minimal *radii* (the center-surface distances) are 6384 km (the Chimborazo's summit) and 6353 km (the Arctic Ocean's floor). An object, moved from the 2-nd spot to the 1-st, will loose $\approx 1\,\%$ of its weight.

  In the ellipsoidal model, the Earth's *equatorial radius* (*semimajor axis*) a, is 6378 km and the *polar radius* (*semiminor axis*) b, is 6357 km. The equatorial and polar *radii of curvature* are $\frac{b^2}{a}$ and $\frac{a^2}{b}$. The *mean radius* is $\frac{2a+b}{3} = 6371$ km.

  The Earth's *authalic* and *volumetric radius* (the radii of the spheres with the same surface area and volume, respectively, as the Earth's ellipsoid) are 6371 and 6371 km; cf. the **characteristic diameters** in Chap. 29.

  In Telecommunications, the *effective Earth radius* is the radius of a sphere for which the distance to the radio horizon, assuming rectilinear propagation, is the same as that for the Earth with an assumed uniform vertical gradient of atmospheric refractive index. For the standard atmosphere, this radius is $\frac{4}{3}$ that of the Earth.
- **Great circle distance**

  The **great circle distance** (or **orthodromic distance**, *air line*) is the shortest distance between points x and y on the Earth's surface measured along a path on this surface. It is the length of the *great circle* arc, passing through x and y, in the spherical model of the planet. Cf. **spherical metric** in Chap. 6.

  Let δ_1, ϕ_1 be the latitude and the longitude of x, and δ_2, ϕ_2 be those of y; let r be the Earth's radius. Here $2r^2 = a^2 + b^2$, where a and b are the equatorial and

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_25

[redacted]

polar radii of the Earth. Then the great circle distance is equal to

$$r \arccos(\sin \delta_1 \sin \delta_2 + \cos \delta_1 \cos \delta_2 \cos(\phi_1 - \phi_2)).$$

In the spherical coordinates (θ, ϕ), where ϕ is the azimuthal angle and θ is the colatitude, the great circle distance between $x = (\theta_1, \phi_1)$ and $y = (\theta_2, \phi_2)$ is

$$r \arccos\left(\cos \theta_1 \cos \theta_2 + \sin \theta_1 \sin \theta_2 \cos(\phi_1 - \phi_2)\right).$$

For $\phi_1 = \phi_2$, the formula above reduces to $r|\theta_1 - \theta_2|$.

The **tunnel distance** between points x and y is the length of the line segment through 3D space connecting them. For a spherical Earth, this line is the chord of the great circle between the points.

The Earth resembles a flattened spheroid with extreme values for the radius of curvature of 6335.4 km at the equator and 6399.6 km at the poles. The *spheroidal distance* between points x and y is their distance in this spheroidal model.

The *geoid* (the shape the Earth would have if it was entirely covered by water and influenced by gravity alone) looks like a lumpy potato; cf. **potato radius**.

- **Loxodromic distance**

  A *loxodrome* (or, *rhumb line*) is a curve on the Earth's surface that crosses each meridian at the same angle. It is the path taken by a ship or plane maintaining a constant compass direction; it is a straight line on the Mercator projection.

  The **loxodromic distance** is a distance between two points on the Earth's surface on the rhumb line joining them. It is never shorter than the great circle distance.

  The **nautical distance** is the length in nautical miles of the rhumb line joining any two places on the Earth's surface. One nautical mile is equal to 1852 m.

- **Continental shelf distance**

  Article 76 of the United Nations Convention on the Law of the Sea (1999) defined the *continental shelf of a coastal state* (its sovereignty domain) as the seabed and subsoil of the submarine areas that extend beyond its *territorial sea* as the natural prolongation of its land territory to the outer edge of the continental margin. It postulated that the **continental shelf distance**, i.e., the **range distance** from the baselines from which the breadth of the territorial sea is measured to above the other edge, should be within 200–350 nautical miles (370–648 km), and gave rules of its (almost) exact determination.

  Territorial sea is a belt of coastal waters extending at most 22 km. Next 370 km form the *exclusive economic zone*; first 22 km of it form the *contiguous zone*.

  Example of arising problems: Canada, Denmark and Russia, all claim a large portion of the Arctic, including North Pole, arguing that underwater Lomonosov Ridge belongs to *their* extended continental shelf.

  Article 47 of the same convention postulated that, for an archipelagic state, the ratio of the area of its waters (sovereignty domain) to the area of its land, including atolls, should be between 1 to 1 and 9 to 1, and elaborated case-by-case rules.

There is no defined bottom underground and upper airspace limit for sovereignty. The international waters/seabed and celestial bodies are the common heritage of mankind for the signatories of the Law of the Sea and Outer Space (1967) treaties.

Among divided islands, the largest one (785,753 km^2) is New Guinea and the smallest one is Koiluoto (200 m × 110 m), shared by Finland and Russia.

- **Port-to-port distance**

  The **port-to-port distance** is the shortest great circle distance between two ports that does not intersect any land contours.

  Officially published **distance between ports** represent the shortest navigable route or longer routes using favorable currents and/or avoiding some dangers to navigation. Reciprocal distances between two ports may differ.

- **Airway distance**

  An *airway* is a designated route in the air. Low altitude (or *victor*) airways are those below 5500 m AMSL (above mean sea level). High altitudes (or *jet*) airways are those above 5500 m AMSL. **Airway distance** is the actual (as opposed to straight line) distance flown by the aircraft between two points, after deviations required by air traffic control and navigation along published routes.

  The *stage length* is the distance of a nonstop leg of an itinerary. *Radar altitude* is the height with respect to the terrain below.

- **Point-to-point transit**

  Point-to-point transit is a route structure (common among low-fare airlines) where a plane, bus or train travels directly to a destination, rather than going through a central hub as in a *spoke-hub network*.

  A **point-to-point telecommunication** is a connection restricted to two endpoints as opposed to a *point-to-multipoint link* used in *hub and switch circuits*; cf. **flower-shop metric** in Chap. 19.

- **Lighthouse distance**

  The **lighthouse distance** is the distance from which the light from the lighthouse is first seen from of a sailboat. This distance (in feet) is $\approx 1.17(\sqrt{h_e}+\sqrt{h_l})$, where h_l is the lighthouse's height above tide level and h_e is the observer's eye level above sea. For $h_l = 0$, it estimates the **distance to horizon**.

- **Optical horizon**

  Optical (or, say, *neutrino*, *gravitational wave*) **horizon** is the farthest distance that any photon (respectively, neitrino and gravitational wave) can freely streem.

- **Distance to horizon**

  The *horizon* is the locus of points at which line of sight is tangent to the surface of the planet. At a height h above the surface of a spherical planet of radius R without atmosphere, the line-of-sight distance to the horizon is $d = \sqrt{(R+h)^2 - R^2}$, and the arc length distance to it along the curved planet's surface is $R\cos^{-1}(\frac{R}{R+h})$.

  Taking the equatorial radius 6378 km of the Earth as a typical value, gives $d \approx 357\sqrt{h}$ m for small $\frac{h}{R}$. Allowing for refraction, gives roughly $d \approx 386\sqrt{h}$ m.

  The *middle distance* is halfway between the observer and the horizon.

[redacted]

- **Radio distances**

  Marconi's law, 1897, claims that the maximum signalling distance of an antenna in meters is cH^2, where H is antenna's height and c is a constant.

  The *electrical length* is the length of a transmission medium or antenna element expressed as the number of wavelengths of the signal propagating in the medium. In coaxial cables and optical fibers, it is $\approx$ 1.5 times the physical length.

  The **electrical distance** is the distance between two points, expressed in terms of the duration of travel of an electromagnetic wave in free space between the two points. The *light microsecond*, $\approx$ 300 m, is a convenient unit of electrical distance.

  The main modes of electromagnetic wave (radio, light, X-rays, etc.) propagation are *direct wave* (line-of-sight), *surface wave* (interacting with the Earth's surface and following its curvature) and *skywave* (relying on refraction in the ionosphere).

  The **line-of-sight distance** is the distance which radio signals travel, from one antenna to another, by a *line-of-sight path*, where both antennas are visible to one another, and there are no metallic obstructions.

  The *radio horizon* is the locus of points in telecommunications at which direct rays from an antenna are tangential to the surface of the Earth. The **horizon distance** is the distance on the Earth's surface reached by a direct wave; due to ionospheric refraction or tropospheric events, it is sometimes greater than the distance to the visible horizon. In television, the **horizon distance** is the distance of the farthest point on the Earth's surface visible from a transmitting antenna.

  The **skip distance** is the shortest distance that permits a radio signal (of given frequency) to travel as a skywave from the transmitter to the receiver by reflection (hop) in the ionosphere.

  If two radio frequencies are used (for instance, 12.5 kHz and 25 kHz in maritime communication), the **interoperability distance** and **adjacent channel separation distance** are the range within which all receivers work with all transmitters and, respectively, the minimal distance which should separate adjacent tunes for narrow-band transmitters and wide-band receivers, in order to avoid interference.

  DX is amateur radio slang (and Morse code) for distance; **DXing** is a distant radio exchange (amplifiers required). Specifically, *DX* can mean **distance unknown**, short for DXing and a far-away station that is hard to hear.

  Radio waves need 138 ms to go round the world and $\approx$ 2.57 s to be reflected from the Moon. **Long delayed echoes** (LDEs) are radio echoes which return to the sender later than 2.7 s after transmission; it is a rare and not explained phenomenon.

- **Ground sample distance**

  In Remote Sensing of the surfaces of terrestrial objects of the Solar System, including the Earth, the **ground sample distance** (or *GSD*, *ground sampling distance*, *ground-projected sample interval*) is the spacing of areas represented by each pixel in a digital photo of the ground from air or space.

For example, in an image with GSD 22 m, provided by UK-DMC2 (a British Earth imaging satellite), each pixel represents a ground area of 22^2 m^2.

- **Map's distance**

  The **map's distance** is the distance between two points on the map (not to be confused with **map distance** from Chap. 23). The length of a curved line feature on a map is usually measured by an *opisometer* (or *curvimeter*).

  The **horizontal distance** is determined by multiplying the map's distance by the numerical scale of the map.

  Map resolution is the size of the smallest feature that can be represented on a surface; more generally, it is the accuracy at which the location and shape of map features can be depicted for a given map scale.

- **Equidistant map**

  An **equidistant map** is a map projection of Earth having a well-defined nontrivial set of *standard lines*, i.e., lines (straight or not) with constant scale and length proportional to corresponding lines on the Earth. Some examples are:

  Sanson–Flamsteed equatorial map: all parallels are straight lines;

  cylindric equidistant map: the vertical lines and equator are straight lines;

  an *azimuthal equidistant map* preserves distances along any line through the central point; a *Werner cordiform map* preserves, moreover, distances along any arc centered at that point.

  Maurer–Close (or *doubly equidistant*) *map* preserves distances from two central points. If those points are identical, the map is azimuthal equidistant.

  A *gnomonic map* displays all great circles as straight lines; so, it preserves the shortest route between two locations.

- **Distance cartogram**

  A **distance** (or *linear*) **cartogram** is a diagram or abstract map in which distances are distorted proportionally to the value of some thematic variable. Typically, it shows the relative travel times and directions from vertices in a network.

- **Tolerance distance**

  In GIS (computer-based Geographic Information System), the **tolerance distance** is the maximal distance between points which must be established so that gaps and overshoots can be corrected (lines snapped together) as long as they fall within it.

- **Space syntax**

  Space syntax is a set of theories and techniques (cf. Hiller–Hanson, 1984) for the analysis of spatial configurations complementing Transport Engineering and geographic accessibility analysis in a GIS (Geographical Information System).

  It breaks down space into components, analyzed as networks of choices, and then represents it by maps and graphs describing the relative connectivity and integration of parts. The basic notions of space syntax are, for a given space:

  isovist (or visibility polygon), i.e., the field of view from any fixed point;

  axial line, i.e., the longest line of sight and access through open space;

  convex space, i.e., the maximal inscribed convex polygon (all points within it are visible to all other points within it).

[redacted]

These components are used to quantify how easily a space is navigable, for the design of settings where way-finding is important such as museums, airports, hospitals. Space syntax has also been applied to predict the correlation between spatial layouts and social effects such as crime, traffic flow, sales per unit area, etc.

- **Defensible space**

  In landscape use, **defensible space** refers to the 30 m zone surrounding a structure that has been maintained/designed to reduce fire danger. The first 9 m ($\approx$ 30 feet) is where vegetation is kept to a minimum combustible mass. The remaining area 9–30 m is the *reduced fuel zone*, where fuels and vegetation should be separated (by thinning, pruning, etc.) vertically and/or horizontally.

- **Sanitation distances**

  The **drinking distance** of a dwelling is its distance from the closest source of water.

  A *latrine* is a communal facility containing (usually many) toilets. It should be at most 50 m away from dwellings to be served and at least 50 m away from communal food-storage and preparation. A latrine should be $\geq$ 30 m from water-storage and treatment facilities, as well as from surface water and shallow groundwater sources. A septic tank should be $\geq$ 15 m from a water supply well.

  The **vertical separation distance** is the distance between the bottom of the drain field of a sewage septic system and the underlying water table. This separation distance allows pathogens (disease-causing bacteria, viruses, or protozoa) in the effluent to be removed by the soil before it comes in contact with the groundwater.

- **Setback distance**

  In land use, a **setback** (or *buffer*) **distance** is the minimum horizontal distance at which a building or other structure must legally be from property lines, or the street, or a watercourse, or any other place which needs protection. Setbacks may also allow for public utilities to access the buildings, and for access to utility meters. Cf. also **buffer distance** and **clearance distance** in Chap. 29.

- **Shy distance**

  Shy distance is the space left between vehicles (or pedestrians and vehicles) as they pass each other.

- **Distance-based numbering**

  The **distance-based exit number** is a number assigned to a road junction, usually an exit from a freeway, expressing in miles (or km) the distance from the beginning of the highway to the exit. A *milestone* (or *kilometer sign*) is one of a series of numbered markers placed along a road at regular intervals.

  The *Kilometer Zero* (or *km* 0) is a particular location (usually in the nation's capital city), from which distances are traditionally measured. For France, it is on the square facing the entrance of Notre Dame cathedral in Paris. For Russia, it is in a short passage connecting Red Square with Manege Square in Moscow.

  Distance-based house addressing is the system when buildings and blocks are numbered according to the distance, i.e., the number of increments (feet or division of miles), from a given baseline. For example, the number $67W430$

in Naperville, US, can express that the house is 67 miles west of downtown Chicago. One of the GIS-inspired guidelines: to use the address $n = \frac{d}{10} + 100$, where d is the distance in feet of the house from the reference point; roughly, $d = \frac{n}{500}$ miles.

Metes and bounds is a traditional system of land description (in Real Estate and town boundary determination) by courses and distances. *Metes* is a boundary defined by the measurement of each straight run specified as **displacement**, i.e., by the distance and direction. *Bounds* refers to a general boundary description in terms of local geography (along some watercourse, public road, wall, etc.). The boundaries are described in a running prose style, all the way around the parcel of land in sequence.

Surveying is the technique of determining the terrestrial and spatial position of points and the distances and angles between them; cf., for example, *Surveyor's Chain measures* among **imperial length measures** in Chap. 27.

- **Driveway distances**

  A *driveway* is a private road giving access from a public way. The main **driveway distances** follow.

  The *throat length* is the distance between the street and the end of the driveway inside the land development. It should be 200–250 feet (about 61–76 m or 15 car lengths) for shopping centers and 25–28, 9–15 m for small developments with or without signalized access.

  The optimal one-way *driveway width* is 4.5–5 m. Driveways entering a roadway at traffic signals should have two outbound lanes (for right and left turns) at least 7 m and an inbound lane at least 4.5 m wide. The normal width of residential driveways is 4.5–7.5 m.

  The *turn radius* is the extent that the edge of a commercial driveway is “rounded” to permit easier entry/exit by turning vehicles. In urban settings, it is 8–15 m.

- **Road sight distances**

  In Transportation Engineering, the *normal visual acuity* is the ability of a person to recognize a letter (or an object) of size 25 mm from a distance of 12 m.

  The **visibility distance** of a traffic control device is the maximum distance at which one can see it, while its **legibility distance** is the distance from which the driver can discern the intended message in order to have time to take the necessary action. For a safety sign, the **distance factor** is the ratio of the observation distance to the size of the symbol or text.

  The **clear sight distance** is the length of highway visible to a driver. The **safe sight distance** is the necessary sight distance needed to a driver in order to accomplish a fixed task. The main safe distances, used in Road Design, are:

  - the **stopping sight distance**—to stop the vehicle before reaching an unexpected obstacle;
  - the *maneuver sight distance*—to drive around an unexpected small obstacle;
  - the *road view sight distance*—to anticipate the alignment (eventually curved and horizontal/vertical) of the road (for instance, choosing a speed);

[redacted]

- the *passing sight distance*—to overtake safely (the distance the opposing vehicle travels during the overtaking maneuver).

The *safe overtaking distance* is the sum of four distances: the passing sight distance, the **perception-reaction distance** (between decision and action; cf. Chap. 25), the distance physically needed for overtaking and the buffer safety distance.

Also, adequate sight distances are required locally: at intersections and in order to process information on traffic signs. A warning “objects in mirror are closer than they appear” can be required on vehicle’s passenger side mirrors.

In a traffic flow, the *spacing* (or *distance headway*) is the distance between corresponding points (front to front) of consecutive vehicles moving in the same lane, while the *clearance* is the spacing minus the length of the leading vehicle. The corresponding time measures are *headway* and *gap*.

- **Road travel distance**

The **road travel** (or *road*, *driving*, *wheel*, *actual*) **distance** between two locations (say, cities) of a region is the length of the shortest road connecting them.

Some GISs (Geographic Information Systems) approximate road distances as the l_p-metric with $p \approx 1.7$ or as a linear function of **great circle distances**; in the US the *distance factor* (multiplier) is ≈ 1.15 in an east–west direction and ≈ 1.21 in the north–south direction. Several relevant notions of distance follow.

The **GPS navigation distance**: the distance directed by GPS (Global Positioning System, cf. **radio distance measurement** in Chap. 29) navigation devices. But this shortest route, from the GPS system point of view, is not always the best, for instance, when it directs a large truck to drive through a tiny village; cf. the Talmudic little boy’s paradox among **distance-related quotes** in Chap. 28.

The **official distance**: the officially recognized (by, say, an employer or an insurance company) driving distance between two locations that will be used for travel or mileage reimbursement. Distance data (shortest paths between locations) are taken from a large web map service (say, MapQuest, Google, Yahoo or Bing) which uses a variation of the *Dijkstra algorithm*; cf. **Steiner ratio** in Chap. 1.

The **distance between zip codes** (in general, postal or telephone area codes) is the estimated driving distance (or driving time) between two corresponding locations.

Time-distance and **cost-distance** are time and cost measures of how far apart places are. The **journey length** is a general notion of distance used as a reference in transport studies. It can refer to, say, the average distance traveled per person by some mode of transport (walk, cycle, car, bus, rail, taxi) or a statutory vehicle distance as in the evaluation of aircraft fuel consumption.

An *odometer* is an instrument that indicates distance traveled by a vehicle. A *hubometer* is such device mounted on the axle of a vehicle, while a *trip meter* is an electronic device recording such distance in any particular journey.

Distance-based (or *mileage-based*, *per-mile*) *pricing* means that vehicle charges are based on the amount a vehicle is driven during a time period.

Dead mileage is when a public transit vehicle operates without carrying or accepting passengers, such as when coming from a garage to begin its first trip of the day.

- **Death/kilometer**

In countries collecting data on distance travelled, the traffic volume is measured in *vehicle-kilometrs* (calculated by odometer readings, traffic counts, driver survey or fuel consumption) or in *person-kilometers*, accounting for vehicle occupancy.

The most objective indicator to describe risk on the road network is fatalities per billion vehicle-kilometrs, or fatalities per billion person-kilometrs.

In the UK, the lifetime risk of dying according to various forms of transport is measured as relative traffic and crude death rate per 100 million vehicle-kilometrs.

In French-speaking journalism, *loi du mort-kilomètre*, i.e., the *law of death-kilometer*, is the principle that information (especially on victims) is more or less important according to its proximity—geographic, temporal, emotional, or social—to the news medium's user. Cf. **distance to death** in Chap. 23 and **distant suffering** in Chap. 28.

- **Horizontal distance**

The **horizontal distance** (or **ground distance**) is the distance on a true level plane between two points, such as scaled off the map (it does not take into account the relief between two points). There are two types of horizontal distance: **straight line distance** (the length of the straight line segment between two points as scaled off the map), and **distance of travel** (the length of the shortest path between two points as scaled off the map, in the presence of roads, rivers, etc.).

The *thalweg* (valley way) of a river or valley is the deepest inline within it.

The *stream gradient* is the slope measured (say, in m/km) by the ratio of drop in a stream per unit distance; the *relief ratio* is such average drop. The *gradient of a road* is the ratio of the vertical to the horizontal distance, measured in m/km or as slop tangent of the angle of the elevation. The *pitch* (or *slope*, *incline*) of a roof is the ratio of the rise to the roof span, expressed in cm/m.

- **Slope distance**

The **slope distance** (or **slant distance**) is the inclined distance (as opposed to the true horizontal or vertical distance) between two points.

In Engineering, the **rollout distance** is the distance that a boulder or rock took to finally reach its resting point after rolling down a slope. The *release height* is the height at which a boulder or rock was released in relation to a slope.

Naismith's rule in mountaineering: eight units of walking flat distance are time-equivalent to one unit of climb on a typical decline 12°.

Craeme *et al*, 2014, claim that the cost for an organism of mass M kg to walk uphill, gaining 100 m in altitude, is $2.94M$ kilojoules.

[redacted]

Walking uphill, humans and animals minimize metabolic energy expenditure; so, at critical slopes, they shift to zigzag walking. Langmuir's hiking handbook advises one to do it at 25°. Llobera and Sluckin, 2007, explain switchbacks in hill trails by the need to zigzag in order to maintain the critical slope, $\approx$ 16° uphill and $\approx$ 12.4° downhill. Skiing and sailing against the wind also require zigzagging.

The west face of Mount Thor, in the Canadian Arctic, is the Earth's greatest *vertical drop*: a uninterrupted wall 1250 m, with an average angle of 105°. The world record for the longest *rappel* (slope descent using ropes), 33 days, was set here in 2006. The world's highest unclimbed mountain is Gangkhar Puensum (7570 m) on the Bhutan-Tibet border. The most dangerous by fatality rate mountains are: Annapurna and K2, 10-th and 2-nd highest ones: 8091 and 8611.

- **Vertical distance**

  The **vertical distance** of a location is its height above or depth below a fixed *vertical datum*, i.e., reference surface, say, the Earth's surface, mean sea level (MSL) or its model. On other planets, the elevations of solid surface are measured relative to the mean radius.

  The terms *elevation* (or *geometric height*), *altitude* (or *geopotential height*) and *depth* are used for points/planes on the ground, in the air and below the surface, respectively. *AMSL*, *AGL*, *AAE* and (in Broadcasting) *HAAT* mean height above MSL, ground level, nearest aerodrome and average (surrounding) terrain, respectively. The *height of an aircraft* is its AGL, i.e., AMSL plus elevation of the ground.

  For measuring the height of land objects, the usual datum used is MSL, defined as the mean hourly water elevation taken over a specific 19 years cycle. Countries tend to choose MSL at one specific point to be used as the standard "zero elevation".

  Depths and tides on a nautical chart are measured relative to *chart datum*, defined as a level below which tide rarely falls: the lowest tide predictable from the effects of gravity, average lowest tide of each day, or (in waters with very low tidal ranges) MSL. Bridge clearances are given relative to a datum based on high tide.

  The *geoid* is a surface of a constant potential which is the best approximation, in a least-square sense, of the global MSL. The **orthometric height** of an object is its vertical distance above the geoid. Main nontidal datums are *gravimetric*, based on integral-mean gravity M between the geoid and the Earth's surface along some levelling line, and *ellipsoidal* one, used in satellite techniques.

  A **gravimetric height** of Earth's surface is $\frac{W_0-W}{M}$, where W, W_0 are the potentials of the Earth's gravity field on Earth's and geoid's levels. In particular, the *rigorous orthometric height* is the curved distance between the geoid and the Earth's surface along the *plumbline* (line orthogonal to the geoid). At present, there is no vertical datum, that is based on this height, due to difficulty in obtaining such M.

The minimal, maximal and average land heights are −418 m at the Dead Sea, 8848 m at the top of Mount Everest and 840 m, while the average and maximal depths of the ocean are 3680 m and 10,911 m (Challenger Deep in Mariana Trench). The surface's points closest (6353 km) and farthest (6384 km) from the Earth's center are the bottom of the Arctic Ocean and the summit of the Andean volcano Chimborazo (6268 m).

- **Prominence**

  In Topography, **prominence** (or *autonomous height*, *relative height*, *shoulder drop*) is a measure of the stature of a summit of a hill or mountain. The prominence of a peak is the minimum height of climb to the summit on any route from a higher peak (called the *parent peak*), or from sea level if there is no higher peak. The lowest point on that route is the *col*. So, the prominence of any island or continental highpoint is equal to its elevation above sea level.

  The highest mountains of the two largest isolated landmasses, Afro-Eurasia (Mount Everest) and the Americas (Aconcagua), have the most prominent peaks, 8848 m and 6962 m. But from its ocean base, the elevation of the Hawaiian volcano Mauna Kea (4205 m) is 10,203 m, and the mountain with the highest (5486 m) elevation from its land base is Mount McKinley (6193 m) in Alaska.

  The *topographic isolation* of a summit is the great circle distance to the nearest point of equal elevation; for Everest, it is 40,008 km (Earth's circumference between the poles). *Spire measure* (or *ORS*, short for *omnidirectional relief and steepness*) is a rough measure of the visual "impressiveness" of a peak. It averages out how high and steep a peak is in all directions above local terrain.

- **Special parallels and meridians**

  A network of *parallels* and *meridians* (lines of **latitude** and **longitude**, cf. Chap. 25) provides a locational system on Earth, using North Pole, South Pole (parallels 90°N and 90°S), rotation axis, and *equatorial plane* (an imaginary plane passing through Earth halfway between the poles and perpendicular to rotation axis).

  The *equator* is the imaginary midline, where the equatorial plane intersects Earth's surface. It is the parallel of 0° latitude separating North and South hemispheres.

  The *Prime meridian* and the *Date line* are internationally agreed at 0° and (with some bends, so as not to cross any land) ≈ 180° longitude; they form a great circle separating the Eastern and Western hemispheres. The point 0°, 0° is located in the Atlantic Ocean ≈ 614 km south of Accra, Ghana.

  A degree of latitude varies from 110.567 km apart at the equator to 111.699 km at the poles; each minute ($\frac{1}{60}$-th of a degree) is ≈ 1 mile. A degree of longitude shrinks from 111.321 km at the equator to 0 at the poles.

  The (solar) *terminator* (or, *circle of illumination*, *grey line*) is the great circle that divides Earth between a light half and a dark half. The *land hemisphere* is the hemisphere containing the largest possible area, $\approx \frac{7}{8}$, of land. It is centered on 47°13′N 1°32′W (in the city of Nantes, France). The other half is the *water hemisphere*.

[redacted]

The *tropic of Cancer* and the *tropic of Capricorn* are parallels at which the Sun is directly overhead at the northern and the southern summer solstice, respectively. Their positions depend on the Earth's axial tilt. The region between them, centered on the equator, is the *tropics*. The regions around 25–30°N and S are *subtropics*, and the regions 30–35°N and S are *horse latitudes* (or *subtropical highs*).

Equatorial and *polar* regions are within a few degrees of the equator or a pole. The *Arctic circle* and the *Antarctic circle* are parallels at which the Sun does not appear above the horizon at the northern and the southern winter solstice.

The longest land, continuous land, continuous sea latitudes are 48°24′53N (10,726 km France–Ukraine–Kazakhstan–China), 78°35′S (7958 km Antarctica), 55°59′S (22,471 km). The longest land, continuous land, continuous sea longitudes are 22°E (13,035 km Russia–China–Thailand), 99°1′30"E (7590 km), 34°45′45"W (15,986 km). The longest continuous land and sea distances along a great circle are 13,573 km (Liberia–Suez Canal–China) and ≈32,000 km.

Many parallels and meridians, often named and/or approximated, represent political boundaries. For example, 49°N latitude is (much of, from British Columbia to Manitoba) the border between Canada and US, 38°N is the boundary between North and South Korea, and 60°S is the northern boundary of Antarctica in the Antarctic Treaty. 2°20′14.025″E longitude is the *Paris meridian* (historic rival of the Prime meridian through Greenwich), 52.5°E is the official meridian of Iran, and ≈ 70°E was agreed in 1941 by Nazi Germany and the Empire of Japan as division of their spheres of interest in Asia. The *Brandt Line* (Brandt, 1980), represents socio-economic and political divide between the "rich North" and the "poor South". It encircles the world at ≈ 30°N latitude, passing between North and Central America, north of Africa and the Middle East, then going north so as to exclude China, Mongolia, Korea and going south so as to include Australia.

- **Remotest places on Earth**

In medieval geographies, *ultima Thule* was any distant place located beyond the borders of the known world. Eratosthenes (c. 276–195 BC), measuring the *oikoumene* (inhabited world), put its northern limit in a mythical island Thule.

The remotest island is uninhabited Bouvet island in the South Atlantic Ocean. Its nearest (1600 km) land is Antarctica and nearest inhabited land is Tristan da Cunha, the remotest inhabited archipelago.

Among other remotest (i.e., lacking normal transportation links) places on the Earth are: Kergelen (France), Pitcairn (UK), Svalbard (Norway) archipelagos, Easter (Chili), Foula (UK), Macquarie (Australia) islands, Motuo (China) county, McMurdo Station (Antarctica-US), La Rinconada (Peru, at an altitude of 5100 km) towns, Alert (Canada, 800 km below the North Pole) village.

The *continental pole of inaccessibility* (*Point Nocean*), the point on land farthest (2514 km) from any ocean, lies in the Xinjiang, China, around 45°22′N 88°11′E. The *oceanic pole of inaccessibility* (*Point Nemo*), the point farthest (2690 km) from any land, lies in the South Pacific Ocean at 48°52.6′S 123°23.6′W.

The *northern pole of inaccessibility* (84°03′N 174°51′W) is the point on the Arctic Ocean pack ice, 661 km from the North Pole, farthest (1094 km) from any land mass. The *southern pole of inaccessibility* (82°06′S 54°58′E) is the point on the Antarctic, 878 km from the South Pole, farthest (1300 km) from the ocean.

For a country, accessibility to its coast from its interior is measured by the ratio of coastline length in meters to land area in km^2. This ratio is the highest (10,100) for Tokelau and the lowest nonzero (0.016) for the Democratic Republic of the Congo. Canada has the longest (202,080 km) coastline.

The largest *antipodal* (diametrically opposite) land masses are the Malay Archipelago–Amazon Basin, and east China + Mongolia, antipodal to Chile + Argentina. Capitals close to being antipodes are: Buenos Aires–Beijing, Madrid–Wellington, Lima–Bangkok, Quito–Singapore, Montevideo–Seoul.

Politically unaccessed areas include isolated people (as Sentinelese and ≈ 100 tribes in dense forests) and unclaimed areas (antarctic Marie Byrd Land, Bir Tawil).

The *geographical centre of Earth* is the superficial barycenter produced by treating the geoid as a sphere and each continent or island as a region of a thin shell of uniform density. It is at 40°52′N 34°34′E (180 km northeast of Ankara).

Counting as different only population centers at >1000 km, the *point of minimum aggregate travel* (or *geometric median*, cf. **Fréchet mean** in Chap. 25) of the world's population lies around Afghanistan-Kashmir. This point is closest, 5200 km of the mean great circle distance, to all humans, and its antipodal point is the farthest from mankind. But the closest, 5600 km, point to the world's entire wealth (measured in GNP) lies in southern Scandinavia.

In terms of altitude, the number of people decreases faster than exponentially with increasing elevation (Cohen–Small, 1998). Within 100 m of sea level, lies 15.6 % of all inhabited land but 33.5 % of the world population live there. Altitude of residence is, via hypoxia, a risk factor for psychological distress in bipolar disorder.

- **Latitudinal distance effect**

  Diamond, 1997, explained the larger spread of crops and domestic animals along an east–west, rather than north–south, axis by the greater longitudinal similarity of climates and soil types.

  Ramachandran and Rosenberg, 2006, confirmed that genetic differentiation increases (and so, cultural interaction decreases) more with *latitudinal distance* in the Americas than with *longitudinal distance* in Eurasia. Randler, 2008: within the same time zone, people in the east get up and go to bed earlier than people in the west.

  Turchin–Adams–Hall, 2006, observed that ≈ 80 % of land-based, contiguous historical empires are wider in the east–west compared to the north–south directions. Three main exceptions—Egypt (New Kingdom), Inca, Khmer—obey a more general rule of expansion within an ecological zone.

  Taylor et al., 2014: polyandry in species is more common in northern latitudes.

  The *latitudinal biodiversity gradient* (LBG) refers to the decrease in in both terrestrial and marine biodiversity, that occurs, the past 30 Ma, from the equator

[redacted]

to the poles for most fauna and flora. Its causes remain unresolved. Deep-time studies indicate that it might characterise cold icehouse climatic regimes,

Mace and Pagel, 1995 and 2004, found the same gradient for the density (number per range) of language groups and cultural variability. Around 60 % of the world's languages are found in the great belts of equatorial forest. Papua New Guinea (14 % of languages), sub-Saharan Africa and India have the largest linguistic diversity. The number of phonemes in a language decrease, but the number of color terms increase, from the equator to the poles.

25.2 Distances in Geophysics

- **Atmospheric visibility distances**

  Atmospheric extinction (or *attenuation*) is a decrease in the amount of light going in the initial direction due to *absorption* (stopping) and scattering (direction change) by particles with diameter 0.002–100 μm or gas molecules. The dominant processes responsible for it are *Rayleigh scattering* (by particles smaller than the wavelength of the incident light) and absorption by dust, ozone O_3 and water. For example, mountains in the distance look blue due to the Rayleigh scattering effect.

  In extremely clean air in the Arctic or mountainous areas, the visibility can reach 70–100 km. But it is often reduced by air pollution and high humidity: haze (in dry air) or mist (moist air). *Haze* is an atmospheric condition where dust, smoke and other dry particles (from farming, traffic, industry, fires, etc.) obscure the sky.

  The World Meteorological Organization classifies the horizontal obscuration into the categories of fog (a cloud in contact with the ground), ice fog, steam fog, mist, haze, smoke, volcanic ash, dust, sand and snow. Fog and mist are composed mainly of water droplets, haze and smoke can be of smaller particle size.

  Visibility of less than 100 m is usually reported as zero. The international definition of *fog*, *mist* and *haze* is a visibility of < 1 km, 1–2 km and 2–5 km.

  In the air pollution literature, **visibility** is the distance at which the contrast of a visual target against the background (usually, the sky) is equal to the threshold contrast value for the human eye, necessary for object identification, while **visual range** is the distance at which the target is just visible. Visibility can be smaller than the visual range since it requires recognition of the object.

  Visibility is usually characterized by either visual range or by the *extinction coefficient* (attenuation of light per unit distance due to four components: scattering and absorption by gases and particles in the atmosphere). It has units of inverse length and, under certain conditions, is inversely related to the visual range.

  Meteorological range (or *standard visibility*, *standard visual range*) is an instrumental daytime measurement of the (daytime sensory) visual range of a target. It is the furthest distance at which a black object silhouetted against a sky would be visible assuming a $2\,\%$ threshold value for an object to be distinguished from the background. Numerically, it is $\ln 50$ divided by the extinction coefficient.

In Meteorology, **visibility** is the distance at which an object or light can be clearly discerned with the unaided eye under any particular circumstances. It is the same in darkness as in daylight for the same air. **Visual range** is defined as the greatest distance in a given direction at which it is just possible to see and identify with the unaided eye in the daytime, a prominent dark object against the sky at the horizon, and at night, a known, unfocused, moderately intense light source.

The International Civil Aviation Organization defines the **nighttime visual range** as the greatest distance at which lights of 1000 candelas can be seen and identified against an unlit background. Daytime and nighttime ranges measure the atmospheric attenuation of contrast and flux density, respectively.

In Aviation Meteorology, the **runway visual range** is the maximum distance along a runway at which the runway markings are visible to a pilot after touchdown. It is measured assuming constant contrast and luminance thresholds.

Oblique visual range (or *slant visibility*) is the greatest distance at which a target can be perceived when viewed along a line of sight inclined to the horizontal.

- **Atmosphere distances**

The **atmosphere distances** are the altitudes above Earth's surface (mean sea level) which indicate approximately the following specific (in terms of temperature, gravity, electromagnetism, etc.) layers of its atmosphere.

Below 1–2 km: *planetary boundary layer*, where winds are directly retarded by surface friction. The reminder of the atmosphere: the *free atmosphere*.

From 8 km: the death zone for human climbers (lack of oxygen).

From the *Armstrong line* (18.900–19.350 km) water boils at 37 °C (low pressure) and a pressure suit is needed.

Below 7–20 km (over the poles and equator, respectively): unstable *troposphere* in which temperature decreases with height (the weather and clouds occur here).

Above the troposphere to $\approx$ 51 km: stable *stratosphere*, where the temperature increases with height (the ozone layer is at 19–48 km). The *tropopause* (its boundary with the troposphere) occurs at a pressure $\approx$ 0.1 bar.

Above the stratosphere to 80–85 km: the *mesosphere*, in which temperature again decreases with height. Above the mesosphere to 500–1000 km: the *thermosphere*, where the temperature again increases with height.

20–100 km: *near space* (or *upper atmosphere*), above airliners but below satellites.

100 km: the *Kármán line* prescribed by Fédération Aéronautique International as the boundary separating Aeronautics and Astronautics, near and outer space.

Above the thermosphere to $\approx$190,000 km : the *exosphere*, where molecules are still gravitationally bound but they can escape into space. Below the exosphere: the *homosphere*, where atmosphere has relatively uniform composition

[redacted]

since turbulence causes a continuous mixing. The reminder of the atmosphere: the *heterosphere*.

From 50–80 to 2000 km: the *ionosphere*, an electrically conducting region. From $\approx$ 160 km upwards: the *anacoustic zone*, where distances between air particles are so great that sound can no longer propagate; high-frequency sounds disappear first.

Up to 6–10 Earth radii on the sunward side: the *magnetosphere*, where Earth's magnetic field still dominates that of the solar wind. *Geospace* is the region from the beginning of ionosphere to the end of magnetosphere.

The altitude of the International Space Station is 278–460 km. 35,786 km: the altitude of geostationary (communication and weather) satellites. For observation and science satellites, it is 480–770 km and 4800–9700 km, respectively. Geocentric orbits with altitudes up to 2000 km, 2000–35,786 km and more than 35,786 km are called low, medium and high Earth orbits, respectively.

From 320,000 km: Moon's (at 356,000–406,700 km) gravity exceeds Earth's.

1,496,000 km = 0.011 AU: Earth's **Hill radius**, where Sun's gravity Earth's.

- **Wind distances**

Examples of wind-related distances follow.

Monin–Obukhov length: a rough measure of the height over the ground, where mechanically produced (by vertical wind shear) turbulence becomes smaller than the buoyant production of turbulent energy (dissipative effect of negative buoyancy). In the daytime over land, it is usually 1–50 m.

The **aerodynamic roughness length** (or *roughness length*) z_0 is the height at which a wind profile assumes zero velocity.

The **wind daily run** is the distance that results by integrating the wind speed, measured at a point, over 24 h. The fastest recorded wind speed near Earth's surface was 318 mph (i.e., 511.76 km/h) in Oklahoma, US, in 1999.

Rossby radius of deformation is the distance that cold pools of air can spread under the influence of the *Coriolis force*, i.e., the apparent deflection of moving objects when they are viewed from a rotating reference frame. It is the length scale at which effects, caused by Earth's rotation and the inertia of the mass experiencing the effect, become as important as buoyancy or gravity wave effects in the evolution of the flow about some disturbance.

The *aerial plankton* carried aloft by winds or convection, consists of bacteria, fungi, spores, pollen and small invertebrates. Even in the upper troposphere (8–15 km altitude), viable bacteria cells represent $\approx$ 20 % of $0.25 - -1$ μm-sized particles.

The *jet streams* are fast flowing, narrow air currents found in the atmosphere. The strongest jet streams are, both west-to-east and in each hemisphere, the *Polar jet*, at 7–12 km above sea level, and the weaker *Subtropical jet* at 10–16 km.

The *atmospheric rivers* are narrow (a few hundred km across but several thousand km long) corridors of atmospheric water vapor transport over mid-latitude ocean regions. They account for over 90 % of such global meridional daily transport.

A *teleconnection* refers to climate anomalies being related to each other at large, say, thousands of km distances. For example, teleconnection between sea level pressure at Tahiti and Darwin, Australia, defines ENSO (El Niño Southern Oscillation).

- **Distances in Oceanography**

Sea-surface height (SSH) is the height of the ocean's surface. **Decay distance**: the distance through which ocean waves travel after leaving the generating area.

Wavelength is the distance between the troughs at the bottom of consecutive waves. A wave's *height* and *amplitude* are its trough-crest and rest-crest distances.

The **significant wave height** (SWH) is the mean height $H_{1/3}$ of the highest third of waves. More modern and very close value is four times the root-mean-square of the surface elevation. Unusual *rogue waves* are defined as those with height $> 2H_{1/3}$; the tallest recorded one (by ship USS Ramapo in 1933) was 34 m. A wave up to $11H_{1/3}$ is possible. Large internal waves occur at thermoline and saltwater-freshwater interface. A *hundred-year wave* is a statistically projected water wave, the height of which, on average, is met or exceeded once in a 100 years for a given location.

The maximum horizontal distance inland and height reached there by tsunami waters are called the **run-up** (or *inundation*) **distance** and **run-up height**. It was 1100 and 525 m for the 1958 Lituya Bay, Alaska, megatsunami, the largest recorded.

Deep water (or *short*, *Stokesian*) **wave**: a surface ocean wave that is traveling in water depth greater than one-quarter of its wavelength; the velocity of deep water waves is independent of the depth. **Shallow water** (or *long*, *Lagrangian*) **wave**: a surface ocean wave of length 25 or more times larger than the water depth.

Littoral (or *intertidal*): the zone between high and low water marks. Sometimes, *littoral* refers to the zone between the shore and water depths of $\approx$ 200 m.

Oceanographic (or *thermal*) **equator**: the zone of maximum sea surface temperature located near (generally, north) the geographic equator. Sometimes, it is defined more specifically as the zone within which the sea surface temperature exceeds 28 °C. Below about 500 m, all of the world's oceans are at about 1.1 °C.

Standard depth: a depth below the sea surface at which water properties should be measured and reported (in m): 0, 10, 20, 30, 50, 75, 100, 150, 200, 250, 300, 400, 500, 600, 800, 1000, 1200, 1500, 2000, 2500, $3000, 4000, \dots, 9000, 10{,}000$.

Charted depth: the recorded vertical distance from the *lowest astronomical tide* (LAT, the lowest low water that can be expected in normal circumstances) to the seabed. **Drying height**: the vertical distance of the seabed that is exposed by the tide, above sea level at LAT. Actual depth of water is height of tide + charted depth or height of tide—drying height. **Tidal range**: the difference between the heights of high water and low water at any particular place. The empirical *rule*

[redacted]

of twelfths assigns 6 h for it and subdivides the range into $1, 2, 3, 3, 2, 1$ twelfths per hour.

The **thermocline**, **halocline** and **pycnocline**: the layers where the water temperature, salinity and density, respectively, change rapidly with depth.

Depth of no motion: a reference depth in a body of water at which it is assumed that the horizontal velocities are practically zero. On a horizontal scale, *ocean fronts* are the boundaries between water masses with different properties.

Plankton (viruses, bacteria, phytoplankton, zooplankton and small pelagic larvae) aggregate at the clines, depth of no motion and persistent ocean fronts. $\approx$ 75 % of the water column's biomass consist of plankton organized in thin (<3–4 m) layers 1–12 km in horizontal extent. Standard proxies for phyto- and zooplankton abundance are chlorofill-a imagery and sound attenuation. Giant (up to 130,000 km^2) bacterial mats float in the oxygen minimum zone off Chili and Peru.

Depth of the effective sunlight penetration: the depth at which $\approx$ 1 % of solar energy penetrates; in general, it does not exceed 100 m. The ocean is opaque to electromagnetic radiation with a small window in the visible spectrum. But it is transparent to acoustic transmission.

Depth of compensation: the depth at which illuminance has diminished to the extent that oxygen production through photosynthesis and oxygen consumption through respiration by plants are equal. The maximum depth for photosynthesis depends on plants and weather. Within the *epipelagic zone* there is enough light for photosynthesis, and thus plants and animals are largely concentrated here.

Below the *mesopelagic zone* lies the **aphotic zone** which is not exposed to sunlight. Organisms there depend on "marine snow" (a continuous shower of mostly organic detritus, decaying creatures and feces, falling from above) and chemosynthesis. The *deep sea* is the layer in the ocean below thermocline, at the depth 1800 m or more.

The **pelagic zone** consists of all the sea other than that near the coast or the sea floor, while the *benthic zone* is the ecological region at the very bottom of the sea.

The ocean is divided into the following horizontal layers from the top down:

- From the surface down 200 m: *epipelagic* (sunlit zone);
- 200–1000 m: *mesopelagic* (twilight zone);
- 1000–4000 m: *bathypelagic* (dark zone);
- 4000–6000 m: *abyssopelagic* (abyss);
- below 6000 m: *hadalpelagic* (trenches).

The **deep sound** (or **SOFAR**, i.e., SOund Fixing And Ranging) **channel** is a layer of ocean water where the speed of sound is at a minimum ($\approx$1480 m/s), because water pressure, temperature and salinity cause a minimum of water density. Sound waves of low frequency, caught and bent here, can travel hundreds of km. In low and middle latitudes, the SOFAR channel axis lies 600–1200 m below the sea surface; it is deepest in the subtropics and comes to the surface in high latitudes.

The **SLD** (sonic layer depth) is the depth of maximum sound speed above this axis. The **best depth** for a submarine to avoid detection is SLD plus 100 m.

Mixing length: the distance at which an *eddy* (a circular movement of water) maintains its identity until it mixes. An eddy can reach 500 km across and persist for months. Cf. the **mean free path** and **diffusion length** in Chap. 24.

Mixed layer depth: the depth of the bottom of the *mixed layer*, i.e., a nearly isothermal surface layer of 40–150 m depth where water is mixed through wave action or thermohaline convection.

Depth of exponential mixing or **depth of homogeneous mixing** refers to a surface turbulent mixing layer in which the distribution of a constituent decreases exponentially, or is constant, respectively, with height.

Depth of frictional resistance: the depth at which the wind-induced current direction is 180° from that of the true wind.

The *fetch* (or *fetch length*): the horizontal distance along open water over which wave-generating wind or waves have traveled uninterrupted. In an enclosed body of water, the *fetch* is the distance between the points of minimum and maximum water-surface elevation. In Meteorology, the *fetch* is the distance upstream of a measurement site, receptor site, or region of interest that is relatively uniform.

The total volume of Earth's water is $\approx$ 1.39 billion km^3 (0.2 % Earth's mass) of which $\approx$ 96 % is liquid. For each sustained 1 °C increase in global temperature, the sea level could rise by 2.3 m. Global sea level will rise by at least 1 m by 2100. At 2016, it is higher than anytime during last 115,000 years.

The ocean on Ganymede (Jupiter's satellite and the largest Solar System's moon) has total volume 40 billion km^3 and depth 800 km (30 % of Ganymede's radius).

- **River length**

  Meaning of **river length**, i.e., the distance between source and mouth, depend on their definitions, *anabranches* (multiple channels), map scale, etc.

  The **maximal river length** is the length of the longest continuous river channel in a given river system, regardless of name. Here, a river's "true source" is considered to be the source of whichever tributary is farthest from the mouth.

  The world's longest rivers are Nile-Kagera (≈6650–6853 km), Amazon-Ucayali-Apurmac (≈6400–6992 km) and Yangtze (≈6300–6418 km).

- **Soil distances**

  Soil is composed of particles of broken rock that have been altered by chemical and environmental processes that include weathering and erosion. It is a mixture of mineral and organic constituents that are in solid, gaseous and aqueous states. A *soil horizon* is a specific layer in the land area that is parallel to the soil surface and possesses physical characteristics which differ from the layers above and beneath. Each soil type usually has 3–4 horizons.

  - *A Horizon* (or *topsoil*): the upper layer (usually 5–20 cm) with most organic matter accumulation and soil life.

[redacted]

- *B Horizon* (or *subsoil*): the deeper layer accumulating by *illuviation* (action of rainwater), iron, clay, aluminum and organic compounds.
- *C Horizon*: the layer which is little affected by soil forming processes.
- *R Horizon*: the layer of partially weathered bedrock at the base of the soil profile.

The current average rate of erosion over the Earth's land is 30 B, where B is the *Bubnoff unit* of speed (for lowering of earth surfaces due to erosion) equal to 1 m in a million year, i.e., removing 1 m^3 of earth from an area of 1 km^2 in 1 year.

The *pedosphere* is the outermost layer of the Earth that is composed of soil and subject to soil formation processes. It lies below the vegetative cover of the biosphere and above the groundwater and *lithosphere* (outermost shell of the Earth).

Larger *Critical Zone* includes vegetation, the pedosphere, groundwater aquifer systems and ends in the bedrock where the biosphere and *hydrosphere* (combined mass of Earth's water) cease to make significant changes to the chemistry.

The *water table* (or *phreatic surface*) is the level at which the groundwater pressure is equal to atmospheric pressure.

The *cryosphere* is the part of the hydrosphere describing the Earth's ice: sea/lake/river ice, snow cover, glaciers, ice caps, ice sheets and frozen ground including permafrost. The Bentley Subglacial Trench in Antarctica is the world's deepest, 2555 m, ice.

The Earth is now in a warm phase of the 5-th (Quaternary) major Ice Age, which started 2.58 Ma ago, The last glacial expansion ended 11,700 years ago with the start of the Holocene. Next one is expected within coming 1500–10,000 years unless anthropogenic global warming will delay it. The planet has warmed by only 0.74 °C since the early 1900s.

- **Frost line (in Earth Science)**

  The **frost line** (or *freezing depth*) is the depth to which the groundwater in soil is expected to freeze. In polar locations with year-round permafrost, the *thaw depth* is the depth to which the permafrost is expected to thaw each summer.

  In tropical regions, *frost line* may refer to the vertical geographic elevation below which frost does not occur. The *climatic snow line* is the point above which snow and ice cover the ground throughout the year; seasonally, snow occurs much lower. Cf. **frost line (in Astrophysics)**.

- **Moho distance**

  The Earth's *oceanic crust* (or *sima* for Si and Mg in basaltic rocks) is the surface, 5–10 km thick, of the ocean basins. The *continental crust* (or *sial* for Si and Al) is the layer of granitic rocks, 20–90 km thick, forming continents and continental shelves. The *Moho interface* (or *Mohorovicić seismic discontinuity*) is the boundary between the crust and the mantle, where the velocity of seismic P-waves increases. The **Moho distance** is the crustal thickness, i.e., the distance from a surface's point to the *Moho interface* beneath it.

The world's lowest sea-drilled point was 10,680 m-deep (in the Gulf of Mexico) under 1259 m of water. The Japanese research vessel *Chikyu*, aiming to the Moho interface, drilled 7740 m below the sea level and 2466 m below the seafloor. Cf. the lowest point on dry land (the shore of Dead Sea: 418 m), deepest cave (Krubera, Caucasus: 2191 m and, Chavé Cave, Mexico, possibly, $\sim$ 2500 m), deepest mine (Mponeng gold mine, South Africa: about 4 km) and deepest drill (Kola Superdeep Borehole: 12,262 m). The temperature rises usually by 1° every 33 m.

The *Curie depth* is the depth (usually 10–50 km) where the temperature reaches the *Curie point* at which rocks lose their ferromagneticic properties.

The Earth's mantle extends from the Moho interface to the mantle-core boundary at $\approx$ 2890 km. The liquid outer core of radius 3480 km (i.e., Mars-sized) contains a solid iron inner core (expanding $\approx$ 0.5 mm per year) of radius 1220 km. The mantle is divided into the upper and lower one at $\approx$ 660 km. Main other seismic boundaries are at about 60–90 km, 50–150 km, 220 km (Hales, Gutenberg and Lehmann discontinuities) and 410 km, 520 km, 710 km.

The *lithosphere* comprises the crust and the rigid portion of the upper mantle that behaves elastically on large time scales. Its thickness is the depth of the isotherm $\approx$1000° of the transition between brittle and viscous behavior. The lithosphere is broken into *tectonic plates* which float on the more plastic part of the mantle, the *asthenosphere*, 100–200 km deep.

The Eurasian and African plates are moving at the rates of 2 and 2.15 cm per year. The maximum earthquakes occur on the boundaries of the major moving plates. The eastern part of Indo-Australian plate is moving north 5.6 cm per year while the western part (India) is moving (north-east) only 3.7 cm per year due to impediment by Himalayas. The sequence of earthquakes in 2012 off the coast of Sumatra indicates that this plate has broken up into 2 or 3 separate plates.

- **Distances in Seismology**

  The Earth's crust is broken into tectonic plates that move around (at some cm per year) driven by the thermal convection of the deeper mantle and by gravity. At their boundaries, plates stick most of the time and then slip suddenly.

  An *earthquake*, i.e., a sudden (several seconds) motion or trembling in the Earth, caused by the abrupt release of slowly accumulated strain, was, from 1906, seen mainly as a rupture (the sudden appearance, nucleation and propagation of a new crack or fault) due to elastic rebound. However, from 1966, it is seen within the framework of slippage along a pre-existing fault or plate interface, as the result of stick-slip frictional instability. One of most important parameters controlling fault instability is the **slip-weakening distance** D_c over which a fault weakens during its seismogenic motion. The coefficient of friction degrades linearly with slip until D_c is reached; then it stays constant.

  So, an earthquake happens when dynamic friction becomes less than static friction. The advancing boundary of the slip region is called the *rupture front*. The standard approach assumes that the fault is a definite surface of tangential displacement discontinuity, embedded in a liner elastic crust.

[redacted]

Most earthquakes occur at near-vertical faults but a magnitude 6.0 earthquake at Kohat, Pakistan, in 1992, moved a 80 km^2 swath of land 30 cm horizontally. Almost all (81 % and 17 %) world's largest earthquakes occur along the *Ring of Fire* (circum-Pacific seismic belt) and the *Alpide belt* (from Java to Sumatra through the Himalayas, the Mediterranean, and out into the Atlantic).

90 % of earthquakes are of tectonic origin, but they can also be caused by volcanic eruption, nuclear explosion and work in a large dam, well or mine. Earthquakes can be measured by **focal depth**, speed of slip, intensity (modified Mercali scale of earthquake effects), magnitude, acceleration (main destruction factor), etc.

The Richter scale of magnitude is computed from the amplitude and frequency of shock waves received by a seismograph, adjusted to account for the **epicentral distance**. An increase of 1.0 of this magnitude corresponds to an increase of 10 times in amplitude of the waves and $\approx$ 31 times in energy; the largest recorded value is 9.5 (Chile, 1960). Asteroid's impact in Yucatan 66 Ma ago was 12.55.

An earthquake first releases energy in the form of shock *pressure waves* that move quickly through the ground with an up-and-down motion. Next come shear S-waves which move along the surface, causing much damage: *Love waves* in a side-to-side fashion, followed by *Rayleigh waves* which have a rolling motion. The earthquake *extinction length* is the distance over which the S-wave energy is decreased by $\frac{1}{e}$.

Distance attenuation models (cf. **distance decay** in Chap. 29), used in Earthquake Engineering for buildings and bridges, postulate acceleration decay with an increase of some **site-source distance**, i.e., the distance between seismological stations and the crucial (for the given model) "central" point of the earthquake.

The simplest model is the *hypocenter* (or focus), i.e., the point inside the Earth from which an earthquake originates (the waves first emanate, the seismic rupture or slip begins). The *epicenter* is the point of the Earth's surface directly above the hypocenter. This terminology is also used for other catastrophes, such as an impact or explosion of a nuclear weapon, meteorite or comet but, for an explosion in the air, the term *hypocenter* refers to the point on the Earth's surface directly below the burst. A list of the main Seismology distances follows.

The **focal depth**: the distance between the hypocenter and epicenter. Earthquake is *shallow-*, *mid-* or *deep-focus* if it is $<$ 70 km, 70–300 km or 300–700 km.

The **hypocentral distance**: the distance from the station to the hypocenter.

The **epicentral distance** (or **earthquake distance**): the **great circle distance** from the station to the epicenter.

The **Joyner–Boore distance** (1981): the distance from the station to the closest point, located over the *rupture surface* (rupturing portion of the fault plane).

The **rupture distance**: the distance from the station to the closest point on the rupture surface. The **seismogenic depth distance**: the distance from the station to the closest point of the rupture surface within the *seismogenic zone*, i.e., the depth range where the earthquake may occur; usually at depth 8–12 km.

The **crossover distance**: the distance on a seismic refraction survey time-distance chart at which the travel times of the direct and refracted waves are the same.

Also used are the distances from the station to:

- the center of static energy release and the center of static deformation of the fault plane;
- the surface point of maximal macroseismic intensity, i.e., of maximal ground acceleration (it can be different from the epicenter);
- the epicenter such that the reflection of body waves from the *Moho interface* (the crust-mantle boundary) contribute more to ground motion than directly arriving shear waves (it is called the *critical Moho distance*);
- the line extending the *fault trace* (top edge of the rupture) in both directions;
- the sources of noise and disturbances: oceans, lakes, rivers, railroads, buildings.

The **space-time link distance** between two earthquakes x and y is defined by

$$\sqrt{d^2(x,y) + C|t_x - t_y|^2},$$

where $d(x,y)$ is the distance between their epicenters or hypocenters, $|t_x - t_y|$ is the time lag, and C is a scaling constant needed to connect distance and time.

The *earthquake distance effect*: at greater distances from its center, the perception of an earthquake weaken and lower frequency shaking dominates it. Many animals hear infrasound of imminent earthquakes and feel primary P-waves.

Another space-time measure for catastrophic events is **distance between landfalls** for hurricanes hitting a US state. It is (Landreneau, 2003) the length of state's coastline divided by the number of hurricanes which have affected it from 1899.

Also, Shennan et al., 2015, compare archeological cultures by both, great circle distance between sites and Euclidean distance between median culture dates.

- **Plume height**

  In a volcanic eruption, **plume height** refers to the highest point the eruptive cloud reaches before it flattens out and begins to drift downwind.

  The Volcanic Explosivity Index (VEI) is a scale measuring known eruptions by their volume of ejecta and plume height from VEI 0 (1000 m^3, < 100 m) to *mega-colossal* (or *super-volcano*) VEI 8 (1000 km^3, > 50 km).

- **Weather distance records**

  For a tornado, maximum width of damage, highest elevation, longest path: 4000 m, 3650 m, 472 km. The longest transport of a surviving human and of an object: 398 m and 359 km (personal check).

[redacted]

Longest path of a tropical cyclone: 13,500 km; highest storm surge: 13 m.

Largest snowflake and hail diameter: 38 and 20 cm. Longest lighting bolt: 190 km.

Greatest minute, hour, day, month, year rainfall: 31.2 mm, 0.3 m, 1.82 m, 9.3 m, 26.47 m. Lowest and highest mean annual precipitation: 0.762 mm and 11.872 m.

- **Extent of Earth's biosphere**

  Life has adapted to every (except, perhaps, ocean vent locales > 130 °C) ecological niche possessing liquid water and a source of free energy (say, sunlight, plate tectonics, water-rock chemistry). The main physical factors are temperature and pressure; their range for known active life as $[-20^\circ, 122^\circ]$C and $[5 \times 10^{-2}, 1.3 \times 10^3]$ bar. But the range, say, $[-30^\circ, 135^\circ]$C looks possible. The acidity/alkalinity range of known life is [0, 11] on the pH scale, from acidic hot springs to soda lakes.

  In Jones–Lineweaver, 2010, the depth 5–10 km of the 122 °C isotherm and the altitude 10–15 km (a *tropopause* boundary of the vertical movement of particles) are the boundaries of active life. In Nussinov–Lysenko, 1991, the boundaries of biosphere are *Moho interface* (say, −30 km) and *Kármán line* (100 km).

  For humans, the typical bounds for main physiologic factors are: core temperature 35–38 °C, serum pH 7.35–7.45, plasma *osmolality* 270–290 mOsm/kg, fasting plasma glucose 3.3–5.6 mmol/l and serum calcium 2.2–2.6 mmol/l. But there are permanent human habitations at mean annual temperatures of 34.4 °C, −19.7 °C and at an altitude of 5.1 km. Birds usually fly at altitudes 0.65–1.8 km but a vulture collided, at 11.3 km, with an aircraft.

  A fish has been spotted at 8372 m down in the Puerto Rico Trench. Deepest multicellular life are worms found at depths up to 3.6 km in gold mines and up to 10.8 km in the Challenger Deep. Stoddard, 2015, expects microbial life up to 19 km deep since high pressure can offset the heat's destructive effects.

  Microbes, supported by chemosynthesis, have been found in cores drilled 5.3 km, in hydrothermal vents at 11 km depth and 2.5 km beneath ocean's floor. Such *deep biosphere* (1–10 % of the world's biomass and the Earth vastest) is expected below the surface of continents and bottom of the ocean up to isotherm 122 °C. Some of such microbes have doubling times in millennia, while *Vibrio natriegens* divide in 9.8 min. The same 19 deep-rock bacteria found to be similar worldwide.

  The ranges for latent life (*cryptobiosis*: reversible state of low or undetectable metabolism) are much larger. Fungi and bacterial spores were found at an altitude 18–41 km. Examples of survival limits follow.

  Some frogs, turtles and snakes survive the winter by freezing solid. A brine shrimp *Artemia* tolerates salt amounts of 25 %. Tardigrades, in cryptobiosis, survive −272 °C, 151 °C (a few minutes), −20 °C (30 years), pressure 6000 bar, radiation 6200 gray of gamma rays and 120 years without water. Fly's larva *Polypedilium vanderplanki* dehydrates, in dry period, to 3 % water content, and it can survive −270 °C, 102 °C, radiation 7000 g and 18 months in outer space vacuum. Some Antarctic fungi and Alpine lichens survive 18 months in Mars-like conditions or space travel.

A parasitic leach *Ozobranchus jantseanus* survives $-196\,^\circ$C (24 h) and $-90\,^\circ$C during 32 months. Archea *Thermococcus gammatolerans* survive 30,000 gray. Bacteria growing under hypergravity 403,627 g were cultivated. A bacterium survived 2.5 years on the Moon. *Deinococcus radiodurans*, listed as the world's toughest bacterium, survive extreme cold, dehydration, vacuum, radiation and acid. A bacteria *Tersicoccus phoenicis* has only been found in two spacecraft assembly clean rooms and is resistant to the methods used to clean such facilities.

Millions of years old (nondormant, just slow metabolizing) microbs, reproducing only every 10,000 years, were found in ocean floor. Bacterial spores were revived after 34,000 years of stasis; it was claimed also for 40 Ma old spores. A 1300 years old lotus seed and 2000 year old seed from extinct Judean date palm were germinated. *Silene stenophylla* was grown from 31,800 years old fruit.

Among the proponents of *panspermia* (the hypothesis that life, via extremophile bacteria and crystallized viruses surviving in space, propagates throughout the Universe), Yang et al., 2009, expect microbe density to be $10^{-3} - 10^{-2}$ cells/m^3 at altitude 100 km and $10^{-6} - 10^{-4}$ at 500 km. A large amount is expected at the altitude of the ISS (278–460 km). Napier–Wickramasinghe, 2010, claim that $10^{14} - 10^{16}$ microorganisms ($\approx$ 10 tonnes) per year are ejected from Earth at survivable temperatures. Organics preserved in cometary amorphous ice and meteorite-formed glass can be transported from one planet to another. Also, microbes, if somehow protected from UV, survive in meteorites with transit time of a few years.

A total of 7.5×10^{15} terrestrial microbes could reach the Moon per year, and the Solar System could be surrounded by an expanding biosphere of radius $>$ 5 parsecs containing $10^{19} - 10^{21}$ microbes. No ubiquitous ultrasmall bacteria were found but large *Bacillus* and eukaryotes (5–100 μ fungal spores) have been isolated from the stratosphere 27 km up, too high to have been lifted from Earth's surface. So, some viable but nonculturable microbes could be incoming from space. Hoover, 2011, found microfossils similar to filamentous prokaryotes in CI1 (Alais, Ivuna and Orgueil) and CM2 (Murchison and Murray) meteorites.

Life on Mars, if any, is expected to be of the same origin (and, perhaps, earlier) as that on Earth, but it would have to be under at least 1 m of soil/rock to survive. Impact of icy comets crashing into Earth billions of years ago could have produced a variety of prebiotic or life-building compounds, including amino acids.

Interstellar panspermia, when the Sun passes a star-forming cloud, and even intergalactic panspermia, when galaxies collide, are debated. But on a cosmic scale, even enthusiasts of panspermia see it as a local, "a few megaparsecs", phenomenon.

[redacted]

25.3 Distances in Astronomy

A *celestial object* (or *celestial body*) is a term describing astronomical objects such as stars and planets. The *celestial sphere* is the projection of celestial objects into their apparent positions in the sky as viewed from the Earth. The *celestial equator* is the projection of the Earth's equator onto the celestial sphere. The *celestial poles* are the projections of Earth's North and South Poles onto the celestial sphere. The *hour circle* of a celestial object is the great circle of the celestial sphere, passing through the object and the celestial poles.

The *ecliptic* is the intersection of the plane, containing the Earth's orbit, with the celestial sphere: seen from the Earth, it is the path that the Sun appears to follow over the course of a year. The *vernal equinox point* (or the *First point in Aries*) is one of the two points on the celestial sphere, where the equator intersects the ecliptic: it is the position of the Sun at the time of the vernal equinox.

In Astronomy, the *horizon* is the horizontal plane through the eyes of the observer. The *horizontal coordinate system* is a celestial coordinate system using the observer's local horizon as the *fundamental plane*, the locus of points having an altitude of 0°. The horizon is the line separating Earth from sky; it divides the sky into the upper hemisphere that the observer can see, and the lower hemisphere that he cannot. The pole of the upper hemisphere (the point of the sky directly overhead) is called the *zenith*; the pole of the lower hemisphere is called the *nadir*.

Positions and distances are calculated with respect to the center of mass of a body, not with respect to the geometric center, or surface. In general, an **astronomical distance** is a distance from one celestial body to another measured in astronomical units or gygameters for planets, light-years (ly) or parsecs for stars, kiloparsecs at the galactic scale, in megaparsecs (Mpc) for nearby galaxies and in terms of redshift for distant galaxies.

The average distance between stars (in a galaxy) is several ly; it is about 4 ly in Milky Way and 6.57 ly in the solar neighborhood. The average distance between galaxies (in a cluster) is several Mpc, while between clusters of galaxies (in a supercluster) it is of order 10 Mpc. The large structures are groups of galaxies, clusters, galaxy clouds (or groups of clusters), superclusters, and supercluster complexes (or galaxy filaments, great walls). A range of structures: galaxies ($\sim$ 10 kpc), groups (0.3–1 Mpc), clusters (few Mpc), superclusters (10–100 Mpc).

The Universe appears as a collection of giant bubble-like voids separated by great walls, with the superclusters appearing as relatively dense nodes. The average density of stars is about 1.4 per 100 billion cubic light-years, i.e., the average distance between them is about 4150 light-years. The mean density of visible matter (i.e., galaxies) in the Universe is $\sim 10^{-31}$ g/cm^3, while 1 g/cm^3 is the mass density of water.

ESA's Herschel space observatory found that our galaxy is threaded with filamentary structures on every length scale, from a few to a hundreds light-years, revealing the "skeleton" of Milky Way. These filaments appear to have a universal width of about one third of a light-year, and they seem to be the preferred sites for stellar birth.

- **Latitude**

  In spherical coordinates (r, θ, ϕ), the **latitude** is the **angular distance** δ from the *xy* plane (*fundamental plane*) to a point, measured from the origin; $\delta = 90^\circ - \theta$, where θ is the **colatitude**.

  In a *geographic coordinate system* (or *earth-mapping coordinate system*), the **latitude** is the angular distance from the Earth's equator to an object, measured from the center of the Earth. Latitude is measured in degrees, from -90° (South Pole) to $+90^\circ$ (North Pole). *Parallels* are the lines of constant latitude. The **colatitude** is the angular distance from the Earth's North Pole to an object.

  The **celestial latitude** is an object's latitude (measured in degrees) on the celestial sphere from the intersection of the fundamental plane with the celestial sphere in a given *celestial coordinate system*. In the *equatorial coordinate system* the fundamental plane is the plane of the Earth's equator; in the *ecliptic coordinate system* the fundamental plane is the plane of the ecliptic; in the *galactic coordinate system* the fundamental plane is the plane of the Milky Way; in the *horizontal coordinate system* the fundamental plane is the observer's horizon.

  Geomagnetic latitude is a parameter analogous to geographic latitude, except that bearing is with respect to the magnetic poles. The intersection between the magnetic and rotation axes of the Earth is located $\approx$ 500 km North from its centre.

- **Longitude**

  In spherical coordinates (r, θ, ϕ), the **longitude** is the **angular distance** ϕ in the *xy* plane from the *x* axis to the intersection of a great circle, that passes through the point, with the *xy* plane.

  In a *geographic coordinate system* (or *Earth-mapping coordinate system*), the **longitude** is the angular distance measured eastward along the Earth's equator from the *Greenwich meridian* (or *Prime meridian*) to the intersection of the meridian that passes through the object. Longitude is measured in degrees, from 0° to 360°. A *meridian* is a great circle, passing through Earth's North and South Poles; the meridians are the lines of constant longitude.

  The apparent longitude shift, 102 m east, at Greenwich is a local effect due to the direction of gravity there, and not a global shift in the longitude system.

  The **celestial longitude** is the longitude of a celestial object (measured in units of time) on the celestial sphere measured eastward, along the intersection of the fundamental plane with the celestial sphere in a given *celestial coordinate system*, from the chosen point. In the *equatorial coordinate system* the fundamental plane is the plane of the Earth's equator; in the *ecliptic coordinate system* it is the plane of the ecliptic; in the *galactic coordinate system* it is the plane of the Milky Way; and in the *horizontal coordinate system* it is the observer's horizon.

[redacted]

- **Declination**

  In the *equatorial* (or *geocentric*) *coordinate system*, the **declination** δ is the **celestial latitude** of a celestial object on the celestial sphere, measured from the celestial equator. Declination is measured in degrees, from $-90°$ to $+90°$.

- **Right ascension**

  In the *equatorial* (or *geocentric*) *coordinate system*, fixed to the stars, the **right ascension** *RA* is the **celestial longitude** of a celestial object on the celestial sphere, measured eastward along the celestial equator from the First point in Aries to the intersection of the hour circle of the celestial object. RA is measured in units of time with 1 h approximately equal to 15°.

  The time needed for one complete cycle of the precession of the equinoxes is called a *Platonic* (or *Great*) *year*); it is 257–258 centuries and slightly decreases. This cycle is important in Astrology. Also, it is close to the Maya calendar's longest cycle—5 *Great Periods* of 5125 years; cf. **distance numbers** in Chap. 29.

  The time (225–250 million Earth years) it takes the Solar System to revolve once around the center of the Milky Way (*Solar circle*) is called the *Galactic year*.

- **Hour angle**

  In the *equatorial* (or *geocentric*) *coordinate system*, fixed to the Earth, the **hour angle** HA is the **celestial longitude** of a celestial object on the celestial sphere, measured along the celestial equator from the observer's meridian to the intersection of the hour circle of the celestial object.

  HA gives the time elapsed since the celestial object's last transit at the observer's meridian (for $HA > 0$), or the time until the next transit (for $HA < 0$).

- **Polar distance (in Geography)**

  In the *equatorial* (or *geocentric*) *coordinate system*, the **polar distance** (or *codeclination*) *PD* is the **colatitude** of a celestial object, i.e., the **angular distance** from the celestial pole to a celestial object on the celestial sphere. Similarly as the **declination** δ, it is measured from the celestial equator: $PD = 90° \pm \delta$. An object on the celestial equator has $PD = 90°$.

- **Ecliptic latitude**

  In the *ecliptic coordinate system*, the **ecliptic latitude** is the **celestial latitude** (in degrees) of a celestial object on the celestial sphere from the ecliptic.

  The object's **ecliptic longitude** is its **celestial longitude** on the celestial sphere measured eastward along the ecliptic from the First point in Aries.

- **Zenith distance**

  In the *horizontal* (or *Alt/Az*) *coordinate system*, the **zenith distance** (or *North polar distance*, *zenith angle*) *ZA* is the object's **colatitude**, measured from the zenith.

- **Altitude**

  In the *horizontal* (or *Alt/Az*) *coordinate system*, the **altitude** *ALT* is the **celestial latitude** of an object from the horizon. It is the complement of the **zenith distance** *ZA*: $ALT = 90° - ZA$. Altitude is measured in degrees.

- **Azimuth**

  In the *horizontal* (or *Alt/Az*) *coordinate system*, the **azimuth** is the **celestial longitude** of an object, measured eastward along the horizon from the North point. Azimuth is measured in degrees, from 0° to 360°.
- **Morning width**

  Morning (or *rise*) **width** is the horizontal angular distance between the rise azimuth of a celestial body and the easterly direction.
- **Elliptic orbit distance**

  The **elliptic orbit distance** is the distance from a body in an elliptical orbit around a central body at a focus to that body. This distance is given by

$$r(\theta) = \frac{a(1 - e^2)}{1 + e\cos\theta} = a(1 - e\cos E),$$

where a, b are the *semimajor* and *semiminor* axes (halves of the major and minor diameters), e is the *eccentricity* $\sqrt{1 - (\frac{b}{a})^2}$, and θ, E are the *true anomaly* (angle between the direction of periapsis and the current position of the body, as seen from the central body) and the *eccentric anomaly* $E = \arccos \frac{e+\cos\theta}{1+e\cos\theta}$. The *mean anomaly* is $F = 1 - e\sin E$. The **orbital distance** is mean $r(\theta)$ over E.

The **periapsis distance** and **apoapsis distance** are the closest and farthest distances $r_{\min} = r(0) = a(1 - e)$ and $r_{\max} = r(\pi) = a(1 + e)$.

For Earth's orbit, $e = 0.01671$; it wire the points at periapsis and apoapsis distances of an elliptical orbit around the Earth, while the *perihelion* and *aphelion* are such points around the Sun. The *periastron* and *apastron* of a double star are the closest and farthest points of the smaller star to its primary. ll reach a minimum, 0.00236, in 0.027 Ma.

The *near-Earth objects* are asteroids, comets, spacecraft, and large meteoroids whose *perigee* is closer than 1.3 AU; the largest such asteroids are 1036 Ganymede and 433 Eros, about 34 km across. The *perigee* and *apogee* are the points at periapsis and apoapsis distances of an elliptical orbit around the Earth, while the *perihelion* and *aphelion* are such points around the Sun. The *periastron* and *apastron* of a double star are the closest and farthest points of the smaller star to its primary.

In addition to a, e and F, the other three main parameters required to uniquely identify a *Kepler orbit* (ellipse, parabola, hyperbola, or straight line) are the *orbital inclination* (angle between reference and orbital planes), the *longitude of the ascending node* (angle Ω from a reference direction to the direction of the ascending node, measured in a reference plane), and the *argument of periapsis* (angle ω from the body's ascending node to its periapsis, measured in the direction of motion).

Sedna, 2012VP113, and other distant detached Kuiper Belt objects exhibit unusual alignment in orbital planes and clustering around $\omega = 0$ if $a > 150$ AU and $\Omega = 113° \pm 13°$ if $a > 250$ AU. Batygin–Brown, 2016, explain it by *resonant coupling* with a hypothetical super-Earth, whose orbit lies in about the

[redacted]

same plane, but perihelion is 180° away from the perihelia of the minor bodies. This *Planet Nine* should have perihelion 200 AU and aphelion 1200 AU; so, $a = 700$ AU.

- **Minimum orbit intersection distance**

  The **minimum orbit intersection distance** (MOID) between two bodies is the distance between the closest points of their gravitational *Kepler orbits*.

  An asteroid or comet is a *potentially hazardous object* (PHO) if its Earth MOID is less than 0.05 AU and its diameter is at least 150 m. Impact with a PHO occurs on average around once per 10,000 years. The only known asteroid whose hazard could be above the background is 1950 DA (of mean diameter 1.2 km) which can, with probability $\frac{1}{300}$, hit Earth on March 16, 2880. Object up to about 25 m in size burn up in the atmosphere, and even the 50 to 80 m objects disintegrate before reaching Earth, although they can create a damaging airburst. The closest known geocentric distance for a comet was 0.0151 AU (Lexell's comet on July 1, 1770).

- **Impact distances**

  After an impact event, the falling debris forms an *ejecta blanket*, i.e., a generally symmetrical apron of ejecta that surrounds crater. About half the volume of ejecta falls within 2 radii from the center of the crater, and over 90 % falls within $\approx$ 5 radii. Beyond it, the debris are discontinuous and are called *distal ejecta*.

  Main parameter of an impact crater is the ratio of *rim-to-floor depth d* to the *rim-to-rim diameter D*. The *simple craters* are small with $\frac{1}{7} \leq \frac{d}{D} \leq \frac{1}{5}$ and a smooth bowl shape. If $D > D_0$, where the *transitional diameter* D_0 scales as the inverse power of the planet's surface gravity, the initially steep crater walls collapse gravitationally downward and inward, forming a complex structure. On Earth, $2 \leq D_0 \leq 4$ km depending on target rock properties; on the Moon, $15 \leq D_0 \leq 20$ km.

  The largest known (diameter of 300 km) and old (2023 Ma ago) *astrobleme* (meteorite impact crater) is Vredefort Dome, 120 km south-west of Johannesburg. It was the world's greatest known single energy release event and largest asteroid known to have impacted the Earth ($\approx$ 10 km). The diameter of MAPCIS crater in Australia is 600 km, but it is not confirmed impact crater.

  Sometimes, the term *impact distance* is used more generally as a **setback distance** from some possible hazard (say, explosion, toxic chemical release, odor from swine facilities) or from the action of some equipment (say, laser, homogenizer); Cf. **standoff distance** and **protective action distance** in Chap. 29.

- **Elongation**

  Elongation (or *digression*) is the angular distance in longitude of a celestial body from another around which it revolves (usually a planet from the Sun).

- **Lunar distance**

  The **lunar distance** is the **angular distance** between the Moon and another celestial object.

  In Astronomy, *new moon* (or *dark moon*) is a lunar phase that occurs at the moment of conjunction in **ecliptic longitude** with the Sun. If, moreover, the Sun, Moon, and Earth are aligned exactly, a solar eclipse occurs. *Full moon* occurs when the Moon is on the opposite side of the Earth from the Sun. If, moreover, the Sun, Earth, and Moon are aligned exactly, a lunar eclipse occurs.

A *supermoon* (or *perigee-syzygy of the Earth-Moon-Sun system*) is the near-coincidence of a full moon or a new moon with the closest approach the Moon makes to the Earth on its orbit, resulting in its largest apparent size.

- **Sun-Earth-Moon distances**

  The Sun, Earth and Moon have masses 1.99×10^{30}, 5.97×10^{24}, 7.36×10^{22} kg and equatorial radii 695,500, 6378, 1738 km, respectively.

  Earth's axial tilt varies 22.1–24.5° about every 41,000 years, its rotation occurs about every 19,000 years and eccentricity cycles 0.003–0.058 about every 0.1 Ma.

  The Earth and the Moon are at a mean distance of 1 AU $\approx 1.496 \times 10^8$ km from the Sun. This distance increases at the present rate $\approx$ 15 cm per year.

  The Moon, at distance 0.0026 AU ($\approx$ 60 Earth radii $R_\oplus$), is within the **Hill radius** (1,496,000 km) of the Earth, but well outside of the **Roche radius** (9496 km).

  Asimov argued that the Earth-Moon system is a double planet because their diameter and mass ratios ($\approx$4:1 and $\approx$81:1) are smallest for a planet in the Solar System. Also, the Sun's gravitational effect on the Moon is more than twice that of Earth's. But the *barycenter* (common center of mass) of the Earth and Moon lies well inside the Earth, $\approx \frac{3}{4}$ of its radius.

  The Moon has a greater tidal influence on the Earth than the Sun. Because of tidal forces, the Moon is receding from the Earth at $\approx$ 3.8 cm per year. So, Earth's rotation is slowing, and Earth's day increases by $\approx$ 23 s every million years (excluding glacial rebounds). At present rate, the Moon's orbital distance will reach, $\approx$ 1 Ga from now, $67R_\oplus$, and Earth's axial tilt will become chaotic.

- **Opposition distance**

  A *syzygy* is a straight line configuration of three celestial bodies A, B, C. Then, as seen from A, B and C are in *conjunction*, and the passage of B in front of C is called *occultation* if the apparent size of B is larger, and *transit*, otherwise. *Appulse* is the closest approach of B and C as seen from A.

  If B and C are planets orbiting the star A, then C said to be in *opposition* to A, and the distance between B and C (roughly, their closest approach) is called their **opposition distance**. It can vary at different oppositions.

  A *greatest conjunction* is a conjunction of Jupiter and Saturn at or near their opposition to the Sun. The "Star of Bethlehem"—thought by Kepler to have appeared c. 7 BC—was theorized to be such triple conjunction and almost occultation of Saturn by Jupiter. The next such occultation will happen in 7541.

  The closest possible distance between Earth and a planet is 38 million km: the minimal opposition distance with Venus. The closest known approach between exoplanets is 0.0016 AU $\approx$ 240,000 km (by Kepler-70b and Kepler-70c).

  The closest known distance between two stars is 80,000 km in the binary HM Cancri; their orbital period is 5.4 min. The orbital period of exoplanet 2MASS J2126-8140 around a red dwarf star TYC 9486-927-1 takes nearly a million Earth

[redacted]

years. The largest and smallest known orbits of a planet around a single star are 6900 AU (by 2MASS J2126-8140) and 0.006 AU (by Kepler-70b).

- **Planetary aspects**

  In Astrology, an *aspect* is an angle (measured by the **angular distance** of ecliptic longitude, as viewed from Earth) the planets make to each other and other selected points in the *horoscope*, i.e., a chart representing the apparent positions and selected angles of the celestial bodies at the time of an event, say, a person's birth. Astrology claims a link between aspects and events in the human world.

  Major aspects are 1–10° (*conjunction*) and 90° (*square*), 180° (*opposition*) for which an *orb* (error) of 5–10° is usually allowed. Then follow $120 \pm 4°$ (*trine*), $60 \pm 4°$ (*sextile*) and (with orb 2°) 150° (*quincunx*), 45°, 135°, 72°, 144°. Other aspects are based on the division of the zodiac circle by 7, 9, 10, 11, 14, 16 or 24.

- **Primary-satellite distances**

  Consider two celestial bodies: a *primary* M and a smaller one m (a satellite, orbiting around M, or a secondary star, or a comet passing by).

  Let ρ_M, ρ_m and R_M, R_m be the densities and radii of M and m. The **Roche radius** (or *Roche limit*, *tidal radius*) of the pair (M,m) is the maximal distance between them within which m will disintegrate due to the tidal forces of M exceeding the gravitational self-attraction of m. This distance is $\approx 1.26 R_M \sqrt[3]{\frac{\rho_M}{\rho_m}}$ or $\approx 2.423 R_M \sqrt[3]{\frac{\rho_M}{\rho_m}}$ if m is rigid or fluid. The *Roche lobe* of a star is the region of space around the star within which orbiting material is gravitationally bound to it.

  The *tidal locking radius* of M is the distance at which the axial and orbital rotations of m become synchronized, i.e., the same side of m always faces M. The Moon is tidally locked by the Earth. Pluto and Charon are mutually tidally locked.

  Let $d(m,M)$ denote the **mean distance** between m and M, i.e., the arithmetic mean of their maximum and minimum distances; let S_m and S_M denote the masses of m and M. The barycenter of (M,m) is the point (in a focus of their elliptical orbits) where M and m balance and orbit each other. The distance from M to the barycenter is $d(m,M)\frac{S_m}{S_m+S_M}$. For the $(Earth, Moon)$ system, it is 4670 km (1710 km below the Earth's surface). Pluto and Charon, the largest of its five moons, form rather a *binary planet*, since their barycenter lies outside of either body.

  The *Hill sphere* of a body is the region in which it dominates the attraction of satellites. The **Hill radius** of m in the presence of M is $\approx d(m,M)\sqrt[3]{\frac{S_m}{3S_M}}$; within it m can have its own satellites. The Earth's Hill radius is 0.01 AU; in the Solar System, Neptune has the largest Hill radius, 0.775 AU.

  The pair (M,m) can be characterized by five **Lagrange points** L_i, $1 \leq i \leq 5$, where a third, much smaller body (say, a spacecraft) will be relatively stable because its centrifugal force is equal to the combined gravitational attraction of M and m, i.e., gravitational force of M and m balance out. These points are:

  L_1, L_2, L_3 lying on the line through the centers of M and m, so that $d(L_3,m) = 2d(M,m)$, $d(M,L_2) = d(M,L_1) + d(L_1,m) + d(m,L_2)$, $d(L_1,m) = d(m,L_2)$ (the satellite SOHO is at the point L_1 of the Sun-Earth system, where the view of the Sun is uninterrupted; the satellites WMAP and Planck are at L_2);

L_4 and L_5 are the points forming equilateral triangles with the centers of M and m. These points are more stable; each of them forms with M and m a partial solution of the unsolved gravitational 3-*body problem*. Objects orbiting at L_4 or L_5 are called *Trojans of Greek* and *Trojan camp*, respectively. The Moon was created 4.5 Ga ago by head-on impact of a Mars-sized Trojan planetoid Theia on the Earth. The only known Sun-Earth Trojan is asteroid 2010 TK7, about 300 m across.

Other instances of the circular restricted 3-body problem are provided by planet–*co-orbital moons* and star–planet–*quasi-satellite* systems. *Co-orbital moons* are natural satellites that orbit at a very similar distance from their parent planet. Only Saturn's system is known to have them; it has three sets.

Orbital resonance occurs when the bodies orbital periods are in a close-to-integer ratio. For example, Pluto-Neptune are in a 2:3 ratio and Jupiter's moons Ganymede-Europa-Io are in a 1:2:4 ratio. Earth and Venus are in a *quasi-resonance* only 0.032 % away from 8:13. A *quasi-satellite* is an object in a 1:1 orbital resonance with its planet that stays close to the planet over many orbital periods. The largest of four known Earth's quasi-satellites is 3753 Cruithne, ≈ 5 km across.

The most tenuously linked long-distance binary in the Solar System is 2001 QW322: two icy bodies ($\approx$ 130 km in diameter) in the Kuiper Belt, at mean distance $> 10^5$ km, orbiting each other at 3 km/h. One of widest binaries is the Alpha Centauri system with Proxima separated from the A/B stars by 15,000 AU.

The elliptic restricted 3-body problem treats the *circumbinary* (orbiting two stars) planets such as Kepler-16b. A planet PH1 was found in a quadruple (binary pair) star system Kepler-64. Systems with up to 5 stars are known; most are triples.

- **Low-energy trajectory**

The *Hohmann transfer orbit* is an elliptical orbit used to transfer between two circular orbits of different radii in the same plane. Currently, it is the favored method for getting a spacecraft into orbit around its destination. Leaving Earth's atmosphere, the craft makes, at high speed, a beeline for the destination; at approach, it slams on the brakes and starts to orbit the destination. Such a transfer is a highly effective, but it is time-specific and expensive, since it requires much energy.

A **low-energy** (or *ballistic capture*) **trajectory** is a route in space that allows a spacecraft to change orbits using very little fuel. In such transfer to a planet, the craft is launched out ahead of the planet's orbital path. It would gradually slow and hold in place, waiting for the planet to swing by—the gravity of the planet pulls it into its orbit. Such transfers are very slow, and useful only for automated probes.

Low-energy transfers follow the *Interplanetary Transport Network* (*ITN*), i.e., a collection of gravitationally determined pathways through the Solar System that

[redacted]

require very little energy (fuel) for an object to follow. The ITN uses **Lagrange points** as locations where trajectories are redirected using little or no energy.

- **Dynamical spacing**

  Let M, m_1, m_2 be the masses of a star and two adjacent planets orbiting it with semimajor axes a_1 and a_2. The *mutual Hill radius* of two planets is

$$R_H = \frac{a_1 + a_2}{2}\sqrt[3]{\frac{m_1 + m_2}{3M}}$$

  and their **dynamical spacing** is (Gladman, 1993; Chambers et al., 1996)

$$\Delta = \frac{|a_2 - a_1|}{R_H}$$

  .

  Fang–Margot, 2013, claim that on average $\Delta = 21.7$, and $\Delta < 10$ leads to instability in giga-year time span. In the Solar System, $\Delta > 26$ for terrestrial planets.

- **Titius–Bode law**

  The **Titius–Bode law**, 1766, is an empirical rule approximating the mean distance d_i of i-th planet from the Sun (its orbital *semimajor axis*) by $\frac{3k+4}{10}$ AU.

  Here 1 AU $\approx 1.5 \times 10^8$ km and $k_1 = 0 = 2^{-\infty}$ (for Mercury), $k_i = 2^{i-2}$ for $i \geq 2$, i.e., Venus, Earth, Mars, Ceres (the largest one in the Asteroid Belt, $\approx \frac{1}{3}$ of its mass), Jupiter, Saturn, Uranus, Pluto. (But Neptune does not fit in the law while Pluto fits Neptune's spot $k = 2^7$.) The best fit for the form (Wurm, 1787) $d_i = AC^{i-2} + B$ is given by $C \approx 1.925, A \approx 0.334, B = 0.382$. Cf. **elliptic orbit distance**.

  In the Solar System, the period ratios between adjacent orbits scatter around the dominant 5 : 2 ratio; it is 3 : 2 for Earth-Venus and 2 : 1 for Mars-Earth.

  A *generalized Titius–Bode relation* $d_i = AC^i$ for some A, C fits even better for many other exoplanet systems showing such preference towards near mean motion resonance; cf. **dynamical spacing**. It helps to locate undetected exoplanets.

  Hamano et al., 2013, claim that between 108 (as Venus) and 150 (as Earth) million km from the Sun, there is a critical distance explaining their difference. Venus has a similar size and bulk composition to those of Earth, but it lacks water. Earth solidified from its molten magma state within several million years, trapping water in rock and under its hard surface, while Venus got more of the Sun's heat and remained in molten state for ≈ 100 Ma giving time for any water to escape.

- **Planetary distance ladder**

  The scale of interstellar-medium dust, *chondrules* (round grains found in stony meteorites, the oldest solid material in the Solar System), *boulders* (rock with grain size of diameter ≥ 256 mm), *planetesimals* (kilometer-sized solid objects in protoplanetary disks) and *protoplanets* (internally melted Moon-to-Mars-sized planetary embryos) is 10^{-6}, 10^{-3}, 10^{0}, 10^{3} and 10^{6} m.

In the Solar System's protoplanetary gas/dust disk, the binary electrostatic coagulation of dust/ice grains resulted in the creation, of planetesimals. Then gravity took over the accretion process. The growth was *runaway* (when $T_1 < T_2$, for growth time scales of the first and second most-massive bodies) at first and then (with $T_1 > T_2$ at some *transition radius*) it became *oligarchic*. A few tens of protoplanets were formed and then, by giant impacts, they were transformed into Earth and the other rocky planets. The process took $\approx$ 90 Ma from $\approx$ 4.57 to $\approx$ 4.48 Ga ago.

Brilliantov et al., 2015: if a system is comprised of particles that, colliding, can merge at low velocities and break at large velocities, the size distribution of particles abundance follow the inverse cubes law. Rare collisions of fast particles lead to a balance between aggregation and fragmentation. It explained Saturn's rings.

- **Potato radius**

The basic shape-types of objects in the Universe are: an irregular dust, rounded "potatoes" (asteroids, icy moons), spheres (planets, stars, black holes), disks (Saturn's rings, galactic disks) and halos (elliptic galaxies, globular star clusters).

At mean radius $R <$ few km, objects (dust, crystals, life forms) have irregular shape dominated by nonmass-dependent electronic forces. Solid objects with $R >$ 200–300 km are gravity-dominated spheres. If both energy E and angular momentum L are exported (by some dissipative processes), the object, if large enough, collapses into a sphere. If only E is exported, the shape is a disk. If neither E, nor L is exported, the shape is a *halo*, i.e., the body is spheroidal.

If R ($R >$ few km) increases, there is a smooth size-dependent transition to more and more rounded potatoes until $\approx$200–300 km, where gravity begins to dominate. Ignoring surface tension, erosion and impact fragmentation, the potato shape comes mainly from a compromise between electronic forces and gravity. It also depends on the density and the structural yield strength of the (rocky or icy) material.

Lineweaver and Norman, 2010, define the **potato radius** R_{pot} as this potato-to-sphere transition radius. They derived $R_{pot} = 300$ km for asteroids (Vesta, Pallas, Ceres have mean radius $R = 265, 275, 475$ km, respectively) and $R_{pot} = 200$ km for icy moons (Hyperion, Mimas, Miranda have $R = 135, 198, 235$ km, respectively). Mimas is the smallest body in the Solar System to be rounded, while Hyperion is the largest nonspherical one; their masses are 3.75×10^{19} and 5.62×10^{18} kg.

In 2006, the IAU (International Astronomical Union) defined a *planet* as a orbiting body which has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a *hydrostatic equilibrium* (nearly round) shape and cleared the neighborhood around its orbit. If the body has not cleared its neighborhood, it is called a *dwarf planet*. The potato radius, at which self-gravity makes internal overburden pressures equal to the yield strengths of the material, marks the boundary of hydrostatic equilibrium used in above IAU definition.

[redacted]

Buchhave et al., 2014: planets smaller than 1.7 Earths are likely to be completely rocky, while those larger than 3.9 Earths are probably gas giants.

- **Frost line (in Astrophysics)**

  In Astrophysics, by analogy with **frost line (in Earth Science)**, the **frost** (or *snow*, *ice*) **line** is the distance from a star (or a nebula's protostar) where hydrogen compounds such as water, ammonia, methane condense into ice grains. It separates an inner region of rocky objects from an outer region of icy objects.

  Water and methane condensate at 180 K and 40 K, respectively. Sun's water-frost and methane-frost lines are roughly at 2.7 AU and 48 AU, i.e., in the Asteroid Belt (between the orbits of Mars and Jupiter) and the Kuiper Belt (at 30–55 AU). On the other hand, inside of $\approx$ 0.1 AU, rocky grains cannot exist: dust evaporates.

  Martin and Livio, 2012, claim that a giant planet like Jupiter should be in the right location outside of the frost line to produce an asteroid belt of the appropriate size, offering the potential for life on a nearby rocky planet like Earth.

- **Solar distances**

  Following a supernova explosion 4570 Ma ago in our galactic neighborhood, the Sun was formed 4567 Ma ago by rapid gravitational collapse of a fragment (about 1 parsec across) of a giant (about 20 parsecs) hydrogen molecular cloud.

  The mean distance of the Sun from Earth is 1 AU $\approx 1.496 \times 10^8$ km. The mean distance of the Sun from the Milky Way core is 27,200 light-years.

  The Sun is more massive than 95 % of nearby stars and its orbit around the Galaxy is less eccentric than $\approx$ 93 % of similar (i.e., of spectral types F, G, K) stars within 40 parsecs. The Sun's mass (99.86 % of the Solar System) is $\approx 1.988 \times 10^{30}$ kg.

  The Sun's radius is 6.955×10^5 km; it is measured from its center to the edge of the *photosphere* ($\approx$ 500 km thick layer below which the Sun is opaque to visible light). The Sun will expand $\approx$ 256 times in 7–8 Ga and then become a white dwarf.

  The Sun does not have a definite boundary, but it has a well-defined interior structure: the *core* extending from the center to $\approx$ 0.2 solar radii, the *radiative zone* at $\approx$0.2–0.8 solar radii, where thermal radiation is sufficient to transfer the intense heat of the core outward, the *tachocline* (transition layer) and the *convection zone*, where thermal columns carry hot material to the surface (photosphere) of the Sun.

  The principal zones of the *solar atmosphere* (the parts above the photosphere) are: temperature minimum, chromosphere, transition region, corona, and heliosphere.

  The *chromosphere*, a $\approx$ 3000 km deep layer, is more visually transparent. The *corona* is a highly rarefied region continually varying in size/shape; it is visible only during a total solar eclipse. The chromosphere-corona region is much hotter than the Sun's surface. Extending further, the corona becomes the *solar wind*, a very thin gas of charged particles that travels through the Solar System.

  The *heliosphere* is the teardrop-shaped region around the Sun created by the solar wind and filled with solar magnetic fields and outward-moving gas.

It extends from ≈ 20 solar radii (0.1 AU) outward 86–100 AU past the orbit of Pluto to the *heliopause*, its outermost edge, where the interstellar medium and solar wind pressures balance. The interstellar medium and solar wind are moving supersonically in opposite directions, towards and away from the Sun. The points, ≈ 80 AU and ≈ 230 AU from the Sun, where the solar wind and interstellar medium become subsonic, are the *termination shock* and *bow shock*, respectively.

The *tidal truncation radius* (100,000–200,000 AU, say, ≈ 2 ly from the Sun) is the outer limit of the *Oort Cloud*. It is the boundary of the Solar System, i.e., Sun's *Hill/Roche sphere*, where its gravity is overtaken by the galactic tidal force.

- **Dyson radius**

  The **Dyson radius** of a star is the radius of a hypothetical *Dyson sphere* around it, i.e., a megastructure (say, a system of orbiting star-powered satellites) meant to completely encompass a star and capture a large part of its energy output. The solar energy, available at distance d (measured in AU) from the Sun, is $\frac{1366}{d^2}$ watts/m^2. The inner surface of the sphere is intended to be used as a habitat.

  For example, at Dyson radius 300×10^6 km from the Sun a continuous structure with ambient temperature 20 °C (on the inner surface) and efficiency 3 % of power generation (by a heat flux to −3 °C on the outer surface) is conceivable.

- **Star's radii**

  The **corotation radius** of a star is the distance from it where the centrifugal force on a particle corotating with it balances the gravitational attraction, i.e., the accretion disk rotates at the same angular velocity as the star.

  The **Bondi-Hoyle accretion radius** is the radius where star's gravitational energy is larger than the kinetic energy and, so, at which material is bound to star.

  The **Hayashi radius** (or *Hayashi limit*) of a star is its maximum radius for a given mass. A star within *hydrostatic equilibrium* (where the inward force of gravity is matched by the outward pressure of the gas) cannot exceed this radius.

  The **Eddington radius** (or *Eddington limit*) of a star is the radius where the gravitational force inwards equals the continuum radiation force outwards, assuming hydrostatic equilibrium and spherical symmetry. A star exceeding it would initiate a very intense continuum driven stellar wind from its outer layers.

  The largest and smallest known stars, the red hypergiant UY Scuti and red dwarf OGLE-TR-122b, have respective radii 1708 ± 192 and 0.12 solar radii.

- **Galactocentric distance**

  A star's **galactocentric distance** (or *galactocentric radius*) is its **range distance** from the galactic center; it may also refer to a distance between two galaxies. The *galactic anticenter* is the point lying opposite, for an observer on Earth, this center.

  The Sun's present galactocentric distance is nearly fixed ≈ 8.4 kiloparsec, i.e., 27,400 light-years, but it may have been 2.5–5 kpc in the past. *Einasto's*

[redacted]

law, 1963, claims that the density $\rho(r)$ of a spherical stellar system (say, a galaxy or its halo) varies as $\exp(-Ar^{\alpha})$, where r is the distance from the center.

- **M31–M33 bridge**

  Braun and Thilker, 2004, discovered that the distance 782,000 light-years between Andromeda (M31) and Triangulum (M33) galaxies is spanned by a link consisting of about 500 million Sun's masses of ionized hydrogen.

  A third of all baryonic matter is in stars and galaxies; another $\frac{1}{3}$ is diffuse and thought to be in filamentary networks spread through space. Remaining $\frac{1}{3}$, called *warm-hot intergalactic medium* (WHIM), is expected to be of intermediate density. The **M31–M33 bridge** consists of WHIM, the first evidence of this medium. Such WHIM bridges are likely remnants of collisions between galaxies.

- **Radii of a star system**

  Given a star system, say, a galaxy or a *globular cluster* (a spherical collection of stars that orbits a galactic core as a satellite), its **half-light radius** (or **effective radius**) hr is the distance from the core within which half the total luminosity from the system, assumed to be circularly symmetric, is received. The **core radius** cr is the distance at which the apparent surface luminosity has dropped by half; so, $cr \leq hr$. If the luminosity profile rises steeply to a break, and then more slowly, we have the **break radius**. In general, **isophotal radius** is the size attributed to the system corresponding to a particular level of surface brightness.

  The **half-mass radius** $r_{0.5}$ is the radius from the core that contains half the total mass of the system. In general, the *Lagrangian radii* are the distances from the center at which various percentages of the total mass are enclosed.

  The **tidal radius** of a globular cluster is the distance from its center at which the external gravitation of the galaxy has more influence over the stars in the cluster than does the cluster itself.

  Unlike a star cluster, all galaxies are filled with and surrounded by a halo of dark matter acting (via gravity) as a sort of glue within and between galaxies. Thin tendrils of dark matter connect nodes of galaxy clusters, creating a cosmic web.

  The **virial radius** R_{vir} of a galaxy is the radius centered on it containing matter at 200 (sometimes, $18\pi^2 \approx 178$ or 130) times the critical (or, mean) density of the Universe. The mass within R_{vir} is a measure of the total mass inside a dark matter halo. Kravtsov, 2011, claim that $r_{0.5} \approx 0.015R_{vir}$. Also, Harris, 2013, explains speed anomalies of Earth's satellites by 0.005–008 % increase of its mass due to a dark matter's disk, 191 km thick and 70,000 km across, around the equator.

  The **Abell radius** of a cluster is $\frac{1.72}{z}$ arcminutes, where z is the cluster's redshift. At least 50 cluster's members should lie within this radius of its centre.

- **Habitable zone radii**

  A maximally Earth-like mean temperature is expected at the distance $\sqrt{\frac{L_{star}}{L_{sun}}}$ AU from a star, where L is the total radiant energy.

  The **habitable zone radii** of a star are the minimal and maximal orbital radii r, R such that liquid water may exist on a *terrestrial* (i.e., primarily composed of silicate rocks or metals) planet orbiting within this range, so that life, constructed from carbon and reliant on liquid water, could develop there in a similar way as on the early Earth. For Sun, $[r,\ R]$ is $[0.99,\ 1.70]$; AU; it includes Earth and Mars. It moves outwards, according to Rushby et al., 2012, 0.1 Au every billion years. For best candidates—orange dwarf stars—HZ is within $[0.5,\ 1]$ AU.

The **Kasting distance** (or *habitable zone distance*) of an exoplanet, at distance d from its star, is an index defined by

$$HZD(d) = \frac{2d - (R + r)}{R - r}.$$

So, $-1 \leq HZD(d) \leq 1$ correspond to $r \leq d \leq R$.

The above notion of surface habitability is modeled from temperature/humidity; the edges r, R of HZ are determined by loss of water and, respectively, by the maximum greenhouse provided by a CO_2 atmosphere. Among known exoplanets of 2–10 Earth's mass of our galaxy, good candidates for Earth-like habitability are GJ 667Cc, HD-85512b, Kepler-438b (24, 36, 470 ly away) orbiting, respectively, stars Gliese 667C, Gliese 370, Kepler-438 in the constellations Scorpius, Vela, Lyra.

Habitable zones are 10–14 times wider for subsurface life. Protected inside a warm mineral-rich rocks, it can be much more typical than Earth's surface life.

Instead of zones, habitability can be defined in terms of, say, geothermal vents on ocean floor, as on Saturn's moon Enceladus.

Petigra et al., 2013: *eta-Earth* of our galaxy, i.e., fraction of Sun-like stars with an *Earth-size* (1–2 Earth's radii) planet orbiting in habitable zones, is $22 \pm 8\,\%$, depending on the definition of HZ. There could be 4×10^{10} habitable Earth-size planets in the Milky Way with the nearest one being within 12 ly.

Our *galactic habitable zone* is (Lineweaver et al., 2004) a slowly expanding region between 7 and 9 kpc of **galactocentric distance**; so, the minimal and maximal radii are 22,000 and 28,000 ly. They used four prerequisites for complex life: the presence of a host star, enough heavy elements to form terrestrial planets, sufficient time (4 ± 1 Ga) for biological evolution and an supernovae-free environment.

Di Stefano et al., 2015, claim that the globular clusters (there are about 160 in Milky Way) could be best places to look for space-faring civilizations.

Most life-friendly are giant elliptical galaxies: their low rate of star formation means less supernovae and gamma bursts. Dwarf irregular galaxies are the least hospitable.

Behroozi–Peeples, 2015, claim that the Hubble volume, ~ 10 Mpc^3, contains $\sim 10^{20}$ Earth-like planets, $\sim 10^{21}$ will be formed in 100 Gyr to 1 Tyr from now and the likehood of us being the only civilisation the Universe will ever have is $< 8\,\%$.

[redacted]

- **Earth similarity index**

  The **Earth similarity index** of a planet P is (Schulze-Makuch et al., 2011):

$$ESI(P) = \prod_{i=1}^{n}(1 - |\frac{x_i(P) - x_i(E)}{x_i(P) + x_i(E)}|)^{\frac{w_i}{n}},$$

  where $x_i(P)$ is a planetary parameter (including surface temperature, escape velocity, mean radius, bulk density), $x_i(E)$ is the reference value for Earth (i.e., 14.85° C, 1,1,1), w_i is a weight (5.58, 0.70, 0.57, 1.07) and n is the number of parameters. ESI(P)$=$ 1, 0.88, 0.85, 0.84, 0.70 for Earth, Kepler-438b, Kepler-296e, GJ 667Cc and Mars. Many exomoons and unconfirmed NASA Kepler candidates rank within $[0.76, 0.96]$. Terrestrial, but only simple extremophilic, life might be possible if ESI(P)$>$ 0.6, while plants/animals may require $>$ 0.8.

  The same authors proposed a *planetary habitability index* based on the presence of a stable substrate, atmosphere, magnetic field, available energy, appropriate chemistry and the potential for holding a liquid solvent, such as 100-km deep ocean beneath the surface of Jupiter's moon Europa and hydrocarbon lakes on Saturn's moon Titan. Unicellular life has been found in the most adverse conditions on Earth. So, the presence of extremophiles on Mars and, with very different biochemistry, on Europa and Titan is plausible. For primary producers (say, plants and cyanobacteria), Earth was more habitable 500 Ma ago, with less seasonal ice and deserts.

  Barnes et al., 2015, habitability index for star-transiting planet, observed by Kepler telescope, is based on orbital period, *transit depth* (planet-star ratio of areas), transit duration and host star's surface gravity, radius and effective temperature.

  Observing oxygen in a planet's atmosphere will indicate photosynthetic life since the photosynthesis is the only known process able to release O_2 in any real quantity. But the importance of oxygen and carbon can be a peculiarity of Earth life. For Oze et al., 2012, low ($<$ 40) hydrogen/methane ratio indicate that life is likely present. Also, infrared, or heat, radiation can indicate an alien civilization.

- **SETI detection ranges**

  SETI (Search for Extra Terrestrial Intelligence) involves using radio telescopes to search for a possible alien radio transmission. The recorded signals are mostly random noise but in 1977 a strong signal (called WOW!) was received at $\leq$ 10 kHz of the frequency $\approx$1420.406 MHz (21 cm) of the hydrogen line. Also, a puzzling radio source SHGb02+14a was observed three times in 2003 at $\approx$1420 MHz.

  SETI detection ranges are the maximal distances over which detection is still possible using given frequency, antenna dish size, receiver bandwidth, etc. They are low for broadband signals from Earth (from 0.007 AU for AM radio up to 5.4 AU for EM radio) but reach 720 light-years for the S-Band of the world's largest (with dish's diameter 305 m) single-aperture radio telescope at Arecibo.

SETI searches in the microwave window 1–10 GHz (the part of the radio spectrum that can pass through the atmosphere), especially around the "waterhole" 1420–1666 MHz (21–18 cm) between hydrogen, H, and hydroxyl, OH.

All known signals with spectral width < 5 Hz arise from artificial sources; so, such extraterrestrial signal will indicate an intelligent civilization. SETI searched those narrow band signals in L-band (1.1–1.9 GHz) from 86 stars in the Kepler field of view hosting most life-promising exoplanets, but not found none. Tarter et al., 2013, deduce from it that the number of *Kardashev type II* (using all energy from their star; our total power consumption today is $\approx 0.01\,\%$ of the sunlight falling on Earth) civilizations in the Milky Way loud in L-band, is less than 1 in a million per sun-like star. The volume V of our galaxy is about $\pi(50{,}000^2)1000 \approx 7.9 \times 10^{12}$ (ly)^3. If N civilizations are distributed there uniformly with spacing d, then $d^3 = \frac{V}{N}$.

Active SETI (or *METI*) consists of sending radio or optical signals into space hoping that they will be picked up by an alien intelligence. The first radio signals from Earth to reach space were produced around 1940 but TV and radio signals decompose into static within 1–2 ly. In 1974 Arecibo telescope sent an elaborate radio signal aimed at the star cluster M13 located 25,000 ly away.

About the perceived risk of revealing the location of the Earth to an alien civilization, METI enthusiasts reply that an advanced civilization within a radius of 100 ly already knows of our existence due to electromagnetic signals leaking from TV, radio and radar. But now, with digital transmissions replacing analogue ones and virtually no radiation escaping into outer space, the Earth become electronically invisible to aliens. Still, a civilization even slightly more advanced than ours could detect the lights of our big cities from up to 500 light years away, using its sun as a gravitational lens. Also, some life (plants, lichens, algae, bacterial mats) can be recognized by its light signature from space.

Besides radio signals and light, nonmicrobial alien life can be discovered by analyzing the output of methane or oxygen in the atmosphere of exoplanets.

- **Voyager 1 distance**

The Voyager 1 is a 722-kg robotic space probe launched by NASA in 1977; it has power to operate its radio transmitters until 2025 but only 69 kB of memory.

It is currently the farthest man-made object from Earth, the first probe to leave the Solar System (in 2012) and the fastest probe (moving at ≈ 17 km/s or 3.6 AU/year). As of January 2016, **Voyager 1 distance** from Earth was 2.1×10^{10} km ≈ 134.4 AU, while for Voyager 2, it was ≈ 110.9 AU.

Voyager 1 is moving now, until at least 2025, through a special region of interstellar space where magnetic fields are rotated away from true magnetic north.

The NASA *Juno* spacecraft to Jupiter holds from 2016 solar power distance record. Its maximum distance from the Sun will be about 832 million km.

Even though the NASA *New Horizons* was launched (in 2006, to see Pluto and Kuiper Belt) far faster, it will never be more distant from Earth than Voyager 1. Close flybys of Saturn and Titan gave Voyager 1 an advantage with their gravity assist. In 2038, New Horizons will reach, at 13 km/s, the distance of 100 AU.

[redacted]

The Earth-Moon distance ($\approx$ 1.28 light-seconds) can be covered, with current technologies, in $\approx$ 8 h. The distance from Earth to other planets ranges from 3 light-minutes to $\approx$ 4 light-hours. At Voyager 1's current rate, a journey to Proxima Centauri (the nearest known star, 4.24 ly away) would take 72,000 years.

Interstellar travel will be possible only with new technology, say, beamed-light sails, hydrogen-fuelled ramjet, nuclear pulse propulsion, warp drive, wormholes.

Human spaceflight beyond the close neighborhood in the Solar System looks, as now, unlikely, because of duration, cost and health threat due to microgravity, radiation and isolation. Also, long (more than a month) sojourns in space produce potentially serious brain anomalies and severe eyesight problems.

- **Earth in space**

The Earth, spinning 0.5 km/s, orbits the Sun at 30 km/s. The Sun orbits the galactic center at 219 km/s and it moves at 16.5 km/s, with respect to the motion of its galactic neighborhood, towards Vega, a star in the constellation Lyra. There are 45 stars within 15 ly of the Solar System, the nearest one at 4.24 ly.

The *Local Bubble* is a cavity, 300–800 ly across (with hydrogen density 0.05 atoms per cm^3, one tenth of the galactic density) in the *Local* (or *Orion-Cygnus*) *Arm* of the Milky Way. The Solar System has been traveling through this Bubble for the last 5–10 Ma and is located now close to its inner rim, about half-way along the Arm's length. From 0.044 to 0.15 Ma ago and for another 0.01–0.02 Ma, the Sun is traversing the *Local Interstellar Cloud* 30 ly across at 23 km/s.

The Milky Way (0.1 Mly across and 1 kly thick) and Andromeda galaxies are 2.5 Mly apart and are approaching at 100–140 km/s. In $4 + 1.3 + 0.1$ Ga (3 consecutive collisions) they will merge to form the *Milkomeda*, new elliptical galaxy in which our Solar System would remain intact but Sun's **galactocentric distance** will be 0.16 Mly. Their stars will not collide but central black holes will merge.

Our *Local Group* (LG) is a *poor* (small and not centered) cluster, 10 Mly across, consisting of Andromeda (M31), Milky Way (MW), Triangulum and about 50 small galaxies. It lies in the outskirts (on a small filament connecting the Fornax and Virgo clusters) of our small Local Supercluster (LSC), 110 Mly across and with a mass 10^{15} suns. The number of galaxies per unit volume, in the LSC, falls off with the square of the distance from its center, near the Virgo cluster.

Tully and Courtois, 2014, defined superclusters (by analogy with drainage basins in Hydrology) via the relative velocities of galaxies: within a given supercluster, all galaxy motions will be directed inward, toward the center of mass. By new definition, LSC, Pavo-Indus Supercluster and Hydra-Centaurus Supercluster, containing *Great Attractor* (GrAt), formtogether "truly local" Supercluster, called *Laniakea*. Laniakea, about of diameter 520 Mly and mass 10^{17} suns, is centered in GrAt, toward which all its galaxies are moving at 630 km/s.

The *Extended Local Group* is the LG plus the "nearby" (3.9 Mpc) Maffei and Sculptor groups. It belongs to our *Local Filament* (LF, or *Coma-Sculptor Cloud*), a branch of the Fornax-Virgo filament of the LSC.

The LF bounds the *Local Void* (LV), extending 60 Mpc from the edge of the LG. The *Local Sheet* (LS) is all LF's matter within 7 Mpc. The Milky Way and Andromeda are encircled by 12 large galaxies arranged in a ring about 24-Mly across.

With respect to the CMBR (cosmic microwave background radiation) filling the Universe almost uniformly, the Solar System, Milky Way, and LG velocities are 369, 600, 627 km/s. *Peculiar velocities* V_{pec} are the deviations from the Hubble expansion, i.e., $V_{pec} = V_{obs} - H_0 d$, where V_{obs} is the observed velocity, d is the distance and H_0 is the Hubble constant, $\approx$ 72 km/s for every Mpc. The Hubble flow, dominating at large distances, is negated by gravity at smaller distances; for example, its recession velocity is < 1 mm/s at the edge of the Solar System.

According to Tully et al., 2007, the Local Sheet is moving as a unit with low internal dispersion; the LG moves at only 66 km/s with respect to the LS. The bulk flow of the LS is sharply discontinuous from the flows of other nearby structures. The vector of this flow has, with respect to the CMBR, amplitude 631 km/s. It can be decomposed into a vector sum of three quasi-orthogonal components: *local* (259 km/s away from the center of the Local Void), *intermediate* (185 km/s to the Virgo cluster) and *large* (455 km/s towards the *Great Attractor* (GrAt)).

All matter within 4.6 Mpc moves away from the Local Void at 268 km/s. It will collide, in $\approx$ 10 Ga, with the nearest adjacent filament, the Leo Spur. The Local Sheet moves toward the Virgo cluster, at the distance 17 Mpc. All matter within 50 Mpc moves at 600 km/s towards overdensities at 200 Mly (GrAt dominated by the Norma cluster) and 600 Mly (Shapley supercluster, roughly behind GrAt). Lebeskind et al., 2015, claim that a dark matter bridge of some 2000 galaxies, bound by voids, stretches from our Local Group to Virgo cluster.

[redacted]

Chapter 26
Distances in Cosmology and Theory of Relativity

26.1 Distances in Cosmology

The *Universe* is defined as the whole space-time continuum in which we exist, together with all the energy and matter within it.

Cosmology is the study of the large-scale structure of the Universe. Specific cosmological questions of interest include the *isotropy* of the Universe (on the largest scales, the Universe looks the same in all directions, i.e., is invariant to rotations), the *homogeneousness* of the Universe (any measurable property of the Universe is the same everywhere, i.e., it is invariant to translations), the density of the Universe, the equality of matter and antimatter, and the origin of density fluctuations in galaxies.

Hubble, 1929, discovered that all galaxies have a positive *redshift*, i.e., all galaxies, except for a few nearby galaxies like Andromeda, are receding from the Milky Way. By the Copernican principle (that we are not at a special place in the Universe), we deduce that all galaxies are receding from each other, i.e., we live in an expanding Universe, and the further a galaxy is away from us, the faster it is moving away (this is now called the *Hubble law*). The *Hubble flow* is the general outward movement of galaxies and clusters of galaxies resulting from the expansion of the Universe. It occurs radially away from the observer, and obeys the Hubble law. The gravitation in galaxies can overcome this expansion, but the clusters and superclusters (largest gravitationally bound objects) only slow the rate of their expansion.

In Cosmology, the prevailing scientific theory about the early development and shape of the Universe is the *Big Bang Theory*. The observation that galaxies appear to be receding from each other, combined with the General Theory of Relativity, leads to the construction that, as one goes back in time, the Universe becomes increasingly hot and dense, then leads to a gravitational singularity, at which all distances become zero, and temperatures and pressures become infinite.

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_26

The term *Big Bang* is used to refer to a hypothesized point in time when the observed expansion of the Universe began. Based on measurements of this expansion, it is currently believed that the Universe has an age of $\approx$ 13.82 Ga (billion years).

In Cosmology (or, more exactly, *astrometry*, the study of the positions, distances between, and motions of astronomical objects) there are many ways to specify the distance between two points, because in the expanding Universe, the distances between comoving objects are constantly changing, and Earth-bound observers look back in time as they look out in distance. But all distance measures somehow measure the separation between events on *radial null trajectories*, i.e., trajectories of photons which terminate at the observer. In general, the **cosmological distance** is a distance far beyond the boundaries of our Galaxy.

The geometry of the Universe is determined by several *cosmological parameters*: the *cosmic scale factor* a, the *Hubble constant* H, the *density* ρ and the *critical density* ρ_{crit} (the density required for the Universe to stop expansion and, eventually, collapse back onto itself), the *cosmological constant* Λ, the *curvature* k *of the Universe*. Many of these quantities are related under the assumptions of a given *cosmological model*. The most common cosmological models are the closed and open *Friedmann–Lemaître cosmological models* and the *Einstein–de Sitter cosmological model*.

This model assumes a homogeneous, isotropic, constant curvature Universe with zero cosmological constant Λ and pressure p. For constant mass M of the Universe, $H^2 = \frac{8}{3}\pi G\rho$, $t = \frac{2}{3}H^{-1}$, $a = \frac{1}{R_C}(\frac{9GM}{2})^{\frac{1}{3}}t^{\frac{2}{3}}$, where $G = 6.67 \times 10^{-11}\ \mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$ is the *gravitational constant*, $R_C = |k|^{-\frac{1}{2}}$ is the *radius of curvature*, and t is the age of the Universe.

The **scale factor** $a = a(t)$ is an expansion parameter, relating the size of the Universe $R = R(t)$ at time t to its size $R_0 = R(t_0)$ at time t_0 by $R = aR_0$.

The *Hubble constant* H is the constant of proportionality between the speed of expansion v and the size of the Universe R, i.e., $v = HR$. This equality is just the *Hubble law* with the Hubble constant $H = \frac{a'(t)}{a(t)}$. This is a linear redshift-distance relationship, where redshift is interpreted as recessional velocity v, typically expressed in km/s.

The current value of the Hubble constant is $H_0 = 71 \pm 4\,\mathrm{km\,s}^{-1}\,\mathrm{Mpc}^{-1}$, where the subscript 0 refers to the present epoch because H changes with time. The *Hubble time* and the **Hubble distance** are defined by $t_H = \frac{1}{H_0} \approx 13.82$ Ga and $D_H = \frac{c}{H_0} \approx 4.24$ Gpc. The *Hubble volume* (or *Hubble sphere*) is the region, about $10^{13}\ \mathrm{Mpc}^3$, of the Universe surrounding an observer beyond which the recessional velocity exceeds the speed c of light, i.e., any object beyond *particle horizon* (4.4×10^{26} m $=$ 47 light-Ga), is receding (due to the expansion of the Universe itself) at a rate greater than c.

The volume of observable Universe is the volume $\approx 4.1 \times 10^{34}$ cubic light-years, or $\approx 3.4 \times 10^{80}\ \mathrm{m}^3$, of Universe with a comoving size of $\frac{c}{H_0}$, i.e., a sphere with radius $\approx$ 14 Gpc (about 3 times larger than that of Hubble volume). It has mass $\approx 1.6 \times 10^{53}$ kg and contains $\approx 10^{23}$ stars (in at least 8×10^{10} galaxies) and $\approx 10^{80}$ atoms.

[redacted]

The mass density ρ (ρ_0 in the present epoch) and the value of the cosmological constant Λ are dynamical properties of the Universe; today $\rho \sim 9.47 \times 10^{-27}$ kg m^{-3} and $\Lambda \sim 10^{-52}$ m^{-2}. They can be made into dimensionless parameters Ω_M and Ω_Λ by $\Omega_M = \frac{8\pi G\rho_0}{3H_0^3}$, $\Omega_\Lambda = \frac{\Lambda}{3H_0^3}$. A third parameter $\Omega_R = 1 - \Omega_M - \Omega_\Lambda$ measures the "curvature of space". These parameters determine the geometry of the Universe if it is homogeneous, isotropic, and matter-dominated.

The velocity of a galaxy is measured by the *Doppler effect*, i.e., the fact that light emitted from a source is shifted in wavelength by the motion of the source. (The Doppler shift is reversed in some *metamaterials*: a light source moving toward an observer appears to reduce its frequency.) A relativistic form of the Doppler shift exists for objects traveling very quickly, and is given by $\frac{\lambda_{obser}}{\lambda_{emit}} = \sqrt{\frac{c+v}{c-v}}$, where λ_{emit} is the emitted wavelength, and λ_{obser} is the shifted (observed) wavelength. The change in wavelength with respect to the source at rest is called the *redshift* (if moving away), and is denoted by the letter z. The relativistic redshift z for a particle is given by $z = \frac{\Delta\lambda_{obser}}{\lambda_{emit}} = \frac{\lambda_{obser}}{\lambda_{emit}} - 1 = \sqrt{\frac{c+v}{c-v}} - 1$.

The cosmological redshift is directly related to the scale factor $a = a(t)$: $z + 1 = \frac{a(t_o)}{a(t_e)}$. Here $a(t_o)$ is the value of the scale factor at the time the light from the object is observed, and $a(t_e)$ is its value at the time it was emitted. It is usually chosen $a(t_o) = 1$, where t_o is the present age of the Universe.

- **Metric expansion of space**

  The **metric expansion of space** is the averaged increase of measured distances between objects in the Universe with time.

  It is not a motion of space and not a motion into pre-existing space. Only distances expand (and contract). The expansion has no center: all distances increase by the same factor, and every observer sees the same expanding cosmos. The observed *Hubble law* quantifies expansion from an observer. Expansion rate between two points in free space 1 m apart is 2.2×10^{-18} m/s.

  The mean distances between widely separated galaxies increase by $\approx$ 1 % every 140 million years. **FLRW metric** models, at large (superclusters of galaxies) scale, this expansion. On the scales of galaxies, there is no expansion since the metric of the local Universe has been altered by the presence of the mass of the galaxy. Full expansion, at the Hubble rate $\approx$7000 km/s, commences only at distances $\approx$ 100 Mpc. Superclusters are expanding but remain *gravitationally bound*, i.e., their expansion rate is decelerated.

  Expansion is thought to start due to cosmic inflation and then, due mainly to inertia. Its rate decelerated about 12 Ga ago due to gravity by dark and normal matter and then, from about 6 Ga ago when dark energy took over, accelerated. Now, for every megaparsec of distance from the observer, the rate of expansion increases by about 74 km/s. When the Universe doubles in volume, the dark energy doubles too. In 10^{11} years our galaxy will be the only one left in the observable Universe.

  The Universe was *radiation-dominated* with the **scale factor** $a(t) \sim t^{\frac{1}{2}}$ first $\approx$ 70,000 years, then *matter-dominated* with $a(t) \sim t^{\frac{2}{3}}$ until $\approx$ 4.5 Ma ago, then *dark-energy-dominated* with $a(t) \sim \exp(Ht)$ and the Hubble constant $H = \sqrt{\frac{8\pi G\rho}{3}} = \sqrt{\frac{\Lambda}{3}}$. In fact, its expansion caused the matter surpass the radiation in energy density and further, when matter and radiation dropped to low concentrations, the repulsive *dark energy* (or *vacuum energy*) overtook the gravity of matter.

The most commonly accepted scenario for the future is the *Big Freeze*: continued expansion results in a universe that asymptotically approaches 0° K and the *Heat Death*, a state of maximum entropy in which everything is evenly distributed. Caldwell, 2003, claimed that the scale factor a will become infinite in the finite future, resulting in *Big Rip*, final singularity in which all distances diverge to ∞.

- **Zero-gravity radius**

For a cluster of mass M, its **zero-gravity** (or *zero-velocity*, *turnover*) **radius** R_V is (Sandage, 1986, and Chernin–Teerikorpi–Baryshev, 2006) the distance r from the cluster's barycenter, where the radial force $\frac{GM}{r^2}$ of the point mass M gravity become equal to the radial force ($G2\rho_V\frac{4\pi}{3}r^3$ divided by r^2) of vacuum antigravity. So,

$$R_V^3 = \frac{3M}{8\pi\rho_V}.$$

Here G is the *gravitational constant* and $\rho_V \approx 7 \times 10^{-30}$ g/cm^3 is the constant density of dark energy inferred from global observations of supernovae 1a.

The **Einstein–Straus radius** R_M is the radius besides which expansion rate reach the global level. It is estimated that $\frac{R_M}{R_V}$ is 1.5–1.7 if the ratio of local and global density of dark energy is 0.1–1. If above ration is 1, then $R_M = R_V(1+z_V)$, where $z_V \approx 0.7$ is the global zero-acceleration redshift.

For the *Local Group* (LG), containing Milky Way and of mass 2–3.5 $\times 10^{12}$ suns, above model corresponds to observed $R_V = 1.3$–1.55 Mpc and $R_M = 2.2$–2.6 Mpc. The Virgo cluster, dominating Local supercluster, contains over 1000 galaxies in a volume slightly larger than LG; its mass is $\approx 10^{15}$ suns and $R_V = 10.3$ Mpc.

- **Hubble distance**

The **Hubble distance** (or **cosmic light horizon**, **Hubble radius**) is an increasing maximum distance $D_H = ct_H$ that a light signal could have traveled since the Big Bang, the beginning of the Universe. Here c is the speed of light and t_H is the *Hubble time* (or *age of the Universe*). It holds $t_H = \frac{1}{H_0}$, where H_0 is the *Hubble constant* which is estimated as $71 \pm 4\, km\, s^{-1} Mpc^{-1}$ at present. So, at present, $t_H \approx 4.32 \times 10^{17}$ s ≈ 13.82 Ga, and $D_H = \frac{c}{H_0} \approx 13.82$ billion light-years $\approx 1.31 \times 10^{26}$ m ≈ 4.24 Gpc, i.e., 4.6×10^{61} **Planck lengths**.

But we are observing now, due to the space expansion, objects much farther away than a static distance 13.82 Gly.

For small $\frac{v}{c}$ or small distance d in the expanding Universe, the velocity is proportional to the distance, and all distance measures, for example, **angular**

[redacted]

diameter distance, **luminosity distance**, etc., converge. In the linear approximation, this reduces to $d \approx zD_H$. But for large $\frac{v}{c}$, the relativistic **Lorentz length contraction** $L = L_0\sqrt{1-(\frac{v}{c})^2}$, where L_0 is a **proper length**, of an object traveling at velocity v relative to an observer, become noticeable to that observer.

Above Hubble radius was measured (by the Wilkinson Microwave Anisotropy Probe) as a **light travel distance** to the source of *cosmic background radiation*. Other estimations: 13.1 Gly (calibrating the distances to supernovae of a standard brightness), 14.3 Gly (measuring radio galaxies of a standard size) and 14.5 Gly (basing on the abundance ratio of uranium/thorium chondritic meteorites, [Dau05]).

- **Cosmic sound horizon**

Cosmic background radiation (CMB) is thermal radiation (strongest in the microwave region of the radio spectrum) filling the observable Universe almost uniformly. It originated $t_r \approx 380{,}000$ years after the Big Bang (or at a redshift of $z = 1100$), at *recombination*, when the Universe (ionized plasma of electrons and *baryons*, i.e., protons and neutrons) cooled to below 3000 K. (Now, the Universe's temperature is $\approx$ 1100 times cooler and its size is $\approx$ 1100 times larger.)

The electrons and protons start to form neutral hydrogen atoms, allowing photons (trapped before by Thomson scattering) to travel freely. During next $\approx$ 100,000 years radiation decoupled from the matter and the Universe became transparent.

The plasma of photons and baryons can be seen as a single fluid. The gravitational collapse around "seeds" (point-like overdensities produced during inflation) into dark matter hierarchical halos was opposed by outward radiation pressure from the heat of photon-matter interactions. This competition created longitudinal (acoustic) oscillations in the photon-baryon fluid, analogic to sound waves, created in air by pressure differences, or to ripples in a pond.

At recombination, the only remaining force on baryons is gravitation, and the pattern of oscillations (configuration of baryons and, at the centers of perturbations, dark matter) became frozen into the CMB. Baryon radiative cooling into gas and stars let this pattern of seeds to grow into structure of the Universe.

More matter existed at the centers and edges of these waves, leading eventually to more galaxies there. Today, we detect the sound waves (regular, periodic fluctuations in the density of the visible baryonic matter) via the primary CMB anisotropies.

These *baryon acoustic oscillations* (BAO) started at $t = 0$ (post-inflation) and stopped at $t = t_r$ (recombination). The **cosmic sound horizon** is the distance sound waves could have traveled. At recombination, it was $\approx c_s t_r \sim 100$ kpc, approximating the speed c_s of sound as $\frac{c}{\sqrt{3}}$.

Expanding by factor $1 + z = 1100$, it is 120–150 Mpc today. It is a standard ruler; an excess of galaxy pairs separated by this horizon was confirmed. Cf. **cosmological distance ladder** and, in Chap. 24, **acoustic metric**.

- **GZK-horizon**

  Greisen and Kuzmin–Zatsepin, 1966, computed that a cosmic ray with kinetic energy over *GZK-limit* (5×10^{19} eV) traveling from its distant, over **GZK-horizon** (50 Mpc $\approx$ 163 Mly) source, will be absorbed (due to slowing interaction with photons of the CMB and associated mean path) and so never observed on Earth.

  Several cosmic rays apparently exceeding *GZK-limit* were observed; this *GZK-paradox* is still unexplained.
- **Comoving distance**

  The standard Big Bang model uses *comoving coordinates*, where the spatial reference frame is attached to the average positions of galaxies. With this set of coordinates, both the time and expansion of the Universe can be ignored and the shape of space is seen as a spatial hypersurface at constant cosmological time.

  The **comoving** (or *cosmological*) **distance** is a distance (denoted χ or d_{comov}) in comoving coordinates between two points in space at a single cosmological time, i.e., the distance between two *nearby* (close in redshift z) objects, which remains constant with epoch if these objects are moving with the Hubble flow.

  The (cosmological) **proper distance** d_{proper} is a distance between two nearby events in the frame in which they occur at the same time. It is the distance measured by a ruler at the time t_o of observation. It holds

  $$d_{comov}(x,y) = d_{proper}(x,y) \cdot \frac{a(t_o)}{a(t_e)} = d_{proper}(x,y) \cdot (1+z),$$

  where $a(t)$ is the **scale factor**. In the time t_o, i.e., at the present, $a = a(t_o) = 1$, and $d_{comov} = d_{proper}$, In general, $d_{proper}(t) = a(t)d_{comov}$, for a cosmological time t.

  The total **line-of-sight comoving distance** D_C from us to a distant object is computed by integrating the infinitesimal $d_{comov}(x,y)$ contributions between nearby events along the time ray from the time t_e, when the light from the object was emitted, to the time t_o, when the object is observed:

  $$D_C = \int_{t_e}^{t_o} \frac{cdt}{a(t)}.$$

  In terms of redshift, D_C from us to a distant object is computed by integrating the infinitesimal $d_{comov}(x,y)$ contributions between nearby events along the radial ray from $z = 0$ to the object: $D_C = D_H \int_0^z \frac{dz}{E(z)}$, where D_H is the **Hubble distance**, and $E(z) = (\Omega_M(1+z)^3 + \Omega_R(1+z)^2 + \Omega_\Lambda)^{\frac{1}{2}}$.

  In a sense, the comoving distance is the fundamental distance measure in Cosmology, since all other distances can simply be derived in terms of it.
- **Proper motion distance**

  The **proper motion distance** (or **transverse comoving distance**, *contemporary angular diameter distance*) D_M is a distance from us to a distant object

[redacted]

defined as the ratio of the actual transverse velocity (in distance over time) of the object to its *proper motion* (in radians per unit time). It is given by

$$D_M = \begin{cases} D_H \frac{1}{\sqrt{\Omega_R}} \sinh(\sqrt{\Omega_R} D_C/D_H), & \text{for } \Omega_R > 0, \\ D_C, & \text{for } \Omega_R = 0, \\ D_H \frac{1}{\sqrt{|\Omega_R|}} \sin(\sqrt{|\Omega_R|} D_C/D_H), & \text{for } \Omega_R < 0, \end{cases}$$

where D_H is the **Hubble distance**, and D_C is the **line-of-sight comoving distance**. For $\Omega_\Lambda = 0$, there is an analytic solution (z is the *redshift*):

$$D_M = D_H \frac{2(2 - \Omega_M(1 - z) - (2 - \Omega_M)\sqrt{1 + \Omega_M z})}{\Omega_M^2(1 + z)}.$$

The proper motion distance D_M coincides with the line-of-sight comoving distance D_C if and only if the curvature of the Universe is equal to zero. The **comoving distance** between two events at the same redshift or distance, but separated in the sky by some angle $\delta\theta$, is equal to $D_M \delta\theta$.

The distance D_M is related to the **luminosity distance** D_L and the **angular diameter distance** D_A by $D_M = (1 + z)^{-1} D_L = (1 + z) D_A$.

- **Luminosity distance**

  The **luminosity distance** D_L is a distance from us to a distant object defined by the relationship between the observed flux S and emitted luminosity L:

$$D_L = \sqrt{\frac{L}{4\pi S}}.$$

  This distance is related to the **proper motion distance** D_M and t the **angular diameter distance** by $D_L = (1 + z) D_M = (1 + z)^2 D_A$, where z is the *redshift*.

  The luminosity distance does take into account the fact that the observed luminosity is attenuated by two factors, the relativistic redshift and the Doppler shift of emission, each of which contributes an $(1 + z)$ attenuation: $L_{obser} = \frac{L_{emiss}}{(1+z)^2}$.

  The *corrected luminosity distance* D_L' is defined by $D_L' = \frac{D_L}{1+z}$.

- **Distance modulus**

  The **distance modulus** is $DM = 5 \ln(\frac{D_L}{10pc})$, where D_L is the **luminosity distance**. The distance modulus is the difference between the *absolute magnitude* (the brightness that star would appear to have if it was at a distance of 10 parsec) and *apparent magnitude* (the actual brightness) of an astronomical object.

  Distance moduli are most commonly used when expressing the distances to other galaxies. For example, the Andromeda Galaxy's *DM* is 24.5, and the Virgo cluster has *DM* equal to 31.7. For a much smaller object (planet, comet or asteroid), the *absolute magnitude* is its apparent visual magnitude at zero phase angle and at unit (1 AU) heliocentric and geocentric distances. The brightest (with peak apparent magnitude −7.5) recorded stellar event was the supernova in 1006.

- **Angular diameter distance**

  The **angular diameter distance** (or *angular size distance*) D_A is a distance from us to a distant object defined as the ratio of an object's physical transverse size to its angular size (in radians). It is used to convert angular separations in telescope images into proper separations at the source. It is special for not increasing indefinitely as $z \to \infty$; it turns over at $z \sim 1$, and so more distant objects actually appear larger in angular size. D_A is related to the **proper motion distance** D_M and the **luminosity distance** D_L by $D_A = \frac{D_M}{1+z} = \frac{D_L}{(1+z)^2}$, where z is the *redshift*.

  The *distance duality* $\frac{D_L(z)}{D_A(z)} = (1+z)^2$ links D_L, based on the apparent luminosity of standard candles (for example, supernovae) and D_A, based on the *apparent size* ("visual diameter" measured as an angle) of standard rulers (for example, **cosmic sound horizon**). It holds for any general **metric theory of gravity** (Chap. 24) in any background in which photons travel on unique null geodesics.

  If the angular diameter distance is based on the representation of object diameter as angle × distance, the **area distance** is defined similarly according the representation of object area as solid angle × distance^2.

- **Einstein radius**

  General Relativity predicts *gravitational lensing*, i.e., deformation of the light from a *source* (a galaxy or star) in the presence of a *gravitational lens*, i.e., a body of large mass M (another galaxy, or a black hole) bending it.

  If the source S, lens L and observer O are all aligned, the gravitational deflection is symmetric around the lens. The **Einstein radius** is the radius of the resulting *Einstein* (or *Chwolson*) *ring*. In radians it is

  $$\sqrt{M\frac{4G}{c^2}\frac{D(L,S)}{D(O,L)D(O,S)}},$$

  where $D(O,L)$ and $D(O,S)$ are the **angular diameter distances** of the lens and source, while $D(L,S)$ is the angular diameter distance between them.

- **Light travel distance**

  The **light travel** (or *look-back*) **distance** is a distance from us to a distant object, defined by $D_{lt} = cD_t$, where D_t is the difference $t_o - t_e$ between the time, when the object was observed, and the time, when the light from it was emitted. The *look-back time* D_t is a **proper time**, but D_{lt} is not a **proper distance**.

  D_{lt} is not a very useful distance, because it is hard to determine t_e, the age of the Universe at the time of emission of the light which we see. Cf. **Hubble radius**.

- **Parallax distance**

  Given an object O viewed along two different lines of sight, its *parallax* is the angle $p = AOB$ between its directions of view from the two ends of a baseline AB. If $AO \approx BO$ and p, AB are small, the distance $A0$ can be easily estimated.

[redacted]

Animals use their two eyes (*stereoopsis*) or two positions of moving head (*motion parallax*) as points A, B. Cf. **animal depth/distance perception** in Chap. 23.

In Astronomy, the **parallax distance** is a distance D_P from us to a distant object (say, a star) defined from measuring of *stellar parallaxes*, i.e., its apparent changes of position in the sky caused by the motion of the observer on the Earth.

Usually, it is the *annual* (or *heliocentric*) *paralllax*, i.e., p is the angle Earth-star-Sun (in arcseconds) subtended at a star by the longest baseline—mean radius $R = 1$ AU of the Earth's orbit around the Sun. So, this distance (in parsecs, corresponding to $p = 1$ arcsec) is given by $D_P = \frac{1}{\tan(p)} \approx \frac{1}{p}$, since p is usually very small.

- **Kinematic distance**

  The **kinematic distance** is the distance to a galactic source, which is determined from differential rotation of the galaxy: the *radial velocity* of a source directly corresponds to its **galactocentric distance**. But the *kinematic distance ambiguity* arises since, in our inner galaxy, any given galactocentric distance corresponds to two distances along the line of sight, *near* and *far* kinematic distances.

  This problem is solved, for some galactic regions, by measurement of their absorption spectra, if there is an interstellar cloud between the region and observer.

- **Radar distance**

  The **radar** (or *target*) **distance** D_R is a distance from us to a distant object, measured by a *radar*, i.e., a high frequency radio pulse sent out for a short interval of time. When it encounters a conducting object, sufficient energy is reflected back to allow radar to detect it. Since radio waves travel in air at close to their speed c (of light) in vacuum, one can calculate the distance D_R of the detected object from the round-trip time t between the transmitted and received pulses as

$$D_R = \frac{1}{2}ct.$$

  In general, *Einstein protocol* is to measure the distance between two objects A and B as $\frac{1}{2}c(t_3 - t_1)$. Here a light pulse is sent from A to B at time t_1 (measured in A), received at time t_2 (measured in B) and immediately sent back to A with a return time t_3 (measured with A).

- **Pulsar dispersion measure**

  A *pulsar* is an extraterrestrial source of radiation having a regular periodicity, detected as short bursts in, usually, the radio region of the spectrum.

  The **pulsar dispersion measure** (DM) is the integrated *column density* of free electrons (their number per unit area) between an observer and a pulsar. DM is shown by a pulse's broadening when it is observed over a finite bandwidth.

  The distance to the pulsar is $\frac{DM}{n_e}$, where n_e is the mean electron density.

- **Cosmological distance ladder**

  For measuring distances to astronomical objects, one uses a kind of "ladder" of different methods; each method applies only for a limited distance, and each method which applies for a larger distance builds on the data of the preceding methods.

  The starting point is knowing the distance from the Earth to the Sun; this distance is called one *astronomical unit* (*AU*), and is roughly 150 million km. Distances in the inner Solar System are measured by bouncing radar signals off planets or asteroids, and measuring the time until the echo is received.

  The next step in the ladder consists of simple geometrical methods; with them, one can go to a few hundred ly. The distance to nearby stars can be determined by their *parallaxes*: using Earth's orbit as a baseline, the distances to stars are measured by triangulation. This is accurate to about 1 % at 50 ly, 10 % at 500 ly.

  Using data acquired by the geometrical methods, and adding *photometry* (measurements of the brightness) and spectroscopy, one gets the next step in the ladder for stars so far away that their parallaxes are not measurable yet. The *distance-luminosity relation* is that the light intensity from a star is inversely proportional to the square of its distance; cf. **distance modulus**.

  For even larger distances, are used *standard candles*, i.e., several types of cosmological objects, for which one can determine their absolute brightness without knowing their distances. *Primary standard candles* are the *Cepheid* variable stars. They periodically change their size and temperature. There is a relationship between the brightness of these pulsating stars and the period of their oscillations, and this relationship can be used to determine their absolute brightness. Cepheids can be identified as far as in the Virgo cluster (60 Mly).

  Secondary standard candles are supernovae 1a (having equal peak brightness), red giant branch stars, active galactic nuclei and entire galaxies. Main other techniques to estimate the **angular diameter distance** to galaxies are *gravitational lensing* (cf. **Einstein radius**) and using *baryon acoustic oscillations* matter clustering (cf. **cosmic sound horizon**) as a standard ruler.

  For very large distances (hundreds of Mly or several Gly), the cosmological redshift and the Hubble law are used. A complication is that it is not clear what is meant by "distance" here, and there are several types of distances used here: **luminosity distance**, **proper motion distance**, **angular diameter distance**, etc.

  Depending on the situation, there is a large variety of special techniques to measure distances in Cosmology, such as *light echo*, *Bondi radar*, *RR Lyrae*, *maser* distances and *secular*, *statistical*, *expansion*, *spectroscopic* parallax distances.

  NASA's Chandra X-ray Observatory measures large distances via the delay of the halo of scattering material (interstellar dust grains) between the source and Earth. Also, frequency-dependent dispersion of *fast radio bursts* indicate the distance via amount of ionized plasma that lies between the burst's origin and Earth.

- **VLBI distance measurements**

  Very-long-baseline interferometry (VLBI) is a technique of radio astronomy.

[redacted]

In VLBI a signal from a compact bright astronomical radio source (say, *pulsar* or *maser*) is collected by an array of radio telescopes on Earth and in near space. The distance between the telescopes is then calculated using the time difference between the arrivals of the signals at different telescopes. Observations of an object made simultaneously by many telescopes are combined, emulating a telescope with a size equal to the maximum distance between the telescopes. VLBI arrays can be combined, increasing the resolution to, say, microarcseconds.

VLBI directly measure distance, via *annual parallax*, and proper motions.

26.2 Distances in Theory of Relativity

The *Minkowski space-time* (or *Minkowski space*, *Lorentz space-time*, *flat space-time*) is the usual geometric setting for the Einstein Special Theory of Relativity. In this setting the three ordinary dimensions of space are combined with a single dimension of time to form a 4D *space-time* $\mathbb{R}^{1,3}$ in the absence of gravity. See, for example, [Wein72] for details.

Vectors in $\mathbb{R}^{1,3}$ are called 4-*vectors* (or *events*). They can be written as (ct, x, y, z), where the first component is the unidirectional *time-like dimension* (c is the speed of light in vacuum, and t is the time), while the other three components are bidirectional *spatial dimensions*. Formally, c is a conversion factor from time to space.

In fact, c is the speed of gravitational waves and any massless particle: the *photon* (carrier of electromagnetism), the *gluon* (carrier of the strong force) and the *graviton* (theoretical carrier of gravity). It is the highest possible speed for any physical interaction in nature and the only speed independent of its source and the motion of an observer.

In the *spherical coordinates*, the events can be written as (ct, r, θ, ϕ), where r $(0 \leq r < \infty)$ is the *radius* from a point to the origin, ϕ $(0 \leq \phi < 2\pi)$ is the azimuthal angle in the xy plane from the x axis (*longitude*), and θ $(0 \leq \theta \leq \pi)$ is the polar angle from the z axis (*colatitude*). 4-vectors are classified according to the sign of their squared *norm*:

$$||v||^2 = \langle v, v \rangle = c^2t^2 - x^2 - y^2 - z^2.$$

They are said to be *time-like*, *space-like*, and *light-like* (*isotropic*) if their squared norms are positive, negative, or equal to zero, respectively. The set of all light-like 4-vectors forms the *light cone*. If the coordinate origin is singled out, the space can be broken up into three domains: domains of *absolute future* and *absolute past*, falling within the light cone, whose points are joined to the origin by time-like vectors with positive or negative value of time coordinate, respectively, and the domain of *absolute elsewhere*, falling outside of the light cone, whose points are joined to the origin by space-like vectors.

A *world line* of an object is the sequence of events that marks its time history. A world line is a *time-like* curve tracing out the path of a single point in the Minkowski space-time, i.e., at any point its *tangent vector* is a time-like 4-vector. All world lines fall within the light cone, i.e., the curves whose tangent vectors are light-like 4-vectors correspond to the motion of light and other particles of zero rest mass.

World lines of particles at constant speed (equivalently, of free falling particles) are called *geodesics*. In Minkowski space they are straight lines. A geodesic in Minkowski space which joins two given events x and y, is the longest curve among all world lines which join these two events. This follows from the **Einstein time triangle inequality** (cf. **inverse triangle inequality** and, in Chap. 5, **reverse triangle inequality**):

$$||x+y|| \geq ||x|| + ||y||,$$

according to which a time-like broken line joining two events is shorter than the single time-like geodesic joining them, i.e., the **proper time** of the particle moving freely from x to y is greater than the **proper time** of any other particle whose world line joins these events. It holds also in Minkowski space extended to any number of spatial dimensions, assuming null or time-like vectors in the same time direction. It is called *twin paradox*.

The *space-time* is a 4D *manifold* which is the usual mathematical setting for the Einstein General Theory of Relativity, which is the generalization of Special Relativity to include gravitation. Here the three spatial components with a single time-like component form a 4D space-time in the presence of gravity. Gravity is equivalent to the geometric properties of space-time, and in the presence of gravity the geometry of space-time is curved. (Bean, 2009, found evidence that over extragalactic distances gravity exerts a greater pull on time than on space.) So, the space-time is a 4D curved manifold for which the tangent space to any point is the Minkowski space, i.e., it is a *pseudo-Riemannian manifold*—a manifold, equipped with a nondegenerate indefinite metric (called **pseudo-Riemannian metric** in Chap. 7) of signature (1,3).

In the General Theory of Relativity, gravity is described by the properties of the local geometry of space-time. In particular, the gravitational field can be built out of a **metric tensor**, a quantity describing geometrical properties space-time such as distance, area, and angle. Matter is described by its *stress-energy tensor*, a quantity which contains the density and pressure of matter. The strength of coupling between matter and gravity is determined by the gravitational constant G.

The *Einstein field equation* is an equation in the General Theory of Relativity, that describes how matter creates gravity and, conversely, how gravity affects matter. A solution of the Einstein field equation is a certain **Einstein metric** appropriated for the given mass and pressure distribution of the matter.

A *black hole* is an astrophysical object that is theorized to be created from the collapse of a neutron or "quark" star. The gravitational forces are so strong in a black hole that they overcome neutron degeneracy pressure and, roughly, collapse to a *singularity* (point of infinite density and space-time curvature). Even light cannot escape the gravitational pull of a black hole within the black hole's its *event horizon*.

[redacted]

Uncharged black holes are called *Schwarzschild* or *Kerr black holes* if their angular momentum is zero or not, respectively. Charged black holes are called *Kerr–Newman* or *Reissner–Nordström black holes* if they are spinning or not, respectively.

Universe and black hole both have singularities—in time and space, respectively. *Naked* (not surrounded by a black hole) singularities were not observed but they might exist also. **Kerr metric** and **Reissner–Nordström metric below** admit such case. Also, a *kugelblitz* is a putative black hole formed from energy as opposed to mass.

Experimentally, General Relativity is still untested for strong fields (such as near neutron-star surfaces or black-hole horizons) or over distances on a galactic scale and larger. Neither Newton law of gravitation was tested below 6×10^{-5} m.

Putative *gravitational waves* (fluctuations in the curvature of space-time propagating as a wave, predicted by Einstein), have been detected in 2014. Also predicted *frame-dragging effect* (the spinning Earth pulls space-time around with it) is under probe. The *geodetic effect*, confirming that space-time acts on matter, was found.

- **Minkowski metric**

The **Minkowski metric** is a **pseudo-Riemannian metric**, defined on the *Minkowski space* $\mathbb{R}^{1,3}$, i.e., a 4D real vector space which is considered as the *pseudo-Euclidean space* of signature $(1,3)$. It is defined by its **metric tensor**

$$((g_{ij})) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

The *line element* ds^2 of this metric are given by

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2.$$

In *spherical coordinates* (ct, r, θ, ϕ), one has $ds^2 = c^2dt^2 - dr^2 - r^2d\theta^2 - r^2 \sin^2 \theta d\phi^2$.

The pseudo-Euclidean space $\mathbb{R}^{1,3}$ of signature $(1,3)$ with the *line element*

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2$$

can also be used as a space-time model of the Special Theory of Relativity.

Above notion of *space-time* (Minkowski, 1908) was the first application of geometry to a nonlength-like quantity. But there were some precursors of such union of space and time. Lagrange, 1797, observed that with time as a 4-th coordinate, "one can regard mechanics as 4-dimensional geometry". Schopenhauer wrote in *On the Fourfold Root of the Principle of Sufficient Reason* (1813): "…it is only by the combination of Time and Space that the representation of coexistence arises." Poe wrote "Space and Duration are one" in *Eureka: A Prose Poem* (1848).

Wells wrote on the first page of *The Time Machine* (1895): ‘Clearly,’ the Time Traveler proceeded, ‘any real body must have extension in four directions: it must have Length, Breadth, Thickness, and Duration ... There is no difference between Time and any of the tree dimensions of Space except that our consciousness moves along it”. *Quechua*, the language of Inca and 8–10 million modern speakers, have a single concept, *pacha*, for the location in time and space.

- **Proper distance and time**

In Relativistic Physics, **proper distance** and **proper time** between any two events are true physical distance and time difference: the spatial distance between them when the events are simultaneous and the temporal distance between them when the events occur at the same spatial location. They are the *invariant* (with respect to Lorentz transformations, describing a transition to a coordinate system associated with a moving body) intervals of a space-like path or pair of space-like separated events, and, respectively, of a time-like path or pair of time-like separated events.

In General Relativity, **proper time** is the pseudo-Riemannian arc length of world lines in 4D-spacetime. In particular, in SR (Special Relativity), it is

$$\tau = \int_P \sqrt{dt^2 - c^{-2}(dx^2 + dy^2 + dz^2)},$$

where t and x, y, z are time and spatial coordinates, while P is the path of the clock in space-time. In the subcase of inertial motion, it become

$$\Delta\tau = \sqrt{(\Delta t)^2 - c^{-2}((\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2)},$$

where Δ means “the change in” between two events. Cf. the **kinematic metric**.

In SR, the **proper distance** between two space-like separated events is

$$\Delta\sigma = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 - c^2(\Delta t)^2}.$$

- **Proper length**

In Special Theory of Relativity, the **proper** (or *rest*) **length** between two space-like separated events is the distance between them, such as measured in an inertial frame of reference in which the events are simultaneous. In contrast to invariant **proper distance**, such simultaneity depends on the observer.

In a flat space-time, the proper length between two events is the proper length of a straight path between them. General Relativity consider the curved space-times in which may be more than one straight path (geodesic) between two events.

So, the general **proper length** is defined as the path integral $\int_P \sqrt{-g_{ij}dx^i dx^j}$, where g_{ij} is the metric tensor for the space-time with signature $(1, 3)$, along the shortest curve joining the endpoints of the space-like path P at the same time.

[redacted]

- **Affine space-time distance**

  Given a space-time (M^4, g), there is a unique affine parametrization $s \to \gamma(s)$ for each light ray (i.e., light-like geodesic) through the observation event p_o, such that $\gamma(0) = p_o$ and $g(\frac{d\gamma}{dt}, U_o) = 1$, where U_o is the 4-velocity of the observer at p_o (i.e., a vector with $g(U_o, U_o) = -1$). In this case, the **affine space-time distance** is the *affine parameter* s, viewed as a distance measure.

  This distance is monotone increasing along each ray, and it coincides, in a small neighborhood of p_o, with the Euclidean distance in the rest system of U_o.

- **Lorentz metric**

  A **Lorentz metric** is a **pseudo-Riemannian metric** (i.e., **nondegenerate indefinite metric**, cf. Chap. 7) of signature $(1, p)$.

  The curved space-time of the General Theory of Relativity can be modeled as a *Lorentzian manifold* (a manifold equipped with a Lorentz metric) of signature $(1, p)$. The *Minkowski space* $\mathbb{R}^{1,p}$ with the flat **Minkowski metric** is a model of it, in the same way as Riemannian manifolds can be modeled on Euclidean space.

  Given a rectifiable non-space-like curve $\gamma : [0,1] \to M$ in the space-time M, the *length* of the curve is defined as $l(\gamma) = \int_0^1 \sqrt{-\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \rangle} dt$. For a space-like curve, we set $l(\gamma) = 0$ and define the **Lorentz distance** between two points $p, q \in M$ as

  $$\sup_{\gamma \in \Gamma} l(\gamma),$$

  if $p \prec q$, i.e., if the set Γ of *future directed* non-space-like curves from p to q is nonempty; otherwise, this distance is 0.

  The **Lorentz–Minkowski distance** is a **pseudo-Euclidean distance** (Chap. 7) $\sqrt{D(x,y)}$, where $D(x,y) = |x_1 - y_1|^2 - \sum_{2 \le i \le n} |x_i - y_i|^2$. The points are called *time-*, *space-*, *null-separated* if $D(x,y)$ is more, less or equal to 0, respectively, i.e., if they can be joined by a time-like, space-like or null path.

- **Distances on causal sets**

  Causal Set Theory is a fundamentally discrete approach to quantum gravity. A *causal set* (or *causet*) is a partially ordered set $(X, \preceq)$, which is *locally finite*, i.e., the *interval* $(x,y) = \{z \in X : x \prec z \prec y\}$ is finite for any $x, y \in X$. A *link* is a pair $x, y \in X$ such that $x \prec y$ and $(x,y) = \emptyset$. A *chain* is a subcauset such that $x \prec y$ or $y \prec x$ for any two its elements x, y.

  Given $x, y \in X$ with $x \prec y$, their **time-like distance** $d_t(x,y)$ is (Brightwell–Gregory, 1991) the *length* (number of links) in any *geodesic* between them, i.e., longest chain between and including x and y. Given two unrelated elements $x, y \in X$, their **naive space-like distance** is defined (Brightwell–Gregory, 1991) as

  $$d_{ns}(x,y) = \min_{u,v \in X : u \prec (x,y) \prec v} d_t(u,v).$$

  Rideout–Wallden, 2013, modified $d_{ns}(x,y)$, replacing the minimum above with an average over suitably selected minimizing pairs.

The elements of X can be seen as events in a discrete space-time, where the partial order represent causal relationship. In a causet embedded in Minkowski space-time, the distance $d_t(x, y)$ is proportional to the **proper time**. A related discrete space-time is a random poset obtained by sampling from a compact domain in a space-time manifold. Cf. also **D-separation in Bayesian network** in Chap. 22.

- **Kinematic metric**

  Given a set X, a **kinematic metric** (or *abstract Lorentzian distance*) is (Pimenov, 1970) a function $\tau : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that, for all $x, y, z \in X$:

  1. $\tau(x, x) = 0$;
  2. $\tau(x, y) > 0$ implies $\tau(y, x) = 0$ (*antisymmetry*);
  3. $\tau(x, y), \tau(y, z) > 0$ implies $\tau(x, z) > \tau(x, y) + \tau(y, z)$ (**inverse triangle inequality** or *anti-triangle inequality*).

  The *space-time* set X consists of *events* $x = (x_0, x_1)$ where, usually, $x_0 \in \mathbb{R}$ is the *time* and $x_1 \in \mathbb{R}^3$ is the *spatial location* of the event x. The inequality $\tau(x, y) > 0$ means *causality*, i.e., x can influence y; usually, it is equivalent to $y_0 > x_0$ and the value $\tau(x, y) > 0$ can be seen as the largest (since it depends on the speed) **proper time** of moving from x to y.

  If the gravity is negligible, then $\tau(x, y) > 0$ implies $y_0 - x_0 \geq ||y_1 - x_1||_2$, and $\tau_p(x, y) = ((y_0 - x_0)^p - ||y_1 - x_1||_2^p))^{\frac{1}{p}}$ (as defined by Busemann, 1967) is a real number. For $p \approx 2$ it is consistent with Special Relativity observations.

  A kinematic metric is not our usual distance metric; also it is not related to the **kinematic distance** in Astronomy.

  But Zapata, 2013, proved that $\sup_z \max(|\tau(x, z) - \tau(y, z)|, |\tau(z, x) - \tau(z, y)|)$ is a continuous metric on a compact part of space-time and it generates the same topology as a nonphysical coordinate-dependent Euclidean distance $\sqrt{\sum_{i=2}^{4} |x_i - y_i|^2}$.

- **Galilean distance**

  The **Galilean distance** is a distance on $\mathbb{R}^n$ defined by

$$|x_1 - y_1| \quad \text{if } x_1 \neq y_1,$$

and by

$$\sqrt{(x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \text{ if } \; x_1 = y_1.$$

The space $\mathbb{R}^n$, equipped with the Galilean distance, is called *Galilean space*.

For $n = 4$, it is a setting for the space-time of classical mechanics according to Galilei–Newton in which the distance between two events taking place at the points p and q at the moments of time t_1 and t_2 is defined as the time interval $|t_1 - t_2|$, while if $t_1 = t_2$, it is defined as the distance between the points p and q.

[redacted]

- **Einstein metric**

  In the General Theory of Relativity, describing how space-time is curved by matter, the **Einstein metric** is a solution to the Einstein field equation

  $$R_{ij} - \frac{g_{ij}R}{2} + \Lambda g_{ij} = \frac{8\pi G}{c^4}T_{ij},$$

  i.e., a **metric tensor** $((g_{ij}))$ of signature $(1, 3)$, appropriated for the given mass and pressure distribution of the matter. Here $E_{ij} = R_{ij} - \frac{g_{ij}R}{2} + \Lambda g_{ij}$ is the *Einstein curvature tensor*, R_{ij} is the *Ricci curvature tensor*, R is the *Ricci scalar*, and T_{ij} is a *stress-energy tensor*. *Empty space* (*vacuum*) is the case of $R_{ij} = 0$.

  Einstein introduced in 1917 the *cosmological constant* Λ to counteract the effects of gravity on ordinary matter and keep the Universe *static*, i.e., with **scale factor** always being 1. He put $\Lambda = \frac{4\pi G\rho}{c^2}$. The static Einstein metric for a homogeneous and isotropic Universe is given by the *line element*

  $$ds^2 = -dt^2 + \frac{dr^2}{(1 - kr^2)} + r^2(d\theta^2 + \sin^2\theta d\phi^2),$$

  where k is the curvature of the space-time. The radius of this curvature is $\frac{c}{\sqrt{4\pi G\rho}}$ and numerically it is of the order 10 Gly. Einstein from 1922 call this model his "biggest blunder" but Λ was reintroduced in modern dynamic models as dark energy.

- **de Sitter metric**

  The **de Sitter metric** is a maximally symmetric vacuum solution to the Einstein field equation with a positive cosmological constant Λ, given by the *line element*

  $$ds^2 = dt^2 + e^{2\sqrt{\frac{\Lambda}{3}}t}(dr^2 + r^2 d\theta^2 + r^2\sin^2\theta d\phi^2).$$

  Expansion of Universe is accelerating at the rate consistent with $\Lambda \sim 10^{-123}$, but Hartie–Hawking–Hertog, 2012, gave a quantum model of it with $\Lambda < 0$.

  The most symmetric solutions to the Einstein field equation in a vacuum for $\Lambda = 0$ and $\Lambda < 0$ are the flat **Minkowski metric** and the **anti de Sitter metric**.

  The n-dimensional *de Sitter space* dS_n and *anti de Sitter space* AdS_n are Lorentzian manifold analogs of elliptic and hyperbolic space, respectively.

  In order to explain the *hierarchy problem* (why the weak nuclear force is 10^{32} times stronger than gravity?), Randall and Sundram, 1999, proposed that Universe is 5D anti de Sitter space AdS_5 with elementary particles, except for the graviton, being on a $(3 + 1)$-D brane or branes. This **Randall–Sundrum metric** is $ds^2 = e^{-2ky}g_{ab}dx^a dy^b + dy^2$, where k is of order the Planck scale ($\sim 10^{-35}$ m) and x^a, y are coordinates in 4D and extra-dimension.

- **BTZ metric**

  The **BTZ metric** (Banados, Teitelboim and Zanelli, 2001) is a black hole solution for $(2+1)$-dimensional gravity with a negative cosmological constant Λ.

  There are no such solutions with $\Lambda = 0$. BTZ black holes without any electric charge are locally isometric to *anti de Sitter space*.

  This metric is given by the *line element*

$$ds^2 = -k^2(r^2 - R^2)dt^2 + \frac{1}{k^2(r^2 - R^2)}dr^2 + r^2 d\theta^2,$$

  where R is the black hole radius, in the absence of charge and angular momentum.
- **Schwarzschild metric**

  The **Schwarzschild metric** is a vacuum solution to the *Einstein field equation* around a spherically symmetric mass distribution; this metric represents the Universe around a black hole of a given mass, from which no energy can be extracted.

  It was found by Schwarzschild, 1915, only a few months after the publication of the Einstein field equation, and was the first exact solution of this equation.

  The *line element* of this metric is given by

$$ds^2 = \left(1 - \frac{r_g}{r}\right) c^2 dt^2 - \frac{1}{\left(1 - \frac{r_g}{r}\right)} dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2),$$

  where $r_g = \frac{2Gm}{c^2}$ is the *Schwarzschild radius* and m is the mass of the black hole.

  This solution is only valid for radii larger than r_g, as at $r = r_g$ there is a coordinate singularity. This problem can be removed by a transformation to a different choice of space-time coordinates, called *Kruskal–Szekeres coordinates*. As $r \to +\infty$, the Schwarzschild metric approaches the **Minkowski metric**.
- **Kottler metric**

  The **Kottler metric** is the unique spherically symmetric vacuum solution to the Einstein field equation with a cosmological constant Λ. It is given by

$$ds^2 = -\left(1 - \frac{2m}{r} - \frac{\Lambda r^2}{3}\right) dt^2 + \left(1 - \frac{2m}{r} - \frac{\Lambda r^2}{3}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

  It is called also the **Schwarzschild–de Sitter metric** for $\Lambda > 0$ and **Schwarzschild–anti de Sitter metric** for $\Lambda < 0$. Cf. **Delaunay metric** in Chap. 7.
- **Reissner–Nordström metric**

  The **Reissner–Nordström metric** is a vacuum solution to the Einstein field equation around a spherically symmetric mass distribution in the presence of a charge; it represents the Universe around a charged black hole. This metric is

[redacted]

given by

$$ds^2 = \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right) dt^2 - \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right)^{-1} dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2),$$

where m is the mass of the hole, e is the charge ($e < m$), and we have used units with the speed c of light and the *gravitational constant* G equal to one.

- **Kerr metric**

  The **Kerr metric** (or **Kerr–Schild metric**) is an exact solution to the Einstein field equation for empty space (vacuum) around an axially symmetric, rotating mass distribution, This metric represents the Universe around a rotating black hole. Its *line element* is given (in *Boyer–Lindquist form*) by

  $$ds^2 = \rho^2\left(\frac{dr^2}{\Delta} + d\theta^2\right) + (r^2 + a^2)\sin^2\theta d\phi^2 - dt^2 + \frac{2mr}{\rho^2}(a\sin^2\theta d\phi - dt)^2,$$

  where $\rho^2 = r^2 + a^2\cos^2\theta$ and $\Delta = r^2 - 2mr + a^2$. Here m is the mass of the black hole and a is the angular velocity as measured by a distant observer.

  The **Schwarzschild metric** is the Kerr metric with $a = 0$. A black hole is rotating if radiation processes are observed outside its *Schwarzschild radius* (the event horizon radius as dependent on the mass only) but inside its *Kerr radius* (where the rotational kinetic energy is comparable with the rest energy). For the Earth, those radii are about 1 cm and 3 m, respectively.

  In 2013, the spin of a black hole was directly measured for the first time: the central black hole of the galaxy NGC 1365 rotates at 84 % of the speed c of light.

- **Kerr–Newman metric**

  The **Kerr–Newman metric** is an exact, unique and complete solution to the Einstein field equation for empty space (vacuum) around an axially symmetric, rotating mass distribution in the presence of a charge, This metric represents the Universe around a rotating charged black hole. Its *line element* is given by

  $$ds^2 = -\frac{\Delta}{\rho^2}(dt - a\sin^2\theta d\phi)^2 + \frac{\sin^2\theta}{\rho^2}((r^2 + a^2)d\phi - adt)^2 + \frac{\rho^2}{\Delta}dr^2 + \rho^2 d\theta^2,$$

  where $\rho^2 = r^2 + a^2\cos^2\theta$ and $\Delta = r^2 - 2mr + a^2 + e^2$. Here m is the mass of the black hole, e is the charge, and a is the angular velocity.

  The Kerr–Newman metric becomes the **Kerr metric** if the charge is 0 and the **Reissner–Nordström metric** if the angular momentum is 0.

- **Ozsváth–Schücking metric**

  The **Ozsváth–Schücking metric** (1962) is a rotating vacuum solution to the field equations having in Cartesian coordinates the form

  $$ds^2 = -2[(x^2 - y^2)\cos(2t) - 2xy\sin(2t)]dt^2 + dx^2 + dy^2 - 2dtdz.$$

- **Static isotropic metric**

  The **static isotropic metric** is the most general solution to the Einstein field equation for empty space (vacuum); this metric can represent a static isotropic gravitational field. The *line element* of this metric is given by

  $$ds^2 = B(r)dt^2 - A(r)dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2),$$

  where $B(r)$ and $A(r)$ are arbitrary functions.
- **Eddington–Robertson metric**

  The **Eddington–Robertson metric** is a generalization of the **Schwarzschild metric** which allows that the mass m, the *gravitational constant* G, and the density ρ are altered by unknown dimensionless parameters α, β, and γ (all equal to 1 in the *Einstein field equation*). The *line element* of this metric is given by

  $$ds^2 = \left(1 - 2\alpha\frac{mG}{r} + 2(\beta - \alpha\gamma)\left(\frac{mG}{r}\right)^2 + \dots\right)dt^2 - \left(1 + 2\gamma\frac{mG}{r} + \dots\right)dr^2 -$$

  $$-r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

- **Janis–Newman–Wincour metric**

  The **Janis–Newman–Wincour metric** is the most general spherically symmetric static and asymptotically flat solution to the Einstein field equation coupled to a massless scalar field. It is given by the *line element*

  $$ds^2 = -\left(1 - \frac{2m}{\gamma r}\right)^{\gamma} dt^2 + \left(1 - \frac{2m}{\gamma r}\right)^{-\gamma} dr^2 + \left(1 - \frac{2m}{\gamma r}\right)^{1-\gamma} r^2(d\theta^2 + \sin^2\theta d\phi^2),$$

  where m and γ are constants. For $\gamma = 1$ one obtains the **Schwarzschild metric**. In this case the scalar field vanishes.
- **FLRW metric**

  The **FLRW** (Friedmann–Lemaître–Robertson–Walker) **metric** is a exact solution to the Einstein field equation for a simply connected, homogeneous, isotropic expanding (or contracting) Universe filled with a constant density and negligible pressure. This metric represents a matter-dominated Universe filled with a *dust* (pressure-free matter); it models the **metric expansion of space**.

  Its *line element* is usually written in the *spherical coordinates* (ct, r, θ, ϕ):

  $$ds^2 = c^2dt^2 - a(t)^2\left(\frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right),$$

  where $a(t)$ is the **scale factor** and k is the *curvature* of the space-time.

[redacted]

- **Vaidya metric**

  The **Vaidya metric** is a inhomogeneous solution to Einstein field equation describing a spherically symmetric space-time composed purely of radially propagating radiation. It has been used to describe the radiation emitted by a shining star, by a collapsing star and by evaporating black hole.

  The Vaidya metric is a nonstatic generalization of the **Schwarzschild metric** and the radiation limit of the **LTB metric**. Let $M(u)$ be the mass parameter; the *line element* of this metric (Vaidya, 1953) is given by

$$ds^2 = -[1 - 2\frac{M(u)}{r}]du^2 + 2dudr + r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

- **LTB metric**

  The **LTB** Lemaître–Tolman–Bondi) **metric** is a solution to the Einstein field equation describing a spherical (finite or infinite) cloud of *dust* (pressure-free matter) that is expanding or collapsing under gravity.

  The LTB metric describes an inhomogeneous space-time expected on very large (Gpc) scale. It generalizes the **FLRW metric** and the **Schwarzschild metric**.

  The *line element* of this metric in the *spherical coordinates* is:

$$ds^2 = dt^2 - \frac{(R')^2}{1 + 2E}dr^2 - R^2(d\theta^2 + \sin^2\theta d\phi^2),$$

  where $R = R(t, r)$, $R' = \frac{\partial R}{\partial r}$, $E = E(r)$. The shell $r = r_0$ at a time $t = t_0$ has an area $4\pi R^2(r_0, t_0)$, and the areal radius R evolves with time as $\frac{\partial R}{\partial t} = 2E + \frac{2M}{R}$, where $M = M(r)$ is the gravitational mass within the comoving sphere at radius r.

- **Kantowski–Sachs metric**

  The **Kantowski–Sachs metric** is a solution to the Einstein field equation, given by the *line element*

$$ds^2 = -dt^2 + a(t)^2dz^2 + b(t)^2(d\theta^2 + \sin\theta d\phi^2),$$

  where the functions $a(t)$ and $b(t)$ are determined by the Einstein equation. It is the only homogeneous model without a 3D transitive subgroup.

  In particular, the Kantowski–Sachs metric with the *line element*

$$ds^2 = -dt^2 + e^{2\sqrt{\Lambda}t}dz^2 + \frac{1}{\Lambda}(d\theta^2 + \sin^2\theta d\phi^2)$$

  describes an anisotropic Universe with two spherical dimensions having a fixed size during the cosmic evolution and exponentially expanding 3-rd dimension.

- **Bianchi metrics**

  The **Bianchi metrics** are solutions to the Einstein field equation for cosmological models that have spatially homogeneous sections, invariant under the action of a 3D Lie group, i.e., they are real 4D metrics with a 3D isometry group, transitive on 3-surfaces. Using the Bianchi classification of 3D Lie algebras over Killing vector fields, we obtain the nine types of Bianchi metrics.

  Each Bianchi model B defines a transitive group G_B on some 3D simply connected manifold M; so, the pair (M, G) (where G is the maximal group acting on X and containing G_B) is one of eight Thurston *model geometries* if M/G' is compact for a discrete subgroup G' of G. In particular, Bianchi type IX corresponds to the geometry S^3. Only the model geometry $S^2 \times \mathbb{R}$ is not realized in this way.

  The Bianchi type I metric is a solution to the Einstein field equation for an anisotropic homogeneous Universe, given by the *line element*

  $$ds^2 = -dt^2 + a(t)^2dx^2 + b(t)^2dy^2 + c(t)^2dz^2,$$

  where the functions $a(t)$, $b(t)$, and $c(t)$ are determined by the Einstein equation. It corresponds to flat spatial sections, i.e., is a generalization of the **FLRW metric**.

  The Bianchi type IX metric, or **Mixmaster metric** (Misner, 1969), exhibits chaotic dynamic behavior near its curvature singularities.
- **Kasner metric**

  The **Kasner metric** is a Bianchi type I metric, which is a vacuum solution to the Einstein field equation for an anisotropic homogeneous Universe, given by

  $$ds^2 = -dt^2 + t^{2p_1}dx^2 + t^{2p_2}dy^2 + t^{2p_3}dz^2,$$

  where $p_1 + p_2 + p_3 = p_1^2 + p_2^2 + p_3^2 = 1$.

  The equal-time slices of Kasner space-time are spatially flat, but space contracts in one dimension (i with $p_i < 0$), while expanding in the other two. The volume of the spatial slices is proportional to t; so, $t \to 0$ can describe either a Big Bang or a *Big Crunch*, depending on the sense of t.
- **GCSS metric**

  A **GCSS** (i.e., **general cylindrically symmetric stationary**) **metric** is a solution to the *Einstein field equation*, given by the *line element*

  $$ds^2 = -fdt^2 + 2kdtd\phi + e^{\mu}(dr^2 + dz^2) + ld\phi^2,$$

  where the space-time is divided into two regions: the interior, with $0 \le r \le R$, to a cylindrical surface of radius R centered along z, and the exterior, with $R \le r < \infty$. Here f, k, μ and l are functions only of r, and $-\infty < t, z < \infty$, $0 \le \phi \le 2\pi$; the hypersurfaces $\phi = 0$ and $\phi = 2\pi$ are identical.

[redacted]

- **Lewis metric**

  The **Lewis metric** is a **cylindrically symmetric stationary metric** which is a solution to the *Einstein field equation* for empty space (vacuum) in the exterior of a cylindrical surface. The *line element* of this metric has the form

$$ds^2 = -fdt^2 + 2kdtd\phi - e^{\mu}(dr^2 + dz^2) + ld\phi^2,$$

  where $f = ar^{-n+1} - \frac{c^2}{n^2a}r^{n+1}$, $k = -Af$, $l = \frac{r^2}{f} - A^2f$, $e^{\mu} = r^{\frac{1}{2}(n^2-1)}$ with $A = \frac{cr^{n+1}}{naf} + b$. The constants n, a, b, and c can be either real or complex, the corresponding solutions belong to the *Weyl class* or *Lewis class*, respectively. In the last case, such metrics form a subclass of the *Kasner type metrics*.
- **van Stockum dust metric**

  The **van Stockum dust metric** is a stationary cylindrically symmetric solution to the Einstein field equation for empty space (vacuum) with a rigidly rotating infinitely long dust cylinder. The *line element* of this metric for the interior of the cylinder is given (in comoving, i.e., corotating, coordinates) by

$$ds^2 = -dt^2 + 2ar^2dtd\phi + e^{-a^2r^2}(dr^2 + dz^2) + r^2(1 - a^2r^2)d\phi^2,$$

  where $0 \leq r \leq R$, R is the radius of the cylinder, and a is the angular velocity of the dust particles. There are three vacuum exterior solutions (i.e., **Lewis metrics**) that can be matched to the interior solution, depending on the mass per unit length of the interior (the *low mass case*, the *null case*, and the *ultrarelativistic case*).

  Under some conditions (for example, if $ar > 1$), the existence of *closed time-like curves* (and, hence, time-travel) is allowed.
- **Levi-Civita metric**

  The **Levi-Civita metric** is a static cylindrically symmetric vacuum solution to the Einstein field equation, with the *line element*, given (in the Weyl form) by

$$ds^2 = -r^{4\sigma}dt^2 + r^{4\sigma(2\sigma-1)}(dr^2 + dz^2) + C^{-2}r^{2-4\sigma}d\phi,$$

  where the constant C refers to the deficit angle, and σ is a parameter.

  In the case $\sigma = -\frac{1}{2}$, $C = 1$ this metric can be transformed either into the *Taub's plane symmetric metric*, or into the *Robinson–Trautman metric*.
- **Weyl–Papapetrou metric**

  The **Weyl–Papapetrou metric** is a stationary axially symmetric solution to the Einstein field equation, given by the *line element*

$$ds^2 = Fdt^2 - e^{\mu}(dz^2 + dr^2) - Ld\phi^2 - 2Kd\phi dt,$$

  where F, K, L and μ are functions only of r and z, $LF + K^2 = r^2$, $\infty < t, z < \infty$, $0 \leq r < \infty$, and $0 \leq \phi \leq 2\pi$; the hypersurfaces $\phi = 0$ and $\phi - 2\pi$ are identical.

- **Bonnor dust metric**

  The **Bonnor dust metric** is a solution to the *Einstein field equation* which is an axially symmetric metric describing a cloud of rigidly rotating dust particles moving along circular geodesics about the z axis in hypersurfaces of $z =$ *constant*. The *line element* of this metric is given by

  $$ds^2 = dt^2 + (r^2 - n^2)d\phi^2 + 2ndtd\phi + e^{\mu}(dr^2 + dz^2),$$

  where, in Bonnor comoving (i.e., corotating) coordinates, $n = \frac{2hr^2}{R^3}$, $\mu = \frac{h^2r^2(r^2-8z^2)}{2R^8}$, $R^2 = r^2 + z^2$, and h is a rotation parameter.

  As $R \to \infty$, the metric coefficients tend to Minkowski values.

- **Weyl metric**

  The **Weyl metric** is a general static axially symmetric vacuum solution to the Einstein field equation given, in Weyl canonical coordinates, by the *line element*

  $$ds^2 = e^{2\lambda}dt^2 - e^{-2\lambda}\left(e^{2\mu}(dr^2 + dz^2\right) + r^2d\phi^2),$$

  where λ and μ are functions only of r and z such that $\frac{\partial^2\lambda}{\partial r^2} + \frac{1}{r}\cdot\frac{\partial\lambda}{\partial r} + \frac{\partial^2\lambda}{\partial z^2} = 0$, $\frac{\partial\mu}{\partial r} = r(\frac{\partial\lambda^2}{\partial r} - \frac{\partial\lambda^2}{\partial z})$, and $\frac{\partial\mu}{\partial z} = 2r\frac{\partial\lambda}{\partial r}\frac{\partial\lambda}{\partial z}$.

- **Zipoy–Voorhees metric**

  The **Zipoy–Voorhees metric** (or *γ-metric*) is a **Weyl metric**, obtained for $e^{2\lambda} = \left(\frac{R_1+R_2-2m}{R_1+R_2+2m}\right)^{\gamma}$, $e^{2\mu} = \left(\frac{(R_1+R_2+2m)(R_1+R_2-2m)}{4R_1R_2}\right)^{\gamma^2}$, where $R_1^2 = r^2+(z-m)^2$, $R_2^2 = r^2 + (z+m)^2$. Here λ corresponds to the Newtonian potential of a line segment of mass density $\gamma/2$ and length $2m$, symmetrically distributed along the z axis.

  The case $\gamma = 1$ corresponds to the **Schwartzschild metric**, the cases $\gamma > 1$ ($\gamma < 1$) correspond to an oblate (prolate) spheroid, and for $\gamma = 0$ one obtains the flat Minkowski space-time.

- **Straight spinning string metric**

  The **straight spinning string metric** is given by the *line element*

  $$ds^2 = -(dt - ad\phi)^2 + dz^2 + dr^2 + k^2r^2d\phi^2,$$

  where a and $k > 0$ are constants. It describes the space-time around a straight spinning string. The constant k is related to the string's mass-per-length μ by $k = 1–4\mu$, and the constant a is a measure of the string's spin. For $a = 0$ and $k = 1$, one obtains the **Minkowski metric** in cylindrical coordinates.

- **Tomimatsu–Sato metric**

  A **Tomimatsu–Sato metric** ([ToSa73]) is one of the metrics from an infinite family of spinning mass solutions to the Einstein field equation, each of which has the form $\xi = U/W$, where U and W are some polynomials.

  The simplest solution has $U = p^2(x^4 - 1) + q^2(y^4 - 1) - 2ipqxy(x^2 - y^2)$, $W = 2px(x^2 - 1) - 2iqy(1 - y^2)$, where $p^2 + q^2 = 1$. The *line element* for it is

[redacted]

given by

$$ds^2 = \Sigma^{-1}\left((\alpha dt + \beta d\phi)^2 - r^2(\gamma dt + \delta d\phi)^2\right) - \frac{\Sigma}{p^4(x^2 - y^2)^4}(dz^2 + dr^2),$$

where $\alpha = p^2(x^2-1)^2+q^2(1-y^2)^2$, $\beta = -\frac{2q}{p}W(p^2(x^2-1)(x^2-y^2)+2(px+1)W)$, $\gamma = -2pq(x^2 - y^2)$, $\delta = \alpha + 4((x^2 - 1) + (x^2 + 1)(px + 1))$, $\Sigma = \alpha\delta - \beta\gamma = |U + W|^2$.

- **Gödel metric**

  The **Gödel metric** is an exact solution to the Einstein field equation with cosmological constant for a rotating Universe, given by the *line element*

$$ds^2 = -(dt^2 + C(r)d\phi)^2 + D^2(r)d\phi^2 + dr^2 + dz^2,$$

where (t, r, ϕ, z) are the usual *cylindrical coordinates.*

The *Gödel Universe* is homogeneous if $C(r) = \frac{4\Omega}{m^2}\sinh^2\left(\frac{mr}{2}\right)$, $D(r) = \frac{1}{m}\sinh(mr)$, where m and Ω are constants. The Gödel Universe is singularity-free. But there are *closed time-like curves* through every event, and hence time-travel is possible here. The condition required to avoid such curves is $m^2 > 4\Omega^2$.

- **Conformally stationary metric**

  The **conformally stationary metrics** are models for gravitational fields that are time-independent up to an overall conformal factor. If some global regularity conditions are satisfied, the space-time must be a product $\mathbb{R} \times M^3$ with a (Hausdorff and paracompact) 3-manifold M^3, and the *line element* of the metric is given by

$$ds^2 = e^{2f(t,x)}(-(dt + \sum_{\mu}\phi_\mu(x)dx_\mu)^2 + \sum_{\mu,\nu} g_{\mu\nu}(x)dx_\mu dx_\nu),$$

where $\mu, \nu = 1, 2, 3$. The conformal factor e^{2f} does not affect the light-like geodesics apart from their parametrization, i.e., the paths of light rays are completely determined by the Riemannian metric $g = \sum_{\mu,\nu} g_{\mu\nu}(x)dx_\mu dx_\nu$ and the one-form $\phi = \sum_\mu \phi_\mu(x)dx_\mu$ which both live on M^3.

In this case, the function f is called the *redshift potential*, the metric g is called the **Fermat metric**. For a static space-time, the geodesics in the Fermat metric are the projections of the null geodesics of space-time.

In particular, the **spherically symmetric and static metrics**, including models for nonrotating stars and black holes, *wormholes*, *monopoles*, *naked singularities*, and (boson or fermion) stars, are given by the *line element*

$$ds^2 = e^{2f(r)}(-dt^2 + S(r)^2dr^2 + R(r)^2(d\theta^2 + \sin^2\theta d\phi^2)).$$

Here, the one-form ϕ vanishes, and the Fermat metric g has the special form

$$g = S(r)^2dr^2 + R(r)^2(d\theta^2 + \sin^2\theta d\phi^2).$$

For example, the conformal factor $e^{2f(r)}$ of the **Schwartzschild metric** is equal to $1 - \frac{2m}{r}$, and the corresponding Fermat metric has the form

$$g = (1 - \frac{2m}{r})^{-2}(1 - \frac{2m}{r})^{-1} r^2 (d\theta^2 + \sin\theta d\phi^2).$$

- **pp-wave metric**

  The **pp-wave metric** is an exact solution to the Einstein field equation, in which radiation moves at the speed c of light. The *line element* of this metric is given (in Brinkmann coordinates) by

  $$ds^2 = H(u, x, y)du^2 + 2dudv + dx^2 + dy^2,$$

  where H is any smooth function. The term “pp” stands for *plane-fronted waves with parallel propagation* introduced by Ehlers–Kundt, 1962.

  The most important class of particularly symmetric pp-waves are the **plane wave metrics**, in which H is quadratic. The *wave of death*, for example, is a *gravitational* (i.e., the space-time curvature fluctuates) plane wave exhibiting a strong nonscalar null curvature singularity which propagates through an initially flat space-time, progressively destroying the Universe.

  Examples of axisymmetric pp-waves include the *Aichelburg–Sexl ultraboost* modeling the motion past a spherically symmetric gravitating object at nearly c, and the *Bonnor beam* modeling the gravitational field of an infinitely long beam of incoherent electromagnetic radiation. The Aichelburg–Sexl wave is obtained by boosting the Schwarzschild solution to the speed c at fixed energy, i.e., it describes a Schwarzschild black hole moving at c. Cf. **Aichelburg–Sexl metric** (Chap. 24).
- **Bonnor beam metric**

  The **Bonnor beam metric** is an exact solution to the Einstein field equation which models an infinitely long, straight beam of light. It is an **pp-wave metric**.

  The interior part of the solution (in the uniform plane wave interior region which is shaped like the world tube of a solid cylinder) is defined by the *line element*

  $$ds^2 = -8\pi m r^2 du^2 - 2dudv + dr^2 + r^2 d\theta^2,$$

  where $-\infty < u, v < \infty$, $0 < r < r_0$, and $-\pi < \theta < \pi$. This is a null dust solution and can be interpreted as incoherent electromagnetic radiation.

  The exterior part of the solution is defined by

  $$ds^2 = -8\pi m r_0^2(1 + 2\log(r/r_0))du^2 - 2dudv + dr^2 + r^2 d\theta^2,$$

  where $-\infty < u, v < \infty$, $r_0 < r < \infty$, and $-\pi < \theta < \pi$.

[redacted]

- **Plane wave metric**

  The **plane wave metric** is a vacuum solution to the Einstein field equation, given by the *line element*

  $$ds^2 = 2dwdu + 2f(u)(x^2 + y^2)du^2 - dx^2 - dy^2.$$

  It is conformally flat, and describes a pure radiation field. The space-time with this metric is called the *plane gravitational wave*. It is an **pp-wave metric**.

- **Wils metric**

  The **Wils metric** is a solution to the *Einstein field equation*, given by

  $$ds^2 = 2xdwdu - 2wdudx + \left(2f(u)x(x^2 + y^2) - w^2\right) du^2 - dx^2 - dy^2.$$

  It is conformally flat, and describes a pure radiation field which is not a *plane wave*.

- **Koutras–McIntosh metric**

  The **Koutras–McIntosh metric** is a solution to the Einstein field equation, given by the *line element*

  $$ds^2 = 2(ax+b)dwdu-2awdudx+\left(2f(u)(ax + b)(x^2 + y^2) - a^2w^2\right) du^2-dx^2-dy^2.$$

  It is conformally flat and describes a pure radiation field which, in general, is not a *plane wave*. It gives the **plane wave metric** for $a = 0$, $b = 1$, and the **Wils metric** for $a = 1, b = 0$.

- **Edgar–Ludwig metric**

  The **Edgar–Ludwig metric** is a solution to the *Einstein field equation*, given by

  $$ds^2 = 2(ax + b)dwdu - 2awdudx+$$

  $$+ \left(2f(u)(ax + b)(g(u)y + h(u) + x^2 + y^2) - a^2w^2\right) du^2 - dx^2 - dy^2.$$

  This metric is a generalization of the **Koutras–McIntosh metric**. It is the most general metric which describes a conformally flat pure radiation (or null fluid) field which, in general, is not a *plane wave*. If plane waves are excluded, it has the form

  $$ds^2 = 2xdwdu-2wdudx+\left(2f(u)x(g(u)y + h(u) + x^2 + y^2) - w^2\right) du^2-dx^2-dy^2.$$

- **Bondi radiating metric**

  The **Bondi radiating metric** describes the asymptotic form of a radiating solution to the Einstein field equation, given by the *line element*

  $$ds^2 = -\left(\frac{V}{r}e^{2\beta} - U^2r^2e^{2\gamma}\right) du^2-2e^{2\beta}dudr-2Ur^2e^{2\gamma}dud\theta+r^2(e^{2\gamma}d\theta^2+e^{-2\gamma}\sin^2\theta d\phi^2),$$

where u is the retarded time, r is the **luminosity distance**, $0 \le \theta \le \pi$, $0 \le \phi \le 2\pi$, and U, V, β, γ are functions of u, r, and θ.

- **Taub–NUT de Sitter metric**

  The **Taub–NUT de Sitter metric** (cf. **de Sitter metric**) is a positive-definite (i.e., Riemannian) solution to the Einstein field equation with a cosmological constant Λ, given by the *line element*

$$ds^2 = \frac{r^2 - L^2}{4\Delta} dr^2 + \frac{L^2 \Delta}{r^2 - L^2}(d\psi + \cos\theta d\phi)^2 + \frac{r^2 - L^2}{4}(d\theta^2 + \sin^2\theta d\phi^2),$$

  where $\Delta = r^2 - 2Mr + L^2 + \frac{\Lambda}{4}(L^4 + 2L^2r^2 - \frac{1}{3}r^4)$, L and M are parameters, and θ, ϕ, ψ are the *Euler angles*.

  If $\Lambda = 0$, one obtains the **Taub–NUT metric** (Chap. 7).

- **Eguchi–Hanson de Sitter metric**

  The **Eguchi–Hanson de Sitter metric** (cf. **de Sitter metric**) is a positive-definite (i.e., Riemannian) solution to the Einstein field equation with a cosmological constant Λ, given by the *line element*

$$ds^2 = \left(1 - \frac{a^4}{r^4} - \frac{\Lambda r^2}{6}\right)^{-1} dr^2 + \frac{r^2}{4}\left(1 - \frac{a^4}{r^4} - \frac{\Lambda r^2}{6}\right)(d\psi + \cos\theta d\phi)^2 +$$

$$+ \frac{r^2}{4}(d\theta^2 + \sin^2\theta d\phi^2),$$

  where a is a parameter, and θ, ϕ, ψ are the *Euler angles*.

  If $\Lambda = 0$, one obtains the **Eguchi-Hanson metric**.

- **Barriola–Vilenkin monopole metric**

  The **Barriola–Vilenkin monopole metric** is given by the *line element*

$$ds^2 = -dt^2 + dr^2 + k^2r^2(d\theta^2 + \sin^2\theta d\phi^2),$$

  with a constant $k < 1$. There is a deficit solid angle and a singularity at $r = 0$; the plane $t =$ *constant*, $\theta = \frac{\pi}{2}$ has the geometry of a cone.

  This metric is an example of a conical singularity; it can be used as a model for *monopole*, i.e., a hypothetical isolated magnetic poles. It has been theorized that such things might exist in the form of tiny particles similar to electrons or protons, formed from topological defects in a similar manner to cosmic strings.

  Cf. **Gibbons–Manton metric** in Chap. 7.

- **Bertotti–Robinson metric**

  The **Bertotti–Robinson metric** is a solution to the Einstein field equation in a universe with a uniform magnetic field. The *line element* of this metric is

$$ds^2 = Q^2(-dt^2 + \sin^2 t dw^2 + d\theta^2 + \sin^2\theta d\phi^2),$$

  where Q is a constant, $t \in [0, \pi]$, $w \in (-\infty, +\infty)$, $\theta \in [0, \pi]$, and $\phi \in [0, 2\pi]$.

[redacted]

- **Wormhole metric**

  A *wormhole* is a hypothetical region of space-time containing a *world tube* (the time evolution of a closed surface) that cannot be continuously deformed to a *world line* (the time evolution of a point).

  Wormhole metric is a theoretical distortion of space-time in a region of the Universe that would link one location or time with another, through a "shortcut", i.e., a path that is shorter in distance or duration than would otherwise be expected. A wormhole geometry can only appear as a solution to the Einstein equations if the *stress-energy tensor* of matter violates the *null energy condition* at least in a neighborhood of the wormhole throat.

  Einstein–Rosen bridge (1935) is a *nontraversable* (unstable) wormhole formed from either black hole or spherically symmetric vacuum regions; it possesses a singularity and impenetrable event horizon. Traversable wormholes, as well as *warp drive* (faster-than-light propulsion system) and *time machines*, permitting journeys into the past, require bending of space-time by *exotic matter* (negative mass or energy).

  Whereas the curvature of space produced by the attractive gravitational field of ordinary matter acts like a converging lens, negative energy acts like a diverging lens. The negative mass required for engineering, say, a wormhole of throat diameter 4.5 m, as in Stargate's inner ring (from TV franchise *Stargate*), is $\approx -3\times 10^{27}$ kg. But oscillating warp and tweaking wormhole's geometry (White, 2012) can greatly reduce it. Also, negative energy can be created (Butcher, 2014) at the centre of a wormhole if its throat is orders of magnitude longer than its mouth.

  Lorentzian wormholes, not requiring exotic matter to exist, were proposed, using higher-dimensional extensions of Einstein's theory of gravity, by Bronnikov–Kim, 2002, and Kanti–Kleihaus–Kunz, 2011. Still, only atomic-scale wormholes would be practical to build, using them solely for superluminal information transmission.

  Lorentzian wormholes can be seen as maximally entangled states of two black holes in a *Einstein–Podolsky–Rosen correlation*, i.e., *nonclassical* one (it cannot be approximated by convex combinations of product states). Maldacena–Susskind, 2013: any two entangled quantum subsystems (cf. Chap. 24) are connected by a such wormhole. For Sonner, 2013, gravity might emerge from quantum entanglement.

- **Morris–Thorne metric**

  The **Morris–Thorne metric** (Morris–Thorne, 1988) is a traversable **wormhole metric** which is a solution to the Einstein field equation with the *line element*

$$ds^2 = e^{\frac{2\Phi(w)}{c^2}} c^2 dt^2 - dw^2 - r(w)^2 (d\theta^2 + \sin^2\theta d\phi^2),$$

  where $w \in [-\infty, +\infty]$, r is a function of w that reaches some minimal value above zero at some finite value of w, and $\Phi(w)$ is a gravitational potential allowed by the space-time geometry. It is the most general static and spherically symmetric metric able to describe a stable and traversable wormhole.

Morris–Thorne–Yurtsever, 1988, stated that two closely spaced ($10^{-9} - 10^{-10}$ m) concentric thin charged hollow spheres the size of 1 AU can create negative energy, required for engineering this wormhole, using the quantum *Casimir effect*.

- **Alcubierre metric**

The **Alcubierre metric** (Alcubierre, 1994) is a **wormhole metric** which is a solution to the Einstein field equation, representing *warp drive space-time* where the existence of *closed time-like curves* is allowed. The Alcubierre construction corresponds to a *warp* (i.e., faster than light) drive in that it causes space-time to contract in front of a spaceship bubble and expand behind, thus providing the spaceship with a velocity that can be much greater than the speed of light relative to distant objects, while the spaceship never locally travels faster than light.

In this case, only the relativistic principle that a space-traveler may move with any velocity up to, but not including or exceeding, the speed of light, is violated.

Also, light has another speed, *phase velocity*, measured by how fast a wavelength's crests move. This speed depends on the material it's moving through.

The *line element* of this metric has the form

$$ds^2 = -dt^2 + (dx - vf(r)dt)^2 + dy^2 + dz^2,$$

where $v = \frac{dx_s(t)}{dt}$ is the apparent velocity of the warp drive spaceship, $x_s(t)$ is spaceship trajectory along the coordinate x, the radial coordinate is $r = ((x - x_s(t))^2 + y^2 + z^2)^{\frac{1}{2}}$, and $f(r)$ an arbitrary function subject to the conditions that $f = 1$ at $r = 0$ (the location of the spaceship) and $f(\infty) = 0$.

Another warp drive space-time was proposed by Krasnikov, 1995. **Krasnikov metric** in the 2D subspace t, x is given by the *line element*

$$ds^2 = -dt^2 + (1 - k(x,t))dxdt + k(x,t)dx^2,$$

where $k(x,t) = 1 - (2 - \delta)\theta_\epsilon(t - x)(\theta_\epsilon(x) - \theta_\epsilon(x + \epsilon - D))$, D is the distance to travel, θ_ϵ is a smooth monotonic function satisfying $\theta_\epsilon(z) = 1$ at $z > \epsilon$, $\theta_\epsilon(z) = 0$ at $z < 0$ and δ, ϵ are arbitrary small positive parameters.

- **Misner metric**

The **Misner metric** (Misner, 1960) is a metric, representing two black holes, instantaneously at rest, whose throats are connected by a *wormhole*. The *line element* of this metric has the form

$$ds^2 = -dt^2 + \psi^4(dx^2 + dy^2 + dz^2),$$

[redacted]

where the *conformal factor* ψ is given by

$$\psi = \sum_{n=-N}^{N} \frac{1}{\sinh(\mu_0 n)} \frac{1}{\sqrt{x^2 + y^2 + (z + \coth(\mu_0 n))^2}}.$$

The parameter μ_0 is a measure of the ratio of mass to separation of the throats (equivalently, a measure of the distance of a loop in the surface, passing through one throat and out of the other). The summation limit N tends to infinity.

The topology of the *Misner space-time* is that of a pair of asymptotically flat sheets connected by a number of wormholes. In the simplest case, it can be seen as a 2D space $\mathbb{R} \times S^1$, in which light progressively tilts as one moves forward in time, and has *closed time-like curves* (so, time-travel is possible) after a certain point.

- **Rotating C-metric**

  The **rotating C-metric** is a solution to the *Einstein–Maxwell equations*, describing two oppositely charged black holes, uniformly accelerating in opposite directions. The *line element* of this metric has the form

$$ds^2 = A^{-2}(x+y)^{-2}\left(\frac{dy^2}{F(y)} + \frac{dx^2}{G(x)} + k^{-2}G(X)d\phi^2 - k^2A^2F(y)dt^2\right),$$

  where $F(y) = -1 + y^2 - 2mAy^3 + e^2A^2y^4$, $G(x) = 1 - x^2 - 2mAx^3 - e^2A^2x^4$, m, e, and A are parameters related to the mass, charge and acceleration of the black holes, and k is a constant fixed by regularity conditions.

  This metric should not be confused with the **C-metric** from Chap. 11.

- **Myers–Perry metric**

  The **Myers–Perry metric** describes a 5D rotating black hole. Its *line element* is

$$ds^2 = -dt^2 + \frac{2m}{\rho^2}(dt - a\sin^2\theta d\phi - b\cos^2\theta d\psi)^2 +$$

$$+\frac{\rho^2}{R^2}dr^2 + \rho^2 d\theta^2 + (r^2 + a^2)\sin^2\theta d\phi^2 + (r^2 + b^2)\cos^2\theta d\psi^2,$$

  where $\rho^2 = r^2 + a^2\cos^2\theta + b^2\sin^2\theta$, and $R^2 = \frac{(r^2+a^2)(r^2+b^2)-2mr^2}{r^2}$. Above black hole is asymptotically flat and has an event horizon with S^3 topology.

  Emparan and Reall, 2001, using the possibility of rotation in several independent rotation planes, found a 5D *black ring*, i.e., asymptotically flat black hole solution with the event horizon's topology of $S^1 \times S^2$.

- **Ponce de León metric**

  The **Ponce de León metric** (1988) is a 5D metric, given by the *line element*

$$ds^2 = l^2dt^2 - (t/t_0)^2 p l^{\frac{2p}{p-1}}(dx^2 + dy^2 + dz^2) - \frac{t^2}{(p-1)^2}dl^2,$$

where l is the 5-th (space-like) coordinate. This metric represents a 5D apparent vacuum. It is not flat but embed the flat 4D **FLRW metric**.

- **Kaluza–Klein metric**

  The **Kaluza–Klein metric** is a metric in the *Kaluza–Klein model* of 5D space-time which seeks to unify classical gravity and electromagnetism.

  Kaluza, 1921 (but sent to Einstein in 1919), found that, if the Einstein theory of pure gravitation is extended to a 5D space-time, the Einstein field equation can be split into an ordinary 4D gravitation tensor field, plus an extra vector field which is equivalent to the Maxwell equation for the electromagnetic field, plus an extra scalar field known as the *dilation* (or *radion*).

  Klein, 1926, assumed that the 5-th dimension (i.e., 4-th spatial dimension) is curled up in a circle of an unobservable size, below 10^{-20} m. Almost all modern higher-dimensional unified theories are based on Kaluza–Klein approach.

  An alternative proposal is that the extra dimension(s) is extended, and the matter is trapped in a 4D submanifold. In a model of a such large extra dimension, the 5D metric of a universe can be written in Gaussian normal coordinates as

$$ds^2 = -(dx_5)^2 + \lambda^2(x_5)\sum_{\alpha,\beta}\eta_{\alpha\beta}dx_\alpha dx_\beta,$$

  where $\eta_{\alpha\beta}$ is the 4D **metric tensor** and $\lambda^2(x_5)$ is any function of the 5-th coordinate.

  In particular, the *STM* (space-time-matter) *theory* (Wesson and Ponce de León, 1992) relate the 5-th coordinate to mass via either $x_5 = \frac{Gm}{c^2}$ or $x_5 = \frac{h}{mc}$, where G is the *Newton gravitational constant* and h is the *Planck constant*.

  The **Ponce de León metric** is a STM solution. In STM (or *induced matter*) theory, the 4D curvature arises not due to the distribution of matter in the Universe (as claims Relativity Theory) but because the Universe is embedded in some higher-dimensional vacuum manifold M, and all the matter in our world can be thought of as being manifestations of the geometrical properties of M.

  Wesson and Seahra, 2005, claim that the Universe may be a 5D black hole. Life is not excluded since in 5D there is no physical plughole and the "tidal" forces are negligible. Suitable manifolds for such STM theory are given by two isometric solutions of the 5D vacuum field equations: *Liu–Mashhoon–Wesson metric* and *Fukui–Seahra–Wesson metric*; both embed 4D **FLRW metric**.

- **Carmeli metric**

  The **Carmeli metric** (Carmeli, 1996) is given by the *line element*

$$ds^2 = dx^2 + dy^2 + dz^2 - \tau^2 dv^2,$$

  where $\tau = \frac{1}{H}$ is the inverse of *Hubble constant* and v is the *cosmological recession velocity*. So, comparing with the **Minkowski metric**, it has τ and velocity v, instead of c and time t. This metric was used in *Carmeli's Relativity Theory* which is intended to be better than General Relativity on cosmological scale.

[redacted]

The Carmeli metric produces the *Tulli–Fisher type relation* in spiral galaxies: 4-th power of the rotation speed is proportional to the mass of galaxy; it obviate the need for dark matter. This metric predicts also cosmic acceleration.

Including *icdt* component of the **Minkowski metric**, gives the **Kaluza–Klein–Carmeli metric** (Harnett, 2004) defined by

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2dt^2 - \tau^2 dv^2.$$

- **Prasad metric**

  A *de Sitter Universe* can be seen as the sum of the external and internal space.

  The internal space has a negative constant curvature $-\frac{1}{r^2}$ and can be characterized by the symmetry group $SO_{3,2}$. The **Prasad metric** of this space is given, in hyperspherical coordinates, by the *line element*

  $$ds^2 = r^2 \cos^2 t(d\chi^2 + \sinh^2 \chi(d\theta^2 + \sin^2 \theta d\phi^2)) - r^2dt^2.$$

  The value $\sin \chi$ is called *adimensional normalized radius* of the de Sitter Universe.

  The external space has constant curvature $\frac{1}{R^2}$ and can be characterized by the symmetry group $SO_{4,1}$. Its metric has the *line element* of the form

  $$ds^2 = R^2 \cosh^2 t(d\chi^2 + \sin^2 \chi(d\theta^2 + \sin^2 \theta d\phi^2)) - R^2dt^2.$$

Part VII
Real-World Distances

[redacted]

Chapter 27
Length Measures and Scales

The term *length* has many meanings: distance, extent, linear measure, span, reach, end, limit, etc.; for example, the length of a train, a meeting, a book, a trip, a shirt, a vowel, a proof. The *length* of an object is its linear extent, while the *height* is the vertical extent, and *width* (or *breadth*) is the side-to-side distance at 90° to the length, wideness. The *depth* is the distance downward, distance inward, deepness, profundity, drop.

The ancient Greek mathematicians saw all numbers as lengths (of segments), areas or volumes. In Mathematics, a **length function** is a function $l : G \to \mathbb{R}_{\geq 0}$ on a group $(G, +, 0)$ such that $l(0) = 0$ and $l(g) = l(-g)$, $l(g + g') \leq l(g) + l(g')$ for $g, g' \in G$.

In Engineering and Physics, "length" usually means "distance". **Unit distance** is a distance taken as a convenient unit of length in a given context.

In this chapter we consider length only as a measure of physical distance. We give selected information on the most important length units and present, in length terms, a list of interesting physical objects.

27.1 Length Scales

The main length measure systems are: Metric, Imperial (British and American), Japanese, Thai, Chinese Imperial, Old Russian, Ancient Roman, Ancient Greek, Biblical, Astronomical, Nautical, and Typographical.

There are many other specialized length scales; for example, to measure cloth, shoe size, gauges (such as interior diameters of shotguns, wires, jewelry rings), sizes for abrasive grit, sheet metal thickness, etc.

Many units express relative or inverse distances. Quantities measured in **reciprocal length** include: radius of curvature, density of a linear feature in an area, surface-area-to-volume ratio, vector's magnitude (in Crystallography and Spectroscopy),

© Springer-Verlag Berlin Heidelberg 2016
M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_27

optical power (cf. **lens distances** in Chap. 29), absorption (or attenuation) coefficient (cf. Chap. 24), gain (in Laser Physics). Common units used for such measures are inverse meter m^{-1} (called *diopter* in Optics), cm^{-1} and cycles per unit length (for spatial frequency).

Some units express nonlength quantities in length terms. For example, the *denudation rate* (wearing down of the Earth's surface) is measured in cm per 1000 years. Cf. also **meter of water equivalent** in Chap. 24.

- **International Metric System**

  The **International Metric System** (or SI, short for *Système International*), also known as MKSA (meter–kilogram–second–ampere), is a modernized version of the system of units, established by the *Treaty of the Meter* from 20 May 1875, which provides a logical and interconnected framework for all measurements in science, industry and commerce. The system is built on a foundation consisting of the following seven *SI base units*, assumed to be mutually independent:

  (1) length: **meter** (m); it is equal to the distance traveled by light in a vacuum in 1/299,792,458 of a second; (2) time: *second* (s); (3) mass: *kilogram* (kg); (4) temperature: *kelvin* (K); (5) electric current: *ampere* (A); (6) luminous intensity: *candela* (cd); (7) amount of substance: *mole* (mol).

  Meter, second, candela are already defined in terms of fundamentals of nature. Kilogram, kelvin, mole and ampere are expected to be redefined in terms of the Planck, Boltzmann, Avogadro constants and the charge of proton, respectively.

  Meter is defined as a **proper length** (the length of the object in its rest frame, cf. Sect. 26.2). So, it is well defined only over short distances where relativistic effects are negligible (cf. **Lorentz length contraction** in Sect. 26.1), and all cosmic distances, given in meters, are approximations.

  Originally, on March 26, 1791, the *mètre* (French for meter) was defined as $\frac{1}{10{,}000{,}000}$ of the distance from the North Pole to the equator along the Dunkirk-Barcelona meridian. The name *mètre* was derived from the Greek *metron* (measure). In 1799 the standard of *mètre* became a meter-long platinum-iridium bar kept in Sèvres, a town outside Paris, for people to come and compare their rulers with. (The metric system, introduced in 1793, was so unpopular that Napoleon was forced to abandon it and France returned to the *mètre* only in 1837.) In 1960–1983, the **meter** was defined in terms of wavelengths.

  The initial metric unit of mass, the *gram*, was defined as the mass of one cubic centimeter of water at its temperature of maximum density. A *metric ton* (or *metric tonne*, *tonne*) is a unit of mass equal to 1000 kg; this non-SI unit is used instead of the SI term *megagram* (10^6 grams). For capacity, the *litre* (liter) was defined as the volume of a cubic decimeter.

- **Metrication**

  The **metrication** is an ongoing (especially, in the US, the UK and Caribbean countries) process of conversion to the **International Metric System**, SI. Only the US, Liberia and Myanmar have not fully switched to SI.

[redacted]

For example, the US uses only miles for road distance signs (milestones). Altitudes in aviation are usually described in feet, and resolutions of output devices are specified in *dpi* (dots per inch). In shipping, nautical miles and knots are used; both are accepted for use with SI.

Hard metric means designing in the metric measures from the start and conformation, where appropriate, to internationally recognized sizes and designs.

Soft metric means multiplying an inch-pound number by a metric conversion factor and rounding it to an appropriate level of precision; so, the soft converted products do not change size. The *American Metric System* consists of converting traditional units to embrace the uniform base 10 used by the Metric System.

Such SI-Imperial hybrid units, used in soft metrication, are, for example, *kiloyard* (914.4 m), *kilofoot* (304.8 m), *mil* or *milli-inch* (25.4 micron), and *min* or *microinch* (25.4 nm). The *metric inch* (2.5 cm $\approx$ 1 inch) and *metric foot* (30 cm) were used in some Soviet computers when building from American blueprints.

In athletics and skating, races of 1500 m or 1600 m are often called *metric miles*. Examples of traditional units adapted to the meter are Chinese *li* = 500 m = 1500 *chi* (Chinese feet), Thai *wa* = 2 m = 4 *sok*, Vietnamese *xich* = 1 m = 1000 *ly*.

- **Meter, in Poetry and Music**

In Poetry, *meter* (or *cadence*) is a measure of rhythmic quality, the regular linguistic sound patterns of a verse or line in it. The meter of a verse is the number of lines, the number of syllables in each line and their arrangement as sequences of *feet*. Each foot is a specific sequence of syllable types—such as unstressed/stressed or long/short. Fussell, 1965, define four types of meter: syllabic, accentual, accentual-syllabic and quantitative, where patterns are based on *syllable weight* (number and/or duration of segments in the rhyme) rather than stress.

Hypermeter is part of a verse with an extra syllable; *metromania* is a mania for writing verses and *metrophobia* is a fear/hatred of poetry.

In Music, *meter* (or *metre*) is the regular rhythmic patterns of a musical line, the division of a composition into parts of equal time, and the subdivision of them. It is derived from the poetic meter of song. Different tonal preferences in voiced speech are reflected in music; it explains why Eastern and Western music differ.

Metrical rhythm is where each time value is a multiple or fraction of a fixed unit (*beat*) and normal accents re-occur regularly providing systematic grouping (*measures*). *Isometre* is the use of a *pulse* (unbroken series of periodically occurring short stimuli) without a regular meter, and *polymetre* is the use of two or more different meters simultaneously, whereas *multimetre* is using them in succession.

A rhythmic pattern/unit is either *intrametric* (confirming the pulses on the metric level), or *contrametric* (syncopated, not following the beat/meter), or *extrametric* (irregular with respect to the metric structure of the piece).

Rhythms/chords with the same multiset of intervals/distances are called *homometric*.

A temporal pattern is *metrically represented* if it can be subdivided into equal time intervals. A *metronome* is any device that produces regular, metrical ticks (beats); *metronomy*: measurement of time by a metronome or, in general, an instrument.

- **Meter-related terms**

  We present this large family of terms by the following examples (besides the unit of length and use in Poetry and Music).

  Metrograph: a device attached to a locomotive to record its speed and the number and duration of its stops. Cf. unrelated *metrography* in Medicine (Chap. 29).

  Metrogon: a high resolution, low-distortion, extra-wide field photographic lens design used extensively in aerial photography.

  The names of various measuring instruments contain *meter* at the end, say, *ammeter*, *gas meter*, *multimeter* (or *volt-ohm meter*).

  Metrosophy: a cosmology based on strict number correspondences.

  Metrology: the science of, or a system of, weights and measures.

  A **metric meterstick** is a rough rule of thumb for comprehending a metric unit; for example, 5 cm is the side of a matchbox, and 1 km is $\approx$ 10 minutes' walk.

  Metering: an equivalent term for a *measurement* (assignment of numbers to objects or events); *micrometry*: measurement under the microscope; *hypsometry*: measurement of heights; *telemetry*: technology that allows remote measurement; *archeometry*: the science of exact measuring referring to the remote past.

  Hedonimetry: the study of happiness as a measurable economic asset; *psychometry*: alleged psychic power enabling one to divine facts by handling objects.

  Psychometrics: the theory and technique of psychological measurement; *psychrometrics*: the determination of physical and thermodynamic properties of gas-vapor mixtures; *biometrics*: the identification of humans by their characteristics or traits; *cliometrics*: the systematic application of econometric techniques and other formal or mathematical methods to the study of history.

  Metric, as a nonmathematical term, is a standard unit of measure (for example, *font metrics* refer to numeric values relating to size and space in the font) or, more generally, part of a system of parameters; cf. **quality metrics** in Chap. 29.

  Antimetric matrix: a square matrix A with $A = -A^T$; an *antimetric electrical network* is one that exhibits antisymmetrical electrical properties.

  Isometropia: equality of refraction in both eyes; *hypermetropia* is farsightedness.

  Isometric particle: a virus which (at the stage of virion capsid) has icosahedral symmetry. *Isometric process*: a thermodynamic process at constant volume.

  Metrohedry: overlap in 3D of the lattices of twin domains in a crystal.

  Multimetric crystallography: to consider (Janner, 1991), in addition to the Euclidean metric tensor, *pseudo-Euclidean tensors* (hyperbolic rotations)

[redacted]

attached to the same basis; cf. **pseudo-Euclidean distance** in Chap. 7 and **multimetric** in Chap. 3.

Metria: a genus of moths of the *Noctuidae* family.

Metrio: Greek coffee with one teaspoon of sugar (medium sweet). In Anthropology, *metriocranic* means having a skull that is moderately high compared with its width, with a breadth-height index 92–98.

Metroid: the name of a series of video games produced by Nintendo and *metroids* are a fictional species of parasitic alien creatures from those games.

Examples of companies with a meter-related name are: Metron, Metric Inc., MetaMetrics Inc., Metric Engineering, Panametric, Prometric, Unmetric, World Wide Metric. *Metric* is also a Canadian New Wave rock band.

- **Metric length measures**

kilometer (km) $= 1000$ meters $= 10^3$ m;
meter (m) $= 10$ decimeters $= 10^0$ m;
decimeter (dm) $= 10$ centimeters $= 10^{-1}$ m;
centimeter (cm) $= 10$ millimeters $= 10^{-2}$ m;
millimeter (mm) $= 1000$ micrometers $= 10^{-3}$ m;
micrometer (or non-SI *micron*; µm) $= 1000$ nanometers $= 10^{-6}$ m;
nanometer (or non-SI 10 *angstroms* Å; nm) $= 1000$ picometers $= 10^{-9}$;
picometer (pm) $= 1000$ femtometers $= 10^{-12}$ m;
femtometer (or non-SI *fermi*; fm) $= 1000$ attometers $= 10^{-15}$ m.

The numbers 10^{3t} ($t = -8, \dots, -1, 1, \dots, 8$) are given by *metric prefixes*: yocto-(y), zepto-(z), atto-(a), femto-(f), pico-(p), nano-(n), micro-(μ), milli-(m), kilo-(k), mega-(M), giga-(G), tera-(T), peta-(P), exa-(E), zetta-(Z), yotta-(Y), respectively, while 10^t ($t = -2, -1, 1, 2$) are given by centi-(c), deci-(d), deca-(da), hecto-(h).

But for plankton's size femto-, pico-, nano-, micro-, meso-, macro-, mega- mean $< 0.2\mu, 0.2$–$2\mu, 2$–$20\mu, 20$–200μ, 0.2–20 mm, 20–200 mm, > 200 mm.

In computers, a *bit* (binary digit) is the basic unit of information, a *byte* (or *octet*) is 8 bits, and 10^{3t} bytes for $t = 1, \dots, 8$ are kilo-(KB), mega-(MB), giga-(GB), ..., yottabyte (YB), respectively. Sometimes (because of $2^{10} = 1024 \approx 10^3$) the binary terms kibi-(KiB), mebi-(MiB), gibibyte (GiB), etc., are used for 2^{10t} bytes.

- **Imperial length measures**

The **Imperial length measures** (as slightly adjusted by a treaty in 1959) are:
(land) *league* $= 3$ international miles;
(international) *mile* $= 5280$ feet $= 1609.344$ m;
(US survey) *mile* $= 5280$ US feet ≈ 1609.347 m;
data (or *tactical*) *mile* $= 6000$ feet $= 1828.8$ m and *radar mile* $= 12.204\ \mu$s (time it takes a radar pulse to travel one data mile forth and back);
(international) *yard* $= 0.9144$ m $= 3$ *feet* $= \frac{1}{2}$ *fathom*;
(international) *foot* $= 0.3048$ m $= 12$ *inches*;
(international) *inch* $= 2.54$ cm $= 12$ *lines*;

(a unit of measure of height of equipment) *rack unit* $= \frac{7}{4}$ inch;

(a unit of measure in advertising space) *agate line* $\approx \frac{1}{14}$ inch;

(a unit of computer mouse movement) *mickey* $= \frac{1}{200}$ inch;

mil (British *thou*) $= \frac{1}{1000}$ inch; *mil* is also an angular measure $\frac{\pi}{3200} \approx 0.001$ radian.

In addition, *Surveyor's Chain measures* are: *furlong* $=$ 10 chains $= \frac{1}{8}$ mile; *chain* $=$ 100 links $=$ 66 feet; *rope* $=$ 20 feet; *rod* (or *pole*) $=$ 16.5 feet; *link* $=$ 7.92 inches. Mile, furlong and fathom come from the slightly shorter Greco-Roman milos (milliare), stadion and orguia, mentioned in the New Testament.

For measuring cloth, old measures are used: *bolt* $=$ 40 yards; *goad* $= \frac{3}{2}$ yard; *ell* $= \frac{5}{4}$ yard $=$ 45 inches; *quarter* $= \frac{1}{4}$ yard; *finger* $= \frac{1}{8}$ yard; *nail* $= \frac{1}{16}$ yard.

Other old English units of length: *barleycorn* $= \frac{1}{3}$ inch; *digit* $= \frac{3}{4}$ inches and *palm*, *hand*, *shaftment*, *span*, *cubit* $= 3, 4, 6, 9, 18$ inches, respectively.

- **Cubit**

  The **cubit**, originally the length of the forearm from the elbow to the tip of the middle finger, was the ordinary unit of length in the ancient Near East which varied among cultures and with time. It is the oldest recorded measure of length.

  The cubit was used, in the temples of Ancient Egypt from at least 2700 BC, as follows: 1 *ordinary Egyptian cubit* $=$ 6 *palms* $=$ 24 *digits* $=$ 45 cm (18 inches), and 1 *royal Egyptian cubit* $=$ 7 *palms* $=$ 28 *digits* $\approx$ 52.6 cm. Relevant Sumerian measures were: 1 *ku* $=$ 30 *shusi* $=$ 25 *uban* $=$ 50 cm, and 1 *kus* $=$ 36 *shusi*.

  Biblical measures of length are the *cubit* and its multiples by $4, \frac{1}{2}, \frac{1}{6}, \frac{1}{24}$ called *fathom*, *span*, *palm*, *digit*, respectively. But the length of this cubit is unknown; it is estimated now as $\approx$ 44.5 cm (as Roman *cubitus*) for the common cubit, used in commerce, and 51–56 cm for the sacred one, used for building.

  The *Talmudic cubit* is 48–57.6 cm. The *pyramid cubit* (25.025 inches $\approx$ 63.567 cm), derived in Newton's Biblical studies, is supposed to be the basic one in the dimensions of the Great Pyramid and far-reaching numeric relations on them.

  Thom, 1955, claim that the *megalithic yard*, 82.966 cm, was the basic unit used for stone circles in Britain and Brittany c. 3500 BC. Butler–Knight, 2006, derived this unit as $1/(360 \times 366^2)$-th of 40,075 km (the Earth's equatorial circumference), linking it to the putative Megalithic 366-degree circle and Minoan 366-day year. Such a "366 geometry" is a part of the pseudoscientific metrology

- **Nautical length units**

  The main **nautical length units** (also used in aerial navigation) are:

  sea league $=$ 3 sea (nautical) miles;

  nautical mile $=$ 1852 m (originally defined as 1 min of arc of latitude);

  geographical mile $\approx$ 1855.32 m (the average distance on the Earth's surface, represented by 1 min of arc along the Earth's equator);

  (international) *short cable length* $= \frac{1}{10}$ nautical mile $=$ 185.2 m $\approx$ 101 fathoms;

[redacted]

(US customary) *cable length* = 120 fathoms = 720 feet = 219.456 m;
fathom = 6 feet = 1.8288 m.

- **Preferred design sizes**

  Objects are often manufactured in a series of sizes of increasing magnitude. In Industrial Design, *preferred numbers* are standard guidelines for choosing such product sizes within given constraints of functionality, usability, compatibility, safety or cost. **Preferred design sizes** are such lengths, diameters and distances.

  Four basic *Renard's series* of preferred numbers divide the interval from 10 to 100 into 5, 10, 20, or 40 steps, with the factor between two consecutive numbers being constant (before rounding): the 5-th, 10-th, 20-th, or 40-th root of 10. Since the **International Metric System** (SI) is decimally-oriented, the International Organization for Standardization (ISO) adopted Renard's series as the main preferred numbers for use in setting metric sizes. But, for example, the ratio between adjacent terms (i.e., notes) in the Western musical scale is 12-th root of 2.

  In the widely used ISO paper size system, the height-to-width ratio of all pages is the *Lichtenberg ratio*, i.e., $\sqrt{2}$. The system consists of formats An, Bn and (used for envelopes) Cn with $0 \leq n \leq 10$, having widths $2^{-\frac{1}{4}-\frac{n}{2}}$, $2^{-\frac{n}{2}}$ and $2^{-\frac{1}{8}-\frac{n}{2}}$, respectively. The above measures are in m; so, the area of An is 2^{-n} m^2. They are rounded and expressed usually in mm; for example, format A4 is 210 × 297 and format B7 (used also for EU and US passports) is 88 × 125.

- **Typographical length units**

  PostScript point = $\frac{1}{72}$ inch = 100 gutenbergs = 0.3527777778 mm;
  TeX point (or *printer's point*) = $\frac{1}{72.27}$ inch = 0.3514598035 mm;
  ATA point (or *Anglo-Saxon point*) = $\frac{1}{72.272}$ inch = 0.3514598 mm;
  point (Didot, European) = 0.37593985 mm, *cicero* = 12 points Didot;
  pica (Postscript, TeX or ATA) = 12 points in the corresponding system;
  twip = $\frac{1}{20}$ of a point in the corresponding system.

  In display systems, *twip* is $\frac{1}{1440}$ inch, and *himetric* is 0.01 mm.

- **Astronomical system of units**

  The **astronomical system of units** (or, formally, *IAU (1976) System of Astronomical Constants*) is a system of measurement developed for use in astronomy by the IAU (International Astronomical Union).

  It is based on units of length (AU), mass ($M_\odot$), and time (day). The speed of light in IAU is defined as $c_0 = 299{,}792{,}458$ m/s.

  The *astronomical unit of time* is the day $D = 24 \times 60^2 = 86{,}400$ s. 365.25 days make up one *Julian year*.

  The *astronomical unit of mass* is the mass of the Sun: $M_\odot = 1.98892 \times 10^{30}$ kg.

  The *astronomical unit of length* (*AU* or *ua*) is $c_0 t_A = 149{,}597{,}870{,}691 \pm 3$ m $\approx$ mean Earth–Sun distance; here t_A is the transit time of light across 1 AU.

  Among derived units are *light-year* $c_0 D \times 365.25 = 0.94607304725808 \times 10^{16}$ m and *parsec* $\frac{648{,}000}{\pi}$ AU $\approx 206{,}264.81$ AU $\approx 3.085677581 \times 10^{16}$ m.

- **Length units in Astronomy**

  The **Hubble distance** (Chap. 26) or *Hubble length* is $D_H = \frac{c}{H_0} \approx 1.31 \times 10^{26}$ m $\approx$ 4.237 Gpc $\approx$ 13.82 Gly (used to measure distances $d > \frac{1}{2}$ Mpc in terms of redshift z: $d = zD_H$ if $z \leq 1$, and $d = \frac{(z+1)^2-1}{(z+1)^2+1} D_H$, otherwise).

  gigaparsec (Gpc) = 10^3 megaparsec (Mpc) = 10^6 kiloparsec (kpc) = 10^9 parsecs;
  hubble (or light-gigayear, light-Ga, Gly) = 10^3 million light-years (Mly);
  siriometer = 10^6 AU $\approx$ 15.813 ly (about twice the Earth-Sirius distance);
  parsec (pc) = $\frac{648{,}000}{\pi} \approx$ 206,265 AU $\approx$ 3.262 light-years = 3.086×10^{16} m;
  light-year (ly) $\approx 9.461 \times 10^{15}$ m $\approx \pi \times 10^7$ light-seconds $\approx$ 0.307 parsec;
  spat (used formerly) = 10^{12} m = 10^3 gigameters $\approx$ 6.685 AU;
  astronomical unit (AU) $\approx 1.496 \times 10^{11}$ m $\approx$ 499 light-seconds;
  light-second $\approx 2.998 \times 10^8$ m (the Earth-Moon distance is $\approx$ 1.28 light-seconds);
  radii of Moon, Earth, Jupiter and Sun: 1737, 6371, 69,911 and 695,510 km;
  picoparsec $\approx$ 30.86 km; cf. other funny units such as *sheppey* 1.4 km (closest distance at which sheep remain picturesque), *beard-second* 5 nm (distance that a beard grows in a second), *microcentury* $\approx$ 52.5 min (lecture's length), *nanocentury* $\approx \pi$ sec.

- **Natural length units**

  Natural units are units of measurement based only on physical constants, for example, the speed c of light, gravitational constant G, reduced Planck constant $\hbar$, *Boltzmann constant* k_B, *Coulomb's constant* k_e, proton's *elementary charge* e, *fine-structure constant* $\alpha = \frac{e^2 k_e}{\hbar c} \approx \frac{1}{137}$ and masses m_e, m_p of electron and proton.

  Planck length (smallest measurable length) is $l_P = \sqrt{\frac{\hbar G}{c^3}} \approx 1.6162 \times 10^{-35}$ m. (The *Stoney length*, used formerly, is $\sqrt{\alpha} l_P$.) l_P is the **reduced Compton wavelength** $\bar{\lambda}_C(m) = \frac{m_P}{m} l_P$, and also half of the **Schwarzschild radius** $r_s(m) = 2Gmc^{-2}$ (Chap. 24) for $m = m_P$ (*Planck mass*) $= \sqrt{\frac{\hbar c}{G}} \approx 2.18 \times 10^{-8}$ kg $\approx$ 22 mg.

  The remaining base Planck units are *Planck time* $t_p = \frac{l_P}{c} \approx 5.39 \times 10^{-44}$ s, *Planck temperature* $T_P = \frac{m_P c^2}{k_B} \approx 1.42 \times 10^{32}$ K, and *Planck charge* $q_P = \sqrt{\frac{\hbar c}{k_e}} = \frac{e}{\sqrt{\alpha}} \approx 1.88 \times 10^{-18}$ C. The *Planck area* A_P is l_P^2, *Planck energy* E_P is $m_P c^2 \approx 1.22 \times 10^{28}$ eV $\approx$ 500 kWh, and *Planck density* ρ_P is $m_P \mathrm{l}_P^{-3} \approx 5.16 \times 10^{96}$ kg/m^3. Only black holes exceed ρ_P; some theories (for example, *Landau poles*) allow to exceed T_P.

  The Planck units come by a normalization of the *geometrized units* for the expressing SI units second, kilogram, kelvin and coulomb as c, $\frac{G}{c^2}$, $\frac{Gk_B}{c^4}$ and $\frac{\sqrt{Gk_e}}{c^2}$ m, respectively.

  The length unit of Quantum Chromodynamics (or strong interactions) is $\bar{\lambda}_C(m_p) \approx 2.103 \times 10^{-16}$ m. The majority of lengths, used in experiments on

[redacted]

nuclear fundamental forces, are integer multiples of $\lambda_C(m_p) = 2\pi\overline{\lambda}_C(m_p) \approx 1.32$ fm.

X unit $\approx 1.002 \times 10^{-13}$ m ≈ 0.1 pm measures wavelengths of X- and gamma rays.

The atomic unit of length is the **Bohr radius** (or *bohr*) $\alpha_0 \approx 5.292 \times 10^{-11}$ m ≈ 53 pm $= 0.53$ Å, the most probable distance between the proton and electron in a hydrogen atom. It is $\alpha^{-1}\overline{\lambda}_C(m_e) = \alpha^{-2}r_e$, where $r_e \approx 2.818 \times 10^{-15}$ m is the **Thomson scattering length** (Chap. 24), i.e., the *classical electron radius*.

In the units of Particle Physics, $1/rmeV = 10^{-9}$/GeV is $\frac{\hbar c}{\text{eV}} = \frac{E_P}{\text{eV}} l_P \approx 1.97 \times 10^{-7}$ m.

- **Length scales in Physics**

In Physics, a **length scale** (or **distance scale**) is a distance range determined with the precision of a few orders of magnitude, within which given phenomena are consistently described by a theory. Roughly, the scales $< 10^{-15}$, 10^{-15}–10^{-6}, $10^{-6} - 10^{6}$ and $> 10^{6}$ m are called *subatomic*, *atomic to cellular* (microscopic), *human* (macroscopic) and *astronomical*, respectively.

Bacteria (and human ovums) are roughly on the geometrical mean (10^{-4} m) of Nature's hierarchy of sizes. Dawkins, 2006, used term **middle world** for our realm between two counterintuitive extreme levels of existence: the microscopic world of quarks/atoms and the Universe at the galactic/universal level. The limit scales correspond to the **Planck length** l_P and **Hubble distance** $\approx 4.6 \times 10^{61}\ l_P$.

In terms of their constituents, Chemistry (molecules, atoms), Nuclear (say, proton, electron, photon), Hadronic (exited states) and Standard Model (quarks and leptons) are applicable at scales $\geq 10^{-10}$, $\geq 10^{-14}$, $\geq 10^{-15}$ and $\geq 10^{-18}$ m.

At the *meso-* (or *nanoscopic*) *scale*, $10^{-9} - 10^{-7}$ m, materials and phenomena can be described continuously and statistically, and average macroscopic properties (say, temperature and entropy) are relevant. At the *atomic scale*, $\sim 10^{-10}$ m $= 1$ Å, the atoms should be seen as separated. The *electroweak scale*, $\sim 10^{-18}$ m (100–1000 GeV, in terms of energy) will be probed by the LHC (Large Hadron Collider). The *Planck scale* (Quantum Gravity), $\sim 10^{-35}$ m ($\sim 10^{19}$ GeV) is not yet accessible.

Belenchia et al., 2015, gave upper bound $10^{-29} - 10^{-22}$ m of *nonlocality scale* for Quantum Gravity induced nonlocality via opto-mechanical quantum oscillators.

Both, uncertainty principle from Quantum Mechanics and gravitational collapse (black hole formation) from classical General Relativity, indicate some minimum length of order the Planck length l_P where the notion of distance loses operational meaning. Also, *doubly special relativity* adds minimum length and maximum energy scales to observer-independent maximum velocity c.

At short distances, classical geometry is replaced by "quantum geometry" described by 2D conformal field theory (CFD). As two points are getting closer together, the vacuum fluctuations of the gravitational field make the distance between them fluctuate randomly, and its mean value tends to a limit, of the order of l_P. So, no two events in space-time can ever occur closer together.

In String Theory, space-time geometry is not fundamental and, perhaps, it only emerges at larger distance scales. The *Maldacena duality* is the conjectured equivalence between an M-theory defined on a ("large, relativistic") space, and a (quantum, without gravity) CFD defined on its (lower dimensional) conformal boundary.

The Big Bang paradigm supposes a minimal length scale and a smooth distribution (homogeneous and isotropic) at a large scale. For Vilenkin et al., 2011, the main theories admitting "before the Big Bang" (*cyclical universe*, *eternal inflation*, *multiverse*, *cosmic egg*) still require a beginning. For Hartle–Hawkings, 1983, time emerged continuously from space after the Universe was at the age t_P.

- **Glashow's snake**

Uroboros, the snake that bites in its own tail, is an ancient symbol representing the fundamental in different cultures: Universe, eternal life, integration of the opposite, self-creation, etc. **Glashow's snake** is a sketch of the cosmic uroboros by Glashow, 1982, arraying four fundamental forces and the distance scales over which they dominate (62 orders of magnitude from the *Planck scale* $\sim 10^{-35}$ m to the *cosmological scale* $\sim 10^{26}$ m) in clock-like form around the serpent. The dominating forces are:

1. gravity: in the *macrocosmos* from cosmic to planetary distances;
2. electromagnetism: from mountains to atoms (say, within $[10^{-10}, 2 \times 10^5]$ m);
3. weak and strong forces: in the *microcosmos* inside the atom (say, $< 10^{-12}$ m).

No objects are known within $[10^{-14}, 10^{-10}]$ m (the largest nucleus and smallest atom). As distances decrease and energies increase, the last three forces become equivalent around the length 10^{-28} m. Then gravity is included (super-unification happens) linking the largest and smallest: the snake swallows its tail.

Also, a symmetry between small and large distances, called *T-duality*, claims: two superstring theories are *T-dual* if one compactified on a space of large volume is equivalent to the other compactified on a space of small volume.

Cosmic inflation (expansion by a factor of at least 10^{78} in volume, to the size of a grain of sand, from 10^{-36} to $\approx 10^{-32}$ second after the Big Bang) may have created the large scale of the Universe out of quantum-scale fluctuations. Strong and weak forces describe both atomic nuclei and energy generation in stars. Cf. **range of fundamental forces** in Chap. 24.

In *Conformal Cyclic Cosmology* (Penrose, 2010), the Universe is a sequence of *aeons* (space-times with **FLRW metrics** g_i), where the future time-like singularity of each aeon is the Big Bang singularity of the next. In an eaon's beginning and end, distance and time do not exist; only *conformal* (preserving angles) geometry holds. Any eaon is attached to the next one by a conformal rescaling $g_{i+1} = \Omega^2 g_i$.

[redacted]

27.2 Orders of Magnitude for Length

In this section we present a selection of such orders of length, expressed in meters.

$1.616252(81) \times 10^{-35}$: Planck length;
10^{-34}: length of a putative *string* in *M-theory* which supposes that all forces and elementary particles arise by vibration of such strings (but there is no even agreement that there are smallest fundamental objects);
1.01×10^{-25}: *Schwarzschild radius* ($\frac{2Gm}{c^2}$: the value below which mass m collapses into a black hole) of an average (68 kg) human;
$10^{-24} = 1$ **yoctometer**: effective cross-section radius of 1 MeV neutrinos is 2×10^{-23};
10^{-22}: a certain quantum roughness starts to show up, while the space appears completely smooth at the scale of 10^{-14};
$10^{-21} = 1$ **zeptometer**: *preons*, hypothetical components of quarks/leptons;
$10^{-18} = 1$ **attometer**: size of up quark and down quarks; sizes of strange, charm and bottom quarks are 4×10^{-19}, 10^{-19} and 3×10^{-20};
$10^{-15} = 1$ **femtometer** (or *fermi*);
1.75×10^{-15} and 1.5×10^{-14}: diameter of the smallest (H, hydrogene) and largest (U, uranium-234) nucleus;
1.68×10^{-15}: diameter of proton, range of the weak nuclear force;
$10^{-12} = 1$ **picometer**: distance between atomic nuclei in a white dwarf star;
10^{-11}: wavelength of the hardest (shortest) X-rays and longest gamma rays;
0.62 Å and 5.2 Å: diameter of the smallest (helium) and largest (caesium) atom;
$10^{-10} = 1$ Å (angstrom): diameter of a typical atom;
0.74 Å and 1000 Å: diameter of the smallest (H_2) and largest (a SiO_2) molecule;
1.54 Å: length of a typical covalent bond (C-C);
3.4 Å: distance between base pairs in a DNA molecule;
$10^{-9} = 1$ **nanometer**: diameter of typical molecule;
10^{-8}: wavelength of softest X-rays and most extreme ultraviolet;
1.1×10^{-8}: diameter of prion (smallest self-replicating biological entity);
in 2012;
9×10^{-8}: human immunodeficiency virus, HIV; in general, capsid sizes of known viruses range from 1.7×10^{-8} (*Porsine circovirus*) to 1.5×10^{-6} (*pithovirus sibericum*);
10^{-7}: size of chromosomes and largest particle fitting through a surgical mask;
2×10^{-7}: limit of resolution of the light microscope;
$3.8–7.6 \times 10^{-7}$: wavelength of visible (to humans) light;
$10^{-6} = 1$ **micrometer** (or *micron*);
$10^{-6} - 10^{-5}$: diameter of a typical bacterium; known (nondormant) bacteria range from $2–3 \times 10^{-7}$ (*Mycoplasma genitalium*) to 7.5×10^{-4} (*Thiomargarita Namibiensis*);
8.5×10^{-6}: size of *Ostreococcus*, the smallest free-living eukaryotic unicellular organism, while the length of a nerve cell of the Colossal Squid can reach 12 m;
10^{-5}: typical size of (a fog, mist, or cloud) water droplet;

10^{-5}, 1.5×10^{-5}, and 2×10^{-5}: widths of cotton, silk, and wool fibers;

2×10^{-4}: approximately, the lower limit for the human eye to discern an object;

5×10^{-4}: diameter of a human ovum, grain of salt;

$10^{-3} = 1$ **millimeter**;

5×10^{-3}: length of average red ant; in general, insects range from 1.39×10^{-4} (*Dicopomorpha echmepterygis*, the smallest animal) to 5.67×10^{-1} (*Phobaeticus chani*); the smallest flying insects are on the order of 10^{-3} in length;

$7.7 \times 10^{-3}, 5 \times 10^{-2}$ and 9.2×10^{-2}: length of the smallest ones: vertebrate (frog *Paedophryne amauensis*), warm-blooded vertebrate (bee hummingbird *Mellisuga helenae*) and primate (lemur *Microcebus berthae*);

8.9×10^{-3}: Schwarzschild radius of the Earth;

$10^{-2} = 1$ **centimeter**;

5.8×10^{-2}: length of uncoiled sperm of the fruit fly *Drosophila bifurca* (it is 20 fly's bodylengths and the longest sperm cell of any known organism);

$10^{-1} = 1$ **decimeter**: wavelength of the lowest microwave and highest UHF radio frequency, 3 GHz;

1 **meter**: wavelength of the lowest UHF and highest VHF radio frequency, 300 MHz;

1.5: average ground level of the Maldives above sea level;

2.77–3.44: wavelength of the broadcast radio FM band, 108–87 MHz;

5.5 and ≈ 3: height of the tallest animal (giraffe) and extinct primate *Gigantopithecus*;

$10 = 1$ **decameter**: wavelength of the lowest VHF and highest shortwave radio frequency, 30 MHz;

20, 33, 37 and 55: lengths of the longest animals (tapeworm *Diphyllobothrium Klebanovski*, blue whale, lion's mane jellyfish and bootlace worm *Lineus longissimus*);

99.6: height of the world's tallest flowering plant, a tasmanian *Eucalyptus* Centurion (after 100 m, the distribution of the products of photosynthesis become impossible);

$100 = 1$ **hectometer**: wavelength of the lowest HF (high radio frequency) and highest MF (medium radio frequency), 3 MHz;

115.5: height of the world's tallest living tree, a californian *sequoia* Hyperion;

139, 324, 541, 830, 1007 and 8.5: heights of the Great Pyramid of Giza, Eiffel Tower in Paris, One World Trade Center in New York, Burj Khalifa skyscraper in Dubai, Kingdom Tower, planned in Jeddah for 2019, and 11,000 years-old Tower of Jericho;

187–555: wavelength of the broadcast radio AM band, 1600–540 kHz;

340: distance which sound travels in air in 1 s;

$10^3 = 1$ km;

2.954×10^3: Schwarzschild radius of the Sun;

$10^4 = 1$ *miriameter* (used formerly): *scandinavian mile* (Norwegian/Swedish *mil*);

[redacted]

8848 and 10,911: the highest (Mount Everest) and deepest (Mariana Trench) points on the Earth's surface;
$5 \times 10^4 = 50$ km: the maximal distance at which the light of a match can be seen (at least 10 photons arrive on the retina during 0.1 s);
$1.11 \times 10^5 = 111$ km: one degree of latitude on the Earth;
$1.5 \times 10^4 - 1.5 \times 10^7$: wavelengths of sound audible to humans (20 Hz to 20 kHz);
1.37×10^5 and 1.9×10^6: length of the world's longest tunnel, Delaware Aqueduct, New York, and of longest street, Jounge Street, Ontario;
2×10^5: wavelength of a typical tsunami;
$10^6 = 1$ **megameter**, thickness of Earth's atmosphere;
2.22×10^6: diameter of Typhoon Tip (northwest Pacific Ocean, 1979), the most intense tropical cyclone on record;
2.37×10^6: diameter of Pluto at 39.54 AU from the Sun; the smallest dwarf planet is Ceres (the largest asteroid in the Asteroid Belt) of diameter 9.42×10^5 and at 2.77 AU;
3.47×10^6: diameter of the Moon;
9.3×10^6 and 2.1×10^7: length of Trans-Siberian Railway and China's Great Wall;
1.28×10^7 and 4.01×10^7: Earth's equatorial diameter and length of the equator;
4.5×10^7: distance from which Earth's good-looking photograph, *The Blue Marbre*, was taken in 1972 by the Apollo 17 mission. Other famous Earth's images are *Earthrise* (1968, by the Apollo 8), *Pale Blue Dot* (0.12 pixel against the space's vastness; 1990, by Voyager 1) and one from Saturn's neighborhood (2013, by NASA's Cassini).
1.4×10^8: mean diameter of Jupiter;
1.67×10^8: diameter of OGLE-TR-122b, the smallest known star;
$\approx 3 \times 10^8$ (299,792.458 km): distance traveled by light in 1 s;
3.84×10^8: Moon's orbital distance from the Earth;
4.002×10^8: the farthest distance a human has ever been from Earth (Appolo 13 mission, 1970, passed over the far side of the Moon);
$10^9 = 1$ **gigameter**;
1.39×10^9: Sun's diameter and orbital distance of a planet with 3.3-h "year";
6.37×10^9: distance at which Earth's gravity becomes $\frac{1}{1{,}000{,}000}$ of that on its surface;
5.83×10^{10}: orbital distance of Mercury from the Sun;
1.496×10^{11} (1 astronomical unit, AU): mean Earth-Sun distance;
$\approx$ 2.8 AU (near the middle of the Asteroid Belt): Sun's water **frost line** (the distance where it is cold enough, $\approx -123\,^\circ$C, for water to condense into ice), separating terrestrial and jovian planets; it is the radius of the inner Solar System;
5.7×10^{11}: length of the longest observed comet tail (Hyakutake, 1996); the Great Comet of 1997 (Hale–Bopp) has biggest known nucleus (> 60 km);
$10^{12} = 1$ **terameter** (formerly, *spat*);
15.8 AU: diameter of the largest known star, red supergiant UY Scuti;
30.1 AU: radius of the outer Solar System (orbital distance of Neptune);

50 AU: distance from the Sun to the *Kuiper cliff*, the abrupt outer boundary of the *Kuiper Belt* (the region of trans-Neptunian objects around Sun).
936 AU: aphelion of Sedna, the farthest known Solar System object (its orbital period is $\approx$ 11,400 years); it and 2012 VP_{113} (having the farthest known—$80, 5$ AU—perihelion in the Solar System) are only known objects within the *Oort Cloud*;
$10^{15} = 1$ **petameter**;
50,000–200,000 AU: distance from the Sun to the boundaries of the Oort Cloud;
52,000 AU $=$ 0.82 ly: the only known close flyby of Solar System by a star (*Scholz's binary* of dwarfs passed the Oort Cloud about 70,000 years ago);
1.3 parsec $\approx 4 \times 10^{16} \approx$ 4.24 ly: distance to Proxima Centauri, the nearest star;
$\approx 6.15 \times 10^{17}$: radius of humanity's radio bubble, caused by high-power TV broadcasts leaking through the atmosphere into outer space;
$10^{18} = 1$ **exameter**;
$1.57 \times 10^{18} \approx$ 50.9 pc, $\approx$ 250 pc, 46 pc: distances to supernova 1987A, to rapidly rotating neutron star Geminga (remains of a supernova 0.3 Ma ago which created the Local Bubble), to IK Pegasi B (nearest known supernova candidate). A supernova within 10–20 pc would be a catastrophic event for the Earth's life;
$2.59 \times 10^{20} \approx$ 8.4 kpc $\approx$ 27,400 ly: distance from the Sun to the geometric center of our Milky Way galaxy (in Sagittarius A^*, "our" supermassive black hole);
12.9 kpc and 52.8 kpc: distances to the closest (Canis Major Dwarf) and the largest (Large Magellanic Cloud) of 26 satellite galaxies of the Milky Way;
$9.46 \times 10^{20} \approx$ 30.66 kpc $\approx 10^5$ ly: diameter of the Milky Way. The largest known galaxy, C 1101, at the center of the cluster Abell 2029, is $\approx$ 6 Mly across;
$10^{21} = 1$ **zettameter**;
$2.23 \times 10^{22} =$ 725 kpc $=$ 2.54 Mly: distance to Andromeda (M31), the closest (and approaching at 100–140 km/s) large galaxy; also, it is the farthest permanently visible to the naked eye object;
$5.7 \times 10^{23} =$ 59 Mly: distance to Virgo, the nearest (and approaching) major cluster;
$10^{24} = 1$ **yottameter**;
$2 \times 10^{24} =$ 60 Mpc $=$ 110 Mly: diameter of the Local (or Virgo) supercluster;
1.2 Gly: maximal size of structure compatible with the cosmological principle (that distribution of matter is homogeneous and isotropic on a large enough scale);
1.3 Gly, 5.6 Gly and 10 Gly: diameters of the *Canes Venatici Supervoid* (largest known), Giant GRB Ring (largest known regular formation) and Hercules-Corona Borealis Great Wall of GRB (largest known superstructure);
7.5 Gly: the gamma ray burst GRB 080319B, farthest visible, $\approx$ 30 s in 2008, to the naked eye object (GRB 980425, at $\approx$ 40 Mpc, is closest known);
13.14 Gly: the gamma ray burst GRB 090429B (farthest event, ever observed);
13.4 Gly ($z = 11.1$): distance to the farthest known object, galaxy GN-z11;
$1.3 \times 10^{26} =$ 13.82 Gly $=$ 4.24 Gpc: **Hubble radius** of the Universe measured as the **light travel distance** to the source of CMB radiation;

[redacted]

$4.4\times10^{26} = 47$ Gly $= 14.4$ Gpc: *particle horizon* (present radius of the Universe measured as a **comoving distance**); it is larger than the Hubble radius, since the Universe is expanding). It is $\approx 2\,\%$ larger than the radius of the *visible universe* including only signals emitted later than $\approx$ 380,000 years after the Big Bang;

The size of whole Universe can be now much larger than the size of the observable one, even infinite, if its curvature is 0. If the Universe is finite but unbounded or if it is nonsimply connected, then it can be smaller than the observable one.

Projecting into the future: the scale of the Universe will be 10^{31} in 10^{14} years (last red dwarf stars die) and 10^{37} in 10^{20} years (stars have left galaxies). If protons decay, their half-life is $\geq 10^{35}$ years; their estimated number in the Universe is 10^{77};

The Universe, in the current *Heat Death* scenario, achieves beyond 10^{1000} years such a low-energy state that quantum events become major macroscopic phenomena, and space-time loses its meaning again, as below the Planck time/length;

The hypothesis of parallel universes estimates that one can find another identical copy of our Universe within the distance $10^{10^{118}}$ m.

Chapter 28
Distances in Applied Social Sciences

In this chapter we present selected distances used in real-world applications of Human Sciences. In this and the next chapter, the expression of distances ranges from numeric (say, in m) to ordinal (as a degree assigned according to some rule) and nominal.

Depending on the context, the distances are either practical ones, used in daily life and work outside of science, or uncountable ones, used figuratively, say, as metaphors for remoteness (being apart, being unknown, coldness of manner, etc.).

28.1 Distances in Perception and Psychology

- **Distance ceptor**

  A **distance ceptor** is a nerve mechanism of one of the organs of special sense whereby the subject is brought into relation with his distant environment.
- **Oliva et al. perception distance**

  Let $\{s_1, \dots, s_n\}$ be the set of stimuli, and let q_{ij} be the conditional probability that a subject will perceive a stimulus s_j, when the stimulus s_i was shown; so, $q_{ij} \geq 0$, and $\sum_{j=1}^{n} q_{ij} = 1$. Let q_i be the probability of presenting the stimulus s_i.

  The **Oliva et al. perception distance** ([OSLM04]) between stimuli s_i and s_j is

$$\frac{1}{q_i + q_j} \sum_{k=1}^{n} \left| \frac{q_{ik}}{q_i} - \frac{q_{jk}}{q_j} \right| .$$

- **Representational dissimilarity matrix**

  Representational dissimilarity matrix (or RDM) is a square matrix indexed horizontally and vertically by the stimuli (or experimental conditions) and containing a dissimilarity index in each cell, which compares the two brain-activity patterns associated with stimuli labeling the row and column.

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_28

[redacted]

If the same activity pattern estimates are used for the vertical and horizontal dimensions, the RDM is symmetric about a diagonal of zeros. If independent pattern estimates are used for the vertical and horizontal dimensions, the RDM contains entries comparing independent pattern estimates for identical stimuli (reflecting measurement noise) along its diagonal, and two alternative dissimilarity estimates for each stimulus pair in symmetric off-diagonal positions.

- **Visual space**

  Visual space refers to a stable perception of the environment provided by vision, while **haptic space** (or *tactile space*) and **auditory space** refer to such internal representation provided by the senses of pressure perception and audition. The geometry of these spaces and the eventual mappings between them are unknown. But Lewin et al., 2012, found that sensitivity to touch is heritable, and linked to hearing. The main observed kinds of distortion of vision and haptic spaces versus physical space follow; the first three were observed for auditory space also.

  - *Distance-alleys*: lines with corresponding points perceived as equidistant, are, actually, some hyperbolic curves. Usually, the parallel-alleys are lying within the distance-alleys and, for visual space, their difference is small at > 1.5 m.
  - *Oblique effects*: performance of certain tasks is worse when the orientation of stimuli is oblique rather than horizontal or vertical.
  - *Equidistant circles*: the **egocentric distance** is direction-dependent; the points perceived as equidistant from the subject lie on egg-like curves, not on circles.

  These effects and **size-distance invariance hypothesis** should be incorporated in a good model of visual space. In a visual space the distance d and direction are defined from the self, i.e., as the **egocentric distance**. There is evidence that visual space is almost affine and, if it admits a metric d, then d is a **projective metric**, i.e., $d(x, y) + d(y, z) = d(x, z)$ for any perceptually collinear points x, y, z.

  The main models for visual space are a Riemannian space of constant negative curvature (cf. **Riemannian color space** in Chap. 21), a general Riemannian/Finsler space, or an affinely connected (so, not metric, in general) space ([CKK03]).

  An *affine connection* is a linear map sending two vector fields into a third one. The expansion of perceived depth on near and its contraction at far distances hints that the mapping between visual and physical space is not affine.

  Amedi et al., 2002, observed the convergence of visual and tactile shape processing in the human lateral occipital complex. The *vOICe technology* (OIC for "Oh I see!") explores cross-modal binding for inducing visual sensations through sound (mental imagery and artificial synesthesia). Some blind people "see" by echolocation. The cane extends peri-hand space of blind users and, in general, extrapersonal or far space can remap as peripersonal or near space when using tools.

  Weber's illusion: two tactile points are perceived as further apart on body parts with higher tactile sensitivity,

- **Length-related illusions**

  The most common optical illusions distort size or length. For example, in the *Müller–Lyer illusion*, one of two lines of equal length appear shorter because of the way the arrows on their ends are oriented. Pigeons and parrots also are susceptible to it. Segall et al., 1963, found that the mean fractional misperception varies cross-culturally from 1.4 % to 20.3 % with maximum for Europeans. Also, urban residents and younger subjects are much more susceptible to this illusion.

  In the *Luckiech–Sander illusion* (1922), the diagonal bisecting the larger, left-hand parallelogram appears to be longer than the diagonal bisecting the smaller, right-hand parallelogram, but is in fact of the same length.

  The perspective created in *Ponzo illusion* (1911) increases the perceived distance and so, compliant with **Emmert's size-distance law**, perceived size increases.

  The *Moon illusion* (mentioned in clay tablets at Nineveh in the seventh century BC) is that the Moon, despite the constancy of its visual angle ($\approx 0.52°$), at the horizon may appear to be about twice the zenith Moon. This illusion (and similar *Sun illusion*) could be cognitive: the zenith moon is perceived as approaching. (Plug, 1989, claim that the *distance to the sky*, assumed unconsciously, is about 10–40 m cross-culturally and independent of the consciously perceived distance.) The *Ebbenhouse illusion*: the diameter of the circle, surrounded by smaller circles, appears to be larger than one of the same circle nearby, surrounded by larger circles.

  In *vista paradox* (Walker–Rupich–Powell, 1989), a large distant object viewed through a window appears to both shrink in size and recede in distance as the observer approaches; a similar framing effect works in the *coffee cup illusion* (Senders, 1966). In the *Pulfrich depth illusion* (1922), lateral motion of an object is interpreted as having a depth component.

  An *isometric illusion* (or *ambiguous figure*) is a shape that can be built of same-length (i.e., isometric) lines, while relative direction between its components are not clearly indicated. The *Necker Cube* is an example.

  The *Charpentier size-weight illusion* (1891): the larger of two graspable/liftable objects of equal mass is misperceived to be less heavy than the smaller.

- **Size-distance invariance hypothesis**

  The SDIH (**size-distance invariance hypothesis**) by Gilinsky, 1951, is that $\frac{S'}{D'} = C\frac{S}{D}$ holds, where S, D are the physical and S', D' are perceived size and distance of visual stimulus, while C is an observer constant. A simplified formula is $\frac{S'}{D'} = 2\tan\frac{\alpha}{2}$, where α is the angular size of the stimulus.

  A version of SDIH is the **Emmert's size-distance law**: $S' = CD'$. This law accounts for *size constancy*: object's size is perceived to remain constant despite changes in the retinal image (more distant objects appear smaller because of perspective). The Müller–Lyer and Ponzo illusions are examples of size constancy.

  The Moon and Ebbenhouse illusions are called **size-distance paradoxes** since they unbalance SDIH. They are misperceptions of visual angle and examples of

[redacted]

distance constancy: distance is perceived constant despite changes in the retinal image.

If an observer's head translates smoothly through a distance K as he views a stationary target point at pivot distance D_p, then the point will appear to move through a displacement W' when it is perceived to be at a distance D'. The **apparent distance/pivot distance hypothesis** (Gogel, 1982): it holds $\frac{D'}{D_p} + \frac{W'}{K} = 1$.

The **size-distance centration** is the overestimation of the size of objects located near the focus of attention and underestimation of it at the periphery.

Hubbard and Baiard, 1988, gave to subjects name and size S of a familiar object and asked imaged distances d_F, d_O, d_V. Here the object mentally looks to be of the indicated size at the *first-sight distance* d_F. The object become, while mentally walking (zooming), too big to be seen fully with zoom-in at the *overflow distance* d_O, and too small to be identified with zoom-out at the *vanishing point distance* d_V. Consistently with SDIH, d_F was linearly related to S. For d_O and d_V, the relation were the power functions with exponents about 0.9 and 0.7. The time needed to imagine d_O increased slower than linearly with the *scan distance* $d_O - d_F$.

Konkle and Oliva, 2011, found that the real-world objects have a consistent visual size at which they are drawn, imagined, and preferentially viewed. This size is proportional to the logarithm of the object's assumed size, and is characterized by the ratio of the object and the frame of space around it. This size is also related to the first-sight distance d_F and to the typical distance of viewing and interaction. A car at a typical viewing distance of 9.15 m subtends a visual angle of 30°, whereas a raisin held at an arm's length subtends 1°. Cf. the **optimal eye-to-eye distance** and, in Chap. 29, the *TV viewing distance* in the **vision distances**.

Similarly, Palmer et al., 1981, found that in goodness judgments of photographs of objects, the $\frac{3}{4}$ *perspective* (or 2.5 *view*, *pseudo-3D*), in which the front, side, and top surfaces are visually present, were usually ranked highest. Cf. the *axonometric projection* in the **representation of distance in Painting**.

- **Egocentric distance**

The **egocentric distance** is the perceived absolute distance from the self (observer or listener) to an object or a stimulus; cf. **subjective distance**. Usually, such visual distance underestimates the actual physical distance to far objects, and overestimates it for near objects. Such distortion decreases in a lateral direction.

In Visual Perception, the *action space* of a subject is 1–30 m; the smaller and larger spaces are called the *personal space* and *vista space*, respectively.

The **exocentric distance** is the perceived relative distance between objects.

- **Distance cues**

The **distance cues** are cues used to estimate the **egocentric distance**.

For a listener at a fixed location, the main auditory distance cues include: *intensity*, *direct-to-reverberant energy ratio* (in the presence of sound reflecting surfaces), *spectrum* and *binaural differences*; cf. **acoustics distances** in Chap. 21.

For an observer, the main visual distance cues include:

- *relative size*, *relative brightness*, *light and shade*;
- *height in the visual field* (in the case of flat surfaces lying below the level of the eye, the more distant parts appear higher);
- *interposition* (when one object partially occludes another from view);
- *binocular disparities*, *convergence* (depending on the angle of the optical axes of the eyes), *accommodation* (the state of focus of the eyes);
- *aerial perspective* (distant objects become bluer and paler), *distance hazing* (distant objects become decreased in contrast, more fuzzy);
- *motion perspective* (stationary objects appear to a moving observer to glide past).

Examples of the techniques which use the above distance cues to create an optical illusion for the viewer, are:

- *distance fog*: a 3*D* computer graphics technique such that objects farther from the camera are progressively more blurred (obscured by haze). It is used, for example, to disguise the too-short **draw distance**, i.e., the maximal distance in a 3D scene that is still drawn by the rendering engine;
- *forced perspective*: a technique to make objects appear either far away, or nearer depending on their positions relative to the camera and to each other.
- *lead room*: space left in the direction the subject is facing or moving.

- **Subjective distance**

  The **subjective distance** (or *cognitive distance*) is a mental representation of actual distance molded by an individual's social, cultural and general life experiences; cf. **egocentric distance**. Cognitive distance errors occur either because information about two points is not coded/stored in the same branch of memory, or because of errors in retrieval of this information.

  For example, the length of a route with many turns and landmarks is usually overestimated. In general, the filled or divided space (distance or area) appears greater than the empty or undivided one. Also, affective signals of threat and disgust increase and decrease, respectively, perceived proximity.

  Human mental maps, used to find out distance and direction, rely mainly, instead of geometric realities, on real landscape understanding, via webs of landmarks.

  Ellard, 2009, suggests that this loss of natural navigation skills, coupled with the unique ability to imagine themselves in another location, may have given modern humans the freedom to create a reality of their own.
- **Geographic distance biases**

  Sources of distance knowledge are either symbolic (maps, road signs, verbal directions) or directly perceived ones during locomotion: environmental features (visually-perceived turns, landmarks, intersections, etc.), travel time/effort.

  They relate mainly to the perception and cognition of **environmental distances**, i.e., those that cannot be perceived in entirety from a single point of view but can still be apprehended through direct travel experience.

[redacted]

Examples of **geographic distance biases** (subjective distance judgments) are:

- observers are quicker to respond to locations preceded by locations that were either close in distance or were in the same region;
- distances are overestimated when they are near to a reference point; for example, intercity distances from coastal cities are exaggerated;
- subjective distances are often asymmetrical as the perspective varies with the reference object: a small village is considered to be close to a big city while the big city is likely to be seen as far away from it;
- traveled routes segmented by features are subjectively longer than unsegmented routes; moreover, longer segments are relatively underestimated;
- increasing the number of pathway features encountered and recalled by subjects leads to increased distance estimates;
- structural features (such as turns and opaque barriers) breaking a pathway into separate vistas strongly increase subjective distance (suggesting that distance knowledge may result from a process of summing vista distances) (turns are often memorized as straight lines or right angles);
- *Chicago-Rome illusion*: belief that some European cities are located far to the south of their actual location; in fact, Chicago and Rome are at the same latitude (42°), as are Philadelphia and Madrid (40°), etc.;
- *Miami-Lima illusion*: belief that US east coast cities are located to the east of the west coast cities of South America; in fact, Miami is 3° west of Lima.

Such illusions could be perceptually based mental representations that have been distorted through normalization and/or conceptual nonspatial plausible reasoning.

Thorndyke and Hayes-Roth, 1982, compared distance judgments of people with navigation- and map-derived spatial knowledge. Navigation-derived route distance estimates were more accurate than Euclidean judgments, and this difference diminished with increased exploration. The reverse was true for map subjects, and no improvement was observed in the map learning.

Turner–Turner, 1997, made a similar experiment in a plane virtual building. Route distances were much underestimated but exploration-derived Euclidean judgments were good; repeated exposure did not help. The authors suggest that exploration of virtual environments is similar to navigation in the real world but with a restricted field of view, as in tunnels, caves or wearing a helmet, watching TV.

Krishna et al., 2008, compared spatial judgments of *self-focused* ("Western") and *relationship-focused* ("Eastern") people. The former ones were more likely to misjudge distance (when multiple features should be considered), to pay attention to only focal aspects of stimuli and ignore the context and background information.

- **Psychogeography**

  Psychogeography is (Debord, 1955) the study of the precise laws and specific effects of the geographical environment, consciously organized or not, on the emotions and behavior of individuals. An example of related notions is a *desire path* (or *social trail*), i.e., a path developed by erosion caused by animal or human footfall, usually the shortest or easiest route between an origin and destination.

Also, the psychoanalytic study of spatial representation within the unconscious construction of the social and physical world is called *Psychogeography*. In general, *Depth Psychology* refer to unconscious-accounting approaches to therapy and research.

- **Psychological Size and Distance Scale**

  The CID (*Comfortable Interpersonal Distance*) scale by Duke and Nowicky, 1972, consists of a center point 0 and eight equal lines emanating from it. Subjects are asked to imagine themselves on the point 0 and to respond to descriptions of imaginary persons by placing a mark at the point on a line at which they would like the imagined person to stop, that is, the point at which they would no longer feel comfortable. CID is then measured in mm from 0.

  The GIPSDS (**Psychological Size and Distance Scale**) by Grashma and Ichiyama, 1986, is a 22-item rating scale assessing interpersonal status and affect. Subjects draw circles, representing the drawer and other significant persons, so that the radii of the circles and the distances between them indicate the thoughts and feelings about their relationship. These distances and radii, measured in mm, represent the **psychological distance** and status, respectively. Cf. related questionnaire on http://www.surveymonkey.com/s.aspx?sm=Nd8c_2fazsxMZfK9ryhvzPlw_3d_3d
- **Visual Analogue Scales**

  In Psychophysics and Medicine, a **Visual Analogue Scale** (or *VAS*) is a self-report device used to measure the magnitude of internal states such as pain and mood (depression, anxiety, sadness, anger, fatigue, etc.) which range across a continuum and cannot be measured directly. Usually, VAS is a horizontal (or vertical, for Chinese subjects) 10 cm line anchored by word descriptors at each end.

  The *VAS score* is the distance, measured in mm, from the left hand (or lower) end of the line to the point marked by the subject. The VAS tries to produce *ratio data*, i.e., ordered data with a constant scale and a natural zero.

  Amongst scales used for pain-rating, the VAS is more sensitive than the simpler verbal scale (six descriptive or activity tolerance levels), the Wong–Baker facial scale (six grimaces) and the numerical scale (levels $0, 1, 2, \dots, 10$). Also, it is simpler and less intrusive than questionnaires for measuring internal states.
- **Psychological distance**

  CLT (*construal level theory*) (Liberman–Trope, 2003, defines **psychological distance** from an event or object as a common meaning of spatial ("where"), temporal ("when"), social ("who") and hypotheticality ("whether") distance from it.

  Expanding spatial, temporal, social and hypotheticality horizons in human evolution, history and child development is enabled by our capacity for *mental construals*, i.e., abstract mental representations. Any event or object can be represented at *lower-level* (concrete, contextualized, secondary) or *higher-level* (abstract, more schematic, primary) construal.

[redacted]

More abstract construals lead to think of more distant (spatially, temporally, socially, hypothetically) objects and vice versa; cf., say, **temporal distance**.

CLT implied that the four dimensions are functionally similar. For example, increase of distance in only one dimension leads to greater moral concern. Zhang and Wang, 2008, observed that stimulating people to consider spatial distance influences their judgments along three other dimensions, but the reverse is not true.

It is consistent with a claim by Boroditsky, 2000, that the human cognitive system is structured around only concepts emerging directly out of experience, and that other concepts are then built in a metaphorical way. Williams and Bargh, 2008, also claim that psychological distance is a derivative of spatial distance. Spatial concepts such as "near/far" are present at 3 to 4 months of age since the relevant information is readily available to the senses, whereas abstract concepts related to internal states are more difficult to understand. Also, spatial relations between oneself, one's caretakers and potential predators have primary adaptive significance.

- **Temporal distance**

  Temporal distance is a type of **psychological distance** altering individuals' construals of future or past events (Trope–Liberman, 2003). Immediate events trigger concrete construals, which are characterized by an emphasis on details.

  People construe events at greater temporal distance in terms of their abstract, central, goal-related features and pro-arguments, while nearer events are treated situation-specifically at a lower level of counter-arguments. Examples are: greater moral concern over a distant future event, more likely forgiveness by a victim of an earlier transgression, more intense affective consumer's reaction when a positive outcome is just missed. Also, temporal distance relates to our conceptions of the self. Actor perspective dominates the simulation of a near-future event; it switches to a third-person depiction when the event is located in the distant future.

  Spronken et al., 2016, found that in mind-wandering, future-oriented thoughts are more positive compared to past-oriented thoughts. Also, thoughts about the distant past and future are more positive than thoughts about the near past and future, and the frequency of positive thoughts increases with temporal distance.

- **Time-distance relation (in Psychology)**

  People often talk about time using spatial linguistic metaphors (a long vacation, a short concert) but much less talk about space in terms of time. This bidirectional but asymmetric relation suggests that spatial representations are primary, and are later co-opted for other uses such as time.

  Casasanto and Boroditsky, 2008, showed that people, in tasks not involving any linguistic stimuli or responses, are unable to ignore irrelevant spatial information when making judgments about duration, but not the converse. So, the metaphorical space-time relationship observed in language also exists in our more basic representations of distance and duration. Mentally representing time as a linear spatial path may enable us to conceptualize abstract (as moving a meeting forward, pushing a deadline back) and impossible (as time-travel) temporal events.

In Psychology, the *Kappa effect* is that among two journeys of the same duration, one covering more distance appears to take longer, and the *Tau effect* is that among two equidistant journeys, one taking more time to complete appears to have covered more distance. Jones–Huang, 1982, see them as effects of *imputed velocity* (subjects impute uniform motion to discontinuous displays) on judgments of both time and space, rather than direct effect of time (distance) on distance (time) judgment.

Fleet–Hallet–Jepson, 1985, found spatiotemporal inseparability in early visual processing by retinal cells. Maruya–Sato, 2002, reported a new illusion (the time difference of two motion stimuli is converted in the illusory spatial offset) indicating interchangeability of space and time in early visual processing. Simner–Mayo–Spiller, 2009, tested ten individuals with time-space synaesthesia.

The differences appear at the level of higher processing because of different representations: space is represented in retinotopic maps within the visual system, while time is processed in the cerebellum, basal ganglia and cortical structures. Evidence from lesion and human functional brain imaging/interference studies point towards the posterior parietal cortex as the main site where spatial and temporal information converge and interact with each other. Cf. also **spatial-temporal reasoning**.

In human-computer interaction, *Fitts' law* claims that the average time taken to position a mouse cursor over an on-screen target is $a + b\log_2(1 + \frac{D}{W})$, where D is the distance to the center of the target, W is the width (along the axis of motion) of the target and a, b represent the start/stop time and device's speed.

People in the West construct diagonal mental timelines (running from bottom left to the top right); those with damaged right side of their brain have trouble imagining past, i.e., timeline's left side.

Núñez, 2012, found that our spatial representation of time is not innate but learned. The Aymara of the Andes place the past in front and the future behind. The Pormpuraaw of Australia place the past in the east and the future in the west. Some Mandarin speaker have the past above and future below.

For the Yupno of Papua New Guinea, past and future are arranged as a nonlinear 3D bent shape: the past downhill and the future uphill of the local river. Inside of their homes, Yupno point towards the door when talking about the past, and away from the door to indicate future. Yupno also have a native counting system and number concepts but they ignore the number-line concept. They place numbers on the line but only in a categorical manner, ignoring line's extension.

- **Symbolic distance effect**

  In Psychology, (symbolic) **distance effect** is the phenomenon that the brain compares two abstract concepts, represented as a continuous space, with higher accuracy and faster reaction time if they differ more on the relevant dimension. For example, the performance of subjects when comparing a pair of positive numbers (x, y) decreases for smaller $|x - y|$ (*behavioral numerical distance effect*).

  The related *magnitude effect* is that performance decreases for larger $\min\{x, y\}$. For example, it is more difficult to measure a longer distance (say,

[redacted]

100 m) to the nearest mm than a short distance (say, 1 cm). Those effects correspond to *Weber law. They* are valid also for congenitally blind people; they learn spatial relations via tactile input (interpreting, say, numerical distance by placing pegs in a peg board).

A current explanation is that there exists a mental line of numbers (MNL) which is oriented from left to right (as $2, 3, 4$) and nonlinear (more mental space for smaller numbers). So, close numbers are easier to confuse since they are represented on the MNL at adjacent and not always precise locations. Possible MNL's, explaining such confusion, are *linear-scalar* (the psychological distance $d(a, a+1)$ between adjacent values is constant but the amount of noise increases as ka) or *logarithmic* (amount of noise is constant but $d(a, a+1)$ decreases logarithmically).

Related *SNARC* (spatial-numerical association of response codes) *effect* is that smaller (or larger) numbers are responded to more easily with responses toward a left (or, respectively, right) location. Also, smaller numbers promote a left-oriented gaze-direction whereas the opposite is true for higher numbers. Similar spatial-musical association *SMARC* and a mental line of pitches were observed. The same (from left to right) mental number line ability was found in three-day-old chicks.

In general, human cognition is often body-based, i.e., conceptual thinking relies on simulations of perceptual symbols. There is now increasing evidence that abstract concepts are also grounded in sensory-motor representations through metaphors. For example, power/status mapped to verticality, physical force to size, social proximity to temperature, and social distance to spatial distance.

- **Law of proximity**

  Gestalt Psychology is a theory of mind and brain of the Berlin School, in which the brain is holistic, parallel and self-organizing. Perceptual organization is composed of grouping and segregation. The visual grouping of discrete elements is determined by proximity, similarity, common fate, good continuation, closure (Wertheimer, 1923), and, more recently, common region, connectedness or synchrony.

  In particular, the **law of proximity** is that spatial or temporal proximity of elements may induce the mind to perceive a collective or totality.

- **Emotional distance**

  The **emotional distance** is the degree of emotional detachment (toward a person, group or events), aloofness, indifference by personal withdrawal, reserve.

  The Bogardus Social Distance Scale (cf. **social distance**) measures the distance between two groups by averaged emotional distance of their members.

  Spatial empathy is the awareness that an individual has to the proximity, activities, and comfort of people surrounding him.

  The *propinquity effect* is the tendency for people to get emotionally involved with those who have higher *propinquity* (physical/psychological proximity) with them, i.e., whom they encounter often. Walmsley, 1978, proposed that emotional involvement decreases as $d^{-\frac{1}{2}}$ with increasing **subjective distance** d.

- **Psychical distance**

  Psychical (or *psychic*) **distance** is a term having no commonly accepted definition. In several dictionaries, it is a synonym for the **emotional distance**. This term was introduced in [Bull12] to define what was called later the **aesthetic distance** (cf. the **antinomy of distance**) as a degree of the emotional involvement that a person, interacting with an aesthetic artifact or event, feels towards it.

  In Marketing, the psychic distance mean the level of attraction or detachment to a particular country resulting from the degree of uncertainty felt about it.

- **Distancing**

  Distancing (from the verb *to distance*, i.e., to move away from or to leave behind) is any behavior or attitude causing to be or appearing to be at a distance.

  Uncountable noon **distantness** (or **farness**) is the state or quality of being distant, remote, far-off, way in the distance. Archaic meaning: distant parts or regions.

  Distancy, **farawayness**, **distaunce** are rare/obsolete synonyms for distance, while **indistancy** is either nearness, or lack (or want) of distance (or separation).

  Self-distance is the ability to critically reflect on yourself and your relations from an external perspective; not to confound with mathematical notions of **self-distance** in Chaps. 1 and 17.

  Outdistancing means to outrun, especially in a long-distance race, or, in general, to surpass by a wide margin, especially through superior skill or endurance.

  In Martial Arts, **distancing** is the selection of an appropriate *combat range*, i.e., distance from the adversary. For other examples of spatial distancing; cf. **distances between people** and, in Chap. 29, **safe distancing** from a risk factor.

  Social distancing during pandemic refers to focused measures to increase the physical distance between individuals, or activity restrictions, such as increasing distance between student desks, canceling sports activities, and closing schools.

  In *Mediation* (a form of alternative dispute resolution), **distancing** is the impartial and nonemotive attitude of the mediator versus the disputants and outcome.

  In Psychoanalysis, **distancing** is the tendency to put persons and events at a distance. It concerns both the patient and the psychoanalyst.

  In Developmental Psychology, **distancing** (Werner–Kaplan, 1964, for deaf-blind patients) is the process of establishing the individuality of a subject as an essential phase (prior to symbolic cognition and linguistic communication) in learning to treat symbols and referential language. For Sigel (1970, for preschool children), **distancing** is the process of the development of cognitive representation: cognitive demands by the teacher or the parent help to generate a child's representational competence. **Distancing from role identities** is the first step of 7-th (individualistic) of nine stages of ego development in Loevinger, 1976.

  In the books by Kantor, **distancing** refers to APD (Avoidant Personality Disorder): fear of intimacy and commitment in confirmed bachelors, "femmes

[redacted]

fatales", etc. **Associational distancing** refers to individual's dissociation with those in the group inconsistent with his desired social identity.

The **distancing language** is phrasing used by a person to avoid thinking about the subject or content of his own statement (for example, referring to death).

Distancing by scare quotes is placing quotation marks around an item (single word or phrase) to indicate that the item does not signify its literal or conventional meaning. The purpose could be to distance the writer from the quoted content, to alert the reader that the item is used in an unusual way, or to represent the writer's concise paraphrasing. **Neutral distancing** convey a neutral writer's attitude, while distancing him from an item's terminology, in order to call attention to a neologism, jargon, a slang usage, etc; sometimes italics are used for it.

Cf. **technology-related distancing**, **antinomy of distance**, **distanciation**.

28.2 Distances in Economics and Human Geography

- **Technology distances**

The **technological distance** between two firms is a distance (usually, χ^2- or **cosine distance**) between their *patent portfolios*, i.e., vectors of the number of patents granted in (usually, 36) technological subcategories. Other measures are based on the number of patent citations, co-authorship networks, etc.

Granstrand's **cognitive distance** between two firms is the **Steinhaus distance** $\frac{\mu(A \Delta B)}{\mu(A \cup B)} = 1 - \frac{\mu(A \cap B)}{\mu(A \cup B)}$ between their technological profiles (sets of ideas) A and B seen as subsets of a *measure space* $(\Omega, \mathcal{A}, \mu)$.

Olsson, 2000, defined the metric space (I, d) of all ideas (as in human thinking), $I \subset \mathbb{R}^n_+$, with some **intellectual distance** d. The closed, bounded, connected *knowledge set* $A_t \subset I$ extends with time t. New elements are, normally, convex combinations of previous ones: *innovations* within gradual technological progress. Exceptionally, breakthroughs (Kuhn's paradigm shifts) occur.

The similar notion of *thought space* (of ideas/knowledge and relationships among them in thinking) was used by Sumi et al., 1997, for computer-aided thinking with text; they proposed a system of mapping text-objects into metric spaces.

Introduced by Patel, 1965, the **economic distance** between two countries is the time (in years) for a lagging country to catch up to the same per capita income level as the present one of an advanced country. Nazarczuk, 2015, adapted this notion as the *regional economic distance* and defined the *relative regional distance* as the mean number of years necessary to achieve the reference area GDP per capita, taking into consideration the growth rate of the reference area.

Introduced by Fukuchi–Satoh, 1999, the **technology distance** between countries is the time (in years) when a lagging country realizes a similar technological structure as the advanced one has now. The basic assumption of the *Convergence Hypothesis* is that this distance between two countries is smaller than the economic one.

- **Production Economics distances**

  In quantitative Economics, a *technology* is modeled as a set of pairs (x, y), where $x \in \mathbb{R}^m_+$ is an *input* vector, $y \in \mathbb{R}^m_+$ is an *output* vector, and x can produce y. Such a set T should satisfy standard economical regularity conditions.

  The **directional distance function** of input/output x, y toward a (projected and evaluated) direction $(-d_x, d_y) \in \mathbb{R}^m_- \times \mathbb{R}^m_+$ is (Chambers–Chung–Färe, 1996)

  $$\sup\{k \geq 0 : ((x - kd_x), (y + kd_y)) \in T\}.$$

  For $d_x = x, d_y = y$, it is a scaled version of the **Shephard input distance function** (Shephard, 1953 and 1970) $\sup\{k \geq 0 : (x, \frac{y}{k}) \in T\}$.

  The *frontier* $f_s(x)$ is the maximum feasible output of a given input x in a given system (or year) s. The **distance to frontier** (Färe–Crosskopf–Lovell, 1994) of a production point (x, y), where $y = g_s(x)$), is $\frac{g_s(x)}{f_s(x)}$.

  The *Malmquist index* measuring the change in TFP (total factor productivity) between periods s, s' (or comparing to another unit in the same period) is $\frac{g'_s(x)}{f_s(x)}$. The *distance to frontier* is the inverse of TFP in a given industry (or of GDP per worker in a given country) relative to the existing maximum (the frontier, usually, US). In general, the term *distance-to-target* is used for the deviation in percentage of the actual value from the planned one.

  Consider a *production set* $T \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ (input, output). The measure of the technical efficiency, given by Briec–Lemaire, 1999, is the point-set distance $\inf_{y \in we(T)} ||x - y||$ (in a given norm $||.||$ on $\mathbb{R}^{n_1+n_2}$) from $x \in T$ to the *weakly efficient set* $we(T)$. It is the set of minimal elements of the poset $(T, \preceq)$ where the *partial order* $\preceq$ ($t_1 \preceq t_2$ if and only if $t_2 - t_1 \in K$) is induced by the cone $K = int(\mathbb{R}^{n_1}_{>0} \times \mathbb{R}^{n_2}_{>0}) + \{0\}$.
- **Distance to default**

  A *call option* is a financial contract in which the buyer gets, for a fee, the right to buy an agreed quantity of some commodity or financial instrument from the seller at a certain time (the expiration date) for a certain price (the *strike price*).

  Let us see a firm's equity E as a call option on the firm's assets A, with the total *liabilities* (debt) L being the strike price, i.e., $E = \max(0, A - L)$ with $A < L$ meaning the firm's default. Applying Black–Sholes, 1973, and Merton, 1974, option pricing formulas, the **distance to default** t periods ahead is defined by

  $$D2D_t = \frac{\ln \frac{A_t}{D} + t(\mu_A - \frac{1}{2}\sigma_A^2)}{\sigma_A \sqrt{t}},$$

  where μ_A is the rate of growth of A and σ_A is its *volatility* (standard deviation of yearly logarithmic returns). A Morningstar's credit score is $cs = \frac{7}{2}(D2D + SS) + 8BR + CC \times \max(D2D, SS, BR)$, where SS, BR and CC are the solvency, business risk and cash flow cushion scores. The resulting credit rating AAA, AA, A, BBB etc., corresponds to cs within $[16, 23), [23, 61), [61, 96)$, etc.

[redacted]

- **Action distance**

  The **action distance** is the distance between the set of information generated by the Active Business Intelligence system and the set of actions appropriate to a specific business situation. Action distance is the measure of the effort required to understand information and to effect action based on that information. It could be the physical distance between information displayed and action controlled.
- **Effective trade distance**

  There is large border effect of political boundaries on the volume of trade and on relative prices. The border introduces costs related to tariffs, market regulations, differences in product packages and languages.

  Engel–Rogers, 1996, showed that the dispersion of prices within a country is orders of magnitude smaller than across countries, and estimated that the US-Canadian border was equivalent to a distance of 120,000 km. McCallum, 1995, found that inter-provincial trade within Canada was, on average, 22 times larger that the trade of any province with any State from US. Cf. **impact of distance on trade**.

  Borraz et al., 2012, showed that the "online border" in E-commerce is equivalent to the average distance from the online warehouse to the offline stores.

  [HeMa02] defined the **effective trade distance** between countries x and y with populations $x_1, \dots, x_m$ and $y_1, \dots, y_n$ of their main agglomerations as

$$\Big(\sum_{1 \le i \le m} \frac{x_i}{\sum_{1 \le t \le m} x_t} \sum_{1 \le j \le n} \frac{y_j}{\sum_{1 \le t \le n} y_t} d_{ij}^r\Big)^{\frac{1}{r}},$$

  where d_{ij} is the bilateral distance (in km) of the corresponding agglomerations x_i, y_i, and r measures the sensitivity of trade flows to d_{ij}.

  As an **internal distance of a country**, measuring the average distance between producers and consumers, [HeMa02] proposed $0.67\sqrt{\frac{area}{\pi}}$.
- **Impact of distance on trade**

  Bilateral trade decreases with distance; this effect slightly increased over the last century. Webb, 2007, claims that an average distance of trade in 1962 of 4790 km changed only to 4938 km in 2000.

  The relationship between shipments and distance, found in Hillberry–Hummels, 2008, is highly nonlinear: at the beginning, there is a sharp reduction in value with distance; but, once a distance-threshold is achieved the negative effect vanishes.

  An example of used measures is the average distance traveled by heavy trucks between actual origins and destinations in their deliveries of commodities.

  Frankel–Rose, 2000, estimated impact of certain distance variables on trade, for example, +340 %, +200 %, +80 %, +0.8 %,-0.2 %,-1.1 % for common currency, common language, common border, economic size (1 % GDP increase), physical size (1 % increase), physical distance (1 % increase), respectively.

  Using the **gravity models** with 16 *CAGE* (cultural, administrative, geographic, economic) distances between countries, Ghemawat, 2004, developed *CAGE Distance Framework* for managers considering international strategies.

His distances are *cultural* (different languages, ethnicities, religions, social norms), *administrative* (absence of shared monetary or political association, institutional weakness), *geographic* (physical remoteness, different climates, lack of common border or waterway access, weak transportation or communication links) and *economic* (difference in consumer incomes, cost and quality of natural, financial, human resources).

Most affected industries are: meat, cereals, tobacco (by linguistic ties), gold, electricity, textile (by preferential trading agreements), electricity, gas, live animals (by physical remoteness). The wealth difference decreases trade in metals, fertilisers, meat, but increases trade in coffee, tea, animal oils, office machines.

- **Long-distance trade routes**

  Examples of such early historic routes are the *Amber Road* (from northern Africa to the Baltic Sea), *Via Maris* (from Egypt to modern day Iran, Iraq, Turkey, Syria), the route *from the Varangians to the Greeks* (from Scandinavia across Kievan Rus' to the Byzantine Empire), the *Incense Road* (from Mediterranean ports across the Levant and Egypt through Arabia to India), Roman-Indian routes, Trans-Saharan trade, *Grand Trunk Road* (from Calcutta to Peshawar) and the *Ancient Tea Route* (from Yunnan to India via Burma, to Tibet and to central China).

  The *Silk Road* was, from the second century BC, a network of trade routes connecting East, South, and Western Asia with the Mediterranean world, North/Northeast Africa and Europe. Extending 6500 km, it enabled traders to transport goods, slaves and luxuries such as silk, other fine fabrics, perfumes, spices, medicines, jewels, as well as the spreading of knowledge, ideas, cultures, plants, animals and diseases. But the Silk Road became unsafe and collapsed in the tenth century after the fall of the Tang Dynasty of China, the destruction of the Khazar Khaganate and, later, the Turkic invasions of Persia and the Middle East.

  During fifth to tenth centuries, the *Radhanites* (medieval Jewish merchants) dominated trade between the Christian and Islamic worlds, covering much of Europe, North Africa, Middle East, Central Asia and parts of India and China. They carried commodities combining small bulk and high demand (spices, perfumes, jewelry, silk). The *Maritime Republics* (mercantile Italian city-states, especially Genoa, Venice, Pisa, Amalfi) dominated long-distance trade during tenth to thirteenth centuries. The spice trade from Asia to Europe became, via new sea routes, a Portuguese monopoly (fifteenth to seventeenth centuries) replaced by the Dutch, and soon after the English and the French. During thirteenth to seventeenth centuries, the *Hanseatic League* v(an alliance of trading cities and their guilds) dominated trade along the coast of Northern Europe.

- **Relational proximity**

  Economic Geography considers to nongeographical types of proximity (organizational, institutional, cognitive, etc.). In particular, **relational proximity** (or trust-based interaction between actors) is an inclusive concept of the benefits

[redacted]

derived from spatially localized sets of economic activities. It generates relational capital through the dynamic exchange of locally produced knowledge.

The five dimensions of relational proximity are proximity: of contact (directness), through time (continuity, stability), in diversity (multiplicity, scope), in mutual respect and involvement (parity), of purpose (commonality).

Individuals are close to each other in a relational sense when they share the same interaction structure, make transactions or realize exchanges. They are *cognitively close* if they share the same conventions and have common values and representations (including knowledge and technological capabilities).

Bouba-Olga and Grossetti, 2007, divide socio-economic proximity into relational one (role of social networks) and *mediation proximity* (role of newspapers, directories, Internet, agencies, etc.). Tranos and Nijkamp, 2013: physical distance and relational proximities have a significant impact on Internet's infrastructure.

- **Migration distance (in Economics)**

  The **migration distance**, in Economic Geography, is the distance between the geographical centers of the municipalities of origin and destination.

  Ravenstein's 2-nd and 3-rd laws of migration (1880) are that the majority of migrants move a short distance, while those move longer distances tend to choose big-city destinations. About 80 % of migrants move within their own country.

  Migration tends to be an act of aspiration; it generally improves migrant's wealth and lifestyle. *Existential migrants* refer to voluntary noneconomic expatriates with "existential wanderlust". Madison, 2006, defines them as seeking greater possibilities for self-actualising, exploring foreign cultures in order to assess own identity, and ultimately grappling with issues of home/belonging in the world generally.

- **Commuting distance**

  The **commuting distance** is the distance (or travel time) separating work and residence when they are located in separated places (say, municipalities).

  The **acceptable commute distance**, in Real Estate, is the distance that can be covered in an acceptable travel time and increases with better connectivity.

- **Consumer access distance**

  Consumer access distance is a distance measure between the consumer's residence and the nearest provider where he can get specific goods or services (say, a store, market or a health service). For example, *food miles* refers to the distance food is transported from the time of its production until it reaches consumers.

  Measures of geographic access and spatial behavior include distance measures (**map's distance**, **road travel distance**, perceived travel time, etc.), **distance decay** (decreased access with increasing distance) effects, transportation availability and *activity space* (the area of $\approx \frac{2}{3}$ of the consumer's routine activities).

  For example, by US Medicare standards, consumers in urban, suburban, rural areas should have a pharmacy within $2, 5, 15$ miles, respectively. The patients residing outside of a 15-miles radius of their hospital are called *distant patients*. Food grown within 100 miles of its point of purchase or consumption is *local food*.

Similar studies for retailers revealed that the negative effect of distance on store choice behavior was (for all categories of retailers) much larger when this behavior was measured as "frequency" than when it was measured as "budget share".

- **Distance decay (in Spatial Interaction)**

  In general, **distance decay** or the **distance effect** (Chap. 29) is the attenuation of a pattern or process with distance. In Spatial Interaction, **distance decay** is the mathematical representation of the inverse ratio between the quantity of obtained substance and the distance from its source.

  This decay measures the effect of distance on accessibility and number of interactions between locations. For example, it can reflect a reduction in demand due to the increasing travel cost. The quality of streets and shops, height of buildings and price of land decrease as distance from the center of a city increases.

  The *bid-rent distance decay* induces, via the cost of overcoming distance, a class-based spatial arrangement around a city: with increasing distance (and so decreasing rent) commercial, industrial, residential and agricultural areas follow.

  In location planning for a service facility (fire station, retail store, transportation terminal, etc.), the main concerns are *coverage standard* (the maximum distance, or travel time, a user is willing to overcome to utilize it) and distance decay (demand for service decays with distance).

  An example of related *size effect*: doubling the size of a city leads usually to a 15 % decrease of resource use (energy, roadway amount, etc.) per capita, a rise of $\approx$ 15 % in socio-economic well-being (income, wealth, the number of colleges, etc.), but also in crime, disease and average walking speed. Bettencourt et al., 2007, observed that "social currencies" (information, innovation, wealth) typically scale superlinearly with city size, while basic needs (water and household energy consumption) scale linearly and transportation/distribution infrastructures scale sublinearly.

  Distance decay is related to **friction of distance** which posits that in Geography, the *absolute distance* (say, in km) requires some amount of effort, money, time and/or energy to overcome. The corresponding cost is called *relative distance*; it describes the amount of social, cultural, or economic connectivity between two places.

- **Gravity models**

  The general **gravity model** for social interaction is given by the *gravity equation*

$$F_{ij} = a\frac{M_i M_j}{D_{ij}^b},$$

  where F_{ij} is the "flow" (or "gravitational attraction", *interaction*, *mass-distance function*) from location i to location j (alternatively, between those locations), D_{ij} is the "distance" between i and j, M_i and M_j are the relevant economic "masses" of i and j, and a, b are parameters. Cf. Newton's **law of universal gravitation**

[redacted]

in Chap. 24, where $b = 2$. The first instances were formulated by Reilly (1929), Stewart (1948), Isard (1956) and Tinbergen (1962).

If F_{ij} is a monetary flow (say, export values), then M is GDP (gross domestic product), and D_{ij} is the distance (usually the **great circle distance** between the centers of countries i and j). For trade, the true distances are different and selected by economic considerations. But the distance is a proxy for transportation cost, the time elapsed during shipment, cultural distance, and the costs of synchronization, communication, transaction. The **distance effect on trade** is measured by the parameter b; it is 0.94 in Head, 2003, and 0.6 in Leamer–Levinsohn, 1994.

If F_{ij} is a people (travel or migration) or message flow, then M is the population size, and D_{ij} is the travel or communication cost (distance, time, money).

If F_{ij} is the force of attraction from location i to location j (say, for a consumer, or for a criminal), then, usually $b = 2$. Reilly's *law of retail gravitation* is that, given a choice between two cities of sizes M_i, M_j and at distances D_i, D_j, a consumer tends to travel further to reach the larger city with the equilibrium point defined by

$$\frac{M_i}{D_i^2} = \frac{M_j}{D_j^2}.$$

- **Nearness principle**

  The **nearness principle** (or Zipf's *least effort principle*, in Psychology) is the following basic geographical heuristic: given a choice, a person will select the route requiring the least expenditure of effort, i.e., path of least resistance.

  This principle is used, for example, in transportation planning and locating of serial criminals: they tend to commit their crimes fairly close to where they live.

  The **first law of geography** (Tobler, 1970) is: "Everything is related to everything else, but near things are more related than distant things".
- **Distances in Criminology**

  Geographic profiling (or *geoforensic analysis*) aims to identify the spatial behavior (target selection and likely *offender's heaven*, i.e., the residence or workplace) of a serial criminal as it relates to the spatial distribution of linked crime sites.

  The **offender's buffer zone** is an area surrounding the offender's heaven, from which little or no criminal activity will be observed; usually, such a zone occurs for premeditated personal offenses. The primary streets and network arterials that lead into the buffer zone tend to intersect near the estimated offender's heaven. A 1 km buffer zone was found for the UK serial rapists. Most personal offenses occur within about 2 km from the offender's heaven, while property thefts occur further away.

  Given n crime sites (x_i, y_i), $1 \le i \le n$ (where x_i and y_i are the latitude and longitude of the i-th site), the *Newton–Swoope Model* predicts the offender's heaven to be within the circle around the point $(\frac{\sum_i x_i}{n}, \frac{\sum_i y_i}{n})$ with the search radius being

$$\sqrt{\frac{\max|x_{i_1}-x_{i_2}|\cdot\max|y_{i_1}-y_{i_2}|}{\pi(n-1)^2}},$$

where the maxima are over (i_1, i_2), $1 \le i_1 < i_2 \le n$. The *Ganter–Gregory Circle Model* predicts the offender's heaven to be within a circle around the first offense crime site with diameter the maximum distance between crime sites.

The *centrographic models* estimate the offender's heaven as a *center*, i.e., a point from which a given function of travel distances to all crime sites is minimized; the distances are the Euclidean distance, the Manhattan distance, the **wheel distance** (i.e., the actual travel path), perceived travel time, etc. Many of these models are the reverse of Location Theory models aiming to maximize the placement of distribution facilities in order to minimize travel costs. These models (*Voronoi polygons*, etc.) are based on the **nearness principle** (*least effort principle*).

The **journey-to-crime decay function** is a graphical **distance curve** used to represent how the number of offenses committed by an offender decreases as the distance from his/her residence increases. Such functions are variations of the center of gravity functions; cf. **gravity models**.

For detection of criminal, terrorist and other hidden networks, there are many data-mining techniques which extract latent associations (distances and **near-metrics** between people) from graphs of their co-occurrence in relevant documents, events, etc. In, say, drug cartel networks, better to remove *betweeners* (not well-connected bridges between groups, as paid police) instead of hubs (kingpins).

Electronic tagging consist of a device attached to a criminal or vehicle, allowing their whereabouts to be monitored using GPS. An *ankle monitor* (or *tether*) is a such tracking device that individuals under house arrest or parole are often required to wear. The *range of a tether* (or *inclusion zone*, 10–50 m) depends usually on the gravity of the crime; it is set by the offender's probation officer.

- **Drop distance**

  In judicial hanging, the **drop distance** is the distance the executed is allowed to fall. In order to reduce the prisoner's physical suffering (to about a third of a second), this distance is pre-determined, depending on his/her weight, by special *drop tables*. For example, the (US state) Delaware protocol prescribes, in pounds/feet, about 252, 183 and 152 cm for at most 55, 77 and at least 100 kg. Unrelated *hanging distance* is the minimum (horizontal) distance needed for hanging a hammock.

  In Biosystems Engineering, a ventilation jet *drop distance* is defined as the horizontal distance from an air inlet to the point where the jet reaches the occupational zone. In Aviation, an airlift *drop distance* (or *drop height*) is the

[redacted]

vertical distance between the aircraft and the drop zone over which the airdrop is executed.

- **Distance telecommunication**

  Distance telecommunication is the transmission of signals over a distance for the purpose of communication. In modern times, this process almost always involves the use of electromagnetic waves by transmitters and receivers.

  Nonelectronic visual signals were sent by fires, beacons, smoke signals, then by mail, pigeon post, hydraulic semaphores, heliographs and, from the fifteenth century, by maritime flags, semaphore lines and signal lamps.

  Audio signals were sent by drums, horns (cf. **long-distance drumming** in Chap. 21) and, from nineteenth century, by telegraph, telephone, and radio.

  Advanced electrical/electronic signals are sent by television, videophone, fiber optical telecommunications, computer networking, analog cellular mobile phones, SMTP email, Internet and satellite phones.

- **Distance supervision**

  Distance supervision refers to the use of interactive distance technology (landline and cell phones, Email, chat, text messages to cell phone and instant messages, video teleconferencing, Web pages) for live (say, work, training, psychological umbrella, mental health worker, administrative) supervision.

  Such supervision requires tolerance for ambiguity when interacting in an environment that is devoid of nonverbal information.

- **Distance education**

  Distance education is the process of providing instruction when students and instructors are separated by physical distance, and technology is used to bridge the gap. *Distance learning* and *distance* (or *online*) degrees are the desired outcomes. A *semi-distance program* is one combining online and residential courses.

  The **transactional distance** (Moore, 1993) is a perceived degree of separation during interaction between students and teachers, and within each group. This distance decreases with *dialog* (a purposeful positive interaction meant to improve the understanding of the student), with larger autonomy of the learner, and with lesser predetermined structure of the instructional program.

  Vygotsky's *zone of proximal development* is the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers.

- **Distance selling**

  Distance selling, as opposed to face-to-face selling in shops, covers goods or services sold without face-to-face contact between supplier and consumer but through *distance communication means*: press adverts with order forms, catalog sales, telephone, tele-shopping, e-commerce (via Internet), m-commerce (via mobile phone). Examples of the relevant legislation are Consumer Protection (Distance Selling) Directive 97/7/EC and Regulations 2000 in EU.

  The main provisions are: clear prior information before the purchase, its confirmation in a durable medium, delivery within 30 days, "cooling-off" period of seven working days during which the consumer can cancel the contract without any reason and penalty. Exemptions are: *Distance marketing* (financial services sold at distance), business-to-business contracts and some purchases (say, of land, or at an auction, or from vending machines).

- **Approximative human-scale distances**

  An **arm's length** is a distance (about 0.7 m, i.e., within **personal distance**) sufficient to exclude intimacy, i.e., discouraging familiarity or conflict; its analogs are: Italian *braccio*, Turkish *pik*, and Old Russian *sazhen*.

  The **reach distance** is the difference between the maximum reach and arm's length. The **striking** (or *handshaking*) **distance** is a short, easily reachable distance.

  The **whiffing** (or **spitting**, **poking**) **distance** is a very close distance.

  A **stone's throw** is a distance of about 25 fathoms (46 m).

  The **hailing** (or **shouting**, **calling**) **distance** is the distance within which the human voice can be heard. **Far cry**: distance estimated in audibility's terms.

  The **Goldilocks distance** is the "just right" one; cf. the **Kasting distance** in Sect. 25.3.

  The **walking distance** is the context-depending distance normally reachable by walking. In Japan, its standard unit is 80 m, i.e., one minute walking time. Some the UK high schools define 2 and 3 miles as the statutory walking distance for children younger and older than 11 years.

  Pace out means to measure distance by *pacing* (walking with even steps).

- **Walkability distances**

  Walkability is (Abley, 2005) the extent to which the environment is friendly to the presence of people living, shopping, visiting, or spending time in an area.

  Walk Score is a walkability index based on the Euclidean distance to amenities such as businesses, schools, parks, and other common destinations within 20 minutes' walk (1 mile $=$ 5280 feet $\approx$ 1.609 km) of a given starting point. The algorithm awards points based on the distance to the closest amenity in each category.

  The number of points declines from maximum to 0 as the distance runs from 0.4 km (5 minutes' walk) to 1.6 km. Each category is weighted equally and the points are summed and normalized to yield a score from 0 to 100.

  Walk Score also measures pedestrian friendliness by analyzing population density, block length and intersection density. Scores 0 to 49 mean that almost all or most daily errands require a car, and scores 90 to 100 signify a "walker's paradis".

  Walkability is a central principle of *New Urbanism*. In particular, the (urban-to-rural) *transect* is (Duany, 2000) a series of zones that transition from the natural landscape to the dense urban core: T_1 (natural), T_2 (rural), T_3 (suburban), T_4 (general urban), T_5 (urban center), T_6 (downtown).

  Block length and building height for $T_3 - T_6$ are, respectively: 700–800, 600, 500, $\leq$ 400 feet and 1–2, 2–3, 2–4, $\geq$ 4 floors. Each zone is *self-similar* in that it contains a similar transition from the edge to the center of the neighborhood.

  A **walking radius** is a typical urban walking distance from a given starting point to transit or before driving. It was estimated as 400 m for the US, and larger for Europe. For the US the blog *Walk Appeal* defines the following urban

[redacted]

walking radii: W_6 (2 miles on a highway), W_5 (0.75 mile on a main street, in T_5), W_4 (0.25 mile on a neighborhood street, in T_4), W_3 (0.1 mile on a suburban street, in T_3), W_2 (250 feet on a subdivision street), W_1 (100 feet in parking lot), W_0 (unwalkable).

- **Optimal eye-to-eye distance**

  The **optimal eye-to-eye distance** between two persons was measured for some types of interaction. For example, such an optimal viewing distance between a baby and its mother's face, with respect to the immature motor and visual systems of the newborn, is 20–30 cm. During the first weeks of life the accommodation system does not yet function and the lens of the newborn is locked at the **focal distance** of about 19 cm. At ages 8–14 months, babies are judging distances well; they fear a distance with mother (separation anxiety stage) and strangers. Also, left-sided cradling/holding preference have been found in humans and great apes.

- **Language style matching**

  During conversation, texting, emailing, and other forms of interactive communication, people unconsciously mimic their partners' language use patterns.

  The *LSM* (language style matching) *score* of a dyad $(1, 2)$ of persons, with respect to a *function word type* k, $1 \leq k \leq 9$, is $LSM_i = 2\frac{\min(l_{1k}, l_{2k})}{l_{1k}+l_{2k}}$, where l_{ik} $(i = 1, 2)$ is the percentage of person i's words of type k. Each dyad's total *LSM* is the mean of its LSM_i across the nine types of *function words*: auxiliary verbs (say, *am*, *will*, *have*), articles, common adverbs (say, *hardly*, *often*), personal pronouns, indefinite pronouns, prepositions, negations, conjunctions (say, *and*, *but*) and quantifiers.

  LSM is high within the first 15–30 s of any interaction and is generally unconscious. Women use conjunctions at much higher rates.

  LSM predicts successful hostage negotiations (Taylor–Thomas, 2008), task group cohesiveness (Gonzales–Hancock–Pennebaker, 2010), and the formation and persistence of romantic relationships (Ireland et al., 2011).

  However (Manson et al., 2013), the probability of diad's cooperation in a post-conversation one-shot prisoner's dilemma, is positively related, instead of LSM, to the convergence of their speech rates (mean syllable duration).

- **Distances between people**

  In [Hall69], four interpersonal bodily distances were introduced: the *intimate distance* for lovers, childrens, pets (from touching to 46 cm), the *personal-casual distance* for conversations among friends (46–120 cm), the *social-consultative distance* for conversations among acquaintances (1.2–3.7 m), and the *public-audience distance* for public speaking (over 3.7 m). To each of those *proxemics distances*, there corresponds an intimacy/confidence degree and appropriated sound level.

  The distance which is appropriate for a given social situation depends on culture, gender and personal preference. For example, under Islamic law, proximity (being in the same room or secluded place) between a man and a woman is permitted only in the presence of their *mahram* (a spouse or anybody from the same sex or a pre-puberty person from the opposite sex). For an average westerner, personal space is about 70 cm in front, 40 cm behind and 60 cm on either side.

In interaction between strangers, the interpersonal distance between women is smaller than between a woman and a man. The bonding hormone oxytocin discourages partnered (but not a single) men from getting close to a female stranger; they, if were given oxytocin, stayed 10–15 cm farther from the attractive woman.

An example of other cues of nonverbal communication is given by angles of vision which individuals maintain while talking. The **people angular distance in a posture** is the spatial orientation, measured in degrees, of an individual's shoulders relative to those of another; the position of a speaker's upper body in relation to a listener's (for example, facing or angled away). Speaker uses 45° open position in order to make listener feel comfortable and direct body point in order to exert pressure. In general, to approach men directly from the front or women from behind is rude. Also, this distance reveals how one feels about people nearby: the upper body unwittingly angles away from disliked persons and during disagreement.

Eye-contact decreases with spatial proximity. Persons stand closer to those whose eyes are shut. The *Steinzor effect* is the finding that members of leaderless discussion groups seated in circles, are most apt to address remarks to or to get responses from persons seated opposite or nearly opposite them, while in the presence of a strong leader, it happens with persons seated alongside or nearly alongside.

Distancing behavior of people can be measured, for example, by the *stop distance* (when the subject stops an approach since she/he begins to feel uncomfortable), or by the *quotient of approach*, i.e., the percentage of moves made that reduce the interpersonal distance to all moves made.

Humans and monkeys with amygdala lesions have much smaller than average preferred interpersonal distance. *Peripersonal*, i.e., within reach of any limb of an individual, *space* is located dorsally in the parietal lobe whereas *extrapersonal* (outside his reach) *space* is located ventrally in the temporal lobe.

- **Death of Distance**

  Death of Distance is the title of the influential book [Cair01] arguing that the telecommunication revolution (the Internet, mobile telephones, digital TV, etc.) initiated the "death of distance" implying fundamental changes: three-shift work, lower taxes, prominence of English, outsourcing, new ways of government control and citizens communication, but also management-at-a-distance and concentration of elites within the "latte belt". Physical distance (and so, Economic Geography) do not matter; we all live in a "global village". Friedman, 2005, announced: "The world is flat". Gates, 2006, claimed: "With the Internet having connected the world together, someone's opportunity is not determined by geography". The proportion of long-distance relationships in foreign relations increased.

  Similarly (see [Ferg03]), steam-powered ships and the telegraph (as railroads previously and cars later) led, via falling transportation/communication costs, to

[redacted]

the “annihilation of distance” in the nineteenth and twentieth centuries. Heine wrote in 1843: “Space is killed by the railways, and we are left with time alone”.

Further in the past, archaeological evidence points out the appearance, ≈ 0.14 Ma ago, long-distance (up to 300 km) trade, and, $\approx$ 0.04 Ma ago, the innovation of projectile weapons and traps which allowed humans to kill large game (and other humans) from a safe distance.

But already Orwell, 1944, dismissed as “shallowly optimistic” the the phrases “airplane and radio have abolished distance” and “all parts of the world are now interdependent”. Heidegger wrote in 1950: “All distances in time and space are shrinking…The peak of this abolition of every possibility of remoteness is reached by television…” but “The frank abolition of all distances brings no nearness”. Edgerton, 2006, claims that new technologies foster self-sufficiency and isolation instead.

Modern technology eclipsed distance only in that the time to reach a destination has (usually) shrunk. Distances still matter for, say, a company’s strategy on the emerging markets (cf. **impact of distance on trade**) and for political legitimacy. “Tyranny of distance” still affects small island states in the Pacific.

Partridge et al., 2007, report that proximity to higher-tiered urban centers (with their higher-order services, urban amenities, higher-paying jobs, lower-cost products) increasingly favors local job growth. Increased access to services and knowledge exchange requires more face-to-face interaction and so, an increase in the role of distance. Economic and innovation activity are highly localized spatially and tend to agglomerate more. Also, the social influence of individuals, measured by the frequency of memorable interactions, is heavily determined by distance. Goldenberg–Levy, 2009, show that the IT (Information Technology) revolution which occurred in the 1990s, increased local social interactions (as email, Facebook communication, baby name diffusion) to a greater degree than long-distance ones.

In military affairs, Boulding, 1965, and Bandow, 2004, argued that twentieth century technology reduced the value of proximity for the projection of military power because of “a very substantial diminution in the cost of transportation of armed forces” and “an enormous increase in the range of the deadly projectile”. It was used as partial justification for the withdrawal of US forces from overseas bases in 2004. But Webb, 2007, counter-argues that any easing of transport is countered by increased strain put upon its modes since both sides will take advantage of the falling costs to send more supplies. Also, the greatest movement of logistics continues to be conducted by sea, with little improvement in speed since 1900.

- **Technology-related distancing**

  The *Moral Distancing Hypothesis* postulates that technology increases the propensity for unethical conduct by creating a **moral distance** between an act and the moral responsibility for it.

  Print technologies divided people into separate communication systems and distanced them from face-to-face response, sound and touch. TV involved audile-tactile senses and made distance less inhibiting, but it exacerbated *cognitive distancing*: story and image are biased against space/place and time/memory.

This distancing has not diminished with computers but interactivity has increased. In terms of Hunter: technology only re-articulates *communication distance*, because it also must be regarded as the space between understanding and not. The collapsing of spatial barriers diminishes economic but not social and cognitive distance.

The *Psychological Distancing Model* in [Well86] relates the immediacy of communication to the number of information channels: sensory modalities decrease progressively as one moves from face-to-face to telephone, videophone, and e-mail. Skype communication is rated higher than phone since it creates a sense of co-presence. People phone with bad news but text with good news.

Online settings tend to filter out social and relational cues. The lack of instant feedback (since e-mail communication is asynchronous) and low bandwidth limit visual/aural cues. For example, moral and cognitive effects of distancing in online education are not known at present. Also, the shift from face-to-face to online communication can diminish social skills, lead to smaller and more fragmented networks and so, increase feeling of isolation and alienation. But it can be only a bias, based on traditional spatiotemporal assumptions that farness translates into an increase in mediation and results in blurring of the communication.

Virtual distance is the perceived distance between individuals when their primary way of communication is not face-to-face. The main markers of virtual distance are physical, operational and affinity distances.

Mejias, 2005, define *epistemological distance* and *ontological distance* between things as the difference, respectively, in degree of knowledge justification and in ability of subjects to act upon things. He argue that we should strive towards ontological nearness, using modern information and communication technologies to manipulate temporal/spatial and epistemological distances to attain this goal.

Mejias, 2007, see some new advantages in "uniform distancelessness", deplored by Heidegger. *Networked proximity* (nearness mediated through new technology) provides shift from physical proximity to informational availability as the main measure of social relevance. It facilitates new kinds of spatially unbound community, and these emerging forms of sociality could be no less meaningful than the older ones. Networked sociality reconfigures distance rather than eliminates it.

28.3 Distances in Sociology and Language

- **Sociometric distance**

  The **sociometric distance** refers to some measurable degree of mutual or social perception, acceptance, and understanding. Hypothetically, greater sociometric distance is associated with more inaccuracy in evaluating a relationship.

[redacted]

- **Social distance**

  In Sociology, the **social distance** is the extent to which individuals or groups are removed or excluded from participating in one another's lives; a degree of understanding and intimacy which characterize personal and social relations generally. This notion was originated by Simmel in 1903; in his view, the social forms are the stable outcomes of distances interposed between subject and object.

  For example (Mulgan, 1991), the centers of global cities are socially closer to each other than to their own peripheries. In general, the notion of social distance is conceptualized in affective, normative or interactive way, i.e., in terms of sympathy the members of a group feel for another group, norms to define in- and outsider, or the frequency/intensity of interactions between two groups.

  The *Social Distance Scale* by Bogardus, 1947, offers the following response items: would marry, would have as a guest in my household, would have as next door neighbor, would have in neighborhood, would keep in the same town, would keep out of my town, would exile, would kill; cf. **emotional distance**. The responses for each (say, ethnic/racial) group are averaged across all respondents which yields (say, racial) distance quotient ranging from 1.00 to 8.00.

  Dodd and Nehnevasja, 1954, attached distances of 10^t m, $0 \leq t \leq 7$, to eight levels of the Bogardus scale. Many studies on conflicts in ex-Yugoslavia consider *ethnic distance* defined via some modified Bogardus scale, i.e., in terms of acceptance of a particular relation with an abstract person from the other group. Caselli and Coleman, 2012, defined *ethnic distance* as the cost to be born by a member of one group to successfully pass himself as a member of the other group.

  An example of relevant models: Akerlof ([Aker97]) defines an *agent* x as a pair (x_1, x_2) of numbers, where x_1 represents the initial, i.e., inherited, social position, and the position expected to be acquired, x_2. The agent x chooses the value x_2 so as to maximize

  $$f(x_1) + \sum_{y \neq x} \frac{e}{(h + |x_1 - y_1|)(g + |x_2 - y_1|)},$$

  where e, h, g are parameters, $f(x_1)$ represents the intrinsic value of x, and $|x_1 - y_1|$, $|x_2 - y_1|$ are the inherited and acquired *social distances* of x from any agent y (with the social position y_1) of the given society.

  Hoffman, Cabe and Smith, 1996, define *social distance* as the degree of reciprocity that subjects believe exists within a social interaction.

- **Rummel sociocultural distances**

  [Rumm76] defined the main sociocultural distances between two persons as follows.

  - **Personal distance**: one at which people begin to encroach on each other's territory of personal space.
  - **Psychological distance**: perceived difference in motivation, temperaments, abilities, moods, and states (subsuming **intellectual distance**).

- **Interests-distance**: perceived difference in wants, means, and goals (including **ideological distance** on socio-political programs).
- **Affine distance**: degree of sympathy, liking or affection between the two.
- **Social attributes distance**: differences in income, education, race, sex, etc.
- **Status-distance**: differences in wealth, power, and prestige (including **power distance**).
- **Class-distance**: degree to which one person is in general authoritatively superordinate to the other.
- **Cultural distance**: differences in meanings, values and norms reflected in differences in philosophy-religion, science, ethics, law, language, and fine arts.

- **Cultural distance**

  The **cultural distance between countries** $x = (x_1, \dots, x_5)$ and $y = (y_1, \dots, y_5)$ (usually, US) is derived (in [KoSi88]) as the following composite index

$$\sum_{i=1}^{5} \frac{(x_i - y_i)^2}{5V_i},$$

where V_i is the variance of the index i, and the five indices represent ([Hofs80]):

1. Power distance (preferences for equality);
2. Uncertainty avoidance (risk aversion);
3. Individualism versus collectivism;
4. Masculinity versus femininity (gender specialization);
5. Confucian dynamism (long-term versus short-term orientation).

The above **power distance** measures the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally, i.e., how much a culture has respect for authority. For example, Latin Europe and Japan fall in the middle range.

But for Shenkar–Luo–Yeheskel, 2008, above distance is merely a measure of how much a country strayed from the core culture of the multinational enterprise. They propose instead (especially, as a regional construct) the *cultural friction* linking goal incongruity and the nature of cultural interaction.

In order to explain multinational enterprise behavior, Kostova, 1999, introduced the *institutional distance* between its home and host countries as the difference in their regulative, cognitive, and normative institutions. Habib–Zurawicki, 2002, consider effects of the *corruption distance*, i.e., such difference in corruption levels.

Wirsing, 1973, defined *social distance* as a "symbolic gap" between rulers and ruled designed to set apart the political elite from the public. It consists of reinforced and validated ideologies (a formal constitution, a historical myth, etc.).

[redacted]

Davis, 1999, theorized social movements (in Latin America) in terms of their shared distance from the state: geographically, institutionally, socially (class position and income level) and culturally.

The Inglehart–Welzel *cultural map of the world* represent countries as points on $\mathbb{R}^2$, in which the two dimensions (*traditional/secular-rational* and *survival/self-expression*) explain $> 70\,\%$ of cross-national variance in 10 indicators.

- **Trust radius**

  People's **trust radius** is the width of the *cooperation circle* (people among whom a certain trust level exists), while *trust level* is the intensity with which people trust other people. The relation between them is weak. Individualism/collectivism is associated with a broader/narrower trust radius. A given level of trust can be broad/narrow if trust is mostly out-group/in-group connoted.

  Delhey–Newton–Welzel, 2011, distinguish *in-group trust* (trust in your family, your neighborhood, people you know personally) from *out-group trust* (people you meet for the first time, people of another religion, of another nationality). Respondents were asked whether "most people can be trusted" or "you need to be very careful in dealing with people." Then this generalized trust was regressed on out-group and in-group trust via the degree of in-group/out-group connotation of the term *most people*. If this term is more in-group connoted, trust radius is narrow, and it is broad otherwise. Quantification of trust radius occurs by estimating linear regression models, separately for all countries in the sample, employing level of trust in *most people* as the dependent variable and scaled measures of in-group and out-group trust levels as the independent variables.

  The trust radius is calculated as (rescaled to fit a [0, 1] scale) the difference of the estimated coefficients for out-group and in-group trust. Swiss residents got the widest trust radius (followed by Italy, Slovenia, Sweden, Australia), while Thai residents rank lowest (with Morocco, Burkina Faso, and China).

  An unrelated term, the *margin of error*, is usually defined as the *radius* (or half the width) *of a confidence interval* for a particular statistic or observational error. *Confidence level* is usually 90 %, 99 % or, typically, 95 %.

- **Political distance**

  A finite metric space $(X = \{x_1, \dots, x_n\}, d)$ can be seen as a *political space* with the points and distance seen as positions (policy proposals) and some **ideological distance**, respectively. Usually, (X, d) is a subspace of $([0,1]^m, ||x - y||_2)$.

  Let $\{v_1, \dots, v_n\}$ be the vote shares of all candidates $\{c_1, \dots, c_n\}$ of an election or, say, allocated seat shares of all competing parties; so, $\sum_{i=1}^{n} v_i = 1$. The *index of political diversity* (Ortuño-Ortin and Weber, 2008) is $\sum_{1 \le i < j \le n} v_i v_j d(x_i, x_j)$.

  The **mean minimum political distance**, cf. http://wiki.electorama.com/wiki, is (the case $m = 1$ of) $\sum_{i=1}^{n} v_i \min_{j \in E} d(x_i, y)$, where $E = \{1 \le i \le n : c_i \text{ is elected}\}$. Cf. **distance-rationalizable voting rule** in Chap. 11.

- **Surname distance model**

  A **surname distance model** was used in [COR05] in order to estimate the preference transmission from parents to children by comparing, for 47 provinces of mainland Spain, the 47 × 47 distance matrices for **surname distance** with those of **consumption distance** and **cultural distance**.

  The distances were l_1-distances $\sum_i |x_i - y_i|$ between the frequency vectors (x_i), (y_i) of provinces x, y, where z_i is, for the province z, either the frequency of the i-th surname (**surname distance**), or the budget share of the i-th product (**consumption distance**), or the population rate for the i-th cultural issue, say, rate of weddings, newspaper readership, etc. (**cultural distance**), respectively.

  Other (matrices of) distances considered there are:

  - *geographical distance* (in km, between the capitals of two provinces);
  - *income distance* $|m(x) - m(y)|$, where $m(z)$ is mean income in the province z;
  - *climatic distance* $\sum_{1 \le i \le 12} |x_i - y_i|$, where z_i is the average temperature in the province z during the i-th month;
  - *migration distance* $\sum_{1 \le i \le 47} |x_i - y_i|$, where z_i is the percentage of people (living in the province z) born in the province i.

  Strong *vertical preference transmission*, i.e., correlation between surname and consumption distances, was detected only for food items.

- **Distance as a metaphor**

  Lakoff and Núñez, 2000, claim that mathematics emerged via conceptual metaphors grounded in the human body, its motion through space and time, and in human sense perceptions: change is motion, arithmetic is motion along a path, etc.

  For them, the mathematical idea of distance comes from the activity of measuring, and the corresponding technique consists of rational numbers and metric spaces. The idea of proximity/connection comes from connecting and corresponds to topological space. The idea of subtraction mathematizes the ordinary idea of distance.

- **Metaphoric distance**

  A **metaphoric distance** is any notion in which a degree of similarity between two difficult-to-compare things is expressed using spatial notion of distance as an implicit bidirectional and understandable metaphor. Some examples are:

  Internet and Web bring people closer: proximity in subjective space is at-handiness;

  professional distance: teacher-student, therapist-patient, manager-employee;

  financial distance: degree of separation in couple's money/property arrangements;

  competitive distance (incomparability) between two airline product offerings;

  metaphoric distance that a creative thinker takes from the problem, i.e., degree of intuitivity, required to evolve/reshape concepts into new ideas.

  The **distance-similarity metaphor** (Montello et al., 2003) is a design principle, where relatedness in nonspatial data is projected onto distance, so that semantically similar documents are placed closer to one another in an

[redacted]

information space. It is the inverse of the Tobler's **first law of geography**; cf. **nearness principle**. This metaphor is used in Data Mining, Pattern Recognition and **Spatialization**.

Comparing the linguistic metaphor *proximity→similarity* with its mental counterpart, Casasanto (2008), found that stimuli (pairs of words or pictures) presented closer together on the computer screen were rated more similar in conceptual judgments of abstract entities or unseen object properties but, less similar in perceptual judgments of visual appearance of faces and objects.

- **Spatial cognition**

  Spatial cognition concerns the knowledge about spatial properties of objects and events: location, size, distance, direction, separation/connection, shape, pattern, and movement. For instance, it consider navigation (locomotion and way-finding) and orientation during it: recognition of landmarks and *path integration* (an internal measuring/computing process of integrating information about movement).

  Spatial cognition addresses also our (spatial) understanding of the World Wide Web and computer-simulated virtual reality.

  Men surpass women on tests of spatial relations, mental rotation and targeting, while women have better fine motor skills and spatial memory for immobile objects and their location. Such selection should come from a division of labor in Pleistocene groups: hunting of mobile prey for men and gathering of immobile plant foods for women. Women's brains are 10–15 % smaller than men's, but their frontal lobe (decision-making, problem-solving), limbic cortex (emotion regulation) and hippocampus (spatial memory) are proportionally larger, while the parietal cortex (spatial perception) and amygdala (emotional memory) are smaller. Men's brains contain stronger front-to-rear connections (suggesting greater synergy between perception and action) while those of women are better connected from left to right (facilitating emotional processing and the ability to infer others' intentions).

  One of the *cultural universals* (traits common to all human cultures) is that men on average travel greater distances over their lifetime. They are less likely than women to migrate within the country of their birth but more likely to emigrate.

- **Size representation**

  Konkle and Oliva, 2012, found that object representations is differentiated along the ventral temporal cortex by their real-world size. Both big and small objects activated most of temporal cortex but fMRI (cf. **brain distances**) voxels with a big- or small-object preference were consistently found along its medial or, respectively, lateral parts. These parts overlapped with the regions known to be active when identifying spaces to interact with (say, streets, elevators, cars, chairs) or, respectively, processing information on tools, ones we usually pick up.

  Different-sized objects have different action demands and typical interaction distances. Big/small preferences are object-based rather than retinotopic or conceptual. They may derive from systematic biases, say, eccentricity biases and size-dependent biases in the perceptual input and in functional requirements for action. For example, over the viewing experience, in the lifetime or over evolutionary time, the smaller objects tend to be rounder, while larger objects tend to extend more peripherally on the retina. Cf. the **size-distance invariance hypothesis** and, in Chap. 29, **neurons with spatial firing properties**.

- **Spatialization**

  Spatialization (Lefebvre, 1991) refer to the spatial forms that social activities and material things, phenomena or processes take on. It includes cognitive maps, cartography, everyday practice and imagination of possible spatial worlds.

  One of the debated definitions of consciousness: it is a notion of self in space and an ability to make decisions based on previous experience and the current situation. Self-awareness permits to evaluate the distance that separates one from his objectives and to adjust his behavior in order to approach those aims.

  We usually give the upper face or upper torso as *egocenter* (spatial seat of self).

  The term *spatialization* is also used for information display of nonspatial data.

- **Spatial reasoning**

  Spatial reasoning is the domain of spatial knowledge representation: spatial relations between spatial entities and reasoning based on these entities and relations.

  As a modality of human thought, spatial reasoning is a process of forming ideas through the spatial relationships between objects (as in Geometry), while verbal reasoning is the process of forming ideas by assembling symbols into meaningful sequences (as in Language, Algebra, Programming). *Spatial intelligence* is the ability to comprehend 2D and 3D images and shapes.

  Luria, 1973, called the ability to derive the abstract concepts from spatially organized heteromodal information, the *quasi-spatial synthesis*.

  Spatial-temporal reasoning (or *spatial ability*) is the capacity to visualize spatial patterns, to manipulate them mentally over a time-ordered sequence of spatial transformations and to draw conclusions about them from limited information.

  Specifically, *spatial visualization ability* is the ability to manipulate mentally 2D and 3D figures. *Spatial skills* is the ability to locate objects in a 3D-world using sight or touch. *Spatial acuity* is the ability to discriminate two closely-separated points or shapes (say, two similar polygons with different numbers of sides).

  Visual thinking (or *visual/spatial learning*, *picture thinking*) is the common (about 60 % of the general population) phenomenon of thinking through visual processing. Spatial-temporal reasoning is prominent among visual thinkers, as well as among *kinesthetic learners* (who learn through body mapping and physical patterning) and *logical thinkers* (mathematical/systems thinking) who think in patterns and relationships and may work without this being pictorially.

  In Computer Science, spatial-temporal reasoning aims at describing, using abstract relation algebras, the common-sense background knowledge on which human perspective of physical reality is based. It provides rather inexpensive reasoning about entities located in space and time.

[redacted]

- **Spatial language**

  Spatial language consists of natural-language spatial relations used to indicate where things are, and so to identify or refer to them. It usually expresses imprecise and context-dependent information about space.

  Among spatial relations there are *topological* (such as on, to, in, inside, at), *path-related* (such as across, through, along, around), *distance-related* and more complex ones (such as right/left, between, opposite, back of, south of, surround).

  A **distance relation** is a spatial relation which specifies how far the object is away from the reference object: near, far, close, etc.

  The **distance concept of proximity** (Pribbenow, 1992) is the area around the RO (reference object) in which it can be used for localization of the LO (local object), so that there is visual access from RO and noninterruption of the spatial region between objects, while LO is less directly related to a different object. Such proximity can differ with physical distance as, for example, in "The Morning Star is to the left of the church". The area around RO, in which a particular relation is accepted as a valid description of the distance between objects, is called the *acceptance area*.

  Pribbenow, 1991, proposed five distance distinctions: *inclusion* (acceptance area restricted to projection of RO), *contact/adjacency* (immediate neighborhood of RO), *proximity*, *geodistance* (surroundings of RO) and *remoteness* (the complement of the proximal region around RO).

  For Jackendorff–Landau, 1992, there are 3 degrees of distance distinctions in English: interior of RO (*in*, *inside*), exterior but in contact (*on*, *against*), proximate (*near*), plus corresponding negatives, such as *outside*, *off of*, *far from*.

  A spatial reference system is mainly egocentric, or relative (such as *right*, *back*) for the languages spoken in industrialized societies, while the languages spoken in small scale societies rely rather on an allocentric, or absolute set of coordinates.

  Language affects our sense of space and time. Guugu Yimithirr in Australia do not have words for relative space, like left and right, but they unusually well keep track where they are, using their terms for north, south, east and west.

  Semantics of spatial language is considered, say, in Linguistics, Cognitive Psychology, Anatomy, Robotics, Artificial Intelligence, Computer Vision and Geographic Information Systems. An example of far-reaching applications is Grove's *clean space*, a neuro-linguistic psychotherapy based on the spatial metaphors produced by (or extracted from) the client on his present and desired "space" (state).

- **Language distance from English**

  Such measures are based either on a typology (comparing formal similarities between languages), or language trees, or performance (mutual intelligibility and learnability of languages). For example, Rutheford, 1983, defined **distance from English** as the number of differences from English in the following three-way typological classification: subject/verb/object order, topic-prominence/subject-prominence and pragmatic word-order/grammatical word-order. It gives distances $1, 2, 3$ for Spanish, Arabic/Mandarin, Japanese/Korean.

Borland, 1983, compared several languages of immigrants by their acquisition of four areas of English syntax: copula, predicate complementation, negation and articles. The resulting ranking was English, Spanish, Russian, Arabic, Vietnamese.

Elder–Davies, 1998, used ranking based on the following three main types of languages: isolating, analytic or root (as Chinese, Vietnamese), inflecting, synthetic or fusional (as Arabic, Latin, Greek), agglutinating (as Turkish, Japanese). It gave ranks $1, 2, 4, 5$ for Romance, Slavic, Vietnamese/Khmer, Japanese/Korean, respectively, and 3 for Chinese, Arabic, Indonesian, Malay.

The **language distance index** (Chiswick–Miller, 1998) is the inverse of the *language score* of the average speaking proficiency, after 24 weeks of instruction, of English speakers learning this language. This score was measured at regular intervals by increments of 0.25; it ranges from 1 (hardest to learn) to 3 (easiest to learn). The score was, for example, $1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00$ for Japanese, Cantonese, Mandarin, Hindi, Hebrew, Russian, French, Dutch, Afrikaans.

In addition to the above distances, based on syntax, and **linguistic distance**, based on pronunciation, see the lexical semantic distances in Chap. 22.

Cf. **clarity similarity** in Chap. 14, **distances between rhythms** in Chap. 21, **Lasker distance** in Chap. 23 and **surname distance model** in Chap. 28.

Translations of the English noun *distance*, for example, into French, Italian, German, Swedish, Spanish, Interlingua, Esperanto are: distance, distanza, distanz, distans, distancia, distantia, distanco.

The word *distance* has Nr. 625 in the list (Wiktionary:Frequency lists/PG/2006/04) of the common English words in the books found on Project Gutenberg. It has Nrs. 835, 1035, 2404 in contemporary poetry, fiction, TV/movie and overall Nrs. 1513 (written), 1546 (spoken). It comes from Latin *distantia* (distance, farness, difference) and *distans*, present participle of *distare*: *di* (apart) + *stare* (to stand).

The **longest English word** (noncoined and nontechnical) is *antidisestablishmentarianism* (28 letters). Examples of funny distance-related words in *Urban Dictionary* (Web-based dictionary of slang in English) are: *distading* (start and give up on many goals in quick succession), *distarnated* (having no friends and being hated by everyone), *distanitus* (illness one suffer from spotting a person which looks really good from across the room but is a butterface at 5 feet distance), *disstance* (space provided when someone is *dissing*, i.e., show disrespect for, someone else).

- **Editex distance**

The main phonetic encoding algorithms are (based on English language pronunciation) *Soundex*, *Phonix* and *Phonex*, converting words into one-letter three-digit codes. The letter is the first one in the word and the three digits are derived using an assignment of numbers to other word letters. Soundex and Phonex assign:

0 to *a, e, h, i, o, u, w, y*; 1 to *b, p, f, v*; 2 to *c, g, j, k, q, s, x, z*; 3 to *d, t*; 4 to *l*; 5 to *m, n*; 6 to *r*.

[redacted]

Phonix assigns the same numbers, except for 7 (instead of 1) to *f* and *v*, and 8 (instead of 2) to *s, x, z*.

The **Editex distance** (Zobel–Dart, 1996) between two words x and y is a cost-based **editing metric** (i.e., the minimal cost of transforming x into y by substitution, deletion and insertion of letters). For substitutions, the costs are 0 if two letters are the same, 1 if they are in the same letter group, and 2, otherwise.

The *syllabic alignment distance* (Gong–Chan, 2006) between two words x and y is another cost-based **editing metric**. It is based on Phonix, the identification of syllable starting characters and seven edit operations.

- **Phone distances**

  A *phone* is a sound segment having distinct acoustic properties, and is the basic sound unit. A *phoneme* is a minimal distinctive feature/unit in the language (a set of phones which are perceived as equivalent to each other in a given language).

  The number of phonemes (consonants) range, among about 7102 languages spoken now, from 11 (6) in Rotokas to 112 (77) in Taa (languages spoken by about 4000 people in Papua New Guinea and Botswana, respectively).

  The main classes of the **phone distances** (between two phones x and y) are:

  – *Spectrogram-based distances* which are physical-acoustic distortion measures between the sound spectrograms of x and y;
  – *Feature-based phone distances* which are usually the **Manhattan distance** $\sum_i |x_i - y_i|$ between vectors (x_i) and (y_i) representing phones x and y with respect to a given inventory of phonetic features (for example, nasality, stricture, palatalization, rounding, syllabicity).

  The **Laver consonant distance** refers to the improbability of confusing 22 consonants among $\approx$ 50 phonemes of English, developed by Laver, 1994, from subjective auditory impressions. (Chen–Wang–Jia–Dang, 2013, considered similar perception distance between two types of Chinese initials.) The smallest distance, 15 %, is between phonemes [*p*] and [*k*], the largest one, 95 %, is, for example, between [*p*] and [*z*]. Laver also proposed a quasi-distance based on the likelihood that one consonant will be misheard as another by an automatic speech-recognition system.

  Each vowel can be represented by a pair (F_1, F_2) of resonant frequencies of the vocal tract (first and second formants). For example, [*u*], [*a*], [*i*] are represented by $(350, 800)$, $(850, 1150)$, $(350, 1700)$ in *mels* (Chap. 21). The International Phonetic Alphabet identifies seven levels of *height* (F_1) and five levels of *backness* (F_2).

  Chang et al., 2013, produced English language map of the brain; they found the set of neurons in the sensorimotor cortex which controls muscles (in the tongue, lips, jaw, larynx) and fires in unique combination for each sound.

- **Phonetic word distance**

  The **phonetic word** (or *pronunciation*, *Levenstein phonological*) **distance** between two words x and y seen as strings of phones is the **Levenstein metric** with costs (Chap. 11): the minimal cost of transforming x into y by substitution, deletion and insertion of phones. Given a **phone distance** $r(u, v)$ on the International Phonetic Alphabet with the additional phone 0 (silence), the cost of substitution of phone u by v is $r(u, v)$, while $r(u, 0)$ is the cost of insertion or deletion of u.

Levenstein orthographic distance (or *LPD*) is the same measure, but operating on letters instead of phonemes. Words with larger mean LPD to (but smaller mean frequency of) its 20 closest neighbors are easier to recognize.

The average adult has a vocabulary of about 40,000–50,000 words.

- **Linguistic distance**

  The **linguistic distance** between two languages is a term loosely used to describe their difference. The mutual intelligibility of the two languages depends on the degree of their lexical, phonetical, morphological, and syntactical similarity.

  The **lexical similarity** is the percentage of *common* (similar in form and meaning) words in their standardized wordlists. English was evaluated to have a lexical similarity of 60 %, 27 %, 24 % with German, French and Russian, respectively. Cf. **language distance index**, **language distance effect**, **Swadesh similarity**.

  Specifically, the **linguistic** (**dialectal**) **distance** between language varieties X and Y is the mean, for a fixed sample S of notions, **phonetic word distance** between *cognate* (i.e., having the same meaning) words s_X and s_Y, representing the same notion $s \in S$ in X and Y, respectively.

  One example of a *dialect continuum* (as *ring species* in Biology) is Dutch-German: their mutual intelligibility is small but a chain of dialects connects them.

- **Swadesh similarity**

  The *Swadesh word list* of a language (Swadesh, 1940–1950) is a list of vocabulary with (usually, 100) basic words which are acquired by the native speakers in early childhood and supposed to change very slowly over time. The **Swadesh similarity** between two languages is the percentage of *cognate* (having similar meaning and sound) words in their Swadesh lists. *Glottochronology* is a method of assessing the temporal divergence of two languages based on this similarity.

  The first 12 items of the original final Swadesh list: *I, you, we, this, that, who?, what?, all, many, one, two*. Cf. the first 12 most frequently used English words: *the, of, and, a, to, in, is, you, that, it, he, was* in all printed material and *I, the, and, to, a, of, that, in, it, my, is, you* across both spoken and written texts.

  Acerbi et al., 2013: the frequency of emotional words declined in English-language books over twentieth century, but the use of fear-related words increased from 1980. The *half-life* of a word is the number of years after which it has a 50 % probability of having been replaced by a new noncognate word; rougly, it is within 750–20,000 years, say, 9000, 3200, 1900. *stab*, *bird*, *we*.

  Pagel et al., 2013, suggest existence of a *proto-Eurasian mother tongue*. For example, they list 15,000 years old words cognate in at least 4 Eurasiatic language families: thou, I, not, that, we, to give, who, this, what, man/male, ye, old, mother, to hear, hand, fire, to pull, black, to flow, bark, ashes, to spit, worm.

[redacted]

- **Language distance effect**

  In Foreign Language Learning, Corder, 1981, conjectured the existence of the following **language distance effect**: where the mother tongue (L1) is structurally similar to the target language, the learner will pass more rapidly along the developmental continuum (or some parts of it) than where it differs; moreover, all previous learned languages have a facilitating effect.

  Ringbom, 1987, added: the influence of the L1 is stronger at early stages of learning, at lower levels of proficiency and in more communicative tasks. But such correlation could be indirect. For example, the written form of modern Chinese does not vary among the regions of China, but the spoken languages differ sharply, while spoken German and Yiddish are close but have different alphabets.
- **Long-distance dependence (in Language)**

  In Language, **long-distance dependence** (or *syntactic binding*) is a construction, including wh-questions (such as "Who do you think he likes"), topicalizations (such as "Mary, he likes"), *easy*-adjectives (such as "Mary is easy to talk to"), relative clauses (such as "I saw the woman who I think he likes")—which permits an element in one position (*filler*) to fulfill the grammatical role associated with another nonadjacent position (*gap*). The *filler-gap distance*, in terms of the number of intervening clauses or words between them in a sentence, can be arbitrary large. Cf. **long range dependence** in Chap. 18.

  An *anaphora* is a subsequent reference to an entity already introduced in discourse. In order to be interpreted, anaphora must get its content from an antecedent in the sentence which in English is usually syntactically local as in "Mary excused *herself*". A **long-distance anaphora** is an anaphora with antecedent outside of its local domain, as in "The players told us stories about *each other*". Its resolution (finding what it refers to) is a hard problem of machine translation.

  The **anaphoric distance** is (Ariel, 1990) the number of words between an anaphora and its antecedent. The **referential distance** (or **textual distance**) is (Givón, 1983) the amount of clauses between them. In general, each text can be represented as a tree in which *discourse units* (normally, clauses) are vertices and *rhetorical relations* (sequence, joint, cause, elaboration, etc.) are edges.

  The **rhetorical distance** between two discourse units is (Fox, 1987) the minimal number of "sequence"-edges on a path between them.

28.4 Distances in Philosophy, Religion and Art

- **Zeno's distance dichotomy paradox**

  This paradox by the pre-Socratic Greek philosopher Zeno of Elea claims that it is impossible to cover any distance, because half the distance must be traversed first, then half the remaining distance, then again half of what remains, and so on.

The paradoxical conclusion is that travel over any finite distance can neither be completed nor begun, and so all motion must be an illusion.

But, in fact, dividing a finite distance into an infinite series of small distances and then adding the all together gives back the finite distance one started with.

- **Space (in Philosophy)**

The present Newton–Einstein notion of **space** was preceded by Democritus' (c. 460–370 BC) *Void* (the infinite container of objects), Plato's (c. 424–348 BC) *Khora* (an interval between being and nonbeing in which Forms materialize) and Aristotle's (380–322 BC) *Cosmos* (a finite system of relations between objects). Cf. **Minkowski metric** (Chap. 26) for the origin of the space-time concept.

Like the Hindu doctrines of Vedanta, Spinoza (1632–77) saw our Universe as a mode under two (among an infinity of) attributes, *Thought* and *Extension*, of *God* (unique absolutely infinite, eternal, self-caused substance, without personality and consciousness). These parallel (but without causal interaction) attributes define how substance can be understood: to be composed of thoughts and physically extended in space, i.e., to have breadth and depth. So, the Universe is deterministic.

For Newton (1642–1727) space was absolute: it existed permanently and independently of whether there is any matter in it. It is a framework of creation, stage setting within which physical phenomena occur. For Leibniz's (1646–1716) space was a collection of relations between objects, given by their distance and direction from one to another, i.e., an idealized abstraction from the relations between individual entities or their possible locations which must therefore be discrete.

For Kant (1724–1804) space is not substance or relation, but a part of an unavoidable systematic framework used by the humans to organize their experiences. Disagreement continues between philosophers over whether space is an entity, a relationship between entities, or part of a conceptual framework.

In *biocentric cosmology* (Lanza, 2007), build on quantum physics, space is a form of our animal understanding and does not have an observer-independent reality, while time is the process by which we perceive changes in the Universe. Also, space-time could be not fundamental, but emerging from a deeper quantum reality.

Free space refers to a perfect vacuum, devoid of all particles; it is excluded by the uncertainty principle. The quantum vacuum is devoid of atoms but contains subatomic short-lived particles—photons, gravitons, etc.

A *parameter space* is the set of values of parameters in a mathematical model.

In Mathematics and Physics, the *phase space* (Gibbs, 1901) is a space in which all possible states of the system are represented as unique points; cf. Chap. 18.

- **Kristeva nonmetric space**

Kristeva, 1980, considered the basic psychoanalytic distinction (by Freud) between pre-Oedipal and Oedipal aspects of personality development. Narcissistic identification and maternal dependency, anarchic component drives, polymorphic erotogenicism, and primary processes characterize the pre-Oedipal.

[redacted]

Paternal competition and identification, specific drives, phallic erotogenicism, and secondary processes characterize Oedipal aspects.

Kristeva describes the pre-Oedipal feminine phase by an enveloping, amorphous, "nonmetric" space (Plato's *khora*) that both nourishes and threatens; it also defines and limits self-identity. She characterizes the Oedipal male phase by a metric space (Aristotle's *topos*); the self and the self-to-space are more precise and well defined in topos. Kristeva insists also on the fact that the semiotic process is rooted in feminine libidinal, pre-Oedipal energy which needs channeling for social cohesion.

Deleuze–Guattari, 1980, divide *multiplicities* (networks, manifolds, spaces) into *striated* (metric, hierarchical, centered, numerical) and *smooth* ("nonmetric, rhizomic, those that occupy space without counting it and can be explored only by legwork").

The above French post-structuralists use the metaphor of *nonmetric* in line with a systematic (but generating controversy) use of topological terms by the psychoanalyst Lacan. In particular, he sought the space J (of *Jouissance*, i.e., sexual relations) as a metric space and used metaphorically the *Heine–Borel theorem* (that closed and bounded subspaces of Euclidean spaces are their only compact subspaces).

Back to Mathematics, when a notion, theorem or algorithm is extended from metric to general distance space, the latter is called **nonmetric space**.

- **Emerson distance between persons**

  We call the **Emerson distance between persons** the separation between "gods", which was required by an American poet and philosopher Ralph Waldo Emerson (1803–82) in his Essay 16 *Manners*: "Let the incommunicable objects of nature and the metaphysical isolation of man teach us independence... We should meet each morning, as from foreign countries, and spending the day together, should depart at night, as into foreign countries. In all things I would have the island of a man inviolate. Let us sit apart as the gods, talking from peak to peak all round Olympus... Lovers should guard their strangeness... Every natural function can be dignified by deliberation and privacy." At the end of his 1836 book *Nature*, Emerson also wrote: "Every spirit builds itself a house; and beyond its house, a world; and beyond its world, a heaven... Build, therefore, your own world."

  Similar dignified separation is mentioned in quotes from the Russian philosopher Mikhail Bakhtin (1895–1975): "The feeling of respect creates a distance, both in relation to the other person, and in relation to one's own self", and the Bohemian-Austrian poet Rainer Maria Rilke (1875–1926): "Once the realization is accepted that even between the closest human beings infinite distances continue, a wonderful living side by side can grow, if they succeed in loving the distance between them which makes it possible for each to see the other whole against the sky."

- **Nietzsche’s Ariadne distance**

  The German philosopher Friedrich Wilhelm Nietzsche (1844–1900), treated distance in a sensual/erotic way. In “On the Genealogy of Morals” (1887) he wrote:

  “The pathos of nobility and distance… the fundamental total feeling on the part of a higher ruling nature in relation to a lower nature, to a ‘beneath’ – that is the origin of the opposition between ‘good’ and ‘bad.’”

  His Zarathustra favors *fernstenliebe* (love of the farthest) over Christian love of the neighbor. Moreover, *fernstenliebe* is to love neither objects, nor ends—but rather, distance/endlessness itself, which makes all distances recur and perpetuate themselves.

  The courtly troubadours of the twelfth century valued eroticization of the unattainable object, while for German romanticism (for example, Novalis, Schopenhauer, Wagner) there can be no satisfaction in erotic relations, or in life itself, as long as distance remains. In Wagner’s opera, Tristan laments: “Blessed nearness, tedious distance.”

  Kuzma, 2013, claims that Nietzsche, by the early 1880s, “rehabilitated erotic distance”, in response to its denigration and the consummatory idealism and passive nihilism of the German romantic tradition. This rehabilitation of courtly love culminated in the concept of an absolute separation and eternal recurrence.

  According to Kuzma, Ariadne in ‘*Thus Spoke Zarathustra* (1883–1885) is not only the symbol of the human soul and life, but Nietzsche’s privileged name for absolute, infinite spatially and eternal distance itself, for an eternity conceived in the absence of every end, any possible object to attain and every Other to love. To desire Ariadne, is to desire the incessant prolongation of longing in the absence of all fulfilment. Zarathustra does not seek rest, consummation, and release, but affirms a sort of metaphysical *coitus reservatus*, the eternal prolongation of boundless and unresolved desire, implying “voluptuousness of the future” and “love of fate”.

  The eternal recurrence requires spatial or temporal infinity. Nietzsche, in his posthumous notes, posits finite matter and infinite cyclical time.

- **Heidegger’s de-severance distance**

  The German philosopher Martin Heidegger (1889–1976), sought space in terms of limit and event of placing, not merely a location. He wrote: “space is something that has been spaced, or made room for, and that which is let into its bounds”.

  His main notion, *Dasein* (Being there), means Being-in-the-world, as opposed to the Cartesian abstract agent, a subject, or the objective world alone. Dasein is revealed by projection into, and engagement with, a personal world, one’s environment. It is *ontically* (in factual existence) closest to itself, yet ontologically farthest.

  For Heidegger, Dasein dwells spatially in the world, but in the *equipmental space* (functional places, defined by Dasein-centered totalities of involvements) rather than in physical, Cartesian space, and this spatiality is characterized by *de-severance*, where “*de-severing* amounts to making the farness vanish—that

[redacted]

is, making the remoteness of something disappear, bringing it close". Not only reducing physical distance, de-severance is the *reach* of Dasein's skilled practical activity.

An entity is *nearby* if it is readily available for some such activity, and *far away* if it is not. Nearness comes into being when the thing is examined. We reach it through things; it is nearness that makes the thingness of the thing appear. Cf. *Heidegger's Topology* (MIT Press, 2007) by Malpas. The following quotes (of 1924, 1954, 1966, 1971) illustrate **Heidegger's de-severance distance**:

"Man, as existing transcendence abounding in and surpassing toward possibilities, is a creature of distance. Only through the primordial distances he establishes toward all being in his transcendence does a true nearness to things flourish in him."

"Longing is the agony of the nearness of the distant."

"Then thinking would be coming-into-the-nearness-of distance."

"What is this uniformity in which everything is neither far nor near – is, as it were, without distance? Everything gets lumped together into uniform distancelessness."

Cf. the **technology-related distancing** and **death of distance** in Sect. 28.2.

French philosopher/writer Maurice Blanchot (1907–2003) considered Nietzsche, Heidegger and Lévinas via their metaphorics of distance. For example, he wrote:

"A distance is synonymous with extreme non-coincidence."

"Far and near are dimensions of what escapes presence as well as absence under attraction of [impersonal] 'it'. It draws away, draws close, the same ghostly affirmation, the same premises of non-presence."

"To the proximity of the most distant, to the pressure of what is lightest, … to the contact of that which is never arrives, it is by friendship that I can respond, … the response of passivity to the non-presence of unknown [stranger]".

- **Lévinas distance to Other**

  We call the **Lévinas distance to Other** a primary distance between the individuals in their face-to-face encounter, which the French philosopher and Talmudic scholar Emmanuel Lévinas (1906–95) discusses in his book *Totality and Infinity*, 1961.

  Lévinas considers the precognitive relation with the Other: the Other, appearing as the Face, gives itself priority, its first demand even before I react to, love or kill it, is: "thou shalt not kill me". This Face is not an object but pure expression affecting me before I start meditating on it and passively resisting the desire that is my freedom. In this asymmetrical relationship—being silently summoned by the exposed Face of the Other ("widow, orphan, or stranger") and responding by responsibility for the Other without knowing that he will reciprocate—Lévinas (in line of *Misnagdim Judaism* ethics) finds the meaning of being human and concerned about justice. For him, this ethical duty is prior to pursuit of knowledge and ontology of nature.

  According to him, before covering the distance separating the *existent* (the lone subject) from the Other, one must first go from anonymous existence to the existent, from “there is” (swarming of points) to the Being (lucidity of consciousness localized here-below). Lévinas’ ethics spans the distance between the foundational chaos of “there is” and the objective or intersubjective world. Ethics marks the primary situation of the face-to-face, whereas morality comes later as a set of rules emerging in the social situation if there are more than two people face-to-face. And, for Lévinas, the scriptural/traditional God is the Infinite Other.

- **Distant suffering**

  Normally, physical distance is inversely related to charitable inclinations. But the traditional morality of “universal” proximity (geographic, age, familial, etc.) and pity not works in contemporary life. Most important actions happen on distance and the *mediation* (capacity of the media to involve us emotionally and culturally) address our concern for the “other”. Cf. **death/kilometer** in Chap. 25.

  The nonuniversal quality of humanity should be constructed. So, mass media, NGO’s, aid agencies, live blogs, and celebrity advocacy use imagery in order to encourage audiences to acknowledge, care and act for far away vulnerable others.

  For Chouliaraki, 2006, the current mediation replaced earlier claims to our “common humanity” by artful stories that promise to make us better people. As suffering becomes a spectacle of sublime artistic expression, the inactive spectator can merely gaze in disbelief. Arising voyeuristic altruism is motivated by self-empowerment: to realize our own humanity while keeping the humanity of the sufferer outside the remit of our judgement and imagination, i.e., keeping **moral distance**. Chouliaraki calls it *narcissistic self-distance* or *improper distance*.

  Silverstone’s (2002) **proper distance in mediation** refers to the degree of proximity required in our mediated interrelationships if we are to create a sense of the other sufficient not just for reciprocity but for a duty of care, obligation and understanding. It should be neither too close to the particularities or the emotionalities of specific instances of suffering, nor too far to get a sense of common humanity as well as intrinsic difference. Cf. **Lévinas distance** and **antinomy of distance**.

  Silverstone and Chouliaraki call to represent sufferers as active, autonomous and empowered individuals. They advocate *agonistic solidarity*, treating the vulnerable other as other with her/his own humanity. It requires “an intellectual and aesthetic openness towards divergent cultural experiences, a search for contrasts rather than uniformity” (Hannerz, 1990). For Arendt, 1978, the imagination enables us to create the distance which is necessary for an impartial judgment,

  But for Dayan, 2007, a climactic Lévinasian encounter with Other is not dualistic: there are many others awaiting my response at any given moment. So, proper distance should define the point from which one is capable of equitably hearing their respective claims, and it involves the reintroduction of actual distance.

[redacted]

- **Moral distance**

  The **moral distance** is a measure of moral indifference or empathy toward a person, group of people, or events. Abelson, 2005, refers to moral distance as the emotional closeness between agent and beneficiary.

  But Aguiar, Brañas-Garza and Miller, 2008, define it as the degree of moral obligation that the agent has towards the recipient. So, for them the social distance is only a case of moral distance in which anonymity plays a crucial, negative role.

  The **ethical distance** is a distance between an act and its ethical consequences, or between the moral agent and the state of affairs which has occurred.

  The (moral) *distancing* is a separation in time or space that reduces the empathy that a person may have for the suffering of others, i.e., that increases moral distance. In particular, **distantiation** is a tendency to distance oneself (physically or socially, by segregation or congregation) from those that one does not esteem. Cf. **distanciation**. On the other hand, the *good distancing* (Sartre, 1943, and Ricoeur, 1995) means the process of deciding how long a given ethical link should be.
- **Simone Weil distance**

  We call **Simone Weil distance** a kind of moral God-cross radius of the Universe which the French philosopher, Christian mystic, social activist and self-hatred Jewess, Simone Weil (1909–43) introduced in "The Distance", one of the philosophico-theological essays comprising her *Waiting for God* (Putnam, New York, 1951).

  She connects God's love to the distance; so, his absence can be interpreted as a presence: "every separation is a link" (Plato's *metaxu*). She wrote: "God did not create anything except love itself, and the means to love... Because no other could do it, he himself went the greatest possible distance, the infinite distance. This distance between God and God, this supreme tearing apart, this agony beyond all others, this marvel of love, is the crucifixion."

  In her peculiar Christian theodicy, "evil is the form which God's mercy takes in this world", and the crucifixion of Christ (the greatest love/distance) was necessary "in order that we should realize the distance between ourselves and God ... for we do not realize distance except in the downward direction". Weil's *God-cross* (or *creator-creature*) distance recalls the old question: can we equate distance from God with proximity to Evil? Her main drive, purity, consisted of maximizing **moral distance** to Evil, embodied for her by "the social, Rome and Israel".

  Cf. Irenaeus (second century) God-humans **epistemic distance**, which must be far enough that belief in God remains a free choice. In Irenaean teodicy, God created both, evil/suffering and free will, allowing us moral choices and development.

  Cf. Pascal's (1669) *God-man-nothing* distances in *Pensées*, note 72: "...what is man in Nature? A nothing in relation to infinity, all in relation to nothing, a central point between nothing and all and infinitely far from understanding either".

Cf. Montaigne's (1580) *nothing-smallest* and *smallest-largest* distances in *Essais, III:11 On the lame*: "Yet the distance is greater from nothing to the minutest thing in the world than it is from the minutest thing to the biggest."

Cf. Tipler's (2007) *Big Bang–Omega Point* time/distance with Initial and Final singularities seen as God-Father and God-Son. Tipler's *Omega point* (technological singularity) is a variation of prior use of the term (Teilhard de Chardin, 1950) as the supreme point of complexity and consciousness: the Logos, or Christ.

Calvin's (1537) *Eucharistic theology* (doctrine on the meaning of bread and vine that Christ offered to his disciples during the last supper before his arrest) also relies on spatial distance as a metaphor that best conveys the separation of the world from Christ and of the earthly, human from the heavenly, divine.

Weil's approach reminds that of the Lurian (about 1570) kabbalistic notions: *tzimtzum* (God's concealment, withdrawal of a part, creation by self-delimitation) and *shattering of the vessels* (evil as impure vitality of husks, produced whenever the force of separation loses its distancing function and giving man the opportunity to choose between good and evil). The purpose is to bridge the distance between Infiniteness of God (or Good) and the diversity of existence, without falling into the facility of dualism (as manicheanism and gnosticism). It is done by postulating intermediate levels of being (and purity) during emanation (unfolding) within the divine and allowing humans to participate in the redemption of the Creation.

So, a possible individual response to the Creator is purification and *ascent*, i.e., the spiritual movement through the levels of emanation in which the coverings of impurity, that create distance from God, are removed progressively.

Besides, the song "From a Distance", written by Julie Gold in 1985 , is about how God is watching us and how, despite the distance (physical and emotional) distorting perceptions, there is still a little peace and love in this world.

- **Golgotha distance**

The exact locations of the Praetorium, where Pilate judged Jesus, and Golgotha, where he was crucified, as well as of the path that Jesus walked, are not known. At present, the Via Dolorosa (600 m from the Antonia Fortress west to the Church of the Holy Sepulchre) in the Old City of Jerusalem, held to be this path.

The first century Jerusalem was about 500 m east to west and 1200 m north to south. Herod's palace (including Praetorium) was about 600 m from Golgotha and 400 m from the Temple. The **Golgotha distance** (total distance from Gethsemane, where Jesus was arrested, to the Crucifixion) was about 1500 m.

Another New Testament's distance is mentioned in Apocalypse: "And the angel thrust in his sickle into the earth, and gathered the vine of the earth, and cast it into the great winepress of the wrath of God. And the winepress was trodden without the city [Jerusalem], and blood came out of the winepress, even unto the horse bridles, by the space of 1600 furlongs [200 miles]" (Revelation 14:19–20). It can hint to the whole length of the land of Israel, computable as 1600 studia.

[redacted]

- **Distance to Heaven**

  Below are given examples of distances and lengths which old traditions related (sometimes as a metaphor) to such notions as God and Heaven.

  In the Hebrew text *Shi'ur Qomah* (*The Measure of the Body*), the height of the *Holy Blessed One* is 236×10^7 *parasangs*, i.e., 14×10^{10} (divine) *spans*. In the Biblical verse "Who has measured the waters in the hollow of his hand and marked off the heaven with a span" (Isaiah 40:12), the size of the Universe is one such span.

  The age/radius of the Universe is 13.82 billion ly. *Sefer HaTemunah* (by Nehunia ben Hakane, first century) and *Otzar HaChaim* (by Yitzchok deMin Acco, thirteenth century) deduced that the world was created *in thought* 42,000 divine years, i.e., $42{,}000 \times 365{,}250 \approx 15.3$ billion human years, ago. It counts, using the 42-letter name of the God at the start of Genesis, that now we are in the 6-th of the 7 cosmic *sh'mitah* cycles, each one being 7000 divine years long. *Tohu va-bohu* (formless and empty) followed and 6000 years ago the creation of the world *in deed* is posited.

  In the Talmud (Pesahim, 94), the Holy Spirit points out to "impious Nebuchadnezzar" (planning "to ascend above the heights of the clouds like the Most High"): "The distance from earth to heaven is 500 year's journey alone, the thickness of the heaven again 500 years…". This heaven is the *firmament* plate, and the journey is by walking. Seven other heavens, each 500 years thick, follow and the feet of the holy Creatures are equal to the whole…Their ankles, wings, necks, heads and horns are each consecutively equal to the whole." Finally, "upon them is the Throne of Glory which is equal to the whole". The resulting journey of 4,096,000 years amounts, at the rate of 80 miles ($\approx$ 129 km) per day, to $\approx$2600 AU, i.e., $\approx \frac{1}{100}$ of the actual distance to Proxima Centauri, the nearest other star. Also, in Talmud, the width of *Jacob's Ladder* (bridge to Heaven that Jacob dreams about, described in the Book of Genesis) is computed as 8000 parasangs.

  On the other hand, *Baraita de Massechet Gehinom* affirms in Section VII.2 that Hell consists of 7 cubic regions of side 300 year's journey each; so, 6300 years altogether. According to the Christian Bible (Chap. 21 of the Book of Revelation), New Heavenly Jerusalem (a city that is or will be the dwelling place of the Saints) is a cube of side 12,000 furlongs ($\approx$ 2225 km), or a similar pyramid or spheroid.

  Islamic tradition (Dawood, Book 40, Nr. 470) also attributes a journey of 71–500 years (by horse, camel or foot) between each *samaa'a* (the ceiling containing one of the seven luminaries: Moon, Mercury, Venus, Sun, Mars, Jupiter, Saturn). Besides 7 heavens (as in Judaism and Hinduism), Shia Islam and Sufism have 7 depths of esoteric meanings of Quran, with only God knowing the 4-th meaning.

  The Vedic text (Pancavimsab Brahmana, c. 2000 BC) states that the distance to Heaven is 1000 Earth diameters and the Sun (the middle one among seven luminaries) is halfway at 500 diameters. A similar ratio 500–600 was expected till the first scientific measurement of 1 AU (mean Earth-Sun distance) by Cassini and Richter, 1672. The actual ratio is $\approx 11{,}728$.

The sacred Hindu number 108 ($=6^2 + 6^2 + 6^2 = \prod_{1\leq i\leq 3} i^i$), connected to the Golden Ratio as the interior angle 108° of a regular pentagon, is traced to the following Vedic values: 108 Sun's diameters for the Earth-Sun distance and 108 Moon's diameters for the Earth-Moon distance. The actual values are (slightly increasing) ≈ 107.6 and ≈ 110.6; they could be computed during an eclipse, since the angular size of the Moon and Sun, viewed from the Earth, is almost identical.

Also, the ratio between the Sun and Earth diameters is ≈ 108.6, but it is unlikely that Vedic sages knew this. In Ayurveda, the devotee's distance to his "inner sun" (God within) consists of 108 steps; it corresponds to 108 beads of *mala* (rosary): by saying beads, the devotee does a symbolic journey from his body to Heaven.

In Bhagavata, the total diameter of the Universe is about 6.4 billion km.

- **Swedenborg heaven distances**

  The Swedish scientist and visionary Emanuel Swedenborg (1688–1772), in Section 22 (Nos. 191–199, *Space in Heaven*) of his main work *Heaven and Hell* (1952, first edition in Latin, London, 1758), posits: "distances and so, space, depend completely on interior state of angels". A move in heaven is just a change of such a state, the length of a way corresponds to the will of a walker, approaching reflects similarity of states. In the spiritual realm and afterlife, for him, "instead of distances and space, there exist only states and their changes".

- **Safir distance**

  According to Islamic law, a traveler may shorten the prayers, combine them, and be permitted to break the fast of Ramadan if the travel (*safir*) exceeds some minimum distance. Hanafi, the largest Sunni school of jurisprudence, define such **safir distance** as 3 days of continuous journey (in the great part of the day and at a moderate speed) or 15 *farsakh* (ancient unit of length, called also *parasang*).

  Three other main schools define it as 2 days of such journey or as 16 farsakh, computed differently. This distance is usually approximated as 80 or 83 km and applied for travel by camel, car, plane or ship. Another strong opinion, by Ibn Taymiyya, claims that safir is not merely a distance but also a state of mind, an exposure to the wilderness; so, any distance customarily considered traveling is safir.

- **Sabbath distance**

  The **Sabbath distance** (or *rabbinical mile*) is a **range distance**: 2000 Talmudic cubits (960–11,152 m, cf. **cubit** in Chap. 27) which an observant Jew should not exceed in a public thoroughfare from any given private place on the Sabbath day. It is about the distance covered by an average man in 18 min.

  Other Israelite/Talmudic length units are: a day's march, parsa, stadium (40, 4, $\frac{2}{15}$ of the rabbinical mile, respectively), and span, hasit, palm, thumb, middle finger, little finger ($\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{6}$, $\frac{1}{24}$, $\frac{1}{30}$, $\frac{1}{36}$ of the Talmudic cubit, respectively).

[redacted]

- **Bible code distance**

  Witztum, Rips and Rosenberg, 1994, claimed to have discovered a meaningful subtext of the Book of Genesis, formed by uniformly spaced letters. The text was seen as written on a cylinder on which it spirals down in one long line. Many reactions followed, including criticism by McKay, Bar-Natan, Bar-Hillel and Kalai, 1999, in the same journal *Statistical Science*.

  The following **Bible code distance** d_t between two letters, that are t positions apart in the text, was used. Let h be the circumference of the cylinder, and let q and r be the quotient and reminder, respectively, when t is divided by h, i.e., $t = qh + r$ with $0 \leq r \leq h - 1$. Then $d_t = \sqrt{q^2 + r^2}$ if $2q \leq h$, and $d_t = \sqrt{(q+1)^2 + (r-h)^2}$, otherwise. It is, approximatively, the shortest distance between those letters along the cylinder surface; cf. **cylindrical distance** (Chap. 20).

- **Distance numbers**

  On Maya monuments usually only one anchor event is dated absolutely, in the linear *Mesoamerican Long Count calendar* by the number of days passed since the mythical creation on August 11, 3114 BC of the fourth world, which completed a *Great Period* of 13 b'ak'tuns ($\approx$ 5125 years) on December 21, 2012.

  The other events were dated by adding to or subtracting from the anchor date some **distance numbers**, i.e., periods from the cyclical 52-year *Calendar Round*.

- **Antinomy of distance**

  The **antinomy of distance**, as introduced in [Bull12] for aesthetic experiences by the beholder and artist, is that both should find the right amount of **emotional distance** (neither too involved, nor too detached), in order to create or appreciate art. The fine line between objectivity and subjectivity can be crossed easily, and the amount of distance can fluctuate in time.

  The **aesthetic distance** is a degree of emotional involvement of the individual, who undergoes experiences and objective reality of the art, in a work of art. It refers to the gap between the individual's conscious reality and the fictional reality presented in a work of art. It means also the frame of reference that an artist creates, by the use of technical devices in and around the work of art, to differentiate it psychologically from reality; cf. **distanciation**.

  Some examples are: the perspective of a member of the audience in relation to the performance, the psychological/emotional distance between the text and the reader, the *actor-character distance* in the Stanislavsky system of acting.

  Antinomy between inspiration and technique (embracement and estrangement) in performance theory is called the *Ion hook* since Ion of Ephesus (a reciter of rhapsodic poetry, in a Platon's dialog) employed a double-consciousness, being ecstatic and rational. The acting models of Stanislavsky and Brecht are, respectively, incarnating the role truthfully and standing artfully distanced from it. Cf. **role distance**.

  [Morg76] defines pastoral ecstasy as the experience of *role-distancing*, or the authentic self's supra-role suspension, i.e., the capacity of an individual to stand outside or above himself for purposes of critical reflection. Morgan concludes:

"The authentic self is an *ontological possibility*, the social self is an *operational inevitability*, and awareness of both selves and the creative coordination of both is the gift of ecstasy". Cf. **Lévinas distance to Other**.

The **historical distance**, in terms of [Tail04], is the position the historian adopts *vis-à-vis* his objects—whether far-removed, up-close, or somewhere in between; it is the fantasy through which the living mind of the historian, encountering the inert and unrecoverable, positions itself to make the material look alive. The antinomy of distance appears because historians engage the past not just intellectually but morally/emotionally. The formal properties of historical accounts are influenced by the affective, ideological and cognitive commitments of their authors.

A variation of the antinomy of distance appears in critical thinking: the need to put some emotional and **epistemic** (intellectual) distance between oneself and ideas, in order to better evaluate their validity and avoid *illusion of explanatory depth* (to fail see the trees for the forest). A related problem is how much distance people must put between themselves and their pasts in order to remain psychologically viable; Freud showed that often there is no such distance with childhoods.

- **Role distance**

In Sociology, Goffman, 1961, using a dramaturgical metaphor, defined **role distance** (or *role distancing*) as actions which effectively convey some disdainful detachment of the (real life) performer from a role he is performing. An example of social role distancing is when a teacher explains to students that his disciplinary actions are due only to his role as a teacher.

Goffman observed that children are able to merge doing and being, i.e., *embracement of the performer's role*, only from 3–4 years. Starting from about 5, their role distance (distinguishing being from doing) appears and expands, especially, at age 8, 11 and adult years.

Besides role embracement and role distance, one can play a role cynically in order to manage the outcomes of the situation (impression management). The most likely cause of role distancing is *role conflict*, i.e., the pressure exerted from another role to act inconsistently from the expectations of the first role.

A *frame* is a type of role (participant, observer, authority, critic, artist, etc.) given to a person in relation to a given event. The **frame distance**, introduced by Heathcote, 1980, in teaching drama, refer to a specific (implied by the frame) responsibility, interest, attitude and behavior of a person/student in this event/drama.

- **Distanciation**

In scenic art and literature, **distanciation** (Althusser, 1968, on Brecht's *alienation effect*) consists of methods to disturb purposely (in order to challenge basic codes and conventions of spectator/reader) the *narrative contract* with him, i.e., implicit clauses defining logic behavior in a story. The purpose is to differentiate art psychologically from reality, i.e., to create some **aesthetic distance**.

[redacted]

For Bakhtin, the mandate to "be outside" that which you create is a matter of subject-subject (as opposed to subject-object) relations. For Shklovsky the distancing of an object sharpens our perception and stimulates senses, thereby arousing us to artistic (as opposed to dull everyday) experience.

One of the distanciation devices is *breaking of the fourth wall*, when the actor/author addresses the spectators/readers directly through an imaginary screen separating them. The *fourth wall* is the conventional boundary between the fiction and the audience. It is a part of the *suspension of disbelief* between them: the audience tacitly agrees to provisionally suspend their judgment in exchange for the promise of entertainment. Cf. **distancing** and **distantiation**.

- **Narrative distance**

  The author creates a persona of *narrator*, who tell the story, usually, from the point of view of first- or third-person. **Narrative distance** is (Genette, 1980) the distance between the narrator and the story's characters, setting, events and objects.

  The closest possible distance—the narrator reports on the thoughts and feeling (even unconscious ones) of the characters, while the farthest distance—reporting only actions and situations. The author can vary this distance; say, the third-person omniscient narrator can zoom in and out of character's perspectives.

- **Ironic distance**

  Rhetorical writer or speaker does not allow audience to maintain an objective or fixed distance from the story. He intrude to distance himself from characters in a story or from his own remarks. **Ironic distance** refers to the narrative irony: distance of knowledge between author/narrator/character/reader.

  As a literary device, irony implies a distance between what is said and what is meant. Irony is also the art of juxtaposing incongruous parts; so, an *ironic distance* also mean the closeness between two things that never meet.

- **Epistemic distance**

  Epistemic (or *intellectual*) **distance** from something refers to the degree of difficulty involved in knowing it. For example, conditional rhetorical constructions are used in discourse, in order to indicate this distance. Mejias (2005) *epistemological distance* between things is the difference in degree of knowledge justification.

- **Representation of distance in Painting**

  In Western Visual Arts, the *distance* is the part of a picture representing objects which are the farthest away, such as a landscape; it is the illusion of 3D depth on a flat picture plane. The *middle distance* is the central part of a scene between the foreground and the background (implied horizon).

  Perspective projection draws distant objects as smaller to provide additional realism by matching the decrease of their visual angle; cf. Chap, 6. A *vanishing point* (or *point of distance*) is a point at which parallel lines receding from an observer seem to converge. (For a meteor shower, the *radiant* is the point in the sky, from which meteors appear to originate.) *Linear perspective* is a drawing with 1–3 vanishing points; usually, they are placed on the horizon and equipartition it.

In a *curvilinear perspective*, there are ≥ 4 vanishing points; usually, they are mapped into and equipartition a *distance circle*. 0-*point perspective* occurs if the vanishing points are placed outside the painting, or if the scene (say, a mountain range) does not contain any parallel lines. Such perspective can still create a sense of *depth* (3D distance) as in a photograph of a mountain range.

In a *parallel projection*, all sets of parallel lines in 3D object are mapped to parallel lines in 2D drawing. This corresponds to a perspective projection with an infinite *focal length* (the distance from the image plane to the projection point).

Axonometric projection is parallel projection which is *orthographic* (i.e., the projection rays are perpendicular to the projection plane) and such that the object is rotated along one or more of its axes relative to this plane. The main case of it, used in Engineering Drawing, is **isometric projection** in which the angles between three projection axes are the same, or $\frac{2\pi}{3}$.

In Chinese Painting, the *high-distance*, *deep-distance* or *level-distance* views correspond to picture planes dominated, respectively, by vertical, horizontal elements or their combination. Instead of the perspective projection of a "subject", assuming a fixed position by a viewer, Chinese classic hand scrolls (up to 10 m in length) used axonometric projection. It permitted them to move along a continuous/seamless visual scenario and to view elements from different angles. It was less faithful to appearance and allowed them to present only three (instead of five) of six surfaces of a normal interior. But in Chinese Painting, the focus is rather on symbolic and expressionist representation.

- **Scale in art**

  In drawing, the **scale** refers to the proportion or ratio that defines the size relationships. It is used to create the illusion of correct size relationships between objects and figures. The *relative scale* is a method used to create and determine the spatial position of a figure or object in the 3D picture plane: objects that are more distant to the viewer are drawn smaller in size. In this way, the relative size of an object/figure creates the illusion of space on a flat 2D picture.

  In an architectural composition, the **scale** is the two-term relationship of the parts to the whole which is harmonized with a third term—the observer. For example, besides the proportions of a door and their relation to those of a wall, an observer measures them against his own dimensions.

  The **scale** of an outdoor sculpture, when it is one element in a larger complex such as the facade of a building, must be considered in relation to the scale of its surroundings. In *flower arrangement* (floral decoration), the **scale** indicates relationships: the sizes of plant materials must be suitably related to the size of the container and to each other.

  The *hierarchical scale* in art is the manipulation of size and space in a picture to emphasize the importance of a specific object. Manipulating the scales was the theme of *Measure for Measure*, an art/science exhibition at the Los Angeles Art Association in 2010. Examples of the interplay of the small and the large in literature are Swift's *Gulliver's Travels* and Carrol's *Through the Looking Glass*. In the cinema, the spectator can easily be deceived about the size of objects, since scale constantly changes from shot to shot.

[redacted]

In Advertising and Packaging, the size changes the meaning or value of an object. The idea that "bigger is better" is validated by the sales of sport utility vehicles, super-sized soft drinks and bulk food at Wal-Mart.

In reverse, the principle "small is beautiful" is often used to champion small, appropriate objects and technologies that are believed to empower people more. For example, small-sized models sell the benefits of diet programs and fitness regimes designed to scale back people's proportions. Examples of Japanese miniaturization culture are bonsai and many small/thin portable devices.

- **Distances in Interior Design**

  In Interior Design, the *scale* refer to how an item relates to the size of the room or the owner, and the *proportion* refer to the shape of an item and how it relates to other objects in the room. The vertical, horizontal, diagonal and curved lines give a room a feeling of formalness, casualness, transition and sensuality, respectively. Other required space relationships are *balance* (equal weight between objects on either side of a room) and *rhythm* (repetition of patterns, color, or line).

  Workplane is the height at which an activity takes place; it is about 90, 75–90 and 75 cm for a kitchen, bath and a dining room or desk. In a kitchen, the perimeter of the *work triangle* formed by sink, cooking surface and refrigerator ideally should be 3.5–7.5 m. In a living room, the triangle of *focal points* to emphasize is formed usually at the door or fireplace, TV, big window, sofa. Other examples of recommended distances are: 35–45 cm between the sofa (or chairs) and coffee table, 60 cm between dining chairs and at least 90 cm for traffic lanes.

  Used in lighting calculations, the **room cavity ratio** (or RCR) is $\frac{5hP}{2A}$, where h, P, A are the ceiling height, perimeter and area of the room. So, $RCR = \frac{5h(l+w)}{lw}$ for a rectangular room of length l and width w.

- **Spatialism**

  Spatialism (or *Spazialismo*) is an art movement founded by Lucio Fontana in Milan in 1947, intended to synthesize space, color, sound, movement and time into a new "art for the Space Age". Instead of the illusory virtual space of traditional *easel* (i.e., of a size and on a material suitable for framing) painting, he proposed to unite art and science to project color and form into real space by the use of up-to-date techniques, say, TV and neon lighting. His *Spatial Concept* series consisted of holes or slashes, by a razor blade, on the surface of monochrome paintings.

- **Spatial music**

  Spatial music refers to music and sound art (especially, electroacoustic), in which the location and movement of sound sources, in physical or virtual space, is a primary compositional parameter and a central feature for the listener.

  Space music is gentle, harmonious sound that facilitates the experience of contemplative spaciousness. Engaging the imagination and generating serenity, it is particularly associated with ambient, New Age, and electronic music.

- **Distance-named cultural products**

  Far Near Distance is the name of the program of the House of World Cultures in Berlin which presents contemporary positions of Iranian artists. Examples of similar use of distance terms in modern popular culture follow.

  “Some near distance” and “Zero/Distance” are the titles of art exhibitions of Mark Lewis (Bilbao, 2003) and Jim Shrosbree (Des Moines, Iowa, 2007). “A Near Distance” is a paper collage by Perle Fine (New York, 1961); “Quiet Distance” is a fine art print by Ed Mell. “Distance” is a Windows/Mac/Linux survival racing game; “Dream Drop Distance” is a video game for Nintendo.

  “Distance” is a Japanese film directed by Hirokazu Koreeda (2001) and an album of Utada Hikaru (her famous ballad is called “Final Distance”). It is also a song by Christina Perry, the stage name of a musician Greg Sanders and the name of the rock/funk band led by Bernard Edwards. “The Distance” is a US film directed by Benjamin Busch (2000), an album by the band “Silver Bullet” and a song by the band “Cake”. “Near Distance” is a musical composition by Chen Yi (New York, 1988) and lyrics by the quartet “Puressence”.

  “Distance to Fault”, “Distance from Shelter”, “Long Distance Calling” are the rock bands. Among popular albums are “The Tyranny of Distance”, “The Great Cold Distance”, “Close the Distance”, “The Distance to Here”, “Love and Distance”, “Long Distance Voyager” and “The Crawling Distance”, “This Magnificent Distance” by the bands “Washington, D.C.”, “Katatonia”, “Go Radio”, “Live”, “The Helio Sequence”, “The Moody Blues” and Robert Pollard, Chris Robinson.

  The terms *near distance* and *far distance* are also used in Ophthalmology and for settings in some sensor devices.
- **Distance-related quotes**
  - “Respect the gods and the devils but keep them at a distance.” (Confucius)
  - “Sight not what’s near through aiming at what’s far.” (Euripides)
  - “It is when suffering seems near to them that men have pity.” (Aristotle)
  - “Distance in space or time weakened all feelings and all sorts of guilty conscience.” “Distance is a great promoter of admiration.” (Denis Diderot)
  - “Our main business is not to see what lies dimly at a distance, but to do what lies clearly at hand.” (Thomas Carlyle)
  - “We can only see a short distance ahead, but we can see plenty there that needs to be done.” (Alan Turing)
  - “The foolish man seeks happiness in the distance; the wise grows it under his feet.”(Julius Robert Oppenheimer)
  - “The very least you can do in your life is to figure out what you hope for. And the most you can do is live inside that hope. Not admire it from a distance but live right in it, under its roof.” (Barbara Kingsolver)
  - “Better is a nearby neighbor, than a far off brother.” (Proverbs 27:10, Bible)
  - “These [patriarchs] all died in faith without receiving the things promised [Canaan, Messiah, Gospel], but they saw them and welcomed them from

[redacted]

a distance, admitting that they were strangers and pilgrims on the earth." (Hebrews 11:13, Bible)

- "By what road", I asked a little boy, sitting at a cross-road, "do we go to the town?"—"This one", he replied, "is short but long and that one is long but short". I proceeded along the "short but long road". When I approached the town, I discovered that it was hedged in by gardens and orchards. Turning back I said to him, "My son, did you not tell me that this road was short?"—"And", he replied, "Did I not also tell you: "But long"?" (Erubin 53b, Talmud)
- "The Prophet Muhammad was heard saying: "The smallest reward for the people of paradise is an abode where there are 80,000 servants and 72 wives, over which stands a dome decorated with pearls, aquamarine, and ruby, as wide as the distance from Al-Jabiyyah [a Damascus suburb] to Sana'a [Yemen]." (Hadith 2687, Islamic Tradition)
- "The closer the look one takes at the world, the greater distance from which it looks back." (Karl Kraus)
- "Telescopes and microscopes are designed to get us closer to the object of our studies. That's all well and good. But it's as well to remember that insight can also come from taking a step back." (New Scientist, March 31, 2012)
- "Where the telescope ends, the microscope begins. Which of the two has the grander view?" (Victor Hugo)
- "Nature uses only the longest threads to weave her patterns." (Richard Feynman)
- "We're about eight Einsteins away from getting any kind of handle on the universe." (Martin Amis)
- "It is true that when we travel we are in search of distance. But distance is not to be found. It melts away. And escape has never led anywhere ... What are we worth when motionless, is the question." (Antoine de Saint-Exupéry)
- "If you want to build a ship, don't drum up people to collect wood and don't assign them tasks and work, but rather teach them to long for the endless immensity of the sea." (Antoine de Saint-Exupéry)
- "Ships at a distance have every man's wish on board." (Zora Neale Hurston)
- "If you've never stared off in the distance, then your life is a shame." (Adam Duritz)
- "Every once in a while, people need to be in the presence of things that are really far away." (Ian Frazier)
- "Only those who will risk going too far can possibly find out how far one can go." (Thomas Stearns Eliot)
- "Distance is to love like wind is to fire ... it extinguishes the small and kindles the great." (source unknown)
- "I could never take a chance of losing love to find romance
  In the mysterious distance between a man and a woman." (Performed by U2)
- "In true love the smallest distance is too great, and the greatest distance can be bridged." (Hans Nouwens)

- "Love is like a landscape which doth stand
  Smooth at a distance, rough at hand." (Robert Hegge)
- "Life is like a landscape. You live in the midst of it but can describe it only from the vantage point of distance." (Charles Lindbergh)
- "Distance between two people is only as one allows it to be." (source unknown)
- "It is only the mountains which never meet." (french proverb)
- "Nothing makes Earth seem so spacious as to have friends at a distance; they make the latitudes and longitudes." (Henri David Thoreau)
- "Distance can endear friendship, and absence sweeteneth it." (James Howell)
- "The word is distance within non-distance, that is, the width of a gap that every letter stresses while bridging it. What is said is always said in relation to what will never be expressed. At these limits we recognize ourselves." (Edmond Jabès)
- "The poet's other readers are the ancient poets, who look upon the freshly written pages from an incorruptible distance. Their poetic forms are permanent, and it is difficult to create new forms which can approach them." (Salvatore Quasimodo)
- "Sad things are beautiful only from a distance ... From a distance of 130 years i'm going to distance myself until the world is beautiful..." (Tao Lin)
- "Dying away into the distance, prose turns into poetry, speech into vocalise, language into music." (Berthold Hoeckner)
- "Everything becomes romantic and poetic, if one removes it to a distance ... Distant philosophy sounds like poetry – for each call into the distance becomes a vowel ... Everything becomes poetry – poem from afar." (Novalis)
- "The appropriated way to determine whether a painting is melodious is to look at it from a distance so as to be unable to comprehend its subject or its lines." (Charles Baudelaire)
- "There is no object so large ... that at great distance from the eye it does not appear smaller than a smaller object near." (Leonardo da Vinci)
- "Distance lends enchantment to the view,
  And robes the mountain in its azure hue." (Thomas Campbell)
- "There are charms made only for distant admiration."
  "Distance has the same effect on the mind as on the eye." (Samuel Johnson)
- "Age, like distance, lends a double charm." (Oliver Wendell Holmes)
- "Distance not only gives nostalgia, but perspective, and maybe objectivity." (Robert Morgan)
- "It is the just distance between partners who confront one another, too closely in cases of conflict and too distantly in those of ignorance, hate and scorn, that sums up rather well, I believe, the two aspects of the act of judging. On the one hand, to decide, to put an end to uncertainty, to separate the parties; on the other, to make each party recognize the share the other has in the same society." (Paul Ricoeur)
- "Authority doesn't work without prestige, or prestige without distance." (Charles de Gaulle)

[redacted]

- “The human voice can never reach the distance that is covered by the still small voice of conscience.” (Mohandas Gandhi)
- “A smile is the shortest distance between two people.” (Victor Borge)
- “The shortest distance between two points is under construction.” (Leo Aikman)
- “A straight line may be the shortest distance between two points, but it is by no means the most interesting.” (Third Doctor from BBC TV series *Doctor Who*)
- “In politics a straight line is the shortest distance to disaster.” (John P. Roche)
- “Fret not where the road will take you. Instead concentrate on the first step. That is the hardest part and that is what you are responsible for. Once you take that step let everything do what it naturally does and the rest will follow. Do not go with the flow. Be the flow.” (Shams Tabrisi)
- “The distance is nothing; only the first step that is difficult.” (Marie du Deffand)
- “A perfect run has nothing to do with distance. It’s when your stride feels comfortable.” (Sean Astin)
- “Fill the unforgiving minute with sixty seconds worth of distance run.” (Rudyard Kipling)
- “The distance between dreams and reality is called Discipline.” (Albert Wright)
- “The distance between who you are and who you might be is closing.” (Jan Chipchase)
- “Everywhere is within walking distance if you have the time.” (Steven Wright)
- “Time is the longest distance between two places.” (Tennessee Williams)
- “There is an immeasurable distance between late and too late.” (Og Mandino)
- “They couldn’t hit an elephant at this distance.” (last words of John Sedgwick, seconds before he was mortally wounded)
- “The distance that the dead have gone does not at first appear;
  Their coming back seems possible for many an ardent year.” (Emily Dickinson)
- “A vast similitude interlocks all . . . All distances of place however wide,
  All distances of time, all inanimate forms, all souls . . . ” (Walt Whitman)

Chapter 29
Other Distances

In this chapter we group together distances and distance paradigms which do not fit in the previous chapters, being either too practical (as in equipment), or too general, or simply hard to classify.

29.1 Distances in Medicine, Anthropometry and Sport

- **Distances in Medicine**

  Some examples from this vast family of physical distances follow.

  In Dentistry, the **interocclusal distance**: the distance between the occluding surfaces of the maxillary and mandibular teeth when the mandible is in a physiologic rest position. The **interarch** and **interridge distances**: the vertical distances between the maxillary and mandibular arches, or, respectively, ridges.

  The **intercanine distance**: the distance between the distal surfaces of the maxillary canines on the curve (the circumference of six maxillary anterior teeth).

  The **interincisor distance**: the distance between the upper and lower incisors.

  The **interproximal distance**: the **spacing distance** between adjacent teeth; *mesial drift* is the movement of the teeth slowly toward the front of the mouth with the decrease of the interproximal distance by wear.

  The **biologic width**: the distance between the deepest point of the gingival sulcus and the alveolar bone crest. The *crown-to-root-ratio*: the ratio of the length of the part of a tooth that appears above the alveolar bone versus what lies below it.

  The **interbrow distance**: the distance between the eyebrows.

  The **interaural** (or *biauricular*) **distance**: the distance between the ears.

  Lumbar spinal canal (LSC) is the tunnel that the spinal cord and nerves pass through. The **LSC diameter** is the longest horizontal distance between any two points on the lumbar edge; normally, it is 15–27 mm.

© Springer-Verlag Berlin Heidelberg 2016

M.M. Deza, E. Deza, *Encyclopedia of Distances*,
DOI 10.1007/978-3-662-52844-0_29

[redacted]

The **rectosacral distance**: the shortest distance from the rectum to the *sacrum* (triangular bone at the base of the spine, inserted between the two hip bones) between the 3-rd and 5-th sacral vertebra. It is at most 10 mm in adults.

The **anogenital distance** (or AGD): the length of the *perineum*, i.e., the region between the anus and genital area (the anterior base of the penis for a male). For a male it is 5 cm in average (twice what it is for a female). ARD is a measure of physical masculinity and, for a male, lower ARD correlates with lower fertility.

The **internipple distance**: the distance between nipples. "Ideally", the nipples and sternal notch form an equilateral triangle with a side of 21 cm, and the nipples are at the middle of the humeral shoulder-elbow distance. The average *areolar diameter* is 38 mm for a mature woman and 25 mm for a male.

The **intercornual distance**: the distance between uterine horns (2–4 cm). The **C-V distance**: clitoris-vagina distance (2.3–3 cm); < 2.5 cm tend to yield reliable orgasms from intercourse alone, while > 3 cm almost exclude it. The *clitoral index* (CI): product of the crosswise (3–4 mm) and lengthwise (4–5 mm) widths of the external portion of the clitoris; CI is a measure of virilization in women. Mean *vaginal depth* (cervix-introitus) is 7–8 cm, increasing to 11–13 cm during sexual arousal. For a human penis, mean *flaccid and erect lengths* are 9.16 and 13.2 cm.

A **pelvic diameter** is any measurement that expresses the diameter of the birth canal in the female. For example, the *diagonal conjugate* (13 cm) joins the posterior surface of the pubis to the tip of the sacral promontory, and the *true* (or *obstetric*, *internal) conjugate* (11.5 cm) is the anteroposterior diameter of the pelvic inlet.

In Obstetrics, the **fundal height** is the size the mother's uterus (the distance between the tops of uterus and pubic bone) used to assess fetal growth and development during pregnancy. The **crown-rump length** is the length of human embryos/fetuses (the distance, determined from ultrasound imagery, from the top of the head to the bottom of the buttocks) used to estimate gestational age.

Metra and uterus are (Greek and Latin) medical terms for the womb. *Metropathy* is any disease of the uterus, say, *metritis* (inflammation), *metratonia* (atony), *metrofibroma*. *Metrometer* is an instrument measuring the womb's size. *Metrography* (or *hysterography*) is a radiographic examination of the uterine cavity filled with a contrasting medium. Cf. **meter-related terms** in Chap. 27.

In Radiography, the **teardrop distance**: the distance from the lateral margin of the pelvic teardrop to the most medial aspect of the femoral head; a widening of ≥ 1 mm indicates excess hip joint fluid and so inflammation. The **intertrochanteric distance**: the distance between femurs. The **interpediculate distance**: the distance between the vertebral pedicles. The **source-skin distance**: the distance from the focal spot on the target of the X-ray tube to the subject's skin.

In *Intubation* (insertion of a tube into a body canal or hollow organ, to maintain an opening or passageway), the **insertion distance**: the distance from the body aperture at which the tubing is advanced. The *French size* of a catheter with external diameter D is $\pi D \approx 3D$; so, 20 F means $D = 6.4$ mm.

In Anesthesia, the **thyromental distance** (or TMD): the distance from the upper edge of the thyroid cartilage (laryngeal notch) to the *menton* (tip of the chin). The **sternomental distance**: the distance from the upper border of the manubrium sterni to the menton. The **mandibulo-hyoid distance**: mandibular length from menton to hyoid. When the above distances are less than 6–6.5 cm, 12–12.5 cm and 4 cm, respectively, a difficult intubation is indicated.

The *depth of anesthesia* is a number expressing the likehood of awareness by the degree of slowing and irregularity in electroencephalogram (EEG) signals. Also, at loss of consciousness, high frequency (12–35 Hz) brain waves are replaced by two (low, < 1 Hz, and alpha, 8–12 Hz) superimposed waves. Even beyond a flat line EEG, some neuronal spikes come to the cortex from the hippocampus.

The **sedimentation distance** (or ESR, *erythrocyte sedimentation rate*): the distance red blood cells travel in 1 h in a sample of blood as they settle to the test tube's bottom. ESR indicates inflammation and increases in many diseases.

The **stroke distance**: the distance a column of blood moves during each heart beat, from the aortic valve to a point on the arch of the aorta.

The **distance between the lesion and aortic valve** being < 6 mm, is an important predictor, available before surgical resection of DSS (discrete subaortic stenosis), or reoperation for recurrent DSS. The **aortomesenteric distance** (between aorta and superior mesenteric artery) correlates with the *body mass index*.

The **aortic diameter**: the maximum diameter of the outer contour of the aorta. It, as well as the cross-sectional diameter of the left ventricle, varies between the ends of the *systole* (the time of ventricular contraction) and *diastole* (the time between contractions). The smallest and largest cardiac dimensions are LVE (left ventricle end-) *systolic* and *diastolic diameters*; the *strain* is the ratio between them.

The **dorsoventral interlead distance** of an implanted pacemaker or defibrillator: the horizontal separation of the right and left ventricular lead tips on the lateral chest radiograph, divided by the *cardiothoracic ratio* (ratio of the cardiac width to the thoracic width on the posteroanterior film).

The **distance factor** is a crude measure $\frac{l}{d} - 1$ of arterial tortuosity, where l is the vessel length and d is the Euclidean distance between its endpoints.

In Nerve Regeneration by transplantation of cultured stem cells, the **regeneration distance** is the distance between the point of insertion of the proximal stump and the tip of the most distal regenerating axon.

The *small-for-size syndrome* (SFSS) is acute liver failure resulting from the transplantation of a too small (usually $< 0.8\,\%$ of recipient weight) *graft* (donor liver).

A **distant flap** is a procedure moving tissue (skin, muscle, bone, or some combination) from one part of the body, where it is dispensable, to another part.

The length of the alimentary (mouth-to-anus) tract is $\approx$ 9 m in a dead and, due to muscle tone, 5–6 m in a leaving human. Transit takes 30–50 h.

[redacted]

In Laser Treatments, the **extinction length** and **absorption length** of the vaporizing beam are the distances into the tissue along the ray path over which 90 % (or 99 %) and 63 %, respectively, of its radiant energy is absorbed.

In Ophthalmic Plastic Surgery, the **marginal reflex distances** MRD_1 and MRD_2 are the distances from the center of the pupil (identified by the corneal reflex created by shining a light on the pupil) to the margin of the upper or lower eyelid, while the **vertical palperbal fissure** is the distance between these eyelids.

The main distances used in Ultrasound Biomicroscopy (for glaucoma treatment) are the **angle-opening distance** (from the corneal endothelium to the anterior iris) and the **trabecular ciliary process distance** (from a particular point on the *trabecular mesh-work* to the *ciliary process*).

In Medical Statistics, *length bias* is a selection bias that can occur when the lengths of intervals are analyzed by selecting random intervals in space or time. This process favors longer intervals, thus skewing the data. For example, screening over-represents less aggressive disease, say, slower-growing tumors.

- **Distances in Oncology**

In Oncology, the **tumor radius** is the mean radial distance R from the tumor origin (or its center of mass) to the tumor-host interface (the tumor/cell colony border). The cell proliferation along $[0, R]$ is ≈ 0 up to some r_0, then increases only linearly up to some r_1, and it happens mainly within $[r_1, R]$.

The **tumor diameter** is the greatest vertical diameter of any section; the *tumor growth* is the geometric mean of its three perpendicular diameters. The *average diameter* is $\frac{L+W+H}{3}$ where L, W, H are the longest length, width and height.

Tubiana, 1986, claims that for each tumor type a critical tumor diameter and mass for metastatic spread exists and this threshold may be reached before the primary tumor is detectable. For breast cancer, metastases were found in 50 % of the women whose primary tumor had a diameter of 3.5 cm, i.e., a mass ≈ 22 g.

In the *tumor, node, metastasis* (TNM) classification, describing the stage of cancer in a patient's body, the parameter T is the **tumor size** (direct extent of the primary tumor) by the categories T-1, T-2, T-3, T-4. In breast cancer, T-1, T-2, T-3 are $< 2, 2$–$5, > 5$ cm and T-4 is a tumor of any size that has broken through the skin, or is attached to the chest wall. A *clinical size* is 10^9–10^{11} cells.

In Oncological Surgery, the **margin distance** (or *margins of resection*) is the distance between a tumor and the ink-marked edge of *tumor bed*, i.e., normal-appearing tissue surrounding tumor that is removed along with it in order to prevent local recurrence. If the margins, as checked by a pathologist under microscope, are *positive* (cancer cells are found in the ink), then more surgery is needed. The margins are *negative* (or *clear, clean*) if no cancer cells are found "close" to the ink.

The **perfusion distance** is the shortest distance between the infusion outlet and the surface of the electrodes during radio-frequency tumor ablation.

In Radiation Oncology, the **maximum heart distance** MHD is the maximum distance of the heart contour (as seen in the beam's eye view of the medial tangential field) to the medial field edge, and the **central lung distance** CLD is the distance from the dorsal field edge to the thoracic wall. An "L-bar" armrest, used to position the arm during breast cancer irradiation, decreases these distances.

A **distant cancer** (or relapse, metastasis) is a cancer that has spread from the original (primary) tumor to distant organs or distant lymph nodes. It can happen by **long-distance dispersal** (Chap. 23) and by dividing of *cancer stem cell*. DDFS (Distant Disease-free Survival) is the time until such an event.

According to Hanahan–Weinberg, 2000, tumor progress via evolution-like process of genetic changes which can be grouped into six hallmarks. Tumorigenesis requires a mutation pathway of 4–6 events among them to occur in the lineage of one cell. Spencer et al., 2006, define *tumor heterogeneity* as

$$\frac{2}{n(n-1)} \sum_{1 \le i < j \le p} n_i n_j d_{ij},$$

where $n = \sum_{1 \le i \le p} n_i$, p, n_i are the numbers of cells in a tumor, of distinct pathways, of cells with i-th pathway, and d_{ij} is the **ancestral path distance** (cf. **pedigree-based distances** in Chap. 23) between i-th and j-th pathways.

Similarly, the *distance to flu pandemic* is, say, the length of mutation pathway for a virus strain to become airborne transmissible among humans.

- **Distances in Rheumatology**

The main such distances (measured in cm to the nearest 0.1 cm) follow.

Occiput wall distance: the distance from the patient's occiput to the wall during maximal effort to touch the head to the wall, without raising the chin above its usually carrying level (when heels and, if possible, the back are against the wall).

Modified Schober test: the distance between two marked points (a point over the spinous process of $L5$ and the point 10 cm above) measured when the patient is extending his lumbar spine in a neutral position and then when he flexes forward as far as possible. Normally, the 10 cm distance increases to $\geq$ 16 cm.

Lateral spinal flexion: the distance from the middle fingertip to the floor in full lateral flexion without flexing forward or bending the knees or lifting the heels and attempting to keep the shoulders in the same place.

Chest expansion: the difference between full expiration and full inspiration, measured at the nipples.

Intermalleolar distance: the distance between the medial malleoli when the patient (supine, the knees straight and the feet pointing straight up) is asked to separate the legs as far as possible.

- **Distance healing**

Distance (or *distant*, *remote*) **healing** is defined (Sicher–Targ, 1998) as a conscious, dedicated act of mentation attempting to benefit another person's physical or emotional well-being at a distance. Cf. **action at a distance** in Chap. 24.

[redacted]

It includes prayer (intercessory, supplicative and nondirected), spiritual/mental healing and strategies purporting to channel some supra-physical energy (*noncontact therapeutic touch*, *Reiki healing*, *external qigong*).

Distant healing is part of popular alternative/complementary medicine but it is highly controversial: some positive results are attributed to a placebo effect. However, such rejection (as well as for homeopathy) is also a matter of belief.

In Chinese Medicine, the life-energy, *qi*, flow through 20 *meridians* connecting about 400 acupuncture points. Many *trigger points* (hyperirritable muscle spots) and *pressure points* in martial arts are related to above points. Many meridians are located along connective-tissue planes between muscles or muscle and bone.

In Ayurveda, the life force, *prana*, flow through $> 72{,}000$ *nadis* connecting *chakras* (intensity points); it also lists 13 internal *srotas* (physical body channels) and, by the number of orifices, 9 external ones for males and 12 for females. The main 13 chakras are evenly spaced, using the **interpupillary distance** as the unit. The root chakra is at the base of the spine, followed by chakras at the sex, navel, solarplexis, heart (two), Adam's apple, chin's tip, nose's tip, third eye, just above the forehead, the crown, one hand length above the head.

Distance medicine technologies are used to transmit/treat patient information, to submit prescriptions, to create distributive patient care and distributive learning. Examples of telephonic communication with patients are in: follow-up care, reminders, interactive systems, screening and access in hospital use.

- **Brain distances**

fMRI (functional Magnetic Resonance Imaging) measures, using MRI technology, brain activity by detecting changes associated with blood flow.

dMRI (diffusion MRI) noninvasively produces *in vivo* images of brain tissues weighted by their water diffusivity. The image intensities at each position are attenuated proportionally to the strength of diffusion in the direction of its gradient. Diffusion in tissues is described by a diffusivity tensor. Tensor data are displayed, for each voxel, by ellipsoids; their length in any direction is the diffusion distance molecules cover in a given time in this direction. The **diffusion tensor distance** is the length from the center to the surface of the diffusion tensor.

In brain MRI, the distances considered for *cortical maps* (i.e., outer layer regions of cerebral hemispheres representing sensory inputs or motor outputs) are: MRI **distance map** from the GW (gray/white matter) interface, **cortical distance** (say, between activation locations of spatially adjacent stimuli), **cortical thickness** (the shortest distance between the GW the boundary and the innermost surface of *pia mater* enveloping the brain) and *lateralization metrics*. In fact, language, for example, tends to be on the left, attention more on the right side of the brain.

The cortical thickness of Einstein's brain is 2.1 mm, while the average one is 2.6 mm; the resulting closer packing of neurons may speed up communication between them. This brain had a relatively larger (and more intricately folded) prefrontal cortex and an unusually high glia-to-neuron ratio. Also, the *corpus* *callosum* is thicker in many areas, indicating greater connectivity between the two hemispheres.

Stereotaxic coordinates of a point p in the human brain are given by a triple (x_p, y_p, z_p) in mm (Talairach–Tournoux, 1988), where the *anterior commissure* is the origin $(0, 0, 0)$. The dimensions x, y, z refer to the left-right (LR), posterior-anterior (PA) and ventro-dorsal (VD or inferior-superior) orthogonal axes with positive values for the right hemisphere, anterior part and dorsal part. The **Talairach distance** of a point p is its Euclidean distance $\sqrt{x_p^2 + y_p^2 + z_p^2}$ from the origin.

Among proto-humans, only Neanderthals had a larger (11.6 %) brain than *Homo sapiens*; we got the brain-size increasing gene 0.037 Ma ago. But over the last 0.03 Ma our brains have been shrinking, while craniums have been increasing. Our brain reaches its full size by age 20 and then shrinks (faster for men) by about 1 % per year. Brain accounts for $\frac{1}{5}$ of the total ($\approx$ 100 W) body energy consumption.

Network models (of human brain in a living person) derived from resting state fMRI and dMRI are similar and strongly correlate. Siegelmann et al., 2015, defined the **connectivity-distance** (or *network depth*) of a brain area, as increase in its average connectivity from sensory cortex inputs, derived from those MRI data. This depth is expected to be a proxy for an increase in number of sequential synapses, processing time and to correlate with degree of abstractness of cognitive behaviors.

- **Dysmetria**

  Dysmetria is a symptom of a cerebellar disorder or syndrome, expressed in a lack of coordination of movement typified by the undershoot (*hypometria*) or overshoot (*hypermetria*) of the intended position with the hand, arm, leg, or eye.

  More generally, *dysmetria* can refer to an inability to judge distance or scale, which is also one of symptoms of dyscalculia. The **distance constancy** (Chap. 28) is poor in schizophrenics; so, their visual perception is lacking in depth.

  Alice in Wonderland syndrome, affecting mainly children, is that objects appear either much smaller (*micropsia*) or larger (*macropsia*) than they are. Micropsia appear also in *Charles Bonnet syndrome*, affecting mainly vision-impaired elderly.

- **Space-related phobias**

  Several **space-related phobias** have been identified: *agoraphobia*, *astrophobia*, *claustrophobia*, *cenophobia*, and *acrophobia*, *bathophobia*, *gephyrophobia*, *megalophobia* which are, respectively, fear of open, celestial, enclosed, empty spaces, and heights, depths, bridges, large/oversized objects. *Autoscopy* (or *out-of-body experience*) is the hallucination of seeing one's own body at a distance.

  Among neuropsychological spatial disorders are: *Balint's syndrome* (inability to localize objects in space), *hemispatial neglect* (bias of attention to and awareness of the side of the hemispheric lesion) and *allochiria* (left-right disorientation).

[redacted]

Topographical disorientation is the inability to orient in the surrounding as a result of focal brain damage. Such agnosia with respect to self, to landmarks, to external environment, to new environments is called *egocentric*, *landmark*, *heading*, *anterograde*, respectively. *Dromosagnosia* is the loss of direction while driving.

In Chap. 28, among applications of **spatial language** is mentioned Grove's *clean space*: a neuro-linguistic psychotherapy based on the spatial metaphors produced by the client on his present and desired "space" (state).

- **Neurons with spatial firing properties**

  Known types of **neurons with spatial firing properties** are listed below; cf. also **spike train distances** in Chap. 23.

  Many mammals have in several brain areas *head direction cells*: neurons which fire only when the animal's head points in a specific direction.

  Place cells are principal neurons in the hippocampus that fire strongly whenever an animal is in a specific location (the cell's *place field*) in an environment.

  Grid cells are neurons in the entorhinal cortex that fire periodically and at very regular distances as an animal walks. Grid cells measure distance while place cells indicate location. But only place cells are sensitive (albeit weakly) to height.

  Spatial view cells are neurons in the hippocampus which fire when the animal views a specific part of an environment. They differ from head direction cells since they represent not a global orientation, but the direction towards a specific object. They also differ from place cells, since they are not localized in space.

  Border cells are neurons in the entorhinal cortex that fire when a border is present in the proximal environment.

  Mirror cells are neurons that fire both when an animal acts and when it observes the same action performed by another.

  Head direction cells of rats are fully developed before pups open their eyes and become mobile. Next to mature are place cells followed by grid cells. All navigational cell types mature before rat adolescence (about 30 days of age).

  The smallest processing module of cortical neurons is a *minicolumn*—a vertical column (of diameter 28–40 microns) through the cortical layers of the brain, comprising 80–120 neurons that seem to work as a team. There are about 2×10^8 minicolumns in humans. Smaller minicolumns (as observed in scientists and in people with autism) mean that there are more processing units within any given cortical area; it may allow for better signal detection and more focused attention.

- **Vision distances**

  The *interocular distance* is the distance between the centers of rotation of the eyeballs of an individual or between the oculars of optical instruments.

  The **interpupillary distance** (or *binocular pupillary distance*) is the distance (50–75 mm) between the centers of the pupils of the two eyes. The *monocular pupillary distance* is the distance from the center of the nose to the pupil.

Stereoacuity is the smallest detectable depth difference that can be seen in binocular vision.

Involuntary *pupil's diameter* dilation of 0.5 mm can indicate interest, attention or sexual stimulation. Straight (homosexual) observers' pupils increase linearly as the amount of clothing on opposite (same, respectively) sex models decreases.

The *near acuity* is the eye's ability to distinguish an object's shape and details at a near distance such as 40 cm; the **distance acuity** is the ability to do it at a far distance such as 6 m. The **distance vision** is a vision for objects that are at least 6 m from the viewer. *Optical near devices* are designed for magnifying close objects and print; **distance devices** are for magnifying things in the distance.

The *near distance* is the distance between the object and *spectacle* (eyeglasses) planes. The *vertex distance*: the distance between a person's glasses (spectacles planes) and their eyes (the corneal). The *infinite distance*: a distance of at least 6 m; so called because rays entering the eye from an object at that distance are practically as parallel as if they came from an infinitely far point.

The *default accommodation distance* (or *resting point of accommodation*, *RPA distance*) is the distance at which the eyes focus if there is nothing to focus on.

The *RPV distance* (or *resting point of vergence*) is the distance at which the eyes are set to *converge* (turn inward toward the nose) when there is no close object to converge on. It averages about 1.15 m when looking straight ahead and in to about 0.9 m with a 30° downward gaze angle. Ergonomists recommend the RPV distance as the eye-screen distance in sustained viewing, in order to minimize eyestrain.

The *least distance of distinct vision* (or *reference seeing distance*) is the minimum comfortable distance (usually, 25 cm) between the eye and a visible object. Ideal focus distances for reading and writing are within 37–62 cm from the eyes.

The **Harmon distance** (or *elbow distance*) is the optimal visual distance for reading and other near work. It is the distance from the elbow on the desk to the first *knuckle* (prominence of a joint connecting the finger to the hand).

The ideal *TV viewing distance* is 1.9 times the screen width, since then this width occupies a 30° angle from the viewing position. For multiple-row seating in the home theater, a viewing angle 26–36° is recommended.

The **Lechner distance** is the optimal viewing distance at which the human eye can best process the details given by High Definition TV resolution. For example, it is about 1.7 or 2.7 m for a 1080 HD TV with a screen size of 42 or 69 inches.

Lateral masking (or *crowding*) is impairment of peripheral object identification by *flankers* (nearby objects). *Critical spacing* (or *crowding distance*) is the minimum target-flanker distance that does not produce crowding of a target of fixed size.

The *throw distance* is the distance that the projector needs to be from the screen to project the optimum image. The *viewing-distance factor* is a ratio of the width of a projected image to the maximum acceptable viewer's distance from it.

[redacted]

The **laser hazard distance** is the safe viewing distance for direct exposure to visible laser beams.

- **Gait distances**

  Gait stride is the distance traveled between successive footfalls of the same foot. It is the double of the *step length* (distance traveled while a foot is on the ground).

  Stride width (or *walking base*) is the side-to-side distance between the line of the two feet. Normally, it is 3–8 cm for adult but it increases with gait instability.

  The *Gait Deviation* (from normality) *Index* GDI is (Schwartz–Rosumalski, 2008) the **standardized Euclidean distance** (Chap. 17) in the 15D gait feature space between the abnormal state vector of a patient and the closest matching normal (mean of controls) state vector.

  The length of cane, when it is needed, should extend the distance from the distal wrist crease to the ground, when the person is placing arms at the sides.

  The average walking speed is 1–1.5 m/s; above 2 m/s, it is more efficient to run. Cadence for normal adults is 100–117 steps/min at preferred speed. As the body moves forward, its center of gravity moves vertically and laterally, with average displacement 5 cm and 6 cm, in a smooth sinusoidal pattern.

  Wearing high heels by women exaggerate some sex specific elements of female gait: greater pelvic rotation, increased lateral pelvic tilt, shorter strides and higher cadence. Millipedes have smooth wave-like polypedal gait. Some millipeds *Illacme* have 750 legs. No species with larger or odd number of legs are known.

  Most insects have a *tripod gait*, with front and back legs on one side moving in sync with the middle leg on the other side. But some dung beetles can gallop.

  Humans, birds and (occasionally) apes walk bipedally. Humans, birds, many lizards and (at their highest speeds) cockroaches run bipedally. But (Alexander, 2004) no animal walks or runs as we do: the trunk erect, almost straight knees at mid-stance, striking the ground with the heel alone and two-peaked force pattern in fast walking. Our walking, but not running, is relatively economical metabolically.

  A *pedometer* is a device, usually portable and electronic or electromechanical, that counts each step a person takes by detecting the motion of the person's hands or hips. Modern activity trackers, such as the Fitbit Tracker, Misfit Wearables and Razer Nabu, count the number of steps and, using it, distance walked or run.

- **Biodistances for nonmetric traits**

  In Physical Anthropology and Human Osteology (including Forensic Anthropology and Paleoanthropology), the *biodistances* (or *biological distances*) are the measures of relatedness between and within human groups, living or past, based on human cranial, skeletal or dental variation.

  Nonmetric traits are skeletal **nonmetric data** (binary, nominal or ordinal, cf. Chap. 17). The main distance statistics used to compare them between populations x and y are *Mahalanobis D^2 statistics*, i.e., square **Mahalanobis**

distance (Chap. 17) and, when the data are incomplete, the **mean measure of divergence**:

$$MMD = \frac{1}{n}\sum_{i=1}^{n}((\phi_{xi} - \phi_{yi})^2 - 4\frac{N_{xi} + N_{yi} + 1}{(2N_{xi} + 1)(2N_{yi} + 1)}).$$

Here, n is the number of traits used in the comparison, ϕ_{xi} and ϕ_{yi} are the transformed frequencies in radians of the i-th trait in the groups x, y, and N_{xi} and N_{yi} are the numbers of individuals scored for the i-th trait in the two groups.

The frequencies ϕ are obtained (in radians) from observed trait frequencies $\frac{k}{n}$ by the *Freeman–Tukey arcsine transformation*. The MMD can be negative. The *standardized MMD* (SMMD) is obtained by dividing MMD by its standard deviation.

- **Body distances in Anthropometry**

  Besides weight and circumference, the main metric (i.e., linear continuous, cf. Chap. 17) measurements in Anthropometry are between some body landmark points or planes. The main vertical distances from a standing surface are:

  - *stature* (to the top of the head);
  - *C7 level height* (to the first palpable vertebra from the hairline down, C7);
  - *acromial height* (to the *acromion*, i.e., the lateral tip of the shoulder);
  - *L5 level height* (to the first palpable vertebra from the tailbone up, L5);
  - *knee height* (to the *patella*, i.e., kneecap plane).

  The genotype gives 60 % of the phenotypic variation of human height (stature). It was about 1.63 and 1.83 m for *Neanderthal* 0.07 Ma ago and *Homo erectus* 1.8 Ma ago. The height of the average modern man ranges from 1.37 (Mbuti people of the Democratic Republic of the Congo) to 1.84 m (the Dutch). There is small (0.15–0.20) correlation between IQ and height within national populations.

  Examples of other body distances are:

  - *sitting height*: the distance from the top of the head to the sitting surface;
  - *popliteal* (or *stool*) height (seated): the distance between the underside of the foot to the underside of the thigh at the knee;
  - *hip breadth* (seated): the lateral distance at the widest part of the hips;
  - *biacromial breadth*: the distance between the acromions;
  - *buttock-knee length* : the distance from the buttocks to the patella;
  - *total foot length*: the maximum length of the right foot;

  In the thigh, there are the longest ones in the human body: bone (*femur*), muscle (*sartorius*) and nerve (*sciatic*).

- **Head and face measurement distances**

  The main linear dimensions of the cranium in Archeology are: lengths (of temporal bone, of tympanic plate, glabella-opistocranion), breadths (maximum cranial, minimum frontal, biauricular, mastoid), heights (of temporal bone, basion-bregma), thickness of the tympanic plate, and bifrontomolar-temporal distance.

[redacted]

Main viscerocranium measurements in Craniofacial Anthropometry are the *head width*, i.e., the (horizontal) maximum breadth of the head above the ears, and the *head length* (or *head depth*): the horizontal distance from the *nasion* (the top of the nose between the eyes) to the *opistocranion* (the most prominent point on the back of the head). The *cephalic index* of a skull is the percentage of width to length.

The **face length** FL: the distance between the *trichion* (midpoint on the forehead) and the *gnathion* (the lowest point of the midline of the lower jaw). It is divided by nasion and *subnasale* lines into three ("ideally", equal) parts. The *upper face height* UFH is the distance between brow and upper lip; LFH is the *lower face height*.

The **intercanthal distance** IC and **outercanthal distance** OC: the distances between inner or outer *canthi* (corners of eyes). The **face width** FW (or *bizygomatic width*) is the maximum distance between lateral surfaces of the cheeks. Let EW be the eye width and NW be the nose width (or **interalar distance**).

According to Schmid–Marx–Samal, 2007, among neoclassical canons of face, proposed by artists, the most related to attractiveness are: IC = NW = EW, FW = 4× NW and nose length = ear length = forehead height = LFH. Among aesthetically pleasing facial golden ratios $\frac{1+\sqrt{5}}{2} \approx 1.618$ are: mouth width (or lips-chin distance) to IC (or NW), ear length to NW and the *superior facial index* $\frac{UFH}{FW}$.

According to Lefevre et al., 2013, the ratios FWHR= $\frac{FW}{UFH}$ and $\frac{FW}{LFH}$ correlate with "maleness" (testosterone in mating context, aggression, status-striving etc.). On average, men have much larger faces (below the pupils), lips and chins; wider cheekbones, jaws and nostrils; and longer lower faces, but much lower eyebrows.

In Face Recognition, the sets of (vertical and horizontal) *cephalofacial dimensions*, i.e., distances between *fiducial* (standard of reference for measurement) facial points, are used. For example, the following five independent facial dimensions are derived in [Fell97] for facial gender recognition: IC, NW, FW and (vertical ones) eye-to-eyebrow distance EB and distance EM between eye midpoint and horizontal line of mouth. "Femaleness" relies on large IC, EB and small NW, FW, EM. In general, a face with larger EB is perceived as baby-like and less dominant.

Humans have the innate ability to recognize and distinguish (friend from foe) between faces from a distance. Facial attractiveness is a cultural construct found in all extant societies, and males strongly prefer neotenous facial features in females.

Pallett–Link–Lee, 2009, claim that Caucasian females with $EM \approx 36\,\%$ of FL and the **interpupillary distance** $\approx 46\,\%$ of FW, have the both, most attractive and average, faces. On the other hand, Japanese standards of beautiful eyes changed with Westernization (comparing Meiji and modern portraits): the mean ratios to *corneal diameter* (horizontal white-to-white distance) of eye height and upper lid-to-eyebrow distance are moved from 0.62 and 2.21 to 0.82 and 1.36.

Among the symmetry pairs, main contributors to attractiveness are the nose, mouth and upper tips of the lips. Comparing 3D facial scans with their mirror images, Djordjevic et al., 2011, found that on average, males and females have 53.5 % and 58.5 % symmetry of the whole face. Cf. **distances from symmetry** in Chap. 21. Alare and pogonio were the most and the least symmetric landmark.

Modifying traditional canons of Facial Plastic Surgery (based on horizontal and vertical planes in 2D), Young, 2008, asserts that the iris, nasal tip and lower lip are the most prominent structures within the eye, nose and mouth. All distances which he proposed as elements of facial beauty are multiples of the diameter of the iris.

- **Gender-related body distance measures**

The main gender-specific body configuration features are:

for females, WHR (*waist-to-hip ratio*), LBR (*leg-to-body ratio*) and BMI (*body mass index*), i.e., the ratio of the weight in kg and squared height in m^2;

for males, height, SHR (*shoulder-to-hip ratio*) and WCR (*waist-to-chest ratio*);

androgen equation (three times the shoulder width minus one times the pelvic width) and HGS (hand grip strength), which are higher for males;

right *second-to-fourth digit* (index to ring finger) *ratio 2D-4D*, which is lower (as well as prenatal testosterone is higher) for males in the same population;

anogenital distance (cf. **distances in Medicine**), which is larger for males;

person's *center of mass* (slightly below the belly button), higher for males.

The *angle of lumbar curvature* θ is the angle between the thoracic back and buttocks. Its mean value is 43.25° for men, 47.19° for women and 45.5° for reproductively viable women. This wedging in third-to-last lumbar vertebra shifts the center of mass of pregnant woman back over hips, reducing hip torque by over 90 %. But excessive θ (hyperlordosis) also lead to muscular fatigue and lower back pain; the optimal θ is 45.5°. Men prefer the women for whom the buttocks protrusion θ is close to optimal and attributable to vertebral wedging, not buttock mass.

The main predictor for developmental instability, increasing with age, is FA (*fluctuating asymmetry*), i.e., the degree to which the size of bilateral body parts deviates from the population mean, aggregated across several traits. Women (men) prefer the odors, faces and voices of men (women) with lower FA. Old men (but not women) with lower facial FA have better cognitive ability and reaction time.

BMI and WHR indicate the percentage of body fat and fat distribution, respectively; they are used in medicine to assess risk factors. A WHR of 0.7 for women and 0.9 for men correlates with general health and fertility. As a cue to female body attractiveness for men, the ideal WHR varies from 0.6 in China to 0.85 in Africa. In general, men tend to prefer BMI 17–20 in women; it correspond, usually, to women of 18–20 yers old.

In Fan et al., 2005, the main visual cue to male body attractiveness is VHI (*volume-to-height index*), i.e., the ratio of the volume in liters and squared height in m^2. Mautz et al., 2013, claim that women prefer taller men with higher SHR

[redacted]

and FPL (flaccid penis length), but attractiveness increased quickly until FPL reached 7.6 cm and then began to slow down. Stulp et al., 2013, found that on average among speed-daters, women choose 25 cm taller men, while men choose only 7 cm shorter women, resulting in suboptimal (19 cm) pair formation.

In terms of somatotype, women prefer mesomorphs (muscular men) followed by ectomorphs (lean men) and endomorphs (heavily-set men).

In terms of BWH (bust-waist-hips), the average Playboy centrifold 1955–1968 has $(90.8, 58.6, 89.3)$ cm, close to the ideal hourglass figures $(90, 60, 90)$ cm and $(36, 24, 36)$ inch. But dietitians advocate waistline 80–85 cm and at most half-height.

In conversation, women are better at detecting mismatch between meaning and *prosody* (intonation and rhythm of speech), but worse at vocabulary's variety. Men's vocal cords are larger and their vocal tracts are longer than women's; so, they speak about an octave lower. In English, women use less nonstandard forms and often use different color terms and descriptive phrases from men. Pirahã (Amazon's tribe) men use larger articulatory space and, say, only men use "s".

Used as obesity indices, WC, ICO$=$WC/height and (proposed by Krakauer–Krakauer, 2012) ABSI$=$WC/(BMI$^{\frac{2}{3}}$height$^{\frac{1}{2}}$) are better predictors of mortality than BMI.

- **Sagittal abdominal diameter**

  Sagittal abdominal diameter (SAD) is the distance between the back and the highest point of the abdomen, measured while standing. It is a measure of visceral obesity. Normally, SAD should be under 25 cm. SAD$>$ 30 cm correlates to insulin resistance and increased risk of cardiovascular and Alzheimer's diseases.

  A related measurement is SAH, the abdominal height as measured in the supine position. *Inter-recti distance* (IRD) is the width of the *linea alba* (a fibrous structure that runs down the midline of the abdomen).

- **Body distances for clothes**

  Humans lost body hair around 1 Ma ago and began wearing clothes $\approx$ 0.17 Ma ago.

  The European standard EN 13402 "Size designation of clothes" defined, in part EN 13402-1, a standard list of 13 body dimensions (measured in cm) together with a method for measuring each one on a person. These are: body mass, height, foot length, arm length, inside leg length, and girth for head, neck, chest, bust, under-bust, waist, hip, hand. Examples of these definitions follow.

  Foot length: horizontal distance between perpendiculars in contact with the end of the most prominent ones, toe and part of the heel, measured with the subject standing barefoot and the weight of the body equally distributed on both legs.

  Arm length: distance from the armscye/shoulder line intersection (acromion), over the elbow, to the far end of the prominent wrist bone (ulna), with the subject's right fist clenched and placed on the hip, and with the arm bent at 90°.

  Inside leg length: distance between the crotch and the soles of the feet, measured in a straight vertical line with the subject erect, feet slightly apart, and the weight of the body equally distributed on both feet.

For clothes where a larger step size is sufficient to select the right product, the standard also defines a letter code: XXS, XS, S, M, L, XL, XXL, 3XL, 4XL or 5XL. This code represents the bust girth for women and the chest girth for men.

Vanity sizing (or *size deflation*) is the marketing phenomenon of ready-to-wear clothing of the same nominal size becoming bigger in physical size over time.

- **Distance handling**

  Distance handling refers to the training of gun dogs (to assist hunters in finding and retrieving game) or sport dogs (for canine agility courses) where a dog should be able to work away from the handler.

  In agility training, the *lateral distance* is the distance that the dog maintains parallel to the handler, and the *send distance* is the distance that the dog can be sent straight away from the handler.

- **Racing distances**

  In Racing, **length** is an informal unit of distance to measure the distance between competitors; for example, in boat-racing it is the average length of a boat.

  The **horse-racing distances** and winning margins are measured in terms of the **horses** (or **lengths**, i.e., lengths of a horse, $\approx$ 8 feet (2.44 m), ranging from half the length to the **distance**, i.e., more than 20 lengths. The *length* is often interpreted as a unit of time equal to $\frac{1}{5}$ second. Smaller margins are: *short-head*, *head*, or *neck*. A *distance flag* is a flag held at a distance pole in a racecourse.

  The distances a horse travels without stops (15–25 km) and it travels in a day (40–50 km) or hour (6 km) were used as Tatar and Persian units of length.

- **Triathlon race distances**

  The **Ironman distance** (or **Ultra distance**) started in Hawaii, 1978, is a 3.86 km swim followed by a 180 km bike and a 42.195 km (*marathon distance*) run.

  The international **Olympic distance** started in Sydney, 2000, is 1.5 km (*metric mile*), 40 km and 10 km of swim, cycle and run, respectively.

  Next to it are the **Sprint distance** 0.75, 20, 5 km, the **Long Course** (or *Half Ironman*) 1.9, 90, 21.1 km and the *ITU long distance* 3, 80, 20 km.

- **Running distances**

  In Running, usually, *sprinting* is divided into 100, 200, 400 m, *middle distance* into 800, 1500, 3000 m and *long distance* into 5, 10 km.

  LSD (long slow distance) is a is a form of aerobic endurance training in running and cycling, in which distances longer, than those of races, are covered, but at a slower pace.

  Fartlek (or *speed play*) is an approach to distance-running training involving variations of pace and aimed at enhancing the psychological aspects of conditioning.

  Race-walking is divided into 10, 20, 50 km, and *relay races* into $4 \times 100, 4 \times 200, 4 \times 300, 4 \times 400$ m. *Distance medley relay* is made up of 1200, 400, 800, 1600 m legs.

[redacted]

Besides track running, runners can compete on a measured course, over an established road (*road running*), or over open or rough terrain (*cross-country running*).

Roughly, 4 units of running distance are time-equivalent to 1 unit of swimming distance. Also, one have to walk about twice the distance to burn the same amount of calories as running it. Running workout times should be multiplied by 3.5 when aiming for a similar training effect from cycling. A multiple of 0.75–1 should be used for an indoor rowing-to-running ratio.

- **Distance swimming**

  Distance swimming is any swimming race $>$ 1.5 km; usually, within 24–59 km.

  DPS (distance per swim stroke) is a metric of swimming efficiency used in training. In Rowing, *run* is the distance the boat moves after a stroke.

- **Distance jumping**

  The four Olympic jumping events are: *long jump* (to leap horizontally as far as possible), *triple jump* (the same but in a series of three jumps), *high jump* (to reach the highest vertical distance over a horizontal bar), and *pole vault* (the same but using a long, flexible pole).

  The world's records, as in 2013: 8.95 , 18.29 , 2.45, and 6.14 m, respectively.

- **Distance throwing**

  The four Olympic throwing events are: shot put, discus, hammer, and javelin.

  The world's records, as in 2013: 23.12 , 74.08 , 86.74 m, and 98.48 m, respectively.

  As in 2013, the longest throws of an object without any velocity-aiding feature are 427.2 m with a boomerang and 406.3 m with a flying ring *Aerobie*.

  Distance casting is the sport of throwing a fishing line with an attached sinker (usually, on land) as far as possible.

  Darts is a sport and a pub game in which darts are thrown at a *dartboard* (circular target) fixed to a wall so that the bullseye is 172.72 cm from the floor. The *oche* (line behind which the throwing player must stand) is 236.86 cm from the dartboard.

- **Archery target distances**

  FITA (Federation of International Target Archery, organizing world championships) **target distances** are 90, 70, 50, 30 m for men and 70, 60, 50, 30 m for women, with 36 arrows shot at each distance. Farthest accurate shot is 200 m.

  In Archery, *brace height* (or *fistmele*) is the distance from the string, when the bow is strung, to the pivot point of the bow's grip, i.e., the handle of a bow.

- **Bat-and-ball game distances**

  The best known bat-and-ball games are bowling (cricket) and baseball. In cricket, the field position of a player is named roughly according to its polar coordinates: one word (*leg, cover, mid-wicket*) specifies the angle from the batsman, and this word is preceded by an adjective describing the distance from the batsman. The *length* of a delivery is how far down the *pitch* (central strip of the cricket field) towards the batsman the ball bounces.

This distance is called *deep* (or *long*), *short* and *silly distance* if it is, respectively, farther away, closer and very close to the batsman. The distance further or closer to an extension of an imaginary line along the middle of the pitch bisecting the stumps, is called *wide* or *fine distance*, respectively.

In baseball, a *pitch* is the act of throwing a baseball toward home plate to start a play. The standard professional *pitching distance*, i.e., the distance between the front (near) side of the pitching rubber, where a pitcher start his delivery, and home plate is 60 feet 6 inches ($\approx$ 18.4 m). The distance between bases is 90 feet.

- **Three-point shot distance**

In basketball, the *three-point line* is an arc at a set radius, called **three-point shot distance**, from the basket. A field goal made from beyond this line is worth three points. In international basketball, this distance is 6.25 m.

Goals in *indoor soccer* are worth 1, 2 or 3 points depending upon distance.

- **Football distances**

In *association football* (or *soccer*), the average distance covered by a player in a men's professional game is 9–10 km. It consists of about 36 % jogging, 24 % walking, 20 % cruising submaximally, 11 % sprinting, 7 % moving backwards and 2 % moving in possession of the ball. The ratio of low- to high-intensity exercise is about 2.2:1 in terms of distance, and 7:1 in terms of time.

In *American football*, one *yard* means usual yard (0.9144 m) of the distance in the direction of one of the two goals. A field is 120 yards long by 53.3 yards wide. A team possessing the ball should advance at least the *distance* (10 yards) in order to get a new set of (4 or 3) *downs*, i.e., periods from the time the ball is put into play to the time the play is whistled over by the officials. *Yardage* is the amount of yards gained or lost during a play, game, season, or career.

- **Golf distances**

In golf, *carry* and *run* are the distances the ball travels in the air and once it lands.

The golfer chooses a golf club, grip, and stroke appropriate to the distance. The *drive* is the first shot of each hole made from the area of *tees* (peg markers) to long distances. The *approach* is used in long- to mid-distance shots.

The *chip* and *putt* are used for short-distance shots around and, respectively, on or near the green. The maximum distance a typical golfer can hit a ball with a particular club is the club's *hitting distance*.

A typical *par* (standard score) 3, 4, 5 holes measure 229, 230–430, $\geq$431 m. The greatest recorded drive distance, carry, shot with one hand are 471, 419, 257 m.

Some manufacturers stress the large range of a device in the product name, say, *Ultimate Distance* golf balls (or softball bates, spinning reels, etc.).

- **Fencing distances**

In combative sports and arts, *distancing* is the appropriate selection of the distance between oneself and a combatant throughout an encounter.

For example, in fencing, the *distance* is the space separating two fencers, while the distance between them is the *fencing measure*.

[redacted]

A *lunge* is a long step forward with the front foot. A *backward spring* is a leap backwards, out of distance, from the lunge position.

The following five distances are distinguished: *open distance* (farther than advance-lunge distance), *advance-lunge distance*, *lunging distance*, *thrusting distance* and *close quarters* (closer than thrusting distance).

In Japanese martial arts, **maai** is the *engagement distance*, i.e., the exact position from which one opponent can strike the other, after factoring in the time it will take to cross their distance, angle and rhythm of attack. In kendo, there are three maai distances: *to-ma* (long distance), *chika-ma* (short distance) and, in between, *itto-ma* $\approx$ 2 m, from which only one step is needed in order to strike.

- **Distance in boxing**

  The distance is boxing slang for a match that lasts the maximum number (10 or 12) of scheduled rounds. The longest boxing match (with gloves) was on April 6–7, 1893, in New Orleans, US: Bowen and Burke fought 110 rounds for 7.3 h.

- **Soaring distances**

  Soaring is an air sport in which pilots fly unpowered aircraft called *gliders* (or *sailplanes*) using currents of rising air in the atmosphere to remain airborne.

  The *Silver Distance* is a 50 km unassisted straight line flight. The *Gold* and *Diamond Distance* are cross-country flights of 300 km and over 500 km, respectively.

  Possible courses—*Straight*, *Out-and-Return*, *Triangle* and 3 *Turnpoints Distance*—correspond to 0, 1, 2 and 3 turnpoints, respectively.

  Using open class gliders, the world records in free distance, in absolute altitude and in gain of height are: 3008.8 km (by Olhmann and Rabeder, 2003), 15,460 m (by Fossett and Enevoldson, 2006) and 12,894 m (by Bikle, 1961). The distance record with a paraglider is 501.1 km (by Hulliet, 2008).

  Eustace jumped in 2014 from a balloon at 41,419 m, setting world records in exit altitude and total free fall distance. But, since Eustace used a stabilizing drogue parachute while Baumgartner, 2012, did not, their free fall distance and vertical speed records are in different categories. A stewardess Vesna Vulović survived in 1972 a fall of 10 km, when JAT Flight 367 was brought down by explosives.

- **Aviation distance records**

  Absolute general aviation world records in flight distance without refueling and in altitude are: 41,467.5 km by Fossett, 2006, and 37,650 m by Fedotov, 1977.

  Distance and altitude records for free manned balloons are, respectively: 40,814 km (by Piccard and Jones, 1999) and 41,425 m (by Eustace, 2014).

  The general flight altitude record is 112,010 m by Binnie, 2004, on a rocket plane.

  The longest (13,804 km during 16 h 55 min) nonstop scheduled passenger route in 2015 was Qantas Flight 8 from Dallas to Sydney.

  The *Sikorsky prize* (US $250,000) will be awarded for the first flight of a human-powered helicopter which will reach an altitude of 3 m, stay airborne for at least 1 min remaining within $10\mathrm{m}\times 10\mathrm{m}$ square. In 2012, a craft (32.2 kg) by a team at the University of Maryland flew 50 s at 61 cm up.

- **Amazing greatest distances**

  Examples of such distances among Guinness world records are the greatest distances:

  - being fired from a cannon (59 m),
  - walked unsupported on tightrope (130 m),
  - run on a static cycle in 1 min (2.04 km),
  - moon-walked (as Michael Jackson) in 1 h (5.125 km),
  - covered three-legged (the left leg of one runner strapped to the right leg of another runner) in 24 h (33 km),
  - jumped with a pogo stick (37.18 km),
  - walked with a milk bottle balanced on the head (130.3 km),
  - covered by a car driven on its side on two wheels (371.06 km),
  - hitchhiked with a fridge (1650 km).

  Amazing race *The* 2904 is to drive the 2904 miles from New York City to San Francisco for $2904 including the vehicle, fuel, food, tolls, repairs and tickets.

- **Isometric muscle action**

  An **isometric muscle action** refers to exerting muscle strength and tension without producing an actual movement or a change in muscle length.

  Isometric action training is used mainly by weightlifters and bodybuilders. Examples of such *isometric exercises*: holding a weight at a certain position in the range of motion and pushing or pulling against an immovable external resistance.

29.2 Equipment Distances

- **Motor vehicle distances**

  The **safe following distance**: the reglementary distance from the vehicle ahead of the driver. For reglementary perception-reaction time of at least 2 s (the *two-second rule*), this distance (in m) should be $0.56 \times v$, where v is the speed (in km/h). Sometimes the *three-second rule* is applied. The stricter rules are used for heavy vehicles (say, at least 50 m) and in tunnels (say, at least 150 m).

  The **perception-reaction distance** (or **thinking distance**): the distance a vehicle travels from the moment the driver sees a hazard until he applies the brakes (corresponding to human perception time plus reaction time). Physiologically, it takes 1.3–1.5 s, and the brake action begins 0.5 s after application.

  The **braking distance**: the distance a motor vehicle travels from the moment the brakes are applied until the vehicle completely stops.

  The (total) **stopping distance**: the distance a motor vehicle travels from where the driver perceives the need to stop to the actual stopping point (corresponding to the vehicle reaction time plus the vehicle braking capability).

[redacted]

The **crash distance**: (or *crushable length*): the distance between the driver and the front end of a vehicle in a frontal impact (or, say, between the pilot and the first part of an airplane to impact the ground).

The **skidding distance** (or *length of the skid mark*): the distance a motor vehicle *skidded*, i.e., slid on the surface of the road (from the moment of the accident, when a wheel stops rolling) leaving a rubber mark on the road.

The **cab-to-frame** (or **cab-to-end**, *CF*, *CE*): the distance from back of a truck's cab to the end of its frame.

The **distance to empty** (or *DTE*) displays the estimated distance the vehicle can travel before it runs out of fuel. The warning lamp start blinking at 80 km.

The **acceleration-deceleration distance** of a vehicle, say, a car or aircraft, is (Drezner–Drezner–Vesolowsky, 2009) the cruising speed v times the travel time. For a large origin-destination distances d, it is $d + \frac{v^2}{2}(\frac{1}{a} + \frac{1}{b})$, where a is the acceleration at the beginning and $-b$ is the deceleration at the end.

- **Aircraft distances**

The maximum distance the aircraft can fly without refueling is called the **maximum range** if it fly with its maximum cargo weight and the **ferry range** if it fly with minimum equipment.

For a warplane, the *combat range* is the maximum distance it can fly when carrying ordnance, and the *combat radius* is a the maximum distance it can travel from its base, accomplish some objective, and return with minimal reserves.

The FAA **lowest safe altitude**: 1000 feet (305 m) above the highest obstacle within a horizontal distance of 2000 feet.

A **ceiling** is the maximum *density altitude* (height measured in terms of air density) an aircraft can reach under a set of conditions.

A *flight level* (FL) is specific barometric pressure, expressed as a nominal altitude in hundreds of feet, assuming standard sea-level pressure datum of 1013.25 hPa.

The *transition altitude* is the altitude above sea level at which aircraft change from the use of altitude to the use of FL's; in the US and Canada, it is 5500 m).

The **gust-gradient distance**: the horizontal distance along an aircraft flight path from the edge of the *gust* (sudden, brief increase in the speed of the wind) to the point at which the gust reaches its maximum speed.

The **distance-of-turn anticipation**: the distance, measured parallel to the anticipated course and from the earliest position at which the turn will begin, to the point of route change.

The **landing distance available** (LDA): the length of runway which is declared available and suitable for the ground run of an airplane landing. The **landing roll**: the distance from the point of touchdown to the point where the aircraft can be brought to a stop or exit the runway. The **actual landing distance** (ALD): the distance used in landing and braking to a complete stop (on a dry runway) after crossing the runway threshold at 50 feet (15.24 m); it can be affected by various operational factors. The FAA **required landing distance** (used for dispatch purposes): a factor of 1.67 of ALD for a dry runway and 1.92 for a wet runway.

The **takeoff run available** (TORA): the *runway distance* (length of runway) declared suitable for the ground run of an airplane takeoff. The **takeoff distance available** (TODA): TORA plus the length of the clearway, if provided. The **emergency distance** (ED or *accelerate-stop distance*): the runway plus *stopway length* (able to support the airplane during an aborted takeoff) declared suitable for the acceleration and deceleration of an airplane aborting a takeoff.

The **arm's distance**: the horizontal distance that an item of equipment is located from the *datum* (imaginary vertical plane, from which all horizontal measurements are taken for balance purposes, with the aircraft in level flight attitude).

In the parachute deployment process, the **parachute opening distance** is the distance the parachute system dropped from pulling to full inflation of the canopy, while the **inflation distance** is measured from *line stretch* (when the suspension lines are fully extended) to full inflation.

Wing's aspect ratio (of an aircraft or bird) is the ratio $AR = \frac{b^2}{S}$ of the square of its span to the area of its planform. If the length of the *chord* (straight line joining the leading and trailing edges of an airfoil) is constant, then AR is length-to-breadth **aspect ratio**; cf. Chap. 1. A better measure of the aerodinamic efficiency is the *wetted aspect ratio* $\frac{b^2}{S_w}$, where S_w is the entire surface area exposed to airflow.

- **Ship distances**

Endurance distance: the total distance that a ship or ground vehicle can be self-propelled at any specified endurance speed.

Distance made good: the distance traveled by the boat after correction for current, *leeway* (the sideways movement of the boat away from the wind) and other errors that may be missed in the original distance measurement.

Log: a device to measure the distance traveled which is further corrected to a distance made good. Hitherto, sea distances were measured in units of a day's sail.

Leg (nautical): the distance traveled by a sailing vessel on a single tack.

Berth: a safety margin to be kept from another vessel or from an obstruction.

Length overall (LOA): the maximum length of a vessels's hull along the waterline.

Length between perpendiculars (LPP): the length of a vessel along the waterline from the main bow perpendicular member to the main stern perpendicular member.

Freeboard: the height of a ship's hull above the waterline. *Draft* (or *draught*): the vertical distance between the waterline and the keel (bottom of the hull).

GM-distance (or *metacyclic height*) of a ship: the distance between its center of gravity G and the *metacenter*, i.e., the projection of the *center of buoyancy* (the center of gravity of the volume of water which the hull displaces) on the centerline of the ship as it heels. This distance, 1–2 m, determines ship's stability.

[redacted]

Distance line (in Diving): a marker (say, 50 m of thin polypropylene line) of the shortest route between two points. It is used, as Ariadne's thread, to navigate back to the start in conditions of low visibility, water currents or *penetration diving* into a space (cave, wreck, ice) without vertical ascent back.

- **Distance-to-fault**

  In Cabling, DTF (**distance-to-fault**) is a test using time or frequency domain reflectometers to locate a fault, i.e., discontinuity caused by, say, a damaged cable, water ingress or improperly installed/mated connectors.

  The amount of time a pulse (output by the tester into the cable) takes for the signal (reflected by a discontinuity) to return can be converted to distance along the line and provides an approximate location of the reflection point.

  Protective *distance relays* respond to the voltage and current. The *impedance* (their ratio) per km being constant, these relays respond to the relay-fault distance.

- **Distances in Forestry**

  In Forestry, the **diameter at breast height** (d.b.h.) is a standard measurement of a standing tree's diameter taken at 4.5 feet ($\approx$ 1.37 m) above the ground. The *diameter at ground line* (d.g.l.) is the diameter at the estimated cutting height. The *diameter outside bark* (d.o.b.) is a measurement in which the thickness of the bark is included, and d.i.b. is a measurement in which it is excluded.

  The *crown height* is the vertical distance of a tree from ground level to the lowest live branch of the crown. The *merchantable height* is the point on a tree to which it is salable. A log is a length of tree suitable for processing into a wood product.

  Optimum road spacing is the distance between parallel roads that gives the lowest combined cost of *skidding* (log dragging) and road construction costs per unit of log volume. The *skid distance* is the distance logs are dragged.

  A *yarder* is a piece of equipment used to lift and pull logs by cable from the felling site to a landing area or to the road's side. The **yarding distance** is the distance from which the yarder takes logs. The **average yarding distance** is the total yarding distance for all turns divided by the total number of turns.

  A *spar tree* is a tree used as the highest anchor point in a cable logging setup. A *skyline* is a cableway stretched between two spar trees and used as a track for a log carriage. The distance spanned by a skyline is called its *reach*.

  Understory is the area of a forest which grows at the lowest height level between the forest *floor* and the *canopy* (layer formed by mature tree crowns and including other organisms). Perhaps, a half of all life on Earth could be found in canopy. The *emergent layer* contains a small number of trees which grow above the canopy.

- **Distance in Military**

  In the Military, the term **distance** usually has one the following meanings:

  the space between adjacent individual ships or boats measured in any direction between foremasts;

  the space between adjacent men, animals, vehicles, or units in a formation measured from front to rear;

the space between known reference points or a ground observer and a target, measured in m (artillery), or in units specified by the observer. This distance along an imaginary straight line from the spotter is called *observer-target distance*.

In amphibious operations, the *distant retirement area* is the sea area located to seaward of the landing area, and the *distant support area* is the area located in the vicinity of the landing area but at considerable distance seaward of it.

Strategic depth refers to the distances between the front lines and the combatants' industrial and population core areas.

In military service, a **bad distance** of the troop means a temporary intention to extract itself from war service. This passing was usually heavily punished and equated with that of *desertion* (an intention to extract itself durably).

In US military slang, BFE (Big Fucking Empty) is an extremely distant or isolated deployment or location; used mostly about the disgust at the distance or remoteness. Also, a *klick* means a distance of 1 km.

- **Interline distance**

  In Engineering, the **interline distance** is the minimum distance permitted between any two buildings within an explosives operating line, in order to protect buildings from propagation of explosions due to the blast effect.

- **Scaled distance**

  The **scaled distance** (SD) is the parameter used to measure the level of vibration from a blast, when effects of the frequency characteristics are discounted.

  The minimum safe distance from a blast to a monitoring location is $SD \times \sqrt{W}$, where W denotes the maximum per delay (instantaneous) charge weight.

- **Standoff distance**

  The **standoff distance** is the distance of an object from the source of an explosion (in Warfare), or from the delivery point of a laser beam (in laser material processing). Also, in Mechanics and Electronics, it is the distance separating one part from another; for example, for insulating (cf. **clearance distance**), or the distance from a noncontact length gauge to a measured material surface.

- **Buffer distance**

  In Nuclear Warfare, the **horizontal buffer distance** is the distance which should be added to the radius of safety in order to be sure that the specified degree of risk will not be exceeded. The **vertical buffer distance** is the distance which should be added to the fallout safe-height of a burst, in order to determine a desired height of burst so that militari significant fallout will not occur.

  The term *buffer distance* is also used more generally as, for example, the buffer distance required between sister stores or from a high-voltage line.

  Cf. **clearance distance** and, in Chap. 25, **setback distance**.

- **Offset distance**

  In Nuclear Warfare, the **offset distance** is the distance the desired (or actual) ground zero is offset from the center of the area (or point) target.

[redacted]

In Computation, the *offset* is the distance from the beginning of a string to the end of the segment on that string. For a vehicle, the **offset** of a wheel is the distance from its hub mounting surface to the centerline of the wheel.

The term *offset* is also used for the **displacement** vector (Chap. 24) specifying the position of a point or particle in reference to an origin or to a previous position.

- **Range of ballistic missile**

  Main **ranges of ballistic missiles** are *short* (at most 1000 km), *medium* (1000–3500 km), *long* (3500–5500 km) and *intercontinental* (at least 5500 km).

  Tactical and *theatre* ballistic missiles have ranges 150–300 and 300–3500 km.
- **Proximity fuse**

  The **proximity fuse** is a fuse that is designed to detonate an explosive automatically when close enough to the target.
- **Sensor network distances**

  The **stealth distance** (or *first contact distance*): the distance traveled by a moving object (or intruder) until detection by an active sensor of the network (cf. **contact quasi-distances** in Chap. 19); the *stealth time* is the corresponding time.

  The **first sink contact distance**: the distance traveled by a moving object (or intruder) until the monitoring entity can be notified via a sensor network.

  The **miss distance**: the distance between the lines of sight representing estimates from two sensor sites to the target (cf. the **line-line distance** in Chap. 4).

  The **sensor tolerance distance**: a **range distance** within which a localization error is acceptable to the application (cf. the **tolerance distance** in Chap. 25).

  The actual distances between some pairs of sensors can be estimated by the time needed for a two-way communication. The positions of sensors in space can be deduced (cf. **Distance Geometry Problem** in Chap. 15) from those distances.
- **Proximity sensors**

  Proximity (or *distance*) **sensors** are varieties of ultrasonic, laser, photoelectric and fiber optic sensors designed to measure the distance from itself to a target. For such laser range-finders, a special *distance filter* removes measurements which are shorter than expected, and which are therefore caused by an unmodeled object. The *blanking distance* is the minimum range of an ultrasonic proximity sensor.

  The *detection distance* is the distance from the detecting surface of a sensor head to the point where a target approaching it is first detected. The *maximum operating distance* is its maximum detection distance from a standard modeled target, disregarding accuracy. The *stable detection range* is the detectable distance range in which a standard detected object can be stably detected with respect to variations in the operating ambient temperature and power supply.

  The *resolution* is the smallest change in distance that a sensor can detect. The *span* is the working distance between measurement range endpoints over which the sensor will reliably measure displacement. The *target standoff* is the distance from the face of the sensor to the middle of the span.

Distance constant of a metereological sensor is the length of fluid flow past required to cause it to respond to 63.2 % (i.e., $1 - \frac{1}{e}$) of a step change in speed.

- **Precise distance measurement**

  The resolution of a TEM (transmission electronic microscope) is about 0.2 nm (2×10^{-10} m). This resolution is 1000 times greater than a light microscope and about 500,000 times greater than that of a human eye which is 576 mega pixel. However, only nanoparticles can fit in the vision field of an electronic microscope.

  The methods, based on measuring the wavelength of laser light, are used to measure macroscopic distances nontreatable by an electronic microscope. But the uncertainty of such methods is at least the wavelength of light, say, 633 nm.

  The recent adaptation of *Fabry–Perot metrology* (measuring the frequency of light stored between two highly reflective mirrors) to laser light permits the measuring of relatively long (up to 5 cm) distances with an uncertainty of only 0.01 nm.

  The main devices used for low accuracy distance measurement are the rulers, engineer's scales, calipers and surveyor's wheels.

- **Laser distance measurement**

  Lasers measure distances without physical contact. They allow for the most sensitive and precise length measurements, for extremely fast recording and for the largest measurement ranges. The main techniques used are as follows.

  Triangulation (cf. **laterations**) is useful for distances from 1 mm to many km. *Pulse measurements*, used for large distances, measure the time of flight of a laser pulse from the device to some target and back. The *phase shift method* uses an intensity-modulated laser beam. *Frequency modulation methods* involve frequency-modulated laser beams. *Interferometers* allow for distance measurements with an accuracy which is far better than the wavelength of the light used.

  The main advantage of laser distance measurement is that laser light has a very small wavelength, allowing one to send out a much more concentrated probe beam and thus to achieve a higher transverse spatial resolution.

- **Radio distance measurement**

  DME **distance measuring equipment**) is an air navigation technology that measures distances by timing the propagation delay of UHF signals to a *transponder* (a receiver-transmitter that will generate a reply signal upon proper interrogation) and back. DME will be phased out by global satellite-based systems: GPS (US), GLONASS (Russia), BeiDou (China) and Galileo (EU).

  The GPS (Global Positioning System) is a radio navigation system which permits one to get her/his position on the globe with accuracy of 10 m. It consists of 32 satellites and a monitoring system operated by the US Department of Defense. The nonmilitary part of GPS can be used by the purchase of an adequate receiver.

  The **GPS pseudo-distance** (or *pseudo-range*) is an approximation (since the receiver clock is not so perfect as the clock of a satellite) of the distance between

[redacted]

a satellite and a GPS receiver by the travel time of a satellite time signal to a receiver multiplied by the propagation time of the radio signal.

The receiver uses **trilatelation** in order to calculate its position (latitude, longitude, altitude) and speed by solving a system of equations using its pseudo-distances from 4 to 7 satellites and their positions. Cf. **radio distances** in Chap. 25.

- **Laterations**

  Lateration (or *ranging*) is the determination of the distance from one location or position to another one. Usually, the term *ranging* is used for moving objects, while *surveying* is used for static terrestrial objects. Active ranging systems operate with unilateral transmission and passive reflections, such as SONAR (SOund Navigation And Ranging), RADAR (RAdio Detection) and LIDAR (Light Detection).

  A *rangefinder* is a device for measuring distance from the observer to a target. Among applications are surveying, navigation, ballistics and photography.

  Triangulation is the process of locating a point P as the third point of a triangle with one known side (say, $[A, B]$ of length l) and two known angles (say, $\angle PAB = \alpha$ and $\angle PBA = \beta$). In $\mathbb{R}^2$, the perpendicular distance between P (say, a ship) and $[A, B]$ (say, a shore) is $\frac{l \sin\alpha \sin\beta}{\sin(\alpha+\beta)}$. Cf. point-line distance in Chap. 4.

  Technically more complicated, **trilateration** is locating a (possibly, moving) object P, using only its distances to known locations A_1, A_2 and A_3 (for example, to stations, beacons or satellites), as the overlap of $2D$ or $3D$ spheres, centered on them and having radii $d(P, A_1), d(P, A_2), d(P, A_3)$, respectively. Using additional stations, as in GPS, permits double-checking of the measurements. Cf. the **metric basis** in Chap. 1.

  More accurate generally, **multilateration** is locating a moving object P, using only two pairs (A_1, A_2), (B_1, B_2) of known locations, as the intersection of two curves defined by the relative distances $d(P, A_1) - d(P, A_2)$ and $d(P, B_1) - d(P, B_2)$, respectively.

- **Transmission distance**

  The **transmission distance** is a **range distance**: for a given signal transmission system (fiber optic cable, wireless, etc.), it is the maximal distance the system can support within an acceptable path loss level.

  For a given network of contact that can transmit an infection (or, say, an idea with the belief system considered as the immune system), the **transmission distance** is the path metric of the graph, in which edges correspond to events of infection and vertices are infected individuals. Cf. **forward quasi-distance** in Chap. 22.

- **Delay distance**

  The **delay distance** is a general term for the distance resulting from a given delay. For example, in a meteorological sensor, the *delay distance* is the length of a column of air passing a wind vane, such that the vane will respond to 50 % of a sudden angular change in wind direction. When the energy of a neutron is measured by the delay (say, t) between its creation and detection, the *delay distance* is $vt - D$, where v is its velocity and D is the source-detector distance.

In evaluations of visuospatial working memory (when the subjects saw a dot, following a 10-, 20-, or 30-s delay, and then drew it on a blank sheet of paper), the *delay distance* is the distance between the stimulus and the drawn dot.

- **Master-slave distance**

  Given a design (say, remote manipulation, surveillance, or data transmission system) in which one device (the *master*) fully controls one or more other devices (the *slaves*), the **master-slave distance** is a measure of distance between the master and slave devices. Cf. also Sect. 18.2.

- **Flow distance**

  In a manufacturing system, a group of machines for processing a set of jobs is often located in a serial line along a path of a transporter system.

  The **flow distance** from machine i to machine j is the total flow of jobs from i to j times the physical distance between machines i and j.

- **Single row facility layout**

  The **SRFLP** (or **single row facility layout problem**) is the problem of arranging (finding a permutation of) n departments (disjoint intervals) with given lengths l_i on a straight line so as to minimize the total weighted distance $\sum_{i=1}^{n-1}\sum_{i=j+1}^{n} w_{ij}d_{ij}$ between all department pairs. Here w_{ij} is the average daily traffic between two departments i and j, and d_{ij} is their **centroid distance**.

  Among applications of SRFLP, there are arranging machines in a manufacturing system, rooms on a corridor and books on a shelf.

- **Distance hart**

  In Technical Drawing, the **distance hart** means the distance from the center (the heart) of an object, as, for example, the distance hart of the toilet seat to the wall.

  The **center-to-center distance** (or **on-center**, *O.C.*) is the distance between the centers of two adjacent members (say, columns or pillars). Cf. **centroid linkage**, **centroid distance** in Chaps. 17, 19 and *center* **gear distance**.

- **Push distance**

  Precise machining of bearing rings should be preceded by part centering. In such a centering system, the **push distance** is the distance the slide must move towards the part in order to push it from its off-center position to the center of rotation.

- **Engine compression distance**

  Piston motors convert the compressed air energy to mechanical work through motion. **Engine compression distance** (or *compression height*) is the distance from the centerline of the wrist pin to the top deck of the piston.

- **Shift distance**

  A *penetrometer* is a device to test the strength of a material, say, soil. The penetrometer (usually cone-shaped) is pressed against material and the depth of the resulting hole is measured. The **shift distance** (or *friction-bearing offset*) is the distance between the cone's base and the mid-height of the friction sleeve above it.

[redacted]

- **Throat distance**

  The *swing* (size) of a drill/boring press is twice the **throat distance**, the distance from the center of the spindle to the column's edge.
- **Collar distance**

  In Mining, the **collar distance** is the distance from the top of the powder column to the collar of the blasthole, usually filled with stemming.
- **Quenching distance**

  Quenching is the rapid cooling of a workpiece; the **quenching distance** is the diameter of smallest hole a flame can travel through.

  The *run-up length* is the distance between initiation of a flame and onset of detonation (supersonic combustion wave). *Markstein number* is the *Markstein length*, measuring the effect of curvature on a flame, divided by the flame thickness.
- **Feeding distance**

  Carbon steel shrinks during solidification and cooling. In order to avoid resulting porosity, a **riser** (a cylindric liquid metal reservoir) provides liquid feed metal until the end of the solidification process.

  A riser is evaluated by its **feeding distance** which is the maximum distance over which a riser can supply feed metal to produce a radiographically sound (i.e., relatively free of internal porosity) casting. The **feeding length** is the distance between the riser and the furthest point in the casting fed by it.
- **Etch depth**

  Laser etching into a metal substrate produces craters. The **etch depth** is the central crater depth averaged over the apparent roughness of the metal surface.
- **Approach distance**

  In metal cutting, the **approach distance** is the linear distance in the direction of feed between the point of initial cutter contact and the point of full cutter contact.
- **Input and output distances**

  The **input distance** d_{in} of a machine M is a distance machine is moved by the *input* (applied on it) force F_{in}. The **output distance** d_{out} is a distance the *output* (exerted by it) force F_{out} results in. The *mechanical advantage* of M is $\frac{F_{out}}{F_{in}} = \frac{d_{in}}{d_{out}}$.

  For example, the *effort* (or *resistance*) *distance* and *load distance* on a lever are the distances from the fulcrum to the resistance and load, respectively.
- **Instrument distances**

  Examples of such distances follow.

  The *K-distance*: the distance from the outside fiber of a rolled steel beam to the web toe of the fillet of a rolled shape.

  The *end distance* and *edge distance* are the distances from a fastener (say, bolt, screw, rivet, nail) to the end and, respectively, edge of treated material.

  The *calibration distance*: the standard distance used in the process of adjusting the output or indication on a measuring instrument.

- **Gear distances**

  Given two meshed gears, the distance between their centers is called the *center distance*. Examples of other distances used in basic gear formulas follow.

  Pitch diameter: the diameter of the *pitch circle* (the circle whose radius is equal to the distance from the center of the gear to the pitch point).

  Addendum: the radial distance between the pitch circle and the top of the teeth.

  Dedendum: the depth of the tooth space below the pitch line. It should be greater than the addendum of the mating gear to provide *clearance*.

  Whole depth: the total depth of a tooth space, equal to addendum plus dedendum.

  Working depth: the depth of engagement (i.e., the sum of addendums) of two gears.

  Backlash: the play between mating teeth.

- **Threaded fastener distances**

  Examples of distances applied to nuts, screws and other threaded fasteners, follow.

  Pitch: the nominal distance between two adjacent thread roots or crests.

  Ply: a single thickness of steel forming part of a structural joint.

  Grip length: the total distance between the underside of the nut to the bearing face of the bolt head.

  Effective nut radius: the radius from the center of the nut to the point where the contact forces, generated when the nut is turned, can be considered to act.

  Effective diameter (or *pitch diameter*): the diameter of an imaginary cylinder coaxial with the thread which has equal metal and space widths.

  Virtual effective diameter: the effective diameter of a thread, allowing for errors in pitch and flank angles. *Nominal diameter*: the external diameter of the threads.

  Major and *minor diameters* are the diameters of imaginary cylinders parallel with the crests of the thread (i.e., the distance, crest-to-crest for an external or root-to-root for an internal thread), or, respectively, just touching the roots of an external (or the crests of an internal) thread.

  Thread height: the distance between thread's minor and major diameters measured radially. *Thread length*: the length of the portion of the fastener with threads.

- **Distance spacer**

  A **distance spacer** is a device for holding two objects at a given distance from each other. Examples of such components are: male-female *distance bolt*, *distance bush*, *distance ring*, *distance plate*, *distance sleeve*, *distance finger*, *distance gauge*.

- **Sagging distance**

  The brazeability of brazing sheet materials is evaluated by their **sagging distance**, i.e., the **deflection** of the free end of the specimen sheet after brazing.

- **Haul distance**

  In Engineering, the **haul distance** is the average distance material is transported from where it originates to where it is deposited.

[redacted]

- **Deflection**

  In Engineering, **deflection** is the degree, in distance or angle, to which a structural part is displaced under a load/stress.

  In general, it can be a specified amount of deviation, say, the distance an elastic body or spring moves when subjected to a force, the amount by which a propagating wave or a projectile's trajectory is bent, and so on.
- **Distances in Structural Engineering**

  Examples of such distances related to superstructures (mainly, bridges and buildings) follow; cf. also **bar-and-joint framework** in Chap. 15.

  For a building, its *length* is the distance between out ends of wall steel lines, *width* is the distance from outside of *eave strut* (piece spanning columns at roof's edge) of one sidewall to outside of eave strut of the opposite sidewall, *height* is the distance from finished floor level to the top outer point of the eave strut.

  A *bay* refers the space between architectural elements. *Bay depth* is the distance from the building's corridor wall to the outside window. *End bay length* is the distance from outside of the outer *flange* (longitudinal part of a beam) of endwall columns to centerline of the first interior frame column. *Interior bay length* is the distance between the centerlines of two adjacent interior main frame columns.

  Clear height (or *head room*) is the vertical distance from the floor to the bottom of the lowest hanging overhead obstruction, allowable for passage.

  A *beam* is a structural element that is capable of withstanding load primarily by resisting bending; *girder* is a support beam used in construction. A *truss* is a framed or jointed structure designed to act as a beam while each of its members is primarily subjected to longitudinal stress only. Given a truss or girder, its *effective length* is the distance between the points of support, *effective depth* is the perpendicular distance between the gravity lines, and *economic depth* is the depth, which will give satisfactory results from all standpoints and involving the least expenditure of money for properly combined first cost, operation, maintenance and repairs.

  Effective span is the distance between supports (centres of bearings) in any structure. The **bearing distance** is the length of a beam between its bearing supports.

  For a bridge, its *effective span* is the center-to-center distance of end pins, *structural height* is the maximum vertical distance from the uppermost point down to the lowest visible point, while the *deck height* is the maximum vertical distance from the *deck* (road bed) down to the ground or water surface.

  Clear headway: the vertical distance from the lowest part of the superstructure to the ground or water surface; it is the measure of height of the tallest vehicle that could pass through the bridge. *Clear waterway*: the horizontal distance over the water, measured perpendicularly to the centrelines of adjacent piers.
- **Clearance distance**

  In Civil Engineering and Safety, a **clearance distance** (or *separation distance*, *clearance*) is a physical distance or unobstructed space tolerance between the loading and structural gauges as, for example, the distance between the lowest point on the vehicle and the road (*ground clearance*).

For vehicles going in a tunnel or under a bridge, the *clearance* is the difference between the *structure gauge* (minimum size of tunnel or bridge) and the vehicles' *loading gauge* (maximum size). A clearance distance can be prescribed by a code or a standard between a piece of equipment containing potentially hazardous material (say, fuel) and other objects (buildings, equipment, etc.) and the public. Or, say, no vehicle should parked nearer than 15 feet (4.6 m) from a fire hydrant.

In general, *clearance* refers to the distance to the nearest "obstacle" as defined in a context. It can be either a *tolerance* (the limit of an acceptable unplanned deviation from the nominal or theoretical dimension), or an *allowance* (planned deviation). Cf. **buffer distance** and **setback distance** in Chap. 25.

- **Creepage distance**

  The **creepage distance** is the shortest path distance along the surface of an insulation material between two conductive parts, while the shortest (straight line) distance between them is *clearance distance*; cf. the general term above.

- **Spark distance**

  The simplest way of measuring high voltages is by their **spark distance** (or **maximum spark length**). It is the length d of the gap between two electrodes in a gas, at which given voltage V becomes the *breakdown voltage*, i.e., starts a discharge or electric arc (a spark jumps over). Spark distance depends on the pressure p of gas and many other factors. The *Paschen's law* estimate V as a function of pd.

- **Humidifier absorption distance**

  The **absorption distance** of a (water centrifugal atomizing) humidifier is the list of minimum clearance dimensions needed to avoid condensation.

- **Spray distance**

  The **spray distance** is the distance maintained between the nozzle tip of a thermal spraying gun and the surface of the workpiece during spraying.

- **Protective action distance**

  The **protective action distance** is the distance downwind from an incident (say, a spill involving dangerous goods which are considered toxic by inhalation) in which persons may become incapacitated.

  The *screening distance* in a forest fire is the downwind distance which should be examined for possible smoke-sensitive human sites. The *spot fire distance* is the maximum distance between a source of firebrands (a group of burning trees) and a potential *spot fire* (a fire started by flying sparks or embers from the main fire). The *response distance* is the distance to fire traveled by fire companies.

  The notion of *mean distance between people and any hazardous event* operates also at a large scale: expanding the living area of human species (say, space colonization) will increase this distance and prevent many human extinction scenarios.

- **Fringe distance**

  Usually, the **fringe distance** is the spacing between *fringes*, for example, components into which a spectral line splits in the presence of an electric

[redacted]

or magnetic field (*Stark* and *Zeeman effects*, respectively, in Physics) or dark and light regions in the interference pattern of light beams (cf., in Chap. 24, *Pendellösung fringes* in **dynamical diffraction distances**).

For an interferometer, the fringe distance is the value $\frac{\lambda}{2\sin\alpha}$, where λ is the laser wavelength and α is the beam angle, while the **shear distance** is the spacing between two, due to the thickness of the plate, reflections.

In Image Analysis, there is also the *fringe distance* (Brown, 1994) between binary images (cf. **pixel distance** in Chap. 21).

- **Ranged weapon**

  A **ranged** (or *distance*) **weapon** is any weapon that can harm targets at distances greater than hand-to-hand distance; it can be thrown or used to fire projectiles at targets. A weapon intended to be used in hand-to-hand combat is called a *melee weapon*. But spears, knives, daggers can be used for both throwing and stabbing.

  The **maximum effective range** of a weapon is the greatest distance fired and able to produce casualties or damage consistently. The **effective weapon distance** is the actual distance (as opposed to maximal range) over which it is usually deployed. For given game and rifle type, the *effective hunting distance* (or *killing distance*) is the maximal range of a "clean kill."

- **Shooting distance**

  The **shooting distance** (or *shot distance*) is the distance achieved by, say, a bullet or a golf ball after a shot.

  The range of a Taser projectile delivering an incapacitating shock is called the *shocking distance*. Longest confirmed sniper kill at 2016 was 2475 m.

  For a shooting range, *firing distance* is the distance between the firing line and the target line. In shooting incident reconstruction, *firing distance* (or *muzzle-to-target distance*) is the distance from the muzzle of the firearm to the victim's clothing.

  In photography, the *shooting distance* is the camera-subject distance.

- **Lens distances**

  A convex lens is converging/magnifying; a concave one is diverging/reducing.

  The **focal distance** (*effective focal length*): the distance from the optical center of a lens (or a curved mirror) to the focus (to the image). Its reciprocal measured in m is called the *diopter* and is used as a unit of measurement of the (refractive) power of a lens; roughly, the magnification power of a lens is $\frac{1}{4}$ of its diopter.

  The *lens effective diameter* is twice the longest lens radius measured from its center to the apex of its edge. The *back focal length* is the distance between the rear surface of a lens and its image plane; the *front focal length* is the distance from the vertex of the first lens to the front focal point.

  Depth of field (DoF): the distance in the object plane (in front of and behind the object) over which the system delivers an acceptably sharp image, i.e., the region where blurring is tolerated at a particular resolution.

  The *depth of focus*: the range of distance in the image plane (the eyepiece, camera, or photographic plate) over which the system delivers an acceptably sharp image.

The *vertex depth* (or *sagitta*) is the depth of the surface curve on a lens measured over a specific diameter. Given a circle, the *apothem* is the perpendicular distance from the midpoint of a chord to the circle's center; it is the radius minus the *sagitta*.

The **working distance**: the distance from the front end of a lens system to the object when the instrument is correctly focused; it is used to modify the DoF. For a flashlight, it is the distance at which the *illuminance* (maximum light falling on a surface) would fall to 0.25 lux as, say, a full moon on a clear night.

The *register distance* (or *flange distance*): the distance between the flange (protruding rim) of the lens mount and the plane of the film image.

The *conjugate image distance* and *conjugate object distance*: the distances along the optical axis of a lens from its principal plane to the image and object plane, respectively. When a converging lens is placed between the object and the screen, the sum of the inverses of those distances is the inverse focal distance.

A *circle of confusion* (CoC) is an optical spot caused by a cone of light rays from a lens not coming to a perfect focus; in photography, it is the largest blur circle that will still be perceived as a point when viewed at a distance of 25 cm.

The *close* (or *minimum*, *near*) *focus distance*: the closest distance to which a lens can approach the subject and still achieve focus.

The **hyper-focal distance**: the distance from the lens to the nearest point (*hyper-focal point*) that is in focus when the lens is focused at infinity; beyond this point all objects are well defined and clear. It is the nearest distance at which the far end of the **depth of field** stretches to infinity (cf. **infinite distance**).

Eye relief: the distance an optical instrument can be held away from the eye and still present the full field-of-view. The *exit pupil width*: the width of the cone of light that is available to the viewer at the exact eye relief distance.

- **Distances in Stereoscopy**

A method of 3D imaging is to create a pair of 2D images by a two-camera system.

The *convergence distance* is the distance between the *baseine* of the camera center to the *convergence point* where the two lenses should converge for good stereoscopy. This distance should be 15–30 times the **intercamera distance**.

The *intercamera distance* (or *baseline length*, *interocular lens spacing*) is the distance between the two cameras from which the left and right eye images are rendered.

The *picture plane distance* is the distance at which the object will appear on the *picture plane* (the apparent surface of the image). The *window* is a masking border of the screen frame such that objects, which appear at (but not behind or outside) it, appear to be at the same distance from the viewer as this frame. In human viewing, the picture plane distance is about 30 times the **intercamera distance**.

- **Distance-related shots**

A film *shot* is what is recorded between the time the camera starts (the director's call for "action") and the time it stops (the call to "cut").

[redacted]

The main **distance-related shots** (camera setups) are:

- *establishing shot*: a shot, at the beginning of a sequence which establishes the location of the action and/or the time of day;
- *long shot*: a shot taken from at least 50 yards (45.7 m) from the action;
- *medium shot*: a shot from 5 to 15 yards (4.6–13.7 m), including a small entire group, which shows group/objects in relation to the surroundings;
- *close-up*: a shot from a close position, say, the actor from the neck upwards;
- *two-shot*: a shot that features two persons in the foreground;
- *insert*: an inserted shot (usually a close up) used to reveal greater detail.

29.3 Miscellany

- **Range distances**

  In Mathematics, *range* is the set of values of a function or variable; specifically, it means the difference (or interval, area) between a maximum and minimum.

  The **range distances** are practical distances emphasizing a maximum distance for effective operation such as vehicle travel without refueling, a bullet range, visibility, movement limit, home range of an animal, etc. For example, the range of a risk factor (toxicity, blast, etc.) indicates the minimal **safe distancing**.

  The **operating distance** (or *nominal sensing distance*) is the range of a device (for example, a remote control) which is specified by the manufacturer and used as a reference. The **activation distance** is the maximal distance allowed for activation of a sensor-operated switch.

- **Spacing distances**

  The following examples illustrate this large family of practical distances emphasizing a minimum distance; cf. **minimum distance**, nearest-neighbor **distance in Animal Behavior**, **first-neighbor distance** in Chaps. 16, 23, 24, respectively.

  The **miles in trail**: a specified minimum distance, in nautical miles, required to be maintained between airplanes. *Seat pitch* and *seat width* are airliner distances between, respectively, two rows of seats and the armrests of a single seat.

  The **isolation distance**: a specified minimum distance required (because of pollination) to be maintained between variations of the same species of a crop in order to keep the seed pure (for example, $\approx$ 3 m for rice).

  The **legal distance**: a minimum distance required by a judicial rule or decision, say, a distance a sex offender is required to live away from school.

  In a restraining order, **stay away** means to stay a certain distance (often 300 yards, i.e., 275 m) from the protected person. A general **distance restriction**: say, a minimum distance required for passengers traveling on some long distance trains in India, or a distance from a voting facility where campaigning is permitted.

The **stop-spacing distance**: the interval between bus stops; such mean distance in US light rail systems ranges from 500 (Philadelphia) to 1742 m (Los Angeles).

The **character spacing**: the interval between characters in a given computer font.

The **just noticeable difference** (JND): the smallest perceived percent change in a dimension (for distance/position, etc.). *Weber's law* states that JND between two stimuli is proportional to their magnitude (and the subject's sensitivity).

- **Cutoff distances**

Given a range of values (usually, a length, energy, or momentum scale in Physics), *cutoff* (or *cut-off*) is the maximal or minimal value, as, for example, Planck units.

A **cutoff distance** is a cutoff in a length scale. For example, infrared and ultraviolet cutoff (the maximal and minimal wavelength that the human eye takes into account) are *long-distance* and *short-distance cutoff*, respectively, in the visible spectrum. Cutoff distances are often used in Molecular Dynamics.

A similar notion of a **threshold distance** refers to a limit, margin, starting point distance (usually, minimal) at which some effect happens or stops. Some examples are the threshold distance of sensory perception, neuronal reaction or, say, upon which a city or road alters the abundance patterns of the native bird species.

- **Quality metrics**

A **quality metric** (or, simply, *metric*) is a standard unit of measure or, more generally, part of a system of parameters, or systems of measurement. This vast family of measures (or standards of measure) concerns different attributes of objects. In such a setting, our distances and similarities are rather "similarity metrics", i.e., metrics (measures) quantifying the extent of relatedness between two objects.

Examples include academic metrics, crime statistics, corporate investment metrics, economic metrics (indicators), education metrics, environmental metrics (indices), health metrics, market metrics, political metrics, properties of a route in computer networking, software metrics and vehicle metrics.

For example, the site http://metripedia.wikidot.com/start aims to build an Encyclopedia of IT (Information Technology) performance metrics. Some examples of nonequipment quality metrics are detailed below.

Landscape metrics evaluate, for example, greenway patches in a given landscape by *patch density* (the number of patches per km^2), *edge density* (the total length of patch boundaries per hectare), *shape index* $\frac{E}{4\sqrt{A}}$ (where A is the total area, and E is the total length of edges), connectivity, diversity, etc.

Morphometrics evaluate the forms (size and shape) related to organisms (brain, fossils, etc.). For example, the roughness of a fish school is measured by its *fractal dimension* $2\frac{\ln P - \ln 4}{\ln A}$ where P, A are its perimeter (m) and surface (m^2).

[redacted]

Management metrics include: surveys (say, of market share, sales increase, customer satisfactions), forecasts (say, of revenue, contingent sales, investment), R&D effectiveness, absenteeism, etc.

Risk metrics are used in Insurance and, in order to evaluate a portfolio, in Finance.

Importance metrics rank the relative influence such as, for example:

- *PageRank* of Google ranking web pages;
- ISI (now Thomson Scientific) *Impact Factor* of a journal measuring, for a given two-year period, the number of times the average article in this journal is cited by some article published in the subsequent year;
- Hirsch's *h-index* of a scholar: the largest number h such that h of his/her publications have at least h citations;
- and his/her $i10$-*index*: the number of publications with at least 10 citations.

- **Heterometric and homeometric**

  The adjective **heterometric** means involving or dependent on a change in size, while **homeometric** means independent of such change.

  Those terms are used mainly in Medicine; for example, *heterometric* and *homeometric autoregulation* refer to intrinsic mechanisms controlling the strength of ventricular contractions that depend or not, respectively, on the length of myocardial fibers at the end of diastole; cf. **distances in Medicine**.

- **Distal and proximal**

  The antipodal notions near (close, nigh) and far (distant, remote) are also termed *proximity* and *distality*.

  The adjective **distal** (or *peripheral*) is an anatomical term of location (on the body, the limbs, the jaw, etc.); corresponding adverbs are: distally, distad.

  For an *appendage* (any structure that extends from the main body), **proximal** means situated towards the point of attachment, while **distal** means situated around the furthest point from this point of attachment. More generally, as opposed to *proximal* (or *central*), *distal* means: situated away from, farther from a point of reference (origin, center, point of attachment, trunk). As opposed to *mesial* it means: situated or directed away from the midline or mesial plane of the body.

  Proximal and distal *demonstratives* are words indicating *place deixis*, i.e., a spatial location relative to the point of reference. Usually, they are two-way as *this/that*, these/those or *here/there*, i.e., in terms of the dichotomy near/far from the speaker. But, say, Korean, Japanese, Spanish, and Thai make a three-way distinction: proximal (near to the speaker), medial (near to the addressee) and distal (far from both). English had the third form, *yonder* (at an indicated distance within sight), still spoken in Southern US. Cf. **spatial language** in Chap. 28.

  A **distal stimulus** is an real-word object or event, which, by some physical process, stimulates the body's sensory organs. Resulting raw pattern of neural activity is called the **proximal stimulus**. *Perception* is the constructing mental representations of distal stimuli using the information available in proximal stimuli.

A *proximate cause* is an event which is closest to, or immediately responsible for causing, some observed result. This exists in contrast to a higher-level *ultimate* (or *distal*) *cause* which is usually thought of as the "real" reason something occurred.

Tinbergen's (1960) *proximate* and *ultimate questions* about behavior are "how" an organism structures function? and "why" a species evolved the structures it has?

- **Distance effect**

  The **distance effect** is a general term for the change of a pattern or process with distance. Usually, it is the result of **distance decay**. For example, a static field attenuates proportionally to the inverse square of the distance from the source.

  Another example of the distance effect is a periodic variation (instead of uniform decrease) in a certain direction, when a *standing wave* occurs in a time-varying field. It is a wave that remains in a constant position because either the medium is moving in the opposite direction, or two waves, traveling in opposite directions, interfere; cf. **Pendellösung length** in Chap. 24.

  The distance effect, together with the size (source magnitude) effect determine many processes; cf. **island distance effect**, **insecticide distance effect** in Chap. 23 and **symbolic distance effect**, **distance effect on trade** in Chap. 28.

- **Distance decay**

  The **distance decay** is the attenuation of a pattern or process with distance. Cf. **distance decay (in Spatial Interaction)** in Chap. 28.

  Examples of distance-decay curves: Pareto model $\ln I_{ij} = a - b \ln d_{ij}$, and the model $\ln I_{ij} = a - bd_{ij}^p$ with $p = \frac{1}{2}$, 1, or 2 (here I_{ij} and d_{ij} are the interaction and distance between points i, j, while a and b are parameters). The *Allen curve* gives the exponential drop of frequency of all communication between engineers as the distance between their offices increases, i.e., face-to-face probability decays.

  A **mass-distance decay curve** is a plot of "mass" decay when the distance to the center of "gravity" increases. Such curves are used, say, to determine an *offender's heaven* (the point of origin; cf. **distances in Criminology**) or the galactic mass within a given radius from its center (using *rotation-distance curves*).

- **Distance factor**

  A **distance factor** is a multiplier of some straight-line distance needed to account for additional data. For example, 10 % increase of aircraft weight implies 20 % increase, i.e., a distance factor of 1.2, in needed take-off distance.

- **Propagation length**

  For a pattern or process attenuating with distance, the **propagation length** is the distance to decay by a factor of $\frac{1}{e}$.

  Cf. **radiation length** and the **Beer–Lambert law** in Chap. 24.

  A **scale height** is a distance over which a quantity decreases by a factor of e.

- **Incremental distance**

  An **incremental distance** is a gradually increasing (by a fixed amount) one.

[redacted]

- **Distance curve**

  A **distance curve** is a plot (or a graph) of a given parameter against a corresponding distance. Examples of distance curves, in terms of a process under consideration, are: **time-distance curve** (for the travel time of a wave-train, seismic signals, etc.), *height-run distance curve* (for the height of tsunami wave versus wave propagation distance from the impact point), *drawdown-distance curve*, *melting-distance curve* and *wear volume-distance curve*.

  A **force-distance curve** is, in SPM (scanning probe microscopy), a plot of the vertical force that the tip of the probe applies to the sample surface, while a contact-AFM (Atomic Force Microscopy) image is being taken. Also, *frequency-distance* and *amplitude-distance* curves are used in SPM.

  The term *distance curve* is also used for charting growth, for instance, a child's height or weight at each birthday. A plot of the rate of growth against age is called the **velocity-distance curve**; this term is also used for the speed of aircraft. Example of a constant rate of growth: in a month, human (hear or body) hair grow 15 and 8.1 mm, while nails (finger and toe) grow 3.5 and 1.6 mm.
- **Distance sensitivity**

  Distance sensitivity is a general term used to indicate the dependence of something on the associated distance. It could be, say, commuting distance sensitivity of households, traveling distance sensitivity of tourists, distance sensitive technology, distance sensitive products/services and so on.
- **Characteristic diameters**

  Let X be an irregularly-shaped 3D object, say, Earth's spheroid or a particle. A **characteristic diameter** (or **equivalent diameter**) of X is the diameter of a sphere with the same geometric or physical property of interest. Examples follow.

  The **authalic diameter** and **volumetric diameter** (*equivalent spherical diameter*) of X are the diameters of the spheres with the same surface area and volume. The **Heywood diameter** is the diameter of a circle with the same projection area. Cf. the **Earth radii** in Chap. 25 and the **shape parameters** in Chap. 21.

  The **Stokes diameter** is the diameter of the sphere with the same gravitational velocity as X, while the **aerodynamic diameter** is the diameter of such sphere of unit density. Cf. the **hydrodynamic radius** in Chap. 24.

  Equivalent *electric mobility*, *diffusion* and *light scattering* diameters of a particle X are the diameters of the spheres with the same electric mobility, penetration and intensity of light scattering, respectively, as X.
- **Characteristic length**

  A **characteristic length** (or *scale*) is a convenient reference length of a given configuration, such as the overall length of an aircraft, the maximum diameter or radius of a body of revolution, or a chord or span of a lifting surface.

  In general, it is a length that is representative of the system (or region) of interest, or the parameter which characterizes a given physical quantity in, say, Heat Transfer or Fluid Mechanics. For complex shapes, it is defined as the volume of the body divided by the surface area. For example, for a rocket engine, it is the ratio of the volume of its combustion chamber to the area of the nozzle's throat, representing the average distance that the products of burned fuel must travel to escape.

- **True length**

  In Engineering Drawing, **true length** is any distance between points that is not foreshortened by the view type. In 3D, lines with true length are parallel to the projection plane, as, for example, the base edges in a top view of a pyramid.

- **Path length**

  In general, a path is a line representing the course of actual, potential or abstract movement. In Topology, a path is a certain continuous function; cf. parametrized **metric curve** in Chap. 1.

  In Physics, **path length** is the total distance an object travels, while *displacement* is the net distance it travels from a starting point. Cf. **displacement**, **inelastic mean free path**, **optical distance** and **dislocation path length** in Chap. 24. In Chemistry, (cell) **path length** is the distance that light travels through a sample in an analytical cell.

  In Graph Theory, **path length** is a discrete notion: the number of vertices in a sequence of vertices of a graph; cf. **path metric** in Chap. 1. Cf. **Internet IP metric** in Chap. 22 for **path length** in a computer network. Also, it means the total number of machine code instructions executed on a section of a program.

- **Middle distance**

  The **middle distance** is a general term. It can be a precise distance (cf. **running distances**), the halfway between the observer and the horizon (cf. **distance to horizon** in Chap. 25), implied horizon of a scene (cf. **representation of distance in Painting** in Chap. 28), or the place that you can see when you are not quite focusing on the world around you (Urban Dictionary of slang). Another example: goals in the middle distance (1 block to 2 miles) are places one might walk to.

- **Long-distance**

  The term **long-distance** usually refers to telephone communication (long-distance call, operator) or to covering large distances by moving (long-distance trail, running, swimming, riding of motorcycles or horses, etc.) or, more abstractly: long-distance migration, commuting, supervision, relationship, etc. For example, a *long-distance relationship* (LDR) is typically an intimate relationship that takes place when the partners are separated by a considerable distance.

  For example, a *long-distance* (or *distance*) *thug* has two meanings: 1. a person that is a coward in real life, but gathers courage from behind the safety of a computer, phone, or through e-mail; and 2. a hacker, spammer, or scam artist that takes advantage of the Internet to cause harm to others from a distance.

  Cf. **long-distance dispersal**, **animal** and **plant long-distance communication**, **long range order**, **long range dependence**, **action at a distance** (in Computing, Physics, along DNA).

  DDD (or *direct distance dialing*) is any switched telecommunication service (like 1+, 0 + +, etc.) that allows a call originator to place long-distance calls directly to telephones outside the local service area without an operator.

[redacted]

The term *short-distance* is rarely used. Instead, the adjective **short range** means limited to (or designed for) short distances, or relating to the near future. Finally, **touching**, for two objects, is having (or getting) a zero distance between them.

- **Long-distance intercourse**

  Long-distance intercourse (coupling at a distance) is found often in Native American folklore: Coyote, the Trickster, is said to have lengthened his penis to enable him to have intercourse with a woman on the opposite bank of a lake.

  A company Distance Labs has announced the "intimate communication over a distance", an interactive installation *Mutsugoto* which draws, using a custom computer vision and projection system, lines of light on a body of a person. Besides light, haptic technology provides a degree of touch communication between remote users. A company Lovotics created *Kissinger*, a messaging device wirelessly sending kisses. Other devices can send hugs, strokes and squeezes over the internet by making companion devices respond with vibrations and temperature changes.

  In Nature, the *acorn barnacle* (small sessile crustacean) have the largest penis-body size ratio (up to 10 when extended) of any animal. The squid *Onykia ingens* have largest ratio among mobile animals. The male octopus *Argonauta* use a modified arm, the *hectocotylus*, to transfer sperm to the female *at a distance*; this tentacle detaches itself from the body and swims—under its own power—to the female.

  Also, many aquatic animals (say, coral, hydra, sea urchin, bony fish) and amphibians reproduce by external fertilization: eggs and sperm are released into the water. Similar transfer of sperm at a distance is pollination (by wind or organisms) in flowering plants. Another example is *in vitro* fertilization in humans.

  The shortest range intercourse happens in anglerfish. The male, much smaller, latches onto a female with his sharp teeth, fuses inside her to the blood-vessel level and degenerates into a pair of testicles. It releases sperm when the female (with about 6 males inside) releases eggs. But female-male pairings of a parasitic worm *Schistosoma mansoni* is monogamous: the male's body forms a channel, in which it holds the longer and thinner female for their entire adult lives, up to 30 years. Two worms *Diplozoon paradoxum* fuse completely for lifetime of cross-fertilization.

- **Go the distance**

  Go the (full) **distance** is a general distance idiom meaning to continue to do something until it is successfully completed.

  An *unbridgeable distance* is a distance (seen as a spatial or metaphoric extent), impossible to span: a wide unbridgeable river, chasm or, in general, differences.

Newnes Radio and RF Engineering Pocket Book

Newnes Radio and RF Engineering Pocket Book

3rd edition

Steve Winder
Joe Carr

Newnes

OXFORD AMSTERDAM BOSTON LONDON NEW YORK PARIS
SAN DIEGO SAN FRANCISCO SINGAPORE SYDNEY TOKYO

Newnes
An imprint of Elsevier Science
Linacre House, Jordan Hill, Oxford OX2 8DP
225 Wildwood Avenue, Woburn, MA 01801-2041

First published 1994
Reprinted 2000, 2001
Second edition 2000
Third edition 2002

Copyright © 1994, 2000, 2002, Steve Winder. All rights reserved

The right of Steve Winder to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988

No part of this publication may be reproduced in any material form (including photocopying or storing in any medium by electronic means and whether or not transiently or incidentally to some other use of this publication) without the written permission of the copyright holder except in accordance with the provisions of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London, England W1T 4LP. Applications for the copyright holder's written permission to reproduce any part of this publication should be addressed to the publisher

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0 7506 5608 5

For information on all Newnes publications
visit our website at www.newnespress.com

Typeset by Laserwords Private Limited, Chennai, India.
Printed and bound in Great Britain

Contents

Preface to second edition xi

Preface to third edition xiii

1 Propagation of radio waves 1
  - 1.1 Frequency and wavelength 1
  - 1.2 The radio frequency spectrum 1
  - 1.3 The isotropic radiator 3
  - 1.4 Formation of radio waves 3
  - 1.5 Behaviour of radio waves 7
  - 1.6 Methods of propagation 13
  - 1.7 Other propagation topics 18
  - References 24

2 The decibel scale 25
  - 2.1 Decibels and the logarithmic scale 25
  - 2.2 Decibels referred to absolute values 25

3 Transmission lines 35
  - 3.1 General considerations 35
  - 3.2 Impedance matching 35
  - 3.3 Base band lines 36
  - 3.4 Balanced line hybrids 36
  - 3.5 Radio frequency lines 37
  - 3.6 Waveguides 45
  - 3.7 Other transmission line considerations 47
  - References 51

4 Antennas 52
  - 4.1 Antenna characteristics 52
  - 4.2 Antenna types 56
  - 4.3 VHF and UHF antennas 60
  - 4.4 Microwave antennas 69
  - 4.5 Loop antennas 73
  - References 78

5 Resonant circuits 79
  - 5.1 Series and parallel tuned circuits 79
  - 5.2 *Q* factor 81
  - 5.3 Coupled (band-pass) resonant circuits 81
  - References 84

6 Oscillators 85
  6.1 Oscillator requirements 85
  6.2 Tunable oscillators 85
  6.3 Quartz crystal oscillators 87
  6.4 Frequency synthesizers 89
  6.5 Caesium and rubidium frequency standards 93
  References 94

7 Piezo-electric devices 95
  7.1 Piezo-electric effect 95
  7.2 Quartz crystal characteristics 97
  7.3 Specifying quartz crystals 101
  7.4 Filters 102
  7.5 SAW filters and resonators 105
  References 109

8 Bandwidth requirements and modulation 110
  8.1 Bandwidth of signals at base band 110
  8.2 Modulation 112
  8.3 Analogue modulation 113
  8.4 Digital modulation 123
  8.5 Spread spectrum transmission 129
  References 131

9 Frequency planning 132
  9.1 International and regional planning 132
  9.2 National planning 132
  9.3 Designations of radio emissions 134
  9.4 Bandwidth and frequency designations 135
  9.5 General frequency allocations 135
  9.6 Classes of radio stations 139
  9.7 Radio wavebands 142
  Reference 142

10 Radio equipment 143
  10.1 Transmitters 143
  10.2 Receivers 148
  10.3 Programmable equipment 157
  References 158

11 Microwave communication 159
  11.1 Microwave usage 159
  11.2 Propagation 159

11.3 *K* factor 161
11.4 Fresnel zones, reflections and multi-path fading 161
11.5 Performance criteria for analogue and digital links 164
11.6 Terminology 165
11.7 Link planning 165
11.8 Example of microwave link plan 165
Reference 166

12 Information privacy and encryption 167
12.1 Encryption principles 167
12.2 Speech encryption 168
12.3 Data encryption 169
12.4 Code division multiple access (CDMA) or spread spectrum 172
12.5 Classification of security 172
References 172

13 Multiplexing 173
13.1 Frequency division multiplex 173
13.2 Time division multiplex (TDM) 174
13.3 Code division multiple access (CDMA) 177
Reference 178

14 Speech digitization and synthesis 179
14.1 Pulse amplitude modulation 179
14.2 Pulse code modulation 179
14.3 ADPCM codecs 181
14.4 The G728 low delay CELP codec 181
14.5 The GSM codec 182
References 182

15 VHF and UHF mobile communication 183
15.1 Operating procedures 183
15.2 Control of base stations 186
15.3 Common base station (CBS) operation 186
15.4 Wide area coverage 187

16 Signalling 194
16.1 Sub-audio signalling 194
16.2 In-band tone and digital signalling 195
16.3 Digital signalling 197
16.4 Standard PSTN tones 198
References 199

17 Channel occupancy, availability and trunking 200
17.1 Channel occupancy and availability 200
17.2 Trunking 201
17.3 In-band interrupted scan (IBIS) trunking 203
17.4 Trunking to MPT 1327 specification 203
References 204

18 Mobile radio systems 205
18.1 Paging 205
18.2 Cordless telephones 206
18.3 Trunked radio 207
18.4 Analogue cellular radio-telephone networks 208
18.5 Global system mobile 209
18.6 Other digital mobile systems 211
18.7 Private mobile radio (PMR) 213
18.8 UK CB radio 213
References 213

19 Base station site management 214
19.1 Base station objectives 214
19.2 Site ownership or accommodation rental? 214
19.3 Choice of site 214
19.4 Masts and towers 215
19.5 Installation of electronic equipment 216
19.6 Earthing and protection against lightning 217
19.7 Erection of antennas 219
19.8 Interference 221
19.9 Antenna multi-coupling 225
19.10 Emergency power supplies 226
19.11 Approval and certification 227
References 227

20 Instrumentation 229
20.1 Accuracy, resolution and stability 229
20.2 Audio instruments 230
20.3 Radio frequency instruments 231
References 235

21 Batteries 236
21.1 Cell characteristics 236
21.2 Non-rechargeable, primary batteries 238
21.3 Rechargeable batteries 242

22 Satellite communications 246
22.1 Earth orbits 246
22.2 Communications by satellite link 248
22.3 Proposed satellite television formats 248
22.4 Global positioning system (GPS) 252
References 255

23 Connectors and interfaces 256
23.1 Audio and video connectors 256
23.2 Co-axial connector 258
23.3 Interfaces 268
Reference 280

24 Broadcasting 281
24.1 Standard frequency and time transmissions 281
24.2 Standard frequency formats 283
24.3 UK broadcasting bands 284
24.4 BBC VHF test tone transmissions 284
24.5 Engineering information about broadcast services 287
24.6 Characteristics of UHF terrestrial television systems 288
24.7 Terrestrial television channels 291
24.8 Terrestrial television aerial dimensions 294
24.9 AM broadcast station classes (USA) 295
24.10 FM broadcast frequencies and channel numbers (USA) 296
24.11 US television channel assignments 299
24.12 License-free bands 301
24.13 Calculating radio antenna great circle bearings 302

25 Abbreviations and symbols 307
25.1 Abbreviations 307
25.2 Letter symbols by unit name 313
25.3 Electric quantities 321

26 Miscellaneous data 323
26.1 Fundamental constants 323
26.2 Electrical relationships 323
26.3 Dimensions of physical properties 324
26.4 Fundamental units 324
26.5 Greek alphabet 325

26.6 Standard units 325
26.7 Decimal multipliers 327
26.8 Useful formulae 327
26.9 Colour codes 334

Index 337

Preface to second edition

This edition of the *Newnes Radio and RF Engineer's Pocket Book* is something special. It is a compendium of information of use to engineers and technologists who are engaged in radio and RF engineering. It has been updated to reflect the changing interests of those communities, and reflects a view of the technology like no other. It is packed with information!

This whole series of books is rather amazing with regard to the range and quality of the information they provide, and this book is no different. It covers topics as diverse as circuit symbols and the abbreviations used for transistors, as well as more complex things as satellite communications and television channels for multiple countries in the English speaking world. It is a truly amazing work.

We hope that you will refer to this book frequently, and will enjoy it as much as we did in preparing it.

John Davies
Joseph J. Carr

Acknowledgements

I gratefully acknowledge the ready assistance offered by the following organizations: Andrew Ltd, Aspen Electronics Ltd, BBC, British Telecommunications plc, Farnell Instruments Ltd, Independent Television Authority, International Quartz Devices Ltd, Jaybeam Ltd, MACOM Greenpar Ltd, Marconi Instruments Ltd, Panorama Antennas Ltd, Radiocommunications Agency, the Radio Authority, RTT Systems Ltd. A special thanks goes to my wife Dorothy for once again putting up with my months of seclusion during the book's preparation.

Preface to third edition

This, the third edition of the *Newnes Radio and RF Engineering Pocket Book* has been prepared with a tinge of sadness. Joe Carr, who edited the second edition, has died since the last edition was published. Although I did not know Joe personally, his prolific writing over recent years has impressed me. His was a hard act to follow.

I have updated this book to be more international. Thus the long tables giving details of British television transmitters have been removed (they are available on the Web). Details of the European E1 multiplexing system have been supplemented by a description of the US and Japanese T1 system. There are many more general updates included throughout.

Steve Winder

1 Propagation of radio waves

1.1 Frequency and wavelength

There is a fixed relationship between the frequency and the wavelength, which is the distance between identical points on two adjacent waves (*Figure 1.1*), of any type of wave: sound (pressure), electromagnetic (radio) and light. The type of wave and the speed at which the wavefront travels through the medium determines the relationship. The speed of propagation is slower in higher density media.

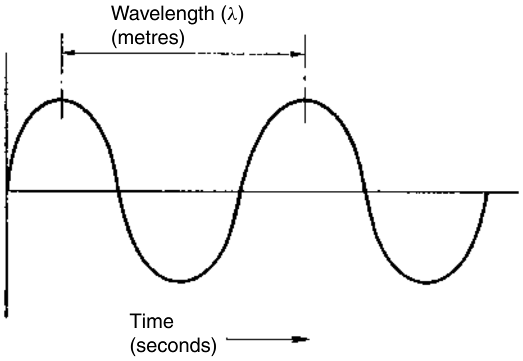

Figure 1.1 Frequency and wavelength

Sound waves travel more slowly than radio and light waves which, in free space, travel at the same speed, approximately 3×10^8 metres per second, and the relationship between the frequency and wavelength of a radio wave is given by:

$$\lambda = \frac{3 \times 10^8}{f} \text{ metres}$$

where λ is the wavelength and f is the frequency in hertz (Hz).

1.2 The radio frequency spectrum

The electromagnetic wave spectrum is shown in *Figure 1.2*: the part usable for radio communication ranges from below 10 kHz to over 100 GHz.

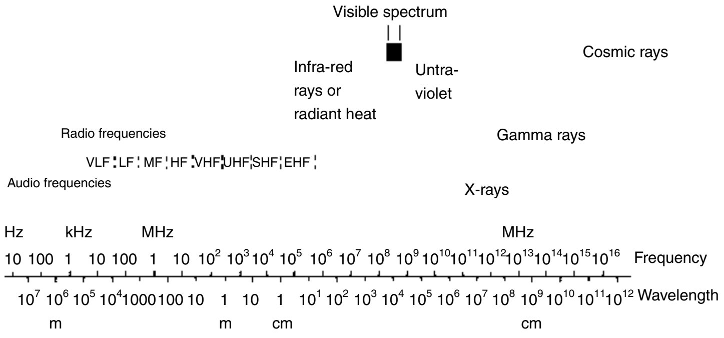
Visible spectrum
Cosmic rays
Infra-red rays or radiant heat
Untra-violet
Radio frequencies
Gamma rays
VLF LF MF HF VHF UHF SHF EHF
Audio frequencies
X-rays
Hz
kHz
MHz
MHz
10 100 1 10 100 1 10 10^2 10^3 10^4 10^5 10^6 10^7 10^8 10^9 10^{10} 10^{11} 10^{12} 10^{13} 10^{14} 10^{15} 10^{16}
Frequency
10^7 10^6 10^5 10^4 1000 100 10 1 10 1 10^1 10^2 10^3 10^4 10^5 10^6 10^7 10^8 10^9 10^{10} 10^{11} 10^{12}
Wavelength
m
m
cm
cm

Figure 1.2 The electromagnetic wave spectrum

The radio spectrum is divided into bands and the designation of the bands, their principal use and method of propagation is shown in *Table 1.1*. Waves of different frequencies behave differently and this, along with the amount of spectrum available in terms of radio communication channels in each band, governs their use.

Table 1.1 Use of radio frequencies

Frequency band	*Designation, use and propagation*
3–30 kHz	Very low frequency (VLF). Worldwide and long distance communications. Navigation. Submarine communications. Surface wave.
30–300 kHz	Low frequency (LF). Long distance communications, time and frequency standard stations, long-wave broadcasting. Ground wave.
300–3000 kHz	Medium frequency (MF) or medium wave (MW). Medium-wave local and regional broadcasting. Marine communications. Ground wave.
3–30 MHz	High frequency (HF). ‘Short-wave’ bands. Long distance communications and short-wave broadcasting. Ionospheric sky wave.
30–300 MHz	Very high frequency (VHF). Short range and mobile communications, television and FM broadcasting. Sound broadcasting. Space wave.
300–3000 MHz	Ultra high frequency (UHF). Short range and mobile communications. Television broadcasting. Point-to-point links. Space wave. Note: The usual practice in the USA is to designate 300–1000 MHz as ‘UHF’ and above 1000 MHz as ‘microwaves’.
3–30 GHz	Microwave or super high frequency (SHF). Point-to-point links, radar, satellite communications. Space wave.
Above 30 GHz	Extra high frequency (EHF). Inter-satellite and micro-cellular radio-telephone. Space wave.

1.3 The isotropic radiator

A starting point for considering the propagation of radio- or lightwaves is the isotropic radiator, an imaginary point source radiating equally in all directions in free space. Such a radiator placed at the centre of a sphere illuminates equally the complete surface of the sphere. As the surface area of a sphere is given by $4\pi r^2$ where r is the radius of the sphere, the brilliance of illumination at any point on the surface varies inversely with the distance from the radiator. In radio terms, the power density at distance from the source is given by:

$$P_\mathrm{d} = \frac{P_\mathrm{t}}{4\pi r^2}$$

where P_t = transmitted power.

1.4 Formation of radio waves

Radio waves are electromagnetic. They contain both electric and magnetic fields at right angles to each other and also at right angles to the direction of propagation. An alternating current flowing in a conductor produces an alternating magnetic field surrounding it and an alternating voltage gradient – an electric field – along the length of the conductor. The fields combine to radiate from the conductor as in *Figure 1.3*.

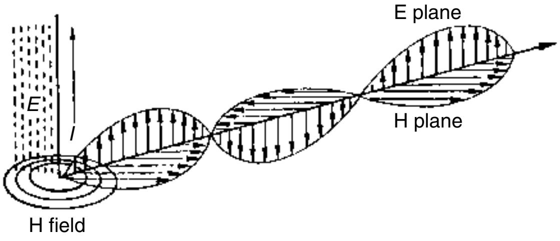

Figure 1.3 Formation of electromagnetic wave

The plane of the electric field is referred to as the E plane and that of the magnetic field as the H plane. The two fields are equivalent to the voltage and current in a wired circuit. They are measured in similar terms, volts per metre and amperes per metre, and the medium through which they propagate possesses an impedance. Where $E = ZI$ in a wired circuit, for an electromagnetic wave:

$$E = ZH$$

where
$E =$ the RMS value of the electric field strength, V/metre
$H =$ the RMS value of the magnetic field strength, A/metre
$Z =$ characteristic impedance of the medium, ohms

The voltage is that which the wave, passing at the speed of light, would induce in a conductor one metre long.

The characteristic impedance of the medium depends on its permeability (equivalent of inductance) and permittivity (equivalent of capacitance). Taking the accepted figures for free space as:

$\mu = 4\pi \times 10^{-7}$ henrys (H) per metre (permeability) and

$\varepsilon = 1/36\pi \times 10^{9}$ farads (F) per metre (permittivity)

then the impedance of free space, Z, is given by:

$$\sqrt{\frac{\mu}{\varepsilon}} = 120\pi = 377\,\text{ohms}$$

The relationship between power, voltage and impedance is also the same for electromagnetic waves as for electrical circuits, $W = E^2/Z$.

The simplest practical radiator is the elementary doublet formed by opening out the ends of a pair of wires. For theoretical considerations the length of the radiating portions of the wires is made very short in relation to the wavelength of the applied current to ensure uniform current distribution throughout their length. For practical applications the length of the radiating elements is one half-wavelength ($\lambda/2$) and the doublet then becomes a dipole antenna (*Figure 1.4*).

When radiation occurs from a doublet the wave is polarized. The electric field lies along the length of the radiator (the E plane) and the magnetic field (the H plane) at right angles to it. If the E plane is vertical, the radiated field is said to vertically polarized. Reference to the E and H planes avoids confusion when discussing the polarization of an antenna.

Unlike the isotropic radiator, the dipole possesses directivity, concentrating the energy in the H plane at the expense of the E plane. It effectively provides a power gain in the direction of the H plane

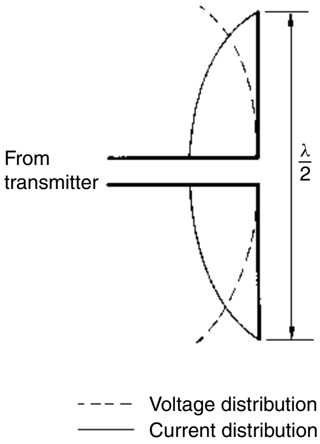
From
transmitter
$\frac{\lambda}{2}$
Voltage distribution
Current distribution

Figure 1.4 Doublet (dipole) antenna

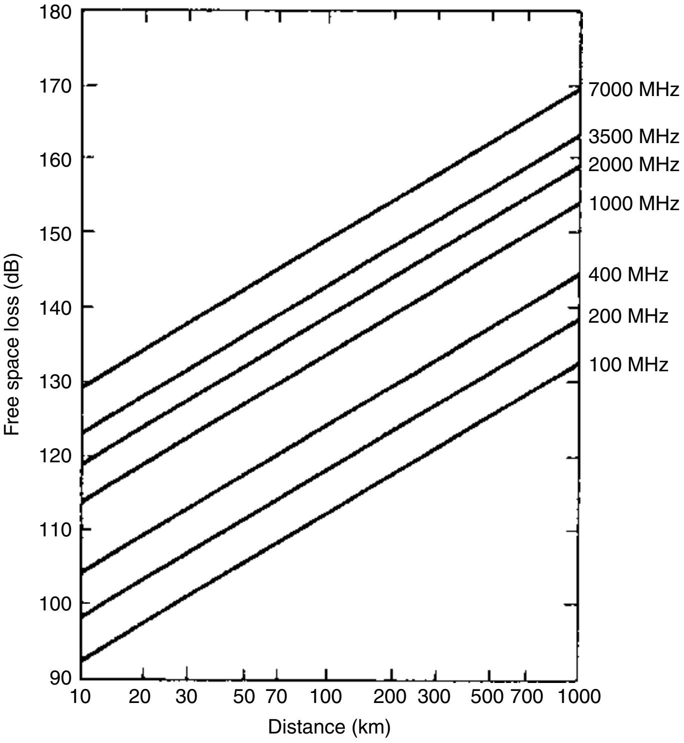

Figure 1.5 Free space loss vs. distance and frequency

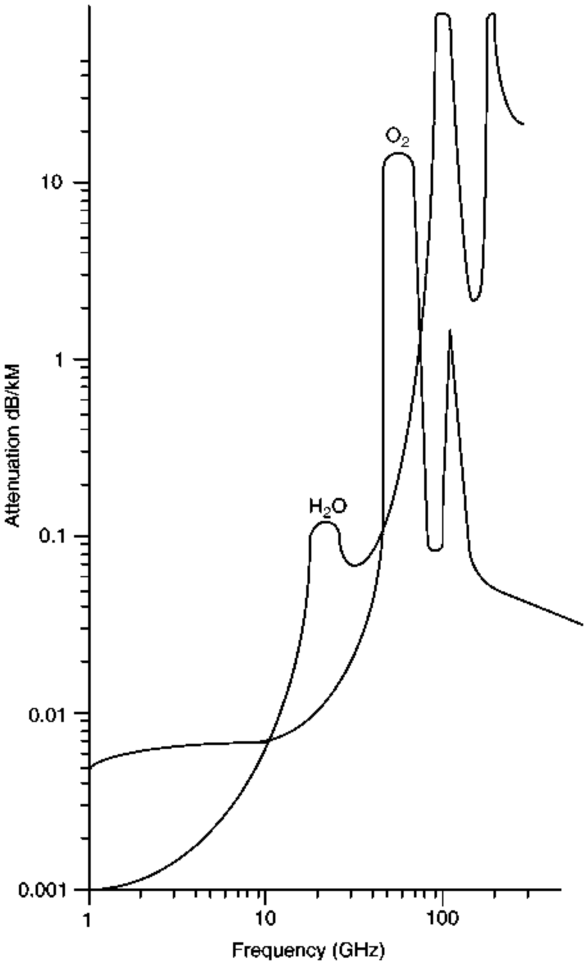

Figure 1.6 Major loss of microwave communications and radar systems due to atmospheric attenuation

compared with an isotropic radiator. This gain is 1.6 times or 2.15 dBi (dBi means dB relative to an isotropic radiator).

For a direct ray the power transfer between transmitting and receiving isotropic radiators is inversely proportional to the distance between them *in wavelengths*. The free space power loss is given by:

$$\text{Free space loss, dB} = 10\log_{10}\frac{(4\pi d)^2}{\lambda^2}$$

where d and λ are in metres, or:

$$\text{Free space loss (dB)} = 32.4 + 20 \times \log_{10} d + 20 \times \log_{10} f$$

where d = distance in km and f = frequency in MHz.

The free space power loss, therefore, increases as the square of the distance and the frequency. Examples are shown in *Figure 1.5*.

With practical antennas, the power gains of the transmitting and receiving antennas, in dBi, must be subtracted from the free space loss calculated as above. Alternatively, the loss may be calculated by:

$$\text{Free space loss (dB)} = 10 \log_{10} \left[\frac{(4\pi d)^2}{\lambda^2} \times \frac{1}{G_\mathrm{t} \times G_\mathrm{r}} \right]$$

where G_t and G_r are the respective actual gains, not in dB, of the transmitting and receiving antennas.

A major loss in microwave communications and radar systems is atmospheric attenuation (see *Figure 1.6*). The attenuation (in decibels per kilometre (dB/km)) is a function of frequency, with especial problems showing up at 22 GHz and 64 GHz. These spikes are caused by *water vapour* and *atmospheric oxygen* absorption of microwave energy, respectively. Operation of any microwave frequency requires consideration of atmospheric losses, but operation near the two principal spike frequencies poses special problems. At 22 GHz, for example, an additional 1 dB/km of loss must be calculated for the system.

1.5 Behaviour of radio waves

1.5.1 Physical effects

The physical properties of the medium through which a wave travels, and objects in or close to its path, affect the wave in various ways.

Absorption

In the atmosphere absorption occurs and energy is lost in heating the air molecules. Absorption caused by this is minimal at frequencies below about 10 GHz but absorption by foliage, particularly when wet, is severe at VHF and above.

Waves travelling along the earth's surface create currents in the earth causing ground absorption which increases with frequency. A horizontally polarized surface wave suffers more ground absorption than a vertically polarized wave because of the 'short-circuiting' by the ground of the electric field. Attenuation at a given frequency is least for propagation over water and greatest over dry ground for a vertically polarized wave.

Refraction and its effect on the radio horizon

As radio waves travel more slowly in dense media and the densest part of the atmosphere is normally the lowest, the upper parts of a wave usually travel faster than the lower. This refraction (*Figure 1.7*) has the effect of bending the wave to follow the curvature of the earth and progressively tilting the wavefront until eventually the wave becomes horizontally polarized and short-circuited by the earth's conductivity.

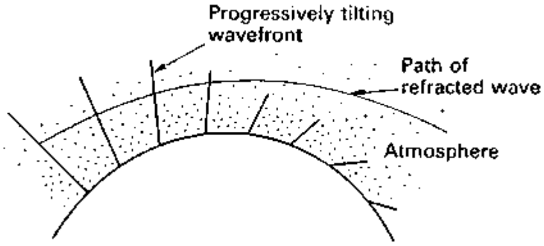

Figure 1.7 Effects of refraction

Waves travelling above the earth's surface (space waves) are usually refracted downwards, effectively increasing the radio horizon to greater than the visual.

The refractive index of the atmosphere is referred to as the *K* factor; a *K* factor of 1 indicates zero refraction. Most of the time *K* is positive at 1.33 and the wave is bent to follow the earth's curvature. The radio horizon is then 4/3 times the visual. However, the density of the atmosphere varies from time to time and in different parts of the world. Density inversions where higher density air is above a region of low density may also occur. Under these conditions the *K* factor is negative and radio waves are bent away from the earth's surface and are either lost or ducting occurs. A *K* factor of 0.7 is the worst expected case.

Ducting occurs when a wave becomes trapped between layers of differing density only to be returned at a great distance from its source, possibly creating interference.

Radio horizon distance at VHF/UHF

The radio horizon at VHF/UHF and up is approximately 15% further than the optical horizon. Several equations are used in calculating the distance. If D^{-} is the distance to the radio horizon, and H is the antenna height, then:

$$D = k\sqrt{H}$$

- When D is in statute miles (5280 feet) and H in feet, then $k = 1.42$.
- When D is in nautical miles (6000 feet) and H in feet, then $k = 1.23$.
- When D is in kilometres and H is in metres, then $k = 4.12$.

Repeating the calculation for the receiving station and adding the results gives the total path length.

Diffraction

When a wave passes over on the edge of an obstacle some of its energy is bent in the direction of the obstacle to provide a signal in what would otherwise be a shadow. The bending is most severe when the wave passes over a sharp edge (*Figure 1.8*).

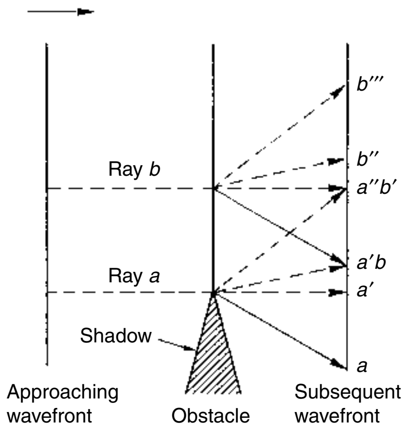

Figure 1.8 Effects of diffraction

As with light waves, the subsequent wavefront consists of wavelets produced from an infinite number of points on the wavefront, rays *a* and *b* in *Figure 1.8* (Huygens' principle). This produces a pattern of interfering waves of alternate addition and subtraction.

Reflection

Radio waves are reflected from surfaces lying in and along their path and also, effectively, from ionized layers in the ionosphere – although most of the reflections from the ionized layers are actually the products of refraction. The strength of truly reflected signals increases with frequency, and the conductivity and smoothness of the reflecting surface.

Multi-path propagation

Reflection, refraction and diffraction may provide signals in what would otherwise be areas of no signal, but they also produce interference.

Reflected – or diffracted – signals may arrive at the receiver in any phase relationship with the direct ray and with each other. The relative phasing of the signals depends on the differing lengths of their paths and the nature of the reflection.

When the direct and reflected rays have followed paths differing by an odd number of half-wavelengths they could be expected to arrive at the receiver in anti-phase with a cancelling effect. However, in the reflection process a further phase change normally takes place. If the reflecting surface had infinite conductivity, no losses would occur in the reflection, and the reflected wave would have exactly the same or opposite phase as the incident wave depending on the polarization in relation to the reflecting surface. In practice, the reflected wave is of smaller amplitude than the incident, and the phase relationships are also changed. The factors affecting the phasing are complex but most frequently, in practical situations, approximately 180° phase change occurs on reflection, so that reflected waves travelling an odd number of half-wavelengths arrive in phase with the direct wave while those travelling an even number arrive anti-phase.

As conditions in the path between transmitter and receiver change so does the strength and path length of reflected signals. This means that a receiver may be subjected to signal variations of almost twice the mean level and practically zero, giving rise to severe fading. This type of fading is frequency selective and occurs on troposcatter systems and in the mobile environment where it is more severe at higher frequencies. A mobile receiver travelling through an urban area can receive rapid signal fluctuations caused by additions and cancellations of the direct and reflected signals at half-wavelength intervals. Fading due to the multi-path environment is often referred to as Rayleigh fading and its effect is shown in *Figure 1.9*. Rayleigh fading, which can cause short signal dropouts, imposes severe restraints on mobile data transmission.

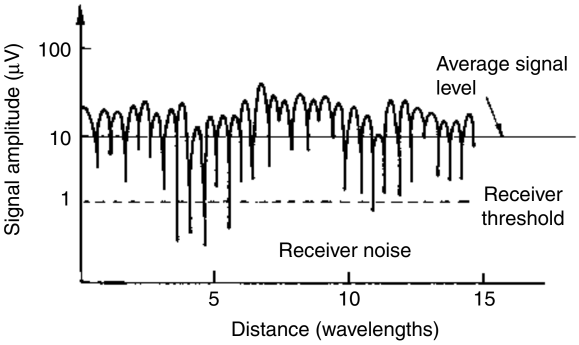

Figure 1.9 Multi-path fading

Noise

The quality of radio signals is not only degraded by the propagation losses: natural or manmade electrical noise is added to them, reducing their intelligibility.

Atmospheric noise includes static from thunderstorms which, unless very close, affects frequencies below about 30 MHz and noise from space is apparent at frequencies between about 8 MHz to 1.5 GHz.

A type of noise with which radio engineers are continually concerned is thermal. Every resistor produces noise spread across the whole frequency spectrum. Its magnitude depends upon the ohmic value of the resistor, its temperature and the bandwidth of the following circuits. The noise voltage produced by a resistor is given by:

$$E_n = \sqrt{4kTBR}$$

where
E_n = noise voltage, V(RMS)
k = Boltzmann's constant
= 1.38×10^{-23} joules/kelvin
T = temperature in degrees K
B = bandwidth of measurement, Hz
R = resistance in ohms

An antenna possesses resistance and its thermal noise, plus that of a receiver input circuit, is a limiting factor to receiver performance.

Noise is produced in every electronic component. Shot noise – it sounds like falling lead shot – caused by the random arrival of electrons at, say, the collector of a transistor, and the random division of electrons at junctions in devices, add to this noise.

Doppler effect

Doppler effect is an apparent shift of the transmitted frequency which occurs when either the receiver or transmitter is moving. It becomes significant in mobile radio applications towards the higher end of the UHF band and on digitally modulated systems.

When a mobile receiver travels directly towards the transmitter each successive cycle of the wave has less distance to travel before reaching the receiving antenna and, effectively, the received frequency is raised. If the mobile travels away from the transmitter, each successive cycle has a greater distance to travel and the frequency is lowered. The variation in frequency depends on the frequency of the wave, its propagation velocity and the velocity of the vehicle containing the receiver. In the situation where the velocity of the vehicle is small compared with the velocity of light, the frequency shift when moving directly towards, or away from, the transmitter is given to sufficient accuracy for most purposes by:

$$f_{\mathrm{d}} = \frac{V}{C} f_{\mathrm{t}}$$

where
f_{d} = frequency shift, Hz
f_{t} = transmitted frequency, Hz
V = velocity of vehicle, m/s
C = velocity of light, m/s

Examples are:

- 100 km/hr at 450 MHz, frequency shift = 41.6 Hz
- 100 km/hr at 1.8 GHz – personal communication network (PCN) frequencies – frequency shift = 166.5 Hz
- Train at 250 km/hr at 900 MHz – a requirement for the GSM pan-European radio-telephone – frequency shift = 208 Hz

When the vehicle is travelling at an angle to the transmitter the frequency shift is reduced. It is calculated as above and the result multiplied by the cosine of the angle of travel from the direct approach (*Figure 1.10*).

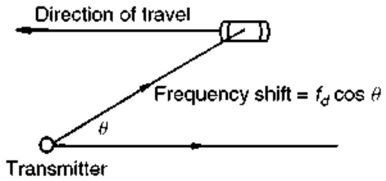
Direction of travel
Frequency shift = $f_d \cos \theta$
θ
Transmitter

Figure 1.10 Doppler frequency shift and angle to transmitter

In a radar situation Doppler effect occurs on the path to the target and also to the reflected signal so that the above formula is modified to:

$$f_{\mathrm{d}} = \frac{2V}{C} f_{\mathrm{t}}$$

where f_{d} is now the total frequency shift.

1.6 Methods of propagation

The effects of all of the above phenomena vary with frequency and are used in the selection of frequencies for specific purposes. The behaviour of waves of different frequencies gives rise to the principal types of wave propagation.

Ground wave propagation

Waves in the bands from very low frequencies (VLF, 3–30 kHz), low frequencies (LF, 30–300 kHz) and medium frequencies (MF, 300–3000 kHz) travel close to the earth's surface: the ground wave (*Figure 1.11*). Transmissions using the ground wave must be vertically polarized to avoid the conductivity of the earth short-circuiting the electric field.

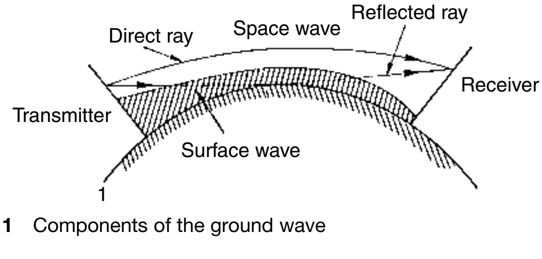
Reflected ray
Direct ray
Space wave
Receiver
Transmitter
Surface wave
1

Figure 1.11 Components of the ground wave

The ground wave consists of a surface wave and a space wave. The surface wave travels along the earth's surface, and is attenuated by ground absorption and the tilting of the wavefront due to diffraction.

The losses increase with frequency and thus VLF radio stations have a greater range than MF stations. The attenuation is partially offset by the replacement of energy from the upper part of the wave refracted by the atmosphere.

The calculation of the field strength of the surface wave at a distance from a transmitter is complex and affected by several variables. Under plane earth conditions and when the distance is sufficiently short that the earth's curvature can be neglected the field intensity is given by:

$$E_{\mathrm{su}} = \frac{2E_0}{d}A$$

where

E_{su} = field intensity in same units as E_0
d = distance in same units of distance as used in E_0
A = a factor calculated from the earth losses, taking frequency, dielectric constant and conductivity into account
E_0 = the free space field produced at unit distance from the transmitter. (With a short (compared with $\lambda/4$) vertical aerial, $2E_0 = 300\sqrt{P}$ mV/m at 1 km where P is the radiated power in kW.) (Terman, 1943)

For a radiated power of 1 kW and ground of average dampness, the distance at which a field of 1 mV/m will exist is given in *Table 1.2*.

Table 1.2 Distance at which a field of 1 mV/m will exist for a radiated power of 1 kW and ground of average dampness

Frequency	*Range (km)*
100 kHz	200
1 MHz	60
10 MHz	6
100 MHz	1.5

The direct and reflected components of the ground wave produce multi-path propagation and variations in received single strength will arise depending on the different path lengths taken by the two components. When the transmitting and receiving antennas are at ground level the components of the space wave cancel each other and the surface wave is dominant. When the antennas are elevated, the space wave becomes increasingly strong and a height is eventually reached where the surface wave has a negligible effect on the received signal strength.

Sky wave propagation

High frequency (HF) waves between 3 MHz and 30 MHz are effectively reflected by ionized layers in the ionosphere producing the sky wave. Medium frequency waves may also be reflected, but less reliably.

The ionosphere contains several layers of ionized air at varying altitudes (*Figure 1.12*). The heights and density of the layers vary diurnally, seasonally and with the incidence of sunspot activity. The E and F_2 layers are semi-permanent while the F_1 layer is normally only present during daytime.

Radio waves radiated at a high angle and reflected by these layers return to earth at a distance from the transmitter. The HF reflection

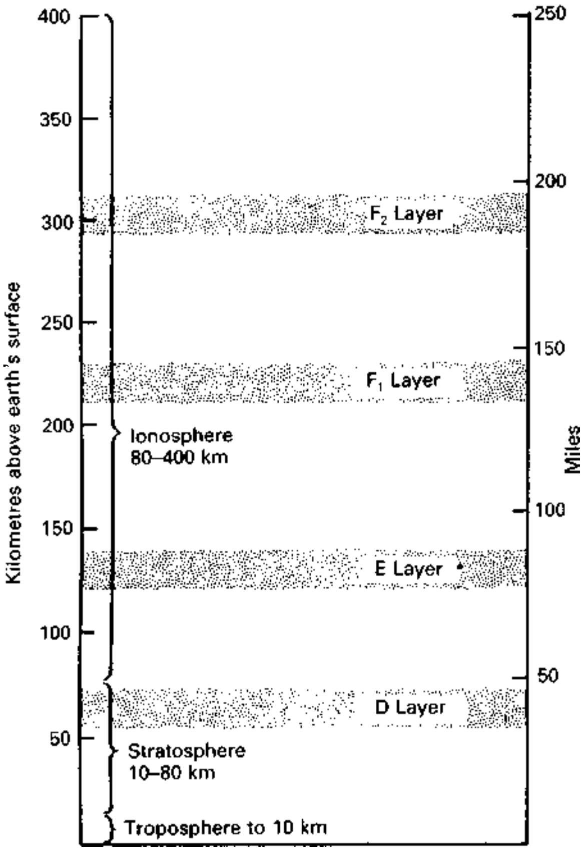
Kilometres above earth's surface
400
350
300
250
200
150
100
50
Miles
250
200
150
100
50
F2 Layer
F1 Layer
E Layer
D Layer
Ionosphere
80–400 km
Stratosphere
10–80 km
Troposphere to 10 km

Figure 1.12 The ionosphere

process is in reality one of refraction in layers possessing a greater free electron density than at heights above or below them. The speed of propagation is slowed on entering a layer and the wave is bent and, if of a suitable frequency and angle of incidence, returned to earth (*Figure 1.13*). The terms used are defined as follows:

- *Virtual height*. The height at which a true reflection of the incident wave would have occurred (*Figure 1.13*).
- *Critical frequency* (f_c). The highest frequency that would be returned to earth in a wave directed vertically at the layer.
- *Maximum usable frequency* (muf). The highest frequency that will be returned to earth for a given angle of incidence. If the angle of incidence to the normal is θ, the muf $= f_c/\cos\theta$.
- *Skip distance*. The minimum distance from the transmitter, along the surface of the earth, at which a wave above the critical frequency will be returned to earth (*Figure 1.12*). Depending on the frequency, the ground wave will exist at a short distance from the transmitter.
- *Sporadic E-layer reflections*. Reflections from the E layer at frequencies higher than those which would normally be returned to earth. They appear to be reflections from electron clouds having sharp boundaries and drifting through the layer. As the name implies the reflections are irregular but occur mostly in summer and at night.

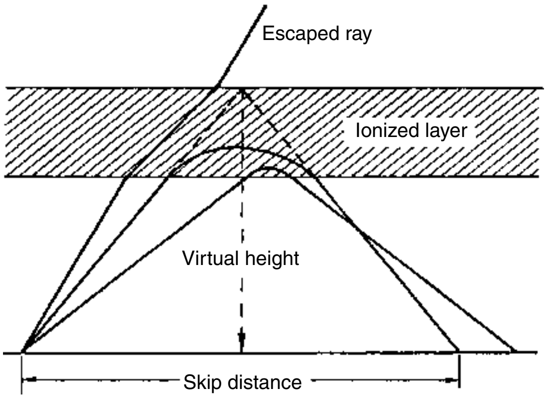

Figure 1.13 Sky wave propagation

Space wave propagation

The space wave travels through the troposphere (the atmosphere below the ionosphere) between the transmitter and the receiver. It contains

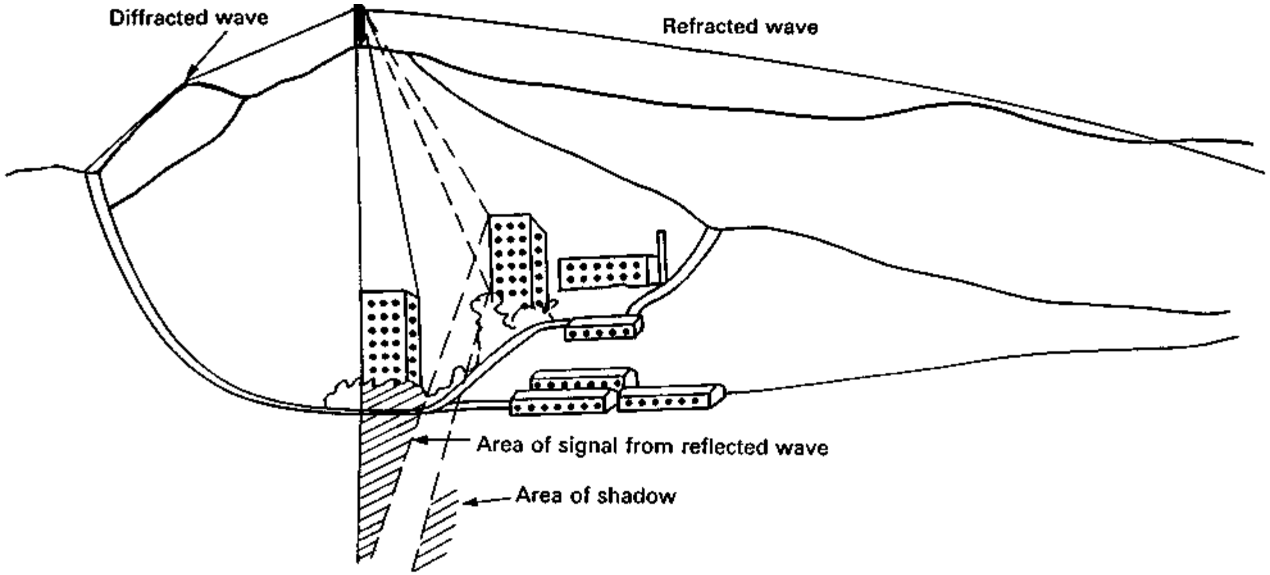
Diffracted wave
Refracted wave
Area of signal from reflected wave
Area of shadow

Figure 1.14 Pictorial representation of radio coverage from a base station

both direct and reflected components (see *Figure 1.11*), and is affected by refraction and diffraction. The importance of these effects varies with frequency, the nature of the terrain and of objects close to the direct path, and the type of communication, e.g. data. Apart from medium-wave broadcasting, space waves are used mainly for communications using frequencies of VHF and upwards.

The range of space waves is the radio horizon. However, places of little or no signal can arise in the lee of radio obstacles. Fortunately, they may be filled with either reflected or diffracted signals as depicted in *Figure 1.14*.

Tropospheric scatter

The tropospheric, or forward, scatter effect provides reliable, over the horizon, communication between fixed points at bands of ultra and super high frequencies. Usable bands are around 900, 2000 and 5000 MHz and path lengths of 300 to 500 km are typical.

The mechanism is not known with certainty but reflections from discontinuities in the dielectric constant of the atmosphere and scattering of the wave by clouds of particles are possibilities. It is an inefficient process, the scattered power being −60 to −90 dB relative to the incident power, so high transmitter powers are necessary. The phenomenon is regularly present but is subject to two types of fading. One occurs slowly and is due to variations of atmospheric conditions. The other is a form of Rayleigh fading and is rapid, deep and frequency selective. It is due to the scattering occurring at different points over a volume in the atmosphere producing multipath propagation conditions.

Troposcatter technique uses directional transmitting and receiving antennas aimed so that their beams intercept in the troposphere at the mid-distance point. To overcome the fading, diversity reception using multiple antennas spaced over 30 wavelengths apart is common.

1.7 Other propagation topics

Communications in the VHF through microwave regions normally takes place on a 'line-of-sight' basis where the radio horizon defines the limit of sight. In practice, however, the situation is not so neat and simple. There is a transition region between the HF and VHF where long distance ionospheric 'skip' occurs only occasionally. This effect is seen above 25 MHz, and is quite pronounced in the 50 MHz region. Sometimes the region behaves like line-of-sight VHF, and at others like HF shortwave.

1.7.1 Scatter

There are a number of scatter modes of propagation. These modes can extend the radio horizon a considerable amount. Where the radio horizon might be a few tens of kilometres, underscatter modes permit very much longer propagation. For example, a local FM broadcaster at 100 MHz might have a service area of about 40 miles, and might be heard 180 miles away during the summer months when *Sporadic-E* propagation occurs. One summer, a television station in Halifax, Nova Scotia, Canada, was routinely viewable in Washington, DC in the United States during the early morning hours for nearly a week.

Sporadic-E is believed to occur when a small region of the atmosphere becomes differentially ionized, and thereby becomes a species of 'radio mirror'. Ionospheric scatter propagation occurs when clouds of ions exist in the atmosphere. These clouds can exist in both the ionosphere and the troposphere, although the tropospheric model is more reliable for communications. A signal impinging this region may be scattered towards other terrestrial sites which may be a great distance away. The specific distance depends on the geometry of the scenario.

There are at least three different modes of scatter from ionized clouds: *back scatter, side scatter*, and *forward scatter*. The back scatter mode is a bit like radar, in that signal is returned back to the transmitter site, or in regions close to the transmitter. Forward scatter occurs when the reflected signal continues in the same azimuthal direction (with respect to the transmitter), but is redirected toward the Earth's surface. Side scatter is similar to forward scatter, but the azimuthal direction might change.

Unfortunately, there are often multiple reflections from the ionized cloud, and these are shown as 'multiple scatter' in *Figure 1.15*. When these reflections are able to reach the receiving site, the result is a rapid, fluttery fading that can be of quite profound depths.

Meteor scatter is used for communication in high latitude regions. When a meteor enters the Earth's atmosphere it leaves an ionized trail of air behind it. This trail might be thousands of kilometres long, but is very short lived. Radio signals impinging the tubular metre ion trail are reflected back towards Earth. If the density of meteors in the critical region is high, then more or less continuous communications can be achieved. This phenomenon is noted in the low VHF between 50 and about 150 MHz. It can easily be observed on the FM broadcast band if the receiver is tuned to distant stations that are barely audible. If the geometry of the scenario is right, abrupt but short-lived peaks in the signal strength will be noted.

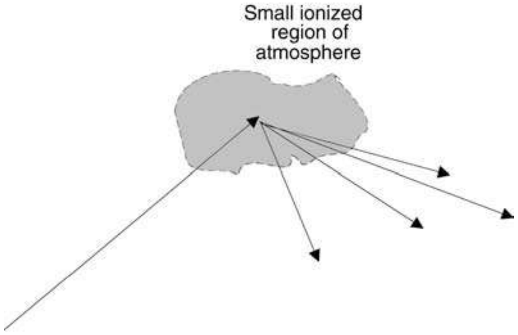

Figure 1.15 Multiple scatter

1.7.2 Refraction modes

Refraction is the mechanism for most tropospheric propagation phenomena. The dielectric properties of the air, which are set mostly by the moisture content, are a primary factor in tropospheric refraction. Refraction occurs in both light or radio wave systems when the wave passes between mediums of differing density. Under that situation, the wave path will bend an amount proportional to the difference in density of the two regions.

The general situation is typically found at UHF and microwave frequencies. Because air density normally decreases with altitude, the top of a beam of radio waves typically travels slightly faster than the lower portion of the beam. As a result, those signals refract a small amount. Such propagation provides slightly longer surface distances than are normally expected from calculating the distance to the radio horizon. This phenomenon is called *simple refraction*, and is described by the *K factor*.

Super refraction

A special case of refraction called super refraction occurs in areas of the world where warmed land air flows out over a cooler sea (*Figure 1.16*). Examples of such areas are deserts that are adjacent to a large body of water: the Gulf of Aden, the southern Mediterranean, and the Pacific Ocean off the coast of Baja, California. Frequent VHF/UHF/microwave communications up to 200 miles are reported in such areas, and up to 600 miles have reportedly been observed.

Ducting

Another form of refraction phenomenon is weather related. Called *ducting*, this form of propagation is actually a special case of super

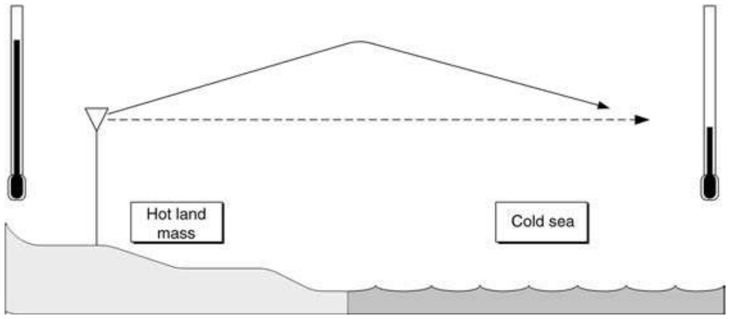

Figure 1.16 An example of super refraction

refraction. Evaporation of sea water causes temperature inversion regions to form in the atmosphere. That is, layered air masses in which the air temperature is greater than in the layers below it (note: air temperature normally decreases with altitude, but at the boundary with an inversion region, it begins to increase). The inversion layer forms a 'duct' that acts similarly to a waveguide. Ducting allows long distance communications from lower VHF through microwave frequencies; with 50 MHz being a lower practical limit, and 10 GHz being an ill-defined upper limit. Airborne operators of radio, radar, and other electronic equipment can sometimes note ducting at even higher microwave frequencies.

Antenna placement is critical for ducting propagation. Both the receiving and transmitting antennas must be either: (a) physically inside the duct (as in airborne cases), or (b) able to propagate at an angle such that the signal gets trapped inside the duct. The latter is a function of antenna radiation angle. Distances up to 2500 miles or so are possible through ducting.

Certain paths, where frequent ducting occurs, have been identified: in the United States, the Great Lakes region to the southeastern Atlantic seaboard; Newfoundland to the Canary Islands; across the Gulf of Mexico from Florida to Texas; Newfoundland to the Carolinas; California to Hawaii; and Ascension Island to Brazil.

Subrefraction

Another refractive condition is noted in the polar regions, where colder air from the land mass flows out over warmer seas (*Figure 1.17*). Called *subrefraction*, this phenomena bends EM waves away from the Earth's surface – thereby reducing the radio horizon by about 30 to 40%.

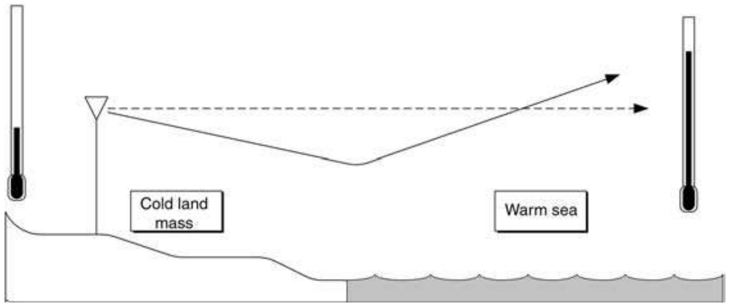

Figure 1.17 An example of subrefraction

All tropospheric propagation that depends upon air-mass temperatures and humidity shows diurnal (i.e. over the course of the day) variation caused by the local rising and setting of the sun. Distant signals may vary 20 dB in strength over a 24-hour period. These tropospheric phenomena explain how TV, FM broadcast, and other VHF signals can propagate great distances, especially along seacoast paths, sometimes weak and sometimes nonexistent.

1.7.3 Great circle paths

A great circle is the shortest line between two points on the surface of a sphere, such that it lays on a plane through the Earth's centre and includes the two points. When translated to 'radiospeak', a great circle is the shortest path on the surface of the Earth between two points. Navigators and radio operators use the great circle for similar, but different, reasons. Navigators use it in order to get from here to there, and radio operators use it to get a transmission path from here to there.

The heading of a directional antenna is normally aimed at the receiving station along its great circle path. Unfortunately, many people do not understand the concept well enough, for they typically aim the antenna in the wrong direction. For example, Washington, DC in the USA is on approximately the same latitude as Lisbon, Portugal. If you fly due east, you will have dinner in Lisbon, right? Wrong. If you head due east from Washington, DC, across the Atlantic, the first landfall would be West Africa, somewhere between Ghana and Angola. Why? Because the great circle bearing 90 degrees takes us far south. The geometry of spheres, not flat planes, governs the case.

Long path versus short path

The Earth is a sphere (or, more precisely, an 'oblique spheroid'), so from any given point to any other point there are two great circle paths: the long path (major arc) and the short path (minor arc). In general, the best reception occurs along the short path. In addition, short path propagation is more nearly ‘textbook’, compared with long path reception. However, there are times when long path is better, or is the only path that will deliver a signal to a specific location from the geographic location in question.

Grey line propagation

The Grey line is the twilight zone between the night and daytime halves of the earth. This zone is also called the *planetary terminator*. It varies up to +23 degrees either side of the north–south longitudinal lines, depending on the season of the year (it runs directly north–south only at the vernal and autumnal equinoxes). The D-layer of the ionosphere absorbs signals in the HF region. This layer disappears almost completely at night, but it builds up during the day. Along the grey line, the D-layer is rapidly decaying west of the line, and has not quite built up east of the line.

Brief periods of abnormal propagation occur along the grey line. Stations on either side of the line can be heard from regions, and at distances, that would otherwise be impossible on any given frequency. For this reason, radio operators often prefer to listen at dawn and dusk for this effect.

1.7.4 Scatter propagation modes

Auroral propagation

The auroral effect produces a luminescence in the upper atmosphere resulting from bursts of particles released from the sun 18 to 48 hours earlier. The light emitted is called the northern lights and the southern lights. The ionized regions of the atmosphere that create the lights form a radio reflection shield, especially at VHF and above, although 15 to 20 MHz effects are known. Auroral propagation effects are normally seen in the higher latitudes, although listeners in the southern tier of states in the USA and Europe are often treated to the reception of signals from the north being reflected from auroral clouds. Similar effects exist in the southern hemisphere.

Non-reciprocal direction

If you listen to the 40 metre (7–7.3 MHz) amateur radio band receiver on the East Coast of the United States, you will sometimes hear European stations – especially in the late afternoon. But when the

US amateur tries to work those European stations there is no reply whatsoever. The Europeans are unable to hear the US stations. This propagation anomaly causes the radio wave to travel different paths dependent on which direction it travels, i.e. an east–west signal is not necessarily the reciprocal of a west–east signal. This anomaly can occur when a radio signal travels through a heavily ionized medium in the presence of a magnetic field, which is exactly the situation when the signal travels through the ionosphere in the presence of the Earth's magnetic field.

References

Belcher, R. *et al*. (1989). *Newnes Mobile Radio Servicing Handbook*. Butterworth-Heinemann, Oxford.

Kennedy, G. (1977). *Electronic Communications Systems*. McGraw-Hill Kogashuka, Tokyo.

Terman, F.E. (1943). *Radio Engineers' Handbook*. McGraw-Hill, London.

2 The decibel scale

2.1 Decibels and the logarithmic scale

The range of powers, voltages and currents encountered in radio engineering is too wide to be expressed on linear scale. Consequently, a logarithmic scale based on the decibel (dB, one tenth of a bel) is used. The decibel does not specify a magnitude of a power, voltage or current but a ratio between two values of them. Gains and losses in circuits or radio paths are expressed in decibels.

The ratio between two powers is given by:

$$\text{Gain or loss, dB} = 10\log_{10}\frac{P_1}{P_2}$$

where P_1 and P_2 are the two powers.

As the power in a circuit varies with the square of the voltage or current, the logarithm of the ratio of these quantities must be multiplied by twenty instead of ten. To be accurate the two quantities under comparison must operate in identical impedances:

$$\text{Gain or loss, dB} = 20\log_{10}\frac{V_1}{V_2}$$

To avoid misunderstandings, it must be realized that a ratio of 6 dB is 6 dB regardless of whether it is power, voltage or current that is referred to: if it is power, the ratio for 6 dB is four times; if it is voltage or current, the ratio is two times (*Table 2.1*).

2.2 Decibels referred to absolute values

While the decibel scale expresses ratios only, if a reference value is added to the statement as a suffix it can be used to refer to absolute values. For example, a loss of 10 dB means a reduction in power to a level equal to one tenth of the original and if the statement is −10 dBm the level referred to is 1/10 of a milliwatt. Commonly used suffixes and, where applicable, their absolute reference levels are as follows. *Table 2.2* shows the relative levels in decibels at 50 ohms impedance.

Table 2.1 The decibel figures are in the centre column: figures to the left represent decibel loss, and those to the right decibel gain. The voltage and current figures are given on the assumption that there is no change in impedance

Voltage or current ratio	*Power ratio*	*dB* ← − + →	*Voltage or current ratio*	*Power ratio*
1.000	1.000	0	1.000	1.000
0.989	0.977	0.1	1.012	1.023
0.977	0.955	0.2	1.023	1.047
0.966	0.933	0.3	1.035	1.072
0.955	0.912	0.4	1.047	1.096
0.944	0.891	0.5	1.059	1.122
0.933	0.871	0.6	1.072	1.148
0.912	0.832	0.8	1.096	1.202
0.891	0.794	1.0	1.122	1.259
0.841	0.708	1.5	1.189	1.413
0.794	0.631	2.0	1.259	1.585
0.750	0.562	2.5	1.334	1.778
0.708	0.501	3.0	1.413	1.995
0.668	0.447	3.5	1.496	2.239
0.631	0.398	4.0	1.585	2.512
0.596	0.355	4.5	1.679	2.818
0.562	0.316	5.0	1.778	3.162
0.501	0.251	6.0	1.995	3.981
0.447	0.200	7.0	2.239	5.012
0.398	0.159	8.0	2.512	6.310
0.355	0.126	9.0	2.818	7.943
0.316	0.100	10	3.162	10.00
0.282	0.0794	11	3.55	12.6
0.251	0.0631	12	3.98	15.9
0.224	0.0501	13	4.47	20.0
0.200	0.0398	14	5.01	25.1
0.178	0.0316	15	5.62	31.6
0.159	0.0251	16	6.31	39.8
0.126	0.0159	18	7.94	63.1
0.100	0.0100	20	10.00	100.0
3.16×10^{-2}	10^{-3}	30	3.16×10	10^{3}
10^{-2}	10^{-4}	40	10^{2}	10^{4}
3.16×10^{-3}	10^{-5}	50	3.16×10^{2}	10^{5}
10^{-3}	10^{-6}	60	10^{3}	10^{6}
3.16×10^{-4}	10^{-7}	70	3.16×10^{3}	10^{7}
10^{-4}	10^{-8}	80	10^{4}	10^{8}
3.16×10^{-5}	10^{-9}	90	3.16×10^{4}	10^{9}
10^{-5}	10^{-10}	100	10^{5}	10^{10}
3.16×10^{-6}	10^{-11}	110	3.16×10^{5}	10^{11}
10^{-6}	10^{-12}	120	10^{6}	10^{12}

Table 2.2 Relative levels in decibels at 50 ohms impedance

dB μV	*Voltage*	*dBV*	*dBm*	*dBW*	*Power*
−20	100 nV	−140	−127	−157	200 aW
−19	115	−139	−126	−156	250
−18	125	−138	−125	−155	315
−17	140	−137	−124	−154	400
−16	160	−136	−123	−153	500
−15	180	−135	−122	−152	630
−14	200	−134	−121	−151	800
−13	225	−133	−120	−150	1 fW
−12	250	−132	−119	−149	1.25
−11	280	−131	−118	−148	1.6
−10	315	−130	−117	−147	2.0
−9	355	−129	−116	−146	2.5
−8	400	−128	−115	−145	3.15
−7	450	−127	−114	−144	4.0
−6	500	−126	−113	−143	5.0
−5	565	−125	−112	−142	6.3
−4	630	−124	−111	−141	8.0
−3	710	−123	−110	−140	10
−2	800	−122	−109	−139	12.5
−1	900	−121	−108	−138	16
0	1 μV	−120	−107	−137	20
1	1.15	−119	−106	−136	25
2	1.25	−118	−105	−135	31.5
3	1.4	−117	−104	−134	40
4	1.6	−116	−103	−133	50
5	1.8	−115	−102	−132	63
6	2.0	−114	−101	−131	80
7	2.25	−113	−100	−130	100
8	2.5	−112	−99	−129	125
9	2.8	−111	−98	−128	160
10	3.15	−110	−97	−127	200
11	3.55	−109	−96	−126	250
12	4.0	−108	−95	−125	315
13	4.5	−107	−94	−124	400
14	5.0	−106	−93	−123	500
15	5.65	−105	−92	−122	630
16	6.3	−104	−91	−121	800
17	7.1	−103	−90	−120	1 pW
18	8.0	−102	−89	−119	1.25
19	9.0	−101	−88	−118	1.6
20	10.0	−100	−87	−117	2.0
21	11.5	−99	−86	−116	2.5
22	12.5	−98	−85	−115	3.15
23	14	−97	−84	−114	4.0
24	16	−96	−83	−113	5.0
25	18	−95	−82	−112	6.3

(*continued overleaf*)

Table 2.2 (*continued*)

dB μV	*Voltage*	*dBV*	*dBm*	*dBW*	*Power*
26	20	−94	−81	−111	8.0
27	22.5	−93	−80	−110	10
28	25	−92	−79	−109	12.5
29	28	−91	−78	−108	16
30	31.5	−90	−77	−107	20
31	35.5	−89	−76	−106	25
32	40	−88	−75	−105	31.5
33	45	−87	−74	−104	40
34	50	−86	−73	−103	50
35	56.5	−85	−72	−102	63
36	63	−84	−71	−101	80
37	71	−83	−70	−100	100
38	80	−82	−69	−99	125
39	90	−81	−68	−98	160
40	100	−80	−67	−97	200
41	115	−79	−66	−96	250
42	125	−78	−65	−95	315
43	140	−77	−64	−94	400
44	160	−76	−63	−93	500
45	180	−75	−62	−92	630
46	200	−74	−61	−91	800
47	225	−73	−60	−90	1 nW
48	250	−72	−59	−89	1.25
49	280	−71	−58	−88	1.6
50	315	−70	−57	−87	2.0
51	355	−69	−56	−86	2.5
52	400	−68	−55	−85	3.15
53	450	−67	−54	−84	4.0
54	500	−66	−53	−83	5.0
55	565	−65	−52	−82	6.3
56	630	−64	−51	−81	8.0
57	710	−63	−50	−80	10
58	800	−62	−49	−79	12.5
59	900	−61	−48	−78	16
60	1 mV	−60	−47	−77	20
61	1.15	−59	−46	−76	25
62	1.25	−58	−45	−75	31.5
63	1.4	−57	−44	−74	40
64	1.6	−56	−43	−73	50
65	1.8	−55	−42	−72	63
66	2.0	−54	−41	−71	80
67	2.25	−53	−40	−70	100
68	2.5	−52	−39	−69	125
69	2.8	−51	−38	−68	160
70	3.15	−50	−37	−67	200
71	3.55	−49	−36	−66	250
72	4.0	−48	−35	−65	315

Table 2.2 (*continued*)

dB µV	*Voltage*	*dBV*	*dBm*	*dBW*	*Power*
73	4.5	−47	−34	−64	400
74	5.0	−46	−33	−63	500
75	5.65	−45	−32	−62	630
76	6.3	−44	−31	−61	800
77	7.1	−43	−30	−60	1 µW
78	8.0	−42	−29	−59	1.25
79	9.0	−41	−28	−58	1.6
80	10 mV	−40	−27	−57	2.0
81	11.5	−39	−26	−56	2.5
82	12.5	−38	−25	−55	3.15
83	14	−37	−24	−54	4.0
84	16	−36	−23	−53	5.0
85	18	−35	−22	−52	6.3
86	20	−34	−21	−51	8.0
87	22.5	−33	−20	−50	10
88	25	−32	−19	−49	12.5
89	28	−31	−18	−48	16
90	31.5	−30	−17	−47	20
91	35.5	−29	−16	−46	25
92	40	−28	−15	−45	31.5
93	45	−27	−14	−44	40
94	50	−26	−13	−43	50
95	56.5	−25	−12	−42	63
96	63	−24	−11	−41	80
97	71	−23	−10	−40	100
98	80	−22	−9	−39	125
99	90	−21	−8	−38	160
100	100	−20	−7	−37	200
101	115	−19	−6	−36	250
102	125	−18	−5	−35	315
103	140	−17	−4	−34	400
104	160	−16	−3	−33	500
105	180	−15	−2	−32	630
106	200	−14	−1	−31	800
107	225	−13	0	−30	1 mW
108	250	−12	1	−29	1.25
109	280	−11	2	−28	1.6
110	315	−10	3	−27	2.0
111	355	−9	4	−26	2.5
112	400	−8	5	−25	3.15
113	450	−7	6	−24	4.0
114	500	−6	7	−23	5.0
115	565	−5	8	−22	6.3
116	630	−4	9	−21	8.0
117	710	−3	10	−20	10
118	800	−2	11	−19	12.5

(*continued overleaf*)

Table 2.2 (*continued*)

dB μV	*Voltage*	*dBV*	*dBm*	*dBW*	*Power*
119	900	−1	12	−18	16
120	1 V	0	13	−17	20
121	1.15	1	14	−16	25
122	1.25	2	15	−15	31.5
123	1.4	3	16	−14	40
124	1.6	4	17	−13	50
125	1.8	5	18	−12	63
126	2.0	6	19	−11	80
127	2.25	7	20	−10	100
128	2.5	8	21	−9	125
129	2.8	9	22	−8	160
130	3.15	10	23	−7	200
131	3.55	11	24	−6	250
132	4.0	12	25	−5	315
133	4.5	13	26	−4	400
134	5.0	14	27	−3	500
135	5.65	15	28	−2	630
136	6.3	16	29	−1	800
137	7.1	17	30	0	1 W
138	8.0	18	31	1	1.25
139	9.0	19	32	2	1.6
140	10	20	33	3	2.0
141	11.5	21	34	4	2.5
142	12.5	22	35	5	3.15
143	14	23	36	6	4.0
144	16	24	37	7	5.0
145	18	25	38	8	6.3
146	20	26	39	9	8.0
147	22.5	27	40	10	10
148	25	28	41	11	12.5
149	28	29	42	12	16
150	31.5	30	43	13	20
151	35.5	31	44	14	25
152	40	32	45	15	31.5
153	45	33	46	16	40
154	50	34	47	17	50
155	56.5	35	48	18	63
156	63	36	49	19	80
157	71	37	50	20	100
158	80	38	51	21	125
159	90	39	52	22	160
160	100	40	53	23	200
161	115	41	54	24	250
162	125	42	55	25	315
163	140	43	56	26	400
164	160	44	57	27	500
165	180	45	58	28	630

Table 2.2 (*continued*)

dB μV	*Voltage*	*dBV*	*dBm*	*dBW*	*Power*
166	200	46	59	29	800
167	225	47	60	30	1 kW
168	250	48	61	31	1.25
169	280	49	62	32	1.6
170	315	50	63	33	2.0
171	355	51	64	34	2.5
172	400	52	65	35	3.15
173	450	53	66	36	4.0
174	500	54	67	37	5.0
175	565	55	68	38	6.3
176	630	56	69	39	8.0
177	710	57	70	40	10
178	800	58	71	41	12.5
179	900	59	72	42	16
180	1 kV	60	73	43	20

Decibel glossary

dBa	stands for dB 'adjusted'. This is a weighted circuit noise power referred to −85 dBm, which is 0 dBa. (Historically measured with a noise meter at the receiving end of a line. The meter is calibrated on a 1000 Hz tone such that 1 mW (0 dBm) gives a reading of +85 dBm. If the 1 mW is spread over the band 300–3400 Hz as random white noise, the meter will read +82 dBa.)
dBa0	circuit noise power in dBa referred to, or measured at, a point of zero relative transmission level (0 dBr). (A point of zero relative transmission level is a point arbitrarily established in a transmission circuit. All other levels are stated relative to this point.) It is preferable to convert circuit noise measurement values from dBa to dBa0 as this makes it unnecessary to know or to state the relative transmission level at the point of measurement.
dBd	used for expressing the gain of an antenna referred to a dipole.
dBi	used for expressing the gain of an antenna referred to an isotropic radiator.
dB μV	decibels relative to 1 microvolt.
dbm	decibels relative to 1 milliwatt. The term dBm was originally used for telephone and audio work and, when used in that context, implies an impedance of 600 Ω, the nominal impedance of a telephone line. When it is desired to define a relative transmission level in a circuit, dBr is preferred.

Table 2.3 Binary decibel values

Bits	*Max. value*	*Decibels (dB)*
1	2	6.02
2	4	12.04
3	8	18.06
4	16	24.08
5	32	30.10
6	64	36.12
7	128	42.14
8	256	48.16
9	512	54.19
10	1024	60.21
11	2048	66.23
12	4096	72.25
13	8192	78.27
14	16 384	84.29
15	32 768	90.31
16	65 536	96.33
17	131 072	102.35
18	262 144	108.37
19	524 288	114.39
20	1 048 576	120.41
21	2 097 152	126.43
22	4 194 304	132.45
23	8 388 608	138.47
24	16 777 216	144.49
25	33 554 432	150.51
26	67 108 864	156.54
27	134 217 728	162.56
28	268 435 456	168.58
29	536 870 912	174.60
30	1 073 741 824	180.62
31	2 147 483 648	186.64
32	4 294 967 296	192.66

dBm0	dBm referred to, or measured at, a point of zero transmission level.
dBmp	a unit of noise power in dBm, measured with psophometric weighting. dBmp $= 10\log_{10}$ pWp $- 90 =$ dBa $- 84 =$ dBm $- 2.5$ (for flat noise 300–3400 Hz). pWp = picowatts psophometrically weighted.
dBm0p	the abbreviation for absolute noise power in dBm referred to or measured at a point of zero relative transmission level, psophometrically weighted.
dBr	means dB ‘relative level’. Used to define transmission level at various points in a circuit or system referenced to the zero transmission level point.

dBrn	a weighted circuit noise power unit in dB referenced to 1 pW (−90 dBm) which is 0 dBrn.
dBrnc	weighted noise power in dBrn, measured by a noise measuring set with 'C-message' weighting.
dBrnc0	noise measured in dBrnc referred to zero transmission level point.
dBu	decibels relative to 0.775 V, the voltage developed by 1 mW when applied to 600 Ω. dBu is used in audio work when the impedance is not 600 Ω and no specific impedance is implied.
dbV	decibels relative to 1 volt.
dbW	decibels relative to 1 watt.

Note: To convert dBm to dB μV add 107 (e.g. $-20\,\text{dBm} = -20 + 107 = 87\,\text{dB}\,\mu\text{V}$.

The beauty of decibel notation is that system gains and losses can be computed using addition and subtraction rather than multiplication and division. For example, suppose a system consists of an antenna that delivers a −4.7 dBm signal at its terminals (we conveniently neglect the antenna gain by this ploy). The antenna is connected to a 40 dB low-noise amplifier (A1) at the head end, and then through a 370 metre long coaxial cable to a 20 dB gain amplifier (A2), with a loss (L1) of −48 dB. The amplifier is followed by a bandpass filter with a −2.8 dB insertion loss (L2), and a −10 dB attenuator (L3). How does the signal exist at the end of this cascade chain?

S1	−4.7 dBm
A1	40.0 dB
A2	20.0 dB
L1	−48.0 dB
L2	−2.8 dB
L3	−10.0 dB
Total:	−5.5 dBm

Converting dBm to watts

$$P = \frac{10^{\text{dBm}/10}}{1000}$$

Converting any dB to ratio

Power levels: $\dfrac{P1}{P2} = 10^{\text{dB}/10}$

Voltage levels: $\frac{V1}{V2} = 10^{\mathrm{dB}/20}$

Current levels: $\frac{I1}{I2} = 10^{\mathrm{dB}/20}$

Binary decibel values

Binary numbers are used in computer systems. With the digitization of RF systems it is necessary to understand the decibel values of binary numbers. These binary numbers might be from an analogue-to-digital converter (ADC or A/D) that digitizes the IF amplifier output, or a digital-to-analogue converter (DAC or D/A) used to generate the analogue signal in a direct digital synthesis (DDS) signal generator.

3 Transmission lines

3.1 General considerations

The purpose of any transmission line is to transfer power between a source and a load with the minimum of loss and distortion in either amplitude, frequency or phase angle.

Electrons travel more slowly in conductors than they do in free space and all transmission lines contain distributed components: resistance, inductance and capacitance. Consequently, lines possess an impedance which varies with frequency, and loss and distortion occur. Because the impedance is not constant over a wide frequency band the insertion loss will not be the same for all frequencies and frequency distortion will arise. A wavefront entering a line from a source takes a finite time to travel its length. This transit time, because of the distributed components, also varies with frequency and creates phase distortion.

3.2 Impedance matching

To transfer the maximum power from a generator into a load the impedance of the load and the internal impedance of the generator – and any intervening transmission line – must be equal.

Figure 3.1 illustrates the simplest case of a generator of internal impedance Z_s equal to 5 ohms and producing an e.m.f. of 20 volts.

When loads of varying impedance, Z_1, are connected the output voltage, V (p.d.) and the power in the load, P_1, varies as follows:

$Z_1 = 5\,\Omega \quad I = 20/10 = 2\,\mathrm{A} \quad V = 10\,\mathrm{V}$
$Z_1 = 3\,\Omega \quad I = 2.5\,\mathrm{A} \quad V = 7.5\,\mathrm{V}$
$Z_1 = 8.33\,\Omega \quad I = 1.5\,\mathrm{A} \quad V = 12.5\,\mathrm{V}$

$P_1 = V^2/\mathrm{Z}_1 = 100/5 = 20\,\mathrm{W}$
$P_1 = V^2/\mathrm{Z}_1 = 56.25/3 = 18.75\,\mathrm{W}$
$P_1 = V^2/\mathrm{Z}_1 = 156.25/8.33 = 18.75\,\mathrm{W}$

When dealing with alternating current and when transmission lines, particularly at radio frequencies, are interposed between the source and the load, other factors than the power transfer efficiency must also be considered.

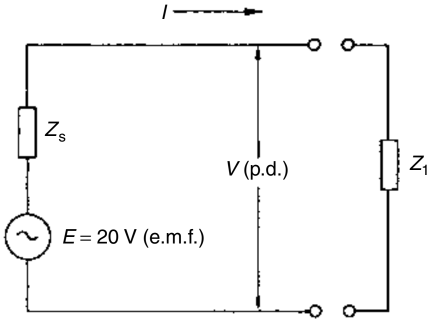

Figure 3.1 Impedance matching

3.3 Base band lines

These are the lines which generally operate at comparatively low frequencies carrying information at base band, e.g. speech, music, video or data. Generally provided by the telecommunications or telephone companies, usually on a rental basis, they are no longer likely to be hard wired, solid copper lines, although these may still be obtainable for lengths below about 10 km within one exchange area. Longer lines will probably be multiplexed, and comprised of radio and optical circuits over part of their length.

Baseband line impedance may vary between 450 Ω and 750 Ω. Nominal impedance is 600 Ω. Most line parameters are specified when measured between 600 Ω non-reactive impedance.

3.4 Balanced line hybrids

Radio transmitters and receivers are often controlled over a two-wire line. To facilitate this a balanced line hybrid circuit, consisting of two transformers connected back to back as in *Figure 3.2*, is inserted between the transmitter and receiver, and the line.

A signal from the receiver audio output is fed to winding L_1 of transformer T_1 which induces voltages across L_2 and L_3. The resultant line current also flows through L_4 and produces a voltage across L_6 which would appear as modulation on the transmitter but for the anti-phase voltage appearing across L_5. To ensure that the voltages cancel exactly a variable resistor, and often a capacitor to equalize the frequency response, is connected between L_2 and L_5.

A signal arriving via the line is applied to the transmitter as modulation; that it is also applied to the receiver poses no problem.

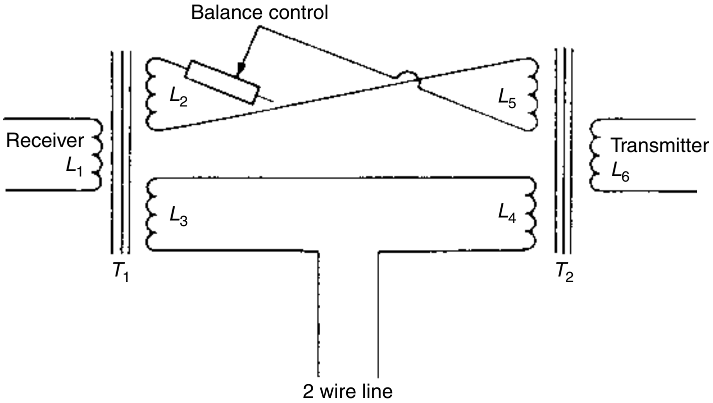

Figure 3.2 Balanced line hybrid

3.5 Radio frequency lines

Radio frequency transmission lines possess similar electrical characteristics to base band lines. However, they may be carrying large powers and the effects of a mismatched load are much more serious than a loss of transferred power. Three types of wire RF line are commonly used: a single wire with ground return for MF and LF broadcast transmission, an open pair of wires at HF and co-axial cable at higher frequencies. Waveguides are used at the higher microwave frequencies. RF lines exhibit an impedance characterized by their type and construction.

3.5.1 Characteristic impedance, Z_0

The physical dimensions of an RF transmission line, the spacing between the conductors, their diameters and the dielectric material between them, determine the characteristic impedance of the line, Z_0, which is calculated for the most commonly used types as follows.

Single wire with ground return (*Figure 3.3(a)*):

$$Z_0 = 138 \log_{10} \frac{4h}{d} \text{ ohms}$$

2-wire balanced, in air (*Figure 3.3(b)*):

$$Z_0 = \frac{276}{\sqrt{\varepsilon_p}} \log_{10} \frac{2s}{d} \text{ ohms}$$

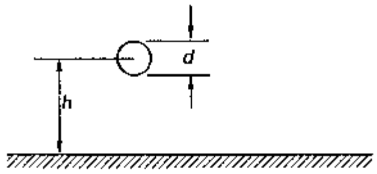

(a) Single wire, ground return line

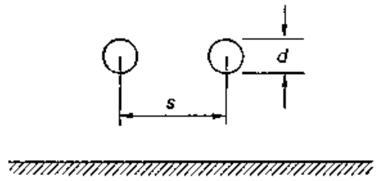

(b) 2-wire balanced air dielectric line

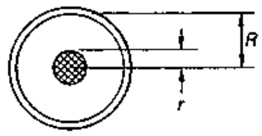

(c) Co-axial cable

Figure 3.3

Co-axial (*Figure 3.3(c)*):

$$Z_0 = \frac{138}{\sqrt{\varepsilon_r}} \log_{10} \frac{R}{r} \text{ ohms}$$

where dimensions are in mm and ε_r = relative permittivity of continuous dielectric.

3.5.2 Insertion loss

The loss in RF cables is quoted in specifications as attenuation in dB per unit length at specific frequencies, the attenuation increasing with frequency. The electrical specifications for cables having a braided outer conductor are given in *Tables 3.1* and *3.2*. Those for the now commonly used foam dielectric, solid outer conductor cables are provided in *Table 3.3*.

Table 3.1 RF cables USA RG and URM series

				Attenuation (dB per 100 ft)						
RG number	*Nominal impedance Z_0 (ohms)*	*Overall diameter (inches)*	*Velocity factor*	*1 MHz*	*10 MHz*	*100 MHz*	*1000 MHz*	*3000 MHz*	*Capacity (pF/ft)*	*Maximum operating voltage RMS*
RG-6A/U	75.0	0.332	0.659	0.21	0.78	2.9	11.2	21.0	20.0	2700
RG-11A/U	75.0	0.405	0.66	0.18	0.7	2.3	7.8	16.5	20.5	5000
RG-12A/U	75.0	0.475	0.659	0.18	0.66	2.3	8.0	16.5	20.5	4000
RG-16/U	52.0	0.630	0.670	0.1	0.4	1.2	6.7	16.0	29.5	6000
RG-34A/U	75.0	0.630	0.659	0.065	0.29	1.3	6.0	12.5	20.5	5200
RG-34B/U	75.0	0.630	0.66	–	0.3	1.4	5.8	–	21.5	6500
RG-35A/U	75.0	0.945	0.659	0.07	0.235	0.85	3.5	8.60	20.5	10 000
RG-54A/U	58.0	0.250	0.659	0.18	0.74	3.1	11.5	21.5	26.5	3000
RG-55/U	53.5	0.206	0.659	0.36	1.3	4.8	17.0	32.0	28.5	1900
RG-55A/U	50.0	0.216	0.659	0.36	1.3	4.8	17.0	32.0	29.5	1900
RG-58/U	53.5	0.195	0.659	0.33	1.25	4.65	17.5	37.5	28.5	1900

(*continued overleaf*)

Table 3.1 (*continued*)

RG number	*Nominal impedance Z_0 (ohms)*	*Overall diameter (inches)*	*Velocity factor*	*Attenuation (dB per 100 ft)*					*Capacity (pF/ft)*	*Maximum operating voltage RMS*
				1 MHz	*10 MHz*	*100 MHz*	*1000 MHz*	*3000 MHz*		
RG-58C/U	50.0	0.195	0.659	0.42	1.4	4.9	24.0	45.0	30.0	1900
RG-59A/U	75.0	0.242	0.659	0.34	1.10	3.40	12.0	26.0	20.5	2300
RG-59B/U	75	0.242	0.66	–	1.1	3.4	12.0	–	21	2300
RG-62A/U	93.0	0.242	0.84	0.25	0.85	2.70	8.6	18.5	13.5	750
RG-83/U	35.0	0.405	0.66	0.23	0.80	2.8	9.6	24.0	44.0	2000
RG-174A/U	50	0.1	0.66	1.83	3.35	11.0	32.0	64.0	30.0	1400
RG-188A/U	50	0.11	0.66	–	–	10.0	30.0	–	–	1200
[*]RG-213/U	50	0.405	0.66	0.16	0.6	1.9	8.0	–	29.5	5000
[†]RG-218/U	50	0.870	0.66	0.066	0.2	1.0	4.4	–	29.5	11 000
[‡]RG-220/U	50	1.120	0.66	0.04	0.2	0.7	3.6	–	29.5	14 000
URM43	50	0.116	0.66	0.396	1.25	3.96	–	–	30.0	2600
URM70	75	0.128	0.66	0.457	1.46	4.63	–	–	20.5	1800
URM76	50	0.116	0.66	0.457	1.46	4.72	–	–	30.0	2600

[*] Formerly RG8A/U.

[†] Formerly RG17A/U.

[‡] Formerly RG19A/U.

Table 3.2 British UR series

UR number	Nominal impedance Z_0 (ohms)	Overall diameter (inches)	Inner conductor (inches)	Capacity (pF/ft)	Maximum operating voltage RMS	Attenuation (dB per 100 ft)				Nearest RG equivalent
						10 MHz	100 MHz	300 MHz	1000 MHz	
43	52	0.195	0.032	29	2750	1.3	4.3	8.7	18.1	58/U
57	75	0.405	0.044	20.6	5000	0.6	1.9	3.5	7.1	11A/U
63*	75	0.855	0.175	14	4400	0.15	0.5	0.9	1.7	
67	50	0.405	7/0.029	30	4800	0.6	2.0	3.7	7.5	213/U
74	51	0.870	0.188	30.7	15 000	0.3	1.0	1.9	4.2	218/U
76	51	0.195	19/0.0066	29	1800	1.6	5.3	9.6	22.0	58C/U
77	75	0.870	0.104	20.5	1 25 000	0.3	1.0	1.9	4.2	164/U
79*	50	0.855	0.265	21	6000	0.16	0.5	0.9	1.8	
83*	50	0.555	0.168	21	2600	0.25	0.8	1.5	2.8	
85*	75	0.555	0.109	14	2600	0.2	0.7	1.3	2.5	
90	75	0.242	0.022	20	2500	1.1	3.5	6.3	12.3	59B/U

All the above cables have solid dielectric with a velocity factor of 0.66 with the exception of those marked with an asterisk, which are helical membrane and have a velocity factor of 0.96.

Table 3.3 Foam dielectric, solid outer conductor co-axial RF cables 50 ohms characteristic impedance

	Superflexible FS-J series		*LDF series*	
Cable type	FSJ1-50A 1/4″	FSJ4-50B 1/2″	LDF2-50 3/8″	LDF4-50A 1/2″
Min. bending radius, in. (mm)	1.0 (25)	1.25 (32)	3.75 (95)	5.0 (125)
Propagation velocity, %	84	81	88	88
Max. operating frequency, GHz	20.4	10.2	13.5	8.8
Att. dB/100 ft (dB/100 m)				
50 MHz	1.27	0.732	0.75	0.479
	(4.17)	(2.40)	(2.46)	(1.57)
100 MHz	1.81	1.05	1.05	0.684
	(5.94)	(3.44)	(3.44)	(2.24)
400 MHz	3.70	2.18	2.16	1.42
	(12.1)	(7.14)	(7.09)	(4.66)
1000 MHz	6.00	3.58	3.50	2.34
	(19.7)	(11.7)	(11.5)	(7.68)
5000 MHz	14.6	9.13	8.80	5.93
	(47.9)	(30.0)	(28.9)	(19.5)
10 000 MHz	21.8	15.0	13.5	N/A
	(71.5)	(49.3)	(44.3)	N/A
Power rating, kW at 25°C (77°F)				
50 MHz	1.33	3.92	2.16	3.63
100 MHz	0.93	2.74	1.48	2.53
400 MHz	0.452	1.32	0.731	1.22
1000 MHz	0.276	0.796	0.445	0.744
5000 MHz	0.11	0.30	0.17	0.29
10 000 MHz	0.072	0.19	0.12	N/A

The claimed advantages of foam dielectric, solid outer conductor cables are:

1. Lower attenuation
2. Improved RF shielding, approximately 30 dB improvement
3. High average power ratings because of the improved thermal conductivity of the outer conductor and the lower attenuation.

A disadvantage is that they are not so easy to handle as braided cables.

3.5.3 Voltage standing wave ratio (VSWR)

When an RF cable is mismatched, i.e. connected to a load of a different impedance to that of the cable, not all the power supplied to the cable is absorbed by the load. That which does not enter the load is reflected back down the cable. This reflected power adds to the incident voltage when they are in phase with each other and subtracts from the incident voltage when the two are out of phase. The result is a series of voltage – and current – maxima and minima at half-wavelength intervals along the length of the line (*Figure 3.4*). The maxima are referred to as antinodes and the minima as nodes.

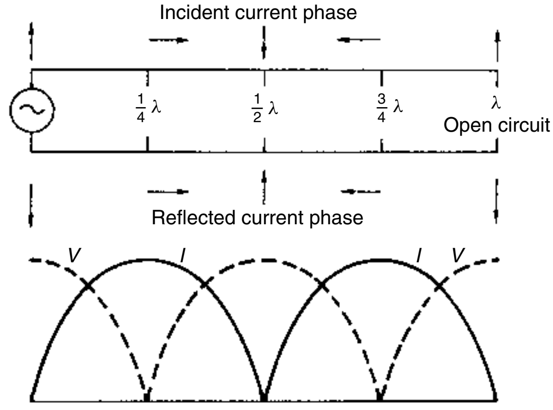

Figure 3.4 Formation of standing waves

The voltage standing wave ratio is the numerical ratio of the maximum voltage on the line to the minimum voltage: VSWR $= V_{\text{max}}/V_{\text{min}}$. It is also given by: VSWR $= R_L/Z_0$ or Z_0/R_L (depending on which is the larger so that the ratio is always greater than unity) where $R_L =$ the load resistance.

The return loss is the power ratio, in dB, between the incident (forward) power and the reflected (reverse) power.

The reflection coefficient is the numerical ratio of the reflected voltage to the incident voltage.

The VSWR is 1, and there is no reflected power, whenever the load is purely resistive and its value equals the characteristic impedance of the line. When the load resistance does not equal the line impedance, or the load is reactive, the VSWR rises above unity.

A low VSWR is vital to avoid loss of radiated power, heating of the line due to high power loss, breakdown of the line caused by high voltage stress, and excessive radiation from the line. In practice, a VSWR of 1.5:1 is considered acceptable for an antenna system, higher ratios indicating a possible defect.

3.5.4 Transmission line filters, baluns and matching circuits

Use can be made of standing waves on sections of line to provide filters and RF transformers. When a line one-quarter wavelength long (a$\lambda/4$ stub) is open circuit at the load end, i.e. high impedance, an effective short-circuit is presented to the source at the resonant frequency of the section of line, producing an effective band stop filter. The same effect would be produced by a short-circuited $\lambda/2$ section. Unbalanced co-axial cables with an impedance of 50 Ω are commonly used to connect VHF and UHF base stations to their antennas although the antennas are often of a different impedance and balanced about ground. To match the antenna to the feeder and to provide a balance to unbalance transformation (known as a balun), sections of co-axial cable are built into the antenna support boom to act as both a balun and an RF transformer.

Balun

The sleeve balun consists of an outer conducting sleeve, one quarter-wavelength long at the operating frequency of the antenna, and connected to the outer conductor of the co-axial cable as in *Figure 3.5*. When viewed from point Y, the outer conductor of the feeder cable and the sleeve form a short-circuited quarter-wavelength stub at the operating frequency and the impedance between the two is very high. This effectively removes the connection to ground for RF, but not for DC, of the outer conductor of the feeder cable permitting the connection of the balanced antenna to the unbalanced cable without short-circuiting one element of the antenna to ground.

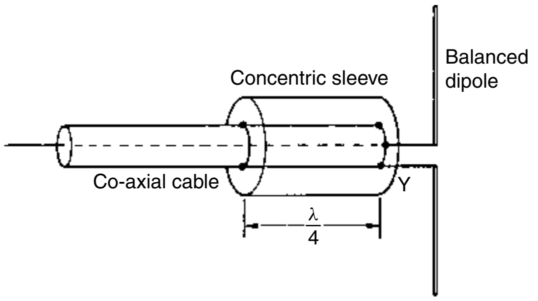

Figure 3.5 Sleeve balun

RF transformer

If a transmission line is mismatched to the load variations of voltage and current, and therefore impedance, occur along its length (standing waves). If the line is of the correct length an inversion of the load impedance appears at the input end. When a λ/4 line is terminated in other than its characteristic impedance an impedance transformation takes place. The impedance at the source is given by:

$$Z_s = \frac{{Z_0}^2}{Z_L}$$

where
Z_s = impedance at input to line
Z_0 = characteristic impedance of line
Z_L = impedance of load

By inserting a quarter-wavelength section of cable having the correct characteristic impedance in a transmission line an antenna of any impedance can be matched to a standard feeder cable for a particular design frequency. A common example is the matching of a folded dipole of 300 Ω impedance to a 50 Ω feeder cable.

Let $Z_s = Z_0$ of feeder cable and $Z_0^{'}$ = characteristic impedance of transformer section. Then:

$$\begin{aligned} Z_0 &= \frac{Z_0^{'\,2}}{Z_L} \\ Z_0^{'} &= \sqrt{Z_0 Z_L} \\ &= \sqrt{300 \times 50} = 122\,\Omega \end{aligned}$$

3.6 Waveguides

At the higher microwave frequencies waveguides which conduct electromagnetic waves, not electric currents, are often used. Waveguides are conductive tubes, either of rectangular, circular or elliptical section which guide the wave along their length by reflections from the tube walls. The walls are not used as conducting elements but merely for containment of the wave. Waveguides are not normally used below about 3 GHz because their cross-sectional dimensions must be comparable to a wavelength at the operating frequency. The advantages of a waveguide over a co-axial cable are lower power loss, low VSWR and a higher operating frequency, but they are more expensive and difficult to install.

In a rectangular waveguide an electromagnetic wave is radiated from the source at an angle to the direction of propagation and is

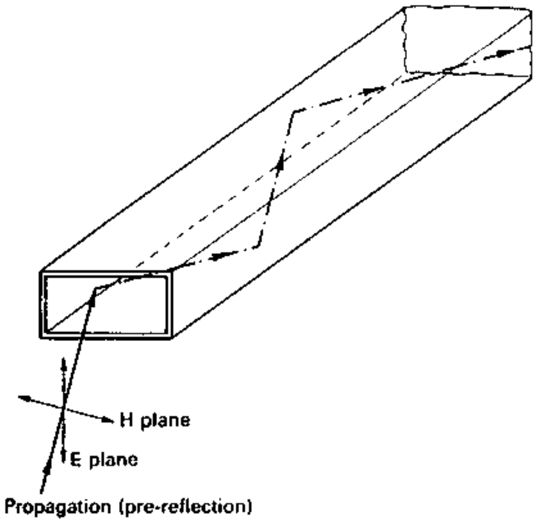

Figure 3.6 Propagation in rectangular waveguide

bounced off the walls (*Figure 3.6*). If the wave were transmitted directly along the length of the guide the electric field would be parallel to one of the walls and be short-circuited by it. Radiating the wave at an angle to the walls creates the maximum field at the centre of the guide and zero at the walls, if the dimensions of the guide are correct for the frequency. However, because the wave does not travel directly along the length of the guide, the speed of propagation is less than in space.

In an electro-magnetic wave the electric and magnetic fields, and the direction of propagation, are mutually perpendicular (see *Figure 1.3*) and such a wave may therefore be thought of as transverse electro-magnetic (TEM). In a waveguide though, because of the short-circuiting effect of the walls, a TEM wave cannot exist. A method of making the wave either transverse electric or transverse magnetic is needed.

When a wave is propagated by a reflection either the magnetic or the electric field is changed. The changed field will now contain the normal component perpendicular to the direction of propagation and a component in its direction, i.e. the wave is no longer wholly transverse. It must be either transverse electric or transverse magnetic. The terminology used to distinguish the type of wave differs: the American system uses the field which behaves as it would in free space to describe the type of wave, e.g. when there is no electric field in the direction of propagation the wave is called TE and the mode with no magnetic field in the direction of propagation, TM; the European system uses the field that is modified and an American TE wave is a European TM wave. In the European system H and E may also be used in lieu of (American) TE and TM.

The behaviour of waves in circular waveguides is similar to that in rectangular guides. Circular waveguides minimize feeder attenuation and are particularly suitable for long vertical runs. A single circular waveguide can carry two polarizations with a minimum isolation of 30 dB. Circular waveguides are recommended where attenuation is critical or where multi-band capability is needed.

Elliptical waveguides have the advantages of flexibility, the availability of long continuous runs and reduced cost.

3.7 Other transmission line considerations

3.7.1 Noise factor of coaxial cable transmission line

Any lossy electrical device, including coaxial cable, produces a noise level of its own. The noise factor of coaxial line is:

$$F_{\mathrm{N(COAX)}} = 1 + \frac{(L-1)T}{290}$$

where

$F_{\mathrm{N(COAX)}}$ is the noise factor of the coax

L is the loss expressed as a linear quantity

T is the physical temperature of the cable in Kelvins

The linear noise factor due to loss can be converted to the noise figure by $10\log(F_{\mathrm{N(COAX)}})$, which can be added to the system noise decibel for decibel.

The attenuation loss figure published in manufacturers' tables is called the *matched line loss* (L_{M}) because it refers to the situation where the load and characteristic impedance of the line are equal. But we also have to consider the *Total Line Loss* (TLL), which is:

$$\mathrm{TLL} = 10\log\left[\frac{B^2 - C^2}{B(1-C^2)}\right]$$

where

B = antilog L_{M}

$C = (\mathrm{SWR}_{\mathrm{LOAD}} - 1)/(\mathrm{SWR}_{\mathrm{LOAD}} + 1)$

$\mathrm{SWR}_{\mathrm{LOAD}}$ is the VSWR at the load end of the line

3.7.2 *Types of coaxial cable*

Coaxial cable consists of two cylindrical conductors sharing the same axis (hence 'co-axial') and separated by a dielectric. For low frequencies (in flexible cables) the dielectric may be polyethylene or polyethylene foam, but at higher frequencies Teflon® and other materials are used. Also used in some applications, notably high powered broadcasting transmitters, are dry air and dry nitrogen.

Several forms of coaxial line are available. Flexible coaxial cable discussed earlier in this chapter is perhaps the most common form. The outer conductor in such cable is made of either braided wire or foil. Again, television broadcast receiver antennas provide an example of such cable from common experience. Another form of flexible or semi-flexible coaxial line is *helical line* (*Figure 3.7*) in which the outer conductor is spiral wound. This type of coaxial cable is usually 2.5 or more centimetres in diameter.

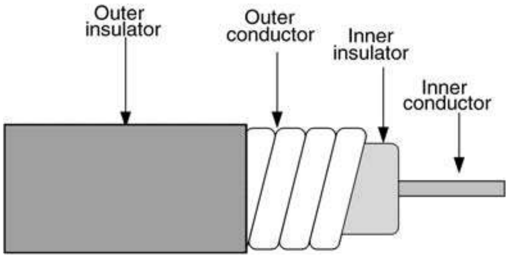
Outer insulator
Outer conductor
Inner insulator
Inner conductor

Figure 3.7 The helical line

Hardline is coaxial cable that uses a thin-walled pipe as the outer conductor. Some hardline coax used at microwave frequencies has a rigid outer conductor and a solid dielectric. *Gas-filled line* is a special case of hardline that is hollow (*Figure 3.8*), the centre conductor

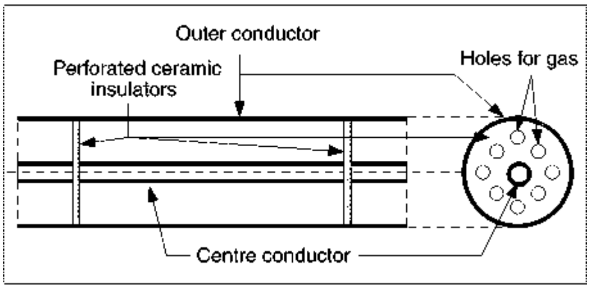
Outer conductor
Perforated ceramic insulators
Holes for gas
Centre conductor

Figure 3.8 Gas-filled line

being supported by a series of thin ceramic or Teflon insulators. The dielectric is either anhydrous (i.e. dry) nitrogen or some other inert gas.

Some flexible microwave coaxial cable uses a solid 'air-articulated' dielectric (*Figure 3.9*), in which the inner insulator is not continuous around the centre conductor, but rather is ridged. Reduced dielectric losses increase the usefulness of the cable at higher frequencies. Double shielded coaxial cable (*Figure 3.10*) provides an extra measure of protection against radiation from the line, and EMI from outside sources from getting into the system.

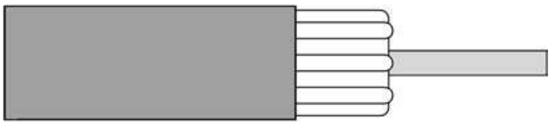

Figure 3.9 Solid 'air-articulated' dielectric

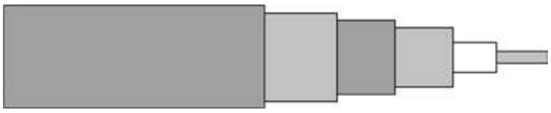

Figure 3.10 Double shielded coaxial cable

3.7.3 Transmission line noise

Transmission lines are capable of generating noise and spurious voltages that are seen by the system as valid signals. Several such sources exist. One source is coupling between noise currents flowing in the outer and inner conductors. Such currents are induced by nearby electromagnetic interference and other sources (e.g. connection to a noisy ground plane). Although coaxial design reduces noise pick-up compared with parallel line, the potential for EMI exists. Selection of high-grade line, with a high degree of shielding, reduces the problem.

Another source of noise is thermal noise in the resistances and conductances of the line. This type of noise is proportional to resistance and temperature.

There is also noise created by mechanical movement of the cable. One species results from movement of the dielectric against the two conductors. This form of noise is caused by electrostatic discharges in much the same manner as the spark created by rubbing a piece of plastic against woollen cloth.

A second species of mechanically generated noise is piezoelectricity in the dielectric. Although more common in cheap cables, one should be aware of it. Mechanical deformation of the dielectric causes electrical potentials to be generated.

Both species of mechanically generated noise can be reduced or eliminated by proper mounting of the cable. Although rarely a problem at lower frequencies, such noise can be significant at microwave frequencies when signals are low.

3.7.4 Coaxial cable capacitance

A coaxial transmission line possesses a certain capacitance per unit of length. This capacitance is defined by:

$$C = \frac{24\varepsilon}{\log(D/d)} \frac{\text{pF}}{\text{Metre}}$$

where

C is the capacitance

D is the outside conductor diameter

d is the inside conductor diameter

ε is the dielectric constant of the insulator.

A long run of coaxial cable can build up a large capacitance. For example, a common type of coax is rated at 65 pF/metre. A 150 metre roll thus has a capacitance of (65 pF/m) (150 m), or 9750 pF. When charged with a high voltage, as is done in performing breakdown voltage tests at the factory, the cable acts like a charged high voltage capacitor. Although rarely if ever lethal to humans, the stored voltage in new cable can deliver a nasty electrical shock and can irreparably damage electronic components.

3.7.5 Coaxial cable cut-off frequency (F_c)

The normal mode in which a coaxial cable propagates a signal is as a transverse electromagnetic (TEM) wave, but others are possible – and usually undesirable. There is a maximum frequency above which TEM propagation becomes a problem, and higher modes dominate. Coaxial cable should not be used above a frequency of:

$$F_{\text{CUT-OFF}} = \frac{1}{3.76(D + d)\sqrt{\varepsilon}}$$

where

F is the TEM mode cut-off frequency

D is the diameter of the outer conductor in mm

d is the diameter of the inner conductor in mm
ε is the dielectric constant

When maximum operating frequencies for cable are listed it is the TEM mode that is cited. Beware of attenuation, however, when making selections for microwave frequencies. A particular cable may have a sufficiently high TEM mode frequency, but still exhibit a high attenuation per unit length at X or Ku-bands.

References

Andrew Antennas, (1991). *Catalogue 35*, Illinois.

British Telecommunications (1992). *Connect Direct, private circuits*. BT, London.

Kennedy, G. (1977). *Electronic Communications Systems*. McGraw-Hill Kogashuka, Tokyo.

Terman, F.E. (1943), *Radio Engineers' Handbook*, McGraw-Hill, London.

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

4 Antennas

4.1 Antenna characteristics

4.1.1 Bandwidth

Stated as a percentage of the nominal design frequency, the bandwidth of an antenna is the band of frequencies over which it is considered to perform acceptably. The limits of the bandwidth are characterized by unacceptable variations in the impedance which changes from resistive at resonance to reactive, the radiation pattern, and an increasing VSWR.

4.1.2 Beamwidth

In directional antennas the beamwidth, sometimes called half-power beamwidth (HPBW), is normally specified as the total width, in degrees, of the main radiation lobe at the angle where the radiated power has fallen by 3 dB below that on the centre line of the lobe (*Figure 4.1A*).

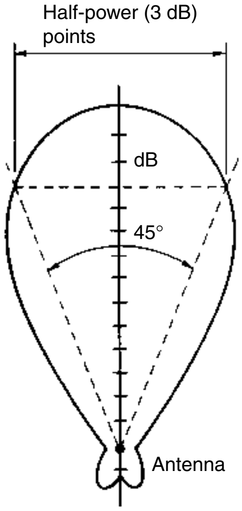

Figure 4.1A Half-power beamwidth

4.1.3 Directivity and forward gain

All practical antennas concentrate the radiated energy in some directions at the expense of others. They possess directivity but are completely passive; they cannot increase the power applied to them. Nevertheless, it is convenient to express the enhanced radiation in some directions as a power gain.

Antenna gain may be quoted with reference to either an isotropic radiator or the simplest of practical antennas, the dipole. There is a difference of 2.15 dB between the two figures. A gain quoted in dBi is with reference to an isotropic radiator and a gain quoted in dBd is with reference to a dipole. When gain is quoted in dBi, 2.15 dB must be subtracted to relate the gain to that of a dipole.

4.1.4 Effective height or length

The current flowing in an antenna varies along its length (see *Figure 1.4*). If the current were uniform along the length of an antenna it would produce a field appropriate to its physical length, and the effective height or length of the antenna would be its physical length. In practice, because the current is not uniform, the effective length is less than the physical length and is given by:

$$l_{\mathrm{eff}} = \frac{l_{\mathrm{phys}} \times I_{\mathrm{mean}}}{I}$$

where

l_{eff} = effective length

l_{phys} = physical length

I = current at feed point

With an antenna which is short in comparison with a wavelength the current can be considered to vary linearly over its length and $I_{\mathrm{mean}} = I/2$. Because the apparent length of a vertical radiator is twice that of its physical dimension due to the mirror image formed below the ground, the effective length of an electrically short vertical antenna may be approximated to be its physical length.

4.1.5 Effective radiated power (erp)

This is the power effectively radiated along the centre line of the main lobe. It is the power supplied to the antenna multiplied by the antenna gain with reference to a dipole.

4.1.6 Radiation resistance and efficiency

The power radiated by an antenna can conveniently be expressed in terms of the value of a resistor which would dissipate the same power that the antenna radiates. This value is referred to as the radiation resistance and is defined as the ratio of the power radiated to the square of the current at the feed point. The efficiency is the ratio of the power radiated to that lost in the antenna. It is given by:

$$\text{eff} = \frac{R_\text{r}}{R_\text{r} + R_\text{L}} \times 100\%$$

where R_r is the radiation resistance and R_L represents the total loss resistance of the antenna. The sum of the two resistances is the total resistance of the antenna and, for a resonant antenna, is also the impedance.

4.1.7 Front-to-back ratio

The ratio, in dB, of the strength of the radiation (or received signal) in the forward (desired) direction to that in the reverse (unwanted) direction. The front-to-back ratio of the antenna shown in *Figure 4.1A* is 13 dB.

4.1.8 Impedance

The impedance of an antenna is that presented to the feeder cable connecting it to the transmitter or receiver. It is the result of the vectorial addition of the inductive, capacitive and resistive elements of the antenna. Each resonant antenna possesses an impedance characteristic of the type, and when an antenna operates at its resonant frequency the reactive elements cancel out and the impedance becomes resistive. The radiation resistance plus the losses in the antenna, i.e. the series resistance of the conductors, the shunt resistance of the base material and losses in nearby objects, form the resistive portion of the impedance.

4.1.9 Polarization

The radiated field from an antenna is considered to be polarized in the plane of the length of the conductors which is the plane of the electric field, the E plane. Confusion arises when reference is made to vertical or horizontal polarization and it is preferable when referring to polar diagrams to use the E and H plane references.

Circular polarization, produced by crossed dipoles or helical wound antennas, is occasionally used for point-to-point work at VHF and above to reduce multi-path propagation losses.

Cross polarization discrimination defines how effectively an antenna discriminates between a signal with the correct polarization, i.e. mounted with the elements in the same plane, and one operating at the same frequency with the opposite polarization. 20 dB is typical.

4.1.10 Radiation pattern

A plot of the directivity of an antenna showing a comparison of the power radiated over 360°. Two polar diagrams are required to show the radiation in the E and H planes. The polar diagrams may be calibrated in either linear (voltage) or logarithmic (decibel) forms.

4.1.11 Voltage standing wave ratio (VSWR)

Most VHF and UHF antennas contain an impedance matching device made up of lengths of co-axial cable. Thus the VSWR (see Chapter 3) of these types of antenna varies with the operating frequency, more so than the bandwidth of the antenna alone would produce. At the centre design frequency, the VSWR should, theoretically, be 1:1 but in practice a VSWR less than 1.5:1 is considered acceptable.

4.1.12 Receive aperture

Receiving antennas also possess a property called aperture, or capture area. This concept relates the amount of power that is delivered to a matched receiver to the power density (watts per square metre). The aperture is often larger than the physical area of the antenna, as in the case of the half-wavelength dipole (where the wire fronts a very small physical area), or less as in the case of a parabolic reflector used in microwave reception. *Figure 4.1B* shows the capture area of a half-wavelength (0.5λ) dipole. It consists of an ellipse with major axes of 0.51λ and 0.34λ. The relationship between gain and aperture is:

$$A_e = \frac{G\lambda^2}{4\pi n}$$

where

A_e is the effective aperture

G is the gain

λ is the wavelength of the signal

n is the aperture effectiveness ($n = 1$ for a perfect no-loss antenna, real values are typically 0.3 to 0.55

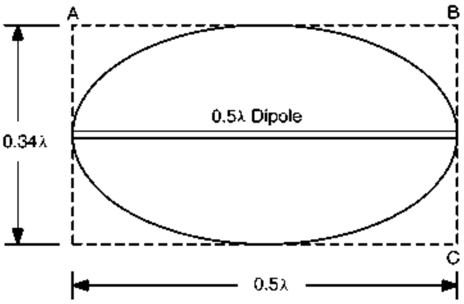
A
B
C
0.5λ Dipole
0.34λ
0.5λ

Figure 4.1B Capture area of half-wavelength (0.5λ) dipole

4.2 Antenna types

4.2.1 The dipole

The half-wavelength (λ/2) dipole as described in Section 1.4 is the antenna on which many others are based. *Figure 4.2* shows the relative radiation in the E and H planes of a dipole in free space.

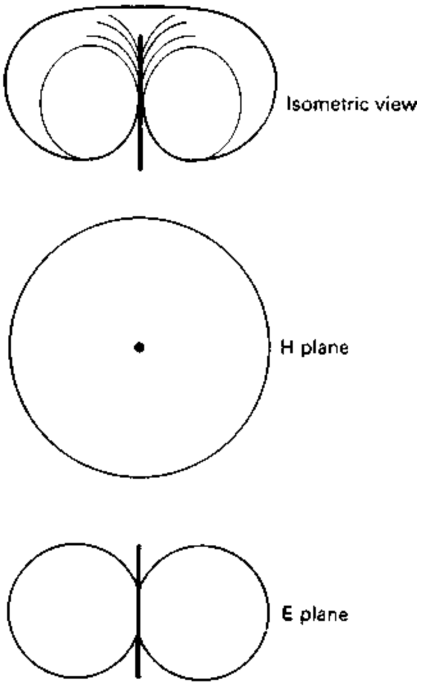
Isometric view
H plane
E plane

Figure 4.2 Radiation patterns of a half-wave dipole

The impedance of a half-wavelength dipole is 72 Ω; that of a full wavelength or folded dipole is 300 Ω.

4.2.2 *The quarter-wavelength vertical radiator*

The quarter-wavelength (λ/4) vertical radiator is a commonly used antenna for MF broadcasting and for VHF and UHF mobile radio applications. It is derived from the λ/2 dipole and it is assumed that a mirror image of the radiator is formed below the ground to complete the λ/2 structure of the dipole as in *Figure 4.3*. The radiation pattern of a λ/4 vertical radiator mounted close to a perfect earth shows a strong similarity to that of a dipole. The effect of reducing the size and conductivity of the ground plane raises the angle of radiation.

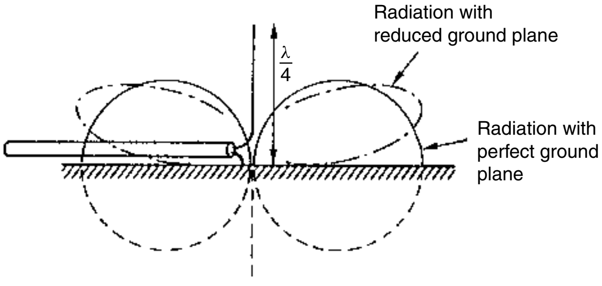
Radiation with reduced ground plane
$\frac{\lambda}{4}$
Radiation with perfect ground plane

Figure 4.3 Quarter-wavelength vertical radiator

The impedance of a perfect λ/4 vertical radiator is 36 Ω but reducing the effectiveness of the ground plane raises the impedance.

4.2.3 *LF, MF and HF antennas*

Because of the physical lengths involved, LF and MF antennas are usually non-resonant and their impedances do not conform to the resistive 70 Ω or 36 Ω of the basic resonant types. The impedance of a non-resonant antenna is usually higher and reactive so an antenna tuning or matching unit is used to couple the antenna efficiently to the transmission line and also act as filter to reduce out-of-band radiations. The matching unit comprises a tuned circuit with either a tap on the coil at the correct impedance point or a separate coupling coil to feed the antenna.

Obtaining an adequate length is always the problem with low frequency antennas and various methods have been used based mainly on the λ/4 radiator. The horizontal section of the inverted L (*Figure 4.4*) extends the effective length but, as the ground wave is much used at

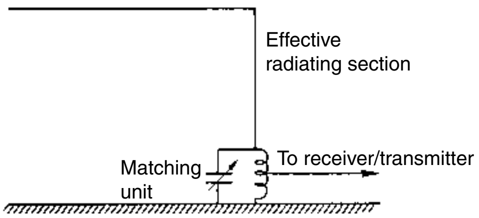

Figure 4.4 Inverted L antenna

the lower frequencies, these antennas are intended for vertical polarization and it is therefore only the down-lead which radiates, or receives, effectively. An alternative method of increasing the effective height of a vertical radiator is to provide a capacitance top where the system of horizontal conductors provides a high capacitance to ground. This prevents the current falling to zero at the top of the antenna, maintaining a higher mean current and so increasing the antenna's effective length.

Dipoles used at HF are mounted horizontally because of their length and have directivity in the horizontal (E) plane. Propagation is mainly by the sky wave and the omni-directional properties in the vertical (H) plane, modified by ground reflections, produce wide angle upwards radiation.

4.2.4 Directional arrays

Broadside array

A broadside array consists of several radiators spaced uniformly along a line, each carrying currents of the same phase. When each radiator has an omni-directional pattern, and the spacing between radiators is less than $3\lambda/4$, maximum radiation occurs at right angles to the line of the array. The power gain is proportional to the length of the array, provided that the length is greater than two wavelengths; this means, effectively, the number of radiators. *Figure 4.5* shows a typical H plane polar diagram for an array with vertically mounted radiators and a spacing of $\lambda/2$.

End-fire array

Physically an end-fire array is identical to a broadside except for the feeding arrangements and the spacing of the elements. In an end-fire array the radiators are fed with a phase difference between adjacent radiators equal in radians to the spacing between them in wavelengths.

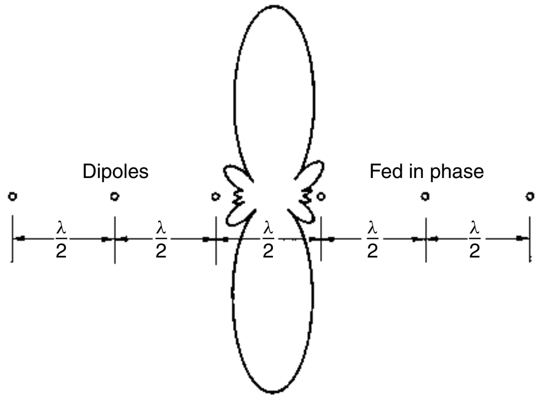

Figure 4.5 Broadside array

A spacing of λ/4 requires a phase shift of 90° between adjacent radiators. *Figure 4.6* shows a typical radiation pattern. An end-fire array concentrates the power in both the E and H planes and the maximum radiation is in the direction of the end of the array with the lagging phase.

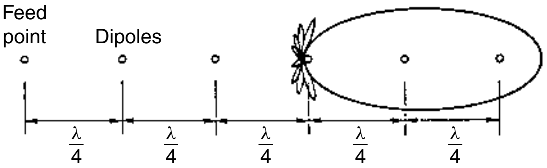

Figure 4.6 End-fire array

Rhombic antenna

A rhombic is a wide band, directional antenna comprised of four non-resonant wire antennas, each several wavelengths long, arranged as shown in *Figure 4.7(a)* which also shows the radiation pattern for each leg of the rhombus. The lobe angle θ can be varied by adjusting the length, in wavelengths, of each radiator. The antenna has greater directivity than a single wire and can be terminated by an appropriate value resistor to ensure non-resonance and a wide bandwidth. However, because it must be terminated in a resistance equal to the characteristic impedance of the conductors, it cannot be more than 50% efficient. It also exhibits considerable side lobes of radiated power. Rhombics are used for sky wave working at HF and more than one frequency is allocated to allow for varying propagation conditions. The conductors of a rhombic are normally horizontal and the horizontal directivity is determined by the tilt angle, β in *Figure 4.7(a)*. If the lobe angle θ is equal to $(90 - \beta)^\circ$ the radiation in the A lobes cancels, while that from the B lobes, which point in the same direction, is added. The resultant pattern in the horizontal plane is shown in *Figure 4.7(b)*. The vertical directivity is controlled by the height of the conductors above the ground.

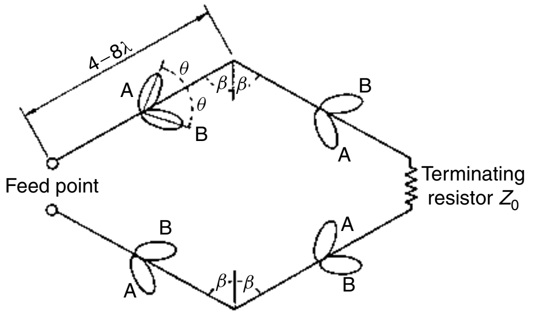

(a) Rhombic antenna

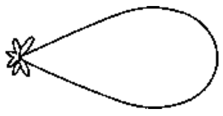

(b) Radiation pattern of rhombic

Figure 4.7

Log-periodic antenna

An alternative, usable from HF through UHF, to the rhombic for wide band operation is the log-periodic antenna. It is comprised of several dipoles of progressive lengths and spacings as in *Figure 4.8*, and is resonant over a wide frequency range and may be mounted with either polarization. The dipoles are fed via the support booms and this construction ensures that the resultant phasing of the dipoles is additive in the forward direction producing an end-fire effect. However, because at any one frequency only a few of the dipoles are close to resonance, the forward gain of the antenna is low considering the number of elements it contains.

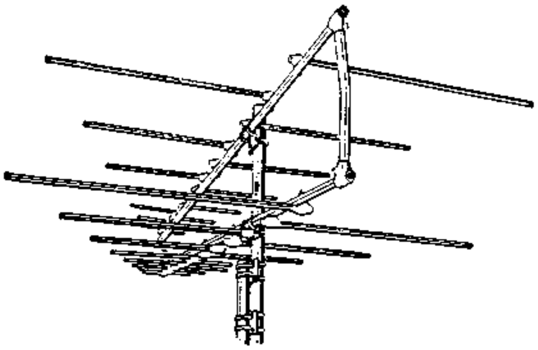

Figure 4.8 Log-periodic antenna at VHF frequency

4.3 VHF and UHF antennas

4.3.1 Base station antennas

Apart from entertainment broadcasting and mobile telephony, most VHF and UHF systems use vertical polarization and a dipole – or to prevent noise from rain static, the folded dipole – with the conductors mounted vertically is a frequently used antenna for VHF and UHF base station installations. Unfortunately it is often mounted on the side of the support structure in a manner which seriously affects its omnidirectional radiation pattern. Where practical, there should be a minimum spacing of one wavelength between the structure and the rearmost element of the antenna.

To obtain a good omni-directional pattern either an end-fed dipole (*Figure 4.9*) or a unipole antenna (*Figure 4.10*) protruding from the top of the mast or tower is the best option. A unipole is a variation of the vertical quarter-wave radiator and provides a low angle of radiation.

To reduce the likelihood of co-channel interference directional antennas are often necessary. The simplest of these is the combination of a $\lambda/2$ dipole and reflector shown in *Figure 4.11*. The reflector is slightly longer than the dipole and spaced one quarter-wavelength from it. The portion of the signal radiated by the dipole in the direction of the reflector is received and re-transmitted by the reflector, with a 180° phase change occurring in the process. The signal re-transmitted

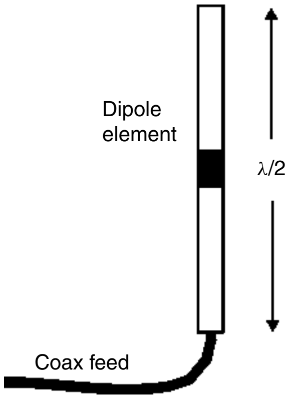

Figure 4.9 Type EDV end-fed antenna (by kind permission of C and S Antennas Ltd)

to the rear of the antenna – the direction of the reflector – cancels the signal from the dipole, that towards the front of the antenna adds to the signal from the dipole giving the radiation pattern shown. The power gain of a dipole and reflector, a two-element array, is 3 dBd.

Directivity can be increased by adding directors forward of the dipole, the result is a Yagi–Uda array. The limit to the number of radiators is set by physical constraints and the reduction of bandwidth produced by their addition. At low VHF, a 3-element array is about the practical maximum, while at 1500 MHz, 12-element arrays are commonplace. As a rule of thumb, doubling the number of elements in an array increases the forward gain by 3 dB. Where the maximum front-to-back ratio is essential the single rod reflector can be replaced by a corner reflector screen.

Co-linear antennas provide omni-directional characteristics and power gain in the H plane. A co-linear consists of a number of dipoles stacked vertically and, in the normal configuration, fed so that they radiate in phase and the maximum power is radiated horizontally. *Figure 4.12* shows alternative feeding arrangements. One advantage of the co-linear is that the horizontal angle of radiation can be tilted to about 15° downwards by changing the phasing of the elements. The gain of a co-linear is limited, because of the physical lengths involved and losses in the feeding arrangements to 3 dBd at VHF and 6 dBd at UHF.

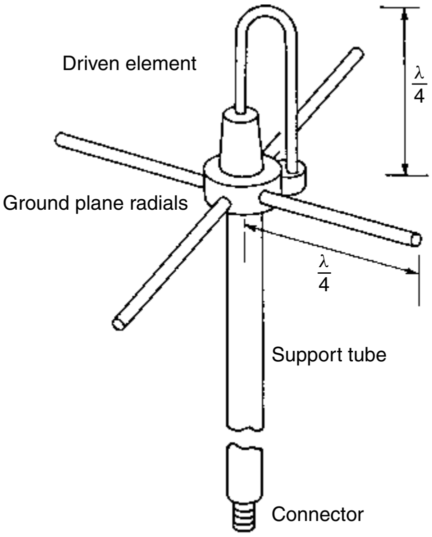
Driven element
$\frac{\lambda}{4}$
Ground plane radials
$\frac{\lambda}{4}$
Support tube
Connector

Figure 4.10 Folded unipole antenna

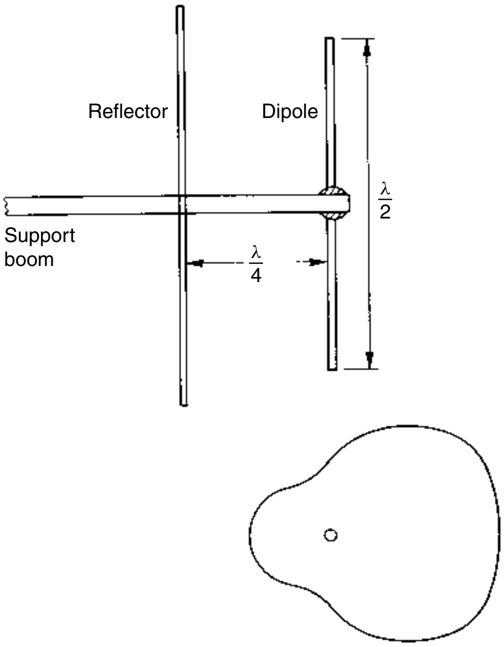
Reflector
Dipole
$\frac{\lambda}{2}$
Support boom
$\frac{\lambda}{4}$

Figure 4.11 Dipole and reflector

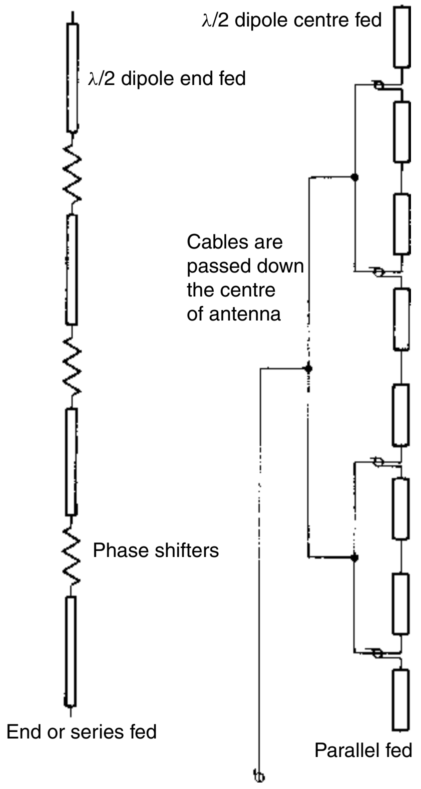
λ/2 dipole centre fed
λ/2 dipole end fed
Cables are passed down the centre of antenna
Phase shifters
End or series fed
Parallel fed

Figure 4.12 General construction of co-linears

Figure 4.13 shows a slot antenna cut into a flat metal sheet. Current (I) injected at the centre of the slot flows around the edge and creates a vertical electric field. The radiated field pattern is like a dipole.

The type of slot antenna typically used for mobile telephony base stations is a cylindrical waveguide with slots cut width-wise. Current flowing along the waveguide creates an electric field along the length of the cylinder. The radiation pattern produced by a slot antenna cut into a cylinder is directional, with the main beam perpendicular to the slot. Using two slot antennas side by side provides radio coverage over a 120° sector. Three pairs of slot antennas placed around a mast gives three sectors that can operate at different frequencies.

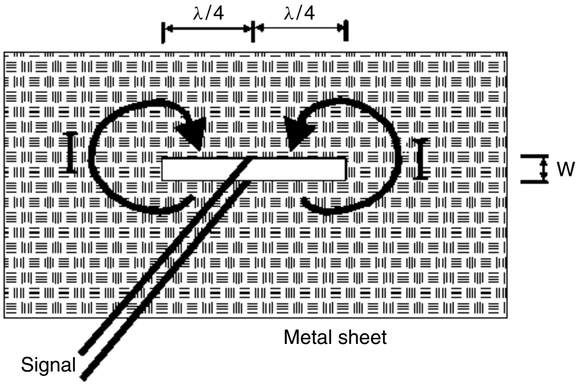

Figure 4.13 Slot antenna

A wide-band alternative to the log-periodic is the conical (discone) antenna (*Figure 4.14*). It provides unity gain, is omni-directional and has a bandwidth of approximately 3:1, depending on the designed frequency range. In practice there has been a tendency to expect these antennas to perform outside their specified bandwidths with unsatisfactory results.

Stacking and baying

A method of increasing an antenna's directivity is to mount two or more antennas vertically above one another (stacking) or side-by-side (baying), and to feed them so that they radiate in phase. Stacking two dipoles vertically increases the directivity in the E plane and baying them increases the directivity in the H plane, approximately halving the beamwidth in each case.

An array of two stacked plus two bayed antennas approximately halves the beamwidth in both planes.

4.3.2 Mobile antennas

The aerial is the least expensive, and most abused, component of a mobile radio installation. Frequently installed in a manner which does not produce optimum performance it can have a profound effect on the performance of the whole installation.

Most mobile antennas consist of a metal rod forming a quarter-wavelength radiator. The ideal mounting position is the centre of a metallic roof, and as the area of the ground plane is reduced the radiation pattern changes and more of the energy is radiated upwards (not always a bad thing in inner city areas); also, the impedance rises.

The effect of the mounting position on the H plane radiation can be dramatic, resulting in ragged radiation patterns and, in some directions, negligible radiation. Advice on the installation of mobile antennas and the polar diagrams produced by typical installations are illustrated in MPT 1362, *Code of Practice for Installation of Mobile Radio Equipment in Land Based Vehicles*.

As the installation moves away from the ideal and the antenna impedance rises a mismatch is introduced between the antenna and the feeder with the consequent production of standing waves on the feeder. Under high VSWR conditions the cable is subject to higher voltage stresses and it also behaves as an aerial radiating some of the reflected power. This spurious radiation adds to the radiation from the antenna in some directions but subtracts from it in others giving rise to jagged radiation patterns or deep nulls in radiated signal.

Mobile antennas providing a small amount of gain, typically 3 dB and obtained by narrowing the radiation lobes, are on the market. These have a length of 5/8 wavelength and, because the extra length makes the impedance capacitive at the operational frequency, a loading coil is inserted at the lower end of the element to cancel the capacitive reactance. An adjustable metallic disk towards the base of the whip is often provided for tuning purposes. Note that gain figures quoted for mobile antennas are usually with reference to a quarter-wave whip.

Low profile antennas

Low profile antennas are available for use at UHF. They have a built-in ground plane approximately 150 mm in diameter and a height of some 30 mm and have obvious applications for use on high vehicles and, although not strictly covert, where a less obtrusive antenna is required. They are fitted with a tuning screw and when adjusted to resonance a VSWR of better than 1.2:1 is quoted by one maker and a bandwidth of 10 MHz at a VSWR of 2:1. *Figure 4.15* shows the radiation pattern for one type.

Motor-cycle antennas

The installations of antennas on motor cycles poses problems because of the absence of a satisfactory ground plane. One frequently used method is to employ a 5/8 wavelength whip and loading coil. Another

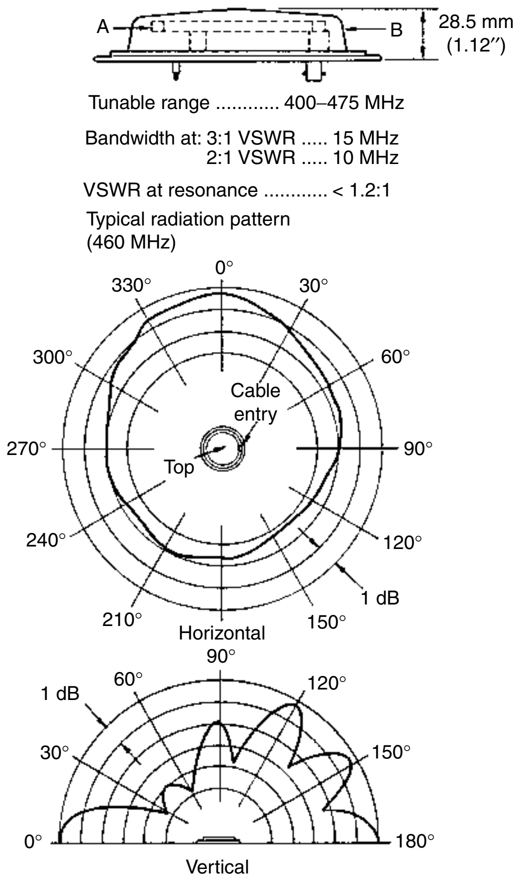

Tunable range .... 400–475 MHz

Bandwidth at: 3:1 VSWR ..... 15 MHz
2:1 VSWR ..... 10 MHz

VSWR at resonance .... < 1.2:1

Typical radiation pattern
(460 MHz)

A
B
28.5 mm
(1.12″)
0°
330°
30°
300°
60°
270°
90°
240°
120°
210°
150°
Cable
entry
Top
1 dB
Horizontal
90°
60°
120°
1 dB
30°
150°
0°
180°
Vertical

Antenna mounted on a 40 cm × 40 cm aluminium plate

Figure 4.15 Low profile UHF antenna (by kind permission of Panorama Antennas Ltd)

method uses a pair of grounded downwards-pointing rods to form the lower half of a dipole.

Hand-portable antennas

Again, because of the lack of a ground plane high performance antennas are difficult to provide for hand-portables, particularly at VHF. Body-worn sets may have an antenna incorporated in the microphone lead but the high current portion of the antenna must then be at a low height and in some directions the radiation must pass through the body, which is highly absorbent, to reach the base station. Helical antennas are frequently used on hand-held sets to reduce the physical length. Useful operating tips are to face the base station when using the radio in low signal areas, while placing the set on a nearby car roof effectively increases the performance of the antenna.

Safety

There are two safety aspects to consider when installing mobile antennas: physical and electrical. The physical considerations are that the antenna must be incapable of inflicting injury when it is in its correct position, and also when it has been bent or damaged. Rear wing mounted antennas need particular care in their positioning; a Band 111 aerial tip is just about eye height when bending over an open boot lid. The same considerations apply to hand-portables, helical antennas being perhaps safer than whips because they are thicker and thus more easily seen. They also have rounded tips.

The electrical dangers are from radiation affecting the body either directly – radiation from a hand-portable helical into the eye is a possible example – or indirectly by affecting electronic equipment. The danger of radiation affecting equipment in the vehicle is increased when the VSWR is high because of increased radiation from the feeder. Advice should be sought from the Radiological Protection Board.

4.4 Microwave antennas

The small antenna elements at microwaves facilitate the construction of highly directive, high gain antennas with high front-to-back ratios.

At frequencies below about 2 GHz, 12- to 24-element Yagi arrays, enclosed in plastic shrouds for weather protection, may be used. At higher frequencies, antennas with dish reflectors are the norm.

The aperture ratio (diameter/wavelength) of a dish governs both its power gain and beamwidth. The power gain of a parabolic dish is given to a close approximation by:

$$\text{Gain} = 10\log_{10} 6(D/\lambda)^2 \times N, \text{dBi}$$

where D = dish diameter and N = efficiency. Dimensions are in metres. The half-power beam width (HPBW) in degrees is approximately equal to $70\lambda/D$.

A microwave antenna with its dish reflector, or parasitic elements in the case of a Yagi type, is a large structure. Because of the very narrow beamwidths – typically 5° for a 1.8 m dish at 2 GHz – both the antenna mounting and its supporting structure must be rigid and able to withstand high twisting forces to avoid deflection of the beam in high winds. Smooth covers, radomes, fitted to dishes and the fibreglass shrouds which are normally integral with Yagis designed for these applications considerably reduce the wind loading and, for some antenna types, increase the survival wind speed.

The electrical performance of a selection of microwave antennas is given in *Table 4.1* and the wind survival and deflection characteristics in *Table 4.2* (Andrew Antennas, 1991).

Table 4.1 2.1–2.2 GHz antennas – electrical characteristics

Type number	*Dia. (m)*	*Gain (dBi)*			*Beam width (deg.)*	*Cross pol. disc (dB)*	*F/B ratio (dB)*	*VSWR max.*
		Bottom	*Mid-band*	*Top*				
Ultra High Performance Antenna, F-Series Unpressurized – Radome Inc.								
Single polarized								
UHP8F-21	2.4	31.9	32.1	32.3	4.2	32	61	1.10
UHP10F-21	3.0	33.7	33.9	34.0	3.6	33	64	1.10
UHP12F-21	3.7	35.4	35.6	35.8	2.9	32	65	1.10
Dual polarized								
UHX8F-21	2.4	31.9	32.1	32.3	4.2	30	58	1.20
UHX10F-21A	3.0	33.8	34.0	34.2	3.6	32	62	1.20
UHX12F-21A	3.7	35.4	35.6	35.8	2.8	32	67	1.20
High Performance Antenna, F-Series Unpressurized – Radome Inc.								
Single polarized								
HP6F-21B	1.8	29.4	29.6	29.8	5.5	30	46	1.12
HP8F-21A	2.4	31.9	32.1	32.3	4.1	30	53	1.12
HP10F-21A	3.0	33.8	34.0	34.2	3.4	32	55	1.12
HP12F-21A	3.7	35.4	35.6	35.8	2.9	32	56	1.12
Standard Antenna, F-Series Unpressurized								
Single polarized								
P4F-21C	1.2	26.4	26.6	26.8	7.6	30	36	1.15
P6F-21C	1.8	29.8	30.0	30.2	4.9	30	39	1.12
P8F-21C	2.4	32.3	32.5	32.7	3.8	30	40	1.12
P10F-21C	3.0	34.0	34.2	34.4	3.4	30	44	1.12
Grid Antenna, F-Series Unpressurized								
Single polarized								
GP6F-21A	1.8	29.6	29.8	30.0	5.4	31	36	1.15
GP8F-21A	2.4	32.0	32.2	32.4	4.0	35	39	1.15
GP10F-21A	3.0	34.0	34.2	34.4	3.3	40	42	1.15
GP12F-21	3.7	35.5	35.7	35.9	2.8	40	44	1.15

With shrouded Yagis and some dishes low loss foam-filled cables are generally used up to about 2 GHz although special connectors may

Table 4.2 Wind survival and deflection characteristics

Antenna types	*Survival ratings*		*Max. deflection in 110 km wind (degrees)*
	Wind velocity (km/h)	*Radial ice (mm)*	
P4F Series			
Without Radome	160	12	0.1
With Radome	185	12	0.1
Standard Antennas (except P4F Series)			
Without Radome	200	25	0.1
With Standard Radome	200	25	0.1
UHX, UMX, UGX Antennas	200	25	0.1

be required. At higher frequencies, air-spaced or pressurized nitrogen-filled cables are frequently used with waveguides as an alternative.

4.4.1 Omnidirectional normal mode helix

The normal mode helix antenna shown in *Figure 4.16* produces an omnidirectional pattern when the antenna is mounted vertically. The diameter (D) of the helical coil should be one-tenth wavelength ($\lambda/10$), while the pitch (i.e. S, the distance between helix loops) is one-twentieth wavelength ($\lambda/20$). An example of the normal mode helix is the 'rubber ducky' antenna used on VHF/UHF two-way radios and scanners.

4.4.2 Axial mode helical antenna

An axial mode helical antenna is shown in *Figure 4.17*. This antenna fires 'off-the-end' in the direction shown by the arrow. The helix is mounted in the centre of a ground plane that is at least 0.8λ across. For some UHF frequencies some manufacturers have used aluminum pie pans for this purpose. The helix itself is made from either heavy copper wire (solid, not stranded) or copper or brass tubing. The copper tubing is a bit easier to work. The dimensions are:

$$D \approx \lambda/3$$

$$S \approx \lambda/4$$

$$\text{Length} \approx 1.44\lambda$$

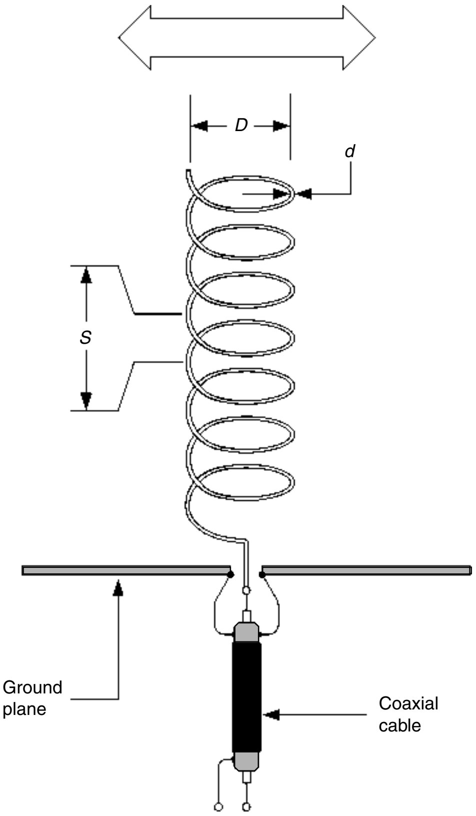

Figure 4.16 Normal mode helix antenna

A ‘rule of thumb’ for the circumference is that maximum gain is obtained when circumference C is:

$$C = 1.066 + [(N - 5) \times 0.003]$$

4.4.3 Small loop antennas

Small loop antennas are used mostly for receiving, although some designs are also used for transmitting. One application for the small loop antenna is radio direction finding (RDF). Another use is for providing a small footprint antenna for people who cannot erect a full sized receiving antenna. Perhaps the greatest use of the small loop antenna is for receiving stations on crowded radio bands. The small loop antenna has very deep nulls that make it easy to null out interfering co-channel and adjacent channel signals.

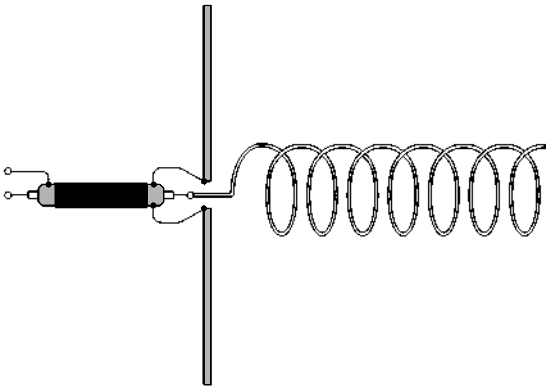

Figure 4.17 Axial mode helical antenna

4.5 Loop antennas

4.5.1 Small loop antennas defined

Large loop antennas are those with overall wire lengths of 0.5λ to more than 2λ. Small loop antennas, on the other hand, have an overall wire length that is much less than one wavelength (1λ). According to a Second World War US Navy training manual such antennas are those with an overall length of $\leq 0.22\lambda$. Jasik's classic 1961 text on radio antennas uses the figure $\leq 0.17\lambda$, while John Kraus (1950) used the figure $\leq 0.10\lambda$. An amateur radio source, *The ARRL Antenna Book*, recommends $\leq 0.085\lambda$ for small loop antennas. For the purposes of our discussion we will use Kraus's figure of $\leq 0.10\lambda$.

A defining characteristic of small loops versus large loops is seen in the current distribution. In the small loop antenna the current flowing in the loop is uniform in all portions of the loop. In the large loop, however, the current varies along the length of the conductor, i.e. there are current nodes and antinodes.

The small loop antenna also differs from the large loop in the manner of its response to the radio signal. A radio signal is a transverse electromagnetic (TEM) wave, in which magnetic and electrical fields alternate with each other along the direction of travel. The large loop, like most large wire antennas, respond primarily to the *electrical field* component of the TEM, while small loops respond mostly to the *magnetic field* component. The importance of this fact is that it means the small loop antenna is less sensitive to local electromagnetic interference sources such as power lines and appliances. Local EMI consists largely of electrical fields, while radio signals have both magnetic and electrical fields. With proper shielding, the electrical response can be reduced even further.

4.5.2 Small loop geometry

Small loop antennas can be built in any of several different shapes (*Figure 4.18*). Popular shapes include hexagonal (*Figure 4.18A*), octagonal (*Figure 4.18B*), triangular (*Figure 4.18C*), circular (*Figure 4.18D*) and square (*Figure 4.18E*).

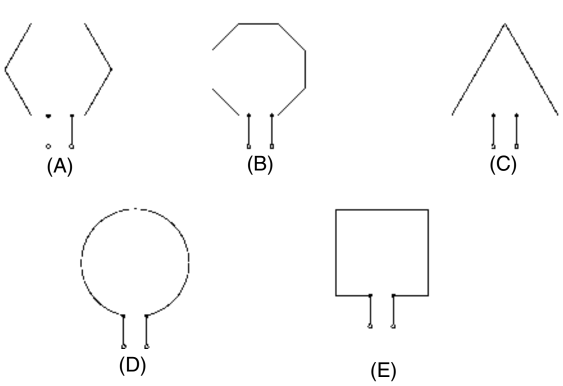

Figure 4.18 Small loop antennas: (A) hexagonal, (B) octagonal, (C) triangular, (D) circular and (E) square

The far-field performance of small loop antennas is approximately equal provided that $A^2 \leq \lambda/100$, where A is the loop area and λ is the wavelength of the desired frequency.

The 'standard' loop used in this discussion is the square loop depicted in *Figure 4.19*. There are two forms of winding used: *depth wound* and *spiral wound*. The difference is that the depth wound has its turns in different parallel planes, while in the spiral wound version all of the turns are in the same plane. The spiral wound loop theoretically has a deeper null than depth wound, but in practical terms there is usually little difference.

The length of each side of the loop is designated A, while the depth or width is designated B. These dimensions will be used in equations shortly. The constraint on the dimensions is that A should be $\leq 0.10\lambda/4$ and $B \leq A/5$.

The loop may be either tuned or untuned. The differences between tuned and untuned will be discussed shortly.

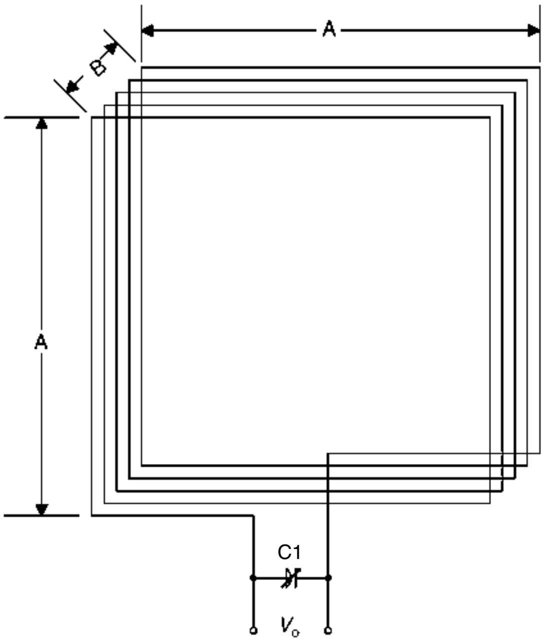

Figure 4.19 Square loop

4.5.3 Small loop antenna patterns

Small loop antennas have patterns opposite those of large loops. The minima, or 'nulls', are perpendicular to the plane of the loop, while the maxima are off the ends. *Figure 4.20* shows the directions of maximum and minimum response. The loop antenna is viewed from above. The nulls are orthogonal to the loop axis, while the maxima are along the loop axis.

The fact that the small loop pattern has nulls perpendicular to the loop axis, i.e. perpendicular to the plane of the loop, is counterintuitive to many people. The advancing radio wave produces alternating regions of high and low amplitude. A *potential difference* exists between any two points. When the loop is aligned such that its axis is parallel to the isopotential lines low signal levels are induced into the loop. If the turns of the loop cut several isopotential lines, a larger signal is induced from this direction.

4.5.4 Signal voltage (V_o) developed by the loop

The actual signal voltage (V_o) at the output of the terminals of an untuned loop is a function of the direction of arrival of the signal (α),

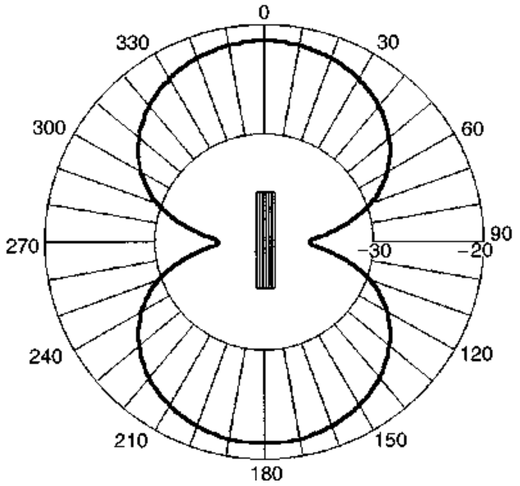

Figure 4.20 Small loop azimuth

as well as the strength of the arriving signal and the design of the loop. The angle α is formed between the loop axis and the advancing isopotential lines of the radio signal.

The output voltage of an untuned loop is:

$$V_o = \frac{2\pi ANE_f \cos\alpha}{\lambda}$$

While the output voltage of a tuned loop is:

$$V_o = \frac{2\pi ANQE_f \cos\alpha}{\lambda}$$

where

V_o is the output voltage of an untuned loop in volts (V)
A is the length of one side of the loop
N is the number of turns in the loop
E_f is the strength of the incoming signal in volts per metre (V/m)
α is the angle between the advancing wavefront and the loop axis
λ is the wavelength of the radio signal in metres (m), i.e. the reciprocal of the frequency ($\lambda = 1/F$)
Q is the loaded Q (figure of merit) of the tuned circuit formed by C1 and the loop inductance (typically 10 to 100)

4.5.5 Effective height

Loop antennas are sometimes described in terms of the *effective height* (H_{eff}) of the antenna. This number is a theoretical construct that compares the output voltage of a small loop with a vertical piece of the same kind of wire that has a height of:

$$H_{\text{eff}} = \frac{2\pi NA}{\lambda}$$

where H_{eff} is the effective height in metres, and all other terms are as defined above.

Loop inductance

A loop antenna essentially forms a coil of wire (or other conductor), so will have inductance. There are several methods for calculating loop inductance, but the most common are the *Grover equation* and the *Patterson equation.*

Grover equation:

$$L_{\mu\text{H}} = \left[K_1 N^2 A \text{Ln}\left[\frac{K_2 AN}{(N+1)B}\right]\right] + K_3 + \left[\frac{K_4(N+1)B}{AN}\right]$$

where all terms are as previously defined, except K_1–K_4, which are defined in *Table 4.3*.

Table 4.3 *K* factors for calculating loop inductance

Shape	K_1	K_2	K_3	K_4
Triangle	0.006	1.1547	0.65533	0.1348
Square	0.008	1.4142	0.37942	0.3333
Hexagon	0.012	2.00	0.65533	0.1348
Octagon	0.016	2.613	0.75143	0.07153

Patterson equation:

$$L\mu\text{H} = (0.00508A) \times \left[2.303 \log\left(\frac{4A}{d}\right) - \phi\right]$$

where

d is the conductor diameter

ϕ is a factor found in *Table 4.4*

Table 4.4 ϕ Factor for calculating loop inductance

Shape	*Factor (ϕ)*
Circle	2.451
Octagon	2.561
Hexagon	2.66
Pentagon	2.712
Square	2.853
Triangle	3.197

Of these equations, most people will find that the Grover equation most accurately calculates the actual inductance realized when a practical loop is built.

References

Andrew Antennas (1991). *Catalog 35*. Illinois.

Belcher, R. *et al*. (1989). *Newnes Mobile Radio Servicing Handbook*. Butterworth-Heinemann, Oxford.

Davies, J. (1987). *Private Mobile Radio – A Practical Guide*. Heinemann, London.

Green, D.C. (1979). *Radio systems, Vols II and III*. Pitman, London.

Kennedy, G. (1977). *Electronic Communication Systems*. McGraw-Hill Kogashuka, Tokyo.

Terman, F.E. (1943). *Radio Engineers' Handbook*. McGraw-Hill, London.

5 Resonant circuits

5.1 Series and parallel tuned circuits

Tuned resonant circuits composed of inductance and capacitance are used to generate alternating voltages of a specific frequency and to select a wanted frequency or band of frequencies from the spectrum. *Figure 5.1* contains the diagrams of series and parallel resonant circuits including the resistances which account for the losses present in all circuits. In practice the greatest loss is in the resistance of the inductor, R_L, with a smaller loss, r_c, occurring in the dielectric of the capacitor.

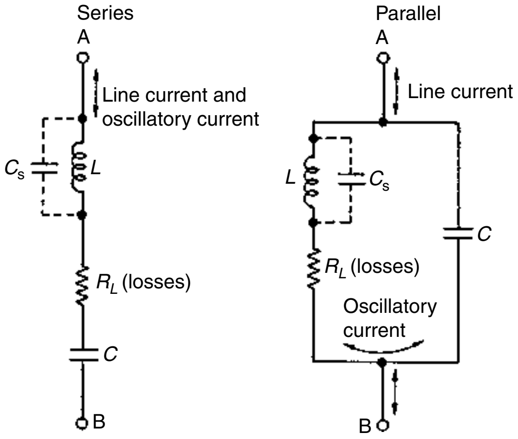
Series
A
Line current and oscillatory current
C_s
L
R_L (losses)
C
B
Parallel
A
Line current
L
C_s
C
R_L (losses)
Oscillatory current
B

Figure 5.1 Series and parallel resonant circuits

5.1.1 Series resonance

Off resonance, the series circuit exhibits a high impedance to a voltage applied across A and B. This impedance is formed by the vectorial addition of the reactances of the inductance and capacitance at the applied frequency plus the resistances. Ignoring the dielectric losses and the very small stray shunt capacitance C_s, the resonant impedance is given by:

$$Z = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$$

where
R = resistance of components in ohms
$\omega = 2\pi \times$ frequency in hertz
L = inductance in henries
C = capacitance in farads

At resonance the reactances cancel out and the impedance falls to approximately the value of the resistance, R, and a maximum line current will flow. *Figure 5.2* shows the response curves of series and parallel circuits near resonance. The resonant frequency is given by:

$$f = \frac{1}{2\pi\sqrt{LC}}$$

where
f = resonant frequency in hertz
L = inductance in henries
C = capacitance in farads

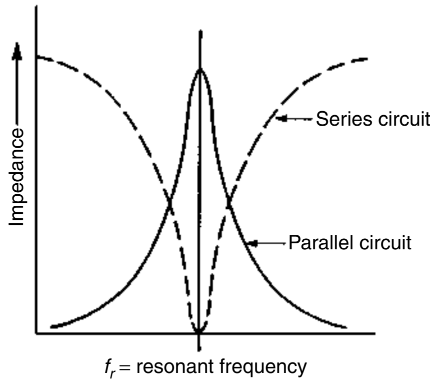

Figure 5.2 Variation of impedance around resonance with series and parallel tuned circuits

5.1.2 Parallel resonance

At resonance, calculated by using the same formula as for a series circuit, the impedance of a parallel tuned circuit is also resistive, and the circulatory current in the circuit is high producing the maximum voltage across the inductance and capacitance. Consequently, at resonance the minimum line current flows. The impedance at resonance or dynamic resistance of a parallel tuned circuit of moderate to high Q is given by:

$$R_d = \frac{L}{CR}$$

where
R_d = dynamic resistance of circuit
R = resistance of components in ohms
L = inductance in henries
C = capacitance in farads

5.2 *Q* factor

The voltages across the inductor and capacitor in a circuit at resonance are substantially opposite in phase (the loss resistances affect the phasing slightly) and cancel each other. The voltage developed across the inductor, usually the lossiest component, compared with the voltage applied in series to the circuit, is a measure of the 'goodness' of the circuit. This ratio is often referred to as the magnification factor, Q, of the circuit. The Q factor is calculated from:

$$Q = \frac{\omega L}{R} \text{ or } \frac{1}{\omega CR}$$

where L and C are in henries and farads respectively.

5.3 Coupled (band-pass) resonant circuits

5.3.1 Methods of coupling

Radio signals carrying intelligence occupy a band of frequencies and circuits must be able to accept the whole of that band whilst rejecting all others. When tuned circuits are coupled together in the correct manner they form such a band-pass circuit. The coupling may be either by mutual inductance between the two inductances of the circuits as in *Figure 5.3* or through discrete electrical components as in *Figure 5.4*.

Mutual inductance can be defined in terms of the number of flux linkages in the second coil produced by unit current in the first coil. The relationship is:

$$M = \frac{\begin{array}{c}\text{flux linkages in 2nd coil}\\ \text{produced by current in 1st coil}\end{array}}{\text{current in 1st coil}} \times 10^{-8}$$

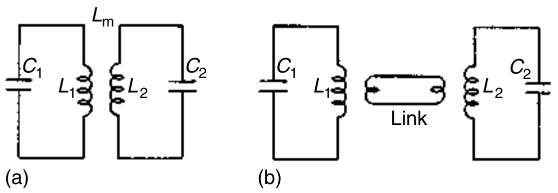

L_m
C_1
L_1
L_2
C_2
(a)
C_1
L_1
Link
L_2
C_2
(b)

Figure 5.3 Mutual inductance (low impedance) coupling

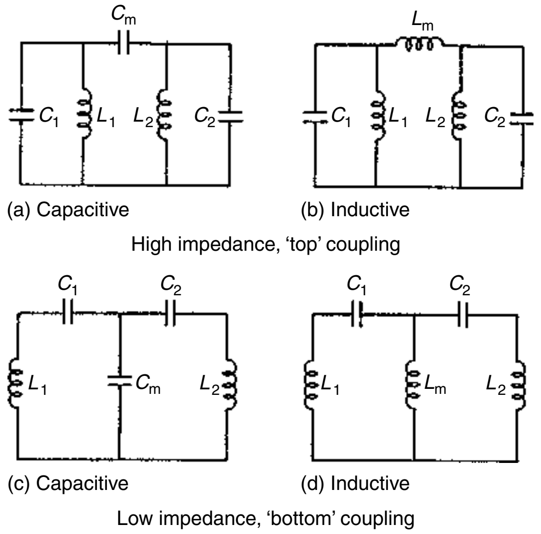

C_m
C_1
L_1
L_2
C_2
(a) Capacitive
L_m
C_1
L_1
L_2
C_2
(b) Inductive
High impedance, 'top' coupling
C_1
C_2
L_1
C_m
L_2
(c) Capacitive
C_1
C_2
L_1
L_m
L_2
(d) Inductive
Low impedance, 'bottom' coupling

Figure 5.4 Electrical coupling

where M = mutual inductance in henries. The e.m.f. induced in the secondary is $e_2 = -j\omega M I_1$ where I_1 is the primary current.

The maximum value of mutual inductance that can exist is $\sqrt{L_1 L_2}$ and the ratio of the actual mutual inductance to the maximum is the coefficient of coupling, k:

$$k = \frac{M}{\sqrt{L_1 L_2}}$$

The maximum value of k is 1 and circuits with a k of 0.5 or greater are said to be close coupled. Loose coupling refers to a k of less than 0.5.

An advantage of coupling using discrete components is that the coupling coefficient is more easily determined. Approximations for k for the coupling methods shown in *Figure 5.4* are:

(a) $k = \dfrac{C_m}{\sqrt{C_1C_2}}$ where C_m is much smaller than (C_1C_2)

(b) $k = \dfrac{\sqrt{L_1L_2}}{L_m}$ where L_m is much larger than (L_1L_2)

(c) $k = \dfrac{\sqrt{C_1C_2}}{C_m}$ where C_m is much larger than (C_1C_2)

(d) $k = \dfrac{L_m}{\sqrt{L_1L_2}}$ where L_m is much smaller than (L_1L_2)

5.3.2 Response of coupled circuits

When two circuits tuned to the same frequency are coupled their mutual response curve takes the forms shown in *Figure 5.5*, the actual shape depending on the degree of coupling.

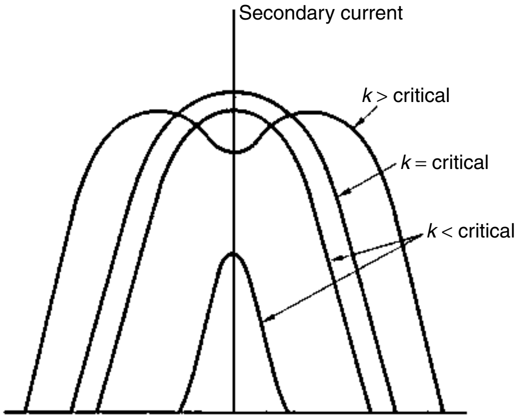

Figure 5.5 Effect of degree of coupling

When coupling is very loose the frequency response and the current in the primary circuit are very similar to that of the primary circuit alone. Under these conditions the secondary current is small and the secondary response curve approximates to the product of the responses of both circuits considered separately.

As coupling is increased the frequency response curves for both circuits widen and the secondary current increases. The degree of coupling where the secondary current attains its maximum possible value is called the critical coupling. At this point the curve of the primary circuit shows two peaks, and at higher coupling factors the secondary response also shows two peaks. The peaks become more prominent and further apart as coupling is increased, and the current at the centre frequency decreases.

References

Langford-Smith, F. (1955). *Radio Designer's Handbook*, Iliffe and Sons, London.
Terman, F.E. (1943). *Radio Engineers' Handbook*. McGraw-Hill, London.

6 Oscillators

6.1 Oscillator requirements

Oscillators generate the frequencies used in radio and electronic equipment. The performance of those which determine the operating frequencies of radio systems is tremendously important. Most oscillators must:

- Generate a precise frequency of high purity.
- Be highly stable, i.e. produce an output constant in frequency and level despite changes in temperature, supply voltage and load.
- Be tunable in frequency.
- Produce minimum noise and microphony (fluctuations in frequency with vibration).

These requirements conflict. A readily tunable oscillator cannot be precise and highly stable, and compromises must be made; either a less stringent specification must be accepted where permissible or the facility for tuning restricted.

Not all oscillators need to produce a pure output, devoid of spurious frequencies. The clock generators in digital circuitry, for instance, produce square waves, but a radio transmitter carrier generator and receiver local oscillator must produce a pure sine wave output if spurious radiations and receiver responses are to be avoided.

6.2 Tunable oscillators

An oscillator is an amplifier with a portion of the output fed back to the input. The feedback must be positive, i.e. in phase with the input, and the loop gain, input back to input via the feedback loop, must be sufficient to overcome the losses in the circuit.

Most radio frequency oscillators – and some audio ones – use inductance and capacitance (LC) tuned circuits as the frequency determining elements. *Figure 6.1* shows two commonly used basic circuits, the Hartley and the Colpitts.

The frequency is determined by the values of L and C_1 (the combined values of C_1 in the Colpitts circuit) and the amount of feedback by the collector choke and C_2. Such circuits produce a very pure output but, principally because the physical dimensions of the frequency

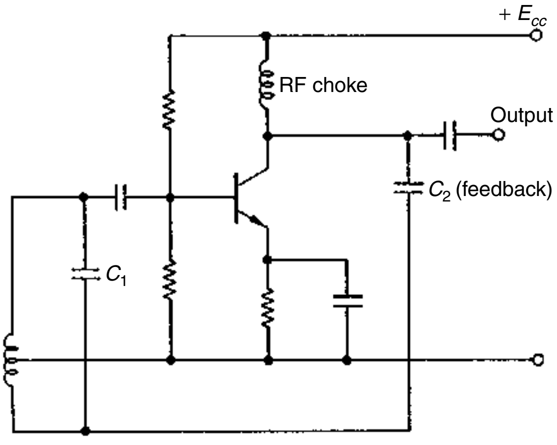

(a) Hartley circuit

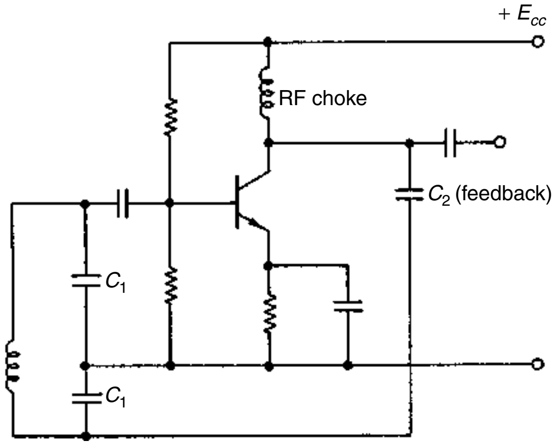

(b) Colpitts circuit

Figure 6.1 Hartley and Colpitts oscillators

determining components change with temperature, the accuracy of the set frequency is doubtful and is not very stable. Temperature compensation can be applied by selecting a capacitor for C_1 with the correct negative temperature coefficient (assuming that the inductance increases with temperature), inserting temperature compensation for the rise in collector current with temperature and stabilizing the supply voltage. In the design of equipment an oscillator should be built into an area of low temperature change.

6.3 Quartz crystal oscillators

The problems of frequency accuracy and stability are largely overcome by using a quartz crystal as the frequency determining element (see Chapter 7).

Figure 6.2(a) shows a circuit for an oscillator using the crystal's parallel resonant mode. In this circuit, the rising voltage developed across R_e on switch-on is applied via C_1 to the base accelerating the rise of current through the transistor. When saturation is reached the voltage across R_e becomes static and the voltage on the base falls, reducing the transistor current. The oscillations are only sustained

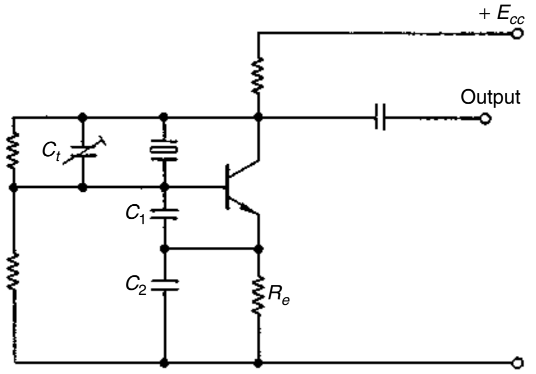

(a) Parallel resonant

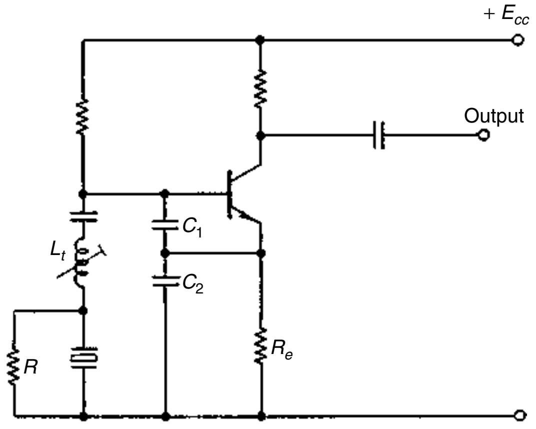

(b) Series resonant

Figure 6.2 Quartz crystal oscillators

at the parallel resonant frequency of the crystal where it presents a high impedance between base and collector. C_t enables the parallel resonance of the crystal to be adjusted to a precise frequency.

Figure 6.2(b) shows a series resonant crystal oscillator and here L_t is the tuning inductor of a Colpitts oscillator. The loop gain is adjusted so that the circuit will oscillate only at the series resonant frequency of the crystal where it presents a very low resistance. At other frequencies the crystal presents an increasing impedance in series with L_t, shunted by R which can be of a low value. When first setting the oscillator, L_t is adjusted, with the crystal short-circuited, for oscillation at a frequency close to that desired. The short-circuit is then removed and L_t used as a fine trimmer.

The same circuit will operate at the parallel resonant frequency of the crystal by making C_1 equal to the crystal load capacitance.

The maximum frequency error permitted by the British Radiocommunications Agency specification MPT 1326 for mobile radio equipment designed for 12.5 kHz channel separation in the band 100–300 MHz is plus or minus 1.5 MHz. This is an overall accuracy of 0.0005% over the temperature range $-10\,^\circ$C to $+55\,^\circ$C. Well-designed standard crystal oscillators meet this specification, but higher stability can be obtained by operating the crystal in an oven at a constant higher temperature.

Until recently equipment which was required to change operating frequency quickly was fitted with several crystals, one for each operating frequency, and a change of frequency was made by selecting the appropriate crystal. Frequency synthesizer circuits are now normally used for such applications.

6.3.1 Overtone oscillators

Piezoelectric crystals can oscillate at more than one frequency. The oscillations of a crystal slab are in the form of *bulk acoustic waves* (BAWs), and can occur at any frequency that produces an odd half-wavelength of the crystal's physical dimensions (e.g. $1\lambda/2$, $3\lambda/2$, $5\lambda/2$, $7\lambda/2$, $9\lambda/2$, where the fundamental mode is $1\lambda/2$). Note that these frequencies are not exact harmonics of the fundamental mode, but are actually valid oscillation modes for the crystal slab. The frequencies fall close to, but not directly on, some of the harmonics of the fundamental (which probably accounts for the confusion). The overtone frequency will be marked on the crystal, rather than the fundamental (it is rare to find fundamental mode crystals above 20 MHz or so, because their thinness makes them more likely to fracture at low values of power dissipation).

The problem to solve in an overtone oscillator is encouraging oscillation on the correct overtone, while squelching oscillations at the fundamental and undesired overtones. Crystal manufacturers can help with correct methods, but there is still a responsibility on the part of the oscillator designer. It is generally the case that overtone oscillators will contain at least one $L-C$ tuned circuit in the crystal network to force oscillations at the right frequency.

6.4 Frequency synthesizers

Frequency synthesizers offer the stability of a quartz crystal oscillator combined with the facility to change operating frequency rapidly. They are essential for equipment operating on trunked or cellular networks where the frequency of the mobiles is changed very rapidly on instructions from the network.

6.4.1 Voltage controlled oscillators

The advent of the variable capacitance diode (varicap), where the capacitance across the diode varies according to the applied DC voltage, made the frequency synthesizer a practicality.

When a varicap diode replaces the tuning capacitor in an oscillator the circuit becomes a DC-voltage-controlled oscillator (VCO). The two varicaps in *Figure 6.3* are used to minimize harmonic production and to obtain a greater capacitance change per volt.

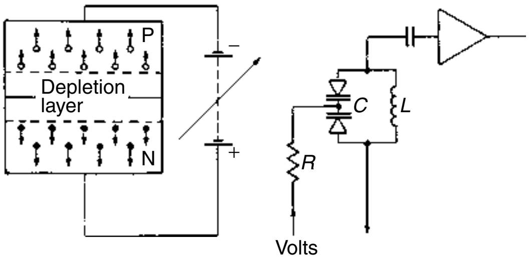

Figure 6.3 Voltage variable capacitance diode

The VCO is the circuit that directly generates the output frequency of a frequency synthesizer but, by itself, inherits the problem of frequency stability. To overcome this, the frequency and phase of the VCO are compared with those of a crystal-controlled high stability oscillator. Any frequency or phase difference between the two oscillators creates a DC voltage of the correct sense to change the frequency of the VCO to agree with that of the crystal oscillator.

While the stability of the crystal oscillator is transferred to the synthesizer output, additional noise is produced close to the operational frequency and the elimination of microphony requires careful physical design.

6.4.2 Phase-locked loops

Figure 6.4(a) and *(b)* are diagrams of a simple phase-locked loop (PLL). The outputs of both the crystal oscillator and the voltage-controlled oscillator are fed to the phase comparator which produces pulses whenever there is a frequency or phase difference between the two inputs. The pulses will be either positive or negative depending on the sense of the difference, and their width is dependent upon the magnitude of the differences. The pulses are then fed to a low pass loop filter which smooths them. If the time constant of the filter is sufficiently long it will completely remove the pulses and produce a DC output proportionate to the input pulse width which is applied to the VCO in the right sense to correct the frequency error. To enable the pulses to swing the VCO frequency in either direction, a small bias voltage of about 4 volts is applied to the varicap.

Identical frequencies have been selected for both oscillators in *Figure 6.4* but in practice this is seldom the case. More frequently, the

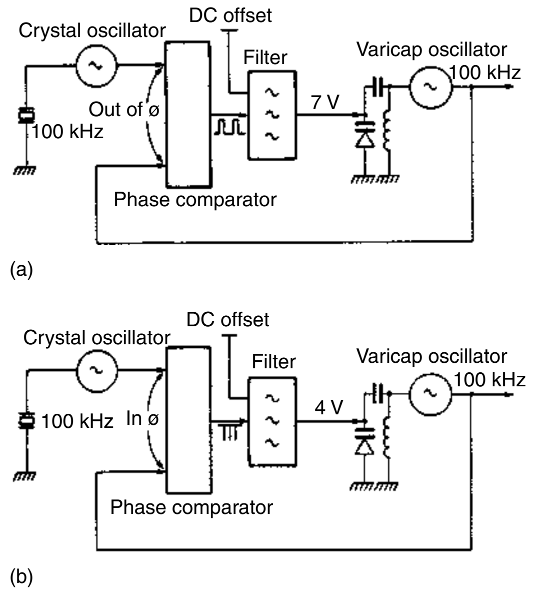
DC offset
Crystal oscillator
Filter
Varicap oscillator
100 kHz
7 V
Out of ø
100 kHz
Phase comparator
(a)
DC offset
Crystal oscillator
Filter
Varicap oscillator
100 kHz
4 V
100 kHz
In ø
Phase comparator
(b)

Figure 6.4 Simple phase-locked loop

VCO runs at a higher frequency than the crystal oscillator and a divider is used to equate the frequencies applied to the phase comparator (*Figure 6.5*). Changing the division ratio provides a convenient means of tuning the oscillator.

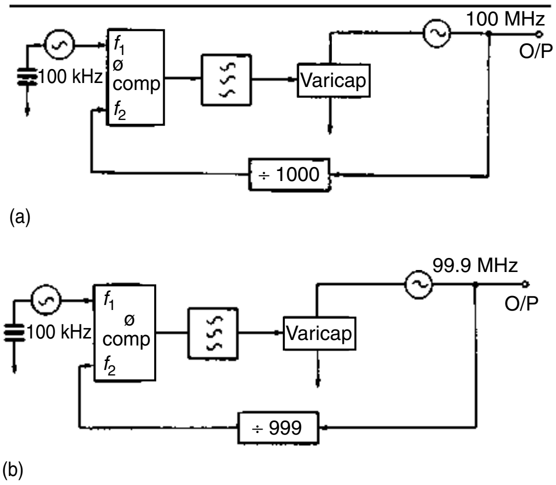

Figure 6.5 Frequency variable phase-locked loop

While the division ratio is 1000, the VCO will run at 100 MHz but if the division ratio is changed to 999 the comparator will produce pulses which, when converted to a DC voltage by the loop filter, will change the frequency of the VCO to 99.9 MHz, and the loop will lock at this new frequency.

The design of the loop filter is critical. Too long a time constant lengthens the settling time when changing frequency, yet if it is too short any deliberate frequency modulation will be removed. In practice, a relatively long time constant is chosen which is shortened by a 'speed up' circuit introduced whenever a channel change is called for.

The above values would enable a radio operating on a system with a channel separation of 100 kHz to change channel, but mobile radio channel separations are 25 kHz, 12.5 kHz or even 6.25 kHz at frequencies from 50 MHz to at least 900 MHz. To change channel at these frequencies a synthesizer must use a high division ratio. With a reference frequency applied to the comparator of 6.25 kHz and an operating frequency of, say, 450 MHz, the frequency select divider must have a ratio of 72 000, and be programmable in steps of 1 with a minimum operating speed of at least 900 MHz. A problem is then that the technology capable of meeting these requirements, emitter coupled logic (ECL), is power hungry, and the preferred LSI low power technology, CMOS, has a maximum operating speed of about 30 MHz. A simple ECL pre-scaler to bring the VCO frequency to about 30 MHz needs a ratio of 20 (500 MHz to 25 MHz). However, every change of 1 in the CMOS divider ratio then changes the total division ratio by 20. The solution is to use a dual modulus pre-scaler.

6.4.3 Dual modulus pre-scaler

The division ratio of the dual modulus pre-scaler (*Figure 6.6*) is programmable between two consecutive numbers, e.g. 50 and 51 (P and $P + 1$) and, in conjunction with two CMOS dividers, ÷A and ÷N, provides a fully programmable divider.

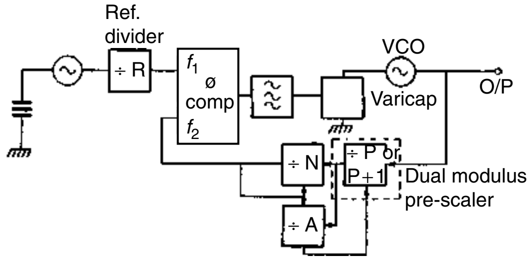

Figure 6.6 Programmable frequency synthesizer

The A and N dividers are pre-loaded counters. These count down and when the count value reaches zero they produce an output which changes the division ratio of the pre-scaler. The total division ratio, N_t, is decided by the initial programmed contents of the A and N counters and the setting of the pre-scaler. The initial content of the A counter must be less than that of the N counter.

Consider the pre-scaler set to divide by $P + 1$. For every count of $P + 1$, the contents of the A and N counters are reduced by 1 until the contents of the A counter are zero. The difference between the original contents of the A and N counters, $N - A$, remains in the N counter, and the total count, N_t, up to now, is $A(P + 1)$. At this point the division ratio of the pre-scaler is changed to P. Now, for every P count, the contents of the N counter are reduced by 1 until zero is reached. Under these conditions the total division ratio is given by:

$$
\begin{aligned}
N_t &= A(P + 1) + (N - A)P \\
&= AP + A + NP - AP \\
&= NP + A
\end{aligned}
$$

For example, let $P = 50$ so $P + 1 = 51$, let $N = 10$ and $A = 7$. Then:

$$N_t = 10 \times 50 + 7 = 507$$

Now, change A to 6:

$$N_t = 10 \times 50 + 6 = 506$$

a change of N_t by 1.

Programming a divider

Example:

$$\text{VCO frequency} = 455.6\,\text{MHz}$$

$$\text{Reference frequency} = 12.5\,\text{kHz}$$

Calculate N_t, and the numbers which must be programmed into the A and N counters, assuming $P = 50$:

1. Calculate $N_t = 455.6\,\text{MHz}/12.5\,\text{kHz} = 36\,448$.
2. Divide N_t by P: $36\,488/50 = 728.96$. Make $N = 728$.
3. For A, multiply fraction by P: $0.96 \times 50 = 48$.
4. Check $N_t = NP + A = 728 \times 50 + 48 = 36\,448$.
   Change A to 47: $NP + A = 728 \times 50 + 47 = 36\,447$.
   $36\,447 \times 12.5\,\text{kHz} = 455.5875\,\text{MHz}$, the adjacent 12.5 kHz channel.

6.4.4 Direct digital synthesis

A method of direct digital frequency synthesis replaces the voltage-controlled oscillator by a numerically controlled oscillator (NCO) where the function of the VCO is digitally synthesized.

The direct digital synthesizer generates an analogue sine wave from digital sine wave samples applied to a digital to analogue (D/A) converter. There are limitations to the method in terms of bandwidth and spectral purity.

6.5 Caesium and rubidium frequency standards

Where extra high stability is required for, say, laboratory standards or in quasi-synchronous wide area coverage systems, oscillators utilizing the atomic resonances of substances like caesium and rubidium, although expensive, may be employed.

Caesium oscillators are used to provide standard frequencies such as 1, 5 and 10 MHz with accuracies of $\pm 7 \times 10^{-12}$ over a temperature range of 0 to 50 °C with a long-term stability of 2×10^{-12}.

Rubidium oscillators are used to provide secondary standards and in some quasi-synchronous radio systems. Their accuracy is less than that of caesium, the long-term drift being of the order of 1×10^{-11} per month.

References

Belcher, R. *et al*. (1989). *Newnes Mobile Radio Servicing Handbook*. Butterworth-Heinemann, Oxford.

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

7 Piezo-electric devices

7.1 Piezo-electric effect

When electrical stress is applied to one axis of a quartz crystal it exhibits the piezo-electric effect: a mechanical deflection occurs perpendicular to the electric field. Equally, a crystal will produce an e.m.f. across the electrical axis if mechanical stress is applied to the mechanical axis. If the stress is alternating – the movement of the diaphragm of a crystal microphone is an example – the e.m.f. produced will be alternating at the frequency of the movement. If the stress alternates at a frequency close to the mechanical resonance of the crystal as determined by its dimensions, then large amplitude vibrations result. Polycrystalline ceramics possess similar qualities.

Quartz crystals used for radio applications are slices cut from a large, artificially grown crystal. The slices are then ground to the appropriate size to vibrate at a desired frequency. The performance of an individual slice – the crystal as the end user knows it – depends upon the angle at which it was cut from the parent crystal.

Each crystal slice will resonate at several frequencies and if the frequency of the stimulus coincides with one of them the output, electrical or mechanical, will be very large.

The vibrations occur in both the longitudinal and shear modes, and at fundamental and harmonic frequencies determined by the crystal dimensions.

Figure 7.1A shows a typical natural quartz crystal. Actual crystals rarely have all of the planes and facets shown. There are three *optical axes* (X, Y and Z) in the crystal used to establish the geometry and locations of various cuts. The actual crystal segments used in RF circuits are sliced out of the main crystal. Some slices are taken along the optical axes, so are called Y-cut, X-cut and Z-cut slabs. Others are taken from various sections, and are given letter designations such as BT, BC, FT, AT and so forth.

7.1.1 Equivalent circuit of a quartz crystal

A quartz crystal behaves similarly to a very high Q tuned circuit and the equivalent circuit of a crystal is shown in *Figure 7.1B*.

C_1 and L_1 are equivalent to the inductance and capacitance of a conventional tuned circuit and R_1 represents the losses in the quartz

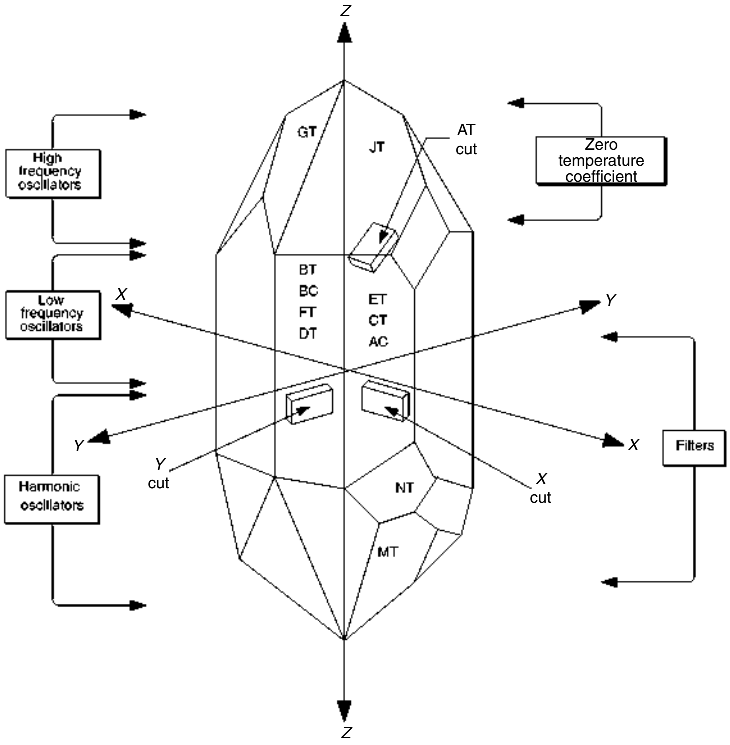
Z
GT
JT
AT cut
Zero temperature coefficient
High frequency oscillators
Low frequency oscillators
Harmonic oscillators
BT
BC
FT
DT
ET
CT
AC
X
Y
Y
X
Y cut
X cut
Filters
NT
MT
Z

Figure 7.1A Natural quartz crystal

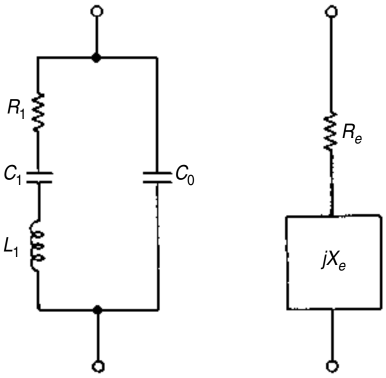
R1
C1
L1
C0
Re
jXe

Figure 7.1B Equivalent circuit of a crystal

and the mounting arrangements. C_0, typically 3–15 pfd, represents the shunt capacitance of the electrodes in parallel with the can capacitance. If the oscillatory current is considered, the resonant frequency is decided by the values of C_0 in series with C_1, L_1 and R_1, and all crystals basically resonate in a series mode. *Figure 7.2* illustrates the changes in impedance close to resonance. However when a high impedance, low capacitance, load is connected across the crystal terminals it behaves as a parallel tuned circuit exhibiting a high resistance at the resonant frequency. A crystal operating in the parallel mode oscillates at a higher frequency than that of series resonance.

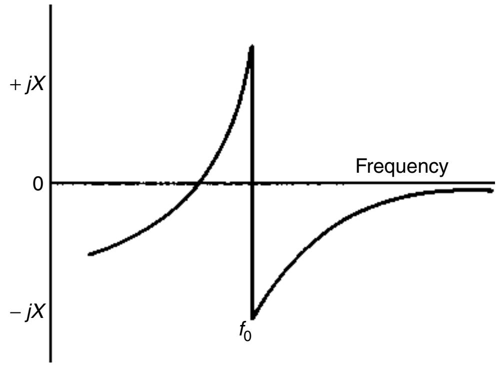

Figure 7.2 Crystal reactance close to resonance

A crystal will resonate at its fundamental frequency or at one or more of its harmonics. As the desired resonant frequency is increased, a crystal slice operating at its fundamental frequency becomes extremely thin and fragile. Consequently, overtone crystals are composed of larger slices of quartz operating close to, but not necessarily at, an exact harmonic of the fundamental frequency. Crystals operating at the 3rd, 5th and 7th harmonics are often employed at frequencies above approximately 25 MHz.

7.2 Quartz crystal characteristics

7.2.1 Resonant frequency

The resonant frequency is determined by the mass of the finished crystal which can be adjusted by grinding and the deposition of gold or other metal onto the crystal faces during manufacture. The adjustment is made to suit the intended method of operation, series or parallel, and at a specific temperature, usually 25°C. When parallel mode is specified, allowance is made for the load or circuit capacitance, usually 20–30 pfd, in parallel with C_0.

7.2.2 Frequency stability

Temperature coefficient

A crystal's resonant frequency varies with temperature and this temperature coefficient is determined by the angle at which the slice was cut from the parent crystal. Commonly used cuts are AT and BT. Because of its better performance AT is the most common.

Typical examples of the temperature coefficients for these are shown in *Figure 7.3*.

The temperature coefficient is specified, usually in parts per million (ppm) per degree C, or as a percentage, over a defined temperature range. The standard European temperature range is −10°C to +60°C. A crystal designed for a restricted temperature range has a better

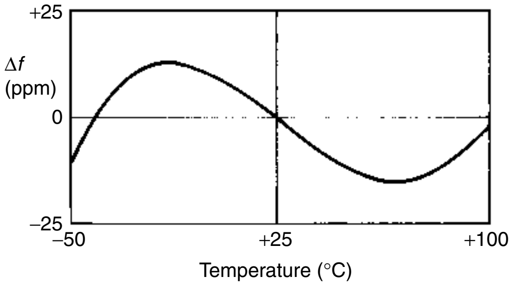

AT cut crystal

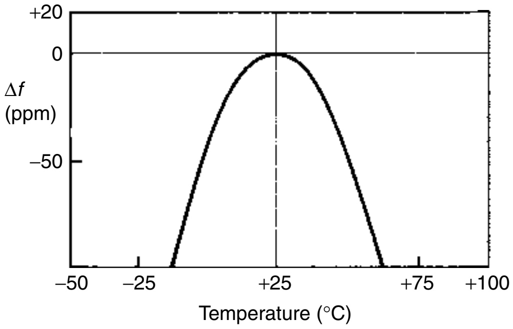

BT cut crystal

Figure 7.3 Frequency vs. temperature curves AT and BT cut crystals

stability over that range than one designed for operation over a wide temperature range will have when used over a restricted range.

For higher frequency stability crystals may be operated in a temperature-controlled oven operating at a more constant high temperature.

Common frequency tolerance specifications are ±0.005% or 0.0025% from −55°C to +105°C. These include the frequency errors from all sources, including the calibration tolerance; thus, the temperature coefficient is slightly better than these figures.

Ageing

The resonant frequency shifts with age from that set at production, following a curve similar to that in *Figure 7.4*. Initially the frequency shift for a given period of time is rapid but slows with age. The frequency may shift in either direction, and although it is possible to specify crystals ageing in one direction – high stability oscillators for quasi-synchronous transmission systems is an application – they are selected from a batch, not specifically manufactured. Once a crystal has been operated, a subsequent long period of inactivity can produce a glitch in the ageing curve followed by a higher rate of change for a short time.

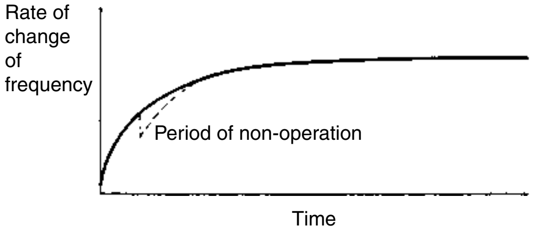

Figure 7.4 Effect of ageing

7.2.3 Load capacitance and pullability

When a crystal is operated in the parallel mode across a low capacitance load the results are a higher frequency and larger output voltage to the load. Increasing the load capacitance causes a reduction in frequency approaching that of series resonance.

The change in frequency that can be achieved by varying the load capacitance – a small trimmer capacitor is often connected across the crystal for this purpose – is the crystal's pullability. A typical pullability is from −1 ppm/pfd to −20 ppm/pfd for a total shunt capacitance of 40 pfd ($C_0 + C_{\text{load}}$).

7.2.4 Activity, effective series resistance (ESR) and Q

All these characteristics are interrelated. A crystal's activity, its vibrational response, can be quoted in terms of the effective series resistance. A higher effective series resistance implies lower activity, lower output and lower Q. The usual range of ESRs is from 20 Ω to 100 Ω although higher values occur in some low frequency crystals. Some manufacturers may quote activity levels for crystals for use in a parallel mode as effective parallel resistance (EPR). The EPR is the value of the resistor which, if connected in lieu of the crystal in an oscillator, would give the same output level as the crystal. The higher the EPR, the greater the crystal activity and Q.

7.2.5 Spurious responses

Crystals will resonate at frequencies other than those of the fundamental and harmonic modes for which they were designed; *Figure 7.5* shows the overtone (harmonic) and some typical spurious responses. The spurious responses of overtone crystals can occur with very little separation from the desired overtone frequency requiring very careful oscillator design if they are to be avoided.

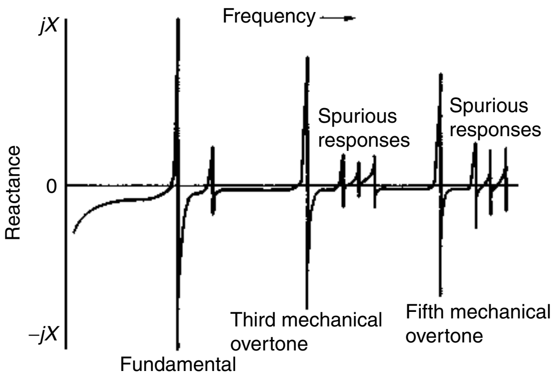

Figure 7.5 Overtone response of a quartz crystal

7.2.6 Case styles

A wide range of mounting styles is available. The American military nomenclature is widely used to describe them and *Figure 7.6* shows the outlines of some of the more popular styles.

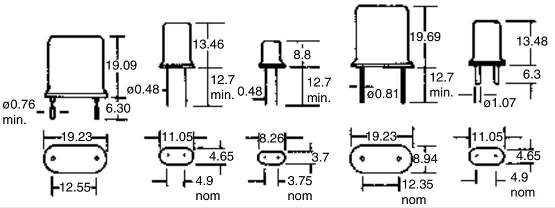
19.09
ø0.76 min.
ø0.48
6.30
19.23
12.55
13.46
12.7 min.
11.05
4.65
4.9 nom
0.48
8.8
12.7 min.
8.26
3.7
3.75 nom
19.69
12.7 min.
ø0.81
19.23
8.94
12.35 nom
13.48
6.3
ø1.07
11.05
4.65
4.9 nom

Solder seal	HC6/U AA	HC18/U J	– –	HC33/U BF	HC25/U CX	MIL IEC
Cold weld	HC36/U DNC	HC43/U DPC	HC45/U EBC	HC47/U AC	HC42/U DQ	MIL IEC
Glass	HC27/U DA	HC26/U CY	– –	– –	HC29/U CZ	MIL IEC
Res weld	HC48/U DNR	HC43/U DPR	HC45/U EBR	HC51/U RF	HC50/U RX	MIL IEC

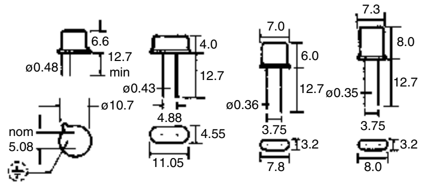
6.6
12.7 min
ø0.48
ø10.7
nom 5.08
4.0
12.7
ø0.43
4.88
4.55
11.05
7.0
6.0
12.7
ø0.36
3.75
3.2
7.8
7.3
8.0
12.7
ø0.35
3.75
3.2
8.0

Solder seal	– –	– –	– –	– –		MIL IEC
Cold weld	HC35/U DKC	– –	– –	– –		MIL IEC
Glass	– –	– –	– –	– –		MIL IEC
Res weld	HC35/U DKR	HC49/4H –	– UM-5	– UM-1		MIL IEC

Figure 7.6 Crystal case styles

7.3 Specifying quartz crystals

The details which must be specified when ordering crystals are:

1. *Frequency*. Normally specified in kHz up to 9999.999 kHz and in MHz from 10.0 MHz upwards except for integer values which are all specified in MHz. The frequency must be described to seven significant figures, otherwise any figure that might follow those given will be taken as zero.
2. *Mounting or holder style*.
3. *Frequency tolerance*. This is the cutting or calibration tolerance acceptable at 25°C. It should be borne in mind that cost rises with increased manufacturing accuracy and a slight adjustment (pullability) is possible in the circuit.

4. *Frequency stability*. Normally specified as a plus or minus value measured over a defined temperature range. A crystal designed for a restricted range has a better performance over that range than one designed for a wider range so it is important not to overspecify.
5. *Temperature range*. The range over which the crystal is required to operate and meet the performance specified in 4. Standard temperature ranges are:

   0 to 5°C
   −10 to 60°C
   −20 to 70°C
   −30 to 80°C
   −40 to 90°C
   −55 to 105°C
   −55 to 125°C

   It is sufficient when ordering from some manufacturers to quote only the lower temperature limit.

   For ovened operation the quoted figure, say 80°C, would denote the oven temperature.
6. *Circuit condition*. This specifies the shunt capacitance that the circuit will place across the crystal in parallel mode operation.
7. *Drive level*. The maximum power that the crystal can safely dissipate. 1 mW is a typical value for crystals used in radio transmitters and receivers.

A typical specification therefore reads:

16.66667 MHz	HC49	20	30	10	30
1	2	3	4	5	6

referring to the items listed above

When the crystal is for operation in series mode, it is usually sufficient to replace the last figure with ‘S’. The drive level is not normally specified in the ordering details.

7.4 Filters

Both quartz and ceramic materials are used in the production of radio frequency filters. Ceramic filters do not have the same performance as quartz but have the advantages of a lower cost. They are used at lower frequencies and where the higher stability and lower spurious responses of quartz are not essential.

Crystal filters are obtainable at frequencies up to about 45 MHz. Most of these use either a number of discrete crystals arranged in the form of a lattice or a monolithic structure. A single crystal will behave as an extremely narrow band filter and it is possible to use a crystal bar in this way down to a very few kilohertz.

The characteristics of filters can be divided into groups affecting the performance (*Figure 7.7*).

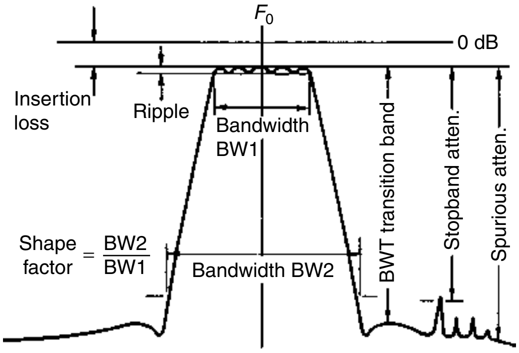

Figure 7.7 Filter characteristics

7.4.1 Passband performance

- *Insertion loss*. The loss at centre frequency, in dB, resulting from the insertion of the filter in a transmission system.
- *Flat loss*. The insertion loss at the frequency of minimum loss within the passband.
- *Attenuation*. The loss of a filter at a given frequency measured in dB.
- *Passband (bandwidth, BW_1)*. The range of frequencies attenuated less than a specified value, typically 3 or 6 dB.
- *Centre frequency* (f_0). The arithmetic mean of the passband limits.
- *Fractional bandwidth*. A specified frequency, typically the minimum loss point or F_0, from which all attenuation measurements are made.
- *Ripple*. The amplitude difference, in dB, between the maximum peak and minimum passband valley. Both the peak and the valley are defined by a surrounding change in slope, i.e. sign of the amplitude response. This is very important as a high ripple, particularly between a peak and the adjacent trough, produces rapid phase changes as the signal moves across the

passband resulting in audio distortion and corruption in digital signals.

7.4.2 *Stopband performance*

- *Attenuation*. The output of a filter at a given frequency relative to the defined insertion loss reference.
- *Stopband*. The range of frequencies attenuated by a greater amount than some specified minimum level of attenuation.
- *Transition band (bandwidth,* BW_2*)*. The range of frequencies differently attenuated between the passband and stopband limits.
- *Shape factor*. The ratio of the bandwidth at some point within the transition region, typically 60 dB, to the specified passband bandwidth. It is given by:

$$\text{Shape factor } \frac{BW_2}{BW_1}$$

- *Spurious attenuation*. The specified minimum level of attenuation received by all non-harmonic related resonances of each crystal resonator within the filter network.

7.4.3 *Time domain performance*

- *Insertion phase*. The phase shift at the output load (measured at the reference frequency) resulting from the insertion of the filter.
- *Differential phase*. The measurement of phase at a given frequency relative to the phase at the reference frequency.
- *Phase linearity*. The phase error in degrees between the phase points and a straight line drawn through the phase points.
- *Group delay*. The time by which a signal will be delayed before it appears at the filter output, i.e. the derivative of phase with respect to frequency.
- *Differential delay*. The measurement of delay at a given frequency relative to the reference frequency.

7.4.4 *Source and load impedance*

- *Source impedance*. The impedance of the circuit driving the filter, measured at the reference frequency.
- *Load impedance*. The impedance of the circuit terminating the filter at its output, measured at the reference frequency.

7.4.5 Non-linear effects

- *Maximum input level*. The driving point power, voltage or current level above which intolerable signal distortion or damage to the device will result.
- *Drive level stability*. The ability of the filter to return within a specified tolerance of its original insertion loss, at a specified drive level, after experiencing changing environmental and/or drive level conditions.
- *Drive level linearity*. The maximum permissible variation in insertion loss, per dB change in drive level, measured over a specific dynamic range.
- *Inband intermodulation distortion*. The attenuation, in dB, of third and higher order signal products, inband, relative to the power level of two signals placed within the passband.
- *Out-of-band intermodulation distortion*. The attenuation, in dB, of third and higher order signal products, inband, relative to the power level of two signals placed in the stopband, or one signal in the transition region and the other in the stopband.

A manufacturer's specifications for two stock 10.7 MHz filters are given in *Table 7.1*.

Table 7.1 Manufacturers' specifications for two stock 10.7 MHz filters

Centre freq.	*Passband width (plus/minus)*	*Attenuation bandwidth*	*Ripple (max)*	*Ins. loss (max)*	*Term. impedance ΩS/pfd*
10.7 MHz	3.75 kHz (6 dB)	8.75 kHz (45 dB) 12.5 kHz (60 dB)	2.0 dB	3.0 dB	1.5k/1
10.7 MHz	7.5 kHz (6 dB)	15.0 kHz (60 dB) 20.0 kHz (80 dB)	2.0 dB	4.0 dB	3k/1

7.5 SAW filters and resonators

The piezo-electric effect that some ceramic materials such as lithium niobate exhibit enables useful applications, such as pressure sensors and audible alarms. When piezo-electric material has an electric field applied across it, usually through metal electrodes, the material changes shape. This often takes the form of local material expansion or contraction (depending on the polarity of the applied voltage). In the case of the audible alarm, the material expansion and contraction modulates the air pressure to give an audible tone if the electric field is alternating at audio frequencies.

One form of the piezo-electric effect is the SAW (surface-acoustic-wave) phenomenon, in which signals travel due to an acoustic wave in the ceramic material. To see this effect the piezo-electric material has one side bonded to a metal plate. The upper surface has metal electrodes applied to it, to provide an electric energizing force. The surface wave is created when a material is forced to expand (by the piezo-electric effect) beneath the metal electrodes. Physical bonds between molecules then force adjacent material to expand and this expansion propagates through the material.

The piezo-electric effect is bi-directional, so that if the material between electrodes changes shape, an electric field is generated between the electrodes. This is the effect seen when a piezo-electric gas lighter is used: a sudden force across a piezo-electric element generates high voltages that force a spark to be produced. In the SAW device, the acoustic wave travels along the material and deforms the piezo-electric material between a second set of electrodes, which are located at the other end of the piezo-electric material. A voltage is then formed between these electrodes. By having two sets of electrodes, known as transducers, at opposite ends of a piece of piezo-electric material, a signal can be transmitted through the material by applying a voltage at one end and detecting it at the other.

The range of frequencies that propagate through the piezo-electric material can be controlled by suitable transmit and receive transducer spacing. The propagation frequencies can be further controlled by applying additional metallized areas (described later) between the transmit and receive transducers. Thus SAW devices could be used to make compact and low-cost filters.

SAW resonators use piezo-electric material that is free to vibrate in one direction. The speed of wave propagation and the dimensions of the material are such that the wave reflects back and forth, resonating at a certain frequency. This can replace quartz crystals in many oscillator circuits, where the frequency accuracy is not critical.

7.5.1 Fundamentals of SAW transversal filters

In its simplest form, a transversal SAW filter consists of two transducers with inter-digital arrays of thin metal electrodes deposited on a piezoelectric substrate, such as quartz or lithium niobate (see *Figure 7.8*). The electrodes that comprise these arrays are arranged to have alternate polarities, so that an RF signal voltage of the proper

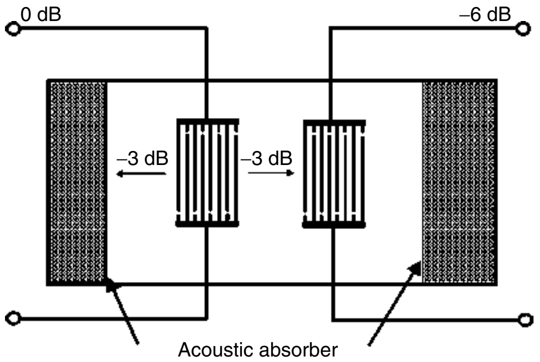

Figure 7.8 Simple transversal SAW filter configuration

frequency applied across them causes the surface of the crystal to alternately expand and contract along its length. This generates the Rayleigh wave, or surface wave, as it is more commonly called.

These inter-digital electrodes are generally spaced at half or quarter wavelength of the operating centre frequency. Since the surface wave or acoustic velocity is 10^5 slower than the speed of light, an acoustic wavelength is much smaller than its free space electromagnetic counterpart. For example, a CW signal at 1000 MHz with a free space wavelength of 300 mm would have a corresponding acoustic wavelength of about 3 μm. This wavelength compressing effect results in the SAW filter's unique ability to provide considerable signal processing or delay in a very small volume.

The wavelength compressing effect of the piezo-electric material has the effect of producing physical limitations at very high and low frequencies. At very high frequencies the electrodes become too narrow to fabricate with standard photo-lithographic techniques. At low frequencies the devices become impracticably large. Consequently, SAW devices are most typically used over the frequency range 10 MHz to about 3 GHz.

The basic SAW transducer is a bi-directional radiator. That is, half of the power (or -3 dB) is directed towards the output transducer while the other half is radiated towards the end of the crystal and is lost. By reciprocity, only half of the intercepted acoustic energy at the output is reconverted to electrical energy; hence, the inherent 6 dB loss associated with this structure (refer to *Figure 7.8*). Numerous second-order effects, such as coupling efficiency, resistive losses, and impedance mismatch, raise the insertion loss of practical filters to about 15 to 30 dB.

A new low-loss structure is the single-phase unidirectional transducer (SPUDT). Unlike the traditional three-phase unidirectional filters, the SPUDT devices are generally as straightforward to fabricate as ordinary bi-directional transducers. In addition, the required impedance matching network is greatly simplified, usually consisting of an L-C network on each port. Most SPUDT structures contain acoustic reflectors within the inter-digital pattern, illustrated schematically in *Figure 7.9* as wider electrodes. These internal reflectors serve to redirect most of the acoustic energy that is normally lost in a conventional bi-directional device. Second-order effects tend to limit the practical insertion loss to about 5 to 12 dB.

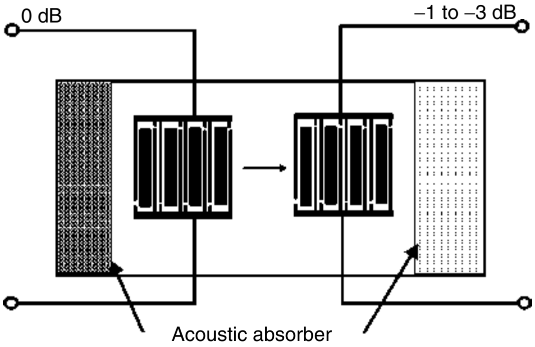

Figure 7.9 Single-phase unidirectional filter configuration

7.5.2 SAW device design

A better understanding of SAW device operation can be obtained by understanding how they are designed. SAW filters use finite impulse response design techniques, which are very similar to those used for digital filters. Hence, the Fourier transform is used to relate the time and frequency responses of the transducers and resultant filter. The desired total frequency response characteristics are used to find the impulse responses for the two transducers. These two impulse response shapes are then etched onto the surface of a metallized piezo-electric substrate to form transducers. When an impulse is applied to the transducer, the output wave shape is the same as the shape of the metallized transducer pattern.

To illustrate this relationship, the following examples may be valuable. First, consider the case of the uniform-overlap (or unweighted) transducer as shown in *Figure 7.10*. Since all fingers have the same

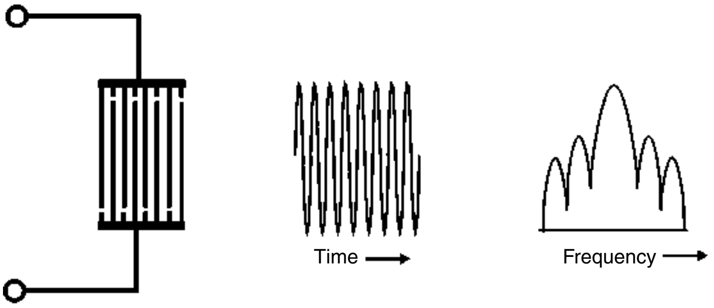

Unweighted transducer

Impulse response

Frequency response

Figure 7.10 Uniform-overlap transducer

length, the impulse response envelope of this transducer is a rectangle. The Fourier transform of a rectangle is a $\sin(x)/x$ function with a 4 dB bandwidth roughly equal to the reciprocal of its time length. This illustrates the important trade-off between bandwidth and size, the narrower the bandwidth: the longer the transducer.

The $\sin(x)/x$ response has 13 dB first side-lobe attenuation, which results in 26 dB of rejection when two of these unweighted transducers are cascaded on the substrate. The 26 dB frequency rejection and poor skirt selectivity make the unweighted transducer suitable for delay lines, but unsuitable for most filter applications where more rejection is needed close to the passband edge.

According to Fourier transform theory, to obtain the rectangular frequency response, a transducer must be built that is infinitely long with electrode lengths that have $\sin(x)/x$ weighting. This is impossible, but the principle behind this example will always hold true: the steeper the skirts (the more rectangular the response) for a given bandwidth, the longer the device must be. The longest possible SAW device would exhibit a narrow bandwidth with a very steep roll-off. The transducer pattern must be limited in length but a gradual tailing off to zero length of the outer electrodes provides close to ideal performance (akin to the windowing function in digital FIR filters).

References

Crystal Product Data Book (1993). International Quartz Devices Ltd, Crewkerne, Somerset.

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

8 Bandwidth requirements and modulation

8.1 Bandwidth of signals at base band

8.1.1 Analogue signals

The amount of information and the speed at which it is transmitted determines the base bandwidth occupied by a signal. For analogue signals, the base bandwidth is the range of frequencies contained in the signal; it is not the same as that occupied by a radio frequency carrier modulated by the signal. Examples of base bandwidths are given in *Table 8.1*.

Table 8.1 Base bandwidths

Application	*Frequency range (Hz)*
Speech	
High fidelity reproduction	15–15 000
Good fidelity	150–7000
Public address	200–5000
Restricted bass and treble	500–4000
Toll quality (good quality telephone line)	300–3400
Communications quality (radio communication)	300–3000
Mobile radio (12.5 kHz channel separation)	300–2700
Music (for FM broadcasting)	30–15 000
Video	60 Hz–4.2 MHz

8.1.2 Digital signals

Bit rate (b/s) and baud rate are terms used to specify the speed of transmitting digital information. Where the duration of all the signalling elements is identical the terms are synonymous, but not where the duration of the information bits differs.

As the term implies, the bit rate is the number of bits transmitted per second but the baud rate (after J.M.E. Baudot, the code's inventor) is the reciprocal of the length, in seconds, of the shortest duration signalling element. *Figure 8.1(a)* shows a binary code pattern where

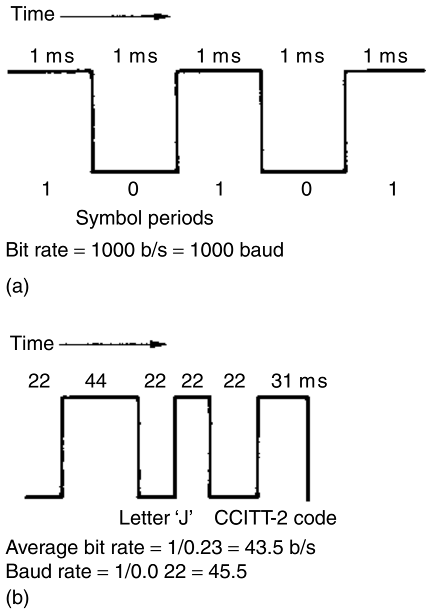
Time
1 ms 1 ms 1 ms 1 ms 1 ms
1 0 1 0 1
Symbol periods
Bit rate = 1000 b/s = 1000 baud
(a)
Time
22 44 22 22 22 31 ms
Letter 'J' CCITT-2 code
Average bit rate = 1/0.23 = 43.5 b/s
Baud rate = 1/0.0 22 = 45.5
(b)

Figure 8.1 Bit rate and baud rate

all the bits are of equal duration, in this case 1 millisecond; the bit rate is 1000 per second and the baud rate is $1/0.001 = 1000$ also.

On telegraphy systems all the bits may not be of the same duration and *Figure 8.1(b)* shows the pattern for the letter 'J' in the CCITT-2 code as used for teletype. In this code a letter is composed of 5 elements, each of 22 ms duration, but each letter is preceded by a space of 22 ms and followed by a mark of 31 ms. The duration of each character is 163 ms – the time for 7.5 elements – but is comprised of only 7 bits.

The baud rate is $1/0.022 = 45.5$ baud.

The average bit duration is $163/7 = 23.29$ ms.

The average bit rate is $1/0.023 = 43.5$ bits per second.

If the bandwidth were under consideration the baud rate, being faster, would be the figure to use.

A stream of binary coded information is composed of pulses where, say, a pulse (mark) represents digit 1 and the absence of a pulse (space) represents digit 0. The highest frequency contained in the information is determined by the bit – or baud – rate. Because a series of 1s or 0s may be consecutive in the data stream, the pulse repetition rate will vary throughout the message although the bit rate will be constant, and over a long period of time as many spaces will be sent as pulses. Transmitting either a stream of spaces or of marks requires no bandwidth; it is only when a change of state occurs that frequencies are produced. In the duration of one cycle (*Figure 8.1(a)*) two bits may be carried and the maximum *fundamental* frequency contained in the wave is one-half the number of bits per second, i.e. the channel capacity, bits/second, equals twice the bandwidth in hertz.

8.1.3 Channel capacity – Hartley – Shannon theorem

Channel capacity as stated by Hartley's law is, in the absence of noise:

$$C = 2\delta f \log_2 N$$

where
C = channel capacity, bits per second
δf = channel bandwidth, Hz
N = number of coding levels (2 in binary system)

When noise is present, the channel capacity calculated according to the Hartley–Shannon theorem is:

$$C = \delta f \log_2(1 + S/N)$$

where S/N = the ratio of total signal power to total noise power at the receiver input within the bandwidth, δf.

8.2 Modulation

For radio transmission, the low frequency information signal is carried on a radio frequency wave and it must change (modulate) that carrier. The modulation may change the amplitude, frequency or phase of the carrier. Modulation aims to achieve:

1. the transfer of information with the minimum distortion or corruption
2. the modulation of the carrier with the minimum loss of power
3. efficient use of the frequency spectrum.

8.3 Analogue modulation

8.3.1 Amplitude modulation (AM)

There are a number of methods of modulation where the amplitude of the carrier is varied by the information signal but the most commonly used is double sideband AM (DSB). *Figure 8.2* shows a radio frequency carrier modulated by a low frequency signal.

The amount or depth of modulation is expressed as percentage ratio, $m\%$, of the maximum to minimum amplitude:

$$\begin{aligned}\text{Mod. depth} &= m\% \\ &= \frac{\text{max. amplitude} - \text{min. amplitude}}{\text{max. amplitude} + \text{min. amplitude}} \times 100\%\end{aligned}$$

When the modulation is increased to the point where the minimum amplitude falls to zero, 100% modulation occurs. Any further increase in modulation produces spurious, out-of-band frequencies (AM splash), a source of interference for other radio users. For this reason, the depth of amplitude modulation is usually limited to 70%.

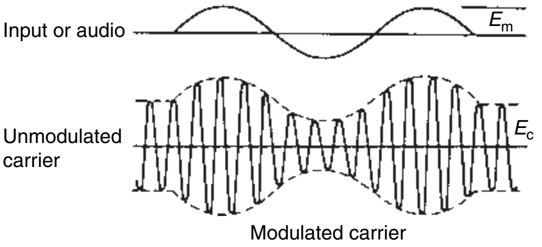

Figure 8.2 Amplitude modulation

An alternative expression for modulation depth is in terms of a modulation index from 0 to 1. The peak carrier voltage in *Figure 8.2* is E_c and the peak modulation voltage, E_m. The modulation index, m, is:

$$m = \frac{E_m}{E_c}$$

Amplitude modulation produces a band of frequencies above and below the carrier frequency – the upper and lower sidebands. The width of each sideband is equal to the highest modulating frequency so the bandwidth of an AM wave is 2 × the highest modulating frequency. To conserve spectrum, the range of modulating frequencies

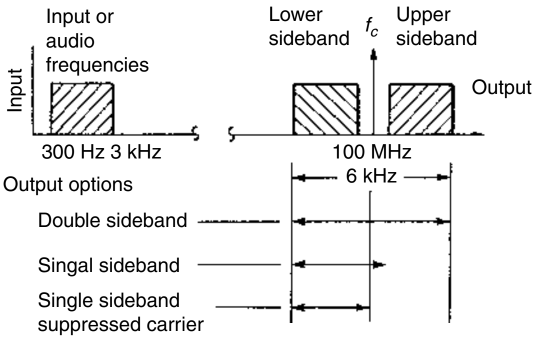
Input or audio frequencies
Input
300 Hz 3 kHz
Lower sideband
f_c
Upper sideband
Output
100 MHz
6 kHz
Output options
Double sideband
Singal sideband
Single sideband suppressed carrier

Figure 8.3 Amplitude modulation – sidebands

is restricted. For example, radio communication quality speech is limited to 300 Hz to 3000 Hz. The bandwidth occupied by a double sideband, amplitude modulated carrier for this service is 6 KHz (*Figure 8.3*).

Power relationships in an AM wave

The total power in an AM wave is the sum of powers of the carrier, the upper sideband and lower sideband:

$$P_t = \frac{E_{carr}{}^2}{R} + \frac{E_{lsb}{}^2}{R} + \frac{E_{usb}{}^2}{R}$$

where all values are RMS and R is the resistance in which the power is dissipated. From the peak voltages shown in *Figure 8.2* the power in the unmodulated carrier is:

$$P_c = \frac{E_c/\sqrt{2}\,^2}{R} = \frac{E_c{}^2}{2R}$$

The power in the sidebands

$$P_{lsb} = P_{usb} = \left(\frac{mE_c/2}{\sqrt{2}}\right)^2 \div R = \frac{m^2 E_c{}^2}{8R}$$

$$= \frac{m^2}{4}\frac{E_c{}^2}{2R}$$

The ratio of the total power in the wave to the carrier power is:

$$\frac{P_t}{P_c} = 1 + \frac{m^2}{2}$$

As m cannot exceed 1, the maximum RMS power in the wave is $P_t = 1.5P_c$; but if m reaches 1, the peak sum of E_c and E_m is $2E_c$ and so the instantaneous peak power is $2P_c$. Circuitry must be capable of handling this power level without distortion.

Double sideband amplitude modulation wastes power and spectrum. Two-thirds of the power is in the carrier which conveys no information and one sideband is discarded in the receiver. Also, the modulation must be accomplished either in the final power amplifier of the transmitter necessitating a high power modulator, or in an earlier, low power stage when all subsequent amplifiers must operate in a linear, but inefficient, mode.

8.3.2 Double sideband suppressed carrier (DSBSC)

In an amplitude modulated wave the carrier conveys no information yet contains 2/3 of the transmitted power. It is possible to remove the carrier by using a balanced modulator (*Figure 8.4*), and improve the power efficiency by this amount.

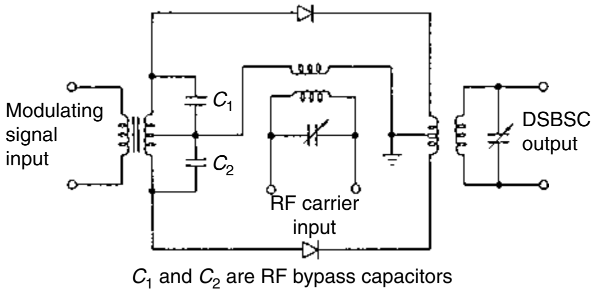

C_1 and C_2 are RF bypass capacitors

Figure 8.4 Balanced modulator

In a balanced modulator, the modulating voltage is fed in push–pull to a pair of matched diodes or amplifiers while the carrier is fed to them in parallel. The carrier components in the output cancel leaving the two sidebands. The result is a double sideband suppressed-carrier (DSBSC) wave, which is not sinusoidal, formed by the sum of the two sidebands. The carrier must be re-introduced in the receiver and its accuracy in both frequency and phase is critical.

8.3.3 Single sideband suppressed carrier (SSB or SSBSC)

The advantages of single sideband suppressed carrier transmission over double sideband AM are:

- removal of the carrier saves 2/3 of the total power
- removal of one sideband saves 50% of the remaining power
- an SSBSC transmitter only produces power when modulation is present
- the occupied bandwidth is halved; a spectrum saving
- the received signal-to-noise ratio is improved by 9 dB for a 100% modulated carrier. Halving the bandwidth accounts for 3 dB, the remainder from the improved sideband power to total power ratio. The S/N ratio improves further with lower modulation levels
- reduced susceptibility to selective fading and consequent distortion.

Two methods of generating a single sideband wave are in general use. One filters out the unwanted sideband after removal of the carrier by a balanced modulator. The other is a phase shift method (*Figure 8.5*). Here, the modulating signal is fed to two balanced modulators with a 90° phase difference. The output from both modulators contains only the sidebands but, while both upper sidebands lead the input carrier voltage by 90°, one of the lower sidebands leads it by 90° and the other lags it by 90°. When applied to the adder, the lower sidebands cancel each other while the upper sidebands add.

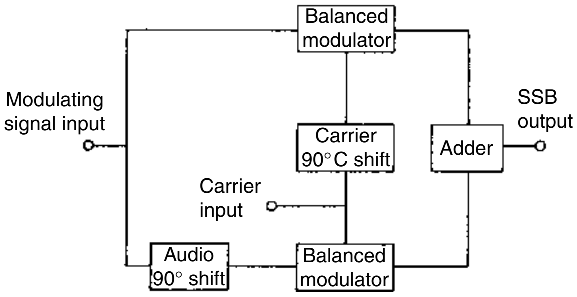

Figure 8.5 Phase shift production of single sideband

A single sideband AM wave modulated by a sinusoid consists of a constant amplitude signal whose frequency varies with the frequency of the modulating wave. Note that this is not the same as FM: the frequency in SSB does not swing to either side of the carrier. It is higher than the carrier frequency if the upper sideband is transmitted, and lower if the lower sideband is selected. The single sideband waveform is sinusoidal and, although the frequency of the reintroduced carrier must be highly accurate (±2 Hz), the phase is unimportant making a single sideband receiver less complex than one for DSBSC.

On some systems a pilot carrier is transmitted and the transmitter output power is then specified in terms of peak envelope power (pep), the power contained in a wave of amplitude equal to the pilot carrier and transmitted sideband power. Where no pilot carrier is transmitted, the power is specified as peak sideband power (psp).

8.3.4 Frequency modulation (FM)

Both frequency and phase modulation (both may be referred to as angle modulation) effectively vary the frequency of the carrier rather than its amplitude. Frequency modulation varies the carrier frequency directly but its amplitude remains constant regardless of the modulating voltage. Angle modulation is employed at VHF and above for both communications and broadcasting services.

When frequency modulated, a carrier frequency either increases or decreases when the modulation voltage is positive and varies in the opposite sense when the modulating voltage is negative (*Figure 8.6*). The amount of modulation, i.e. the 'deviation' of the carrier from its nominal frequency, is proportional to the amplitude, and not the frequency, of the modulating voltage. The modulation index, M, is defined as the deviation divided by the modulating frequency:

$$M = \frac{f_\mathrm{d}}{f_\mathrm{m}} \text{ or } \frac{\omega_\mathrm{d}}{\omega_\mathrm{m}}$$

where
f_d = deviation in hertz
f_m = modulating frequency in hertz

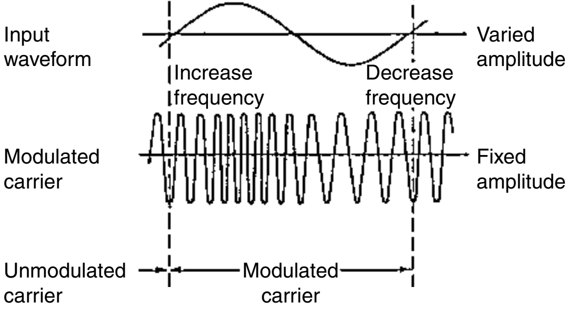

Figure 8.6 Frequency modulation

$\omega_d = 2\pi f_d$ = deviation in radians
$\omega_m = 2\pi f_m$ = modulating frequency in radians

The 'deviation ratio', D, is given by:

$$D = \frac{f_{d(max)}}{f_{m(max)}}$$

The peak deviation is the maximum amount of modulation occurring (equivalent to 100% amplitude modulation). With FM, this is limited only by the need to conserve spectrum; there is no technical limit where distortion occurs, as with 100% AM. The maximum permitted deviation for a service is determined by regulation.

The bandwidth of a frequency modulated signal is made up of the carrier and a series of sidebands, sometimes referred to as sidecurrents, spaced apart from each other at the modulating frequency. The number of sidebands is proportional to the modulation index and their amplitudes decrease with spacing from the carrier. It is generally considered satisfactory to transmit those sidebands $M + 1$ in number, with amplitudes greater than 10% of that of the carrier for that modulation index (*Figure 8.7*). The sidebands occur on both sides of the carrier and:

$$\text{1st order sidebands} = f_c \pm f_m$$

$$\text{2nd order sidebands} = f_c \pm 2f_m, \text{ etc.}$$

When the modulation index approaches 6 a good approximation of the bandwidth (δf) required for an FM transmission is $2(f_d + f_m)$ Hz. For example, speech 300–3000 Hz, max deviation 15 kHz (relevant to a VHF, 50 kHz channel spacing system):

$$\delta f = 2(f_d + f_m) = 2(15 + 3)\,\text{kHz} = 36\,\text{kHz}$$

$$M = 15/3 = 5$$

The bandwidth is also given by $\delta f = f_m \times$ highest needed sideband $\times$ 2. From Bessel functions (*Figure 8.7*), a modulation index of 5 requires the 6th order sideband to be transmitted ($M + 1$). Therefore:

$$\delta f = f_m \times 6 \times 2 = 3 \times 6 \times 2 = 36\,\text{kHz}$$

For specific values of M the carrier of an FM wave disappears. The successive disappearances and the modulation index are given in *Table 8.2*.

Table 8.2 Successive disappearances and modulation index

Order of disappearance	*Modulation index M*
1	2.40
2	5.52
3	8.65
4	11.79
5	14.93
6	18.07
$n(n > 6)$	$18.07 + \pi(n - 6)$

Systems where the modulation index exceeds $\pi/2$ are considered to be wide band FM (WBFM), those with a modulation index lower than $\pi/2$, narrow band FM (NBFM). The bandwidth of an NBFM signal is $2f_{\mathrm{m(max)}}$.

Pressure on spectrum necessitates narrower channel spacings for communications and, at VHF, 12.5 kHz is normal with ±2.5 kHz as the maximum permitted deviation. Lower standards of performance with a restricted modulation index and a highest modulating frequency of 3 kHz have had to be accepted.

For transmitters used on 12.5 kHz channel spaced systems the highest modulating frequency is, in practical terms, 2700 Hz because the specification requires the frequency response to fall above 2.55 kHz.

The modulation index on such systems, assuming a highest modulating frequency of 3 kHz, is $2.5/3 = 0.8333$ ($<\pi/2$ and so system is NBFM) and the amplitude of the 2nd order sideband is $<10\%$ of carrier amplitude (*Figure 8.7*), so the bandwidth $= 2f_{\mathrm{m(max)}} = 6$ kHz.

8.3.5 Phase modulation

The end result of phase modulation is frequency modulation, but the method of achieving it and the definition of the modulation index is different. Phase modulation is used in VHF and UHF transmitters where the carrier frequency is generated directly by a crystal oscillator. The frequency of a crystal oscillator can be varied by only a few radians but if the oscillator frequency is multiplied to produce the final carrier frequency the phase variation is also multiplied to produce a frequency deviation.

In frequency modulation deviation is proportional to the modulating voltage, but in phase modulation the frequency deviation is proportional to both the modulating voltage and frequency.

The phase modulation, ϕ_d radians, equals the modulation index so, for phase modulation, the modulation index is $\phi_d = \omega_d/\omega_m$.

Phase modulation has, therefore, a frequency response for the deviation that rises at 6 dB per octave of the modulating frequency. The flat frequency response of FM can be produced on a phase modulated transmitter by installing a filter with a response falling at 6 dB per octave in the audio amplifier, and the rising response of PM can be produced in an FM transmitter with a rising response filter.

8.3.6 Pre- and de-emphasis

Phase modulation, and FM modified to give a rising frequency response (pre-emphasis), offer an improved signal-to-noise ratio in the receiver. The +6 dB per octave response produced in the transmitter is restored to a flat response in the receiver by a −6 dB per octave filter in the audio circuitry which reduces both the enhanced levels of the higher speech frequencies and the high frequency noise (de-emphasis).

8.3.7 Merits of amplitude and frequency modulation

The advantages and disadvantages of AM and FM are given in *Table 8.3*.

Table 8.3 Advantages and disadvantages of AM and FM

	Advantages	*Disadvantages*
AM	Simple modulators and de-modulators	Susceptible to man-made noise
	Narrower bandwidth than wide-band FM	Audio strength falls with decreasing RF signal strength Inefficient power usage Limited dynamic range Transmitter output power not easily adjusted
FM	Less susceptible to noise Constant audio level to almost the end of radio range	Wider bandwidth
	Capture effect in receiver	Capture effect may be undesirable, e.g. aviation communications
	More power-efficient Transmitter output power easily adjustable	

8.3.8 Stereo FM radio

An important advantage of FM over AM radio broadcasting is the availability of stereo sound. This requires the transmission and reception of the left and right audio signals, to produce a stereo image. However, the introduction of stereo had to be achieved while still catering for large numbers of listeners who wanted to continue receiving in mono. The introduction of stereo FM radio had to go unnoticed by anyone using old mono radios. Without this condition, a simpler method could have been chosen.

The standard FM stereo system used by broadcasters around the world is illustrated in the block diagram of *Figure 8.8*. It uses frequency division multiplexing (FDM) to combine the two signals of the left and right channels. The signals are filtered to limit the bandwidth to 15 kHz. The left (L) and right (R) signals are then added to produce a sum signal and subtracted one from the other to produce a difference signal.

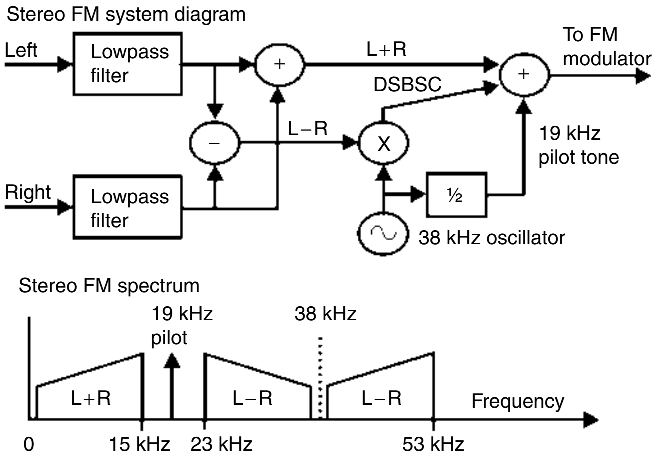
Stereo FM system diagram
Left
Lowpass filter
+
L+R
To FM modulator
DSBSC
+
−
L−R
X
19 kHz pilot tone
Right
Lowpass filter
½
38 kHz oscillator
Stereo FM spectrum
19 kHz pilot
38 kHz
L+R
L−R
L−R
Frequency
0
15 kHz
23 kHz
53 kHz

Figure 8.8 FM stereo transmitter

The sum signal provides a monophonic signal, which provides a baseband signal for the frequency modulator. This was the technique used in mono FM and thus was the obvious choice for stereo FM, to allow backward compatibility. An mono FM radio can receive this signal and recover the combined L and R channels, thereby satisfying the requirement for providing unchanged service to mono radios.

The difference signal is used to amplitude modulate a 38 kHz sinewave. By utilizing a balanced mixer, double sideband suppressed carrier (DSBSC) is generated. However, the modulation method must take into account the ease of demodulation. In particular, demodulating a DSBSC signal can be difficult. Both frequency and phase of the carrier are needed to perform faithful demodulation.

In the stereo system the DSBSC demodulation problem is dealt with by including a 19 kHz pilot tone in the broadcast. This tone is generated by a divide-by-two frequency converter circuit, which takes the 38 kHz carrier and produces the 19 kHz pilot tone. The 19 kHz pilot tone falls midway in the spectral region between the mono sum signal (up to 15 kHz) and below the DSBSC difference signal information. The DSBSC signal extends from 23 kHz to 53 kHz, since the input modulating signals are band limited to 15 kHz. The DSBSC output is added to the baseband (L and R sum) signal and the 19 kHz pilot tone before being sent to the FM modulator.

A mono FM receiver ignores the stereo information by using a filter after its FM demodulator to block everything above 15 kHz. It passes the combined L and R channel signal, which is monophonic.

A stereo receiver has an additional circuit after the FM demodulator to detect and demodulate the DSBSC signal. The stereo receiver detects a 19 kHz pilot tone and uses this to generate a 38 kHz signal. This is then used to demodulate the DSBSC signal that carries the L and R channel difference information. The stereo receiver then has both the sum and difference signals, which is all that is needed to recreate the separate left and right signals. Separation is achieved by adding and subtracting sum and difference signals.

The noise power spectral density of a demodulated FM signal tends to increase with the square of the modulation frequency. This is why pre-emphasis is used to boost the high frequency baseband signals for maintaining the signal-to-noise ratio of the transmitted signal. However, this means that there will be more noise in the 23 kHz to 53 kHz band used for the difference signal than for the 0–15 kHz band used for the sum signal. Consequently a significantly higher input signal level is required to receive a stereo transmission compared with a mono signal for the same output signal-to-noise ratio. Thus stereo reception requires far higher radio signal levels than for mono reception and is more susceptible to interference from other radio sources.

8.4 Digital modulation

8.4.1 Data processing

Filtering

A data pulse with a sharp rise- and fall-time produces harmonic frequencies and requires a wide bandwidth if it is to maintain its shape during transmission. Consequently, to transmit the data over a limited bandwidth the pulses must be shaped to reduce the harmonic content as much as possible without impairing the intelligibility of the signal. This is accomplished by the use of low-pass (Gaussian) filters of which the result is a string of smoother pulses, often referred to as 'tamed' (*Figure 8.9(b)*). Tamed FM permits high data rates within a limited channel bandwidth whilst maintaining acceptable adjacent channel interference levels.

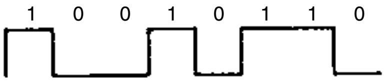

(a) Unprocessed 'raw' data

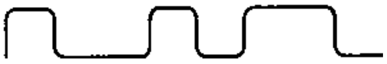

(b) Filtered 'tamed' data

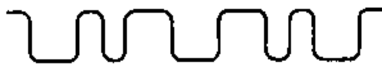

(c) Manchester encoded, 'tamed' data

Figure 8.9 Data processing

Manchester encoding

A serious problem with the transmission of binary data is that unless the clocks in the transmitter and receiver are synchronous the digits become confused, particularly where a continuous string of 1s or 0s occurs. Manchester encoding makes 1s change state from 1 to 0, and 0s from 0 to 1 during each digit period (*Figure 8.9(c)*), facilitating the synchronization of the clocks and rendering the digits more easily recognizable.

Gray coding

For binary data the number of signalling levels is not restricted to two (mark and space). Multi-level signalling has the advantage that one signalling element carries the information for more than one information bit, thus reducing the bandwidth requirement.

If the number of levels is increased to m, where $m = 2^n$ (i.e. $n = \log_2 m$), the m-level data symbol is represented by n binary digits of 0 or 1 (see quaternary phase shift keying, Section 8.4.5). For example, in a quaternary data signal, $m = 4 = 2^2$ giving the binary sequences 00, 01, 10, 11 for each signalling level. This process is Gray coding where there is only one digit difference for each transition between adjacent levels. The number of levels is not restricted to four and 16 level quadrature amplitude modulation (QAM), where $n = 4$, is an efficient system.

In a Gray coded signal where each element contains n bits ($n = \log_2 m$) and the signalling element rate is B bauds (elements per second), the transmission rate is $B \log_2 m$ bits per second.

8.4.2 On/off and amplitude shift keying

On/off keying

The earliest modulation method. A continuous radio frequency wave (CW) is interrupted in a recognizable pattern (Morse code). To provide audibility the carrier is heterodyned with a beat frequency oscillator (BFO) in the receiver. The use of a modulated continuous wave (MCW) eliminates the need for a BFO but the bandwidth of the signal is increased. The problem with on/off keying is the lack of a reference level. If the signal strength temporarily falls below the sensitivity threshold of the receiver it appears to the operator as a series of spaces.

Binary amplitude shift keying (ASK or BASK)

This shifts the level of an audio frequency subcarrier which then modulates a radio frequency carrier.

Because the level of a subcarrier is changed, AM sidebands are produced. Also, because the keyed waveform is non-sinusoidal harmonics occur. The occupied sub-carrier bandwidth for ASK is:

$$\text{Bandwidth} = 2B$$

where B = bit repetition rate (bits/second).

When the RF carrier is modulated its bandwidth is $2(f_c + B)$ where f_c = subcarrier frequency.

8.4.3 Frequency shift keying (FSK)

Although used for conveying digital information, frequency shift keying in reality employs frequency modulation. In its original form, developed for HF transmission, FSK changes the carrier frequency to indicate a 1 or a 0 but retains the nominal carrier frequency as a reference and to represent a mark. A downwards shift of carrier frequency by 170 Hz represents a space in the HF radio system.

Table 8.4 Standard ITU-T frequencies for various bit rates

Bit rate (bps)	*Frequencies*		*Subcarrier frequency*
	0	*1*	
600	1700	1300	1500
1200	2100	1300	1700
up to 300	1180	980	1080
	1850	1650	1750

Modern FSK uses two different modulation frequencies to represent 1s and 0s. If intersymbol interference (ISI) is to be avoided the separation of the tones must be more than half the bit rate, and a factor of 0.7 is often used. Standard ITU-T frequencies for various bit rates are given in *Table 8.4*.

The base bandwidth requirement is:

$$\text{Bandwidth} = f_2 - f_1 + 2B$$

The bandwidth of a modulated carrier is:

$$\text{RF bandwidth} = 2(f_2 + B)(\text{narrow band FM})$$

Minimum shift keying (MSK) is a form of FSK where the frequency deviation is equal to half the bit rate.

Gaussian minimum phase shift keying (GMSK)

Similar to MSK, the Gaussian filters improve the adjacent channel performance against a small cost (approximately 1%) in ISI while achieving high data transmission rates.

8.4.4 Fast frequency shift keying (FFSK)

Fast frequency shift keying may either amplitude or frequency modulate the carrier. In binary FFSK, the data is changed in a modem to tones of 1800 Hz to represent binary 0 and 1200 Hz to represent binary 1. During transmission a binary 1 consists of 1 cycle of 1200 Hz, f_1, and a 0, $1\frac{1}{2}$ cycles of 1800 Hz, f_2, i.e. a bit rate of 1200 bps. For acceptable intersymbol interference the distance between the tones cannot be less than half the bit rate and the 600 Hz separation in FFSK represents the fastest signalling speed – hence the description – and minimum bandwidth. For this reason it is sometimes called minimum frequency shift keying (MFSK). The base bandwidth is the same as for FSK, but the RF bandwidth depends upon the system deviation. For example, on a 12.5 kHz channel spaced system carrying FFSK and complying with Radiocommunications Agency Code of Practice MPT 1317, deviation = 60% of system deviation = 1.5 kHz:

$$f_1 = 1200\,\text{Hz}$$

$$f_2 = 1800\,\text{Hz}$$

$$B = 1200\,\text{bps}$$

$f_{\max} = f_2 + B = 3\,\text{kHz}$ and the mod. index $m = \text{max. dev.}/f_{\max} < \pi/2$ so the system is NBFM and the bandwidth $= 2(f_2 + B) = 6\,\text{kHz}$.

Where more data states than binary are to be transmitted, multistate FSK (M-ary FSK) is also possible where M may be up to 32 states.

Both FFSK and M-ary FSK are well suited to radio transmission as the change of state occurs while the signals are passing through zero, avoiding sudden phase changes (*Figure 8.10(a)*).

The minimum distance between the tones used in FFSK of 0.5 times the bit rate is not ideal for immunity to intersymbol interference (ISI).

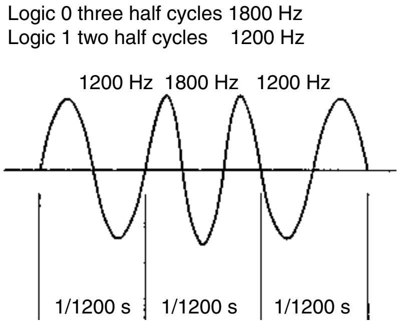
Logic 0 three half cycles 1800 Hz
Logic 1 two half cycles 1200 Hz
1200 Hz
1800 Hz
1200 Hz
1/1200 s
1/1200 s
1/1200 s

(a) FFSK modulation

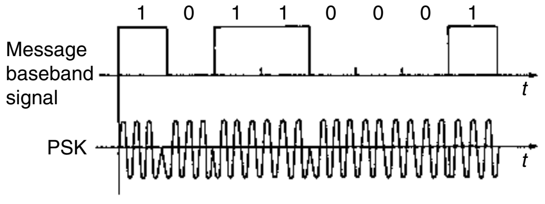
1
0
1
1
0
0
0
1
Message baseband signal
t
PSK
t

(b) PSK modulation

Figure 8.10 FFSK and PSK

A separation of 0.7 times the bit rate would improve the ISI but increase the bandwidth.

8.4.5 Phase shift keying

There are several variants of phase shift keying. Binary phase shift keying (BPSK or PSK) changes the phase of the carrier by 180° at the zero crossing point (*Figure 8.10(b)*). No carrier frequency is present with PSK as half the time the carrier is multiplied by $+1$ and the other half by -1 and cancels out, but the reference phase of the carrier must be re-inserted at the receiver. The bandwidths occupied are the same as for ASK, i.e.:

$$\text{Baseband} = 2B$$

$$\text{RF bandwidth (AM or NBFM)} = 2(f_c + B)$$

Differential phase shift keying (DPSK) advances the phase 90° or 270° at each change of logic state (*Figure 8.11*). Changing phase only at a

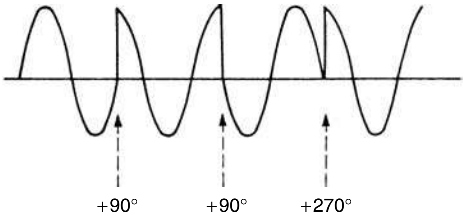

Data	Phase change
0	+90°
1	+270°

(a) DPSK modulation

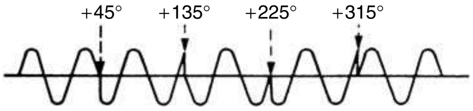

Dibit	Phase change
00	+45°
01	+135°
11	+225°
10	+315°

(b) QPSK modulation

Figure 8.11 DPSK and QPSK

change of logic state saves bandwidth which, for DPSK, is equal to the bit rate.

An important advantage of both FFSK and PSK over FSK is that because the moment of change is predefined it is possible to recover data more accurately. However, the transition between signalling states is not smooth requiring large and rapid phase shifts. Multilevel systems with less phase shift between elements are preferable.

Quaternary, or quadrature, phase shift keying (QPSK) is a four-level Gray-coded signalling method with 90° phase shift between adjacent signalling elements (*Figure 8.11(b)*). If the signal is considered as a vector the points at $\pi/4$ (45°), $3\pi/4$ (135°), $5\pi/4$ (225°) and $7\pi/4$ (315°) represent the transition points between states and the binary data.

8.5 Spread spectrum transmission

The spread spectrum technique spreads the carrier containing the information over a very wide bandwidth, typically 1.25 MHz, using pseudo-noise generation techniques as described in Chapter 12. The transmitter uses what is in effect a digital key to spread the bandwidth and the receiver is equipped with an identical key for despreading. A number of users with different keys can occupy the same band at the same time. The system operates well in poor signal-to-noise or high interference environments.

A continuous wave (CW) transmission concentrates all the radiated energy on a single frequency (*Figure 8.12*). Amplitude modulation and narrow band FM widen the radiated bandwidth, reducing the energy at the carrier frequency and per kHz. Wide band FM carries the process a stage further until with spread spectrum the band width is increased to the extent that the signal almost disappears into the noise floor.

Spreading of the bandwidth is achieved by multiplying the digitally modulated signal by a spreading code at a much higher bit rate (100–1000 times the signal bit rate). This is done by combining the signal with the output of a random code generator running at 2 or 3 orders of magnitude faster than the binary signal rate. *Figure 8.13* is a block diagram of a spread spectrum system. A clock running at the spreading rate R_c is used to drive both the spreading generator and, after frequency division, the data encoder. The carrier is first BPSK modulated by the encoded data and then in a balanced modulator (the spreading correlator) by the high rate code from the spreading generator. The resultant transmitted bits are referred to as chips to distinguish them from data bits. In the receiver, the clock pulses R_c are recovered and used to drive both the despreading generator and decoder.

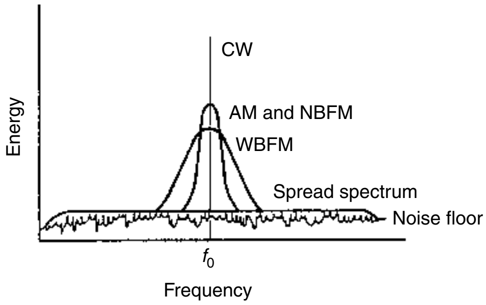

Figure 8.12 Comparative energy dispersal

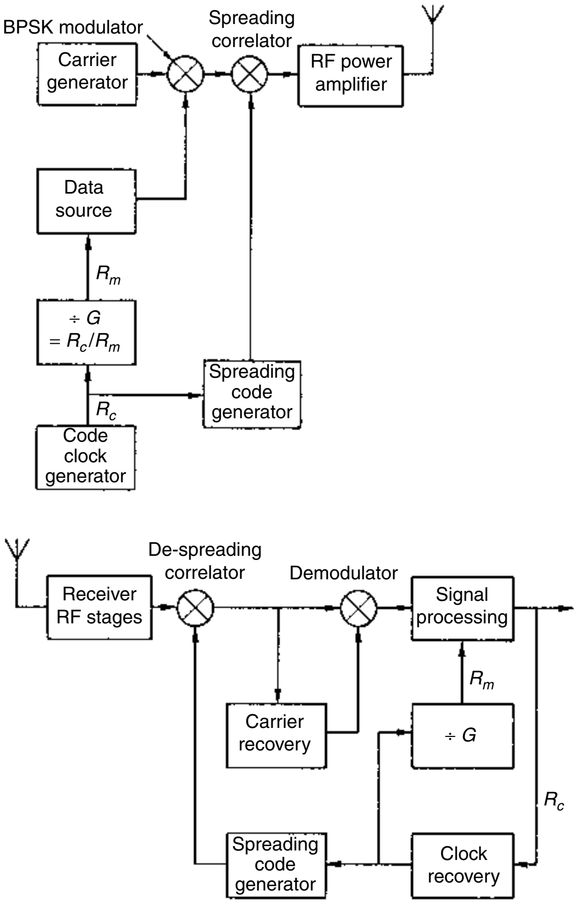
BPSK modulator
Spreading correlator
Carrier generator
RF power amplifier
Data source
R_m
÷ G = R_c/R_m
Spreading code generator
R_c
Code clock generator
De-spreading correlator
Demodulator
Receiver RF stages
Signal processing
R_m
Carrier recovery
÷ G
R_c
Spreading code generator
Clock recovery

Figure 8.13 CDMA (spread spectrum) system

Many users can be accommodated by allocating each a unique spreading code. It is common to use a pseudo-noise (PN) generator to multiply the bit rate and then to modulate the carrier with either FSK or PSK.

Although spread spectrum is a digital system, in quality of signal there are similarities with analogue:

Analogue	*Digital*
Signal/noise ratio	Energy per bit, E_b/N_o
Intelligibility, signal/ noise + distortion	Bit error rate

References

Belcher, R. *et al.* (1989). *Newnes Mobile Radio Servicing Handbook*. Butterworth-Heinemann, Oxford.

Kennedy, G. (1977). *Electronic Communications Systems*, McGraw-Hill Kogashuka, Tokyo.

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

9 Frequency planning

9.1 International and regional planning

The International Telecommunications Union (ITU) administers the planning and regulation of the radio frequency spectrum on a worldwide basis through the World Administrative Radio Conferences.

For planning purposes the world is divided into three regions as shown in *Figure 9.1*. The boundaries are formed by geographical features suited to the purpose such as seas, high mountain ranges or uninhabited remote areas.

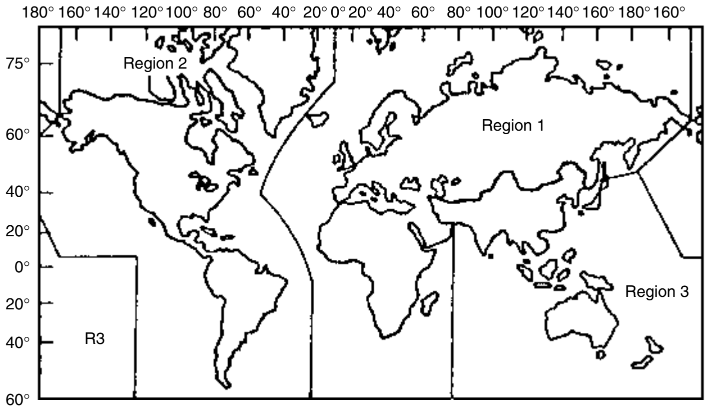

Figure 9.1 ITU defined regions. For purposes of international allocations of frequencies the world has been divided into three regions

The World Administrative Radio Conferences decide the use of blocks of the spectrum, e.g. sound broadcasting, television, marine communications, and the countries permitted to use the blocks for those allotted purposes.

9.2 National planning

Having been allocated blocks of frequencies for a particular type of use, the administration of each country determines the method of allocating frequency bands from each block to the user categories within their country. Every country has a radio regulatory department within its administration. In the UK this is the Radiocommunications Agency of the Department of Trade and Industry.

9.2.1 The role of the Radiocommunications Agency

The aims of the Radiocommunications Agency in the UK, and, in general, the radio regulatory bodies of other countries are:

1. To ensure that the radio frequency spectrum is used in ways which maximize its contribution to national social and economic welfare, having regard to safety of life factors.
2. To make the maximum amount of spectrum available for commercial use.
3. To provide an expert service to government as a whole in the field of radio regulation.

The first stage of national planning is the assignment of the radio frequencies (channels) within a geographic area. Both the allocation of the blocks to users and the geographic assignment of channels if not wisely carried out can result in spectrum pollution – intermodulation products are one source (see Chapter 19) – and unacceptable interference between services and users. The incorrect allocation of blocks may not only cause interference in the home country but, even at VHF and UHF, between adjacent countries. Incorrect assignment of channels causes a more local problem but, depending on the frequencies involved, the undesirable effects can spread over a wide area.

The second stage of national planning is the assignment of discrete channels for use on multi-user sites where the selection of incompatible frequencies causes interference, receiver de-sensitization and, possibly, blocking, and intermodulation products.

Not all channels are allocated directly by the regulatory body. Blocks of channels, usually comprised of two frequencies, may be issued to responsible user organizations: the Home Office for the police and fire services is an example. These organizations then become responsible for the frequency planning and allocation within their user group.

Additional to the allocation of frequencies, the Radiocommunications Agency through the licensing procedure regulates the use of base station sites, the maximum transmitter power, and antenna directivity for each service. It also prepares and publishes technical specifications with which all equipment must comply. A list of the current specifications, of which single copies can be obtained, is available from

The Information and Library Service, Radiocommunications Agency, Wyndham House, 189 Marsh Wall, London E14 9SX.

The Radio Investigation Service (RIS) is the branch of the Agency which, in addition to investigating interference, inspects all radio installations prior to commissioning and, if in order, issues an inspection certificate. No station is permitted to operate without a certificate and may not be modified subsequent to the issue of the certificate.

Current policy throughout Europe is leading towards the de-regulation of radio communications while safeguarding the protection from interference. Allocation of frequencies by pricing is also under consideration on the basis that a scarce resource, the spectrum, will be allocated to the users having the greatest need.

9.3 Designations of radio emissions

Radio emissions should be expressed in a three-symbol code form, which defines the exact nature of carrier, signal and transmitted information. The first symbol defines the carrier, the second symbol defines the signal, and the third symbol defines the information.

First symbol

A	Double-sideband amplitude-modulated
B	Independent sideband amplitude-modulated
C	Vestigial sideband amplitude-modulated
D	Amplitude- and angle-modulated simultaneously, or in a predefined sequence
F	Frequency modulated
G	Phase modulated
H	Single-sideband, full carrier
J	Single-sideband, suppressed carrier
K	Amplitude-modulated pulse sequence
L	Width-modulated pulse sequence
M	Position phase modulated pulse sequence
N	Unmodulated carrier
P	Unmodulated pulse sequence
Q	Pulse sequence in which carrier is angle-modulated during the pulse period
R	Single-sideband, reduced or variable level carrier
V	Pulse sequence with a combination of carrier modulations, or produced by other means
W	Carrier is modulated by two or more of angle, amplitude, and pulse modes, simultaneously or in a defined sequence
X	Other cases

Second symbol

0	No modulating signal
1	Digital signal without modulating sub-carrier
2	Digital signal with modulating sub-carrier
3	Analogue signal
7	Two or more channels with digital signals
8	Two or more channels with analogue signals
9	Composite system with one or more channels of digital signals and one or more channels of analogue signals
X	Other cases

Third symbol

A	Aural telegraph
B	Automatic telegraph
C	Facsimile
D	Data
E	Telephony (and sound broadcasting)
F	Television
N	No information transmitted
W	Combination of any of the above
X	Other cases

9.4 Bandwidth and frequency designations

A four symbol code should be used to express bandwidth and frequency to three significant figures. A letter to denote the unit of frequency is placed in the position of the decimal point, where the letters and bandwidths are:

Letter	Bandwidth
H	Below 1000 Hz
K	Between 1 and 999 kHz
M	Between 1 and 999 MHz
G	Between 1 and 999 GHz

So, a frequency of 120 Hz is 120H, while a frequency of 12 Hz is 12H0 etc.

9.5 General frequency allocations

VLF, LF, MF (frequency in kHz)

10.0	140.5	Fixed; maritime; navigation
140.5	283.5	Broadcast

255.0	526.5	Radio navigation; fixed
526.0	1606.5	Broadcast
1606.5	1800.0	Maritime and land mobile; fixed
1810.0	1850.0	Amateur (shared in UK)
1850	2000	Amateur
1850	2045	Fixed; mobile
2045	2173.5	Maritime mobile; fixed
2160	2170	Radiolocation
2173.5	2190.5	Mobile
2190.5	2194	Maritime
2194	2625	Fixed; mobile
2300	2498	Broadcast
2625	2650	Maritime mobile
2650	2850	Fixed; mobile
2850	3155	Aero mobile

HF (frequency in kHz)

3155	3400	Fixed; mobile
3200	3400	Broadcast
3400	3500	Aero mobile
3500	3800	Amateur, fixed; mobile
3800	4000	Amateur (region 2 only)
3800	3900	Fixed; mobile
3800	3950	Aero mobile
3950	4000	Fixed; broadcast
4000	4063	Fixed; maritime mobile
4063	4438	Maritime mobile
4438	4650	Fixed; mobile
4650	4750	Aero mobile
4750	5060	Fixed; mobile; broadcast
5060	5480	Fixed; mobile
5450	5730	Aero mobile
5730	5950	Fixed; mobile
5950	6200	Broadcast
6200	6525	Maritime mobile
6525	6765	Aero mobile
6765	7000	Fixed; mobile
7000	7100	Amateur
7100	7300	Amateur (region 2 only)
7100	7300	Broadcast (regions 1 and 3)
7300	8195	Fixed
8100	8815	Maritime mobile
8815	9040	Aero mobile

9040	9500	Fixed
9500	10 000	Broadcast
10 000	10 100	Aero mobile
10 100	11 175	Fixed
10 100	10 150	Amateur
11 175	11 400	Aero mobile
11 400	11 650	Fixed
11 650	12 050	Broadcast
12 050	12 230	Fixed
12 230	13 200	Maritime mobile
13 200	13 360	Aero mobile
13 360	13 600	Fixed
13 600	13 800	Broadcast
13 800	14 000	Fixed
14 000	14 350	Amateur
14 350	15 000	Fixed
15 000	15 100	Aero mobile
15 100	15 600	Broadcast
15 600	16 360	Fixed
16 360	17 410	Maritime mobile
17 410	17 550	Fixed
17 550	17 900	Broadcast
17 900	18 030	Aero mobile
18 030	18 068	Fixed
18 068	18 168	Amateur
18 168	18 780	Fixed
18 780	18 900	Maritime mobile
18 900	19 680	Fixed
19 680	19 800	Maritime mobile
19 800	21 000	Fixed
21 000	21 450	Amateur
21 450	21 850	Broadcast
21 850	21 870	Fixed
21 870	22 000	Aero mobile
22 000	22 855	Maritime mobile
22 855	23 200	Fixed; mobile
23 200	23 350	Aero mobile
23 350	24 890	Fixed; mobile
24 890	24 990	Amateur
25 010	25 070	Fixed; mobile
25 070	25 210	Maritime mobile
21 210	25 550	Fixed; mobile

25 550	25 670	Radio astronomy
25 670	26 100	Broadcast
26 100	26 175	Maritime mobile
26 175	28 000	Fixed; mobile
28 000	29 700	Amateur
29 700	30 000	Fixed; mobile

VHF, UHF (frequencies in MHz)

30.0	50.0	Fixed; mobile
47.0	68.0	Broadcast (TV)
50.0	52.0	Amateur (UK)
50.0	54.0	Amateur (regions 2 and 3)
68.0	74.8	Fixed; mobile
70.0	70.5	Amateur (UK)
74.8	75.2	Aero navigation
75.2	87.5	Fixed; mobile
87.5	108	Broadcast (FM)
108	118	Aero navigation
118	137	Aero mobile
137	138	Spacecraft; satellites
138	144	Aero mobile; space research
144	146	Amateur
146	148	Amateur (regions 2 and 3 only)
146	174	Fixed; mobile
156	174	Maritime mobile
174	230	Broadcast (TV)
220	225	Amateur (USA)
230	328.6	Fixed; mobile
328.6	335.4	Aero navigation
335.4	400	Fixed; mobile
400	410	Space research; meteorology
410	430	Fixed; mobile
430	440	Amateur; radiolocation
440	470	Fixed; mobile
470	855	Broadcast (TV)
855	1300	Fixed; mobile
902	928	Amateur (USA)
934	935	Citizens band (UK)
1240	1325	Amateur
1300	1350	Aero navigation
1350	1400	Fixed; mobile
1400	1429	Space (uplink); fixed
1429	1525	Fixed; mobile

1525	1600	Space (downlink)
1600	1670	Space (uplink)
1670	1710	Space (downlink)
1710	2290	Fixed; mobile
2290	2300	Space (downlink); fixed
2300	2450	Amateur; fixed
2310	2450	Amateur (UK)
2300	2500	Fixed; mobile
2500	2700	Fixed; space (downlink)
2700	3300	Radar
3300	3400	Radiolocation; amateur
3400	3600	Fixed; space (uplink)
3600	4200	Fixed; space (downlink)
4200	4400	Aero navigation
4400	4500	Fixed; mobile
4500	4800	Fixed; space (downlink)
4800	5000	Fixed; mobile
5000	5850	Radio navigation; radar
5650	5850	Amateur
5850	7250	Fixed; space (uplink)
7250	7900	Fixed; space (downlink)
7900	8500	Fixed; mobile; space
8500	10 500	Radar; navigation
10 000	10 500	Amateur
10 700	12 700	Space (downlink); fixed
12 700	15 400	Space (uplink); fixed
17 700	20 000	Space (up/down); fixed
24 000	24 250	Amateur

9.6 Classes of radio stations

AL	Aeronautical radionavigation land station
AM	Aeronautical radionavigation mobile station
AT	Amateur station
AX	Aeronautical fixed station
BC	Broadcasting station, sound
BT	Broadcasting station, television
CA	Cargo ship
CO	Station open to official correspondence exclusively
CP	Station open to public correspondence
CR	Station open to limited public correspondence
CV	Station open exclusively to correspondence of a private agency

DR	Directive antenna provided with a reflector
EA	Space station in the amateur-satellite service
EB	Space station in the broadcasting-satellite service (sound broadcasting)
EC	Space station in the fixed-satellite service
ED	Space telecommand space station
EE	Space station in the standard frequency-satellite service
EF	Space station in the radiodetermination-satellite service
EG	Space station in the maritime mobile-satellite service
EH	Space research space station
EJ	Space station in the aeronautical mobile-satellite service
EK	Space tracking space station
EM	Meteorological-satellite space station
EN	Radionavigation-satellite space station
EO	Space station in the aeronautical radionavigational-satellite service
EQ	Space station in the maritime radionavigation-satellite service
ER	Space telemetering space station
ES	Station in the intersatellite service
EU	Space station in the land mobile-satellite service
EV	Space station in the broadcasting-satellite service (television)
EW	Space station in the earth exploration-satellite service
EX	Experimental station
EY	Space station in the time signal-satellite service
FA	Aeronautical station
FB	Base station
FC	Coast station
FL	Land station
FP	Port station
FR	Receiving station only, connected with the general network of telecommunication channels
FS	Land station established solely for the safety of life
FX	Fixed station
GS	Station on board a warship or a military or naval aircraft
LR	Radiolocation land station
MA	Aircraft station
ME	Space station
ML	Land mobile station
MO	Mobile station
MR	Radiolocation mobile station

MS	Ship station
ND	Non-directional antenna
NL	Maritime radionavigation land station
OD	Oceanographic data station
OE	Oceanographic data interrogating station
OT	Station open exclusively to operational traffic of the service concerned
PA	Passenger ship
RA	Radio astronomy station
RC	Non-directional radio beacon
RD	Directional radio beacon
RG	Radio direction-finding station
RM	Maritime radionavigation mobile station
RT	Revolving radio beacon
SM	Meteorological aids station
SS	Standard frequency and time signal station
TA	Space operation earth station in the amateur-satellite service
TB	Fixed earth station in the aeronautical mobile-satellite service
TC	Earth station in the fixed-satellite service
TD	Space telecommand earth station
TE	Transmitting earth station
TF	Fixed earth station in the radiodetermination-satellite service
TG	Mobile earth station in the maritime mobile-satellite service
TH	Earth station in the space research service
TI	Earth station in the maritime mobile-satellite service at a specified fixed point
TJ	Mobile earth station in the aeronautical mobile-satellite service
TK	Space tracking earth station
TL	Mobile earth station in the radiodetermination-satellite service
TM	Earth station in the meteorological-satellite service
TN	Earth station in the radionavigation-satellite service
TO	Mobile earth station in the aeronautical radionavigation-satellite service
TP	Receiving earth station
TQ	Mobile earth station in the maritime radionavigation-satellite service
TR	Space telemetering earth station
TS	Television, sound channel
TT	Earth station in the space operation service

TU	Mobile earth station in the land mobile-satellite service
TV	Television, vision channel
TW	Earth station in the earth exploration-satellite service
TX	Fixed earth station in the maritime radionavigation-satellite service
TY	Fixed earth station in the land mobile-satellite service
TZ	Fixed earth station in the aeronautical radionavigation-satellite service

9.7 Radio wavebands

Frequency band	*Frequency*	*Wavelength*	*Waveband definition*
VLF	3 to 30 kHz	1 00 000 to 10 000 m	myriametric
LF	30 to 300 kHz	10 000 to 1000 m	kilometric
MF	300 to 3000 kHz	1000 to 100 m	hectometric
HF	3 to 30 MHz	100 to 10 m	decametric
VHF	30 to 300 MHz	10 to 1 m	metric
UHF	300 to 3000 MHz	1 to 0.1 m	decimetric
SHF	3 to 30 GHz	10 to 1 cm	centimetric
EHF	30 to 300 GHz	1 to 0.1 cm	millimetric
EHF	300 to 3000 GHz	0.1 to 0.01 cm	decimillimetric

Reference

Pannell, W.M. (1979). *Frequency Engineering in Mobile Radio Bands.* Granta Technical Editions, Cambridge.

10 Radio equipment

10.1 Transmitters

10.1.1 Transmitter functions

The functions of all transmitters and the terminology used to describe them, irrespective of the modulation method, are:

1. To generate the radio frequency carrier and amplify it to an appropriate power level; the RF power output.
2. To modulate the carrier with the intelligence to the pre-determined level; the modulation depth for AM, the deviation for FM or PM. The process must introduce the minimum noise and distortion, and prevent the modulation from exceeding the permitted level.
3. Radiate the minimum signals at frequencies outside the permitted bandwidth. Out-of-band or spurious radiation is strictly controlled by the Radiocommunications Agency MPT specifications.

10.1.2 Amplitude-modulated transmitters

Figure 10.1 is a block diagram of an amplitude modulated transmitter; in this case a quartz crystal oscillator generates the carrier frequency, although a frequency synthesizer could equally well be used. The carrier frequency in an AM transmitter is usually generated either at the transmitted frequency or one of its subharmonics; 2nd, 3rd or 6th subharmonic frequencies are typical choices depending on the final frequency.

The output of the oscillator is amplified to the level of the specified power output and if the oscillator runs at a subharmonic a frequency multiplier stage will be included, as in *Figure 10.1*, in the amplifier chain before the final stage, the power amplifier (PA). A tuned filter in the aerial circuit removes from the output unwanted frequencies which might cause interference with other users. A matching circuit correctly matches the impedance of the filter circuit to that of the aerial to ensure maximum power transfer.

In the audio circuits, the speech input from the microphone is processed by controlling the range of frequencies it contains and limiting its amplitude. This eliminates the risk of over-modulation and the production of out-of-band frequencies. Over-modulation produces out-of-band frequencies in any transmitter, but with AM once 100% modulation is exceeded frequencies are produced across a wide range of the spectrum: a disastrous situation for other users. After processing, the audio is amplified in an AM transmitter to a high level and applied to the RF power amplifier to vary its output in the form illustrated in *Figure 8.2*.

AM transmitters are power inefficient. The RF PA cannot operate in class C. It must be linear so as not to distort the speech, and the audio, when it is the power amplifier which is modulated, must be amplified to a high power. If the modulation is applied at a lower power level (in an early stage of the amplifier chain) all the subsequent amplifiers must operate in a linear, but inefficient, manner.

The power output of an AM transmitter is normally specified in terms of the RMS value of the carrier power but the average and peak powers will depend on the depth of modulation, 100% modulation producing a peak power of twice the carrier peak power (see Chapter 8).

10.1.3 Angle modulated transmitters

Figure 10.2 is a block diagram of a frequency modulated transmitter using a frequency synthesizer for carrier generation. The frequency is generated at the final frequency and RF amplifiers raise the power level to that specified for the transmitter output. When a crystal oscillator is used for carrier generation, it must operate at a very low frequency because a quartz crystal oscillator can be frequency modulated by a few radians only. Several stages of the RF amplifier chain then operate as frequency multipliers. Similar filter and aerial matching circuits to those of an AM transmitter are necessary in the output arrangements of the transmitter.

The audio processing circuitry is similar to that for AM transmitters but the modulation (deviation) is applied to a stage operating at a very low RF power level (directly to the VCO in a synthesizer-equipped transmitter and immediately following the crystal oscillator in a direct crystal controlled one); high power audio is not necessary. An additional simple circuit in an FM transmitter may be included to enhance the higher audio frequencies at a rate of 6 dB/octave. This is pre-emphasis and has the merit of improving the level of speech to noise at the receiver (signal-to-noise ratio). The audio level must still be limited because, while the effect of over-deviation is not as disastrous as over-modulation in an AM transmitter, increasing deviation produces a steadily increasing range of frequencies, known as side-currents, outside the permitted bandwidth.

Angle modulated transmitters are more power efficient than amplitude modulated because the modulation is applied at a low power level and also, as no audio frequencies are directly present in the RF amplifier and PA stages, these can operate efficiently in class C.

Phase modulated (PM) transmitters, because the phase shift with modulation is very small, generate the carrier at a very low frequency and use direct crystal control. The frequency is then multiplied many times, thirty-two to thirty-six is common, up to the final frequency. After multiplication, the phase modulation which was originally a few radians has effectively become frequency modulation. The phase modulation process produces pre-emphasis inherently. As far as the user is concerned there is no practical difference between phase modulation and frequency modulation with added pre-emphasis.

10.1.4 Transmitter specifications

In the UK, the Radiocommunications Agency issues specifications with which all equipment must comply. Other countries have their own regulatory bodies, such as the FCC in the USA. These specifications are concerned principally with the prevention of interference and obtaining the maximum use of the frequency spectrum. The characteristics defined in the MPT specifications and other features which affect the user, apart from the physical dimensions, are:

- *Supply voltage.*
- *Operational frequency band.*
- *Modulation method.*
- *Channel separation.*
- *RF power output and impedance.* Output ranges from about 0.5 W to 5 W for hand-portables and 5 W to 25 W for mobiles. The maximum power permitted on a system will be specified in the Licence. Output impedance is commonly 50 Ω.
- *Spurious emissions.* The level of these is critical for the prevention of interference with other users on different frequencies. The limit for VHF and UHF is a maximum of 0.25 μW.
- *Residual noise.* Not always quoted by manufacturers, it is the noise level existing on an unmodulated carrier. A typical figure is better than −40 dB referred to full deviation.
- *Audio frequency distortion.* Typically <3% and usually measured with a modulating frequency of 1 kHz at 60% modulation.
- *Audio frequency response.* This is the variation of modulation level over the audio frequency spectrum. Typically within +1 dB to −3 dB over a frequency range of 300 to 3000 Hz (2.55 kHz for 12.5 kHz

channel spacing equipment). It may be quoted with reference to a pre-emphasis curve.

- *Switching bandwidth.* This is the frequency range over which the transmitter will operate without retuning and without degradation of performance. Much equipment is now specified to cover a complete frequency band, e.g. 146–174 MHz, without retuning.

10.2 Receivers

10.2.1 Receiver functions

A receiver's functions are:

- Detect a weak signal; the minimum level, which may be as low as 0.25 microvolts, defines the receiver sensitivity.
- Amplify a received signal and maintain the information contained in a minimum strength signal at a minimum of 12 dB above the electrical noise level (signal-to-noise ratio). If the audio distortion produced in the receiver is also taken into account the above figure becomes the signal-to-noise + distortion (Sinad) ratio. As the signal is increased, the ultimate Sinad should attain 50 to 55 dB.
- Separate the wanted signal from any unwanted ones which may be very close in frequency (the adjacent channel may be 12.5 kHz away at UHF); the selectivity.
- Recover the information from the carrier; demodulation.
- Amplify the audio information to a level suitable for operating a loudspeaker; the audio power output. The audio amplification must introduce the minimum distortion.
- Disenable the audio amplifiers in the absence of signal to cut out the electrical noise. This is done by the mute or squelch circuit.

10.2.2 Types of receiver

It is possible to amplify directly the incoming RF signal to a level suitable for demodulation. This is done in a tuned radio frequency (TRF) receiver, but these are seldom used today because of the problems of obtaining sufficient selectivity and gain at one radio frequency, and the difficulty of retuning a number of RF stages to change frequency. Almost all receivers designed for analogue communications now operate on the superheterodyne principle where the incoming radio frequency is converted to a lower, more manageable, intermediate frequency (IF). The fixed IF means that only the oscillator and, possibly, one RF amplifier stage need retuning for a change of channel.

The lower frequency of the IF facilitates the acquisition of adequate gain with stability and selectivity.

There is little difference in the layout of receivers for AM and FM except that the circuits perform their functions differently. *Figure 10.3* is a block diagram of a typical FM receiver with a crystal controlled local oscillator.

When two frequencies are applied to a non-linear circuit such as the mixer, they combine to produce other frequencies, their sum and difference being the strongest. The superheterodyne mixes a locally generated frequency with the received signal to produce a, usually lower, frequency retaining the modulation of the received signal. Commonly used values for this intermediate frequency (IF) are 465 kHz for MF and HF receivers and 10.7 MHz for VHF and UHF. At these fixed frequencies the necessary high amplification with low noise and stability and the required selectivity are easier to obtain. The local oscillator (injection) frequency = signal frequency ± IF frequency.

Superheterodyne receivers are susceptible to a particular form of interference. Assume a local oscillator frequency of 149.3 MHz is mixed with a wanted signal of 160 MHz to produce the IF frequency of 10.7 MHz. A signal of 138.6 MHz would also combine with the local oscillator to produce 10.7 MHz. The frequency of this spurious response is the image, or second channel frequency. The IF amplifier cannot discriminate against it so some degree of selectivity must also be provided in the RF amplifier and input circuitry. Intermediate frequencies are chosen which are a compromise between ease of obtaining adjacent channel discrimination and image frequency rejection.

The IF amplifier contains a block filter, either crystal or ceramic (see *Figure 7.7*), necessary to discriminate between channels adjacent in frequency. The design of the filter is crucial. It must be wide enough to accommodate the band of frequencies present in the modulation plus an allowance for frequency drift and its response over this band must be uniform with minimal ripple, particularly if data is to be received, yet its response must be of the order of -100 dB at the frequency of the adjacent channel.

Apart from the demodulator the main difference between AM and FM receivers lies in the operation of the IF amplifier. The IF stages in an AM receiver are linear, although their gain is variable. Part of the IF amplifier output is rectified and used to control the gain to provide automatic gain control (AGC). An increase of signal above a predetermined level, with delayed AGC, causes a reduction in IF gain maintaining a sensibly constant audio output level.

The IF amplifier in a FM receiver possesses a very high gain, some 100 dB, and is non-linear. On receipt of a signal, or even with only the receiver noise, it runs into limitation, cutting off both positive and negative peaks of the signal or noise. This gives FM its constant level audio output over a wide range of signal levels. It also produces the capture effect where a strong signal, fully limiting, completely removes a weaker signal. A signal difference of some 6 dB is required to provide effective capture.

Most receivers employ only one change of frequency, single superheterodynes, but double superheterodynes are occasionally used at VHF and above. A double superheterodyne changes the frequency twice, perhaps to 10.7 MHz for the first IF and then, using a fixed frequency crystal second local oscillator, to a lower, often 1.2 MHz or thereabouts, second IF. The result is greater gain and selectivity but the incorporation of a second oscillator and mixer increases the number of possible spurious responses.

A variation of a very old receiver circuit, the Autodyne, forgotten in about 1914, is now finding favour in receivers for digitally modulated signals. Under its new names of Homodyne or zero-IF receiver it lends itself to the purpose. In this type of receiver the local oscillator runs at the same frequency as the incoming signal, hence the zero-IF. Frequency or phase shifts of the incoming signal representing the data emerge from the mixer at base band and are applied directly to the processing circuits.

In a communications receiver the noise generated in the aerial and RF stages in the absence of a signal is amplified to what may be, when demodulated, an unacceptable level. To eliminate the annoyance the loudspeaker is switched off during no-signal periods by a squelch or mute circuit. The mute circuit rectifies the high frequency noise at the demodulator and uses it to switch off the audio amplifier. When a signal is received the limiting action of the IF amplifier of an FM receiver depresses the noise in favour of the signal. Some AM mobile receivers use additional FM circuitry to provide improved mute action.

10.2.3 Noise figure

An ideal receiver would generate no noise and the signal-to-noise ratio, in a receiver of given bandwidth, would be determined by the level of the signal at the base of the antenna compared with the noise produced in the antenna. The noise factor of the ideal receiver is the number of times the signal power must exceed the antenna noise power to produce a 1:1 ratio at the receiver. It is given by (see Section 1.5.1):

$$\frac{e^2_{\text{antenna}}(\text{e.m.f.})}{4kTBR}$$

When the receiver input impedance is matched to the antenna impedance, half the power is dissipated in the antenna and the noise factor is 2 (3 dB). In a practical receiver, the noise generated in the RF amplifier is the most significant, and the noise figure is the sum of the RF amplifier noise plus all the preceding losses. Figures between 4 and 6 dB are common and the higher the noise figure, the worse the receiver sensitivity.

Signal-to-noise ratio (SNR, S/N or S_N)

Receivers are evaluated for quality on the basis of *signal-to-noise ratio* (S/N or 'SNR'), sometimes denoted S_N. The goal of the designer is to enhance the SNR as much as possible. Ultimately, the minimum signal level detectable at the output of an amplifier or radio receiver is that level which appears just above the noise floor level. Therefore, the lower the system noise floor, the smaller the *minimum allowable signal.*

Noise factor, noise figure and noise temperature

The noise performance of a receiver or amplifier can be defined in three different, but related, ways: *noise factor* (F_N), *noise figure* (NF) and *equivalent noise temperature* (T_E); these properties are definable as a simple ratio, decibel ratio or Kelvin temperature, respectively.

Noise factor (F_N). For components such as resistors, the noise factor is the ratio of the noise produced by a real resistor to the simple thermal noise of an ideal resistor.

The noise factor of a radio receiver (or any system) is the ratio of output noise power (P_{NO}) to input noise power (P_{NI}):

$$F_N = \left[\frac{P_{NO}}{P_{NI}}\right]_{T=290/K}$$

In order to make comparisons easier the noise factor is usually measured at the standard temperature (T_o) of 290 K (standardized room temperature); although in some countries 299 K or 300 K are commonly used (the differences are negligible). It is also possible to define noise factor F_N in terms of the output and input signal-to-noise ratios:

$$F_N = \frac{S_{NI}}{S_{NO}}$$

where

S_{NI} is the input signal-to-noise ratio
S_{NO} is the output signal-to-noise ratio

Noise figure (NF). The *noise figure* is the frequency used to measure the receiver's 'goodness', i.e. its departure from 'idealness'. Thus, it is a *figure of merit*. The noise figure is the noise factor converted to decibel notation:

$$\mathrm{NF} = 10 \log(F_N)$$

where

NF is the noise figure in decibels (dB)
F_N is the noise factor
log refers to the system of base-10 logarithms

Noise temperature (T_e). The noise 'temperature' is a means for specifying noise in terms of an equivalent temperature. That is, the noise level that would be produced by a resistor at that temperature (expressed in degrees Kelvin). Evaluating the noise equations shows that the noise power is directly proportional to temperature in degrees Kelvin, and also that noise power collapses to zero at the temperature of Absolute Zero (0 K).

Note that the equivalent noise temperature T_e is *not* the physical temperature of the amplifier, but rather a theoretical construct that is an *equivalent* temperature that produces that amount of noise power in a resistor. The noise temperature is related to the noise factor by:

$$T_e = (F_N - 1)T_o$$

and to noise figure by

$$T_e = KT_o \log^{-1}\left[\frac{\mathrm{NF}}{10}\right] - 1$$

Noise temperature is often specified for receivers and amplifiers in combination with, or in lieu of, the noise figure.

Noise in cascade amplifiers

A noise signal is seen by any amplifier following the noise source as a valid input signal. Each stage in the cascade chain amplifies both signals and noise from previous stages, and also contributes some additional noise of its own. Thus, in a cascade amplifier the final stage sees an input signal that consists of the original signal and noise amplified by each successive stage plus the noise contributed by earlier stages. The overall noise factor for a cascade amplifier can be calculated from *Friis' noise equation*:

$$F_N = F_1 + \frac{F_2 - 1}{G1} + \frac{F_3 - 1}{G1G2} + \cdots + \frac{F_N - 1}{G1G2 \cdots G_{n-1}}$$

where

F_N is the overall noise factor of N stages in cascade
F_1 is the noise factor of stage-1
F_2 is the noise factor of stage-2
F_N is the noise factor of the *n*th stage
$G1$ is the gain of stage-1
$G2$ is the gain of stage-2
G_{n-1} is the gain of stage $(n - 1)$.

As you can see from Friis' equation, the noise factor of the entire cascade chain is dominated by the noise contribution of the first stage or two. High gain, low noise radio astronomy RF amplifiers typically use low noise amplifier (LNA) circuits for the first stage or two in the cascade chain. Thus, you will find an LNA at the feedpoint of a satellite receiver's dish antenna, and possibly another one at the input of the receiver module itself, but other amplifiers in the chain might be more modest (although their noise contribution cannot be ignored at radio astronomy signal levels).

The matter of signal-to-noise ratio (S/N) is sometimes treated in different ways that each attempts to crank some reality into the process. The signal-plus-noise-to-noise ratio (S + N/N) is found quite often. As the ratios get higher, the S/N and S + N/N converge (only about 0.5 dB difference at ratios as little as 10 dB). Still another variant is the SINAD (signal-plus-noise-plus-distortion-to-noise) ratio. The SINAD measurement takes into account most of the factors that can deteriorate reception.

10.2.4 Signal-to-noise ratio and bandwidth

To obtain the maximum signal-to-noise ratio the bandwidth of every circuit must be designed to admit its operational band of frequencies only. The wider its bandwidth the more noise a circuit admits, and the more the bandwidth exceeds that needed, the worse becomes the signal-to-noise ratio. There is a linear ratio between bandwidth and noise power admitted: doubling the bandwidth doubles the noise power.

Improvements in local oscillator crystal frequency stability has resulted in improved signal-to-noise ratios by reducing the necessary width of the IF filter.

Demodulation, the recovery of the audio from the IF bandwidth affects the signal-to-noise ratio. The relationship is complex, but consider two examples. First, for 12.5 kHz channel spacing FM:

The audio frequency range is 300 Hz to 3000 Hz
Bandwidth, $b = 2700\,\text{Hz}$
The modulation index $M = f_{\text{d(max)}}/f_{\text{m(max)}}$
Signal/noise out $= 3M^2/2b\times$ signal/noise in
For a 12.5 kHz PMR channel:
    $M = 2500/3000 = 0.83$
    $b = 2.7\,\text{kHz}$
    $3M^2/2b = 0.38$

A 0.38 times reduction in power is $-4.8\,\text{dB}$. Demodulation in this case worsens the signal-to-noise ratio by some 5 dB, and if a signal-to-noise ratio of 12 dB is required at the loudspeaker, 17 dB is needed at the input to the demodulator.

When channel separations were 25 kHz, and deviation 5 kHz, the situation was:

Audio frequency range, 300 Hz to 3000 Hz
Bandwidth, $b = 2700\,\text{Hz}$
For a 50 kHz channel:
    $M = 5000/3000 = 1.66$
    $b = 2.7\,\text{kHz}$
    $3M^2/2b = 1.54$

A 1.54 times gain in power is $+1.8\,\text{dB}$. In this case the demodulation improved the signal-to-noise ratio slightly. Reducing the bandwidth would have the same effect.

For AM, the signal-to-noise ratio is dependent on the modulation depth. The demodulation process reduces the signal-to-noise ratio by

6 dB but the recovered audio is less than the IF bandwidth by a factor of 3:1 which compensates for the demodulation loss. Reducing the modulation depth degrades the signal-to-noise ratio by 6 dB for every halving of the modulation depth.

10.2.5 Receiver specifications

The important features of receiver specifications are:

- *Sensitivity*. The minimum signal to which a receiver will respond. For an AM receiver, the generally accepted standard is the signal (30% modulated with sinusoidal tone, either 400 Hz or 1 kHz) required to provide an audio output of 50 mW. For an FM receiver, the standard is the unmodulated signal required to produce a 20 dB reduction in noise. A typical figure is 0.25 μV (p.d.) for 20 dB quieting. However, sensitivity is often quoted in terms of either the signal-to-noise ratio or Sinad so, in modern parlance, sensitivity and signal-to-noise ratio are sometimes considered to be synonymous.
- *Signal-to-noise ratio (may be quoted as Sinad, signal-to-noise and distortion)*. Typically 0.3 μV (p.d.) for 12 dB Sinad.
- *Spurious response attenuation*. Typically better than 80 dB.
- *Adjacent channel selectivity*. Better than 65 dB at 12.5 kHz channel spacing.
- *Cross modulation*. The modulation, in the receiver, of a wanted signal by a stronger, unwanted signal. It is usually caused by non-linearity in the receiver RF stages.
- *Blocking and de-sensitization*. The reduction in sensitivity of a receiver due to overloading of the RF stages when a strong signal is applied. A blocked receiver may take an appreciable time to recover.
- *Audio frequency response*. Typically within +1 dB to −3 dB of a 6 dB/octave de-emphasis curve from 300 Hz to 3000 Hz (2.55 kHz for 12.5 kHz channel spacing, above which the response falls more rapidly).
- *Audio output*. Typically 3–4 W. For hand-portables, 100–500 mW. (Distortion may also be quoted, typically better than 5%.)
- *Switching bandwidth*. As for transmitter.
- *Duplex separation*. For a receiver/transmitter combination, the minimum separation between the receiving and transmitting frequencies which will permit duplex operation with minimal degradation of the receiver performance.

10.3 Programmable equipment

The current trend is for many of the functions of radio equipment – even down to the control buttons on the front panel – to be software controlled. In one manufacturer's equipment the only screwdriver-adjustable control is that for setting maximum deviation; this is insisted upon by the Radiocommunications Agency for use in the UK.

Software control enables radio sets to be cloned so that once a single piece of equipment is programmed with its frequencies, power output, selective calling, etc., the whole of a fleet of mobiles can be identically programmed in a short time. Some of the functions can be programmed by the user, but not all, and different manufacturers permit different degrees of programming.

The functions are allocated degrees of priority which determine which functions can be changed by each class of person. Again, some of the priorities may be altered by the user but the highest priority functions are installed in a programmable read only memory (PROM) by the manufacturer and can only be altered by replacing the PROM, in most instances a job for the manufacturer.

The essential equipment for major reprogramming is an IBM or equivalent computer and the equipment manufacturer's software which is supplied as a package complete with hardware interface and instructions.

Cloning may be carried out using an inter-connecting cable between the equipment, or restricted reprogramming via a unit supplied by the equipment manufacturer for the purpose.

The functions which may be programmed from a computer are:

- Channel frequencies.
- Transmitter power and deviation.
- Receiver squelch setting (referred to Sinad ratio). Selective calling details including encode or decode only, type of signalling, extension of first tone and call sign.
- Transmission time out timer.
- Channel busy light. Transmitter inhibit on busy channel.
- Alert tones. Tone decoder indicator.
- Low battery indicator. Battery saver.

Restricted reprogramming may include:

- Channel frequencies, including any offsets.
- Lock out of any channels.

- Channel spacing.
- Channel search priority.
- Frequency stability.
- Signalling type.
- Control functions.
- Synthesizer reference crystal (the frequency can be trimmed to compensate for ageing).
- Transmitter power.
- CTCSS (encode and decode).
- Timers.
- Alerts.

Pagers are also programmable. The functions which can be changed include:

- Code number.
- Alert tone/vibrator/repeat.
- Urgency.
- Out of range warning.
- Printer on/off.
- Language of display.

References

Belcher, R. *et al*. (1989). *Newnes Mobile Radio Servicing Handbook*. Butterworth-Heinemann, Oxford.

Langford-Smith, F. (1955). *Radio Designer's Handbook*, Iliffe, London.

11 Microwave communication

11.1 Microwave usage

Microwaves are loosely considered to be those at frequencies between 1 GHz (30 cm) and 100 GHz (0.3 cm). Principal ground communications usage is for point-to-point links carrying information and control signals on systems such as multiplexed communication – including data networks, telemetry and mobile radio. The propagation of the higher frequencies, 30–100 GHz, is being studied for possible future use in micro-cellular radio-telephone systems. These higher frequencies are also used for satellite communications and radar applications. Microwave frequencies up to 3 GHz have now been reserved for mobile use.

11.2 Propagation

The path loss is higher at microwave frequencies than at VHF and UHF: *Figure 1.5* charts the free space loss between isotropic radiators for microwaves at frequencies of 1, 2, 3.5 and 7 GHz over distances of 10 to 1000 km. The free-space loss between isotropic radiators is given by

$$\text{Free space loss, d}B = 32.4 + 20\log_{10} d + 20\log_{10} f$$

where d is in km and f in MHz.

The free-space loss between practical antennas is given by:

$$\text{Loss, dB} = 10\log_{10}\left(\frac{(4\pi d)^2}{\lambda^2}\frac{1}{G_\text{t}G_\text{r}}\right)$$

where

d = path length, metres
λ = wavelength, metres
G_t = power gain of transmitting antenna
G_r = power gain of receiving antenna

The antenna gains are expressed relative to an isotropic radiator (not in dB).

Absorption varies with atmospheric humidity, and as energy is also absorbed by the ground the path height therefore has an effect on the losses. Absorption by rain is a factor at the higher frequencies but, although considered to be insignificant below about 3 GHz, it has an indirect effect. Wet foliage, for example, produces considerable absorption at frequencies as low as 450 MHz.

Waves of millimetric lengths are special cases and narrow bands of very high absorption due to resonance effects exist at 22 and 183 GHz for water vapour, and 60 and 119 GHz for oxygen. Non-resonant attenuation occurs due to scatter from rain, hail and snow. The attenuation increases with frequency as the wavelength approaches the dimensions of a raindrop. Bands of very low absorption, ‘atmospheric windows’, where the water vapour and oxygen attenuations are very low occur at 37, 97, 137 and 210 GHz.

Objects close to a path may severely affect the received signal and a proposed path must be examined for the likely effects of these at the planning stage. The effects may be due either to diffraction bending the wave away from the line of sight between the antennas, or reflection causing multi-path signals.

The additional losses in a microwave path caused by objects either intruding into the first Fresnel zone or close to it, and which exhibit ‘knife-edge’ or ‘smooth-sphere’ characteristics, are shown in *Figure 11.1.* A negative F/F_1 indicates an intrusion. When the path clearance exceeds 0.6 times the first Fresnel zone radius, the free space loss is achieved.

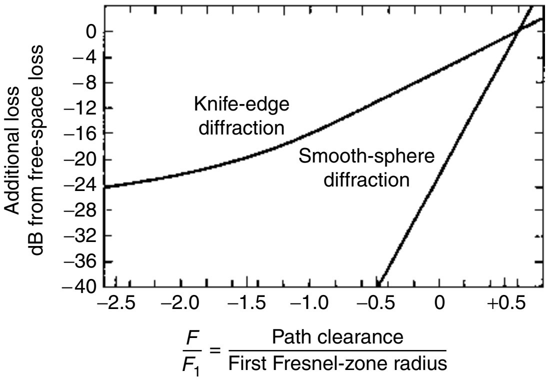

Figure 11.1 Knife-edge and smooth-sphere diffraction

Refraction also affects microwave propagation and temperature inversions may cause ducting resulting in a loss of signal and, possibly, interference. When ducting occurs, a layer of air of low refractive index is formed between two highly refractive layers and a wave may be trapped between them. Under these conditions a signal can be carried an abnormal distance before it can return to Earth resulting in loss of signal at the intended receiver and possibly interference at a distant one on the same frequency.

11.3 *K* factor

The degree and sense of refraction of wave is related to a factor known as K. Refraction normally bends the wave downwards, extending the range, but the refractive index of the atmosphere varies from place to place and with time and height. At times, during an atmospheric inversion for instance, and in some places on the earth the effect is reversed:

When no bending occurs, $K = 1$.
When $K > 1$, the bending is downwards, effectively increasing the Earth's radius.
When $K < 1$, the bending is upwards, effectively reducing the Earth's radius.

For most of the time $K > 1$, 1.33 being accepted as the normal factor, but for small periods K may be less than one.

For a point-to-point link, the path between the transmitting and receiving sites should be a clear line of sight, although by making allowances for refraction, the Earth's radius has, up to now, generally been considered to be effectively increased by the factor of 1.33. However, in the interests of reliability a decrease to 0.7 times the radius (the minimum K factor considered likely to occur) is now often used in link planning. When a link has been planned using a higher K factor, a temporary reduction of K not only reduces the radio horizon but effectively raises objects close to the path, possibly to the point where they become significantly close to the first Fresnel zone.

It is usual when planning a link to plot the profile of a path on paper with curved horizontal graduations to represent the amended Earth's radius. The radio beam can then be drawn as a straight line between the antenna locations. *Figure 11.2* is an example.

11.4 Fresnel zones, reflections and multi-path fading

Signals which arrive at the receiver by more than one path as the results of reflection or diffraction may arrive in any phase relationship to the direct wave. When they arrive anti-phase to the direct wave, cancellations result. The intensity and phase of the spurious signal may not be constant, thus providing random multi-path fading.

Where a carefully drawn profile of a link path shows there to be a clear line of sight, the effect of waves reflected or diffracted from objects close to the line of the direct wave must then be considered. The effect of these indirect waves can be predicted by calculating where the reflection occurs in relation to a series of ellipsoids which can be drawn around the line-of-sight path between the transmitting and receiving antennas. These ellipsoids, known as the Fresnel zones, contain the points where reflected waves will follow a path of constant length, as shown in *Figure 11.3*.

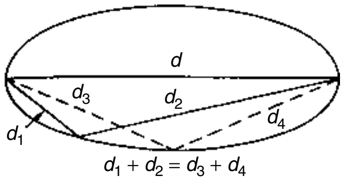

Figure 11.3 Fresnel zone: reflected path lengths

Waves reflected at the odd-numbered Fresnel zones will travel an odd number of half-wavelengths further than the direct wave but, because a 180° phase change usually occurs in the reflection process, will arrive at the receiver in phase with the direct wave. Waves reflected at even-numbered zones will arrive anti-phase to the direct wave with a cancelling effect. The effect of reflected waves diminishes with reflections from the higher order zones. The radius of a Fresnel zone in metres at the point of intrusion is given by:

First zone, F_1:

$$F_1 = 31.6\sqrt{\frac{\lambda d_1 d_2}{d}}$$

or

$$F_1 = 548\sqrt{\frac{d_1 d_2}{fd}}$$

where

F_1 = radius in metres at point of intrusion
$d_1 + d_2 = d$ (path length in km)
λ = wavelength in metres
f = frequency in MHz

Second zone:

$$F_2 = \sqrt{2} \times F_1$$

Third zone:

$$F_3 = \sqrt{3} \times F_1$$

and so on.

The degree of reflection from an object depends on its nature, the greatest reflection occurring from smooth flat ground or water. Where a path lies over the sea, variations in the path length of a reflected wave due to tides may render a path unusable. When the height of the antenna closest to the sea is varied, the *effect* of the reflected wave passes through a series of minima and maxima and adjustment of the height of that antenna can reduce or, occasionally, overcome the effect.

Atmospheric conditions change giving rise to fading and variations of the multi-path effects. The reliability of a link may be crucial to the success of a complete system and, where a critical path in terms of performance exists, long-term tests are advisable to ensure that variations of propagation do not reduce the reliability to an unacceptable level. Paths which contain obstacles in the line of sight which will cause additional losses are obviously suspect. So are those where objects or large stretches of water or flat ground which might produce diffraction or reflections of the wave lie close to the line of sight.

11.5 Performance criteria for analogue and digital links

The transmission quality for analogue modulated systems is based on the signal-to-noise ratio. The noise is specified relative to a standard test tone level and is commonly expressed as either picowatts psophometrically weighted (pWp) or decibels (dB) of C-message weighted noise above a reference noise level of −90 dBm (defined as 0 dBrnc0). Typical objectives range from 28 dBrnc0 for long-haul routes to 34 dBrnc0 for short-haul routes. When the signal fades the noise increases until the threshold noise level is reached. When the threshold (typically 55 to 58 dBrnc0) is exceeded, the transmission quality is considered unacceptable.

With digitally modulated systems the bit error rate (BER) is the measure of transmission quality. The bit error rate is the number of bit errors per total received averaged over a period of time. If the transmission rate is 10 Mbits per second and 100 bit errors occur over a 100 second period, the BER is 10^{-7}, an average of 1 error in 10^6 bits. The acceptable level of transmission is determined by the type of traffic.

For PCM voice traffic, bit errors manifest themselves as clicks and a threshold of 10^{-6} (1 click approximately every 15 seconds) is usually considered acceptable. At this threshold the speech is intelligible, but beyond it the clicks become annoying and intelligibility falls rapidly.

For data with error correction a higher BER of 10^{-8} is normally acceptable (Communications International, 1989).

11.6 Terminology

A number of different units are used worldwide to define the performance and transmission levels of a radio relay system (for decibel definitions see Chapter 2). Important international definitions and units are:

- *Zero transmission reference point*. This is a point arbitrarily established in a transmission circuit, with all other levels in the circuit being stated with reference to this point. Its relative level is 0 dBr.
- *Standard test tone*. The standard test tone for use at audio circuit points is defined as a power of 1 milliwatt (0 dBm) at a frequency of 1000 Hz applied at the zero transmission reference level point.

11.7 Link planning

Planning a link involves producing a profile of the path and calculating the net loss in the system to arrive at a transmitter output power which will produce the designed signal-to-noise ratio at the receiver. A simple example using the profile of *Figure 11.2* is shown in Section 11.8.

Transmission lines and waveguides are discussed in Chapter 3 and microwave antennas in Chapter 4.

11.8 Example of microwave link plan

Frequency: 2000 MHz		
Antenna type, station A: P6F-17C	height agl.	20 m
Antenna type, station B: P6F-17C	height agl.	20 m
Feeder type, station A: LDF5P50A	loss, dB/100 m	6.46
Total length, antenna to equipment		30 m
Feeder type, station B: LDF5P50A	loss, dB/100 m	6.46
Total length, antenna to equipment		30 m

Performance

Path length: 20 km, therefore clear path loss	124.0 dBi
Obstruction loss	0.0 dB
Feeder loss, station A:	1.9 dB
Feeder loss, station B:	1.9 dB
Feeder tail loss, total for link	1.5 dB
Connector loss, total for link	1.5 dB
Total loss	**130.8 dB**
Gain, antenna A	28.6 dB
Gain, antenna B	28.6 dB
Total gain	**57.2 dB**
Nett loss (total loss − total gain)	**73.6 dB**
Receiver threshold for max. signal/noise	−125.0 dBW
Design fade margin	+30.0 dB
Design receiver input level (threshold − fade margin)	**−95.0 dB**
Transmitter output power (receiver input − nett loss)	**−21.4 dBW**

Reference

Krzyczkowski, M. (1989). *Communications International*, August.

12 Information privacy and encryption

12.1 Encryption principles

Radio communication was never secret, but since the advent of fast frequency scanning receivers the ability to overhear, even on cellular radio telephones, is within easy reach of anyone. Privacy systems are, however, available which will deter the casual listener and gain time against the determined eavesdropper.

Messages, which may be either speech or data, are encrypted to prevent both eavesdropping and the injection of spurious information. The aim is to make the encryption and decryption as easy and inexpensive as possible for authorized users and time-consuming, difficult and costly for the eavesdropper (cryptanalyst). *Figure 12.1* shows a standard cryptographic channel.

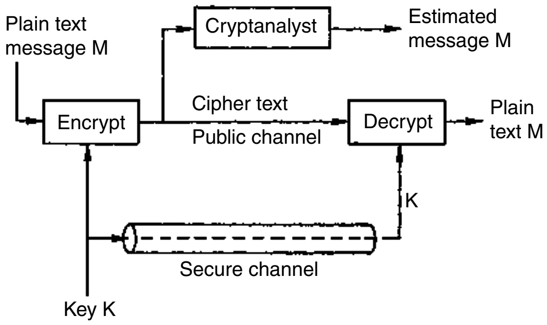

Figure 12.1 Cryptographic channel

A plain text message M (speech, written or digital) is encrypted by mixing with a key K to produce a cipher text. The cipher text may be transmitted over a channel which is accessible to the public and hence to the cryptanalyst. The key is issued via a secure channel to the authorized recipient who uses it to decipher the message. The cryptanalyst without access to the key attempts to derive the maximum information from the cipher text to enable him or her to estimate the content of the message.

One key may be used continuously or for long periods or, to increase the cryptanalyst's confusion, the key may be changed frequently, perhaps even for each character of the message. A sequence of key changes which repeats after a fixed number of characters produces what is known as periodic encryption.

Encryption may be either symmetrical or asymmetrical. Symmetrical encryption uses the same key for both encryption and decryption. Asymmetrical encryption uses a different key for each process, thus providing for different levels of authorization. Encryption keys may be supplied to many persons who are authorized to transmit encrypted messages but decryption keys may be issued to only a few authorized recipients.

12.2 Speech encryption

The encryption of speech offers fewer possibilities than does the encryption of written or digital data messages. The simplest method of encrypting speech is scrambling by inverting the speech frequencies; *Figure 12.2* shows this process.

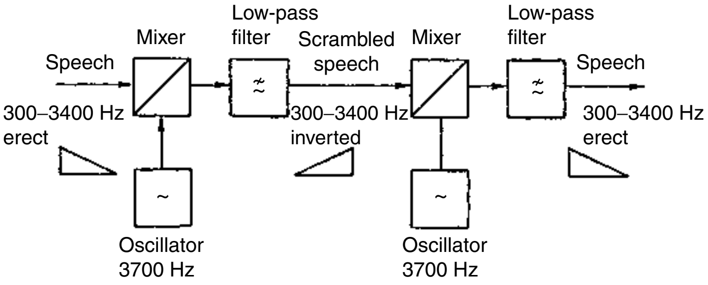

Figure 12.2 Speech inversion

The speech, contained in the band 300–3400 Hz, is mixed with a key frequency of 3700 Hz producing an erect, upper side band from 4000–7100 Hz, and an inverted, lower side band where the 300 Hz components of speech have become 3400 Hz and the 3400 Hz have been inverted to 300 Hz. The upper side band is rejected by a lowpass filter and the inverted lower side band is transmitted. In the receiver the scrambled speech is mixed again with 3700 Hz to produce an erect side band – the original non-inverted speech message.

Simple inverted speech is easily unscrambled. There is little choice of key frequency and if the eavesdropper uses a slightly different frequency the pitch is changed but the speech is readable. Also, if the centre frequencies only of the inverted band are selected by means of a band-pass filter, inverted speech becomes intelligible.

A sophistication which renders the speech more secure divides the speech band into sections and transmits them separately using a different key frequency for each band (audio frequency hopping). The divisions of the speech band may also be treated as blocks and transposed in time according to a user-programmable pattern to create further confusion in the mind of the eavesdropper.

The most up-to-date methods of speech scrambling convert the speech into digital form by either pulse code modulation (PCM) or some other method. The digits corresponding to the speech may then be either transmitted as frequency modulation, e.g. FFSK, on an analogue radio system or, possibly after further encryption, transmitted directly on a digital system. Digitized speech creates improved security not only by the digitization itself but by offering the higher encryption capabilities of data.

The price to be paid for security with analogue encryption is a degradation of the received signal-to-noise ratio by 9 dB, effectively reducing the range of a transmitter by approximately 40%.

12.3 Data encryption

Digital data may be encrypted by changing the digits, perhaps adding superfluous digits, and transmitting the resultant cipher message either in blocks of a fixed size or as a stream.

Block encryption treats the blocks in the same way as in the encryption of speech, with different keys being used for each block – or each character – and the blocks re-distributed in time.

Stream encryption has no fixed block size and each plain text bit, M_i, is encrypted with the ith element, K_i, of a pseudo-random, sometimes called pseudo-noise (PN), key.

Figure 12.3 shows two methods of generating pseudo-random keys or pseudo-noise. The first, *Figure 12.3(a)*, operates as follows. At each clock pulse the contents of the pre-loaded four-stage shift register are stepped forward from left to right. Immediately after the shift, the output bit is fed back into the 1st and 2nd stages. It introduces a new bit into stage 1, and is added by modulo 2 addition to the new content of stage 2, producing a new set of contents. The initial loading of 1100 emerges as a pseudo-random 15 bit sequence which then repeats. The periodicity of the sequence is given by:

$$\text{Sequence length, characters} = 2^N - 1$$

where $N =$ the number of stages in the shift register.

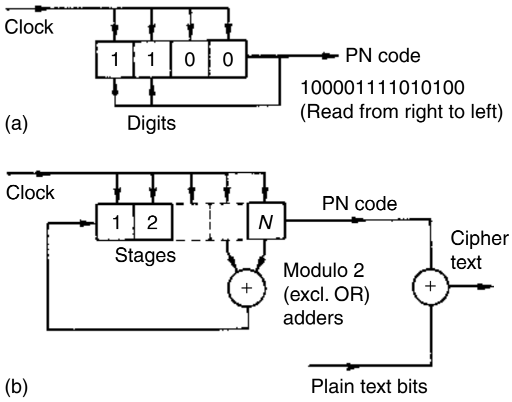

Code length = $2^N - 1$
e.g. $N = 5$, length = 31 characters

Figure 12.3 Pseudo-noise (PN) generation

For a four-stage register, therefore, the sequence repeats after 15 bits and is shown in *Figure 12.3(a).*

A more commonly used method combines the outputs of two or more of the earlier stages in a modulo 2 adder and feeds the result back to the input of the register as in *Figure 12.3(b).*

To form the cipher text the resultant pseudo-random key is mixed with the original data message in a second modulo 2 adder. If the clock rate for the shift register is the same as the bit rate of the plain text message, the plain text bits are exchanged for those of the modulo 2 sum, but if the shift register runs faster than the plain text bit rate, additional bits are added into the cipher text. This is more common and extends the time taken by a cryptanalyst to estimate the message. The price to be paid for the improved security is either a slower effective bit rate for the message or a higher overall bit rate and hence an increased bandwidth requirement.

Mixing the cipher text with the output from an identical PN generator in the receiver recovers the original text.

Modulo 2 addition

A modulo 2 adder is an exclusive-OR gate which produces a logic 1 output whenever either of the inputs is at logic 1 and the other is at logic 0. When both inputs are identical, the exclusive-OR gate produces a logic 0 output.

The truth table for an exclusive-OR gate is:

Input *A*	*B*	*Output* *Y*
0	0	0
0	1	1
1	0	1
1	1	0

Modulo addition is not limited to two inputs. Any quantity of binary numbers may be added: if there is an odd number of logic 1s in a column, the adder produces a logic 1 output, if an even number, i.e. no remainder in the binary addition, the output is logic 0.

Synchronous encryption

The key is generated independently of the message from a previously loaded register. If a character is lost during transmission of a synchronous text, resynchronization of transmitter and receiver key generators is necessary before transmission can continue.

Self-synchronous encryption

The key for each character is derived from a fixed number of previous characters of the plain text message. For example, the shift register is pre-loaded with the plain text characters so that in a four-stage register the key used for encrypting the 4th character will be the 4th previous message character. If a self-synchronized transmission loses a character, the system automatically re-synchronizes the same number of characters (in this case four) later.

Written messages may be encrypted using one of the classical mechanical methods of rearrangement of the letters before digital encryption.

When sufficient RF channels exist frequency hopping is a further possibility, and the spread spectrum technique, where the signal energy is spread over a very wide band of frequencies, not only offers very high security but also makes detection of the signal difficult. The shift register techniques described above are also used for generation of the frequency hopping sequence and the spreading of the base band frequencies.

12.4 Code division multiple access (CDMA) or spread spectrum

The extension of pseudo-random key or noise generation is code division multiple access or spread spectrum transmission, described in Chapter 8. The spread spectrum technique provides an extremely high level of security by reducing the radiated energy at any one frequency to very little above the ambient noise level by spreading the transmission over a very wide band. The transmitter uses what is in effect an extended digital key to spread the bandwidth and the receiver is equipped with an identical key for de-spreading. The transmission almost disappears into the noise and, without the appropriate key, the existence of a spread spectrum signal is very difficult to detect.

12.5 Classification of security

Unconditionally secure

Those systems where the cryptanalyst has insufficient information to estimate the content of the cipher regardless of the amount of time and computation facilities available. This is only realistic when using what is known as a one-time pad where the key is used once and once only.

Computationally secure

Encryption systems are specified in terms of the amount of time taken by a cryptanalyst to estimate the cipher's contents using the state of the art techniques. Unless an extremely long periodicity is used for a progressive key – months and even years in some instances – requiring many stages in the shift register, it is possible for a cryptanalyst who knows, or can estimate, a small part of the message to calculate all the parameters necessary to decipher the message. However, stream encryption with a pseudo-random key approaches perfect secrecy for a finite number of messages.

References

Chambers, W.G. (1985). *Basics of Communications and Coding*. Oxford University Press, Oxford.
Sklar, B. (1988). *Digital Communications*. Prentice Hall, Englewood Cliffs, NJ.

13 Multiplexing

Multiplexing enables several information (speech or data) channels to be carried simultaneously over one bearer, a wide band, single frequency microwave radio link for example. Both frequency and time division multiplex are common methods. On trunked radio systems where channels are allocated to users on demand the multiplexing is referred to as frequency- or time-division multiple access (FDMA, TDMA).

Frequency division multiplexing (FDM) used to be a common technique for sharing analogue trunk circuits between 12 (or multiples of 12) separate channels. This is an analogue technique that been superseded by the use of digital trunk circuits, where digital multiplexing is employed. FDM is still used on some satellite and microwave links, although most links are now using digital techniques. Some systems such as cellular telephones use separate transmit and receive frequencies, which is known as frequency division duplexing (FDD).

13.1 Frequency division multiplex

Frequency division multiplex divides a broad band of frequencies into slots, each wide enough to accommodate an information channel. This is achieved by amplitude modulating a higher frequency subcarrier with each speech signal to form groups of channels.

Each speech channel contains frequencies between 300 and 3400 Hz plus, in some systems, an out-of-band signalling tone of 3825 Hz and a guard band. Each channel modulates a base band subcarrier spaced at 4 kHz intervals upwards from 64 kHz. This produces an upper and lower side band from each channel (*Figure 13.1*). The carrier and upper side band are removed by filters and the lower (inverted) side band is transmitted. At the receiver, the base band frequencies are again mixed with the same subcarrier frequency to restore the original speech. The subcarrier frequencies are maintained to an accuracy of ±1 Hz which creates the ±2 Hz frequency translation error quoted in some telephone line specifications.

Twelve such channels form a ITU-T basic group B occupying the band between 60 and 108 kHz. This basic group B may now be

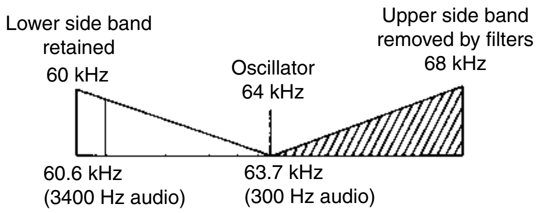
Lower side band
retained
60 kHz
Oscillator
64 kHz
Upper side band
removed by filters
68 kHz
60.6 kHz
(3400 Hz audio)
63.7 kHz
(300 Hz audio)

(a) Production of side bands in frequency division multiplex

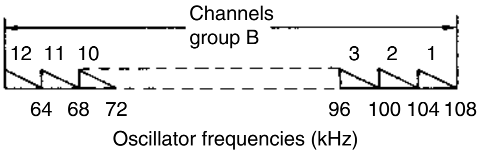
Channels
group B
12 11 10
3 2 1
64 68 72
96 100 104 108
Oscillator frequencies (kHz)

(b) Production of basic group B

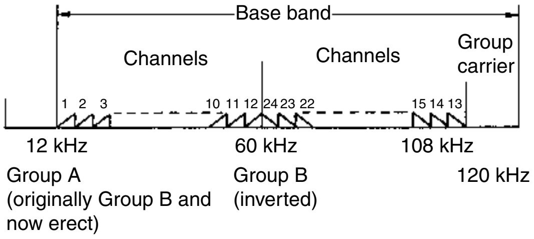
Base band
Channels
Channels
Group
carrier
1 2 3
10 11 12
24 23 22
15 14 13
12 kHz
60 kHz
108 kHz
Group A
(originally Group B and
now erect)
Group B
(inverted)
120 kHz

(c) Production of groups A and B

Figure 13.1 Frequency division multiplex

mixed with 120 kHz to produce a lower side band of 12 to 60 kHz, now basic group A. Filters leave 60 to 108 kHz free for a new basic group B.

The process may be repeated by using five basic groups to modulate still higher-frequency carriers, to produce super- and hyper-groups.

For FDM data communications, the bearer circuit bandwidth of 3000 Hz is divided into 12 channels each of 240 Hz bandwidth. Data is transmitted at 110 bits per second allowing a send-and-receive channel in each block of 240 Hz.

13.2 Time division multiplex (TDM)

A time division multiplex system conveys digital data, and speech must first be converted to data. TDM allocates short-duration time slots within a wider time frame to each information channel. For example, a continuous stream of data sent over a link at a rate of 2400 bit/s could convey the information contained in four 600 bit/s channels in short sequential bursts.

If the duration of one input bit is 1/600 s or 1.666 ms, a seven-bit character occupies 11.66 ms and 85 such characters can be sent per second. If the transmitted rate can be speeded up each bit sent at 2400 bit/s has a duration of 416 μs and 343 seven-bit characters can be sent per second. Such a system is shown in *Figure 13.2*. The data is stored in the buffers at the transmitter and the clock pulses are applied to each store/gate sequentially allowing one character from each data channel to be transmitted at a rate of 2400 bit/s. Perfect synchronization must be maintained between all channels and the transmitter and receiver to avoid data errors.

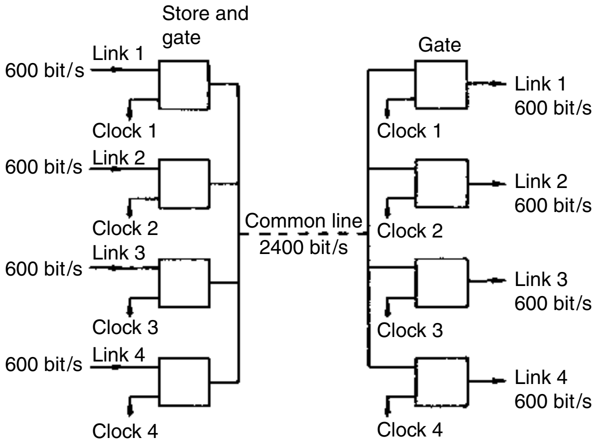
Store and gate
Link 1
600 bit/s
Clock 1
Link 2
600 bit/s
Clock 2
Link 3
600 bit/s
Clock 3
Link 4
600 bit/s
Clock 4
Common line
2400 bit/s
Gate
Link 1
600 bit/s
Clock 1
Link 2
600 bit/s
Clock 2
Link 3
600 bit/s
Clock 3
Link 4
600 bit/s
Clock 4

Figure 13.2 Time division multiplex

The principle is illustrated in *Figure 13.3* where a 1 s time frame at 2400 bit/s contains 343 time slots, each of 2.92 ms and containing a character from a specific information channel. Every second, therefore:

channel A would occupy slots 1, 5, 9, . . ., 337

channel B would occupy slots 2, 6, 10, . . ., 338

channel C would occupy slots 3, 7, 11, . . ., 339

channel D would occupy slots 4, 8, 12, . . ., 340

This leaves three blank slots; in practice slots are also allocated for preamble, address and synchronization purposes.

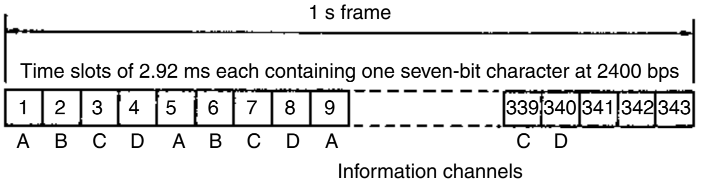
1 s frame
Time slots of 2.92 ms each containing one seven-bit character at 2400 bps
1
2
3
4
5
6
7
8
9
339
340
341
342
343
A B C D A B C D A
C D
Information channels

Figure 13.3 TDM time frame

13.2.1 European E1 multiplexing

The digital base band signal adopted in Europe operates at 64 kbit/s and the multiplexed signal at 2048 kbit/s. An eight-bit word or sample of a PCM voice channel (see Chapter 14) occupies 3.9 μs, and the interval between successive samples of a channel is 125 μs, the time frame duration. Therefore, the number of channels (time slots) that can be accommodated in one frame is $125/3.9 = 32$. Thirty of the slots are used for information channels and two for control purposes (*Figure 13.4*).

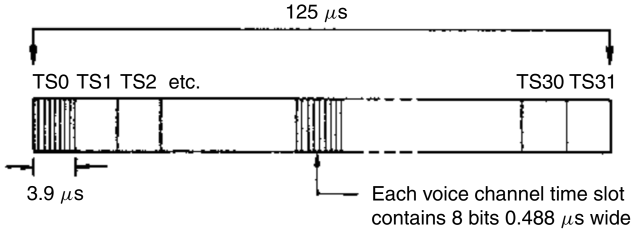
125 μs
TS0 TS1 TS2 etc.
TS30 TS31
3.9 μs
Each voice channel time slot contains 8 bits 0.488 μs wide

Figure 13.4 TS0, TS16 are used for signalling; TS1–TS15 and TS17–TS31 are used for voice channels

In the European E1 system, timeslots are numbered 0 to 32. Timeslot zero (TS0) is used for synchronization and TS16 is used for signalling. The signalling channel is therefore operating at 64 kbit/s. Digitized voice is carried in the remaining 30 timeslots. Voice channels 1 to 15 are carried over TS1–15. Voice channels 16–30 are carried over TS17–31.

13.2.2 American T1 (D1) multiplexing

The American T1 system is a 1.544 Mbit/s transmission system. The data format should be referred to as DS-1, but T1 is sometimes used instead. Timeslots (TS), or channels, in this system are numbered 1 to 24; there are 23 available to carry traffic and TS24 is used for synchronization. Each timeslot carries 8 bits, sampled at 8 kHz. An additional bit is sent at the end of each frame (known as the frame-bit or F-bit) so there are 24, 8-bit timeslots, plus 1 bit = 193 bits per frame.

The synchronization word sent in TS24 is coded as 1 0 1 1 1 Y R 0. The Y and R are dependent upon the data multiplexer and are for the manufacturer's use.

Signalling information is transmitted using 'stolen' bits from each channel, rather than using a separate timeslot as in the European E1 system. Bits are not stolen from timeslots in every frame, because this would degrade the quality of a voice channel. Instead, the least significant bit of every sixth frame is used for signalling. Thus the effective word length of the encoded voice is 7.833 bits, rather than 8 bits of the E1 system.

There are two virtual signalling channels (A and B) created by the use of stolen bits. Signalling Channel A is derived from bits stolen from timeslots during the sixth frame. Signalling Channel B is derived from bits stolen from timeslots during the twelfth frame. This is repeated for every superframe. Each channel has a data rate of 24 bits per 12 frames (1.5 ms), or 16 kbit/s, giving 32 kbit/s signalling in total. Although this data rate is half that of the E1 system, there are only 23 voice channels instead of the 30 used in the E1 system.

The F-bit is used to identify the frame, which is important when extracting the signalling information. Because only one in six frames are used for signalling, it is vital to know which ones carry the information and which ones carry voice traffic. Also there are two signalling channels and these must be separated at the receiver end of the system. A six bit data pattern of 1 0 0 0 1 1 is transmitted, followed by the inverse: 0 1 1 1 0 0. The complete 12-frame sequence is known as a superframe.

13.3 Code division multiple access (CDMA)

Spread spectrum transmission (described in Chapter 8) is a form of multiplexing. In addition to high security it permits multiple occupation of the wide – typically 1.25 MHz – frequency band. A number of users possessing keys of low correlation can occupy the same band at the same time. The system operates well in poor signal-to-noise or high interference environments.

Digital signals can be multiplexed by coding each source individually. Instead of transmitting logical 1's and 0's, a data pattern is assigned to each bit. The pattern for a logical 1 could be a ten-bit word (e.g. 1011010011); a separate pattern would represent a logical 0. Correlation (i.e. pattern matching) between a known signal and a signal received with a certain code gives an indication of whether the bit received is a 1 or a 0. This technique allows the occasional error in the data path without affecting the received bit. More often, it allows several users to transmit at the same frequency. Each user is allocated a different data pattern so that only the correct person will be able to decode the signal; all other signals will be ignored.

Reference

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

14 Speech digitization and synthesis

14.1 Pulse amplitude modulation

The digitization of analogue waveforms by pulse code modulation is accomplished in two stages. First the waveform is sampled to produce pulse amplitude modulation (PAM). Short-duration samples are taken at regular intervals which are long compared with the sampling time but short in relation to the highest signal frequency. The result is a train of pulses whose amplitude envelope is the same as the analogue waveform. The envelope (*Figure 14.1*) will contain:

- clock frequency f_c, the sampling rate
- all the signal frequencies contained in the waveform from the lowest, f_1, to the highest, f_2
- upper and lower side bands, $(f_c - f_2)$ to $(f_c - f_1)$ and $(f_c + f_1)$ to $(f_c + f_2)$
- harmonics of f_c and the upper and lower side band frequencies
- a DC component equal to the mean value of the PAM waveform.

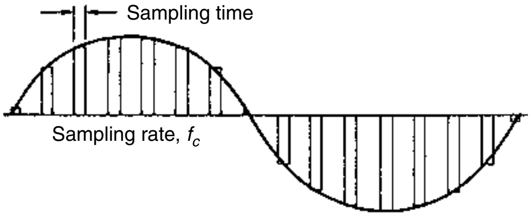

Figure 14.1 Pulse amplitude modulation

The envelope contains the original signal frequencies and can be demodulated by a low-pass filter which will pass f_2 but not the clock frequency. The clock frequency must therefore be higher than $2f_2$. For line communications in the UK a clock frequency of 8 kHz is used with a maximum modulating frequency of 3400 Hz.

14.2 Pulse code modulation

To overcome the susceptibility of PAM signals to corruption of the amplitude waveform by noise and interference, the waveform is processed further to produce pulse code modulation (PCM) before transmission.

In this process the magnitude of the PAM samples with respect to a fixed reference is quantized and converted to a digital (data) signal. Quantizing rounds off the instantaneous sample pulse amplitude to the nearest one of a number of adjacent voltage levels. *Figure 14.2* illustrates the process for an eight-level system. In the figure the amplitude at t_0 is 2, between 4 and 5 at t_1, between 5 and 6 at t_2, etc. After quantization the values would be 2 at t_0, 5 at t_1, 5 at t_2 and so on. The difference between the amplitude levels and the rounded-off values is the quantization noise or distortion.

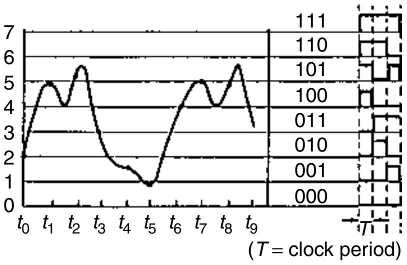

Figure 14.2 Eight-level pulse code modulation

The binary pulse train – leaving a one-bit synchronizing space between each number – for this example would be as in *Figure 14.3.*

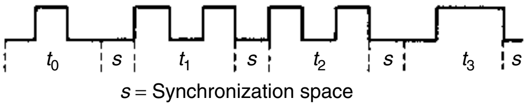

Figure 14.3 Binary pulse train (from *Figure 14.2*)

The number of quantizing levels is 2^n and the highest denary number represented is $(2^n - 1)$ where n is the number of bits used to represent each sample. If each train of pulses is accompanied by one synchronizing bit, the number of bits per sample is $(n + 1)$. If the sampling rate is f_c, the transmitted bit rate is $(n + 1)f_c$. For example, an eight-bit word – including the synch. bit – is commonly used to represent a sample, so, with a clock frequency of 8 kHz:

$$\text{number of quantizing levels} = 2^7 = 128$$

$$\text{sampling rate, } f_c = 8\,\text{kHz}$$

$$\text{therefore transmitted bit rate} = (n + 1)f_c$$

$$= 64\,\text{kb/s}$$

The maximum frequency of the pulses will be when transmitting alternate 1s and 0s, and the occupied bandwidth, $\frac{1}{2}$ × bit rate, 32 kb/s.

Quantization as described above is linear, i.e. the spacing of the quantization levels is the same over the range of pulse amplitudes. This produces a poor signal-to-noise ratio for low-level signals which is improved by the use of more closely-spaced levels at small signal amplitudes than at large amplitudes. Where non-linear quantization is used the most significant bit of the sample character identifies the polarity of the signal.

If linear quantization is used then about 12 bits per sample are needed to give a good voice quality, which requires a bit rate of about 96 kbits/s. For coding speech using non-linear quantization 8 bits per sample is sufficient, which requires a bit rate of 64 kbits/s. Two slightly different non-linear PCM codecs were standardized in the 1960s. In America u-law coding is the standard, but in Europe A-law coding is used.

14.3 ADPCM codecs

Adaptive differential pulse code modulation (ADPCM) codecs are waveform codecs which quantize the difference between the speech signal and a prediction of the speech signal. If the prediction is accurate, then the difference between the real and predicted speech samples will be small. The signal will be accurately quantized with fewer bits than would be needed using standard PCM. At the decoder the quantized difference signal is added to the predicted signal to give the reconstructed speech signal.

The performance of the codec is aided by using adaptive prediction and quantization. The predictor and difference quantizer adapt themselves to the changing characteristics of the speech being coded.

The CCITT (now ITU-T) standardized a 32 kbits/s ADPCM, known as G721, in the 1980s. The quality of the reconstructed speech was almost as good as that from standard 64 kbits/s PCM codecs. Since then G726 and G727 codecs operating at 40, 32, 24 and 16 kbits/s were standardized.

14.4 The G728 low delay CELP codec

At bit rates of around 16 kbits/s and lower the quality of waveform codecs falls rapidly. Thus, at these rates, CELP codecs and their derivatives, tend to be used. CELP stands for Code Excited Linear Prediction. A code representing a voice pattern is transmitted, rather than coding short samples of voice signal. The code is calculated by looking at the long-term changes in the voice and then predicting how the voice pattern will change. Since the pattern takes some time to establish, CELP codecs produce a delay.

A backward adaptive CELP codec was standardized in 1992 as G728. This codec uses backward adaptation in its calculations. This means that, rather than buffer 20 ms or so of the input speech for its calculations, the pattern of previous speech is used. This means that the G728 codec has to analyse fewer samples than traditional CELP codecs. The G728 codec uses only 5 samples giving it a total delay of less than 2 ms.

14.5 The GSM codec

The 'Global System for Mobile communications' (GSM) is a digital mobile radio system which is extensively used throughout Europe, and also in many other parts of the world. The GSM codec provides good quality speech, although not as good as the slightly higher rate G728 codec.

The GSM full rate speech codec operates at 13 kbits/s and uses a regular pulse excited (RPE) codec. This codec has short-term and long-term predictors to estimate how the voice pattern will change over time.

The input speech is split up into 20 ms long sections, called frames, and for each frame a set of eight filter coefficients are calculated for the short-term predictor. Each frame is then divided into four 5 ms sub-frames, and for each sub-frame the encoder calculates the delay and gain needed for the codec's long term predictor. Finally, after both short and long term prediction has been applied, the residual signal is quantized for each sub-frame. The filter coefficients for the short term predictor, the delay and gain of the long-term predictor and the coded residual signal are all transmitted. The receiver uses all these values to reconstruct the signal.

References

Kennedy, G. (1977). *Electronic Communications Systems*, McGraw-Hill Kogashuka, Tokyo.

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*, Butterworth-Heinemann, Oxford.

15 VHF and UHF mobile communication

Mobile communication operating throughout the VHF and UHF bands is expanding rapidly and, although the fastest growth has been in the radio-telephone field, interest in private mobile radio (PMR) is undiminished. Most of the procedures described for PMR are applicable to many other branches of radio communication, so traditional PMR is considered here. Other systems are discussed in later chapters.

A private mobile radio system, comprising a base station and mobiles, is one that is effectively owned by the user and, under the conditions of the licence, may only be operated by the user's own staff for his or her own business. Airtime cannot be leased to other persons.

15.1 Operating procedures

Frequencies are normally allocated in pairs, one for the up-link to the base station and one for the down-link to the mobile (not to be confused with a radio link used for control purposes). Such a pair of frequencies, spaced sufficiently apart to permit simultaneous transmission and reception by a station, comprises the radio channel. Occasionally, for special purposes and for small, low-power, and possibly temporary, operations a single frequency only may be allocated. The mountain rescue teams are an example.

The methods of operating are as follows.

Single-frequency simplex

This method uses a common frequency for transmission and reception by all stations operating on the system (*Figure 15.1*). Transmission

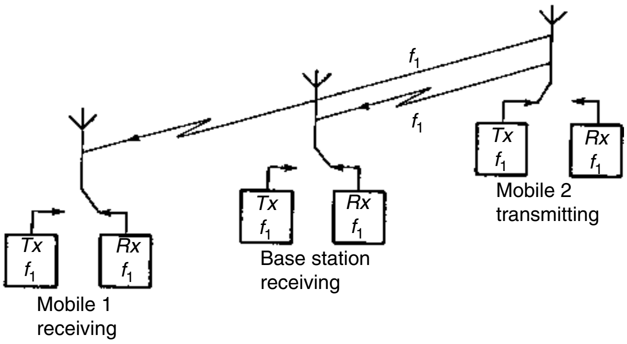

Figure 15.1 Single-frequency simplex

and reception cannot take place simultaneously at a station, and a receiver is switched off whenever the transmit switch is operated. This prevents blocking of the receiver by the transmitter and acoustic feedback occurring. The method allows all stations within range to hear both sides of a conversation and to relay messages to more distant stations; an obvious advantage for mountain rescue.

At the end of a transmission an operator must say ‘over’ and switch off the transmitter to hear the reply. A conversation must end with the word ‘out’ so that other stations are aware that the system is unoccupied.

Two-frequency simplex

Separate frequencies are used for transmission and reception but whilst a station is transmitting its receiver is still switched off (*Figure 15.2*). Mobiles hear only the base station and, therefore, the relaying of messages is not possible. A further disadvantage is that because mobiles hear only the base station, they may be unaware of the system occupancy and transmit, interfering with an existing conversation. The advantage of two-frequency simplex is the avoidance of receiver blocking or de-sensitization, not only from the associated transmitter but also, at base stations where several channels within the same band are located, from nearby transmitters.

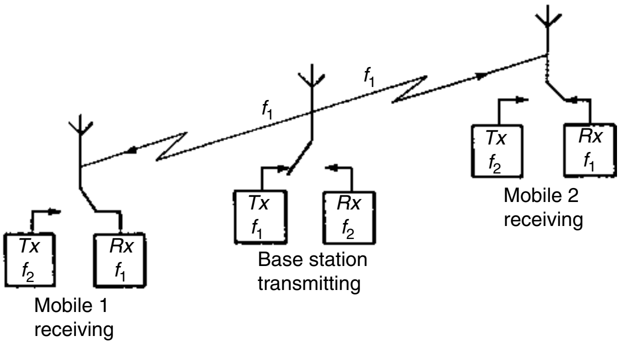

Figure 15.2 Two-frequency simplex

Duplex and semi-duplex

Separate frequencies are used for transmission and reception and, in full duplex, all stations can transmit and receive simultaneously as in a two-way telephone conversation (*Figure 15.3*). While a station is transmitting, its receiver audio output is switched from the loudspeaker

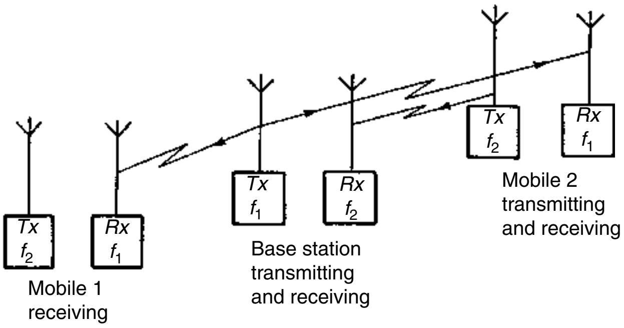

Figure 15.3 Duplex

to an earpiece to prevent acoustic feedback. A mobile cannot receive other mobiles directly but full duplex enables all stations to break in on a conversation in an emergency or to query part of a message; it also facilitates the use of talk-through where mobiles can speak to each other via the base station. To maintain awareness of system occupancy the base station may transmit a series of pips as an engaged signal during pauses in the despatcher's speech.

Many systems operate semi-duplex where only the base station operates a duplex procedure and the mobiles use a simplex procedure. This avoids the higher cost of duplex mobiles and offers most of the facilities of duplex, except that a despatcher cannot break in on a transmitting mobile.

Open channel/'All informed'

All mobiles hear all the calls from control, i.e. no selective calling is in operation.

Selective calling

Mobile receivers remain quiescent until specifically addressed; the opposite of open channel working. Individual mobiles, groups and a whole fleet may be addressed.

Auto-acknowledgement

When selectively called a mobile automatically transponds, sending the code for its address and, possibly, its status information. Mobiles can only acknowledge when individually addressed; auto-acknowledgement on group and fleet calls is prevented to avoid mobiles transponding simultaneously.

Status updating

The transmission, automatically or manually, of the data denoting the mobile's current status.

Call stacking

The storage of calls from mobiles and their presentation in call order to the despatcher. Arrangements are usually made to raise urgent calls to the top of the stack with an enhanced display.

15.2 Control of base stations

Where adequate radio signals over the desired service area of the system can be provided by one base station sited at the control point, it can be easily controlled, either directly from the front panel or over a multi-way cable.

In many instances, to obtain adequate coverage, the base station must be sited on high ground remote from the control point. Then, either a land line or a radio link, both of which will today probably be digital and the link microwave, must be used for control. A land line will most likely be rented but radio links are favoured by those users who insist on the complete system being under their direct control.

15.3 Common base station (CBS) operation

An economic method of providing mobile communication for users with a small quantity of mobiles and light traffic is a common base station or community repeater system. Although still referred to as private mobile radio, the base station is shared by several users who pay a fixed subscription for the service irrespective of airtime used. The station is controlled by radio using tone-controlled talk-through or what is sometimes called reverse frequency trigger (not normally permitted on a true PMR system). *Figure 15.4* shows the layout of a single-station CBS system. Each participant's office contains a fixed transmitter/receiver operating on the same frequencies as a mobile and equipped with CTCSS. Different CTCSS tones are assigned to each user to ensure privacy. All office stations must use a directional antenna so that they access only the base station to which they subscribe.

When a user transmits, from either the office station or a mobile, the base station, on receipt of a signal containing a valid CTCSS tone,

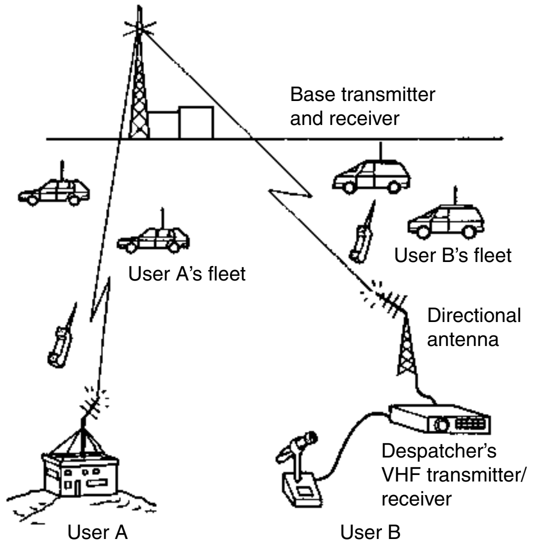
Base transmitter and receiver
User A's fleet
User B's fleet
Directional antenna
Despatcher's VHF transmitter/ receiver
User A
User B

Figure 15.4 Common base station system

enters a talk-through mode. The caller's CTCSS tone is retransmitted and, in turn, opens the mute on all that user's mobiles allowing communication via talk-through. Mobiles are equipped with 'busy' lamps to indicate whenever the system is engaged, and transmission time-out timers to prevent excessively long calls excluding other users.

15.4 Wide area coverage

Where the desired service area for one radio channel is larger than can be covered by a single base station the way in which the base stations are to be controlled requires special consideration. Crucial aspects are the presentation of the best received signal from a mobile to the control operator or despatcher and the selection of the transmitter most likely to provide the best signal at the location of the mobile.

Receiver voting

Well-proven receiver voting circuits which present the best received signal to the despatcher have been used for many years. These circuits sample the signals received from a mobile at each base station and, by means of coded information–which may be either digital or in the form of continuous tones–enable equipment at the control centre to automatically select the best. The selection may be made by comparing either the signal-to-noise ratio or the signal strength of the received signal. If the information is to be used solely to select the best signal in terms of readability, the signal-to-noise ratio is probably the better characteristic to use, but if the information is also to be used to select a base transmitter, the signal strength could be considered more satisfactory. Some systems utilize both types of information.

A typical 3-station voting system is shown in *Figure 15.5* where signal sampling and vote encoding occur at the receiving sites and the coded information is passed over the base station control system to the control centre. This method is necessary when the selection is made on a signal strength basis, but where the signal-to-noise ratio is used for the selection the sampling and encoding can be done at the control centre taking into account the noise occurring in the control network. Receiver voting systems operate very quickly, and changes

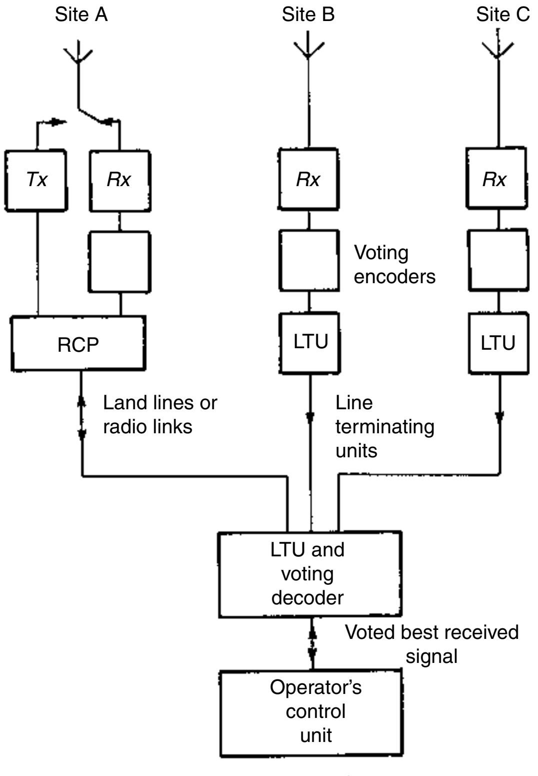
Site A
Site B
Site C
Tx
Rx
Rx
Rx
Voting encoders
RCP
LTU
LTU
Land lines or radio links
Line terminating units
LTU and voting decoder
Voted best received signal
Operator's control unit

Figure 15.5 Radio scheme with one transmit/receive plus two receiver-only stations with voting

of the selected receiving site may occur several times during a message without the despatcher's awareness.

The broadcast transmit and receive paths are not always reciprocal, for instance when low power hand-portables are integrated with higher power vehicular mobiles. In these circumstances the use of additional receiver-only fill-in stations is an economic and satisfactory proposition.

Base transmitter control poses a problem much more difficult to resolve. The selection of a base transmitter to communicate with a mobile whose precise whereabouts may be unknown, the broadcasting of messages to all, or groups of, mobiles whose locations may be widespread, and the provision of talk-through between mobiles are all facilities required on major systems, and difficult to provide satisfactorily. Apart from trunking, which can be economic only on very large networks, there are three traditional methods of operating the base transmitters on wide area schemes: manual selection, automatic selection and simultaneous transmission from more than one transmitter.

Manual transmitter selection

On many systems the despatchers select the transmitters manually. It is the simplest and least expensive method but has serious disadvantages:

1. Making the selection may entail trying a number of transmitters before sending the message, increasing the operator's work load and wasting air time.
2. Mobiles outside the service area of the selected transmitter may call, interrupting an existing conversation, either because they are unaware that the system is engaged or have received a poor signal that they believe may have been intended for them. Transmitting bursts of 'engaged' pips sequentially over the unselected transmitters during pauses in the despatcher's speech alleviates the first situation; the use of selective calling overcomes the second.
3. Broadcast messages must be transmitted on each transmitter in turn and talk-through between mobiles which are not in the service area of the same transmitter is not practicable.

Automatic transmitter selection

Selecting the transmitters automatically, or semi-automatically, is an improvement over manual selection. On a system where the mobiles are not equipped with selective calling and automatic acknowledgement of a call, automatic selection of transmitters can only occur on receipt of a call from a mobile. The transmitter through which to reply is then selected by the receiver voting system at the same time as it selects the signal to present to the despatcher. The selection is made at the start of the call and, because of the switching times involved, the transmitter selected is retained for the duration of the call. On these open channel systems, a calling despatcher must initially manually select a transmitter.

On systems where selective calling and auto-acknowledgement of a call are provided, the system can be made fully automatic by transmitting the data corresponding to a mobile's call sign from each transmitter in turn until a satisfactory acknowledgement is received. The successful transmitter is then retained for the duration of the conversation and, at its conclusion, is usually the one used to commence another call.

Such a system overcomes the disadvantage of the need to manually select the transmitter but the problems with mobiles outside the service area of the selected transmitter, and of broadcasting and talk-through, remain.

Simultaneous transmission

Operationally, simultaneous transmission from all sites is ideal and, under various names such as Spaced Carrier, Simulcast and Quasi-synch., has been around since the mid-1940s. Its operational value is proven but systems require special care in their planning, adjustment and subsequent maintenance.

An early form of simultaneous transmission was the amplitude modulated spaced carrier system. Used very successfully at VHF on systems using 25 kHz channel separation, the transmitter carrier frequencies were separated by 7 kHz – above the mobile receivers' audio pass-band. With the reduction of channel spacings to 12.5 kHz, spaced carrier operation on this basis was no longer possible and alternatives are to either synchronize, or very nearly synchronize, the carrier frequencies. There are, however, undesirable effects of synchronous and quasi-synchronous transmission but, with care, these can be reduced to an acceptable level. They are:

1. The beat note between transmitters being audible in the mobile receiver.
2. Variations in signal level due to interference patterns between signals from more than one transmitter.
3. Distortion due to audio phase differences and differing modulation levels when signals of comparable strength are received from more than one transmitter.

The beat note is easily dealt with. It is rendered unobjectionable either by placing it outside the mobile receiver audio pass-band – which is the usual method – or by synchronizing the transmitter carrier frequencies so that no beat note is produced. Synchronization, however, raises other problems, which are particularly severe at VHF but less so at UHF.

The spaced carrier system placed the beat note above the receiver pass-band, but modern systems place it below. This means that the beat note between the lowest and highest carrier frequencies must be below about 150 Hz if it is to be unobtrusive. Tests have shown the optimum carrier separation to be from 0.5 Hz to 4 Hz between any two transmitters using amplitude modulation, and from 5 Hz to 40 Hz between adjacent frequency transmitters using angle modulation.

Two transmitters on the same or closely-spaced frequencies produce deep nulls in the received signal in the areas where they provide almost equal strength signals. This is a natural phenomenon and the effect can only be reduced by the correct siting of stations and antenna configurations. Because of the longer wavelength the effect is more detrimental at VHF than UHF. Where the frequencies are quasi-synchronous the interference pattern is continually moving, but with synchronized carriers the pattern is virtually stationary and at low band VHF wavelengths it is possible to stop a vehicle in a place of semi-permanent zero signal. While moving slowly in an area of equal signal strengths from two transmitters, the cancellations become very objectionable. At 450 MHz the distance between the nulls is so short that it is almost impossible to remain in one, and while moving they are unnoticed. Strong signals in the overlap areas minimize the time that the signals fall below the receiver noise threshold at each cancellation. They are the key to reducing the annoyance from the cancellations. Amplitude modulated systems have an advantage in that the receivers produce less severe bursts of noise during the signal nulls.

The audio distortion, provided the equipment does not introduce significant additional harmonic distortion, is attributable to audio phase differences in the signals received from more than one transmitter. For the distortion to be severe, the received signals must differ by less than 6 dB on an angle modulated system; capture effect in the receiver removes the audible effects at greater differences.

The phase differences arise from differing audio characteristics in all the circuits in the path including land lines, radio links and control equipment. Common to all systems are the phase differences due to the different path lengths from the control centre to a receiving mobile, but where land lines or multiplexed circuits are used for the control of the base stations, variable bulk and group delays and frequency translation errors can present serious difficulties. Tests have shown AM to be slightly more tolerant than angle modulation in respect of this type of distortion, phase delays of 100 μs being acceptable with AM compared with 70 μs with FM.

Differing path lengths between the control centre to the transmitting sites can be equalized by installing audio delay circuits at the transmitters. The path lengths from each transmitting site to all the places where equal signals occur must then be less than about 21 km (equivalent to 70 μs) for acceptable quality. Modern techniques enable delays to be dynamically equalized to compensate for variations in the path.

Figure 15.6 shows the layout of a multi-station scheme including the audio delays. However, signals do not confine themselves to neat

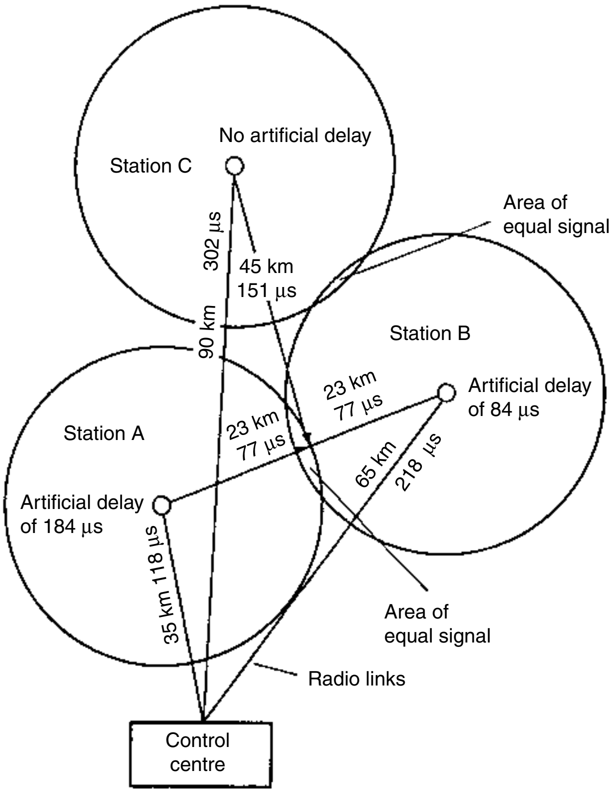
No artificial delay
Station C
302 μs
90 km
45 km
151 μs
Area of
equal signal
Station B
Station A
23 km
77 μs
23 km
77 μs
Artificial delay
of 84 μs
65 km
218 μs
Artificial delay
of 184 μs
35 km 118 μs
Area of
equal signal
Radio links
Control
centre

Figure 15.6 Quasi-synchronous transmitter system

circles and the worst situation is where the signals from two transmitters arrive at a mobile more or less equal in strength and phase. In this situation, apart from the ripple caused by the carrier frequency offset of the transmitters, a satisfactory signal would still be received but, occasionally, the presence of a weaker signal from a third, distant station (C in *Figure 15.6*) intrudes during the cancellation periods and, because of its long path length, introduces severe distortion. The area where this situation occurs is often small and the areas of overlap can usually be moved slightly by adjustment of either the transmitter power or antenna directivity.

16 Signalling

Many radio receivers are required to respond to instructions sent to them over the radio channel to which they are tuned. Several methods of signalling these instructions are in use, from the continuous tones used for controlling the receiver mute operation to the rapid data to initiate channel changes and transmitter power control on cellular radio-telephones.

16.1 Sub-audio signalling

The slowest signalling system uses continuous sub-audio frequency tones. Known as continuous tone controlled signalling (previously squelch) system (CTCSS) its performance, in the UK, is specified in Radiocommunications Agency specification MPT 1306. The most common use for the system is to control receiver mute opening. Permitting a mute to open only on receipt of an authorized signal, its use enables privacy between users to be maintained on shared systems, common base station systems for example, and reduces the annoyance factor from interference in the absence of a signal. Thirty-two tones are permitted and assigned by the Radiocommunications Agency (*Table 16.1*).

Table 16.1 CTCSS tones and modulation levels

67.0	110.9	146.2	192.8
71.9	114.8	151.4	203.5
77.0	118.8	156.7	210.7
82.5	123.0	162.2	218.1
88.5	127.3	167.9	225.7
94.8	131.8	173.8	233.6
103.5	136.5	179.9	241.8
107.2	141.3	186.2	250.3

The tones are transmitted at very low modulation levels (*Table 16.2*).

Table 16.2 Modulation levels

System channel spacing	*Amplitude modulation Modulation depth (%)*	*Angle modulation Peak deviation ± Hz*
25 kHz	10 to 20	400 to 800
12.5	10 to 20	200 to 400

16.2 In-band tone and digital signalling

16.2.1 Dual tone multi-frequency (DTMF)

Developed for telephone work, dual tone multi-frequency signalling has crept into some radio systems but is slow to operate compared with other in-band systems. Two tones are transmitted simultaneously to represent each digit. In the UK they conform to the MF4 code (*Table 16.3*).

Table 16.3 MF4 code of DTMF tones

Digit	*Frequencies (Hz)*	
1	697	1209
2	697	1336
3	697	1477
4	770	1209
5	770	1336
6	770	1477
7	852	1209
8	852	1336
9	852	1477
0	941	1336
*	941	1209
#	941	1477

Some push button telephones signal to the local exchange by generating a combination of two frequencies. Each row and each column of the push button keypad is connected to an oscillator of set frequency. When any push button is pressed, tones corresponding to its row and column frequencies are therefore generated. The row and column oscillator frequencies are as shown below:

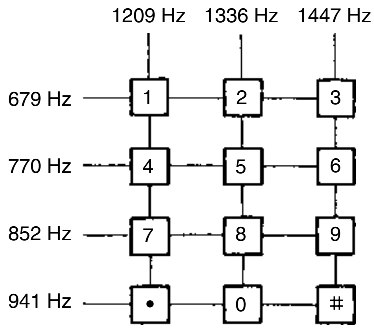

16.2.2 5-tone signalling

So-called 5-tone signalling uses a series of tones – usually five but other quantities are used in some systems. The system is popular for selective calling and addressing purposes (*Table 16.4*). There are a range of tones established by various organizations but, although the tones in each series may be assigned to each digit, equipment manufacturers sometimes amend the code and the usage. The relevant Radiocommunications Agency document is MPT 1316, *Code of Practice, Selective Signalling for use in the Private Mobile Radio Services*.

Table 16.4 5-tone codes

Standard	*ZVEI*	*DZVEI*	*EIA*	*CCIR/ EEA*	*CCITT*	*EURO*
Tone duration (ms)	70	70	33	100/ 40	100	100
Pause duration (ms)	0	0	0	0	0	0
Code/digit	Tone frequency (Hz)					
0	2400	970	600	1981	400	980
1	1060	1060	741	1124	697	903
2	1160	1160	882	1197	770	833
3	1270	1270	1023	1275	852	767
4	1400	1400	1164	1358	941	707
5	1530	1530	1305	1446	1209	652
6	1670	1670	1446	1540	1335	601
7	1830	1830	1587	1640	1477	554
8	2000	2000	1728	1747	1633	511
9	2200	2200	1869	1860	1800	471
Repeat	2600	2400	459	2110	2300	1063
Alarm	2800	2600		2400		
Free-tone						1153
Group tone	970	825	2151	1055		

ZVEI	Zuverein der Electronisches Industrie. Designed to operate on 20 kHz channel spaced systems. On 12.5 kHz channel spaced systems transmission of 2800 Hz creates difficulty and depressed ZVEI (DZVEI) was adopted and recommended in MPT 1316.
EIA	Electrical Industries Association.
CCIR	Committee Consultatif International Radio Communication. The longer tone duration offers robustness against corruption but is slow to operate. Originally designed for marine use, each digit was transmitted twice to ensure reliability.
EEA	Electronic Engineering Association. Recommended in MPT 1316.
CCITT	International Telegraph and Telephone Consultative Committee.

16.3 Digital signalling

FFSK

Many analogue modulated systems use digital signalling in the form of 1200 bit/s fast frequency shift keying which is faster than 5-tone but less robust.

DCS

Digitally coded squelch (DCS) is a digital, but slower, alternative to CTCSS. It is not in current use in the UK. DCS continuously repeats the constant binary bit pattern of a 23 bit word. Golay error coding reduces the number of data bits to 12 of which 3 initiate the Golay sequence leaving 9 information bits, i.e. a 9 bit word. With a 9 bit word, $2^9 = 512$ codes are possible but, because of problems with the frequencies produced, the codes with 7 or more 1s or 0s in succession are discarded leaving 104 ‘clean codes’. The transmitted rate is 134 bit/s and the highest fundamental frequency is therefore 67 Hz. The lowest fundamental frequency is produced when 6 zeros follow 6 ones = $1/6$th of the fundamental frequency = 11.7 Hz. The important harmonics are:

$$\begin{aligned} \text{third} &= 201\,\text{Hz} \\ \text{fifth} &= 335\,\text{Hz} \\ \text{seventh} &= 469\,\text{Hz} \end{aligned}$$

The seventh harmonic is well within the audio pass band of most receivers but at 3 octaves from the fundamental is relatively weak.

16.4 Standard PSTN tones

	Signal			
	UK		*USA*	
Tone	*Frequency*	*Sequence*	*Frequency*	*Sequence*
Number unobtainable	400 Hz	Continuous	500 + 600 Hz	Interrupted every second
Number busy	400 Hz	0.75 s on, 0.75 s off	480 + 620 Hz	Interrupted once per second
System busy	400 Hz	0.4 s on, 0.35 s off, 0.225 s on, 0.525 s off	480 + 620 Hz	Interrupted twice per second
Dial tone	50 Hz 33.3 Hz		350 + 440 Hz	
Ringing tone	400 Hz, or 400 + 450 Hz, modulated by 17, 25, or 50 Hz	0.4 s on, 0.2 s off, 0.4 s on, 2 s off	440 + 480 Hz	Interrupted once per second
Pay tone	400 Hz	0.125 s on, 0.125 s off		
Time tones	900 Hz	0.15 s on, 0.85 s off. Three times every three minutes		
Recorder warning			1400 Hz	One burst, every 15 s

References

Belcher, R. *et al*. (1989). *Newnes Mobile Radio Servicing Handbook*. Butterworth-Heinemann, Oxford.

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

17 Channel occupancy, availability and trunking

17.1 Channel occupancy and availability

The usefulness of any communications network depends on successful calls. Assuming that no calls fail for technical reasons, the success rate depends on the amount of traffic on the channel and how well it is managed. Channel occupancy is a measure of the amount of traffic carried and is affected by:

1. the number of subscribers (mobiles on a radio network)
2. the number of calls per subscriber in a given time
3. the average duration of each message.

The occupancy of any system varies hourly, daily and possibly seasonally. To be successful the network must be able to cope with the traffic during the periods of highest occupancy and, when considering occupancy it is usual to examine the conditions over one hour, the 'busy hour'. The occupancy of a channel can be expressed as a percentage:

$$\text{Channel occupancy} = \frac{nct}{36}\%$$

where

n = number of subscribers (on a mobile radio system, mobiles on watch)
c = call rate (number of calls per subscriber in 1 hour)
t = average message duration in seconds

Occupancy may also be expressed in erlangs where continuous traffic, i.e. 100% occupancy over a specified period, is 1 erlang. Less than 100% occupancy is expressed as a decimal fraction, e.g. 30% occupancy is 0.3 erlang. Demand may also be expressed in traffic units, A:

$$A = \frac{Ct}{T} \text{ erlangs}$$

where C is the number of calls in time T and t is the mean call duration.

The amount of congestion occurring on a system can be defined by the grade of service it offers. Grade of service is expressed as a percentage in terms of the probability of meeting either congestion or waiting more than a specified delay before a call can be established. Erlang's formula is:

$$\text{Grade of service} = \frac{A^N/N!}{1 + A + A^2/2! + A^3/3! + \cdots + A/N!}$$

where A is the number of traffic units (erlangs) and N is the number of outlets (trunked channels).

The probability that delayed calls will have to wait more than a specified time ($W > t$) is:

$$P(W > t) = \exp\left(-\frac{(C - A)t}{H}\right)$$

where
A = traffic units (erlangs)
C = number of trunked channels
H = mean call duration in busy hour

The above assumes that:

1. the number of users is infinite
2. the intervals between calls are random
3. call durations are random
4. call set-up times are negligible
5. delayed calls are queued so that they are dealt with on a first come, first served basis.

The number of potential radio users is continuously growing and any technique that makes more effective use of the finite radio spectrum is welcome. Whenever technology has permitted, higher frequencies and reduced channel spacings have been employed in the attempt to meet the demand. A method of making more effective use of the available spectrum is trunking.

17.2 Trunking

The generally accepted term for the automatic allocation of communication channels to subscribers on demand is trunking, and a 'trunked radio system' is one in which many users share the use of a common pool of radio channels. Channels from the pool are allocated to users on demand and as they become unoccupied; no channels are allocated to specific users or groups of users.

When making a telephone call, provided the speech quality is good it is immaterial which route the call takes. The same applies to radio; providing the message reaches the right person quickly, the radio frequency which carries it does not matter.

Most trunked radio systems use a calling channel on which calls are requested. When a call is established and accepted a working channel from the pool is allocated for the duration of the call and at its termination the channel is returned to the pool for re-allocation to another user. The important principle is that any user has access to any free channel. This has the advantage that more mobiles per channel can be accommodated, or a better grade of service, i.e. less waiting time for access, can be provided. Basic trunked systems are unsuitable where everyone using the system needs to be a party to every call but the latest trunked mobile radio system arrange, on receipt of a call, to assign a working channel to a particular group of mobiles where they operate within the service area of one base station. They then operate on an 'all informed' basis.

Trunking provides the most benefit when the number of channels required is greater than the number assigned to the system. The effective gain in the number of mobiles that can be accommodated on a trunked system is illustrated by the following example:

> *A single radio channel with an efficient signalling system offering a 30% grade of service can accommodate 90 mobiles with a mean waiting time of 20 seconds if each mobile makes 1 call. Twenty non-trunked channels cater for 1800 mobiles. A 20 channel trunked system under the same conditions can accommodate 3430 mobiles with a mean waiting time of 16.4 seconds, a trunking gain of over 170 mobiles per channel.*

The mobile transmitter/receivers used on trunked systems are more complex than those used on systems where channels are permanently allocated because the mobiles must either scan the channels searching for an available one or one containing a call directed to them, or listen continuously on a channel designated for calling. In the latter case when a call is initiated, a complex 'handshake' procedure occurs and the mobile is automatically switched to the channel allocated for that conversation. The band III public access mobile radio (PAMR) networks use a calling channel and many common base station (CBS) systems use channel scanning methods with no calling channel.

Working channels may be assigned for the duration of conversation, which is message trunking, or for the duration of a transmission, which is transmission trunking. With either type of trunking the channel is not returned to the pool immediately the signal from the mobile is lost; some waiting time is essential to cater for short-term signal dropouts but with message trunking the channel is retained for a longer period of time to allow the complete conversation to take place. On a very busy system this is wasteful and, on a system with very rapid signalling, transmission trunking may be used. Then the use of the working channel is lost very soon after the cessation of transmission and a new working channel assigned on its resumption. Some networks automatically migrate from message trunking to transmission as traffic increases.

17.3 In-band interrupted scan (IBIS) trunking

IBIS trunking is a method which is finding favour for common base station applications. All channels are used for communication with no dedicated calling channel, thus making the most effective use of the channels available.

On a CBS system all users access the base station, which operates talk-through, by radio. Each user is allocated his own continuously coded squelch system (CTCSS) tone to access the base station and to prevent overhearing by other users. A trunked common base station contains the transmitters and receivers for a number of radio channels and when the system is quiescent transmits bursts of an in-band (speech band) free-tone – 1953 Hz is one that is used – on one channel only. The mobiles, which are fitted with decoders to recognize both the free-tone and their CTCSS tone, scan all channels looking for one containing free-tone onto which they lock. Should the mobile operator press the transmit button the transmitter will only be enabled when the mobile receiver is on a channel containing either the free-tone or its CTCSS code. As soon as the channel becomes occupied the base station commences to radiate free-tone on another channel and mobiles of other user groups re-lock onto the new free channel. When either a mobile or the control station, which is a fixed mobile, makes a call it is on the channel radiating free-tone and the transmission contains the CTCSS code allocated to that user group. The repeater re-transmits the CTCSS code which activates the calling mobile and all other mobiles within that group. Mobiles of other user groups sensing the alien CTCSS tone leave the channel and continue scanning for a free channel. If all channels are engaged, the caller hears a ‘busy’ tone and must try again later.

17.4 Trunking to MPT 1327 specification

Specification MPT 1327 produced by the Radiocommunications Agency lays down a signalling standard for trunked private land mobile radio systems. The system uses fast frequency shift keying (FFSK) at 1200 bit/s with time division random access. The mobiles operate two-frequency semi-duplex and the base stations or trunking system controller (TSC) operate full duplex. A calling channel is used for call establishment but may be either dedicated or nondedicated. After a call is established messages are transmitted over a traffic (working) channel. The basic components of the signalling format, in sequence, are:

1. The link establishment time (LET), a period of transmission of undefined modulation.
2. A preamble comprising a sequence of bit reversals 1010...10, minimum 16 bits and ending with a binary zero for the receiver data demodulator to achieve synchronization.
3. A message. A contiguous transmission of a synchronization sequence, address codeword and, where appropriate, data codewords.
4. A hangover bit, H, of either binary zero or binary one, added to the last transmitted message identifying the end of signalling transmission.

The message synchronization sequence for both control and message channels consists of 16 bits, the first bit being a binary one on a control channel and binary zero on a traffic channel.

Messages, whether address or data, are transmitted in 64 bit codewords. The first bit, bit 1, indicates the status of the codeword. Binary one denotes an address word, binary zero, a message word. Bits 2–48 are the information field and bits 49–64 are check bits.

References

Macario, R.C.V. (1991). *Personal and Mobile Radio Systems*. Peter Peregrinus, London.

Parsons, J.D. and Gardiner, J.G. (1989). *Mobile Communication Systems*. Blackie, London and Glasgow.

Radiocommunications Agency (1991). MPT 1327, *A Signalling Standard for Trunked Private Land Mobile Radio Systems*.

Radiocommunications Agency (1993). MPT 1318, *Engineering Memorandum, Trunked Systems in the Land Mobile Radio Service*.

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

18 Mobile radio systems

18.1 Paging

Where an immediate reply to a message is not essential, paging systems, which are basically a one-way path from controller to mobile, are an economic option. Paging is also spectrally economic, only one frequency being required and messages being of very short duration. The recipient of a call may be merely alerted by means of a tone emitted by the paging receiver or, if silent alert is required, a vibration. Alternatively, pagers displaying, and storing, alpha-numeric messages are readily available. Message handling services are also offered on some paging systems.

On-site paging systems are usually owned or rented by the user and consist of either a radio base station or, in the case of a non-radio system, an inductive loop around the perimeter of the service area. The details of on-site paging systems and the Radiocommunications Agency specifications governing them are:

RF frequencies:

on-site, 26 MHz and 31 MHz bands, MPT 1365
49.0–49.5 MHz, MPT 1335
160–161 MHz
458–459 MHz, MPT 1305

Channel spacings:

12.5 kHz in the VHF bands and 25 kHz in UHF band

Modulation:

FM, FFSK data

Wide area, national or international systems are generally owned by a network provider to whom the user subscribes for the service. There are frequencies allocated in the high VHF and UHF bands. The appropriate specifications are MPT 1308 for receivers and MPT 1325 for transmitters.

The Post Office Code Standardization Group (POCSAG) standard which has been adopted by the CCITT allows the transmission of messages at a rate of 10 calls per second and a capacity of 1 million pagers with up to four alternative addresses per pager. The recommended bit speed is 512 bit/s with direct FSK deviating the transmitter frequency by ±4.5 kHz. Positive deviation indicates a binary zero and negative deviation a binary one.

In America the Golay sequential code is widely used for paging addressing. The format for tone-only paging consists of 12 address bits followed by 11 parity bits for each 23 bit word. Two bit rates are used, 300 bit/s for address codewords and 600 bit/s for message codewords.

There are European paging and message systems in operation including Euromessage and the frequency agile European Radio Message Service (ERMES).

18.2 Cordless telephones

CT1

The first generation of cordless telephone, still very popular in the UK, is designed to serve the domestic environment with a range of 100 m. The base station transmits on one of eight frequencies between 1.6 MHz and 1.8 MHz, and the handset on one of eight frequencies in the 47 MHz band. Frequency modulation is used with a deviation of ±4 kHz at the base station and 2.5 kHz at the handset. MPT 1322 applies and permits an erp of 10 mW at both the base station and the handset.

CT2

The second generation of cordless telephone. The services, under various names, telepoint, phonezone, etc. were planned to provide for the general public a lower cost alternative to the cellular radiotelephone networks which were seen at the time as a businessman's preserve. However, the low market uptake caused their demise in the UK although services are thriving in some European and Far Eastern countries.

One hand-portable transmitter/receiver can operate to a local base station installed in the home or office, or through one of a number of multi-channel base stations with a range of approximately 200 m installed in public places. While away from the local base station the subscriber must initiate a call: calls cannot be made from the PSTN to a CT2 subscriber's handset.

The operational frequencies are in the band 864.1 to 868.1 MHz and employ time division multiple access. Speech is digitized at 32 kbit/s, stored, and then transmitted at 64 kbit/s in 1 ms slots. This leaves the alternate 1 ms slots available for the digitized and stored speech of a reply. Duplex operation is achieved in this way on a single radio frequency.

Inter-operability between networks was to be ensured by specification MPT 1375 *Common Air Interface*.

Digital European cordless telephone (DECT)

A pan-European system complying with an ETSI standard, DECT operates in the 1880 to 1900 MHz band. It offers data handling facilities and the ability for a subscriber to receive calls while away from the local base station. The techniques are similar to those used for GSM although, because the mobile is virtually stationary, the constraints on data transmission are less severe and no hand-off is required. The 20 MHz RF bandwidth is divided into 13 carriers spaced at 1.7 MHz intervals, each carrier containing 12 TDMA channels with GMSK modulation.

CT3 is developed from DECT and operates in the band 800 to 900 MHz. Each 8 MHz section of that band is divided into 1 MHz blocks, each containing sixteen 1 ms time slots.

18.3 Trunked radio

Bank III trunked radio

So-called because it occupies part of the now redundant television band III, 174 to 225 MHz, and often referred to as private mobile radio, this is a subscriber access system. The network has almost nationwide coverage in the UK, arranged on both local and regional bases. The network permits radio communication between a fixed office station and mobiles, mobile to mobile, and limited access from mobiles to the PSTN.

The relevant Radiocommunications Agency specifications are:

MPT 1323	*Angle-modulated Radio Equipment for Use at Base and Mobile Stations Private Mobile Radio Service Operating in the Frequency Band 174–225 MHz.*
MPT 1327	*A Signalling Standard for Trunked Private Land Mobile Radio Systems.*
MPT 1343	*System Interface Specification for Radio Units to be Used with Commercial Trunked Networks Operating in Band III sub-band 2.*
MPT 1347	*Radio Interference Specification for Commercial Trunked Networks Operating in Band III sub-band 2.*

MPT 1352 *Test Schedule. For Approval of Radio Units to be Used with Commercial Trunked Networks Operating in Band III sub-band 2.*

Frequency modulation is used with FFSK signalling, a channel spacing of 12.5 kHz and TDMA techniques (Section 17.4).

Trans-European trunked radio (TETRA)

A new European standard for digital trunking. The system is aimed at PMR users and will employ TDMA with four users sharing a 25 kHz radio channel with the option of two users on a 12.5 kHz channel. Speech will be digitally encoded at 4.8 kbit/s and transmitted at 36 kbit/s (gross bit rate including error correction overhead) using $\pi/4$ differential quaternary phase shift keying (DQPSK).

Digital short range radio (DSRR)

This utilizes trunking techniques but with no infra-structure of control or switching centres; mobiles work to mobiles via a locally installed repeater. It offers great potential for short-term hire for local events. The band 933 to 935 MHz has been allocated to the service and accommodates 76 traffic and 2 control channels separated by 25 kHz. The operating procedure may be either single-frequency simplex or two-frequency semi-duplex. Mobiles remain on standby waiting to receive their selective signalling code (SSC) on one of the two control channels; mobiles with even serial numbers normally listen on control channel 26 and those with odd serial numbers on control channel 27. A mobile wishing to initiate a call first scans for a free traffic channel and, having obtained one, selectively addresses the intended recipients of the call over the relevant control channel. The recipient mobiles are then instructed automatically by the system to go to the appropriate traffic channel.

The modulation is GMSK at 4 kbit/s for addresses and 16 kbit/s for speech and data messages. The maximum transmitter erp is 4 W.

18.4 Analogue cellular radio-telephone networks

Analogue networks are becoming obsolete and are being replaced by digital cellular networks. There are no analogue networks operating in the UK and the frequency bands that they used to occupy are now reassigned to GSM operators.

However some countries still have an analogue service based on AMPS or TACS. These provide national coverage with roaming facility and hand-off from cell to cell as a mobile travels across the network, calling for more complex signalling than local trunking.

The control system must know at all times the location of all operational mobiles. It does this by continuously monitoring signal strength and instructing the mobile to change channel as it crosses cell boundaries.

To permit the maximum re-use of frequencies mobiles are instructed by the system controller to reduce transmitted power to the minimum necessary for acceptable communications in either eight or five steps, depending on the mobile classification, to 0.01 W. Mobiles are classified by their maximum effective radiated power output:

Class 1	10 W erp, vehicle mounted
Class 2	4 W erp, transportable
Class 3	1.6 W erp, transportable
Class 4	0.6 W erp, hand-portable

The channel separation is 25 kHz with a peak speech deviation of ±9.5 kHz. Signalling is by Manchester encoded PSK at 8 kbit/s with 6.4 kHz deviation.

To discriminate between stations transmitting on the same RF channel an FM supervisory audio tone (SAT) is transmitted concurrently with speech at ±1.7 kHz deviation. There are three such tones: 5970, 6000 and 6030 Hz.

An FM signalling tone (ST) of 8 kHz transmitted for 1.8 seconds (deviation = ±6.4 kHz) indicates the clear-down of a conventional telephone call.

The advanced mobile phone service (AMPS) is the American cellular system designed for 30 kHz channel spacing with a 10 kbit/s signalling rate. TACS is based on the AMPS system but modified for reduced channel spacing. Narrower channel spacing systems (NAMPS and NTACS) are also found in some countries.

18.5 Global system mobile

Global system mobile (GSM), formerly known as Groupe Speciale Mobile, is the pan-European digital cellular radio-telephone service. The operational requirements for GSM are severe, e.g. it must operate satisfactorily to a person walking, or in a slowly moving vehicle, in a street where much of the furnishings will introduce multi-path fading, and operate to a train travelling at 250 km/hr where Doppler frequency shift becomes significant. To reduce the corruption, a high degree of error detection and correction must be applied which increases the occupied bandwidth. To compensate, to some extent, for this, the system takes full advantage of the redundancy in speech to reduce the bandwidth during synthesis.

GSM operates full duplex in the band 890–915 MHz, up-link and 935–960 MHz, down-link.

A combination of FDMA and TDMA is employed. Each allocated band of 25 MHz is divided into 125 carriers spaced 200 kHz apart. The subdivision of each transmitted bit stream into 8 TDM time slots of 540 μs gives 8 channels per carrier and 1000 channels overall. Modulation is Gaussian minimum shift keying (GMSK).

Speech synthesis is by speech codec using linear predictive coding. It produces toll quality speech at 13 kbit/s.

18.5.1 Emergency locator service

The FCC in the US has placed a requirement on mobile phone operators to provide an emergency location service, referred to as E911. The aim is to locate people who need emergency services, but are not able to indicate their location either through injury or through lack of local knowledge. The locator service may also deter hoax calls (or help to catch people making hoax calls) and could have commercial applications in providing directions to businesses such as garages and restaurants.

There are two options: handset based solutions or network-based solutions. The handset-based solutions generally rely on signals picked up from the global positioning system (GPS). Network-based solutions rely on the timing difference in the arrival of signals from neighbouring cell sites. In network based solutions, the E911 requirement is for the mobile phone to be located within 100 metres at least 67% of the time and to within 300 metres in 95% of cases. In handset-based solutions, the location distances are 50 metres and 100 metres, respectively.

18.5.2 Data over GSM

Standard GSM data service allows a maximum of 9600 baud data rate transmission. An enhanced data rate of 14.4 kbps is also possible with some terminals.

The data link can be either synchronous or asynchronous and the data is conveyed through the network using an ISDN-based data protocol, the ITU-T standard V.110 rate adaptation scheme. Asynchronous transmission requires start and stop bits to be added, so the net data rate is reduced to 7680 bits per second (bps). Synchronous transmission potentially allows a faster data rate, but error checking is employed that requests re-transmission of data when errors are identified and this can produce a lower overall data throughput.

18.5.3 High-speed circuit switched data (HSCSD)

HSCSD allows higher data rates by combining up to four 14.4 kbps data channels together. A maximum data rate of 57.6 kbps is possible.

18.5.4 General packet radio service (GPRS)

GPRS combines up to eight 14.4 kbps channels to give data rates of up to 115.2 kbps. At these rates, Internet browsing is possible. GPRS supports X.25 and IP based transmission, which includes wireless application protocol (WAP) data.

18.6 Other digital mobile systems

Second generation mobile systems are digital systems. The GSM system has been described in Section 18.5. This section describes the GSM1800 system, the personal communications system (PCS1900, sometimes known as personal communications network or PCN) and third generation systems.

18.6.1 GSM1800 (DCS1800)

The GSM1800 system operates in the 1840–1880 MHz band, with a carrier spacing of 200 kHz. Each carrier gives multiple access through time division multiple access (TDMA). It is a lower power version of GSM900, which operates in the 900 MHz band. A GSM1800 handset has a maximum power output level of 1 watt, rather than the 2 watts of a GSM900 handset.

Voice coding can be either: half-rate using VSELP at 5.6 kbps, full-rate using RPE-LTP at 13 kbps, or enhanced full rate (EFR) using ACELP at 12.2 kbps. Convolutional coding is used at a rate of $\frac{1}{2}$ to reduce the effects of noise and errors on the radio channel. The radio carrier modulation is Gaussian minimum shift keying (GMSK).

18.6.2 PCS1900 (IS-136, or PCN)

The personal communications system (PCS) is based on GSM and used time division multiplexing to provide separate channels over each radio frequency. Convolutional coding is used at a rate of $\frac{1}{2}$ to reduce the effects of noise and errors on the radio channel. Each carrier is spaced 30 kHz apart and offset quadrature phase-shift keying (OQPSK) is used to modulate the carrier.

The voice-coder (vocoder) used by PCS1900 is either full-rate using VSELP at 7.95 kbps or enhanced full rate (EFR) that uses ACELP at 7.4 kbps. The EFR scheme was developed for use in PCS1900 and has since been adopted by both GSM900 and GSM1800 networks because it improves the audio quality. Previously, GSM used either a half-rate (HR) or full-rate (FR) vocoder.

18.6.3 Third generation systems

The so-called 3G systems have been described using a number of terms: UMTS, IMT-2000, cdma-2000, w-cdma, etc. There are variations between the proposed standards, but in general they all use code division multiple access (CDMA), convolutional coding, quadrature phase shift keying and time or frequency division duplexing.

Third generation systems will operate in the 1885–2025 MHz band and in the 2110–2200 MHz band. Initially the lower band between 1980 and 2010 MHz will not be allocated, nor will the upper band between 2170 and 2200 MHz. However, the full bands will be available when existing services in these bands can be moved (such as PCS1900, PHS and DECT).

The code division multiple access (CDMA) modulation scheme allows several channels to occupy the same frequency band. CDMA is a form of spread spectrum encoding. The digitized speech signals are coded using a unique code, known as a spreading code, so that one bit of speech code produces many encoded bits. When demodulated and mixed with the same spreading code, the original data is recovered. Receivers that use a different spreading code are unable to recover this data, but they are able to recover data intended for them.

18.7 Private mobile radio (PMR)

These systems are privately owned or rented and may be used only by the owner's own staff for the purpose of the owner's business. The leasing of air-time is forbidden. Common base station or community repeater systems where several users subscribe to, and use, the station is often also referred to as PMR.

18.8 UK CB radio

27 MHz band: 27.60125 to 27.99125 MHz

40 channels at 10 kHz spacing.
Max. erp, 2 W; max. transmitter output power, 4 W.
Antenna: single rod or wire, 1.5 m overall length, base loaded.
If mounted higher than 7 m transmitter output to be reduced at least 10 dB.
Modulation: FM only, deviation ±2.5 kHz max.

934 MHz band: 934.025 to 934.975 MHz

20 channels at 50 kHz spacing (may be reduced later to 25 kHz).
If synthesizer used spacing may be 25 kHz on precise channel frequencies specified.
Max. erp, 25 W; max. transmitter output power 8 W. If antenna integral, max. erp, 3 W.
Antenna: may have up to four elements, none exceeding 17 cm.
If mounted higher than 10 m, transmitter output to be reduced at least 10 dB.
Modulation: FM only, deviation ±5.0 kHz max.

Spurious emissions: for both bands, not exceeding 0.25 μW except for specified frequency bands where the limit is 50 n W.

For latest full specifications see publications MPT 1320 (27 MHz band) and MPT 1321 (934 MHz band) obtainable from the Information and Library Service of the Radiocommunications Agency.

References

Belcher, R. *et al.* (1989). *Newnes Mobile Radio Servicing Handbook*. Butterworth-Heinemann, Oxford.
Macario, R.C.V. (1991). *Personal and Mobile Radio Systems*. Peter Peregrinus, London.
Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

19 Base station site management

The information in this chapter applies mainly to sites for stations operating at VHF and above. Some conditions will be different for stations working at lower frequencies due to the generally physically larger transmitters and antennas.

19.1 Base station objectives

The objectives of a radio base station are:

1. To provide an adequate service over the whole of the desired service area in a well-defined manner. Special consideration must be given to those areas where, because of their location, weak signals or a high ambient electrical noise level might be expected.
2. To create the minimum interference with other radio users and be designed to receive the minimum from other users.
3. To be electromagnetically compatible with neighbouring installations, possibly on the same site.
4. To create the minimum impact on the environment.

19.2 Site ownership or accommodation rental?

The question of whether to privately develop a base station site or rent space on someone else's site may be resolved by the planning authorities rather than the developer. Before planning consent will be given the developer must normally prove that no existing site will meet the requirements. However, assuming that land is available and planning consent might be forthcoming, other factors to consider are costs, both initial and recurring, and access for services, materials during building and tower erection and for maintenance during bad weather (*Table 19.1*).

19.3 Choice of site

The first stage in selecting a site is to look for an existing site which might appear to cover the service area. Information may be available from the operators of such sites regarding their coverage, and coverage tests may be made or computer predictions obtained. For a

Table 19.1 Factors affecting the decision to own or rent

	Ownership	*Rental*
Initial costs	Legal charges Land clearance Provision of services Erection of buildings and mast or tower Installation of equipment Provision of standby power	Legal charges Connection fee Installation of equipment Provision of standby power
Recurring costs	Maintenance of site Power Insurance (other than radio equipment)	Rental Normally included in rental Normally included in rental
Other considerations	Very long commissioning time Site under user's control Space may be leased to other users creating revenue	Very short commissioning time Site under landlord's control Accommodation will be shared with other users (install equipment in locked cabinet to preserve security)

receiving site the ambient noise level at the base station must also be considered. It should be realized that a site, or an antenna, can be too high, particularly when coverage is required at very short range or if interference from other users is likely to occur. Antennas radiate the minimum signal along the line of their elements so that vertically polarized antennas, i.e. with the E plane vertical, provide very little, if any, signal immediately below them. This creates a hole in the coverage, 'the Polo mint effect', close to the antenna.

Where there is capacity on an existing site the possibilities are:

1. Sharing an existing lightly used radio channel, e.g. common base station.
2. Sharing equipment space in an existing building or providing own building or extension on a communal site.
3. Sharing an antenna by means of multi-coupling or erecting own antenna on existing structure.

19.4 Masts and towers

Apart from short, unguyed poles mounted on the side of buildings, antenna supports fall into two types: guyed masts and self-supporting towers.

A guyed mast may be a single pole, or a square or triangular section lattice structure. The range of single-pole masts includes those of about 4 m maximum height mounted on chimney stacks and secured by wire lashings, and poles supported on the ground or the roof of a building. The latter types may be up to 30 m high. They are usually constructed from steel or aluminium hollow tubular sections – the bottom section is invariably steel – with a set of guys at about every 5 m. Mast erection is specialist work but to estimate the approximate height that a site can accommodate, a four-guy, single-pole mast requires a minimum spacing for the anchorage points in the form of a square with sides equal to half the height of the mast. The guys may be of steel, possibly plastic-coated, or more likely of synthetic fibre. Steel guys may affect the directivity of antennas in close proximity, and if the guys corrode at their fixing points, intermodulation interference may result.

Self-supporting towers come in all heights up to many hundreds of feet and may be made of a variety of materials from steel sections to concrete.

The factors determining the type of structure to be erected are the number and types of antenna, and the site conditions.

Antenna considerations:

- Physical dimensions and space requirements.
- Weight.
- Wind loading.
- Directivity. For example, the beam width of a 6.5 GHz, 3.0 m dia. microwave dish is 0.9 degrees. A tower supporting such an antenna must twist, therefore, less than about 0.3 degrees in the highest wind likely to be experienced at the site.
- Access for riggers to antenna mounting positions.

Site conditions:

- Availability of space for tower footings, and guys and anchorages if a mast is contemplated.
- Stability and type of ground.
- Weather conditions. High winds. Accumulation of ice and snow. Build-up of ice increases wind loading of antenna – and of mast or tower – considerably. There is no point in using grid microwave antennas where icing is likely to occur.
- Aesthetic and planning consent considerations.

19.5 Installation of electronic equipment

For radio equipment, safety is one of the first priorities and in addition to the electrical considerations, physical hazards must be avoided.

Adequate working space must be provided around the equipment racks. Much equipment is now wall-mounted but tall, floor-standing cabinets must be firmly bolted down to prevent toppling when units are partially withdrawn for servicing and cables need to be routed safely. Overhead cable trays are generally considered the most satisfactory method but, to avoid interaction, it is recommended that cables for the various functions be segregated to reduce cross-talk and interference. Where cabling enters a cabinet at floor level, mounting the cabinet on a hollow plinth and running the cables through the plinth avoids damage to the cables and persons tripping over them.

For low voltage installations (see classifications below) the IEE wiring regulations apply. For high voltage installations, recommendations are given in BS 6701: part 1: 1990 which also covers accommodation, lighting levels and access arrangements.

Classification of installations:

- *Extra low voltage*. Normally not exceeding 50 V AC or 120 V DC.
- *Low voltage*. Exceeding 50 V AC or 120 V DC but not exceeding 1000 V AC or 1500 V DC between conductors, or 600 V AC or 900 V DC between conductors and earth.
- *High voltage*. Exceeding 1000 V AC or 1500 V DC between conductors, or 600 V AC or 900 V DC between conductors and earth.

19.6 Earthing and protection against lightning

Recommendations for earthing are given in BS 7430: 1991. This document covers the earthing of equipment and the principles of earthing for protection against lightning. All equipment metalwork must be bonded together and connected to the electricity supply earth point, the main earth terminal. In addition a connection must be made to the earthing system provided for protection against the effects of lightning.

Antenna systems by their nature are vulnerable to lightning strikes. Nearby taller structures may reduce the risk but precautions must still be taken. The zone of protection – a cone with its apex at the tip of the protecting structure and its base on the ground forming an angle of 45° to the perpendicular – does not necessarily protect structures above 20 m high. BS 6651: 1985 is the *Code of Practice for the Protection of Structures against Lightning*. Lightning protection begins at the top of the mast or tower where, ideally, the highest item should be a finial bonded to the tower. However, a finial mounted alongside an omnidirectional antenna will affect is radiation pattern. Grounded antennas are preferred and the outer conductor of each feeder cable must be bonded to the tower at the top and bottom of its vertical run. Grounding kits for the purpose are obtainable. The feeder cables must also be bonded to an earthing bar at their point of entry to the equipment building. The codes of practice should be consulted regarding the routes to be taken by earthing tape and the methods of jointing. An important point is that neither very sharp bends nor ‘U’ bends are created.

The legs of the tower must be bonded together with earthing tape and each leg connected to an earthing system which may consist of a buried plate or rods. Several rods may be necessary for each subsystem to attain the specified earth resistance of not more than 10 ohms. The rods should, except in rock, be driven into ground which is not made up or backfilled or likely to dry out. Where several rods are necessary to achieve the specified earth resistance these should be spaced well apart, the reduction in earth resistance being small with parallel rods closely spaced. Joints to facilitate testing of each subsystem separately must be provided either above ground or in a purpose-built pit or chamber. *Figure 19.1* shows a typical earthing system.

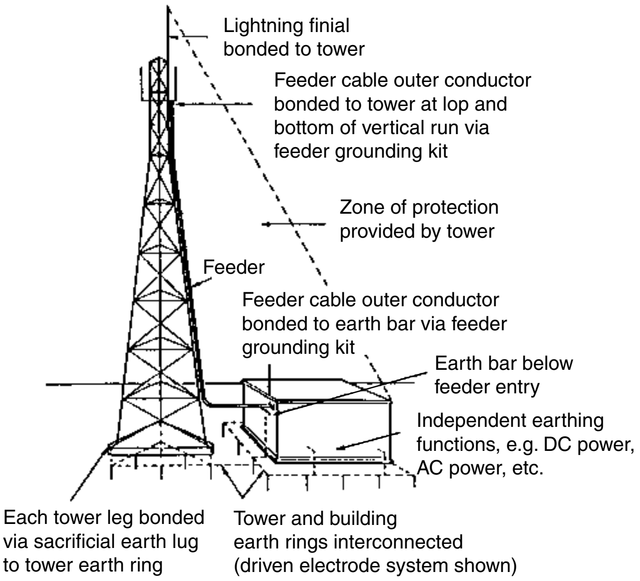
Lightning finial bonded to tower
Feeder cable outer conductor bonded to tower at lop and bottom of vertical run via feeder grounding kit
Zone of protection provided by tower
Feeder
Feeder cable outer conductor bonded to earth bar via feeder grounding kit
Earth bar below feeder entry
Independent earthing functions, e.g. DC power, AC power, etc.
Each tower leg bonded via sacrificial earth lug to tower earth ring
Tower and building earth rings interconnected (driven electrode system shown)

Figure 19.1 Typical example of good earthing practice (© Crown Copyright 1991, Radiocommunications Agency)

The system should be tested regularly, at least annually, and the results recorded in a lightning protection system log book. The test comprises:

1. Measurement of the resistance to earth of each termination network and each earth electrode.
2. Visual inspection of all conductors, bonds and joints or their measured electrical continuity.

BS 7430 recommends the method of testing.

Other vulnerable systems are overhead cables for either the electricity supply or telephone and control purposes.

19.7 Erection of antennas

The installation of the antenna probably influences the performance of a radio system more than any other part. The heights involved render supervision of the installation difficult and, although antennas are relatively inexpensive, labour charges raise the cost of their replacement, and bad weather may delay rigging work. It is therefore important that good quality antennas are specified and that the initial installation is correct. In addition to the physical mounting arrangements, the dispositions of antennas on the structure are important. Some aspects are detailed below and MPT 1331 contains more information.

19.7.1 Directivity

The radiation pattern diagrams provided in the antenna manufacturers' catalogues are produced by an isolated antenna. They can be greatly affected by the proximity of the support structure, e.g. if a dipole is mounted close to the leg of a tower, the leg behaves as a reflector and the radiation pattern of the antenna approaches that of two-element Yagi. The less directional the antenna, the worse the effect. Omnidirectional antennas should not be mounted on the side of a mast or tower if it can be avoided. If an omni-directional antenna must be so mounted it should be spaced two wavelengths from it. This is difficult to achieve at the lower frequencies considering the antenna dimensions, weight and windloading. The rear element of a Yagi antenna must be one wavelength clear of the structure. To obtain a nominally omni-directional pattern a better, albeit expensive, solution is to mount a number of antennas, phased correctly, around the tower.

All antennas should point directly away from the structure, i.e. at right angles to the side of the tower.

The proximity of metalwork also adversely affects the VSWR of an antenna, causing standing waves with their additional voltage stresses on the feeder and possibly producing interference.

19.7.2 Practical aspects of antenna installation

All installation components must be made from compatible materials, electrically dissimilar metals, e.g. steel and brass, brass and aluminium, being avoided at all costs; they will corrode and cause intermodulation. Preferred materials are aluminium, and galvanized or stainless steel. Mounting brackets must be secure – remembering that antennas and feeders vibrate – but not overtightened so as to distort and weaken antenna support booms.

Connectors of the correct impedance, and preferably with a captive centre pin, are vital. Type N connectors are available at either 50 or 70 ohms impedance, are robust and can be used at frequencies up to 10 GHz. UHF 83 series connectors, although very robust, are of imprecise impedance and, from first-hand experience of interference aggravated by a UHF 83 used to connect an antenna feeder to a receiver tail, should not be used on base station installations. Assembly instructions for a range of connectors are shown in Chapter 23. All joints must be waterproofed, preferably by first wrapping with a self-amalgamating tape for two inches either side of the joint, then a layer of PVC tape covered by a further layer of Sylglass or Denso tape. When installing feeder cables the required length should be loosely coiled and taken up the tower and paid out as it is secured, working from top to bottom. It must not be dragged from a drum on the ground. Feeders must be cleated at the intervals specified by the manufacturer. Route cables where they are easily accessible but least likely to suffer physical damage.

Health and safety

Apart from the physical risks to riggers, there are radiation hazards. Large amounts of RF power may be radiated from antenna systems and research is continuing into its effects on health. The most recent recommendations of the National Radiological Protection Board and the Department of Trade and Industry should be sought and followed. Riggers should be aware of the power levels present on a structure before climbing, and equipment switched off where levels are considered unsafe.

19.7.3 Antenna checking and fault finding

The majority of VHF and UHF antennas present a short-circuit across the antenna and to ground as a means of reducing their vulnerability to static charges. An ohmmeter applied to the lower end of the feeder will indicate a circuit through the feeder and antenna but may also indicate a short-circuit or low resistance fault, water in the feeder for instance. Most antenna system faults increase the VSWR present on the feeder cable. A good antenna, operating within its design bandwidth, exhibits a VSWR of 1:1. When connected to a length of co-axial cable this may be raised and a VSWR of 1.5:1 for the system is generally considered acceptable. However, faults can be masked by conditions occurring on the feeder, and a measurement taken with a standing wave, or forward and reflected power meter at the bottom of the feeder, may not indicate the true condition of the antenna. A measurement taken with an accurate dummy resistive load connected to the top of the feeder will prove the cable, and a measurement of the VSWR should also be taken directly at the antenna. A record of the VSWR at installation and any subsequent measurements is helpful.

19.8 Interference

There are two sources of radio interference induced by antenna systems. One occurs from strong signals radiated by either the system's own transmitter antenna or from a co-sited system operating on a close neighbouring frequency. The other source is intermodulation, 'rusty bolt effect', the mixing of two or more signals to produce the interfering frequency.

19.8.1 Antenna isolation

One solution to the problem of direct radiation is to space the antennas so that there is sufficient isolation between them. Because minimum radiation is present immediately below and above the ends of the elements of vertically polarized antennas, maximum isolation occurs when such antennas are mounted in a vertical line. Maximum radiation, and hence minimum isolation, occurs when antennas are mounted side-by-side. The degree of isolation depends on the spacing but a figure of 40–45 dB between a transmitter and a receiver antenna, and 20–25 dB between transmitting antennas, should be the target. To achieve 25 dB isolation requires a vertical separation of 0.9 wavelengths between the centres of vertically polarized dipoles; 45 dB requires a vertical separation of 3 wavelengths. The spacing for horizontal separation of vertically polarized antennas is 2.5 wavelengths for 25 dB isolation and 25 wavelengths for 45 dB. To preserve this isolation the feeder cables should also be separated over their routes.

Antenna multi-coupling

A precise method of controlling the RF isolation between mutually sited systems where the frequencies are moderately spaced is for several systems to share one antenna by using multi-coupling techniques (see Section 19.9).

19.8.2 Intermodulation

Any two or more RF signals applied to a non-linear device intermodulate, that is, they combine to form additional frequencies. *Table 19.2* lists the combinations for two input frequencies. It must be remembered that the side frequencies produced by modulation of the original carriers will also be present.

Table 19.2 Low order intermodulation products

2nd	$A+B$	5th	$3A+2B$	7th	$4A+3B$
	$A-B$		$3B+2A$		$4B+3A$
			$3A-2B^{*}$		$4A-3B^{*}$
3rd	$2A+B$		$3B-2A^{*}$		$4B-3A^{*}$
	$2A-B^{*}$		$4B+A$		$5A+2B$
	$2B+A$		$4A+B$		$5B+2A$
	$2B-A^{*}$		$4A-B$		$5A-2B$
			$4B-A$		$5B-2A$
4th	$2A+2B$				$6A+B$
	$2A-2B$	6th	$5A+B$		$6B+A$
	$2B+2A$		$5B+A$		$6A-B$
	$2B-2A$		$5A-B$		$6B-A$
	$3A+B$		$5B-A$		
	$3A-B$		$4B+2A$		
	$3B+A$		$4B+2B$		
	$3B-A$		$4B-2A$		
			$4B-2B$		
			$3A+3B$		
			$3A-3B$		
			$3B-3A$		

*indicates in-band products. No eighth-order products fall in-band but ninth-order in-band products are produced by:

$5A-4B$

$5B-4A$

19.8.3 Control of intermodulation

A clue to the source of interference will be obtained from examination of the relationship of all frequencies produced on site, and also those produced nearby. That the side frequencies produced by modulation must be considered was proved in an intermodulation situation where the interference was only received when a nearby band III television transmitter radiated a peak white signal.

Corroded joints in metalwork close to the site, receiver RF stages subjected to very large signals and transmitter output stages are all non-linear devices and can create intermodulation.

Regular maintenance of the metalwork around the site reduces the chances of intermodulation from this source, but it is difficult to see and eliminate all corrosion on an inspection. Should interference from this cause be suspected after elimination of the other sources, a successful method of locating the offending joint requires a receiver tuned to the interfering signal and fitted with a whip antenna. While receiving the interference, a very sharp null in signal strength occurs when the tip of the whip points directly at the source of the signal.

Receiver RF amplifiers are usually designed to cope with small signals. When a strong signal radiated from a nearby transmitter is received the stage may well be overloaded and driven into non-linearity, or even into a blocked state where the receiver is effectively dead and may take some seconds to recover after the signal is removed. In the non-linear state the receiver is in an ideal condition to create intermodulation. If this situation is suspected a method of identifying it is to install a variable RF attenuator in the antenna feed to the suspect receiver. With zero attenuation the interference will be received but, if the receiver is the cause, increasing attenuation will produce little reduction of the interference until a point is reached where the interference suddenly disappears – the receiver at this point has re-entered a linear state. If the interference reduces gradually to zero with increased attenuation, the source is elsewhere.

Most amplifiers are designed to be linear, that is, the output signal level will follow that of the input signal. However, with a sufficiently high input signal overloading occurs, the amplifier becomes non-linear and compression of the signal results (*Figure 19.2*). The point where 1 dB of compression occurs is a commonly referred to amplifier parameter. If the input level is increased further the gain of the amplifier is reduced until saturation is reached when the output level can no longer increase.

At the onset of non-linearity harmonics and intermodulation, if any other frequencies are present, are produced. The strength of these rises rapidly with an increase of input, and the point where an extension of the almost linear portion of their curve crosses an extension of the linear gain line of the amplifier is the third-order intercept – usually it is only the third-order products which are considered for this purpose.

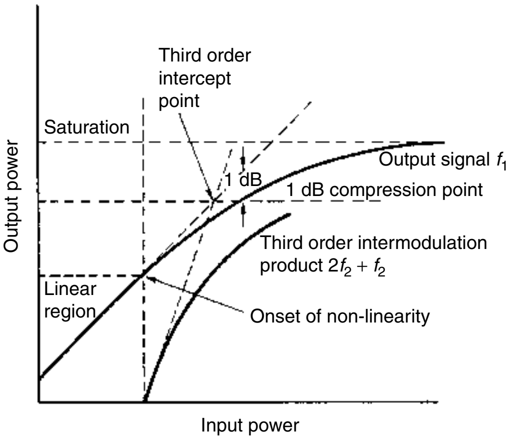

Figure 19.2 Production of intermodulation

The class C operated output stages of transmitters are by design non-linear. Strong signals from a neighbouring transmitter or antenna applied via the antenna feed will mix with the transmitted frequency to create intermodulation products. Increased isolation between antennas and feeders may be the simplest remedy. Alternatively, a circulator or filters may be connected to the output of the transmitter,

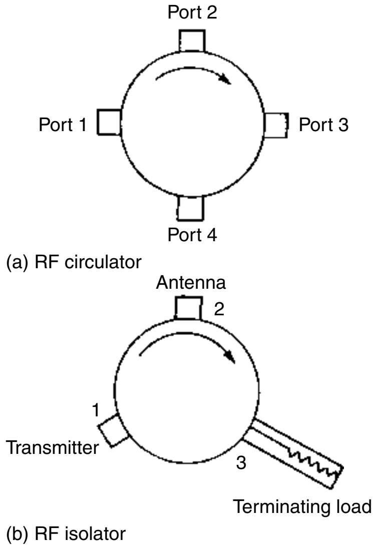
Port 2
Port 1
Port 3
Port 4
Antenna
2
1
Transmitter
3
Terminating load

(a) RF circulator

(b) RF isolator

Figure 19.3 RF circulator and isolator

and the possibility of direct radiation from transmitters must not be ignored.

A circulator is a uni-directional device with either three or four ports as shown in *Figure 19.3(a)*. A signal entering at port 1 will leave at port 2 but not at ports 3 or 4. Similarly, a signal entering at port 2 will leave at port 3 but not ports 1 or 4. Circulators are used for combining the outputs of transmitters for application to a single antenna but if a three-port circulator is connected as in *Figure 19.3(b)* it forms an isolator. Typical isolation over a 1.5% bandwidth is 20–30 dB with an insertion loss of 0.7 dB.

19.9 Antenna multi-coupling

The sharing of antennas and feeders between a number of co-sited services is not only a precise method of controlling frequency isolation, it is an economic solution to the problems of antenna and tower management. An antenna may be shared between a number of receivers or transmitters, or shared simultaneously by both receivers and transmitters. Receiver sharing requires a splitter, i.e. a filter, to separate the frequencies for each receiver and an amplifier to make up the filter losses. Transmitters can share an antenna through circulator/isolators and filters. To obtain the required selectivity, stability

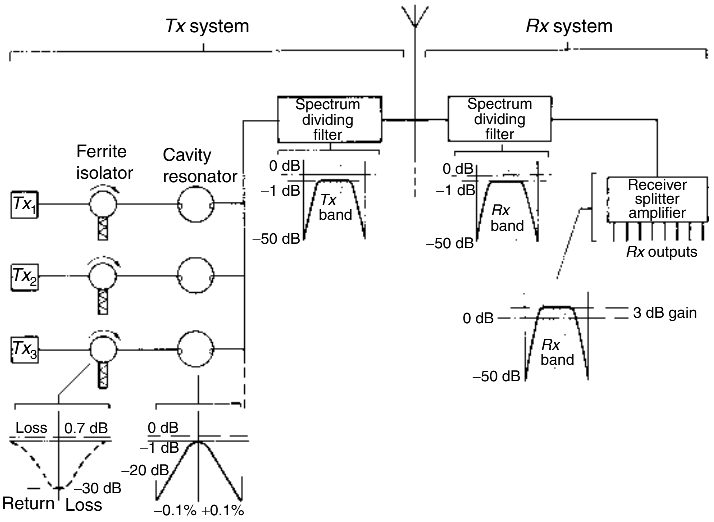
Tx system
Rx system
Spectrum dividing filter
Spectrum dividing filter
Ferrite isolator
Cavity resonator
0 dB
–1 dB
Tx band
–50 dB
0 dB
–1 dB
Rx band
–50 dB
Receiver splitter amplifier
Rx outputs
*Tx*1
*Tx*2
*Tx*3
0 dB
3 dB gain
Rx band
–50 dB
Loss
0.7 dB
–30 dB
Return
Loss
0 dB
–1 dB
–20 dB
–0.1% +0.1%

Figure 19.4 Typical sub-band transmitter/receiver sharing (© Crown Copyright 1991, Radiocommunications Agency)

and power handling capability the filters are often solidly constructed cavity resonators with an insertion loss of 1 dB and a bandwidth of 0.1% to the −20 dB points. *Figure 19.4* shows the layout of a transmitter/receiver sharing system.

19.10 Emergency power supplies

The installation of equipment to provide power for all the equipment on a site during a failure of the public supplies has declined in recent years. But although lengthy power failures seem to be a thing of the past, they could recur, and electricity supplies do still fail for short periods. To provide a complete supply for large station necessitates a standby generator. On smaller stations batteries may be capable of supplying the demands of the electronic equipment and emergency lighting for one or two hours. The trend on multi-user shared sites appears to be for the individual user to provide the emergency power, derived from batteries, for his own equipment. Factors to consider are:

1. Standby motor generators are expensive. They need housing, ventilation and fuel storage. They take time to start up, 45 seconds being typical.
2. Uninterruptible power supplies (UPS). Where no interruptions can be tolerated the most efficient supply is provided by batteries alone but the equipment must be designed to operate directly from a DC, low voltage supply.

   To supply power at 220/240 V AC an inverter is the accepted method. Double inverters permanently charge a battery from the AC supply and the battery drives an oscillator to reproduce the AC supply voltage, a very inefficient process. Also, should any item of the inverter break down the supply is lost. This is overcome in some inverters by a switch which transfers the load to the public supply when a failure of the inverter occurs. Single inverters charge a battery and supply the equipment directly from the public supply until a failure occurs. When this happens the supply changes over to the battery and oscillator. Single inverters are more efficient but there is a short break in supply while the change-over occurs. Adequate stabilization and filtering to avoid disturbances on the supply must be provided. One possible hazard with all inverter supplies is that where more than one inverter is used, the outputs of the inverters are unlikely to be in phase and a 480 V difference could exist between the supplies for two items of equipment.

For all battery-operated supplies correct charging of the battery is crucial. Nickel–cadmium batteries require constant-current charging with a fall to the trickle rate when fully charged and sophisticated circuitry is needed to identify the precise point of full charge (see Section 21.3.2).

Lead acid batteries produce hydrogen if over-charged so their storage must be designed so that no danger, either to personnel or of explosion, exists. For reliability, natural ventilation is preferred to forced. Recommendations for the accommodation of lead acid batteries other than sealed are given in BS 6133: 1985. Batteries which gas also need more frequent maintenance. Because of the reduction in performance of batteries at low temperatures the ambient temperature of the battery storage room should not fall below 4°C, and to reduce water evaporation lead acid batteries should not be run continuously above 40°C.

The period of time over which a lead acid battery must be fully recharged determines the capacity of the battery more than the period of time over which it must supply power. Because of the need to reduce the charge rate – to avoid gassing – to a very low level as the battery approaches full charge, the last 15% of capacity takes a very long time to acquire (see Section 21.3).

19.11 Approval and certification

No base radio station may be operated in the UK prior to inspection of the installation by the Radio Investigation Service of the Radiocommunications Agency and the issue of their certificate of approval. The installation may not subsequently be modified without re-approval and certification.

References

Belcher R. *et al.* (1989). *Newnes Mobile Radio Servicing Handbook.* Butterworth-Heinemann, Oxford.

BS 6651: 1985. *Code of Practice for the Protection of Structures against Lightning.*

BS 6701: Part 1:1990. *Code of Practice for Installation of Apparatus Intended for Connection to Certain Telecommunications Systems.*

BS 7430: 1991. *Code of Practice for Earthing. Earthing of Telecommunications Installations.* International Telecommunications Union.

MPT 1331. *Code of Practice For Radio Site Engineering.*

MPT 1351. *Code of Practice For Repeater Operation at Communal Sites.*

MPT 1367. *Guidance Note to Legal Requirements Covering the Installation and use of Radio Apparatus and other Apparatus Generating Radio Frequency Emissions.*

MPT 1368. *Code of Practice for the Inspection of a Land Mobile Radio System for Conformity with the Wireless Telegraphy Act Licence and Performance Specification.*

Note: MPT documents are available from the Radiocommunications Agency (UK).

20 Instrumentation

20.1 Accuracy, resolution and stability

All measurements are subject to error and two instruments applied to the same piece of equipment under test may give a different answer. Tolerances must therefore be accepted. The errors arise from the following sources:

1. Human error, e.g. precision in reading a scale, use of incorrect instrument or range setting for the purpose.
2. The accuracy to which the instrument is able to display the result of a measurement or, in the case of a generator, the frequency or output level.
3. Accuracy of calibration.
4. Tolerances in the components used in the construction of the instrument. Variations in the load applied to the instrument.
5. Variations caused by long-term drift in the values of components.
6. Variations due to temperature and supply voltage fluctuations, and the warm-up time required by some instruments.
7. An effect on the circuit under test by the connection of the instrument.

There is an important difference between the accuracy and the resolution of an instrument. The accuracy is a statement of the maximum errors which may occur from the causes in statements 3 to 6 above. In instrument specifications, stability defined by statements 5 and 6 is usually quoted separately.

The accuracy of analogue measuring instruments is normally quoted as a percentage of full scale deflection (FSD). This is the accuracy of the instrument movement and components plus the scale calibration. The scale graduations, though, may not permit the user to determine the reading to the accuracy of the instrument perhaps because the graduations are cramped or parallax reading error occurs. These factors decide the resolution or precision of reading.

The accuracy of instruments with digital displays is usually quoted as a percentage of the reading plus or minus one count or one digit. While digital instruments are generally more accurate than their analogue counterparts, the fact that the least significant figure may be in error affects the resolution. *Figure 20.1* shows how this can arise. Most digital instruments use a gating process to switch the input to

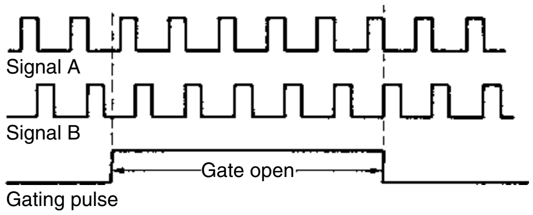

Figure 20.1 Gating error introduced by signal phasing

the measuring circuitry for the appropriate period of time. The gating time itself may vary and affect the accuracy, but even with a perfectly accurate and stable gating time, the phase of the input signal at the time of switching affects the number of pulses passing through the gate and thus the resolution.

Performance figures taken from a number of manufacturer's catalogues are listed below. They show only the more important features of the specifications and are typical of those for high quality instruments used in radio work.

20.2 Audio instruments

20.2.1 Output power meters

Those for radio applications measure RMS audio power and are usually calibrated in watts and dBW or dBm. They usually contain a dummy load resistor of adjustable impedance as a substitute for the receiver loudspeaker when in use.

20.2.2 Distortion factor and Sinad meters

These may be separate instruments or combined, and may also be incorporated within an audio output power meter. Both operate on the same principle. An audio input amplifier, with automatic gain control to give a fixed reference level at its output, applies the output from the unit under test, first to a notch filter to remove the test frequency of 1000 Hz, then to an AC voltmeter. The voltmeter is calibrated in percentage distortion or dB Sinad.

In testing a receiver Sinad ratio, the instrument is connected across the loudspeaker or equivalent dummy load. An RF signal generator connected to the antenna socket is adjusted to apply a moderately high level signal modulated with a 1 kHz tone – which must be pure and match the notch filter in the Sinad meter – at 60% of the system peak modulation. To be correct, the receiver volume control is adjusted to give the rated audio output but if only Sinad sensitivity is being measured a lower level is acceptable. The Sinad meter will now display the combined level of all the frequencies except 1000 Hz present in the receiver's output. These are the noise and distortion products. With the high level input applied these will be low and the meter deflection will be small indicating a high Sinad ratio. The RF input level is now reduced until the meter reads 12 dB (the standard reference) and the RF level, μV or dBm, is the Sinad sensitivity for the receiver.

The oscillator to produce the test tone may be included in the meter and filters may be provided to comply with various weightings.

20.2.3 Audio signal generators

The instrument provides an output for synchronizing to an external standard and a TTL compatible output. A high power version capable of delivering 3 watts into a 3 Ω load in the audio band is available.

20.3 Radio frequency instruments

20.3.1 RF power meters

Direct reading RF power meters either contain a non-reactive load or use an external load and may be calibrated in watts or dBm. RF calorimeters convert the RF energy into heat and measure the temperature of the heated element. At low powers, 'dry' calorimeters are used but their long thermal time constant inhibits their use at high power levels. To measure high powers 'flow' calorimeters, where a fluid flows around a closed system and the output temperature of the fluid is measured, are used. Power can be determined from:

$$P = F(T_{\text{out}} - T_{\text{in}})c(T)$$

where

P = power

F = mass flow rate of the fluid

$c(T)$ = the fluid's specific heat

T_{in} = temperature of the fluid entering the load

T_{out} = temperature of the fluid after being heated by the load

Thruline type instruments require an external load which may be either an antenna system or a load resistor. This type of meter reads forward and reflected power and, in some instruments, VSWR directly.

20.3.2 *RF signal generators*

The range of instruments designed for use on specific systems, e.g. cellular and digitally modulated, is so wide that manufacturers' catalogues must be consulted for each application. In addition to the accuracy of the carrier frequency and output level, RF leakage, spectral purity and modulation noise levels are important. The output level may be calibrated in μV, dBV or dBm, and may refer to either an unterminated instrument (p.d.) or terminated in a load equal to the output impedance of the generator (e.m.f.). If the instrument is calibrated in p.d., the output voltage must be halved when the instrument is terminated in an equal impedance.

The 50 ohm output impedance of many signal generators matches the input impedance of most VHF and UHF receivers directly. To simulate the impedance of antennas at HF and below a dummy antenna is usually inserted between the generator and the receiver under test. *Figure 20.2* shows the circuit of a standard dummy antenna.

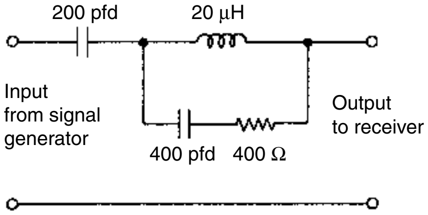

Figure 20.2 Standard dummy antenna

20.3.3 *Frequency counters*

Instruments are necessary to measure not only the precise frequencies of transmitters and receiver local oscillators but also those of low frequency signalling tones. Consequently they have a wide operating range.

The output of a transmitter must never be connected directly into a counter. The signal should either be obtained off-air or through a 'sniffer', appropriate for the frequency range, which siphons off a fragment of the transmitter output. It is also essential to ensure that the frequency displayed is the fundamental and not a harmonic or some other spurious frequency.

20.3.4 *Modulation meters*

The remarks in the above section relating to the connection of a transmitter and the selection of the fundamental frequency apply equally to modulation meters.

20.3.5 *Spectrum analysers*

Today's tight regulation of spectrum usage makes a spectrum analyser an essential tool for the precise alignment of radio equipment and for the measurement of modulation products and the noise content of signals.

A spectrum analyser is essentially a superheterodyne receiver with an adjustable IF bandwidth which sweeps across a portion of the frequency spectrum in synchronism with the horizontal trace of a cathode ray tube display. A signal at any frequency within the swept band will, while it remains within the IF passband, appear as a vertical displacement of the display trace. The design of the filter is crucial but must be a compromise. An analyser using multiple filters permits fast measurement but low resolution; a single filter offers high resolution but slower response time. Also, the shape factor of a wide filter inhibits the display of low level signals close to the centre frequency of the filter and a narrow filter, while permitting the display of these signals, may fail to display transients. The bandwidth determines the sweep speed, narrower bandwidths requiring slower speeds.

The possibility exists of spurious signals, e.g. intermodulation, being produced within the analyser with large input signals. A method of reducing the possibility is to insert a circulator and a notch filter

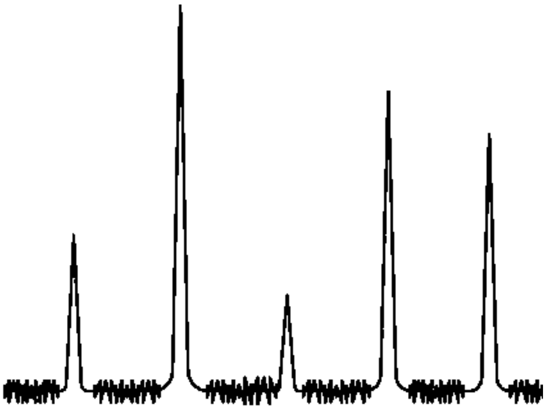

Figure 20.3 Amplitude-vs-frequency spectrum analyser display

between the unit under test and the analyser input to reduce the fundamental frequency amplitude. The circulator must be installed between the filter and the unit under test to absorb the reverse power produced by the impedance variations of the filter.

Many instruments are programmable and suitable for incorporation in automatic testing (ATE) systems, and instruments with similar features to those described are also offered in a combination test set form.

20.3.6 Network analysers

A network analysers examines *incident, reflected* and *transmitted* signals through a circuit or device, and displays the magnitude and phase of these signals. A spectrum analyser, on the other hand, measures only one channel, and displays magnitude and frequency.

Figure 20.4 shows an optical analogy to the network analyser. A 'device' of different optical density than ambient is in the path of an incident ray of light (R). When the light hits the surface, part of it is reflected (A) and part is transmitted through the 'devices' (even though refracted a bit).

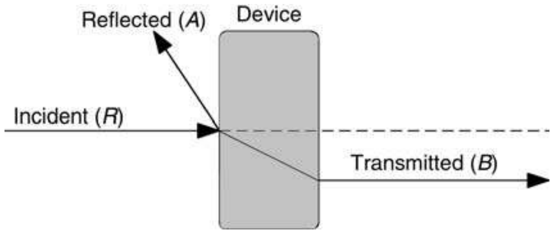

Figure 20.4 Optical analogy to network analyser

A network analyser consists of a three-channel RF receiver and a display. The incident signal is considered the *reference signal*, so is designated the R-channel. The other two channels receive the reflected signal on the A-channel, and the transmitted signal on the B-channel.

The uses of scalar and vector network analysers differ from the uses of spectrum analysers. The spectrum analyser measures external signals of unknown frequency and modulation type. Even when a tracking generator is added, to allow the spectrum analyser to perform stimulus-response tests, the spectrum analyser cannot do the job of the network analyser. The network analyser, by contrast, contains a known signal source, and is capable of sweeping a range of frequencies and power output levels. It can also perform ratio measurements.

References

Belcher, R. *et al*. (1989). *Newnes Mobile Radio Servicing Handbook*. Butterworth-Heinemann, Oxford.

Bird Electronic Corp., Catalog GC-92.

Farnell Instruments Ltd, Catalogue, 1993.

Marconi Instruments Ltd, Test and Measurement Instrument Systems 1993/94.

RTT Systems Ltd, Catalogue, 1993.

Winder, S.W. (2001). *Newnes Telecommunications Pocket Book*. Butterworth-Heinemann, Oxford.

21 Batteries

21.1 Cell characteristics

Batteries are composed of cells which exhibit characteristics peculiar to their chemical constituents and construction. Common features, important to users, are their ability to store energy within a small space and with the least weight, and to release it at an adequate rate for the purpose under consideration.

21.1.1 Capacity

The amount of energy a battery can store is measured in ampere hours (Ah) at a specified discharge rate. For large cells this is usually the 10 hour rate but American practice, which is now almost universal, at least for smaller cells, is to use the 1 hour rate. It is important, therefore, to be certain which rate is referred to. The capacity reduces as the rate of discharge increases. Thus a battery of 60 ampere hours capacity at the 10 hour rate will provide 6 amps for ten hours before reaching the point at which it is considered to be discharged. If a current of 12 amps is taken, the battery will become discharged in less than 5 hours, and if the current is 3 amps, it will last longer than 20 hours. The rate of discharge is often referred to in terms of the C-rate which may be expressed in several ways. 1C, C or C_1 are numerically the same as the rated capacity, e.g. a 500 mA Nicad cell supplying 500 mA, which may be expressed as C, 1C or C_1 continuously will be discharged in approximately 1 hour. If the cell supplies current at a reduced rate, 0.5 C, 0.5 C_1, or C/2, i.e. 250 mA, it will last approximately 125 minutes. The subscript, e.g. C_5, indicates the hourly discharge rate.

The terminal voltage at which a cell is considered discharged also varies with the discharge rate; a lead acid cell discharged at 1C is considered to be discharged when the terminal voltage falls to 1.75 V. At C_{10} the cell is considered to be discharged at 1.85 V.

The capacity of a battery or cell may also be specified at a given ambient temperature, usually 20°C. Lower temperatures reduce the effective capacity and maximum current off-take, higher temperatures increase them slightly.

For radio use battery duration may be quoted in terms of standby, receive and talk time. The duty cycle obviously varies from user to user but a useful standard for a radio-telephone is 90% standby, 5% receive and 5% transmit. On an open channel PMR system, 80%/15%/5% is more typical. Measuring the current drains during these activities enables the battery requirements to be calculated.

21.1.2 Internal resistance

The maximum instantaneous current which a battery can deliver is determined by its internal resistance. In this respect a battery behaves like any other generator (see Section 21.3.2), where increasing load currents produce an increasing voltage drop across the internal resistance. The internal resistance of a battery is seldom specified, but for a battery in good condition it is extremely low (one quoted figure is 15 mΩ for a fully charged 500 mAh Nicad cell) and the voltage drop in the connecting leads will govern the maximum withdrawable current. Equally, the resistance of the meter used will affect the measurement of charge or discharge currents. If using an analogue meter, the older low resistance types are preferable (I keep a model 40 Avometer, 0.03 Ω int. res. on 12 amp range, for the purpose). Low current-rating fuses also present a resistance higher than that of a battery (4 Ω measured for a 250 mA fuse, 0.5 Ω for a 1 A) and, probably, the connecting leads.

21.1.3 Power:weight and volume ratios

It is the battery that now limits the size to which radio equipment can be reduced. Recent developments have increased the power to weight and power to volume ratios which are possible. Typical ratios are referred to in the sections dealing with individual battery types.

21.1.4 Recharging conditions

The initial charge rate is the current flowing through a discharged battery to replace the charge in a specified time. Unless supplied from a constant current charger the current will fall as the battery voltage rises, but as full charge is approached the charge rate is usually reduced to a trickle or finishing charge rate.

Trickle charging maintains the cells in a fully charged condition by passing a very small current through them sufficient merely to replace any self-discharge losses through leakage.

Finishing charge is a rate to which the charging current is reduced when a battery reaches about 85% of its full capacity. It is a rate at which gassing is unlikely to occur.

Float charging maintains the cell voltage at its nominal while it is supplying continuous and variable loads.

21.2 Non-rechargeable, primary batteries

While rechargeable batteries are the obvious choice for use in equipment which is in use continuously, disposable types have economic and logistic advantages for some applications – the batteries are more expensive long term but no charger is needed. Also primary batteries have up to 4 times the available capacity of their nickel cadmium equivalents with less weight. Details of disposable batteries with some standard nickel-cadmium rechargeable equivalents are given in *Table 21.1*.

Lithium primary batteries are now available and have the following qualities:

1. A high cell voltage of 3.6 V on load. The voltage is constant – after an initial fall when load is first applied – until discharged when a rapid fall occurs.
2. An operating temperature range of −55°C to +75°C is possible.
3. Energy densities up to 630 mWh/g. A standard AA size cell has a power/weight ratio of 340 mWh/g.
4. Long shelf life. Ten years at room temperature for a 10% fall in capacity is envisaged.

They have a disadvantage in that fire, explosion or severe burns may result if the batteries are mistreated. They must not be recharged, disassembled, heated above 100°C, incinerated or the contents exposed to water.

One use of lithium batteries is as a power source for memories but protection against charging is needed. *Figure 21.1* shows a typical circuit.

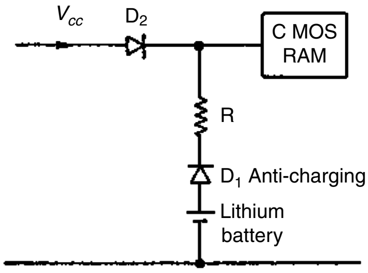

Figure 21.1 Typical lithium battery application

21.3 Rechargeable batteries

21.3.1 Lead acid batteries

Lead acid batteries, whether of free electrolyte or low-maintenance sealed construction, have a nominal terminal voltage on load of 2.0 V per cell. This voltage falls on load in a gradual curve, shown in *Figure 21.2*, until the discharged voltage of between 1.75 and 1.85 volts per cell is reached.

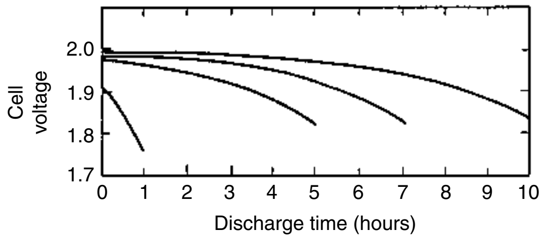

Figure 21.2 Typical lead acid cell discharge characteristics

A discharged battery will, because of its inefficiencies, require a recharge equal to the amperes × hours discharged +11%, e.g. a cell discharged at 5 amps for 10 hours will require a recharge of 55.5 ampere hours with a constant current at the 10 hour rate. Because of the reducing current as full charge is approached a recharge time of 1.4 to 1.5 times the capacity to be restored is more practical. The final on-charge cell voltage can increase to approximately 2.7 volts. Gassing occurs and hydrogen is liberated when the cell voltage reaches 23 V but provided the charging current is sufficiently low above this point gassing will be avoided. This lower charge rate, the 'finishing rate' can be applied by maintaining the charging voltage at about 2.4 volts when the battery will automatically limit the charging current. The specific gravity of the electrolyte in a fully charged cell is between 1.205 and 1.215.

The trickle charge current must be low enough to avoid gassing. A current of 7% of the 10 hour capacity is typical.

Float charging should maintain a cell voltage of approximately 2.2 volts.

The power:weight ratio of lead acid batteries is poor, a small (4 Ah), 6.0 V, sealed lead acid battery having a power:weight ratio of 26 mWh/g.

21.3.2 Nickel-cadmium (Nicad) batteries

At present the Nicad is probably the most commonly used rechargeable battery for portable applications. A standard size AA cell has a power:weight ratio of 27 mWh/g. The on-load Nicad cell voltage, after an initial fall from the on-charge voltage of between 1.3 and 1.4 V, remains substantially constant at about 1.2 V until the discharged voltage of 0.9 to 1.1 V is reached. Thereafter the voltage falls rapidly. This is illustrated in *Figure 21.3*. While the constant voltage is ideal during discharge it poses a problem in that it is difficult to reliably measure the intermediate state of charge which created difficulties with recharging. A Nicad battery which is repeatedly partially discharged and then recharged may, after many cycles, behave as though it were fully discharged when the repeated recharge condition is reached (the memory effect) and, with a fixed time charger, the possibility of over-charging is present. One solution fully discharged all batteries after use to a predetermined level, typically 1.1 V per cell, and then recharged them at a constant current for a fixed period of time. Unfortunately, this procedure shortened the life of the batteries; *Figure 21.4* shows the life expectancy of a cell with repeated discharges. The present solution is to charge the batteries automatically to the fully charged state and then reduce the current to the trickle charge level. Batteries which are subjected to repeated partial discharge may then be occasionally fully discharged to obviate the memory effect.

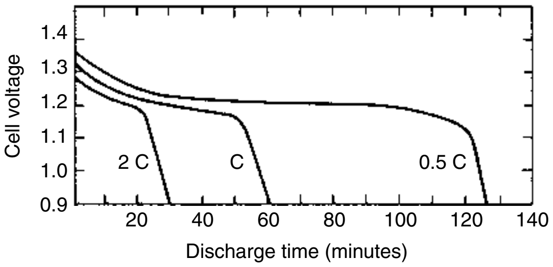

Figure 21.3 Typical Nicad cell discharge characteristics

Constant current, automatic charging is recommended. Chargers vary in complexity, some detecting the end-of-charge point by sensing a variation of voltage. At end of charge the cell voltage first rises and then falls slightly as in *Figure 21.5*. More sophisticated chargers also sense the cell case temperature which rises during charge. Batteries are available for standard charging at the ten hour rate where 14 to

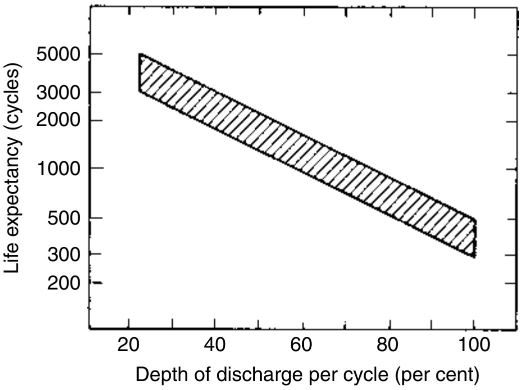

Figure 21.4 Nicad cell: effect of repeated discharge vs. cell life

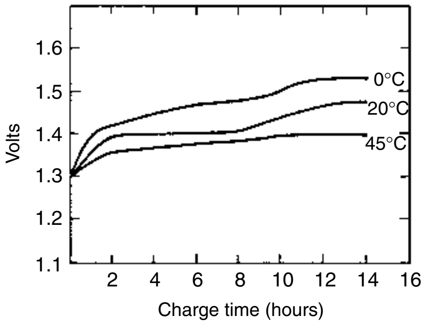

(a) Standard charge, C/10 × 15 hrs

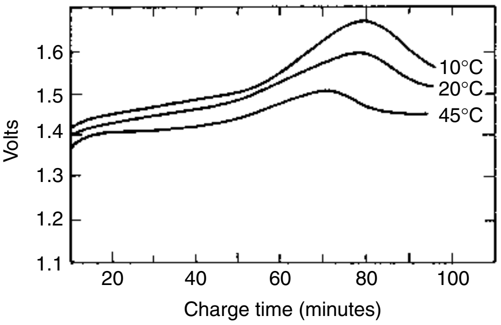

(b) Rapid charge, 1 C × 90 mins

Figure 21.5 Nicad cell: discharging characteristics

15 hours will be required to recharge a fully discharged battery, fast charging at 5 C and rapid charging at C rate.

Cells may be fitted with a re-sealing one-way vent which opens at about 200 psi and closes at about 175 psi to relieve any excess internal pressure caused by a fault or abuse.

22 Satellite communications

22.1 Earth orbits

Communications satellites are required to illuminate the earth with radio signals and their orbits are chosen according to the size and location of the part of the earth's surface they must light up.

A satellite orbiting the earth is continuously pulled by a centripetal force, in this case gravity, towards the centre of the earth. It is also pulled by centrifugal force to leave its orbit at a tangent. When these opposing forces are equal in magnitude the satellite is in a stable orbit. There is, then, for a given height (the radius of the path minus the radius of the earth, 6378 km), a velocity at which the conditions for stable orbit apply, and which determines the orbiting time.

22.1.1 Geostationary orbits

Satellites relay information from ground stations, either fixed or mobile, or between satellites. It is an advantage for some purposes, therefore, to use satellites with an orbit time identical to that of the earth so that no tracking from the ground stations is needed. Geostationary satellites have the same angular velocity as the earth making them appear to be stationary. Their height is 35 788 km and four such satellites cover the earth from latitude 81.3°N to 81.3°S as in *Figure 22.1*.

The disadvantages of geostationary satellites are that they are in a high earth orbit (HEO) resulting in a signal delay of 240 ms for the complete go and return path. Also they are in an equatorial orbit so that signals to the higher latitudes travel at a shallow angle to the earth's surface rendering them unsuitable for mobile use where communications must be achieved at street level in cities.

As each satellite covers a large portion of the earth, the design of their antennas to permit repeated frequency re-use is important and antennas with small, steerable footprints have been developed for this purpose. An advantage, in addition to the lack of tracking, is that the shadowing by the earth is minimal so that solar power cells receive almost continuous illumination.

22.1.2 Elliptical orbits

A polar orbit, where the satellite follows a North/South track, provides the opportunity to survey the earth in a series of strips. The satellites

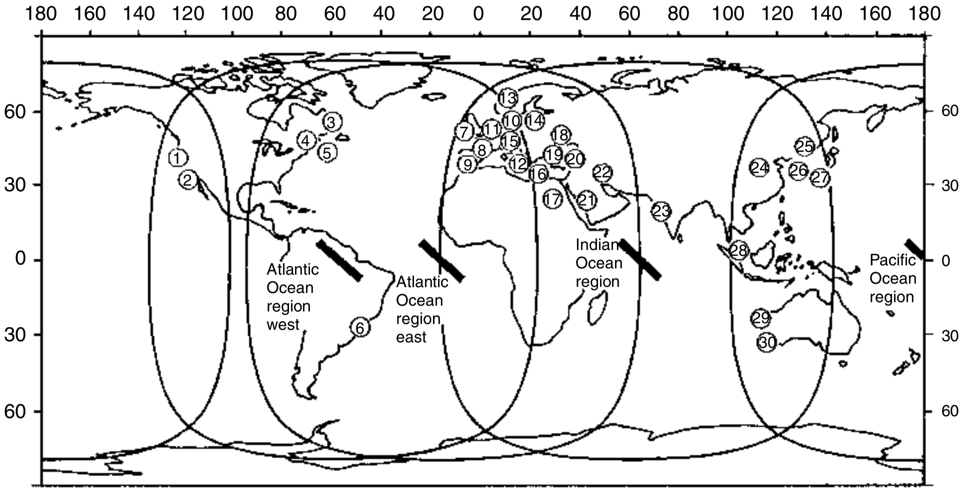
180
160
140
120
100
80
60
40
20
0
20
40
60
80
100
120
140
160
180
60
30
0
30
60
Atlantic Ocean region west
Atlantic Ocean region east
Indian Ocean region
Pacific Ocean region
①
②
③
④
⑤
⑥
⑦
⑧
⑨
⑩
⑪
⑫
⑬
⑭
⑮
⑯
⑰
⑱
⑲
⑳
㉑
㉒
㉓
㉔
㉕
㉖
㉗
㉘
㉙
㉚

Land earth station services	Inmarsat-A	Inmarsat-C	Inmarsat-Aero
1. Niles Canyon	●		●
2. Santa Paula	●	●	●
3. Laurentides			●
4. Southbury	●	●	●
5. Staten Island	●		
6. Tangua	●		
7. Goonhilly	●	●	●
8. Pleumeur Bodou	●	●	
9. Ausseguel			●
10. Blavand		●	
11. Burum	●	●	
12. Fucino	●		
13. Elk	●	●	●
14. Psary	●		
15. Raisting	●	●	
16. Thermopylae	●		
17. Maadi	●		
18. Odessa	●		
19. Anatolia	●		
20. Ata	●		
21. Jeddah	●		
22. Boumehen	●		
23. Arvi	●		
24. Beijing	●		
25. Nakhodka	●		
26. Kumsan	●		
27. Yamaguchi	●		●
28. Singapore	●	●	●
29. Perth	●	●	●
30. Ganagara	●		

Figure 22.1 Immarsat coverage and GES locations

used for this purpose are in low each orbit (LEO) and consequently have a high velocity. Their orbit time is approximately 1.5 hours and between successive orbits the earth has rotated 22.5°. Sixteen orbits are therefore needed to scan the earth's surface. Until recently polar orbits were used only for optical surveillance but now several projects for radio communication are either in the early stages of installation or under development. These use a number of satellites so that for mobile communications, for instance, there is a satellite continuously in view. Tracking of these extremely fast satellites by the ground stations, which may themselves be moving, along with Doppler effect has been a major obstacle which is now being overcome.

22.2 Communications by satellite link

Satellites are radio links and receive signals from the ground and other satellites which they must re-transmit. The signals from ground stations are comparatively weak and require high power amplification for onward transmission. As the satellite's receive and transmit antennas must be close together, the possibility of RF instability prohibits on-frequency repetition. The up-link, from ground station to satellite, must therefore be converted to another before re-transmission. The up-link frequency is normally higher than the down-link and the frequency converter is referred to as a 'down-converter'.

The frequency bands used for communications purposes are listed in *Table 22.1*.

Satellite television

The broadcasting of television may be via either communications satellites or Direct Broadcasting by Satellite (DBS) satellites. The positions of non-DBS satellites relative to the UK are shown in *Figure 22.2* and *Table 22.2* lists the European channels. *Figure 22.3* shows the world allocations of DBS satellites and *Tables 22.3* and *22.4* list the channel frequencies and national allocations. The frequency plan for the Astra satellite is in *Table 22.5*.

22.3 Proposed satellite television formats

Most current European satellite television programmes (non-DBS) are broadcast as fairly standard PAL signals, FM modulated into the satellite channel. DBS transmissions, and those from Astra satellite, will probably be of a multiplexed analogue component (MAC) format. In MAC, data corresponding to sound tracks and subtitles, etc., an

Table 22.1 Communications satellite frequencies

Frequency band (GHz)	*Link*	*European telecom links*	*International telecom links*
1.5–1.6	Down	Mobile	
1.6–1.7	Up	Mobile	
3.4–4.2	Down	Fixed	Fixed
4.5–4.8	Down	Fixed	Fixed
5.9–7.0	Up	Fixed	Fixed
10.7–11.7	Down	Fixed (+ Non-DBS television)	
11.7–12.5	Down	Fixed (+ television)	
12.5–12.75	Down	Fixed (private links)	
12.75–13.75	Up	Fixed (private links)	
14.0–14.8	Up	Fixed	
17.3–18.3	Up	Fixed	
17.7–20.2	Down	Fixed	
20.2–21.2	Down	Mobile	
27.0–30.0	Up	Fixed	
30.0–31.0	Up	Mobile	
22.5–23.55		Allocated for intersatellite links	
32.0–33.0		Allocated for intersatellite links	
54.25–58.2		Allocated for intersatellite links	
59.0–64.0		Allocated for intersatellite links	

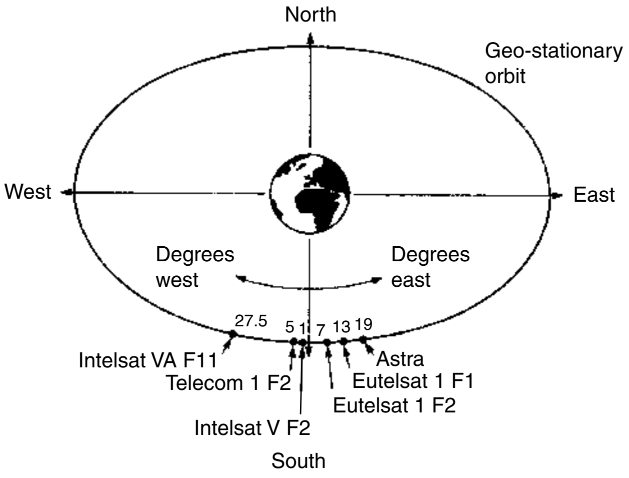
North
Geo-stationary orbit
West
East
Degrees west
Degrees east
27.5
5
1
7
13
19
Intelsat VA F11
Telecom 1 F2
Intelsat V F2
Astra
Eutelsat 1 F1
Eutelsat 1 F2
South

Figure 22.2 Non-DBS satellite television: positions of the main non-DBS satellites relevant to the UK

Table 22.2 European satellite television channels broadcast via communications satellites

Channel	Frequency (GHz)	Polarization	Audio sub-carrier	Video standard	Satellite
3Sat	11.175	H	6.65	PAL	Eutelsat 1 F1
Anglovision	11.515	V	6.6	PAL	Intelsat VA-F11
Arts Channel	11.135	H	6.6	PAL	Intelsat VA-F11
BBC1/2	11.175	H	6.65	PAL	Intelsat VA-F11
CanalJ	12.564	V	5.8	PAL	Telecom 1 F2
Children's Channel	11.015	H	6.6	PAL	Intelsat VA-F11
CNN	11.155	V	6.65	PAL	Intelsat VA-F11
Filmnet	11.140	V	6.6	PAL	Eutelsat 1 F1
Moscow 1	3.675	–	7.0	SECAM	Gorizont-12
Moscow 2	3.0	–	7.0	SECAM	Gorizont-7
Infofilm & Video	11.015	H	6.6	PAL	Intelsat F2
La Cinq	12.606	V	5.8	SECAM	Telecom 1 F2
Lifestyle	11.135	H	6.6	PAL	Intelsat VA-F11
M6	12.648	V	5.8	SECAM	Telecom 1 F2
MTV	10.975	H	6.65	PAL	Intelsat VA-F11
Norsk Rikskringkasti	11.644	H	Digital	C-MAC	Eutelsat 1 F2
Premiere	11.015	H	6.6	PAL	Intelsat VA-F11
RAI-Uno	11.007	H	6.6	PAL	Eutelsat 1 F1
RTL-Plus	11.091	V	6.65	PAL	Eutelsat 1 F1
SAT1	11.507	V	6.65	PAL	Eutelsat 1 F1
Satellite Information	11.575	H	Digital	B-MAC	Intelsat VA-F11
Screensport	11.135	H	6.6	PAL	Intelsat VA-F11
Skychannel	11.650	H	6.65	PAL	Eutelsat 1 F1
Superchannel	10.674	V	6.65	PAL	Eutelsat 1 F1
SVT-2	11.178	H	Digital	C-MAC	Intelsat F2
SVT-2	11.133	H	Digital	C-MAC	Intelsat F2
Teleclub	11.987	V	6.5	PAL	Eutelsat 1 F1
TV5	11.472	H	6.65	PAL	Eutelsat 1 F1
Worldnet	11.512	H	6.65	PAL	Eutelsat 1 F1
Worldnet	11.591	H	6.6	SECAM	Eutelsat 1 F1
Worldnet	12.732	V	5.8	NTSC	Telecom 1 F2

analogue signal corresponding to chrominance and an analogue signal corresponding to luminance are transmitted separately in each broadcast line of the picture.

In order to achieve multiplexing of the three parts of the format, time compression of chrominance and luminance signals occurs before

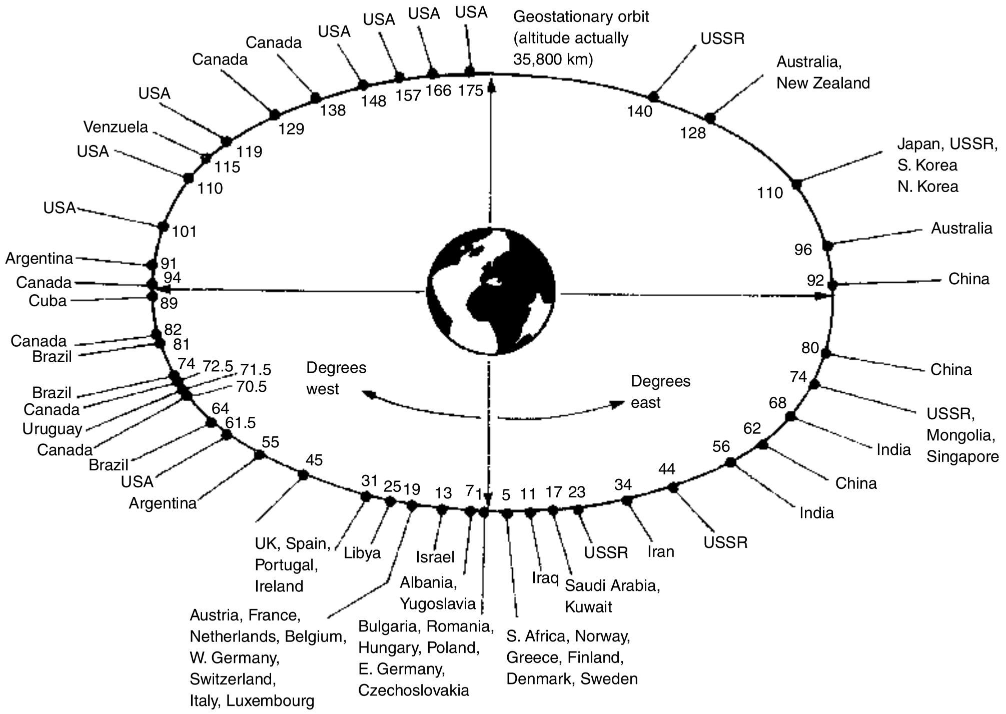
Geostationary orbit
(altitude actually
35,800 km)
Degrees
west
Degrees
east
USA
175
USA
166
USA
157
USA
148
Canada
138
Canada
129
USA
119
Venzuela
115
USA
110
USA
101
Argentina
91
Canada
94
Cuba
89
Canada
82
Brazil
81
74
72.5
71.5
70.5
Brazil
Canada
Uruguay
Canada
64
61.5
Brazil
USA
Argentina
55
45
UK, Spain,
Portugal,
Ireland
31
Libya
25
Austria, France,
Netherlands, Belgium,
W. Germany,
Switzerland,
Italy, Luxembourg
19
Bulgaria, Romania,
Hungary, Poland,
E. Germany,
Czechoslovakia
13
Israel
7
Albania,
Yugoslavia
1
5
S. Africa, Norway,
Greece, Finland,
Denmark, Sweden
11
Iraq
17
Saudi Arabia,
Kuwait
23
USSR
34
Iran
44
USSR
56
India
62
China
68
India
74
USSR,
Mongolia,
Singapore
80
China
92
China
96
Australia
110
Japan, USSR,
S. Korea
N. Korea
128
Australia,
New Zealand
140
USSR

Figure 22.3 World allocations of DBS satellite positions

Table 22.3 DBS television channels

Channel number	Frequency (GHz)		
1	11.72748	21	12.11108
2	11.74666	22	12.13026
3	11.76584	23	12.14944
4	11.78502	24	12.16862
5	11.80420	25	12.18780
6	11.82338	26	12.20698
7	11.84256	27	12.22616
8	11.86174	28	12.24534
9	11.88092	29	12.26452
10	11.90010	30	12.28370
11	11.91928	31	12.30288
12	11.93846	32	12.32206
13	11.95764	33	12.34124
14	11.97682	34	12.36042
15	11.99600	35	12.37960
16	12.01518	36	12.39878
17	12.03436	37	12.41796
18	12.05354	38	12.43714
19	12.07272	39	12.45632
20	12.09190	40	12.47550

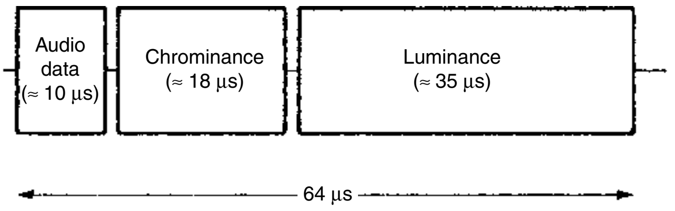

transmission, and they must be reconstituted at the receiver. Two main variations of the MAC format have been selected for European broadcasters, D-MAC and D2-MAC. They differ basically in the number of data channels, and hence the overall bandwidth required. Both modulate video signals in FM, and data signals in duobinary FM.

22.4 Global positioning system (GPS)

This system uses an American government satellite network comprised of 24 satellites in three circular orbits at a height of 20 000 km and with twelve-hour periods. At least six satellites should be visible from any point around the earth at any one time.

Table 22.4 Direct broadcast by satellite (DBS) television. European allocation of satellite positions, channel and polarizations

		Orbital position		
Band	*5° E*	*19° W*	*31° W*	*37° W*
11.7–12.1 GHz RH polarized	*Turkey:* CH. 1, 5, 9, 13, 17 *Greece:* CH. 3, 7, 11, 15, 19	*France:* CH. 1, 5, 9, 13, 17 *Luxembourg:* CH. 3, 7, 11, 15, 19	*Eire:* CH. 2, 6, 10, 14, 18 *UK:* CH. 4, 8, 12, 16, 20	*San Marino:* CH. 1, 5, 9, 13, 17 *Lichtenstein:* CH. 3, 7, 11, 15, 19
11.7–12.1 GHz LH polarized	*Finland:* CH. 2, 6, 10 *Norway:* CH. 14, 18 *Sweden* CH. 4, 8 *Denmark:* CH. 12, 16, 20	*Germany:* CH. 2, 6, 10, 14, 18 *Austria:* CH. 4, 8, 12, 16, 20	*Portugal:* CH. 3, 7, 11, 15, 19	*Andorra:* CH. 3, 8, 12, 16, 20
12.1–12.5 GHz RH polarized	*Cyprus:* CH. 21, 25, 29, 33, 37 *Iceland, etc.:* CH. 23, 27, 31, 35, 39	*Belgium:* CH. 21, 25, 29, 33, 37 *Netherlands:* CH. 23, 27, 31, 35, 39		*Monaco:* CH. 21, 25, 29, 33, 37 *Vatican* CH. 23, 27, 31, 35, 39
12.1–12.5 GHz LH polarized	*Nordic group** CH. 22, 24, 26, 28 30, 32, 36, 40 *Sweden:* CH. 34 *Norway:* CH. 38	*Switzerland:* CH. 22, 26, 30, 34, 38 *Italy:* CH. 24, 28, 32, 36, 40	*Iceland:* CH. 21, 25, 29, 33, 37 *Spain:* CH. 23, 27, 31, 35, 39	

*Wide beam channels: Denmark, Finland, Norway, Sweden.

Each satellite transmits continuously updated information about its orbit on two frequencies, 1227 MHz and 1575 MHz. One radio channel carries two pseudo-random codes, one very long and the other very short. The second channel is modulated only with the short code. The codes enable the satellite to be positively identified, and the distance from a receiver on, or close to, the earth to be calculated.

A receiver's position, in three-dimensional space, is identified by measuring and calculating the distance from three satellites. The short codes provide the initial fix but increased precision is obtained from the long codes. There is an error due to time variations arising from various sources of which satellite speed is one, but taking a measurement from a fourth satellite enables a correction factor to be applied.

Receivers which utilize only the short PN codes provide a resolution accurate to about 100 m. Those which can process the long codes provide a fix accurate to about 45 m.

References

Lewis, G.E. (1988). *Communication Services via Satellite*. BSP Professional Books, Oxford.

Long, M. *The 1993/1994 World Satellite Annual*. Mark Long Enterprises Inc., Ft Lauderdale, USA.

23 Connectors and interfaces

23.1 Audio and video connectors

Audio connectors

The DIN standards devised by the German Industrial Standards Board are widely used for the connection of audio equipment. The connectors are shown below. The 3-way and 5-way 45 are the most common and connections for those are listed.

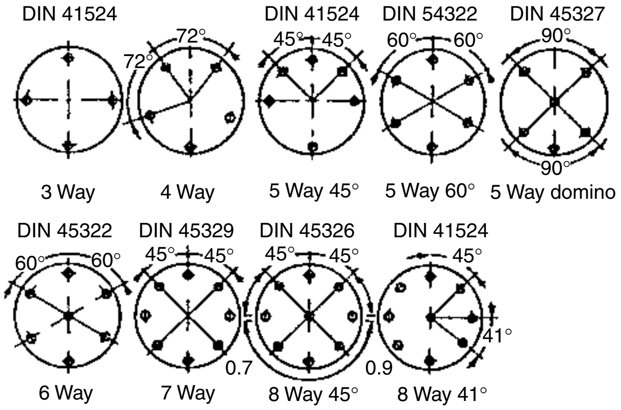
DIN 41524
72°
72°
DIN 41524
45°
45°
DIN 54322
60°
60°
DIN 45327
90°
90°
3 Way
4 Way
5 Way 45°
5 Way 60°
5 Way domino
DIN 45322
60°
60°
DIN 45329
45°
45°
DIN 45326
45°
45°
DIN 41524
45°
41°
0.7
0.9
6 Way
7 Way
8 Way 45°
8 Way 41°

	Mono	*Stereo*
Microphone	Input 1 O V 2 Pin 3 available for polarizing voltage	Input LH 1 Input RH 4 O V 2 Pins 3 and 5 available for polarizing voltage
Tape recorder inputs and monitor outputs	Input 1 O V 2 Output 3	Input LH 1 Input R4 4 O V 2 Output LH 3 Output RH 5

	Mono	*Stereo*
Tape recorder replay output	Output, low Z1 O V 2 Output, high Z3	Output LH, low Z1 Output RH, low Z4 O V 2 Output LH, high Z3 Output RH, high Z5
Amplifiers	Output to tape 1 O V 2 Input from tape 3	Output LH 1 Output RH 1 O V 2 Input LH 3 Input RH 5

Variations on the above exist between different manufacturers.

Videorecorder/televisions/camera connectors

Standard pin configurations for videorecorders, televisions and videocameras are shown below. Many follow standard DIN connector pinouts, but videocamera and SCART connectors differ significantly.

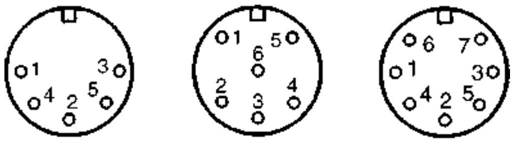

(a) 5-pin DIN

1,4 Audio in
2 Chassis
3,5 Audio out or stereo
1 L audio in
2 Chassis
3 L audio out
4 R audio in
5 R audio out

(b) 6-pin DIN

1 AV select, For VCR high = all outputs, low = all inputs, Opposite for TV set
2 Video in/out
3 Chassis
4 L audio in/out
5 12 V
6 R audio in/out

(c) 7-pin DIN

1 L audio in
2 Chassis
3 L audio out
4 R audio in
5 R audio out
6 Remote control data
7 Chassis

SCART (BS 6552)

SCART connectors, also known as Peritelevision or Euroconnector connectors, feature two control systems which allow remote control over the television's or videorecorder's functions.

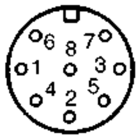

(d) 8-pin DIN

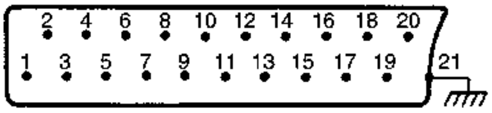

(e) SCART

1 L audio in
2 Remote control data
3 R audio in
4 Chassis (audio)
5 Chassis (remote control)
6 Chassis (video)
7 Chassis (audio)
8 Video in

1 R audio out
2 R audio in
3 L audio out
4 Chassis (audio)
5 Chassis (B video)
6 L video in
7 B video in
8 Source switching
9 Chassis (G video)
10 Data bus
11 G video in
12 Data bus
13 Chassis (data bus)
14 Chassis (data bus)
15 R video in
16 Fast video blanking
17 Chassis (composite video)
18 Chassis (fast video blanking)
19 Composite video out
20 Composite video in

The simplest is a source switching input (pin 8) in which an external source (videorecorder, computer, etc.) can, by issuing a 12 volt signal, cause the television to switch to baseband inputs.

A more complex control system, called *domestic data bus* (D^2B), is given through pins 10 and 12, in which serial data can be passed between controlling microprocessors in the television and external equipment. No standard yet exists for D^2B.

23.2 Co-axial connector

The most commonly used connectors for RF cables are:

Type	*Impedance, Z_0 (ohms)*	*Max. VSWR to (frequency GHz)*	*Maximum proof RF voltage*	*Notes*
N	50 or 75	1.30 (4)	1.5 kV (5 MHz at sea level)	Screw together

Type	*Impedance, Z_0 (ohms)*	*Max. VSWR to (frequency GHz)*	*Maximum proof RF voltage*	*Notes*
BNC	50 or 75	1.01 (1) 1.30 (4)	1 kV (5 MHz at sea level)	Miniature. Bayonet fitting
TNC	50 or 75	As BNC	As BNC	Miniature. Robust screw together
SMB	50	1.46 (4)	1.5 kV (At sea level. Cable dependent)	Sub-miniature. Snap together
SMC	50	1.41 (4) 1.69 (10)	As SMB	Sub-miniature. Screw together
SMD	50	As SMB	As SMB	Sub-miniature. Push together
BT 43	75			Developed from SMB range for use in telecomms and data transmission
7–16	50	1.3 (5 GHz)	2.7 (connector)	Suitable for medium to high power applications in the cellular and broadcast industries. Screw together
PL259/ SO239/ UHF 83	50	–	500 (pk)	Non-constant imp. High VSWR makes it unsuitable for use above 144 MHz and for extending RF cables. Very robust. Screw together
C	75	–	–	Bayonet fitting

Type	*Impedance, Z_0 (ohms)*	*Max. VSWR to (frequency GHz)*	*Maximum proof RF voltage*	*Notes*
F	50	–	–	American CCTV connector used on some 144 MHz hand-portable transceivers. Plugs use inner conductor of cable as centre pin
Belling Lee	50	–	–	British TV antenna connector. Aluminium versions may corrode when used outdoors
GR	50	1 MHz max. frequency	–	Constant imp. sexless connector
Phono	–	–	–	American connector for audio use

Examples of assembly instructions for co-axial RF connectors (by kind permission of M/ACOM Greenpar Ltd.)

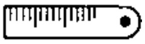

Measuring instrument – a rule is shown, but better results are obtained by using a Vernier gauge

Stout trimming blade, suitable for cutting copper wire braid

Crimping tool

Soldering iron

Side cutters, also for trimming braid

Spanner, of the relevant size for the connector

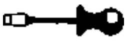

Small screwdriver

Hacksaw, sometimes appropriate for semi-rigid cable, although for repetitive operations a power trimmer should be considered

Assembly instructions Type N

Cable types:

50 ohm: PSF1/4M (BBC), RG 8A/U, RG 213/U, URM 67
75 ohm: RG 11A/U, RG 63B/U, RG 114A/U, URM 64

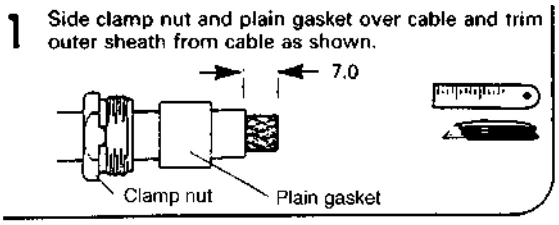

1 Side clamp nut and plain gasket over cable and trim outer sheath from cable as shown.

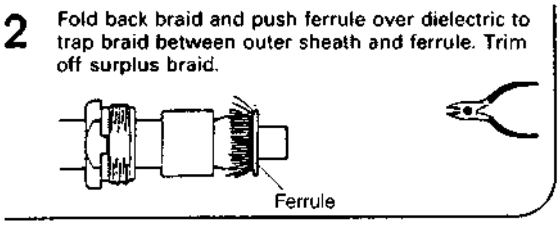

2 Fold back braid and push ferrule over dielectric to trap braid between outer sheath and ferrule. Trim off surplus braid.

3 Trim back dielectric and check the length of the protruding centre conductor.

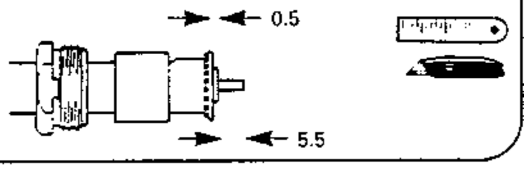

4 Tin centre conductor, then slide rear insulator over dielectric, to butt against ferrule.

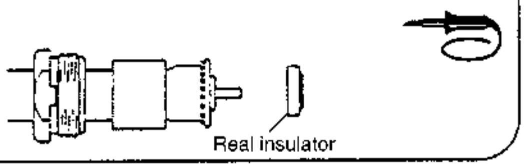

5 Fit contact (male for plugs, female for jacks) onto centre conductor. Hold cable and contact tightly together and solder.

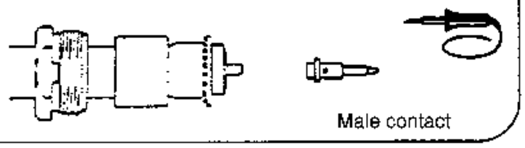

6 Slide plain gasket and clamp nut up to ferrule trapping braid. Fit front insulator over contact to butt against rear insulator and press sub-assembly into body as far as possible.

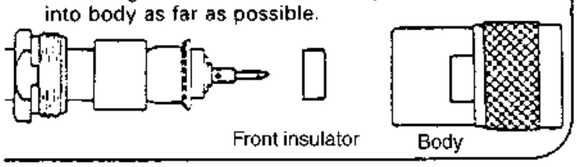

7 Engage and tighten clamp nut.

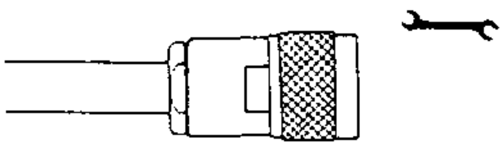

1 Slide crimp sleeve over cable

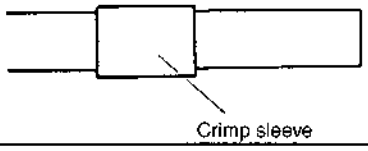

2 Trim back outer sheath and braid to dimensions shown. Any foils adjacent to the dielectric should be left in place but trimmed 1.5 mm from the face of the dielectric

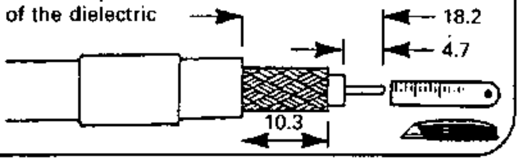

3 Fit contact (male for plugs, female for jacks) over centre conductor to butt against dielectric and crimp.

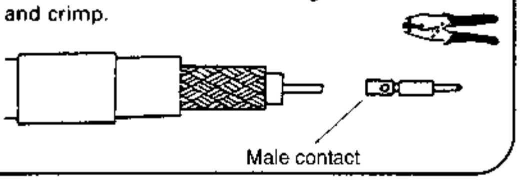

4 Push fit sub-assembly into body until contact clicks into insulator. In all cases, braids (and any intermediate foils) slide over ferrule while any foil adjacent to the dielectric should enter the ferrule bore.

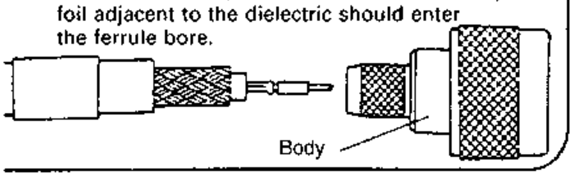

5 Slide crimp sleeve forward over braid and ferrule to butt against rear of the connector and crimp.

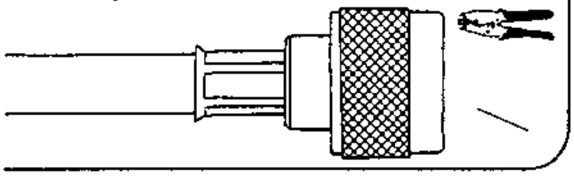

Assembly instructions Type BNC

Cable types:
RG 58C/U, RG 141A/U, URM 43, URM 76

1 Slide clamp nut and plain gasket over cable and trim outer sheath from cable, as shown.

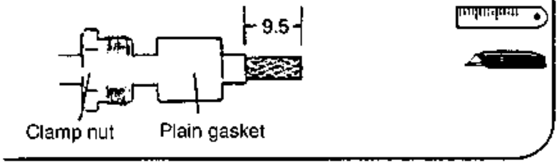

2 Fold back braid and push ferrule over dielectric to trap braid between outer sheath and ferrule. Trim off surplus braid.

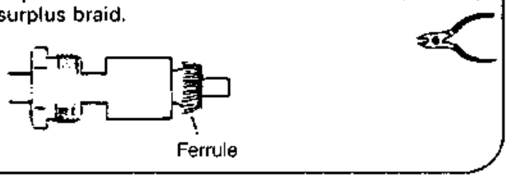

3 Trim back dielectric and check the length of the protruding centre conductor.

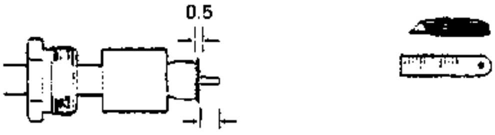

4 Tin centre conductor, then slide rear insulator over dielectric, to butt against ferrule.

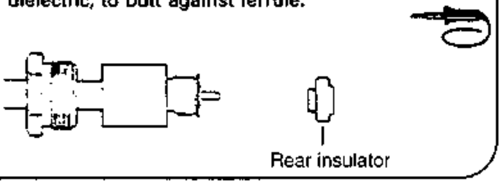

5 Fit contact (male for plugs, female for jacks) onto centre conductor, with collar pressed into recess in rear insulator. Hold cable and contact tightly together, and solder.

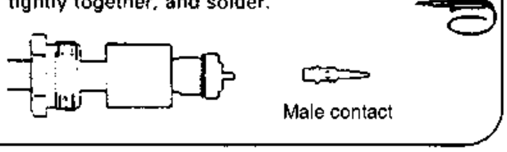

6 Slide plain gasket and clamp nut up to ferrule, trapping braid. Fit front insulator over contact to butt against rear insulator and press sub-assembly into body as far as possible.

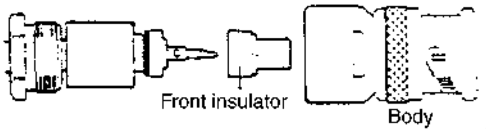
Front insulator
Body

7 **Engage and tighten clamp nut**

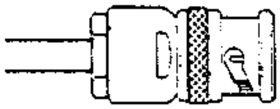

1 Slide metal crimp sleeve over cable, trim outer sheath from cable as shown.

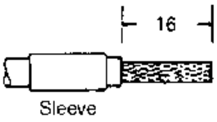

2 Trim back braid and dielectric to the dimensions shown.

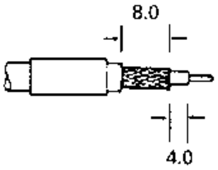

3 Fit contact over centre conductor to butt against the dielectric, then crimp.

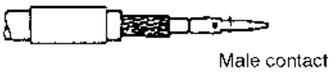
Male contact

4 Press sub-assembly into body, until contact clicks into place and ensuring that the knurled ferrule is inserted between the dielectric and braid.

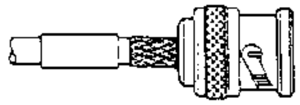

5 Slide the sleeve along the cable, until it butts against the body sub-assembly. Crimp, using the tool listed below.

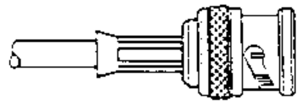

Note: A plug is shown, but these instructions are relevant to both plugs and jacks. The shape of contacts and insulators may also vary from the drawings shown.

Assembly instructions Types SMB/SMC/SMD

Cable types:
TM 3306, RG174A/U, RG188A/U, RG316/U
TM 3263, RG178B/U, RG196A/U, URM 110

1 Slide clamp nut, a washer, a gasket and the other washer over the cable trim outer sheath to dimension shown.

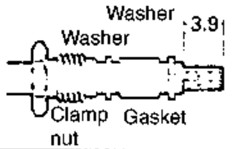

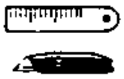

2 Fold back braid. Push ferrule over dielectric to trap braid between outer sheath and ferrule.

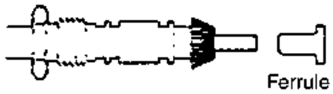

3 Trim off surplus braid. Trim dielectric flush with ferrule and check length of centre conductor. Tin centre conductor.

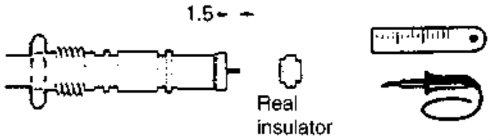

4 Side rear insulator over centre conductor until it butts against ferrule.

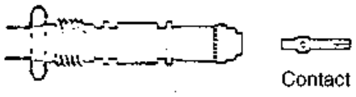

5 Fit contact onto centre conductor until it butts against rear insulator. Hold cable and contact tightly together and solder.

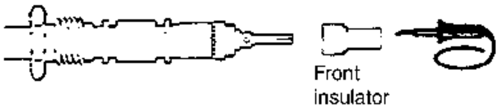

6 Fit front insulator over contact until it butts against internal shoulder.

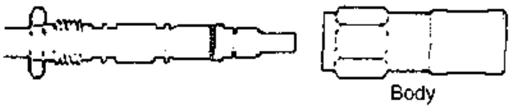

7 Press sub-assembly into body as far as possible. Engage and tighten clamp nut.

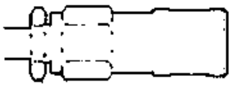

1 Place crimp sleeve over sheat and trim cable to the dimensions shown. Ensure that the centre conductor is not damaged.

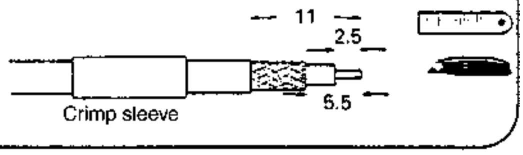

2 Place contact on centre conductor and crimp.

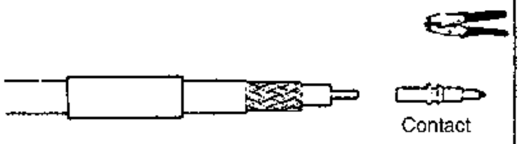

3 Push contact into body sub-assembly. Ensure the contact 'clicks' into rear insulator with the ferrule between the braid and the dielectric.

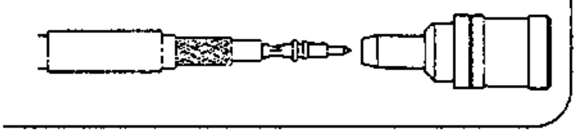

4 Slide crimp sleeve over the braid and crimp.

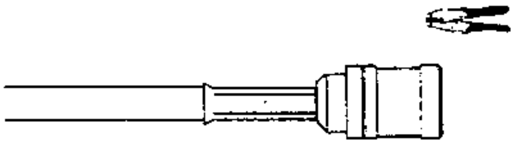

23.3 Interfaces

23.3.1 Connectors and connections

Data interchange by modems

When transmitting and receiving data across telephone or other circuits, the equipment which actually generates and uses the data (e.g. a computer or VDU terminal) is known as *data terminating equipment* (DTE). The equipment which terminates the telephone line and converts the basic data signals into signals which can be transmitted is known as *data circuit-terminating equipment* (DCE). As far as the user is concerned the interface between DTE and DCE is the most important. ITU-T recommendation V24 defines the signal interchanges and functions between DTE and DCE; these are commonly known as the 100 series interchanges circuits:

Interchange circuit		*Data*		*Control*		*Timing*	
Number	*Name*	*From DCE*	*To DCE*	*From DCE*	*To DCE*	*From DCE*	*To DCE*
101	Protective ground or earth						
102	Signal ground or common return						
103	Transmitted data		•				
104	Received data	•					
105	Request to send				•		
106	Ready for sending			•			
107	Data set ready			•			
108/1	Connect data set to line				•		
108/2	Data terminal ready				•		
109	Data channel received line signal detector			•			
110	Signal quality detector			•			
111	Data signalling rate selector (DTE)				•		
112	Data signalling rate selector (DCE)			•			
113	Transmitter signal element timing (DTE)						•

Interchange circuit		*Data*		*Control*		*Timing*	
Number	*Name*	*From DCE*	*To DCE*	*From DCE*	*To DCE*	*From DCE*	*To DCE*
114	Transmitter signal element timing (DCE)					•	
115	Receiver signal element timing (DCE)					•	
116	Select stand by				•		
117	Standby indicator			•			
118	Transmitted backward channel data		•				
119	Received backward channel data	•					
120	Transmit backward channel line signal				•		
121	Backward channel ready			•			
122	Backward channel received line signal detector			•			
123	Backward channel single quality detector			•			
124	Select frequency groups				•		
125	Calling indicator			•			
126	Select transmit frequency				•		
127	Select receive frequency				•		

Interchange circuit		*Data*		*Control*		*Timing*	
Number	*Name*	*From DCE*	*To DCE*	*From DCE*	*To DCE*	*From DCE*	*To DCE*
128	Receiver signal element timing (DTE)						●
129	Request to receive				●		
130	Transmit backward tone				●		
131	Received character timing					●	
132	Return to non-data mode				●		
133	Ready for receiving				●		
134	Received data present			●			
191	Transmitted voice answer				●		
192	Received voice answer			●			

Modem connector pin numbers

The connectors used with 100 series interchange circuits and its pin assignments are defined by international standard ISO 2110 and are (for modems following the ITU-T recommendations V21, V23, V26, V26bis, V27 and V27bis) as follows:

Pin number	Interchange circuit numbers			Pin Name
	V21	V23	V26/ V27	
1	*1	*1	*1	Ground
2	103	103	103	TXD
3	104	104	104	RXD
4	105	105	105	RTS
5	106	106	106	RFS
6	107	107	107	DSR
7	102	102	102	Signal return
8	109	109	109	Signal DET
9	*N	*N	*N	
10	*N	*N	*N	
11	126	*N	*N	STF
12	*F	122	122	
13	*F	121	121	
14	*F	118	118	
15	*F	*2	114	
16	*F	119	119	
17	*F	*2	115	
18	141	141	141	
19	*F	120	120	
20	108/1-2	108/1-2	108/1-2	DTR
21	140	140	140	
22	125	125	125	Call ind.
23	*N	111	111	
24	*N	*N	113	
25	142	142	142	

Notes:

*1 Pin 1 is assigned for connecting the shields between tandem sections of shielded cables. It may be connected to protective ground or signal ground.

*F Reserved for future use.

*N Reserved for national use.

9-Pin connector

Pin number	*Pin name*	*Description*
1	DCD	Data carrier detect
2	RXD	Receive data
3	TXD	Transmit data
4	DTR	Data terminal ready
5	GND	Ground
6	DSR	Data set ready
7	RTS	Ready to send
8	CTS	Clear to send
9	RI	Ring indicator

Automatic calling

A similar series of interchange circuits is defined in ITU-T recommendation V25 for automatic calling answering between modems over the telephone network. This is the 200 series interchange circuits:

Interchange circuit

Number	*Name*	*From DCE*	*To DCE*
201	Signal ground	•	•
202	Call request		•
203	Data line occupied	•	
204	Distant station connected	•	
205	Abandon call	•	
206	Digit signal (2^0)		•
207	Digit signal (2^1)		•
208	Digit signal (2^2)		•
209	Digit signal (2^3)		•
210	Present next digit	•	
211	Digit present		•
213	Power indication	•	

RS 232C

The EIA equivalent of CCITT V24 interface is the RS 232C specification, which similarly defines the electrical interface between DTE and DCE. Although the two have different designations, they are to all practical purposes equivalent. The RS 232C interchange circuits are:

Interchange circuit		*Data*		*Control*		*Timing*	
Mnemonic	*Name*	*From DCE*	*To DCE*	*From DCE*	*To DCE*	*From DCE*	*To DCE*
AA	Protective ground						
AB	Signal ground/common return						
BA	Transmitted data		●				
BB	Received data	●					
CA	Request to send				●		
CB	Clear to send			●			
CC	Data set ready			●			
CD	Data terminal ready				●		
CE	Ring indicator			●			
CF	Received line signal detector			●			
CG	Signal quality detector			●			
CH	Data signal rate selector (DTE)				●		
CI	Data signal rate selector (DCE)			●			
DA	Transmitter signal element timing (DTE)						●
DB	Transmitter signal element timing (DCE)					●	
DD	Receiver signal element timing (DCE)					●	

Interchange circuit		Data		Control		Timing	
Mnemonic	*Name*	*From DCE*	*To DCE*	*From DCE*	*To DCE*	*From DCE*	*To DCE*
SBA	Secondary transmitted data		●				
SBB	Secondary received data	●					
SCA	Secondary request to send				●		
SCB	Secondary clear to send			●			
SCF	Secondary received line signal detector			●			

RS 449

The EIA RS 232C standard, although the most common, is by no means perfect. One of its main limitations is the maximum data rate −18.2 K baud. Various improved interchange circuits (RS 422, RS 423) have been developed. The RS 449 standard is capable of very fast data rates (up to 2 Mbaud):

Interchange circuit		*Data*		*Control*		*Timing*	
Mnemonic	*Name*	*From DCE*	*To DCE*	*From DCE*	*To DCE*	*From DCE*	*To DCE*
SG	Single ground						
SC	Send common						
RC	Receive common						
IS	Terminal in service				●		
IC	Incoming call			●			
TR	Terminal ready				●		

Interchange circuit			*Data*		*Control*		*Timing*	
	Mnemonic	*Name*	*From DCE*	*To DCE*	*From DCE*	*To DCE*	*From DCE*	*To DCE*
Primary channel	DM	Data mode			●			
	SD	Send data		●				
	RD	Receive data	●					
	TT	Terminal timing						●
	ST	Send timing					●	
	RT	Receive timing					●	
	RS	Request to send				●		
	CS	Clear to send			●			
	RR	Receiver ready			●			
	SQ	Signal quality			●			
	NS	News signal				●		
	SF	Select frequency				●		
	SR	Signalling rate selector				●		
	SI	Signalling rate indicator			●			
Secondary channel	SSD	Secondary send data		●				
	SRD	Secondary receive data	●					
	SRS	Secondary request to send				●		
	SCS	Secondary clear to send			●			
	SRR	Secondary receiver ready			●			
	LL	Local loopback				●		
	RL	Remote loopback				●		
	TM	Test mode			●			
	SS	Select standby				●		
	SB	Standby indicator			●			

Centronics interface

Most personal computers use the Centronics parallel data transfer to a printer. The pin connections of the connector, abbreviations and signal descriptions are shown.

All signals are standard TTL, although not all signals necessarily exist in any given interface.

Pin number	*Abbreviation*	*Signal description*
1	STROBE	Strobe
2	DATA1	Data line 1
3	DATA2	Data line 2
4	DATA3	Data line 3
5	DATA4	Data line 4
6	DATA5	Data line 5
7	DATA6	Data line 6
8	DATA7	Data line 7
9	DATA8	Data line 8
10	ACKNLG	Acknowledge data
11	BUSY	Busy
12	PE	Paper end
13	SLCT	Select printer
14	AUTO FEED XT	Automatic line feed at end of line
15	NC	No connection
16	OV	Logic ground
17	CHASSIS GND	Printer chassis (not necessarily the same as logic ground)
18	NC	No connection
19 to 30	GND	Single ground
31	INIT	Initialize
32	ERROR	Error
33	GND	Signal ground
34	NC	No connection
35	Logic 1	Logic 1
36	SLCT IN	Select input to printer

RS 232C/RS 449/V24 comparison

RS 232C		*RS 449*		*V24*	
AB	Signal ground	SG	Signal ground	102	Signal ground
		SC	Send common	102a	DTE common
		RC	Receive common	102b	DCE common
		IS	Terminal in service		
CE	Ring indicator	IC	Incoming call	125	Calling indicator
CD	Data terminal ready	TR	Terminal ready	108/2	Data terminal ready
CC	Data set ready	DM	Data mode	107	Data set ready
BA	Transmitted data	SD	Send data	103	Transmitted data
BB	Receive data	RD	Receive data	104	Received data
DA	Transmitter signal element Timing (DTE source)	TT	Terminal timing	113	Transmitter signal element Timing (DTE source)
DB	Transmitter signal element Timing (DCE source)	ST	Send timing	114	Transmitter signal element Timing (DCE source)
DD	Receiver signal element timing	RT	Receive timing	115	Receiver signal element timing (DCE source)
CA	Request to send	RS	Request to send	105	Request to send
CB	Clear to send	CS	Clear to send	106	Ready for sending
CF	Received line single detector	RR	Receiver ready	109	Data channel received line signal detector

CG	Signal quality detector	SQ	Signal quality	110	Data signal quality detector
		NS	New signal		
		SF	Select frequency	126	Select transmit frequency
CH	Data signal rate selector (DTE source)	SR	Signalling rate selector	111	Data signalling rate selector (DTE source)
CI	Data signal rate selector (DCE source)	SI	Signalling rate indicator	112	Data signalling rate selector (DCE source)
SBA	Secondary transmitted data	SSD	Secondary send data	118	Transmitted backward channel data
SBB	Secondary received data	SRD	Secondary receive data	119	Received backward channel data
SGA	Secondary request to send	SRS	Secondary request to send	120	Transmit backward channel line signal
SCB	Secondary clear to send	SCS	Secondary clear to send	121	Backward channel ready
SCF	Secondary received line signal detector	SRR	Secondary receiver ready	122	Backward channel received line signal detector
		LL	Local loopback	141	Local loopback
		RL	Remote loopback	140	Remote loopback
		TM	Test mode	143	Test indicator
		SS	Select standby	116	Select standby
		SB	Standby indicator	117	Standby indicator

Reference

M/ACOM Greenpar Ltd, catalogue 1993.

24 Broadcasting

24.1 Standard frequency and time transmissions

Frequency (mHz)	*Wavelength (m)*	*Code*	*Station location*	*Country*	*Power (kW)*
60 kHz	5000	MSF	Rugby	England	–
–	–	WWVB	Colorado	USA	–
75 kHz	4000	HBG	–	Switzerland	–
77.5 kHz	3871	DCF77	Mainflingen	DDR	–
1.5	200	HD210A	Guayaquil	Ecuador	–
2.5	120	MSF	Rugby	England	0.5
–	–	WWV	Fort Collins	USA	2.5
–	–	WWVH	Kekaha	Hawaii	5
–	–	ZLF	Wellington	New Zealand	–
–	–	RCH	Tashkent	USSR	1
–	–	JJY	–	Japan	–
–	–	ZUO	Olifantsfontein	South Africa	–
3.33	90.09	CHU	Ottawa	Canada	3
3.81	78.7	HD201A	Guayaquil	Ecuador	–
4.5	66.67	VNG	Victoria	Australia	–
4.996	60.05	RWM	Moscow	USSR	5
5	60	MSF	Rugby	England	0.5
–	–	WWVB	Fort Collins	USA	10
–	–	WWVH	Kekaha	Hawaii	10
–	–	ATA	New Delhi	India	–
–	–	LOL	Buenos Aires	Argentina	2
–	–	IBF	Turin	Italy	5
–	–	RCH	Tashkent	USSR	1
–	–	JJY	–	Japan	–
–	–	ZUO	Olifantsfontein	South Africa	–
5.004	59.95	RID	Irkutsk	USSR	1
6.10	49.2	YVTO	Caracas	Venezuela	–
7.335	40.9	CHU	Ottawa	Canada	10
7.5	40	VNG	Lyndhurst	Australia	5

Frequency (mHz)	*Wavelength (m)*	*Code*	*Station location*	*Country*	*Power (kW)*
7.6	39.4	HD210A	Guayaquil	Ecuador	–
8	37.5	JJY	–	Japan	–
8.1675	36.73	LQB9	Buenos Aires	Argentina	5
9996	30.01	RWM	Moscow	USSR	–
10	30	MSF	Rugby	England	0.5
–	–	WWVB	Fort Collins	USA	10
–	–	WWVH	Kekaha	Hawaii	10
–	–	BPM	Xian	China	–
–	–	ATA	New Delhi	India	–
–	–	JJY	–	Japan	–
–	–	LOL	Buenos Aires	Argentina	2
–	–	RTA	Novosibirsk	USSR	5
–	–	RCH	Tashkent	USSR	1
10.004	29.99	RID	Irkutsk	USSR	1
12	25	VNG	Lyndhurst	Australia	10
14.67	20.45	CHU	Ottawa	Canada	3
14.996	20.01	RWM	Moscow	USSR	8
15	20	WWVB	Fort Collins	USA	10
–	–	WWVH	Kekaha	Hawaii	10
–	–	LOL	Buenos Aires	Argentina	2
–	–	RTA	Novosibirsk	USSR	5
–	–	BPM	Xian	China	–
–	–	ATA	New Delhi	India	–
–	–	JJY	–	Japan	–
15.004	19.99	RID	Irkutsk	USSR	1
16.384	18.31	–	Allouis	France	2000
15.55	17.09	LQC20	Buenos Aires	Argentina	5
20	15	WWVB	Fort Collins	USA	2.5
100	3	ZUO	Olifantsfontein	South Africa	–

24.2 Standard frequency formats

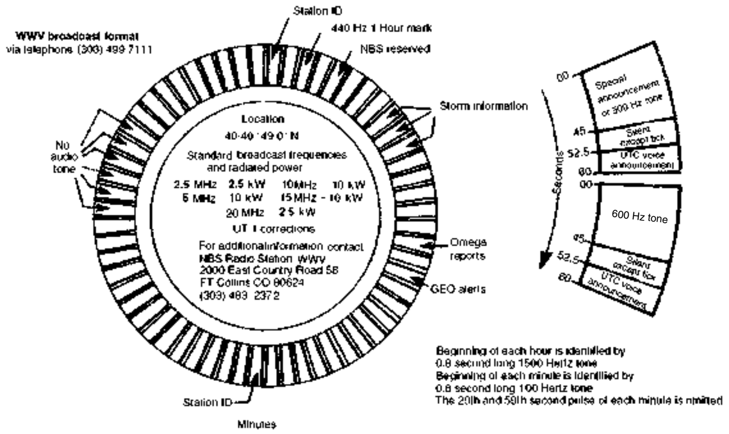
WWV broadcast format
via telephone (303) 499 7111
Station ID
440 Hz 1 Hour mark
NBS reserved
Storm information
No audio tone
Location
40·40 ' 49·0 ' N
Standard broadcast frequencies and radiated power
2.5 MHz 2.5 kW 10 MHz 10 kW
5 MHz 10 kW 15 MHz - 10 kW
20 MHz 2.5 kW
UT 1 corrections
For additional information contact
NBS Radio Station WWV
2000 East Country Road 58
FT Collins CO 80624
(303) 483 2372
Omega reports
GEO alerts
Station ID
Minutes
Seconds
00
45
52.5
60
Special announcement or 500 Hz tone
Silent except tick
UTC voice announcement
600 Hz tone

Beginning of each hour is identified by
0.8 second long 1500 Hertz tone
Beginning of each minute is identified by
0.8 second long 100 Hertz tone
The 29th and 59th second pulse of each minute is omitted

MSF Rugby

Time is inserted in the 60 kHz transmission in two ways, illustrated below.

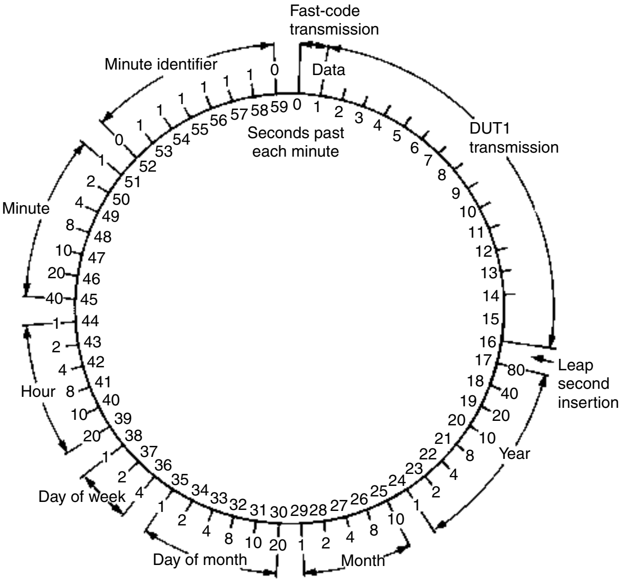
Fast-code transmission
Data
Minute identifier
DUT1 transmission
Seconds past each minute
Minute
Hour
Day of week
Day of month
Month
Year
Leap second insertion
0 1 1 1 1 1 1 0
1 2 4 8 10 20 40
1 2 4 8 10 20
1 2 4
1 2 4 8 10 20
1 2 4 8 10
1 2 4 8 10 20 40 80
59 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

Slow code time and date information is transmitted between the 17th and 59th seconds of the minute-long cycle, in normal BCD coding. *Fast code* time and date BCD coded information is inserted into a 500 ms window in the first second of each minute-long cycle, as illustrated below.

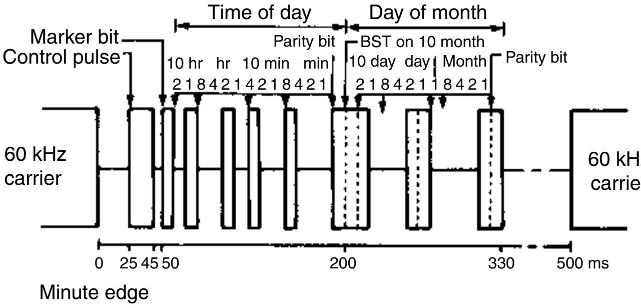
Time of day
Day of month
Marker bit
Control pulse
Parity bit
BST on 10 month
Parity bit
10 hr hr 10 min min
10 day day Month
2 1 8 4 2 1 4 2 1 8 4 2 1
2 1 8 4 2 1 1 8 4 2 1
60 kHz carrier
60 kHz carrier
0 25 45 50 200 330 500 ms
Minute edge

24.3 UK broadcasting bands

Frequency band	*Frequency*	*Use*
Long wave	150–285 kHz (2000–1053 m)	AM radio
Medium wave	525–1605 kHz (571–187 m)	AM radio
Band II (VHF)	88–108 MHz	FM radio
Band IV (UHF)	470–582 MHz (channels 21 to 34)	TV
Band V (UHF)	614–854 MHz (channels 39 to 68)	TV
Band VI (SHF)	11.7–12.5 GHz (channels 1 to 40)	Satellite TV

24.4 BBC VHF test tone transmissions

Transmission starts about 4 minutes after the end of Radio 3 programmes on Mondays and Saturdays.

Time min.	*Left channel*	*Right channel*	*Purpose*
–	250 Hz at zero level	440 Hz at zero level	Identification of left and right channels and setting of reference level
2	900 Hz at +7 dB	900 Hz at +7 dB, antiphase to left channel	Adjustment of phase of regenerated subcarrier (see Note 4) and check of distortion with L−R signal only
6	900 Hz at +7 dB	900 Hz +7 dB, in phase with left channel	Check of distortion with L + R signal only
7	900 Hz at +7 dB	No modulation	Check of L to R cross-talk
8	No modulation	900 Hz at +7 dB	Check of R to L cross-talk
9	Tone sequence at −4 dB: 40 Hz 6·3 kHz 100 Hz 10 kHz 500 Hz 12.5 kHz 1000 Hz 14 kHz This sequence is repeated	No modulation	Check of L-channel frequency response and L to R cross-talk at high and low frequencies
11′40″	No modulation	Tone sequence as for left channel	Check of R-channel frequency response and R to L cross-talk at high and low frequencies

Time min.	*Left channel*	*Right channel*	*Purpose*
14′20″	No modulation	No modulation	Check of noise level in the presence of pilot
15′20″	End of test transmissions		

Notes:

1. This schedule is subject to variation or cancellation to accord with programme requirements and essential transmission tests.
2. The zero level reference corresponds to 40% of the maximum level of modulation applied to either stereophonic channel before pre-emphasis. All tests are transmitted with pre-emphasis.
3. Periods of tone lasting several minutes are interrupted momentarily at one-minute intervals.
4. With receivers having separate controls of subcarrier phase and crosstalk, the correct order of alignment is to adjust first the subcarrier phase to produce maximum output from either the L or the R channel and then to adjust the crosstalk (or ‘separation’) control for minimum crosstalk between channels.
5. With receivers in which the only control of crosstalk is by adjustment of subcarrier phase, this adjustment of subcarrier phase, this adjustment should be made on the crosstalk checks.
6. Adjustment of the balance control to produce equal loudness from the L and R loudspeakers is best carried out when listening to the announcements during a stereophonic transmission, which are made from a centre-stage position. If this adjustment is attempted during the tone transmissions, the results may be confused because of the occurrence of standing-wave patterns in the listening room.
7. The outputs of most receivers include significant levels of the 19-kHz tone and its harmonics, which may affect signal-level meters. It is important, therefore, to provide filters with adequate loss at these frequencies if instruments are to be used for the above tests.

24.5 Engineering information about broadcast services

Information about all BBC services as well as advice on how best to receive transmissions (including television) can be obtained from:

British Broadcasting Corporation
Engineering Liaison
White City
201 Wood Lane
London W12 7TS

Telephone: 020 8752 5040

Transmitter service maps for most main transmitters can also be supplied, but requests for maps should be accompanied by a stamped addressed A4 sized envelope.

Similarly, information about all IBA broadcast services can be obtained from:

Radio:
Radiocommunications Agency
Wyndham House
189 Marsh Wall
London E14 9SX

Telephone: 020 7211 0211

Television:
The Independent Television Commission
Kings Worthy Court
Kings Worthy
Winchester
Hants SO23 7QA

Telephone: 01962 848647

24.6 Characteristics of UHF terrestrial television systems

24.6.1 World systems

System	*Number of lines*	*Channel width (MHz)*	*Vision bandwidth (MHz)*	*Vision/sound separation (MHz)*	*Vestigial side-band (MHz)*	*Vision modulation*	*Sound modulation*	*Field frequency*
A	625	8	5.5	+6	1.25	Neg.	FM	50
B	625	8	5.0	+5.5	1.25	Neg.	FM	50
C	625	8	5.0	+5.5	0.75	Neg.	FM	50
D	625	8	6.0	+6.5	1.25	Pos.	AM	50
E	625	8	6.0	+6.5	0.75	Neg.	FM	50
F	525	6	4.2	+4.5	1.25	Neg.	FM	60

A – UK and Eire
B – Eastern Europe
C – Most of Western Europe, Australia, New Zealand
D – France
E – Russia and Eastern Europe
F – USA, most of Central and South America, Japan

24.6.2 European systems

Country	*System*	*Colour*
Austria	C	PAL
Belgium	B	PAL
Bulgaria	No UHF system	
Cyprus	C	PAL
Czechoslovakia	E	SECAM
Denmark	No UHF system	PAL
Finland	C	PAL
France	D	SECAM
Germany	C	PAL

Country	*System*	*Colour*
German DR	C	SECAM
Greece	B	PAL
Holland	C	PAL
Hungary	E	SECAM
Iceland	No UHF system	
Ireland	A	PAL
Italy	C	PAL
Luxembourg	D	SECAM
Malta	B	PAL
Monaco	D	SECAM
Norway	C	PAL
Poland	E	SECAM
Portugal	C	PAL
Romania	E	PAL
Spain	C	PAL
Sweden	C	PAL
Switzerland	C	PAL
Turkey	No UHF system	
UK	A	PAL
USSR	E	SECAM
Yugoslavia	B	PAL

24.6.3 UK 625-line television system specification

Channel bandwidth	8 MHz
Upper sideband (vision signal)	5.5 MHz
Lower sideband (vision signal)	1.25 MHz
Vision modulation	AM negative
Sound modulation	FM
Sound deviation (max.)	±50 kHz
Sound pre-emphasis	50 μs
Sound carrier relative to vision carrier	+
Aspect ratio	4:3
Blanking and black level	76%
White level	20% peak
Sync. level	100% peak
Video bandwidth	5.5 MHz

Field frequency	50 Hz
Line frequency	15 625 Hz
Field sync. signal	5 equalizing then 5 broad pulses, followed by 5 equalizing pulses in 7.5 line periods
Field sync. and flyback intervals	2 × 25 line periods
Line period (approx.)	64 μs
Line syn. pulses (approx.)	4.7 μs
Line blanking (approx.)	12 μs
Field sync. pules (broad)	27.3 μs
Field sync. pulses (equalizing)	2.3 μs
Colour subcarrier frequency	4.43361875 MHz
Burst duration	2.25 μs
Burst amplitude	equal to sync
Burst phase	180° ± 45°

24.6.4 UK 625-line television system field blanking details

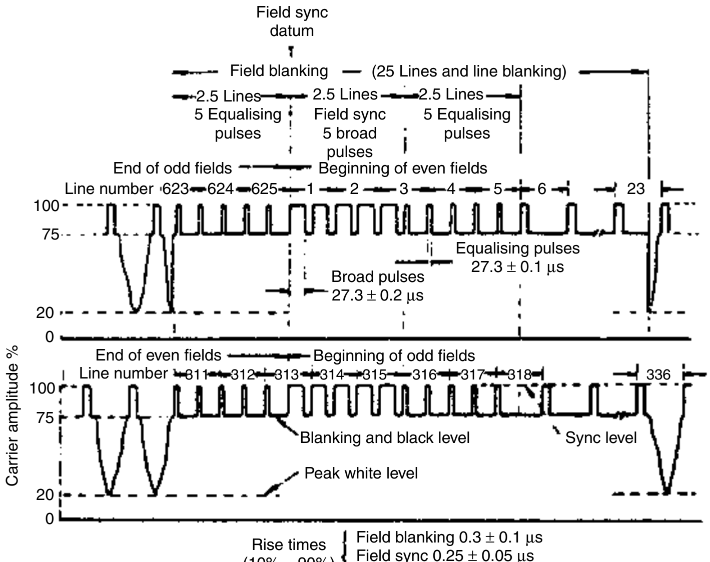
Field sync
datum
Field blanking — (25 Lines and line blanking)
2.5 Lines
5 Equalising
pulses
2.5 Lines
Field sync
5 broad
pulses
2.5 Lines
5 Equalising
pulses
End of odd fields
Beginning of even fields
Line number 623 624 625 1 2 3 4 5 6 23
100
75
20
0
Equalising pulses
27.3 ± 0.1 μs
Broad pulses
27.3 ± 0.2 μs
Carrier amplitude %
End of even fields
Beginning of odd fields
Line number 311 312 313 314 315 316 317 318 336
100
75
20
0
Blanking and black level
Sync level
Peak white level
Rise times
(10% – 90%)
Field blanking 0.3 ± 0.1 μs
Field sync 0.25 ± 0.05 μs
Equalising pulses 0.25 ± 0.05 μs

24.7 Terrestrial television channels

UK

	Frequency (MHz)			Frequency (MHz)	
Channel	*Vision*	*Sound*	*Channel*	*Vision*	*Sound*
39	615.25	621.25	54	735.25	741.25
40	623.25	629.25	55	743.25	749.25
41	631.25	637.25	56	751.25	757.25
42	639.25	645.25	57	759.25	765.25
43	647.25	653.25	58	767.25	773.25
44	655.25	661.25	59	775.25	781.25
45	663.25	669.25	60	783.25	789.25
46	671.25	677.25	61	791.25	797.25
47	679.25	685.25	62	799.25	805.25
48	687.25	693.25	63	807.25	813.25
49	695.25	701.25	64	815.25	821.25
50	703.25	709.25	65	823.25	829.25
51	711.25	717.25	66	831.25	837.25
52	719.25	725.25	67	839.25	845.25
53	727.25	733.25	68	847.25	853.25

Republic of Ireland

	Frequency (MHz)			Frequency (MHz)	
Channel	*Vision*	*Sound*	*Channel*	*Vision*	*Sound*
IA	45.75	51.75	IF	191.25	197.25
IB	53.75	59.75	IG	199.25	205.25
IC	61.75	67.75	IH	207.25	213.25
ID	175.25	81.25	IJ	215.25	221.25
IE	183.25	189.25			

South Africa

	Frequency (MHz)			*Frequency (MHz)*	
Channel	*Vision*	*Sound*	*Channel*	*Vision*	*Sound*
4	175.25	181.25	9	215.25	221.25
5	183.25	189.25	10	223.25	229.25
6	191.25	197.25	11	231.25	237.25
7	199.25	205.25	13	247.43	253.43
8	207.25	213.25			

Australia

	Frequency (MHz)			*Frequency (MHz)*	
Channel	*Vision*	*Sound*	*Channel*	*Vision*	*Sound*
0	46.25	51.75	6	175.25	180.75
1	57.25	62.75	7	182.25	187.75
2	64.25	69.75	8	189.25	194.75
3	86.25	91.75	9	196.25	201.75
4	95.25	100.75	10	209.25	214.75
5	102.25	107.75	11	216.25	221.75
5A	138.25	143.75			

New Zealand

	Frequency (MHz)			*Frequency (MHz)*	
Channel	*Vision*	*Sound*	*Channel*	*Vision*	*Sound*
1	45.25	50.75	6	189.25	194.75
2	55.25	60.75	7	196.25	201.75
3	62.25	67.75	8	203.25	208.75
4	175.25	180.75	9	210.25	215.75
5	182.25	187.75			

USA

Channel	*Frequency (MHz)* *Vision*	*Frequency (MHz)* *Sound*	*Channel*	*Frequency (MHz)* *Vision*	*Frequency (MHz)* *Sound*
2	55.25	59.75	43	645.25	649.75
3	61.25	65.75	44	651.25	655.75
4	67.25	71.75	45	657.25	661.75
5	77.25	81.75	46	663.25	667.75
6	83.25	87.75	47	669.25	673.75
7	175.25	179.75	48	675.25	679.75
8	181.25	185.75	49	681.25	685.75
9	187.25	191.75	50	687.25	691.75
10	193.25	197.75	51	693.25	697.75
11	199.25	203.75	52	699.25	703.75
12	205.25	209.75	53	705.25	709.75
13	211.25	215.75	54	711.25	715.75
14	471.25	475.75	55	717.25	721.75
15	477.25	481.75	56	723.25	727.75
16	483.25	487.75	57	729.25	733.75
17	489.25	493.75	58	735.25	739.75
18	495.25	499.75	59	741.25	745.75
19	501.25	505.75	60	747.25	751.75
20	507.25	511.75	61	753.25	757.75
21	513.25	517.75	62	759.25	763.75
22	519.25	523.75	63	765.25	769.75
23	525.25	529.75	64	771.25	775.75
24	531.25	535.75	65	777.25	781.75
25	537.25	541.75	66	783.25	787.75
26	543.25	547.75	67	789.25	793.75
27	549.25	553.75	68	795.25	799.75
28	555.25	559.75	69	801.25	805.75
29	561.25	565.75	70	807.25	811.75
30	567.25	571.75	71	813.25	817.75
31	573.25	577.75	72	819.25	823.75
32	579.25	583.75	73	825.25	829.75
33	585.25	589.75	74	831.25	835.75
34	591.25	595.75	75	837.25	841.75
35	597.25	601.75	76	843.25	847.75
36	603.25	607.75	77	849.25	853.75
37	609.25	613.75	78	855.25	859.75
38	615.25	619.75	79	861.25	865.75
39	621.25	625.75	80	867.25	871.75
40	627.25	631.75	81	873.25	877.75
41	633.25	637.75	82	879.25	883.75
42	639.25	643.75	83	885.25	889.75

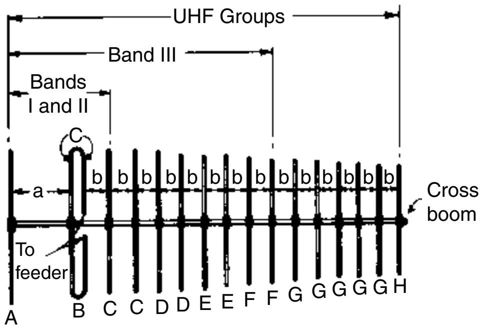

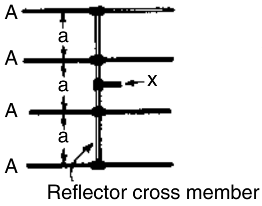

Pattern of general-purpose Yagi array to be used in conjunction with the dimensions given below.

24.9 AM broadcast station classes (USA)

The US AM broadcast band is 540 kHz to 1700 kHz, with 10 kHz channel spacings with centre frequencies divisible by ten (e.g. 780 kHz or 1540 kHz). Other countries in the western hemisphere operate on either 10 kHz channel spacings with frequencies ending in ‘5’, or 9 kHz spacing. The Domestic Class is generally the class of station defined in 47 CFR Section 73.21. The Region 2 Class is generally the class of station as defined in the Region 2 [Western Hemisphere] AM Agreement. This class also corresponds to the class in the 1984 US–Canadian AM Agreement and the 1986 US–Mexican Agreement.

24.9.1 Class A station

A Class A station is an unlimited time station (that is, it can broadcast 24 hours per day) that operates on a clear channel. The operating power shall not be less than 10 kilowatts (kW) or more than 50 kW.

24.9.2 Class B station

A Class B station is an unlimited time station. Class B stations are authorized to operate with a minimum power of 250 watts and a maximum power of 50 kW. (If a Class B station operates with less than 250 W, the RMS must be equal to or greater than 141 mV/m at 1 km for the actual power.) If the station is authorized to operate in the expanded band (1610 to 1700 kHz), the maximum power is 10 kW.

24.9.3 Class C station

A Class C station is an unlimited time station that operates on a local channel. The power shall not be less than 250 W nor more than 1 kW. Class C stations that are licensed to operate with 100 W may continue to operate as licensed.

24.9.4 Class D station

A Class D station operates either daytime, limited time, or unlimited time with a night-time power less than 250 W and an equivalent RMS antenna field less than 141 mV/m at 1 km for the actual power. Class D stations shall operate with daytime powers not less than 0.250 kW nor more than 50 kW. *Note*: If a station is an existing daytime-only station, its class will be Class D.

24.10 FM broadcast frequencies and channel numbers (USA)

From US Code USC 47 CFR 73. The FM broadcast band consists of that portion of the radio frequency spectrum between 88 MHz and 108 MHz. It is divided into 100 channels of 200 kHz each. For convenience, the frequencies available for FM broadcasting (including those assigned to non-commercial educational broadcasting) are given numerical designations which are shown in the table below:

Frequency (MHz)	*Channel No.*
88.1	201
88.3	202
88.5	203
88.7	204
88.9	205
89.1	206
89.3	207
89.5	208
89.7	209
89.9	210
90.1	211
90.3	212
90.5	213

Frequency (MHz)	*Channel No.*
90.7	214
90.9	215
91.1	216
91.3	217
91.5	218
91.7	219
91.9	220
92.1	221
92.3	222
92.5	223
92.7	224
92.9	225
93.1	226
93.3	227
93.5	228
93.7	229
93.9	230
94.1	231
94.3	232
94.5	233
94.7	234
94.9	235
95.1	236
95.3	237
95.5	238
95.7	239
95.9	240
96.1	241
96.3	242
96.5	243
96.7	244
96.9	245
97.1	246
97.3	247
97.5	248
97.7	249
97.9	250
98.1	251
98.3	252
98.5	253

Frequency (MHz)	*Channel No.*
98.7	254
98.9	255
99.1	256
99.3	257
99.5	258
99.7	259
99.9	260
100.1	261
100.3	262
100.5	263
100.7	264
100.9	265
101.1	266
101.3	267
101.5	268
101.7	269
101.9	270
102.1	271
102.3	272
102.5	273
102.7	274
102.9	275
103.1	276
103.3	277
103.5	278
103.7	279
103.9	280
104.1	281
104.3	282
104.5	283
104.7	284
104.9	285
105.1	286
105.3	287
105.5	288
105.7	289
105.9	290
106.1	291
106.3	292
106.5	293

Frequency (MHz)	*Channel No.*
106.7	294
106.9	295
107.1	296
107.3	297
107.5	298
107.7	299
107.9	300

24.11 US television channel assignments

Channel No.	*Frequency (MHz)*
1	(No Ch. 1 assigned)
2	54–60
3	60–66
4	66–72
5	76–82
6	82–88
7	174–180
8	180–186
9	186–192
10	192–198
11	198–204
12	204–210
13	210–216
14	470–476
15	476–482
16	482–488
17	488–494
18	494–500
19	500–506
20	506–512
21	512–518
22	518–524
23	524–530
24	530–536
25	536–542
26	542–548
27	548–554

Channel No.	*Frequency (MHz)*
28	554–560
29	560–566
30	566–572
31	572–578
32	578–584
33	584–590
34	590–596
35	596–602
36	602–608
37	608–614
38	614–620
39	620–626
40	626–632
41	632–638
42	638–644
43	644–650
44	650–656
45	656–662
46	662–668
47	668–674
48	674–680
49	680–686
50	686–692
51	692–698
52	698–704
53	704–710
54	710–716
55	716–722
56	722–728
57	728–734
58	734–740
59	740–746
60	746–752
61	752–758
62	758–764
63	764–770
64	770–776
65	776–782
66	782–788
67	788–794

Channel No.	*Frequency (MHz)*
68	794–800
69	800–806

Notes:

1. In Alaska, television broadcast stations operating on Channel 5 (76–82 MHz) and on Channel 6 (82–88 MHz) shall not cause harmful interference to and must accept interference from non-Government fixed operations authorized prior to 1 January 1982.
2. Channel 37, 608–614 MHz is reserved exclusively for the radio astronomy service.
3. In Hawaii, the frequency band 488–494 MHz is allocated for non-broadcast use. This frequency band (Channel 17) will not be assigned in Hawaii for use by television broadcast stations.

24.12 License-free bands

In Europe there are a number of license-free bands that can be used by anyone. However, there are restrictions on the use of these bands, both in terms of the application and the transmitted power. In particular, the manufacturer of transmitters operating in these bands must certify the equipment as meeting the required emission limits.

49.82 to 49.98 MHz – general purpose
173.2 to 173.35 MHz – telemetry. Note: 173.225 MHz is for short range alarms only
433.05 to 434.79 MHz – telemetry and vehicle security
458.5 to 458.95 MHz – industrial or commercial telemetry
868 to 870 MHz – general purpose.

Worldwide agreement has resulted in two bands, at 2.4 GHz and 5 GHz, being allocated to wireless LANs. Both of these bands are for industrial, scientific and medical (ISM) use.

The 2.4 GHz allocation covers 2.4 to 2.4835 GHz. Restrictions on the use of this ISM band include the use of carrier frequency hopping in step multiples of 1 MHz. This band is used by wireless LANs operating under IEEE802.11B and Bluetooth.

The IEEE802.11B standard uses 5 MHz carrier frequency steps, with some international variations. Europe allows the use of 13 channels (channel 1 to 13, with centre frequencies 2412 MHz to 2472 MHz), but the USA only allows the use of 11 channels (channel 1 to 11, with centre frequencies 2412 MHz to 2462 MHz). In Japan only channel 14 is allowed (no hopping), using a carrier centre frequency of 2477 MHz.

Bluetooth uses 79 hop frequencies, with multiples of a 1 MHz step size. The lowest carrier frequency is at 2402 MHz (channel 0) and the highest is 2480 MHz (channel 78). The frequency hopping pattern is psuedo-random, with 1600 hops per second.

The 5 GHz allocation is actually two sub-bands, one covering 5.15 to 5.35 GHz and the other covering 5.47 GHz to 5.725 GHz. The higher frequency ISM band allows 1 watt radiated power, rather than the 200 mW limit of the lower frequency band. As with the 2.4 GHz ISM band, sources in the 5 GHz band must use carrier frequency hopping with 5 MHz steps.

24.13 Calculating radio antenna great circle bearings

Aiming radio antennas to target a particular area of the world requires calculation of the *great circle bearing* between your location and the other stations' location. That bearing is calculated from some simple spherical trigonometry using a hand-held calculator or a computer program. Before talking about the maths, however, we need to establish a frame of reference that makes the system work.

24.13.1 Latitude and longitude

The need for navigation on the surface of the Earth caused the creation of a grid system uniquely to locate points on the surface of our globe. *Longitude* lines run from the north pole to the south pole, i.e. from north to south.

The reference point (longitude zero), called the *prime meridian*, runs through Greenwich, England. The longitude of the prime meridian is 0 degrees. Longitudes west of the prime meridian are given a plus sign (+), while longitudes east of the prime are given a minus (−) sign. If you continue the prime meridian through the poles to the other side of the Earth it has a longitude of 180 degrees. Thus, the longitude values run from −180 degrees to +180 degrees, with ±180 degrees being the same line.

The observatory at Greenwich is also the point against which relative time is measured. Every 15 degree change of longitude is equivalent to a one hour difference with the Greenwich time. To the west, subtract one hour for each 15 degrees and to the east add one

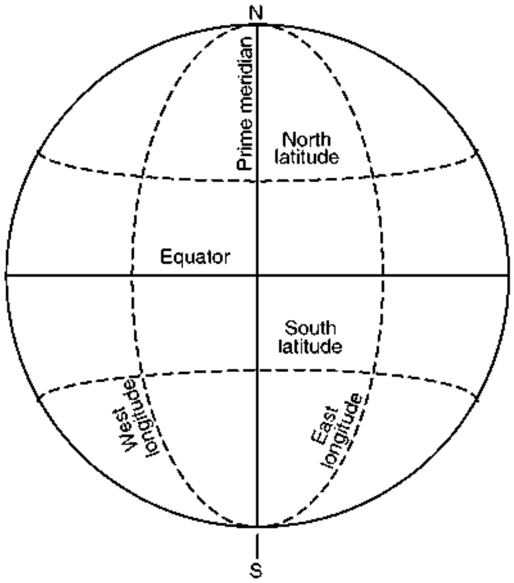

Figure 24.1 Lines of longitude and latitude

hour for each 15 degrees. Thus, the time on the East Coast of the United States is −5 hours relative to Greenwich time. At one time, we called time along the prime meridian *Greenwich mean time* (GMT), also called *Zulu time* to simplify matters for CW operators.

Latitude lines are measured against the equator, with distances north of the equator being taken as positive, and distances south of the equator being negative. The equator is 0 degrees latitude, while the north pole is +90 degrees latitude and the south pole is −90 degrees latitude.

Long ago navigators learned that the latitude can be measured by 'shooting' the stars and consulting a special atlas to compare the angle of certain stars with tables that translate to latitude numbers. The longitude measurement, however, is a bit different. For centuries sailors could measure latitude, but had to guess longitude (often with tragic results). In the early eighteenth century, the British government offered a large cash prize to anyone who could design a chronometer that could be taken to sea. By keeping the chronometer set accurately to Greenwich mean time, and comparing GMT against local time (i.e. at a time like high noon when the position of the sun is easy to judge), the longitude could be calculated. If you are interested in this subject, then most decent libraries have books on celestial navigation.

24.13.2 The great circle

On the surface of a globe, a curved line called a *great circle path* is the shortest distance between two points.

Consider two points on a globe: ‘A’ is your location, while ‘B’ is the other station’s location. The distance ‘D’ is the great circle path between ‘A’ and ‘B’.

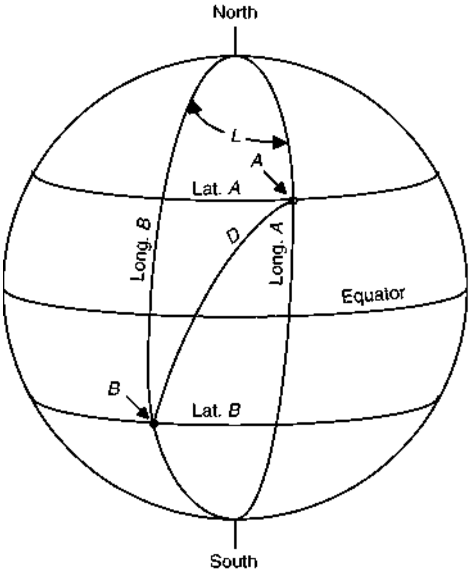

Figure 24.2 Great circle path

The great circle path length can be expressed in either degrees or distance (e.g. miles, nautical miles or kilometres). To calculate the distance, it is necessary to find the difference in longitude (L) between your longitude (LA) and the other station’s longitude (LB): $L = LA - LB$. Keep the signs straight. For example, if your longitude (LA) is 40 degrees, and the other station’s longitude (LB) is -120 degrees, then $L = 40 - (-120) = 40 + 120 = 160$. The equation for distance (D) is:

$$\cos D = (\sin A \times \sin B) + (\cos A \times \cos B \times \cos L)$$

where:

D is the angular great circle distance
A is your latitude
B is the other station’s latitude.

To find the actual angle, take the arccos of the above equation, i.e.

$$D = \arccos(\cos D)$$

In the next equation you will want to use D in angular measure, but later on will want to convert D to miles. To do this, multiply D in degrees by 69.4. Or, if you prefer metric measures, then $D \times 111.2$ yields kilometres. This is the approximate distance in statute miles between 'A' and 'B'.

To find the bearing from true north, work the equation below:

$$C = \arccos\left[\frac{\sin B - (\sin A \times \cos D)}{(\cos A \times \sin D)}\right]$$

However, this equation won't always give you the right answer unless you make some corrections.

The first problem is the 'same longitude error', i.e. when both stations are on the same longitude line. In this case, $L = LA - LB = 0$. If LAT $A >$ LAT B, then $C = 180$ degrees, but if LAT $A <$ LAT B, then $C = 0$ degrees. If LAT $A =$ LAT B, then what's the point of all these calculations?

The next problem is found when the condition $-180^\circ \leq L \leq +180^\circ$ is not met, i.e. when the absolute value of L is greater than 180°, $\text{ABS}(L) > 180^\circ$. In this case, either add or subtract 360 in order to make the value between ±180°:

$$\text{If } L > +180, \text{ then } L = L - 360$$

$$\text{If } L < -180, \text{ then } L = L + 360$$

One problem seen while calculating these values on a computer or hand calculator is the fact that the $\sin(X)$ and $\cos(X)$ cover different ranges. The $\sin(X)$ function returns values from 0° to 360°, while the $\cos(X)$ function returns values only over 0° to 180°. If L is positive, then the result bearing C is accurate, but if L is negative then the actual value of $C = 360 - C$. The following test is necessary:

```
If $L < 0$ then
$L = 360 - L$
Else $L = L$
End if
```

Another problem is seen whenever either station is in a high latitude near either pole (±90°), or where both locations are very close together, or where the two locations are antipodal (i.e. on opposite points on the Earth's surface). The best way to handle these problems is to use a different equation that multiplies by the cosecant of D (i.e. cosec(D)), rather than dividing by sine of D (i.e. sin(D)).

25 Abbreviations and symbols

25.1 Abbreviations

Many abbreviations are found as either capital or lower case letters, depending on publishers' styles. Symbols should generally be standard, as shown.

A	Ampere or anode
ABR	Auxiliary bass radiator
a.c.	Alternating current
Ack	Acknowledgement
A/D	Analogue to digital
ADC	Analogue to digital converter
Ae	Aerial
a.f.	Audio frequency
a.f.c.	Automatic frequency control
a.g.c.	Automatic gain control
a.m.	Amplitude modulation
AMPS	Advanced mobile phone system
ANSI	American National Standards Institute
ASA	Acoustical Society of America
ASCII	American Standard Code for Information Interchange
a.t.u.	Aerial tuning unit
AUX	Auxiliary
a.v.c.	Automatic volume control
AWG	American Wine Gauge
b	Base of transistor
BAF	Bonded acetate fibre
B & S	Brown & Sharpe (U.S.) wire gauge
b.p.s.	Bits per second
BR	Bass reflex
BSI	British Standards Institution
BW	Bandwidth
C	Capacitor, cathode, centigrade, coulomb
c	Collector of transistor, speed of light
CB	Citizen's band
CCD	Charge coupled device
CCIR	International Radio Consultative Committee

CCITT	International Telegraph and Telephone Consultative Committee
CCTV	Closed circuit television
CDMA	Code division multiple access
chps	Characters per second
CISPR	Comité International Special Des Peturbations (radio interference standards body)
CLK	Clock signal
CMOS	Complementary metal oxide semiconductor
CrO_2	Chromium dioxide
CPU	Central processor unit
CTCSS	Continuous tone controlled signalled system
CTD	Charge transfer device
c.w.	Continuous wave
D	Diode
d	Drain of an f.e.t.
D/A	Digital to analogue
DAC	Digital to analogue converter
dB	Decibel
d.c.	Direct current
DCC	Double cotton covered
DCE	Data circuit-terminating equipment
DCS	Digital Communications Service
DF	Direction finding
DIL	Dual-in-line
DIN	German standards institute
DMA	Direct memory access
DPDT	Double pole, double throw
DPSK	Differential phase shift keying
DPST	Double pole, single throw
dsb	or dsbam. Double sideband amplitude modulation
DSRR	Digital short range radio
DTE	Data terminal equipment
DTL	Diode-transistor logic
DTMF	Dual tone multi-frequency
DX	Long distance reception
e	Emitter of transistor
EAROM	Electrically alterable read only memory
ECL	Emitter coupled logic
EDGE	Enhanced Data for GSM Evolution
e.h.t.	Extremely high tension (voltage)

EIRP	Effective Isotropic Radiated Power
e.m.f.	Electromotive force
en	Enamelled
EPROM	Erasable programmable read only memory
EQ	Equalization
EROM	Erasable read only memory
ERP	Effective radiated power
ETACS	Extended total access communications system
F	Farad, fahrenheit or force
f	Frequency
FDD	Frequency Division Duplex
FDM	Frequency division multiplex
FDMA	Frequency division multiple access
Fe	Ferrous
FeCr	Ferri-chrome
f.e.t.	Field effect transistor
FFSK	Fast frequency shift keying
f.m.	Frequency modulation
f.r.	Frequency response or range
f.s.d.	Full-scale deflection
FSK	Frequency shift keying
G	Giga (10^9)
g	Grid, gravitational constant
GMSK	Gaussian minimum shift keying
GSM	Global system mobile
H	Henry
h.f.	High frequency
Hz	Hertz (cycles per second)
I	Current
IB	Infinite baffle
i.c.	Integrated circuit
IF	Intermediate frequency
IHF	Institute of High Fidelity (U.S.)
I^2L(HL)	Integrated injection logic
i.m.d.	Intermodulation distortion
i/p	Input
i.p.s.	Inches per second
K	Kilo, in computing terms ($= 2^{10} = 1024$), or degrees Kelvin
k	Kilo (10^3) or cathode
L	Inductance or lumens
l.e.d.	Light emitting diode

l.f.	Low frequency
LIN	Linear
LOG	Logarithmic
LS	Loudspeaker
LSI	Large scale integration
l.w.	Long wave (approx. 1100–2000 m)
M	Mega (10^{6})
m	Milli (10^{-3}) or metres
m.c.	Moving coil
MHz	Megahertz
mic	Microphone
MOS	Metal oxide semiconductor
MPU	Microprocessor unit
MPX	Multiplex
MSC	Mobile Switching Centre
MSK	Minimum shift keying
m.w.	Medium wave (approx. 185–560 m)
n	Nano (10^{9})
NAB	National Association of Broadcasters
Ni-Cad	Nickel-cadmium
n/c	Not connected; normally closed
n/o	Normally open
NMOS	Negative channel metal oxide semiconductor
NRZ	Non-return to zero
o/c	Open channel; open circuit
o/p	Output
op-amp	Operational amplifier
p	Pico (10^{-12})
PA	Public address
PABX	Private automatic branch exchange
PAL	Phase alternation, line
p.a.m.	Pulse amplitude modulation
PCB	Printed circuit board
PCM	Pulse code modulation
PCN	Personal communications network
PCS	Personal communication system
PLA	Programmable logic array
PLL	Phase locked loop
pm	Phase modulation
PMOS	Positive channel metal oxide semiconductor
P.P.M.	Peak programme meter
p.r.f.	Pulse repetition frequency

PROM	Programmable read only memory
PSK	Phase shift keying
PSS	Packet SwitchStream
PSTN	Public Switched Telephone Network
PSU	Power supply unit
PTFE	Polytetrafluoroethylene
PU	Pickup
PUJT	Programmable unijunction transistor
Q	Quality factor; efficiency of tuned circuit, charge
QAM	Quadrature (or quaternary) amplitude modulation
QPSK	Quadrature (or quaternary) phase shift keying
R	Resistance
RAM	Random access memory
RCF	Recommended crossover frequency
r.f.	Radio frequency
r.f.c.	Radio frequency choke (coil)
RFI	Radio frequency interference
RIAA	Record Industry Association of America
r.m.s.	Root mean square
ROM	Read only memory
RTL	Resistor transistor logic
R/W	Read/write
RX	Receiver
S	Siemens
s	Source of an f.e.t.
s/c	Short circuit
SCR	Silicon-controlled rectifier
s.h.f.	Super high frequency
SI	International system of units
S/N	Signal-to-noise.
SPL	Sound pressure level
SPDT	Single pole, double throw
SPST	Single pole, single throw
ssb	Single sideband amplitude modulation
ssbdc	Single sideband diminished carrier
ssbsc	Single sideband suppressed carrier
SSI	Small scale integration
s.w.	Short wave (approx. 10–60 m)
s.w.g.	Standard wire gauge
s.w.r.	Standing wave ratio
T	Tesla

TACS	Total access communications system
TDD	Time division duplex
TDM	Time division multiplex
TDMA	Time division multiple access
t.h.d.	Total harmonic distortion
t.i.d.	Transient intermodulation distortion
TR	Transformer
t.r.f.	Tuned radio frequency
TRS	Transmitter repeater station
TTL	Transistor transistor logic
TTY	Teletype unit
TVI	Television interface; television interference
TX	Transmitter
UART	Universal asynchronous receiver transmitter
u.h.f.	Ultra high frequency (approx. 470–854 MHz)
u.j.t.	Unijunction transistor
ULA	Uncommitted logic array
V	Volts
VA	Volt-amps
v.c.a.	Voltage controlled amplifier
v.c.o.	Voltage controlled oscillator
VCT	Voltage to current transactor
v.h.f.	Very high frequency (approx. 88–216 MHz)
v.l.f.	Very low frequency
VSB	Vestigial side band
VSWR	Voltage standing wave ratio
VU	Volume unit
W	Watts
WAP	Wireless application protocol
Wb	Weber
W/F	Wow and flutter
WML	Wireless mark-up language
w.p.m.	Words per minute
X	Reactance
Xtal	Crystal
Z	Impedance
ZD	Zener diode

25.2 Letter symbols by unit name

Unit	*Symbol*	*Notes*
ampere	A	SI unit of electric current.
ampere (turn)	At	SI unit of magnetomotive force.
ampere-hour	Ah	
ampere per metre	$\mathrm{A\,m^{-1}}$	SI unit of magnetic field strength.
angstrom	Å	1 Å = 10^{-10} m.
apostilb	asb	1 asb$(1/\pi)$cd $\mathrm{m^{-2}}$ A unit of luminance. The SI unit, candela per square metre, is preferred.
atmosphere:		
standard atmosphere	atm	1 atm = 101 325 N $\mathrm{m^{-2}}$.
technical atmosphere	at	1 at = 1 kgf $\mathrm{cm^{-2}}$.
atomic mass unit (unified)	u	The (unified) atomic mass unit is defined as one-twelfth of the mass of an atom of the ^{12}C nuclide. Use of the old atomic mass unit (amu), defined by reference to oxygen, is deprecated.
bam	b	1 b10^{-28} $\mathrm{m^2}$.
bar	bar	1 bar = 100 000 N $\mathrm{m^{-2}}$.
baud	Bd	Unit of signalling speed equal to one element per second.
becquerel	Bq	1 Bq = 1 $\mathrm{s^{-1}}$. SI unit of radioactivity.
bel	B	
bit	b	
British thermal unit	Btu	

Unit	*Symbol*	*Notes*
calorie (International Table calorie)	$\mathrm{cal_{IT}}$	1 $\mathrm{cal}^{\circ} - = 4.1868$ J. The 9th Conférence Générale des Poids et Mesures adopted the joule as the unit of heat, avoiding the use of the calorie as far as possible.
calorie (thermochemical calorie)	cal	1 cal = 4.1840 J. (See note for International Table calorie.)
candela	cd	SI unit of luminous intensity.
candela per square inch	$\mathrm{cd\ in^{-2}}$	Use of the SI unit, candela per square metre, is preferred.
candela per square metre	$\mathrm{cd\ m^{-2}}$	SI unit of luminance. The name nit has been used.
candle		The unit of luminous intensity has been given the name *candela*; use of the word *candle* for this purpose is deprecated.
centimetre	cm	
circular mil	cmil	1 cmil = $(\pi/4)10^{-6}\ \mathrm{in}^2$.
coulomb	C	SI unit of electrical charge.
cubic centimetre	$\mathrm{cm^3}$	
cubic foot	$\mathrm{ft^3}$	
cubic foot per minute	$\mathrm{ft^3\ min^{-1}}$	
cubic foot per second	$\mathrm{ft^3\ s^{-1}}$	
cubic inch	$\mathrm{in^3}$	
cubic metre	$\mathrm{m^3}$	
cubic metre per second	$\mathrm{m^3\ s^{-1}}$	
cubic yard	$\mathrm{yd^3}$	
curie	Ci	Unit of activity in the field of radiation dosimetry.
cycle	c	
cycle per second	$\mathrm{c\,s^{-1}}$	Deprecated. Use hertz.
decibel	dB	
degree (plane angle)	°	

Unit	*Symbol*	*Notes*
degree (temperature):		Note that there is no space between the symbol ° and the letter. The use of the word *centigrade* of the Celsius temperature scale was abandoned by the Conférence Générale des Poids et Mesures in 1948.
degree Celsius	°C	
degree Fahrenheit	°F	
degree Kelvin		See Kelvin.
degree Rankine	°R	
dyne	dyn	
electronvolt	eV	
erg	erg	
erlang	E	Unit of telephone traffic.
farad	F	SI unit of capacitance.
foot	ft	
footcandle	fc	Use of the SI unit of illuminance, the lux (lumen per square metre), is preferred.
footlambert	fL	Use of the SI unit, the candela per square metre, is preferred.
foot per minute	$\mathrm{ft\ min^{-1}}$	
foot per second	$\mathrm{ft\ s^{-1}}$	
foot per second squared	$\mathrm{ft\ s^{-2}}$	
foot pound-force	$\mathrm{ft\ lb_f}$	
gal	Gal	$1\,\mathrm{Gal} = 1\,\mathrm{cm\ s^{-2}}$.
gallon	gal	The gallon, quart, and pint differ in the US and the UK, and their use is deprecated.
gauss	G	The gauss is the electromagnetic CGS (Centimetre Gram Second) unit of magnetic flux density. The SI unit, tesla, is preferred.
gigaelectronvolt	GeV	

Unit	*Symbol*	*Notes*
gigahertz	GHz	
gilbert	Gb	The gilbert is the electromagnetic CGS (Centimetre Gram Second) unit of magnetomotive force. Use of the SI unit, the ampere (or ampere-turn), is preferred.
grain	gr	
gram	g	
gray	Gy	$1\ \mathrm{Gy} = 1\,\mathrm{J\ kg^{-1}}$. SI unit of absorbed dose.
henry	H	
hertz	Hz	SI unit of frequency.
horsepower	hp	Use of the SI unit, the watt, is preferred.
hour	h	Time may be designated as in the following example; $9^{\mathrm{h}}46^{\mathrm{m}}30^{\mathrm{s}}$.
inch	in	
inch per second	$\mathrm{in\ s^{-1}}$	
joule	J	SI unit of energy.
joule per Kelvin	$\mathrm{J\,K^{-1}}$	SI unit of heat capacity and entropy.
Kelvin	K	SI unit of temperature (formerly called *degree Kelvin*). The symbol K is now used without the symbol°.
kiloelectronvolt	KeV	
kilogauss	kG	
kilogram	kg	SI unit of mass.
kilogram-force	$\mathrm{kg_f}$	In some countries the name *kilopond* (kp) has been adopted for this unit.
kilohertz	kHz	
kilohm	kΩ	
kilojoule	kJ	
kilometre	km	

Unit	*Symbol*	*Notes*
kilometre per hour	$\mathrm{km\ h^{-1}}$	
kilopond	kp	See kilogram-force.
kilovar	kvar	
kilovolt	kV	
kilovoltampere	kVA	
kilowatt	kW	
kilowatthour	kWh	
knot	kn	$1\,\mathrm{kn} = 1\,\mathrm{nmi\ h^{-1}}$.
lambert	L	The lambert is the CGS (Centimetre Gram Second) unit of luminance. The SI unit, candela per square metre, is preferred.
litre	l	
litre per second	$\mathrm{l\,s^{-1}}$	
lumen	lm	SI unit of luminous flux.
lumen per square foot	$\mathrm{lm\ ft^{-2}}$	Use of the SI unit, the lumen per square metre, is preferred.
lumen per square metre	$\mathrm{lm\ m^{-2}}$	SI unit of luminous excitance.
lumen per watt	$\mathrm{lm\ W^{-1}}$	SI unit of luminous efficacy.
lumen second	lm s	SI unit of quantity of light.
lux	lx	$1\ \mathrm{lx} = 1\,\mathrm{lm\ m^{-2}}$. SI unit of illuminance.
maxwell	Mx	The maxwell is the electromagnetic CGS (Centimetre Gram Second) unit of magnetic flux. Use of the SI unit, the weber, is preferred.
megaelectronvolt	MeV	
megahertz	MHz	
megavolt	MV	
megawatt	MW	
megohm	MΩ	
metre	m	SI unit of length.
mho	mho	$1\,\mathrm{mho} = 1\ \Omega^{-1} = 1\,\mathrm{S}$.
microampere	μA	
microbar	μbar	

Unit	*Symbol*	*Notes*
microfarad	μF	
microgram	μg	
microhenry	μH	
micrometre	μm	
micron		The name *micrometre* (μm) is preferred.
microsecond	μS	
microwatt	μW	
mil	mil	1 mil = −0.001 in.
mile		
nautical	nmi	
statute	mi	
mile per hour	$\mathrm{mi\ h^{-1}}$	
milliampere	mA	
millibar	mbar	mb may be used.
milligal	mGal	
milligram	mg	
millihenry	mH	
millilitre	ml	
millimetre	mm	
conventional millimetre of mercury	mm Hg	1 mm Hg = 133.322 $\mathrm{N\ m^{-2}}$.
millimicron		The name *nanometre* (nm) is preferred.
millisecond	ms	
millivolt	mV	
milliwatt	mW	
minute (plane angle)	$\ldots'$	
minute (time)	min	Time may be designated as in the following example: $9^{h}46^{m}30^{s}$.
mole	mol	SI unit of amount of substance.
nanoampere	nA	
nanofarad	nF	
nanometre	nm	
nanosecond	ns	
nanowatt	nW	
nautical mile	nmi	

Unit	*Symbol*	*Notes*
neper	Np	
newton	N	SI unit of force.
newton metre	Nm	
newton per square metre	$\mathrm{N\,m^{-2}}$	See pascal.
nit	nt	$1\,\mathrm{nt} = 1\,\mathrm{cd\ m^{-2}}$. See candela per square metre.
oersted	Oe	The oersted is the electromagnetic CGS (Centimetre Gram Second) unit of magnetic field strength. Use of the SI unit, the ampere per metre, is preferred.
ohm	Ω	SI unit of electrical resistance.
ounce (avoirdupois)	oz	
pascal	Pa	SI unit of pressure or stress. $1\,\mathrm{Pa} = 1\,\mathrm{N\ m^{-2}}$.
picoampere	pA	
picofarad	pF	
picosecond	ps	
picowatt	pW	
pint	pt	The gallon, quart, and pint differ in the US and the UK, and their use is deprecated.
pound	lb	
poundal	pdl	
pound-force	$\mathrm{lb_f}$	
pound-force feet	$\mathrm{lb_f\ ft}$	
pound-force per square inch	$\mathrm{lb_f\ in^{-2}}$	
pound per square inch		Although use of the abbreviation psi is common, it is not recommended. See pound-force per square inch.

Unit	*Symbol*	*Notes*
Quart	qt	The gallon, quart, and pint differ in the US and the UK, and their use is deprecated.
rad	rd	Unit of absorbed dose in the field of radiation dosimetry.
revolution per minute	$\mathrm{r\ min^{-1}}$	Although use of the abbreviation rpm is common, it is not recommended.
revolution per second	$\mathrm{r\ s^{-1}}$	
roentgen	R	Unit of exposure in the field of radiation dosimetry.
second (plane angle)	$\ldots''$	
second (time)	s	SI unit of time. Time may be designated as in the following example: $9^{\mathrm{h}}46^{\mathrm{m}}30^{\mathrm{s}}$.
siemens	S	SI unit of conductance. $1\,\mathrm{S} = 1\ \Omega^{-1}$.
square foot	ft^2	
square inch	in^2	
square metre	m^2	
square yard	yd^2	
steradian	sr	SI unit of solid angle.
stilb	sb	$1\,\mathrm{sb} = 1\,\mathrm{cd\ cm^{-2}}$. A CGS unit of luminance. Use of the SI unit, the candela per square metre, is preferred.
tesla	T	SI unit of magnetic flux density. $1\,\mathrm{T} = 1\,\mathrm{Wb\ m^{-2}}$.
tonne	t	$1\,\mathrm{t} = 1000\,\mathrm{kg}$.
(unified) atomic mass unit	u	See atomic mass unit (unified).
var	var	Unit of reactive power.
volt	V	SI unit of electromotive force.
voltampere	VA	SI unit of apparent power.
watt	W	SI unit of power.
watthour	Wh	

Unit	*Symbol*	*Notes*
watt per steradian	$\mathrm{W\ sr^{-1}}$	SI unit of radiant intensity.
watt per steradian square metre	$\mathrm{W\ (sr\ m^2)^{-1}}$	SI unit of radiance.
weber	Wb	SI unit of magnetic flux. $1\,\mathrm{Wb} = 1\,\mathrm{V\,s}$.
yard	yd	

25.3 Electric quantities

Quantity	*Symbol*	*Unit*	*Symbol*
Admittance	Y	siemens	S
Angular frequency	ω	hertz	Hz
Apparent power	S	watt	W
Capacitance	C	farad	F
Charge	Q	coulomb	C
Charge density	ρ	coulomb per square metre	$\mathrm{C\,m^{-2}}$
Conductance	G	siemens	S
Conductivity	κ, γ, σ	siemens per metre	$\mathrm{S\,m^{-1}}$
Current	I	ampere	A
Current density	j, J	ampere per square metre	$\mathrm{A\,m^{-2}}$
Displacement	D	coulomb per square metre	$\mathrm{C\,m^{-2}}$
Electromotive force	E	volt	V
Energy	E	joule	J
Faraday constant	F	coloumb per mole	$\mathrm{C\,mol^{-1}}$
Field strength	E	volt per metre	$\mathrm{V\,m^{-1}}$
Flux	ψ	coulomb	C
Frequency	v, f	hertz	Hz
Impedance	Z	ohm	Ω
Light, velocity of in a vacuum	c	metre per second	$\mathrm{m\,s^{-1}}$
Period	T	second	s
Permeability	μ	henry per metre	$\mathrm{H\,m^{-1}}$
Permeability of space	μ_o	henry per metre	$\mathrm{H\,m^{-1}}$
Permeance	Λ	henry	H

Quantity	*Symbol*	*Unit*	*Symbol*
Permittivity	ε	farad per metre	$\mathrm{F\,m^{-1}}$
Permittivity of space	ε_o	farad per metre	$\mathrm{F\,m^{-1}}$
Phase	ϕ	–	–
Potential	V, U	volt	V
Power	P	watt	W
Quality factor	Q	–	–
Reactance	X	ohm	Ω
Reactive power	Q	watt	W
Relative permeability	μ_r	–	–
Relative permittivity	ε_r	–	–
Relaxation time	τ	second	s
Reluctance	R	reciprocal henry	$\mathrm{H^{-1}}$
Resistance	R	ohm	Ω
Resistivity	ρ	ohm metre	Ω m
Susceptance	B	siemens	S
Thermodynamic temperature	T	kelvin	K
Time constant	τ	second	s
Wavelength	λ	metre	m

26 Miscellaneous data

26.1 Fundamental constants

Constant	*Symbol*	*Value*
Boltzmann constant	k	$1.38062 \times 10^{-23}\ \mathrm{J\,K^{-1}}$
Electron charge, proton charge	e	$\pm 1.60219 \times 10^{-19}\ \mathrm{C}$
Electron charge-to-mass ratio	e/m	$1.7588 \times 10^{11}\ \mathrm{C\,kg^{-1}}$
Electron mass	m_e	$9.10956 \times 10^{-31}\ \mathrm{kg}$
Electron radius	r_e	$2.81794 \times 10^{-15}\ \mathrm{m}$
Faraday constant	F	$9.64867 \times 10^{4}\ \mathrm{C\,mol^{-1}}$
Neutron mass	m_n	$1.67492 \times 10^{-27}\ \mathrm{kg}$
Permeability of space	μ_o	$4\pi \times 10^{-7}\ \mathrm{H\,m^{-1}}$
Permittivity of space	ε_o	$8.85419 \times 10^{-12}\ \mathrm{F\,m^{-1}}$
Planck constant	h	$6.6262 \times 10^{-34}\ \mathrm{J\,s}$
Proton mass	m_p	$1.67251 \times 10^{-27}\ \mathrm{kg}$
Velocity of light	c	$2.99793 \times 10^{8}\ \mathrm{m\,s^{-1}}$

26.2 Electrical relationships

Amperes × ohms = **volts**
Volts ÷ amperes = **ohms**
Volts ÷ ohms = **amperes**
Amperes × volts = **watts**
$(\text{Amperes})^2 \times \text{ohms} =$ **watts**
$(\text{Volts})^2 \div \text{ohms} =$ **watts**
Joules per second = **watts**
Coulombs per second = **amperes**
Amperes × seconds = **coulombs**
Farads × volts = **coulombs**
Coulombs ÷ volts = **farads**
Coulombs ÷ farads = **volts**
Volts × coulombs = **joules**
$\text{Farads} \times (\text{volts})^2 =$ **joules**

26.3 Dimensions of physical properties

Length: metre [L]. Mass: kilogram [M]. Time: second [T]. Quantity of electricity: coulomb [Q]. Area: square metre [L^2]. Volume: cubic metre [L^3].

Velocity: metre per second	[LT^{-1}]
Acceleration: metre per second^2	[LT^{-2}]
Force: newton	[MLT^{-2}]
Work: joule	[ML^2T^{-2}]
Power: watt	[ML^2T^{-3}]
Electric current: ampere	[QT^{-1}]
Voltage: volt	[$ML^2T^{-2}Q^{-1}$]
Electric resistance: ohm	[$ML^2T^{-1}Q^{-2}$]
Electric conductance: siemens	[$M^{-1}L^{-2}TQ^2$]
Inductance: henry	[ML^2Q^{-2}]
Capacitance: farad	[$M^{-1}L^{-2}T^2Q^2$]
Current density: ampere per metre^2	[$L^{-2}T^{-1}Q$]
Electric field strength: volt per metre	[$MLT^{-2}Q^{-1}$]
Magnetic flux: weber	[$MLT^2T^{-1}Q^{-1}$]
Magnetic flux density: tesla	[$MT^{-1}Q^{-1}$]
Energy: joule	[ML^2T^{-2}]
Frequency: hertz	[T^{-1}]
Pressure: pascal	[$ML^{-1}T^{-2}$]

26.4 Fundamental units

Quantity	*Unit*	*Symbol*
Amount of a substance	mole	mol
Charge	coulomb	C
Length	metre	m
Luminous intensity	candela	cd
Mass	kilogram	kg
Plane angle	radian	rad
Solid angle	steradian	sr
Thermodynamic temperature	kelvin	K
Time	second	s

26.5 Greek alphabet

Capital letters	*Small letters*	*Greek name*	*English equivalent*	*Capital letters*	*Small letters*	*Greek name*	*English equivalent*
Α	α	*Alpha*	a	Ν	ν	*Nu*	n
Β	β	*Beta*	b	Ξ	ξ	*Xi*	x
Γ	γ	*Gamma*	g	Ο	o	*Omicron*	*ǒ*
Δ	δ	*Delta*	d	Π	π	*Pi*	p
Ε	ε	*Epsilon*	e	Ρ	ρ	*Rho*	r
Ζ	ζ	*Zeta*	z	Σ	σ	*Sigma*	s
Η	η	*Eta*	é	Τ	τ	*Tau*	t
Θ	θ	*Theta*	th	Υ	υ	*Upsilon*	u
Ι	ι	*Iota*	i	Φ	ϕ	*Phi*	ph
Κ	κ	*Kappa*	k	Χ	χ	*Chi*	ch
Λ	λ	*Lambda*	l	Ψ	ψ	*Psi*	ps
Μ	μ	*Mu*	m	Ω	ω	*Omega*	ö

26.6 Standard units

Ampere Unit of electric current, the constant current which, if maintained in two straight parallel conductors of infinite length of negligible circular cross-section and placed one metre apart in a vacuum, will produce between them a force equal to 2×10^{-7} newton per metre length.

Ampere-hour Unit of quantity of electricity equal to 3 600 coulombs. One unit is represented by one ampere flowing for one hour.

Candela Unit of luminous intensity. It is the luminous intensity, in the perpendicular direction, of a surface of $1/600\,000\,\text{m}^{-2}$ of a full radiator at the temperature of freezing platinum under a pressure of 101 325 newtons m^{-2}.

Coulomb Unit of electric charge, the quantity of electricity transported in one second by one ampere.

Decibel (dB) Unit of acoustical or electrical power ratio. Although the bel is officially the unit, this is usually regarded as being too large, so the decibel is preferred. The difference between two power levels is P_1 and P_2, is given as

$$10 \log_{10} \frac{P_1}{P_2} \text{ decibels}$$

Farad Unit of electric capacitance. The capacitance of a capacitor between the plates of which there appears a difference of potential of one volt when it is charged by one coulomb of electricity. Practical units are the microfarad (10^{-6} farad), the nanofarad (10^{-9} farad) and the picofarad (10^{-12} farad).

Henry Unit of electrical inductance. The inductance of a closed circuit in which an electromotive force of one volt is produced when the electric current in the circuit varies uniformly at the rate of one ampere per second. Practical units are the microhenry (10^{-6} henry) and the millihenry (10^{-3} henry).

Hertz Unit of frequency. The number of repetitions of a regular occurrence in one second.

Joule Unit of energy, including work and quantity of heat. The work done when the point of application of a force of one newton is displaced through a distance of one metre in the direction of the force.

Kilovolt-ampere 1000 volt-amperes.

Kilowatt 1000 watts.

Light, velocity of Light waves travel at 300 000 kilometres per second (approximately). Also the velocity of radio waves.

Lumen m^{-2}, lux Unit of illuminance of a surface.

Mho Unit of conductance, see Siemens.

Newton Unit of force. That force which, applied to a mass of one kilogram, gives it an acceleration of one metre per second per second.

Ohm Unit of electric resistance. The resistance between two points of a conductor when a constant difference of potential of one volt, applied between these two points, produces in the conductor a current of one ampere.

Pascal Unit of sound pressure. Pressure is usually quoted as the root mean square pressure for a pure sinusoidal wave.

Siemens Unit of conductance, the reciprocal of the ohm. A body having a resistance of 4 ohms would have a conductance of 0.25 siemens.

Sound, velocity of Sound waves travel at 332 metres per second in air (approximately) at sea level.

Tesla Unit of magnetic flux density, equal to one weber per square metre of circuit area.

Volt Unit of electric potential. The difference of electric potential between two points of a conducting wire carrying a constant current of one ampere, when the power dissipated between these points is equal to one watt.

Volt-ampere The product of the root-mean-square volts and root-mean-square amperes.

Watt Unit of power, equal to one joule per second. Volts times amperes equals watts.

Weber Unit of magnetic flux. The magnetic flux which, linking a circuit of one turn, produces in it an electromotive force of one volt as it is reduced to zero at a uniform rate in one second.

26.7 Decimal multipliers

Prefix	*Symbol*	*Multiplier*	*Prefix*	*Symbol*	*Multiplier*
tera	T	10^{12}	centi	c	10^{-2}
giga	G	10^{9}	milli	m	10^{-3}
mega	M	10^{6}	micro	μ	10^{-6}
kilo	k	10^{3}	nano	n	10^{-9}
hecto	h	10^{2}	pico	p	10^{-12}
deka	da	10	femto	f	10^{-15}
deci	d	10^{-1}	atto	a	10^{-18}

26.8 Useful formulae

Boolean Algebra (laws of)

Absorption:

$$\mathrm{A} + (\mathrm{A.B}) = \mathrm{A}$$
$$\mathrm{A.(A} + \mathrm{B)} = \mathrm{A}$$

Annulment: $\mathrm{A} + 1 = 1$
$\mathrm{A}.0 = 0$

Association: $(\mathrm{A} + \mathrm{B}) + \mathrm{C} = \mathrm{A} + (\mathrm{B} + \mathrm{C})$
$(\mathrm{A}.\mathrm{B}).\mathrm{C} = \mathrm{A}.(\mathrm{B}.\mathrm{C})$

Commutation: $\mathrm{A} + \mathrm{B} = \mathrm{B} + \mathrm{A}$
$\mathrm{A}.\mathrm{B} = \mathrm{B}.\mathrm{A}$

Complements: $\mathrm{A} + \bar{\mathrm{A}} = 1$
$\mathrm{A}.\bar{\mathrm{A}} = 0$

De Morgan's: $\overline{(\mathrm{A} + \mathrm{B})} = \bar{\mathrm{A}}.\bar{\mathrm{B}}$
$\overline{(\mathrm{A}.\mathrm{B})} = \bar{\mathrm{A}} + \bar{\mathrm{B}}$

Distributive: $\mathrm{A}.(\mathrm{B} + \mathrm{C}) = (\mathrm{A}.\mathrm{B}) + (\mathrm{A}.\mathrm{C})$
$\mathrm{A} + (\mathrm{B}.\mathrm{C}) = (\mathrm{A} + \mathrm{B}).(\mathrm{A} + \mathrm{C})$

Double negation: $\bar{\bar{\mathrm{A}}} = \mathrm{A}$

Identity: $\mathrm{A} + \mathrm{O} = \mathrm{A}$
$\mathrm{A}.1 = \mathrm{A}$

Tautology: $\mathrm{A}.\mathrm{A} = \mathrm{A}$
$\mathrm{A} + \mathrm{A} = \mathrm{A}$

Capacitance

The capacitance of a parallel plate capacitor can be found from

$$C = \frac{0.885\,KA}{d}$$

C is in picofarads, K is the dielectric constant (air = 1). A is the area of the plate in square cm and d the thickness of the dielectric.

Calculation of overall capacitance with:
Parallel capacitors $-\mathrm{C} = \mathrm{C}_1 + \mathrm{C}_2 + \cdots$
Series capacitors $-\dfrac{1}{\mathrm{C}} = \dfrac{1}{\mathrm{C}_1} + \dfrac{1}{\mathrm{C}_2} + \cdots$

Characteristic impedance

$$\text{(open wire) } Z = 276 \log \frac{2D}{d} \text{ ohms}$$

where

$$\left.\begin{aligned} D &= \text{wire spacing} \\ d &= \text{wire diameter} \end{aligned}\right\} \text{in same units.}$$

$$\text{(coaxial) } Z = \frac{138}{\sqrt{(K)}} \log \frac{d_o}{d_i} \text{ ohms}$$

where K = dielectric constant, d_o = outside diameter of inner conductor, d_i = inside diameter of outer conductor.

Dynamic resistance

In a parallel-tuned circuit at resonance the dynamic resistance is

$$R_d = \frac{L}{Cr} = Q\omega L = \frac{Q}{\omega C} \text{ ohms}$$

where L = inductance (henries), C = capacitance (farads), r = effective series resistance (ohms). Q = Q-value of coil, and $\omega = 2\pi \times$ frequency (hertz).

Frequency – wavelength – velocity

(See also Resonance)
The velocity of propagation of a wave is

$$v = f\lambda \text{ metres per second}$$

where f = frequency (hertz) and λ = wavelength (metres).

For electromagnetic waves in free space the velocity of propagation v is approximately 3×10^8 m/sec, and if f is expressed in kilohertz and λ in metres

$$f = \frac{300\,000}{\lambda} \text{ kilohertz} \qquad f = \frac{300}{\lambda} \text{ megahertz}$$

or

$$\lambda = \frac{300\,000}{f} \text{ metres} \qquad \lambda = \frac{300}{f} \text{ metres}$$

f in kilohertz $\qquad$ f in megahertz

Impedance

The impedance of a circuit comprising inductance, capacitance and resistance in series is

$$Z = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$$

where R = resistance (ohms), $\omega = 2\pi \times$ frequency (hertz). L = inductance (henries), and C = capacitance (farads).

Inductance

Single layer coils

$$L(\text{in microhenries}) = \frac{a^2N^2}{9a + 10l} \text{ approximately}$$

If the desired inductance is known, the number of turns required may be determined by the formula

$$N = \frac{5L}{na^2}\left[1 + \sqrt{\left(1 + \frac{0.36n^2a^3}{L}\right)}\right]$$

where N = number of turns, a = radius of coil in inches, n = number of turns per inch. L = inductance in microhenries (μH) and l = length of coil in inches.

Calculation of overall inductance with:
Series inductors $-\mathrm{L} = \mathrm{L}_1 + \mathrm{L}_2 + \cdots$
Parallel inductors $-\frac{1}{\mathrm{L}} = \frac{1}{\mathrm{L}_1} + \frac{1}{\mathrm{L}_2} + \cdots$

Meter conversions

Increasing range of ammeters or milliammeters
Current range of meter can be increased by connecting a shunt resistance across meter terminals. If R_m is the resistance of the meter; R_s the value of the shunt resistance and n the number of times it is wished to multiply the scale reading, then

$$R_\mathrm{s} = \frac{R_\mathrm{m}}{(n-1)}$$

Increasing range of voltmeters
Voltage range of meter can be increased by connecting resistance in series with it. If this series resistance is R_s and R_m and n as before, then $R_\mathrm{s} = R_\mathrm{m} \times (n-1)$.

Negative feedback

Voltage feedback

$$\text{Gain with feedback} = \frac{A}{1 + Ab}$$

where A is the original gain of the amplifier section over which feedback is applied (including the output transformer if included) and b is the fraction of the output voltage fed back.

$$\text{Distortion with feedback} = \frac{d}{1 + Ab} \text{ approximately}$$

where d is the original distortion of the amplifier.

Ohm's Law

$$I = \frac{V}{R} V = IR \;\; R = \frac{V}{I}$$

where I = current (amperes), V = voltage (volts), and R = resistance (ohms).

Power

In a d.c. circuit the power developed is given by

$$W = VI = \frac{V^2}{R} = I^2 R \text{ watts}$$

where V = voltage (volts), I = current (amperes), and R = resistance (ohms).

Power ratio

$$P = 10 \log \frac{P_1}{P_2}$$

where P = ratio in decibels, P_1 and P_2 are the two power levels.

Q

The Q value of an inductance is given by

$$Q = \frac{\omega L}{R}$$

Radio horizon distance

The radio horizon at VHF/UHF and up is approximately 15% further than the optical horizon. Several equations are used in calculating the distance. If D is the distance to the radio horizon, and H is the antenna height, then:

$$D = k\sqrt{H}$$

1. When D is in statute miles (5280 feet) and H in feet, then $K = 1.42$.
2. When D is in nautical miles (6000 feet) and H in feet, then $K = 1.23$.
3. When D is in kilometres and H is in metres, then $K = 4.12$.

Reactance

The reactance of an inductor and a capacitor respectively is given by

$$X_{\mathrm{L}} = \omega L \text{ ohms} \qquad X_{\mathrm{C}} = \frac{1}{\omega C} \text{ ohms}$$

where $\omega = 2\pi \times$ frequency (hertz), $L =$ inductance (henries), and $C =$ capacitance (farads).

The total resistance of an inductance and a capacitance in series is $X_{\mathrm{L}} - X_{\mathrm{C}}$.

Resistance

Calculation of overall resistance with:
Series resistors $- \mathrm{R} = \mathrm{R}_1 + \mathrm{R}_2 + \cdots$
Parallel resistors $- \dfrac{1}{\mathrm{R}} = \dfrac{1}{\mathrm{R}_1} + \dfrac{1}{\mathrm{R}_2} + \cdots$

Resonance

The resonant frequency of a tuned circuit is given by

$$f = \frac{1}{2\pi\sqrt{LC}} \text{ hertz}$$

where $L =$ inductance (henries), and $C =$ capacitance (farads). If L is in microhenries (μH) and C is picofarads, this becomes

$$f = \frac{10^6}{2\pi\sqrt{LC}} \text{ kilohertz}$$

The basic formula can be rearranged

$$L = \frac{1}{4\pi^2 f^2 C} \text{ henries} \qquad C = \frac{1}{4\pi^2 f L} \text{ farads}$$

Since $2\pi f$ is commonly represented by ω, these expressions can be written

$$L = \frac{1}{\omega^2 C} \text{ henries} \qquad C = \frac{1}{\omega^2 L} \text{ farads}$$

Time constant

For a combination of inductance and resistance in series the time constant (i.e. the time required for the current to reach 63% of its final value) is given by

$$\tau = \frac{L}{R} \text{ seconds}$$

where L = inductance (henries), and R = resistance (ohms).

For a combination of capacitance and resistance in series the time constant (i.e. the time required for the voltage across the capacitance to reach 63% of its final value) is given by

$$\tau = CR \text{ seconds}$$

where C = capacitance (farads), and R = resistance (ohms).

Transformer ratios

The ratio of a transformer refers to the ratio of the number of turns in one winding to the number of turns in the other winding. To avoid confusion it is always desirable to state in which sense the ratio is being expressed: e.g. the 'primary-to-secondary' ratio $n_\mathrm{p}/n_\mathrm{s}$. The turns ratio is related to the impedance ratio thus

$$\frac{n_\mathrm{p}}{n_\mathrm{s}}\sqrt{\frac{Z_\mathrm{p}}{Z_\mathrm{s}}}$$

where n_p = number of primary turns, n_s = number of secondary turns, Z_p = impedance of primary (ohms), and Z_s = impedance of secondary (ohms).

Wattage rating

If resistance and current values are known,

$$W = I^2 R \text{ when } I \text{ is in amperes}$$

or

$$W = \frac{\text{milliamps}^2}{1\ 000\ 000} \times R$$

If wattage rating and value of resistance are known, the safe current for the resistor can be calculated from

$$\text{milliampers} = 1.000 \times \sqrt{\frac{\text{watts}}{\text{ohms}}}$$

Wavelength of tuned circuit

Formula for the wavelength in metres of a tuned oscillatory circuit is: $1885\sqrt{LC}$, where L = inductance in microhenries and C = capacitance in microfarads.

26.9 Colour codes

26.9.1 Resistor and capacitor colour coding

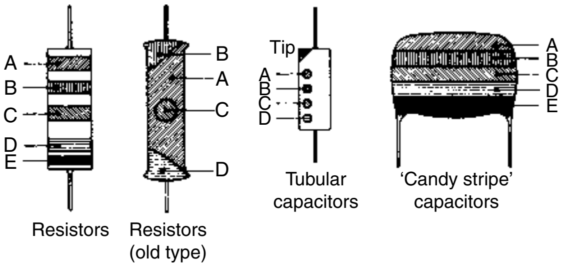
A
B
C
D
E
B
A
C
D
Tip
A
B
C
D
A
B
C
D
E
Resistors
Resistors (old type)
Tubular capacitors
'Candy stripe' capacitors

Tantalum capacitors

	1	**2**	**3**	**4**
Black	–	0	× 1	10 V
Brown	1	1	× 10	
Red	2	2	× 100	
Orange	3	3	–	
Yellow	4	4	–	6.3 V
Green	5	5	–	16 V
Blue	6	6	–	20 V
Violet	7	7	–	
Grey	8	8	× 0.01	25 V
White	9	9	× 0.1	3 V

(Pink 35 V)

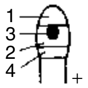

26.9.2 Resistor and capacitor letter and digit code table (BS 1852)

Resistor values are indicated as follows:

0.47 Ω	marked	R47
1 Ω		1R0
4.7 Ω		4R7
47 Ω		47R
100 Ω		100R
1 kΩ		1K0
10 kΩ		10K
10 MΩ		10M

A letter following the value shows the tolerance.

F = ±1%; G = ±2%; J = ±5%; K = ±10%; M = ±20%; R33M = 0.33Ω ± 20%; 6K8F = 6.8 kΩ ± 1%.

Capacitor values are indicated as:

0.68 pF	marked	p68
6.8 pf		6p8
1000 pF		1n0
6.8 nf		6n8
1000 nF		1μ0
6.8 μF		6μ8

Tolerance is indicated by letters as for resistors. Values up to 999 pF are marked in pF, from 1000 pf to 999 000 pF (= 999 nF) as nF (1000 pF = InF) and from 1000 nF (= 1 μF) upwards as μF.

Some capacitors are marked with a code denoting the value in pF (first two figures) followed by a multiplier as a power of ten (3 = 10^3). Letters denote tolerance as for resistors but C = ±0.25 pf. E.g. 123 J = 12 pF × 10^3 ± 5% = 12 000 pF (or 0.12 μF).

STAN GIBILISCO

THE ILLUSTRATED DICTIONARY OF Electronics

EIGHTH EDITION

Audio/Video
Consumer Electronics
Wireless Technology

The
Illustrated Dictionary of
Electronics

This page intentionally left blank.

The
Illustrated Dictionary of Electronics

Eighth Edition

Stan Gibilisco
Editor-in-Chief

McGraw-Hill
New York Chicago San Francisco Lisbon London Madrid
Mexico City Milan New Delhi San Juan Seoul
Singapore Sydney Toronto

Cataloging-in-Publication Data is on file with the Library of Congress

McGraw-Hill
A Division of The **McGraw-Hill** *Companies*

Copyright © 2001 by The McGraw-Hill Companies, Inc. All rights reserved. Printed in the United States of America. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a data base or retrieval system, without the prior written permission of the publisher.

1 2 3 4 5 6 7 8 9 0 AGM/AGM 0 9 8 7 6 5 4 3 2 1

P/N 0-07-137237-7

ISBN 0-07-137236-9

The sponsoring editor for this book was Scott Grillo, and the production supervisor was Pamela Pelton. It was set in Bookman by Techbooks.

Printed and bound by Quebecor/Martinsburg.

McGraw-Hill books are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information, please write to the Director of Special Sales, McGraw-Hill, Two Penn Plaza, New York, NY 10121-2298. Or contact your local bookstore

Information in this book has been obtained by the publisher from sources believed to be reliable. However, neither the publisher nor the authors guarantee the accuracy or completeness of any information published herein. Neither the publisher nor the authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that the publisher and authors are supplying information but are not attempting to render professional services in any way, shape or form. If such services are required, the assistance of an appropriate professional should be sought.

This book is printed on recycled, acid-free paper containing a minimum of 50 percent recycled de-inked fiber.

To Tony, Tim, and Samuel
from Uncle Stan

This page intentionally left blank.

Contents

Preface *ix*
Acknowledgments *xi*

Dictionary 1
Appendix A Schematic Symbols 773
Appendix B Tables and Data 787

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

This page intentionally left blank.

Preface

The Illustrated Dictionary of Electronics—8th Edition has been revised, clarified, and updated, reflecting technological advances of recent years. New definitions have been added in the fields of wireless technology, robotics, and artificial intelligence. Every effort has been made to be concise and accurate, without "talking down" to the reader.

Many definitions contain cross references (indicated in ALL CAPITALS); these provide recommended additional information or allow comparison with related terms. Expressions of special significance are printed in italics. Electronics abbreviations are included in the text; the full terms are stated as definitions.

While an effort has been made to avoid superfluous mathematics, equations are sometimes necessary to completely and effectively define a term. Mathematics beyond the high-school level has not been used.

Appendix A contains the standard symbols used in electrical and electronic diagrams. These symbols are used in illustrations throughout this dictionary. Appendix B contains the following data tables:

1. Conversion between electrical systems
2. Greek alphabet
3. Mathematical functions and operations
4. Prefix multipliers
5. Resistor color code

Suggestions for future editions are welcome.

Stan Gibilisco
Editor-in-Chief

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

This page intentionally left blank.

Acknowledgments

Illustrations in this book were generated with *CorelDRAW.* Some clip art is courtesy of Corel Corporation, 1600 Carling Avenue, Ottawa, Ontario, Canada K1Z 8R7.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

This page intentionally left blank.

The
Illustrated Dictionary of
Electronics

This page intentionally left blank.

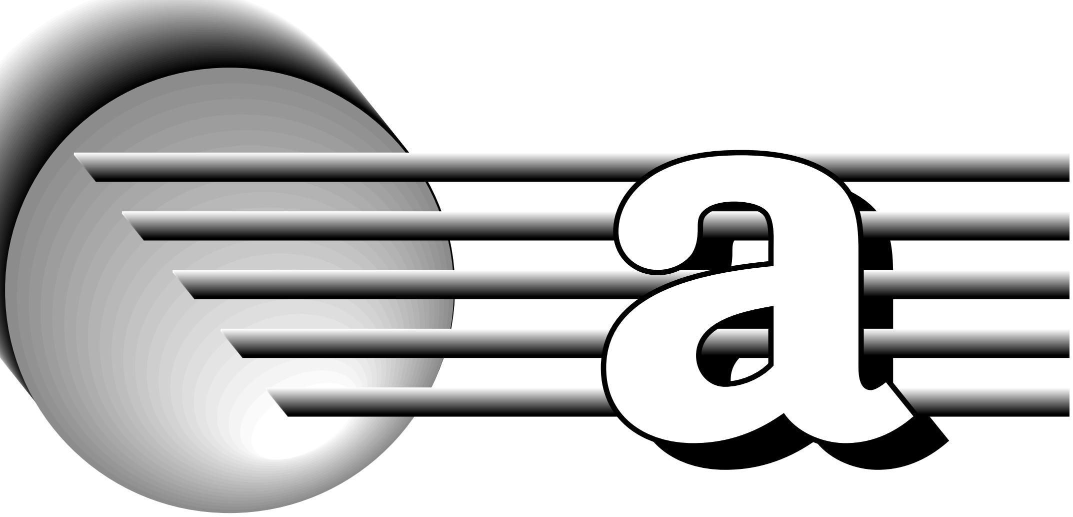

A **1.** Symbol for GAIN. **2.** Symbol for AREA. **3.** Symbol for AMPERE (SI unit for current).

A− Symbol for negative terminal of filament-voltage source in a vacuum-tube circuit.

A+ Symbol for positive terminal of filament-voltage source in a vacuum-tube circuit.

a **1.** Abbreviation of ATTO- (prefix). **2.** Abbreviation of AREA. **3.** Abbreviation of ACCELERATION. **4.** Abbreviation of ANODE. **5.** Obsolete abbreviation of cgs prefix AB-.

aA **1.** Abbreviation of *attoampere.* **2.** Obsolete for ABAMPERE.

AAAS Abbreviation for *American Association for the Advancement of Science.*

AAC Abbreviation of *automatic aperture control* (NASA).

AAS Abbreviation of *advanced antenna system* (NASA).

AASR Abbreviation of *airport and airways surveillance radar.*

AB Abbreviation of *acquisition beacon* (NASA).

A-B In sound and acoustics, the direct comparison of two sources of sound by alternately turning on one and the other.

ab- **1.** Prefix that transforms the name of a practical electrical unit to that of the equivalent electromagnetic cgs unit (e.g., ABAMPERE, ABOHM, ABVOLT). See individual entries of such cgs units. **2.** Abbreviation for ABSOLUTE.

abac A graphic device for the solution of electronics problems. Also see ALIGNMENT CHART.

abampere The unit of current in the cgs electromagnetic system. One abampere equals 10 amperes and corresponds to 1 abcoulomb per second.

Abbe condenser **1.** In microscopy, a special two-piece lens that has enhanced light-gathering power. **2.** A similar focusing device in an electromagnetic antenna.

abbreviated dialing In telephone systems, special circuits requiring fewer-than-normal dialing operations to connect subscribers.

abc **1.** Abbreviation of AUTOMATIC BASS COMPENSATION, a system for boosting the volume of bass sounds at low amplifier gain. **2.** Abbreviation of AUTOMATIC BIAS CONTROL. **3.** Abbreviation of AUTOMATIC BRIGHTNESS CONTROL. **4.** Abbreviation of AUTOMATIC BRIGHTNESS COMPENSATION.

abcoulomb The unit of electrical quantity in the cgs electromagnetic system. One abcoulomb equals 10 coulombs and is the quantity of electricity that flows past any point in a circuit in one second when the current is one abampere.

aberration **1.** Distortion from perfect shape in a lens or reflecting mirror or antenna dish. **2.** A small error in the determination of the direction of a source of electromagnetic energy, on account of the motion of the source and/or the detecting apparatus. **3.** A small displacement in the apparent positions of the stars from month to month on account of the earth's orbital motion.

ABETS Acronym for *airborne beacon electronic test set* (NASA).

abfarad The unit of capacitance in the cgs electromagnetic system. One abfarad equals 10^9 farads and is the capacitance across which a charge of 1 abcoulomb produces a potential of 1 abvolt.

abhenry The unit of inductance in the cgs electromagnetic system. One abhenry equals 10^{-9} henry

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

and is the inductance across which a current that changes at the rate of 1 abampere per second induces a potential of 1 abvolt.

ABL Abbreviation of *Automated Biology Laboratory* (NASA).

abmho The obsolete unit of conductance and of conductivity in the cgs electromagnetic system. Replaced with ABSIEMENS.

abnormal dissipation Power dissipation higher or lower than the customary level, usually an overload.

abnormal oscillation **1.** Oscillation where none is desired or expected, as in an amplifier. **2.** Oscillation at two or more frequencies simultaneously when single-frequency operation is expected. **3.** Oscillation at an incorrect frequency. **4.** Parasitic oscillation.

abnormal propagation **1.** The chance shifting of the normal path of a radio wave, as by displacements in the ionosphere, so that reception is degraded. **2.** Unintentional radiation of energy from some point other than the transmitting antenna. **3.** Propagation over a path or in a direction not expected.

abnormal reflections Sharp, intense reflections at frequencies higher than the critical frequency of the ionosphere's ionized layer.

abnormal termination The shutdown of a running computer program or other process. Caused by the detection of an error by the associated hardware that indicates that some ongoing series of actions cannot be executed correctly.

abnormal triggering The false triggering or switching of a circuit or device, such as a flip-flop, by some undesirable source instead of the true trigger signal. Electrical noise pulses often cause abnormal triggering.

abohm The unit of resistance and of resistivity in the cgs electromagnetic system. One abohm equals 10^{-9} ohms and is the resistance across which a steady current of 1 abampere produces a potential difference of 1 abvolt.

abort To deliberately terminate an operation, experiment, process, or project before it has run its normal course.

AB power pack **1.** A portable dry-cell or wet-cell array containing both A and B batteries in one package. **2.** An ac-operated unit in one package for supplying A and B voltages to equipment normally operated from batteries.

abrasion machine An instrument for determining the abrasive resistance of a wire or cable.

abrasion resistance A measure of the ability of a wire or wire covering to resist mechanical damage.

ABS A basic programming abbreviation for the absolute value (of a number, variable, or expression).

abscissa **1.** The independent variable in a function. **2.** The axis (usually horizontal) on the graph of a function that indicates the independent variable.

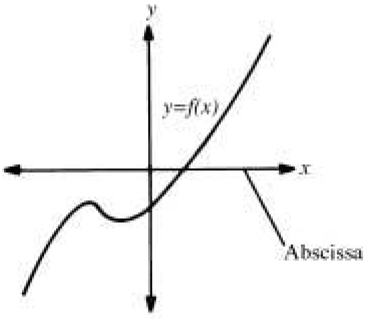

abscissa

absence-of-ground searching selector A rotary switch that searches for an ungrounded contact in a dial telephone system.

absiemens The unit of conductance or conductivity in the cgs electromagnetic system. One absiemens equals 10^9 siemens and is the conductance through which a potential of 1 abvolt forces a current of 1 abampere.

absolute **1.** A temperature scale in which zero represents the complete absence of heat. Units of measure are same as units on Celsius and Fahrenheit scales. See ABSOLUTE SCALE. **2.** Independent of any arbitrarily assigned units of measure or value.

absolute accuracy The full-scale accuracy of a meter with respect to a primary (absolute) standard.

absolute address In a digital computer program, the location of a word in memory, as opposed to location of the word in the program.

absolute code A computer code in which the exact address is given for storing or locating the reference operand.

absolute coding In computer practice, coding that uses absolute addresses.

absolute constant A mathematical constant that has the same value wherever it is used.

absolute delay The time elapsing between the transmission of two synchronized signals from the same station or from different stations, as in radio, radar, or loran. By extension, the time interval between two such signals from any source, as from a generator.

absolute digital position transducer A digital position transducer whose output signal indicates absolute position. (See ENCODER.)

absolute efficiency The ratio X_x/X_s, where X_x is the output of a given device, and X_s is the output of an ideal device of the same kind under the same operating conditions.

absolute encoder system A system that permits the encoding of any function (linear, nonlinear, continuous, step, and so on) and supplies a nonambiguous output.

absolute error The difference indicated by the approximate value of a quantity minus the actual value. This difference is positive when the approximate value is higher than the exact value, and it is negative when the approximate value is lower than the exact value. Compare RELATIVE ERROR.

absolute gain Antenna gain for a given orientation when the reference antenna is isolated in space and has no main axis of propagation.

absolute humidity The mass of water vapor per unit volume of air. Compare RELATIVE HUMIDITY.

absolute instruction A computer instruction that states explicitly and causes the execution of a specific operation.

absolute magnitude For a complex number quantity, the vector sum of the real and imaginary components (i.e., the square root of the sum of the squares of those components). Also see ABSOLUTE VALUE and IMPEDANCE.

absolute maximum rating The highest value a quantity can have before malfunction or damage occurs.

absolute maximum supply voltage The highest supply voltage that can be applied to a circuit without permanently altering its characteristics.

absolute measurement of current Measurement of a current directly in terms of defining quantities. **1.** TANGENT GALVANOMETER method: Current is proportional to the tangent of the angle of deflection of the needle of this instrument. Deflection depends on torque, resulting from the magnetic field produced by current in the galvanometer coil acting against the horizontal component of the earth's magnetic field. **2.** ELECTRODYNAMOMETER method: With this 2-coil instrument, current is determined from the observed deflection, the torque of the suspension fiber of the movable coil, and the coil dimensions.

absolute measurement of voltage Measurement of a voltage directly in terms of defining quantities. **1.** CALORIMETRIC method: A current-carrying coil immersed in water raises the temperature of the water. The difference of potential that forces the current through the coil then is determined in terms of the equivalent heat energy. **2.** Disk-electrometer method: In this setup, a metal disk attached to one end of a balance beam is attracted by a stationary disk mounted below it, the voltage being applied to the two disks. The other end of the beam carries a pan into which accurate weights are placed. At balance, the voltage is determined in terms of the weight required to restore balance, the upper-disk area, and the disk separation.

absolute minimum resistance The resistance between the wiper and the nearer terminal of a potentiometer, when the wiper is as close to that terminal as physically possible. All potentiometers have two such specifications, one for each end terminal.

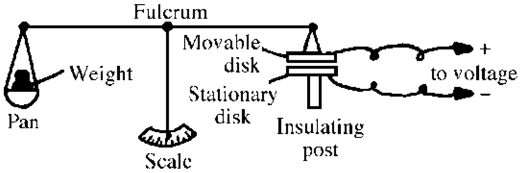

absolute measurement of voltage, 2

absolute Peltier coefficient The product of the absolute Seebeck coefficient and absolute temperature of a material.

absolute pitch A tone in a standard scale, determined according to the rate of vibration, independent of other tones in the range of pitch.

absolute pressure Pressure (force per unit area) of a gas or liquid determined with respect to that of a vacuum (taken as zero).

absolute-pressure transducer A transducer actuated by pressure from the outputs of two different pressure sources, and whose own output is proportional to the difference between the two applied pressures.

absolute scale **1.** A scale in which the zero value indicates the lowest physically possible value that a parameter can attain. **2.** A standard scale for measurement of a quantity. **3.** A universally agreed-upon scale for the determination of a variable quantity. **4.** The Kelvin temperature scale. **5.** The Rankine temperature scale.

absolute Seebeck coefficient The quotient, as an integral from absolute zero to the given temperature, of the Thomson coefficient of a material divided by its absolute temperature.

absolute spectral response The frequency output or response of a device in absolute power units (such as milliwatts) as opposed to relative units (such as decibels).

absolute system of units A system of units in which the fundamental (ABSOLUTE) units are those expressing length (l), mass (m), charge (q), and time (t). All other physical units, including practical ones, are then derived from these absolute units.

absolute temperature Temperature measured on either the Kelvin or Rankine scales, where zero represents the total absence of heat energy.

absolute temperature scale **1.** The Kelvin temperature scale, in which the divisions are equal in size to 1° Celsius, and the zero point is absolute zero, the coldest possible temperature, approximately –273.16° Celsius. **2.** The Rankine temperature scale, in which the divisions are equal in size to 1° Fahrenheit, and the zero point is absolute zero or approximately –459.7° Fahrenheit.

absolute tolerance The value of a component as it deviates from the specified or nominal value. It is usually expressed as a percentage of the specified value.

absolute units Fundamental physical units (see ABSOLUTE SYSTEM OF UNITS) from which all others are derived. See, for example, AMPERE, OHM, VOLT, and WATT.

absolute value The magnitude of a quantity without regard to sign or direction. The absolute value of a is written $|a|$. The absolute value of a positive number is the number itself; thus, $|10|$ equals 10. The absolute value of a negative number is the number with its sign changed: $|-10|$ equals 10.

absolute-value circuit A circuit that produces a unipolar signal in response to a bipolar input and in proportion to the absolute value of the magnitude of the input.

absolute-value computer A computer in which data is processed in its absolute form; i.e., every variable maintains its full value. (Compare to INCREMENTAL COMPUTER.)

absolute-value device In computer practice, a device that delivers a constant-polarity output signal equal in amplitude to that of the input signal. Thus, the output signal always has the same sign.

absolute zero The temperature –273.16°C (−459.7°F and 0 Kelvin). The coldest possible temperature, representing the complete absence of heat energy.

absorbed wave A radio wave that dissipates in the ionosphere as a result of molecular agitation. This effect is most pronounced at low and medium frequencies.

absorptance The amount of radiant energy absorbed in a material; equal to 1 minus the transmittance.

absorption The taking up of one material or medium by another into itself, as by sucking or soaking up. Also, the retention of one medium (or a part of it) by another medium, through which the first one attempts to pass. See, for example, ABSORBED WAVE, ABSORPTION COEFFICIENT, DIELECTRIC ABSORPTION. Compare ADSORPTION.

absorption band See ABSORPTION SPECTRUM.

absorption circuit A circuit that absorbs energy from another circuit or from a signal source—especially a resonant circuit, such as a wavemeter or wavetrap.

absorption current In a capacitor, the current resulting from absorption of energy by the dielectric material.

absorption dynamometer A power-measuring instrument in which a brake absorbs energy from a revolving shaft or wheel.

absorption fading Fading of a radio wave, resulting from (usually) slow changes in the absorption of the wave in the line of propagation.

absorption frequency meter See WAVEMETER.

absorption line See ABSORPTION SPECTRUM.

absorption loss **1.** Transmission loss caused by dissipation of electrical energy, or conversion of it into heat or other forms of energy. **2.** Loss of all or part of a skywave because of absorption by the ionosphere. Also called *ionospheric absorption* or *atmospheric absorption.*

absorption marker A small blip introduced onto an oscilloscope trace to indicate a frequency point. It is so called because it is produced by the action of a frequency-calibrated tuned trap, similar to an absorption wavemeter.

absorption modulation Amplitude modulation of a transmitter or oscillator by means of an audio-frequency-actuated absorber circuit. In its simplest form, the modulator consists of a few turns of wire coupled to the transmitter tank coil and connected to a carbon microphone. The arrangement absorbs energy from the transmitter at a varying rate as the microphone changes its resistance in accordance with the sound waves it receives.

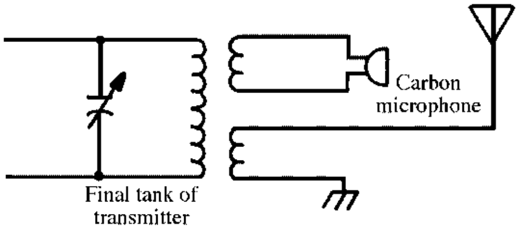

absorption modulation

absorption spectrum For electromagnetic waves, a plot of absorption coefficient (of the medium of propagation) versus frequency. Also called EMISSION SPECTRUM.

absorption trap See WAVETRAP.

absorption wavemeter A resonant-frequency indicating instrument that is inductively coupled to the device under test.

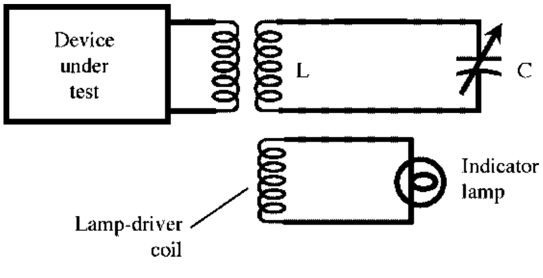

absorption wavemeter

absorptivity In audio and microwave technologies, a measure of the energy absorbed by a given volume of material.

A-B test Comparison of two sounds by reproducing them in alternating succession.

abvolt The unit of potential difference in the cgs electromagnetic system. One abvolt equals 10^{-8} V and is the difference of potential between any two points when 1 erg of work is required to move 1 abcoulomb of electricity between them.

abwatt The unit of power in the cgs electromagnetic system. One abwatt equals 10^{-7} W and is the power corresponding to 1 erg of work per second.

ac **1.** Abbreviation of ALTERNATING CURRENT. **2.** Abbreviation of ATTITUDE CONTROL. **3.** Abbreviation of AERODYNAMIC CENTER. **4.** A suffix meaning AUTOMATIC CALCULATOR or AUTOMATIC COMPUTER.

a/c **1.** Abbreviation of AIRCRAFT. **2.** Abbreviation of AIR CONDITIONING.

Ac Symbol for ACTINIUM.

ACA Abbreviation of *automatic circuit analyzer.*

ac base current Symbol, $I_{B(ac)}$. The ac component of base current in a bipolar transistor.

ac base resistance Symbol, $R_{B(\mathrm{ac})}$. The dynamic base resistance in a bipolar transistor.

ac base voltage Symbol, $V_{B(\mathrm{ac})}$. The ac component of base voltage in a bipolar transistor. It is the ac input signal voltage in a common-emitter amplifier or emitter-follower amplifier.

ac bias In a tape recorder, the high-frequency current that passes through the recording head to linearize operation.

acc **1.** Abbreviation of AUTOMATIC CHROMINANCE CONTROL. **2.** Abbreviation of AUTOMATIC COLOR COMPENSATION. **3.** Abbreviation of ACCELERATION.

ac cathode current Symbol, $I_{K(\mathrm{ac})}$. The ac component of cathode current in an electron tube.

ac cathode resistance Symbol, $R_{K(\mathrm{ac})}$. The dynamic cathode resistance in an electron tube. $R_{K(\mathrm{ac})}$ equals $\mathrm{d}V_K/\mathrm{d}I_K$ for a constant value of V_G.

ac cathode voltage Symbol, $V_{K(ac)}$. The ac component of cathode voltage in an electron tube. It is the ac output signal voltage in cathode-follower and grounded-grid amplifiers.

accelerated life test A test program that simulates the effects of time on devices or apparatus, by artificially speeding up the aging process.

accelerated service test A service or bench test in which equipment or a circuit is subjected to an extreme condition in an attempt to simulate the effects of average use over a long time.

accelerating conductor or relay A conductor or relay that prompts the operation of a succeeding device in a starting mode according to established conditions.

accelerating electrode In a cathode-ray tube or klystron, the electrode to which the accelerating voltage is applied.

accelerating time The elapsed time that starts when voltage is applied to a motor, and ends when the motor shaft reaches maximum speed.

accelerating voltage A positive high voltage applied to the accelerating electrode of a cathode-ray tube to increase the velocity of electrons in the beam.

acceleration at stall The angular acceleration of a servomotor at stall, determined from the stall torque and the moment of inertia of the motor's rotor.

acceleration derivative Acceleration (a) expressed as the second derivative of distance (s) with respect to time (t): a equals d^2s/dt^2.

acceleration potential See ACCELERATING VOLTAGE.

acceleration switch A switch that operates automatically when the acceleration of a body to which it is attached exceeds a predetermined rate in a given direction.

acceleration time The time required by a computer to take in or deliver information after interpreting instructions. Compare ACCESS TIME.

acceleration torque During the accelerating period of a motor, the difference between the torque demanded and the torque actually produced by the motor.

acceleration voltage The potential between accelerating elements in a vacuum tube, the value of which determines average electron velocity.

accelerometer A transducer whose output voltage is proportional to the acceleration of the moving body to which it is attached.

accentuation The emphasis of a desired band of frequencies, usually in the audio-frequency spectrum.

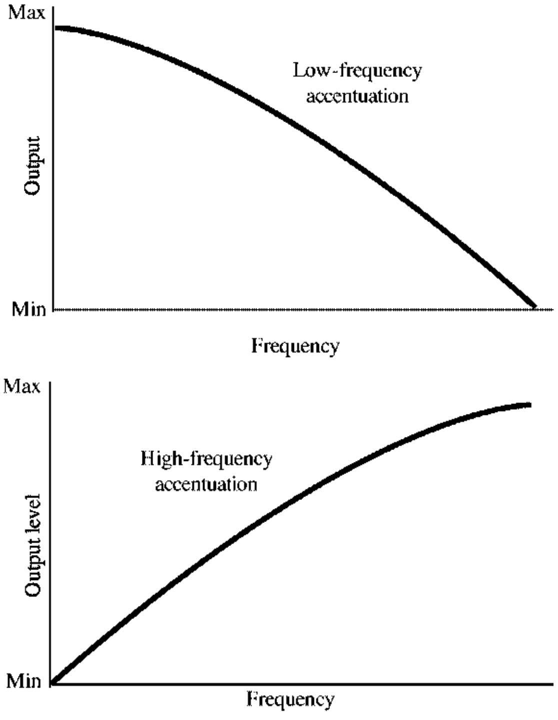

accentuation

accentuator A circuit or device, such as a filter, tone control, or equalizer, used to emphasize a band of frequencies, usually in the audio-frequency spectrum. Also see ACCENTUATION.

acceptable-environmental-range test A test to disclose the environmental conditions that equipment can endure while maintaining at least the minimum desired reliability.

acceptable quality level Abbreviation, AQL. A percentage that represents an acceptable average of defective components allowable for a process, or the lowest quality that a supplier is permitted to regularly present for acceptance.

acceptance sampling plan A probabilistic method of sampling a quantity of units from a lot, and determining from the sample whether to accept the lot, reject the lot, or perform another sampling.

acceptance test A test performed on incoming equipment or on submitted samples to determine if they meet tester's or supplier's specifications.

acceptor **1.** Any device or circuit, such as a series-resonant circuit, that provides relatively easy transmission of a signal, in effect accepting the signal. **2.** A hole-rich impurity added to a semiconductor to make the latter p-type. It is so called because its holes can accept electrons. Compare DONOR.

acceptor circuit See ACCEPTOR, **1.**

acceptor impurity See ACCEPTOR, **2.**

access **1.** To gain entrance to something, such as the interior of the cabinet of a high-fidelity amplifier. **2.** In a computer, the action of going to a specific memory location for the purpose of data retrieval. **3.** A port or opening into a piece of equipment, placed there to make the equipment easy to maintain and repair.

access arm A mechanical device that positions the read/write mechanism in a computer storage unit.

access control register A register that is part of a computer protection system that prevents interference between different software modules.

access method A method of transferring information or data from main storage to an input/output unit.

access right The access status given to computer system users that indicates the method of access permitted (e.g., read a file only or write to a file).

access time The time required by a computer to begin delivering information after the memory or storage has been interrogated.

accidental error An unintentional error committed by a person making measurements and recording data.

accidental triggering The undesired chance-operation of a flip-flop or other switching circuit caused by a noise pulse or other extraneous signal.

ac collector current Symbol, $I_{C(ac)}$. The ac component of collector current in a bipolar transistor.

ac collector resistance Symbol, $R_{C(ac)}$. The dynamic collector resistance of a bipolar transistor. $R_{C(ac)}$ equals $\mathrm{d}V_C/\mathrm{d}I_C$ for a constant value of base current I_B (in a common-emitter circuit) or emitter current I_E (in a common-base circuit).

ac collector voltage Symbol, $V_{C(ac)}$. The ac component of collector voltage in a bipolar transistor. The ac output signal voltage in a common-emitter or common-base amplifier.

accompanying audio channel The RF signal that supplies television sound. Also called *Cochannnel sound frequency.*

ac component In a complex wave (i.e., one containing both ac and dc), the alternating, fluctuating, or pulsating part of the combination. Compare DC COMPONENT.

accordion A printed-circuit connector contact with a Z-shaped spring that allows high deflection with low fatigue.

ac-coupled flip-flop A flip-flop that is operated by the rise or fall of a clock pulse.

ac coupling Transformer coupling or capacitive coupling, which transmit ac, but not dc. Compare DIRECT COUPLING.

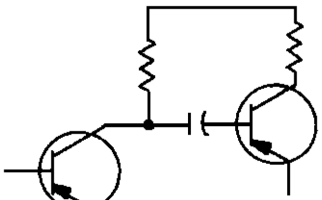

Capacitive

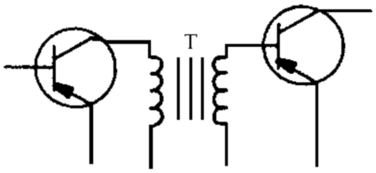

Inductive

ac coupling

accumulator **1.** In a digital computer, a circuit or register device that receives numbers, totals them, and stores them. **2.** Storage battery.

accuracy **1.** Precision in the measurement of quantities and in the statement of physical characteristics. **2.** Degree of precision. Usually expressed, in terms of error, as a percentage of the specified value (e.g., 10 V plus or minus 1%), as a percentage of a range (e.g., 2% of full scale), or as parts (e.g., 100 parts per million).

accuracy rating The maximum error in an instrument, given as a percentage of the full-scale value.

accw Abbreviation of ALTERNATING-CURRENT CONTINUOUS WAVE.

ac/dc Abbreviation of ALTERNATING CURRENT/DIRECT CURRENT. Pertains to equipment that will operate from either ac utility power or a dc power source. A notebook computer is a good example.

ac directional overcurrent relay A relay that works on a specific value of alternating overcurrent that is rectified for a desired polarity.

ac drain current Symbol, $I_{D(ac)}$. The ac component of drain current in a field-effect transistor.

ac drain resistance Symbol, $R_{D(ac)}$. The dynamic drain resistance in a field-effect transistor; $R_{D(ac)}$ equals dV_D/dI_D for a constant value of gate voltage V_G.

ac drain voltage Symbol, $V_{D(ac)}$. The ac component of drain voltage in a field-effect transistor. The ac output signal voltage in a common-source FET amplifier.

ac dump The removal of all ac power from a system or component.

ac emitter current Symbol, $I_{E(ac)}$. The ac component of emitter current in a bipolar transistor.

ac emitter resistance Symbol, $R_{E(ac)}$. The dynamic emitter resistance of a bipolar transistor; $R_{E(ac)}$ equals dV_E/dI_E for a constant value of base current I_B (in an emitter-follower circuit) or collector voltage V_{CC} (in a common-base circuit).

ac emitter voltage Symbol, $V_{E(ac)}$. The ac component of emitter voltage in a bipolar transistor. The ac input signal voltage in a common-base amplifier; the ac output signal voltage in an emitter-follower amplifier.

ac equipment An apparatus designed for operation from an ac power source only. Compare DC EQUIPMENT and AC/DC.

ac erasing In tape recording, the technique of using an alternating magnetic field to erase material already recorded on the tape.

ac erasing head Also called *ac erase head.* In tape and wire recording, a head that carries alternating current to erase material already recorded on the tape or wire. Also see AC ERASING.

acetate Cellulose acetate, a tough thermoplastic material that is an acetic acid ester of cellulose. It is used as a dielectric and in the manufacture of photographic films.

acetate base **1.** The cellulose acetate film that served as the base for the magnetic oxide coating in early recording tape. Most such tapes today are of polyester base. **2.** The cellulose acetate substrate onto which certain photosensitive materials are deposited for lithographic reproduction. Also see ACETATE and ANCHORAGE.

acetate tape Recording tape consisting of a magnetic oxide coating on a cellulose acetate film. Also see ACETATE BASE.

ac gate voltage Symbol, $V_{G(ac)}$. The ac component of gate voltage in a field-effect transistor. The ac input signal voltage.

ac generator **1.** A rotating electromagnetic machine that produces alternating current (e.g., a dynamo or alternator). **2.** An oscillator or combination of an oscillator and an output amplifier.

ac grid voltage Symbol, $V_{G(ac)}$. The ac component of control grid voltage in an electron tube. The ac input signal voltage in a common-cathode amplifier or cathode follower.

A channel The left channel of a two-channel stereo system.

achieved reliability A statement of reliability based on the performance of mass-produced parts or systems under similar environmental conditions. Also called OPERATIONAL RELIABILITY.

achromatic **1.** Without color. In a TV image, the tones from black through gray to white. The term occasionally refers to black-and-white television, although MONOCHROMATIC is more often used in this sense.

achromatic locus Also called *achromatic region.* An area on a chromaticity diagram that contains all points, representing acceptable reference white standards.

achromatic scale A musical scale without accidentals.

ACIA Abbreviation of *asynchronous communications interface adapter.*

acicular Pertaining to the shape of magnetic particles on recording tape. Under magnification, these particles look like thin rods.

acid A substance that dissociates in water solution and forms hydrogen (H) ions (e.g., sulfuric acid). Compare BASE, **2.**

acid depolarizer Also called *acidic depolarizer.* An acid, in addition to the electrolyte, used in some primary cells to slow the process of polarization.

ac line A power line that delivers alternating current only.

ac line filter A filter designed to remove extraneous signals or electrical noise from an ac power line, while causing virtually no reduction of the power-line voltage or power.

ac line voltage The voltage commonly delivered by the commercial power line to consumers. In the United States, the two standards are 117 V and 234 V (~ about 5 percent). The lower voltage is used by most appliances; the higher voltage is intended for appliances and equipment that draws high power, such as electric ovens, cooking ranges, clothes dryers, and amateur-radio amplifiers. In Europe, 220 V is the common standard.

aclinic line Also called *magnetic equator.* An imaginary line drawn on a map of the world or of an area that connects points of zero inclination (dip) of the needle of a magnetic compass.

ACM Abbreviation for *Association for Computing Machinery.*

ac magnetic bias See AC BIAS.

ac meter A meter that is intended to work only on alternating current or voltage. Such meters include iron-vane and rectifier types.

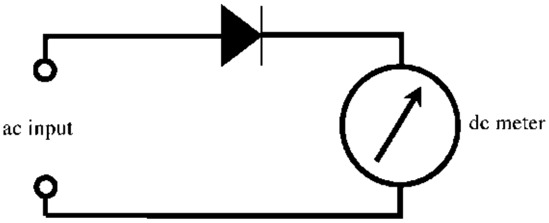

ac meter

ac noise **1.** Electromagnetic interference originating in the ac power lines. **2.** Electrical noise of a rapidly alternating or pulsating nature.

ac noise immunity In computer practice, the ability of a logic circuit to maintain its state, despite excitation by ac noise.

acous Abbreviation for ACOUSTIC.

acoustic Pertaining to audible sound disturbances, usually in air (versus audio-frequency currents or voltages).

acoustic absorption The assimilation of energy from sound waves passing through or reflected by a given medium.

acoustic absorption loss That portion of sound energy lost (as by dissipation in the form of heat) because of ACOUSTIC ABSORPTION.

acoustic absorptivity The ratio of sound energy absorbed by a material to sound energy striking the surface of the material.

acoustic attenuation constant The real-number component of the complex acoustical propagation constant, expressed in nepers per unit distance.

acoustic burglar alarm An alarm that receives the noise made by an intruder. The alarm device responds to the impulses from concealed microphones.

acoustic capacitance The acoustic equivalent of electrical capacitance.

acoustic clarifier In a loudspeaker system, a set of cones attached to the baffle that vibrate to absorb and suppress sound energy during loud bursts.

acoustic communication Communications by means of sound waves. This can be through the atmosphere, or it can be through solids or liquids, such as a taut wire, a body of water, or the earth.

acoustic compliance COMPLIANCE in acoustic transducers, especially loudspeakers. It is equivalent to electrical capacitive reactance.

acoustic consonance An effect that occurs when two objects are near each other but not in physical contact, and both have identical or harmonically related resonant frequencies. An example is shown by two tuning forks with identical fundamental frequencies. If one fork is struck and then brought near the other, the second fork will begin vibrating. If the second fork has a fundamental frequency that is a harmonic of the frequency of the first fork, the second fork will vibrate at its own resonant frequency. See HARMONIC, RESONANCE.

acoustic coupling Data transfer via a sound link between a telephone and a pickup/reproducer. Was once common in computer terminals and facsimile machines. This scheme has been largely replaced by hard wiring and optical coupling.

acoustic damping The deadening or reduction of the vibration of a body to eliminate (or cause to die out quickly) sound waves arising from it.

acoustic delay line Any equivalent of a special transmission line that introduces a useful time delay between input and output signals. In one form, it consists of a crystal block or bar with an input transducer at one end and an output transducer at the other. An electrical input signal in the first transducer sets up sound waves that travel through the interior of the crystal; the piezoelectric reaction of the crystal to sound vibrations sets up an output voltage in the second transducer. The delay is caused by the time required for the acoustic energy to travel the length of the crystal bar.

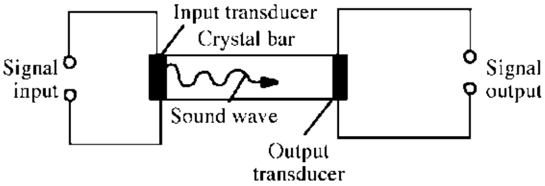

acoustic delay line

acoustic depth finder A direct-reading device for determining the depth of a body of water, or for locating underwater objects via sonic or ultrasonic waves transmitted downward and reflected back to the instrument.

acoustic dispersion Variation of the velocity of sound waves, depending on their frequency.

acoustic elasticity **1.** In a loudspeaker enclosure, the compressibility of air behind the vibrating cone of the speaker. **2.** In general, the compressibility of any medium through which sound passes.

acoustic electric transducer A transducer, such as a microphone or hydrophone, that converts sound energy into electrical energy. Compare ELECTRICAL/ACOUSTIC TRANSDUCER. Also see ACOUSTIC TRANSDUCER.

acoustic feedback A usually undesirable effect that occurs when sound waves from a loudspeaker (or other reproducer) reach a microphone (or other input transducer) in the same system.

This can cause an amplifier to oscillate, with a resultant rumbling, howling, or whistling.

acoustic filter Any sound-absorbing or transmitting arrangement, or combination of the two, that transmits sound waves of desired frequency while attenuating or eliminating others.

acoustic frequency response The sound-frequency range as a function of sound intensity. A means of describing the performance of an acoustic device.

acoustic generator A device that produces sound waves of a desired frequency and/or intensity. Examples are electrical devices (headphones or loudspeakers operated from a suitable oscillator, buzzer, bell, or flame) and mechanical devices (tuning forks, bells, string, or whistles).

acoustic grating A set of bars or slits that are parallel to one another and arranged a fixed distance apart so that an interference pattern forms as sound passes through. Used to determine the wavelength of acoustic waves.

acoustic homing system **1.** A system that uses a sound signal for guidance purposes. **2.** A guidance method in which a missile homes in on noise generated by a target.

acoustic horn A tapered tube (round or rectangular, but generally funnel-shaped) that directs sound and, to some extent, amplifies it. So called to distinguish it from a microwave horn.

acoustic howl See ACOUSTIC FEEDBACK.

acoustician **1.** A person skilled in acoustics (an acoustics technician). **2.** An AUDIOLOGIST.

acoustic impedance Unit, ACOUSTIC OHM. The acoustic equivalent of electrical impedance. Like the latter, acoustic impedance is the total opposition encountered by acoustic force. Also like electrical impedance, acoustic impedance has resistive and reactive components: ACOUSTIC RESISTANCE and ACOUSTIC REACTANCE.

acoustic inductance Also called *inertance.* The acoustic equivalent of electrical inductance.

acoustic inertance See ACOUSTIC INDUCTANCE.

acoustic inhibition See AUDITORY INHIBITION.

acoustic intensity See SOUND INTENSITY.

acoustic interferometer An instrument that evaluates the frequency and velocity of sound waves in a liquid or gas, in terms of a standing wave set up by a transducer and reflector as the frequency or transducer-to-reflector distance varies.

acoustic labyrinth A loudspeaker enclosure whose internal partitions form a maze-like path or "tube" lined with sound-absorbing material. The tube effectively runs from the back of the speaker down to where it terminates in a MOUTH or PORT that opens at the front of the enclosure. The labyrinth provides an extremely efficient reproduction system because of its excellent acoustic impedance-matching capability.

acoustic lens A system of barriers that refracts sound waves the way that an optical lens does with light waves.

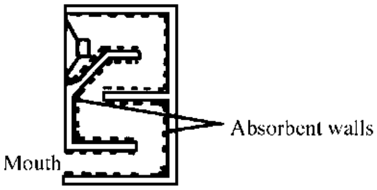

acoustic labyrinth

acoustic line Baffles or other such structures within a speaker that act as the mechanical equivalent of an electrical transmission line to enhance the reproduction of very low bass frequencies.

acoustic load A device that serves simultaneously as the output load of an amplifier and as a transducer of electrical energy into acoustic energy (e.g., headphones or a loudspeaker).

acoustic memory In a computer, a volatile memory element employing an acoustic delay line, often incorporating quartz or mercury as the transmission and delay element.

acoustic mirage A type of sound distortion in which the listener experiences the illusion of two sound sources when there is only one. The phenomenon is caused by the effect of a large temperature gradient in the air or water through which the sound passes.

acoustic mode Crystal-lattice vibration without producing an oscillating dipole.

acoustic noise Interferential (usually disagreeable) sounds carried by the air (or other propagation medium) to the ear or to an acoustic transducer. This is in contrast to electrical noise, which consists of extraneous current or voltage impulses and is inaudible until converted into sound.

acoustic ohm The unit of acoustic resistance, reactance, or impedance. One acoustic ohm equals the volume velocity of 1 cm/s produced by a sound pressure of 1 microbar (0.1 Pa). Also called *acoustical ohm.*

acoustic phase constant The imaginary-number component of the complex acoustic propagation constant expressed in radians per second or radians per unit distance.

acoustic phase inverter A bass reflex loudspeaker enclosure.

acoustic pressure **1.** The acoustic equivalent of electromotive force, expressed in dynes per square centimeter; also called *acoustical pressure.* **2.** Sound pressure level.

acoustic propagation The transmission of sound waves, or subaudible or ultrasonic waves, as a disturbance in a medium, rather than as an electric current or electromagnetic field.

acoustic radiator A device that emits sound waves. Examples are the cone of a loudspeaker, the diaphragm of a headphone, and the vibrating reed of a buzzer.

acoustic radiometer An instrument for measuring the intensity of a sound wave (see SOUND INTENSITY) in terms of the unidirectional steady-state pressure exerted at a boundary as a result of absorption or reflection of the wave.

acoustic reactance Unit, ACOUSTIC OHM. The imaginary-number component of ACOUSTIC IMPEDANCE. It can take the form of ACOUSTIC CAPACITANCE or ACOUSTIC INDUCTANCE.

acoustic reflectivity The ratio F_r/F_i, where F_r is the rate of flow of sound energy reflected from a surface and F_i is the rate of flow of sound energy incident to the surface.

acoustic refraction The deflection of sound waves being transferred obliquely between media that transmit sound at different speeds.

acoustic regeneration See ACOUSTIC FEEDBACK.

acoustic resistance Unit, ACOUSTIC OHM. The real-number component of ACOUSTIC IMPEDANCE. The opposing force that causes acoustic energy to be dissipated in the form of heat. It is attributed to molecular friction in the medium through which sound passes. See ACOUSTIC OHM.

acoustic resonance In an enclosed chamber with walls that reflect sound waves, resonance that occurs at certain wavelengths because the echoes combine in and out of phase. Speaker enclosures almost always have resonance at certain frequencies. This effect can be used to an advantage when it is necessary to get good bass (low-frequency) response from a relatively small speaker.

acoustic resonator **1.** A chamber, such as a box, cylinder, or pipe, in which an air column resonates at a particular frequency. **2.** A piezoelectric, magnetostrictive, or electrostrictive body that vibrates at a resonant audio frequency that is governed by the mechanical dimensions of the body when an audio voltage at that frequency is applied.

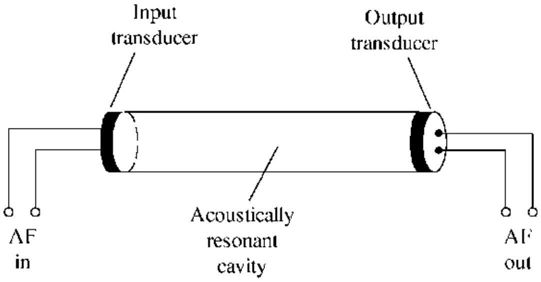

acoustic resonator, 1

acoustics **1.** The physics of sound. The study and applications of acoustic phenomena. **2.** The qualities of an enclosure or sound chamber (room, auditorium, or box) that describe how sound waves behave in it.

acoustic scattering The spreading of a sound wave in many directions as a result of diffraction, reflection, or refraction.

acoustic suspension A loudspeaker design that allows exceptional low-frequency reproduction for a fairly small physical size. An airtight enclosure is used to increase the tension on the speaker cone.

acoustic system **1.** A coordinated array of acoustic components (e.g., acoustic filters, resonators, etc.) that responds to sound energy in a predetermined manner. **2.** An audio-frequency system in which sound energy is converted into electrical energy, processed, and then reconverted into sound energy for a clearly defined purpose.

acoustic telegraph A telegraph that gives audible signals, as opposed to visual signals or printed messages.

acoustic transducer **1.** Any device, such as headphones or a loudspeaker, for converting audio-frequency electrical signals into sound waves. **2.** Any device, such as a microphone, for converting sound waves into alternating, pulsating, or fluctuating currents.

acoustic transmission The direct transmission of sound energy without the intermediary of electric currents.

acoustic transmission system A set of components designed to generate acoustic waves.

acoustic transmissivity Also called *acoustic transmitivity*. The ratio e_t/e_i, where e_t is the sound energy transmitted by a medium, and e_i is the incident sound energy reaching the surface of the medium. Acoustic transmissivity is proportional to the angle of incidence.

acoustic treatment Application of sound-absorbing materials to the interior of an enclosure or room to control reverberation.

acoustic wave The traveling vibration, consisting of molecular motion, via which sound is transmitted through a gas, liquid or solid. Usually refers to sound waves in air.

acoustic wave filter See ACOUSTIC FILTER.

acoustoelectric effect The generation of a voltage across the faces of a crystal by sound waves traveling longitudinally through the crystal.

acoustoelectronics A branch of electronics concerned with the interaction of sound energy and electrical energy in devices, such as surface-wave filters and amplifiers. In such devices, electrically induced acoustic waves travel along the surface of a piezoelectric chip and generate electrical energy. Also called *praetersonics* and *microwave acoustics*.

ac plate current Symbol, $I_{P(ac)}$. The ac component of plate current in a vacuum tube.

ac plate resistance Symbol, $R_{P(ac)}$. The dynamic plate resistance of an electron tube. $R_{P(ac)}$ equals dE_P/dI_P, where E_P is the plate voltage and I_P is the plate current, for a constant value for grid voltage E_G.

ac plate voltage Symbol, $E_{P(ac)}$. The ac component of plate voltage in an electron tube. The ac output-signal voltage in a common-cathode amplifier.

ac power Symbol, P_{ac}. Unit, watt (W). The power acting in an ac circuit, P_{ac} equals $EI \cos q$, where E is in volts, I in amperes, and q is the phase angle. Compare DC POWER. Also see POWER.

ac power supply A power unit that supplies ac only (e.g., ac generator, vibrator-transformer, oscillator, or inverter). Compare DC POWER SUPPLY.

acquisition **1.** The gathering of data from transducers or a computer. **2.** Locating the path of an orbiting body for purposes of collecting telemetered data. **3.** Orienting an antenna for optimum pickup of telemetered data.

acquisition and tracking radar An airborne or ground radar, which locks in on a strong signal and tracks the body that reflects (or transmits) the signal.

acquisition radar A radar that spots an oncoming target and supplies position data regarding the target to a fire-control or missile-guidance radar, which then tracks the target.

acr **1.** Abbreviation of AUDIO CASSETTE RECORDER. **2.** Abbreviation of AUDIO CASSETTE RECORDING SYSTEM.

ac reclosing relay The controlling component in an alternating-current circuit breaker. It causes the breaker to reset after a specified period of time.

ac relay A relay designed to operate on alternating current without chattering or vibrating.

ac resistance Pure resistance in an ac circuit. Unlike reactance and impedance, which are also forms of opposition to the flow of current, ac resistance introduces no phase shift.

acronym A word formed from letters or syllables taken from other applicable words of a multiword term. Acronyms are convenient for naming new devices and processes in electronics. Usually, a term is considered an acronym only when it is spelled in all-capital letters; once the term is accepted and popularized, it is written as a conventional word and is no longer thought of as an acronym. For example, *LASER* was once an acronym for light amplification by the stimulated emission of radiation. By the popularization process, the acronym became a conventional word from which other terms (such as the verb “lase”) were derived.

acrylic resin A synthetic resin used as a dielectric and in electronic encapsulations. It is made from acrylic acid or one of its derivatives.

ACS Abbreviation of *automatic control system.*

ac source current Symbol, $I_{S(ac)}$. The ac component of source current in a field-effect transistor.

ac source resistance Symbol, $R_{S(ac)}$. The dynamic source resistance in a field-effect transistor; $R_{S(ac)}$ equals dV_S/dI_S for a constant value of V_G.

ac source voltage Symbol, $V_{S(ac)}$. The ac component of source voltage in a field-effect transistor. The ac output-signal voltage in a source-follower (grounded-drain) FET amplifier.

acss Abbreviation of *analog computer subsystem.*

ac time overcurrent relay A device with a certain time characteristic, which breaks a circuit when the current exceeds a certain level.

actinic rays Short-wavelength light rays in the violet and ultraviolet portion of the spectrum that give conspicuous photochemical action.

actinism The property whereby radiant energy (such as visible and ultraviolet light, X-rays, etc.) causes chemical reactions.

actinium Symbol, Ac. A radioactive metallic element. Atomic number, 89. Atomic weight, 227.

actinodielectric Exhibiting a temporary rise in electrical conductivity during exposure to light.

actinoelectric effect The property whereby certain materials (such as selenium, cadmium sulfide, germanium, and silicon) change their electrical resistance or generate a voltage on exposure to light. Also see ACTINODIELECTRIC.

actinometer An instrument for measuring the direct heating power of the sun’s rays or the actinic power of a light source.

action current A small transient current that flows in a nerve in the human body as a result of stimulation.

activate To start an operation, usually by applying an appropriate enabling signal.

activation **1.** Supplying electrolyte to a battery cell to prepare the cell for operation. **2.** Causing the acceleration of a chemical reaction.

activation time In the activation of a battery cell (see ACTIVATION, **1**), the interval between addition of the electrolyte and attainment of full cell voltage.

activator A substance added to an accelerator (see ACCELERATOR, **3**) to speed the action of the accelerator.

active Pertaining to a circuit or device that requires a power supply for its operation. This differs from a passive circuit or device, which operates with no external source of power.

active antenna An antenna that uses a small whip, loop, or ferrite loopstick with a high-gain amplifier for receiving at very-low, low, medium, and high radio frequencies (approximately 9 kHz to 30 MHz).

active area The forward-current-carrying portion of the rectifying junction of a metallic rectifier.

active arm See ACTIVE LEG.

active balance In telephone repeater operation, the sum of return currents at a terminal network balanced against the local circuit or drop resistance.

active chord mechanism Abbreviation, ACM. In robots, an electromechanical gripper capable of conforming to irregular objects. It has a structure similar to the human spine, with numerous small, rigid links connected by hinges.

active communications satellite A satellite containing receivers (which pick up beamed electromagnetic signals from a ground point and amplify them) and transmitters (which send signals back to the surface of the earth). Also called *active comsat.* Compare PASSIVE COMMUNICATIONS SATELLITE.

active component **1.** A device capable of some dynamic function (such as amplification, oscillation, or signal control) that usually requires a power supply for its operation. Examples include bipolar transistors, field-effect transistors, and integrated circuits. Compare PASSIVE COMPONENT. **2.** In an ac circuit, a quantity that contains no reactance so that the current is in phase with the voltage.

active component of current See ACTIVE CURRENT.

active computer A computer in an installation or network that is processing data.

active comsat See ACTIVE COMMUNICATIONS SATELLITE.

active control system A device or circuit that compensates for irregularities in the operating environment.

active current In an ac circuit, the current component that is in phase with the voltage. This is in contrast to reactive current, which is not in phase with the voltage, and is “inactive,” with respect to power in the circuit. The active current is equal to the average power divided by the effective voltage.

active decoder An automatic ground-station device that gives the number or letter designation of a received radio beacon reply code.

active device **1.** An electronic component, such as a transistor that needs a power supply, and/or that is capable of amplifying. **2.** Broadly, any device (including electromechanical relays) that can switch (or amplify) by application of low-level signals.

active electric network A network containing one or more active devices or components, usually amplifiers or generators, in addition to passive devices or components.

active element The driven or RF-excited element in a multielement antenna or antenna array.

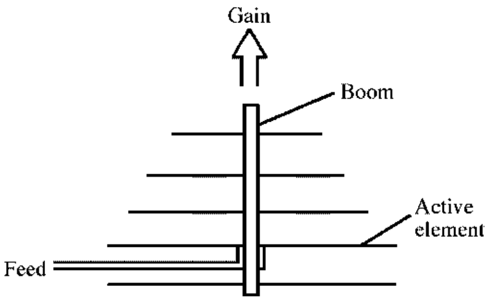

active element

active file A computer file in use (i.e., one that is being updated or referred to).

active filter A bandpass, bandstop, highpass or lowpass filter, consisting of resistors, capacitors, and operational amplifiers, arranged to pass a desired frequency response. Commonly used at audio frequencies.

active infrared detection Detection of infrared rays reflected from a target to which they were beamed.

active jamming Transmission or retransmission of signals for the purpose of disrupting communications.

active junction A pn junction in a semiconductor device that has been created by a diffusion process.

active leg An element within a transducer that changes one or more of its electrical characteristics in response to the input signal of the transducer. Also called *active arm.*

active lines In a U.S. television picture, the lines (approximately 488) that make up the picture. The remaining 37 of the 525 available lines are blanked and are called INACTIVE LINES.

active material **1.** In a storage cell, the chemical material in the plates that provides the electrical action of the cell, as distinguished from the supporting material of the plates themselves. **2.** A radioactive substance. **3.** The phosphor coating of a cathode-ray tube screen. **4.** The material used to coat an electron-tube cathode.

active mixer A signal mixer using one or more active components, such as transistors or integrated circuits. An active circuit provides amplification, input-output isolation, and high input impedance, in addition to the mixing action. Compare PASSIVE MIXER.

active modulator A modulator using one or more active components, such as transistors or integrated circuits. An active circuit provides gain, input-output isolation, and high input impedance, in addition to modulation. Compare PASSIVE MODULATOR.

active network See ACTIVE ELECTRIC NETWORK.

active pressure The electromotive pressure that produces a current in an ac circuit.

active pull-up An arrangement using a transistor as a pull-up resistor replacement in an integrated circuit, providing low output impedance and low power consumption.

active RC network **1.** A resistance-capacitance (RC) circuit that contains active components (transistors or integrated circuits), as well as passive components (capacitors and resistors). **2.** An RC network in which some or all of the resistors and capacitors are simulated by the action of active components.

active repair time The time during which maintenance is done on a system and the system is out of operation.

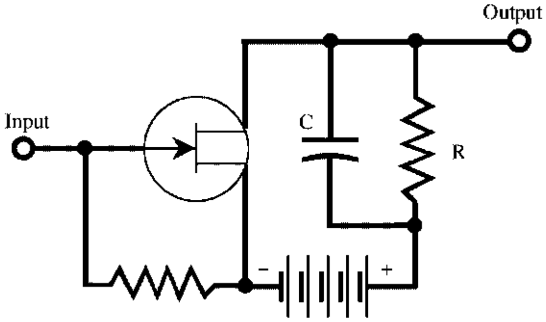

active RC network

active satellite See ACTIVE COMMUNICATIONS SATELLITE.

active sensor In an electronic security system, a transducer that generates an electromagnetic field or acoustic-wave field, and detects changes in the field resulting from the presence or movement of objects in the vicinity.

active substrate In an integrated circuit, a substrate consisting of single-crystal semiconductor material into which the components are formed; it acts as some or all of the components. This is in contrast to a substrate consisting of a dielectric, where the components are deposited on the surface.

active system A radio and/or radar system that requires transmitting equipment to be carried in a vehicle.

active tracking system A system in which a transponder or responder on board a vehicle retransmits information to tracking equipment (e.g., *azusa, secor*).

active transducer **1.** A transducer that contains an active device, such as a transistor or integrated circuit, for immediate amplification of the sensed quantity. **2.** A transducer that is itself an active device.

active wire In the armature of a generator, a wire experiencing induction and, therefore, is delivering voltage.

activity **1.** Intensity of, as well as readiness for, oscillation in a piezoelectric crystal. **2.** Radioactive intensity. **3.** Intensity of thermal agitation. **4.** Thermionic emission of electrons.

activity ratio The ratio of active to inactive records in a computer file.

ac transducer A transducer that either requires an ac supply voltage or delivers an ac output signal—even when operated from a dc supply.

ac transmission The use of an alternating voltage to transfer power from one point to another, usually from generators to a distribution center, and generally over a considerable distance.

actual ground The ground as “seen” by an antenna. The actual ground surface is not necessarily in the same physical location as the true ground surface (i.e., the earth itself). An actual ground can be an artificial ground plane, such as that provided in some antenna structures. Actual ground can also be modified by nearby rooftops, buildings, guy wiring, and utility wiring.

actual height The highest altitude where radio wave refraction actually occurs.

actual power Also called active or AVERAGE POWER. Symbol, P_{avg}. In a resistive circuit under sine-wave conditions, average power is the product of the rms voltage and the rms current. It is also equal to half the product of the maximum current and maximum voltage.

actuating device A device or component that operates electrical contacts to affect signal transmission.

actuating system **1.** An automatic or manually operated system that starts, modifies, or stops an operation. **2.** A system that supplies energy for ACTUATION.

actuating time Also called *actuation time.* The time interval between generation of a control signal, or the mechanical operation of a control device, and the resulting ACTUATION.

actuation **1.** The starting, modification, or termination of an operation or process. **2.** Activation of a mechanical or electromechanical switching device.

actuator An electromechanical device that uses electromagnetism to produce a longitudinal or rotary thrust for mechanical work. It is often the end (load) device of a servosystem.

ACU Abbreviation of *automatic calling unit.*

ac voltage A voltage, the average value of which is zero, that periodically changes its polarity. In one cycle, an ac voltage starts at zero, rises to a maximum positive value, returns to zero, rises to a maximum negative value, and finally returns to zero. The number of such cycles per second is termed the *ac frequency.*

ac voltmeter See AC METER.

acyclic machine Also called ACYCLIC GENERATOR. A dc generator in which voltage induced in the active wires of the armature is always of the same polarity.

A/D Abbreviation for ANALOG-TO-DIGITAL. See ANALOG-TO-DIGITAL CONVERSION.

Ada A microcomputer language designed primarily for use in multi-computer systems, where each small computer communicates with the others, providing some of the advantages of a larger computer.

Adam A communications code word sometimes used for phonetic verbalizing of the letter A. More commonly, ALPHA is used.

adapter **1.** A fitting used to change either the terminal scheme or the size of a jack, plug, or socket to that of another. **2.** A fitting used to provide a transition from one type or style of conductor to another (e.g., waveguide to coaxial line). **3.** An

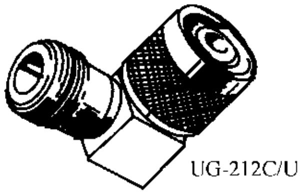

adapter

auxiliary system or unit used to extend the operation of another system (e.g., a citizens-band adapter for a broadcast receiver).

adaptive communication A method of communication that adjusts itself according to the particular requirements of a given time.

adaptive suspension vehicle Abbreviation, ASV. A specialized robot that moves on mechanical legs, rather than on wheels. It generally has six legs and resembles an insect. It is designed to move over extremely irregular or rocky terrain, and to carry a human passenger.

adaptivity The ability of a system to respond to its environment by changing its performance characteristics.

adc Abbreviation of ANALOG-TO-DIGITAL CONVERTER.

Adcock antenna A directional antenna system consisting of two vertical antennas, spaced in such a way that the whole array behaves like a loop antenna. Its members are connected and positioned so that it discriminates against horizontally polarized waves, and delivers output that is proportional to the vector difference of signal voltages induced in the two vertical arms.

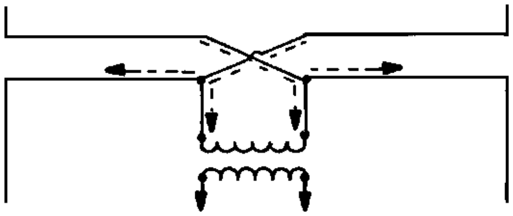

Adcock antenna

Adcock direction finder A radio direction-finding system based on the directivity of the ADCOCK ANTENNA.

Adcock radio range A radio range system with four ADCOCK ANTENNAS situated at the corners of a square, and a fifth antenna at the center of the square.

add-and-subtract relay A stepping relay that can be switched either uprange (add) or downrange (subtract).

addend In a calculation, any number to be added to another. Compare AUGEND.

addend register In a digital computer, the register that stores the addend.

adder **1.** In a digital computer, the device or circuit that performs binary addition. A HALF ADDER is a two-input circuit that can produce a sum output and a carry output, but it cannot accommodate a carry signal from another adder. A FULL ADDER can accommodate a carry input, as well as two binary signals to be added. Also see ANALOG ADDER. **2.** A circuit in a color TV receiver that amplifies the receiver primary matrix signal.

additive **1.** The character or characters added to a code to encipher it. **2.** In a calculation, an item that is to be added. **3.** An ingredient, usually in a small quantity, added to another material to improve the latter in quality or performance.

additive color A color formed by combining the rays from two or three primary-colored lights onto a single neutral surface. For example, by projecting a red and a green beam onto a neutral screen, a yellow additive color results.

additive primaries Primary colors that form other colors in a mixing of light (see ADDITIVE COLOR), but are not themselves formed by mixing other additive primaries. For example, red, green, and blue are the additive primaries used in color television. Through appropriate mixing, these colors can be used to generate an unlimited variety of other colors. Compare SUBTRACTIVE PRIMARIES, which form the color spectrum by mixing pigments rather than lights. In additive systems, each superimposed primary color increases the total light output from the reflecting (viewing) surface; in subtractive systems, each superimposed primary decreases the total reflectivity. Thus, equal combination of additive primaries produces gray or white, and equal combination of subtractive primaries produces gray or black.

addition record An extra data store created in a computer during processing.

address **1.** In computer operations, a usually numerical expression designating the location of material within the memory or the destination of such material. **2.** The accurately stated location of information within a computer; a data point within a grid, matrix, or table; a station within a network. **3.** In computer operations, to select the location of stored information.

address comparator A device that ensures that the address being read is correct.

address computation In digital computer operations, the technique of producing or modifying only the address part of an instruction.

address field In a computer, the part of the instruction that gives the address of a bit of data (or a word) in the memory.

address generation The programmed generation of numbers or symbols used to retrieve records from a randomly stored direct-access file.

address indirect An address that specifies a storage location that contains another address.

address memory The memory sections in a digital computer that contain each individual register.

address modification In computer operations, altering only the address portion of an instruction; if the command or instruction routine is then repeated, the computer will go to the new address.

address part In a digital computer instruction, the part of an expression that specifies the location. Also called ADDRESS FIELD.

address register In a computer, a register in which an address is stored.

add/subtract time In a computer, the time required to perform addition or subtraction, excluding the time required to get the quantities from storage and to enter the sum or difference into storage.

add time In computer operations, the time required to perform addition, excluding the time required to get the quantities from storage and to enter the sum into storage.

a/d converter A device that changes an analog quantity into a digital signal. See ANALOG-TO-DIGITAL CONVERSION.

ADF Abbreviation of AUTOMATIC DIRECTION FINDER.

ADI Abbreviation of ALTERNATE DIGIT INVERSION.

adiabatic damping In an accelerator (see ACCELERATOR, **1**), reduction of beam size as beam energy is increased.

adiabatic demagnetization A technique using a magnetic field to keep a substance at a low temperature, sometimes within a fraction of a degree of absolute zero.

adjacency A character-recognition condition in which the spacing reference lines of two characters printed consecutively in line are closer than specified.

adjacent- and alternate-channel selectivity The selectivity of a receiver or radio-frequency (RF) amplifier, with respect to adjacent-channel and alternate-channel signals. That is, the extent to which a desired signal is passed, and nearby unwanted signals are rejected.

adjacent audio channel See ADJACENT SOUND CHANNEL.

adjacent channel The channel (frequency band) immediately above or below the channel of interest.

adjacent-channel attenuation The reciprocal of the *selectivity ratio* of a radio receiver. The selectivity ratio is the ratio of the sensitivity of a receiver (tuned to a given channel) to its sensitivity in an adjacent channel or on a specified number of channels removed from the original.

adjacent-channel interference In television or radio reception, the interference from stations on adjacent channels. A common form arises from the picture signal in the next higher channel and the sound signal in the next lower channel.

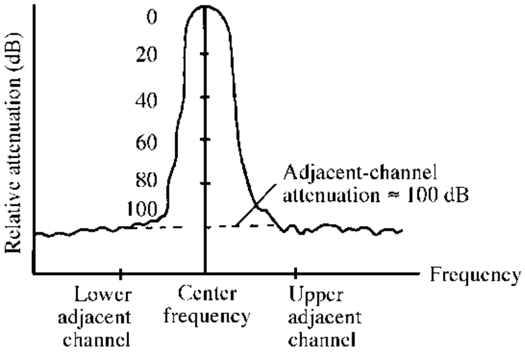

adjacent-channel attenuation

adjacent-channel selectivity The extent to which a receiver or tuned circuit can receive on one channel and reject signals from the nearest outlying channels.

adjacent sound channel In television, the radio-frequency (RF) channel containing the sound modulation of the next lower channel.

adjacent video carrier In television, the radio-frequency (RF) carrier containing the picture modulation of the next higher channel.

adjustable component Any circuit component whose main electrical value can be varied at will (e.g., a variable capacitor, inductor, resistor, or load).

adjustable instrument **1.** An instrument whose sensitivity, range, or response can be varied at will (e.g., multirange meter or wideband generator). **2.** An instrument that requires adjustment or manipulation to measure a quantity (e.g., bridge, potentiometer, or attenuator).

adjustable motor tuning An arrangement that allows the motor tuning of a receiver to be confined to a portion of the frequency spectrum.

adjustable resistor A wirewound resistor in which the resistance wire is partially exposed to allow varying the component's value.

adjustable voltage divider A wirewound resistor with terminals that slide on exposed resistance wire to produce various voltage values.

adjusted circuit A circuit in which leads that are normally connected to a circuit breaker are shunted so that current can be measured under short-circuit conditions without breaker tripping.

adjusted decibels Noise level (in decibels) above a reference noise level (designated arbitrarily as zero decibels) measured at any point in a system with a noise meter that has previously been adjusted for zero (at reference), according to specifications.

admittance Symbol, *Y*. Unit, siemens (formerly mho). The property denoting the comparative ease with which an alternating current flows through a circuit or device. Admittance is the reciprocal of impedance (*Z*): $Y = 1/Z$.

adp **1.** Abbreviation of AMMONIUM DIHYDROGEN PHOSPHATE, a piezoelectric compound used for sonar crystals. **2.** Abbreviation of AUTOMATIC DATA PROCESSING.

adsorption Adhesion of a thin layer of molecules of one substance to the surface of another without absorption. An example is adsorption of water to the surface of a dielectric. This term is often confused with ABSORPTION because the spellings of the two words are almost identical. Compare ABSORPTION.

adu Abbreviation of *automatic dialing unit.*

advanced-class license An amateur-radio license conveying all operating privileges, except for a few small bands that are allocated to extra-class licensees. The second-highest class of amateur license.

advance information Data published prior to the actual production or availability of a manufactured component, circuit, or system. Advance information is often only an approximate reflection of the expected characteristics of a device.

advance wire A resistance wire used in thermocouples and precision applications. It is an alloy of copper and nickel, which has high resistivity and a negligible temperature coefficient of resistance.

aeolight A glow lamp using a cold cathode and a mixture of inert gases. Because its illumination can be regulated with an applied signal voltage, it is sometimes used as a modulation indicator for motion-picture sound recording.

aerial See ANTENNA.

aerial cable A wire or cable run through the air, using support structures, such as towers or poles.

aerodiscone antenna A miniature discone antenna designed for use on aircraft.

aerodynamics The science dealing with forces exerted by air and other gases in motion—especially upon bodies (such as aircraft) moving through these gases.

aerogram See RADIOGRAM.

aeromagnetic Pertaining to terrestrial magnetism, as surveyed from a flying aircraft.

aeronautical advisory station A civil defense and advisory communications station in service for the use of private aircraft stations.

aeronautical broadcasting service The special service that broadcasts information regarding air navigation and meteorological data pertinent to aircraft operation.

aeronautical broadcast station A station of the aeronautical broadcasting service.

aeronautical fixed service A fixed radio service that transmits information regarding air navigation and flight safety.

aeronautical fixed service station A station that operates in the aeronautical fixed service.

aeronautical ground station A land station that provides communication between aircraft and ground stations.

aeronautical marker-beacon signal A distinctive signal that designates a small area above a beacon transmitting station for aircraft navigation.

aeronautical marker-beacon station A land station that transmits an aeronautical marker-beacon signal.

aeronautical mobile service A radio service consisting of communications between aircraft, and between aircraft and ground stations.

aeronautical radio-beacon station An aeronautical radio-navigation land station that transmits signals used by aircraft and other vehicles to determine their position.

aeronautical radionavigation services Services provided by stations transmitting signals used in the navigation of aircraft.

aeronautical radio service A service that encompasses aircraft-to-aircraft, aircraft-to-ground, and ground-to-aircraft communications important to the operation of aircraft.

aeronautical station A station on land, and occasionally aboard ship, operating in the aeronautical mobile service.

Aeronautical Telecommunication Agency The agency that administers the operation of stations in the aeronautical radio service.

aeronautical telecommunications Collectively, all of the electronic and nonelectronic communications used in the aeronautical service.

aeronautical utility land station A ground station in an airport control tower that provides communications having to do with the control of aircraft and other vehicles on the ground.

aeronautical utility mobile station At an airport, a mobile station that communicates with aeronautical utility land stations and with aircraft and other vehicles on the ground.

aerophare See RADIO BEACON.

aerospace **1.** The region encompassing the earth's atmosphere and extraterrestrial space. **2.** Pertaining to transport and travel in the earth's atmosphere and in outer space. This includes aircraft, orbiting space vessels, and interplanetary spacecraft.

AES Abbreviation for *Audio Engineering Society.*

AEW Abbreviation of *airborne (or aircraft) early warning.*

aF Abbreviation of ATTOFARAD.

AF Abbreviation of AUDIO FREQUENCY.

AFC **1.** Abbreviation of AUTOMATIC FREQUENCY CONTROL. **2.** Abbreviation of AUDIO-FREQUENCY CHOKE.

affirmative In voice communications, a word often used for "yes"—especially when interference is present or signals are weak.

AFIPS Acronym for *American Federation of Information Processing Societies.*

afpc Abbreviation of *automatic frequency/phase control.*

AFSK Abbreviation of AUDIO-FREQUENCY-SHIFT KEYING.

afterglow The tendency of the phosphor of a cathode-ray-tube screen to glow for a certain time after the cathode-ray beam has passed. Also see PERSISTENCE.

afterpulse An extraneous pulse in a multiplier phototube (photomultiplier), induced by a preceding pulse.

AF transformer See AUDIO-FREQUENCY TRANSFORMER.

a/g Abbreviation of AIR-TO-GROUND.

AGC Abbreviation of AUTOMATIC GAIN CONTROL.

AGE Abbreviation of AEROSPACE GROUND EQUIPMENT.

agent An active force, condition, mechanism, or substance that produces or sustains an effect. Thus, a sudden voltage rise is a triggering agent in certain bistable circuits; arsenic is a doping agent in semiconductor processing; the slow cooling of a heated metal to improve ductility is an ANNEALING AGENT.

aging **1.** An initial run of a component or circuit over a certain period of time shortly after manufacture to stabilize its characteristics and performance. **2.** The changing of electrical characteristics or of chemical properties over a protracted period of time.

agonic line An imaginary line connecting points on the earth's surface at which a magnetic needle shows zero declination (i.e., points to true geographic north).

AGREE Acronym for *Advisory Group on Reliability of Electronics Equipment.*

Ah Abbreviation of AMPERE-HOUR. Depending on the standard used, the abbreviation can be amp-hr, a-h, a-hr, or A-h.

aH Abbreviation of ATTOHENRY.

aided tracking In radar and fire control, a system in which manual correction of target tracking error automatically corrects the rate of movement of the tracking mechanism.

AIEE Abbreviation for *American Institute of Electrical Engineers*, now consolidated with the IRE, forming the IEEE.

AIP Abbreviation for *American Institute of Physics.*

air The mixture of gases that constitutes the earth's atmosphere and figures prominently in the manufacture and operation of numerous electronic devices. By volume, air contains about 21 percent oxygen, 78 percent nitrogen, and lesser amounts of argon, carbon dioxide, helium, hydrogen, krypton, neon, and xenon. It also contains varying amounts of water vapor, and in smoggy areas, carbon monoxide and the oxides of sulfur and nitrogen.

airborne intercept radar A type of short-range radar used aboard fighter and interceptor aircraft for tracking their targets.

airborne long-range input Equipment aboard aircraft, for the purpose of facilitating the use of long-range missiles.

airborne noise See ACOUSTIC NOISE.

airborne radar platform Surveillance and altitude-finding radar used aboard aircraft.

air capacitor A capacitor in which air is the dielectric between two sets of conductive plates. Also called *air-dielectric capacitor.*

aircarrier aircraft station On an aircraft, a radio station that is involved in carrying people for hire or in transporting cargo.

air cell A primary electrochemical cell in which the positive electrode is depolarized by reduced oxygen in the air.

air cleaner See DUST PRECIPITATOR.

air column The open space inside an acoustic chamber, pipe, or horn.

air-cooled component A component, such as a power transistor, that is cooled by circulating air, compared with one cooled by a circulating liquid, such as water or oil.

air-cooled transistor A transistor (particularly a power transistor) from which the heat of operation is drawn away, through radiation and convection, into the surrounding air. The transistor is usually mounted on a heatsink or fitted with fins.

air-cooled tube An electron tube from which heat is drawn away, mainly via convection, into the surrounding air. A device called a *chimney* can be placed around the tube, through which air is blown by a fan. Cool air enters through the bottom of the assembly, and hot air escapes from the top.

air-core inductor A coil of wire wound around a hollow cylindrical form or in a loop, designed to introduce inductive reactance into a circuit or system. In practice, the maximum attainable inductance is approximately 1 mH. This type of inductor is used in some wireless transmitters, receivers, and antenna networks. The component can be designed for high current-carrying capacity by using heavy-gauge wire and a large winding radius. The magnetic lines of flux extend considerably beyond the interior of the coil, especially along the winding axis. This increases the likelihood of mutual inductance between the coil and surrounding electrical components, devices, or circuits.

air-core transformer A transformer without a ferromagnetic core, so called because air is the only material at the center of (and immediately surrounding) the transformer coils.

aircraft bonding The practice of solidly connecting, for electrical purposes, the metal parts of an aircraft, including the engine.

aircraft flutter Rapid, repetitive fading and intensifying of a received radio or television signal, resulting from reflections of the signal by passing aircraft.

aircraft station A nonautomatic radio communications station installed on an aircraft.

air-dielectric coax A special type of COAXIAL CABLE designed to have minimum loss. The space between inner and outer conductors is mostly empty (i.e., air-filled). Some such cables are sealed and filled with an inert gas. The inner conductor is held away from the inner wall of the outer conductor by beads, washers, or a spiral-wound filament of high-grade dielectric material, such as polyethylene.

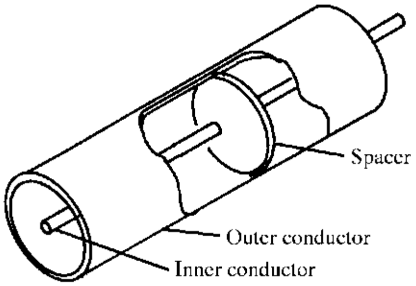

air-dielectric coax

air environment Pertaining to communications equipment aboard aircraft.

airflow The path or movement of air in, through, or around an electronic device or piece of equipment—especially pertaining to an AIR-COOLED COMPONENT.

air gap **1.** A narrow space between two parts of a magnetic circuit (e.g., the gap in the core of a filter choke). Often, this gap is filled with a nonmagnetic material, such as plastic, for mechanical support. **2.** The space between two or more magnetically coupled or electrostatically coupled components. **3.** A device that gets its name from the narrow gap between two small metal balls, needle points, or blunt rod tips therein. When an applied voltage is sufficiently high, a spark discharges across the gap.

air/ground control radio station A station for aeronautical telecommunications related to the operation and control of local aircraft.

air-insulated line **1.** An open-wire feeder or transmission line. Typically, the line consists of two parallel wires held apart by separators (bars or rods of high-grade dielectric material) situated at wide intervals. **2.** AIR-DIELECTRIC COAX.

air-moving device A mechanical device, such as a specially designed fan or blower, used to facilitate air cooling of electronic components.

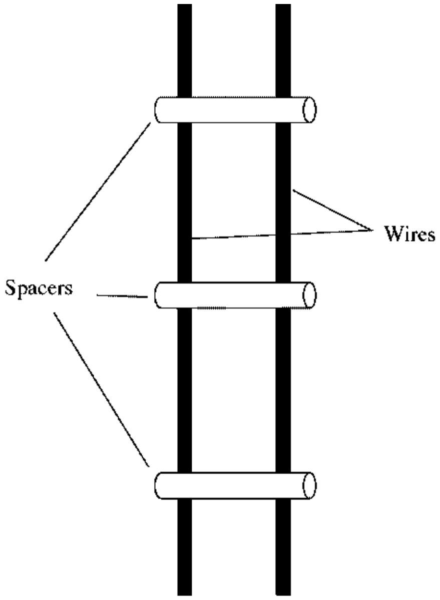

air-insulated line, 1

airport beacon A radio or light beacon that marks the location of an airport.

airport control station A station that provides communications between an airport control tower and aircraft in the vicinity.

airport surveillance radar An air-traffic-control radar that scans the airspace within about 60 miles (approximately 100 kilometers) of an airport, and displays in the control tower the location of all aircraft below a certain altitude and all obstructions in the vicinity.

air-position indicator An airborne computer system that, using airspeed, aircraft heading, and elapsed time, furnishes a continuous indication of the position of the aircraft. The indication is affected by high-altitude winds. Compare GROUND-POSITION INDICATOR.

air-to-air communication Radio transmission from one aircraft to another in flight. Compare AIR-TO-GROUND COMMUNICATION and GROUND-TO-AIR COMMUNICATION.

air-to-ground communication Radio transmission from an aircraft in flight to a station located on the ground. Compare AIR-TO-AIR COMMUNICATION and GROUND-TO-AIR COMMUNICATION.

air-to-ground radio frequency The carrier frequency, or band of such frequencies, allocated for transmissions from an aircraft to a ground station.

airwaves **1.** Radio waves. The term is slang, but is widely used. It probably came from the public's mistaken notion that radio signals are propagated by the air. **2.** Skywaves.

Al Symbol for ALUMINUM.

alabamine See ASTATINE.

alacratized switch A mercury switch in which the tendency of the mercury to stick to the parts has been reduced.

alarm **1.** An electronic security system. **2.** A silent and/or audible alert signal transmitted by an electronic security system when an intrusion occurs. **3.** A silent and/or audible signal that informs personnel of the occurrence of an equipment malfunction.

alarm circuit A circuit that alerts personnel to a system malfunction, a detected condition, or an intruder.

alarm condition **1.** An intrusion or equipment malfunction that triggers an alarm circuit. **2.** The operation of an alarm circuit that occurs in response to an intrusion or equipment malfunction.

alarm hold A device that keeps an alarm sounding once it has been actuated.

alarm output The signal sent from an alarm circuit to a siren, buzzer, computer, or other external device to alert personnel to an ALARM CONDITION.

alarm relay A relay that is actuated by an alarm device.

A-law A form of companding law frequently used in European electronics (the mu-law is more often used in North America). A nonlinear transfer characteristic in companding circuits. It can be continuous, or can be a piecewise linear approximation of a continuous function.

A-law companded Companding by means of an 8-bit binary code following the A-LAW, a specific companding function.

albedo For an unpolished surface, the ratio of reflected light to incident light. It can vary from 0.0 to 1.0, or from 0 to 100 percent.

albedograph An instrument for measuring the albedo of planets.

ALC Abbreviation of AUTOMATIC LEVEL CONTROL.

alerting device An audible alarm that includes a self-contained solid-state audio oscillator. Powered from the ac line or a battery, the device produces a raucous noise when actuated.

Alexanderson antenna A very-low-frequency (VLF) and low-frequency (LF) vertically polarized antenna, designed to minimize ground losses in structures of manageable height. It usually consists of several wires, each quarter-wave resonant with a loading coil, and all connected together at the apex of a tower. The antenna is fed between the ground and the base of one of the wires.

Alford antenna A loop antenna, in a square configuration, with the corners bent toward the center to lower the impedance at the current nodes.

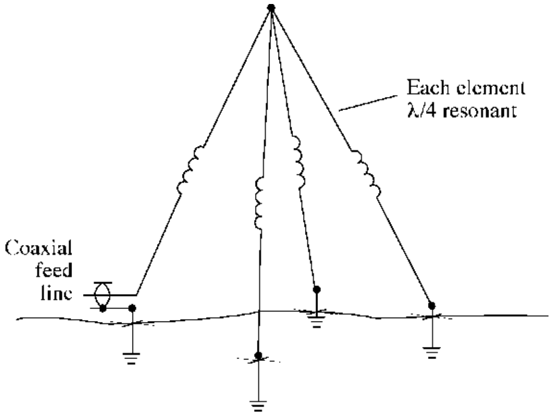

Alexanderson antenna

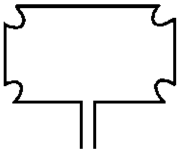

Alford antenna

algebraic adder In computer operations, an adder that provides the algebraic sum, rather than the arithmetic sum, of the entered quantities.

algebraic operation A form of electronic calculator operation, in which the keystrokes proceed in an intuitive sequence, following the way in which the calculation is written down. Compare REVERSE POLISH NOTATION.

algebraic sum The sum of two or more quantities with consideration of their signs. Compare ARITHMETIC SUM.

algorithm A step-by-step procedure for solving a problem, (e.g., the procedure for finding the square root of a number). It can be expressed in a line-by-line instruction set or as a flowchart.

algorithmic language A computer language used to describe a numeral or algebraic process.

alias A label that is an alternate term for items of the same type; a label and several aliases can identify the same data element in a computer program.

aliasing **1.** In analog-to-digital (A/D) conversion, a false output signal that results from a sampling rate that is too slow. Ideally, the sampling rate is at least twice the highest input signal frequency. **2.** Sawtooth-like irregularities, also called *jaggies*, which are sometimes introduced into a bit-mapped computer image when it is changed in size.

aliasing noise A form of signal distortion caused by a signal with an excessive bandwidth.

align **1.** To adjust (i.e., to preset) the circuits of an electronic system, such as a receiver, transmitter, or test instrument, for predetermined response. **2.** To arrange elements in a certain precise orientation and spacing, relative to each other, as in a Yagi antenna. **3.** To orient antennas so that they are in line of sight, with respect to each other.

alignment The process of ensuring that equipment, components, or systems are adjusted, both physically and electronically, for the most efficient possible performance.

alignment chart A line chart for the simple solution of electronic problems. It is so called because its use involves aligning numerical values on various scales, the lines intersecting at the solution on another scale. Also called *nomograph.*

alignment pin A pin or protruding key, usually in the base of a removable or plug-in component, to ensure that the latter will be inserted correctly into a circuit. Often, the pin mates with a keyway, notch, or slot.

alignment tool A specialized screwdriver or wrench (usually nonmagnetic) used to adjust padder or trimmer capacitors or inductor cores.

alive See LIVE.

alkali See BASE, **2.**

alkali metals Metals whose hydroxides are bases (alkalis). The group includes cesium, francium, lithium, potassium, rubidium, and sodium.

alkaline battery **1.** A battery composed of alkaline cells and characterized by a relatively flat discharge curve under load.

alkaline cell A common non-rechargeable electrochemical cell that employs granular zinc for the negative electrode, potassium hydroxide as the electrolyte, and a device called a *polarizer* as the positive electrode. Produces approximately 1.5 volts under no-load conditions. The geometry of construction is similar to that of the zinc–carbon cell, but it can deliver current effectively at lower temperatures. Cells of this type have shelf lives longer than zinc–carbon cells; they also have greater energy-storage capacity per unit volume, but they are more expensive than zinc–carbon cells. They are used in calculators, transistor radios, and cassette tape and compact-disc players. Compare ZINC–CARBON CELL.

alkaline-earth metals The elemental metals barium, calcium, strontium, and sometimes beryllium, magnesium, and radium, some of which are used in vacuum tubes.

alkaline earths Substances that are oxides of the alkaline-earth metals. Some of these materials are used in vacuum tubes.

all-diffused A type of INTEGRATED CIRCUIT in which both active and passive elements have been fabricated by diffusion and related processes.

Allen screw A screw fitted with a six-sided (hexagonal) hole.

Allen wrench A tool used to tighten or loosen an Allen screw. It is a hexagonal rod and is available in various sizes.

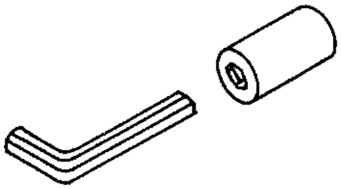

Allen screw and wrench

alligator clip A spring-loaded clip with jagged teeth, designed to be used for temporary electrical connections.

allocate **1.** To assign (especially through legislation) operating frequencies or other facilities or conditions needed for scientific or technical activity; see, for example, ALLOCATION OF FREQUENCIES. **2.** In computer practice, to assign locations in the memory or registers for routines and subroutines.

allocated channel A frequency channel assigned to an individual or group.

allocated-use circuit **1.** A circuit in which one or more channels have been authorized for the exclusive use of one or more services. **2.** A communications link assigned to users needing it.

allocation of frequencies See RADIO SPECTRUM.

allocator A telephone system distributor associated with the finder control group relay assembly. It reserves an inactive line-finder for another call.

allophone A variation in the sound of a phoneme, depending on what comes before and/or after the phoneme in the course of speech. Important in speech recognition and synthesis. There are 128 different phoneme variations in the English language. See PHONEME.

alloter relay A telephone system line-finder relay that reserves an inactive line-finder for the next incoming call from the line.

allotropic Pertaining to a substance existing in two forms.

alloy A metal that is a mixture of several other metals (e.g., brass from copper and zinc), or of a metal and a nonmetal.

alloy deposition In semiconductor manufacture, depositing an alloy on a substrate.

alloy-diffused transistor A transistor in which the base is diffused and the emitter is alloyed. The collector is provided by the semiconductor substrate into which alloying and diffusion are affected. Compare ALLOY TRANSISTOR and DIFFUSE TRANSISTOR.

alloy diode A junction-type semiconductor diode in which a suitable substance (such as p-type) is alloyed into a chip of the opposite type (such as n-type) to form the junction. Also called *alloy-junction diode.*

alloy junction In a semiconductor device, a positive/negative (pn) junction formed by alloying a suitable material (such as indium) with the semiconductor (silicon or germanium).

alloy transistor A transistor whose junctions are created by alloying. Also see ALLOY JUNCTION.

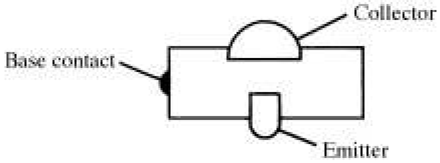

alloy transistor

all-pass filter Also called *all-pass network.* A filter that (ideally) introduces a desired phase shift or time delay, but has zero attenuation at all frequencies.

all-relay central office In telephone service, an automatic central-office switchboard that uses relay circuits to make line interconnections.

all-wave Pertaining to a wide operating-frequency range. Few systems are literally all-wave. For example, a so-called "all-wave radio receiver" might cover 500 kHz to 30 MHz only.

all-wave antenna An antenna that can be operated over a wide frequency range with reasonable efficiency and preferably without needing readjustment. Examples are the DISCONE ANTENNA and the LOG-PERIODIC ANTENNA.

all-wave generator A signal generator that will supply output over a wide range of frequencies.

all-wave receiver A radio receiver that can be tuned over a very wide range of frequencies, such as 10 kHz to 70 MHz.

allyl plastics Plastics, sometimes used as dielectrics or for other purposes in electronics, based on resins made by polymerization of monomers (such as diallyl phthalate) that contain allyl groups.

alnico Coined from the words aluminum, nickel, and cobalt. An alloy used in strong permanent magnets, it contains the constituents noted plus (sometimes) copper or titanium.

alpha **1.** Symbol, α. The current gain of a common-base-connected bipolar transistor. It is the ratio of the differential of collector current to the differential of emitter current; $\alpha = dI_C/dI_E$. For a junction transistor, alpha is always less than unity, but very close to it. **2.** In voice communications, the phonetic representation of the letter A.

alphabet The set of all characters in a natural language.

alphabetic coding In computer practice, an abbreviation system for coding information to be fed into the computer. The coding contains letters, words, and numbers.

alphabetic-numeric Also called *alphabetical-numerical* and *alphanumeric.* In computer operations, pertaining to letters of the alphabet and special characters, and to numerical digits.

alpha cutoff frequency Also called *alpha cutoff.* In a bipolar transistor circuit, the frequency at which the alpha (current gain) becomes 0.707 (70.7 percent) of its value at 1 kHz. A bipolar transistor can have considerable gain at its alpha cutoff. This specification denotes how rapidly a transistor loses gain as the frequency increases, an important consideration in the design of radio-frequency (RF) amplifiers. See ALPHA. Compare GAIN BANDWIDTH PRODUCT.

alpha decay The decay of a substance in which the nuclei of the atoms emit alpha particles, resulting in a change of the atomic number and atomic weight of the substance over a period of time.

alphanumeric See ALPHABETIC-NUMERIC.

alphanumeric code In computer operations or in communications, a code composed of, or using, both letters and numbers.

alphanumeric readout A type of digital readout that displays both letters and numerals.

alpha particle A nuclear particle bearing a positive charge. Consisting of two protons and two neutrons, it is given off by certain radioactive substances. Compare BETA RAYS and GAMMA RAYS.

alpha system An alphabetic code-signaling system.

alphatron An ionizing device in which the radiation source is an emitter of alpha particles.

alteration An inclusive-OR operation.

alternate channel In communications, a channel situated two channels higher or lower than a given channel. Compare ADJACENT CHANNEL.

alternate-channel interference Interference caused by a transmitter operating in the channel beyond an adjacent channel. Compare ADJACENT-CHANNEL INTERFERENCE.

alternate digit inversion In multiplex equipment, a method of switching the binary signals to the opposite state, in accordance with A-law companding.

alternate frequency A frequency allocated as an alternative to a main assigned frequency and used under certain specified conditions.

alternate-mark inversion signal A signal that conveys bits in which the successive signals are of opposite polarity (positive, then negative, then positive, etc.). They are equal in absolute value amplitude.

alternate mode The technique of displaying several signals on an oscilloscope screen by rapidly switching the signals in sequence at the end of each sweep.

alternate routing A secondary, or backup, communications path, used when primary (normal) routing is impossible.

alternating-charge characteristic In a nonlinear capacitor, the relationship between the instantaneous charge and the instantaneous value of an alternating voltage.

alternating current Abbreviation, ac. A current that periodically reverses its direction of flow. In one cycle, an alternation starts at zero, rises to a maximum positive level, returns to zero, rises to a maximum negative level, and again returns to zero. The number of such cycles completed per second is termed the *ac frequency.* Also see CURRENT.

alternating-current continuous wave An amplitude-modulated signal resulting from the operation of an oscillator or RF amplifier with raw ac voltage.

alternating current/direct current See AC/DC.

alternating-current erasing head See AC ERASING HEAD.

alternating-current pulse A short-duration ac wave.

alternating-current transmission **1.** The propagation of alternating currents along a length of conductor—especially for power-transfer purposes. **2.** A means of picture transmission in which a given signal strength produces a constant value of brightness for a very short time.

alternating voltage Also called *alternating-current voltage.* See AC VOLTAGE.

alternation In ac practice, a half cycle. In a complete cycle, there are two alternations, one in the positive direction and one in the negative direction.

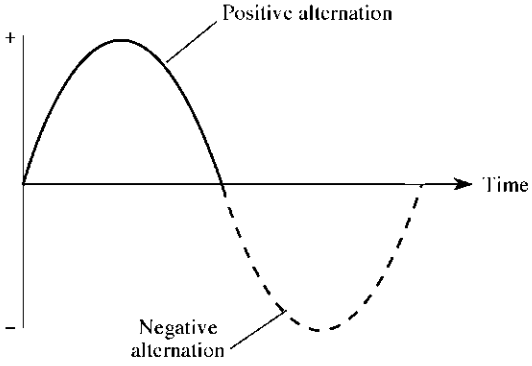

alternation

alternative denial A NOT-AND operation.

alternator Any mechanically driven machine for generating ac power. Sometimes specifically one having a permanent-magnet rotor, such as a *magneto.*

altimeter station An airborne transmitter whose signals are used to determine the altitude of aircraft.

altitude **1.** The vertical distance of an object above sea level. **2.** The vertical distance of an object above the earth's surface. **3.** The angle, measured in degrees, with respect to the horizon, at which a highly directional antenna is pointed.

altitude delay In a plan-position-indicating type of radar, the sync delay introduced between transmission of the pulse and start of the trace on the indicator screen to eliminate the altitude circle in the display.

ALU Abbreviation of ARITHMETIC AND LOGIC UNIT.

alumel An alloy used in the construction of one type of THERMOCOUPLE. It is composed of nickel (three parts) and aluminum (one part).

alumina An aluminum-oxide ceramic used in electron tube insulators and as a substrate in the fabrication of thin-film circuits.

aluminum Symbol, Al. An elemental metal. Atomic number, 13. Atomic weight, 26.98. Aluminum is widely used in electronics, familiar instances being chassis, wire, shields, semiconductor doping, and electrolytic-capacitor plates.

aluminum antimonide Formula, AlSb. A crystalline compound useful as a semiconductor dopant.

aluminized screen A television picture-tube screen with a thin layer of aluminum deposited on its back to brighten the image and reduce ion-spot formation.

Am Symbol for AMERICIUM.

A/m Abbreviation of *ampere per meter:* the SI unit of magnetic field strength.

AM **1.** Abbreviation of *amplitude modulator.* **2.** Abbreviation of AMPLITUDE MODULATION.

amalgam An alloy of a metal and mercury. Loosely, any combination of metals.

amateur **1.** A nonprofessional, usually noncommercial devotee of any technology (i.e., a hobbyist). **2.** A licensed radio operator legally authorized to operate a station in the AMATEUR SERVICE.

amateur band Any band of radio frequencies assigned for noncommercial use by licensed radio amateurs (see AMATEUR, **2**). In the United States, numerous such bands are above 1.8 MHz (160 meters). Also see AMATEUR SERVICE and AMATEUR STATION.

amateur call letters Call letters assigned by a government licensing authority—especially to amateur stations. Call-letter combinations consist of a letter prefix denoting the country in which the station is situated, plus a number designating the location within the country, and two or more letters identifying the particular station. For example: W6ABC: W (or K) = United States, 6 = California, and ABC = identification of individual licensee (issued alphabetically, except under special circumstances).

amateur callsign See AMATEUR CALL LETTERS.

amateur extra-class license The highest class of amateur-radio operator license in the United States. It conveys all operating privileges.

amateur radio **1.** A general term, referring to the practice of operation, experimentation, and other work in and related to the amateur service. **2.** The hardware that comprises an amateur radio station. **3.** A radio receiver, transmitter, or transceiver that is specifically designed for operation in the amateur bands.

amateur radio operator Also called *radio ham* or *ham radio operator.* An individual licensed to transmit radio signals in the amateur service.

amateur service A two-way radio service, existing purely for hobby purposes (i.e., without pecuniary interest).

amateur station A radio station licensed in the AMATEUR SERVICE.

amauroscope An electronic aid to the blind, in which photocells in a pair of goggles receive light images. Electric pulses proportional to the light are impressed upon the visual receptors of the brain through electrodes in contact with nerves above each eye.

amber A yellow or brown fossil resin that is historically important in electronics. It is the first material reported to be capable of electrification by rubbing (Thales, 600 BC). Also, the words *electricity*, *electron*, and *electronics* are derived from the Greek name for amber, *elektron.*

ambience The acoustic characteristic of a room, in terms of the total amount of sound reaching a listener from all directions.

ambient An adjective meaning "surrounding." Often used as a noun in place of the adjective-noun combination (thus, "10 degrees above ambient," instead of "10 degrees above ambient temperature").

ambient humidity The amount of moisture in the air at the time of measurement or operations in which dampness must be accounted for.

ambient level The amplitude of all interference (acoustic noise, electrical noise, illumination, etc.) emitted from sources other than that of a signal of interest.

ambient light Also called *ambient illumination.* Room light or outdoor light incident to a location at the time of measurement or operations.

ambient-light filter In a television receiver, a filter mounted in front of a picture-tube screen to minimize the amount of ambient light reaching the screen.

ambient noise **1.** In electrical measurements and operation, background electrical noise. **2.** In acoustical measurements and operations, audible background noise.

ambient pressure Surrounding atmospheric pressure.

ambient temperature The temperature surrounding apparatus and equipment (e.g., room temperature).

ambient-temperature range **1.** The range over which ambient temperature varies at a given location. **2.** The range of ambient temperature that will cause no malfunction of, or damage to, a circuit or device.

ambiguity **1.** Any unclear, illogical, or incorrect indication or result. **2.** The seeking of a false null by a servo. **3.** In digital computer operations, an error resulting from improper design of logic.

ambiguous count In digital counters, a clearly incorrect count. See ACCIDENTAL TRIGGERING.

ambisonic reproduction A close approximation of the actual directional characteristics of a sound in a given environment. The reproduced sound almost exactly duplicates the sound in the actual environment in which it was recorded.

American Morse code (Samuel F. B. Morse, 1791–1872). Also called *Railroad Morse.* A telegraph code, at one time used on wire telegraph lines in the United States. It differs from the *Continental code*, also called the *International Morse Code*, which is used in radiotelegraphy. Compare CONTINENTAL CODE.

American National Standards Institute Acronym, ANSI. An industrial group in the United States that encourages companies to manufacture devices and equipment in accordance with certain standards. The objective is to minimize hardware incompatibility problems.

American Radio Relay League A worldwide organization of amateur radio operators, headquartered in Newington, Connecticut. The official publications are the monthly magazines, *QST* and *QEX.* They also publish numerous books and other educational materials.

American Standards Association Abbreviation, ASA. At one time, the name of the national association in the U.S. devoted to the formation and dissemination of voluntary standards of dimensions, performance, terminology, etc. See ANSI.

American wire gauge Abbreviation, AWG. Also called *Brown and Sharpe gauge* or *B & S gauge.* The standard American method of designating wire sizes. Wire is listed according to gauge number from 0000 (460 mils diameter) to 40 (3.145 mils diameter).

americium Symbol, Am. A radioactive elemental metal first produced artificially in the 1940s. Atomic number, 95. Atomic weight, 243.

AM/FM receiver A radio set that can receive either amplitude-modulated or frequency-modulated signals. Usually, a band switch incorporates the demodulation-selection circuitry so that as the frequency range is changed, the appropriate detector is accessed.

AM/FM transmitter A radio transmitter whose output signal can be frequency- or amplitude-modulated by a panel selector switch.

AM/FM tuner A compact radio receiver unit that can handle either amplitude- or frequency-modulated signals, and delivers low-amplitude output to a high-fidelity audio power amplifier. Compare AM TUNER and FM TUNER.

Character	*Symbol*	*Character*	*Symbol*
A	. —	U	. . —
B	— . . .	V	. . . —
C	. . .	W	. — —
D	— . .	X	. — . .
E	.	Y	. . . .
F	. — .	Z	. . . .
G	— — .	1	. — — .
H	. . . .	2	. . — . .
I	. .	3	. . . — .
J	— . — .	4	. . . . —
K	— . —	5	— — —
L	——	6	. . . . . .
M	— —	7	— — . .
N	— .	8	— . . . .
O	. .	9	— . . —
P	. . . . .	0	——
Q	. . — .	period	. . — — . .
R	. . .	comma	. — . —
S	. . .	question mark	— . . — .
T	—		

American Morse Code

American Wire Gauge (AWG) Diameters

AWG	Millimeters	Inches	AWG	Millimeters	Inches
1	7.35	0.289	21	0.723	0.0285
2	6.54	0.257	22	0.644	0.0254
3	5.83	0.230	23	0.573	0.0226
4	5.19	0.204	24	0.511	0.0201
5	4.62	0.182	25	0.455	0.0179
6	4.12	0.163	26	0.405	0.0159
7	3.67	0.144	27	0.361	0.0142
8	3.26	0.128	28	0.321	0.0126
9	2.91	0.115	29	0.286	0.0113
10	2.59	0.102	30	0.255	0.0100
11	2.31	0.0909	31	0.227	0.00894
12	2.05	0.0807	32	0.202	0.00795
13	1.83	0.0720	33	0.180	0.00709
14	1.63	0.0642	34	0.160	0.00630
15	1.45	0.0571	35	0.143	0.00563
16	1.29	0.0508	36	0.127	0.00500
17	1.15	0.0453	37	0.113	0.00445
18	1.02	0.0402	38	0.101	0.00398
19	0.912	0.0359	39	0.090	0.00354
20	0.812	0.0320	40	0.080	0.00315

AMI See ALTERNATE-MARK INVERSION SIGNAL.

A-minus Also, A-. The negative terminal of an A battery, or pertaining to the part of a circuit connected to that terminal.

ammeter An instrument used to measure the amount of current (in amperes) flowing in a circuit.

ammeter shunt A resistor connected in parallel with an ammeter to increase its current range. Also see AYRTON-MATHER GALVANOMETER SHUNT.

ammeter-voltmeter method The determination of resistance or power values from the measurement of voltage (E) and current (I). For resistance, $R = E/I$; for power, $P = EI$.

ammonium chloride Formula, NH_4Cl. The electrolyte in the carbon-zinc type of primary cell. Also called SAL AMMONIAC.

AMNL Abbreviation of AMPLITUDE-MODULATION NOISE LEVEL.

amortisseur winding **1.** A winding that acts against pulsation of the magnetic field in an electric motor. **2.** A winding that acts to prevent oscillation in a synchronous motor.

amorphous substance A noncrystalline material.

amp **1.** Slang for AMPERE. **2.** Slang for AMPLIFIER—especially in audio high-fidelity applications.

ampacity Current-carrying capacity expressed in amperes.

amperage The strength of an electric current (i.e., the number of amperes).

ampere (Andre Marie Ampere, 1775-1836). Abbreviations, A (preferred), a, amp. The SI base unit of current intensity (I). The ampere is the constant current that, if maintained in two straight parallel conductors of infinite length and of negligible circular cross section and placed 1 meter apart in a vacuum, would produce between the conductors a force of 2×10^{-7} newton per meter. One ampere flows through a 1-ohm resistance when a potential of 1 volt is applied; thus $I = E/R$. Also see MICROAMPERE, MILLIAMPERE, NANOAMPERE, and PICOAMPERE.

ampere balance A device consisting of two conductors in which the force between them (caused by current) is balanced against the gravitational force exerted on an object in the gravitational field of the earth. Used for the precise determination of current of large dimension, or of the size of the ampere.

ampere-hour Abbreviations: Ah, amp-hr. The quantity of electricity that passes through a circuit in one hour when the rate of flow is one ampere. Also see BATTERY CAPACITY.

ampere-hour meter An instrument for measuring ampere-hours. It contains a small motor driven by the current being measured and which moves a point on an ampere-hour scale. The motor speed is proportional to the current. The position of the pointer is proportional to current and elapsed time.

Ampere's law Current flowing in a wire generates a magnetic flux that encircles the wire in the clockwise direction when the current is moving away from the observer.

ampere-turn Symbol, NI. A unit of magnetomotive force equal to 1 ampere flowing in a single-turn coil. The ampere-turns value for any coil is obtained by multiplying the current (in amperes) by the number of turns in the coil.

Amperian whirl The stream of electrons in a single-turn, current-conducting wire loop acting as an elementary electromagnet.

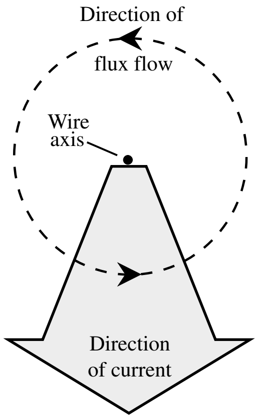

Ampere's Law

amp-hr One style of abbreviating AMPERE-HOUR. Also, Ah.

amplidyne A dynamo-like rotating dc machine that can act as a power amplifier because the response of the output voltage to changes in field excitation is quite rapid. Used in servo systems.

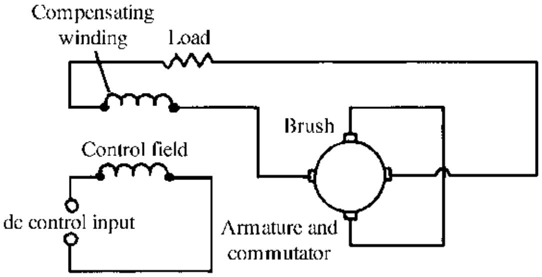

amplidyne

amplification **1.** The process of increasing the magnitude of a signal. This entails an input signal controlling a local power supply to produce a larger output signal. Depending on the kind of input and output signals, amplification can be categorized as CURRENT, VOLTAGE, POWER, or some combination of these. **2.** The qualitative signal increase resulting from the process in **1. 3.** The quantitative signal increase (resulting from the process in **1**), expressed as a factor (such as 100) or in terms of *decibels* (dB). See AMPLIFICATION FACTOR and DECIBEL.

amplification factor **1.** The ratio of the output voltage, current, or power to the input voltage, current, or power of an AMPLIFIER circuit. For voltage or current, this ratio has meaning only when the input and output impedances are identical. **2.** The number of decibels by which an AMPLIFIER circuit increases the amplitude of a signal. For voltage or current, this figure has meaning only when the input and output impedances are identical. See DECIBEL. **3.** The ALPHA or BETA of a bipolar transistor. **4.** In the operation of an electron tube, the ratio of the derivative (instantaneous rate of change) of the plate voltage to the derivative of the grid voltage, for zero change in plate current.

amplified ALC Abbreviation, AALC. An automatic-level-control (ALC) system that uses the amplification of the fed-back control signal. It is used in RF power amplifiers, particularly single-sideband (SSB) linear amplifiers, to prevent overmodulation and nonlinearity.

amplified back bias A declining voltage developed across a fast-time-constant circuit in an amplifier stage and fed back into a preceding stage.

amplifier Any device that increases the magnitude of an applied signal. It receives an input signal and delivers a larger output signal that, in addition to its increased amplitude, is a replica of the input signal. Also see CURRENT AMPLIFIER, POWER AMPLIFIER, and VOLTAGE AMPLIFIER.

amplifier diode Any semiconductor that can provide amplification in a suitable circuit or microwave system. See DIODE AMPLIFIER.

amplifier distortion A change in the waveform of a signal, arising within an amplifier that is operated in compliance with specified conditions.

amplifier input **1.** The terminals and section of an amplifier that receive the signal to be amplified. **2.** The signal to be amplified.

amplifier noise Collectively, all extraneous signals present in the output of an amplifier when no working signal is applied to the amplifier input terminals.

amplifier nonlinearity A condition in which the amplifier output signal does not exhibit a linear relationship to the corresponding input signal. Some amplifiers are designed to operate in a linear manner at all times, but many amplifier types need not function in this manner to be effective. Also see AMPLIFIER DISTORTION and LINEAR AMPLIFIER.

amplifier output **1.** The terminals and section of an amplifier that deliver the amplified signal for external use. **2.** The amplified signal.

amplifier power The power level of the output signal delivered by an amplifier (also called OUTPUT POWER), or the extent to which the amplifier increases the power of the input signal (also called POWER AMPLIFICATION).

amplifier response The performance of an amplifier throughout a specified frequency band. Factors usually included are gain, distortion, amplitude versus frequency, and power output.

amplify To perform the functions of amplification (see AMPLIFICATION, **1**).

amplifying delay line A delay line that causes amplification of signals in a circuit intended for pulse compression.

amplistat A self-saturating magnetic amplifier.

amplitron A backward-wave amplifier used in microwave circuits.

amplitude The extent to which an alternating or pulsating current or voltage swings, positively and negatively, from zero or from a mean value.

amplitude-controlled rectifier A thyratron- or thyristor-based rectifier circuit.

amplitude density distribution A mathematical function giving the probability that, at a given instant in time, a fluctuating voltage has a certain value.

amplitude distortion In an amplifier or network, the condition in which the output-signal amplitude exhibits a nonlinear relationship to the input-signal amplitude.

amplitude error **1.** The error in measuring the amplitude of a signal, normally expressed as a percentage of signal amplitude or as a percentage of full scale. **2.** The frequency at which the output amplitude of a signal is in error by 1% with amplitude at 10% of full scale.

amplitude factor For an ac wave, the ratio of the peak value to the rms value. The amplitude factor of a sine wave is equal to the square root of 2 = 1.4142136.

amplitude fading In the propagation of electromagnetic waves, a condition in which the amplitudes of all components of the signal (i.e., carrier and sidebands) increase and decrease uniformly. Compare SELECTIVE FADING.

amplitude/frequency response Performance of an amplifier throughout a specified range, as exhibited by a plot of output-signal amplitude versus frequency for a constant-amplitude input signal.

amplitude gate A transducer that transmits only those portions of an input wave that lie within two close-spaced amplitude boundaries; also called *slicer.*

amplitude limiter A circuit, usually with automatic gain control (AGC), that keeps an amplifier output signal from exceeding a certain level, despite large variations in input-signal amplitude. A dc-biased diode performs passive limiting action via *clipping.*

amplitude-modulated generator A signal generator whose output is amplitude modulated. Usually, this instrument is an RF generator that is modulated at an audio frequency.

amplitude-modulated transmitter A radio-frequency transmitter whose carrier is varied in amplitude, according to the rate of change of some data-containing signal (such as voice, music, facsimile, television pictures, control signals, or instrument readings).

amplitude modulation Abbreviation, AM. A method of conveying intelligence in wireless communications and broadcasting. The modulating-signal energy appears at sideband frequencies above and below, and very close to, the carrier frequency. These sideband signals carry all the information. The extent of modulation is expressed as a percentage, from 0, which represents an unmodulated carrier, to 100, which represents full modulation. In a signal modulated 100 percent, one-third of the power is used to convey the data; the other two-thirds is consumed by the carrier. This form of modulation is essentially outmoded, although it is still used in the standard broadcast band from 535 to 1605 kHz. See FREQUENCY MODULATION, PHASE MODULATION, SINGLE SIDEBAND.

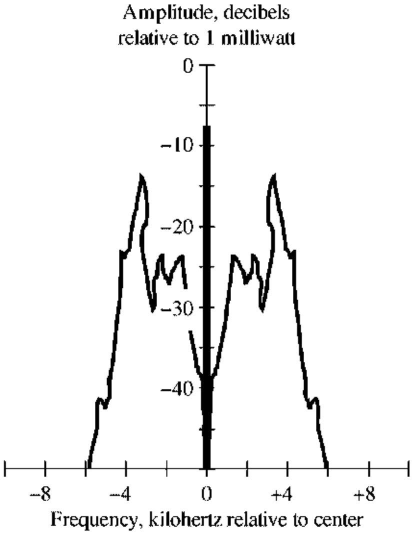

amplitude modulation

amplitude-modulation noise Spurious amplitude modulation of a carrier wave by extraneous signals and random impulses, rather than by the intended data-containing signal.

amplitude noise In radar, amplitude fluctuations of an echo returned by a target. This noise limits the precision of the system.

amplitude of noise The level of random noise in a system. The amplitude of noise is measured in the same way that signal amplitude is measured.

amplitude range The maximum-to-minimum amplitude variation of a signal. It can be expressed as a direct numerical ratio or in decibels.

amplitude response The maximum output obtainable at various frequencies over the range of an instrument operating under rated conditions.

amplitude selection The selection of a signal, according to its correspondence to a predetermined amplitude or amplitude range.

amplitude separator In a television receiver, a circuit that separates the control pulses from the composite video signal.

amplitude suppression ratio The ratio of an undesired output of a frequency-modulated (FM) receiver to the desired output, when the test signal is amplitude modulated and frequency modulated simultaneously.

amplitude-versus-frequency distortion Distortion resulting from varying gain or attenuation of an amplifier or network, with respect to signal frequency.

AMTOR A form of amateur-radio data communications, in which the accuracy of a group of characters in a message is checked periodically by the receiving station. If an error appears likely, then the receiving station sends an instruction to the transmitting station to retransmit that particular group of characters. Characters are sent in bunches with pauses for possible inquiries from the receiving station.

AM tuner A compact radio receiver unit that handles amplitude-modulated signals and delivers low-amplitude audio output to a high-fidelity amplifier. Compare AM/FM TUNER and FM TUNER.

amu Abbreviation of *atomic mass unit.*

amusement robot An electromechanical robot, often computer-controlled, that is intended for use as a toy.

AN- A prefix designator used by American military services to indicate commonality.

anacoustic Pertaining to the lack of sound or absence of reverberation or transmission of sonic waves.

analog **1.** A quantity that corresponds, point for point or value for value, to an otherwise unrelated quantity. Thus, voltage is the analog of water pressure, and current is the analog of water flow. **2.** Varying over a continuous range and, therefore, capable of attaining an infinite number of values or levels. Compare DIGITAL.

analog adder An analog circuit or device that receives two or more inputs and delivers an output equal to their sum.

analog adder/subtracter An analog circuit or device that receives two or more inputs and delivers an output equal to their sum or difference (in any combination), as desired.

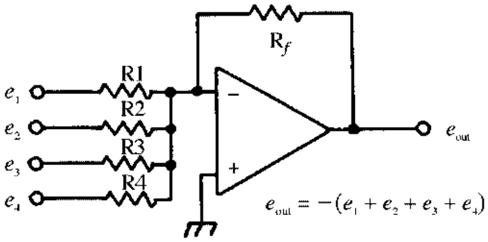

analog adder (inverting)

analog channel In an ANALOG COMPUTER, an information channel in which the extreme limits of data magnitude are fixed, and the data can have any value between the limits.

analog communications Any form of communications in which a carrier, generally an electromagnetic wave or high-frequency current, is varied in a continuous and controlled way by a data-containing signal. See ANALOG, **2.**

analog computer A computer in which input and output quantities are represented as points on continuous (or small-increment) scales. To represent these quantities, the computer uses voltages or resistances that are proportional to the numbers to be worked on. When the quantities are nonelectrical (such as pressure or velocity), they are made analogous by proportional voltages or resistances.

analog data **1.** Data represented in a quantitatively analogous way. Examples are the deflection of a movable-coil meter, the positioning of a slider on a slide rule, and the setting of a variable resistor to represent the value of a nonelectrical quantity. Also see ANALOG. **2.** Data displayed along a smooth scale of continuous values (as by a movable-coil meter), rather than in discrete steps (as by a digital meter).

analog differentiator An analog circuit or device whose output waveform is the derivative of the input-signal waveform, with respect to time.

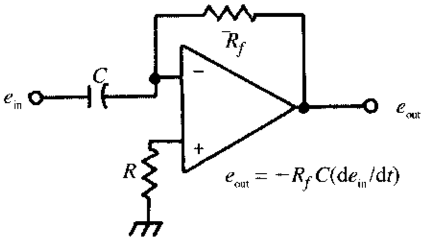

analog differentiator

analog divider An analog circuit or device that receives two inputs and delivers an output equal to their quotient.

analog electronics Electronic techniques and equipment that is based on uniformly changing signals, such as sine waves, and often having continuous-scale indicators, such as D'Arsonval meters. Compare DIGITAL ELECTRONICS.

analog information Approximate numerical information, as opposed to digital information, which is assumed to be exact.

analog integrator An analog circuit or device whose output waveform is the integral of the input signal waveform, with respect to time.

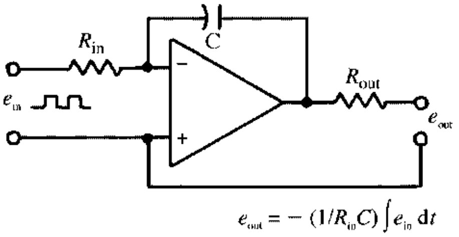

analog integrator

analog inverting adder An analog adder that delivers a sum with the opposite sign to that of the input quantities.

analog meter An indicating instrument that uses a movable-coil arrangement or the equivalent, causing a rotating pointer to indicate a particular value on a graduated printed scale. Compare DIGITAL METER.

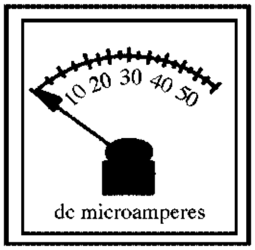

analog meter

analog multiplexer **1.** A multiplexer used with analog signals (see MULTIPLEXER). **2.** An analog time-sharing circuit.

analog multiplier An analog circuit or device that receives two or more inputs and delivers an output equal to their product.

analog network A circuit that permits mathematical relationships to be shown directly by electric or electronic means.

analogous pole In a PYROELECTRIC MATERIAL, the end or face having the positive electric charge.

analog output An output quantity that varies smoothly over a continuous range of values, rather than in discrete steps.

analog record Also called *analog recording.* A record or recording method in which some property of the recorded material, such as displacement or magnetization, varies over a continuous range that is relative to time and/or physical position.

analog recorder Any recorder, such as a recording oscillograph, potentiometric recorder, electroencephalograph, electrocardiograph, or lie detector, that produces an analog record. The counterpart is a digital recorder, which produces a readout in discrete numbers (printed or visually displayed).

analog representation Representation of information within a smooth, continuous range, rather than as separate (discrete) steps or points.

analog signal A signal that attains an infinite number of different amplitude levels, as opposed to one that can attain only a finite number of levels as a function of time.

analog subtracter An analog circuit or device that receives two inputs and delivers an output equal to their difference.

analog summer See ANALOG ADDER.

analog switch A switching device that will only pass signals that are faithful analogs of transducer parameters.

analog-to-digital conversion **1.** A process in which an analog signal (such as a voice waveform) is changed into a digital or binary signal that conveys the same information. This process is commonly used in digital computers to encode sounds and images. It is also used in communications systems to improve efficiency, minimize the necessary bandwidth, and optimize the signal-to-noise ratio. **2.** A process in which continuous mechanical motion is encoded into a digital or binary electronic signal.

analog-to-digital converter Any circuit or device that performs ANALOG-TO-DIGITAL CONVERSION.

analysis **1.** The rigorous determination of the constants and modes of operation for electronic equipment. Compare SYNTHESIS. **2.** A branch of mathematics dealing with point sets, relations, and functions.

analytical engine A primitive mechanical calculating machine, invented in 1833 by Charles Babbage.

analyzer **1.** Any instrument that permits analysis through close measurements and tests (e.g., distortion analyzer, WAVE ANALYZER, or gas analyzer). **2.** A computer program used for debugging purposes; it analyzes other programs and summarizes references to storage locations. **3.** An analysis interface to an oscilloscope.

anastigmatic yoke Also called *full-focus yoke.* In a television (TV) receiver, a deflection yoke with a cosine winding for better focus at the edges of the picture.

anchorage In plastic recording tape, the adhesion of the magnetic oxide coating to the surface of the tape.

ancillary equipment Equipment that does not directly enter into the operation of a central system. Examples are input/output components of a computer and test instruments attached to a system.

AND circuit In digital systems and other switching circuits, a logic gate whose output is high (logic 1) only when all input signals are high. Otherwise the output is low (logic 0). Compare OR CIRCUIT.

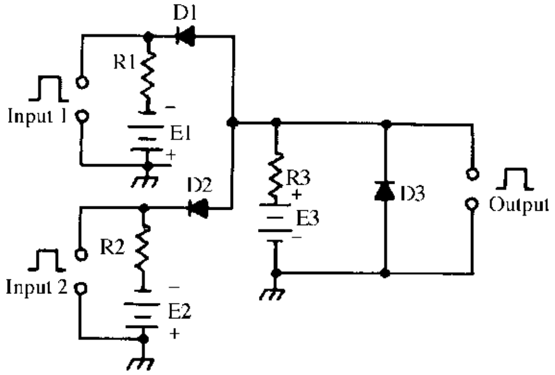

AND circuit

Anderson bridge An ac bridge circuit with six impedances, permitting the value of an unknown inductance to be determined in terms of a standard capacitance.

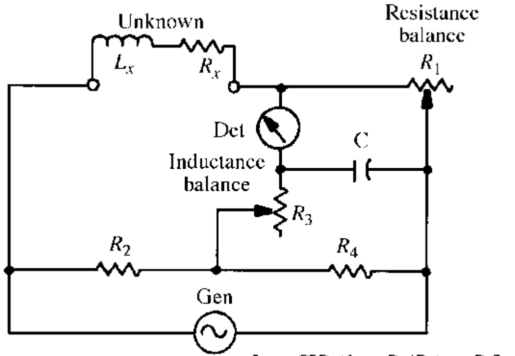

$$L_x = C[R_3(1 + R_2/R_4) + R_2]$$
$$R_x = (R_1 R_2)/R_4$$

Anderson bridge

AND gate **1.** AND circuit. **2.** In a TV receiver, an AND circuit that holds the keyed-AGC signal off until a positive horizontal flyback pulse and a horizontal sync pulse appear simultaneously at the input.

android A sophisticated robot built in humanoid form. Usually, it propels itself by rolling on wheels or on a track drive. A rotatable head contains position sensors, a machine vision system, and/or a machine hearing system. Mechanical arms are equipped with end effectors to perform various tasks. The most advanced androids have self-contained computer control systems.

anechoic Pertaining to the absence of echoes. Examples: ANECHOIC CHAMBER, *anechoic enclosure,* or *anechoic room.*

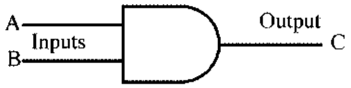

Truth Table

A	B	C
0	0	0
0	1	0
1	0	0
1	1	1

AND gate

anechoic chamber An enclosure that does not reflect sound waves that approach its walls. Such a chamber is used to test certain audio devices.

anemograph An electromechanical device that produces a recording of wind speed versus time. Generally, it consists of an ANEMOMETER connected to a PEN-AND-INK RECORDER via a suitable electronic interface.

anemometer An instrument that measures or indicates wind speed, or speed and direction (velocity).

angel **1.** An extraneous image, usually of short duration, on a cathode-ray-tube (CRT) display. The term applies particularly to anomalies in a radar image caused by low-atmospheric reflection, birds, or other mobile objects. **2.** Air-deployed metallic debris, also known as *chaff,* designed to create radar echoes as a decoy or diversion tactic.

angle jamming A radar jamming technique in which the return echo is jammed with a signal containing improper azimuth or elevation angle components.

angle modulation Variation of the angle of a sine-wave carrier in response to the modulating source, as in FREQUENCY MODULATION and PHASE MODULATION.

angle noise In radar reception, the interference resulting from variations in the angle at which an echo arrives from the target.

angle of arrival The angle which the line of propagation of an incoming radio wave makes with the surface of the earth. Compare ANGLE OF DEPARTURE.

angle of azimuth The horizontal angle between the viewer and object or target, usually measured clockwise from north.

angle of beam The angle enclosing most of the transmitted energy in the radiation from a directional antenna. It is usually measured between the *half-power points* in the main lobe of the directional pattern. This angle can be measured in the horizontal (azimuth) plane or in the vertical (elevation) plane.

angle of conduction **1.** Also called *angle of flow.* The number of degrees of an excitation-signal cycle during which output (drain, collector or plate) current flows in an amplifier circuit. **2.** The number of degrees of any sine wave at which conduction of a device (e.g., a diode) begins.

angle of convergence **1.** In any graphical representation, the angle formed by any two lines or plots that come together at a point. **2.** The angle formed by the light paths of two photocells focused on the same object.

angle of declination The angle between the horizon and a descending line. Compare ANGLE OF ELEVATION.

angle of deflection In a cathode-ray tube, the angle between the electron beam at rest and a new position resulting from deflection.

angle of departure The angle, relative to the horizon, made by the line of propagation of a transmitted radio wave. Compare ANGLE OF ARRIVAL.

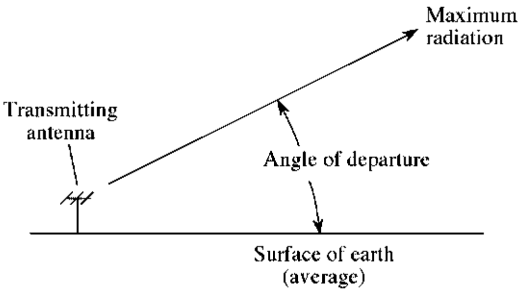

angle of departure

angle of depression See ANGLE OF DECLINATION.

angle of divergence In a cathode-ray tube, the angle formed by the spreading of an undeflected electron beam as it extends from the gun to the screen.

angle of elevation The angle that an ascending line subtends, with respect to the horizon. Compare ANGLE OF DECLINATION.

angle of flow See ANGLE OF CONDUCTION.

angle of incidence The angle, measured relative to the perpendicular (orthogonal) to a surface or boundary, subtended by an approaching ray. Compare ANGLE OF REFLECTION and ANGLE OF REFRACTION.

angle of lag The phase difference (in degrees or radians) whereby one component follows another in time, both components being of the same frequency. Compare ANGLE OF LEAD. Also see PHASE ANGLE.

angle of lead The phase difference (in degrees or radians) whereby one component precedes another in time, both components being of the same frequency. Compare ANGLE OF LAG. Also see PHASE ANGLE.

angle of radiation **1.** The angle, measured with respect to the horizon, at which the principal lobe of an electromagnetic wave leaves a transmitting antenna. **2.** The angle, measured relative to the horizon, of a receiving or transmitting antenna's optimum sensitivity.

angle of reflection The angle, measured relative to the perpendicular (orthogonal) to a surface, subtended by a ray leaving the surface after having been reflected from it. Compare ANGLE OF INCIDENCE.

angle of refraction The angle, measured relative to the perpendicular (orthogonal) to a boundary between two different media, subtended by a ray leaving the boundary after having been refracted thereat. Compare ANGLE OF INCIDENCE.

angle tracking noise Noise in a servo system that results in a tracking error.

angstrom (Anders J. Angstrom, 1814–1874). A unit of length used to describe certain extremely short waves and microscopic dimensions; 1 angstrom equals 10^{-4} microns (10^{-10} meters).

angular deviation loss The ratio of microphone or loudspeaker response on the principal axis of response to the response at a designated angle from that axis. Expressed in decibels.

angular difference See PHASE ANGLE.

angular displacement In an ac circuit, the separation, in degrees, between two waves. See PHASE ANGLE.

angular frequency The frequency of an ac signal, expressed in *radians per second (rad/sec)* and approximately equal to $6.28f$, where f is the frequency in Hertz.

angular length Length, as along the horizontal axis of an ac wave or along the standing-wave pattern on an antenna, expressed as the product of radians and wavelength.

angular-mode keys On a calculator or computer, the DEG, RAD, and GRAD keys for expressing or converting angles in DEGREES, RADIANS, and GRADS, respectively.

angular phase difference For two sinusoidal waves, the phase difference, expressed in degrees or radians.

angular rate In navigation, the rate of bearing change, expressed in degrees or radians.

angular resolution The ability of a radar to distinguish between two targets by angular measurement.

angus pen recorder An instrument that makes a permanent record of the time whenever a channel is used.

anharmonic oscillator An oscillating device in which the force toward the balance point is not linear, with respect to displacement.

anhysteresis The magnetization of a material by a unidirectional field containing an alternating field component of gradually decreasing amplitude.

anhysteretic state The condition of a substance after it has been subjected to a strong magnetic field, the intensity of which alternates in direction and diminishes gradually to zero.

animism A belief or philosophy, held especially in Eastern civilizations, such as Japan, that all things contain an essence of life. This theory renders irrelevant the question of whether or not machines, such as computers and robots can be "alive."

anion A negative ion. Also see ION.

anisotropic Pertaining to the tendency of some materials to display different magnetic and other physical properties along different axes.

ANL Abbreviation of AUTOMATIC NOISE LIMITER.

anneal To heat a metal to a predetermined temperature and let it cool slowly. The operation prevents brittleness and often stabilizes electrical characteristics.

annealed laminations Core laminations for transformers or choke coils that have been annealed.

annealed shield A magnetic shield for cathode-ray tubes, that has been processed by annealing.

annealed wire Soft-drawn wire that has been subjected to annealing.

annotations **1.** Marking on copies of original engineering-installation documents to show changes made during the installation. **2.** Any set of comments or notes accompanying a program, an equipment or system, or a process.

annular **1.** Pertaining to the region between two concentric circles that lie in the same plane; ring-shaped. **2.** Pertaining to two or more concentric circles that lie in a common plane.

annular conductor A number of wires stranded in three concentric layers of alternating twists around a hemp core.

annular transistor A mesa transistor in which the base and collector take the form of concentric rings around a central emitter.

annulling network A subcircuit that shunts a filter to cancel reactive impedance at the extreme ends of the pass band of the filter.

annunciation relay A relay that indicates whether or not a circuit is carrying current.

annunciator A device that produces loud sound and/or conspicuous light to attract attention (e.g., the electronic siren in an automotive security system).

anode **1.** The positive electrode of a vacuum tube or solid-state device (i.e., the electrode toward which electrons move during current flow). **2.** In an electrochemical cell, the electrode that loses electrons by oxidation. This is usually the negative electrode.

anode balancing coil Mutually coupled windings used to maintain equal currents in parallel anodes operating from a common transformer terminal.

anode current Current flowing in the anode circuit of a device.

anode efficiency Also called *plate efficiency.* In a power amplifier using an electron tube, the ratio P_o/P_i, where P_o is the output power in watts and P_i is the dc anode power input in volt-amperes.

anode power input Symbol, $P_{A(input)}$. The product of anode current and anode voltage.

anode power supply The ac or positive dc power supply unit that delivers current and voltage to the anode of a device.

anode saturation The point beyond which a further increase in anode voltage does not produce an increase in anode current.

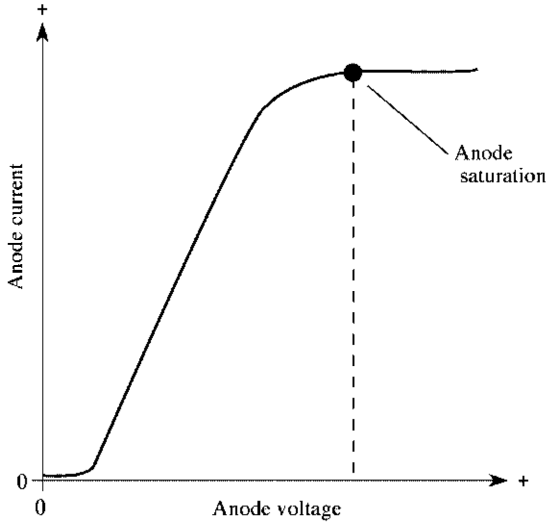

anode saturation

anode strap In a multicavity magnetron, a metal strap connecting the anodes.

anode terminal **1.** In a diode, the terminal to which a positive dc voltage must be applied for forward bias. Compare CATHODE TERMINAL. **2.** In a diode, the terminal at which a negative dc voltage appears when the device is used as an ac rectifier. Compare CATHODE TERMINAL. **3.** The terminal that is connected internally to the anodic element of any device.

anode voltage Symbol, E_A or V_A. The difference in potential between the anode and cathode of a device.

anodic Pertaining to the anode of a device, or to anode-like effects.

anodizing An electrolytic process in which a protective oxide film is deposited on the surface of a metallic body acting temporarily as the anode of the electrolytic cell.

anomalous dispersion Dispersion of electromagnetic radiation that is characterized by a decrease in refractive index with increase in frequency.

anomalous propagation **1.** The low-attenuation propagation of UHF or microwave signals through atmospheric layers. **2.** Unusual, bizarre, or unexplainable electromagnetic-wave propagation (e.g., apparent F-layer ionospheric effects in the FM broadcast band). **3.** Rapid fluctuation of a sonar echo because of variations in propagation.

anoxemia toximeter An electronic instrument for measuring or alerting against the onset of anoxemia (deficiency of oxygen in the blood)—especially in airplane pilots.

AN radio range A navigational facility entailing four zones of equal signal strength. When the aircraft deviates from course, an aural Morse-code signal, A (DIT DAH) or N (DAH DIT) is heard; but when the aircraft is on course, a continuous tone is heard.

ANSI Acronym for *American National Standards Institute.*

AN signal The signal provided by an AN radio range to apprise aircraft pilots of course deviation.

answerback The automatic response of a terminal station to a remote-control signal.

answer cord In a telephone system, the cord used for answering subscribers' calls and incoming trunk calls.

answering machine A device that automatically answers a telephone and records an audio message from the caller.

answer lamp A telephone switchboard lamp that lights when an answer cord is plugged into a line jack; it switches off when the telephone answers and lights when the call is completed.

ant Abbreviation of ANTENNA.

antenna In a communications system, a specialized transducer that converts incoming electromagnetic fields into alternating electric currents having the same frequencies (*receiving antenna*), or converts an alternating current at a specific frequency into an outgoing electromagnetic field at the same frequency (*transmitting antenna*). An antenna can be a simple wire or rod, or a complicated structure. Thousands of geometries and specifications are possible. The optimum antenna type for a given situation depends on the communications frequency, the distance to be covered, and various other factors.

antenna ammeter An RF ammeter, usually of the thermocouple type, employed to measure current flowing to a transmitting antenna.

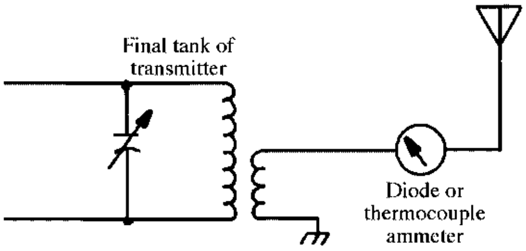

antenna ammeter

antenna amplifier **1.** A radio-frequency amplifier, often installed at the antenna, used to boost signals before they reach a receiver (also called an *RF preamplifier*). **2.** Occasionally, the first RF amplifier stage of a receiver, also known as the *front end.*

antenna array See ARRAY.

antenna bandwidth The frequency range throughout which an antenna will operate at a specified efficiency without needing alteration or adjustment.

antenna beamwidth A measure of the extent to which a directional antenna focuses a transmitted electromagnetic field, or focuses its response to incoming electromagnetic fields. Expressed as the angle in degrees between opposite *half-power points* in the main lobe of the directional pattern. Usually determined in the horizontal plane, but occasionally in the vertical plane.

antenna coil The primary coil of the input RF transformer of a receiver, or the secondary coil of the output RF transformer of a transmitter.

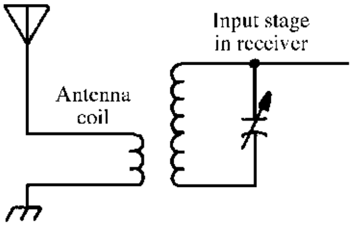

antenna coil

antenna coincidence The condition in which two directional antennas are pointed directly toward each other.

antenna-conducted interference Extraneous signals generated in a transmitter or receiver and presented to the antenna, from which they are radiated.

antenna core A ferrite rod or slab around which a coil of wire is wound to act as a self-contained antenna, usually in a miniature receiver.

antenna coupler A device consisting of an inductor, RF transformer, or a combination of inductor(s) and capacitor(s), used to match the impedance of an antenna to that of a transmitter or receiver. Also known as a *transmatch* or *antenna tuner.*

antenna coupling Inductive and/or capacitive coupling used to optimize the transfer of energy from an antenna to a receiver, or from a transmitter to an antenna.

antenna current **1.** Radio-frequency current flowing from a transmitter into an antenna. **2.** Radio-frequency current flowing from a receiving antenna into a receiver.

antenna detector A circuit that warns aircraft personnel that they are being observed by radar. It picks up the radar pulses and actuates a warning light or other device.

antenna diplexer A coupling device that permits several transmitters to share one antenna without troublesome interaction. Compare ANTENNA DUPLEXER.

antenna directivity The directional characteristics of a transmitting or receiving antenna, usually expressed qualitatively (e.g., *omnidirectional*, *bidirectional*, or *unidirectional*). A more precise expression is ANTENNA BEAMWIDTH.

antenna director In a directional antenna, a PARASITIC ELEMENT situated in front of the radiator and separated from it by an appropriate fraction of a wavelength. Its function is to intensify radiation in the direction of transmission. Compare ANTENNA RADIATOR and ANTENNA REFLECTOR.

antenna duplexer A circuit or device permitting one antenna to be shared by two transmitters without undesirable interaction.

antenna effect The tendency of wires or metallic bodies to act as antennas (i.e., to radiate or receive radio waves).

antenna efficiency The ratio of radio-frequency energy supplied to a wireless transmitting antenna, to the energy radiated into space. Electrically, the radiation resistance of the antenna (R_R) appears in series with loss resistance (R_L). The efficiency *Eff* of the antenna can be determined by the following formula:

$$Eff = R_R/(R_R + R_L)$$

As a percentage,

$$Eff_{\%} = 100\ (R_R/(R_R + R_L)$$

The efficiency is always less than 1 (100 percent) because, in practice, the loss resistance can never be reduced to zero.

antenna factor A factor (in decibels) added to an RF voltmeter reading to find the true open-circuit voltage induced in an antenna.

antenna field The electromagnetic field immediately surrounding an antenna.

antennafier Low-profile antenna/amplifier device, sometimes used with portable communications systems. Also called an *active antenna.*

antenna front-to-back ratio For a directional antenna, the ratio of field strength in front of the antenna (i.e., directly forward in the line of maximum directivity) to field strength in back of the antenna (i.e., 180 degrees from the front), as measured at a fixed distance from the radiator. It is usually specified in decibels.

antenna gain For a given antenna, the ratio of signal strength (received or transmitted) to that obtained with a comparison antenna, such as a simple dipole. Generally specified in decibels.

antenna ground system The earth, counterpoise, guy wires, radials, and/or various conducting objects in the vicinity of an antenna which, taken together, form the radio-frequency (RF) ground system against which the antenna operates. Some antennas require an extensive ground system to function efficiently; others need no ground system.

antenna/ground system An arrangement embodying both an antenna and a low-resistance connection to the earth, as opposed to an antenna system that involves no connection to earth.

antenna height **1.** The height of an antenna above the surface of the earth immediately beneath the driven element(s). **2.** The height of an antenna above the effective radio-frequency (RF) ground immediately beneath the driven element(s). **3.** The height of an antenna above average terrain, determined against the mean altitude of a number of points on the earth's surface that lie within a certain radius of the antenna structure. Also called *height above average terrain (HAAT).*

antenna impedance The complex-number impedance that an antenna presents to a transmission line. It can vary over a tremendous range, and depends on the antenna type, antenna size, antenna height, operating frequency, and various other factors.

antenna-induced potential Also called *antenna-induced microvolts.* The voltage across the open-circuited terminals of an antenna.

antenna lens Also called *lens antenna.* A radiator designed to focus microwave energy in much the same manner that an optical lens focuses light rays. Lens antennas are made from dielectric materials and/or metals.

antenna loading **1.** The insertion of inductance in antenna elements to lower the resonant frequency of the system without necessarily making the system physically larger or the elements longer. **2.** The insertion of capacitance in antenna elements to raise the resonant frequency of the system without necessarily making the system physically smaller or the elements shorter.

antenna lobe A well-defined region in the radiation pattern of an antenna in which radiation is most intense, or in which reception is strongest. Also see ANTENNA PATTERN.

antenna matching The technique of establishing a satisfactory relationship between the antenna impedance and the transmission-line or transmitter-output impedance, for maximum transfer of power into the antenna. Also, the matching of antenna impedance to receiver-input impedance, for delivery of maximum energy to the receiver.

antennamitter An antenna/oscillator combination that serves as a low-power transmitter.

antenna pattern A polar plot of antenna performance that shows field strength versus angle of azimuth, with the antenna at the center. It is usually specified in the horizontal plane.

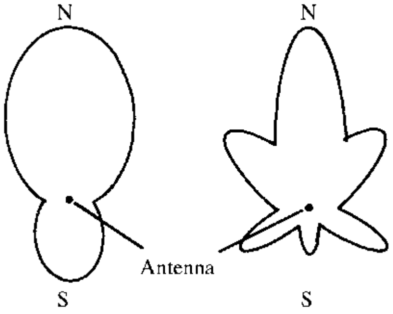

antenna patterns

antenna polarization The orientation of electric lines of flux, with respect to the surface of the earth, for which an antenna is most efficient. A vertical antenna radiates and receives vertically polarized waves. A horizontal antenna radiates and receives horizontally polarized waves broadside to itself, and vertically polarized waves at high elevation angles off its ends. In other directions, the polarization is slanted at various angles.

antenna power Symbol, P_{ant}. The RF power developed in an antenna by a transmitter; P_{ant} equals I^2R, where I is the antenna current and R is the antenna resistance at point I is measured.

antenna power gain The ratio of the maximum effective radiated power (ERP) from a wireless transmitting antenna to the ERP from a reference antenna, expressed in decibels (dB). If the ERP from an antenna under test is P_T watts and the ERP from the reference antenna is P_R watts, then the gain G_{dB} is:

$$G_{dB} = 10 \log_{10} (P_T/P_R)$$

Power gain is always measured in the direction in which the test antenna performs the best. The reference antenna, usually a dipole, is chosen with a gain assumed to be unity, or 0 dB. Gain relative to a dipole is expressed in dBd (decibels relative to a dipole). Alternatively, the reference antenna can be an isotropic radiator, in which case the gain is expressed in dBi (decibels relative to an isotropic radiator). Gain figures in dBd and dBi differ by a constant amount as follows:

$$G_{dBi} = 2.15 + G_{dBd}$$

antenna preamplifier A highly sensitive amplifier used to enhance the gain of a receiver. It is usually used at the very high frequencies and above.

antenna radiation The propagation of radio waves by a transmitting antenna.

antenna radiator The element of an antenna that receives RF energy from the transmitter and radiates waves into space. Also known as the *driven element.* Compare ANTENNA DIRECTOR and ANTENNA REFLECTOR.

antenna range **1.** The frequency band, communication distance characteristically covered, or other continuum of values that specify the operating limits of an antenna. **2.** The region immediately surrounding an antenna in which tests and measurements usually are made. Sometimes called ANTENNA FIELD.

antenna reflector In a directional antenna, a PARASITIC ELEMENT situated behind the radiator and separated from the latter by an appropriate fraction of a wavelength. Its function is to intensify radiation in the direction of transmission. Compare ANTENNA DIRECTOR and ANTENNA RADIATOR.

antenna relay In a radio station, a low-loss, heavy-duty relay that enables the antenna to be switched between transmitter and receiver.

antenna resistance The resistive component of ANTENNA IMPEDANCE.

antenna resonant frequency The frequency, or narrow band of frequencies, at which an antenna's impedance appears resistive.

antenna stage **1.** The first RF amplifier stage of a receiver. **2.** Occasionally, the final RF amplifier of a transmitter.

antenna switch In a radio station, a low-loss, heavy-duty switch that enables the antenna to be connected to transmitter, receiver, or safety ground.

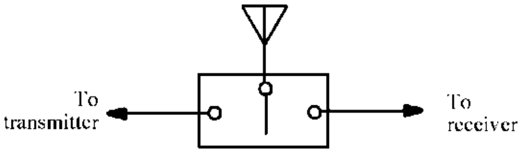

antenna switch (single-pole, double-throw)

antenna system Collectively, an antenna and all of the auxiliary electrical and mechanical devices needed for its efficient operation, including couplers, tuners, transmission lines, supports, insulators, and rotator.

antenna terminals **1.** The points at which a transmission line is attached to an antenna. **2.** The signal input terminals of a receiver. **3.** The signal output terminals of a transmitter.

antennaverter An antenna and converter combined into a single circuit, intended for connection to the antenna terminals of a receiver to allow operation on frequencies outside the band for which the receiver has been designed.

antenna wire **1.** The radiator element of a wire-type antenna. **2.** A strong solid or stranded wire (e.g., hard-drawn copper, copper-clad steel, or phosphor-bronze) used for antennas.

anthropomorphism The perception, by people, of machines as having human qualities. This can lead to emotional attachment to hardware, such as computers and robots. The more sophisticated the apparatus, in general, the more powerful this perception can become.

antialiasing filter A low-pass or bandpass filter that limits the bandwidth of an input signal to prevent *aliasing* and its effects. See ALIASING, **1**.

anticapacitance switch A switch whose members are thin blades and stiff wires widely separated to minimize capacitance between them.

anticathode The target electrode of an X-ray tube.

Anticipatory Sciences A group of *futurists*, people who attempt to predict the course of technology. Some futurists believe that progress will continue until, for example, homes become fully automated and artificial intelligence reaches a level comparable to human intelligence. Other futurists believe that such things are highly improbable.

anticlutter circuit A supplementary circuit in a radar receiver that minimizes the effect of extraneous reflections that would obscure the image of the target.

anticlutter gain control In a radar receiver, a circuit that automatically raises the gain of the receiver slowly to maximum after each transmitter pulse to reduce the effect of clutter-producing echoes.

anticoincidence Noncoincidental occurrence of two or more signals. Compare COINCIDENCE.

anticoincidence circuit In computers and control systems, a circuit that delivers an output signal only when two or more input signals are not received simultaneously. Compare COINCIDENCE CIRCUIT. Also see NAND CIRCUIT.

anticoincidence operation An exclusive-OR operation.

anticollision radar A vehicular radar system that is used to minimize the probability of a collision with another vehicle, whether or not that other vehicle has a similar system.

antiferroelectric **1.** Pertaining to the property wherein the polarization curve of certain crystalline materials shows two regions of symmetry. **2.** A material that exhibits the aforementioned property.

antiferromagnetic Pertaining to the behavior of materials in which, at low temperatures, the magnetic moments of adjacent atoms point in opposite directions.

antihunt The condition in which *hunting* is counteracted, usually by removing overcorrection in automatic control or compensation systems.

antihunt circuit **1.** A circuit that minimizes or eliminates hunting. Also see ANTIHUNT. **2.** In a television (TV) receiver, a circuit that stabilizes an automatic frequency control (afc) system.

antijamming Pertaining to communications systems that are resistant to, or that counteract, the effects of *jamming.*

antilogarithm Abbreviated, antilog or $\log^{-1}$. The number corresponding to a given logarithm. For example, $\log 10{,}000 = \log 10^4 = 4$, and thus antilog $4 = 10^4 = 10{,}000$.

antilogous pole In a PYROELECTRIC MATERIAL, the end that becomes negatively charged as the temperature rises.

antimagnetic Pertaining to materials having extremely low RETENTIVITY.

antimatter Pertaining to particles that are the counterparts of conventional particles (i.e., positrons instead of electrons, antineutrons instead of neutrons, and antiprotons instead of protons). When a particle meets its antiparticle, the two annihilate, releasing energy. Also see ANTIPARTICLE.

antimicrophonic See NONMICROPHONIC.

antimony Symbol, Sb. A metalloidal element. Atomic number, 51. Atomic weight, 121.76. Often used as n-type dopant in semiconductor manufacture.

antineutrino The antiparticle of the NEUTRINO, emitted as a result of radioactive decay.

antineutron An uncharged particle with a mass equal to that of the neutron, but with a magnetic moment in the direction opposite that of the neutron.

antinode A point of maximum amplitude in a standing wave.

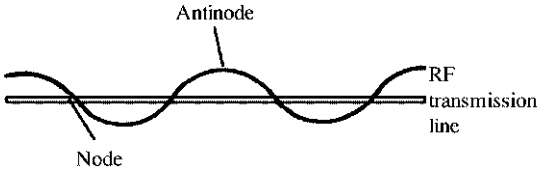

antinode

antinoise carrier-operated circuit A circuit that cuts off the audio output of a receiver while the station transmitter is in use. This can be accomplished in the automatic-gain-control (AGC) circuit of the receiver, or in the speaker or audio line. The circuit is actuated by energy from the transmitted signal.

antinoise microphone Any microphone that discriminates against acoustic noise (e.g., a *lip microphone* or *throat microphone*).

antinucleon A particle with the mass of a nucleon, but with the opposite electrical charge and direction of magnetic moment. Compare NUCLEON.

antioxidant A material, such as a lacquer coat or an inactive oxide layer, that prevents or slows oxidation of a material exposed to air.

antiparticle A subatomic particle opposite in character to conventional particles, such as electrons, neutrons, protons. Antiparticles constitute *antimatter.* Also see ANTINEUTRINO, ANTINEUTRON, ANTINUCLEON, ANTIPROTON, and POSITRON.

antiphase The property of being in phase opposition (180 degrees out of phase).

antipincushioning magnets In some television (TV) receivers, a pair of corrective magnets in the deflection assembly on the picture tube that eliminate *pincushion distortion* (disfigurement of the raster so that it resembles a pincushion—a rectangle with its sides bowed in).

antiproton A subatomic particle with a mass equal to that of the proton, but with opposite electrical charge.

antiquark An ANTIPARTICLE of a QUARK.

antirad substance A material that protects against damage caused by atomic radiation.

antiresonance **1.** Parallel resonance. **2.** The condition of being detuned from a resonant frequency.

antiresonant circuit See PARALLEL-RESONANT CIRCUIT.

antiresonant frequency **1.** The resonant frequency of a parallel-resonant circuit. **2.** In a piezoelectric crystal, the frequency at which impedance is maximum (as in a parallel-resonant circuit).

antisidetone Pertaining to the elimination in telephone circuits of interference between the microphone and earphone of the same telephone.

antistickoff voltage The low voltage applied to the coarse synchro control transformer rotor winding in a dual-speed servo system to eliminate ambiguous behavior in the system.

antitransmit/receive switch Abbreviated *ATR.* In a radar installation, an automatic device to prevent interaction between transmitter and receiver.

antivirus program A computer program or utility designed to detect and eliminate viruses and Trojan horses in a computer system.

antivoice-operated transmission Radio communications that use a voice-activated circuit as a transmitter interlock during reception on the companion receiver.

apc **1.** Abbreviation of *automatic picture control.* **2.** Abbreviation of AUTOMATIC PHASE CONTROL.

aperiodic Characterized by a lack of predictable repetitive behavior. For example, the *sferics* or "static" electromagnetic interference caused by lightning.

aperiodic current The unidirectional current that follows an electromagnetic disturbance in an LCR circuit, in which R is equal to or higher than the critical circuit resistance.

aperiodic damping Damping of such a high degree that the damped system, after disturbance, comes to rest without oscillation or hunting.

aperiodic discharge A discharge in which current flowing in an LCR circuit is unidirectional, rather than oscillatory. For this condition, $1/LC$ is less than or equal to $R^2/4L^2$.

aperiodic function A nonrepetitive function (e.g., a hyperbolic trigonometric function).

aperture **1.** The larger, normally open end of a horn antenna or horn loudspeaker. **2.** An opening in an opaque disk or mask that passes a predetermined amount of light or other radiant energy. **3.** The portion of a directional antenna through which most of the radiated energy passes.

aperture angle For an antenna or telescope or microscope, the half angle formed by the radius of the detecting instrument, as viewed from the source.

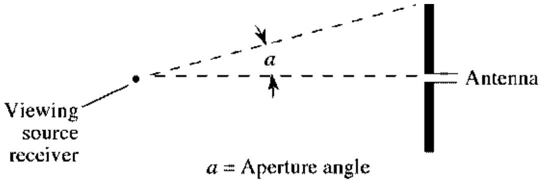

aperture angle

aperture antenna An antenna whose beamwidth depends on the size of a horn, reflector, or lens.

aperture compensation In a television (TV) camera, the minimizing of APERTURE DISTORTION by widening the video-amplifier passband.

aperture distortion In a television (TV) camera tube, a form of distortion that occurs when the scanning beam covers several mosaic elements simultaneously. This condition, caused by excessive beam thickness, results in poor image resolution.

aperture mask In a three-gun color picture tube, a thin, perforated sheet mounted behind the viewing screen to ensure that a particular color phosphor will be excited only by the beam for that color. Also called *shadow mask.*

aperture synthesis In telescopes, a method of obtaining high resolution using several small antennas separated by great distances. The small antennas are moved around to simulate the resolving power of a much larger antenna that would, in practice, be impossible or impractical to construct.

aphelion **1.** The point at which a solar-orbiting satellite attains its highest altitude. It occurs once for every complete orbit. At this point, the satellite travels slower than at any other point in the orbit. **2.** The altitude, measured from the sun's surface or the sun's center, of a solar-orbiting satellite at its most distant point.

APL Abbreviation for *A Programming Language.* A high-level computer language designed for ease of use, and characterized by the requirement for a special character set.

apl **1.** Abbreviation of *average picture level.* **2.** Abbreviation of *automatic phase lock.*

A plus Also, A+. The positive terminal of an A battery. Also, pertaining to the part of a circuit connected to that terminal.

apogee **1.** The point at which an earth-orbiting satellite attains its highest altitude. It occurs once for every complete orbit. At this point, the satellite travels slower than at any other point in the orbit. **2.** The altitude, measured from the earth's surface or the earth's center, of an earth-orbiting satellite at its most distant point.

A power supply A term sometimes used to denote the unit that supplies energy to a vacuum-tube filament. Compare B POWER SUPPLY.

apparent bearing In radio-direction finding, the uncorrected direction from which a signal appears to arrive.

apparent power In an ac circuit, the power value obtained by multiplying the current by voltage (P equals IE), with no consideration of the effects of phase angle. Compare TRUE POWER.

apparent power loss The loss in an ammeter or voltmeter, caused by the imperfection of the instrument. At full scale, the ammeter has a certain voltage across its terminals; the apparent power loss is the current multiplied by this voltage. A voltmeter carries a small current; the apparent power loss is the product of the current and the indicated voltage.

appearance potential The potential through which an electron must move to produce a certain ion from the atom with which it is associated.

applause meter An instrument consisting essentially of a microphone, audio amplifier, and indicating meter (reading directly in sound level). It is so called because of its familiar use in measuring audience response, as indicated by loudness of applause.

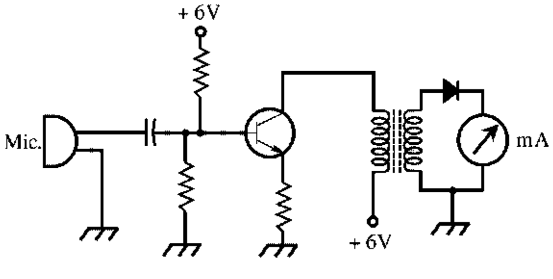

applause meter

Applegate diagram For a velocity-modulated tube, a plot of the positions of electron bunches in the drift space versus time.

Appleton layer Collectively, the F1 and F2 layers of the ionosphere, at a height between 150 and 400 kilometers above the surface of the earth.

apple tube A color picture tube, used in television, with the red, blue, and green phosphor in vertical strips.

appliance Electrical equipment in general. This might include any home-operated device.

application A task or job for which an electronic device or system is used. It especially pertains to personal-computer software that has practical usefulness.

application factor A factor involved in determining the failure rate of a circuit or system affected by unusual operating conditions.

application schematic diagram A diagram of pictorial symbols and lines that illustrate the interrelationship of functional circuit blocks in a specific program mode.

applicators **1.** In dielectric heating, the electrodes between which the dielectric body is placed and the electrostatic field developed. **2.** In medical electronics, the electrodes applied to a patient undergoing diathermy or ultrasonic therapy.

applied voltage The voltage presented to a circuit point or system input, as opposed to the voltage drop resulting from current flow through an element.

applique circuit A circuit for adapting equipment to a specialized job.

approach-control radar A radar installation serving a ground-controlled approach (GCA) system.

approximate data **1.** Data obtained through physical measurements. Such data can never be exact; all measurements are subject to error. **2.** Loosely estimated data or imprecise calculations.

AQL Abbreviation of ACCEPTABLE QUALITY LEVEL. A statistically defined quality level, defined in terms of percent defective, accepted on an average of 95 percent of the time.

Aquadag A tradename for a material that consists of a slurry of fine particles of graphite. Aquadag forms a conductive coating on the inside and outside walls of some cathode-ray tubes.

aqua pura Pure water; in most instances, distilled water. Formula, H_2O. Pure water is a nonconductor with a dielectric constant of about 81.

Ar Symbol for ARGON.

arbitrary function fitter A circuit or device, such as a potentiometer, curve changer, or analog computer element, providing an output current or voltage that is some preselected function of the input current or voltage.

arc **1.** A luminous sustained discharge between two electrodes. Because it is sustained, rather than intermittent, an arc is distinguished from a spark discharge, the latter being a series of discharges (sparks)—even when it appears continuous. **2.** In graphical presentations, a section of curved line, as of a circle.

arc angle The angle in degrees traced out by a circular arc if the center point of the circle is considered to be the vertex of an angle formed by two rays intersecting the arc at designated points.

arc cosecant Abbreviated arc csc or $\csc^{-1}$. **1.** The inverse of the *cosecant* function. **2.** The angle, in

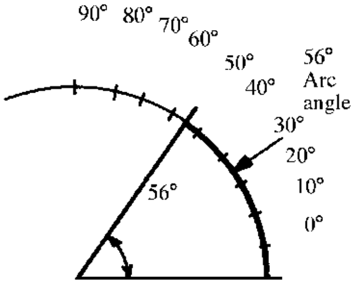

arc angle

radians or degrees, corresponding to a given cosecant.

arc cosine Abbreviated arc cos or $\cos^{-1}$. **1.** The inverse of the *cosine* function. **2.** The angle, in radians or degrees, corresponding to a given cosine.

arc cotangent Abbreviated arc cot or $\cot^{-1}$. **1.** The inverse of the *cotangent* function. **2.** The angle, in radians or degrees, corresponding to a given cotangent.

arc failure **1.** Damage to, and/or failure of, insulation or a dielectric as a result of ARCOVER. **2.** Failure of make-and-break contacts through damage caused by arcover.

arc function An *inverse trigonometric function.* See ARC COSECANT, ARC COSINE, ARC COTANGENT, ARC SECANT, ARC SINE, and ARC TANGENT.

arc furnace A high-temperature electric furnace in which heat is produced by one or more electrical arcs.

architecture The functional design elements of a computer—especially the components of the *central processing unit (CPU)* and the manner in which these elements interact.

archived file A computer file stored on some backup medium, such as magnetic tape, diskette, or CD-ROM (compact disk, read-only memory), rather than being held on the hard disk. Such a file will be apart from the operating system's catalog of current files, but can be reconstituted as needed.

archives A complete, periodically updated set of ARCHIVED FILES.

arcing See ARCOVER.

arcing contacts Make-and-break contacts between which an arc occurs when they are separated.

arcing ring A metal ring placed around an insulator in a high-voltage electrical system. This keeps an arc from charring or breaking the insulator.

arcing time The elapsed time between the breaking of contacts and the end of the arc between the contacts.

arc lamp An electric lamp in which a brilliant arc jumps between the tips of two rods (originally carbon).

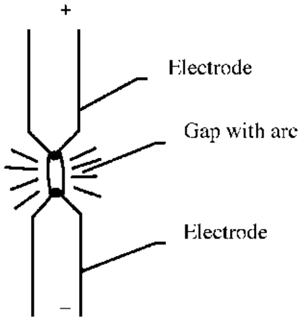

arc lamp

arc length The length along a given arc, usually a part of the circumference of a circle. If the circle has circumference C and the arc measures x degrees, then the arc length is $Cx/360$ units.

arc minute See MINUTE.

arc oscillation Oscillations that can occur when opening relay contacts arc.

arcover The occurrence of an electrical ARC between electrodes, contacts, or capacitive plates.

arcover voltage The voltage at which disruptive discharge occurs, typically accompanied by an arc.

arc resistance The ability of a material, usually a dielectric, to resist damage from arcing. This property is commonly expressed as the length of time between the start of the arc and the establishment of a conductive path through the material.

arc secant Abbreviated arc sec or $\sec^{-1}$. **1.** The inverse of the *secant* function. **2.** The angle, in radians or degrees, that corresponds to a given secant.

arc second See SECOND.

arc sensor A device for detecting visible arcs and excessive reflected power in microwave systems.

arc sine Abbreviated arc sin or $\sin^{-1}$. **1.** The inverse of the *sine* function. **2.** The angle, in radians or degrees, that corresponds to a given sine.

arc suppression Extinguishing an arc discharge. Disruptive arcs in electronic circuits are suppressed by means of auxiliary diodes or resistor-capacitor networks.

arc-suppressor diode A semiconductor diode used to prevent arcing between make-and-break contacts.

arc tangent Abbreviated arc tan or $\tan^{-1}$. **1.** The inverse of the *tangent* function. **2.** The angle, in radians or degrees, corresponding to a given tangent.

arcthrough The puncturing of a material by an arc.

area code In the United States, a three-digit number that indicates the location, according to specified assigned districts, of a telephone subscriber. When making a long-distance call, the area code of the desired station must be given in addition to the seven-digit telephone number.

area protection Coverage of a defined region, in terms of area or volume, by an alarm system.

area redistribution A scheme to determine the effective duration of an irregularly shaped pulse. A rectangle is constructed whose height is equal to the peak height of the pulse, as displayed on an oscilloscope. The rectangle width is adjusted until the area of the rectangle is the same as the area under the curve representing the pulse. The width of the rectangle then represents the effective duration of the pulse.

area search The scanning of a large group of computer records for those of a major category or class.

area sensor A transducer, used with an alarm system, that protects a defined region or volume, such as an office or bedroom.

Argand diagram Named after Jean Robert Argand, (1768-1822) of Geneva, for his work on the graphical representation of complex numbers. A graphical illustration of a complex number in the form *A* + *jB*, where the real-number (*A*) axis is perpendicular to the imaginary-number (*jB*) axis. The value *j* is the square root of -1, the unit imaginary number. The axes are perpendicular, usually with the *A* axis horizontal. The length of the line from the point (0,0) to the point (*A*,*jB*) is the amplitude of the vector **X** = *A* + *jB*. The direction is specified as the angle, in degrees or radians, of the vector measured counterclockwise from the *A* axis.

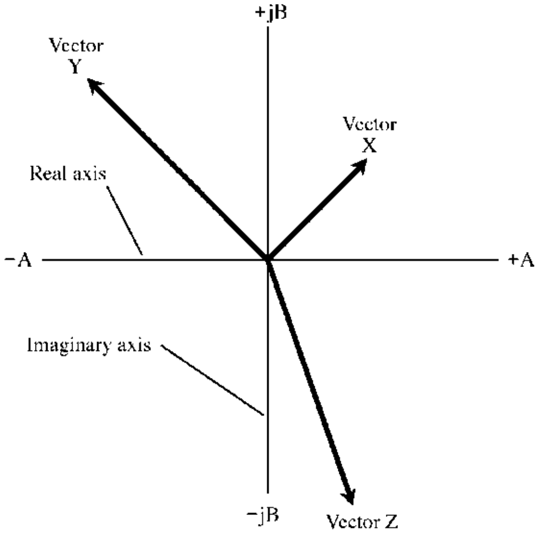

Argand diagram

argon Symbol, Ar. An inert gaseous element. Atomic number, 18. Atomic weight, 39.94. Argon, present in small amounts in the earth's atmosphere, is used in various specialized devices, such as lasers.

argon laser A laser whose tube is filled with argon gas. It generates coherent light at specific wavelengths that are characteristic of elemental argon.

argument **1.** The direction angle of a polar vector. **2.** An independent variable whose value determines the value of a function.

arithmetic address An address obtained by performing an arithmetic operation on another address.

arithmetic and logic unit Abbreviation, ALU. The part of a digital computer containing the circuits that perform calculations and logic operations; distinguished from mass storage, input/output, and peripheral units.

arithmetic circuit Also called *arithmetic element.* In a digital computer, a circuit that is involved in the execution of calculations. Included are adders, storage registers, accumulators, subtracters, and multipliers.

arithmetic mean The average of a group of quantities, obtained by dividing their sum by the number of quantities.

arithmetic operation In digital computer practice, a numerical process performed: addition, subtraction, multiplication, division, comparison.

arithmetic progression A mathematical series in which each term following the first is obtained by adding a constant quantity to the preceding one. For example, *S* = 1, 2, 3, 4, . . . *n*. Compare GEOMETRIC PROGRESSION.

arithmetic shift In a digital computer, the multiplication or division of a quantity by a power of the base used in the notation.

arithmetic sum The sum of two or more quantities disregarding their signs. Compare ALGEBRAIC SUM.

arithmetic symmetry A filter response that is exactly symmetrical about the center frequency when the frequency scale is linear.

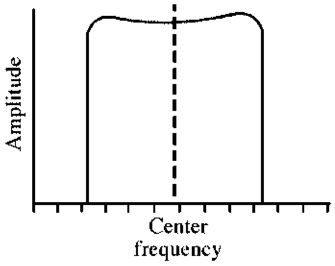

arithmetic symmetry

arm **1.** Any of the distinct branches of a circuit or network. Also called *leg.* **2.** A movable element in a device, usually containing a contact for switching.

armature **1.** The rotating member of a motor. **2.** The rotating member of some types of electro-mechanical generator. **3.** The movable member of a relay, bell, buzzer, or gong. **4.** The movable member of an actuator. **5.** The soft-iron *keeper* placed across the poles of a permanent magnet to conserve power.

armature coil A coil of insulated wire wound on a ferromagnetic core to provide the electromagnetic properties of an armature. In a motor or generator, the armature coil is distinguished from the FIELD COIL.

armature core The ferromagnetic core upon which the armature coil of a motor or generator is wound.

armature gap **1.** In a motor or generator, the space between an armature core and the pole of a field magnet. **2.** In a relay, the space between the armature and the relay-coil core.

armature hesitation A momentary delay in the movement of a relay.

armature-hesitation contact chatter Undesired (usually rapid, repetitive) making and breaking of relay contacts. Generally caused by armature hesitation.

armature-impact contact chatter Undesired (usually rapid, repetitive) making and breaking of relay contacts, caused by *contact bounce* when the armature strikes the relay core (closure) or backstop (opening).

armature relay A relay that uses an electromagnet to pull a lever toward or away from a set of fixed contacts.

armature travel The distance traveled by an armature during relay operation.

armor A protective metal cable covering.

Armstrong FM system (Edwin H. Armstrong, 1890–1954). A phase-shift method of frequency modulation. See PHASE MODULATION.

armature voltage control A means of controlling motor speed by changing the applied armature winding voltage.

armchair copy An amateur radio term for reception of exceptionally clear signals.

arming the oscilloscope sweep Enabling an oscilloscope to trigger on the next pulse by closing a switch.

Armstrong oscillator (Edwin H. Armstrong, 1890–1954). An oscillator circuit that uses inductive feedback between the output and input. Either the output coil or the input coil can be tuned to set the oscillator frequency. The amount of positive feedback is controlled by varying the coupling between the coils.

Armstrong superheterodyne circuit See SUPERHETERODYNE CIRCUIT.

Armstrong superregenerative circuit See SUPERREGENERATIVE CIRCUIT.

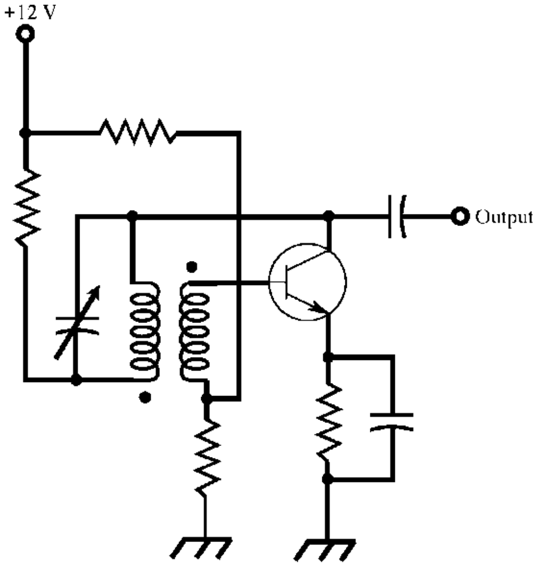

Armstrong oscillator

ARPA Acronym for *Advanced Research Projects Agency,* a subsidiary of the U.S. Department of Defense.

array **1.** A directive antenna that consists of an assembly of properly dimensioned and spaced elements, such as radiators, directors, and reflectors. **2.** A coordinated group or matrix of components, such as diodes, resistors, memory cells, etc., often enclosed in one capsule. **3.** Subscripted variables representing data arranged so that a program can examine the array and extract data relevant to a particular subscript.

array device A group of similar or identical components that are connected together in a certain fashion, to perform a specific task.

arrester **1.** A device used to protect an installation from lightning. It consists of a varistor or an air gap connected between an antenna or power line and an earth ground. The device passes little or no current under ordinary conditions, but passes heavy current to ground during a lightning stroke. Also called LIGHTNING ARRESTER. **2.** A self-restoring protective device used to reduce voltage surges on power lines.

ARRL Abbreviation for *American Radio Relay League.*

arrowhead A wideband, log-periodic antenna with linear polarization.

ARS Abbreviation of *Amateur Radio Service.*

arsenic Symbol, As. A metalloidal element. Atomic number, 33. Atomic weight, 74.91. Arsenic is familiar as an n-type dopant in semiconductor processing.

ARSR Abbreviation of *air route surveillance radar.*

articulation A measure of the effectiveness of voice communications, expressed as the percentage of speech units understood by the listener when the effect of context is negligible.

artificial antenna See DUMMY ANTENNA.

artificial ear A microphone-type sensor, equivalent to the human ear, used to measure sound pressures.

artificial echo **1.** In radar practice, the reflections of a transmitted pulse returned by an artificial target. **2.** A signal from a pulsed radio-frequency (RF) generator, delayed to simulate an echo.

artificial ground The effective ground provided by the radials or disk of a ground-plane antenna, as opposed to actual ground (the earth itself). Compare TRUE GROUND.

artificial horizon In aircraft instrumentation, a device that displays lines showing the position of the aircraft in flight, with reference to the horizon.

artificial intelligence Abbreviation, AI. **1.** A specialized field of computer science overlapping with electronics, biology, physiology, and other sciences, concerning attempts to develop advanced computer systems that can emulate the processes of the human mind. **2.** The ability of a computer to learn from its mistakes, refine its own processes, and perhaps ultimately reason in a humanlike manner.

artificial ionization An artificial reflecting layer that is created in the atmosphere to provide a skip condition.

artificial language A language that is not commonly used, but has been devised for efficiency in a particular situation—especially in a computer system.

artificial life **1.** The ultimate endpoint of ARTIFICIAL INTELLIGENCE, wherein machines acquire qualities, such as wisdom and the capability to feel emotions. The state of the art is currently nowhere near this point. **2.** A hypothetical machine or set of machines with lifelike qualities, including human-level intelligence, wisdom, and emotion.

artificial stimulus An electronic method of robot guidance and navigation using radar, sonar, vision systems, edge detection, and/or beacons.

artificial transmission line A network of capacitors and inductors with characteristics similar to those of the more bulky transmission line it replaces in tests and measurements. It also serves as a time-delay or phase-shift device and as a pulse-forming network.

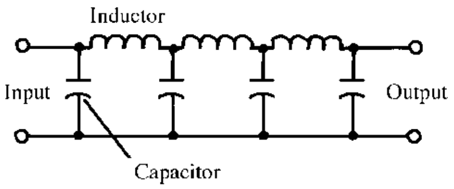

artificial transmission line

artificial voice A device used to test and calibrate noise-canceling microphones, consisting essentially of a small loudspeaker that has a baffle whose acoustical properties simulate those of the human head.

artos stripper A machine that cuts and strips wire for the fabrication of multiconductor cables.

artwork **1.** In the manufacture of printed circuits, the scaled drawings from which the mask or etch pattern is obtained photographically. **2.** Collectively, the illustrations depicting an electronic circuit, device, or system.

As Symbol for ARSENIC.

ASA Abbreviation of AMERICAN STANDARDS ASSOCIATION.

asbestos A nonflammable fibrous material consisting of calcium and magnesium silicates that is used for high-temperature insulation.

A-scan A radar-screen presentation in which the horizontal time axis displays distance or range, and the vertical axis displays the amplitude of signal pulse and echo pulses.

ascending node For a satellite orbiting the earth or another planet, any point at which the groundtrack crosses the equator as it moves from the southern hemisphere into the northern hemisphere. This node generally changes for each succeeding orbit, because the earth or planet rotates underneath the orbit of the satellite. Compare DESCENDING NODE.

ascending pass For a specific point on the earth's surface, the time during which an artificial communications satellite is accessible when its latitude is moving northward. The duration of accessibility depends on the altitude of the satellite, and on how close its groundtrack comes to the earth-based point. Compare DESCENDING PASS.

ASCII Acronym (pronounced "*ask*-ee") for *American Standard Code for Information Interchange.*

ASI Abbreviation for *American Standards Institute.*

A-scope A radar system that displays an A-SCAN.

Askarel A synthetic, nonflammable liquid dielectric.

aspect ratio The width-to-height ratio of a video image, generally three units high by four units wide.

asperities On the surface of an electrode, tiny points at which the electric field is intensified and from which discharge is highly probable.

ASR **1.** Abbreviation of AIRBORNE (or AIRPORT) surveillance radar. **2.** Abbreviation of AUTOMATIC SEND/ RECEIVE.

ASRA Acronym for *automatic stereophonic recording amplifier.*

assemble **1.** To gather subprograms into a complete digital computer program. **2.** To translate a

Symbols for ASCII teleprinter code

First four signals	*Last three signals*							
	000	*001*	*010*	*011*	*100*	*101*	*110*	*111*
0000	NUL	DLE	SPC	0		P	/	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	/	l	/
1101	CR	GS	–	=	M	]	m	}
1110	SO	RS	.	>	N		n	~
1111	SI	US	|	?	O	-	o	DEL

ACK: acknowledge
BEL: bell
BS: back space
CAN: cancel
CR: carriage return
DC1: device control no. 1
DC2: device control no. 2
DC3: device control no. 3
DC4: device control no. 4
DEL: delete
DLE: data link escape
ENQ: enquiry
EM: end of medium
EOT: end of transmission
ESC: escape
ETB: end of transmission block
ETX: end of text

FF: form feed
FS: file separator
GS: group separator
HT: horizontal tab
LF: line feed
NAK: do not acknowledge
NUL: null
RS: record separator
SI: shift in
SO: shift out
SOH: start of heading
SPC: space
STX: start of text
SUB: substitute
SYN: synchronous idle
US: unit separator
VT: vertical tab

symbolic program language into a machine (binary) language program by substituting operation codes and addresses.

assembly **1.** A finished unit that can be either a practical working model or a dummy, a prototype, or a final model; an integrated aggregation of subunits. **2.** A low-level computer source-code language that uses crude mnemonics that are easier to remember than the *machine-language* equivalents.

assembly language A source code that uses mnemonic instructions. (See ASSEMBLY, **2.**)

assembly program The program that operates on a symbolic-language program to produce a machine language program in the process of assembly. Also called *assembler.*

assembly robot A form of *industrial robot* that puts hardware together. Such a robot is generally a component of an *automated integrated manufacturing system (AIMS).* The robot can do repetitive work at high speed and precision for long periods of time.

assign To reserve part of a computing system for a specific purpose, normally for the duration of a program run.

assigned frequency The radio carrier frequency or band of frequencies designated for a transmitting station by a licensing authority. Also see RADIO SPECTRUM.

associative memory Computer memory in which locations are identified by content, rather than by specific address.

assumed decimal point A decimal point that does not occupy an actual computer storage space, but is used by the computer to align values for calculation; the decimal point is assumed to be at the right unless otherwise specified.

astable Having two temporary states; BISTABLE.

astable circuit A circuit that has two unstable states, and whose operation is characterized by alternation between those states at a frequency determined by the circuit constants.

astable multivibrator A free-running multivibrator. The common circuit uses two bipolar or field-effect transistors, their inputs and outputs being cross coupled. Conduction switches alternately between the two.

astatic **1.** Without fixed position or direction. **2.** In a state of neutral equilibrium.

astatic galvanometer A galvanometer with a movable element consisting of two identical magnetized needles mounted nonparallel on the same suspension. Each needle is surrounded by a coil. The coils are wound in opposite directions, and are connected in series to the current source. A large permanent magnet provides the field against which the needle assembly rotates. The instrument functions independently of the geomagnetic field.

astatine Symbol, At. A radioactive elemental halogen produced from radioactive decay. Atomic number, 85. Atomic weight, 210. Formerly called *alabamine.*

A station One of the two stations in the transmitting system of LORAN (long-range navigation).

astigmatism A focusing fault in a cathode-ray tube (CRT), in which electrons in different axial planes focus at different points.

ASTM Abbreviation for *American Society for Testing and Materials.*

astrionics The design, production, and application of electronic devices and systems for use in space vehicles and space navigation.

astronomical unit Abbreviation, AU. A unit of distance equal to 1.496×10^8 kilometers (9.296×10^7 miles). Approximately equal to the mean distance between the earth and the sun.

A supply See A POWER SUPPLY.

asymmetrical cell A photocell exhibiting ASYMMETRICAL CONDUCTIVITY.

asymmetrical communications **1.** Two-way communications in which the volume of transmitted data is much greater in one direction than in the other. **2.** Two-way communications in which the speed of transmitted data is much greater in one direction than in the other. Compare SYMMETRICAL COMMUNICATIONS.

asymmetrical conductivity A condition in which a device conducts well in one direction, but poorly in the other direction. A rectifier diode is a common example of a component that exhibits this effect.

asymmetrical distortion In a binary system, lengthening or shortening of one of the states, by comparison to the theoretical or ideal duration.

asymmetrical FET A FIELD-EFFECT TRANSISTOR in which the source and drain cannot be interchanged without degrading performance.

asymmetrical multivibrator An unbalanced multivibrator (i.e., one in which the circuit halves are not identical). If the time constants of the halves are different, the output pulses will be short and widely separated.

asymmetrical sideband See VESTIGIAL SIDEBAND.

asymmetrical sideband transmission See VESTIGIAL SIDEBAND TRANSMISSION.

asymmetrical wave A wave whose upper (positive half-cycle) and lower (negative half-cycle) portions have different amplitudes or shapes. Also called *asymmetric wave.*

asymmetry control An adjustment in a device intended for measuring the pH (acidity/alkalinity). This corrects the inaccuracies that results from the differences between the electrodes.

asymptote In analytical geometry, a fixed straight line or ray *L* with a special relationship to a curve or part of a curve *K* that recedes to infinity. As the distance from the origin (0,0) increases without limit, the separation between *K* and *L* approaches zero, but *K* and *L* never actually meet.

asymptotic breakdown voltage A voltage that will cause *dielectric breakdown* if applied continuously for a sufficiently long time.

asymptotic expression An expression having a very small error in terms of percentage.

asynchronous **1.** Not synchronous, i.e., nonrecurrent (as in out-of-phase waves). **2.** A mode of computer operation in which the completion of one operation starts another.

asynchronous device A device not regulated by the system in which it is used, as far as its operating frequency or rate is concerned.

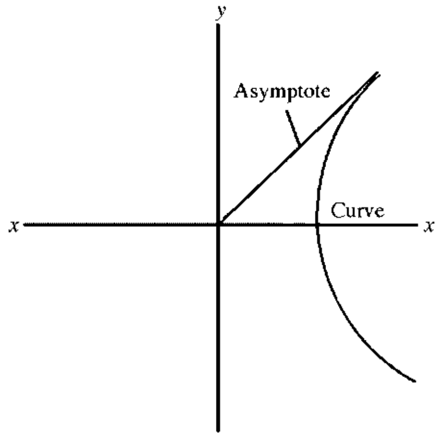

asymptote

asynchronous input In digital circuitry, any flip-flop input at which a pulse can affect the output independently of the clock.

asychronous motor An ac motor whose speed is not proportional to the supply frequency.

asynchronous transmission Data transmission in which each character or symbol begins with a start signal and ends with a stop signal. This eliminates the need for the data to be sent at a uniform speed.

asynchronous vibrator In a vibrator-type portable power supply, a vibrator that only makes and breaks the primary circuit of the step-up transformer. This is in contrast to the synchronous vibrator, which also makes and breaks the secondary circuit in synchronism with the primary. Also called NONSYNCHRONOUS VIBRATOR.

AT A quartz crystal cut wherein the angle between the x-axis and the crystal face is 35 degrees.

At Symbol for astatine.

AT-cut crystal A piezoelectric crystal cut at a 35-degree angle, with respect to the optical axis of the quartz. The frequency of such a crystal does not appreciably change with variations in temperature.

atmosphere **1.** The gas surrounding a planet, particularly the air sheathing the earth. **2.** Abbreviation, atm. A unit of pressure equal to 1.013×10^6 dynes per square centimeter (about 14.7 pounds per square inch).

atmospheric absorption **1.** The conversion of electromagnetic energy into heat, with resulting loss, as the energy passes through the earth's atmosphere. The extent of this effect depends on the wavelength. **2.** See ABSORPTION LOSS, **2.**

atmospheric absorption noise Noise, principally above 1 GHz, resulting from atmospheric absorption (see ABSORPTION LOSS, **2**).

atmospheric bending The refraction or reflection of electromagnetic waves by the troposphere or ionosphere. See ATMOSPHERIC REFLECTION.

atmospheric duct A tropospheric stratum, often associated with temperature inversions, lake effects, or weather fronts, through which electromagnetic energy at ultra-high and microwave frequencies is efficiently propagated for long distances.

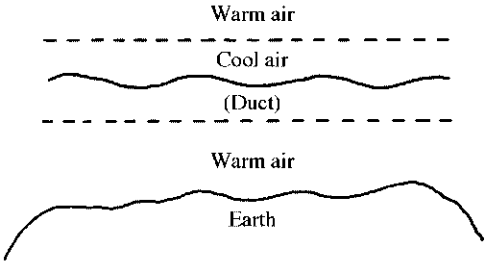

atmospheric duct

atmospheric electricity Static electricity present in the atmosphere, which evidences itself in disturbance of radio communications and in displays of lightning.

atmospheric noise Receiver noise resulting from ATMOSPHERIC ELECTRICITY. Also called *sferics* or *static.*

atmospheric pressure Abbreviation, atm press. **1.** The pressure exerted by the earth's atmosphere, as indicated by a barometer at sea level; normally between 29 and 31 inches of mercury. **2.** A pressure of 1.013×10^6 dynes per square centimeter. See ATMOSPHERE, **2.**

atmospheric radio wave See SKYWAVE.

atmospheric radio window The band of frequencies (approximately 10 MHz to 10 GHz), including radio waves that can penetrate the earth's troposphere and ionosphere.

atmospheric reflection The return of a radio wave to earth, resulting from reflection by an ionized portion of the atmosphere.

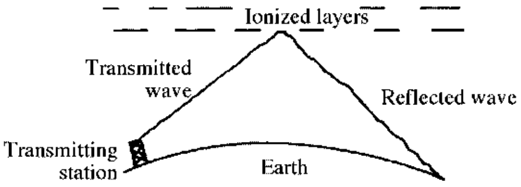

atmospheric reflection

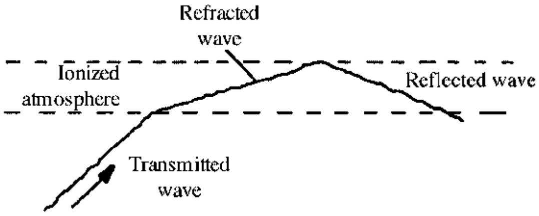

atmospheric refraction

atmospheric refraction **1.** Downward bending of radio waves as a result of variations in the dielectric constant of the troposphere. **2.** Downward bending of radio waves in the ionosphere, resulting in long-range propagation at high frequencies.

atmospheric scatter **1.** The scattering of very-high frequency (VHF) and ultra-high frequency (UHF) radio waves by the lower atmosphere. **2.** Communication via scattering of VHF and UHF radio waves in the lower atmosphere.

atmospherics See ATMOSPHERIC NOISE.

atom **1.** The smallest material particle that displays the unique characteristics and properties of an element. Atoms consist of a dense, positively charged central nucleus, around which less-massive, negatively charged electrons "swarm" at definite levels called *shells.* Also see BOHR ATOM and RUTHERFORD ATOM. **2.** In a computer-compiling operation, an operator or operand.

atomechanics The physics of electron movement.

atomic battery A battery in which atomic energy is converted into electrical energy.

atomic charge The electrification (i.e., the electron charge) exhibited by an ion.

atomic clock Also called *atomic time standard.* A highly accurate electronic clock, driven by the characteristic oscillations of certain atoms.

atomic energy Energy released by the FUSION or FISSION of atomic nuclei. Also see ATOMIC POWER.

atomic fission See FISSION.

atomic frequency The natural vibration frequency of an atom.

atomic fusion See FUSION.

atomic pile See REACTOR, **2.**

atomic mass unit Abbreviated amu. A unit that expresses the relative mass of an elemental *isotope.* One amu is equal to 1/12 of the atomic mass of carbon 12 (C12). A neutron has a mass of roughly one amu.

atomic migration The transfer or "wandering" of a valence electron between or among atoms in a single molecule.

atomic number The number of protons in the nucleus of an atom. Also, the number of electrons if the atom is electrically neutral. For example, the atomic number for copper is 29, indicating 29 protons in the nucleus. An electrically neutral atom of copper has 29 electrons. The atomic number uniquely identifies an element.

atomic radiation The emission of radiant energy by radioactive substances.

atomic reactor See REACTOR, **2.**

atomic theory The scientific theory that all matter is composed ultimately of *atoms,* which are the smallest particles retaining the identity of an element. Atoms combine to form *molecules*, the smallest particles that retain the identity of a compound. Atoms themselves contain minute *subatomic particles*, some of which carry electric charges. See BOHR ATOM and RUTHERFORD ATOM.

atomic time **1.** A means of time determination that makes use of the resonant vibrations of certain substances, such as cesium. **2.** Synchronized astronomical time, as determined by an ATOMIC CLOCK.

atomic unit of energy In a hydrogen atom, the potential energy of the electron in the lowest-energy shell, as averaged over a certain length of time. The shell represents the mean energy of the electron.

atomic weight **1.** The mass of a particular atom in ATOMIC MASS UNITS (amu). **2.** A number characterizing the average mass of individual atoms for a specific isotope of an element. Thus, carbon 12 (C12) has an atomic weight of 12, oxygen 16 (O16) has an atomic weight of approximately 16, and uranium 238 (U238) has an atomic weight of about 238.

atomistics The science of the atom and atomic energy. Also called *atomics.*

attack **1.** The rise of a pulse from zero to maximum amplitude. **2.** The time required for a pulse to rise from zero to maximum amplitude. **3.** The initialization of a circuit voltage or current for a certain purpose, such as an automatic gain control. **4.** The rise of a musical note from zero to full volume.

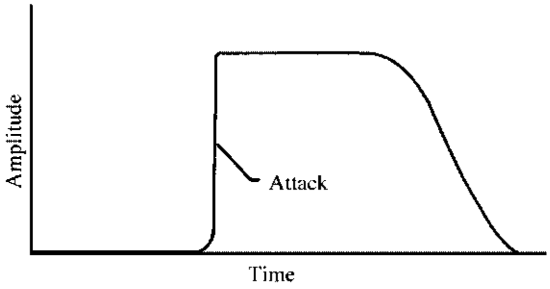

attack

attack time The time required for an applied signal that suddenly increases in amplitude to reach 63.2 percent of its final, stable value.

attemperator An automatic temperature-controlling device; a *thermostat.*

attention display A computer-generated chart or graph, displayed as an alert signal concerning a particular situation.

attenuate To reduce in amplitude.

attenuation A reduction of signal amplitude.

attenuation characteristic Also called *attenuation constant.* **1.** In an amplifier, network, or component, the decrease in signal amplitude as a function of frequency, usually expressed in decibels per octave. **2.** In a transmission line, the decrease in signal amplitude per unit length. Usually expressed in decibels per 100 feet, decibels per mile, or decibels per kilometer.

attenuation constant See ATTENUATION CHARACTERISTIC.

attenuation distortion A type of distortion characterized by variation of attenuation with frequency within a given frequency range.

attenuation equalizer An equalizer that stabilizes the transfer impedance between two ports at all frequencies within a specified frequency band.

attenuation-frequency distortion Distortion characterized by the attenuation of the frequency components in a complex waveform. Frequency-sensitive RC networks (such as a Wien bridge) exhibit this type of distortion when they attenuate a fundamental and each harmonic unequally.

attenuation network A combination of components (R, C, or L singly or in any necessary combination) that provide constant signal attenuation with negligible phase shift throughout a frequency band.

attenuation ratio The ratio indicating a relative current, voltage, power or energy decrease. For example, for voltage, $E_{input}/E_{output} = 6/2 = 3{:}1 = 3$.

attenuator A device for reducing signal amplitude in precise, predetermined steps, or smoothly over a continuous range. A network of resistors, capacitors, or both. The simplest attenuator consists of one or more noninductive resistors.

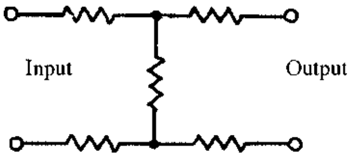

attenuator

attitude The position of an aircraft or space vehicle relative to a (usually terrestrial) reference point, often determined with electronic instruments.

atto- Abbreviated, a. A prefix meaning 10^{-18} or multiplication by 10^{-18}.

attofarad Abbreviation, aF. An extremely small unit of low capacitance; 1 aF equals 10^{-18} F.

attracted-disk electrometer A device to measure potential difference consisting of two parallel metal disks—one of which is connected to a tension spring. The force between the disks indicates the magnitude of the electric field.

attraction The drawing together or pulling toward, as in the attraction between electric charges or magnetic poles. Dissimilar charges and poles attract each other (electric plus to minus, magnetic north to south). Compare REPULSION.

ATV Abbreviation of *amateur television*, used in the Amateur Radio Service.

AU Abbreviation of ASTRONOMICAL UNIT.

Au Symbol for GOLD.

audibility The quality of being detectable by the human ear. In a healthy listener, the *threshold of audibility* is extremely low; at the threshold, the pressure of a sound wave varies from normal by approximately 10^{-4} pascal. The frequency range of human audibility extends roughly from 20 Hz to 20 kHz.

Sound	*Audibility (dB)*
Threshold of hearing	0
Whisper	10–20
Electric fan at 10 feet	30–40
Running water at 10 feet	40–60
Speech at 5 feet	60–70
Vacuum cleaner at 10 feet	70–80
Passing train at 50 feet	80–90
Jet at 1000 feet altitude	90–100
Rock band on stage	110–120
Air hammer at 5 feet	130–140

audibility table

audibility curve A graph (such as the *Fletcher-Munson curve*) that depicts the range of human hearing in terms of frequency versus the sound pressure at the threshold of AUDIBILITY.

audible Detectable by the human ear.

audible alarm device An ANNUNCIATOR that produces an easily identifiable sound in response to an ALARM CONDITION in a security system.

audible frequency See AUDIO FREQUENCY.

audible tone A vibration of air molecules that can be detected by the human ear, and with periodic properties, such as a sine-wave vibration.

audio **1.** Pertaining to the spectrum of frequencies corresponding to the human hearing range (about 20 Hz to 20 kHz), or to equipment or performance associated with that spectrum. **2.** Any disturbance, such as a current or compression wave, falling within the range of about 20 Hz to 20 kHz. **3.** AUDIO FREQUENCY.

audio amplifier See AUDIO-FREQUENCY AMPLIFIER.

audio band The range (band) of audio frequencies.

audio channel **1.** The portion of a complex signal or waveform used to convey audio information exclusively. **2.** The audio-frequency section of a transmitter or receiver (as opposed to the radio-frequency section). **3.** A radio channel of fixed frequency that is reserved for voice communications.

audio clipping Brute-force limiting of the amplitude of an audio signal, usually accomplished using semiconductor diodes to prevent the positive and negative peak amplitudes from exceeding a certain level.

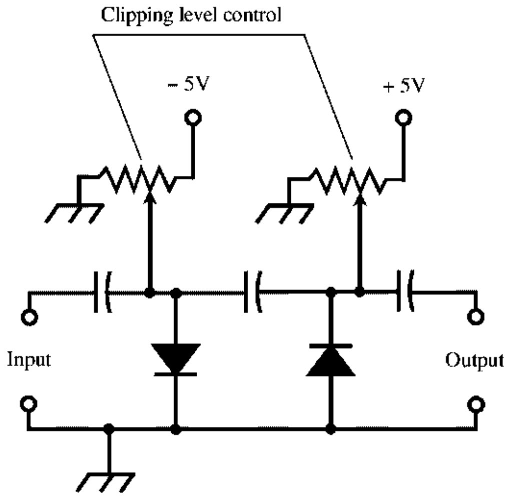

audio clipping

audio component The audio-frequency portion of any wave or signal.

audio converter A circuit in which a received radio-frequency (RF) signal is heterodyned with a local RF oscillator signal to produce an audio-frequency (AF) beat-note output. The beat note is then amplified by an AF amplifier. It is used especially by amateur radio operators in the reception of continuous-wave (CW) radiotelegraphy, radioteletype, and packet radio at high frequencies.

audio frequency A frequency lying within the audible spectrum. Abbreviated AF. See AUDIO-FREQUENCY SPECTRUM.

audio-frequency amplifier An amplifier that operates in part or all of the frequency range 20 Hz to 20 kHz. High-fidelity amplifiers function over a somewhat wider range (e.g., 10 Hz to 50 kHz).

audio-frequency choke An inductor (usually having a ferromagnetic core) that blocks audio-frequency current, but passes direct current.

audio-frequency feedback **1.** Electrical FEEDBACK (positive and/or negative) that affects audio-frequency circuits. **2.** ACOUSTIC FEEDBACK.

audio-frequency filter A filter of any type that operates on any part of the frequency range 20 Hz to 20 kHz.

audio-frequency meter An instrument to measure frequencies in the audio-frequency spectrum (approximately 20 Hz to 20 kHz). Three types are commonly used:

- ***Analog*** Gives direct indications of frequency on the scale of a D'Arsonval meter; the usual range is 20 Hz to 100 kHz.
- ***Digital*** Gives direct indications of frequency by means of readout lamps; the usual range is 1 Hz to 15 MHz. This instrument is useful also as a radio-frequency meter.
- ***Bridge*** Consists of a frequency-sensitive bridge, such as a Wien bridge, with a null-indicating meter. The operator balances the bridge and reads the unknown frequency from the dial of the balance control.

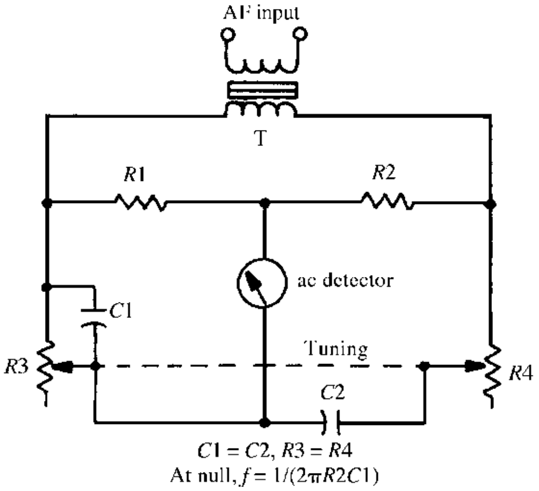

audio-frequency meter

audio-frequency noise Any electrical noise signal causing interference within the audio-frequency spectrum.

audio-frequency oscillator See AUDIO OSCILLATOR.

audio-frequency peak limiter Any circuit or device, such as a biased diode, that performs the function of audio limiting.

audio-frequency-shift keying Abbreviation, AFSK. Frequency-shift keying that is done at audio frequencies (below approximately 20 kHz) rather than at radio frequencies. There are two audio sine-wave signals, one for logic 1 (*high* or *mark*) pulses and the other for the logic 0 (*low* or *space*) pulses. This scheme is commonly used with telephone modems where the signal bandwidth is severely limited by circuit characteristics. At typical data speeds in twisted-pair telephone lines (usually 28.8 or 57.6 kbps), signals of this type sound like a hiss or roar. Compare FREQUENCY-SHIFT KEYING.

audio-frequency-shift modulator A modulator for audio-frequency-shift keying of a signal.

audio-frequency spectrum The band of frequencies extending from roughly 20 Hz to 20 kHz. High-fidelity component specifications extend this range somewhat in both directions (e.g., from 10 Hz to 50 kHz).

audio-frequency transformer Abbreviation, AF transformer. A device used for the purpose of matching impedances at frequencies within the range of human hearing (up to approximately 20 kHz). This ensures the most efficient possible transfer of power between stages of audio amplification, between an amplifier and a speaker or headset, or between a microphone and an audio preamplifier. These transformers are available with various power ratings and impedance-matching ratios. Some devices are tailored to have a certain attenuation-versus-frequency response. At audio frequencies, transformers are physically similar to the alternating-current transformers used in power supplies. They are wound on laminated or powdered-iron cores. Compare RADIO-FREQUENCY TRANSFORMER.

audio-frequency transistor A transistor that is usually used only at audio frequencies.

audiogram A graph used to rate hearing, used by audiologists and audiometrists.

audio image In a direct-conversion receiver, a response to a signal on one side of (above or below) the local-oscillator (LO) frequency, when the operator is listening to a signal on the other side of the LO frequency. These responses are reduced or eliminated in single-signal receivers.

audio-level meter An ac meter for monitoring signal amplitude in an audio-frequency system. It can indicate in volts, decibels, volume units (VU), or arbitrary units, and is often permanently connected in the circuit.

audio limiter A limiter or clipper operated in the audio-frequency (AF) channel of a receiver or transmitter to hold the output-signal amplitude constant, or to minimize the effect of noise peaks.

audiologist A person skilled in testing hearing (i.e., in using audiometers and other electronic instruments) and evaluating their indications for the fitting of hearing aids.

audiometer An instrument used for hearing tests, which consists of a specialized audio-frequency (AF) amplifier with calibrated attenuators, output meter, and signal source.

audiometrist A person skilled in the use of audiometers and other electronic instruments that measure sound and human hearing, and who deals with attendant health and behavior problems. Compare ACOUSTICIAN and AUDIOLOGIST.

audio mixer An amplifier circuit for blending two or more audio-frequency (AF) signals, such as those delivered by microphones or receivers.

audio oscillator **1.** An oscillator that delivers an output signal in the frequency range 20 Hz to 20 kHz. **2.** An audio-frequency (AF) signal generator. Some instruments of this type operate above and below the limits of the common audio-frequency spectrum (e.g., 1 Hz to 1 MHz).

audio output The output of an audio-frequency oscillator or amplifier. It can be measured in terms of peak or rms volts, amperes, or watts.

audiophile A sound-reproduction hobbyist.

audio power Alternating-current power at frequencies roughly between 20 Hz and 20 kHz. When used in connection with transmitters and other modulated radio-frequency (RF) equipment, the term refers to modulator power output.

audio response unit A device that links digitized responses, held in computer storage, to a telephone set or line to answer incoming calls and inquiries.

audio signal generator See AUDIO OSCILLATOR, **2.**

audio spectrum The range of sine-wave frequencies detectable by the human ear when they occur as acoustic vibrations. This range is about 20 Hz to 20 kHz.

audio squelch A squelch circuit that operates only on the audio channel of a receiver.

audio system **1.** The portion of any electronic assembly that is used to process sound. **2.** Special computer equipment capable of storing and processing digitized audio-frequency (AF) data.

audiotape Magnetic tape for the recording and reproduction of data in the audio-frequency (AF) range.

audio taper In potentiometers, a semilogarithmic variation of resistance versus rotation. Used in volume and tone controls for audio circuits. At midposition (the halfway point), the counterclockwise portion of the device has about 1⁄10 the resistance of the clockwise portion. A listener will hear sound at half-volume because of the logarithmic nature of the human audibility curve.

audio-visual Pertaining to a combination of sound and sight (e.g., television and sound motion pictures).

auditory backward inhibition A subjective phenomenon, in which a sound is erased from the memory of a listener by a second sound arriving about 60 milliseconds later.

auditory inhibition The tendency of sound waves to be partially or totally canceled by the ears/mind of a listener, depending on the waves' intensity, relative phase, and/or direction of impact.

auditory mirage See ACOUSTIC MIRAGE.

audit trail A history of the processes relating to a record, transaction, or file in a computer system. Created during the routine processing of data, the trail is stored as a file. The audit trail allows auditing of the system or the subsequent recreation of files.

augend In a calculation, the number to which another is to be added. Compare ADDEND.

augend register In a digital computer, the register that stores the augend. Compare ADDEND REGISTER.

aural Pertaining to sound actually heard, as opposed to sound that exists only as audio-frequency currents or waves.

aurora A phenomenon sometimes called the *northern lights* or *southern lights*, as seen in the night sky. In the Northern Hemisphere, it is known as *Aurora Borealis*; in the Southern Hemisphere, it is called *Aurora Australis.* It generally occurs a few hours after a solar flare, when charged particles, emitted from the sun, arrive at the earth, and are accelerated in the vicinity of the the geomagnetic poles.

auroral absorption Radio wave absorption by an aurora.

auroral flutter Rapid fading of a signal at high or very high frequencies, so-called because it often imparts a fluttering quality to the signal that is caused by phase distortion and Doppler shift when the waves are reflected from the aurora.

auroral interference **1.** Interference to high-frequency radio propagation and also occasionally to medium-frequency and low-frequency propagation, caused by the activity of the aurora. **2.** Auroral flutter on a signal.

auroral opening A condition in which radio communication becomes possible via AURORAL PROPAGATION. It can occur when communication between two points is normally impossible at a certain frequency. Auroral openings allow long-distance communication well into the very-high-frequency (VHF) spectrum.

auroral propagation Reflection of radio signals from aurora that occur during geomagnetic storms. Theoretically, auroral propagation is possible when the aurora are active, between any two points on the earth's surface from which the same part of the aurora lie on a line of sight. This type of propagation seldom occurs when one end of the

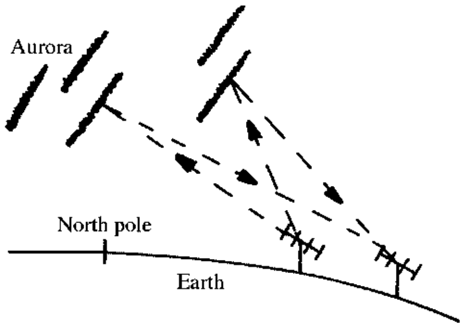

auroral propagation

circuit is at a latitude less than 35 degrees north or south of the equator. Auroral propagation can take place at frequencies well above 30 MHz. It is characterized by deep, rapid fading and random phase modulation of reflected signals.

auroral reflection The return of electromagnetic waves that have been beamed toward an aurora. Most often observed between 15 MHz and 150 MHz.

authorized access switch A device that disables a security system in a defined region or volume so that authorized personnel can enter without triggering an alarm condition.

authorized channel The carrier frequency or band assigned to a transmitting station by a licensing authority. Also see RADIO SPECTRUM.

autoalarm A device that is actuated from a received signal to alert a radio or computer network operator to the existence of a message.

autobaud **1.** In digital communications, a function that allows the equipment to adjust itself to the speed of the terminal. **2.** Any digital communications equipment capable of automatically adjusting to the speed of the terminal.

autocondensation The application of radio-frequency (RF) energy to the human body for medical purposes. The living organs serve as an impedance or load, across which the RF is applied.

autoconduction The application of radio-frequency (RF) currents into the body, by placing the living organ inside a coil and supplying the coil with RF. Used for medical purposes.

autocorrelation function A measure of the similarity between delayed and undelayed versions of a signal, expressed as a delay function.

autodyne reception Radio reception of cw signals by means of an oscillating detector. This is in contrast to heterodyne reception, in which a local oscillator (LO) generates an audio beat note with the cw signal in a separate detector.

autoionization A two-phase process of atomic ionization. The atom is excited beyond its ionization potential, and then it is allowed to deionize, causing the emission of an electron. The result is a positively charged atom (positive ion).

automated communications The transfer of data without the use of operating personnel; generally done with computers connected to communications equipment.

automated guided vehicle Abbreviation, AGV. A robot cart that runs without a driver. It uses an electric engine and is guided by the magnetic field produced by a current-carrying wire embedded in the floor or pavement. Alternatively, the robot can run on a track.

automated home A residence in which many, or most, of the routine chores are done by computers and/or robots. Examples of such tasks are dishwashing, doing the laundry, mowing the lawn, blowing snow, and vacuuming the floors.

automated integrated manufacturing system Acronym, AIMS. An assembly line or factory that uses robots, often controlled by one or more computers, to perform specific tasks that result in the production of various hardware items.

automatic Self-regulating, independent of human intervention. Some periodic adjustment might be needed.

automatic base bias A method of obtaining base bias in a bipolar transistor, where a resistor develops a voltage drop because of the current flowing through it. The resistor is usually placed in the emitter circuit, raising the emitter above ground potential.

automatic bass compensation Also called *bass boost.* In audio high-fidelity systems, a resistor-capacitor (RC) network that increases the relative amplitude of the bass at low volume levels. This compensates for the ear's inefficiency at low frequencies. The function can be automatically actuated by the setting of the volume control, or it can be switched manually on and off.

automatic bias In an amplifier, dc base/gate/grid bias obtained from the voltage drop produced by collector/drain/plate current flowing through a resistor common to the input and output. This resistor is usually shunted by a capacitor and placed in the emitter/source/cathode circuit.

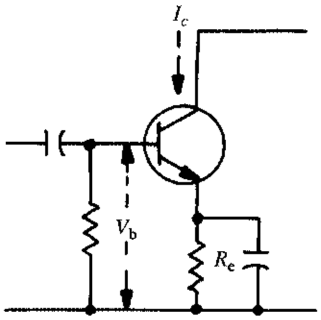

automatic bias

automatic brightness control A circuit that uses the same principles used in AUTOMATIC GAIN CONTROL (AGC) to hold steady the average brightness of a television (TV) picture.

automatic carriage Typewriters, automatic key punches, and other devices that can control automatically the spacing and feeding of paper, cards, and forms.

automatic check **1.** In a digital computer, the automatic inspection of operation and performance by a self-contained subsystem. **2.** The circuit or device for performing this inspection.

automatic chrominance control In a color television (TV) receiver, a subcircuit that controls the gain of the chrominance bandpass amplifier by automatically adjusting its bias.

automatic circuit breaker Any device that opens a circuit automatically when the flow of current becomes excessive. The breaker generally resets automatically after a specified length of time, or after power has been temporarily removed from the circuit.

automatic coding The use of a computer to determine the steps for solving a problem, before the actual program for the problem is written. This can help software engineers develop long and/or complex computer programs.

automatic contrast control A circuit that automatically adjusts the gain of the video IF and RF stages of a television (TV) receiver to preserve good picture contrast.

automatic controller In servo systems, any of several circuits or devices that samples a variable signal, compares it with a standard (reference) signal, and delivers a control or correction signal to an actuator.

automatic crossover **1.** Current limiting in a power supply. **2.** A device that switches a circuit from one operating mode to another automatically when conditions change in a predetermined manner.

automatic current limiter A circuit or device for holding the output current of a power supply to a safe value during overload.

automatic current regulator A circuit or device that holds the output current of a generator or power supply to a predetermined value, in spite of wide variations in load resistance.

automatic cutout A device that shuts down a circuit or system when the safe limits of operation are exceeded. A circuit breaker is an example of such a device, as is a thermostat in a power amplifier.

automatic data processing Abbreviation, ADP. The use of computers and accessories for calculations and tabulations using data gathered automatically by the system.

automatic degausser A system for automatically demagnetizing the picture tube in a color television (TV) receiver.

automatic dialing unit Abbreviated, ADU. A device that automatically generates dialing digits. Many telephone sets have these devices, some of which can be programmed for several different telephone numbers, including country codes and area codes.

automatic dictionary A computer system component that substitutes codes for words and phrases in information retrieval systems. In language-translating systems, it provides word-for-word substitutions.

automatic direction finder Abbreviated ADF. A specialized receiver/antenna combination for automatically showing the direction from which a signal arrives.

automatic error correction A technique of correcting transmission errors using error-detecting and error-correcting codes and, usually, automatic retransmission.

automatic exchange A transmission exchange in which interterminal communications are accomplished without operators.

automatic focusing A method of focusing a picture tube automatically, in which a resistor connects the focusing anode to the cathode; thus, no external focusing voltage is necessary.

automatic frequency control Abbreviation, AFC. A system that keeps a circuit automatically tuned to a desired signal frequency. A detector (such as a discriminator) operated from the tuned circuit delivers a dc output voltage only when the circuit is operating above or below the signal frequency; otherwise, it has zero dc output. The dc output, when present, alters the capacitance of a varactor in the tuned circuit to retune the stage to the desired frequency.

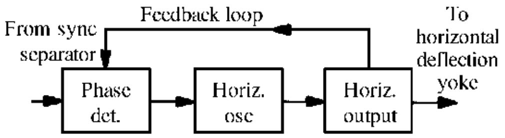

automatic frequency control (used in typical TV receiver)

automatic gain control Abbreviated AGC. A system that holds the output of a receiver or amplifier substantially constant despite input-signal amplitude fluctuations. A rectifier samples the ac signal output and delivers a dc signal proportional to that output. The dc signal is filtered, and the smoothed-out voltage is applied in correct polarity as bias to one or more preceding stages to reduce their gain. The stronger the signal entering the system, the greater the reduction in gain. As a result, weak signals are amplified much more than strong ones. Various forms of this scheme are used in many types of amplifiers and communications systems.

automatic gate bias A method of obtaining gate bias in a FET, where a resistor develops a voltage drop because of the current flowing through it. The resistor is usually placed in the source circuit, raising the source above ground potential.

automatic height control In a television (TV) receiver, a system that automatically maintains the height of the picture, despite signal-amplitude fluctuations, power-line voltage changes, and gain variations.

automatic intercept A telephone answering machine. It allows messages to be recorded when the subscriber is not able to answer the telephone.

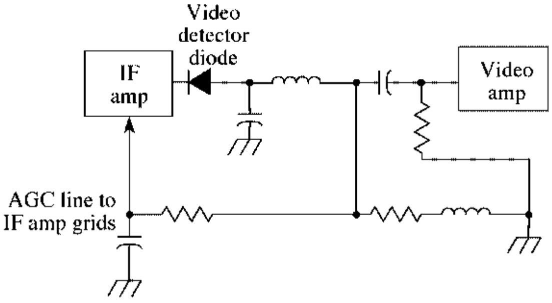

**automatic gain control
(used in typical TV receiver)**

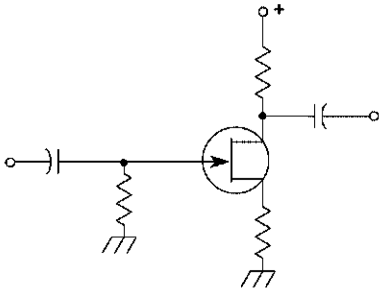

automatic gate bias

automatic interrupt A program interruption caused by hardware or software acting in response to some event independent of the program.

automatic level compensation See AUTOMATIC GAIN CONTROL.

automatic level control Abbreviation, ALC. **1.** A circuit that adjusts the input gain of a magnetic-tape recording device to compensate for changes in the loudness of the sound reaching the microphone. **2.** A form of AUTOMATIC GAIN CONTROL used in single-sideband (SSB) radio transmitters to maintain linearity while increasing the level of the average power relative to the peak power.

automatic line feed In the digital transmission of printed matter, the automatic insertion of a line feed (LF) character immediately following every carriage return (CR) character.

automatic modulation control Abbreviation, AMC. In a frequency-modulated (FM) radio transmitter, a form of AUTOMATIC GAIN CONTROL that regulates the gain of the audio amplifiers to compensate for fluctuating audio input amplitude. This prevents overdeviation while optimizing signal intelligibility.

automatic noise limiter Abbreviation, ANL. Any of several circuits for clipping noise peaks exceeding a predetermined maximum received-signal amplitude.

automatic phase control In a color television (TV) receiver, a circuit that synchronizes the burst signal with the 3.58-MHz color oscillator.

automatic pilot An electronic device, often computer-controlled, that automatically keeps a ship, airplane, or space vehicle on course.

automatic polarity In an electronic metering device, a means of automatically switching the input polarity of the instrument when the input signal polarity is shifted. Also called *bipolar operation.*

automatic programming See AUTOMATIC CODING.

automatic protective device A circuit or device (such as a fuse, circuit breaker, limiter, or regulator) that protects another circuit or device by automatically removing, reducing, or increasing the current or voltage during overload or underload.

automatic radio compass See AUTOMATIC DIRECTION FINDER.

automatic ranging In a metering device, the automatic adjustment or optimization of the full-scale range to compensate for large changes in the input parameter.

automatic regulation 1. *Voltage regulation.* In a power supply, the automatic holding of the output voltage to a constant value, despite variations in the input voltage or load resistance. **2.** *Current regulation.* In a power supply, the automatic holding of the output current to a constant value, despite variations in the input voltage or load resistance.

automatic relay The relaying of messages automatically from one station to another via intermediate points, without the need for human operators.

automatic repeater station A station that receives signals and simultaneously retransmits them, usually on a different frequency.

automatic reset 1. The self-actuated restoration of a circuit or device to a given state (e.g., the state of rest). **2.** A circuit or device that restores another circuit or device to a given state.

automatic scanning 1. The automatic (usually repetitive) tuning or adjustment of a circuit or system throughout a given frequency range. In a radio receiver, the system can be programmed to pause or stop at occupied channels, passing over vacant ones; or it can be programmed to pause or stop at vacant channels, passing over occupied ones. **2.** The repetitive sweep of a cathode-ray-tube (CRT) electron beam.

automatic scanning receiver Also called PANORAMIC RECEIVER. A radio receiver that is automatically tuned (usually repetitively) over a frequency band. Such a receiver either homes in on a signal when one is found, or displays on a cathode-ray-tube (CRT) screen the distribution of signals in the band.

automatic secure voice communications A wideband and narrowband voice-digitizing application to a security network that provides encoded voice communications.

automatic send/receive set A teletypewriter or terminal that is capable of receiving and transmitting.

automatic sensitivity control **1.** A self-actuating circuit using principles similar to those used in AUTOMATIC GAIN CONTROL. It varies the sensitivity of the radio-frequency (RF) and intermediate-frequency (IF) sections of a receiver in inverse proportion to the strength of a received signal. **2.** In a bridge null detector, a circuit similar to the one described in **1**, which operates ahead of the detector, varying the sensitivity of the latter automatically.

automatic sequencing The ability of a digital computer to perform successive operations without additional instructions from the operator.

automatic short-circuiter A device that automatically short-circuits the commutator bias in some single-phase commutator motors.

automatic short-circuit protection A circuit that allows the output of a power supply to be short-circuited without damage to the components in the supply. It usually consists of a current-limiting device.

automatic shutoff A switching arrangement that automatically shuts off a device or circuit under certain specified conditions.

automatic switch center A telephone-switching network that routes calls to their destinations without the need for a human operator.

automatic target control For a vidicon television camera tube, a circuit that automatically adjusts the target voltage in proportion to brightness of the scene.

automatic telegraph reception Telegraph reception providing a direct printout of the received information, without intervention by an operator.

automatic telegraph transmission Telegraph transmission originating from tapes, disks, or other records, rather than from a hand-operated key.

automatic telegraphy Communications that utilize automatic telegraph transmission and reception.

automatic time switch A time-dependent circuit or device that opens or closes another circuit at the end of a predetermined time interval.

automatic tracking A method of keeping a radar beam automatically fixed on a target.

automatic trip A circuit breaker that automatically opens a circuit.

automatic tuning A process whereby a circuit tunes itself to a predetermined frequency upon receiving a command signal.

automatic voltage regulator A circuit that keeps the output of a power supply constant, despite the load resistance or input voltage to the supply.

automatic volume control Abbreviated AVC. The use of AUTOMATIC GAIN CONTROL in an audio amplifier system.

automatic zero In an electronic meter, a means of automatically setting the indicator to zero in the absence of an input signal.

automation **1.** The control of machines or processes by self-correcting electronic systems. See ROBOT. **2.** The use of robots and/or computers, rather than human beings, to perform repetitive tasks. **3.** The use of robots and/or computers to assist human beings in industrial, office, governmental, and educational work.

automaton A simple robot that performs a task or set of tasks without artificial intelligence (AI). These machines have existed for decades. Compare ANDROID.

automonitor In digital computer operations, to require the machine to supply a record of its information-handling operations. Also, the program for such instructions.

automotive battery A set of four or eight rechargeable lead-acid cells connected in series and housed in a common enclosure. The electrolyte is a free-flowing liquid acid. Provides approximately 6 volts (four cells) or 12 volts (eight cells) under no-load conditions when fully charged. The high mass results in large energy-storage capacity. These batteries must be handled with care and always kept in an upright position to prevent spillage of the acid. See LEAD-ACID BATTERY, LEAD-ACID CELL.

autonomous robot A self-contained robot with an independent computerized control system. It moves under its own power, usually by rolling on wheels or a track drive. Compare INSECT ROBOT.

autopatch A remotely controllable device that interconnects a radio-communications system into a telephone network.

autopilot A self-correcting control and guidance device for the automatic management of an aircraft or missile.

autoranging See AUTOMATIC RANGING.

autosyn A device or system that operates on the principle of the synchronous ac motor, in which the position of the rotor in one motor (the transmitter) is assumed by the rotor in a distant motor (the receiver) to which the first is connected.

auto tracking A method of controlling the output voltages of many different power supplies simultaneously.

autotransducter A type of magnetic amplifier whose power windings serve also as control windings.

autotransformer A single-winding transformer in which the primary coil is a fraction of the entire winding for voltage step-up, or the secondary coil is a fraction of the entire winding for voltage step-down.

auxiliary circuit A circuit that is supplementary to the main system.

auxiliary contacts In switches and relays, contacts that are supplementary to the main contacts and are usually actuated with them.

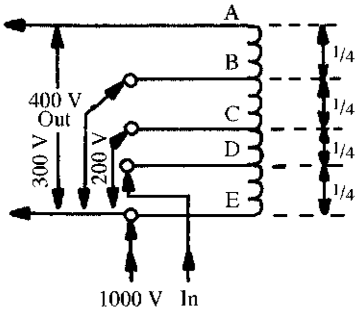

autotransformer

auxiliary equipment **1.** Also known as *peripherals.* An apparatus not directly governed by the central processing unit of a digital computer, such as a printer or personal robot. **2.** Peripheral equipment in any system. **3.** Backup equipment.

auxiliary memory In a digital computer, a unit that is supplementary to the main memory, which it augments.

auxiliary receiver Also called *standby receiver.* In a radio communications system, a receiver that is available for use if the main receiver fails.

auxiliary relay **1.** A standby relay. **2.** A relay whose operation supports that of another relay. **3.** A relay that is actuated by the operation of another relay.

auxiliary switch **1.** A standby switch. **2.** A switch wired in series or parallel with another switch. **3.** A switch that is operated by another switch.

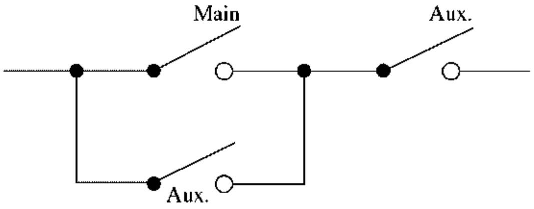

auxiliary switch, 2

auxiliary transmitter Also called *standby transmitter.* In a radio communications system, a transmitter that is available for use if the main transmitter fails.

a/v Abbreviation of AUDIO-VISUAL.

aV Abbreviation of *attovolt.*

availability The proportion of time during which an apparatus is operating correctly. It is usually given as a percentage.

available conversion gain The ratio of the input power to the output power of a transducer or converter. It is generally given in decibels.

available gain The ratio P_o/P_i, where P_i is the available power at the input of a circuit and P_o is the available power at the output.

available line The percentage of the length of a facsimile scanning line that is usable for picture signals.

available power The mean square of the open-circuit terminal voltage of a linear source divided by four, times the resistive component of the source impedance. The available power is the maximum power delivered to a load impedance, equal to the conjugate of the internal impedance of the power source.

available power gain In a power transistor, the ratio of available transistor output power to the power available from the generator. It depends on the generator resistance, but not on the transistor load resistance.

available signal-to-noise ratio The ratio P_s/P_n, where P_s is the available signal power at a given point in a system and P_n is the available random-noise power at that point.

available time **1.** The time during which a computer is available and ready for immediate use. **2.** The amount of time a computer is available to an individual.

avalanche The phenomenon in semiconductors operated at high reverse bias voltage, whereby carriers acquire sufficient energy to produce new electron-hole pairs as they collide with atoms. The action causes the reverse current to increase sharply.

avalanche breakdown In a semiconductor P-N junction, a condition that occurs when the reverse bias voltage exceeds a certain value. If the electric field in the vicinity of the junction becomes strong enough, charge carriers are dislodged from the atoms and the carriers (electrons and holes) flow freely across the P-N junction in the opposite direction from normal. The mini-

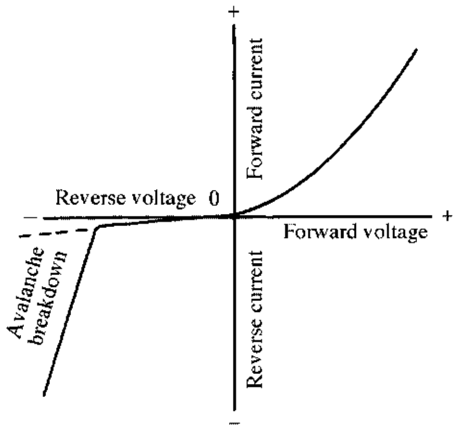

avalanche breakdown

mum reverse-bias voltage required to cause this phenomenon varies among different kinds of diodes. Some diodes are manufactured to have precise avalanche voltages. See ZENER DIODE.

avalanche conduction In a semiconductor junction, the enhanced reverse-bias conduction caused by a condition of AVALANCHE.

avalanche current The high current that flows through a semiconductor junction when AVALANCHE occurs.

avalanche diode See ZENER DIODE.

avalanche impedance The reduced impedance of a diode during avalanche.

avalanche noise Electrical noise generated in a junction diode operated at the point at which avalanche just begins.

avalanche transistor A transistor that operates at a high value of reverse-bias voltage, causing the pn junction between the emitter and base to conduct because of avalanche breakdown.

avalanche voltage In a semiconductor P-N junction, the minimum applied reverse-bias voltage that produces AVALANCHE BREAKDOWN.

AVC Abbreviation of *automatic volume control.*

avdp Abbreviation for *Avoirdupois*, a weight-measurement scheme that is used in English-speaking countries and is based on the pound.

average absolute pulse amplitude The average (disregarding algebraic sign) of the absolute amplitudes of a pulse, taken over the duration of the pulse.

average brightness The average brilliance of a television (TV) picture, cathode-ray-tube (CRT) computer display, or oscilloscope image.

average calculating operation The operating time considered typical for a computer calculation (i.e., one that is longer than an addition and shorter than a multiplication); it is frequently taken as the average of nine additions and one multiplication.

average current Abbreviation, I_{avg}. The average value of alternating current flowing in a circuit. Taking polarity into account, this value is zero for a pure sine wave. For other waveforms, it can vary. When polarity is not considered, the sine-wave value of I_{avg} is equal to 0.637 times I_{pk}, the peak value of current; $I_{avg} = 0.637\ I_{pk}$.

average life See MEAN LIFE.

average noise figure The ratio of the total noise output from a circuit to the thermal noise output at 290 degrees Kelvin. It is usually expressed in decibels, with the noise taken at all frequencies.

average power The average value of power in an ac circuit. In a resistive circuit, it is the square of the effective (rms) current times the resistance; $P_{avg} = (I_{rms})^2R$ (for sine waves).

average pulse amplitude Also called *effective pulse amplitude.* The value obtained by integrating the pulse amplitude, with respect to time, from the start of the pulse to its end, then dividing this integral by the pulse duration.

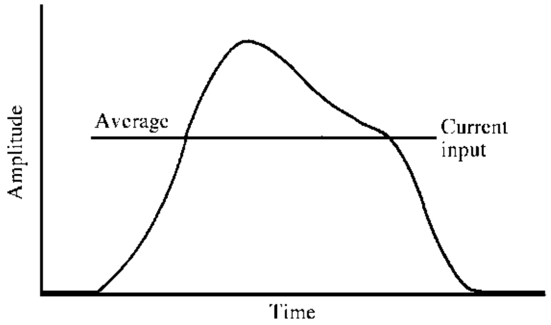

average pulse amplitude

average rectified current Abbreviation, I_{avg}. The average value of rectifier output current before filtering. For a full-wave rectifier with a sine wave input and a resistive load, I_{avg} is equal to the maximum current I_m multiplied by 0.637.

average rectified voltage Abbreviation, E_{avg}. The average value of rectifier output voltage before filtering. For a full-wave rectifier with a sine-wave input and a resistive load, E_{avg} is maximum voltage E_m multiplied by 0.637.

average value **1.** The *arithmetic mean* of two or more quantities. **2.** The *geometric mean* of two or more quantities. **3.** The *harmonic mean* of two or more quantities. **4.** In ac operation, the average current, voltage, or power.

average voltage Abbreviation, E_{avg}. The average value of ac voltage in a circuit. Taking polarity into account, this value is zero for a pure sine wave. For other waveforms, it can vary. When polarity is not considered, the sine-wave value of E_{avg} is equal to 0.637 times E_{pk}, the peak value of voltage.

avg Abbreviation of *average.*

aviation channels Frequency channels assigned to the AVIATION SERVICES.

Aviation services The radio-communication services used by aeronautical-mobile and radio navigation personnel.

avigation Acronym for *aviation navigation.* Aircraft navigation by means of electronic equipment.

avionics Acronym for *aviation electronics.* The design, production, and application of electronic devices and systems for use in aviation, navigation, and astronautics.

Avogadro's constant (Amedeo Avogadro, 1776–1856.) Symbol, NA. The number of molecules in a kilogram-molecular weight of any substance; NA equals 6.025×10^{26} (kg-mole)$^{-1}$.

A voltage The filament voltage in a vacuum-tube circuit.

aW Abbreviation of *attowatt.*

AWG Abbreviation of AMERICAN WIRE GAUGE.

AX.25 A signal format used in some digital communications systems, notably amateur *packet radio.*

axial leads The centrally located leads emanating from the ends of cylindrical components, such as resistors and diodes.

axial ratio The ratio of the minor to major axes of a waveguide's polarization ellipse.

axis **1.** A coordinate in a graphical presentation or display (e.g., horizontal and vertical axes in a rectangular coordinate system). **2.** The real or imaginary straight line around which a body rotates, or the line that passes through the center of a symmetrical arrangement (line of symmetry).

axis of abscissas The horizontal axis (x-axis) of a rectangular-coordinate graph or screen. Compare AXIS OF ORDINATES.

axis of imaginaries The vertical axis of the complex plane in which rectangular vectors lie. Compare AXIS OF REALS.

axis of ordinates The vertical (y-axis) of a rectangular-coordinate graph or screen. Compare AXIS OF ABSCISSAS.

axis of reals The horizontal axis of the complex plane in which rectangular vectors lie. Compare AXIS OF IMAGINARIES.

Ayrton-Mather galvanometer shunt A step-adjustable universal shunt resistor for varying the sensitivity of a galvanometer. It has the virtue of keeping the galvanometer critically damped. The shunt is also useful in multirange milliammeters, microammeters, and ammeters. The sensitive meter movement is never without a shunting resistor during range switching.

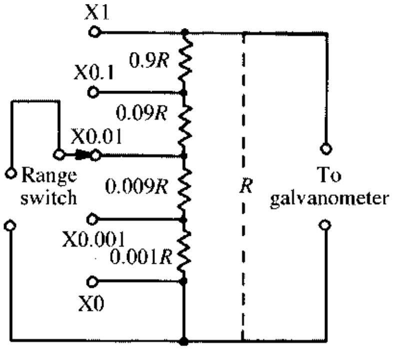

Ayrton-Mather galvanometer shunt

Ayrton-Perry winding A noninductive winding comprising two inductors conducting current in opposite directions; the opposing flow cancels the magnetic field.

azel display A plan-position display that incorporates two different radar traces on a single cathode-ray tube (CRT), one giving bearing, the other elevation.

azimuth Also called *compass direction.* Angular measurement in the horizontal plane, clockwise from north. It is important in radio and television communications, navigation, direction finding, land surveying, and radar.

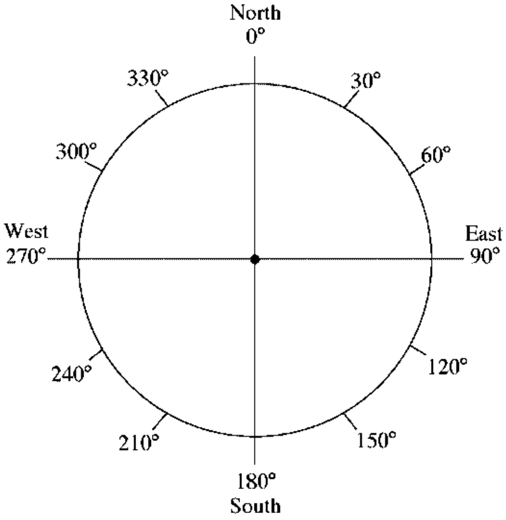

azimuth

azimuth alignment In a tape recorder, the alignment of record and playback head gaps so that their centerlines are parallel.

azimuth blanking In a radar system, blacking-out of the image as the antenna sweeps across a specified range of azimuth angles. Effectively eliminates nuisance echoes from stationary, permanent objects (such as tall buildings or communications towers).

azimuth resolution In a radar system, the minimum azimuth separation of two targets whose range (distance from the station) are equal that is required for the system to show two echoes, rather than one. It is generally measured in degrees.

azusa An electronic tracking system, in which a single station provides slant range and two direction cosines for a distant airborne object. This accurately defines the coordinates of the distant object in three-dimensional space.

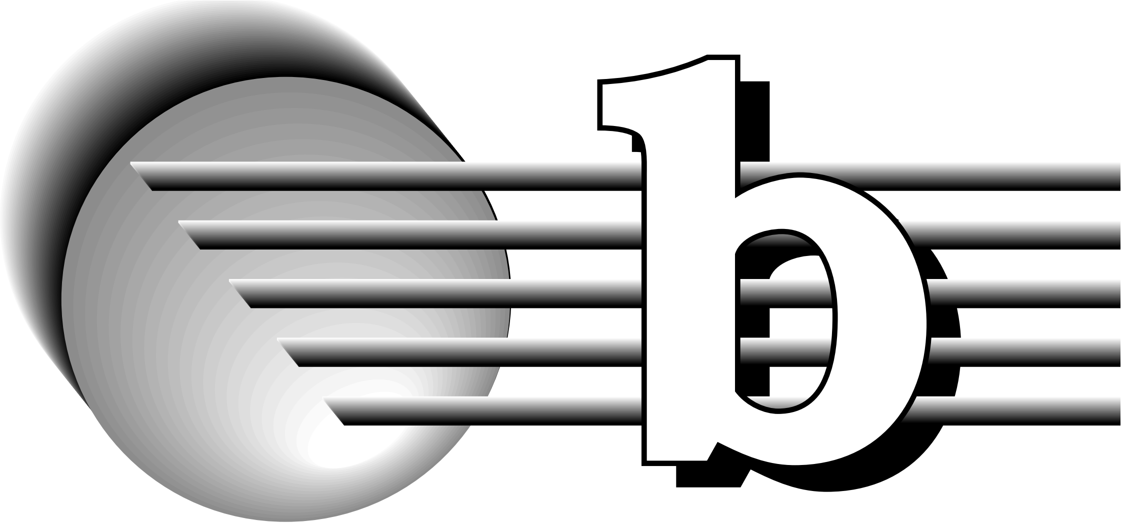

B **1.** Symbol for SUSCEPTANCE. **2.** Symbol for FLUX DENSITY. **3.** Abbreviation of BATTERY. **4.** Symbol for BORON. **5.** Symbol for base of transistor (see BASE, **1**). **6.** Abbreviation of BASS. **7.** Abbreviation of BEL. **8.** Anode voltage or main operating voltage in any circuit (when used with sign). Also see B VOLTAGE.

b **1.** Symbol for SUSCEPTANCE. **2.** Symbol for base of transistor (see BASE, **1**). **3.** Abbreviation of BASS. **4.** Symbol for BARN.

B&S See AMERICAN WIRE GAUGE.

B5-cut crystal A piezoelectric plate cut from a quartz crystal in such a way that the face of the plate is at an angle, with respect to the z-axis of the crystal. This type of crystal has good frequency stability under conditions of changing temperature.

BA Abbreviation of BATTERY. Also see B and BAT.

Ba Symbol for BARIUM.

babbit A relatively soft, tin-base alloy of various compositions. One composition contains 7.4% antimony, 3.7% copper, and 88.9% tin.

babble Interference caused by *crosstalk* from a number of channels.

babble signal A jamming signal containing babble components. See BABBLE and JAMMING.

BABS Abbreviation of BLIND-APPROACH BEACON SYSTEM.

baby monitor A short-range radio transmitter and receiver that can be used to listen at a distance to the sounds in an infant's room. The transmitter contains a sensitive microphone, a whip antenna, and a power supply. The unit can be placed on a table or desk, or even on the floor near the baby's crib. The receiver is similar to a handheld "walkie-talkie." It is battery-powered and can be carried around. It has an inductively loaded, short "rubber duckie" antenna similar to the antennas on cordless telephone sets. The receiver can pick up signals from the transmitter at distances of up to about 200 feet. The radio-frequency signals pass easily through walls, ceilings, and floors.

back bias **1.** A feedback signal (negative or positive). **2.** Reverse bias (also see BIAS). **3.** A reverse bias voltage, obtained from a voltage divider connected between a voltage source and ground.

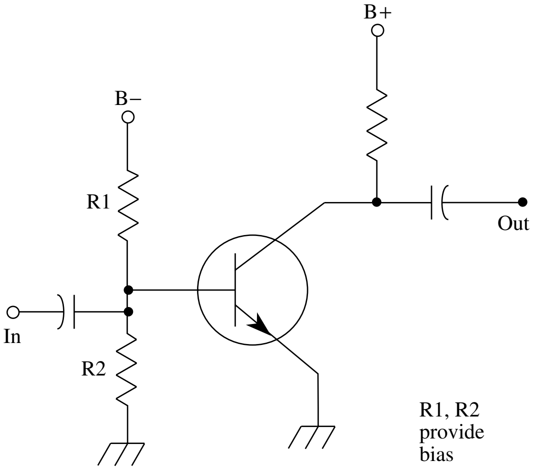

back bias, 3

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

backbone A form of transmission line with capacitive connections between the generator and the load.

back conduction Conduction of current in the reverse direction, as across a semiconductor junction that is reverse-biased.

back contact A contact that closes a circuit when a relay, switch, or jack is in its normal rest position.

back current Symbol, I_b. The normally small current flowing through a reverse-biased pn semiconductor junction. Also called *reverse current* and *inverse current.* Compare FORWARD CURRENT.

back diode A semiconductor diode that is normally back-biased (reverse-biased).

back echo An echo resulting from the rear lobe of an antenna radiation pattern.

back emf See BACK VOLTAGE.

Back-Goudsmit effect See ZEEMAN EFFECT.

background **1.** The context or supporting area of a picture (e.g., the background of a television picture). **2.** Background noise.

background control In a color television receiver, a potentiometer used to set the dc level of the color signal at one input of the three-gun picture tube.

background count Residual response of a radioactivity counter in an environment as free as practicable of radioactivity. This background is caused largely by cosmic rays and inherent radioactivity of surrounding buildings and other bodies.

background job A low-priority, relatively long-running computer program that can be interrupted so that a higher-priority program can be run.

background noise Electrical noise inherent to a particular circuit, system, or device that remains when no other signal is present.

background processing In a computer, the running of programs having low priority.

background radiation Nuclear radiation from materials in the environment. Also see BACKGROUND COUNT.

background response The response of a radiation detector to background radiation.

backing store In a computer, a device that stores large amounts of information. In most small computers, this is done via MAGNETIC DISK and/or MAGNETIC TAPE. A backing store can also be an optical storage medium, such as CD-ROM (compact disk, read-only memory).

backlash **1.** Slack or lag in action of moving parts. Example: delay between initial application of a force (such as that required to turn a knob) and movement of a part or device (e.g., a potentiometer or variable capacitor). **2.** On a mechanical analog tuning dial, an arc within which slack or lag is discernible.

backloaded horn A loudspeaker enclosure in which the front of the speaker cone feeds sound directly into the listening area, and the rear of the cone feeds sound into the same area through a folded horn.

backloading In a cascaded series of amplifiers, the tendency of loading effects to be passed to earlier stages. A change in the output impedance of a final amplifier circuit, for example, could also result in a change in the output impedance of the driver circuit, and perhaps even in a change in the output impedance of the predriver.

back lobe In the pattern of a directional antenna, the lobe directly opposite the major lobe, representing the radiation or response in or from a direction 180 degrees from that in which the gain is greatest.

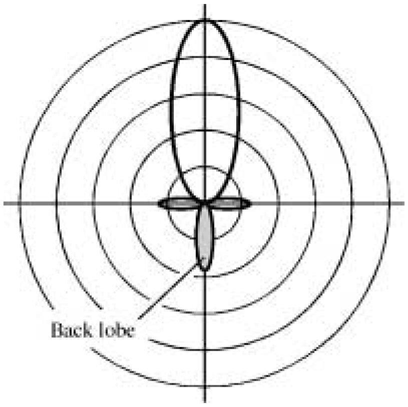

back lobe

backplate A flat electrode in a television (TV) camera tube that receives the stored-charge image via capacitive coupling.

back porch In a television (TV) horizontal sync pulse, the time interval between the end of the rise of the blanking pedestal and the beginning of the rise of the sync pulse. That portion of the flat top of the blanking pedestal behind the sync pulse. Compare FRONT PORCH.

back-porch effect In transistor operation, the continuation of collector-current flow for a short time after the input signal has fallen to zero.

back-porch tilt The departure of the top edge of a back porch from true horizontal.

back pressure sensor A device that detects and measures the torque that a motor is applying, and produces a signal whose amplitude is proportional to the torque. This signal can be used for various purposes. In a robotic device, for example, the sensor output can be fed back to the motor control to limit the applied force.

back resistance Symbol, R_{b}. The resistance of a reverse-biased pn semiconductor junction. Also called REVERSE RESISTANCE.

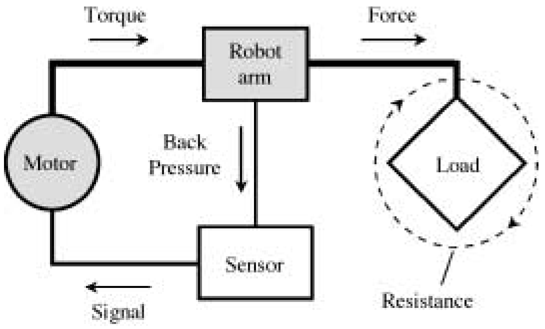

back pressure sensor

back scatter Scattering of a wave back toward a radio transmitter from points beyond the skip zone. This phenomenon is caused by ionospheric reflection. Compare FORWARD SCATTER.

backstop A contact or barrier (such as a screw or post) that serves to limit the BACKSWING of the armature of a relay.

backswing **1.** The tendency of a pulse to overshoot, or reverse direction after completion. Backswing is measured in terms of the overshoot amplitude as a percentage of the maximum amplitude of the pulse. **2.** The extent to which a relay armature moves back from a contact when the relay contacts are open.

back-to-back connection The connection of diodes or rectifiers in reverse parallel (i.e., the anode of one to the cathode of the other) across a signal line to pass both half cycles of ac in certain control circuits.

back-to-back sawtooth A symmetrical sawtooth wave in which the rise slope is equal to the fall slope. Also called *triangular wave* and *pyramidal wave.*

backup **1.** An element, such as a circuit component, that is used to replace a main component, in case of main-component failure. **2.** Any process or scheme that serves to maintain operation of a system in case of main-component failure. **3.** A battery that maintains volatile memory data stored in one or more integrated circuits. **4.** A computer file, or set of files, stored in a nonvolatile medium, such as diskettes or magnetic tape, to prevent catastrophic data loss in the event of hard-disk failure. **5.** A battery or alternative power source that keeps an alarm system operational in the event of a utility power failure.

backup battery **1.** In a computer or microcomputer-controlled electronic device, a source of voltage to preserve volatile memory data if the power is removed. **2.** A battery used for powering a system in the event that the main power source should fail.

backup facility In an electrical or communications system, a facility that is intended for use when the primary, or main, facility is not operational.

back voltage **1.** Voltage induced in an inductor by the flow of current through the inductor, so called because its polarity is opposite to that of the applied voltage. Also called *counter emf.* **2.** A voltage used to obtain bucking action (e.g., the voltage used to zero the meter in an electronic voltmeter circuit). **3.** Reverse voltage applied to a semiconductor junction.

backwall In a pot core, the plate or disk that connects the sleeve and center post to close the magnetic circuit.

backward diode A semiconductor diode manufactured in such a way that its high-current flow occurs when the junction is reverse biased. Such a diode is also a negative-resistance device.

backward-wave oscillator Abbreviation, BWO. A microwave oscillator tube similar to the *traveling-wave tube.* Like the traveling-wave tube, the BWO contains a helical transmission line. In the electron beam, electron bunching results from interaction between the beam and the electromagnetic field, and reflection occurs at the collector. The wave moves backward from collector to cathode, and oscillation is sustained because the backward wave is in phase with the input. Output is taken from the cathode end of the helix.

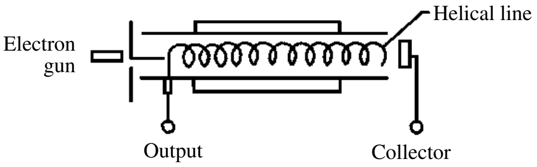

backward-wave oscillator

back wave The oscillator signal present in an amplifier-keyed, continuous-wave (CW), Morse-code transmitter. Normally, this signal is at the same frequency as the transmitter output, but is not sufficiently strong to be radiated over the air.

back-wave radiation The condition wherein a back wave is strong enough to be heard on a continuous-wave (CW) keyed signal at the receiving station. This results from ineffective amplifier keying.

baffle A board on which a loudspeaker is mounted to separate acoustic radiation from the back of the cone from radiation emanating from the front. The baffle improves bass response by increasing the wavelength (lowering the frequency) at which phase cancellation occurs.

baffle plate **1.** See BAFFLE. **2.** A metal plate mounted in a waveguide to reduce the cross-sectional area.

bail A wire loop or chain that holds one member of a two-member assembly to prevent loss (e.g., the short chain holding the dust cap of a jack).

Bakelite The trade name for a specialized plastic dielectric material. Its chemical composition is *phenol-formaldehyde* resin.

baker An obsolete phonetic alphabet code word for letter B. BRAVO is commonly used instead.

baking-out In the process of evacuating a system, the procedure of heating the system to a high temperature to drive out gases occluded in the glass and metal parts.

balance **1.** See BRIDGE. **2.** To null a bridge or similar circuit. **3.** To equalize loads, voltages, or signals between two circuits or components. **4.** In a high-fidelity stereo sound system, a control or set of controls that adjusts the relative loudness of the left and right channels. **5.** Alignment of a *balanced modulator* for minimum carrier output amplitude. **6.** A condition in which two branches of a circuit have identical impedances, relative to ground.

balance coil **1.** A type of autotransformer that enables a three-wire ac circuit to be supplied from a two-wire line. A series of taps around the center of the winding enables the circuit to be compensated for unequal loads. **2.** See BALANCING COIL.

balance control A variable component, such as a potentiometer or variable capacitor, that is used to balance bridges, null circuits, or loudspeakers.

balanced Having identical impedances, with respect to ground.

balanced amplifier Any amplifier with two branches that have identical impedances, with respect to ground. Usually, the two branches are in phase opposition (180 degrees out of phase).

balanced antenna An antenna system where two halves are exact replicas of each other, geometrically and electrically. Such an antenna normally must either be fed with a balanced transmission line or with a coaxial cable and balun.

balanced antenna system A balanced antenna, fed with a balanced transmission line, that has currents of equal magnitude in each side. An example is a half-wave dipole at uniform height above electrical ground, fed at the center with parallel-wire line. It is important that the transmission line runs away from the antenna at a right angle for at least ¼ wavelength, preferably ½ wavelength or more, to prevent line imbalance caused by currents induced from the radiated field.

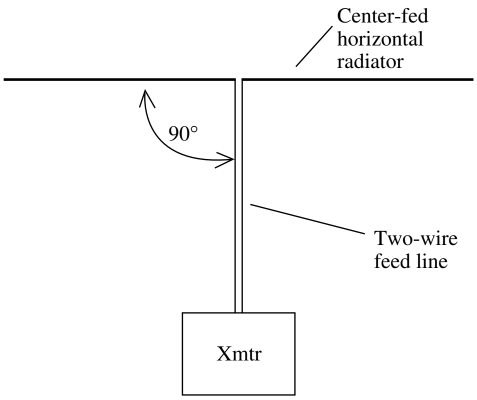

balanced antenna system

balanced bridge Any four-leg bridge circuit in which all legs are identical in all electrical respects.

balanced circuit **1.** A circuit that has its electrical midpoint grounded, as opposed to the *single-ended circuit*, which has one side grounded. **2.** A bridge circuit in the condition of null.

balanced converter See BALUN.

balanced currents Currents with the same value. In the two conductors of a balanced transmission line, these currents are equal in amplitude and opposite in phase at every point along the line.

balanced delta A set of coils or generators in a three-phase system, connected so that the currents in any two coils differ in phase by 120 degrees.

balanced detector A symmetrical demodulator, such as a full-wave diode detector or a discriminator.

balanced electronic voltmeter An electronic voltmeter circuit in which two matched transistors are connected in a four-arm bridge arrangement. The drift in one half of the circuit opposes that in the other half; the resulting drift of the zero point is virtually eliminated.

balanced filter A filter consisting of two identical sections, one in each branch of a balanced system, such as a parallel-wire transmission line.

balanced input An input circuit whose electrical midpoint is grounded. Compare SINGLE-ENDED INPUT.

balanced input transformer An input transformer in which the center tap of the primary winding is grounded.

balanced line A pair of parallel wires that possesses a uniform characteristic impedance. The two conductors are of the same material and have identical diameters. The distance between them is constant. In a balanced two-wire line, the currents in the two conductors are of equal amplitude and opposite phase.

balanced lines In high-fidelity audio systems, a cable that consists of two parallel conductors surrounded by a single braid. The parallel wires carry the audio-frequency (AF) signals, and the braid is grounded for shielding.

balanced loop antenna A loop antenna with a grounded electrical midpoint, determined by the junction of two identical series-connected capacitors shunting the loop.

balanced low-pass filter A low-pass filter used in a balanced system or balanced transmission line.

balanced method A system of instrumentation in which a zero-center scale is used. The reading can be either side of the zero reading.

balanced modulator A symmetrical modulator circuit using bipolar transistors, field-effect transistors, an integrated circuit, or diodes as principal components, that delivers an output signal containing the sidebands, but not the carrier. It is commonly used to generate a double-sideband (DSB) signal that can be filtered to obtain a single-sideband (SSB) signal.

balanced multivibrator A switching oscillator circuit in which the two halves are identical in configuration, and as nearly identical as practicable in performance.

balanced network Any network intended to be used with a balanced system or balanced transmission line. It is characterized by a pair of terminals, each of which shows the same impedance with respect to ground.

balanced oscillator A PUSH-PULL OSCILLATOR.

balanced output Output balanced against ground (e.g., where the electrical midpoint of the output circuit is grounded).

balanced output transformer **1.** A push-pull output transformer with a center-tapped primary winding. **2.** An output transformer with a grounded center tap on its secondary winding.

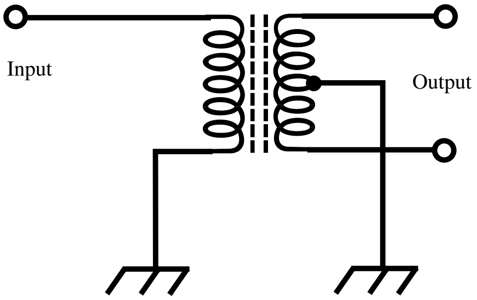

balanced output transformer, 2.

balanced probe A probe, such as one for an electronic voltmeter or oscilloscope, that has a balanced input and (usually) a single-ended output.

balanced-tee trap A wavetrap constructed in a T configuration, with a resonant section in each conductor of a balanced transmission line.

balanced telephone line A telephone transmission line that has two sides, similar to a balanced radio-frequency transmission line. Either side has the same impedance, with respect to ground.

balanced termination A load device (or the practice of using such a device) in which the sections provide identical termination for each of the sections or conductors of a balanced system, such as a balanced line.

balanced-to-unbalanced transformer See BALUN.

balanced transmission line See BALANCED LINE.

balanced varactor tuning A two-varactor, back-to-back circuit for adjusting the value of a capacitor using an applied dc voltage. This arrangement has an advantage over a single-varactor (unbalanced) circuit, because high-tuned-circuit *Q* is maintained and harmonic generation is reduced.

balanced voltages In any symmetrical system, such as a balanced line or push-pull circuit, two or more input or output voltages that are adjusted to have the same amplitude and (usually) opposite phase.

balanced-wire circuit A circuit or conductor system with identical halves that are symmetrical, with respect to ground and to other conductors.

balancing circuit See BUCKING CIRCUIT.

balancing coil In a receiver, a center-tapped antenna coil that is balanced to ground to eliminate MARCONI EFFECT.

ballast **1.** A component that is used to stabilize the current flow through, or operation of, a circuit, stage, or device. **2.** An iron-core choke connected in series with one of the electrodes in a fluorescent or other gas-discharge lamp.

ballast resistor **1.** A nonlinear inductive power resistor whose voltage-current (EI) characteristic is such that current through the resistor is independent of voltage over a useful range. This feature enables the ballast resistor to act as an automatic voltage regulator when it is simply connected in series with a power supply and load. **2.** A small (usually high-resistance) resistor operated in series with a glow lamp, such as a neon lamp, to prevent overload.

ballast transformer A misnomer often used in place of BALLAST, **2.**

ballistic galvanometer An undamped galvanometer that is used particularly to observe electric charges by noting the single throw resulting from the momentary flow of current through the galvanometer coil.

ballistics The electronics-supported science concerned with the motion of projectiles and similar bodies in air or space.

balloon antenna A vertical antenna consisting of a wire or wires held aloft by a captive balloon. Occasionally, used by radio amateurs and shortwave listeners at low and medium frequencies. A potentially dangerous antenna because of large static-electric buildup, a tendency to attract lightning, the possibility of its breaking loose, and the risk of accidental contact with high-voltage power lines.

balop Contraction of BALOPTICON.

balopticon An opaque-picture projecting system in which the picture is viewed by a television (TV) camera, such as a vidicon, and displayed by a picture tube. Also called *balop.*

balun A specialized impedance-matching radio-frequency (RF) transformer. It is a wideband device, usually providing a 1:1 or 1:4 impedance ratio and available in several different forms. It is so called because it has an unbalanced input suitable for coaxial transmission lines, and a balanced output suitable for dipole, Yagi, and quad antennas.

banana jack The female half of a two-part quick-connector combination. Splicing of a circuit is completed by inserting a BANANA PLUG into this jack.

banana plug The male half of a two-part quick-connector combination, with sides usually composed of flat springs that ensure contact with the female BANANA JACK into which it is inserted.

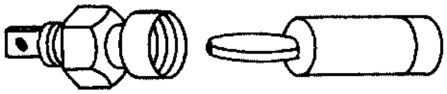

banana jack and plug

band **1.** A continuous range of radio or television communications frequencies or wavelengths, usually designated by the lowest and highest frequencies, or the approximate wavelength (e.g., the 20-meter amateur radio band). **2.** A set of discrete radio or television frequency channels within a specified range (e.g., the standard AM broadcast band). **3.** A range of wavelengths for infrared, visible, ultraviolet, X-ray, or gamma-ray energy. **4.** A range of energy levels. **5.** A colored stripe on a resistor or capacitor that forms part of the code that indicates component value and tolerance.

band center **1.** In a given radio or television communications band, the arithmetic mean of the lowest and highest frequencies. **2.** In a given band, the geometric mean of the longest and shortest wavelengths.

band-elimination filter See BAND-REJECTION FILTER.

band gap In any atom, the difference in electron energy between the conduction and valence bands.

bandpass **1.** The frequency limits between which a BANDPASS FILTER or BANDPASS AMPLIFIER transmits ac energy with negligible loss. **2.** The ability to allow passage of signals at a given frequency or band of frequencies while blocking other signals. Compare BANDSTOP.

bandpass amplifier An amplifier that is tuned to pass only those frequencies between preset limits.

bandpass coupling A coupling circuit with a flat-topped frequency response so that a band of frequencies, rather than a single frequency, is coupled into a succeeding circuit. Also see BANDPASS, **1.**

bandpass filter Any resonant circuit, or combination of resonant circuits, designed to discriminate against all frequencies except a specific frequency f_0, or a band of frequencies between two limiting frequencies f_0 and f_1. In a parallel inductance-capacitance (LC) circuit, the device exhibits high impedance at the desired frequency or frequencies and a low impedance at unwanted frequencies. In a series configuration, the filter has a low impedance at the desired frequency or frequencies and a high impedance at unwanted frequencies. Compare BAND-REJECTION FILTER, HIGH-PASS FILTER, LOW-PASS FILTER.

bandpass flatness The degree to which a bandpass device's attenuation-versus-frequency curve is a straight line with zero slope within the passband.

band pressure level The net acoustic pressure of a sound source within a specified frequency range (band).

band-rejection filter Also called a *band-stop filter*. Any resonant circuit, or combination of resonant circuits designed to discriminate against a specific frequency f_0, or a band of frequencies between two limiting frequencies f_0 and f_1. In a parallel inductance-capacitance (LC) circuit, the device exhibits high impedance at the desired frequencies, and a low impedance at the unwanted frequency or range of frequencies. In a series configuration, the filter has a low impedance at the desired frequencies and a high impedance at the unwanted frequency or range of frequencies. Compare BANDPASS FILTER, HIGH-PASS FILTER, LOW-PASS FILTER, NOTCH FILTER.

band selector Any switch or relay that facilitates switching the frequency of a radio transmitter, receiver, or transceiver among various bands.

bandset capacitor In some older communications receivers, a variable capacitor is used to preset the tuning range in each band to correspond to graduations on the tuning dial. This capacitor is a trimmer or padder operated in conjunction with the main tuning capacitor.

bandspreading In some older communications receivers, the process of widening the tuning range within a given frequency band to cover the entire dial. Otherwise, the band would occupy only a portion of the dial, and tuning would be difficult. It is usually accomplished with a BANDSPREAD TUNING CONTROL whose range is preset via the main tuning control and/or a BANDSET CAPACITOR.

bandspread tuning control An analog adjustment in some older communications receivers that allows continuous tuning over a desired band of frequencies. This control is separate from the main tuning control.

bandstop **1.** The frequency limits between which a BAND-REJECTION FILTER blocks, or greatly attenuates, ac energy. **2.** The ability to suppress or block signals of a given frequency or band of frequencies, while allowing signals of other frequencies to pass with little or no attenuation. Compare BANDPASS.

bandstop filter See BAND-REJECTION FILTER.

band suppression **1.** The property of blocking, or greatly attenuating, signals within a specific frequency band. **2.** The frequency limits between which a device or circuit rejects or blocks ac energy, while passing energy at other frequencies with negligible loss.

band-suppression filter See BAND-REJECTION FILTER.

bandswitch A low-reactance selector switch (usually rotary) that facilitates changing the tuning range of a radio receiver, transmitter or transceiver from one band of frequencies to another.

bandswitching In a receiver, transmitter, or test instrument, the process of switching self-contained tuned circuits to change from one frequency spectrum to another within the range of the device's intended operation.

bandwidth **1.** For a communications or data signal, a measure of the amount of spectrum space the signal occupies. Usually, it is given as the difference between the frequencies at which the signal amplitude is nominally 3 dB down with respect to the amplitude at the center frequency. These frequencies represent the half-power points of the amplitude-versus-frequency function. In general, the bandwidth increases as the data rate (in bits per second, baud, or words per minute) increases. **2.** Also called NECESSARY BANDWIDTH. The minimum amount of spectrum space normally required for effective transmission and reception of a communications or data signal. **3.** See BANDPASS, **1.**

bank A collection of usually similar components used in conjunction with each other, usually in a parallel configuration. Some examples are resistor bank, lamp bank, and transformer bank.

banked transformers Parallel-operated transformers.

bankwound coil A coil wound in such a way that most of its turns are not side by side, thus reducing the inherent distributed capacitance.

bar **1.** Abbreviation, b. The cgs unit of pressure, in which 1 b = 10^5 pascals per square centimeter. **2.** A horizontal or vertical line produced on a television (TV) screen by a bar generator and used to check linearity. **3.** A thick plate of piezoelectric crystal. **4.** A solid metal conductor, usually uninsulated, of any cross section. **5.** A silicon ingot from which semiconductor devices can be fabricated.

BAR Abbreviation of BUFFER ADDRESS REGISTER.

bar code A printed pattern that contains data that can be recovered by laser scanning. It is commonly used for the pricing and identification of store merchandise. It can also be used by an assembly or maintenance robot as an aid to identifying tools.

bar-code reader A laser scanning device that recovers the data from a tag that contains a BAR CODE. The laser beam moves across the tag. The beam is reflected from the white regions between the lines, but is absorbed by the dark lines themselves. This produces modulation of the reflected beam by the data contained in the tag.

bare conductor A conductor with no insulating covering, a common example being bare copper wire.

bar generator A special type of radio-frequency signal generator that produces horizontal or vertical bars on the screen of a television receiver. It is used in adjustment of horizontal and vertical linearity.

bar graph A graphical presentation of data, in which numerical values are represented by horizontal bars of width that correspond to the values. This type of graph is nonstandard in the sense that the ordinate is horizontal, whereas it is usually vertical. Compare COLUMNAR GRAPH.

bar-graph meter See BAR METER.

barium Symbol, Ba. An elemental metal of the alkaline-earth group. Atomic number, 56. Atomic weight, 137.36. It is present in some compounds used as dielectrics (e.g., barium titanate).

barium-strontium oxides The combined oxides of barium and strontium used as coatings of vacuum-tube cathodes to increase electron emission at relatively low temperatures.

barium strontium titanate A compound of barium, strontium, oxygen, and titanium that is used as a ceramic dielectric material. It exhibits ferroelectric properties and is characterized by a high dielectric constant.

barium titanate Formula, $BaTi0_2$. A ceramic used as the dielectric in ceramic capacitors. It exhibits high dielectric constant and some degree of ferroelectricity.

Barkhausen effect The occurrence of minute jumps in the magnetization of a ferromagnetic substance as the magnetic force is increased or decreased over a continuous range.

Barkhausen interference Interference that results from oscillation because of the BARK-HAUSEN EFFECT.

bar magnet A relatively long permanent magnet in the shape of a bar with a rectangular or square cross section.

bar meter A digital meter that displays a quantity, such as signal strength, incrementally, using a set of LEDs or LCDs arranged in a straight line. Its main advantage is that it has no moving parts, yet (unlike direct-readout digital meters) gives the viewer some impression of the way a rapidly fluctuating quantity changes. Its chief

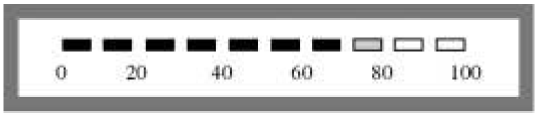

bar meter

disadvantage is that it does not provide a precise indication.

barn Symbol, b. A non-SI unit of nuclear cross section equal to 100 square femtometers or 10^{-24} square centimeters. This unit is approved as compatible with SI (International System of Units).

Barnett effect The development of a small amount of magnetization in a long iron cylinder that is rotated rapidly about its longitudinal axis.

barograph A recording barometer, using either a drum recorder (pen recorder) or a computer to store the data as a function of atmospheric pressure versus time.

barometer An instrument for measuring atmospheric pressure.

barometer effect A relation that appears to exist between the intensity of *cosmic rays* and the atmospheric pressure. It is an inverse relation; that is, increasing pressure seems to correlate with reduced intensity of cosmic rays. It is said to be approximately to 1 or 2% per centimeter of mercury.

barometric pressure The atmospheric pressure, usually given in inches of mercury. The average barometric pressure at the surface of the earth is just under 30 inches of mercury.

bar pattern A series of spaced lines or bars (horizontal, vertical, or both) produced on a television picture screen by means of a BAR GENERATOR. It is useful in adjusting horizontal and vertical linearity of the picture.

barrage array An antenna array in which a string of collinear elements are vertically stacked. The end quarter wavelength of each string is bent in to meet the end quarter wavelength of the opposite radiator to improve balance.

barrage jamming The jamming of many frequencies, or an entire band, at the same time.

barrell distortion Television picture distortion consisting of horizontal and vertical bulging.

barrier **1.** The carrier-free space-charge region in a semiconductor pn junction. **2.** An insulating partition placed between two conductors or terminals to lengthen the dielectric path.

barrier balance The state of near equilibrium in a semiconductor pn junction (after initial junction forming), entailing a balance of majority and minority charge carrier currents.

barrier capacitance **1.** The capacitance in a bipolar transistor between the emitter and collector. It varies with changes in applied voltage, and also with the junction temperature. **2.** The capacitance across any pn junction that is reverse-biased.

barrier height The difference in voltage between opposite sides of a barrier in a semiconductor material.

barrier layer See BARRIER, **1.**

barrier-layer cell A photovoltaic cell, such as the copper oxide or selenium type, in which photons striking the barrier layer produce the potential difference.

barrier potential The apparent internal dc potential across the barrier (see BARRIER, **1**) in a pn junction.

barrier strip A terminal strip having a barrier (see BARRIER, **2**) between each pair of terminals.

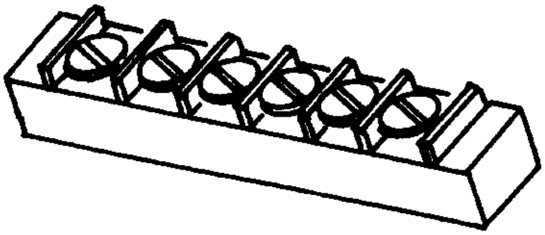

barrier strip

barrier voltage The voltage required for the initiation of current flow through a pn junction.

Bartlett force See EXCHANGE FORCE.

baryon A subatomic particle made up of three quarks.

base **1.** In a bipolar transistor, the intermediate region between the emitter and collector, which usually serves as the input or controlling element of transistor operation. **2.** A substance that dissociates in water solution and forms hydroxyl (OH) ions. For example, sodium hydroxide. **3.** The constant figure upon which logarithms are computed (10 for common logs, 2.71828 for natural logs). **4.** The radix of a number system (e.g., base 10 for the decimal system, base 8 for the octal system, base 16 for the hexadecimal system, and base 2 for the binary system). **5.** A fixed nonportable radio communications installation.

base address The number in a computer address that serves as the reference for subsequent address numbers.

baseband The frequency band of the modulating signal in a communications, broadcast, or data transmitter. For voice communications, this is generally the range of voice frequencies necessary for intelligible transmission. For high-fidelity music broadcasting, it is approximately the range of human hearing. For fast-scan television, it ranges up to several megahertz. It can be restricted or expanded, depending on the nature of the transmitted signal. See BASEBAND FREQUENCY RESPONSE.

baseband frequency response **1.** The amplitude-versus-frequency characteristic of the audio-frequency (AF) or composite video section of a transmitter that defines the BASEBAND, or range of modulating frequencies. **2.** The range of frequencies over which a radio transmitter can be modulated to convey information. For single sideband (SSB), it is approximately 300 Hz to 3 kHz; for high-fidelity, frequency-modulated (FM) music transmission, it is about 10 Hz to 20 kHz or 30 kHz; for fast-scan television, it consists of frequencies up to several megahertz. This range is determined by bandpass and/or lowpass filters in the AF or composite video section of the transmitter.

base bias The steady dc voltage applied to the base electrode of a transistor to determine the operating point along the transistor characteristic curve.

base-bulk resistance The resistance of the semiconductor material in the base layer of a bipolar transistor.

base-charging capacitance In the common-emitter connection of a bipolar transistor, the internal capacitance of the base-emitter junction.

base current Symbol, I_B. Current flowing through the base electrode of a bipolar transistor. Also see AC BASE CURRENT and DC BASE CURRENT.

base electrode See BASE, **1.** Also called *base element.*

base element **1.** Base electrode. **2.** One of the basic metals, such as iron or tin, that are not generally considered precious (as opposed to NOBLE).

base-e logarithm See NAPIERIAN LOGARITHM.

base film The plastic substrate of a magnetic recording tape.

base frequency **1.** The frequency of the principal, or strongest, component in a complex signal or waveform; also called *basic frequency.* **2.** The frequency of operation of a base-station transmitter when the receiver is tuned to a second channel.

base-input circuit A common-collector circuit, common-emitter circuit, or emitter follower.

base insulator A stout dielectric insulator, used to support a heavy conducting element and keep the conductor isolated from other possible conductors or conductive paths.

base line In visual alignment procedures involving an oscilloscope and radio-frequency (RF) sweep generator, a zero-voltage reference line developed by the generator as a horizontal trace on the oscilloscope screen.

baseline stabilizer A clamping circuit that holds the reference voltage of a waveform to a predetermined value. Also called DC RESTORER.

base-loaded antenna A usually vertical antenna or radiating element, the electrical length of which is adjusted by means of a loading coil or tuned circuit in series with, and positioned at the bottom of, the antenna or radiator.

base material In printed circuits, the dielectric material used as a substrate for the metal pattern. Also called *base medium.*

base notation The numbering or radix system used in any application (as octal, decimal, binary, or hexadecimal).

base number See BASE, **4.**

base pin One of the straight prong-like terminals on an electrical or electronic component; it is used to provide support for the device and to allow a physical connection between the socket terminal, into which it fits, and one of the internal electrodes of the device.

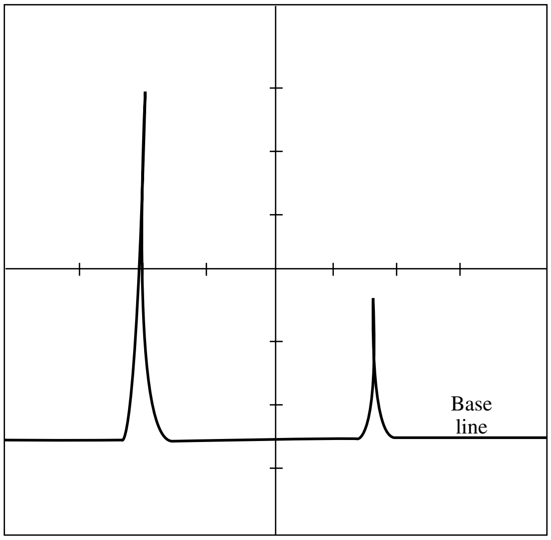

base line

base plate The chassis plate upon which components are mounted before wiring.

base potential See BASE VOLTAGE.

base region See BASE, **1.**

base resistance Symbol, R_B. Resistance associated with the base electrode of a bipolar transistor. Also see AC BASE RESISTANCE and DC BASE RESISTANCE.

base resistor The external resistor connected to the base of a bipolar transistor. In the common-emitter circuit, the base resistor is analogous to the gate resistor of a field-effect transistor (FET) circuit.

base spreading resistance Symbol, r_{BB}. In a bipolar transistor, the bulk-material resistance of the base region between the collector junction and emitter junction.

base station The head station or fixed home station in a communication network.

base-10 logarithm Abbreviation, $\log_{10}$. A logarithm based on the decimal number 10. If $\log_{10}(x) = y$, then $10^y = x$. Base-10 logarithms are commonly used in engineering. Compare NAPIERIAN LOGARITHM.

base voltage Symbol, V_B. The voltage at the base electrode of a bipolar transistor. Also see AC BASE VOLTAGE and DC BASE VOLTAGE.

BASIC Acronym for BEGINNER'S ALL-PURPOSE SYMBOLIC INSTRUCTION CODE, a relatively primitive, but versatile and easy-to-learn computer language developed at Dartmouth College.

basic frequency **1.** The FUNDAMENTAL FREQUENCY of a signal, as opposed to one of its harmonics. **2.** See BASE FREQUENCY, **1.**

basic protection Devices and procedures essential to minimize the risk of damage to electronic equipment, and/or injury or death to its operators, as a result of lightning. Hardware provisions include a substantial earth ground, heavy-gauge grounding wire, lightning arrestors for antennas, and transient suppressors for power connections. The safest procedure is to disconnect and ground all antennas, and unplug all equipment from utility outlets, during electrical storms and/or when the apparatus is not in use. Radio communications equipment with outdoor antennas, in particular, should not be operated during thunderstorms.

basket The structure that supports the cone in an acoustic loudspeaker.

basket-weave coil A type of single-layer inductor in which adjacent turns do not parallel each other around the circumference, but zigzag oppositely as a strand does in the woven pattern of a basket. This reduces distributed capacitance.

bass Low audio frequencies (AF) corresponding to low-frequency musical notes or sounds.

bass boost **1.** The special emphasis given to low audio frequencies (the bass notes) by selective circuits in audio systems. **2.** The technique of increasing the loudness of the bass, relative to the higher audio frequencies, to render a more faithful reproduction of sound at low volume levels.

bass compensation See BASS BOOST, **2.**

bass control **1.** A manually variable potentiometer for adjusting bass boost of an amplifier or sound system. **2.** The arrangement of components that are required to achieve amplitude variation of bass in an audio signal.

bass port In a loudspeaker, a hole in the cabinet that enhances the low-frequency (bass) sound output. Used in high-fidelity audio systems.

bass-reflex enclosure A loudspeaker cabinet with a critically dimensioned duct or port that allows back waves to be radiated in phase with front waves, thus averting unwanted acoustic phase cancellation.

bass-reflex loudspeaker A loudspeaker mounted in a bass reflex enclosure. Also see ACOUSTICAL PHASE INVERTER.

bass-resonant frequency The low frequency at which a loudspeaker or its enclosure displays resonant vibration.

bass roll-off **1.** The attenuation of the low-frequency (bass) component in a high-fidelity audio signal. **2.** A control that allows adjustable attenuation of the low-frequency component in a high-fidelity audio signal.

bass suppression In speech transmission, the removal of all frequencies below about 300 Hz, on the assumption that those frequencies contribute little to intelligibility. This suppression permits the speech level to be increased without overmodulating a transmitter. It also allows smaller audio transformers to be used because transformer core size must increase as the frequency it passes decreases.

bassy In audio and high-fidelity applications, a sound in which the low-frequency components, below about 500 Hz, are overly predominant.

BAT Abbreviation of BATTERY.

batch fabrication process The manufacture of devices in a single batch from materials of uniform grade. Particularly, the manufacture of a large number of semiconductor devices from one batch of semiconductor material by means of carefully controlled, identical processes.

batch processing In digital-computer operations, the processing of quantities of similar information during a single run.

bat-handle switch A toggle switch, the lever of which is relatively long and thick, and is shaped like a baseball bat.

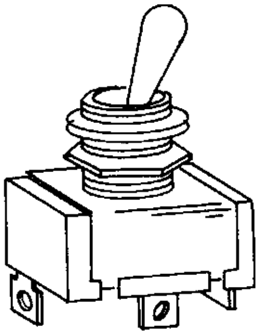

bat-handle switch

bathtub capacitor A (usually oil-filled) capacitor housed in a metal can that looks like a miniature bathtub.

bathyconductorgraph An instrument that is used to measure the electrical conductivity of seawater.

bathythermograph An instrument that plots a graph of temperature versus depth in a body of water, such as a lake or an ocean.

batten Supporting bars or braces that hold a loudspeaker in place within its cabinet, and/or that hold the cabinet panels in place.

battery Abbreviations, B, BA. BAT. A device consisting of two or more interconnected electrochemical or photovoltaic cells that generate dc electricity. The cells can be connected in series to supply a desired voltage, in parallel to supply a desired current-delivering capability, or in series-parallel to obtain a desired voltage and current-delivering capability. Also see CELL, EDISON BATTERY, LEAD-ACID BATTERY, PHOTO-

VOLTAIC CELL, PRIMARY BATTERY, and STORAGE BATTERY.

battery acid **1.** A chemical acid, such as sulfuric acid, used as the electrolyte of a battery. **2.** Colloquially, any cell or battery electrolyte, whether acid, base, or salt.

battery capacity The current-supplying capability of a battery, usually expressed in ampere-hours (Ah).

battery cell See CELL, **1.**

battery charger **1.** A specialized dc power supply, usually embodying a stepdown transformer, rectifier, and filter. It is used to charge a storage battery from an ac power line. **2.** A motor-generator combination used to charge a storage battery from an ac power line. **3.** A combination of solar cells, generators, or other voltaic transducers, that are used to charge a storage battery with dc obtained from a nonelectrical energy source.

battery clip **1.** A heavy-duty metallic clamp that is used for quick, temporary connection to a large cell terminal, such as that of a lead-acid storage battery. **2.** A small connector of the snap-fastener type, used for quick connection to a small power source, such as a transistor-radio battery.

battery eliminator A specialized dc power supply, usually embodying a transformer, rectifier, and filter, that permits battery-powered equipment to be operated from an ac power line.

battery holder **1.** A case or container of any kind for holding a cell or battery. **2.** A shelf for holding a cell or battery. **3.** A small, metal bracket-type device for holding a cell or battery between two contacts.

battery life **1.** The ampere-hour or watt-hour capacity of a battery. **2.** The number of times that a rechargeable electrochemical battery can be cycled before it becomes unusable. **3.** The nominal length of time (e.g., hours, days, or weeks) that an electrochemical battery will function effectively in a given application before it must be discarded or recharged.

battery memory See MEMORY DRAIN.

battery receiver A usually portable radio or television receiver operated from self-contained batteries.

battery substitute See BATTERY ELIMINATOR.

bat wing On a television (TV) or frequency-modulation (FM) broadcast receiving antenna, a metallic element with a shape that resembles that of a bat's wing.

baud A unit of communications processing speed in telegraphy and digital data communications systems. Often confused with *bits per second* (bps). Baud refers to the number of times per second that a signal changes state. The speed in bps is generally higher than the speed in baud, sometimes by a factor of several times. Compare BITS PER SECOND.

Baudot code A machine communications code that uses five parallel binary digits of equal length, the interpretation of which depends on the history of the previous transmission or an additional case bit.

baud rate **1.** A colloquial expression for data speed in BAUD. **2.** Colloquial, and technically inaccurate, expression for data speed in BITS PER SECOND.

Baume (Antione Baume, 1728–1804). Abbreviation, Be. Pertaining to the *Baume scales* for hydrometers. The two such scales are for liquids heavier than water and for liquids lighter than water.

bay One of several sections of a directional antenna array.

bayonet base The insertable portion of a plug-in component (e.g., a lamp) that has a projecting pin that fits into a slot or keyway in the shell of the socket into which the component is inserted.

bayonet socket A socket with a suitably slotted shell for receiving the bayonet base of a plug-in component.

bazooka A linear BALUN, in which a quarter wavelength of metal sleeving surrounds a coaxial feeder, and is shorted to the outer conductor of the feeder to form a shorted quarter-wave section.

bb Abbreviation of BLACKBODY.

BBC Abbreviation of *British Broadcasting Corporation.*

BBM Abbreviation of BREAK BEFORE MAKE.

b-box The index register of a computer.

BC Abbreviation of BROADCAST.

BCD Abbreviation of BINARY-CODED DECIMAL.

BCFSK Abbreviation of BINARY CODE FREQUENCY-SHIFT KEYING.

B channel One of the channels of a two-channel stereophonic system. Compare A CHANNEL.

BCI Abbreviation of BROADCAST INTERFERENCE.

BCL Amateur radio abbreviation of BROADCAST LISTENER.

BCN Abbreviation of BEACON.

BCO Abbreviation of BINARY-CODED OCTAL.

BCST Abbreviation of BROADCAST.

BDC Abbreviation of BINARY DECIMAL COUNTER.

B display A radar display in which the target is represented by a bright spot on a rectangular-coordinate screen. Compare A DISPLAY and J DISPLAY.

Be Symbol for BERYLLIUM.

Be Abbreviation of BAUME.

beacon **1.** A beam of radio waves, or a radio signal, that is used for navigation and/or direction finding. **2.** A transmitter that radiates a beam of radio waves, or a radio signal, as an aid in navigation and/or direction finding. **3.** A signal transmitted continuously on a specific frequency, to help radio operators ascertain propagation conditions. **4.** A station or transmitter that generates and radiates a signal to help radio operators determine propagation conditions. **5.** In robotics, a device or system that aids in navigation. For example, tricorner reflectors can be positioned in strategic locations, and a mobile robot equipped with a scanning infrared laser. The robot controller determines the distance to any given reflector by measuring the time required for the laser beam to return. In this way, two mirrors can allow the robot to locate its position in two dimensions; three mirrors can facilitate position determination in three-dimensional space.

beacon direction finder A direction finder using a signal received from a beacon station.

beacon receiver A receiver that is specially adapted for the reception of beacon signals (see BEACON, **1** and **3**).

beacon station **1.** A station broadcasting beacon signals (see BEACON, **1** and **3**) for direction finding, navigation, and/or determination of radio-wave propagation conditions. **2.** Sometimes, a radar transmitting station.

beacon transmitter A transmitter specially adapted for the transmission of beacon signals (see BEACON, **1** and **3**).

bead **1.** A small ferromagnetic ring that is used as a passive decoupling choke by slipping it over the input power leads of a circuit or stage, or around a coaxial transmission line. **2.** A magnetic memory element in a ferrite-core matrix.

beaded coax A low-loss, coaxial transmission line, in which the inner conductor is separated from the outer conductor by means of spaced dielectric beads.

beaded support A plastic or dielectric bead that is used to support the inner conductor of an air-insulated transmission line of coaxial construction.

bead thermistor A thermistor consisting essentially of a small bead of temperature-sensitive resistance material into which two leads are inserted.

beam **1.** The more-or-less narrow pattern of radiation from a directional antenna. **2.** A directional antenna—especially a YAGI ANTENNA. **3.** The stream or cloud of electrons emitted by the cathode in an electron tube—especially a BEAM POWER TUBE.

beam alignment **1.** The lining-up of a directional transmitting antenna with a directional receiving antenna for maximum signal transfer. **2.** In a television (TV) camera tube, the lining-up of the electron beam so that it is perpendicular to the target. **3.** In a cathode-ray tube, the positioning of the electron rays so that they converge properly on the screen, regardless of the deflection path.

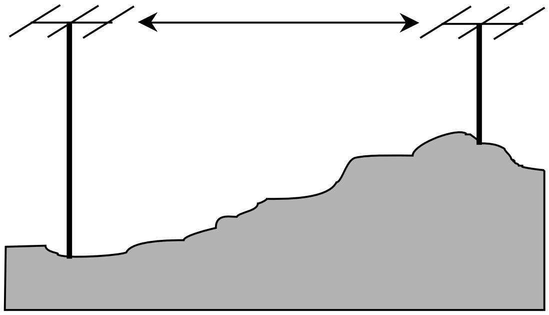

beam alignment

beam angle In the radiation from an antenna, the direction of most intense radiation, the side limits of which are determined by the points at which the field strength drops to half the value in the principal direction.

beam antenna **1.** A multielement directional antenna, consisting of a half-wave driven dipole and one or more parasitic elements. See YAGI ANTENNA. **2.** Any directional antenna used for transmitting and receiving radio-frequency (RF) signals.

beam bender **1.** In a television (TV) picture tube, the ion-trap magnet. **2.** Deflection-plate correction device or circuit.

beam bending Deflection of an electron beam by electric or magnetic fields.

beam blanking See BLANK, **2.**

beam convergence The meeting, at a shadow-mask opening, of the three electron beams in a three-color television picture tube. See BEAM ALIGNMENT, **3.**

beam coupling A method of producing an alternating current between two electrodes by passing a density-modulated beam of electrons between the electrodes. This, in effect, demodulates the electron beam, recovering the information.

beam crossover Either of the half-power points in the beam of a directional antenna, usually in the horizontal plane. The reference point is considered to be the direction of maximum radiation.

beam current The current represented by the flow of electrons in the beam of a cathode-ray tube.

beam cutoff In an oscilloscope or television picture tube, the complete interruption of the electron beam, usually as a result of highly negative control-grid bias.

beam deflector A deflection plate in an oscilloscope tube.

beam efficiency In a cathode-ray tube, the ratio of the number of electrons generated by the gun to the number reaching the screen. The efficiency is high in electromagnetic-deflection tubes and lower in electrostatic-deflection tubes.

beam lead In an integrated circuit, a relatively thick and strong lead that is deposited in contact with portions of the thin-film circuit. It provides stouter connections than continuations of the thin film would provide.

beam-lead isolation In an integrated circuit, reduction of distributed capacitance and other interaction through use of beam leads.

beam modulation See INTENSITY MODULATION.

beam parametric amplifier A PARAMETRIC AMPLIFIER in which the variable-reactance component is supplied by a modulated electron beam.

beam-positioning magnet In a three-gun color television picture tube, a permanent magnet that is used to position one of the electron beams correctly, with respect to the other two.

beam power tube A tetrode or pentode vacuum tube, in which special deflector plates concentrate the electrons into beams in their passage from cathode to plate. The beam action greatly increases plate current at a given plate voltage. It is used in some radio-frequency (RF) power amplifiers.

beam-rider control system A missile-guidance system in which a control station sends a radio beam to a missile. The beam is moved in such a way that as the missile stays within the beam, it hits the target.

beam-rider guidance **1.** An aircraft landing guidance system, in which the aircraft follows a radio beam in its glide path. **2.** The circuitry in a guided missile using a beam-rider control system.

beam splitter A device used to divide a light beam (as by a transparent mirror) into two components, one transmitted and the other reflected; hence, a BEAM-SPLITTING MIRROR.

beam splitting In radar, a method of calculating the mean azimuth of a target from the azimuth at which the target is first revealed by one scan, and the azimuth at which the target information ceases.

beam-splitting mirror In an oscilloscope-camera system, a tilted, transparent mirror that allows rays to pass horizontally from the oscilloscope screen to the camera and to be reflected vertically to the viewer's eye.

beamwidth of antenna The angular width of the main lobe of the pattern of radiation from a directional antenna. Generally, it is measured between the half-power points in the horizontal plane. Occasionally, it is measured in the vertical plane.

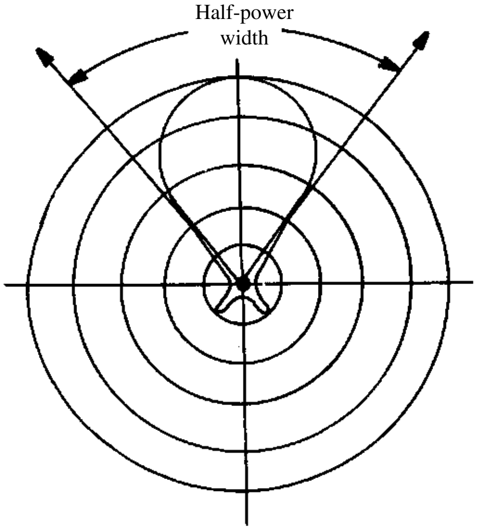

beamwidth of antenna

bearing The direction of an object or point expressed in degrees within a 360° horizontal clockwise boundary, with the center of the circle serving as the observation point.

bearing resolution In radar operations, the minimum horizontal separation of two targets, in degrees, that permits the individual targets to be displayed as two echoes, rather than one.

beat Any one of the series of pulsations constituting a beat note, which results from heterodyning one signal against another.

beat frequency Either of two frequencies f_{C1} and f_{C2} resulting from the mixing of two signals of different frequencies f_A and f_B. Frequency f_{C1} is the sum of the two input frequencies; $f_{C1} = f_A + f_B$. Frequency f_{C2} is the difference; $f_{C2} = f_A - f_B$ when f_A is the higher of the two input frequencies.

beat-frequency oscillator Abbreviation, BFO. An oscillator used to set up audible beat frequencies with an incoming received signal and installed in the intermediate-frequency (IF) stages of a superheterodyne communications receiver. For single-sideband (SSB) reception, the BFO is set at the frequency of the received suppressed carrier. In continuous-wave (CW) Morse code reception, the BFO is set at a frequency that differs from that of the incoming signal by about 400 to 1000 Hz. The resulting tone has an audio frequency equal to the difference between the BFO frequency and the received signal carrier frequency. For reception of frequency-shift-keyed (FSK) signals, the BFO is set to such a frequency that the resulting audio beat notes are appropriate for the mark and space inputs of a terminal unit or modem.

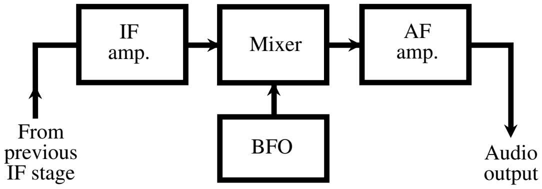

beat-frequency oscillator

beating **1.** Also called *heterodyning.* The combination of signals of different frequencies resulting in sum and difference frequencies. **2.** The fluttering noise heard when two audio tones, very close in frequency and very similar in amplitude, are emitted at the same time.

beat marker In the visual (oscilloscopic) alignment of a tuned circuit, a marker pip that results from the beat note between the sweep-generator signal and the signal from a marker oscillator.

beat note The sum or difference frequency that results from the heterodyning of two signals or, under some conditions, of more than two signals.

beat-note reception **1.** Reception in which a radio-frequency carrier is made audible by heterodyning it with a beat-frequency oscillator (BFO) to produce an audible beat note. **2.** Superheterodyne reception (see SUPERHETERODYNE CIRCUIT).

beat tone A beat note in which the frequency is within the range of hearing.

beaver tail A flat or elongated radar beam, wide in the azimuth plane. Primarily used to determine the altitude of a target. The beam is moved up and down to find the target elevation.

Becquerel effect A phenomenon in which a voltage is produced when radiant energy, such as infrared, visible light, ultraviolet, or X-rays, falls on one electrode in an electrolytic cell.

bedspring A directional antenna consisting of a broadside array with a flat reflector and one or more helical driven elements.

beep A test or control signal, usually of single tone and short duration.

beeper **1.** A pocket- or hand-carried transceiver—especially one for maintaining two-way contact with personnel who are away from their base. **2.** An acoustic transducer that produces a beep in response to an input signal.

beetle A urea formaldehyde plastic used as a dielectric material and as a container material.

bel Abbreviation, B. The basic logarithmic unit (named for Alexander Graham Bell) for expressing gain or loss ratios. One bel is equivalent to a power gain of 10. Also see DECIBEL.

bell An electric alarm device consisting of a metallic gong that emits a ringing sound when it is struck by an electrically vibrated clapper.

Bellini-Tosi direction finder A direction finder in which the sensing element consists of two triangular vertical antennas crossed at right angles, the antennas being open at the top and accordingly not acting as conventional coil antennas.

bell-shaped curve A statistical curve (so called from its characteristic shape) that exhibits a *normal distribution* of data. Typically, the curve describes the distribution of errors of measurement around the real value.

bell transformer A (usually inexpensive) stepdown transformer that operates an electric bell or similar alarm or signaling device from the ac power line.

bell wire Insulated 18-gauge (AWG) solid copper wire, so called because of its principal early use in the wiring of electric-bell circuits.

belt generator Also known as a *Van de Graaff generator.* A very-high-voltage electrostatic generator, a principal part of which is a fast-traveling endless belt of dielectric material. At the lower end, charges of one sign are sprayed on the belt at 10 to 100 kV dc and are carried to the inside of a hollow metal sphere at the upper end, where they are removed and spread to the surface of the sphere, which they raise to a potential up to several million volts.

benchmark A test standard to measure product performance.

benchmark routine A routine designed to evaluate computer software and/or hardware, producing a good indication of how well the software or hardware will perform in real-life situations. In particular, tests *instructions per second* and *throughput*, thereby producing an indication of the overall computer power in applications, such as word processing, database, spreadsheet, graphics, animation, and mathematical calculations.

bench test An extensive checkout of a piece of equipment in the test laboratory—either to find an intermittent problem, or to check for reliability.

bend An angular shift in the lengthwise direction of a waveguide.

bending effect **1.** The downward refraction of a radio wave by the ionosphere. **2.** The low-atmosphere turning of a radio wave downward by temperature discontinuity and atmospheric inversions.

Benito A continuous-wave method of measuring the distance of an aircraft from the ground, involving the transmission of an audio-modulated signal from ground and the retransmission back to ground by the aircraft. The phase shift between the two signals is proportional to the distance to the aircraft.

bent antenna An antenna that has its driven element bent, usually near the ends and at right angles, to conserve space.

bent gun A television picture tube neck arrangement having an electron gun that is slanted to direct the undesired ion beam toward a positive electrode, but which allows the electron beam to pass to the screen. This prevents the ion beam from "burning" a permanent spot on the phosopor of the screen.

BeO Formula for beryllium oxide. Also see BERYLLIA.

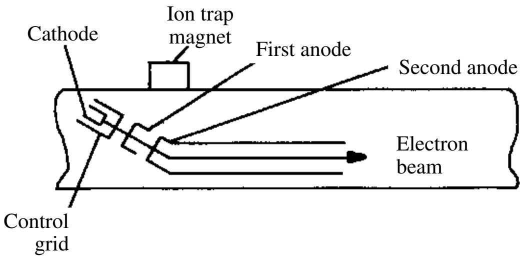

bent gun

berkelium Symbol, Bk. A radioactive elemental metal produced artificially. Atomic number, 97. Atomic weight, 247.

beryllia Formula, $Be0$. Beryllium oxide, used in various forms as an insulator and structural element (as in resistor cores).

beryllium Symbol, Be. An elemental metal. Atomic number, 4. Atomic weight, 9.01218. Beryllium is present in various dielectrics and alloys used in electronic components.

Bessel functions Sophisticated mathematical functions for dealing with periodic electronic phenomena in which the waveform often displays decrement. Also called *cylindrical functions.*

beta Symbol, β. The current gain of a common-emitter bipolar transistor stage. It is the ratio of the induced change of collector current to the applied change of base current: $\beta = dIC/dIB$.

beta circuit The output-input feedback circuit in an amplifier.

beta cutoff frequency The frequency at which the current amplification of a bipolar transistor falls to 70.7% of its low-frequency value.

beta particles Minute radioactive subatomic bits identical to the electron or positron, and emitted by some radioactive materials. Also see BETA RAYS.

beta rays Rays emitted by the nuclei of radioactive substances, consisting of a stream of beta particles (i.e., electrons or positrons) that move at velocities up to 299.8 million meters per second. Compare ALPHA PARTICLE and GAMMA RAYS.

beta-to-alpha conversion For a bipolar transistor, the conversion of current amplification expressed as beta (β) to current amplification expressed as alpha (α): $\alpha = \beta/(\beta + 1)$.

betatron A particle accelerator in which injected electrons are given extreme velocity by being propelled in circular paths in a doughnut-shaped glass container. The term comes from the fact that high-speed electrons constitute BETA PARTICLES.

beta videocassette recorder The earliest scheme for videocassette recording, developed by Sony corporation in the 1970s. Compare *VHS videocassette recorder.*

beta zinc silicate phosphor Formula, $(Zn0 + Si0_2)$:Mn. A phosphorescent substance used to coat the screen of a cathode-ray tube. The fluorescence is green-yellow.

BeV Abbreviation of *billion electronvolts.* Also see ELECTRONVOLT, GEV, MEV, and MILLION ELECTRONVOLTS. This abbreviation has been supplanted by the SI (International System of Units) abbreviation GeV, for GIGAELECTRONVOLTS.

bevatron An accelerator (see ACCELERATOR, **1**) similar to the synchrotron, which accelerates particles to levels greater than 10 GeV.

Beverage antenna (Harold H. Beverage.) A nonresonant, directional long-wire antenna, erected a few feet above ground and run in a straight line for one to several wavelengths. It is generally used for reception at low and medium frequencies, the best response is to vertically polarized signals arriving from one or both directions in line with the wire. It can be left unterminated for bidirectional response, or it can be terminated at its far end by a noninductive resistor of about 600 ohms for a unidirectional response.

beyond-the-horizon propagation See FORWARD SCATTER.

bezel A faceplate for an electronic instrument, usually having a fitted rim and cutouts for knobs, switches, jacks, etc.

BFO Abbreviation of BEAT-FREQUENCY OSCILLATOR.

BG Abbreviation of BIRMINGHAM WIRE GAUGE. Also abbreviated BWG.

B-H curve A plot showing the B and H properties of a magnetic material. Magnetizing force H is plotted along the horizontal axis, and flux density B is plotted along the vertical axis.

B-H loop See BOX-SHAPED LOOP.

B-H meter Any instrument for displaying or evaluating the hysteresis loop of a magnetic material.

bhp Abbreviation of *brake horsepower.*

Bi Symbol for BISMUTH.

bias **1.** Any parameter of which the value is set to a predetermined level to establish a threshold or operating point. Although it is common to think of bias currents and bias voltages, other parameters (e.g., capacitance, resistance, illumination, magnetic intensity, etc.) can serve as biases. **2.** In a high-fidelity audio system, a circuit in a tape recorder/player that optimizes performance for a particular type of recording tape.

bias current A steady, constant current that presets the operating threshold or operating point of a circuit or device, such as a transistor, diode, or magnetic amplifier. Compare BIAS VOLTAGE.

bias current drift The ratio of a change in input bias current to a change in ambient temperature, generally expressed in nanoamperes per degree Celsius.

bias distortion Distortion caused by operation of a tube or transistor with incorrect bias so that the response of the device is nonlinear.

biased diode A diode having a dc voltage applied in either forward or reverse polarity. Current flows readily through the forward-biased diode; the reverse-biased diode appears as an open circuit. The biased diode is the basis of clippers, limiters, slicers, and similar circuits.

biased off In a circuit or device, the state of cutoff caused by application of a control-electrode bias. Examples include collector-current cutoff (when the dc base bias of a bipolar transistor reaches a critical value), and drain-current cutoff (when the dc gate bias reaches a critical value in a field-effect transistor).

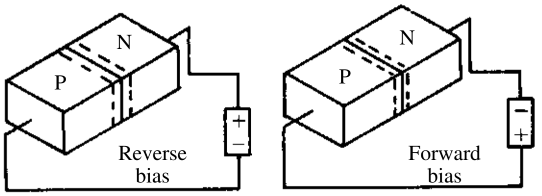

biased diodes

biased search A scheme that a mobile robot can use to find its way to a destination or target, by deliberately searching off to the side and then homing in as the approach progresses. It is so called because the general nature of the initial error (bias) is known, although its exact extent need not be known.

bias oscillator In a magnetic recorder, an oscillator operated at a frequency in the 40-kHz to 100-kHz range to erase prerecorded material and bias the system magnetically for linear recording.

bias resistor A usually fixed resistor, such as the source resistor in a field-effect-transistor (FET) circuit or the emitter resistor in a bipolar-transistor circuit, across which a desired bias voltage is developed by current flowing through the resistor.

bias set A control, such as a potentiometer or variable autotransformer, that facilitates manual adjustment of the dc bias of a circuit.

bias stabilization **1.** The maintenance of a constant bias voltage, despite variations in load impedance or line voltage. It is usually accomplished by means of automatic voltage regulation. **2.** The stabilization of transistor dc bias voltage by means of resistance networks or through the use of barretters, diodes, or thermistors.

bias supply **1.** Batteries that provide bias voltage or current for bipolar or field-effect transistors. **2.** A line-operated unit for supplying dc bias and consisting of a transformer, rectifier, and high-grade filter.

bias voltage A steady voltage that presets the operating threshold or operating point of a circuit or device, such as a transistor. Compare BIAS CURRENT.

bias windings The dc control windings of a saturable reactor or magnetic amplifier.

biconical antenna A form of broadband antenna, consisting of two conical sections joined at the apexes. The cones are at least ¼ wavelength in diagonal height. The vertex angles of the cones can vary, although the apex angle is usually the same in each cone. The vertex angle affects the feedpoint impedance. Such an antenna radiates, and responds optimally to, signals with polarization parallel to the axis of the cones.

biconical horn antenna A double-horn microwave antenna that radiates along relatively sharp front and back beams.

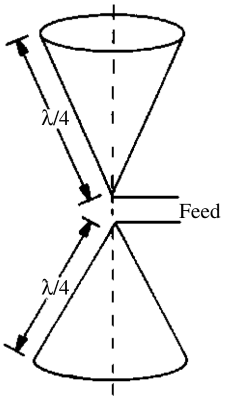

biconical antenna

bidecal base The 20-pin base of a cathode-ray tube. Also see DIHEPTAL, DUODECAL, and MAGNAL.

bidirectional Radiating or receiving (usually equally) from opposite directions (e.g., front-and-back radiation from an antenna or loudspeaker, or front-and-back pickup with an antenna or microphone).

bidirectional antenna An antenna with a directional pattern that consists of maximum lobes 180 degrees apart.

bidirectional bus In computers, a data path over which both input and output signals are routed.

bidirectional bus driver In a microcomputer, a signal-driving device that permits direct connection of a buffer-to-buffer arrangement on one end (the interface to I/O, memories, etc.) and data inputs and outputs on the other. This device permits bidirectional signals to pass and provides drive capability in both directions.

bidirectional counter A counter that can count consecutively up from a given number or down from that number. Also called UP-DOWN COUNTER.

bidirectional current A current that flows in both directions. Utility alternating current (ac) is a common example.

bidirectional loudspeaker A loudspeaker that delivers sound waves to the front and rear.

bidirectional microphone A microphone that picks up sound waves equally well from the front and rear.

bidirectional transistor A symmetrical transistor (i.e., one in which the two main current-carrying electrodes can be interchanged without influencing device performance). Some field-effect transistors (FETs) are of this type; the drain and the source can be interchanged.

bifilar electrometer An electrometer in which the sensitive element consists of two long platinized-quartz fibers. When an electric potential is applied, the fibers separate by a distance proportional to the voltage.

bifilar resistor A wirewound resistor with two oppositely wound filaments. The nature of the winding tends to cancel the inductance, making the device useful at a much higher frequency than an ordinary wirewound resistor.

bifilar transformer A transformer in which unity coupling is approached by interwinding the primary and secondary coils (i.e., the primary and secondary turns are wound side by side and in the same direction).

bifilar winding **1.** A method of winding a coil (such as a resistor coil) in the shape of a coiled hairpin so that the magnetic field is self-canceling and the inductance is minimized. **2.** A method of winding transformers to minimize leakage reactance.

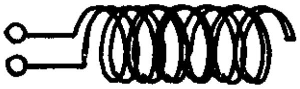

bifilar winding

bifurcated contact A forked contact whose parts act as two contacts in parallel for increased reliability.

bilateral amplifier An amplifier that transmits or receives in either direction equally well (i.e., the input and output can be exchanged at will).

bilateral antenna A bidirectional antenna, such as a loop antenna or a half-wave dipole.

bilateral element A circuit element or component (as a capacitor, resistor, or inductor) that transmits energy equally well in either direction. Compare UNILATERAL ELEMENTS.

bilateral network A network, usually passive and either balanced or unbalanced, that has BILATERAL SYMMETRY. Thus, the input and output terminals can be exchanged without affecting the performance of the network in any way.

bilateral symmetry **1.** Exhibiting symmetry, with respect to a vertical line or plane. **2.** For a network, having the property that if the input and output are reversed, the circuit behavior remains precisely the same. See BILATERAL NETWORK. **3.** For an amplitude-versus-frequency response curve, having the property that the right-hand and left-hand halves are mirror images of each other.

billboard antenna A phased group of dipole antennas that lie in one plane. A reflector might be used behind the entire array.

bilobe pattern An antenna radiation pattern consisting of two major lobes in a given plane, usually the horizontal plane. Often the lobes exist in opposite directions relative to each other, as in a half-wave dipole. But they can be at varying angles, as in a long-wire antenna.

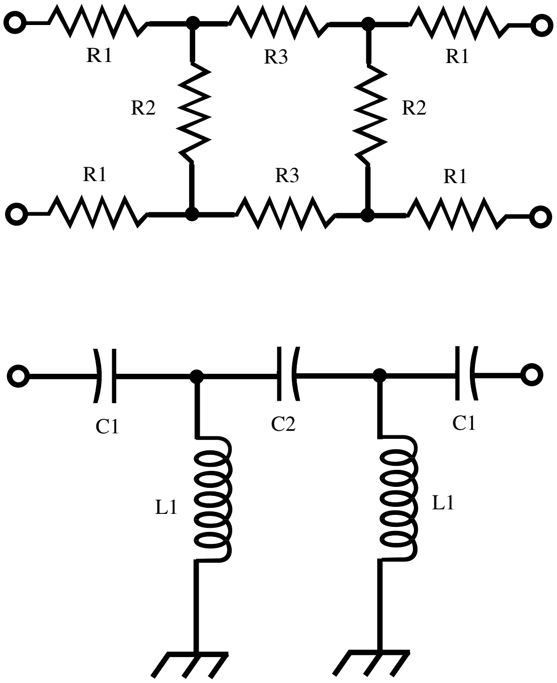

bilateral network

bimetal A union of two dissimilar metals—especially those having a different temperature coefficient of expansion. The two are usually welded together over their entire surface.

bimetallic element A strip or disk of bimetal. When the element is heated, it bends in the direction of the metal that has the lower temperature coefficient of expansion; when cooled, it unbends. Usually, an electrical contact is made at one extreme or the other so that the element can serve as a thermostat.

bimetallic switch A temperature-sensitive switch based on a bimetallic element.

bimetallic thermometer A thermometer based on a bimetallic element that is mechanically coupled (as through a lever and gear system) to a pointer that moves over a temperature scale.

bimetallic thermostat A thermostat in which a bimetallic element closes or opens a pair of switch contacts.

bimorphous cell A piezoelectric transducer that consists of two crystal plates, such as Rochelle salt, bound intimately face to face. In a crystal microphone, vibration of the transducer results in a voltage output; in a crystal headphone, an ac signal voltage impressed on the transducer causes vibratory mechanical motion.

BiMOS A combination of bipolar and MOSFET transistors in an integrated circuit. Thus, a typical BiMOS device can have MOSFET input for high impedance and bipolar output for low impedance.

binant electrometer An electrometer in which a thin platinum vane ("the needle") is suspended within two halves of a metal pillbox-shaped container. The halves or binants are biased with a dc voltage of 1 to 12 V, and the unknown voltage is applied to the vane. It is also called DUANT ELECTROMETER and HOFFMAN ELECTROMETER.

binary **1.** Pertaining to the base-2 number system. Thus, binary arithmetic uses two digits: 0 and 1. **2.** Pertaining to two-element chemical compounds.

binary arithmetic Mathematical operations performed using only the digits 0 and 1.

binary cell In a computer memory, an element that can display either of two stable states.

binary chain A cascade of binary elements, such as flip-flops, each unit of which affects the stable state of the succeeding unit in sequence.

binary channel Any channel whose use is limited to two symbols.

binary code A system of numbers representing quantities by combinations of 1 and 0; a binary-number system.

binary-coded decimal notation In digital computer operations, a system of notation in which each digit of a decimal number is represented by its binary equivalent. Thus, the decimal number 327 in BCD notation becomes 0011 0010 0111. (By contrast, in pure binary notation, 327 is 101000111.)

binary-coded octal notation A method of numbering in which each base-8 digit is represented by a binary number from 000 to 111.

binary-controlled gate circuit A gate circuit controlled by a binary stage. An example is a gating transistor that receives its on/off pulses from a flip-flop.

binary counter A counter circuit consisting of a cascade of bistable stages. Each stage is a scale-of-two counter because its output is on for every second input pulse. At any instant, the total binary count in a multistage counter thus is shown by the on and off states of the various stages in sequence.

binary decoder A device or stage that accepts binary signals on its input lines, and provides a usually exclusive output (representing a decimal digit, for example).

binary digit See BIT.

binary number system The base-two system of notation. This system uses only two symbols, 0 and 1, and accordingly is easily applied to two-position switches, relays, and flip-flops.

binary preset switch In a binary counter or binary control circuit, a selector switch that allows the circuit to be preset to deliver an output pulse only after a predetermined number of input pulses.

binary relay See BISTABLE RELAY.

binary scaler In its simplest form, a single two-stage device, such as a flip-flop, which functions as a divide-by-two counter, because one output pulse results from every two input pulses. Higher-order scaling is obtained by cascading stages.

binary search A system of search entailing the successive division of a set of items into two parts and the rejection of one of the two until all items of the sought-for kind are isolated.

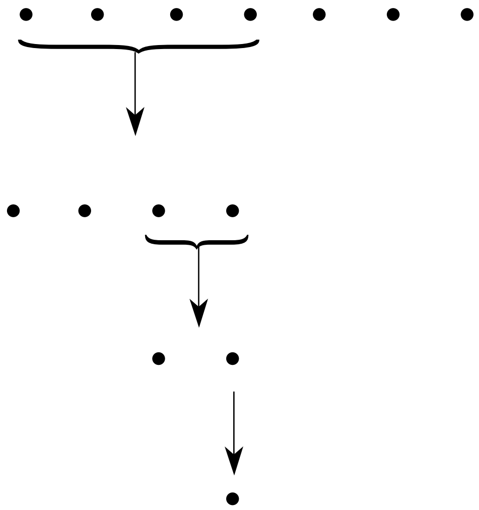

binary search

binary signal Any signal that can attain either of two states. Such a signal is always a digital signal.

binary-to-decimal conversion **1.** The automatic conversion of a number represented by a series of binary pulses into the corresponding decimal number, which then is displayed by a readout device. **2.** The arithmetic operation of converting a binary number into a decimal number; this can be done by noting the powers of 2 represented by the various binary digits in a number, and then adding the decimal values of these powers.

binary word A binary numeral that has a particular meaning, agreed upon by convention. For example, the letters A through Z can be represented by binary numbers 00001 through 11010; a word can be represented by several blocks of five digits.

binaural Literally, two-eared. In sound recording and reproduction, the transcription of a broad sound source using two microphones spaced at approximately the distance between the ears on a human head, and played back using headphones to re-create the stereo effect. The technique evolved into multichannel stereophonic reproduction.

binaural machine hearing Also called *stereo machine hearing.* The ability of a machine, such as a robot, to sense the direction and distance to a source of sound, using two acoustic transducers and a computer to process their output signals. The machine determines the location of the sound source by comparing the relative amplitude and phase of the signals from the two transducers. It functions according to the same principle as human hearing, in which a person can determine the general direction and distance to a sound source by subconsciously comparing the relative amplitude and phase of the sounds arriving at the left and right ears.

binaural sound The equivalent of a listener hearing a concert through a pair of earholes; it takes earphones to reproduce the signal. If speakers are substituted for the earphones, the listener hears monophonically, as if standing back several feet from the earholes.

binder A material (such as lacquer) that acts as a holder and cohesive medium for the particles of another material. It is used in carbon resistors, ceramic dielectric bodies, powder cores, and resistive and metallic paints.

binding energy A property of the nucleus of an atom. The binding energy of a nucleus is equal to the difference between the nuclear weight and the sum of the weights of the lighter particles making up the nucleus. The nucleus is stable when the binding energy is high.

binding force Any one of the electrostatic forces that bind crystals together.

binding post A screw-type terminal of various styles, often having a hole into which a wire or tip can be inserted and gripped. It is used for temporary indoor connections only.

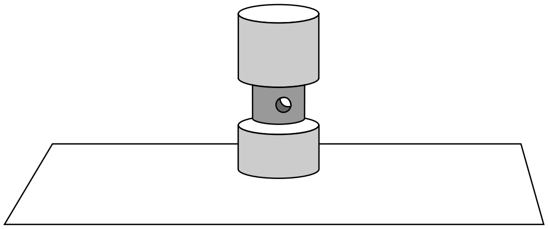

binding post

binistor A semiconductor switching device that exhibits two stable states and also negative resistance.

binocular machine vision Also called *stereoscopic machine vision.* The ability of a machine vision system to provide depth and perspective data. Uses two optical sensors spaced a fixed distance apart. The left sensor sees a slightly different image than the right sensor. These two images are combined and processed by a computer, allowing the machine (such as a mobile robot) to determine the distances to various objects in its environment. Functions on the same principle as stereoscopic human vision.

bin picking In robotics, the selection of a particular object from a container (bin) in which there are many objects. Can be done using object recognition, bar coding, or passive transponders. It requires a sensor, operating in conjunction with a computer that processes the sensed data and controls the movements of the robot.

binomial An algebraic expression containing two terms joined by a plus or minus sign. Examples: $a^2 + b^2$, $3x^3 - 6x$.

binomial theorem The theorem, proven by Isaac Newton, permits a binomial to be raised to any desired power without performing the multiplications. In electronics, *power series* are convenient for expressing such expressions.

biochemical cell A fuel-cell energy source in which electricity is generated chemically through the oxidation of biological substances. Also called *biochemical fuel cell.*

biochip **1.** A natural, living organism with a physical structure that in some way resembles that of an electronic *integrated circuit (IC).* **2.** A theoretical possibility, according to some scientists, but not yet a practical reality: An IC manufactured by a laboratory process that mimics the way in which nature builds living organisms. A form of *artificial life,* harnessed for electronic and/or computing applications.

bioelectricity **1.** Electric currents in living tissues, generated by the organism and not applied by external means. **2.** The science or study of such currents.

bioelectrogenesis The study and application of electricity generated by living animals, including humans, in the powering and control of electronic devices.

bioelectronics Electronics in relation to the life sciences—especially the electronic instrumentation of biological experiments.

bioengineering **1.** The engineering of equipment, such as electron microscopes, electroencephalographs, centrifuges, irradiators, etc., for study and experimentation in the life sciences. **2.** The engineering of equipment, such as pacemakers, hearing aids, X-ray apparatus, shock-therapy units, etc., for aid or support-of-life processes.

biofeedback A technique in which changes in skin temperature and resistance are detected and displayed by an electronic device.

biofeedback monitor A system that provides an indication of skin temperature and resistance to a user. Because skin temperature and resistance are affected by emotions, such as fear, nervousness, anger, etc., these monitors might be of value to people who wish to gain improved control of their emotions, and thus perhaps minimize the physiological effects of stress.

biological robot Believed by some researchers to be possible, but not yet a practical reality: A living organism created by biological cloning, whose brain has been programmed exactly as a computer is programmed.

biological shield An absorbent shield that blocks or attenuates ionizing radiation to protect personnel working near radioactive materials.

bioluminescence **1.** The emission of light by a living organism. **2.** The light itself so produced by living organisms.

biomechanism An electromechanical device that simulates the workings of some part of a living being's body. Examples are electromechanical hands, arms, and legs. Such a device is often difficult to distinguish from its biological counterpart when obscured by clothing.

biomechatronics A contraction of the words *bio*logy, *mecha*nics and elec*tronics*. Research, development and manufacturing that encompasses aspects of all three fields. This is especially important in robotics.

biometrics Mathematics, and in particular, statistics and probability, applied to biology.

biometric security system An advanced intrusion-prevention system that measures biological characteristics of the people who are authorized to enter a property. Such a machine can employ vision systems, object recognition, and/or pattern recognition to check a person's face. The machine might use speech recognition to identify people by the waveforms of their voices. It might record a hand print, a fingerprint, or an iris print, or a combination of all these things. A powerful computer analyzes the data obtained by the sensors and determines whether the person is authorized to enter the premises.

bionics The study, design, and application of microelectronic systems that simulate the functions of living organisms.

biotelemetry The use of telemetry to collect data from living organisms or to direct their movement.

biotelescanner An instrument that monitors body functions via radio, from a great distance.

Biot/Savart law A principle of electromagnetism that expresses the intensity of magnetic field H in the vicinity of a long, straight wire carrying a steady current I. The basic formula is $H = 2I/r$, where H is in oersteds, I is in amperes, and r is the distance in centimeters from the wire.

bip Abbreviation of *binary image processor.*

biphase half-wave rectifier An alternative term for FULL-WAVE RECTIFIER; also, each leg of a two-diode full-wave rectifier.

BIPM Abbreviation of *International Bureau of Weights and Measures.*

bipolar The condition of possessing two pole sets. In a conventional (non-FET) transistor, one pole set exists between the base and collector, and another pole set exists between the base and emitter.

bipolar driving unit A magnetic headphone or loudspeaker in which both poles (north and south) of a magnet actuate a diaphragm or lever.

bipolar operation See AUTOMATIC POLARITY.

bipolar transistor A two-junction transistor whose construction takes the form of a pnp or an npn "sandwich." Such devices are current-operated, compared with field-effect transistors, which are voltage-operated. The bipolar transistor (of which the familiar npn and pnp types are examples) uses both electron and hole conduction.

biquinary code A variety of binary-coded-decimal notation in which seven bits are used to represent each decimal digit. A number is written in two groups of bits: a two-bit group followed by a five-bit group. The positional values are 5 and 0 for the two-bit group, and 4, 3, 2, 1, and 0 for the five-bit group.

biquinary decade A decade counter that consists of a binary stage, followed by a quinary stage.

bird **1.** Slang for orbiting SATELLITE. **2.** Slang for *guided missile.*

birdie **1.** A spurious beat note in a superheterodyne receiver. So called because of the characteristic chirping sound it makes as the operator tunes by the frequency on which it occurs. **2.** A parasitic oscillation in a radio transmitter, also called a *spurious emission* or *spur.*

Birmingham wire gauge Abbreviation, BWG. Also called *Stubs gauge.* A method of designating the various sizes of solid wire. BWG diameters are somewhat larger than corresponding AMERICAN WIRE GAUGE diameters for a given wire-size designator.

Birmingham Wire Gauge (BWG) Diameters

BWG	Millimeters	Inches
1	7.62	0.300
2	7.21	0.284
3	6.58	0.259
4	6.05	0.238
5	5.59	0.220
6	5.16	0.203
7	4.57	0.180
8	4.19	0.165
9	3.76	0.148
10	3.40	0.134
11	3.05	0.120
12	2.77	0.109
13	2.41	0.095
14	2.11	0.083
15	1.83	0.072
16	1.65	0.064
17	1.47	0.058
18	1.25	0.049
19	1.07	0.042
20	0.889	0.035

bismuth Symbol, Bi. A metallic element. Atomic number, 83. Atomic weight, 209.

bismuth flux meter A flux meter in which the sensor contains a length of bismuth wire, which acts as a magnetoresistor.

bismuth thermocouple A thermocouple that uses the junction between bismuth and antimony wires. Used in thermocouple-type meters.

bistable Having two stable states.

bistable device Any device, such as a flip-flop, the operation of which exhibits two stable states and which can be switched at will from one state to the other.

bistable multivibrator A multivibrator, the operation of which exhibits two stable states. More commonly known as a FLIP-FLOP. These circuits are abundant in digital electronic equipment. Compare ASTABLE MULTIVIBRATOR and MONOSTABLE MULTIVIBRATOR.

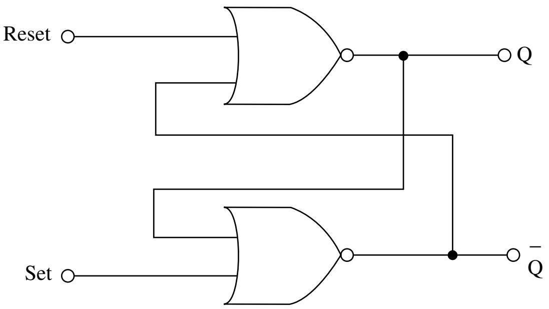

bistable multivibrator

bistable relay A relay that has two stable states: open and closed. Successive actuating pulses open and close the relay, two consecutive pulses being required to return the relay to a given state. Also called *binary relay*, *relay flip-flop*, and *electromechanical flip-flop.*

bistatic radar A radar set in which the transmitting and receiving antennas are separate.

bistate Having two states. Example: the performance of a FLIP-FLOP.

bit An acronym formed from the words *binary digit.* The smallest or elementary unit of data in digital electronics. Represented either by logic 0 (low) or logic 1 (high). These states can be represented by any dichotomy, such as off/on, false/true, minus/plus, dark/bright, red/green, etc.

BIT Abbreviation of *built-in test.*

bit density The number of digital bits per unit area or volume, as the number of bits per square centimeter of magnetic tape.

BITE Abbreviation of built-in test equipment.

bit rate The speed in BITS PER SECOND (bps) at which digital data bits are transmitted or handled.

bit-slice processor A microprocessor whose word or byte capacity is achieved through the use of interrelated smaller capacity processors (e.g., a 16-bit unit derived from eight 2-bit "slices").

bits per second Abbreviation, bps. An expression of digital data speed. Commonly used in computer communications. This unit is often confused with, and improperly called, the *baud.* There is generally a difference between the speed of a signal in baud, and the speed of the same signal in bps. Compare BAUD.

bitter pattern A pattern produced in a suspension of ferromagnetic powder in the presence of an imperfection in a magnet. The pattern appears as an irregularity that is easy to see.

Bjerknes' equation An expression for the total (primary plus secondary) decrement of a tuned circuit, based on measurements of the tank current at the resonant frequency and at a frequency near resonance.

BK **1.** Radiotelegraph signal for BREAK. **2.** Abbreviation of BREAK-IN.

Bk Symbol for BERKELIUM.

black-and-white Also called *monochrome* and *gray-scale.* Any system of image reproduction, transmission, or reception in which the image is composed of opaque elements (black) and white or bright areas, as in noncolor television reception.

black area An area in which there is only an encrypted signal.

blackboard system A method via which computers can recognize, and to some extent determine the meaning of, spoken words and visual images. Incorporates machine vision and/or machine hearing in conjunction with artificial intelligence (AI). Incoming voices and/or images are digitized and entered into a large-capacity random-access memory (RAM). The data is evaluated by sophisticated software to determine the most logical or probable interpretations of the sounds and images.

blackbody An ideal surface or object, that completely absorbs energy of any wavelength that strikes it. Such an object is a theoretically perfect radiator of energy at all wavelengths.

blackbody radiation Electromagnetic radiation from a heated ideal BLACKBODY. This radiation is conceived as covering the entire ELECTROMAGNETIC FREQUENCY SPECTRUM. It can be expressed graphically as a characteristic curve with a peak at a wavelength that depends on the absolute temperature of the object. As the absolute temperature increases, the peak occurs at progressively shorter wavelengths (higher frequencies). This enables radio astronomers to get a reasonably good idea of the temperatures of distant celestial objects, such as planets.

black box **1.** Any "box" or "block" that can be included in an analysis or synthesis based upon the BLACK-BOX CONCEPT. **2.** Any functional unit (such as a module) whose operating characteristics are known, and that can be inserted into a system in development or maintenance operations. **3.** Any subcircuit or stage that can be specified in total as required in a system, in terms of its known or prescribed performance, but whose internal structure need not be known.

black-box concept A technique for development of equivalent circuits and of considering their operation. The "box" has a pair of input terminals and a pair of output terminals; one input terminal is often common to one output terminal. The contents of the box need not be known, but from the input and output current and voltage relationships, its nature can be determined. Moreover, from the available input signal and desired output signal, the internal circuit of the box can be specified. Integrated circuits (ICs) are often treated as black boxes by engineers designing complex electronic equipment.

black compression Attenuation of the level of dark areas in a television picture.

blacker than black The video-signal amplitude region above the level that just darkens the screen. Signal information (such as control pulses) in this region are therefore not seen.

black light **1.** Ultraviolet radiation—especially when used to cause visible fluorescence in certain materials. **2.** A lamp that produces a principal portion of its radiation in the ultraviolet region, causing visible fluorescence of certain substances. Such lamps are used in some scientific experiments, and also for creating special effects at presentations or parties. It is hazardous to look directly at the output of such a lamp with unprotected eyes.

blackout **1.** A complete interruption of ac utility power to numerous customers at the same time. **2.** A complete cessation of ionospheric radio-wave propagation, such as might be caused by a solar flare. **3.** Complete blanking of the screen of an oscilloscope or picture tube.

black reference In a television signal, the blanking level of pulses, beyond which the sync pulse is in the *blacker-than-black* region.

black reference level In a television signal, the voltage threshold of the BLACK REFERENCE (i.e., its level above zero volts).

black transmission A system of picture or facsimile transmission in which the maximum copy darkness corresponds to the greatest amplitude (in an amplitude-modulated transmitter) or the lowest instantaneous frequency (in a frequency-modulated transmitter). Compare WHITE TRANSMISSION.

blank **1.** A piezoelectric plate cut from a quartz crystal, but not yet finished to operate at a desired frequency. **2.** To obscure or interrupt a signal or electron beam (usually momentarily), as in z-axis blanking in an oscilloscope. **3.** A silicon wafer cut from a large slab, containing dopants only. **4.** A magnetic diskette or tape on which nothing is recorded. **5.** An optical diskette on which nothing is recorded. **6.** A location (such as a symbol or space) that is used to verify proper data character grouping and values.

blanketing A form of radio interference accompanied by severe degradation of reception, virtually unaffected by tuning, over a wide range of frequencies. An example is ac line noise caused by an arcing power transformer or electrical appliance in the vicinity of a receiving antenna. It tends to occur most often at low, medium, and high frequencies.

blanking Obscuring or momentary elimination of a signal (see BLANK, **2**).

blanking interval The short period during which the electron beam of a cathode-ray tube is cut off so that the beam can return to its start position without creating a trace on the screen.

blanking level The discrete, predetermined level (usually a threshold voltage) at which BLANKING occurs.

blanking pedestal In the horizontal pulse of a television signal, the lower portion between zero volts and the blanking level.

blanking pulse A pulse that produces momentary blanking (see BLANK, **2**).

blanking time The time interval during which the electron beam of a cathode-ray tube is interrupted by a blanking signal.

blank tape **1.** Magnetic tape that has never been subjected to the recording process and that is substantially free from noise. **2.** Magnetic tape from which all preexisting information has been erased.

blasting **1.** Severe overloading of a sound system, usually caused by setting the volume control at or near maximum and then applying a significant input signal to the amplifier. Accompanied by distortion, in its worst form, it can cause damage to speakers and/or headsets. **2.** In a communications receiver, the result of a strong signal coming in unexpectedly when the automatic gain control (AGC) has been switched off, and the audio-frequency (AF) and radio-frequency (RF) gain controls are set high for reception of weak signals.

bleeder A resistor or group of resistors, used permanently to drain current from charged capacitors. It establishes the predetermined initial load level for a power supply or signal source, and it serves as a safety device in high-voltage power supplies.

bleeder current The current normally flowing through a bleeder.

bleeder divider A network of resistors, series-strung across the output of a power supply or its regulator. As a load resistor, the bleeder improves regulation and protects against no-load voltage surges. The resistor junctions allow various voltages to be drawn from the supply.

bleeder power Power dissipated as heat in a bleeder.

bleeder resistor See BLEEDER.

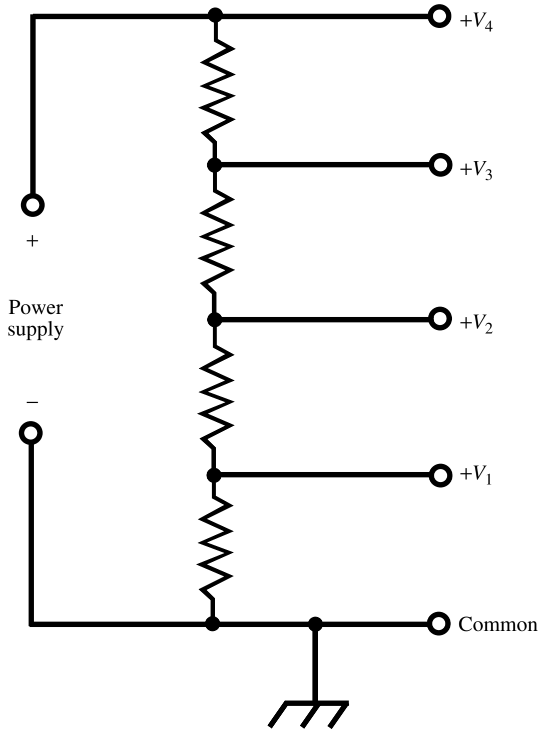

bleeder divider

bleeder temperature The operating temperature in a bleeder. It is generally high because of power dissipation in the form of heat.

bleeding whites A flowing of the white areas of a television picture into the black areas; an overload condition.

blemish See BURN.

blind flight The flying of aircraft entirely by means of instruments and electronic communications.

blind landing Landing of an aircraft entirely by means of instruments and electronic commnications.

blind zone **1.** In radar operations, an area that gives no echoes. **2.** Skip zone (see ZONE OF SILENCE).

blip **1.** The pulse-like figure on a radar scan, indicating the transmission or reflection (see A-SCAN and J-SCAN). Also called PIP. **2.** In visual alignment of a tuned circuit using a sweep generator and marker generator, the pulse or dot produced on the response curve by the marker signal. **3.** A short, momentary signal pulse, such as a single Morse dot.

BLIP Abbreviation for *background-limited infrared photoconductor.*

blip-scan ratio The number of radar scans necessary to show a visible blip, or echo, on a radar screen.

Bloch functions Solutions of the *Schrodinger wave equation* for a single electron surrounded by an electric field. The field varies periodically with distance from the source.

Bloch wall The transition layer between adjacent ferromagnetic domains (see DOMAIN).

block **1.** A group of data words or digits. **2.** A group of memory storage spaces. **3.** A circuit that operates as an identifiable unit. **4.** The symbol for a circuit, stage, unit, or device in a BLOCK DIAGRAM.

block diagram A simplified diagram of an electronic system, in which circuits, stages, units, or devices are shown as two-dimensional boxes with the internal wiring and detail circuitry omitted. This makes it possible to clearly show the interconnection among circuits, stages, units or devices. It also provides a concise rendition of the overall functional concept of the system.

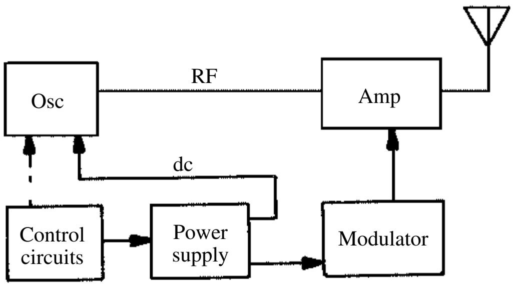

block diagram
(of a radio transmitter)

blocked impedance The input impedance of a transducer, whose output load is a theoretically infinite impedance.

blockette In a computer, the subdivision of a character block that is handled as a unit during data transfer.

blocking action Obstruction of circuit action, usually abrupt, through internal action or by the application of an external signal. Thus, the operation of an amplifier can be blocked (output reduced to zero) by an input signal or by excessive feedback, either of which overloads the input.

blocking capacitor A capacitor inserted into a circuit to prevent the passage of direct current while easily passing alternating current.

blocking choke Any inductor, such as a choke coil, that is used to prevent the flow of an alternating current while allowing direct current to pass with little resistance.

blocking interference Radio interference from signals strong enough to reduce the receiver output through blocking action.

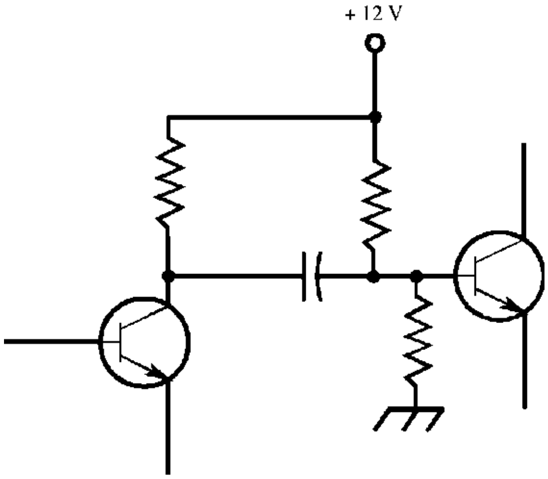

blocking capacitor

blocking oscillator An oscillator that turns itself off after one or more cycles. It does this as a result of an accumulation of negative charge on its input electrode (base of a bipolar transistor or gate of a field-effect transistor). The action is repetitive. In the self-pulsing type of blocking oscillator, a series of pulses consisting of trains of sine waves with intervening spaces is generated. In the single-swing type of blocking oscillator, the output consists of a series of single cycles with long intervals between them.

blocking oscillator synchronization **1.** In the BLOCKING OSCILLATOR used in the vertical deflection circuit of a television receiver, the oscillator is synchronized with vertical sync pulses arriving in the video signal. **2.** Synchronization of the repetition rate of any blocking oscillator with a suitable external control signal.

blocking system In a telephone system, a method of dealing with the condition of having more subscribers than connection paths. Allocation is made on a demand basis. If all channels are in use, it is impossible to make new calls. This prevents excessive degradation of the quality of existing connections.

block length The number of characters, bits, or words that compose a defined unit word or character group.

block transfer The conveyance of a word or character grouping in a computer register to another register or a peripheral device.

blooming On a cathode-ray-tube (CRT) screen, an enlargement of the electron-beam spot, caused by poor focusing. This results in poor image resolution.

blooper **1.** A radio receiver that is in oscillation, and is transmitting a signal that causes interference. **2.** A parasitic oscillation in a radio transmitter. **3.** In broadcasting, a statement in which a radio or television announcer makes an embarrassing error or breach of etiquette.

blow The opening of a fuse or circuit breaker as a result of excessive current.

blower A fan used to remove heat from electronic circuits. These are often used in tube-type radio-frequency (RF) power amplifiers, where much heat is generated, and in computers to cool the microprocessor and surrounding components.

blowout **1.** An alternate term for BURNOUT. **2.** The forceful opening of a circuit breaker. **3.** The extinguishing of an arc.

blowout coil An electromagnet that provides a field to extinguish an arc.

blowout magnet A permanent magnet that provides a field to extinguish an arc.

blst Abbreviation of *ballast.*

blue-beam magnet In a color television picture-tube assembly using three electron guns, a small permanent magnet to adjust the static convergence of the beam for blue phosphor dots.

blue box An accessory device (sometimes unlawfully used) that generates tones that switch a telephone circuit in the placing of calls.

blue glow **1.** In a neon lamp, a bluish light that results from high-voltage arcing. **2.** The normal color of the gas discharge in an argon glow lamp. **3.** The bluish glow between anode and cathode of a gassy vacuum tube. **4.** The normal color of the discharge that fills a mercury-vapor tube.

blue gun The electron in a three-gun color picture tube, the beam from which strikes the blue phosphor dots.

blueprint **1.** A type of contact-print reproduction in which a sheet of sensitized paper is exposed to an image on a translucent or transparent film, under strong light, and is then developed and fixed. Although this process is still used to reproduce electronic illustrations and typescripts, it has been superseded largely by other (dry) processes. **2.** Loosely, any plan or design for the development of a system.

blue restorer In a three-gun color television circuit, the dc restorer in the blue channel.

blue ribbon program A computer program that has been hand-prepared and debugged completely before its first computer run.

blue video voltage The signal voltage presented to the grid of the blue gun of a three-gun color picture tube.

blurring **1.** BLOOMING. **2.** A defocusing of a television picture or oscilloscope trace. **3.** An obscuring of a signal by echoes or trailing (e.g., the slow decrement of a Morse code signal element).

B-minus Also called *B-negative.* The negative terminal of a B-power supply.

BNC Abbreviation of *bayonet Neill-Concelman.* A type of coaxial connector that can be quickly con-

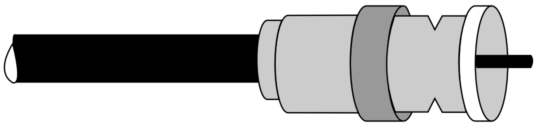

BNC

nected and disconnected. It is commonly used with test equipment.

B-negative Alternative expression for B-MINUS.

BNL Abbreviation of *Brookhaven National Laboratory.*

BO Abbreviation of *beat oscillator.* Also abbreviated BFO.

board **1.** A panel containing patch jacks. **2.** A printed circuit.

boat A type of crucible in which a semiconductor material is melted and sometimes processed. The material of which the boat is made (e.g., graphite) does not react with or contaminate the semiconductor material.

bobbin **1.** A usually nonmetallic spool on which a coil is wound. **2.** The form onto which the voice coil of a loudspeaker is wound.

Bode plot A pair of curves plotted to the same frequency axis, one showing the gain of a network or amplifier and the other showing its phase shift. Phase and amplitude of active and passive networks can be exhibited. Also called *Bode curve* and *Bode diagram.*

body-antenna effect The tendency of the human body to act as a receiving antenna when a finger is touched to the antenna input terminal of a receiver or when a hand (or the whole body) is brought close enough to the circuit to provide capacitive coupling.

body capacitance Capacitance between the body of the operator (as one plate of an equivalent capacitor) and a piece of electronic equipment (as the other plate). This phantom capacitance is often the cause of detuning and of the injection of interfering signals and noise because the body acts as a pickup antenna.

body electrode **1.** An electrode attached to the human body (or to the body of a laboratory animal) to conduct body-generated currents to an instrument, as in cardiography, electroencephalography, and myography. **2.** An electrode attached to the human body (or to the body of a laboratory animal) to conduct currents into the body, as in shock therapy and skin-resistance measurement.

body leakage Leakage of current through the bulk or body of a dielectric material, as opposed to SURFACE LEAKAGE.

body temperature In a thermistor, a rating that represents the temperature measured on the surface of the device. It is any combination of ambient temperature, power dissipation, and operation of the internal heater element (if the thermistor has one).

bof Abbreviation of *barium oxide ferrite.*

boffle A loudspeaker enclosure consisting of stretched screens that are sound absorbing and elastic.

bogie Also called *bogey.* **1.** The exact value of a specified characteristic. Thus, if resistance is given as 1 kΩ ±0.5%, the bogie value is 1 kΩ. **2.** The average value (i.e., the ARITHMETIC MEAN). **3.** A false or unidentified echo on a radar screen.

Bohr atom The concept of the nature of the atom, proposed by Niels Bohr in 1913 partly to explain why the electrons in the Rutherford atom do not fly off into space or fall into the nucleus. The Bohr theory places the electrons in permissible orbits where they cannot radiate energy (see BOHR RADIUS). They can radiate or absorb energy, however, if they go to a lower orbit or to a higher orbit, respectively. Compare RUTHERFORD ATOM.

bohrium Symbol, Bh. Also called *unnilseptium* (Uns). Atomic number, 107. The most common isotope has atomic weight 262. Classified as a transition metal. It is human-made and is not known to occur in nature.

Bohr radius Symbol, a_0. A physical constant whose value is approximately 5.291772×10^{-11} meter.

boiling point Abbreviation, bp. The temperature at which a liquid vaporizes. The boiling point of water in air at a pressure of one atmosphere is 100°C or 212°F.

bolometer Any device that is essentially a small, nonrectifying, temperature-sensitive resistor that can be used for heat sensing, radio-frequency power measurement, curve changing, demodulation, circuit protection, etc. Included in this category are the BARRETTER, the THERMISTOR, and the wire-type FUSE.

bolometer bridge A dc bridge in which a bolometer is one of the four arms. The bridge is balanced first with the bolometer cold. The bolometer then is excited with a radio-frequency (RF) current, whereupon the resultant heating changes the bolometer resistance. The bridge is rebalanced for the new resistance. The RF power driving the bolometer is determined according to a predetermined function of bridge settings versus RF input power.

Boltzmann constant Symbol, k. A figure that enters into the calculation of thermionic emission and of thermal noise factor. It represents the temperature equivalent of work function, in electron volts per Kelvin (eV/K) or joules per Kelvin (J/K). The values are approximately:

$$k = 8.617 \times 10^{-5}\ \text{eV/K} = 1.38 \times 10^{-23}\ \text{J/K}$$

Boltzmann's principle A description of the statistical distribution of large numbers of tiny particles under the influence of a force, such as an electric or magnetic field. When the system is in statistical equilibrium, the number of particles in any portion of the field is given by:

$$N_{\mathrm{E}} = N_{0e}^{-E/kT}$$

where E is the potential energy of a particle in the observed area, N_0 is the number of particles per unit volume in a part of the field where E is zero, k is the BOLTZMANN CONSTANT, T is the absolute temperature of the system of particles, and e is approximately equal to 2.718.

bombardment The usually forceful striking of a target with rays or a stream of particles.

bond **1.** An area in which two or more items are securely and intimately joined. **2.** The attractive force that holds an atomic or subatomic particle or particle group together.

bonded-barrier transistor A bipolar transistor in which the connection at the base region is alloyed.

bonded negative-resistance diode A diode that displays a negative-resistance characteristic over part of its current curve. This results from avalanche breakdown.

bond energy In a molecule, the energy necessary to break an atomic bond.

bonding **1.** The formation of bonds between adjacent atoms in a crystalline material, such as a semiconductor. See specifically COVALENT BINDING FORCES, IONIC BINDING FORCES, and METALLIC BINDING FORCES. **2.** The secure fastening together of conducting surfaces, as by soldering or brazing, to produce a high-conductance, leak-free continuum.

bond strength The minimum stress required to separate a material from another to which it is bonded.

bone-conduction transducer A device used in place of the earphone in a hearing aid to convey sound energy to the bone structure of the head.

Bongard problem A method of evaluating how well a machine vision system can differentiate among patterns. Similarities and differences are noted between objects in two sets of boxes. It was developed for object-recognition systems, mainly for use in intelligent robots.

book capacitor A variable capacitor in which the metal plates are bonded along one edge and separated from each other by means of mica sheets. The capacitance is varied by opening and closing the assembly book fashion. It is used as a padder or trimmer.

Boolean algebra A system of symbolic logic. Statements are represented as symbols, usually variables such as *x*, *y*, and *z*. The logical AND operation is represented by multiplication; the logical inclusive OR operation is represented by addition; the logical NOT operation is represented by a minus sign or a line over the element symbol. The system has rules, definitions and axioms via which *theorems* can be derived. Used by engineers in the design of digital electronic circuits.

Boolean truth table

		AND	*NOT*	*OR*
x	*y*	*xy*	*x*	*x* + *y*
0	0	0	1	0
0	1	0	1	1
1	0	0	0	1
1	1	1	0	1

Boolean function In mathematical logic, a function that makes use of BOOLEAN ALGEBRA.

Boolean theorems

1. $x + 0 = x$ (additive identity)
2. $x1 = x$ (multiplicative idenity)
3. $x + 1 = 1$
4. $x0 = 0$
5. $x + x = x$
6. $xx = x$
7. $(x')' = x$ (double negation)
8. $x + x' = 1$
9. $x'x = 0$
10. $x + y = y + x$ (commutativity of addition)
11. $xy = yx$ (commutativity of multiplication)
12. $x + xy = x$
13. $xy' + y = x + y$
14. $x + y + z = (x + y) + z = x + (y + z)$ (associativity of addition)
15. $xyz = (xy)z = x(yz)$ (associativity of multiplication)
16. $x(y + z) = xy + xz$ (distributivity)
17. $(x + w)\,(y + z) = xy + xz + wy + wz$ (distributivity)

boom **1.** A horizontal support for a microphone, enabling the microphone to be suspended over a sound source, but out of the sight of a camera. **2.** A horizontal support for a small antenna that is undergoing tests or sampling the field of another antenna. **3.** The supporting element in a Yagi, quad, or log-periodic antenna. It establishes the center of gravity and directional axis of the radiation pattern. The driven element(s) and parasitic element(s) are attached, usually at right angles.

boost capacitor In the damper circuit of a television receiver, the capacitor that is used to boost the B-plus voltage. Also called *booster capacitor.*

boost charge A high-current, short-interval charge used to revitalize a storage battery quickly. Also called *booster charge.*

booster **1.** Any device used to increase the amplitude of a signal (e.g., as an amplifier or preamplifier) or of an energy source (e.g., to boost the output of a power supply). **2.** A radio-frequency preamplifier used ahead of a television receiver.

booster battery **1.** A battery used to forward bias a diode detector into a favorable region of its conduction curve, or to bias a bolometer into the square-law region of its response. **2.** A battery supplying power to a booster.

booster gain The amplification (usually in terms of voltage gain) provided by a booster (see especially BOOSTER, **2**).

boot **1.** The powering-up routine in a digital computer, in which the machine executes a series of programs to get itself ready for use. **2.** The resetting of a computer, by pressing certain keyboard keys (e.g., CTRL-ALT-DEL), pressing a reset button, or by powering-down, waiting about two minutes, and then powering-up again. **3.** To install a computer diskette and instruct the computer to execute one or more programs on the diskette. **4.** A usually flexible protective nipple or jacket pulled over a cable or connector, so called from its resemblance to a foot boot.

boot loader A form of computer program that operates on the BOOTSTRAP ROUTINE.

bootstrap A technique for making a device or process achieve a condition through its own actions; see BOOTSTRAP CIRCUIT, for example.

bootstrap circuit A specialized form of follower circuit that presents very high input impedance. Its chief feature is the return of the control-element resistor to a tap on the source or emitter resistor. The technique takes its name from the figurative notion that such a circuit "lifts its input impedance by its own bootstraps."

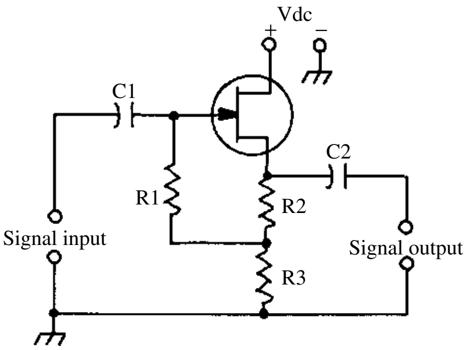

bootstrap circuit
(with junction-type field-effect transistor)

bootstrap routine **1.** Also called *bootstrap program.* In a digital computer, and especially in a personal computer, the routine that the machine follows when first powered-up. See BOOT, **1.** **2.** In a digital computer, a routine in which the first few instructions put in storage are later used to complete the routine, as supplemented by some operator instruction. **3.** A portion of a computer program that is used to establish an alternate version of the program.

borax-aluminum cell An electrolytic cell that consists essentially of an aluminum electrode and a lead electrode in a saturated solution of sodium tetraborate (borax). After electroforming, such a cell can be used either as a rectifier or as an electrolytic capacitor.

boric acid Formula, H_3BO_3. A compound used variously in electronics—especially as the electrolyte in electrolytic capacitors.

bornite Formula, Cu_5FeS_4. A natural mineral that is a sulfide of copper and iron. Its crystalline structure made it important in early semiconductor diodes (crystal detectors).

boron Symbol, B. A metalloidal element. Atomic number, 5. Atomic weight, 10.82. It is used as a dopant in semiconductor processing.

bot **1.** Abbreviation for *beginning of tape.* **2.** Abbreviation of *bottom.*

bottoming Excessive movement of the cone of a loudspeaker or the diaphragm of a headphone so that the magnet or supporting structure is struck by the moving-coil piston assembly. It produces a clapping sound, particularly on bass (low-frequency) audio peaks.

bounce **1.** The springback or vibration of the armature of a relay on closure. **2.** An abnormal, abrupt change in the brightness of the image in a television receiver or cathode-ray-tube (CRT) computer monitor.

boundary **1.** In a polycrystalline substance, the area of contact between adjacent crystals. **2.** The area of meeting of two regions (such as n and p) in a semiconductor.

boundary defect A condition in which a piezoelectric crystal has two regions, intersecting in a plane, with different polarizations.

boundary effect In audio systems, a pheno-menon in which the proximity of an acoustic transducer to a flat surface enhances the pickup and/or transmission of sound. Occurs because of reflection of acoustic waves from the surface.

bound charge The portion of the electric charge on a conductor that does not escape to ground when the conductor is grounded. This occurs because of induction from neighboring charge carriers. Compare FREE CHARGE.

bound electron An electron held tightly in its orbit within an atom so that it is not ordinarily free to drift between atoms and contribute to electric current flow.

bow-tie antenna A center-fed antenna in which the two horizontal halves of the radiator are triangular plates that resemble a bow tie. A flat reflector consisting of closely spaced horizontal wires is mounted behind the triangles.

bow-tie test An oscilloscope-display checkout of a single-sideband (SSB) signal, in which the appearance of the display indicates the signal quality. The transmitter output signal is fed to the vertical deflection plates of the oscilloscope. The exciter audio output is fed to the horizontal sweep input of the scope.

boxcars Long pulses with short separating spaces between them.

box-shaped loop The characteristic square-loop hysteresis curve (B-H loop) that result when a sine wave of current is used to magnetize a sample of magnetic material. In this plot, which covers all four quadrants, the horizontal axis (H) displays magnetizing force, and the vertical axis (B) displays magnetization. Also see HYSTERESIS.

Boys radiomicrometer A detector for radiant energy. The device consists of a thermocouple and a galvanometer. When energy falls on the thermocouple, a voltage is produced, and this is measured by the galvanometer.

bp **1.** Abbreviation of BOILING POINT. **2.** Abbreviation of BANDPASS.

bpi Abbreviation of *bits per inch.*

B-plus Also called *b-positive.* **1.** Symbol, B+. The positive dc voltage required for certain electrodes of vacuum tubes, transistors, etc. **2.** The positive terminal of a B power supply.

B positive See B-PLUS.

B power supply A name used sometimes for the unit that supplies high-voltage dc energy to a vacuum tube plate or screen circuit.

bps Abbreviation of BITS PER SECOND.

Br Symbol for BROMINE.

bracketing A troubleshooting routine characterized by isolating progressively smaller areas in a circuit or chain of stages until the defective subcircuit or stage is located.

Bradley detector A locked-oscillator circuit that was once used as an FM detector.

braid **1.** A woven network of fine metal wires used for grounding purposes. It is usually made of fine copper conductors. The increased surface-area-to-volume ratio improves the conductivity, at radio frequencies, over a single conductor that has the same cross-sectional area. Braid can be tinned (saturated with solder) to retard corrosion. **2.** It is also called a *shield.* The outer conductor in prefabricated coaxial cable.

braided wire A length of braid. Used for grounding or shielding purposes.

brain waves Alternating or pulsating voltages that are caused by electrical activity in the brain of an animal or human being. The voltages can be picked up by electrodes attached to the scalp, and amplified to be viewed on a cathode-ray-tube (CRT) screen, heard by headphones or speakers, or traced by an electroencephalograph.

branch **1.** Any one of the separate paths of a circuit. With respect to the layout of its components, a branch can be series, parallel, series-parallel, parallel-series, or any combination of these. It is also called a LEG. **2.** See BRANCH CIRCUIT.

branch circuit In electrical wiring, a group of outlets served through a single cutout from a source of power-line ac voltage. The source can be a distribution center, subdistribution center, main, or submain. Interior lighting circuits are usually branch circuits because many lights are connected to one circuit controlled by a single fuse or circuit breaker.

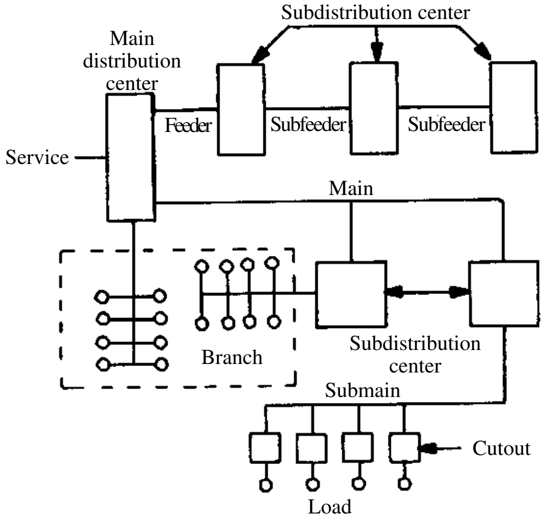

branch circuit (enclosed in broken lines)

branch current Current flowing through a branch of a circuit, whose magnitude, with respect to the total current of the circuit depends on the nature of the branch.

branched In molecular polymers, the condition of side chains being attached to the main chain.

branched windings Forked windings of a polyphase transformer.

branching In robotics and artificial intelligence (AI), a set of routines or programs containing points at which a computer must select from among two or more alternatives. Such routines are used in critical processes, such as the manufacture of precision equipment.

branch point See JUNCTION POINT.

branch voltage The voltage, or voltage drop, across a branch of a circuit.

brass **1.** An alloy of copper and zinc that is widely used in electronics. Compared to annealed copper, this metal has four times the resistivity (or ¼ the conductivity), half the temperature coefficient, more than twice the tensile strength, and a lower melting point (900°C). **2.** A colloquialism for an old-fashioned, straight telegraph key.

brass pounder **1.** Colloquialism for telegraph operator or radiotelegraph operator. **2.** A radio amateur who handles large amounts of message traffic, particularly via Morse code. **3.** A radio amateur proficient in Morse code operation.

Braun electroscope An electroscope consisting essentially of a fixed metal vane to which a movable needle is fastened at a pivot. The repulsion between the two, when an electric charge is applied, causes the needle to move over a calibrated scale.

bravo Phonetic representation of the letter B.

brazing The joining of two metal (usually iron or steel) parts together with a suitable melted copper-alloy metal. Compare SOLDERING.

breadboard **1.** A perforated board, a chassis, or any basic framework on which electronic components can be mounted and quickly wired for the preliminary test of a circuit. It is so called because the first such foundation units of this sort actually were wooden breadboards. **2.** Any preproduction electronic prototype circuit. **3.** To set up a circuit on a breadboard.

breadboard model **1.** The preliminary model of an electronic device, often built on a breadboard (see BREADBOARD, **1**). **2.** Loosely, any prototype.

break **1.** An open circuit. **2.** To open a circuit. **3.** In communications, a word indicating a desire to transmit on a wavelength already occupied by radio traffic. **4.** See BREAK-IN, **1**.

break-before-make contacts Contacts, especially in a rotary selector switch, that open one circuit before closing the next one.

breakdown **1.** Failure of a circuit or device, caused mainly by excessive voltage, current, or power. A sudden high current, however, does not always indicate failure. **2.** AVALANCHE BREAKDOWN. **3.** The separation of an electronics problem or project into its constituent parts for an easier solution.

breakdown diode See ZENER DIODE.

breakdown region The region, in a pn junction, in which avalanche breakdown occurs.

breakdown strength See DIELECTRIC STRENGTH.

breakdown voltage **1.** The voltage at which current suddenly passes in destructive amounts through a dielectric. **2.** The voltage at which a gas suddenly ionizes, as in a gas tube. **3.** The voltage at which the reverse current of a semiconductor junction suddenly rises to a high value (nondestructive if the current is limited). See AVALANCHE BREAKDOWN.

break-in **1.** A technique of radio communication in which one station interrupts a transmission from another station, rather than waiting until the end of the latter's transmission. **2.** Also called *full break-in.* In a radio communications transceiver or transmitter/receiver combination, extremely rapid transmit/receive switching, approaching *full duplex* communications. Every pause in transmission, even of only a few milliseconds, creates a "receive window" allowing reception between spoken words or Morse code elements. **3.** BURN-IN.

breaking current The momentary current that flows when the contacts of a switch or relay are broken.

break-in keying A system of radiotelegraph keying in which the receiver is in operation whenever the key is open. See BREAK-IN, **2.**

break-in operation In radiotelegraph or single-sideband (SSB) communications, the practice of interrupting at any time to "talk back" to the other transmitting station. This operation is made possible by high-speed transmit/receive switching. See BREAK-IN, **2.**

break-in relay An electromechanical or solid-state relay that enables break-in operation. Largely supplanted by solid-state switching devices.

breakover point In a silicon-controlled rectifier, the source-voltage value at which the load current is suddenly triggered to its steep climb. Also called TRIGGERING POINT.

breakover voltage In a silicon-controlled rectifier with open gate circuit, the anode voltage at which anode current is initiated.

breakpoint A point in a computer program when, for the purpose of obtaining information for the program's analysis, the sequence of operations is interrupted by an operator or a monitor program.

breakpoint frequencies The upper- and lower-frequency points at which the gain-versus-frequency response of an amplifier or network departs from flatness.

breakpoint instruction An instruction that stops a computer.

breakthrough **1.** A new discovery, insight, or solution to a problem that results in an advancement in the state of the art. **2.** See PUNCHTHROUGH. **3.** See BREAKDOWN, **1. 4.** See AVALANCHE BREAKDOWN.

break time The time taken for a relay to drop out completely or a switch to open. Compare MAKE TIME.

breathing Slow, rhythmic pulsations of a quantity, such as current, voltage, brightness, beat note, etc.

breezeway In a sync pulse in NTSC color television, the part of the back porch between the trailing edge of the pulse and the color burst.

B-register An index register in a computer for storing words that are used to change an instruction before it is executed by the program.

Bremsstrahlung radiation The radiation emitted by a charged particle whose speed is altered when it passes through the electric field in the vicinity of an atomic nucleus.

brevity code A code not intended to conceal information, but to shorten the number of characters in a message or data file. The Q SIGNALS are an example of a brevity code used in communications. In computer data transfer and communications, brevity codes allow compression, speeding up the transfer rate and reducing the storage space for a given amount of data.

Brewster angle From BREWSTER'S LAW, the polarizing angle at which the reflected and refracted rays of incident light are perpendicular to each other.

Brewster's law (Sir David Brewster, 1781–1868). For any dielectric reflector, the relationship in which the refractive index is equal to the tangent of the polarizing angle.

bridge **1.** A network, usually consisting of four branches, connected so that an input signal can be applied between two opposite points and the output taken between the other two opposite points. When the component values are in a certain ratio, the voltage between the output points is zero, and the bridge is said to be balanced or set to null. **2.** A circuit such as that described in (**1**) used for electrical measurements. **3.** An audio or servo amplification system in which the load is driven from two outputs having opposite polarity, neither of which are at ground potential. **4.** A communications path between or among two or more networks. This allows the subscribers in any network to obtain data from, or send data to, any other network, in effect creating a network of networks.

bridge balance control A potentiometer, variable capacitor, or variable inductor that is used to adjust a bridge circuit to balance.

bridge-connected amplifier **1.** A dc amplifier stage in which the transistors and resistors are connected in a four-arm bridge circuit, with respect to dc. When the bridge is initially balanced, all dc is eliminated in the output load. The input signal unbalances the bridge, which results in an amplified output signal in the load. **2.** An amplifier pair having opposing outputs across which a load can be bridged to obtain twice the power output of either amplifier alone.

bridged differentiator See HALL NETWORK.

bridge detector The output-indicating device (e.g., meter, oscilloscope, or headphones) that indicates whether a bridge is balanced or unbalanced. Also called *null detector* or *null indicator.*

bridged integrator A null network that consists of a two-stage resistance-capacitance (RC) integrator circuit bridged by a capacitor. This network produces a shallow null at a single frequency determined by the *R* and *C* values in the integrator. Compare HALL NETWORK.

bridged-tee attenuator An attenuator consisting of a tee section, between the input and output of which is bridged a single-series arm.

bridged-tee circuit Any circuit (of resistors, capacitors, inductors, or a combination of these) that consists of a tee section, bridged by a single-series section, from input to output.

bridged-tee null network A bridged-tee circuit of resistance (*R*) and capacitance (*C*), proportioned so that at some setting of the *R* and *C* values, the output of the circuit is zero.

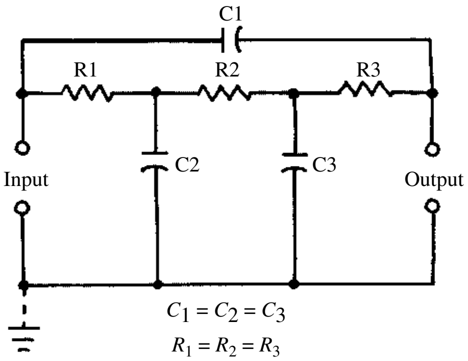

$C_1 = C_2 = C_3$

$R_1 = R_2 = R_3$

$f_{\text{null}} = \dfrac{1.732}{2\pi R_1 C_1}$

bridged integrator

bridged-tee oscillator A low-distortion oscillator circuit whose frequency is determined by a bridged-tee null network inserted into the negative-feedback path of the circuit.

bridge feedback A combination of current feedback and voltage feedback around an amplifier circuit. It is so called because, in the feedback circuit, the resistors and the output resistance of the amplifier form a four-arm bridge.

bridge generator The power source (e.g., a battery or oscillator) that supplies the signal to a BRIDGE used for electrical measurements.

bridge indicator See BRIDGE DETECTOR.

bridge oscillator See BRIDGE GENERATOR.

bridge rectifier A full-wave rectifier circuit in which four rectifying diodes are connected in a bridge configuration. Each half-cycle of ac input is rectified by a pair of diodes in opposite quarters of the bridge and in series with each other. The bridge does not require a transformer with a center-tapped secondary, as does the FULL-WAVE, CENTER-TAP RECTIFIER circuit.

bridge source See BRIDGE GENERATOR.

bridge-type meter A frequency-sensitive bridge (such as the Wien bridge) that can be used to measure audio frequency. Because the bridge can be balanced at only one frequency at a time,

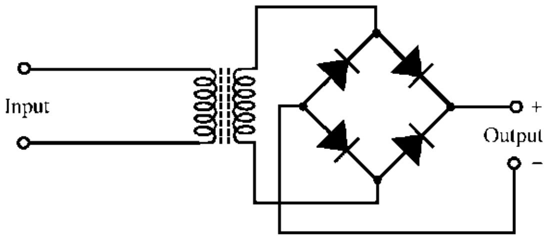

bridge rectifier

its adjustable arm can be calibrated to read the frequency directly.

bridge-type impedance meter An impedance-measuring circuit in which unknown impedance *Z* is connected in series with a calibrated variable resistor R. An ac voltage is applied to the series circuit. The separate voltage drops across the resistor and impedance are measured successively as the value of *R* is varied. When the two voltage drops are identical, *Z* equals *R*, and the impedance can be read from a calibrated dial on the variable-resistor control.

bridge-type oscillator A resistance-capacitance (RC) tuned oscillator in which a Wien bridge is used as the frequency-determining circuit in the feedback loop.

bridge-type power meter **1.** See BOLOMETER BRIDGE. **2.** A four-arm bridge specially designed to operate at radio frequencies. At null, the impedance of the unknown is read directly from the balancing dial or calculated from bridge constants. This instrument is used to measure the impedance of circuit components, antennas, and transmission lines.

bridge-type SWR meter A four-arm bridge that is specially designed to operate at radio frequencies. At null, the standing-wave ratio (SWR) is calculated from the bridge resistance values or read from a direct-reading scale on the null-indicating meter.

bridging amplifier An amplifier whose input impedance is so high that it can be considered infinite for practical purposes. Thus, the amplifier can be connected across a load or line without significantly affecting the operation of the system.

bridging coupler A voltage-dependent resistor that permits an occasionally used device (such as a bell) to be connected permanently across a regularly used device (such as a telephone) without continuously short-circuiting the latter. Thus, the bridging coupler ordinarily has very high resistance; but when the line voltage is momentarily raised, the resistance lowers and the occasionally used device is actuated (e.g., the bell rings).

bridging gain The gain of a bridging amplifier expressed as the ratio (in decibels) of the power developed in the amplifier load to the power in the load to which the input terminals of the amplifier are connected.

bridging loss The loss that results from the shunting of a speaker, microphone, earphone, or other transducer by a resistor, capacitor, or inductor. Generally, the loss is expressed as a power ratio in decibels.

Briggsian logarithm (Henry Briggs, 1556-1631). A base-10 logarithm, generally known as a COMMON LOGARITHM. Compare NAPIERIAN LOGARITHM.

brightness SI unit, candela per square meter (cd/m^2); cgs unit, lambert (L). The quantity of light, per unit area, emitted or reflected perpendicular to a light-emitting surface.

brightness control **1.** In a computer monitor, television receiver, or oscilloscope, a potentiometer that varies the negative bias voltage on the control grid of the cathode-ray tube (CRT). The brightness of the image is inversely proportional to this negative bias voltage. **2.** The control of the brightness of an illuminated area.

brilliance See BRIGHTNESS.

brilliance control **1.** The BRIGHTNESS CONTROL in a television receiver or computer monitor. **2.** The brightness control in a cathode-ray oscilloscope. **3.** A control for adjusting the level of the tweeter output in a speaker system.

British Standard wire gauge Abbreviation, NBS SWG. A classification of wire sizes sometimes used in England, Australia, and New Zealand. The higher the number, the thinner the wire. The designator does not take into account any coatings on the wire, such as enamel, rubber, or plastic insulation. In the United States, the *American wire gauge* is more often used. See AMERICAN WIRE GAUGE.

British Standard Wire Gauge (NBS SWG) Diameters

NBS SWG	Millimeters	Inches
1	7.62	0.300
2	7.01	0.276
3	6.40	0.252
4	5.89	0.232
5	5.38	0.212
6	4.88	0.192
7	4.47	0.176
8	4.06	0.160
9	3.66	0.144
10	3.25	0.128
11	2.95	0.116
12	2.64	0.104
13	2.34	0.092
14	2.03	0.080
15	1.83	0.072
16	1.63	0.064
17	1.42	0.056
18	1.22	0.048
19	1.02	0.040
20	0.91	0.036
21	0.81	0.032
22	0.71	0.028
23	0.61	0.024
24	0.56	0.022
25	0.51	0.020
26	0.46	0.018
27	0.42	0.0164
28	0.38	0.0148
29	0.345	0.0136
30	0.315	0.0124
31	0.295	0.0116
32	0.274	0.0108

NBS SWG	Millimeters	Inches
33	0.254	0.0100
34	0.234	0.0092
35	0.213	0.0084
36	0.193	0.0076
37	0.173	0.0068
38	0.152	0.0060
39	0.132	0.0052
40	0.122	0.0048

British thermal unit Abbreviation, Btu. The amount of heat required to raise the temperature of a pound of water by one degree Fahrenheit, in an ambient environment of slightly greater than 39°F.

broadband Also called *wideband.* Possessing a characteristic wide bandwidth or range of operating frequencies. This term can be applied at audio frequencies (AF) or radio frequencies (RF), and is frequently used to describe the performance of oscillators, amplifiers, antennas, and various types of networks. The term can also be applied to describe the nature of electromagnetic emissions or noise. Examples are given in the following several definitions. Compare NARROWBAND.

broadband amplifier An amplifier that has very wide frequency response, such as 10 Hz to 10 MHz. Examples are an instrument amplifier and a video amplifier.

broadband antenna An antenna that operates satisfactorily over a comparatively wide band of frequencies without requiring retuning at individual frequencies. Examples are the log-periodic and discone antennas.

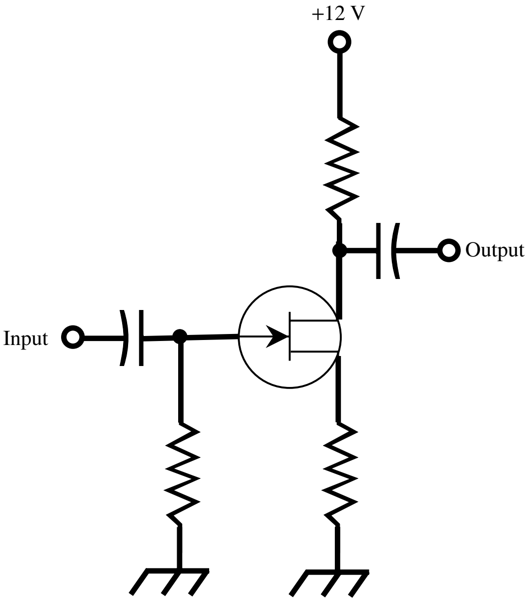

broadband amplifier

broadband electrical noise Electrical noise that is present over a wide frequency spectrum (e.g., 3 kHz to 30 MHz).

broadband I-F An intermediate-frequency (IF) amplifier or amplifier chain. The wide frequency response is important when an increased bandpass is preferred to high selectivity, as in high-fidelity radio tuners.

broadband interference Interference, other than noise, that is present over a wide band of frequencies. An example is over-the-horizon shortwave radar, recognizable by its characteristic “woodpecker” sound in communications receivers at high frequencies.

broadband Klystron A Klystron oscillator with a broadbanded tuned circuit.

broadband tuning Receiver tuning characterized by a selectivity curve having a pronounced flat top or broad nose that passes a wide band of frequencies. Also called *broadband response.*

broadcast **1.** A radio-frequency transmission of an intelligence-bearing signal that is directed to numerous unspecified receiving stations. **2.** The transmission or dissemination of signals to a large, unspecified number of receiving stations.

broadcast band Any band of frequencies allocated for broadcasting (see BROADCAST SERVICE, **1**), but particularly the U.S. standard amplitude-modulation (AM) and frequency-modulation (FM) radio broadcast bands at 535 to 1605 kHz (AM) and 88 to 108 MHz (FM).

broadcasting The dissemination of signals for reception by the general public, not for communications purposes.

broadcast interference Abbreviation, BCI. Interference to normal reception by broadcast receivers, usually arising from signals emitted by other stations.

broadcast receiver A receiver intended primarily to pick up standard broadcast stations. Also see BROADCAST BAND.

broadcast service **1.** Any radio transmitting service (including television) that exists for the purpose of sending out electromagnetic signals for general reception, rather than addressing them to specific receiving stations. **2.** The service provided by a station operating in the broadcast band.

broadcast station Any station in the broadcast service, but especially one assigned to operate in the standard U.S. broadcast bands. Also called *broadcasting station.*

broadcast transmitter A radio transmitter designed specifically for, and operated in, the broadcast service.

broad response Slow deflection of an indicator, such as a meter, over a relatively wide range of values of the input quantity.

broadside In a perpendicular direction; for example, broadside radiation from an antenna.

broadside antenna See BROADSIDE ARRAY.

broadside array Also called *broadside antenna.* A phased group of antennas arranged so maximum radiation occurs in directions perpendicular to the plane containing the driven elements. This requires that all of the antennas be fed in phase. The elements can be half-wave dipoles or full-wave, center-fed conductors. Full-wave elements have a slight gain over half-wave elements. At high frequencies, this type of array is usually constructed from two driven antennas. At very-high and ultra-high frequencies there can be several driven antennas. The antennas can each consist of a single element, or they can be Yagis, loops, or other systems with individual directive properties. In general, the larger the number of elements in the entire array, the greater the gain and directivity.

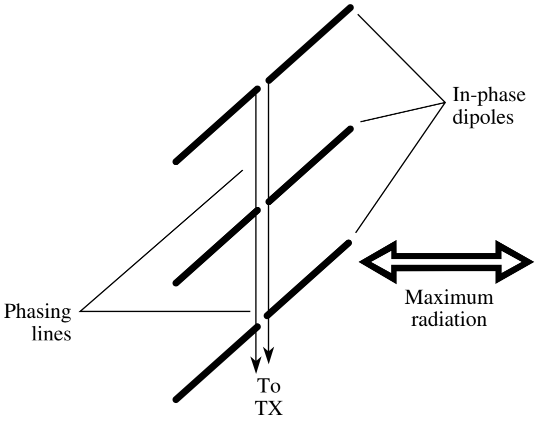

broadside array

broad tuning Tuning that is characterized by pronounced signal width, often resulting in adjacent-channel interference. A common cause of such impaired selectivity is low *Q* in the tuned circuit(s).

Broca galvanometer A device consisting of an astatic magnetic arrangement, with a coil enclosing central consequent poles. The device is characterized by fast response and high sensitivity.

bromine Symbol, Br. A nonmetallic element of the halogen family. Atomic number, 35. Atomic weight, 79.90.

bronze An alloy of copper and tin that has various uses in electronics. Also see PHOSPHOR BRONZE.

Brown and Sharpe gauge See AMERICAN WIRE GAUGE.

Brownian movement (Robert Brown, 1773–1858). Random movement of microscopic particles—especially in solutions. It occurs because of collisions of molecules with the particles. Einstein showed, in his early work, a connection between this movement and the *Boltzmann constant.*

brownout A deliberate lowering of line voltage by a power company to reduce load demands. Minor events of this type often pass unnoticed by the average consumer. More pronounced events produce observable effects, such as shrinkage of television and cathode-ray-tube (CRT) computer-display images.

Bruce antenna A vertical collinear array that consists of several resonant sections connected by short, rigid, parallel-conductor stubs. The currents in the radiating sections are in phase. Maximum radiation and response occur broadside to the antenna (omnidirectional in the horizontal plane). Polarization is vertical. The antenna produces gain at low radiation and response angles, and is commonly used in repeater installations and fixed communications stations at very-high frequencies (VHF) and ultra-high frequencies (UHF).

brush A usually metal or carbon strip, blade, or block, that slides in contact with another part, as in a motor commutator.

brush discharge Also called *Saint Elmo's fire.* A cloud of repelled ions around the tip of a pointed conductor charged to a high voltage. It often produces a visible glow in the air.

brush holder The housing for a brush in a motor, generator, rheostat, slip-ring junction in a rotating data-transmission system, etc.

brute force **1.** The transmission of a signal of excessive or unnecessary power. **2.** An inefficient approach to a problem, which might solve the problem, but requires far more energy, effort, or computer memory/storage space than the minimum needed to accomplish the same result.

brute-force filter A pi-type lowpass dc power supply filter, so called because of the extremely large inductances and capacitances that are generally used.

brute supply An unregulated power supply.

B-scope A cathode-ray tube (CRT), used in radar, that presents a B DISPLAY.

B service A teletype communication system operated by the Federal Aviation Administration (FAA).

B-supply The dc power supply that provides anode operating voltages, such as plate and screen voltages in a vacuum-tube radio-frequency (RF) power amplifier.

BT-cut crystal A piezoelectric plate cut from a quartz crystal at an angle of rotation (relative to the x-axis) of -49°. It has a zero temperature coefficient of frequency at approximately 25°C. Also see CRYSTAL AXES and CRYSTAL CUTS.

Btu Abbreviation of BRITISH THERMAL UNIT.

BuAer Abbreviation of *Bureau of Aeronautics.*

bubble memory In digital-computer practice, a special type of static magnetic memory. The magnetic material is divided into regions that are magnetized in different directions. So called because the flux lines of the tiny magnetized regions are shaped somewhat like, and move around after the fashion of, bubbles on the surface of a glass of soda.

bubble shift register A shift register that uses a magnetic bubble (see BUBBLE MEMORY) that can be moved sequentially from electrode to electrode on a wafer.

bubbling See MOTORBOATING.

bucket A computer memory or a designated location in such a memory.

bucking The process of counteracting one quantity, such as a current or voltage, via series or parallel application of a similar quantity that has opposite polarity (180 degrees out of phase).

bucking circuit **1.** A circuit used to obtain bucking action. The simplest form is a battery and potentiometer that supply a variable voltage of polarity opposite to that of the voltage to be bucked. A more sophisticated form is an ac transformer, the secondary of which is connected in series and out of phase with the ac utility line. **2.** The zero-set circuit in an electronic voltmeter.

bucking coil A coil placed and positioned so that its magnetic field partially or completely cancels the field of another coil. Troublesome hum fields sometimes are neutralized with such a coil.

bucking voltage See BACK VOLTAGE, **2.**

buckling The warping of storage-battery plates, usually resulting from excessive charge or discharge.

buckshot In an amplitude-modulated (AM) or single-sideband (SSB) radio transmission, broadband signal splatter caused by excessive modulation, or detuned multiplier circuits.

buffer **1.** An amplifier used principally to match two dissimilar impedance points and isolate one stage from a succeeding one in a cascaded system, and thus to prevent undesirable interaction between the two. **2.** In a digital computer, a storage site used temporarily during data transfers to compensate for differences in data flow rates. **3.** In digital-computer operations, a follower stage that is used to drive a number of gates without overloading the preceding stage.

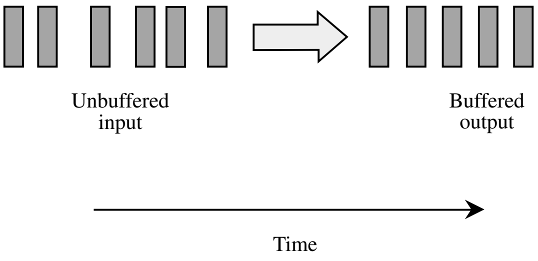

buffer, 2.

buffer amplifier See BUFFER, **1.**

buffer capacitor A high-voltage fixed capacitor that is placed across a transformer secondary to suppress voltage spikes and sharp waveforms—especially when the input is a square wave.

buffer circuit **1.** In a data system that uses a keyboard, an electronic circuit that allows the operator to type ahead of the data output. **2.** See BUFFER, **1**, **2** and **3.**

buffered output An output (power, signal, etc.) that is delivered from the generating device through an isolating stage, such as a buffer amplifier. This arrangement protects the device from variations in the external load. Compare UNBUFFERED OUTPUT.

buffer storage **1.** A buffer that is used to interface between data systems with different rates of transmission. **2.** See BUFFER, **2.**

bug **1.** Slang for WIRETAP, **1. 2.** Slang for circuit fault, **1**. **3.** A semiautomatic key that some radiotelegraph operators use to send Morse code.

bug key See BUG, **3.**

building-block technique The process of assembling electronic equipment by quickly connecting together already completed stages (in the form of boxes or blocks) and supplying power and signals to the setup. Also called *modular technique* and *modular construction.*

building-out circuit A short section of transmission line shunting another line; it is used for impedance matching. Also called *building-out section.*

buildup **1.** The process whereby the voltage of a rotating generator starts at a point that is determined by the residual magnetism of the machine, and gradually increases to a voltage representing the point at which the resistance line crosses the magnetization curve. **2.** The (usually gradual) accumulation of a quantity (e.g., the buildup of charge in a capacitor).

bulb A globe-like container having any of a number of characteristic shapes from spherical to tubular and usually evacuated, for enclosing the elements of an electron device, such as a vacuum tube, gas tube, photocell, or lamp.

bulge **1.** A nonlinear attenuation-versus-frequency curve in a transmission line. **2.** A localized nonlinearity in a function.

bulk The body or mass of a semiconductor specimen, as opposed to junctions within the specimen. Current flows through a junction, but it can also flow, more or less, through the mass of semiconductor wafer into which the junction has been formed.

bulk effect An effect, such as current, resistance, or resistivity, observed in the overall body of a sample of material, as opposed to a region within the material or on its surface. Thus, a silicon diode can display junction resistance (i.e., resistance offered by a junction processed in a wafer of silicon), as well as bulk resistance (i.e., the effective resistance of all paths around the junction, through the mass of the wafer). Compare SURFACE EFFECT.

bulk-erased tape Recording tape whose signal content has been removed via a bulk eraser.

bulk-erase noise **1.** The residual magnetic impulses that remain on a magnetic tape after it has been bulk-erased. **2.** Noise generated by bulk-erased tape when the latter passes through deenergized record or erase heads in a tape machine.

bulk eraser A type of power-line-frequency degausser that erases an entire reel of magnetic tape without requiring that the tape be unreeled and passed continuously under an erase head. This saves considerable time, but often leaves some BULK-ERASE NOISE on the tape. Also called BULK DEGAUSSER.

bulletin board In personal computing or amateur packet communications, a system that allows subscribers to leave messages for each other for access via a *modem* or *terminal node controller.*

bulletin station A station intended for the transmission of bulletins of interest to certain parties, such as military personnel or amateur radio operators. An example is W1AW in Newington, Connecticut, an amateur radio station that transmits bulletins and code practice.

buncher In a Klystron, a cavity resonator that contains two grids mounted parallel to the electron stream. The electrostatic field of the grids alternately accelerates and retards the electrons, velocity-modulating the stream into bunches.

buncher grids In a Klystron, the closely spaced grids that velocity-modulate the electron beam into successive bunches.

buncher resonator In a velocity-modulated tube, such as a Klystron, the input cavity resonator.

buncher voltage The radio-frequency (RF) grid-to-grid voltage in the buncher resonator of a Klystron.

bunching The production of electron bunches in a velocity-modulated tube, such as a Klystron. Also see BUNCHER.

bunch stranding A technique for combining several thin wires into a single thick wire. Often used in guy wires and electrical conductors to improve tensile strength and flexibility. At radio frequencies, bunch stranding also improves electrical conductivity by increasing the ratio of surface area to cross-sectional area. This minimizes losses caused by *skin effect.*

Bunet's formula A formula for calculating the inductance of a multilayer air-core coil that has a diameter less than three times the length:

$$L = a^2N^2/(9a + 10l + 8.4c + 3.2cl/a)$$

where N is the number of turns, a is the average coil radius, c is the winding thickness, and l is the length of the coil.

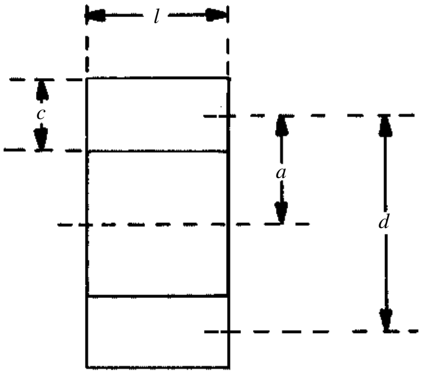

Bunet's formula

Bunsen cell A cell consisting of a zinc rod in a sulfuric acid solution contained in a porous pot, which is in a nitric acid solution. The zinc rod serves as the negative pole; the positive pole is a piece of hard carbon. The cell produces about 1.9 volts and delivers relatively high current.

burden See VOLTAGE BURDEN.

burn **1.** A blemish on the screen of a cathode-ray tube (CRT), caused by destruction of the phosphor there. This results from prolonged focusing of an intense electron beam in one spot. **2.** A blemish on the screen of a television picture tube, usually resulting from ions that reach the screen when the ion trap is not working correctly.

burn-in A long, thorough, carefully controlled preliminary test of a component, device, or system, to stabilize its electrical characteristics after manufacture and to ensure that it will function according to rated specifications. An important test for equipment whose reliability must be guaranteed, such as an emergency communications transceiver.

burnout **1.** Failure of a conductor or component caused by overheating from excess current or voltage. **2.** The open-circuiting of a fuse. **3.** Electrical failure of any type.

burst **1.** The abrupt ionization of the gas in an ionization chamber by cosmic rays. **2.** An abrupt increase in the amplitude of a signal. Also, the type of signal that results from burst action. **3.** See COLOR BURST.

burst amplifier In a color-television receiver, the amplifier that separates the burst pulse from the video signals and amplifies the former. See COLOR BURST.

burst gate timing In a color-television receiver, the timing of the gating pulse with the input signal of the burst amplifier.

burst generator A signal generator delivering a burst output (see BURST, **2**) for testing various types of equipment. Its output is intermediate between sine waves and square waves, and is convenient for rapidly appraising the performance of such devices as amplifiers, filters, electronic switches, transducers, and loudspeakers.

burst transmission A short transmission at high speed. This method of transmission saves time, but increases the necessary bandwidth of a signal by the same factor as the ratio of the high speed to the original speed.

bus **1.** A main conductor in a circuit. A bus can be high in the sense that its potential is above or below ground, or it can be low or at ground reference. **2.** In computer operations, a common group of paths over which input and output signals are routed.

bus driver A buffering device designed to increase the driving capability of a microprocessor, which itself might be capable of driving no more than a single load.

business machine Any piece of electronic or electromechanical equipment used mainly, or entirely, for business purposes. Examples are photocopiers, facsimile (fax) machines, printers, and computers.

busing The parallel interconnection of circuits.

busy test A check conducted to find out whether or not a certain telephone subscriber line is in use.

busy tone Also called *busy signal.* An intermittent tone that indicates that the subscriber line being called is in use.

Butler oscillator An oscillator that consists of a two-stage amplifier with a quartz crystal in the positive-feedback path from output to input.

butterfly capacitor A plate-type variable capacitor that has two stator sections and a single rotor section common to the two stators. External connections are made to the stators only. Thus, no wiping contact is required to the rotor, and the troubles associated with such a contact are avoided. The butterfly capacitor is actually two variable capacitors in series. The unit is so called from the shape of its rotor.

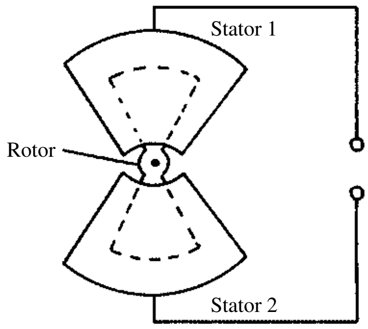

butterfly capacitor

butterfly circuit A combination of a butterfly capacitor and a ring, of which the stator plates of the capacitor are an integral part. The resulting structure is a compact variable-frequency tuned circuit. The ring supplies the inductance, and the butterfly supplies the capacitance. It is also called *butterfly tank* and *butterfly resonator.*

Butterworth filter A high-pass, low-pass, bandpass or band-rejection filter, characterized by a flat passband (absence of passband ripple) and high attenuation at frequencies far removed from the passband.

Butterworth function A mathematical function that is used in the design of a BUTTERWORTH FILTER.

button **1.** Usually, a small switch that is actuated by finger pressure. It is also called pushbutton and pushbutton switch. Sometimes, the term is applied only to the insulated knob or pin which is pushed to operate the switch. **2.** A tiny lump of impurity material, placed on the surface of a semiconductor wafer for alloying with the wafer to form a junction. See ALLOY JUNCTION. **3.** The carbon element(s) in a BUTTON MICROPHONE.

button capacitor A button-shaped ceramic or silvered-mica fixed capacitor. Because of its disk shape and mode of terminal connection, it offers very low internal inductance.

button microphone A microphone in which a button-shaped carbon element is attached to a diaphragm, which is set into vibration by sound waves. This motion causes the button resistance to vary, modulating a direct current that passes

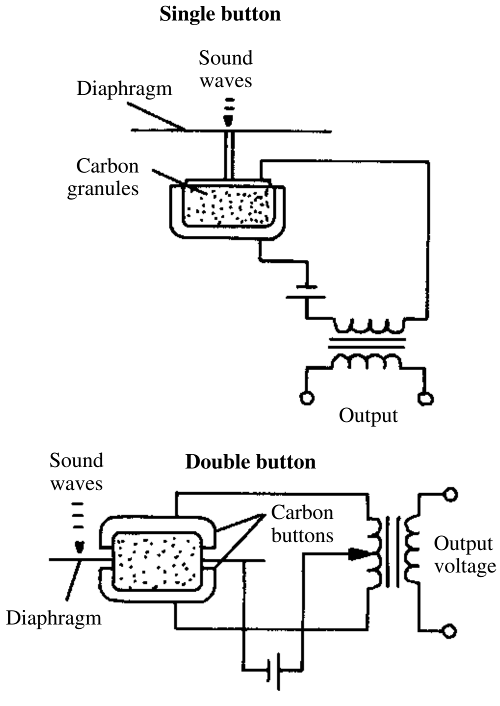

button microphone

through the button. A single-button microphone has only one button, whereas a double-button microphone has two—one mounted on each side of the center of the diaphragm.

buzz **1.** A low-pitched rough sound with high-frequency components, usually the result of electrical interference from nonsinusoidal voltages generated by neighboring equipment or devices. **2.** The waveform associated with such a sound. **3.** Fastening two conducting surfaces by a KELLIE BOND.

buzzer A nonringing device used principally to generate sound other than that achievable with sine waves. In an electromechanical vibrating-reed buzzer, the reed acts as an armature, which is mounted close to the core of an electromagnet. At quiescence, the reed rests against a stationary contact. When voltage is applied to the electromagnet, the reed is attracted to the core, moving away from the contact; but this breaks the circuit, the magnetism ceases, and the reed springs back to the contact. The action is repeated continuously at a frequency that depends on the reed dimensions and its distance from the core.

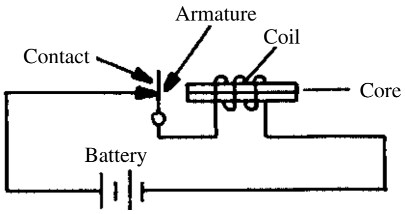

buzzer

BV Abbreviation of BREAKDOWN VOLTAGE.

B voltage The dc voltage required by certain electrodes of vacuum tubes and transistors. It especially pertains to voltages required by the plate and screen of a vacuum tube, as opposed to the filament voltage and control-grid voltage.

bw **1.** Abbreviation of *bandwidth.* 2. Abbreviation of *black-and-white.*

BWA Abbreviation of *backward-wave amplifier.*

BWG Abbreviation of BIRMINGHAM WIRE GAUGE.

BWO Abbreviation of BACKWARD-WAVE OSCILLATOR.

BX Symbol and abbreviation for armored and insulated flexible electrical cable.

bypass A route (either intended or accidental) through which current easily flows around a component or circuit instead of through it.

bypass capacitor A capacitor that is used to conduct an alternating current around a component or group of components. Often the ac is removed from an ac/dc signal, the dc being free to pass through the bypassed component.

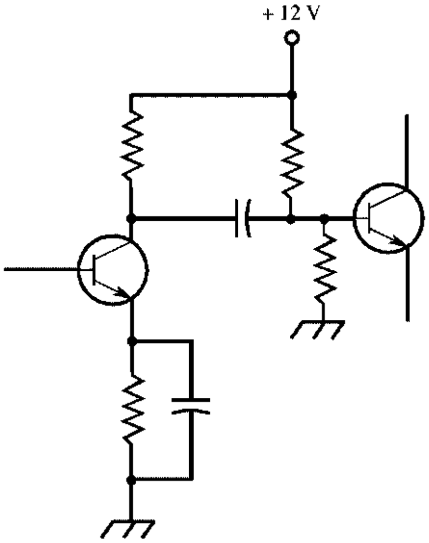

bypass capacitor

B-Y signal In a color television receiver, the color-difference signal which, when combined with a luminance (Y) signal, forms a blue primary signal for the three-gun picture tube.

byte In digital-computer and data-communications operations, a unit of data consisting of eight contiguous bits. In packet communications, the term *octet* is often used.

C

C **1.** Abbreviation of CAPACITANCE. **2.** Symbol for COLLECTOR of a transistor. **3.** Symbol for CARBON. **4.** Abbreviation of CELSIUS. **5.** Symbol for COULOMB. **6.** Abbreviation of CALORIE.

c **1.** Abbreviation of CENTI. **2.** Abbreviation of CENTS. **3.** Symbol for CAPACITANCE. **4.** Symbol for SPEED OF LIGHT in a vacuum.

Ca Symbol for CALCIUM.

cabinet An enclosure for a piece of apparatus. It might or might not incorporate electromagnetic shielding.

cable **1.** A usually flexible (but sometimes rigid) medium, via which electrical power or signals are transferred. Although the term is occasionally applied to a single conductor, especially when it is a braid or weave of a number of wires, cable usually means a bundle of separate, insulated wires or strands of fiberoptic material. **2.** CABLEGRAM.

cable address A code word that specifies the recipient of a CABLEGRAM.

cable assembly A special-purpose cable with connectors.

cable attenuation Reduction of signal intensity along a cable, usually expressed in decibels per foot, hundred feet, mile, etc.

cable capacitance Capacitance between conductors in a cable or between conductors and the outer sheath of a cable. **2.** Sometimes, capacitance between a cable and earth.

cable clamp A support device for cable runs in equipment and systems.

cable communications Telegraphy or telegraphy via a (usually undersea) cable.

cable connector A connector, such as a coaxial fitting, that joins cable circuits or connects a cable to a device.

cabled wiring Insulated leads connecting circuit points; they are tied together with lacing cord or with spaced fasteners.

cablegram A (usually printed) message transmitted or received via undersea cable. Compare RADIOGRAM and TELEGRAM.

cable loss See CABLE ATTENUATION.

cable run The path taken by a cable.

cable splice **1.** An electrical attachment between two sections of cable that has identical or similar construction, with or without the use of connectors. **2.** To electrically attach two sections of cable that have identical or similar construction, with or without the use of connectors.

cable tie A short piece of wire or plastic that holds wires or cables in a bundle.

cable TV See COMMUNITY-ANTENNA TELEVISION.

cache memory A short-term, high-speed, high-capacity computer memory. Similar to a scratchpad or read-write memory.

CAD Acronym for *computer-aided design.*

CAD/CAM Acronym for *computer-aided design and manufacturing.*

cadmium Symbol, Cd. A metallic element. Atomic number, 48. Atomic weight, 112.41. Many electronic structures are cadmium plated for protection.

cadmium borate phosphor Formula, ($CdO + B_2O_3$): Mn. A substance used as a phosphor coating on the screen of cathode-ray tubes. The characteristic fluorescence is green-orange.

cadmium cell Also called *Weston standard cell.* An electrochemical standard cell used as a reference voltage source. Produces 1.0186 volt at 20°C.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

cadmium plating The process of coating a conductor or component with cadmium to increase its resistance to corrosion.

cadmium selenide photocell A photoconductive cell in which cadmium selenide is the light-sensitive material.

cadmium silicate phosphor Formula, ($CdO + SiO_2$). A substance used as a phosphor coating on the screen of cathode-ray tubes; the characteristic fluorescence is orange-yellow.

cadmium standard cell See STANDARD CELL.

cadmium sulfide photocell A photoconductive cell in which cadmium sulfide is the light-sensitive material.

cadmium tungstate phosphor Formula, $CdO + WO_3$. A substance used as a phosphor coating on the screen of cathode-ray tubes; the characteristic fluorescence is light blue.

cage A completely shielded enclosure, such as a screen room, which is covered with a grounded fine-mesh conductive screen on all sides.

cage antenna An antenna, usually center-fed and balanced, that consists of multiple parallel conductors arranged in a cylindrical cage configuration. The cage results in a much broader bandwidth than is the case with an antenna made up of a single conductor. Cage antennas are typically used at frequencies between about 10 and 200 MHz.

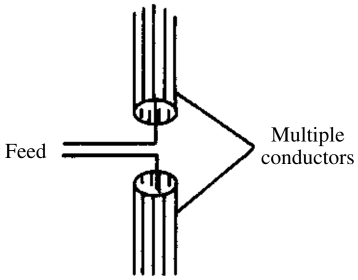

cage antenna

CAI Abbreviation for *computer-assisted instruction.*

CAL An acronym for *conversional algebraic language*, a general-purpose problem-oriented computer programming language used in time-sharing systems.

calcium Symbol, Ca. A metallic element of the alkaline-earth group. Atomic number, 20. Atomic weight, 40.08.

calcium phosphate phosphor Formula, $Ca_3(PO_4)_2$. A substance used as a phosphor coating on the screen of long-persistence cathode-ray tubes; the characteristic fluorescence is white, as is the phosphorescence.

calcium silicate phosphor Formula, ($CaO + SiO_2$): Mn. A substance used as a phosphor coating on the screen of cathode-ray tubes; the characteristic fluorescence ranges from green to orange.

calcium tungstate phosphor Formula, $CaWO_4$. A substance used as a phosphor coating on the screen of short-persistence cathode-ray tubes; the characteristic fluorescence is blue, as is the phosphorescence.

calculate To perform the steps of an intricate mathematical operation. Compare COMPUTE.

calculating punch A data-processing peripheral that reads punched cards, makes calculations, and punches new data into those cards or new cards.

calculator A machine that performs mathematical operations, especially arithmetic. Typically, the device is a small box with buttons and a miniature numeric display. Used only in mathematical applications. In contrast, a COMPUTER can be used for a much wider variety of jobs, such as word processing, graphics, and data-base. Many personal computers have calculator programs; the “buttons” are actuated by pointing and clicking with a mouse.

calculus **1.** The symbology and rules comprising a system of logic, such as BOOLEAN ALGEBRA. **2.** A branch of mathematical analysis concerned with rates of change and accumulation. See DIFFERENTIAL CALCULUS and INTEGRAL CALCULUS.

calendar age The age of a piece of equipment, measured since the date of manufacture. Specified in years, months, and days. The actual manufacture date might alternatively be given.

calendar time The time available in a working period [i.e., a 40-hour work week represents a calendar time of 120 hours (five days times 24 hours per day)].

calibrate To compare and bring into agreement with a standard.

calibrated measurement **1.** A measurement made with an instrument that has been calibrated with a standard reference source. **2.** A measurement that is corrected for instrument error.

calibrated meter An analog or digital meter that has been adjusted to agree as closely as possible with a reference source.

calibrated scale **1.** A scale whose graduations have been carefully checked for accuracy (i.e., they correspond to the true values of the quantity that they represent). The scale is graduated to read directly in units of the quantity, such as milliamperes, kilohertz, volts, etc. **2.** A scale with fixed, plain numeric graduations (e.g., 0 to 100) that do not directly indicate the magnitude of a quantity, but that can be converted to various quantities via graphs, nomographs, tables, or charts. See CALIBRATION CURVE.

calibrated sweep In an oscilloscope, a sweep circuit calibrated to indicate sweep frequency or time at all control settings.

calibrated triggered sweep In an oscilloscope, a triggered sweep circuit calibrated in terms of sweep time or frequency.

calibration **1.** Determining the accuracy with which an instrument indicates a quantity. **2.** Determining the degree to which the response of a circuit or device corresponds to desired performance. **3.** Marking a scale to show actual values of a quantity in the form of a direct readout. For example, the scale of an analog meter might be calibrated in milliamperes (mA) from 0 to 50 in increments of 1 mA.

calibration accuracy **1.** A quantitative expression of the agreement between the value of a quantity, as indicated by an instrument, and the true value. Usually expressed as the maximum percentage of the true value by which the indicated value can be expected to deviate in either direction (e.g., ±0.5 percent). **2.** The precision of a direct-reading meter in terms of its full-scale deflection (e.g., ±2.0 percent of full scale).

calibration curve A graph showing the relation between the actual values of a quantity and the setting or indication of an instrument or component. Usually plotted in rectangular coordinates.

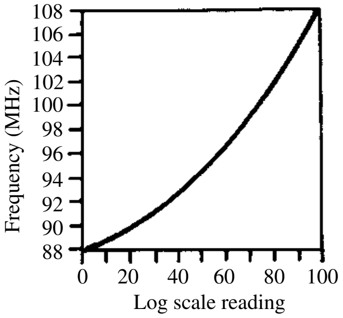

calibration curve

calibration marker A pip or blip, superimposed on a pattern displayed on a cathode-ray-tube (CRT) screen, to identify a point closely as to frequency, voltage, distance, or some similar term.

calibrator A device used to perform a calibration (e.g., a signal generator).

calibrator crystal A highly accurate and stable quartz crystal, used in an oscillator as a frequency standard. An example is the 100-kHz crystal oscillator and harmonic generator used in some communications receivers.

californium Symbol, Cf. A radioactive element produced artificially. Atomic number, 98. Atomic weight, 251.

call **1.** In communications, a transmission by a station for the purpose of either alerting a particular receiving station for which there is a message, or alerting all receiving stations to prepare them for a general broadcast message. **2.** In a computer program, a branch to a closed subroutine; also, to branch to such a subroutine.

call direction code Abbreviation, CDC. In telegraph networks, a special code that, when transmitted to a terminal, causes the teleprinter to be automatically switched on.

calling sequence **1.** Computer program instructions needed to establish the conditions for a call (see CALL, **2**). **2.** Subroutine instructions providing a link to the main program.

call instruction A computer program instruction that makes a program controller branch to a subroutine; it also locates and identifies the parameters needed for the subroutine's execution. Also known as *subroutine call.*

call letters Letters and/or numbers assigned to, and used to identify, licensed radio stations.

calorie Abbreviation, cal or C. The amount of heat energy, at a pressure of 1 atmosphere, that will raise the temperature of 1 gram of water by 1 degree Celsius.

calorimeter An instrument for measuring heat energy. By adaptation, a calorimeter can be used to measure radio-frequency (RF) power—especially at microwave frequencies (see CALORIMETRIC POWER METER).

calorimeter system See CALORIMETRIC POWER METER.

calorimetric power meter A specialized form of wattmeter, in which the power to be measured is dissipated in an oil or water bath that has a known and fixed mass. The wattage is determined indirectly, by measuring the extent to which the temperature of the liquid increases in a certain amount of time.

CAM **1.** Abbreviation of *computer-aided manufacturing.* **2.** Abbreviation of *content-addressed memory.*

cambric Finely woven cotton or linen used for insulation. One type of spaghetti (conductor insulation), for example, is varnished cambric tubing.

camera cable A multiwire cable that conducts the video signal from a television camera to control equipment.

camera chain In television, the camera and the equipment immediately associated with it, excluding the transmitter and its peripherals.

camera signal The output signal delivered by a television camera.

camera tube Any video pickup tube, such as an iconoscope or orthicon, that converts light reflected by a scene into a corresponding television signal.

Campbell bridge A circuit that is used for comparing mutual inductance with capacitance.

camp-on In a telephone system, a method of engaging a line that is busy until it becomes available for use.

CAN Abbreviation of CANCEL CHARACTER.

can **1.** A metal enclosure or container roughly resembling a tin can (though not necessarily cylindrical), used for shielding or potting components. **2.** Colloquial expression for HEADPHONE.

Canada balsam A transparent cement derived from the turpentine distilled from balsam fir resin. It is useful in optical technology and in certain areas of electro-optics.

Canadian Standards Association The Canadian equivalent of the *National Bureau of Standards* in the United States. An agency that publishes agreed-on standards for industries.

cancel character **1.** IGNORE CHARACTER. **2.** A control character indicating that the associated data is erroneous.

cancellation The elimination of one quantity by another, as when a voltage is reduced to zero by another voltage of equal magnitude and opposite sign.

candela Symbol, cd. The SI unit of luminous intensity; 1 cd represents 1⁄60 of the radiating power of one square centimeter of a perfect radiator at the temperature of freezing platinum.

candle Abbreviation, c. Also called *international candle.* A unit of light intensity that is the value of emission by the flame of a sperm-whale-oil candle burning at the rate of 7.776 grams per hour.

candle power Abbreviation, cp. Luminous intensity in international candles: the luminous intensity resulting from the burning of a sperm-whale-oil candle at 7.776 grams per hour.

candoluminescence White light produced without extreme heat.

cannibalization The deliberate use of parts from operational equipment to temporarily repair or maintain other equipment. It is a last-resort, emergency measure.

cap **1.** Abbreviation of CAPACITANCE. **2.** Abbreviation of CAPACITOR.

capacimeter See CAPACITANCE METER.

capacitance Symbol, C. Unit, farad. The property exhibited by two conductors separated by a dielectric, whereby an electric charge becomes stored between the conductors. Capacitance is thought of as analogous to mechanical elasticity. Also see FARAD.

capacitance bridge A four-arm ac bridge for gauging capacitance against a standard capacitor. In its simplest form, it has a standard capacitor in one arm and resistors in the other three.

capacitance coupling The transfer of ac energy between two circuits or devices by a capacitor or capacitance effect. Also see COUPLING.

capacitance diode See VARACTOR.

capacitance divider An alternating-current voltage divider that uses capacitors, rather than resistors. It is used in certain oscillators, such as the Colpitts type.

capacitance filter A filter consisting of only a high-capacitance capacitor. Because the capacitor cannot discharge instantaneously, it tends to maintain its voltage and smooth out the ripples in the voltage applied to it.

capacitance-inductance bridge A combination ac bridge that can be used for either capacitance or inductance measurement. Both capacitance and inductance can be measured in terms of a standard capacitance; however, some of these bridges use standard inductors in the inductance-measuring mode.

capacitance meter A direct-reading meter for measuring capacitance. In most available types, a stable ac voltage is applied to the meter circuit, to which an unknown capacitor is connected in series; meter deflection is roughly proportional to the reactance of the capacitor. Also called MICROFARAD METER.

capacitance ratio In a variable capacitor, the ratio of maximum to minimum capacitance.

capacitance relay A relay circuit that operates from a small change in its own capacitance. It consists of an RF oscillator whose tank capacitance is very low. When a finger is brought near the circuit's short pickup antenna, the attendant increase in capacitance detunes the oscillator, activating the relay. Also called PROXIMITY RELAY and PROXIMITY SWITCH.

capacitance-resistance bridge A combination ac bridge that can be used for either capacitance or resistance measurement. The unknown resistance is measured against a standard resistor; the unknown capacitance against a standard capacitor.

capacitance sensor See CAPACITANCE TRANSDUCER.

capacitive amplifier See DIELECTRIC AMPLIFIER.

capacitive attenuator An ac attenuator whose elements are capacitors in any desired combination of fixed and/or variable units. The desired attenuation is afforded by the capacitance ratio.

capacitive coupling A means of coupling between circuits that uses a series capacitor for direct-current blocking. The signal passes through the capacitor, but the blocking effect allows different bias voltages to be applied to the two stages.

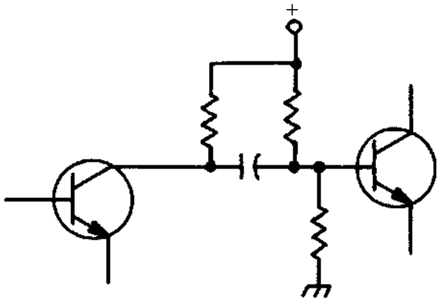

capacitive coupling

capacitive diaphragm A metal plate deliberately placed in a waveguide to introduce capacitive reactance and, thereby, cancel an inductive reactance.

capacitive-discharge ignition An electronic ignition system for automotive engines. Provides nearly constant high voltage, regardless of engine speed. A dc-to-dc step-up converter charges a large capacitor (typically to 300 volts) when the distributor breaker points are closed; when they are open, the capacitor discharges through the ignition coil, thereby generating an ignition pulse of several thousand volts.

capacitive division Reduction of an ac voltage by a capacitive voltage divider.

capacitive feedback Feeding energy back from the output to the input of an amplifier or oscillator through a capacitor.

capacitive-input filter A smoothing filter for ac power supplies, in which the element closest to the rectifier is a capacitor, regardless of the components or circuits placed subsequently.

capacitive load A load consisting of a capacitor or a predominantly capacitive circuit.

capacitive loading In an antenna, the addition of capacitance in series with the element(s). This raises the resonant frequency for a radiator having a given physical length. It can also serve to increase the physical length required for a radiator having a specified resonant frequency. Compare INDUCTIVE LOADING.

capacitive post A protrusion inside a waveguide for the purpose of introducing capacitive reactance to cancel an inductive reactance.

capacitive potentiometer See CAPACITIVE VOLTAGE DIVIDER.

capacitive pressure sensor A pressure sensor that uses a radio-frequency oscillator and a pair of metal plates separated by dielectric foam. The circuit is designed so a change in the capacitance between the plates causes the oscillator frequency to change. This change is sensed. A signal is sent to an analog-to-digital converter (ADC) and then to a microcomputer that calculates the extent of the pressure.

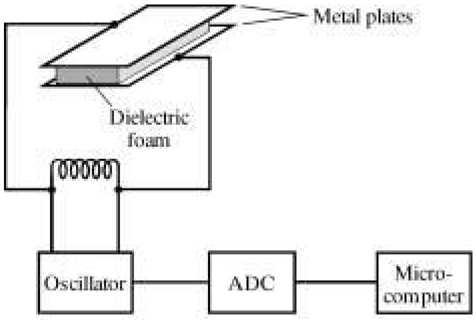

capacitive pressure sensor

capacitive proximity sensor A transducer used in mobile robots that detects the presence of certain kinds of objects. It consists of an oscillator whose frequency is determined by an inductance-capacitance (LC) circuit to which a metal plate is connected. When a conducting or partially conducting object comes near the plate, the mutual capacitance changes the oscillator frequency. This change is detected and sent to the robot controller.

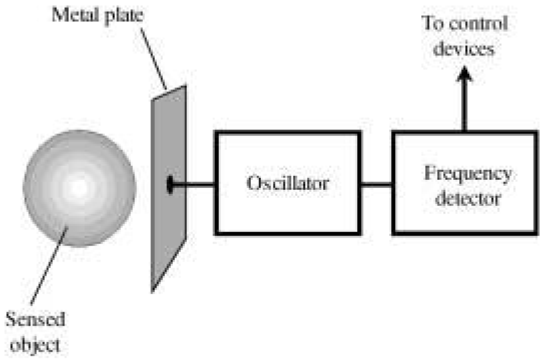

capacitive proximity sensor

capacitive reactance Symbol, X_C. Unit, ohm. The reactance exhibited by an ideal capacitor, considered as a negative pure-imaginary quantity; $X_C = -j/(6.28fC)$, where f is the frequency in hertz, C is the capacitance in farads, and j is the unit imaginary number (the square root of –1). Alternatively, f can be specified in megahertz and C in microfarads. In a pure capacitive reactance, current leads voltage by 90 degrees. Also see CAPACITANCE, CAPACITOR, and REACTANCE.

capacitive speaker See ELECTROSTATIC SPEAKER.

capacitive transducer A transducer consisting essentially of a refined variable capacitor whose value is varied by a quantity under test, such as pressure, temperature, liquid level, etc.

capacitive tuning Variable-capacitor tuning of a circuit.

capacitive voltage divider A capacitive attenuator usually consisting of two series-connected capacitors whose values are such that an applied ac voltage is divided across them in the desired ratio.

capacitive welding An electronic welding system in which energy stored in a capacitor is discharged through the joint to be welded. This develops the heat necessary for the operation.

capacitive window A pair of capacitive diaphragms used in a waveguide to introduce capacitive reactance.

capacitor A passive electronic-circuit component consisting of, in basic form, two metal electrodes or plates separated by a dielectric (insulator).

capacitor amplifier See DIELECTRIC AMPLIFIER.

capacitor antenna See CONDENSER ANTENNA.

capacitor bank A network of capacitors connected in combination, yielding a desired characteristic.

capacitor braking The connection of a capacitor to the winding of a motor after the removal of power, to speed up the process of braking.

capacitor color code See COLOR CODE.

capacitor decade See DECADE CAPACITOR.

capacitor-discharge ignition CAPACITIVE-DISCHARGE IGNITION.

capacitor filter In a direct-current power supply, a filter consisting simply of a capacitor connected in parallel with the rectifier output.

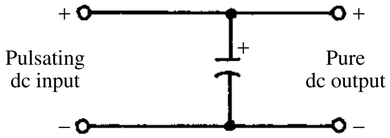

capacitor filter

capacitor-input filter A filter whose input component is a capacitor. The capacitor-input power-supply filter is distinguished by its relatively high dc output voltage, but somewhat poorer voltage regulation, compared with the CHOKE-INPUT FILTER.

capacitor leakage Direct current flowing through the dielectric of a capacitor. In a good nonelectrolytic capacitor, this current is normally less than 1 microampere. In an electrolytic capacitor, it can be up to several milliamperes, depending on the capacitance and the applied voltage.

capacitor loudspeaker See ELECTROSTATIC SPEAKER.

capacitor microphone See CONDENSER MICROPHONE.

capacitor motor An ac motor that uses a capacitor in series with an auxiliary field winding for starting purposes. Initially out-of-phase current in the auxiliary field (starting winding) causes a rotating field that turns the rotor. When the rotor reaches a safe speed, a centrifugal switch disconnects the capacitor and auxiliary field, and the motor continues running as an induction motor.

capacitor series resistance The ohmic loss in a capacitor. It results partly from conductor losses, and partly from losses in the dielectric material.

capacitor substitution box An enclosed assortment of selected-value capacitors arranged to be switched one at a time to a pair of terminals. In troubleshooting and circuit development, any of several useful fixed capacitance values can be thus obtained.

capacitor voltage **1.** The voltage at the terminals of a capacitor. **2.** The maximum voltage rating of a capacitor.

capacitor voltmeter See ELECTROSTATIC VOLTMETER.

capacity **1.** A measure of a cell's or battery's ability to supply current during a given period. **2.** CAPACITANCE. **3.** The number of bits or bytes a computer storage device can hold. **4.** The limits of numbers that a register can process.

capacity lag In an automatic control system, a delay caused by the storing of energy by the components. For example, in a heating system, capacity lag results from the time taken to heat the air or fluid after the thermostat turns on the heat.

capillary electrometer A sensitive voltage indicator, consisting of a column of mercury in a transparent capillary tube, in which is suspended a small drop of acid. When a voltage is applied to both ends of the mercury column, the acid drop moves toward the low-potential end of the column over a distance proportional to the voltage.

capstan The driven spindle or shaft of a magnetic tape recorder or transport.

capture area The effective ability of a radio antenna to pick up electromagnetic signals. The larger the capture area, the greater the antenna gain.

capture effect **1.** In frequency-modulation (FM) radio receivers, the effect of domination by the stronger of two signals, or by the strongest of several signals, on the same frequency. **2.** In an automatic-frequency-control system, the tendency of the receiver to move toward the strongest of several signals near a given frequency. **3.** In general, the tendency of one effect to totally predominate over other effects of lesser amplitude.

capture ratio A measure of frequency-modulation (FM) tuner selectivity: The amplitude difference, in decibels, between unwanted signals and the one being tuned in.

carbon Symbol, C. A nonmetallic element. Atomic number, 6. Atomic weight, 12.011. Carbon, besides being an invaluable material in electronics, is an important constituent of organic compounds.

carbon arc The arc between two electrified pencils of carbon or, as in an arc converter, between a carbon pencil and a metal electrode.

carbon brush A contact made of carbon or some mixture of carbon and another material, used in motors, generators, variable auto-transformers, rheostats, and potentiometers.

carbon-button amplifier An audio-frequency amplifier having as the active component an earphone whose diaphragm is attached to a carbon microphone button. The input signal applied to the earphone makes its diaphragm vibrate. The vibrating button modulates a local direct current. Amplification results from the large ratio of modulated local current to input-signal current.

carbon-composition resistor A non-inductive resistor made from a mixture of finely powdered carbon with a non-conductive substance, usually phenolic. The resulting clay-like material is pressed into a cylindrical shape, and wire leads are inserted in the ends. The resistance depends on the ratio of carbon to the non-conducting material, and on the physical distance between the wire leads. This type of resistor is useful from direct current to ultra-high radio frequencies. Compare FILM RESISTOR, WIREWOUND RESISTOR.

carbon/disk rheostat A rheostat consisting of a stack of carbon disks or washers, arranged so that a controllable pressure can be exerted on the stack. As a knob is turned, a screw increases or decreases the pressure, varying the total resistance of the stack.

carbon-film resistor A stable resistor whose resistance element is a film of carbon, vacuum-deposited on a substrate, such as a ceramic.

carbonization The application of a coat of carbon onto an electrode, either by electroplating or by any other means.

carbon microphone A microphone that includes one or two carbon buttons. See BUTTON MICROPHONE.

carbon-paper recorder A recorder in which a signal-actuated stylus writes, by impression only, through a sheet of carbon paper onto a plain sheet underneath. This eliminates the need for an ink-carrying stylus.

carbon-pile regulator A voltage regulator in which a stack of carbon disks or washers is in series with the shunt field. The pile resistance and field current depend on pressure applied to the pile by a wafer spring acting through a movable iron armature. Voltage drops increase the pressure and voltage rises decrease the pressure, thus regulating the generator with which it is associated.

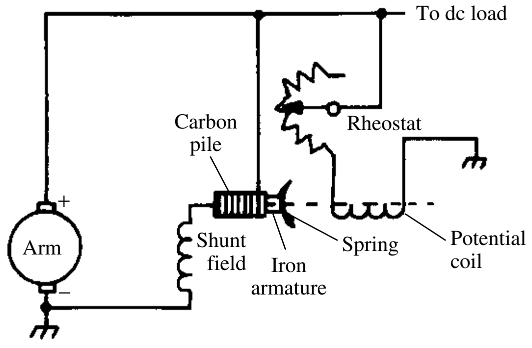

carbon-pile regulator

carbon-pile rheostat See CARBON-DISK RHEOSTAT.

carbon recording **1.** A record made with a CARBON-PAPER RECORDER. **2.** The use of a carbon-paper recorder in data acquisition, facsimile, communications, and similar applications.

carbon resistor A resistor made from carbon, graphite, or some composition that contains carbon.

carbon/silicon-carbide thermocouple A thermocouple that is a junction between carbon and silicon carbide.

carbon transfer recording A method of facsimile reception in which the image is reproduced by carbon particles sprayed on the paper, a process controlled by the received signal.

carbon-zinc cell See ZINC-CARBON CELL.

Carborundum Formula, SiC. Trade name for a synthetic silicon carbide used as a semiconductor, refractory, or abrasive. Also see SILICON CARBIDE.

Carborundum crystal Trade name for a characteristically superhard crystal of silicon carbide.

Carborundum varistor A voltage-dependent resistor made from Carborundum.

carcinotron A special kind of oscillator tube used at ultra-high and microwave frequencies.

card **1.** A usually thin, rectangular board containing a PRINTED CIRCUIT, often equipped with an edge connector that makes it easy to install, remove, or replace. Common in electronic and computer equipment having modular construction. **2.** The usually flat, thin insulating strip on which a resistor element is wound.

cardiac monitor An electronic device that displays or records electrical impulses from the heart for medical observation or diagnosis.

cardiac pacemaker An electrical cardiac stimulator that causes the heart to beat at certain intervals. Used when the patient has heart disease that prevents the heart from regulating itself.

cardiac stimulator An electronic device (sometimes implanted in the subject) that supplies electric pulses to stimulate heart action. Also called DEFIBRILLATOR and PACEMAKER.

card image In memory storage, the data contained on a single card.

cardiogram ELECTROCARDIOGRAM.

cardiograph ELECTROCARDIOGRAPH.

cardioid diagram A polar response curve in the shape of a cardioid pattern.

cardioid microphone A microphone with a (roughly) heart-shaped sound-field pickup pattern.

cardioid pattern A radiation/response pattern with one sharp null in the direction opposite the single main lobe. The lobe is extremely broad. In two dimensions, the curve is shaped somewhat like a “Valentine” heart.

cardiotachometer A device that indicates the pulse rate.

cardistimulator See CARDIAC STIMULATOR.

Carey-Foster bridge A special version of the slide-wire bridge that is useful for measuring an unknown resistance, whose value is close to that of a standard resistance.

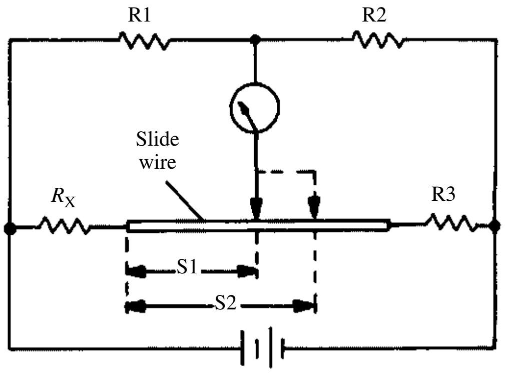

Carey-Foster bridge

Carey-Foster mutual inductance bridge An ac bridge that permits the measurement of mutual inductance in terms of a standard capacitor.

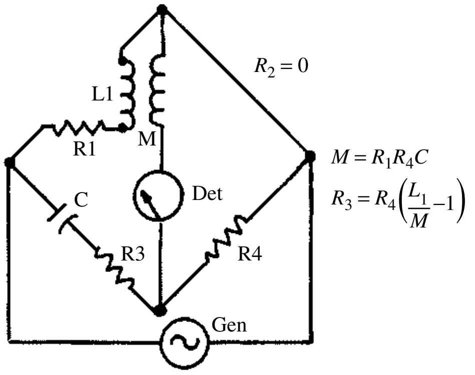

Carey-Foster mutual inductance bridge

carnauba wax A wax obtained from the Brazilian wax palm. Used as an electrical insulator, and as the dielectric in some electrets.

Carnot theorem In thermodynamics, the proposition that in a reversible cycle, all available energy is converted into mechanical work. Also called *Carnot's principle.*

carrier 1. See CARRIER WAVE. **2.** See CHARGE CARRIER.

carrier amplifier See DIELECTRIC AMPLIFIER.

carrier beating 1. The mixing of two radio-frequency carriers that are separated by a small amount of frequency, resulting in an audible tone in a receiver. **2.** A heterodyne in a facsimile or television signal, resulting in a pattern of cross hatches in the received image.

carrier choke A radio-frequency (RF) choke, inserted in a line to block a carrier component.

carrier chrominance signal For conveying color television information, sidebands of a modulated chrominance subcarrier.

carrier color signal For conveying color information in color television transmission, the sidebands of a modulated chrominance subcarrier (plus the unsuppressed chrominance subcarrier) added to the monochrome signal.

carrier concentration In a semiconductor material, the number of charge carriers per unit volume.

carrier control 1. The modification, adjustment, or switching of a carrier wave. **2.** Adjustment of a circuit or device by means of a carrier wave.

carrier current The current component of a carrier wave, or the amplitude of that current. Compare CARRIER POWER and CARRIER VOLTAGE.

carrier-current communication See WIRED WIRELESS.

carrier-current control 1. Control of the current component in a carrier wave. **2.** Remote control by means of wired wireless.

carrier-current receiver See WIRED-RADIO RECEIVER.

carrier-current relay A radio-frequency (RF) relay circuit, operated over a wire line by means of a transmitter.

carrier-current transmitter See WIRED-RADIO TRANSMITTER.

carrier deviation See CARRIER SWING.

carrier dispersion In a semiconductor, the spreading out of electrons and holes that leave the emitter simultaneously, but arrive at the collector at various times after following different paths.

carrier frequency The center frequency of a CARRIER WAVE.

carrier-frequency pulse A pulse that contains radio-frequency oscillation.

carrier-frequency range The band of carrier frequencies over which a transmitter or signal generator can operate.

carrier injection The apparent emission (injection) of electrons or holes into a semiconductor when a voltage is applied to the junction.

carrier leak 1. A point at which carrier-wave energy escapes a circuit or enclosure. **2.** The residual carrier voltage present in the output of a carrier-suppressing circuit.

carrier level The amplitude of an unmodulated carrier wave.

carrier lifetime In a semiconductor, the interval before an injected current carrier (see CARRIER INJECTION) recombines with an opposite carrier and ceases to be mobile.

carrier line In carrier-current systems (see WIRED WIRELESS), the line or cable conducting the carrier-wave energy.

carrier mobility Symbol, μ. In a semiconductor material, the average drift velocity of electrons and holes per unit electrostatic field.

carrier noise Modulation of a carrier when there is no input from the modulator itself; unwanted modulation.

carrier noise level The noise signal amplitude that results from unintentional fluctuations of an unmodulated carrier.

carrier-on-light transmission A form of transmission in which many different signals are sent simultaneously by modulating a beam of light at multiple frequencies.

carrier-on-microwave transmission A form of transmission in which many different signals are sent simultaneously by modulating a microwave signal at multiple lower frequencies.

carrier-on-wire transmission A form of transmission in which many different signals are sent at the same time over a wire, by using radio-frequency carriers. Also called CARRIER-CURRENT COMMUNICATIONS or WIRED RADIO.

carrier oscillator In a single-sideband receiver, the radio-frequency (RF) oscillator that supplies the missing CARRIER WAVE.

carrier power The actual power represented by a radio-frequency (RF) carrier applied to an antenna, measured by either the direct or indirect method. The direct method involves determination of power according to the formula $P = I^2R$, where I is antenna current and R is antenna resistance at the point of current measurement. The indirect method involves determination of power according to the formula $P = EIF$, where E and I are antenna voltage and current, and F is a factor less than 1.0, whose value depends on the type of modulation used.

carrier power-output rating The power delivered by an unmodulated transmitter or generator to the normal load or its equivalent.

carrier shift In an amplitude-modulated transmitter or generator, the undesired change of average carrier voltage during modulation.

carrier-shift indicator An instrument for detecting carrier shift. It usually contains only a pickup coil, semiconductor diode, and dc milliammeter in series. Meter deflection is steady until carrier shift is detected; then, the needle fluctuates.

carrier signaling In wire telephony, the use of carrier-wave signals to operate such functions as dialing, ringing, busy signal, etc.

carrier storage In a semiconductor device, the tendency of mobile carriers to stay near a junction for a short time after the junction voltage has been removed or reversed in polarity.

carrier suppression The elimination of the carrier in an amplitude-modulated signal so that only the sideband energy remains.

carrier swing In frequency-modulated or phase-modulated transmission, the total deviation (lowest to highest instantaneous frequency) of the carrier wave.

carrier system The transmission of many signals over one circuit, accomplished by modulating various different carriers at different frequencies. Different signals can use different modulation methods.

carrier telegraphy **1.** Continuous-wave telegraphy by WIRED WIRELESS. **2.** Wired-wireless telegraphy in which a radio-frequency carrier is modulated by an audio-frequency keying wave.

carrier telephony Telephone communication by WIRED WIRELESS.

carrier terminal **1.** At each end of a carrier-current line or cable, the equipment for generating, modifying, or utilizing the carrier energy. **2.** In a balanced modulator, the point of carrier insertion.

carrier-to-noise ratio The ratio of carrier amplitude to noise-voltage amplitude.

carrier transmission Transport of information by a carrier, as by an amplitude-modulated radio wave that carries the low-frequency information as the AF modulation envelope and delivers it to the demodulator at the receiving station.

carrier-type dc amplifier A high-frequency ac amplifier, ahead of which is operated a generator and transducer. A dc voltage applied to the transducer modulates the carrier supplied by the generator; the amplifier boosts the modulated wave, and the resultant output is rectified at a level higher than that of the dc input signal.

carrier voltage The voltage component of a carrier wave; also, the amplitude of this component. Compare CARRIER CURRENT and CARRIER POWER.

carrier wave A sine wave that is modulated to convey information in wireless and cable communications systems. The lowest frequency normally used for wireless signal transmission is 9 kHz, corresponding to a wavelength of approximately 33 km. The highest frequency is less well defined; some systems make use of visible light waves, whose wavelengths are as short as approximately 4×10^{-7} m. For modulation to work effectively, the carrier must have a frequency at least 10 times the highest frequency of the modulating signal.

carry **1.** In adding a column of figures, the digit added to the column at the left when the sum exceeds one less than the radix value. **2.** In digital computers and counters, a pulse that corresponds to the arithmetic operation in which a figure is carried to the next column in addition.

carrying capacity The ability of a conductor, such as copper wire, to carry current safely (expressed in maximum amperes).

carry-complete signal In an arithmetic computation by a computer, an adder-produced signal indicating that the pertinent carries have been generated.

carry system A communications system in which several carries occupy one circuit.

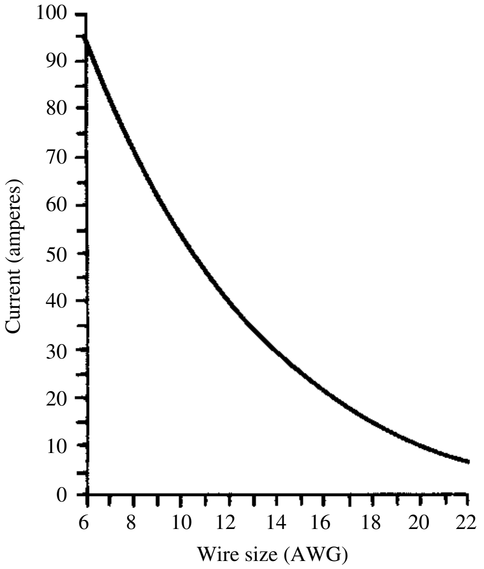

carrying capacity
(of some AWG sizes of copper wire)

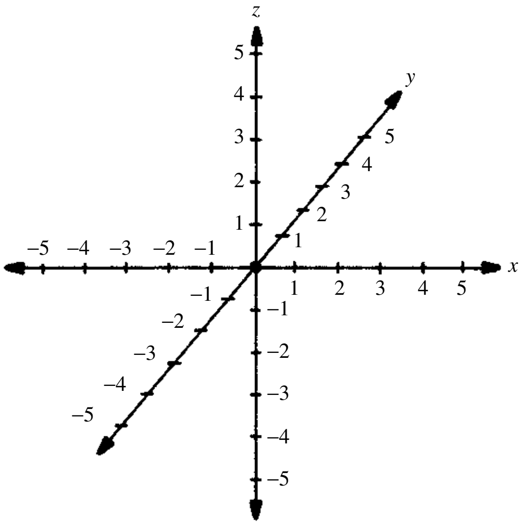

Cartesian three-space

carry time The time taken for a digital computer or counter to perform a carry operation (See CARRY, **2**).

Cartesian coordinate geometry Also called *rectangular coordinate geometry.* In robotic systems, a movement scheme in two or three dimensions. The position of the robot arm is determined by linear coordinates, relative to an origin point. These coordinates are specified along linear axes—each of which is perpendicular to the others at the origin. See CARTESIAN COORDINATES, CARTESIAN PLANE, and CARTESIAN THREE-SPACE.

Cartesian coordinates Also called *rectangular coordinates.* A mathematical system that uniquely defines the position of a point on a plane, in space, or in general, in an *n*-dimensional hyperspace when *n* is a whole number greater than 3. There are *n* axes for *n* dimensions, each axis intersects all the others at a single point, called the origin. The axes are mutually perpendicular at this origin. The axes are scaled in units with the origin having coordinate values that are all equal to zero (usually). Positive values go along the axes in one direction; negative numbers go in the opposite direction for each axis. Usually, the axes are graduated in equal-sized units. The system gets its name from the mathematician Rene Descartes.

Cartesian plane A linear, two-dimensional coordinate plane commonly used for graphing equations in one variable.

Cartesian three-space A linear, three-dimensional graph-coordinate system used for rendering equations in one or two variables.

Cartesian three-space graph A three-dimensional graph that shows an equation in one or two variables. Three-space graphs are often displayed more clearly by means of computer graphics, in which the entire display can be rotated to show the characteristics of the surface resulting from a given equation or function.

Cartesian n-space The coordinate space defined by a Cartesian system of *n* coordinates, where *n* is a whole number of 2 or greater.

cartridge **1.** The replaceable transducer assembly of a microphone. **2.** A magnetic-tape magazine. Also see TAPE CARTRIDGE. **3.** A removable computer mass-storage medium, containing a tape, magnetic diskette, or optical diskette. **4.** An insulating tube housing a fuse, semiconductor component, resistor, capacitor, or other part.

cartridge fuse A fuse consisting of a fusible wire enclosed in a cartridge, having a ferrule at each end for plug-in connection.

cascadable Capable of, or designed for, being connected in cascade with other similar or identical components.

cascade **1.** Components or stages connected and operated in sequence, as in a three-stage amplifier. The components or stages are often but not necessarily identical. **2.** To form a cascade.

cascade control **1.** In an automatic control system, a controller whose setting is varied by the output of another controller. **2.** An automatic control system in which the control units are connected in stages, so that one unit must operate before the next one can function.

cascaded amplifier A multistage amplifier in which the stages are forward-coupled in succession.

cascaded carry In digital computer practice, a system of performing the carry operation (see CARRY) in which the $n + 1$ place receives a carry pulse only when the nth place has received carry information to generate the pulse.

cascade thermoelectric device A thermoelectric component or circuit that consists of several cascaded sensors (see CASCADE, **1**).

cascade voltage doubler A voltage-doubler circuit (see VOLTAGE DOUBLER) consisting of two diode-capacitor combinations in cascade. Unlike the conventional voltage-doubler circuit with two capacitors in the output, the cascade voltage doubler has one in the input and one in the output.

cascode A high-gain, low-noise, high-input-impedance amplifier circuit, consisting of a grounded-emitter or grounded-source input stage coupled directly to a grounded-base or grounded-gate output stage.

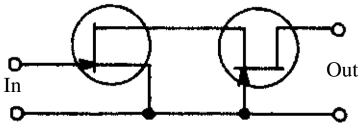

cascode (field-effect transistor arrangement)

case temperature The temperature at a designated point on the outside surface of a component's case or housing.

Cassegrain antenna A dish antenna that uses CASSEGRAIN FEED.

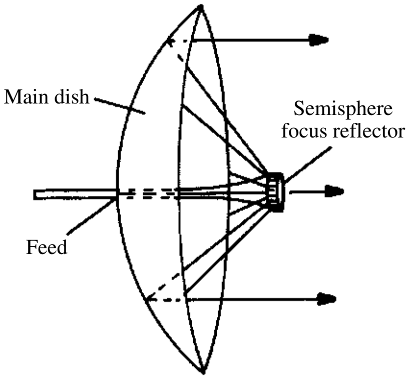

Cassegrain antenna

Cassegrain feed A dish-antenna feed system in which the feed point is located at the center of the dish itself. For transmission, the radio-frequency energy emerges from a waveguide and is directed to a small convex reflector at the focal point of the dish. The small reflector directs the signal back to the dish, spreading the energy out to cover the entire surface of the dish. The dish reflects the energy again and collimates it in the desired direction of propagation. For reception, the process is reversed; the dish focuses the energy on the small reflector, which propagates it back to the feed point.

cassette **1.** A holder (magazine) of reels of magnetic tape that is itself a mechanical subassembly, which can be easily inserted into and removed from a tape deck. **2.** A lightweight holder of photographic film or X-ray plates (before, during, and after exposure).

castor oil A viscous insulating oil extracted from castor beans. Highly refined castor oil is used as an impregnant in some oil-filled capacitors. Dielectric constant, 4.3 to 4.7. Dielectric strength, 380 V/mil.

CAT Abbreviation of COMPUTERIZED AXIAL TOMOGRAPHY.

catalysis The process whereby an agent, called a catalyst, enhances a chemical reaction without entering into the reaction. Catalysts are used in electronics, for example, to promote the setting of resins in potting and encapsulating operations.

catalytic agent A substance that accomplishes catalysis.

cataphoresis As caused by the influence of an electrostatic field, the migration toward the cathode of particles suspended in a liquid.

catastrophic failure **1.** Sudden, unexpected failure of a component or circuit. **2.** Failure that can result in the breakdown of an entire system. Also called *catastrophic breakdown.*

catcher In a Klystron, the second reentrant cavity. (See KLYSTRON.)

catcher diode A diode that is connected to regulate the voltage at the output of a power supply. The cathode is connected to a source of reference voltage. If the anode, connected to the source to be regulated, becomes more positive than the cathode, the diode conducts and prevents the regulated voltage from rising more than 0.3 volt above the reference voltage (for germanium diodes) or 0.6 volt above the reference voltage (for silicon diodes).

catcher grids In a Klystron, the grids through which the bunched electrons pass on their way from the buncher to the collector. Catcher grids absorb energy from the bunched electrons and present it to the collector circuit.

category In a computer system, a group of magnetic disk volumes containing information related by a common application.

category storage A computer-file storage section that contains a number of categories and used by an operating system.

catenation See CONCATENATION.

cathode **1.** The negative electrode of a device (i.e., the electrode from which electrons move when a current passes through the device). **2.** In an electrochemical cell, the electrode that gains electrons. This is generally the positive electrode. **3.** In a vacuum tube, the electron-emitting electrode (filament or indirectly heated cathode sleeve).

cathode current Symbol I_K. The current flowing in the cathode circuit of a tube. Cathode current is the total of grid, plate, screen, and suppressor currents, and can have an ac and a dc component.

cathode dark current The electron emission from the photocathode of a camera tube when there is no illumination.

cathode element In a vacuum tube, an indirectly heated emitter of electrons. Also see CATHODE, **2.**

cathode emission **1.** The giving up of electrons by the cathode element of a device, such as a vacuum tube. Electrons can be emitted by either hot or cold cathodes, depending on the tube. **2.** Collectively, electrons released by a cathode.

cathode heating time The time required for the temperature of a tube cathode to increase from cold to its maximum specified operating temperature after the cathode current has been initiated. Also called cathode warmup time.

cathode luminous sensitivity For a photomultiplier tube, the cathode's sensitivity to light. This sensitivity figure is the ratio of photocathode current to incident light flux.

cathode-ray oscillograph An instrument that provides a permanent record, by photographic or other means, of the image on the screen of a cathode-ray tube.

cathode-ray oscilloscope See OSCILLOSCOPE.

cathode rays Invisible rays emanating from the cathode element of an evacuated tube operated with a high voltage between the anode and cathode. Cathode rays (electrons) cause certain substances, PHOSPHORS, to glow upon striking them.

cathode-ray scanning tube Any tube in which an electron beam is deflected horizontally and vertically to scan an area. These include oscilloscope tubes, some computer monitors, radar displays, and television camera tubes.

cathode-ray tube **1.** An evacuated tube containing an anode and cathode that generates cathode rays when operated at high voltage. **2.** An oscilloscope tube. **3.** A picture tube.

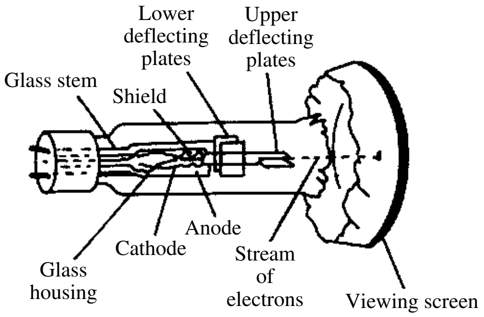

cathode-ray tube

cathode terminal **1.** In a diode (semiconductor or tube), the terminal to which a negative dc voltage must be applied for forward-biasing the diode. Compare ANODE TERMINAL. **2.** In a diode, the terminal at which a positive dc voltage appears when the diode acts as an ac rectifier. Compare ANODE TERMINAL. **3.** The terminal connected internally to the cathode element of device. **4.** In a vacuum tube, an indirectly heated electron emitter.

cathode voltage Symbol, E_K. The voltage between ground (or B-minus) and the cathode of a tube; it can have both ac and dc components.

cathodic protection A method of preventing corrosive galvanic action in underground metal pipes or the submerged hulls of ships. The part to be protected is used as the cathode of a circuit through which a direct current is passed in the direction opposite to that which caused the corrosion, thus counteracting it.

cathodofluorescence Fluorescence resulting from a material's exposure to cathode rays.

cathodoluminescence In a vacuum chamber in which a metal target is bombarded with high-velocity electrons (cathode rays), the emission of radiation of a wavelength characteristic of the metal.

cation A positive ion. Also see ION.

CAT scanner The X-ray apparatus for COMPUTERIZED AXIAL TOMOGRAPHY.

CATV Abbreviation of COMMUNITY-ANTENNA TELEVISION (usually cable television).

caustic soda electrolyte Symbol, $NaOH$. Sodium hydroxide solution, as used in some secondary cells and experimental devices.

cavitation The local formation of cavities in a fluid used in ultrasonic cleaning because of the reduction in pressure at those points.

cavitation noise In an ultrasonic cleaner, the noise resulting from the collapse of bubbles produced by cavitation.

cavity A metallic chamber (can) in which energy is allowed to reflect, sometimes resulting in resonance.

cavity filter A microwave (usually band rejection) filter consisting of a resonant cavity and associated coupling devices.

cavity frequency meter See CAVITY WAVEMETER.

cavity impedance The impedance across a cavity at a particular frequency. At resonance, the cavity impedance is purely resistive.

cavity laser A laser that employs a resonant cavity filled with gas, such as helium/neon or argon, and a pair of reflectors. Resonance occurs between the reflectors, one of which is totally reflective and the other of which is approximately 95 percent reflective. Output is from the partially reflective end of the device.

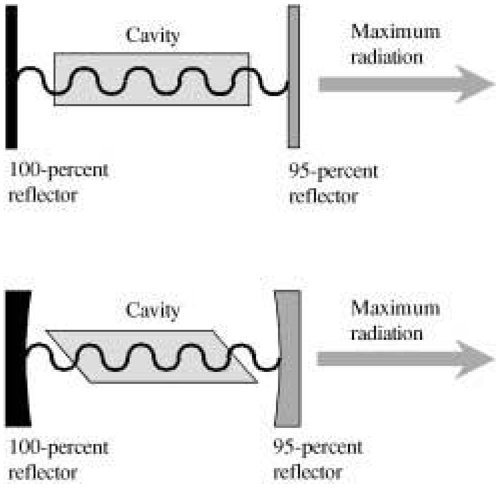

cavity laser

cavity magnetron A magnetron whose anode is a series of resonant cavities.

cavity oscillator An oscillator with a cavity-tuned circuit.

cavity radiation Energy radiated from a tiny hole in an otherwise sealed chamber. The radiation occurs at all electromagnetic wavelengths; the greater the temperature within the chamber, the greater the frequency at which the radiation has its maximum amplitude.

cavity resonance The phenomenon whereby a hollow cavity resonates; specifically, resonance in small metal cavities at microwave frequencies.

cavity resonator See RESONANT CAVITY.

cavity wavemeter An absorption wavemeter whose adjustable element is a tunable resonant cavity into which radio-frequency (RF) energy is injected through a waveguide or coaxial cable. Such an instrument is useful at microwave frequencies.

CB Abbreviation of CITIZENS BAND.

Cb Symbol for COLUMBIUM.

C_B Symbol for BASE CAPACITANCE of a transistor.

C band The band of radio frequencies between 3.9 and 6.2 GHz.

C_C Symbol for collector capacitance of a transistor.

cc **1.** Alternative abbreviation of *cubic centimeter.* The International Organization for Standardization recommends cm^3. **2.** Abbreviation of COTTON-COVERED.

CCA Abbreviation of CURRENT-CONTROLLED AMPLIFIER.

CCD Abbreviation of CHARGE-COUPLED DEVICE.

CCIS Abbreviation of COMMON-CHANNEL INTERFACE SIGNALING.

CCIR Abbreviation of *Comite Consultatif International des Radiocommunications (International Radio Consultative Committee).*

CCIT Abbreviation of *Comite Consultatif International Telegrafique (International Telegraph Consultative Committee).*

CCITT Abbreviation of *Comite Consultatif International Telegrafique et Telephonique (International Telegraph and Telephone Consultative Committee).*

CCS **1.** Abbreviation of CONTINUOUS COMMERCIAL SERVICE. **2.** Abbreviation of *common-channel signaling.*

CCTV Abbreviation of CLOSED-CIRCUIT TELEVISION.

CCTV monitor A video monitor that receives a signal from a CCTV transmitter.

CCTV signal The picture signal in a CCTV system. It can be either a modulated radio-frequency signal or a composite video signal.

ccw Abbreviation of COUNTERCLOCKWISE.

CD Abbreviation of COMPACT DISK.

Cd Symbol for CADMIUM.

cd Abbreviation of CANDELA.

CD-4 A method of obtaining quadraphonic reproduction on a phonograph disk using modulated carriers with frequencies above the human hearing range.

CDI Abbreviation of CAPACITOR-DISCHARGE IGNITION.

C display A radar display showing the target as a dot whose coordinates represent the bearing (horizontal) and angle of elevation (vertical). Compare A DISPLAY, J DISPLAY, and K DISPLAY.

cd/m^2 Candelas per square meter, the SI unit of luminance.

CD-ROM Abbreviation of COMPACT-DISK READ-ONLY MEMORY.

Ce Symbol for CERIUM.

C_e Symbol for EMITTER CAPACITANCE of a transistor.

ceiling **1.** The maximum possible power output from a transmitter. **2.** The maximum possible current or voltage that a circuit can deliver. **3.** In aviation, the level of the cloud base.

ceilometer An instrument for measuring ceiling (cloud height).

cel In animated graphics, an individual image or frame.

cell **1.** A single (basic) unit for producing dc electricity by electrochemical or photovoltaic action, as in a battery or a solar panel. Also see PRIMARY CELL, STANDARD CELL, STORAGE CELL. **2.** An addressable, one-word-capacity storage element in a computer memory. **3.** The geographic region covered by a specified repeater in a cellular communications network. See CELLULAR COMMUNICATIONS. **4.** An electrostatic charge dipole in the atmosphere, usually occurring in or near thunderstorms. **5.** A thunderstorm.

cell constant The surface area of the electrodes in a cell divided by the distance between them. The basic linear units must be the same: for example, square centimeters for surface area and centimeters for distance.

cell counter A bioelectronic instrument used to count blood cells and other minute particles.

cell reversal A condition that can occur in some rechargeable electrochemical cells and batteries, such as *nickel-cadmium batteries.* It most often results from neglecting to recharge the cell or battery when it has become fully discharged.

cell-type enclosure A room designed to prevent the entrance or escape of radio-frequency (RF) electromagnetic fields, characterized by double-walled copper-mesh shielding.

cellular coil A coil having a crisscross (usually multilayer) winding. Examples: lattice-wound coil, honeycomb coil, basket-weave coil.

cellular communications A radio, telephone, or television communications network that makes use of numerous fixed repeaters. Subscribers use mobile or portable transceivers that are always within range of at least one repeater. The most common form is known as *cellular telephone* or *cellular mobile radio telephone.*

celluloid A thermoplastic dielectric material that is a blend of cellulose nitrate and camphor. Dielectric constant, 4 to 7. Dielectric strength, 250 to 780 V/mil.

cellulose acetate A plastic dielectric material used as a substrate for magnetic tapes, photographic film, and similar applications. Dielectric constant, 6 to 8. Dielectric strength, 300 V to 1 kV/mil. Also see ACETATE.

cellulose acetate base See ACETATE BASE.

cellulose acetate butyrate A thermoplastic dielectric material that is an acetic and butyric acid ester of cellulose.

cellulose acetate tape See ACETATE TAPE.

cellulose nitrate The nitric acid ester of cellulose, a plastic insulating material.

cellulose propionate A thermoplastic molding material that is a propionic acid ester of cellulose.

Celsius scale A temperature scale in which 0 degrees is the freezing point of water, and 100 degrees the boiling point of water. Also called CENTIGRADE SCALE. Compare ABSOLUTE SCALE, and FAHRENHEIT SCALE.

cent An audio-frequency interval of 0.01 (1⁄100) of a half step. A half step is the frequency difference between two immediately adjacent keys on a piano.

center channel In high-fidelity stereo, a phantom sound source that appears to exist midway between the left and right speakers or earpieces. The effect is caused by identical, or nearly identical, signals in the left and right channels.

center-fed antenna An antenna in which the feeders are connected to the center of the radiator.

center feed **1.** Attaching a feeder or transmission line to the center of the radiator of an antenna. **2.** Connection of signal-input terminals to the center of a coil. **3.** Descriptive of paper tape whose feed holes are aligned with character hole centers. Compare ADVANCE FEED TAPE.

center frequency **1.** The frequency, in a communications receiver, that is midway between the lower and upper 3-dB-down amplitude points. **2.** The average frequency of a modulated carrier. **3.** The carrier frequency of a modulated signal, whether or not the carrier is suppressed.

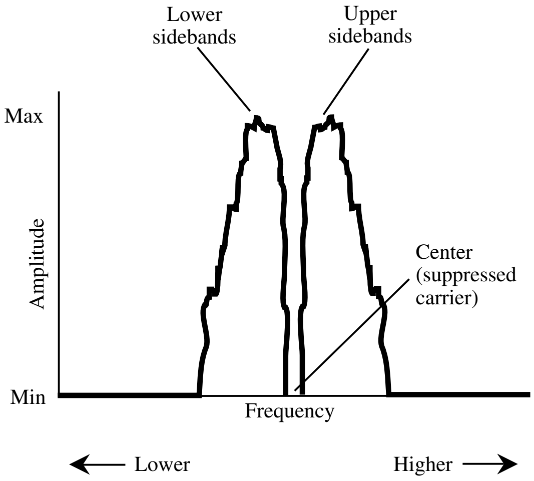

center frequency, 3.

centering control In an oscilloscope circuit, a potentiometer used to position the image on the screen (particularly in the center). Separate controls are provided for horizontal and vertical centering.

center loading In an inductively loaded antenna, placement of the loading coil(s) at or near a point or points midway between the feed point and the end(s) of the radiating element.

center of beam **1.** In a directional antenna system, the direction, denoted by a straight ray, where the signal strength or response is the greatest. **2.** In a beam of visible light, the geometric center of the spot produced when the beam strikes a surface perpendicular to the beam. **3.** In a beam of visible light, the axis within the beam where the intensity is greatest.

center of channel The frequency that is midway between the lowest and highest frequency components of a communications channel.

center of radiation The point from which the energy radiated by an object appears to arrive.

center tap A connection made to the centermost turn of a coil or to the center-value point of a resistor, filament, or capacitor pair.

center-tapped coil See CENTER-TAPPED WINDING.

center-tapped filament A tube or lamp filament that has a tap at its center.

center-tapped inductor An inductor that has a tap at half the total number of turns (the physical center of the winding).

center-tapped potentiometer A potentiometer that has a tap at half the total resistance of the resistance element.

center-tapped resistor A fixed resistor that has a tap at half the total resistance.

center-tapped transformer A transformer that has one or more center-tapped windings.

center-tapped winding A winding that has a tap at half the total number of turns (the physical center of the winding).

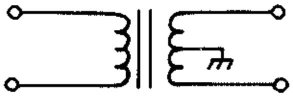

center-tapped winding

center tracking frequency In three-frequency alignment (tracking) of a circuit, the frequency between the upper and lower frequency limits (alignment or tracking points of the circuit).

center-zero meter A meter that has its zero point at the center of the scale (e.g., a dc galvanometer).

centi- Abbreviation, c. Prefix meaning hundredth(s) (10^{-2}).

centigrade scale See CELSIUS SCALE.

centimeter Abbreviation, cm. A unit of length equal to 10^{-2} meter, or 0.3937 inch.

centimeter-gram-second system Abbreviation, cgs. A system of units, now seldom used, in which the centimeter is the fundamental unit of length, the gram is the fundamental unit of mass, and the mean solar second is the fundamental unit of time. Electrical units in the cgs system fall into two categories: *electrostatic* and *electromagnetic.* The names of cgs electrostatic units have the prefix *stat-* (e.g., STATAMPERE, STATVOLT, etc.). Cgs electromagnetic units have the prefix *ab-* (e.g., ABAMPERE, ABVOLT, etc.).

centimetric waves See MICROWAVES.

centipoise A cgs measure of the dynamic viscosity of liquids. Equal to 10^{-2} poise.

central office In telephone systems, a switching network at which numerous circuits or subscriber lines converge.

central processing unit Abbreviation, CPU. In a digital computer, the section containing the arithmetic and logic unit (ALU), control circuits, and internal memory circuits. Also called *central processor.*

Central Radio Propagation Laboratory A government laboratory that studies radio propagation and collects, correlates, and analyzes data for predicting propagation conditions. The organization also studies methods of measuring propagation.

centrifugation potential An electric potential that occurs in a colloidal solution when the solution is centrifuged.

centrifugal switch A switch actuated by rotational motion (e.g., the automatic disconnection switch in a capacitor motor).

centripetal force The force that draws the mass of a rotating body toward the axis of rotation.

ceramal See CERMET.

ceramet seal See CERAMIC-TO-METAL SEAL.

ceramic-based microcircuit A tiny circuit printed or deposited on a ceramic substrate.

ceramic capacitor A component made with sheets of metal stacked alternately with wafers of ceramic. This material, like mica, has low loss, and therefore allows for high efficiency. For low values of capacitance, only one layer of ceramic is needed, and two metal plates can be glued to a disk of porcelain, one on each side. Alternatively, a tube or cylinder of ceramic can be employed, and metal ink applied to the inside and outside of the tube. These capacitors have values ranging from a few picofarads to about 0.5 µF. Their voltage ratings are comparable to those of paper capacitors. Compare ELECTROLYTIC CAPACITOR, MICA CAPACITOR, PAPER CAPACITOR, PLASTIC-FILM CAPACITOR, TANTALUM CAPACITOR.

ceramic dielectric **1.** A ceramic used as a dielectric in capacitors. Examples: barium titanate, barium strontium titanate, and titanium dioxide. Ceramic dielectrics provide high dielectric constant. **2.** A ceramic used as an insulator. Examples: isolantite, porcelain, and steatite.

ceramic filter A resonant filter similar to a crystal filter, but using a piezoelectric ceramic material.

ceramic magnet A permanent magnet made of a magnetic ceramic material, such as mixtures of barium oxide and iron oxide.

ceramic microphone A microphone that uses a CERAMIC PIEZOELEMENT to convert sound waves into electrical impulses.

ceramic piezoelement A component that uses a piezoelectric ceramic material. Examples: ceramic filter, ceramic microphone, ceramic phono pickup, ceramic transducer, and electrostrictive transducer. Also called PIEZOELECTRIC CERAMIC.

ceramic resistor A carborundum resistor whose value is voltage-dependent. It usually displays a negative temperature coefficient of resistance (but a positive coefficient is available) and a negative voltage coefficient of resistance.

ceramics **1.** Clay-based materials used as dielectrics and insulators in electronics. Examples: barium titanate, titanium dioxide, porcelain, isolantite, and steatite. **2.** The science and art of using and developing ceramics.

ceramic-to-metal seal A bond in which ceramic and metal bodies are joined, for example, the bonding of a metal lead to a ceramic disk, through which it passes to provide a leak-proof seal. Also called *ceramet seal.*

ceramic transducer A transducer that uses a CERAMIC PIEZOELEMENT to translate such parameters as pressure and vibration into electrical pulses.

ceramic tube A high-temperature vacuum tube that uses a ceramic material, instead of glass, as the envelope; the tube offers low losses at high frequencies.

Cerenkov radiation Light emanating from a transparent material that is traversed by charged particles, whose speed is higher than the speed of light through the material.

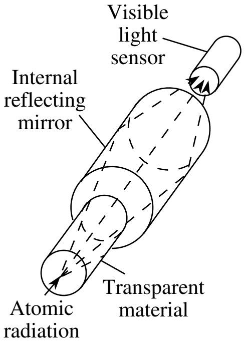

Cerenkov radiation

Cerenkov rebatron device An apparatus for generating radio-frequency energy by passing an electron beam through a piece of dielectric having a small aperture.

ceresin wax A yellow or white wax obtained by refining *ozocerite.* Used as an insulant and sealant against moisture. Dielectric constant, 2.5 to 2.6.

cerium Symbol, Ce. A metallic element of the rare-earth group. Atomic number, 58. Atomic weight, 140.13.

cerium metals A group of metals belonging to the rare-earth group: cerium, lanthanum, neodymium, praseodymium, promethium, and samarium.

cermet An alloy of a ceramic, such as titanium carbide, and nickel, a metal. A thin film of cermet is used as a resistive element in some microcircuits. Cermet is an acronym for *ceramic metal.*

certified tape A magnetic recording tape that has been thoroughly checked and found to have no flaws.

cesium Symbol, Cs. A metallic element of the alkali-metal group. Atomic number, 55. Atomic weight, 132.91. The oscillations of this element's atoms have been used as atomic time standards. The element is used in some phototubes as the light-sensitive material, and in some arc lamps.

cesium-vapor lamp A low-voltage arc lamp used as an infrared source.

Cf Symbol for CALIFORNIUM.

cgs Abbreviation of CENTIMETER-GRAM-SECOND.

chad The punched-out particle(s) constituting refuse from paper-tape punching.

chadded tape Punched paper tape in which the chad is left partially attached to the tape's punched holes.

chadless tape Punched paper tape without CHAD.

chafe **1.** An area that has been abraded by rubbing or scraping. **2.** To produce a chafe.

chaff Strips of metal foil used to create radar interference or ambiguity in locating a target by multiple reflections of the beam. Also called MIRROR.

chain broadcasting Simultaneous transmissions from a number of broadcast transmitters connected together in a network by wire line, coaxial cable, or microwave link.

chain calculation As performed by a calculator, a calculation that can be entered as it would normally be written (i.e., without the need for regrouping operands).

chain printer In the readout channel of a digital computer, a high-speed printer carrying printer's type on a revolving chain.

chain radar system A number of radar stations along a missile-flight path that are connected in a communications or control network.

chain reaction A reaction (as in nuclear fission) that is self-sustaining or self-repeating. Unless controlled from outside, such a reaction runs to destruction.

chain switch A switch that is actuated by pulling a light metal chain. Successive pulls turn the switch alternatively on and off.

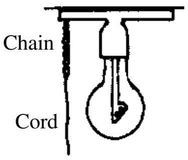

chain switch

change dump In computer operation (especially in debugging), the display of the names of locations that have changed following a specific event.

change file See TRANSACTION FILE.

change of control In a sequence of computer records being processed, a logical break that initiates a predetermined action, after which processing continues.

changer In a high-fidelity disk player, a device that allows several disks to be played, one after the other, without the need for manually exchanging the disks.

change record A computer record that changes information in a related master record. Also called *transaction record.*

change tape See TRANSACTION TAPE.

channel **1.** A frequency (or band of frequencies) assigned to a radio or television station. **2.** See KEYWAY. **3.** A subcircuit in a large system [e.g., the radio-frequency (RF) channel of a receiver, the vertical-amplifier channel of an oscilloscope, or the modulator channel of a radio transmitter]. **4.** The end-to-end electrical path through the semiconductor body in a field-effect transistor. **5.** One of the independent audio circuits in a stereo sound system (e.g., the left channel or the right channel).

channel analyzer A (usually multiband) continuously tunable instrument, similar to a tuned radio receiver, used in troubleshooting radio communications circuits by substituting a perfect channel for one that is out of order.

channel balance The state in which the apparent amplitude of two or more channels is identical.

channel bank In a transmission system, the terminal equipment used for the purpose of multiplexing the individual channels.

channel capacity The fullest extent to which a channel can accommodate the information (frequencies, bits, words, etc.) to be passed through it.

channel designator A name, number, or abbreviation given to a channel in a communications system.

channel effect The possible current flow through a high impedance between the collector and emitter in a bipolar transistor.

channel frequency The CENTER FREQUENCY of a communications channel.

channeling Multiplex transmission in which separate carriers within a sufficiently wide frequency band are used for simultaneous transmission.

channelizing The subdivision of a relatively wide frequency band into a number of separate subbands.

channel reliability **1.** The proportion of time, usually expressed as a percentage, that a communications channel is useful for its intended purpose. **2.** The relative ease with which communications can be carried out over a particular channel.

channel reversal In stereo reproduction, interchanging the left and right channels.

channel-reversing switch In a stereo system, a switch that allows channel reversal without the need for reorienting speaker cables or connectors.

channel sampling rate The rate at which individual channels are sampled. For example, in the electronic switching of an oscilloscope, the number of times per second each input-signal channel is switched to the instrument.

channel selector A switch or relay used to put any of a series of channels into functional status in a system.

channel separation **1.** The spacing between communications channels, expressed in kilohertz. **2.** In stereo reproduction, the degree to which the information on one channel is separate from the other; usually expressed in decibels.

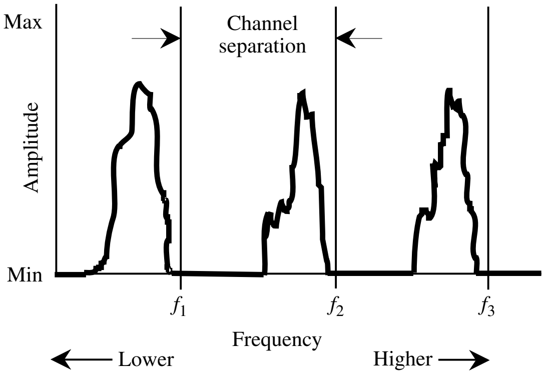

channel separation, 1.

channel slot On a carrier modulated by numerous signals, the position or frequency of a specific modulating signal.

channel shift The interchange of communications channels (e.g., the shift from a calling frequency to a working frequency).

channel strip A fixed-channel amplifier for a television receiver.

channel time slot In a frame of transmitted information, such as a television picture, a time interval designated to a channel for the transmission of a character signal or other information.

channel-to-channel connection A device, such as a channel adapter, used to transfer data rapidly between any two channels of two digital computers, at the data speed of the slower channel.

channel-utilization index An indication of the extent to which channel capacity is used. For a given channel, the index is the ratio of information rate to channel capacity, each expressed in units per second.

channel wave An acoustic wave that travels within a region or layer of a substance because of a physical difference between that layer and the surrounding material. An example of a channel wave is the propagation of sound over a still lake.

channel width In a frequency channel, the difference $f_2 - f_1$, where f_1 is the lower-frequency limit and f_2 is the upper-frequency limit of the channel.

chapter A self-contained computer program section.

character **1.** One of the symbols in a code. **2.** In computer operations, a digit, letter, or symbol used alone or in some combination to express information, data, or instructions.

character code In a communications or computer system, the combination of elements (e.g., bits) representing characters.

character crowding A reduction of the time interval between successive characters—especially those read from tape.

character density The number of characters that can be stored in a given length or surface area of a medium. On a magnetic tape, it might be specified in characters per millimeter; on a magnetic disk, it might be specified in characters per square millimeter.

character emitter A coded-pulse generator in a digital computer.

character generator A device that converts coded information into readable alphanumeric characters.

characteristic **1.** A quantity that characterizes (typifies) the operation of a device or circuit. Examples are emitter current, output power, and frequency deviation. **2.** In floating point notation, the exponent.

characteristic curve A curve showing the relationship between an independent variable and a dependent variable, with respect to the parameter(s) for a device or circuit. Example: the collector voltage-collector current characteristic curve of a transistor.

characteristic distortion **1.** In a digital signal, pulse distortion caused by the effects of the previous pulse or pulses. **2.** Distortion in the characteristic curve of a component or device.

characteristic frequency The frequency peculiar to a given channel, service, or response.

characteristic impedance Symbol, Z_0. **1.** Theoretically, the impedance that would be simulated by a given two-conductor or coaxial line of uniform construction, if that line were of infinite length. This value is determined by the materials used for the two conductors, the dielectric used to insulate the two conductors, the diameters of the conductors, and the spacing between them. **2.** In practice, for a transmission line or waveguide terminated with a load that produces no standing waves on the line, the ratio of radio-frequency (RF) voltage to RF current. This ratio is the same at all points along the length of a perfectly matched line, and depends on the physical construction of the line. Coaxial lines typically have Z_0 between 50 and 100 ohms. Twinlead is available with 75-ohm and 300-ohm Z_0 values. Open-wire line has Z_0 between 300 and 600 ohms, depending on the spacing between the conductors, and also on the type of dielectric (insulating material) employed to keep the spacing constant between the conductors. **3.** Experimentally, the value of impedance that, if it terminates a transmission line or waveguide, results in no reflected power from the load end of line. This is always a pure resistance; that is, it contains no reactance.

characteristic overflow In floating-point arithmetic, the condition that occurs when a characteristic exceeds the upper limit specified by a program or computer.

characteristic spread The range of values over which a characteristic extends. For example, if an amplifier's output ranges from 15 W to 25 W, its characteristic spread is 10 W.

characteristic underflow In floating-point arithmetic, the condition that occurs when a characteristic exceeds the lower limit specified by a program or computer.

character modifier In address modification, a constant (compare VARIABLE) that refers to a specific character's location in memory.

character-oriented A computer in which character locations, rather than words, can be addressed.

character printer A computer output device that prints matter in the manner of a conventional typewriter.

character reader Also called an *optical scanner.* In a digital computer, an input device that can read printing and script directly.

character recognition The reading of a written or printed character by a computer, including its identification and encoding.

character sensing The detection of characters by a computer input device. This can be done galvanically, electrostatically, magnetically, or optically.

character set The set of characters in a complete language, or in a communications system.

character signal The set of elements or bits representing a character in a digital transmission system. The signal can also represent the quantizing value of a sample.

characters per minute An expression of the speed of transmission of a digital signal. The number of characters (on average) transmitted in a period of one minute. In Morse code (CW) transmission, this is generally taken as the number of times the word *paris* plus the subsequent space, multiplied by six (five letters and one space following), can be sent in one minute.

characters per second An expression of the speed of transmission of a digital signal. The number of characters (on average) transmitted in a period of one second.

character string A one-dimensional character array [i.e., a list of characters that, when printed or displayed, would appear in a row or column, but not both (as in a matrix)].

character subset A classification of characters within a set.

Charactron A cathode-ray readout tube that displays letters, numbers, and symbols on its screen. More commonly called a *monitor.*

charcoal tube In a system for producing a high vacuum, a trap containing activated charcoal, which is heated to dull red, then cooled by liquid air to absorb gases.

charge **1.** A quantity of electricity associated with a space, particle, or body. **2.** To electrify a space, particle, or body (i.e., to give an electric charge). **3.** To store electricity, as in a storage battery or capacitor. Compare DISCHARGE.

charge carrier **1.** An ELECTRON whose movement constitutes a flow of electric current. **2.** An electron deficiency (HOLE) whose movement constitutes a flow of electric current. **3.** Any particle, such as a charged atom (ION), PROTON, ALPHA PARTICLE, or BETA PARTICLE, whose movement constitutes a flow of electric current.

charge-coupled device Abbreviation, CCD. A form of analog-to-digital converter that generates a digital signal output representing an analog image input. The transfer of stored charges provides the method of operation. Used in machine vision systems and in numerous scientific applications.

charge density The degree of charge or current-carrier concentration in a region.

charged particle **1.** See CHARGE CARRIER. **2.** See ION.

charged voltage **1.** The voltage across a fully charged capacitor. **2.** The terminal voltage of a fully charged storage cell.

charge holding See CHARGE RETENTION.

charge of electron The negative electric charge carried by a single electron. Approximately equal to 1.602×10^{-19} coulombs.

charger **1.** See BATTERY CHARGER. **2.** Any device or circuit that charges a capacitor.

charge retention **1.** The holding of an electric charge by a cell or battery when no current is being drawn from it. **2.** A measure of the ability of a cell or battery to maintain an electric charge when no current is drawn from it. Often specified in terms of *shelf life.* **3.** The holding of a charge by a capacitor.

charge-storage tube A cathode-ray tube that holds a display of information on its screen until the operator removes it by pressing an erase button.

charge-to-mass The ratio of the electric charge to the mass of a subatomic particle.

charge-to-mass ratio of electron The ratio of the charge (e) of the electron to the mass (m_e) of the electron, in coulombs per kilogram (C/kg). For an electron at rest, e/m_e is approximately equal to 1.602×10^{-19} C divided by 9.11×10^{-31} kg = 1.76×10^{11} C/kg.

charge transfer **1.** The switching of an electric charge from one capacitor to another. **2.** The capture of an electron by a positive ion from a neutral atom of the same kind, resulting in the ion becoming a neutral atom, and the previously neutral atom becoming a positive ion.

charge transfer device A semiconductor in which an electric charge is moved from location to location. Applications include delay lines, video signal processing, and signal storage.

charging **1.** The process of storing electrical energy in a capacitor. **2.** The process of storing electrochemical energy in a storage cell or battery.

charging current **1.** The current flowing into a capacitor. **2.** The current flowing into a previously discharged storage cell.

charging rate **1.** The rate at which charging current flows into a storage cell or battery, expressed in amperes or milliamperes. For most cells and batteries, the rate is greatest initially, when the cell or battery is depleted or nearly depleted; the rate decreases as the cell or battery becomes charged. **2.** The instantaneous rate at which charging current flows into a capacitor or capacitance-resistance circuit, expressed in amperes, milliamperes, or microamperes.

charged voltage **1.** The voltage across a fully charged capacitor. **2.** The terminal voltage of a fully charged storage cell.

Charlie Phonetic alphabet code word for the letter C.

chassis A (usually metal) foundation on which components are mounted and wired.

chassis ground A ground connection made to the metal chassis on which the components of a circuit are mounted. When several ground connections are made to a single point on the chassis, a COMMON GROUND results.

chatter **1.** A rapidly repetitive signal, caused by interruption or variation of a current (usually interference). **2.** Extraneous vibration, as of the armature in a relay.

chatter time The interval between the instant that contacts close (for example, in a relay) and the instant at which chatter ends.

cheater cord An extension cord used to conduct power to a piece of equipment (especially a television receiver) by temporarily bypassing the safety switch or interlock. Use of such a cord presents a potentially fatal shock hazard to personnel using, or working on, the equipment.

Chebyshev filter Also spelled *Tschebyscheff* or *Tschebysheff.* A form of inductance-capacitance (LC) lowpass, highpass, bandpass, or band-rejection filter, characterized by an attenuation-versus-frequency curve with ripple in the passband.

check **1.** A test generally made to verify condition, performance, state, or calculations; specifically, in computer operations, it applies to operands or results. **2.** The usually abrupt halting of an action.

check bit A binary CHECK DIGIT.

check character In a group of characters, one whose value depends on the other characters, which it checks when the group is stored or transferred.

check digit Also called *check number.* In computer operations, a number added to a group of digits, forming a code that identifies entities in the system (including personnel) and can be used for verification. The check digit is the remainder when the number code (for example, 459) is divided by a fixed number (for example, 5); in this case, the check digit (the remainder of 459/5) is 4, and the amended code number is 4594.

check indicator An indication, made via a video display, that something has been shown to be invalid according to a check.

checking program Also called *checking routine.* For debugging purposes, a diagnostic computer program capable of detecting errors in another program.

checkout A test routine that ascertains whether or not a circuit or system is functioning according to specifications.

checkout routine A routine used by programmers to debug programs.

checkpoint A point in a digital-computer program at which sufficient information has been stored to allow restarting the computation from that point.

checkpoint dump The process of recording details of a computer program run. This process might be necessary in the event of a system failure that requires reconstruction of a program or programs.

checkpointing The writing of a computer program in such a manner that, during a program run, information is frequently dumped as insurance against possible loss in the event of a system failure.

check problem A presolved problem used to check the operation of a digital computer or program.

check register In some digital computers, a register in which transferred information is stored so that it can be checked against the same information as it is received a second time.

check routine A special program designed to ascertain if a program or computer is operating correctly. Also see CHECK PROBLEM.

checksum Used as part of a summation check, a sum derived from the digits of a number. For example, the checksum of 23,335 is 16. Also called HASH TOTAL.

check symbol For a specific data item, a digit or digits obtained by performing an arithmetic check on the item, which it then accompanies through processing stages for the purpose of checking it.

check total See CONTROL TOTAL.

check word A check symbol in the form of a word added to, and containing data from, a block of records.

chelate Pertaining to cyclic molecular structure in which several atoms in a ring hold a central metallic ion in a COORDINATION COMPLEX.

chemical deposition The coating of a surface with a substance resulting from *chemical reduction* of a solution. In mirror making, for example, formaldehyde reduces a solution of silver nitrate, and deposits metallic silver on the surface of polished glass. Also see CHEMICALLY DEPOSITED PRINTED CIRCUIT and CHEMICAL REDUCTION.

chemical detector See ELECTROLYTIC DETECTOR.

chemical effect An alteration in the chemical makeup of a substance or solution, resulting from the passage of an electric current through it. Examples include electrolysis, electroplating, and the reduction of ores.

chemical energy Energy that is stored in the chemical bonds of a material or solution. An example is the stored energy in terms of watt hours in an electrolytic cell.

chemical load An arrangement of a chemical material or device for the passage of electricity through it. Examples: electroplater, electrolytic cell for the production of hydrogen gas, and storage battery.

chemically deposited printed circuit A printed circuit in which the pattern of metal lines and areas are chemically deposited on a substrate.

chemically pure Abbreviation, CP. Free from impurities.

chemical rectifier See ELECTROLYTIC CELL.

chemical reduction The process of making a chemical compound (usually in solution) into a metal, by removing the nonmetallic component from the compound. For example, when copper oxide is heated in the presence of hydrogen (a reducting agent), the oxygen (the nonmetallic component) is driven out, and copper (along with some water) remains.

chemical resistor See ELECTROLYTIC RESISTOR.

chemical switch See ELECTROCHEMICAL SWITCH.

CHIL Abbreviation for *current-hogging injection logic.* A form of bipolar digital logic technology.

chip **1.** An INTEGRATED CIRCUIT. **2.** A small slab, wafer, or die of dielectric or semiconductor material, on which a subminiature component or circuit is formed or deposited.

chip capacitor A subminiature capacitor formed on a chip.

chip resistor A subminiature resistor formed on a chip.

chip tray A chad receptacle located at a card or paper tape punching site.

Chireix-Mesny antenna A high-frequency (HF) beam antenna, in which each dipole section constitutes one side of a diamond. Cophased horizontal and vertical components of current flow in each of the diagonals, and radiation is broadside to the plane of the driven element.

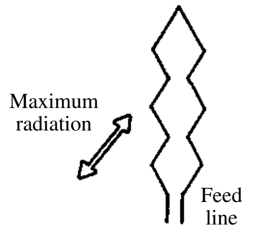

Chireix-Mesny antenna

chirp A rapid change in the frequency of a continuous-wave Morse-code signal. The chirp usually occurs at the beginning of each dot or dash, and can go up or down in frequency. Chirp occurs because of a change in the output impedance of an oscillator as it is keyed. Modern code transmitters do not exhibit significant chirp.

chirp modulation A form of modulation in which the frequency of a signal is deliberately changed in a systematic way. Used in some radar systems.

chirp radar A radar system that uses CHIRP MODULATION.

Chladni's plates Conducting plates that are used to evaluate the nature of a vibration in a solid material. The plates are clamped to the material, and sand is sprinkled on the surface. This produces patterns that indicate the nature of the vibrations.

chlorinated diphenyl A synthetic organic substance used as an impregnant in some oil-filled capacitors.

chlorinated naphthalene See HALOWAX.

chlorine Symbol, Cl. A gaseous element of the halogen family. Atomic number, 17. Atomic weight, 35.453.

choke **1.** To restrict or curtail passage of a particular current or frequency by means of a discrete component, such as a choke coil. **2.** See CHOKE COIL.

choke air gap A fractional-inch opening in the iron core of a filter choke, usually filled with wood or plastic. The gap prevents saturation of the core when the choke coil carries maximum rated direct current.

choke coil **1.** A large-value inductor that provides a high impedance to alternating current (ac), while offering virtually no opposition to direct current (dc). **2.** In radio-frequency (RF) applications, an inductor that provides a high impedance to RF signals while showing low impedance for audio-frequency (AF) signals and direct currents (dc).

choke-coupled modulation An amplitude-modulation (AM) scheme, in which the modulator is coupled to the radio-frequency (RF) amplifier through a shared iron-core choke coil.

choke flange At the end of a waveguide, a flange in which a groove forms a CHOKE JOINT.

choke-input filter A filter whose input component is an inductor (choke). The choke-input power-supply filter is distinguished by its superior voltage regulation, compared with the CAPACITOR-INPUT FILTER.

choke joint A joint connecting two waveguide sections and permitting efficient energy transfer without requiring electrical contact with the inside wall of the waveguide.

chopped dc See INTERRUPTED DC.

chopped mode In a single-gun cathode-ray-tube (CRT) oscilloscope, a technique for sequentially displaying several signals that are not referenced to the oscilloscope sweep.

chopped signal An ac or dc signal that is periodically interrupted, as by means of a CHOPPER.

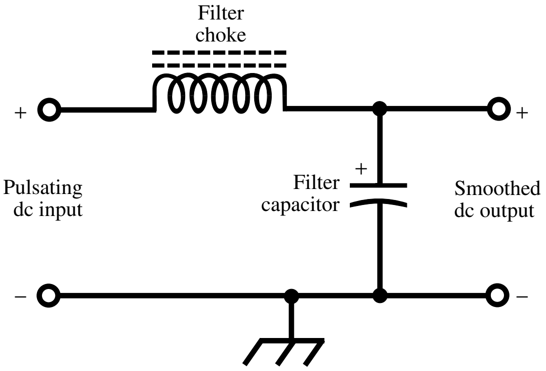

choke-input filter

chopper A device or circuit that interrupts a direct current (dc) at some predetermined rate. Ideally, such a device is characterized by distinct on and off operation.

chopper amplifier A circuit that amplifies the output of a CHOPPER. Used in conjunction with a CHOPPER CONVERTER in dc amplification.

chopper converter A device that interrupts a direct current (dc), and changes it to a pulsating, rectangular-wave current or voltage that can be handled by a stable ac amplifier and rectified to supply amplified dc.

chopper power supply Also called *power inverter.* A circuit that delivers high-voltage ac from a dc source. The input is typically 12 volts dc, and the output is usually 117 volts rms ac. These devices facilitate the use of small appliances such as computers, television sets, and communications radios in portable and mobile environments. The output of a low-cost power inverter is generally not a good sine wave. More sophisticated inverters produce good sine waves and have a frequency close to 60 Hz.

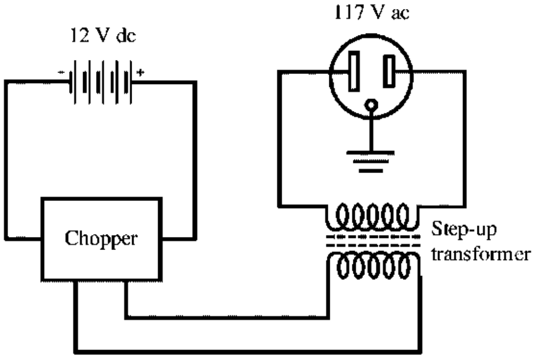

chopper power supply

chopper stabilization **1.** Stabilization of direct-current (dc) amplification by using a CHOPPER CONVERTER ahead of a stable ac amplifier, and rectifying the amplifier output. **2.** In a regulated power supply, use of a CHOPPER AMPLIFIER at the control-circuit input to improve regulation.

chopper-stabilized amplifier See CHOPPER AMPLIFIER and CHOPPER STABILIZATION, **1.**

chopper transistor A transistor that provides rapid and repeated on/off switching of direct current (dc), in the manner of an electromechanical interrupter. See CHOPPER.

chopping frequency The frequency at which a chopper interrupts a signal.

chord **1.** A harmonious mixture of musical tones of various frequencies. **2.** A straight line that joins two points on a curve (such as an arc of a circle). **3.** The width of an airfoil.

chord organ An electronic organ that will sound a musical chord when a key is pressed (see CHORD, **1**).

choreographer program A computer program similar to one originally written by Charles Lecht of Lecht Sciences, Inc. The computer operator gives commands that cause a human form, portrayed on the display screen, to make various movements. Used in animated computer graphics.

chorus Signals at very low radio frequencies (VLF), natural in origin, that sweep upward in frequency. Believed to result from lightning-generated electromagnetic fields that circulate in the magnetosphere (earth's magnetic field). The term is derived from the sound the signals make in high-gain audio-frequency (AF) amplifiers connected directly to VLF receiving antennas.

Christiansen antenna A radio-telescope antenna for obtaining high resolution. Two straight arrays are placed at an angle, intersecting approximately at their centers. The resulting interference pattern has extremely narrow lobes.

Christmas tree A tree-like pattern on the screen of a television receiver, caused by loss of horizontal synchronization.

chroma The quality of a color: *hue* and *saturation.*

chroma circuit In color television, one of several circuits whose ultimate purpose is to produce a color component on the screen.

chroma-clear raster In color television reception, the clear raster resulting from a white video signal, or from operation of the chroma circuits of the receiver (as if they were receiving a white transmission). Also called *white raster.*

chroma control In a color television receiver, a rheostat or potentiometer that permits adjustment of color saturation through variation of the chrominance-signal amplitude before demodulation.

chromatic fidelity See COLOR FIDELITY.

chromaticity **1.** The state of being chromatic (see CHROMA). **2.** A quantitative assessment of a color in terms of dominant or complementary wavelength and purity.

chromaticity coordinate For a color sample, the ratio of any one of the three tristimulus values (primary colors) to the sum of the three.

chromaticity diagram A rectangular-coordinate graph in which one of the three CHROMATICITY COORDINATES of a three-color system is plotted against another coordinate.

chromaticity flicker Flicker caused entirely by chromaticity fluctuation (see CHROMATICITY, **2**).

chromel A nickel-chromium alloy with some iron content, used in thermocouples.

chromel-alumel junction A thermocouple that uses wires of the alloys *chromel* and *alumel.*

chromel-constantan thermocouple A thermocouple consisting of a junction between wires or strips of *chromel* and *constantan.* Typical output is 6.3 mV at 100°C.

chrome plating The process of coating a metal with chromium. Generally protects against corrosion.

chrome recording tape Also called *chrome tape* or *chromium tape.* Tape that is manufactured from the compound chromium dioxide. Noted for its ability to faithfully record and reproduce music.

chrominance In color television, the difference between a reproduced color and a standard reference color of the same luminous intensity.

chrominance amplifier In a color television circuit, the amplifier separating the chrominance signal from the total video signal.

chrominance cancellation On a black-and-white picture tube screen, cancellation of the fluctuations in brightness caused by a chrominance signal.

chrominance-carrier reference In color television, a continuous signal at the frequency of the chrominance subcarrier; it is in fixed phase with the color burst and provides modulation or demodulation phase reference for carrier-chrominance signals.

chrominance channel In color television, a circuit devoted exclusively to the color function, as opposed to audio and general control channels.

chrominance component In the NTSC color television systems, either of the components (I-signal or Q-signal) of the complete chrominance signal.

chrominance demodulator In a color television receiver, a demodulator that extracts video-frequency chrominance components from the chrominance signal, and a sine wave from the chrominance subcarrier oscillator.

chrominance gain control A rheostat or potentiometer in the red, green, and blue matrix channels of a color television receiver, used to adjust the primary-signal amplitudes.

chrominance modulator In a color television transmitter, a device that generates the chrominance signal from the I and Q components and the chrominance subcarrier.

chrominance primary One of the transmission primaries (red, green, and blue) upon which the chrominance of a color depends.

chrominance signal The signal component in color television that represents the hues and saturation levels of the colors in the picture.

chrominance subcarrier In color television, the 3579.545-kHz signal that serves as a carrier for the I- and Q-signals.

chrominance-subcarrier oscillator In a color television receiver, a crystal-controlled oscillator that generates the subcarrier signal (see CHROMINANCE SUBCARRIER).

chrominance video signals Output signals from the red, green, and blue channels of a color television camera or receiver matrix.

chromium Symbol, Cr. A metallic element. Atomic number, 24. Atomic weight, 51.996. Commonly used as a plating for metals to improve resistance to corrosion.

chronistor An elapsed-time indicator in which current, flowing during a given time interval, electroplates an electrode. The duration of the interval is determined from the amount of deposit.

chronograph **1.** An instrument that provides an accurate time base along the horizontal axis of its permanent record. **2.** Stopwatch.

chronometer A precision clock. Electronic chronometers often use a highly accurate and stable crystal oscillator, followed by a string of multivibrators to reduce the crystal frequency to an audio frequency (such as 1 kHz) that drives the clock motor.

chronoscope An instrument for precisely measuring small time intervals.

CHU Call letters of the Canadian time-signal station whose primary frequency is 7.335 MHz.

Ci Symbol for INPUT CAPACITANCE.

CIE Abbreviation for *International Commission on Illumination.*

cinching In a reel of magnetic tape, the slipping of tape as force is applied.

cinematograph See KINEMATOGRAPH.

cipher A code used for the purpose of preventing interception of a message by third parties.

circ **1.** Abbreviation of *circuit.* **2.** Abbreviation of *circular.*

circle graph Also called a *pie graph.* A representational device consisting of a disk subdivided into various triangular areas (radiating from the center of the circle), which are proportional to represented quantities.

circle of confusion A circular image of a point source of light, resulting from an aberration in an optical system.

circle of declination The graduated circular scale of a *declinometer.*

circlotron amplifier A high-powered microwave amplifier of the one-port, cross-field, nonlinear type using a magnetron.

circuit **1.** A closed path through which current flows from a generator, through various components, and back to the generator. (An electronic circuit is often a combination of interconnected subcircuits.) **2.** The wiring diagram of an electronic device or system.

circuit analysis The careful determination of the nature and behavior of a circuit and its various parts. It can be theoretical, practical, or both. Compare CIRCUIT SYNTHESIS.

circuit analyzer See CIRCUIT TESTER.

circuit board A panel, plate, or card on which electronic components are mounted and interconnected to provide a functional unit.

circuit breaker A resettable fuse-like device that is designed to protect a circuit against overloading. In a typical circuit breaker, the winding of an electromagnet is connected in series with the load circuit and with the switch contact points. Excessive current through the magnet winding causes the switch to be opened.

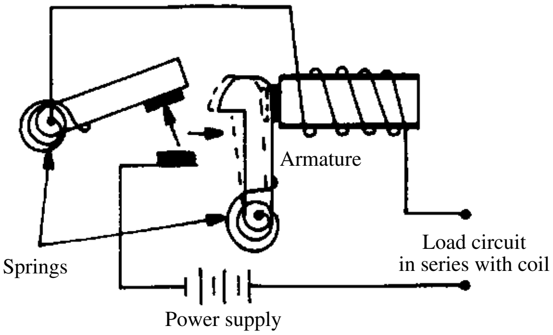

circuit breaker

circuit capacitance The total capacitance (lumped, distributed, and stray) present in a circuit.

circuit capacity **1.** The ability of a circuit to handle a quantity (such as current, voltage, frequency, power, etc.) safely and efficiently. **2.** The maximum value of some parameter at which a circuit can function safely and efficiently (e.g., a circuit capacity of 50 A). **3.** The number of channels that can be accommodated simultaneously by a circuit.

circuit component **1.** Any of the electronic devices or parts (capacitors, resistors, transistors, etc.) that are connected through wiring to form a circuit. **2.** An electrical quantity required for, or arising from, circuit operation. Examples: input voltage, feedback current, stray capacitance, and circuit noise.

circuit diagram A drawing in which symbols and lines represent the components and wiring of an electronic circuit. Also called CIRCUIT

SCHEMATIC, SCHEMATIC DIAGRAM, and WIRING DIAGRAM.

circuit dropout A momentary interruption of circuit operation, often caused by a break in the circuit.

circuit efficiency A quantitative measure of the effectiveness of circuit operation, customarily expressed as the ratio of the useful output power to the total input power.

circuit element See CIRCUIT COMPONENT, **1.**

circuit engineer An electronics engineer who specializes in circuit analysis, circuit synthesis, or both.

circuit fault **1.** Malfunction of a circuit. **2.** An error in circuit wiring.

circuit hole A perforation within the conductive area of a printed-circuit board, for the insertion and connection of a pigtail, terminal, etc., or for connecting the conductors on one side of the board with those on the other.

circuit loading Intentionally or unintentionally drawing power from a circuit.

circuit noise **1.** Electrical noise generated by a circuit in the absence of an applied signal. **2.** In wire telephony, electrical noise as opposed to acoustic noise.

circuit noise level The ratio of circuit-noise amplitude to reference-noise amplitude, expressed in decibels above the reference amplitude.

circuit-noise meter A meter that measures the intensity of the noise generated within a circuit.

circuit parameter See CIRCUIT COMPONENT, **2.**

circuit protection Automatic safeguarding of a circuit from damage from overload, excessive drive, heat, vibration, etc. Protection is afforded by various devices and subcircuits, ranging from the common fuse to sophisticated limiters and breakers.

circuit reliability A quantitative indication of the ability of a circuit to provide dependable operation as specified. See MEAN TIME BEFORE FAILURE and MEAN TIME BETWEEN FAILURES.

circuitry **1.** Collectively, electronic and electrical circuits. **2.** A detailed plan of a circuit and its subcircuits. **3.** Collectively, the components of a circuit.

circuit schematic See CIRCUIT DIAGRAM.

circuit simplification **1.** In circuit analysis, the reduction of a complex circuit to its simplest representation to minimize labor and to promote clarity. Thus, through application of Kirchhoff's laws, a complicated circuit could theoretically be reduced to a single generator in series with a single impedance. **2.** In circuit synthesis, the arrangement of a circuit so as to provide desired performance with the fewest components and least-complex wiring.

circuit switching In telephony, a method of connection in which a single circuit is maintained between two subscribers for the entire duration of the call. The signal path does not change. The connection is maintained even during periods of silence (no data transmitted by either subscriber). Compare PACKET SWITCHING.

circuit synthesis The development of a circuit under the guidance of theoretical or practical knowledge of basic electronics principles and component parameters. Compare CIRCUIT ANALYSIS.

circuit tester An instrument for checking the performance of electronic circuits. Often consists of a specialized continuity tester, but occasionally it includes a dynamic performance tester.

circuit tracking The alignment and/or pretuning of circuits for identical or optimum response. It applies especially to cascaded circuits, whose variable elements, such as tuned inductance-capacitance (LC) networks, must follow each other in step when ganged together.

circular angle The angle described by a radius vector as it rotates counterclockwise around a circle.

circular antenna A half-wave horizontally polarized antenna, whose driven element is a rigid conductor bent into a circle with a break opposite the feed point. Also called *halo antenna.* Used primarily at very-high frequencies (VHF).

circular electric wave An electromagnetic wave with circular electric lines of flux. An example is the field in the immediate vicinity of a CIRCULAR ANTENNA.

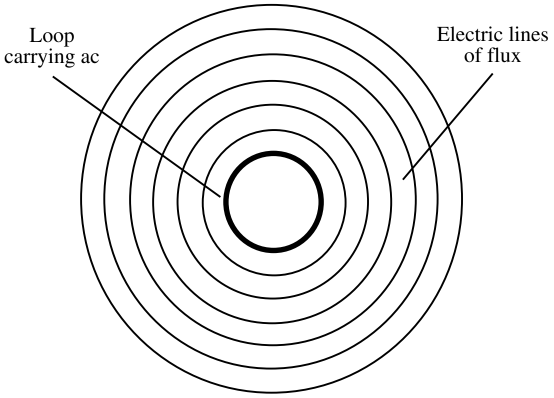

circular electric wave

circular functions Trigonometric functions of the angle described by a vector rotating counterclockwise around a circle. Also see COSINE, COSECANT, COTANGENT, SECANT, SINE, and TANGENT.

circular magnet See RING MAGNET.

circular magnetic wave An electromagnetic wave in which the magnetic lines of flux are circular. An example is the field in the immediate vicinity of a straight-conductor antenna.

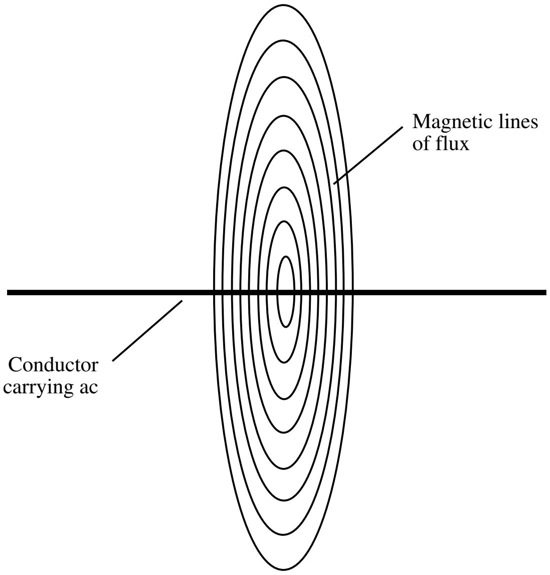

circular magnetic wave

circular mil A unit of cross-sectional area equivalent to 0.785 millionths of a square inch, or the area of a circle having a diameter of 0.001 inch. Generally, the circular mil is used to specify the cross-sectional area of a conductor, such as wire.

circular mil foot A unit of volume in which the length is 1 foot and the cross-sectional area is 1 circular mil.

circular polarization A form of electromagnetic-wave polarization in which the orientation of the electric flux rotates continuously and uniformly as the wave propagates through space. Circular polarization can occur in either a clockwise or counterclockwise sense.

circular radian The angle enclosed by two radii of a unit circle and subtended by a unit arc. Equal to about 57.296 angular degrees.

circular scan A radar scan in which the electron-beam spot describes a circle centered around the transmitting antenna.

circular sweep In an oscilloscope, a sweep obtained when the horizontal and vertical sinusoidal deflecting voltages have the same amplitude and frequency, but are out of phase by 90 degrees (1⁄4 cycle).

circular trace An oscilloscope pattern consisting of a circle obtained with a circular sweep of the electron beam.

circular waveguide A waveguide with a circular cross section.

circulating register In a digital computer, a register in which digits are taken from locations at one end and returned to those at the other end.

circulating tank current The alternating current that oscillates between the capacitor and inductor within a tank circuit.

circulator A multi-terminal coupler in which microwave energy is transmitted in a particular direction from one terminal to the next.

circumvention In a security or alarm system, the evasion of detection. Can be done by physically avoiding regions of coverage, or by defeating the system electronically.

cis A prefix meaning “on this side of.” For example, the *cislunar field* is the field on this side of the moon.

Citizen Band Abbreviation, CB. A band of radio frequencies allocated for two-way communication between private citizens (apart from amateur and commercial services).

Citizens Radio Service Two-way radio communication in a CITIZEN BAND. In the United States, the FCC licenses users of this service without requiring them to take an examination.

C/kg Abbreviation of *coulombs per kilogram,* the unit for electron charge-to-mass ratio.

C/kmol Abbreviation of *coulombs per kilomole,* the unit for the Faraday constant.

ckt Abbreviation of CIRCUIT.

Cl Symbol for CHLORINE.

cl Abbreviation of CENTILITER.

cladding The bonding of one metal to another to minimize or prevent corrosion. A common example is copper-clad steel wire, ideal for use in radio-frequency antenna systems. The copper provides excellent conduction, and the steel provides high tensile strength with a minimum of wire stretching.

clamper A device that restricts a wave to a predetermined dc level. Also called DC RESTORER.

clamping **1.** Fixing the operation of a device at a definite dc level. Also see CLAMPER. **2.** In television, establishing a fixed level for the picture signal at the start of each scanning line.

clamping circuit See CLAMPER.

clamping diode A diode used to fix the voltage level of a signal at a particular reference point.

clapper In a bell, the ball or hammer that strikes the bell; in an electric bell, it is affixed to the vibrating armature.

Clapp-Gouriet oscillator A *Colpitts oscillator* in which a capacitor is connected in series with the inductor. The circuit offers high frequency stability in the presence of input and output capacitance variations.

Clapp oscillator A series-tuned hybrid *Colpitts oscillator,* having a tuning capacitor in series with the inductor, rather than in parallel with the inductor. The circuit allows the use of a smaller tuning capacitor, resulting in improved stability.

Clark cell See ZINC STANDARD CELL.

class-A amplifier An amplifier whose bias is set at approximately the midpoint of the characteristic curve. Output electrode current flows during the

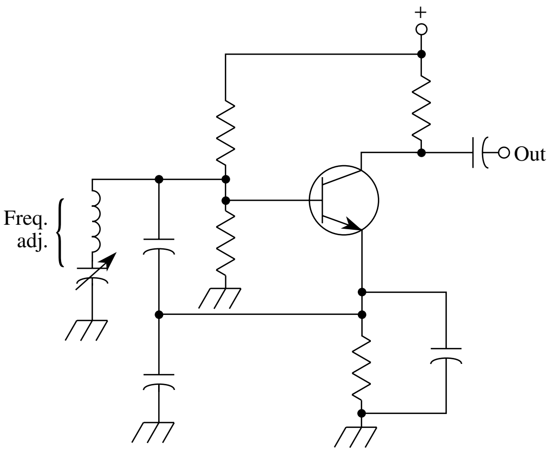

Clapp oscillator

complete ac driving-voltage cycle. The input signal never drives the device into the nonlinear portion of the characteristic curve.

class-AB amplifier Either a CLASS-AB$_1$ AMPLIFIER or a CLASS-AB$_2$ AMPLIFIER.

class-AB$_1$ amplifier An amplifier whose bias is adjusted to a level between that of a class-A amplifier and that of a class-AB$_2$ amplifier. Output electrode current flows during the entire ac driving-voltage cycle. The input signal drives the device into the nonlinear portion of the characteristic curve during part of the cycle.

class-AB$_2$ amplifier An amplifier whose bias is adjusted to a level between that of a class-AB$_1$ amplifier and that of a class-B amplifier. Output electrode current flows during more than 50 percent, but less than 100 percent, of the input signal cycle.

class-AB modulator A modulator whose output stage is a class-AB$_1$ or class-AB$_2$ amplifier.

class-A modulator A circuit for obtaining amplitude-modulated signals; essentially a class-A amplifier with two inputs, one for the carrier and the other for the modulating signal.

class-A operation The operation of a transistor, field-effect transistor, or vacuum tube, in which the collector, drain, or plate current flows during the entire signal cycle.

class-B amplifier An amplifier whose bias is adjusted to operate at the cutoff point in the characteristic curve. Output current flows during approximately 50 percent of the input signal cycle. Efficiency is higher than that of a class-A amplifier.

class-B modulator A push-pull modulator whose output stage is a class-B amplifier.

class-B operation The operation of a transistor, field-effect transistor, or vacuum tube, in which the collector, drain, or plate current flows for approximately half the signal cycle.

class-C amplifier An amplifier whose input-electrode bias is adjusted for operation at a point considerably beyond cutoff. Output current flows during less than half of the input signal cycle. Such an amplifier requires comparatively high driving power, but is capable of excellent efficiency. Commonly used in continuous-wave (CW), amplitude-modulated (AM), and frequency-modulated (FM) radio transmitters.

class-C operation The operation of a transistor, field-effect transistor, or vacuum tube, in which the collector, drain, or plate current flows for significantly less than half the signal cycle.

class-D telephone A telephone restricted to use by emergency services, such as fire departments and guard alarm installations.

classical electron radius Abbreviated r_e. The quantity expressed as $e^2/(m_e c^2)$, where e is the electron's charge in electrostatic units, m_e is its rest mass, and c is the speed of light. The value r_e is equal to approximately 2.82×10^{-13} cm or 2.82×10^{-15} m.

clean room A room for the assembly or testing of critical electronic equipment. The term is derived from the extraordinary steps taken to remove dust and other contaminating agents. The personnel wear carefully cleaned garments (or disposable clothing), gloves, caps, and masks; in some situations, they are required to walk between ceiling and floor ducts of a vacuum system upon entering the room.

cleanup process In the process of electron tube evacuation, a technique used to remove residual and occluded gases from the vacuum apparatus and from the device being evacuated.

clear **1.** In computer operations, to restore a switching element (e.g., a flip-flop) or a memory element to its standard (e.g., zero) state. **2.** In computer practice, an asynchronous input.

clearance The distance between two live terminals, or between one live terminal and ground.

clear band In optical character recognition, the part of a document that must remain unprinted.

clear channel **1.** A channel in the standard amplitude-modulation (AM) broadcast band that is designated to only one station within the area covered by the signal from that station. **2.** In television broadcasting, a channel for which there are no restrictions on the nature of the programming.

clear memory A function in a calculator or small computer that erases the contents of the memory.

clear raster The raster on the screen of a television picture tube in the absence of a signal, noise, or faulty beam deflection.

cleavage In a crystalline substance, the quality of splitting along definite planes. Also, a fragment resulting from such a cleft.

click filter See KEY-CLICK FILTER.

click method An emergency technique for rendering an electric current audibly detectable, by making and breaking the circuit carrying the current to a headset or earphone. A single click results from each make and each break. Also see TIKKER.

click suppressor See KEY-CLICK FILTER.

climate chamber A test chamber that provides accurately controlled temperature, humidity, and/or barometric pressure, for evaluating the performance of electronic components and circuits. Also called ENVIRONMENTAL TEST CHAMBER.

climatometer An instrument incorporating a hygrometer and bimetallic thermometer, whose dial pointers intersect to indicate comfort zones (best temperature-to-humidity ratio).

clinometer An electromechanical device that measures the steepness of a slope. When the device is level (horizontal), the output voltage is zero. If the device is tipped in one direction, a negative voltage is produced; if it is tipped in the other direction, a positive voltage is produced. The output voltage is proportional to the angle at which the device is tipped. Used in mobile robots.

clip A pinch-type connector whose jaws are normally held closed by a spring.

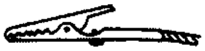

clip

clipped-noise modulation Modulation of a jamming signal through clipping action to increase the sideband energy and resulting interference.

clipper A circuit whose output voltage is fixed at a value for all input voltages higher than a predetermined value. Clippers can flat-top the positive, negative, or both positive and negative peaks of an input voltage.

clipper amplifier An amplifier operated so that the positive, negative, or both positive and negative peaks are clipped in the output signal. The clipping action results from feeding a regular symmetric waveform into an amplifier so that on negative excursion extremes, the stage is cut off; on positive excursion extremes, the amplifier is driven into saturation.

clipper limiter A device that delivers an output signal whose amplitude range corresponds to input-signal voltages between two predetermined limits. It can be used as a *noise limiter* with an element or elements that clip all pulses whose amplitudes are greater than the signal being processed.

clipping **1.** Leveling off (flat-topping) a signal peak at a predetermined level. Also see CLIPPER. **2.** In audio practice, the loss of syllables or words because of cutoff periods in the operation of the circuit (usually caused by overdriving a stage).

clock In a digital computer or controller, the device or circuit that supplies timing pulses to pace the operation of the system.

clocked flip-flops A master-slave arrangement of direct-coupled flip-flops. Information entered into the master unit when the input-trigger pulse amplitude is high is transferred to the slave unit when the amplitude is low.

clock frequency In a digital computer or control, the reciprocal of the period of a single cycle, expressed in terms of the number of cycles occurring in one second of time (hertz, kilohertz, or megahertz).

clock generator A test-signal generator that supplies a chain of pulses identical to those supplied by the clock of a digital computer.

clock module A complete plug-in or wire-in digital unit whose readout indicates time of day or elapsed time. Connected to a suitable power supply, it serves as either a clock or timer.

clock pulse A time-base pulse supplied by the clock of a digital computer, expressed as a period whose reciprocal is frequency.

clock rate See CLOCK FREQUENCY.

clock track On a magnetic tape or disk for data storage, a track containing read or write control (clock) pulses.

clockwise Abbreviation, cw. Rotation in a right-hand direction around a circle, starting at the top. Compare COUNTERCLOCKWISE.

clockwise-polarized wave An elliptically polarized electromagnetic wave whose electric-intensity vector rotates clockwise, as observed from the point of propagation. Compare COUNTER-CLOCKWISE-POLARIZED WAVE.

clone A machine manufactured by a relatively unknown company that performs all the same functions, in basically the same way, as another machine manufactured by a well-known, major corporation. The term is used especially in reference to computers and computer peripherals. If a device is compatible with a certain computer, then clones of that device are generally compatible with that computer. Also, the device is likely to be compatible with all clones of the computer.

close coupling Also called *tight coupling.* In a transformer, the placement of the primary and secondary coils as close together as possible for maximum energy transfer. Compare LOOSE COUPLING.

closed capacitance The value of a variable capacitor whose rotor plates are completely meshed with the stator plates. Compare OPEN CAPACITANCE.

closed circuit A continuous unbroken circuit (i.e., one in which current can flow without interruption). Compare OPEN CIRCUIT.

closed-circuit cell A primary cell, such as the early gravity cell, designed for heavy and polarization-free service.

closed-circuit communication Communication between units only within a defined, hard-wired system, not extending to other units or systems.

closed-circuit security system An electronic security or alarm system, consisting of subsystems interconnected so that a disturbance anywhere in the circuit will result in an alarm signal pinpointing the location of the disturbance.

closed-circuit signaling Signaling accomplished by raising or lowering the level of a signaling current flowing continuously in a circuit.

closed-circuit television Abbreviation, CCTV. A usually in-plant television system, in which a transmitter feeds one or more receivers through a cable.

closed core A magnetic core generally constructed in an "O" or "D" configuration to confine the magnetic path to the core material. Compare OPEN CORE.

closed-core choke A choke coil wound on a CLOSED CORE. Also called CLOSED-CORE INDUCTOR.

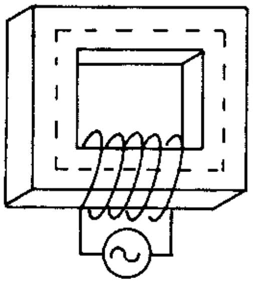

closed-core choke

closed-core transformer A transformer wound on a CLOSED CORE.

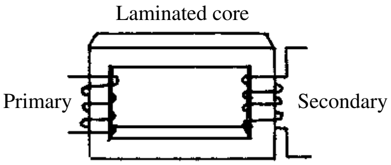

closed-core transformer

closed loop **1.** The feedback path in a self-regulating control system. An oscillator, for example, is a closed-loop amplifier. **2.** A loop within a program that would continue indefinitely, except for an external exit command.

closed-loop bandwidth The frequency at which the gain of a closed-loop circuit (see CLOSED LOOP, **1**) drops 3 decibels from the direct-current or midband value.

closed-loop control system A control system in which self regulation is obtained by means of a feedback path (see CLOSED LOOP). An example is a voltage regulator, in which a rise in output voltage is fed back to the input. This changes the input voltage and reduces the output voltage to its correct value. Compare OPEN-LOOP CONTROL SYSTEM.

closed-loop input impedance The input impedance of an amplifier that has feedback.

closed-loop output impedance The output impedance of an amplifier that has feedback.

closed-loop voltage gain The voltage gain of an amplifier that has feedback.

closed magnetic circuit A magnetic circuit in which the flux is uninterrupted, as in a ferromagnetic core, which has no air gap. Also see CLOSED CORE.

closed subroutine In a digital computer program, a subroutine that can be accessed and left by branch instructions, such as GOSUB and RETURN in the high-level language BASIC.

close-spaced array A beam antenna in which the elements (radiator, director, and reflector) are spaced less than a quarter-wavelength apart.

close-talk microphone A microphone that must be placed close to the mouth. Such a microphone is less susceptible to background noises than an ordinary microphone, and is useful in environments where the ambient noise level is high.

closing rating A specification for closure conditions in a relay, including duty cycle and contact life (total guaranteed closures before contact failure).

closure **1.** The act of closing or being closed (e.g., *switch closure* or *relay closure*). **2.** Circuit completion (i.e., the elimination of all discontinuities).

cloud The mass of electrons constituting the space charge in a vacuum tube.

cloverleaf antenna An omnidirectional transmitting antenna in which numerous horizontal, four-element radiators (stacked vertically, a quarter-wavelength apart) are arranged in the shape of a four-leaf clover.

C/L ratio See LC RATIO.

clutter Extraneous echoes that interfere with the image on a radar display.

clutter gating In radar operations, a switching process that causes the normal video to be displayed in regions free of clutter, and the video indicating target movement to be displayed only in cluttered areas.

Cm Symbol for CURIUM.

cm Abbreviation of CENTIMETER.

c.m. Abbreviation of CIRCULAR MIL.

cm^2 Abbreviation of *square centimeter.*

cm^3 Abbreviation of *cubic centimeter.*

C_{max} Abbreviation of *maximum capacitance.*

C meter See CAPACITANCE METER.

C_{min} Abbreviation of *minimum capacitance.*

CML Abbreviation of CURRENT-MODE LOGIC.

CMOS Abbreviation of COMPLEMENTARY METAL-OXIDE SEMICONDUCTOR.

CMR See COMMON-MODE REJECTION.

CMRR See COMMON-MODE REJECTION RATIO.

CMV See COMMON-MODE VOLTAGE.

C network A circuit with three impedances connected in series, the free leads being connected to a pair of terminals and the two internal junctions, to another pair of terminals.

Co Symbol for COBALT.

C_o Symbol for OUTPUT CAPACITANCE.

coalesce In computer operations, to create one file from several.

coarse adjustment Adjustment of a quantity in large increments. Compare FINE ADJUSTMENT.

coarse-chrominance primary See Q SIGNAL.

coastal bending A change in the horizontal direction of a line-of-sight radio wave when it crosses a coastline.

coast station In the Maritime Mobile Radio Service, a land station that communicates with shipboard stations.

coating **1.** The application of a substance to another substance by means of electroplating, electrophoresis, or similar process, for the purpose of protecting the material, isolating it from the environment, or improving the conductivity of an electrical connection to some other object. **2.** The magnetic material on a recording tape. **3.** In a computer system, the magnetic material on a magnetic diskette or hard disk.

coating thickness On magnetic tape or magnetic disks, the depth of the magnetic coating applied to the base.

coax Abbreviation of COAXIAL CABLE or COAXIAL LINE.

coaxial antenna A half-wave vertical antenna that is center-fed by coaxial cable. The cable runs upward through a 1⁄4-wave section of tubing that composes the lower half of the antenna. The outer conductor of the cable is connected to this tubing through a shorting disk at the top. The inner conductor of the cable is connected to a 1⁄4-wave vertical radiator that is insulated from, and that extends upward from the top of, the lower section.

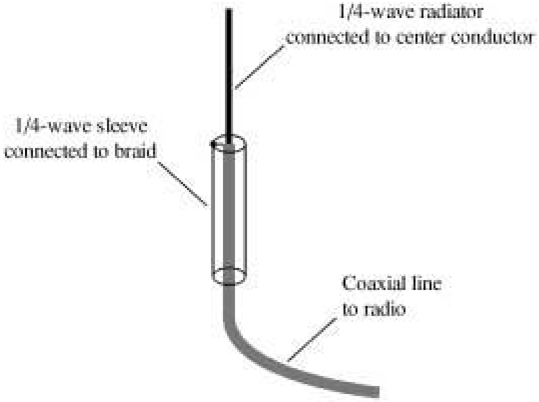

coaxial antenna

coaxial cable An unbalanced cable consisting of two concentric conductors: an inner wire and an outer, braided sleeve. The inner and outer conductors are separated by a dielectric, usually solid or foamed polyethylene. The outer conductor is generally grounded while the inner conductor carries the signals. This cable is used in community-antenna television (CATV) networks, and as a transmission line connecting antennas to radio transmitters, receivers, and transceivers at low, medium, high, and very-high frequencies. It is also used in some high-fidelity sound systems—especially to connect microphones, compact-disc players, tape players, tuners, and speakers to audio amplifiers.

coaxial cable
(From left to right: insulating jacket, woven outer conductor, low-loss insulating sleeve, inner conductor.)

Characteristics of prefabricated coaxial transmission lines.

Type	*Characteristic impedance, (ohms)*	*Velocity factor*	*Outside dia. (in.)*	*Picofarads per foot*
RG-8/U	52	0.66	0.41	29.5
RG-9/U	51	0.66	0.42	30.0
RG-11/U	75	0.66	0.41	20.6
RG-17/U	52	0.66	0.87	29.5
RG-58/U	54	0.66	0.20	28.5
RG-59/U	73	0.66	0.24	21.0
RG-174/U	50	0.66	0.10	30.8
hard line	50	0.81	0.50	25.0
(1/2-inch)	75	0.81	0.50	16.7
hard line	50	0.81	0.75	25.0
(3/4-inch)	75	0.81	0.75	16.7

coaxial capacitor **1.** A somewhat uncommon, but highly effective, capacitor that uses two telescoping sections of tubing. It works because there is a certain effective surface area between the inner and the outer tubing sections. A sleeve of plastic dielectric is placed between the sections of tubing. This allows the capacitance to be adjusted by sliding the inner section in or out of the outer section. Coaxial capacitors are especially useful in antenna systems for tuning and/or impedance matching. Their values are generally from a few picofarads up to about 100 pF. **2.** A short length of coaxial cable that is used as a capacitor rather than a transmission line because of the inherent capacitance between its center conductor and braid. See COAXIAL CABLE.

coaxial cavity A cavity consisting of a cylindrical metal chamber housing a central rod. The cavity can be tuned to resonance by means of a piston.

coaxial connector A device used to splice coaxial line or to connect a coaxial line to a transmitter, receiver, or other piece of apparatus.

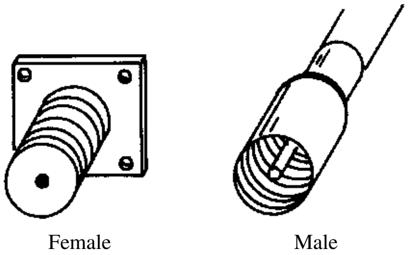

coaxial connector

coaxial diode A semiconductor diode housed in a cylindrical metal shell acting as one contact, and provided with a recessed, concentrically mounted end pin, which serves as the other contact.

coaxial driver See COAXIAL SPEAKER.

coaxial filter **1.** A filter that uses a coaxial cable as a tuned circuit. **2.** A filter designed to be used in a coaxial transmission line.

coaxial jack A female receptacle or connector, whose concentric terminals have the same spacing as a male coaxial-cable connector designed to fit it.

coaxial line A signal transmission line consisting of COAXIAL CABLE.

coaxial-line frequency meter A microwave absorption wavemeter (see WAVEMETER) with input and output receptacles for insertion into a coaxial line.

coaxial-line oscillator See CONCENTRIC-LINE OSCILLATOR.

coaxial loudspeaker See COAXIAL SPEAKER.

coaxial plug A male connector whose concentric terminals have the same spacing as a female coaxial cable connector designed to fit it.

coaxial receptacle A coaxial connector, such as a coaxial jack or plug. Receptacles are installed in equipment, whereas plugs are usually attached to the end of coaxial cables.

coaxial relay A relay designed to connect and disconnect, or to interchange, coaxial cables in a transmission line without disturbing the characteristic impedance of the line.

coaxial speaker Also called *coaxial driver* and *coaxial loudspeaker.* A large low-frequency speaker and a small high-frequency speaker mounted concentrically, the smaller within the larger. When used with a crossover network, this arrangement provides fairly good wide-range audio-frequency response, and saves physical space, compared with the use of separate speakers.

coaxial stub **1.** A length of coaxial cable acting as a branch to another coaxial cable. Commonly used for impedance matching. **2.** A length of coaxial cable, usually cut to 1⁄4 or 1⁄2 wavelength, and connected across a coaxial transmission line to act as a WAVETRAP. Commonly used to reject strong interfering signals.

coaxial switch A switch designed to connect and disconnect, or to interchange, coaxial cables in a transmission line without disturbing the characteristic impedance of the line.

coaxial tank A tank circuit consisting of a rod within a cylinder. The tank is usually tuned by a small variable capacitor connected between the rod and cylinder at one end of the combination. Generally used at ultra-high frequencies (UHF).

coaxial-tank oscillator A stable, self-excited oscillator that uses a COAXIAL TANK. Also see CONCENTRIC-LINE OSCILLATOR.

coaxial transistor A transistor in which a semiconductor wafer is mounted centrally in a metal cylinder (the base connection) and is contacted on opposite faces by the emitter and collector whiskers, which are axially mounted.

coaxial transmission line A transmission line that is a COAXIAL CABLE.

coaxial wavemeter A type of absorption wavemeter in which the tunable element is a section of coaxial line (i.e., a metal cylinder surrounding a metal rod). An internal short-circuiting disk is moved along the cylinder to connect its inner wall to selected points along the rod's length, thereby varying the resonant frequency. The instrument is useful at microwave frequencies.

cobalt Symbol, Co. A metallic element. Atomic number, 27. Atomic weight, 58.94.

cochannel interference Interference between similar signals transmitted on the same channel.

Cockcroft-Walton accelerator A proton accelerator in which nuclei of hydrogen atoms are given high velocity through a straight tube by a high dc voltage.

codan Any of several muting (SQUELCH) systems. In particular, a squelch circuit that suppresses noise in a sensitive receiver equipped with automatic gain control (AGC). The receiver is quiet until a carrier of predetermined strength is received. The name is an acronym for *carrier-operated device antinoise.*

codan lamp A lamp that alerts a radio operator that a signal of satisfactory strength is being received. Also see CODAN.

code **1.** A set of symbols for communications (e.g., the Morse code of radiotelegraphy and wire telegraphy in which dots and dashes correspond to letters, numbers, and marks of punctuation). **2.** In a computer program, symbolically represented instructions. **3.** ENCODE.

codec In encoding and decoding equipment, a *coder/decoder*, usually in a single package and operating at 8 kHz for an input signal with a passband of 3100 Hz (300 to 3400 Hz).

code character **1.** The representation of character in a particular code form. **2.** A sequence of dots and dashes in the Morse code.

code conversion The translation of a coded signal from one form of code to another.

coded decimal digit A number expressed in binary form (computer code), that is, in terms of zeros and ones only.

code-directing characters Characters added to a message to indicate how and where it is going.

coded program See PROGRAM.

coded signal **1.** A wire- or radiotelegraph signal in which secrecy is achieved by using letters in cipher groups, instead of straight language. **2.** SCRAMBLED SIGNAL.

coded stop See PROGRAMMED HALT.

code elements The smallest identifiable parts that compose a digital code. For example, in computer code, the elements are ones and zeroes (*high* and *low* logic states); in Morse code, they are *dots* and *dashes*.

code holes In a punched card or tape, holes representing data.

code line A written computer program instruction.

code machine Any one of several devices for recording or reproducing code signals.

code position The part of a data medium (e.g., card row) reserved for data.

code-practice oscillator A simple keyed audio oscillator intended for practicing Morse code.

coder **1.** In computer operations, a person who prepares instructions from flow charts and procedures devised by a programmer. **2.** A device that delivers coded signals.

code receiver A radiotelegraph receiver.

code ringing A method of ringing a telephone subscriber in a predetermined manner to convey a certain message.

code segment The instruction part of computer storage associated with a process. Compare DATA SEGMENT and DUMP SEGMENT.

code set The collection of codes representing all of the characters in a language.

code speed See KEYING SPEED.

code transmitter **1.** A radiotelegraph transmitter. **2.** A tape-operated keyer for wire telegraphy or radiotelegraphy.

code word See PHONETIC ALPHABET CODE WORD.

coding **1.** Performing the service of a CODER. **2.** Writing instructions for a digital computer; a part of programming.

coding check A pencil-and-paper verification of a routine's validity.

coding sheet A form on which program instructions are written prior to input.

codiphase radar A radar system that uses beam forming, signal processing, and a phased-array antenna.

codistor A voltage-regulating semiconductor device.

coefficient **1.** A factor in an indicated product. Thus, in $4y$, 4 is the coefficient of y. **2.** A parameter that indicates a specific characteristic of some component or device (e.g., COEFFICIENT OF COUPLING or COEFFICIENT OF REFLECTION).

coefficient of coupling Symbol k. The ratio of MUTUAL INDUCTANCE between two inductors to the maximum possible (theoretical) value of mutual inductance. This ratio is always greater than or equal to 0 (no coupling between inductors), and less than or equal to 1 (perfect coupling between inductors).

coefficient of current detection See CURRENT-DETECTION COEFFICIENT.

coefficient of reflection A measure of the amount of electromagnetic field reflected in a transmission line from the load feed point. The coefficient of reflection is equal to the square root of the reflected power divided by the forward power.

coercive force The demagnetizing force required to remove residual magnetism from a material.

coercivity See COERCIVE FORCE.

cogging Nonuniform rotation of a motor armature. The velocity increases as an armature coil enters the magnetic field and decreases as it leaves the field.

coherence In electromagnetic radiation, a condition in which all the wavefronts are in phase. This results in high energy concentration, and makes possible the long-distance transmission of infrared, visible light, and ultraviolet, because the rays are almost perfectly parallel. It also makes possible the extreme radiation intensity characteristic of some LASER devices.

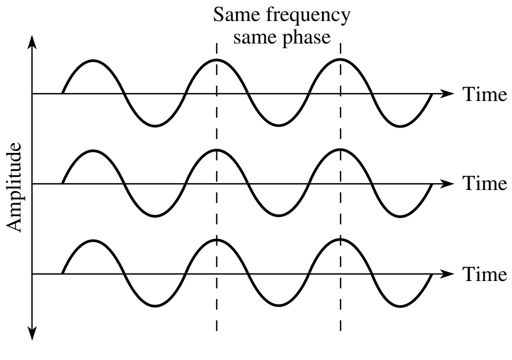

coherence

coherent bundle A bundle of optical fibers, such that the individual fibers are in the same relative positions at either end of the bundle.

coherent carrier A carrier that agrees in frequency and phase with a reference signal.

coherent electroluminescent device See LASER DIODE.

coherent light Visible light in which the phase relationship between successive waves is such that the beam consists of parallel rays that provide a high concentration of energy. Also see LASER.

coherent-light radar See COLIDAR.

coherent oscillator In a radar system, an oscillator that provides a COHERENT REFERENCE.

coherent-pulse operation Pulse operation characterized by a fixed phase relationship between pulses.

coherent radiation Radiation characterized by COHERENCE.

coherent reference A stable reference frequency with which other signals are phase locked for coherence.

coherent transponder A transponder in which the frequency and phase of the input and output signals have a fixed relationship.

coil A long conductor or group of conductors wound into a tight helical package, often in several layers on a cylindrical form. This takes advantage of the resulting concentration of magnetic flux, maximizing the inductance that can be obtained in a component of limited physical size. Further increases in inductance can be realized by the use of ferromagnetic core materials. See also INDUCTOR.

coil antenna See LOOP ANTENNA.

coil checker An alternating-current (ac) meter or simple bridge for checking inductors. Such instruments usually only indicate inductance values, but some list readings of resistance or approximate inductor Q factor.

coil dissipation The power wasted in a coil as heat. Generally, this dissipation or loss is proportional to the resistance of the coil, and to the square of the current passing through the coil.

coil form The insulating support around which an air-core coil is wound.

coil loading The insertion of one or more inductors into a transmission line or antenna element, for the purpose of impedance matching, alteration of the resonant frequency, or both.

coil magnification factor The Q factor of an inductor. Generally given by the ratio X_L/R_L, where X_L is the inductive reactance of the coil in ohms, and R_L is the resistance of the coil in ohms.

coil neutralization See INDUCTIVE NEUTRALIZATION.

coil resistance The resistance of a coil (inductor), as distinct from its reactance. It is almost entirely the result of ohmic loss in the wire from which the coil is manufactured.

coilshield A metal can designed to provide efficient electrostatic and electromagnetic shielding of a coil, preventing unwanted inductive coupling to other components.

coincidence The simultaneous occurrence of two or more signals. Compare ANTICOINCIDENCE.

coincidence amplifier An amplifier that delivers an output signal only when two or more input signals occur simultaneously.

coincidence circuit See AND CIRCUIT.

coincidence counter A circuit or device, such as a gate, that delivers an output pulse only when two or more input pulses occur simultaneously; the output pulses go to a device that counts them.

coincidence detector See AND CIRCUIT.

coincidence gate See AND GATE.

coincident-current selection Selection of a magnetic core (in a core memory or similar device) by applying two or more currents simultaneously.

coin shooting Searching for coins and similar small, buried metallic objects using a METAL LOCATOR.

coke A porous material obtained from the destructive distillation of coal. It is valued for the production of carbon components for electronics, such as dry-cell electrodes and motor brushes.

cold **1.** Pertaining to an electrical circuit, component, or terminal that is at ground potential. **2.** A term denoting a bad solder joint. **3.** Pertaining to an unheated electrode or element. See COLD CATHODE.

cold alignment The alignment of a tracking system (especially of its tuned circuits) when the system is not in operation, as when transistor power is off. Also called QUIET ALIGNMENT.

cold cathode **1.** In an electron tube, a cathode that emits electrons without being heated. **2.** A cathode electrode operated at a temperature below ambient temperature.

cold chamber An enclosure in which electronic equipment can be tested at selected, precise low temperatures. Compare OVEN.

cold flow The (usually gradual) change in the dimensions of a material, such as plastic in a molded part.

cold junction In a thermocouple system, an auxiliary thermocouple connected in series with the hot thermocouple, and immersed in ice or operated at ambient temperature.

cold light Light produced without significant heat, as from the ionization of a gas by a high voltage (as in neon bulbs and fluorescent lamps), or by electroluminescence, bioluminescence, cathodoluminescence, or a similar phenomenon.

cold pressure welding Welding sometimes used in the fabrication of electronic equipment, in which the metal parts to be joined are pressed together tightly to the point of deformation, whereupon they become welded.

cold resistance The resistance of an unheated electronic component. Compare HOT RESISTANCE.

cold rolling A method of manufacturing an inductor core so that the magnetic grains are all arranged lengthwise.

cold solder joint A solder joint in which insufficient heat has been applied, resulting in a bad connection.

cold spot **1.** An area of a circuit or component whose temperature is ordinarily lower than that of the surrounding area. **2.** A node of current or voltage. Compare HOT SPOT.

cold weld A welded joint produced by means of COLD PRESSURE WELDING.

colidar An optical radar system using unmodulated, coherent (laser-produced) light. The term is an acronym for *coherent light detection and ranging.*

collate In data processing, to produce an ordered set from two or more similarly ordered sets (as punched cards).

collator In a punched-card system, a device that collates (see COLLATE) punched cards.

collector **1.** In a bipolar transistor, the electrode toward which emitted current carriers travel. **2.** In a Klystron, the final electrode toward which electrons migrate after passing through the buncher and catcher. **3.** In an iconoscope, a cylindrical electrode around the circumference of the tube, which gathers and conducts away the electrons released by the mosaic. **4.** The final (target) electrode in a backward-wave or traveling-wave tube. **5.** A computer program segment that collates compiled segments so that they can be loaded into the computer.

collector capacitance **1.** Symbol, C_C. The capacitance of the collector junction in a bipolar transistor. **2.** The capacitance of the collector electrode in a Klystron, iconoscope, backward-wave tube, or traveling-wave tube.

collector current **1.** Symbol, I_C. The current flowing in the collector circuit of a bipolar transistor. Also see AC COLLECTOR CURRENT and DC COLLECTOR CURRENT. **2.** Current flowing in the collector circuit of a Klystron, iconoscope, backward-wave tube, or traveling-wave tube.

collector-current cutoff See COLLECTOR CUTOFF.

collector cutoff In a bipolar transistor, the condition in which the collector current is cut off (i.e., reduced to the residual value). Also see CUTOFF CURRENT.

collector cutoff current See CUTOFF CURRENT.

collector-diffusion isolation A method of making integrated circuits that contain bipolar transistors. Provides electrical separation of the transistors in a semiconductor integrated circuit.

collector dissipation Symbol, P_C. In a bipolar transistor, the power dissipation of the collector electrode. The collector dc power dissipation is the product of collector current and collector voltage: $P_C = V_C I_C$.

collector efficiency In a bipolar transistor circuit, the ratio of ac power output to dc collector-power input.

collector family For a bipolar transistor, a group of collector current versus collector voltage curves. Each is plotted for a particular value of base current (common-emitter circuit) or emitter current (common-base circuit).

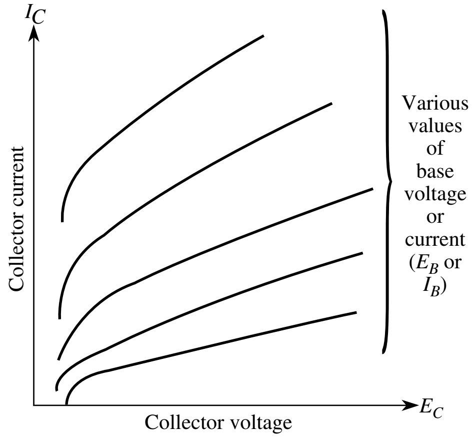

collector family

collector junction In a bipolar transistor, the junction between collector and base layers.

collector mesh In a cathode-ray storage tube, a flat, fine wire screen that attracts and conducts away the secondary electrons knocked out of the storage mesh by the electron beam.

collector multiplication In a bipolar transistor, an increase in the number of electrons at the collector electrode, caused by a momentary alteration of the charge density of the collector junction by injected carriers reaching the junction.

collector resistance In a bipolar transistor, the internal resistance of the collector junction. See AC COLLECTOR RESISTANCE and DC COLLECTOR RESISTANCE.

collector ring **1.** A rotating, brush-contacted ring electrode connected to one end of a coil in an ac generator. **2.** A similar ring which, with a brush, serves as a connection to a rotating element, as in a signal-gathering system. **3.** The collector electrode in an iconoscope.

collector transition capacitance The capacitance between the collector and base of a bipolar transistor under normal operating conditions. This capacitance has a limiting effect on the operating frequency of a bipolar device.

collector voltage Symbol, V_C. In a bipolar transistor, the voltage on the collector electrode. See AC

COLLECTOR VOLTAGE and DC COLLECTOR VOLTAGE.

collimated rays Electromagnetic waves made parallel or nearly parallel. This can be done by means of a reflector, a lens, or a laser.

collimation **1.** The process of rendering electromagnetic rays parallel. **2.** Adjustment of the line of sight of an instrument, such as a level or transit.

collimation equipment Optical-alignment equipment.

collimator A device for producing parallel rays of light or other radiation. In electronics, the most common example is a *dish antenna.*

collinear antenna A broadside directional antenna consisting of two or more half-wave radiators; the current is kept in phase in each section by quarter-wave stubs between each radiating section. The radiators are stacked end to end horizontally or vertically. Also called FRANKLIN ANTENNA.

Collins coupler A single-section, pi-filter circuit, used to match a radio transmitter to a wide range of antenna impedances. Also called *pi coupler* and *Collins network.*

collodion A viscous solution of pyroxylin and a solvent (such as acetone, alcohol, or ether) sometimes used as a binding agent for coils and other components.

cologarithm Abbreviation, colog. The logarithm of the reciprocal of a number; colog $x = \log(1/x) = \log x^{-1} = -\log x$.

color A perceived characteristic, and a direct function, of visible-light wavelength. Seen by the human eye as a spectrum of hues, ranging from red at the longest visible wavelengths, through orange, yellow, green, blue, indigo, and finally violet at the shortest visible wavelengths. See HUE.

coloration In audio applications, a blending of sounds as a result of mixing among components at different frequencies. Sometimes this is done deliberately; in other instances, it is undesirable.

color balance In a color television receiver, adjustment of the beam intensities of the individual guns of a three-gun picture tube. Compensates for the difference in light emissivity of the red, green, and blue phosphors on the tube screen.

color bar-dot generator A radio-frequency (RF) signal generator that produces a bar or dot pattern on the screen of a color television picture tube. Used for testing and alignment.

color-bar pattern A color television test pattern of vertical bars—each of a different color.

color breakup A transient separation of a color television picture into its red, green, and blue components, as a result of a sudden disturbance of viewing conditions (blinking of eyes, moving of head, intermittent blocking of screen, etc.).

color burst As a phase reference for the 3.579545-MHz oscillator in a color television receiver, approximately nine cycles of the chrominance subcarrier added to the back porch of the horizontal blanking pedestal in the composite color signal.

color carrier See CHROMINANCE SUBCARRIER.

colorcast A color television broadcast.

color code **1.** A system that uses colored stripes or dots to mark the nominal values and other characteristics on capacitors, resistors, and other components. **2.** A code that represents the various frequencies being used by radio-control modelers in competition, and used on flags attached to transmitters, for example, as a safeguard against jamming.

color coder See COLOR ENCODER.

color contamination In a color television system, faulty color reproduction resulting from incomplete separation of the red, green, and blue channels.

color-coordinate transformation In a color television system, the computation (performed electrically in the system) of the tristimulus (primary) values with reference to one set of primaries, from the same colors derived from another set of primaries.

color depth An expression for the extent to which an image can accurately render color. Generally expressed in bits or in number of colors. Some systems can reproduce millions of different colors.

color-difference signal Designated B-Y, G-Y, and R-Y. The signal resulting from reducing the amplitude of a color signal by an amount equal to the luminance-signal amplitude. Also see B-Y SIGNAL, G-Y SIGNAL, and R-Y SIGNAL.

color dot **1.** A phosphor spot on the screen of a color television picture tube. **2.** One of the spots stamped on a capacitor, indicating the capacitance, voltage, and tolerance (see COLOR CODE, **1**). **3.** A spot stamped on a resistor, indicating the number of zeros to be added to the value indicated by the color bands.

color edging In a color television picture, an aberration consisting of false color at the boundaries between areas of different color.

color encoder In a color television transmitter, the circuit or channel in which the camera signals and the chrominance subcarrier are combined into the color-picture signal.

color equation A mathematical means of determining the resultant color obtained by adding primary colors in various proportions.

color fidelity The faithfulness with which a color television system, lens, or film reproduces the colors of a scene.

color filter A transparent plate or film that transmits light of a desired color, and eliminates or attenuates all other colors.

color flicker In a color television system, image instability that occurs when the luminance and chromaticity both fluctuate.

color fringing In a color television picture, false color around objects, sometimes causing them to appear separated into different colors.

color generator A special radio-frequency (RF) signal generator to adjust or troubleshoot a color television receiver. The color signals it delivers are identical to those produced by a broadcast station.

color graphics Computer graphics displayed in color on a cathode-ray tube (CRT) or liquid-crystal display (LCD).

colorimeter A device used to quantitatively measure the color intensity of a sample relative to a standard.

colorimetric A characteristic of visible light, representing the wavelength concentration. Refers to the perceived color of a light beam.

colorimetry The science and art of color measurement.

color killer In a color television receiver, a circuit that, in the absence of a color signal, delivers a negative bias to cut off the bandpass amplifier.

color match In photometry, the condition in which color agreement exists between the halves of an area. Also see COLOR MATCHING.

color matching The art of selecting colors that are identical in hue, saturation, and brilliance. This can be done with the unaided eye or with the help of an instrument.

color media Substances that transmit essentially one color of visible light while blocking other colors.

color meter A photoelectric instrument for measuring color values, and comparing and matching colors.

color mixture An additive combination of two or more colors. Thus, red + yellow = orange, blue + red = violet, red + blue + green = white, etc.

color oscillator The oscillator in a color television receiver that coordinates the color response. This oscillator is operated at 3.579545 MHz, to within plus or minus 10 Hz.

color palette In a color video image, the total number of possible colors that can be displayed.

color phase In color television, the phase difference between an I or Q chrominance primary signal and the chrominance carrier reference.

color-phase diagram In color television, a quadrant diagram showing (for each of the three primary and complementary colors) the difference in phase between the color-burst signal and the chrominance signal, as well as the peak amplitude of the chrominance signal. Also shown are the peak amplitude and polarity of both in-phase and quadrature components required for the chrominance signals. For color TV receiver adjustment, the color-phase diagram is displayed, in effect, by a VECTORSCOPE when a suitable signal from a color generator is applied to the receiver.

color picture signal **1.** In color television and/or computer graphics, an electrical signal containing components corresponding to the hue, saturation, and brilliance of a fixed or changing visual image. **2.** In color television, the combination of chrominance and luminance signals minus blanking and sync signals.

color picture tube A specialized type of cathode-ray tube (CRT), used in color television receivers and computer displays. Three different images are produced: one in red, one in blue, and one in green. The three monochrome images are combined to form a complete color image.

color primaries **1.** Also called *additive primaries* or *primary colors.* In color television, the hues red (R), green (G), and blue (B). When these colors are mixed in various ratios, any visual color can result. **2.** Also called *subtractive primaries* or *primary pigments.* In color printing, the hues magenta (M), cyan (C), and yellow (Y). These roughly correspond to red (R), blue (B) and yellow (Y). Sometimes black (K) is also included. When these pigments are mixed in various ratios, any visual pigment can result.

color purity The ratio of wanted to unwanted components in a color. In a pure color, there are no components other than those required to produce the color. Color, in this context, includes white, black, and all shades of gray.

color-purity magnet A permanent magnet on the neck of a color television picture tube, used to help ensure color purity by maintaining proper displacement of the electron beam.

color registration In color television reception, the precise superimposition of red, green, and blue so that the composite is free from COLOR EDGING.

color rendering index A mathematical expression defining the effect of the color of a light source on an object. For example, in red light, a blue object appears nearly black.

color sampling rate The number of times per second that each primary color is sampled in a color television receiver.

color saturation A measure of the purity of a hue. The extent to which a hue is without a white component; 100% saturation indicates a complete absence of white.

color sensing In machine vision systems, the ability to distinguish between light of different wavelengths. Usually done with red, green and blue color filters and three separate cameras.

color sensitivity **1.** The degree of which a photosensitive device, such as a photocell or camera tube, responds to various colors of light. **2.** The degree to which photographic film responds to various colors of light.

color signal See COLOR PICTURE SIGNAL.

color spectrum The band of electromagnetic energy containing visible light; it extends from red (at the longest wavelengths) to violet (at the shortest). Commonly measured in *nanometers (nm),* where 1 nm = 10^{-9} m. Also expressed in *Angstroms*, where 1 Angstrom = 10^{-10} m = 0.1 nm. In order of decreasing wavelength, the colors are red at 750 to 700 nm (7500 to 7000

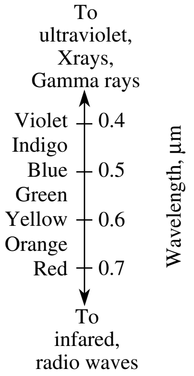

color spectrum

Angstroms), orange, yellow, green, blue, indigo, and violet at 410 to 390 nm (4100 to 3900 Angstroms).

color subcarrier A modulated monochrome signal whose sidebands convey color information.

color-sync signal See COLOR BURST.

color system Also called *RGB color model.* A means of representing a color in terms of mathematical coordinates. This can be done in three dimensions because there are three COLOR PRIMARIES. Each color primary is represented by an axis. Any COMPOSITE COLOR can be represented by a unique vector. The relative amount of each color primary is given by the length of the composite-color vector components along each axis.

color television Television in which the picture approximates natural color. It operates on the basis of mixing three primary colors (red, blue, and green) of phosphors on the picture tube screen.

color television receiver A television receiver designed to reproduce color pictures.

color television signal The signal transmitted by a color television transmitter, containing all of the information needed to reproduce a complete, full-color, moving image.

color transmission The television transmission of a picture in color.

color triad On the screen of a color picture tube, one of the color cells, each of which contains one of the three phosphor dots: red, green, and blue.

color triangle A triangle that can be inscribed on a chromaticity diagram to reveal the chromaticity range resulting from adding the three color primaries.

color TV signal The complete signal (video, color, and sync components) required for transmitting a picture in color.

color weather radar A computer-enhanced radar rendition of weather patterns, usually showing various intensities of precipitation as different colors. Commonly, areas of precipitation show up as violet, blue, green, yellow, orange, and red, in order of increasing intensity.

Colpitts oscillator A radio-frequency (RF) oscillator that uses a single, untapped inductor. A combination of two fixed capacitors in series is connected in parallel with the inductor. The feedback is controlled by the ratio of capacitances. A permeability-tuned coil or a roller inductor can be used to obtain variable-frequency operation. Stability is enhanced when the output of the oscillator is taken from the emitter or source portion of the circuit. To prevent the output signal from being short-circuited to ground, an RF choke is connected in series with the emitter or source. Compare HARTLEY OSCILLATOR.

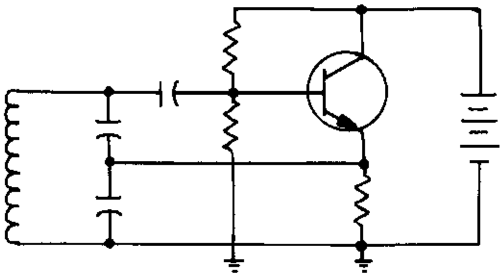

Colpitts oscillator

columbium Symbol, Cb. The former name of the metallic element *niobium.* Atomic number, 41. Atomic weight, 92.9064.

column See CARD COLUMN.

columnar graph A graphical presentation of data, in which the ordinates are represented by vertical columns whose height depends on the value. Commonly used in *presentation graphics*, but less common in *analytical graphics.*

column binary Binary number representation on punched cards, wherein consecutive digits correspond to consecutive column punching positions.

column speaker An acoustic speaker with a long cabinet, so that a large column of air is used for resonating or reinforcing purposes. This type of speaker radiates over a wide azimuth angle, while providing a narrow beam in the elevation plane.

column split On a punched card machine, the device for reading, as two separate characters or codes, two parts of a single column.

COM **1.** Abbreviation for *communications port.* **2.** Abbreviation for *computer output on microfilm.*

coma An aberration that causes the beam spot on the screen of a cathode-ray tube to resemble a comet.

coma lobes An aberration in the radiation or response pattern of a dish antenna that occurs when the radiating element is not exactly at the focal point of the reflector. When the directional pattern is altered by moving the driven element, rather than turning the entire antenna, these lobes appear.

comb amplifier An arrangement of several sharply tuned bandpass amplifiers whose inputs are connected in parallel and whose outputs are separate; the amplifiers separate various frequencies from a multifrequency input signal. The name is derived from the comb-like appearance of the response pattern of various output peaks displayed along a frequency-base axis.

comb filter A selective device that passes several narrow bands of frequencies within a larger band, while rejecting frequencies in between the narrow bands. So called because its frequency-response curve resembles the teeth of a comb when observed on a spectrum analyzer. Also see COMB AMPLIFIER.

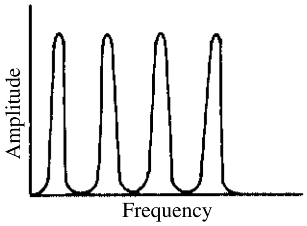

comb filter response

comb generator **1.** A signal generator that provides outputs at evenly spaced frequencies. So called because, on a spectrum analyzer, its output looks like the teeth of a comb. **2.** A transmitter with many spurious signals at its output.

combination **1.** A functional, usually stationary, installation consisting of two or more pieces of equipment. Examples: transmitter/receiver combination, motor/generator combination, and tuner/amplifier combination. **2.** In mathematics, a selection of several factors from a group, without regard to order. Thus, from the group ABC, the three possible combinations are AB, AC, and BC. Compare PERMUTATION.

combinational circuit Two or more basic logic circuits, combined in such a way that the output state depends entirely on the input states.

combination bridge A bridge that affords two or more classes of measurement, usually selectable by means of a function switch. Examples: capacitance-inductance bridge, and capacitance-resistance bridge.

combination cable A cable that has conductors grouped in pairs, threes, quads, or similar arrangements.

combination feedback See CURRENT-VOLTAGE FEEDBACK.

combination microphone Two or more microphones combined into one unit.

combination speaker Two or more loudspeakers combined into one (e.g., a COAXIAL SPEAKER).

combination tone An acoustic tone resulting from the combination of two other acoustic tones. If the original tones have frequencies f_1 and f_2 (where f_1 is higher than f_2), then the first-order combination frequencies are $f_1 + f_2$ and $f_1 - f_2$. Higher-order combination tones can result from mixing among the original tones and the first-order combination tones.

combinatorial logic A form of digital logic, in which the output states depend on the input states, but on no other factor.

combined head See READ-WRITE HEAD.

combined reactance The net reactance (X) in a circuit, obtained by vectorially adding the inductive reactance (X_L) and the capacitive reactance (X_C).

combiner A circuit or device for mixing various signals to form a new signal. Also see MIXER.

combiner circuit In a color television camera, the circuit that combines the chroma and luminance with the sync.

comeback A spurious response in a bandpass or band-rejection filter, at a frequency well above or below the passband or stopband.

command **1.** In computer operations, the group of selected pulses or other signals that cause the computer to execute a step in its program. **2.** Instruction.

command chain Part of a computer operation carried out independently as a series of input/output instructions.

command control In automation, electronic control, and computer operations, the performance of functions in response to a transmitted signal.

command destruct signal A signal for instigating the destruction of a missile in flight.

command guidance system A system in which a guided missile and its target are both tracked by radar.

command language A computer language made up of command operators.

command link In a command guidance system, the section that transmits missile-steering commands.

command network A radio communications network in which the chain of command is rigorously defined and followed.

command reference The current or voltage to which a feedback signal is referenced in a control system or servomechanism.

comment A statement written into a computer program for a documentation, rather than implementation (e.g., to describe the purpose of a step or subroutine).

comment field A record or file in which instructions or explanations are given.

commercial data processing A commercial (rather than industrial, scientific, or personal) application of data processing.

commercial-level security See LEVEL-2 SECURITY.

commercial killer A usually remote-controlled, electronic relay for disabling a radio or television receiver during advertisements.

commercial language A computer programming language for commercial applications (payroll, for example).

common **1.** Grounded. **2.** Pertaining to a connection shared by several different points in a circuit or system. **3.** See COMMON GROUND.

common area A computer storage area usable by several programs or segments within a program.

common-base circuit A bipolar transistor circuit in which the transistor base is the common (or grounded) electrode. Also called *grounded-base circuit.*

common battery **1.** A battery shared by two or more different circuits or pieces of equipment. **2.** In wire telephony, a central office battery that supplies the entire system.

common-battery office In wire telephony, a central office that provides a common battery.

common business-oriented language See COBOL.

common-capacitor coupling The process of coupling one tuned circuit to another by means of a capacitor that is common to both circuits.

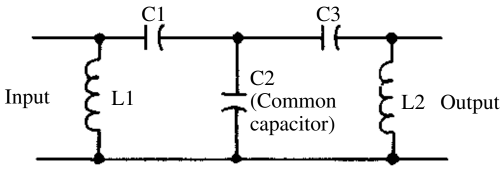

common-capacitor coupling

common-carrier fixed station A fixed radio station that provides public service.

common-cathode circuit A tube circuit in which the cathode is the common (or grounded) electrode. Also called *grounded-cathode circuit.*

common-channel interference Radio interference resulting from two stations transmitting on the same channel. It is characterized principally by beat-note (heterodyne whistle) generation, and suppression or capture of the weaker signal by the stronger one.

common-collector circuit A bipolar-transistor circuit in which the collector is the common (or grounded) electrode. Also called *grounded-collector circuit* and EMITTER FOLLOWER.

common communications carrier A communications company authorized by the licensing agency to furnish public communications.

common-component coupling See COMMON-CAPACITOR COUPLING, COMMON-INDUCTOR COUPLING, and COMMON-RESISTOR COUPLING.

common-drain circuit A field-effect transistor circuit in which the drain terminal is the common (or grounded) electrode. Also called *grounded-drain circuit* and SOURCE FOLLOWER.

common-emitter circuit A bipolar transistor circuit in which the emitter is the common (or grounded) electrode. Also called *grounded-emitter circuit.*

common-gate circuit A field-effect transistor circuit in which the gate is the common (or grounded) electrode. Also called *grounded-gate circuit.*

common-grid circuit A tube circuit in which the control grid is the common (or grounded) electrode. Also called *grounded-grid circuit.*

common ground A single ground-point connection shared by several portions of a circuit.

common impedance A single impedance shared by parts of a circuit. Because currents from the various parts flow through this impedance simultaneously, coupling (desired or undesired) can occur between them.

common-impedance coupling See COMMON-CAPACITOR COUPLING, COMMON-INDUCTOR COUPLING, and COMMON-RESISTOR COUPLING.

common-inductor coupling The process of coupling one tuned circuit to another by means of an inductor that is common to both circuits.

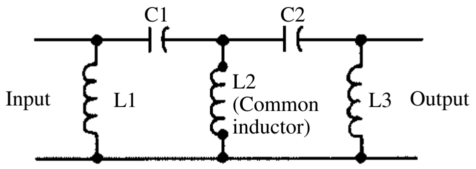

common-inductor coupling

common language A language recognized by all the equipment in a data processing system.

common logarithm Abbreviation, log$_{10}$. Also called *base-10 logarithm.* A logarithm in which the base number is 10. Also see LOGARITHM.

common mode Pertaining to signals or signal components that are identical in amplitude and duration.

common-mode characteristics In an operational amplifier, characteristics denoting amplifier performance when a common signal is applied to inverting and noninverting inputs.

common-mode gain The voltage gain of a differential amplifier with a common-mode input.

common-mode impedance input The impedance between ground and one of the inputs of a differential amplifier. Compare COMMON-MODE INPUT IMPEDANCE.

common-mode input capacitance In a differential amplifier, the internal capacitance of the common-mode input circuit.

common-mode input circuit In a differential amplifier, the input circuit between ground and the inputs connected together.

common-mode input impedance In a differential amplifier, the open-loop impedance between ground and the inputs connected together. Compare COMMON-MODE IMPEDANCE INPUT.

common-mode input signal A signal applied to the common-mode input circuit of a differential amplifier (i.e., to both inputs connected together). Compare COMMON-MODE SIGNAL.

common-mode input voltage In a differential amplifier, the maximum voltage that can be applied safely between ground and the inputs connected together.

common-mode interference A form of interference that occurs across the terminals of a grounded system.

common-mode rejection The extent to which a differential amplifier will reject a signal presented simultaneously to both inputs in phase, or of two signals identical in amplitude, frequency, and phase applied separately to the two inputs. Also see COMMON-MODE REJECTION RATIO.

common-mode rejection ratio In a differential amplifier, the extent to which the amplifier cancels undesired signals. It is the ratio of the differential gain to the common-mode gain. Also see COMMON-MODE REJECTION.

common-mode signal The algebraic average of two signals applied simultaneously to the two ends of a balanced circuit, such as a differential amplifier. Compare COMMON-MODE INPUT SIGNAL.

common-mode voltage The part of the input that is common to both inputs of a differential amplifier circuit. It is quantitatively defined as the arithmetic mean of the voltages at the inputs.

common-mode voltage gain See COMMON-MODE GAIN.

common-mode voltage range The range limited by the maximum nonsaturating input voltage that can be applied to both inputs of an operational amplifier.

common pool An assigned memory store, utilized by two or more circuits or systems.

common-resistor coupling The process of coupling one circuit to another by means of a resistor that is common to both circuits.

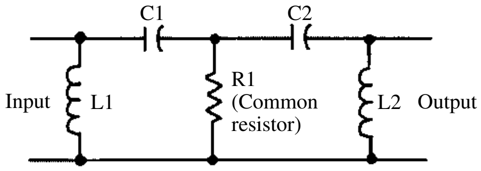

common-resistor coupling

common-source circuit A field-effect transistor circuit in which the source terminal is the common (or grounded) electrode. Also called *grounded-source circuit.*

common-user channels Communication channels open to all licensees in a particular service.

communication band A band of frequencies whose use is authorized expressly for communications, rather than for other services (such as broadcasting, education, remote control, etc.).

communication channel **1.** In radio or wire service, a (usually auxiliary) channel for direct exchange of information between units of the service (e.g., a "talking circuit" between a broadcast studio and the transmitter house). **2.** A data transmission channel between two points (e.g., a remote terminal and a central computer system).

communication link **1.** Collectively, the equipment providing a communication channel between two transmitters. **2.** Data terminal equipment.

communication protocol The specifications of a digital signal, including the speed in bits per second (bps) or bauds, the code type, the bit duration, the mark-to-space ratio, etc.

communications The science and art of using and developing electronic equipment and processes for the transmission and reception of information.

communications common carrier An organization licensed to provide public communication services.

communications network An organization of transmitting and receiving stations for the reliable exchange of intelligence. Also called *net.*

communications receiver A general-coverage or multiband radio receiver, designed primarily for listening to amateur, weather, or other nonbroadcast stations. Compare BROADCAST RECEIVER.

communications satellites Satellites in earth orbit that provide propagation paths (e.g., by reflection or retransmission) for radio waves between terrestrial transmitters and receivers. Also see ACTIVE COMMUNICATIONS SATELLITE and PASSIVE COMMUNICATIONS SATELLITE.

community-antenna television Abbreviation, CATV. A system in which an advantageously located receiving station receives television signals, amplifies them if necessary, and distributes them in the community served by the system. Commonly called *cable TV.*

commutating capacitor **1.** In a flip-flop circuit, a capacitor connected in parallel with the cross-coupling resistor to accelerate the transition from one stable state to the other. Also called *speedup capacitor.* **2.** A capacitor connected in parallel between silicon-controlled rectifier (SCR) stages to momentarily reverse the current going through the SCR, thereby causing the SCR to go into the cutoff condition.

commutation **1.** In a direct-current (dc) generator, periodic reversal of the current in the armature coils as the coils alternately pass the north and south poles of the magnetic field. When the ends of each coil are connected to opposite bars of the commutator, the electrical polarity at the commutator brushes remains constant. **2.** In a thyratron or silicon-controlled rectifier (SCR) circuit, momentarily reversing the polarity to cut the device off.

commutator **1.** In a direct-current (dc) motor, generator, or rotating selector, an arrangement of parallel metal bars or strips on a rotating drum. As the drum turns, the bars contact one or more brushes that are in sliding contact with the commutator. **2.** An electronic circuit that switches a single input sequentially to a series of output terminals, or switches a number of inputs sequentially to a single pair of output terminals.

commutator ripple The pulsating voltage superimposed on the direct-current (dc) voltage delivered by an unfiltered dc generator.

compact disc Abbreviation, CD. A digital, high-density optical disc, used in high-fidelity stereo sound systems. Also used to store computer data. The information is encoded as tiny pits on the surface of the disc, and is recovered by a laser, a sensor and a digital-to-analog (D/A) converter. These disks have largely superseded magnetic tapes, and have rendered long-playing vinyl disks and turntables obsolete. See also COMPACT-DISK READ-ONLY MEMORY.

compact-disk read-only memory Abbreviation, CD-ROM. A digital COMPACT DISC used for the long-term storage of computer data and/or software programs. Usually the same size as a high-fidelity stereo disk, it can hold over 600 megabytes of data. Although data can be read from the disk, it cannot be overwritten.

compander Term for *compressor/expander.* In the transmission and reception of audio-frequency (AF) intelligence, a system that uses an amplitude compressor at the transmitter and an amplitude expander at the receiver. The compressor reduces the dynamic range before transmission, and the expander restores it after reception. Provides improved signal-to-noise ratio under marginal communications conditions. Also increases the ratio of average power to peak power. See COMPANDING.

companding A process in which a signal is compressed at the transmitting end of a circuit and expanded at the receiving end, yielding a signal like the original at the receiver output. Signals are more efficiently transmitted when they are compressed because the average power increases, relative to the peak power. This improves the average signal-to-noise ratio for weak signals. See COMPANDER.

companding law The mathematical function used for companding. It is an output-amplitude versus input-amplitude function for the compression at the transmitter, and the inverse of this function for the expansion at the receiver.

companion keyboard An auxiliary keyboard connected to a regular keyboard and operated remotely.

companionship machine A computer or robot with sufficient machine intelligence to provide entertainment and mental stimulation for humans.

comparator **1.** An integrated circuit (IC) with two inputs, called A and B. The device compares the voltages that appear at these inputs. If the input voltage at A is significantly greater than the input voltage at B, the output is about +5 V. If the input voltage at A is not greater than the input voltage at B, the output voltage is about +2 V. These ICs are used to actuate, or trigger, other devices such as relays and electronic switching circuits. **2.** In general, any circuit that compares some characteristic of two input signals and produces an output that depends on the relationship between the inputs. **3.** An instrument for checking the condition of a component by comparing it directly with an identical component of known quality has a scale reading in percentage deviation, or simply "GO/NO-GO." Examples: capacitor comparator, resistor comparator, coil comparator.

compare In computer operations, a relational test performed on two quantities to determine their relative magnitude, including an indication of the test result and, sometimes, the taking of action. Example: the process and acton resulting from execution of the statement "IF A > B THEN GO TO LINE 250."

comparison **1.** An expression of the relationship between two voltages, currents, phase angles, component values, or other quantities in an electrical or electronic circuit or system. **2.** An examination of different data bits or items, which results in a conclusion about some aspect of their relationship.

comparison bridge A bridge designed especially for the quick comparison of components (e.g., the comparison of resistors with a standard resistor, inductors with a standard inductor, and capacitors with a standard capacitor).

comparison measurement A measurement in which a quantity or component is compared with a known, similar quantity or component value, rather than having the measurement displayed directly by a meter. Examples: bridge measurements, potentiometric measurements, and frequency matching.

compass **1.** Any of several instruments for determining direction on the earth's surface [e.g., magnetic (mariner's) compass and gyrocompass]. **2.** A radio direction finder. **3.** An instrument for drawing circles.

compatibility **1.** A desirable condition in which devices or systems can function efficiently together, without any modification of equipment. **2.** In computer operations, a desirable condition in which different computers can run the same software, without any modification of hardware or software.

compatible color television A color-television system whose transmissions can be received in black and white by any ordinary monochromatic receiver.

compatible integrated circuit A hybrid integrated circuit (IC) that has an active element inside the integrated structure and a passive element deposited on its insulated outer surface.

compensated amplifier A wideband amplifier whose frequency range is extended by special components and circuit modifications. Also see COMPENSATING CAPACITOR and COMPENSATING COIL.

compensated diode detector A diode detector in which a positive dc voltage from the automatic-gain-control (AGC) rectifier is applied to the diode anode. The voltage is always proportional to the signal carrier. The arrangement allows the diode to handle a heavily modulated AM signal without producing excessive distortion.

compensated-impurity resistor A resistor consisting of a diffused semiconductor material to which are added controlled amounts of n- or p-type dopants (impurities).

compensated-loop direction finder A direction finder whose loop antenna is complemented by another antenna for polarization-error compensation.

compensated semiconductor A doped semiconductor material in which the acceptor impurity cancels the effects of the donor impurity.

compensated volume control A combination volume-tone control that provides bass boost at low volume levels to compensate for the ear's deficiency at low frequencies.

compensating capacitor **1.** A capacitor that has a temperature coefficient of capacitance numerically equal to, but having the opposite sign from, that of another capacitor in a tank or other circuit. When the capacitors are connected in parallel, a temperature-induced value change in the main capacitor is balanced by an equal and opposite change in the compensating capacitor; the net capacitance of the circuit does not change. This greatly reduces frequency drift. **2.** In a video amplifier, a large capacitance connected between ground and a tap on the collector or drain resistor to boost low-frequency response. Compare COMPENSATION COIL. **3.** A usually low-capacitance capacitor of known temperature coefficient, operated in combination with a main capacitor to reduce capacitance/temperature drift of the latter to zero or to some desired positive or negative value.

compensating diode A junction diode used to temperature-stabilize a transistor circuit. It is usually forward-biased in the base-bias network of the transistor.

compensating filter **1.** A selective filter used for the purpose of eliminating some irregularity in the frequency distribution of received energy. **2.** A filter used to change the wavelength distribution of electromagnetic energy.

compensating resistor **1.** A low-value resistor of known temperature coefficient, connected in series with a main resistor to reduce the resistance/temperature drift to zero, or to some desired positive or negative value. **2.** See TRIMMER RESISTOR.

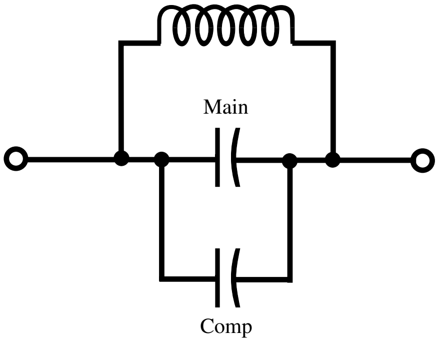

compensating capacitor, 1.

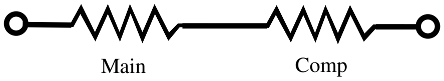

compensating resistor, 1.

compensation Adjusting a quantity, manually or automatically, to obtain precise values, or to counteract undesired variations. Example: temperature compensation of electronic components. For illustration, see COMPENSATING CAPACITOR, **1**.

compensation coil In a video amplifier, an inductor connected in series with the collector or drain resistor, or in the coupling path between stages, or both, to boost high-frequency response.

compensation filter See COMPENSATING CAPACITOR, **2**.

compensation signal A signal recorded on a tape track containing computer data, to ensure that the tape plays back at exactly the correct speed at all times.

compensation theorem An impedance (*Z*) in a network can be replaced by a generator having zero internal impedance, and whose generated voltage equals the instantaneous potential difference produced across *Z* by the current flowing through it. Compare MAXIMUM POWER TRANSFER THEOREM, NORTON'S THEOREM, RECIPROCITY THEOREM, SUPERPOSITION THEOREM, and THEVENIN'S THEOREM.

compensator A device or circuit that facilitates the adjustment of a quantity, manually or automatically, to obtain precise values, or to counteract undesired variations.

compilation time The period during which a program is compiled, as distinct from RUN TIME.

compile **1.** To unify computer subroutines into an all-encompassing program. **2.** To gather information or data together into a single file or file set.

compiler In computer operations, a program that changes a HIGH-LEVEL LANGUAGE, such as BASIC, C, C++, COBOL, or FORTRAN, into MACHINE LANGUAGE. A compiler must be written especially for the high-level language being used.

compiler language Any computer language that serves as an interface between the operator and the computer.

compiler program A program that converts compiler language into machine language.

compiling routine In digital computer operation, a routine permitting the computer itself to construct a program to solve a problem.

complement **1.** The difference between a number and the radix (modulus or base) of the number system. For example, the complement of 7 is equal to 3 (because 10 – 7 = 3) in the decimal (radix-10) number system. **2.** Also called *ones complement.* In computer operations, a representation of the negative value of a binary number. All the available digits are set to 1, and then the number in question is subtracted. For example, the complement of 101 is equal to 010 (because 111 – 101 – 010); the complement of 1001 is equal to 0110 (because 1111 – 1001 = 0110).

complementary A Boolean operation whose result is the same as that of another operation, but with the opposite sign; thus, OR and NOR operations are complementary.

complementary colors **1.** In the additive color system, two colors that produce light gray or white when combined. **2.** In the subtractive color system, two pigments that produce dark gray or black when combined. **3.** Colors or pigments that are opposite each other on the color wheel.

complementary constant-current logic A form of bipolar logic with high operating speed and high component density.

complementary metal-oxide semiconductor Also sometimes called *complementary metal-oxide silicon.* Acronym, CMOS (pronounced "seamoss"). A digital integrated-circuit (IC) technology, in which logic gates are formed by n-channel/p-channel pairs of metal-oxide-semiconductor field-effect transistors (MOSFETs) fabricated on a substrate. Noted for high speed and low current drain.

complementary operator The logical negation (NOT) operation.

complementary pushpull circuit See COMPLEMENTARY-SYMMETRY CIRCUIT.

complementary rectifier In the output circuit of a magnetic amplifier, nonsaturating half-wave rectifier elements.

complementary silicon-controlled rectifier A silicon-controlled rectifier that has polarity opposite from the usual silicon-controlled rectifier.

complementary-symmetry circuit A bipolar-transistor circuit that uses an npn and pnp transistor. The transistors conduct during opposite half-cycles of the input signal, the result being that push-pull output is provided with a single-ended input; no phase-splitting input circuit is required. The complementary-symmetry circuit offers very low output impedance, permitting a loudspeaker voice coil (or other low-impedance load) to be operated directly without a coupling transformer.

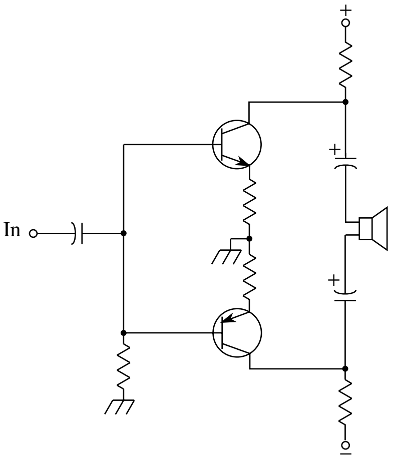

complementary-symmetry circuit

complementary tracking A control system in which several secondary (slave) devices are controlled by a primary (master) device.

complementary transistors A transistor pair of opposite polarity operated in a complementary-symmetry circuit or its equivalent.

complementary wave An electromagnetic wave in a transmission line that occurs as a result of reflection. Any impedance discontinuity will result in complementary waves.

complementer A logic circuit that provides an output pulse when there is no input pulse, and vice versa. Also called INVERTER and NOT CIRCUIT.

complement number In a base-n number system, for a given positive integer p less than n, the positive integer m such that $m + p = n$. For example, in the decimal (base-10) system, the complement of 4 is 6, the complement of 7 is 3, and the complement of 9 is 1. In the hexadecimal (base-16) number system, the complement of 4 is 12, the complement of 7 is 9, and the complement of 9 is 7.

complement-number handling A computer system in which the operations are carried out via the complements of the input numbers.

complement-setting technique A process of determining the number of pulses required to complete the switching of a counter circuit when it is started at some state higher than full zero. The number of pulses required for completion is equal to the number that represents the starting state's complement.

complete carry In digital computer operation, a system permitting all carries to generate carries.

complete circuit See CONTINUOUS CIRCUIT.

complete modulation Modulation to the maximum extent possible while maintaining acceptable circuit or system operation.

complete operation In computer operations, the condition in which the machine rigorously follows program instructions.

complete routine A vendor-supplied computer program that is usable without modification.

complex function **1.** A mathematical function of a complex-number variable. **2.** An integrated circuit (IC) containing two or more subcircuits that perform an operation more complicated that of any one of the circuits alone.

complex notation Notation taking into consideration both the real-number and imaginary-number components of a quantity. Thus, impedance (Z) is a complex quantity that includes a resistive (real) component (R) and a reactive (imaginary) component (jX). See COMPLEX NUMBER and COMPLEX OPERATOR.

complex number A number expressed in complex notation (e.g., $a + jb$, where a and b are real numbers and j is the COMPLEX OPERATOR). Can also be expressed as a point or a vector in an ARGAND DIAGRAM.

complex operator The unit imaginary number, represented as j by engineers and as i by mathematicians. This number is defined mathematically as the positive square root of -1.

complex parallel permeability An expression of the permeability of an inductor core under actual operating conditions, assuming zero loss in the conductors of the coil winding. A parallel combination of reactance and resistance.

complex periodic wave A periodic wave composed of a sine-wave fundamental and certain harmonics in specific proportions.

complex permeability An expression of inductor-core permeability, obtained from the mathematical ratio of the magnitudes of the vectors representing the induction and electromagnetic field strength within the core.

complex plane A Cartesian coordinate system with real numbers on the horizontal axis and imaginary numbers on the vertical axis. Used for vectorial representation of complex numbers. See ARGAND DIAGRAM.

complex quantity A quantity containing both real and imaginary components. Example: Impedance (Z) is a complex combination of resistance R (a real component) and reactance jX (an imaginary component): $Z = R + jX$.

complex radar target A radar target that is large enough in theory to be detected by radar, but, because of its geometry, cannot be detected. This effect is the result of phase combinations of signal components reflected from various surfaces on the target.

complex series permeability An expression of complex permeability of an inductor core under actual operating conditions, assuming zero loss in the conductors of the coil winding. A series combination of reactance and resistance.

complex steady-state vibration Periodic vibration with more than one sine-wave component.

complex tone An audio tone made up of more than one sine-wave component.

complex variable A variable having real and imaginary parts.

complex waveform The shape of a COMPLEX PERIODIC WAVE. It is the resultant of the individual sine-wave components (i.e., of the fundamental and the harmonics).

complex-wave generator A signal generator whose output signal is any of several selectable waveforms and frequencies (or repetition rates). Also see FUNCTION GENERATOR.

compliance **1.** The ease with which a material can be flexed or bent, an important characteristic of transducers (such as loudspeakers). Expressed in cm/dyne, compliance is the reciprocal of stiffness, and is the acoustical or mechanical equivalent of capacitance. **2.** A measure of the output impedance of a switched-current signal source. Generally given as maximum current for a certain change in the voltage.

compliance range The voltage range required to maintain a constant current throughout a load-resistance range.

compliance voltage The range over which the output voltage of a constant-current power supply must swing in order to maintain a steady current in a varying load.

compliance-voltage range The output voltage range of a constant-current power supply.

component **1.** A device or part used in a circuit to obtain some desired electrical action [e.g., a resistor (passive component) or an integrated circuit (active component)]. Also see ACTIVE COMPONENT and PASSIVE COMPONENT. **2.** An attribute inherent in a device, circuit, or performance (e.g., the REACTIVE COMPONENT of an inductor). **3.** A specified quantity or term (e.g., the WATTLESS COMPONENT of ac power). **4.** A piece of equipment in a high-fidelity sound system.

component density The number of components (see COMPONENT, **1**) in an electronic assembly of a given physical volume.

component failure rate **1.** The percentage of components, out of a specified group, that can be expected to fail within a specified length of time. **2.** The frequency with which a given component, in a certain application, can be expected to fail.

component layout The mechanical arrangement of components (see COMPONENT, **1**) in an electronic assembly.

component stress The electrical or mechanical strain to which a component is subjected. In general, the greater the stress, the higher the component failure rate.

composite cable A cable containing other cables of different types.

composite circuit A circuit handling telegraphy and telephony simultaneously without causing mutual interference.

composite color A color that is not one of the COLOR PRIMARIES, but instead, consists of a combination of the three color primaries.

composite color signal The complete color television signal, including all picture, color, and control components.

composite conductor A set of wires connected in parallel. The wires are often, but not necessarily, of identical size and constitution.

composite current A current having both alternating-current (ac) and direct-current (dc) components; an alternating current superimposed on a direct current. Also called *fluctuating current.*

composite curve A curve or pair of curves showing two modes of operation, as of biased and unbiased conditions.

composite filter A filter consisting of more than one section. The sections might be, but often are not, identical.

composite video signal The television picture signal containing picture information and sync pulses.

composite-video-signal distortion Distortion of the composite video signal as evidenced by overshooting, ringing, and sync-pulse shortening.

composite voltage A voltage having both alternating-current (ac) and direct-current (dc) components; an ac voltage superimposed on a dc voltage. Also called *fluctuating voltage.*

composite wave filter Two or more wave filters (not necessarily of the same type) operated in cascade.

composition resistor A resistor made from a mixture of materials, usually finely powdered carbon and a binder.

compound A substance in which the atoms of two or more elements have united chemically to form a molecule. For example, an atom of cadmium (Cd) and one of sulfur (S) combine to form a molecule of cadmium sulfide (CdS).

compound connection A direct connection of two transistors, the amplified output of the first being further amplified by the second. The connection provides extremely high current gain. Also called DARLINGTON PAIR.

compound generator A generator that has both series and shunt fields. Also called *compound-wound generator.*

compound horn A horn reflector used for transmission of microwave energy. The faces of the horn approach four geometric plane surfaces as the distance from the center increases.

compound modulation A system of successive modulation, the modulated wave from one step becoming the modulating wave in the next. Also called *multiple modulation.*

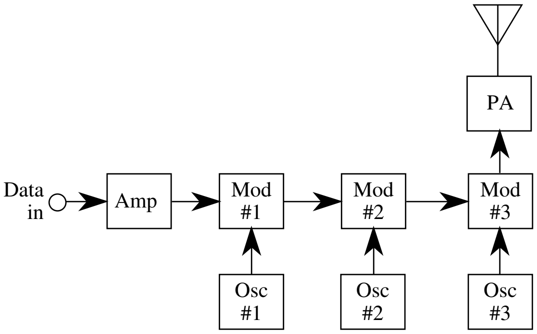

compound modulation

compound motor An electric motor having both series and shunt fields. Also called *compound-wound motor.*

compound transistor Two or more transistors directly coupled in the same envelope for increased amplification. Also see COMPOUND CONNECTION.

compound-wound generator See COMPOUND GENERATOR.

compound-wound motor See COMPOUND MOTOR.

compress **1.** In communications, to reduce or minimize the bandwidth of a signal. **2.** In communications, the processing of a signal to increase low-level components and thereby raise the average power level relative to the peak power level. **3.** In computer operations, to reduce or minimize the number of bits in a digital signal or file, while still retaining all the essential information. Compare EXPAND.

compressed-air capacitor A high voltage air-dielectric capacitor enclosed in a case in which the air pressure is held at several atmospheres. The device exploits the dielectric strength of compressed air, which is higher than that of air at normal pressure.

compressed-air speaker A speaker that uses an airtight chamber to enhance the acoustic reproduction at certain frequencies.

compression **1.** In communications, the reduction or minimization of signal bandwidth. **2.** In communications, the processing of a signal to increase low-level components and thereby raise the average power level relative to the peak power level. Usually, a logarithmic function. **3.** In computer operations, the reduction or minimization of the number of bits in a digital signal or file, while still retaining all the essential information. Compare EXPANSION.

compression ratio In a system using COMPRESSION, the ratio A_1/A_2, where A_1 is the gain (or transmission) at a reference-signal level and A_2 is the gain (or transmission) at a specified higher signal level.

compression wave A wave disturbance that travels via longitudinal motion of particles in a medium. Sound waves through air are the most common example.

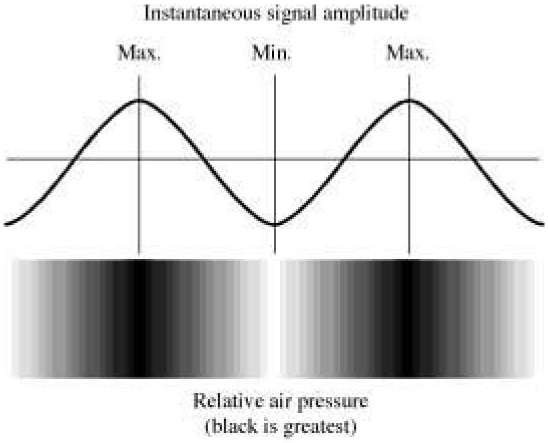

compression wave

compressor A circuit or device that limits the amplitude of its output signal to a predetermined value, despite wide variations in input signal amplitude.

compressor driver unit A loudspeaker that works into an air space connected by a throat to a horn, rather than by driving a diaphragm.

Compton diffusion An effect that occurs when a photon and electron collide. Some of the energy from the photon is transferred to the electron. On a large scale, such collisions result in diffusion of electromagnetic waves.

Compton effect The increase in wavelength (decrease in frequency) of X-rays scattered by the electrons of lighter atoms bombarded with the X-rays.

Compton shift See COMPTON EFFECT.

compute To perform a mathematical operation by means of a relatively simple process. Thus, a digital computer solves intricate problems using simple arithmetic steps. Compare CALCULATE.

computer A device or machine for performing mathematical operations on data, and producing the results as information or control signals. There are numerous types, the most common being the *digital computer.*

computer-aided design Abbreviation, CAD. The use of computers in conceiving, developing, and perfecting new products.

computer-aided manufacturing Abbreviation, CAM. The use of automated manufacturing systems, such as assembly lines, that are partially or totally controlled by computers.

computer antibody Also called *vaccine.* A small subprogram designed to eliminate viruses from computer systems.

computer-assisted instruction Abbreviation, CAI. The use of computers as teaching and training aids.

computer code See MACHINE LANGUAGE.

computer consciousness The degree to which a machine can be considered aware of its own existence. Until recently, this idea was considered ridiculous. But as microprocessor power continues to grow, some researchers now consider it worth thinking about.

computer-controlled catalytic converter A microprocessor-controlled system for automatically supervising gaseous emissions exhausted by a motor vehicle. An oxygen sensor monitors the exhaust stream, and the associated electronic system adjusts the air-to-fuel ratio of the carburetor to reduce smog-producing pollutants in the exhaust.

computer diode A semiconductor diode having low capacitance and fast RECOVERY TIME, thus suiting it to rapid switching in computer circuits and to very-high-frequency applications.

computer engineer A person skilled in the theory and application of computers, related equipment, and associated mathematics.

computer file See FILE.

computer game See VIDEO GAME.

computer graphics **1.** The use of computers to assist in drawing and drafting, and in the processing of video images such as photographs. **2.** Broadly, any computer-generated or computer-processed image.

computer instruction See INSTRUCTION.

computer interfacing apparatus The equipment used to connect a computer to other systems, and to peripherals.

computerized axial tomography Abbreviation, CAT. A multiple X-ray system that enables the observer to obtain cross-sectional images of the internal organs of the body.

computer map A blueprint, used in conjunction with machine vision, sonar, radar or beacons, that a mobile robot can use as a navigational aid. One or more such blueprints are stored in the robot controller's main memory.

computer music **1.** Music that is composed by a computer. **2.** See MUSICAL INSTRUMENT DIGITAL INTERFACE.

computer program See PROGRAM.

computer programmer A person skilled in devising and/or writing the routines that a digital computer uses to solve problems or process data.

computer storage tube A cathode-ray tube in which the electron beam scans and stores information in thousands of memory cells on a target. A cell "remembers" by acquiring and holding an electrostatic charge when it is struck by the beam from the writing gun. Information taken is read out of a cell by a second beam from the reading gun.

computer system A central processor and its associated online and offline peripherals, such as a monitor, modem, printer, optical scanner, magnetic disk drives, CD-ROM drive, and tape backup.

computer technician A professional skilled in building, repairing, and maintaining computers, and who, occasionally, designs them. Usually works under the supervision of a computer engineer.

computer terminal **1.** A teleprinter or video display unit and keyboard, used by human operator(s) of a computer. **2.** An interface between a computer and its human operator(s).

computer/TV interface A device or circuit for delivering the output of a digital computer to a standard television receiver so that the latter can serve as a GRAPHIC TERMINAL.

computer virus A deliberately created and disseminated subprogram or piece of programming code, that electronically spreads through computer systems and hinders operation. Usually diverts the computer(s) from intended functions; sometimes causes a catastrophic malfunction. Often exists undetected, being transferred from one computer to another by means of diskettes or software.

computer word See WORD.

computing amplifier See OPERATIONAL AMPLIFIER.

computing machine See COMPUTER.

concatenation **1.** A method of speed control for a 3-phase motor in which two induction motors are operated with their shafts coupled together. The stator of the first motor is connected to the 3-phase supply, and the slip rings of this motor are connected to the field of the second motor. The slip rings of the second motor are connected to the three ganged sections of a Y-rheostat used for adjusting the speed. **2.** Arrangement of a set into a series.

concentrated-arc lamp A brilliant low-voltage lamp, containing nonvaporizing electrodes in an inert-gas atmosphere. An arc across the electrodes creates the light source.

concentrated winding A coil winding that has a large number of turns in a small space.

concentration cell An electrolytic cell in which two electrodes are immersed in solutions of the same compound but having different combinations. The voltage is usually very small, 0.1 volt or less.

concentration gradient Between points in a semiconductor, the difference in electron or hole concentration.

concentric cable See COAXIAL CABLE.

concentric capacitor A fixed or variable capacitor whose plates are concentric cylinders. Also called *concentric-plate capacitor.*

concentric jack See COAXIAL JACK.

concentric line See COAXIAL LINE.

concentric-line oscillator A stable, self-excited oscillator whose frequency-determining tank consists principally of a section of concentric (coaxial) line. Used primarily at ultra-high frequencies (UHF).

concentric plug See COAXIAL PLUG.

concentric receptacle See COAXIAL RECEPTACLE.

concentric tank See COAXIAL TANK.

concentric-wound coil A combination of two or more coils wound on top of, and insulated from, each other.

conceptual modeling A technique for solving problems by devising a mathematical model based on the results of an experiment; experiments performed on the model are used to verify its validity.

concurrent conversion In computer operations, running conversion and conventional programs together. Also see CONVERSION PROGRAM.

concurrent processing See MULTIPROGRAMMING.

condenser **1.** An obsolete term for CAPACITOR. **2.** A mirror or lens for concentrating light (on an object, for example). **3.** Something that condenses a gas or vapor. **4.** See CONDENSER MICROPHONE.

condenser antenna A two-wire horizontal antenna system in which the radiator is a wire situated above a counterpoise.

condenser microphone Also called *capacitor microphone.* A microphone in which a tightly stretched metal diaphragm forms one plate of an

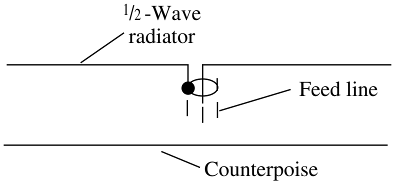

condenser antenna

air capacitor, and a closely situated metal plug forms the other plate. A dc bias voltage is applied to the arrangement. Impinging sound waves cause the diaphragm to vibrate, varying the capacitance and causing the output current to fluctuate accordingly.

condensing routine In computer operations, a program that compresses data. See COMPRESSION, **3**.

condensite A plastic insulating material whose base is phenol formaldehyde resin.

conditional Pertaining to a quantity or phenomenon that depends on some external factor, and is therefore subject to change.

conditional branch A point in a computer program where a relational test is performed, and the statement line in which the test is made is left so that an out-of-sequence instruction can be implemented. Such a branch might be made, for example, following a statement, such as "IF Z = Y THEN GO TO LINE 380."

conditional branch instruction The instruction in a computer program that causes a CONDITIONAL BRANCH.

conditional implication operation A Boolean operation in which the result of operand values X and Y are such that the output is high only if input X is high and input Y is low. Also called *inclusion* or *if-then operation.*

conditional jump See CONDITIONAL BRANCH.

conditional stop instruction In a computer program, an instruction that can cause a halt in the run, as dictated by some specified condition.

conditional transfer See CONDITIONAL BRANCH.

condition code A set of constraints for a computer program; sets limits on what can be done with the computer under certain circumstances.

conditioning **1.** The process of making equipment compatible for use with other equipment. Generally involves some design or installation changes. **2.** Interfacing.

Condor A continuous-wave navigational system that produces a cathode-ray-tube display for automatically determining the bearing and distance from a ground station.

conductance Symbol, *G*. Unit, siemens. The ability of a circuit, conductor, or device to conduct electricity. Conductance in siemens is the reciprocal of resistance: $G = 1/R$, where R is the resistance in ohms.

conducted heat Heat transferred by conduction through a material substance, as opposed to *convection* and *radiation.* A heatsink conducts dissipated energy away from a transistor, whereupon convection and radiation allow heat to escape from the sink.

conduction **1.** The propagation of energy through a medium, depending on the medium for its travel. **2.** The transfer of electrons through a wire. **3.** The transfer of holes through a P-type semiconductor material. **4.** Heat transfer through a material object (see CONDUCTED HEAT).

conduction angle See ANGLE OF CONDUCTION.

conduction band In the arrangement of energy levels within an atom, the band in which a free electron can exist; it is above the *valence band* in which electrons are bound to the atom. In a metallic atom, conduction and valence bands overlap; but in semiconductors and insulators, they are separated by an energy gap.

conduction current **1.** The electromagnetic-field flow that occurs in the direction of propagation. A measure of the ease with which the field is propagated. **2.** Current in a wire or other conductor.

conduction-current modulation In a microwave tube, cyclic variations in the conduction current; also, the method of producing such modulation.

conduction electron See FREE ELECTRON.

conduction error In a temperature-acutated transducer, error caused by conduction of heat between the sensor and the mounting.

conduction field An energy field that exists in the vicinity of an electric current.

conductive coating A conducting layer applied to the glass envelope of a cathode-ray tube, such as an oscilloscope tube or picture tube.

conductive coupling See DIRECT COUPLING.

conductive material See CONDUCTOR.

conductive pattern The pattern of conductive lines and areas in a printed circuit.

conductivity Unit, S/m (siemens per meter). An expression of conductance per unit length of a material; the reciprocal of *resistivity.*

conductivity meter A device for measuring electrical conductivity. Generally, such a device is calibrated in siemens.

conductivity modulation In a semiconductor, the variation in conductivity that results from a variation of charge-carrier density.

conductivity-modulation transistor A transistor in which the bulk resistivity of the semiconductor material is modulated by minority carriers.

conductor **1.** A material that allows charge carriers (usually electrons) to move with ease among atoms. Examples are metals, electrolytes, and ionized gases. Substances vary widely in their suitability as conductors; the conductivity of commercial copper, for example, is almost twice that of aluminum. Compare INSULATOR. **2.** An individual conducting wire in a cable, insulated or uninsulated.

conduit A hollow tube, made of plastic or metal, through which wires, cables, and other transmission media are fed.

cone The conical diaphragm of a (usually dynamic) loudspeaker.

cone antenna An antenna in which the radiator is a sheet-metal cone or a conical arrangement of rods or wires.

Conelrad An early amplitude-modulation (AM) broadcast protocol, intended for use in the event of a nuclear war. Now replaced by the EMERGENCY BROADCAST SYSTEM.

cone marker A UHF marker beacon whose conical energy lobe radiates vertically from a radio-range beacon station. Aircraft in flight use such markers to accurately locate the beacon station.

cone of protection The zone surrounding a lightning rod, in which the chances of a lightning strike are greatly reduced. The cone has an apex angle of 45°, relative to the rod. Objects entirely within this cone are unlikely to be struck (although it is still possible).

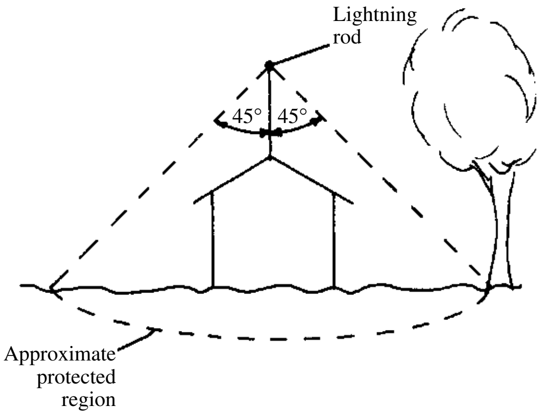

cone of protection

cone of silence A small zero-signal zone directly over a low-frequency radio-range beacon. The zone is the product of the combined directive properties of the beacon transmitting antenna and the antenna on an aircraft.

cone speaker A loudspeaker having a sound-producing cone (diaphragm) made of specially treated paper or other material, as opposed to a loudspeaker having a flat diaphragm.

confetti On a color TV screen, color spots caused by chrominance-amplifier noise.

confidence The probability that a predicted result will occur.

confidence factor Confidence, expressed either as a fraction (between 0 and 1) or as a percentage.

confidence interval The range over which a parameter can vary so that a given confidence factor is maintained.

confidence level See CONFIDENCE FACTOR.

confidence limitations The maximum and minimum points of a confidence interval. Outside the confidence-limitation points, the confidence factor drops below the required minimum.

configuration **1.** The characteristic arrangement of components in an electronic assembly, or of the equipment symbols in the corresponding circuit diagram. **2.** Computer system.

configuration state In a computer system, an expression of the availability status of a device for a given application. A *configured-in* device is available; a *configured-out* device is available, but is restricted to certain users; a *configured-off* device is unavailable.

configuration table Within a computer's operating system, a table that provides the configuration state for various system units.

configured-in See CONFIGURATION STATE.

configured-off See CONFIGURATION STATE.

configured-out See CONFIGURATION STATE.

conformance The degree to which a quantity or variable corresponds to a standard or to expectations.

conformance error The extent (usually expressed as a percentage) to which conformance is lacking.

conical antenna See CONE ANTENNA.

conical horn A horn (antenna, loudspeaker, or sound pickup) having the general shape of a cone: the cross-sectional area varies directly as the square of the horn's axial length.

conical monopole antenna An unbalanced broadband antenna that derives its name from its shape. It is usually constructed from wire and must be operated against a good radio-frequency (RF) ground.

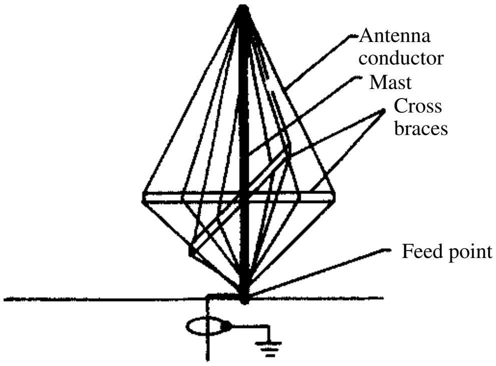

conical monopole antenna

conical scanning In radar transmission, a method of scanning in which the beam describes a cone, at the apex of which is the antenna.

conic sections The geometric plane figures that result from the intersection of a cone with a plane. These figures are the *circle*, the *ellipse*, the *parabola*, and the *hyperbola.*

conjugate For a given complex number $A + jB$, the quantity $A - jB$. When complex conjugates are multiplied together, the result is $A^2 + B^2$.

conjugate branches In a network, two branches of such a nature that a signal in one has no effect on the other.

conjugate bridge A bridge in which the detector and generator occupy positions opposite to those in a conventional bridge of the same general type.

conjugate impedance For a given complex impedance, $R + jX$, where R is the resistive component and jX is the reactive component, the impedance: $R - jX$. The resistance is identical; the reactance is of equal magnitude, but opposite sign (capacitive as opposed to inductive, or vice versa).

conjunction The logical AND operation.

connect To provide an electrical path between two points.

connection The point at which two conductors are physically joined.

connective An operation symbol written between operands.

connector **1.** A device that provides electrical connection. **2.** A fixture (either male or female) attached to a cable or chassis for quickly making and breaking one or more circuits. **3.** A symbol that connects points on a flowchart.

conoscope A device that uses focused polarized light to examine crystals (as in checking the optical axis of a quartz crystal).

consequent poles The poles of an equivalent single magnet that is formed when two magnets are aligned with their two identical poles together. Thus, when the two north poles are placed together, the consequent poles are a south pole at each end and a north pole at the center.

conservation of energy **1.** The preservation of the potential for work by a given quantity of energy—even when it undergoes a change in form within a system. **2.** The *law of conservation of energy*, which states that energy can be neither created nor destroyed, but only changed in form.

console **1.** The main station or position for the control of electronic and/or computer equipment. **2.** The equipment at a fixed location. **3.** An equipment-containing cabinet that stands on the floor. **4.** Equipment permitting communication with a computer. Also called *dumb terminal.*

consonance **1.** Harmony between audio tones. **2.** Acoustical or electrical resonance between bodies or circuits that are not physically connected.

constant **1.** A quantity whose value remains fixed, such as the speed of light in a vacuum. Compare VARIABLE. **2.** The value of a component specified for use in a particular electronic circuit. **3.** An electronic component, particularly a capacitance or inductance. **4.** In a computer program, data items that remain unchanged for each run.

constant-amplitude recording In sound recording, the technique of holding the maximum amplitude of the signal steady as the frequency changes.

constantan An alloy of copper and nickel used in some thermocouples and standard resistors.

constantan-platinum thermocouple A thermocouple that uses the junction between constantan and platinum wires, which is contained in thermocouple-type meters.

constant area As allocated by a computer program, an area of memory that holds constants.

constant bandwidth In a broadband tuned circuit, bandwidth that does not change with frequency.

constant current A current that undergoes no change in value as it flows through a changing resistance. Compare CONSTANT VOLTAGE.

constant-current characteristic A condition in which the current through a circuit remains constant—even if the voltage across the circuit increases or decreases.

constant-current curve A graph in which the dependent variable is an electric current that levels off at, or approaches, a specific maximum. An example is the collector-current versus collector-voltage curve for a bipolar transistor.

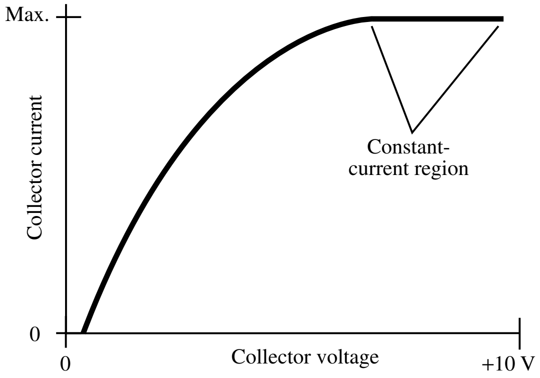

constant-current curve

constant-current drive Driving power obtained from a constant-current source.

constant-current modulation See CHOKE-COUPLED MODULATION.

constant-current power supply See CONSTANT-CURRENT SOURCE.

constant-current sink See CURRENT SINK.

constant-current source A power supply whose current remains steady during variations in load resistance. Also called *constant-current supply* and *current-regulated supply.* Compare CONSTANT-VOLTAGE SOURCE.

constant-current supply See CONSTANT-CURRENT SOURCE.

constant-current transformer A transformer supplied from a constant-voltage source that automatically delivers a constant current to a varying secondary load.

constant-k filter Also called a *Zobel filter.* A filter section in which Z_1Z_2 equals k^2 at all frequencies, where Z_1 is the impedance of the series element and Z_2 is the impedance of the shunt element.

constant-power dissipation line A line connecting points on a family of current-voltage characteristic curves, the points corresponding to the maximum power that can safely be dissipated by the device to which the curves apply.

constant-resistance network A circuit of resistors that, when terminated in a resistance load, presents a constant resistance to a driving source under various conditions of operation.

constant-speed motor **1.** Also called a *shunt motor.* A motor whose speed varies little, or not at all, with variations in the armature current. **2.** A motor that runs at an unvarying speed through the action of associated automatic electronic control circuitry.

constant voltage A voltage that does not change as the load resistance varies. Compare CONSTANT CURRENT.

constant-voltage, constant-current supply A combination current-regulated and voltage-regulated power supply; delivers constant current to low load resistances and constant voltage to high load resistances.

constant-voltage drive Driving power obtained from a CONSTANT-VOLTAGE SOURCE.

constant-voltage source A power supply whose output voltage remains steady during variations in load current. Also called *constant-voltage supply* and *voltage-regulated supply.*

constant-voltage transformer A special transformer used to reduce variations in power-line voltage. A capacitor in the device causes a winding to resonate at the line frequency (e.g., 60 Hz). This tends to maintain a more constant current than would be the case in an ordinary transformer.

construct A source (user's) computer program statement that, when implemented, produces a predetermined effect.

consumer reliability risk **1.** The chance a consumer takes when buying a component or piece of equipment that has not been subjected to quality-assurance/quality-control (QA/QC) testing. **2.** An expression of the failure rate for a consumer item.

contact **1.** A conducting body (such as a button, disk, or blade) that serves to close an electric circuit when pressed against another conductor. Example: switch contact and spring contact. **2.** The state of being touched together, as when two conductors are brought into contact to close a circuit.

contact arc The arc that initially occurs when current-carrying contacts are separated.

contact area **1.** The face of an electrical contact. **2.** The common area shared by two conductors in mutual contact.

contact bounce The springing apart or vibration of contacts upon making or breaking.

contact chatter The abnormal vibration of mating contacts, caused by contact bounce or by an extraneous alternating current.

contact-closure input The input circuit of a device, such as a control-system amplifier, that is actuated by the closing of switching contacts. Compare CONTACT-OPEN INPUT.

contact combination The set of contacts on a switch or electronic relay.

contact detector A rectifier or demodulator, composed of two dissimilar materials in contact with each other. Semiconductor diodes are of this general type. Some contact-detector action can be obtained with two dissimilar fine wires (such as copper and iron) by touching their tips lightly together.

contact EMF Short for *contact electromotive force*; also called *contact potential.* A low direct-current (dc) voltage that is sometimes generated by the contact of two dissimilar materials.

contact follow The tendency of relay contacts to follow the actuating signals.

contact force **1.** The force with which relay contacts close with a given amount of coil current. **2.** The force with which a pair of relay contacts are held together when current flows through the coil. **3.** In a mercury-wetted relay, the force exerted by the mercury on the contacts as the relay closes.

contact gap The distance between contacts when they are open.

contact load **1.** The power dissipated by a load that is connected to a power supply through a closed set of contacts. **2.** The current passing through a set of closed contacts.

contact microphone A microphone placed in direct contact with a vibrating surface for pickup. Actuated by the vibration of a solid, rather than by the movement of air molecules.

contact miss **1.** The improper alignment of contacts in a switch or relay. **2.** The condition of relay contacts not lining up properly.

contact modulator An electromechanical CHOPPER.

contact-open input The input circuit of a device, such as a control-system amplifier, that is actuated by the opening of switching contacts. Compare CONTACT-CLOSURE INPUT.

contactor A switch used for frequent opening or closing of a circuit. An example is a relay contactor used for keying a transmitter.

contactor noise **1.** Electrical noise that is the product of make-and-break contact action or fluctuations in conduction when the contacts are

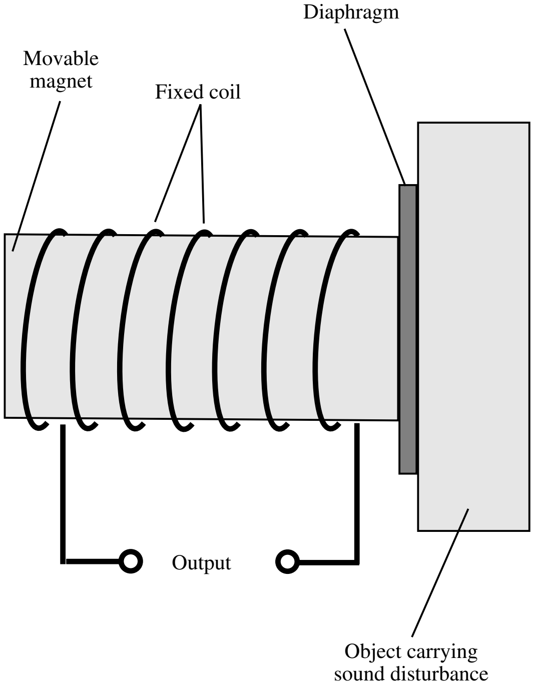

contact microphone

closed. **2.** Sounds coming directly from contacts that are opening and closing.

contact potential The small direct-current (dc) voltage that results from the bombardment of an electrode by electrons, when the electrode has no external voltage applied to it.

contact pressure The pressure that holds contacts together.

contact protector A component (such as a diode, capacitor, resistor, or combination of these) that serves to suppress contact arcing.

contact rating The maximum current, voltage, and/or power specified for a given set of contacts.

contact rectifier A rectifier consisting of two dissimilar materials in direct contact. Examples: copper and copper oxide, magnesium and copper sulfide, selenium and aluminum, and germanium and indium.

contact resistance The resistance of the closed contacts of switches, relays, and other similar devices. Normally, this is a very small resistance.

contact separation See CONTACT GAP.

contact strip See TERMINAL STRIP.

contact switch An electromechanical switch that uses contacts to make and break a circuit, as compared with an electronic switch that uses semiconductor devices.

contact travel The distance over which a relay or switch contact must move to close a circuit.

contact wetting The use of mercury (a conducting liquid) to improve the action of a relay contact or contacts.

contact wipe A sliding motion between closed contacts. Helps to establish a good connection and to keep the contact surfaces clean.

container file See CONTROLLING FILE.

contaminated material **1.** A semiconductor material containing some undesired substance. **2.** A material unintentionally made radioactive.

contamination **1.** The presence of an impurity in a substance. **2.** The addition of a radioactive material to a substance. **3.** In a coaxial cable, the tendency for the jacket material to bleed through the outer braid into the dielectric, resulting in increased loss.

content-addressed storage In a computer, memory- or data-storage locations identified by content (see CONTENTS), instead of by address. Also called *associative storage.*

contention The result of interference among more than one transmitting station on the same communications channel.

contents **1.** The data in a computer random-access memory (RAM). **2.** The data in a specific storage location, such as on a hard disk, diskette, or CD-ROM.

context **1.** The environment in which a word is used in a natural language (such as English, Spanish, or Russian). Important in speech recognition and speech synthesis. **2.** The environment in which a string of characters, composing a data unit or word, is used in a computer program.

Continental code A version of the Morse code used internationally in radiotelegraphy. Also called *International Morse code* and *general service code.* Compare AMERICAN MORSE CODE.

continuity A condition of being uninterrupted—especially pertaining to current flowing in an electrical or electronic circuit.

continuity test A test of the completeness of an electrical path. Ideally, the only concern is whether the circuit is open or closed, but sometimes circuit resistance is also of interest.

continuity tester A device (such as an ohmmeter, battery and buzzer, and battery and lamp) with which a continuity test can be made.

continuity writer The person who prepares copy for a radio or television broadcaster.

continuous carrier A medium (such as a radio-frequency wave) that will convey information (as when the carrier is modulated) with no disruption of the medium itself.

continuous circuit An uninterrupted circuit.

Continuous Commercial Service Abbreviation, CCS. A category in which safe operating parameters are listed for electronic components and communications equipment operated over long, uninterrupted periods. Compare INTERMITTENT COMMERCIAL AND AMATEUR SERVICE.

Character	*Symbol*
A	•–
B	–•••
C	–•–•
D	–••
E	•
F	••–•
G	––•
H	••••
I	••
J	•–––
K	–•–
L	•–••
M	––
N	–•
O	–––
P	•––•
Q	––•–
R	•–•
S	•••
T	–
U	••–
V	•••–
W	•––
X	–••–
Y	–•––
Z	––••
0	–––––
1	•––––
2	••–––
3	•••––
4	••••–
5	•••••
6	–••••
7	––•••
8	–––••
9	––––•
Period	•–•–•–
Comma	––••––
Query	••––••
Slash	–••–•
Dash	–••••–
Break (pause)	–•••–
Semicolon	–•–•–•
Colon	–––•••

Continental code

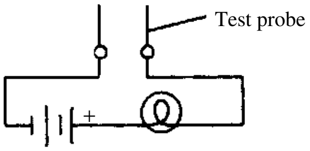

continuity tester

continuous duty The requirement of a device to sustain a 100-percent duty cycle for a prolonged period of time.

continuous-duty rating A maximum current, voltage, or power rating for equipment operated for extended periods at a 100-percent duty cycle.

continuous load A load that requires a continuous feed for a prolonged period of time.

continuous memory See NONVOLATILE MEMORY.

continuous-path motion In robotics, machine movement that occurs in a smooth fashion, rather than in discrete steps. Allows precise positioning of a mechanical arm or gripper.

continuous power The maximum sine-wave power that an amplifier can deliver for 30 seconds.

continuous recorder An instrument that provides an uninterrupted recording.

continuous recording A record made on a continuous sheet or tape, instead of on separate sheets or tapes. An example is a continuous-playing tape used for repeated public announcements.

continuous spectrum **1.** The range of all electromagnetic frequencies between a specified lower limit f_1 and a specified upper limit f_2. **2.** A range of electromagnetic frequencies that exhibits similar behavior between its lower and upper limits.

continuous stationery Also called *fanfold paper.* The pack of paper a line printer uses. It consists of sheets connected by perforated or tear-off edges, folded in accordion fashion. It usually has tear-off perforated strips along either side to facilitate feed through the printer mechanism.

continuous variable A variable that can attain any value within a specific range of values. An example is a frequency within the 75- to 80-meter amateur radio band, from 3.5 to 4.0 MHz.

continuous wave Abbreviation, CW. **1.** A periodic wave, such as a radio-frequency (RF) carrier, that is not interrupted at any point between its normal start and termination, and that is unmodulated. **2.** An RF carrier that is interrupted digitally with a keying device according to some code (such as Morse), for the purpose of conveying information.

continuous-wave laser See CW LASER.

continuous-wave radar See CW RADAR.

contour A control on an audio reproduction system that increases the base and treble amplitudes at low levels to compensate for the ear's natural losses in these ranges. Alternatively, this control can attenuate signals in the 3-kHz region, where the human ear is most sensitive.

contours of equal loudness See AUDIBILITY CURVES.

CONTRAN A computer language that requires no compiler, or translating, interface between the operator and the machine. The programming is done in a language similar to machine language.

contrast **1.** In a video image, the degree to which adjacent areas of a picture are differentiated. Insufficient contrast makes for a "flat" picture; ex-

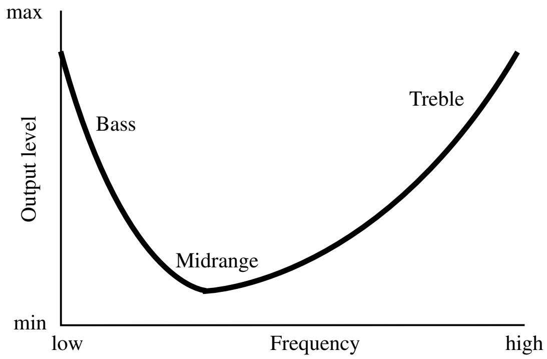

contour

cessive contrast, a "hard" picture. **2.** In optical character recognition, the degree to which a character is distinguishable from its background.

contrast control A potentiometer for adjusting the gain of the video in a television receiver or cathode-ray-tube (CRT) computer display and, accordingly, the image contrast.

contrast range In an image or pattern, the brightness range from the lightest to the darkest parts.

contrast ratio In a video image, the ratio of maximum to minimum luminance.

control **1.** An adjustable component, such as a rheostat, potentiometer, variable capacitor, or variable inductor, that allows some quantity to be varied at will. **2.** A test or experiment conducted simultaneously with another similar test conducted under conditions lacking the factor under consideration. Thus, if 100 resistors coated with a special varnish are tested at 120°F, 100 identical unvarnished resistors could be tested (as a control) under the same conditions; in this way, the effect of the varnish would be ascertainable. **3.** As a computer function, understanding and implementing instructions or carrying out tasks, according to specific conditions.

control ampere-turns The ampere-turns of the control winding in a magnetic amplifier.

control block A storage block for control information in a computer.

control bus In a digital computer, the electrical conductors linking the central-processing-unit (CPU) control register to the memory circuits.

control card A card that provides control information for a computer.

control character A character (bit group) used to start the control of a peripheral.

control characteristic A representation (such as a collector-current versus collector-voltage curve) depicting the extent to which the value of one quantity affects or controls the value of another.

control circuit **1.** A circuit in which one signal or process is made to control another signal or process. **2.** In a digital computer, a circuit that handles and interprets instructions and commands, particularly in the arithmetic and logic unit (ALU).

control computer A computer that receives signals concerning the parameters in some process, and responds with signals that control those parameters.

control counter See CONTROL REGISTER.

control data **1.** In a computer record having a key, information used to put the records in some sequence. **2.** Information affecting a routine's selection or modification.

control electrode An electrode to which an input signal can be applied to control an output signal. Common examples are the base of a bipolar transistor, the gate of a field-effect transistor, and the inputs of a logic gate.

control field **1.** In direct-current generators of the amplifying type, an auxiliary field winding used for feedback and regulation, in contrast to the self-excited field winding (which is the conventional field winding of the generator). **2.** A computer record field containing control data.

control flux In an amplidyne, magnetic flux generated by current flowing through the control winding.

control grid See GRID, **1**.

control-grid bias The negative dc voltage applied between ground and the control grid of a vacuum tube to establish the operating point.

control language Within the operating system of a computer, the command set that the operator or programmer uses to control the running of a program or the operation of peripherals. Also called *job control language* or *system control language.*

control language interpreter See CONTROL LANGUAGE and INTERPRETER.

controlled avalanche diode Also called *avalanche diode* or *Zener diode.* A diode that has a well-defined avalanche voltage. Used primarily for voltage regulation in power supplies.

controlled-carrier modulation See QUIESCENT CARRIER OPERATION.

controlled-carrier transmission See QUIESCENT CARRIER OPERATION.

controlled rectifier A rectifier whose dc output can be varied by adjusting the voltage or phase of a signal applied to the control element. See SILICON-CONTROLLED RECTIFIER.

controller **1.** The control signal of an electronic control (or servo), system. **2.** A device, such as a specialized variable resistor, used to adjust current or voltage. **3.** A computer that oversees and controls the operation of a robot or fleet of robots.

controller function The control of the movements of a servo system.

controlling file A computer storage area encompassing several complete magnetic disk cylinders; its size can be changed to accommodate a number of files.

control loop See CONTROL TAPE.

control mark See TAPE MARK.

control panel **1.** An accessible surface on which are mounted switches, buttons, potentiometers, meters, digital indicators, monitoring devices, and other apparatus essential to regulating and supervising an electronic system. **2.** The console that a computer operator or programmer uses to communicate with the central processing unit (CPU).

control plate The metallic plate or disk that serves as the antenna of a CAPACITANCE RELAY or TOUCHPLATE RELAY.

control program A program that arranges computer-operation programs in a certain order. Puts information in the computer memory for later use.

control rectifier A semiconductor diode device, used for the purpose of switching large currents. A small control signal can provide switching of high-power devices.

control register In a computer, the register that stores the address of the next instruction in the program being run.

control sequence The order in which instructions are executed in a digital computer.

control stack In a computer system, a unit of hardware having storage locations and used to perform arithmetic, assist in allocating memory to programs, and to control internal processes.

control statement In a programming language, an instruction that causes some action to be taken, as specified by a condition; it is also applicable to source program statements that affect the compiler's operation without modifying the machine code.

control tape Punched paper or plastic tape in the form of a closed loop and used to control printing devices. Also called *control loop.*

control total For a file or record group, a total derived during an operation; it is used to verify that all the records have been processed similarly.

control transfer The situation in which the control unit of a digital computer leaves the main sequence of instructions and takes its next instruction from an out-of-sequence address.

control transfer instruction See BRANCH INSTRUCTION.

control-voltage winding In a servomotor, the winding that receives a varying voltage of a phase different from that applied to the fixed-voltage windings.

control winding In a magnetic amplifier, the winding that conducts the control-signal current.

control word A word (a bit group) stored in a computer memory and used for a control function.

convection The flow of a gas or liquid that results in the transfer of heat from one location to another.

convection cooling The removal of excess heat from a component, such as a power vacuum tube or transistor, via upward movement of surrounding air that has been heated by the component.

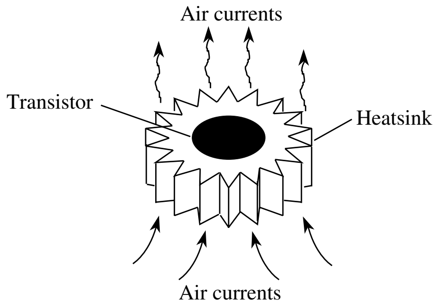

convection cooling

convection current **1.** The motion of current carriers or a charge across the surface of a conductor or dielectric. **2.** Air currents rising above a heat source or heated body.

convective discharge The continuous high-voltage current discharge across a spark gap.

convectron A device that indicates the angle, with respect to the vertical, based on convection cooling of a straight wire. The temperature difference is greatest when the angle is 0 degrees (the wire is vertical); the temperature difference decreases as the angle increases, reaching a minimum at 90 degrees (when the wire is horizontal).

convenience outlet **1.** In North America, a wall outlet providing a nominal 117 volts alternating current (ac) at 60 Hz for common household appliances. **2.** An outlet in a laboratory that provides power for a certain application.

conventional current The notion that current flows from the positive pole to the negative pole in an electric circuit. This representation is used most often by physicists. Electron flow is opposite to conventional current flow; positively charged particles, such as holes, move in the same direction as the conventional current.

convergence **1.** The eventual meeting of values or bodies at some point (sometimes at infinity, as in certain mathematical series). **2.** The intersection point of the beams from separate electron guns in a cathode-ray tube (CRT).

convergence coil One of a pair of coils used in a color television receiver to produce dynamic beam convergence (see CONVERGENCE, **2**).

convergence control In a color television receiver, a potentiometer in the high-voltage circuit for convergence adjustment (see CONVERGENCE, **2**).

convergence electrode An electrode that provides an electrostatic field for converging electron beams. Compare CONVERGENCE MAGNET.

convergence frequency The frequency of the last member of a spectrum series.

convergence magnet An assembly that provides a magnetic field to converge electron beams. Compare CONVERGENCE ELECTRODE.

convergence phase control In a three-gun color picture tube, a variable resistor or variable inductor used to adjust the phase of the dynamic convergence voltage.

convergence plane **1.** In a color picture tube, the plane in which the red, green, and blue beams all focus. **2.** In a cathode-ray tube, the plane in which the electron beam reaches its sharpest focus.

convergent series A mathematical series that approaches a specific, finite numerical value as the number of terms increases. Thus, the series 0.3 + 0.03 + 0.003 + . . . approaches a limiting value of 1⁄3. Compare DIVERGENT SERIES and INFINITE SERIES.

converging lens A lens having a real focus for parallel rays; generally a *convex lens.*

conversational compiler In computer operations, a compiler that, using the CONVERSATIONAL MODE of operation, shows the programmer whether or not each statement entered into the computer is valid, and whether or not to proceed with the next instruction.

conversational mode High-level computer operation or programming, in which the computer gives responses to the operator's input.

conversion **1.** The deliberate mixing of radio-frequency (RF) signals to produce signals at the sum and/or difference frequencies. **2.** The process of changing direct current (dc) to alternating current (ac). **3.** The process of changing low-voltage dc to high-voltage dc. **4.** The changing of a computer file to another format and, possibly, transferring it to a different storage medium (e.g., from tape to internal memory). **5.** The processing of a program or file written for one computer or application into a form suitable for another computer or application.

conversion efficiency In a converter (see CONVERTER, **1**), the ratio of output-signal amplitude to input-signal amplitude. For example, in a superheterodyne converter, a large intermediate-frequency (IF) output for a low radio-frequency (RF) input indicates high conversion efficiency.

conversion equipment In a computer system, an offline device for transferring data from one medium to another [e.g., a disk-to-tape converter (tape drive)]. Also called CONVERTER.

conversion exciter An exciter for transmitters, in which an output signal of a desired frequency is obtained by beating the output of a variable-frequency self-excited oscillator with the output of a fixed-frequency oscillator (such as a crystal oscillator).

conversion gain Amplification as a byproduct of conversion. See CONVERSION EFFICIENCY.

conversion loss Conversion gain of less than 1.

conversion program In computer operations, a program for data conversion (see CONVERSION, **4** and **5**).

conversion rate Also called *sampling rate.* The number of samples per second taken by an ANALOG-TO-DIGITAL CONVERTER.

conversion time In digital computer operation, the time required for the machine to read out all the digits in a coded word.

conversion transconductance See CONVERSION EFFICIENCY.

convert **1.** To perform frequency conversion (see CONVERSION, **1**). **2.** To perform voltage conversion (see CONVERSION, **2** and **3**). **3.** In computer operations, to change information from one number base to another. **4.** To perform data conversion (see CONVERSION, **4** and **5**).

converter **1.** A heterodyne mixer in which two input signals of different frequency are mixed to yield a third (output) signal of yet a different frequency. **2.** A machine for converting direct current (dc) to alternating current (ac) (e.g., a chopper converter). **3.** A transistor circuit for converting a low-voltage dc to higher-voltage dc. **4.** Conversion equipment. **5.** A circuit or device that changes analog data to digital data or vice versa.

converter amplifier See CHOPPER AMPLIFIER.

converter stage A circuit used principally to mix two signals (such as a received signal and local-oscillator signal in a superheterodyne receiver), and deliver the resultant signal.

convexo-concave Pertaining to a lens having a convex face of greater curvature than its concave face.

coolant A liquid (often water or oil) used to remove heat from an electronic component.

Coolidge X-ray tube An X-ray tube containing a heated filament (with focusing shield) and a slanting tungsten target embedded in a heavy copper anode.

cooling Maintenance of the operating temperature of an electronic component or system at a safe level. Common devices for cooling are heatsinks, circulating or forced air, and circulating liquid.

coordinate bond A covalent bond that consists of a pair of electrons supplied by only one of the atoms joined by the bond.

Coordinated Universal Time Abbreviation, UTC. Astronomical time at the Greenwich meridian (zero degrees longitude). The UTC day begins at 0000 hours and ends at 2400 hours. Based on the mean, or average, synodic (sun-based) rotational period of the earth. The earth is slightly behind UTC near June 1, and is slightly ahead near October 1.

coordinate digitizer A device or circuit that encodes a coordinate graph into digital signals for storage or transmission.

coordinate of chromaticity See CHROMATICITY COORDINATE.

coordinates A set of axes with points that can be uniquely defined or located on a line, in a plane, or in space. See CARTESIAN COORDINATES and POLAR COORDINATES.

coordinate system A mathematical means of uniquely defining or locating a point on a line, in a plane, or in space. The most common coordinates are CARTESIAN COORDINATES (also called *rectangular coordinates*), consisting of numbered lines intersecting at right angles.

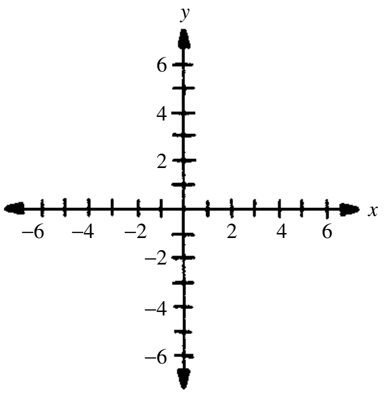

coordinate system (Cartesian)

coordination complex An ion or compound having a central (usually metallic) ion combined by coordinate bonds with a definite number of surrounding groups, ions, or molecules.

coplanar array A set of antennas that lie in the same plane, and are fed by a common transmission line.

copper Symbol, Cu. A metallic element. Atomic number, 29. Atomic weight, 63.546. An excellent conductor of electricity and heat, commonly used in the manufacture of wires and cables.

copper-clad wire Iron or steel wire plated with copper.

copper-constantan thermocouple A thermocouple consisting of a junction between wires or strips of copper and constantan. Typical output is 4.24 mV at 100°C.

copper loss Power (I^2R) loss in copper wires, cables, and/or coils.

copper-oxide diode A small diode in which the semiconductor material is copper oxide. Such diodes, widely used before the ready availability of selenium and silicon, are still occasionally found in meter-rectifier service.

copper-oxide modulator An amplitude modulator whose action is derived from the nonlinear conduction characteristic of copper-oxide diodes.

copper-oxide photocell A photoelectric cell in which the light-sensitive material is copper oxide.

copper-oxide rectifier A rectifier in which the semiconductor material is copper oxide. Rectifiers of this type are suitable for low-voltage service; they were widely used before the advent of germanium, silicon, and selenium rectifiers.

copper pyrites See CHALCOPYRITE.

copper-sulfide rectifier A rectifier in which the unilateral junction is between copper-sulfide and magnesium elements. Like the copper-oxide rectifier, the copper-sulfide unit was once widely used in low-voltage applications.

copy **1.** Also called *hard copy*. Printed or written text. **2.** In communications, a qualitative expression of the extent to which received data is intelligible (e.g., a radio operator's signal report, "You are solid (perfect) copy."). **3.** To duplicate data in a storage system, the original being in another system, or in a different location in the same system. **4.** An exact duplicate of data in any form.

copying telegraph A descriptive term for a *facsimile* system.

Corbino disk A variable resistor consisting of a semiconductor disk capable of exhibiting the CORBINO EFFECT. The disk is inserted into an adjustable magnetic field, which serves as the control medium.

Corbino effect A phenomenon similar to the HALL EFFECT, in which a current flows around a disk carrying a radial current when the disk is inserted into a magnetic field whose lines of flux are perpendicular to the disk. Compare HALL EFFECT.

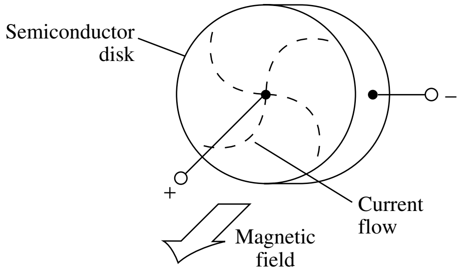

Corbino effect

cord **1.** A length of flexible, insulated cable, usually having two or three conductors. **2.** Tough, insulating string (e.g., *dial cord* or *lacing cord*).

cordless **1.** Descriptive of a plug without a flexible cord. **2.** Pertaining to radio-frequency (RF) or infrared short-range links for communications and control (e.g., a *cordless telephone* set).

cordless keyboard A computer keyboard that employs an infrared (IR), very-high frequency (VHF), or ultra-high-frequency (UHF) transmitter and receiver. Commonly used with so-called Web TV systems and in presentations using a display projection system. Operates according to the same electronic scheme as a CORDLESS MOUSE.

cordless modem See WIRELESS MODEM, 3.

cordless mouse A hand-controlled computer mouse that employs an infrared (IR), a very-high frequency (VHF), or an ultra-high-frequency (UHF) transmitter and receiver. The transmitter is inside the device, and the receiver is contained either inside the computer main unit, or in a small box attached to the computer main unit by a cord. The box can be placed somewhere out of the way; for example, at the back of the desk. Then the mouse can be moved around freely. This link is effective at distances of up to 20 or 30 feet.

cordwood A type of construction in which electronic components are sandwiched perpendicularly between layers of components. So called because it looks somewhat like stacked cordwood.

cordwood module A module containing discrete components mounted perpendicularly between two parallel printed circuits.

core **1.** The body or form on which a coil or transformer is wound. Can be made of ferromagnetic or dielectric material. The properties depend on the application. **2.** CORE MEMORY.

core dump Dumping core memory content to an output peripheral. Also see DUMP.

coreless induction heater An induction heater in which the body to be heated receives energy directly from the field of the energizing coil (there is no intervening core). Compare CORE-TYPE INDUCTION HEATER.

core loss Loss of energy in a magnetic core, caused by eddy currents and hysteresis in the core material.

core memory An older memory technology, consisting of a series of small ringshaped magnetic cores, into or out of which data can be written or read by changing the magnetization of the cores.

core plane A usually flat assembly of special magnetic cores, through which pass associated current-conducting wires to provide a CORE MEMORY.

core saturation The condition in which a core of magnetic material accommodates the maximum number of magnetic lines characteristic of that material. Increasing the magnetizing force produces no additional magnetization.

core shift register A shift register that uses special magnetic cores as bistable components. See CORE MEMORY.

core storage A high-speed magnetic core storage unit. Also see CORE MEMORY and CORE PLANE.

core transformer A transformer whose coils are wound around a ferromagnetic core.

core wrapping The placing of an insulating layer over an inductor or transformer core. This minimizes the chances of short-circuiting between the windings and the core material.

core-type induction heater An induction heater in which the body to be heated is magnetically linked, by a core, to the energizing coil. Compare CORELESS INDUCTION HEATER.

corner **1.** An abrupt turn in the axis of a waveguide. **2.** The line, and the region in the vicinity thereof, at which two intersecting plane surfaces meet (e.g., the reflector screen of a CORNER-REFLECTION ANTENNA). The plane surfaces are usually perpendicular to each other. **3.** The point, and the region in the vicinity thereof, at which three intersecting plane surfaces meet. Generally, the plane surfaces are mutually perpendicular. **4.** The passband frequency limit(s) of a bandpass, band-rejection, high-pass, or low-pass filter. **5.** A sharp bend in the attenuation-versus-frequency curve of a bandpass, band-rejection, high-pass, or low-pass filter, depicting the limit(s) of the passband.

corner diffraction **1.** The bending of sound waves around a corner. **2.** The bending of radio-frequency (RF) energy around an object, when the wavelength is great, compared with the size of the object.

corner effect A rounding off of the frequency response of a filter at the corner(s) [i.e., at the limit(s) of the passband].

corner frequency See CORNER, **4**.

corner reflection The reflection of a beam of light (or of microwave energy or other short-wavelength energy) from a corner reflector, so the beam leaves the reflector in exactly the opposite direction from which it approaches. See CORNER REFLECTOR, **2**.

corner-reflection antenna A directional antenna consisting of a dipole radiator situated at the apex formed by two nonparallel, flat reflecting sheets or a single folded sheet. See CORNER REFLECTOR, **1**.

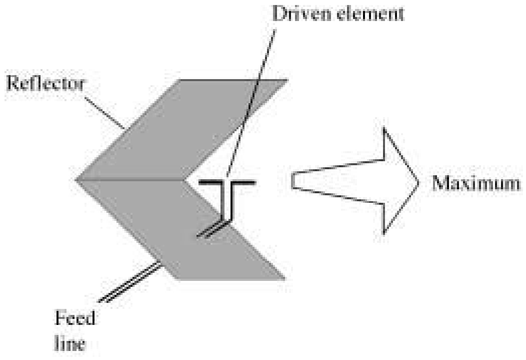

corner-reflection antenna

corner reflector **1.** An antenna with a half-wave driven element and a reflector made of wire mesh, screen, or sheet metal that resembles an open folder. The flare angle of the reflecting element is about 90 degrees. The antenna is used at ultra-

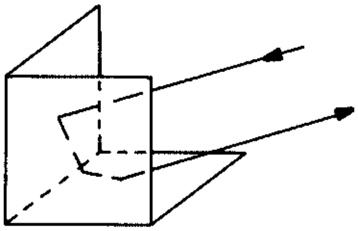

corner reflector, 2

high and microwave frequencies for television reception and satellite communications. Sometimes several half-wave dipoles are fed in phase and arranged along a common line with a single, elongated reflector. **2.** Also called *tricorner reflector*. A set of three flat metal surfaces or screens, attached together in a manner identical to the way two walls meet the floor or ceiling in a room. Such a device, if it is at least several wavelengths across, returns electromagnetic energy in exactly the same direction from which it arrives. Devices of this type are used as radar dummy targets and in optical and infrared (IR) wireless ranging systems.

corona A luminous discharge in the space surrounding a high-voltage conductor; caused by ionization of the air. The discharge constitutes a loss of energy.

corona effect The production of a luminous discharge, especially at the end of a pointed terminal, when the voltage gradient reaches a critical value.

corona failure A form of high-voltage failure, resulting from the erosion of an object (such as an electrical insulator) by corona.

corona loss Loss caused by energy dissipation through a corona. It occurs as a result of the emission of electrons from the surface of electrical conductors at high potentials, and depends on the curvature of the conductor surface, with most emission occurring from sharp points and the least from surfaces with a large radius of curvature. It is often accompanied by a blue glow and a crackling or hissing sound.

corona resistance The length of time that an insulating material can withstand a specified level of field-intensified ionization before completely breaking down.

corona shield A shield surrounding a high-voltage point to prevent corona by redistributing the electric flux.

corona starting voltage The minimum voltage between two electrodes, or on a single electrode in free space, at which corona occurs.

corona voltmeter A voltmeter used to measure the peak value of a voltage in terms of corona discharge. It consists of a metal tube in which a central wire is mounted, the parts being connected to the voltage source. The air density in the tube is varied until corona occurs.

corpuscle A tiny particle. It was the name given to the ELECTRON by some early experimenters and theorists.

correction **1.** The addition of a factor that provides greater accuracy in a measurement. **2.** A change in the calibration of an instrument to increase the accuracy.

correction factor A percentage, or numerical factor, added to or subtracted from a reading to provide a greater degree of accuracy. Often used with instruments known to be inaccurate by a certain amount.

corrective feedback Feedback that is used to correct (bring to a prescribed level) a quantity constituting the input to a system.

corrective maintenance The repair of a circuit or system after it has malfunctioned or broken down.

corrective network A network that improves the performance of the circuit into which it is inserted.

corrective stub A combination tuning-matching stub used in some antenna systems. It matches the resistive component of the antenna impedance to the characteristic impedance of a feed line, and also eliminates any reactance that might be present at the antenna feed point.

correed relay A sealed reed relay used as a high-speed switching device in communications equipment.

correlation A statistical expression or measure of the degree to which two sets of data are related. Can be given qualitatively (high-positive, low-positive, zero, low-negative, or high-negative) or quantitatively (as a number between –1 and 1). Does not necessarily imply causation.

correlation detector A detector that compares a signal of interest with a standard signal at every point, delivering an output that is proportional to the correspondence between the two signals.

correlation distance The smallest distance between two antennas that results in fading of signals under conditions of tropospheric propagation. It is used at very-high frequencies (VHF) and above, to determine the maximum range over which communications can be carried out reliably.

correlation tracking A method of target tracking in which phase relationships are used to determine positions.

correspondence The ability of a binocular machine vision system to tell when both of its optical sensors are processing an image from the same object; also, the ability of the system to keep both sensors tracking the same object.

corrosion-resistant Pertaining to materials that are treated to be immune to corrosion by the elements. Such substances are preferable for use in marine or tropical environments, where corrosion is especially severe.

corruption The altering of data or a code as a result of a program error or machine fault.

COS Abbreviation of COMPLEMENTARY-SYMMETRY CIRCUIT.

cosecant Abbreviation, csc. A trigonometric function; csc $q = c/a$, where c is the hypotenuse of a right triangle and a is the side opposite q. The cosecant is the reciprocal of sine: csc $q = 1/\sin q$.

cosecant-squared antenna A radar antenna that radiates a COSECANT-SQUARED BEAM.

cosecant-squared beam A radar beam whose intensity varies directly with the square of the cosecant of the angle of elevation.

cosech Abbreviation of HYPERBOLIC COSECANT. Also abbreviated as csch.

cosh Abbreviation of HYPERBOLIC COSINE.

cosine Abbreviation, cos. A trigonometric function; cos $q = b/c$, where b is the side adjacent to q and c is the hypotenuse of the right triangle.

cosine law The brightness in any direction from a perfectly diffusing surface is proportional to the cosine of the angle between the direction vector and a vector perpendicular to the surface.

cosine wave A periodic wave that follows the cosine of the phase angle. It has a shape identical with a SINE WAVE, but differs by 90 degrees of phase.

cosine yoke A magnetic-deflection yoke that has nonuniform windings for improved focus at the edges of a television picture. Also called *anastigmatic yoke* and *full-focus yoke.*

cosmic noise Radio noise produced by signals from extraterrestrial space.

cosmic rays Extremely penetrating rays consisting of streams of atomic nuclei entering the earth's atmosphere from outer space.

COS/MOS IC An integrated circuit (IC), such as an operational amplifier, utilizing metal-oxide-semiconductor (MOS) field-effect transistors in a complementary-symmetry (COS) arrangement.

cost analysis In a commercial or industrial organization, ascertaining the expense associated with a service, process, or job.

cot Abbreviation of COTANGENT.

cotangent Abbreviation, cot. A trigonometric function; cot $q = b/a$, where a is the side adjacent to q and b is the side opposite q (in a right triangle). Cotangent is the reciprocal of tangent: cot $q = 1/\tan q$.

coth Abbreviation of HYPERBOLIC COTANGENT.

Cotton-Mouton effect See KERR MAGNETO-OPTICAL EFFECT.

Cottrell process Dust precipitated by high voltage. Dust in the air is made to flow through a grounded metal chamber that contains a wire maintained at high voltage. The dust particles become charged and adhere to the chamber walls, from which they are later collected.

coul-cell A coulometer of the electrolytic-cell type.

coulomb (Charles Augustin Coulomb, 1736–1806). Abbreviation, C. The unit of electrical charge quantity, equal to the charge contained in 6.24×10^{18} electrons. A current of one ampere (1 A) represents 1 coulomb per second (C/s).

Coulomb's law The force between two electrically charged objects is directly proportional to the product of the charge quantities in coulombs, and inversely proportional to the square of the distance between the charge centers. This force is an attraction for opposite charges, and a repulsion for similar charges.

coulometer An instrument that measures electrical charge quantity in coulombs. A typical version keeps a cumulative count of coulombs (ampere-seconds) by integrating current, with respect to time. Also called *coulombmeter.*

Coulter counter See CELL COUNTER.

count **1.** The number of pulses tallied by a counting system in a given period of time. **2.** A single response by a radioactivity counter. **3.** A record of the number of times an instruction or subroutine in a computer program is executed (by increasing the value of a variable by one, as stated in a FOR-NEXT loop, for example).

countdown A decreasing count of time units remaining before an event or operation occurs showing time elapsed and time remaining.

counter **1.** A circuit, such as a cascade of flip-flops, that tracks the number of pulses applied to it and usually displays the total number of pulses. **2.** A mechanism, such as an electromechanical indicator, that tracks the number of impulses applied to it and displays the total. **3.** An electronic switching circuit, such as a flip-flop or stepping circuit, that responds to sequential input pulses applied to it, giving one output pulse after receiving a certain number of input pulses.

counter- Prefix meaning "opposite to" or "contrary to." Examples: *counter EMF, counterclockwise.*

counterclockwise Abbreviation, ccw. Pertaining to rotational motion in a sense opposite that of a typical analog clock. Movement is to the left at the top of the rotational circle, and to the right at the bottom of the circle. Compare CLOCKWISE.

counterclockwise-polarized wave An elliptically polarized electromagnetic wave whose electric-intensity vector rotates counterclockwise as observed from the point of propagation. Compare CLOCKWISE-POLARIZED WAVE.

counter efficiency The sensitivity of a radiation counter or scintillation counter to incident X-rays or gamma rays.

counterelectromotive cell A cell used to counteract a direct-current voltage.

counter EMF See BACK VOLTAGE and KICKBACK.

counter-meter A radioactivity instrument, such as a Geiger counter, that indicates the number of radioactive particles per unit time.

counterpoise A means of obtaining a radio-frequency (RF) ground by using a grid of wires or tubing in a plane parallel to the earth's surface or to average terrain. The radius of the grid is usually at least 0.25 wavelength, but might be smaller if the feed-point impedance of the antenna is very high.

counterpoise ground system A counterpoise with a radius such that resonance is obtained with a quarter-wavelength antenna operated at a height of more than 0.25 wavelength above actual ground. Usually such a system consists of three or four radials measuring 0.25 wavelength each, and extending outward from the base of the antenna nearly parallel to the average terrain.

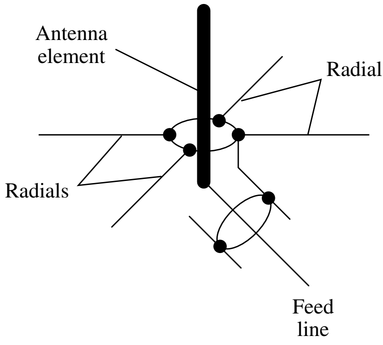

counterpoise ground system

counter tube A tube, such as the Geiger-Meuller tube, in which a penetrating radioactive particle ionizes a gas and produces an output pulse.

counter voltage See BACK VOLTAGE and KICKBACK.

counting-type frequency meter A direct-reading analog or digital frequency meter that indicates the number of pulses (or cycles) per second applied to it.

count-remaining technique See COMPLEMENT-SETTING TECHNIQUE.

couple Two dissimilar metals in contact with each other or immersed in an electrolyte.

coupled circuits Circuits between which energy is transferred electrostatically, electromagnetically, by some combination of the two, or by direct connection.

coupled impedance The impedance that a circuit “sees” when it is coupled to another circuit. Thus, when the secondary of a transformer is terminated with an impedance, the primary “sees” a combination of that impedance and its own.

coupler A device for transferring energy between two circuits and using capacitive coupling, direct coupling, inductive coupling, or some combination of these.

coupling **1.** Also called *electrostatic coupling* or *capacitive coupling.* The linking of two circuits or devices by electric flux. **2.** Also called *magnetic coupling* or *inductive coupling.* The linking of two circuits or devices by magnetic flux. **3.** Also called *direct coupling.* The linking of two circuits or devices by direct connection. **4.** Also called *resistive coupling.* The linking of two circuits or devices through a resistance. **5.** Also called *optical coupling.* The linking of two circuits or devices through an optoisolator.

coupling aperture A hole in a waveguide that is used to transmit energy to the waveguide, or receiving energy from outside the waveguide.

coupling capacitor A capacitor used to conduct ac energy from one circuit to another. Also see CAPACITIVE COUPLING.

coupling coefficient See COEFFICIENT OF COUPLING.

coupling diode A semiconductor diode connected between the stages of a direct-coupled amplifier. When the diode is connected in the correct polarity, it acts as a high resistance between stages when there is no signal, and does not pass the high dc operating voltage from one stage to the next. When a signal is present, the diode resistance decreases, and the signal gets through.

coupling efficiency A measure of the effectiveness of a coupling system (i.e., the degree to which it delivers an undistorted signal of correct amplitude and phase).

coupling loop **1.** A single turn of a coupling transformer. **2.** A small loop inserted into a waveguide to introduce microwave energy.

coupling probe A usually short, straight wire or pin protruding into a waveguide to electrostatically introduce microwave energy into the waveguide. It acts like a miniature whip antenna.

coupling transformer A transformer used primarily to transfer alternating-current (ac) energy electromagnetically into or out of a circuit.

covalent binding forces In a crystal, the binding forces resulting from the sharing of valence electrons by neighboring atoms.

covalent bonding The binding together of the atoms of a material as a result of shared electrons or holes.

coverage **1.** The area within which a broadcast or communication station can be reliably heard. **2.** The shielding effectiveness of a coaxial cable.

coversed sine Abbreviation, covers. The trigonometric functional equivalent of the *versed sine* of the complement of an angle [i.e., the difference between the sine of an angle and unity (1)]. Thus, covers $q = 1 - \sin q$.

CP Abbreviation of *chemically pure.*

cp **1.** Abbreviation of CANDLE POWER. **2.** Abbreviation of *central processor.*

cps **1.** Abbreviation of CYCLES PER SECOND. Cycles per second, to denote ac frequency, has been supplanted by HERTZ. **2.** Abbreviation of *characters per second.*

CPU Abbreviation of CENTRAL PROCESSING UNIT.

CQ A general call signal used in radio communication, especially by amateur stations, to invite a response from any station that hears it.

Cr Symbol for CHROMIUM.

cracked-carbon resistor A high-stability resistor in which the resistance material is particulate carbon.

cracker A hacker with malicious intent (also see HACKER). Such a person attempts to gain access to computer systems or databases in order to steal something or inflict damage. Examples include theft, erasure, or mutilation of data; fraudulent debiting of bank accounts; alteration of credit information; and identity theft.

cradle guard See GUARD WIRE.

cradlephone A telephone set in which the microphone and earphone are mounted on opposite ends of a handle. This handle, called the *receiver*, rests on the crossmember of a stand connected to a base containing the dial and ringing circuits. Also called *French phone, French telephone*, and *handset.*

crash **1.** A condition in which a computer or network server becomes inoperative because of a software or memory-management problem. **2.** In a computer hard disk or diskette drive, contact of the read/write head with the surface of a disk or platter. Usually, it is the result of excessive physical vibration or shock.

crate A foundation unit into which modules are plugged to establish a circuit.

crawl **1.** See CREEPING COMPONENT. **2.** The credits (names of staff and their contribution to content) superimposed and moving on a television picture at the end of a program.

crazing The formation of tiny cracks in materials, particularly in such dielectrics as plastic and ceramic.

creep See COLD FLOW.

creepage Current leakage across the surface of a dielectric.

creeping component A quantity, such as current, voltage, or frequency, that slowly changes in value with time.

crest factor See AMPLITUDE FACTOR.

crest value The maximum amplitude of a composite current or voltage.

crest voltmeter A peak-reading (or sometimes peak-responsive) voltmeter.

crippled mode The mode of operation for a computer or other hardware in which some of the components are inoperable. Compare GRACEFUL DEGRADATION.

crisscross neutralization See CROSS-CONNECTED NEUTRALIZATION.

crisscross rectifier circuit A conventional bridge rectifier circuit configured in such a way that two of the diodes are connected in crisscross fashion between the input and output terminals.

critical angle **1.** In radio communications, an angle of departure that a transmitted electromagnetic field subtends, with respect to the horizon at the transmitting (TX) point, below which the ionosphere will reliably return the signal to the earth, and above which the ionosphere will not reliably return the signal. This angle (shown by the double arc marked *X* in the drawing) depends on the frequency of the transmitted electromagnetic wave, and also on ionospheric conditions. **2.** For an electromagnetic wave or ray approaching a boundary at which the index of refraction abruptly decreases, the minimum angle of incidence (relative to a line perpendicular to a plane tangent to the boundary) at which the energy is totally reflected.

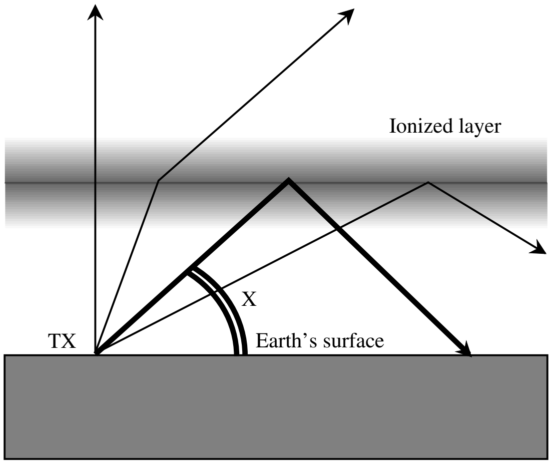

critical angle, 1

critical characteristic A parameter that has a disproportionate effect on other variables. A small change in this characteristic can result in a large change in the operating conditions of a circuit or system.

critical component A component or part that is especially important in the operation of a circuit or system.

critical coupling The value of coupling at which maximum power transfer occurs. Increasing the extent of coupling beyond the critical value decreases power transfer.

critical damping The value of damping that yields the fastest transient response without overshoot.

critical dimension The cross-sectional size of a waveguide that determines its minimum usable frequency.

critical failure A component or circuit failure that results in shutdown of a system, or a malfunction that results in improper operation.

critical field The smallest magnetic-field intensity in a magnetron that keeps an electron, emitted from the cathode, from reaching the anode.

critical frequency For a particular layer of the ionosphere, the high frequency at which a vertically propagated wave is no longer reflected back to the earth.

critical inductance In a choke-input power-supply filter, the minimum inductance that will maintain a steady value of average load current.

critical potential The potential difference required for an electron to excite or ionize an atom with which it collides.

critical voltage The voltage at which a gas ionizes.

critical wavelength The wavelength that corresponds to CRITICAL FREQUENCY.

CRO Abbreviation of *cathode-ray oscilloscope.*

Crookes dark space In a glow-discharge tube, the narrow dark space next to the cathode. Also see CROOKES TUBE.

Crookes tube A glow-discharge tube containing an anode, cathode, and a small amount of gas under low pressure.

cross antenna An antenna in which two (usually equal-length) horizontal radiators cross each other at right angles and are connected together to a feeder at their point of intersection. It takes its name from its horizontal-cross shape.

cross assembler A program used with one computer to translate instructions for another computer.

crossband operation **1.** Communications in which two frequency bands are used. Station X, for example, might transmit on frequency f_A in band A and receive on frequency f_B in band B; station Y would then transmit on f_B and receive on f_A. **2.** In satellite communications, the use of two frequency bands to facilitate full-duplex operation and to allow the satellite transponder to effectively function. The transponder receives signals from the earth within a specific frequency band, and converts this entire band of signals to a set of signals that occupies an equal amount of spectrum space on another frequency band. The converted signals are then retransmitted back to earth.

crossbar switch A three-dimensional array of switch contacts in which a magnetic selector chooses individual contacts, according to their coordinates in the matrix.

cross bearings A method of radionavigation, in which directional readings are taken from a receiving station (such as a ship or aircraft) for two fixed transmitting stations whose locations are known. Lines are drawn on a map from the transmitting stations, in directions 180 degrees opposite the bearings obtained from the receiving station. The intersection point of these lines is the location of the receiving station.

cross beat A spurious frequency arising from CROSS MODULATION.

cross-check To compare the result of a calculation or computer routine with the result obtained by a different method.

cross color In the chrominance channel of a color television receiver, crosstalk interference caused by monochrome signals.

cross-connected neutralization Neutralization of a push-pull amplifier by feedback through two capacitors—each connected from the output circuit of one transistor to the input circuit of the other.

cross-coupled multivibrator A multivibrator circuit in which feedback is provided by a coupling capacitor between the output of the second stage and the input of the first stage; the stages are forward-coupled by a capacitor of the same value.

cross coupling **1.** The state of being cross-coupled (see, for example, CROSS-COUPLED MULTIVIBRATOR). **2.** Undesired coupling between two circuits.

cross current A current that flows in the opposite direction from some other current.

crossed-pointer indicator **1.** Also called *crossed-needle meter.* A combination of two analog metering instruments in one case. Each needle has its own independently calibrated scale. A third scale corresponds to the intersection point of the needles. Commonly used in directional wattmeters that simultaneously show forward power, reflected power, and standing-wave ratio (SWR). **2.** A two-pointer meter used in aircraft to show the position of the aircraft, relative to the glide path.

crossed-wire thermoelement Two wires or strips of dissimilar metals joined or twisted at a point that constitutes a thermoelectric junction. In usual operation, a high-frequency current is passed through one wire, and a proportional direct-current (dc) voltage, generated by thermoelectric action, appears at the other wire.

cross flux The magnetic flux component that is perpendicular to the flux produced by field magnets.

cross-hair pattern A television test pattern consisting of a single vertical line and a single horizontal line, which form a simple cross. The pattern resembles the cross hairs of an optical instrument.

crosshatch generator A modulated radio-frequency (RF) signal generator that produces a crosshatch pattern on a picture-tube screen.

crosshatch pattern A grid of horizontal and vertical lines produced on a picture-tube screen by a cross-hatch generator. It is used in checking horizontal and vertical linearity.

cross modulation **1.** A type of radio-frequency interference (RFI) between two strong stations that are close in frequency. The desired carrier is modulated by the interfering signal. **2.** The production of signals by rectifier junctions in pipes and wiring near a radio receiver. These objects pick up waves and deliver energy at a different frequency, which finds its way into the receiver. Also called *external cross modulation.* **3.** The interaction between signals of different frequency when they magnetize a core of nonlinear magnetic material. Also see CROSSTALK.

cross-modulation factor An expression of the amount of cross modulation (or crosstalk) present in a particular instance. It is equal to M_1/M_2, where M_1 is the modulation percentage that a modulated wave produces in a superimposed unmodulated wave, and M_2 is the modulation percentage of the modulated wave.

cross-neutralized circuit See CROSS-CONNECTED NEUTRALIZATION.

crossover **1.** In a circuit diagram, a point at which lines representing wires intersect, but are not connected. **2.** In a characteristic curve, point at which the plot crosses an axis or operating point. **3.** See CROSSOVER NETWORK.

crossover distortion Distortion of a characteristic at a crossover point (see CROSSOVER, **2**); for example, a bend in the curve where the plot of a waveform passes through zero.

crossover frequency The frequency at which a crossover network delivers equal power to the two circuits it supplies.

crossover network Following final amplification in a sound-reproduction system, an outboard filter circuit that facilitates delivery of the low and high audio frequency (AF) components to the correct speakers.

crossover point See CROSSOVER, **2**.

crossover S-curve The S-shaped image obtained on an oscilloscope screen during sweep-generator alignment of a frequency-modulation (FM) detector. In correct alignment, the exact center of the S-curve (the crossover point) coincides with the zero point on the screen.

cross product Also called *vector product.* For vectors **A** and **B** having lengths A and B, respectively, and subtending an angle θ relative to each other, the cross product **A** × **B** points in a direction perpendicular to the plane containing both **A** and **B**. The length of **A** × **B** is equal to AB sin θ.

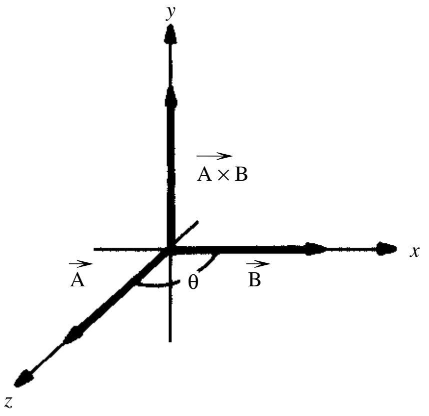

cross product

cross-sectional area **1.** The surface area of a face of a conductor after cutting through it at a right angle. Specified in square inches, square millimeters, or circular mils. **2.** The total of the cross-sectional areas of all the wires in a stranded conductor.

cross-sectional testing In quality assurance and quality control (QA/QC), a method of checking a large lot of units or components. Instead of testing every device, a fraction of the devices is tested. The sampling is taken uniformly from the group (e.g., every fifth unit).

crosstalk Undesired transfer of signals between or among telephone lines, data lines, or system components. In computer operations, this effect places a practical limit on the lengths of parallel data cables.

crosstalk coupling Undesired coupling between circuits, caused by crosstalk.

crosstalk factor See CROSS-MODULATION FACTOR.

crosstalk level The amplitude of crosstalk, usually expressed in decibels above a reference level.

crosstalk loss Loss of energy caused by crosstalk.

crowbar An action producing a high overload on a circuit protection device.

crowfoot **1.** A pattern formed by the cracking or crazing of solid plastics of solidified encapsulating compounds, so called from its resemblance to a bird's footprint. **2.** In a gravity battery cell, the zinc electrode, so called from its resemblance to a bird's foot.

CRT Abbreviation of CATHODE-RAY TUBE.

crud **1.** Broadband electrical noise, originating inside and/or outside a system. **2.** Undesired signals that interfere with a desired signal.

cryogenic device A device that exhibits unique electrical characteristics (such as superconductivity) at extremely low temperatures.

cryogenic motor A motor designed for operation at extremely low temperatures.

cryoelectronics The study of the behavior of electronic devices, circuits, and systems at extremely low temperatures.

cryogenics The branch of physics dealing with the behavior of matter at temperatures approaching absolute zero. Also concerned with methods of obtaining such temperatures in controlled environments.

cryosar A semiconductor switch utilizing low-temperature avalanche breakdown.

cryoscope An instrument used to determine freezing point.

cryostat A chamber for maintaining a very low temperature for cryogenic operations. Also see CRYOGENICS.

cryotron A switching device consisting essentially of a straight tantalum wire, around which a single-layer control coil is wound. The magnetic field generated by control current flowing through the coil causes the tantalum wire to become superconductive at a temperature of approximately 4.4 degrees K.

cryotronics Low-temperature electronics, concerned with such phenomena as superconductivity. The term is an acronym from cryogenics and electronics. Also see CRYOGENICS.

cryptanalysis The breaking of ciphers.

crypto- A prefix added to words, that implies encoding for the purpose of changing or hiding the meaning of a message or signal.

cryptography The creating and writing of ciphers.

cryptology The art and science of creating, writing, unscrambling, and breaking ciphers.

crystal **1.** A material distinguished by the arrangement of its atoms into a redundant pattern called a *lattice* that presents, in fragments of various sizes, a characteristic polyhedral shape. Common shapes include cubes, parallelepipeds, and hexagonal prisms. **2.** A fragment of material as defined in (1). **3.** A plate or bar cut from a piece of piezoelectric material.

crystal amplifier **1.** A semiconductor diode circuit using carrier storage. Transistor action and, accordingly, pulse amplification is obtained by alternately making one electrode of the diode an emitter or collector. **2.** Archaic term for TRANSISTOR.

crystal audio receiver An audio radar receiver, consisting of a crystal detector and audio-amplifier stages.

crystal axes The imaginary lines traversing a piezoelectric crystal, along which (or perpendicular to which) plates are cut for oscillators, resonators, or transducers.

crystal calibrator A crystal oscillator used to generate harmonic checkpoints for frequency calibration. Common fundamental calibrator frequencies are 100 kHz and 1 MHz.

crystal capacitor See VARACTOR.

crystal control The control of the operating frequency of a circuit by means of a piezoelectric crystal.

crystal-controlled receiver A superheterodyne radio receiver whose local oscillator is crystal controlled.

crystal-controlled transmitter A radio transmitter whose master oscillator is crystal controlled.

crystal counter A device for counting the frequency of subatomic particles, based on their ability to change the conductivity of a crystal. The particles can be photons, electrons, protons, neutrons, or the nuclei of atoms.

crystal current Current flowing through a crystal; specifically, the radio-frequency (RF) current flowing through a quartz plate in a crystal-controlled oscillator.

crystal cuts The classification of piezoelectric plates according to the angle at which they were cut from a quartz crystal. Common cut designations are AT, BT, CT, DT, X, Y, and Z. Various cuts afford such complementary factors as frequency, temperature, and thickness. Also see CRYSTAL AXES.

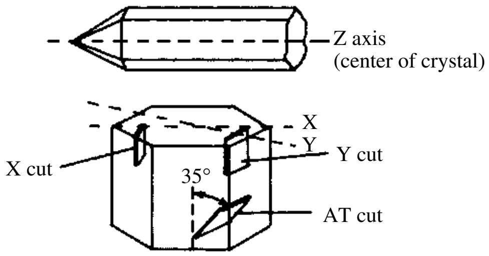

crystal cuts

crystal detector A rudimentary form of semiconductor diode consisting of a mounted lump of mineral (the crystal) in contact with a springy wire ("cat's whisker"). The point of the wire is moved to various points of contact on the crystal surface until the most-sensitive rectifying spot is found.

crystal diffraction The tendency of electromagnetic waves to be scattered when passing through a crystal material.

crystal diode Archaic term for SEMICONDUCTOR DIODE. Also see GALLIUM-ARSENIDE DIODE; GERMANIUM DIODE; JUNCTION DIODE; LASER DIODE; POINT-CONTACT DIODE; SELENIUM DIODE; SIGNAL DIODE; SILICON DIODE.

crystal earphone An earphone in which the transducer is a piezoelectric crystal. Electrical impulses applied to the crystal vary its shape and cause a vibration that is transmitted to a diaphragm; this in turn produces corresponding sound waves.

crystal filter See CRYSTAL RESONATOR.

crystal headphone See CRYSTAL EARPHONE.

crystal holder A fixture specially designed to hold a piezoelectric crystal; it ensures minimum distortion of crystal dimensions and minimum residual capacitance, inductance, and resistance.

crystal imperfection A flaw in the lattice structure of a crystal.

crystal lattice The orderly, redundant pattern of atoms and molecules within a crystalline material; it is a characteristic of a given material.

crystal-lattice filter A crystal resonator in which piezoelectric crystals are used to give a desired shape to the filter response curve.

crystalline material A material exhibiting the characteristic properties of a crystal (see CRYSTAL, **1**).

crystallogram An X-ray photograph or other record of crystal structure.

crystallography The science dealing with crystals and their properties (see CRYSTAL, **1**).

crystal loudspeaker A loudspeaker whose transducer is a piezoelectric crystal. Electrical impulses applied to the crystal vary its shape and cause vibrations that are transmitted to a diaphragm or cone, which produces corresponding sound waves.

crystal meter A rectifier-type ac meter using a semiconductor diode in series with a dc milliammeter or microammeter.

crystal microphone A microphone whose transducer is a natural or synthetic piezoelectric crystal. Sound waves striking the crystal (directly or via a diaphragm) vary its shape, making it produce an audio-frequency (AF) output voltage.

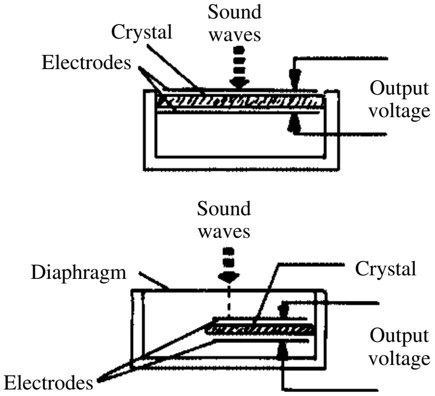

crystal microphones

crystal mixer A mixer (converter) circuit utilizing the nonlinearity of a semiconductor diode to mix signals.

crystal operation **1.** The characteristics of a piezoelectric crystal in a particular circuit. **2.** Crystal frequency control.

crystal oscillator An oscillator whose operating frequency is determined by the dimensions of an oscillating piezoelectric quartz-crystal plate. Compare SELF-EXCITED OSCILLATOR.

crystal oven A constant-temperature chamber for stabilizing the frequency of a quartz crystal by maintaining its operating temperature at a fixed point.

crystal photocell A photoelectric cell in which the light-sensitive material is a crystalline substance, such as germanium, selenium, silicon, etc.

crystal pickup A phonograph pickup whose transducer is a natural or synthetic piezoelectric crystal. The crystal is attached (either directly or through a mechanical linkage) to a stylus, whose movement in the disk groove varies the shape of the crystal. The resultant vibration generates a corresponding audio-frequency (AF) output voltage across the crystal.

crystal probe A radio-frequency (RF) probe, whose rectifying element is a semiconductor diode.

crystal pulling **1.** The extraction of a single crystal from a molten mass of crystalline material. Single crystals are used for high-quality semiconductor devices. Also see CZOCHRALSKI METHOD, SINGLE CRYSTAL, and SINGLE-CRYSTAL MATERIAL. **2.** The use of an inductor or capacitor in a crystal-controlled radio-frequency (RF) oscillator circuit to allow adjustment of the frequency over a small range.

crystal receiver See CRYSTAL SET.

crystal rectifier **1.** A semiconductor diode used for the purpose of rectifying alternating current (dc), usually in a power supply.

crystal resistor A temperature-sensitive resistor made from silicon, and exhibiting a positive temperature coefficient of resistance.

crystal resonator A highly selective resonant circuit in which the center frequency is the resonant frequency of a piezoelectric quartz-crystal plate.

crystal sensor See CRYSTAL TRANSDUCER.

crystal set A simple radio receiver that uses a tuned circuit, semiconductor-diode detector, and earphones.

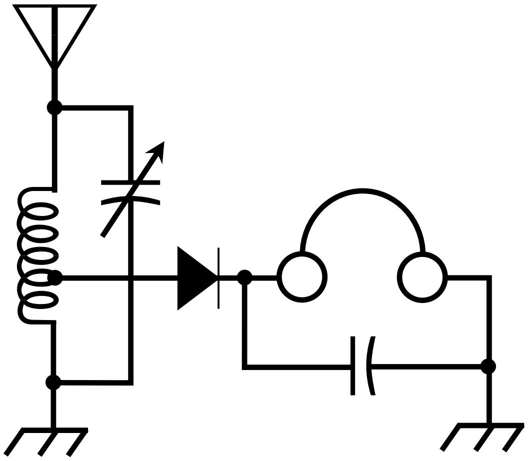

crystal set

crystal slab See QUARTZ BAR.

crystal socket **1.** A low-capacitance, low-loss socket for a piezoelectric crystal. **2.** A socket for a semiconductor diode.

crystal tester **1.** An oscillator used to check quartz crystals. Most such units check only the crystal's ability to oscillate; more elaborate ones also check crystal current, frequency, temperature coefficient, activity, filter action, etc. **2.** An instrument for checking the electrical characteristics of semiconductor diodes. **3.** An instrument for checking the performance of piezoelectric ceramics.

crystal tetrode A transistor having four elements: emitter, collector, and two bases.

crystal transducer A transducer using a piezoelectric crystal as the sensitive element. Examples: crystal earphone, crystal loudspeaker, crystal microphone, and crystal pickup.

crystal triode See TRANSISTOR.

Cs Symbol for CESIUM.

CS Abbreviation of COMPLEMENTARY SYMMETRY. Also COS.

C_S **1.** Symbol for *standard capacitance.* **2.** Symbol for *source capacitance.*

csc Abbreviation of COSECANT.

C scan See C DISPLAY.

csch Abbreviation of HYPERBOLIC COSECANT.

C scope A cathode-ray tube used in radar to provide a C DISPLAY.

CT-cut crystal A piezoelectric plate cut from a quartz crystal at an angle of rotation around the X-axis of +38°. Such a plate has a zero temperature coefficient of frequency at 25°C. Also see CRYSTAL AXES and CRYSTAL CUTS.

CTL Abbreviation of *complementary-transistor logic.*

Cu Symbol for COPPER.

cube **1.** A regular polyhedron with six identical square faces and eight vertices. At each vertex, three edges converge at mutual right angles. **2.** The third power of a number; thus the cube of n is written n^3.

cube tap An electrical adapter, in which a set of male prongs and three sets of female contacts are on the sides of a molded cube. Allows three appliances to be used with a single electrical socket.

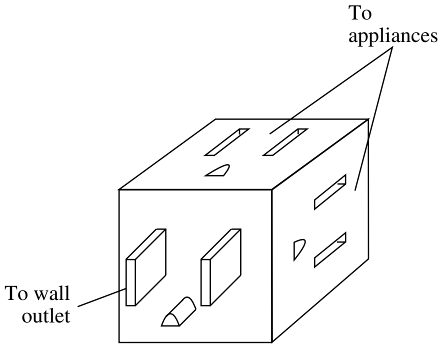

cube tap

cubical antenna An antenna in which the elements form the outline of a geometric cube or rectangular prism. The most common example is the QUAD ANTENNA.

cubical quad antenna See QUAD ANTENNA.

cubic equation A polynomial equation of the third degree. Its general form is $ax^3 + bx^2 + cx + d = 0$.

cue A condition or signal that alerts an operator, circuit or system to act in a specific manner.

cue circuit A device for transmitting cues used in program control.

cueing receiver **1.** A (usually miniature) radio receiver used to pick up cues. Example: a receiver carried by a technician, actor, or lecturer. **2.** A receiver or other pickup circuit that receives a cuing pulse, which it uses to set another circuit.

cu ft Abbreviation of *cubic foot* or *cubic feet.*

cu in Abbreviation of *cubic inch* or *cubic inches.*

cumulative error In a sum or other final value, the total error that has accumulated from the individual errors in the terms. Also called *systematic error.*

cup core A coil core that also forms a magnetic shield around the coil.

cuprous-oxide rectifier See COPPER-OXIDE RECTIFIER.

cur Abbreviation of CURRENT.

curie Abbreviation Ci. A unit of radioactivity; 1 curie is the amount of radiation from (or in equilibrium with) 1 gram of radium. Also equivalent to 3.7×10^{10} atomic breakdowns per second.

Curie point **1.** The temperature above which a ferromagnetic material loses its magnetism or becomes paramagnetic. **2.** The temperature at which the ferroelectric properties of a substance disappear.

curie temperature As a magnetized substance is heated, the lowest temperature at which magnetization is lost. It is generally measured in degrees Celsius or degrees Kelvin. For iron, this temperature is 760 degrees Celsius; for nickel, it is 356 degrees Celsius.

Curie's law For a paramagnetic substance, the ratio of the magnetization to the magnetizing force is inversely proportional to the absolute temperature.

Curie-Weiss law Above the Curie point, the susceptibility of a paramagnetic material varies inversely as the excess of temperature above the Curie point increases. This law is invalid for applications at or below the Curie point.

curium Symbol, Cm. A radioactive metallic element produced artificially. Atomic number, 96. Atomic weight, 247.

current Symbol, I or i. The movement of charge carriers, such as electrons, holes, or ions. Also see AMPERE.

current amplification **1.** An electronic process in which the instantaneous, average, or peak magnitude of a current is increased. **2.** The extent to which a current increases in a circuit; the ratio (always greater than one) of output current to input current, I_{out}/I_{in}. Also called *current gain.*

current amplifier An amplifier operated primarily to increase a signal current. Compare POWER AMPLIFIER and VOLTAGE AMPLIFIER.

current antinode See CURRENT LOOP.

current attenuation **1.** The reduction of current amplitude along a line. **2.** The extent to which a current decreases in a line or circuit; the ratio (always less than one) of output current to input current, $I_{\text{out}}/I_{\text{in}}$.

current balance An instrument for determining the size of the ampere. This is done by measuring the force between two current-carrying conductors.

current-balance switch A switch or relay, operated by the existence of a difference between two currents.

current-carrying capacity The maximum current (usually expressed in amperes) that a conductor or device can safely conduct.

current coil The series coil in a nonelectronic wattmeter. Compare POTENTIAL COIL.

current-controlled amplifier Abbreviation, CCA. An amplifier in which gain is controlled by means of a current applied to a control-input terminal.

current density The current (usually expressed in amperes per square centimeter) passing through a cross-sectional area of a conductor.

current drain **1.** The current supplied to a load by a generator or generator-equivalent. **2.** The current required by a device for its operation; also, the current taken by the device during standby periods.

current echo Reflected current in a transmission line that is not terminated in an impedance exactly matching its characteristic impedance.

current-fed antenna An antenna in which the transmission line is attached to the radiator at a current loop (voltage node). Compare VOLTAGE-FED ANTENNA.

current feed **1.** The delivery of power to a device or circuit at a point where current dominates. Compare VOLTAGE FEED. **2.** In an antenna, feeding it at a current maximum.

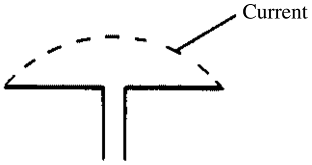

Current feed

current feedback **1.** A feedback signal consisting of current fed from the output to the input circuit of an amplifier. **2.** A system or circuit for obtaining current feedback.

current-feedback pair A two-stage, direct-coupled transistor amplifier having direct-current shunt-series feedback.

current flow Charge carriers passing through a solid, liquid, gas, or vacuum. Also see CURRENT and CURRENT DENSITY.

current gain See CURRENT AMPLIFICATION.

current hogging **1.** An undesirable condition that sometimes takes place when two or more transistors are operated in parallel. One device tends to do all the work, taking all the current. The result can be destruction of that device. **2.** The tendency of one component in a group of identical parallel-connected components to dissipate most of the power.

current-hogging injection logic Acronym, CHIL. A form of bipolar digital logic, similar to current-hogging logic but having the greater density characteristic of injection logic.

current instruction register A register in which are held instructions ready for execution by a program controller.

current lag A circuit condition in which current variations are delayed by up to 180 degrees of phase relative to voltage variations. Compare CURRENT LEAD.

current lead A circuit condition in which current variations occur earlier than voltage variations by up to 180 degrees of phase. Compare CURRENT LAG.

current limiting The controlling of current so that it does not exceed a desired value.

current-limiting resistor A series resistor inserted into a circuit to limit the current to a prescribed value.

current loop A point on a transmission line or antenna radiator at which the current reaches a local maximum. Compare CURRENT NODE.

current meter A usually direct-reading instrument, such as an ammeter, milliammeter, or microammeter, used to measure current strength. Also see ELECTRONIC CURRENT METER.

current-meter operation The operation of a voltmeter as a current meter by connecting it to respond to the voltage drop across a resistor that carries the current of interest.

current-mode logic In computer operations, transistor logic in which the transistors operate in the unsaturated mode.

current node A point on a transmission line or antenna radiator at which the current reaches a local minimum. Compare CURRENT LOOP.

current noise Electrical noise produced by current flowing through a resistor.

current probe A transformer usually having a snap-around, one-turn coil that picks up energy from a conductor and couples it into an alternating-current ammeter.

current rating **1.** A specified value of operating current. **2.** See CURRENT-CARRYING CAPACITY.

current-regulated supply See CONSTANT-CURRENT SOURCE.

current regulation The stabilization of current at a predetermined level or value.

current regulator See BARRETTER.

current relay A relay actuated by specific values of pickup and dropout current.

current saturation In the operation of a device (such as a transistor, saturable reactor, or magnetic amplifier), the leveling off of current at a value beyond which no further increase occurs—even though an input parameter is further increased.

current sense amplifier An amplifier used to increase the sensitivity of, or to decrease the loading of, a current-sensing component.

current sensing Sampling a current (e.g., when the voltage drop across a series resistor is used as a proportional indication of the current flowing through the resistor).

current-sensing resistor A low-value resistor inserted into a circuit primarily for current sensing.

current sensitivity In a current meter or galvanometer, current (in amperes or fractions thereof) per scale division.

current-sheet inductance Symbol, L_S. The low-frequency inductance of a single-layer coil, calculated with the formula $L_S = (0.10028\ a^2N^2)/s$, where L_S is in microhenrys, a is the coil radius in inches, N is the total number of turns, and s is the coil length in inches.

current shunt **1.** A resistor connected in parallel with a voltmeter to convert it into an ammeter. **2.** A resistor connected in parallel with the input of a voltage amplifier to make the response of the amplifier proportional to input-signal current.

current sink A circuit or device through which a constant current can be maintained.

current-sinking logic A form of bipolar digital logic. Current flows from one stage to the input of the stage immediately before.

current-squared meter An ammeter or milliammeter whose deflection is proportional to the square of the current.

current-stability factor In a common-base connected bipolar transistor, the ratio dI_E/dI_C, where I_E is the emitter current and I_C is the collector current.

current strength The magnitude of electric current (see CURRENT) (i.e., the number of carriers flowing past a given point per unit time, expressed in coulombs per second or in amperes).

current transformer **1.** A transformer used to increase or decrease current flow. A primary-to-secondary step-up turns ratio reduces the current; a primary-to-secondary step-down turns ratio increases the current. **2.** A particular transformer (as in 1) used to change the range of an alternating-current milliammeter or ammeter.

current vector In a vector diagram, a line with an arrowhead (vector) showing the magnitude and phase of a current. Compare VOLTAGE VECTOR.

current-voltage feedback In an amplifier or oscillator, the process of applying some of the output current and voltage to the input. This feedback might be in phase (positive) or out of phase (negative), with respect to the input.

cursor **1.** A marker that indicates the position where a character can be entered in a video alphanumeric display. Commonly used in computers and word processors. **2.** The sweeping line on a radar display. **3.** The movable marker on a slide rule.

curve trace **1.** A device that supplies a special variable test voltage to a component or circuit under test, at the same time supplying a sweep voltage to an oscilloscope. The component's output voltage is also presented to the oscilloscope. As a result, the response curve of the component appears on the oscilloscope screen. **2.** A device that produces a permanent record (photographic or graphic) of an electrical phenomenon. Also called OSCILLOGRAPH or RECORDER.

curvilinear trace A trace made on paper with curved vertical lines. The lines are curved to match the arc through which the recording pen swings.

cut-in angle In a semiconductor rectifier circuit, a phase angle slightly greater than zero degrees, at which current conduction begins. Compare CUTOUT ANGLE.

Cutler antenna A parabolic-dish antenna, in which the driven element consists of a waveguide that has two apertures on opposite sides of a resonant cavity.

Cutler feed An aircraft antenna feed system in which radio-frequency (RF) energy is fed to the reflector by a resonant cavity at the end of a waveguide.

Cutler tone control A dual resistance-capacitance (RC) filter circuit of the general bridged-tee variety. Variation of the series leg provides adjustable treble boost; variation of the shunt leg provides adjustable bass boost.

cutoff **1.** The process of reducing some operating parameter, such as collector current, to zero by adjusting the bias at the input electrode. **2.** The point on the characteristic curve of an amplifying device, at which the output current drops to zero under no-signal conditions. **3.** The lowest frequency at which a waveguide will efficiently function. **4.** The frequency or frequencies corresponding to the point or points in a filter response, at which the attenuation is three decibels greater than the lowest attenuation within the passband. See also CUTOFF FREQUENCY.

cutoff attenuator A variable, nondissipating attenuator consisting of a variable length of waveguide used at a frequency below cutoff.

cutoff bias In a transistor or vacuum-tube circuit, the value of control-electrode bias that produces output current cutoff.

cutoff current Symbol, I_{co}. In a transistor, the small collector current that flows when the emitter current is zero (common-base circuit) or when the base current is zero (common-emitter circuit).

cutoff frequency **1.** Symbol, f_{co}. The high frequency at which the current-amplification factor of a transistor drops to 70.7% of its 1-kHz value. **2.** In a filter, amplifier, or transmission line, the frequency point(s) at which transmission loss or filter rejection begins. It is generally specified as the half-power point(s), or the point(s) at which the attenuation is three decibels, relative to the lowest attenuation. Examples: the high-frequency cutoff of an amplifier and the upper and lower cutoff points of a bandpass filter.

cutoff limiting Output-peak clipping that results from overdrive in an amplifying device. Compare SATURATION LIMITING.

cutoff potential See CUTOFF BIAS.

cutoff voltage See CUTOFF BIAS.

cutoff wavelength **1.** The wavelength corresponding to cutoff frequency. **2.** For a waveguide, the ratio of the velocity of electromagnetic waves in free space (3×10^8 meters per second) to the cutoff frequency of the waveguide in Hz. The result is thus expressed in meters.

cutout **1.** A device, such as a circuit breaker, that automatically disconnects a circuit, usually to prevent overload, but occasionally to prevent underload. **2.** Emergency switch. **3.** Fuse.

cut-out angle In a semiconductor rectifier circuit, a phase angle slightly less than 180 degrees at which current conduction ceases. Compare CUT-IN ANGLE.

cutout base A fuse block.

cut rate **1.** The speed at which a cutter moves across the surface of a blank vinyl disk during the recording process. **2.** The number of cut lines per inch in a vinyl disk recording.

CW **1.** Abbreviation of CONTINUOUS WAVE. **2.** Abbreviation of CLOCKWISE.

CW filter In a communications receiver, a highly selective filter in the intermediate-frequency (IF) or audio-frequency (AF) stage. The bandwidth is typically 200 Hz to 500 Hz; some audio filters can be set for bandwidths as low as about 50 Hz.

CW laser A laser that emits energy in an uninterrupted stream, rather than in pulses.

CW monitor See KEYING MONITOR.

CW oscillator **1.** In a radio receiver, a variable-frequency oscillator that heterodynes a radiotelegraph signal in the intermediate-frequency (IF) amplifier chain, to make audible the continuous-wave dits and dahs. **2.** Sometimes, an external variable-frequency radio-frequency (RF) oscillator, whose output beats against the actual carrier of a continuous-wave radiotelegraph signal, making it audible as dits and dahs. **3.** An unmodulated, unkeyed oscillator.

CW radar A radar system in which radio-frequency (RF) energy is transmitted continuously.

CW reference signal A sinusoidal radio-frequency (RF) signal, used to control the conduction time of a synchronous demodulator in color television.

C_x Symbol for UNKNOWN CAPACITANCE.

cyan Blue-green, one of the three primary pigments.

cyber- A prefix that indicates relevance to, or involvement with, computers, computer systems, and electronic control systems.

cybernetics The study of control system theory in terms of the relationship between animal and machine behavior.

Cyber Sapiens An expression for a computer or robot with artificial intelligence (AI) on the forefront of current technology.

cyberspace **1.** Alternative expression for INFORMATION SUPERHIGHWAY. **2.** Alternative expression for VIRTUAL REALITY.

cyborg Acronym of the words *cybernetic* and *organism.* **1.** A human being with at least one artificial body part, such as a *prosthesis* (artificial limb). **2.** A human being who is largely composed of robotic body parts.

cycle **1.** Abbreviation, c. One complete, 360-degree revolution of the current or voltage vector in an alternating-current (ac) wave. An ac frequency of 1 cycle per second is 1 Hz (see HERTZ). **2.** A complete sequence of operations.

cycle counter A device that totals the number of cycles of a phenomenon repeated during a given period.

cycle index The number of times that a particular cycle has been, or must be, iterated in a computer program.

cycle index counter A variable that indicates how often a cycle of computer program instructions has been executed. In a program, for example, this can be accomplished by increasing, through instruction, the value of a location's content every time a loop operation is performed.

cycle life The total number of charge-discharge cycles a rechargeable cell or battery can tolerate before becoming useless.

cycle reset To change the value of a cycle count (making it zero or some other value).

cycle shift See CYCLIC SHIFT.

cycles per second Abbreviation, cps. Archaic term for HERTZ.

cycle time Pertaining to an operation, the duration of a complete cycle.

cycle timer A timer that switches a circuit or device on and off, according to a predetermined cycle. Also called *programmed timer.*

cyclic code See GRAY CODE.

cyclic memory In computer operations, a memory whose locations can only be accessed points in a cycle, as of a magnetic diskette.

cyclic shift The moving of data out of one end of a storage register and reentering it character-by-character or bit-by-bit at the other end in a closed loop (e.g., 87654 cyclically shifted one place to the right becomes 48765).

cyclic variations Periodic changes in the features of the ionosphere, occurring on a daily, seasonal, or sunspot-related basis. These changes are fairly predictable.

cycling The tendency of a parameter to oscillate back and forth between two different values.

cyclogram A method of showing the relationship between two signals on an oscilloscope. The two signals must have a fixed phase relationship.

cyclotron A type of particle accelerator. An applied electromagnetic field, acting together with an intense applied magnetic field, cause charged subatomic particles to travel with increasing velocity in a spiral path between two semicircular metal boxes called *dees.* When the particles go fast enough in the correct path, they are expelled and strike a target in their path.

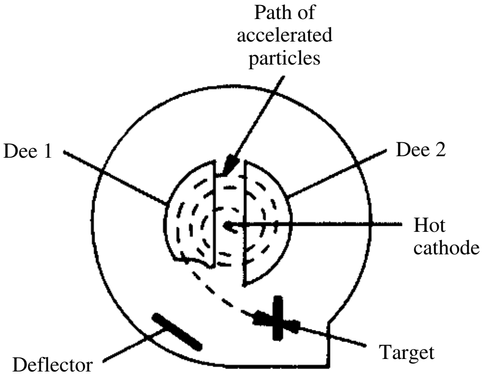

cyclotron

cyclotron frequency The angular frequency of a charged particle in a cyclotron. The cyclotron frequency depends on the number of times per second the magnetic field of the device is reversed.

cyclotron radiation An electromagnetic field produced by the circular movement of charged particles in a fluctuating magnetic field.

cylinder In computer operations, the combination of equal-radius tracks on the platters of a hard disk.

cylinder magnet A permanent magnet in the shape of a cylinder.

cylindrical capacitor See CONCENTRIC CAPACITOR.

cylindrical contour The most common curvature of the face of a magnetic tape recording head; it is a section of a cylinder having a constant radius of 0.5 inch to 1 inch.

cylindrical coordinate geometry A scheme for robot-arm movement. There are three coordinates, called *reach, angle,* and *elevation.* It allows precise positioning of a robot end effector within a region consisting of two concentric cylinders and all the volume in between.

cylindrical coordinates A method of locating a point in three-space in which height, distance, and angle are used to uniquely define points.

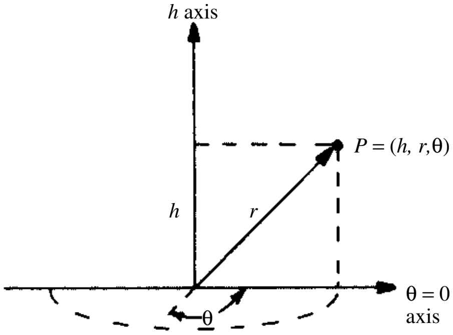

cylindrical coordinates

cylindrical magnet See CYLINDER MAGNET.

cylindrical wave An electromagnetic wave whose field surfaces are nearly perfect cylinders.

cylindrical waveguide A waveguide resembling a round pipe.

cylindrical winding A method of coil winding in which the wire is formed into a helix. There might be only one layer, or there might be several layers. The length of the coil is greater than the diameter. Also called a *linear winding.*

Czochralski method A technique for obtaining a relatively large single crystal from a substance, such as the semiconductors germanium and silicon. The method consists essentially of dipping a seed crystal into a molten mass of the same substance, then slowly withdrawing it while rotating it.

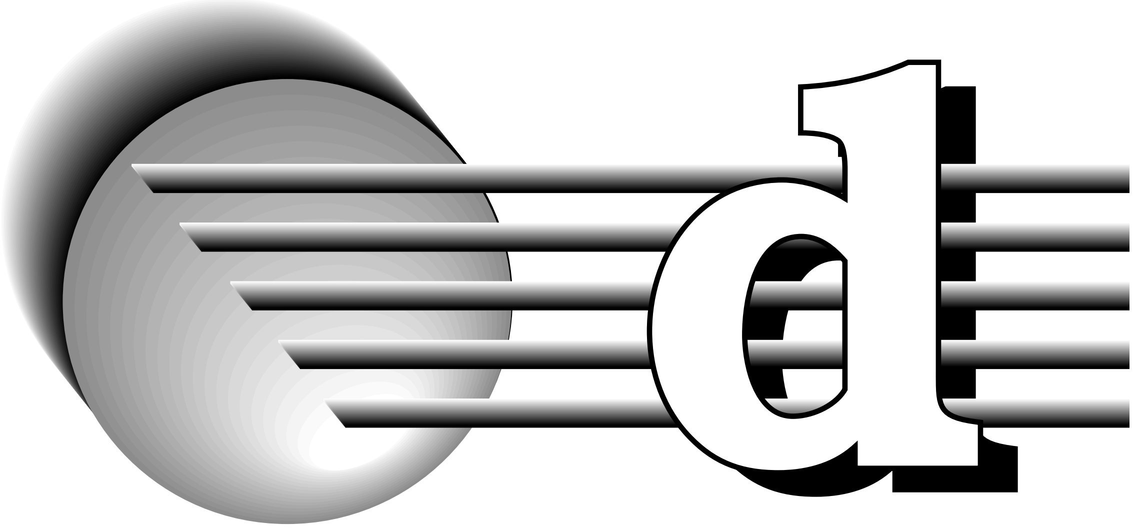

D **1.** Symbol for DEUTERIUM. **2.** Symbol for ELECTRIC DISPLACEMENT. **3.** Symbol for ELECTRIC FLUX DENSITY. **4.** Symbol for DISSIPATION FACTOR. **5.** Symbol for *drain* (see DRAIN, **3**). **6.** Abbreviation of DISSIPATION. **7.** Symbol for *determinant.* **8.** Symbol for DIFFUSION CONSTANT.

d **1.** Abbreviation of DECI. **2.** Symbol for DIFFERENTIAL. **3.** Symbol for distance. **4.** Symbol for DENSITY. **5.** Symbol for *drain* (see DRAIN, **3**). **6.** Abbreviation of DISSIPATION. **7.** Abbreviation of day. **8.** Abbreviation of DEGREE. **9.** Abbreviation of diameter. **10.** Abbreviation of DRIVE.

D/A Abbreviation of DIGITAL-TO-ANALOG. See DIGITAL-TO-ANALOG CONVERSION.

dA **1.** Symbol for DIFFERENTIAL OF AREA. **2.** Symbol for *differential of amplification.* **3.** Seldom-used abbreviation of *deciampere.*

da Abbreviation of DEKA.

DAC Abbreviation of DIGITAL-TO-ANALOG CONVERTER.

DACI Abbreviation of *direct adjacent-channel interference.*

DAGC Abbreviation of DELAYED AUTOMATIC GAIN CONTROL.

daisy chain A method of transferring a signal in a computer from one stage to the next.

daisy wheel A form of printing device consisting of a disk having several dozen radial spokes, each of which has a character molded on its face. The disk rotates to the proper position in the printing process, and a hammer strikes the spoke to press the molding against the ribbon and paper.

DAM Abbreviation of *data-addressed memory.*

Damon effect The change that the susceptibility of ferrite undergoes under the influence of high RF power.

damped galvanometer A galvanometer with a provision for overswing limiting or oscillation prevention.

damped loudspeaker A loudspeaker in which undesirable excursions are prevented by damping in the associated amplifier or speaker circuit.

damped meter **1.** A meter with a provision for overswing limiting or oscillation prevention. **2.** A meter that is protected during transport by a shorting bus between the two meter terminals.

damped natural frequency **1.** The frequency at which a damped system having one degree of freedom will oscillate after momentary application of a transient force. **2.** In the presence of damping, the rate at which a sensing element oscillates freely.

damped oscillations Oscillations in which the amplitude of each peak is lower than that of the preceding one; the oscillation eventually dies out (the amplitude becomes zero). Compare CONTINUOUS WAVE.

damped speaker See DAMPED LOUDSPEAKER.

damped wave A wave whose successive peaks decrease in amplitude (i.e., it decays), eventually reaching an amplitude of zero. Compare CONTINUOUS WAVE and UNDAMPED WAVE.

damped-wave decay See DECREMENT, **1**.

dampen To cause the amplitude of a signal to decay.

damper See DAMPING DIODE.

damper diode See DAMPING DIODE.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

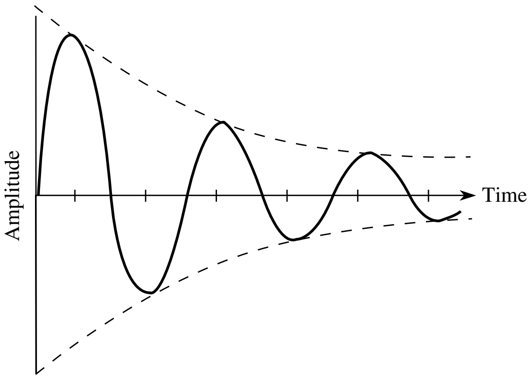

damped wave

damper winding A special short-circuited motor winding that opposes pulsation or rotation of the magnetic field.

damping **1.** See DAMPING ACTION. **2.** In a loudspeaker, sound-absorbent material used to minimize resonant effects within the enclosure.

damping action **1.** Quenching action. **2.** The prevention of overswing, dither, or flutter in a meter or loudspeaker (see DAMPED GALVANOMETER, DAMPED LOUDSPEAKER, DAMPED METER). **3.** The prevention of oscillation or ringing in a circuit. **4.** Inhibition of the vibration of an acoustic transducer to prevent ringing and other unwanted effects.

damping coefficient A figure expressing the ratio of the damping in a system to critical damping.

damping diode A diode used to prevent oscillation in an electric circuit (e.g., the diode that prevents ringing in the power supply of a television receiver). Also called *damper.*

damping factor **1.** Symbol, a. For a coil of inductance L and RF resistance R in a damped-wave circuit, the value $R/2L$, where L is in henrys and R in ohms. **2.** Abbreviation, F_o. For a torque motor, the ratio of the stall torque to the no-load rotational speed.

damping magnet A permanent magnet so situated, with respect to a moving conductor, disk, or plate, that the resulting field opposes the movement.

damping ratio See DAMPING COEFFICIENT.

damping resistance **1.** The value of shunt resistance required to prevent ringing in a coil. **2.** The value of resistance required for critical damping of a galvanometer.

damping resistor **1.** A shunt across a coil to prevent ringing. **2.** A resistor used to provide critical damping of a galvanometer.

Daniell cell A nonpolarizing primary wet cell with zinc (negative) and copper (positive) electrodes. The zinc plate is in a porous cup containing a weak zinc-sulfate solution with a little sulfuric acid; the cup is in a jar filled with a saturated copper-sulfate solution in which the copper electrode is immersed. Typical voltage for the cell is 1.1 V.

daraf The unit of ELASTANCE. Elastance in darafs is the reciprocal of capacitance in farads.

dark conduction The flow of dark current in a photoconductive or glow-discharge device.

dark current The usually tiny current flowing through a darkened photoconductive cell, phototransistor, or glow-discharge device.

dark discharge The occurrence of a discharge in a gas, without the production of visible light.

dark-spot signal A spurious signal generated by some camera tubes, arising from secondary-emission effects.

dark-trace tube An oscilloscope tube on whose white screen a long-persistence magenta image is traced by the electron beam. Illuminating the screen with bright light intensifies the image.

Darlington amplifier A high-gain amplifier that uses a COMPOUND CONNECTION of two bipolar transistors.

Darlington pair See COMPOUND CONNECTION.

D'Arsonval current A large, low-voltage, high-frequency current at one time thought to be therapeutic.

D'Arsonval meter A electromechanical analog meter, in which a coil turns on jeweled pivots between the poles of a strong magnet and against the force of spiral springs. A pointer is attached to the coil. The pointer moves over a calibrated scale.

D'Arsonval movement The mechanism of a D'Arsonval meter.

DART Abbreviation of *data analysis recording tape.*

dart leader A flow of electrons along a path traveled by a lightning stroke, preceding a second stroke. The dart leader, if any, occurs a few milliseconds after the first stroke. Several strokes could occur, each preceded by a dart leader, within less than 1 second.

dash The longer of the two characters (DOT and DASH) of the telegraph code. The duration of the dash is three times longer than that of a dot.

dashpot A delayed-action device in which the movement of a piston is slowed by air or a liquid in a closed cylinder.

dashpot relay A time-delay relay assembly in which the delay is obtained with a DASHPOT.

DAT **1.** Abbreviation of DIGITAL AUDIO TAPE. **2.** Abbreviation of *diffused-alloy transistor.*

data **1.** A collection of digital bits (binary digits) with informational content (e.g., a computer file, a digital image, or a digital sound recording). **2.** General expression for information, especially in encoded or written form.

data acquisition The reception and gathering of data (see DATA COLLECTION and DATA SYSTEM, **1**).

data-acquisition system A computer or dumb terminal used to gather data from one or more external points.

data analysis display unit A video display peripheral for online data analysis.

data area A computer memory area that holds data only (i.e., one that does not contain program instructions).

data bank A data file stored in a direct-access storage device, which can be drawn from by many system users through remote terminals.

database **1.** A computer file containing often-used information (e.g., names and addresses, or electronic part numbers). **2.** A popular form of computer software that allows users to create, maintain, and modify information.

data block A set of data bits, comprising an identifiable item.

data bus A conductor or medium over which digital data is transmitted from one place to another within a computer.

data carrier storage A medium of data storage outside of a computer (e.g., a magnetic disk).

data code A set of abbreviations or codes for data characters or words.

data collection The pickup of signals representing test data and their transmission to a computer, data processor, or recorder. Also see DATA SYSTEM, **1**.

datacom Acronym for DATA COMMUNICATION.

data communication The transmission and reception of data signals between or among points in a system.

data communication terminal A computer peripheral providing an input and output link to a central computer system, and that can be used offline for other functions.

data compression **1.** The process of reducing the size of a data file by eliminating redundancies. **2.** The process of minimizing the length of a data transmission by eliminating redundancies. **3.** The process of reducing the bandwidth of a data transmission. **4.** The process of reducing the dynamic amplitude range of a data transmission.

data control The automatic control of incoming and outgoing data in a data processing system.

data conversion The process of changing data from one form to another, e.g., from analog to digital (A/D), digital to analog (D/A), parallel to serial, or serial to parallel.

data converter **1.** A circuit or device for performing DATA CONVERSION. **2.** An analog-to-digital (A/D) converter. **3.** A digital-to-analog (D/A) converter. **4.** A parallel-to-serial converter. **5.** A serial-to-parallel converter.

data description The description of a unit of data, as included in a computer source program.

data display A device, such as a cathode-ray tube (CRT) or liquid-crystal display (LCD), that presents data for visual examination. Compare DATA PRINTOUT.

data element **1.** A component of a data signal (e.g., a number, letter, symbol, or the equivalent electrical pulses). **2.** A device or circuit for acquiring or processing data. **3.** A unit of data (e.g., a field in a file).

data-flow diagram A block diagram showing the movement of data through a data-processing system.

data format The form of data in a record or file (e.g., character format or numerical format).

data gathering See DATA COLLECTION.

data-handling capacity **1.** The amount of data that can be stored in a memory circuit. **2.** The amount of data that can be transmitted over a certain medium. **3.** The rate at which data can be transferred under certain conditions.

data-handling system A system that gathers, routes, transmits, or receives data, but does not necessarily process it.

data item A logical element (character, byte, or bit) describing a characteristic of a record used by a system for which there is a specific application.

data level Descriptive, through a programming language, of the relative weight of logical elements (data items) in a computer record. Also called *data hierarchy.*

data link The portion of a computer system that gathers data and, if necessary, converts it to a form acceptable by a computer.

data matrix Variables and their possible values stored as a series of columns and rows of values in a computer memory.

data name An operand specified in a computer source program.

data pickup **1.** A transducer that collects data signals from a source; it converts nonelectrical data into corresponding electrical signals and delivers its output to a data processing system. **2.** Data acquisition.

data playback The reproduction of data signals stored by some method of data recording.

data plotter See X-Y PLOTTER.

data printout **1.** A device that prints a record of data or the results of a computation. **2.** A permanent printed record, usually of a calculation or computation—especially the printed output of a computer peripheral device.

data processing Work performed on acquired data, as in solving problems, making comparisons, classifying material, organizing files. Usually done by a computer.

data-processing equipment A digital computer and the peripheral equipment needed to collate, store, analyze, and reduce data.

data-processing machine A computer or system used to collate, store, analyze, and reduce data, as opposed to a computer or system used primarily to solve problems or perform routine tasks. Also called *data processor.*

data-processing system An electronic system for automatic data processing. It can be based on analog and/or digital techniques.

data processor See DATA-PROCESSING MACHINE.

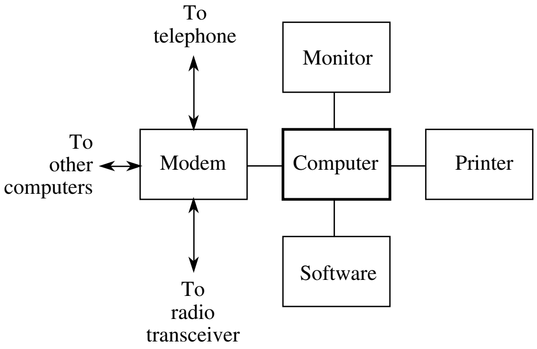

data-processing system

data receiver At a particular point in a data-processing system, a circuit or device for receiving data from a data transmitter.

data reception Receiving data signals from some point within or outside a data-processing system.

data-reception system A data receiver and its associated equipment.

data record A computer-processed record containing a data unit.

data recorder A machine for storing data acquired in the form of electrical signals (see DATA RECORDING).

data recording **1.** The preservation of data signals by some process, such as magnetic-disk encoding, optical-disk encoding, or tape recording, for future use or as a backup. **2.** A record of data signals, as on magnetic tape.

data reduction The summarization of a mass of electronically gathered data.

data-reduction system A system used to minimize the amount of data necessary to convey given information.

data representation Values and data as described by numerals, symbols, and letters (e.g., computer program instructions).

data segment As related to a particular computer process, a subunit of allocated storage containing data only.

data selector/multiplexer A digital circuit that has several or many input signals, and feeds one of them onto a common line.

data set A device that connects a data processor to a telegraph or telephone line.

data signal **1.** A signal (such as one of binary bit combinations) that can represent data as numbers, letters, or symbols. **2.** A signal current or voltage proportional to some sampled quantity, and that can be used to actuate indicating instruments during tests or measurements.

data statement A computer source program statement identifying a data item and specifying its format.

data storage The preservation of data, particularly computer files, for long periods of time in nonvolatile form (no source of power is required to ensure that the data remains intact).

data storage media Hardware that preserves data, particularly computer files, for long periods of time in nonvolatile form (no source of power is required to ensure that the data remains intact). Common media include magnetic disks, magnetic tape, and optical disks.

data synchronizer A device used to synchronize data transmission within a computing or processing system.

data system **1.** An arrangement for collecting, recording, and routing data in the form of electrical signals. **2.** An arrangement for processing data (i.e., for correlating, computing, routing, storing, etc.).

data terminal A remote input/output device connected to a central computer.

data throughput In a computer system, the amount of data per unit time (bytes, kilobytes, megabytes, gigabytes, or terabytes per second or minute) that can be transferred from one place to another.

data transducer In tests and measurements, a transducer that converts a monitored phenomenon into electrical quantities that can be used for computer analysis or calculations.

data transmission Sending data signals from a pickup point or processing stage to another point within a data-processing system; also, sending such signals to points outside the system.

data-transmission system A data transmitter and its associated equipment.

data transmission utilization measure The ratio of the useful data output of a data-transmission system to the total data input.

data transmitter A circuit or device for sending data from point to point within or outside of a data-processing system.

data unit Characters in a group that are related in a way that makes them a meaningful whole (e.g., a text word, or an object such as a circle in vector graphics).

data value A measure of the amount of information contained in a certain number of data bits. The greater the ratio of the actual information to the number of bits, the higher the data value.

data words In digital computer operations, words (bit groups) representing data, rather than program instructions.

DAVC Abbreviation of DELAYED AUTOMATIC VOLUME CONTROL.

David Phonetic alphabet code word for letter D.

daylight effect The modification of transmission paths during the day because of ionization of the upper atmosphere by solar radiation.

daylight lamp An incandescent lamp whose filament is housed in a blue glass bulb, which absorbs some red radiation and transmits most of the green, blue, and violet. So called because the spectral output resembles that of typical daylight.

daylight range The distance over which signals from a given transmitter are consistently received during the day.

DB **1.** Abbreviation of DIFFUSED BASE of a transistor. **2.** Abbreviation of DOUBLE BREAK (relay).

dB **1.** Abbreviation of DECIBEL or *decibels.* **2.** Symbol for *differential of susceptance.*

dBa Abbreviation of ADJUSTED DECIBELS.

dBc Abbreviation of *decibels referred to the carrier.*

DBD Abbreviation of *double-base diode.*

dBd The power gain of an antenna in the direction of maximum radiation, compared to the radiation in the favored direction of a half-wave dipole in free space receiving the same amount of power. Expressed in decibels.

dBi The power gain of an antenna in the direction of maximum radiation, compared to the radiation from a theoretical isotropic antenna in free space receiving the same amount of power. Expressed in decibels.

dBj The level of an RF signal, in decibels, relative to 1 millivolt.

dBk Abbreviation of DECIBELS REFERRED TO 1 KILOWATT.

DBM Abbreviation of *database management.*

dBm Abbreviation of DECIBELS REFERRED TO 1 MILLIWATT.

dBm0 Signal level in dBm, referred to a zero-transmission level.

dBm0p Noise in dBm0, measured with set phosphometric weighting.

dB meter A usually high-impedance ac voltmeter with a scale reading directly in decibels.

dBmp The level in dBm, measured with phosphometric weighting. Generally equal to dBm –2.5, for a noise level that is flat within the communications audio range.

dBmr Decibels measured with respect to zero transmission level.

dBmV Abbreviation of DECIBELS REFERRED TO 1 MILLIVOLT.

dBrap Abbreviation of DECIBELS ABOVE REFERENCE ACOUSTIC POWER (10^{-6} W).

dBrn Abbreviation for *decibels above reference noise.* A level of 0 dBrn is defined as noise power of 10^{-9} W (1 nanowatt).

dBrnc Noise power in dBrn for a circuit with message weighting c.

dBrnc0 Noise in dBrnc measured with respect to zero transmission level.

dBV Abbreviation of DECIBELS REFERRED TO 1 VOLT.

dBW Abbreviation of DECIBELS REFERRED TO 1 WATT.

dBx Abbreviation of DECIBELS ABOVE REFERENCE COUPLING.

dC Symbol for *differential of capacitance.*

dc **1.** Abbreviation of DIRECT CURRENT. **2.** Abbreviation of *direct-coupled.*

dc-ac converter A circuit that converts a dc input voltage into an ac output voltage, with or without step-up or step-down. Also called INVERTER.

dc alpha The current amplification factor (ALPHA) of a common-base transistor stage for a dc input (emitter) signal. Compare DC BETA.

dc amplifier **1.** A *direct-coupled amplifier.* 2. An amplifier for boosting direct-current signals.

dc balance **1.** Adjustment of a circuit or device for dc stability or dc null. **2.** Adjustment of a circuit for dc stability during gain changes. **3.** A potentiometer or other variable component used to stabilize or null a dc circuit.

dc bar See DC BUS.

dc base current Symbol, $I_{B(dc)}$. The static direct current in the base element of a bipolar transistor.

dc base resistance Symbol, $R_{B(dc)}$. The static dc resistance of a bipolar transistor's base element; $R_{B(dc)} = V_B/I_B$.

dc base voltage Symbol, $V_{B(dc)}$. The static dc voltage at the base element of a bipolar transistor.

dc beta The current amplification factor (BETA) of a common-emitter-connected transistor for a dc input (base) signal. Compare DC ALPHA.

dc block A coaxial section that has a capacitance in series with the inner or outer conductor, or both, to block dc while passing RF. Compare DC SHORT.

dc bus A supply conductor carrying direct current only.

dcc Abbreviation of double cotton covered (wire).

dc cathode current Symbol, $I_{K(dc)}$. The static direct current in the cathode element of an electron tube.

dc cathode resistance Symbol, $R_{K(dc)}$. The static dc resistance of the cathode path of an electron tube.

dc cathode voltage Symbol, $V_{K(dc)}$. The static dc voltage at the cathode of an electron tube.

dc circuit breaker A circuit breaker operated by direct-current overload or underload, depending on its design and application.

dc collector current Symbol, $I_{C(dc)}$. The static direct current in the collector element of a bipolar transistor.

dc collector resistance Symbol, $R_{C(dc)}$. The static dc resistance of a bipolar transistor's collector element; $R_{C(dc)} = V_C/I_C$.

dc collector voltage Symbol, $V_{C(dc)}$. The static dc voltage at the collector element of a bipolar transistor.

dc component In a complex wave (i.e., one containing both ac and dc), the current component having an unchanging polarity. The dc component constitutes the mean (average) value around which the ac component alternates, pulsates, or fluctuates.

dc converter A dynamoelectric machine for converting low-voltage dc into higher-voltage dc. It is essentially a low-voltage dc motor coupled me-

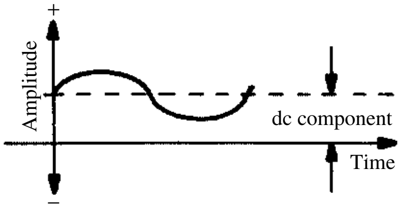

dc component

chanically to a higher-voltage dc generator. Compare DC INVERTER.

dc coupling See DIRECT COUPLING.

dc drain current Symbol, $I_{D(dc)}$. The static direct current in the drain element of a field-effect transistor.

dc drain resistance Symbol, $R_{D(dc)}$. The static dc resistance of an FET's drain element; $R_{D(dc)} = V_D/I_D$.

dc drain voltage Symbol, $V_{D(dc)}$. The static dc voltage at the drain element of a field-effect transistor.

dc dump In digital computer operation, removing dc power from a computer, which would eradicate material stored in a volatile memory.

dc emitter current Symbol, $I_{E(dc)}$. The static direct current in the emitter element of a bipolar transistor.

dc emitter resistance Symbol, $R_{E(dc)}$. The static dc resistance of a bipolar transistor's emitter element; $R_{E(dc)} = V_E/I_E$.

dc emitter voltage Symbol, $V_{E(dc)}$. The static dc voltage at the emitter element of a bipolar transistor.

dc equipment Apparatus designed expressly for operation from a dc power supply. Compare AC EQUIPMENT and AC/DC.

dc erase head In a magnetic recorder, a head supplied with a dc current for the purpose of removing data.

dc error voltage In a television receiver, the dc output of the phase detector, which is used to control the frequency of the horizontal oscillator.

dc gate current Symbol, $I_{G(dc)}$. The very small static direct current in the gate element of a field-effect transistor.

dc gate resistance Symbol, $R_{G(dc)}$. The very high, static dc resistance of an FET's gate element; $R_{G(dc)} = V_G/I_G$.

dc gate voltage Symbol, $V_{G(dc)}$. The static dc voltage at the gate element of a field-effect transistor.

dc generator **1.** A rotating machine (dynamo) for producing direct current. Also see DYNAMOELECTRIC MACHINERY. **2.** Generically, a device that produces direct current: batteries, photocells, thermocouples, etc.

dc generator amplifier A special type of generator that provides power amplification. The input signal energizes the field winding of a constant-speed machine; because the output voltage is proportional to field flux and armature speed, a high output voltage is obtained. Also see AMPLIDYNE.

dc grid bias Steady dc control-grid voltage used to set the operating point of an electron tube.

dc grid current Symbol, $I_{G(dc)}$. The static direct current in the control-grid element of an electron tube.

dc grid resistance Symbol, $R_{G(dc)}$. The static dc resistance in the control-grid element of an electron tube; $R_{G(dc)} = V_G/I_G$.

dc grid voltage Symbol, $V_{G(dc)}$. The static dc voltage at the control grid of an electron tube.

dc inserter In a television transmitter, a stage that adds the dc pedestal (blanking) level to the video signal.

dc inverter An electrical, electronic, or mechanical device that converts dc to ac. Also called INVERTER.

dcl Abbreviation of *dynamic load characteristic.*

dc leakage The unintended flow of direct current.

dc leakage current **1.** The direct current that normally passes through a correctly polarized electrolytic capacitor operated at its rated dc working voltage. **2.** The zero-signal reverse current in a semiconductor pn junction.

DCM Abbreviation of DIGITAL CAPACITANCE METER.

D/CMOS Combination of DMOS and CMOS on a monolithic chip.

dc motor A motor that operates from direct current only.

dc noise Noise heard during the playback of magnetic tape that was recorded while direct current was in the record head.

dc noise margin In a digital or switching circuit, the difference $V_o - V_i$, where V_o is the output-voltage level of a driver gate and V_i is the input threshold voltage of a driven gate.

dc operating point For a bipolar transistor, field-effect transistor, or vacuum tube, the static, zero-signal dc voltage and current levels.

dc overcurrent relay A relay or relay circuit actuated by dc coil current rising above a specified level. Compare DC UNDERCURRENT RELAY.

dc overvoltage relay A relay or relay circuit actuated as a result of the dc coil voltage rising above a specified level. Compare DC UNDERVOLTAGE RELAY.

dc patch bay A patch bay in which the dc circuits of a system are terminated.

dc picture transmission In television, transmission of the dc component of the video signal; this component corresponds to the average illumination of the scene.

dc plate current Symbol, $I_{P(dc)}$. The static direct current in the plate element of an electron tube.

dc plate resistance Symbol, $R_{P(dc)}$. The static dc resistance of the internal plate-cathode path of an electron tube; $R_{P(dc)} = V_P/I_P$.

dc plate voltage Symbol, $V_{P(dc)}$. The static dc voltage at the plate electrode of an electron tube.

dc positioning Alignment of the spot on the screen of an oscilloscope tube, by means of adjustable dc voltages applied to the horizontal and vertical deflecting plates.

dc power Symbol, P_{dc}. Unit, watt. The power in a dc circuit; $P_{dc} = EI$, where E is in volts and I is in amperes. Compare AC POWER. Also see POWER.

dc power supply A power unit that supplies direct current only. Examples: battery, transformer/rectifier/filter circuit, dc generator, and photovoltaic cell. Compare AC POWER SUPPLY.

dc relay A relay having a simple coil and core system for closure by direct current, which can be rectified ac.

dc resistance Resistance offered to direct current, as opposed to in-phase ac resistance.

dc resistivity The resistivity of a sample of material measured using a pure dc voltage under specified conditions (physical dimensions, temperature, etc.).

dc restoration The reinsertion of the dc component into a signal from which the component has been extracted through a capacitor or transformer.

dc restorer A circuit that reinserts the average dc component of a signal after the component has been lost because the signal passed through a capacitor or transformer.

DCS Abbreviation of DORSAL COLUMN STIMULATOR.

dc shift A shift in the DC OPERATING POINT.

dc short A coaxial fitting providing a dc path between the center and outer conductors, while permitting radio-frequency (RF) current to flow easily through the coaxial section. Compare DC BLOCK.

dc signaling A signaling procedure that uses direct current as the medium (e.g., simple wire telegraphy or telephony).

dc source **1.** DC GENERATOR. **2.** A live circuit point from which one or more direct currents can be taken.

dc source current Symbol, $I_{S(dc)}$. The static direct current in the source element of a field-effect transistor.

dc source resistance Symbol, $R_{S(dc)}$. The static dc resistance of an FET's source element.

dc source voltage Symbol, $V_{S(dc)}$. The static dc voltage at the source element of a field-effect transistor.

DCTL Abbreviation of DIRECT-COUPLED TRANSISTOR LOGIC.

dc-to-dc inverter See DC INVERTER.

dc transducer **1.** A transducer that depends on direct current for its operation (i.e., it has a dc power supply whose output is modulated by the sensed phenomenon). **2.** A transducer that converts a direct current into some other form of energy, such as heat, pressure, or sound.

dc transformer A dc-to-dc converter providing voltage step-up. The applied dc is usually first converted to ac, which is then stepped up by a transformer. The higher-voltage ac is then rectified to produce a high dc output voltage.

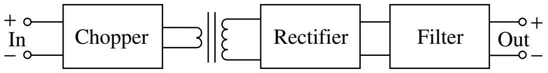

dc transformer

dc transmission **1.** Sending dc power from a generating point to a point of use. **2.** In television transmission, the retention of the dc component in the video signal.

dc tuning voltage The capacitance-varying dc voltage applied to a varactor in an inductance-capacitance (LC) tuned circuit.

dcu Abbreviation of decimal counting unit.

dc undercurrent relay A relay or relay circuit that is actuated as a result of the dc coil current dropping below a specified level. Compare DC OVERCURRENT RELAY.

dc undervoltage relay A relay or relay circuit that is actuated as a result of the dc voltage dropping below a specified level. Compare DC OVERVOLTAGE RELAY.

dcv Abbreviation of DC VOLTS or DC VOLTAGE.

dc voltage Abbreviation, dcv. A voltage that does not change in polarity, an example being the voltage delivered by a battery or dc generator. Also see VOLTAGE.

dc working voltage Abbreviation, dcwv. The rated dc voltage at which a component can be operated continuously with safety and reliability.

dc working volts Abbreviation, dcwv. The actual value, expressed in volts, of a DC WORKING VOLTAGE.

dcwv Abbreviation of DC WORKING VOLTAGE.

dD Symbol for *differential of electric displacement.*

DDA Abbreviation of *digital differential analyzer.*

DDD Abbreviation of DIRECT DISTANCE DIALING (telephone).

D display See D SCOPE.

DE Abbreviation of *decision element.*

dE Symbol for *differential of voltage.*

deac In frequency-modulation (FM) receivers, a device used for deemphasis. The name is short for *deaccentuator.*

deactuating pressure For an electrical contact, the pressure at which contact is made or broken as the pressure reaches the level of activation.

dead **1.** Unelectrified. **2.** Lacking electromagnetic signals or fields. **3.** Electrically or mechanically inoperative.

dead band **1.** A radio-frequency band on which no signals are heard. **2.** A range of values for which an applied control quantity (e.g., current or voltage) has no effect on the response of a circuit.

deadbeat The state wherein a moving body (such as the pointer of a meter or the voice coil of a loudspeaker) comes to rest without overswing or oscillation.

deadbeat galvanometer See DEADBEAT INSTRUMENT.

deadbeat instrument A meter or recorder that is highly damped to ensure that overswing or oscillation does not occur.

deadbeat meter See DEADBEAT INSTRUMENT.

dead break An unreliable contact of a relay, caused by insufficient pressure.

dead circuit A circuit that is electrically disabled.

dead end The unused end of a tapped coil (i.e., the turns between the end of the coil and the last turn used).

dead-end tower A supporting tower for an antenna or transmission line that can withstand stresses caused by loading or pulling.

dead file A computer file that is not in use, but is being kept in a record.

dead front panel A metal panel that, for safety and desensitization, is completely insulated from voltage-bearing components mounted on it; it is often grounded.

dead interval See DEAD TIME.

dead line A deenergized line or conductor.

dead period See DEAD TIME.

dead room An anechoic room in which acoustic tests and studies are made.

dead short A short circuit with extremely low (virtually no) resistance from dc into the radio-frequency spectrum.

dead space See DEAD BAND.

dead spot **1.** An area in which radio waves from a particular station are not received. **2.** On a vacuum-tube cathode (directly or indirectly heated), a spot from which no electrons are emitted.

dead stretch The tendency of insulating materials to permanently retain their approximate dimensions after having been stretched.

dead time **1.** DOWN TIME. **2.** An interval during which there is no response to an actuating signal. **3.** In a computer system, an interval between related events that is allocated to prevent interference between the events.

dead volume In a pressure transducer, the zero-stimulus volume of the pressure port cavity.

dead zone See ZONE OF SILENCE.

debatable time Computer time that cannot be placed in any other category.

debounced switch A switch in sensitive computer or control systems that has circuitry for eliminating the electrical effects of bounce (see BOUNCE, **1**).

de Broglie waves Electromagnetic waves that are believed to be associated with moving particles (such as electrons, protons, and neutrons).

debug **1.** To eliminate errors in, and maximize the efficiency of, a computer program or group of programs. **2.** To optimize the design and construction of electronic equipment.

debugging A process by which engineers eliminate the flaws in a circuit, machine, or computer program.

debugging aid routine A computer program used to test other programs.

debugging period The time interval following completion of a software design, a hardware interconnection, or the manufacture of a piece of electronic equipment, during which errors and imperfections are sought and corrected.

debunching In a velocity-modulated tube, such as a Klystron, a beamspreading space-charge effect that destroys electron bunching.

Debye length The maximum distance between an electron and a positive ion over which the electron is influenced by the field of the ion.

Debye shielding distance See DEBYE LENGTH.

deca- A prefix that indicates multiplication by 10.

decade **1.** A frequency band whose upper limit is 10 times the lower limit. Example: 20 Hz to 200 Hz. **2.** A set of 10 switched or selectable components in which the total value is 10 times that of individual values. Example: a *decade capacitor.* Also called DECADE BOX. **3.** A group, sometimes a unit of access, of 10 computer storage locations.

decade amplifier An amplifier or preamplifier whose gain can be adjusted in increments of 10 (×1, ×10, ×100, etc.).

decade box A group of components that provides values in 10 equal steps selected by a switch or jacks. For compactness, the components and the associated hardware are enclosed in a box or can. See, for example, DECADE CAPACITOR.

decade capacitor A composite capacitor whose value is variable in 10 equal steps. For example, the values might be set at 100 picofarads (pF), 200 pF, 300 pF, etc., up to 1000 pF. Compare DECADE INDUCTOR and DECADE RESISTOR.

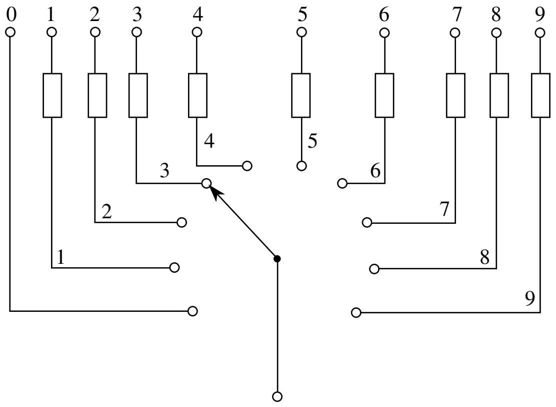

decade box

decade counter A counter (see COUNTER, **1**, **2**) in which the numeric display is divided into sections, each having a value 10 times that of the next and displaying a digit from zero to nine.

decade inductor An inductor whose value is variable in 10 equal steps. Compare DECADE CAPACITOR and DECADE RESISTOR.

decade resistor A resistor whose value is variable in 10 equal increments. Compare DECADE CAPACITOR and DECADE INDUCTOR.

decade scaler A scale-of-10 electronic counter (i.e., a circuit delivering one output pulse for each group of 10 input pulses).

decametric waves Waves in the 10- to 100-meter band (30 to 3 MHz).

decay **1.** The decrease in the value of a quantity, e.g., current decay in a resistance-capacitance circuit. **2.** The gradual, natural loss of radioactivity by a substance.

decay characteristics **1.** The decay of a parameter; usually an exponential function. **2.** The persistence time in a storage oscilloscope.

decay curve A curve, usually logarithmic, representing the function of quantity versus time for a signal decrement, the decrement of radioactivity, or other natural process.

decay rate A quantitative expression for the rapidity with which a quantity decreases. Generally listed in decibels per second (dB/s) or decibels per millisecond (dB/ms).

decay time The time required for pulse amplitude to fall from 90% to 10% of the peak value. Also called FALL TIME.

Decca A 70- to 130-kHz CW radio navigation system (British).

decelerated electron A high-speed electron that is abruptly decelerated upon striking a target, causing X-rays to be emitted.

decelerating electrode A charged electrode that slows the electrons in an electron beam.

deceleration Acceleration that results in a decrease in speed.

deceleration time **1.** The time taken by magnetic tape to stop moving after the last recording or playback has finished. **2.** The time taken by a mechanical data storage medium, such as a hard disk, to come to rest after completion of a read or write operation, or on powering-down.

decentralized data processing Data processing in which the computing equipment is distributed among managerial subgroups.

deception A method of producing misleading echoes in enemy radar.

deception device A radar device, or radar-associated device, for deception.

deci- Abbreviation, d. A prefix meaning one-tenth (10^{-1}). Examples: DECIBEL, DECIMETER.

decibel Abbreviation, dB. A practical unit of relative gain. In terms of power, the relative gain in decibels is equal to:

$$\text{Gain (dB)} = 10 \log_{10}(P_{out}/P_{in}),$$

where P_{out} is the output power and P_{in} is the input power. For voltage, if the input and output impedances are the same, the gain in decibels is given by:

$$\text{Gain (dB)} = 20 \log_{10}(V_{out}/V_{in}),$$

where V_{out} is the output voltage and V_{in} is the input voltage. For current, if the input and output impedances are the same, the gain in decibels is given by:

$$\text{Gain (dB)} = 20 \log_{10}(I_{out}/I_{in}),$$

where I_{out} is the output current and I_{in} is the input current. Losses are indicated by negative dB gain values.

decibels above reference acoustic power Abbreviation, dBrap. The ratio of a given acoustic power level to a lower reference acoustic power level, specified in decibels.

decibels above reference noise Abbreviation, dBrn. The ratio of the noise level at a selected point in a circuit to a lower reference noise level, in decibels.

decibels referred to 1 millivolt Abbreviation, dBmV. The relative voltage level of a signal when compared with a 1-mV signal measured at the same terminals.

decibels referred to 1 milliwatt Abbreviation, dBm. The ratio, in decibels, of an applied power level to the power level of 1 mW.

decibels referred to 1 volt Abbreviation, dBV. The ratio, in decibels, of a given voltage to 1 V, expressed in decibels.

decibels referred to 1 watt Abbreviation, dBW. The ratio of a given power level to the power level of 1 W, expressed in decibels.

decider See DECISION ELEMENT.

decigram A unit of mass equal to 0.1 gram.

deciliter A unit of volume equal to 0.1 liter, or 10^{-4} cubic meter.

decilog A unit equal to 0.1 times the common logarithm of a ratio.

decimal **1.** Pertaining to the base-10 number system (see DECIMAL NUMBER SYSTEM). **2.** A base-10 numerical fraction, represented by figures to the right of the radix point (decimal point), and arranged serially according to negative powers of 10. Examples: $0.12 = 1.2 \times 10^{-1}$, $0.00135 = 1.35 \times 10^{-3}$.

decimal attenuator An attenuator circuit whose resistances are chosen for attenuation in decimal steps. Thus, one section provides attenuation in steps of 0.1 times the applied voltage, another in steps of 0.01 times the applied voltage, another in steps of 0.001 times the applied voltage, etc.

decimal code A method of defining numbers, in which each place has a value of ten times that immediately to the right.

decimal-coded digit **1.** A numeral from 0 to 9. **2.** A numeral in the DECIMAL NUMBER SYSTEM. **3.** A binary representation of a decimal value from 0 to 9.

decimal digit A numeral from 0 to 9.

decimal equivalent The decimal number equal to a given fraction (e.g., the decimal equivalent of $^{21}/_{64}$ is 0.3281).

decimal fraction See DECIMAL, 2.

decimal notation See DECIMAL NUMBER SYSTEM.

decimal number system The familiar base-10 or radix-10 number system, in which the digits 0 through 9 represent values according to their position, relative to the decimal point (also called the *radix point*). Positions to the left of the point represent successive positive powers of 10, and those to the right represent successive negative powers of 10.

decimal point The radix point in a decimal number. It serves to separate the integral part from the fractional part of the number.

decimeter waves See MICROWAVES.

decimetric waves Electromagnetic waves having lengths ranging from 0.1 meter to 1 meter (3000 MHz to 300 MHz). Also known as *ultrahigh frequency (UHF).*

decineper A natural-logarithmic unit equal to 0.1 *neper.*

decipher See DECODING, **3**.

decision **1.** A choice based on the evaluation and comparison of data, and the identification of a specified objective. **2.** In digital computer operations, the automatic selection of the next step in a sequence, on the basis of data being compared by a relational test.

decision box A block on a computer flowchart indicating the point at which a decision (see DECISION, **2**) must be made as to which of several branches the program will take.

decision elements See LOGIC CIRCUITS.

decision instruction A computer program instruction to compare the values of operands and take an appropriate action, as per the BASIC instruction "IF A = B THEN GO TO (line number)."

decision procedure In decision theory, a series of calculations made to optimize the speed or efficiency of a process, or to minimize risk, failure, cost, etc.

decision theory A statistical discipline concerned with identifying and evaluating choices and alternatives, and determining the best sequence of steps to take in reaching an objective.

decision tree In decision theory, a diagram showing alternative choices, so called from its resemblance to a tree with branches.

decision value A value that defines the boundary between two intervals in the encoding process.

deck **1.** See TAPE DECK. **2.** A pack of punched cards in a computer file.

declarative macroinstruction As part of an assembly language, instructions to the compiler to do something or record a condition without affecting the object program.

declarative statement A computer source program instruction specifying the size, format, and kind of data elements and variables in a program for a compiler.

declination **1.** The angle representing the deviation of magnetic north from true north; it is the angle subtended by a freely turning magnetic needle and the meridian. Compare INCLINATION. **2.** Celestial latitude.

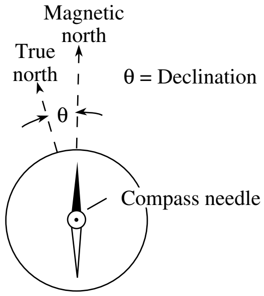

declination, 1

declinometer An instrument for measuring declination.

decode **1.** To unscramble a coded message. **2.** In digital computer operations, to deliver a specific output from character-coded inputs. **3.** In a multiplex system, the separation of the subcarrier from the main carrier.

decoder A circuit or device for performing DECODING.

decoder/demultiplexer A circuit that places an input signal on a selected output line.

decoder/driver An integrated circuit containing a decoder and driver.

decoding **1.** In computer and data-processing operations, DIGITAL-TO-ANALOG CONVERSION. **2.** The conversion to English of a message received in a code. **3.** Translating a message from a secret code (i.e., deciphering a message). **4.** The automatic conversion of a signal into the appropriate switching action (as the enabling of a transmitter or receiver by a tone in a selective calling system).

decoding circuit A circuit intended for the purpose of translating a code into ordinary language.

decollator An offline computer device for separating the parts of output continuous stationery sets. Also see CONTINUOUS STATIONERY.

decommutation The extraction of a signal component from the composite signal, resulting from commutation.

decommutator A circuit or device for performing decommutation, including demodulators, demultiplexers, and signal separators.

decoupler A device that isolates two circuits so that a minimal amount of coupling exists between them.

decoupling The elimination or effective minimization of coupling effects, as in decoupling amplifier stages to prevent interaction through a common power-supply lead.

decoupling capacitor **1.** A capacitor that provides a low-impedance path to ground to prevent undesired stray coupling among the circuits in a system. **2.** The capacitive member of a resistance-capacitance (RC) decoupling filter.

decoupling filter A resistance-capacitance (RC) filter, usually inserted into a common dc line in a multistage amplifier to prevent interstage feedback coupling through the common impedance of the line.

decoupling network One or more decoupling filters.

decoupling resistor The resistive member of a resistance-capacitance (RC) decoupling filter.

decoy In radar, an object that provides misleading reflections. Also see CHAFF.

decreasing function A function whose curve has a negative slope at all points in the domain.

decrement **1.** Also called *logarithmic decrement.* The rate at which a damped wave dies down. The decrement value is the natural (base-*e*) logarithm of the ratio of two successive peaks of the same polarity. **2.** A quantity used to lessen the value of a variable. **3.** To lower the value (of a register, for example) by a single increment.

decremeter An instrument for measuring the decrement of a radio wave.

decremeter capacitor A variable capacitor for use in a decremeter. The rotor plates are shaped so that equal angular rotations correspond to the same decrement at all settings. Thus, the percentage of capacitance change for a given angle of rotation is constant throughout the capacitance range.

decryption The conversion of an encrypted signal from a cipher into plain text, graphics, or other commonly recognizable form. Also see CIPHER. Compare ENCRYPTION.

decryption key An algorithm, or a set of algorithms, that converts an encrypted signal from a cipher into plain text, graphics, or other commonly recognizable form. Each cipher has its own unique algorithm or set of algorithms for this purpose. The signal cannot be decrypted unless all the components of the key are present.

dedicated Assigned exclusively to a certain purpose [e.g., a dedicated facsimile (fax) line].

deductive logic A form of symbolic logic used to demonstrate that a certain conclusion will always follow, given a certain set of circumstances. The logic of digital circuits is deductive. Compare INDUCTIVE LOGIC.

dee In a cyclotron, one of the D-shaped chambers in and between which particles accelerate in a spiral path to high velocity.

dee line In a cyclotron, a support for the dee, with which it forms a resonant circuit.

deemphasis In frequency modulation, the introduction of a low-pass characteristic (response falls as modulating frequency increases) to complement the rising response of preemphasis. Also called *postemphasis* or *postequalization.* Compare PREEMPHASIS.

deemphasis amplifier An amplifier used to remove the high-frequency preemphasis applied to signals prior to broadcasting, multiplexing, tape recording, or telemetering. Also see DEEMPHASIS and PREEMPHASIS.

deemphasis circuit A low-pass filter that provides deemphasis in an FM receiver.

deemphasis network See DEEMPHASIS CIRCUIT.

deenergize To take a circuit or device out of operation (i.e., to remove its power or signal excitation).

deep cycle Pertaining to a rechargeable cell or battery that can operate until it is almost completely discharged. It generally has a high ampere-hour capacity.

deep-diffused junction A pn junction made by diffusing the impurity material deep in the semiconductor wafer. Compare SHALLOW-DIFFUSED JUNCTION.

deep discharge The nearly complete discharge of a cell or battery; usually done prior to recharging.

deep-space net A radar system intended for constant monitoring of spacecraft.

defeating **1.** The disabling or circumvention of an alarm or security system, leaving the protected property vulnerable to intrusion. **2.** The dangerous, and potentially lethal, disabling of a safety device in an electrical or electronic system.

defect **1.** Absence of an electron (hence, presence of a hole) in the lattice of a semiconductor crystal. **2.** An abnormality of design, construction, or performance of an electronic circuit or device. **3.** In a computer system, a hardware or software fault that could be the eventual cause of a failure. **4.** A flaw in a crystalline substance.

defect conduction In a semiconductor material, conduction via holes.

deferred addressing Indirect addressing in which a preset counter makes several references to find a desired address.

deferred entry An entry into a computer subroutine, delayed because of a delay in the exit from a control program.

deferred exit An exit from a computer subroutine, delayed because of a particular command.

defibrillation Use of a CARDIAC STIMULATOR to halt fibrillation of the heart, as caused by electric shock.

defibrillator See CARDIAC STIMULATOR.

definite-purpose component A component designed for a specific use, rather than for a wide range of possible applications (e.g., a video detector diode, as opposed to a general-purpose diode). Compare GENERAL-PURPOSE COMPONENT.

definition **1.** Clarity of a video image (i.e., one having good contrast and faithful tones). **2.** Good intelligibility of reproduced sounds.

deflecting coil One of a set of external coils carrying sawtooth currents, which provide electromagnetic deflection of the cathode-ray beam in picture tubes, camera tubes, radar display tubes, sonar display tubes, and some oscilloscopes. Also called *deflection coil.*

deflecting electrode An electrode, such as a deflecting plate, used to alter the direction an electron beam. Also called *deflection electrode.*

deflecting plate In a cathode-ray tube, a plate that attracts or repels the electron beam, causing the spot to move horizontally or vertically on the screen. Also called *deflection plate.*

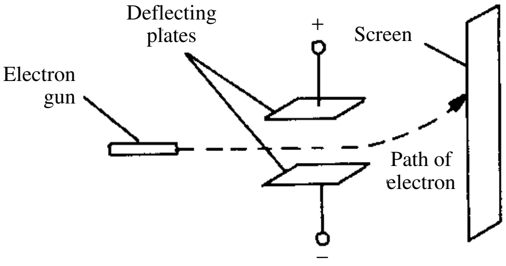

deflecting plate

deflecting torque The torque required to move the pointer of a meter, or the pen or mirror of a recorder.

deflection **1.** In a cathode-ray tube, movement of the electron beam by electric or magnetic fields. **2.** Movement of the pointer of a meter or the pen or mirror of a recorder by an applied current or voltage.

deflection factor Symbol, G. The reciprocal of DEFLECTION SENSITIVITY.

deflection plane In a cathode-ray tube, the plane perpendicular to the axis of the tube. This plane contains the electromagnetic and/or electrostatic lines of flux that result in deflection of the electron beam.

deflection coil See DEFLECTING COIL.

deflection electrode See DEFLECTING ELECTRODE.

deflection plate See DEFLECTING PLATE.

deflection polarity In a cathode-ray tube, the polarity of the voltage applied to a particular deflecting plate to move the electron beam in a particular direction.

deflection sensitivity Symbol, S. A quantitative measure of the extent to which the input voltage will displace the electron beam on the screen of an electrostatic cathode-ray tube. Expressed in volts per centimeter (V/cm) or volts per inch (V/in).

deflection voltage The potential difference between the deflection plates of an electrostatic cathode-ray tube. It is used to control the direction of the electron beam striking the phosphor screen.

deflection yoke An assembly of deflection coils in picture and camera tubes, and in some magnetically deflected oscilloscope tubes. The usual combination is two series-connected horizontal deflection coils and two series-connected vertical deflection coils.

deflector **1.** A beam-forming plate in a beam-power tube. **2.** A deflection plate in a cathode-ray tube. **3.** A deflection coil or yoke in a picture tube, camera tube, or magnetic-deflection oscilloscope tube. **4.** A mechanical attachment for improving the angle of radiation of a loudspeaker by spreading the higher-frequency waves.

defocusing Blurring of the image on the screen of a cathode-ray tube, caused by spreading of the electron beam.

deformation potential The voltage generated when a crystal lattice is subjected to pressure. An example is the voltage produced by a crystal microphone when acoustic waves strike the crystal.

defruiting The elimination of non-synchronized echoes in a radar system.

deg Abbreviation of DEGREE.

degassing During the evacuation of a vacuum tube or similar device, the removal of gas, including that which has bonded to the glass and metal parts.

degauss See DEMAGNETIZE.

degausser **1.** A circuit that performs DEGAUSSING. **2.** A device for bulk erasing magnetic tape; also called a *bulk tape eraser.*

degaussing **1.** The demagnetization of an object; in particular, the removal of all residual magnetism. **2.** The erasure of data from a magnetic or magneto-optical data-storage medium.

degaussing circuit In a color television receiver, a circuit including a thermistor, voltage-dependent resistor, and coil for automatically demagnetizing the picture tube when the receiver is switched on.

degaussing coil A coil carrying an alternating current; the resulting magnetic field demagnetizes objects that have become accidentally magnetized.

degeneracy In microwave practice, the appearance of a single resonant frequency for two or more modes in a resonator.

degenerate modes In microwave operations, a set of modes with the same resonant frequency or propagation constant.

degenerate parametric amplifier An inverting parametric amplifier, in which the two signals are of the same frequency, which is half the pump frequency.

degenerate semiconductor A semiconductor that behaves like a metal over a wide range of temperatures.

degeneration In an amplifier, the technique of feeding a portion of the output back to the input out of phase with the input signal, to improve fidelity at the expense of gain. Also called *negative feedback* or *inverse feedback.* Compare REGENERATION.

degenerative resistor An unbypassed emitter resistor in a common-emitter bipolar-transistor circuit, or an unbypassed source resistor in a common-source field-effect transistor circuit. Signal current flowing through the resistor produces negative feedback current (degeneration), which reduces the gain of the stage, but increases the linearity of the transfer characteristic.

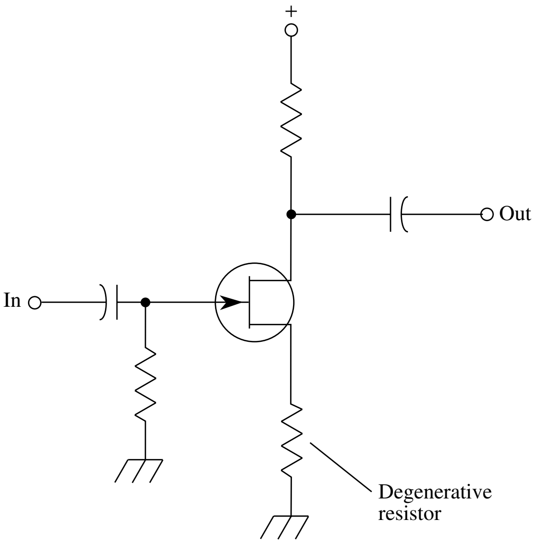

degenerative resistor

degradation **1.** Gradual deterioration in the condition or performance of a circuit or device. **2.** In a computer system, compromised performance caused by component failure.

degradation failure Failure occurring at the terminal point of degradation.

degraded operation See DEGRADATION.

degreaser See ULTRASONIC CLEANING TANK.

degree **1.** A unit of circular angular measurement equal to $\frac{1}{360}$ of the circumference of a circle. Also called GEOMETRIC DEGREE. **2.** A unit of temperature measurement. See DEGREE ABSOLUTE, DEGREE CELSIUS, DEGREE CENTIGRADE, DEGREE FAHRENHEIT, and DEGREE REAUMUR.

degree absolute Symbol, K. The unit of temperature on the absolute scale. Also see ABSOLUTE SCALE.

degree Celsius Symbol, °C. The unit of temperature on the CELSIUS SCALE.

degree centigrade Symbol, °C. The unit of temperature on the *centigrade scale* (now called CELSIUS SCALE).

degree Fahrenheit Symbol, °F. The unit of temperature on the FAHRENHEIT SCALE.

degree of current rectification For a rectifier, the ratio of the average direct output current to the root-mean-square (rms) alternating input current.

degree of voltage rectification For a rectifier, the ratio of the average direct-current (dc) output voltage to the root-mean-square (rms) alternating-current (ac) input voltage.

degree Reaumur Symbol, °R. The unit of temperature on the REAUMUR SCALE.

degrees of freedom **1.** The ways in which a point can move or a system can change. In three-dimensional space, a rigid body has six degrees of freedom: motion in three linear directions, and rotation around three linear axes extending through its center. **2.** The ways in which a robot arm can move, including linear motion and rotational motion.

degrees of rotation A measure of the extent to which a robot joint, or a set of robot joints, can be turned. Some reference axis is always used; angles are specified in degrees, relative to that axis.

degrees-to-radians conversion The conversion of angles in degrees to angles in radians. To change degrees to radians, multiply degrees by 0.01745. Compare RADIANS-TO-DEGREES CONVERSION.

deion circuit breaker A circuit breaker in which the arc occurring when the contacts open is quickly extinguished by an external magnetic device.

deionization The conversion of an ionized substance, such as a gas, to a neutral (non-ionized) state. The process changes the ions into uncharged atoms.

deionization potential The voltage at which an ionized substance becomes deionized; for example, the voltage at which a glow discharge is extinguished when the gas ions become neutral atoms at that voltage. Also called *extinction potential.*

deionization time The time required for an ionized gas to become neutral after the removal of the ionizing voltage.

deionization voltage See DEIONIZATION POTENTIAL.

deionize To restore to an electrically neutral condition (i.e., to convert ions to neutral atoms, as in the deionization of the gas when the discharge in a glow tube is extinguished).

deka- A prefix meaning ten(s) (e.g., DEKAMETER).

dekahexadecimal number system See HEXADECIMAL NUMBER SYSTEM.

delamination The splitting apart, in layers, of an insulating material, such as mica or bonded plastic film.

delay **1.** The interval between the instant at which a signal or force is applied or removed and the instant at which a circuit or device subsequently responds in a specified manner. **2.** The time required for a signal to traverse a given medium, such as air, mercury, or quartz.

delay action Response occurring some time after a stimulus has been applied or removed (e.g., the retarded opening of a delayed-dropout relay).

delay circuit **1.** A circuit, such as a resistance-capacitance (RC) or resistance-inductance (RL) combination, that introduces a time delay. **2.** See DELAY LINE.

delay coincidence circuit A coincidence circuit (see AND CIRCUIT) triggered by two pulses, one of which lags behind the other.

delay counter In a digital computer, a device that halts a program run long enough for an operation to be completed.

delay distortion **1.** Distortion resulting from variations in the phase delay of a circuit or device at different points in its frequency range. **2.** In a facsimile (fax) signal, variations in the delay of different frequency components of the signal.

delayed AGC See DELAYED AUTOMATIC GAIN CONTROL.

delayed automatic gain control An automatic gain control circuit that operates only when the signal amplitude exceeds a predetermined threshold level, thus providing maximum amplification of weaker signals.

delayed automatic volume control See DELAYED AUTOMATIC GAIN CONTROL.

delayed break In relay or switch operation, contacts separating some time after the switch has been thrown or the relay deenergized. Compare DELAYED MAKE.

delayed close See DELAYED MAKE.

delayed closure See DELAYED MAKE.

delayed contacts Contacts that open or close at a predetermined instant after their activating signal is applied or removed.

delayed drop-in See DELAYED MAKE.

delayed dropout See DELAYED BREAK.

delayed loop In security applications, a circuit or system that registers an alarm some time after intrusion is first detected. The delay can usually be selected or preadjusted.

delayed make In relay or switch operation, contacts closing some time after the switch has been thrown or the relay has been energized. Compare DELAYED BREAK.

delayed open See DELAYED BREAK.

delayed PPI Plan-position indicating radar having a delayed time base.

delayed pull-in See DELAYED MAKE.

delayed repeater A repeater that receives and stores information, and retransmits the information later, in response to a switching or interrogation signal.

delayed repeater satellite An active communications satellite that acts as a delayed repeater (i.e., it receives and records information at one time and retransmits it at a later time).

delayed sweep **1.** In an oscilloscope or radar, a sweep that starts at a selected instant after the signal under observation has started. **2.** The (usually calibrated) circuit for producing a sweep, as defined in (**1**).

delayed updating Updating a computer record or record set so that the record fields are left unchanged until all other changes attendant to the pertinent event are processed.

delay equalizer A network that corrects DELAY DISTORTION.

delay-frequency distortion Distortion caused by variation of envelope delay within a frequency band.

delay line A device (not always a line) that introduces a time lag in a signal. The lag is the time required for the signal to pass through the device, minus the time necessary for the signal to traverse the same distance through a wire, cable, optical fiber, or free space.

delay-line memory In a digital computer, a memory that uses a delay line, associated input- and output-coupling devices, and an external regenerative-feedback path. Information is kept stored by causing it to recirculate in the line by regeneration.

delay-line register In a digital computer, a register that operates in the manner of a DELAY-LINE MEMORY and has a register length (capacity) of an integral number of words.

delay-line storage See DELAY-LINE MEMORY and DELAY-LINE REGISTER.

delay multivibrator See MONOSTABLE MULTIVIBRATOR.

delay-power product Unit, watt-second. The figure of merit for an integrated circuit (IC) gate. Increasing gate power reduces propagation delay. Also called PROPAGATION DELAY-POWER PRODUCT.

delay relay A relay that opens or closes at the end of a predetermined time interval.

delay switch A switch having delayed make, delayed break, or both.

delay time **1.** The interval between the instant a voltage or current is applied and the instant a circuit or device operates. **2.** In an output pulse, the interval between the instant an ideal pulse is applied to the input of a system and the instant the output pulse reaches 10% of its maximum amplitude. **3.** The time elapsed between the presentation of a pulse to the input of a delay line and the appearance of the pulse at the output.

delay timer **1.** A timer that starts or stops an operation after a prescribed length of time. **2.** A delay relay or switch.

delay unit In a radar system, a circuit for delaying pulses.

delete **1.** To erase or blank out a signal. **2.** The elimination from a computer file of a record or record group. **3.** To remove a computer program from memory or storage.

deletion record In the master file of a digital computer, a new record that causes existing ones to be deleted.

delimiter In digital computer operations, a character limiting a sequence of characters of which it is not itself a member.

Dellinger effect The sudden disappearance of a radio signal as a result of an abrupt increase in atmospheric ionization caused by a solar eruption.

deliquescent material A material that absorbs enough moisture from the air to get wet. For example; calcium chloride, a deliquescent material, is often used to keep electronic equipment dry. Compare HYGROSCOPIC MATERIAL.

delta circuit A three-phase electrical circuit with no common ground.

delta connection A triangular connection of coils or load devices in a three-phase system, so called from its resemblance to the Greek letter delta. Compare WYE-CONNECTION.

delta-matched antenna See WYE-MATCHED IMPEDANCE ANTENNA.

delta-matched impedance antenna See WYE-MATCHED IMPEDANCE ANTENNA.

delta matching transformer In a WYE-MATCHED IMPEDANCE ANTENNA, the fanned-out (roughly delta-shaped) portion of the two-wire feeder at its point of connection to the radiator. It matches the impedance of the feeder to that of the radiator.

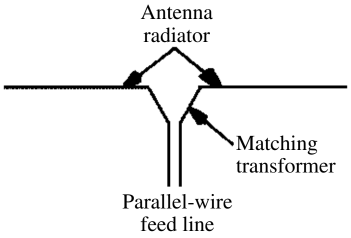

delta-matching transformer

delta modulation The conversion of an analog signal into a digital pulse train that can be decoded to yield the original analog signal.

delta network See DELTA CONNECTION.

delta pulse-code modulation In wire or radio communications, the conversion of an audio signal into a digital pulse train.

delta quantity An increment (i.e., the difference between two values of a variable).

delta rays The emission of secondary electrons as a result of radioactivity.

delta-sigma modulation A method of analog-to-digital conversion. The output is a pulse density function of the input. The input can be obtained by low-pass filtering of the output.

delta tune Also called *receiver incremental tuning (RIT).* In high-frequency (HF) communications transceivers, a control that allows the receiver frequency to be adjusted up to several kilohertz higher or lower than the transmitter frequency.

delta waves Brain waves having a frequency less than 9 Hz. Also see ELECTROENCEPHALOGRAPH and ELECTROENCEPHALOGRAM.

Deluc's pile See DRY PILE.

dem Abbreviation of DEMODULATOR.

demagnetization curve The portion of a magnetic hysteresis curve, showing reduction of demagnetization.

demagnetization effect The phenomenon in which uncompensated magnetic poles at the surface cause a reduction of the magnetic field inside a sample of a material.

demagnetize To remove magnetism from an object, either temporarily or permanently.

demagnetizer See DEGAUSSER.

demagnetizing current The half-cycle of an alternating current (or polarity of a direct current) flowing through a coil wound on a permanent magnet (as in a headphone, permanent-magnet loudspeaker, or polarized relay), that reduces the magnetic field.

demagnetizing force **1.** A magnetic force whose direction reduces the residual induction of a magnetized material. **2.** An effect that reduces the magnetism of a permanent magnet, such as high temperature or a physical blow.

demand factor In the use of electric power, the ratio of the consumer's maximum demand to the actual power consumed.

demand processing Descriptive of a system that processes data as it is available, without storing it.

demarcation strip An interface between a terminal unit and a carrier line.

Dember effect The appearance of a voltage between regions in a semiconductor when one of the regions is illuminated.

demodulation The process of retrieving the information (modulation) from a modulated carrier. In receivers and certain test instruments, this process is called DETECTION.

demodulator **1.** A circuit that recovers the information from a modulated analog or digital signal. In radio communications, such a device is usually called a DETECTOR. **2.** In computer communications, a device that performs ANALOG-TO-DIGITAL CONVERSION of incoming online signals.

demand read (write) Inputting or outputting data blocks to or from a central processor, as needed for processing.

demodulator probe A diode probe that removes the modulation envelope from an applied amplitude-modulated signal, and presents the envelope to a voltmeter or oscilloscope.

demonstrator A device used to show and teach the way in which a component, circuit, or system operates.

DeMorgan's theorem A rule of sequential or digital logic. It states that the negation of (A AND B), for any two statements A AND B, is equivalent to NOT A OR NOT B. Also, the negation of (A OR B) is equivalent, logically, to NOT A AND NOT B.

demultiplexer A circuit or device that separates the components of a multiplexed signal transmitted over a channel.

demultiplexing circuit See DEMULTIPLEXER.

denary band A band in which the highest frequency is 10 times the lowest frequency.

dendrite **1.** The branching (tree-like) structure formed by some materials, such as semiconductors, as they crystallize. **2.** The branching portion of a nerve cell; hence, the corresponding circuit element in the electronic model of such a cell.

dendritic growth **1.** Dendrite (see DENDRITE, **1**). **2.** The process of growing long, flat semiconductor crystals.

dendron See DENDRITE, **2**.

dens Abbreviation of DENSITY.

dense binary code A binary representation system, in which any possible combination of characters is assigned some correspondent.

densitometer An instrument for measuring the density of a body.

density **1.** Mass per unit volume of a material. **2.** Concentration of charge carriers or of lines of flux. **3.** The number of items per unit volume, area, distance, or time.

density modulation Modulation of the density, with respect to time, of electrons in an electron beam.

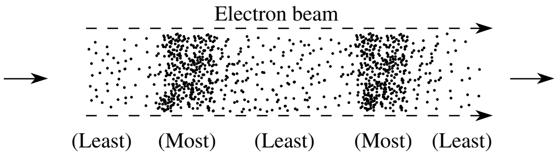

density modulation

density of electrons The concentration of electrons (i.e., the number per unit volume, area, distance, or time).

density packing A figure indicating the quantity of bits per inch or per centimeter, stored on a magnetic tape.

dependent equations Equations that are alike and have an infinite number of solutions. Compare INDEPENDENT EQUATIONS and INCONSISTENT EQUATIONS.

dependent linearity Linearity (especially in its deviation from an ideal slope) as a dependent variable.

dependent variable A changing quantity whose value at any instant is governed by the value at that instant of another changing quantity (the independent variable). Compare INDEPENDENT VARIABLE.

depletion-enhancement-mode MOSFET A metal-oxide-semiconductor field-effect transistor (MOSFET) designed for zero gate-bias voltage. An ac gate signal voltage drives the MOSFET alternately into the depletion mode (negative signal half-cycle) and enhancement mode (positive signal half-cycle). Compare DEPLETION-TYPE MOSFET and ENHANCEMENT-TYPE MOSFET.

depletion field-effect transistor A field-effect transistor whose operation is based on the control of depletion layer width.

depletion layer See BARRIER, **1**.

depletion-layer capacitance See JUNCTION CAPACITANCE.

depletion-layer rectification Rectification provided by a semiconductor junction.

depletion-layer transistor A transistor whose action depends on modulation of current carriers in a space-charge region (depletion layer).

depletion mode Operation characteristic of the DEPLETION-TYPE MOSFET.

depletion region See BARRIER, **1**.

depletion-type MOSFET A metal-oxide-semiconductor field-effect transistor (MOSFET) in which the channel directly under the gate electrode is narrowed by a negative gate voltage (in an n-channel device) or by a positive gate voltage (in a p-channel device).

depolarization **1.** In a primary cell, the removal of the agents that have caused polarization. **2.** The addition of a polarization-inhibiting substance to the electrolyte of a primary cell.

depolarizer A substance that retards polarization in an electrochemical cell. An example is the manganese dioxide used in dry cells.

depolarizing agent See DEPOLARIZER.

deposition The application of a layer of one substance (usually a metal) to the surface of another (the substrate), as in evaporation, sputtering, electroplating, silk-screening, etc.

depth finder See ACOUSTIC DEPTH FINDER.

depth indicator **1.** A sounding instrument for determining the depth of a body of water. **2.** On an ACOUSTIC DEPTH FINDER, the meter that indicates the depth of water.

depth of cut On a phonograph disk, the depth of the recorded groove.

depth of discharge Abbreviation, DOD. In a rechargeable cell or battery, a measure of the extent to which discharging has occurred. It is generally specified as a percentage. For example, if the DOD of a 10-ampere-hour (10-AH) battery is 80 percent, then 8 AH have been used up, and 2 AH remain before recharging will be necessary.

depth of heating In dielectric heating, the depth of heat penetration in the sample when both electrodes are applied to one of its faces.

depth of modulation The degree to which a carrier wave is modulated.

depth of penetration The extent to which a skin-effect current penetrates the surface of a conductor.

depth sounder See ACOUSTIC DEPTH FINDER.

de-*Q* **1.** To reduce the *Q* of a component or tuned circuit. **2.** To inhibit laser action during an interval when an ion population excess is pumped up.

derating To reduce an operating parameter (e.g., current, voltage, power) as another factor (such as temperature) increases, to ensure efficient, reliable, and safe operation.

derating curve A graph that shows the extent to which a quantity (such as allowable power dissipation) must be reduced as another quantity (such as temperature) increases.

derating factor The amount by which a current, power, or voltage must be decreased to ensure safe and efficient operation of a circuit or device in a given environment (temperature, altitude, humidity, etc.). Also see DERATING and DERATING CURVE.

derivative **1.** A mathematical expression indicating the rate at which a function changes, with respect to the independent variable. See DERIVATIVE FUNCTION. **2.** The slope of a line tangent to a curve at a given point. **3.** The output signal of a DIFFERENTIATOR, relative to the input signal.

derivative action In a control system, an action producing a corrective signal proportional to the rate of change (derivative) of the controlled variable.

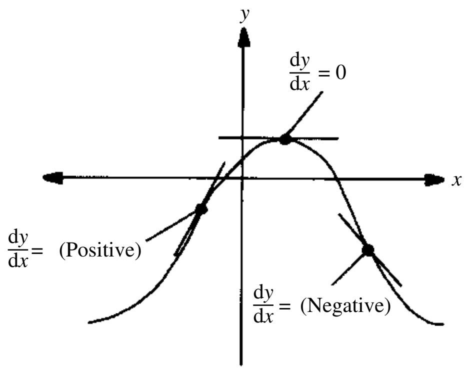

derivative

derivative control A method of automatic control, actuated according to the number of errors per second.

derivative function For a mathematical function $f(x)$, the function $f'(x) = \mathrm{d}f(x)/\mathrm{d}x$, over the domain of f. For any specific point x_0 in the domain of f, the value of $f'(x_0)$ is equal to the slope of a line tangent to f at the point $(x_0, f(x_0))$.

derived center channel The sum or difference of the left and right channels in a stereophonic system.

Dershem electrometer A variation of the *quadrant electrometer.* In the Dershem instrument, the needle (to which a small mirror is attached) rotates within slots cut in the quadrant plates and, therefore, can never accidentally touch the plates.

descending node For a satellite orbiting the earth or another planet, any point at which the groundtrack crosses the equator as it moves from the northern hemisphere into the southern hemisphere. This node generally changes for each succeeding orbit, because the earth or planet rotates underneath the orbit of the satellite. Compare ASCENDING NODE.

descending pass For a specific point on the earth's surface, the time during which an artificial communications satellite is accessible when its latitude is moving southward. The duration of accessibility depends on the altitude of the satellite and on how close its groundtrack comes to the earth-based point. Compare ASCENDING PASS.

description A data element that is part of a record and is used to identify it.

desensitization **1.** The process of making a circuit or device less responsive to small values of a quantity. **2.** Also called *desensing.* In a communications receiver, an unwanted, often intermittent reduction in front-end gain, caused by an extremely strong local signal.

desensitize **1.** To reduce the sensitivity of a receiver. **2.** To reduce the gain of an amplifier. **3.** To reduce the small-quantity response of an instrument.

desiccant A compound, such as cobalt chloride, used for the purpose of keeping enclosed items dry.

design **1.** A unique, planned arrangement of electronic components in a circuit, in accordance with good engineering practice, to achieve a desired end result. **2.** A unique layout of components or controls, in accordance with good engineering practice, esthetics, and (often) ergonomics. **3.** Invention. **4.** Plan. **5.** To produce a design, as defined in **1**, **2**, **3**, or **4**.

designation Within a computer record, coded information identifying the record so that it can be handled accordingly.

design-center rating A specified parameter that, if not exceeded, should provide acceptable average performance for the greatest number of the components so rated.

design compatibility The degree to which a transmitter and receiver are designed for the rejection of unwanted electromagnetic noise.

design engineer An engineer who is skilled in the creation of new designs and in the comparative analysis of designs.

design-maximum rating See MAXIMUM RATING.

design-proof test A performance test made on a newly completed circuit or device to determine the suitability of the design.

Desk-Fax A facsimile transceiver that can be placed on a desktop, used for wire or radio transmission and reception of still images.

desk microphone A microphone equipped with a stand that sits on a table or desktop. It allows the operator to use both hands for equipment adjustment, taking notes, etc.

desktop computer A personal computer designed for nonportable use, usually equipped with a built-in hard disk, one or more diskette drives, a CD-ROM drive, and a fax/modem. It generally uses an external cathode-ray-tube display, printer, and keyboard. The power supply is intended for use with 117-volt utility circuits.

desolder To unsolder joints, usually with a special tool that protects delicate parts and removes melted solder by suction.

destaticization A chemical process used to minimize the retention of electrostatic charges by certain substances.

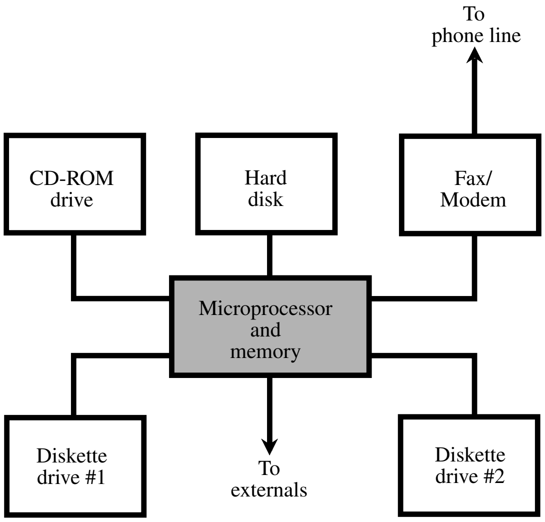

desktop computer
(main unit only)

destination **1.** The point in a system to which a signal of any sort is directed. **2.** In communications, a receiving station.

destination file A computer file that receives data output during a specific program run.

destination register In a digital computer, a register into which data is entered.

Destriau effect Light emission resulting from the action of an alternating electric field on phosphors embedded in a dielectric.

destructive addition A computer logic operation in which the sum of two operands appears in the memory location occupied by one of the operands.

destructive breakdown A breakdown in which the effects are irreversible (e.g., permanent damage to a dielectric by excessive applied voltage).

destructive interference Interference resulting from the addition of two waves that have the same frequency, but opposite phase.

destructive read In a computer or calculator, the condition in which reading the answer erases the data (as from a location) used in the calculation.

destructive test A test that unavoidably destroys the test sample. Compare NONDESTRUCTIVE TEST.

DETAB A COBOL-based computer programming language permitting the programmer to present problems as decision tables.

detail constant Pertaining to a video signal, the ratio V_H/V_L, where V_H is the amplitude of high-frequency components, and V_L is the amplitude of the low-frequency reference component.

detected error In a computer system, an error that is identified, but remains uncorrected until final output is available.

detection **1.** See DEMODULATION. **2.** The sensing of a change in the operating parameters of a circuit or system.

detection range In security applications, the radius within which transducers or sensors can be expected to reliably operate. This radius varies, depending on the environment, the sensitivity of the receiving circuits and transducers, and the strength of the transmitted signal (if any).

detectophone A device for eavesdropping on a conversation. The device can use a tape recorder or a tiny radio transmitter.

detector **1.** In radio communications, a device or circuit that extracts the information from a modulated carrier. Also sometimes called a *demodulator.* **2.** A device that senses a signal or condition and indicates its presence.

detector balanced bias In a radar system, bias obtained from a controlling circuit and used to reduce or eliminate clutter.

detector bias Steady dc voltage applied to a detector to set its operating point.

detector blocking In a regenerative receiver, a phenomenon in which a strong signal tends to pull the detector oscillator into phase with itself, thereby causing the detector to oscillate at the signal frequency.

detector circuit A demodulator circuit (i.e., one used to recover the intelligence from a modulated carrier).

detector probe See DEMODULATOR PROBE.

detector pull-in See DETECTOR BLOCKING.

detector stage In a receiver or instrument, the separate stage that contains the detector circuit. Some systems, such as a superheterodyne receiver, have more than one detector. Also see FIRST DETECTOR and SECOND DETECTOR.

detent A mechanical stop used on a rotary switch to hold the switch pole securely in each selected position.

detune **1.** To adjust a circuit to some frequency other than its resonant frequency. **2.** To set the frequency of a receiver or transmitter to some point other than the frequency normally used. **3.** To stagger-tune a receiver intermediate-frequency system.

detuning Tuning to a point above or below the frequency to which a device or system is normally (or initially) adjusted (usually the resonant frequency of the device).

detuning stub A device used for the purpose of coupling a feed line to an antenna, while choking off currents induced on the feed line as a result of the near-field radiation of the antenna.

deupdating Producing an earlier form of a computer file by substituting older records for current ones.

deuterium Symbol, D, d, H^2, or ^2H. Also called *heavy hydrogen.* The hydrogen isotope having a nucleus consisting of one proton and one neutron.

deuterium oxide Symbol, D_2O. Also called *heavy water.* This compound has wide use in nuclear reactors.

deuteron The nucleus of a deuterium atom.

deuton See DEUTERON.

deutron See DEUTERON.

deviation **1.** In a frequency-modulated (FM) radio signal, the instantaneous amount of carrier frequency shift away from the unmodulated frequency. It is usually expressed in kilohertz; directly proportional to the amplitude of the modulating signal, up to a certain maximum that depends on the bandwidth allowed. **2.** The maximum instantaneous carrier frequency shift in a FM signal. **3.** The extent or amount by which a quantity drifts from its proper value.

deviation distortion In a frequency-modulation (FM) receiver, distortion resulting chiefly from discriminator nonlinearity and restricted bandwidth.

deviation ratio In a frequency-modulated (FM) signal, the ratio between the highest modulating frequency and the maximum carrier deviation.

deviation sensitivity For a frequency-modulation (FM) receiver, the smallest deviation that will produce a specified audio output power. Expressed in kilohertz, or as a percentage of rated deviation of the receiver, measured with the receiver set for maximum gain.

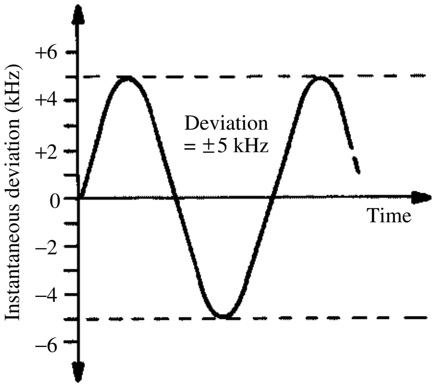

deviation

device **1.** A simple or complex discrete electronic component. **2.** A subsystem used as a unit, and regarded as a single component.

device complexity The number of components in an integrated circuit.

device independence A characteristic of a computer, that allows operation independent of the types of input/output devices used.

dew point For a gas containing water vapor (typically air), the highest temperature at which the vapor condenses as the gas is cooled. The dew point depends on the amount of vapor in the gas.

dew-point recorder An instrument for determining and recording the temperature at which water vapor in the air condenses to a liquid.

DF Abbreviation of DIRECTION FINDER.

DF antenna An antenna that is mechanically rotatable or has an electrically rotatable response pattern for use with a direction finder.

DF antenna system Two or more DF antennas arranged for maximum directivity and maneuverability, together with associated feeders and couplers.

D flip-flop A delayed flip-flop. The state of the input determines the state of the output during the following pulse, rather than during the current pulse.

dg Abbreviation of *decigram.*

dia Abbreviation of *diameter.*

diac A two-terminal, bilateral, three-layer semiconductor device that exhibits negative resistance. When the applied voltage exceeds a critical value, the device conducts.

diagnosis **1.** Determination of the cause and location of a hardware malfunction. **2.** In computer operations, determination of the cause of a system operation error.

diagnostic routine **1.** An efficient sequence of diagnostic tests for rapid, foolproof trouble-shooting of electronic hardware. **2.** A computer software package intended for debugging programs, or for finding the cause of a hardware or operating-system malfunction. Also called *diagnostic*, *diagnostic program*, or *diagnostic utility.*

diagnostic test **1.** A test made primarily to ascertain the cause of dysfunction in electronic equipment. Compare PERFORMANCE TEST. **2.** To apply a diagnostic routine to hardware faults, or to implement one to prevent such a fault.

diagnotor In digital computer operations, a troubleshooting routine combining both diagnosis and editing.

diagram A (usually line) drawing depicting a circuit, assembly, or organization. See, for example, BLOCK DIAGRAM and CIRCUIT DIAGRAM.

dial **1.** A graduated scale, arranged horizontally, vertically, in a circle, or over an arc. Used to show the distance through which a variable component (such as a potentiometer, variable capacitor, or switch) has been adjusted. A pointer can move over the scale, or the scale can be moved past a stationary pointer. **2.** The graduated face of a meter. **3.** In a telephone system, to press the keys or actuate the tones that establish contact with another subscriber.

dial cable A flexible cable or belt conveying motion on the shaft of an adjustable component (such as a potentiometer or variable capacitor) to a dial.

dial-calibrated attenuator A variable attenuator with a dial reading directly in decibels.

dial-calibrated capacitor A variable capacitor with a dial reading directly in picofarads.

dial-calibrated inductor A variable inductor with a dial reading directly in microhenrys.

dial-calibrated potentiometer A potentiometer with a dial reading directly in output volts, percentage of input voltage, number of turns (when resistance is a linear function), or other quantity.

dial-calibrated resistor A variable resistor with a dial reading directly in ohms, kilohms, or megohms.

dial-calibrated rheostat See DIAL-CALIBRATED RESISTOR.

dial cord A form of dial cable. Cord usually designates a fabric string, whereas a cable is a flexible, braided wire.

dial knob The knob used to turn a dial under a pointer, or to turn a pointer over a dial scale.

dial lamp See DIAL LIGHT.

dial light A small lamp sometimes used to illuminate a dial. Can also serve as a pilot light.

dial lock A small mechanism used to lock a dial at a particular setting to prevent further turning.

dialer See AUTOMATIC DIALING UNIT.

dialing key In a telephone system, a dial that uses keys, rather than a rotary dial.

dial jack In a telephone system, a set of jacks that facilitates interconnections between dial cords and external lines.

dial light A lamp or light-emitting diode placed in the dial mechanism of a radio receiver, transmitter, or transceiver. Allows the dial to be read in dim light or in darkness.

dialog equalizer In sound transmission and recording, a high-pass filter that reduces low-frequency response during dialog and extreme closeups.

dial pulse An interruption of the direct current in a telephone system when the dial contacts of the calling telephone open. The number of such interruptions corresponds to the digit dialed.

dial scale The graduated portion of a dial.

dial system **1.** See DIAL TELEPHONE SYSTEM. **2.** The arrangement of dials and knobs that facilitates adjustment of electronic equipment.

dial telephone A telephone set in which a numbered rotatable disk is used to produce the switch interruptions that cause generation of the transmitted multidigit telephone numbers.

dial telephone system The complete automatic circuit, including central-office facilities, for dial telephone operation.

dial tone In a telephone system, a constant hum or whine heard before dialing, indicating that the system is operational.

dial-up In a telephone system, the calling of one subscriber by another, using a dial system.

diam Abbreviation of *diameter.*

diamagnetic Pertaining to a material having magnetic permeability less than unity.

diamagnetism The state of having magnetic permeability less than unity. A material with this property reduces the flux density of a magnetic field, relative to the flux density in air or in free space.

diamond antenna Also called *rhombic antenna.* A nonresonant wideband directional antenna whose horizontal wire elements are arranged in the shape of a diamond (rhombus). The arrangement is fed at one corner, the opposite corner being terminated with a noninductive resistor.

diamond lattice The orderly internal arrangement of atoms in a redundant pattern in crystalline materials, such as germanium or silicon.

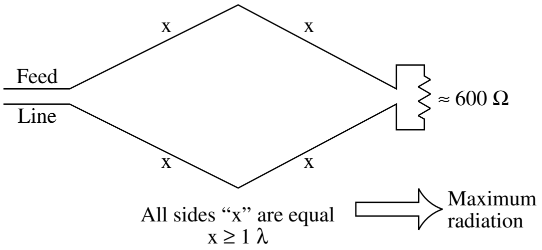

diamond antenna

diamond stylus A phonograph "needle" having as its point a small, ground diamond.

diapason **1.** Either of the two principal stops (open and closed) of an electronic organ that cover the entire range of the instrument. When one is used, a note played is automatically sounded in several octaves. **2.** Tuning fork.

diaphony See DISSONANCE.

diaphragm A usually thin metal or dielectric disk used as the vibrating member in headphones, loudspeakers, and microphones, and as the pressure-sensitive element in some sensors and barometers.

diaphragm gauge A sensitive gas pressure gauge using a thin metal diaphragm stretched flat. Increments of pressure move the diaphragm, relative to a nearby electrode, varying the capacitance between the two.

diathermic Pertaining to a substance that efficiently transfers heat or infrared energy.

diathermotherapy The use of diathermy in the treatment of various physiological disorders.

diathermy **1.** In medicine and physical therapy, the production of heat in subcutaneous (below the skin) tissues by means of high-frequency radio waves. **2.** A radio-frequency (RF) power oscillator and associated equipment used to produce heat in subcutaneous tissues.

diathermy interference Radio-frequency interference (RFI) resulting from the operation of unshielded and/or unfiltered diathermy equipment.

diathermy machine See DIATHERMY, **2**.

diatomic Having two atoms (e.g., a DIATOMIC MOLECULE).

diatomic molecule A molecule (such as that of oxygen) composed of two atoms. Compare MONATOMIC MOLECULE.

dibble A mathematical function in which a number (usually an integer) is doubled, and then one is added to the result. Thus, dibble $n = 2n + 1$.

dibit A combination of two binary digits (bits). The four possible dibits are 00, 01, 10, and 11.

dice Plural of DIE, **1**, **3**.

dichotomizing search Also called *binary search.* In digital computer operations, locating an item in a table of items that are arranged by key values in serial order. The required key is compared with a key halfway through the table; according to this relational test, half of the table is accepted and again divided for comparison, etc. until the keys match and the item is found.

dichotomy Characterized by the usually repetitive branching into two sets, groups, or factions.

dichroism Also called *dichromatism.* **1.** The property of a crystal showing different colors, depending on which axis corresponds to the line of sight. **2.** The property of a solid taking on different colors as the thickness of the transmitting layer changes. **3.** The property of a liquid changing color, according to solution concentration.

dichromate cell An electrolytic cell consisting of electrodes of carbon and zinc. The zinc electrode is immersed in a diluted solution of sulfuric acid, and the carbon electrode in a solution of potassium dichromate.

dicing The cutting of a semiconductor melt, crystal wafer, or other material into dice (see DIE).

dictionary A table of specifications for the size and format of computer file operands, and data names for field and file types.

die **1.** A small wafer of useful electrical material, such as a semiconductor or a precision resistor chip. **2.** A casting designed to mold molten metal into a specific configuration until the metal hardens. **3.** Any small object of roughly cubical proportions. **4.** To lose power or energy completely, usually unintentionally. **5.** In a computer program, to produce unpredicted and useless results following an initial run.

die bonding The bonding of dice or chips to a substrate.

die casting Making a casting by forcing molten metal (such as an aluminum alloy, lead, tin, or zinc) under high pressure into a die or mold.

dielectric A material that is a nonconductor of electricity; especially, a substance that facilitates the storage of energy in the form of an electric field. Such materials are commonly used in capacitors and transmission lines.

dielectric absorption The ability of certain dielectric materials to retain some of their electric charge—even after being momentarily short-circuited. Capacitors with this property must be shorted out continuously for a certain length of time before the dielectric has completely discharged.

dielectric amplifier A voltage amplifier circuit in which the active component is a capacitor having a nonlinear dielectric. A signal voltage applied to the capacitor varies the capacitance, thus varying the current. The modulated current flows through a load resistor, developing an output-signal voltage higher than the input-signal voltage.

dielectric antenna An antenna in which some or all of the radiating element is made of a dielectric material, such as polystyrene. Primarily used at microwave frequencies.

dielectric breakdown Sudden, destructive conduction through a dielectric when the applied voltage exceeds a critical value.

dielectric breakdown voltage The voltage at which DIELECTRIC BREAKDOWN occurs in an insulating material. Varies, depending on the particular dielectric substance.

dielectric capacity See DIELECTRIC CONSTANT.

dielectric constant Symbol, k. For a dielectric material, the ratio of the capacitance of a two-plate capacitor using the dielectric material, to the capacitance of the equivalent capacitor with dry air as a dielectric. Also called *inductivity* and *specific inductive capacity.*

dielectric current **1.** Current flowing over the surface of a dielectric material in response to a varying electric field. **2.** Current flowing through a dielectric as a result of its finite insulation resistance.

dielectric dissipation For a dielectric material in which an electric field exists, the ratio of the lost (dissipated) electrical energy to the recoverable electrical energy.

dielectric dissipation factor The cotangent of the dielectric phase angle, also equal to the reciprocal of the *Q* factor.

dielectric fatigue In some dielectric materials subjected to a constant voltage, the deterioration of dielectric properties with time.

dielectric guide A waveguide made from a solid dielectric, such as polystyrene.

dielectric heater A high-frequency power generator used for DIELECTRIC HEATING.

dielectric heating The heating and forming of a dielectric material, such as a plastic, by temporarily making the material the dielectric of a two-plate capacitor. This capacitor is connected to the output of a high-power radio-frequency (RF) generator. Losses in the dielectric cause its heating. Compare INDUCTION HEATING.

dielectric hysteresis See DIELECTRIC ABSORPTION.

dielectric isolation In a monolithic integrated circuit (IC), the isolation of circuit elements from each other by a dielectric film, as opposed to isolation by reverse-biased pn junctions.

dielectric lens A molded piece of dielectric material used to focus microwaves. Its operation is analogous to that of an optical lens.

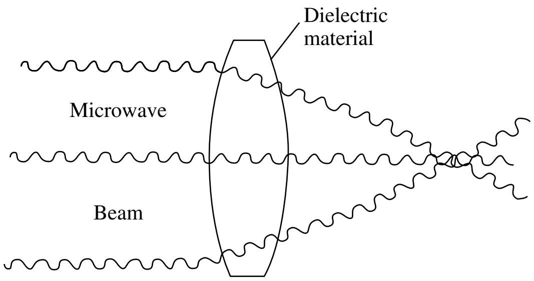

dielectric lens

dielectric loss For a dielectric material subjected to a changing electric field, the rate of transformation of electric energy into heat.

dielectric loss angle Ninety degrees minus the DIELECTRIC PHASE ANGLE.

dielectric loss factor For a dielectric material, the product of the dielectric constant and the tangent of the dielectric loss angle.

dielectric loss index See DIELECTRIC LOSS FACTOR.

dielectric matching plate A dielectric plate used in some waveguides for impedance matching.

dielectric mirror A reflector containing a number of layers of dielectric material. Its action depends on electromagnetic energy being partially reflected from the interfaces between materials having unequal indexes of refraction.

dielectric phase angle For a dielectric material, the angular phase difference between a sinusoidal voltage applied to the material and the component of the resultant current having the same period as that of the voltage.

dielectric phase difference See DIELECTRIC LOSS ANGLE.

dielectric polarization The effect characterized by the slight displacement of the positive charge in each atom of a dielectric material, with respect to the negative charge, under the influence of an electric field.

dielectric power factor The cosine of the dielectric phase angle, or the sine of the dielectric loss angle.

dielectric puncture voltage See DIELECTRIC BREAKDOWN VOLTAGE.

dielectric rating The breakdown voltage, and sometimes the power factor, of the dielectric material used in a device, such as a relay, motor, or switch.

dielectric ratings Electrical characteristics of a dielectric material: breakdown voltage, power factor, dielectric constant, etc.

dielectric resistance See INSULATION RESISTANCE.

dielectric rigidity See DIELECTRIC STRENGTH.

dielectric-rod antenna A unidirectional antenna that uses a dielectric substance to obtain power gain.

dielectric soak See DIELECTRIC ABSORPTION.

dielectric strain The distorted internal state of a dielectric, caused by the influence of an electric field. Also called DIELECTRIC STRESS.

dielectric strength The highest voltage a dielectric can withstand before DIELECTRIC BREAKDOWN occurs. Usually expressed in volts or kilovolts per mil of material thickness.

dielectric stress The distortion of electron orbits in the atoms of a dielectric material subjected to an electric field.

dielectric susceptibility For a polarized dielectric, the ratio of polarization to electric intensity.

dielectric tests Laboratory experiments performed to determine the dielectric characteristics of a substance—especially the dielectric constant and the dielectric breakdown voltage.

dielectric waveguide See DIELECTRIC GUIDE.

dielectric wedge A wedge-shaped dielectric slug placed inside a waveguide for impedance matching.

dielectric wire A small dielectric waveguide that acts as a wire to carry signals between points in a circuit.

Dietzhold network A four-terminal, shunt m-derived circuit used in some wideband amplifiers.

Dietzhold peaking In some wideband amplifiers, frequency compensation obtained with a shunt m-derived network (see DIETZHOLD NETWORK).

difference amplifier See DIFFERENTIAL AMPLIFIER.

difference channel In a stereophonic amplifier, an audio channel that handles the difference between signals in the right channel and those in the left channel.

difference detector A detector whose output is the difference between two simultaneous input signals.

difference frequency A signal frequency produced by mixing or heterodyning of signals at two other frequencies. If the lower input signal frequency is f_1 and the higher input signal frequency is f_2, then the difference frequency f_d is equal to $f_2 - f_1$.

difference of potential The absolute value of the algebraic difference of voltages at two points of different electrical potential. Thus, the difference of potential between a +5-V point and a –5-V point is +5 – (–5) V = 10 V. Also see POTENTIAL DIFFERENCE.

difference quantity See INCREMENT.

difference signal **1.** The resultant signal obtained by subtracting, at every instant for at least one full cycle, the amplitudes of two signals. **2.** The difference of the left- and right-channel outputs in a stereo system.

differential **1.** A device, consisting of a gear system, that adds or subtracts angular motions and delivers the result. **2.** A gear system in which the motion of a shaft is transferred to two other shafts aligned with each other and perpendicular to the first shaft. **3.** One of two coils arranged to produce opposite polarities at a point in a circuit. **4.** Pertaining to a difference between two signals or quantities.

differential amplifier A circuit, usually an operational amplifier, that amplifies the voltage difference between two input signals. The instantaneous output voltage is equal to some constant multiple of the difference between the instantaneous input voltages.

differential analyzer An analog computer that solves differential equations using integrators.

differential angle For a mercury switch, the angle between operation and release positions.

differential capacitor A dual variable capacitor with two identical stator sections, and a single rotor section that turns into one stator section and out of the other. The capacitance of one section decreases while that of the other increases.

differential coil See DIFFERENTIAL, **3**.

differential comparator A linear integrated circuit (IC) that delivers an output proportional to the difference between two input signals.

differential compound dc generator A compound-wound dc generator in which the magnetomotive force of the series field opposes that of the shunt (main) field.

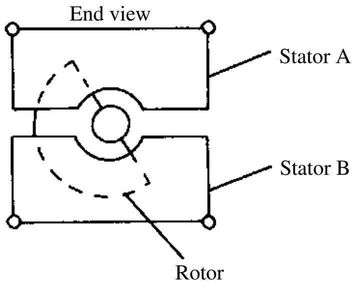

differential capacitor

differential compound dc motor A compound-wound dc motor in which the magnetomotive force of the series field coil opposes that of the shunt (main) field coil.

differential cooling Reducing temperature at different points on a surface at different rates.

differential delay The difference $d_{\max} - d_{\min}$ across a frequency band, where $d_{\max}$ is the maximum frequency delay and $d_{\min}$ is the minimum frequency delay.

differential discriminator A device that passes pulses, whose amplitudes are between two predetermined values above or below zero.

differential distortion In an automatic-gain-control circuit, distortion from effects that cause shunting of the diode load resistor.

differential flutter Fluctuations in the speed of a magnetic tape that are nonuniform in different parts of the tape.

differential gain In a differential amplifier, the average gain of the two sections of the amplifier. Compare DIFFERENTIAL UNBALANCE.

differential gain control A circuit or device for setting the gain of a radio receiver in terms of an anticipated change in signal strength, to reduce the receiver output signal differential.

differential galvanometer A galvanometer in which currents in two similar coils neutralize each other; thus, there is zero deflection when the currents are equal.

differential gap The smallest range of values that a controlled variable must take to change a three-position controller's output from on to off, or vice versa.

differential heating Increase of temperature at different points on a surface at different rates.

differential impedance See DIFFERENTIAL-INPUT IMPEDANCE.

differential induction coil An induction coil having two differentially wound primary coils.

differential input In a differential amplifier, the circuit between input terminals 1 and 2, as opposed to the circuit between input 1 or input 2 and ground.

differential-input amplifier A differential amplifier whose output is proportional to the difference between two input signals—each applied between an input terminal and common ground.

differential-input capacitance In a differential amplifier, the capacitance between the input terminals.

differential-input impedance In a differential amplifier, the impedance between the input terminals.

differential-input measurement For a differential amplifier, a floating measurement made between the input terminals.

differential-input rating In an operational amplifier, the greatest difference signal that can be placed between the inputs while allowing proper operation.

differential-input resistance In a differential amplifier, the resistance between the input terminals.

differential-input voltage In a differential amplifier, the signal voltage presented to the floating input terminals.

differential-input voltage range In a differential amplifier, the range of signal voltages that can be applied between the differential input terminals without overdriving the amplifier.

differential input-voltage rating The maximum differential-input voltage that can be applied safely to a differential amplifier.

differential instrument A galvanometer or other meter in which deflection results from the differential effect of currents flowing in opposite directions through two identical coils. Also see DIFFERENTIAL GALVANOMETER.

differential keying A system of break-in keying, in which the oscillator stage of a transmitter containing a keyed amplifier is disabled when the key is open to prevent interference with the receiver at the keying station, and is enabled when the key is closed.

differential-mode gain In an operational amplifier, the ratio, in decibels, between the output voltage and the differential input voltage.

differential-mode input In an operational amplifier in differential mode, the difference between the two input signal voltages.

differential-mode signal In a balanced three-terminal circuit, such as the input of a differential amplifier, a signal applied between the floating (ungrounded) input terminals.

differential multiplexer An analog multiplexer that selects both the high and low portion of the input signal.

differential nonlinearity Incremental error from an ideal analog output difference when the input is changed by a certain value. Generally expressed as a fraction of full-scale output.

differential permeability The derivative of normal induction, with respect to magnetizing force.

differential phase In a television system tested with a low-level, high-frequency sine-wave signal (f_1) superimposed on a low-frequency, sine-wave signal (f_2), the difference in phase shift of f_1 throughout the system for two specified levels of f_2.

differential phase-shift keying Keying of a carrier by varying the carrier phase.

differential pressure The difference in pressure between two points.

differential-pressure transducer A transducer that delivers an output proportional to the difference between two sensed actuating pressures.

differential protective relay A differential relay that operates to protect equipment or personnel when the difference between the two actuating quantities reaches a prescribed level.

differential receiver A synchro differential that receives the electrical output of two synchro transmitters. The receiver can subtract one input voltage from the other.

differential relay A relay actuated by the difference between two currents or voltages.

differential selsyn A selsyn in which the position assumed by the rotor is proportional to the sum of rotor and stator field values.

differential stage See DIFFERENTIAL AMPLIFIER.

differential synchro See DIFFERENTIAL RECEIVER and DIFFERENTIAL TRANSMITTER.

differential transducer A dual-input, single-output sensor, such as a pressure transducer, that is actuated by two sensed quantities and delivers an output proportional to their difference.

differential transformer A variable inductance transformer having a (usually cylindrical) core that is moved in and out to provide adjustable coupling between the interwound primary and secondary windings. This permits adjustment of the amplitude and phase of the transformer output voltage, with respect to the input voltage.

differential transmitter A synchro differential connected to a synchro transmitter. In a synchro receiver supplied by this combination, the change in rotor position is the algebraic difference between the transmitter-rotor position and the differential-rotor position.

differential unbalance For a differential amplifier, the average difference in gain between the two amplifier sections. Compare DIFFERENTIAL GAIN.

differential voltage **1.** The voltage difference between the input signals to a differential device. **2.** The breakdown voltage minus the operating voltage for a lamp.

differential voltage gain **1.** The ratio, in decibels, between the differential output and differential input voltages of an amplifier. **2.** The instantaneous ratio, in decibels, between the rate of change of the output signal voltage and the rate of change of the input signal voltage in an amplifier.

differential-wound field In a motor or generator, a field winding having series and shunt coils whose fields are opposing.

differentiate **1.** To produce an output signal, the instantaneous amplitude of which is proportional to the instantaneous rate of change of the input amplitude. **2.** To determine the derivative of a mathematical function.

differentiating circuit See DIFFERENTIATING NETWORK.

differentiating network A four-terminal resistance-capacitance (RC) network whose output voltage is the derivative of the input voltage, with respect to time. Compare INTEGRATING NETWORK.

differentiation **1.** The processing of an input signal to create an output signal whose voltage waveform represents the derivative, with respect to time, of the input voltage waveform. **2.** The process of computing a mathematical derivative.

differentiator **1.** See DIFFERENTIATING NETWORK. **2.** An operational amplifier whose output waveform is the mathematical derivative of the input waveform.

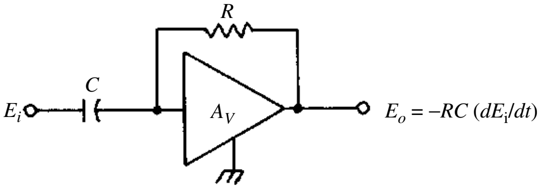

differentiator, 2

diffracted wave A wave or ray of energy undergoing DIFFRACTION.

diffraction **1.** Interference of one part of an energy beam with another part when the beam is deflected along two or more paths having different lengths. When this happens with visible light, dark and light bands or colored bands appear. This effect is responsible for the rainbow-like appearance of light reflected from the surface of a compact disc. **2.** The bending of electromagnetic waves around an object. This effect explains why radio signals can propagate around large obstructions, such as buildings and hills. The effect becomes more pronounced as the wavelength increases (the frequency decreases). **3.** The bending of acoustic waves around an object. This effect explains why sound propagates around large obstructions, such as buildings. The effect becomes more pronounced as the wavelength increases (the frequency decreases).

diffraction grating A transparent plate containing thousands of parallel lines or grooves spaced extremely close together. Light passing through the slits between the lines produces a rainbow spectrum as a result of DIFFRACTION.

diffraction spectrum **1.** The spectrum produced in visible light by a diffraction grating. **2.** The distribution of energy at various frequencies, produced by diffraction of electromagnetic waves. **3.** The distribution of energy at various frequencies, produced by diffraction of acoustic waves.

diffractometer An instrument for measuring the diffraction of radiation, such as light or X-rays.

diffuse **1.** To produce or cause DIFFUSION. **2.** Energy that is diffused.

diffused-alloy transistor See DRIFT-FIELD TRANSISTOR.

diffused-base transistor A bipolar transistor in which the base region has been diffused into the semiconductor wafer. Also see DIFFUSED JUNCTION.

diffused device A semiconductor device in which the junction is produced by diffusion (see DIFFUSION, **1**). Examples: DIFFUSED-BASE TRANSISTOR, DIFFUSED DIODE, DIFFUSED-JUNCTION RECTIFIER, and DIFFUSED-MESA TRANSISTOR.

diffused diode A semiconductor diode having a diffused junction.

diffused-emitter-and-base transistor A transistor in which n and p materials both have been diffused into the semiconductor wafer to provide emitter and base junctions. Also see DIFFUSION, **1** and DIFFUSED TRANSISTOR.

diffused junction In a semiconductor device, a pn junction formed by diffusing a gas into a semiconductor at a high temperature that is below the melting point of the semiconductor. Typically, a gas containing an n-type impurity is diffused into p-type semiconductor material. Compare ALLOY JUNCTION.

diffused-junction rectifier A semiconductor rectifier using a diffused junction.

diffused-junction transistor See DIFFUSED-BASE TRANSISTOR, DIFFUSED-MESA TRANSISTOR, and DIFFUSED TRANSISTOR.

diffused-layer resistor In an integrated circuit, a resistor produced by diffusing a suitable material into the substrate.

diffused-mesa transistor A transistor whose base is a n-type layer diffused into a p-type wafer (the remaining p-type material serving as the collector); its emitter is a small p-type area diffused into or alloyed with the n-layer. Unwanted diffused portions are etched away, leaving the transistor in a mesa shape.

diffused planar transistor A diffused transistor in which emitter, base, and collector electrodes are exposed at the face of the wafer, which has an oxide layer to forestall leakage between surface electrodes.

diffused resistor See DIFFUSED-LAYER RESISTOR.

diffused sound **1.** Sound distributed so that its energy flux is the same at all points. **2.** Sound whose source is difficult to locate or seems to shift, as that heard from out-of-phase stereo channels.

diffused transistor A transistor in which one or both electrodes are created by diffusion. See DIFFUSED JUNCTION.

diffused-junction transistor See DIFFUSED-BASE TRANSISTOR, DIFFUSED-MESA TRANSISTOR, and DIFFUSED TRANSISTOR.

diffusion **1.** In the fabrication of semiconductor devices, the slow, controlled introduction of a material into the semiconductor, for example, the high-temperature diffusion of a n-type impurity (from a gas containing it) into a p-type wafer to form a diode. **2.** The random velocity and movement of current carriers in a semiconductor, resulting from a high-density gradient. **3.** The characteristic spreading of light reflected from a rough surface or transmitted through a translucent material. **4.** The spreading-out of sound waves, for example when reflected from acoustic baffles. **5.** The migration of atoms from one substance to another, as in the spreading of one gas throughout another.

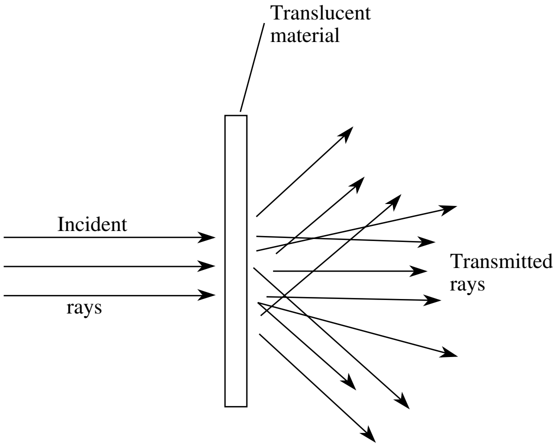

diffusion, 3

diffusion bonding A method of joining different substances by diffusing atoms of one into the other. This technique is employed in the manufacture of certain semiconductor diodes, transistors, and other devices.

diffusion capacitance The current-dependent capacitance of a forward-biased semiconductor junction.

diffusion current Current resulting from the diffusion of carriers within a substance (see DIFFUSION, **2**).

diffusion length In a semiconductor junction, the distance a current carrier travels to the junction during carrier life.

diffusion process **1.** The technique of processing semiconductor devices by diffusion (see DIFFUSION, **1**). **2**. Producing a high vacuum by means of diffusion (see DIFFUSION PUMP).

diffusion pump A pump for fast, efficient creation of a high vacuum in electron tubes and similar devices. In one form, the pump, in conjunction with a force pump, uses mercury vapor as the pumped medium. Gas molecules evacuated from the device diffuse into a chamber, where condensing mercury vapor traps and carries them off.

diffusion theory The notion that, in a homogeneous medium, current density is directly proportional to the gradient of particle flux density.

diffusion transistor A transistor whose operation is based principally on the diffusion of current carriers (see DIFFUSION, **2**).

diffusor In acoustics, a device or structure deliberately installed to spread sound waves throughout a region.

dig-in angle A stylus angle of 90 degrees, relative to the surface of a phonograph disc. Compare DRAG ANGLE.

DIGIRALT Acronym for *digital radar altimetry.* A system that utilizes digital techniques to enhance the accuracy of an altimeter using radar.

digit A single symbol in a numbering system (e.g., 0 through 9 in the decimal system, or 0 or 1 in the binary system), whose value depends on its position in a group and on the radix of the particular system used.

digital **1.** Pertaining to components, circuits, or systems that use signals having an integral number of discrete levels or values, rather than signals, whose levels or values vary over a continuous range. **2.** Pertaining to a numeric readout or display. **3.** See BINARY, **1**.

digital annunciator An annunciator that gives an alphanumeric digital display of information, as well as sounding an alarm.

digital audio tape Abbreviation, DAT. A magnetic tape intended for recording digitally encoded audio data. Used in some high-fidelity applications, and also for computer data storage.

digital barometer An electronic barometer providing a digital readout.

digital capacitance meter Abbreviation DCM. A meter with a digital readout for measuring capacitance values.

digital cellular See PERSONAL COMMUNICATIONS SERVICE.

digital circuit A circuit affording a dual-state switching operation (i.e., on or off, high or low, etc.). Also called *binary circuit.*

digital communications Radio or wire communications using a dual-state mechanism (on/off, positive/negative, or modulated/unmodulated) to represent information.

digital comparator A comparator that presents two digital values, one for each of the quantities being compared.

digital computer A high-speed, electronic machine for performing mathematical operations, file management, machine control, or other "intelligent" functions, and whose basic internal operations (data storage, comparing, and computation) are based on semiconductor devices assuming one of two states (on or off, high or low). Compare ANALOG COMPUTER.

digital data Information represented and processed in the form of combinations of digits (0 and 1, in the binary system).

digital-data cable A cable designed to conduct high-speed digital pulses with minimal distortion and loss.

digital data-handling system A system that accepts, sorts, modifies, classifies, or records digital data, displaying the final result or passing the data to a computer.

digital delay circuit A device that stores digitized audio data, and releases it after a specified delay.

digital device **1.** A digital integrated circuit (IC). **2.** Any circuit or system that operates by digital means.

digital differential analyzer Abbreviation, DDA. A digital computer that can perform integration using specialized circuitry.

digital display A presentation of information (such as the answer to a problem) in the form of actual digits, as opposed to one in the form of, for example, a meter deflection. See, for example, DIGITAL-TYPE METER.

digital divider In a computer, a device that can divide (i.e., provide a quotient and remainder using dividend and divisor signals).

digital electrometer An electrometer that has a digital current or voltage indicator.

digital electronics The branch of electronics concerned with components, circuits, and systems that use signals having an integral number of discrete levels or values, as opposed to signals whose levels or values vary over a continuous range. Compare ANALOG ELECTRONICS.

digital frequency meter A direct-reading frequency meter using high-speed electronic switching circuits and a digital readout. Such instruments read frequency from less than 1 Hz to many gigahertz.

digital HIC A hybrid integrated circuit (HIC) designed for digital applications. Also see DIGITAL INTEGRATED CIRCUIT.

digital IC See DIGITAL INTEGRATED CIRCUIT.

digital incremental plotter A device that can draw, according to signals received from a computer, graphs depicting solutions to problems.

digital information See DIGITAL DATA.

digital information display See DIGITAL DISPLAY.

digital integrated circuit An integrated circuit (IC) intended for binary operations, such as switching, gating, etc. Compare LINEAR INTEGRATED CIRCUIT.

digital integrator A device that can perform integration, in which increments in input variables, and an output variable, are represented by digital signals.

digital logic A form of Boolean algebra, consisting of negation, conjunction, and disjunction, in which the binary digit 1 has the value "true" and 0 the value "false" (in positive logic) or vice versa (in negative logic). Digital logic is the basis by which all digital devices function.

digital-logic module **1.** A circuit that performs digital operations. **2.** A logic gate.

digital meter A meter that produces a readout in discrete blocks or directly as numerals. The first, more primitive and less precise type, is known as a BAR METER. The second, more sophisticated type can resolve to several significant digits and often includes a fixed or floating radix point. This scheme eliminates the need for personnel to interpolate the reading on a scale. There is little chance for error on the part of the technician or engineer, because the readout is straightforward. Another advantage is the fact that there are no moving parts to wear out or be damaged by physical shock. Compare ANALOG METER.

digital multimeter Abbreviation, DMM. A voltohm-milliammeter producing a digital readout of measured values.

digital multiplex **1.** The combination of several or many digital signals into a single digital signal. **2.** Also called *digital demultiplex.* The reverse process from that defined in 1, in which the original signals are obtained from the combination signal. **3.** Communication using the techniques defined in 1 and 2.

digital multiplex equipment Equipment that accomplishes digital multiplexing or the reverse process, digital demultiplexing.

digital multiplier In a digital computer, a device that produces a product signal from multiplier and multiplicand signals.

digital output An output signal of digital pulses representing a number equal or proportional to the value of a corresponding input signal.

digital panel meter A numeric-readout meter whose relatively small size allows mounting on a panel.

digital phase shifter A phase shifter actuated by a digital control signal.

digital photometer An electronic photometer providing a digital readout of illumination values.

digital power meter An electronic wattmeter providing a digital readout of measured power.

digital readout An indicating device that displays a sequence of numerals that represent a measured value.

digital recording A system for tape-recording high-fidelity sound. The audio is converted from analog to binary digital form, and the binary digits (bits) are recorded on magnetic tape.

digital representation The use of digital signals to represent information as characters or numbers.

digital rotary transducer A device that delivers a digital output signal proportional to the rotation of a shaft.

Digital Satellite System Abbreviation, DSS. Trade name for a satellite television (TV) system developed by RCA. The analog signal is changed into digital pulses at the transmitting station via analog-to-digital (A/D) conversion. The digital signal is amplified and uplinked to a geostationary satellite. The satellite has a transponder that receives the signal, converts it to a different frequency, and downlinks it back to the earth. The downlink is picked up by a portable dish that can be placed on a balcony or patio, on a rooftop, or in a window. A tuner selects the channel that the subscriber wants to watch. The digital signal is amplified. If necessary, digital signal processing (DSP) can be used to improve the quality of reception under marginal conditions. The digital signal is changed back into analog form, suitable for viewing on a conventional TV set, via digital-to-analog (D/A) conversion.

digital signal A signal having an integral number of discrete levels or values, as opposed to a signal whose levels or values vary over a continuous range.

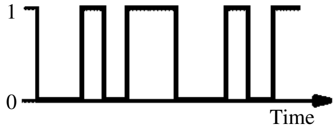

digital signal

digital signal processing Abbreviation, DSP. A method of signal enhancement that operates by eliminating confusion between digital states. This improves dynamic range and frequency response, reduces the number of errors, and virtually eliminates noise. It is used extensively in digital communication and recording, often in conjunction with analog-to-digital (A/D) and digital-to-analog (D/A) conversion to enhance the quality of analog signals and recordings.

digital sound Sound recording and reproduction accomplished with digital, rather than analog, signals. Advantages include wideband frequency response, superior dynamic range, and relative immunity to noise.

digital speech communications A system of voice communications, in which the analog voice signal is encoded into digital pulses at the transmitter, and decoded at the receiver.

digital subtractor In a computer, a device that produces an output signal whose value is equal to the difference of the values of two input signals.

digital switching Routing operations carried out on digital signals to establish communications links between specified system users.

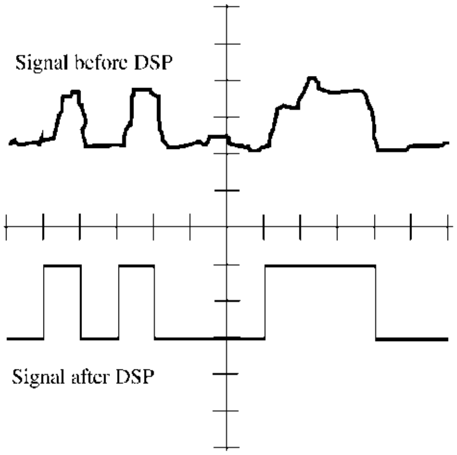

digital signal processing

digital television **1.** A television system in which the picture information is encoded into digital form at the transmitter, and decoded at the receiver. **2.** A form of television picture transmission that functions according to picture motion, rather than absolute brightness.

digital temperature indicator See DIGITAL THERMOMETER.

digital thermometer An electronic thermometer that provides a digital readout of temperature.

digital-to-analog conversion The conversion of a digital quantity into an analog representation, such as shown by a performance curve. Compare ANALOG-TO-DIGITAL CONVERSION.

digital-to-analog converter A circuit or device that performs DIGITAL-TO-ANALOG CONVERSION.

digital transmission **1.** A method of signal transmission in which the modulation occurs in defined increments, rather than over a continuous range. **2.** A message that is sent in digital form.

digital-type meter An indicating instrument in which a row of numeral indicators displays a value. Compare ANALOG-TYPE METER.

digital voltmeter Abbreviation, DVM. An electronic voltmeter having a direct numerical readout, rather than an analog display.

digital wattmeter See DIGITAL POWER METER.

digital compression In digital computer operation, the process of representing data with an economy of characters to reduce file size.

digit current In digital computer operations, the current associated with writing or reading a digit into or out of a memory cell.

digit delay element A logic element (gate) whose output signal lags the input signal by one digit period.

digit filter A device for detecting designations. See DESIGNATION.

digitize **1.** To express the results of an analog measurement in digital units. **2.** To convert an analog signal into corresponding digital pulses.

digitizer See ANALOG-TO-DIGITAL CONVERTER.

digit period In a digital circuit or system, the time interval between the start of one digital pulse and the start of the next pulse.

digit place See DIGIT POSITION.

digit plane In a matrix-type computer memory, the plane within a three-dimensional array of memory storage elements representing a DIGIT POSITION.

digit position The ordinal position of a digit in a numeral, the first position being occupied by the least-significant digit (e.g., 7 is in the third position in the numeral 756).

digit pulse A pulse that energizes magnetic core memory elements representing a digit position in several words.

digitron A display in which all of the characters lie in a single, flat plane.

digit time The duration of a digit signal in a series of signals.

digit time slot In digital communications, the interval of time assigned to one bit or one digit.

digit-transfer bus In a digital computer, a main line (of conductors) that transfers information among various registers; it does not handle control signals.

diheptal CRT base The 14-pin base of a cathode-ray tube. Also see BIDECAL, DUODECAL, and MAGNAL.

DIIC Abbreviation for *dielectric-isolated integrated circuit.* Several separate integrated-circuit wafers are contained in a single package, and kept electrically insulated by layers of dielectric.

dilatometer An instrument used to measure expansion.

dimension **1.** Any measurable quantity, such as distance, time, temperature, humidity, etc. **2.** An axis in the three-dimensional Cartesian coordinate system. **3.** An independent variable in a function of one or more variables.

dimensional analysis A mathematical procedure whereby an equation involving quantities with different units is verified as being *dimensionally correct.* The original variables are replaced with fundamental quantities, such as resistance (R), current (I), length or displacement (d), and time (t), applicable to electrical systems. The equation is dimensionally correct if it can be shown that the left and right sides of the equation are identical.

dimensional ratio In magnetism, the ratio of the longest diameter of an elongated ellipsoid of revolution to the shortest.

dimensional stability Nonvariance or little variance in the shape and size of a medium (such as film) during the processing of that material.

dimensionless quantity A quantity that is merely a real number. Example: logarithm, exponent, numerical ratio, etc. In contrast are physical quantities: 3 volts, 5000 hertz, 10 amperes, etc.

diminished radix complement See COMPLEMENT.

dimmer An electronic device used for controlling the brightness of incandescent lamps. Using amplified control, the device enables high-wattage lamp loads to be smoothly adjusted via a small rheostat or potentiometer. A photoelectric-type dimmer automatically controls lamps in accordance with the amount of daylight.

dimmer curve The function of a light-dimmer voltage output as a function of setting on a linear scale.

DIN Abbreviation for *Deutsche Industrie Normenausschuss.* A German association that sets standards for the manufacture and performance of electrical and electronic equipment, as well as other devices.

D indicator In radar operations, an indicator combining type B and C indicators (see B DISPLAY and C DISPLAY).

Dingley induction-type landing system An aircraft landing system that provides lateral and vertical guidance; instead of radio, it uses the magnetic field surrounding two horizontal cables laid on or under either side of the runway.

diode A two-element device containing an anode and a cathode, and providing unidirectional conduction. The many types are used in such devices as rectifiers, detectors, peak clippers, mixers, modulators, amplifiers, oscillators, and test instruments.

diode action **1.** The characteristic behavior of a diode (i.e., rectification and unidirectional conduction). **2.** Two-electrode rectification or unidirectional conductivity in any device other than a diode (e.g., asymmetrical conductivity between the collector and base of a transistor).

diode amplifier **1.** A parametric amplifier employing a varactor. **2.** An amplifier utilizing hole-storage effects in a semiconductor diode. **3.** A negative-resistance amplifier using a tunnel diode.

diode array A combination of several diodes in a single housing.

diode assembly See DIODE ARRAY.

diode bias A steady direct-current (dc) voltage applied to a diode to establish its operating point.

diode capacitance The capacitance existing at the p-n junction of a semiconductor diode when the junction is reverse-biased. The capacitance generally varies, depending on the reverse-bias voltage.

diode capacitor **1.** A capacitor normally operated with a diode. **2.** A voltage-variable capacitor utilizing the junction capacitance of a semiconductor diode (e.g., a varactor).

diode-capacitor memory cell A high-value capacitor in series with a high-back-resistance semiconductor diode. A data pulse forward-biases the diode and charges the capacitor, which remains charged, thus holding the data bit, because of the long time constant of the high capacitance and the high back resistance of the diode.

diode characteristic The current-versus-voltage curve for a diode.

diode checker An instrument for testing semiconductor diodes. There are two forms: A static checker, which measures forward and reverse current; and a dynamic checker (see DYNAMIC DIODE TESTER), which displays the entire diode response curve on an oscilloscope screen.

diode chopper A chopper using an alternately biased diode as the switching element.

diode clipper A clipper using one or more diodes. A single biased diode will limit the positive or negative peak of an applied alternating-current (ac) voltage, depending on diode polarity and bias. Two biased diodes with opposing polarity will clip both peaks. Also see LIMITER.

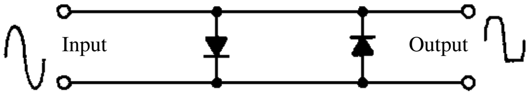

diode clipper

diode converter See DIODE MIXER.

diode current The forward or reverse current flowing through a diode.

diode current meter A direct-current (dc) milliammeter or microammeter with a semiconductor-diode rectifier that allows the measurement of alternating current (ac).

diode curve changer A diode or network of diodes used to make a linear current-voltage curve acquire some nonlinear shape.

diode demodulator See DEMODULATOR PROBE and DIODE DETECTOR.

diode detector A detector circuit in which a diode demodulates a signal. The diode, a simple device, provides linear response at high signal amplitudes, but affords no amplification.

diode feedback rectifier **1.** In a rectified-carrier, negative-feedback system for an amplitude-modulated (AM) transmitter, the diode that rectifies the modulated carrier and provides the audio envelope for use as negative-feedback voltage. This voltage is applied to the speech amplifier/modulator channel to reduce distortion, noise, and hum, at the same time providing automatic modulation control. **2.** The diode that rectifies a part of the signal at the output of an audio amplifier and provides a proportional direct-current (dc) voltage for use as bias in an automatic-gain-control (AGC) circuit.

diode field-strength meter A simple meter for measuring the intensity of a radio-frequency (RF) electromagnetic field. It consists of a short whip antenna, an inductance-capacitance (LC) tuned circuit, a diode detector, and a direct-current (dc) microammeter. The deflection of the meter is roughly proportional to the RF signal voltage.

diode gate A passive switching circuit of biased diodes. Also see AND CIRCUIT and OR CIRCUIT.

diode impedance The vector sum (resultant) of the resistive and reactive components of a diode. In a semiconductor diode, the inductive component of reactance is almost entirely the inductance of leads and electrodes, whereas the capacitive component of reactance is the shunting capacitance between leads and electrodes, plus the voltage-variable capacitance of the pn junction. The resistive component is almost entirely the voltage-variable resistance of the pn junction.

diode isolation A means of insulating an integrated-circuit chip from its substrate. The chip is surrounded by a pn junction that is reverse-biased. This prevents conduction between the chip and the substrate.

diode lamp See LASER DIODE.

diode laser See LASER DIODE.

diode light source See LASER DIODE.

diode limiter See DIODE CLIPPER.

diode load **1.** The current drawn from a diode acting as a rectifier or demodulator. **2.** The output (load) resistor into which a diode operates.

diode load resistance The required value for a diode load resistor.

diode load resistor A resistor usually connected to the output of a diode rectifier or diode detector.

diode logic Digital circuitry, such as AND and OR circuits, using diodes as the principal components.

diode matrix In some digital devices, a grid of wires, the intersections of some being interconnected through diodes, whose polarities determine circuit operation. A series of AND circuits is provided by this arrangement, which acts as a high-speed rotary switch when it is supplied with input pulses.

diode mixer A frequency converter that operates via the nonlinearity of semiconductor diodes.

diode noise limiter A noise limiter circuit having one or more biased diodes.

diode oscillator An oscillator based on the negative resistance or breakdown characteristics of certain diodes, such as high-reverse-biased germanium diodes, tunnel diodes, Gunn diodes, and four-layer diodes. It is generally used at microwave frequencies.

diode pack A device containing more than one diode. An example is the full-wave bridge-rectifier integrated circuit.

diode peak detector A diode detector whose load resistance is high at modulation frequencies; the voltage across the resistance is proportional to the peak amplitude of the modulated signal.

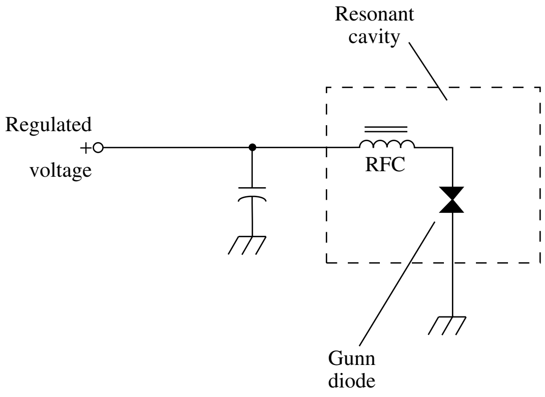

diode oscillator, 1

diode peak voltmeter A diode-type alternating-current (ac) voltmeter, in which the deflection of the direct-current (dc) milliammeter or microammeter is proportional to the peak value of the applied ac voltage.

diode probe A test probe containing a diode used as either a rectifier or demodulator.

diode recovery time The interval during which relatively high current continues to flow after the voltage across a semiconductor junction has been abruptly switched from forward to reverse. Recovery time is attributable to DIODE STORAGE.

diode rectification Conversion of alternating current (ac) to pulsating direct current (dc) by diode action.

diode rectifier **1.** A diode device that converts alternating current (ac) to pulsating direct current (dc) in a power supply. **2.** A small-signal diode device that converts ac to dc in the automatic-gain-control (AGC) circuit of a superheterodyne receiver. Also called *AGC rectifier.*

diode resistor **1.** A resistor usually operated with a diode. **2.** A voltage-variable resistor utilizing the (usually forward) resistance of a semiconductor diode.

diode storage The charge carriers (electrons and holes) remaining within a pn junction for a short time after forward bias has been either removed or switched to reverse polarity.

diode storage time See DIODE RECOVERY TIME.

diode switch See DIODE GATE.

diode sync separator A diode used in a television receiver circuit to separate and deliver the sync pulses from the composite video signal.

diode temperature stabilization **1.** Keeping the temperature of a diode at a constant level. **2.** Using the temperature-resistance characteristic of a forward-biased semiconductor diode to stabilize a circuit (such as a transistor amplifier stage) (i.e., to prevent variations caused by temperature changes).

diode tester See DIODE CHECKER.

diode transistor **1.** See UNIJUNCTION TRANSISTOR. **2.** A semiconductor diode whose operation simulates that of a transistor by means of pulsed operation that alternately makes the single junction an emitter or collector. **3.** A transistor connected to operate solely as a diode.

diode-transistor logic Abbreviation, DTL. Logic circuitry in which a diode is the logic element and a transistor acts as an inverting amplifier.

diode-type meter A rectifier-type alternating-current (ac) meter consisting of a semiconductor diode(s) and a direct-current (dc) milliammeter or microammeter. The diode rectifies the ac input, the resulting dc deflecting the meter.

diode varactor A conventional semiconductor diode or rectifier used as a makeshift varactor (voltage-variable capacitor).

diode variable resistor See DIODE VARISTOR.

diode varistor A conventional diode used as a makeshift varistor (voltage-variable resistor).

diode voltage reference See ZENER VOLTAGE REFERENCE.

diode voltage regulator See ZENER VOLTAGE REGULATOR.

DIP Abbreviation of DUAL IN-LINE PACKAGE.

dip **1.** A distinct decrease in the value of a varying quantity, followed by an increase [e.g., the sudden drop, followed by a rise, in collector current when a bipolar-transistor radio-frequency (RF) power amplifier is tuned through resonance]. **2.** Also called *magnetic inclination.* The slanting of a compass needle, resulting from the orientation of the geomagnetic lines of flux, with respect to the earth's surface. It varies, depending on magnetic latitude.

dip adapter An external accessory that allows a radio-frequency (RF) signal generator to be used as a DIP METER.

dip coating **1.** Applying a protective coat of insulating material to a conductor or component by dipping it into the liquid material, then draining and drying it. Compare SPRAY COATING. **2.** The coat applied in this way.

dip encapsulation Embedding a component or circuit in a protective block of insulating material (such as a plastic) while the material is in a liquid state, and then allowing the material to harden in ambient air or in an oven.

dip impregnation Saturating a component or material (such as absorbent film) with a substance (such as oil or wax) by dipping or vacuum forcing.

diplexer A coupler that permits two or more transmitters to operate simultaneously into a single antenna.

diplex operation **1.** Simultaneous transmission or reception of two signals using a single antenna. **2.** Simultaneous transmission or reception of two signals on a single carrier.

diplex reception The reception of signals while transmitting with the same antenna.

diplex transmission The transmission of signals while receiving with the same antenna.

dip meter A tunable radio-frequency (RF) instrument that, by means of a sharp dip of an indicating meter, indicates resonance with an external circuit under test. Specific names are derived from the active component used: *grid-dip meter*, *gate-dip meter*, etc.

dip needle See INCLINOMETER.

dipolar Also, *bipolar.* Possessing two poles (usually electric or magnetic).

dipolarization See DEPOLARIZATION.

dipole **1.** A pair of electrically opposite charge poles separated by a specific distance. **2.** A pair of magnetically opposite poles separated by a specific distance. **3.** See DIPOLE ANTENNA. **4.** See FOLDED DIPOLE.

dipole antenna Also called *dipole* and *doublet.* A half-wavelength radiator fed at the center with a two-wire or coaxial transmission line. Each "leg" of the antenna is one-quarter wavelength long. Such an antenna can be oriented horizontally or vertically, or at a slant. The radiating element is usually straight. For a straight wire radiator, properly insulated at the ends and placed well away from obstructions, the length L_{ft} (in feet) at a design frequency f (in megahertz) is approximately

$$L_{\text{ft}} = 467/f$$

The length L_{m} (in meters) is close to

$$L_{\text{m}} = 143/f$$

Because of its simplicity, this antenna is popular among shortwave listeners and radio amateurs, especially at frequencies below 10 MHz. A full-size antenna of this type has a feed-point impedance of approximately 73 ohms, purely resistive. Compare FOLDED DIPOLE.

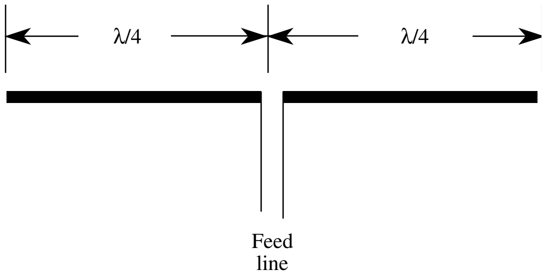

dipole antenna

dipole disk feed A method of coupling radio-frequency energy to a disk-shaped antenna. The energy is applied to a dipole located adjacent to the disk.

dipole feed A method of coupling radio-frequency energy to an antenna by means of a half-wave dipole. The dipole is directly fed by the transmission line, and the dipole radiates energy to the rest of the system.

dip oscillator The oscillator that provides the signal for a DIP METER.

dipotassium tartrate Abbreviation, DKT. An organic piezoelectric material.

dipped component A discrete electronic component that has been given a protective coating by dipping into a suitable material (such as oil, varnish, or wax) and draining off the surplus.

dipper Collective term for resonance-type instruments, such as a DIP METER or DIP ADAPTER.

dipper interrupter A cyclic switching device in which a contact pin is part of a revolving wheel partially immersed in mercury.

dipping **1.** The application of a protective coating or impregnant to a component by immersing it in a suitable material. Also see DIP COATING, DIP ENCAPSULATION, and DIP IMPREGNATION. **2.** In a resonant (tuned) amplifier circuit, the adjustment of the resonant circuit for minimum current through the amplifying device.

dipping needle See INCLINOMETER.

dip soldering **1.** Soldering leads or terminals by dipping them into molten solder and then removing excess solder. **2.** Tinning printed-circuit patterns by dipping the boards into molten solder or placing them in contact with the surface of a solder bath. **3.** Soldering leads in printed circuits by the methods defined in (2).

DIP switch A switch (or group of miniature switches) mounted in a dual-inline package (DIP) for easy insertion into an integrated-circuit socket or printed-circuit board.

direct-access storage device A computer memory in which data access time is unaffected by the data location. Also called *random-access memory device.*

direct-acting recorder See GRAPHIC RECORDER.

direct-acting recording instrument See GRAPHIC RECORDER.

direct address The actual address of a computer storage location (i.e., the one designated by machine code 0. Also called *absolute address* or *real address.*

direct capacitance The capacitance between two points in a circuit, as opposed to the capacitance between either point and other objects (including ground).

direct allocation In digital computer operations, to specify the necessary memory locations and peripherals for a particular program when it is written.

direct coding Computer programming in machine language.

direct control Control of one machine by another, for example, the control of a computerized mobile robot by a central computer system.

direct-conversion receiver A heterodyne receiver in which the incoming radio-frequency (RF) signal

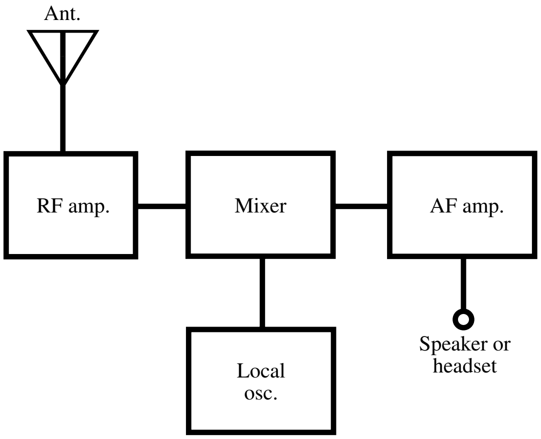

direct-conversion receiver

is amplified, then mixed with the RF output of a tunable local oscillator, producing an audio-frequency (AF) beat note. The AF is amplified; audio filtering can be added. Although the direct-conversion receiver somewhat resembles the superheterodyne type, it has no intermediate-frequency (IF) chain, and does not normally provide single-signal reception. Also see ZERO-BEAT RECEPTION.

direct-coupled amplifier An amplifier in which the output circuit of one stage is wired directly to the input circuit of the following stage (i.e., there is no intervening capacitor or transformer). Such an amplifier can handle alternating-current (ac) or direct-current (dc) signals, and has wide frequency response.

direct-coupled transistor logic Abbreviation, DCTL. In digital computer and switching circuits, a logic system using only direct-coupled transistor stages.

direct coupling Direct connection of one circuit point to another for signal transmission (i.e., without intermediate capacitors or transformers). Because coupling devices aren't used, direct coupling provides transmission of direct current (dc), as well as alternating current (ac).

direct current **1.** Abbreviation, dc. A current that always flows in the same direction (i.e., the polarity never reverses). The current might be constant, as from a battery or a regulated power supply; it might be pulsating, as from an unfiltered rectifier. **2.** Pertaining to current that always flows in the same direction. **3.** Descriptive of a voltage, resistance, or other parameter under conditions in which there is a usually constant current that always flows in the same direction.

direct-current amplifier An amplifier for boosting direct-current (dc) signals, as opposed to dc voltage signals.

direct-current bar See DC BAR.

direct-current beta See DC BETA.

direct-current block See DC BLOCK.

direct-current bus See DC BUS.

direct-current circuit breaker See DC CIRCUIT BREAKER.

direct-current component See DC COMPONENT.

direct-current converter See DC CONVERTER.

direct-current coupling See DC COUPLING.

direct-current dump See DC DUMP.

direct-current equipment See DC EQUIPMENT.

direct-current erase head See DC ERASE HEAD.

direct-current generator See DC GENERATOR.

direct-current inverter See DC INVERTER.

direct-current leakage See DC LEAKAGE.

direct-current motor See DC MOTOR.

direct-current noise See DC NOISE.

direct-current power See DC POWER.

direct-current relay See DC RELAY.

direct-current resistance See DC RESISTANCE.

direct-current shift See DC SHIFT.

direct-current short See DC SHORT.

direct-current signaling See DC SIGNALING.

direct-current source See DC SOURCE.

direct-current transducer See DC TRANSDUCER.

direct-current transformer See DC TRANSFORMER.

direct-current transmission See DC TRANSMISSION.

direct digital control In a digital computer, multiplexing or time sharing among a number of controlled loops.

direct display unit A cathode-ray-tube (CRT) peripheral that displays data recalled from memory.

direct-distance dialing A form of telephone service that allows dialing of long-distance numbers without involving a human operator.

direct drive **1.** Pertaining to electromechanical accessories for electronic equipment. **2.** The transmission of power directly from a source (such as a motor) to a driven device without intermediate gears, belts, or clutches.

direct-drive robot A robot that uses the minimum possible number of gears and other drive systems.

direct-drive torque motor In a positioning or speed-control system, a servoactuator connected directly to the driven load.

direct-drive tuning A tuning or adjusting mechanism in which the shaft of the variable component (such as a potentiometer or variable capacitor) is turned directly by a knob (i.e., without gearing, dial cable, or similar linkage).

directed number A number having direction as well as magnitude; a vector quantity.

direct electromotive force A direct-current (dc) voltage that does not fluctuate or pulsate.

direct emf See DIRECT ELECTROMOTIVE FORCE.

direct ground **1.** A ground connection made by the shortest practicable route. Compare INDIRECT GROUND. **2.** An earth ground.

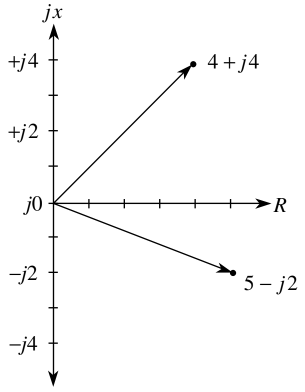

directed number

direct induced current A transient current induced in the same direction as the induction current when it is interrupted.

directing antenna See DIRECTIONAL ANTENNA.

direct-input circuit A circuit, especially an amplifier, whose input is wired directly to the input electrode of the active device (i.e., without a coupling capacitor or transformer).

direct-insert subroutine In digital computer operations, a subroutine directly inserted into a larger instruction sequence. It must be rewritten at every point it is needed.

direct instruction A computer program instruction that indicates the location of an operand in memory.

directional **1.** Depending on direction or orientation. **2.** Having a concentration in an identifiable direction. **3.** Pertaining to a transducer in which radiation, or sensitivity, is concentrated in certain directions at the expense of radiation or sensitivity in other directions.

directional antenna An antenna that transmits and receives signals more effectively in some directions than in others. Also called *beam*, *beam antenna*, and *directive antenna.*

directional array **1.** A directional antenna having a set of elements assembled in such a way that their combined action shapes the radiation into a unidirectional pattern. **2.** A group of antennas spaced and phased to produce unidirectional radiation and reception patterns.

directional beam **1.** An antenna whose radiation or reception pattern strongly favors a specific direction. **2.** The radiation or reception pattern of such an antenna.

directional characteristic The precise directional properties of an antenna or transducer.

directional CQ In amateur radio, a transmission that invites replies only from stations in a certain direction or in a particular city, state, or country.

directional coupler A microwave device that couples an external system to waves traveling through the coupler in one direction.

directional diode A high-back-resistance semiconductor diode inserted into a direct-current (dc) signal circuit or control circuit. Permits unidirectional current flow.

directional filter In carrier-current transmission, a filter that halves the frequency band, one half being for transmission in one direction, and the other half being for transmission in the opposite direction.

directional gain Symbol, k_S. The ratio of the power that would be radiated by a loudspeaker if the free-space axial sound pressure were constant over a sphere, to the actual radiated power. Usually expressed in decibels.

directional homing A scheme for locating the source of a radio signal. An effort is made to keep the bearing of the target or guiding station constant. Therefore, the search path is as direct (as nearly a straight line) as practicable.

directional horn See DIRECTIVE HORN.

directional hydrophone A hydrophone whose response pattern strongly favors one direction.

directional lobe In the spatial response pattern of a device, such as an antenna or loudspeaker, a portion showing emphasized response in a given direction.

directional microphone A microphone that strongly favors sound emanating from in front of it.

directional pattern See DIRECTIVITY PATTERN.

directional phase shifter A phase-shifting circuit in which the characteristics are different in one direction, as compared with the other direction.

directional power relay A relay that is actuated when the monitored power reaches a prescribed level in a given direction.

directional relay See POLARIZED RELAY.

directional response For any form of transducer, a radiation or sensitivity pattern that is concentrated in certain directions.

directional separation filter See DIRECTIONAL FILTER.

directional transducer A device that senses or emits some effect to an extent that depends on the direction from which the effect comes. Directional effects are often, but not always, accompanied by gain in the favored direction(s). Examples: *directional microphone*, *directional speaker*, and *directional antenna.*

directional variation of radio waves Changes in the field strength of radio waves, depending on the direction. There are various causes, including antenna directivity, ground characteristics, ionospheric factors, weather conditions, and the presence of obstructing objects.

directional wattmeter A device that can measure radio transmitter output power and can also give an indication of how well an antenna is matched to a transmission line. Such meters fall into two categories. One type has a single scale, calibrated in watts, and sometimes also in milliwatts or kilowatts (switch selectable). The meter reads either forward power or reflected power, depending on the position of a switch or rotatable internal element. Another type has two needles in a single enclosure, with a different calibrated scale for each needle. Both of these scales are graduated in watts, and sometimes also in milliwatts or kilowatts. One needle/scale indicates forward power and the other needle/scale indicates reflected power. There is a third scale, calibrated for the point where the two needles cross. This scale indicates the standing-wave ratio (SWR). See also CROSSED-POINTER INDICATOR.

direction angle In radar operations, the angle between the center of the antenna baseline and a line going to the target.

direction finder A receiver specially adapted to show the direction from which a signal is received, thus revealing the direction of the receiver with respect to the transmitting station, and vice versa. In its simplest form, it is a receiver with a loop antenna that is rotatable over a map or compass card. For increased accuracy, checks are made with signals from two transmitting stations; the exact location of the receiver is pinpointed by triangulation.

direction finding The taking of bearings by means of a direction finder.

direction of lay In a multiconductor cable, the lateral direction of winding of the topmost conductors as they recede from the observer; called *left-hand lay* or *right-hand lay.* If the cable is viewed from either end, left-hand lay is equivalent to conductors that rotate clockwise as they recede from the viewer; right-hand lay is equivalent to conductors that rotate counterclockwise as they recede from the viewer.

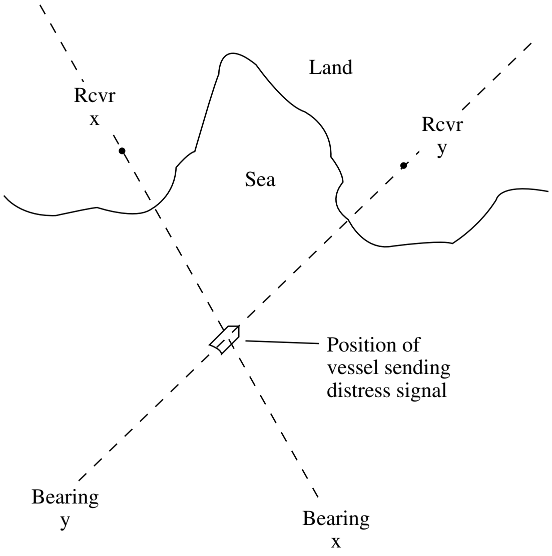

direction finding

direction of polarization The direction of the electrostatic field in a linearly polarized wave.

direction of propagation The direction in which energy moves from a transmitter, or between equivalent points in a sector of space under consideration.

direction rectifier In a control system, a rectifier whose direct-current (dc) output voltage has a magnitude and polarity dependent on the magnitude and phase of an alternating-current (ac) selsyn error voltage.

direction resolution **1.** The smallest difference in azimuth that a direction-finding device can detect. **2.** The smallest angular separation between two targets that allows a radar set to show two separate echoes rather than a single echo.

directive In a computer source program, a statement directing the compiler in translating the program into machine language without being translated itself. Also called *control statement.*

directive antenna An antenna designed for best propagation or reception in one (often steerable) horizontal direction. Also called *beam antenna* and *directional antenna.*

directive gain For a directional antenna, a rating equal to $12.566(Pr/Pt)$, where Pr is the radiated power per steradian in a given direction and Pt is the total radiated power.

directive horn A microwave antenna having the shape of a (usually rectangular) horn.

directivity **1.** In an antenna, a directional response. **2.** The degree to which the radiation or sensitivity of a transducer is concentrated in certain directions. **3.** The angle between the half-power points of a directive antenna in the azimuth plane. **4.** In an antenna system, the ratio, in decibels, between the power in the favored direction and the power in the exact opposite direction; also called front-to-back ratio. **5.** The forward power gain of an antenna, with respect to a dipole in free space. **6.** The forward power gain of an antenna, with respect to an isotropic radiator in free space.

directivity diagram A graph of the radiation/response pattern of a beam antenna or other directional device, usually in a horizontal or vertical plane. Also see DIRECTIVITY PATTERN.

directivity factor **1.** A measure of the directivity of an antenna or transducer. **2.** In acoustics, the ratio, in decibels, between the gain in the maximum direction and the gain in the minimum direction, for a transducer, such as a speaker or microphone.

directivity index **1.** For an acoustic-emitting transducer, the ratio, in decibels, of E_1 to E_2, where E_1 is the average intensity over an entire sphere surrounding the transducer, and E_2 is the intensity on the acoustic axis. **2.** For an acoustic pickup transducer, the ratio, in decibels, of E_1 to E_2, where E_1 is the average response over an entire sphere surrounding the transducer, and E_2 is the response on the acoustic axis.

directivity of antenna For a beam antenna, the ratio E_{max}/E_{avg}, where E_{max} is the maximum field intensity at a selected distance from the antenna and E_{avg} is the average field intensity at the same distance.

directivity of directional coupler The ratio, in decibels, of P_1 to P_2, where P_1 is the power at the forward wave-sampling terminals (measured with a forward wave in the transmission line) and P_2 is the power at the terminals when the wave is reversed in direction.

directivity pattern The calculated or measured radiation or response pattern (transmission or reception) of an antenna, microphone, loudspeaker, or similar device, with particular attention to the directional features of the pattern.

directivity signal A spurious output signal resulting from finite directivity in a coupler.

direct light Light rays traveling directly from a source to a receptor or target without reflection.

directly grounded Connected to earth or to the lowest-potential point in a circuit, without any intervening resistance or reactance.

directly heated cathode A vacuum-tube filament. It is so called because, when heated, it becomes the cathode of the tube (i.e., the emitter of electrons).

directly heated thermistor A thermistor whose temperature changes with the surrounding temperature, and also as a result of power dissipation in the device itself. Compare INDIRECTLY HEATED THERMISTOR.

directly heated thermocouple A meter thermocouple heated directly by signal currents passing through it. Compare INDIRECTLY HEATED THERMOCOUPLE.

direct measurement Immediate measurement of a quantity, rather than determining the value of the quantity through adjustments of a measuring device (e.g., measuring capacitance with a capacitance meter, rather than with a bridge). Compare INDIRECT MEASUREMENT.

direct memory access Abbreviation, DMA. The transfer of data from a computer memory to some other location, without the intervention of the central processing unit (CPU).

direct numerical control In a computer or data system, the capability for distributing information among numerically controlled machines whenever desired.

director In a multielement directional antenna, an element that is usually mounted in front of the radiator element, and that is phased and spaced to direct the radiation forward. The director functions in conjunction with the reflector element, which is usually mounted behind the radiator.

directory See DICTIONARY.

direct pickup The broadcasting, especially in television, of events at the same time as they occur (e.g., without recording/reproduction).

direct piezoelectricity The production of a piezoelectric voltage by mechanically stressing a suitable crystal.

direct playback In audio or video recording, the reproduction of a recording without additional processing (e.g., the playing of an original recorded tape, rather than a tape that has been mass produced).

direct-point repeater A relay-operated telegraph repeater. The received signals actuate the relay, which switches the second line.

direct-radiator loudspeaker A loudspeaker whose cone or diaphragm is directly coupled to the air.

direct ray An electromagnetic ray (wave) that reaches a receiver without reflection or refraction, and without encountering obstructions.

direct recording 1. A record produced by a graphic recorder. **2.** The technique of producing such a record.

direct-recording instrument A device, such as a graphic recorder, that directly produces a permanent record (such as an inked trace) of the variations of a quantity.

direct resistance coupling A form of coupling in which the output of the first amplifying device is connected through a resistor directly to the input of the second device. The resistance value can vary; sometimes the connection is a short circuit.

directrix A fixed line to which a curve is referred (e.g., the axis of a parabola).

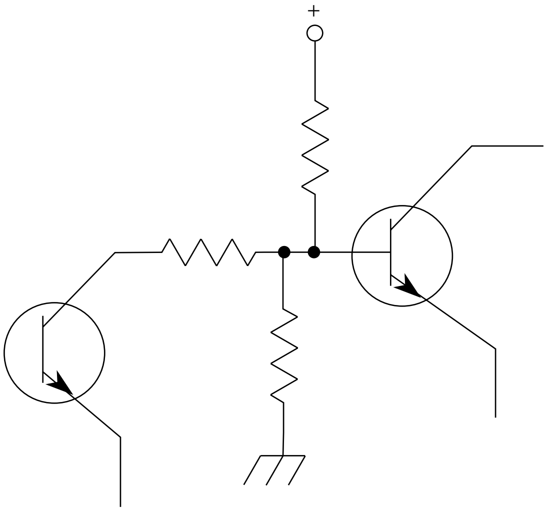

direct resistance coupling

direct scanning In television, the sequential viewing of parts of a scene by the camera—even though the entire scene is continuously illuminated.

direct serial file organization A technique of organizing files stored in a direct access device, in which a record can be chosen by number and amended where it is without altering other members of the file.

direct sound wave A sound wave arriving directly from its source—especially a wave within an enclosure that is not affected by reflection.

direct substitution **1.** An exact component replacement. **2.** Installing an exact component replacement.

direct synthesizer A device for producing random, rapidly changing frequencies for security purposes. A reference oscillator provides a comparison frequency; the output frequency is a rational-number multiple of this reference frequency.

direct voltage See DC VOLTAGE.

direct wave A wave that travels from a transmitter to a receiver without being reflected by the ionosphere or the ground. Compare SKYWAVE.

direct Wiedemann effect Twisting force (torque) in a wire carrying current in a longitudinal magnetic field. Occurs because of interaction between the longitudinal field and the circular magnetic field around the wire.

direct-wire circuit A communications or control line of wires connecting a transmitter (or control point) and a receiver (or controlled point) without an intermediary, such as a switchboard.

direct-writing recorder See GRAPHIC RECORDER.

direct-writing telegraph **1.** See PRINTING TELEGRAPH. **2.** See TELAUTOGRAPH.

dis- A prefix meaning "deprived of." For the formation of electronic terms, the prefix must be distinguished from *un-*, meaning "not." For example, a *discharged* body is one that was charged, but has been emptied of its charge. An *uncharged* body is one that ordinarily or presently is not charged.

disable **1.** To deliberately render a circuit, device, or system inoperative. **2.** In digital computer operations, to defeat a software or hardware function.

disc See DISK.

disc engraving **1.** Recording sound by cutting a groove in a record disc. **2.** The groove resulting from such a process.

discharge The emptying or draining of electricity from a source, such as a battery or capacitor. The term also denotes a sudden, heavy flow of current, as in DISRUPTIVE DISCHARGE. Compare CHARGE.

discharge current **1.** Current flowing out of a capacitor. **2.** Current flowing out of a cell—especially a storage cell. Compare CHARGING CURRENT.

discharge key See DISCHARGE SWITCH.

discharge lamp A gas-filled tube or globe in which light is produced by ionization of the gas between electrodes. Familiar examples are the neon bulb and fluorescent tube.

discharge phenomena The effects associated with electrical discharges in gases, such as luminous glow.

discharge potential See IONIZATION POTENTIAL.

discharger **1.** A short-circuiting tool for discharging capacitors. **2.** A spark gap or other device for automatically discharging an overcharged capacitor.

discharge rate **1.** The current that can be supplied by an electrochemical cell or battery reliably during its discharging cycle. Usually expressed in milliamperes or amperes. **2.** An expression of the speed with which a battery is being discharged at a specific point in time. It is usually specified in amperes or milliamperes.

discharge switch A switch for connecting a charged capacitor to a resistor or other load, through which the capacitor discharges. In some circuits, when the switch is in its resting position, it connects the capacitor to the charging source.

discharge voltage See IONIZATION POTENTIAL.

discharging **1.** The conversion of chemical energy to electrical energy by an electrochemical cell or battery. **2.** The release of stored electrical energy from a capacitor, or from a network containing capacitors.

discharging tongs See DISCHARGER, **1**.

discone antenna An antenna consisting of a horizontal metal or wire-mesh disk above a metal or wire-mesh cone. The antenna has an omnidirectional radiation pattern in the horizontal plane, and provides a good match to a coaxial transmission line over a frequency range of several octaves. Commonly used at very-high frequencies (VHF) and ultra-high frequencies (UHF).

disconnect **1.** To separate leads or connections, thereby interrupting a circuit. **2.** A type of connector whose halves can be pulled apart to open a cable or other circuit quickly. **3.** To open a switch or relay.

disconnector See DISCONNECT, **2** and DISCONNECT SWITCH.

disconnect signal A signal sent over a telephone line, ending the connection.

disconnect switch A switch whose main function is to open a circuit quickly (either manually or automatically) in the event of an overload.

discontinuity **1.** A break in a conductor. **2.** A point at which the impedance in a transmission line abruptly changes.

discontinuous wave trains See DAMPED WAVES.

discrete **1.** Complete and self-contained, as opposed to a part of something else. **2.** Composed of individual, separate members.

discrete capacitor Capacitance that is entirely self-contained, rather than being electrically distributed. Also called LUMPED CAPACITOR. Compare DISTRIBUTED CAPACITANCE.

discrete circuit A circuit comprised of discrete components, such as resistors, capacitors, diodes, and transistors, not fabricated into an integrated circuit.

discrete component A self-contained device that offers one particular electrical property in lumped form (i.e., concentrated at one place in a circuit, rather than being distributed). A discrete component is built especially to have a specific electrical property, and exists independently, not in combination with other components. Examples: disk capacitor, toroidal inductor, and carbon-composition resistor. Compare DISTRIBUTED COMPONENT.

discrete device Any component or device that operates as a self-contained unit.

discrete element A discrete device that forms part of a larger system.

discrete inductor An inductive component that is entirely self-contained, rather than being electrically spread out. Also called *lumped inductor.* Compare DISTRIBUTED INDUCTANCE.

discrete information source A source of data containing a finite number of individual elements, rather than a continuously variable parameter.

discrete part See DISCRETE COMPONENT.

discrete resistor A resistive component that is entirely self-contained, rather than being electrically spread out. Also called *lumped resistor.* Compare DISTRIBUTED RESISTANCE.

discrete sampling Sampling of individual bits or characters, one or more at a time.

discrete thin-film component A discrete component produced by the thin-film process (e.g., thin-film capacitor, thin-film potentiometer, etc.).

discretionary wiring A method of interconnecting the components and circuits on a semiconductor wafer for optimum performance. This requires a separate analysis and wiring pattern for every chip.

discrimination **1.** Sharp distinction between electrical quantities of different value. **2.** The detection of a frequency-modulated (FM) signal (i.e., the delivery of an audio signal corresponding to the frequency or phase variations in the FM carrier).

discriminator A second detector for frequency-modulated (FM) signals, in which two diodes are operated from the center-tapped secondary of a special intermediate-frequency (IF) transformer. The circuit is balanced for zero output when the instantaneous received signal frequency is at the unmodulated carrier frequency; the circuit delivers output when the instantaneous received signal frequency swings above or below the unmodulated carrier frequency. Also see FOSTER-SEELEY DISCRIMINATOR and TRAVIS DISCRIMINATOR.

discriminator transformer The special input transformer in a DISCRIMINATOR.

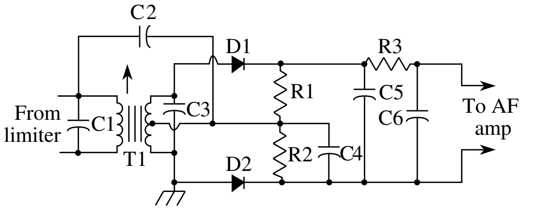

discriminator

discriminator tuner A device that tunes a discriminator to a selected subcarrier.

discriminator tuning device See DISCRIMINATOR TUNER.

dish See DISH ANTENNA.

dish antenna A transmitting and/or receiving antenna consisting of a driven element and a large reflector. The reflector has the shape of a shallow, circular section of a paraboloid or sphere. The feed point is at the focus of the reflector. This antenna, noted for its high directivity and gain, is used mainly at ultra-high and microwave frequencies for communications and satellite television. Large antennas of this type are used in some radio telescopes.

dish-type construction A type of panel-and-chassis construction in which the chassis is fastened vertically to the back of the panel.

disintegration **1.** The destructive breakdown of a material. **2.** The stripping of a vacuum-tube cathode of its emissive coating (see DISINTEGRATION VOLTAGE). **3.** The decay of a radioactive material.

disintegration voltage The anode voltage at which the cathode of a gas tube begins to be stripped of its electron-emitting material. For safety and reasonable tube life, the anode working voltage must be between the ionization and disintegration values.

disintegrator An ultrasonic device for reducing crystals or particles to fine suspensions.

disjunction The logical inclusive-OR operation.

disk **1.** A flat, circular plate (e.g., *rectifier disk*). **2.** See DISKETTE. **3.** See HARD DISK. **4.** See COMPACT DISC. **5.** See CD-ROM. **6.** A phonograph record or the equivalent unrecorded blank.

disk capacitor A fixed (usually two-plate) capacitor consisting of a disk of dielectric material on whose faces are deposited metal-film plates.

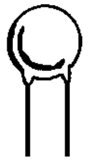

disk capacitor

disk coil See DISK WINDING.

disk dynamo A rudimentary direct-current (dc) generator, in which a copper disk rotates between the poles of a permanent magnet. The outer edge of the disk becomes positively charged; the center of the disk becomes negatively charged.

diskette A magnetic recording disk used for microcomputer data storage. Housed in a square, flat case. In personal computing, there are two sizes: 5.25 inches square and 3.5 inches square. The 5.25-inch version is flexible and is sometimes called a *floppy disk.*

disk files An information-storage system in which data are recorded on rotating magnetic disks.

disk generator **1.** See DISK DYNAMO. **2.** A disk-type electrostatic generator.

disk memory A common misnomer for DISK STORAGE.

Disk Operating System Abbreviation, DOS. Any of several command-driven operating systems commonly used in IBM-compatible personal computers.

disk pack In disk files, a set of disks that can be handled as a single unit.

disk recorder A device for recording (and usually also playing back) sound or other signals on record disks.

disk recording **1.** Recording sound or other signals on disks. **2.** A disk resulting from such a recording. See DISK.

disk rectifier A semiconductor rectifier (such as copper-oxide, selenium, magnesium-copper-sulfide, or germanium type) in which the active material is deposited on a metal disk.

disk resistor A resistor consisting of a resistive material deposited on a metal disk; or a disk of resistive material. In the latter, electrodes are plated on the faces of the disk, one or more of which are held between clips or screws for connections.

disk storage In digital computer systems, an online or offline data storage scheme, in which information is recorded on the magnetic coating of a rotating disk or set of disks. See DISKETTE, and HARD DISK.

disk system A sound-motion-picture system using audio disks synchronized with the film.

disk thermistor A thermistor having the general shape of a disk.

disk-type storage See DISK STORAGE.

disk varistor A varistor having the general shape of a disk.

disk winding An armature or coil winding that is flat, rather than cylindrical. Also called *disk coil, pancake coil,* and *spiral coil.*

dislocation A crystal region in which the arrangement of atoms does not have the perfect lattice structure of the crystal.

dispersion **1.** The property of a material that causes energy at different wavelengths to pass through it at different speeds. **2.** The separation of a wave into its various component frequencies (as when white light is broken up into the color spectrum by a prism). **3.** The scattering of a microwave beam when it strikes an obstruction. **4.** The scattering of sound or ultrasound as it emanates from an acoustic transducer. **5.** A suspension of finely divided particles within another substance.

dispersive medium A medium that disperses a wave passing through it.

displacement **1.** A change in the position of a point, particle, figure, or body. **2.** The vector representing a change in the position of a point, particle, figure, or body. **3.** Movement of a member through a specified distance.

displacement current **1.** An alternating current proportional to the rate of change of an electric field, and existing in addition to usual conduction current. **2.** The current flowing into a capacitor immediately after application of a voltage. This current continues to flow, although continually diminishing in value, until the capacitor becomes fully charged.

displacement of porches In a television signal, the amplitude difference between the front porch and back porch of a horizontal sync pulse.

displacement of vectors Vector rotation through a specified number of angular degrees or radians.

displacement transducer A transducer in which movement (displacement) of a rod, armature, core, reed, or other object converts mechanical energy into proportionate electrical energy.

display **1.** Visually observable presentation of information, such as data entered into a computer, an answer to a problem solved by a computer, the value of a measured quantity, or a graph of a function. **2.** The screen in a computer system or terminal that visually portrays text and graphical data. In laptop, notebook, and portable computers, this is usually a *liquid-crystal display (LCD)*; in desktop computers and terminals, it is usually a *cathode-ray tube (CRT).* **3.** To portray information in a visual manner (e.g., as text, numerals, symbols, or graphic images).

display blanking See DISPLAY INHIBIT.

display console In a computer system, a peripheral that is used to access and display data being processed or stored; often, it is a unit with a cathode-ray tube (CRT), keyboard, and light pen.

display control An interface device between a central processor and several visual display units (terminals).

display dimming See DISPLAY INHIBIT.

display inhibit In a digital meter, the blanking or dimming of the display when the instrument is not being used. It is used to conserve battery energy.

display loss The ratio P_1/P_2, where P_1 is the minimum input-signal power that can be detected by an ideal output device at the output of a receiver, and P_2 is the minimum input-signal power value seen by an operator using an output device with the same receiver. Also called *visibility factor.*

display mode **1.** A particular method of presenting a display. For example, a character display on a video unit might consist of bright characters on a dark background, or dark characters on a light background. **2.** An operating mode for a particular device, in which a display is used.

display module A self-contained unit with circuitry and readouts for indicating a numerical count.

display primaries Also called *primary colors.* In a color television receiver, the colors red, green, and blue. When mixed correctly, these three colors can produce any visible hue.

display-storage tube A special cathode-ray tube in which patterns and other information can be stored for later viewing. The tube has two electron guns: a writing gun and a reading (viewing) gun.

display unit A device that presents information for visual reading. Included are analog and digital meters, cathode-ray tubes, data printers, graphic recorders, etc. Also see DISPLAY CONSOLE.

display visibility The ease with which a display can be read by an operator.

display window **1.** In a panoramic display, the width of the presented frequency band in hertz. **2.** The panel opening through which the indication of a display unit appears.

displayed part That portion of a number displayed in the readout of a calculator or computer. There might be digits that are not displayed, but which the machine might take into consideration when making calculations. For example, in a 10-digit calculator display, the number 245.789378214895 would be displayed as 245.7893782. Depending on the calculator design, the machine might truncate (disregard) the undisplayed digits (14895), or take the undisplayed digits into account when making calculations.

disposable component A circuit component or machine part that is so inexpensive that it is more cost-effective to discard it than to repair it when it fails.

disruptive discharge Sudden, heavy current flow through a dielectric material when it fails completely under electric stress.

dissector A transducer that samples an illuminated image point by point.

dissector tube A camera tube using a flat photocathode, upon which the image is focused by the lens system. Electromagnetic deflection from external coils provides scanning. Electrons pass sequentially from the image cathode to a scanning tube at the opposite end of the camera tube. Also called *Farnsworth dissector tube* and *orthiconoscope.*

dissipation **1.** The consumption of power, often without contributing to a useful end, and usually accompanied by the generation of heat. **2.** In an amplifier, the difference between the collector, drain, or plate input power and the usable output power.

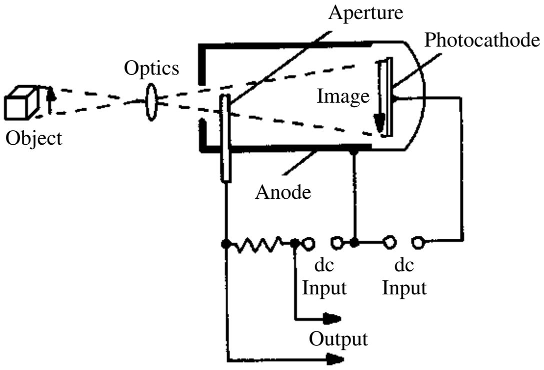

dissector tube

dissipation constant For a thermistor, the ratio of the change in power dissipation to a corresponding change in body temperature.

dissipation factor **1.** For a dielectric material, the tangent of the dielectric loss angle. Also called *loss tangent.* **2.** Symbol, *D.* For an impedance (such as a capacitor), the ratio of resistance to reactance; $D = R/X$. It is the reciprocal of the *figure of merit (Q).*

dissipation line A resistive section of transmission line, used for dissipating power at a certain impedance. Two parallel lengths of resistance wire are terminated by a large, noninductive resistor that has a value equal to the characteristic impedance of the line.

dissipator **1.** A device used primarily to consume power (i.e., a power sink). **2.** A device for removing heat generated by a device's operation (e.g., a heatsink attached to a power transistor).

dissociation The condition that characterizes electrolytes (certain acids, bases, or salts in water solution) in which the molecules of the material break up into positive and negative ions.

dissonance The unpleasant effect (especially in music) produced by nonharmonious combinations of sounds.

dissymmetrical network A network having unequal input and output image impedances.

dissymmetrical transducer A transducer having unequal input and output image impedances.

distance-double law A theoretical rule for determining the rate at which sound intensity decreases as distance increases. Under ideal conditions, when the distance from a sound source is doubled, the sound pressure is reduced to one-fourth of its original level, a reduction of 6 decibels. This is analogous to the *inverse-square law* for visible light and other radiant energy.

distance mark On a radar screen, a mark indicating the distance from the radar set to the target.

distance measurement Also called *ranging.* A method or system that allows a robot to navigate in its environment. It also allows a central computer to track the locations of robots under its control. Can use radar, sonar, visible light, or infrared.

distance-measuring equipment In radionavigation, a system that measures the distance of the interrogator to a transponder beacon in terms of the transmission time to and from the beacon.

distance protection The use of a protective device within a specified electrical distance along a circuit.

distance relay In circuit protection, a relay that operates to remove power when a fault occurs within a predetermined distance along the circuit.

distance resolution **1.** Qualitatively, the ability of a ranging system to differentiate between two objects or beacons that are almost, but not quite, the same distance away. See RANGING. **2.** Quantitatively, the minimum radial separation of objects or beacons necessary for a ranging system to tell them apart. **3.** For two targets having the same azimuth bearing, the minimum difference in range for which a radar display renders them as distinct blips.

distant control See REMOTE CONTROL.

distorted-drive multiplier A frequency multiplier whose excitation signal is a peaked wave that has been predistorted to decrease the angle of flow in the device, thus increasing its efficiency.

distorted nonsinusoidal wave A nonsinusoidal wave whose ideal shape (square, rectangular, sawtooth, etc.) has been altered.

distorted sine wave A wave that is approximately of sinusoidal shape (i.e., it is not an exact plot of a sine wave because of the presence of harmonics).

distortion **1.** Deformation of a signal waveform. **2.** The additional deformation of a signal exhibiting a less-than-ideal waveshape when it passes through a circuit. Some distortion originates within the signal generator itself; other forms result from circuits and devices transmitting the signal. **3.** Any degradation in the quality of a high-fidelity audio signal. **4.** See TOTAL HARMONIC DISTORTION.

distortionless **1.** Having no distortion. **2.** Having a propagation velocity that does not depend on frequency.

distortion meter An instrument for measuring harmonic distortion. It consists of a highly selective band-rejection filter (notch filter) that removes the fundamental frequency of the signal under test, and a sensitive voltmeter that can be switched between the filter input and the filter output. The distortion percentage is determined from the ratio between filter-output and filter-input voltages.

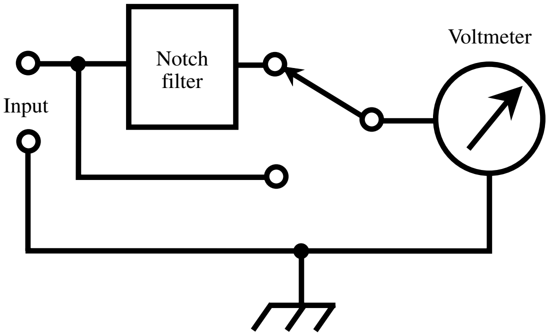

distortion meter

distortion tolerance The maximum amount of distortion that can be present in a signal without making it useless. This varies over wide limits. The maximum harmonic distortion that might be acceptable in a high-fidelity sound system could be less than 0.1% total, whereas in some applications of ac power, 10% would be acceptable.

distress frequency A radio frequency on which an emergency signal is transmitted. Ships at sea and aircraft over the sea use 500 kHz (by international agreement). In Citizen Band communications, channel 9 has been set aside for emergency use.

distress signal A signal indicating that trouble exists at the transmitting station and imploring aid from the recipient. The international radiotelegraph distress signal is the three-letter combination *SOS*; the international radiotelephone distress signal is the word *mayday*, the phonetic equivalent of the French *m'aidez* (help me).

distributed Existing over a measurable interval, area, or volume; not concentrated in a single place or places.

distributed amplifier A wideband, untuned amplifier whose active devices are spaced (distributed) along parallel, artificial delay lines consisting of coils that act in combination with the input and output capacitances of the active devices. Adding active devices to the lineup increases the gain. Commonly used as preamplifiers for television receivers.

distributed capacitance Symbol, C_d. Capacitance that is dispersed throughout a component or system, rather than being lumped in one place. An example is the distributed capacitance of a coil.

distributed component An electrical property that is spread throughout a circuit or device, rather than being concentrated at one point, as in a discrete component. For example, DISTRIBUTED CAPACITANCE and DISTRUBUTED INDUCTANCE are spread along the length of a transmission line. Another example is the DISTRIBUTED RESISTANCE of a wire coil. Distributed components are often unintended, but they can be useful. Compare DISCRETE COMPONENT and LUMPED COMPONENT.

distributed constant See DISTRIBUTED COMPONENT.

distributed-constant delay line A delay line whose capacitance and inductance are distributed throughout the line. Compare LUMPED-CONSTANT DELAY LINE.

distributed inductance Symbol, L_d. Inductance that is dispersed throughout a system or component, rather than being lumped in one place, such as in a coil (e.g., the inductance of an antenna or capacitor).

distributed network **1.** A network in which electrical properties (such as resistance, inductance, and capacitance) are distributed over a measurable interval, area, or volume. **2.** A network whose characteristics do not depend on frequency within a given range.

distributed-parameter network A network composed of distributed components, rather than lumped components.

distributed pole In a motor or generator, a pole having a DISTRIBUTED WINDING.

distributed resistance Symbol, R_d. Resistance that is dispersed throughout a component or circuit, rather than being lumped in one place, such as in a resistor. An example is the high-frequency resistance of an antenna system.

distributed-shell transformer A transformer having two complete closed cores that are perpendicular to each other.

distributed winding In a motor or generator, a winding that is placed in several slots (rather than in one slot) under a pole piece.

distributing amplifier An amplifier having a single input and two or more outputs that are isolated from each other; it distributes signals to various points.

distributing cable **1.** In cable television, the cable connecting the receiver to the transmission cable. **2.** In power service, the cable running between a feeder and a consumer's house.

distribution **1.** The selective delivery of a quantity (e.g., *power distribution*). **2.** In statistical analysis, the number of times particular values of a variable appear. Also called *frequency distribution.*

distribution amplifier A low-output-impedance power amplifier that distributes a radio, television, or audio signal to a number of receivers or speakers.

distribution cable See DISTRIBUTING CABLE.

distribution center **1.** The central point from which a signal is routed to various points of use. **2.** In electric power operations, the point at which generation, conversion, and control equipment is operated to route power to points of use.

distribution factor For a polyphase alternator, the factor by which the total voltage V_T can be determined in terms of the coil voltage V_C and the number of coils n: $V_T = nV_C$. Distribution factor $k_d = (\sin(sd/2))/(s \times \sin(d/2))$, where s is the number of slots per phase per pole, and d is the angle between adjacent slots.

distribution function In statistical analysis, the function $F(x)$ expressing the probability that F takes on a value equal to or less than x.

distribution switchboard **1.** A switchboard through which signals can be routed to or among various points. **2.** A switchboard for routing electric power to points of use.

distribution transformer A step-down transformer used to supply low-voltage alternating-current (ac) utility power to one or more consumers from a high-voltage line.

distributor **1.** See COMMUTATOR. **2.** A switching device consisting of a rotating blade and a number of contacts arranged in a circle. Accomplishes sequential switching of a voltage to a number of points in a circuit. A common example is the distributor in the ignition system of an automotive engine.

disturbance An undesired variation in, or interference with, an electrical or physical quantity.

disturbed-one output In digital computers, the one output of a magnetic cell that has received only a partial write pulse train because it was last written into. Compare UNDISTURBED-ONE OUTPUT.

disturbed-zero output In digital computers, the zero output of a magnetic cell that has received only a partial write pulse train since it was last read from. Compare UNDISTURBED-ZERO OUTPUT.

dither **1.** Vibrate; quiver. **2.** The condition of vibration or quivering (e.g., the dither of a meter pointer). **3.** To blend pixels in a digitized image to obtain various shades and colors.

divergence **1.** The tendency of a collimated beam of energy to spread out. **2.** The extent to which a collimated beam of energy spreads out, generally measured in seconds of arc, minutes of arc, angular degrees, or angular radians.

divergence loss Loss of transmitted sound energy, resulting from spreading.

diverging lens A lens having a virtual focus for parallel rays; generally a *concave lens.*

diversity **1.** The property of consisting of two or more independent components or media. **2.** See DIVERSITY RECEPTION. **3.** See DIVERSITY TRANSMISSION.

diversity factor **1.** A measure of the degree to which a system exhibits unity among its constituents. **2.** The sum of the requirements of each constituent of a system, divided by the total requirement of the system.

diversity gain **1.** Signal gain achieved by using two or more receiving antennas. **2.** Signal gain achieved by using two or more transmitting antennas.

diversity reception Also called *dual-diversity reception.* A method of minimizing the effects of fading in ionospheric communication at high frequencies (HF). Accomplished using two receivers whose antennas are 5 to 10 wavelengths apart. Each receiver, tuned to the same signal, feeds a common audio amplifier. The timing of the fading is different at the two antennas because of phasing effects. The composite signal, therefore, fades less than either of the component signals. Some diversity systems use three or more antennas and receivers to reduce the effects of fading even further; this is sometimes called *multiple-diversity reception.*

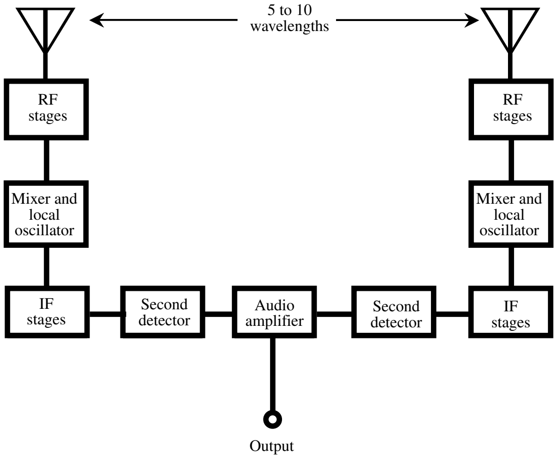

diversity reception

diversity transmission Also called *dual-diversity transmission* or *multiple-diversity transmission.* A scheme similar to DIVERSITY RECEPTION, except applied at the transmitting end of a communication circuit. The signals from two or more transmitters, at identical frequencies, are fed to antennas spaced several wavelengths apart.

diverter-pole generator A well-regulated direct-current (dc) generator, whose shunt winding is on the main field pole, the series winding being on a diverter pole whose flux opposes that of the main pole.

divide-by-seven circuit A three-stage binary circuit having feedback from stage three to stage one. Stage three is switched on by the fourth input pulse; at that time, the feedback pulse switches on stage one, simulating one input pulse and reducing the usual counting capacity from eight to seven.

divide-by-two circuit A circuit that delivers one output pulse for each two successive input pulses (i.e., a *flip-flop*).

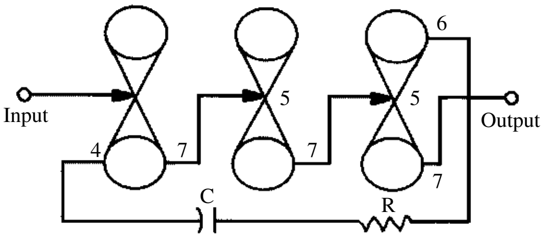

divide-by-seven circuit

divided-carrier modulation Modulation obtained by adding two identical frequency carriers that are 90 degrees out of phase.

divided circuit A *parallel circuit.*

divided equipment A system of modular electronic components interconnected with cables. A simple example is a radio receiver having an external power supply and external loudspeaker.

divider **1.** See VOLTAGE DIVIDER. **2.** See FREQUENCY DIVIDER. **3.** See PULSE-COUNT DIVIDER. **4.** A computing circuit or device for performing mathematical division.

divider probe A test probe that divides an applied signal voltage by some factor (such as 2, 5, or 10) to place it within the range of the instrument with which the probe is used.

dividing network See CROSSOVER NETWORK.

division **1.** Separating a quantity into a number of equal parts, as indicated by the divisor. **2.** Voltage division (see VOLTAGE DIVIDER). **3.** Frequency division (see FREQUENCY DIVIDER). **4.** Pulse-count division (see PULSE-COUNT DIVIDER).

division of vectors **1.** The quotient of two rectangular vectors determined by the principle of rationalization in algebra (i.e., by multiplying the numerator and denominator of the indicated division by the conjugate of the denominator, simplifying, and performing the division). **2.** To find the quotient of two polar vectors: the quotient of their moduli and the difference of their arguments.

dj Abbreviation of *diffused junction.*

DKT Abbreviation of *dipotassium tartrate.*

D layer A layer of the ionosphere that is below the E layer; its altitude is approximately 60 kilometers.

dm Abbreviation of *decimeter.*

DMA Abbreviation of DIRECT MEMORY ACCESS and *direct memory addressing.*

DME Abbreviation of DISTANCE-MEASURING EQUIPMENT.

DMM Abbreviation of DIGITAL MULTIMETER.

DMOS Abbreviation for *double-diffused metal-oxide semiconductor*, a type of field-effect transistor that exhibits extremely low capacitance and low source-drain resistance when conducting.

DNL See DIFFERENTIAL NONLINEARITY.

DNS Abbreviation of *Doppler navigation system.*

doctor To use unconventional (sometimes substandard) methods in fixing a circuit or device or in correcting a bad design.

document **1.** In digital computer operations, especially in file maintenance, a form that provides information pertinent to a transaction. Also see TRANSACTION. **2.** To perform documentation (see DOCUMENTATION, **2**). **3.** A computer text file.

documentation **1.** Paperwork explaining the scope of programs and how they can be optimized. **2.** Annotating a computer program at critical points during its writing (e.g., so that the purpose of various segments are understood). A measure of good programming, documentation becomes especially valuable for program modification or debugging.

document reader An electronic device that reads printed cards, usually for data entry into a computer.

dog **1.** A malfunctioning circuit or device. **2.** The cause of a circuit or device malfunction.

doghouse An enclosure for antenna loading inductors and other resonating components, placed at the base of a vertical broadcasting tower.

dog whistle See ULTRASONIC WHISTLE.

Doherty amplifier A highly efficient linear radio-frequency (RF) amplifier in which a *carrier tube* and a *peak tube* operate jointly, both receiving amplitude-modulated RF excitation. During unmodulated intervals, the carrier tube supplies carrier power to the load, while the peak tube, biased to cutoff, idles. On positive modulation peaks, the peak tube supplies output power that combines with that of the carrier tube, the increase in power corresponding to the condition of full modulation of the carrier. On negative modulation peaks, the peak tube does not supply power, and the output of the carrier tube is reduced to zero.

Dolby An electronic method of improving the audio reproduction quality of magnetic-tape systems. The gain is increased for low-level sounds during the recording process. During playback, the gain of the low-level sounds is reduced back to its original level.

Dolby A A Dolby system with four frequency ranges, operated independently. It is used mostly by recording professionals.

Dolby B A modified form of Dolby A, with only one band of noise-reducing circuitry. It is used primarily by hobbyists.

Dolezalek electrometer See QUADRANT ELECTROMETER.

dolly **1.** A low, wheeled frame or platform for transporting electronic equipment. **2.** A tool with which one end of a rivet is held while the head is hammered out of the other end.

DOM Abbreviation of DIGITAL OHMMETER.

domain **1.** A region of unidirectional magnetization in a magnetic material. **2.** A region of unidirectional polarization in a ferroelectric material. **3.** A region in which a variable is confined.

domestic electronics Also called *consumer electronics.* The branch of electronics concerned with appliances, automatic controls, protective devices, entertainment systems, communications devices, and other equipment for the home.

domestic induction heater A household cooking utensil heated by currents induced in it. A primary coil (connected to the power line) is imbedded in the utensil, which acts as a short-circuited secondary coil.

dome tweeter A speaker designed for high-frequency (treble), high-fidelity audio, and often functional at frequencies considerably above the limit of the human hearing. Characterized by a convex diaphragm. Usually part of an assembly including a woofer and midrange speaker.

dominant In statistical analysis, the nature of any quantity that imposes its effects even in the presence of other quantities.

dominant mode In a waveguide, the propagation mode exhibiting the lowest cutoff frequency.

dominant wave In a waveguide, the wave having the lowest cutoff frequency.

dominant wavelength For visible light of a given hue, the wavelength at which the emitted energy is the greatest.

donor An electron-rich impurity added to a semiconductor to make it into an n-type material. So called because it donates its excess electrons. Compare ACCEPTOR.

donor atom An atom having an excess electron. When a substance having such atoms is added to an intrinsic semiconductor, the extra electron is donated, making the semiconductor into an n-type material.

donor impurity A substance whose atoms have excess electrons, and that donates electrons to the atomic structure of the semiconductor crystal to which it is added. Donor elements make semiconductors into n-type materials. Also see DONOR ATOM. Compare ACCEPTOR IMPURITY.

do-nothing instruction A computer program instruction that causes no action to be taken. Can be used to provide space for future program updating, or to fill out a block of instructions, as needed by a compiler. Also called *dummy instruction.*

don't-care state In a logic function or gate, an input digit whose state (high or low) does not affect the output.

donut capacitor A flat, ring-shaped capacitor.

donut coil See TOROIDAL COIL.

donut crystal A relatively large, zero-temperature-coefficient piezoelectric quartz crystal cut in the form of a torus with the y-axis passing through the center of rotation.

donut magnet See RING MAGNET.

donut pattern The three-dimensional radio-frequency (RF) radiation/response pattern of a free-space straight antenna element measuring ½ wavelength, neglecting the effects of ground and nearby objects.

doohickey A usually unnamed device—especially one used to achieve some significant modification of circuit performance.

doorknob capacitor A high-voltage fixed capacitor, so called from its round package, which somewhat resembles a doorknob.

doorknob tube A special UHF vacuum tube, so called from its characteristic shape. The unique design provides short electron-transit time and low interelectrode capacitance. Largely replaced in recent years by semiconductor devices.

dopant An impurity added in controlled amounts to a semiconductor to make it an n-type or p-type material. Also see ACCEPTOR and DONOR.

dope To add impurities to a semiconductor material. Doping allows the manufacture of n-type or p-type semiconductors with varying degrees of conductivity. In general, the greater the extent of doping, the higher the conductivity.

doped junction In a semiconductor device, a junction produced by adding a dopant to the semiconductor melt.

doping Adding a dopant to a semiconductor to alter the way it conducts current.

doping agent See DOPANT.

doping compensation Opposite doping (i.e., adding a donor impurity to p-type semiconductor material or adding an acceptor impurity to n-type semiconductor material).

doping gas A gas diffused into a semiconductor material to dope it. For example, phosphorus pentoxide gas can be used to create an n-type region in a p-type silicon chip.

doping level The relative concentration of impurity added to a semiconductor material to obtain a certain resistivity and polarity. The greater the doping level, the lower the resistivity.

Doppler broadening In a spectrum, the spreading out or blurring of a spectral line caused by DOPPLER EFFECT, in turn resulting from motion of molecules, atoms, or other particles in the medium.

Doppler cabinet A loudspeaker enclosure with which a vibrato effect is achieved by rotating or reciprocating either the loudspeaker or a baffle board; the length of the sound path is altered cyclically.

Doppler effect A change in the frequency of a wave that occurs when the source and observer are in relative motion. The frequency of the wave increases (the wavelength shortens) as the source and observer approach each other; the frequency decreases (the wavelength becomes greater) as the source and observer recede from each other. This effect is often observed with sound waves, as when the pitch of an automobile horn seems to rise as the car approaches and to fall as the car passes. The effect is also observable in electromagnetic radiation at all wavelengths. It affects satellite communication and space communication.

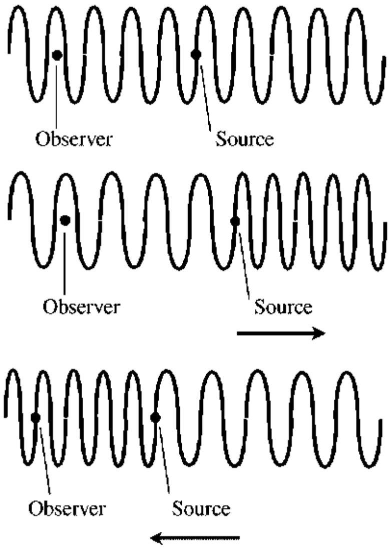

Doppler effect

Doppler enclosure See DOPPLER CABINET.

Doppler radar A radar that uses the change in carrier frequency of the signal returned by a moving target (approaching or receding) to measure its velocity. Used by law enforcement officers to determine the speed of moving vehicles. Also used by meteorologists to evaluate air circulation patterns in thunderstorms, and to determine wind speeds in hurricanes and tornadoes.

Doppler ranging See DORAN.

Doppler shift The extent to which the frequency or wavelength of a signal changes because of DOPPLER EFFECT. Can be measured in Hertz (for frequency) or in meters (for wavelength). In astronomy, the shift is also measured as displacement of absorption or emission lines in an infrared, visible, or ultraviolet spectrum.

Doppler's principle See DOPPLER EFFECT.

doran A continuous-wave trajectory-measuring system utilizing Doppler shift (see DOPPLER EFFECT). The name is a contraction of *doppler ranging.*

dorsal column stimulator Abbreviation, DCS. A machine that generates radio-frequency energy that is applied to human tissues for the temporary relief of pain.

dosage meter See DOSIMETER.

dose The total quantity of radiation received upon exposure to nuclear radiation or X-rays.

dosimeter An instrument for measuring the amount of exposure to nuclear radiation or X-rays.

dot **1.** The shorter of the two characters (dot and dash) of the telegraph code. The dot, a short sound, mark, or perforation, is one-third the length (duration) of a dash. Compare DASH. **2.** One of the small spots of red, green, or blue phosphor on the screen of a color-television picture tube or cathode-ray-tube (CRT) computer display. **3.** A small spot of material alloyed with a semiconductor to form an alloy junction. **4.** The junction of two lines on a schematic diagram, representing a wired connection; also called *solder dot.*

dot AND Externally connected circuits or functions whose combined outputs result in an AND function. Compare DOT OR.

dot-and-dash telegraphy Telegraphy (wire or radio) by means of dot and dash characters.

dot cycle One period of an alternation between two signaling conditions, each of which is of unit duration (e.g., a unit mark followed by a unit space).

dot encapsulation A method of packaging cylindrical components by pressing them into the holes of perforated disks; interconnections are made, to complete a circuit, on each face of the disks.

dot generator A special radio-frequency (RF) signal generator used to produce a pattern of red, green, and blue dots on the screen of a color television receiver.

dot matrix A rectangular array of spaces, some of which are filled in to form alphanumeric and punctuation characters.

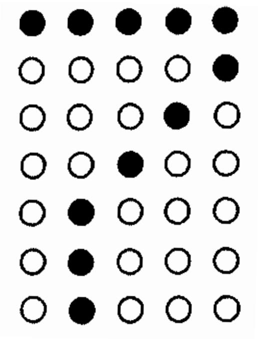

dot matrix

dot-matrix display A display that shows characters in dot-matrix form.

dot-matrix printer A computer output peripheral that prints characters and images on paper as a fine grid of dots. A print head, containing several pins, presses the ribbon against the paper as it moves laterally across each line. Can be used to print text and/or graphics.

dot movement pattern The movement of the red, green, and blue dots on the screen of a color television picture tube as the red, green, and blue magnets and the lateral magnet are adjusted for convergence of the dots at the center. The blue dots move horizontally or vertically; the red and green dots, diagonally.

dot OR Externally connected circuits or functions whose combined outputs result in an OR function. Compare DOT AND.

dot pattern In color television testing with a dot generator, dots of color (a red group, green group, and blue group) produced on the screen. With overall beam convergence, the three groups blend to produce white.

dot-sequential system The color television system in which the image is reproduced by means of primary-color dots (red, green, blue) sequentially activated on the screen of the picture tube. Compare FIELD-SEQUENTIAL SYSTEM and LINE-SEQUENTIAL SYSTEM.

double-amplitude-modulation multiplier A modulating system in which a carrier is amplitude-modulated first by one signal and then by a second signal. The resulting signal is fed to a detector, the output of which contains the product of the two modulating signals.

double-anode diode A semiconductor diode having two anodes and a common cathode.

double armature An armature (such as that of a dynamotor or a two-voltage generator) that has two separate windings on a single core, and has two separate commutators.

double-balanced mixer See BALANCED MIXER.

double-balanced modulator See BALANCED MODULATOR.

double-base diode See UNIJUNCTION TRANSISTOR.

double-base junction transistor A junction transistor having the usual emitter, base, and collector electrodes, plus two base connections, one on either side of the base region. The additional base connection acts as a fourth electrode to which a control voltage is applied. Also called *tetrode transistor.*

double-beam CRT See DUAL-BEAM OSCILLOSCOPE.

double-beam oscilloscope See DUAL-BEAM OSCILLOSCOPE.

double-bounce calibration In radar operations, a calibration technique for determining zero-beat error. Round-trip echoes are observed, the correct range being the difference between the two echoes.

double-bounce signal A signal that is received after having been reflected twice.

double-break contacts The member of a set of contacts that is normally closed on two others. Compare DOUBLE-MAKE CONTACTS.

double-break switch A switch that opens a previously closed circuit at two points simultaneously on closing. Compare DOUBLE-MAKE SWITCH.

double bridge See KELVIN DOUBLE BRIDGE.

double buffering In the input/output operation of a computer peripheral, the use of two memory areas for temporary storage.

double-button microphone A carbon microphone having two buttons mounted on each side of the center of a stretched diaphragm, and connected in push-pull. Also see BUTTON MICROPHONE.

double-channel duplex Two-way communication over two independent channels. One station transmits on one channel, and the other station transmits on the other channel. The result is conversation-mode communications, in which one operator can interrupt the other at any time; both receivers are always operational.

double-channel simplex A system of communication in which two channels are used. One station transmits on one channel, and the other station transmits on the other channel. Interruption is not possible because whenever either operator transmits, the station receiver is muted.

double-checkerboard pattern In a magnetic core memory, the maximum noise that appears when half of the half-selected cores are in the one state and the others are in the zero state. Also called *worst-case noise pattern.*

double circuit tuning A circuit whose output and input are tuned separately. Such tuning provides increased selectivity when the input and output are resonant at the same frequency, and decreased selectivity when they are tuned to different frequencies. Also see DOUBLE-TUNED AMPLIFIER.

double clocking A phenomenon that occurs in some digital circuits when the input pulse is nonuniform, and appears as two pulses to the device. The device is thus actuated at twice the desired frequency.

double-coil direction finder A radio direction finder (RDF) using an antenna that consists of two identical, perpendicular coils. The directivity of the antenna is the resultant of the directivity of the individual coils.

double conversion **1.** Two complete frequency conversions in a superheterodyne system. For example, the incoming signal might be converted to a 9-MHz first intermediate frequency (first IF); at a later stage, this signal might be converted to a 455-kHz second IF. The high first IF widely separates the signal from the image; the low second IF allows superior selectivity to be achieved at a reasonable cost. Also called *dual conversion.* **2.** Pertaining to a superheterodyne receiver with two intermediate frequencies.

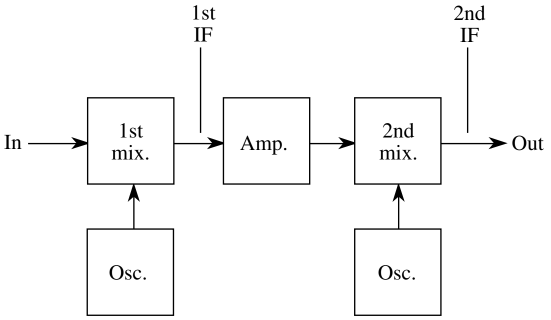

double conversion

double-conversion receiver Also called *double-conversion superheterodyne.* A superheterodyne receiver using DOUBLE CONVERSION to achieve optimum selectivity and image rejection.

double-current generator **1.** A dynamo-type generator supplying both alternating current (ac) and direct current (dc) from one armature winding. **2.** A rotary converter operating on dc and delivering ac.

double-diamond antenna A broadband antenna consisting of two rhomboid plates, one attached to each side of the feeder.

double-diffused epitaxial mesa transistor A transistor in which a thin mesa crystal is overlaid on another mesa crystal. Also called *epitaxial-growth mesa transistor.*

double-diffused transistor See DIFFUSED-EMITTER-AND-BASE TRANSISTOR.

double diode See DUODIODE.

double-diode limiter A limiter in which two diodes are connected back to back in parallel, to limit both peaks of an alternating-current (ac) signal.

double-doped transistor See GROWN-JUNCTION TRANSISTOR.

double edit In audio tape recording, to make two changes in a given span of the tape. For example, a producer might dislike the wording of a certain sentence, and re-record the sentence. Then, changing his or her mind, the producer might record the original sentence back over the re-recorded sentence. These changes increase the risk of audible irregularities appearing in the final recording.

double emitter follower See COMPOUND CONNECTION.

double-ended amplifier See PUSH-PULL AMPLIFIER and DOUBLE-ENDED CIRCUIT.

double-ended circuit A symmetrical circuit (i.e., one having identical halves, each operating on a half-cycle of the input signal). Example: a push-pull amplifier.

double-extended Zepp antenna A horizontal, collinear, center-fed antenna, in which each section measures 0.65 wavelength. This antenna gives increased gain over that of the Zepp and double Zepp (see DOUBLE ZEPP ANTENNA).

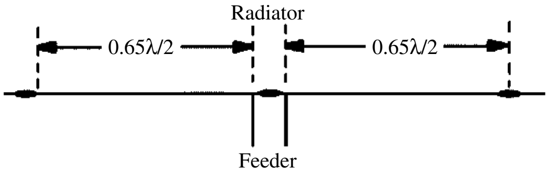

double-extended Zepp antenna

double-hump resonance curve A resonant response that is flattened by double tuning; it exhibits two resonance peaks. Also see DOUBLE-TUNED AMPLIFIER.

double-hump wave See DOUBLE-PULSE WAVE.

double image Two overlapping television pictures, one usually fainter than the other. Caused by the signal arriving over two different paths (one possibly attributable to reflection of the wave) and, hence, at different instants in time. The fainter image is called a GHOST.

double insulation The use of two layers of insulation on a conductor, made of different materials.

double ionization Ionization resulting from an electron colliding with an ion. In a gas, for example, a neutral atom might collide with an electron, which can knock an electron out of the atom. The atom then becomes a positive ion; it might in turn be bombarded by an electron, releasing still another electron.

double-junction photosensitive semiconductor See PHOTOTRANSISTOR.

double layer See HELMHOLTZ DOUBLE LAYER.

double local oscillator A mixer system in which a local oscillator generates two accurate radio-frequency (RF) signals separated by a few hundred hertz. The difference frequency is used as a reference in some applications.

double-make contacts A set of normally open contacts of which one closes against two others simultaneously. Compare DOUBLE-BREAK CONTACTS.

double-make switch A switch that closes a previously open circuit at two points simultaneously. Compare DOUBLE-BREAK SWITCH.

double moding In microwave operations, the abrupt changing of frequency at irregular intervals.

double modulation Using a modulated carrier to modulate another carrier of a different frequency.

double-play tape A thin magnetic recording tape that has approximately twice the playing time of the usual tape. Although the playing time is longer, double-play tape is more subject to jamming and stretching than standard-thickness recording tape.

double-pole Having two poles or switchable circuits (e.g., a double-pole switch).

double-pole, double-throw switch or relay Abbreviation, DPDT. A switch or relay having two contacts that can be closed simultaneously in one of two directions, to close or open two circuits.

double precision The use of two computer words to represent a single number to gain precision.

double-pole, single-throw switch or relay Abbreviation, DPST. A switch or relay having two contacts that can be closed in only one direction, to simultaneously close or open two circuits.

double precision hardware Within a computer, arithmetic units permitting the use of double-precision operands, sometimes also accommodating floating-point arithmetic.

double-precision number In digital computer operations, a number represented by two words for greater precision.

double pulse reading Pertaining to a magnetic core in a computer memory, recording bits as two states held simultaneously by one core having two areas that can be magnetized with alternate polarities. For example, positive-negative could represent zero, and negative-positive could represent one.

double-pulse wave An alternating-current (ac) wave having two successive positive peaks followed by two successive negative peaks within each cycle. The output voltage of a varistor bridge has such a waveshape for an ac input.

double-pulsing station A loran station that transmits at two pulse rates upon receiving two pairs of pulses.

double pumping A method of obtaining increased peak output power from a laser by pumping it for a comparatively long interval and then immediately pumping it for a short interval.

doubler **1.** A circuit or device for multiplying a frequency by two (see FREQUENCY DOUBLER). **2.** A circuit or device for multiplying a voltage by two (see VOLTAGE DOUBLER).

double probe A test probe that multiplies an applied signal voltage by two, so it can be handled more effectively by the instrument with which the probe is used.

double punching In perforating a punched card, putting two holes in one column; it is an error if it occurs in a field of a card that is part of a record.

double rail A form of logic system in which two lines are used, with three possible states. The output can be high, low, or undecided.

double response **1.** Two-point response, as that associated with tuning a receiver to a signal and then to its image. **2.** See DOUBLE-HUMP RESONANCE CURVE.

double screen A cathode-ray tube having a two-layer screen on which there is an additional, long-persistence coating of a different color.

double shield Two independent electromagnetic shields for a circuit enclosure or cable. The shielding structures are concentric, and can be connected together at a single point (the common point).

double sideband Abbreviation, DSB. In a modulated signal, the presence of both sidebands.

double-sideband suppressed carrier Abbreviation, DSSC. A transmission technique in which both sideband products of modulation are transmitted and the carrier is suppressed. Compare LOWER SIDEBAND, SUPPRESSED CARRIER, UPPER SIDEBAND, and SUPPRESSED CARRIER.

double-sideband system A modulation or demodulation system utilizing both sidebands, with or without the carrier.

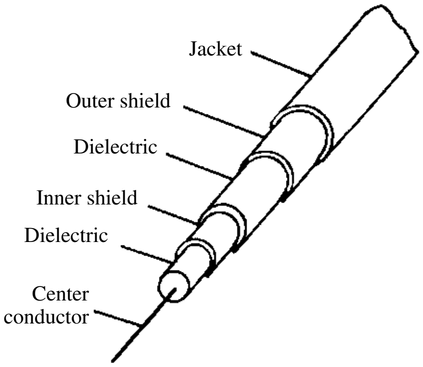

double shield

double-sideband transmitter A modulated transmitter using a double-sideband system.

double signal In reception, the property of having a signal on either side of the carrier frequency, as in a direct-conversion receiver. The two signals represent the sum and difference of the local oscillator signal and the actual received signal. Because the two signals convey identical information, the phenomenon is wasteful of received spectrum, and degrading to receiver selectivity and sensitivity.

double-signal receiver A receiver, such as a direct-conversion type, in which the signals occur in replicated form on either side of the local oscillator signal. Compare SINGLE SIGNAL.

double-spot tuning In a superheterodyne receiver, tuning in the same signal at two different places on the dial, a condition caused by image response.

double squirrel-cage induction motor A polyphase induction motor having a double squirrel-cage rotor. The rotor slots contain two bars, an upper bar having low reactance (being near the air gap) and high resistance, and a lower bar having high reactance and low resistance. This motor has low starting current, high starting torque, and a full-load slip of less than 5%.

double-stream amplifier A traveling-wave tube in which microwave amplification results from the interaction of two electron beams of different average velocity.

double-stub tuner Two stubs (see STUB) connected in parallel with a transmission line and usually spaced 0.375 wavelength (135 electrical degrees) apart; it is used as an impedance matcher.

double superheterodyne See DOUBLE-CONVERSION SUPERHETERODYNE.

double superheterodyne reception See DOUBLE-CONVERSION SUPERHETERODYNE.

double-surface transistor See COAXIAL TRANSISTOR.

doublet See DIPOLE ANTENNA.

double-throw Operating in opposite directions as selected (e.g., a double-throw relay or switch).

double-throw circuit breaker A circuit breaker that closes in both its pull-in and dropout positions.

double-throw switch or relay A switch or relay having two ganged poles.

double-trace recorder See DOUBLE-TRACK RECORDER, **2**.

double tracing Displaying two signals simultaneously on the screen of an oscilloscope through the use of an electronic switch.

double-track recorder **1.** A tape recorder whose head is positioned so that separate recordings can be made as two tracks on the tape. **2.** A graphic recorder that produces two separate parallel tracings.

doublet trigger A two-pulse, constant-spaced trigger signal used for coding.

double-tuned amplifier An amplifier whose input and output circuits are both tuned.

double-tuned circuit A circuit, such as an amplifier or filter, using separate input and output tuning. Also see DOUBLE CIRCUIT TUNING and DOUBLE-TUNED AMPLIFIER.

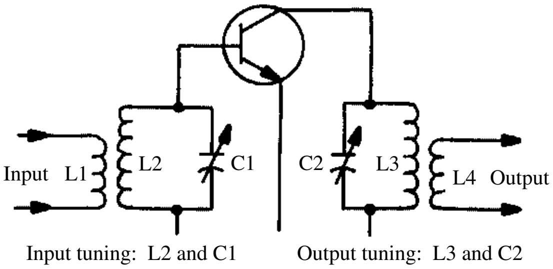

double-tuned circuit

double-tuned detector A form of frequency-modulation (FM) discriminator with two resonant circuits. One is tuned slightly higher than the channel center frequency, and the other is tuned an equal amount below the center.

double-vee antenna A broadband, modified dipole antenna resembling two vees in line. Also see VEE ANTENNA.

double-winding generator A dynamo-type generator having separate armature windings for supplying two voltages, either of which can be direct (dc voltage) or alternating (ac voltage).

double-wye rectifier A heavy-load circuit using six rectifier diodes, each conducting for 120 degrees of the cycle. An interphase winding is used. The circuit is equivalent to two three-phase, half-wave rectifiers connected in parallel.

double-Y rectifier See DOUBLE-WYE RECTIFIER.

double Zepp antenna A usually horizontal, straight, center-fed, full-wavelength antenna.

Also called *two half waves in phase.* Its name was derived because it is, in fact, two Zepp antennas forming a collinear array.

doubling **1.** Producing the second harmonic of a signal. **2.** In communication, unintentional simultaneous transmission by both operators, resulting in missed information. **3.** In a speaker, distortion resulting in large amounts of second-harmonic output.

doubly balanced modulator See BALANCED MODULATOR.

doughnut capacitor See DONUT CAPACITOR.

doughnut coil See TOROIDAL COIL.

doughnut crystal See DONUT CRYSTAL.

doughnut magnet See RING MAGNET.

down convert In superheterodyne conversion, to heterodyne a signal to an intermediate frequency lower than the signal frequency. Compare UP CONVERT.

down lead See LEAD-IN.

downlink The signal sent down from an active communications satellite to the earth, usually on a different frequency than the signal sent up. See UPLINK.

downlink beamwidth The angle subtended between the half-power points of the downlink signal from an active communications satellite.

downlink frequency The frequency of the downlink signal from an active communications satellite. Usually, the downlink signals occupy a certain band of frequencies, anywhere from several kilohertz to several megahertz wide.

downlink power **1.** The output power of the downlink transmitter in an active communications satellite. **2.** The effective radiated power (ERP) of the downlink signal from an active communications satellite.

down time A period of time during which electronic equipment is completely inoperative (for any reason).

downturn A usually sudden dip in a performance curve. Compare UPTURN.

downward modulation Modulation in which the average carrier component decreases during modulation. Example: amplitude modulation of a transmitter in which the antenna current decreases during modulation. Compare UPWARD MODULATION.

DP Abbreviation of DATA PROCESSING.

DPDT Abbreviation of *double-pole, double-throw* (switch or relay).

DPM **1.** Abbreviation of *digital power meter.* **2.** Abbreviation of DIGITAL PANEL METER. **3.** Abbreviation of *disintegrations per minute.*

DPS Abbreviaton of *disintegrations per second.*

DPST Abbreviation of *double-pole, single-throw* (switch or relay).

dr Abbreviation of *dram.*

drag **1.** A retarding force, caused by friction, acting on a moving body in contact with another moving or stationary body or medium. **2.** A retarding force introduced by an applied magnetic or electric field.

drag angle In disk recording, an angle of less than 90° between the stylus and the disk. The acute angle causes the stylus to drag instead of digging in.

drag cup A cup of nonmagnetic metal (usually copper or aluminum) that, when rotated in a magnetic field, acquires a voltage proportional to the speed of rotation. The device is often used as a brake.

drag-cup motor A servomotor whose shaft has a copper or aluminum drag cup that rotates in the field of a two-phase stator. Eddy currents set up in the cup by the field winding produce torque; braking action, direction control, and speed control are obtainable by means of associated electronics.

drag magnet In a motor-type meter, a braking magnet (i.e., one used to reduce speed through eddy-current effects). Also called *retarding magnet.*

drain **1.** The current or power drawn from a signal or power source. **2.** A load that absorbs current or power. **3.** The electrode in a field-effect transistor (FET) from which the output is usually taken; equivalent to the collector of a bipolar transistor.

drainage equipment Devices and systems for protecting circuits against transients generated by circuit breakers and similar safety devices.

drain-coupled multivibrator An oscillator that uses two field-effect transistors (FETs) in the circuit equivalent of a collector-coupled bipolar-transistor multivibrator. The drain of one stage is capacitance-coupled to the gate of the other stage.

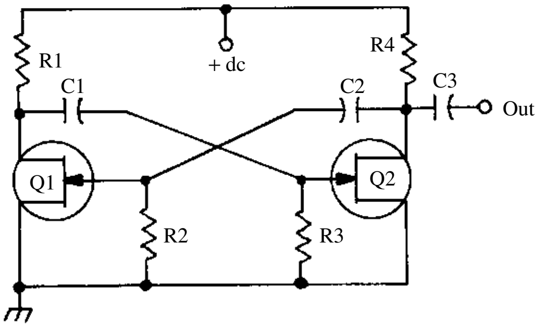

drain-coupled multivibrator

D-region A low region of the ionosphere beneath the E-region, whose ionization varies with the inclination of the sun. The greatest ionization is at midday; the layer disappears at night.

dress The (usually experimental) arrangement of leads for optimum circuit operation (minimum capacitance, best suppression of oscillation, minimum pickup, etc.).

dressed contact A contact having a permanently attached locking spring member.

drift **1.** Within a conductor or semiconductor, the controlled, directed movement of charge carriers resulting from an applied electric field. **2.** A usually gradual and undesirable change in a quantity, such as current, as a result of a disturbing factor, such as temperature or age.

drift current In a semiconductor, the current resulting from a flow of charge carriers in the presence of an electric field. The charge carriers are electrons in n-type material and holes in p-type material.

drift field The inherent internal electric field of a DRIFT-FIELD TRANSISTOR.

drift-field transistor An alloy-junction, bipolar, radio-frequency (RF) transistor for which the impurity concentration is graded from high on the emitter side of the base wafer to low on the collector side. This creates an internal drift field that accelerates current carriers and raises the upper frequency limit of the transistor.

drift-matched components Active or passive components that have been closely matched in terms of the drift of one or more parameters, with respect to time, temperature, etc.

drift mobility For current carriers in a semiconductor, the average drift velocity per unit electric field.

drift space **1.** In a vacuum tube, a space that is nearly free of alternating-current (ac) fields from the outside, and in which the repositioning of electrons is governed by the space-charge forces and the velocity distribution of the electrons. **2**. In a Klystron, the space between buncher and catcher cavities in which there is no field.

drift speed The average velocity of charge carriers moving through a medium.

drift transistor See DRIFT-FIELD TRANSISTOR.

drift velocity The net velocity of a charged particle (electron, hole, or ion) in the direction of the field applied to the conducting medium.

drift voltage The usually gradual change in voltage resulting from such causes as internal heating. Also called voltage drift.

drip loop In a transmission line for an antenna or power service, a loop near the point of entry to the building for the purpose of allowing condensation or rain water to drip off.

drip-proof motor A motor with ventilating apertures arranged so that moisture and particles cannot enter the machine.

drip-tight enclosure A housing designed to prevent entry of rain, snow, and dust; it also prevents accidental contact with the enclosed apparatus or machinery.

drive **1.** To excite (i.e., to supply with input-signal current, power, or voltage) (see DRIVING CURRENT, DRIVING POWER, and DRIVING VOLTAGE). **2.** Input-signal excitation (see DRIVING CURRENT, DRIVING POWER, and DRIVING VOLTAGE). **3.** A device that moves a recording medium (e.g., *tape drive* and *diskette drive*). **4.** The transmission of mechanical energy from one place to another (e.g., *motor drive*).

drive array A set of two or more hard-disk drives in a computer system. They function together to minimize the possibility of data loss. Such a system can also increase the amount of fast-access data storage.

drive belt A continuous belt used to transmit mechanical energy from a driving pulley to a driven pulley.

drive circuit **1.** A circuit used to provide the excitation to a motor. **2.** An amplifier that supplies drive to a more powerful amplifier.

drive control In a television receiver, the potentiometer used to adjust the ratio of horizontal pulse amplitude to the level of the linear portion of the sawtooth scanning-current wave.

driven element In a multielement antenna, an element to which electromagnetic energy is fed directly, as opposed to a PARASITIC ELEMENT, which is excited by a nearby radiator element.

driven-element directive antenna A multielement directional antenna whose elements are driven from the feed line (i.e., no element is parasitic). Compare PARASITIC-ELEMENT DIRECTIVE ANTENNA.

driven multivibrator A multivibrator whose operation or frequency is controlled by an external synchronizing or triggering voltage. Compare FREE-RUNNING MULTIVIBRATOR.

driven single sweep A single oscilloscope sweep that is initiated by the signal under observation.

drive pattern A pattern of interference in a facsimile system that is caused by improper synchronization of the recording spot.

driven sweep An oscilloscope sweep that is initiated by the signal under observation.

drive pin A pin used to prevent a record from slipping on the rotating turntable of a recorder or reproducer. It is similar to, and located near, the center pin of the turntable.

drive pulse In digital computer operations, a pulse that magnetizes a cell in a memory bank.

driver **1.** A device that supplies a useful amount of signal energy to another device to ensure its proper operation (e.g., a current driver for a magnetic-core memory, an oscillator driving a loudspeaker). **2.** A power amplifier stage that supplies signal power to a higher-powered amplifier stage. **3.** In a digital computer, a stage that increases the output current or power of another stage (e.g., a clock driver). **4.** The cone and magnet of a dynamic speaker.

driver element In a multielement directive antenna, the element excited directly by the feeder, the other elements (directors and reflectors) being parasitic.

driver impedance **1.** The output impedance of a driver stage. **2.** The impedance "seen" from the driven stage of an amplifier, through the driver transformer, to the driver stage. It is the vector sum of driver reactance and resistance.

driver inductance In an amplifier's driver transformer, the inductance, as "seen" looking through the transformer from the driven stage into the driver stage.

driver resistance In an amplifier's driver transformer, the resistance "seen" looking through the transformer from the driven stage into the driver stage.

driver stage An amplifier stage whose chief purpose is to supply excitation (input-signal current, power, or voltage) to the next stage. Also see DRIVER.

driver transformer The transformer that couples a driver stage to a driven stage. Example: the interstage transformer inserted between the collector of a single-ended driver transistor and the two bases of a push-pull power-output stage in an audio amplifier.

driving current In a power amplifier, the input signal current required to produce a given amount of output power.

driving-point admittance The reciprocal of DRIVING-POINT IMPEDANCE.

driving-point impedance The input impedance of a network.

driving power In a power amplifier, the input signal power required to produce a given amount of output power.

drive wire The wire forming the coil around the toroidal cell in a magnetic core memory; supplies pulses that magnetize the cell.

driving-range potential In cathodic protection, the difference of potential between the anode and (protected) cathode.

driving signal **1.** Drive (see DRIVE, **2**). **2.** In television, time-scanning signals (line-frequency pulses and field-frequency pulses) at the pickup location.

driving spring In a stepping relay, the spring that moves the wiper blades.

driving voltage In a power amplifier, the input signal voltage required to produce a given amount of output power.

DRO Abbreviation of DIGITAL READOUT.

drone A pilotless radio-controlled aircraft without a human pilot.

drone cone An undriven loudspeaker cone that is mounted in a bass-reflex enclosure with other speakers. Also called PASSIVE RADIATOR.

droop **1.** A dip in the graph of a function. **2.** In a pulse train, the decrease in mean amplitude (in percent of maximum amplitude) at a given instant after attainment of maximum amplitude.

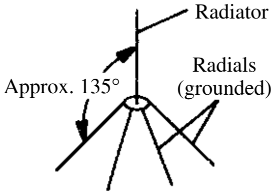

drooping radials

drooping radials In a ground-plane antenna, radials that slope downward to provide a transmission-line impedance match. The slope angle depends on the characteristic impedance of the line; typically, the angle is between 45 degrees and 70 degrees, relative to horizontal.

drop **1.** In wire communications, the line connecting a telephone cable to a subscriber's building. **2.** See VOLTAGE DROP.

drop bar A device that automatically grounds or short-circuits a capacitor when the door of a protective enclosure is opened.

drop cable See DISTRIBUTING CABLE, **1**.

drop channel In a communications system utilizing several channels, a channel that is not used.

drop-in The unintentional creation of bits when a magnetic storage device is being read from or written into. Compare DROP-OUT, **4**.

drop indicator In a signaling system, such as an annunciator, a hinged flap that drops into view when the signaling device is actuated.

drop-out **1.** The opening of a relay or circuit breaker. **2.** In digital computer operations, variation in signal level of the reproduced tape-recorded data. Such variation can result in errors in data reproduction. **3.** In the production of monolithic circuits, a special image placed at a desired point on the photomask. **4.** Digit loss during a read or write operation involving a magnetic storage device.

dropout current See DROPOUT VALUE.

dropout power See DROPOUT VALUE.

dropout value The level of current, power, or voltage at which a device, such as a circuit breaker or relay, is released.

dropout voltage See DROPOUT VALUE.

dropping resistor A series resistor providing a voltage reduction equal to the voltage drop across itself. For example, a 1000-ohm resistor in series with a 45-V battery, and carrying a current of 10 mA, will provide a voltage reduction equal to 10 V ($IR = 0.01 \times 1000 = 10$ V), thus dropping the 45 V to 35 V.

drop relay In a telephone system, a relay that is activated by the ringing signal. The relay is used to switch on a buzzer, light, or other device.

drop repeater A repeater intended for a termination of a communications circuit in a telephone system.

dropsonde A parachute-supported radiosonde dropped from a high-flying aircraft.

drop-tracks The tracks of radioactive particles made visible by moisture in an ionization chamber.

drop wire A wire that runs from a building to a pole (for line extension) or to a cable terminal (for cable extension).

drum **1.** A rotating cylinder coated with a magnetic material on which digital information can be recorded in the form of tiny magnetized spots. These spots are read as the drum rotates under pickup heads, or erased when the stored information is no longer needed. **2.** In some graphic recorders, facsimile receivers, etc., a rotating cylinder carrying the recording sheet.

drum capacitor See CONCENTRIC CAPACITOR.

drum controller The device that regulates the recording process on a drum memory.

drum mark On a track of a magnetic drum, a character that signifies the end of a character group.

drum memory In digital computers, a memory based on a magnetic drum (see DRUM, **1**). They have been largely replaced in recent years by electronic random-access memory, in the form of integrated circuits (ICs) and/or PCMCIA standard adapter cards.

drum parity The degree of accuracy in a drum recording/reproducing system.

drum programmer A device for sequencing operations. Its heart is a rotating drum, around whose surface contacts or points can be placed to actuate or terminate operations at selected times.

drum receiver A facsimile receiver using recording paper or photographic film wound around a revolving drum.

drum recorder A graphic recorder in which the record sheet is wound around a rotating drum.

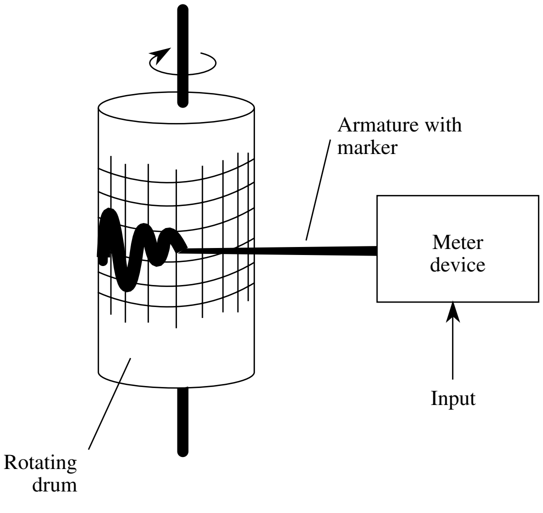

drum recorder

drum resistor A resistor consisting of a hollow cylinder of resistive material. Such a resistor can be cooled by circulating air or liquid through the cylinder.

drum speed The speed, usually measured in revolutions per minute (rpm), of the rotating drum in a graphic recorder, facsimile transmitter, or facsimile receiver.

drum storage The storage of data as magnetic impulses on a cylindrical, or drum-shaped, medium. Largely supplanted in recent years by magnetic disks, optical disks, and magnetic tapes.

drum switch A sequential switch whose contacts are pins or teeth placed at points around the outside of a revolving drum.

drum transmitter A facsimile transmitter in which the sheet bearing the material to be transmitted is wound around a revolving drum.

drum-type controller A motor-driven drum switch arranged to time various operations through sequential switching.

drum varistor A varistor that is a hollow cylinder of nonlinear resistance material. This varistor can be cooled by circulating air or liquid through it.

drum winding In a motor or generator, an armature whose conductors are on the outer face of the core, the two branches of a turn lying under adjacent poles of opposite polarity.

drunkometer An instrument for testing the extent of alcoholic intoxication. It electronically measures blood alcohol content through analysis of the subject's breath.

dry In an electric cell, a term used to describe an electrolyte that is semiliquid or solid.

dry battery A battery of dry cells.

dry cell **1.** A Leclanche primary cell in which the positive electrode is carbon, the negative electrode is zinc, and the electrolyte is a gel of ammonium chloride and additives. Also see CELL and PRIMARY CELL. **2.** A cell whose electrolyte is a gel or paste.

dry circuit A circuit in which the maximum voltage is 50 mV and the maximum current 200 mA.

dry-contact rectifier See DRY-DISK RECTIFIER.

dry contacts Contacts that neither make nor break a circuit.

dry-disk rectifier A solid-state rectifier, such as a copper-oxide, magnesium-copper-sulfide, or selenium type, that consists of a metal disk coated with a semiconductor material. The name was originally used to distinguish this rectifier from the *wet electrolytic rectifier.*

dry electrolytic capacitor An electrolytic capacitor whose electrolyte is a paste or solid. Compare WET ELECTROLYTIC CAPACITOR.

dry flashover voltage The breakdown voltage between electrodes in dry air when all insulation is clean and dry.

dry pile A voltaic pile containing numerous disks silvered or tinned on one face and covered with manganese dioxide on the other.

dry reed A metal contact, generally used as a relay or switch, that moves toward or away from another fixed contact under the influence of a magnetic field.

dry-reed relay See DRY-REED SWITCH.

dry-reed switch A switch consisting of two thin, metallic strips (reeds) hermetically sealed in a glass tube. The tube is surrounded by a coil of wire. When a current flows in the coil, a magnetic field affects the reeds. In the *normally open dry-reed switch*, the magnetic field causes the reeds to come together and close the circuit. In the *normally closed dry-reed switch*, the magnetic field causes the reeds to separate, opening the circuit. Compare MERCURY-WETTED REED RELAY.

dry run **1.** The preliminary operation of equipment for testing and appraisal. Such a procedure precedes putting the equipment into regular service. **2.** A step-by-step, paper-and-pencil "run" of a computer program before it is machine-implemented.

dry shelf life The life of a battery cell stored without its electrolyte.

dry-transfer process A method of transferring printed-circuit patterns and panel labels from sheets by rubbing them onto the substrate or panel.

dry-type forced-air-cooled transformer A DRY-TYPE TRANSFORMER that is cooled by convection of air circulated by a blower or fan. This increases the amount of power that the transformer can safely handle.

dry-type self-cooled transformer A DRY-TYPE TRANSFORMER that is cooled by natural air circulation (convection), without the use of a blower or fan.

dry-type transformer A transformer that, rather than being immersed in oil, is cooled entirely by the circulation of air.

DSB Abbreviation of DOUBLE SIDEBAND.

DSBSC Abbreviation of DOUBLE-SIDEBAND SUPPRESSED CARRIER. Also abbreviated DSSC.

dsc Abbreviation of *double silk covered* (wire).

D scope A radar whose display resembles that of a C scope, the difference being that blip height gives an approximation of the distance.

D service A Federal Aviation Agency service providing radio broadcasts of weather data, notices to aircraft personnel, and other advisory messages.

D-shell connector A multi-pin connector (either male or female) with a characteristic shape that ensures proper pin placement. Commonly used for computer ports, and for connections of peripherals to electronic equipment.

dsp Abbreviation of *double silver plated.*

DSR Abbreviation of DYNAMIC SPATIAL RECONSTRUCTOR.

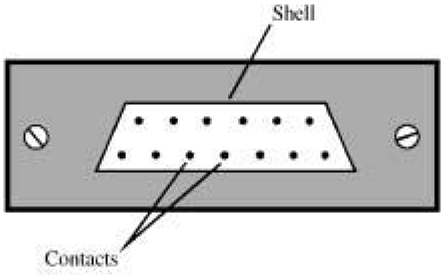

D-shell connector

DSS Abbreviation of *direct station selection* (telephone).

DSSC Abbreviation of DOUBLE-SIDEBAND SUPPRESSED CARRIER. Also abbreviated DSBSC.

DT Abbreviation of DATA TRANSMISSION.

DTA Abbreviation of *differential thermoanalysis.*

DT-cut crystal A piezoelectric plate cut from a quartz crystal at an angle of rotation about the z-axis of –53 degrees. It has a zero temperature coefficient of frequency at approximately 30 degrees Celsius. Also see CRYSTAL AXES and CRYSTAL CUTS.

DTL Abbreviation of DIODE-TRANSISTOR LOGIC.

DTn Abbreviation of DOUBLE TINNED.

DTS **1.** Abbreviation of DATA-TRANSMISSION SYSTEM. 2. Abbreviation of *digital telemetry system.*

DU Abbreviation of DUTY CYCLE.

dual **1.** Pertaining to a combination of two components such as diodes, transistors, etc., in a single housing. The components are often carefully matched. Compare QUAD. **2.** Pertaining to a device or circuit that behaves in a manner analogous to that of another operating with component and parameter counterparts. Thus, a current amplifier can be the dual of a voltage amplifier; a series-resonant circuit, the dual of a parallel-resonant circuit; or a field-effect transistor, the dual of a bipolar transistor.

dual-beam CRT A cathode-ray tube having two separate electron guns, for use in a dual-beam oscilloscope.

dual-beam oscilloscope Also called *dual-trace oscilloscope.* An oscilloscope having two electron guns and deflection systems; it can display two phenomena on the screen simultaneously for comparison.

dual capacitor **1.** Two fixed capacitors combined in a single housing, sometimes sharing a common capacitor plate. **2.** A two-section, ganged variable capacitor.

dual-channel amplifier An amplifier having two separate, independent channels (e.g., a stereo high-fidelity audio amplifier).

dual-cone speaker A speaker designed for a wide range of audio frequencies. One cone responds to the bass (low) and midrange audio frequencies, and a smaller cone responds to the treble (high) audio frequencies.

dual diode A discrete component consisting of two diodes in one package.

dual-diversity receiver A receiver or receiver system for DIVERSITY RECEPTION.

dual-diversity reception See DIVERSITY RECEPTION.

dual-emitter transistor A low-level silicon pnp chopper transistor of the planar passivated epitaxial type; it has two emitter electrodes.

dual-frequency calibrator A secondary frequency standard providing two fundamental test frequencies (e.g., 100 kHz and 1 MHz).

dual-frequency induction heater An induction heater whose work coils carry energy of two different frequencies. The coils heat the work either simultaneously or successively.

dual gate **1.** A digital integrated circuit (IC) consisting of two logic gate units. **2.** Pertaining to a field-effect transistor (FET) with two gates or gate electrode connections.

dual-gate FET A field-effect transistor with two gate (input) electrodes.

dual-gate MOSFET A metal-oxide-semiconductor field-effect transistor (MOSFET) with two gate (input) electrodes.

dual-inline package Abbreviation, DIP. A flat, molded integrated-circuit (IC) package having terminal pins along both long edges.

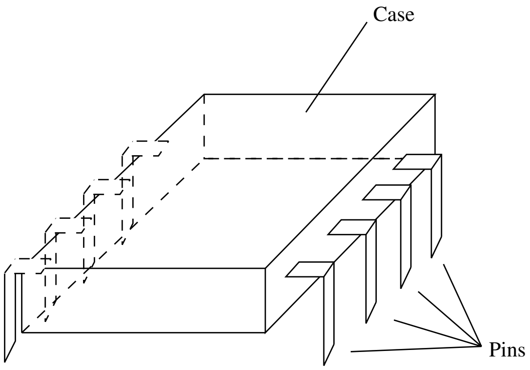

dual-inline package

duality **1.** The condition of being dual (see DUAL). This can be an aid in the design of certain circuits requiring complementary parameters, e.g., current-operated circuit analogs of voltage-operated circuits. **2.** See DUALITY OF NATURE.

duality of nature **1.** Any of various situations in which a phenomenon exhibits two distinct and different natures. A commonly cited example is the dual model of light. In some instances, visible light behaves like a barrage of particles, but in other environments it appears to be a wave effect. Another example is the dual model of electrostatic energy, behaving as point charges in some scenarios and as force fields in other situations. **2.** The tendency of a set of principles to be duplicated in sense by predictable analogies, as between inductance and capacitance, electrostatics and magnetics, etc.

dual local oscillator See DOUBLE LOCAL OSCILLATOR.

dual meter A meter having two meter movements and scales in a single case; the arrangement permits simultaneous monitoring of two quantities.

dual modulation The modulation of a single carrier or subcarrier by two different types of modulation—each carrying different information.

dual network A network that is the dual of another network having complementary parameters. For example, a common-emitter, current-sensitive, bipolar-transistor circuit is the dual of a common-source, voltage-sensitive, field-effect-transistor (FET) circuit. Also see DUALITY.

dual operation In digital logic, the operation resulting from inverting all of the digits. Every 1 is replaced with a 0, and vice versa.

dual-output power supply A power supply with two outputs. Often, one output is positive and the other is negative. In some cases, one output consists of alternating current (ac) and the other consists of direct current (dc).

dual pickup In disk reproduction, a pickup having two styli, one for large-groove records and one for fine-groove records.

dual potentiometer A ganged assembly of two potentiometers. The resistance values might or might not be the same.

dual preset counter A preset counter that will set alternately to two different numbers.

dual rail See DOUBLE RAIL.

dual resistor See DUAL POTENTIOMETER and DUAL RHEOSTAT.

dual rheostat A ganged assembly of two rheostats. The resistance values might or might not be the same.

dual stereo amplifier **1.** A two-channel audio amplifier for stereophonic audio applications. **2.** A two-channel linear integrated circuit (IC) for stereophonic audio applications.

dual-system loudspeaker See TWO-WAY SPEAKER.

dual trace In a cathode-ray oscilloscope, the use of two separate electron beams, which can show two different signals simultaneously on a single screen.

dual-trace recorder See DOUBLE-TRACK RECORDER, **2**.

dual-track recorder See DOUBLE-TRACK RECORDER.

dual recording In digital computer operations, updating two sets of master files simultaneously.

dual use The use of a communications system for two modes of data transfer at the same time.

Duant electrometer See BINANT ELECTROMETER.

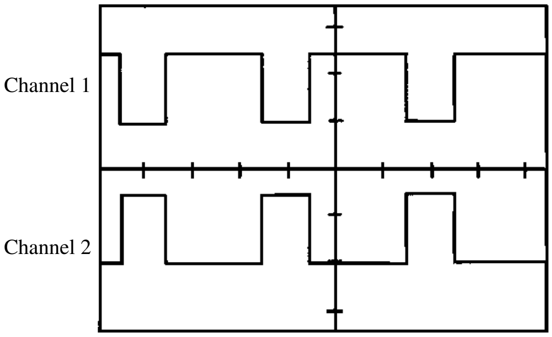

dual trace

duct 1. A narrow propagation path, sometimes traveled by microwaves, created by unusual atmospheric conditions. **2.** A pipe or channel for cables and wires.

dubbing The adding of sound to a recorded magnetic tape, record disk, or film (e.g., replacing the sound track of a film in one language with that of another language).

dubnium Symbol, Db. Also clalled *unnilpentium* (Unp) and *hahnium* (Ha). Atomic number, 105. The most common isotope has atomic weight 262. Classified as a transition metal. It has a half-life on the order of a few seconds to a few tenths of a second (depending on the isotope), is human-made, and is not known to occur in nature.

duct 1. A narrow propagation path, sometimes traveled by microwaves, created by unusual atmospheric conditions. **2.** A pipe or channel for cables and wires. **3.** An opening, vent, or other airway used for various purposes, such as cooling and acoustic wave transmission.

ducted port An opening in a speaker cabinet that has an airway (duct) extending several inches into the cabinet. It improves the quality of sound from a speaker system by equalizing the air pressure inside and outside the cabinet. Also provides resonant audio effects at frequencies that depend on the dimensions of the duct.

duct effect see TROPOSPHERIC DUCTING.

ductilimeter An instrument used for measuring the ductility of metals.

ducting The confinement of a radio wave to a duct (see DUCT, **1**) between two layers of the atmosphere or between an atmospheric layer and the earth.

Duddell arc A carbon copper arc circuit that produces audible continuous waves. Consists of a series inductance-capacitance (LC) circuit shunting an electric arc.

Duerdoth's multiple feedback system In an amplifier, feedback through several paths to improve response over that afforded by single-path feedback. In a simple application of multiple feedback, a single external loop is augmented with unbypassed emitter resistors in the amplifier stages.

Duerdoth's stability margin A feedback-amplifier stability margin equal to a 6-dB increase in gain at low and high frequencies over beta values between 0.3 and somewhat less than 2. For higher beta values, Duerdoth adopts an angular margin (for example, 15°); below $\beta = 0.3$, no danger of instability is present.

dummy 1. A nonoperative model of a piece of equipment, usually assembled with dummy components (see DUMMY COMPONENT, **1**) for the purpose of developing a layout. **2.** DUMMY ANTENNA, DUMMY COMPONENT, or DUMMY LOAD. **3.** Part of a computer program that, rather than being useful for the problem at hand, only serves to satisfy some other format or logic requirement.

dummy antenna 1. A nonradiating device that serves as a load for a transmitter (i.e., it takes the place of the regular antenna during tests and adjustments of the transmitter). **2.** A device containing a network of discrete inductive, capacitive, and resistive elements, inserted between a radio-frequency signal generator and receiver to simulate a standard antenna.

dummy component 1. A nonoperative component used in developing a layout or package. **2.** A nonoperative component fraudulently included in a piece of equipment (e.g., an unwired transistor in a receiver circuit, a common occurrence during the early days of the transistor, when a 10-transistor radio brought more money than an 8-transistor radio, without regard to the circuit itself).

dummy instruction In a computer, a command that serves no operational purpose, other than to fill a format requirement.

dummy load 1. A load device, usually consisting of resistance without reactance, used to terminate a power generator or power amplifier during adjustments and tests. The load resistance is equal to the output impedance of the generator or amplifier. **2.** See DUMMY ANTENNA.

dummy resistor A power-type resistor used as a dummy load.

dump 1. In digital-computer operations, to transfer, completely or partially, the contents of memory into a peripheral. **2.** To switch off all power to a computer, deliberately or accidentally, thereby losing what is in the volatile memory.

dump and restart During a halt in a computer program run, to backtrack to the last dump point and use the data there to resume the run. Also see DUMP POINT.

dump check In digital-computer operations, the checking of all digits being transferred (see DUMP, **1**) to prevent errors when they are retransferred.

dumping To transfer the output at various stages in a computer program run to an external storage medium, so it will be available (in case of a failure) for the program's resumption from a point other than at the beginning.

dump point In writing a computer program, a point at which instructions are given to transfer data processed thus far to a storage medium that would be unaffected by a software or hardware failure. Also see DUMP AND RESTART and DUMPING.

dumping resistor **1.** See BLEEDER. **2.** A resistor having the minimum resistance permissible in a given situation. Used to discharge a capacitor, it acts to provide an alternative path to a potentially destructive short circuit.

duo Any pair of matched components, usually in a single package.

duodecal CRT base The 12-pin base of a cathode-ray tube. Also see BIDECAL, DIHEPTAL, and MAGNAL.

duodecal socket A 12-pin tube socket. Also see DUODECAL CRT BASE.

duodecimal **1.** Having 12 possibilities, states, choices, etc. **2.** Pertaining to the DUODECIMAL NUMBER SYSTEM. **3.** A number or numeral in the DUODECIMAL NUMBER SYSTEM.

duodecimal number system A system of numbering in which the radix, also called the base or modulus, is 12. The system uses the digits 0 through 9, plus two other characters (usually A and B) to represent 10 and 11. Thus, counting proceeds as 0, 1, 2, . . . , 9, A, B, 10, 11, 12, . . . , 19, 1A, 1B, 20, 21, 22, etc. At one time, some people seriously proposed that this system replace the DECIMAL NUMBER SYSTEM for general use.

duodiode See DUAL DIODE.

duolateral coil A multilayer, lattice-wound coil (see UNIVERSAL WINDING) in which the turns in successive layers are staggered slightly. Also called *honeycomb coil.*

duopole A two-pole all-pass device.

duplex **1.** A mode of communication in which two channels are used so that either operator in a conversation can interrupt the other at any time. **2.** The transmission of two messages over a single circuit, at the same time.

duplex artificial line In wire telephony, a balancing network that simulates the impedance of the actual line and the remote terminal equipment; it prevents an outgoing transmission from interfering with the local receiver.

duplex cable A cable consisting of a twisted pair of insulated stranded-wire conductors.

duplex channel A channel used for wire or radio DUPLEX OPERATION.

duplex communication See DUPLEX OPERATION.

duplex diode See DUAL DIODE.

duplexer In radar operations, a device operated by the transmitted pulse to automatically switch the antenna from the receiver to the transmitter.

duplexing assembly In a radar system, a device that automatically makes the receiver unresponsive to the outgoing transmitted signal while allowing incoming signals to reach the receiver easily. Also see TRANSMIT-RECEIVE SWITCH.

duplex computer system An installation of two computer systems, one standing by to take over in case the other fails.

duplex operation The simultaneous operation of a transmitter and receiver at a single location. This becomes possible (without mutual interference) through the use of two sufficiently separated carrier frequencies.

duplex system A system composed of two identical equipment sets—either of which will perform the intended function while the other stands by.

duplication check In digital-computer operations, the checking of an operation by doing it twice, using different methods in each case, to ensure the accuracy of results.

duplication house A professional person or company who makes high-quality copies of tape recordings (either audio or video). Charges vary, depending on the type and length of the recording.

duplicate To transfer data from one storage location to another. Compare DUMP, **1**.

dural See DURALUMIN.

duralumin An alloy of aluminum, copper, magnesium, manganese, and silicon. It offers strength with minimal weight.

duration control A potentiometer or variable capacitor for adjusting the duration of a pulse.

duration time The period during which a pulse is sustained (i.e., the interval between turn-on and turn-off time).

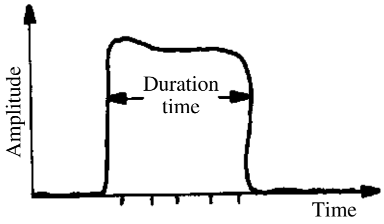

duration time

during cycle The interval during which a timer is in operation.

durometer An instrument for measuring the hardness of a material.

dust collector See DUST PRECIPITATOR.

dust core A magnetic core for radio-frequency coils consisting of very minute particles of iron or an alloy, such as Permalloy.

dust cover A removable, usually plastic cover for electronic and computer equipment, to protect hardware during periods of nonuse.

dust-ignition-proof motor A motor whose housing completely prevents the entry of dust, virtually eliminating the danger of fine dust sparking inside the machine.

dust precipitator An electrostatic device for removing dust, lint, and other particles from the air. It consists essentially of a pair of screens or wires through which the air passes; a potential of several thousand volts is maintained between them. The particles acquire a charge, then stick to the oppositely charged screen.

Dutch metal A copper-zinc alloy.

duty cycle **1.** The proportion or percentage of time during which a device, circuit, or system is operating or handling power. For example, when a radiotelegraph transmitter is keyed on and off to send Morse code, the duty cycle is approximately 50 percent; with frequency-shift keying, the duty cycle is 100 percent. **2.** The conditions under which an electrochemical cell or battery is used. In particular, the proportion or percentage of time during which current is drawn from the cell or battery for the purpose of operating an electrical or electronic circuit, device, or system.

duty cyclometer A direct-reading instrument for measuring duty cycle.

duty factor **1.** The ratio P_{avg}/P_{pk}, where P_{avg} is the average power in a system and P_{pk} is the peak power. **2.** The product of the duration and the repetition rate of regularly recurring pulses comprising a carrier.

duty ratio See DUTY FACTOR, 1.

DVM Abbreviation of DIGITAL VOLTMETER.

DVOM Abbreviation of DIGITAL VOLT-OHM-MILLIAMMETER.

dwell meter An instrument that shows the period (or angle) during which contacts remain closed.

dwell switching Switching action in which the contacts are held closed (or a circuit kept on) for specified periods, as opposed to MOMENTARY SWITCHING.

dwell tachometer A combination dwell meter/tachometer for automobile engine testing and adjustment. The dwell meter allows observation and adjustment of the ignition point cam angle; the tachometer shows the motor speed in revolutions per minute (rpm).

DX **1.** Radiotelegraph abbreviation meaning *long distance* or *foreign country.* **2.** A communication or broadcast station located far away and/or in a foreign country. **3.** Abbreviation of DUPLEX.

DXer An amateur radio operator who prefers to communicate with stations far away and/or in foreign countries.

Dy Symbol for DYSPROSIUM.

dyadic operation A binary operation (i.e., one using two operands).

dyn Abbreviation of DYNE.

dyna- A prefix (combined form) meaning *power* (e.g., dynamometer and dynatron).

dynamic acceleration Acceleration whose magnitude and direction are constantly changing.

dynamic allocation In multiprogramming, a system in which a monitor program assigns peripherals and areas of memory to a program.

dynamic analogy A mathematical similarity between or among various phenomena involving the motion of particles.

dynamic base current See AC BASE CURRENT.

dynamic base resistance See AC BASE RESISTANCE.

dynamic base voltage See AC BASE VOLTAGE.

dynamic behavior **1.** The behavior of a component, device, or system when signals are applied, as opposed to *static behavior* under no-signal conditions. **2.** The behavior of a device or system involving the motion of particles over a period of time.

dynamic braking A technique for stopping a motor quickly using a resistor (the dynamic braking resistance) connected across the spinning armature. The resistor dissipates the energy generated by the motor, producing a damping action that results in braking.

dynamic characteristic The performance characteristic of a device or circuit under alternating-current (ac) operating conditions, as opposed to the *static characteristic*, when only direct current (dc) flows.

dynamic check **1.** A test made under actual operating conditions of a device or circuit. **2.** A test made with an alternating-current (ac) applied signal, rather than with direct-current (dc) quantities.

dynamic collector current See AC COLLECTOR CURRENT.

dynamic collector resistance See AC COLLECTOR RESISTANCE.

dynamic collector voltage See AC COLLECTOR VOLTAGE.

dynamic contact resistance In relay or switch contacts, variation in the electrical resistance of the closed contacts because of variations in contact pressure.

dynamic convergence In a color picture tube, the meeting of the three beams at the aperture mask during scanning.

dynamic current amplification Abbreviation, DCA. An expression of gain in a bipolar transistor. Specified as the ratio of the difference in collector current I_C to the difference in base current I_B

$$\text{DCA} = dI_C/dI_B$$

Geometrically, the dynamic current amplification at a given bias point is the slope of a line tangent to the I_C-versus-I_B curve at that point.

dynamic curve A characteristic curve that accounts for the presence of resistance in series with the device to which the curve applies.

dynamic debugging Any debugging operation performed on a computer system during a normal-speed program run.

dynamic decay Decay resulting from such factors as ion charging in a storage tube.

dynamic demonstrator A teaching aid consisting of a board displaying an electronic circuit, behind which is mounted the actual circuit. Various circuit components (especially adjustable ones) are mounted on the front of the board, in clear view at places where their circuit symbols appear. Pin jacks at important test points in the circuit allow connection of a meter, signal generator, and oscilloscope leads for testing or demonstrating the circuit.

dynamic deviation The difference between ideal output and actual output of a circuit or device operating with a reference input that changes at a constant rate and is free of transients.

dynamic diode tester An instrument that displays the response curve (or family of curves) of a diode on a calibrated oscilloscope screen. The horizontal axis of the screen indicates voltage, the vertical axis shows current, and zeros for both quantities are at center screen. Also see DYNAMIC RECTIFIER TESTER.

dynamic drain current See AC DRAIN CURRENT.

dynamic drain resistance See AC DRAIN RESISTANCE.

dynamic drain voltage See AC DRAIN VOLTAGE.

dynamic dump A dump that occurs during a program run. See DUMPING.

dynamic electric field An electric field whose intensity is constantly changing, either periodically or in a complex way.

dynamic emitter current See AC EMITTER CURRENT.

dynamic emitter resistance See AC EMITTER RESISTANCE.

dynamic emitter voltage See AC EMITTER VOLTAGE.

dynamic equilibrium **1.** The state of balance between constantly varying quantities. **2.** The tendency of two current-carrying circuits to maintain at a maximum the magnetic flux linking them.

dynamic error In a periodic signal delivered by a transducer, an error resulting from the restricted dynamic response of the device.

dynamic flip-flop A flip-flop (bistable multivibrator) that is kept on by recirculating an alternating-current (ac) signal. The device can be switched on or off by a single pulse. Compare STATIC FLIP-FLOP.

dynamic focus Compensation for defocusing, caused by the electron beam sweeping in an arc across a flat color picture-tube screen; the method uses an alternating-current (ac) focusing-electrode voltage.

dynamic gate voltage See AC GATE VOLTAGE.

dynamic impedance The impedance of a device (such as a transistor or diode) when it is operating with an applied alternating-current (ac) signal, as opposed to its *static resistance* with only direct current (dc) applied.

dynamic limiter A limiter, such as is used in frequency-modulation (FM) receivers, that maintains the output-signal level, despite appreciable excursions of input-signal amplitude.

dynamic loudspeaker See DYNAMIC SPEAKER.

dynamic magnetic field A magnetic field whose intensity is constantly changing, either periodically or in a complex way.

dynamic memory A usually random-access data storage method in which the memory cells must be electrically refreshed periodically to avoid the loss of held data.

dynamic microphone A microphone in which a small coil attached to a vibrating diaphragm or cone moves in a uniform magnetic field to generate the output signal.

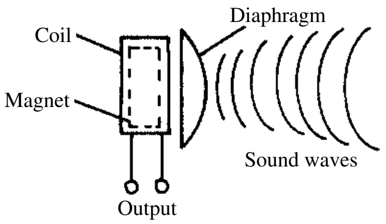

dynamic microphone

dynamic mutual conductance See DYNAMIC TRANSCONDUCTANCE.

dynamic noise suppressor A noise limiter consisting of an audio filter whose bandwidth is directly proportional to signal strength (i.e., it is varied automatically by signal amplitude).

dynamic operating line A curve displaying the control function of a device. For example, the collector-current-versus-base-current curve of a bipolar transistor is drawn between the limits of saturation and cutoff.

dynamic output impedance The output impedance of a power supply, as "seen" by the load.

dynamic pickup A phonograph pickup whose stylus causes a small coil to vibrate in the field of a permanent magnet. Works on the same principle as the DYNAMIC MICROPHONE.

dynamic printout A printout that occurs as a single function, actuated by one command, and completing itself in one operation.

dynamic problem checking A method of checking the solution obtained by an analog computer, to see that it makes sense (is not absurd).

dynamic programming A method of problem solving in which continual checks are made to ensure accuracy or conformance to a certain set of rules.

dynamic range **1.** In high-fidelity audio, the ratio of the loudest sound to the faintest sound that can be reproduced without significant distortion or noise. It is usually expressed in decibels. **2.** In a communications receiver, a measure of the ability to receive both weak and strong signals without excessive noise, distortion, desensitization or other undesirable effects. It is expressed in various ways, typically in decibels. **3.** The ratio between the loudest and faintest sounds, or between the strongest and weakest signals, encountered in a given environment or situation. It is usually expressed in decibels.

dynamic rectifier tester An instrument that displays the response curve of a rectifier on a calibrated oscilloscope screen. During the test, the rectifier receives an alternating-current (ac) voltage with a low positive peak and high negative peak, both corresponding to the rated forward and reverse voltages (respectively) of the rectifier. The horizontal axis of the screen indicates voltage, the vertical axis indicates current, and zeros for both quantities are at center screen.

dynamic regulation In an automatically regulated system, such as a voltage-regulated power supply, the transient response of the system. Dynamic regulation is determined from maximum overshoot and recovery time when the load or line value is suddenly changed.

dynamic regulator A circuit or device providing dynamic regulation.

dynamic reproducer **1.** See DYNAMIC MICROPHONE. **2.** See DYNAMIC PICKUP. **3.** See DYNAMIC SPEAKER.

dynamic resistance See DYNAMIC IMPEDANCE.

dynamic run See DYNAMIC CHECK, **1**. See also DYNAMIC DEBUGGING.

dynamics The study of bodies, charges, fields, forces, or pulses in motion. Compare STATICS.

dynamic sequential control In digital computer operation, the computer's changing the sequence of instructions during a run.

dynamic source current See AC SOURCE CURRENT.

dynamic source resistance See AC SOURCE RESISTANCE.

dynamic source voltage See AC SOURCE VOLTAGE.

dynamic spatial reconstructor Abbreviation, DSR. An advanced x-ray machine, developed at the Mayo Clinic, that displays organs in three-dimensional views in motion, and allows them to be electronically dissected without actually operating on the patient.

dynamic speaker A loudspeaker in which a small coil (voice coil), attached to a diaphragm or cone and carrying an audio-frequency signal current, moves back and forth in a permanent magnetic field and, accordingly, causes the diaphragm or cone to vibrate (emit sound). Compare MAGNETIC SPEAKER.

dynamic stability A measure of the ability of a robot to maintain its balance while in motion.

dynamic stop As caused by a computer program instruction, a loop indicating the presence of an error.

dynamic storage See DYNAMIC MEMORY.

dynamic subroutine A form of computer subroutine that allows the derivation of other subroutines in various forms.

dynamic test See DYNAMIC CHECK.

dynamic transconductance Transconductance determined from alternating-current (ac) signal parameters, rather than from direct-current (dc) parameters.

dynamic transducer A coil-and-magnet device that converts mechanical vibration into electric currents, or vice versa. Common examples include most microphones, headphones, and loudspeakers.

dynamic transfer characteristic An input-output characteristic determined, with respect to the load of a transfer device. Also see DYNAMIC CHARACTERISTIC.

dynamic transistor tester **1.** An instrument for checking the alternating-current (ac) gain of a transistor, rather than its direct-current (dc) beta. **2.** An instrument for determining the condition of a transistor from its performance in a simple oscillator circuit. **3.** An instrument that displays a transistor response curve, or a family of such curves, on a calibrated oscilloscope screen. Also see DYNAMIC DIODE TESTER and DYNAMIC RECTIFIER TESTER.

dynamo A mechanical generator of electricity; typically a rotating machine.

dynamoelectric machinery Rotating electric machinery. Examples: amplidynes, generators, dynamotors, rotary converters.

dynamometer **1.** See ELECTRODYNAMOMETER. **2.** A device for mechanically measuring the output power of a motor.

dynamometer ammeter See ELECTRODYNAMOMETER.

dynamometer voltmeter See ELECTRODYNAMOMETER.

dynamophone A dynamometer (see DYNAMOMETER, **2**) that uses two telephone circuits to measure the twist of a shaft.

dynamostatic machine A machine driven by alternating-current (ac) or direct-current (dc) power for the generation of static electricity.

dynamotor A (usually small) self-contained motor-generator. The motor and generator portions are enclosed in a common housing, giving the machine the appearance of a simple motor.

dynaquad A pnpn four-layer semiconductor device with three terminals, similar to the silicon-controlled rectifier or thyristor.

dynatron A form of vacuum tube that displays a negative-resistance characteristic, resulting in oscillation at ultra-high and microwave frequencies.

dynatron frequency meter A heterodyne-type frequency meter using a dynatron oscillator.

dyne Abbreviation, d. A unit of force. One dyne (10^{-5} newton) is the force that will give a mass of 1 gram an acceleration of 1 centimeter per second per second. Compare NEWTON.

dyne-centimeter See ERG.

dyne-five In the Giorgi mks system, a unit of force equal to 1 newton.

dyne per square centimeter Abbreviation, d/cm^2. A unit of pressure equal to 0.1 pascal (9.869 × 10^{-7} atmosphere).

dyne-seven A unit of force equal to 10^7 dynes.

dynistor A semiconductor diode that continues to conduct after the forward voltage is reduced below the normal threshold point. To stop the conduction, a reverse voltage must be applied, or voltage must be entirely removed from the device. It is used in switching applications.

dynode In a photomultiplier tube, any of several slanting electrodes that receives a beam of electrons generated by the light-sensitive cathode and reflects it, along with secondary electrons. This amplifies the beam; the process is repeated several times. Thus, the emission from the cathode is greatly amplified when it reaches the plate.

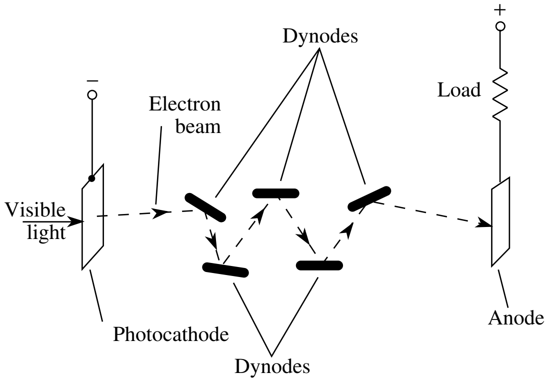

dynode

dysprosium Symbol, Dy. An element of the rare-earth group. Atomic number, 66. Atomic weight, 162.50. Dysprosium is a highly magnetic substance.

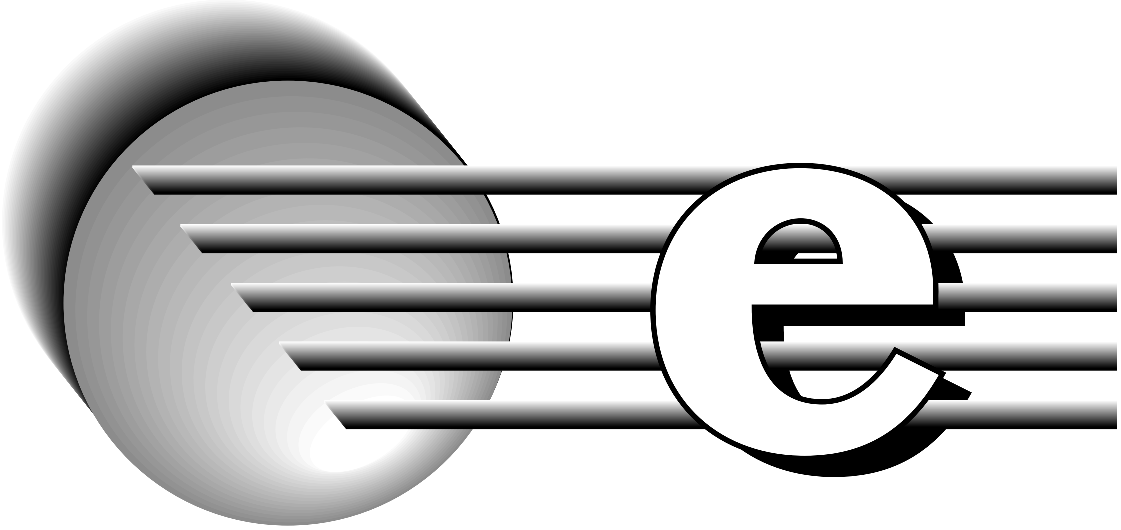

E **1.** Symbol for VOLTAGE. **2.** Symbol for ELECTRIC FIELD STRENGTH. **3.** Abbreviation of EMITTER. **4.** Symbol for prefix EXA. **5.** Symbol for ENERGY.

e **1.** Symbol for VOLTAGE. **2.** Abbreviation of EMITTER. **3.** Symbol for ELECTRON CHARGE. **4.** Symbol for the natural logarithm base (approximately equal to 2.71828). **5.** Symbol for ECCENTRICITY. **6.** Abbreviation of ERG.

EAM Abbreviation of *electronic accounting machine.*

E and M terminals The output and input leads in some signaling systems. Also called *E and M leads.*

early early sound Sound propagated through solids and/or liquids that reaches a pickup device (such as a microphone) before the sound propagated through the air. In general, sound waves travel faster as the medium becomes more dense.

early-failure period The period immediately after manufacture of a device, during which the failure rate (caused by defects in equipment or workmanship) is high.

early-warning radar Abbreviation, EWR. A radar system that produces immediate warning when enemy aircraft enter the monitored area.

earphone **1.** *Headphone* (usually a single unit). **2.** Telephone receiver. **3.** A miniature acoustic transducer that is small enough to be inserted into the ear.

earpiece See EARPHONE, **3**.

earth **1.** The ground. **2.** An electrical connection to the earth (see GROUND CONNECTION, **2**). **3.** In space communications, the planet Earth.

earth connection See GROUND CONNECTION, **2**.

earth currents **1.** Electric currents induced in the earth by current flowing through underground or underwater cables. **2.** Electric currents flowing through the earth between ground connections of electrical equipment.

earth ground **1.** A common connection to an electrode buried in the earth so that good conductivity is maintained between the common circuit point and the earth itself. **2.** A rod driven into the surface of the earth for use as a common circuit connection.

earth inductor A magnetometer consisting of a coil that is rotated in the earth's magnetic field. It delivers an alternating-current (ac) voltage proportional to the field strength. Also called *generating magnetometer.*

earth-moon-earth See MOONBOUNCE.

earth resonance A resonant effect at extremely low frequencies, caused by reflection of currents within the earth. Resonant currents have been tested as a means of communicating with submarines worldwide.

earth's magnetic field Also called *geomagnetic field.* The natural magnetic field whose lines of flux extend from north to south. The earth's magnetic poles, also called the *geomagnetic poles*, do not exactly coincide with the geographic poles. The field somewhat resembles that of a bar magnet.

Eastern Standard Time Abbreviation, EST. Local mean time at the 75th meridian west of Greenwich.

east-west effect The phenomenon in which the number of cosmic rays approaching earth near the equator from the west is greater than that from the east by 10 percent.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

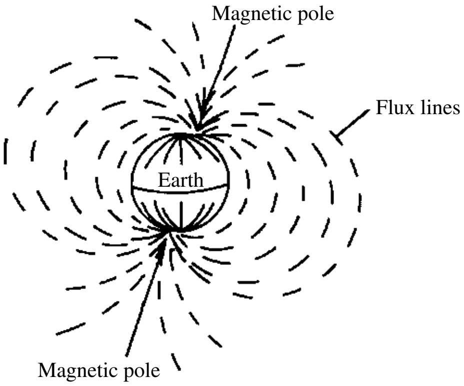

earth's magnetic field

E_{avg} Symbol for AVERAGE VOLTAGE.

E_b Symbol for BATTERY VOLTAGE.

E bend In a waveguide, a smooth change in the direction of the axis, which remains parallel to the direction of polarization.

EBI Abbreviation of EQUIVALENT BACKGROUND INPUT.

ebiconductivity Conductivity resulting from electron bombardment.

ebonite Hard rubber used as an insulant. Dielectric constant: 2.8. Dielectric strength: 30 to 110 kV/mm.

EBR Abbreviation of ELECTRON-BEAM RECORDING.

EBS Abbreviation of ELECTRON-BOMBARDED SEMICONDUCTOR.

EBS amplifier An amplifying device using an electron-bombarded semiconductor. The electron beam is modulated by the input signal, and the modulated resistance of the semiconductor target modulates a relatively heavy current to provide an amplified output. Current gains on the order of 2000 are possible.

ec Abbreviation of ENAMEL-COVERED (in reference to wire).

eccentric circle See ECCENTRIC GROOVE.

eccentric groove On a phonograph record, an off-center groove in which the stylus rides at the end of the recording, where it causes the tone arm to trip the record-changing mechanism.

eccentricity **1.** The condition of being off center, intentionally or not. It is often a consideration in the behavior of dials, potentiometers, and servomechanisms. **2.** On a phonograph record, the condition in which the spiral recording groove and the center hole of the disk are not concentric. **3.** A quantitative expression for the extent to which an ellipse is elongated.

eccentricity of orbit The extent to which the orbit or path of a satellite differs from a circle. A circular orbit has eccentricity zero. As the orbit becomes elliptical and deviates more and more from a perfect circle, the eccentricity increases. When the eccentricity reaches 1, the object takes a parabolic path through space. If the eccentricity exceeds 1, the path is a hyperbola.

ECCM Abbreviation of ELECTRONIC COUNTER-COUNTERMEASURES.

ECDC Abbreviation of *electrochemical diffused collector.*

ECG Abbreviation of electrocardiogram. (Also, EKG.)

ECG telemetry Use of a radio telemetering system to monitor the heart function of a person from a distance.

echelon **1.** A level of calibration accuracy, the highest echelon being the national standard for the particular measurement involved. **2.** A level of maintenance in which lower ordinal numbers refer to less-critical tasks, and higher ordinal numbers refer to tasks requiring progressively higher skills and technological expertise.

echelon grating A diffraction grating with extremely high resolution. Generally useful only over a small range of wavelengths.

echo **1.** A signal that is reflected back to the point of origin. **2.** A reflected or delayed signal component that arrives at a given point behind the main component. **3.** A radar blip, indicating an object or thundershower. **4.** Reflection of the signal on a telephone line, caused by improper impedance matching, or by overload of the system by too many subscribers attempting to use the system at the same time. **5.** In audio systems, a circuit that causes sounds to repeat one or more times, at intervals ranging from a fraction of a second to several seconds. **6.** The effect produced by a circuit, as defined in 5.

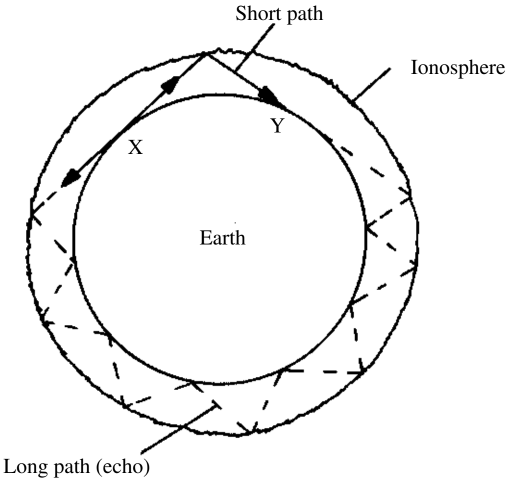

echo

echo area The area of a target that will return a radar signal as an echo.

echo attenuation In a bidirectional wire-communication circuit equipped with repeaters or multiplexers, the attenuation of echo currents set up by conventional operation.

echo box A resonant-cavity device used to test a radar set. Part of the transmitted energy enters the box, which retransmits it to the receiver. The signal reaching the receiver is a slowly decaying transient whose intensity eventually falls below the level that can be displayed on the screen; the time required to reach this level indicates radar performance.

echo chamber A reverberation chamber, electronic recording device, or room for acoustic tests or for simulating sonic delays.

echo check In data communication, a means of checking the accuracy of received data by sending it back to the transmitting station for comparison with the original data.

echo depth sounder See ACOUSTIC DEPTH SOUNDER.

echo eliminator **1.** A device that quiets a navigational instrument after receipt of a pulse, to prevent reception of a subsequent, delayed pulse. **2.** In a two-way telephone circuit, a voice-operated device that suppresses echo currents caused by conversation currents going in the opposite direction.

echoencephalograph An ultrasonic medical instrument that allows viewing of internal organs. Used for diagnostic purposes in certain situations, instead of the X-ray machine.

echogram In acoustics, a graph of the sound decrement in an enclosure. Time is plotted on the horizontal axis; signal amplitude is plotted on the vertical axis. An intense pulse is transmitted from a speaker; a microphone picks up the echoes and sends them to a pen recorder or microprocessor.

echo intensifier A device used at a radar target to boost the intensity of reflected energy.

echo interference Radio interference resulting from a reflected signal arriving slightly later than the direct signal.

echo matching In an echo-splitting radar system, the trial-and-error orientation of the antenna to find the direction from which the pulse indications are identical.

echo ranging An ultrasonic method of determining the bearing and distance of an underwater object.

echo send In an audio mixer, an output for delivering signals to external systems, such as an echo box (see ECHO, **5**). It can also provide an auxiliary output for a second set of speakers, a tape recorder, etc.

echo sounder See ACOUSTIC DEPTH SOUNDER.

echo splitting Separating a radar echo into two parts so that a double indication appears on the radar screen.

echo suppression In a telephone circuit, a device that chokes off reflected waves, thereby minimizing audible echo.

echo suppressor See ECHO ELIMINATOR.

echo talk Echo in a telephone system that results in distracting interference.

echo wave A reflected wave, such as a radio wave reflected alternately between earth's surface and the ionosphere.

ECL Abbreviation of EMITTER-COUPLED LOGIC.

eclipse effect A decrease in the critical frequency of the E and F1 layers of the ionosphere during a solar eclipse.

ecliptic orbit Any orbit that lies in the same plane as the orbit of the earth around the sun (the *ecliptic plane*). The ecliptic plane is slanted about 23.5° from the plane of the earth's equator.

ECM Abbreviation of ELECTRONIC COUNTERMEASURES.

ECO Abbreviation of ELECTRON-COUPLED OSCILLATOR.

econometer An instrument for continuously monitoring the amount of carbon dioxide in (factory) flue gases.

E core A transformer or transducer core having the shape of an E. Coils can be wound on one, two, or all three of the crosspieces.

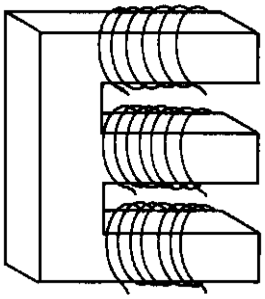

E core

ECPD Abbreviation of *Engineer's Council for Professional Development.*

ECTL Abbreviation of *emitter-coupled transistor logic.*

EDD Abbreviation of ENVELOPE-DELAY DISTORTION.

eddy current A circulating current induced in a conducting material by a varying magnetic field, often parasitic in nature. Such a current can, for example, flow in the iron core of a transformer.

eddy-current device A brake, coupling, clutch, drag cup, drive unit, or similar device whose operation is based on the generation of torque, pull, or opposition by the action of eddy currents.

eddy-current heating Heating caused by eddy-current loss in a material.

eddy-current loss Power loss resulting from eddy currents induced in nearby structures by an electromagnetic field. Eddy currents in the core of a transformer give rise to such loss.

edge connector A terminal block with a number of contacts, attached to the edge of a printed-circuit board for easy plugging into a foundation circuit.

edge control In the manufacture of paper, a robotic system for maintaining the width of a sheet by sensing the edges and correcting the machine accordingly. Transducers that sense the passing edges deliver output signals proportional to variations from standard width.

edge detection The ability of a machine vision system to locate and follow boundaries. Used extensively in mobile robots.

edge effect The extension of electric lines of flux between the outer edges of capacitor plates. This portion of the interplate field contributes a small amount of capacitance. Because the lines of flux are not confined to the space between plates, they can cause capacitive coupling with external bodies.

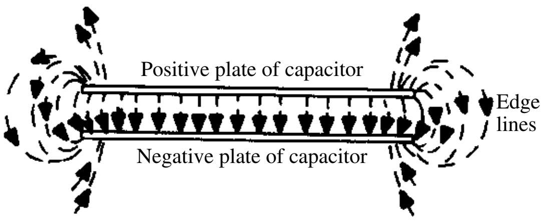

edge effect

edge-punched card In digital computer operations, a punched card whose edge is perforated in a narrow column, the center being used for written annotation.

edgewise meter A meter having a curved horizontal scale; this arrangement allows mounting the instrument edgewise in a panel.

edgewise-wound coil A coil made of a flat metal strip cut in the shape of a coil spring. The design allows the use of clips to vary the inductance, but this advantage is often offset by the coil's high distributed capacitance.

edging In a color television picture, extraneous color of a different hue than the objects around whose edges it appears.

Edison base A threaded base on light bulbs, cone-type heaters, and some pilot lamps.

Edison battery A group of Edison cells connected in series, parallel, or both, and contained in a single package with two electrodes.

Edison cell A secondary (storage) cell in which the active positive plate material consists of nickel hydroxide held in steel tubes assembled into a steel grid; the active negative plate material is powdered iron oxide mixed with cadmium; the electrolyte is potassium hydroxide. The open-circuit voltage of the cell is typically 1.2 V at full charge.

Edison distribution system A three-wire, 110- to 250-volt direct-current (dc) power-distribution system.

Edison effect Thermionic emission of negatively charged particles (electrons) from a hot filament sealed in an evacuated bulb; they are attracted by a cold, positively charged metal plate in the bulb.

E display A radar display in which the horizontal coordinate indicates distance, and the vertical coordinate indicates elevation.

edit **1.** In tape recording, the modifying of the recorded material by deleting (cutting out or erasing), adding (splicing or overrecording), or changing the sequence of the material by physically or magnetically altering the tape. **2.** In digital computer operations, to make data ready for processing.

edit decision list In the editing of a digital audio/video presentation, a record of every change (cut and paste). The list is automatically made by the computer and stored on disk for later reference if needed.

editing **1.** Alteration of a magnetic-tape recording by means of splicing. **2.** Alteration of data stored in memory, either by adding information, removing information, changing information, or (usually) a combination of these operations.

EDL Abbreviation of EDIT DECISION LIST.

EDM Abbreviation of ELECTRICAL-DISCHARGE MACHINING.

EDP Abbreviation of *electronic data processing.*

EDPC Abbreviation of ELECTRONIC DATA-PROCESSING CENTER.

EDPM Abbreviation of ELECTRONIC DATA-PROCESSING MACHINE.

EDPS Abbreviation of ELECTRONIC DATA-PROCESSING SYSTEM.

EDT **1.** Abbreviation of *ethylene diamine tartrate* (a synthetic piezoelectric material). **2.** Abbreviation of *Eastern Daylight Time.*

EDU Abbreviation of *electronic display unit.*

educational robot A robot that can be programmed for the purpose of teaching its users something. Popular as an educational toy for children.

EDVAC Acronym for *Electronic Discrete Variable Automatic Computer*, a development of the University of Pennsylvania.

EE Abbreviation of ELECTRICAL ENGINEER or ELECTRONICS ENGINEER.

E_{eff} Symbol for EFFECTIVE VOLTAGE.

EEG Abbreviation of ELECTROENCEPHALOGRAM.

EEPROM Abbreviation of ELECTRICALLY ERASABLE PROGRAMMABLE READ-ONLY MEMORY. See PROM and ROM.

EES Abbreviation of EARLY EARLY SOUND.

effective acoustic center The apparent point of propagation of spherically divergent sound waves radiated by an acoustic generator.

effective actuation time The total actuation time of a relay (i.e., the sum of the initial actuation time and subsequent intervals of contact chatter).

effective address The address a computer uses in implementing an instruction (i.e., one not necessarily coinciding with the address given in the instruction).

effective ampere An effective current of 1 ampere. Also see EFFECTIVE CURRENT.

effective antenna length See ELECTRICAL LENGTH.

effective antenna resistance The radiation resistance of an antenna, as measured at the input point.

effective bandwidth The bandwidth of an ideal bandpass filter, which, at a reference frequency, has the same transfer ratio as an actual bandpass filter under consideration; it also has the same current and voltage characteristics.

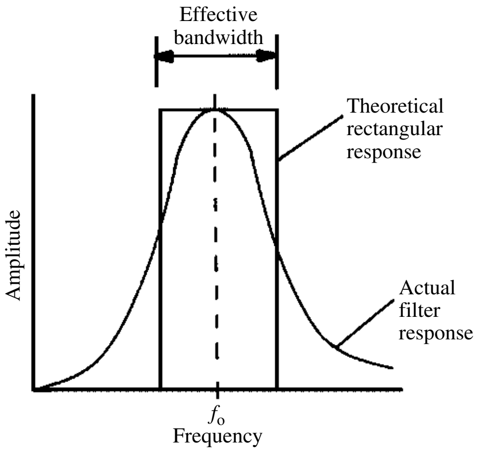

effective bandwidth

effective capacitance The actual capacitance between two points in a circuit resulting from the combination of inherent, lumped, and stray capacitances.

effective conductivity Conductivity measured between the parallel faces of a unit cube of a material.

effective confusion area In a radar system, an area in which interference makes it impossible to see whether a target is present.

effective current Symbol, I_{eff}. The root-mean-square (rms) value of alternating current (see EFFECTIVE VALUE). For a sinusoidal current, $I_{eff} = 0.707 I_{max}$, where I_{max} is the maximum value of the current. Also called *rms current.*

effective cutoff See EFFECTIVE CUTOFF FREQUENCY.

effective cutoff frequency For a filter or similar device operated between specified impedances, the frequency at which insertion loss is higher than the loss at a specified reference frequency in the passband.

effective field intensity The root-mean-square (rms) value of the field-strength voltage, averaged for all points at a horizontal distance of one mile from a transmitting antenna.

effective height The height of an antenna in terms of its performance as a transmitter or receiver of electromagnetic energy.

effective input capacitance The actual operative capacitance present at the input terminals of a circuit or device, caused by the shunt capacitance of the terminals themselves and the net capacitance of the circuit connected to the terminals.

effective internal resistance In an electrochemical cell or battery, a resistance that originates within the electrolyte and electrodes. This resistance is low when the current drain is low; it rises as the current drain increases. It limits the maximum current that the cell or battery can deliver.

effective isolation The condition of components or circuits being so well isolated or shielded that no significant direct coupling, capacitive coupling, or inductive coupling exists between them.

effective instruction The machine-language version of an instruction given in a computer program, as produced by resident software.

effectively bonded The condition afforded by an extremely low-resistance union between two conducting surfaces that are solidly fastened together.

effectively grounded The condition of being connected to earth or to the low-potential end of a circuit by means of an extremely low-resistance connection.

effective parallel capacitance Inherent capacitance that manifests itself in parallel with two circuit points in combination with any lumped capacitance.

effective parallel resistance **1.** The leakage resistance that manifests itself in parallel with a dielectric (e.g., the leakage resistance of a capacitor). **2.** Parallel-resistance effects caused by stray shunt-resistance components.

effective percentage of modulation For a complex waveform, an expression of the equivalent percentage of modulation by a pure sine wave. Given a certain proportion of power in the sidebands with modulation by a complex signal, the effective percentage of modulation is that percentage which, when the modulating signal is sinusoidal, results in the same proportion of power in the sidebands.

effective phase angle In alternating-current (ac) circuits, the phase angle, with respect to waveforms for current and voltage. When both waveforms are sinusoidal, the effective phase angle is the actual phase angle. But when harmonics are present in current or voltage, the angles differ, the difference being greater in capacitive circuits than in inductive circuits.

effective radiated power Abbreviation, ERP or P_{ER}. **1.** A measure of the performance of a wireless transmitting antenna. Suppose a test antenna, A_T, is set up and the field strength in its favored direction at a frequency f is measured at a distance d in free space. Let the field strength thus measured be F watts per square meter. Suppose A_T is replaced with an isotropic radiator, A_I, and the field strength at the same frequency f is measured at the same distance d in free space. Let the radio-frequency (RF) power at the feed point of the isotropic radiator A_I be varied until the field strength is F, the same as it was with the test antenna A_T. Let this RF power be symbolized P. Then P is defined as the effective radiated power (ERP or P_{ER}) of the test antenna, A_T. **2.** The figure defined as in (1), measured in some direction other than the favored direction of a test antenna.

effective resistance **1.** In a coupled circuit, the sum of the actual resistance of the circuit and the reflected resistance of the load. **2.** See EFFECTIVE ANTENNA RESISTANCE.

effective series inductance Inherent (distributed) inductance acting in series with other components in a circuit. The inherent inductance of the wire in a wirewound resistor, for example, manifests itself in series with the resistance of the device.

effective series resistance Inherent (distributed) resistance acting in series with other components in a circuit. Thus, the inherent resistance of the wire in a coil appears in series with the inductance of the coil. Likewise, a capacitor has an effective series resistance because of the resistance of the leads, plates, and connections.

effective shunt capacitance See EFFECTIVE PARALLEL CAPACITANCE.

effective shunt resistance See EFFECTIVE PARALLEL RESISTANCE.

effective sound pressure The root-mean-square (rms) value of instantaneous sound pressure at one point in a sound cycle.

effective speed of transmission In telegraphy (wire or radio) and in electronic data transmission, the transmission speed (characters per minute, bits per second, etc.) that can be reliably maintained for a given period.

effective thermal resistance The effective temperature rise (in degrees per watt of dissipation) of a semiconductor junction above an external reference temperature that is at equilibrium.

effective time For a computer, the time during which useful work is performed.

effective transmission speed See EFFECTIVE SPEED OF TRANSMISSION.

effective value The root-mean-square (rms) value of an alternating-current (ac) quantity. The effective value an alternating current produces in a pure resistance has the same heating effect as the equivalent direct current. See also ROOT MEAN SQUARE.

effective volt An effective potential of one root-mean-square (rms) volt. Also see EFFECTIVE VOLTAGE.

effective voltage Symbol, E_{eff}. The root-mean-square value of alternating-current (ac) voltage (see EFFECTIVE VALUE). For a sinusoidal voltage, $E_{eff} = 0.707E_{max}$, where E_{max} is the maximum value of the voltage. Also called *rms voltage.*

effective wavelength Wavelength in terms of measured frequency and effective propagation velocity.

effects processor In audio systems, a circuit that produces various sound effects via digital signal processing.

efficiency **1.** The ratio of useful power or energy output to total power or energy input to a device or system. **2.** The proportion of applied audio-frequency (AF) power that a loudspeaker converts into acoustic energy. **3.** See ELECTRICAL EFFICIENCY.

efficiency modulation A system of amplitude modulation in which the efficiency of a radio-frequency (RF) power amplifier is varied at an audio-frequency (AF) rate.

efficiency of rectification For a rectifier, the ratio of the direct-current (dc) output voltage to the peak value of alternating-current (ac) input voltage. For percent efficiency, the ratio is multiplied by 100.

efflorescence The giving up of water by a substance upon exposure to air. Some materials used in electronics exhibit this property. Common efflorescent compounds are hydrated ferrous carbonate, ferrous sulfate, and sodium carbonate.

efflorescent material A material exhibiting efflorescence. Compare DELIQUESCENT MATERIAL.

E field **1.** An electric field. **2.** the electric-field component of an electromagnetic wave.

EFL Abbreviation of *emitter-follower logic.*

E_g Symbol for GENERATOR VOLTAGE.

EHD Abbreviation of *electrohydrodynamic(s).*

EHF Abbreviation of EXTREMELY HIGH FREQUENCY.

EHP Abbreviation of *effective horsepower.*

E-H tee A waveguide junction in which E- and H-plane tee junctions intersect the main waveguide at the same point. Also see WAVEGUIDE TEE.

E-H tuner An impedance-transforming E-H tee with two arms that are terminated in tunable plungers for critical adjustments. See WAVEGUIDE PLUNGER.

EHV Abbreviation of *extra-high voltage.*

E_i Symbol for INPUT VOLTAGE.

EIA Abbreviation of *Electronics Industries Association.*

eight-level code A code, such as the *American Standard Code for Information Interchange (ASCII),* in which each character is represented by eight bits.

E indicator A radar elevation display, in which the horizontal scale shows range and the vertical scale shows elevation.

Einstein equation The equation depicting the interconversion of mass and energy; $E = mc^2$, where E is energy (ergs), m is mass (grams), and c is the speed of light in a vacuum (centimeters per second).

einsteinium Symbol, Es. A radioactive element produced artificially. Atomic number, 99. Atomic weight, 252.

Einstein/Dehaas effect The tendency for an iron or steel cylinder to become magnetized as it rotates.

Einstein shift The decrease in frequency and loss of energy experienced by quanta acted upon by gravitation.

Einthoven string galvanometer A simple galvanometer in which a silvered glass filament carrying current is mounted in a magnetic field set up by either a permanent magnet or an electromagnet. The current causes the filament to be deflected through a distance proportional to current strength, the deflection being observed through a microscope.

EIT Abbreviation of *engineer-in-training.*

either-or operation The logical inclusive-OR operation.

EJC Abbreviation of *Engineers' Joint Council.*

eka-aluminum See GALLIUM.

eka-silicon See GERMANIUM.

EKG Abbreviation of ELECTROCARDIOGRAM. (Also, ECG.)

EKG telemetry See ECG TELEMETRY.

EL Abbreviation of ELECTROLUMINESCENCE.

elapsed time **1.** In data-processing and computer operations, what appears to be the duration of a process, compared with actual processing time, as measured by internal clocks, for example. **2.** The accumulated time, usually expressed in hours, minutes, and seconds, that an operation takes or a machine runs.

elapsed-time meter An instrument that indicates the time an electronic device or system has been in operation. Most such meters are based on electric clockwork that runs only while the system is in operation, holding the count during shutdown periods. Also see ELECTROLYTIC ELAPSED-TIME METER.

elastance Unit, daraf. The opposition of a capacitor to being charged. It is the reciprocal of capacitance.

elastic collision A collision between two charged particles in which neither loses energy—even though they are deflected from their normal paths.

elasticity **1.** The ability of a body to return to its original shape after being deformed. See YOUNG'S MODULUS. **2.** See ELASTIVITY.

elastic limit The maximum stress that can be tolerated by a material without being permanently deformed.

elastic wave A wave in an elastic medium, such as air or water; thus, a wave that is mechanically produced.

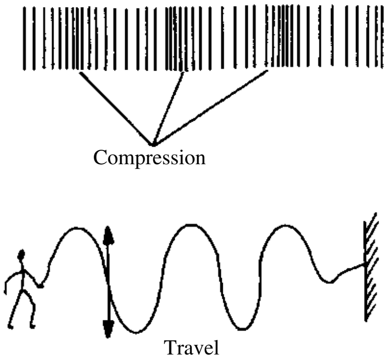

elastic waves

elastivity **1.** Specific elastance (i.e., the elastance in darafs per cubic unit of a dielectric). **2.** The ratio of electric stress to displacement.

elastomer A compressible, conducting substance used in pressure sensing. In one arrangement, an array of electrodes is connected to each side of a pad of the material. When pressure appears at some point, the material compresses, lowering the resistance in the pressure zone. A microprocessor determines the location and extent of the pressure.

E layer A layer in the ionosphere that is 50 to 90 miles above the surface of the earth, and is noted for returning radio waves at high and very-high frequencies.

elbow bend A 90-degree bend in a waveguide.

ELD Abbreviation of *edge-lighted display.*

electra A radio-navigational system in which equal-intensity signal zones (usually 24) are provided.

electralloy A nonmagnetic alloy used in the manufacture of radio hardware, such as chassis.

electre See ELECTRUM.

electrepeter A device used to change the direction of an electric current.

electret A device whose heart is a dielectric disk or slab that is permanently polarized electrically, and so possesses a permanent electric field. The electrical equivalent of the permanent magnet. Certain waxes, ceramics, and plastics acquire permanent polarization after they have been heated, then cooled slowly in an intense electric field.

electret microphone Also called *electret capacitor microphone.* A microphone in which sound waves cause an electret to vibrate and generate audio-frequency (AF) output current. An internal power supply (such as a battery) supplies the necessary voltage.

electric **1.** Pertaining to electricity and its various manifestations. **2.** See ELECTROSTATIC.

electric absorption See DIELECTRIC ABSORPTION.

electric accounting machine A self-contained data-processing machine that is neither a computer nor a computer peripheral.

electrical-acoustical transducer A transducer, such as a headphone, sonic applicator, or buzzer, that converts electrical energy into sound energy. Compare ACOUSTICAL-ELECTRICAL TRANSDUCER.

electrical angle The angle assumed at any instant by the rotating vector representing an alternating current or voltage. A complete cycle is divided into 360 electrical degrees. Thus, for an alternating-current (ac) sine wave, the angle is 0 degrees for zero and positive-going, 90 degrees for positive maximum, 180 degrees for zero and negative-going, and 270 degrees for negative maximum.

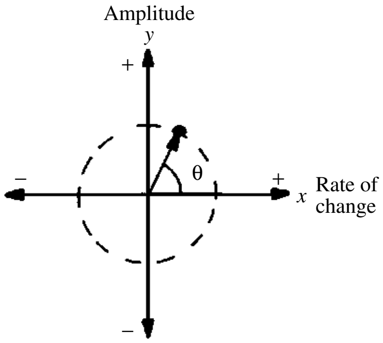

electrical angle

electrical attraction The attraction between two oppositely charged bodies or particles. Compare ELECTRICAL REPULSION.

electrical axis In a quartz crystal, the axis through opposite corners of the hexagonal cross section. The various electrical axes are *x*, *x*’, and *x*” (or *x*-*x*, *x*’-*x*’, and *x*”-*x*”). The electrical axis is perpendicular to the mechanical axis, which runs through the crystal’s length. Also see CRYSTAL AXES and X-AXIS, **2**.

electrical bail An action in which a special switch changes contact position and locks itself in that position after a station has been actuated, at the same time releasing a previously actuated station.

electrical bandspread In a tuned circuit, bandspread obtained by changing values of inductance or capacitance, rather than by mechanical gearing.

electrical bias A current maintained in a relay coil (sometimes an auxiliary coil) to keep the relay partially closed, thus sensitizing it. Compare MECHANICAL BIAS.

electrical boresight In radar operations, the tracking axis, as determined by an electrical test, such as one involving a sharp null or sharp peak response.

electrical center The point at which an adjustable component (variable resistor, variable inductor, etc.) has exactly half its total value. This point does not always coincide with the physical center.

electrical conductance See CONDUCTANCE.

electrical conduction The flow of charge carriers through a material. The degree of conduction is indicated by the material’s value of conductance.

electrical conductivity See CONDUCTIVITY.

electrical coupling The coupling of two or more circuits or elements by means of electric-field effects.

electrical degree **1.** In a periodic waveform, the length of time corresponding to $1/360$ of the time for completion of one cycle. **2.** In space, that distance representing $1/360$ of the wavelength in the medium through which electromagnetic energy travels.

electrical discharge The flow of current out of a voltage reservoir, such as a battery or capacitor.

electrical discharge in gases The phenomenon of electric conduction (current) by a gas, caused by sudden breakdown as a result of gas ionization. The discharge is often accompanied by light, as in the red glow of a neon bulb.

electrical-discharge machining A method of machining metals in which the metal is vaporized by an arc formed between an electrode and the metal workpiece (anode). In this way, metal is removed in tiny bits from the surface of the workpiece.

electrical distance Distance in terms of the time required for an electromagnetic wave to travel between two points in a particular medium.

electrical drainage The diversion of electric currents away from underground pipes to prevent corrosion by electrolysis.

electrical efficiency The ratio of the output of an electrical or electronic device to the total input. It can be expressed as a decimal or percentage. For example, for a bipolar transistor amplifier, the percent efficiency is equal to $100(P_{\mathrm{out}}/P_{\mathrm{in}})$, where P_{in} is the collector input in volt-amperes, and P_{out} is the output power in watts.

electrical elasticity See CAPACITANCE.

electrical element See ELEMENT, **2**.

electrical energy Energy in the form of electricity (see ELECTRICITY, **1**). The term is often used in place of ELECTRICITY.

electrical engineer Abbreviation, EE. A trained professional skilled in applying physics and mathematics to electricity, and in the theory and application of basic engineering and related subjects. Of particular interest to the EE are the generation and distribution of electrical energy and the design and application of electromechanical devices. Compare ELECTRONICS ENGINEER.

electrical erosion In electrical contacts, the loss of metal as a result of the evaporation or transfer of metal during switching.

electrical filter A bandpass, band-rejection, high-pass, or low-pass filter that operates by electrical means. Examples: Butterworth filter and Chebyshev filter.

electrical forming See ELECTROFORM, **1**.

electrical gearing In an electromechanical system, such as a servo, the condition in which an output shaft is electrically rotated at a speed different from that of an input shaft.

electrical glass High-temperature insulating materials made from glass fibers.

electrical inertia See INDUCTANCE.

electrical initiation **1.** Starting an action (electrical or nonelectrical) by means of an electrical signal. **2.** Using an enabling pulse.

electrical instrument A device for measuring an electrical quantity (such as voltage, current, or power).

electrical interlock Also called an *interlock switch* or *door-interlock switch.* A door- or lid-operated switch connected in series with the power switch of a piece of equipment. The interlock causes power to be removed from internal circuits whenever the door is opened, the lid lifted, or the case removed. This minimizes the chance for electric shock to occur to service personnel.

electrical length The length, in wavelengths, of an antenna or transmission line. The electrical length usually differs from the actual length because of ground-capacitance effects, end effects, and the speed of electromagnetic waves in conductors and/or dielectrics.

electrical load A device connected to a source of electricity (generator, amplifier, network, etc.) for a useful purpose (heat, work, etc.).

electrically connected Connected via direct path, such as through a wire, resistance, inductance, or capacitance.

electrically erasable PROM A programmable read-only memory (PROM) that can be erased by an electrical signal, rather than by exposure to ultraviolet light. Also see PROM and ROM.

electrically variable capacitor See VOLTAGE-VARIABLE CAPACITOR.

electrically variable inductor An inductor whose value varies inversely with the amount of direct current that is caused to flow through it or through an auxiliary winding on the same core.

electrically variable resistor See VOLTAGE-DEPENDENT RESISTOR.

electrical nature of matter The general behavior of matter as a complex interplay of waves and particles. Also see ELECTRON THEORY OF MATTER, WAVE MECHANICS, and WAVE THEORY OF MATTER.

electrical network A circuit containing two or more components (including generators and loads), usually arranged in some pattern.

electrical noise Extraneous currents and/or voltages that interfere with desired electrical quantities. Compare ACOUSTIC NOISE.

electrical polarity The distinct difference observable in electrification, designated positive (or plus) and negative (or minus). Negative electrification is generally characterized by a surplus of electrons; positive electrification is characterized by a deficiency of electrons.

electrical quantity **1.** See COULOMB and QUANTITY, **3**. **2.** An electrical unit (e.g., AMPERE, OHM, VOLT, and WATT).

electrical repulsion The mutual repulsion of bodies or particles having similar electric charges. Two positively charged objects will repel each other, as will two negatively charged objects.

electrical reset An electromechanical device for resetting a relay that ordinarily remains in the position resulting from actuation.

electrical resistance The in-phase current-retarding effect that all conductors exhibit to some extent. Also see RESISTANCE.

electrical resistivity See RESISTIVITY.

electrical resolver A synchro whose rotor has two perpendicular windings in addition to another winding.

electrical scan A method of changing the orientation of the major lobe of an antenna. The antenna is kept physically stationary, but the phase/amplitude relationships of the signals applied to different driven elements are varied.

electrical sheet Sheet iron or steel used for motor laminations.

electrical system **1.** The overall configuration of electrical elements for a set of apparatus. **2.** The wiring system that supplies power to a set of devices. **3.** One of several methods of quantizing electrical properties, such as METER-KILOGRAM-SECOND (mks), CENTIMETER-GRAM-SECOND (cgs) electromagnetic, cgs electrostatic, and the INTERNATIONAL SYSTEM OF UNITS (SI).

electrical taste See GALVANIC TASTE.

electrical technology The theory and practical application of electricity. Taught as a subengineering major, usually in two-year colleges that award the degree of Associate in Arts (AA) or Associate in Science (AS).

electrical time constant For a torque motor, the ratio of armature inductance to effective armature resistance. Compare MECHANICAL TIME CONSTANT.

electrical transcription **1.** A phonograph record made electrically, as opposed to one made mechanically. **2.** A radio program in which such a record is played. **3.** Any direct mechanical or electrical recording of an audio signal.

electrical transducer **1.** A transducer that converts a nonelectrical phenomenon into a proportional current, voltage, or frequency. **2.** A transducer that converts electricity in one form to electricity in another [e.g., a transducer actuated by direct-current (dc) voltage, delivering an alternating-current (ac) output, whose frequency is proportional to the dc input voltage].

electrical twinning A defect in which two quartz crystals intergrow in such a way that the electrical sense of their axes becomes reversed. Compare OPTICAL TWINNING.

electrical unit A standard for measuring an electrical quantity (e.g., ampere, ohm, volt, watt, siemens, etc.).

electrical wavelength The distance between one point in an electromagnetic wave cycle and the next identical point. This is usually expressed as the separation between points where the instantaneous amplitude of the electric field is zero and increasing positively. This quantity depends on the velocity factor, ν, of the medium through, or along, which the field propagates. The electrical wavelength also depends on the frequency, f, of the energy. In a transmission line with a velocity factor ν (given as a fraction rather than as a percentage), the electrical wavelength λ is given by

$$\lambda = 3.00 \times 10^8 \nu / f$$

where s is in meters, ν is in meters per second, and f is in hertz. Often, this formula is modified for values of f expressed in megahertz (Mhz) rather than in hertz. The equation in this case becomes

$$\lambda = 300 \nu / f$$

The wavelength of a signal in a radio-frequency (RF) transmission line is always less than the wavelength of the same signal in free space.

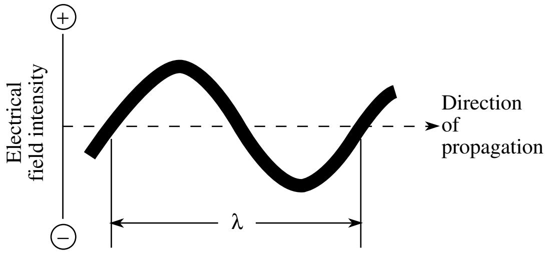

electrical wavelength

electrical zero **1.** A zero-output or minimum-output point resulting from the adjustment of a bridge or other zero-set circuit. **2.** In a meter whose pointer is mechanically set to some point above or below the zero on the scale, the zero setting obtained when the meter is deflected to scale zero by a current or voltage. **3.** For a synchro, the position at which the amplitudes and time phase of the outputs are defined.

electric and magnetic double refraction See KERR ELECTRO-OPTICAL EFFECT and KERR MAGNETO-OPTICAL EFFECT.

electric arc A sustained luminous discharge in the space between two electrodes. Compare ELECTRIC SPARK.

electric aura See ELECTRIC WIND.

electric balance See BRIDGE, **1** and **2**.

electric bell See BELL.

electric brazing A method of brazing in which electric current generates the required heat.

electric breakdown **1.** The usually sudden ionization of a gas by an electric field and the accompanying heavy current flow through the gas. **2.** The (destructive) puncture of a dielectric by the strain produced by high voltage. Also see DIELECTRIC STRENGTH. **3.** The usually nondestructive, abrupt increase in semiconductor junction current at a high reverse voltage. See, for example, AVALANCHE BREAKDOWN.

electric breakdown voltage **1.** The voltage at which avalanche effect occurs. **2.** Dielectric strength.

electric breeze See ELECTRIC WIND.

electric buzzer See BUZZER. Compare ELECTRONIC BUZZER.

electric calculator An electrically driven machine for performing mathematical operations. Its electromechanical nature distinguishes it from the electronic calculator, which features no moving parts, other than keys. Also see CALCULATOR.

electric catfish A fish native to tropical and northern Africa, capable of delivering a strong electric shock.

electric cell See CELL, **1**.

electric chair An electrode-bearing chair used in some states for administering the death penalty via high-voltage electricity. See ELECTROCUTION.

electric charge Potential energy as the electrification of a body or component. For a capacitance of C farads charged to a potential of E volts, the charge Q, in coulombs, is equal to the product CE. Also see ENERGY STORED IN CAPACITOR.

electric chronograph An instrument for accurately recording time intervals.

electric chronometer A precision electric or electronic timepiece. Also see ELECTRIC CLOCK and ELECTRONIC CLOCK.

electric circuit A network of interconnected components and devices, often including a source of electric power. Current flowing through a circuit is acted on by components, which produce specific desired effects.

electric clock A clock driven by electric current. Electric clocks fall into two categories: those driven by synchronous alternating-current (ac) motors, and those driven by stepping mechanisms that usually operate from direct current (dc).

electric column See VOLTAIC PILE.

electric conduction The flow of current carriers through a conductor.

electric constant Also called *permitivity of vacuum.* The fixed electrical permitivity of free space, the value of which is 8.8542×10^{-12} farad per meter.

electric contact See CONTACT, **1**, **2**.

electric controller An adjustable device for modifying the operating voltage or power of a component or system. Compare ELECTRONIC CONTROLLER.

electric cooling **1.** Cooling via PELTIER EFFECT. **2.** See ELECTROSTATIC COOLING. **3.** Forced-air cooling (of equipment) by electric blowers or fans.

electric current The phenomenon wherein charge carriers move in a directed manner through a material or vacuum. In most electrical conductors, current results from movement of electrons. In a semiconductor material, electric current can result from the movement of holes, as well as electrons; the proportion of holes to electrons depends on the nature of the semiconductor. In a gas or electrolyte, current consists of a flow of ions. In certain situations, electric currents can result from the movement of positrons, protons, anti-protons, alpha particles, and various atomic nuclei.

electric current density See CURRENT DENSITY.

electric delay line See DELAY LINE.

electric density See ELECTRIC SPACE DENSITY and ELECTRIC SURFACE DENSITY.

electric dipole A pair of equal charges having opposite polarity and separated by a fixed distance.

electric discharge See ELECTRICAL DISCHARGE.

electric-discharge lamp See DISCHARGE LAMP.

electric disintegration See ELECTRIC DISPERSION.

electric dispersion In a colloidal suspension, dispersion accomplished by passing an electric current through the material.

electric displacement The movement of a body or particle in response to an electric current or field.

electric double refraction See KERR ELECTRO-OPTICAL EFFECT.

electric dust precipitator See DUST PRECIPITATOR.

electric eel An eel (fish) capable of delivering a disabling shock on contact.

electric elasticity See ELASTIVITY, **1**, **2**.

electric endosmosis See ELECTRO-OSMOSIS.

electric eye A sensing device that uses a radiant energy beam to detect objects. It generally uses a laser diode as the beam source, and a photoelectric cell, phototransistor, or photovoltaic cell as the beam detector. The output is used to control some external machine or system.

electric fence A wire fence through which an electric current is passed. Anyone touching the fence will receive a shock. It is used in some prisons, and also by cattle ranchers to keep people or animals contained.

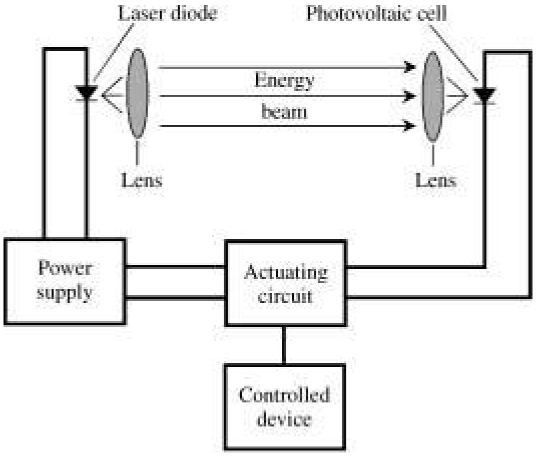

electric eye

electric fidelity The frequency response of a circuit or device.

electric field The space surrounding an electric charge or charged body, in which electric energy acts (electric lines of flux fill the space).

electric field intensity See ELECTROSTATIC FIELD INTENSITY.

electric field strength **1.** Symbol, *E.* In an electromagnetic wave, the amplitude of the electric component of the field, expressed in volts per meter. **2.** Dielectric strength.

electric-field vector See ELECTRIC-FIELD STRENGTH, **1**.

electric filter See ELECTRIC WAVE FILTER.

electric fish Fish capable of generating intense electric shocks (e.g., *electric catfish* and *electric eel*).

electric flux See ELECTROSTATIC FLUX.

electric flux density Symbol, *D.* In an electric field, the number of lines of flux per unit area, usually expressed in coulombs per square meter.

electric flux lines The direction of the electric field in the vicinity of a charged object. The field is denoted by means of "lines of flux" or "lines of force," with each line representing a designated electric field intensity. The closer together the lines, the more intense the field in a given region.

electric focusing See ELECTROSTATIC FOCUSING.

electric force The force exerted by an electrically charged particle or an electric field.

electric forces The forces exerted by electric charges or electric fields. Also see UNIT ELECTROSTATIC CHARGE.

electric furnace An electrically heated chamber (sometimes heated by an electric arc) used in ore reduction, carbide manufacture, and other high-temperature processes.

electric generator A device for producing electricity. Thus, many different devices, such as bat-

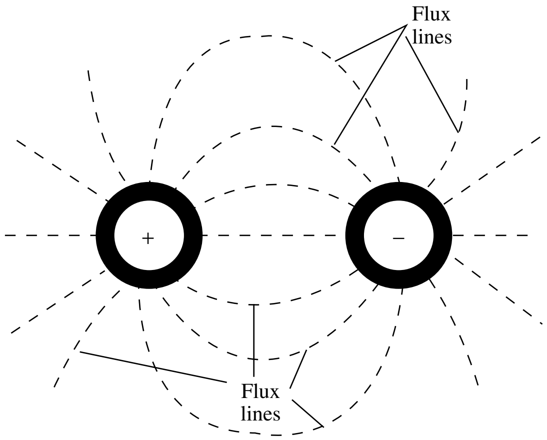

electric flux lines

teries, dynamos, oscillators, solar cells, and thermocouples, are classed as generators.

electric glow The light (usually pale blue) that occasionally accompanies an electric discharge in air.

electric guitar A guitar whose acoustic vibrations are converted by a transducer to an electrical signal for amplification.

electric hygrometer An instrument for measuring humidity in terms of the moisture-sensitive resistance of a sensor. A relatively simple direct-current (dc) circuit is used. Compare ELECTRONIC HYGROMETER.

electric hysteresis See ELECTROSTATIC HYSTERESIS.

electrician A professional person who installs and services electrical equipment and wiring.

electric image For solving certain problems involving electricity, an array of electrical points forming an image of certain other electrical points.

electricity **1.** Phenomena resulting from the existence of stationary or moving electric charge carriers, such as electrons, holes, and ions. **2.** A branch of physics concerned with phenomena resulting from the existence of stationary or moving charge carriers. **3.** Electrical energy or power. **4.** The voltage existing at utility outlets. In the North America, this is nominally an alternating-current (ac) voltage of 117 or 234 V at 60 Hz.

electric lamp An electric-powered device used primarily as a source of light. Common types are arc, incandescent, and fluorescent.

electric light **1.** Visible light produced by means of electricity. **2.** Electric lamp.

electric lines of flux Lines of flux associated with an electric charge and constituting the charge's electric field.

electric machine A mechanical device for generating static electricity. See, for example, ELECTROSTATIC GENERATOR, VAN DE GRAAFF GENERATOR, and WIMSHURST MACHINE.

electric meter **1.** An instrument such as an ammeter, voltmeter, or wattmeter, used to indicate an electrical quantity (usually directly). **2.** See KILOWATT-HOUR METER.

electric mirror See ELECTRON MIRROR.

electric moment In an electric field of unit intensity, the maximum torque exerted on an electric dipole.

electric motor A machine that converts electrical energy into mechanical work. The familiar form is a machine in which an armature rotates between the poles of a field magnet, mechanical energy being produced at the armature's revolving shaft.

electric needle A needle electrode carrying high-frequency current; it is used in surgery to cut tissue and sear it immediately to prevent bleeding.

electric network See ELECTRICAL NETWORK.

electric organ See ELECTRONIC ORGAN.

electric oscillations The alternate flow of electric charges in opposite directions, occurring at a defined frequency or frequencies.

electric osmosis See ELECTRO-OSMOSIS.

electric piano See ELECTRONIC PIANO.

electric polarization **1.** The orientation of flux lines in an electric field. **2.** The orientation of the electric field component in an electromagnetic field.

electric potential See ELECTROSTATIC POTENTIAL.

electric power Symbol, *P*. Unit, watt. The rate at which electrical energy is used. Power is energy per unit time; in the context of electricity, it is expressed as the product of current and voltage. In terms of heat losses, it is often expressed as I^2R (current in amperes squared, multiplied by resistance in ohms).

electric precipitator See DUST PRECIPITATOR.

electric probe A pin or rod inserted into an electrostatic field to sample it, or into an electromagnetic field to sample its electric component. See, for example, WAVEGUIDE PROBE. Compare MAGNETIC PROBE.

electric radiation **1.** The radiation of energy by means of electric waves. **2.** The energy so radiated.

electric recording Inkless recording on paper by direct use of an electric current. The two principal types are: (1) A current-carrying stylus burns away (in a fine line) the metallic coating of the recording paper, exposing the dark underlying layer. (2) A stylus delivers current that produces a line by means of electrolysis in a special paper (see ELECTROLYTIC RECORDER). Compare ELECTROSTATIC PRINTER.

electric reset See ELECTRICAL RESET.

electric residue A residual electric charge, such as might remain on a capacitor after it has been incompletely discharged.

electric rings Colored rings formed on a plate by the electrolytic deposition of substances, such as copper and some peroxides.

electric screen See ELECTROSTATIC SCREEN.

electric shield See ELECTROSTATIC SCREEN.

electric shock A physiological reaction caused by the passage of electric current through living tissue. When slight, it is characterized by tingling sensations and involuntary contractions of the muscles; a severe shock can cause paralysis, unconsciousness, heart fibrillation, and/or burns. If heart fibrillation occurs or if burns are severe, death can result.

electric spark A momentary, luminous discharge of electricity in the space between two electrodes. Compare ARC, **1**.

electric steel Steel that has been processed in an electric furnace.

electric strain See DIELECTRIC STRAIN.

electric strain gauge A device for detecting the strain that a certain stress produces in a body. Typically, such a gauge consists of one or more fine insulated wires cemented to the surface under test. As the surface becomes strained, the wire stretches, undergoing a change in electrical resistance that is proportional to the change in strain.

electric strength See DIELECTRIC STRENGTH.

electric stress See DIELECTRIC STRESS.

electric stroboscope See ELECTRONIC STROBOSCOPE.

electric surface density The ratio of the electric charge on a surface to the area of the surface. Compare ELECTRIC VOLUME DENSITY.

electric tachometer See ELECTRONIC TACHOMETER.

electric telemeter A device that sends metered data over a wire transmission line to a remote point for monitoring.

electric thermometer See ELECTRONIC THERMOMETER.

electric transcription See ELECTRICAL TRANSCRIPTION.

electric transducer A transducer that responds to non-electric energy, then delivers a proportional electric current or voltage to another circuit.

electric tuning A means of adjusting the frequency of a receiver, transmitter, transceiver, or oscillator, without the use of mechanical devices. An example is the use of a varactor diode, whose capacitance varies with voltage.

electric utilities The power companies that supply electric current for industrial and consumer use.

electric vane A small demonstration device consisting of a rotor having several spokes terminating in points. The rotor is mounted on a pivot bearing and, when it is connected to a source of high voltage, spins from the force of electricity escaping from the points into the surrounding air.

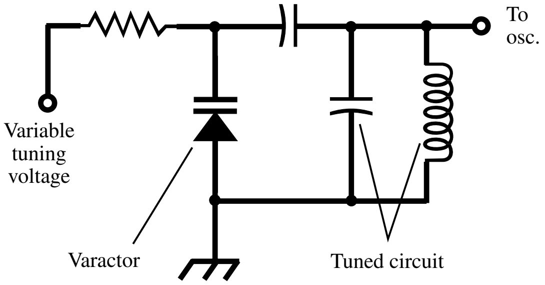

electric tuning

electric vector In an electromagnetic field, the vector representing the electric component. It is perpendicular to the magnetic vector.

electric volume density The ratio of the electric charge in a space to the volume of the space. Compare ELECTRIC SURFACE DENSITY.

electric watch A small timepiece driven by a tiny, self-contained electrochemical cell that drives an electrical escapement or other stepping mechanism. Compare ELECTRONIC WATCH.

electric wave See ELECTROMAGNETIC WAVES.

electric-wave filter A circuit or device for separating signals of one frequency from those of other frequencies.

electric whirl See ELECTRIC VANE.

electric wind **1.** Air currents set up by electrons escaping from the sharp point of a high-voltage electrode. **2.** The outward-rushing plasma (solar wind) ejected by the sun and traveling through space.

electrification **1.** Generating an electric charge in a body, as in charging a glass rod by rubbing it with a silk cloth. **2.** Providing electric service (e.g., RURAL ELECTRIFICATION). **3.** The conversion of a system from purely mechanical to electrical or electromechanical.

electroacoustic Pertaining to devices and systems that exhibit both electrical and acoustic effects. For example, speakers and microphones are electroacoustic devices.

electroacoustic device A device that transfers energy by converting it from electrical to acoustic form or vice-versa.

electroacoustic transducer A transducer that converts sound vibrations into electrical pulsations or, conversely, one that converts electricity into sound.

electroanalysis Chemical analysis performed by electrolytic methods.

electroanesthesia Anesthesia produced by an electric current going through some part of the body.

electroballistics A branch of physics concerned with the electrical or electronic measurement of the velocities and trajectories of projectiles.

electrobath A solution in which ELECTROPHORESIS or ELECTROPLATING is done.

electrobiology Biology concerned with electrical phenomena in living organisms.

electrobioscopy The examination of a body for viability by inducing muscular contractions with an electrical impulse.

electrocapillarity The production of capillary effects by means of electricity. See, for example, CAPILLARY ELECTROMETER.

electrocardiogram Abbreviation, ECG or EKG. A record made by an electrocardiograph of the changes in potential caused by the heartbeat; used as a diagnostic aid.

electrocardiograph An instrument that records changes in electrical potential caused by the heartbeat.

electrocardiophonograph A medical instrument that detects and records the impulses of the heart.

electrocatalysis Catalytic action produced by electricity (see CATALYSIS).

electrocautery **1.** In medicine, a cauterizing instrument consisting essentially of a platinum wire (at the tip of an insulated probe) that is heated by an electric current. Also see ELECTRIC NEEDLE. **2.** Cauterizing with an electrocautery.

electrochemical deterioration An electrochemical reaction that results in the permanent or temporary failure of a device.

electrochemical diffused-collector transistor A high-current pnp transistor in which metal has replaced the etched-away mass of p material, providing a built-in heatsink.

electrochemical equivalent In electrolysis or electroplating, a constant (*Z*) for the metal in plates. For a given metal, *Z* is the mass (in grams) of the metal deposited by 1 coulomb of electricity.

electrochemical junction transistor See SURFACE-BARRIER TRANSISTOR.

electrochemical measurements **1.** Measurements made on chemical substances with electrical instruments to determine such factors as conductivity, pH, dielectric strength, dielectric constant, etc. **2.** Measurements of electrical or electronic phenomena in terms of electrochemical response (e.g., current drain in terms of weight of plated metal, or voltage in terms of gas breakdown).

electrochemical polarization The disabling of a primary cell caused by gas products deposited around or on one of the electrodes.

electrochemical recording See ELECTRIC RECORDING.

electrochemical reduction Extracting a metal from a compound by means of electrolysis.

electrochemical series See ELECTROMOTIVE SERIES.

electrochemical switch A static, ionic alternating-current (ac) switch consisting of an anode, cathode, and a control electrode, all immersed in an electrolyte. A positive control-electrode voltage switches the device on, initiating ion current from anode to cathode through the liquid.

electrochemical transducer A transducer that converts chemical changes into electrical quantities, or vice versa. Examples: soil-acidity probe and electrolytic elapsed-time meter.

electrochemistry The branch of chemistry, overlapping with physics, concerned with chemical action arising from the effect of electricity on substances, and electrical effects produced by chemical action.

electrochromic display A display that operates by means of electric fields, which control the light-transmission and light-reflection characteristics in different regions of the material.

electrochronometer A precision electric or electronic clock. Also see ELECTRIC CLOCK and ELECTRONIC CLOCK.

electrocoagulation Use of a high-frequency alternating current to solidify tissue, as in arresting bleeding.

electrocorticogram See ELECTROENCEPHALOGRAM.

electroculture Acceleration or modification of plant growth through the application of electricity to plants, seeds, or soil.

electrocution **1.** Death resulting from electric shock. **2.** An electric shock inflicted for the purpose of causing death (e.g., in an ELECTRIC CHAIR).

electrode A body, point, or terminal in a device or circuit that delivers electricity, or to which electricity is applied. A positive electrode is usually an ANODE; a negative electrode is usually a CATHODE.

electrode admittance The admittance encountered by current flowing through an electrode; the property is entirely that of the electrode and is the reciprocal of ELECTRODE IMPEDANCE. Consists of a real-number component (ELECTRODE CONDUCTANCE) and an imaginary-number component that is 90 degrees out of phase (ELECTRODE SUSCEPTANCE).

electrode capacitance The capacitance between an electrode and a reference body, such as ground or another electrode.

electrode characteristic The mathematical function or graph of electrode current versus electrode voltage.

electrode conductance The conductance encountered by current flowing through an electrode; the property is entirely that of the electrode and is the reciprocal of ELECTRODE RESISTANCE.

electrode current Current entering or leaving an electrode.

electrode dark current See DARK CURRENT.

electrode dissipation The power lost in the form of heat in an electrode.

electrode drop Voltage drop resulting from ELECTRODE RESISTANCE.

electrode efficiency In an electrolytic cell, the ratio of the yield of metal deposited to the maximum possible theoretical yield. This value ranges between 0 and 1 (0 percent and 100 percent).

electrode impedance The impedance encountered by alternating current flowing through an electrode; the property is entirely that of the electrode and is the reciprocal of ELECTRODE ADMITTANCE. Consists of a real-number component (ELECTRODE RESISTANCE) and an imaginary-number component that is 90 degrees out of phase (ELECTRODE REACTANCE).

electrodeless discharge Discharge in a gas tube that is not directly connected to a power source. A familiar example is the glow of a neon lamp held in a strong radio-frequency (RF) electromagnetic field.

electrodeposit **1.** To deposit a substance by electrical action. Also see ELECTROPHORESIS and ELECTROPLATING. **2.** A deposit that is formed on an electrode by electrophoresis or electroplating.

electrodeposition The electrical application of a layer of one material (such as a metal) on the surface of another (the substrate) (e.g., electroplating, evaporation, and sputtering).

electrode potential See ELECTRODE VOLTAGE.

electrode reactance The imaginary-number component of ELECTRODE IMPEDANCE.

electrode resistance The resistance encountered by current flowing through an electrode; the property is entirely that of the electrode and is the reciprocal of ELECTRODE CONDUCTANCE.

electrodermography A method of monitoring the functions of the human body by measuring the resistance between two electrodes placed on the surface of the skin.

electrode susceptance The imaginary-number component of ELECTRODE ADMITTANCE.

electrode voltage The voltage between an electrode and a reference point, such as ground or another electrode.

electrodiagnosis **1.** The diagnosis of a disease or disorder through the use of electromedical instruments. **2.** Troubleshooting the electrical portion of electromechanical equipment.

electrodialysis See ELECTRO-OSMOSIS.

electrodissolution Dissolving a constituent substance of an immersed electrode by means of electrolysis.

electrodynamic Pertaining to electricity in motion (i.e., current flow and its accompanying electric and magnetic fields).

electrodynamic braking Stopping a tape-deck motor quickly by applying a braking voltage. In this method, direct braking current flows through the shaded-pole alternating-current (ac) reel motor.

electrodynamic instrument See ELECTRODYNAMOMETER.

electrodynamic loudspeaker See DYNAMIC SPEAKER.

electrodynamics The branch of physics concerned with moving electric charge carriers, such as electrons, holes, and ions. Also concerned with the interaction of electrical and mechanical phenomena. Important in the design and manufacture of devices, such as motors and generators.

electrodynamic speaker See DYNAMIC SPEAKER.

electrodynamism See ELECTRODYNAMICS.

electrodynamometer An indicating meter whose movable coil rotates between two stationary coils. All three coils are connected in series, and the magnetic fields of the two stationary coils are additive. This meter produces a positive indication for alternating current (ac), as well as for direct current (dc). Can be adapted for use as an ammeter, voltmeter, or wattmeter.

electroencephalogram Abbreviation, EEG. A record made by an ELECTROENCEPHALOGRAPH, showing changes in electric potential resulting from bioelectric action in the brain. The record is used as a diagnostic aid.

electroencephalograph An instrument that produces a record of voltage changes resulting from the brain's bioelectricity.

electroencephaloscope A type of oscilloscope used to pick up, amplify, and display changes in potential caused by the brain's bioelectric action.

electroextraction Extracting a substance from a mixture (e.g., a metal from an ore) by an electrical process, such as electrolysis.

electroform **1.** To precondition a material or device (e.g., a semiconductor junction) by passing a current through it for a specified period. **2.** To form articles by electrodepositing material on a mold or core.

electrogalvanized Electroplated with zinc.

electrogastrogram A recording of the electrical impulses and other functions of the stomach, for medical diagnostic purposes.

electrograph **1.** A picture transmitting or receiving device (see FACSIMILE RECEIVER and FACSIMILE TRANSMITTER). **2.** A device used for the electrolytic etching or transfer of designs.

electrographic recording A method of producing a visible pattern or record, using electrodes to create discharge through an insulating material.

electrographite Synthetic graphite prepared by heating carbon in an electric furnace.

electrojet A region of high current concentration in the sky near bright auroral displays or along the magnetic equator.

electrokinetic energy The energy of electric charge carriers in motion; a form of kinetic energy. The current in a wire (a flow of electrons) is a common manifestation of this form of energy. Compare ELECTROSTATIC ENERGY.

electrokinetics A branch of electricity concerned with (1) the behavior of moving charged particles (such as ions and molecules) and bodies in motion, and (2) the generation of static charges by moving liquids or solids in contact with each other.

electroless process Plating a metal from a solution of one of its salts without using electricity.

electroluminescence The ability of certain phosphors to emit light continuously when an alternating-current (ac) voltage is applied to them.

electroluminescent cell A device for generating light by electroluminescence. It consists of a luminescent-phosphor layer and two transparent metal films. An alternating-current (ac) voltage applied between the films causes the phosphor to glow through the transparent metal.

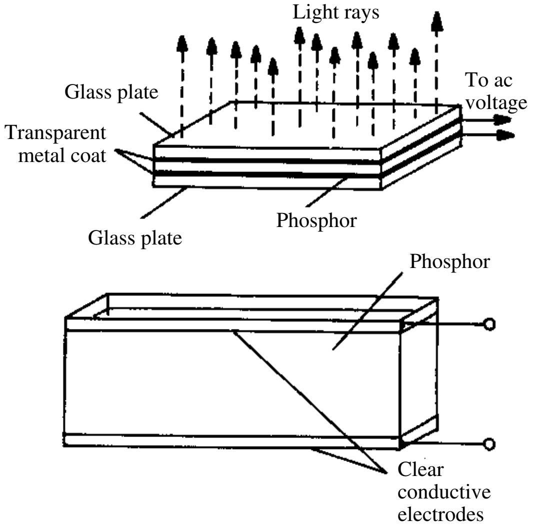

electroluminescent cell

electroluminescent lamp A lamp using one or more electroluminescent panels.

electroluminescent panel A panel that forms a complete electroluminescent cell. It delivers low-intensity visible light when an alternating-current (ac) voltage is applied to it. Available in various sizes and shapes.

electrolysis **1.** The action whereby a current passing through a conductive solution (electrolyte) produces a chemical change in the solution and the electrodes. **2.** An electrical method of destroying hair roots.

electrolyte A substance that ionizes in solution. Electrolytes conduct electricity; in batteries, they are instrumental in producing electricity by chemical action.

electrolytic **1.** Containing an electrolyte substance. **2.** See ELECTROLYTIC CAPACITOR.

electrolytic capacitor A component that is manufactured by rolling up aluminum foil strips separated by paper saturated with electrolyte. When a polarized voltage is applied, the aluminum oxidizes. The oxide layer forms the dielectric for the capacitor. The layer is extremely thin, producing high capacitance per unit volume. This type of capacitor must be connected with the proper polarity in a circuit. The component can have values up to thousands of microfarads, and some units can handle thousands of volts. They are used in audio circuits and in power supplies. Compare CERAMIC CAPACITOR, MICA CAPACITOR, PAPER CAPACITOR, PLASTIC-FILM CAPACITOR, TANTALUM CAPACITOR.

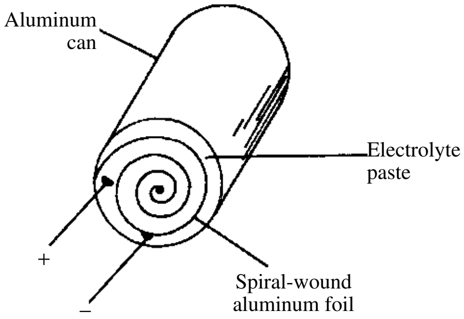

electrolytic capacitor

electrolytic cell An electrochemical device containing an electrolyte and at least two electrodes. Included in this category are voltaic cells, electrolytic capacitors, and electrolytic resistors.

electrolytic conduction Electric current flowing through an electrolyte, an action characterized by (1) positive electrolyte ions migrating to the negative electrode, where they acquire electrons; (2) negative ions migrating to the positive electrode, where they lose electrons; and (3) current flow in the external circuit, which consists of conventional electron flow (current in the electrolyte is a movement of ions).

electrolytic conductivity Conductance of an electrolyte. It is the conductance of a cube of the electrolyte, measuring one centimeter along each edge, at a specified temperature.

electrolytic corrosion Corrosion caused by an applied voltage, or that is accelerated by the voltage. Compare GALVANIC CORROSION.

electrolytic current meter See VOLTAMMETER.

electrolytic dissociation See DISSOCIATION.

electrolytic elapsed-time meter An instrument that indicates the time that equipment has been in operation in terms of the amount of metal electroplated on the cathode of an electrolytic cell by energy consumed during the period.

electrolytic gas A gas produced by electrolysis. Examples are hydrogen (H) and oxygen (O), generated in a ratio of two to one, by the electrolysis of water (H_2O).

electrolytic iron Very pure iron obtained by electrolytic refining.

electrolytic polarization In electrolysis, the tendency for the products to recombine. In an electrolytic cell, this effect interferes with the performance of the cell, reducing the voltage.

electrolytic potential The difference of potential that appears between a metal electrode in an electrolyte and the electrolyte immediately surrounding it. Also see ELECTROMOTIVE SERIES.

electrolytic recorder A data recorder that uses a paper impregnated with a chemical that turns dark when an electric current passes through the paper from the point of a stylus.

electrolytic rectifier A rectifier consisting of an aluminum electrode and a lead or carbon electrode in a solution of borax or sodium bicarbonate, or in a solution of ammonium citrate, ammonium phosphate, and potassium citrate. Also called *chemical rectifier.*

electrolytic refining Extracting or purifying metals by electrolysis.

electrolytic resistor An emergency resistor made by immersing two wire leads in an electrolyte; the weaker the solution, the higher the resistance.

electrolytic switch See ELECTROCHEMICAL SWITCH.

electrolyze To subject something to electrolytic action.

electrolyzer A cell used in the production of various materials by electrolysis. See, for example, ELECTROCHEMICAL REDUCTION and ELECTROLYTIC REFINING.

electromagnet **1.** A coil of insulated wire wound around an iron or steel cylinder, intended for use as a magnet. When current flows through the coil, a magnetic field develops, in effect rendering the cylinder a strong bar magnet. **2.** Any device that exhibits magnetism only while an electric current flows through it.

electromagnetic Exhibiting both electric and magnetic properties (e.g., an *electromagnetic wave*).

electromagnetic attraction **1.** The attraction of iron or steel to an electromagnet. **2.** The attraction of an electromagnetic pole to the opposite pole of another electromagnet. A unit pole attracts another unit pole 1 centimeter away with a force of 1 dyne or 10^{-5} newton. Compare ELECTROMAGNETIC REPULSION.

electromagnetic communication **1.** Any form of communication using a combination of electric and magnetic phenomena. Examples include wire telegraphy, wire telephony, radiotelegraphy, radiotelephony, facsimile, and television. **2.** Electronic communication via electromagnetic fields (i.e., radio communication).

electromagnetic compatibility **1.** The ability of a set of electronic devices to work together without being adversely affected by each other's electromagnetic fields. **2.** In radio communication, the relative immunity of a device or devices to the effects of electromagnetic fields.

electromagnetic complex A system that produces electromagnetic radiation.

electromagnetic component **1.** The magnetic component of an electromagnetic wave, which is perpendicular to the electrostatic component, and can be thought of as the wave's current component. **2.** A device operated by electromagnetism, such as a coil-type relay or a current-operated field magnet.

electromagnetic consonance An effect that takes place when two antenna elements, both having identical or nearly identical resonant frequencies, are in close proximity. If one antenna is fed with energy at its resonant frequency, currents will be induced in the other antenna, and it, too, will radiate. Parasitic arrays, such as the Yagi antenna and the quad antenna, operate on this principle. See RESONANCE.

electromagnetic constant **1.** Symbol, *c*. The propagation speed of electromagnetic waves in a vacuum, approximately equal to 299,792 kilometers per second or 186,282 miles per second. Also called SPEED OF LIGHT and VELOCITY OF LIGHT. **2.** The propagation speed of electromagnetic waves in a particular medium. Equal to vc, where v is the VELOCITY FACTOR of the medium, and c is the speed of electromagnetic waves in a vacuum (299,792 kilometers per second or 186,282 miles per second).

electromagnetic coupling See INDUCTIVE COUPLING.

electromagnetic crack detector An instrument that uses electromagnetic fields to find cracks in iron or steel.

electromagnetic CRT A cathode-ray tube using electromagnetic deflection.

electromagnetic cylinder See SOLENOID.

electromagnetic deflection In a television picture tube and some oscilloscopes, deflection of the electron beam by the magnetic fields of external horizontal- and vertical-deflection coils. Compare ELECTROSTATIC DEFLECTION.

electromagnetic deflection coil See DEFLECTION COILS.

electromagnetic delay line See DELAY LINE.

electromagnetic energy Energy in the form of electric and magnetic fields. A radio wave traveling through space, for example, has electric and magnetic components, between which energy oscillates.

electromagnetic energy conversion The conversion of electrical energy into mechanical work and vice versa, through the intermediary of an electromagnetic field.

electromagnetic environment A region in which electric and magnetic fields are present.

electromagnetic field A combination of alternating electric and magnetic fields. The electric lines of flux are perpendicular to the magnetic lines of flux at every point in space. The field propagates in a direction perpendicular to both the electric and magnetic lines of flux. The frequency of oscillation can range from a fraction of one Hz to many quadrillions of Hz. In order from longest wavelength (lowest frequency) to shortest wavelength (highest frequency), effects of this type include radio waves, infrared, visible light, ultraviolet, X rays, and gamma rays.

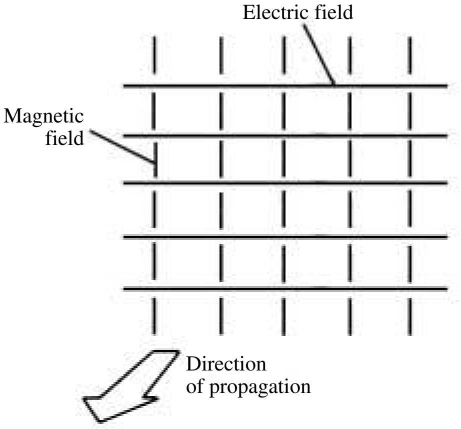

electromagnetic field

electromagnetic flux The magnetic field surrounding a coil or conductor carrying an electric current.

electromagnetic focusing In a television picture tube, electron-beam focusing obtained by varying the direct current flowing through an external focusing coil.

electromagnetic force The force that causes a conductor to be displaced from its position in a magnetic field when it conducts current.

electromagnetic frequency spectrum The frequency range of electromagnetic fields including radio waves, infrared, visible light, ultraviolet, X rays, and gamma rays. Ranges from a fraction of one Hz to many quadrillions of Hz.

electromagnetic horn radiator A horn used to radiate microwave energy. Also see HORN ANTENNA.

electromagnetic induction Inducing a voltage in a circuit or conductor by causing alternating current to flow in another nearby circuit or conductor. Compare ELECTROSTATIC INDUCTION.

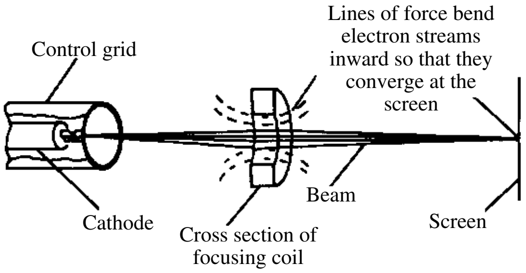
Control grid
Lines of force bend electron streams inward so that they converge at the screen
Cathode
Cross section of focusing coil
Beam
Screen

electromagnetic focusing

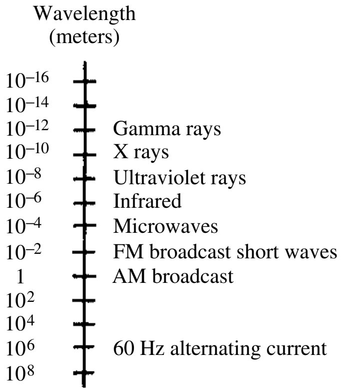

electromagnetic frequency spectrum

electromagnetic inertia The tendency for the current in a circuit to lag the voltage at high frequencies.

electromagnetic interference Abbreviation, EMI. A phenomenon in which electronic devices upset each other's operation. Computers, television receivers, telephone sets, high-fidelity sound equipment, and certain medical devices can malfunction because of strong radio-frequency fields such as those from a nearby broadcast transmitter. The EMI is usually the result of improper or ineffective shielding in the affected device or system.

electromagnetic lens A coil or coil system whose magnetic field causes an electron beam passing through it to converge or diverge as a light beam does in passing through an optical lens. Compare ELECTROSTATIC LENS. Also see ELECTROMAGNETIC FOCUSING.

electromagnetic mass The mass that a moving electric charge is thought to possess.

electromagnetic microphone A microphone in which sound energy is converted into proportionate electrical energy by electromagnetism. Common examples are the *dynamic microphone* and *velocity microphone.*

electromagnetic mirror A reflector of electromagnetic waves (e.g., antenna elements, ionospheric layers, buildings, and hills).

electromagnetic momentum The momentum of a moving electric charge, comparable to that of matter in motion. Electromagnetic momentum is the product of electromagnetic mass and charge velocity.

electromagnetic oscillograph **1.** An electromechanical data recorder for tracing the waveform or variations of a signal. **2.** See ELECTROMECHANICAL OSCILLOSCOPE.

electromagnetic oscilloscope **1.** An oscilloscope using electromagnetic deflection. **2.** See ELECTROMECHANICAL OSCILLOSCOPE.

electromagnetic pump A device used for moving conducting or semiconducting fluids. When a current is passed through the fluid, a force is exerted on the molecules of the fluid because of the magnetic field set up by the current.

electromagnetic pulse Abbreviation, EMP. **1.** A broadband electromagnetic field emitted in a short, intense burst from a lightning stroke or nuclear explosion. This field can disrupt the operation of, and (in some cases) cause damage to, electrical and electronic apparatus. **2.** In electromagnetic induction, the displacement of an electron in a conductor by the magnetic field.

electromagnetic radiation The propagation of electromagnetic fields through space; it normally occurs at approximately 299,792 kilometers per second or 186,282 miles per second.

electromagnetic reaction The reaction between magnetic fields. Also see ELECTROMAGNETIC ATTRACTION and ELECTROMAGNETIC REPULSION.

electromagnetic reconnaissance In military applications, the use of electromagnetic apparatus to detect potential enemy activity in a certain geographic region.

electromagnetic relay See ELECTROMECHANICAL RELAY.

electromagnetic repulsion The repulsion of a pole of an electromagnet by the pole of another electromagnet (north pole opposing north pole, south opposing south). Compare ELECTROMAGNETIC ATTRACTION.

electromagnetics A branch of physics concerned with the theory and application of electromagnetism.

electromagnetic screen See ELECTROMAGNETIC SHIELD.

electromagnetic shield **1.** A partition, can, or box made of magnetic material (iron, steel, or special alloy) enclosing a magnetic component. The magnetic flux generated by the component is confined by the shield, thus preventing interference with external components. Likewise, external magnetic fields are prevented from reaching the component. **2.** A grounded partition, metal sheet, wire braid, or other barrier that prevents electromagnetic fields from passing through. Commonly used in electronic equipment to prevent ELECTROMAGNETIC INTERFERENCE. Also used in COAXIAL CABLE to confine electromagnetic fields to the transmission line.

electromagnetic shielding The use of an ELECTROMAGNETIC SHIELD to prevent undesired interaction among electrical and electronic devices and systems.

electromagnetic spectrum See ELECTROMAGNETIC FREQUENCY SPECTRUM and ELECTROMAGNETIC WAVELENGTH SPECTRUM.

electromagnetic switch **1.** A switch actuated by magnetism produced by control current flowing through a coil wound on an iron core. **2.** See ELECTROMECHANICAL RELAY.

electromagnetic theory of light The theory that light consists of electromagnetic waves that are similar to radio waves, but of shorter wavelength.

electromagnetic tube A cathode-ray tube using electromagnetic deflection (e.g., a television picture tube).

electromagnetic unit Abbreviation, emu. A unit of measure in the electromagnetic system of CENTIMETER-GRAM-SECOND (cgs) units.

electromagnetic vibrator See INTERRUPTER.

electromagnetic wavelength spectrum The wavelength range of electromagnetic fields, including radio waves, infrared, visible light, ultraviolet, X rays, and gamma rays. It ranges from many kilometers to a tiny fraction of one millimeter.

electromagnetic-wave polarization The orientation of the electric flux lines in an electromagnetic (EM) field, especially a field propagating through space. The polarization is generally parallel with the active element of a radio transmitting or receiving antenna. Thus, a vertical antenna radiates and receives fields with *vertical polarization*, and a horizontal antenna radiates and receives fields having *horizontal polarization.* Some antennas radiate and receive an EM field whose polarization continually and rapidly rotates. This is *elliptical polarization.* If the rate of rotation is constant, it is *circular polarization.*

electromagnetic waves Waves produced in a conductor or in space by the acceleration or oscillation of electric charge carriers. Such waves have an electric and a magnetic component acting at right angles to each other. The waves propagate at right angles to both the electric and magnetic flux lines.

electromagnetism **1.** Magnetism resulting from the movement of electric charge carriers (e.g., the magnetic field surrounding a coil of wire carrying an electric current). **2.** See ELECTROMAGNETICS.

electromagnetizer A magnetizer using continuous direct current (dc) as the magnetic-field source.

electromechanical Descriptive of any device that converts energy from electrical to mechanical form or vice-versa. Examples are the motor and the generator.

electromechanical amplifier An amplifier that converts an electrical input signal into mechanical motion (vibratory or rotary), which it then converts back into an electrical output signal of higher current, voltage, or power.

electromechanical chopper A vibrator-type interrupter used primarily to chop direct current, converting it into a square-wave signal, whose amplitude is proportional to current strength. Also see CHOPPER.

electromechanical counter A device that indicates the number of pulses that have been applied to it. Typically, it has a series of dials—each capable of displaying the numerals 0 to 9 in sequence, one for each decade in the count. The dials are geared together, the train being operated by the stepping action of an electromagnetic escapement. Compare ELECTRONIC COUNTER.

electromechanical energy The energy stored by an inductor or capacitor in an electromechanical device.

electromechanical filter See ULTRASONIC FILTER, **1**.

electromechanical flip-flop See BISTABLE RELAY.

electromechanical frequency meter A usually direct-reading instrument for measuring frequency in the lower and middle portions of the audio spectrum. It works via the mechanical motion resulting from the applied signal. The two varieties are the movable-iron type and the reed type.

electromechanical modulator See CHOPPER.

electromechanical oscillator An oscillator consisting of an electromechanical amplifier provided with regenerative feedback.

electromechanical oscilloscope A galvanometer-type instrument for displaying a varying or alternating current or voltage. The signal is applied to a meter movement having a movable coil, which swings or vibrates in response to the signal. A tiny mirror cemented to the coil reflects a beam of light to a rotating mirror that sweeps the beam across a translucent screen on which the image is produced.

electromechanical recorder An instrument in which a pen or stylus is moved on a sheet of paper by the varying signal current or voltage being recorded.

electromechanical rectifier A rectifier in which a moving part, such as a vibrating reed or rotating commutator-slip-ring unit, is driven by alternating current (ac) to close the circuit during positive or negative ac half-cycles, thus rectifying the ac. Compare ELECTROLYTIC RECTIFIER and ELECTRONIC RECTIFIER.

electromechanical relay An electromagnetic switch consisting of a multiturn coil wound on an iron core near an armature with a movable end contact. When control current flows through the coil, it becomes magnetized and attracts the armature, closing the movable contact against a stationary one.

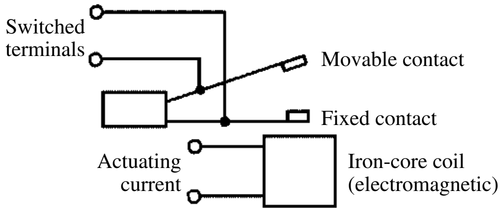

electromechanical relay

electromechanical timer A device for automatically timing a process or an observed event. Most such timers are based on an accurate clock (electric or spring driven) that opens or closes contacts at predetermined instants. Compare ELECTRONIC TIMER.

electromechanical transducer A transducer that translates mechanical signals directly into electrical ones or vice versa, without the intermediary of active devices, such as transistors or integrated circuits.

electromechanical valve A usually poppet-type valve for gases or liquids. The valve is operated by electromagnetic action and is often aided by an electronic (servo) circuit.

electromechanics The theory and application of electromechanical devices.

electromedical engineering The branch of electronics engineering concerned with the theory, design, and application of electronic equipment for medical diagnosis or treatment.

electromedical equipment Electrical or electronic equipment used in medical diagnosis or treatment.

electromerism Ionization in gases.

electrometallurgy The branch of metallurgy concerned with the use of electricity (especially in electrolysis) to separate or purify metals or to furnish heat for metallurgical processes.

electrometer A specially designed, highly sensitive electronic voltmeter used to measure extremely low voltages and, indirectly, extremely low currents. It is sometimes used as a galvanometer.

electrometer amplifier A stable low-noise amplifier for increasing the sensitivity of an electrometer.

electrometry The science of electrical measurements.

electromigration The movement of atoms in a substance from one place to another, because of interaction between electrons and ions in the presence of electric currents. This effect can cause the eventual deterioration of certain semiconductor devices.

electromotion Motion produced by electric charges or electrons.

electromotive force Abbreviation, emf. Electrical pressure, the potential that causes charge carriers to move through a substance or circuit. See VOLTAGE.

electromotive series A list of metals arranged according to the potential between the surface of the metals and an electrolyte into which they are immersed. Some metals acquire a positive potential (with respect to hydrogen, for which the potential is zero) and others, a negative potential.

electromotor A generator or motor, depending upon the context in which the term is used.

electromyogram The record produced by an ELECTROMYOGRAPH.

electromyograph An instrument for detecting, measuring, and analyzing the weak electrical currents generated by muscular activity.

electromyography The monitoring and analysis of the electrical activity of human muscles.

electron The subatomic particle that carries the unit negative charge of electricity. The electron has a mass of 9.109×10^{-31} kilogram and carries a charge of 1.602×10^{-19} coulomb.

electron acceleration See ELECTRON MOTION, **2**.

electronarcosis Loss of consciousness caused by passing a weak current through the brain. Useful in treating certain mental disorders, the process is somewhat similar to ELECTROSHOCK.

electron attachment The bonding of an electron to a neutral atom to form a negative ion. Also see ANION and ION.

electron avalanche See AVALANCHE and ELECTRON MULTIPLICATION.

electron band **1.** An emission line in the spectrum of an element or compound, caused by the movement of electrons from higher to lower energy levels within the atoms. **2.** An absorption line in the spectrum of an element or compound, caused by the movement of electrons from lower to higher energy levels within the atoms.

electron beam See ELECTRON STREAM, **1**.

electron-beam bender Any element that causes intentional deflection of the electron stream in a cathode-ray tube.

electron-beam focusing Reducing the size of the spot produced by the electron beam in a cathode-ray tube or television picture tube. This is accomplished by adjusting the direct-current (dc) bias voltage on a focusing electrode.

electron-beam generator **1.** See ELECTRON GUN. **2.** A tube, such as a Klystron, in which velocity modulation of the electron beam generates extremely high radio frequencies.

electron-beam instrument An instrument, such as an oscilloscope, based on a cathode-ray tube.

electron-beam machining Welding or shaping materials by controlled electron beams.

electron-beam magnetometer A magnetometer in which the magnetic field under measurement deflects the electron beam in a cathode-ray tube over a distance that is proportional to field intensity.

electron-beam recording In digital computer operations, a technique whereby the output of a computer is recorded on microfilm by an electron beam.

electron-beam scanning tube A tube in which an electron beam strikes a sensitized screen to produce a spot of light, which is deflected electrically or magnetically across a screen. Examples are oscilloscope tubes, storage tubes, and television picture tubes.

electron-beam tube Electron tubes (such as beam-power tubes, klystrons, oscilloscope tubes, and television picture tubes), in which an electron beam is generated and controlled.

electron-beam welding A method of welding in which an electron beam is focused on the workpiece to heat it.

electron-bombarded semiconductor A semiconductor wafer, plate, or junction that is acted on by an electron beam; it alters the resistance of the semiconductor to control the current in an external circuit.

electron-bombarded semiconductor amplifier See EBS AMPLIFIER.

electron bunching See BUNCHING.

electron charge See ELEMENTARY CHARGE.

electron cloud A mass of free electrons.

electron diffraction Diffraction that occurs when a beam of electrons passes through a crystal material. Fast-moving electrons have wavelike properties; the wavelength depends on the speed of the particles. This effect can also occur with beams of other particles, such as neutrons, protons, or alpha particles.

electron drift **1.** The movement of an electron from atom to atom in a conductor, as caused by the influence of an applied voltage. **2.** In a semiconductor, directed electron movement. Also see DRIFT, **1**.

electronegative Having negative electrification or polarity (see ELECTRICAL POLARITY). Compare ELECTROPOSITIVE.

electron emission The emission of electrons into surrounding space by a material. Depending on the material, this effect can be initiated by application of heat, light, torsion, electron impact, a high-voltage field, and other actions.

electron flow See ELECTRON DRIFT.

electron g-factor A physical constant that expresses the ratio of electron magnetic moment to the Bohr magneton, and equal to approximately 1.00116. Also called *free-electron g-factor.*

electron-gas binding forces See METALLIC BINDING FORCES.

electron-gas bonding See BONDING, **1** and METALLIC BINDING FORCES.

electron gun A composite electrode for generating an electron beam (see ELECTRON STREAM, **1**) in a vacuum. In a cathode-ray tube, the gun comprises a heated cathode, control electrode, accelerating electrodes, and a focusing electrode.

electron-hole pair In a semiconductor, an electron and a related hole. Each electron in the conduction band has a counterpart in the valence band, a vacancy (hole) left by the electron's moving to the conduction band.

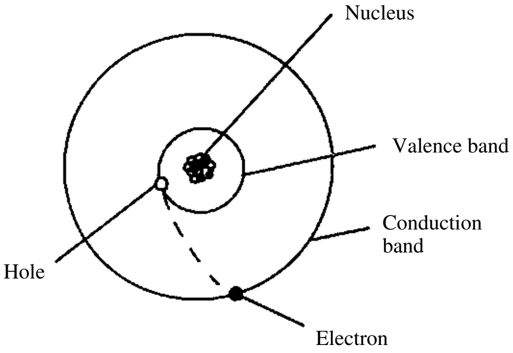

electron-hole pair

electronic **1.** Descriptive of any component, device, or system that functions, according to the principles of ELECTRONICS. **2.** Pertaining to electrons.

electronic adder A circuit (such as an operational amplifier) for performing arithmetic addition. In such a circuit, the output-signal amplitude is the sum of the input-signal amplitudes. Also see ADDER and ANALOG ADDER.

electronic aid An electronic device or circuit that contributes to the operation of a nonelectronic device or system; a pH meter, for example, is an electronic aid to chemistry.

electronics aide See ELECTRONICS TECHNICIAN.

electronic attenuator An attenuator in which the variable resistor is the output section of an active device, such as a transistor. Varying the direct-current (dc) bias of the input section varies the resistance of the output section.

electronic autopilot A servo system for detecting and automatically correcting an aircraft's flight path.

electronic balance An electronic scale, which uses a sensitive current-measuring device in conjunction with a movable tension device.

electronic brain **1.** A high-end computer. **2.** A robot controller. **3.** A system that possesses or uses artificial intelligence (AI).

electronic breadboard **1.** A thin, usually nonmetallic board or card having prepunched holes for the quick assembly of electronic circuits for test and evaluation. **2.** Any circuit prototype that is manually wired during the experimental phase of product development.

electronic bug **1.** A semiautomatic telegraph key that produces dots via an oscillating mechanical arm. The operator produces dashes manually. **2.** A telegraph keying device that automatically generates dots and dashes. Also see ELECTRONIC KEY, **1**. **3.** An undetermined source of problems or improper operation in an electronic circuit.

electronic buzzer **1.** A mechanical buzzer driven by a direct-current (dc) amplifier. **2.** An oscillator circuit that produces a sound similar to that of a mechanical buzzer.

electronic calculating punch A machine that punches on a card the result of calculations it has performed on data it has read from another punched card.

electronic calculator A fully electronic machine for performing mathematical calculations. The simplest machines perform basic arithmetic; more sophisticated ones can do operations with trigonometric, logarithmic, hyperbolic, and other scientific functions. Basic machines are available in department stores for about five dollars. It is usually powered via a small cell or battery, or by a small photovoltaic panel.

electronic camouflage The use of electronics by a target craft to minimize or prevent the reflection of radar echoes.

electronic carillon An electronic system that produces sounds resembling those of a bell carillon.

electronic chime See ELECTRONIC CARILLON.

electronic circuit An electric circuit containing active electronic components, such as transistors and integrated circuits, as opposed to a circuit containing only passive electrical components (such as resistors, switches, heating elements, etc.).

electronic clock **1.** An electric clock whose motor is driven by a constant-frequency oscillator (crystal or tuning fork type), followed by multivibrators and amplifiers. **2.** Any electronic timing circuit that produces pulses at predetermined intervals for the purpose of regulating the operation of other circuits, subsystems, or assemblies.

electronic commutator See COMMUTATOR, **2**.

electronic conduction A flow of electric current resulting from the movement of electrons among atoms in a conductor.

electronic control **1.** The science of automatically controlling machines and devices by means of electronic circuits. **2.** A circuit or device that provides automatic electronic control.

electronic controller A controller (see CONTROLLER, **2**) having no moving parts. For automatic operation, such a device often contains a circuit that senses control signals, compares them with a signal standard, and automatically adjusts the output control power accordingly.

electronic counter A fully electronic circuit that indicates the number of pulses that have been applied to it (see COUNTER, **1**). Unlike the electromechanical counter, the electronic counter has no moving parts and is therefore capable of extremely high-speed, noiseless operation.

electronic counter-countermeasures Abbreviation, ECCM. Military procedures for interfering with a foe's electronic countermeasures.

electronic countermeasures Abbreviation, ECM. Interference with enemy radio and radar emissions by electronic means. Also see JAMMING.

electronic coupling Coupling via electronic effects or devices.

electronic crowbar A switch that prevents destructive currents from flowing through the components of a circuit.

electronic current meter A current meter that uses an amplifier ahead of an analog or digital indicator to provide increased sensitivity.

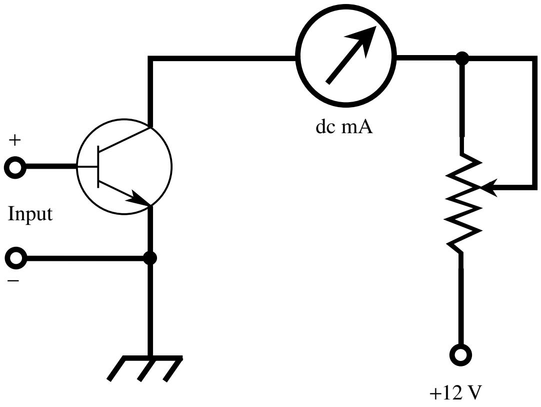

electronic current meter

electronic data processing Abbreviation, EDP. See DATA PROCESSING and ELECTRONIC INFORMATION PROCESSING.

electronic data-processing center Abbreviation, EDPC. An installation of electronic equipment and accessories for processing and storing data, usually in digital form. Also see DATA PROCESSING and ELECTRONIC INFORMATION PROCESSING.

electronic data-processing machine Abbreviation, EDPM. A device, such as an electronic computer, used in the automatic processing of data, usually in digital form.

electronic data-processing system **1.** A unique arrangement of machines for processing data. **2.** The sequence of steps in, and the underlying rationale for, the processing of data by automated equipment.

electronic deception See DECEPTION and DECEPTION DEVICE.

electronic differential analyzer See ANALOG DIFFERENTIATOR, DIGITAL DIFFERENTIAL ANALYZER, and DIFFERENTIAL ANALYZER.

electronic differentiator A circuit that performs mathematical differentiation. Also see DIFFERENTIATOR, **2**.

electronic divider **1.** An electronic device for performing arithmetic division. In a digital computer, such a divider can be a sequence of flip-flops, each of which produces a single output for every two input pulses. In an analog computer, the output signal amplitude is equal to the quotient of two input-signal amplitudes. **2.** See FREQUENCY DIVIDER. **3.** A voltage divider using active components, rather than resistors.

electronic dust precipitator See DUST PRECIPITATOR.

electronic efficiency A quantitative expression for the effectiveness of an electron beam as a medium of power transmission. The electronic efficiency, in percent, is equal to $100P_{out}/P_{in}$, where P_{out} is the (output) power delivered by the beam, and P_{in} is the (input) power supplied to the beam.

electronic equivalent of gravity In equations for the acceleration, velocity, and distance traveled for an electron, the factor equal to eF/m, where e is the electron charge, F is the potential gradient of field, and m is the electron mass. Also see ELECTRON MOTION, **2**.

electronic flash **1.** A device containing a circuit that uses an electronic flash tube as a light source for photography or other purposes. Also called *photoflash.* **2.** A bright momentary light burst produced by the equipment described in **1**, above.

electronic flash tube A tube used to produce brilliant bursts of light in photoflash units, stroboscopes, and laser exciters. A flash tube usually contains xenon gas, which is fired by a high-voltage pulse.

electronic frequency meter **1.** An instrument that gives direct readings of frequency in hertz, kilohertz, or megahertz on an analog scale or as a digital readout. **2.** Any device that indicates the operating frequency of another device, directly or indirectly, when used for such purpose.

electronic frequency synthesizer An instrument that supplies a number of selectable frequencies derived from one or more internally generated fixed frequencies.

electronic gas A collection of free electrons whose behavior resembles that of a gas.

electronic gate **1.** A logic gate that operates by electronic means. **2.** A security system, consisting of a mechanical gate controlled electronically. Similar to an electronic garage door.

electronic guitar See ELECTRIC GUITAR.

electronic heating The production of heat in an object via high-frequency energy. The two principal methods are *dielectric heating* and *induction heating.*

electronic hygrometer An electrical device for measuring relative humidity, whose sensitivity and stability have been increased by the addition of active amplifying devices.

electronician See ELECTRONICS TECHNICIAN.

electronic induction See ELECTROSTATIC INDUCTION.

Electronic Industries Association Abbreviation, EIA. An American association of electronic manufacturers and engineers. It sets standards, disseminates information, provides industry-government liaison, and maintains public relations for the industry.

electronic information processing The use of electronic equipment (especially digital computers and attendant devices) to perform mathematical operations on data entered into the system in the form of electrical signals. Also see DATA PROCESSING.

electronic instrument An instrument whose circuit uses active devices for increased sensitivity over that of the electrical counterpart, and for minimum loading of a device under test. Compare ELECTRICAL INSTRUMENT.

electronic integrator A active device (such as an operational amplifier) for performing mathematical integration. Also see INTEGRATOR, **2**.

electronic intelligence **1.** Information exchanged by electronic means. Examples: radio messages, radar information, and computer data. **2.** The faculties of reasoning and decision making, as apparently simulated by a high-level computer.

electronic interference The malfunctioning of a device because of nearby currents, voltages, or electromagnetic fields.

electronic inverter An electronic device for converting direct current (dc) to alternating current (ac). Typically, an inverter is a transistorized square-wave oscillator inductively coupled to ac output terminals. The dc to be inverted energizes the transistors, which perform the switching function at the rate determined by the components of the circuit. Also see INVERTER, **1**.

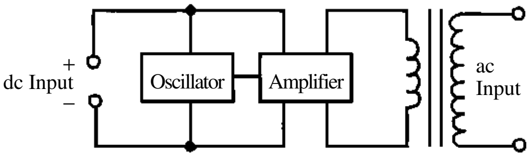

electronic inverter

electronic jamming The deliberate transmission of electromagnetic energy for the purpose of interfering with the operation of a device or devices.

electronic key **1.** For telegraphy (radio or wire), an electronic circuit that generates a continuous string of accurately spaced and timed Morse code dots or dashes, depending on its lever's position (right or left). **2.** For telegraphy (radio or wire), a keyboard that produces perfectly timed Morse code output that corresponds to the operator's typed input. **3.** An electronic device for opening an electronic lock.

electronic keyer See ELECTRONIC KEY, **1**.

electronic lock A lock that will open only after application of a special coded sequence of signals.

electronic mail Also called *e-mail.* A communications system that allows people to leave digital text messages for each other. It is popular among users of personal computers. Operates through the telephone lines using terminal emulation software and a modem. It can also be used via amateur radio packet communications.

electronic microammeter See FET CURRENT METER and TRANSISTOR CURRENT METER.

electronic microvoltmeter See MICROVOLTMETER.

electronic milliammeter See FET CURRENT METER and TRANSISTOR CURRENT METER.

electronic millivoltmeter A millivoltmeter that uses an amplifier ahead of an analog or digital indicator to provide high input impedance and increased sensitivity.

electronic multimeter A *voltohm-milliammeter* that uses active amplifying devices. Also see ELECTRONIC INSTRUMENT.

electronic multiplier A device, such as a Hall generator, whose output is equal (or proportional) to the product of two inputs (i.e., it can perform arithmetic multiplication).

electronic music **1.** Music produced by a combination of electronic oscillator, amplifier, and loudspeaker. A number of successful instruments have been developed. See, for example, ELECTRONIC CARILLON, ELECTRONIC ORGAN, ELECTRONIC PIANO, and THEREMIN. **2.** The electronically amplified sounds of conventional musical instruments.

electronic organ A musical instrument with a keyboard similar to that of a conventional organ, in which tones produced by oscillators or electrically driven reeds are processed and amplified for delivery to a system of loudspeakers.

electronic packaging See ENCAPSULATION.

electronic part A lowest replaceable unit, or component, in an electronic circuit.

electronic phase meter An electronic instrument for measuring phase difference. Direct readings, in degrees of lead or lag, are visible on an analog scale or as a digital readout.

electronic photoflash A transistorized light-intensity meter. Also see ELECTRONIC INSTRUMENT.

electronic piano A musical instrument having the keyboard of a conventional piano and provided with electronic amplification.

electronic power supply A power supply using transistors or integrated circuits for stabilization and output control.

electronic precipitator See DUST PRECIPITATOR.

electronic product Any commercially manufactured electronic device, intended for purchase by the public, by industry, or by government.

electronic profilometer An electronic instrument for measuring surface roughness.

electronic ratchet A stair-step circuit or other arrangement functioning in the manner of an equivalent electromechanical stepping switch. Also see COMMUTATOR.

electronic reconnaissance In military applications, the use of electronic systems to locate enemy installations (such as radio stations, guided-missile sites, and radar bases).

electronic rectifier A rectifier that uses active devices to change alternating current (ac) to direct current (dc).

electronic regulator A voltage regulator that uses active electronic circuits, as opposed to a reactor-type or electromechanical device. See, for example, VOLTAGE REGULATOR.

electronic relay **1.** A switching circuit that uses one or more transistors, and performs the relay function without moving parts. **2.** An electronic component designed to switch when gating signals are applied (e.g., triac, diac, or silicon-controlled rectifier).

electronic resistor **1.** The effective internal collector-emitter resistance of a common-emitter bipolar-transistor stage. **2.** The effective internal drain-source resistance of a field-effect transistor (FET) stage.

electronics The branch of physics concerned with the behavior and application of electric charge carriers in components, devices and systems that accomplish amplification, oscillation, signal processing, and/or switching.

electronics engineer A trained professional skilled in the physics and mathematics of electronics, and in the theory and application of basic engineering and related subjects. Compare ELECTRICAL ENGINEER.

electronics service person An electronics technician skilled in repairing and maintaining electronic equipment. Also called *electronics service technician.*

electronic shutter See KERR CELL.

electronics technician A professional skilled in building, testing, repairing, or maintaining electronic equipment.

electronics technology The theory and practical application of electronics. Taught as a subengineering major, usually in two-year junior colleges or technical institutes, awarding the degree of associate in arts (AA) or associate in science (AS).

electronic stethoscope A stethoscope employing a miniature microphone, amplifier, and earphones. The amplifier gain is continuously controllable, and its bandwidth often selectable for emphasizing particular heart sounds and other body noises.

electronic stimulator A device for applying controlled electrical pulses to the body to stimulate muscles or nerves during diagnosis or therapy.

electronic stroboscope A stroboscope that uses a rate-calibrated oscillator, rather than a mechanical contactor to generate pulses that strobe the lamp.

electronic subtracter An electronic circuit for performing arithmetic subtraction.

electronic surge A sudden, large increase in the current in a conductor. Can be caused by an electromagnetic pulse; can also occur when utility power is restored following a blackout.

electronic switch **1.** A nonmechanical device, such as a flip-flop or gate, whose characteristic on-off operation can be used to make and break an electric circuit. Compare CONTACT SWITCH. **2.** A device using electronic gating and sequencing circuits to present several signals alternately to the single input of an oscilloscope, allowing simultaneous viewing of the signals.

electronic tachometer An instrument for measuring angular velocity, usually in revolutions per minute (rpm). Ideally, the response is independent of sensor voltage amplitude, showing only the number of pulses per unit time reaching the meter circuit.

electronic thermal conductivity The thermal-conductivity component, resulting from the transfer of heat by electrons and holes.

electronic thermometer An instrument for measuring temperature as a result of variations in a temperature-sensitive component, such as a resistor, thermocouple, or thermistor.

electronic timer An electronic circuit or device for automatically timing a process or observed event. Most are based on the time constant of a stable resistance-capacitance (RC) circuit. Compare ELECTROMECHANICAL TIMER.

electronic tube See ELECTRON TUBE.

electronic tuning Variation of the resonant frequency of a device or circuit by changing the bias voltage or current of a controlling electronic component.

electronic voltmeter A voltmeter that uses electronic amplification ahead of the indicating meter to provide high input impedance and increased sensitivity. Also see FET VOLTMETER and TRANSISTOR VOLTMETER.

electronic voltohmmeter A voltohmmeter that uses electronic amplification ahead of the indicating meter to provide high input impedance and increased sensitivity.

electronic typewriter A typewriter with a microcomputer that provides features, such as buffering, automatic repeat, erase memory, and paragraph/page memory.

electronic warfare The use of electronic systems for military purposes, including interfering with an enemy's use of similar systems.

electronic watch **1.** A watch whose movement is a tiny, high-frequency, alternating-current (ac) motor driven by a stable oscillator. **2.** Any miniature timepiece incorporating solid-state circuitry, but especially one using a digital readout.

electronic wattmeter A wattmeter in which an amplifier is used for increased sensitivity.

electronic waveform synthesizer A signal generator that delivers an alternating or pulsating signal whose waveform can be tailored by means of adjustable circuit components.

electron lens A device that focuses an electron beam in a manner similar to the focusing of light rays by a glass lens. Also see ELECTROSTATIC LENS, ELECTROMAGNETIC LENS, and WAVEGUIDE LENS.

electron magnetic moment The energy per unit flux density available in an electron. Approximately equal to 9.2848×10^{-24} joule per tesla.

electron mass See MASS OF ELECTRON AT REST.

electron microscope A microscope in which the source of illumination is an electron beam focused by electromagnetic lenses. It allows much greater magnification than is possible with optical microscopes.

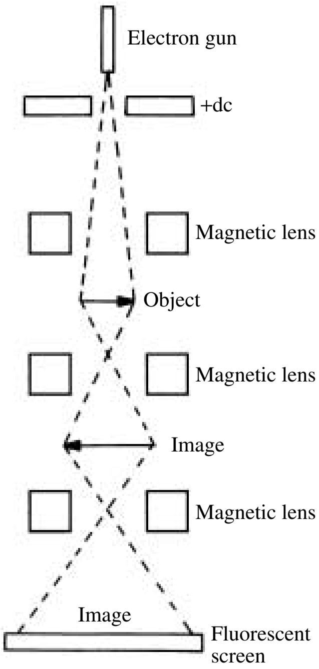

electron microscope

electron mirror A reflector of electrons—especially a dynode element in a photomultiplier tube or electron-multiplier tube.

electron motion **1.** The movement of electrons in a conductor, semiconductor, or space, as the result of electric or magnetic attraction or repulsion. **2.** The movement of an electron as a charged mass. In an electric field, this movement simulates that of a free-falling body in a gravitational field.

electron multiplication **1.** In a gas discharge, the production of additional electrons as a result of collisions between electrons, atoms, and molecules. **2.** The increased production of electrons in a semiconductor when avalanche occurs.

electron-multiplier tube **1.** A vacuum tube utilizing a sequence of secondary emissions for increased current amplification. Electrons from the cathode strike a positively biased dynode with a force that dislodges secondary electrons, which, upon joining those first emitted, are reflected to a second positive dynode that contributes more secondary electrons, reflecting the total to a third positive dynode, etc. The last dynode in the chain reflects the enhanced beam to an anode collector that passes the high current to an external circuit. **2.** See PHOTOMULTIPLIER TUBE.

electronography Printing by means of the electrostatic transfer of ink from a printing plate across a gap to an impression cylinder.

electron optics See ELECTRO-OPTICS.

electron orbits See ELECTRON SHELLS.

electron oscillator A device in which oscillation is obtained by causing electrons to move in an oscillatory path, to travel in bunches, etc. Examples: klystron, magnetron, and traveling-wave tube.

electron pair Two electrons from adjacent atoms, which sometimes share the same orbits, but always produce a bond between two adjacent atoms.

electron-pair bond The bond between an electron pair.

electron physics The physics of electronics, usually from a highly theoretical viewpoint.

electron-proton magnetic moment ratio A physical constant whose value is approximately equal to 658.211, derived from the division of the magnetic moment of the electron by that of the proton.

electron recoil The recoil of an electron from a photon it has collided with.

electron rest mass See MASS OF ELECTRON AT REST.

electron scanning Deflection of an electron beam. See, for example, ELECTROSTATIC DEFLECTION and ELECTROMAGNETIC DEFLECTION.

electron shells The spheres, concentric with the nucleus of an atom, that represent the median distances from the nucleus around which electrons migrate.

electron spin The rotation of an electron (i.e., around its axis). This motion is independent of the electron's movement around the nucleus of an atom.

electron stream **1.** The beam of electrons generated by the electron gun in a cathode-ray tube.

2. The electrons moving between the cathode and plate in an electron tube.

electron-stream instrument See ELECTRON-BEAM INSTRUMENT.

electron-stream meter An oscilloscope (or cathode-ray tube alone) used as a device for making measurements.

electron-stream transmission efficiency The ratio of the current through a positive electrode to the current impinging on it. In a tube, for example, some electrons are absorbed by the plate, and others are reflected.

electron telescope A telescope using a combination of a glass lens, photocathode, and electrostatic focusing. Light from the object is focused on the photocathode by the lens, the electrons emitted being focused electrostatically on a phosphorescent viewing screen.

electron transit time The time required for an electron to travel a given distance. For a vacuum tube, the upper frequency limit of operation is governed by the time required for an electron to reach the plate after leaving the cathode. Transit time is usually stated in fractions of a microsecond.

electron tube An evacuated or gas-filled chamber in which electrons are emitted (usually by a hot cathode) and controlled (usually by a voltage applied to a grid electrode).

electron unit See ELEMENTARY CHARGE.

electron velocity The velocity acquired by an electron that moves between two points having a given potential difference. Also see ELECTRON MOTION, **2**.

electronvolt Abbreviation, eV. The energy acquired by a unit charge moving through a potential difference of one volt; it is equal to approximately 1.6022×10^{-19} joule.

electron-wave tube A tube, such as a klystron or traveling-wave tube, in which electrons traveling at different velocities interact with each other, modulating the electron stream.

electro-oculogram A recording of the voltage that is found between the anterior and posterior parts of the eyeball.

electro-optical transistor A phototransistor or pair of phototransistors in a single package. Used in electronic circuits to sense changes in light levels.

electro-optical valve See KERR CELL.

electro-optic radar A form of radar that makes use of visual apparatus for locating a target.

electro-optics The branch of electronics dealing with related electrical and optical phenomena: photoelectricity, light generation, laser technology, light amplification, etc. It is also concerned with electronic phenomena that are analogous to optical phenomena, such as electronic focusing, reflection, refraction, diffraction, etc. Also called *electron optics* and *optoelectronics*.

electro-osmosis Causing liquids to flow by applying an electric field across the walls of a porous plug. The force exerted by the field on ions in the liquid causes it to flow.

electropad The skin-contacting electrode of an electrocardiograph.

electropathy See ELECTROTHERAPY.

electrophilic Pertaining to the tendency to seek electrons.

electrophobia An irrational fear of electricity, a psychological condition sometimes exhibited by victims of serious electric shock.

electrophonic effect Sound heard by a person when an alternating current is passed through some part of the body.

electrophoresis The movement of dielectric particles through a liquid in which they are suspended, produced by the electric field between electrodes immersed in the suspension.

electrophoresis equipment 1. Any device intended for the purpose of depositing a dielectric material onto a metal by means of electrophoresis. **2.** Any equipment in which electrophoresis occurs.

electrophoresis scanner A device that senses the movement of charged particles caused by electrophoresis effects.

electrophoretic deposition A type of deposition in which a low-voltage direct current passing through a colloidal suspension of dielectric polymer particles deposits them as a coating on a metallic body (the anode in the process). It can provide a better coating than one obtained with spray painting or dipping.

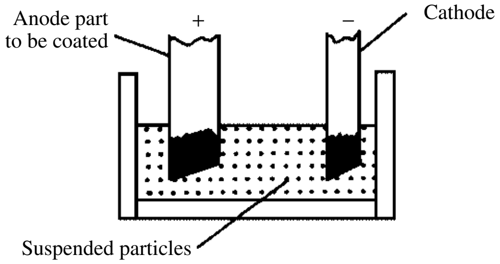

electrophoretic deposition

electrophorus A simple device used to demonstrate electrostatic generation and induction. It consists of a smooth metal plate at the end of an insulating handle and an accompanying cake of resin or hard rubber. The cake is rubbed with cat's fur, making it negatively charged. The metal plate is touched to the charged cake; by induction, it acquires a bound positive charge on the face that touched the cake and a free negative charge on the opposite face. When the plate is lifted and its top face touched momentarily with the finger or grounded, the negative charge leaks off, often with a sharp spark, but the positive charge remains.

electrophotographic process See XEROGRAPHY.

electrophotography The production of photographs by means of electricity. See XEROGRAPHY.

electrophotometer A light-intensity meter using a photoelectric sensor and a meter, but usually not incorporating an amplifier. Compare ELECTRONIC PHOTOMETER.

electrophrenic respiration A system of inducing respiration in which one or both of the phrenic nerves (i.e., of the diaphragm) are stimulated electrically to produce contractions of the diaphragm muscles.

electrophysiology **1.** The study of electrical processes in the human body. **2.** The study of how electrical impulses affect, and are produced by, body organs.

electroplaques In electric fish, small voltage-generating cells that are connected in series-parallel networks.

electroplate **1.** To cause one metallic substance to adhere to the surface of another through the effects of electrolysis. **2.** A metal plating deposited via electrolysis.

electroplating Depositing one metal on the surface of another by electrolytic action.

electropolar Having electrical polarity (either positive or negative).

electropolishing An electrolytic method of smoothing a rough metal surface. The workpiece to be polished becomes the anode of an electrolytic cell in which electrolytic action dissolves tiny surface irregularities.

electropositive Having positive electrification or polarity. Compare ELECTRONEGATIVE.

electropotential series See ELECTROMOTIVE SERIES.

electropsychrometer An electronic instrument for humidity measurements.

electroreduction In electrolysis, reduction of the cathode electrode.

electrorefining The refining of metals by means of electrolysis.

electroretinograph An instrument used to measure the electrical response of the human retina to light.

electroretinography The process of detecting and measuring electrical impulses from the retina.

electroscope An instrument for detecting electric charges and fields. The common type uses a pair of gold-leaf strips hung from the end of a metal rod in a glass tube or jar. When the exposed end of the rod is brought near a charged object, the leaves repel each other and spread apart.

electrosection The use of an arc-generating device for making surgical incisions.

electrosensitive recording See ELECTRIC RECORDING.

electroshock **1.** The system of creating a controlled electric shock in the brain as a treatment for certain mental disorders. **2.** The electric shock used in the therapy described in **1**.

electrospinograph An instrument that senses and records electrical impulses from the spinal cord.

electrostatic Pertaining to stationary electric charges and fields, and their application.

electrostatic actuator A device for measuring the sensitivity of a microphone. Electrostatic charges produce forces on the diaphragm of the microphone, and the resulting output is recorded.

electrostatic amplifier See DIELECTRIC AMPLIFIER.

electrostatic charge See ELECTRIC CHARGE.

electrostatic component The electric component of an electromagnetic wave. It is perpendicular to the magnetic component and can be thought of as the wave's voltage component.

electrostatic constant See ELECTRIC CONSTANT.

electrostatic convergence See ELECTROSTATIC FOCUSING.

electrostatic cooling Accelerated cooling of a body through the application of an intense electrostatic field. The body must be in a free convection state, a corona discharge must be present, and the field must not be uniform.

electrostatic copier A document-copying apparatus that uses electrostatic effects to reproduce printed material.

electrostatic coupling See CAPACITANCE COUPLING.

electrostatic deflection In a cathode-ray tube, deflection of the electron beam by the electrostatic fields between pairs of internal horizontal and vertical deflecting plates. It is primarily used in laboratory oscilloscopes. Compare ELECTROMAGNETIC DEFLECTION.

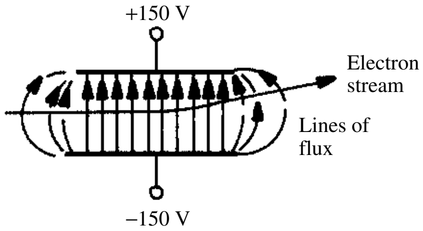

electrostatic deflection

electrostatic electrometer See KELVIN ABSOLUTE ELECTROMETER.

electrostatic electrophotography See XEROGRAPHY.

electrostatic energy The potential energy contained in an electrostatic field (e.g., the energy in a charged capacitor). Compare ELECTROKINETIC ENERGY.

electrostatic field A stationary electric field.

electrostatic field intensity An expression of the strength of an electrostatic field. Usually expressed in volts per meter, millivolts per meter, or microvolts per meter.

electrostatic flux The flux existing around an electric charge or a charged body.

electrostatic focusing In a cathode-ray tube, electron-beam focusing achieved by varying the dc bias voltage on a focusing electrode. Compare ELECTROMAGNETIC FOCUSING.

electrostatic galvanometer A galvanometer operating on the principle of the electrostatic voltmeter.

electrostatic generator A device for producing high-voltage electric charges; e.g., a *Van de Graaff generator.*

electrostatic headphone A device similar to an electrostatic speaker, but held against the head for private listening. Incoming audio signals cause attraction and repulsion among charged plates, resulting in acoustic vibration.

electrostatic hysteresis The tendency of some dielectrics (especially ferroelectric materials) to saturate and retain a portion of their polarization when an alternating electric field to which they are exposed reverses polarity. This causes the charge to lag behind the charging force.

electrostatic induction The charge acquired by a body inserted into an electric field. Compare ELECTROMAGNETIC INDUCTION.

electrostatic instrument See ELECTROSTATIC VOLTMETER.

electrostatic lens An assembly of deflecting plates or cylinders, whose electric field causes an electron beam to converge or diverge in much the same way as a visible light beam passing through an optical lens. Compare ELECTROMAGNETIC LENS.

electrostatic loudspeaker See ELECTROSTATIC SPEAKER.

electrostatic memory A memory unit in which an information bit is stored as an electric charge.

electrostatic memory tube A cathode-ray tube in which information bits are stored in capacitive cells swept by the scanning electron beam.

electrostatic microphone See CAPACITOR MICROPHONE.

electrostatic phase shifter See PHASE-SHIFTING CAPACITOR.

electrostatic potential In an electric field, the potential energy represented by the voltage between the two elements creating the field, or between any two points within the field.

electrostatic precipitator See DUST PRECIPITATOR.

electrostatic printer A computer output peripheral in which the printing medium, a fine dust, is fused by heat onto paper that has been charged according to the data being represented.

electrostatic process **1.** Any process that uses electrostatic action. **2.** A method of photography in which visual images are converted to electrostatic images.

electrostatic recording A method of recording that employs a signal-controlled electric field.

electrostatic relay A high-input-impedance relay consisting of two polarity-controlled contacts; opposite charges on the contacts close the relay, and like charges open it.

electrostatics The branch of electricity concerned with electrical charges at rest. Compare ELECTRODYNAMICS and ELECTROKINETICS.

electrostatic screen A shield against electric flux consisting of a number of straight, narrowly separated rods or wires joined at only one end. The shield has little effect on magnetic flux. Also called *Faraday shield.*

electrostatic separator A device for separating fine particles from a mixture by exposing the mixture to an intense electrostatic field.

electrostatic series A list of materials arranged in this sequence: any one of them becomes positively electrified when rubbed with another lower in the list, or negatively electrified when rubbed with another higher in the list. Compare ELECTROMOTIVE SERIES.

electrostatic shield Any metallic enclosure designed to confine an electric field.

electrostatic speaker A loudspeaker whose vibrating diaphragm is one of two plates in a large air-dielectric capacitor, the other being a closely situated metal plate (or plug). An audio voltage applied to the plates causes them to vibrate. Also called *capacitive loudspeaker* and *capacitor loudspeaker.*

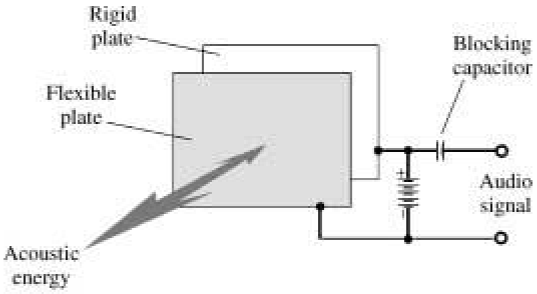

electrostatic speaker

electrostatic sprayer An equipment for spray painting in which fine droplets of paint are attracted by an electrostatic field to the surface to be coated.

electrostatic storage See ELECTROSTATIC MEMORY.

electrostatic stress **1.** Stress in the vicinity of a charged body or particle. **2.** See DIELECTRIC STRESS.

electrostatic transducer See CAPACITANCE TRANSDUCER.

electrostatic tube of flux The space between electric lines of flux going through adjacent points on the boundary of a given area in an electric field.

electrostatic tweeter A small electrostatic speaker for reproducing high-frequency sounds. Compare WOOFER.

electrostatic unit Abbreviation, esu. A unit of measure in the electrostatic system of cgs units. Also see CENTIMETER-GRAM-SECOND. Compare ELECTROMAGNETIC UNIT.

electrostatic vector See ELECTRIC VECTOR.

electrostatic voltmeter An indicating meter whose movement consists of a stationary metal plate near a rotating metal plate. A voltage applied to the plates charges them, and the attraction between them causes the movable member to rotate against the torque of a returning spring over an arc proportional to the voltage.

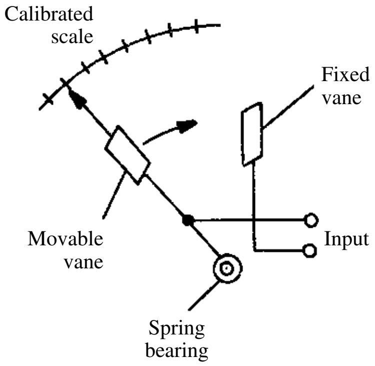

electrostatic voltmeter

electrostatography See XEROGRAPHY.

electrosteel See ELECTRIC STEEL.

electrostenolysis The deposition of certain metals from a solution in capillary tubes when an electric current passes through the solution.

electrostimulation Electrical excitation of nerves for the relief of pain.

electrostriction In certain materials, the physical contraction that occurs when a voltage is applied. Compare MAGNETOSTRICTION.

electrostrictive ceramic A ceramic exhibiting ELECTROSTRICTION when voltage is applied.

electrostrictive relay A relay in which the movable contact is carried by a bar of electrostrictive material, such as barium titanate. A control voltage deforms the material, causing the contacts to close.

electrosurgery Surgery, sometimes bloodless, achieved with diathermy-like equipment. See DIATHERMY, **2**.

electrosynthesis Chemical synthesis produced by means of electric currents or fields.

electrotape Also called *electronic tape measure.* Any device that measures distance by electronic means, such as radar or sonar.

electrotechnology See ELECTRICAL TECHNOLOGY and ELECTRONICS TECHNOLOGY.

electrotellurograph An instrument for measuring ground currents.

electrotherapeutics See ELECTROTHERAPY.

electrotherapy The treatment of disorders or diseases by electrically induced heat—especially by DIATHERMY.

electrothermal **1.** Pertaining to electrically generated heat. **2.** Pertaining to a combination of electricity and heat.

electrothermal device A device whose operation depends on the heat generated by an electric current (e.g., a bolometer, hot-wire ammeter, thermocouple, or varistor).

electrothermal expansion element A thermostatic element, such as a bimetallic strip, whose expansion is used in heat-sensitive switches.

electrothermal instrument A hot-wire or thermocouple-type meter.

electrothermal recorder See ELECTRIC RECORDING, **1** and THERMAL RECORDER.

electrothermic See ELECTROTHERMAL.

electrothermics The study and application of the heating effects of electricity in conductors and junctions.

electrotitration In chemistry, the completion of titration, as indicated by an electrical measurement, such as of the resistance of the solution being titrated.

electrotonic Pertaining to ELECTROTONUS.

electrotonus Modification of a nerve's sensitivity by passing a constant current through it.

electro-ultrafiltration In physical chemistry, filtering a colloidal suspension by electro-osmosis.

electrovalence **1.** The number of charges acquired by an atom gaining electrons. **2.** The number of charges forfeited by an atom losing electrons. **3.** Valence resulting from electron transfer between atoms and the resulting creation of ions.

electrovalency See ELECTROVALENCE.

electrovalent bond See IONIC BOND.

electrowin To recover (win) a metal from a solution of its salts by means of electrolysis.

electrum A natural alloy of gold and silver.

element **1.** See ELECTRODE. **2.** A circuit component intended for a specific purpose. **3.** A specific part of an antenna array (e.g., *driven element* or *parasitic element*). **4.** A fundamental, unique substance whose atoms are of only one kind (examples: aluminum, carbon, silicon, and sulfur). There are more than 100 elements, some man-made. Elements combine to form compounds. **5.** A circuit, such as an AND gate, that can be taken as a unit because it performs a special function. **6.** In digital computer operations, a subunit of a category that cannot be further categorized [e.g., a *bit* (word element) or a *record* (file element)].

elemental area In a facsimile or television picture, a scanning line segment as long as the line's width.

elemental charge See ELEMENTARY CHARGE.

elemental semiconductor A semiconductor containing one undoped chemical element.

elementary charge Symbol, *e.* Also called *unit electric charge.* The electric charge of a single electron or proton. This charge is approximately equal to 1.6022×10^{-19} coulomb.

elementary particle **1.** A minute charged or uncharged particle within the atom (i.e., electron, proton, neutron, quark, etc.). **2.** In theory, a subatomic particle that cannot be broken down into smaller particles.

element error rate In communications or data transfer, the ratio n_i/n_t, where n_i is the number of elements received incorrectly and n_t is the number of elements transmitted.

element spacing **1.** The spacing between radiator, director, and reflector elements in a directional antenna. **2.** The spacing between the internal electrodes of a vacuum tube.

elevation Angular position (in degrees) of a point above the horizontal.

elevation-position indicator A type of radar display simultaneously indicating the elevation of, and the line-of-sight distance to, the target.

elevator control **1.** An electronic system for automatically stopping an elevator and opening the doors. Various safety functions are included, an example being the reopening of a closing door when a passenger steps into the car. **2.** In an aircraft, the mechanical, electronic, or electromechanical devices or circuits involved in actuation of the elevators.

ELF Abbreviation for EXTREMELY LOW FREQUENCY.

eliminator **1.** A device or circuit acting as a surrogate for an inconvenient or undesirable component (e.g., battery eliminator). **2.** A device for removing or minimizing an undesirable signal or quantity (e.g., harmonic eliminator); interference eliminator.

ell A coaxial fitting that is a right-angle line section with a coaxial connector at each end. It takes its name from its L shape.

ellipse A geometric figure having the Cartesian-plane formula $(x - x_0)^2/a^2 + (y - y_0)^2/b^2 = 1$, where a and b are constants, and x_o and y_0 represent the center point.

elliptical filter See ELLIPTIC FILTER.

elliptical function See ELLIPTIC FUNCTION.

elliptical load line For any amplifier with an output transformer, a load line in the shape of an ellipse, obtained when the load connected to the output element is reactive, rather than purely resistive.

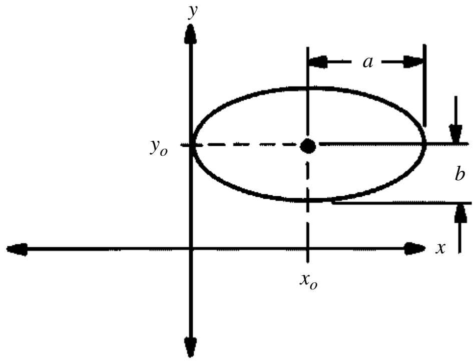

ellipse

elliptically polarized wave An electromagnetic wave in which the rotation of the electric-intensity vector at one point describes an ellipse.

elliptical orbit A satellite orbit that is not a perfect circle. In theory, all satellites deviate slightly from perfectly circular orbits. Sometimes a satellite is deliberately put into an orbit that is greatly elongated. The closer the satellite is to the earth, the faster it moves.

elliptical polarization Polarization characterized by elliptical rotation of the wave vector at a given point.

elliptical stylus In a phonograph (turntable) system, a stylus with a characteristic ellipsoidal shape.

elliptic filter A band-pass, band-stop, high-pass, or low-pass inductance-capacitance (LC) filter, designed according to an ELLIPTIC FUNCTION. Characterized by a steep attentuation-versus-frequency cutoff response with ripple in both the passband and the stopband.

elliptic function A function, similar to the Chebyshev and Butterworth functions, used in the design of certain selective filters. The elliptic function results in a better filter magnitude response than the Chebyshev or Butterworth functions in some applications. See ELLIPTIC FILTER.

elongation A form of modulation distortion resulting from multipath propagation. Some of the paths result in greater propagation delay than other paths; this causes the modulation envelope to spread out. The higher the modulating frequency, the greater the effect.

ELSE A word used in a BASIC computer program that provides an instruction based on a relational test and, in this respect, is related to IF-THEN, ON-GOTO, etc. It specifies the operation to be done if the conditions given in the same program line don't occur.

ELSIE Abbreviation of *electronic letter-sorting and indicator equipment.*

EM **1.** Abbreviation of EFFICIENCY MODULATION. **2.** Abbreviation of ELECTROMAGNETIC(S). **3.** Abbreviation of *electromagnetic iron.* **4.** Abbreviation of ELECTROMAGNETIZER. **5.** Abbreviation of ELECTRON MICROSCOPE. **6.** Abbreviation of EXPOSURE METER. **7.** Abbreviation of *electromotive.*

$\mathbf{E_m}$ **1.** Symbol for MAXIMUM VOLTAGE. **2.** Symbol for MAXIMUM JUNCTION FIELD.

emanation **1.** Emission of electrons. **2.** Emission of radioactive particles or ionizing radiation. **3.** Emission of electromagnetic energy.

$\mathbf{E_{max}}$ Symbol for MAXIMUM VOLTAGE.

embedded path A means of guiding a mobile robot along a specific route. One common scheme uses a buried, current-carrying wire that produces a magnetic field. The robot can sense and follow this field. Colored paints and tapes can also be used in conjunction with machine vision systems. Compare EDGE DETECTION.

embedded training The inclusion of training/tutorial programs in computerized equipment that assist users in the operation of the equipment.

embedding See ENCAPSULATION.

embossed-foil printed circuit A printed circuit made by pressing the pattern from metal foil into the insulating substrate and then removing the surplus foil.

embossed-groove recording **1.** A phonograph record into which grooves are embossed, rather than scribed. **2.** Recording sound by embossing grooves on record disks.

embossing stylus The rounded-tip stylus used to make an embossed-groove recording.

EMC Abbreviation of ELECTROMAGNETIC COMPATIBILITY.

EME Abbreviation of *earth-moon-earth.* See MOONBOUNCE.

$\mathbf{e/m_e}$ The ratio of the elementary electron charge to its mass: 1.7588×10^{11} coulombs per kilogram. Also see CHARGE-MASS RATIO.

Emergency Broadcast System In the United States, a general plan for dissemanating information via broadcast stations in the event of a national emergency.

emergency channel A communication channel allocated for emergency service.

emergency communication Radio or other electronic transmission and reception of urgent messages (e.g., distress signals, storm warnings, etc.).

emergency equipment **1.** Apparatus kept in standby status for immediate operation when regularly used equipment fails. **2.** Equipment, especially vehicular, for use in emergency situations. Examples are ambulances, fire-fighting trucks and equipment, etc.

emergency power supply An alternating-current (ac) or direct-current (dc) power unit kept in standby status for immediate use when the regular power supply fails.

emergency service A communications service devoted exclusively to emergency communication.

EMF, emf Abbreviation of ELECTROMOTIVE FORCE.

emf standard See STANDARD CELL.

EMG **1.** Abbreviation of ELECTROMYOGRAM. **2.** Abbreviation of ELECTROMYOGRAPH.

EMI Abbreviation of ELECTROMAGNETIC INTERFERENCE.

E microscope See ELECTRON MICROSCOPE.

emission **1.** The ejection of particles, especially electrons, from a material. **2.** Waves radiated from any source (as from a transmitting antenna or from an amplifier stage). **3.** The emanation of radiant, electromagnetic, acoustical, electrical, or magnetic energy.

emission code A system of abbreviating the various types of radio emission. See EMISSION MODE.

emission frequency **1.** In communications, the carrier frequency of the transmitted signal as it is radiated from the antenna or fed into a transmission line. **2.** The actual frequency or frequency range of a signal as it is transmitted or radiated. This might be the carrier frequency. **3.** The frequency of energy in an emission band or bands in a spectrum.

emission lines In a spectrum, radiation intensity peaks that appear as bright lines in a visible display. In a radio-frequency spectrum, the emission lines occur as sharp peaks in radiated energy at specific wavelengths.

emission mode Any of various official classifications of radio communication emission types. Emissions are designated according to the modulation method used (e.g., continuous waves, amplitude modulation, single-sideband with suppressed carrier, frequency modulation, pulse modulation, etc.).

emission power **1.** The rate at which energy is radiated from an object. **2.** In radio communication, the transmitter output power.

emission spectrum The radiation spectrum of a substance that emits energy (e.g., the light spectrum of an incandescent metal).

emission types See EMISSION MODE.

emission velocity The initial velocity of an electron as it leaves an emitting surface.

emission wavelength **1.** In communications, the carrier wavelength of the transmitted signal as it is radiated from the antenna or fed into a transmission line. **2.** The actual wavelength or wavelength range of a signal as it is transmitted or radiated. This might or might not be the carrier wavelength. **3.** The wavelength of energy in an emission band or bands in a spectrum.

emissive power The rate at which a surface emits energy of all wavelengths in all directions, per unit area of radiating surface, regardless of temperature.

emissivity For a radiating source, the ratio W_1/W_2, where W_1 is the energy emitted by the source at a particular temperature, and W_2 is the energy emitted by a blackbody (i.e., a theoretically perfect radiator) at the same temperature.

emittance For an energy-radiating source, the radiated power per unit area of radiating surface.

emitted electron An electron that has left an atom of a material and has escaped into surrounding space or entered a neighboring material.

emitter **1.** A body that discharges particles or waves (see EMISSION). **2.** In a semiconductor device, the area, region, or element from which current carriers are injected into the device. In a transistor symbol, the emitter is that electrode shown with an arrowhead.

emitter-base junction In a bipolar transistor, the boundary between base and emitter regions.

emitter bias Emitter current or voltage maintained to set the operating point of a bipolar transistor.

emitter bulk resistance The portion of the resistance of the semiconductor material in a transistor that affects emitter resistance.

emitter-coupled logic A bipolar form of digital logic, abbreviated ECL.

emitter-coupled multivibrator A two-transistor multivibrator circuit in which the emitters share a common resistor.

emitter-coupled phase inverter A transistor phase inverter in which the out-of-phase component is taken from the collector and the in-phase component from the emitter resistor (of the same transistor). Another transistor is often used to amplify the in-phase component so that both outputs are equal in magnitude.

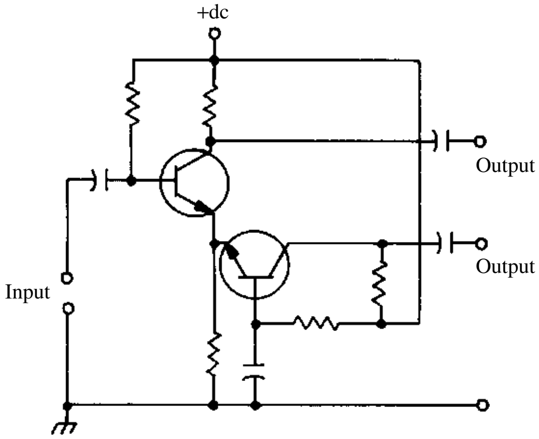

emitter-coupled phase inverter

emitter current Symbol, I_E. The current in the emitter electrode of a bipolar transistor.

emitter degeneration In a transistor amplifier, current degeneration obtained by use of an unbypassed emitter resistor. The arrangement results in virtually distortion-free amplification at a sacrifice in voltage gain.

emitter follower A transistor circuit in which the input signal is applied to the base, and the output signal is taken from the emitter resistor. Gain is always less than unity; output impedance is low.

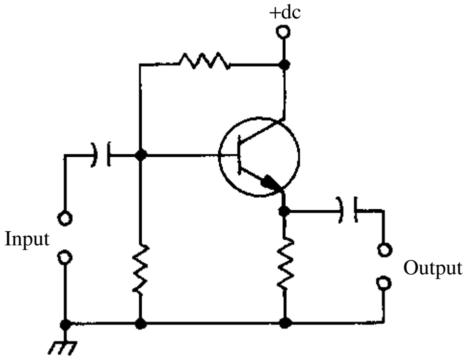

emitter follower

emitter-input circuit See COMMON-BASE CIRCUIT.

emitter junction See EMITTER-BASE JUNCTION.

emitter resistance Symbol, R_E. **1.** The resistance of the emitter electrode in a bipolar transistor. **2.** External resistance connected to a transistor's emitter terminal.

emitter stabilization In a common-emitter transistor stage, an emitter resistor that stabilizes the circuit against temperature variations.

emitter-to-base junction See EMITTER-BASE JUNCTION.

emitter voltage Symbol, V_E. The voltage at the emitter electrode of a bipolar transistor.

EMP, emp **1.** Abbreviation of ELECTROMAGNETIC PULSE. **2.** Abbreviation of *electromagnetic power.*

emphasis Modification of the amplitude-versus-frequency output or response of an audio circuit, for the purpose of optimizing signal intelligibility.

emphasizer An audio-frequency device with a specially tailored response, intended to maximize intelligibility of a voice.

Empire cloth Varnished cambric used as an insulating sheet or tape.

empirical Observable; derived from experimentation.

empirical curve A curve plotted from data acquired from observations, tests, and calculations, rather than from mathematical laws or other theory.

empirical design The design of electronic circuits by cut-and-try methods and, to some extent, through intuition arising from experience (i.e., practical as opposed to theoretical design).

empirical probability Probability estimated from experience and observations. This method is often used in quality-control and reliability procedures.

empty medium A computer storage medium, such as a magnetic tape or disk, that is ready to accept data (i.e., rather than being completely blank, it contains the signals necessary for processing the to-be-added data).

EMU, emu Abbreviation of ELECTROMAGNETIC UNIT(S).

emulator In computer engineering, a sophisticated device that substitutes for a similar device or stage in the computer, and thereby provides a basis for experimenting and troubleshooting without disturbing the equivalent part of the computer.

E_n Symbol for *voltage remaining at null.*

enable To initiate the operation of a circuit or device by applying a pulse or trigger signal.

enable pulse **1.** A pulse that initiates the operation of a circuit or device. **2.** A binary pulse that augments a write pulse to make a magnetic core change state.

enabling gate A digital device that regulates the length of a pulse for specialized use.

enameled wire Wire that is insulated by a thin coat of baked enamel. Commonly used in coil winding because the thin enamel allows for a maximum number of turns in a given volume for a given wire gauge.

encapsulant A material, such as potting resin, used to embed (encapsulate) a component, circuit, or device.

encapsulated circuit A component, circuit, or device embedded in plastic or wax (see ENCAPSULATION).

encapsulated component An electronic part that is embedded in plastic or wax (see ENCAPSULATION).

encapsulating material See ENCAPSULANT.

encapsulation The embedding of a circuit or component in a solid mass of plastic or wax. The mold or container remains as part of the assembly after the plastic or wax has solidified. Protects against the environment, and/or against the effects of physical vibration. Compare POTTING.

encephalogram See ELECTROENCEPHALOGRAM.

encephalograph See ELECTROENCEPHALOGRAPH.

enciphered facsimile Facsimile communications that have been rearranged or scrambled at the transmitting location so that it cannot be intercepted by a third party. A deciphering device is needed at the receiver end of the circuit.

enclosure **1.** A cabinet, case, or other housing for electronic equipment, such as a receiver, transmitter, or test instrument. **2.** A specially designed housing for a loudspeaker.

encode **1.** To convert signals or data into a desired (usually digital) form. Also called CODE. **2.** To equip a transmitter with a tone-producing device (encoder). **3.** To develop and apply an encoding system to a group of transceivers or transmitters of a communications network.

encoder **1.** An analog-to-digital or digital-to-analog converter. **2.** An electromechanical device for translating the angular position of a rotating shaft into a corresponding series of digital pulses. Also see SHAFT-ANGLE ENCODER. **3.** A device for encoding data (see ENCODE). **4.** A machine with a keyboard for printing characters that can be read by optical character recognition (OCR) equipment. **5.** A tone generator used as a receiver enabler in the transmitters of a communications network.

encoding **1.** The translation, either by a machine or by a human operator, of a spoken or written language into digital code. **2.** Any function performed by an ENCODER.

encryption The conversion of a signal from plain text, graphics, or other commonly recognizable form into a cipher. See also CIPHER. Compare DECRYPTION.

end-around carry In a computer, a carry produced in the most significant position, causing a carry into the least-significant position.

end-around shift In digital-computer operations, the transfer of characters from one end of a register to the other end. Also called LOGICAL SHIFT.

end bell **1.** The part of a motor housing that supports the bearing and protects internal rotating parts. **2.** A clamping part fastened to the back of a plug or receptacle. **3.** Either of the two frames of a transformer that contains the mounting lugs.

end bracket See END BELL, **2**.

end cell A cell intended for series operation in conjunction with a storage battery. As the voltage of the battery drops, the end cell can be added into the circuit.

end effect **1.** In a tapped coil, losses because of induced currents flowing in the inductance and distributed capacitance of the unused end of the coil. **2.** EDGE EFFECT in a capacitor. **3.** An effective capacitance at the ends of an antenna, resulting from air discharge. This lowers the resonant frequency slightly below that predicted by theory. The effect is exaggerated by the proximity of objects, such as trees and buildings, or when an antenna is placed close to the earth.

end effector The device or tool connected to the end of a robot arm (e.g., a gripper, screwdriver, drill, or soldering iron).

end-fed antenna An antenna whose lead-in or feeders are attached to an end of the radiator.

end feed A method of feeding electromagnetic fields to an antenna by connecting the transmission line to the end. Ordinarily, the antenna must be a multiple of 0.5 wavelength long for end feed to be effective.

end-fire antenna See END-FIRE ARRAY.

end-fire array Also called *end-fire antenna.* A phased antenna in which the greatest radiation/response takes place off one or both ends. The array consists of two or more parallel driven elements, all of which lie in a single plane. A typical system might consist of two half-wave dipoles, fed 90 degrees out of phase and spaced one-quarter wavelength apart in free space. This produces a unidirectional directivity pattern. Two elements might be driven in phase and spaced 1 wavelength apart, producing a bidirectional pattern. These systems show some power gain, in their favored directions, compared to a single half-wave dipole. The larger the number of elements, with optimum phasing and spacing, the greater the gain.

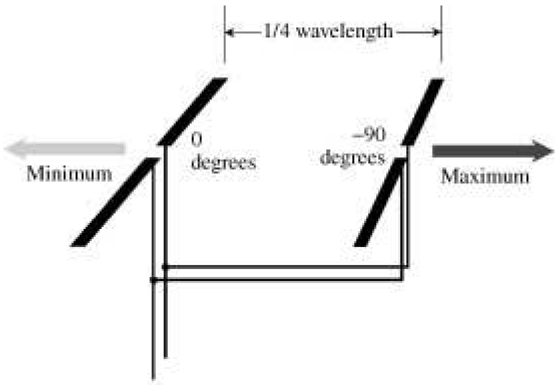

end-fire array

end-fire directivity In a directive antenna, beaming a signal along the plane of the antenna (i.e., off its ends).

end instrument A device capable of converting intelligence into electrical signals or vice versa, and that needs to be connected to only one terminal of a loop.

end item A final, completed product or component.

endless loop Also called *infinite loop.* A computer programming bug that causes the machine to go in an indefinite, and often useless, logical circle. For example, suppose that at line 180, the computer encounters the command GOTO 250, meaning “Go to line 250,” but line 250 gives the command GOTO 180. Once the computer gets to line 180, it enters a loop in which nothing is accomplished, and from which the only escape is intervention by the operator (e.g., terminating the program).

end mark In digital-computer operations, a signal or code indicating the close of an information unit.

endodyne reception See ZERO-BEAT RECEPTION.

end-of-charge voltage For a rechargeable cell or battery, the voltage at full charge (i.e., just after disconnection of the charging apparatus and before use).

end-of-data mark A code or character signaling that all the data in a computer storage medium has been read or used.

end-of-discharge voltage For a rechargeable cell or battery, the voltage at the termination of a discharging cycle, immediately before the unit is taken out of use and the charging apparatus is connected.

end-of-field mark In computer operations, a “flag” code that signals when the end of a field has been reached.

end-of-file mark In computer operations, a code instruction that signals when the last record in a file has been read.

end-of-line unit The last device or circuit in a chain.

end-of-message character A character or code signaling the end of a message.

end-of-run The end of a computer program or program run, as indicated by the program.

end-of-tape mark A physical marker at the end of a magnetic tape (e.g., something that can be sensed by methods other than that used to read the tape).

end-of-tape routine A computer program that handles the processing needed after the last record on a reel of magnetic tape has been reached.

end-on armature A relay armature that moves in the direction of the core’s axis.

end-on directional antenna See END-FIRE ANTENNA and END-FIRE ARRAY.

endoradiograph An X-ray picture, derived or enhanced by the introduction of substances into the body.

endoradiosonde A tiny pill-enclosed transducer and radio transmitter for sensing physiological conditions in the stomach and intestines; it transmits corresponding signals to instruments outside.

endothermic reaction A chemical reaction producing cold (i.e., one in which kinetic energy is lost). Compare EXOTHERMIC REACTION.

end-plate magnetron A magnetron whose oscillation intensity is increased by a positive and a negative end plate, the electric field between them causes the electrons to move axially while spinning.

end point **1.** For a precision potentiometer, the shaft position between the last and first positions of measurement. **2.** The point at which the useful life of a device can be considered spent. **3.** The point at which a time interval or operational sequence ends. **4.** The end-point voltage of a primary or Edison storage cell. **5.** For a

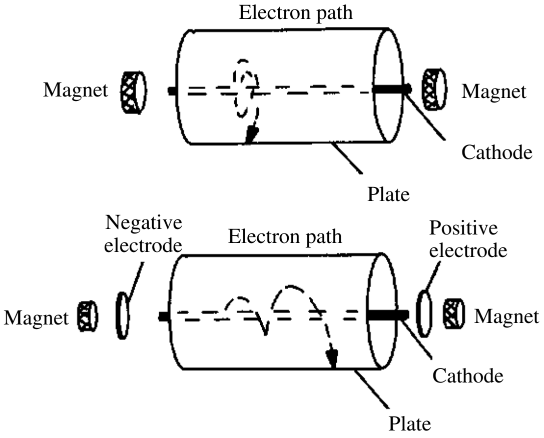

end-plate magnetron

lead-acid storage cell, the specific-gravity value of the electrolyte at which the cell is considered in need of recharging (nominally 1.150 to 1.175).

end-point control A form of quality control in which the end item is checked for defects.

end-point sensitivity A means of expressing the sensitivity of a meter or other indicating device: the ratio, in decibels, between the input signal required to produce a full-scale or maximum reading and the smallest detectable input signal.

end-point voltage The voltage of a battery or cell terminal when the device is no longer useful.

end resistance In a rheostat or potentiometer, the resistance between the wiper and the end terminal when the wiper is set to the end point of the device.

end-resistance offset In a potentiometer, the resistance between the wiper and an end terminal when they are in contact.

end-scale deflection See END-SCALE VALUE.

end-scale value For an indicating meter, the electrical quantity indicated at the last graduation on the scale.

end section Either the input or output section of a multisection filter.

end setting **1.** The fully clockwise or fully counterclockwise setting of a rotatable control. **2.** The minimum or maximum setting of a control.

end shield In a magnetron, a shield that confines the space charge to the interaction space.

end spaces The cavities at either end of the anode block in a multicavity magnetron tube; they terminate all the anode-block cavity resonators.

end use The intended application of a circuit or device.

energize To apply operating power and input signals to a circuit or device.

energized The condition of a circuit or device that is powered or excited.

energy Symbol, W. Common units: joule, watt-hour, and kilowatt-hour. **1.** The capacity for doing work. Some common forms of energy are electrical, mechanical, and chemical. Also see CONSERVATION OF ENERGY, KINETIC ENERGY, and POTENTIAL ENERGY. **2.** The work performed by electric power. The unit used by utility companies is the kilowatt-hour (kWh), equal to the product *Pt*, where *P* is power in kilowatts and *t* is the period (hours) during which the power is used.

energy-band diagram A diagram depicting the various energy levels within the atom of a conductor, semiconductor, or insulator.

energy barrier The natural potential gradient across a semiconductor junction. In the absence of an applied voltage, the gradient, not measurable from the outside, prevents total interaction between the n- and p-type materials.

energy cell **1.** A usually small primary or secondary cell—especially the kind used in hearing aids and electronic watches. **2.** A capacitive-type direct-current (dc) source (see ENERGY-STORAGE DEVICE, **2**).

energy consumption **1.** The conversion of energy from one form to another by a component, circuit, system, or machine, in the process of performing some useful task. **2.** The amount of energy involved in the process defined in **1**.

energy conversion The transformation of energy from one form to another. See also CONSERVATION OF ENERGY and ENERGY TRANSFORMATION.

energy-conversion device A component, circuit, system, or machine that changes energy from one form to another. See also CONSERVATION OF ENERGY.

energy density **1.** For an energy-producing cell, such as an electrochemical cell, the ratio of available energy to cell mass. It is expressed in joules per gram or in watt-hours per kilogram. **2.** For an energy-producing cell, the ratio of available energy to cell volume. Expressed in joules per cubic centimeter or in watt-hours per cubic centimeter.

energy gap In the energy-level diagram for a semiconductor or insulator, the region between valence and conduction bands representing the minimum energy required to make the electron pass from the valence to the conduction band (i.e., to become a current carrier). Also called *forbidden energy band.*

energy level A constant-energy state, such as one of the energy levels of an electron in an atom.

energy-level diagram **1.** A diagram showing the energy levels (in electronvolts) of electrons in the various shells of an atom. **2.** A diagram showing variations in power that correspond to variations in current in a channel.

energy loss In any system, the energy that is unavoidably lost (i.e., it is not converted into useful work). Also see ENTROPY and POWER LOSS.

energy of a charge The energy level of an electrostatic charge. It is $QV/2$ ergs, where Q is the quantity of electricity in coulombs, and V is the potential in volts.

energy product An expression of the effectiveness of a permanent magnet. The magnetic flux density is multiplied by the magnetic field strength to obtain the energy product, specified in gauss-oersteds.

energy redistribution A mathematical process for determining the effective duration of an irregular pulse. The instantaneous power output of the irregular pulse is integrated from the start to the end of the pulse. Then, a rectangular pulse is constructed having the same peak power and the same total energy content (area under the power curve). The length of this rectangular pulse is considered to be the effective duration of the irregular pulse.

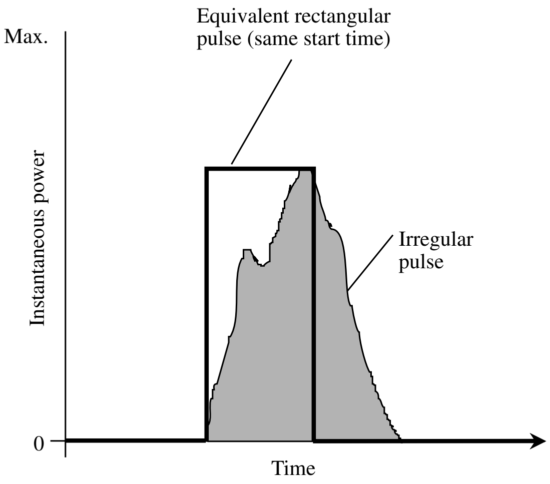

energy redistribution

energy state The condition of an electron, as expressed by its position and velocity, with respect to the position and velocity of other electrons.

energy-storage capacitor A usually high-value capacitor used primarily to store the charge used to fire a lamp (as in a photoflash unit), create a spark discharge (as in electronic ignition), or perform some similar function.

energy-storage device **1.** See CAPACITOR. **2.** A small, electrochemical component offering very high capacitance (e.g., several farads) and low leakage current (less than 1 pA). It has a number of applications, including long-interval timing, power-supply filtering, and energy-cell service. Its active ingredients are compressed powders.

energy stored in capacitor The electrical energy in the field between the plates of a charged capacitor. In this instance, energy $W = CE^2/2$, where W is the energy in joules, C is the capacitance in farads, and E is the voltage in volts.

energy stored in inductor The magnetic energy in the field surrounding an inductor carrying current. In this instance, energy $W = LI^2/2$, where W is the energy in joules, L is the inductance in henrys, and I is the current in amperes.

energy transformation The conversion of one form of energy into another, as with a transducer.

engine analyzer An instrument for checking the performance of an automobile engine. In addition to measuring voltage and resistance throughout a car's electrical system, the instrument measures engine speed, cam dwell angle, and other factors.

engineer **1.** A person who designs machines, circuits, and other devices. **2.** A person who develops methods of utilizing machines, circuits, or other devices more efficiently, or for new applications. **3.** To design or implement an apparatus.

engineering The science of applying scientific laws to technical problems and designing practical devices. Also see ELECTRICAL ENGINEER and ELECTRONICS ENGINEER.

enhanced-carrier demodulation A method of reducing distortion in the demodulation of amplitude-modulated (AM) signals. A properly phased and synchronized local carrier is added to the signal in the demodulator.

enhancement mode Operation characteristic of an ENHANCEMENT-TYPE MOSFET.

enhancement-type MOSFET A metal-oxide semiconductor field-effect transistor (MOSFET) in which the channel directly under the gate electrode is widened (enhanced) by a negative gate voltage in the n-channel unit or by a positive gate voltage in the p-channel unit. Compare DEPLETION-TYPE MOSFET.

ENIAC An electronic computer developed at the University of Pennsylvania. The name is an acronym for *Electronic Numerical Integrator And Calculator.*

ENIC Abbreviation of *voltage negative-impedance converter.*

enrichment In a mixture of different isotopes of a given element, the increase in the relative concentration of one particular isotope.

ensemble **1.** A collection of devices that functions together as a complete unit. **2.** In music recording, the ability of all the musicians to hear each other during the session. **3.** A set of random mathematical functions, all starting at the same point.

ENSI Abbreviation of EQUIVALENT-NOISE-SIDEBAND INPUT.

entladungsstrahlen Ultraviolet radiation emitted by electric arcs. At atmospheric pressure, the wavelength is approximately 40 to 90 nanometers, depending on the arc length. The term is derived from the German word for *discharge rays.*

entrainment Providing a path for gases to escape from an electrochemical cell or battery.

entrance delay In security applications, a delay that allows authorized people time to leave the protected area after activating the alarm system, or to deactivate the system after entering the protected area. The delay is approximately 30 to 45 seconds.

entropy **1.** In all closed physical systems, the measure of energy wasted. According to the second law of thermodynamics, for example, supplied heat can never be converted entirely into work. **2.** In communications, the amount of information in a message, defined as the base-10 logarithm of the number of equivalent messages that can exist. **3.** A natural process in which the energy in the universe tends to become more uniformly distributed with the passage of time.

entropy coding A form of digital encoding that minimizes redundancy, thereby increasing the amount of data in a given amount of memory or storage space.

entry **1.** A unit of computer input or output information. **2.** A data item in a table or list. **3.** A computer source program statement. **4.** In a computer program, the address of the first instruction.

entry condition A condition that must be specified before a computer program is run (e.g., establishing operand values).

entry-level system **1.** The least-sophisticated computer that will perform the things that a user requires. **2.** A simple electronic or computer system (e.g., an amateur radio transciever or personal computer, intended for ease of operation, and from which the user expects to upgrade to a more powerful system at a later date).

entry point In a computer program, the first instruction to be implemented, or a point during the run when data can be entered.

envelope **1.** On a graph, the imaginary line joining successive signal peaks. In the graph for an amplitude-modulated signal, the line reproduces the modulating wave. **2.** The enclosure of a transistor or integrated circuit. **3.** The glass shell of a vacuum tube.

envelope delay In a tuned amplifier, time delay introduced in the envelope of a modulated signal by varying the phase of the envelope with the modulating frequency. This delay varies directly with the amount by which the sidebands shift, with respect to the carrier frequency.

enveloped file A computer file with labels permitting it to be handled by a computer of a type different from that used to make the file.

environmental conditions See ENVIRONMENTAL FACTORS.

environmental factors Aspects of the space immediately surrounding and sometimes influencing electronic equipment. Examples: altitude, dust, light, moisture, noise, pressure, shock, temperature, and vibration.

environmentally sealed Sealed against the effects of adverse environmental factors.

environmental test chamber See CLIMATE CHAMBER.

E_o **1.** Symbol for OUTPUT VOLTAGE. **2.** Symbol for *zero reference voltage.*

EOF Abbreviation of *end of file.*

EOL Abbreviation of *end of line.*

EOLM Abbreviation of *electro-optical light modulator.*

EOR Abbreviation of END OF (program) RUN.

EOS Abbreviation of *electro-optical system(s).*

EOT Abbreviation of *end of tape.*

EOTS Abbreviation of *electro-optical tracking system.*

E_P **1.** Symbol for PLATE VOLTAGE. **2.** Symbol for PEAK VOLTAGE.

EP Abbreviation for EXTENDED PLAY.

ephemeris time Time measured with respect to the orbit of the earth around the sun. Initiated in the year 1900 AD.

epipolar navigation A scheme for position sensing and navigation that uses an artificially intelligent vision system. Allows calculation of position and velocity, based on changes in the visualized direction, size, and shape of an object whose actual location, size, and shape are precisely known. It is used in some mobile robots.

episcotister A mechanical light beam modulator. The device consists of a series of rotating disks having transparent and opaque sections that alternately interrupt and pass the light beam at an audio-frequency rate.

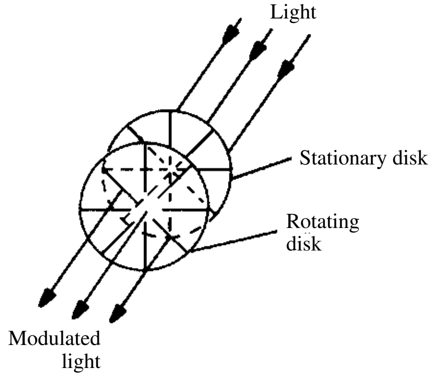

episcotister

epitaxial Pertaining to, or having the property of, EPITAXY.

epitaxial deposition The tendency of certain materials to grow on a semiconductor substrate under certain conditions.

epitaxial device A semiconductor device built by means of EPITAXIAL GROWTH.

epitaxial film A film of single-crystal semiconductor material deposited onto a single-crystal semiconductor substrate.

epitaxial growth Growing monocrystalline silicon on a silicon wafer by precipitating silicon from a gas in which the wafer is placed. Epitaxy is secured between the precipitate and the wafer.

epitaxial layer A semiconductor layer exhibiting epitaxy. Also see EPITAXIAL GROWTH.

epitaxial mesa transistor See DOUBLE-DIFFUSED EPITAXIAL MESA TRANSISTOR.

epitaxial planar transistor A planar transistor having an epitaxially grown collector on a low-resistivity substrate, and a diffused base and emitter.

epitaxial process See EPITAXIAL GROWTH PROCESS.

epitaxial transistor A transistor in which an epitaxial layer (into which a base region later is diffused and an emitter region alloyed) is grown on the face of a semiconductor wafer, which serves as the collector. Also see DOUBLE-DIFFUSED EPITAXIAL MESA TRANSISTOR.

epitaxy The condition in which atoms in a thin film of single-crystal semiconductor material grown on the surface of the same kind of wafer continue their characteristic alignment. Also see EPITAXIAL GROWTH.

E plane The plane of an antenna containing the electric field.

E plane bend See E BEND.

E-plane tee junction A waveguide junction whose structure changes in the plane of the electric field.

epoxy resin A synthetic resin used to encapsulate electronic equipment, or as a cement. Epoxy resins are based on ethylene oxide or its derivatives.

EPROM Abbreviation of *erasable programmable read-only memory.*

EPU **1.** Abbreviation of *electronic power unit.* **2.** Abbreviation of *emergency power unit.*

Eq **1.** Abbreviation of *equation.* **2.** Abbreviation of EQUALIZER or EQUALIZATION.

equal alternations Positive and negative half-cycles of a wave that have identical shape and amplitude.

equal-energy source A light source that has a constant emission rate (energy per unit wavelength).

equal-energy white The color of light emitted by a source radiating equally the wavelengths of the visible-light spectrum.

equal heterodyne In a beat-frequency system, the condition in which the outputs of the two heterodyning oscillators are identical.

equality circuit A logic circuit that, when two numbers are put into it, outputs logic 1 if the numbers are equal, and logic 0 if the numbers are not equal.

equalization **1.** The use of an EQUALIZER to make the frequency response of a line, amplifier, or other device uniform over a given frequency range. **2.** The use of an EQUALIZER to modify the frequency response of a line, amplifier, or other device.

equalizer A circuit or device, such as a compensated attenuator, that allows the user to tailor the frequency response of a line, amplifier, or other device. Sometimes used in sophisticated high-fidelity stereo amplifier systems, to obtain a desired bass/midrange/treble frequency output.

equalizer circuit breaker A form of circuit breaker that trips in the event of unbalance in an electrical system.

equalizing current A current that flows in the circuit of two compound generators connected in parallel.

equalizing network A circuit used to equalize a line.

equalizing pulses In a television signal waveform, several pulses (preceding and following the vertical sync pulse and having a repetition rate of twice the power-line frequency) that start the vertical retrace at the correct instant for good interlace.

equal-loudness curves See AUDIBILITY CURVES.

equal vectors Vectors having the same magnitude and the same direction. They do not necessarily originate at the same point. Compare IDENTICAL VECTORS.

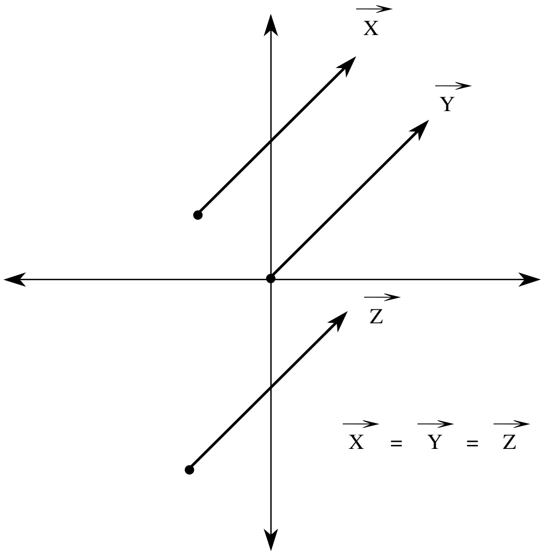

equal vectors

equation solver A (usually analog) computer for solving linear simultaneous equations or for determining the roots of polynomials.

equatorial orbit A satellite orbit that lies in the plane of the earth's equator.

equiphase surface Any surface in a wave, over which the field vectors at a particular instant have either 0° or 180° phase difference.

equiphase zone The space region in which two radionavigation signals show no phase difference.

equipment **1.** Collectively, apparatus or components designated for a specific purpose (e.g., *radio equipment*). **2.** A functional electronic unit, such as a test instrument, receiver, or memory unit.

equipment chain A system consisting of series-connected circuits or devices.

equipment ground An electrical ground connection intended to reduce the chances of electric shock. An equipment ground does not necessarily constitute a good radio-frequency ground; it serves only to eliminate potential differences among the individual units in a system.

equipment life The period during which electronic equipment functions according to specifications; it is terminated at an END POINT.

equipment test A usually preliminary, qualifying test of electronic equipment.

equipotential Having a potential difference of zero; being at the same voltage level.

equipotential line Between two charged plates, the locus (an imaginary line) of points having the same potential, with respect to the plates.

equipotential surface A surface on which all points have the same electrical potential.

equisignal Pertaining to signals having equal intensity.

equisignal localizer See TONE LOCALIZER.

equisignal radio-range beacon For aircraft guidance, a radio-range beacon that transmits two distinct signals that are received by aircraft with equal intensity only in certain directions.

equisignal surface The "surface" around a transmitting antenna formed by points of equal field intensity.

equisignal zone The region in which two radionavigation signals have identical amplitude.

equivalence The condition existing when one network can be substituted for another without disturbing currents, impedances, and voltages at the terminals.

equivalent absorbing power See EQUIVALENT STOPPING POWER.

equivalent absorption Unit, sabin. The rate at which a surface absorbs sound energy.

equivalent binary digits For a given decimal number or specific character, the corresponding binary digits (bits).

equivalent bit rate The number of binary digits (bits) that can be sent in a given unit of time, such as one second, in a digital communications system.

equivalent capacitance The value of a single lumped capacitance, that would cause the same action as the capacitance distributed throughout a circuit.

equivalent circuit A circuit that has the same overall current, impedance, phase, and voltage relationships as a more-complicated counterpart that it usually replaces for analysis.

equivalent component density For a circuit in which discrete components are not used or are not evident, the volume of that circuit divided by the number of discrete components that would be required if the circuit used them.

equivalent conductivity The conductivity of a solution that contains 1 gram equivalent of the solute in the space between electrodes 1 centimeter apart.

equivalent dark-input current For a photoelectric device, the illumination required for an output current equal to the DARK CURRENT of the device.

equivalent decrement The value of decrement in a damped wave that would result in the same amount of interference at a receiver as the interference caused by the sidebands of an amplitude-modulated signal.

equivalent delay line A comparatively simple network, such as a resistance-capacitance (RC) circuit, that will provide the attenuation and phase characteristics of an ideal delay line.

equivalent delta In a three-phase system, a delta-connected circuit that is equivalent to a given wye-connected circuit, from the standpoint of impedance and phase. Also see DELTA CONNECTION and WYE CONNECTION. Compare EQUIVALENT WYE and WYE-EQUIVALENT CIRCUIT.

equivalent differential input capacitance For a differential amplifier, the equivalent input capacitance (see EQUIVALENT CAPACITANCE) at one input (inverting or noninverting) when the opposite input is grounded.

equivalent differential input impedance For a differential amplifier, the equivalent input impedance at one input (inverting or noninverting) when the other input is grounded.

equivalent differential input resistance For a differential amplifier, the equivalent input resistance at one input (inverting or noninverting) when the other input is grounded.

equivalent equations Two equations for an unknown that have the same root.

equivalent four-wire system A two-wire line over which full-duplex operation is obtained by use of frequency division.

equivalent height See VIRTUAL HEIGHT.

equivalent impedance **1.** The value of a single lumped impedance that would cause the same action as the impedance distributed throughout a circuit. **2.** An impedance that draws current of the same strength and phase as that drawn by an impedance it replaces.

equivalent inductance The value of a single lumped inductance that would cause the same action as the inductance distributed throughout a circuit.

equivalent input offset current For a differential amplifier, the difference between currents flowing into the inverting and noninverting inputs when the output voltage is zero.

equivalent input offset voltage For a differential amplifier, the input voltage required to reduce the output voltage to zero.

equivalent input wideband noise voltage For a differential amplifier, the ratio V_o/G_v, where V_o is the root-mean-square (rms) output-noise voltage, and G_v is the direct-current (dc) voltage gain.

equivalent length of antenna **1.** The electrical length of an antenna, as measured in degrees or wavelengths. **2.** The free-space length of an antenna. **3.** The length d (in feet) of a quarter-wave resonant antenna at a specific frequency f (in megahertz), given by the formula $d = 234/f$. **4.** The length d (in feet) of a half-wave resonant antenna at a specific frequency f (in megahertz), given by the formula $d = 468/f$.

equivalent length of electric dipole The distance, measured in a straight line, separating the points that represent the charge centers of an electric dipole.

equivalent length of feed line The electrical length of a feed line as measured in degrees or wavelengths. Generally, this is equal to $1/v$ times the line length in free-space wavelengths, where v is the VELOCITY FACTOR of the line, expressed as a fraction between 0 and 1.

equivalent length of magnet The distance separating the poles of a magnet. In a bar magnet, these poles are not exactly at the ends. The actual equivalent length is about 83% of the length of the bar magnet.

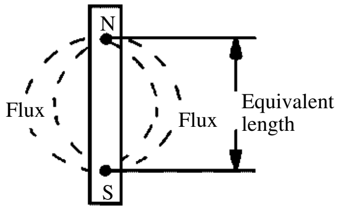

equivalent length of magnet

equivalent loudness The actual intensity, in decibels, of a given sound whose apparent loudness changes with frequency.

equivalent network A network that can replace a more-complex network for analysis purposes.

equivalent noise input The value of modulated luminous flux that, when applied to a photoelectric device, produces a root-mean-square (rms) output current equal to the device's rms noise current.

equivalent-noise-sideband input Abbreviation, **ENSI.** A specification for receiver noise characteristics. Numerically, $ENSI = 0.3E_s(P_n/P_s)^{1/2}$, where E_s is the voltage of an unmodulated radio-frequency (RF) carrier applied to the receiver, P_n is the resulting noise-output power of the receiver (measured with an rms meter), and P_s is the noise-output power measured with the RF signal 30% amplitude modulated at 400 Hz with a 400-Hz bandpass filter inserted between the receiver output terminals and the meter.

equivalent noise temperature For a component having resistance, the temperature (degrees absolute) at which a theoretically perfect resistor having the resistance of the component would generate the same noise the component generates at room temperature.

equivalent optics The analogy between certain optical lenses and prisms and the electrostatic deflection of an electron beam. Thus, when the upper deflecting plate in an electrostatic deflection system is made negative and the lower plate is positive, the beam is deflected downward, like horizontal light rays bent by a prism. When both plates are made equally negative, the beam converges to a point, as light rays do when they pass through a double convex lens. When both plates are made equally positive, the beam spreads out, as do light rays passing through a double concave lens.

equivalent permeability The permeability of a component made of certain materials, compared with that of a component having the same reluctance, shape, and size, but made of different materials.

equivalent reactance The value of a single lumped reactance that would cause the same action as the reactance distributed throughout a circuit.

equivalent resistance The value of a single lumped resistance that would cause the same action as the resistance distributed throughout a circuit.

equivalent series and parallel circuits Series and parallel circuits in which current, voltage, phase, and frequency relationships are identical. Any series circuit can be transformed into an equivalent parallel circuit.

equivalent series resistance The equivalent resistance acting in series with circuit components.

equivalent sine wave A sine wave of the same frequency and effective voltage as a given wave.

equivalent stopping power For a material in the path of radioactive particles, the thickness of the material that produces the same energy loss as that produced by one centimeter of air.

equivalent time The effective duration of some phenomenon, such as a pulse.

equivalent volt See ELECTRONVOLT.

equivalent wye In a three-phase system, a wye-connected circuit that is equivalent to a given delta-connected circuit from a standpoint of impedance and phase. Also see DELTA

CONNECTION, EQUIVALENT DELTA, WYE CONNECTION, and WYE-EQUIVALENT CIRCUIT.

equivalent Y See EQUIVALENT WYE.

equivocation A condition in which the meaning of data depends on certain parameters.

ER Abbreviation of ECHO RANGING.

Er Symbol for ERBIUM.

E_r Symbol for voltage drop across a resistance.

erasable storage In computer operations, any storage medium holding information that can be erased.

erasable PROM A programmable read-only memory (PROM) from which the data can be removed, usually by exposure to ultraviolet light. Also see PROM.

erase To obliterate or remove a signal, especially a recorded one, as in the erasure of recorded material from a magnetic tape or the data from a computer disk.

erase button A pushbutton that actuates the circuit supplying a signal that erases stored material (as the display on a storage oscilloscope).

erase current In an electromagnetic erase head, the current flowing through the coil of the head. In most instances, it is a high-frequency current (usually the regular bias current), but it can be as low as 60 Hz, as long as the speaker does not respond to what remains of it on the tape after erasure.

erase head In a tape recorder, a head used to erase recorded material from tape. It can contain a permanent magnet (see ERASE MAGNET) or an electromagnet whose coil carries erase current.

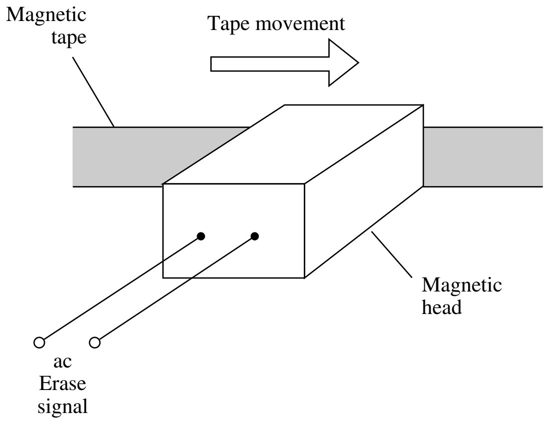

erase head

erase magnet In a tape recorder, a magnet used to erase recorded material from tape. Because the strength of the magnet is greater than that of the magnetized areas on the tape, erasure is complete (the tape left demagnetized).

erase oscillator In a tape recorder, a high-frequency (typically 30 to 80 kHz) oscillator that supplies erase current.

eraser See BULK ERASER.

erase signal A signal that causes recorded material to be erased (see ERASE and ERASE CURRENT).

erasing speed The rate at which successive storage elements are erased, as in a charge-storage tube.

erasure **1.** In tape-recording and digital-computer operations, the process of erasing a recorded signal (see ERASE). **2.** An erasure accomplished by the process described in 1, above.

erbium Symbol, Er. A metallic element of the rare-earth group. Atomic number, 68. Atomic weight, 167.26.

E region See E LAYER.

e register In a computer, a register used in double-precision calculations.

E_r-E_y signal In color television, the resultant signal that is the difference between the original full-red and E_y signals.

ERG Abbreviation of ELECTRORETINOGRAPHY.

erg Abbreviation, *e*. A unit of work. It is the work done by a force of one dyne (10^{-5} newton) acting through a distance of one centimeter.

ergograph An instrument used to measure and record work done by muscles.

ergometer An instrument for measuring energy consumed or work accomplished.

ergon See ERG.

E_{rms} Symbol for ROOT-MEAN-SQUARE VOLTAGE.

ERP Abbreviation of EFFECTIVE RADIATED POWER.

error **1.** In calculations and measurements, the difference between a true value and an observed or calculated value. **2.** In electronic circuits, especially those of automatic control systems, the difference between a required (or reference signal) level and the actual signal level. **3.** In communications, a discrepancy between the transmitted data and the received data.

error accumulation The adding-up of maximum possible error when measurements are repeatedly made. Generally, the maximum plus-or-minus error per measurement is multiplied by the number of measurements.

error amplifier An amplifier for boosting error current or voltage.

error-checking code An error-correcting or error-detecting code.

error-correcting code An error-detecting code that, in addition to the function indicated by its name, indicates the correct code.

error-correcting telegraph A digital communications system in which an improbable or incorrect character is not accepted. In the event that such a character is received, the receiver instructs the transmitter to send that character again.

error correction **1.** The restoration of mutilated, corrupted, or missing data in a digital system, especially in magnetic data storage media, such as tapes and disks. **2.** In digital communications, any scheme in which the receiver (destination) automatically eliminates (to the greatest possible extent) errors in data from a transmitter (source). For example, the destination can instruct the source to repeat questionable characters or words.

error-correction routine In computer operations, a series of programmed instructions to detect and correct errors in files. A common example is a spell-checking program for word-processed document files.

error current An error signal that is a feedback current for automatically correcting a system.

error curve A bell-shaped curve that describes the distribution of errors in measurement around a true value.

error-detecting code In computer operations, a character-coding system that ensures that an impossible combination (forbidden characters) will be generated by an error (for error detection).

error-detecting routine A computer program that detects errors by checking the validity of data.

error detection and feedback In computer operations, a system in which an error (sensed by an error-detecting code) automatically generates a request to repeat the suspect signal.

error detector A sensor that responds to an error signal by delivering a signal proportional to the error.

error diagnostics As performed by a compiler, detecting and indicating the presence of errors in source language statements.

error interrupt A computer program halt caused by a software or hardware error and accompanied by a display of what has happened.

error list As produced by a compiler, a list of source language statement faults.

error message During a computer program run, a statement (displayed on a peripheral) of what is in error.

error of measurement The positive or negative difference between the value of an actual measurement and the true (or most probable) value.

error range For a data item, the range of values over which it will cause an error.

error rate In data transmission, the ratio of errors transmitted to the data transmitted.

error-rate damping Damping that involves adding to an error signal another signal that is proportional in rate of change.

error ratio **1.** In a received message, the number of incorrect characters divided by the total number of characters. Can be represented as a fraction between 0 and 1 or as a percentage by multiplying the fraction by 100. **2.** A measure of distortion for digital signal communications. The number of inaccurately received bits divided by the total number of received bits.

error routine A computer program segment that is input when an error is detected so that an appropriate action is taken (correct the error, repeat the process, etc.).

error-sensing circuit A circuit that samples the output current or voltage of a power supply, amplifier, or control system, compares this output with a standard value, and delivers a feedback (correction) signal whose amplitude is proportional to the difference (error).

error signal In a servo system, an output signal whose value is proportional to the difference between the actual operating quantity of the system and a standard reference quantity. The signal is fed back to the input of the system for automatic correction.

error tape In data processing, a record tape designed and used for storing errors for subsequent study.

error voltage An error signal that is a feedback voltage for automatically correcting a system.

Es Symbol for EINSTEINIUM.

Esaki diode See TUNNEL DIODE.

E scope See E DISPLAY.

escape character In computer operations, a character indicating that the next character belongs in a new group.

escapement A (usually oscillating) mechanical or electromechanical device that stores energy (often in a spiral spring) on one swing, and returns that energy on the next swing. Such a mechanism advances a shaft progressively in a clock or watch, and in some control equipment.

escape velocity **1.** The minimum velocity (about 25,000 miles per hour or seven miles per second) required for a space vehicle to completely escape the gravitational field of the earth. **2.** The minimum velocity required for a space vehicle to completely escape the gravitational field of a planet or star. **3.** The minimum velocity required for an electron to escape the electrical influence of an atomic nucleus.

escutcheon A usually decorative plate that frames an opening or covers a panel in a piece of equipment (e.g., the escutcheon of a radio tuning dial).

ESD Abbreviation of ENERGY-STORAGE DEVICE.

ESG Abbreviation of *electronic sweep generator.*

Esnault-Pelterie formula A formula for approximately calculating the inductance of a single-layer solenoidal coil:

$$L = 0.1008(a^2n^2)/(s + 0.92a)$$

where L is the coil inductance in microhenrys, a is the radius of the coil in inches, s is the length of the coil in inches, and n is the number of turns in the winding. The formula is accurate to 0.1 percent for all values of $2a/s$ between 0.2 and 1.5.

ESS Abbreviation of *electronic switching system.*

EST Abbreviation of EASTERN STANDARD TIME.

esu Abbreviation of ELECTROSTATIC UNIT(S).

ET Abbreviation of EPHEMERIS TIME.

ETC Abbreviation of *electronic temperature control.*

etchant Any substance such as cupric chloride, ferrous chloride, or hydrochloric acid, used in etching.

etched circuit A circuit produced by etching the metallic coating of a substrate to provide the required pattern of conductors and terminals to which discrete components are soldered.

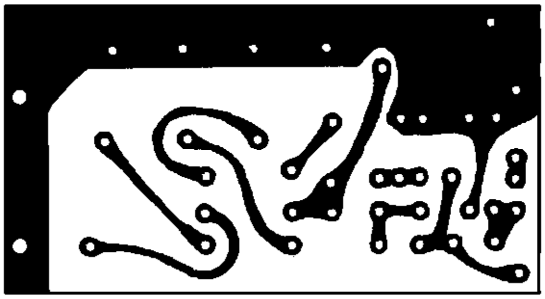

etched circuit

etch factor The ratio of the depth to the width of an etched track in an etched circuit.

etching **1.** Chemically eating away a metal to form a desired pattern, such as an etched circuit. **2.** Thinning a quartz-crystal plate by slowly eroding one or both of its faces with hydrofluoric acid to fine-tune the resonant frequency.

ET-cut crystal A piezoelectric plate cut from a quartz crystal at an angle of +66°, with respect to the z-axis. Also see CRYSTAL AXES and CRYSTAL CUTS.

ether **1.** Also called *luminiferous ether.* A nonviscous fluid once thought to fill space, convey waves (radio, light, etc.), and sustain fields. **2.** A volatile liquid occasionally used in electronics as a solvent [e.g., ethyl oxide $(C_2H_5)_2O$].

ether drift The postulated motion between a material body and the ether (see ETHER, **1**). The concept was checked by Michelson and Morley, who failed to find that the earth moves relative to the ether. This eventually led to scientific rejection of the so-called *ether theory* of the propagation of light.

ethical slave A machine, especially a smart robot, that is treated in the manner of a slave, based on the notion that a machine cannot have "feelings." Some researchers fear that the use of ethical slaves could lead to technological nightmares. For example, robots might be used as soldiers in a marauding offensive army; the commanders could rationalize that there is nothing immoral about the war because there is no loss of life on their side.

E transformer A differential transformer whose primary is wound on the center leg of an E core, the secondaries being wound on the outer legs.

Ettinghausen effect A phenomenon somewhat like the HALL EFFECT. It occurs when a metal strip, carrying current longitudinally, is placed into a magnetic field perpendicular to the plane of the strip: corresponding points on opposite edges of the strip exhibit different temperatures.

Eu Symbol for EUROPIUM.

eudiometer **1.** An instrument for measuring the amount of oxygen in the air. **2.** An instrument for analyzing gases.

eureka **1.** See CONSTANTAN. **2.** The ground transponder beacon in the British rebecca-eureka radar navigational system (see REBECCA-EUREKA SYSTEM).

europium Symbol, Eu. An element of the rare-earth group. Atomic number, 63. Atomic weight, 151.96.

eutectic **1.** A form of reaction in which mixed liquids solidify when cooled. **2.** The solid substance resulting from a reaction as defined in **1**.

eutectic alloy A metallic alloy with a specific melting point, made via eutectic process.

eutectic bond A connection between two dissimilar metals, facilitated by a third metal alloyed, via eutectic process, to the adjoining faces.

eV Symbol for ELECTRONVOLT.

evacuation The removal of air or other gases from a tube or chamber, specifically, the envelope of a vacuum tube that houses the internal elements.

evaporation **1.** A technique for electrically depositing a film of a selected metal on a metallic or nonmetallic surface. A filament of the metal to be deposited is heated by an electric current in a vacuum chamber, which makes filament particles travel to the (nearby) object to be coated, where they condense as a film. In an alternate method, a piece of the metal to be deposited is laid on or wrapped around a filament of some other metal. **2.** Electron emission by a hot cathode.

evaporation theory The theory that electrons will acquire sufficient escape velocity to leave a material when the energy acquired by (or imparted to) the electron exceeds the work function of the material. Also see WORK FUNCTION.

E vector The vector that represents the electric component of an electromagnetic wave.

even-even nucleus An atomic nucleus containing an even number of protons and an even number of neutrons. An example is the alpha particle, or helium nucleus, which contains two protons and two neutrons.

even harmonic In a complex waveform, an even-numbered multiple of the fundamental frequency. Compare ODD HARMONIC.

even line In a television picture, an even-numbered member of the 262.5 horizontal lines scanned by the spot in developing the even-line field. Compare ODD LINE.

even-line field On a television screen, the complete field obtained when the spot has traced all the even-numbered lines. Compare ODD-LINE FIELD.

even parity check A check to verify the presence of an even number of ones or zeros in a group of bits.

event An occurrence that affects the state of a computer file.

event counter Any device that measures the number of specified events taking place within a certain interval of time.

evolution Extracting a root of a number (e.g., square root, cube root, etc.).

E wave In microwave operations, the transverse magnetic (TM) wave. Also see WAVEGUIDE MODES.

EWR Abbreviation of EARLY-WARNING RADAR.

EWS Abbreviation of *early-warning system.*

E_X **1.** Symbol for voltage drop across a reactance. **2.** Symbol for EXCITATION ENERGY.

exa- Symbol, E. A prefix meaning 10^{18} (International System of Units).

exalted-carrier reception In radio reception, overcoming the effects of selective fading by maintaining the carrier at a high amplitude. This is accomplished before demodulation by removing the carrier from an amplitude-modulated or phase-modulated signal, amplifying it, and reinserting it at a higher amplitude with the sidebands.

exc **1.** Abbreviation of EXCITER. **2.** Abbreviation of EXCITATION.

except gate A logic gate that delivers an output pulse when an input pulse is present at one or more of a set of input terminals, and absent from one or more of another set of input terminals. Also called *exclusive-OR element.*

excess charge The amount of overcharge for a storage battery.

excess conduction In a semiconductor, current conduction by excess electrons.

excess electron **1.** An electron that, when introduced into an atom, results in a negative ion. **2.** An electron resulting from the addition of a donor impurity to a semiconductor substance.

excess meter A meter that integrates the amount of power in excess of some predetermined level.

excess minority carriers The number of minority carriers in excess of the normal equilibrium number in a semiconductor material.

excess modified index of refraction Symbol, M. For waves transmitted through a refracting medium, a modified index of refraction greater than unity.

excess noise Electrical noise caused by current in a semiconductor material.

excess sound pressure Unit, dyne/cm^2. In a medium conducting sound waves, the quantity $P_i - P_s$, where P_i is total instantaneous pressure at a given point in the medium, and P_s is static pressure in the absence of the sound waves.

excess-three code A computer code derived from binary notation by adding binary three (i.e., 0011) to each four-bit group. Thus, decimal seven is 1010 in the code (it is 0111 in binary). Unlike the binary representation for zero, the excess-three representation (0011) contains two ones, a feature that distinguishes actual zero from a machine fault.

exchange **1.** To reverse the contents of two memory banks. For example, if the memory banks are called A and B, an exchange is the placing of the contents of memory A into memory B, and the placing of the contents of memory B into memory A. The original contents are removed. **2.** A two-way sequence of data transmissions. **3.** A designated location in a telephone circuit.

exchange line A telephone line.

exciplex In a laser, a method of adjusting the color by means of chemical reactions in organic dyes.

excitant The electrolyte in a voltaic cell.

excitation **1.** Supplying input-signal driving current, driving power, or driving voltage. **2.** Input-signal driving current, driving power, or driving voltage.

excitation anode In a mercury-pool tube, an auxiliary anode whose operation maintains the cathode spot when no output current is being drawn from the tube.

excitation current **1.** Input-electrode current in an excited transistor amplifier. **2.** Grid current in an excited vacuum-tube amplifier. **3.** Current flowing in the circuit of the excitation anode of a mercury-pool tube. **4.** Current flowing in the exciter circuit of an alternator. **5.** Shunt-field current in a motor.

excitation energy **1.** Symbol, E_X. In artificial transmutation, the energy of a nucleus when protons of less than maximum energy have been emitted from the atom. **2.** Electrical energy required by a transducer.

excitation purity In color television, complete saturation of a hue (i.e., there is no contamination by other colors, and the saturated hue is distributed uniformly).

excitation voltage **1.** The signal voltage that achieves, or is required for, excitation (see EXCITATION, **1**). **2.** The value of driving voltage.

excitator An electrical discharger.

excited atom An atom in which one or more electrons have been pushed out of their normal orbits into higher ones by energy applied from the outside.

excited-field speaker A dynamic speaker in which the magnetic field is provided, not by a permanent magnet, but by direct current flowing through a large coil of wire wound around the speaker core. The coil usually acts as a filter choke in the power supply of the attendant amplifier or receiver.

excited state In artificial transmutation, the state of the nucleus when protons of less-than-maximum energy have been emitted from the atom. The energy of the protons, in this instance, is greater than the ground state.

exciter **1.** An amplifier or oscillator (or a system of such units) that supplies the input (driving) signal to the output amplifier in a radio transmitter or similar device. **2.** A small direct-current (dc) generator that supplies direct current to the field winding of an alternating-current (ac) generator. **3.** See INDUCTION COIL.

exciter lamp **1.** A concentrated-filament, high-intensity incandescent lamp used in sound-on-film recording and reproduction and in some types of electromechanical television. **2.** In a facsimile transmitter, the lamp illuminating what is being scanned.

exciter relay In an electromechanical generator, the relay that activates the direct-current (dc) field excitation during machine startup.

exciter response **1.** A change in the exciter voltage of a motor when the field-circuit resistance changes. **2.** A change in the operating conditions of a radio frequency exciter, as a result of a change in the impedance at the input of the final amplifier.

exciter unit See EXCITER.

exciting current **1.** The output current produced by the exciter of a generator (see EXCITER, **2**). **2.** The field current of a dynamo-type generator. **3.** Primary current in an unloaded transformer.

exciting power **1.** The output power produced by an exciter. **2.** The input-signal power required for full output from a power amplifier. Also called DRIVING POWER.

exciting voltage **1.** Input-signal voltage. **2.** The input-signal-voltage amplitude required for full rated output from a power amplifier. Also called DRIVING VOLTAGE. **3.** The output voltage produced by an exciter.

exciton In a semiconductor or dielectric, a bound electron-hole pair.

excitron A mercury-pool rectifier whose arc is initiated mechanically (e.g., by means of a magnetic plunger in the tube).

exclusion principle The rule that only one particle of a particular kind can occupy a given quantum state at one time.

exclusive-NOR A logic function where the output is 1 if both inputs are 1 or both are 0 (same). The output is 0 if one input is 0 and the other is 1 (different). Compare EXCLUSIVE OR.

exclusive-OR A logic function in which the output is 1 when the two inputs are different, and is 0 when the two inputs are the same.

excursion **1.** A change in the value of a quantity in a given direction. **2.** In an oscillatory system, a body's moving away from the point of equilibrium or mean position.

execution A computer's performance of the operations required by an instruction.

execution time The length of time required for a computer to complete a designated operation.

executive routine In computer operations, a program that controls and processes other routines. Also called *monitor program.* Compare MONITOR SYSTEM.

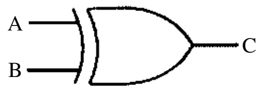

A	B	C
0	0	0
0	1	1
1	0	1
1	1	0

exclusive-OR

exhaust analyzer An instrument for examining the exhaust fumes of an internal combustion engine, to measure the presence of noxious materials and to evaluate air-to-fuel ratio and combustion efficiency.

exhaustion See EVACUATION.

exit **1.** In computer operations, the last instruction in a program or program segment, often taking a subroutine back to the main program. **2.** To leave a computer application, routine or subroutine.

exoskeleton A robot that resembles a suit of armor, and that greatly magnifies the force of physical movements. A human operator occupies the interior. Thus, for example, the operator might use the machine to throw a football 2500 yards, or to run 50 miles an hour, or to smash through walls. Primarily a tool of science fiction writers, this machine is within the scope of current robotic technology.

exosphere The extreme outer layer of the earth's atmosphere.

exothermic Pertaining to a chemical or electrochemical reaction in which heat is given off. Compare ENDOTHERMIC.

exothermic reaction In a chemical reaction, the production of positive reaction energy (i.e., kinetic energy is gained). Compare ENDOTHERMIC REACTION.

exp **1.** Symbol for EXPONENTIAL. **2.** Abbreviation of EXPERIMENT(AL).

expand **1.** In communications, to increase the bandwidth of a signal, restoring it to normal bandwidth after it has been compressed. **2.** In communications, to increase the dynamic range of a signal. **3.** In computer operations, to restore a file to full or normal size after it has been compressed. **4.** To widen the scale of a meter. **5.** To widen (or magnify a portion of) the trace of an oscilloscope beam. Compare COMPRESS.

expandable Capable of being built up into larger circuits or systems.

expandable gate In digital logic, a gate that can be provided with an unlimited number of input lines by electrical interconnection with other gates.

expanded memory In personal computer systems, memory beyond the basic 640 kilobytes (640 kb), up to one megabyte (1 MB). This memory resides in integrated circuits (ICs) in the computer, and is normally volatile (i.e., it is not retained when power is removed). Compare EXTENDED MEMORY.

expanded-scale meter A meter having a scale designed to display a narrow range of values. Such a meter used for monitoring the 117-V power line might have a scale reading 100 to 140 V, instead of a conventional scale beginning at zero.

expanded sweep **1.** In an oscilloscope, speeding up the deflection of the beam during a selected portion of the trace. **2.** The circuit for the action described in **1**.

expander A circuit for increasing the dynamic range over which a signal or quantity can vary. A typical example is the volume expander, a device that greatly increases the amplitude of strong signals while weakening, or having no effect on, signals of low amplitude.

expansion **1.** In communications, the process of increasing the bandwidth of a signal, restoring it to normal bandwidth after it has been compressed. **2.** In communications, a process in which stronger components are amplified more than weak ones, restoring a signal to its normal dynamic range after it has been compressed. **3.** In computer operations, the restoration of a file to full or normal size after it has been compressed. **4.** The widening of a meter scale. **5.** The widening or magnification of an oscilloscope trace. Compare COMPRESS.

expansion chamber A cloud chamber for viewing the paths of radioactive particles. It consists of a closed glass cylinder containing humid air and a piston. An electrostatic field is applied through the cylinder, the piston is pulled quickly, and the volume of the chamber expands. The temperature inside falls below the dew point, a cloud is formed, and droplets of water condense on ions, making their paths visible for observation or photography through the cylinder walls.

expansion ratio In communications, the inverse of COMPRESSION RATIO.

expansion time For an expansion chamber, the interval during which expansion occurs. The interval is kept short to ensure that the temperature will drop low enough for vapor condensation, and to minimize the possibility of continuing gas motion distorting the track of a particle.

expectation In probability theory, the middle value (average or mean) of a random variable.

expendable A component or system that, for economy, is best discarded instead of repaired when it fails. Also called DISPOSABLE COMPONENT.

experiment One or a series of carefully planned tests carried out under controlled conditions to obtain data or to check performance.

experimental chassis See ELECTRONIC CHASSIS.

experimental model A prototype of an electronic circuit or device, produced solely for operational tests or as a model against which theory and design can be checked.

experimental service A special, nonamateur radio service intended for on-the-air testing of new methods and equipment.

experimental station A station specially licensed to operate on specific frequencies in the experimental service.

expert system Also called *rule-based system.* A form of artificial intelligence (AI) that allows a computer or smart robot to act as a highly talented specialist in a specific field. An example is the use of a computer to help a physician diagnose a complex disease. A smart robot might be used as a surgical assistant.

exploring coil A pickup coil for sensing a signal or magnetic field. Sometimes called a *sniffer.*

exploring electrode **1.** A sampling electrode sealed in a discharge tube for measuring ionization at the point of insertion. **2.** Broadly, a test probe.

explosion-proof device A device that is housed and operated so that its sparking, heating, or production of radiant energy will not cause materials in the environment to explode.

exponent A number written as a superscript indicating the power to which another number (called the *base*) is to be raised. For example, 2^2 is the square (second power) of 2; x^3 is the cube (third power) of x.

exponential **1.** A base (such as the natural number e) modified by an exponent. **2.** Related to a change in value as determined by an exponent. Thus, using increments for x in the equation $y = e^x$ produces an exponential curve.

exponential curve A curve based on powers of a number (such as for $y = e^x$). Also see EXPONENTIAL, EXPONENTIAL DECREASE, and EXPONENTIAL INCREASE.

exponential damping Damping action described by an exponential curve.

exponential decay See EXPONENTIAL DECREASE.

exponential decrease The continuous reduction in the value of a quantity, according to the equation $y = e^{-x}$, which depicts the natural decay curve.

exponential function A function, such as $f(x) = e^x$, that varies exponentially. See, for example, EXPONENTIAL DECAY, EXPONENTIAL GROWTH, and EXPONENTIAL SERIES.

exponential horn A horn of circular or rectangular cross section, whose cross-sectional area S at any point x feet along its axis is given by the formula $S = S_0 e^{mx}$, where S_0 is the cross-sectional area at the throat, e is the natural logarithm base (approximately 2.71828), and m is the horn flaring constant.

exponential increase The continuous increase of a quantity, according to the equation $y = e^x$, which depicts the natural growth curve.

exponential line A transmission line whose characteristic impedance varies exponentially with its electrical length.

exponential quantity A quantity involving an exponential (e.g., $3e^x$).

exponential series A mathematical series based on exponential expressions. Example: $e^x = 1 + x + (x^2/2!) + (x^3/3!) + (x^4/4!) + \ldots$

exponential sweep In cathode-ray-tube (CRT) devices, such as oscilloscopes, a beam sweep that starts fast and slows exponentially.

exponential transmission line See EXPONENTIAL LINE.

exponential waveform Any waveform in which the rate of change in the amplitude is directly or inversely proportional to the instantaneous amplitude. The absolute value of the derivative of such a waveform is equal to the absolute value of the instantaneous amplitude, multiplied by a constant that depends on the amplitude units.

exposure **1.** The total amount of radiation received in a given area, or by a given sample, or by a person, over a specified length of time. **2.** The extent to which a photographic film has been darkened or otherwise modified by visible light, infrared, ultraviolet, or X rays.

exposure meter **1.** A usually simple instrument for measuring light intensity—especially for photographic purposes. A common form consists of a self-generating photocell connected to a direct-current microammeter. **2.** A device that indicates the amount of ionizing radiation that has been received over a given period of time.

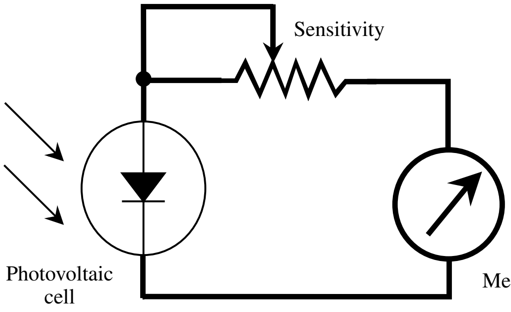

exposure meter

expression control A volume control in an electronic organ.

extended double Zepp antenna See DOUBLE-EXTENDED ZEPP ANTENNA.

extended memory In personal computer systems, memory beyond the first megabyte (1 MB). There is no limit in theory to the extent of this memory, although it is limited by current technology. Used in high-level programs—especially those involving graphical applications, artificial intelligence, or intensive calculations. Resides in integrated circuits (ICs) in the computer, and is normally volatile (i.e., it is not retained when power is removed). Compare EXPANDED MEMORY.

extended octaves Audio-frequency tones above or below the normal range of an electronic musical instrument. Special circuits must be added to make the extended octaves available.

extended play Pertaining to a recorded phonograph disc that provides a longer playing time than conventional discs of the same size and recording speed.

extender A substance added to an encapsulant to make it go further.

extensimeter See EXTENSOMETER.

extension cable A flexible, low-capacitance (usually concentric) cable for connecting part of one circuit to part of another. Extension cables are available with a variety of end connectors.

extension cord A flexible power cord having a male plug on one end and female receptacle on the other.

extension loudspeaker An auxiliary loudspeaker serving areas in which the main speakers can't be adequately heard.

extensometer An instrument used to measure small amounts of expansion, contraction, or deformation.

exterior label On a diskette or tape cartridge used for computer data storage, a written identification on the housing or cartridge, as opposed to the label, which is recorded on the diskette or tape itself.

external armature In a dynamo-type machine, an armature that rotates around the outside of the field magnets, as opposed to the usual (inside) arrangement.

external capacitor A high-value capacitor connected externally to an oscillator or sweep generator to lower its frequency.

external circuit A circuit or subcircuit connected and external to a main equipment.

external controls **1.** Control devices that are connected to, but operated away from, a main circuit. **2.** Manual or screwdriver-adjusted controls that are mounted on the panel of an equipment, as opposed to those mounted in the case or behind the panel.

external critical damping resistance The value of external resistance that must be connected to a galvanometer or other meter to produce critical damping.

external damping device A resistor or short-circuiting bar connected temporarily between the terminals of a meter to keep its movement immobile during transportation.

external feedback Negative or positive feedback through a separate path outside of and around the main circuit. Example: negative feedback through a resistance-capacitance (RC) path between the output terminals and the input terminals of an amplifier.

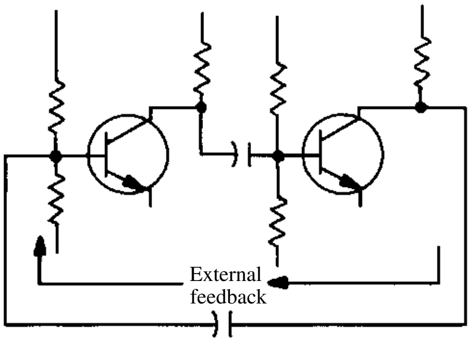

external feedback

external impedance Load impedance (i.e., an impedance connected to the output terminals of a generator or amplifier).

external load See EXTERNAL IMPEDANCE.

external loudspeaker See EXTERNAL SPEAKER.

externally caused chatter In a relay, contact chatter caused by mechanical vibration outside of the relay.

externally caused failure Failure of a circuit or component resulting from unfavorable environmental factors.

external memory In computer operations, a memory unit outside of the computer mainframe.

external power supply A power supply unit situated apart from the powered equipment. Such separation is helpful in eliminating the disturbing effects of heat, hum, and vibration associated with internal power units.

external *Q* For a microwave tube, the quantity $1/(1/Q_1 + 1/Q_2)$, where Q_1 is the loaded Q and Q_2 is the unloaded Q.

external S-meter A signal-strength meter connected to a receiver, but not installed in its panel.

external speaker A loudspeaker that doesn't share an enclosure with an amplifier, receiver, or other device that drives it. Such isolation is helpful in eliminating the undesirable effects of vibration and acoustic feedback.

external storage In computer operations, storage media (such as magnetic diskettes or tapes) that are outside of the computer.

extinction potential See DEIONIZATION POTENTIAL.

extinction voltage See DEIONIZATION POTENTIAL.

Extra-class license An amateur-radio license that conveys all available amateur operating privileges in the United States. The highest class of amateur license.

extract **1.** To remove a signal or quantity from some product containing it, or from its source. Examples: extracting a fifth harmonic from a complex signal, extracting the direct-current (dc) component from a signal containing both alternating current (ac) and dc. **2.** To derive a factor (e.g., to extract a root). **3.** To separate certain classes of information from an aggregate of information.

extract instruction In computer operations, the instruction to generate a new word by the serial arrangement of designated segments of specified words.

extractor **1.** A circuit or device for removing a signal (or a signal component) from another circuit or device. A demodulator probe, for example, extracts the modulation from a modulated signal. **2.** A device for removing used active devices from a circuit board. Such extractors can also employ heat to desolder, as well as remove the devices.

extraneous component A usually undesired inherent effect that results from the physical nature of a component or device. Examples: distributed capacitance of a coil, internal inductance of a capacitor.

extraneous emission Undesired emission from a transmitter (e.g., excessive harmonics).

extraneous response The unintended response of a circuit or device (e.g., image response in a superheterodyne communications receiver).

extraneous root In the solution of an equation derived from another equation, one or more roots that satisfy the derived equation but not the original one.

extraneous signal A superfluous and potentially interference-causing signal.

extranuclear Outside the nucleus of an atom.

extraordinary ray Of the two rays resulting from the double refraction of electromagnetic waves, the one that does not follow the usual laws of refraction. Also see X WAVE.

extraordinary wave See X WAVE.

extrapolar **1.** Outside of electrical or magnetic poles. **2.** Not between electrical or magnetic poles.

extrapolation Estimation of values beyond the range of available data. An example is the extension of a curve beyond its final plotted point to determine a value for a variable. There is always some margin for error, which increases as the process is extended further and further beyond the range of actual data values. Commonly done by computers (e.g., prediction of the probable path of a hurricane 12, 24, 36, and 48 hours in advance).

extrared See INFRARED.

extraviolet See ULTRAVIOLET.

extreme **1.** The lowest or highest value of a quantity. **2.** The lowest or highest point of the dependent variable in the range of a function. **3.** An unusual value for a parameter (e.g., temperature extreme and pressure extreme).

extremely high frequency Abbreviation, EHF. A frequency near the upper limit of the radio-frequency spectrum—especially one in the 30- to 300-GHz band.

extremely low frequency **1.** Abbreviation, ELF. Pertaining to a signal or current within the audio-frequency (AF) range, but not used for audio applications. **2.** Electromagnetic emissions from a cathode-ray tube (CRT), resulting from the currents in the deflecting coils and/or electron beam. **3.** The 60-Hz electromagnetic field generated by utility power lines and wiring.

extrinsic base-resistance/collector-capacitance product Units: seconds, milliseconds, and microseconds. For a bipolar transistor, the product R_BC_C, where R_B is the base resistance and C_C is the collector capacitance. This product is a time constant that determines the high-frequency operating limit of the transistor.

extrinsic conductance For a material, the conductance resulting from impurities or such external factors as environmental conditions.

extrinsic properties For a semiconductor material, properties resulting from doping (e.g., altered resistivity or majority/minority carrier differentiation). Also see EXTRINSIC SEMICONDUCTOR.

extrinsic semiconductor A semiconductor material, such as germanium or silicon, to which a controlled amount of a suitable impurity material has been added to give the semiconductor a desired resistivity and polarity. Compare INTRINSIC SEMICONDUCTOR.

extrinsic transconductance Symbol, g_m. For a bipolar transistor, the first derivative of collector current, with respect to base-emitter voltage. It is the ratio of a small change in collector current (dI_C) to the small change in base-emitter voltage (dV_{BE}) that produced it, collector voltage being constant; $g_m = dI_C/dV_{BE}$.

extrusion The process of forming a material such as metal or plastic, by forcing it through dies. Many pieces of electronic hardware are mass produced in this manner. Examples are insulating rods and tubes, metal cans, and metal tubing.

eye-in-hand system In robotics, a scheme that uses a vision system, servo, and microprocessor to precisely manipulate an END EFFECTOR, such as a robot gripper ("hand"). A camera (the "eye") is contained in the end effector itself, and is designed to work at close range (i.e., approximately one meter down to a fraction of a millimeter).

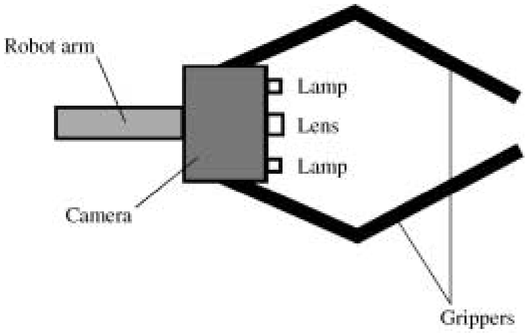

eye-in-hand system

eyelet connection A connection made by fastening conductors together with an eyelet or by soldering leads or pigtails to an eyelet.

eyepiece A small lens system for viewing an oscilloscope screen through a camera setup.

E_Z Symbol for *voltage drop across an impedance.*

E zone A portion of the earth including most of the eastern hemisphere. When propagation forecasts are made, this region is one of three longitude zones specifie

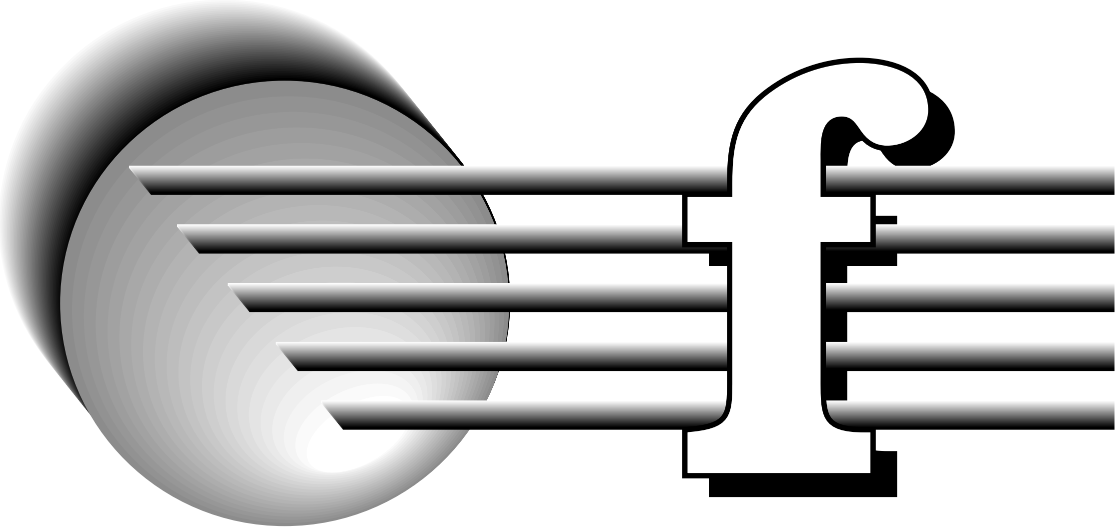

F **1.** Symbol for FORCE. **2.** Symbol for FLUORINE. **3.** Abbreviation of FAHRENHEIT. **4.** Abbreviation of FARAD. **5.** Abbreviation of FERMI. **6.** Symbol for FOCAL LENGTH. **7.** Symbol for FUSE. **8.** Symbol for FARADAY CONSTANT.

f **1.** Abbreviation of FEMTO. **2.** Symbol for FREQUENCY. **3.** Symbol for FUNCTION.

F_0 Symbol for DAMPING FACTOR.

F1 layer The lower part of the ionosphere's F REGION. Also called *F1 region.*

F2 layer The upper part of the ionosphere's F REGION. Also called *F2 region.*

fA Abbreviation of FEMTOAMPERE.

fabrication tolerance The amount of variation that can be tolerated in the manufacture of components.

Fabry-Perot interferometer A resonant cavity, often used with lasers, that has mirrors at each end; the interferometer produces the optical equivalent of standing waves.

face **1.** A flat crystal surface whose orientation can be expressed as its position relative to other faces. **2.** The viewer's side of a screen. **3.** The scale part of a meter.

face material In a tape recorder, the plastic used to coat the face of a head.

face-parallel cut See Y-CUT CRYSTAL.

face-perpendicular cut See X-CUT CRYSTAL.

face side The side of pressure-sensitive insulating tape that is coated with adhesive.

facom A radionavigation system that operates by means of phase comparison at low frequencies. Effective over long distances and under poor conditions.

facsimile Also called *fax.* The transmission and reception, through the medium of radio or by wire, of permanent pictures, writing, and other graphic material.

facsimile receiver The complete device or system that selects, amplifies, and demodulates a picture signal received up from the air, wires, or cable, and uses the elements of this signal to reproduce the picture. Also see FACSIMILE.

facsimile recorder The machine that puts a transmitted facsimile image on paper.

facsimile transmitter The complete device or system that generates signals depicting graphic material (pictures, writing, printing, etc.) and sends them to a distant point via cable, wire lines, or radio for subsequent reproduction. Also see FACSIMILE.

factor **1.** A data element that is an operand in an arithmetic operation. **2.** To find the two or more numbers whose product is the number being factored. **3.** One of two or more numbers whose product is the number being factored.

factorial Symbol, !. For an integer *n,* the product of all positive integers up to and including *n.* Thus, $5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$. The term *n*! is read "*n* factorial."

factor of merit See FIGURE OF MERIT.

factor of safety See SAFETY FACTOR.

fade in To gradually increase an audio or video signal—especially for recording.

fade out **1.** To gradually decrease an audio or video signal—especially for recording. **2.** The complete disappearance of a radio communications signal. See FADING.

fader **1.** In sound amplification systems, an attenuator circuit that enables the operator to fade out one signal and fade in another. Ordinarily, a fader does not provide mixing action. **2.** In an au-

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

tomotive high-fidelity sound system, a control that adjusts the volume of the front speaker set and rear speaker set. Usually, this is a single knob; increasing the front volume decreases the rear volume, and vice-versa.

fading **1.** Repeated increases and decreases of radio communications signal amplitude at the reception point. Fading of a radio wave results from multiple propagation paths from transmitter to receiver, causing the received waves to arrive in constantly varying phase relationships. **2.** In a cell or battery power supply, a gradual loss of power-delivering capability that occurs during the discharge cycle.

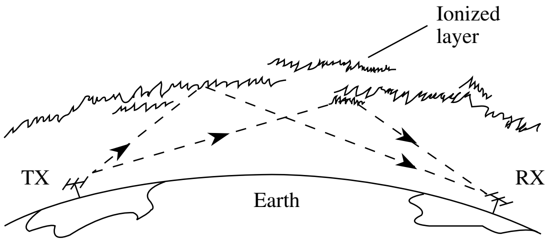

fading

factory automation The use of robotic devices, to the greatest extent possible, as laborers in manufacturing plants.

Fahnestock clip A flat, sheet-metal spring clip for holding a wire (usually in a temporary breadboard setup).

Fahrenheit scale A temperature scale on which the freezing point of water is 32 degrees, and the boiling point of water is 212 degrees. Compare ABSOLUTE SCALE and CELSIUS SCALE.

failsafe Pertaining to devices or circuits that, upon failure, cause no damage or serious malfunction.

failsoft In a computer, a system in which operation is maintained—even in the event of partial failure. Efficiency is reduced but the computer does not completely shut down.

failure The condition wherein a circuit, system, or device is not operating correctly.

failure analysis **1.** The process of determining the failure rate for a component, system, or device. **2.** The process of determining the cause of a failure.

failure mode The particular way in which a failure of equipment or a method occurs.

failure unit A unit of machine or device failure: one failure per billion (10^9) hours of operation.

fall-in The time when synchronous speed is attained in a synchronous motor.

falling characteristic A NEGATIVE RESISTANCE characteristic.

fall time **1.** Decay time. **2.** The time required for the amplitude of a pulse to decrease from 90 percent to 10 percent of its peak amplitude. Compare RISE TIME.

false add A logic add (i.e., addition without carries).

false alarm **1.** Improper operation of an electronic security system, resulting in actuation of the device when no breach of security has occurred. **2.** In radar, the presence of a false echo that causes the attendant circuits or personnel to act as though an enemy target is present.

false error A condition in which a computer system erroneously signals the existence of an error.

false precision See MISLEADING PRECISION.

false retrieval The incorrect specification of criteria for information to be selected for retrieval so that an unwanted item of data is selected. Also called *false drop.*

family Any group of components, circuits, ratings, or characteristics, classed together because of some common or analogous feature or application. Examples: IC family, family of curves, and family of equations.

family of curves A group of curves plotted on the same axes, that depict the performance of a circuit or device at several levels of a third parameter (e.g., curves showing transistor collector current vs. collector voltage for several levels of base current).

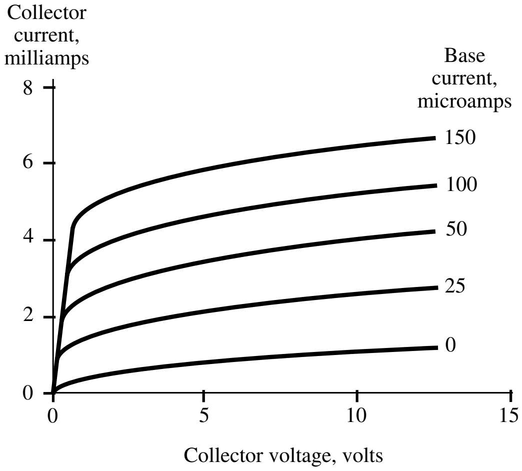

family of curves

fan antenna See DOUBLE-V ANTENNA.

fan-in **1.** A number of inputs entering a common input terminal. **2.** In digital computer operations, the number of inputs that can be accommodated by a logic circuit. Compare FAN-OUT.

fan-in circuit A circuit having a number of input lines entering a common input point. Also see FAN-IN, **1**.

fan-out **1.** One common output terminal feeding a number of output lines. **2.** In digital computer operations, the number of outputs that can be fed by a logic circuit. Compare FAN-IN.

fan-out circuit A circuit in which a number of output lines leave a common output terminal. Also see FANOUT, **1**.

fantasy robot A robot as portrayed in science fiction or fantasy. Such robots are usually, but not always, *androids* (that is, they have humanoid form). Some fictional robot characteristics are later realized. Science fiction can even give roboticists ideas for future designs.

farad (Michael Faraday, 1791–1867) Abbreviation, F. The basic unit of capacitance. A capacitor has a capacitance of 1 F when a charge of 1 volt per second across the capacitor produces a current of 1 ampere through it.

faraday An electrical quantity approximately equal to 9.65×10^4 coulombs; it is the quantity of electricity required in electrolysis to free 1 gram atomic weight of a univalent element. The equivalent, and preferred, unit is the COULOMB. Also called *Faraday constant.*

Faraday cage See ELECTROSTATIC SCREEN.

Faraday constant See FARADAY.

Faraday cylinder A cylindrical metal shield placed around electrical equipment or circuits to prevent electromagnetic fields from affecting the equipment or circuits.

Faraday rotation **1.** A change in the polarization of an electromagnetic wave as it passes through a medium. **2.** Constant and random fluctuations in the polarization of radio signals as they pass through the earth's ionosphere. This phenomenon affects signals having linear polarization at frequencies well below 10 GHz. It has little or no effect on circularly polarized signals, or on signals at frequencies above 10 GHz. It can be a problem with some satellite communications links.

Faraday effect See MAGNETO-OPTICAL ROTATION.

Faraday's disk dynamo See DISK DYNAMO.

Faraday shield See ELECTROSTATIC SCREEN.

Faraday's law The voltage induced in a conductor moving in a magnetic field is proportional to the rate at which the conductor cuts the magnetic lines of flux.

Faraday's laws of electrolysis **1.** In electrolysis, the mass of a substance liberated from solution is proportional to the strength and duration of the current. **2.** For different substances liberated by the same current in a certain time, the masses are proportional to the electrochemical equivalents of the substances. Also see ELECTROCHEMICAL EQUIVALENT; ELECTROLYSIS, **1**; and ELECTROLYTE.

faradic current The lopsided alternating current produced by an induction coil.

faradmeter An alternate term for MICROFARAD METER.

far field **1.** The region beyond the near field of an antenna (see NEAR FIELD, **1**). **2.** The region beyond the near field of a loudspeaker (see NEAR FIELD, **2**).

far infrared Also called far IR. The lower-frequency portion of the infrared (IR) spectrum.

farming robot A robot that performs labor in an agricultural setting. Examples: fruit-picking robot, field-watering robot, cow-milking robot. Such machines are usually overseen by a central computer, via which the farmer can monitor and control the robots' actions.

far zone See FAR FIELD, **1**, **2**.

fast access storage In a computer memory, the section from which information can be most quickly accessed, depending on the relative speed of other system devices.

fast-break, fast-make relay A relay that opens and closes rapidly.

fast-break, slow-make relay A relay that opens rapidly and closes slowly.

fast charge The rapid charging of a rechargeable cell or battery, particularly of a nickel-cadmium (NICAD) or nickel-metal-hydride (NiMH) device.

fast diode See COMPUTER DIODE.

fast drift The rapid change of a quantity or setting, usually in one direction. Compare SLOW DRIFT.

fast-food robot A robot that prepares and/or serves items in a fast-food establishment. Robots have not yet gained wide acceptance in this setting, mainly because of technical difficulties and prohibitive cost. Another problem is that some customers are put off or intimidated by robots.

fast-forward Abbreviation, FF or FFWD. In a tape recorder, a mechanism for running the tape through the machine rapidly.

fast groove The informationless groove between tracks on a disk recording.

fast-make/fast-break relay A relay that closes and opens rapidly.

fast-make, slow-break relay A relay that closes rapidly and opens slowly.

fast-forward playback In a videotape recorder, the playing back of the tape at faster than real-life speed. It allows the viewer to move quickly ahead in a program, and also to watch the images so as to know when to resume normal replay.

fast-reverse playback In a videotape recorder, the playing back of the tape rapidly backwards. It allows the viewer to move quickly back in a program, and also to watch the images so as to know when to resume normal replay.

fast time constant **1.** The property of responding quickly to changes in input parameters. **2.** In radar, a method of defeating attempts at jamming by modification of the receiving circuitry.

fathometer See ACOUSTIC DEPTH FINDER.

fathom **1.** To measure the depth of a body of water, as in the use of sonar for this purpose. **2.** A unit of length (distance) equal to six feet.

fatigue **1.** The degradation of the performance of circuits or materials with time. **2.** The tendency of bodies and materials to weaken, deform, or fracture under repeated strain.

fault **1.** A defective point or region in a circuit or device. **2.** A failure in a circuit or device.

fault current **1.** A momentary current surge. **2.** A leakage current.

fault finder A troubleshooting instrument or device (e.g., a multimeter).

fault resilience **1.** A design scheme for an electronic or computer device or system so that if a component or circuit fails, the system will continue to operate, although perhaps at reduced efficiency. The operator is notified of the problem so that it can be repaired with minimal downtime. **2.** In a computer system, the property of being as nearly sabotage-proof as possible.

fault tolerance Total redundancy in an electronic or computer system so that if a component or circuit fails, the system will continue to function at full efficiency. Every component has a backup that automatically takes over in case of failure. The operator is notified of the problem, so the defective part or circuit can be replaced while its backup keeps the circuit working continuously at 100-percent capacity.

Faure plate A storage battery plate consisting of a lead grid containing a chemical electrolytic paste.

fax Abbreviation of FACSIMILE.

fc Abbreviation of FOOT-CANDLE.

f_c Abbreviation of CARRIER FREQUENCY.

FCC See FEDERAL COMMUNICATIONS COMMISSION.

f_{co} Abbreviation of CUTOFF FREQUENCY.

F connector A type of antenna feedline connection especially common on television receivers and videocassette recorders.

F display See F SCAN.

FDM Abbreviation of *frequency-division multiplex.*

FE Abbreviation of FERROELECTRIC. See FERROELECTRICITY.

Fe Symbol for IRON.

feasibility study The procedures for evaluating the potential gains in applying a computer system to a job or to an organization's process, or in modifying or replacing an existing system.

FEB Abbreviation of FUNCTIONAL ELECTRONIC BLOCK.

Federal Communications Commission Abbreviation, FCC. Established in 1934, the U.S. Government agency that regulates electronic communications. The FCC succeeded the Federal Radio Commission (FRC), which was established in 1927; the FRC succeeded the Radio Division of the Bureau of Navigation in the Department of Commerce, whose jurisdiction over radio began in 1912.

feed **1.** To supply power or a signal to a circuit or device. **2.** The method of supplying such a signal or power. See, for example, PARALLEL FEED and SERIES FEED. **3.** To cause data to be entered into a computer for processing.

feedback **1.** The transmission of current or voltage from the output of a circuit or device back to the input, where it interacts with the input signal to modify operation of the device. Feedback is positive when it is in phase with the input, and is negative when it is out of phase. **2.** To input the result at one point in a series of operations to another point; the method allows a system to monitor its actions and make necessary corrections.

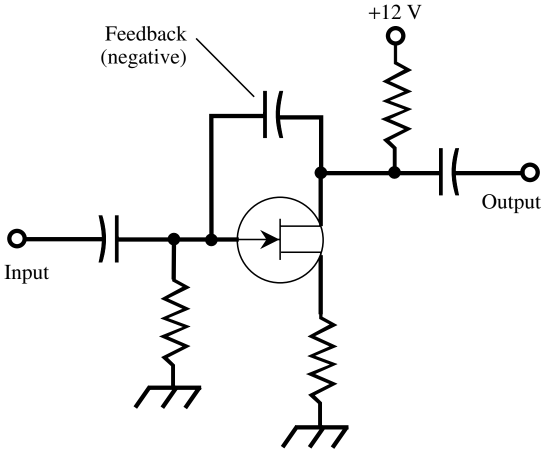

feedback, 1.

feedback amplifier **1.** An amplifier whose performance (especially frequency response) is modified by means of positive, negative or both positive and negative feedback. **2.** An amplifier placed in the feedback path of another circuit to increase the amplitude of feedback.

feedback attenuation **1.** In an operational-amplifier circuit, the attenuation in the voltage from output to input. **2.** In an audio-frequency or radio-frequency amplifier circuit, the reduction of feedback by electronic means.

feedback bridge A bridge circuit in the feedback channel of an amplifier or oscillator.

feedback capacitance **1.** A capacitance through which feedback current is coupled from the output to the input of a circuit or system. **2.** The interelectrode capacitance of a vacuum tube.

feedback control **1.** The variable component (potentiometer or variable capacitor) used to adjust the level of feedback current or voltage. **2.** The control of circuit performance by means of feedback.

feedback cutter A device used for the purpose of cutting grooves in phonograph disks. Feedback is used to provide a flat frequency response.

feedback factor For a feedback amplifier, the quantity $1 - bA$, where A is the open-loop gain of the amplifier and b is the FEEDBACK RATIO.

feedback input current In a feedback network, the current drawn by the feedback input. This current affects the design parameters of a network.

feedback loop The part of a circuit that provides controlled feedback in an operational-amplifier circuit.

feedback oscillator A circuit in which oscillation is obtained by feeding a portion of the output of an active amplifying device back to the input circuit by inductive coupling. Also called *tickler oscillator.*

feedback path A path over which feedback, either positive or negative, can occur in a circuit. The feedback can be intentionally produced, or it can be undesirable.

feedback percentage Symbol, n. In a feedback circuit, the percentage of output voltage that is fed back; $n = 100V_f/V_o$, where V_f is the feedback voltage and V_o is the open-loop output voltage. Compare FEEDBACK RATIO.

feedback ratio For a feedback system, the ratio V_f/V_o, where V_f is the voltage that is fed back and V_o is the open-loop output voltage of the system.

feedback rectifier See DIODE FEEDBACK RECTIFIER.

feedback regulator In a controlled-feedback circuit, the device that determines the amount of feedback.

feedback resistance **1.** The internal base resistance of a point-contact transistor. **2.** The resistance in a feedback loop.

feedback transfer function The transfer function of a feedback loop exclusively.

feedback winding A special winding on a magnetic amplifier or saturable reactor, for the introduction of feedback currents.

feeder **1.** A conductor or set of conductors that carries electric power from one point to another. **2.** The transmission line connecting a transmitter to an antenna.

feeder cable **1.** A communication cable running in a primary route from a central station (or in a secondary route from a main feeder cable) as a means of making connections to distribution cables. **2.** In a cable television system, the cable carrying transmission from the head end to the trunk amplifier. Also called TRUNK CABLE.

feeder loss Loss of energy resulting from resistance in, or radiation from, feeder lines.

feeding In character recognition, a system in which documents go into the transport of a character reader at a steady, specified rate.

feed pitch The distance between feed holes.

feed reel The tape supply reel of a tape recorder.

feedthrough **1.** The usually undesirable transmission of a signal through a circuit without being processed by the circuit, because of unavoidable capacitive coupling, for example. **2.** Contraction of FEEDTHROUGH COMPONENT.

feedthrough capacitor A capacitor whose design is like that of a feedthrough terminal; it is mounted in a hole in a chassis. The center screw or wire is the "high" terminal of the capacitor, to which connections can be made above or below the chassis. The body of the device is the "low" terminal of the capacitor; it is soldered to the chassis or secured with a nut.

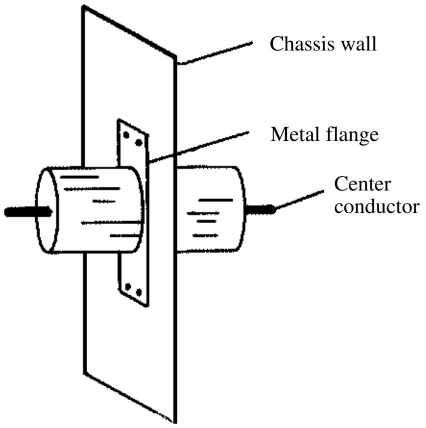

feedthrough capacitor

feedthrough component A passive device permanently installed in a panel or plate (e.g., a FEEDTHROUGH CAPACITOR or FEEDTHROUGH INSULATOR).

feedthrough insulator An insulator mounted tightly in a hole in a wall or chassis, and provided with a center hole for a lead.

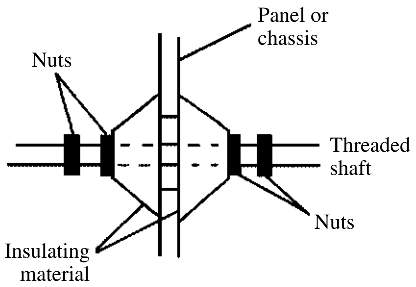

feedthrough insulator

feedthrough terminal A terminal mounted tightly in a hole in a chassis or wall; it consists of a screw going through a feedthrough insulator. Connections can be made to either end of the screw.

FE-EL Abbreviation of *ferroelectric-electroluminescent.*

feeler **1.** A wire or blade contact (e.g., a finger that senses holes in a punched card). **2.** The wire portion of a point-contact diode.

Felici mutual-inductance balance An inductive null circuit for determining mutual inductance (M_x) in terms of a standard mutual inductance (M_s). The secondary coils of two mutual-inductance circuits are connected in phase opposition. The standard mutual inductor, which is variable, is adjusted for null. At null, $M_x = M_s$.

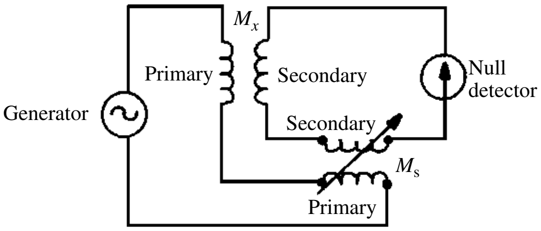

Felici mutual-inductance balance

female plug A plug whose contacts are separated by a recess into which the prongs of a mating male plug are inserted. Compare HERMAPHRODITE PLUG and MALE PLUG.

femto- Abbreviation, f. A prefix meaning quadrillionth (10^{-15}).

femtoampere Abbreviation, fA. A unit of extremely low current; 1 fA equals 10^{-15} A.

femtofarad Abbreviation, fF. A unit of extremely low capacitance; 1 fF equals 10^{-15} F.

femtovolt Abbreviation, fV. A unit of extremely low voltage; 1 fV equals 10^{-15} V.

fence A system or string of early warning radar stations.

Fermat's principle The principle that the path of a ray of radiant energy between two points in any medium is the shortest distance between those points in that medium. This path is also the path of least propagation time. This might not necessarily be a straight line. On a flat plane, the path of least time between any two points is always a straight line, assuming constant velocity of propagation. However, on a sphere, the path of least time between two points is represented by a great circle.

fermi Abbreviation, F. An extremely small unit of length and wavelength, equal to one quadrillionth (10^{-15}) of a meter.

Fermi-Dirac distribution function The probability that an electron will be in a certain quantum state under conditions of thermal equilbrium. This is determined as the probability that the electron will be in a given energy band at a particular instant of time.

fermium Symbol, Fm. A radioactive metallic element that is artificially produced. Atomic number, 100. Atomic weight, 257.

ferpic Acronym for *ferroelectric ceramic picture* device. An image-storing device containing a photoconductive film, transparent electrodes, and a ferroelectric ceramic, in layers.

ferreed A form of magnetic switching device, similar to a reed relay, that maintains its position indefinitely without the need for a continuous current.

ferret A vehicle or craft equipped for determining the locations of enemy radar transmitters.

ferri- A prefix used to denote magnetic properties.

ferric oxide Formula, Fe_2O_3. A red oxide of iron used to coat magnetic recording tape.

ferristor A high-frequency magnetic amplifier using a ferroresonant circuit.

ferrite A high-resistance magnetic material consisting principally of ferric oxide and one or more other metals. After being powdered and sintered, ferrites exhibit low eddy-current loss at high frequencies and make ideal core material for inductors and switching elements. Also used in television deflection yokes and in miniature antennas. Also see FERROSPINELS.

ferrite antenna See FERRITE-ROD ANTENNA.

ferrite bead **1.** A magnetic storage device in the form of a bead of ferrite powder fused onto the signal conductors of a memory matrix. **2.** A tiny ring of ferrite that can be slipped over a wire or cable to choke off radio-frequency (RF) currents.

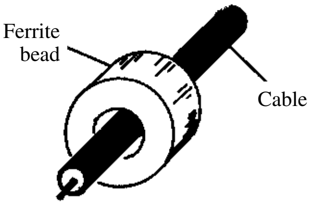

ferrite bead

ferrite core A coil or switching-element core made from a ferrite; specifically, in a core memory, a small magnetic toroid that can retain its polarity when charged by a pulse.

ferrite core memory A magnetic memory in which ferrite cores are interconnected by a network of input and output wires.

ferrite isolator A microwave device that permits energy to pass with negligible loss in one direction through a waveguide or coaxial line, while absorbing energy passing in the opposite direction.

ferrite limiter A device used in the antenna circuit or front end of a receiver, to prevent overload while maintaining a linear response. Used mostly at ultra-high and microwave frequencies.

ferrite loopstick See FERRITE-ROD ANTENNA.

ferrite memory A static memory using ferrite cores. See CORE MEMORY.

ferrite-rod antenna Also called *loopstick antenna*. A small antenna that can be used for wireless reception at frequencies below approximately 20 MHz. This antenna consists of a coil wound on a solenoidal, high-permeability, powdered-iron core, usually less than 20 centimeters (cm) long and 1 cm in diameter. A series or parallel capacitor, in conjunction with the coil forms a tuned circuit. The operating frequency is determined by the resonant frequency of the inductance-capacitance (LC) combination. Response is maximum off the sides of the coil, and a sharp null occurs off the ends. This antenna has narrow bandwidth. The null can be oriented to minimize system response to undesired local signals or humanmade noise. Compare SMALL LOOP ANTENNA.

ferrite switch A device that regulates the flow of power through a waveguide. The electric-field vector is rotated, resulting in a high degree of attenuation when actuated, but little or no attenuation when not activated.

ferroelectric **1.** Producing ferroelectricity. **2.** A ferroelectric material.

ferroelectric amplifier See DIELECTRIC AMPLIFIER.

ferroelectric capacitor A capacitor in which a ferroelectric material is the dielectric.

ferroelectric cell See FERROELECTRIC CAPACITOR.

ferroelectric crystal A crystal of ferroelectric material.

ferroelectric flip-flop A flip-flop based on the hysteresis of a ferroelectric capacitor. Compare FERRORESONANT FLIP-FLOP.

ferroelectricity Electric polarization in certain crystalline materials. The effect is analogous to the magnetization of a ferromagnetic material by a magnetic field.

ferroelectric-luminescent Pertaining to a ferroelectric cell that emits light.

ferroelectric material A nonlinear dielectric material capable of producing ferroelectricity. Examples: barium titanate, barium strontium titanate, potassium dihydrogen phosphate, guanadine aluminum sulfate hexahydrate (GASH), Rochelle salt, and triglycene sulfate.

ferromagnetic **1.** Pertaining to a substance that conducts a magnetic field with relative ease. **2.** Pertaining to a material in which a magnetic-field change causes a voltage, which in turn results in a measurable current flow.

ferromagnetic-core inductor A coil of wire designed to introduce inductive reactance into a circuit or system, wound around a core consisting of ferromagnetic material that greatly increases the inductance for a given number of turns. With proper design and choice of core material, the Q factor (selectivity attainable) can be extremely high. The core confines much of the magnetic flux within itself. When the coil current exceeds a certain level, *core saturation* occurs, and further increases in the current will not produce a corresponding increase in magnetic flux. This decreases the effective inductance, and reduces the efficiency because power is dissipated as heat in the core.

ferromagnetic material A substance that concentrates magnetic lines of flux relative to their concentration in free space. Iron, powdered iron, and ferrite are common examples.

ferromagnetic resonance The point at which the permeability of a magnetic material peaks at a microwave frequency.

ferromagnetic spinels Highly permeable and resistive ceramic-like materials. The low eddy-current losses and high permeability of these materials suit them for use as cores in radio-frequency (RF) transformers and inductors. Also see FERRITE.

ferromagnetic tape Magnetic tape used for winding closed transformer cores.

ferrometer An instrument for testing hysteresis and permeability in steel and iron.

Ferron detector See IRON-PYRITES DETECTOR.

ferroresonant circuit An inductance-capacitance (LC) circuit in which the coil is a saturable reactor. Because of coil nonlinearity, the circuit is resonant at only one value of alternating-current (ac) voltage, and exhibits both negative resistance and bistable operation.

ferroresonant counter A digital counter using ferroresonant flip-flops, rather than semiconductor devices.

ferroresonant flip-flop A flip-flop using one or two ferroresonant circuits instead of semiconductor devices. See BISTABLE MULTIVIBRATOR. Compare FERROELECTRIC FLIP-FLOP.

ferroresonant shift register A shift register using ferroresonant circuits instead of semiconductor devices.

ferrosoferric oxide See MAGNETITE.

ferrospinels See FERROMAGNETIC SPINELS.

ferrous Pertaining to a substance that contains iron and is magnetizable.

Ferroxcube A nonmetallic ferromagnetic material having high permeability and resistivity, and a Curie point near room temperature. These characteristics make the material suitable for the cores of radio-frequency (RF) inductors and transformers, and for high-frequency magnetic shields.

Fessenden oscillator In underwater communications, a transmitter of acoustic waves.

FET Abbreviation of FIELD-EFFECT TRANSISTOR.

fetch An operation in a computer run in which the location of the next instruction is taken from memory and changed if necessary; it then goes to the control register.

FET current meter An ammeter, milliammeter, or microammeter having a self-contained amplifier that uses field-effect transistors. Also see ELECTRONIC CURRENT METER.

FET op-amp **1.** An operational amplifier composed of field-effect transistors and associated components. **2.** An operational amplifier having a field-effect transistor in its input stage.

FET voltmeter A voltmeter using a field-effect transistor amplifier for high-impedance input. Also see ELECTRONIC VOLTMETER.

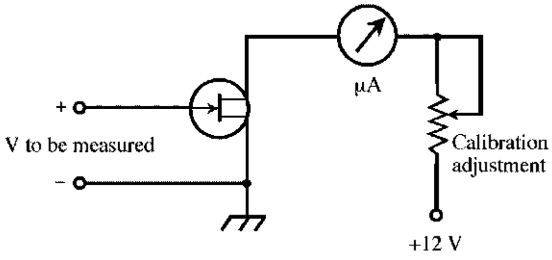

FET voltmeter

FET VOM A volt-ohm-milliammeter (VOM) using a field-effect transistor amplifier for increased sensitivity and high input impedance.

FF Abbreviation of FLIP-FLOP.

fF Abbreviation of FEMTOFARAD.

FFI Abbreviation of FUEL-FLOW INDICATOR.

fhp Abbreviation of FRACTIONAL HORSEPOWER.

fiber **1.** A tough, vulcanized insulating material. Dielectric constant, 2.5 to 5. Dielectric strength, 2 kV/mm. **2.** A thin thread of a material. **3.** Also called *optical fiber.* A light-conductive transparent filament; see FIBEROPTICS, **1**.

fiber electrometer An instrument for measuring small quantities of electricity. It consists of a thin thread, such as one of plasticized quartz, hanging freely between two knife-edged metal pieces that are charged by the electricity being measured. The charge draws the fiber away from its position of rest. The movement can be observed with a microscope. A special form of this instrument, using two fibers, is the *bifilar electrometer.*

fiber needle A soft phonograph needle made from a fiber. It produces less disk wear than other styli, but is short-lived.

fiber metallurgy A process in which metallic fibers or filaments are grown.

fiberoptic bundle A cable of optical fibers. See FIBEROPTICS, **1**.

fiberoptic coupling Also called *optical coupling.* A method of coupling in which a light-conducting fiber is placed between a light (signal) source and a photoreceptor. Also see FIBEROPTICS, **1**, **2**.

fiberoptics **1.** Also called *optical fibers.* Extruded materials, such as certain plastic filaments, that provide paths for light. **2.** The science of developing and using communications systems that use optical fibers to transfer data.

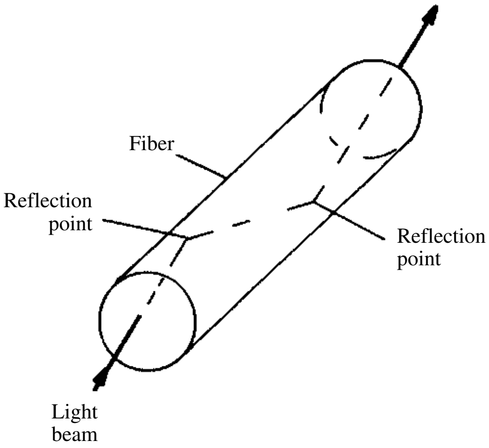

fiberoptics

fiberoptic scrambler A fiberscope in which a center section of fibers in the core is deliberately disoriented before the bundle is encapsulated; when cut, one half can decode the image encoded by the other half. See FIBERSCOPE.

fiberscope A flexible bundle of optical fibers having a lens at each end; it is used to view areas that are otherwise inaccessible to view.

fiber stylus See FIBER NEEDLE.

fibre Alternate (Brit.) spelling of FIBER.

fibrillation Dangerous, irregular beating of the heart that often follows electric shock. Also see CARDIAC STIMULATOR and DEFIBRILLATION.

fidelity The degree to which a circuit or device transmits a signal without distorting it. Pertains especially to acoustic devices and stereo audio equipment.

field **1.** A volume of space in which a force is operative. See, for example, ELECTRIC FIELD and MAGNETIC FIELD. **2.** Half of a video image, consisting either of all the even-numbered lines, or all the odd-numbered lines. **3.** A computer record subdivision containing an information unit (e.g., a bank account record might have deposits as a field).

field circuit breaker A circuit breaker designed to control the field excitation of a motor or other device.

field coil **1.** The winding on the field pole of a motor or generator. **2.** The winding on the pole of an electrodynamic speaker. **3.** The main coil of a relay. **4.** The fixed coil in an electrodynamometer.

Field Day In amateur radio, an annual contest sponsored by the American Radio Relay League (ARRL) to simulate emergency communication conditions. It occurs during the last full weekend in June.

field direction The direction in which an electric field or magnetic field exerts its force.

field effect The phenomenon in which the flow of current carriers in a solid substance is controlled by an external electric field. A useful application is the FIELD-EFFECT TRANSISTOR (FET).

field-effect tetrode Also called *dual-gate FET.* A field-effect transistor (FET) in which the gate electrode is split into two parts, each connected to a separate external lead. The reverse bias between the channel and either gate lead affects the conductivity through the device.

field-effect transistor Abbreviation, FET. A monolithic semiconductor amplifying device in which a high-impedance GATE electrode controls the flow of current carriers through a thin bar of silicon (rarely, germanium) called the CHANNEL. Ohmic connections made to the ends of the channel constitute SOURCE and DRAIN electrodes. Also see JUNCTION FIELD-EFFECT TRANSISTOR, METAL-OXIDE SEMICONDUCTOR FIELD-EFFECT TRANSISTOR, and CHANNEL JUNCTION FIELD-EFFECT TRANSISTOR.

field-effect varistor A nonlinear dual-terminal semiconductor device capable of maintaining a value of current for a range of voltages.

field-emission microscope An instrument for examining the atomic structure of high-melting-point metals; it magnifies more than 2 million times. The metal to be examined is made into a needle that is subjected to 5 to 30 kV; electrons emitted by the tip of the needle form an image on a fluorescent screen.

field forcing A method of controlling a motor by changing the magnetic field in the windings.

field frequency In television, the product of frame frequency and fields per frame (in the United States, 60 per second).

field intensity **1.** The strength of an electric or magnetic field. **2.** The strength of an electromagnetic field, usually expressed in microvolts per meter or millivolts per meter.

field-intensity meter See FIELD-STRENGTH METER.

field ionization The tendency for atoms to be ionized in a gas by a high-intensity electric field. The ionization occurs mostly near the poles of the electric field.

field-ion microscope A high-resolution field-emission microscope that uses helium ions instead of electrons. The ions are repelled by the tip of the metal needle under observation, forming an image on a fluorescent screen. Also see FIELD-EMISSION MICROSCOPE.

field length Record field size in applicable units, usually bytes, characters, or words.

field magnet **1.** The permanent magnet in a dynamic speaker. **2.** A similar magnet in an earphone, generator, microphone, motor, phono pickup, transducer, etc.

field pickup **1.** A probe or sensor for insertion into an electric or magnetic field. **2.** An on-location radio or television program (i.e., one coming from outside the studio). Also called remote or NEMO.

field resistor A resistive component consisting of an insulated form with a thin layer of conductive material.

field rheostat The rheostat whose setting determines the amount of current flow through the field coil of a motor or generator.

field scan A form of television scanning in which the lines are scanned alternately.

field-sequential system A color-television system in which the image is reproduced by means of primary color fields (red, green, and blue) flashed sequentially on the screen of the picture tube. Compare DOT-SEQUENTIAL SYSTEM and LINE-SEQUENTIAL SYSTEM.

field strength See FIELD INTENSITY.

field-strength meter An instrument for measuring the radio-frequency (RF) voltage of a signal reaching a chosen location. The instrument consists essentially of a radio detector equipped with a portable antenna and an output meter.

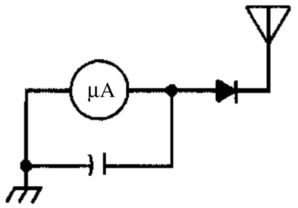

field-strength meter

field telephone A rugged, portable telephone system for outdoor use.

field test A test of equipment under actual operating conditions (i.e., outside the laboratory or factory).

field winding See FIELD COIL, **1**, **2**.

fig. Abbreviation of *figure* (usually in reference to an illustration).

figure-8 pattern **1.** A bidirectional antenna pattern whose cross section, in a specified plane, resembles the numeral 8. **2.** A Lissajous figure resembling the numeral 8.

figure of merit **1.** For a capacitor, inductor, or tuned circuit, the ratio of reactance to resistance. Also called *Q factor* or *Q.* **2.** For a magnetic amplifier, the ratio of power amplification to control time constant. **3.** For a transistor, the *gain-bandwidth factor.*

fil Abbreviation of FILAMENT.

filament In a vacuum tube or incandescent lamp, the thin wire heated by electric current; it emits electrons, light, and heat. The filament is the cathode in a filament-type tube, but serves only to heat indirectly the cathode sleeve in an indirectly heated tube.

filament battery See A BATTERY.

filament choke A radio-frequency (RF) choke operated in the filament lead of an electron tube. Such chokes are necessary in filament-type tubes in RF power amplifiers.

filament circuit The circuit carrying filament current.

filament coil See FILAMENT WINDING.

filament current Symbol, I_F. The current flowing through the filament of a vacuum tube.

filament emission Electrons emitted directly by the filament in an electron tube (thermionic emission) or the amount of such emission.

filament hum A hum signal caused by voltage induced in a circuit by the alternating-current (ac)-operated filaments (heaters) of tubes or by the filament wiring.

filament lag The time delay in the heating and cooling of an ac tube or lamp filament as filament-current changes polarity.

filament power supply A source of power, usually alternating-current, for heating the filament of a vacuum tube or tubes.

filament resistance **1.** The resistance of the filament in an electron tube or incandescent lamp. **2.** The resistance of an external dropping resistor in the filament circuit of a tube or lamp.

filament transformer A step-down transformer that supplies power exclusively to the filament (heater) of an electron tube.

filament-type bolometer A BOLOMETER in which the sensitive element is a wire filament. Examples: barretter, incandescent lamp, and wire fuse.

filament voltage Symbol, V_F. The voltage across the filament of an electron tube.

filament winding In a power transformer, the coil that supplies heating power to the filament of a vacuum tube. Also called FILAMENT COIL.

file An organized collection of records related by a common format, data source, or application.

file conversion Converting data files from one form to another, often for the purpose of making them compatible with other computers.

file gap An area of a data medium that signifies the end of a file; it can also mark the start of another file.

file identification A code that identifies a file.

file label File identification in which the first record in the file is a set of characters unique to the file; it conveys such information about (for example, a tape file as a description of content, generation number, reel number, date of writing, etc.). Also called HEADER LABEL.

file layout How the contents of a file are organized; usually defined by the system or specified by a program.

file maintenance To delete, add, or correct records in a file. Unlike updating, which is done to reflect changes in events recorded in the file, maintenance ensures that the contents of the file are accurate records of the necessary data.

file management A method of storing and recalling data from computer storage media, such as magnetic disks.

filename In a computer file label, the alphanumeric character set that identifies and describes the file. It generally consists of one to eight characters, often including an extension of one to three characters. In some computer operating systems, filenames can be much longer (e.g., up to 32 characters). See FILENAME EXTENSION.

filename extension A group of one, two, or three characters following the main body of a FILENAME, and separated from it by a period (.). It generally denotes the application or purpose of the file. The extension .BAK, for example, denotes a backup file; .TXT denotes a text file; .BMP denotes a bit-mapped graphics file.

file organization The way words, bits, or records are physically arranged in the storage medium for a file, possibly including the method of access (serial, alphabetical, random, etc.).

file-oriented programming Computer programming that uses a general file and record-control program to simplify I/O coding.

file-oriented system A system having file storage as its basis.

file print A hard copy (printout) of the contents of a file.

file processing The operations associated with making and using files.

file protection Preventing the possibility of writing over data files before they are made available for use. It is usually done by having a program check file labels.

file reconstitution Restoring a partially or completely damaged file by updating a previous generation of the file using a file of interim transactions.

file recovery Following the interruption of file processing because of system failure, the procedure for reestablishing the file's condition as necessary for the resumption of processing without losing accuracy.

file section Part of a file in certain consecutive locations on a storage medium.

file security Protective and security measures (e.g., the issuance of clearances, status markers, etc.) as they relate to computer files.

file set A collection of interrelated files stored consecutively in a magnetic disk volume (package).

fill The percentage of lines in a cable that are actually in use at a given time.

filler A nonessential data part used, for example, to bring a record to a standard size.

film See THIN-FILM MEMORY.

film capacitor A capacitor in which the electrodes are plated or deposited on the faces of a thin film of plastic or other dielectric material.

film chain A system designed for the transmission of movies over a television system. This requires synchronization of the movie frame rate with the television scanning rate.

film frame A single picture on a strip of motion picture film.

film-frame blanking interval The interval during which a film frame is blanked out as motion-picture film moves through a camera, projector, or pickup. The blanking action allows a frame to move into position without blurring the image as seen by viewers.

film integrated circuit A monolithic circuit whose elements are films formed on an insulating substrate.

film pickup A photocell, photodiode, or phototransistor circuit used to pick up recordings from the sound track of motion-picture film.

film reader A device for converting data on film into digital form for a computer.

film recorder An apparatus that records data as a sound pattern on film. Compare FILM REPRODUCER.

film reproducer An apparatus that plays back data recorded on photographic film. Compare FILM RECORDER.

film resistor Also called *carbon-film resistor* or *metal-film resistor.* An electronic component consisting of a mixture of ceramic, carbon, and/or metal, applied to a cylindrical form in a thin layer to obtain a desired value of resistance. The cylindrical form is made of an insulating substance, such as porcelain. The film can be deposited on this form by various methods, and the value tailored as desired. Metal-film units can be made to have nearly exact values. Film type resistors usually have low to medium-high resistance. A major advantage of film-type resistors is that they, like carbon-composition units, do not have much inductance or capacitance. A disadvantage, in some applications, is that they cannot handle as much power as the more massive carbon-composition units, or as wirewound types. Compare CARBON-COMPOSITION RESISTOR, WIREWOUND RESISTOR.

film scanning The conversion of a movie into a form suitable for transmission by television.

film speed **1.** The speed at which motion-picture film moves intermittently through a camera, projector, or pickup, measured in feet or frames per second. **2.** A measure of film's light sensitivity, given as an ASA (American Standards Association) or DIN (European) number; in either system, the higher the number, the greater the light sensitivity.

filter **1.** A circuit or device that passes alternating currents at some frequencies while attenuating or blocking currents at other frequencies. Examples: BANDPASS FILTER, BAND-REJECTION FILTER, HIGH-PASS FILTER, and LOW-PASS FILTER. **2.** An inductance-capacitance (LC) or resistance-capacitance (RC) circuit for removing the ripple from the output of a power-supply rectifier. **3.** A transparent disk with special optical properties placed in front of a camera lens for a special photographic effect. **4.** A character patterned to control the elimination or selection of characters in another pattern. **5.** A device or program that separates information, according to certain specifications or characteristics. **6.** A machine word that specifies the elements to be treated in another machine word; also called MASK.

filter attenuation In a selective filter, the power, current, or voltage loss, in decibels, that occurs within the passband.

filter attenuation band In a selective filter, the frequency band(s) outside the passband; that is, the frequency range over which signals are significantly attenuated.

filter bank In audio applications, a set of bandpass filters, each of which covers a specific portion of the audio-frequency (AF) spectrum. There is some (minimal) overlap between the passbands of the filters, so an AF signal of a specific frequency will always pass through at least one of the filters, but will never pass through more than two of the filters.

filter capacitor A capacitor that provides capacitive reactance in a wave filter or power-supply filter while also blocking direct current.

filter center A place where information is modified for transmission to aircraft pilots. Such information can include weather data, course changes, or other instructions.

filter choke An inductor that provides inductive reactance in a wave filter or power-supply filter while affording relatively easy conduction of direct current.

filter crystal A piezoelectric crystal used in a CRYSTAL RESONATOR.

filter cutoff The frequency or frequencies at which the transmission figure of a filter is below its maximum value by a prescribed amount, usually 3 dB, representing the half-power point(s).

filter discrimination The amount of fluctuation in the insertion loss of a bandpass, band-rejection, high-pass, or low-pass selective filter. The fluctuation is measured at various points in the filter passband.

filter inductor See FILTER CHOKE.

filter passband The frequency range over which a selective filter passes signals with minimum attenuation.

filter reactor See FILTER CHOKE.

filter slot In a waveguide, a slot that acts as a choke to suppress undesirable modes.

filter stopband See FILTER ATTENUATION BAND.

filter transmission band See FILTER PASSBAND.

fin **1.** A metal disk or plate attached to a component for the purpose of radiating heat. **2.** A projection in an irregular heatsink.

final amplifier Also called *output amplifier* or *power amplifier.* In a wireless transmitter, the amplifier that boosts the signal power to the level to be emitted over the airwaves. The input is provided by the driver; the output is connected to the system, to a transverter, or to a network cable power amplifier. In some transmitters, the output amplifier requires no tuning; in others, adjustment of the output circuitry is necessary to obtain proper coupling to the antenna system. In a single-sideband (SSB) or amplitude-modulated (AM) transmitter, all amplifiers following the modulator must be linear. In transmitters intended for use only in digital modes or for frequency modulation (FM), power amplifiers need not be linear.

final result A result displayed at the end of a data processing operation. Compare INTERMEDIATE RESULT.

finder The switch or group of relays that selects the path for a call going through a telephone switching system. Also called LINE FINDER.

fine adjustment Adjustment of a quantity in small increments or as a smooth, continuous variation. Compare COARSE ADJUSTMENT.

fine-chrominance primary See I SIGNAL.

fine frequency control A variable component, such as a potentiometer or variable capacitor, that permits a signal or response frequency to be varied over a small increment; it is often used in conjunction with a coarse frequency control.

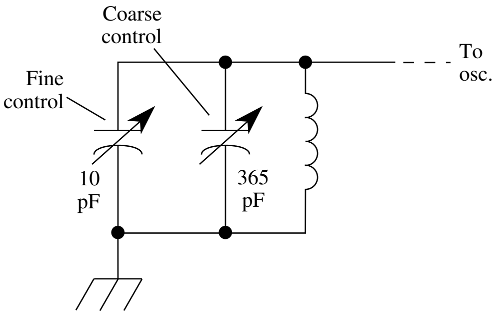

fine frequency control

fine-groove record See MICROGROOVE RECORD.

fine index In computer operations, a secondary, supplemental index used with a main, or gross, index when the latter does not adequately detail the differences between the items being indexed.

fine-motion planning The scheme used by a robotic end effector ("hand") to get in exactly the correct position. It uses machine vision, tactile sensors, pressure sensors, or high-precision displacement transducers.

finger See FEELER, **1**.

finger plethysmograph A device that senses and records the resistance through the human finger during various parts of the heart cycle.

finger rules See FLEMING'S LEFT-HAND RULE, FLEMING'S RIGHT-HAND RULE, and RIGHT-HAND RULE FOR WIRE.

finish lead The lead attached to the last turn of a coil. Also called *outside lead.* Compare START LEAD.

finished blank The end product in the crystal manufacturing process, often including electrodes.

finishing The careful handwork and testing involved in bringing a crystal blank to a condition that is acceptable as finished, according to specifications.

finishing rate The rate of charging a battery, as the battery approaches a full charge. Generally, the finishing rate is less than the normal charging rate.

finite Pertaining to that which has defined limits.

finite sample space In statistics, a sample space having definite limits.

finite series A mathematical series having a limited number of terms. Example: $1 + 0.1 + 0.01 + 0.001$, a finite series containing four terms and whose sum is equal to 1.111.

finned surface The irregular surface of a heatsink. The ratio of surface area to volume is greater than with a flat surface; this increases the cooling effectiveness of the heatsink.

fins Metal vanes radiating from components that dissipate large amounts of power as heat.

FIR Abbreviation of FAR INFRARED.

fire A transition from non-conduction to conduction in an ionizing switching device.

fire control The aiming and firing of guns automatically via radar and associated electronic systems.

firefighter robot A robotic device or system intended to protect lives and property from fire. Such a robot can be simple, such as an automatic sprinkler system actuated via heat sensors and/or smoke detectors. More sophisticated systems might use autonomous or remotely controlled androids that perform the same functions as human firefighters.

Fire Underwriter's regulations See NATIONAL ELECTRIC CODE.

firing The pulse that initiates conduction in an ionization switching device.

firing angle **1.** For a magnetic amplifier, the angular distance through which the input-voltage vector rotates before the core is driven into saturation. **2.** For a silicon-controlled rectifier (SCR), the point, as an angle (in degrees or radians), along the control voltage half-cycle at which the SCR fires.

firing circuit Any circuit, such as a phase shifter, that permits adjustment of the firing angle of a silicon-controlled rectifier (SCR) or similar device, or which delivers the required pulse or other signal to initiate firing.

firmware **1.** Programs (software) in nonvolatile computer memory [e.g., in a read-only memory (ROM), that can only be changed by replacement with an alternate unit]. **2.** Unalterable internal interconnections that determine what a computing device or system can do. Also called MICROPROGRAM.

first detector In a superheterodyne circuit, the signal frequency detector. Compare SECOND DETECTOR.

first filter capacitor The input capacitor in a capacitor-input power-supply filter.

first Fresnel region A portion of a directional transmitted electromagnetic ray, shaped generally like a paraboloid with the apex at the transmitter and the axis in the direction of transmission. Any point in the first Fresnel zone is in such a position that the sum of the lengths of the paths from the point to the receiver, and the point to the transmitter, is no greater than 0.5 wavelength more than the distance from the transmitter to receiver.

first harmonic The fundamental frequency in a complex waveform from which multiples are generated.

first-in/first-out Acronym, FIFO. A read-write memory commonly used as a buffer to smooth out the flow of data bits in a digital system. The output bits are in the same order (sequence) as the input bits. If a data bit *x* goes into the FIFO before data bit *y*, then *x* will always emerge from the FIFO before *y*. Compare PUSHDOWN STACK.

first-in/last-out See PUSHDOWN STACK.

first law of thermodynamics Quantities of heat can be converted into mechanical work, and vice versa. Also see MECHANICAL EQUIVALENT OF HEAT.

first level address See ABSOLUTE ADDRESS.

first selector The selector that responds to the first-digit dial pulses when a telephone number is called.

fishbone antenna An untuned, wideband directional antenna of the general end-fire type. Consists of a number of collector antennas, each capacitively coupled to the resistor-terminated transmission line in collinear pairs. It is so called from its resemblance to the skeleton of a fish.

fishpaper A chemically treated, vulcanized-fiber paper used for electrical insulation.

fish pole A microphone boom that can be held in the hand.

fist In radiotelegraphy and wire telegraphy, an operator's manual sending style.

fitting A device intended to mechanically fasten a wire or cable in place.

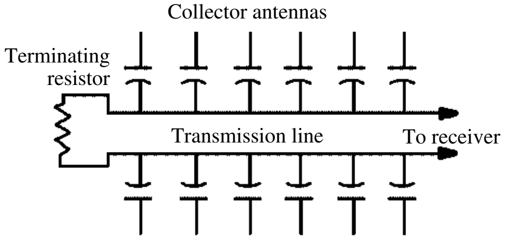

fishbone antenna

five-element code A five-impulse telegraph code that describes a character (e.g., Baudot, which also includes start and stop elements).

five-layer device A semiconductor device containing four pn junctions. Examples: diac and triac.

five-level code A teletype code that utilizes five binary elements to define a character.

fix **1.** In direction finding, the point at which two lines of direction intersect. **2.** In electronics maintenance, to repair successfully. **3.** To subject an in-process circuit board to a solution or other medium to stop a photographic action permanently. Also, the solution in which such photosensitive materials can be immersed to halt development.

fixed bias Bias voltage or current supplied from a fixed external source (such as a battery or power supply) independent of the operation of the biased device. Compare AUTOMATIC BIAS.

fixed block length Blocks of data having a fixed number of words or characters, as required because of hardware limitations or a program instruction. Compare VARIABLE BLOCK LENGTH.

fixed capacitor A nonadjustable capacitor (i.e., one having a single unalterable value).

fixed component Any component (e.g., a capacitor, inductor, or resistor that has one unalterable value).

fixed contact The stationary contact in a relay or switch. Compare MOVABLE CONTACT.

fixed-crystal detector A simple crystal detector in which the point of the contact wire is permanently placed in contact with a sensitive spot on the surface of the crystal.

fixed field Fields in records organized so that those containing similar information in each record are the same length and in the same relative position in the record. Compare VARIABLE FIELD.

fixed form coding The coding of source languages so that each part of the instruction is in a fixed field.

fixed-frequency amplifier An amplifier that is pretuned to operate on one frequency, or in a relatively narrow band of frequencies.

fixed-frequency oscillator An oscillator that is preset to operate on one frequency. Such an oscillator can be self-excited or controlled (crystal, fork, magnetostriction, etc.).

fixed-frequency receiver A receiver that is pretuned to receive signals of one frequency.

fixed-frequency transmitter A transmitter that is pretuned to radiate signals of one frequency. Such a transmitter can contain a self-excited or crystal-controlled oscillator.

fixed inductor A nonadjustable inductor (i.e., one having an unalterable value of inductance).

fixed-length record A record in which word or character size is constant. Compare VARIABLE LENGTH RECORD.

fixed logic Applicable to computers or peripherals whose logic can only be altered internally by changing connections.

fixed memory A nonvolatile readout computer memory that can only be altered mechanically.

fixed placement file A file that has been allocated a fixed location in storage.

fixed-point system A notation system in which a single set of digits represents a number, and the radix point (in the decimal system, the decimal point) can only be placed in one position for the value being expressed. Also see FLOATING-POINT CALCULATION.

fixed resistor A nonadjustable resistor (i.e., one having an unalterable value of resistance).

fixed-sequence robot A robot that performs one task or set of tasks, making exactly the same movements every time. The sequence is programmed in the robot controller via firmware. Many assembly robots and toy robots are of this type. Compare FLEXIBLE AUTOMATION.

fixed station A radio station operating from a stationary point; one that is not mobile.

fixed-step potentiometer A potentiometer whose output is varied in one or more discrete steps by fixed-resistor sections. Also see POTENTIOMETER.

fixed wireless 1. A cellular communications network in which some, or all, of the end users are at fixed locations. This is increasingly popular as an alternative to conventional hard-wired telephone and Internet access. 2. The use of wireless equipment, such as cell phones and wireless-modem–equipped computers, at fixed (rather than mobile or portable) locations.

fixed word length Applicable to the organization of information in storage in which each computer word stored has a fixed number of characters or bits.

fixture A piece of hardware used in equipment setups (e.g., microwave couplers, joints, sections, etc.).

fL Abbreviation of FOOT-LAMBERT.

flag **1.** A piece of information added to a data item that gives information about the data item. **2.** A bit added to a character or word to delineate a field boundary. **3.** An indication that an operation is complete and need not be done by the program. **4.** An indicator identifying the members of mixed sets. **5.** A character that signals the presence of some condition (e.g., an *error flag* indicates that a data item caused an error).

flag event A program condition that causes a flag to be set.

flag line An input pulse to a microprocessor that depends on specific external instructions. Indicates a certain condition or change of state.

flagpole antenna **1.** Any of several vertical UHF or VHF antennas consisting of a radiator mounted atop a coaxial pipe or cable (see, for example, COAXIAL ANTENNA). It takes its name from its resemblance to a flagpole. **2.** A vertical antenna formed by shunt-feeding a pole already in existence, such as a flagpole. This can be done to conceal the antenna.

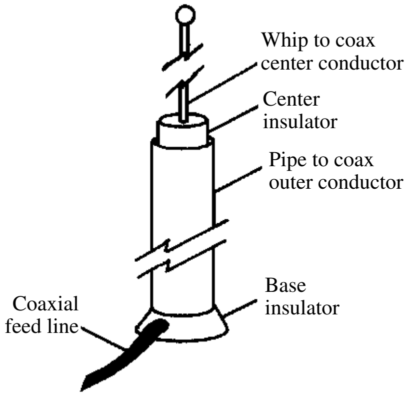

flagpole antenna, 1

flag terminal A form of terminal that does not require soldering for electrical contact. A protruding "flag" is crimped around the conductor.

flame alarm A (usually photoelectric or thermoelectric) device or circuit for detecting a flame and actuating an alarm.

flame control A (usually photoelectric or thermoelectric) device or circuit for sensing and automatically controlling the height of a flame, such as a gas pilot.

flame-failure control A FLAME CONTROL that automatically cuts off the fuel if the flame goes out.

flame microphone A microphone in which two electrodes in a flame undergo a change in electrical resistance when the flame is influenced by sound waves, thus modulating current passing between the electrodes.

flange **1.** A flat, protruding edge used for fastening a connector or plug to the chassis of a piece of equipment. **2.** In a waveguide, a coupling used for connection to another section of waveguide, or to a horn or other external device.

flange focus The focal length of a lens, based on the distance from the mounting flange to the focal plane.

flanging In sound systems, a means of obtaining special effects using a COMB FILTER. Creates an "extraterrestrial," digitized sound. Commonly used by rock music bands.

flanking Modification of the response of a selective filter, resulting from the parallel connection of two or more similar filters.

flap See DROP INDICATOR.

flap attenuator A waveguide attenuator consisting of a sheet of resistance material inserted transversely into the waveguide through a slot.

flare **1.** The hyperbolic cross section of a horn antenna or loudspeaker. **2.** A transient or stationary bright area with (usually) a central pip on the screen of an otherwise blank oscilloscope or television picture tube.

flare angle The gradual change in a waveguide's diameter over its length.

flare factor The angle at which the faces of a horn speaker are curved or turned outward.

flaring constant Symbol, m. A number expressing the degree of flare (see FLARE, **1**) in a horn antenna or loudspeaker. The value of m is equal to $0.6931/d$, where d is the distance along the axis required for the cross-sectional area A of the horn to double. If d is given in feet (ft), for example, then A is in units of square feet (ft^2) and m is in units of reciprocal feet (1/ft). If d is in centimeters (cm), then A is in units of square centimeters (cm^2) and m is in units of reciprocal centimeters (1/cm).

flash **1.** A photographic camera flash. **2.** To vaporize a metal (such as magnesium) in an electron tube being evacuated, to absorb gases. **3.** Flashover.

flash arc In a vacuum tube, a sudden high-current arc between cathode and plate at high plate voltages; it can short-circuit the plate power supply.

flashback voltage The maximum inverse voltage that causes the gas in a tube to ionize.

flash delay A device that automatically postpones the operation of a FLASHTUBE until a predetermined instant, such as the moment when a moving object arrives at a particular point before a camera.

flasher An electrical or electronic device or circuit that flashes a light or a series of lights sequentially.

flasher LED A light-emitting diode (LED) that, when connected to a low-voltage direct-current (dc) source, emits light that flashes at a basic rate of a few pulses per second.

flashlamp **1.** See FLASHTUBE. **2.** A small portable light operated from self-contained cells; a flashlight or lantern.

flashlight See FLASHLAMP, **2**.

flash memory High-capacity, nonvolatile random-access memory, used in some computers in place of an electromechanical hard drive. This type of memory allows a computer to store, access, and run large data files or programs much faster than is possible with a conventional hard drive. See also NONVOLATILE MEMORY, RANDOM-ACCESS MEMORY.

flashover The sudden discharge of electrical energy between electrodes or conductors, often accompanied by light; it is usually the result of excessive voltage.

flashover voltage **1.** The peak voltage at which FLASHOVER occurs. **2.** The voltage at which disruptive discharge occurs between electrodes and across the surface of an insulating material.

flash plating Electroplating in which a thin layer is deposited quickly.

flash test Insulation testing by applying a higher-than-normal voltage for a short time.

flashtube A straight or coiled glass tube filled with gas and provided with electrodes. When a high voltage is applied to the electrodes, the tube emits a brilliant flash of light.

flat cable Also called *ribbon cable.* A cable whose flexible conductors are molded side by side in a flexible, flat ribbon of plastic (such as polyethylene).

flat-compounded generator A compound-wound generator whose windings are proportioned so that the full-load and no-load voltages are identical.

flat fading Fading of a radio signal that occurs independently of frequency; all frequency components of the signal fade to the same extent at the same time.

flat file A computer file containing unfolded documents.

flat frequency response Relatively equal response to all fixed-point frequencies within a given spectrum, exhibited by an amplifier or other circuit that must transmit a band of frequencies.

flat line A transmission line in which there are no standing waves, or for which the standing-wave ratio is very low.

flat pack An integrated circuit package consisting of a square or rectangular flat housing, with pins projecting straight outward from the edges.

flat response A response characteristic in which the dependent variable is substantially constant over a specified range of values of the independent variable. For example, in amplifier operation, an output signal whose component fundamental frequencies and their harmonics are in the same proportion as those of the input signal being amplified.

flat-ribbon line A transmission line (feeder) consisting of two flexible conductors molded in a flexible, flat ribbon of plastic, such as polyethylene. Also called *twinlead.*

flattening The leveling-off or blunting of a normally peaked or curved response, often caused by signal saturation within a circuit. Sine-wave clipping is an example.

flat top **1.** The horizontal radiating portion of an antenna. **2.** See FLAT-TOP ANTENNA. **3.** Of an amplifier, to distort by clipping of the positive half-cycles.

flat-top antenna An antenna having a horizontal wire or wires at the top to lower the resonant frequency and increase the bandwidth.

flat-top beam A bidirectional, end-fire antenna consisting of two close-spaced dipoles center-fed out of phase. Also see KRAUS ANTENNA.

flat-topping The positive-peak clipping of a modulation envelope that occurs when an amplifier is overdriven or when a signal is overmodulated. This is in contrast to deliberate clipping. Flat-topping results in "splatter" and is, therefore, an undesirable characteristic in an amplitude-modulated signal.

flat-top response The ability to uniformly transmit frequencies in a given band.

flat transmission line **1.** A transmission line that is free of standing waves. Also see MATCHED TRANSMISSION LINE. **2.** See FLAT-RIBBON LINE.

flaw An irregularity in a substance that can result in problems, such as mechanical failure or poor equipment performance.

flaw detector An instrument that uses ultrasonic waves to detect internal flaws in metal. The waves are reflected by flaws.

F layer See F REGION.

fLb Abbreviation of FOOT-LAMBERT.

Fleming/Kennelly law Under conditions at or near magnetic saturation in a ferromagnetic substance, the reluctivity is directly proportional to the intensity of the magnetic field.

Fleming's generator rule See FLEMING'S RIGHT-HAND RULE.

Fleming's left-hand rule A simple way of indicating certain relationships in the behavior of electric generators and motors. If the thumb, index finger, and middle finger of the left hand are positioned so that they are at right angles to each other, the thumb will point in the direction of force or motion when the index finger is pointed in the direction of flux; the middle finger points in the direction of current flow. Compare FLEMING'S RIGHT-HAND RULE.

Fleming's motor rule See FLEMING'S LEFT-HAND RULE.

Fleming's right-hand rule A simple way of indicating certain relationships in the behavior of electric generators and motors. If the thumb, index finger, and middle finger of the right hand are positioned so that they are at right angles to each other, the middle finger points in the direction of an induced voltage, the thumb in the direction of the motion of a conductor, and the index finger in the direction of the magnetic field. Compare FLEMING'S LEFT-HAND RULE.

Fletcher-Munson curves A set of curves depicting the uneven frequency response of human hearing. Also called AUDIBILITY CURVES.

flexible automation The ability of a robot or system to do various different tasks. Changing from one task to another is simply a matter of changing software. Compare FIXED-SEQUENCE ROBOT.

flexible collodion A viscous solution of pyroxylin (cellulose nitrates) used sometimes as a binder for coils.

flexible contact A contact made from flat, metal spring stock; it is usually bent or curved. Also called *spring contact.*

flexible coupling A device for joining two shafts and conveying rotary motion from one to the other; it is elastic, so the shafts need not be exactly aligned with each other.

flexible flat cable See FLAT CABLE.

flexible manufacturing system A roboticized manufacturing plant that can turn out a variety of different products. One or more central computers oversee the operation of the facility. Such factories are commonly used in the production of electronic devices (such as printed circuits, calculators, and portable radios).

flexible resistor An insulated, wirewound resistor that can be bent, coiled, or knotted.

flexible shaft A control shaft that can be bent somewhat while still allowing easy adjustment.

flex life A measure of how much bending a conductor or other flexible object can take without breaking.

flexode A diode that is flexible in that its junction can be changed (i.e., reversed without reversing its leads, its resistance being variable from the forward- to backward-resistance value).

flicker **1.** A tendency for a video image to appear, disappear, and reappear, or to increase and decrease in intensity frequently. **2.** The effect created by such action (as in a flickering light).

flicker frequency The number of times the screen illumination flashes on and off in the projection of a motion picture. It is 48 per second (twice the frame rate) in conventional movie projectors; for each frame, the screen is blanked once when the frame is pulled into position and once again during projection of the frame.

flight control Electronic monitoring and control of an aircraft in flight.

flight path The course planned for an aircraft's flight.

flight-path computer A computer that controls the course of an aircraft in flight, from takeoff to landing.

flight-path deviation The departure of an aircraft in flight from the course in the flight plan. Also see FLIGHT PATH.

flight-path-deviation meter An instrument that provides a visual indication of the departure of an aircraft in flight from the course in the flight plan.

flight telerobotic servicer A remotely controlled robot used to maintain and repair space vehicles and satellites. The machine can be controlled by a computer program or by a human operator. The most-sophisticated machines can use TELEOPERATION and/or TELEPRESENCE so that a human can perform dangerous work without being placed at personal risk.

flight test **1.** To test airborne electronic equipment in actual flight. **2.** Any test made as in **1**.

Flinders bar In a magnetic compass, a metal bar that corrects for the vertical component (inclination) of the earth's magnetic field. The bar must be designed differently in different geographic locations, because the inclination varies from place to place. Inclination is greatest near the geomagnetic poles, and is zero at the geomagnetic equator.

flint glass A hard, bright, lead glass. Dielectric constant, 7 to 9.9. Dielectric strength, 30 to 150 kV/mm. Also see GLASS.

flip chip A monolithic semiconductor device (such as a diode, transistor, or integrated circuit), in which bead-like terminals are provided on one face of the chip for bonding.

flip-chip bonding A scheme for making connections between a semiconductor chip and a header, in which leads are not run between chip and header. Instead, bead-like projections are electrodeposited as terminals around one face of the chip, which is then registered with the header terminals and bonded to them.

flip-flop **1.** See BISTABLE MULTIVIBRATOR. **2.** A two-position relay that locks in alternate positions upon receiving successive actuating pulses.

flip-flop key In a video display, a key that, when pressed, allows viewing of one half of the screen and then the other.

flip-flop memory A bistable computer memory that stores bits of data as flip-flop states.

flip-flop relay See BISTABLE RELAY.

float charging The constant charging of a storage battery, keeping the battery at or near the fully charged state at all times.

floated battery A storage battery connected in parallel with a generator, which supplies the load; the battery, always completely charged, helps during high-current demands.

floating **1.** To float a storage battery; see FLOAT. **2.** An ungrounded device or circuit that is not connected to a source of voltage. **3.** Not loaded or driven. **4.** Not fixed in position. **5.** A dedicated ground connection that remains isolated from the common circuit ground.

floating address See RELATIVE ADDRESS.

floating charge See TRICKLE CHARGE.

floating control **1.** A potentiometer, such as a gain control, installed with its shaft insulated from ground and, accordingly, subject to body-capacitance effects. **2.** A type of automatic control in which the rate of final control element movement depends on the amount that the controlled variable deviates from a prescribed value.

floating ground See FLOATING, **5**.

floating input An ungrounded input circuit.

floating-input measurement See DIFFERENTIAL-INPUT MEASUREMENT.

floating instrument An instrument whose signal terminals are above ground.

floating I/O port An input/output (I/O) terminal that is not loaded or being driven.

floating junction A junction (in a semiconductor device, for example), that has no net current flowing through it.

floating neutral A circuit with a variable common voltage reference.

floating paraphase inverter A dual-transistor adaptation of the paraphase inverter. The second stage receives its input signal from a tap on the load resistor of the first stage and provides the additional phase shift that is required.

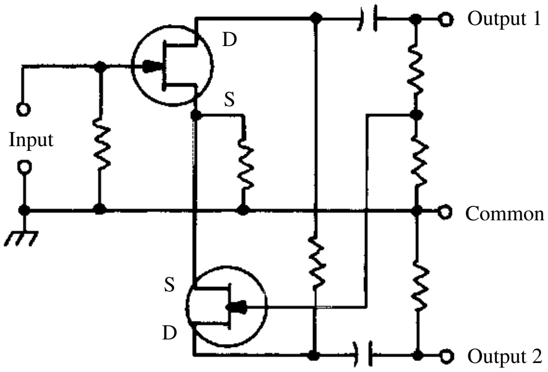

floating paraphase inverter

floating-point calculation An electronic calculation using a floating point number [i.e., a number whose value is represented by two sets of digits, a fixed-point part (see FIXED-POINT SYSTEM) and a radix (base number) with an exponent]. Important in the operation of microprocessors, especially in computer systems.

floating-point number A number expressed in the floating-point system.

floating-point package Computer-vendor software that enables that computer to perform floating-point calculations.

floating-point system A system of notation in which a number n is represented by two sets of numbers: a fixed-point part (see FIXED-POINT SYSTEM) a, the radix (base number) r, and an exponent b as follows: $n = a \times r^b$. For example, in the floating-point system, 623 can be written 6.23×10^2. Floating-point numbers can be stored economically (in terms of memory) and in magnitudes that might otherwise be beyond the capacity of the computer to operate upon with relatively consistent accuracy. Compare FIXED-POINT SYSTEM.

floating probe A test electrode (wire or plate) inserted in a discharge tube at a desired point to sample the potential gradient, but which acquires a misleading negative charge, with respect to the gas cloud, because electrons (traveling faster than the positive ions) tend to accumulate on the probe.

floating zero A control system in which the reference point is easily moved.

floating zone In a semiconductor ingot undergoing purification, a molten zone in which impurities float. The material in the zone is melted by the radio-frequency (RF) field of an external heating coil, which is passed along the ingot to move the molten zone to one end, picking up impurities along the way and concentrating them at the end that is later sawed off.

float switch A switch operated by a float, such as in a sump pump.

flocking **1.** Particulate felt used on phonograph turntables to protect disks from being scratched. **2.** To coat with flocking.

flood gun In a storage (image-holding) oscilloscope, the electron gun that sprays the storage target with low-velocity electrons and makes the image visible on the viewing screen. The gun is mounted next to one pair of deflection plates. Compare WRITING GUN.

floor stand A support for a microphone, consisting of a heavy base that rests on the floor, and an adjustable, vertical boom that allows the microphone to be set at various heights.

floppy disk A flexible magnetic disk used in recording, as in computer and data system storage. It usually refers to a 5.25-inch diskette.

flow **1.** The movement of current carriers under the influence of an electric field. **2.** See ANGLE OF CONDUCTION. **3.** A series of interrelated events in a time sequence.

flow angle See ANGLE OF CONDUCTION.

flowchart **1.** A diagram depicting the logic steps in a digital-computer program. **2.** A diagram showing the flow of material through a sequence of processes.

flow direction The method of delineating antecedent and successor events on a flowchart; usually arrows and flowlines connecting the events in the way a page is read (top to bottom, left to right).

flowed-wax disk A form of recording disk, in which wax is melted onto a plastic or metal base. The grooves are cut in the wax layer.

flowline A line showing flow direction on a flowchart.

flowmeter An instrument for measuring liquid flow rate.

flow relay A relay that is actuated by a predetermined rate of fluid flow.

fluctuating current See COMPOSITE CURRENT.

fluctuating voltage See COMPOSITE VOLTAGE.

fluid absorption See LIQUID ABSORPTION.

fluid analogy The comparison of electric current flow to the movement of a simple fluid. Also see WATER ANALOGY.

fluid capacitor See WATER CAPACITOR.

fluid computer A digital computer that uses fluid logic elements (i.e., one that contains no electronic circuits or moving parts).

fluid damping Use of a viscous fluid to damp a mechanical member's movement.

fluid-flow alarm An electronic circuit that actuates an alarm when fluid flowing through pipes or other channels changes from a predetermined rate.

fluid-flow control A servo system that automatically maintains or adjusts liquid flow through pipes or other channels.

fluid-flow gauge See FLUID-FLOW METER.

fluid-flow indicator See FLUID-FLOW METER.

fluid-flow meter An instrument that indicates fluid flow rate through pipes or other channels.

fluid-flow switch In a fluid-cooled system, a switch that actuates an alarm when the fluid slows or stops.

fluidics **1.** A form of digital logic in which circuits operate by means of fluid flow. **2.** A branch of physics concerned with the behavior of fluids; more commonly called *fluid dynamics.*

fluid-level control A servo system that automatically maintains the level of a fluid in a tank.

fluid-level gauge An electronic system that provides direct readings of the level of a fluid in a tank.

fluid-level indicator See FLUID-LEVEL GAUGE.

fluid logic Logic operations carried out by varying the flow and pressure of a gas or liquid in a circuit of channels. Also see FLUID COMPUTER.

fluid ounce (U.S.) Abbreviation, fl. oz. A unit of volume equal to 2.957×10^{-5} cubic meters, or 0.02957 liter. A quart is 32 ounces; a gallon is 128 ounces.

fluid-pressure alarm An electronic circuit that actuates an alarm when fluid pressure rises or falls beyond set limits.

fluid-pressure control A servo system that automatically maintains or adjusts fluid pressure in pipes or other channels.

fluid-pressure gauge See FLUID-PRESSURE METER.

fluid-pressure indicator See FLUID-PRESSURE METER.

fluid-pressure meter An instrument that indicates the pressure of a fluid in a pipe or other channel.

fluid valve See ELECTROMECHANICAL VALVE.

fluorescence The property of some materials to glow when excited by a stimulus, such as ultraviolet, X rays, or an electron beam. Compare PHOSPHORESCENCE.

fluorescent lamp See FLUORESCENT TUBE.

fluorescent materials Materials that glow when irradiated, but cease to glow when the source of excitation is removed. An example is the phosphor coating on the screen of a cathode-ray tube (CRT).

fluorescent screen A transparent or translucent plate (such as the end of a cathode-ray tube or fluoroscope) coated with phosphors that glow when struck by an electron beam, or by high-energy electromagnetic radiation, such as ultraviolet or X rays.

fluorescent tube A mercury-vapor glow lamp distinguished by having a glass tube whose inner wall is coated with a phosphor that emits light when excited by the ultraviolet glow discharge in the vapor.

fluorescent X rays X rays reradiated by the atoms of a material that has absorbed X radiation. During initial exposure, energy absorbed from the radiation raises the energy level of electrons in the atoms; when the electrons return to their normal energy levels, they reradiate some of the absorbed energy.

fluorine Symbol, F. A gaseous element of the halogen family. Atomic number, 9. Atomic weight, 18.998.

fluoroscope A device used for viewing the internal structures of objects. A screen coated with material that fluoresces when exposed to X rays is mounted in one end of a light-tight viewing hood. When an object is placed between the screen and an X-ray tube, an image is produced on the screen. In medical applications, this device has been supplanted by methods that do not use ionizing radiation; *nuclear magnetic resonance imaging (NMRI)* is one example.

fluoroscopy The art of using a fluoroscope in the inspection of materials and parts or in medical examinations.

flush A form of mounting in which there is little or no protrusion from the panel surface.

flutter **1.** In a high-frequency superheterodyne receiver, a rapid fluctuation in signal strength, caused by tuning and detuning of the oscillator stage. This usually results from poor direct-current (dc) power-supply regulation. **2.** Repetitive, rapid fluctuations in the output of a sound reproducer. Also see WOW. **3.** An echo effect sometimes observed in rooms or auditoriums of poor acoustic design.

flutter bridge A bridge-type instrument for measuring flutter in constant-speed machines, such as sound recording and reproducing devices.

flutter rate The frequency of flutter, in cycles per second (Hertz).

flux **1.** Theoretical lines of force that extend in all directions from an electric charge (electric flux) or from a magnetic pole (magnetic flux). **2.** A material that makes metals more amenable to being joined by soldering. **3.** The number of photons that pass through a surface for a given time.

flux density Symbol, *B.* Unit, tesla. The degree of concentration of magnetic lines of force. One tesla represents a flux density of one volt-second per square meter (V•s/m^2).

fluxgate A device that controls the azimuth bearing of a directional system by means of interaction with the geomagnetic field.

fluxgate magnetometer A magnetic compass for robot guidance. Uses coils to sense changes in artificially generated reference fields. Output from the sensors is sent to a computer that calculates the robot's position, based on the orientation and intensity of the lines of flux in the reference fields.

flux graph A device that graphically records the intensity of a magnetic field around a permanent magnet or electromagnet, or around an inductor carrying a current.

flux leakage See MAGNETIC LEAKAGE.

flux lines The theoretical lines of force in an electric or magnetic field.

flux linkage The passage of lines of force set up by one component through another component, so as to enclose most of the penetrated component's volume.

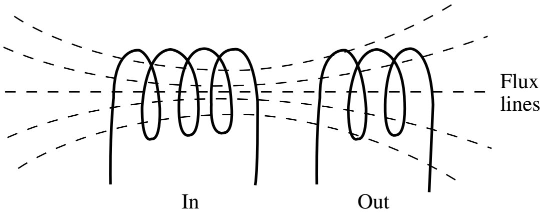

flux linkage

fluxmeter An instrument for measuring magnetic flux density. Also called *gaussmeter.*

flux refraction The tendency for magnetic lines of flux to change direction at the boundary between substances having different permeability. Flux refraction resembles refraction of electromagnetic radiation in or at a boundary between substances having different indices of refraction.

flyback **1.** The abrupt drop or reversal of a current or voltage that was previously increasing (e.g., the rapid fall of a sawtooth wave). Also see KICKBACK. **2.** The duration of the drop of a current or voltage that was previously increasing, for a sawtooth or similar wave. **3.** In an oscilloscope or picture tube, the rapid return of the beam to its starting position.

flyback checker An apparatus that senses the presence of short or open circuits in motors, transformers, and generators, by measuring the amount of flyback (kickback).

flyback power supply See KICKBACK POWER SUPPLY.

flyback time The time taken for the electron beam in an oscilloscope tube, picture tube, or camera tube to return to its starting point after it has reached the point of maximum deflection.

flyback transformer In a television receiver circuit, the horizontal output transformer. The unit supplies horizontal scanning voltage and kickback voltage, which is rectified to produce the high-voltage direct-current (dc) anode potential. Also see FLYBACK and KICKBACK POWER SUPPLY.

flying eyeball An undersea exploration robot consisting of a television camera, illumination lamps, and thrusters (such as jets or propellers). A cable, which also serves as a tether, sends data to a human operator, and allows the operator to control the movements of the robot. In some cases, the tether/cable can be replaced by an infrared or visible-light laser data link.

flying-spot tube A tube, such as a camera tube, in which a rapidly deflected spot of light scans an image on a transparent screen; the spot is projected through the picture to a photomultiplier.

fly's-eye lens A lens consisting of hundreds of much smaller lenses. Used in microelectronic circuit fabrication to produce many images of the same circuit.

flywheel effect **1.** In an inductance-capacitance (LC) tank circuit, the completion of a partial input wave cycle at the resonant frequency, resulting from the storage and release of energy. This provides a nearly perfect sine-wave output for class-AB, class-B, and class-C radio-frequency (RF) power amplifiers. **2.** In an LC tank circuit, the action in which energy continues to oscillate between the capacitor and inductor after an input signal has been removed. The oscillation stops when the tank-circuit finally loses the energy absorbed. The lower the inherent resistance of the circuit, the longer the decrement (decay time).

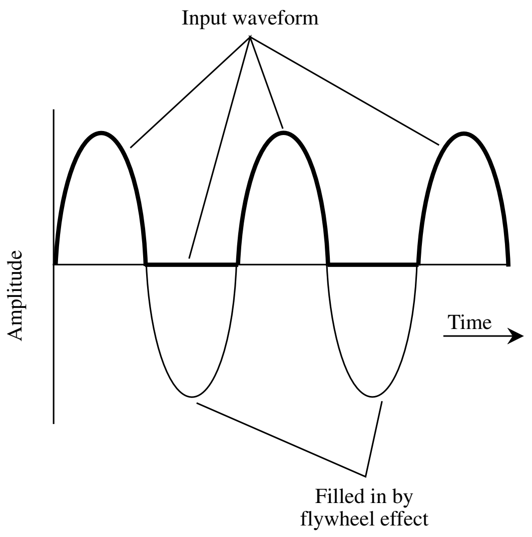

flywheel effect, 1.

flywheel synchronization A form of television scanning synchronization used when the received signal is very weak. The synchronization signals from the transmitter are sensed by the receiver, which then produces its own local pulses based on the rate of received pulses.

flywheel tuning A tuning dial mechanism in which the control shaft has a flywheel for the smoother tuning action afforded by the added momentum.

Fm Symbol for FERMIUM.

FM Abbreviation of FREQUENCY MODULATION.

f_m Abbreviation of MODULATION FREQUENCY.

FM-AM Pertaining to equipment that will operate with either amplitude-modulated or frequency-modulated signals.

FM-AM multiplier A method of frequency multiplication using both amplitude and frequency modulation of a carrier wave.

FM broadcast band The 88- to 108-MHz frequency band, within which channels spaced 200 kHz apart occupy positions from 88.1 to 107.9 MHz.

FM detector See DISCRIMINATOR, RATIO DETECTOR, and SLOPE DETECTOR.

FM-FM Frequency modulation by one or more FM subcarriers.

FM limiter In a frequency-modulation circuit, a stage which holds the amplitude of the FM signal to a constant value. The limiter can be active (e.g., an amplifier-limiter transistor) or passive (e.g., a diode clipper).

FM multiplex See MULTIPLEX ADAPTER.

FM noise Unintentional modulation of a frequency-modulated transmitter, resulting from noise in the audio-input stages.

FM-PM A system of modulation in which a carrier is phase modulated by frequency-modulated subcarriers.

FM radar A radar system in which the signal is frequency modulated; the distance to the target is measured in terms of the beat note between transmitted and reflected waves.

FM repeater A two-way radio system composed of a simultaneously operating receiver and transmitter, the latter of which retransmits (usually on a different frequency) all signals picked up by the receiver. The system is usually tower- or hilltop-mounted, and is used to extend the range of two-way units in a communications network.

FM stereo The use of multiplex methods to transmit and receive stereophonic programs in an FM channel. Also see MULTIPLEX ADAPTER.

FM tuner A compact radio receiver that handles frequency-modulated (FM) signals, and delivers its low-amplitude audio output to a high-fidelity system. Compare AM TUNER and AM-FM TUNER.

focal length Symbol, F. The distance from the center of a lens or dish antenna to the principal focus. Also see PRINCIPAL FOCUS.

focus **1.** The point at which rays converge. Also see PRINCIPAL FOCUS. **2.** To bring rays to a point of convergence.

focus coil See FOCUSING COIL.

focus control In an oscilloscope or television circuit, the potentiometer that controls the voltage on the focusing electrode of the cathode-ray tube and, accordingly, the sharpness of the image.

focus grid **1.** The focusing electrode in an electrostatic cathode-ray tube. **2.** The focusing electrode in an oscilloscope tube.

focusing Bringing a ray of particles or energy to a common point. This can be done using lenses, deflecting coils, deflecting plates, or reflecting devices. Focusing can be done with acoustic waves, electromagnetic waves, and theoretically with any kind of disturbance propagated through any medium.

focusing anode See FOCUSING ELECTRODE.

focusing coil An external coil used to focus an electron beam in a cathode-ray tube. Also see ELECTROMAGNETIC FOCUSING.

focusing electrode The internal electrode (grid or ring) used to focus the electron beam in a cathode-ray tube. Also called focus electrode. Also see ELECTROSTATIC FOCUSING and FOCUS GRID, **1**, **2**.

focusing magnet A permanent magnet assembly for focusing the electron beam in a cathode-ray tube.

foil **1.** The thin conductive strips on a printed-circuit board. **2.** Also known as *tape.* Thin metal supplied in strips, intended for use in certain security systems. It can be installed in closed loops at potential points of entry.

foil capacitor A capacitor whose plates are sheets or strips of metal foil separated by a dielectric film.

foil coil See FOIL-WOUND COIL.

foil conductor A conductor that is a strip of metal foil, rather than wire. Also see FOIL PATTERN.

foil electroscope See LEAF ELECTROSCOPE.

foil pattern The pattern of thin metal circuit paths that constitute the "wiring" of a printed circuit. Also see ETCHED CIRCUIT and PRINTED CIRCUIT.

foil-wound coil A coil wound with metal foil (usually aluminum or copper) instead of wire. Such coils substantially reduce the weight of large transformers and filter chokes.

foldback **1.** See FOLDBACK CURRENT LIMITING. **2.** In audio recording, the routing of sound (via an audio mixer) to some other destination in addition to the recording medium. Example: playing an electronic organ while singing, recording the arrangement on tape, and also listening to it (organ and voice) in a headset.

foldback current limiting In a power supply, a method of automatically reducing the output current to a safe level when the load current exceeds the maximum recommended value. This action protects both the power supply and the powered equipment.

folded dipole A half-wavelength, center-fed antenna constructed of two parallel wires with their ends connected together. It has the same length as a simple dipole antenna, but the feed-point impedance is four times that of the ordinary dipole. Instead of approximately 73 ohms, the folded dipole presents a resistive impedance of about 300 ohms. This makes the folded antenna desirable for use with high-impedance, parallel-wire transmission lines. It also can be used to obtain a good match with 75-ohm coaxial cable when four antennas are connected in phase, or with 50-ohm coaxial cable when six antennas are connected in phase. Compare DIPOLE ANTENNA.

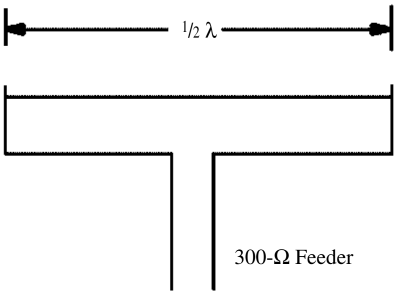

folded dipole

folded horn A loudspeaker having a horn whose flare is divided into several zigzagging chambers; that is, the horn is, in effect, folded to squeeze a required length into a small cabinet.

folded-horn enclosure See LABYRINTH SPEAKER.

folded pattern An oscilloscope image having an elongated time axis obtained by successive horizontal sweeps—each placed slightly lower on the

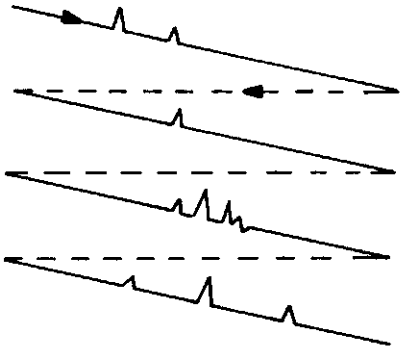

folded pattern

screen than the preceding one. The folded-pattern technique provides a time axis several times longer than the screen width.

folding frequency In a system where sampling is made at uniform frequency increments, the frequency corresponding to half the sampling rate in hertz.

foldover Distortion characterized by the horizontal or vertical overlapping of a television picture.

follower A single-stage, active circuit characterized by zero phase reversal, and voltage gain less than unity. The *emitter follower* is also called a *common-collector circuit*; the *source follower* is also known as a *common-drain circuit.* Characterized by moderate to high input impedance, and low output impedance over a wide band of frequencies.

follower drive In a servo system, the drive that mechanically follows the master drive.

following blacks In a television picture, the effect in which a moving white object has a black border following it.

following whites In a TV picture, the effect in which a moving dark object has a white border following it.

follow-up motor See SERVOMOTOR.

font The physical shape and size of the letters and numbers in an alphanumeric system.

font reticle In optical character recognition, an overlay reference pattern of lines used to check the size and configuration of an input character, the size of punctuation marks, and spacing between lines and characters.

food-service robot Any robot that is used for the purpose of packaging, preparing, and/or serving food.

foot Abbreviation, ft. A unit of linear measure in the English system equal to 0.3048 meter.

foot-candle Abbreviation, fc. A unit of illuminance; 1 fc is the amount of direct light emitted by 1 candela (see CANDLEPOWER) that falls on 1 square foot of a surface on which every point is 1 foot away from the source. In the International System of Units, the unit is *lux* (lumens per square meter). Compare METER-CANDLE.

foot-candle meter A light meter whose scale reads directly in foot-candles.

foot-lambert Abbreviation, fL. A unit of luminance; the average brightness of a surface that emits or reflects 1 lumen per square foot. The Standard International (S.I.) unit is the *candela per square meter* (cd/m^2); 1 fL = 3.426 cd/m^2.

foot-pound Abbreviation, ft-lb. In the English system, a unit of energy equal to 1 pound displaced through 1 foot in the direction of the exerting force. The Standard International (S.I.) unit is the *joule* (j); 1 ft-lb = 1.356 j.

foot-pound-second system See FPS SYSTEM OF UNITS.

foot switch A switch operated by the foot, generally used for the purpose of turning a playback or recording system on and off. Often used for taking dictation.

forbidden band See ENERGY GAP.

forbidden character code An error-finding code using forbidden characters: combinations of prohibited bits. Also called *forbidden combination.*

forbidden energy band See ENERGY GAP.

force **1.** Symbol, F. Units: newton, dyne, poundal. The agency or influence that accomplishes work. **2.** An operator interjection made during a program run that causes the computer to execute a branch instruction; it is usually necessary when a condition responsible for halting a program must be bypassed.

forced coding Programming that minimizes the time required to retrieve information from storage. Also called *minimum latency programming* or *minimum access programming.*

forced oscillations Oscillations in a circuit, such as in an inductance-capacitance (LC) tank, that result from continuously applied alternating-current (ac) excitation. Compare FREE OSCILLATIONS.

foreground job A relatively high-priority, short-running program that is carried out by interrupting a low priority, long-running program. Compare BACKGROUND JOB.

force pump In a multistage vacuum system, the first pump that reduces the pressure considerably below atmospheric pressure. Also see DIFFUSION PUMP and VACUUM PUMP.

force summing device A transducer element that is physically moved by a force being transduced.

foreshortened addressing In control computers, the mixing of available storage by using simplified addressing instructions.

fork oscillator An audio-frequency oscillator controlled by a tuning fork. The dimensions of the fork determine its vibration frequency and, accordingly, the frequency of the oscillator.

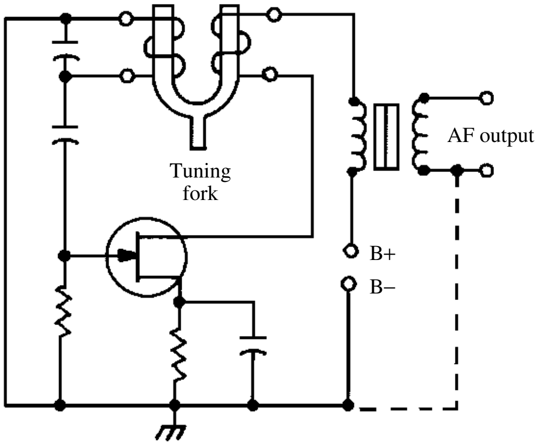

fork oscillator

form **1.** The core or frame upon which an inductor is wound. **2.** A vessel, such as a mold, used in the shaping stage of a manufacturing process.

formaldehyde Formula, HCHO. A colorless, pungent gas that is a constituent of many well-known plastic insulating materials (see PHENOLFORMALDEHYDE PLASTICS).

formant **1.** The audio-frequency range in which the sound of a spoken syllable is concentrated. **2.** Any general group of audio frequencies.

formant filter In an electronic organ, an audio filter that changes the waveshape of a tone so that the tone will have the desired characteristics.

format **1.** The form in which data is presented (e.g., the arrangement of characters, fields, words, totals, etc.). **2.** To prepare a computer disk or tape so that it will accept data.

form factor **1.** The SHAPE FACTOR for a filter or tuned circuit. **2.** For a half-cycle of an alternating-current (ac) quantity, the ratio of the root-mean-square (rms) value to the average value.

form feed **1.** A mechanical system that positions paper being supplied to a line printer. **2.** The FF character that initiates advancement of printout paper in a printer. **3.** The advancement of printout paper in a printer.

form feed character In a control loop, a character (symbol, FF) used on printing devices for controlling form feed.

forming See ELECTROFORM, **1**.

form stop An automatic device that stops a printer when the paper runs out.

FORTH A high-level computer programming language used in certain robots, automated factories, medical electronic devices, and electronic games. It was originally developed in the 1970s to facilitate computer control of equipment in astronomical observatories.

FORTRAN A high-level computer programming language developed in the 1950s, and still used in some scientific and mathematical applications. It is not especially useful for the control of electronic or mechanical devices.

fortuitous conductor A medium that creates an unwanted electrical path.

fortuitous distortion Waveform distortion that results from causes other than characteristic effects or bias effects.

forward AGC Automatic gain control provided by special transistors whose transconductance decreases with increasing emitter current, and vice versa. Compare REVERSE AGC.

forward-backward counter A counter that runs forward to perform addition and backward to perform subtraction.

forward bias Forward voltage or current in a transistor or semiconductor diode.

forward-blocking state For a silicon-controlled rectifier, the off state, during which the forward bias is so much less than the forward breakover voltage that only small off-state current flows.

forward breakover voltage **1.** For a semiconductor pn junction, the smallest forward voltage at which appreciable conduction occurs. This is about 0.3 V for germanium and 0.6 V for silicon. **2.** For a silicon-controlled rectifier, the forward voltage value at which the device abruptly switches on.

forward characteristic The current-voltage response of a semiconductor junction that is biased in the forward (high-conduction) direction. Compare REVERSE CHARACTERISTIC.

forward compatibility standards Standards developed to make programs for one computer system usable for additional or replacement equipment.

forward conduction The increased current conduction through a pn junction that is forward biased. Compare REVERSE CONDUCTION.

forward current Symbol, I_f. The increase in current flow through a pn junction that is forward biased. Compare REVERSE CURRENT.

forward current-transfer ratio The current gain of a bipolar transistor (alpha for the common-base connection and beta for the common-emitter connection).

forward power **1.** In a transmission line, the power leaving the generating source, as measured by a directional wattmeter at that location. **2.** The power arriving at the load at the terminating end of a transmission line.

forward propagation by ionospheric scatter See FORWARD SCATTER.

forward propagation by tropospheric scatter Abbreviation, FPTS. A method of transmitting part of a radio signal beyond the horizon using the scattering effect of the troposphere. Also see FORWARD SCATTER and TROPOSPHERE.

forward resistance Symbol, R_f. The resistance of a forward-biased pn junction. Also see FORWARD BIAS. Compare REVERSE RESISTANCE.

forward-reverse ratio See FRONT-TO-BACK RATIO.

forward scatter Also called *forward propagation by ionospheric scatter.* The scattering of a radio wave in the normal direction of propagation to points beyond the skip zone. The phenomenon occurs because of waves returned from regions in the ionosphere. Compare BACK SCATTER.

forward transconductance Symbol, g_{fs}. For a common-source-connected FET, the ratio of a drain-current differential to the differential of gate-to-source voltage that produces it; $g_{fs} = 1000(dI_D/dV_{GS})$, where g_{fs} is in microsiemens, I_D is the drain current in milliamperes, and V_{GS} is the gate-to-source voltage in volts.

forward voltage Symbol, E_f or V_f. Voltage whose polarity causes maximum current to flow through a pn junction. Compare REVERSE VOLTAGE.

forward voltage drop The voltage across a semiconductor junction that is biased in the forward (high-conduction) direction. Compare REVERSE VOLTAGE DROP.

FOSDIC Acronym for *film optical scanning devices for input to computer.*

Foster-Seeley discriminator A discriminator circuit in which the diodes are operated from a single-tuned, center-tapped secondary of the input transformer. The center tap is also capacitively coupled to the top of the transformer's primary coil. Compare TRAVIS DISCRIMINATOR.

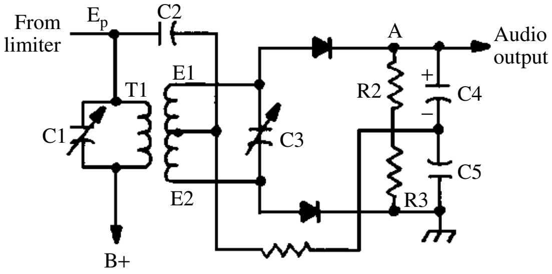

Foster-Seeley discriminator

Foucault currents See EDDY CURRENTS.

four-address instruction A computer instruction in which the address is comprised of four addresses: two for operands, one for the result of the operation, and one for the upcoming instruction.

four-channel sound system Also called *quadraphonic sound system.* A high-fidelity, stereophonic sound reproduction system, in which there are four channels, rather than the usual two. The channels are generally designated left front, left rear, right front, and right rear. The four-channel system is an enhancement of two-channel stereo.

four-dimensional continuum **1.** In relativistic theory, the space-time continuum. There are three spatial dimensions and one time dimension. A point in the continuum can be uniquely defined by three space coordinates and one time coordinate. **2.** Any continuum that requires four and only four coordinates to uniquely determine the position of a point.

Fourier analysis Use of the FOURIER SERIES to evaluate the components of a complex wave, or to define a complex wave in terms of its components.

Fourier series An infinite mathematical series that shows any periodic function to be a combination of sine terms and cosine terms. Any complex wave (e.g., a square wave) consists of fundamental and harmonic sine-wave components. In simplified form, the series is:

$$f(x) = A_0/2 + A_1 \cos x + B_1 \sin x + A_2 \cos 2x + B_2 \sin 2x + A_3 \cos 3x + B_3 \sin 3x + \ldots$$

In general, the more terms to which the series is calculated, the better the approximation of a complex waveform.

four-layer diode A dual-terminal npnp device that is usable as a bistable switch, sawtooth or pulse generator, memory device, etc.

four-layer transistor A transistor in which the wafer or block has four processed regions; however, the device might have only three terminals. Some examples are the silicon-controlled rectifier, silicon-controlled switch, and thyristor.

four-level laser A laser identical to the three-level laser, except for the addition of one excited state.

four-phase system A two-phase system in which the center taps of the coils are interconnected. Also called QUARTER-PHASE SYSTEM.

four-terminal network A network having two input terminals and two output terminals. One input terminal can be internally connected to one output terminal (as when a common ground is present), but this is not mandatory.

four-space **1.** A mathematical space in which four coordinates (w,x,y,z) are necessary to uniquely define a point. **2.** Three-dimensional space with the addition of time as a fourth dimension; coordinates are, for example, (x,y,z) for space and t for time.

four-space coordinates **1.** A system of coordinates for uniquely determining points in four-space. Such a system can be Euclidean or non-Euclidean, as with space of any number of dimensions. **2.** The set of numbers that defines a particular point uniquely in four-space; for example, $P = (3,-15,0,-7)$.

four-sphere The set of all points equidistant from a given point P in four-space. Formula is $w^2 + x^2 + y^2 + z^2 = r^2$, where r is the radius and the coordinates are (w,x,y,z) in the Euclidean, Cartesian system.

four-track recording A tape recording in which four channels are recorded in two adjacent tracks on the tape. Usually, tracks number 1 and 3 are in the forward direction, and tracks number 2 and 4 are in the reverse direction.

four-track tape A magnetic tape with four parallel sound paths.

four-wire wye system A three-phase system in which three wires supply the respective phases, a fourth being the neutral conductor.

F_p Symbol for POWER-LOSS FACTOR.

fp Abbreviation of FREEZING POINT.

FPC Abbreviation of *Federal Power Commission.*

FPI Abbreviation of FUEL-PRESSURE INDICATOR.

FPIS Abbreviation of *forward propagation by ionospheric scatter.* See FORWARD SCATTER.

fpm Abbreviation of *feet per minute.*

fps **1.** Abbreviation of *feet per second.* **2.** Abbreviation of *frames per second.* **3.** Abbreviation of *foot-pound-second (fps),* a chiefly British system of units.

fps system of units The British system of units of measurement that uses the foot for length, the pound for mass, and the second for time. Compare CENTIMETER-GRAM-SECOND and INTERNATIONAL SYSTEM OF UNITS.

FPTS Abbreviation of FORWARD PROPAGATION BY TROPOSPHERIC SCATTER.

Fr Symbol for FRANCIUM.

fr Abbreviation of FRANKLINE.

fractional exponent An exponent indicating that a number is to be raised to a fractional power (e.g., $10^{4/3}$). The numerator of the exponent indicates the power to which the base number must be raised; the denominator of the exponent indicates the root that must be taken of the result. Thus, *na/b* is equal to the *b*th root of *na*.

fractional gain Amplification less than unity. A notable example is the transfer function of a source follower or emitter follower.

fractional horsepower Any power rating lower than one horsepower (1 hp). Also see HORSEPOWER.

fractional uncertainty See RELATIVE UNCERTAINTY.

frame **1.** A single, complete video image, scanned in 1/30 second in conventional television receivers. **2.** A single motion-picture (film) image. **3.** In packet communications, a fundamental unit of data. The three types are called *information (I) frame*, *supervisory (S) frame*, and *unnumbered (U) frame*. **4.** One of a recurring cycle of pulses. **5.** In pulse-code modulation (PCM), a cyclic word group including a sync signal. **6.** A complete commutator cycle. **7.** A digital representation of a set of objects, useful in robotics and artificial intelligence (AI).

frame alignment The condition in which the receiver, or receiving apparatus, is in correct alignment with the signal to be received. In television, for example, this results in true rendition of the picture. Incorrect frame alignment (misalignment) might result in the picture being split with the top and bottom interposed. For other types of signals, misalignment would result in garbled reception.

frame-alignment signal In television, a transmitted signal that is used to ensure that frame alignment occurs in the receiver. It is a form of synchronizing pulse.

frame-alignment time slot In a transmitted television frame, an interval of time that is used for the purpose of transmitting a frame-alignment signal. There might or might not be other signal information transmitted during this time interval.

frame frequency The number of frames of a motion-picture film that come into position per unit of time in a camera, projector, or pickup.

frame of reference Geometric relationships used to describe the location of a body in space.

frame rate See FRAME FREQUENCY.

frame-repetition rate See FRAME FREQUENCY.

frame roll Momentary vertical roll in a television picture.

frame synchronizing signal **1.** In pulse amplitude modulation (PAM), a coded pulse indicating initiation of a commutation frame. **2.** In pulse-code modulation (PCM), a signal used to identify an information frame.

framing **1.** Synchronization of the vertical component of a video signal so that the top and bottom of the transmitted and received pictures line up. **2.** The process of lining up the top and bottom of a movie picture. **3.** Alignment of the characters in a digital alphanumeric transmission.

francium Symbol, Fr. A radioactive metal element of the alkali-metal group. Produced artificially through radioactive disintegration. Atomic number, 87. Atomic weight, 223.

Frankenstein scenario A theme often depicted in science fiction, in which intelligent machines seize power from their human creators. With the rapid advance of technology, especially in robotics and artificial intelligence (AI), some people, including a few educated researchers, believe this scenario is within the realm of possibility. Most scientists think it is highly improbable.

Franklin antenna A vertical collinear array that produces omnidirectional gain because of phasing among the individual components.

frankline (Benjamin Franklin, 1706–1790) Abbreviation, fr. A name that has been suggested for the unit of electric charge; 1 fr is the charge that exerts a force of 1 dyne on an equal charge at a distance of 1 centimeter in a vacuum.

Franklin oscillator A dual-terminal, audio/radio-frequency (AF/RF) oscillator circuit. Consists of a two-stage, resistance-capacitance (RC) coupled amplifier, with a tuned inductance-capacitance (LC) tank in the input gate circuit, and with capacitive feedback from the second drain to the tank.

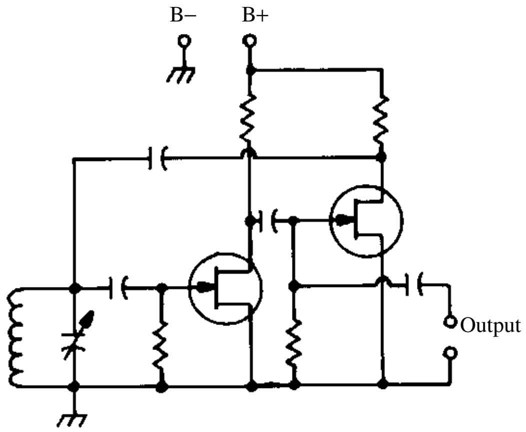

Franklin oscillator

Fraunhofer region The area surrounding a radiating antenna, throughout which the energy appears to come from a single point located near the actual antenna.

free air resonance For a speaker, the resonant frequency or frequencies exhibited when the device is not mounted in a cabinet.

free carrier A free electron or, in a semiconductor material, the equivalent hole. Also see ELECTRON and HOLE.

free charge The portion of a charge on a conductor that, being unaffected by a neighboring charge, will escape to ground when the conductor is grounded. Compare BOUND CHARGE.

free electron **1.** An electron situated in one of the outer orbits of an atom, held loosely by the nucleus. Because free electrons can easily escape the attraction of atomic nuclei, they will drift among atoms if the material is subjected to an electric potential. The result is electric current. Also see ELECTRON and BOUND ELECTRON. **2.** An electron that is not associated with any atomic nucleus.

free field Data organized in a storage medium in such a way that a data item or field can be anywhere in the medium. Compare FIXED FIELD.

free impedance For a transducer, the input impedance produced by a perfectly short-circuited load.

free magnetic pole A magnetic pole that is so well isolated from its opposing pole that it experiences little or no influence from the latter.

free magnetism A theoretical medium or fluid to which magnetic effects are conventionally given. The sum of free magnetism in any given object is always zero. Within any small part of the field, the free magnetism is thought of as flux lines. This theoretical medium can be any nonmagnetic material.

free net In radio communications, a network in which stations are free to communicate with other stations in the net without constant supervision by the net control station. Such communication is carried out on a frequency slightly above or below that of the net's formal operation.

free oscillations Oscillations in a circuit, such as an inductance-capacitance (LC) tank, that continue after excitation has been removed. Also see FLYWHEEL EFFECT. Compare FORCED OSCILLATIONS.

free path In a gas tube, the path taken by an electron as it collides with atoms. Also see MEAN FREE PATH.

free-power supply **1.** A simple tuned radio-frequency (RF) detector diode, used to rectify a radio signal and supply small amounts of direct current for the operation of low-powered transistor circuits. **2.** See SOLAR BATTERY.

free reel The supply reel of a magnetic-tape recorder.

free-running frequency The frequency at which a synchronized generator, such as a multivibrator or self-excited oscillator, will operate when the synchronizing voltage is removed.

free-running multivibrator See ASTABLE MULTIVIBRATOR and UNCONTROLLED MULTIVIBRATOR.

free space Empty space; a theoretical ideal.

free-space loss Radio transmission loss disregarding variable factors (a theoretical condition).

free-space pattern The ideal directivity pattern of an antenna that is situated many wavelengths above ground. In use, this pattern is modified by reflections from ground.

free speed The angular velocity of an unloaded motor.

free-standing display In a computer system, a remote display unit for prompting peripheral operators.

freezing point Abbreviation, fp. The temperature at which a liquid starts becoming a solid at normal pressure. Compare MELTING POINT.

F region Also called *F layer.* A region of the ionosphere with an altitude at night of approximately 175 miles. In daytime, the region splits into the lower *F1 region* and the higher *F2 region.* This layer is primarily responsible for long-distance propagation of radio waves at high frequencies (3 MHz to 30 MHz). At times it returns waves at frequencies as high as about 70 MHz.

Fremodyne detector A frequency-modulation (FM) detector that is essentially a conventional amplitude-modulation (AM) circuit detuned to one side of resonance (slope-tuned) to demodulate a frequency-modulated signal. Also see SLOPE DETECTOR.

French phone See CRADLEPHONE.

freqmeter Contraction of FREQUENCY METER.

frequency Symbol, f. The rate at which a phenomenon is repeated. The basic unit of frequency is the *Hertz (Hz)*, which represents one complete cycle per second. Common units encountered in electronics are the *kilohertz (kHz)*, *megahertz (MHz)*, and *gigahertz (GHz)*, where 1 kHz = 10^3 Hz, 1 MHz = 10^6 Hz, and 1 GHz = 10^9 Hz. Occasionally, the *terahertz (THz)* is used; 1 THz = 10^{12} Hz.

frequency-agile radar A radar system in which the transmitter frequency is shifted in a predetermined pattern for the purpose of avoiding detection. A frequency-agile radar system, with a complex frequency control program, is very difficult to jam.

frequency allocation **1.** The assignment of frequencies to radio and allied services by the licensing authority (in the United States, the Federal Communications Commission). **2.** A specific assignment of a frequency or a band of frequencies. Also see RADIO SPECTRUM.

frequency band A given range of frequencies, usually specified for some application (e.g., the band allocated for standard radio broadcast service). Also see BAND.

frequency bias An intentional change in the frequency of a transmitted signal.

frequency bridge 1. Any alternating-current bridge, such as the *Wien bridge* or *resonance bridge*, that can be nulled at only one frequency for a given set of bridge-arm values. **2.** Any alternating-current bridge that is used to measure unknown frequencies.

frequency calibrator A device, such as a crystal oscillator, that provides a signal of precise frequency with which other signals can be compared. Also see SECONDARY FREQUENCY STANDARD.

frequency changer 1. A superheterodyne converter (see CONVERTER). **2.** A motor-generator in which the output voltage has the same value as the input voltage, but is of a different frequency. **3.** See FREQUENCY-MULTIPLYING TRANSFORMER. **4.** See FREQUENCY MULTIPLIER.

frequency-change signaling See FREQUENCY-SHIFT KEYING.

frequency channel A relatively narrow segment of a frequency band allocated to a station in a particular service. The bandwidth of the channel depends on the type of modulation used, the type of data to be transmitted, and the speed or fidelity of the data to be transmitted.

frequency comparator A device, such as an oscilloscope or zero-beat indicator, used to check one frequency against another. Also see FREQUENCY COMPARISON.

frequency comparison The observation of a current or voltage of one frequency for similarities in that of another frequency. Comparisons (as in frequency matching) can be made by audio means, by visual means, or both. Common instruments used are oscilloscopes, beat-note detectors, and beat-note meters.

frequency-compensated attenuator An attenuator, such as one in an electronic voltmeter or wideband oscilloscope, that has been modified by the addition of capacitors or inductors to achieve reasonably flat response over a wide range of frequencies.

frequency compensation The modification of a circuit, such as an amplifier or attenuator, by the addition of capacitors or inductors to tailor its response at specified frequencies.

frequency control 1. An adjustable component (potentiometer, variable capacitor, or variable inductor) with which the frequency or frequency response of a circuit is adjusted. **2.** A device, such as a quartz crystal or tuning fork, that automatically sets the frequency of an oscillator.

frequency conversion The process of changing a signal from one frequency to another, usually without altering the signal bandwidth. In some cases, a signal is turned "upside down" by this process [e.g., an upper-sideband (USB) signal might be changed to a lower-sideband (LSB) signal]. Generally accomplished by means of a MIXER.

frequency converter 1. An active or passive device for changing the frequency of a signal. **2.** The mixer in a superheterodyne circuit.

frequency correction Manual or automatic resetting of a deviated frequency to its original value.

frequency counter An instrument that counts signal cycles or pulses over a standard time base (a frequency measurement). Often used to accurately measure the frequencies of radio or television signals; in this application, it is a precision FREQUENCY METER.

frequency cutoff See CUTOFF FREQUENCY.

frequency detector See FREMODYNE DETECTOR.

frequency deviation 1. The degree to which a frequency changes from a prescribed value. Thus, if the frequency of a 1000-Hz oscillator drifts between 990 and 1010 Hz, the deviation is ±10 Hz. **2.** In a frequency-modulated (FM) signal, the amount of instantaneous frequency shift above and below the unmodulated carrier frequency.

frequency-deviation meter In frequency-modulation (FM) communications operations, a meter that gives a direct reading of frequency deviation resulting from a modulating signal. It uses either a tuned circuit or a frequency comparator.

frequency difference 1. In a superheterodyne circuit, the difference between the signal frequency and the oscillator frequency. **2.** In any beat-frequency operation, the quantity $f_2 - f_1$, where f_2 is the higher frequency and f_1 is the lower frequency. Compare FREQUENCY SUM.

frequency discriminator See DISCRIMINATOR.

frequency distortion A form of distortion in which the amplification of some frequencies is different from that of others.

frequency distribution See DISTRIBUTION, **2**.

frequency diversity The transmission and reception of signals at two or more frequencies for the purpose of reducing the effects of fading. It is generally used in long-distance, high-frequency circuits.

frequency divider A circuit or device whose output frequency is a fraction of the input frequency. Compare FREQUENCY MULTIPLIER.

frequency-dividing network See CROSSOVER NETWORK.

frequency-division multiplex A form of multiple-signal parallel transmission in which a single carrier is modulated by two or more signals simultaneously.

frequency doubler A circuit that multiplies an input frequency by two. If a doubler's input circuit is tuned to frequency f, then its output circuit is generally tuned to $2f$. Frequency doubling is performed by various nonlinear devices, including transistors, varactors, and biased diodes.

frequency drift An undesired, usually gradual, change in the frequency of a signal from its intended frequency or channel; expressed in hertz or kilohertz.

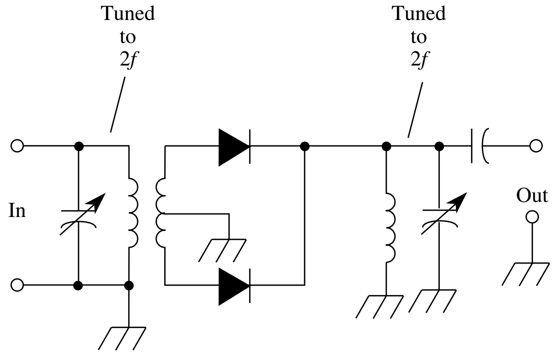

frequency doubler

frequency function See PROBABILITY DENSITY FUNCTION.

frequency indicator **1.** A device that indicates when a phase or frequency is common to two alternating currents. **2.** The display or dial that shows the operating frequency of a radio receiver or transmitter. **3.** See FREQUENCY METER.

frequency keying See FREQUENCY-SHIFT KEYING.

frequency meter An instrument for measuring the frequency of an alternating current. The several different types are used in different applications. Also see AUDIO FREQUENCY METER, FREQUENCY COUNTER, POWER-FREQUENCY METER, and WAVEMETER.

frequency-modulated radar See FM RADAR.

frequency modulation Abbreviation, FM. A method of conveying intelligence in wireless communications and broadcasting. The amplitude of the carrier remains constant, and the instantaneous frequency varies. One scheme for obtaining this type of modulation is to apply the modulating signal to a varactor in an oscillator. Another method is to modulate the phase of the oscillator signal. This causes small fluctuations in the frequency as well, because any instantaneous phase change shows up as an instantaneous frequency change (and vice versa). Also see AMPLITUDE MODULATION, PHASE MODULATION, SINGLE SIDEBAND.

frequency modulation deviation **1.** In frequency modulation (FM), the largest difference between the instantaneous signal frequency and the unmodulated carrier frequency. **2.** The maximum bandwidth of an FM signal at its audio modulation amplitude peak.

frequency modulator **1.** A circuit or device that modulates the frequency of an oscillator. **2.** The modulator section of an FM transmitter.

frequency monitor A device used (often continuously) to check the frequency of a signal (e.g., a frequency-deviation meter used in radio broadcast stations or a frequency meter used in electric-generating stations).

frequency multiplier A circuit or device whose output frequency is a multiple of the input frequency. See, for example, FREQUENCY DOUBLER.

frequency-multiplying amplifier See MULTIPLIER AMPLIFIER.

frequency-multiplying transformer A magnetic amplifier that generates harmonics of the supply frequency. The effect results from the nonlinearity of the transformer core material.

frequency offset **1.** The difference between an actual frequency and the desired frequency. **2.** In a communications transceiver, the difference between the receiver frequency and the transmitter frequency. In some modes, such as single-sideband (SSB), the offset is normally zero. In other modes, notably continuous-wave (CW) Morse code, the offset is normally several hundred Hz. **3.** See FREQUENCY SPLIT, **1**.

frequency overlap **1.** A common band of frequencies between two adjacent channels in a communications system. **2.** A common frequency region between two assigned bands. **3.** A condition in which parts of the sidebands of two signals occupy the same range of frequencies.

frequency pulling A change in the frequency of a circuit, especially of a self-excited oscillator, resulting from the detuning effects of an external circuit, device, or condition (such as body capacitance or a change in the temperature).

frequency pushing An effect in which a current change in a source oscillator causes a shift in source frequency.

frequency quadrupler See QUADRUPLER, **2**.

frequency quintupler See QUINTUPLER, **2**.

frequency range **1.** A communication system's frequency transmission limits, beyond which the power output is attenuated below a specified amount. **2.** The frequency band or bands within which a radio transmitter, receiver, or transceiver is designed to operate.

frequency ratio counter See FREQUENCY RATIO METER.

frequency ratio meter A meter that indicates the ratio between two frequencies, and is useful in the quick identification of harmonics.

frequency record A phonograph test disk containing recordings of various frequencies at specified amplitudes.

frequency rejection The elimination, usually by a filter, of a single frequency (or narrow band of frequencies) from a mixture of frequencies. Compare FREQUENCY TRANSMISSION.

frequency relay A frequency-sensitive relay (see SELECTIVE RELAY, **1**).

frequency response A performance characteristic that describes the operation of a device or circuit over a specified range of signal frequencies (e.g., the gain-versus-frequency characteristic of an amplifier).

frequency-response recorder A graphic recorder that automatically plots a frequency-response curve for a device under test.

frequency run A test, or test sequence, that determines the loss characteristics of a circuit as a function of the operating frequency.

frequency scanning **1.** A controlled fluctuation of the transmitter frequency in a frequency-agile radar or communications system. **2.** In a programmable, digital communications receiver or transceiver, a form of simultaneous digital monitoring of two or more channels. **3.** The frequency-response change in a spectrum analyzer.

frequency scaler See SCALER.

frequency-selection sensor A sensor that passes or rejects phenomena at certain frequencies while ignoring those at other frequencies.

frequency-selective relay See SELECTIVE RELAY, **1**.

frequency-sensitive bridge A bridge, such as the *Wien bridge* or *resonance bridge*, that can be balanced at only one frequency for a given set of bridge-arm values.

frequency separator In a television receiver, the circuit that separates horizontal- and vertical-scanning sync pulses.

frequency-shift keying Abbreviation, FSK. A method of digital signal transmission. The logic 1 (*high* or *mark*) pulses are sent at a specific carrier frequency, and the logic 0 (*low* or *space*) pulses are transmitted at another frequency slightly higher or lower than the logic 1 pulses. This is the most primitive form of frequency modulation (FM). The difference between the mark and space frequencies is called the *shift*, and is usually between 100 and 1000 Hz. Compare AUDIOFREQUENCY-SHIFT KEYING.

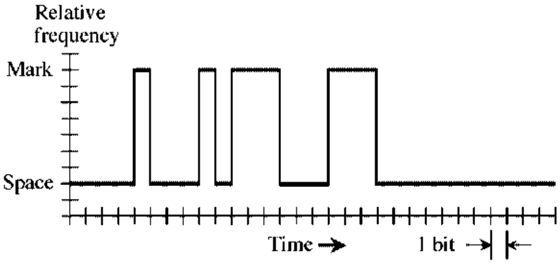

frequency-shift keying

frequency-shift radar See DOPPLER RADAR.

frequency span The difference $f_2 - f_1$, where f_1 is the lowest frequency in a given range of frequencies and f_2 is the highest frequency. Compare FREQUENCY SPREAD.

frequency spectrum All electromagnetic radiation, from longest to shortest wavelengths, within a set of specified limits.

frequency split **1.** The difference between the receiver frequency and the transmitter frequency in a communications repeater. **2.** See FREQUENCY OFFSET, **2**.

frequency spotting The setting-up of signals at reference frequencies (usually harmonics of a standard-frequency oscillator), and their use in identifying unknown frequencies. Also see FREQUENCY CALIBRATOR.

frequency spread The ratio f_2/f_1, where f_1 is the lowest frequency in a given range of frequencies and f_2 is the highest frequency. Compare FREQUENCY SPAN.

frequency stability The degree to which a frequency remains constant during variations in temperature, current, voltage, and similar factors. It is specified in Hertz (Hz), kilohertz (kHz), or megahertz (MHz), or in parts per million per unit of the variable parameter.

frequency standard A signal source of a precise frequency, against which other signal sources can be calibrated. See specifically PRIMARY FREQUENCY STANDARD and SECONDARY FREQUENCY STANDARD.

frequency sum In a beat-frequency system, the quantity $f_1 + f_2$, where f_1 is the lower frequency and f_2 is the higher frequency. Compare FREQUENCY DIFFERENCE, **2**.

frequency swing See FREQUENCY DEVIATION, **1**, **2**.

frequency synthesizer A generator of signals at a precise frequency or set of frequencies, generally adjustable in discrete frequency steps. It is used for test or communications purposes. The signals are derived from a single-frequency source, usually a crystal oscillator. Also see SIGNAL SYNTHESIZER.

frequency tolerance The acceptable amount by which a frequency can vary from its intended value. The tolerance can be specified as a percentage of the stated frequency, a certain number of parts per million, or a certain number of hertz (Hz), kilohertz (kHz), or megahertz (MHz). Example: 3.675000 MHz ±10 Hz.

frequency-to-voltage converter A device or circuit that delivers an output voltage that is proportional to the input frequency.

frequency translation **1.** The conversion of a given frequency band from one part of the electromagnetic spectrum to another, without changing the actual separation of channels or the overall width of the band. **2.** See FREQUENCY CONVERSION.

frequency transmission The passage of a frequency or band of frequencies from a mixture of frequencies through a filter or other circuit. Compare FREQUENCY REJECTION.

frequency tripler See TRIPLER, **2**.

frequency-variation method A method of determining the figure of merit (Q) of a tuned circuit by varying the frequency of the applied test voltage from resonance (f_r) to a high point (f_2) and a low point (f_1). At the high and low points, the circuit voltage is 0.707 times the voltage at resonance. The figure of merit then is calculated from the formula $Q = f_r/(f_2 - f_1)$.

frequency-voltage converter See FREQUENCY-TO-VOLTAGE CONVERTER.

frequency-wavelength conversion See WAVELENGTH-PERIOD-FREQUENCY RELATIONSHIPS.

fresnel (A.J. Fresnel, 1788–1827) A unit of frequency equal to 10^{12} Hz. Also called *terahertz* and abbreviated THz.

fresnel lens A usually square plastic sheet with progressively thicker concentric areas; its effect is similar to that of an automotive headlight lens.

Fresnel number A measure of the relative effects of diffraction in an optical lens. The Fresnel number is equal to the radius of the lens divided by the product of the light wavelength and the lens focal length, all measured in the same units.

Fresnel region For a radio-frequency transmitting antenna, the zone between the antenna and the FRAUNHOFER REGION. The size of the Fresnel region depends on the wavelength of the radiated energy.

friction The resistance to mechanical motion when one material is rubbed against another. Friction was one of the earliest sources of human-made electricity (see FRICTIONAL ELECTRICITY and ELECTRIC MACHINE). Electrical resistance, opposing the flow of current, is analogous to friction.

frictional electricity Static electricity generated by rubbing one material against another.

frictional electric machine See ELECTRIC MACHINE.

frictional error The change in parameters of a phonograph pickup, resulting from friction with the disk surface.

frictional loss A decrease or impairment in the efficiency with which energy is converted into useful work, caused by friction between moving parts.

fringe area The region in which a signal falls to the minimum field strength necessary for satisfactory communication.

fringe howl In a regenerative detector, a howling sound that occurs when the transistor first begins to oscillate, obscuring the signal. The term is used because the circuit is operated at the fringe of oscillation.

fringing See EDGE EFFECT.

Fritch Trade name (American Telephone & Telegraph Co.) for *frequency-selective switch.*

fritting A condition in which electrical contact corrosion creates a small hole, through which molten contact material passes to form a conductive bridge.

front contact The movable contact of a relay.

front end 1. The first radio-frequency (RF) amplifier stage in a radio or television receiver. **2.** The converter portion of a superheterodyne communications receiver (i.e., the RF amplifier, first detector, and local oscillator). Compare REAR END.

front layer photocell See RECTIFIER PHOTOCELL.

front porch In a television horizontal sync pulse, the interval between the end of the sync pulse and the fall of the blanking pedestal. Compare BACK PORCH.

front projection In big-screen video, a scheme in which the images from a set of bright cathode-ray tubes (CRTs) are projected onto a reflective screen, in a manner similar to the way the film is projected in a movie theater.

front-surface mirror Also called *first-surface mirror.* A mirror that has its reflective material on the front, instead of on the back.

front-to-back ratio Abbreviation, f/b. An expression of the ability of a unidirectional antenna to concentrate its radiation or response in its favored direction. This specification is nearly always given in decibels (dB). The field strength in the favored direction is compared with the field strength exactly opposite the favored direction at the same distance from the antenna in free space, at the same frequency, and with the same power applied to the antenna feed point. Measurements can be made with a calibrated field-strength meter. Compare FRONT-TO-SIDE RATIO.

front-to-side ratio Abbreviation, f/s. An expression of the directivity of a unidirectional or bidirectional antenna system. This specification is nearly always given in decibels (dB). The field strength in the favored direction(s) is compared with the field strength at right angles to the favored direction(s) at the same distance from the antenna in free space, at the same frequency, and with the same power applied to the antenna feed point. Measurements can be made with a calibrated field-strength meter. Compare FRONT-TO-BACK RATIO.

frost alarm A device or circuit that responds to the presence of frost and actuates an alarm. Such alarms are sensitive to temperature, moisture, or both.

FRUGAL Acronym for *FORTRAN rules used as a general applications language.*

FRUSA Acronym for *flexible rolled-up solar array*, such as the type used in spacecraft and communications satellites.

F scan In radar operations, a display in which a central blip represents the target at which the antenna is pointed; horizontal and vertical displacement of the blip indicate corresponding horizontal and vertical aiming errors.

FSK Abbreviation of FREQUENCY-SHIFT KEYING.

FSM Abbreviation of FIELD-STRENGTH METER.

FS meter See FIELD-STRENGTH METER.

FSR Abbreviation of *feedback shift register.*

ft Abbreviation of *foot* or *feet.*

FT-cut crystal A piezoelectric plate cut from a quartz crystal at an angle of +57°, with respect to the z-axis. Also see CRYSTAL AXES and CRYSTAL CUTS.

ft-Lb Abbreviation of FOOT-LAMBERT.

ft-lb Abbreviation of FOOT-POUND.

Fuchs antenna A simple antenna consisting of a single-wire radiator without feeder or transmission line, connected directly to the transmitter. It is usually an odd multiple of 0.25 wavelength long. When a good radio-frequency (RF) ground is used, this antenna can be effective at high frequencies, although part of its radiated field is often inside the transmitter building.

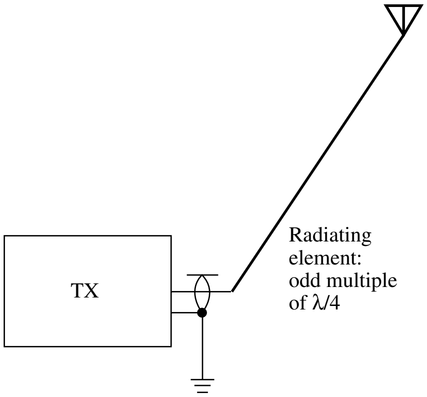

Fuchs antenna

fuel alarm A sensing circuit that actuates an alarm when the fuel in a tank or reservoir falls to a prescribed level.

fuel cell A generator that produces electricity directly from a reaction between fuel substances, such as hydrogen and oxygen.

fuel-flow alarm An electronic circuit that actuates an alarm when fuel flow changes from a prescribed value.

fuel-flow control A servo system that automatically maintains or corrects the flow rate of a fuel.

fuel-flow gauge See FUEL-FLOW METER.

fuel-flow indicator See FUEL-FLOW METER.

fuel-flow meter An instrument for measuring fuel flow rate.

fuel-flow switch A switch that is actuated by fuel flowing in pipes or other channels.

fuel gauge An instrument consisting of a transducer that senses the level of liquid fuel in a tank and delivers a proportional output current or voltage, and an electric meter whose needle is deflected in proportion to the current or voltage and, therefore, to the fuel level. Alternatively, the meter can be a direct-readout digital device, showing the number of gallons remaining in the tank, and/or the extent to which the tank is full (fraction or percentage).

fuel meter See FUEL GAUGE.

fuel-pressure indicator An instrument for measuring fuel pressure in pipes or other channels.

fuel-pressure meter See FUEL-PRESSURE INDICATOR.

full adder In a digital computer, an adder circuit that can handle the carry signal, as well as the binary elements that are to be added. Also see ADDER and CARRY. Compare HALF ADDER.

full bridge A bridge-rectifier circuit in which each of the four arms contains a diode. By comparison, the three-quarter bridge contains a resistor in one arm; the half bridge, resistors in two arms; and the quarter bridge, resistors in three arms.

full-duplex system In data communications, a system that transmits data in both directions simultaneously and continuously. Compare HALF-DUPLEX SYSTEM.

full-focus yoke See COSINE YOKE.

fullhouse A multichannel radio-control model plane system that allows the use of a realistic complement of working control surfaces.

full-load current The output current from a source when the load is maximum (that is, the load resistance is minimum).

full-load power The power drawn from a source when the load is maximum (that is, the load resistance is minimum).

full-load voltage The output voltage of a source when full power is drawn [i.e., when the load is maximum (that is, the load resistance is minimum)].

full-load wattage See FULL-LOAD POWER.

full-power frequency response The highest frequency at which a signal can fluctuate at full voltage (peak-to-peak) without causing distortion of more than a certain specified amount.

full-range speaker See MONORANGE SPEAKER.

full scale **1.** The operating range of an instrument. **2.** In an analog meter, the quantity indicated by maximum deflection of the needle (usually at the extreme right-hand end of the calibrated scale). **3.** Transducer output as a function of highest allowable input stimulus.

full-scale current Symbol, I_{FS}. In a digital-to-analog converter, the maximum current that can occur at the output.

full-scale error For an electrical indicating instrument, the rated full-scale input signal minus the actual input signal that causes a full-scale deflection. Thus, the predictable error in an instrument, expressed as a percentage of the full-scale reading.

full-scale frequency Generally expressed in Hertz (Hz) or kilohertz (kHz). The maximum frequency at which a voltage-to-frequency converter can operate while remaining within its specifications.

full-scale sensitivity The current, voltage, or power required to deflect a meter mechanism to full scale.

full-scale symmetry Expressed in microamperes (mA). The mathematical difference between the full-scale current outputs in a complementary-output digital-to-analog converter.

full track A recording track covering the full width of a magnetic tape.

full-track head A tape-recorder head having a gap that covers the full width of the tape.

full-track recording Usually applicable to quarter-inch or narrower magnetic recording tape, a one-track recording made by a head that magnetizes essentially the entire width of the tape.

full-wave bridge rectifier See BRIDGE RECTIFIER.

full-wave, center-tap rectifier A circuit in which the center-tapped secondary winding of a transformer operates two rectifier diodes, each on an alternate half-cycle of secondary voltage. The frequency of the ripple in the direct-current (dc) output is equal to twice the alternating-current (ac) input frequency. Compare BRIDGE RECTIFIER.

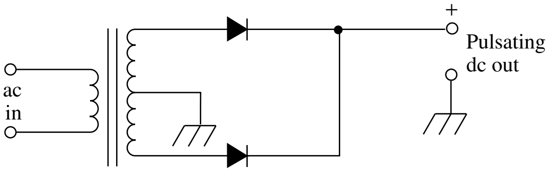

full-wave, center-tap rectifier

full-wave detector A detector circuit using two diodes in a full-wave, center-tap rectifier configuration.

full-wave doubler See FULL-WAVE VOLTAGE DOUBLER.

full wavelength Symbol, λ. The distance that corresponds to 360 degrees of phase as an electromagnetic (EM) field is propagated. In free space, it is related to the frequency by a simple equation:

$$L_{\mathrm{ft}} = 984/f$$

where L_{ft} represents λ in feet, and f represents the frequency in megahertz. If λ is expressed in meters, then the formula is:

$$L_{\mathrm{m}} = 300/f$$

Where L_{m} represents the displacement in meters. In general, if ν is the velocity factor (expressed as a ratio) in a given medium, then:

$$L_{\mathrm{ft}} = 984\nu/f$$

and

$$L_{\mathrm{m}} = 300\nu/f$$

Compare HALF WAVELENGTH, QUARTER WAVELENGTH.

full-wave loop antenna A loop antenna with a circumference of one wavelength, fed at a break. The loop can be circular, square, triangular, or occasionally some odd shape. The most efficient configuration is the circular loop. Maximum radiation occurs perpendicular to the plane of the loop.

full-wave rectifier A rectifier that delivers a half-cycle of pulsating direct-current (dc) output voltage for each half-cycle of applied alternating-current (ac) voltage. The successive output half-cycles have the same polarity. See, specifically, BRIDGE RECTIFIER and FULL-WAVE, CENTER-TAP RECTIFIER. Compare HALF-WAVE RECTIFIER.

full-wave vibrator **1.** In a vibrator-type power supply, an interrupter that closes contacts on both ends of its swing, thus causing direct current (dc) to flow through the transformer in alternate directions. **2.** A vibrator-type rectifier that closes in both directions.

full-wave voltage doubler A voltage-doubler circuit whose direct-current (dc) output has a ripple of twice the alternating-current (ac) supply frequency. Compare HALF-WAVE VOLTAGE DOUBLER.

function **1.** A mapping between two sets of quantities or points *A* and *B*, such that: (1) For each *y* in *B*, there exists at least one corresponding *x* in *A*; and (2) For each *x* in *A*, there exists exactly one *y* in *B*. In this case *y* is said to be a function of *x*; this can be written as $y = f(x)$. The set *A* is called the *domain* of *f*; the set *B* is called the *range* of *f*. **2.** A mathematical expression, using symbols, relating variables (e.g., the expression $x - y = z$ is a function of variables *x*, *y*, and *z*). **3.** The behavior and application for which a device or system is designed. **4.** Part of a computer instruction specifying the operation to be done.

functional blocks Combinations of substances or components that perform specific tasks in an electronic circuit. An example is a tuned circuit, containing inductive reactance, capacitive reactance, and resistance.

functional character See CONTROL CHARACTER.

functional design Design specifications encompassing a description of how system elements will interrelate, and what their logic design will be.

functional diagram FUNCTIONAL DESIGN represented in graphic form; that is, as an illustration or set of illustrations.

functional electronic block Abbreviation, FEB. A complete integrated circuit. See INTEGRATED CIRCUIT.

functional end (FE) point In a system operating from a battery power supply, the lowest voltage at which the equipment will properly operate. As a battery discharges, the voltage decreases; when the voltage drops to the FE point, the battery must be replaced or recharged.

functional test A performance test of a device or circuit, to see that it behaves as intended in the environment in which it is to be used.

function generator **1.** A signal generator whose output is any of several selectable waveforms (e.g., sine, square, triangular, step-pulse) and frequencies (or repetition rates). **2.** An analog computer circuit that produces a variable based on a mathematical function and one or more input variables.

function key **1.** In digital communications, a keyboard key used to control the form in which a message will be received. **2.** On a computer keyboard, any of 12 keys (usually designated F1 through F12) that activates special functions. The precise action of a given key depends on the program being run.

function polling A polling technique in which a disabled device signals its condition and specifies the remedy.

function switch In a multifunction instrument, such as a voltohm-milliammeter, the switch that permits selection of the various functions.

function table **1.** A table of mathematical function values. **2.** Hardware or software that translates one representation of information into another. **3.** A routine that allows a computer to use the values of independent variables to determine the value of a dependent variable.

fundamental Contraction of FUNDAMENTAL FREQUENCY.

fundamental component The FUNDAMENTAL FREQUENCY of a complex wave.

fundamental frequency **1.** The lowest frequency in a complex wave containing harmonic energy. **2.** In a radio or television transmitter, the intended frequency of operation. **3.** In acoustics and audio applications, the predominant pitch of a musical tone.

fundamental group A set of trunk lines in a telephone system, through which zone centers are interconnected.

fundamental mode See DOMINANT MODE.

fundamental suppression Removal of the fundamental frequency from a complex wave, leaving only the harmonics, as in the operation of a null network adjusted to the fundamental frequency.

fundamental units Base units of an absolute system of units. Example: the meter (m), the kilogram (kg), and the second (s) in the mks system.

fundamental wavelength The wavelength that corresponds to the FUNDAMENTAL FREQUENCY of a wave or signal.

fuse A safety device consisting of a wire of low-melting-point metal. When current passing through the wire exceeds a prescribed (safe) level, the resulting heat melts the wire and opens the circuit, protecting equipment from damage. See PROXIMITY FUSE.

fuse box A set of electrical fuses, usually enclosed in a metal box.

fused junction In a semiconductor, a junction produced by alloying metals to the semiconductor material.

fused junction See ALLOY JUNCTION.

fuse resistor See FUSIBLE RESISTOR.

fuse wire The low-melting-point wire used in fuses. See FUSE.

fusible resistor A low-value resistor that also serves as a fuse in certain appliances, such as television receivers.

fusing current The specified current level at which a wire of a given diameter and material composition will melt.

fusion **1.** In acoustics, pertaining to delayed or reflected waves that arrive within approximately $\frac{1}{25}$ of a second of the direct wave. So called because the human ear/brain "fuses" (blends) sounds together when they are separated by less than about $\frac{1}{25}$ second. If the delay is longer, the ear/brain usually perceives an echo instead. **2.** In a nuclear reaction, the uniting of two atomic nuclei, accompanied by the release of energy.

future labels In a computer system, program instruction labels that refer to locations not designated as absolute addresses by a compiler or assembler.

futurist A person who tries to anticipate or predict, based on current technology and trends, what will be accomplished in a given field in the next several years or decades.

fuzz A form of deliberate distortion in the waveform produced by an electric guitar.

fuzzbox A circuit that distorts the waveform produced by an electric guitar, for the purpose of creating various musical sound effects.

fuzz buster Slang for a specialized mobile radio receiver, used by drivers of vehicles to signal the presence of law-enforcement radar equipment.

fV Abbreviation of *femtovolt.*

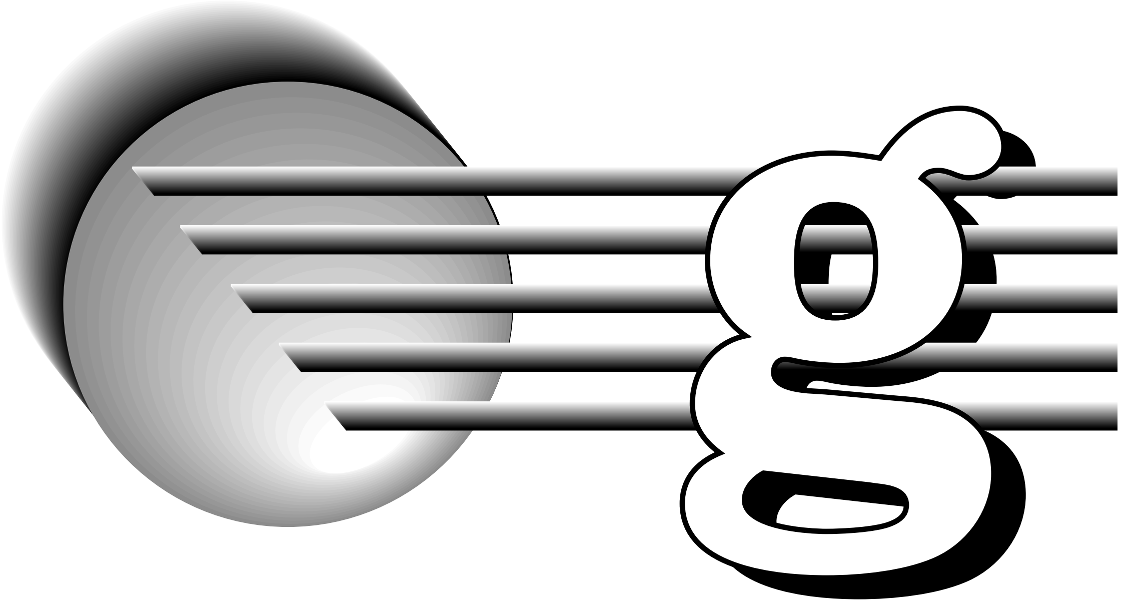

G **1.** Symbol for CONDUCTANCE. **2.** Abbreviation of GIGA-. **3.** Symbol for DEFLECTION FACTOR. **4.** Symbol for PERVEANCE. **5.** Symbol for GRAVITATIONAL CONSTANT. **6.** Symbol for GENERATOR. **7.** Symbol for GATE. **8.** Abbreviation for GAUSS.

g **1.** Symbol for CONDUCTANCE. **2.** Abbreviation of GRAM. **3.** Subscript for GATE. **4.** Subscript for GENERATOR. **5.** Symbol for GRAVITY.

GA Radiotelegraph abbreviation of "Go ahead."

G/A Abbreviation of *ground-to-air.*

Ga Symbol for GALLIUM.

GaAs **1.** Formula for *gallium arsenide.* **2.** Pertaining to semiconductor devices based on gallium arsenide.

GA coil A special form of coil, wound with extra space among the turns to reduce the distributed capacitance.

gadget **1.** A device or component. **2.** A superfluous or makeshift device.

gadolinium Symbol, Gd. A metallic element of the rare-earth group. Atomic number, 64. Atomic weight, 157.25.

gage See GAUGE.

gain The extent to which a component, circuit, device, or system increases current, voltage, or power. Applicable especially to active devices, such as transistors and integrated circuits (ICs), and to amplifiers and filters that use them. Also used to express the directional properties of some antenna systems. Usually specified in decibels (dB). See AMPLIFICATION, CURRENT AMPLIFICATION, DECIBEL, VOLTAGE AMPLIFICATION, and POWER AMPLIFICATION.

gain bandwidth product Symbol, f_T. The frequency at which the gain of a bipolar transistor is equal to 1 (no amplification or loss) in the common-emitter configuration. The f_T represents an absolute upper limit for the frequency at which a bipolar transistor will works as an amplifier. Any attempt to design an amplifier using a bipolar transistor at a frequency higher than its f_T will inevitably fail. Compare ALPHA CUTOFF FREQUENCY.

gain control **1.** To adjust the gain of an amplifier. **2.** A potentiometer used to adjust amplifier gain.

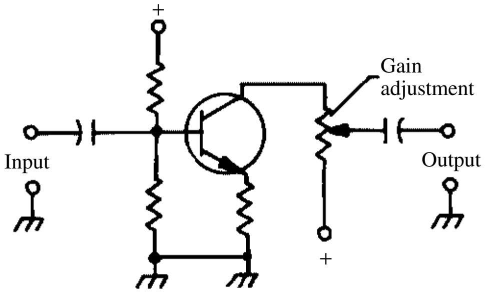

gain control

gain function A function between two currents or voltages in a circuit with gain.

gain-level linearity The quantitative measure of the extent to which the gain of a device depends on the signal level. The level is found by compar-

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

ing the output to the input level over a range of input signal levels.

gain reduction The drop in gain of an amplifier at high- and low-frequency extremes.

gain sensitivity control See DIFFERENTIAL GAIN CONTROL.

gain stability The degree to which the gain of a system remains constant during changes in related factors, such as temperature, supply power, and loading.

gain temperature coefficient The extent to which the full-scale current varies over a certain temperature range, expressed in parts per million per degree Celsius (ppm/°C).

galactic noise Radio noise propagated from the plane of our galaxy, and especially from the center, located in the direction of the constellation Sagittarius. It is of significance in space communications and radio astronomy.

galena Formula, PbS. Natural lead sulfide, which in nature takes the form of bluish-gray, cubical crystals.

gallium Symbol, Ga. Atomic number, 31. Atomic weight, 69.72. One of the constituents of the semiconductor compound GALLIUM ARSENIDE.

gallium arsenide Formula, GaAs. A compound of gallium and arsenic, used as a semiconductor material. It is noted for its low-noise characteristics.

gallium-arsenide diode A diode in which the semiconductor material is processed gallium arsenide.

gallium-arsenide varactor A low-noise, microwave varactor in which the semiconductor material is gallium arsenide.

gallium-phosphide diode A light-emitting diode in which the semiconductor material is processed gallium phosphide.

galloping ghost A form of radio-control system in which the elevation and rudder can be moved to the desired extent.

Galton whistle A device for producing high-frequency acoustic waves (ultrasound), similar to a common dog whistle.

galvanic cell Generic term for any electrochemical primary voltaic cell.

galvanic corrosion Corrosion that occurs on one of two dissimilar metals when they are immersed in an electrolyte. Caused by battery action between them. Compare ELECTROLYTIC CORROSION.

galvanic couple See VOLTAIC COUPLE.

galvanic current A very small direct current such as that produced by dissimilar metals in acid or by nervous reaction in living tissue.

galvanic pile See VOLTAIC PILE.

galvanic series A list of metals and alloys arranged in order of the most to least likely to oxidize in a given environment.

galvanic skin response Abbreviation, GSR. The variations in electrical resistance of the (usually human) skin. This phenomenon is a useful indicator in physiology, psychology, and criminology.

galvanic taste A sharp, metallic taste experienced when a small electric current is passed through the tip of the tongue.

galvanism (After Luigi Galvani, 1737–1798) The production of an electric current by chemical action, as in a battery.

galvanize To coat steel with zinc to forestall corrosion.

galvanometer A sensitive, bi-directional current meter. Used in various electrical tests—especially as a null indicator in bridge operation. Also see MICROAMMETER.

galvanometer constant The number by which a galvanometer reading must be multiplied in order to obtain the current in microamperes, milliamperes, or amperes.

galvanometer recorder A graphic recorder in which a mirror in a movable-coil galvanometer reflects a beam of light to a passing strip of photographic film.

galvanometer shunt A resistor placed in parallel with a galvanometer to decrease its sensitivity. Also see SHUNT RESISTOR.

galvanometry The use of galvanometers to determine the intensity and direction of electric currents.

galvanoplastics The science of ELECTROPHORESIS and ELECTROPLATING.

galvanoscope An instrument for detecting and showing the direction of very weak electric currents.

galvanotherapy The use of electric currents to produce heat in the body of a human or animal.

gamma ferric oxide A form of coating used in formulation of magnetic recording tape.

gamma match A linear transformer for matching an unbalanced (usually coaxial) feed line to a balanced (usually half-wave) antenna. The outer conductor of the cable is connected to the center of the radiator, and an extension of the center conductor runs for a short distance parallel to the radiator, making a right-angle bend before connecting to the radiator.

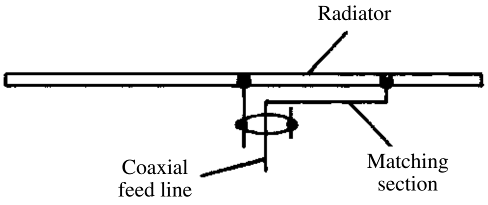

gamma match

gamma rays High-energy, ionizing radiation emitted by radioactive substances; similar to X rays, but of a shorter wavelength.

gamma section See GAMMA MATCH.

gang To mechanically couple components (pots, switches, etc.) for operation by a single knob.

gang capacitor A variable capacitor consisting of sections mounted on the same shaft for simultaneous variation. It is usually specified by the number of sections (e.g., four-gang capacitor). Compare GANGED CAPACITORS.

ganged capacitors Separate variable capacitors mechanically connected together (e.g., by belt or gear drive) for simultaneous variation. Compare GANG CAPACITOR.

ganged potentiometers Separate potentiometers mechanically connected together (e.g., by belt or gear drive) for simultaneous variation. Compare GANG POTENTIOMETER.

ganged rheostats See GANGED POTENTIOMETERS.

ganged switches Separate switches mechanically connected together for simultaneous operation. Compare MULTISWITCH.

ganged tuning Simultaneous tuning of separate circuits by means of ganged capacitors or ganged potentiometers.

gang potentiometer A potentiometer consisting of sections mounted on the same shaft for simultaneous variation. Usually specified according to the number of sections (e.g., *dual potentiometer*).

gang printer In digital computer and data processing operations, an electromechanical printer capable of printing an entire line at one time.

gang punch **1.** To punch identical or nonvarying information into the cards of a group. **2.** A machine for this operation.

gang rheostat See GANG POTENTIOMETER.

gang switch See MULTISWITCH.

Gantt chart A chart of activity versus time used in industry as an aid in making decisions regarding the allocation of resources for specific activities [e.g., as applied to PERT (project evaluation and review techniques)].

gap **1.** A space between electrodes or magnetic poles. **2.** A device consisting essentially of separated electrodes (e.g., *spark gap*). **3.** A relatively narrow space cut in iron cores to provide a break in a magnetic circuit. Also see SLOT, **1**. **4.** The opening between the opposite poles of a tape recorder or playback head.

gap arrester A lightning arrester consisting of a number of metal cylinders separated by air gaps.

gap coding A system in which silent periods are inserted, according to a specific timing code, into a transmission.

gap depth In a magnetic recording head, the depth of the gap (taken perpendicular to the face). Compare GAP WIDTH.

gap digit A digit that contributes no intelligence to the word in which it appears (e.g., a *parity bit*).

gap energy The energy represented by the forbidden gap in electron energy levels (e.g., between the M-valence band and the N-conduction band in a material).

gap filling Modification of an antenna for the purpose of eliminating nulls in the directional pattern.

gap insulation See SLOT INSULATION, **1**, **2**.

gap loss In a reproducing head, the loss that occurs because of the GAP DEPTH.

gap-type protector A spark gap used to protect equipment from high-voltage transients.

gap voltmeter See NEEDLE GAP and SPHERE GAP.

gap width In a magnetic recording head, the width of the gap (taken parallel with the face). Compare GAP DEPTH.

garbage **1.** In digital computer operations, a colloquialism for useless or incorrect data. **2.** Colloquialism for unreadable signals or severe intermodulation in a radio communications circuit. **3.** Colloquialism for an unsound theory.

garble **1.** Garbled matter. **2.** Also called *scramble.* To purposely render communications or data unintelligible to everyone, except the intended recipient(s). See SCRAMBLER CIRCUIT.

garbled matter Confused communications or data, usually resulting from distortion in a circuit or system. Also called GARBLE.

garbler See SCRAMBLER CIRCUIT.

garnet maser A maser that uses natural or synthetic garnet as the stimulated material. Also see YTTRIUM-IRON-GARNET.

gas One of the states of matter, characterized by molecules that are widely separated and are in continual, relatively rapid motion. Because it is a fluid, a gas will readily conform to a container of any shape. Gases can readily be compressed and liquefied. Compare LIQUID, PLASMA, and SOLID.

gas amplification In a radiation-counting device, the ratio, in decibels, of the charge collected to the charge produced in the gas.

gas breakdown The ionization of a gas by means of high voltage. The intensity of the electric field prevents recombination of ions. Collisions among atoms cause further ionization. Thus, the gas becomes a good conductor of current.

gas cell A cell whose operation depends on gas absorption by the electrodes.

gas cleanup Loss of pressure in a gas-filled tube, eventually leading to failure. Caused by gas ions forming compounds with metal parts or with the glass envelope.

gas detector A device for sensing presence of various gases in the air—especially toxic or explosive gases (such as chlorine, hydrogen, or methane).

gaseous conduction The conduction of an electric current through an ionized gas.

gaseous phototube A PHOTOTUBE containing a small amount of a gas suitable for ionic conduction.

gaseous voltage regulator A gas-filled diode across which the voltage drop is substantially constant during the gas discharge and which accordingly delivers a constant output voltage.

gas-filled cable A sealed cable filled with an inert gas that serves as a low-loss dielectric and moisture barrier.

gas-filled counter tube A radiation counter tube containing a gas that ionizes when irradiated. See GEIGER-MUELLER TUBE.

gas-filled lamp **1.** An incandescent lamp filled with a gas, such as nitrogen, for improved performance. **2.** Discharge lamp.

gas-filled tube An enclosure filled with gas that is subjected to an electric potential, intended to produce specific effects, such as light emission or voltage regulation.

gas-flow alarm An electronic circuit that actuates an alarm when the flow of gas through a pipe changes from a predetermined rate.

gas-flow control A servo system for automatically maintaining or adjusting the flow of gas through pipes.

gas-flow gauge See GAS-FLOW METER.

gas-flow indicator See GAS-FLOW METER.

gas-flow meter An instrument that indicates the rate of gas flow through a pipe, in terms of volume per unit time (e.g., cubic feet per minute).

gas-flow switch In a gas-circulating system, a switch that actuates an alarm when the gas flow rate changes.

gas focusing In a cathode-ray tube, a technique by which a gas is used for the purpose of focusing an electron beam. The ionization of the gas causes the electron beam to be made more narrow.

GASH Acronym for *guanidine aluminum sulfate hexahydrate*, an organic crystalline material used as the dielectric in certain ferroelectric capacitors and ferroelectric memory elements.

gas laser A LASER that uses a gas or mixture of gases (instead of a solid rod) as the stimulated medium. Some of the gases used are argon, carbon dioxide, helium, krypton, and neon.

gas maser A MASER in which the stimulated material is a gas, such as ammonia.

gas multiplication See GAS AMPLIFICATION.

gas noise Electrical noise resulting from the undirected motion of gas molecules in a gas-filled tube.

gas-pressure alarm An electronic circuit that actuates an alarm when gas pressure rises or falls.

gas-pressure control A servo system for automatically maintaining or adjusting gas pressure in pipes or other channels.

gas-pressure gauge See GAS-PRESSURE METER.

gas-pressure indicator See GAS-PRESSURE METER.

gas-pressure meter An instrument that indicates gas pressure in a pipe or container, but provides no means for automatically correcting the pressure.

gas ratio For a gas-filled tube, the ratio I_i/E_i, where I_i is the ion current and E_i is the ionization potential.

gas sensor Any element, such as the filament in a hot-filament gas detector, that responds to the presence of a gas in the environment and activates the detector or alarm circuit.

gassing **1.** The generation of gas by a storage battery, especially while it is being charged. **2.** The generation of gas during electrolysis.

gas sniffer See GAS DETECTOR.

gaston A device intended for the purpose of modulating an aircraft signal, making the signal difficult to jam. The signal is randomly modulated by noise from the device.

gas tube See GAS-FILLED TUBE.

gas-tube lightning arrester A lightning arrester, consisting of a special gas diode. The tube has virtually infinite resistance at low voltages, but provides a low-resistance path to ground when the high voltage of a lightning stroke ionizes the gas.

gas-tube oscillator A relaxation oscillator using a two-element gas tube, such as a neon lamp, as the breakdown device.

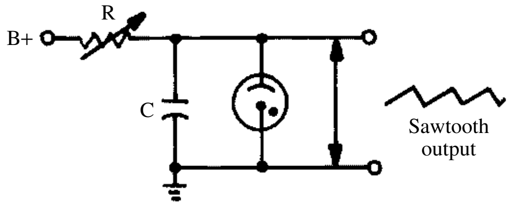

gas-tube oscillator

gas valve See ELECTROMECHANICAL VALVE.

gas X-ray tube An X-ray tube in which the positive ions of a gas bombard the cathode, which emits electrons.

gate **1.** A device or circuit that has no output until it is triggered into operation by one or more enabling signals, or until an input signal exceeds a predetermined threshold amplitude. **2.** The input (control) electrode of a field-effect transistor or thyristor device (e.g., silicon-controlled rectifier). **3.** A signal that triggers the passage of other signals through a circuit.

gate array Basic gates arranged in a pattern on a chip; the gates can be interconnected during manufacture to form a unit that performs whatever function is needed.

gate circuit **1.** An electronic switching circuit (see GATE, **1**). **2.** The circuit associated with the gate electrode of a field-effect transistor.

gate-controlled switch A device similar to a silicon-controlled rectifier or thyristor. A negative current, applied to the gate, switches the device off.

gate current Symbol, I_G. Current flowing in the gate (control) circuit of a semiconductor device. The current is finite in thyristors, but is almost zero in some types of field-effect transistors.

gated amplifier An amplifier whose input is effectively switched on and off by gating signals.

gated buffer A low-voltage, high-current driver, used for differentiation in a multivibrator circuit.

gated flip-flop A flip-flop in which it is impossible for both outputs to be low simultaneously.

gate-dip meter A dip meter using a field-effect-transistor oscillator with the indicating microammeter in the gate circuit.

gate-dip oscillator See GATE-DIP METER.

gated multivibrator A rectangular-wave generator that produces a gate voltage when triggered.

gate-drain voltage Symbol, V_{GD}. In a field-effect transistor, the maximum voltage permitted between the gate and drain electrodes.

gated sweep **1.** In radar, a sweep whose initiation and duration are closely controlled to eliminate echoes in the image. **2.** A circuit providing the action described in **1**.

gate electrode See GATE, **2**.

gate impedance **1.** The impedance of the gate electrode of a field-effect transistor with respect to the other electrode, which serves as the return. **2.** The impedance of the gate winding of a magnetic amplifier.

gate leakage current See GATE REVERSE CURRENT.

gate nontrigger voltage Symbol, V_{Gnt}. For a thyristor, the direct-current (dc) voltage applied between the gate and the cathode, above which the device fails to maintain its rated blocking voltage.

gate power dissipation Symbol, P_G. In a silicon controlled rectifier, the power consumed by the gate-cathode path.

gate-protected MOSFET A metal-oxide-semiconductor field-effect transistor in which the gate electrode is protected from accidental burnout via built-in Zener diodes, connected back-to-back.

gate pulse **1.** A pulse applied to the gate electrode to actuate a gate-controlled semiconductor device. **2.** An actuating pulse in a gate circuit.

gate recovery time Symbol, t_{Gr}. For a silicon-controlled rectifier, an extension of the reverse recovery time: the interval following application of the reverse voltage required before the forward blocking voltage can be reapplied and then blocked by the device.

gate reverse current Symbol, I_{GSS}. In a field-effect transistor, reverse current in the gate-source circuit. Also called *gate leakage current.*

gate signal **1.** The input or control signal applied to the gate electrode of a semiconductor device. **2.** An actuating signal in a gate circuit.

gate-source breakdown voltage Symbol, BV_{GSS}. The voltage at which the gate junction of a junction field-effect transistor (JFET) enters avalanche.

gate-source pinchoff voltage Symbol, V_P. In a field-effect transistor, the gate-source voltage at which the conduction channel just closes.

gate-source voltage Symbol, V_{GS}. In a field-effect transistor, the direct-current (dc) voltage between the gate and source electrodes.

gate terminal **1.** The terminal connected to the gate semiconductor in a field-effect transistor. **2.** The terminal, or terminals, connected to the input or inputs of a digital-logic network.

gate trigger current In a gate-controlled semiconductor switch, the current flowing in the gate circuit when the device is being switched on by a gate trigger voltage.

gate trigger voltage In a gate-controlled semiconductor switch, the trigger voltage required to actuate the device.

gate turn-off current In a gate-controlled semiconductor switch, the low value of gate current that flows when the device is being switched off. Turn-off current varies with collector (anode) current.

gate turn-off voltage In a gate-controlled semiconductor switch, the low value of gate voltage that causes the device to switch off.

gate voltage **1.** The voltage applied to the gate electrode of a field-effect transistor. See GATE-SOURCE VOLTAGE. **2.** The instantaneous gate-cathode voltage in a silicon-controlled rectifier. **3.** The voltage across the gate winding of a magnetic amplifier.

gate winding In a magnetic amplifier, a winding that produces gating action.

gating **1.** The process of using one signal to switch another (or part of another) on or off for a desired interval. **2.** Selecting a part of a wave for observation or for control purposes.

gauge **1.** Any device, such as a METER, used for the purpose of measuring the magnitude of a quantity. **2.** Wire data and measurements (see WIRE GAUGE, **1**, **2**, **3**). **3.** Sheet metal thickness (e.g., 10 gauge).

Gauss (Karl F. Gauss, 1777–1855) Unit of magnetic flux density, equivalent to one line of flux per square centimeter. The SI (preferred) unit of magnetic flux density is the TESLA (webers per square meter); 1 gauss equals 10^{-4} teslas (symbol, T). Also see FLUX DENSITY.

Gaussian curve See BELL-SHAPED CURVE.

Gaussian distribution In statistics, the symmetrical distribution described by a bell-shaped curve. Also called NORMAL DISTRIBUTION.

Gaussian function A mathematical function used in the design of lowpass filters. The Gaussian response is characterized by low band-edge selectivity, high transition bandwidth, negligible step response overshoot, and low step-response delay time.

Gaussian noise Electrical noise whose amplitude-versus-frequency characteristic is described by the GAUSSIAN DISTRIBUTION.

Gaussian waveform A waveform that results in minimal side lobes in a pulse-compression system.

gaussmeter See FLUXMETER.

Gauss' theorem Across any closed surface within an electric field, the total flux is approximately equal to 12.566 times the enclosed quantity of electric charge.

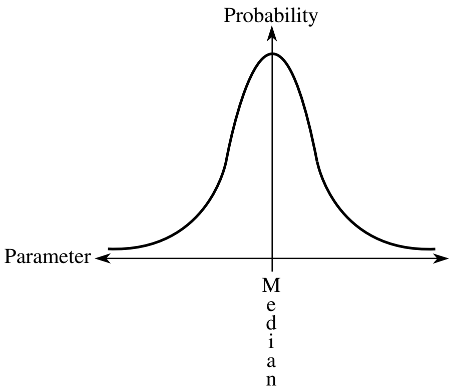

Gaussian distribution

gauze resistor See WOVEN RESISTOR.
GAVRS Abbreviation of *gyrocompass attitude vertical reference system.*
GCA Abbreviation of GROUND-CONTROLLED APPROACH.
GCI Abbreviation of GROUND-CONTROLLED INTERCEPTION.
GCM Abbreviation of *gyrocompass module.* See GYROCOMPASS.
GCT Abbreviation of GREENWICH CIVIL TIME.
Gd Symbol for GADOLINIUM.
G display See G SCAN.
GDO **1.** Abbreviation of GRID-DIP OSCILLATOR. **2.** Abbreviation of GATE-DIP OSCILLATOR.
Ge Symbol for GERMANIUM.
gear **1.** Collectively, electronic equipment. **2.** A toothed wheel commonly used in mechanical devices.
gearmotor An electric motor with a gear train for speed changing.
gear-wheel pattern A pattern produced on an oscilloscope by intensity-modulating a circular trace. A signal of unknown frequency f_x is applied to the intensity-modulation (z-axis) input. The signal produces corrugations in the trace. If there are *n* corrugations around the trace, and if the trace itself completes f_s revolutions per second, then the unknown frequency f_x, in Hz, is equal to nf_s. Compare SPOT-WHEEL PATTERN.

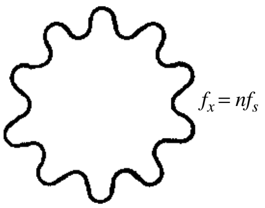

gear-wheel pattern

Geiger counter A radioactivity rate-counting instrument based on the GEIGER-MUELLER TUBE. Pulses from the tube drive a transistor, which, in turn, drives a meter or digital counter to indicate the count.

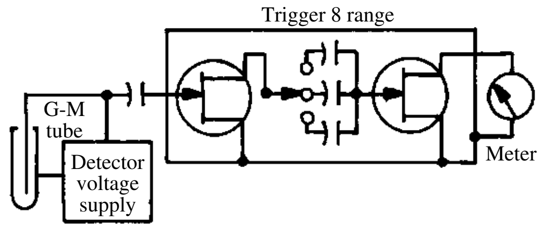

Geiger counter

Geiger-Mueller counter See GEIGER COUNTER.
Geiger-Mueller region For a GEIGER-MUELLER TUBE, the range of voltages within which the output pulse amplitude is constant, regardless of ionizing radiation intensity.
Geiger-Mueller threshold The lowest voltage in the Geiger-Mueller region.
Geiger-Mueller tube A gas-filled radiation detector/counter tube consisting of a straight wire as an anode, surrounded by a cylindrical cathode. The tube is biased by high voltage. Radiation penetrating the tube ionizes the gas; each ionizing event causes an output pulse.
Geiger point counter See POINT COUNTER.
Geiger region See GEIGER-MUELLER REGION.
Geiger threshold See GEIGER-MUELLER THRESHOLD.
Geissler tube A simple gas-filled glow-discharge tube with metal electrodes sealed in each end. When a sufficiently high voltage is applied between the electrodes, the highly rarefied gas ionizes and glows with the color associated with the particular gas used.
gel A substance equivalent to colloidal solution in the solid phase (e.g., silica gel).
gel battery A rechargeable electrochemical BATTERY designed for use with portable electronic and computer equipment. Noted for its ability to deliver high current for short periods, when necessary, and also to deliver moderate current continuously throughout its discharge cycle.
gen Abbreviation of GENERATOR.
genemotor Contraction of *generator/motor*, a (usually battery-driven) dynamotor that has separate motor and generator windings on the same armature core.
general class license An amateur-radio license that conveys some privileges in the high-frequency bands, and all operating privileges in the very-high-frequency region and above. An examination of moderate difficulty is required.

General Packet Radio Service A form of packet radio especially designed for mobile use. Subscribers pay for the volume of data they send and receive, rather than for the time they spend accessing the network. This optimizes the efficiency of the network, both from an engineering standpoint and a financial standpoint. See PACKET COMMUNICATIONS.

general-purpose bridge See UNIVERSAL BRIDGE.

general-purpose component A component designed or used for a wide range of applications. For example, a general-purpose germanium diode is useful as a detector, mixer, limiter, clipper, meter rectifier, automatic-gain-control (AGC) rectifier, and curve changer.

general-purpose computer A computer that can be used in a number of applications for which it was not specifically designed.

general-purpose diode A small-signal semiconductor diode that is useful for a variety of applications, such as detection, light-duty rectification, limiting, logic switching, etc.

general-purpose function generator A nonspecialized function generator that is capable of generating a variety of different waveforms.

general-purpose program A program for the solution of a class of problems or for a specific problem, according to certain parametric values. Also called *general routine.*

general-purpose relay Any relay that can be used in various situations, such as for switching alternating or direct currents.

general-purpose tester An instrument, such as a voltohm-milliammeter, that offers several test capabilities.

general-purpose transistor A transistor that can be used in several applications, such as audio amplification, detection, and oscillation.

general service code See CONTINENTAL CODE.

generate **1.** To produce a signal or carrier wave. **2.** To convert some non-electrical form of energy (usually mechanical) into electrical energy. **3.** To develop subroutines from parameters applied to skeletal coding. **4.** To use a program generator to produce a specialized version of a general-purpose program.

generated address An address developed by program instructions for later use by that program.

generated noise **1.** Electrical noise caused by battery action (i.e., between dissimilar metals) in a component, such as in a potentiometer. **2.** Electrical noise caused by small output variations of generating devices (rotating machines, vibrators, etc.). Also called *generator noise.*

generating magnetometer See EARTH INDUCTOR.

generating station An electric power station.

generating voltmeter An instrument based on a rapidly spinning variable capacitor. A direct-current (dc) voltage applied to the capacitor is converted into an alternating current (ac) by the varying capacitance; the ac is proportional to the voltage.

generation **1.** The production of a signal or carrier wave. **2.** The conversion of some non-electrical form of energy (usually mechanical) into electrical energy. **3.** The number of recording steps between a master recording and a copy. **4.** A copy of data in any form (e.g., tape recording, disk file, and photocopy).

generation number A number that identifies the age of a file; it is included in the file label on the disk or tape containing the file.

generator **1.** Symbol, G. Any signal source. **2.** A rotating machine for producing electricity. **3.** An electronic device for converting direct current into alternating current of a specific frequency and waveshape. **4.** In computer operation, a routine (akin to a compiler) that will produce a program to perform a specific version of some general operation by implementing skeletal coding, according to specific parameters (e.g., *sort generator*).

generator efficiency The ratio of consumed power to delivered power in a generator. It is usually expressed as a percentage.

generator noise Electrical noise caused by a rotating generator. Also see GENERATED NOISE, **2**.

generator-type microphone A microphone that produces an output voltage without the need for a supply voltage. Examples: ceramic, crystal, dynamic, electret, and velocity types.

generator-type transducer A transducer that converts mechanical motion into an electrical signal of a proportional voltage. In such a transducer, an armature or conductor moves in a magnetic field.

generic A form of software collection. Several specialized software packages can be derived from the generic collection, for use in different systems.

geodesic **1.** On a surface, the shortest path between two points. **2.** The shortest path between two geographical locations, measured over the surface of the earth. Also called *geodetic* and *great circle.*

geodesy The branch of applied mathematics concerned with the precise dimensions of the earth.

geodetic system The application of a computer to seismographic studies for the purpose of reducing drilling and mining costs.

geomagnetic field See EARTH'S MAGNETIC FIELD.

geomagnetism The earth's magnetism. Also see EARTH'S MAGNETIC FIELD.

geometric capacitance The ratio of the free charge of a capacitor to the voltage across its terminals.

geometric mean The *n*th root of the product of *n* quantities.

geometric progression A mathematical sequence in which each term after the first is obtained by multiplying the preceding one by a constant quantity (e.g., 1, 2, 4, 8, 16, 32, 64, . . .). Also called *geometric sequence.*

geometric symmetry In a bandpass or band-rejection filter, a condition in which the response is identical on either side of the center frequency. Also called *mirror-image symmetry.*

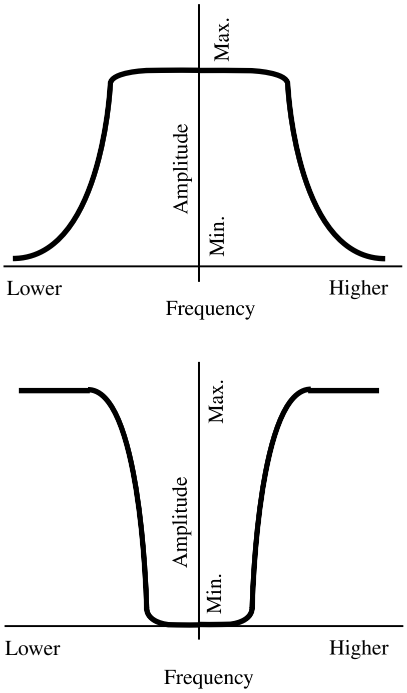

geometric symmetry

george box In an intermediate-frequency amplifier, a device used to reject jamming signals. Any jamming signal with an amplitude lower than a certain minimum is rejected.

geostationary orbit An orbit in which a satellite revolves around the earth exactly once a day, so it remains over the same place on the earth all the time. The altitude must be 22,300 miles, and the orbit must lie in the plane of the earth's equator.

geostationary satellite A satellite in GEOSTATIONARY ORBIT. It is always in the same spot in the sky from any given observing point. A geostationary satellite provides coverage over about 40 percent of the earth, and is commonly used for communications and weather-observation purposes.

germanium Symbol, Ge. A metalloidal element. Atomic number, 32. Atomic weight, 72.59. Used in semiconductor diodes, photocells, rectifiers, and transistors.

germanium diode A diode in which the semiconductor material is specially processed germanium.

germanium dioxide Formula, GeO_2. A gray or white powder obtainable from various sources; it is reduced in an atmosphere of hydrogen or helium to yield germanium, a semiconductor material.

germanium junction diode A germanium diode that contains a pn junction.

germanium photocell A photoconductive cell consisting of a reverse-biased germanium point-contact diode or germanium junction diode.

germanium point contact The contact between a pointed metal wire and a germanium wafer, as in a point-contact diode or point-contact transistor.

germanium rectifier A power rectifier that contains a germanium pn junction.

germanium transistor A transistor in which germanium is the semiconductor material. Such a transistor has lower internal resistance and greater temperature drift than a silicon transistor.

German silver A copper-nickel-zinc alloy used in some resistance wires. Also called NICKEL SILVER.

GEV Abbreviation of *ground effect vehicle.*

GeV Abbreviation of GIGAELECTRONVOLT.

gfi Abbreviation of GROUND-FAULT INTERRUPTER.

g-force See GRAVITY, **2**.

g_{FS} Symbol for FORWARD TRANSCONDUCTANCE.

G/G Abbreviation of GROUND-TO-GROUND.

ghost In television reception, a slightly displaced image appearing on the screen simultaneously with its twin (the false member of a double image).

ghost signal Any signal (such as an undesired reflection) that produces a ghost.

GHz Abbreviation of GIGAHERTZ.

G_i Symbol for INPUT CONDUCTANCE.

Gibson girl A portable radio transmitter, powered by an integral crank-operated generator, introduced during World War II for pilots forced down at sea.

giga- Abbreviation, G. **1.** A prefix meaning billion (10^9). **2.** In digital data applications, a prefix meaning 2^{30}.

gigabit A unit of digital data, equal to 2^{30} bits or 1024 megabits. Also see BIT and MEGABIT.

gigabyte A unit of digital data, equal to 2^{30} bytes or 1024 megabytes. Also see BYTE and MEGABYTE.

gigacycle See GIGAHERTZ.

gigaelectronvolt Abbreviation, GeV. A large unit of voltage; 1 GeV equals 10^9 eV. Also see BEV, ELECTRONVOLT, MEV, and MILLION ELECTRON VOLTS.

gigahertz Abbreviation, GHz. A unit of ultra-high frequency; 1 GHz equals 1,000,000,000 Hz = 10^9 Hz.

gigaohm A unit of extremely high resistance, reactance, or impedance, equal to 1,000,000,000 ohms = 10^9 ohms.

GIGO Abbreviation of *garbage in = garbage out,* an expression signifying that incorrect or improper input to a computer will produce meaningless output.

gilbert (William Gilbert, 1540–1603) A unit of magnetomotive force, equal to 1.26 times the number of ampere-turns. The SI (preferred) unit of magnetomotive force is the ampere (symbol, A); 1 gilbert = 0.796 A.

gilbert per centimeter See OERSTED.

gimbal A suspension device whose orientation can be changed without affecting the attitude of the body being suspended.

gimmick **1.** Colloquialism for any unnamed device. Also see GADGET. **2.** Colloquialism for any tricky manipulation or design. **3.** A low-value capacitor made by twisting two short pieces of insulated wire together.

gimp Colloquialism for the tinsel and cloth conductor used in some earphone cords.

Giorgi system The *meter-kilogram-second (mks)* system of units.

GJD Abbreviation of GERMANIUM JUNCTION DIODE.

glass A hard, brittle, amorphous, and usually transparent substance that is largely silicon dioxide. Glass has a multitude of uses in electronics, and there are several kinds, each having different electrical properties. The dielectric constant ranges from about 4 to 10; the dielectric strength ranges from about 20 to 300 kilovolts per millimeter.

glass arm A stiffness of the wrist or forearm, somewhat resembling writer's cramp, sometimes experienced by radiotelegraph operators or wire telegraph operators after prolonged use of a hand key.

glass bulb The glass enclosure of electron tubes and incandescent lamps.

glass capacitor A capacitor that uses thin glass as the dielectric, and usually has plates consisting of metal electroplated or electrodeposited on opposite faces of the glass. Also see MOLDED GLASS CAPACITOR.

glass diode A semiconductor diode molded in glass.

glass electrode A probe used with a pH meter; it consists of a thin-walled glass tube containing potassium chloride and mercurous chloride. Also see CALOMEL ELECTRODE.

glass envelope See GLASS BULB.

glassivation A procedure for encapsulating semiconductor devices in glass or other dielectric material.

glass-metal seal See GLASS-TO-METAL SEAL.

glass plate capacitor See GLASS CAPACITOR.

glass shell See GLASS BULB.

glass-to-metal seal A bond between glass and metal in electronic devices, such as vacuum tubes, feedthrough terminals, and glass capacitors.

glass tube A vacuum tube whose elements are housed in an evacuated glass envelope.

glide path The guidance beam used by aircraft making instrument landings.

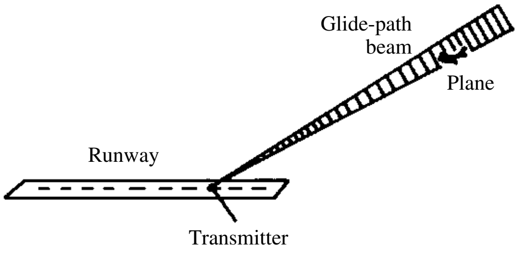

glide-path transmitter

glide-path transmitter A radio-frequency transmitter that produces a guidance beam for aircraft landing purposes. The aircraft follows the beam toward the runway.

glide slope See GLIDE PATH.

G line A microwave conductor consisting of a round wire coated with a dielectric.

glitch **1.** In a television image, a narrow, horizontal interference bar that moves vertically. **2.** A very short and unwanted high-amplitude transient that recurs irregularly in an electronic system.

glitter **1.** In radar, an echo or set of echoes that fluctuates rapidly in intensity because of motion in the target. **2.** A system in which moving devices are used to confuse enemy radar systems.

gloss factor For a reflecting surface, the ratio of reflected light in a selected direction to reflected light in all directions.

glossmeter An instrument for determining GLOSS FACTOR.

glow discharge The luminous electrical discharge resulting from the passage of current through ionized gas in a partially evacuated tube. The color of the glow is characteristic of the particular gas used.

glow-discharge microphone A device that produces audio-frequency currents from the action of sound waves in a glow-discharge tube.

glow-discharge tube A partially evacuated tube that contains two or more electrodes. The rarefied gas in the tube glows when a sufficient voltage is applied to the electrodes. See DISCHARGE LAMP, FLUORESCENT TUBE, and NEON BULB.

glow lamp See DISCHARGE LAMP.

glow modulator tube A gas tube whose luminous output can be modulated by an audio input signal.

glow potential The voltage at which glow discharge just begins in a gas-filled tube.

glow switch In fluorescent light circuits, an electron tube containing two bimetal strips that make mutual contact when heated by the glow discharge.

glow tube See DISCHARGE LAMP, FLUORESCENT TUBE, GLOW MODULATOR TUBE, NEON BULB, and STROBOTRON.

glow voltage See BREAKDOWN VOLTAGE, **2**.

glucinium See BERYLLIUM.

gluon A subatomic particle that is believed to bind quarks together (coined by Prof. Murray Gell-Mann of California Institute of Technology).

GM Abbreviation of *Geiger-Mueller* (see, for example, GEIGER-MUELLER TUBE).

gm Abbreviation of *gram.* Also abbreviated g.

g_m Symbol for TRANSCONDUCTANCE.

g-m Abbreviation of GRAM-METER.

gm-cal Abbreviation of GRAM-CALORIE.

gm-cm Abbreviation of GRAM-CENTIMETER.

G-M counter See GEIGER COUNTER.

GMT Abbreviation of GREENWICH MEAN TIME.

G/M tube See GEIGER-MUELLER TUBE.

gnd Abbreviation of GROUND.

g_o Symbol for OUTPUT CONDUCTANCE.

gold Symbol, Au. A precious metallic element. Atomic number, 79. Atomic weight, 196.967. Electrical contacts that must have low radio-frequency resistance, and that must be relatively immune to corrosion, are often plated with gold.

gold-bonded diode A germanium point-contact diode having a fine gold wire whose point is bonded to the germanium wafer. Its principal features are high forward current and almost constant, low reverse current.

gold doping The diffusion of gold into the base and collector regions of a diffused-mesa transistor; it shortens carrier storage time.

golden ratio A set of proportions used in the design of some speaker cabinets. The width (W) is $\frac{5}{8}$ of the height (H); the depth (D) is $\frac{5}{8}$ of the width. These correspond to a ratio $H:W:D = 1.000:0.625:0.391$. These proportions are thought by some acoustics engineers to result in the best possible sound quality.

gold-leaf electroscope See ELECTROSCOPE.

Goldschmidt alternator An early dynamo for generating radio-frequency power. The high-frequency energy was not generated directly by the machine, but by resonant circuits and frequency-multiplying interaction between components.

Golf Standard phonetic alphabet code word for the letter G.

goniometer **1.** Generically, any radio direction finder. **2.** An inductive coupler having a secondary coil rotated by a dial calibrated to read azimuth. The coupler, when used with a suitable antenna system, comprises a direction finder. **3.** A device for electrically varying the directional pattern of an antenna.

go-no test A test that indicates only acceptance or rejection of a device. No diagnosis is made.

GOTO In computers and programmable calculators, an instruction that, followed by a suitable label, directs the program to that label.

goto circuit In a digital-logic circuit, a device that senses the direction of electric current.

goto pair A pair of diodes connected in reverse series used in digital-logic circuits.

governor **1.** A device that prevents a motor or engine from running faster than a certain speed. **2.** Any device that limits a circuit parameter.

g parameters Conductance parameters obtained for the equivalent-pi model of a transistor: g_{BE}, g_{GC}, g_{CE}, and g_m.

gpc Abbreviation of GERMANIUM POINT CONTACT.

GPI Abbreviation of GROUND-POSITION INDICATOR.

GPRS Abbreviation for GENERAL PACKET RADIO SERVICE.

gr Abbreviation of *grain(s).*

graceful degradation A computer programming technique used to prevent debilitating breakdown by operating the system—even though several subsystems have malfunctioned; also known as *crippled mode.*

grad A unit of angular measurement equal to 0.9 degree.

graded-base transistor See DIFFUSED-BASE TRANSISTOR.

graded filter A power-supply filter that supplies direct-current output at various points in the filter sequence. Thus, the points in the powered equipment that can tolerate the least ripple are connected to the filter output, and those that can tolerate appreciable ripple are connected to the filter input; fairly critical points are connected to an intermediate position in the filter, such as at the junction of two chokes.

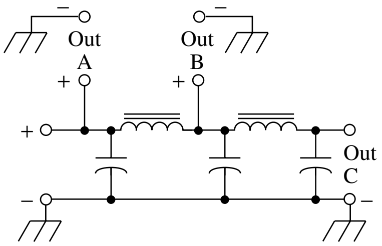

graded filter

graded-junction transistor A grown-junction transistor in which the temperature of the melt and the rate at which the crystal is pulled from it are closely controlled as the n and p layers are formed.

gradient The rate at which a variable quantity increases or decreases. See, for example, VOLTAGE GRADIENT.

gradient microphone A microphone whose output varies with sound pressure. Also see PRESSURE MICROPHONE.

Graetz bridge A full bridge rectifier (i.e., one having a diode in each arm).

Graffiti **1.** Trade name for a specialized written alphabet used for entering data into a handheld computer using a stylus. **2.** The entry of data into a handheld computer using a stylus. See HANDHELD COMPUTER.

grain boundary In a polycrystalline solid, a boundary between single crystalline regions.

gram **1.** Abbreviation, g. A unit of mass and weight, equal to 0.001 kilogram or 0.0353 ounce. **2.** A suffix meaning something drawn (written), or recorded, as in *radiogram* or *electrocardiogram.*

gram atom See GRAM ATOMIC WEIGHT.

gram atomic weight Also called *gram-equivalent.* The quantity of an element with a mass in grams that is equal to the atomic weight of the element. For example, copper has an atomic weight of 63.546; therefore, one gram atomic weight of copper has a mass of 63.546 grams.

gram-calorie Abbreviation, gm-cal. The amount of heat required to raise the temperature of one gram of water by one degree Celsius.

gram-centimeter Abbreviation, gm-cm. The work done by a force of one gram exerted over a distance of one centimeter. Also see JOULE.

gram-equivalent See GRAM ATOMIC WEIGHT.

grammar **1.** The sequence of words and/or abbreviations in a communication or part of a communication. **2.** The sequence of codes and/or commands in a high-level computer programming language.

gramme armature See GRAMME RING.

Gramme ring A type of armature for a motor or generator, consisting of an iron ring onto which is wound a coil of wire, each turn being connected to a commutator bar.

gram-meter Abbreviation, g-m. A unit of work equal to a force of one gram exerted over a distance of one meter. Compare JOULE.

gram-molecular weight See MOL.

gram molecule See MOL.

gramophone Archaic term for *phonograph.*

grandfather cycle A backup scheme in a magnetic reproduction system. The original records are retained for a period of time so that new copies of high precision can be made in case of loss.

grandfather file An original copy of a file on a magnetic disk or tape, retained as a source for reconstruction as needed. Usually, three generations of a file (grandfather, father, and son) are kept, each identified by a generation number. Sometimes the terms *grandparent*, *parent*, and *child* are used instead. See GENERATION NUMBER.

grand synthesizer A hypothetical child with a mind perfectly attuned to artificial intelligence (AI), who grows up to revolutionize the whole industry; a "grand master of AI."

granular carbon Carbon in the form of fine granules, used in the button of a carbon microphone.

granularity **1.** In a digital device, the smallest increment that can be differentiated. **2.** The limit of detail in a reproduction system.

granule One of many narrow frequency subbands, the combination of which composes a complete signal. Each subband carries its own specific data.

graph **1.** A presentation of data, particularly a depiction of the manner in which one variable or set of variables changes, with respect to another. Can be in the form of discrete points, curves, bars, columns, pie-shaped slices, etc. **2.** A curve or set of curves in a coordinate system.

graphical analysis The solution of problems through the use of graphic devices, such as vector diagrams, load lines, Nyquist plots, topological flow diagrams, etc.

graphical harmonic analysis See SCHEDULE METHOD.

graphical user interface Acronym, GUI (pronounced "gooey"). An operating system or software that makes it easy for lay people to use a computer. Commands are given by making choices from among items displayed on the screen. Popular versions use symbols, called *icons*, along with a pointing device, such as a *mouse* or *trackball.*

graphic documentation Records of data in the form of graphs, charts, tables, diagrams, etc.

graphic equalizer A device for tailoring the amplitude-versus-frequency response of a high-fidelity audio system. Consists of a splitter, several audio filters, and a mixer. The gain of each filter is adjustable via a slide potentiometer. The potentiometers are arranged on the front panel in such a way that their relative positions show the approximate shape of the response curve.

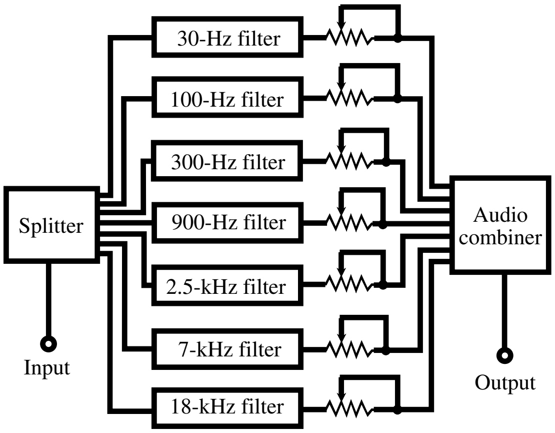

graphic equalizer

graphic instrument See GRAPHIC RECORDER.

graphic-level recorder An instrument that makes a permanent recording of signal amplitude as a function of some independent variable (such as time, frequency, or direction).

graphic panel In process control, a panel of illuminated lights or dials that display the status of a process.

graphic recorder An instrument in which a signal-driven pen or stylus makes a permanent record of a quantity on graph paper. The paper passes at a controlled speed beneath the pen.

graphics **1.** Diagrams, charts, photos, tables, or similar, often symbolic, artwork used to convey information. **2.** The video display in a computer system. **3.** Computer software designed specifically to create and edit illustrations.

graphic solution **1.** A method of depicting the solution(s) to a problem or equation by means of graphs. **2.** The process of solving problems or equations with the aid of graphs.

graphic terminal A display or plotter that provides visual output of a computer run.

graphite A soft form of carbon used in resistors, attenuators, contacts, brushes, vacuum-tube plates, cathode-ray tube coatings, etc.

graphite-line resistor An emergency, makeshift resistor consisting of a pencil line drawn on a piece of paper. The heavier the line for a given width and length, the lower its resistance.

graphophone Archaic term for *phonograph.*

grasping planning A scheme that a robot arm and gripper use to get hold of a particular object. It can use a vision system, a bar-code reader, tactile sensing, and/or proximity sensing. The robot controller (computer) must be programmed to recognize the input from these devices, and to seek out the object.

grass The background noise (noise floor) on the display of a spectrum analyzer; it also appears on certain types of radar displays.

grasshopper fuse A special type of spring-operated fuse. When it burns out, it actuates an alarm that alerts personnel of a possible circuit or system malfunction.

graticule Calibrated gridwork, as on the face of an oscilloscope or spectrum-analyzer screen.

grating A set of parallel, closely spaced, equidistant conductors or bars. When an electromagnetic field that has a wavelength comparable to the conductor spacing passes through the plane containing the conductors, diffraction occurs, producing an interference pattern. An adaptation of this device, called a DIFFRACTION GRATING, is used to separate visible light into its constituent spectral colors.

grating reflector A metal antenna reflector consisting of numerous parallel, straight, closely spaced conductors. When the conductor spacing is much smaller than the wavelength, the set behaves like a solid sheet of metal.

Gratz rectifier A form of full-wave rectifier circuit in a three-phase, alternating-current system.

gravitational constant Symbol, g. The acceleration produced by the attraction of a unit mass at unit distance; $g = 6.673 \times 10^{-11}\ \mathrm{N{\bullet}m^2/kg^2}$.

gravitational wave See GRAVITY WAVE.

gravity **1.** The universal force of attraction between material bodies—especially that force evidenced by the earth's drawing of bodies toward its center, causing them to have weight. **2.** Abbreviation, g. The rate at which a free-falling mass accelerates in a vacuum at the earth's surface; equal to 9.802 meters per second per second ($\mathrm{m/s^2}$). **3.** The rate at which a free-falling mass accelerates in a vacuum in the vicinity, or at the surface, of an astronomical object, such as a planet or star.

gravity cell An electrochemical cell in which the positive electrode is made of copper and the negative electrode is made of zinc. The copper electrode is placed at the bottom of a jar, and the zinc electrode is placed at the top. The jar is half filled with copper sulfate solution, and then filled with zinc sulfate solution. The solutions remain separate because copper sulfate has a higher specific gravity than zinc sulfate.

gravity wave A disturbance in a gravitational field, such as might be caused by a collapsing star. These waves might emanate from black holes or rapidly spinning neutron stars.

gray body A radiating body exhibiting constant spectral emissivity at all wavelengths. That is, the emitted energy is the same at all wavelengths and all frequencies.

Gray code A computer code in which the expressions representing sequential numbers differ in only one bit.

gray scale A reference scale for use in black-and-white television and video display images, consisting of several defined levels of brightness with neutral color.

Gray scale:
hypothetical 16-shade binary codes.

Code	Relative shade	Percent brightness
0000	Black	0.00
0001		6.67
0010	Very dark gray	13.33
0011		20.00
0100	Dark gray	26.67
0101		33.33
0110	Medium-dark gray	40.00
0111		46.67
1000	Medium gray	53.33
1001		60.00
1010	Medium-light gray	66.67
1011		73.33
1100	Light gray	80.00
1101		86.67
1110	Off-white	93.33
1111	White	100.00

gray tin A form of tin that exhibits some properties of a semiconductor at temperatures below 18 degrees Celsius.

greatest lower bound The largest value of a parameter that can be obtained without changing some characteristic of a circuit, program, or system.

Greek alphabet The 24-letter alphabet of the Greek language. Virtually all of letters are used as symbols in electronics and related sciences.

Green Book A specialized format for *compact-disk read-only memory (CD-ROM)* computer data storage media, developed by Sony and Philips. Requires the use of a *Compact Disk-Interactive (CD-I)* player for data retrieval. See also CD-ROM, ORANGE BOOK, RED BOOK, and YELLOW BOOK.

green gun The electron gun in a color cathode-ray tube whose correctly adjusted beam strikes only the green phosphors on the screen.

green video voltage In a three-gun color cathode-ray-tube circuit, the green-signal voltage, which actuates the green gun.

Greenwich Civil Time Abbreviation, GCT. Mean time counted from mean midnight at Greenwich, England, the location of zero meridian.

Greenwich Mean Time Abbreviation, GMT. Mean solar time at zero degrees longitude, also called the *Greenwich meridian* because it passes through Greenwich, England. In recent years, GMT has been supplanted by COORDINATED UNIVERSAL TIME (UTC) as the basis of standard time throughout the world.

grid **1.** The prime control electrode in a vacuum tube. Usually, it is a coil or mesh, but it can have other forms. Also called *control grid.* **2.** Any electrode in a vacuum tube placed between the cathode and the anode (plate) (e.g., *screen grid* and *suppressor grid*). **3.** Two sets of straight, uniformly spaced, parallel conducting wires or rods, one set perpendicular to the other. The conductors are electrically connected at all crossing points. Used as an electrostatic or electromagnetic shield, or as a reflector of electromagnetic waves. **4.** Two sets of uniformly spaced parallel lines, one set perpendicular to the other, used as a system of coordinates or as a basis for physical measurements.

grid capacitor **1.** A capacitor in series with the grid of a vacuum tube, used for blocking purposes. **2.** A bypass capacitor in a grounded-grid tube type amplifier. **3.** The capacitor in the grid tank circuit of a tube type oscillator or amplifier.

grid-cathode capacitance Symbol, *CGK*. The internal capacitance between the control grid and cathode of an electron tube. Also called INPUT CAPACITANCE.

grid characteristic The grid-current-versus-grid-voltage performance curve for a vacuum tube.

grid circuit The external circuit associated with the control grid of a vacuum tube.

grid current Symbol, I_G. Current flowing between the control grid and cathode in a vacuum tube.

grid cylinder The metal cylinder that acts as a control grid in a cathode-ray tube.

grid-dip meter **1.** A dip meter that contains a vacuum-tube oscillator; the indicating microammeter is in the grid circuit. **2.** Loosely, any frequency-sensitive wavemeter that indicates resonance by a marked dip in input (base, grid, and gate) current.

grid-dip oscillator See GRID-DIP METER.

grid dissipation **1.** The amount of power given up as heat in the grid circuit of a vacuum-tube amplifier. **2.** The maximum amount of power that a tube can safely dissipate as heat in the grid.

grid drive See GRID EXCITATION.

grid-driving power The signal power required by the control grid of a power tube.

grid emission Electron or ion emission by the control grid of a vacuum tube.

grid excitation Signal voltage or power applied to the control grid in a vacuum-tube amplifier circuit.

grid impedance Symbol, Z_G. The internal impedance of the grid-cathode path in a vacuum tube.

grid input impedance The impedance of the grid input section of a vacuum-tube circuit. It is a complex combination of grid impedance and the impedance of input-circuit components.

gridistor A special form of field-effect transistor with several channels.

grid-limiter resistor A resistor connected in series with the grid of a tube to limit grid current during the positive half-cycle of grid-signal voltage.

grid limiting The cutting off of plate current in a vacuum tube, with consequent limiting action, by means of a high, negative grid voltage developed by overdriving the grid.

grid loading effect The tendency of the internal grid-cathode path of a vacuum tube to load a tuned circuit—especially when the grid draws current.

grid locking A vacuum-tube fault in which the grid potential has become permanently positive because of excessive grid electron emission.

grid mesh The mechanical structure of a grid (e.g., gauze or a metal screen).

grid neutralization See GRID-NEUTRALIZED AMPLIFIER.

grid-neutralized amplifier A neutralized radio-frequency power amplifier in which the neutralizing capacitor is connected from the plate of the tube to the free end of a center-tapped grid-tank coil.

grid north In the grid system of navigation, the direction most nearly corresponding to geographic north.

grid-plate capacitance See PLATE-GRID CAPACITANCE.

grid pool tube A gas-discharge tube in which the cathode is a pool of mercury.

grid power loss Driving-power loss in the grid-input circuit of a power amplifier.

grid resistor A high-value resistor connected between the control grid and ground in a vacuum-tube amplifier circuit.

grid return The circuit path through which the control grid of a vacuum tube is returned to ground or to the negative grid bias supply.

grid-separation circuit A vacuum-tube circuit in which the control grid is grounded. See COMMON-GRID CIRCUIT.

grid swing The peak-to-peak variation of a grid excitation signal.

grid tank A resonant inductance-capacitance circuit operating in the control-grid circuit of a vacuum tube. Compare PLATE TANK.

grid tank capacitance The capacitance required to tune a GRID TANK to resonance.

grid tank inductance The inductance of the coil in a GRID TANK.

grid tank voltage The alternating-current (ac) voltage developed across the grid tank of a vacuum-tube circuit.

grid tuning Tuning of a vacuum-tube circuit by varying the capacitance, inductance, or both in the GRID TANK.

grid voltage **1.** Symbol, V_G. The direct-current (dc) bias voltage applied to the control grid of a vacuum tube. **2.** Symbol, $V_{G(ac)}$. The voltage of the radio-frequency (RF) or audio-frequency (AF) signal in the grid circuit.

grille A covering for an acoustic speaker, used primarily to protect the speaker cone, but also for esthetic appeal.

grille cloth A durable fabric often used for speaker grilles in high-fidelity sound systems. It transmits sound at all audio frequencies, but protects the speaker(s) and provides an attractive physical appearance.

gripper See ROBOT GRIPPER.

grommet An elastic washer inserted through a hole in a chassis to prevent accidental grounding of a conductor or to reduce wear on a cord or cable exiting the chassis.

groove **1.** See KEYWAY. **2.** The fine, spiral line cut into a phonograph disc when it is manufactured.

groove angle On a phonograph disk, the angle between the walls of the unmodulated groove.

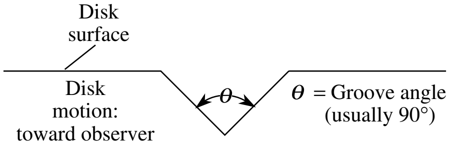
Disk surface
Disk motion: toward observer
θ
θ = Groove angle (usually 90°)

groove angle

groove speed In a phonograph recording or reproducing system, the speed of the cutter or needle, with respect to the disc. The speed is greatest near the outer edge of the disk, and least near the center.

gross content The overall amount of data contained in a message. It can be expressed in bits, bytes, words, or other units.

gross index One of a pair of indexes, used to give a reference in the *fine index*, a supplement; both indexes are used to locate computer records in storage.

gross-motion planning The method(s) that a robot uses to navigate in a general area without running into objects, knocking things over, falling down stairs, or losing its balance. It is often performed using a COMPUTER MAP of the work environment.

ground **1.** The earth in relation to electricity and magnetism. **2.** An electrical connection to the earth. **3.** The return point in a circuit. **4.** A short-circuit to the earth or to a circuit return point. **5.** A short-circuit to the metal chassis, case, or panel of a piece of equipment.

ground absorption The absorption (and resulting loss) of radio-frequency electromagnetic energy by the earth.

ground bus A conductor connected to an earth ground, and to which devices in a system are individually connected. The common ground points (e.g., chassis) of the individual devices are not directly connected to each other, so *ground loops* are avoided. This scheme minimizes the probability of ELECTROMAGNETIC INTERFERENCE to or from the system. Compare GROUND LOOP.

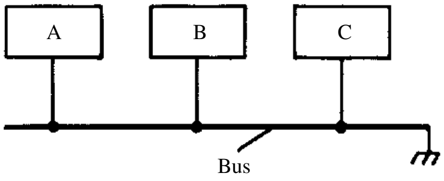

ground bus

ground clamp A device that provides a mechanical and electrical bond between a conductor and a ground rod or pipe. It is generally capable of passing a large amount of current.

ground clutter **1.** In a surface-based radar system, echoes from nearby buildings, hills, and other objects, producing blips or blobs near the center of the screen. **2.** In radar operations, an interference pattern on the screen, caused by accidental grounding in the system.

ground conductivity The ease with which the earth will carry electrical current. Saltwater has the best conductivity; dark, moist soil and fresh water (because of the mineral content) rate from fair to good. Sandy soil has the poorest ground conductivity.

ground conduit A pipe housing one or more ground leads.

ground connection **1.** A low-resistance connection to the earth. **2.** The common point, such as a chassis, to which zero-potential terminals of circuit components are connected.

ground constants The conductivity and dielectric constant of the earth for a particular kind of terrain and soil at a given location. Affects the behavior of radio antenna systems and electrical utility systems. Usually, high conductivity (low resistance) and low dielectric constant are desirable characteristics.

ground-controlled approach Abbreviation, GCA. In air navigation, a ground radar system that provides information for radio-directed aircraft approaches.

ground-controlled interception Abbreviation, GCI. A ground radar system by which an aircraft can be directed to intercept enemy aircraft.

ground current **1.** A direct electric current flowing into the earth from an electrical or electronic device, or into a device from the earth. **2.** An alternating current flowing between the earth and a device. **3.** An electric current flowing through the earth between two points. **4.** A current flowing in the normal ground (low-potential) line of a circuit.

ground detector A device that indicates whether or not a given circuit point is at direct-current ground.

grounded antenna See MARCONI ANTENNA.

grounded-base circuit See COMMON-BASE CIRCUIT.

grounded-cathode circuit See COMMON-CATHODE CIRCUIT.

grounded-collector circuit See COMMON-COLLECTOR CIRCUIT.

grounded-drain circuit See COMMON-DRAIN CIRCUIT.

grounded-emitter circuit See COMMON-EMITTER CIRCUIT.

grounded-gate circuit See COMMON-GATE CIRCUIT.

grounded-grid circuit See COMMON-GRID CIRCUIT.

grounded outlet An outlet with a receptacle having a ground contact that can be connected to equipment-grounding conductors.

grounded-source circuit See COMMON-SOURCE CIRCUIT.

grounded system A set of electrical conductors or a transmission line in which one conductor is deliberately grounded.

ground effect **1.** Modification or distortion of the ideal free-space directivity pattern of an antenna by reflections from, and absorption by, the earth. **2.** Effects of the earth on radio-wave propagation (e.g., the production of a *reflected wave* and a *surface wave*, neither of which can exist if the earth is not part of the signal path). **3.** Effects on antenna behavior, such as modification of the impedance and resonant frequency, caused by the proximity of the earth.

ground efficiency In an antenna system, the quality of the ground circuit. For some antenna systems, such as a balanced dipole at great height, this is not a consideration. Ideal ground efficiency (100%) results in zero ground loss.

ground environment **1.** See GROUND CONSTANTS. **2.** The ground characteristics in the vicinity of an unbalanced antenna working against ground. **3.** In aviation, the set of ground-based installations.

ground fault **1.** Loss of a ground connection. **2.** A short-circuit to ground.

ground-fault interrupter Abbreviation, GFI. A fast-acting electronic circuit breaker that opens the power-line circuit breaker to prevent electric shock or equipment damage when the path of current flow is through the earth.

grounding electrode A device, such as a ground plate or ground rod, that facilitates low-resistance connections to the earth.

grounding plate A metal plate connected to the earth, on which a person stands to discharge static electricity from the body.

grounding rod See GROUND ROD.

ground insulation Electrical insulation used between adjacent energized and grounded parts, such as transformer windings and metal cores.

ground level See GROUND STATE.

ground loop A closed current path resulting from improper grounding of the components in a system. A loop is formed when two devices are connected to each other, and also to separate earth grounds or to a single earth ground via conductors of appreciable length. The loops can act as antennas, increasing the likelihood of ELECTROMAGNETIC INTERFERENCE to or from the system. The use of a single GROUND BUS is preferred.

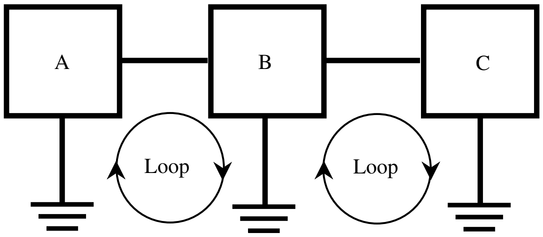

ground loop

ground mat A grid or network of conductors, connected to earth ground, for the purpose of improving the earth conductivity.

ground-mounted vertical antenna A vertical radiator mounted at the earth's surface or at the surface of a body of water, and fed with coaxial cable. The radiator can be any physical length, but it is tuned to resonance at the operating frequency.

The center conductor of the cable is connected to the base of the radiator, and the shield is connected to a ground system. Radial wires, usually buried just below the surface, minimize losses. The antenna can be made resonant on several frequencies by inserting multiple loading coils or traps in the radiator. The chief advantages of this antenna are unobtrusiveness and ease of installation. The main problems are marginal efficiency and susceptibility to human-made electromagnetic noise. See also RADIAL, 1. Compare GROUND-PLANE ANTENNA.

ground noise **1.** Electrical noise that results from a faulty ground connection. **2.** Background noise. **3.** In wire circuits, such as a telephone system, electrical noise that results from fluctuations in ground current.

ground plane **1.** A metal plate or a system of horizontal rods or wires mounted high on a mast, at the base of a vertical antenna, to provide a radio-frequency ground at a point several wavelengths above the surface of the earth. Also see GROUND-PLANE ANTENNA. **2.** In noise and interference tests, a sheet metal structure used to simulate the skin of an aircraft or missile. **3.** On a circuit board, a thin metallic sheet, usually bound to the underside, that serves as a common ground and RF shield.

ground-plane antenna A vertical radiator operated against a system of quarter-wave radials, elevated at least a quarter wavelength above the earth's surface and fed with coaxial cable. The center conductor of the cable is connected to the base of the radiator, and the shield is connected to three or four quarter-wave radials that run outward from the base of the radiator. At shorter wavelengths, a disk or cone can replace the radials. The chief advantages of this antenna are simplicity and high efficiency. The main problem is susceptibility to human-made electromagnetic noise. See also RADIAL, 1. Compare GROUND-MOUNTED VERTICAL ANTENNA.

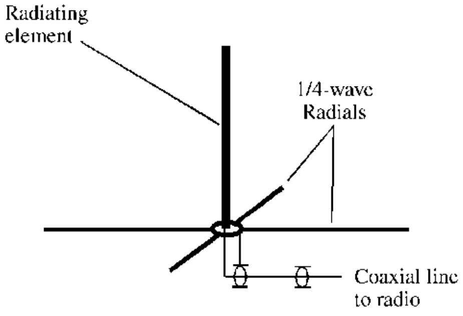

ground-plane antenna

ground plate A metal plate buried in the earth to provide a low-resistance ground connection.

ground-position indicator Abbreviation, GPI. A computer system that gives a continuous indication of an aircraft's position in terms of heading, elapsed time, and speed, with respect to the surface. This provides a more useful indication of the aircraft position than an *air-speed indicator* because it is not affected by high-altitude winds.

ground potential See ZERO POTENTIAL, **3**.

ground protection The use of a GROUND-FAULT INTERRUPTER.

ground-reflected wave A radio wave component that results from ground reflection.

ground reflection The reflection of a radio wave by the earth.

ground resistance The direct-current resistance of a connection to the earth, or the resistance between two points through the earth. The magnitude of the resistance depends on several factors: composition of the soil, amount of moisture, soil electrolytic action, and the area of contact with the earth.

ground return **1.** The point or path used to return a circuit to ground for completion. **2.** In radar, echoes returned from the earth's surface (including reflections from objects on it).

ground-return circuit A circuit, such as a single-wire telephone line, in which earth ground forms one leg of the circuit. Compare METALLIC CIRCUIT.

ground rod A strong metal rod driven deep into the earth as a point of ground connection.

ground speed The speed of an aircraft or missile, relative to the surface of the earth.

ground state The least-energy level of all possible states in a system.

ground support equipment Electronic surface-based apparatus upon which the functioning of a weapons system is dependent.

ground switch A switch for grounding an outside antenna during idle periods. Also called *lightning switch.* Ideally, antennas should be disconnected from equipment, as well as grounded, when not in use.

ground-to-air communication Radio or radar transmission from a land station to an aircraft in flight.

ground-to-ground **1.** Pertaining to communications between land-based stations. **2.** Pertaining to missiles intended for use between points on the surface of the earth.

ground-to-ground communication Communications between land-based stations.

groundtrack For an earth-orbiting satellite, the path followed by the point on the earth's surface from which the spacecraft appears directly overhead. For most satellites, this path shifts toward the west for each succeeding orbit, because the earth rotates eastward underneath the satellite. For geostationary satellites, the point never changes position on the surface. For satellites in equatorial orbits, the track follows the equator.

ground wave In wireless communications and broadcasting, an electromagnetic (EM) wave that consists of three distinct components: the *direct wave* (also called the *line-of-sight wave*), the *reflected wave,* and the *surface wave.* The direct wave is significant only when the transmitting and receiving antennas are connected by a line through free space without obstructions. The reflected wave, after returning from the earth's surface or a human made structure, combines with the direct wave (if any) at the receiving antenna. The surface wave travels in electrical contact with the earth. This occurs only with vertically polarized EM fields at frequencies below about 15 MHz. Below about 300 kHz, the surface wave propagates for hundreds or even thousands of miles. Sometimes the surface wave is called the *ground wave.*

ground wire A conductor between an equipment and a ground connection, either for circuit completion or for safety.

group **1.** A series of computer storage locations containing a specific record or records. **2.** The data in these locations. **3.** A record set having a common key value in a sorted file.

group busy In a telephone system, an audio signal indicating that all of the lines in a group are in use.

group code In digital communications, an error-detecting code used to verify a character group transferred between terminals.

group delay In a modulated signal, a delay in the transmission of data.

grouped-frequency operation In a two-wire communications system, the grouping of directional signals into certain frequency bands.

grouped records A set of data records in which the key of one record identifies the entire set.

grouping **1.** The arrangement of data into blocks or sets. **2.** On a phonograph disc, the insertion of gaps in the arrangement of grooves. **3.** Any periodic irregularity in the spacing of a data transmission. **4.** The bunching of grooves on a disc recording. **5.** In a facsimile system, occasional spacing errors between recorded lines. **6.** A mass of data arranged into groups, according to common characteristics.

group mark **1.** In telegraphy, an indicator that signals the end of a data unit. **2.** A character indicating the end of a character group; usually, it is a logical record that is addressed and processed as a unit.

group velocity The velocity at which a group of waves or a pulse is propagated.

Grove cell A closed-circuit primary cell in which the positive electrode, platinum, is immersed in nitric acid; the negative electrode, zinc, is immersed in sulfuric acid. The nitric acid is held in a porous cup, surrounded by a larger jar of sulfuric acid.

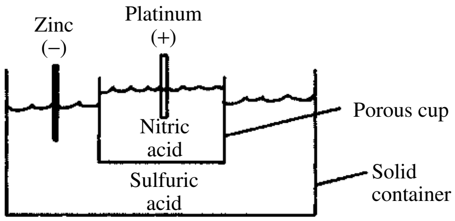

Grove cell

growler **1.** An electromechanical troubleshooting device that indicates the location of short circuits and grounds (especially in electric motors) by emitting a growling or rumbling sound. **2.** Any tester that provides an audible signal, which indicates electrical continuity.

grown-diffused transistor A transistor that is made by first growing the emitter and collector regions as a crystal, into which the base region is later diffused while the crystal is being pulled.

grown diode A semiconductor diode created by growing a layer of p-type material into n-type material (or vice versa) as the single-crystal material is being pulled from the melt.

grown junction A pn junction produced by adding impurities in various amounts to a crystal while it is being pulled from molten semiconductor material.

grown-junction diode See GROWN DIODE.

grown-junction photocell A grown-junction diode used as a photoconductive cell.

grown-junction transistor A transistor made by adding n-type and p-type impurities successively to a crystal in its molten state, then slicing the resulting npn formations from the finished crystal.

G-scan A rectangular radar display consisting of a laterally centered blip that "grows wings" as a target approaches. Horizontal and vertical displacement of the blip indicate horizontal and vertical aiming errors.

G scope See G SCAN.

GSR Abbreviation of GALVANIC SKIN RESPONSE.

G-string antenna In microwave operations, a communications path provided by a dielectric-coated wire that behaves like an extremely low-loss coaxial line with its outer conductor removed to infinity. A horn at each end matches the line to the transmitter and receiver. The term comes

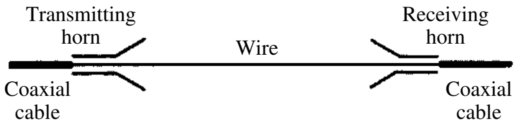

G-string antenna

from the first initial of Dr. George Groubau, inventor of the device, and the stringy appearance of the wire.

GTO Abbreviation of GOTO.

guard band A narrow unoccupied band of frequencies at the upper and lower limits of an assigned channel; its purpose is to prevent adjacent-channel interference by ensuring adequate separation between channels.

guard circle On a phonograph disk, an inner groove that prevents collision of the pickup with the spindle at the center of the disc.

guard circuit An auxiliary circuit added to an alternating-current bridge to compensate for the effects of stray capacitance in the bridge arms. One of its several forms is the WAGNER GROUND.

guarded input An input-terminal arrangement in which one terminal, maintained at the proper potential, shields the entire input-terminal combination.

guarding A method of short-circuiting a leakage current to ground. On a printed-circuit board, guarding is usually accomplished by the use of a large conducting foil surface near critical components.

guard relay A relay that ensures that only one linefinder will be connected to a line circuit when other line relays are in operation.

guard ring A metal ring (or other configuration) surrounding, but separate from, a charged body or terminal, for the purpose of evenly distributing the electric charge over the latter's surface.

guard shield A shield that encloses the input circuit of an amplifier or instrument.

guard terminal In a GUARDED INPUT, the terminal that shields the combination.

guard wire A grounded wire that is intended to catch and ground a broken high-voltage line.

Gudden-Pohl effect The tendency of an ultraviolet irradiated phosphor to glow momentarily when subjected to an electric field.

GUI Acronym for GRAPHICAL USER INTERFACE. Can be spelled out or pronounced "gooey."

guidance Electronic control of the path or course of a robot, missile, or other vehicle.

guidance system The complete electromechanical system for control of a robot, missile, or other vehicle. It consists of hardware and software. The hardware includes beacons, sensors, drive systems, rockets, etc. The software interprets data from, and transmits commands to, the hardware. The nature of the hardware and software depend on the application.

guidance tape In a guided missile, a magnetic tape containing computer instructions for steering the missile in a designated course.

guide See WAVEGUIDE.

guide connector See WAVEGUIDE CONNECTOR.

guided missile A missile whose progress to a target is controlled electronically by signals from a control station or by sensing equipment aboard the missile.

guided propagation A form of radio-wave propagation in which air masses of different temperatures or humidity levels cause refraction and/or reflection of electromagnetic waves, guiding signals over long distances with very little attenuation. Commonly observed at very-high and ultra-high frequencies.

guide elbow See WAVEGUIDE ELBOW.

guide flange See WAVEGUIDE FLANGE.

guide gasket See WAVEGUIDE GASKET.

guide junction See WAVEGUIDE JUNCTION.

guide load See WAVEGUIDE LOAD.

guide slot See KEYWAY.

guide wavelength See WAVEGUIDE WAVELENGTH.

Guillemin effect The tendency for a strip of ferromagnetic material to become straight in a strong magnetic field. This is a form of MAGNETOSTRICTION.

Guillemin line In radar operations, a special pulse-forming network for controlling modulation pulse duration.

guillotine capacitor A variable capacitor in which a sliding (instead of rotary) plate moves between two stator plates. Its name results from its resemblance to the infamous beheading apparatus.

gulp Several bytes of digital information.

gun See ELECTRON GUN.

Gunn diode A semiconductor diode that can operate as an oscillator in the ultra-high-frequency (UHF) and microwave parts of the radio spectrum. Oscillation takes place as a result of a negative-resistance effect in which, within a certain range of applied voltages, the current decreases as the voltage increases. The device is not particularly efficient. Only a small fraction of the input power results in useful signal output. The frequency and oscillation stability are sensitive to changes in temperature and bias voltage. See GUNN EFFECT, NEGATIVE RESISTANCE.

Gunn effect A semiconductor phenomenon named after J. Gunn of International Business Machines (IBM) who first observed and studied it in the 1960s. In certain types of diodes having a negative-resistance characteristic, oscillation occurs when the applied voltage is within a certain range. See GUNN DIODE, NEGATIVE RESISTANCE.

Gunn-effect circuit Any circuit exploiting the Gunn effect, especially a GUNN OSCILLATOR.

Gunn oscillator A discrete semiconductor microwave oscillator using a GUNN DIODE.

Gunnplexer A microwave transmitter consisting of a Gunn-diode oscillator connected to a horn antenna. Usually, the entire device is a compact, self-contained unit. The output power is low. When the device is placed at the focal point of a large dish antenna with the horn pointing toward the dish reflector (conventional dish feed), considerable transmission range is possible. Used primarily by experimenters, and in wireless communications links. See GUNN DIODE.

gutta percha A hard, rubberlike, organic insulating material. Dielectric constant, 3.3 to 4.9. Dielectric strength, 203 to 508 kV/in.

guyed tower In radio communications or in microwave links, a structure that is supported by one or more sets of guy wires to add strength and to prevent collapse.

guying The support of a radio communications or microwave-link tower by the use of one or more sets of guy wires.

guy insulator Also called *egg insulator.* An insulator designed to electrically break a guy wire while maintaining its ability to support a structure. Such an insulator has two slots with holes placed at right angles, in such a position that the wire will not separate even if the insulator breaks. The stress exerted on the insulator is compression, and the insulating material is stronger under this type of stress than under tension (pulling).

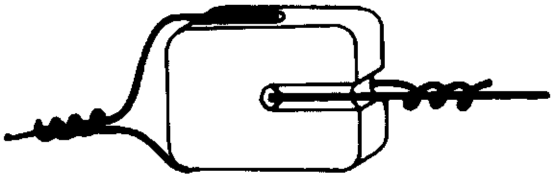

guy insulator

guy wire A bracing wire for antenna masts or towers.

gyrator An active (usually cascaded-transistor) device exhibiting nonreciprocal phase shift. It provides, among other functions, the simulation of inductance using capacitors.

gyro Contraction of GYROSCOPE.

gyro- A prefix meaning “pertaining to gyroscopes,” “containing a gyroscope,” or “behaving like a gyroscope.”

gyrocompass A type of compass in which a spinning gyroscope, acted upon by the earth’s rotation, causes the device to point to true north. Compare MAGNETIC COMPASS.

gyrofrequency The natural frequency of rotation of charged particles around the earth’s magnetic lines of flux.

gyromagnetic Pertaining to the magnetic properties of rotating electric charges (e.g., the effect of electrons spinning inside an atom).

gyromagnetic effect The tendency of a rotating body to become magnetized because of the magnetic field of the earth.

gyropilot See AUTOPILOT.

gyroscope A device that consists of a spinning wheel mounted in a gimbal. The shaft of the wheel will point in one direction, despite the movement of the earth beneath it.

gyrostat See GYROSCOPE.

G-Y signal In a color-television circuit, the signal representing primary green (G) minus luminance (Y). A primary green signal is obtained when the G-Y signal is combined with the luminance (Y) signal. Compare B-Y SIGNAL and R-Y SIGNAL.

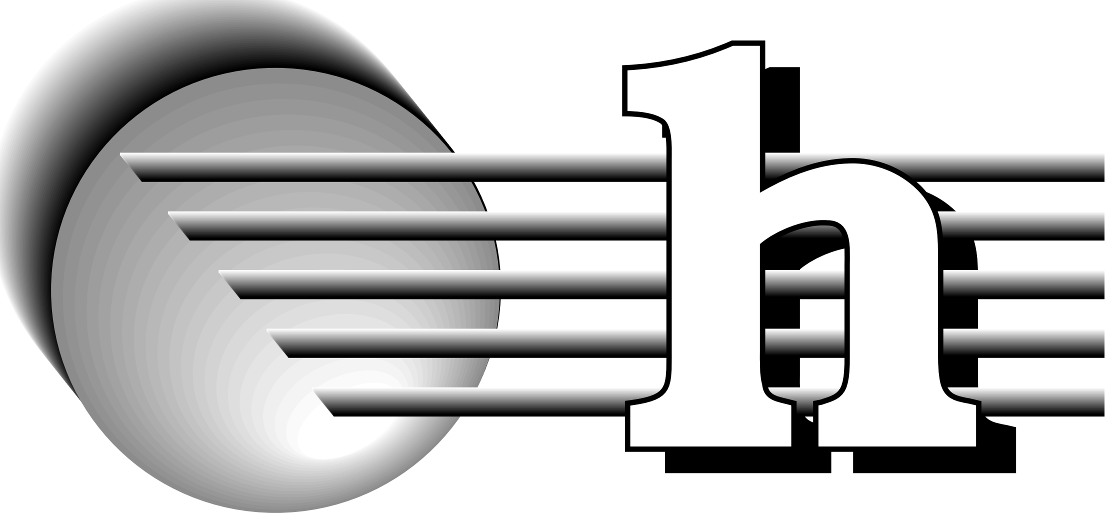

H **1.** Symbol for *magnetic field strength.* **2.** Symbol for MAGNETIZING FORCE. **3.** Symbol for HYDROGEN. **4.** Symbol for UNIT FUNCTION. **5.** Abbreviation of HORIZONTAL. (Also, hor and horiz.) **6.** Symbol for HENRY. **7.** Symbol for HARMONIC.

h **1.** Abbreviation of prefix HECTO-. **2.** Symbol for the PLANCK CONSTANT. **3.** Abbreviation of HOUR.

Haas effect See FUSION, **1**.

hack In computer networking, to access, and sometimes to change, sensitive data without authorization.

hacker A person knowledgeable in computer networking who uses his or her expertise to access, and sometimes alter, sensitive data. This is illegal and can sometimes be destructive. The intent, is rarely malicious, however, the activity is viewed as a challenge. Compare CRACKER.

HACKER program A computer program developed by Gerry Sussman as an early experiment with artificial intelligence (AI), to see how the machine would deal with complex decision-making problems.

hadron A subatomic particle consisting of quarks.

hafnium Symbol, Hf. A metallic element. Atomic number, 72. Atomic weight, 178.49. Readily emits electrons.

hahnium See DUBNIUM.

hailer **1.** A marine microphone-amplifier-speaker system for calling to other boats or persons ashore. **2.** A comparable system for land vehicles, such as police cars. Also see MEGAPHONE, **1**.

hair See HAIRLINE.

hair hygrometer A device for measuring relative humidity, in which a stretched hair is the moisture-sensitive element.

hairline A fine line used as an index or a graticule marker in a precision instrument.

hairpin coil A quarter-turn coil, so called from its resemblance to a hairpin.

hairpin coupling coil A hairpin coil used as a low-impedance primary or secondary coil for input or output coupling.

hairpin match A form of impedance-matching network used at the feed point of a half-wave dipole antenna. A short length of open-wire transmission line, short-circuited at the far end, is connected in parallel with the antenna at the feed point.

hairpin pickup A short, doubled length of wire that acts as a pickup coil at very-high and ultra-high frequencies.

hairspring A fine, usually spiral spring—especially the one in a movable-coil meter or the one connected to the balance wheel of a watch or clock.

hair-trigger Pertaining to extreme sensitivity of response, such as the tendency of a switching device to change state when excited by a weak pulse.

hair wire **1.** An extremely thin wire filament in a lamp or bolometer. **2.** Very small gauge wire (e.g., #44).

hal Abbreviation of HALOGEN.

half-add The sum of two binary digits, in which the carry operation is omitted. Thus 0 + 0 = 0, 0 + 1 = 1, 1 + 0 = 1, and 1 + 1 = 0.

half-adder In digital systems, an adder circuit that can handle the two binary bits that are to be added, but that cannot accommodate a carry signal. Compare FULL ADDER.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

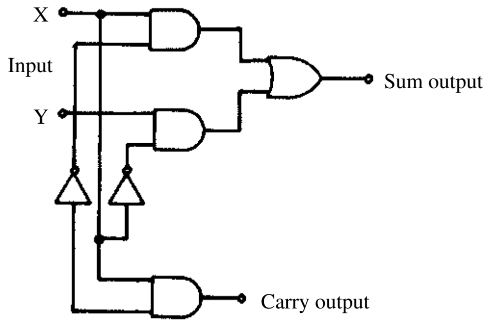

half-adder

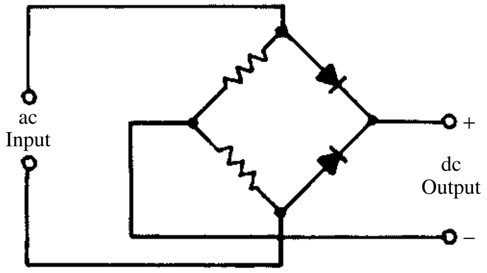

half-bridge

half-bridge A bridge rectifier that has diodes in two arms and resistors in the other two.

half-cell A voltaic cell consisting of a single electrode immersed in an electrolyte and having a definite difference of potential; it is, in effect, half of a primary cell. Also see HELMHOLTZ DOUBLE LAYER.

half-cycle Half of a complete alternation (i.e., 180 degrees of phase).

half-cycle magnetizer A magnetizer using half-cycles of rectified alternating current as the magnetic-field source.

half-duplex channel A communications channel in a HALF-DUPLEX SYSTEM.

half-duplex system In data communications, a system that transmits data in both directions, but not simultaneously. Compare FULL-DUPLEX SYSTEM.

half-lattice crystal filter A band-pass crystal-filter circuit using two piezoelectric crystals in a four-arm bridge. Also see CRYSTAL RESONATOR.

half-nut In a facsimile receiver, a device that guides the lead screw.

half-power point In a response curve or directional pattern, such as for a selective filter or a unidirectional antenna, the points on each side of maximum at which the power is 3 dB below the peak value.

half-power width In a directional antenna system, an expression of beamwidth. It is usually listed as the horizontal-plane angle, in degrees, between the half-power points in the main lobe of the directive pattern.

half-step **1.** In audio engineering, the frequency interval between two sounds, whose ratio is 1.06:1. **2.** The difference in pitch between the notes produced by two adjacent keys on a piano.

half tap A bridging circuit or device that can shunt another circuit with the least electrical disturbance.

half-track recorder A magnetic tape recorder that applies signals to both halves of a tape with a head that covers only half the tape's width in each of two directions. Also called *dual-track recorder.*

half-track tape Magnetic tape recorded by a HALF-TRACK RECORDER.

half-wave Half of a complete wave (i.e., a complete rise and fall in one direction). Its graphic representation is similar in appearance to that for a half-cycle.

half-wave antenna An antenna whose radiator measures an electrical half wavelength from end to end. Such a radiator is about 5 percent less than a free-space half-wavelength long, because of capacitive effects and the velocity factor of the conductor.

half-wave chopper A chopper that closes a circuit during only half the switching signal cycle.

half-wave dipole A center-fed antenna whose radiator measures 0.5 electrical wavelength from end to end. Also see DIPOLE ANTENNA.

half-wave doubler See HALF-WAVE VOLTAGE DOUBLER.

half-wave feeder See HALF-WAVE TRANSMISSION LINE.

half-wavelength Symbol $\lambda/2$. The distance that corresponds to 180 degrees of phase as an electromagnetic (EM) field is propagated in free space, it is related to the frequency by a simple equation:

$$L_{ft} = 492/f$$

where L_{ft} represents $\lambda/2$ in feet, and f represents the frequency in megahertz. If $\lambda/2$ is expressed in meters, then the formula is:

$$L_m = 150/f$$

where L_m represents the displacement in meters. In general, if v is the velocity factor (expressed as a ratio) in a given medium, then:

$$L_{ft} = 492v/f$$

and

$$L_m = 150v/f$$

Compare FULL WAVELENGTH, QUARTER WAVELENGTH.

half-wave loop antenna A loop antenna having a circumference of 0.5 wavelength with a break opposite the feed point. It is, in effect, a HALF-WAVE DIPOLE bent into a circle or square (although any symmetrical configuration can be used). The circle is the most efficient configuration.

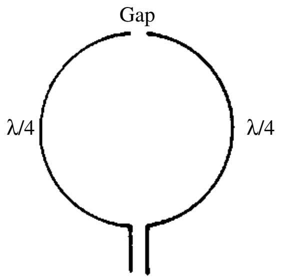

half-wave loop antenna

half-wave radiator An antenna consisting of a single, usually straight, active element that measures an electrical half wavelength from end to end. It is therefore a resonant element. A simple half-wavelength (λ/2) conductor with a high length-to-diameter ratio measures approximately 95 percent of λ/2 in free space. The element can be much shorter than free-space λ/2 yet remain λ/2-resonant if inductance is inserted in series with the radiator. The element can be much longer than free-space λ/2 yet remain λ/2-resonant if distributed capacitances are inserted in series with the radiator.

half-wave rectification The conversion of alternating current (ac) to direct current (dc) during half of each ac cycle. Also see HALF-WAVE RECTIFIER.

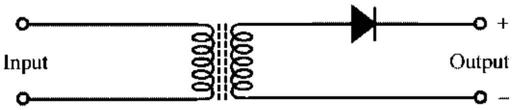

half-wave rectifier

half-wave rectifier A rectifier that delivers a half-cycle of direct-current (dc) output for every other half-cycle of applied alternating-current (ac) voltage. Because the successive dc half-cycles are 180 degrees apart, they have the same polarity. Compare FULL-WAVE RECTIFIER.

half-wave transmission line A transmission line measuring 0.5 electrical wavelength at the transmission frequency. The physical length is somewhat less than a free-space half wavelength because of the VELOCITY FACTOR of the line.

half-wave vibrator A vibrator (see INTERRUPTER) whose reed operates against only one stationary contact. Compare FULL-WAVE VIBRATOR.

half-wave voltage doubler A voltage-doubler circuit whose direct-current (dc) output has a ripple frequency equal to that of the alternating-current (ac) supply. Although its output is harder to filter than that of a full-wave doubler, this circuit has the advantage of a common ground. Compare FULL-WAVE VOLTAGE DOUBLER.

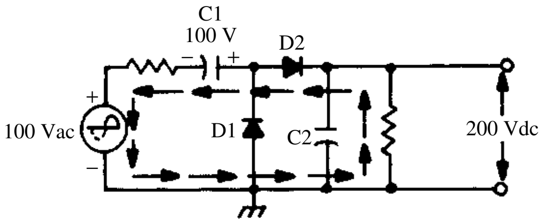

half-wave voltage doubler

halide A compound of a HALOGEN. Examples: sodium iodide, used as a scintillating crystal; ammonium chloride, used as the electrolyte in a dry cell.

halide crystal A halogen-compound crystal, such as mercuric iodide and sodium iodide, useful in detecting radioactivity.

Hall coefficient For a current-carrying conductor, the constant relationship between the Hall (transverse electric) field and the magnetic flux density.

Hall constant For a current-carrying conductor, the constant of proportionality k given by the equation $k = e/(im)$, where e is the transverse electric field (Hall field), i is the current density, and m is the magnetic field strength.

Hall effect A phenomenon observed in thin strips of metal and in some semiconductors. When a strip carrying current longitudinally is placed in a magnetic field that is perpendicular to the strip's plane, a voltage appears between opposite edges of the strip that, although feeble, will force a current through an external circuit. The voltage is positive in some metals (such as zinc) and negative in others (such as gold). Also see ETTINGHAUSEN EFFECT, NERNST EFFECT, and RIGHT-LEDUC EFFECT.

Hall-effect modulator A device that uses the HALL EFFECT to modulate a signal, or to mix two signals.

Hall-effect multiplier A device based upon the Hall generator and used in analog mathematical operations, such as multiplication and the extraction of roots.

Hall field The transverse electric field of a conductor carrying current in a magnetic field.

Hall generator A semiconductor device exhibiting the HALL EFFECT. It is a thin wafer or film of indium antimonide or indium arsenide with leads on opposite edges.

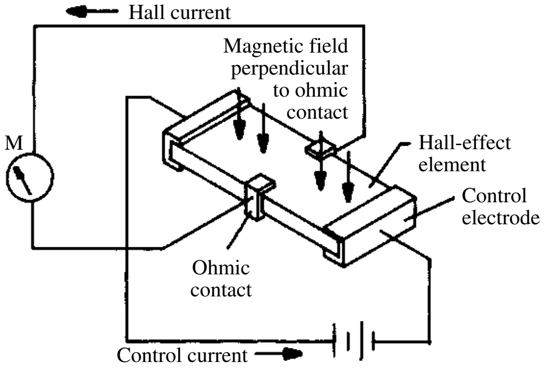

Hall generator

Hall mobility For a conductor or semiconductor, the product of conductivity and the HALL CONSTANT.

Hall network A resistance-capacitance null circuit whose general configuration is two cascaded high-pass tee-sections bridged by a high resistance. The circuit can be tuned with one potentiometer.

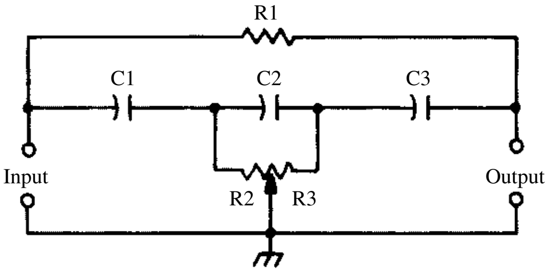

Hall network

Hallwacks effect The phenomenon (observed by Hallwacks in 1888) in which ultraviolet light falling on a polished zinc plate causes a negatively charged electroscope to which it is connected to discharge.

hallucination In complex computers and artificially intelligent systems, the generation or appearance of data for no apparent reason.

halo See AFTERGLOW and PERSISTENCE.

halo antenna A horizontally polarized antenna, consisting of a circular half-wave dipole whose ends are capacitance loaded. Commonly used at very high frequencies (VHF).

halogen Abbreviation, hal. A group of five very active nonmetallic elements whose similar chemical properties put them in group VIIA of the periodic table; they are *astatine*, *bromine*, *chlorine*, *fluorine*, and *iodine.*

Halowax A chlorinated naphthalene wax used as an impregnant for paper capacitors. Dielectric constant, 3.4 to 5.5. Resistivity, 10^{13} to 10^{14} ohm-cm.

halt A stop during the execution of a computer program run, often resulting from a HALT INSTRUCTION.

halt instruction An instruction in a computer program that causes a break in the program's execution, as by BASIC's STOP command, for example.

ham Colloquialism for AMATEUR RADIO operator.

ham radio See AMATEUR RADIO.

Hamilton's principle Also called the *principle of least action.* Motion tends to occur in such a way that the integral of the product of kinetic energy and elapsed time is minimal.

hammer **1.** The striking member in a WHEEL PRINTER. **2.** The clapper in an electric bell or gong.

hammer-and-wheel See WHEEL PRINTER.

Hamming code An error-correction code used in some digital communications circuits.

hand capacitance Also called *body capacitance.* Capacitive coupling effects between a circuit and the human body (e.g., as evidenced between an operator's hand and a device having extremely high impedance and poor grounding and/or shielding).

hand generator An electric generator operated by turning a hand crank.

handheld computer Also called *personal digital assistant (PDA)* or *palmtop computer.* The names PalmPilot and Palm are proprietary (Palm Computing, Inc.) and refer to specific families of handheld computers, although they might someday become generic and refer to handheld computers in general. A battery-powered portable computer, smaller than a notebook computer, and used for simple tasks such as note-taking and record keeping. Some units incorporate wireless modems for connection to the Internet. Others include paging, wireless fax, videoconferencing capability, remote-control capability, and other features. Many units can recognize a specialized form of handwriting so users can enter data with a penlike device called a *stylus.*

Handie-Talkie Abbreviation, HT. Tradename for a portable transceiver small enough to be held in the hand during operation.

hand key Also called *brass pounder.* An old-fashioned, hand-operated telegraph key, operated by manual downward pressure.

handoff In cellular communications networks, the changeover of reception from, and transmission to, a mobile or portable set from one repeater to another as the subscriber moves from one cell into another. When the subscriber is moving rapidly—for example, driving along a freeway, such transfers occur relatively often. When a subscriber is moving slowly, for example, walking along a trail, such transfers occur rarely. When a subscriber is in a fixed location, such transfers do not normally occur.

hand-operated device A device manipulated directly and manually by the operator's hand(s). Also called *manual device.*

hand receiver **1.** A single earphone that must be held against the ear. **2.** A telephone receiver.

hand rules See FLEMING'S LEFT-HAND RULE, FLEMING'S RIGHT-HAND RULE, and RIGHT-HAND RULE FOR WIRE.

handset See CRADLEPHONE.

handshaking **1.** A controlled, periodic exchange of synchronizing pulses between a digital transmitter and receiver. **2.** In a digital communications system, a method of error correction. The receiver detects nonstandard or improbable character sequences, and instructs the transmitter to repeat them for double-checking.

hand-type pointer In an electric meter, a spear-like pointer (resembling the hand of a clock), as opposed to a knife-edged pointer.

hand-wired Pertaining to electronic equipment wired by hand, rather than being assembled on printed-circuit boards. This form of construction is rarely seen nowadays, except in some radio-frequency power amplifiers.

hang AGC An automatic-gain-control (AGC) circuit whose action is sustained for a brief interval after an actuating signal has passed, an advantage in some applications. Also called *fast-attack/slow-release AGC.*

hangover In sound operations, the blurring or smearing of low-frequency (bass) notes by a poorly damped or poorly mounted loudspeaker.

hangup **1.** In phonograph operation, the state in which the same material is played repetitiously (i.e., the stylus does not move toward the spindle). **2.** In digital-computer operations, an unexpected break during a program run as a result of software or hardware failure. Sometimes called UNEXPECTED HALT.

H antenna See LAZY-H ANTENNA.

hard copy **1.** In digital computer operations, a readable document (printout) of material being translated to a form understood by a computer. **2.** Generally, written or typed documents, as opposed to data on other media, such as diskettes, tapes, CD-ROM, etc.

hard disk An electromechanical data storage medium commonly used in personal computers. Consists of several rigid disks, called *platters*, coated with ferromagnetic material.

hard-drawn wire High-tensile-strength unannealed wire.

hard dump See HARDWARE DUMP.

hard magnetic material High-retentivity magnetic material. Also see RETENTIVITY.

hardness **1.** The property that causes a material to resist penetration, deformation, scratches, etc. **2.** The penetrative ability of ultraviolet rays, X rays, or other ionizing radiation. Generally, the radiation hardness increases as the wavelength decreases, and as the photon or particle energy increases.

hardness tester A device for measuring the hardness of a solid in terms of the force required to penetrate its surface. Also see HARDNESS, **1**.

hard radiation In general, any radiation with high penetrating power. Usually, this term is used in reference to short-wavelength (high-energy) ultraviolet rays or X rays.

hard solder Solder that melts at a comparatively high temperature. Compare SOFT SOLDER.

hard vacuum A nearly perfect vacuum, that is, a medium essentially devoid of atomic or subatomic particles.

hardware **1.** Collectively, electronic circuit components and associated fittings and attachments. **2.** In a computer system, the electronic and electromechanical components (e.g., integrated circuits, keyboards, and disk drives) associated with operation. Compare SOFTWARE.

hardware availability ratio A figure depicting the availability of a computer system to do productive work; as a percentage, it is given by the formula:

$$A = 100(ta - td)/ta,$$

where A is the availability ratio, ta is the operational time, and td is the downtime over a specified time period.

hardware check A check on data being transferred within a computer, as done by hardware (e.g., a parity check).

hardware cloth A finely woven wire screen sometimes used in place of a metal plate for an antenna element, an antenna reflector, or a shielded enclosure. Especially useful when free-air circulation is required.

hardware dump During a computer program run, data sent to a storage device for later evaluation; it occurs at the time of a failure. Also called AUTOMATIC HARDWARE DUMP.

hardware engineer A person who designs and perfects the actual electronic circuitry in a system. The hardware engineer is not involved with the programming of the system.

hardware recovery A computer system's ability (through software or hardware) to recover from a failure (i.e., to proceed from the point of failure).

hardware serviceability ratio See HARDWARE AVAILABILITY RATIO.

hardwire **1.** To construct a circuit for direct-current conductivity. **2.** A circuit exhibiting direct-current conductivity over a complete, closed path.

hard-wire telemetry See WIRE-LINK TELEMETRY.

hard wiring **1.** In computer systems, functions or programs built directly into the machine hardware. In order to alter such functions or programs, the system wiring and/or components must be physically changed. **2.** A system interconnected entirely by wires and cables, and using no free-space links, such as radio or infrared.

hard X rays High-frequency (shortwave) X rays. Such radiation has high penetrating power. Compare SOFT X RAYS.

harmonic **1.** Symbol, H. In a complex sound or signal wave, a component whose frequency is a multiple of the FUNDAMENTAL FREQUENCY by a whole-number factor of 2 or more. **2.** Pertaining to whole-number multiples of the FUNDAMENTAL FREQUENCY of a sound or signal, as defined in **1**.

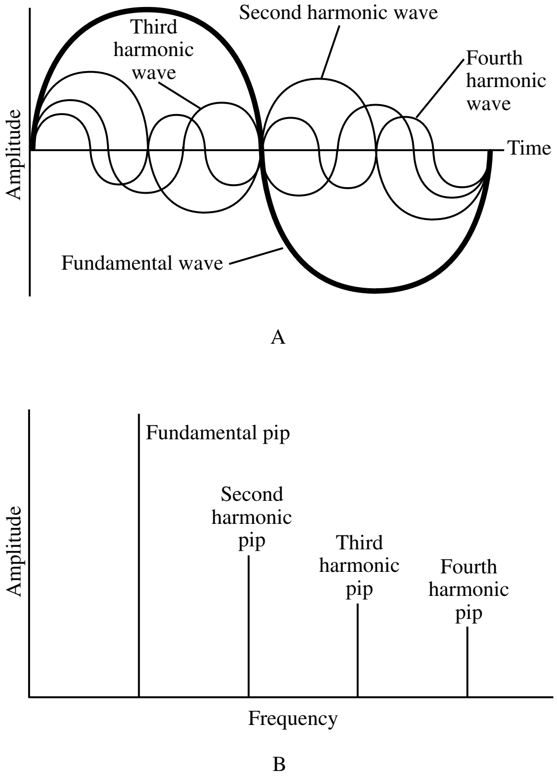

harmonic: waves (A) and spectrum-analyzer display (B)

harmonic accentuation Increasing the amplitude of harmonic components in a complex wave using filters, amplifiers, or special modes of operation.

harmonic accentuator A circuit or device, such as a harmonic amplifier or bandpass filter, for emphasizing signal harmonics.

harmonically related bands In communications, frequency bands arranged so that the frequencies in one band are harmonics of the frequencies in another band. An example of bands that are precisely related in this way are 4.1 to 4.3 MHz and 8.2 to 8.6 MHz. Various amateur radio bands are harmonically related to some extent, such as 40 meters (7.0 to 7.3 MHz) and 20 meters (14.0 to 14.35 MHz).

harmonic amplifier An amplifier, such as one used with a frequency standard, used to increase the amplitude of weak harmonics. Also see HARMONIC ACCENTUATION.

harmonic analysis **1.** The evaluation of the harmonic content of a complex wave. See, for example, HARMONIC WAVE ANALYZER; SCHEDULE METHOD, **2**; SPECTRUM ANALYZER; and WAVE ANALYZER. **2.** See FOURIER ANALYSIS.

harmonic analyzer See HARMONIC WAVE ANALYZER, SPECTRUM ANALYZER, and WAVE ANALYZER.

harmonic antenna An antenna operated at a harmonic of the lowest frequency at which it is resonant. For example, a half-wave dipole cut for 7.0 MHz, but used for transmitting and receiving at 21.0 MHz, is functioning at the third harmonic.

harmonic attenuation Reduction of the amplitude of harmonic components in a complex wave using filters, tuned amplifiers, or special modes of operation.

harmonic attenuator A circuit, device, or method of operation (such as a filter, tuned amplifier, special biasing, or special bypassing) for reducing the amplitude of harmonics.

harmonic component See HARMONIC.

harmonic composition See HARMONIC DISTRIBUTION.

harmonic content The amount of harmonic energy present in a complex wave. Also see HARMONIC-DISTORTION PERCENTAGE and HARMONIC RATIO.

harmonic-cut crystal Also called *overtone crystal.* A quartz crystal that, when operated in the proper circuit, oscillates at a harmonic of the (fundamental) frequency dictated by its thickness.

harmonic detector A detector tuned to respond to a harmonic of a signal.

harmonic distortion **1.** The generation of harmonics by the circuit or device by which the signal is processed. **2.** The deformation of the original signal that results from the action described in **1**. **3.** The disproportionate reproduction of a signal's harmonic components.

harmonic distortion meter See DISTORTION METER.

harmonic-distortion percentage In a signal containing harmonics, the harmonic energy as a percentage of the total signal energy (fundamental plus all harmonics). Also called *total harmonic distortion (THD).*

harmonic distribution For a given signal, the various frequencies and amplitudes of its harmonics, specified within a certain range of frequencies.

harmonic elimination The complete removal of one or more harmonics from a complex wave using a filter or special mode of operation.

harmonic eliminator A circuit or device, such as a band-suppression filter, for removing harmonics.

harmonic filter **1.** A bandpass filter for transmitting one or more harmonics of a complex input wave. **2.** A band-suppression filter for removing one or more harmonics of a complex input wave.

harmonic frequency **1.** In a complex wave, the frequency of a component that is a multiple of the FUNDAMENTAL FREQUENCY by a whole-number factor of two or more. **2.** A frequency that is a whole-number (two or more) multiple of another frequency to which it is referred. Compare NONHARMONIC FREQUENCY.

harmonic generator **1.** An oscillator operated so that it generates strong harmonics of the fundamental frequency. **2.** See FREQUENCY MULTIPLIER. **3.** See HARMONIC AMPLIFIER.

harmonic intensification See HARMONIC ACCENTUATION.

harmonic intensifier See HARMONIC ACCENTUATOR.

harmonic interference Interference resulting from the harmonics of radio or test signals.

harmonic motion Periodic motion typified by a swinging pendulum and illustrated by the plot of a sine wave.

harmonic oscillator A crystal oscillator whose output frequency is a harmonic of the crystal frequency.

harmonic percentage See HARMONIC-DISTORTION PERCENTAGE.

harmonic producer **1.** An oscillator that uses a tuning fork to establish the fundamental frequency. The output can be an odd or even harmonic of this frequency. **2.** See FREQUENCY MULTIPLIER. **3.** A nonlinear circuit used in a calibrator to generate markers at integral multiples of the fundamental frequency.

harmonic ratio **1.** In a complex wave, the ratio of harmonic energy to total signal energy (fundamental plus all harmonics). **2.** In a complex wave, the ratio of harmonic energy to fundamental-frequency energy.

harmonic reducer See HARMONIC ATTENUATOR.

harmonic reduction See HARMONIC ATTENUATION.

harmonic resonance Resonance of an antenna or a circuit at a whole multiple of the applied signal frequency.

harmonic ringing In wire telephony, the use of alternating-current signal harmonics for selective ringing.

harmonic series of tones A set of audio-frequency tones in which the frequencies can be specified by *f*, 2*f*, 3*f*, 4*f*, and so on.

harmonic suppression See HARMONIC ELIMINATION.

harmonic suppressor See HARMONIC ELIMINATOR.

harmonic tolerance The harmonic content permissible in a given system.

harmonic totalizer An instrument for measuring *total harmonic distortion.* See, for example, DISTORTION METER.

harmonizer A circuit that changes the frequency of an audio signal, or produces an output at several audio frequencies from an input having only one audio frequency. Used in sound recording for special effects.

harness A tied bundle of wires or cables for wiring electronic equipment.

harp antenna A vertical antenna consisting of a number of wires that fan out from point to point along a horizontal supporting wire.

hartley A unit of digital information equivalent to 3.32 bits. Used in certain computer applications.

Hartley oscillator A radio-frequency (RF) oscillator that uses a single inductor with a tap on the windings to provide the feedback. The amount of feedback is controlled by the position of the coil tap. A variable capacitor in parallel with the inductor determines the oscillating frequency and allows for frequency adjustment. The circuit uses about 25 percent of its output power to produce feedback. The other 75 percent of the power can be delivered to external circuits or devices. Compare COLPITTS OSCILLATOR.

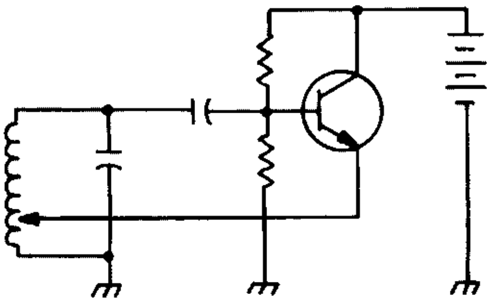

Hartley oscillator

hash **1.** Electrical noise, especially wideband noise with a characteristic hissing sound in a radio receiver. **2.** Undesirable or purposefully meaningless information, as used in a hash total (checksum).

hash filter A radio-frequency filter for eliminating HASH noise in a radio receiver.

hassium Symbol, Hs. Also called *unniloctium* (Uno). Atomic number, 108. The most common isotope has atomic weight 265. Classified as a transition metal. It is human-made and not known to occur in nature.

hat **1.** Also called *capacitance hat.* A small disk or set of wires attached to the end(s) of an antenna radiator, lowering the resonant frequency and increasing the usable bandwidth. **2.** A procedure for randomizing data.

hash total See CHECKSUM.

hatchdot pattern A television test pattern consisting of a crosshatch pattern with dots around its outer edges and one dot at its center.

hatted code A form of code in which randomization is used to maximize the difficulty of breaking the code.

Hay bridge An alternating-current bridge for measuring the inductance and Q of an inductor in terms of resistance, frequency, and a standard capacitance.

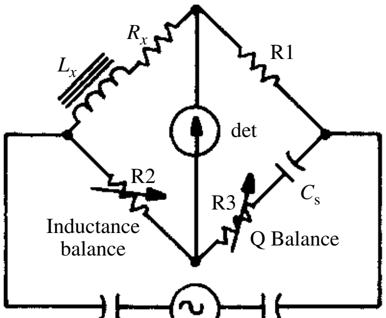

Hay bridge

haywire Loose, disorderly, or apparently careless wiring.

haz Abbreviation of HAZARD.

hazard Abbreviation, haz. A dangerous or potentially dangerous circuit, device, material, method, situation, or system (e.g., *electric-shock hazard*).

H beacon A form of homing beacon with an omnidirectional radiation pattern and a radio-frequency output of between 50 W and 2 kW.

H bend See H-PLANE BEND.

HCD Abbreviation of *hard-copy device.*

HCM Abbreviation of HALF-CYCLE MAGNETIZER.

HDB-3 code Abbreviation of HIGH-DENSITY BIPOLAR-3 CODE.

HDF Abbreviation of HIGH-FREQUENCY DIRECTION FINDER.

HDTV Abbreviation of HIGH-DEFINITION TELEVISION.

H-display See H-SCAN.

He Symbol for HELIUM.

head **1.** The top or operating portion of a device (e.g., microphone head or dynamic-speaker head). **2.** In magnetic recording and reproduction, the magnetic device (transducer) that delivers or picks up recorded impulses. **3.** In a hard disk or diskette drive, a transducer that delivers and picks up recorded data.

head alignment **1.** Positioning the cone of a dynamic speaker so that the voice coil moves freely (i.e., without rubbing against the core). **2.** Positioning a magnetic-recorder head so that a proper relationship to the moving tape is maintained.

head amplifier A self-contained amplifier or preamplifier in the head of a microphone or sound-on-film pickup.

head degausser A device used for the purpose of demagnetizing the head of a tape recorder. Unwanted magnetization can build up because of direct-current components in the driving signal.

head demagnetizer See HEAD DEGAUSSER.

head end In a television network or system, the location from which signals are sent to subscribers.

header **1.** A (usually glass) disk or wafer through which one or more leads pass and to which they are fully sealed. Can be used as the terminal base of an enclosed plug-in unit, such as a miniature coil, filter, or similar components. Also see GLASS-TO-METAL SEAL. **2.** A data set placed before other sets as a means of identifying them and, possibly, including control data pertinent to the sets so identified.

header capacitance Capacitance between or among the leads in a header (see HEADER, **1**).

header label A header recorded on a magnetic tape file (see HEADER, **2**).

head gap **1.** In computer disk or tape drives, the distance between the head and the magnetic medium. **2.** In audio operations, the spacing between tape-unit head electrodes; also called *gap width.*

heading The direction taken by a vehicle with reference to some point (such as a radio beacon, true north, or magnetic north).

headlight In radar operations, a small rotating antenna.

headphone A small acoustic transducer worn against the ear for listening to music without disturbing others, or for monitoring live or recorded material without being disturbed by noise in the environment. Also see RECEIVER, **2**.

headphone amplifier An audio-frequency amplifier designed and operated primarily to supply a signal to headphones.

headphone receiver A portable radio receiver, usually for AM and/or FM broadcast, consisting of a pair of headphones or a headset with the radio built into it.

head room **1.** In a high-fidelity sound system, the extent, measured in decibels, to which an amplifier can be operated beyond the zero point on its volume-unit (VU) meter without causing objectionable distortion on sound peaks. **2.** In tape recording, the region between the maximum recording level specified by the manufacturer of the equipment, and the amplitude at which tape overload occurs. It is specified in decibels.

headset An assembly consisting of one or two earphones, a headband, and a flexible cord. Also see HEADPHONE and RECEIVER, **2**.

head stack In magnetic recording, an assembly of two or more heads for multitrack service. Also see HEAD, **2**.

head station See BASE STATION.

head-to-tape contact In magnetic-tape recording or playback, physical contact between the tape and the head.

hearing aid A miniature audio-frequency device that amplifies sound for people with impaired hearing. It consists of a microphone, a high-gain amplifier, and an earphone or bone-conduction transducer.

hearing-aid battery A physically small battery designed for use with hearing aids. Such a battery is usually of the lithium type, or some other type that has long life under conditions of low current drain.

hearing loss A measure of hearing impairment. Generally expressed as the ratio, in decibels, of an individual's threshold of hearing to the normal threshold of hearing. Also see AUDIOLOGIST, AUDIOMETER, and AUDIOMETRIST.

heart fibrillation A condition in which the heart muscle twitches at random, rather than pumping blood normally. This can be caused by an electric shock through the heart of 100 mA to 300 mA. If normal heart function is not restored, death will follow.

heart pattern See CARDIOID PATTERN.

heart telemetry See ECG TELEMETRY.

heat A form of energy transferred by conduction, convection, or radiation between two bodies having different temperatures. The amount of heat is expressed in degrees, British thermal units, calories, joules, or kelvins.

heat aging **1.** The degeneration of a substance, aggravated by high temperatures. **2.** A test that indicates the immunity of a substance to degeneration because of high temperatures.

heat coil A device that disconnects a circuit when the temperature reaches a certain minimum level.

heat detector A sensor of heat. See, for example, BOLOMETER, INFRARED DETECTOR, MICRORADIOMETER, RADIOMETER, THERMISTOR, THERMOCOUPLE, and THERMOPILE.

heated-pen recorder See THERMAL RECORDER.

heated-stylus recorder See THERMAL RECORDER.

heated-wire flowmeter See HOT-WIRE FLOWMETER.

heated-wire sensor A hot wire used to discriminate between substances, according to how they affect its heating. See, for example, GAS DETECTOR (also usable as a vacuum gauge), HOT-WIRE ANEMOMETER, and HOT-WIRE MICROPHONE.

heat engine A machine that converts heat energy into mechanical energy.

heater **1.** The filament of an indirectly heated vacuum tube. **2.** The filament in an indirectly heated thermistor.

heater-voltage coefficient The amount of frequency change per volt of fluctuation in the filament voltage of a Klystron.

heat exchanger A device or system that removes heat from a hot body and transfers it to another body or to the surrounding air.

heat-eye tube An infrared-sensitive device used for the purpose of locating objects in visible darkness. The tube consists of a cathode-ray device that is sensitive to infrared radiation.

heat gradient The temperature difference between two points on a body, divided by the distance between the two points.

heating depth See DEPTH OF HEATING.

heating effect The production of heat (power loss) by electric current flowing in a conductor.

heating element **1.** See HEATER. **2.** The resistance element (such as a strip or coil) that generates heat in an electric-heating device.

heat loss **1.** Heat emitted by conduction, convection, or radiation from a body at a relatively high temperature. **2.** Power loss as a result of the heating effect of an electric current.

heat of fusion The amount of heat required to melt a unit mass of a solid that has reached its melting point.

heat of radioactivity Heat generated during the process of radioactive disintegration.

heat of reaction In a chemical or electrochemical reaction, the heat (in calories) absorbed or released.

heat of vaporization The amount of heat required to convert 1 gram of a liquid to a vapor without raising its temperature.

heat radiator See HEATSINK.

heat rays See INFRARED RAYS.

heat remover **1.** See HEATSINK. **2.** A forced-air or forced-liquid cooling system.

heat-resistant glass See PYREX.

heat-sensitive resistor See THERMISTOR.

heat-sensitive switch A make-and-break device, such as a thermostat, that is actuated by a change in temperature.

heat-shrink tubing An insulated flexible sleeving made from a plastic that shrinks permanently for a tight fit when heated; it is commonly used at the joint between a cable and connector.

heatsink A heat exchanger in the form of a heavy, metallic mounting base or a set of radiating fins. It conducts heat away from such devices as power transistors, heavy-duty resistors, or power tubes, and dissipates the heat into the surrounding environment via convection and radiation.

heatsink resistance The opposition offered by a heatsink to the flow of heat.

heat therapy **1.** The use of radio-frequency heating for therapeutic purposes. Also see DIATHERMY. **2.** The use of infrared rays for therapeutic purposes.

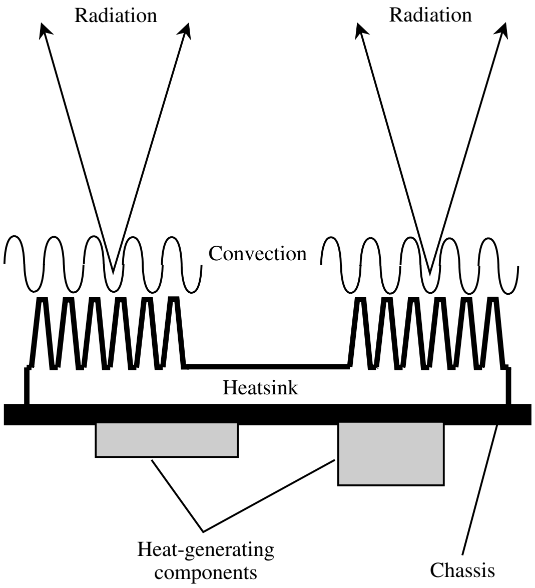

heatsink

heat transfer The movement of heat from one point to another by absorption, conduction, convection, or radiation.

heatronic Pertaining to the heating of a dielectric material subjected to a high voltage.

heat unit **1.** See BRITISH THERMAL UNIT. **2.** See CALORIE. **3.** See KELVIN.

heat waves See INFRARED RAYS.

heat writer See THERMAL RECORDER.

Heaviside-Campbell bridge A form of mutual-inductance bridge. Mutual inductance is determined without regard to the operating frequency.

Heaviside layer See KENNELLY-HEAVISIDE LAYER.

heavy hydrogen An isotope of hydrogen. The term is applied to *deuterium*, whose nucleus consists of one proton and one neutron, and also to *tritium*, whose nucleus consists of one proton and two neutrons.

heavy metal A metal having a specific gravity of 5.0 or higher. Examples: iron (7.85 to 7.88), lead (11.3), nickel (8.6 to 8.9), mercury (13.6), platinum (21.4).

heavy water Formula, D_2O. Water in which deuterium (HEAVY HYDROGEN), rather than ordinary hydrogen has combined with oxygen.

hecto- Abbreviation, h. A prefix meaning hundred(s), (i.e., 10^2).

hectometric wave An electromagnetic field whose wavelength is on the order of hundreds of meters (i.e., at least 100 meters, but less than 1000 meters). The frequency ranges from 300 kHz to 3 MHz.

hectowatt Abbreviation, hW. A unit of power equal to 100 watts. Seldom used; power in this range is usually expressed in terms of the WATT or the KILOWATT.

heelpiece A part of an electronic relay that provides mechanical support for the armature.

Hefner candle A unit of luminous intensity equal to 0.9 candela; the standard (German) is the HEFNER LAMP.

Hefner lamp A standard light source whose luminous intensity is 0.9 candela. It burns amyl acetate (banana oil) and its flame has been the standard of the HEFNER CANDLE, a unit of luminous intensity devised in Germany. Also see CANDLE POWER and LUMINOUS INTENSITY.

height control In a television receiver circuit, the potentiometer or rheostat that controls the vertical dimension of the picture by varying the amplitude of vertical scanning pulses.

height finder An altitude-measuring radar system.

height-position indicator Abbreviation, HPI. A radar displaying the height of a target, its angular elevation, and the slant range.

Heil oscillator An oscillator based on a special tube consisting of a heated cathode, first anode, metal cylinder, and second anode. Electrons emitted by the cathode pass through a hole in the first anode and become a beam, which passes through the cylinder and strikes the second anode (collector). Electron bunching in the cylinder causes energy to be transferred to a tank circuit between the cylinder and anodes.

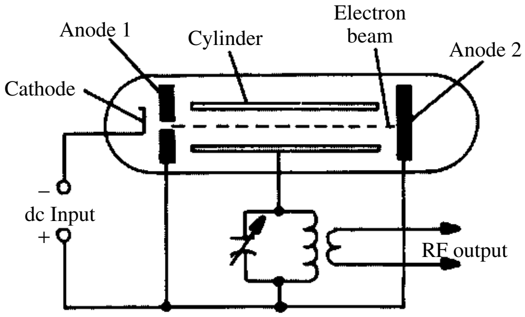

Heil oscillator

Heisenberg uncertainty principle See UNCERTAINTY PRINCIPLE.

hekto- See HECTO-.

heliacal cycle See SUNSPOT CYCLE.

helical antenna A spring-shaped antenna mounted perpendicular to a flat metal-plate reflector, an arrangement that produces circularly polarized waves in a narrow beam. It is used primarily at ultra-high and microwave radio frequencies.

helical-beam antenna See HELICAL ANTENNA.

helical line The helix in a backward-wave oscillator or traveling-wave tube.

helical potentiometer A potentiometer whose resistance element is a wire wound into a coil of several turns. The slider moves over the wire (or the larger coil) from one end to the other as the slider or coil is turned through several complete revolutions. Also called MULTITURN POTENTIOMETER.

helical scanning Radar scanning by an antenna that moves vertically as it moves horizontally, producing a spiral motion to the radiated beam.

helical sweep See SPIRAL SWEEP, **1**, **2**.

helical transmission line See HELICAL LINE.

helicone An antenna used at ultra-high and microwave frequencies, consisting of a helical radiator within a cone-shaped reflector.

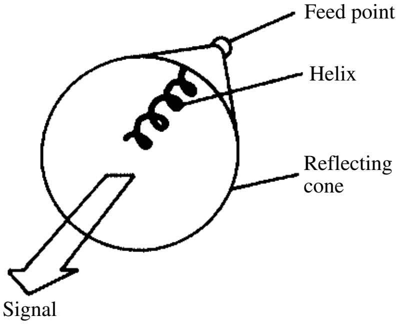

helicone

helionics The science of converting solar heat into electrical energy. The term is an acronym from *helio-* (pertaining to the sun) and *electronics.*

heliostat **1.** A servo-controlled motor-driven device that drives a mirror to keep sunlight trained upon a specific target. **2.** By extension, any similar device to keep a solar cell pointed to the sun.

helitron A form of oscillator used at ultra-high and microwave frequencies. The output frequency is variable over a wide range.

helium Symbol, He. A gaseous element. Atomic number, **2**. Atomic weight, 4.0026.

helium group The six inert gases in group 0 of the periodic table: *argon*, *helium*, *krypton*, *neon*, *xenon*, and *radon.*

helium-neon laser A laser in which the lasing substance is a mixture of helium and neon. Produces a characteristic brilliant red visible output. Also see HELIUM and NEON.

helix **1.** A single-layer coil. **2.** That which is coil-shaped (i.e., spiral in configuration). **3.** See HELICAL ANTENNA. **4.** See HELICAL LINE.

helix line See HELICAL LINE.

helix recorder An information recorder using a spiral method of scanning. The recording medium is usually drum-shaped.

Helmholtz coil A device consisting of two crossed-field primary windings in which an inductively coupled secondary winding rotates. The primary windings carry currents that differ in phase by 90 degrees. Rotating the secondary coil provides 360 degrees of continuously variable phase shift.

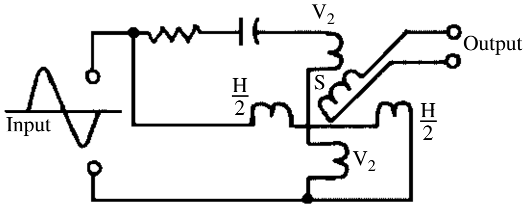

Helmholtz coil

Helmholtz double layer An intermolecular layer between a metal and an electrolyte in which it is immersed. It is formed when the adhesive force between the metal and electrolyte decreases the surface tension of the metal, causing positive ions to migrate from the metal into the liquid. The metal, charged negatively, and the electrolyte, charged positively, form a capacitor whose dielectric is the Helmholtz layer.

Helmholtz resonator An acoustic (sound) chamber whose geometry, in combination with the size of a small opening, results in resonance at a specific frequency.

HEM Abbreviation of *hybrid electromagnetic* (see, for example, HYBRID ELECTROMAGNETIC WAVE).

hemimorphic Pertaining to an object with ends that have unlike faces.

He-Ne laser See HELIUM-NEON LASER.

henry Symbol, H. The standard unit of inductance. It is the inductance exhibited by a closed circuit in which one volt is produced by a current changing uniformly at one ampere per second. This is a large unit of inductance; more common units are the MILLIHENRY and the MICROHENRY.

hermaphroditic plug A plug that has the prongs of a male plug and the recessed contacts of a female plug. Compare FEMALE PLUG and MALE PLUG.

hermetically sealed Constructed in manufacture so as to be permanently closed against the entry of air or other gases, dust, and moisture.

hermetic seal A permanent, air-tight seal that effectively prevents corrosion from elements in the outside environment.

herringbone pattern A pattern of interference in a television picture, so named because of its resemblance to the skeleton of a fish.

Herschel-Quincke tube An acoustic device that demonstrates sound interference. The device has two hollow cylinders, one of which can be adjusted in length. At the far end of the apparatus, the cylinders are joined together. The resultant amplitude depends on the difference in length between the cylinders. Sound wavelengths can be measured using this apparatus.

hertz Abbreviation, Hz. The standard unit of frequency (of periodic phenomena, such as alternating or pulsating currents); 1 Hz = 1 cycle per second. One Hz is an extremely small unit of frequency; more common units are the KILOHERTZ, the MEGAHERTZ, and the GIGAHERTZ.

Hertz antenna An ungrounded halfwave antenna fed by a transmission line attached to one end or to the center of the radiator. See, for example, CENTER-FED ANTENNA and END-FED ANTENNA. Compare MARCONI ANTENNA.

Hertz effect Ionization of a gas, produced by intense ultraviolet radiation.

Hertzian antenna See HERTZ ANTENNA.

Hertzian oscillator See HERTZ OSCILLATOR.

Hertzian radiation Radiation of electromagnetic (radio) waves.

Hertzian waves Electromagnetic waves in the radio spectrum, with wavelengths longer than those of infrared, visible light, ultraviolet, X rays, or gamma rays.

Hertz oscillator A damped-wave generator of oscillations, used by Hertz in his demonstration of radio waves in 1888 (verifying the earlier prediction by James Clerk Maxwell). The oscillator contains a spark gap supplied by an induction coil, attendant coils, capacitors (in the prototype, Leyden jars), and two large metal plates.

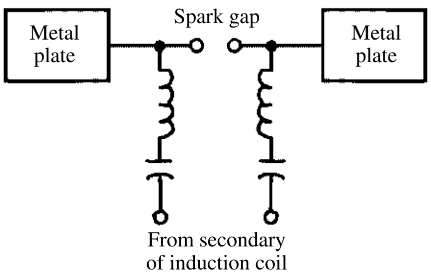

Hertz oscillator

Hertz vector A single vector that specifies the electromagnetic field (electric and magnetic components) of a radio wave.

hesitation As distinct from a halt, a brief break in a computer program run during which internal operations are occurring, such as data transfer to a peripheral.

heterochromatic Consisting of different frequencies, wavelengths, or colors. Compare MONOCHROMATIC.

heterodyne **1.** To beat one alternating-current signal against another to produce one or more beat-frequency signals. Also see BEAT FREQUENCY and BEAT NOTE. **2.** The whistle produced when two signals very close in frequency are mixed in a radio receiver. **3.** To combine radio signals in a mixer, the purpose of which is to produce a sum or difference signal for further processing.

heterodyne detection **1.** Signal detection by beating the incoming signal against one produced by a local oscillator. In this way, an unmodulated signal is made audible (the beat note is an audio frequency). **2.** Signal detection by a superheterodyne circuit.

heterodyne detector **1.** A detector that makes a radio-frequency (RF) signal audible by beating it against the RF signal of a local oscillator, the product being an audio-frequency (AF) beat note. **2.** The FIRST DETECTOR or SECOND DETECTOR in a SUPERHETERODYNE CIRCUIT. **3.** A combination linear detector and local RF oscillator used to detect and measure the frequency of test signals. Also see HETERODYNE FREQUENCY METER.

heterodyne eliminator See WHISTLE FILTER.

heterodyne filter See WHISTLE FILTER.

heterodyne frequency The frequency of the signal obtained by beating one signal against another.

heterodyne frequency meter A frequency-measuring device that contains a variable-frequency oscillator (VFO), a mixer, and an indicator such as an analog meter. The oscillator frequency is adjusted until zero beat is reached with the signal source. This condition is shown by a dip in the meter indication. An audio amplifier can be coupled to the output of the device instead of a meter; in this case the heterodyne appears as an audible tone whose frequency drops to zero when the oscillator frequency is equal to the signal frequency being measured.

heterodyne method See HETERODYNE, **1**.

heterodyne oscillator A signal generator whose output is the beat product of outputs from two internal oscillators. The output frequency can be either the sum or the difference of the oscillator frequencies, as selected by output filtering or tuning. See BEAT-FREQUENCY OSCILLATOR.

heterodyne reception Radio reception (especially in telegraphy) by means of the beat-note process. Also see HETERODYNE DETECTION, **1** and HETERODYNE DETECTOR, **1**.

heterodyne repeater A REPEATER in which the received signals are converted to another frequency before transmission.

heterodyne-type frequency meter See HETERODYNE FREQUENCY METER.

heterodyne wave analyzer A type of audio-frequency (AF) wave analyzer. The input signal is heterodyned in a balanced modulator with the signal from an internal tunable oscillator. One of the resulting sidebands is passed through a sharp bandpass filter, whose output actuates an alternating-current (ac) voltmeter. The internal oscillator is tuned slowly so that different components of the balanced-modulator output sideband pass successively through the filter. The amplitude-versus-frequency function of the input signal is determined by noting the meter readings as the internal oscillator is tuned.

heterodyne wavemeter See HETERODYNE FREQUENCY METER.

heterodyne whistle See HETERODYNE, **2**.

heterogeneous Pertaining to a group of objects or devices that have differing characteristics.

heterogeneous radiation Any broadband form of radiation. In particular, broadbanded radio waves, infrared radiation, visible light, ultraviolet, X rays, or gamma rays.

heterolysis The HYDROLYSIS of a compound into two oppositely charged ions.

heteropolar generator An electric generator in which the active conductors move through magnetic fields, first in one direction, then in the other direction. This is done by means of rotation in a nonuniform magnetic field. Most generators in common use are of this kind.

heuristic knowledge In artificial intelligence (AI), a form of machine knowledge in which a computer learns from its mistakes. As a complex program is repeatedly run over a period of time, the number of errors per run approaches zero.

heuristic program In artificial intelligence (AI), a program with which the computer solves a problem by trial and error, often learning in the process so that mistakes will not be repeated on subsequent runs.

Heusler's alloys Ferromagnetic alloys that contain one or more non-magnetic metals (such as aluminum, copper, or manganese).

hexadecimal number system An alphanumeric, base-16 system of number notation used in some computers. The system uses the usual digits plus the letters A through F to represent the numbers 10 through 15 (each place can only hold one symbol).

hex inverter A collection of six digital inverters, or NOT gates, contained within one package, usually an integrated circuit.

Hf Symbol for HAFNIUM.

HF Abbreviation of HIGH FREQUENCY.

Hg Symbol for MERCURY.

HH beacon In radionavigation, a nondirectional homing beacon.

hi **1.** Contraction of HIGH. **2.** Radiotelegraph symbol for a laugh, often verbalized by radio amateurs.

HIC Abbreviation of HYBRID INTEGRATED CIRCUIT.

HIDM Abbreviation of HIGH-INFORMATION DELTA MODULATION.

hierarchical password protection Also called *multilevel password protection.* A security feature that prevents unauthorized use of a computer, network, or database. The password levels allow users various degrees of control over the host machine.

hi-fi **1.** Contraction of HIGH-FIDELITY. **2.** In video recording, the addition of sound having high fidelity.

high **1.** Pertaining to a circuit point or condition at some potential above ground. **2.** The logical digit 1. **3.** The condition of having relatively large magnitude (e.g., HIGH FREQUENCY and HIGH VOLTAGE). **4.** Pertaining to the upper portion of a range, as in HIGH BAND or HIGH FREQUENCY. **5.** Characterized by greater-than-normal response or performance, as in HIGH *Q* or HIGH FIDELITY.

high band **1.** The higher or highest frequency band used in communications, testing, or processing, when several bands are available. **2.** The very-high-frequency (VHF) television channels 7 through 13. **3.** The communications frequency range from about 144 MHz to about 170 MHz.

high boost In sound recording and reproduction, the emphasis of high frequencies in an operating spectrum. Also called HIGH-FREQUENCY COMPENSATION.

high-C circuit A tuned circuit having high capacitance and low inductance at a given frequency. Such a circuit is characterized by high selectivity and low voltage. Compare HIGH-L CIRCUIT. Also see LC RATIO.

high contrast In an image, a limited range of gray values between black and white, or a similar condition in a color image (overbright whites, little shadow detail). Also see CONTRAST.

high definition In facsimile or television, a condition of minute detail so that the original scene is faithfully reproduced.

high-definition television Abbreviation, HDTV. A method of getting enhanced detail into a television (TV) picture and for obtaining better audio quality, compared with standard analog TV. There are several different schemes. The most noticeable feature is the crispness of the picture. This is vividly apparent in big-screen installations, which have traditionally suffered from image blurring. A standard TV picture has 525 lines per frame, but HDTV systems can have more than twice that number. Another important difference is that HDTV is digital; this minimizes susceptibility to interference. Interlacing is used in some systems.

high-density bipolar-3 code A communications or digital code in which two logic highs (ones) can occur in sequence, without the need for an intervening logic low (zero) to separate them.

high-efficiency linear amplifier A LINEAR AMPLIFIER with higher operating efficiency than is obtainable with conventional class-B linear amplifiers. Efficiencies on the order of 60% at 100% modulation are possible.

high-energy materials See HARD MAGNETIC MATERIALS.

high-energy particle **1.** A SUBATOMIC PARTICLE that has been given high velocity by a particle accelerator. **2.** High-speed subatomic particles emitted by the sun during a solar flare, or arriving from outer space.

high-energy physics The discipline dealing with the characteristics, properties, and applications of HIGH-ENERGY PARTICLES.

higher-order language See HIGH-LEVEL LANGUAGE.

high fidelity Abbreviation, hi-fi. Pertaining to an audio-frequency system that is very faithful to the signal it is processing (i.e., one characterized by extremely low distortion and wide frequency response).

high frequency Abbreviation, HF. Pertaining to frequencies in the 3- to 30-MHz band (wavelengths from 10 to 100 meters). Also see RADIO SPECTRUM.

high-frequency alternator A dynamo for generating radio-frequency energy.

high-frequency bias In a tape recorder, a high-frequency sinusoidal signal superimposed on the signal being recorded, for improving linearity and dynamic range.

high-frequency compensation See HIGH BOOST.

high-frequency converter See SHORTWAVE CONVERTER.

high-frequency crystal See HARMONIC CRYSTAL.

high-frequency direction finder Abbreviation, HDF. A direction finder operated at high radio frequencies (i.e., between about 3 MHz and 30 MHz).

high-frequency heating Electronic heating of materials by high-frequency energy. See, for example, DIELECTRIC HEATING and INDUCTION HEATING.

high-frequency resistance See RADIO-FREQUENCY RESISTANCE.

high-frequency speaker See TWEETER.

high-frequency trimmer **1.** In older high-frequency communications receivers, a low-value variable capacitor operated in parallel with a usually front-panel tuning capacitor to set the high-frequency end of the tuning range. See, for example, OSCILLATOR TRIMMER. **2.** A small variable capacitor used in conjunction with a larger tuning capacitor, the function of which is to permit precision tuning of the larger device.

high-impedance-state output current Pertaining to tests that ensure that an integrated circuit will not overload a bus line.

high-impedance voltmeter A voltmeter having an input impedance of at least several megohms.

high-information delta modulation A companded form of delta modulation, operating at comparatively low sample rate.

high-L circuit A tuned circuit having high inductance and low capacitance at a given frequency. Such a circuit is characterized by low selectivity and high voltage. Compare HIGH-C CIRCUIT. Also see LC RATIO.

high-level audio signal An audio-frequency signal that has been preamplified (e.g., the output of a compact-disc player). Compare LOW-LEVEL AUDIO SIGNAL.

high-level input current **1.** Pertaining to the testing of intertransistor leakage in an integrated circuit (IC) having multiple emitter inputs. **2.** The current into an IC input at minimum high-level voltage.

high-level language Also called *higher-order language.* A computer programming language in which the operator is easily able to communicate with the machine. It generally serves as an interface between a human programmer and the MACHINE LANGUAGE. Examples are BASIC, C, C++, COBOL, and FORTRAN.

high-level modulation In an amplitude-modulated transmitter, introduction of the audio at the final stage of radio-frequency amplification, permitting 100% modulation of the full-power signal.

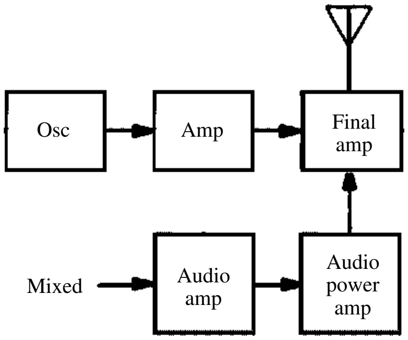

high-level modulation

high-level output current **1.** Pertaining to the testing of drive capability and fanout of an integrated circuit (IC). **2.** The current flowing from an IC output when in the high state.

high-level recovery Hardware recovery using data not involved in the failure, such as that on a magnetic storage medium. Also see HARDWARE RECOVERY.

highlight **1.** A bright area in a television picture. **2.** In computer data processing, the defining or setting-off of a block of data (such as text), with the intention of relocating, editing, or deleting it.

high-noise-immunity logic Abbreviation, HNIL. A form of bipolar digital logic designed for minimal sensitivity to noise. Also known as *high-threshold logic (HTL).*

high order Descriptive of the relationship between bits or digits in a word or number. Of two digits, the one holding the higher place value is the high-order digit (e.g., 2 is the high-order digit in 25).

high-pass filter A combination of capacitance, inductance, and/or resistance, intended to produce large amounts of attenuation below a certain frequency and little or no attenuation above that frequency. The frequency at which the transition occurs is called *cutoff*. At cutoff, the power attenuation is 3 dB with respect to the minimum attenuation. At frequencies above cutoff, the power attenuation is less than 3 dB. At frequencies below cutoff, the power attenuation is more than 3 dB. The simplest circuit consists of a parallel inductor or a series capacitor. The inductance-capacitance (LC) circuit has a combination of parallel inductors and series capacitors. In the resistance-capacitance (RC) circuit, resistors are substituted for the inductors. Compare BAND-PASS FILTER, BAND-REJECTION FILTER, LOW-PASS FILTER.

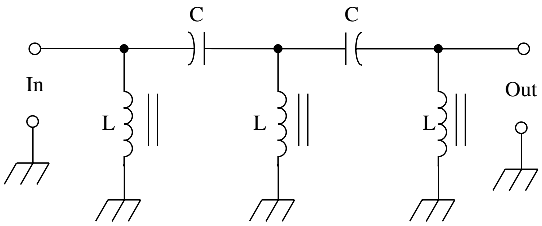

high-pass filter

high-pass-filter method A method of measuring the total harmonic distortion (THD) percentage using a high-pass filter to separate the harmonics from the fundamental. The output voltage V_o of the filter is compared with the input voltage V_i; then THD = $100V_o/V_i$.

high-performance Pertaining to apparatus designed for continuous operation with maximum reliability.

high-performance navigation system Acronym HIPERNAS. An electromechanical guidance system that is purely inertial and self-compensating.

high-potential test A high-voltage test of insulation, in which the applied voltage is continuously increased until the breakdown point of the dielectric is reached.

high-power rectifier A rectifier designed for high-voltage, high-current operation.

high Q For a component or circuit, a high value for the ratio X/R (reactance to resistance). This is a relative term because a particular numerical value of Q, considered high in one situation, might be regarded as low under other circumstances. Also see FIGURE OF MERIT.

high-resistance joint In the wiring of electronic equipment, a joint or connection between conductors that is poorly made, thereby introducing a high resistance between the parts.

high-resistance voltmeter A voltmeter having an input resistance of at least several megohms.

high-speed carry In computer operation, a carry into a column causing a carry out, circumventing the usual intermediate adding circuit.

high-speed diode See COMPUTER DIODE.

high-speed flip-flop A flip-flop having short switch-on and switch-off time.

high-speed oscilloscope An oscilloscope with excellent high-frequency and unit-function response. It can reproduce high-speed pulses faithfully.

high-speed relay A relay with a short make or short break interval.

high-speed transistor See SWITCHING TRANSISTOR.

high tension Pertaining to utility power-transmission lines on which there are very high voltages, typically 100 kilovolts (100 kV) or more.

high-tension line A power-transmission line carrying a very high voltage. It is generally used for the transfer of electric power over long distances.

high-threshold logic See HIGH-NOISE-IMMUNITY LOGIC.

high voltage **1.** A voltage considerably higher than those ordinarily encountered in a particular application. The term is comparative; a few hundred volts might be considered high in one situation, but low in another. **2.** In a cathode-ray tube, the voltage that accelerates the beam electrons. **3.** In a television receiver, the picture-tube anode voltage. **4.** In a vacuum-tube power amplifier, the plate supply voltage.

high-voltage probe A very-high-resistance probe for measuring high voltages with a low-range voltmeter.

highway A path over which multiple signals are propagated using time-division multiplexing.

HILAC Acronym for *heavy-ion linear accelerator.*

hill-and-dale recording See VERTICAL RECORDING.

hinged-iron instrument An alternating-current meter whose input transformer core is hinged in two parts. By means of a thumb trigger, the core can be opened, then closed around the current-carrying conductor that induces magnetism in the core; a secondary coil delivers current to the meter. Also called *clamp ammeter* or *clamp voltmeter.*

HIPERNAS Acronym for HIGH-PERFORMANCE NAVIGATION SYSTEM.

hipernick A high-permeability alloy of iron and nickel.

hipot Contraction of *high potential.* See HIGH-POTENTIAL TEST.

hiss **1.** A high-pitched sound rich in sibilants (*s*, *sh*, and *z* sounds) produced by random high-frequency fluctuations in current. **2.** The characteristic, high-pitched background noise (as in 1) accompanying super-regeneration. **3.** Internally generated noise in a communications receiver, amplified by the audio-frequency stages and appearing at the speaker or headphones.

hiss filter See HASH FILTER.

hit **1.** The occurrence of a lightning stroke at a specific point on the ground. Also called *direct hit.* **2.** The coincidence of two pulses.

H lines Magnetic lines of flux.

HLL Abbreviation of HIGH-LEVEL LANGUAGE.

H network A network of five impedances: two connected in series between the upper input and output terminals, two between the lower input and output terminals, and one shunted between the junctions of the series-connected impedances. Also called *H pad, balanced tee network*, and *balanced tee pad.*

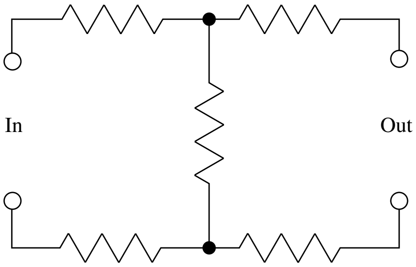

H network

Ho Symbol for HOLMIUM.

hobby robot A robot intended for amusement and, sometimes, for education. Such robots often take humanoid form. Some are programmable, and can give demonstrations, play musical instruments, and do other complex routines.

hockey-stick lead On a capacitor, resistor, or other component, a pigtail lead that is given a single crimp for easy insertion into a printed-circuit board.

hodoscope An instrument consisting essentially of closely spaced ion counters, for studying the path of an ionizing particle.

Hoffmann electrometer See BINANT ELECTROMETER.

hog horn A form of horn antenna used in microwave applications. It is generally used in the feed system of a dish antenna. The horn opening points in the direction of the feed waveguide.

hold **1.** To retain data in a storage device after the data has been duplicated in another location or device. **2.** A momentary halt of an operation or process. **3.** In a television receiver, a control that stabilizes the vertical or horizontal synchronization.

hold circuit **1.** See HOLDING CIRCUIT. **2.** In a television receiver, the circuit associated with the hold control(s). Also see HORIZONTAL-HOLD CONTROL and VERTICAL-HOLD CONTROL.

hold control See HORIZONTAL-HOLD CONTROL and VERTICAL-HOLD CONTROL.

hold current Symbol, I_h. The minimum current that will keep a normally open relay closed or a normally closed relay open.

hold electrode In a mercury switch, the electrode that is in permanent contact with the mercury.

holding beam In an electrostatic cathode-ray storage tube, the electron beam that generates replacement charges for those that were stored on the dielectric surface and then lost.

holding circuit In an electromechanical relay, a separate circuit that, when energized, keeps the relay actuated.

holding coil In an electromechanical relay, the extra coil that is associated with the holding circuit.

holding current **1.** Current in the holding coil of a relay. **2.** In a gas tube, the minimum current required to maintain ionization.

holding gun In an electrostatic cathode-ray storage tube, the electron gun that generates the holding beam.

hold mode A condition in which the output state of a digital-logic circuit remains unchanged while the input signals are removed.

hold-off voltage The highest voltage that can be applied to a flashtube without causing it to fire.

holdover The flow of current through the ionized path created by an electric arc.

hold time **1.** The time permitted for a weld to harden in resistance welding. **2.** In digital communications, the time for which a signal is maintained at a certain input after changing state at another specified input.

hole **1.** In a semiconductor atom, the vacancy resulting from the loss of an electron. When an electron is lost, so is its negative charge, leaving an equivalent net positive charge. This charge, like that of an electron, can move as a current carrier. **2.** The punched-out portion of a chassis or panel, through which wires can be passed or components mounted.

hole conduction In a semiconductor material, electrical conduction as a result of HOLE CURRENT.

hole current In a semiconductor material, the electrical current resulting from the movement of positive charge carriers (holes). Also see HOLE.

hole density The degree of concentration of holes in a semiconductor. Also see HOLE.

hole-electron pair In a semiconductor, a hole and a related electron. Each electron in the conduction band has a counterpart in the valence band, a vacancy (HOLE) left by the movement of the electron to the conduction band.

hole injection The creation of mobile holes in a semiconductor by applying an electric charge. Also see HOLE.

hole injector **1.** The emitter electrode of a bipolar pnp transistor. **2.** The metal whisker of a point-contact diode having an n-type wafer. **3.** The player of a forward-biased junction diode.

hole mobility The ease with which a hole moves within a semiconductor. Also see CARRIER MOBILITY.

hole storage See CARRIER STORAGE.

hole storage factor In a bipolar transistor biased to saturation, the amount of storage charge caused by excess base current.

hole trap In a semiconductor, an impurity that can cancel holes by releasing electrons to fill them.

hollow coil A coreless inductor.

hollow conductor Tubing used as a low-loss conductor at radio frequencies.

hollow core A core that is not solid throughout—especially one that has a central mounting hole.

holmium Symbol, Ho. A metallic element of the rare-earth group. Atomic number, 67. Atomic weight, 164.93. Forms highly magnetic compounds.

holocamera A camera for making holograms. Also see HOLOGRAM, **1** and HOLOGRAPHY.

hologram **1.** A wavefront recording made on photographic film by the process of HOLOGRAPHY. By changing the frequency of the light transmitted, various magnifications of the image can be obtained. Produces a true three-dimensional image. **2.** A visible, three-dimensional display projected in the air or underwater by means of lasers. They are often used at outdoor music concerts and other events.

holography A method of producing a wavefront recording of an object illuminated by laser light. The result, an interference pattern, appears meaningless when viewed in ordinary diffuse light. But when a point source of illumination is used, especially a laser, an image appears that is convincingly three-dimensional.

homeostasis The condition of being in static equilibrium.

home station See BASE STATION.

homing **1.** Guidance by means of an electronic beacon. The vehicle maintains a course toward the beacon. **2.** Guidance by means of some form of emission from a target object. The emission can be acoustic or electromagnetic energy.

homing antenna A direction-finding antenna—especially one on a mobile vehicle.

homing beacon A station radiating a beam for use in direction finding by mobile vehicles.

homing device A receiving device mounted on a mobile vehicle, and that continuously indicates the direction of a selected transmitting station that is the vehicle's destination.

homing relay A stepping relay that returns to its starting position after each switching sequence. Also see STEPPING SWITCH.

homing station See HOMING BEACON.

homodyne reception See ZERO-BEAT RECEPTION.

homogeneous **1.** Uniform in structure; similar at all points or locations. **2.** Consisting of many identical elements.

homogous field A field whose lines of flux in one plane pass through a single point.

homolysis The decomposition of a compound into a pair of neutral atoms or radicals.

homomorphism A one-to-one correspondence between the elements of two sets.

homopolar Pertaining to the union of atoms of the same polarity; nonionic.

homopolar generator A direct-current (dc) generator whose poles have the same polarity, with respect to the armature. Thus, no commutator is necessary.

homopolar magnet A magnet whose pole pieces are concentric.

homunculus In artificial intelligence (AI), a computer or robot that exhibits characteristics of a living being; especially, an ANDROID.

honeycomb coil A multilayer coil having a UNIVERSAL WINDING.

honeycomb winding See UNIVERSAL WINDING.

honker Also called *midrange speaker.* A loudspeaker that favors the middle audio frequencies. Compare TWEETER and WOOFER.

hood A light shield for a cathode-ray tube; it allows the screen to be viewed with a minimum of interference from room light.

Hooke's law Strain is proportional to the stress that produces it, as long as the ELASTIC LIMIT is not exceeded.

hook switch A switch that closes a circuit when a headset or handset is lifted from the resting position. The common telephone receiver uses such a switch.

hook transistor A four-layer pnpn semiconductor device, in which the outer p and n layers serve as emitter and collector, the inner n layer being the base. This places a p layer between the base and collector, resulting in a transistor that provides high alpha as a result of carrier multiplication by the additional junction in the collector layer.

hookup See SCHEMATIC DIAGRAM.

hookup wire Flexible, insulated wire used in the wiring of some electrical and electronic devices.

hoop antenna See CAGE ANTENNA.

hoot stop During a computer program run, a loop made evident by a sound signal.

hop In long-distance radio communications, the transmission of a wave and its subsequent return to the earth from the ionosphere; it is of importance mainly at low, medium, and high frequencies.

hor Abbreviation of HORIZONTAL. (Also, *H* and *horiz.*)

horiz Occasional abbreviation of HORIZONTAL. The usual form is *hor*; another alternate is *H.*

horizon **1.** For a specific location, the circle on the celestial sphere midway between the zenith (the point directly overhead) and the nadir (the point directly underfoot). **2.** Also called *visual horizon.* The set of points, as viewed from a particular location, where the sky and the earth appear to meet (i.e., the last visible part of the earth's surface from a given observation point). **3.** Also called *radio horizon.* The set of points, from a particular location at a given radio frequency, representing the maximum communications range via the ground wave under normal conditions. Also see ARTIFICIAL HORIZON.

horizontal **1.** Pertaining to objects or effects in a plane perpendicular to a line connecting the zenith (the point directly overhead) and the nadir (the point directly underfoot). **2.** Pertaining to that which is parallel to an assumed flat surface. **3.** Pertaining to width deflection on a cathode-ray tube.

horizontal AFC In a television receiver circuit, *automatic frequency control (AFC)* of the horizontal sweep. It keeps the receiver's horizontal scanning in step with that of the camera at the transmitting station.

horizontal amplification Gain provided by the horizontal channel of a device, such as an oscilloscope, cathode-ray electrocardiograph, or television receiver. Compare VERTICAL AMPLIFICATION.

horizontal amplifier A circuit or device that provides HORIZONTAL AMPLIFICATION. Compare VERTICAL AMPLIFIER.

horizontal angle of radiation For an antenna, the direction of maximum radiation in the horizontal plane (see HORIZONTAL, **1**), provided as an azimuth angle measured clockwise from geographic north.

horizontal angle of deviation In a communications circuit, the angular difference, in degrees, between the compass direction from which a received signal arrives, and the great-circle path connecting the receiving station with the transmitting station.

horizontal axis The axis that is parallel to an assumed horizontal surface (of the earth, for example) or the one so represented in a diagram. Also see X-AXIS.

horizontal beamwidth In a directional antenna system, the angle, measured in the horizontal plane, between the half-power points in the major lobe.

horizontal blanking See HORIZONTAL RETRACE BLANKING.

horizontal-blanking pulse In a television signal, the rectangular pedestal-shaped pulse that occurs between the active horizontal lines to achieve horizontal retrace blanking. Compare VERTICAL-BLANKING PULSE.

horizontal centering control See CENTERING CONTROL.

horizontal channel The system of amplifiers, controls, and terminations that constitute the path of the horizontal signal in an equipment, such as an oscilloscope or graphic recorder. Compare VERTICAL CHANNEL.

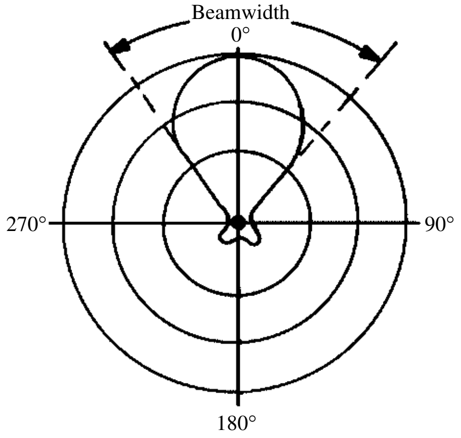

horizontal beamwidth

horizontal-convergence control In a color television receiver, the variable component for adjusting the horizontal dynamic convergence voltage.

horizontal coordinates See CARTESIAN COORDINATES.

horizontal deflection In a cathode-ray tube, the lateral movement of the electron beam on the screen. Compare VERTICAL DEFLECTION.

horizontal deflection coils The pair of coils that produces the magnetic field to horizontally deflect the electron beam in an electromagnetic cathode-ray tube. Also see DEFLECTION COIL.

horizontal-deflection electrodes See HORIZONTAL DEFLECTION COILS and HORIZONTAL DEFLECTION PLATES.

horizontal deflection plates In an electrostatic cathode-ray tube (typical of laboratory oscilloscopes and some early television picture tubes), a pair of plates that produces an electric field to horizontally deflect the electron beam. Compare VERTICAL DEFLECTION PLATES.

horizontal directivity The radiation or reception pattern of a directional antenna in the horizontal plane.

horizontal-drive control See DRIVE CONTROL.

horizontal dynamic convergence During the scanning of a horizontal line in a color picture tube, convergence of the electron beams at the aperture mask. Compare VERTICAL DYNAMIC CONVERGENCE.

horizontal field strength The field strength of signals passing through an antenna in a horizontal plane. Compare VERTICAL FIELD STRENGTH.

horizontal-field-strength diagram A plot of horizontal field strength, usually in polar form. Compare VERTICAL-FIELD-STRENGTH DIAGRAM.

horizontal flowcharting Flowcharting the movement of documents or files, rather than the data bits themselves, through a digital system.

horizontal frequency In television circuits, the horizontal scanning frequency [i.e., the frequency at which the horizontal lines are traced (generally 15.750 kHz)].

horizontal frequency response The gain-vs-frequency characteristic of the horizontal channel of an oscilloscope or graphic recorder. Compare VERTICAL FREQUENCY RESPONSE.

horizontal gain At a specified frequency, the overall amplification of the horizontal channel of an oscilloscope or graphic recorder. Compare VERTICAL GAIN.

horizontal-gain control A control, such as a potentiometer, for adjusting horizontal gain. Compare VERTICAL-GAIN CONTROL.

horizontal-hold control In a television receiver, the control for adjusting the horizontal oscillator frequency to prevent horizontal tearing of the picture. Compare VERTICAL-HOLD CONTROL.

horizontal hum bars Dark, horizontal interferential bars in a television picture, caused by HUM interference.

horizontal linearity The precision of gain and deflection in the horizontal channel of an oscilloscope, graphic recorder, or television receiver. A linear picture is a faithful (undistorted) reproduction of the original image. Compare VERTICAL LINEARITY.

horizontal-linearity control In an oscilloscope or television receiver, the control with which horizontal linearity is adjusted. Compare VERTICAL-LINEARITY CONTROL.

horizontal line frequency See HORIZONTAL FREQUENCY.

horizontal lock See HORIZONTAL-HOLD CONTROL.

horizontally polarized wave An electromagnetic wave whose electric lines of flux are parallel to the plane of the horizon. Compare VERTICALLY POLARIZED WAVE.

horizontal multivibrator In a television receiver, a 15.750-kHz multivibrator that generates the horizontal sweep signal.

horizontal oscillator In a TV receiver, the oscillator (usually a multivibrator) that generates the horizontal sweep signal. Compare VERTICAL OSCILLATOR.

horizontal output stage In a television receiver, an output amplifier following the horizontal oscillator. Compare VERTICAL OUTPUT STAGE.

horizontal output transformer In a television receiver, the output transformer in the horizontal-oscillator-output amplifier section. Also called FLYBACK TRANSFORMER.

horizontal polarization Pertaining to an electromagnetic wave whose electric lines of flux are horizontal. In general, when the radiating element of an antenna is horizontal, the electric lines of flux in the transmitted waves are horizontal, and the antenna is most sensitive to incoming signals whose electric lines of flux are horizontal. Compare VERTICAL POLARIZATION.

horizontal positioning control See CENTERING CONTROL.

horizontal quantity The quantity measured along the X-axis of a graph represented by the horizontal deflection of an oscilloscope beam. Compare VERTICAL QUANTITY.

horizontal recording See LATERAL RECORDING.

horizontal repetition rate See HORIZONTAL FREQUENCY.

horizontal resolution In a television picture, the number of picture elements (*pixels*) that can be discerned in a horizontal scanning line. Compare VERTICAL RESOLUTION.

horizontal retrace In a cathode-ray device, such as an oscilloscope or television receiver, the rapid return of the electron beam to its starting point after completing a horizontal sweep of the screen. Compare VERTICAL RETRACE.

horizontal retrace blanking In oscilloscopes and television receivers, the automatic cutoff of the electron beam during a horizontal retrace period, preventing an extraneous line on the screen during the period. Compare VERTICAL RETRACE BLANKING.

horizontal scanning **1.** The lateral sweeping of the electron beam in a cathode-ray tube. **2.** The sampling of x-axis values in a repetitive or nonrepetitive sweep of that axis.

horizontal scanning frequency See HORIZONTAL FREQUENCY.

horizontal sensitivity The signal voltage required at the input of a horizontal channel for full horizontal deflection. Also see HORIZONTAL GAIN. Compare VERTICAL SENSITIVITY.

horizontal signal A signal serving as a horizontal quantity. Compare VERTICAL SIGNAL.

horizontal sweep **1.** In a cathode-ray tube, the horizontal movement of the spot on the screen; in particular, the movement from left to right, during which a line of the image is formed on the screen. **2.** The circuit that produces horizontal sweep.

horizontal sweep frequency The frequency at which horizontal sweep occurs; in a television receiver, it is generally 15.750 kHz. Also called *horizontal sweep rate.* Compare VERTICAL SWEEP FREQUENCY.

horizontal sweep rate See HORIZONTAL SWEEP FREQUENCY.

horizontal-sync discriminator In a television receiver, a circuit that compares horizontal sync-pulse phase with the phase of the signal from the horizontal sweep oscillator.

horizontal synchronization In a television receiver, synchronization of the horizontal component of scanning with that of the transmitting camera. Also see HORIZONTAL SYNC PULSE. Compare VERTICAL SYNCHRONIZATION.

horizontal sync pulse In a video signal, the pulse that synchronizes the horizontal scanning component in a television receiver with that of the camera; it also triggers horizontal retrace and blanking. Also see BACK PORCH. Compare VERTICAL SYNC PULSE.

horizontal wave See HORIZONTALLY POLARIZED WAVE.

horizontal-width control See WIDTH CONTROL, **1**, **2**.

horn A radiating device that is essentially a cylindrical or rectangular pipe, whose surface flares from a narrow entry to a wide exit. See, for example, HORN ANTENNA, HORN SPEAKER, and MEGAPHONE.

horn antenna An antenna used for transmission and reception of signals at microwave frequencies. There are several configurations; most have a square or rectangular cross section. The antenna provides a unidirectional radiation and response pattern, with the favored direction coincident with the opening of the horn. The feed line consists of a waveguide that joins the antenna at the narrowest point (throat) of the horn. This type of device is used alone in microwave transmitters known as *Gunnplexers*. The device works well in the feed systems of large dish antennas; the horn is pointed toward the center of the dish, in the opposite direction from the favored direction of the dish. See also DISH ANTENNA, GUNNPLEXER.

horn cutoff frequency The lowest frequency at which an exponential horn will function properly.

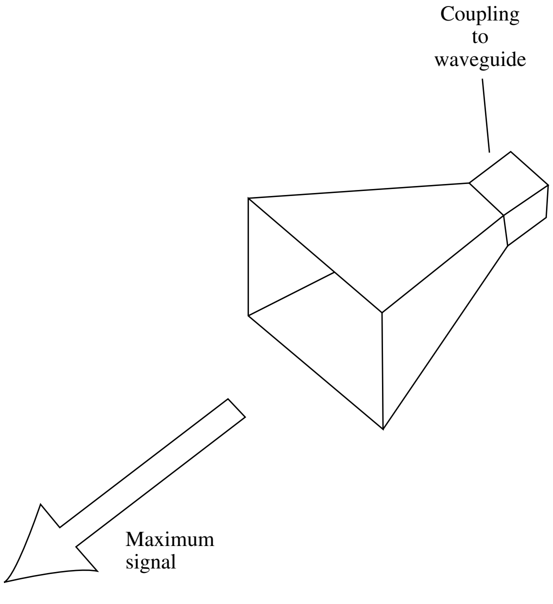

horn antenna

horn loading In a sound-transmission system, a form of propagation that makes use of a horn-shaped speaker.

horn mouth The wider (radiating) end of a horn antenna or speaker. Compare HORN THROAT.

horn radiator See HORN ANTENNA.

horn speaker A loudspeaker using a horn in conjunction with an acoustic transducer. It generally produces sound of low fidelity, and is intended for use in marginal environments, such as outdoor public-address (PA) systems.

horn throat The narrower (input) end of a horn antenna or speaker. Compare HORN MOUTH.

horsepower Abbreviation, hp. A unit of power equal to 746 watts. It is generally used to express mechanical power delivered by engines and electric motors.

horsepower-hour Abbreviation, hp-hr. A unit of energy or work represented by the equivalent of one horsepower delivered for a period of one hour; equal to 746 watt-hours.

horseshoe coil See HAIRPIN COIL and HAIRPIN COUPLING COIL.

horseshoe magnet A (usually permanent) magnet having the shape of a horseshoe or a U and a rectangular cross section.

host A programmable computer that gathers and stores the information from all of the data-entry terminals in a system.

HOT **1.** Abbreviation of HORIZONTAL OUTPUT TRANSFORMER. **2.** Abbreviation of HORIZONTAL OUTPUT TUBE.

hot **1.** Pertaining to an object from which heat energy is transferred to the environment by radiation, convection, and/or conduction. **2.** Charged to an electrical potential, either positive or negative, with respect to common ground. **3.** Pertaining to a conductor that carries a high radio-frequency (RF) current or voltage.

hot carrier In a semiconductor, a charge carrier (electron or hole) whose energy is higher than that of majority carriers normally encountered in the same material.

hot-carrier diode A semiconductor diode having a metal base that receives HOT CARRIERS from a semiconductor layer. The unit has a fast switching speed because there are virtually no minority carriers—either injected or stored.

Hotel Phonetic alphabet code word for the letter H.

hot-filament gas detector A gas detector in which the sensor is a heated filament acting as one arm of a Wheatstone bridge circuit. The bridge, previously balanced, becomes unbalanced when gas impinges upon the filament and changes its resistance.

hot junction The heated junction in a two-junction thermocouple circuit. Compare COLD JUNCTION.

hot line **1.** An energized wire, transmission line, or busbar. **2.** A private communications channel (wire or radio) kept in constant readiness for instant use between persons of high authority.

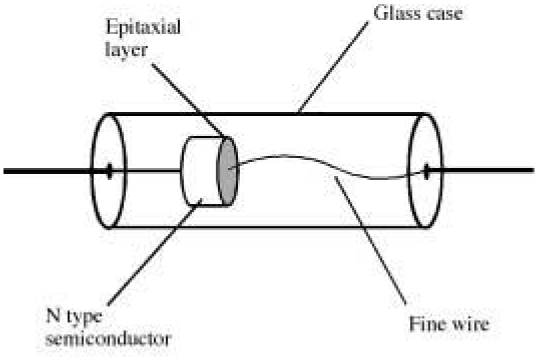

hot carrier diode

hot-pen recorder See THERMAL RECORDER.

hot plate A metal device, usually heated by means of electricity, used for the purpose of conducting certain experiments.

hot resistance The resistance of a component during its operation (i.e., after it has been heated by ambient temperature or internal power dissipation). Compare COLD RESISTANCE.

hot spark A brilliant flash seen when a capacitor discharges through a spark gap in a vacuum.

hot spot **1.** In a circuit or component, an area whose temperature is ordinarily higher than that of the surrounding area. **2.** A local current or voltage maximum on an antenna or transmission line. **3.** In communications operations, a geographic location in which reception is markedly better than in other nearby places or from which the transmitted signal appears to be stronger.

hot-strip ammeter A current meter similar to the hot-wire meter, except that it has a heated metallic strip instead of a heated wire.

hot-stylus recorder See THERMAL RECORDER.

hot-tip writing The use of a heated-tip stylus in graphic recording. Also see THERMAL RECORDER.

hot-wire ammeter See HOT-WIRE METER.

hot-wire anemometer An electrical anemometer whose indication is based on the cooling effect of the wind on a heated filament.

hot-wire flowmeter An instrument for determining the rate of flow of a gas in a pipe or other channel. The circuit is similar to that of the gas detector and hot-wire anemometer.

hot-wire gasmeter See GAS DETECTOR.

hot-wire instrument See HOT-WIRE METER.

hot-wire meter A meter in which current heats a wire, stretching it so that it moves a pointer across a scale over a distance proportional to the magnitude of the current.

hot-wire microphone A microphone in which sound waves vary the temperature of a heated wire and, accordingly, its electrical resistance.

hot-wire relay A time-delay relay in which actuating current heats a wire, causing it to expand, eventually opening or closing the contacts. Also see DELAY RELAY.

hot-wire sensor See HEATED-WIRE SENSOR.

hot-wire transducer See HOT-WIRE MICROPHONE.

hour Abbreviation, h; (sometimes, hr). A unit of time measure equal to 60 minutes or 3600 seconds. Compare MINUTE, **1** and SECOND, **1**. Also see TIME.

housekeeping In digital-computer operations, the part of a program that attends to chores (e.g., setting variables to zero), rather than being involved in making computations for a solution.

howl A discordant sound produced in headphones or a loudspeaker, usually as a result of acoustic or electrical feedback.

howler **1.** An audio-frequency alarm device. **2.** A sound-emitting test device (see GROWLER, **1**, **2**).

howl repeater A form of electric feedback in which a hum or howl occurs because of oscillation. The term is used to describe an oscillating conduction in a wire-communications-system repeater.

hp Abbreviation of HORSEPOWER.

h-p Abbreviation of HIGH PRESSURE.

H pad See H NETWORK.

***h* parameters** Parameters of the four-terminal network equivalent of a transistor. They are *hybrid parameters* (thus, *h*) because of their appearance in mesh and nodal equations. The basic *h* parameters are h_{11}, input resistance with output short-circuited; h_{12}, reverse voltage ratio with input open-circuited; h_{21}, forward current gain with output short-circuited; and h_{22}, output conductance with input open-circuited.

h particle A positive hydrogen ion or proton obtained by bombarding a hydrogen atom with alpha particles or high-velocity positive ions.

HPF Abbreviation of *highest probable frequency.*

hp-hr Abbreviation of HORSEPOWER-HOUR.

HPI Abbreviation of HEIGHT-POSITION INDICATOR.

H plane The plane of the magnetic field of an antenna, or the magnetic component of an electromagnetic field. Compare E PLANE.

H-plane bend In a waveguide, a smooth change in the direction of the axis perpendicular to the direction of polarization.

H-plane tee junction A waveguide tee junction, whose structure changes in the magnetic-field plane. Also see WAVEGUIDE JUNCTION and WAVEGUIDE TEE.

hr Abbreviation of HOUR. (Also, h.)

H scan A radar display in which the target is represented by two close blips, approximating a line, whose shape is proportional to the sine of the target's angle of elevation.

H scope See H SCAN.

HSM Abbreviation of *high-speed memory.*

HTL Abbreviation of HIGH-THRESHOLD LOGIC.

hub The hole in the center of a magnetic tape reel.

hue The quality of having a particular visible color; an attribute that depends on the wavelength(s) of light emitted or reflected. Thus, electromagnetic energy at a wavelength of 700 nanometers (nm) has a red hue; energy at a wavelength of 400 nm has a violet hue. Compare SATURATION, **2**.

hue control Also called *tint control.* In a color television receiver, a control that allows adjustment of the color wavelength, but does not affect the saturation (intensity).

hum **1.** Alternating current having a frequency of 60 Hz. **2.** Residual ripple in the output of a power supply, having a frequency of 60 Hz or 120 Hz, depending on the type of rectifier circuit used. **3.** An electromagnetic field of long wavelength, usually originating from utility lines and having a primary frequency of 60 Hz. **4.** The effects of low-frequency electromagnetic fields or currents, such as moving horizontal bars on a television screen. **5.** An acoustic disturbance of long wavelength (low pitch).

human engineering **1.** Also called *user-friendliness.* The extent to which an electronic or electromechanical device or system is easy to operate. **2.** The branch of engineering devoted to interfacing human beings with the machines and instruments they operate. Both a science and an art, the discipline is concerned with the safest and most efficient design, arrangement, and operation of equipment.

human interface The interface between a sophisticated electronic device and a human operator.

human-made interference Also called *artificial interference.* Electromagnetic interference to radio and television receiving systems or to data terminals, originating from artificial sources (such as radio transmitters, certain electrical appliances, and internal-combustion engines). Of the numerous different forms, some affect only a narrow band of frequencies or set of frequencies, and others affect a wide band of frequencies.

humanoid robot A robot that bears structural resemblance to a human being (e.g., has arms, a head, and perhaps legs). In its most advanced form, such a robot is an ANDROID.

hum-balance potentiometer A potentiometer connected across an alternating-current power supply, with its slider grounded. At a certain setting, hum interference is nulled.

hum bars See HORIZONTAL HUM BARS.

hum bucking The reduction of hum interference by introducing an alternating-current voltage of the same frequency and amplitude as the hum, but opposite in phase.

hum-bucking coil An auxiliary coil used in conjunction with the field and voice coils of an electrodynamic speaker. Reduces hum interference via HUM BUCKING.

hum field The magnetic field surrounding a conductor carrying hum-frequency alternating current.

humidity The amount of moisture in the air. Also see ABSOLUTE HUMIDITY and RELATIVE HUMIDITY.

humidity meter See ELECTRIC HYGROMETER and ELECTRONIC HYGROMETER.

humidity sensor A pickup whose resistance or capacitance varies proportionally with ambient humidity.

hum interference Electrical interference resulting from HUM in any of its various electrical forms (see HUM, **1–4**).

hum loop A ground loop that results in undesired hum in the output of an amplifier.

hummer A nonelectronic audio oscillator similar to the fork oscillator, but using a thick, metal reed, instead of a tuning fork. A carbon microphone button attached to the reed provides the feedback path necessary for sustained oscillation.

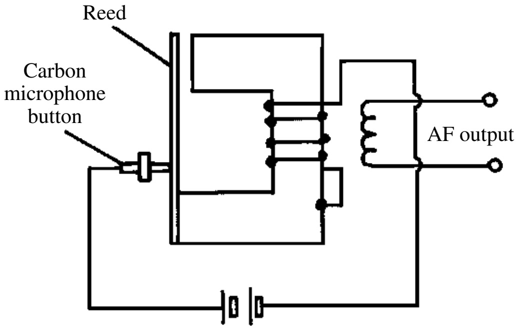

hummer

hum modulation Undesirable modulation of a radio signal or audio amplifier output signal by HUM interference.

hump **1.** Either the positive or the negative half-cycle of a sine wave; that is, either the portion from 0 degrees to 180 degrees, or the portion from 180 degrees to 360 degrees. **2.** A curve whose graph is a multiple of the sine function, and whose domain is restricted to a half-cycle (0 to 180 degrees or 180 to 360 degrees). See, for example, DOUBLE-HUMP RESONANCE CURVE and DOUBLE-HUMP WAVE.

hunting An undesirable condition in which an electrical or electromechanical system oscillates back and forth, relative to some mean mode of operation ("hunts" for the mode), sometimes eventually settling down at the mode. It can occur in a servo system because of overcompensation, lack of hysteresis, or both.

Huygens' principle The observation that an advancing wave is the resultant of secondary waves that arise from points in the medium that have already been passed.

HV Abbreviation of HIGH VOLTAGE.

H vector A vector representing the magnetic field of an electromagnetic wave. Compare E VECTOR.

hW Abbreviation of HECTOWATT.

H-wave mode In a waveguide, a mode of transmission in which the electric lines of flux are at right angles to the direction of the waveguide. Also called *transverse-electric (TE) mode.*

hy Occasional abbreviation of HENRY. The SI abbreviation and symbol, H, is preferred.

hybrid Descriptive of a device that is an offspring of other devices or a product of dissimilar technologies (but using elements of each). See, for example, HYBRID JUNCTION and HYBRID COIL.

hybrid active circuit An active circuit (such as an amplifier, oscillator, or switch) using a combination of two dissimilar active devices (e.g., transistors and vacuum tubes).

hybrid coil A special type of bridging transformer used in wire telephony to prevent self-oscillation in a repeater amplifier that operates in both directions.

hybrid computer A computer system incorporating more than one major computer technology. Examples: analog/digital and digital/neural-network.

hybrid electromagnetic wave Abbreviation, HEM wave. An electromagnetic wave whose electric-field and magnetic-field vectors are both in the direction of propagation.

hybrid IC See HYBRID INTEGRATED CIRCUIT.

hybrid integrated circuit Abbreviation, HIC. An integrated circuit embodying both integrated and microminiature discrete components (i.e., one combining both monolithic and thin-film construction).

hybrid junction **1.** See MAGIC TEE. **2.** A four-terminal device, such as a resistor circuit, special transformer, or waveguide assembly, in which a signal applied to one pair of terminals divides and appears at only the two adjacent terminals.

hybrid microcircuit A microcircuit containing diffused or thin-film elements interconnected with separate chip elements.

hybrid parameters See H PARAMETERS.

hybrid ring A hybrid waveguide junction (see HYBRID JUNCTION, **2**) consisting essentially of a reentrant line with four side arms. Used as an equal power divider.

hybrid-tee See HYBRID JUNCTION, **1**.

hybrid thin-film circuit A microcircuit in which semiconductor devices and discrete components are attached to passive components and conductors that have been electrodeposited on a substrate.

hybrid transformer See HYBRID COIL.

hydroacoustic Pertaining to the sound of fluids, especially water, under pressure.

hydroacoustic transducer A transducer that converts energy from the high-pressure flow of a fluid into acoustic energy.

hydrodynamic pressure The pressure of a fluid in motion. Compare HYDROSTATIC PRESSURE.

hydroelectric Pertaining to the production of electricity by water power, as by a generator turned by water turbines.

hydroelectric machine A device for generating electricity from high-pressure steam escaping from a series of jets.

hydroelectric power See WATER POWER.

hydrogen Symbol, H. A gaseous element. Atomic number, **1**. Atomic weight, 1.00794. Used in making semiconductor materials, it is the lightest and most abundant element in the universe. Compare DEUTERIUM and TRITIUM.

hydrogen atmosphere The nonoxidizing atmosphere in which semiconductor materials are melted and processed, and in which semiconductor crystals are grown. Occasionally, helium is used instead of hydrogen.

hydrogen atom A single atom of the element hydrogen, consisting of one electron and one proton.

hydrogen-ion concentration See pH.

hydrogen lamp A glow-discharge lamp that produces light by means of the ionization of rarefied hydrogen gas. Visible light is emitted at discrete wavelengths.

hydrokinetic Pertaining to fluids in motion or the forces behind such motion.

hydrolysis The process whereby chemical substances become ionized in water solution, producing electrolytes.

hydromagnetics See MAGNETOHYDRODYNAMICS.

hydromagnetic wave In a fluid, a wave in which the energy is propagated via magnetic and dynamic modes.

hydrometer An instrument for measuring the specific gravity of liquids.

hydrophone An underwater sound-to-electricity transducer (microphone).

hydropower See WATER POWER.

hydrostatic pressure The pressure of a fluid at rest. Compare HYDRODYNAMIC PRESSURE.

hygristor A resistor, the resistance of which depends on humidity. A common application is in the measurement of relative atmospheric humidity.

hygrogram A graphic rendition of atmospheric humidity versus time, made by a HYGROGRAPH.

hygrograph A device that produces a graphic recording of atmospheric humidity as a function of time. Consists essentially of a HYGROMETER connected to a permanent data-storage device, such as a computer or pen recorder.

hygrometer An instrument for measuring humidity. Also see ELECTRIC HYGROMETER, ELECTRONIC HYGROMETER, and HAIR HYGROMETER.

hygroscopic material A material that absorbs moisture from the air, but not enough to get wet (e.g., lime and silk). Compare DELIQUESCENT MATERIAL.

hygrostat A humidity-sensitive relay or switching circuit.

hygrothermograph A graphic recorder indicating humidity and temperature on the same chart, both as functions of time.

hyperacoustical zone In the upper atmosphere, a region in which the distance between air molecules is comparable to the wavelengths of audible sound. As the altitude increases within this zone, the upper-frequency limit of effective sound propagation decreases. At altitudes above this zone, no audible sound can be propagated.

hyperbola A conic-section curve satisfying the equation:

$$(x - x_0)^2/a^2 - (y - y_0)^2/b^2 = 1$$

where x and y are the independent and dependent variables, x_0 and y_0 are the coordinates of the center, a is half the length of the major axis, and b is half the length of the minor axis.

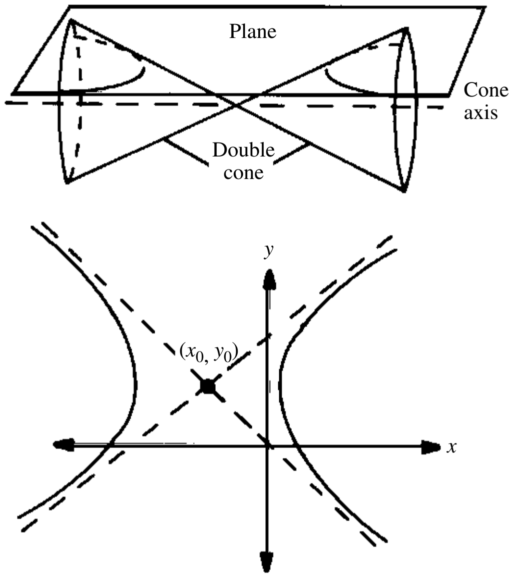

hyperbola

hyperbolic angle An angle subtended by a sector of a hyperbola in a manner analogous to that in which a circular angle is subtended by an arc of a circle.

hyperbolic-cosine horn See CATENOIDAL HORN.

hyperbolic error **1.** In an interferometer, a miscalculation in the direction of arrival of a signal. The signal from one antenna in the system can be assumed to be in phase with the signal from another antenna, when actually the two components differ by an integral number of whole wavelengths. **2.** The angular error, in degrees, minutes, or seconds of arc, resulting from a miscalculation of phase in an interferometer.

hyperbolic face contour See HYPERBOLIC GRIND.

hyperbolic functions The nonperiodic functions of a hyperbolic angle. Hyperbolic functions are related to the hyperbola in the same manner that common trigonometric functions are related to the circle. Hyperbolic functions are useful in calculations involving traveling waves on transmission lines, attenuator design, and wave-filter design.

hyperbolic grind The shape (approximately hyperbolic) to which the face of a magnetic recording head is ground. It provides optimum contact with the tape and ensures good high-frequency response.

hyperbolic horn A horn antenna whose cross-sectional area is a hyperbolic function of the distance along the axis.

hyperbolic logarithm See NAPIERIAN LOGARITHM.

hyperbolic navigation A radionavigation system in which the operator of an aircraft or boat determines position by comparison of two received signals. The two transmitters radiate signals from known positions and with known timing characteristics. The time delay from each transmitter is determined, resulting in two hyperbolic curves on a map. The point of intersection of the curves is the location of the aircraft or ship.

hyperbolic radian A unit of measure derived from a hyperbolic angle. A hyperbolic radian is the hyperbolic angle that encloses an area of 0.5 when the distance along the x-axis to the hyperbola is unity. Also see HYPERBOLIC ANGLE. Compare CIRCULAR RADIAN.

hyperbolic trigonometry The branch of mathematics dealing with the theory and application of hyperbolic angles and their functions.

hypercardioid microphone A unidirectional microphone with exceptional sensitivity in front, and minimal responsiveness from the sides and rear.

hypercardioid pattern A directional CARDIOID PATTERN with accentuated responsiveness or emission in the favored direction (front), and greatly suppressed responsiveness or emission in other directions (sides and rear).

hyperfocal distance The shortest distance to which a lens can be focused without degrading definition at infinity.

hyperfrequency waves See MICROWAVES.

hypernik See HIPERNICK.

hyperon Any one of various particles having a mass greater than that of a neutron or proton.

hyperpolarization The production of an increased voltage across a biological membrane.

hypersonic Pertaining to speeds of at least five times that of sound. In air at normal atmospheric pressure, such speeds are more than approximately 1700 meters, or 5600 feet, per second.

hypersonic speed Any speed greater than five times the speed of sound in a given medium.

hypersyn motor A high-efficiency, high-power-factor synchronous motor combining the advantages of the *direct-current-excited synchronous motor* (stiffness), the *hysteresis motor* (synchronizing torque), and the *induction motor* (high starting torque).

hypervelocity Velocity in excess of 3 kilometers per second.

hypotenuse The side of a right triangle opposite to the right angle.

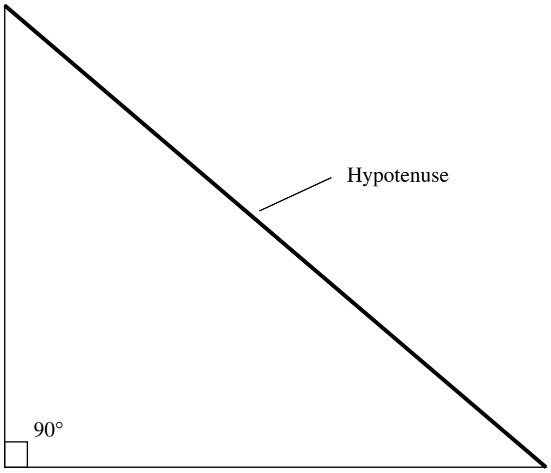

Hypotenuse

hypothesis An idea, concept, or system that seems true or workable, but must be subjected to logical analysis and/or practical testing to prove its validity.

hypsometer An altimeter in which a thermistor (connected to a battery and current meter) is immersed in a boiling liquid. Because the liquid's boiling point is proportional to altitude, it affects the resistance of the thermistor and, hence, the deflection of the meter.

hysteresigram The hysteresis-curve record produced by a HYSTERESIGRAPH.

hysteresigraph A graphic recorder that displays or records the hysteresis curve for a material. Also see HYSTERESISCOPE.

hysteresimeter See HYSTERESIS METER.

hysteresis **1.** The tendency of a magnetic material to saturate and retain some of its magnetism after the alternating magnetic field to which it is subjected reverses polarity, thus causing magnetization to lag behind the magnetizing force. **2.** A similar electrostatic action in a ferroelectric dielectric material. **3.** In a servo system, the condition in which a variable quantity lags the effect or stimulus that causes it to change. The plot for this is a double-line HYSTERESIS CURVE.

hysteresis brake A brake whose retarding action comes from hysteresis in a permanent-magnet motor.

hysteresis clutch A magnetic clutch whose output torque (for synchronous drive or continuous slip) comes from hysteresis in a permanent-magnet motor.

hysteresis coefficient In a sample of iron whose volume is one cubic centimeter, the energy in ergs dissipated during one cycle of magnetization. Also called *coefficient of hysteresis.*

hysteresiscope An oscilloscope that is specially designed to display the hysteresis curve of a material. Compare HYSTERESIGRAPH.

hysteresis curve A response curve depicting hysteresis in a magnetic material, a dielectric, or a servo system. A graph of the extent to which a variable quantity lags the effect or stimulus that causes it to change (e.g., a curve showing the effect of response delay in a thermostatically controlled heating/cooling system).

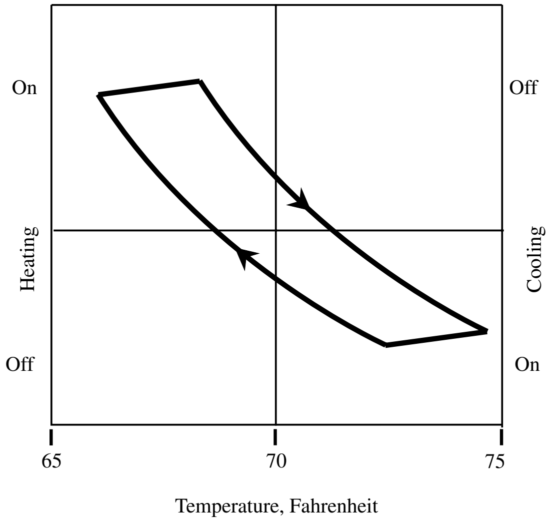

hysteresis curve

hysteresis cycle A complete hysteresis curve.

hysteresis distortion Signal distortion in iron-core components, such as coupling transformers, resulting from hysteresis in the iron.

hysteresis error In a meter, a difference in indications for increasing and decreasing current, an effect caused by hysteresis in iron meter parts.

hysteresis heater An induction heater in which heating results from hysteresis loss in the load.

hysteresis loop See HYSTERESIS CURVE.

hysteresis loss Power loss caused by hysteresis in a magnetic material exposed to an alternating magnetic field, or in a dielectric material exposed to an alternating electric field. It is characterized by the generation of heat.

hysteresis meter An instrument that determines the hysteresis loss in a ferromagnetic material in terms of the torque produced when the material is rotated in a magnetic field, or vice versa.

hysteresis motor A synchronous motor that does not require direct-current excitation, and does not have salient poles. It is started by means of hysteresis losses that the rotating magnetic field causes in the secondary.

hysteretic constant For a ferromagnetic material, hysteresis loss in ergs per cubic centimeter of material per cycle of magnetization.

hysteretic loss See HYSTERESIS LOSS.

hystoroscope A device used to determine the magnetic characteristics of a material.

Hz Abbreviation of HERTZ.

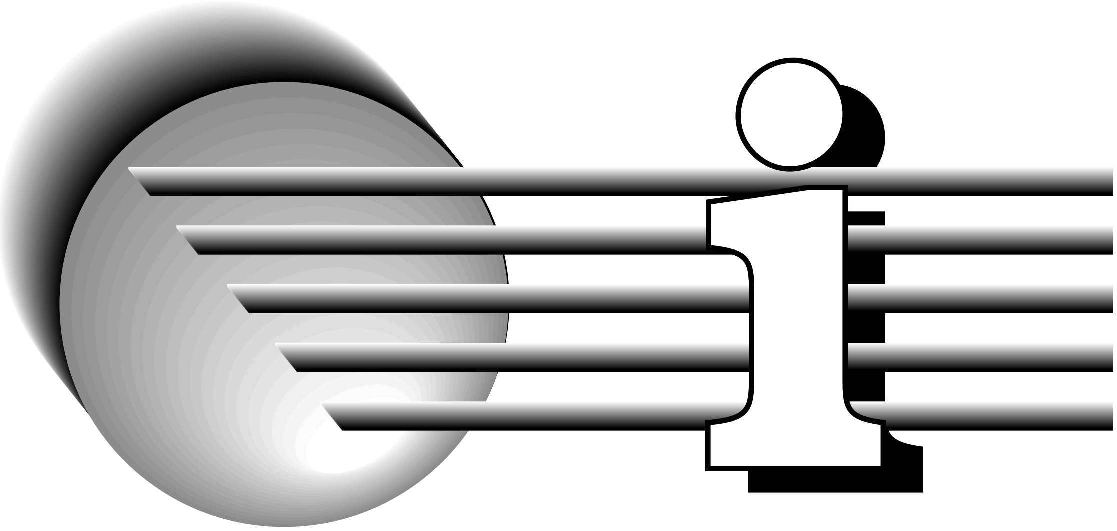

I **1.** Symbol for CURRENT. **2.** Symbol for INTRINSIC SEMICONDUCTOR. **3.** Symbol for LUMINOUS INTENSITY. **4.** Symbol for IODINE.

i **1.** Symbol for the unit imaginary number, the square root of -1. (Also, *j*.) **2.** Subscript for INSTANTANEOUS VALUE. **3.** Symbol for INTRINSIC SEMICONDUCTOR. **4.** Symbol for ANGLE OF INCIDENCE. **5.** Symbol for INSTANTANEOUS CURRENT. (Also, I_i.) **6.** Symbol for a unit vector parallel to the x-axis. **7.** Symbol for INCIDENT RAY.

I_{ac} Symbol for the ac component of a COMPOSITE CURRENT.

I_{AF} Symbol for AUDIO-FREQUENCY CURRENT.

IAGC Abbreviation of INSTANTANEOUS AUTOMATIC GAIN CONTROL.

IAVC Abbreviation of INSTANTANEOUS AUTOMATIC VOLUME CONTROL.

I_B **1.** Symbol for *plate power-supply current.* **2.** Occasional symbol for *plate current* (usually, I_P).

IBM Abbreviation for *International Business Machines* Corporation.

IC **1.** Abbreviation of INTEGRATED CIRCUIT. **2.** Abbreviation of INTERNAL CONNECTION.

I_C **1.** Symbol for transistor COLLECTOR CURRENT. **2.** Occasional symbol for *grid current* (usually, I_G).

ICAD Abbreviation of *integrated control and display.*

ICAS Abbreviation of INTERMITTENT COMMERCIAL AND AMATEUR SERVICE.

I_{CBO} Symbol for the static reverse collector (leakage) current in a common-base connected transistor with an open-circuited emitter.

ICBS Abbreviation of *interconnected business system.*

I_{CEO} Symbol for the static reverse collector (leakage) current in a common-emitter-connected transistor with an open-circuited base.

ice loading **1.** In an antenna, power-line system, or other structure, the additional stress caused by accumulation of ice. **2.** The weight or thickness of ice a structure can safely withstand.

ice-removal circuit A high-voltage, low-frequency power supply used to heat certain antennas to melt ice that accumulates on them.

ICET Abbreviation of *Institute for the Certification of Engineering Technicians* (National Society of Professional Engineers).

ICME Abbreviation of *International Conference on Medical Electronics.*

I_{co} Symbol for the collector cutoff current (static leakage current) of a bipolar transistor (see CUTOFF CURRENT).

icon In a graphical computer interface, a symbol that aids the user in recognizing a selection that can be made.

iconoscope A camera tube in which an electron beam scans a photomosaic on which the image is focused. The light-sensitive droplets of the mosaic form tiny capacitors with the insulated, metallic backplate of the mosaic, each capacitor becoming charged by the light of the picture. As the electron beam scans the mosaic, each capacitor discharges as the beam strikes it, delivering an output pulse proportional to the light intensity at that spot in the picture.

IC tester An instrument for checking the operation of an integrated circuit.

ICW Abbreviation of INTERRUPTED CONTINUOUS WAVE.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

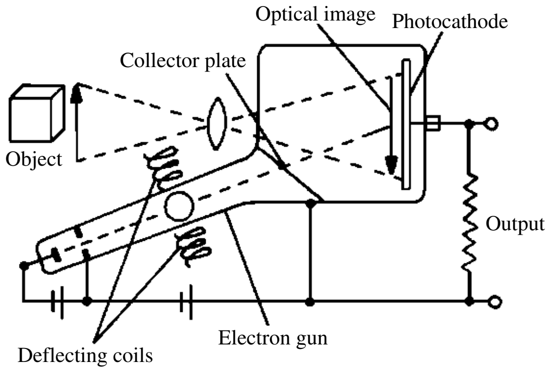

iconoscope

ID Abbreviation of INSIDE DIAMETER.

IDA Abbreviation of *integrodifferential analyzer.*

I_{dc} Symbol for the dc component of a COMPOSITE CURRENT.

ideal Pertaining to a circuit, device, material, or manner of operation that is assumed to conform to the theoretical best-case example; it is not usually realized in real life. Thus, an ideal reactance has no inherent resistance.

ideal capacitor A capacitor having zero dielectric loss and a constant value of capacitance at all alternating-current frequencies.

ideal component A theoretical component that is assumed to be 100-percent efficient and to have no extraneous properties. Compare PRACTICAL COMPONENT.

ideal crystal A piezoelectric crystal that acts as a theoretically perfect tuned circuit, that is, an ideal capacitor and inductor.

ideal inductor An inductor having zero loss and a constant value of inductance at all alternating-current frequencies.

I demodulator In a color-television receiver circuit, a demodulator that receives the chrominance and 3.58-MHz oscillator signals and delivers a video output corresponding to color in the picture.

identical vectors Equal vectors that have the same initial point or point of application.

identification **1.** In radar operations, the (often automatic) determination of the target's identity. See, for example, IDENTIFICATION, FRIEND OR FOE. **2.** In digital computer operations, a symbol or set of symbols within a label identifying a unit of data or its location.

identification beacon **1.** A beacon used for the determination of a particular geographic location. **2.** An automatically transmitted station-identification signal or code, usually superimposed on the regular transmission in the form of a subcarrier or subaudible signal.

identification division The division (one of four) in a COBOL program that describes and identifies the program being compiled.

identification, friend or foe Abbreviation, IFF. A technique in which a radar station transmits an interrogating signal and the station questioned replies automatically with a suitable pulse or other signal if it is aboard a friendly aircraft or vessel. If it is aboard an enemy vehicle, the station gives no reply or sends an unsatisfactory one.

identifier A data file identification label in an input/output device, or a label that identifies a specific storage location.

identity element A logic element that, upon receipt of two input signals, provides an output that is logic 1 (high) only if the input signals are both logic 1 (high) or logic 0 (low).

idiochromatic Possessing the photoelectric properties of a true crystal.

I display See I SCAN.

idle channel noise A measurement of wideband noise. The total energy at an output of an integrated circuit or other communications stage with that output grounded (idle).

idle character A digital character that conveys no information, but helps maintain synchronization between the transmitter and receiver. Sometimes called a *blank.*

idler wheel In a phonograph or magnetic tape recorder, an auxiliary, rubber-coated wheel that transfers rotary motion from the motor pulley to the platter or the rim of the capstan flywheel.

idle time The period during which data processing equipment, although operable, is not in use.

idling Standby equipment operation, as when vacuum-tube filaments in a radio-frequency power amplifier are kept hot—even when the amplifier is not actually being used to transmit a signal.

idling current The current flowing in a device during a standby period, as opposed to operating current. Also called STANDBY CURRENT.

idling frequency In a parametric amplifier, the difference between the signal frequency and pump frequency.

idling power See STANDBY POWER.

idling voltage The voltage required by or measured in a device that is in standby mode.

$I_{D(off)}$ Symbol for drain cutoff current in a field-effect transistor.

IDOT Abbreviation of *instrumentation online transcriber.*

IDP **1.** Abbreviation of INDUSTRIAL DATA PROCESSING. **2.** Abbreviation of INTEGRATED DATA PROCESSING. **3.** Abbreviation of INTERMODULATION-DISTORTION PERCENTAGE.

I_{DSS} Symbol for DRAIN CURRENT AT ZERO GATE VOLTAGE in a field-effect transistor.

I_E Symbol for EMITTER CURRENT.

IEC Abbreviation of *integrated electronic component.*

IEE Abbreviation of *Institution of Electrical Engineers* (British).

IEEE Abbreviation of *Institute of Electrical and Electronics Engineers.*

IES Abbreviation of *Illuminating Engineering Society.*

IF Abbreviation of INTERMEDIATE FREQUENCY.

IF amplifier See INTERMEDIATE-FREQUENCY AMPLIFIER.

I_{FB} In an integrated circuit, current limit feedback or feedback current.

IF channel See INTERMEDIATE-FREQUENCY CHANNEL.

IF converter The converter (first detector-oscillator) section of a superheterodyne radio receiver.

IFF Abbreviation of IDENTIFICATION, FRIEND OR FOE.

IF gain **1.** The amplification provided by the intermediate-frequency (IF) channel of a superheterodyne receiver, usually specified in decibels. **2.** A control that allows adjustment of the amplification in the IF channel of a superheterodyne radio receiver.

IF interference See INTERMEDIATE-FREQUENCY INTERFERENCE.

IFIPS Abbreviation of *International Federation of Information Processing Societies.*

IF selectivity See INTERMEDIATE-FREQUENCY SELECTIVITY.

IF strip A (sometimes removable) circuit section containing a complete intermediate-frequency channel.

IF/THEN/ELSE In computer systems and artificial intelligence, a two-way choice that is made during the execution of a program. For example, in finding the absolute value of a number, the logic proceeds as follows: "IF $x < 0$, THEN multiply x by -1; otherwise (ELSE) leave it alone."

IF transformer See INTERMEDIATE-FREQUENCY TRANSFORMER.

I_G Symbol for GATE CURRENT.

I_G Symbol for GRID CURRENT.

IGFET Abbreviation of INSULATED-GATE FIELD-EFFECT TRANSISTOR.

ignition coil A small open-core transformer having a high step-up turns ratio for converting 6- or 12-volt battery potential to the high voltage needed in an automotive ignition system.

ignition interference Electrical noise generated by the ignition system of an internal combustion engine.

ignition potential **1.** The minimum voltage needed to cause electrical discharge in an internal combustion engine. **2.** See BREAKDOWN VOLTAGE, **1**, **2**.

ignition reserve The extra voltage provided by the starter, as compared with the voltage actually needed for ignition of an internal-combustion engine.

ignition system An electrical or electronic system that supplies the high voltage in an automotive engine. See, for example, CAPACITOR-DISCHARGE IGNITION SYSTEM.

ignition voltage See IGNITION POTENTIAL.

ignitron A form of rectifier tube that contains ionized mercury. It is used with some extremely high-voltage power supplies.

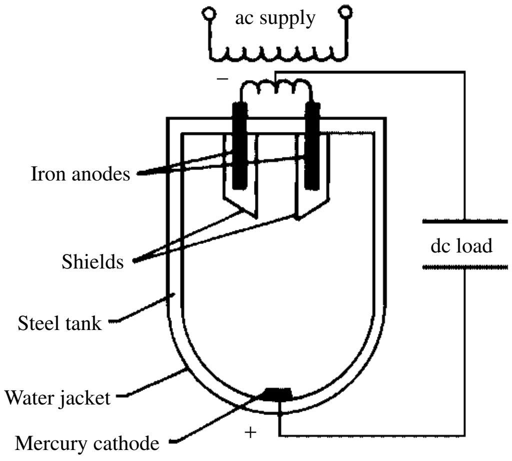

ignitron

ignore character **1.** A character used as a signal to inhibit an action. **2.** A digital character that is ignored by the system.

I_{GSS} Symbol for GATE REVERSE CURRENT in a field-effect transistor.

IGY Abbreviation for INTERNATIONAL GEOPHYSICAL YEAR.

I_H **1.** Symbol for HOLD CURRENT. **2.** Symbol for HOLDING CURRENT.

IHF Abbreviation of INHIBIT FLIP-FLOP.

IHFM Abbreviation for *Institute of High Fidelity Manufacturers.*

ihp Abbreviation of INDICATED HORSEPOWER.

ihp-hr Abbreviation of INDICATED HORSEPOWER-HOURS.

I_i **1.** Symbol for INPUT CURRENT. **2.** Symbol for INSTANTANEOUS CURRENT (also, i).

I_K Symbol for CATHODE CURRENT.

I_L Symbol for current in an inductor.

I^2L Abbreviation of INTEGRATED INJECTION LOGIC. (Also IIL.)

illegal character **1.** A character or bit group that is, according to some standard, invalid. **2.** A bit group that represents a symbol in a character set.

illuminance The amount of luminous flux received per unit surface area, measured in lux (lumens per square meter).

illuminant-C In color television, the reference white that closely resembles average daylight.

illuminated pushbutton See LIGHTED PUSHBUTTON.

illuminated switch See LIGHTED SWITCH.

illumination **1.** Visible light. **2.** The condition of being irradiated by visible light. **3.** The intensity of visible light irradiating an object or region.

illumination control A photoelectric circuit that automatically regulates electric lights, according to the amount of daylight.

illuminometer A device for measuring the intensity of visible light, or, less often, infrared or ultraviolet. In its simplest form, the meter consists of a photovoltaic cell, a potentiometer for adjusting the sensitivity, and a microammeter. More sophisticated devices include direct-current amplification.

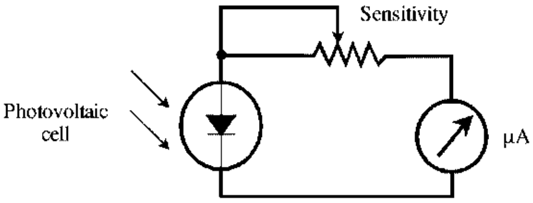

illuminometer

ILS Abbreviation of INSTRUMENT LANDING SYSTEM.

IM Abbreviation of INTERMODULATION.

I_m **1.** Abbreviation of *maximum current.* (Also, I_{max}). **2.** Abbreviation of *meter current.*

image **1.** In a superheterodyne circuit, a spurious response whose frequency differs from that of the desired signal by twice the intermediate frequency. **2.** A picture on the screen of a television receiver or computer display. **3.** A pattern on the screen of an oscilloscope tube. **4.** A picture on the mosaic of a television camera tube. **5.** A duplicate of a computer storage area that is in another part of storage or on another medium.

image admittance The reciprocal of IMAGE IMPEDANCE.

image antenna An imaginary "mirror" antenna below the effective radio-frequency (RF) ground surface, at a depth equal to the height of the true radiating antenna above effective RF ground.

image attenuation constant The real-number part of the IMAGE TRANSFER CONSTANT. Also see IMAGE PHASE CONSTANT.

image compression Any of various methods by which the number of bytes in a digital image can be reduced, without significant degradation of image quality. The two most common methods are *JPEG (Joint Photographic Experts Group) image compression* and *fractal image compression.* Other schemes include *MPEG (Moving Picture Experts Group) image compression* and *Indeo*, developed by Intel Corporation. Fractal image compression can provide a compression factor of as much as 100:1 without significant degradation of image quality.

image converter **1.** A device that changes an invisible image into a visible image. Examples include the *snooperscope*, an infrared-to-visible converter, and photographic apparatus for infrared, ultraviolet, and X-ray wavelengths. **2.** A tube that operates as an image converter.

image dissector See DISSECTOR TUBE.

image effect The effect of reflection of electromagnetic waves from the ground. An IMAGE ANTENNA appears to radiate from a point beneath the effective RF ground plane. The depth of the image antenna below the effective RF ground plane is equal to the height of the actual antenna above the effective RF ground plane.

image frequency The frequency of the image response in a superheterodyne radio receiver. See IMAGE, **1**.

image impedance The property of a network in which the load impedance is "seen," looking into the output terminals with the generator connected to the opposite end, and the generator impedance is "seen," looking into the input terminals with a load connected to the opposite end.

image intensification An increase of the brightness of the display on a cathode-ray tube.

image intensifier A device that increases the brightness of the spot produced when a beam of electrons or X rays hits a fluorescent screen.

image interference A type of interference that can occur in superheterodyne circuits. It occurs when there is a sufficiently strong signal on the IMAGE FREQUENCY.

image orthicon See ORTHICON.

image phase constant The imaginary-number part of the IMAGE TRANSFER CONSTANT. Also see IMAGE ATTENUATION CONSTANT.

image potential The potential energy of a charged particle at a given distance from a metal surface. The metal surface acts in a way similar to a mirror; a reflected image of the object, having equal charge but opposite polarity, is formed on the other side of the surface.

image ratio See SIGNAL-TO-IMAGE RATIO.

image rejection In a superheterodyne radio receiver, the suppression or elimination of IMAGE INTERFERENCE by means of a selective circuit, such as a radio-frequency preamplifier.

image response In a superheterodyne receiver, an undesired response to signals removed from the desired frequency by twice the intermediate frequency.

image transfer constant A number depicting the transfer of power by an impedance network. It has the same value, regardless of the direction of transmission through the network. Also see IMAGE ATTENUATION CONSTANT and IMAGE PHASE CONSTANT.

imaginary axis In a vector diagram of complex impedance, the axis of the imaginary-number component (jX).

imaginary number A real-number multiple of the positive square root of -1. Engineers depict the positive square root of -1 by the lowercase letter *j*; mathematicians use the lowercase letter *i*. Engineers write imaginary numbers in the form jX, and mathematicians write Xi, where X is a real number.

imaginary-number component The imaginary-number part of a COMPLEX NUMBER.

I_{max} Abbreviation of *maximum current.*

IM distortion meter INTERMODULATION METER.

IM distortion percentage See INTERMODULATION-DISTORTION PERCENTAGE.

I_{min} Abbreviation of *minimum current.*

imitation The transmission of false signals for purposes of deception. For example, during wartime, the signals from an enemy station might be recorded and retransmitted.

immediate access **1.** The ability of a computer to store and retrieve data in a minimal amount of time. **2.** Computer storage that can be accessed in a minimal amount of time.

immediate address An instruction address that is used as data by that instruction.

IM meter See INTERMODULATION METER.

immitance Impedance or admittance; a contraction of IMPEDANCE and ADMITTANCE. Example: a *negative-immitance circuit.*

immortal knowledge Knowledge that can be kept indefinitely in the form of detailed computer data files. Some scientists and sociologists believe that this will eventually have a profound effect on the way that future generations view history.

impact excitation See SHOCK EXCITATION.

impact strength **1.** The ability of a component or material to withstand mechanical shock loading. **2.** The work required to fracture the material under shock loading.

IMPATT diode Acronym for *impact avalanche transit time* diode. A microwave semiconductor (silicon or gallium arsenide) diode exhibiting negative resistance resulting from the combined effects of charge-carrier transit time and impact avalanche breakdown. It is used as an oscillator or amplifier.

IMPATT oscillator A microwave oscillator that uses an IMPATT DIODE.

impedance Symbol, Z. Unit, ohm. The total opposition offered by a circuit or device to the flow of alternating current. It is the vector sum of RESISTANCE and REACTANCE. This is a COMPLEX NUMBER whose real-number component is resistance R, and whose imaginary-number component is reactance jX; mathematically, $Z = R + jX$.

impedance angle The angle between the resistance and impedance vectors in an IMPEDANCE TRIANGLE.

impedance arm The network branch that contains one or more impedances, as opposed to an arm that contains only resistance or (predominantly, reactance). Also called *impedance leg.*

impedance branch See IMPEDANCE ARM.

impedance bridge **1.** An alternating-current (ac) bridge (commonly operated at 1 kHz) used to measure resistance, inductance, capacitance, and resistive components associated with inductors and capacitors, from which impedance can be calculated. **2.** Sometimes, an ac half-bridge circuit in which an unknown impedance is compared with a known resistance. **3.** A radio-frequency bridge circuit whose balancing element reads impedance directly in ohms.

impedance bump A discontinuity in the characteristic impedance of a radio-frequency transmission line. It is often caused by the use of improper splicing techniques.

impedance coil See CHOKE COIL.

impedance converter See IMPEDANCE TRANSFORMER.

impedance-coupled amplifier An amplifier using capacitor/coil combinations for interstage and output load coupling.

impedance drop In an alternating-current circuit, the complex sum of the resistance drop and reactance drop.

impedance ground A ground connection in which the impedance at the operating frequency is determined by a network of resistors, capacitors, and/or inductors.

impedance leg See IMPEDANCE ARM.

impedance magnetometer A device that is used for measuring small local variations in the intensity of the earth's magnetic field. A small change in the intensity of the magnetic field will cause a change in impedance of a nickel-iron wire having high permeability.

impedance match The condition (for maximum power transfer) when the transmitting impedance equals the receiving impedance, or when a suitable transformer is inserted between different impedances for matching purposes. Also see IMPEDANCE MATCHING.

impedance matching **1.** The adjustment or modification of two impedances so that they are identical, that is, the two resistive components are equal, and the two reactive components are equal. **2.** The insertion of a suitable transformer or network between circuits having different impedances, for the purpose of optimizing power transfer.

impedance-matching network A network of discrete components, often adjustable, that is used to match a circuit having a certain impedance to a circuit having a different impedance. An example is the inductive-capacitive (LC) coupler, also called a *transmatch*, commonly used to match a radio transmitter to an antenna system.

impedance-matching transformer See IMPEDANCE TRANSFORMER.

impedance meter See Z METER.

impedance plethysmograph An electronic device used to measure changes in the chemical content of body cells.

impedance poles See POLES OF IMPEDANCE.

impedance ratio The quotient of two impedances that are related in some situation, such as impedance match or impedance mismatch. The impedance ratio of a transformer is equal to the square of the turns ratio.

impedance transformer **1.** A transformer for converting an impedance to a different value. The turns ratio is equal to the square root of the impedance ratio. **2.** An emitter follower or source follower circuit, used primarily to match a high impedance to a lower impedance. **3.** A short-circuited transmission-line section used to match or convert impedances at radio frequencies.

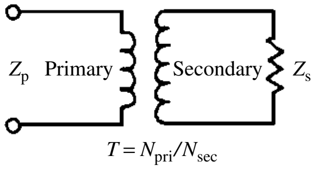

impedance transformer

impedance triangle A triangular vector diagram in which the impedance vector is the hypotenuse, and the reactance and resistance vectors are the perpendicular sides.

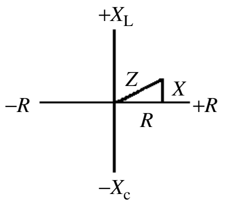

impedance triangle

impedance vector In a vector diagram, the resultant (vector) representing the combined reactance and resistance vectors. Also see IMPEDANCE TRIANGLE.

impedance zeros See ZEROS OF IMPEDANCE.

imperative macroinstruction A MACROINSTRUCTION used to create object (machine language) program instructions.

imperative statement A source language program instruction that is converted into object program (machine language) instructions.

imperfection A fault in the lattice of a crystal. Also see CRYSTAL LATTICE.

implementation **1.** Putting a system into operation and evaluating its performance. **2.** Doing all that is necessary to install a device or system.

implosion The inward collapse of an evacuated chamber, such as the envelope of a cathode-ray tube. It is the opposite of explosion, an outward bursting.

impregnant A substance for the IMPREGNATION of electronic components. Examples: waxes, oils, liquid plastics, and varnish.

impregnation The process of filling the spaces within a material or structure with an insulating compound. Various electronic components, such as capacitors, inductors, transformers, transistors, and diodes, are impregnated for protection and stability.

impressed voltage The voltage applied or presented to a circuit or device.

impulse **1.** A signal of extremely short duration (e.g., a noise burst). **2.** A brief, unidirectional surge in voltage (usually) or current.

impulse excitation Driving a tank circuit with a single pulse, then allowing it to oscillate at its own frequency until another driving pulse arrives.

impulse frequency In a digital telephone, the number of pulse periods per second, corresponding to a dialed digit, generated by the pulse springs.

impulse generator See MARX GENERATOR.

impulse noise Electrical noise from short-duration pulses, such as those produced by an automotive ignition system.

impulse ratio The ratio V_p/V_s, where V_p is the breakdown (or flashover or sparkover) voltage caused by an impulse, and V_s is the corresponding voltage at the crest of the power-frequency cycle.

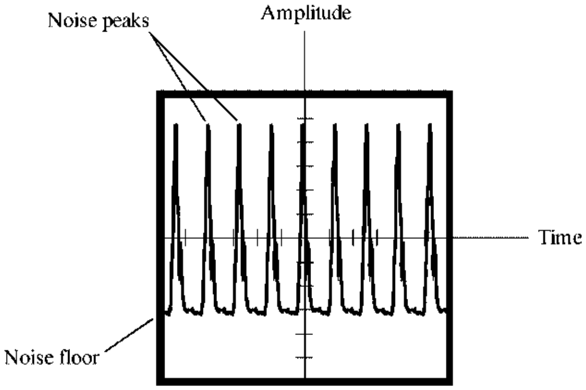

impulse noise

impulse relay A relay that is able to close or open completely when driven by a short pulse.

impulse speed The switching rate of a telephone dialing device as it transmits pulses.

impulse timer A synchronous-motor-driven timer whose cams can control many circuits; it can advance by a number of specified increments, as controlled by an integral stepping mechanism.

impulse transmission A method of transmission in which defined impulses are used to denote changes in signal content or format.

impurity A substance added to an intrinsic semiconductor to alter its electrical properties.

impurity atom In a processed semiconductor material, an atom of an IMPURITY material that produces either n- or p-type properties to the intrinsic semiconductor.

impurity density In the manufacturing process of a semiconductor material, the amount of impurity added to the original semiconductor.

impurity ion In a crystal, an ion in a space between atoms, or one taking the place of an atom.

impurity level **1.** The energy existing in a semiconductor material as a result of doping (the addition of an impurity). **2.** See IMPURITY DENSITY.

impurity material See IMPURITY.

IN Abbreviation for INTELLIGENT NETWORK.

In Symbol for INDIUM.

I_n Symbol for the *n*th value of current in a series of values.

in **1.** Abbreviation of INPUT. **2.** Abbreviation of INCH.

inaccuracy **1.** The state or condition of instrument error. **2.** The difference between the actual value of a parameter and the value indicated by an instrument. **3.** The percentage of instrument error.

inactive leg Within a transducer, an electrical component whose characteristic remains unchanged when the stimulus (quantity being transduced) is applied; specifically, a Wheatstone bridge element in a transducer.

inactive lines In a conventional television picture, blanked lines that do not contribute to the visible part of the image. Approximately half of these lines are at the top of the screen; the other half are at the bottom.

inactive time The period during which a radioactivity counter is insensitive to ionizing agents.

incandescence The state of glowing from intense heat, as when a metal becomes white hot from an electric current flowing through it.

incandescent lamp A filament-type lamp. The filament becomes so hot when an electric current passes through it that it glows brilliantly.

inch Abbreviation, in. A unit of linear measure in the English system; 1 in. = 2.54 centimeters = 0.0254 meter.

inching See JOGGING.

inch-pound Abbreviation, in-lb. A unit of work equal to a force of 1 pound exerted over a distance of 1 inch. Compare FOOT-POUND.

incident **1.** A failure in a computer system requiring the intervention of an operator in removing or revising the job involved. **2.** Pertaining to a ray or rays of energy striking a surface or boundary, or encountering a device (such as an antenna or photovoltaic cell).

incidental AM Undesired amplitude modulation in a frequency-modulated signal. Compare INCIDENTAL FM.

incident field intensity The field strength of an electromagnetic field as it arrives at a receiving antenna.

incidental FM Undesired frequency modulation in an amplitude-modulated signal. Compare INCIDENTAL AM.

incidental time Computer time devoted to other than program runs or program development.

incident light The light that strikes or enters an altering device or medium. See INCIDENT RAY.

incident power In a transmission line, the power that reaches the end of the line. Compare REFLECTED POWER.

incident ray The ray that strikes the surface of a reflecting, refracting, or absorbing body. Compare REFLECTED RAY and REFRACTED RAY.

incident wave **1.** A wave propagated to the ionosphere. Compare REFLECTED WAVE and REFRACTED WAVE. Also see IONOSPHERE and IONOSPHERIC PROPAGATION. **2.** A wave that encounters a change (in density, for example) in a propagation medium or the transition point between media.

in-circuit tester An instrument that permits the checking of components (especially transistors) without removing them from the circuit in which they are wired.

inclination **1.** At a given geographic location, the angle between the earth's magnetic lines of flux and the horizon. This angle ranges from 0 degrees at the geomagnetic equator to 90 degrees at the geomagnetic poles. Compare DECLINATION, 1. **2.** The angle between the plane of a satellite orbit and the plane containing the equator of the earth, planet, or star around which the satellite revolves. Generally measured as the angle at which the satellite crosses the equator while passing from the southern hemisphere to the northern hemisphere. This angle is between 0 and 90 degrees for satellites in non-retrograde orbits, and between 90 and 180 degrees for satellites in retrograde orbits.

inclinometer An instrument for measuring INCLINATION. One form consists of a magnetic needle mounted so that it can swing inside a vertically mounted circular scale.

inclusive-OR Also called *OR operation.* A logical operation between two operands, the result of which depends on rules for combining bits in each position within the operands: an output of 1 results if one or both of the bits have a value of 1; zero only if both are zero. Compare EXCLUSIVE-OR.

incoherent light Electromagnetic radiation in the visible spectrum in which the waves are not aligned (not in phase) and can be of more than one wavelength.

incoherent radiation Electromagnetic radiation in which the waves are not aligned (not in phase) and can be of more than one frequency or wavelength.

incoming inspection The examination of equipment and materials as they enter a factory or laboratory, for the purpose of identifying damaged or faulty units.

incoming line A line that enters a device, facility, or stage. Compare OUTGOING LINE.

Incompleteness Theorem A mathematical theorem of significance in artificial intelligence. For any consistent set of axioms (postulates), there are always more true statements than provable statements.

incomplete program Also called *incomplete routine.* A computer program of generalized steps that must be augmented with specific requirements to be implemented for a given operation.

inconsistency Contradictory computer statements, as detected by the program.

inconsistent equations A set of equations that have no common solution.

Increductor A specialized radio-frequency magnetic amplifier or saturable reactor.

increment **1.** The difference between two successive values of a variable. **2.** A small change in a quantity, such as mass, distance, or time. **3.** A quantity to be added to another quantity. **4.** The difference in value between adjacent hash marks on a meter scale.

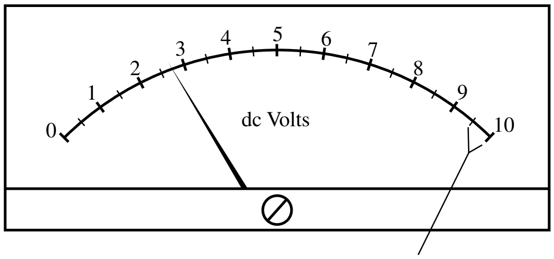

increment, 4

incremental computer A computer that operates on changes in variables. Example: differential analyzer.

incremental digital recorder A magnetic tape recorder that moves the tape across the record head in increments.

incremental display A device that converts digital data into a form for display (characters; graphs).

incremental inductance The inductance exhibited by an inductor, such as an iron-core choke, carrying a direct current.

incremental permeability The permeability exhibited by a material when an alternating-current magnetizing force is superimposed upon a direct current.

incremental plotter A device that, by direction of a computer program, provides the results of a program run in the form of curves or points on a curve, along with annotational characters.

incremental representation For incremental computers, a method of representing variables in terms of changes in the variables.

incremental sensitivity The smallest change in a quantity under measurement that can be detected by the instrument used.

ind **1.** Abbreviation of *indicator.* **2.** Abbreviation of INDUCTANCE (more commonly, *L*). **3.** Abbreviation of INDUCTOR (More commonly, *L*).

indefinite integral See INTEGRAL, **1**.

Indeo Also called *Digital Video Interactive (DVI).* A method of IMAGE COMPRESSION developed by Intel Corporation.

independent equations A set of equations having one common solution.

independent events In probability and statistics, the case where the occurrence of one event has no effect on the occurrence of another.

independent failure A circuit failure whose direct cause is not related to malfunctions elsewhere in the system.

independent mode In tracking supplies, an optional method of operation in which the separate units are adjustable independently of each other. Compare TRACKING MODE.

independent variable A changing quantity whose value at any instant is not governed by the value of any other quantity. Compare DEPENDENT VARIABLE.

index **1.** A reference line, hair, or point (e.g., a *file index*). **2.** In mathematics, an exponent. **3.** A ratio of one quantity to another, as *index of refraction.* **4.** In a computer memory, a table of references in a key sequence; it can be addressed to find the addresses of other data items. **5.** A number that is used to select a specific item within an array of items in memory.

index counter Also called *tape counter.* In a magnetic tape recorder, a (usually electromechanical) counting device that the operator can refer to in order to find material on the tape.

indexed address During or preceding the execution of a computer program instruction, an address that is modified by the content of an index register.

indexing **1.** An information retrieval technique used with files on a direct-access storage medium or on tables in memory. **2.** To modify an instruction using an index word.

index of modulation In frequency modulation, the ratio of carrier frequency deviation to modulating frequency.

index of refraction Symbol, *n*. The ratio v_1/v_2, where v_1 is the speed of energy propagation in the first medium through which the energy passes, and v_2 is the speed in the second medium.

index register Abbreviation, XR. In digital computer operations, a register holding a modifier that allows data to be directly addressed (each program refers to an index register when addressing storage locations). Also called MODIFIER REGISTER.

index word A word (bit group) containing a modifier that will be added to a basic instruction when it is executed during a program run.

India mica High-grade mica mined in India. Its excellent dielectric properties make it useful for capacitor stacks, high-*Q* radio-frequency circuits, and other critical applications.

indicated horsepower Abbreviation, ihp. Horsepower calculated from data or ratings, as opposed to measured horsepower.

indicated horsepower-hours Abbreviation, ihp-hr. Horsepower-hours based on calculation of indicated horsepower.

indicating fuse A fuse that provides some signal (such as a protruding pin) to show that it has blown.

indicating instrument An instrument, such as a meter, that provides direct readings of a measured quantity, as opposed to an instrument, such as a bridge, that must be manipulated and whose operation must often be followed by calculations.

indicating lamp A lamp that is marked or coded so that when it is on or off it conveys information.

indicator **1.** Meter (see METER, **1**). **2.** See MONITOR. **3.** See ANNUNCIATOR. **4.** In a computer, a device that can be set by a specific condition (e.g., by a negative result or error indicator).

indicator probe A test probe having a built-in meter.

indicial response Symbol, I(T). The sum of the transient and steady-state responses to a unit function.

indirect addressing In computer programming, a technique in which the address in an instruction refers to a different location containing another address, that can specify yet another address or an operand. Also called *multilevel addressing.*

indirect coupling Collectively, capacitive and inductive coupling, as opposed to direct coupling.

indirect ground An unintentional ground connection (e.g., accidental grounding of part of a circuit) or one obtained through a roundabout path. Compare DIRECT GROUND.

indirect light Light that has been reflected from one or more surfaces. Compare DIRECT LIGHT.

indirectly controlled Influenced by a directly controlled parameter, but not itself directly controlled.

indirectly grounded Connected to earth or to the lowest-potential point in a system inadvertently or through a roundabout path (e.g., by means of an indirect ground). Compare DIRECTLY GROUNDED.

indirectly heated cathode An electron-tube cathode consisting of a cylindrical or rectangular sleeve coated with a substance that is a rich emitter of electrons; it is heated by a filament inside the cylinder.

indirectly heated thermistor A thermistor whose temperature is changed by a built-in heater (filament) operated by the control current.

indirectly heated thermocouple A meter thermocouple heated by a small heater (filament) through which the signal current passes.

indirect material A semiconductor substance in which electrons move from the conduction band to the valence band in discrete jumps or steps.

indirect measurement The measurement of a quantity by comparing it with a similar quantity, using an instrument that requires adjustment or manipulation (rather than a simple meter). For example, resistance can be measured with a bridge, instead of an ohmmeter. Compare DIRECT MEASUREMENT.

indirect piezoelectricity In a piezoelectric crystal, the application of a voltage for the purpose of producing a strain on the crystal. A piezoelectric buzzer operates on this principle.

indirect scanning A method of video scanning, in which a fast-moving spot of light scans the film or an object and is passed through the film (or reflected by the object) to a photocell.

indirect wave **1.** In communications, a wave that arrives at a receiver after having traveled via refraction, reflection, or both. **2.** A radio wave propagated via the ionosphere. **3.** A wave reflected from some object, such as the moon or a meteor trail. **4.** A wave received from a satellite, originating from a distant earth station.

indium Symbol, In. A metallic element. Atomic number, 49. Atomic weight, 114.82. Used as a dopant in semiconductor processing.

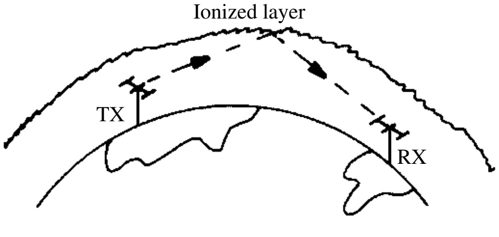

indirect wave

indoor antenna An antenna erected and operated in a building, but kept away from other objects as much as possible.

indoor radiation Electromagnetic radiation from the part of an antenna feeder or lead-in that is inside the transmitter building.

indoor transformer A power service transformer that, for protection against the elements, is installed inside the building it serves.

induced Brought about by the influence of a magnetic or electric field.

induced charge An electric charge produced in a body by the electric field surrounding another charge.

induced current An alternating current established in one circuit by the alternating magnetic field of another circuit. Also see INDUCTION.

induced EMF See INDUCED VOLTAGE.

induced failure A form of component failure that occurs because of operation beyond the normal specifications.

induced voltage An alternating voltage set up across one circuit (especially a coil) by the alternating magnetic field of another circuit. Also see INDUCTION.

inductance Symbol, *L.* Unit, henry. In a conductor, device, or circuit, the inertial property (caused by an induced reverse voltage) that opposes the flow of current when a voltage is applied; it opposes a change in current that has been established. Also see HENRY, INDUCTION, and MUTUAL INDUCTANCE.

inductance bridge An alternating-current bridge for measuring inductance in terms of a standard inductance or a standard capacitance. See, for example, HAY BRIDGE, MAXWELL BRIDGE, and OWEN BRIDGE.

inductance-capacitance Abbreviation, *LC.* **1.** A combination of inductance and capacitance in a circuit, such as a filter, a parallel-resonant circuit, or a series-resonant circuit. **2.** Pertaining to a device for measuring inductance and capacitance (e.g., *LC bridge* and *LC meter*).

inductance-capacitance bridge An alternating-current bridge for measuring inductance and capacitance only.

inductance-capacitance filter A filter composed of inductors and capacitors. Also called *LC filter.*

inductance-capacitance meter A direct-reading meter for measuring inductance and capacitance.

inductance-capacitance-resistance Abbreviation, *LCR.* **1.** A combination of inductance, capacitance, and resistance in a circuit, such as a basic tuned circuit. The resistive component represents loss in the inductor and capacitor. **2.** Pertaining to a device for measuring inductance, capacitance, and resistance (e.g., *LCR bridge* and *LCR meter*).

inductance-capacitance-resistance bridge See IMPEDANCE BRIDGE.

inductance coil See INDUCTOR.

inductance filter A filter using only an inductor, usually a coil of wire.

inductance-resistance time constant The time constant *t* (see ELECTRICAL TIME CONSTANT) of a circuit containing, ideally, only inductance and resistance. Mathematically, $t = LR$, where t is in seconds, L is the inductance in henrys, and R is the resistance in ohms. Also called *LR time constant.*

inductance standard A highly accurate, stable inductor used in precision measurements. Also see PRIMARY STANDARD and SECONDARY STANDARD.

induction **1.** The ability of an alternating, pulsating, or otherwise changing current flowing in one circuit to set up a current in a nearby circuit. The circuits need not be physically connected, but need only be linked by magnetic lines of flux. Also see SELF-INDUCTION. **2.** The phenomenon whereby a body becomes electrically charged by the field surrounding a nearby charged body. Also see ELECTRIC CHARGE.

induction coil A special high-voltage step-up transformer having an open core and a vibrator-interrupter in series with the primary winding, which carries direct current from a battery. The current is broken up into short pulses by the interrupter, and a high alternating-current voltage is generated in the secondary winding.

induction compass A compass whose indications depend on current induced in a coil revolving in the earth's magnetic field. Compare GYROCOMPASS and MAGNETIC COMPASS.

induction factor The ratio of total current to nonproductive current in an alternating-current circuit.

induction field The portion of an electromagnetic field that returns to a radiator, such as a coil, as opposed to the RADIATION FIELD.

induction frequency converter A mechanical device used for converting a signal at a fixed frequency to a signal at another fixed frequency.

induction furnace A furnace in which high-frequency magnetic fields induce currents in metal ores, causing the ore to become hot enough to melt.

induction heater A high-power, radio-frequency generator designed especially for induction heating.

induction heating The heating of metallic work samples by placing them in (but insulated from) a WORK COIL carrying current from a high-power radio-frequency generator. The workpiece is heated by radio-frequency currents induced in it as a result of the intense alternating magnetic field within the coil. Compare DIELECTRIC HEATING.

induction loss Loss of energy from a current-carrying conductor because of inductive coupling to a nearby conductor.

induction modulator See ELECTROMECHANICAL MODULATOR.

induction motor An electric motor in which the stator's rotating magnetic field makes the rotor revolve.

induction speaker An acoustic loudspeaker in which an audio-frequency current is passed through a diaphragm or coil located in a constant magnetic field. This results in movement of the diaphragm or coil.

induction transducer See INDUCTIVE TRANSDUCER.

induction-type landing system See DINGLEY INDUCTION-TYPE LANDING SYSTEM.

induction welding Welding in which the heating current flowing in the workpieces is induced by an electromagnetic field.

inductive capacitor A wound capacitor in which the inductance of the roll is controlled and specified. Such a capacitor is useful in compact filters and in single-frequency bypassing, where the reactive components are supplied by the capacitor. Compare NONINDUCTIVE CAPACITOR.

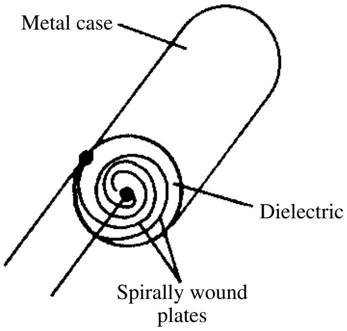

inductive capacitor

inductive circuit **1.** A circuit in which inductance predominates. **2.** A (theoretical) circuit containing inductance only.

inductive coupling The transfer of energy between two inductors (or inductive devices) by a linking electromagnetic field. Also see COEFFICIENT OF COUPLING, COUPLING, INDUCTION, and MUTUAL INDUCTANCE.

inductive feedback See MAGNETIC FEEDBACK.

inductive heater See INDUCTION HEATER.

inductive heating See INDUCTION HEATING.

inductive kick See BACK VOLTAGE and KICKBACK.

inductive load A load device that approaches a pure inductive reactance (e.g., loudspeaker and electric motor).

inductive loading In an antenna, the addition of inductance in series with the element(s). This reduces the resonant frequency for a radiator having a given physical length. It can also serve to reduce the physical length required for a radiator having a specified resonant frequency. Compare CAPACITIVE LOADING.

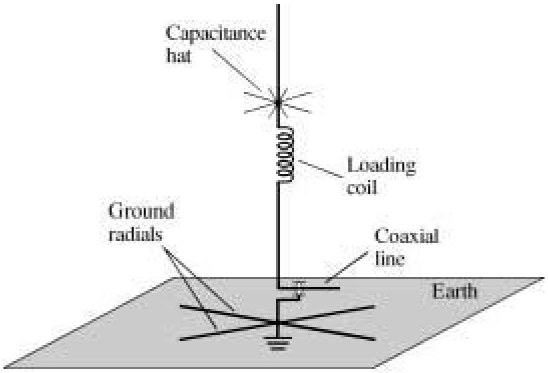

inductive loading

inductive logic A form of reasoning that demonstrates that a certain conclusion is highly probable, given a certain set of circumstances. This is of interest to researchers in artificial intelligence (AI). Compare DEDUCTIVE LOGIC.

inductive microphone A microphone in which sound waves vibrate a conductor or coil in a strong magnetic field, producing a corresponding alternating-current output by the resulting induction. Example: *dynamic microphone.*

induction neutralization Neutralization of a vacuum-tube radio-frequency power amplifier, via negative feedback from the output to the input through coupling coils.

inductive reactance Symbol, X_L. Unit, ohm. The reactance exhibited by an ideal inductor, considered as a positive imaginary-number quantity; $X_L = j6.28f_L$, where X_L is in ohms, f is the frequency in Hertz, L is the inductance in henrys, and j is the unit imaginary number (the square root of -1). Alternatively, f can be specified in megahertz, and L in microhenries. In a pure inductive reactance, current lags 90 degrees behind voltage. Also see INDUCTANCE, INDUCTION, INDUCTOR, and REACTANCE.

inductive switching Switching operations in a circuit containing an inductor. Switching time is influenced by the INDUCTANCE-RESISTANCE TIME CONSTANT of the inductor; overall operation is affected by the back voltage generated by the inductor.

inductive transducer A transducer in which the sensed phenomenon causes a change in inductance (or reluctance), which, in turn, causes a proportional change in output current, voltage, frequency, or bridge balance. Compare CAPACITIVE TRANSDUCER, CRYSTAL TRANSDUCER, MAGNETIC TRANSDUCER, and RESISTIVE TRANSDUCER.

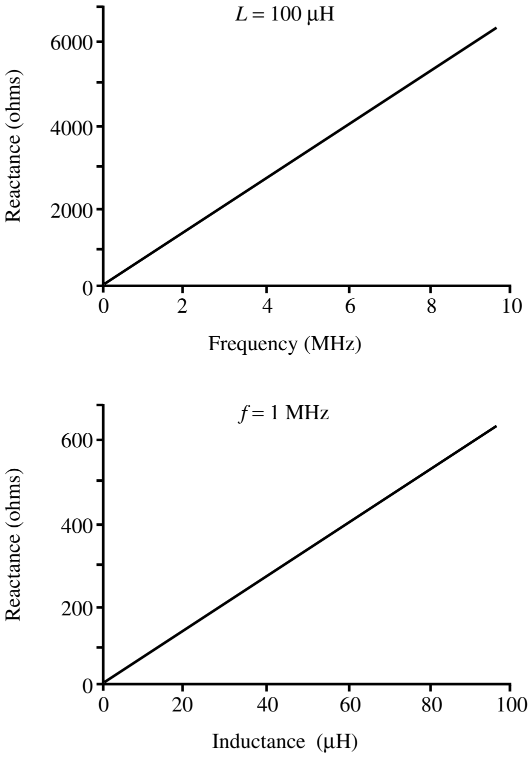

inductive reactance

inductive trimmer See TRIMMER INDUCTOR.

inductive tuning Also called *permeability tuning.* In a radio receiver, transmitter or transceiver, the adjustment of frequency by changing the inductance of a coil having a movable core.

inductivity See DIELECTRIC CONSTANT.

inductometer An instrument for measuring inductance in terms of the resonant frequency of an INDUCTANCE-CAPACITANCE (*LC*) circuit, in which *L* is the unknown inductance and *C* is calibration capacitance.

inductor A coil of wire wound according to various designs, with or without a core of ferromagnetic material, to concentrate the magnetic flux resulting from current flowing in the wire. The coiling of the wire and/or the addition of a ferromagnetic core increases the self-inductance compared with that of a straight wire having the same length. Also see INDUCTANCE; INDUCTION, **1**; and SELF-INDUCTANCE, **1**.

inductor alternator See ALTERNATOR.

inductor amplifier See MAGNETIC AMPLIFIER.

inductor decade See DECADE INDUCTOR.

inductor microphone See INDUCTIVE MICROPHONE.

inductors in parallel See PARALLEL INDUCTORS.

inductors in parallel-series See PARALLEL-SERIES INDUCTORS.

inductors in series See SERIES INDUCTORS.

inductors in series-parallel See SERIES-PARALLEL INDUCTORS.

inductor substitution box An enclosed assortment of common-value inductors that can be switched, one at a time, to a pair of terminals. In troubleshooting and circuit development, any of several useful fixed inductances can be thus obtained.

industrial data processing Abbreviation, IDP. The application of digital computers and associated equipment to industrial problems, through the classification, sorting, storing, and manipulation of information.

industrial electronics The branch of electronics concerned with manufacturing processes and their control, and with the operation and safeguarding of factories.

industrial instrumentation **1.** Supplementing an industrial process with electrical and electronic measuring instruments. **2.** The instruments used for the purpose defined in **1**.

industrial television Abbreviation, ITV. A usually closed-circuit television (CCTV) system, used as an adjunct to a manufacturing process, or as a means of communication or surveillance within an industrial plant.

industrial robot A robotic device used in industrial applications (e.g., mining, construction, manufacturing, or laboratory work).

industrial tube An (often heavy-duty) highly reliable vacuum tube designed expressly for industrial service, such as high-power radio or television broadcasting.

ineffective time The period during which an otherwise operational computer is not being used effectively because of delays or idle time.

inelastic collision A collision between charged particles in which one gains energy and the other loses energy.

inert gas A gas that does not readily react with other elements. Inert gases include *argon*, *helium*, *krypton*, *neon*, and *xenon.* Such gases are often used in hermetically sealed devices to retard corrosion.

inertance See ACOUSTIC INDUCTANCE.

inertia The tendency of a body at rest to remain at rest unless acted on by an outside force. Also, the tendency for a body in motion to maintain that motion unless acted on by an outside force. Compare MOMENTUM.

inertia in electric circuit The condition in a circuit containing inductance, in which a current change lags behind a voltage change (analogous to mechanical inertia; see INERTIA).

inertial guidance A system that automatically guides missiles and satellites in a desired trajectory without the need for continuous control by signals from a station.

inertia relay A time-delay relay whose operation is slowed by the addition of weights or other attachments.

inertia switch A switch that can sense a disturbance of its inertia.

inference engine A circuit that gives instructions to a computer or robot, by applying programmed rules to commands issued by a human operator. Comprises the functional portion of an EXPERT SYSTEM.

infinite Pertaining to a quantity or region that has no defined limits.

infinite baffle A loudspeaker baffle having no openings for the passage of sound from the front to the back of the speaker cone.

infinite-impedance detector A detector that offers the very high input impedance of a gate-source circuit and the large-signal capabilities of a diode detector. Audio-frequency output is taken across the source resistor, which is bypassed for radio-frequency signals. There is no drain resistor. Drain current increases with the input signal from a very low value at zero signal level.

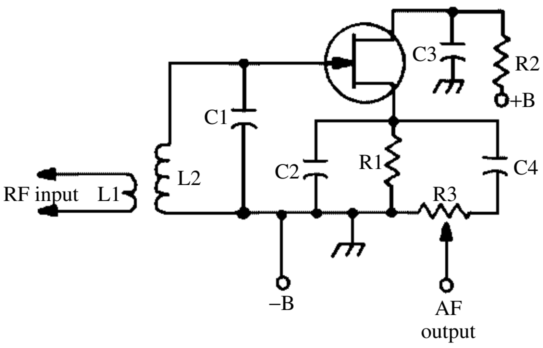

infinite-impedance detector

infinite line See INFINITE TRANSMISSION LINE.

infinite regress A reasoning pattern (either human or machine-based) that is fallacious because it defines or explains something in terms of itself.

infinite sample space In statistics, a sample space having no definite limits.

infinite series A mathematical series in which the number of terms is limitless. For example, ⅙ = 0.1 + 0.06 + 0.006 + 0.0006 + . . .

infinitesimal **1.** A quantity, such as a differential, that approaches zero as the limit. **2.** Pertaining to a quantity whose magnitude is extremely small or negligible. **3.** Pertaining to an extremely small change in a quantity or measured value.

infinite transmission line A theoretical transmission line with normal characteristics, but extending away from the signal generator or receiver for a limitless distance.

infinity Symbol, ∞. A quantity that is unlimited in duration or dimension. A quantity that increases without limit is sometimes said to "approach infinity."

infix notation A system of logical operation notation wherein operands are separated by operators, thus, A & B, where the ampersand means AND. Compare PREFIX NOTATION.

Infobahn See INFORMATION SUPERHIGHWAY.

infobond On a printed circuit board, a form of wiring on the side opposite the components. The wiring is used in place of the foil normally on such a circuit board.

information **1.** Collectively, data or communications, excluding the symbols or signals used to describe, present, or store them. **2.** The result of data processing (i.e., that which is derived from the compilation, analysis, and distillation of data).

information bits In an encoded signal, data characters or digits that can be treated to give information (excluding control characters).

information center A storage bank designed for use by many different subscribers, via computer.

information channel A channel through which data and associated signals are transmitted and received.

information feedback system In message transmission, a control system in which intelligence received at a terminal is returned to the sending unit for automatic verification.

information gate A device or circuit that opens and closes an information channel.

"information processor" species Anything that uses data to derive conclusions, to produce other data, or to take specific actions, and whose functioning can be explained entirely on the basis of data-processing operations. This includes computers and smart robots. Many (but not all) scientists believe that animals are also included; some believe that human beings qualify as well.

information retrieval In digital computer and data-processing operations, the categorizing and storage of information and the automatic recall of specific file items. Also see ACCESS TIME.

information separator An indicator that separates items of information or fields in a (usually variable-length) record.

information storage In digital computer and data processing operations, holding information in memory pending retrieval.

information superhighway **1.** General expression for a worldwide network consisting of computers (personal, educational, industrial, and government) interconnected by telephone lines. **2.** See INTERNET. **3.** A massive, evolving, somewhat controversial data communication network linking computers, television, and telephone systems. It uses high-speed, high-volume data links. Communication technologies include fiberoptics, radio-frequency repeaters, microwaves, geostationary satellites, and low-earth-orbit (LEO) satellite systems.

information word A character group representing stored information and managed, as a unit, by hardware or software.

infra- Prefix meaning *below* or *lower than* (e.g., INFRARED).

infrablack region In a composite video signal, the blacker-than-black region (see BLACKER THAN BLACK).

infradyne receiver A superheterodyne receiver in which the intermediate frequency is the sum of the signal and oscillator frequencies, rather than their (usual) difference.

infrared Pertaining to electromagnetic energy in a band whose wavelength is longer than that of visible light, but shorter than that of microwave energy.

infrared communication Communication by keying or modulating infrared rays.

infrared counter-countermeasure A military tactic in which action is taken against an enemy infrared countermeasure.

infrared countermeasure A military tactic using countermeasure methods to cripple enemy infrared equipment.

infrared detector A device that senses the presence of infrared energy. Some such detectors are bolometers, radiometers, radiomicrometers, and photocells.

infrared-emitting diode Abbreviation, IRED. A semiconductor diode, such as the gallium-arsenide type, that emits infrared rays when a current passes through the p-n junction in the forward direction.

infrared guidance A navigation and reconnaissance system using infrared rays.

infrared homing The method whereby a guided missile uses infrared rays to guide it to its target.

infrared light See INFRARED RAYS.

infrared motion detector See MOTION DETECTOR and INFRARED.

infrared photography Photography in which the scene is illuminated with infrared light or emits infrared rays, and the film is infrared sensitive.

infrared radiation See INFRARED RAYS.

infrared rays Radiation at frequencies in the INFRARED region. Also (somewhat mistakenly) called *heat rays.*

infrared remote control **1.** The use of an infrared link, usually over short line-of-sight distances, for the purpose of controlling the operation of electronic equipment. A common example is the local remote control of a television receiver or high-fidelity system. **2.** A small box, containing buttons, a transmitter and an infrared-emitting diode (IRED), used for local remote control of devices, such as television receivers and high-fidelity sound systems.

infrared spectrum The region of the electromagnetic spectrum in which INFRARED radiation is found. This band lies between the microwave radio spectrum and the visible-light spectrum.

infrared therapy The use of infrared rays by physicians and other practitioners to treat certain disorders.

infrared waves Electromagnetic waves whose lengths are greater than those of visible light waves, but less than those of microwaves.

infrared window Any portion of the infrared spectrum in which energy is easily transmitted through the lower atmosphere of the earth.

infrasonic Pertaining to acoustic disturbances whose frequencies are below the range of human hearing (less than about 20 Hz).

infrasonic intrusion detector A system that detects the presence of extremely low-frequency acoustic disturbances, and sends a signal to an alarm. Such INFRASONIC waves can be caused by various actions such as walking on a wooden floor, opening or closing a door, etc.

infrasonics The branch of physics dealing with INFRASONIC phenomena.

infrasound Acoustic disturbances in the air, whose frequencies are lower than about 20 Hz, and whose wavelengths are longer than about 55 feet (17 meters).

inharmonic distortion Distortion in which the frequencies of extraneous components are not harmonically related to the fundamental frequency. It is sometimes experienced when a tone-burst signal is applied to a loudspeaker.

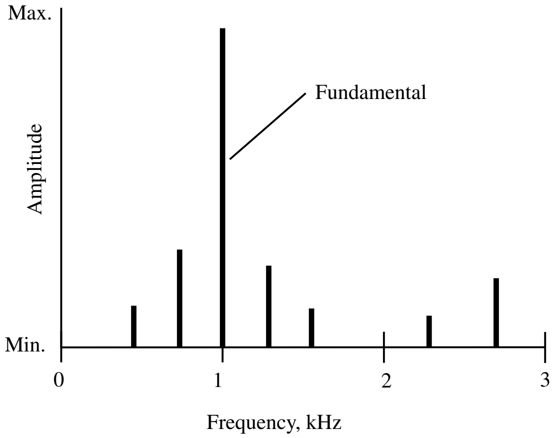

inharmonic distortion

inherent component A (usually extraneous) property possessed by a device because of its internal peculiarities. Thus, an inductor has inherent capacitance; a capacitor has inherent inductance.

inherited error In an extended calculation, an error carried through from one of the earlier steps.

inhibit **1.** In digital computer and logic operations, to prevent an action or block the input of data by means of a pulse. **2.** To delay an action or process.

inhibit gate A pulse-actuated gate circuit that acts as an INHIBITOR.

inhibitor **1.** A device or circuit that produces a pulse or signal that prevents an action, or that blocks data input. **2.** An additive, such as an organic liquid, that delays the hardening of a mixture, such as an encapsulating compound.

inhibit pulse In a computer, a drive pulse that prevents other pulses from changing the direction of magnetization in the cells of a magnetic core memory.

inhibit signal In digital computer and logic operations, the signal that causes an INHIBIT action.

initial drain **1.** The current supplied by a battery or cell at its rated voltage. **2.** The current delivered by a rechargeable battery or cell when it is put to use immediately after receiving a full charge.

initial failure The first failure occurring in the operation of a circuit or device.

initial instructions A resident computer routine used to aid program loading. Also called *initial order.*

initial ionizing event In the operation of a radioactivity counter, the first event that starts the chain of similar events constituting the count.

initialization A computer program instruction that sets the value of a variable to zero.

initial permeability Permeability in the low magnetization region of a material.

initial time delay Abbreviation, ITD. In acoustics, the elapsed time between the instant the direct sound wave is first heard, and the instant the first echoes (reflected sound waves) arrive.

initiate See TRIGGER.

injection **1.** Introducing a signal into a circuit or device. **2.** Introducing charge carriers (electrons or holes) into a semiconductor.

injector **1.** An element or electrode for INJECTION. **2.** A device or circuit that injects a signal into another device or circuit.

injector electrode See INJECTOR, **1**.

ink bleed In the printing of matter for optical character recognition, ink flow around the characters, often making them unrecognizable to the reader.

inkjet galvanometer A galvanometer whose movement controls the pressure of a jet of ink for making a recording on a paper chart. Also see LIQUID-JET OSCILLOGRAPH.

inkjet printer A printer commonly used with personal computers, in which images are created by jets of ink sprayed directly onto the paper. Noted for low operating noise level, high image resolution, and excellent color-reproduction capability.

ink-mist recorder A graphic recorder in which the line is traced by a mist of ink.

ink recorder A graphic recorder using a pen-and-ink stylus.

ink squeeze-out In the printing of matter for optical character recognition, the squeezing of ink from a character's center.

ink-vapor recorder See INK-MIST RECORDER.

in-lb Abbreviation of INCH-POUND.

inlead The part of an electrode that passes through the external shell or case of a component.

inline procedure The main portion of a COBOL computer program, responsible for the primary operations.

inline processing The action peculiar to a system that processes data almost immediately upon receipt (i.e., one that need not be capable of storing a lot of unprocessed data).

inline readout In digital computer operations, a readout device that displays digits side-by-side horizontally.

inline subroutine A subroutine that must be written each time it is needed, as compared with one that can be accessed by a program branch.

inline tuning Tuning of all the stages of a channel, such as an intermediate-frequency amplifier, to the same frequency.

inner conductor The inner wire or rod of a coaxial cable or coaxial tank. It generally carries the signal, and is isolated from the surrounding environment by the grounded OUTER CONDUCTOR.

inorganic Consisting of materials other than carbon compounds; therefore, it is not related to living things.

inorganic electrolyte Any electrolyte that is completely inorganic: containing no compounds of carbon.

in phase The condition in which alternating or pulsating waves or wave phenomena are in step with each other at all points. Compare OUT-OF-PHASE.

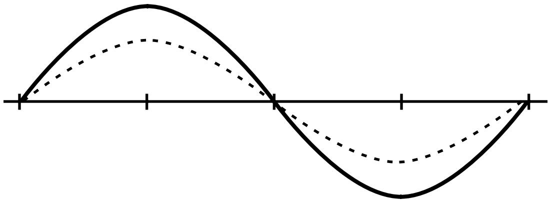

in phase

in-phase carrier See I-PHASE CARRIER.

in-phase current Resistive current in an ac circuit (i.e., current in phase with voltage). Compare QUADRATURE CURRENT.

in-phase feedback Feedback in phase with a main signal. Also called POSITIVE FEEDBACK and REGENERATION.

in-phase voltage A voltage that is in phase with another (reference) voltage.

in-plant system An automatic data communications system within a specific building or complex.

input **1.** Energy or information delivered or transferred to a circuit or device. **2.** The terminals of a device or circuit to which energy or information is applied. **3.** To deliver or transfer energy or information to a circuit or device (as to input data from a computer peripheral to memory).

input admittance Symbol, Y_i. The internal admittance of a circuit or device, as "seen" from the input terminals; the reciprocal of input impedance. Compare OUTPUT ADMITTANCE.

input area In a computer memory, an area set aside for data input from a source other than a program.

input bias current The input bias required by an operational amplifier.

input capacitance Symbol, C_i. **1.** The internal capacitance of a circuit or device, as "seen" from the input terminals. Compare OUTPUT CAPACITANCE. **2.** The grid-cathode capacitance of a vacuum tube.

input capacitor **1.** In a capacitance-coupled circuit, the input coupling capacitor. Compare OUTPUT CAPACITOR. **2.** The first capacitor in a capacitor-input filter (i.e., that capacitor electrically nearest the rectifier output electrode).

input choke The first choke in a choke-input filter (i.e., that choke electrically nearest the rectifier output electrode, when no preceding capacitor is used).

input circuit The circuit or subcircuit constituting the input section of a network or device. Compare OUTPUT CIRCUIT.

input clamp current The current from an input when the input is in a state below ground potential. A test for the input clamp diode.

input conductance Symbol, G_i. The internal conductance of a circuit or device, as "seen" from the input terminals; it is the reciprocal of INPUT RESISTANCE. Compare OUTPUT CONDUCTANCE.

input coupling capacitor See INPUT CAPACITOR, **1**.

input coupling transformer See INPUT TRANSFORMER.

input current Symbol, I_i. **1.** The current delivered to a circuit or device. **2.** Current flowing in the input leg or electrode of a circuit or device.

input device **1.** A device, such as an input transformer, that couples energy or information to a circuit or device. Compare OUTPUT DEVICE. **2.** A device through which another device receives data.

input equipment Collectively, input devices used with a computer.

input error voltage In an operational amplifier, the error voltage at the input terminals when a feedback loop operates around the amplifier.

input extender A diode network that provides increased fan-in for a logic circuit. Also see FAN-IN, **1**.

input gap In a velocity-modulated tube, the gap in which the electron stream is initially modulated.

input guarding A method of eliminating stray coupling among inputs in an integrated circuit. A shield is provided at the input; it is driven to follow along with the input voltage. This ensures low loss and minimum errors resulting from unwanted coupling.

input impedance Symbol, Z_i. The internal impedance of a circuit or device, as "seen" from the input terminals. Compare OUTPUT IMPEDANCE.

input limited The processing time limitation imposed by an input unit on the speed of a program run.

input noise current At the input of an integrated circuit, the root-mean-square (rms) or peak-to-peak (pk-pk) noise current existing within a specified range of frequencies.

input noise current density The noise current, usually expressed as a root-mean-square (rms) value, in a band 1 Hz wide around a given frequency.

input noise voltage At the input of an integrated circuit, the root-mean-square (rms) or peak-to-peak (pk-pk) noise voltage existing within a specified range of frequencies.

input noise voltage density The noise voltage, usually expressed as a root-mean-square (rms) value, in a band 1 Hz wide around a given frequency.

input offset current In an operational amplifier, the difference between the currents going to the input terminals when the output is zero.

input offset voltage In an operational amplifier, the potential that has to be applied between the input terminals for a zero output voltage.

input/output Abbreviation, I/O. **1.** Data transmitted to, or received from, a computer. **2.** A terminal through which data is transmitted to, or received from, a device.

input/output bound A condition affecting a system in which the time consumed by input and output operations is greater than that required for other processes.

input/output buffer A computer memory area specifically reserved for the receipt of data coming from or going to a peripheral.

input/output control The part of a computer system that coordinates activity between a central processor and peripherals.

input/output equipment **1.** In digital computer operations, devices for entering information into the computer or for reading information from it. Examples: keyboard, mouse, display, and optical scanner. **2.** In robotics and artificial intelligence, a data link between a controller and one or more robots, and/or between or among two or more controllers.

input/output isolation Arrangement or operation of a circuit or device so that there is no direct path between input and output terminals around the circuit or device. Also see ISOLATION.

input/output module See INPUT/OUTPUT EQUIPMENT.

input/output routine A routine for simplifying the programming of standard input/output equipment operations.

input/output switching The allocation of more than one channel to peripherals for communications with a central processor.

input/output voltage differential At a given load current, the potential difference that is necessary for an integrated circuit to operate according to its output voltage specifications.

input power Symbol, P_i. **1.** The power presented to the input terminals of a circuit or device. Also called POWER INPUT. Compare OUTPUT POWER. **2.** The operating power of a circuit or device (i.e., the power-supply requirement).

input protection In an integrated circuit, a means of preventing damage to the device from excessive voltage at the input, such as transient spikes or the result of malfunctioning of some other circuit.

input record **1.** A computer record of immediate interest that is ready for processing. **2.** During a computer program run, a record read into memory from an input device.

input recorder A device that makes a permanent record of the signals or data input to a circuit or system.

input register In a computer, a register that receives data from a peripheral relatively slowly and then passes it on to a central processor at a faster speed as a sequence of informational units. Also see REGISTER.

input resistance Symbol, R_i. The internal resistance of a circuit or device, as "seen" from the input terminals. Compare OUTPUT RESISTANCE.

input resonator In a velocity/modulated tube, the resonator in which electron bunching occurs.

input routine A computer program section that manages data transferal between an external storage medium and a memory input area.

input section **1.** See INPUT ROUTINE. **2.** See INPUT AREA.

input sensitivity **1.** The level of input-signal amplitude that will result in a certain signal-to-noise ratio at the output of a device. The specified signal-to-noise ratio is usually 10 or 20 dB. **2.** The level of input signal in a frequency-modulated device, required to produce a specified amount of noise quieting. The specified level of noise quieting is usually 20 dB. Alternatively, 12-dB SINAD (ratio of signal to the level of noise and distortion) can be specified. **3.** The minimum level of input voltage required to actuate a logic gate.

input signal The signal (current, voltage, and power) presented to the input terminals of a circuit or device for processing.

input tank In a double-tuned stage of a transmitter or power generator, the tank circuit in which the input signal is resonated. This is generally the base or gate circuit. Compare OUTPUT TANK.

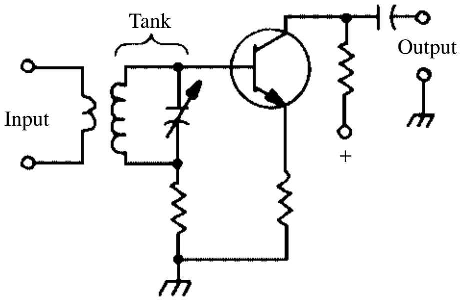

input tank

input terminals Terminals (usually a pair) associated with the input section of a circuit or device. Compare OUTPUT TERMINALS.

input transformer The transformer that delivers signal voltage or power to the input circuit of a network or device. Compare OUTPUT TRANSFORMER.

input uncertainty The combination of all parameters that result in adverse behavior in an operational amplifier.

input unit In a digital computer, the device or circuit that receives information from peripherals.

input voltage **1.** Symbol, E_i or V_i. The voltage presented to a circuit or device. Compare OUTPUT VOLTAGE, **1**. **2.** The voltage across the input leg or electrode of a circuit or device. Compare OUTPUT VOLTAGE, **2**.

input-voltage drift For an integrated circuit (IC), the time- and temperature-dependent change in output voltage divided by the IC's open-loop voltage gain.

input-voltage offset For a differential amplifier, the input signal voltage at the differential input that results in zero output voltage.

input-voltage range The range, in volts, over which the input voltage can fluctuate in an integrated circuit so that the common-mode rejection ratio (CMRR) specifications are not exceeded.

input winding The signal winding of a magnetic amplifier.

inquiry A programmed request for information from storage in a computer.

inquiry display terminal A video display/keyboard terminal used to make an inquiry to a computer system, and display the response.

inquiry station A terminal from which an inquiry can be sent to a central computer.

inrush The initial surge of current that occurs when voltage is first applied to the primary winding of a transformer with no load connected.

inscribe To convert data to a form on a document that is readable by a character-recognition device, as through the use of magnetic ink, for example.

insect robot A member of a fleet of robots, all of which are under the control of a single computer.

The term arises because the system functions like an anthill or beehive, in which the individual machines are "stupid," but the system as a whole is "smart." Such robots often have six legs, like insects. Compare AUTONOMOUS ROBOT.

insert A (usually metallic) bushing that can be molded into a plastic part (or pressed into it after molding is completed) to provide a bearing sleeve or threaded hole.

insert core A ferromagnetic core whose position can be adjusted to vary the inductance of the coil surrounding it.

insert edit **1.** In magnetic tape recording, a section of tape on which new audio is recorded over existing audio. **2.** The process of recording new audio over existing audio in a defined interval on a magnetic tape.

insertion gain In a circuit or system, the gain resulting from the amplifier inserted into the system; it is usually expressed in decibels. Compare INSERTION LOSS.

insertion loss Loss of energy or gain by placing certain devices or subcircuits (filters, impedance matchers, etc.) in a circuit. It is usually expressed in decibels. Also see INSERTION RESISTANCE.

insertion phase shift The difference in phase produced by a circuit installed in an electrical transmission line.

insertion resistance The resistance of a component or instrument that is introduced into a circuit. Thus, the internal resistance of a microammeter becomes an insertion resistance in the circuit in which the meter is connected for current measurement.

inside antenna See INDOOR ANTENNA.

inside diameter Abbreviation, ID. The innermost diameter of a body or figure having two concentric diameters. Compare OUTSIDE DIAMETER.

inside lead See START LEAD.

inside radiation See INDOOR RADIATION.

inside spider A voice-coil centering device within a loudspeaker.

inst **1.** Abbreviation of INSTRUMENT or INSTRUMENTATION. **2.** Abbreviation of INSTANT.

instability Inconsistency in the operation of a circuit or device, in the parameters of a device, or in an electrical quantity. It can be attributed to a number of causes, including temperature, loading, age, humidity, negative resistance, and radioactivity.

installation tape number An identification number given to a reel of magnetic tape by the processing facility.

instant Abbreviation, inst. The point in time at which an event occurs, or at which a quantity reaches a particular value.

instantaneous Occurring at a specified moment, or instant, of time.

instantaneous amplitude The amplitude, specified in amperes, volts, or watts, of a signal, specified at a particular moment in time.

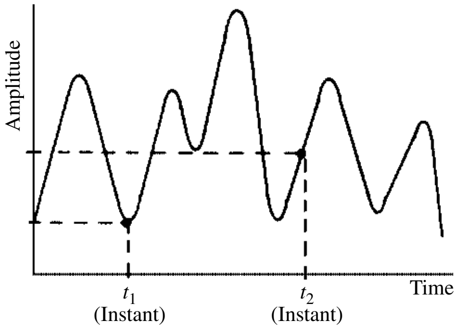

instantaneous amplitude

instantaneous automatic gain control Abbreviation, IAGC. An automatic gain control whose operation almost immediately follows a change in signal amplitude.

instantaneous automatic volume control Abbreviation, IAVC. An instantaneous automatic gain control system for the immediate control of volume in receivers and audio-frequency amplifiers.

instantaneous companding A form of companding that operates according to the instantaneous amplitude of the input signal.

instantaneous contacts Timer contacts that open or close almost immediately upon application of the control signal.

instantaneous current Symbol, i or I_i. The value of an alternating or fluctuating current at a particular instant in the cycle.

instantaneous disc A phonograph disc that can be played back immediately after being recorded.

instantaneous frequency The frequency of a signal at a particular moment in time. The instantaneous frequency changes in frequency-modulated or phase-modulated signals.

instantaneous power **1.** In a single-sideband, suppressed-carrier signal, the power at a specified instant in time. It varies between zero and the peak envelope power (PEP). **2.** The output power of an audio amplifier at a specified instant in time.

instantaneous power output The rate of power delivery to a load at a given instant.

instantaneous relay A relay, such as a fully electronic type (having no moving parts), that shows virtually no delay in its operation.

instantaneous sample A measurement obtained by INSTANTANEOUS SAMPLING.

instantaneous sampling The measurement of wave or signal amplitude at a specific moment in time. See, for example, INSTANTANEOUS CURRENT and INSTANTANEOUS VOLTAGE.

instantaneous speech power In the output of an audio amplifier, the instantaneous value of power in a speech wave, as opposed to that in a sine wave. Also see INSTANTANEOUS VALUE and SPEECH POWER, **1**.

instantaneous value The magnitude of a fluctuating value at a selected instant in time. See, for example, INSTANTANEOUS CURRENT, INSTANTANEOUS POWER, INSTANTANEOUS SPEECH POWER, and INSTANTANEOUS VOLTAGE. Compare AVERAGE VALUE, and EFFECTIVE VALUE.

instantaneous voltage Symbol, e or E_i. The value of an alternating or fluctuating voltage at a particular instant in the cycle.

instant loop In electronic security applications, a circuit that actuates an alarm without delay when an intrusion is detected.

instruction In digital computer operations, a set of bits defining an operation. Consists of an operation code specifying the operation to be performed, one or more operands or their addresses, and one or more modifiers or their addresses (to modify the operand or its address).

instruction address In a computer memory, the address of a location containing an instruction.

instruction address register Also called *program counter.* A register that holds instruction addresses so that the retrieval of the instructions from memory can be controlled during a program run.

instruction code Also called INSTRUCTION SET. The symbols and characters that compose the syntax of a computer programming language.

instruction format In a computer's basic machine code, the part that specifies how characters or digits are used to represent the codes within the machine's instruction set.

instruction modification In a computer instruction, a change in the instruction code that makes the computer do a different operation when the routine containing the code is encountered again.

instruction register A register in a computer containing the address of the current instruction. Also called CONTROL REGISTER (abbreviation, CR).

instruction set **1.** The range of commands that form a programming language. **2.** See INSTRUCTION CODE.

instruction storage A memory circuit that stores computer instructions or programs.

instruction time The time required for a control unit to analyze and implement a computer program instruction.

instruction word In digital computer programming, a word containing the instruction code (type of operation to be performed) and the address part (location of the associated data in storage).

instrument A device for measuring electrical quantities or the performance of electronic equipment. A *meter* provides a direct indication; other devices, such as a *bridge*, must be adjusted, the measured quantities being determined from one or more adjustments (sometimes augmented with calculations).

instrumental error See INSTRUMENT ERROR.

instrument amplifier Also called INSTRUMENTATION AMPLIFIER. A high-gain, wideband amplifier that increases the sensitivity of an instrument (such as an oscilloscope, meter, or graphic recorder).

instrument-approach system See INSTRUMENT LANDING SYSTEM.

instrumentation Planning and providing instruments and instrument systems for the collection and, sometimes, storage and analysis of data.

instrumentation amplifier **1.** A form of integrated-circuit voltage amplifier designed for high linearity, high input impedance, and high common-mode rejection. It is intended for use with electronic instruments. **2.** See INSTRUMENT AMPLIFIER.

instrument chopper A refined chopper for converting a direct-current (dc) signal to alternating current (ac) for an ac instrument, such as a voltmeter or recorder.

instrument error Discrepancy in measured quantities resulting from inaccuracy of the instrument used, insertion resistance, environmental factors, operator error, etc.

instrument flight Also called *blind flight.* Aircraft flight guided by navigational instruments and signals alone. Required when visibility is extremely poor.

instrument fuse A fast-acting, low-current fuse used to protect a sensitive instrument, such as a galvanometer, milliammeter, and/or microammeter.

instrument lamp A light or lamp that illuminates the face of an instrument to facilitate viewing in the dark.

instrument landing Also called *blind landing.* Aircraft landing guided entirely by instruments. Required when visibility is poor and when landing is imperative at a given location at a given time.

instrument landing station The radio or radar station in a blind-landing system (see INSTRUMENT LANDING SYSTEM).

instrument landing system Abbreviation, ILS. The complete instrument and signal system (on the ground or in aircraft) required for an INSTRUMENT LANDING.

instrument multiplier See MULTIPLIER PROBE, **1**.

instrument preamplifier An external, sensitive amplifier for an instrument that has an internal input amplifier. Also see INSTRUMENT AMPLIFIER.

instrument relay See METER RELAY.

instrument resistance See METER RESISTANCE.

instrument shunt A resistance connected in parallel with a current-measuring instrument, used to increase the range of currents that can be measured.

instrument transformer A transformer used to change the range of an alternating-current meter.

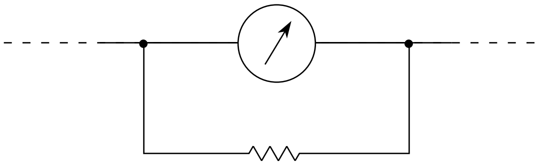

instrument shunt

For ammeters, it is called a *current transformer*; for voltmeters, it is called a *potential transformer.*

insulant A nonconducting material, used to prevent the flow of electric current between or among points. See INSULATOR, **1**.

insulated Isolated from conductors by an INSULANT.

insulated-gate field-effect transistor Abbreviation, IGFET. See METAL-OXIDE SILICON FET.

insulated resistor A resistor around which is molded a nonconducting material, such as vitreous enamel or a plastic.

insulating tape Electrical insulation in the form of a thin, usually adhesive, strip of fabric, paper, or plastic.

insulation **1.** A coating of dielectric material that prevents a short circuit between a conductor and the surrounding environment. **2.** The application of a dielectric coating to an electrical conductor. **3.** Electrical separation between or among different components, circuits, or systems.

insulation breakdown Current leakage through, and rupture of, an insulating material because of high-voltage stress.

insulation ratings Collectively, the dielectric constant, dielectric strength, power factor, and resistivity of an insulating material. Sometimes included are such physical properties as rupture strength, melting point, etc.

insulation resistance The very high resistance exhibited by a good insulating material. It is expressed in megohms (or higher units of resistance) for a sample of material of stated volume or area.

insulation system Collectively, the materials needed to insulate a given electronic device.

insulator **1.** A material that, ideally, conducts no electricity; it can, therefore, be used for isolation and protection of energized circuits and components (also see DIELECTRIC). Actually, no insulator is perfectly nonconductive (see, for example, INSULATION RESISTANCE). **2.** A molded piece of solid insulating material, used to electrically isolate conductors—especially in antenna systems and power transmission lines. **3.** Any body made from an insulating material.

insulator arcover A sudden arc, or flow of current, over the surface of an insulator, because of excessive voltage.

integer A positive or negative whole number, as opposed to a fraction or mixed number.

integral **1.** Also called *indefinite integral* and *antiderivative.* For given mathematical function *f*, function *g*, whose derivative is equal to *f*. **2.** Also called *definite integral.* The area under a curve of a function, between two vertical lines defined by two specific points in the domain of the function. **3.** The part of a number to the left of the radix point. **4.** Pertaining to integers (positive or negative whole numbers) or quantities that can be represented by integers.

integral action In automatic control operations, a control action delivering a corrective signal proportional to the time that the controlled quantity has differed from a desired value.

integral contact In a relay or switch, a contact that carries current to be switched.

integral-horsepower motor A motor rated at one horsepower.

integral multiple A whole multiple of a number. Thus, a harmonic is an integral multiple of fundamental frequency *f*: 2*f*, 3*f*, 4*f*, etc.

integral number See INTEGER.

integrate **1.** To perform the function of mathematical or electrical INTEGRATION. **2.** To construct a circuit on a piece of semiconductor material.

integrated Constructed on a single piece of material, such as a semiconductor wafer.

integrated amplifier An audio-frequency (AF) amplifier having a preamplifier, intermediate amplifier, and output amplifier on a single chassis.

integrated capacitor In an integrated circuit, a fixed capacitor in which one plate is a layer of material diffused into the substrate, the dielectric is a thin-oxide film grown on top of the first layer, and the other plate is a metal layer deposited on top of the oxide film.

integrated circuit Abbreviation, IC. A circuit whose components and connecting "wires" are made by processing distinct areas of a chip of semiconductor material, such as silicon. Classified according to construction (e.g., *monolithic IC*, *thin-film IC*, *hybrid IC*).

integrated data processing Abbreviation, IDP. The detailed electronic classification, sorting, storage, and mathematical processing of data within a coordinated system of equipment, usually at one location.

integrated electronics The branch of electronics that is concerned with the design and fabrication of integrated circuits.

integrated resistor See DIFFUSED-LAYER RESISTOR.

Integrated Services Digital Network Abbreviation, ISDN. A communications network or connection intended primarily for Internet access via telephone lines. Allows significantly higher data speed than is possible with a conventional analog connection. In addition, it is possible to use a digital system, such as a computer, online simultaneously with an analog voice conversation.

integrating circuit See INTEGRATING NETWORK.

integrating galvanometer A device for measuring the change in electric flux produced in a coil in an electric field. Even very slow changes can be measured.

integrating meter An instrument whose indication is a summation (usually) of an electrical quantity that is time-dependent (e.g., *ampere-hour meter* and *watt-hour meter*).

integrating motor An electric motor that follows the integral of the input signal. The angle of rotation of the motor shaft is equal to the integral of the input waveform.

integrating network A four-terminal network whose output voltage is proportional to the time integral of the input voltage. It can be a passive resistance-capacitance (RC) circuit or it can use an operational amplifier. Compare DIFFERENTIATING NETWORK.

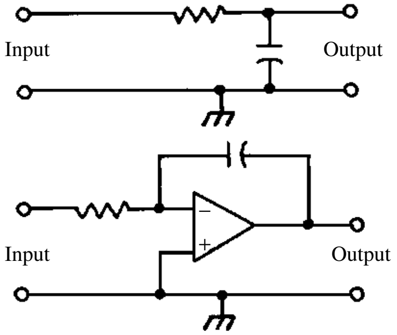

integrating networks

integrating photometer A photometer whose reading is the average candlepower at all angles in one plane.

integration **1.** The process of determining a mathematical function when its derivative is given. **2.** The processing of a signal by an INTEGRATOR circuit. **3.** Collectively, the processes by which an INTEGRATED CIRCUIT is manufactured.

integrator **1.** See INTEGRATING NETWORK. **2.** A device having an output variable whose value is proportional to the integral of one variable, with respect to another, or is proportional to the integral of an input variable, with respect to elapsed time.

intelligence **1.** Meaningful data that modulates a carrier [e.g., the voice or music in a frequency-modulated (FM) radio signal, or the image in a television signal]. **2.** Also called *machine intelligence.* The quality of a system or device, especially a computer, that allows it to "learn" (i.e., to better its capability by repeatedly operating on a given problem).

intelligence bandwidth **1.** The bandwidth necessary to convey a specified amount of data within a certain period of time. **2.** The total bandwidth of one complete signal channel in a communications or broadcast system.

intelligence signal **1.** A signal that conveys data or information. **2.** The modulating waveform in a communications or broadcast transmission.

intelligent network Abbreviation, IN. **1.** In general, any advanced and sophisticated communications network, particularly a broadband digital network. **2.** A network designed to readily accommodate new technologies and services, such as videoconferencing, interactive television, or remote control.

intelligent terminal A computer terminal (e.g., an input/output video display/keyboard unit) that through its circuitry (i.e., by use of a microprocessor) has some data-processing ability.

intelligibility tests Tests that measure the coherence of electronically reproduced speech.

intensification of image See IMAGE INTENSIFICATION.

intensifying ring In some electrostatic cathode-ray tubes, an internal metal ring serving as an extra anode to accelerate the beam and, thus, brighten the image.

intensity The degree or extent of a phenomenon (such as amplitude, brightness, loudness, power, force, etc.).

intensity control In an oscilloscope circuit, the potentiometer that adjusts the direct-current voltage on the control electrode of the cathode-ray tube and, accordingly, the brightness of the image. Also called BRIGHTNESS CONTROL and BRILLIANCE CONTROL.

intensity level **1.** A measure of sound magnitude, expressed in decibels, with respect to a value of one microwatt per square centimeter (10^{-6} W/cm^2) at sea level in the atmosphere. **2.** The setting of the brightness control in a cathode-ray-tube device.

intensity modulation **1.** Modulation of electron-beam intensity in a cathode-ray tube. Also called *z-axis modulation.* **2.** Sometimes, the video-signal modulation in a television image.

intensity-modulation amplifier The z-axis amplifier in an oscilloscope. Also see INTENSITY MODULATION.

interaction The (sometimes mutual) influence of one circuit or device on the behavior of another, as in induction.

interactive display A computer display device with which its operator can supply data to the computer in response to what is displayed. Example: *touch screen.*

interactive graphics A computer graphics system using a cathode-ray tube to draw or modify three-dimensional representations.

interactive photovoltaic system A solar-power plant that operates in conjunction with the utility companies. Energy is sold to the companies during times of daylight and minimum usage, and is bought back from the companies at night or during times of heavy usage. The principal advantage of this system is that the user can keep using electricity (by buying it all from the utilities) if the solar-energy system breaks down. But such a system does not provide the independence from utility companies that some users desire. Compare STAND-ALONE PHOTOVOLTAIC SYSTEM.

interactive program A computer program in which the machine and its operator engage in two-way communication. Most personal computing software is of this type, in contrast to programs that carry out all their functions without operator intervention (other than initialization).

interactive television Television provided to consumers, in which viewers can transmit data, as well as receive it. For example, a survey might be conducted in which viewers are polled and send in their responses. Another example: products might be ordered while viewing an advertisement.

interactive mode See CONVERSATIONAL MODE.

interbase resistance The internal resistance between the bases of a unijunction transistor.

interblock A part of a computer program or a hardware device that will prevent interference between parts of a computer system.

interblock space **1.** On a magnetic tape, the space between recordings, caused by starting and stopping the tape. **2.** On magnetic tape used as a computer storage medium, the interval between recorded blocks.

intercarrier receiver A television (TV) receiver circuit in which video, sound, and sync components of the composite TV signal are amplified together in the radio-frequency (RF), intermediate-frequency (IF), and video IF stages; then they are separated in the video detector and video amplifier stages. Compare SPLIT-SOUND RECEIVER.

intercept receiver In military service, a search receiver tuned over a wide band of frequencies to locate and evaluate enemy signals.

interchangeability The ability of one component to substitute directly for another component of the same kind. Example: capacitor interchangeability, transistor interchangeability. Also see REPLACEMENT.

intercharacter space The three-unit interval between letter symbols in telegraphy. Compare INTERWORD SPACE.

intercom A comparatively simple two-way telephone or low-power radio system for use on the premises of a home or business.

intercommunicator See INTERCOM.

interconnection **1.** A mutual connection of separate circuits. **2.** The interconnection of two or more separate power-generating systems.

interdigital contacts A pair of contacts with “fingers” that are plated, printed, or deposited on the surface of a resistor material or semiconductor substrate. The fingers of each contact are interconnected at one end, the fingers of one contact being interleaved with those of the other.

interdigital tube A magnetron having a cathode surrounded by anode segments that are alternately interconnected at opposite ends in the manner of INTERDIGITAL CONTACTS.

interelectrode capacitance Capacitance between or among electrodes—especially between the plate and control grid of a vacuum tube.

interelement capacitance Internal pn-junction capacitance in a semiconductor device, such as a diode or transistor.

interface **1.** The circuitry that interconnects and provides compatibility between a central processor and peripherals in a computer system. **2.** Collectively, the hardware and software that allows a computer to interact with its operator. **3.** To provide an efficient pathway for data between two devices or systems. **4.** The meeting of surfaces or regions in a material. **5.** The surface of a body that mates with another body similar or identical to it.

interface resistance See CATHODE INTERFACE.

interface routine A computer program routine that links one system to another.

interfacial connection A connection that runs through a printed-circuit board and joins circuit joints on opposite faces of the board.

interference **1.** In communications, degradation of reception caused by noise or undesired signals. **2.** The interaction of acoustic or electromagnetic waves from more than one source, especially when they are of the same frequency, producing a characteristic INTERFERENCE PATTERN of high-amplitude and low-amplitude regions.

interference attenuator A device or mode of operation that reduces the amplitude of interference.

interference eliminator A filter, wavetrap, or similar device that removes interfering signals or noise. Also see INTERFERENCE.

interference filter See INTERFERENCE ELIMINATOR.

interference pattern A regular pattern of high-amplitude and low-amplitude regions, lobes, or bands, produced when waves of identical frequency from two or more sources combine in varying phase. Such patterns can be observed with sound, radio waves, infrared, visible light, ultraviolet, X rays, and gamma rays. The phenomenon is of interest in acoustic engineering, the design of radio antenna systems, and in physics (particularly optics).

interference stub A length of twin-lead feeder cut to appropriate length, connected to the antenna-input terminals of a television receiver, and short-circuited at the opposite end. A stub of the correct length resonates at the frequency of an interfering signal and, acting as a wavetrap, keeps it out of the receiver. Also see STUB.

interference trap A wavetrap that suppresses interfering signals at the rejection frequency of the trap.

interferometer **1.** A radio telescope having two antennas spaced at a distance of many wavelengths, providing much greater resolution than a single antenna. Pioneered by M. Ryle of England and J.L. Pawsey of Australia. **2.** Any device that displays an INTERFERENCE PATTERN for testing or experimental purposes.

interfix A method used in information-retrieval systems that eliminates ambiguity in the responses to inquiries by describing the relationship between keywords in a record.

Interframe A method of digital IMAGE COMPRESSION developed by *MPEG (Moving Picture Experts Group).* It operates by eliminating redundant data from between image frames. Compare INTRAFRAME.

interharmonic beats Beat notes produced by various combinations of the harmonics of a signal.

interim storage See TEMPORARY STORAGE.

interior label On a magnetic tape used as a computer-storage medium, a label recorded at the beginning of the tape. Compare EXTERIOR LABEL.

interior protection **1.** In electronic security applications, a set of sensors contained entirely within the region to be protected. **2.** The installation and operation of a security system whose sensors are all within the region to be protected.

interlace A form of data storage in which portions of the data are stored in alternate locations in the tape or disk.

interlaced field A video image field produced by INTERLACED SCANNING.

interlaced scanning In the display of a video image, the alternate presentation of the even- and odd-line fields. This process increases the obtainable image resolution for a given refresh rate, but can result in "jerkiness" of the image when rapid motion is portrayed.

interlace factor A number expressing the extent to which two fields are interlaced. Also see INTERLACED SCANNING.

interleaving In multiprogramming, the inclusion in a program of segments of another program so that both can be effectively executed simultaneously.

interlock switch See ELECTRICAL INTERLOCK.

intermediate amplifier See BUFFER.

intermediate frequency Abbreviation, IF. In a superheterodyne circuit, the frequency of the signal that results from beating the incoming signal with the signal produced by the local oscillator.

intermediate-frequency amplifier In a superheterodyne circuit, the fixed-frequency amplifier that boosts the intermediate-frequency signal. Also see INTERMEDIATE FREQUENCY.

intermediate-frequency channel Usually, the intermediate-frequency amplifier in a superheterodyne circuit, but sometimes including the second detector, automatic gain control (AGC), and oscillator stages.

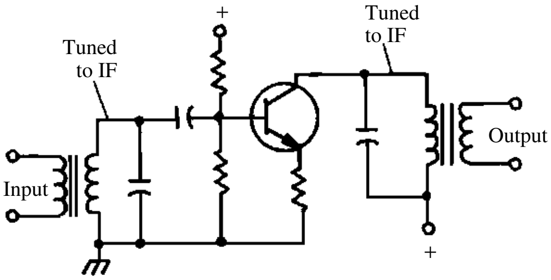

intermediate-frequency amplifier

intermediate-frequency converter See IF CONVERTER.

intermediate-frequency interference Interference from signals at the intermediate frequency of a receiver or instrument.

intermediate-frequency selectivity The selectivity of an intermediate-frequency (IF) channel alone, usually determined by the characteristics of the bandpass filter(s) in the IF channel.

intermediate-frequency transformer A coupling transformer designed for use in an intermediate-frequency amplifier.

intermediate-puck drive In a tape recorder, a speed-reducing drive system in which an intermediate wheel conveys motion from the motor shaft to the rim of the flywheel.

intermediate puck wheel See IDLER WHEEL.

intermediate repeater In wire telephony, a repeater inserted into a line or trunk at some point other than the end.

intermediate result Obtained during a program run or the execution of a subroutine, a result that is used again as an operand in deriving the final result.

intermediate section Any of the internal sections of a multisection filter. Thus, the middle section of a three-section filter.

intermediate storage In a computer system, a storage medium for temporarily holding totals or working figures. Also called WORK AREA.

intermediate subcarrier A modulated or unmodulated subcarrier that modulates either a carrier or another intermediate subcarrier.

intermittent **1.** Pertaining to a circuit fault, such as an open or short circuit, that occurs some of the time, but not all the time. **2.** See INTERMITTENT CONDITION. **3.** Pertaining to a phenomenon that is observed some of the time, but not all the time; sporadic. **4.** Pertaining to a DUTY CYCLE greater than zero, but less than 100 percent; usually between 25 and 50 percent.

intermittent commercial and amateur service Abbreviation, ICAS. Operation of equipment, such as radio transmitters, for short, irregular periods, as in amateur (hobbyist) activity or infrequent commercial service. ICAS ratings are higher than continuous commercial service (CCS) ratings. Compare CONTINUOUS COMMERCIAL SERVICE.

intermittent condition A defect in a circuit or device that causes erratic and unreliable operation. The cause of such a problem is often difficult to determine.

intermittent dc See INTERMITTENT DIRECT CURRENT.

intermittent direct current A regularly pulsed unidirectional current. Also called PULSATING DIRECT CURRENT.

intermittent duty A DUTY CYCLE of less than 100 percent, but greater than zero. Generally, an operating duty cycle of 25 to 50 percent.

intermittent-duty rating The dissipation or power rating of a component, circuit, or system, under conditions of intermittent use, usually a 25-percent to 50-percent DUTY CYCLE.

intermittent operation Operation characterized by often long nonoperating intervals. Intermittent operation is often random, whereas on-off operation tends to be regular.

intermittent signal An interrupted signal resulting from the intermittent operation of a circuit or device.

intermodulation Abbreviation, IM. **1.** The (usually undesired) modulation of one signal by another, caused by nonlinear processing of the signals. **2.** The heterodyning of components in the sidebands produced by an amplitude-modulated (AM) or single-sideband (SSB) transmitter.

intermodulation distortion Abbreviation, IMD. **1.** Distortion products in the output of an amplitude-modulated (AM) or single-sideband (SSB) transmitter, caused by heterodyning of components in the sidebands. **2.** Distortion products in the output of an audio amplifier, caused by heterodyning of the fundamental components. **3.** The extent to which distortion as defined in 1 occurs. See INTERMODULATION-DISTORTION PERCENTAGE.

intermodulation-distortion meter See INTERMODULATION METER.

intermodulation-distortion percentage Abbreviation, IDP. The degree to which a low-frequency test signal modulates a higher-frequency test signal when both are applied simultaneously (in a prescribed amplitude ratio) to a device under test; $IDP = 100(b - a)/a$, where a is the peak-to-peak amplitude of the unmodulated high-frequency wave and b is the peak-to-peak amplitude of the modulated high-frequency wave.

intermodulation meter An instrument for measuring percentage of intermodulation distortion (IMD). The instrument combines a dual-frequency signal generator, filter circuits, and percent-of-modulation meter. Also see INTERMODULATION-DISTORTION PERCENTAGE.

intermodulation noise Electrical noise produced in one channel by signals in another; it is caused by INTERMODULATION.

internal absorptance The ratio of flux absorbed in a substance to the flux leaving at the entry surface of the substance. It does not include energy reflected at the entry surface.

internal amplification In a radioactivity counter tube, current enhancement resulting from cumulative ionization initiated by an ionizing particle.

internal arithmetic In a digital computer, arithmetic operations performed in the computer, as opposed to those performed by peripherals.

internal impedance The impedance in a device, as opposed to that added from the outside. Compare INTERNAL RESISTANCE.

internal input impedance The impedance in a circuit or device, as "seen" from the input terminals. Compare INTERNAL OUTPUT IMPEDANCE.

internal input resistance The resistance in a circuit or device, as "seen" from the input terminals. Compare INTERNAL OUTPUT RESISTANCE.

internal noise Electrical noise generated within a circuit, as opposed to that picked up from outside. Such noise comes from transistors, diodes, integrated circuits, resistors, and any other component through which current flows.

internal output impedance The impedance in a circuit or device, as "seen" from the output terminals. Compare INTERNAL INPUT IMPEDANCE.

internal output resistance The resistance in a circuit or device, as "seen" from the output terminals. Compare INTERNAL INPUT RESISTANCE.

internal resistance **1.** The resistance of a device, as opposed to added resistance. See, for example, METER RESISTANCE. **2.** In a cell or battery, the equivalent resistance, resulting from imperfect conductivity of the electrolyte and electrodes, which limits the maximum deliverable current.

internal thermal shutdown In an integrated circuit, the junction temperature at which thermal shutdown occurs. It is generally indicated in degrees Celsius (°C).

internal transmittance The ratio of flux reaching the exit surface of a material to the flux leaving the entry surface. Reflection is not taken into account. The sum of internal transmittance and INTERNAL ABSORPTANCE is always equal to 1.

international broadcast station A shortwave broadcast station transmitting programs for international reception between 6 and 26.6 MHz.

international callsign The call letters of a station, assigned within a country, according to the method of arrangement (identifying letters, or letters and numerals) prescribed by the International Telecommunication Union.

international candle See CANDLE.

international coulomb A unit of electrical quantity, equal to 0.99985 absolute coulomb.

international farad A unit of capacitance, equal to 0.99952 absolute farad.

international henry A unit of inductance, equal to 1.00018 absolute henry.

international joule A unit of energy, equal to 1.00018 absolute joule.

International Morse Code See CONTINENTAL CODE.

international ohm A unit of electrical resistance, equal to 1.000495 absolute ohm. The other international units are derived from this value.

International Radio Consultative Committee Abbreviation, CCIR (*Comite Consultatif International Radiodiffusion*). An international organization reporting to the International Telecommunication Union, and studying technical operations and tariffs of radio and television.

International Steam Table calorie A unit of heat energy equal to 4.1868 joules.

International System of Units Abbreviation, SI (*Systeme International d'Unites*). The system of units of measurement established in 1960 under the Treaty of the Meter. The base units are as follows.

–METER (m), length: 1,650,763.73 times the wavelength of the light emitted in a vacuum of krypton 86

–KILOGRAM (kg), mass: the mass of the protype kilogram kept at Sevres, France

–SECOND (s), time: the duration of 9,192,631,770 periods of the radiation that corresponds to the transition between the two hyperfine levels of the ground state of cesium 133

–KELVIN (K), thermodynamic temperature: $\frac{1}{273.16}$ the thermodynamic temperature of the triple point of water

–AMPERE (A), electric current: the current that, flowing through two infinitely long parallel wires in a vacuum and separated by 1 meter, produces a force of 2×10^{-7} newton per meter of length between the wires

–CANDELA (cd), luminous intensity: the luminous intensity of $\frac{1}{600,000}$ square meter of perfectly radiating surface at the temperature of freezing platinum

International Telecommunication Union Abbreviation, ITU. An international, nongovernmental organization devoted to standardizing worldwide communications practices and procedures.

International Telegraph and Telephone Consultative Committee See CCITT.

International Telegraph Consultative Committee See CCIT.

international units A system of electrical units, based on the resistance through a specified quantity and configuration of the element mercury. The INTERNATIONAL OHM forms the basis for the international system of units.

international volt A unit of electrical potential, equal to 1.00033 absolute volt.

international watt A unit of power, equal to 1.00018 absolute watt.

Internet A worldwide, interconnected system of computer networks. Originated in the late 1960s as *ARPAnet* (Advanced Research Projects Agency network). Still used extensively by educators and scientists, but gaining popularity among personal-computer users. Sometimes called the INFORMATION SUPERHIGHWAY.

interphone An intercom aboard a mobile vehicle.

interpolation Finding a value that falls between two values listed in a table, indicated by a dial, plotted on a graph, derived by estimate, or given by intermediate calculation. For example, if a linear variable capacitor has a value of 100 pF when its dial is set to 10, and 140 pF when the dial is set to 20, then the capacitance when the dial reads 15 (midway between 10 and 20) can be assumed to be 120 pF (midway between 100 pF and 140 pF). When functions are not linear, interpolation is usually not exact.

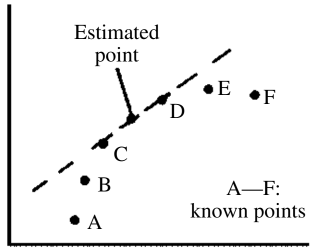

interpolation

interpolation meter See INTERPOLATION-TYPE INSTRUMENT.

interpolation oscillator A frequency-measuring signal generator with a built-in crystal oscillator, whose harmonics provide calibration points on the tuning dial. The dial provides a continuous indication of generator frequencies between crystal-harmonic points. Also see INTERPOLATION-TYPE INSTRUMENT.

interpolation-type instrument An instrument, such as a meter or signal generator, that is used to transfer an accurate quantity point from a standard to another instrument and to provide a range of values between such points. A secondary standard is sometimes used as an interpolation-type instrument (see, for example, INTERPOLATION OSCILLATOR).

interpole motor A direct-current motor with small auxiliary poles (*interpoles*) between its main field poles. The interpoles reduce sparking at the commutator.

interpreter A computer program that can convert instructions given in a high-level language (BASIC, for example) into the machine language that a computer uses; if it is not resident in the computer's nonvolatile memory, it must be loaded each time the machine is activated.

interrecord gap See INTERBLOCK SPACE.

interrupt A break in a computer program, as when a background job is interrupted so that a foreground job can be run. Also see BACKGROUND JOB and FOREGROUND JOB.

interrupted commercial and amateur service See INTERMITTENT COMMERCIAL AND AMATEUR SERVICE.

interrupted continuous wave Abbreviation, ICW. A continuous wave that is interrupted at regular intervals, as in the chopping of a wave at a regular rate. Compare CONTINUOUS WAVE (CW) and MODULATED CONTINUOUS WAVE (MCW).

interrupted dc A direct current or voltage that is periodically started and stopped by switching or chopping.

interrupter contacts Auxiliary contacts operated directly by the armature of a stepping switch.

interruption frequency See QUENCHING FREQUENCY.

interruption-frequency oscillator See QUENCH OSCILLATOR.

interrupt signal The signal that causes a break (INTERRUPT) in a computer program.

intersatellite communication **1.** Communication between or among satellites. **2.** Communication between two earth-based stations, using two or more satellites. **3.** Communication between an earth-based station and a satellite-based station, using at least one intermediate relaying satellite.

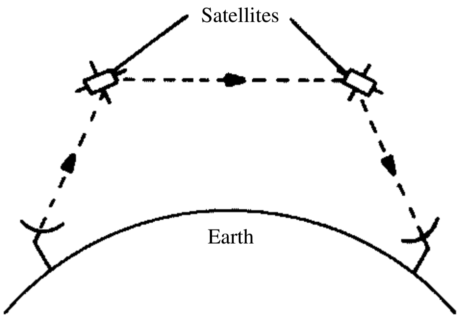

intersatellite communication

intersecting storage ring A device for producing great amounts of energy. It is similar to a vacuum tube. High-speed subatomic particles are fed to a ring-shaped evacuated structure in opposite directions. The two particle beams collide at various points, yielding high energy.

intersection The logical AND operation.

interstage capacitor A coupling capacitor used between two circuit stages.

interstage coupling The transfer of a signal between two circuit stages, such as those of an amplifier. Common forms of interstage coupling include *direct coupling*, *capacitive coupling*, *transformer coupling*, *diode coupling*, and *optoisolator coupling*.

interstage diode A semiconductor coupling diode used between two circuit stages.

interstage transformer A coupling transformer used between two circuit stages. It provides direct-current isolation, and also can match purely resistive impedances.

intersymbol interference In a digital communications signal, a condition in which a given symbol overlaps with one or more other symbols (either immediately preceding it or immediately following it), upsetting the ability of the receiver to decipher signals in certain time intervals. The phenomenon is sometimes a problem in time-division multiplexing—especially at data speeds near the maximum for the system.

intersystem A power-generating network of interconnected separate systems.

intersystem communications Communications between computer systems, either through direct linking of central processors, or by mutual use of peripherals and input/output channels.

intertie See INTERCONNECTION, **2**.

interval **1.** The amount of separation between successive points, events, or quantities. **2.** The continuous range of values between two defined points. **3.** A specific period of time, with defined beginning and ending points.

intervalometer A timing device for operating equipment over a precisely defined time interval.

interval timer A device that provides power to an equipment for a precise interval upon application of a simple initiating signal or action. See also INTERVALOMETER.

interword space The seven-unit interval between words or code groups in telegraphy. Compare INTERCHARACTER SPACE.

intoxication tester See DRUNKOMETER.

intracoding The coding of data using only data that it contains.

Intrafax Western Union's private facsimile system.

Intraframe A method of digital IMAGE COMPRESSION developed by *MPEG (Moving Picture Experts Group)* and *JPEG (Joint Photographic Experts Group)*. It operates by eliminating redundant data within image frames. Compare INTERFRAME.

intrinsic-barrier diode See PIN DIODE.

intrinsic-barrier transistor A bipolar transistor with a layer of intrinsic semiconductor between one of its pn junctions.

intrinsic concentration The number of minority carriers exceeding the normal equilibrium number in a semiconductor.

intrinsic conduction The flow of electron/hole pairs in an intrinsic semiconductor subjected to an electric field.

intrinsic flux A quantity equal to the product of the intrinsic flux density and the cross-sectional area in a magnet.

intrinsic flux density The increased flux density of a magnet in its actual environment, as compared with the flux density resulting from the same magnetizing force in a perfect vacuum.

intrinsic mobility Electron mobility in an intrinsic semiconductor. Also see CARRIER MOBILITY and MOBILITY.

intrinsic *Q* The value of the *Q*, also known as the FIGURE OF MERIT, for an unloaded circuit. This value is generally higher than the value when a load is connected to the circuit.

intrinsic semiconductor A semiconductor whose characteristics are identical to those of a pure crystal of the material. In this condition, the semiconductor is nearly an insulator. Example: highly purified germanium or silicon before n- or p-type impurities have been added. Compare EXTRINSIC SEMICONDUCTOR.

intrusion alarm A set of electronic sensors and associated circuitry composing a system that detects and warns of the presence of unauthorized personnel within a specific region.

intrusion sensor A sensitive pickup (such as a photocell, ultrasonic detector, or capacitive transducer) that responds to a nearby body by delivering an actuating signal to an intrusion alarm.

INV Abbreviation of INVERTER.

inv Abbreviation of INVERSE.

Invar A nickel-steel alloy (36% nickel) having a low temperature coefficient of linear expansion (1 ppm/°C). Invar is used in electronic equipment where mechanical distortion resulting from temperature changes must be negligible, and in magnetostrictive circuits (see MAGNETOSTRICTION).

inverse **1.** Opposite in nature (e.g., an INVERSE CHARACTERISTIC). **2.** Of opposite sign (e.g., a negative current or voltage). **3.** An operation of opposite kind; thus, subtraction is the inverse of addition, and division is the inverse of multiplication.

inverse beta The beta of a transistor operated with the emitter and collector interchanged.

inverse bias See REVERSE BIAS.

inverse characteristics The characteristics of a bipolar transistor when operated with the emitter and collector reversed.

inverse conduction See REVERSE CONDUCTION.

inverse cube law A principle relating the intensity of an effect to the reciprocal of the cube of the distance from the source. The magnetic field around a solenoidal coil of wire obeys this principle.

inverse current See REVERSE CURRENT.

inverse-distance law The inverse-square law applied to the propagation of radio waves, assuming that the waves do not encounter obstacles.

inverse feedback See DEGENERATION.

inverse fourth-power law A rule of propagation for certain complex forms of energy: $I = k/d^4$, where I is the intensity of the field, d is the distance from the source, and k is a constant.

inverse impedances Impedances (Z_1 and Z_2) that are the reciprocal of another impedance (Z_3), satisfying the relationship $Z_1Z_2 = (Z_3)^2$. Also called RECIPROCAL IMPEDANCES.

inverse leakage The flow of a small static reverse current in semiconductor devices.

inverse-parallel circuit See BACK-TO-BACK CIRCUIT and BACK-TO-BACK CONNECTION.

inverse peak voltage See PEAK INVERSE VOLTAGE.

inverse piezoelectric effect Mechanical movement in a piezoelectric material, caused by application of voltage.

inverse resistance See REVERSE RESISTANCE.

inverse resonance See PARALLEL RESONANCE.

inverse-square law The energy or power intensity of a phenomenon is inversely proportional to the square of the distance from the source. This is often applied to quantitative reasoning about radiant energy, electromagnetic energy, and acoustic energy. Thus, if the distance doubles, the energy or power drops to ¼ its previous value.

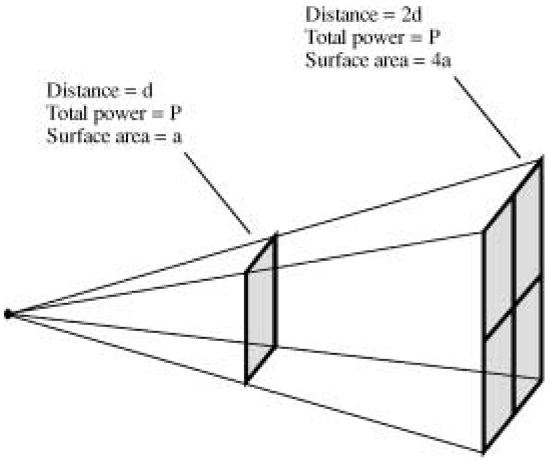

inverse-square law

inverse trigonometric function An angle expressed in terms of a given trigonometric function, followed by the exponent -1 or preceded by "arc" (sin^{-1} is the equivalent of *arcsin*).

inverse voltage **1.** The negative voltage at the anode of a rectifier during the negative half-cycle of alternating-current (ac) input. **2.** The voltage across a power-supply filter capacitor during the negative half-cycle of ac input. **3.** Semiconductor-junction reverse voltage.

inverse Wiedeman effect See DIRECT WIEDEMAN EFFECT.

inversion **1.** A reversal of the normal vertical temperature gradient of the atmosphere, often resulting in long-distance tropospheric radio-wave propagation. **2.** Speech scrambling (see SCRAMBLER CIRCUIT). **3.** Phase inversion (see PHASE INVERTER). **4.** Changing direct current into alternating current, often increasing the voltage (see INVERTER, **1**).

inverted amplifier A push-pull, grounded-gate, field-effect-transistor (FET) amplifier.

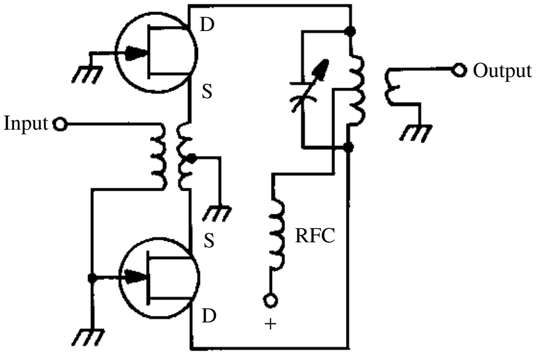

inverted amplifier

inverted-L antenna An antenna having a horizontal radiator and a vertical feeder or lead-in attached to one end of the radiator. The entire arrangement resembles an upside-down L. The overall length is generally ¼ to ½ wavelength.

inverted speech See SCRAMBLED SPEECH.

inverter **1.** Also called *power inverter.* A device that converts direct current (dc) into alternating current (ac), often of a much higher voltage (e.g., 12 Vdc into 117 Vac). **2.** A logic circuit that provides an output pulse that is a negation of the input pulse. Also called a COMPLEMENTER or a NOT CIRCUIT. **3.** See PHASE INVERTER.

inverting adder An analog adder circuit that is provided with an amplifier for a 180° phase shift.

inverting amplifier An amplifier providing a 180° phase shift between input and output.

inverting connection Connection to the inverting input terminals of a differential amplifier or operational amplifier. Also see INVERTING INPUT. Compare NONINVERTING INPUT.

inverting input In a differential amplifier or operational amplifier, the input circuit that produces a phase reversal between the input and output. Compare NONINVERTING INPUT.

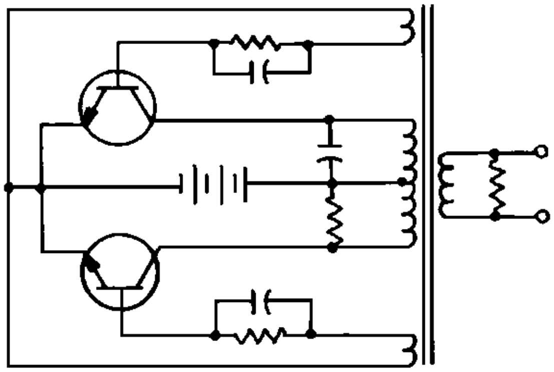

inverter, 1

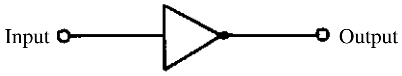

In	Out
0	1
1	0

inverter, 2

inverting transponder In a communications satellite, a transponder in which the downlink band is "upside-down" in frequency relative to the uplink band. That is, the highest downlink frequency corresponds to the lowest uplink frequency, and the lowest downlink frequency corresponds to the highest uplink frequency. Compare NONINVERTING TRANSPONDER. See also DOWNLINK, TRANSPONDER, UPLINK.

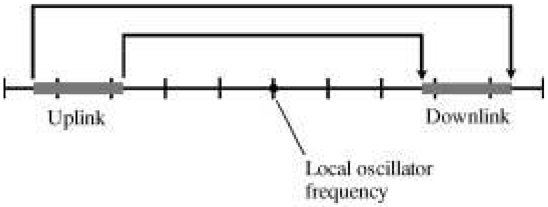

inverting transponder

invisible failure In a computer system, a hardware or software failure whose effect on the system is unnoticeable in a given application. A failure that is invisible in one application might be vividly apparent in some other application.

invister A unipolar semiconductor material, capable of operation at very high frequencies.

involution Raising a number to a power: squaring, cubing, etc. Compare EVOLUTION.

inward-outward dialing Also called *direct dialing.* In a telephone system, a method of dialing in which calls can be made to and from branch exchanges, without operator assistance.

I/O Abbreviation of *input/output* (see INPUT/ OUTPUT EQUIPMENT).

Io Symbol for OUTPUT CURRENT.

iodine Symbol, I. A nonmetallic element of the halogen family. Atomic number, 53. Atomic weight, 126.905. Also see HALOGEN.

ion A charged atom [i.e., one that has gained one or more electrons (a negative ion, or *anion*) or lost one or more electrons (a positive ion, or *cation*)].

ion burn A spot burned on the screen of a cathode-ray tube by negative ions from the cathode striking a single point on the faceplate with high intensity for long periods.

ion concentration **1.** The number of ions, expressed as a percentage or as a number per unit volume, in a substance. **2.** Ionization density in the atmosphere.

ion exchange resins Granular resins that contain acid or base groups, and that trade ions with salts in solutions. The resins play a part in the purification of water for various industrial processes.

ionic binding forces In a crystal, the binding forces that occur when valence electrons of one atom are joined to those of a neighboring atom whose outer shell they fill.

ionic bond In a solid, a bond between atoms, formed as a result of the attraction between oppositely charged atoms (ions).

ionic conduction Conduction, as in a gas or electrolyte, by ion migration (positive to the cathode, negative to the anode).

ionic crystal A crystal whose lattice is held together by the electric forces between ions. Also see IONIC BINDING FORCES and IONIC BOND.

ionic current Current caused by ion movement in a gas or liquid. Also see ION, IONIZATION, and IONIC CONDUCTION.

ionic semiconductor A semiconductor in which the carrier is an ion, as opposed to an electron or hole.

ionic switch See ELECTROCHEMICAL SWITCH.

ionization **1.** The loss or gain of one or more electrons by an atom. Also see ANION, CATION, and ION. **2.** The formation or existence of significant numbers of ions in a gas, liquid, or solid (e.g., ionization of the upper atmosphere).

ionization arc The electrical discharge resulting from the ionization of a material because of high voltage.

ionization chamber An enclosure containing a gas and a pair of electrodes between which a high voltage is applied. Radiation, such as X rays or radioactive particles, passing through the walls of the chamber ionize the gas, creating an ionization current that is proportional to the intensity of the radiation.

ionization current **1.** Current in an ionized gas (such as air). **2.** Current flowing in an electrolyte. **3.** Current in an ionization chamber, Geiger-Mueller tube, or similar gaseous device. **4.** In a gas tube, current flowing after the ignition potential has been reached. **5.** Negative grid current resulting from gassiness in a vacuum tube.

ionization density The extent to which ionization exists in an ionized layer of the atmosphere. The higher the ionization density, the greater the effect on radio waves—especially at frequencies below about 150 MHz.

ionization gauge A form of vacuum tube that can be used to measure the hardness of a vacuum. It consists of a cathode, an anode (plate), and a positively charged grid. Plate current flows as a result of ionization of the atoms within the tube. The more nearly perfect the vacuum, the lower the plate current.

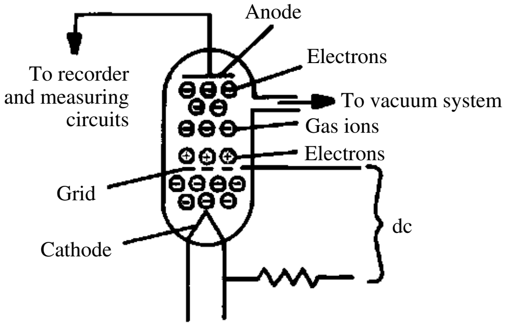

ionization gauge

ionization potential The voltage at which a substance (especially a gas) ionizes. Also called (for a gas) *ignition potential* (see BREAKDOWN VOLTAGE).

ionization pressure In an ionized gas, the pressure increase resulting from the ionization, as compared with the same volume and mass of gas when not ionized.

ionization resistance See CORONA RESISTANCE.

ionization smoke detector A device that senses the presence of smoke or other particles as a result of changes in the IONIZATION CURRENT through the air. When the ionization current suddenly changes, a signal is sent to an alarm circuit.

ionization time The interval, usually measured in microseconds or milliseconds, between the instant that an ionizing potential is applied to a gas, and the instant at which the gas begins to ionize. Compare DEIONIZATION TIME.

ionization trail The ionized path of a meteor as it passes through the upper atmosphere.

ionize To cause the electrons in a substance, particularly a gas, to move freely from atom to atom.

ionized gas A gas whose atoms, under the influence of a strong electric field or IONIZING RADIATION, have become positive or negative ions.

ionized layer See KENNELLY-HEAVISIDE LAYER.

ionized liquid See ELECTROLYTE.

ionizing radiation **1.** Any high-energy electromagnetic radiation that causes ionization in a gas through which the field passes. Examples: ultraviolet, X rays, and gamma rays. **2.** High-speed atomic nuclei (e.g., protons or alpha particles).

ion migration The movement of ions through a solid, liquid, or gas because of the influence of an electric field.

ionosphere **1.** Any of several ionized regions at specific altitudes above the earth's surface. These layers cause absorption and refraction of electromagnetic (EM) fields at some radio frequencies. The D layer exists at an altitude of about 30 miles (50 km) and is ordinarily present only on the daylight side of the planet. This layer does not contribute to, and in fact sometimes hinders, wireless communications. The E layer, about 50 miles (80 km) above the surface, also exists mainly during the day, but nighttime ionization is sometimes observed. The E layer can facilitate medium-range radio communication at certain frequencies. The uppermost regions are the F1 and F2 layers. The F1 layer, normally present only on the daylight side of the earth, forms at about 125 miles (200 km) altitude; the F2 layer exists at about 180 miles (300 km) more or less around the clock. Sometimes the distinction between the F1 and F2 layers is ignored, and they are spoken of together as the F layer.

ionospheric disturbance See IONOSPHERIC STORM.

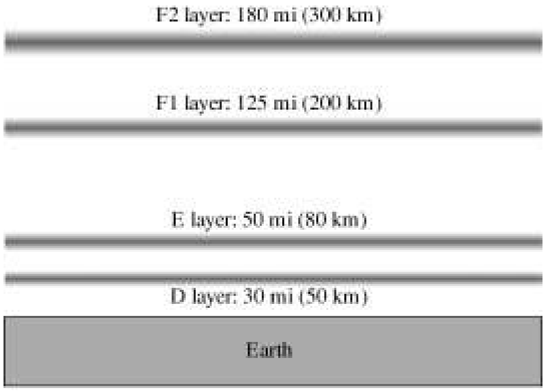

ionosphere

ionospheric forecasting Predicting ionospheric conditions. Radio propagation data is derived from such predictions.

ionospheric layers The respective layers of the ionosphere: *D layer* (at an altitude of about 30 mi or 50 km), *E layer* (at an altitude of about 50 mi or 80 km), the *F1 layer* (at an altitude of about 125 mi or 200 km), and the *F2 layer* (at an altitude of about 180 mi or 300 km). The F1 layer is generally present only in the daytime.

ionospheric propagation Propagation of radio waves by means of reflection or refraction by the ionosphere. Also see HOP; INCIDENT WAVE; IONOSPHERE; MULTIHOP PROPAGATION; REFLECTED WAVE, **1**; REFRACTED WAVE; and SKYWAVE.

ionospheric storm Turbulence in the ionosphere, usually accompanied by a magnetic storm and caused by high-speed particles emitted from an eruption on the sun.

ion sensor A device whose operation is based on the detection of ions and the delivery of a proportionate voltage. Examples are the *Geiger counter*, *halogen gas leak detector*, *mass spectrometer*, and *vacuum gauge*.

ion spot See ION BURN.

ion trap See BENT-GUN CRT.

ion-trap magnet An external (usually double) magnet used with a television picture tube to deflect the ion beam away from the screen. This prevents ION BURN.

I/O port That part of a computer providing, via a connector, a point through which data can enter from, or exit to, peripheral equipment.

I_P **1.** Abbreviation of PLATE CURRENT. **2.** Abbreviation of PEAK CURRENT.

I-phase carrier In color television, a carrier separated by 57 degrees from the color subcarrier.

I-picture A video image that is coded using only data that it contains.

ipm Abbreviation of INCHES PER MINUTE.

ips Abbreviation of INCHES PER SECOND.

IR **1.** The product of current and resistance (see, for example, IR DROP). **2.** Abbreviation of INSULATION RESISTANCE. **3.** Abbreviation of INFRARED.

Ir Symbol for IRIDIUM.

I_r Symbol for CURRENT in a resistor.

IRAC Abbreviation for *Interdepartment Radio Advisory Committee* (a federal government group in the United States).

IR diode See INFRARED DIODE.

IR drop The voltage drop (*E*) across a resistance (*R*) when there is a current (*I*) through the resistor; according to Ohm's law, $E = IR$.

IRE Abbreviation for *Institute of Radio Engineers*, the predecessor of the IEEE (*Institute of Electrical and Electronics Engineers*).

IRED Abbreviation for INFRARED-EMITTING DIODE.

I_{RF} Symbol for RADIO-FREQUENCY CURRENT.

iridescence A sparkling, colorful appearance in a material, resulting from refraction, internal reflection, and interference in light waves passing through the substance. It is especially noticeable in quartz and certain gems.

iridium Symbol, Ir. A metallic element of the platinum group. Atomic number, 77. Atomic weight, 192.22.

iron Symbol, Fe. A ferromagnetic, metallic element. Atomic number, 26. Atomic weight, 55.847. Iron (and its special form, steel) is widely used in magnetic circuits.

iron/constantan thermocouple A thermocouple consisting of a junction between wires or strips of IRON and CONSTANTAN.

iron core A transformer or choke core made from iron or steel. The core is usually laminated to reduce eddy-current loss.

iron-core coil An inductor having an iron or steel core, usually laminated to reduce eddy-current loss.

iron-core IF transformer An intermediate-frequency transformer having a core of powdered iron, a form of iron that has the advantage of high permeability while greatly minimizing eddy currents.

iron-core transformer A transformer whose coils are wound on a core of laminated iron or steel.

iron loss Power lost in the iron cores of transformers, inductors, and electrical machinery as a result of eddy currents and hysteresis.

iron magnet A permanent magnet consisting of magnetized iron or a mixture of iron and nickel.

iron oxide A compound of iron and oxygen, whose most familiar form is common rust. The several variants have characteristics that depend on the number of iron and oxygen atoms in the iron-oxide molecule. See, for example, MAGNETITE and RED OXIDE OF IRON.

iron pyrites Formula, FeS_2. Natural iron sulfide that occurs as bright yellow crystals in its natural state.

iron-vane meter An alternating-current meter whose movable element, a soft iron vane, carries the pointer and pivots near a similar, stationary vane. The vanes are mounted in a multiturn coil of wire. The current flows through the coil, the resulting magnetic field magnetizing the vanes. Because the magnetic poles of the vanes are identical, they repel each other; the movable vane is deflected (against the torque of returning springs) over an arc proportional to the current, carrying the pointer over the scale.

irradiance The amount of radiant flux impinging on a unit surface area; it is generally specified in watts per square meter (W/m^2).

irradiation **1.** Exposure of a device to radioactivity or X rays. **2.** The total radiant power density that is incident upon a receiving surface.

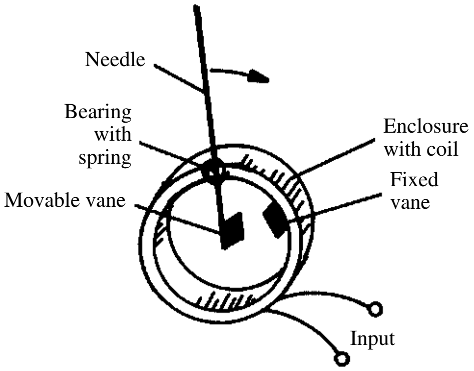

iron-vane meter

irrational number A number that cannot be expressed as the quotient of two integers. Its decimal expansion is nonterminating and nonrepeating.

irregularity **1.** The condition of being nonuniform, or rapidly fluctuating, rather than constant. **2.** A departure from normal operating conditions. **3.** Nonuniformity in a surface. **4.** Nonuniform distribution of matter. **5.** Nonuniform distribution of data.

irregular wave A wave disturbance, or the plot of such a disturbance versus time, that has a complex, periodic, repeating nature.

IR viewer A device that allows observation of images at infrared wavelengths. See SNIPERSCOPE and SNOOPERSCOPE.

I_s **1.** Symbol for source current in a field-effect transistor. **2.** Symbol for *screen current* in a vacuum tube.

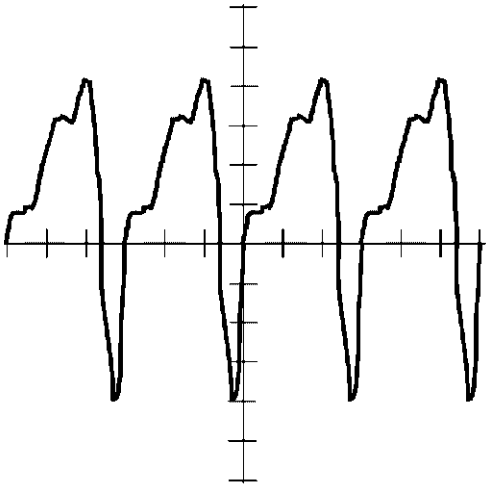

irregular wave

ISDN Abbreviation for INTEGRATED SERVICES DIGITAL NETWORK.

ISCAN Abbreviation of *inertialess steerable communications antenna.*

I-scan A radar display in which the target is shown as a complete circle, whose radius is proportional to the distance to the target.

I-signal With the Z-signal, one of the two signals that modulates the chrominance subcarrier in color television. The I-signal results from mixing a B-Y signal (with -0.27 polarity) and an R-Y signal (with +0.74 polarity).

Isinglass Thinly laminated mica.

ISO Abbreviation for *International Standards Organization.*

ISO 9660 A standard format for producing CD-ROM (COMPACT DISK READ-ONLY MEMORY) mass storage media for use with computers. It is a part of the YELLOW BOOK scheme.

isobar **1.** An atom whose nucleus has the same weight as that of another atom but differs in atomic number. **2.** On a weather map, a line connecting points of equal pressure. Also see BAR, **1**.

isochromal phenomena **1.** Effects occurring at regular time intervals. **2.** Effects of equal duration.

isochromatic Also *orthochromatic.* **1.** The quality of having or producing natural visible-light hues. **2.** Color sensitivity excluding a response to red.

isochronal See ISOCHRONE.

isochrone On a map, a line connecting points of constant time difference in radio-signal reception. It is useful in radiolocation and radionavigation.

isochronous Having identical resonant frequencies or wavelengths.

isoclinic line See ACLINIC LINE.

isodose Pertaining to points receiving identical dosage of radiation.

isodynamic line On a map of the geomagnetic field (the earth's magnetic field), a line connecting points of equal flux density.

isoelectric Having a potential difference of zero.

isoelectronic Having the same number of electrons.

isogonal **1.** See ISOGONIC LINE. **2.** Having uniform magnetic declination at all points.

isogonic line On a map of the geomagnetic field (the earth's magnetic field), a line connecting points of equal magnetic declination.

isolantite An insulating ceramic. Dielectric constant, 6.1.

isolated **1.** Electrically insulated. **2.** Separated in such a way that interaction does not take place.

isolated input **1.** An ungrounded input. **2.** An input circuit with a blocking capacitor to prevent the passage of direct current.

isolated location In a computer, a storage location that is hardware-protected from being addressed by a user's program.

isolating amplifier See BUFFER, **1**.

isolating capacitor A series capacitor inserted in a circuit to pass an alternating-current signal while blocking direct current. Also called a BLOCKING CAPACITOR.

isolating diode A diode used (because of its unidirectional conduction) to pass signals in one direction, but block them in the other direction.

isolating resistor A high-value resistor connected in series with the input circuit of a voltmeter or oscilloscope to protect the instrument from stray pickup. In most voltmeters, this resistor is built into the probe.

isolating transformer A power transformer, usually having a 1:1 turns ratio, for isolating equipment from direct connection to the power line.

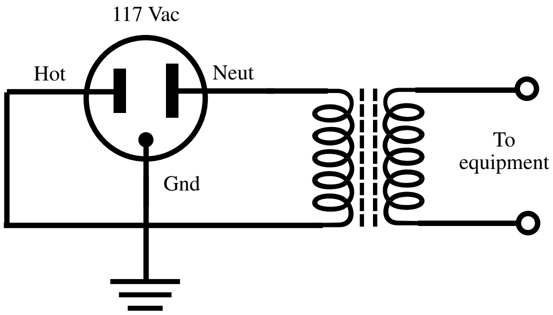

isolating transformer

isolation The arrangement or operation of a circuit so that signals in one portion are not transferred to (nor affect) another portion.

isolation amplifier See ISOLATING AMPLIFIER.

isolation capacitor See ISOLATING CAPACITOR.

isolation diode **1.** In an integrated circuit, a reverse-biased diode that is formed in the substrate to prevent cross-coupling and grounds. **2.** See ISOLATING DIODE.

isolation resistor See ISOLATING RESISTOR.

isolation transformer See ISOLATING TRANSFORMER.

isolator See OPTOELECTRONIC COUPLER.

isolith A form of monolithic integrated circuit, in which the semiconductor is removed in certain places for the purpose of isolating different parts of the circuit.

isomagnetic Having equal magnetic intensity.

isomer A material that has the same atomic number or chemical formula as some other substance, but, because of a difference in the atomic structure, is an entirely different substance. An example is carbon; it can be either *graphite* (by far the more common form) or *diamond.*

isophote On a graph of visible-light intensity, a curve joining points of equal brightness.

isoplanar An integrated-circuit configuration in which insulating barriers or metal oxides are fabricated among the bipolar elements.

isothermal process A physical or chemical process in which there is no temperature change as other factors vary. Compare ENDOTHERMIC REACTION and EXOTHERMIC REACTION.

isotope An atom having the same number of protons as another atom, thereby composing the same chemical element, but having a different number of neutrons. Thus, *deuterium* is an isotope of *hydrogen.* Some isotopes are radioactive [e.g., *carbon 14* (an isotope of the more-common *carbon 12*)]. The two extra neutrons in carbon 14 make it less stable than carbon 12.

isotropic antenna A theoretically ideal antenna that transmits and/or receives electromagnetic fields equally well, and with 100-percent efficiency, in all directions in three-dimensional free space.

isotropic radiator A theoretically ideal radiating element that transmits electromagnetic fields equally well, and with 100-percent efficiency, in all directions in three-dimensional free space.

I_{sup} Symbol for *suppressor current.*

I(t) Symbol for INDICIAL RESPONSE.

ITD Abbreviation for INITIAL TIME DELAY.

item **1.** Component. **2.** Any one of a number of similar or identical components, circuits, or systems.

iteration Repeating a series of arithmetic operations to arrive at a solution to a problem. Computers are commonly programmed to do this thousands, millions, or billions of times. Such a program must include a statement of acceptable accuracy so that it knows when to leave the iteration loop.

iterative impedance In a network consisting of identical, cascaded sections, the input impedance of a section to which the output impedance of the preceding section is made equal.

iterative routine A program or subroutine that provides a solution to a problem by iteration.

iterative transfer constant Symbol, *P.* A property of ITERATIVE IMPEDANCE networks. If I_i is the network input current and I_o is the network output current, then $P = \log_e(I_o/I_i)$.

ITU Abbreviation for *International Telecommunication Union.*

ITV **1.** Abbreviation of INDUSTRIAL TELEVISION. **2.** Abbreviation for INTERACTIVE TELEVISION.

I-type semiconductor See INTRINSIC SEMICONDUCTOR.

I_X Symbol for *current in a reactance.*

I_Y Symbol for *current in an admittance.*

I_Z Symbol for *current in an impedance.*

-ize A suffix used, with some liberty, to form verbs from nouns. In electronics, this commonly refers to procedures or processes (e.g., to ANODIZE, ELECTROCIZE, PLASTICIZE, or TRANSISTORIZE).

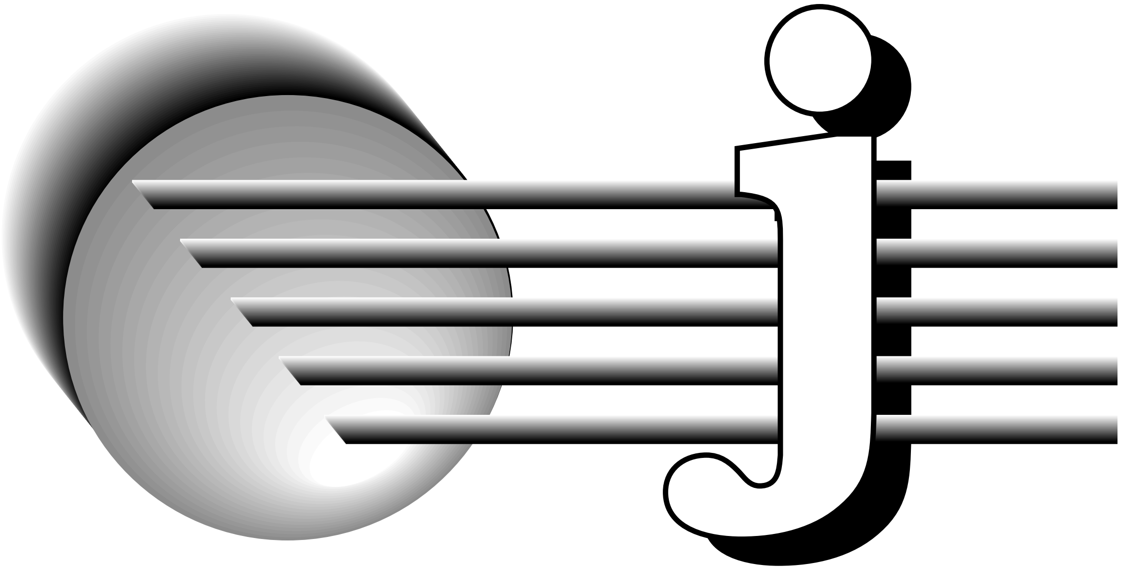

J **1.** Abbreviation for JOULE. **2.** Symbol for JACK or CONNECTOR. **3.** Symbol for EMISSIVE POWER.

j **(*j* operator)** The square root of –1; an imaginary number (usually denoted *i* in mathematics). Assigned to reactance values depicted on the vertical axis of the resistance-reactance (RX) plane in impedance vector diagrams, and whose currents are 90 degrees out of phase with the current in the resistive part of an alternating-current circuit.

jack A receptacle for a plug. A plug (a male connector) is inserted into a jack (a female connector) to complete a circuit or removed from it to break a circuit.

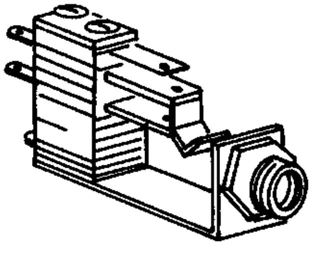

jack

jack box A (usually metallic) box or can used to hold, shield, or protect a jack or group of jacks.

jacket **1.** The outer covering on a cable, as opposed to the insulation or dielectric separating the individual conductors within the cable. **2.** An insulating outer case or wrapper on a component, such as a capacitor. **3.** A shield can or shield box. **4.** A heat-radiating or water-conducting enclosure used in cooling a power vacuum tube.

jack panel A (usually metallic) panel in which a number of jacks are mounted, usually in some order or sequence as denoted by labels.

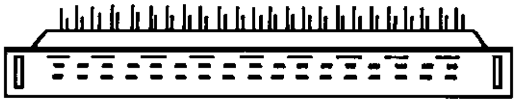

jack panel

jackscrew In a two-piece connector, a screw for mating or separating the halves of the connector.

Jacob's law A principle concerning the behavior of motors. An electric motor develops maximum power when $E_i = 2E_{bk}$, where E_i is the applied voltage and E_{bk} is the back voltage.

JAES Abbreviation for *Journal of the Audio Engineering Society.*

jaff Colloquial term for radar jamming that combines electronic and chaff techniques.

jag Distortion caused by temporary loss of synchronization between the scanner and recorder in a facsimile system.

jam input **1.** A means of setting a logic line to the desired condition by directly applying the desired high or low voltage. **2.** A voltage applied to a logic line to force it high or low.

jammer **1.** A radio transmitter or station used for the purpose of JAMMING communications be-

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

tween or among other stations. **2.** A radio operator who engages in the practice of deliberately JAMMING communications between or among other stations.

jamming The deliberate use of countermeasures, such as malicious transmission of interfering signals, to obstruct communications.

jamming effectiveness The extent to which JAMMING is able to disrupt a service. It can be expressed quantitatively as the ratio of jamming signal voltage to jammed signal voltage. It can also be determined according to the percentage of data that is effectively obliterated.

JAN Abbreviation of JOINT ARMY-NAVY.

janet A system for point-to-point communication via meteor-trail forward scatter. It is generally used at very high frequencies (VHF).

Jansky noise Wideband, high-frequency electromagnetic noise generated by objects in interstellar and intergalactic space.

J antenna An end-fed half-wave antenna having a quarter-wave, parallel-wire matching section. The entire antenna, when oriented vertically, resembles the letter J.

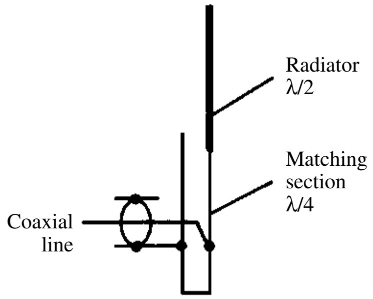

J antenna

Janus antenna array (from *Janus*, an ancient Roman god.) A Doppler-navigation antenna array radiating forward and backward beams.

jar **1.** (From *Leyden jar*) An obsolete unit of capacitance equal to 1/900 microfarad. **2.** The container for the elements of a storage cell.

JASA Abbreviation for *Journal of the Acoustical Society of America.*

J-carrier system In carrier-current (wired/wireless) telephony, a broadband system that provides 12 telephone channels at frequencies up to 140 kHz.

JCET Abbreviation of *Joint Council on Educational Television.*

JCL Abbreviation of *job control language.*

J-display A radar display having a circular time base. The transmitted pulse and reflected (target) pulse are spaced around the circumference; distances can be measured circumferentially between them.

JEDEC Acronym for *Joint Electron Device Engineering Council.*

jerk The rate of change of acceleration; the third derivative of displacement.

JETEC Acronym for *Joint Electron Tube Engineering Council.*

jewel bearing A low-friction bearing used in electric meters and other sensitive devices. It takes its name from a jewel pivot (such as a sapphire) in the groove of which rides the pointed end of a rotating shaft. Also called *jeweled bearing* or *jewel.*

jezebel A passive sonobuoy used in military applications. It detects enemy submarine noises, and transmits them by radio to a monitoring station.

JFET Abbreviation of *junction field-effect transistor.*

JHG Abbreviation of JOULE HEAT GRADIENT.

jig A device constructed especially for the purpose of holding an equipment or circuit board during its repair.

jitter A (usually small and rapid) fluctuation in a phenomenon, such as a quantity or wave, because of noise, mechanical vibration, interfering signals, or similar internal or external disturbances. It is used especially in reference to cathode-ray-tube (CRT) displays.

J/K Abbreviation for *joule(s) per Kelvin*, the SI unit of entropy; also the unit for the Boltzmann constant.

J-K flip-flop A transistor-resistor flip-flop stage producing an output signal even when both inputs are in the logic 1 state (high). It is so called because its input terminals are labeled J and K.

J/(kg·K) Abbreviation for *joule(s) per kilogram Kelvin*, the SI unit of specific heat capacity.

job A unit of computer work, usually consisting of several program runs.

job control language An operating system language used to describe the control requirements for jobs within the system.

job control program A program that uses control language statements and implements them as instructions controlling a job in an operating system.

job control, stacked See SEQUENTIAL-STACKED JOB CONTROL.

job flow control To control the order of jobs being processed by a computer to make the most efficient use of peripherals and central processor time—either manually or by an operating system.

job library A series of related sets of data that will be loaded for a given job.

job-oriented terminal A terminal that produces data in computer-ready form.

job statement A control statement identifying the beginning of a series of job control statements for a job.

job step The execution of a computer program according to a job control statement; several job steps can be specified by a job.

job stream In a processing system, a group of consecutively run jobs.

jogging Rapid, repetitive switching of power to a motor to advance its shaft by small amounts. Also called *inching.*

Johnson counter See RING COUNTER.

Johnson curve A spectral curve (important in appraising solar cell performance) for air mass zero (i.e., for conditions beyond the earth's atmosphere).

Johnson-Lark-Horowitz effect The resistivity gained by a metal or degenerate semiconductor (one in which conduction is nearly equal to that of a simple metal) because of electron scattering by impurity atoms.

Johnson noise See THERMAL NOISE.

Johnson-*Q* feed system See *Q*-ANTENNA.

join Also called *disjunction.* The logical inclusive-OR operation.

joined actuator A form of multiple circuit breaker in which the opening of one circuit results in the opening of all circuits.

joint See JUNCTION, **1**.

joint circuit A communications circuit shared by two or more services.

joint communications Communication facilities being used by more than one service of the same country.

joint denial The logical NOR (NOT-OR) operation.

joint-force sensor A feedback servo device that prevents a robot joint from exerting excessive force. It works by detecting the mechanical resistance the robot arm encounters. If the resistance becomes too great, the joint force is reduced or removed.

jolt **1.** Colloquialism for KILOVOLT. **2.** Colloquialism for ELECTRIC SHOCK. **3.** Colloquialism for TRANSIENT. **4.** Colloquialism for a lightning discharge.

Joly transformer A frequency-tripling transformer whose frequency-multiplying action depends on the nonlinearity of the magnetic induction curve of the core material.

Jones plug A special form of polarized receptacle having numerous contacts.

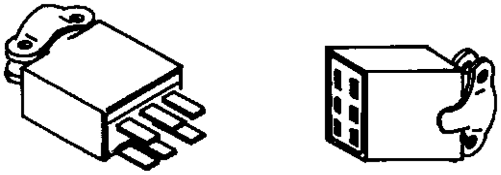

Jones plug

***j* operator** See *J.*

Josephson effect The phenomenon, predicted by Brian Josephson, wherein a current flows across the gap between the tips of two superconductors brought close together; a high-frequency wave is generated.

Joshi effect The phenomenon whereby current in a gas changes as the result of irradiation by light.

joule Abbreviation, J. The SI unit of work. One joule is the work performed when the point of application of one newton is moved one meter in the direction of the applied force. Also see NEWTON.

Joule calorimeter A heat-measuring device that operates electrically.

Joule constant See MECHANICAL EQUIVALENT OF HEAT.

Joule effect **1.** The heat resulting from current flowing through a resistance. **2.** See MAGNETOSTRICTION.

joule heat See JOULE EFFECT, **1**.

joule heat gradient The rate of change in the temperature of a resistive object through which a current flows.

joule meter An integrating wattmeter producing readings in joules.

Joule's law The rate at which heat is produced by current flowing in a constant-resistance circuit is proportional to the square of the current.

journal A file of messages within an operating system, providing information for restarts and historical analysis of system functioning.

joystick A two- or three-dimensional potentiometer with a movable lever, allowing control of a parameter, according to the position of the lever in the up/down and left/right directions. It is often used in computer games for the purpose of manipulating images on a screen. In some such devices, the lever can be rotated clockwise or counterclockwise to obtain additional functions.

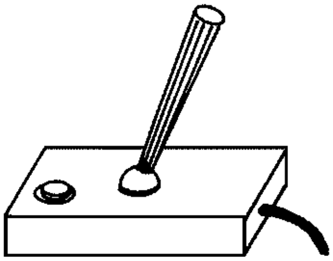

joystick

JPEG Abbreviation of *Joint Photographic Experts Group.*

JPEG image compression An image-compression scheme that eliminates redundancies, greatly reducing the necessary digital storage space. The process is fast enough to allow animated graphics at moderate speed.

JPL Abbreviation of *Jet Propulsion Laboratory.*

J-pole antenna See J ANTENNA.

J rule During transitions of orbital electrons from higher to lower energy states (accompanied by the emission of photons), changes in the inner quantum number can only be by a factor of 0 or +1, or −1.

J scan See J DISPLAY.

J scope A radar device that produces a J DISPLAY.

JSR Abbreviation of *jump to subroutine.*

JTAC Abbreviation of *Joint Technical Advisory Committee.*

judder In facsimile transmission, distortion caused by movements of the transmission or reception equipment.

juice Colloquialism for electric voltage or current.

jukebox An automatic phonograph (usually found in public places) that contains a large assortment of records.

Juliet Phonetic representation for the letter J, used in voice communications.

jump **1.** To purposely provide a short circuit around some component or circuit. **2.** Also called *branch.* In digital computer operations, a programming instruction specifying the memory location of the next instruction and directing the computer to it.

jump, conditional See CONDITIONAL BRANCH.

jump, unconditional See UNCONDITIONAL BRANCH INSTRUCTION.

jumper A short piece of wire (usually flexible, insulated, and equipped with clips) for jumping a component or circuit. See JUMP, **1**.

jump instruction See BRANCH INSTRUCTION.

junction **1.** A joint (connection) between two conductors. **2.** The region of contact between semiconductor materials of opposite type (e.g., *pn junction*). **3.** A waveguide fitting used to attach a branch waveguide to a main waveguide at an angle.

junction barrier See DEPLETION REGION.

junction battery A nuclear battery in which a silicon pn junction is irradiated by strontium 90.

junction box A (usually metal) protective box or can into which several conductors are brought together and connected.

junction capacitance In a semiconductor pn junction, the internal capacitance across the junction; it is of special interest when the junction is reverse-biased. Also called *barrier capacitance.*

junction capacitor See VOLTAGE-VARIABLE CAPACITOR.

junction diode A semiconductor diode created by joining an n-type region and p-type region as a wafer of semiconductor material, such as germanium or silicon.

junction field-effect transistor Abbreviation, JFET. A field-effect transistor in which the gate electrode consists of a pn junction.

junction filter A combination of separate low- and high-pass filters having a common input, but separate outputs. The filter is used to separate two frequency bands and transmit them to different circuits.

junction laser See LASER DIODE.

junction light source See LIGHT-EMITTING DIODE.

junction loss **1.** The loss that occurs in a telephone circuit at connecting points. **2.** Loss in a semiconductor pn junction. **3.** Loss occurring at an electrical connection because of poor bonding.

junction photocell A photoconductive or photovoltaic cell that is essentially a light-sensitive JUNCTION DIODE.

junction point **1.** A point at which two or more conductors, components, or circuits join in an electrical connection. **2.** The point in a computer routine at which one of several choices is made.

junction rectifier A semiconductor rectifier that is, in effect, a heavy-duty junction diode, or the equivalent of several heavy-duty junction diodes in combination.

junction station A microwave relay station joining one or more microwave radio legs to the main route or through route.

junction transistor A transistor in which the emitter and collector consists of junctions (see JUNCTION, **2**) between p and n semiconductor regions. Compare POINT-CONTACT TRANSISTOR.

junctor In a crossbar system, a circuit that bridges frames of a switching unit and terminates in a switching device on each frame.

Jungian-world theory The theory that societies keep repeating the same mistakes, generation after generation. Of interest to some researchers in artificial intelligence. It has been suggested that supercomputers might help find solutions to recurring social problems.

justification **1.** The alignment of text along the left margin, the right margin, or both margins. **2.** A method of altering the speed of a digital signal so that it can be received by equipment designed for a different data speed. The rate of speed is lower after this process is applied. Also called *pulse stuffing.*

justify **1.** To adjust the printing of words for aligned left and/or right margins. **2.** In computer operations, to shift an item in a register so that the most- or least-significant digit is at the corresponding end of the register.

just-operate value The current or voltage level at which a relay or similar device just closes. Also called *just-close value.*

just-release value The current or voltage level at which a relay or similar device just opens. Also called *just-open value.*

just scale A musical scale of three consecutive triads, the highest note of each being the lowest note of the next. Each triad has the ratio 4:5:6 or 10:12:15.

jute Tar-saturated fiber, such as hemp, used as a protective covering for cable.

jute-protected cable A cable whose outer covering is a wrapping of JUTE or similar material.

juxtaposed elements Components placed or mounted side by side.

juxtaposed images Images (e.g., those on the screen of a dual-beam cathode-ray tube) that are close to each other for simultaneous viewing, but do not overlap at any point.

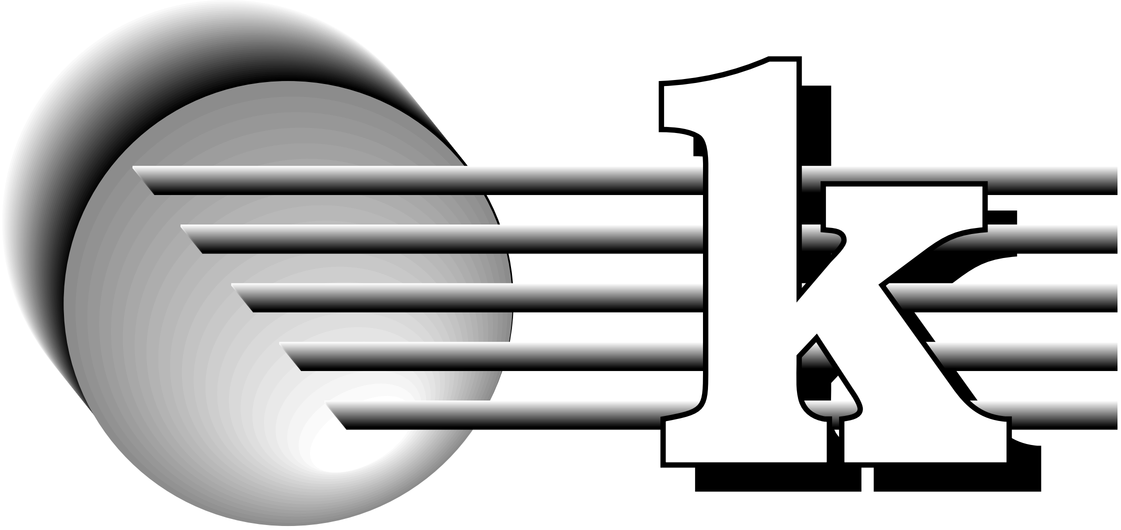

K **1.** General symbol for CONSTANT. **2.** Symbol for POTASSIUM. **3.** Symbol for KELVIN. **4.** Radiotelegraph symbol for *go ahead* or *over.* **5.** Symbol for CATHODE. **6.** Symbol for KILOBYTE. **7.** Abbreviation of KILOHM (k is preferred). **8.** Abbreviation of KILO-.

k **1.** Abbreviation of KILO-. **2.** Symbol for the BOLTZMANN CONSTANT. **3.** General symbol for a mathematical constant. **4.** Symbol for DIELECTRIC CONSTANT. **5.** Abbreviation for KILOHM.

kA Abbreviation of KILOAMPERE.

Karnaugh map A logic chart showing switching-function relationships and used in computer logic analysis to determine quickly the simplest form of logic circuit for a given function. The Karnaugh map is sometimes regarded as a tabular form of the Venn diagram.

Kansas City standard A frequency-shift modulation standard for a computer/tape-recorder interface. Also called *byte standard.*

KB **1.** Abbreviation of KEYBOARD. **2.** Abbreviation of KILOBYTE.

K band The 18- to 27-GHz band of frequencies, pertaining to their use in radar applications.

kcal Abbreviation of KILOCALORIE.

K carrier system A four-wire carrier-current telephone system using frequencies up to 60 kHz and providing 12 channels.

kCi Abbreviation of KILOCURIE.

kcs Abbreviation of *kilocharacter per second* (1000 characters per second).

K display See K SCAN.

KDP Abbreviation of POTASSIUM DIHYDROGEN PHOSPHATE, a ferrolelectric material.

keeper A small iron bar placed across the poles of a permanent magnet to forestall demagnetization.

K electron In certain atoms, one of the electrons whose orbit is nearest the nucleus.

Kel-f Abbreviation of *polymonochlorotrifluorethylene,* a high-temperature insulating material.

Kellie bond **1.** The junction of two mated electrically conductive surfaces that are held together by an adhesive that exhibits negligible resistance thermally and electrically when set. **2.** To make a thermally conductive joint, as from a heatsink to a chassis or component.

kelvin (Lord Kelvin, 1824–1907) Symbol, K. The SI unit of thermodynamic temperature; 1 K = 1/273.16, the thermodynamic temperature of the triple point of water.

Kelvin absolute electrometer An electrostatic voltmeter consisting of a movable metal plate (or plates) and a stationary metal plate (or plates)

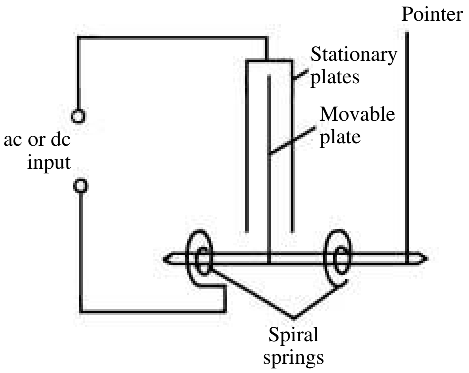

Kelvin absolute electrometer

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

between which voltage is applied. The movable plate(s) is displaced over a distance that is proportional to the potential, against the torque of a return spring.

Kelvin balance An apparatus for measuring current in terms of magnetic pull. A coil is attached to each end of the beam of a balance, and coils ride directly above two stationary coils. Current flows through all coils, making one pair attractive and the other repellent, thus unbalancing the beam. Balance is restored by sliding a weight whose position along a graduated scale indicates the current strength.

Kelvin contacts Electrical contacts designed to eliminate the effect of lead resistance on the accuracy of measurement. Two leads run to each test point, one lead carrying the test signal and the other leading directly to the measuring instrument.

Kelvin double bridge A special bridge for measuring very low resistance (0.1 ohm or less). The arrangement of the bridge reduces the effects of contact resistance that causes significant error when such low resistances are connected to conventional resistance bridges.

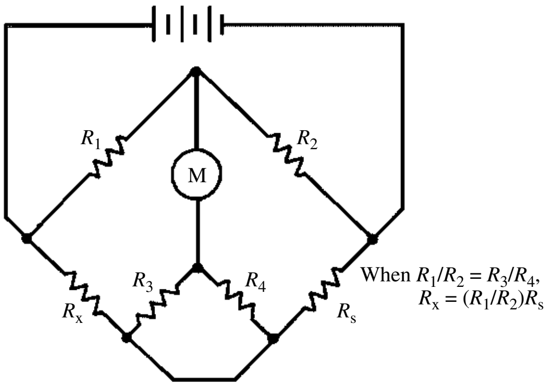

Kelvin double bridge

Kelvin replenisher A static generator consisting of two pairs of concentric, rotating semicircular conductors connected to brushes. The machine can be regarded as a rotating electric doubler.

Kelvin scale See ABSOLUTE SCALE.

Kelvin temperature scale See ABSOLUTE SCALE.

Kelvin voltmeter An electrostatic voltmeter in which an assembly of figure-8-shaped metal plates rotates between the plates of a stationary assembly when a voltage is applied between the assemblies. The length of the arc of rotation is proportional to the electrostatic attraction and, thus, to the applied voltage.

Kendall effect Distortion in a facsimile record caused by unwanted modulation produced by a carrier signal.

Kennelly-Heaviside layer (A.E. Kennelly, 1861–1939; Oliver Heaviside, 1850–1925) An early name for ionized regions in the upper atmosphere. These regions reflect and refract radio waves at certain frequencies. There are several layers at various altitudes. Also see IONOSPHERE and IONOSPHERIC LAYERS.

keraunograph A meteorological instrument for detecting distant electrical storms. In its simplest form, it consists of a galvanometer connected in series with an antenna and ground.

keraunophone A radio-receiving KERAUNOGRAPH.

kernel Inside an electrical conductor, a line along which the magnetic field strength is zero. Generally, this line is near the center of the conductor.

Kerr cell A nitrobenzene-filled cell that makes use of the KERR ELECTRO-OPTICAL EFFECT. It can function as an electric light shutter or control.

Kerr electro-optical effect The tendency of certain dielectric materials to become double-refracting in an electric field.

Kerr magneto-optical effect The tendency of glass and some other solids and liquids to become double-refracting in a magnetic field.

Kerst induction accelerator See BETATRON.

keV Abbreviation of KILOELECTRONVOLT.

Kew magnetometer A special magnetometer used to measure the intensity of the earth's magnetic field, and also the magnetic declination at a given point on the earth's surface. The device is designed for very high accuracy, using magnifying lenses.

key **1.** See DECRYPTION KEY. **2.** A projection or pin that guides the insertion of a plug-in component into a holder or socket. **3.** A digit or digits used to locate or identify a computer record (but not necessarily part of the record). **4.** A specialized hand-operated switch, used to make and break a circuit repetitively to form the dot and dash signals of Morse code telegraphy. Primarily of historical interest. Used by some amateur radio operators for hobby purposes. **5.** Slang for *principal* or *main.*

keyboard An array of lettered or numbered, low-torque push buttons, used to enter information into a computer, telegraph, teletypewriter, or automatic control system.

keyboard computer A digital computer in which the input device is an electrical keyboard of the typewriter or calculator type.

keyboard entry The operation of a keyboard to enter information into a computer for processing.

keyboard keyer A device for automatically sending Morse code using a typewriter-like keyboard, rather than a paddle or straight key. Each key on the keyboard, when pressed, produces the complete character and a space following it. Most keyboard keyers have buffers to allow typing well ahead of the code being sent, with insertion of all the correct spaces. The speed range is usually from about 5 words per minute (wpm) to 60 or 70 wpm, although some keyboard keyers are programmed for speeds over 100 wpm.

keyboard lockout A keyboard interlock in a data transmission circuit that prevents data from being transmitted while the transmitter of another station on the same circuit is operating.

keyboard send-receive unit A teletypewriter lacking an automatic input device.

key cabinet In a telephone system, a facility that shows a subscriber which lines are busy and which lines are open.

key chirp A chirping sound in a received signal, resulting from slight frequency shift when a radiotelegraph transmitter is keyed. It does not occur with a well-designed transmitter.

key-click filter An inductance-capacitance (LC) or resistance-capacitance (RC) filter for smoothing a keying wave to eliminate KEY CLICKS. It functions by optimizing the rise and decay times of the keyed waveform.

key clicks Excessive bandwidth of a radiotelegraph signal that can result when a keyed signal has rise and decay times that are too rapid. Produces characteristic clicking or popping sounds, with resulting interference, in receivers tuned to frequencies near that of the transmitted signal. A KEY-CLICK FILTER can eliminate this.

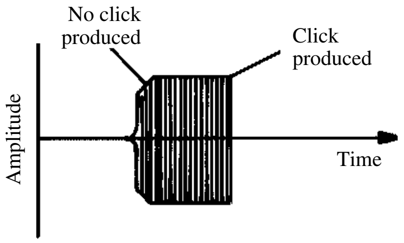

key clicks

keyed AGC A controlled automatic gain control (AGC) system in a television receiver circuit. The AGC acts when the horizontal sync pulse appears; it is inactive between pulses. This prevents unwanted actuation of the AGC by noise transients and picture-signal elements.

keyed clamp A clamping circuit that uses a control signal to determine the clamping time.

keyed interval In a transmission system that is keyed periodically, an interval beginning with a change in state and having a duration of the shortest time between changes in state.

keyed rainbow generator For testing of color-television receivers, a signal generator that produces a rainbow color pattern on the screen (i.e., a set of 10 vertical color bars representing the spectrum, with blank bars in between). The pattern results from gating the 3.56-MHz oscillator in the receiver at a frequency of 189 kHz.

keyer An automatic device for keying a radiotelegraph transmitter or wire telegraph circuit. The keyer can operate from perforated tape, an embossed disk, magnetic tape, or other similar recording.

keyer adaptor A modulated-signal detector that produces a direct-current signal having an amplitude sympathetic with the modulation; it provides the keying signal for a frequency-shift exciter in radio facsimile transmission.

key escrow A controversial system in which the government is provided with certain components of decryption keys to all communications ciphers, according to laws that would allow the government to eavesdrop on private communications or transactions after getting a court order. The key components are held in a secure place, that is, in "escrow," unless and until the necessary court order is obtained. See DECRYPTION KEY.

keying **1.** The modulation of a carrier by switching it on and off. It is commonly used in radiotelegraphy. **2.** The modulation of a carrier by switching its frequency between two defined values. It is also called FREQUENCY-SHIFT KEYING (FSK) and it is used in data transmission. **3.** The modulation of a carrier with an audio tone that is switched on and off. The carrier can be modulated via amplitude modulation (AM), frequency modulation (FM), pulse modulation, or any other form that will convey the audio tone. It is occasionally used in radiotelegraphy at very-high frequencies (VHF) and above. Also called *audio keying.* **4.** The modulation of a carrier with an audio tone whose frequency is switched between two defined values. The carrier can be modulated via amplitude modulation (AM), frequency modulation (FM), pulse modulation, or any other form that will convey the audio tone. It is commonly used in data transmission at VHF and above. Also called AUDIO FREQUENCY-SHIFT KEYING (AFSK).

keying chirp A rapid change in the frequency of a continuous-wave signal, occurring at the beginning of each code element. In the receiver, the resulting sound is a chirp.

keying error rate In data transmission, the ratio of incorrectly keyed signals to the total number of signals keyed.

keying filter See KEY-CLICK FILTER.

keying frequency **1.** In audio-keyed radiotelegraphy, the audio frequency (tone) of the dot and dash signals (as opposed to the carrier frequency). **2.** In radiotelegraphy, the transmission speed (see KEYING SPEED). **3.** The number of times per second that a black-line signal occurs while an object is scanned in a facsimile system.

keying monitor A simple detector used by an operator to listen to the keying of a radiotelegraph transmitter.

keying speed The speed (in words per minute) of a telegraph or radio-telegraph transmission.

keying transients **1.** Transients arising from the keying of a radiotelegraph transmitter or wire telegraph circuit. **2.** Transients that arise from the repetitive making and breaking of any circuit.

keyless ringing In a telephone system, ringing that begins as soon as the calling plug is put in the appropriate jack on the jack panel.

key pulse In telephone operations, a signaling system in which the desired numbers are entered by pressing corresponding pushbuttons or keys.

key punch A keyboard-operated machine for recording information by perforating a tape or cards.

keyshelf A shelf that supports manually operated telephone switchboard keys.

key station The master (control) station in a communications or control network.

keystoning A form of video image distortion in which the top of the picture is wider than the bottom, or vice versa. Thus, the image area is shaped like a trapezoid, rather than a rectangle.

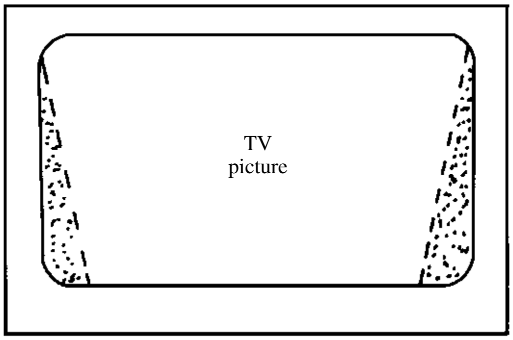

keystoning

key switch **1.** A lockable switch that is operated by inserting and turning a key in it. **2.** A switch having a long handle that transmits motion to the mechanism through a cam. **3.** See KEY, **1**. **4.** The separate short-circuiting switch sometimes mounted on the base of a telegraph key (see KEY, **1**). **5.** A switch that actuates the keys of an electronic organ.

key-to-disk unit A keyboard-to-magnetic-disk data-processing unit.

keyway A groove or slot into which a mating key slides to position a plug-in component (see KEY, **2**).

keyword In information retrieval systems, the significant word in the title describing a document (e.g., in the title "A Primer on French Cuisine," the word *cuisine* would be the keyword, the others having no singular significance).

kg Abbreviation of KILOGRAM.

kgC Abbreviation of KILOGRAM-CALORIE.

kgm Abbreviation of KILOGRAM-METER.

kg/m^3 Symbol for KILOGRAMS PER CUBIC METER, the SI unit of density.

K-H layer See KENNELLY-HEAVISIDE LAYER.

kHz Abbreviation of KILOHERTZ.

kick **1.** To place into sudden operation, as by the quick, forcible closure of a switch or the rapid application of an enabling pulse. **2.** See TRIGGER. **3.** Colloquialism for an abrupt, momentary electric shock.

kickback **1.** The counter EMF that appears across an inductor when current is interrupted. **2.** See FLYBACK.

kickback power supply A high-voltage power supply using the flyback principle. See FLYBACK.

kidney joint A waveguide coupling used in radar. The joint is flexible, or can consist of an air gap, to allow rotation of the antenna.

Kikuchi lines A characteristic spectral pattern produced by the electrons scattered when an electron beam strikes a crystal.

killer **1.** A pulse or other signal used to disable a circuit temporarily (e.g., a blanking pulse). **2.** In personal computing, an application of such importance that it alone serves as the motivation for someone to purchase or upgrade a system. Common examples are word processing and online communications.

killer circuit **1.** A circuit that disables some function of a system, such as the audio in a television receiver. **2.** The blanking circuit in a radar receiver. **3.** A circuit that prevents responses to side-lobe signals in a repeater or transponder.

Kilo Pronunciation, *KEY-low.* Phonetic alphabet code word for the letter K.

kilo- Abbreviation, K or k. **1.** A prefix meaning thousand(s). **2.** In digital data applications, a prefix meaning 2^{10} (1024).

kiloampere Abbreviation, kA. A unit of current equal to 1000 amperes.

kilobit A unit of digital data equal to 2^{10} (1024) bits. Also see BIT.

kilobyte Abbreviation, K or KB. A unit of digital data equal to 2^{10} (1024) bytes. Also see BYTE.

kilocalorie Abbreviation, kcal. A large unit of heat; 1 kcal equals 1000 calories. See CALORIE.

kilocurie Abbreviation, kCi. A large unit of radioactivity equal to 3.71×10^{13} disintegrations per second; 1 kCi equals 1000 curies. Also see CURIE, MEGACURIE, MICROCURIE, and MILLICURIE.

kilocycle See KILOHERTZ.

kiloelectronvolt Abbreviation, keV. A large unit of electrical energy equal to 1000 electronvolts. See ELECTRONVOLT.

kilogauss A large unit of magnetic flux density; 1 kilogauss equals 1000 gauss or 0.1 tesla.

kilogram Abbreviation, kg. The SI base unit of mass; it is equal to 1000 grams.

kilogram-calorie The heat required to raise 1 kilogram of water 1°C.

kilogram-meter Abbreviation, kgm. A unit of mechanical energy (work); 1 kgm is the energy required to raise a mass of 1 kilogram vertically by a distance of 1 meter (equal to 7.2334 foot-pounds). Also see JOULE.

kilohertz Abbreviation, kHz. A unit of frequency; 1 kHz equals 1000 Hz.

kilohm Symbol, kΩ. A unit of high resistance, reactance, or impedance; 1 kΩ equals 1000 ohms.

kilojoule Abbreviation, kJ. A unit of energy or work; 1 kJ equals 1000 joules. See JOULE.

kilolumen Abbreviation, klm. A unit of luminous flux equal to 1000 lumens. See LUMEN.

kilomega- See GIGA-.

kilomegahertz See GIGAHERTZ.

kilometer Abbreviation, km. A large metric unit of linear measure; 1 km equals 1000 meters (3280.8 feet).

kilo-oersted Abbreviation, kOe. A unit of magnetic field strength; 1 kOe equals 1000 oersteds. See OERSTED.

kiloroentgen Abbreviation, kr. A large unit of radioactive radiation; 1 kr equals 1000 roentgens. See ROENTGEN.

kilorutherford Abbreviation, krd. A large unit of radioactivity equal to 10^9 disintegrations per second.

kilovar A compound term coined from *kilo-* and *VAR* (the abbreviation of *volt-amperes reactive*). It is equal to a reactive power of 1000 watts.

kilovar-hour A large unit of reactive electrical energy, equivalent to 1000 reactive watts manifested for a period of one hour.

kilovolt Abbreviation, kV. A unit of high voltage; 1 kV equals 1000 V.

kilovolt-ampere Abbreviation, kVA. A unit of high power that gives the TRUE POWER in a direct-current circuit and the APPARENT POWER in an alternating-current circuit; 1 kVA equals 1000 W. Also see DC POWER.

kilovolt-ampere reactive See KILOVAR.

kilovoltmeter A voltmeter designed to measure thousands of volts (kilovolts).

kilowatt Abbreviation, kW. A unit of high power; 1 kW equals 1000 watts. Also see WATT.

kilowatt-hour Abbreviation, kWh. A common unit of electrical energy; 1 kWh equals 1000 watt hours, or the equivalent of 1000 watts dissipated for a period of one hour. Also see ENERGY, KILOWATT-HOUR, POWER, WATT-HOUR, and WATT-SECOND.

kilowatt-hour meter A motorized meter for recording (electrical) power consumption in kilowatt-hours. Also see KILOWATT-HOUR.

kinematograph A motion picture camera. Also called CINEMATOGRAPH and KINETOGRAPH.

kine **1.** See KINESCOPE, **1**. **2.** See KINESCOPE RECORDING.

kinescope **1.** The picture tube in a television receiver. **2.** See KINESCOPE RECORDING.

kinescope recorder A film or tape apparatus for recording television pictures.

kinescope recording A motion-picture or videotape made from the screen (or taken from the circuit) of a television picture tube.

kinetic energy The energy associated with particles, bodies, or electric charge carriers in motion.

kinetograph See KINEMATOGRAPH.

kinetoscope A motion-picture projector.

kiosk A computer and peripherals set up for the purpose of multimedia use by the general public. It generally uses a touch screen for inputting data and must be ruggedly constructed to tolerate rough treatment.

Kirchhoff's first law The sum of the currents flowing out of a point in a direct-current circuit equals the sum of the currents flowing into that point.

Kirchhoff's laws (Gustav Robert Kirchhoff, 1824–1887) Two laws of electric circuits that account for the behavior of certain networks. See KIRCHHOFF'S FIRST LAW and KIRCHHOFF'S SECOND LAW.

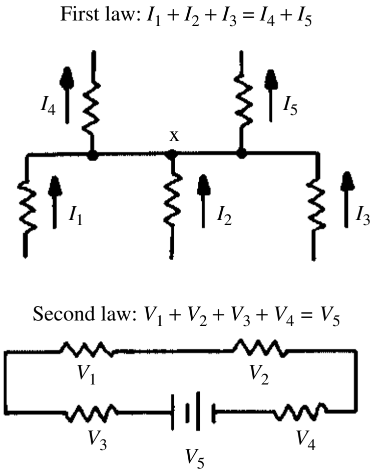
First law: $I_1 + I_2 + I_3 = I_4 + I_5$
I_4
I_5
x
I_1
I_2
I_3
Second law: $V_1 + V_2 + V_3 + V_4 = V_5$
V_1
V_2
V_3
V_5
V_4

Kirchhoff's laws

Kirchhoff's second law The algebraic sum of all the voltage drops around a direct-current circuit (including supply voltages) is always equal to zero.

kit A selection of components, associated equipment, supplies (such as wire and hardware), and instructions for constructing a piece of electronic equipment.

kite-supported antenna A longwire antenna that uses a kite as a support for the far (nonstation) end. A tether is used to reduce the chance that the kite will fly away with wire attached. It was used by Marconi in early experiments with radio. Radio amateurs and shortwave listeners sometimes use this scheme at low and medium frequencies. It is a dangerous antenna because of electrostatic buildup, a tendency to attract lightning, the possibility of its breaking loose, and the risk that it might contact utility lines.

kJ Abbreviation of KILOJOULE.

k-line programming A method by which an artificially intelligent robot can learn as it does a job, so it will have an easier time doing the same job in the future. The robot controller actually learns from the robot's mistakes.

Klipsch horn A loudspeaker that includes a folded low-frequency horn housed in a corner enclosure.

klm Abbreviation of KILOLUMEN.

kludge **1.** A crude, useless, or grossly inefficient machine or process. **2.** A hastily contrived prototype of a circuit or device, put together for the purpose of testing a concept, but not intended as a representative of a production unit.

klydonograph A device that photographically records the voltage gradient in the presence of an electric field.

klystron A microwave tube whose operation is based on the velocity modulation of an electron beam by buncher and cavity reentrant cavities.

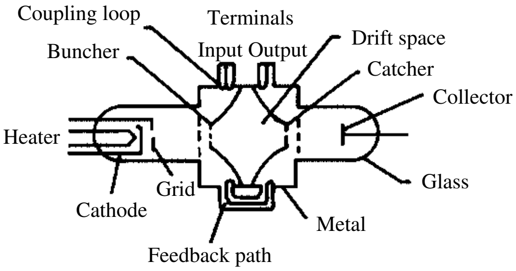

klystron

klystron amplifier A microwave amplifier using a klystron.

klystron harmonic generator A frequency-multiplying power amplifier using a klystron. It is used at microwave radio frequencies.

klystron oscillator A klystron operated as a self-excited microwave oscillator.

klystron repeater A microwave amplifier in which a klystron inserted in a waveguide boosts the amplitude of an incoming signal.

km Abbreviation of KILOMETER.

knee **1.** A sharp bend in a response curve for a device, usually indicating the onset of conduction, saturation, cutoff, pinchoff, or limiting action. It applies especially to semiconductor diodes and transistors. **2.** The characteristics of a device when it is operated at a point in the vicinity of a sharp bend in its response curve.

knee noise Electrical noise generated by rapidly repeating current fluctuations at the knee in a Zener diode.

knife-edge diffraction The lessening of atmospheric signal attenuation when the signal passes over a sharp obstacle and is diffracted.

knife switch A switch composed of one or more flat blades roughly resembling knife blades, which are slid firmly between the jaws of pinching contacts to close a circuit.

knob **1.** A (usually round and insulated) finger dial for adjusting a variable electronic component, such as a potentiometer, variable capacitor, or rotary switch. **2.** A solid round insulator usually having a low diameter to height ratio. **3.** A small ball- or rod-shaped electrode or protuberance.

knocker A fire-control radar subassembly of synchronizing and triggering circuits.

knockout An area in a metal box or chassis that is easily removed by tapping or knocking to provide an opening.

knot A unit of speed, corresponding to 1 nautical mile per hour. A speed of 1 knot is about 1.15 statute miles per hour; a speed of 1 statute mile per hour is about 0.868 knots. It is used by mariners for specifying speeds at sea, and also occasionally by meteorologists in specifying wind speeds.

knowledge The data in a computer and in mass-storage media, accumulated over time and capable of being put to practical use.

kOe Abbreviation of KILO-OERSTED.

Kolster decremeter An absorption wavemeter with a movable scale; it permits measurement of the decrement of a radio wave.

Kooman antenna A unidirectional antenna consisting of stacked, full-wave, center-fed driven elements, and a reflecting screen.

Korner Killer Trade name for an acoustically absorbent object that reduces sound echoes that can occur in enclosed rooms. The name results because the device works best when placed in a corner (where two walls meet).

Kovar An alloy of cobalt, iron, and nickel. It is used mostly in glass-to-metal seals because it has characteristics of both kinds of material.

Kozanowski oscillator A positive-grid vacuum-tube UHF oscillator circuit using two tubes having cylindrical elements, and a pair of parallel-wire tanks.

Kr Symbol for KRYPTON.

kr Abbreviation of KILOROENTGEN.

K radiation X rays emitted from an atom when an electron becomes a K electron.

kraft paper Strong brown paper used for insulation and as the dielectric of paper capacitors.

Kramer system A system of three-phase motor control providing constant horsepower, and having a direct-current (dc) motor coupled to the shaft of a wound-rotor three-phase induction motor. The dc supply for the motor also supplies a rotary converter. The speed-control rheostat is connected in series with the field of the motor and the dc power supply.

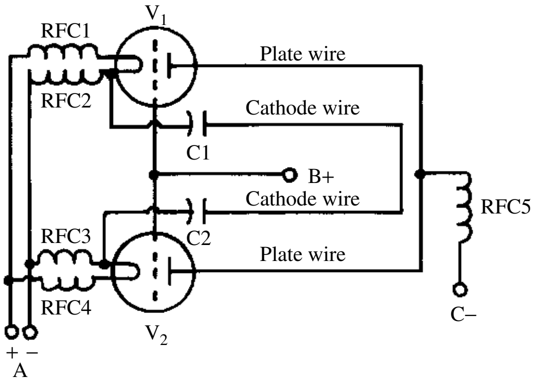

Kozanowski oscillator

Kraus antenna A bidirectional, flat-top beam antenna consisting of a pair of closely spaced dipoles. Several such sections can be connected in series by crisscrossing the wires at voltage loops.

krd Abbreviation of KILORUTHERFORD.

Kryptol A mixture of clay, graphite, and silicon carbide. It is used in electric heater elements because of its low resistance and high melting point.

krypton Symbol, Kr. An inert, gaseous element. Atomic number, 36. Atomic weight, 83.80. Krypton is present in trace amounts in the earth's atmosphere.

K scan In radar operations, a modified A-scan used in aiming antennas in which two pips are displayed; their relative amplitudes indicate the antenna-aiming error.

K series A series of spectral lines for the shortest wavelengths of radiation from the innermost electron shell of a radiating atom.

KSR Abbreviation of *keyboard send-receive* unit.

Ku band A band of microwave frequencies between approximately 12 and 18 GHz.

Kundt's law The index of refraction of a medium does not change continuously with wavelength in the absorption bands.

Kundt tube A device used to measure the speed of sound in gases under various conditions. Suspended particles in the gas form standing waves that can be easily seen. Knowing the frequency of the disturbance and the distance between nodes of the standing waves, the speed can be determined. The pressure and density of the gas, as well as temperature and humidity, affect the speed.

kurchatovium See RUTHERFORDIUM.

kV Abbreviation of KILOVOLT.

kVA Abbreviation of KILOVOLT-AMPERE.

kVAR Abbreviation of REACTIVE KILOVOLT-AMPERE.

kVARh Abbreviation of KILOVAR-HOUR.

kW Abbreviation of KILOWATT.

kWh Abbreviation of KILOWATT-HOUR.

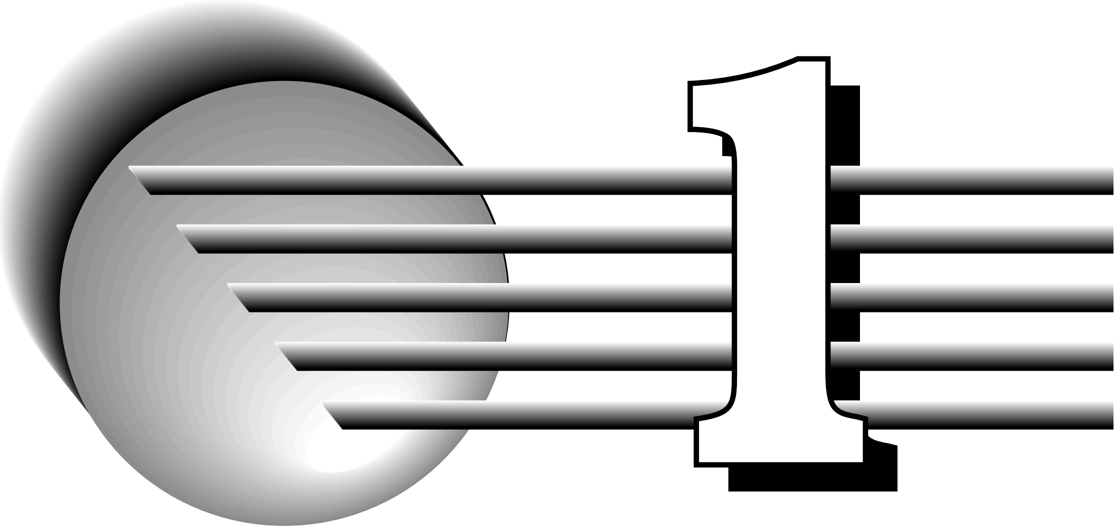

L **1.** Symbol for INDUCTANCE. **2.** Symbol for MEAN LIFE. **3.** Abbreviation of LOW. **4.** Resembling the capital letter L in physical shape. **5.** Symbol for LAPLACE TRANSFORM.

l **1.** Symbol for LENGTH. **2.** Abbreviation of LITER. **3.** Subscript for LOW. **4.** Abbreviation of LUMEN; also abbreviated *lm* (preferred) and *lum.*

La Symbol for LANTHANUM.

label **1.** A symbolic group of characters that identifies an area of memory, an item of data, a file, or a record. **2.** A name assigned to a source program instruction step to identify the step as a coding entry point, or to make the step usable as a reference point for entry to the routine or subroutine in which it appears.

label group A collection of labels, usually of the same type, held in an operating system.

label identifier Within a label, a character set used to name the kind of item labeled.

label record A record identifying a file recorded on a magnetic storage medium (e.g., magnetic tape).

label set A collection of labels having a common label identifier.

labile oscillator A frequency-controlled local oscillator.

laboratory conditions The environmental, mechanical, and electrical parameters characteristic of controlled conditions. Actual operating conditions can be much different.

laboratory-grade instrument An instrument having the high accuracy and stability that suit it to precision measurements in a laboratory. Also called PRECISION INSTRUMENT. Compare SERVICE-TYPE INSTRUMENT.

laboratory power supply A regulated direct-current source whose adjustable output is less than 10 kV at no more than 500 W.

laboratory standard See PRIMARY STANDARD and SECONDARY STANDARD.

labyrinth speaker A loudspeaker whose enclosure (a wooden cubicle) includes a folded pipe or acoustic transmission line (behind the speaker); the inner walls are lined with a sound-absorbent material. When the pipe, which is open-ended, is half as long as the wavelength of the frequency being reproduced, the sound emerging from the open end is in phase with that radiated by the front of the speaker and, therefore, reinforces it.

laced wiring Circuit wiring in which wires or cables run parallel in bundles that are tied together with LACING CORD.

lacing cord Strong, sometimes waxed cord used to tie together wires running parallel in a bundle. Also see LACED WIRING.

lacquer disk See CELLULOSE-NITRATE DISK.

lacquer-film capacitor A fixed capacitor with a plastic film dielectric; the film is applied as liquid lacquer to the metal foil.

lacquer master A master recording made on a CELLULOSE-NITRATE DISK.

lacquer original See LACQUER MASTER.

ladar Abbreviation of LASER DOPPLER RADAR. Also abbreviated *lopplar.*

ladder attenuator See LADDER-TYPE ATTENUATOR.

ladder filter A form of delay line or filter. It generally consists of series and parallel impedances, either in a balanced or unbalanced form.

ladder line See AIR-INSULATED LINE.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

ladder network A network consisting of several L sections in cascade. See L SECTION, **1**, **2**, **3**.

ladder-type attenuator An attenuator consisting of a ladder network equipped with a switching circuit for selecting the output at various sections.

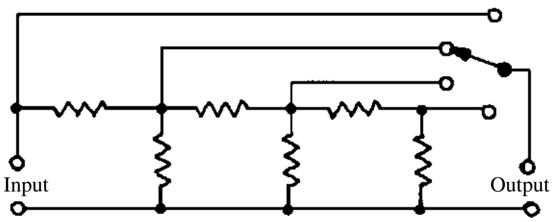

ladder-type attenuator

LAFOT Coded weather broadcasts aired every six hours by the U.S. Weather Bureau through marine radiotelephone broadcasting stations for the Great Lakes region.

lag In computations relating the phase of alternating-current signals, the extent to which one quantity follows another in time (e.g., the current lags the voltage by 90 degrees in a pure inductive reactance). Compare LEAD.

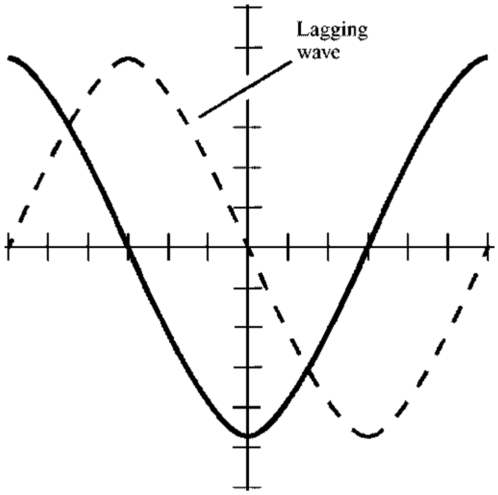

lag

lagged-demand meter A meter with a built-in time delay.

lagging current Current that follows voltage (in time). Also see LAG.

lagging load A load in which current lags behind voltage (i.e., an inductive load). Compare LEADING LOAD.

lag network A phase-shifting circuit containing series-resistance and shunt-capacitance arms. It produces a lagging phase shift. Compare LEAD NETWORK.

lambda wave An electromagnetic disturbance that travels along the surface of an object. An example is the surface wave characteristic of low-frequency propagation.

lambert Symbol, L. The centimeter-gram-second (cgs) unit of luminance, equal to the brightness of an ideal diffusing surface that radiates or reflects light at 1 lumen per square centimeter. The SI (preferred) unit of luminance is the candela per square meter (cd/m^2); 1 lambert equals 10^4 cd/m^2. Also see CANDELA.

Lambert's law of illumination The illumination of a surface by a point light source is inversely proportional to the square of the distance between the surface and the source. If the surface is not perpendicular to the rays, the illumination is proportional to the cosine of the angle of incidence.

laminated armature An armature for a motor or generator, made of stacked laminations.

laminated contact A switch contact consisting of a number of laminations—each contacting a conducting counterpart.

laminated core A core for a transformer, choke, relay, or similar device, made of stacked laminations.

laminated disk A layered recording disk.

laminated pole A pole within a motor, generator, relay, electromagnet, or similar device, made of stacked laminations.

lamination **1.** A relatively thin sheet of metal cut to a required shape to be stacked with other similar sheets to form a laminated core or pole. **2.** A relatively thin sheet of plastic that is bonded together and heat-formed with other similar sheets to produce a sheet or piece of desired thickness and strength.

lamp A device for converting electrical energy into visible light. The term includes a number of devices (e.g., arc lamp, fluorescent tube, incandescent lamp, mercury-vapor lamp, and neon bulb).

lamp-bank resistor A makeshift heavy-duty resistor consisting of several incandescent lamps arranged so that they can be switched in various series, parallel, and series-parallel combinations to vary the resistance provided by the filaments.

lampblack Carbon obtained from soot deposited by a smoky flame. The substance is used as the basic material for some resistors.

lamp cord A two-wire insulated cord, used with low-wattage alternating-current appliances at 117 volts. The wire is usually stranded copper equivalent to American Wire Gauge (AWG) #16.

lamp dimmer See DIMMER.

lamp driver A usually single-stage circuit for amplifying a small pulse to drive an indicator lamp.

lamp extractor A special tool used to insert or extract miniature lamps for electronic equipment.

lamp jack A receptacle with a spring release that holds a small incandescent bulb. The bulb is removed and replaced by pushing and twisting.

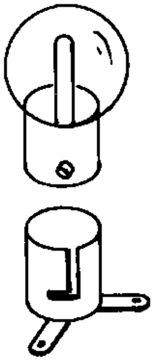

lamp jack

lamp-type expander A volume expander in which the tungsten filament in an incandescent lamp serves as the nonlinear resistor.

lamp-type readout For counters, calculators, and digital meters, a readout device in which each digit is indicated by a lamp.

LAN Abbreviation of LOCAL AREA NETWORK.

land **1.** The flat, reflective surface between pits on a compact disc (CD). Compare PIT, **1**. **2.** The thin vinyl wall between grooves on a phonograph record. **3.** A bonding point in a microcircuit. **4.** Pertaining to earthbound communications stations.

Land camera See POLAROID CAMERA.

landing beacon The aircraft landing-beam transmitter. Also see LANDING BEAM.

landing beam A highly directional airport radio signal beamed upward to guide aircraft landing during conditions of poor visibility.

landline A telephone or telegraph circuit completed with wires.

landmark beacon Any beacon that is not an airway or airport beacon.

land mobile service Two-way radio service between a base station and mobile land vehicles, or among mobile land vehicles.

land mobile station A radio station aboard a mobile, earthbound vehicle.

land return Ground reflection of radar signals back to the transmitter.

land station A fixed ground station.

Langevin ion An electrically charged particle, such as a grain of dust or droplet of water, resulting from the accumulation of ions.

Langmuir dark space In a luminous gas discharge, the dark region around a negatively charged probe inserted into the positive column. Compare CROOKES' DARK SPACE.

Langmuir's law See CHILD'S LAW.

language In digital-computer operations, any one of the detailed systems for representing data, instructions, and procedures through the use of symbols and symbol sequences. See MACHINE LANGUAGE, ASSEMBLY LANGUAGE, COBOL, FORTRAN, and BASIC.

language laboratory An electronic contribution to the teaching and learning of languages. It consists of recordings in a language being studied and all the equipment associated with recording, playback, and monitoring. Students listen to the speech of experts in the language record, listen to, and later erase their own utterances in the language.

language translation **1.** The conversion of statements in one computer language to equivalent statements in another. **2.** The conversion of one written natural language into another (e.g., French to Russian) by means of a computer program.

language translator **1.** An assembly program, compiler, or other routine used for translation between computer languages. **2.** A high-level program that allows a computer to translate one written natural language into another (e.g., Chinese to Italian).

L antenna See INVERTED-L ANTENNA.

lantern battery A moderate-sized electrochemical battery usually rated at 6 volts. Derives its name from its original use as a power source for portable lamps. One type has spring contacts on the top. Another type has thumbscrew terminals. The nonrechargeable battery consists of 4 zinc–carbon or alkaline cells in series. Some varieties are rechargeable, consisting of nickel–cadmium (NiCd) or nickel–metal-hydride (NiMH) cells. This type of battery can provide enough energy to operate a low-power radio transceiver. Two units connected in series, or four units in series-parallel, make a 12-volt battery that can power a small portable Citizen Band (CB) or amateur radio station.

lanthanum Symbol, La. An elemental metal of the rare-earth group. Atomic number, 57. Atomic weight, 138.906.

lanyard A wire or cable used to quickly pull apart the halves of a quick-disconnect connector.

lap A device used for grinding piezoelectric crystals for resonance at a desired frequency.

lap dissolve The simultaneous fading out of one televised scene while another is fading in so that one is apparently dissolving into the next. It is also applicable to motion pictures.

lapel microphone A small microphone that is clipped to a lapel of a user's jacket or coat.

lap joint An overlapping splice of two conductors.

Laplace transform Symbol, *L*. An operator that reduces the work of solving certain differential equations by permitting them to be handled by simpler algebraic methods.

lapping Fine-tuning quartz crystal plates by moving them over a flat plate coated with a liquid abrasive.

laptop computer See NOTEBOOK COMPUTER.

lap winding In a motor or generator armature, a winding in which the opposite ends of each coil are connected to the adjoining segments of the commutator.

lap wrap **1.** A form of asbestos cloth wire insulation. **2.** A method of wrapping with electrical tape, in which there is considerable overlap among the turns of the tape.

large calorie See KILOGRAM-CALORIE.

large loop antenna A single-turn open or closed loop, usually having a circumference of 0.5 wavelength or 1 wavelength. With a half-wavelength open or closed loop, maximum radiation occurs in the plane of the loop. A full-wavelength closed loop exhibits maximum radiation and response along the axis. This type of antenna can be used for wireless transmitting and receiving applications. Either the open or the closed half-wavelength loop exhibits a slight power loss relative to a dipole, but the full-wavelength loop shows a small gain over a dipole in its favored directions. Compare SMALL LOOP ANTENNA.

large-scale integration Abbreviation, LSI. The inclusion of more than 100 transistors, performing various individual, but interrelated circuit functions, on a single integrated-circuit chip.

large signal A relatively high-amplitude signal that traverses so large a part of the operating characteristic of a device that nonlinear portions of the characteristic are usually encountered. Compare SMALL SIGNAL.

large-signal analysis The rigorous study of circuits and devices that process large signals.

large-signal component **1.** A coefficient or parameter such as amplification, transconductance, or dynamic resistance, measured under conditions of large-signal operation. Also see LARGE SIGNAL and LARGE-SIGNAL EQUIVALENT CIRCUIT. **2.** A device designed for operation at high signal levels.

large-signal equivalent circuit For a given transistor circuit, the equivalent circuit at high signal levels (i.e., at amplitudes approaching saturation and cutoff levels). Also see EQUIVALENT CIRCUIT.

large-signal operation The use of a circuit or device at signal levels sufficiently high so that nonlinear portions of the characteristic are usually encountered. Compare SMALL-SIGNAL OPERATION.

large-signal transistor See POWER TRANSISTOR.

large-signal voltage gain In an integrated-circuit amplifier, the voltage gain under open-loop conditions, determined as the difference in the output voltage divided by the difference in the input voltage. It is usually specified in volts per millivolt or volts per microvolt.

Larmor orbit The path followed by a charged particle in a constant magnetic field. Because of interaction between the external field and the field generated by the particle, the charged particle travels in a circular path.

laryngaphone See THROAT MICROPHONE.

LASCR Abbreviation of LIGHT-ACTIVATED SILICON-CONTROLLED RECTIFIER.

LASCS Abbreviation of LIGHT-ACTIVATED SILICON-CONTROLLED SWITCH.

lase To emit coherent electromagnetic energy in the visible-light spectrum. See LASER.

laser Acronym for *light amplification by stimulated emission of radiation.* A device that produces coherent radiation in the visible-light range, that is, between 750 and 390 nanometers (one nanometer is 10^{-9} meter). Some devices that produce coherent radiation in the infrared, ultraviolet, or X-ray parts of the spectrum are also referred to as lasers. Lasers can be either continuous or pulsed, and are characterized by coherent, monochromatic emissions. The peak intensity ranges from a few microwatts to many megawatts.

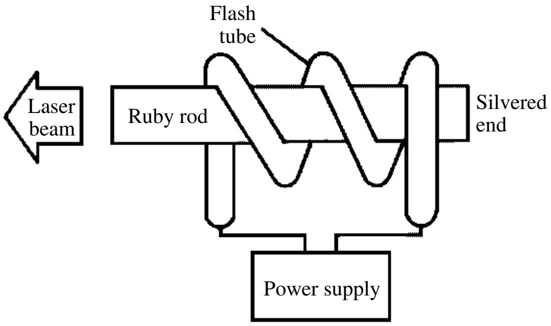

laser

laser beam The radiation from a laser—especially if the divergence is very low, that is, the rays are almost perfectly parallel, resulting in minimal divergence.

laser-beam communication A form of coherent infrared or optical communication in which a laser beam is the link between transmitting and receiving stations. Also see LASER, LASER DIODE, and LIGHT-BEAM COMMUNICATION.

laser capacitor An energy-storage capacitor used to discharge-fire the exciter lamp of a laser. Also see LASER.

laser cavity An optical-resonant cavity that results in the emission of coherent light.

laser cutting A method of using a laser for severing materials.

laser diode A special form of semiconductor light-emitting diode (LED), usually of the gallium-arsenide type, that emits coherent light when a voltage is applied to its terminals. Also see LASER.

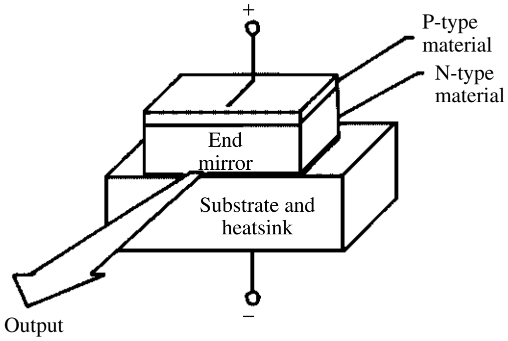

laser diode

laser disk A method of reproducing sound in which a laser is used to recover the sound from a compact disk.

laser Doppler radar Acronyms, *ladar* or *lopplar.* A form of Doppler radar using the light beam of a laser instead of radio waves.

laser eye surgery A method of repairing the retina of the eye without cutting the eyeball, using laser beams to push loose retina tissue back into place.

laser gun A colloquial term for a weapon that makes use of a laser as a device of destruction.

laser optical videodisc system A system in which a low-powered laser reads audio and video information from a videodisc and delivers it to a television receiver.

laser ranger A radar-like device using intense light (instead of microwaves).

laser show A three-dimensional, midair display having motion, made by using lasers in various combinations.

laser surgery The application of a laser in medicine for the purpose of assisting in, or actually performing, operations on human subjects.

laser welding Welding (especially of tiny pieces) with the heat produced by a laser beam.

lasing The emission of coherent electromagnetic energy in the visible-light spectrum. See LASER.

lat Abbreviation of LATITUDE.

latch **1.** A feedback loop in a symmetrical digital circuit, such as a flip-flop, used to maintain a given state. **2.** A simple logic-circuit storage element that consists of two gates as a unit. **3.** To maintain a closed (energized) state in a pair of relay contacts after initial energization from a single electrical pulse. See LATCHING RELAY.

latching current In a thyristor, the minimum value of anode current (slightly higher than the holding current) that will sustain conduction immediately after switch-on.

latching relay An electromechanical or fully electronic relay that locks into whichever mode it is energized for (on or off).

latch-on relay See LOCKING RELAY.

latchup In a transistor switching circuit, the abnormal condition in which the collector voltage remains at its switched-on level after the transistor is switched to cutoff from saturation.

latch voltage The input voltage at which a flip-flop changes states.

late contacts Relay contacts that are operated following the movement of other contacts during the relay's operation.

latency **1.** The time taken by a digital computer to deliver information from its memory. **2.** In a serial storage system, the access time less the word time.

latent image **1.** In a storage tube, a stored image that is not yet visible. **2.** An image stored in the mosaic of an iconoscope. **3.** The image that will appear when photographic film or paper is developed.

lateral chromatic aberration An aberration affecting the sharpness of off-axis color television images.

lateral compliance In phonograph reproduction, the ease with which the stylus can move laterally as it follows the groove. Also see COMPLIANCE and LATERAL RECORDING.

lateral-correction magnet In a color picture tube, a magnet operated with a set of pole pieces attached to the focus element of the blue gun; it controls horizontal positioning of the blue beam for convergence.

lateral magnet See LATERAL-CORRECTION MAGNET.

lateral recording A disc recording in which the groove undulates from side to side. Compare VERTICAL RECORDING.

latitude Abbreviation, lat. Angular distance measured around the earth's circumference to the north and south from the equator. Compare LONGITUDE.

latitude effect The tendency of the earth's magnetic field to decrease the number of charged subatomic particles that reach the surface of the earth near the equator, as compared with the number reaching the surface at other latitudes.

Latour alternator See BETHENOD ALTERNATOR.

lattice **1.** The orderly internal pattern (matrix) of atoms in a crystal. Also see CRYSTAL LATTICE. **2.** A symmetrical arrangement of components in a network (such as an attenuator, a filter, or a bridge circuit).

lattice filter A lattice network having reactance in its arms that makes it a selective circuit.

lattice network See LATTICE, **2**.

lattice section See LATTICE, **2**.

lattice structure See LATTICE, **1**.

lattice winding See UNIVERSAL WINDING.

laue diagram A pattern of spots on a photographic plate produced by the scattering of high energy radiation as it falls on a thin crystal. This pattern is used to determine the nature of the crystal material.

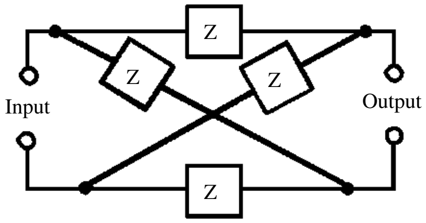

lattice filter

launching The energy transference from a cable into a waveguide.

lavalier microphone A small microphone that can be hung from the user's neck on a cord or chain.

law **1.** A general, verifiable statement that describes the behavior of entities or the relationships between phenomena or concepts. The product of inductive reasoning that follows many observations and controlled experiments (e.g., *first law of thermodynamics*, *inverse-square law*, *Kirchhoff's laws*, and *Ohm's law*. **2.** The nature of the change of a dependent variable, particularly as depicted by a response curve (e.g., *square law*).

LAWEB Civilian weather bulletins issued every six hours from ship and shore positions along the Great Lakes during the sailing season.

lawn mower **1.** A facsimile term for a helix recording mechanism. **2.** A radar receiver preamplifier.

law of a curve See LAW, **2**.

law of averages In probability and statistics, a principle stating that for a large sampling of events, the numerical probability value will be more closely approached than when the sampling is small. Compare LAW OF LARGE NUMBERS.

law of charges Different electric charges attract each other, and similar charges repel each other.

law of electric charges See LAW OF CHARGES.

law of electromagnetic induction See LENZ'S LAW.

law of electrostatic attraction See COULOMB'S LAWS.

law of electrostatic repulsion See COULOMB'S LAWS.

law of first wavefront In acoustics, a rule stating that the perceived direction from which a sound arrives is determined by the first wavefront that reaches the listener's ears.

law of induction See FARADAY'S LAW.

law of inverse squares See INVERSE-SQUARE LAW.

law of large numbers In probability and statistics, a principle stating that with a large sample, the sample average will approximate the population average. It is often erroneously called LAW OF AVERAGES.

law of magnetism Different magnetic poles attract each other, and similar magnetic poles repel each other.

law of natural decay See EXPONENTIAL DECREASE.

law of natural growth See EXPONENTIAL INCREASE.

law of normal distribution Gauss' law of the frequency distribution of a repetitive function, describing the probability of deviations from the mean.

law of octals Chemical activity occurs between two atoms lacking eight valence electrons, and continues until the requirement of eight electrons is satisfied for all but the first orbit, where only two electrons are required. Of interest in the study of semiconductors.

law of radiation See QUANTUM THEORY.

law of reflection For a ray of energy striking a smooth reflective surface, the angle of reflection is equal to the angle of incidence, with respect to a plane tangent to the surface at the point of incidence.

Lawrence accelerator See CYCLOTRON.

lawrencium Symbol, Lr (occasionally Lw). A short-lived radioactive element produced artificially from californium. Atomic number, 103. Atomic weight, about 260.

lay See DIRECTION OF LAY.

layer **1.** A complete coil winding consisting of turns laid side by side (not on top of each other). **2.** In a semiconductor device, a region having unique electrical properties (e.g., *n layer*). **3.** A region of the ionosphere. See IONOSPHERIC LAYERS. **4.** The tape on a reel or in a cassette, encompassing one complete turn (rotation). **5.** In general, a single stratum of a stratified medium.

layer-to-layer transfer In a roll of magnetic tape, unwanted transfer of data between adjacent turns on the reel. If severe, this transfer can cause drop-in or drop-out in a computer. In audio applications, it can sometimes be heard as a delayed echo or a faint sound occurring just prior to the actual recorded sound.

layer winding A coil winding in which the turns are arranged in two or more concentric layers.

layerwound coil An inductor wound in layers, one on top of the other. Also see LAYER, **1**. Compare SINGLE-LAYER COIL.

layout The arrangement of components on a chassis, printed circuit board, or panel.

lazy-H antenna An antenna consisting of two vertically stacked collinear elements, producing both horizontal and vertical directivity.

Lb Abbreviation of LAMBERT; also, *L* (preferred).

lb Abbreviation of POUND.

L band A radio-frequency band extending from 390 MHz to 1.55 GHz. For subdivisions of this band, see L_C BAND, L_F BAND, L_K BAND, L_L BAND, L_P BAND, L_S BAND, L_T BAND, L_X BAND, L_Y BAND, and L_Z BAND.

LC **1.** Abbreviation of LIQUID CRYSTAL; also abbreviated *lix.* **2.** Abbreviation of INDUCTANCE-CAPACITANCE. **3.** Symbol for LC CONSTANT.

L carrier In a telephone system, a carrier having a frequency between approximately 68 kHz and 10 MHz. It can be used in wire-transmission or radio links.

L_C band A section of the L BAND extending from 465 MHz to 510 MHz.

LC bridge See INDUCTANCE-CAPACITANCE BRIDGE.

LC constant Abbreviation, LC. The product of the inductance and capacitance required for resonance at a given frequency.

LCD Abbreviation of LIQUID-CRYSTAL DISPLAY.

LC filter See INDUCTANCE-CAPACITANCE FILTER.

L circuit See L NETWORK.

LC meter See INDUCTANCE-CAPACITANCE METER.

LCR See INDUCTANCE-CAPACITANCE-RESISTANCE.

LC ratio In a tuned circuit, the ratio of inductance to capacitance.

LCR bridge See IMPEDANCE BRIDGE, **1**.

L_d Symbol for DISTRIBUTED INDUCTANCE.

LDF See LOW-FREQUENCY DIRECTION FINDER.

L display Also called *L scan.* A radar display in which the target appears as two horizontal traces, one extending from a vertical timebase to the right, the other to the left.

lead Pronunciation, *leed.* **1.** A conductor (usually a wire) leading to or emerging from a terminal or electrode. **2.** In computations relating phase, the extent to which one quantity precedes another (e.g., current leads voltage by 90 degrees in a pure capacitance). Compare LAG.

lead Pronunciation, *led.* Symbol, Pb. A heavy metallic element. Atomic number, 82. Atomic weight, 207.2. It can be used as a shield against atomic radiation, and has various applications in electronics (e.g., as electrodes in batteries and as a component of solder). See LEAD-ACID BATTERY, FUSE, and SOLDER.

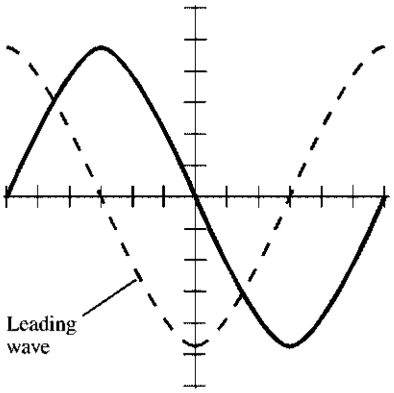

lead, 2.

lead–acid battery A set of two or more lead–acid cells connected in series, usually housed in a common enclosure. Some batteries of this type, notably automotive batteries, are made from sets of lead–acid cells having a free-flowing liquid acid. Other cells have a semisolid "paste" electrolyte. These batteries are popular in consumer electronic devices that require a moderate amount of current. They are also used in uninterruptible power supplies (UPSs) for personal computers. See LEAD–ACID CELL.

lead–acid cell A rechargeable electrochemical cell having an electrolyte of sulfuric acid. The electrodes are lead (negative) and lead dioxide (positive). Produces about 1.5 volts under no-load conditions when fully charged. A large cell of this type can store several tens of ampere-hours. Smaller units have less capacity but more versatility. Their main advantage is reasonable cost, considering that they can be charged and discharged many times.

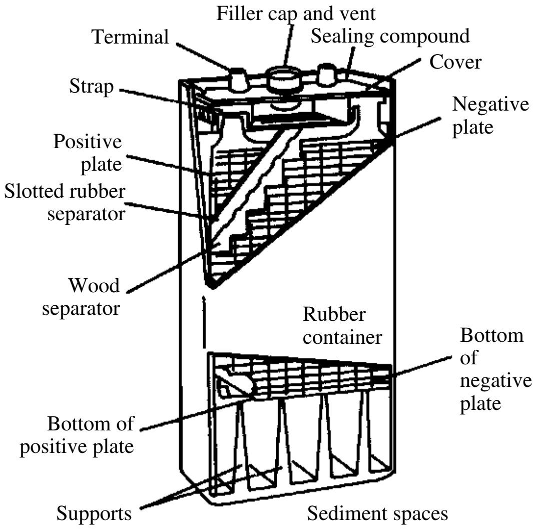

lead-acid cell

lead cell **1.** A lead-acid cell. **2.** A lead-sulfide photocell; see LEAD SULFIDE, **1**.

lead dress See DRESS.

leader **1.** The blank section at the beginning of a magnetic tape. It is usually made of plastic. **2.** A record, preceding a group of records, that identifies the group and provides other data pertinent to the group. **3.** In a lightning stroke, the initial movement of electrons or positive ions, creating the ionized path that allows discharge.

lead frame The metal frame holding the leads of a circuit package (DIP) in place before encapsulation.

lead-in The wire connecting an antenna to a receiver or transmitter.

leading current Current that precedes voltage in time. Also see LEAD.

leading edge The rising edge of a pulse; compare TRAILING EDGE.

leading ghost A twin television image to the left of the original on the screen.

leading load A load in which current leads voltage (i.e., a capacitive load).

lead-in groove Around the outer edge of a phonograph record, a blank spiral groove that leads the stylus into the first groove of the recording. Compare LEAD-OUT GROOVE.

leading whites In a television picture, an abnormal condition in which the leading (left) edge of a black object has a white border.

leading-zero suppression In a digital meter, the blanking out of all zeros to the left of the decimal point.

lead-in spiral See LEAD-IN GROOVE.

lead-in tube A tube of insulating material, such as plastic or ceramic, used to conduct an antenna lead-in through a wall.

lead-in wire **1.** A single wire, used as a feed line for a shortwave receiving antenna. **2.** The feed line for a television receiving antenna. **3.** A feed line for a transmitting antenna.

lead network A phase-shift circuit containing series capacitance and shunt resistance; it produces a leading phase shift. Compare LAG NETWORK.

lead-out groove Around the inner edge of a phonograph record, a blank spiral groove leading into the eccentric or locked groove.

lead-over groove On a phonograph record containing several recorded tracks, a blank groove that conducts the stylus from the end of one recording to the beginning of the next.

lead screw **1.** A threaded rod that guides the cutter across the surface of a disc during its recording. **2.** In facsimile transmission, a threaded shaft that moves the drum or scanning mechanism lengthwise.

lead sulfate Formula, $PbSO_4$. An insulating compound formed in a LEAD-ACID CELL by the chemical action between the lead in the plates and the sulfuric-acid electrolyte. If the sulfate is not broken down during charging of the cell, it will eventually ruin the cell.

lead sulfide Formula, PbS. A compound of lead and sulfur used as the light-sensitive material in some photoconductive cells.

lead zirconate titanate A synthetic piezoelectric material.

leaf electroscope An electroscope using a pair of gold leaves or a single gold leaf and a solid strip of metal.

leak **1.** The loss of energy through a stray path not intended for conduction. **2.** A point from which energy is lost through a stray path not intended for conduction.

leakage **1.** The small current that flows through an electrical insulator. **2.** The electromagnetic field that is radiated or received by a feed line that should theoretically have 100-percent shielding.

leakage current The zero-signal current flowing across a reverse-biased semiconductor junction.

leakage flux Collectively, magnetic lines of flux around a transformer that do not link the primary and secondary coils.

leakage inductance Self-inductance caused by LEAKAGE FLUX. Leakage inductance is effectively in series with the primary or secondary winding of a transformer.

leakage resistance **1.** In an imperfect insulator, the ohmic resistance, calculated by dividing the voltage across the insulator by the current flowing through the insulator. **2.** The quotient of voltage and current in a reverse-biased semiconductor junction.

leakage radiation Radiation from parts of a system, as compared with that from the true radiator.

leakage reactance Inductive reactance caused by leakage inductance in the primary or secondary circuit of a transformer.

leakance The conductance of an insulating material, measured in siemens. It is equal to the reciprocal of the leakage resistance in ohms.

leaky **1.** Descriptive of a capacitor in which the dielectric material is not a perfect insulator. **2.** Descriptive of imperfect shielding in a coaxial transmission line. **3.** Descriptive of a waveguide with imperfect shielding.

leaky dielectric See LEAKY INSULATOR.

leaky insulator An insulator that conducts significant current at a specified (test) voltage.

leaky waveguide A waveguide that has imperfect shielding, allowing some electromagnetic field to escape.

leapfrogging In radar, a phasing process that eliminates false echoes resulting from the signals of other radar sets.

leapfrog test A test performed on different locations by a computer program in memory; it moves to another memory area to continue tests on other locations.

learning The ability of an artificially intelligent machine to improve or expand its knowledge with time. This can occur as a result of accumulation of information; it can also occur in systems that track their errors to avoid repeating them.

leased line A communications circuit reserved exclusively for a specific user.

least-significant bit Abbreviation, LSB. The digit with the lowest place value in a binary number.

least-significant character Abbreviation, LSC. In positional notation, the extreme right-hand character in a group of significant characters. See POSITIONAL NOTATION.

least-significant digit Abbreviation, LSD. The digit in a number that is at the extreme right (i.e., the one having the lowest place value).

least upper bound The smallest value of a parameter that can be obtained without changing some characteristic of a circuit, program or system.

Lecher frame A sturdy assemblage of LECHER WIRES.

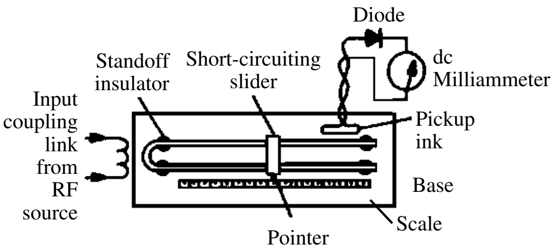

Lecher frame

Lecher lines See LECHER WIRES.

Lecher oscillator A self-excited radio-frequency oscillator using Lecher wires in place of an inductance-capacitance (LC) tank circuit. Also see LINE-TYPE OSCILLATOR.

Lecher wires A circuit segment consisting of two parallel wires or rods joined by a coupling loop on one end, the other end being open. A short-circuiting bar is moved along the wires to vary the effective length of the circuit. Radio-frequency energy is inductively coupled into the system through the loop, and the bar is slid along to various response points, as indicated by a meter or lamp coupled to the wires. The frequency can be determined by measuring the distance between adjacent response points. It is used at very-high frequencies (VHF), ultra-high frequencies (UHF), and microwave frequencies. Also called *Lecher frame* and *Lecher lines.*

Leclanche cell See DRY CELL, **1**.

LED See LIGHT-EMITTING DIODE.

LED phototransistor isolator An optoelectronic isolator in which the light source is a light-emitting diode and the light-sensitive component is a phototransistor.

LEDE Abbreviation for LIVE END/DEAD END.

LEF See LIGHT-EMITTING FILM.

left-hand lay See DIRECTION OF LAY.

left-hand motor rule See FLEMING'S LEFT-HAND RULE.

left-hand polarized wave See COUNTERCLOCKWISE-POLARIZED WAVE.

left-hand taper Potentiometer or rheostat taper in which most of the resistance is in the counterclockwise half of rotation as viewed from the front. Compare RIGHT-HAND TAPER.

left justified An item of data that occupies consecutive locations in storage, starting at the left-hand end of its area; empty locations might appear consecutively at the right-hand end if the item needs fewer positions than have been provided.

left shift A shift operation in which the digits of a word are displaced to the left; the effect is multiplication in an arithmetic shift.

leg **1.** Any one of the distinct branches of a circuit or network; also called ARM or BRANCH. **2.** In a computer program, a path in a routine or subroutine.

L electron In certain atoms, an electron whose orbit is outside of, and nearest to, the orbit of the K electrons.

Lenard rays See CATHODE RAYS.

length **1.** The number of bits or characters in a record, word, or other data unit. **2.** The end-to-end dimension of a device, circuit, line, etc. **3.** The start-to-finish duration of a time interval.

length to fault In cable or line measurements from the home station, the distance (i.e., the cable or line length) to the point at which a fault, such as a short circuit or ground, is located.

lens **1.** A usually circular piece of transparent material with one or both surfaces curved in cross section, used (through its refractive properties) to focus or spread rays that pass through it. Lenses for visible light must be transparent, but those for other radiation, such as radio waves, need not transmit light. Also see ANTENNA LENS. **2.** In a cathode-ray tube, one or more high-voltage electrodes for focusing an electron beam to a fine point on the screen.

lens antenna See ANTENNA LENS.

lens disk A NIPKOW DISK having a lens in each hole.

lens speed The light-transmitting ability of a lens, given as an f-stop number: the lens' focal length divided by its diameter.

lens turret A multiple-lens mount on a camera that can be rotated for quick lens interchange.

LEO Acronym for LOW EARTH ORBIT.

LEO satellite system A set of several dozen low earth orbit (LEO) satellites in polar orbits spaced strategically around the globe such that, for any point on the earth, there is always at least one satellite in range. The satellites can relay messages throughout the fleet. Any two points on the surface are always able to make contact through the satellites.

Lepel discharger See QUENCHED SPARK GAP.

letter-identification word See PHONETIC ALPHABET CODE WORD.

letters patent See PATENT.

letters shift In a text communications terminal, a control character that causes all of the following characters to occur in the lower case. This can be an automatic or a manual control character.

let-through current The current conducted by a circuit breaker during a short-circuit.

level-0 security In communications, the complete lack of security measures. Anyone can listen in on a conversation at any time, provided they are willing to spend the money and/or time to obtain the necessary equipment. Examples are amateur ("ham") radio and Citizens' Band (CB) voice communications.

level-1 security Also called *wire-equivalent security.* In communications, the implementation of security measures such that the circuit is protected to the extent of a typical hard-wired link. The encryption is anticipated to be unbreakable for at least 12 months, and preferably for 24 months or more. The technology is updated at least every 12 months, and preferably every 6 months.

level-2 security Also called *commercial-level security.* In communications, the implementation of security measures such that the circuit is deemed safe for ordinary commercial transactions. The encryption is anticipated to be of such a nature that engineers believe it would take a hacker at least 10 years, and preferably 20 years or more, to break the cipher. The technology should be updated at least every 10 years, but preferably every 3 to 5 years, and more often if possible.

level-3 security Also called *mil-spec security.* In communications, the implementation of security involving the most sophisticated forms of encryption and personnel restriction that a government can muster.

level **1.** The amplitude at which a device is functioning or at which a phenomenon occurs (e.g., collector-current level, or received signal level). **2.** The minimum amplitude at which a phenomenon occurs; also called *threshold amplitude.* **3.** A functional plateau or echelon.

level clipper See CLIPPER.

level compensator **1.** An automatic gain control (AGC) that effectively reduces amplitude variations in a received signal. **2.** An automatic gain control in telegraph receiving equipment.

level control **1.** The adjustment of amplitude or threshold. **2.** A potentiometer or other variable component for adjusting the amplitude or threshold of a quantity.

level indicator See VOLUME INDICATOR.

level translator Any circuit or device that alters the voltage levels of input signals. An example is a converter that changes positive-logic signals to negative-logic signals.

level-triggered flip-flop A flip-flop that responds to voltage level, rather than to the frequency of an input signal.

lever switch **1.** A switch designed for rapid making and breaking of a circuit. **2.** A radiotelegraph key.

Lewis antenna A form of antenna used at ultrahigh and microwave frequencies. It resembles a horn antenna.

Leyden bottle See LEYDEN JAR.

Leyden jar [Leyden, Holland (also Leiden), site of the invention in 1745 by Peiter van Musschenbroek, 1692–1761.] The first practical capacitor. In modern form, it is a glass jar covered inside and out with metal foil and has a rod topped by a metal ball that touches the inner foil. It is still used occasionally in classrooms for demonstrating static electricity. The Leyden jar was co-invented by van Musschenbroek and invented independently by E. G. von Kleist of Pomerania, among others.

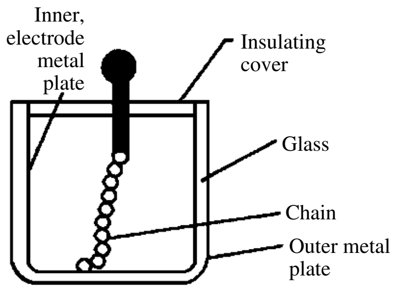

Leyden jar

Leyden phial See LEYDEN JAR.

Leyden vial See LEYDEN JAR.

LF Abbreviation of LOW FREQUENCY.

L_F band A section of the L BAND extending from 1.35 to 1.45 GHz.

L fitting See ELL.

LHD Abbreviation of LOAD/HAUL/DUMP.

Li Symbol for LITHIUM.

librarian program A computer program controlling a LIBRARY.

library In digital-computer and data-processing operations, the permanent storage of data or instructions. Also called *permanent mass storage.*

libration fading In earth-moon-earth (EME) communications, also known as *moonbounce,* rapid and deep fading, accompanied by phase modulation, that takes place because the moon does not always show the earth exactly the same portion of its surface. The moon "wobbles" slightly back and forth on its axis. This causes changes in the relative phases of signals reflected from the topographic features (mountains, in particular) on the lunar surface.

licrystal An acronym from *liquid* and *crystal.* See LIQUID CRYSTAL.

lidar See LIGHT DETECTION AND RANGING.

lie detector See POLYGRAPH.

life **1.** The duration of useful service (or of operation before failure) of electronic equipment. **2.** For a non-rechargeable cell or battery, the length of time it will last in a given application before it must be discarded and replaced. **3.** In robotics and artificial intelligence (AI), a general term that refers to qualitative similarities between machines and animate creatures, including human beings.

life test An assessment of the life of electronic equipment—either by means of full-time test runs or accelerated time tests.

lifetime See CARRIER LIFETIME.

lifter A device in a magnetic tape recorder that removes the tape from the recording and playback heads under fast-forward and rewind conditions.

light Visible electromagnetic radiation occurring in the wavelength band of about 750 nanometers (red light) to 390 nanometers (violet light). Included sometimes in the category of light are infrared and ultraviolet rays.

light-activated silicon-controlled rectifier Abbreviation, LASCR. A silicon-controlled rectifier that functions both as a photosensor and a heavy-duty bistable electronic switch, allowing high currents to be switched by means of a light beam.

light-activated silicon-controlled switch Abbreviation, LASCS. A pnpn device that acts simultaneously as photocell and electronic switch.

light adaption **1.** The process whereby the eye adjusts itself to an increase or decrease in illumination. **2.** Similar action in photoelectric devices. **3.** The length of time required for the eye to adjust itself to an increase or decrease in illumination.

light amplifier A solid-state amplifier using an input electroluminescent cell and an output photocell, or some similar pair of components. The device is essentially an optoelectronic coupler with gain.

light-beam communication A system of communication in which a beam of light between transmitting and receiving stations is modulated or interrupted to convey intelligence. A laser is commonly used because it has minimal beam divergence, allowing maximum communication range.

light-beam meter An electric meter using a LIGHT-BEAM POINTER.

light-beam pointer A slender beam of light that replaces the pointer in a moving-coil meter. The light comes from a small incandescent lamp and is reflected by a mirror attached to the coil; when the coil moves, a spot of light moves over the scale of the meter.

light-beam receiver The receiver in a LIGHT-BEAM COMMUNICATION system.

light-beam recorder A graphic recorder using a light-beam pointer. In this device, a small spot of light traces a pattern on moving photographic film, which is subsequently developed to produce a permanent record.

light-beam transmitter The transmitter in a LIGHT-BEAM COMMUNICATION system.

light cable A cable, consisting of numerous thin optical fibers, through which light can be transmitted for communication or control purposes. See, for example, LIGHT-WAVE TELEPHONY.

light chopper A device that modulates a light beam by interrupting it repetitively.

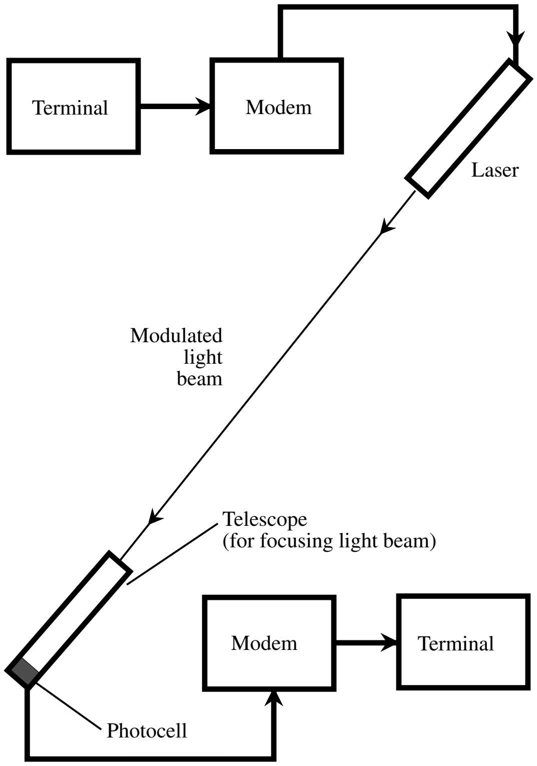

light-beam communication

light detection and ranging Acronym, lidar. A navigation and surveillance system in which laser light scans in a manner similar to that of RADAR.

light dimmer See DIMMER.

lighted pushbutton See LIGHTED SWITCH.

lighted switch A pushbutton switch containing a pilot light that glows to show when the switch is on. Also called *illuminated switch.*

light-emitting diode Abbreviation, LED. A semiconductor device that emits visible light when forward biased. Also see LASER DIODE.

light-emitting film A thin phosphor film that becomes luminescent when a high-frequency voltage is applied across its surface. Also see ELECTROLUMINESCENCE and ELECTROLUMINESCENT CELL.

light flasher An electronic circuit or simple automatic flasher switch for flashing a lamp at regular intervals.

light flicker See LOAD FLICKER.

light flux See LUMINOUS FLUX.

light hood See HOOD.

light-induced electricity See PHOTOELECTRICITY.

light load A load that is a fraction of the usual value for a given application. That is, its resistance or impedance is several times higher than normal.

light meter An electronic instrument for measuring the intensity of light. It generally consists of a photodiode, a battery, and a direct-current microammeter connected in series. A direct-current amplifier can be used to increase the sensitivity.

light microsecond A unit of electrical distance; the distance that light, or any electromagnetic disturbance, travels in free space in 1 microsecond. Approximately equal to 300 meters.

light modulation Variation of the instantaneous brightness of a visible light beam in synchronization with the instantaneous amplitude of a modulating signal. Also see LIGHT MODULATOR.

light modulator A device with which a beam of light can be modulated by an electrical signal.

light negative Pertaining to negative photoconductivity, the decrease in conductivity of a photosensitive material under illumination. Compare LIGHT POSITIVE.

lightning The discharge that occurs between positive and negative poles in the atmosphere. Common in and near areas where heavy rainfall is occurring. It also can occur in snow storms, in sand storms, and over erupting volcanoes. Generally, the negative pole is in a cloud and the positive pole is at the surface of the earth, resulting in a flow of electrons from cloud to ground. Some lightning occurs as a flow of electrons from ground to cloud, or between two clouds. Such discharges sometimes attain current levels of more than 1,000,000 amperes.

lightning arrester A device that bypasses high-voltage pulses from most nearby lightning discharges to the earth, helping to protect electronic equipment connected to an outdoor antenna or power line. Is not a total guarantee of protection, however.

lightning conductor **1.** A system for protecting buildings from lightning strikes. A common system includes a lightning rod, heavy conductor, and ground rod. The ground rod is placed at least six feet from the base of a building and is at least eight feet long. **2.** The conductor between the lightning rod and ground rod in a system, as defined in **1**.

lightning detector See KERAUNOGRAPH and KERAUNOPHONE.

lightning rod A protective device mounted on the outside of structures, consisting of a pointed, grounded metal rod that will conduct a lightning discharge to earth.

lightning strike A direct hit of an object by a lightning stroke. It usually causes extensive damage to electrical and electronic equipment through which the discharge passes.

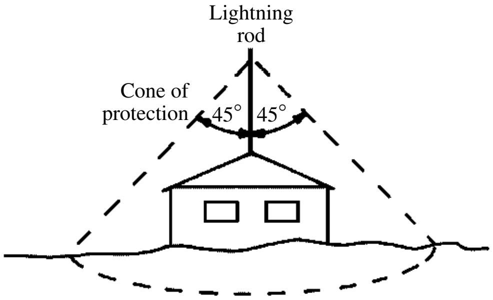

lightning rod

lightning stroke The discharge that occurs with lightning. The peak current is typically several tens of thousands of amperes, but in some cases can exceed 1,000,000 amperes. A stroke can consist of one discharge or several individual discharges in rapid succession.

lightning switch See GROUND SWITCH.

light-operated relay See PHOTOELECTRIC RELAY.

light-operated switch A PHOTOELECTRIC RELAY, or a switch operated by such a relay.

light pen A probe containing a tiny photosensor in its tip. The tip of the light pen is touched to the screen of a cathode-ray tube to sense the beam when it passes the spot of contact. It is used as an input device in some computers and terminals.

light pipe **1.** An OPTICAL FIBER. **2.** A cable consisting of numerous optical fibers in a bundle. See FIBEROPTICS, **1**.

light positive Pertaining to positive photoconductivity, when the conductivity of a photosensitive material increases under illumination. Compare LIGHT NEGATIVE.

light quantum See PHOTON.

light ray A thin beam of light. Theoretically, a ray emerges from a point source (i.e., it has no width).

light receiver See LIGHT-BEAM RECEIVER.

light relay A photoelectric device that operates a relay, according to fluctuations in the intensity of a light beam.

light-sensitive cathode Also called *photocathode.* A cathode that emits electrons when exposed to light.

light-sensitive diode A semiconductor diode usable as a photoconductive cell. Such diodes are available as both junction and point-contact types.

light-sensitive material A photoconductive or photoemissive substance.

light-sensitive resistor See PHOTOCONDUCTIVE CELL.

light sensor **1.** A light-sensitive device, such as a photocell, photodiode, phototransistor, or phototube. **2.** A light-sensitive substance, such as cesium, selenium, silicon, cadmium selenide, or lead sulfide.

light source Any generator of light. Under some conditions, the source is regarded as a point.

light spectrum See ELECTROMAGNETIC THEORY OF LIGHT.

light-spot scanner Also called *flying-spot scanner.* A television camera using (as a source of illumination) a spot of light that scans what is to be televised.

light transmitter See LIGHT-BEAM TRANSMITTER.

light valve **1.** An electromechanical device for varying the intensity of light passing through its adjustable aperture. **2.** See KERR CELL.

light-wave telephony Telephone communication by means of modulated-light transmission, usually through an OPTICAL FIBER.

light-year Abbreviation, lt-yr. Pertaining to astronomy, a unit of distance equal to the distance traveled by light in one year in a vacuum: $9.460\ 55 \times 10^{15}$ meters (5.878×10^{12} miles).

likelihood In probability and statistics, the chance that an event will occur or that an outcome will be realized. Also see PROBABILITY, **1**, **2**.

lim Abbreviation of LIMIT.

Lima Pronunciation, *LEE-ma.* Phonetic alphabet word for the letter L.

limen A unit that has been proposed as the minimum audible change in frequency that can be detected by at least half of a group of listeners.

limit **1.** The lowest or highest frequency in a band. **2.** In mathematics, a fixed value that a variable approaches. **3.** The upper and lower extremes in any performance range or value range.

limit bridge A bridge used to check a component (e.g., resistance, capacitance, or inductance) in terms of the tolerance limits, rather than the nominal (named) value, of that component. Also see BRIDGE, **2**.

limited integrator A circuit that integrates two input signals until the corresponding output signal exceeds a certain limit.

limited stability A characteristic of a circuit or system, allowing proper operation only if the input signal and applied voltages are within certain maximum and minimum limits.

limiter A device or circuit whose output-signal amplitude remains at some predetermined level, despite wide variations in input-signal amplitude.

limiting The restriction of the maximum peak amplitude of a signal to a designated level.

limiting amplifier An amplifier that automatically holds the output-signal level to a prescribed value.

limiting current In electrolysis, the highest current that conducts under certain conditions of ion concentration. This current depends on the electrolyte material, the concentration of the electrolyte in solution, the electrode substance, and the size of the electrolytic cell.

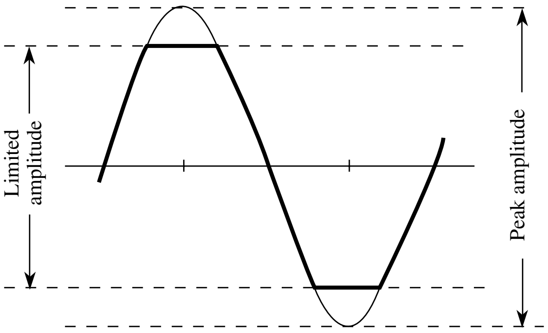

limiting

limiting error The anticipated maximum value of the absolute error in a computation.

limiting resistor See CURRENT-LIMITING RESISTOR.

limiting resolution As a measure of video image resolution, the maximum number of lines for picture height that can be discriminated on a test chart.

limit switch A switch that is actuated when a monitored quantity (e.g., current, voltage, or illumination) reaches the limit of its range.

line **1.** A wire, cable, or waveguide, along which electrical or electromagnetic energy travels from one defined place to another. **2.** One lengthwise path in which a force, such as electricity or magnetism, is evidenced. Such a *line of flux* has theoretically zero width.

line advance **1.** The physical separation between the centers of adjacent scanning lines in a television system. **2.** Line feed in a text data transmission system.

line amplifier An amplifier in a telephone line or similar channel, or one feeding such a line from the input end.

linear **1.** In a straight line. **2.** In the manner of a straight line. Thus, linear response is indicated when one quantity varies directly with another; the graph of this response is a straight line (i.e., one of constant slope). **3.** The characteristic of a signal that is a replica of another (e.g., an amplifier output signal of the same waveform as that of the input signal).

linear absorption coefficient A number expressing the extent to which the intensity of an X-ray beam is reduced per centimeter of the material through which it passes.

linear accelerator A device in which subatomic particles are accelerated in a straight line through a long tube. This action is in contrast with that occurring in a circular accelerator, such as a CYCLOTRON.

linear algebra A branch of mathematics that deals with the solving of linear equations or sets of linear equations.

linear amplifier **1.** An amplifier for which a linear relationship exists between input and output parameters (e.g., a high-fidelity audio amplifier). **2.** A class-AB radio-frequency power amplifier that does not distort the envelope of an amplitude-modulated (AM) or single-sideband (SSB) signal. It is commonly used by amateur radio operators.

linear array A directional antenna having equally spaced, in-line elements.

linear circuit **1.** A circuit whose output is a faithful reproduction of the input. See LINEAR AMPLIFIER, **1** and LINEAR DETECTOR. **2.** A circuit whose performance is linear. See LINEAR RESPONSE, **1**.

linear decrement In a damped wave, a constant decrease in amplitude with time, as opposed to a decrease that is a logarithmic or otherwise nonlinear function of time. Compare DECREMENT.

linear detector A detector whose output/input relationship is linear. Also see LINEAR, **2**, **3**; LINEAR CIRCUIT, **1**; and LINEAR RESPONSE, **1**, **2**, **3**.

linear differential transformer A device that converts the physical position of an object into an output voltage or current. The voltage or current is directly proportional to the displacement.

linear distortion Amplitude distortion in which the output and input signal envelopes are disproportionate (in the absence of spurious frequencies).

linear equation See FIRST-DEGREE EQUATION.

linear function **1.** In Cartesian two-space, a function of the form $y = mx + b$, where the coordinates are (x,y), the slope is m (any real number) and b is the point on the y axis at which the graph crosses the axis. **2.** Any function in any number of dimensions whose graph appears as a straight line in the Cartesian system of coordinates.

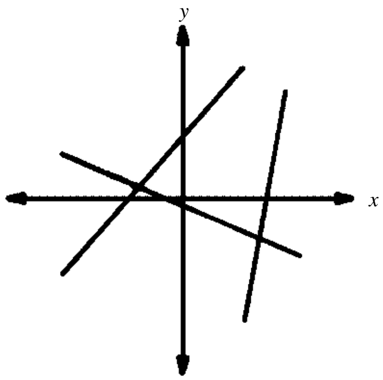

linear function

linear IC See LINEAR INTEGRATED CIRCUIT.

linear integrated circuit An integrated circuit designed for analog operations (such as signal amplification, oscillation, nondigital regulation, analog instrumentation, and similar applications). Compare DIGITAL INTEGRATED CIRCUIT.

linearity **1.** The degree to which performance or response approaches the condition of being linear, expressed in percent or parts per million. Also see LINEAR AMPLIFIER, **1**; LINEAR CIRCUIT, **1**, **2**; LINEAR OSCILLATOR, **1**; LINEAR RESPONSE, **1**, **2**, **3**; and LINEAR TAPER. **2.** In a cathode-ray-tube image, absence of compression or stretching of any portion of the image; that is, an undistorted reproduction.

linearity control In a cathode-ray-tube display, the potentiometer used to correct image linearity. See LINEARITY, **2**.

linearity error **1.** The difference between a theoretically linear function and the actual function, as observed under experimental conditions. **2.** The degree of nonlinearity in an amplifier that is supposed to be linear.

linear modulation **1.** Modulation in which the instantaneous amplitude of the input signal is directly proportional to the instantaneous amplitude of the output signal. **2.** Modulation in which the instantaneous amplitude of the input signal is inversely proportional to the instantaneous amplitude of the output signal. **3.** Modulation in which the instantaneous amplitude of the input signal is directly proportional to the frequency or phase deviation of the output signal.

linear motor A motor in which the stator and rotor are parallel and straight.

linear oscillator **1.** An oscillator whose alternating-current output amplitude varies linearly with its direct-current input. **2.** A line-type oscillator.

linear programming A method of determining the optimum value for a certain set of linear equations. Generally, this is done by finding the point on a plane in space closest to some point not on the plane.

linear quantizing A method of quantizing in which all of the intervals are of equal size or duration. An example is a linear analog-to-digital converter circuit.

linear reflex detector See INFINITE-IMPEDANCE DETECTOR.

linear response **1.** A response in which the value of the dependent variable is equal or directly proportional to the value of the independent variable. Thus, the graph of the response function is a straight line over the range of normal operating values. Compare LOGARITHMIC RESPONSE, **1** and SQUARE-LAW RESPONSE. **2.** A type of response in which a quantity (such as current) varies directly with another quantity (such as voltage). Compare LOGARITHMIC RESPONSE, **2**. **3.** Low-distortion response. Also see HIGH FIDELITY.

linear scale A scale in which all of the divisions represent the same differential and are equally spaced. On a linear scale, a given difference

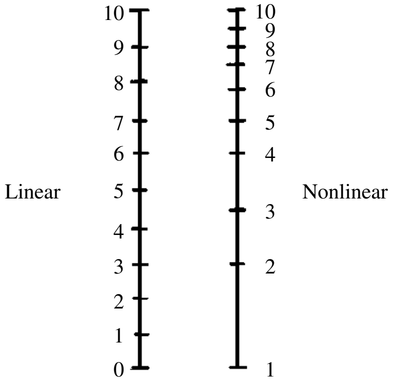

linear scale

always has the same physical length, no matter where on the scale it appears. For example, the interval between 3 and 4 on a linear scale is the same as any interval between *x* and *x* + 1, where *x* is any real value on the scale.

linear sweep In a television or oscilloscope circuit, the scanning of the electron beam across the screen at a constant speed. Also see LINEAR, **1**; LINEAR RESPONSE, **1**; and LINEARITY, **2**.

linear taper In a potentiometer or rheostat, resistance variation that is directly proportional to shaft rotation. Thus, half the total resistance corresponds to movement of the shaft over half the arc of full rotation. Compare LOG TAPER. Also see TAPER.

linear time base For an oscilloscope, the base provided by sweeping the electron beam horizontally at a uniform rate. Also see LINEAR SWEEP.

linear track On a video tape, the track that contains audio information that accompanies the video data. It was so named because it is a single, straight track, in contrast to the video tracks, which are angled.

linear tracking In a turntable system, a design scheme in which the lateral movement of the stylus, as the disc is played, occurs in a straight line, rather than in an arc.

linear transformer A radio-frequency transformer consisting of a section of transmission line.

linear variable differential transformer Abbreviation, LVDT. A differential transformer exhibiting linear response. Also see LINEAR RESPONSE, **1**, **2**.

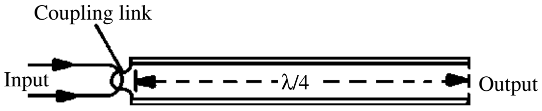

linear transformer

line balance The degree of electrical similarity between transmission line conductors, or between a conductor and ground.

line-balance converter A device used to isolate the outer conductor at the end of a coaxial line from ground.

line characteristic distortion Fluctuations in the duration of received signal impulses in text data communications, caused by changing current transitions in the wire circuit.

line circuit The telephone system relay equipment associated with stations connected to a switchboard.

line code A code between the digits in processing equipment and the pulses representing the digits in a line transmission.

line conditioning In data communications, the modification of private or leased lines by adding compensating reactances to reduce amplitude variations or phase delays over a band of frequencies.

line contact stylus A needle used to reproduce stereo high-fidelity sound from vinyl discs. It has a characteristic oblate ellipsoidal shape.

line coordinate A symbol identifying a specific row of cells in a matrix; a specific cell can be located with an additional column coordinate.

line cord A flexible two- or three-wire insulated cable connecting equipment to the power line by means of a plug that mates to a standard electrical outlet.

line current **1.** Current flowing from a power line into equipment. **2.** Current flowing in a transmission line. **3.** Current flowing into a parallel-resonant circuit.

line diffuser A circuit that creates minor vertical oscillations of the spot on a television screen, making the individual scanning lines less noticeable.

line driver An integrated circuit capable of transmitting logic signals through long lines.

line drop The voltage drop along a line supplying power to a device.

line equalizer See EQUALIZER.

line fault A discontinuity in a transmission line, resulting in signal loss at the receiving end of a circuit.

line feed In a text data transmission system, the movement of the paper, platen, or cursor to allow for printing or displaying an additional line of text.

line filter **1.** A circuit that can be inserted in the utility power cord of an appliance, device, or system for transient suppression and the minimization of electromagnetic interference (EMI). The filter employs series inductors, parallel capacitors, and in some cases, a clamping device. Installed between a radio transmitter and the utility lines, such a filter can choke off radio-frequency (RF) current and help keep utility wiring from acting as an antenna. Installed in the

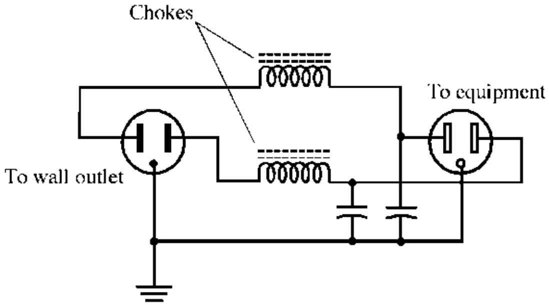

line filter

power cords of home entertainment equipment, such as stereo audio amplifiers, line filters can keep RF from entering the apparatus through the power supply. **2.** A circuit containing parallel capacitors and/or series inductors, installed in a twisted-pair telephone line to reduce or eliminate the effects of EMI and transients in a telephone set.

linefinder A switching device that finds one of a group of calling telephone lines and connects it to a trunk, connector, or selector.

line frequency **1.** The frequency of power-line voltage in the United States, 60 Hz. **2.** The rate at which the horizontal lines are traced in a video image. Also see HORIZONTAL FREQUENCY.

line group **1.** A group of signals sent by wire transmission. **2.** The frequency spectrum occupied by a group of signals sent by wireless transmission.

line leakage Resistance between insulators of two wires in a telephone line loop.

line loss The sum of energy losses in a transmission line.

lineman A technician who works mainly with telephone or telegraph lines.

line matching transformer A device that provides an impedance match in an audio circuit, and also can adapt balanced to unbalanced audio transmission lines (or vice versa).

line noise **1.** Electrical noise (as received by a radio) arising from fluctuations of current or voltage in a power line. **2.** Noise in a data transmission line.

line of flux **1.** A line (usually curved) depicting the points of equal-intensity field strength in the vicinity of an electric charge or a charged body. Also see FLUX. **2.** A line depicting the points of equal-intensity field strength in the vicinity of a magnetic pole or a magnetized body. Also see FLUX.

line-of-sight communication Radio communications between points located so that a straight line between them does not pass through the earth, or through any major obstructions. Also see LINE-OF-SIGHT DISTANCE.

line-of-sight distance The maximum distance over which an ultra-high-frequency (UHF) or microwave signal can be directly transmitted along the surface of the earth. It is slightly more than the maximum optical line-of-sight distance.

line oscillator See LINE-TYPE OSCILLATOR.

line plug The plug terminating a line cord. Also see MALE PLUG.

line printer A machine that prints the results of a computer run, line by line.

line radio See WIRED RADIO.

line regulation Automatic stabilization of power-line voltage.

line-sequential system The color television system in which the image is reproduced by means of primary color lines (red, green, and blue) sequentially beamed across the screen of the picture tube. Compare DOT-SEQUENTIAL SYSTEM and FIELD-SEQUENTIAL SYSTEM.

lines of cleavage See CLEAVAGE.

lines oscillator See LINE-TYPE OSCILLATOR.

line supervision In electronic security systems, a method of monitoring circuit characteristics to detect possible tampering.

line switch **1.** The main power-line switch to a system. **2.** Within a piece of electronic equipment, the switch that opens and closes the circuit to the incoming power line.

line-type amplifier A radio-frequency amplifier in which the tuned circuits are transmission lines consisting of parallel wires, rods, or tubing, or of coaxial cable sections.

line-type oscillator A radio-frequency oscillator in which the tuned circuits are transmission lines consisting of parallel wires, rods, or tubing, or of coaxial cable sections.

line unit In a wire data transmission system, the terminal unit, or device that converts the text signals into electrical impulses and vice versa.

line voltage **1.** The voltage between the conductors in a power line. **2.** The voltage between the conductors of a transmission line.

line-voltage drop See LINE DROP.

line-voltage monitor See POWER-LINE MONITOR.

linguistics The study of languages, including structure, symbology, and phonetics.

link **1.** The small coupling coil used in link coupling. **2.** A communication path between two radio facilities for the purpose of extending the range of one, as between a remote pickup point and a broadcast transmitter. **3.** A data connection between two different computers. **4.** The act or process of creating a signal path or data connection, as defined in **1**, **2**, or **3**. **5.** In a digital computer, a branch instruction, or an address in such an instruction, used to leave a subroutine to return to some point in the main program.

linkage Coupling between separated conductors or devices through the medium of electric or magnetic lines of flux.

link circuit A closed-loop coupling circuit having two coils of a few turns of wire; each coil is placed near one of the circuits to be coupled.

link coupling Low-impedance coupling via a small (usually one-turn) input or output coil fed by a twisted pair or a coaxial line.

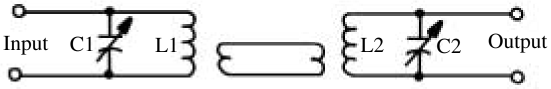

link coupling

linked subroutine A subroutine, entered by a branch instruction from a main routine, that executes a branch instruction returning control to the main routine.

link fuse A fuse consisting of an exposed length of fuse wire.

link neutralization Neutralization achieved by out-of-phase current fed back via link coupling from the output to the input of an amplifier. Also called INDUCTIVE NEUTRALIZATION.

lin-log receiver A radar receiver whose amplitude response is linear for small signals, but logarithmic for large ones.

lip microphone A small microphone operated close to or in contact with the lips.

liquid A state of matter characterized by a level of molecular motion intermediate between that of gases and solids; liquids have the ability (like gases) to take the shape of a container and are only slightly compressible. Compare GAS, PLASMA, and SOLID. Also see STATE OF MATTER.

liquid absorption For a solid material, such as dielectric, the ratio of the weight of liquid absorbed by the material to the weight of the material.

liquid capacitor See WATER CAPACITOR.

liquid cell See ELECTROLYTIC CELL.

liquid conductor See ELECTROLYTE.

liquid cooling Use of circulating water, oil, or other fluid to remove heat from components or equipment, such as microprocessors or power amplifiers.

liquid crystal A liquid exhibiting some of the characteristics of a crystal. Also see NEMATIC CRYSTAL and SMECTIC CRYSTAL.

liquid-crystal display Abbreviation, LCD. A flat-panel display noted for its thin profile, light weight, and low power consumption. The simplest devices are used in calculators, meters, wristwatches, and radios. More sophisticated displays are used in computers and portable video units. This type of display can operate at a much lower voltage than a cathode-ray tube (CRT). This makes it ideal for portable electronic systems in which batteries are used. Older displays of this type can be difficult to read from certain viewing angles. This situation has improved in recent years with the advent of *active-matrix*, also known as *thin-film-transistor* (TFT), displays.

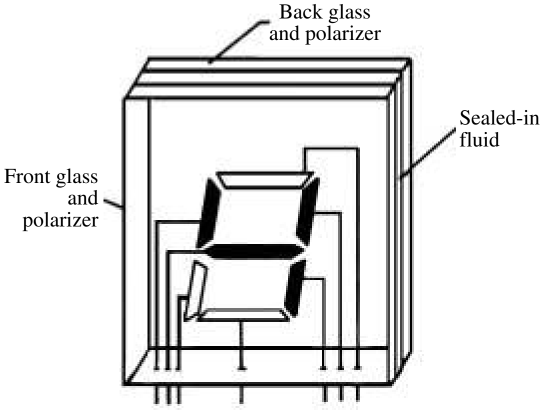

liquid-crystal display

liquid-filled transformer A transformer filled with a protective liquid insulator, such as oil.

liquid-flow alarm An electronic circuit that actuates an alarm when the flow of a liquid through pipes or other channels changes from a desired rate.

liquid-flow control A servo system that automatically maintains or corrects the rate of liquid flow through pipes or other channels.

liquid-flow gauge See LIQUID-FLOW METER.

liquid-flow indicator See LIQUID-FLOW METER.

liquid-flow meter An instrument that indicates the rate at which a liquid moves through pipes or other channels.

liquid-flow switch In a liquid-cooled system, a switch that actuates an alarm when the liquid slows or stops.

liquid-jet oscillograph A graphic recorder using an ink-jet galvanometer to trace the pattern on a paper chart.

liquid laser A laser in which the active material is a liquid.

liquid-level alarm An electronic device that actuates visual or audio signal devices when the surface of a liquid inside a tank rises or falls to a predetermined level.

liquid-level control A servo system that automatically maintains the liquid in a tank at a predetermined level.

liquid-level gauge An electronic system that provides direct readings of the level of a liquid in a tank.

liquid-level indicator See LIQUID-LEVEL GAUGE.

liquid-level meter See LIQUID-LEVEL GAUGE.

liquid load See WATER LOAD.

liquid-pressure alarm An electronic circuit that actuates an alarm when the pressure of a liquid changes.

liquid-pressure control A servo system that automatically maintains or corrects liquid pressure in pipes or other channels.

liquid-pressure gauge See LIQUID-PRESSURE METER.

liquid-pressure indicator See LIQUID-PRESSURE METER.

liquid-pressure meter An instrument that provides direct readings of liquid pressure in a pipe or other channel.

liquid-pressure switch A switch that actuates an external circuit or device when the pressure of a liquid changes.

liquid rheostat See WATER RHEOSTAT.

liquid valve See ELECTROMECHANICAL VALVE.

liser An oscillator that produces an extremely pure microwave carrier signal.

LISP A digital-computer language written in the form of lists. A program can be directly interpreted as data, and vice versa. The entire language is derived from a few basic functions. Programs are easy to debug. It is used in artificial intelligence (AI) research. Also see LANGUAGE.

Lissajous figure Any one of several curves resulting from the combination of two harmonically related sine waves. These figures are familiar in electronics; they are obtained when signals are applied simultaneously to both axes of an oscilloscope. It is also called *Lissajous pattern.*

list **1.** To print serially the records in a file or in memory. **2.** To print (instruct a computer to display) every item of input data in a program. **3.** A one-dimensional array of numbers.

listener fatigue Physiological symptoms, such as headaches, caused by prolonged listening to certain sounds (e.g., a pure sine wave or poorly reproduced music).

listening angle In stereo sound reproduction, the angle between the speakers, with respect to the listener. This angle can vary from zero to 180 degrees. Larger angles result in increased apparent channel separation.

listening test The subjective evaluation of audio equipment by listeners.

liter Abbreviation, l. A metric unit of volume equal to 1.0567 U.S. liquid quarts or 0.908 U.S. dry quart. A liter is the volume of 1 kilogram of water at 4¯C and under a pressure of 1 pascal.

literal operands It is usually applicable to source language instructions, operands that specify the value of a constant, rather than an address of a location in which the constant is stored.

lithium Symbol, Li. An element of the alkali-metal group, and the lightest elemental metal. Atomic number, 3. Atomic weight, 6.941.

lithium battery See LITHIUM CELL.

lithium cell A type of electrochemical cell with exceptional energy-to-mass ratio and long shelf life. There are several variations in the chemical makeup; they all contain lithium, a light, highly reactive metal. These units can be manufactured to supply 1.5 to 3.5 volts, depending on the particular chemistry used. These cells can last for years in very-low-current applications such as memory backup.

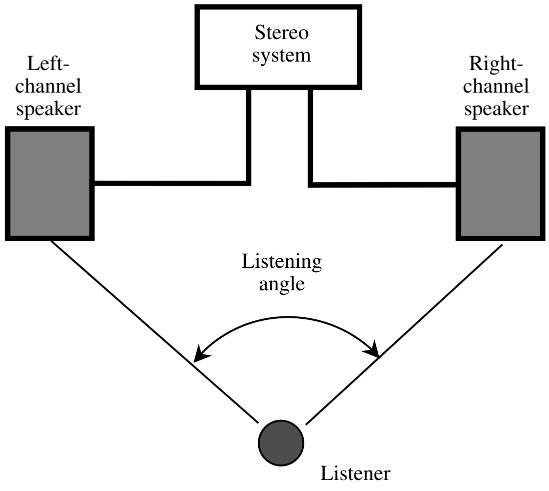

listening angle

Litzendraht wire See LITZ WIRE.

Litz wire A woven wire having a number of copper strands, each separately enameled to insulate it from the others. The wire is woven so that inner strands come to the surface at regular intervals. It is noted for its low losses at radio frequencies.

live **1.** Electrically activated (i.e., sustaining voltage or current). **2.** Being broadcast as it occurs. **3.** Acoustically reflective, as in a LIVE ROOM (contrasted with one that is acoustically absorbent).

live end **1.** In a recording or broadcasting studio, the part of the room in which the acoustic concentration is greatest. **2.** In a utility circuit, the wire or terminal that carries 117 volts alternating current (the ungrounded end).

live end/dead end Pertaining to a room that is acoustically reflective (live) at one end, and acoustically absorbent (dead) at the other end.

live room A room with little or no acoustically absorbent material in its ceiling, walls, and floor, with the result that echoes and reverberation are pronounced. Compare DEAD ROOM.

lix Abbreviation of LIQUID CRYSTAL.

L_K band A section of the L BAND that extends from 1.150 to 1.350 GHz.

L_L band A section of the L BAND that extends from 510 to 725 MHz.

LLL Abbreviation of LOW-LEVEL LOGIC.

lm Preferred abbreviation of LUMEN.

lm/ft^2 Abbreviation of *lumens per square foot.* Also see LUMEN.

lm-hr Abbreviation of LUMEN-HOUR.

lm/m² Abbreviation of *lumens per square meter.* Also see LUMEN and LUX.

lm/W Abbreviation of *lumens per watt*, a unit of luminosity. Also see LUMEN.

ln Representation of the natural (base-*e*) logarithm function (see NAPIERIAN LOGARITHM). Also written *loge*.

L network An impedance-matching circuit, filter, or attenuator whose schematic representation resembles an inverted letter L.

LO Abbreviation of LOCAL OSCILLATOR.

lo Abbreviation of *low*, usually as a prefix or subscript. Also abbreviated L.

load **1.** Also called *electrical load.* A device or circuit that is operated by the electrical power output of another device or circuit. Examples: speaker, light bulb, power amplifier, antenna system. **2.** Also called *mechanical load.* The mechanical power output demanded of a machine—especially a motor. **3.** To fill an internal computer storage with information from an external storage [e.g., from a magnetic disk to a computer's random access memory (RAM)]. **4.** To add inductors and/or capacitors to an antenna system to alter the characteristics of the system—especially the resonant frequency. **5.** To adjust the output circuit of a radio-frequency power amplifier for optimum energy transfer to the antenna system.

load-and-go Automatic coding in which a user's (source) program is translated automatically into machine language and stored.

load capacitance **1.** The capacitance present in an electrical load (see LOAD, **1**). **2.** A capacitance used as an electrical load.

load capacity **1.** In pulse-code modulation, the level at which a sine-wave signal has peaks coinciding with the plus/minus virtual decision values of the encoder. **2.** The maximum number of signals that a medium or line can carry without serious degradation of reception.

load circuit **1.** The circuit that forms the load, or power-consuming portion, of a system. **2.** A circuit that facilitates transfer of power to a load.

load coil See WORK COIL.

load current The current flowing in a load. See LOAD.

load division A method of connecting two or more power sources to a single load, for optimum power transfer.

loaded antenna An antenna incorporating a LOADING COIL and/or LOADING DISK to increase its electrical (effective) length. See LOADING COIL and LOADING DISK.

loaded line A transmission line in which inductors or capacitors are inserted at appropriate points to alter the characteristics of the line.

load end The end of a transmission line to which a radiator or receiver is connected.

load flicker Fluctuations in the brightness of lamps, caused by intermittent loading of the power line by other devices.

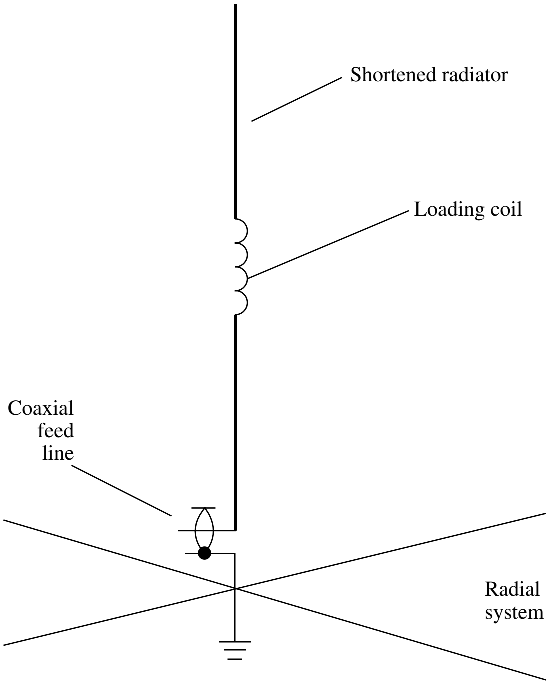

loaded antenna

load/haul/dump Abbreviation, LHD. A remote-controlled or computer-controlled robot used in mining and construction work. Under the direction of a human operator or a computer, it loads, hauls, and dumps materials (hence its name). It can use various navigation methods, including beacons, computer maps, position sensors, and vision systems.

load impedance Symbol, Z_L. The impedance presented by a load connected to a generator or other source.

loading **1.** The matching of source impedance to load impedance, usually by means of the introduction of an inductance or capacitance into the load itself. **2.** Any form of impedance matching. **3.** The addition of inductance and/or capacitance to an antenna system to alter the characteristics of the system—especially the resonant frequency. **4.** The modification of the acoustic impedance of a loudspeaker.

loading coil An inductor inserted in a circuit to increase its total inductance or to provide some special effect, such as canceling capacitive reactance. See LOADED ANTENNA and PUPIN COIL.

loading disk Also called *capacitance hat.* A metal disk mounted atop a vertical antenna to increase its effective length and thereby lower its resonant frequency. It also increases the bandwidth of the system.

loading factor The ratio of source impedance to load impedance before the introduction of loading circuits.

loading inductance **1.** The inductance present in a load. **2.** An inductance used as a load.

loading routine Also called *loading program.* A routine permanently in memory; it allows a program to be loaded into memory from an external storage medium.

load life The longevity of a device in terms of the number of hours it can withstand its full power rating.

load line In a group of voltage-current (*EI*) curves, a line connecting points of equal resistance (*E*/*I*) that are equal to a particular value of load resistance (impedance).

load power The power dissipated in a load.

load regulation Automatic stabilization of load resistance (impedance) at a constant value.

load resistance **1.** The resistance present in a load. **2.** A resistance used as load.

load stabilizer A device for holding load current or load voltage to a constant value.

loadstone Alternate spelling of LODESTONE.

load termination The load connected to the output of a circuit or device as the terminal element in a circuit or system.

load voltage The voltage developed across a load.

load-voltage stabilization Automatic regulation of load voltage.

load wattage See LOAD POWER.

lobe In the directivity pattern of a transducer, a figure, such as a circle or ellipse enclosing an area of intensified response. Applicable especially to antennas.

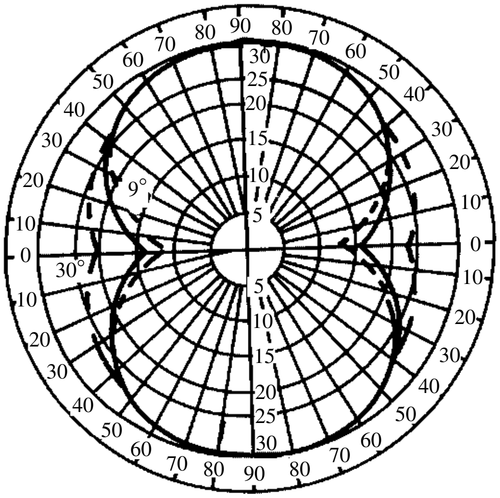

lobe (pattern of horizontal half-wave antenna)

lobing **1.** In a transmitting or receiving antenna, the effect of ground reflection, resulting in phase reinforcement (lobes) at some elevation angles and phase opposition (nulls) at other angles. **2.** The pattern of secondary maxima in the radiation response of a directional antenna.

local action Electrolysis between separate areas of a single electrode immersed in an electrolyte. The action is caused by impurities at different spots in the electrode metal, causing one spot to act as an anode and the other as a cathode, thereby creating a small battery cell.

local alarm In security applications, a visible and/or audible alarm that can be seen and/or heard easily throughout the protected zone.

local area network Abbreviation, LAN. A group of computers and/or terminals that are linked together within a relatively small geographic area, such as a college campus. Interconnections are usually made via cable. There are several different configurations, called *topologies.*

local battery In wire telephony, a battery installed on the subscriber's premises.

local broadcast station A standard broadcast station licensed in the local service. See LOCAL CHANNEL and LOCAL STATION.

local channel A channel in the standard amplitude-modulation (AM) broadcast band, intended to serve only the area near the station. Transmitter power and operating time are restricted to prevent long-distance propagation.

local control The control of a radio transmitter from the site (in contrast to remote control).

local feature focus In robotic systems, the use of only a small portion of the available image data to perform a function. The robot computer recognizes and acts on the data it needs, disregarding the rest.

local feedback Feedback within a circuit stage.

localizer A radionavigation transmitter whose signal guides aircraft to the centerline of a runway.

local oscillator Abbreviation, LO. In a wireless transmitter, one of the oscillators that contributes to the signal that is ultimately modulated and transmitted. The LO output is mixed with the output of a variable-frequency oscillator (VFO). The LO frequency can be switchable at increments of several hundred kilohertz (500 kHz and 1000 kHz are typical), facilitating band changes when mixed with the VFO output. In some transmitters, the LO can operate at a single, fixed frequency if the VFO tunes over an exceptionally wide range.

local program A program that originates at the same single broadcast station from which it is transmitted.

local reception The reception of signals from local stations. Compare LONG-DISTANCE COMMUNICATIONS.

local side The group of circuits and components associated with a communications terminal at a given location.

local station A station situated within the same general area as the receiver, as opposed to a distant station.

local system library A computer program library containing standard software associated with a specific system.

local transmission The sending of signals to receivers in the same general locality as the transmitter, as opposed to long-distance transmission.

local trunk In a telephone system, the interconnecting line between local and long-distance lines.

location In digital computer operation, a memory position (often a register) specified by an address and usually described in terms of the basic storage unit a particular system uses (e.g., a character is a location in a character-oriented machine).

location counter A register in the control section of a computer containing the address of the instruction being executed.

locked groove A continuous blank groove around the inside of a phonograph record. When the disc is done playing, this groove keeps the stylus from running into the label or sliding across the disc.

locked oscillator **1.** A fixed-frequency oscillator, such as a crystal-controlled oscillator. **2.** See BRADLEY DETECTOR.

lock-in A state of synchronism, as when a self-excited oscillator is synchronized (locked-in) with a standard-frequency generator.

lock-in amplifier A detector that makes use of a balanced amplifier. The output is the difference between the collector or drain currents of the two devices.

locking circuit See HOLDING CIRCUIT.

locking relay See LATCHING RELAY.

lock-in relay See LATCHING RELAY.

lock-out **1.** To prevent a hardware unit or routine from being activated (e.g., when there would be a conflict between operations using the same areas of memory). **2.** A safeguard against an attempt to refer to a routine in use.

lock-up relay An electromagnetic relay that can be locked in the actuated state nonmechanically (i.e., by means of an electromagnet or permanent magnet).

locus The set of all points located by stated conditions (e.g., the locus of secondary points that are all equidistant from a primary point is a sphere).

lodestone A natural magnet; a form of the mineral magnetite. Also spelled *loadstone.*

log **1.** Abbreviation of LOGARITHM. **2.** A continuous record of communications kept by a station, or a record of the operation of an equipment.

$\log_{10}$ Abbreviation of *common logarithm* (base-10 logarithm). Also called *Briggsian logarithm.*

logarithm Abbreviation, log. The power y to which a fixed number a, called the *base,* must be raised to equal a given number x. Suppose $x = a^y$, where a, x, and y are real numbers. Then, $\log_a x = y$. The most common logarithmic bases are 10 and the transcendental number e, approximately equal to 2.71828. See COMMON LOGARITHM and NAPIERIAN LOGARITHM.

logarithmic amplifier An amplifier whose output-signal amplitude is proportional to the logarithm of the input-signal amplitude.

logarithmic curve A graphical representation of a logarithmic function, having the form $y = k \log_a x$, where k is a nonzero real-number constant, and a is a positive real number (the logarithmic base).

logarithmic decrement See DECREMENT.

logarithmic graph Also called *log-log graph.* A graph in which the x and y axes are both incremented logarithmically. Compare SEMILOGARITHMIC GRAPH.

logarithmic horn A horn whose diameter varies directly, according to the logarithm of the displacement along the axis. See HORN.

logarithmic mean See GEOMETRIC MEAN.

logarithmic meter A current meter or voltmeter whose deflection is proportional to the logarithm of the quantity under measurement. The increments on the scale of such an instrument are closer together in the upper portion.

logarithmic rate of decay See EXPONENTIAL DECREASE.

logarithmic rate of growth See EXPONENTIAL INCREASE.

logarithmic response **1.** Response in which the value of a dependent variable is at every point proportional to the logarithm of the independent variable. **2.** A type of response in which a quantity (such as current) varies directly with the logarithm of another quantity (such as voltage).

logarithmic scale A graduated scale in which the coordinates are positioned, according to the logarithm of the actual distance from the origin.

logarithmic voltmeter See LOGARITHMIC METER.

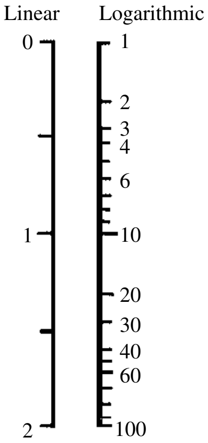

logarithmic scale

$\log_e$ Abbreviation of *logarithm to the base e* (the NAPIERIAN LOGARITHM). Also written *ln*.

logic **1.** In digital circuits, the mathematics dealing with the truth or falsity of statements (represented by variables) and their combinations. Also see SYMBOLIC LOGIC. Generally, "truth" is represented by the binary digit 1, and "falsity" is represented by the binary digit 0. **2.** Collectively, the switching circuits and associated hardware for implementing digital functions (see 1), such as AND, NAND, NOR, OR, etc.

logical decision During a computer program run, a choice between alternatives based on specified conditions. For example, one alternative path in a routine might be selected because an intermediate result was negative.

logical diagram A schematic diagram showing the interconnection between gates of a logic circuit.

logical equivalence The condition in which two logical statements have identical truth value for all possible combinations of truth value of their constituents.

logical file A data set composing one or several logical records.

logical implication For logical statements *x* and *y*, the condition that *y* is true whenever *x* is true: If *x*, then *y*.

logical operation **1.** An operation using logical operators: AND, NOR, OR, and NAND. **2.** A processing operation in which arithmetic is not involved (e.g., a shift).

logical operator A word or symbol representing a logic function operating on one or more operands. Examples: NOT, OR, AND, NOR, and NAND.

logical shift A shift operation in which digits in a word are moved to the left or right in circular fashion; digits displaced at one end of the word are returned at the other. Also called CYCLIC SHIFT.

logic array In logic circuits, a redundant arrangement of identical components in a single package.

logic circuit In digital systems, a gating or switching circuit that performs such logical operations as AND, NAND, NOR, OR, and XOR. It can use diodes, transistors, charge-coupled devices, tunnel diodes, thyristors, ferroelectric elements, magnetic-core elements, or a combination of these. It usually consists of diodes and transistors fabricated onto an integrated-circuit (IC) chip.

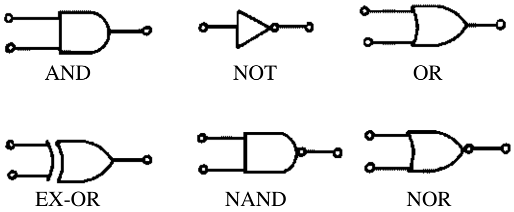

logic circuits (or gates)

logic comparison An operation in which two operands are compared for equal value.

logic connectives Words connecting operands in a logic statement; the truth or falsity of the statements can be determined from their content and the connectives' meanings.

logic diagram **1.** A graphic representation of a logic function (e.g., AND, NAND, NOR, OR, and XOR). **2.** The design of a device or system represented by graphic symbols for logic elements and their relationships.

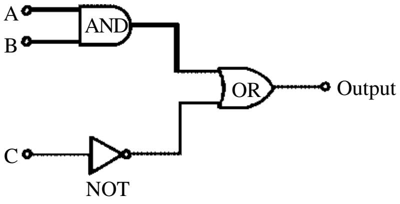

A	B	C	$AB+\overline{C}$
0	0	0	1
0	0	1	0
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	1

logic diagram and truth table

logic diode See COMPUTER DIODE.

logic flowchart The logical steps in a program or subroutine represented by a set of symbols.

logic function An expression for an operation involving one or a combination of logic operators.

logic gate See LOGIC CIRCUIT.

logic input current In an integrated circuit, the input current to the logic gate at a certain voltage value.

logic input levels The range of voltages over which the logic trip level occurs, from low to high or high to low. It is usually expressed in volts for the low state and the high state; for example, low is –1 to +2 volts, high is +4 to +6 volts.

logic instruction A command to execute a logical function.

logic level **1.** One of the two logic states 0 or 1 (on or off, high or low). **2.** Of the two logic states, that which represents the "true" condition. **3.** The voltage amplitude of digital signals in a logic system.

logic probe A test probe with a built-in amplifier and (usually) indicating LEDs; its tip is touched to various test nodes in a digital logic circuit to trace logic levels and pulses.

logic relay See BISTABLE RELAY.

logic swing In a logic circuit, the difference between the voltage corresponding to the high state and the voltage corresponding to the low state.

logic symbol **1.** A symbol used to represent a logic element in a circuit diagram. **2.** A symbol used to represent a logic connective.

log-log graph See LOGARITHMIC GRAPH.

LOGO A high-level computer programming language used for robot control and as an education aid. It is especially useful for teaching children how to operate computers and computerized robots. Statements are simple enough so that elementary-school children can learn to write short programs.

log-periodic antenna Also called *log-periodic dipole array (LPDA).* A broad-spectrum, multiband directional antenna in which the lengths and spacing of the radiator and elements increase logarithmically from one end of the antenna to the other.

log polar navigation A computerized navigation system in which polar-coordinate data is converted into rectangular-coordinate data. In the transformation process, the logarithm of the radius (range) is taken. This results in improved image resolution at close range, although it sacrifices resolution at greater ranges.

log taper In a potentiometer or rheostat, resistance variations that correspond to the logarithm of shaft rotation, or vice versa. Compare LINEAR TAPER. Also see TAPER.

long Abbreviation of LONGITUDE.

long-distance communication **1.** Radio communication between stations separated by distances too great for ground-wave propagation to be effective. **2.** In telephone service, communications that require the dialing of an area code in addition to the local exchange number.

long-distance loop A direct telephone line connecting a subscriber's station to a long distance switchboard.

long-distance reception Reception of radio signals from stations beyond the range of ground-wave propagation. See also LONG-DISTANCE COMMUNICATION, **1**.

long-distance transmission Transmission of radio signals to points beyond the range of ground-wave propagation. See also LONG-DISTANCE COMMUNICATION, **1**.

longitude Abbreviation, long. Angular displacement, measured in degrees around the earth's circumference, to the east and west of the prime meridian that passes through Greenwich, England. Compare LATITUDE. Also see MERIDIAN.

longitude effect The variation (caused by the earth's rotation and magnetic field) of the strength of cosmic rays arriving at different longitudes on the surface of the earth.

longitudinal current Current flowing in the same direction in the parallel wires of a pair (the return circuit is via ground).

longitudinal parity Parity associated with bits recorded on one track of a magnetic storage medium to indicate whether the number of bits is even or odd.

longitudinal redundancy A computer condition, generally affecting magnetic tape records, in which the bits in each track of a record do not meet the required parity, as determined by a LONGITUDINAL REDUNDANCY CHECK.

longitudinal redundancy check A parity check performed on a block of characters or bits (for example, on a track of a magnetic disk). A parity character is generated and transmitted as the last character of the block; thus, each longitudinal block has either even or odd parity.

longitudinal wave A wave in which the movement of particles in a medium is parallel with the direction of propagation.

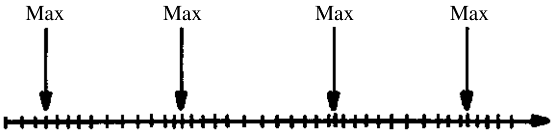

longitudinal wave

long line **1.** A single-wire antenna whose length is greater than the length of the wave fed to it for propagation. Also see LONG-WIRE ANTENNA. **2.** In wire telegraphy, an electrical line that has great physical length. **3.** In electronics theory, a transmission line of indeterminate length, but whose characteristics remain stable and predictable to infinity.

long-persistence screen A cathode-ray-tube screen on which the image remains for a time after the electron beam has passed.

long-play Abbreviation, LP. Descriptive of phonograph discs designed to play at 33.3 revolutions per minute (rpm). Also see MICROGROOVE RECORD.

long-play record See MICROGROOVE RECORD.

long-range navigation See LORAN.

long-range radar Radar that can detect targets at distances of 200 miles or more.

long skip Ionospheric radio-wave propagation, usually via the F layer, between or among stations separated by large geographic distances. The wave angles of departure (from transmitting stations) and arrival (at receiving stations) are very small relative to the horizon. Compare SHORT SKIP.

long-term drift Gradual change in the value of a quantity, such as voltage or frequency, observed over a long period, in contrast to that noted for a brief interval. Compare SHORT-TERM DRIFT.

long-term input offset voltage stability Expressed in microvolts per month. The extent to which the input offset voltage in an integrated circuit stays stable over long periods of time.

long-term stability Stability reckoned over a period of weeks, months or years, as contrasted to that noted for brief intervals of time (minutes or hours).

long throw A speaker design term that describes a woofer moving through long excursions; the objective is to provide good low-frequency response with low distortion.

long waves Low-frequency radio waves, particularly those in the frequency range of 30 kHz to 300 kHz (10 km to 1 km).

long-wire antenna A horizontal or sloping wire antenna measuring a full wavelength or more and fed at a high-current point or at one end. As the wire is made longer, the main radiation/response lobes get more nearly in line with the antenna, and their amplitudes increase. As the wire is made shorter, the main lobes get farther from the axis of the antenna, and their amplitudes decrease. This antenna can produce high gain and excellent low-angle radiation, provided it is straight, is at least several wavelengths long, and is clear of obstructions. However, it cannot be rotated conveniently to change the direction in which maximum gain occurs. Also, a great deal of real estate is needed at medium and high frequencies.

lookup A computer programming technique in which a data item identified by a key is selected from an array.

loop **1.** An electrical circuit consisting of elements connected in series. **2.** In a standing-wave system, a maximum-response point (e.g., *current loop* and *voltage loop*). Compare NODE, **2**. **3.** See LOOP ANTENNA. **4.** A signal path (e.g., *feedback loop*). **5.** A one- or two-turn coil for low-impedance coupling. Also see LINK, **1**. **6.** In a computer program run, the repetitious execution of a series of instructions that terminates when some specified condition is satisfied by a relational test, at which point the next instruction in the main program is obeyed.

loop antenna **1.** A small portable receiving antenna in the form of a wire coil. **2.** A half-wave conductor bent into a circle or square. The conductor is broken at the point opposite the feed point. It can be used for transmitting and receiving. **3.** A full-wavelength, continuous conductor bent into a circle or square. It can be used for transmitting and receiving.

loop checking A method of checking the accuracy of data transmitted over a data link by returning signals received at one terminal to the transmitting terminal for comparison with the original data.

looped amplification See FEEDBACK FACTOR.

looping plug A double phone-plug unit for simultaneously plugging into two phone jacks. Completes (loops) the circuit between the two jacks.

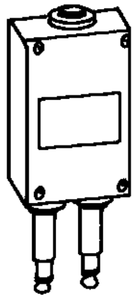

looping plug

loop-input signal A signal introduced into a feedback control loop.

loop pulsing The regular, intermittent breaking of the direct-current path at the transmitting end of a transmission line; also called DIAL PULSING.

loop resistance The resistance of the electrical path around a complete loop (see LOOP, **1**).

loop response time In a security system, the length of time between the first sensing of an abnormal condition (e.g., an intrusion) and the recognition of that condition by the controller.

loopstick antenna See FERRITE-ROD ANTENNA.

loop test A means of locating a discontinuity in a circuit by creating a closed loop, including the suspected fault point.

loose coupling Coupling that transfers only small amounts of energy, as when a primary and secondary coil are spaced so far apart that the coefficient of coupling is small. Compare CLOSE COUPLING.

loosely coupled twin In computer operations, a system in which two processors, each having its own operating system, are used with switches so that they can use common peripherals. Also see SWITCH.

lopplar Acronym for *laser Doppler radar.* Also abbreviated as ladar.

lorac A radio-navigation system that operates by means of phase comparison. Similar to LORAN. Trade name of Seismograph Service Corporation.

loran A long-range radionavigation system in which two pairs of ground stations transmit pulsed signals that are used by aircraft and ships to determine their positions. The name is an acronym for *long-range navigation.*

loran C A radionavigation system that operates at a frequency of 100 kHz. It operates on the hyperbolic principle.

loran D A radionavigation system similar to loran C. It is used by aircraft, operates independent of ground stations, and prevents unwanted enemy detection of aircraft position.

Lorentz force For a charge Q moving with velocity v in a magnetic field B and an electric field E, the force $F = Q(E + (v \times B))$.

loss Energy that is dissipated without doing useful work in a circuit or system. See POWER LOSS.

loss angle For an insulating material, 90 degrees minus the PHASE ANGLE.

loss index For an insulating material, the product of the POWER FACTOR and the DIELECTRIC CONSTANT.

lossless data compression A process in which the number of bits in a data file is reduced by eliminating redundancies, without sacrificing any of the precision or detail in the file. Compression of text and programs must usually be lossless. Image compression need not generally be lossless. Compare LOSSY DATA COMPRESSION.

lossless line A perfect transmission line (i.e., one having no resistance loss, no dielectric loss, and no radiation loss). It is not realizable in practice, but useful in some theoretical calculations.

loss tangent See DISSIPATION FACTOR, **1**.

lossy data compression A process in which the number of bits in a data file (especially a digital image) is reduced by eliminating redundancies, with some sacrifice of precision or detail. In image compression, some loss can usually be tolerated, allowing larger compression ratios than would be possible if zero loss were mandatory. Lossy compression is not generally acceptable for text files and programs. Compare LOSSLESS DATA COMPRESSION.

lossy line A line or cable having comparatively high or excessive attenuation per unit length.

loudness The amplitude of sound, especially in audio reproduction equipment, such as a high-fidelity stereo amplifier. Also called VOLUME.

loudness control See COMPENSATED VOLUME CONTROL.

loudness curves See AUDIBILITY CURVES.

loudness switch/button In a high-fidelity audio amplifier, a switch or button that can be actuated when music is played at low loudness. Increases the volume of the bass relative to the midrange and treble.

loudspeaker A transducer that converts electrical impulses into sound waves of sufficient volume to be heard easily by a number of listeners situated at some distance from the device. Also called *speaker.*

loudspeaker damping See DAMPED LOUDSPEAKER.

loudspeaker dividing network See CROSSOVER NETWORK.

low **1.** The logical digit 0. **2.** Of relatively small magnitude (e.g., LOW VOLTAGE and LOW-FREQUENCY). **3.** In a rechargeable cell or battery, the condition of being near the end of the discharge cycle.

low band **1.** The low or lowest frequency band used in communications, testing, or processing in a given situation. **2.** Television channels 2 to 6 (54 to 88 MHz). **3.** In two-way radio operations, those radio channels between 30 MHz and about 70 MHz.

low battery **1.** The condition of a battery needing replacement or recharging. **2.** An indicating device that shows when a battery needs to be replaced or recharged.

low-battery bias current The current into a designated pin of an integrated circuit required for proper operation of the LOW-BATTERY INDICATOR.

low-battery indicator A device, such as a light-emitting diode (LED) and associated circuitry, that gives a visible indication of the condition of LOW BATTERY.

low-capacitance probe A test probe in which capacitance has been minimized to reduce loading and detuning of the circuit under test.

low earth orbit Acronym, LEO. An artificial satellite orbit that is comparatively low in altitude, resulting in a short orbital period (in some cases less than 2 hours). See also LEO SATELLITE SYSTEM.

low-energy criterion See VON HIPPEL BREAKDOWN THEORY.

lower sideband Abbreviation, LSB. In an amplitude-modulated wave, the lower band of frequencies equal to the difference between the carrier frequency and the modulating frequency. Compare UPPER SIDEBAND.

lower sideband suppressed carrier Abbreviation, LSSC. A single-sideband transmission technique in which the lower sideband is transmitted, but the upper sideband and carrier are suppressed. Compare DOUBLE SIDEBAND SUPPRESSED CARRIER and UPPER SIDEBAND SUPPRESSED CARRIER.

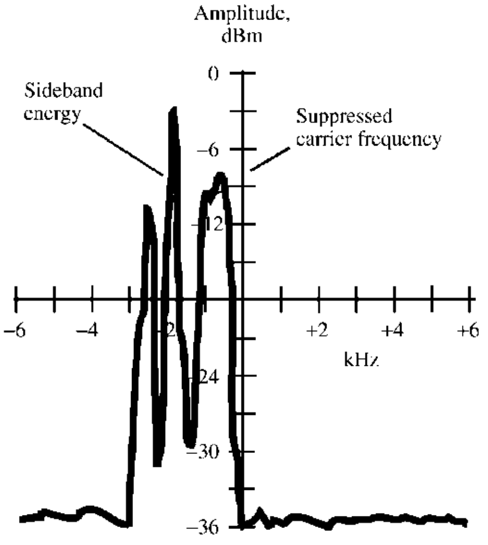

lower sideband

lowest usable frequency Abbreviation, LUF. The lowest frequency that can be used successfully at a given time for communication via the ionosphere. Compare MAXIMUM USABLE FREQUENCY.

low filter A highpass filter that removes low-frequency audio noise from the modulating waveform of a broadcast station. The result is a lower level of transmitted hum and rumble.

low-frequency Abbreviation, LF. **1.** Pertaining to radio frequencies in the band from 30 kHz to 300 kHz (wavelengths from 10 kilometers to 1 kilometer). Also see RADIO SPECTRUM. **2.** Pertaining to audio frequencies below 500 Hz.

low-frequency compensation **1.** In video-amplifier design, special measures, such as use of high coupling and bypass capacitances, to boost low-frequency gain. **2.** Use of special circuits to increase the low-frequency response of an audio amplifier. Also see BASS BOOST, **1**, **2**.

low-frequency direction finder Abbreviation, LDF. A direction finder operated in or below the standard amplitude-modulation (AM) broadcast band, that is, below 1.605 MHz.

low-frequency padder See OSCILLATOR PADDER.

low-frequency parasitics Parasitic oscillations of a frequency lower than that being processed by the amplifier or generated by the oscillator in which they occur.

low-level **1.** A logic term for the more negative of the two (binary) logic levels. **2.** Having an amplitude that is below that normally available in comparable circuits or systems. **3.** In computer operations, pertaining to programming languages (such as assembly language or machine language) that control the machine, but do not directly interface with the operator.

low-level audio signal In audio operations, a signal that has not been amplified by any means (e.g., the output of a dynamic microphone). Compare HIGH-LEVEL AUDIO SIGNAL.

low-level contact A switch or relay contact intended for use with low values of current and voltage.

low-level input current **1.** A test used to check an input pull-up resistor in an integrated circuit to ensure that the fan-in is as specified. **2.** The current flowing from an input when the highest low-level output voltage specified is applied to the input of the device.

low-level language A computer programming language in which each instruction has only one equivalent machine code. Examples are *machine language* and *assembly language.* Compare HIGH-LEVEL LANGUAGE.

low-level logic Abbreviation, LLL. In digital-computer operations, any logic system that operates at low voltage or current levels.

low-level modulation Modulation of a radio or television transmitter at a stage preceding the final radio-frequency (RF) power amplifier.

low-level output current **1.** A test to ascertain that the fan-out and current-sinking capability of an integrated circuit are as specified. **2.** The current flowing into an output with input conditions that cause the output to be at logic low.

low-level signal **1.** A signal with small amplitude. **2.** A signal with peak-to-peak voltage so low that it does not drive an amplifier circuit out of the linear range of operation.

low-loss material A material, particularly a dielectric, having low electrical loss at a given frequency. Also see LOSS.

low-noise Pertaining to circuits, especially weak-signal communications receiving amplifiers and converters, designed to generate the smallest possible amount of internal noise.

low-noise down converter In a satellite television receiving system, a circuit that converts the signals from the dish antenna to frequencies that correspond to the channels on a conventional television set.

low order The lesser-value place(s) of characters or digits in the hierarchy of a group (number or word). For example, 5 and 6 are low-order digits in the number 123,456.

low-order position The extreme right-hand (least significant) position in a number or word.

low-pass filter A combination of capacitance, inductance, and/or resistance, intended to produce large amounts of attenuation above a certain frequency and little or no attenuation below that frequency. The frequency at which the transition occurs is called *cutoff.* At cutoff, the power attenuation is 3 dB with respect to the minimum attenuation. At frequencies below cutoff, the power attenuation is less than 3 dB. At frequencies above cutoff, the power attenuation is more than 3 dB. The simplest circuit consists of a series inductor or a parallel capacitor. The inductance-capacitance (LC) circuit has a combination of series inductors and parallel capacitors. In the resistance-capacitance (RC) circuit, resistors are substituted for the inductors. Compare BAND-PASS FILTER, BAND-REJECTION FILTER, HIGH-PASS FILTER.

low power Abbreviation, LP. Power considerably lower than that ordinarily encountered in a

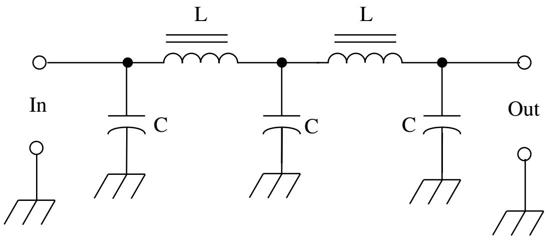

low-pass filter

particular application. The term is arbitrary; several hundred watts might be regarded as low power in one situation, whereas a fraction of a watt would be implied in another.

low-print recording tape Magnetic tape that is less susceptible to print-through than conventional tape.

low Q For a component or circuit, a low quotient for the ratio of reactance to resistance (X/R). This is a relative term because a particular Q value considered low in one situation might be high in other circumstances. Also see FIGURE OF MERIT, **1**.

low tension See LOW VOLTAGE.

low voltage **1.** A voltage considerably lower than that ordinarily encountered in a particular application. The term is arbitrary; several hundred volts might be regarded as low in one situation, and a fraction of a volt would be implied in another. **2.** In a television receiver, the supply voltage applied to all points other than the high-voltage circuit or the picture tube.

low-voltage rectifier In a television receiver, the rectifier that supplies power for the low-voltage stages. See LOW VOLTAGE, **2**.

LP **1.** Abbreviation of LOW POWER. **2.** Abbreviation of LONG PLAY. **3.** Abbreviation of *low pressure.*

L pad An attenuator consisting of one series arm and one shunt arm, arranged in such a way that the schematic representation of the circuit resembles an inverted capital letter L. It is noted for its constant input resistance or impedance as the amount of attenuation is varied.

LPB Abbreviation of LIGHTED PUSHBUTTON.

L_P band A section of the L BAND extending from 390 to 465 MHz.

lpm Abbreviation of *lines per minute*: the output speed of a line printer.

lpW Abbreviation of *lumens per watt*; lm/W is preferred.

Lr Symbol for LAWRENCIUM.

L + R, L – R The sum and difference of the left (L) and right (R) channel signals in a stereo high-fidelity sound system. The L + R signal is the in-phase combination of the two channels; the L – R signal is the out-of-phase combination.

L regulator See L-TYPE VOLTAGE REGULATOR.

LRR Abbreviation of LONG-RANGE RADAR.

LR time constant See INDUCTANCE-RESISTANCE TIME CONSTANT.

LSA diode Abbreviation for *limited-space-charge-accumulation diode.* A solid-state diode that acts as a microwave oscillator.

LSB **1.** Abbreviation of LOWER SIDEBAND. **2.** Abbreviation of LEAST-SIGNIFICANT BIT.

L_S band A section of the L band that extends from 900 to 950 MHz.

LSC Abbreviation of LEAST-SIGNIFICANT CHARACTER.

LSD Abbreviation of LEAST-SIGNIFICANT DIGIT.

L section **1.** A filter section whose schematic representation has the general shape of an inverted capital letter L. **2.** An attenuator circuit whose schematic representation has the general shape of an inverted capital letter L. **3.** A network section consisting of a series (input) impedance arm and a shunt (output) impedance arm. See L PAD. **4.** A right-angle bend in coaxial cable (see ELL).

LSI See LARGE-SCALE INTEGRATION.

LSSC Abbreviation of LOWER SIDEBAND SUPPRESSED CARRIER.

L_T band A section of the L BAND extending from 780 to 900 MHz.

LTROM Abbreviation of LINEAR-TRANSFORMER READ-ONLY MEMORY.

L-type antenna See INVERTED-L ANTENNA.

L-type voltage regulator A simple voltage regulator containing a series current-limiting resistor and shunt regulator (zener diode, VR tube, voltage-dependent resistor, etc.). The schematic representation resembles an inverted capital letter L.

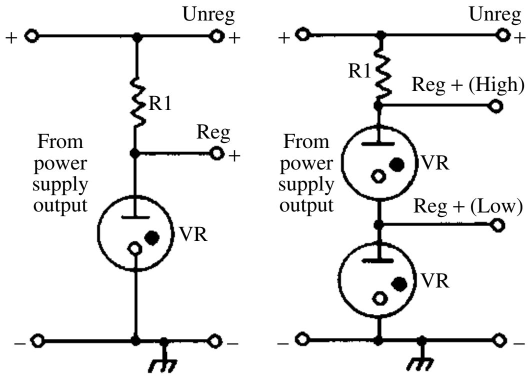

L-type voltage regulator

lt-yr Abbreviation of LIGHT-YEAR.

Lu Symbol for LUTETIUM.

Lucalox General Electric's translucent ceramic; its chief constituent is polycrystalline alumina. The material has many applications in electro-optics.

Lucite Trade name for METHYL METHACRYLATE RESIN.

Luddites During the Industrial Revolution, people who sabotaged automated factory equipment because they believed the machines would put them out of work. The word comes from the name of Ned Ludd, their supposed leader. The term is sometimes applied nowadays to people who fear computers, robots, and other technological innovations because of real or imagined threats to personal job security.

LUF Abbreviation of LOWEST USABLE FREQUENCY.

lug **1.** A contact attached to the end of a wire lead to facilitate connection to a binding post. **2.** A contact attached to a terminal strip, to which wire leads are soldered.

lum Abbreviation of LUMEN. The preferred (SI) form is lm.

lumen Abbreviation, lm, and sometimes l or lum. The SI unit of luminous flux; it is equal to the light that is emitted in one steradian (the unit solid angle) by a uniform point source of one candela. Also see CANDLE POWER, ILLUMINANCE, LUMINOUS INTENSITY, SOLID ANGLE, and STERADIAN.

lumen-hour Abbreviation, lm-hr. The amount of light that a source having a luminous flux of one LUMEN delivers in a time period of one hour.

luminaire A complete and self-contained lighting system, for television-studio use or photographic use. The kit includes all of the needed parts and accessories.

luminance The amount of light emitted or scattered by a surface. This property is expressed in *candelas per square meter* (cd/m^2).

luminance channel In a color television circuit, the channel that processes the Y SIGNAL.

luminance signal See Y SIGNAL.

luminescence The production of visible light, but not heat, by a material stimulated by radiation or electron bombardment. See ELECTROLUMINESCENT CELL and LUMINESCENT SCREEN.

luminescent cell See ELECTROLUMINESCENT CELL.

luminescent screen A cathode-ray tube whose screen is coated with a material that glows under the influence of ionizing radiation, X rays, or electron beams.

luminiferous ether See ETHER, **1**.

luminosity The luminous efficiency of radiant energy, as given by the ratio of luminous flux to radiant flux (lumens per watt) for a specific wavelength.

luminosity factor Abbreviated K, and expressed in lumens per watt. The luminous intensity divided by the actual radiant intensity at a given wavelength of visible light.

luminous energy The energy in visible electromagnetic radiation.

luminous flux The rate of transfer or flow of luminous energy.

luminous intensity Luminous flux through a unit solid angle, expressed in candelas. Also see CANDELA.

lumistor An amplifier or coupling device in which the input signal varies the brilliance of a lamp, electroluminescent cell, or light-emitting diode, and a photocell (or other light-sensitive device) picks up the fluctuating light and uses it to modulate an output current. In a compact lumistor, the light-emitting and light-sensing components are separate layers in a wafer or block of material. Compare LIGHT AMPLIFIER.

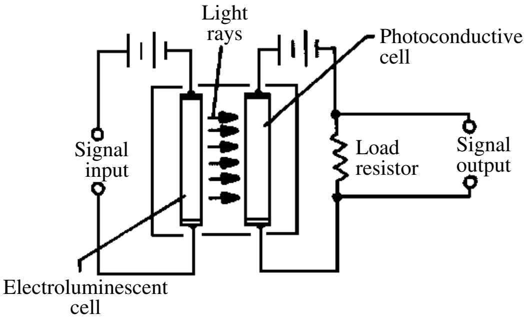

lumistor

lumped Pertaining to a property that is concentrated at or around a single point, rather than being distributed through a circuit (e.g., *lumped capacitance* and *lumped inductance*).

lumped capacitor See DISCRETE CAPACITOR.

lumped component A discrete component (i.e., one that is self-contained). Compare DISTRIBUTED COMPONENT.

lumped constant The total value of any single electrical property in an electrical or electronic component.

lumped-constant delay line A delay line having discrete capacitance and inductance components. Compare DISTRIBUTED-CONSTANT DELAY LINE.

lumped impedance A reactance and/or resistance manifested in a definite location. Examples are ordinary components, such as capacitors, inductors, and resistors.

lumped inductor See DISCRETE INDUCTOR.

lumped parameter Any circuit parameter that can be considered as a discrete parameter—even if it is not made up of a single component.

lumped resistor See DISCRETE RESISTOR.

lutetium Symbol, Lu. A metallic element of the rare-earth group. Atomic number, 71. Atomic weight, 174.967.

lux The unit of illuminance, equivalent to one LUMEN per square meter.

Luxemberg effect The generation of interference by cross-modulation of two or more signals whose paths intersect in the same region of the ionosphere.

luxmeter A device for measuring visible illuminance.

LV Abbreviation of LOW VOLTAGE.

LVDT Abbreviation of LINEAR VARIABLE DIFFERENTIAL TRANSFORMER.

Lw Symbol for LAWRENCIUM; more commonly *Lr.*

lx Abbreviation of LUX.

L_X band A section of the L BAND that extends from 950 MHz to 1.15 GHz.

L_Y band A section of the L BAND that extends from 725 to 780 MHz.

L_Z band A section of the L BAND that extends from 1.450 to 1.550 GHz.

LZT Abbreviation of LEAD ZIRCONATE-TITANATE, a ceramic used in electronics.

m

M **1.** Abbreviation of prefix MEGA-. **2.** Symbol for MUTUAL INDUCTANCE. **3.** Symbol for MODIFIED INDEX OF REFRACTION.

m **1.** Abbreviation of prefix MILLI-. **2.** Symbol for MASS. **3.** Abbreviation of METER. **4.** Abbreviation of MILE. (Also, mi.) **5.** Symbol for MODULATION COEFFICIENT.

m^2 Abbreviation of *square meter*, the SI unit of area.

m^3 Abbreviation of *cubic meter*, the SI unit of volume.

MA **1.** Abbreviation of MAGNETIC AMPLIFIER. (Also, *magamp.*) **2.** Abbreviation of MEGAMPERE.

mA Abbreviation of MILLIAMPERE.

Mache unit A unit of radioactivity equivalent to 13.47 disintegrations per second (3.64×10^{-10} curie) per liter. It represents the concentration of radon gas per liter (when all radiation is absorbed) that will result in a saturation current of 10^{-3} esu (not to be confused with MACH NUMBER).

machine address See ABSOLUTE ADDRESS.

machine code See MACHINE LANGUAGE.

machine cycle In a machine whose operation is periodic, a complete sequence constituting a period of operation.

machine error In a computer or data-processing system, an error attributable to a hardware failure, rather than to a software fault.

machine instruction A computer program instruction written in MACHINE LANGUAGE.

machine knowledge General term for data stored in an artificially intelligent computer system, and the ability of the computer to use that data in meaningful ways.

machine language Computer program instructions and data represented in binary form. In the hierarchy of programming languages, it is the lowest; the computer works directly with it. All high-level languages are translated to machine language by an assembler, compiler, interpreter, or monitor system.

machine learning In artificial intelligence, a computer's ability to learn through repeated calculations for particular problems.

machine logic **1.** The way that a computer's functional parts are interrelated. **2.** The facility whereby a computer solves problems.

machine operation The performance by a computer of a built-in function (e.g., subtraction).

machine operator A person participating in implementing and overseeing the processing of computer programs.

machine word In computer operations, the address of a memory location composed of the full number of bits normally handled by each register of the machine.

machining In industrial robotics, the mechanical modification of parts during assembly. Examples: drilling, welding, sanding, polishing, and painting.

Mach number For a medium such as air, the ratio of the speed of a body in motion to the speed of sound in the medium. (Not to be confused with MACHE UNIT.)

macro **1.** A control shortcut, in which a function requiring the actuation of several switches is abbreviated, via a microcomputer, so that it can be executed by actuating only one or two switches. **2.** Abbreviation of MACROINSTRUCTION.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

macro- Prefix denoting extremely large. Compare MICRO-.

macro assembly program An assembly program whose source statements are translated to several machine-language instructions.

macroinstruction A source program instruction that becomes several machine-language instructions when operated on by a compiler.

macroknowledge In artificial intelligence, knowledge in the large sense (i.e., knowledge about information). Example: a set of definitions in an expert system. Compare MICROKNOWLEDGE.

macroprogram A computer program consisting of macroinstructions.

macrosonics The theory and applications of high-amplitude sound waves.

madistor A component that produces changes in current by means of magnetic-field effects. It is used as an oscillator or amplifier.

MADT Abbreviation of MICROALLOY DIFFUSED TRANSISTOR.

MAG Abbreviation of MAXIMUM AVAILABLE GAIN.

magamp Acronym for MAGNETIC AMPLIFIER.

magazine A tape or film cartridge.

magenta One of the primary pigments used in color printers. It has a pinkish-red hue.

magnal CRT base An 11-pin base typical of many cathode-ray tubes.

magnesium Symbol, Mg. A metallic element. Atomic number, 12. Atomic weight, 24.305.

magnesium fluoride phosphor A substance used as a phosphor coating on the screen of a very-long-persistence cathode-ray tube. The fluorescence and phosphorescence are orange.

magnesium silicate phosphor A substance used as a phosphor coating on the screen of a cathode-ray tube. The fluorescence is orange-red.

magnesium tungstate phosphor A substance used as a phosphor coating on the screen of a cathode-ray tube. The fluorescence is very light blue.

magnet A device or body of material that has the ability to attract to itself pieces of iron and other magnetic metals, and the ability to attract or repel other magnets. Also see ELECTROMAGNET, PERMANENT MAGNET, and TEMPORARY MAGNET.

magnet armature See KEEPER.

magnet battery A group of several magnets placed together in parallel (i.e., with similar poles touching or resting nearby) to act as a single magnet.

magnet charger A device that produces an intense magnetic field for restoring weakened magnets or for making new magnets.

magnetic **1.** Pertaining to MAGNETISM. **2.** Possessing MAGNETISM. **3.** Capable of being magnetized. **2.** See MAGNETIC MATERIAL.

magnetic air-gap A space between two magnetic poles, either the same (in which case the force is repulsive) or opposite (in which case the force is attractive).

magnetic amplifier An iron-core device that uses the principle of the saturable reactor to obtain amplification. In its simplest form, it consists of input and output coils wound on a core of square-loop magnetic metal. The input coil consists of two identical windings connected in series-opposition so that currents in the output winding cannot induce voltage in the input winding. The output coil is connected in series with a load and an alternating-current (ac) supply. A small ac signal applied to the input winding causes a large change in the impedance of the output winding and, therefore, a large change in the voltage across the load.

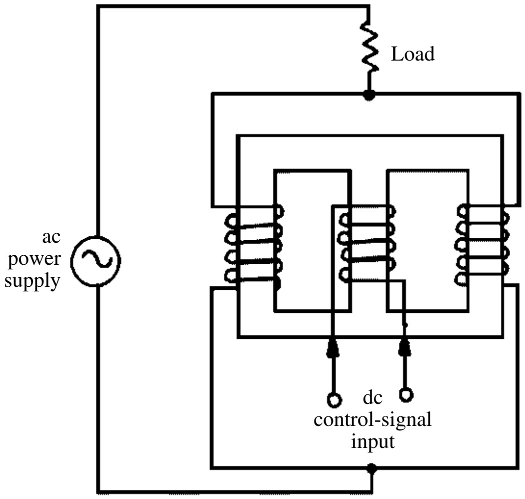

magnetic amplifier

magnetic analysis See MASS SPECTROMETER.

magnetic attraction **1.** The force that causes a magnetic pole to draw to itself an opposite magnetic pole. Thus, a north pole attracts a south pole, and a south pole attracts a north pole. Compare MAGNETIC REPULSION. **2.** The force that causes a magnetic pole to draw to itself a piece of magnetic material, such as iron or steel.

magnetic axis A straight line joining the poles of a magnet.

magnetic bearing The azimuth, or compass direction, measured with respect to magnetic north (the direction of the north geomagnetic pole). It is usually expressed in degrees and can be read directly from a compass.

magnetic azimuth An azimuth bearing relative to magnetic north (the direction of the north geomagnetic pole).

magnetic balance An instrument for measuring the force, either attractive or repulsive, between two objects that are magnetized, or between a magnet and a magnetic substance. The device can also be used for measuring the intensity of a magnetic field, either from a permanent magnet, an electromagnet, or from the earth.

magnetic bias A steady magnetic force applied to another magnetic field to set the latter's quiescent point (e.g., sensitizing a relay by using a permanent magnet to lower the relay draw-in point).

magnetic blowout **1.** The extinction of an electric arc by a strong magnetic field. **2.** The apparatus for accomplishing the action described in **1**.

magnetic bottle A container envisioned for atomic fusion reactions, and that would consist of a magnetic field. Conventional containers cannot withstand the extremely high temperatures involved in atomic fusion.

magnetic braking See ELECTROMAGNETIC BRAKING.

magnetic bridge An instrument comparable to the WHEATSTONE BRIDGE, used to measure magnetic permeability.

magnetic bubble memory See BUBBLE MEMORY.

magnetic capacity The maximum magnetization a given material can receive.

magnetic card A computer storage medium in the form of a card that can be selectively magnetized or imprinted with magnetic ink to represent data.

magnetic cartridge A variable-reluctance phonograph pickup. As the stylus moves in the groove, the vibrations are translated into electric currents by a magnet and coil.

magnetic centering Centering the beam in a television picture tube by means of an electromagnetic focusing coil, a permanent magnet, or both.

magnetic character A letter, numeral, or other symbol written or printed in (visible) magnetic ink for its automatic sensing or reading in computing and signaling operations.

magnetic circuit The closed path determined by a line of magnetic flux or by a set of lines of flux.

magnetic clutch A clutch in which the magnetism of one rotating member causes a second member to lock in and rotate. There need not be physical contact between the two.

magnetic coil The winding in an electromagnet or similar device.

magnetic compass A direction-indicating device using a horizontally suspended magnetic needle as the indicator. The needle tends to point in the direction of the north geomagnetic pole. Compare GYROCOMPASS.

magnetic component See ELECTROMAGNETIC COMPONENT.

magnetic conductivity See PERMEABILITY.

magnetic constant The absolute permeability of free space. It is approximately 1.26×10^{-6} H/m.

magnetic controller A controller that uses electromagnets for some of its functions.

magnetic core The iron core of an electromagnet, choke, transformer, relay, or similar device.

magnetic coupling See INDUCTIVE COUPLING.

magnetic course In navigation, a course referenced to geomagnetic north, rather than geographic north.

magnetic crack detector See ELECTROMAGNETIC CRACK DETECTOR.

magnetic creeping A gradual increase in the magnetization of a material under the influence of a steady magnetizing force.

magnetic cycle **1.** For a material in an alternating magnetic field, the change in magnetic flux as a function of time. **2.** The change in the magnetic-field polarity of the earth. This polarity reverses every few thousand years.

magnetic damping The production of a damping effect or drag in a machine or meter by means of magnetic action on a moving member, in accordance with LENZ'S LAW.

magnetic declination See DECLINATION.

magnetic deflection See ELECTROMAGNETIC DEFLECTION.

magnetic density The concentration of magnetic flux in a region, expressed as the number of lines per unit area of cross section.

magnetic dip At a particular location on the earth's surface, the angle between the terrestrial magnetic field and a horizontal line.

magnetic dipole **1.** A molecule or particle with a north and south magnetic pole. **2.** Any pair of adjacent north and south magnetic poles.

magnetic direction finder Abbreviation, MDF. A type of compass operated by an electric signal delivered by a gyrostabilized magnetic-compass movement.

magnetic disk A rotating disk coated with a layer of magnetic material for the recording, storage and retrieval of information. They are available in various sizes, configurations, and storage capacities. Commonly used with personal computers. Also see DISKETTE and HARD DISK.

magnetic doublet See DOUBLET, **2**.

magnetic drive A device in which mechanical movement is conveyed from one moving part to another by means of a magnetic clutch.

magnetic drum See DRUM.

magnetic effect of electric current The presence of a magnetic field around a conductor carrying electric current.

magnetic equator Also called *geomagnetic equator.* An imaginary circle around the earth, along which a magnetic needle shows no dip. It is near, but slightly displaced from, the geographic equator, and is midway between the geomagnetic poles.

magnetic feedback Feedback by means of inductive coupling between the output and input circuits of a system. It can be positive or negative.

magnetic field The space around a magnetic pole or magnetized body in which magnetic energy acts.

magnetic field intensity See MAGNETIC INTENSITY.

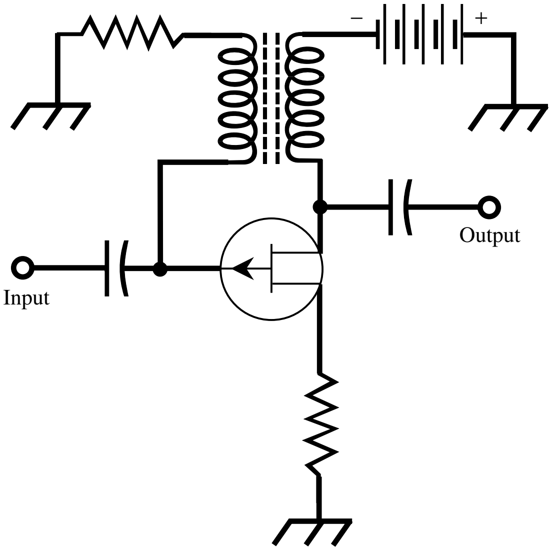

magnetic feedback

magnetic field strength See MAGNETIC INTENSITY.

magnetic-field viewer A device for visually examining a magnetic field. It consists of a clear plastic watchcase filled with iron-oxide particles in liquid suspension. When it is placed within a magnetic field, the particles align themselves in the direction of the magnetic lines of flux.

magnetic-film memory A magnetic memory in which memory cells consist of a thin film (thick film in some instances) of a magnetic material deposited on a substrate. Information is written into and read out of the cell through coils. Also called *thin-film memory.*

magnetic flip-flop A bistable multivibrator using magnetic amplifiers or square-loop cores in place of transistors.

magnetic flux The intensity of a magnetic field through a given area. The unit of magnetic flux is the weber, and the symbol is F. It can loosely be expressed as the number of lines passing through a region of a certain area or of a unit area, such as one square meter. See FLUX.

magnetic flux density See FLUX DENSITY.

magnetic flux linkage The passage of magnetic lines of flux through separate materials or circuits, thereby coupling them magnetically.

magnetic focusing See ELECTROMAGNETIC FOCUSING.

magnetic force The force exerted by a magnet on a body of magnetic material, or on another magnet, within its field.

magnetic friction **1.** See HYSTERESIS, **1**. **2.** The resistance experienced by a magnetic material moving in a magnetic field.

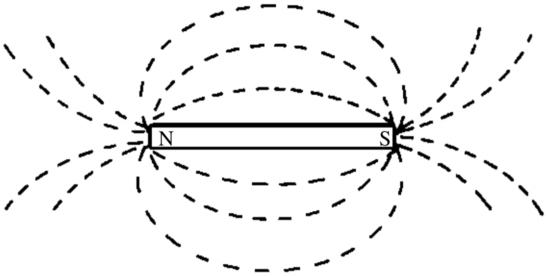

magnetic flux

magnetic gap A space separating the materials in a magnetic circuit. This break is either an air space or one filled with a comparatively thin piece of nonmagnetic material (e.g., the gap in a choke-coil core).

magnetic head See MAGNETIC PICKUP HEAD and MAGNETIC RECORDING HEAD.

magnetic hysteresis See HYSTERESIS, **1**.

magnetic inclination See MAGNETIC DIP.

magnetic induction **1.** The magnetization of a magnetic material, such as iron or steel, when it is placed in a magnetic field. **2.** The induction of an alternating voltage in a conductor by a nearby alternating magnetic field. Also see ELECTROMAGNETIC INDUCTION.

magnetic ink Writing or printing ink that is a suspension of finely divided particles of magnetic material. Also see MAGNETIC CHARACTER.

magnetic instability **1.** The tendency of a magnetic recording medium to deteriorate with time. **2.** Any fluctuation in the intensity of a magnetic field.

magnetic intensity The free-space strength of a magnetic field at a particular point. Specifically, the force (in dynes) that the magnetic field would exert on a unit magnetic pole placed at that point.

magnetic iron oxide See MAGNETITE.

magnetic leakage The usually undesired extension of magnetic flux beyond the confines of a magnetic body, such as the core of a choke.

magnetic lens See ELECTROMAGNETIC LENS.

magnetic line of flux See LINE OF FLUX, **2**.

magnetic load An electromagnetic device operating on the output of an electrical source. Such devices include actuators, alarms, electromagnets, magnetic tapes and disks, relays, and loudspeakers.

magnetic loudspeaker See MAGNETIC SPEAKER, **1**, **2**.

magnetic material **1.** A material, such as magnetite, that exhibits natural magnetism. **2.** A material, such as iron or steel, that is capable of being magnetized.

magnetic media Any medium that stores data as tiny magnetic fields; in particular, MAGNETIC DISK or MAGNETIC TAPE.

magnetic memory **1.** See RETENTIVITY. **2.** A digital memory circuit using magnetic fields to store data bits. Example: BUBBLE MEMORY.

magnetic meridian The circle of the celestial sphere that passes through the zenith and earth's magnetic poles.

magnetic mine A naval mine detonated by a magnetic switch that is closed by the proximity of the steel hull of a ship.

magnetic modulator A core-type device that is somewhat similar to a magnetic amplifier used for amplitude modulation. Modulating current passes through the control winding, and the carrier current through the output winding.

magnetic moment Unit, joule per tesla. For a magnet, the product of pole strength and the distance between poles.

magnetic needle The pivoted magnetic pointer in a magnetic compass.

magnetic north See NORTH MAGNETIC POLE.

magnetic oxide Iron oxide used as the sensitive coating of magnetic recording tape.

magnetic pickup **1.** A phonograph pickup of the variable-reluctance type (see VARIABLE-RELUCTANCE PICKUP). **2.** A magnetic transducer (such as a phono cartridge, tape recording head, or similar input element).

magnetic-pickup head In a tape recorder, the transducer that receives up magnetic impulses from the passing tape and converts them into alternating currents. These currents are amplified to obtain the original sound. Compare MAGNETIC RECORDING HEAD.

magnetic-plate wire Wire in which a magnetic metal has been plated on top of a nonmagnetic metal.

magnetic poles **1.** The points in a MAGNET at which the magnetic lines of flux converge. **2.** The points on the earth at which the geomagnetic lines of flux converge. See NORTH MAGNETIC POLE and SOUTH MAGNETIC POLE.

magnetic pressure See MAGNETOMOTIVE FORCE.

magnetic printing **1.** Also called *magnetic print-through.* In a recording material, such as magnetic tape, the transfer of information from one part of the material to another part (or from one medium to another) by the magnetic field of the recorded material. This phenomenon, which is also called PRINT-THROUGH, sometimes occurs in recording tape on a reel. **2.** Conventional lithography, letterpress, or other reproduction process in which MAGNETIC INK is used.

magnetic print-through See MAGNETIC PRINTING, **1**.

magnetic probe A loop or coil inserted in an electromagnetic field to sample the magnetic component. See, for example, WAVEGUIDE PROBE. Compare ELECTRIC PROBE.

magnetic recording **1.** The recording of sounds or data by varying the magnetization of a medium, such as a magnetic disk or tape. **2.** A magnetic medium on which data has been recorded.

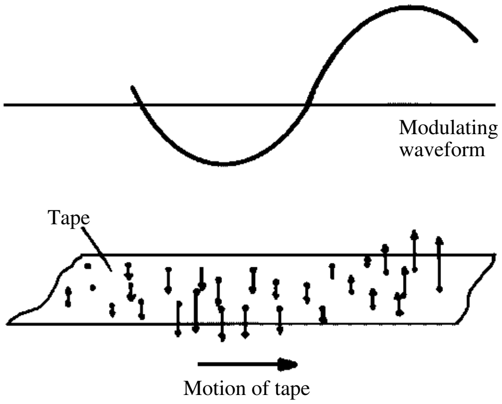

magnetic recording

magnetic recording head In a tape recorder, the transducer that receives current impulses (analogs of the original sound vibrations) from an amplifier and converts them into magnetic impulses that magnetize spaces on the passing tape. Compare MAGNETIC PICKUP HEAD.

magnetic recording medium **1.** A magnetic cylinder, disk, drum, tape, or wire used in the recording of sound or data. **2.** The sensitive material with which any of these is coated.

magnetic relay A relay having a permanent magnet in whose field a coil, bar, or reed moves to open or close a pair of contacts.

magnetic remanence See RESIDUAL MAGNETISM.

magnetic repulsion The force that causes a magnetic pole to push away a similar magnetic pole, although they are not in mutual contact. Thus, two north poles repel each other, and so do two south poles. Compare MAGNETIC ATTRACTION.

magnetic-resonance accelerator See CYCLOTRON.

magnetics **1.** Collectively, magnetic components and equipment. **2.** Collectively, magnetic materials. **3.** A branch of physics dealing with magnets and magnetism.

magnetic saturation The condition in which a magnetic material passes all of the magnetic lines of flux that its permeability allows. Increasing the intensity of the magnetizing force will produce no increase in magnetization.

magnetic scan See ELECTROMAGNETIC DEFLECTION.

magnetic screen See ELECTROMAGNETIC SHIELD.

magnetic shield See ELECTROMAGNETIC SHIELD and MAGNETIC SHIELDING.

magnetic shielding **1.** Enclosing a magnetic field to confine its flux, thus preventing interaction with outside bodies. **2.** Devices (such as boxes, cans, or shells of iron, steel, or a magnetic alloy) used for the purpose described in **1**.

magnetic shift register A shift register using magnetic flip-flops.

magnetic shunt A device that allows the useful magnetic flux of an instrument's magnet to be controlled. The device consists of a piece of magnetic material near the magnet in an electrical measuring instrument.

magnetic south See MAGNETIC SOUTH POLE.

magnetic speaker **1.** A loudspeaker that is essentially an enlarged earphone with a horn that conveys and intensifies the sound from the vibrating diaphragm. **2.** A loudspeaker in which the vibration of a diaphragm or reed in the field of a permanent magnet is conveyed by a pin to a paper or composition cone. Compare DYNAMIC SPEAKER.

magnetic storage **1.** A data bank or memory that stores information in the form of magnetic fields. **2.** The data on a magnetic tape or disk.

magnetic storm A disturbance in the earth's magnetic field that typically follows a solar flare. Often causes interference to radio communications at low, medium and high frequencies.

magnetic strip A strip of powdered iron or ferrite on the back of an identification card, bank cash card, or credit card, that carries a code to identify the account number and verify that the secret entry code (if any) is correct, or that the credit is good.

magnetic susceptibility See SUSCEPTIBILITY.

magnetic switch **1.** In security systems, a switch kept open by the presence of a magnet attached to a door, window, or other movable object. When the object is moved, the magnet moves away from the switch, closing the switch and actuating an alarm. **2.** A REED SWITCH operated by a magnetic field.

magnetic tape Plastic tape coated with a film of magnetic material; it can be magnetized along its length to record sounds, video signals, and computer information.

magnetic-tape core A strip of magnetic metal wound spirally to create a toroid (donut) shape. Such construction is sometimes used in choke or transformer cores. Also see TOROID.

magnetic tape deck See TAPE DECK.

magnetic tape drive See TAPE TRANSPORT.

magnetic tape head See MAGNETIC PICKUP HEAD and MAGNETIC RECORDING HEAD.

magnetic tape library In a computer installation, the place where magnetic tape files are kept, or magnetic tape files and the records needed to utilize them.

magnetic tape parity As a safeguard against losing information bits during the transfer of information between magnetic tape and a memory device, a technique in which an extra bit is generated and added to characters under certain conditions, to make the output uniform temporarily. Lack of uniformity in output then serves as an error indicator. The original quantity is recovered by dropping the extra bit following a parity check.

magnetic tape reader A tape deck for playing back data on magnetic tape.

magnetic tape recorder A recorder-reproducer using magnetic tape.

magnetic test coil See SEARCH COIL.

magnetic thick film A film of magnetic material at least 10^{-6} meter in thickness, deposited on a substrate. Compare MAGNETIC THIN FILM.

magnetic thin film A film of magnetic material, less than 10^{-6} meter in thickness, deposited on a substrate. Compare MAGNETIC THICK FILM.

magnetic transducer A transducer that uses a coil, magnet, or both, to convert displacement into variable magnetic fields or electric currents. Common varieties are the inductance type, transformer type, and generator type. Compare CAPACITIVE TRANSDUCER, CRYSTAL TRANSDUCER, and INDUCTIVE TRANSDUCER.

magnetic tuning In a microwave oscillator, a means of tuning in which a ferrite rod in the cavity resonator is made to have adjustable magnetization so that the resonant frequency of the cavity varies and the frequency of the oscillator is thus adjustable. It is also used at ultra-high frequencies (UHF) and occasionally at very-high frequencies (VHF).

magnetic-vane meter See IRON-VANE METER.

magnetic vector In an electromagnetic field, the vector representing the magnetic component. It is perpendicular to the electric vector.

magnetic viscosity A property of certain materials, described in terms of the time required to magnetize a given substance to a specified level.

magnetic whirl One of the circular magnetic lines of flux around a straight conductor that carries electric current.

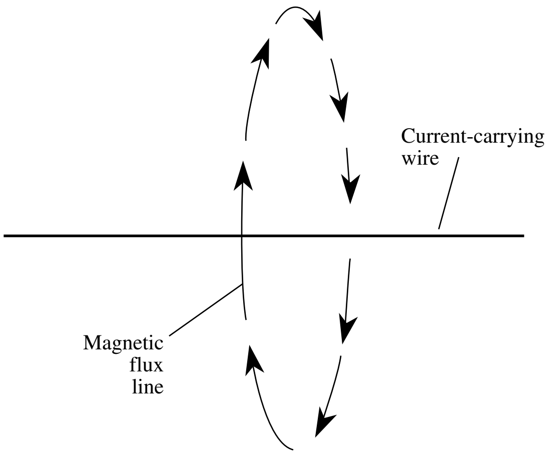

magnetic whirl

magnetic wire The thin wire used in wire recording and playback. See WIRE RECORDER.

magnetism The property of having or causing a magnetic field. It occurs when magnetic dipoles are aligned and when electric charge carriers are in motion.

magnetite A natural magnetic oxide of iron. Also called LODESTONE.

magnetization curve A curve depicting the magnetization of a material versus the applied magnetizing force. See, for example, HYSTERESIS CURVE.

magnetizer A device for magnetizing magnetic materials, as in the making of permanent magnets. Also see MAGNET CHARGER. Compare DEMAGNETIZER.

magnetizing current **1.** A current that sets up a magnetic field of useful intensity. **2.** The half-cycle of an alternating current or the polarity of a direct current flowing through a coil wound on a permanent magnet (as in a headphone, permanent-magnet loudspeaker, or polarized relay) that increases magnetic field strength. Compare DEMAGNETIZING CURRENT. **3.** The field current of a dynamo.

magnetizing force **1.** Magnetomotive force (in gilberts) divided by spatial distance (in meters). **2.** The intensity of a magnetic field that causes a material to become magnetized.

magnet keeper See KEEPER.

magnet meter See MAGNET TESTER.

magnet motor See PERMANENT-MAGNET MOTOR.

magneto See PERMANENT-MAGNET GENERATOR.

magnetocardiogram Abbreviation, MCG. A record, made by a MAGNETOCARDIOGRAPH, of the pulsating magnetic field of the heart. It is used as a diagnostic aid.

magnetocardiograph An instrument that produces a record of the pulsating magnetic field generated around the torso by natural ion currents in the heart.

magnetoelectric generator See MAGNETOGENERATOR.

magnetofluid mechanics See MAGNETOHYDRODYNAMICS.

magnetofluidynamics See MAGNETOHYDRODYNAMICS.

magnetogasdynamics See MAGNETOHYDRODYNAMICS.

magnetoionic duct A propagation path for radio waves between two points that have the same geomagnetic longitude on the surface of the earth. The radio waves tend to travel with the geomagnetic lines of flux at some frequencies under certain conditions.

magnetoionics The study of the effects of the geomagnetic field on the propagation of radio waves.

magnetogenerator See PERMANENT-MAGNET GENERATOR.

magnetograph An instrument for automatically recording a magnetic field.

magnetohydrodynamic generator A device using magnetohydrodynamic principles to generate electric power directly from gases. In the generator, a hot gas is passed through an intense magnetic field; a pair of collector plates picks up electrons from the ionized gas.

magnetohydrodynamic gyroscope A gyroscope whose spin is obtained by a rotating magnetic field circulating a conducting fluid, such as mercury, around a closed loop. Also see MAGNETOHYDRODYNAMICS.

magnetohydrodynamic power generator See MAGNETOHYDRODYNAMIC GENERATOR.

magnetohydrodynamics Abbreviation, MHD. The theory and application of phenomena produced by electrically conductive fluids and gases in electric and magnetic fields.

magnetometer An instrument for measuring the strength and direction of magnetic fields.

magnetomotive force Abbreviation, mmf. Unit, ampere. The phenomenon that is sometimes descriptively called *magnetic pressure.* It is analogous to electromotive force (and to water pressure) and is the agent that produces a magnetic field.

magneton See BOHR MAGNETON.

magneto-optical rotation The tendency of a magnetic field to rotate the plane of polarization of light passing through a substance. Also see KERR MAGNETO-OPTICAL EFFECT.

magneto-optical technology A computer data-storage technology that uses lasers to guide the read/write head in a magnetic disk drive. This greatly increases the amount of data that can be effectively stored on, and retrieved from, a magnetic disk.

magneto-optical valve See KERR MAGNETO-OPTICAL EFFECT.

magnetopause The high-altitude limit of the MAGNETOSPHERE.

magnetoplasmadynamics See MAGNETOHYDRODYNAMICS.

magnetoresistance The phenomenon whereby the resistance of a material, such as a semiconductor, changes when it is exposed to a magnetic field. Also see MAGNETORESISTOR.

magnetoresistor A material (such as bismuth wire, indium antimonide, or indium arsenide) whose resistance varies with the strength of a magnetic field in which it is placed.

magnetosphere In the upper atmosphere, a region extending thousands of kilometers from the earth, in which charged particles are trapped by the earth's magnetic field.

magnetostatic field A stationary magnetic field, such as that produced by a permanent magnet.

magnetostriction The expansion or contraction of a bar or rod of magnetic material (such as Invar, Monel metal, Nichrome, nickel, or Stoic metal) in proportion to the strength of an applied magnetic field. Magnetostrictive vibration in such a rod is

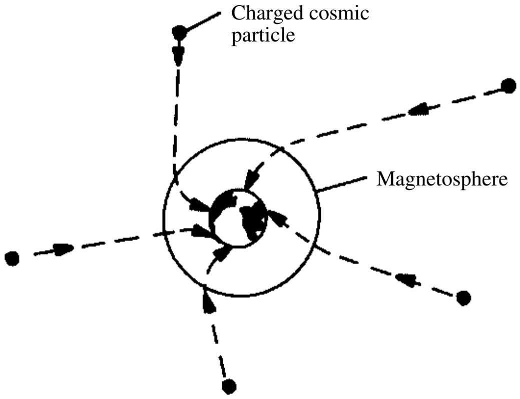

magnetosphere

comparable to piezoelectric vibration in a quartz crystal.

magnetostriction filter See ULTRASONIC FILTER, **1**.

magnetostriction oscillator An oscillator whose frequency is controlled by a magnetostrictive rod (see MAGNETOSTRICTION). The dimensions of the rod and the type of metal it contains determine its vibration frequency and, accordingly, the operating frequency of the oscillator.

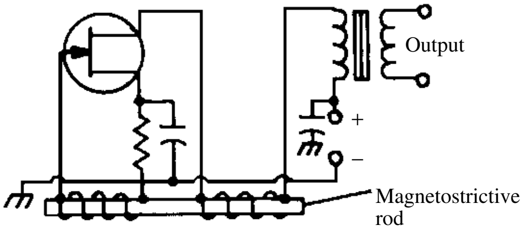

magnetostriction oscillator

magnetostrictive delay line A delay line in which the signal is propagated through a magnetostrictive rod. Also see MAGNETOSTRICTION.

magnetostrictive microphone A microphone in which sound vibrations produce changes in a magnetostrictive element, which, in turn, are converted into output-voltage changes. Also see MAGNETOSTRICTION.

magnetostrictive transducer A transducer in which some phenomenon, such as vibration or pressure, produces changes in a magnetostriction element, which in turn are converted into output-voltage changes. Also see MAGNETOSTRICTION.

magnet protector See KEEPER.

magnetrode The trademark of a radio-frequency device for externally producing hyperthermia (elevated temperature) inside the body, as in heating a cancerous tumor for therapeutic purposes.

magnetron A microwave vacuum tube consisting of a diode (with a cylindrical anode) through which the field of a powerful external permanent magnet passes. The magnetic field causes electrons leaving the cathode to travel in spiral paths between the electrodes. This action gives the tube a negative-resistance characteristic, resulting in oscillation when the tube is connected in an appropriate circuit. Some magnetrons have a built-in resonant cavity.

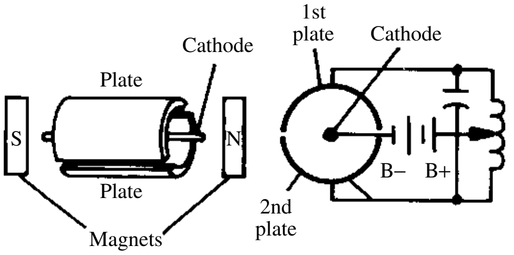

magnetron

magnet steel A high-retentivity alloy of chromium, cobalt, manganese, steel, and tungsten, used in the manufacture of permanent magnets.

magnet tester An instrument used to measure the flux of a magnet. Also see FLUXMETER.

magnet wire Insulated wire (usually solid copper) of 14 to 40 gauge, so called because of its original major use in winding the coils of electromagnets.

magnitude **1.** General expression for degree, size, or extent. **2.** Signal strength (amplitude). **3.** For a number or vector quantity, the absolute value or length. **4.** A measure of the relative or absolute brightness of celestial objects.

mAh Abbreviation of *milliampere-hour.*

main British expression for the alternating-current (ac) utility power available in a house or building.

main bang **1.** In a radar display, the pip or pulse resulting from the actual transmitted signal. This pulse is blanked out. **2.** In a spectrum analyzer, the pip corresponding to a frequency of zero, and caused by the local oscillator.

mainframe **1.** The chassis containing the central processor and arithmetic and logic circuits for a large computer. **2.** The general term for a large, powerful computer.

main lobe Also called *major lobe.* In a directional antenna system, the portion of the directivity pattern representing the greatest transmitted signal gain and/or the greatest received signal response. Also see MINOR LOBE, SIDE LOBE.

main memory The principal (immediate process) memory unit in a digital computer or data-processing system.

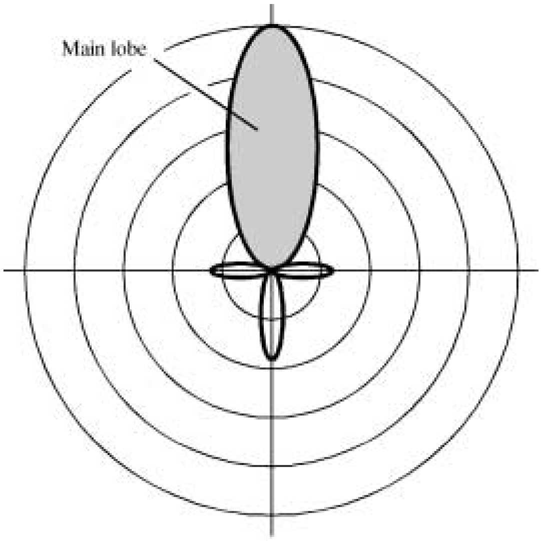

main lobe

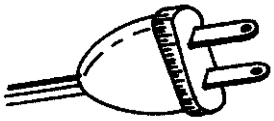

male plug

main path In a computer program, the sequence of instruction execution disregarding the execution of subroutines.

main program The part of a computer program other than a subroutine.

main routine See MAIN PROGRAM.

mains **1.** In a power-distribution center, the lines that supply the entire system. An example is the set of lines leading into a house. **2.** The utility wires and associated outlets in a house or building.

maintenance The process of keeping a system, circuit, or component in operating condition, with minimal down time.

maintenance routine A computer program used by computer service personnel for diagnosis during a regular service interval.

major beats The principal beats produced in a beat-note system; they are usually the sum and/or difference of two fundamental frequencies. Compare MINOR BEATS.

major face In a hexagonal quartz crystal, one of the three larger faces. Compare MINOR FACE.

majority carrier The predominant charge carrier in processed semiconductor material. Electrons are the majority carriers in n-type material; holes are the majority carriers in p-type material. Compare MINORITY CARRIER.

majority logic A logic gate in which the output is high whenever the majority of its inputs is high, regardless of which inputs are high. Thus, in a five-input gate of this type, the output is high when any three or more of the inputs are high.

major lobe See MAIN LOBE.

major loop The principal path for the circulation of information or control signals in an electronic system (e.g., *major feedback loop*). Compare MINOR LOOP.

make **1.** The closing of a pair or set of contacts. **2.** To close a pair or set of contacts.

make-before-break contacts A pair of contacts in which the movable arm closes with the next contact before breaking with the previous one. Compare BREAK-BEFORE-MAKE CONTACTS.

make time The time required for a relay to latch completely, or for a switch (either mechanical or electronic) to close completely. Compare BREAK TIME.

male plug A plug having one or more protruding contacts in the form of pins, blades, or prongs. Compare FEMALE PLUG and HERMAPHRODITIC PLUG.

Malter effect The tendency for a layer of semiconductor having a high secondary emission ratio to become positively charged when bombarded by electrons. This occurs when a thin insulator separates the semiconductor from a metal plate. The insulator must be very thin (on the order of 10^{-7} meters). This results in a potential difference of up to about 100 volts.

manganese Symbol, Mn. A metallic element. Atomic number, 25. Atomic weight, 54.938.

manganese-dioxide depolarizer In a dry cell, manganese dioxide mixed with powdered carbon, the mixture being a depolarizing agent. Also see DEPOLARIZER.

manganin A low-temperature-coefficient alloy used in making wire for precision resistors. A typical composition is: copper (84 percent), manganese (12 percent), and nickel (4 percent).

manipulator A robot arm and end effector, as used in mechanical processes.

man-made interference See HUMAN-MADE INTERFERENCE.

man-made static See HUMAN-MADE INTERFERENCE.

manometer An instrument for measuring gas or vapor pressure—especially at low levels.

manpack A portable radio transceiver that can be used while walking.

mantissa **1.** The portion of a logarithm to the right of the decimal point. Thus, in 3.952502 ($\log_{10}$ 8964), the mantissa is 0.952502. **2.** The fixed point part of a number in scientific notation; thus, in 4×10^3, the mantissa is 4.

manual **1.** Actuated or operated directly by mechanical means, rather than automatically. **2.** A book, or a set of online information files, detailing the operation and maintenance procedures for a device or system.

manual input Use of a keyboard, mouse, trackball, or other electromechanical input device to enter data into a computer program or system.

manual operation In data processing, an operation in which automatic machines are not involved.

manual telegraphy Telegraphy that consists signals transmitted by a hand-operated key and recorded by hand (pen, pencil, or typewriter).

manual tuning Tuning performed entirely by adjusting variable circuit components by hand.

manual word generator A device by which an operator can originate information words for input into computer memory.

manufacturing automation protocol In a factory using computer-controlled robots, the set of standards for data communication between the robots and the controller and/or between individual robots. It keeps the factory operating smoothly.

MAR Abbreviation of MEMORY-ADDRESS REGISTER.

Marconi antenna A quarter-wave radio transmitting or receiving antenna operated against an earth ground.

Marconi effect The undesired tendency of an entire receiving antenna system, including lead-in or feeders, to act as a MARCONI ANTENNA.

margin **1.** A gap or space between two objects, such as adjacent plates of a capacitor. **2.** Clearance. **3.** The maximum error that can be tolerated without risk of improper or abnormal operation. **4.** In a teletypewriter, the range of adjustments in which the error frequency is acceptable.

marginal relay A relay having a small difference between its on and off currents or voltages.

marginal test As performed on equipment in a computer installation, a test to either determine the cause of an intermittent malfunction, or verify an equipment's operating tolerances.

marine broadcast station A coastal station that broadcasts information of interest to shipping: time, weather, ocean currents, etc.

marine radio Radio communications between seagoing vessels or between vessels and shore stations.

marine radiobeacon station A land-based radionavigation station whose transmitted signals are used for taking bearings.

mariner's compass See MAGNETIC COMPASS.

mark **1.** In telegraphy, the dot or dash portion of a character, as opposed to the dead space between such portions. **2.** The intelligence part of a similar signal (such as sound, light, etc.). **3.** The high (logic 1) state represented by a binary bit, as opposed to the low (logic 0) state. **4.** A character identifying the end of a data set. Also called MARKER (see MARKER, **2**).

marker **1.** A pip that indicates a particular frequency on a response curve displayed on an oscilloscope screen. **2.** A character that identifies the end of a data set. Also called MARK (see MARK, **4**).

marker beacons Individual coded-signal transmitters placed along a radio range and indicating features of the course marked by them.

marker frequency **1.** A known frequency that can be used to identify a spot-frequency harmonic of a frequency-standard signal. **2.** A known accurate signal used to identify the limit of a radio band. **3.** The frequency at some point on a response curve as identified by a marker pip (see MARKER).

marker generator An oscillator that supplies a marker pip (see MARKER).

mark hold In telegraphy, an unmodulated signal meaning information is not being sent.

mark reading The reading by an optical scanning device of marks made in specific areas of a document; the process also includes the marks' conversion to digital signals for input to a computer.

mark scanning See MARK READING.

mark sensing A process similar to MARK READING, except that the marks are sensed electrically.

marker trap A wave trap that supplies a dip-type marker pip when used in conjunction with a radio-frequency test oscillator (see MARKER).

market scanner Also called *bar-code reader.* A device that scans a black-bar binary label printed on a carton or other package (or magazine), and indicates the price of the merchandise on the readout of the checkout register.

mark-to-space ratio In radiotelegraphy, the ratio of the duration (mark) of a dot to the interval (space) between successive dots.

Marx generator An impulse-type high-voltage direct-current generator circuit in which several capacitors are charged in parallel through a high-resistance network. When the capacitor voltage reaches a critical high value, discharge occurs in series through spark gaps, producing a high-voltage pulse for each discharge.

maser A low-noise microwave amplifying device in which a microwave input signal causes high-energy-state molecules of ammonia or ruby to fall to the low-energy state and, as a result, to emit large amounts of energy as an output signal. The name is an acronym for *microwave amplification by stimulated emission of radiation.*

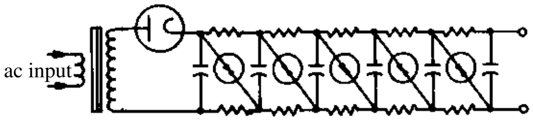

Marx generator

mask **1.** A kind of stencil through which plating, electrodeposition, or diffusion can be done. **2.** The viewing screen, or GRATICULE, of an oscilloscope. **3.** To obliterate a signal with a stronger one. **4.** A bit or character pattern used to change or extract bit positions in another pattern.

masking **1.** The use of a MASK of any type. **2.** The tendency of one effect or phenomenon to obscure another. It applies especially in audio systems, where certain sounds impair the ability of a listener to hear other sounds that occur at the same time. **3.** The extent to which one effect or phenomenon obscures another.

Masonite Masonite Corporation's tough fiberboard used for panels and bases of some electronic equipment.

mass The quantity of matter in a body. Like weight, mass is expressed in kilograms in the metric (SI) system and in pounds in the English system. For a given piece of material, mass can be determined by dividing the weight by the acceleration of gravity.

mass data Data in excess of the maximum amount that can be stored in the main (internal) storage unit of a digital computer (i.e., that which can only be accommodated by external media such as magnetic disks or tapes).

mass-energy equation Energy (E) is the product of a given mass (m) and the square of the speed of light (c^2): $E = mc^2$. It is also called the *Einstein equation.*

mass number **1.** Symbol, A. A number representing the total of neutrons and protons in the nucleus of an atom. The approximate mass of an atom is equal to $A \times m_p$, where m_p is the total proton (rest) mass. **2.** The number indicating the sum of nuclear protons and neutrons in an atom. It is usually written following the symbol for the atom: thus, U238 is uranium having 238 nucleons. An isotope of an element will have a different mass number than that of the normal atom.

mass of electron at rest Symbol, m_e. The amount of matter in an electron; $m_e = 9.1093897 \times 10^{-31}$ kg.

mass of neutron at rest Symbol, m_n. The mass of a neutron in the nucleus of an atom; $m_n = 1.6749286 \times 10^{-27}$ kg.

mass of proton at rest Symbol, m_p. The mass of a proton in the nucleus of an atom; $m_p = 1.6726231 \times 10^{-27}$ kg.

mass resistivity **1.** The resistance of a wire one meter long having a mass of one gram. It varies, depending on the composition of the wire. **2.** The resistance of a wire one mile long having a weight of one pound. It varies, depending on the composition of the wire.

mass spectograph An instrument used to analyze chemical compounds and mixtures in terms of their distinctive mass spectra, exhibited by ionized samples of the materials in a magnetic field.

mass spectrometer Abbreviation, MS. An instrument that permits rapid analysis of chemical compounds via the MASS SPECTRUM.

mass spectrum An electron spectrum that can be used to identify a chemical element. Different elements have nuclei with different charge-to-mass ratios. This results in each element having a unique mass spectrum.

mass storage In a computer system, a magnetic or optical storage medium capable of holding large amounts of data. Examples: magnetic diskette, magneto-optical diskette, external hard disk, compact-disk read-only memory (CD-ROM), and magnetic tape.

mass unit See ATOMIC MASS UNIT.

master **1.** The primary or main element or device in a system. **2.** A primary data medium or recording from which copies are made. **3.** A primary reference standard. See the following several definitions.

master clock **1.** In a digital computer, the primary generator of timing pulses. **2.** A standard time clock that drives other (slave) clocks, or to which clocks of lesser accuracy can be referred.

master console In a computer system, an equipment with panel instruments and controls, which permits operations to be governed, monitored, and controlled by a human operator.

master control **1.** The main control circuit in a system. **2.** A point from which signals or programs are distributed in a communications or broadcast system.

master data Also called *archives.* In a computer record, data elements that remain unaltered for a long time, and from which copies are made.

master file A computer file of data used routinely and remaining unchanged for a long time.

master gain control The principal gain control in an audio amplifier or mixer [i.e., the one used to adjust the gain (volume) of the entire system].

master instruction tape Magnetic tape on which various related computer programs are recorded.

master library tape See MASTER PROGRAM FILE.

master oscillator Abbreviation, MO. The main oscillator in an electronic system (e.g., the oscillator stage in an oscillator-amplifier type of radio transmitter). This oscillator can be either self-excited or crystal-controlled.

master oscillator-power amplifier Abbreviation, MOPA. A type of transmitter or signal generator in which a frequency-determining oscillator drives a power amplifier, which in turn delivers an output signal. Because the oscillator is isolated from the output load, this arrangement has greater stability than one in which the oscillator alone supplies power to the load.

master pattern The etching pattern used for manufacture of a batch of identical printed-circuit boards.

master program file A reel of magnetic tape on which is recorded the programs regularly used in a data-processing installation. It is also called *master library tape.*

master record In a data-processing system, the current record (usually stored on a disk or tape) that will be used for the next computer run.

master relay A relay that operates other (slave) relays. Compare SLAVE RELAY.

master station See KEY STATION.

master switch A switch that can actuate or deactuate an entire installation or system.

master tape **1.** In sound recording and reproduction, a magnetic tape that contains material from which other tapes and discs can be made. **2.** In automation, a magnetic tape on which is recorded the basic signal sequence for controlling a process and other recorders. **3.** In data processing, a magnetic tape that must not be erased.

master volume control See MASTER GAIN CONTROL.

masurium See TECHNETIUM.

MAT Abbreviation of MICROALLOY TRANSISTOR.

match **1.** To mate devices, signals, impedances, etc. for optimum compatibility in terms of signal transfer, equipment interfacing, and other optimizing qualities. **2.** The condition of being compatibly mated, physically or electrically.

matched components Circuit components (capacitors, coils, diodes, resistors, transistors, etc.) that are carefully selected for similar or particularly compatible operating characteristics.

matched filter **1.** A filter with input and output impedances matched to the input line and output load, respectively. **2.** A filter designed for separating a signal with a particular waveform from other signals and noise.

matched impedance A usually non-reactive impedance that has the same value as that of another impedance with which it is operated. Maximum power is transferred between impedances that are matched.

matched load A purely resistive load, the impedance of which is the same as the characteristic impedance of the feed line. This results in optimum power transfer from the line to the load.

matched pair A pair of matched components offered in a single package.

matched transmission line A transmission line terminated in a purely resistive impedance whose value is identical to the characteristic impedance of the line. Such a line transfers all of its energy to its load without reflection; no standing waves are on the line.

matching pad An inductance-capacitance (LC) network for matching the impedance of a load to the output impedance of a signal generator.

matching stub See STUB.

matching transformer An audio-frequency (AF) or radio-frequency (RF) transformer used to match one purely resistive impedance to another.

matchtone A transistorized, single-frequency audio oscillator that can be used to monitor transmitted radiotelegraph signals. The carrier wave from the transmitter is rectified by a small semiconductor diode, whose direct-current output powers the oscillator.

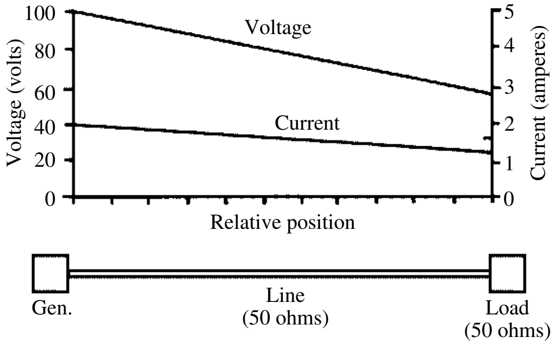

matched transmission line

Mateucci effect The generation of a potential difference in a helically wound, ferromagnetic wire when its magnetization fluctuates.

mathematical check A test of the validity of the result of an arithmetic process (by using alternate methods, for example).

mathematical logic **1.** A branch of mathematics that involves the theoretical behavior of various systems of reasoning. **2.** See BOOLEAN ALGEBRA. **3.** See DIGITAL LOGIC.

mathematical model See MODEL, **2**.

mathematical subroutine Within a computer program, a subroutine serving as an arithmetic function (i.e., one for performing an operation not integral to the monitor program).

matrix **1.** A high-speed switching or memory array used in counters and computers. **2.** Generally, any two-dimensional array of objects. **3.** A device for solving linear simultaneous equations, consisting of a rectangular array of coefficients.

matrix printer See WIRE PRINTER.

mat switch A form of PRESSURE SENSOR used in some security systems. When weight appears on the mat, switches close, actuating an alarm.

matter The building material of the universe that occupies space and has mass that can be measured. See, for illustration, ATOMIC THEORY and STATES OF MATTER.

matter waves See DE BROGLIE WAVES.

max Abbreviation of MAXIMUM.

maxima Points along a curve at which a function reaches a local maximum value. Also see MAXIMA AND MINIMA.

maxima and minima **1.** The loops and nodes of current or voltage on an antenna or transmission line. **2.** The lobes and nulls in a directivity pattern. **3.** The bright and dark bands in a visible-light interference pattern. **4.** In radar reflections, regions of localized maximum and minimum

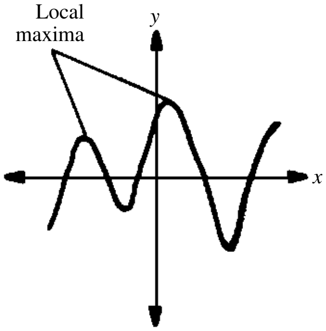

maxima

intensity. **5.** The study and solution of maximum, minimum, and inflection points on the curve of a function.

maximal flatness For an amplifier or network, the condition in which peaks are not present in the normal passband response.

maximum Abbreviation, max. The highest value in a range or set. Also see MAXIMA AND MINIMA and PEAK.

maximum available gain Abbreviation, MAG. The amplification provided by a circuit or device whose input and output impedances are correctly matched to source and load.

maximum current **1.** Symbol, I_m or I_{max}. The highest value reached by an alternating-current half-cycle or by a pulse current. Also called PEAK CURRENT. **2.** The highest value of current in a series of current values.

maximum power **1.** Symbol, P_m or P_{max}. The highest value of power that an equipment can be called upon to supply. **2.** The highest value of power in a series of measurements or calculations.

maximum-power discharge current For a cell or battery, the current at which the greatest amount of power is delivered.

maximum power output See MAXIMUM POWER, **1**.

maximum power transfer The condition in which the largest amount of power is delivered by a source to a load.

maximum power transfer theorem Maximum power is transferred from a generator to a load when the impedance of the load equals the internal impedance of the generator. Compare COMPENSATION THEOREM, NORTON'S THEOREM, RECIPROCITY THEOREM, SUPERPOSITION THEOREM, and THEVENIN'S THEOREM.

maximum rating **1.** The highest value of a quantity (e.g., current, voltage, or power) that can safely be used with a given device. **2.** The highest value of a quantity afforded by a given device (e.g., *maximum capacitance* of a variable capacitor).

maximum record level **1.** In a magnetic tape, magnetic disk, or phonograph disc, the highest amplitude of input signal that can be recorded with an acceptable amount of distortion. **2.** The recording-head current or power that results in third-harmonic distortion of three percent.

maximum signal level **1.** In an amplitude-modulated signal, the peak power. **2.** In an amplitude-modulated facsimile or television system, the amplitude that results in a black or white picture (depending on whether the highest amplitude produces black or white).

maximum undistorted power output Abbreviation, MUPO. The highest power that an active amplifying device will deliver before significant distortion occurs.

maximum usable frequency Abbreviation, MUF. The highest frequency that can be used successfully at a given time, between two specific geographic locations, for communication via the ionosphere.

maximum voltage **1.** Abbreviation, E_m, E_{max}, V_m, or V_{max}. The peak value reached by an alternating-current voltage half-cycle, or by a voltage pulse. **2.** The highest value of voltage in a series of voltage measurements or calculations.

maximum wattage See MAXIMUM POWER.

maxterm form In mathematical calculations, the factored form of a function, expressed as a product of sums. For example, the maxterm form of $f(x) = x^2 + 5x + 6$ is $f(x) = (x + 2)(x + 3)$.

maxwell Symbol, Mx. The cgs unit of magnetic flux, equivalent to one line of flux or 10^{-8} weber.

Maxwell bridge A four-arm alternating-current bridge for measuring inductance against a standard capacitance.

Maxwell's equations A set of four equations developed by James Clerk Maxwell in 1864 and 1873, describing vector quantities pertaining to points in space subjected to varying electric and magnetic forces. Through his classic presentation, Maxwell

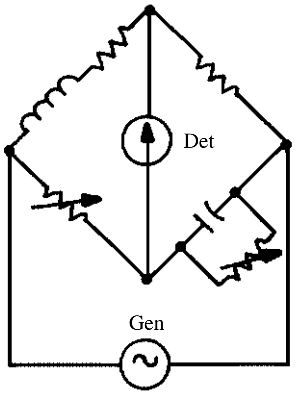

Maxwell bridge

predicted the existence of electromagnetic waves, whose later discovery made radio possible.

Maxwell's law Also called *Maxwell's rule.* Every part of an electric circuit is acted upon by a force tending to move it in the direction that results in the maximum magnetic flux being enclosed.

maxwell-turn A unit of magnetic coupling (linkage) equal to 1 maxwell per turn of wire in a coil linked by magnetic flux. Also see MAXWELL.

mayday In radiotelephony, a word spoken as an international distress signal equivalent to SOS in radiotelegraphy. The word is the phonetic equivalent of the French *m'aidez* (help me).

MB Abbreviation of MIDBAND.

Mb Abbreviation of MEGABAR.

MBB Abbreviation of MAKE BEFORE BREAK.

MBM Abbreviation of *magnetic bubble memory.*

MBO Abbreviation of MONOSTABLE BLOCKING OSCILLATOR.

MBS Abbreviation of *magnetron beam switching.*

mc **1.** Symbol for MILLICURIE (mCi is preferred). **2.** Symbol for METER-CANDLE.

Mc **1.** Symbol for MEGACURIE (MCi is preferred). **2.** Obsolete abbreviation of *megacycle(s)*, a term superseded by MEGAHERTZ.

MCG Abbreviation of MAGNETOCARDIOGRAM.

McLeod gauge An instrument for measuring gas under low pressure. A measured volume of the gas under test is first compressed (to a lower known volume) to a pressure more easily measured via a mercury manometer and the application of Boyle's law.

MCM Abbreviation of MONTE CARLO METHOD.

McProud test A simple test for checking the tracking efficiency of a phonograph pickup and arm for microgroove discs.

Mc/s Obsolete abbreviation of *megacycle(s) per second*, a term superseded by MEGAHERTZ.

MCS **1.** Abbreviation of *Master of Computer Science.* **2.** Abbreviation of *missile control system.*

MCW Abbreviation of MODULATED CONTINUOUS WAVE.

Md Symbol for MENDELEVIUM.

MDAS Abbreviation of *medical data acquisition system.*

m-derived filter A filter whose inductance (L) and capacitance (C) values are derived by multiplying those of a constant-k filter by a factor m between zero and 1. This factor is a function of the ratio f_i/f_c, where f_i is the frequency of infinite attenuation, and f_c is the cutoff frequency. This type of filter exhibits sharper response than the equivalent constant-k filter.

MDI Abbreviation of MAGNETIC DIRECTION INDICATOR.

M-display See M-SCAN.

MDS Abbreviation of MINIMUM DISCERNIBLE SIGNAL.

m_e Symbol for MASS OF ELECTRON AT REST.

Meacham oscillator A highly stable radio-frequency oscillator consisting of an amplifier provided with a feedback circuit containing a four-arm bridge, one arm of which is a quartz crystal, and another, a tungsten-filament lamp acting as a nonlinear resistor. Also called *bridge-stabilized oscillator.*

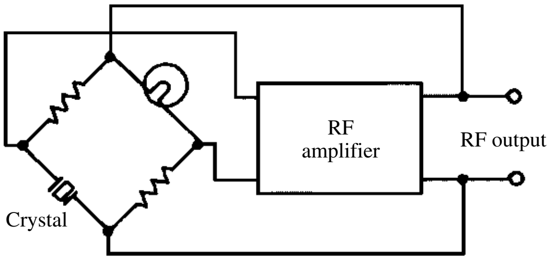

Meacham oscillator

mean **1.** A general term meaning *average.* **2.** See ARITHMETIC MEAN. **3.** See GEOMETRIC MEAN.

mean charge **1.** In an object that is nonuniformly charged, the average charge per unit distance, area, or volume. **2.** In a capacitor carrying a fluctuating current, the average amount of charge held by the plates.

mean free path **1.** In acoustics, the average distance that sound waves travel before striking a barrier or reflecting surface. **2.** The average distance that sound waves travel between reflections (echoes) in a chamber. **3.** In a gas tube, the average of all the free paths of electrons at a specified temperature.

mean life **1.** Symbol, L. The average life of a radioactive substance [i.e., the time taken for $1/e$ (e = base of natural logarithms) of the substance to disintegrate]. **2.** The time required for excess carriers injected into a semiconductor to recombine with carriers of opposite sign. Also called *average life.*

mean proportional See GEOMETRIC MEAN.

mean time before failure Abbreviation, MTBF. The average length of time that a component or system will perform before the first failure occurs. It is generally specified in hours.

mean time between failures Abbreviation, MTBF. The average length of time that a component or system will perform before failure occurs—either initially or after repair or replacement. It is generally specified in hours.

measured A quantity that is presented to an instrument for measurement.

measured service Any service in which charges are assessed per unit-time usage block. Online computer services are a common example. In some cases, other factors, such as distance, affect the cost per unit time; most long-distance telephone services fall into this category.

measurand Any quantity that is measured with an instrument.

measurement **1.** The process by which the magnitude, extent, or duration of a parameter is found. **2.** The value of a parameter, as obtained, according to **1**.

measurement error The difference between the measured value of a quantity and its true value. Also see NEGATIVE ERROR OF MEASUREMENT and POSITIVE ERROR OF MEASUREMENT.

measurement range In a measuring device, the range within which the error is smaller than a specified value.

mechanical analogs Familiar mechanical devices, systems, or effects with which certain electrical counterparts can be compared for ease in teaching or understanding (e.g., inductance compared with mass, capacitance with elasticity, voltage with pressure, and current with velocity).

mechanical axis In a quartz crystal, the axis perpendicular to the faces of the hexagon. Also see Y-AXIS, **2**.

mechanical bandspread Bandspread tuning obtained by reduction-ratio gearing of the tuning mechanism. Compare ELECTRICAL BANDSPREAD.

mechanical bias **1.** A steady pull applied by a spring to the armature of a relay to sensitize it by decreasing the distance that the armature must move to close the contacts. **2.** Bending of a relay frame to position the armature closer to the magnet for the purpose defined in **1**.

mechanical damping Damping action obtained entirely by mechanical devices (such as weights, dashpots, etc.).

mechanical equivalent of heat The amount of mechanical work required to produce a unit quantity of heat. For example, 4.183 joules can be converted into 1 calorie of heat.

mechanical equivalent of light The expression of luminous energy in equivalent power units. In practical measurements, this is taken as the total power output of a lamp minus the power absorbed by a transparent jacket used to remove the infrared and ultraviolet rays.

mechanical filter See ULTRASONIC FILTER, **1**.

mechanical joint A union of electrical conductors consisting exclusively of a junction or splice made without brazing, soldering, or welding.

mechanical load An electromechanical device that uses the output of an electrical source. Such devices include actuators, brakes, clutches, meters, motors, and relays.

mechanical rectifier A vibrator or commutator used to change an alternating current into a direct current by selecting and passing only positive or negative half-cycles. Also see ELECTROMECHANICAL RECTIFIER.

mechanical scanner **1.** A mechanical device for scanning an object or scene and breaking it into horizontal lines that are converted to signals. **2.** A device that scans the reproducer lamp in a mechanical television receiver. See, for example, NIPKOW DISK.

mechanical switch A switch actuated by moving or sliding a lever, pressing a button, or otherwise applying mechanical pressure.

mechanical time constant For a torque motor, the ratio of moment of inertia to damping factor. Compare ELECTRICAL TIME CONSTANT.

mechanical wave filter See ULTRASONIC FILTER, **1**.

mechanics The branch of physics concerned with forces and motion and the laws of gases and liquids. It is subdivided into *kinematics* and *kinetics*.

mechatronics Combination of the words *mechanics* and *electronics*, referring to the use of electromechanical devices (especially robots) in manufacturing. The term was originally coined in Japan.

median **1.** The middle value in a sequence of numbers. For example, in the series: 1, 2, 3, 4, 5, 6, 7, the median is 4. Compare ARITHMETIC MEAN and GEOMETRIC MEAN. **2.** In a statistical distribution, the value s in the domain so that the area under the curve for all values less than s is equal to the area under the curve for all values greater than s.

medical electronics See ELECTROMEDICAL ENGINEERING.

medical robot **1.** A robot used in a doctor's office, or in a hospital to assist doctors and nurses. There are various applications, some of which have provoked controversy (e.g., *robotic surgical assistant*). It generally performs simple, noncritical tasks. It has been suggested as a means of entertaining hospital patients—especially children. **2.** See BIOMECHANISM.

medium In a computer system, that storage device onto or into which data is recorded for input into memory (e.g., magnetic disk, magnetic tape, optical disk, etc.).

medium-frequency Abbreviation, MF. Pertaining to frequencies in the range 300 kHz to 3 MHz, representing wavelengths from 1000 meters to 100 meters.

medium of propagation The substance (or vacuum) through which electromagnetic energy is transmitted (e.g., outer space, the atmosphere, or a dielectric material).

medium-scale integration A method of manufacturing integrated circuits, in which there are at least 10, but less than 100, individual gates on each chip. Abbreviated MSI.

medium-scan television A television (TV) communications medium in which the scanning rate is slowed down compared to regular (fast-scan) TV, but is faster than the commonly used slow-scan TV. It provides some conception of motion, although not as realistic as fast-scan TV.

medium tension Medium voltage. A relative term, but generally referring to common alternating-current utility voltage (e.g., 117 volts or 234 volts).

medium wave Abbreviation, MW. Pertaining to wavelengths corresponding to medium frequencies (see MEDIUM-FREQUENCY) (i.e., those in the 100- to 1000-meter range).

meg **1.** Colloquialism for MEGOHM(s). **2.** Colloquialism for MEGABYTE(s).

mega- Abbreviation, M. **1.** A prefix meaning *million(s)*, (i.e., 10^6 or 1,000,000). **2.** In digital data applications, a prefix meaning 2^{20} or 1,048,576.

megabar Abbreviation, Mb. A cgs unit of high pressure. 1 Mb = 10^6 bars = 10^{11} pascals. Also see BAR, **1**.

megabit A unit of digital data equal to 2^{20} (1,048,576) bits. Also see BIT.

megabyte Abbreviation, M or MB. A unit of digital data equal to 2^{20} (1,048,576) bytes. Also see BYTE.

megacurie Abbreviation, MCi. A large unit of radioactivity equal to 3.71×10^6 disintegrations per second; 1 MCi = 10^6 curies. Also see CURIE.

megacycle See MEGAHERTZ.

megaelectronvolt Abbreviation, MeV. A large unit of electrical energy; 1 MeV = 10^6 eV. Also see ELECTRONVOLT.

megahertz Abbreviation, MHz. A unit of frequency; 1 MHz = 10^6 Hz = 1,000,000 Hz.

megampere Abbreviation, MA. A unit of high current; 1 MA = 10^6 A = 1,000,000 A.

megaphone **1.** A hand-held microphone/amplifier/loudspeaker used to amplify the voice of a person who must be heard over an appreciable area. **2.** A simple horn for amplifying the voice.

megarutherford Abbreviation, Mrd. A large unit of radioactivity equal to 1 trillion (10^{12}) disintegrations per second; 1 Mrd = 10^6 rd = 1,000,000 rd. Also see RUTHERFORD.

megavolt Abbreviation, MV. A unit of extremely high voltage; 1 MV = 10^6 V = 1,000,000 V.

megavolt-ampere Abbreviation, MVA. A unit of extremely high reactive power; 1 MVA = 10^6 VA = 1,000,000 VA. Also see VOLT-AMPERE.

megawatt Abbreviation, MW. A unit of high power; 1 MW = 10^6 W = 1,000,000 W. Also see WATT.

megawatt-hour Abbreviation, MWh. A large unit of electrical energy or of work; 1 MWh = 10^6 Wh = 1,000,000 Wh = 3.6×10^9 joules. Also see WATT-HOUR.

megger An instrument containing an internal high-voltage direct-current power supply, used for measuring high values of resistance. Compare MEGOHMMETER.

meg-mike **1.** Colloquialism for MEGOHM-MICROFARAD(s). **2.** Colloquialism for MEGOHM-FARAD(s).

megohm Symbol, M. A unit of high resistance, reactance, or impedance; 1 M = 10^6 ohms = 1,000,000 ohms.

megohm-farads For a large capacitor, the product of leakage resistance (megohms) and capacitance (farads). Also see MEGOHM-MICROFARADS.

megohmmeter A special ohmmeter for measuring resistances in the megohm range.

megohm-microfarads For a capacitor, the product of leakage resistance (megohms) and capacitance (microfarads). The figure is an expression for the relative insulation resistance of a capacitor.

Meissner circuit An oscillator tuned by means of LECHER WIRES (parallel-conductor resonant circuits). It is used primarily at ultra-high frequencies (UHF).

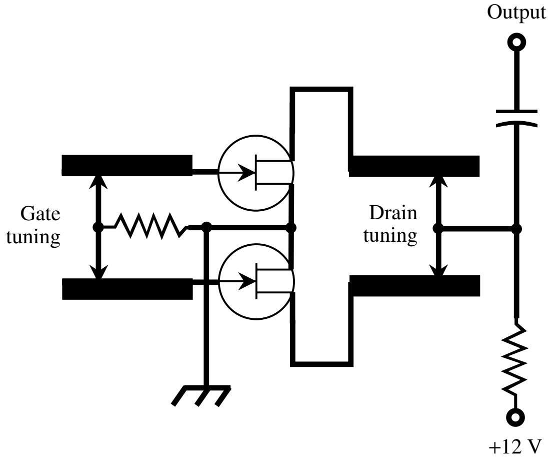

Meissner circuit

Meissner effect In a superconductive material, the abrupt loss of magnetism when the temperature of the material is reduced to a value below that required for superconductivity.

Meissner oscillator See MEISSNER CIRCUIT.

meitnerium Symbol, Mt. Also called *unnilenium* (Une). Atomic number, 109. The most common isotope has atomic weight 266. Classified as a transition metal. It is human-made and not known to occur in nature.

mel An expression of apparent or perceived sound pitch. A tone of 1 kHz, at a level of 40 dB, with respect to the threshold of hearing, represents 1 mel. The perceived pitch depends, to some extent, on the intensity of the sound, as well as on the actual frequency.

M electron In certain atoms, one of the electrons whose orbits are outside of and nearest to those of the L electrons.

meltback process The technique of remelting a doped semiconductor material and allowing it to refreeze to form a grown junction.

meltback transistor A grown-junction transistor produced by the MELTBACK PROCESS.

melting point Abbreviation, mp. The temperature at which a solid starts becoming liquid at a pressure of one atmosphere. Compare FREEZING POINT and MIXTURE MELTING POINT.

memory **1.** The section of a digital computer that records and holds data until it is necessary. In personal computers, the term generally refers to RANDOM-ACCESS MEMORY and READ-ONLY MEMORY, contained in integrated circuits (ICs). Compare STORAGE. **2.** See MEMORY DRAIN.

memory address register In computer storage, a register in which is stored the address of operands in other locations.

memory area A portion of computer memory reserved for a specific type of data. Also called *area.*

memory capacity As a function of the number of memory locations available, the number of bytes that can be stored. It is usually specified in kilobytes, megabytes, or gigabytes. Also see GIGABYTE, KILOBYTE, and MEGABYTE.

memory cycle **1.** The period of execution of a sequence of operations. **2.** The complete operational cycle for inputting data to memory or retrieving it.

memory dialing In a telephone set, a feature that allows rapid dialing of stored digits. The simplest version is the “redial” feature, in which the most recently dialed number is rapidly dialed at the touch of a button. Some sets can store several different numbers, usually including area codes, and sometimes country codes as well.

memory drain Also called *battery memory.* A phenomenon occasionally exhibited by nickel–cadmium cells and batteries, in which the useful ampere-hour capacity is reduced even though the unit is not physically damaged. The depth of the charge cycle decreases to a fraction of its rated value. The problem can usually be overcome by discharging the cell or battery fully, then recharging fully, and repeating the process several times. See also NICKEL–CADMIUM.

memory dump In computer operations, to either print out what is stored in some of or all of the memory locations or transfer the data from a bank of memory cells to some external storage medium.

memory effect See MEMORY DRAIN.

memory guard In a computer, hardware or software that keeps certain memory locations from being addressed by a program being run.

memory location In a computer memory, a place where an information unit (word or character) can be stored; the stored information can be retrieved by appropriate addressing instructions.

memory organization packets In artificial intelligence (AI) and expert systems, a method of arranging computer memory into general rules or statements. The statements are used by software to derive models, forecasts, diagnoses, etc.

memory power Computer memory efficiency in terms of data processing (cycle) speed.

memory protection A hardware device in a multiple programming computer that prevents programs from being altered by other operating programs in the installation.

memory register In a digital computer, a register used in all instruction and data transfers between the memory and other sections of the machine.

memory unit See MEMORY.

mendelevium Symbol, Md. A radioactive element produced artificially. Atomic number, 101. Atomic weight, 258 (approx.).

menu In computer operations, a list of commands for using various functions of the system.

MEP Abbreviation of *mean effective pressure.*

mercuric iodide Formula, HgI_2. A compound whose crystals are useful at room temperature as detectors in high-resolution gamma-ray spectroscopy.

mercury Symbol, Hg. A metallic element. Atomic number, 80. Atomic weight, 200.59. The only metal that is liquid at room temperature. It is used extensively in switches, certain high-voltage rectifiers, high-vacuum pumps, and thermometers.

mercury arc The arc discharge occurring in mercury vapor between solid or liquid (mercury) electrodes. The discharge emits ultraviolet radiation.

mercury-arc rectifier A heavy-duty rectifier tube utilizing ionized mercury vapor. The two general types are MERCURY-VAPOR RECTIFIER and MERCURY-POOL RECTIFIER.

mercury battery Also called *mercuric-oxide battery.* A set of two or more mercury cells stacked one atop the other, electrically connected in series. The resulting battery has a cylindrical shape. A set of four cells provides approximately 5.4 volts under no-load conditions; a battery of seven cells provides 9.5 volts; a battery of nine cells provides 12 volts. See MERCURY CELL.

mercury cadmium telluride Formula HgCdTe. An alloy used as a semiconductor in certain transistors, integrated circuits, and infrared detectors.

mercury cell Also called *mercuric-oxide cell.* An electrochemical cell having a button-like shape, small enough to fit inside a wristwatch. The unit is housed in a steel container and has a mercuric-oxide cathode, amalgamated-zinc anode, and potassium hydroxide and zinc-oxide electrolyte. The potential difference under no-load conditions is 1.35 volts, with a high ratio of stored energy per unit mass. The cell has a flat discharge curve; the voltage remains essentially constant until the charge is almost depleted, and then the voltage drops rapidly. There has been a decrease in the use of mercury cells and batteries in recent years, because mercury is toxic. Unless these cells are discarded in a special way, the mercury from them can cause dangerous contamination of soil and water.

mercury delay line A delay line in which delay is obtained by propagating the signal through a pipe of mercury.

mercury diffusion pump A vacuum diffusion pump using mercury vapor.

mercury displacement relay A form of switching relay in which the electrical contact is made by moving mercury.

mercury-jet switch A multipoint switch using a jet of mercury instead of the conventional wiper arm, for high-speed operation and reduced wear.

mercury memory A recirculating memory using a mercury delay line. Also see DELAY LINE and DELAY-LINE MEMORY.

mercury-pool cathode In certain industrial electron tubes, such as the ignitron, a cathode electrode consisting of a pool of mercury.

mercury-pool rectifier A type of mercury-arc rectifier whose cathode is a pool of mercury. In one type, the arc is initiated by tilting the tube momentarily to bring the mercury into contact with a third electrode, thus causing a starting current to flow through the pool. In another type, the *ignitron*, a starter electrode is in continual contact with the mercury.

mercury pump See MERCURY DIFFUSION PUMP.

mercury rectifier See MERCURY-POOL RECTIFIER and MERCURY-VAPOR RECTIFIER.

mercury relay A relay in which at least one of the contacts is mercury.

mercury storage See MERCURY MEMORY.

mercury switch A switch consisting essentially of two or more stiff wire electrodes and a drop of mercury hermetically sealed in a glass tube. Tilting the tube causes the mercury to flow toward one end, where it immerses the electrodes, providing a conductive path between them.

mercury-vapor lamp A glow lamp emitting blue-green light that causes ionization of mercury vapor by an electric current.

mercury-vapor rectifier A tube-type high-voltage diode rectifier containing a small amount of mercury that vaporizes and ionizes during tube operation.

mercury-vapor tube **1.** See MERCURY-VAPOR LAMP. **2.** See MERCURY-VAPOR RECTIFIER.

mercury-wetted reed relay A reed relay in which the reeds are wetted with mercury in a pool by capillary action. The film of mercury forms a tiny bridge when the reeds open; when this bridge separates, a clean, high-speed break occurs without contact bounce. Compare DRY-REED SWITCH.

merge **1.** In computer operations, to make a single set or file from two or more record sets. **2.** In word processing, to create a corrected master recording from two input media: the original master recording and the recording that contains the corrections.

meridian **1.** A great circle passing through earth's geographic poles and a given point on the surface of the earth. **2.** A line of longitude on a map or globe. Also see TIME ZONE, ZERO MERIDIAN, and ZONE TIME.

mesa A flat-topped, protruding region in a semiconductor wafer. The mesa is produced by etching the surrounding part of the material. Some bipolar transistors are manufactured in this way.

mesa diffusion A method of manufacturing bipolar transistors. The different semiconductor materials are first diffused together. Then part of the resulting wafer is etched away, resulting in a mesa shape.

mesa transistor A diffused planar transistor in which the silicon area around the base has been etched away to reduce collector-to-base capacitance; the base-emitter region remains elevated like a high plateau (mesa).

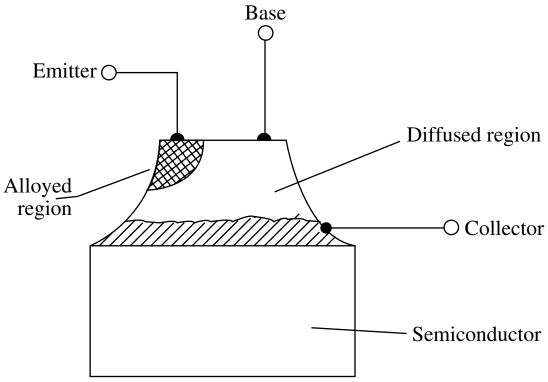

mesa transistor

MESFET A form of field-effect transistor combining depletion-mode and enhancement-mode properties. A Schottky barrier forms the gate electrode.

mesh **1.** A combination of the elements that form a closed path in a network. **2.** The closed figure (such as the delta or star) obtained by connecting polyphase windings together. **3.** A grid, screen, or similar structure in a vacuum tube.

mesh equations Equations describing fully the current and voltage relations in a network of meshes (see MESH, **1**).

Mesny circuit A push-pull ultra-high-frequency (UHF) oscillator whose gate or base tank is a pair of parallel wires short-circuited by a slider; the drain or collector tank is a similar pair of wires. The frequency is varied by moving the sliders along the wires.

mesochronous A condition for signals in which significant instants pass at identical average speeds, such as bits per second.

meson An unstable nuclear particle first observed in cosmic rays. A meson can be electrically positive, negative, or neutral. Its mass lies between that of the electron and proton.

mesotron See MESON.

message **1.** In communications, a body of information sent from a source (transmitter) to a destination (receiver). **2.** Data entered into a transaction-processing system.

message exchange In a digital communications channel, a hardware unit that carries out certain switching functions that would otherwise have to be done by a computer.

message switching system A data communications system having a central computer that receives messages from remote terminals, stores them, and transfers them to other terminals as needed.

metadyne See DC GENERATOR AMPLIFIER.

metal An elemental material that exhibits several familiar properties (such as luster, ductility, malleability, good electrical and heat conductivity, relatively high density, and the ability to emit electrons). Common examples are aluminum, copper, gold, lead, and silver. Compare METALLOID and NONMETAL.

metal-base transistor A bipolar transistor in which the base is a metal film, and the emitter and collector are films of n-type semiconductor material.

metal-ceramic construction The building of certain electronic components by bonding ceramic parts to metal parts. Also see CERMET.

metal-film resistor A fixed or variable resistor in which the resistance element is a film of a metal alloy deposited on a substrate such as a plastic or ceramic.

metal finder See METAL LOCATOR.

metallic binding forces In a crystal, the binding electrostatic force between cations and electrons. Also called *electron-gas binding forces.*

metallic bonding See BONDING, **1** and METALLIC BINDING FORCES.

metallic circuit A circuit, such as a two-wire telephone line, in which earth ground is not a part of the circuit. Compare GROUND-RETURN CIRCUIT.

metallic crystal A crystal substance in which positive ions and free electrons exist; it is, therefore, a good electrical conductor.

metallic insulator A short-circuited quarter-wave section of transmission line that acts as an insulator at the quarter-wavelength frequency.

metallicize To make a circuit fully metallic, as when two wires are used instead of one wire and a ground connection. (Not to be confused with METALLIZE.)

metallic rectifier A dry rectifier using a metal disk or plate coated with a material (such as selenium, an oxide, or a sulfide).

metallic tape Recording tape made from metal, rather than from plastic. Noted for its excellent audio-reproduction characteristics.

metallize To treat, coat, or plate with a metal. (Not to be confused with METALLICIZE.)

metallized capacitor A capacitor in which each face of a dielectric film is metallized to form conductive plates.

metallized-paper capacitor A paper-dielectric capacitor whose plates are metal areas electrodeposited on each side of a paper film.

metallized-polycarbonate capacitor A fixed capacitor in which the dielectric is a polycarbonate plastic film, and the plates are metal areas electrodeposited on each face of the film.

metallized resistor See METAL-FILM RESISTOR.

metal locator An electronic device for locating metal deposits, pipes, or wires underground, in walls, or under floors. It operates via the disturbance that these objects cause to a radio-frequency or magnetic field.

metalloid An element that has some of the properties of a metal. Examples of metalloidal elements widely used in electronics are antimony, arsenic, germanium, silicon, and tin.

metal master See ORIGINAL MASTER.

metal negative See ORIGINAL MASTER.

metal-oxide resistor A resistor in which the resistance material is a film of tin oxide deposited on a substrate.

metal-oxide semiconductor field-effect transistor Abbreviation, MOSFET. A field-effect transistor in which the gate electrode is not a pn junction (as in the junction field-effect transistor), but a thin metal film insulated from the semiconductor channel by a thin oxide film. Gate-control action is entirely electrostatic. Also called *insulated-gate field-effect transistor.* Also see DEPLETION-TYPE MOSFET, DEPLETION-ENHANCEMENT-TYPE MOSFET, and ENHANCEMENT-TYPE MOSFET.

metal-oxide varistor A VOLTAGE-DEPENDENT RESISTOR in which the resistance material is a metallic oxide, such as zinc oxide.

metal-plate rectifier See METALLIC RECTIFIER.

metal tube A vacuum tube housed in a metal envelope for self-shielding and mechanical ruggedness.

metamer A visible-light beam that is identical in color (hue), but different in concentration (saturation), with respect to a reference color.

meteor-burst signals Momentary signals, or increases in signal strength, resulting from reflection of electromagnetic energy from meteor ionization trails. See METEOR SCATTER, **1**.

meteor ionization trail A cloud of ions left in the upper atmosphere as a meteor passes. This cloud tends to reflect radio signals at certain frequencies for a short period of time. During a meteor shower, there could be a sufficient number of such trails to allow continuous over-the-horizon communication when other over-the-horizon modes are unusable.

meteorograph An instrument for the simultaneous measurement of various meteorological phenomena such as temperature, humidity, etc.

meteorology The science of the atmosphere, especially the study of weather and climate. (Not to be confused with METROLOGY.)

meteor-scatter propagation The reflection of radio signals from the ionized trails produced by meteors as they pass through the upper atmosphere. This can result in over-the-horizon radio communication or reception. A meteor produces a trail that persists for a few tenths of a second up to several seconds, depending on the size of the meteor, its speed, and the angle at which it enters the atmosphere. This is not sufficient time for the transmission of very much information, but during a meteor shower, ionization can be almost continuous. Meteor-scatter propagation has been observed at frequencies considerably above 30 MHz.

meteor-trail reflections Momentary reflection of signals by the ionized trails of meteors passing through a signal path.

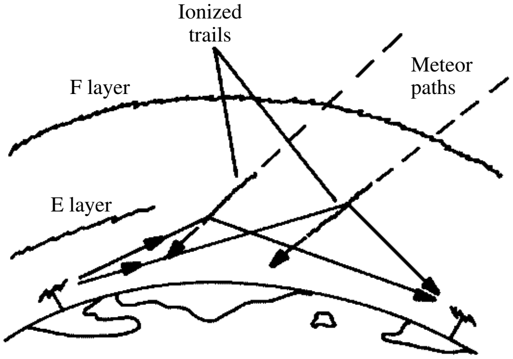

meteor-trail reflections

meter **1.** An instrument for measuring and indicating the value of a particular quantity. See, for example, CURRENT METER and VOLTMETER. **2.** Abbreviation, m. A unit of linear measure and of electrical wavelength, equivalent to 1.65076373×10^6 wavelengths (in a vacuum) of the radiation corresponding to the transition between the two levels of the krypton-86 atom, and approximately equal to 39.37 inches. **3.** To supply in specific increments or by a governed amount.

meter alignment See VISUAL ALIGNMENT.

meter-ampere A unit of transmitted radio signal intensity. Determined by multiplying the antenna current (in amperes) by the height (in meters) of the antenna above ground.

meter-candle Abbreviation, mc. A metric unit of illuminance, equivalent to the illumination on a surface 1 meter from a light source of 1 candle power. Compare FOOT-CANDLE and LUX.

meter equivalent The number of meters equal to a given English measure of length (e.g., the meter equivalent of 3 feet is approximately 0.9144).

meter-kilogram-second Abbreviation, mks. The system of units in which the meter is the standard unit of length, the kilogram is the standard unit of mass, and the second is the standard unit of time. Compare CENTIMETER-GRAM-SECOND and INTERNATIONAL SYSTEM OF UNITS.

meter multiplier See MULTIPLIER RESISTOR.

meter protector A nonlinear resistor, such as a varistor or semiconductor diode, used to prevent overswing in an electric meter by limiting the current flowing through it.

meter rating The maximum reading on a meter, at or below which the accuracy is within a specified limit, but above which the error might exceed that limit.

meter rectifier A light-duty semiconductor diode or bridge circuit, used to change alternating current (ac) to direct current (dc) for deflection of a D'Arsonval-type dc milliammeter or microammeter.

meter relay A sensitive relay that is essentially a moving-coil meter, whose pointer closes against a stationary contact mounted at some point along the scale.

meter resistance Symbol, R_m. The internal resistance of an electric meter. In a simple D'Arsonval meter, it is the resistance of the movable coil. In more-complicated meter circuits, it is the resistance of the parallel combination of the coil and METER SHUNT.

meter scale factor See SCALE FACTOR, **1**.

meter sensitivity See VOLTMETER SENSITIVITY.

meter shunt A resistor connected in parallel with an ammeter, milliammeter, or microammeter to increase the range of currents that the device can measure.

meter torque See DEFLECTING TORQUE.

meter-type relay See METER RELAY.

methyl methacrylate resin Also known by the trade name *Lucite*. A plastic insulating material. Dielectric constant, 2.8 to 3.3. Dielectric strength, 20 kV/mm.

metre Abbreviation, m. Alternate spelling of *meter* when used to specify displacement or wavelength. See METER, **2**.

metric system The decimal system of weights and measures based on the meter, kilogram, and second. Also see CENTIMETER-GRAM-SECOND, METER-KILOGRAM-SECOND, and INTERNATIONAL SYSTEM OF UNITS.

metric ton Abbreviation, MT. A metric unit of weight equal to 1000 kilograms or 1.1023 English tons.

metric waves British designation for electromagnetic energy having wavelengths ranging from 10 meters down to 1 meter, corresponding to frequencies from 30 MHz up to 300 MHz.

metrology The science of weights and measures, including electrical standards and electronic instruments and measurements. (Not to be confused with METEOROLOGY.)

metronome A mechanical or electronic device that produces audible beats (ticks). It is commonly used in setting the tempo for music, and for audibly timing certain processes.

MeV Abbreviation of MEGAELECTRONVOLT(s).

MEW Abbreviation of MICROWAVE EARLY WARNING.

mF Abbreviation of MILLIFARAD.

MF **1.** Abbreviation of MEDIUM FREQUENCY. **2.** Abbreviation of MIDFREQUENCY.

MFD Abbreviation of MAGNETOFLUID DYNAMICS (see MAGNETOHYDRODYNAMICS).

MFSK Abbreviation of MULTIPLE FREQUENCY-SHIFT KEYING.

Mg Symbol for MAGNESIUM.

mg Abbreviation of MILLIGRAM.

MGD Abbreviation of MAGNETOGASDYNAMICS (see MAGNETOHYDRODYNAMICS).

MHD Abbreviation of MAGNETOHYDRODYNAMICS.

MHD generator See MAGNETOHYDRODYNAMIC GENERATOR.

MHD gyroscope See MAGNETOHYDRODYNAMIC GYROSCOPE.

MHD power generation See MAGNETOHYDRODYNAMIC GENERATOR.

mho Obsolete term for the standard unit of electrical conductance. See SIEMENS.

mhp Abbreviation of MILLIHORSEPOWER.

MHz Abbreviation of MEGAHERTZ.

mi Abbreviation of MILE. (Also, m.)

MIC Abbreviation of MICROWAVE INTEGRATED CIRCUIT.

mic Abbreviation of MICROPHONE.

mica A dielectric mineral of complex silicate composition, easily separated into numerous thin, transparent sheets. It is widely used as a capacitor dielectric and high-temperature electrical insulator. Dielectric constant, 2.5 to 7. Dielectric strength, 50 to 220 kV/mm. Also see MUSCOVITE.

mica capacitor A component that is made by alternately stacking metal sheets and layers of mica, or by applying silver ink to sheets of mica. The metal sheets are wired together into two meshed sets, forming the two terminals of the capacitor. This type of capacitor is noted for low loss. Voltage ratings can be up to several thousand volts if thick sheets of mica are used. But these capacitors are physically bulky in proportion to their capacitance. The main application is in radio receivers and transmitters. Compare CERAMIC CAPACITOR, ELECTROLYTIC CAPACITOR, PAPER CAPACITOR, PLASTIC-FILM CAPACITOR, TANTALUM CAPACITOR.

MICR Abbreviation of MAGNETIC INK CHARACTER RECOGNITION.

micro- **1.** A prefix meaning MILLIONTH(S) (i.e., 10^{-6}). **2.** A prefix meaning extremely small (as in microstructure). Compare MACRO-.

microalloy diffused transistor Abbreviation, MADT. A MICROALLOY TRANSISTOR having a uniform base region that is diffused into the wafer before the emitter and collector electrodes are produced by alloying.

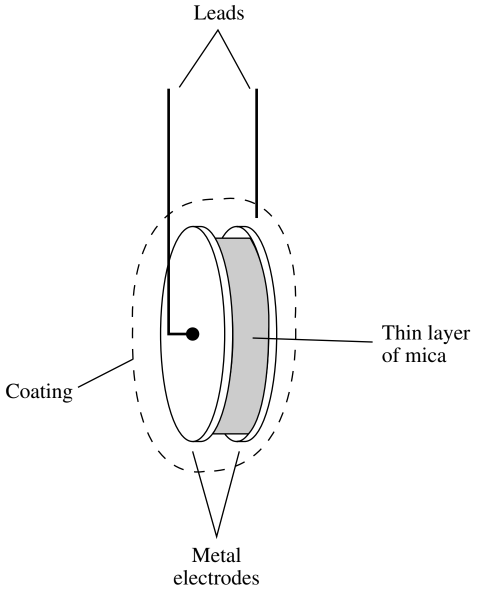

mica capacitor

microalloy transistor Abbreviation, MAT. A transistor having tiny emitter and collector electrodes that are formed by alloying a thin film of impurity material with a collector pit and emitter pit facing each other on opposite surfaces of the semiconductor wafer. Also see SURFACE-BARRIER TRANSISTOR.

microammeter A usually direct-reading instrument used to measure current in the microampere range. Also see CURRENT METER.

microampere A small unit of current, equal to 10^{-6} (0.000001) A.

microbalance A sensitive electronic weighing device. One type uses one or more servo amplifiers for the balancing operation.

microbar A cgs unit of low pressure, equal to 10^{-6} b or 0.1 pascal. Also see BAR, **1** and MILLIBAR.

microbarograph A barograph that is sensitive to small changes in pressure.

microbeam An energy beam (ray) having extremely small cross section.

microcircuit An extremely small circuit fabricated upon and within a substrate, such as a semiconductor chip. Also see INTEGRATED CIRCUIT.

microcode **1.** See MICROINSTRUCTION. **2.** A code for MICROPROGRAMMING.

microcomponent A tiny component in an electronic circuit. Examples are the resistors, capacitors, diodes, and transistors fabricated onto an integrated-circuit chip.

microcomputer **1.** A computer contained within a single integrated circuit (IC). The simplest such devices perform elementary functions and are available for a few dollars. More sophisticated devices control radio receivers and transmitters, television sets, automobiles, aircraft, and robots. The most advanced devices can be programmed to provide electrical impulses to control erratically functioning body organs, move the muscles of paralyzed persons, and transcribe speech to writing or vice versa. Compare MICROPROCESSOR. **2.** In general, any small computer.

microcrystal A crystal that is invisible to the naked eye.

microcurie A small unit of radioactivity equal to 3.71×10^4 disintegrations per second or 10^{-6} curie. Also see CURIE.

microelectrode **1.** An electrode used in MICROELECTROLYSIS. **2.** A tiny electrode, especially one of those used in integrated circuits and in certain biological applications.

microelectrolysis Electrolysis of tiny amounts of material. Also see ELECTROLYSIS, **1**.

microelectronic circuit **1.** A tiny electronic circuit other than an INTEGRATED CIRCUIT (i.e., one assembled in a small space with small discrete or integrated components). **2.** See MICROCIRCUIT.

microelectronic device See MICROELECTRONIC CIRCUIT.

microelectronics The branch of electronics dealing with extremely small components and circuits fabricated on substrates. Also see INTEGRATED CIRCUIT.

microelectrophoresis Electrophoresis of single particles.

microelectroscope A very sensitive electroscope used to detect minute quantities of electricity.

microelement A tiny component (capacitor, resistor, coil, semiconductor device, or transformer) mounted on a wafer and used in a MICROCIRCUIT.

microelement wafer A microwafer on which a microelement is mounted or deposited.

microfarad Abbreviation, μF. A unit of capacitance equal to 10^{-6} (0.000001) F.

microfarad meter **1.** A dynamometer-type meter that indicates the value of a capacitor directly in microfarads. Such instruments operate from an alternating-current power line. **2.** A direct-reading capacitance meter.

microfiche A method of storing printed information on small film cards. The pages are reduced and arranged in order from left to right and top to bottom. The card is inserted into a projecting machine to allow retrieval of the information. The photographic method is similar to that used in MICROFILM.

microfilm A method of storing printed or photographic information. The pages are reduced and arranged sequentially on a strip of film, usually 35-mm size. The film is inserted into a projecting device for retrieval of the information.

microgalvanometer A highly sensitive GALVANOMETER.

microgauss A magnetic unit equal to 10^{-6} (0.000001) gauss.

microgram A metric unit of weight or mass equal to 10^{-6} (0.000001) gram.

microgroove record A phonograph disc with a very fine groove (200 to 300 per inch), designed for playback at 33⅓ revolutions per minute (rpm).

microhenry Symbol, μH. A unit of inductance, equal to 10^{-6} (0.000001) H.

microhm Symbol, μΩ. A unit of low resistance, reactance, or impedance, equivalent to 10^{-6} (0.000001) ohm.

microhm-centimeter A unit of low resistivity, equal to 10^{-6} (0.000001) ohm-cm. See OHM-CENTIMETER and RESISTIVITY.

microhmmeter An instrument for measuring ultra-low resistance. Such an instrument must have a special provision for canceling the effects of contact and lead resistance.

microinch A unit of linear measure equal to 10^{-6} (0.000001) inch.

microinstruction A machine-code instruction that controls the operation of a computer directly (i.e., it is a “wired-in” instruction, or one set by DIP switches, independent of programs loaded into the machine).

microknowledge In artificial intelligence (AI), detailed machine knowledge. It includes logic rules, computer programs, and data in memory. Compare MACROKNOWLEDGE.

microliter A unit of volume, equal to 10^{-6} (0.000001) liter.

microlock A special form of phase-locked-loop system, used especially with radar to improve the signal-to-noise ratio.

micromanipulator A machine that permits handling tiny parts in very small areas. An example of its use is in placing connections close together in microcircuits.

micrometer **1.** An instrument for measuring very small thicknesses, diameters, etc. **2.** Also called *micron.* The SI unit of length, equal to 10^{-6} (0.000001) meter, or 10^{-3} (0.001) millimeter.

micromho See MICROSIEMENS.

micromicro- See PICO-.

micromicrofarad See PICOFARAD.

micromicrohenry See PICOHENRY.

micromicron A unit of linear measure equal to 10^{-12} meter, or 10^{-6} (0.000001) micrometer.

micromillimeter See NANOMETER.

microminiature Pertaining to an extremely small body, component, or circuit; the last adjective in the sequence of those describing size: *standard*, *small*, *midget*, *miniature*, *subminiature*, and *microminiature.*

micromodule A small, encapsulated circuit, consisting of smaller components. The components can be discrete, can consist of integrated circuits, or can be a combination of both. The module is easily removed and replaced by means of a plug-in socket.

micron See MICROMETER.

microphone A transducer that converts sound waves, especially speech and music, into electrical voltage analogs.

microphone amplifier A sensitive, low-distortion, low-noise amplifier used in voice wireless transmitters and public address systems. Most amplifiers of this type have a tailored frequency response, passing audio between about 300 Hz nd 3000 Hz, and attenuating audio outside this a range. The range 300 Hz to 3000 Hz is sufficient to convey intelligible voice signals, and also allows for audio-frequency-shift keying (AFSK) and slow-scan television (SSTV) audio input.

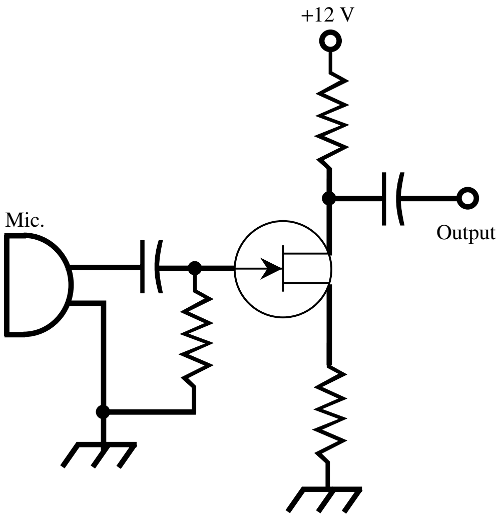

microphone amplifier

microphone boom A device used to hang a microphone, with the base out of the way. It is often used in radio broadcasting.

microphone hummer See HUMMER.

microphone input In an audio amplifier, a jack or other receptacle provided for connection to an external microphone. It can also be used with other low-level audio apparatus. The jack is connected to a MICROPHONE AMPLIFIER that provides high gain with minimum internal noise.

microphone oscillator See HUMMER.

microphonics Ringing (electrical noises) set up by the vibration of a component having loose or movable elements. For example, ringing noises are generated in some circuit boards when they receive a physical blow.

microphonograph A recorder of very low-intensity sound.

microphonoscope An electronic stethoscope, using amplification to enhance the response.

microphotograph An extremely small photograph, often of a pattern or mask used in producing transistors and integrated circuits. Not to be confused with *photomicrograph*, a photograph taken through a microscope.

microphotometer A sensitive instrument for measuring small-area light intensity.

microphysics The branch of physics concerned with atoms, molecules, and subatomic particles.

micropower Extremely small amounts of power—especially the very low direct-current supply power required by some transistors.

microprocessor The integrated circuit (IC), also known as a *chip*, that coordinates the actions of a computer and does the calculations. It is located on the motherboard (sometimes called the *logic board*). These devices get more powerful every year. Physically, this translates to an increasing number of digital switching transistors per chip. The number of digital switches that can be fabricated onto a semiconductor chip of a particular size is ultimately limited by the structure of matter. Compare MICROCOMPUTER.

microprogram **1.** In computer operations, a routine of microinstructions that provides a computer a specific function, independent of those established by programs being run or by the monitor program. Also see MICROINSTRUCTION. **2.** In the direction of a computer, use of a routine that is stored specifically in the memory, instead of elsewhere.

microprogramming In the direction of a computer, the use of a routine that is stored specifically in the memory, instead of elsewhere.

micropulsation Also called *micropulse.* A pulse of extremely short duration.

microradiometer A sensitive detector of heat and infrared radiation, consisting essentially of a thermopile carried by the moving coil of a galvanometer.

microrutherford A unit of radioactivity equal to one disintegration per second or 10^{-6} (0.000001) rutherford. Also see RUTHERFORD.

microsecond A unit of time measure equal to 10^{-6} (0.000001) second.

microsiemens A unit of conductance equal to 10^{-6} (0.000001) siemens.

microspectrophotometer An extremely sensitive spectrophotometer for examining light from tiny areas.

microstrip A microwave component that is, in effect, a single-wire transmission line operating above ground.

microsyn A device that translates rotational position into an electrical signal. Similar to a SELSYN. It is used for such purposes as rotator-direction reading.

microsystems electronics The technology of electronic systems using tiny electronic components. Also see INTEGRATED CIRCUIT, MICROELECTRONIC CIRCUIT, MICROELEMENT, MICROELEMENT WAFER, and MICROWAFER.

microvolt A unit of low voltage, equal to 10^{-6} (0.000001) volt.

microvolter An accurate, external attenuator (usually for an audio signal generator) providing stepped and continuously variable output in microvolts and millivolts.

microvoltmeter A usually direct-reading instrument used to measure voltages in the microvolt range. An input amplifier boosts the test voltage sufficiently to deflect the indicating meter.

microvolts per meter A unit of radio-frequency (RF) field strength. It refers to the RF voltage (in microvolts) between an antenna and ground, divided by the height of the antenna (in meters) above ground. Compare MILLIVOLTS PER METER.

microvolts per meter per mile A means of expressing absolute radio-frequency (RF) field strength. Generally, the numerical value is based on the field strength, in MICROVOLTS PER METER, at a distance of 1 statute mile (5280 feet) from the source.

microwafer A wafer of insulating material, such as a ceramic, on which one or more microelements are mounted and terminals deposited or plated.

microwatt A unit of low power, especially electrical power, equal to 10^{-6} (0.000001) watt.

microwattage See MICROPOWER.

microwattmeter An instrument for measuring power in the microwatt range. Such an instrument obtains its sensitivity from a built-in input amplifier.

microwave See MICROWAVES.

microwave security system A circuit using microwave radio-frequency energy to detect intruders. When an object moves within the field, the intensity of the field changes at one or more pickup sensors, triggering an alarm.

microwave acoustics See ACOUSTOELECTRONICS and ACOUSTIC DELAY LINE.

microwave dish A dish antenna for use at microwave frequencies.

microwave early warning Abbreviation, MEW. A high-power early warning radar system that affords large traffic-handling capacity and long range.

microwave filter A bandpass filter built into a waveguide for use at microwave frequencies.

microwave frequencies The general expression for radio frequencies above the ultra-high range, that is, 3 GHz or more, but below the frequencies of infrared energy. This corresponds to radio wavelengths of 10 centimeters or less.

microwave integrated circuit Abbreviation, MIC. An integrated circuit designed for use at microwave frequencies.

microwave lens See WAVEGUIDE LENS.

microwave mirror A reflector of microwaves.

microwave oven A device consisting essentially of a radio-frequency heater using a magnetron oscillator. It produces microwave energy that causes heating of certain substances via excitation of the molecules.

microwave plumbing Collectively, the waveguides, tees, elbows, and similar fixtures and connections used in microwave setups.

microwave radio relay The use of microwaves to relay radio, television, and control signals from point to point.

microwave refractometer An instrument using microwaves (around 10 GHz) to measure the refractive index of the atmosphere.

microwave region See MICROWAVE FREQUENCIES and MICROWAVES.

microwave relay See MICROWAVE RADIO RELAY.

microwave relay system A series of microwave transmitter-receiver stations for relaying communications in several line-of-sight hops.

microwaves Radio-frequency electromagnetic energy at wavelengths shorter than about 10 centimeters, but longer than the wavelengths of infrared energy. See also MICROWAVE FREQUENCIES.

microwave spectrum See MICROWAVE FREQUENCIES and MICROWAVES.

microwave transistor A transistor whose semiconductor properties and special fabrication enable it to operate at microwave frequencies.

microwave tube A KLYSTRON, MAGNETRON, or similar tube, used to generate or amplify microwave radio-frequency signals.

midband Abbreviation, MB. The region whose limits are immediately above and below a MIDFREQUENCY; the limits are usually specified for a particular case.

midband frequency See MIDFREQUENCY.

midfrequency The center frequency in a specified band of frequencies.

midget Of reduced size (smaller than *small* and larger than *miniature*).

MIDI Acronym for MUSICAL INSTRUMENT DIGITAL INTERFACE.

midpoint voltage The voltage at the terminals of a cell or battery when it has been discharged halfway (i.e., when the amount of energy used up is equal to the amount of energy remaining).

midrange Pertaining to audio frequencies in the middle of the human hearing range, where the ear is the most sensitive. These frequencies lie between the BASS and TREBLE.

midrange horn A MIDRANGE SPEAKER equipped with a flared horn to give the device a unidirectional sound-emission pattern. It is used primarily in high-power systems and by popular music bands or high-end audio enthusiasts.

midrange speaker A loudspeaker operating most efficiently at frequencies in the middle of the audio spectrum. Such a speaker is intermediate in performance between a WOOFER and a TWEETER.

midsection The center section of a multisection filter having an odd number of sections; thus, the second section of a three-section filter.

migration See ION MIGRATION.

mike **1.** See MICROPHONE. **2.** See MICROFARAD. **3.** See MICROMETER.

MIL Abbreviation of *military.*

mil **1.** A small unit of linear measure; 1 mil = 10^{-3} (0.001) inch = 0.0254 mm. **2.** Thousand, as in *n* parts per mil.

mile Abbreviation, m or mi. A large unit of linear measure, 1 mi = 1.609 km = 5280 feet.

military robot A robot designed and used for the purpose of executing some task in warfare. The two general types are: human-operated and computer-controlled. An example of a human-operated military robot is an aircraft that is flown by remote control by a ground-based pilot. The same robot, or a whole fleet of them, might be flown by a computer using sophisticated EXPERT SYSTEMS.

mill A telegraph operator's typewriter.

Miller oscillator A crystal oscillator circuit in which the crystal is connected between the control electrode and ground. The tuned tank is connected in the output circuit. The internal capacitance of the active device provides feedback coupling. Sometimes called *conventional crystal oscillator.*

milli- Abbreviation, m. A prefix meaning thousandth(s): 10^{-3} (0.001).

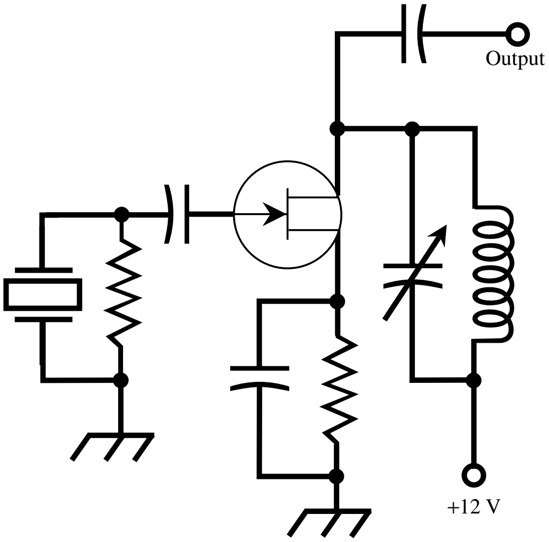

Miller oscillator

milliammeter A usually direct-reading instrument for measuring current in the milliampere range. Also see CURRENT METER.

milliampere Abbreviation, mA. A unit of current equal to 10^{-3} (0.001) ampere.

milliampere-hour Abbreviation, mAh. A unit of low current drain or charging rate, equal to 10^{-3} (0.001) ampere-hour. Also see AMPERE-HOUR and BATTERY CAPACITY.

millibar Abbreviation, mb. A unit of low pressure equal to 10^{-3} (0.001) bar or = 100 pascals.

millicurie Abbreviation, mCi. A small unit of radioactivity equal to 3.71×10^7 disintegrations per second, or 10^{-3} (0.001) curie. Also see CURIE.

millifarad Abbreviation, mF. A seldom-used unit of capacitance, equal to 10^{-3} (0.001) farad or 1000 microfarads.

milligram Abbreviation, mg. A metric unit of weight equal to 10^{-3} (0.001) gram.

millihenry Abbreviation, mH. A unit of inductance, equal to 10^{-3} (0.001) henry.

millihorsepower Abbreviation, mhp. A unit of power equal to 10^{-3} (0.001) horsepower or 0.746 watt. Also see HORSEPOWER.

millilambert Abbreviation, mL. A small unit of brightness equal to 10^{-3} (0.001) lambert.

milliliter Abbreviation, ml. A metric unit of volume equal to 10^{-3} (0.001) liter.

millimaxwell Abbreviation, mMx. A small unit of magnetic flux equal to 10^{-3} (0.001) maxwell or 10^{-11} weber.

millimeter Abbreviation, mm. A metric unit of linear measure equal to 10^{-3} (0.001) meter or 0.03937 inch.

millimeter equivalent The number of millimeters equal to a given English measure fraction (e.g., the millimeter equivalent of $\frac{5}{16}$ inch is 7.937).

millimeter waves Wavelengths between 0.6 and 10 mm (frequencies from 30 to 500 GHz).

millimicro See NANO-.

millimicrofarad See NANOFARAD.

millimicrohenry See NANOHENRY.

millimicron Abbreviation, mm. A unit of wavelength equal to 10^{-3} micron or one nanometer (10^{-9} meter).

millimilliampere See MICROAMPERE.

millimole Abbreviation, mmol. A unit in chemistry equal to 10^{-3} (0.001) mole.

milliohm A small unit of resistance, reactance, or impedance, equal to 10^{-3} (0.001) ohm.

milliohmmeter An ohmmeter for measuring resistances in the milliohm range.

million electronvolt(s) See MEGAELECTRON-VOLT.

milliphot A unit of illumination equal to 10^{-3} (0.001) phot.

millipuffer See PUFFER.

milliradian Abbreviation, mrad. A unit of angular measure equal to 10^{-3} (0.001) radian.

milliroentgen Abbreviation, mr. A small unit of radioactive dosage; 1 mr = 10^{-3} (0.001) roentgen = 2.57976×10^{-7} Ci/kg.

millirutherford Abbreviation, mrd. A small unit of radioactivity equal to 1000 disintegrations per second; 1 mrd = 10^{-3} (0.001) rutherford.

millisecond Abbreviation, ms or msec. A small unit of time equal to 10^{-3} (0.001) second.

millitorr Abbreviation, mT. An obsolete unit of low pressure equal to 10^{-3} (0.001) torr, or 0.133322 pascal. Also see TORRICELLI.

millivolt Abbreviation, mV. A unit of voltage equal to 10^{-3} (0.001) volt.

millivoltmeter A usually direct-reading instrument for measuring low electric potential. Its sensitivity is provided by a high-gain amplifier operated ahead of the indicating meter.

millivolt potentiometer Abbreviation, MVP. A potentiometer-type null instrument for accurately measuring small direct-current voltages, such as those delivered by a thermocouple. Also see POTENTIOMETRIC VOLTMETER.

millivolts per meter Abbreviation, mV/m. A unit of radio-frequency (RF) field strength. It refers to the RF voltage (in millivolts) developed between an antenna and ground, divided by the height (in meters) of the antenna above ground. Compare MICROVOLTS PER METER.

milliwatt Abbreviation, mW. A unit of power equal to 10^{-3} (0.001) watt.

milliwattmeter An instrument for measuring power in milliwatts. Such instruments usually obtain their sensitivity from a built-in preamplifier.

Mills cross A radio-telescope antenna, consisting of two collinear or phased arrays with a common intersecting lobe. The result is high resolution.

mil-spec security See LEVEL-3 SECURITY.

min **1.** Abbreviation of MINIMUM. **2.** Abbreviation of MINUTE.

mineral An element or compound that occurs naturally in the earth's crust. Most minerals are crystalline and many of these have found use in electronics. Some have been produced artificially.

mineral oil A natural liquid insulant derived from petroleum. Dielectric constant, 2.7 to 8.0. Power factor, 0.08 to 0.2 percent at 1 kHz.

mineral-oil capacitor An oil capacitor whose paper dielectric has been impregnated with mineral oil, which is also the filler.

miniature Very small (smaller than midget and larger than subminiature).

miniature cell An electrochemical cell of very small size (e.g., a *button cell* of the kind used in cameras and watches).

miniaturization The technology of minimizing the physical size of a circuit or system, while maintaining its ability to accomplish a given task.

minicalculator A pocket-size electronic calculator.

minicomputer General term for a computer that is more sophisticated than a MICROCOMPUTER, but less powerful than a MAINFRAME.

mini connector A jack or plug having two or three conductors, and measuring ⅛ (0.125) inch in diameter. It is commonly used with audio equipment.

minifloppy A smaller than standard flexible magnetic disk (floppy).

minima Points along a curve at which a function reaches a local minimum value. Also see MAXIMA AND MINIMA.

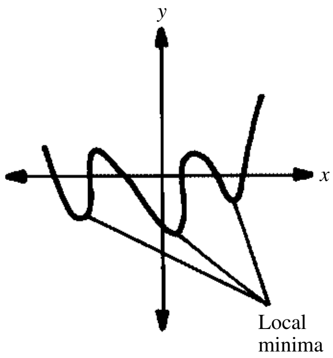

minima

minimum Abbreviation, min. The smallest value in a range or set. Also see MAXIMA AND MINIMA.

minimum detectable signal A signal whose intensity is just higher than the threshold of detection.

minimum discernible signal Abbreviation, MDS. The lowest input-signal amplitude that will produce a discernible output signal in a radio receiver.

miniscope A very-small-sized, lightweight oscilloscope.

minitrack A system used to track an earth satellite, using signals transmitted to the satellite by a line of ground radio stations.

minometer A radioactivity-measuring instrument composed of an ionization chamber and a string galvanometer.

minor beats Secondary or extraneous beats produced in a beat-note system, caused by various sum and difference frequency byproducts of the heterodyne process. Compare MAJOR BEATS.

minor bend A bend in a rectangular waveguide, made without twisting.

minor cycle See WORD TIME.

minor face In a hexagonal quartz crystal, one of the three smaller faces. Compare MAJOR FACE.

minority carrier The type of charge carrier present in relatively small numbers in a processed semiconductor material. Electrons are minority carriers in p-type material; holes are minority carriers in n-type material. Compare MAJORITY CARRIER.

minor lobe In a directional antenna system, any lobe other than the main lobe(s). Such a lobe represents reduced sensitivity and/or power gain relative to the main lobe. Also see MAIN LOBE, SIDE LOBE.

minor loop A subordinate path for the circulation of information or control signals in an electronic system (e.g., *minor feedback loop*). Compare MAJOR LOOP.

minute **1.** Abbreviation, min. A unit of measure of time equal to 60 seconds or 1⁄60 hour. **2.** Also called *minute of arc.* Symbol (′). A unit of arc measure equal to 1⁄60 angular degree or 60 seconds. **3.** General term meaning "extremely small."

MIR Abbreviation of *memory-information register.*

mirror **1.** A device consisting chiefly of a highly polished or silvered surface that reflects a large part of the radiation (such as light) striking it. **2.** Radar-interference material (see CHAFF). **3.** To reflect, as by a mirror.

mirror galvanometer A galvanometer in which a mirror is moved by the coil. The mirror either reflects a spot of light along an external scale, or it reflects the scale, which is then read through a small telescope.

mirror-galvanometer oscillograph See ELECTROMECHANICAL OSCILLOSCOPE.

mirror image **1.** An image or curve that is exactly reversed relative to a straight line or flat plane, compared to a reference image or curve. Compare BILATERAL SYMMETRY. **2.** For a quarter-wave Marconi antenna, the extra quarter-wave element supplied by the earth. **3.** For an antenna at a distance *d* above a ground plane, an effective antenna at an equal distance *d* below the ground plane.

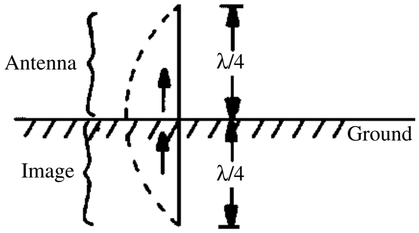

mirror image, 2

mirror-reflection echo A false radar echo or set of echoes, caused by reflection of the radar beam from a plane surface prior to its encountering the target or targets. The beam can also be reflected from one target to another.

mirror-type meter A meter whose movable coil carries a small mirror (rather than a pointer) that reflects a beam of light to produce a spot on a translucent scale.

misaligned head In a tape recorder, a record or a pickup head that is incorrectly oriented, with respect to the passing tape.

misfire Failure of a gas tube or mercury-arc tube to ignite at the correct instant.

misleading precision In electronic calculations and data recording, greater precision than the instruments or conditions justify. Also see SIGNIFICANT FIGURES.

mismatch The condition resulting from joining two circuits, or connecting a line to a circuit, in which the impedances are substantially different.

mismatched impedances Impedances that are unequal, and thus do not satisfy the conditions for maximum power transfer.

mismatch factor For a load not perfectly matched to a driving circuit, the ratio of current flowing in the load to the current that would flow in the load if its impedance were perfectly matched to the output impedance of the driving circuit.

mismatch loss For a load that is mismatched to a source, the ratio P_1/P_2, where P_1 is the power a matched load would absorb from the source, and P_2 is the power actually absorbed by the mismatched load.

mistor A variable-resistance device, used to detect the presence of a magnetic field and to measure magnetic-field strength.

MIT Abbreviation of MASTER INSTRUCTION TAPE.

mix **1.** To produce a beat signal (either the sum or the difference frequency) from two input signals. **2.** The proportion of powdered iron and other inert substances in a ferromagnetic transformer core. Different mixes result in different operating characteristics.

mixdown A method of combining recorded sound from two or more audio tracks, and recording the result onto an audio tape or disc. It is used to create special audio effects.

mixed-base notation A number system in which the base (radix) alternates between two digit positions. Also called *mixed radix notation.*

mixed calculation A mathematical calculation or expression in which more than one operation is used.

mixed modulation Modulation of several kinds coexisting in a system. Thus, a small amount of undesired frequency modulation might accompany amplitude modulation, or vice versa.

mixed number A number having integral (whole) and fractional parts (e.g., 3.14159).

mixer **1.** A device, such as a transistor or semiconductor diode, used to mix two input signals and deliver an output equal to their difference and/or sum (see MIXING). **2.** See AUDIO MIXER. **3.** Any device that combines two or more signals, yielding one output signal whose nature is determined by the characteristics of the circuit.

mixer noise Electrical noise that occurs in a MIXER.

mixing Combining several signals so that some desired mixture of the original signals is obtained. Compare MODULATION.

mixture **1.** A combination of two or more signals that retain their characteristics—even when they interact to produce beat-frequency products. **2.** A

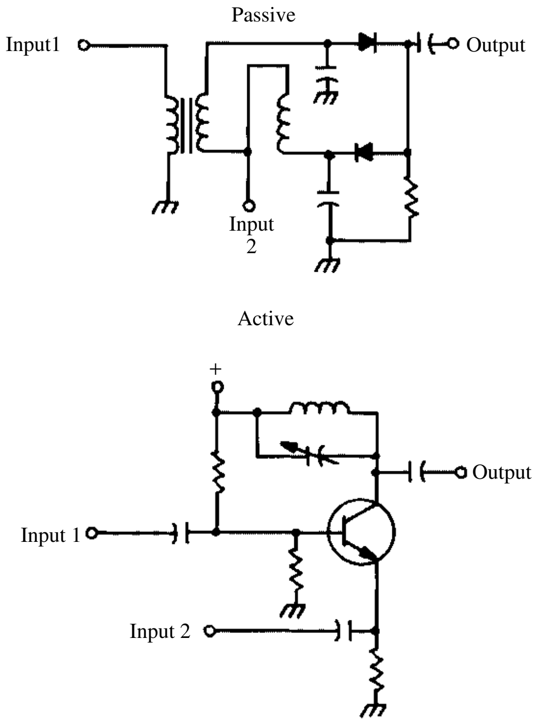

mixer

diffusion of one substance throughout another, without a solution or a chemical reaction resulting.

mixture melting point Abbreviation, mmp. The temperature at which a mixture of solid substances starts turning into a liquid at 1 atmosphere of pressure. This melting point depends upon the melting points of the substances and their relative concentration in the mixture. Also see MELTING POINT and MIXTURE.

mks Abbreviation of METER-KILOGRAM-SECOND.

mL Abbreviation of MILLILAMBERT.

ml Abbreviation of MILLILITER.

mm Abbreviation of MILLIMETER.

mmf Abbreviation of MAGNETOMOTIVE FORCE.

mmol Abbreviation of MILLIMOLE.

mmp Abbreviation of MIXTURE MELTING POINT.

mmv Abbreviation of MONOSTABLE MULTIVIBRATOR.

Mn Symbol for MANGANESE.

m_n Symbol for MASS OF NEUTRON AT REST.

mnemonic **1.** Pertaining to MEMORY or to memory systems. **2.** A memory code or device.

mnemonic code In computer operations, a programming code, such as assembly language, that, although easily remembered by the programmer, requires subsequent conversion to machine language.

mntr Abbreviation of MONITOR. (Also, mon.)

MO Abbreviation of MASTER OSCILLATOR.

Mo Symbol for MOLYBDENUM.

mobile communications **1.** Radio communications between or among stations on board moving or stationary land, waterborne, or airborne vehicles. **2.** Radio communications between at least one fixed station and one or more moving or stationary land, waterborne, or airborne stations.

mobile radio service See MOBILE COMMUNICATIONS.

mobile receiver A radio, television, or other receiver aboard a moving or stationary land, waterborne, or airborne vehicle.

mobile-relay station A fixed station that receives a signal from a MOBILE STATION and retransmits it to one or more other mobile stations.

mobile station A station installed and operated aboard a moving or stationary vehicle. The vehicle might be on land, under water, or in the air. Compare PORTABLE STATION.

mobile transmitter A radio, television, or other transmitter aboard a moving or stationary vehicle. The vehicle might be on land, under water, or in the air.

mobility See CARRIER MOBILITY.

mockup See DUMMY, **1**.

mod **1.** Abbreviation of MODULATOR. **2.** Abbreviation of MODULUS. **3.** Abbreviation of MODIFICATION.

mode **1.** One of the ways a given resonant system can oscillate. **2.** One of the ways that electromagnetic energy can be propagated through a device or system. See MODES OF PROPAGATION. **3.** The method via which intelligence is conveyed in a communications or broadcast signal. See EMISSION MODE. **4.** Resonance of sound waves within an acoustic chamber.

mode coupling The exchange or interaction of energy between identical modes (see MODE, **1**, **2**).

mode filter A waveguide filter that separates waves of different propagation mode, but of the same frequency (see MODES OF PROPAGATION).

model **1.** A working or mockup version of a circuit, system, or device, illustrative of the final version. **2.** A mathematical representation of a process, device, circuit, or system. Example: the Rutherford model of the atom.

modeling The creation of an object in three-dimensional computer graphics.

modem See MODULATOR-DEMODULATOR.

mode purity In a modulated radio-frequency signal, the condition in which no undesirable types of modulation exist. For example, a frequency-modulated signal in which there is zero amplitude modulation.

moderator A substance, such as graphite or heavy water, used to slow neutrons in an atomic reactor. Also see ACCELERATOR, **1** and REACTOR, **2**.

modes of propagation The configurations in which microwave energy can be transmitted through a WAVEGUIDE.

modes of resonance In a microwave cavity, the configurations in which resonant oscillation can exist, depending on the way the cavity is excited.

modification **1.** Changing the configuration of a circuit, device, or system, usually to a minor extent, to tailor its characteristics for a specific purpose. **2.** Changing some aspect of a signal for a specific purpose (e.g., reducing the emission bandwidth to allow more signals to fit within a given band of frequencies). **3.** In computer operations, changing program addresses and instructions by performing logic and arithmetic on them, as if they were data. Also see PROGRAM MODIFICATION.

modified alternate mark inversion A signal that is similar to alternate mark inversion (AMI), but contains certain differences that are specified by a rigorous set of standards for the particular signal.

modifier A data item used to change a computer program instruction so that it can be used to implement different successive operations. Also see PROGRAM MODIFICATION.

modifier register See INDEX REGISTER.

modify To perform a MODIFICATION to a circuit, device, system, signal, program address, etc.

moding A fault characterized by oscillation of a MAGNETRON in undesirable modes.

modular technique See BUILDING-BLOCK TECHNIQUE.

modulated amplifier A usually high-frequency amplifier whose output is varied in some way for the purpose of conveying intelligence. Compare MODULATED OSCILLATOR. In the amplitude modulation of an amplifier, there is little or no disturbance of the carrier frequency. Also see MODULATION.

modulated beam **1.** An electron beam (as in a cathode-ray tube), whose intensity is varied by a desired signal. **2.** A light beam whose intensity is varied for communications or control purposes.

modulated carrier A carrier wave whose amplitude, frequency, or phase is varied to convey intelligence.

modulated continuous wave Abbreviation, mcw. A high-frequency carrier wave modulated by a continuous, lower-frequency wave, as in MCW telegraphy.

modulated CW See MODULATED CONTINUOUS WAVE.

modulated electron beam See MODULATED BEAM, **1**.

modulated light beam See MODULATED BEAM, **2**.

modulated oscillator A usually high-frequency oscillator whose output is varied in some way to convey intelligence. Compare MODULATED AMPLIFIER. Also see MODULATION.

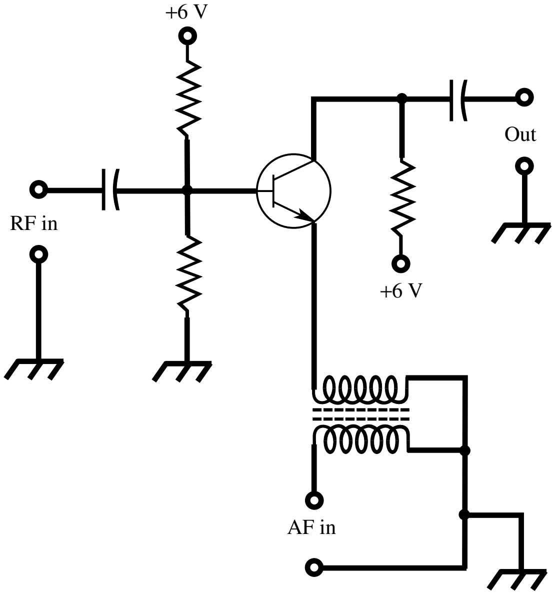

modulated amplifier

modulated-ring pattern See GEAR-WHEEL PATTERN and SPOT-WHEEL PATTERN.

modulated stage A transmitter, amplifier, or oscillator in which the signal information is impressed on the carrier.

modulated wave See MODULATED CARRIER.

modulatee A stage or circuit upon which modulation is impressed (e.g., a MODULATED AMPLIFIER or a MODULATED OSCILLATOR).

modulating electrode **1.** In an oscilloscope, an electrode (usually the intensity electrode) to which a signal can be applied to intensity-modulate the electron beam. **2.** In a cathode-ray tube, the electrode to which the video signal is applied.

modulating signal Intelligence that modulates a carrier (e.g., binary data in radioteletype, or music in broadcasting).

modulation Combining two signals with the result that some aspect of one signal (the carrier) is varied by and in sympathy with the other (the modulating signal). Usually, the carrier has a frequency considerably higher than that of the modulating signal.

modulation bars A form of television interference in which an amplitude-modulated signal causes horizontal bars, alternating light and dark, to appear on the picture screen. The higher the modulating frequency, the closer together the bars appear. In severe cases, the bars completely obliterate the picture.

modulation capability The maximum percentage of modulation a transmitter will permit before

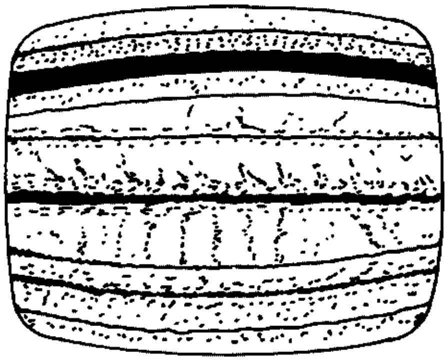

modulation bars

nonlinearity occurs. Also see MODULATION LINEARITY.

modulation characteristic For an amplitude-modulated wave, the ratio of the instantaneous amplitude of the modulated signal to the instantaneous modulating voltage.

modulation code In a modulated transmitter, a system of modulation in which certain signal variations or pulses represent particular characters. Examples are ASCII, BAUDOT CODE, and MORSE CODE.

modulation coefficient Symbol, *m*. A figure expressing the depth (extent) to which a signal is amplitude-modulated. For a signal in which the upward modulation is equal to the downward modulation, $m = (E_m - E_c)/E_c$, where E_c is the peak-to-peak voltage of the unmodulated signal, and E_m is the peak-to-peak voltage of the modulated signal. For full (100%) modulation, $m = 1$.

modulation depth See DEPTH OF MODULATION.

modulation distortion **1.** In a modulated signal, envelope distortion introduced by the modulation process or by the receiver circuit. **2.** External cross modulation (see CROSS MODULATION, **1**).

modulation envelope See ENVELOPE, **1**.

modulation-envelope distortion Undesirable distortion in the ENVELOPE of the modulating intelligence in an amplitude-modulated or single-sideband signal at the output of a radio transmitter.

modulation factor See MODULATION COEFFICIENT.

modulation frequency Abbreviation, f_m. The frequency of a modulating signal.

modulation linearity In a modulated signal, the degree to which the instantaneous signal amplitude or frequency follows the instantaneous amplitude of the modulating signal.

modulation meter See PERCENTAGE-MODULATION METER.

modulation monitor **1.** A linear detector with a pickup coil (or antenna) and headphones for listening to an amplitude-modulated signal. **2.** See PERCENTAGE-MODULATION METER. **3.** A device that displays the modulation characteristics of a signal (e.g., the envelope for amplitude modulation and single sideband or the frequency-vs. time function for frequency modulation).

modulation noise See NOISE BEHIND THE SIGNAL.

modulation percentage A measure of the extent to which a signal is amplitude modulated. It is expressed as the MODULATION COEFFICIENT multiplied by 100.

modulation ratio For a modulated signal, the quotient M_r/M_i, where M_r is the percentage of radiated-signal modulation, and M_i is the percentage of current modulation.

modulator A device or circuit for modulating a carrier.

modulator cell See KERR CELL.

modulator crystal A transparent piezoelectric crystal to which a signal voltage can be applied to modulate a beam of polarized light passing through it.

modulator-demodulator Also called *modem.* **1.** A device that converts binary data to audio-frequency-shift-keyed (AFSK) analog signals, and vice-versa. It is commonly used to interface computers and terminals with telephone lines and radio transceivers. **2.** A circuit or device, such as a

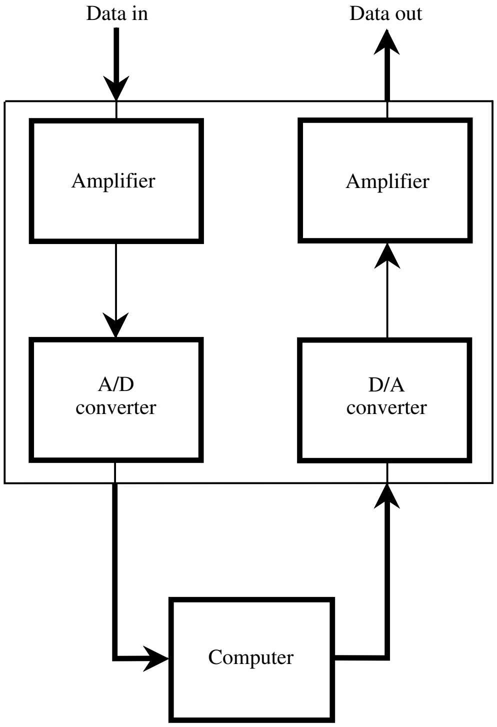

modulator-demodulator

biased diode or diode bridge, that can perform either modulation or demodulation.

modulator driver An amplifier stage that delivers excitation current, voltage, or power to a modulator stage.

module An assembly containing a complete self-contained circuit (or subcircuit), often miniaturized and made for plug-in operation.

modulometer Any instrument, such as a percentage-modulation meter, used to measure the degree of modulation of a signal. Often, it can also measure other signal characteristics (e.g., carrier shift, extraneous amplitude modulation, and extraneous frequency modulation).

modulo n check In computer operations, a technique for verifying the validity of a number used as an operand. The number being so checked is divided by another number to provide a remainder (check digit) that goes with the number. After the number is, for example, transmitted through some part of a computer system, it is again divided by the original divisor, and if the remainder is the check digit, the data has retained its integrity.

modulus **1.** Absolute magnitude. Also see ABSOLUTE VALUE and IMPEDANCE. **2.** Abbreviation, mod. In computer operations, a whole number that indicates the number of states a counter sequences through in each cycle. **3.** Abbreviation, mod. A number (constant or coefficient) expressing the degree to which some property is possessed by a material or body (e.g., *modulus of elasticity*, *shear modulus*, and *bulk modulus*). **4.** A constant by which a logarithm to one base must be multiplied to obtain a logarithm of the same number to another base.

modulus of elasticity The stress-to-strain ratio in a material under elastic deformation.

moire In a television or facsimile picture, an effect produced by the convergence of straight lines. When the lines are nearly parallel to the scanning lines, the converging lines appear irregular.

moisture meter See ELECTRIC HYGROMETER and ELECTRONIC HYGROMETER.

mol Abbreviation of MOLE.

molar conductance See MOLECULAR CONDUCTANCE.

molar polarization Any molecule in an electric field undergoes a small displacement of the positive and negative electrical centers. This results in an electric dipole.

molar solution A solution, such as an electrolyte, containing 1 mol of solute per liter of solvent. Compare NORMAL SOLUTION.

mold **1.** To form matter into a desired shape, as by pouring liquefied material into a container or liquefying the material in the container, then allowing the liquid material to solidify. In *hot molding*, the material is melted in the container and then cooled to hardness; in *cold molding*, the material is shaped without heat and it solidifies with time. **2.** The hollow container used to shape a material, as in **1**.

molded capacitor A capacitor that is molded into a protective body of insulating material. Also see MOLD, **1**, **2** and MOLDED COMPONENT.

molded ceramic capacitor A ceramic-dielectric capacitor enclosed in a molded housing. Also see MOLD, **1**, **2**; MOLDED CAPACITOR; and MOLDED COMPONENT.

molded coil See MOLDED INDUCTOR.

molded component A part (such as a capacitor, coil, or resistor) that is completely enclosed in a protective material (such as a plastic) that is molded around it. Also see MOLD, **1**, **2**.

molded electrolytic capacitor A solid-dielectric electrolytic capacitor enclosed in a molded housing. Also see MOLD, **1**, **2**; MOLDED CAPACITOR; MOLDED COMPONENT; and SOLID-ELECTROLYTIC CAPACITOR.

molded glass capacitor A glassplate-dielectric capacitor enclosed in a molded glass housing. Also see GLASS CAPACITOR; MOLD, **1**, **2**; MOLDED CAPACITOR; and MOLDED COMPONENT.

molded inductor An inductor that is molded into a protective housing of insulating material. Also see MOLD, **1**, **2** and MOLDED COMPONENT.

molded mica capacitor A mica-dielectric capacitor enclosed in a molded housing. Also see MICA CAPACITOR; MOLD, **1**, **2**; MOLDED CAPACITOR; and MOLDED COMPONENT.

molded mud A molding compound that has inferior electrical characteristics. Also see MOLD, **1**, **2** and MOLDED COMPONENT.

molded paper capacitor A paper-dielectric capacitor enclosed in a molded housing. Also see MOLD, **1**, **2**; MOLDED CAPACITOR; MOLDED COMPONENT; and PAPER CAPACITOR.

molded porcelain capacitor A capacitor enclosed in a body of molded porcelain. Also see MOLD, **1**, **2**; MOLDED CAPACITOR; MOLDED COMPONENT; and PORCELAIN.

molded resistor A resistor that is molded in a protective housing of insulating material. Also see MOLD, **1**, **2** and MOLDED COMPONENT.

molded transistor A transistor that is encapsulated in a protective molding compound, such as epoxy resin. Also see MOLD, **1**, **2** and MOLDED COMPONENT.

mole Abbreviation, mol. **1.** The amount of substance in a system containing as many specified entities (atoms, molecules, ions, subatomic particles, or groups of such particles) as there are atoms in 12 grams of carbon 12. **2.** It is also called the *Avogadro constant.* A unit of quantity in chemistry, equal to approximately 6.022×10^{23}.

molectronics See MOLECULAR ELECTRONICS.

molecular circuit See MONOLITHIC INTEGRATED CIRCUIT.

molecular conductance For a solution, such as an electrolyte, the product of specific conductivity and the volume (in liters) of a solution that contains 1 gram molecule of the solute. Also see SOLUTE; SOLUTION, **1**; and SOLVENT, **1**, **2**.

molecular conductivity See MOLECULAR CONDUCTANCE.

molecular electronics The technique of processing a single block of material so that separate areas perform the functions of different electronic components. The entire block constitutes a circuit (e.g., a MONOLITHIC INTEGRATED CIRCUIT).

molecular magnets According to the molecular theory of magnetism, the elemental magnets formed by individual molecules.

molecular theory of magnetism Each molecule in a piece of magnetic metal is itself a magnet (possessing a north and a south pole). These tiny magnets are thought to be normally oriented at random, but when the material is magnetized by an external force, they align themselves with each other.

molecular weight Abbreviation, mol wt. In a molecule of a substance, the sum of the atomic weights of the constituent atoms.

molecule The basic particle of a compound; each molecule usually contains two or more atoms. For example, the formula $AgNO_3$ represents silver nitrate, each molecule of which contains one atom of silver (Ag), one atom of nitrogen (N), and three atoms of oxygen (O).

moletronics See MOLECULAR ELECTRONICS.

mol wt Abbreviation of MOLECULAR WEIGHT.

molybdenum Symbol, Mo. A metallic element. Atomic number, 42. Atomic weight, 95.94. It is used in the grids and plates of certain vacuum tubes.

moment The tendency to produce motion around a point, as by torque, or the product of a quantity and the distance to a point. The moment of force is expressed as the product *Fd*, where *F* is force and *d* is distance.

momentary-contact switch A switch that maintains contact only while it is held down. Such a device is usually a *pushbutton switch*, although it might be a *toggle switch*, a *slide switch*, or a *lever switch.*

momentary switching Switching of short duration, often characterized by a quick make and break immediately following activation of the switch. Compare DWELL SWITCHING.

moment of inertia For a torque motor, the inertia of the armature around the axis of rotation. Also see MOMENT.

mon **1.** Abbreviation of MONITOR. **2.** Abbreviation of MONAURAL.

monatomic **1.** Pertaining to a molecule with only one atom. **2.** Pertaining to a molecule with only one replaceable atom or radical.

monatomic molecule A molecule having a single atom (e.g., argon, helium, and neon). Compare DIATOMIC MOLECULE.

monaural **1.** Pertaining to an audio system having one channel. **2.** Pertaining to hearing with one ear, as opposed to BINAURAL.

monaural recorder A single-track recorder, as opposed to a stereophonic recorder.

Monel metal An alloy of nickel (67%), copper (28%), iron, manganese, and other metals (5%). Its resistivity is approximately 42 microhm-centimeters at 20°C.

monimatch An amateur version of the reflected-power meter and SWR meter.

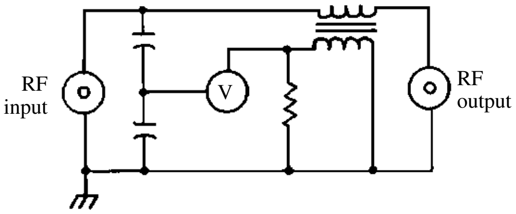

monimatch

moniscope A special cathode-ray tube that produces a stationary picture for testing television equipment. Its name is a contraction of monitor and scope.

monitor **1.** A device that allows the sampling of a signal or quantity. Examples: *line-voltage monitor*, *television monitor*, and *modulation monitor.* **2.** A cathode-ray-tube (CRT) computer display.

monitor head A separate playback head included in some tape recorders for listening to the tape as it is being recorded.

monitoring The act, process, or technique of observing an action while it is in progress or checking a quantity while it is varying. Examples: *carrier monitoring*, *modulation monitoring*, and *line-voltage monitoring.*

monitoring amplifier An auxiliary amplifier used in monitoring an audio-frequency system.

monitoring antenna A usually small pickup antenna used with a signal monitor or monitoring receiver.

monitoring key In a telephone system, a device used to listen to a two-way conversation.

monitoring receiver A radio or television receiver used specifically to monitor a transmission directly.

monitoring station In a security system, a central control location from which personnel can observe the input from sensors, cameras, and other devices at remote locations throughout the secured area.

monitor system A computer program usually stored in the read-only memory (ROM) supplied by the hardware vendor. It controls the implementation of programs written by the user, and the operation of peripherals associated with program runs and inputting or outputting data to or from memory. Also called *executive program.*

monkey chatter The characteristic sound of a single-sideband (SSB) signal when the receiver is mistuned, or when the receiver is set for the wrong sideband.

monk's cloth A coarse drapery fabric sometimes used to soundproof the walls and ceiling of a radio studio or recording booth.

mono See MONAURAL.

mono- Prefix meaning *single.*

monobrid circuit An integrated circuit in which either several monolithic IC chips are interconnected to form a larger, single-package circuit, or monolithic IC chips are interwired with thin-film components into a single-package circuit. The name is a contraction of *monolithic hybrid.*

monochromatic **1.** Being of one color (hue) in nature. **2.** Being of a single wavelength in nature (pertaining to radiation of any kind). **3.** Pertaining to black-and-white television.

monochromaticity Consisting of one color of visible light. The brightness can vary from black to maximum.

monochromatic power density At a given temperature, the energy radiated per square centimeter of blackbody surface per second per unit wavelength range. Also see BLACKBODY and BLACKBODY RADIATION.

monochromatic sensitivity Sensitivity to light of one color only.

monochrome television Black-and-white television.

monoclinic crystal A crystal having three axes of unequal length; two of them intersect obliquely and are perpendicular to the third [e.g., the type of crystal found in one form of sulfur (monoclinic sulfur)].

monocrystalline material See SINGLE-CRYSTAL MATERIAL.

monode A one-element device, such as a filament-type lamp, thermistor, voltage-dependent resistor, barretter, etc.

monogroove stereo A method of making a stereophonic phonograph disc in which both channels are recorded as a single groove.

monolayer A thin film having a thickness of one molecule.

monolithic integrated circuit An integrated circuit (IC) formed in a single block or wafer of semiconductor material. The name is derived from the Greek monolithos ("one stone"). Compare HYBRID INTEGRATED CIRCUIT and THIN-FILM INTEGRATED CIRCUIT.

monometallic Containing or using only one metal.

monomolecular film See MONOLAYER.

monophonic recorder See SINGLETRACK RECORDER.

monophonic system A single-channel sound system. Compare STEREO SYSTEM.

monopole antenna See QUARTER-WAVE MONOPOLE.

monopulse In radar and electronic-navigation operations, using one pulse to determine azimuth and elevation simultaneously.

monorange speaker A loudspeaker that reproduces most of the full audio range. Also called *extended range speaker.* Compare TWEETER and WOOFER.

monostable Having one stable state.

monostable blocking oscillator Abbreviation, MBO. A blocking oscillator that behaves somewhat like a one-shot multivibrator. The oscillator delivers a single output pulse each time it receives an input (trigger) pulse.

monostable multivibrator A multivibrator that delivers one output pulse for each input (trigger) pulse. Also called *one-shot circuit* and *single-shot multivibrator.* Compare ASTABLE MULTIVIBRATOR and BISTABLE MULTIVIBRATOR.

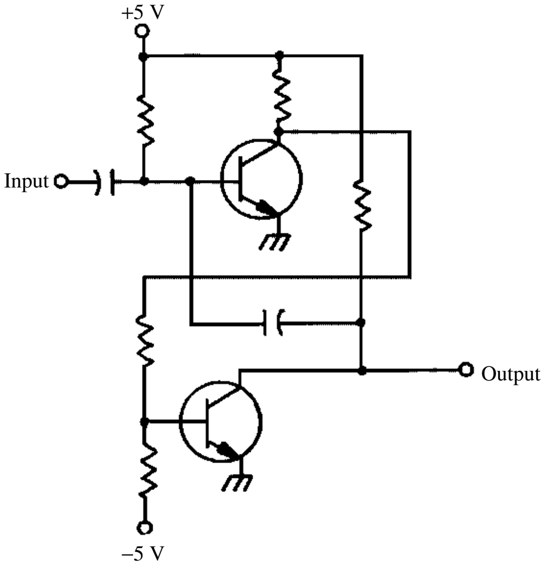

monostable multivibrator

monostatic reflectivity The property whereby, for certain reflectors (such as a tricorner reflector), all incident rays are reflected in exactly the opposite direction from which they arrive.

monotone A sound or series of sounds having a single, constant pitch (frequency).

monotonicity In a digital-to-analog converter, a condition where the output either remains the same or increases for any single increase in the input code.

monovalent See UNIVALENT.

Monte Carlo method **1.** The use of statistical sampling in the approximate solution of an engineering problem. **2.** In computer operations, the construction of mathematical models from randomly selected components taken from representative statistical populations.

Moog synthesizer An electronic device that can be made to simulate virtually any sound, including that of musical instruments and the human voice, through the use of several audio oscillators, whose output can be controlled to produce tones of various harmonic content, duration, attack, and decay periods.

moonbounce Also called *earth-moon-earth (EME).* Radio communication, usually at very-high frequencies (VHF) or ultra-high frequencies (UHF), in which the moon is used as a passive reflector. This is a popular mode among some amateur radio operators.

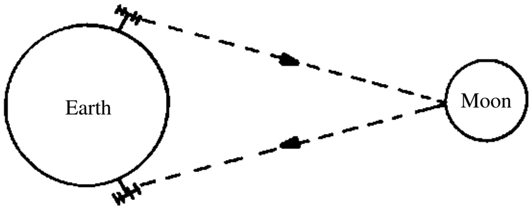

moon bounce

MOPA Abbreviation of MASTER OSCILLATOR/POWER AMPLIFIER.

morphological electronics See MOLECULAR ELECTRONICS.

Morse **1.** See MORSE CODE. **2.** Telegraphy (wire or radio). **3.** To signal by means of the Morse code.

Morse code Either of two similar binary codes used in radio and wire telegraphy. It uses short pulses (dots or dits) and long pulses (dashes or dahs) to represent letters of the alphabet, numerals, and punctuation marks. It usually refers to the CONTINENTAL CODE, but occasionally it refers to the AMERICAN MORSE CODE.

MOS Abbreviation of *metal-oxide-semiconductor.*

mosaic **1.** See PHOTOMOSAIC. **2.** The pattern of tiny photoelectric particles in a television camera tube that convert the image into electric charges.

mosaic crystal A form of imperfect crystal. The defects have certain properties, one of which is to cause additional energy levels in semiconductor materials manufactured from such crystals.

MOS capacitor A capacitor using metal-oxide-semiconductor (MOS) technology. It is used in MOS integrated circuits. A silicon substrate forms one electrode (generally p-type material). An oxide layer forms the dielectric. An electrode forms the other plate of the capacitor. The capacitance might be variable by changing the applied voltage at the metal gate electrode.

MOSFET Abbreviation of METAL-OXIDE-SEMICONDUCTOR FIELD-EFFECT TRANSISTOR.

MOSROM Abbreviation of *metal-oxide-semiconductor read-only memory* (see READ-ONLY MEMORY).

MOST Abbreviation of *metal-oxide-semiconductor transistor.*

most significant character Abbreviation, MSC. In positional number representation, the leftmost character in a significant group, such as a word.

most significant digit Abbreviation, MSD. In a number, the leftmost digit that is not zero (zero being insignificant in this context).

mother **1.** A term or prefix referring to a component that supports (or appears to support) other similar components (e.g., MOTHERBOARD). **2.** A term referring to the source from which samples or components are derived (e.g., MOTHER CRYSTAL). **3.** A mold that has been electroformed from a master phonograph disc.

motherboard In a computer or data-processing device, the circuit board on which most of the main circuitry is mounted.

mother crystal A natural quartz crystal from which is produced the piezoelectric plates and other components used in electronics.

motion detector A device for sensing the movement or stopping of a body, such as a rotating shaft. Various sensors are used in different detectors: magnetic, photoelectric, capacitive, etc.

motion frequency The natural frequency (especially that of oscillation) of a servo.

Motion Picture Experts Group Abbreviation, MPEG. A data-compression standard used for animated digital video. It is useful in computer videoconferencing and telecommuting, and for the development of multimedia presentations.

motion-picture pickup In television operations, a camera (and the technique for using it) for picking up scenes directly from motion-picture film.

motion sensor In security systems, a set of devices that produces an output signal whenever anything moves within a certain area. Such equipment might use infrared, ultrasound, microwaves, capacitive effects, thermal sensors, air-current detectors, sound detectors, video cameras, or a combination of these.

motor **1.** A machine for converting electrical energy into mechanical energy. **2.** The driving mechanism of a loudspeaker.

motorboard The basic mechanism of a tape recorder, embodying motor, flywheel, capstan, rollers, etc., assembled on a board or panel.

motorboating A repetitive, low-frequency popping or puffing noise sometimes observed in malfunctioning audio circuits. It was so named because of its resemblance to the sound of a motorboat.

motorboating filter In an audio amplifier circuit, a simple filter installed to prevent

MOTORBOATING caused by unwanted positive feedback. Also see DECOUPLING FILTER.

motor capacitor See MOTOR-RUN CAPACITOR and MOTOR-START CAPACITOR.

motor constant The ability, expressed numerically, of a torque motor to convert electric input power into torque.

motor converter An electromechanical device for converting alternating current to direct current.

motor-driven relay An electromechanical relay whose contacts are opened and closed by a rotating motor.

motor effect Magnetic force between adjacent current-carrying conductors.

motor-generator A combination of motor and generator in a single machine assembly. A common arrangement is a low-voltage motor turning a high-voltage generator. The machine shafts can be coupled, or the motor can turn the generator through a belt, chain, or gear train.

motor meter A meter in which the movable element is essentially a continuously rotating motor. See, for example, KILOWATT-HOUR METER.

motor-run capacitor A power-factor-boosting capacitor connected (together with its auxiliary winding) in parallel with the main winding in an induction motor. Also see CAPACITOR MOTOR.

motor-speed control A method of controlling the speed of a motor by varying the magnitude and/or phase of its current. Electronic devices usually use diode, diac, or triac circuits.

motor-start capacitor A capacitor that, with an auxiliary winding, is switched into a motor circuit during the starting period, and is automatically disconnected (with the winding) after the motor reaches normal running speed. Also see CAPACITOR MOTOR.

motor starter See STARTING BOX.

mould See MOLD.

mount **1.** A mechanical device with which a component is attached to a circuit board or chassis. **2.** The attachment of a circuit board to a chassis. **3.** The hardware with which an antenna is attached to a mast. **4.** The hardware with which a microphone is attached to a boom or other support. **5.** In general, any attaching hardware.

Mountain standard time Abbreviation, MST. Local mean time at the 105th meridian west of Greenwich. Also see GREENWICH MEAN TIME, STANDARD TIME, TIME ZONE, and COORDINATED UNIVERSAL TIME.

mounting flange The portion of a speaker frame, usually made of metal, that is equipped with holes for attachment to a cabinet or panel.

mouse A pointing device commonly used with personal computers. It is about the size of a candy bar. The computer operator moves the display pointer or cursor by sliding the mouse on a flat surface. Functions and commands are executed by pressing a button (clicking) on the device.

mouth The radiating end of a horn (antenna, loudspeaker, etc.).

MOV Abbreviation of METAL-OXIDE VARISTOR.

movable-coil meter See D'ARSONVAL METER.

movable-coil modulator See ELECTROMECHANICAL MODULATOR.

movable contact The traveling contact in a relay or switch. Compare FIXED CONTACT.

movable-iron meter See IRON-VANE METER.

movies on demand In television, a system that makes it possible for subscribers to choose the programs they want to see, and also to choose the viewing times for the programs.

moving-coil galvanometer A galvanometer whose movable element is a coil of fine wire suspended or pivoted between the poles of a magnet.

moving-coil microphone See DYNAMIC MICROPHONE.

moving-coil motor The driving mechanism of a moving-coil DYNAMIC SPEAKER.

moving-coil pickup See DYNAMIC PICKUP.

moving-coil speaker See DYNAMIC SPEAKER.

moving-conductor microphone See VELOCITY MICROPHONE.

moving-diaphragm meter A headphone used as a sensitive indicator in alternating-current bridge measurements.

moving element In an electromechanical device, the portion that moves physically under variable operating conditions.

moving-film camera An oscilloscope camera in which the film is drawn past the lens continuously at a constant speed, rather than being advanced frame by frame, as in a motion-picture camera.

moving-iron meter See IRON-VANE METER.

moving-vane meter See IRON-VANE METER.

mp Abbreviation of MELTING POINT.

m_p Symbol for MASS OF PROTON AT REST.

MPEG Abbreviation for MOTION PICTURE EXPERTS GROUP.

MPG Abbreviation of *microwave pulse generator.*

mph Abbreviation of *miles per hour.* Also, mi/h.

MPO Abbreviation of *maximum power output.*

mps **1.** Abbreviation of *meters per second.* Also, m/s. **2.** Abbreviation of *miles per second.* Also, mi/s.

MPT Abbreviation of MAXIMUM POWER TRANSFER.

MPX Abbreviation of MULTIPLEX.

MR Abbreviation of MEMORY REGISTER.

mrad Abbreviation of MILLIRADIAN.

MRIA Abbreviation of *Magnetic Recording Industry Association.*

mrm Abbreviation of *milliroentgens per minute.*

MS Abbreviation of MASS SPECTROMETER.

m^2/s Abbreviation of *meters squared per second.* The unit of kinematic viscosity.

M scan In radar operations, a modified A-scan display in which a pedestal signal is moved along the base line to a point where it coincides with the base line of the reflected signal to determine the distance to the target.

msec Abbreviation of MILLISECOND. Also, ms.

msg Abbreviation of MESSAGE.

MSI Abbreviation of MEDIUM SCALE INTEGRATION.

MST Abbreviation of MOUNTAIN STANDARD TIME.

MT Abbreviation of METRIC TON.

MTR Abbreviation of MAGNETIC TAPE RECORDER.

mtr Alternate abbreviation of METER.

MTS Abbreviation for MULTICHANNEL TELEVISION SOUND.

M-type backward-wave oscillator A broadband, voltage-tuned oscillator in which the electrons interact with a backward traveling radio-frequency wave. Compare O-TYPE BACKWARD-WAVE OSCILLATOR.

mu Symbol, μ. **1.** Abbreviation of the prefix MICRO-. **2.** Expression for AMPLIFICATION FACTOR. **3.** Expression for PERMEABILITY. **4.** Expression for MICRON. **5.** Expression for ELECTRIC MOMENT. **6.** Expression for INDUCTIVITY. **7.** Expression for MAGNETIC MOMENT. **8.** Expression for MOLECULAR CONDUCTIVITY.

μ **1.** Symbol for MU.

μA Abbreviation of MICROAMPERE.

μ_B Symbol for BOHR MAGNETON.

μb Abbreviation of MICROBAR.

μCi Abbreviation of MICROCURIE.

μ_e Symbol for ELECTRON MAGNETIC MOMENT.

μF Abbreviation of MICROFARAD.

μg Abbreviation of MICROGRAM.

μH Abbreviation of MICROHENRY.

μin Abbreviation of MICROINCH.

μl Abbreviation of MICROLITER.

μmm Abbreviation of *micromillimeter* (see NANOMETER).

μ_n Symbol for NUCLEAR MAGNETON.

μ_o Symbol for FREE-SPACE PERMEABILITY CONSTANT.

μ_p Symbol for PROTON MAGNETIC MOMENT.

μP Abbreviation of MICROPROCESSOR.

μrd Abbreviation of MICRORUTHERFORD.

μS Abbreviation of MICROSIEMENS.

μs Abbreviation of MICROSECOND. (Also, μsec.)

μV Abbreviation of MICROVOLT.

μV/m Abbreviation of MICROVOLTS PER METER.

μW Abbreviation of MICROWATT.

MUF Abbreviation of MAXIMUM USABLE FREQUENCY.

mu factor See MU, **2**.

mu law In communications, a companding standard generally used in the United States and Canada.

Muller tube A tube that conducts by means of ionization of an internal gas. It is used for radiation detection.

multiaddress In a computer, a multiple memory location.

multiband amplifier A radio-frequency power amplifier capable of operation on more than one frequency band. The bands are often (but not always) harmonically related. Compare WIDEBAND AMPLIFIER.

multiband antenna A single antenna, such as one used in amateur radio and television reception, operated in several frequency bands or channels. The bands are often (but not always) harmonically related.

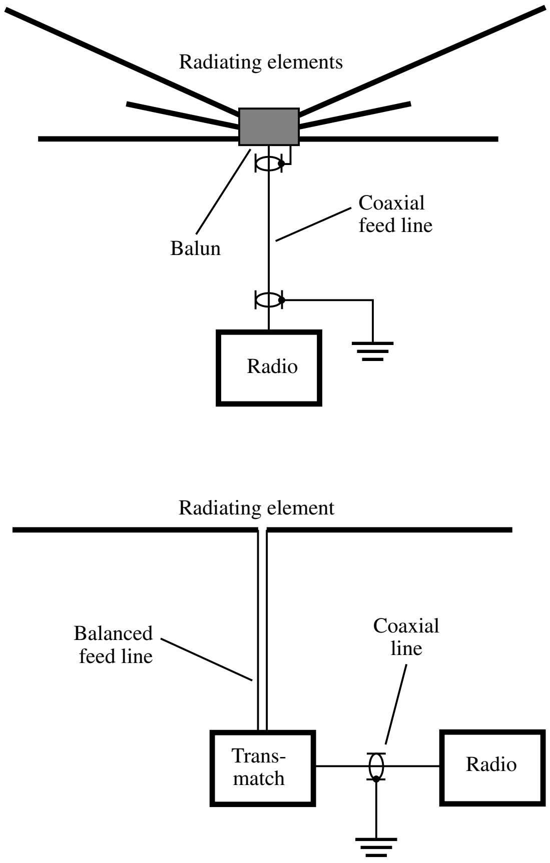

multiband antenna

multiband device A device (such as a tuner, receiver, transmitter, or test instrument) that operates in several selectable frequency bands.

multiband oscilloscope See WIDEBAND OSCILLOSCOPE.

multiband receiver A communications receiver capable of operation on more than one frequency band. Such a receiver might, but does not necessarily, provide continuous coverage over a wide range of frequencies. Compare WIDEBAND RECEIVER.

multicasting The use of two frequency-modulation (FM) broadcast stations, or one FM station and one television station, to broadcast separately the two channels of a stereo program. The program is picked up simultaneously with two receivers.

multicavity magnetron A magnetron whose anode block has two or more cavities.

multicellular horn A loudspeaker in front of which are placed rectangular cells through which the sound passes.

multichannel Pertaining to a radio-communication system that operates on more than one channel at the same time. The individual channels might contain identical information, or they might contain different signals.

multichannel analyzer A test instrument, such as a spectrum analyzer, that splits an input into several channels for testing.

multichannel television sound Abbreviation, MTS. In television broadcasting, audio transmitted on more than one channel to provide stereo sound to subscribers.

multichip circuit A MICROCIRCUIT composed of interconnected active and passive chip-type components.

multichip integrated circuit An INTEGRATED CIRCUIT composed of circuit elements on separate, interconnected chips.

multicontact switch A switch having more than two contacting positions.

multicoupler An impedance-matching device used to couple several receivers to a single antenna.

multielement antenna A directive antenna having more than one element. Such antennas include phased arrays and parasitic arrays. Common examples are the LOG-PERIODIC ANTENNA, the QUAD ANTENNA, and the YAGI ANTENNA.

multiemitter transistor 1. A bipolar transistor having more than one emitter. **2.** A bipolar power transistor having several emitters connected in parallel in the transistor structure.

multiframe A set of frames in a signal that are adjacent to each other in time (consecutive) and positioned according to an alignment signal.

multigun CRT A cathode-ray tube (CRT) having more than one electron gun.

multihop propagation Propagation of a radio wave by several successive reflections between the ionosphere and the surface of the earth.

multilayer circuit A circuit consisting of several sections printed or deposited on separate layers, which are subsequently stacked in a sandwich-like manner.

multilayer coil A coil in which the turns of wire are wound in several complete layers, one on top of the other. Compare SINGLE-LAYER COIL.

multilevel password protection See HIERARCHICAL PASSWORD PROTECTION.

multimedia 1. In computing, the use of video, voice, music, electromechanical control, and/or data transfer at the same time. Useful in education, entertainment, business, and gaming applications. **2.** The use of voice, images, data, and/or video in a communications system or network, and, in particular, in wireless applications. An example is videoconferencing between a corporate office and the user of a handheld computer equipped with a wireless modem.

multimedia computer A personal computer designed especially for multimedia use. It includes a sound board, speakers, a microphone, and a CD-ROM (compact-disk read-only memory) drive. It often has a large-screen, high-resolution monitor (17 inches or greater diagonal measure).

multimeter A meter that allows measurement of different quantities (e.g., current, voltage, and resistance); the functions are usually made available through a selector switch.

multimode operation 1. In radio communication, the use of two or more transmitters simultaneously, operating in different modes [e.g., one using single-sideband (SSB) and another using frequency modulation (FM)]. **2.** The transmission of visible light or infrared energy through an optical fiber in more than one mode at the same time. **3.** The operation of any device in more than one of its modes simultaneously.

multipactor A microwave switching tube capable of operating at high power levels. Characterized by high operating speed.

multipath cancellation A phenomenon sometimes observed in radio-wave propagation. Separate signal components arrive at the receiver in equal amplitude, but opposite phase. Also see MULTIPATH FADING and MULTIPATH RECEPTION.

multipath delay In MULTIPATH RECEPTION, the lag between signal components arriving over different paths.

multipath effect At a receiver, the difference in arrival time of multipath signals. Also see MULTIPATH DELAY.

multipath fading In radio communication, variations in the received signal that result from phasing various components of the transmitted signal that are propagated over different paths. At low, medium, and high frequencies, this effect is usually the result of ionospheric fluctuation. At very-high and ultra-high frequencies, it could occur as a result of changes in the state of the intervening atmosphere or, occasionally, reflection from objects, such as aircraft.

multipath reception Reception of a signal over more than one path. This often results in fading (see MULTIPATH FADING) and other undesirable effects, such as ghosting in television reception.

multipath transmission 1. Transmission of a signal over two or more paths. Also see MULTIPATH RECEPTION. **2.** See MULTIHOP PROPAGATION.

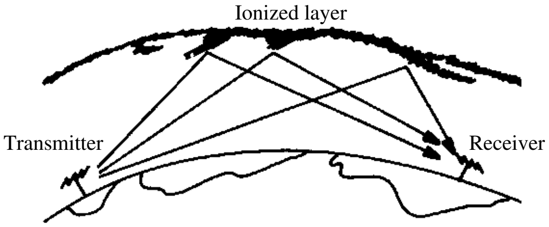

multipath reception

multiphase system See POLYPHASE SYSTEM.

multiple access system A timesharing data processing system that can be used by a number of people at remote locations, usually through peripherals other than conventional terminals (e.g., cash registers linked to a computer system for inventory control).

multiple-address code In computer operations, an instruction code that requires the user to specify the address of more than one operand per instruction.

multiple address instruction A computer program instruction specifying the address of more than one operand.

multiple break **1.** Interruption of a circuit at several points. **2.** Contact bounce.

multiple-chip circuit See HYBRID INTEGRATED CIRCUIT.

multiple connector In a flowchart, a symbol showing the merging of several flowlines.

multiple ionization Successive ionization, as when an ion repeatedly collides with electrons. Also see DOUBLE IONIZATION.

multiple-length number In computer operations, a number occupying more than one register.

multiple-loop feedback system A feedback system using more than one feedback loop. An example is a tunable audio amplifier (see PARALLEL-TEE AMPLIFIER) in which a negative-feedback path provides tuning, while a positive-feedback path sharpens selectivity.

multiple modulation See COMPOUND MODULATION.

multiple precision arithmetic An arithmetic process using several words (bit groups) for an operand to ensure accuracy.

multiple programming **1.** Computer operation in which several programs in memory share peripherals and processor time. **2.** Programming a computer so that several logic or arithmetic operations can be carried out at the same time.

multiple-purpose meter See MULTIMETER.

multiple-purpose tester A multimeter sometimes combined with some other instrument, such as a test oscillator.

multiple reel file In a data-processing system, a magnetic tape data file of more than one reel.

multiple speakers A group of more than two loudspeakers, usually operated from a single amplifier system.

multiple-stacked broadside array A stacked array in which a number of collinear elements are stacked above and below each other. Also see BROADSIDE ARRAY and COLLINEAR ANTENNA.

multiple station Descriptive of a communications network in which more than one terminal is used.

multiple-unit steerable antenna A shortwave antenna system intended to prevent or minimize the effects of fading in received signals. It consists principally of a number of rhombic antennas feeding the receiver. The system utilizes the various waves arriving at different angles. An electric steering system causes the antennas to be selected automatically in the best combination and with their outputs in the proper phase. Also see DIVERSITY RECEPTION.

multiple winding See DRUM WINDING.

multiplex Pertaining to a communications or broadcasting system in which *multiplexing* is used. See MULTIPLEXING, **1**, **2**, **3**.

multiplex adapter A special circuit (or auxiliary unit) used in frequency-modulated radio receivers for stereophonic reception from a station transmitting a multiplex broadcast. Also see MULTIPLEXING, **1** and MULTIPLEX STEREO.

multiplex code The transmission of multiple signals over a single medium, using a code (such as Morse, Baudot, or ASCII). See MULTIPLEXING, **2**, **3**.

multiplex data terminal A computer terminal acting as a modem by virtue of its accepting and transferring signals between its input/output devices and a data channel.

multiplexer A device that allows two or more signals to be transmitted simultaneously on a single carrier wave, communications channel, or data channel. See MULTIPLEXING, **2**, **3**.

multiplexing **1.** A process in which a comparatively low-frequency carrier is modulated, then mixed with a signal that has a higher frequency. Also see SUBCARRIER and SUBSIDIARY COMMUNICATION AUTHORIZATION. **2.** The simultaneous transmission of numerous relatively low-frequency signals on a single carrier having a higher frequency. Also called *frequency-division multiplexing.* **3.** The transmission of numerous signals on a single channel by breaking each signal into timed fragments (intervals) and transmitting the fragments in a rotating sequence. Also called *time-division multiplexing.* **4.** In certain digital light-emitting-diode (LED) displays, the lighting of various parts of the display in a rapidly rotating sequence.

multiplex stereo The use of *multiplexing* to broadcast both channels of a stereophonic program on a single carrier wave. See MULTIPLEXING, **1**, **2**.

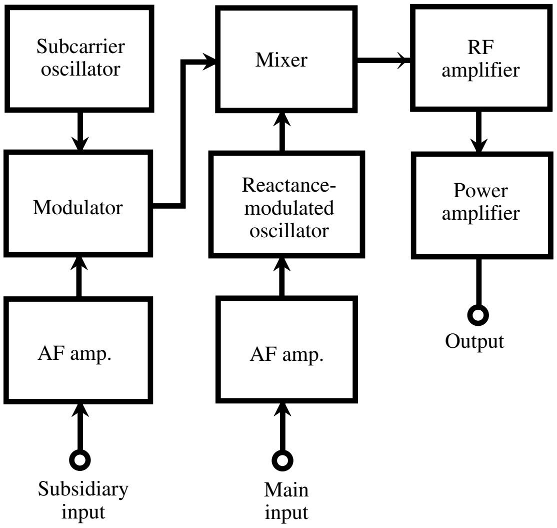

multiplexing, 1.

multiplex telegraphy **1.** A system of wire telegraphy in which two or more messages are sent simultaneously in one or both directions. **2.** A system of radiotelegraphy in which two or more messages can be sent simultaneously on the same carrier wave.

multiplex telephony **1.** A system of wire telephony in which two or more messages can be sent simultaneously in one or both directions over the same line. **2.** A system of radiotelephony in which two or more messages can be sent simultaneously on the same carrier wave.

multiplication **1.** The arithmetic process whereby a certain factor is added to itself the number of times indicated by another factor (the multiplier). **2.** A method of increasing a quantity, magnitude, or rate by some desired factor. See, for example, FREQUENCY MULTIPLIER and VOLTAGE MULTIPLIER.

multiplier **1.** See FREQUENCY MULTIPLIER. **2.** See VOLTAGE MULTIPLIER. **3.** A circuit or device for performing arithmetic multiplication. **4.** See VOLTMETER MULTIPLIER.

multiplier amplifier A frequency-multiplying amplifier (such as a doubler, tripler, or quadrupler), whose output circuit is tuned to an integral multiple of the input frequency. Compare STRAIGHT-THROUGH AMPLIFIER.

multiplier phototube See PHOTOMULTIPLIER TUBE.

multiplier prefix A prefix that, when affixed to the name of a quantity (e.g., Hertz and byte) indicates the amount, usually a power of 10 or a power of 2, by which that quantity is to be multiplied. Also called *prefix* and *prefix multiplier.* See PREFIXES table in Appendix B. Also see, for example, GIGA-, KILO-, MEGA-, MICRO-, MILLI-, NANO-, and PICO-.

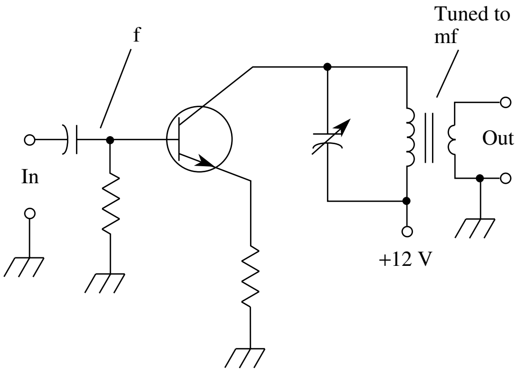

multiplier amplifier

multiplier probe **1.** A resistor-type test probe for a voltmeter, which increases the range of the meter and, therefore, acts as a voltmeter multiplier. **2.** A voltage-multiplier type of test probe for a voltmeter, which multiplies the test voltage before it is applied to the instrument (e.g., voltage-doubler probe).

multiplier register In a computer, a register that retains a number during a multiplication operation.

multiplier resistor A resistor connected in series with a current meter (usually a milliammeter or microammeter) to make it a voltmeter. Also called VOLTMETER MULTIPLER.

multiplier tube See PHOTOMULTIPLIER TUBE.

multipoint circuit A transmission circuit or system, in which information can be entered or retrieved at two or more locations.

multipoint distribution service In radio broadcasting, the transmission of programs via microwave links from a central facility to various local stations.

multipolar Pertaining to the presence of more than two magnetic poles.

multipolar machine A motor or generator having a number of poles in its armature and field.

multiposition relay A relay having more than two positions of closure. See, for example, SELECTOR RELAY.

multiposition switch A switch having more than two contacting positions. See, for example, SELECTOR SWITCH and STEPPING SWITCH.

multiprocessing **1.** In a computer system, the running of more than one program at once. **2.** The use of more than one microprocessor simultaneously in a single computer. Also called *parallel processing.* **3.** See TIME SHARING.

multiprogramming Also called *multitasking.* In a computer system, a technique that allows two or more programs to be executed at the same time.

multipurpose meter An electric meter that performs several functions, usually available through a function selector. Examples: VOLTAMMETER, VOLT-OHMMETER, and VOLT-OHM-MILLIAMMETER.

multirange instrument An instrument provided with several separate ranges, usually selectable by range switching (e.g., five-range voltmeter, two-scale ammeter, and four-band oscillator).

multisection filter A filter having two or more selective sections connected in cascade.

multisegment magnetron See MULTICAVITY MAGNETRON.

multiskip propagation See MULTIHOP PROPAGATION.

multistage device A device having several stages operating in cascade or otherwise coordinated with each other (e.g., a five-stage amplifier).

multistage feedback Feedback (positive or negative) between several stages in a system, as opposed to feedback between the output and input of a single stage.

multistage oscillation Oscillation resulting from positive feedback between or among two or more stages of an amplifier chain, as opposed to oscillation occurring between the output and input of a single amplifier stage.

multistage X-ray tube An X-ray tube providing electron acceleration by means of successive ring-shaped anodes—each biased to a higher voltage than the preceding one.

multiswitch A switch having a number of poles and contacts.

multitester An instrument, such as a multimeter or a combined signal generator and oscilloscope, that performs a number of different test functions.

multitrack recording **1.** A recording on two or more tracks (e.g., *multitrack disk* and *multitrack tape*). **2.** Making a recording on a tape or disc with two or more tracks.

multiturn loop antenna A small loop antenna (whose diameter measures less than approximately 10 degrees of phase in free space) that has two or more turns of wire or tubing. The loop has significant inductance. A capacitor, connected in parallel or in series with the loop, can produce a narrowband frequency response.

multiturn potentiometer A potentiometer whose shaft must be rotated through several complete revolutions to cover the full resistance range.

multivalent Having a valence greater than 2. Compare UNIVALENT.

multivibrator A circuit usually containing two transistors in a resistance-capacitance (RC) coupled amplifier, whose output is capacitance-coupled to the input. The two stages switch each other alternately in and out of conduction at a frequency determined by the *R* and *C* values. Also see ASTABLE MULTIVIBRATOR, BISTABLE MULTIVIBRATOR, and MONOSTA-BLE MULTIVIBRATOR.

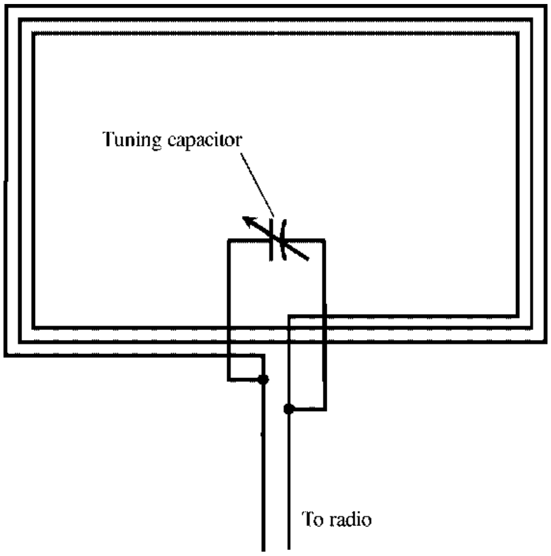

multiturn loop antenna

multiwire antenna An antenna having more than one wire in the radiating section. In early flat-top antennas, such wires were usually connected together at one end.

multiwire doublet antenna A doublet antenna having more than one wire in its radiator. A common form is the *folded dipole*, in effect two closely spaced dipole radiators connected together at both ends, one being center-fed.

Mumetal A high-permeability alloy of iron and nickel, valued especially for use as a magnetic shield for cathode-ray tubes.

Munsell color system A system for specifying colors in terms of hue, saturation, and brilliance, according to charts. Also see COLOR MATCHING.

Muntz metal See YELLOW METAL.

MUPO Abbreviation of MAXIMUM UNDISTORTED POWER OUTPUT.

Murray-loop bridge A specialized form of Wheatstone bridge in which two of the resistance arms are supplied by a two-wire line (such as a telephone line). By means of resistance measurements made with the bridge, the distance from an office or station to a ground fault on the line can be determined.

MUSA Abbreviation of *multiple-unit steerable antenna.*

muscovite Formula, $KH_2Al_2(SiO_4)_3$. A high-grade variety of mica having low dielectric loss.

Musical Instrument Digital Interface Acronym, MIDI. A computer language used in electronic music. It “tells” the computer when to play a note, how long to play it, and how loud to play it. It also sets the tempo of the music, based on how long a quarter-note lasts. It controls the operation of a music synthesizer, and allows two or more synthesizers to communicate.

musical quality See TIMBRE.

musical scale A series of tones between a given tone and its second harmonic, scaled in intervals of ⅛ octave. The most common example can be found on a piano. Variations exist (e.g., *13-note scale*, *six-note scale*, and *five-note scale*).

music chip An integrated circuit for producing various musical effects (such as tones, percussion, etc.).

music power For a power amplifier, the short-term output power obtained in the reproduction of music waveforms, in contrast to root-mean-square (rms) or effective power output.

music synthesizer A set of oscillators, usually operated with a computer, used to create or playing electronic music. Also see MOOG SYNTHESIZER and MUSIC INSTRUMENT DIGITAL INTERFACE.

music under Pertaining to low-volume, continuous, background music, often added to presentations, radio and television advertisements, educational programs, etc.

muting **1.** Disabling a receiver or amplifier under no-signal or weak-signal conditions. **2.** Softening or muffling a sound.

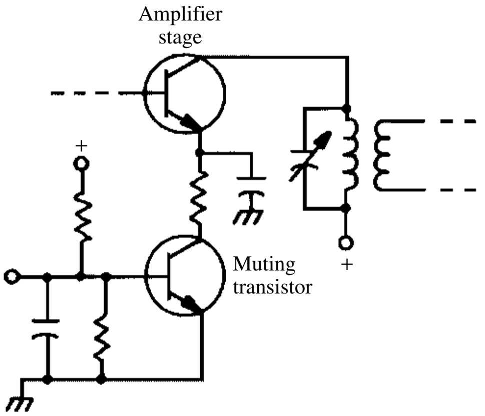

muting

muting circuit **1.** An interference-preventing device that automatically shuts off a radio receiver during the operation of a transmitter, and vice versa. **2.** A squelch circuit. **3.** In a stereo receiver, an electronic element that cuts off all audio when no signal is present or when the receiver is being tuned between carriers.

muting switch A switch or relay that cuts off a receiver during periods of transmission, or when reception is not desired. Generally, a cutoff voltage is applied to one of the intermediate-frequency stages to accomplish muting. In some cases, the audio-frequency circuit is disabled.

mutual antenna coupling Electromagnetic coupling between or among antennas when they are placed too close together. Usually, it is an undesirable phenomenon.

mutual capacitance Inherent capacitance between two conductors.

mutual-capacitance attenuator An attenuator that, in its simplest form, is essentially a shielded, two-plate variable capacitor.

mutual conductance See TRANSCONDUCTANCE.

mutual impedance An impedance shared by two or more branches of a circuit.

mutual inductance Symbol, *M.* Unit, henry. The property shared by neighboring inductors or inductive devices that enables induction to occur. A mutual inductance of 1 henry is present when a current change of 1 ampere per second in one inductor induces 1 volt across another inductor. Also see INDUCTANCE.

mutual-inductance attenuator An attenuator consisting essentially of two coupled coils (input and output), whose spacing can be gradually changed.

mutual-inductance bridge See CAREY-FOSTER MUTUAL-INDUCTANCE BRIDGE.

mutual induction The action whereby the magnetic field produced by alternating current in one conductor produces a voltage in another isolated conductor.

mutual interference **1.** See ADJACENT-CHANNEL INTERFERENCE. **2.** Any kind of interference between or among radio-frequency communications circuits.

mutually exclusive events Two or more events (or data points) so that the occurrence of one prevents the occurrence of the other(s). That is, two events cannot take place simultaneously.

MV **1.** Abbreviation of MEGAVOLT. **2.** Abbreviation of MULTIVIBRATOR.

mV Abbreviation of MILLIVOLT.

MV Abbreviation of MEDIUM VOLTAGE.

MVA Abbreviation of MEGAVOLT-AMPERE.

mV/m Abbreviation of millivolts per meter.

mVP Abbreviation of MILLIVOLT POTENTIOMETER.

MW Symbol for MEGAWATT.

mW Symbol for MILLIWATT.

mw Abbreviation of MEDIUM WAVE.

MWH Abbreviation of MEGAWATT-HOUR.

mW RTL Abbreviation of *milliwatt resistor-transistor logic* or *low-power resistor-transistor logic.*

Mx Abbreviation of MAXWELL.

Mycalex Trade name for an insulating material consisting of mica bonded with glass. Dielectric constant, 6 to 8. Resistivity, 10^{13} ohm-cm.

Mylar A DuPont registered trademark. A tough, plastic insulating material commonly used as a magnetic tape base. Dielectric constant, 2.8 to 3.7. Dielectric strength, 7000 V/mil.

Mylar capacitor A capacitor in which the dielectric film is Mylar.

Mylar tape Magnetic recording tape using a Mylar film as the substrate.

mym Abbreviation of MYRIAMETER.

myoelectricity **1.** Bioelectric pulses of 10- to 1000-mV amplitude produced by muscular activity and detectable by electrodes attached to the skin. Also see ELECTROMYOGRAM, ELECTROMYOGRAPH, and ELECTROMYOGRAPHY. **2.** The study of phenomena, as defined in **1**.

myogram See ELECTROMYOGRAM.

myograph See ELECTROMYOGRAPH.

myography See ELECTROMYOGRAPHY.

myriameter Abbreviation, mym. A metric unit of linear measure equal to 10 kilometers (10^4 meters).

myriametric waves British designation for electromagnetic energy having wavelengths from 100 kilometers down to 10 kilometers, corresponding to frequencies from 3 kHz up to 30 kHz.

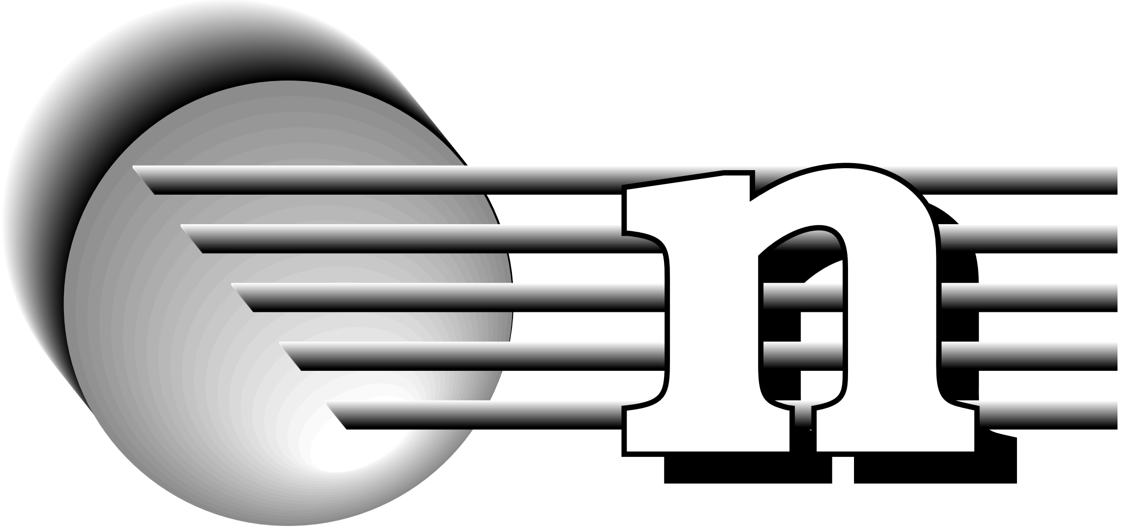

N **1.** Symbol for NUMBER. **2.** Symbol for NITROGEN. **3.** Abbreviation for NEWTON. **4.** Abbreviation of NUMBER (also, No.). **5.** Symbol for the set of natural (counting) numbers (0, 1, 2, 3, . . .).

n **1.** Abbreviation of prefix NANO-. **2.** Symbol for NUMBER. **3.** Symbol for a term (e.g., a multiplier) having an assigned value. **4.** Symbol for part of an expression or operator (as in $2n$, n - 1, n^2, etc.). **5.** Symbol for INDEX OF REFRACTION. **6.** Symbol for amount of substance (unit, mole).

NA **1.** Abbreviation of *not available* (as on a specifications sheet). **2.** Abbreviation of *not applicable.*

N_A Symbol for AVOGADRO'S CONSTANT.

Na Symbol for SODIUM.

nA Abbreviation of NANOAMPERE.

NAB Abbreviation of NATIONAL ASSOCIATION OF BROADCASTERS.

NAB curve In audio-frequency operations, the standard magnetic tape playback equalization curve developed by the National Association of Broadcasters (NAB).

NAE Abbreviation of *National Academy of Engineering.*

NAND circuit Also called *NOT-AND circuit.* A binary digital circuit whose output is high (logic 1) if any of the input signals are low (logic 0); the output is low only when all the inputs are high. The performance of the NAND circuit is the inverse of that of the AND circuit.

NAND gate A gate that performs the function of a NAND circuit.

nano- Abbreviation, n. A prefix meaning billionth(s) (i.e., 10^{-9}).

nanoampere Abbreviation, nA. A unit of current; 1 nA equals 10^{-9} ampere.

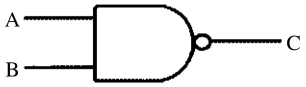

A	B	C
0	0	1
0	1	1
1	0	1
1	1	0

NAND circuit

nanofarad Abbreviation, nF. A unit of capacitance; 1 nF equals 10^{-9} farad.

nanohenry Abbreviation, nH. A unit of low inductance; 1 nH equals 10^{-9} henry.

nanometer Abbreviation, nm. A unit of short wavelength or displacement; 1 nm equals 10^{-9} meter.

nanosecond Abbreviation, ns. A time interval of 10^{-9} second.

nanotechnology The design and construction of superminiature electronic and electromechanical devices. For example, microscopic, computer-controlled robots might someday be used as programmable antibodies to fight infectious diseases, or to build integrated circuits, atom by atom.

nanovolt Abbreviation, nV. A unit of low voltage; 1 nV equals 10^{-9} volt.

nanovoltmeter A sensitive voltmeter for measuring potential in the nanovolt range.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

nanowatt Abbreviation, nW. A unit of low power; 1 nW equals 10^{-9} watt.

nanowattmeter A meter for measuring power in the nanowatt range.

NAP Abbreviation of *nuclear auxiliary power.*

napier See NEPER.

Napierian base Symbol, *e.* An irrational number, equal to approximately 2.718282, and used as the NAPIERIAN LOGARITHM base.

Napierian logarithm Abbreviation, ln or $\log_e$. A logarithm to the base *e* (see NAPIERIAN BASE). Also called NATURAL LOGARITHM. Compare COMMON LOGARITHM.

NAPU Abbreviation of *nuclear auxiliary power unit.*

narrative A computer program statement that, rather then being an instruction, merely describes the purpose of what follows (usually the steps in a routine or a block of instructions) as a debugging or program modification aid. In a program written in BASIC, such a statement is preceded by the abbreviation REM, for remark. Also called COMMENT.

narrowband **1.** A frequency band in which the difference between upper and lower limits is small compared with bandwidths typical of the service specified. **2.** Pertaining to a radio-frequency emission whose bandwidth is limited or restricted. **3.** Pertaining to a circuit or device that operates over a comparatively small range of frequencies. **4.** In a digital network, a channel having a bandwidth of 64 kbps or less.

narrowband amplifier An amplifier whose passband is restricted to a fraction of the frequency spread common to the amplifier's application.

narrowband FM See NARROWBAND FREQUENCY MODULATION.

narrowband frequency modulation Abbreviation, NBFM or NFM. Frequency modulation in which the maximum instantaneous frequency deviation is no greater than the highest modulating frequency.

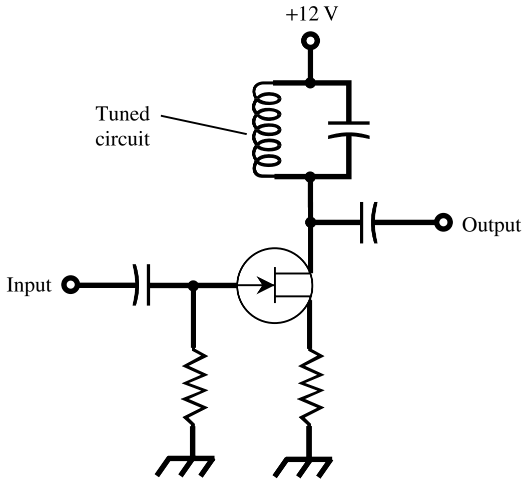

narrowband amplifier

narrowband interference Signal interference whose bandwidth is narrow, compared with that of the circuit affected.

narrowband voice modulation Abbreviation, NBVM. A scheme via which an analog voice signal is compressed into a band of frequencies narrower than that normally required for effective communication.

narrow bandwidth Pertaining to a radio-frequency emission whose bandwidth is comparatively narrow (e.g., a small fraction of the available spectrum space in the frequency band being used).

narrow-sector recorder A directional radio receiver for locating sources of atmospheric noise.

NARTB Abbreviation of *National Association of Radio and Television Broadcasters.*

NAS Abbreviation of *National Academy of Sciences.*

National Association of Broadcasters Abbreviation, NAB. A countrywide organization of radio and television broadcasters.

National Bureau of Standards Abbreviation, NBS. An agency in the U.S. that maintains values for physical constants in the Standard International (SI) System of Units. It also maintains radio broadcast stations that transmit standard time and frequency signals.

National Electrical Manufacturers Association Abbreviation, NEMA. A countrywide organization of manufacturers of electrical and electronic equipment and supplies.

National Electric Code Abbreviation, NEC. Safety regulations and procedures issued by the National Fire Protection Association for the installation of electrical wiring and equipment in the United States. Although the code is advisory from the Association's standpoint, it is enforced to various degrees by local authorities.

National Television Standards Committee Abbreviation, NTSC. A U.S. organization of television companies and other interested organizations. It developed the original black-and-white and color television standards that were approved by the Federal Communications Commission.

natural-decay curve See EXPONENTIAL DECREASE.

natural antenna frequency The fundamental resonant frequency of an electromagnetic antenna.

natural disintegration **1.** The decay of a radioactive substance as a result of the continuous emission of particles and rays. **2.** Also called *half life.* The time required for half of a quantity of a radioactive substance to decay into a different isotope or element.

natural electricity **1.** Atmospheric electricity. **2.** The electricity in living organisms, that is, BIOELECTRICITY.

natural frequency See NATURAL RESONANT FREQUENCY.

natural-growth curve See EXPONENTIAL INCREASE.

natural interference Interference from atmospheric and celestial sources, as opposed to human-made interference.

natural language A spoken or written human language (such as French or Japanese), as opposed to a computer language (such as machine language, C++, or LISP). Translation between natural language and machine language is important in optical character recognition, speech recognition, and speech synthesis.

natural logarithm See NAPIERIAN LOGARITHM.

natural magnet A material, such as magnetite (lodestone), found in nature and exhibiting permanent magnetism.

natural magnetism Magnetism found in some natural materials (see NATURAL MAGNET) and in the earth itself.

natural number **1.** Any nonnegative whole number, that is, a member of the set $N = (0, 1, 2, 3, \ldots)$. **2.** See NAPIERIAN BASE.

natural period The time required for one complete wave cycle to occur in a device at its NATURAL RESONANT FREQUENCY.

natural radiation Noise in the form of radiation emitted by natural radioactive substances, cosmic rays, etc. Also called BACKGROUND RADIATION.

natural resonance Resonance resulting from the unique physical constants of a body, circuit, or system. Also see NATURAL RESONANT FREQUENCY.

natural resonant frequency **1.** The frequency at which a circuit or device responds with maximum amplitude to applied signals. **2.** The frequency at which a circuit or device generates maximum energy. **3.** The frequency at which an object vibrates at maximum amplitude.

natural wavelength The wavelength corresponding to the NATURAL RESONANT FREQUENCY of a circuit, device, or object.

nautical mile A nautical unit of linear measure equal to 1.852 kilometers (1.1508 statute miles).

nav **1.** Abbreviation of *navigation.* **2.** Abbreviation of *navigational.*

Navaglobe A radionavigation system used at very-low or low frequencies over long distances.

NAVAIDS Abbreviation of NAVIGATIONAL AIDS.

navar A radar system in which a ground radar scans the immediate vicinity of an airport, observes the flight activity, and transmits such observations to aircraft in flight. The name is a contraction of the term *navigation and ranging.*

navigation aid An electronic device or system, such as radar or radio direction finding, that assists in the navigation of vehicles on land, at sea, or in the air.

navigation beacon A beam that provides aircraft and ships with navigational aid.

NAWAS Abbreviation of *National Attack Warning System.*

Nb Symbol for NIOBIUM.

nb Abbreviation of NARROWBAND.

NBFM Abbreviation of NARROWBAND FREQUENCY MODULATION. Also abbreviated NFM.

NBS Abbreviation for NATIONAL BUREAU OF STANDARDS.

NBTDR Abbreviation of *narrowband time-domain reflectometry.*

nbw Abbreviation of NOISE BANDWIDTH.

NC **1.** Abbreviation of NORMALLY CLOSED. **2.** On drawings, abbreviation of *no connection.* **3.** Abbreviation of NUMERICAL CONTROL.

N/C Abbreviation of NUMERICAL CONTROL.

nc Abbreviation of *no connection.*

n-channel junction field-effect transistor Abbreviation, NFET. A junction field-effect transistor in which the gate junction is formed on a bar or die of n-type semiconductor material.

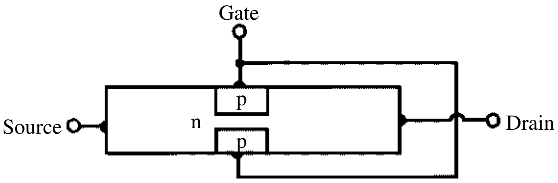

n-channel junction field-effect transistor

n-channel MOSFET A metal-oxide-semiconductor (MOS) field-effect transistor in which the channel is n-type material.

NCMT Abbreviation of NUMERICALLY CONTROLLED MACHINE TOOL.

NCS Abbreviation of *net control station(s).*

N curve A negative-resistance voltage-current curve having a shape roughly resembling an uppercase letter N. Compare S CURVE.

Nd Symbol for NEODYMIUM.

N display A radar display in which the target is represented by a pair of vertical blips, whose amplitude indicates target direction and whose position along the horizontal base line (as determined by lining up a pedestal signal with the blips) is read from the pedestal-adjustment calibration.

n-doped zinc-oxide ceramic A nonlinear-resistance material used in the manufacture of some voltage-dependent resistors.

NDR Abbreviation of NONDESTRUCTIVE READ.

NDT Abbreviation of NONDESTRUCTIVE TEST.

Ne Symbol for NEON.

near-end crosstalk Crosstalk originating at, or near, the telephone subscriber line in which the interference is noticed.

near field **1.** The radiation field within a radius of 1 wavelength from a transmitting antenna. **2.** The sound field near a loudspeaker or other reproducer.

near infrared Pertaining to electromagnetic radiation of 3 to 30 micrometers in wavelength.

near ultraviolet That part of the ultraviolet spectrum nearest the visible light wavelengths. These are the least-penetrating ultraviolet rays. They are also called *soft ultraviolet.*

near video on demand In television, a local service in which various programs/videos are broadcast at specified times.

near zone See NEAR FIELD, **1**.

NEB Abbreviation of *noise equivalent bandwidth.*

NEC Abbreviation of NATIONAL ELECTRIC CODE.

necessary bandwidth The minimum bandwidth needed (with a given emission) to transmit information at a required rate and of a required quality. In general, as data speed increases (as measured in bits per second, baud, or words per minute), the minimum required bandwidth increases. For slow-speed Morse code signals, it can be as small as a few Hertz; for fast-scan television, it is several megahertz.

neck The straight portion of a cathode-ray-tube (CRT) envelope (i.e., the part between the base and the flared portion).

NEDA Abbreviation of *National Electronic Distributors' Association.*

needle **1.** The stylus of a phonograph cartridge. **2.** The pointer of an indicating meter. **3.** One of the electrodes in a voltage-measuring spark gap (see NEEDLE GAP). **4.** The slender, pointed metal tip of a test probe.

needle chatter See NEEDLE TALK.

needle drag In disc recording and reproduction, friction between the needle (stylus) and disc. Also called *stylus drag.*

needle electrode See NEEDLE, **3**, **4**.

needle gap A spark gap composed of two needles having an adjustable air gap between their points. An unknown high voltage is measured in terms of the gap width necessary for sparking.

needle memory A computer memory in which the dual-state elements are thin magnetic needles.

needle pointer The pointer of a meter or compass.

needle pressure See VERTICAL STYLUS FORCE.

needle probe See NEEDLE-TIP PROBE.

needle scratch In disc recording and reproduction, noise resulting from vibration of the needle (stylus) because of an irregular groove surface. Also called *surface noise.*

needle talk Direct radiation of sound by the stylus of a phonograph pickup. Also called *needle chatter.*

needle telegraph A telegraph in which the Morse code characters are converted into magnetic needle deflections.

needle test point See NEEDLE, **4**.

needle-tip probe A test probe that terminates in a sharp point. Also see NEEDLE, **4**.

needle voltmeter See NEEDLE GAP.

neg Abbreviation of NEGATIVE.

negate **1.** To insert the NOT operation in front of a digital expression. **2.** To change logic 1 to logic 0 (high to low) or vice versa.

negation The logical NOT operation (see NOT) in digital systems. Also see NAND CIRCUIT, NOR CIRCUIT, NOR GATE, NOT CIRCUIT, and NOT-OR CIRCUIT.

negation element In a computer system, a device that can give the reverse of a condition, event, or signal.

negative **1.** Possessing NEGATIVE ELECTRIFICATION. **2.** In voice communications, a word often used for "no," especially when interference is present or signals are weak. **3.** Pertaining to a real number less than zero. **4.** An image whose shadings are opposite to those in the scene.

negative acceleration A decrease in speed or velocity; also called *deceleration.*

negative acknowledgment character In a handshaking or forward-error-correction (FEC) system, a response by the receiving station that indicates a missed bit or bits.

negative angle **1.** An angle in the third or fourth quadrant in a system of rectangular coordinates. **2.** An angle measured clockwise from the positive *x*-axis in a rectangular coordinate system. Compare POSITIVE ANGLE.

negative bias A steady, negative direct-current voltage or current applied continuously to an electrode of a device, such as a transistor to establish the operating point. Compare POSITIVE BIAS.

negative bus See NEGATIVE CONDUCTOR.

negative charge An electric charge consisting of a quantity of NEGATIVE ELECTRIFICATION. Also see CHARGE, **1**; ELECTRIC CHARGE; and UNIT ELECTROSTATIC CHARGE. Compare POSITIVE CHARGE.

negative conductor The conductor or wire connected to the negative terminal of a current, power, or voltage source. Compare POSITIVE CONDUCTOR.

negative electricity See NEGATIVE CHARGE and NEGATIVE ELECTRIFICATION.

negative electrification Electrification characterized by an excess of electrons. For example, when a glass rod is rubbed with a silk cloth, the cloth becomes negatively charged because electrons are transferred from the glass to the cloth. Similarly, when a neutral atom acquires an extra electron, the atom becomes negatively charged because it has an excess of electrons. Compare POSITIVE ELECTRIFICATION.

negative electrode **1.** An electrode connected to the negative terminal of a current, power, or voltage source. **2.** The negative output terminal of a current, power, or voltage source, such as a battery or generator.

negative error of measurement An error of measurement in which the difference between a measured value and the true or most probable value is negative. Compare POSITIVE ERROR OF MEASUREMENT.

negative exponent In mathematical notation, an exponent indicating that a number is to be raised to a negative power. Raising a number *x* to a negative power $-n$ means taking the reciprocal of the number raised to the power *n*. That is, $x^{-n} = 1/x^n$; for example, $10^{-2} = 1/10^2 = 1/100 = 0.01$.

negative feedback Feedback that is out of phase with the input signal. Also called INVERSE FEEDBACK, DEGENERATION, and DEGENERATIVE FEEDBACK. Compare POSITIVE FEEDBACK.

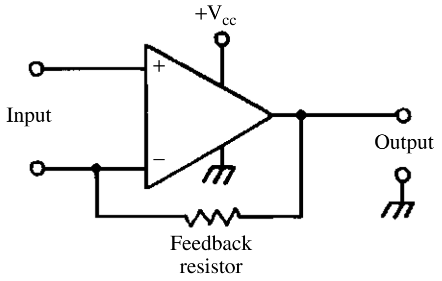

negative feedback

negative-feedback amplifier An amplifier in which negative feedback is used to improve performance or modify response.

negative function A trigonometric function having the negative sign. In a rectangular coordinate system, the sine function is negative in the third and fourth quadrants, the cosine in the second and third, and the tangent in the second and fourth. Compare POSITIVE FUNCTION.

negative gain A misnomer for FRACTIONAL GAIN, arising because fractional gain can be expressed in negative decibels.

negative ghost In a television picture, a ghost with negative (see NEGATIVE, **4**) shading.

negative ground In a direct-current power system, the connection of the negative pole to common ground.

negative image **1.** A picture in which the blacks, whites, and shades in between are the reverse of those in the scene (see NEGATIVE, **4**). **2.** An abnormal image that has the reverse shading described in **1**.

negative impedance An impedance that displays the same behavior as that of NEGATIVE RESISTANCE.

negative ion An atom with an excess of electrons and, consequently, a net negative charge. Also called ANION. Compare POSITIVE ION.

negative-ion generator A device for generating negative ions and circulating them into the surrounding air. This is believed by some people to be beneficial to health.

negative lead See NEGATIVE CONDUCTOR.

negative-lead filtering Power-supply filtering in which the choke coils and capacitors are in the negative direct-current lead, rather than in the positive lead (the usual position). One advantage of this arrangement is the lower insulation requirement of the choke.

negative light modulation In television image transmission, the condition in which transmitted power is increased by a decrease in the initial intensity of light. Compare POSITIVE LIGHT MODULATION.

negative line See NEGATIVE CONDUCTOR.

negative logic **1.** Binary logic in which a high negative state represents logic 1, and a low negative state represents logic 0. **2.** Binary logic in which a low positive state represents logic 1, and a high positive state represents logic 0. Compare POSITIVE LOGIC.

negative measurement error See NEGATIVE ERROR OF MEASUREMENT.

negative modulation Amplitude-modulated television transmission in which the transmitted power decreases as image brightness increases. Compare POSITIVE MODULATION.

negative modulation factor For an amplitude-modulated wave having unequal positive and negative modulation peaks, a ratio expressing the maximum negative deviation from the average for the envelope. Compare POSITIVE MODULATION FACTOR.

negative number A real number less than zero (i.e., one to which the minus sign is assigned).

negative peak The maximum negative instantaneous current or voltage in an alternating-current waveform.

negative-peak voltmeter An electronic meter for measuring the NEGATIVE PEAK voltage of an alternating-current waveform.

negative phase-sequence relay A relay that responds to the negative phase sequence in a polyphase circuit. Compare POSITIVE PHASE-SEQUENCE RELAY.

negative picture modulation See NEGATIVE MODULATION.

negative picture phase In a television signal, the picture-signal voltage swing from zero to negative, in response to an increase in brightness. Compare POSITIVE PICTURE PHASE.

negative plate The negative member of a battery cell; electron flow is from the plate through the external circuit.

negative pole See NEGATIVE ELECTRODE, **1**, **2**.

negative positive zero Abbreviation, NP0. Pertaining to temperature-compensating capacitors having a temperature coefficient of capacitance that changes sign within a specified temperature range.

negative potential **1.** The potential measured at a negative electrode, with respect to the positive electrode or to ground. **2.** Potential less than that of the earth as a reference.

negative resistance **1.** A decrease in voltage across a device as the current through it increases, and an increase in voltage as the current decreases. **2.** A decrease in current through a device as the voltage across it increases, and an increase in current as the voltage decreases. This is opposite to the behavior of an ohmic (positive) resistance. Also see N-CURVE, NEGATIVE-RESISTANCE REGION, NEGATIVE RESISTOR, and S CURVE.

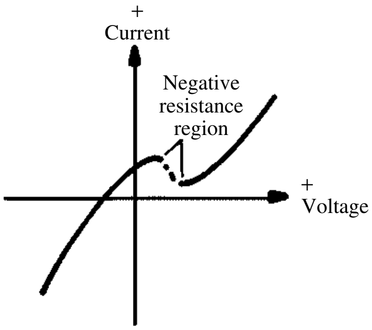

negative resistance

negative-resistance amplifier A simple circuit in which a negative-resistance device, such as a TUNNEL DIODE, cancels the positive resistance of the circuit, causing amplification or oscillation.

negative-resistance device See NEGATIVE RESISTOR.

negative-resistance diode **1.** See TUNNEL DIODE. **2.** A reverse-biased germanium diode (and occasionally a silicon diode) that exhibits NEGATIVE RESISTANCE. **3.** A special diode tube that, when operated at ultrahigh frequencies, exhibits negative resistance because of transit-time effects.

negative-resistance magnetron A split-anode magnetron operated at a combination of anode voltage and magnetic field strength corresponding to cutoff; it exhibits negative resistance to voltage applied symmetrically between the anode halves. The frequency of oscillation is determined by an external tank circuit.

negative-resistance oscillator An oscillator that consists of a negative-resistance device connected across a tuned circuit. The arrangement oscillates because the negative resistance cancels the positive resistance (losses) of the tuned circuit. See, for example, NEGATIVE-RESISTANCE MAGNETRON and TUNNEL-DIODE OSCILLATOR.

negative-resistance region In certain devices, a portion of the voltage-versus-current curve having negative slope (i.e., the current decreases as the voltage increases, or vice versa). Also see N CURVE, NEGATIVE RESISTANCE, and S CURVE.

negative-resistance repeater A repeater that produces gain by means of NEGATIVE RESISTANCE effects.

negative resistor Any device exhibiting NEGATIVE RESISTANCE.

negative space charge The cloud of electrons (negative particles) in the region surrounding an emitter, such as the hot cathode of a vacuum tube.

negative temperature coefficient Abbreviation, NTC. A number expressing the amount by which a quantity (such as the rating of a component) decreases when the temperature is raised. The coefficient is stated as a percentage, or as a certain number of parts per million (ppm) per degree temperature rise. Compare POSITIVE TEMPERATURE COEFFICIENT and ZERO TEMPERATURE COEFFICIENT.

negative terminal See NEGATIVE ELECTRODE, **2**.

negative torque In an electric motor, a torque that acts against the operating torque.

negative transmission In a television or facsimile system, the condition in which brighter light corresponds to lower transmitted power, and dimmer light corresponds to higher transmitted power.

negative valence The valence of a negative ion. Also see VALENCE.

negator A logical NOT element [i.e., one that outputs the complement of an input bit (1 for 0 and vice-versa)].

negatron The term that specifically differentiates the familiar ELECTRON from a POSITRON (positive electron).

NEI Abbreviation of *noise equivalent input.*

neighborhood node A local site at which a fiberoptic television (video) signal is converted to a cable signal.

NEL Abbreviation of *National Electronics Laboratory.*

NELA Abbreviation of *National Electric Light Association.*

N electron In certain atoms, an electron whose orbit is outside of, and nearest to, those of the M electrons.

NEMA Abbreviation of *National Electrical Manufacturers' Association.*

nematic crystal A normally transparent liquid crystal that becomes opaque when an electric field is applied to it, and becomes transparent again when the field is removed. The crystal material is cut in the form of a letter or numeral and provided with a reflecting backplate for display readouts in calculators, watches, and various electronic devices.

nematic-crystal display A device in which an electrically controlled film of nematic-crystal material is used to transmit and interrupt light from a lamp or from a reflecting mirror, in this way displaying characters in whose shape the film has been formed.

nematic liquid In a liquid-crystal display (LCD), a normally clear liquid that becomes opaque in the presence of an electric field.

nemo A radio or television program that is picked up from a location outside the studio. Also called *field pickup* or *remote.*

neodymium Symbol, Nd. A metallic element of the rare-earth group. Atomic number, 60. Atomic weight, 144.24.

neodymium-YAG laser A laser that uses neodymium and yttrium-aluminum-garnet (YAG) crystal. It is used in medical applications, such as surgery and various other jobs where high precision is required. Generally, it is a low-to-medium-power laser.

neomatachograph An instrument used in psychology to measure complex reaction time. Compare NEOMATACHOMETER.

neomatachometer An instrument used in psychology to measure simple reaction time. Compare NEOMATACHOGRAPH.

neon Symbol, Ne. An inert-gas element. Atomic number, 10. Atomic weight, 20.179. Neon, present in trace amounts in the earth's atmosphere, is used in some glow tubes, readout devices, and indicator lamps.

neon bulb A (usually small) neon-filled gas diode. It has a characteristic pink glow and is ignited by a firing voltage for the particular unit. Also called *neon glow lamp* and *neon tube.*

neon-bulb flip-flop A flip-flop circuit (bistable multivibrator) using two neon bulbs as the bistable components.

neon-bulb gate A gate circuit containing a neon bulb biased below the firing point. A trigger voltage added to the bias voltage raises the applied voltage and fires the bulb, producing an output pulse.

neon-bulb logic Logic circuits composed of neon-bulb gates.

neon-bulb memory See NEON-BULB STORAGE.

neon-bulb multivibrator A multivibrator using two neon bulbs as the switching components.

neon-bulb oscillator A simple relaxation oscillator consisting essentially of a neon bulb, capacitor, resistor, and direct-current supply. The frequency of the sawtooth-wave output depends principally on the capacitance and resistance values. The maximum operating frequency is limited to about 5 kHz by the deionization time of the gas.

neon-bulb overmodulation indicator The application of a neon-bulb overvoltage indicator as a monitor for amplitude-modulated radio-frequency signals. The bulb flashes each time the modulation percentage exceeds a predetermined value. Also called *neon-bulb modulation alarm.*

neon-bulb overvoltage indicator A relatively simple circuit in which a neon bulb flashes each time a voltage monitored by the circuit exceeds a predetermined value. The flash shows that the portion of the voltage presented to the bulb has exceeded the firing potential.

neon-bulb peak indicator See NEON-BULB OVERVOLTAGE INDICATOR.

neon-bulb ring counter A ring counter composed of neon-bulb flip-flops. The maximum counting speed is limited by the deionization time of the neon bulbs to approximately 300 events per second.

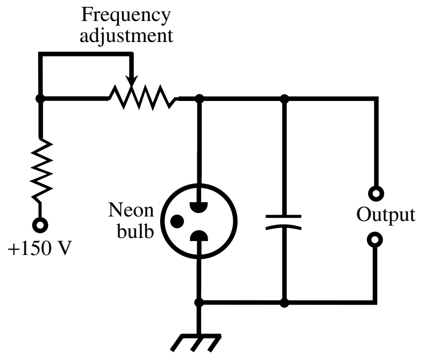

neon-bulb oscillator

neon-bulb sawtooth generator A relatively simple relaxation oscillator using a neon bulb, capacitor, and resistor. The output is a sawtooth wave, whose frequency is determined principally by the capacitance and resistance values.

neon-bulb scale-of-two circuit A scale-of-two circuit (frequency halver) using neon bulbs as the bistable elements.

neon-bulb storage A storage (memory) device composed of neon bulbs. A fired bulb (representing a bit of stored information) remains fired until turned off by an erase signal.

neon-bulb stroboscope A stroboscope in which a neon bulb supplies the light flashes. The circuit is essentially that of the neon-bulb oscillator, the flash rate being continuously variable by an adjustable frequency control.

neon-bulb voltage regulator A simple circuit utilizing the constant voltage drop across a fired neon bulb as a regulated voltage. The usual circuit configuration is a neon bulb and current-limiting resistor in series with a power supply.

neon-bulb volume indicator A neon-bulb overvoltage indicator used in some tape recorders to show when the volume exceeds a predetermined level—especially when the volume is high enough to cause an unacceptable amount of distortion.

neon glow lamp See NEON BULB.

neon lamp See NEON BULB.

neon pilot lamp A neon bulb used as a pilot lamp operated from the power-line circuit of an electronic equipment. Also called *neon pilot light.*

neon tube See NEON BULB.

NEP Abbreviation of NOISE EQUIVALENT POWER.

NEPD Abbreviation of *noise equivalent power density.*

neper Abbreviation, Np. A Napierian-logarithmic unit expressing a ratio of power levels; $Np = \log_e (P_1/P_2)^{1/2}$, where P_1 and P_2 are the power values being compared and e is the Napierian logarithm base, equal to approximately 2.71828. The neper is related to the decibel (dB), a similar unit based on common logarithms, in the following manner: 1 Np = 8.686 dB; 1 dB = 0.1151 Np.

neptunium Symbol, Np. A radioactive metallic element produced artificially. Atomic number, 93. Atomic weight, 237.05.

Nernst effect The appearance of a voltage between the opposite edges of a metal strip that is conducting heat longitudinally when the strip is placed in a magnetic field perpendicular to the plane of the strip.

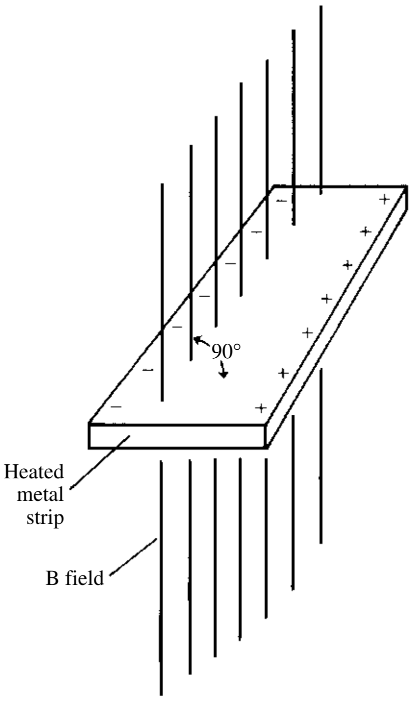

Nernst effect

Nernst-Ettinghausen effect In a piezoelectric crystal, the tendency for a temperature gradient to exist as a result of applied electromagnetic fields.

Nernst lamp An incandescent lamp in which the filament material is a mixture of magnesia and certain rare oxides (such as yttria). A surrounding current-carrying coil heats the filament, lowering its resistance and causing it to glow, after which it continues to operate on a low current.

nerve center The most essential part of a control system or communication network.

nerve current In an animal organism, the small electrical current that flows along nerve fibers.

nervous breakdown The upset of a function in a digital computer (which becomes disabled as a result).

NESC Abbreviation of *National Electrical Safety Code.*

nested loops See NESTING.

nesting In digital computers and data processing, the inclusion of one or more loops or subroutines inside a larger loop or subroutine. Nested loops, for example, can be used to assign values to double-subscripted variables.

NET Abbreviation of *noise equivalent temperature.*

net **1.** An effective, useful, or resultant quantity (e.g., NET CAPACITANCE). **2.** Colloquialism for *communications network.* See NETWORK, **2**.

net authentification A password or other special identification used in a radio network.

net capacitance The resultant capacitance in a circuit in which capacitances act in combination with each other.

net component The total value of two or more passive components of the same sort. See, for example, NET CAPACITANCE, NET CURRENT, NET IMPEDANCE, NET INDUCTANCE, NET POWER, NET REACTANCE, NET RESISTANCE, and NET VOLTAGE.

net current The current flowing in a circuit in which currents aid or oppose each other.

net gain For an amplifier, the amount of gain remaining after all losses in the device have been subtracted.

net impedance The impedance of a circuit in which impedances act in combination with each other.

net inductance The inductance of a circuit in which inductances act in combination with each other.

net loss For an amplifier or other system, the algebraic sum of gains and losses between two points in the system.

net power The resultant power observed when power signals aid or oppose each other in a single circuit or system.

net radiometer A device for measuring the difference in intensity between radiation entering the earth's surface and radiation leaving it. This is generally visible light or infrared radiation from the sun.

net reactance Symbol, X_t. The combined inductive reactance (X_L) and capacitive reactance (X_C) in a circuit or device.

net resistance The resistance of a circuit in which resistances act in combination with each other.

net voltage The resultant voltage at a point where voltages aid or oppose each other.

network **1.** A circuit arrangement of electronic components, sometimes redundant in its design [e.g., resistance-capacitance (RC) network]. **2.** A group of interconnected computers, communications stations, or other facilities, often organized for simultaneous operation and data transfer. **3.** To conduct research or gather information using a group of interconnected computers, communications stations, or other facilities.

network analog A circuit or circuits representing variables and used to express and solve a mathematical relationship between the variables.

network analysis The rigorous examination of a network to determine its properties and performance. Compare NETWORK SYNTHESIS.

network analyzer An analog or digital circuit for simulating and analyzing a network (see NETWORK, **1**) by means of measurements and calculations from the model so obtained.

network assurance The construction and operation of a network in such a manner that failures can be quickly and easily located and corrected.

network calculator An analog or digital device for determining the component values and performance of a given network (see NETWORK, **1**).

network constant The value of a passive component (capacitance, inductance, resistance, etc.) used in a network (see NETWORK, **1**).

network filter A transducer that passes or rejects signals, depending on their frequency.

network interface unit Abbreviation, NIU. A device, provided to each subscriber, that connects telephones, television sets, and personal computers to an electrical or fiberoptic cable. Envisioned as an integral component of the multimedia INFORMATION SUPERHIGHWAY.

network provisioning The control of a network by means of computer software, rather than via manual rearrangement of system devices.

network relay **1.** A relay that provides protection of the circuits in a network (see NETWORK, **1**); a circuit breaker. **2.** In a communications network (see NETWORK, **2**), the reception and retransmission of a message by a human operator.

network synthesis The design and fabrication of a network (see NETWORK, **1**) by rigorous engineering methods to achieve a prescribed performance.

network theorems See COMPENSATION THEOREM, MAXIMUM POWER TRANSFER THEOREM, NORTON'S THEOREM, RECIPROCITY THEOREM, SUPERPOSITION THEOREM, and THEVENIN'S THEOREM.

network topology The analysis of networks (see NETWORK, **1**) with signal-flow diagrams. Components of the diagram represent signal paths, open loops, closed loops, and node points.

network transfer function A function describing the overall processing of energy by a network (see NETWORK, **1**); it is equal to V_o/V_i, where V_i is the input voltage and V_o is the output voltage of the network.

Neuman's law A property of mutual inductances. For a given orientation and environment for two inductors, the value of the mutual inductance does not change, regardless of the magnitude, frequency, or phase of the currents in the coils. That is, mutual inductance is subject only to the physical environment surrounding the coils.

neuristor A two-terminal semiconductor device that simulates the behavior of a neuron (nerve cell) and allows machines to duplicate some of the neurological phenomena observed in the human body.

neuroelectricity Low-voltage electricity in the nervous system of a human being or animal.

neuron A nerve cell in a living organism.

neurosurgery assistance robot An electromechanical device, controlled by a human operator and assisted by computers, that helps to perform simple neurosurgical operations requiring extreme precision and steadiness. One procedure for which a robot has been used is the drilling of holes in the skull prior to brain surgery.

neutral **1.** Having no electric charge; thus, an atom is normally neutral because its internal positive charges neutralize its internal negative charges. **2.** Devoid of voltage (e.g., a neutral line). **3.** Pertaining to a chemical, especially a solution, that is neither acidic nor alkaline.

neutral bus See NEUTRAL WIRE.

neutral circuit **1.** A deenergized circuit (i.e., a "dead" one). **2.** In a teletypewriter system, a circuit in which current flows in one direction.

neutral conductor See NEUTRAL WIRE.

neutral ground A ground connection to a neutral wire or to the neutral point of a circuit.

neutralization The process of balancing out positive feedback in an amplifier to prevent self-oscillation. Also see NEUTRALIZING CAPACITOR and NEUTRALIZING COIL.

neutralization indicator A neon bulb or radio-frequency meter used during neutralization of an amplifier. When the amplifier is properly neutralized, the bulb or meter will show zero signal with no input to the amplifier, indicating that the circuit is not oscillating.

neutralize To eliminate positive feedback in a radio-frequency amplifier.

neutralized amplifier An amplifier in which neutralization has been performed to prevent self-oscillation.

neutralizing capacitor In a capacitively neutralized circuit, a small capacitor that serves as a coupler of signal energy from the output back to the input in reverse phase, to cancel self-oscillation of the circuit.

neutralizing circuit Any component or set of components that is used to neutralize a radio-frequency amplifier.

neutralizing coil In an inductively neutralized circuit, a small coupling coil that picks up signal energy from the output and applies it in reverse phase to the input, to cancel circuit self-oscillation.

neutralizing tool A nonconducting screwdriver-like device for adjusting a neutralizing capacitor or coil. It is usually made of fiber or plastic.

neutralizing voltage The feedback voltage that cancels self-oscillation in the process of neutralization.

neutralizing wand See NEUTRALIZING TOOL.

neutral line See NEUTRAL WIRE.

neutral relay See UNPOLARIZED RELAY.

neutral wire In a polyphase power-transmission system, the line (wire) that does not carry current until the system is unbalanced.

neutrodon See NEUTRODYNE.

neutrodyne A radio-frequency amplifier that is neutralized by feeding energy through a small capacitor from a tap on the secondary coil of the collector or drain output transformer back to the base or gate.

neutron An uncharged atomic particle having a mass approximately equal to that of the proton. The neutron is present in the nucleus of every atom, except that of hydrogen.

neutron age 16.7% of the mean square displacement of a neutron as it moves through a specified energy range.

neutron rest mass See MASS OF NEUTRON AT REST.

new candle See CANDELA.

newton (Sir Isaac Newton.) Symbol N. The SI unit of force. A force of 1 newton imparts an acceleration of 1 meter per second per second to a mass of 1 kilogram. The newton is equal to 10^5 dynes.

Newton's laws of motion Three natural laws discovered in 1686 by Sir Isaac Newton. First law: A body at rest or in motion tends to remain in that state, unless it is acted upon by some force. Second law: A body tends to accelerate or decelerate when it is acted upon by a force, the acceleration or deceleration being directly proportional to the force and inversely proportional to the mass of the body. Third law: For every action or acting force, there is an equal and opposite reaction or reacting force.

nexus An interconnection point in a system.

NF Abbreviation of NOISE FIGURE.

nF Abbreviation of NANOFARAD.

NFET Abbreviation of N-CHANNEL JUNCTION FIELD-EFFECT TRANSISTOR.

NFM Abbreviation of NARROWBAND FREQUENCY MODULATION. Also abbreviated NBFM.

NFM reception Reception of a signal having NARROWBAND FREQUENCY MODULATION. Standard discriminators and ratio detectors can be used. Slope detection can be achieved with an amplitude-modulation (AM) receiver by tuning slightly to one side of the center frequency of the FM signal.

NFM transmission Transmission of a signal via NARROWBAND FREQUENCY MODULATION. A simple method consists of frequency modulating the master oscillator of the transmitter at the modulation frequency. Phase modulation can also be used.

NFPA Abbreviation of *National Fire Protection Association* (see NATIONAL ELECTRIC CODE).

NFQ Abbreviation of *night frequency.*

NG **1.** Abbreviation of NEGATIVE GLOW. **2.** Abbreviation of *no good*, used for marking inoperative or malfunctioning components.

nH Abbreviation of NANOHENRY.

nhp Abbreviation of NOMINAL HORSEPOWER.

Ni Symbol for NICKEL.

nibble In computer operations, a four-bit word.

NICAD Acronym for NICKEL–CADMIUM.

NiCd Chemical symbol for NICKEL–CADMIUM.

Nichrome A nickel-chromium alloy used in the form of a wire or strip for resistors and heater elements.

nickel Symbol, Ni. A metallic element. Atomic number, 28. Atomic weight, 58.69. Nickel is familiar as an alloying metal in some resistance wires and as the material used in some electron-tube elements.

nickel–cadmium Acronym, NICAD. Chemical symbol, NiCd. A mixture used in certain rechargeable electrochemical cells and batteries. In recent years, these cells and batteries have been largely replaced by nickel–metal-hydride (NiMH) and lithium-based units. These batteries must be disposed of in a special way because of the toxic cadmium they contain. Compare NICKEL–METAL-HYDRIDE.

nickel–cadmium battery **1.** A battery of nickel–cadmium cells. **2.** Loosely, a nickel–cadmium cell. See also NICKEL–CADMIUM.

nickel-cadmium cell A small, rechargeable, electrochemical cell. The anode is cadmium, the cathode is nickel hydroxide, and the electrolyte is potassium hydroxide. See also NICKEL–CADMIUM.

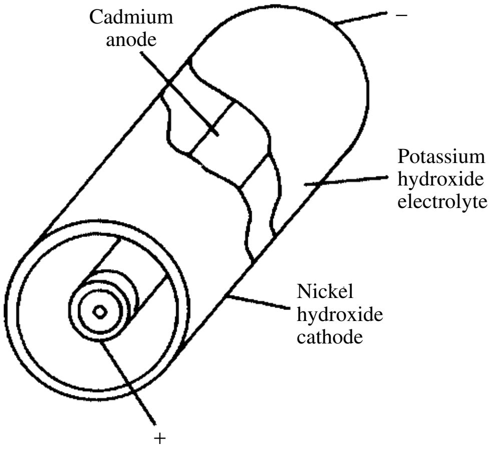

nickel-cadmium cell

nickel hydroxide Chemical symbol, NiOH. A compound used in certain rechargeable electrochemical power supplies. Examples are nickel–cadmium (NiCd or NICAD) and nickel–metal-hydride (NiMH) batteries, used in older notebook computers.

nickel–iron Chemical symbol, NiFe. A mixture used for a specialized rechargeable electrochemical cell in which the active positive plate material consists of nickel hydroxide. The active negative plate material is powdered iron oxide mixed with cadmium, and the electrolyte is potassium hydroxide.

nickel-iron battery See EDISON BATTERY.

nickel-metal–hydride Chemical symbol, NiMH. A mixture used in certain rechargeable electrochemical cells and batteries. This type of cell or battery is less likely than nickel–cadmium units to develop *battery memory* or *memory drain,* in which the ampere-hour capacity is reduced because of repeated partial discharging. In addition, the chemicals in NiMH cells and batteries are less toxic than those in NICADs. Compare NICKEL–CADMIUM.

nickel–metal-hydride battery **1.** A small, rechargeable, nickel-based electrochemical battery, similar to the nickel–cadmium (NICAD) type. **2.** Loosely, a nickel–metal-hydride cell. See NICKEL–METAL-HYDRIDE.

nickel–metal-hydride cell A small, rechargeable, nickel-based electrochemical cell, similar to the nickel–cadmium type. See also NICKEL–METAL-HYDRIDE.

nickel-oxide diode A diode fabricated from nickel-oxide semiconductor material.

nickel silver An alloy of copper, nickel, and zinc, sometimes used for making resistance wire. Also called GERMAN SILVER.

Nicol prism An optical component for producing or analyzing plane-polarized light. It consists of two prisms of Iceland spar cemented together. Light entering the device strikes the interface, where the ordinary ray is totally reflected and the extraordinary ray passes through, both rays being plane-polarized perpendicular to each other.

NIDA Abbreviation of *numerically integrating differential analyzer.*

NIF Abbreviation of NOISE IMPROVEMENT FACTOR.

NiFe Chemical symbol for NICKEL–IRON.

night effect A phenomenon sometimes observed at frequencies below approximately 500 kHz. Direction-finding signals received between sunset and the following sunrise appear to come from a transmitter that moves slowly back and forth.

night-effect errors In radio direction-finding, inaccurate or uncertain readings resulting from the NIGHT EFFECT.

night range The distance over which signals from a given transmitter are consistently received after sunset.

ni junction In a semiconductor device, the junction between an n-type layer and an intrinsic layer.

nil **1.** Colloquialism for *negligible.* **2.** British expression for *zero.* Also, *nought.* **3.** Colloquialism for *nothing.*

NiMH Chemical symbol for NICKEL–METAL-HYDRIDE.

niobium Symbol, Nb. A metallic element chemically resembling tantalum. Atomic number, 41. Atomic weight, 92.91.

NIPO Abbreviation of *negative input/positive output.*

NIR Abbreviation of NEAR INFRARED.

nit In digital-computer operations, a unit of data equal to 1.44 bits.

nitrocellulose See CELLULOSE NITRATE.

nitrogen Symbol, N. A gaseous element. Atomic number, 7. Atomic weight, 14.007. Nitrogen is the most-abundant component (about 78%) of the Earth's atmosphere.

NIU Abbreviation of NETWORK INTERFACE UNIT.

Nixonite A trade name for cellulose acetate, a plastic.

Nixonoid A trade name for cellulose nitrate, a plastic.

NJCC Abbreviation of *National Joint Computer Conference.*

n layer A semiconductor layer doped to provide current carriers that are predominantly electrons. Compare P LAYER.

n-level logic **1.** A multilevel form of logic, with *n* different possible states. **2.** In a computer, the connection of up to *n* logic gates.

NLR **1.** Abbreviation of *nonlinear resistance.* **2.** Abbreviation of NONLINEAR RESISTOR.

NLS Abbreviation of NO-LOAD SPEED.

N/m^2 Abbreviation of *newtons per meter squared* (pascals).

Nm^2/kg Newton meters squared per kilogram, the SI unit of the *gravitational constant.*

NMAA Abbreviation of *National Machine Accountants Association.*

NMOS A metal-oxide semiconductor device made on a p-type substrate whose active carriers, electrons, migrate between n-type source and drain contacts.

NMR **1.** Abbreviation of NUCLEAR MAGNETIC RESONANCE. **2.** Abbreviation of NORMAL-MODE REJECTION.

n-n junction In a semiconductor device, especially an integrated circuit, the junction between two n-type regions having somewhat different properties (sometimes designated n_1 and n_2).

NO Abbreviation of NORMALLY OPEN.

No Symbol for NOBELIUM.

No. Abbreviation of NUMBER.

no-address instruction In digital computer operations, an instruction requiring no reference to storage or memory for its execution.

nobelium Symbol, No. A radioactive element produced artificially. Atomic number, 102. Atomic weight, approximately 259 (varies with isotope).

Nobili's rings See ELECTRIC RINGS.

noble Chemically inert or inactive. For example, *noble metals* oxidize less rapidly than *base metals.*

noble gas An inert rare gas (such as argon, helium, krypton, neon, or xenon). It is used in electronic glow devices.

noble metal A comparatively nonreactive metal (such as gold, silver, or platinum).

noctovision A television transmission system using infrared rays instead of visible light to scan the object. This makes it possible to televise images in complete visual darkness (hence, the prefix *nocto-*, meaning "night").

no-charge machine fault time Unproductive computer time resulting from errors or a malfunction.

nodal point See NODE.

node **1.** The terminal point at which two or more branches of a circuit meet, or a point that is common to two circuits. **2.** In a standing-wave system, a zero point or minimum point, e.g., current node. Compare LOOP, **1. 3.** A database management system expression defining the location of information about a record, user, field, etc. **4.** A point at which a satellite crosses the equator. See also ASCENDING NODE, DESCENDING NODE.

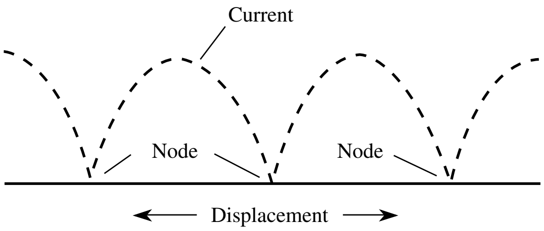

node, 2

nodules **1.** Oxide particles that protrude above the surface of magnetic tape. **2.** In a planar pattern describing radiation or pickup characteristics (as for antennas, microphones, loudspeakers), a small peak aligned in a direction other than that of the main lobe.

no-field release In the starting box for a shunt motor, the electromagnet that normally holds the arm in full-running position; it is connected in series with the field winding. When the field current is lost, the arm is released, disconnecting the armature for safety. Compare NO-VOLTAGE RELEASE.

noise **1.** A random-frequency current or voltage signal extending over a considerable frequency spectrum and having no useful purpose, unless it is intentionally generated for test purposes. **2.** Dissonant, interferential sound; unlike harmonious sound, it is disagreeable. **3.** In audio operations, unwanted hiss and/or hum. **4.** Extra bits or bytes that must be removed from digital data before it can be useful.

noise abatement The elimination or reduction of noise intensity—especially a measure in a program concerned with noise pollution in the environment.

noise analysis The measurement of the amplitude and spectral distribution of noise and the determination of its character.

noise analyzer An instrument for evaluating the nature of noise in a communications system. See, for example, NOISE METER. Noise analyzers are sometimes adapted for vibration analysis.

noise-balancing system A bridge circuit inserted between a receiver and antenna for balancing out interferential signals resulting from nearby power-line leaks or similar causes.

noise bandwidth Abbreviation, NBW. A figure obtained by dividing the area under the power-output-vs.-frequency curve of a device by the power amplitude at the noise frequency of interest.

noise behind the signal Noise caused by, but exclusive of, a signal.

noise blanker A device that cuts off one of the intermediate-frequency stages of a radio receiver during a noise pulse. The noise blanker is effective against high-amplitude impulses of short duration.

noise-canceling antenna A specialized receiving antenna system that uses two elements, one intended for receiving primarily the desired signal(s) and the other intended for receiving primarily local human-made noise. The outputs from the two elements are combined so that the noise impulses appear out of phase, but of equal amplitude. Because one element is located in a favorable position for reception of desired signals while the other element is not, the desired signals do not cancel at the receiver input. As a result, the signal-to-noise ratio is greatly improved, compared with the use of the signal antenna alone.

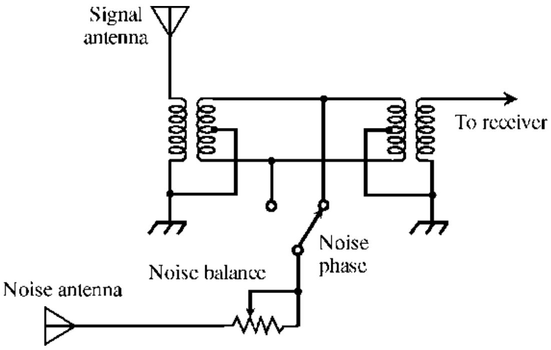

noise-canceling antenna

noise-canceling microphone A microphone that discriminates against background sounds. It is usually directional and relatively insensitive, requiring the user to talk directly into it at close range.

noise clipper A biased-diode circuit used as an automatic noise limiter. The device cuts off all signals above a predetermined amplitude on the theory that noise peaks are high-level transients in an otherwise uniform signal. Noise is reduced at the sacrifice of system reproduction fidelity.

noise criteria An expression for the level of ambient acoustic noise.

noise current Noise-generated current.

noise-current generator A noise generator that supplies a useful current. Compare NOISE-VOLTAGE GENERATOR.

noise digit A digit (usually zero) generated during normalization of a floating-point number. See NORMALIZE.

noise diode A reverse-biased semiconductor diode that produces a standard noise voltage.

noise elimination The nearly complete removal of noise effects from a system. Noise can never be eliminated altogether because the movement of electrons and atoms generates some electrical and thermal noise. However, in some digital systems, the effects of noise can be almost totally overcome. Compare NOISE SUPPRESSION.

noise equivalent power Abbreviation, NEP. The power that produces an rms signal-to-noise ratio of 1 in a detector.

noise factor For a circuit, especially a communications receiver or weak-signal amplifier, the ratio R_1/R_2, where R_1 is the signal-to-noise power ratio of an ideal circuit, and R_2 is the signal-to-noise ratio of the circuit under test. Compare NOISE FIGURE.

noise figure The NOISE FACTOR of a circuit, expressed in decibels. If N is the noise factor expressed as a ratio, then noise figure N_{dB} can be determined by $N_{dB} = 10 \log_{10} N$.

noise filter A filter designed to suppress noise that would otherwise enter an electronic circuit (e.g., a power-line noise filter).

noise floor **1.** In a receiver, the level of noise in microvolts that determines the weakest signal that can be heard or accurately received. **2.** In a spectrum analyzer, the level of noise that determines the weakest signal that will be visibly displayed.

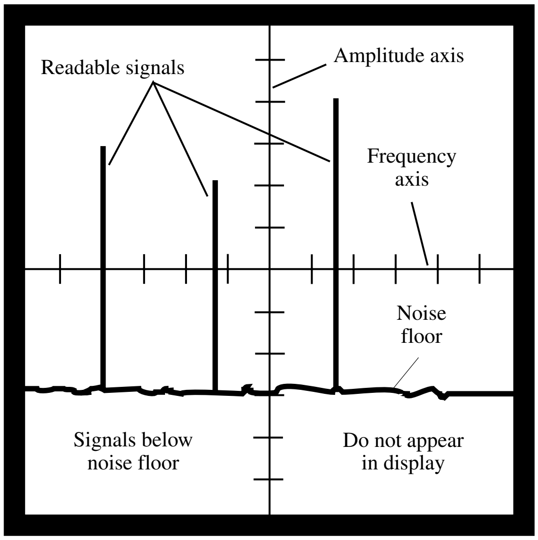

noise floor, 2.

noise generator A device for generating precise amounts of noise voltage for test purposes.

noise grade **1.** The relative level of radio-frequency background noise, over all electromagnetic frequencies, in a particular geographic location. The noise grade is generally lowest near the poles and highest near the equator. **2.** The mathematical function of relative electromagnetic noise intensity versus latitude and longitude.

noise immunity The degree to which a circuit or device is insensitive to extraneous energy—especially noise signals.

noise-improvement factor Abbreviation, NIF. For a radio receiver, the ratio SN_i/SN_o, where SN_i is the input signal-to-noise ratio and SNo is the output signal-to-noise ratio.

noise killer **1.** See AUTOMATIC NOISE LIMITER. **2.** See NOISE FILTER. **3.** See NOISE BLANKER.

noiseless alignment See VISUAL ALIGNMENT.

noise level **1.** The amplitude of ambient electrical noise generated outside an electronic system of interest. **2.** The amplitude of electrical noise generated in an electronic system of interest. **3.** The intensity of ambient acoustic noise.

noise limiter See AUTOMATIC NOISE LIMITER.

noise margin In a binary logic circuit, the difference between operating and threshold voltages.

noise-measuring set See NOISE METER.

noise meter An instrument for measuring acoustic noise level. It consists essentially of a sensitive, multirange voltmeter provided with a microphone, amplifier, and attenuators. The meter scale reads noise level directly in decibels.

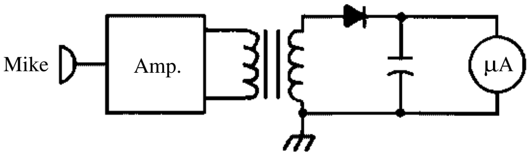

noise meter

noise power The power component of a noise signal.

noise power ratio The ratio of noise power at the output of a circuit (such as a receiver) to the noise power at the input.

noise pulse A random short-duration noise burst whose amplitude exceeds the average peak noise level.

noise quieting In a radio receiver, the reduction (in decibels) of background noise, with respect to a signal of interest.

noise ratio See NOISE POWER RATIO.

noise-reducing antenna A receiving antenna having a balanced transmission line and usually some form of noise-balancing system for reducing electrical noise picked up by the antenna.

noise reduction See NOISE SUPPRESSION.

noise residue The residual output (see NULL VOLTAGE, **1**) of a balanced bridge, caused entirely by noise.

noise silencer A noise-limiting circuit that removes noise-pulse transients with little or no effect on the signal from which the noise is removed. Compare NOISE CLIPPER.

noise source See NOISE GENERATOR.

noise spike See NOISE PULSE.

noise suppression **1.** In communications, the reduction of noise amplitude to a level that is noncompetitive with desired signals. **2.** In audio recording and reproduction, reduction of unwanted noise (e.g., hiss) to the greatest extent possible without degrading the fidelity of the desired audio. Compare NOISE ELIMINATION.

noise suppressor A device for eliminating electrical noise or reducing its amplitude. See, for example, AUTOMATIC NOISE LIMITER and NOISE BLANKER.

noise temperature At a given frequency, the temperature of a passive system that has the same noise power per unit bandwidth as that observed at the terminals of a device under test.

noise voltage The voltage component of an electrical noise signal.

noise-voltage generator A signal generator that supplies an alternating-current waveform containing random-frequency pulses of relatively uniform distribution over a given frequency spectrum. Compare NOISE-CURRENT GENERATOR.

noisy mode During normalization of a floating-point number, the generation of digits, excluding zero, as part of the fixed-point part (see NORMALIZE).

NOL **1.** Abbreviation of *National Ordnance Laboratory.* **2.** Abbreviation of *Naval Ordnance Laboratory.*

no-load current **1.** Output-electrode current (e.g., drain, plate, or collector current) when a device is not delivering output to an external load. **2.** Current flowing in the primary winding of an unloaded transformer.

no-load losses Losses in an unloaded transformer (see NO-LOAD CURRENT, **2**).

no-load speed The rotational speed of an unloaded motor.

no-load voltage The open-circuit output voltage of a power supply, amplifier, generator, or network.

nominal **1.** Named, rated, or specified. The nominal value of a speaker, for example, might be 8 ohms, even though the actual impedance value depends on the frequency of the applied signal. **2.** Approximate, and specified as a typical example only, for the purpose of identifying the operating or value range. For example, an automotive circuit might have a nominal rating of 12 volts—even though it can be operated at 10 volts to 14.6 volts.

nominal band In a facsimile signal, the waveband extending between zero and the maximum frequency of modulation.

nominal bandwidth **1.** For a filter, the difference $f_{c2} - f_{c1}$, where f_{c1} is the nominal lower cutoff frequency and f_{c2} is the nominal upper cutoff frequency. **2.** For an allocated communication channel, the total bandwidth, including upper and lower guard frequencies. **3.** The intended and specified bandwidth of a given channel, regardless of the bandwidth of the signal on that frequency at any given time.

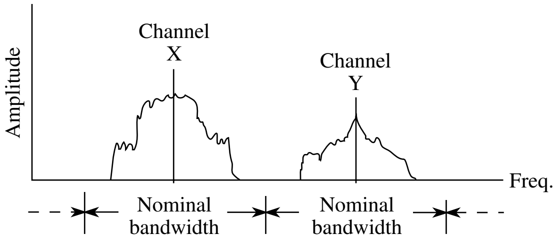

nominal bandwidth, 2

nominal capacitance The rated ("label") value of a capacitor. Also see NOMINAL VALUE.

nominal current The rated ("nameplate") value of required current or of current output. Also see NOMINAL VALUE.

nominal horsepower The rated ("nameplate") horsepower of a machine, such as a motor. Also see NOMINAL VALUE.

nominal impedance The rated impedance of a circuit or device. Also see NOMINAL VALUE.

nominal inductance The rated ("label") value of a coil's inductance. Also see NOMINAL VALUE.

nominal line pitch In a television raster, the average center-to-center separation between adjacent lines.

nominal line width **1.** For a television raster, the factor $1/n$, where n is the number of lines per unit width for the direction in which the lines progress. **2.** In facsimile, the average center-to-center separation of scanning or recording lines.

nominal power factor The rated ("nameplate") value of power factor of a device. Also see NOMINAL VALUE.

nominal power rating The rated ("nameplate") value of power output, power drain, or power dissipation. Also see NOMINAL VALUE.

nominal *Q* The rated ("nameplate") value of Q of a capacitor, inductor, transformer winding, or tank circuit. Also see NOMINAL VALUE.

nominal rating See NOMINAL, **1**, **2**.

nominal resistance The rated ("label") value of a resistor or similar device. Also see NOMINAL VALUE.

nominal speed The highest speed of a data-processing unit or system, disregarding slowdowns because of factors other than computational operations.

nominal value A named, specified, rated, or labeled value, given without reference to tolerance.

This can differ significantly from the actual value. For example, the nominal value of a capacitor might be 100 pF; but if the tolerance is ±10 percent, the actual capacitance might be any value between 90 pF and 110 pF.

nominal voltage **1.** The rated ("nameplate") value of required voltage or of voltage output. Also see NOMINAL VALUE. **2.** In a cell or battery, the MIDPOINT VOLTAGE.

nomogram See ALIGNMENT CHART.

nomograph See ALIGNMENT CHART.

nomography The geometric representation of a mathematical function or relation by means of alignment charts.

nonaccountable time The period during which a computer system is unavailable to the user (because of a power outage, for example).

nonarithmetic shift See LOGICAL SHIFT.

nonblinking meter A digital meter that does not alternate, or oscillate, between two different values when the measured parameter is between two discrete values. Instead, the display is rounded off to the nearest value, and the display remains at that value continuously.

nonblocking system In a telephone communications network, a system that ensures a circuit will be completed when necessary. That is, at no time is it impossible for a connection to be made. Under conditions of extremely heavy usage, the quality of communications might be degraded, but the connection will not be cut off.

nonbridging contact In a switch or relay, a movable contact that leaves one stationary contact before contacting another.

nonchargeable battery A primary battery (i.e., one that cannot ordinarily be recharged). An example is a battery of common zinc-carbon or alkaline cells.

noncoherent Pertaining to electromagnetic radiation in which the wave disturbances are not all precisely aligned in frequency and phase.

nonconductor See DIELECTRIC.

noncontact temperature measurement The use of infrared or optical electronic equipment to measure the temperature of bodies without touching them.

noncorrosive flux A solder flux that does not corrode the metals to which solder is applied.

noncrystalline Pertaining to materials that possess none of the characteristics of crystals. Compare CRYSTALLINE MATERIAL.

nondestructive read In digital computer and counter operation, the process of reading data without erasing it as a result. The name is also applied to the readout device.

nondestructive test Abbreviation, NDT. A test that does little or no irreversible harm to the test sample. Compare DESTRUCTIVE TEST.

nondeviated absorption Absorption that slows waves by a negligible amount; also, normal sky-wave absorption.

nondirectional antenna An antenna that displays equally intense radiation or equally sensitive reception in all directions within a specified plane. An example is a vertical dipole antenna, which is nondirectional in the horizontal plane.

nondirectional microphone A microphone that responds equally well to sound from any direction; an omnidirectional microphone.

nondissipative load A purely reactive load. In such a load, the only power consumed is that which is dissipated in the inherent resistance (losses) of the load.

nondissipative stub A stub that exhibits only slight losses; it consumes no power, except that dissipated in small, inherent losses. Also see STUB.

nonelectrical Not electrical in nature. The term is commonly used to designate the mechanical parts of electromechanical systems, such as robots and servomechanisms.

nonelectrolyte A substance that does not ionize in water solution. Compare ELECTROLYTE.

nonelectronic meter A meter that uses no electronic devices (such as transistors, liquid crystals, light-emitting diodes, or integrated circuits). Also called *conventional meter.*

nonequivalence operation See EXCLUSIVE-OR OPERATION.

nonerasable storage In digital-computer and data-processing operations, storage media that cannot be erased under ordinary circumstances. A common example is CD-ROM (compact-disk read-only memory).

nonferrous metal A metal or alloy that does not contain iron, and is not related to iron in the sense that it is not attracted to magnets.

nonflammable Pertaining to a material that is resistant to burning.

nonharmonic frequency A frequency that has no integral numerical relationship to another frequency of interest. Compare HARMONIC FREQUENCY.

nonharmonic oscillations Parasitic oscillations that do not occur at the fundamental frequency nor at any harmonic frequency of an oscillator or amplifier in which they appear.

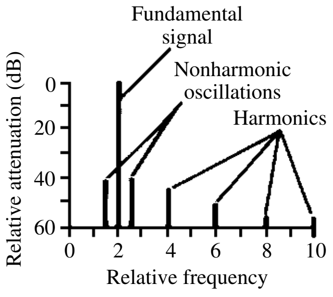

nonharmonic oscillations

nonillion The number 10^{30}, so called because when written out, the number contains nine groups of three zeros (following the first 1000).

noninductive capacitor A wound capacitor in which the edges of one of the spiral windings are connected together to minimize the inductance of the roll. Compare INDUCTIVE CAPACITOR.

noninductive resistor A wirewound resistor constructed so that the magnetic field of the coil is self-canceling. Therefore, the inductance is practically eliminated.

noninterlaced scanning In the display of a video image, the presentation of all the raster lines in a single scan. It is commonly used in cathode-ray-tube (CRT) computer monitors. This process minimizes "jerkiness" in rapidly moving images.

noninverting connection Connection to the noninverting input of a differential or operational amplifier. Also see NONINVERTING INPUT. Compare INVERTING CONNECTION.

noninverting input In a differential or operational amplifier, the input that provides an output signal in phase with the input. Compare INVERTING INPUT.

noninverting transponder In a communications satellite, a transponder in which the downlink band is "rightside-up" in frequency relative to the uplink band. That is, the highest downlink frequency corresponds to the highest uplink frequency, and the lowest downlink frequency corresponds to the lowest uplink frequency. Compare INVERTING TRANSPONDER. Also see DOWNLINK, TRANSPONDER, UPLINK.

nonionic **1.** Possessing none of the properties of ions. **2.** Electrically neutral.

nonionizing radiation Electromagnetic radiation that does not cause ionization of gases under a given set of conditions. Examples: radio signals, television signals, and visible light.

nonlinear **1.** Pertaining to components, circuits, or devices in which the instantaneous output signal amplitude is not directly proportional to the instantaneous input signal amplitude. The graph of instantaneous output versus instantaneous input is a curve, not a straight line. Example: CLASS-C AMPLIFIER. **2.** Pertaining to components, circuits, or devices in which a specified value is not directly proportional to some other specified value. Example: NONLINEAR CAPACITOR.

nonlinear bridge See VOLTAGE-SENSITIVE BRIDGE.

nonlinear capacitor A capacitor whose value varies nonlinearly with applied voltage. Also see VOLTAGE-VARIABLE CAPACITOR, **1**, **2**.

nonlinear coil See SATURABLE REACTOR.

nonlinear dielectric A material (such as processed barium-strontium titanate) whose dielectric constant varies with applied voltage.

nonlinear distortion Distortion caused by nonlinear response of an amplifier or component. This causes different parts of the signal to be amplified or transmitted by different amounts; therefore, the amplitude variations in the output signal differ from those in the input signal.

nonlinear inductor See SATURABLE REACTOR.

nonlinearity **1.** The condition of being NONLINEAR. In an amplifier, this means that the output signal is not a faithful reproduction of the input signal, and distortion occurs. **2.** A measure of the extent to which a circuit is nonlinear. Expressed as a percentage of peak-to-peak full-scale output, the maximum extent to which the output differs from a perfect reproduction of the input.

nonlinearity error An error in received signals resulting from nonlinearity in one or more of the stages in the communications circuit.

nonlinear mixing The mixing of signals as a result of the nonlinear response of a device (such as a semiconductor diode operated in its square-law region) through which they are passed simultaneously. Also see MIXER and MIXING.

nonlinear network A circuit that produces distortion in an input waveform; the output and input waves are not related by a linear function.

nonlinear quantizing A method of signal quantizing in which the intervals are not all the same size or duration.

nonlinear resistor A resistor whose value varies with applied voltage. Also see VOLTAGE-DEPENDENT RESISTOR.

nonlinear response Any response for which the corresponding plot is not a straight line; doubling the independent variable, for example, does not double the dependent variable.

nonloaded *Q* See UNLOADED *Q*.

nonmagnetic **1.** Possessing no magnetism. **2.** Incapable of being magnetized.

nonmathematical Pertaining to materials and methods that rely upon physical description and qualitative procedures instead of mathematical development, prediction, and quantitative procedures.

nonmetal An elemental material devoid of the properties exhibited by metals (e.g., luster, good ductility, electrical conductivity, heat conductivity, and malleability). Examples: carbon, phosphorus, sulfur. Compare METAL and METALLOID.

nonmetallic conduction Collectively, ionic conduction in liquids and gases, conduction in dielectrics by small leakage currents, and thermionic conduction in a vacuum.

nonmicrophonic Without microphonic properties, e.g., a nonmicrophonic integrated circuit does not produce electrical ringing when physically struck.

nonnumeric character A character that is not a numeral, i.e., a symbol or letter.

nonohmic response **1.** Nonlinear resistance or reactance. Compare OHMIC RESPONSE. **2.** See NEGATIVE RESISTANCE.

nonoscillating detector A detector devoid of positive feedback action and, therefore, unable to generate a signal on its own. Compare OSCILLATING DETECTOR.

nonplanar **1.** Existing in three spatial dimensions. **2.** Pertaining to a circuit that cannot be fabricated on a two-dimensional board without the use of jumper wires.

nonpolar **1.** Having no pole(s). **2.** Pertaining to atoms that share electrons to complete their outer shells. **3.** Not polarized nor requiring polarization. Example: a 100-pF disk ceramic capacitor is nonpolar because it can be used in circuits without consideration of voltage polarity.

nonpolar crystal A crystal in which lattice points are identical.

nonpolarized electrolytic capacitor An electrolytic capacitor that has no definite negative and positive terminals and, consequently, can be used in alternating-current circuits, as well as in direct-current circuits. See also NONPOLAR, **3**.

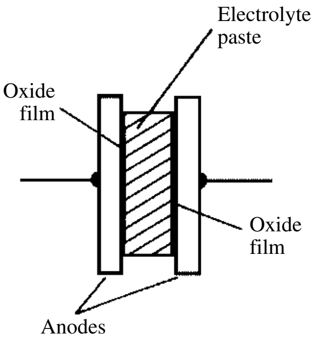

nonpolarized electrolytic capacitor

nonpolarized reactor A saturable reactor in which the lines of flux produced in the three-leg core by the coils on the two outer legs oppose each other in the center leg. When direct current (dc) is passed through the coil on the center leg to saturate the core, operation remains the same for either dc polarity. Compare POLARIZED REACTOR.

nonpolarized relay See UNPOLARIZED RELAY.

nonprint code In telegraphy, a code used to start teleprinter functions excluding printing.

nonreactive circuit A circuit containing pure resistance only.

nonrechargeable battery See NONCHARGEABLE BATTERY.

nonrecurrent Pertaining to phenomena that do not repeat periodically. Thus, a single sweep in an oscilloscope is nonrecurrent.

nonrecurrent sweep See NONREPETITIVE SWEEP.

nonregenerative detector A detector having no regenerative feedback. Such a detector is stable, but relatively insensitive. Compare REGENERATIVE DETECTOR.

nonregenerative receiver A radio receiver in which no local signal whatever is generated.

nonrepetitive phenomena See NONRECURRENT.

nonrepetitive sweep In an oscilloscope, a single horizontal sweep of the electron beam, initiated either by the operator or by the signal under observation. Also called SINGLE SWEEP. Compare RECURRENT SWEEP.

nonreset timer A timer that must be reset manually.

nonresident routine A computer routine not permanently stored in memory. Compare RESIDENT ROUTINE.

nonresonant **1.** Pertaining to a resonant circuit or device operated at some frequency other than one of its resonant frequencies. Thus, reactance is present at the operating frequency. See RESONANCE. **2.** Pertaining to a circuit or device that exhibits pure resistance (without reactance) over a wide range of frequencies.

nonresonant lines Transmission lines so dimensioned and operated that they do not resonate at the operating frequency.

nonresonant load An alternating-current load that is either purely resistive or is detuned from the fundamental and harmonic frequencies of the source from which it is operated.

nonreturn to zero In the magnetic recording of digital data, the system in which the current flowing in the write-head coil is sustained (i.e., does not return to zero) after the write pulse.

nonsalient pole A nonprojecting (often flush) pole. Compare SALIENT POLE.

nonsaturated color **1.** Visible light that consists of energy at more than one wavelength. **2.** A color that contains some white, in addition to the pure color.

nonsaturated logic A logic circuit in which transistors are prevented from saturating. This results in higher operating speed than SATURATED LOGIC using the same transistors.

nonshorting switch A multiple-throw switch that disconnects one circuit before completing another; that is, no two poles are ever connected simultaneously.

nonsinusoidal waveform A waveform whose curve cannot be represented by the equation $y = c \sin a(x + b)$, where a, b, and c are constants, y is the dependent variable (usually instantaneous amplitude or frequency), and x is the independent variable (usually time), for any real-number values of a, b, and c. Examples: BACK-TO-BACK SAWTOOTH, SAWTOOTH, and SQUARE WAVE. The COSINE WAVE is sinusoidal, being a SINE WAVE shifted in phase by 90 degrees.

nonsymmetrical wave See ASYMMETRIC WAVE.

nonsynchronous Unrelated in cyclic quality to other such qualities in the system.

nonsynchronous network A communications network in which the clocks are not all synchronized.

nonsynchronous vibrator A power-supply vibrator that is essentially a single-pole, double-throw switch providing no mechanical rectification. A separate rectifier must be used. Compare

VIBRATOR-TYPE RECTIFIER. Also see VIBRATOR-TYPE POWER SUPPLY.

nontechnical Pertaining to circuits, devices, systems, or phenomena described in lay terms, using concise graphics and little or no mathematics. An example is the simplified explanation of the operation of a spread-spectrum radio transmitter.

nontonal components See NOISE, **3**.

nontrigger voltage For a thyristor, the maximum gate-to-cathode voltage that can be applied without triggering the device. The amplitude of interferential signals, including noise, must be below this level to prevent accidental triggering.

nonuniform field An electric or magnetic field whose intensity is not the same at all points.

nonvolatile memory Memory whose data is retained even when power is removed for extended periods. This type of memory requires no backup power source. The main advantage of this type of memory is the fact that the data is not lost in case of a power interruption. Memory should not be confused with storage. Magnetic, magneto-optical, or optical disks (including hard drives) are storage media, not memory. Memory data can be stored and retrieved much faster than storage data, because memory uses no mechanical parts. Also see RANDOM-ACCESS MEMORY, READ-ONLY MEMORY. Compare VOLATILE MEMORY.

nonvolatile storage A computer storage medium in which the data does not require a source of power to be retained. Examples: MAGNETIC DISK, MAGNETIC TAPE, and COMPACT-DISK READ-ONLY MEMORY.

no-op instruction An instruction that commands a computer to perform no operation, other than to proceed to the following instruction.

NOR circuit Also called NOT-OR CIRCUIT. In computer and control operations, a circuit that delivers a zero output signal, except when two or more input signals are zero. The NOR circuit function is the inverse of that of the OR circuit.

NOR gate A gate that performs the functions of a NOR circuit.

norm The average or ambient condition.

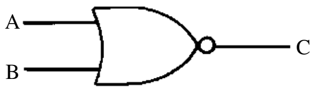

A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

NOR circuit

normal **1.** Pertaining to the most commonly observed set of conditions or parameters. **2.** Standard. **3.** Perpendicular; oriented at right angles. **4.** Pertaining to a NORMAL DISTRIBUTION. **5.** Pertaining to a NORMAL SOLUTION. **6.** Pertaining to an atom at its lowest energy state. See NORMAL STATE OF ATOM.

normal curve See BELL-SHAPED CURVE.

normal distribution In a statistical evaluation, a probability distribution represented by the so-called bell-shaped curve. The maximum probability occurs at the 50-percent value.

normal-distribution curve See BELL-SHAPED CURVE.

normal electrode A standard electrode used in electrode-potential measurements.

normal fault An unintended path between the hot terminal of a load and ground.

normal fault plus grounded neutral fault A combination of NORMAL FAULT and GROUNDED NEUTRAL FAULT.

normal glow discharge In a glow-discharge tube, the discharge region between the Townsend discharge and the abnormal glow in which current increases sharply, but a constant voltage drop is maintained across the tube.

normal impedance A transducer's input impedance when the load impedance is zero.

normal induction curve A saturation curve for a magnetic material. Also see BOX-SHAPED LOOP and SATURABLE REACTOR.

normalize In computer programming, to use floating-point numbers to modify the fixed-point part of a number so that it is within a desired range.

normalized admittance The quantity $1/Z_n$, where Z_n is NORMALIZED IMPEDANCE.

normalized frequency The unitless number represented by the ratio f/f_r, where f_r is a reference frequency and f is a frequency of interest. Response plots are sometimes conveniently drawn on the basis of normalized frequency, the reference (or resonant) frequency being indicated as **1**, twice the reference frequency as **2**, etc.

normalized impedance A value of impedance divided by the characteristic impedance of a waveguide.

normally closed Abbreviation, NC. Pertaining to a switch or relay whose contacts are closed when the device is at rest. Compare NORMALLY OPEN.

normally open Abbreviation, NO. Pertaining to a switch or relay whose contacts are open when the device is at rest. Compare NORMALLY CLOSED.

normal mode Pertaining to a device or system operated in its usual or most common manner.

normal mode A state of acoustic resonance in an enclosure, such as a speaker cabinet or a room.

normal-mode rejection Abbreviation, NMR. In a digital direct-current voltmeter, the level of noise on the applied voltage that will be rejected by the instrument. Compare COMMON-MODE REJECTION.

normal position In a switch or relay, the state of the contacts when the device is at rest.

normal solution A solution, such as an electrolyte, in which the amount of dissolved material is chemically equivalent to 1 gram-atomic weight of hydrogen per liter of the solution. Compare MOLAR SOLUTION.

normal state of atom The condition in which an atom is at its lowest energy level. For the hydrogen atom, for example, the state in which the electron is in the lowest-energy orbit.

normal-through A feature in an audio PATCH BAY or PATCH PANEL that connects two sockets by default. The top socket and the one immediately below it are connected, even when a patch cord is not plugged into either of them.

northern lights See AURORA.

north magnetic pole The north pole of the equivalent bar magnet constituted by the EARTH'S MAGNETIC FIELD. The north magnetic pole lies close to the geographic north pole. Compare SOUTH MAGNETIC POLE.

north pole **1.** See NORTH MAGNETIC POLE. **2.** The earth's geographic north pole. **3.** See NORTH-SEEKING POLE.

north-seeking pole Symbol, N. The so-called north pole of a magnet. When the magnet is suspended horizontally, this pole points in the direction of the earth's north magnetic pole. Compare SOUTH-SEEKING POLE.

Norton's equivalent An equivalent circuit based on NORTON'S THEOREM, replacing a Thevenin equivalent for a current-actuated device, such as a bipolar transistor. Also see THEVENIN'S THEOREM.

Norton's theorem With reference to a particular set of terminals, any network containing any number of generators and any number of constant impedances can be simplified to one constant-current generator and one impedance. The equivalent circuit will deliver to a given load the same current that would flow if the output terminals of the original circuit where short-circuited. Compare COMPENSATION THEOREM, MAXIMUM POWER TRANSFER THEOREM, RECIPROCITY THEOREM, SUPERPOSITION THEOREM, and THEVENIN'S THEOREM.

NOT In binary logic, an operation that changes high to low and vice versa. Also see NAND CIRCUIT, NOR CIRCUIT, NOR GATE, NOT CIRCUIT, and NOT-OR CIRCUIT.

NOT-AND circuit See NAND CIRCUIT.

notation The way that numbers, quantities, or formulas are represented (e.g., *binary notation*, *Polish notation*, and *scientific notation*).

notch A dip in frequency response, typical of a band-suppression (band-elimination) filter or other frequency-rejection circuit. Compare PEAK, **3**.

notch amplifier An amplifier containing a notch filter or other arrangement that permits it to reject one frequency or a given band of frequencies while passing all higher and lower frequencies.

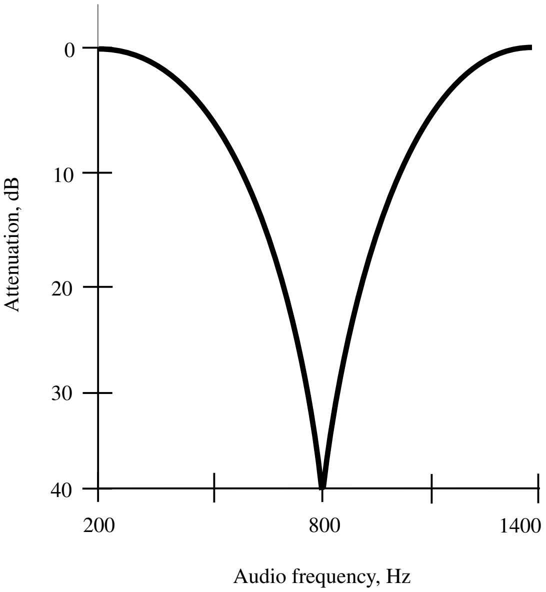

notch

notch antenna An antenna with a slot in the radiating surface, for the purpose of obtaining a directional response.

notcher See NOTCH FILTER.

notcher-peaker A circuit or device that can be set to perform either as a NOTCH FILTER or PEAK FILTER.

notch filter A circuit that exhibits high attenuation at and near a single frequency and little or no attenuation at all other frequencies. This type of device is used in some radio communications receivers, and can reduce interference caused by strong, unmodulated carriers within the passband. The notch frequency is adjustable, so that the deep null can be tuned to any frequency within the receiver passband. A properly designed circuit can produce attenuation in excess of 40 dB in the center of the notch. Some sophisticated types, especially audio designs, can provide more than 60 dB of attenuation at the notch frequency. Audio notch filters employ operational amplifiers with resistance-capacitance (RC) circuits. In some audio notch filters, the notch width (sharpness) and frequency are both adjustable. Compare BAND-REJECTION FILTER.

notch gate In radar, a gate that determines the minimum and maximum range.

notch sweep An oscilloscope sweep that expands only a small portion (notch) of the pattern on the screen, leaving the portions on either side of the notch untouched. Thus, the first dozen or so cycles might appear at the normal sweep speed, the next two cycles expanded, and the remaining two or three at normal sweep speed.

NOT circuit A logic circuit that provides an output pulse when there is no input pulse, and vice versa. Also called COMPLEMENTER, NEGATOR, and INVERTER.

note See BEAT NOTE.

notebook computer A portable personal computer, also called a *laptop computer.* It is about the size of a typical three-ring notebook, and generally contains a DISKETTE DRIVE, a HARD DISK, a MODEM, and attachments for peripherals, such as printers. It uses rechargeable batteries and can be operated for approximately two to six hours between battery charges.

NOT gate A digital circuit that inverts a logical condition—either from high (logic 1) to low (logic 0) or vice versa. Also called an *inverter.*

NOT-OR circuit A logical OR CIRCUIT combined with a NOT CIRCUIT.

novelty calculator See SPECIAL-PURPOSE CALCULATOR.

November Phonetic alphabet code word for the letter N.

novice **1.** A beginner class of amateur radio license. **2.** Any beginner or inexperienced practitioner.

no-voltage release In the starting box for a shunt motor, the electromagnet that normally holds the arm in full-running position. It is connected directly across the power line to disconnect the motor in the event of power failure. When the arm is released, it falls to its off position, thereby preventing burnout that would result if the motor were left connected to the line in the full-running position when power resumed. Compare NO-FIELD RELEASE.

noys scale A scale of apparent acoustic noise, based on a linear function instead of the more common logarithmic function.

Np **1.** Symbol for NEPTUNIUM. **2.** Abbreviation of NEPER.

N_p Symbol for number of primary turns in a transformer.

n-phase system A polyphase system having *n* phases.

npin transistor A junction transistor having an intrinsic layer between a p-type base and an n-type collector. The emitter is a second n-type layer on the other side of the base.

N plant See NUCLEAR POWER PLANT.

n-plus-one address instruction A computer program instruction containing two addresses, one of which specifies the location of an upcoming instruction to be executed.

NPM Symbol for *counts per minute.*

npnp device A semiconductor switching device having three junctions. Examples: FOUR-LAYER DIODE, and SILICON-CONTROLLED RECTIFIER. Also called *pnpn device.*

npn transistor A bipolar transistor in which the emitter and collector layers are n-type semiconductor material, and the base layer is p-type semiconductor material. Compare PNP TRANSISTOR.

NP0 Abbreviation of NEGATIVE POSITIVE ZERO.

NP0 capacitor A fixed capacitor exhibiting temperature-compensating ability over a wide temperature range, in which the coefficient has negative, positive, and zero values.

NPS Symbol for *counts per second.*

N radiation X rays emitted as a result of an electron becoming an N electron.

NRD Abbreviation of NEGATIVE-RESISTANCE DIODE.

N region See N LAYER.

NRZ Abbreviation of NONRETURN TO ZERO.

N_s Symbol for number of secondary turns in a transformer.

ns Abbreviation of NANOSECOND.

N scan See N DISPLAY.

N scope Colloquialism for a radar set using an N DISPLAY.

nsec Alternate abbreviation of NANOSECOND.

Ns/m² Newton-seconds per meter squared, the unit of dynamic viscosity.

n-space A coordinate system in *n* variables. It is generally of mathematical interest. The coordinates are written $(x_1, x_2, x_3, \ldots, x_n)$ and are called *ordered n-tuples.*

NSPE Abbreviation of *National Society of Professional Engineers.*

NTC Abbreviation of NEGATIVE TEMPERATURE COEFFICIENT.

nth harmonic An unspecified harmonic, having a frequency of *n* times the fundamental frequency, where *n* is some positive integer. Also see HARMONIC and HARMONIC FREQUENCY.

nth term An unspecified term in a mathematical sequence or series.

NTP Abbreviation of NORMAL TEMPERATURE AND PRESSURE.

NTSC Abbreviation of *National Television Standards Committee.*

NTSC color signal The color-television signal specified by the National Television Systems Committee. In the signal, the phase of a 3.58-MHz signal varies with the instantaneous hue of the transmitted color, and the amplitude varies with the instantaneous saturation of the color.

NTSC triangle On a chromaticity diagram, a triangle whose sides encompass the range of colors obtainable from the additive primaries.

NTSC-type generator A special radio-frequency signal generator for color-television tests. It provides separate, individually selected color bars that are fully saturated. The signals are strictly in accordance with NTSC standards.

n-type conduction In a semiconductor, current flow consisting of electron movement. Compare P-TYPE CONDUCTION.

n-type material Semiconductor material that has been doped with a donor-type impurity and, consequently, conducts a current via electrons. Germanium, for example, when doped with arsenic, becomes n-type. Compare P-TYPE MATERIAL.

n-type semiconductor See N-TYPE MATERIAL.

nuclear battery See ATOMIC BATTERY.

nuclear bombardment In nucleonics, the bombarding of the nucleus of an atom with subatomic particles, usually neutrons.

nuclear charge The net positive charge of the nucleus of an atom.

nuclear clock A chronometer based on the rate of disintegration of a radioactive material.

nuclear energy Energy resulting from the splitting of the nucleus of an atom or from the fusion of nuclei. Also see ATOMIC ENERGY, ATOMIC POWER, NUCLEAR FISSION, NUCLEAR FUSION, NUCLEAR REACTOR, and NUCLEUS.

nuclear fission A nuclear reaction resulting from the bombardment of nuclei in the atoms of certain radioactive materials. The bombardment with neutrons creates two new nuclei (by splitting) and several new neutrons that split several other nuclei, producing still more nuclei and neutrons, etc. The result is a *chain reaction* that can lead to a violent explosion if not checked. Compare NUCLEAR FUSION.

nuclear force A strong attraction that holds together pairs of nucleons in an atomic nucleus. This prevents an electric charge of protons from driving the nucleus apart. Nuclear force acts only over very minute distances. At greater distances, electrostatic repulsion is stronger.

nuclear fusion A nuclear reaction resulting from the violent collision of the nuclei of the atoms of a hydrogen isotope (such as deuterium) at extremely high temperature. The process produces more energy than does NUCLEAR FISSION, and leaves no hazardous radioactive waste.

nuclear magnetic resonance An atomic phenomenon in which a particle, such as a proton, in a steady magnetic field "flips over" when an alternating magnetic field is applied perpendicular to the steady field.

nuclear magnetic resonance imaging Abbreviation, NMRI. The use of NUCLEAR MAGNETIC RESONANCE effects to produce a picture of internal body organs. Using computers, three-dimensional renditions can be generated. It is useful in medicine as a diagnostic aid.

nuclear medicine A branch of medicine involving the use of radioactive isotopes in diagnosing and treating disease. A radioactive isotope is put inside the body and it tends to accumulate in certain areas. Abnormal concentration of radioisotopes might indicate abnormal body activity in a certain area.

nuclear pile See NUCLEAR REACTOR.

nuclear power plant A power-generating plant using a NUCLEAR REACTOR.

nuclear reaction **1.** A reaction in which a heavy atomic nucleus is split into two or more lighter nuclei, with an accompanying release of radiant energy. Also called NUCLEAR FISSION. **2.** A reaction in which two or more light nuclei combine to form a heavier nucleus, accompanied by the release of radiant energy. Also called NUCLEAR FUSION.

nuclear reactor **1.** A device in which nuclear fission can be initiated and controlled. At the center of the reactor is a core of nuclear fuel, such as a fissionable isotope of uranium. The core is surrounded by a graphite moderator jacket that is, in turn, surrounded by a coolant jacket; the whole is surrounded by a thick concrete shield. Neutron-absorbing rods are inserted through various walls to different depths in the fuel to control the reaction. Also called *atomic pile.* **2.** A controlled nuclear fusion device, not yet perfected, but under development. It would provide all the benefits of an atomic pile (fission reactor), but would be more efficient and would not produce hazardous radioactive waste.

nuclear recoil An observable vibration of an atomic nucleus when it disintegrates.

nuclear resonance The condition wherein a nucleus absorbs a gamma ray emitted by an identical nucleus.

nuclear service robot A remotely controlled (teleoperated) robot used for general work in environments where the level of radioactivity is too high for humans (e.g., the maintenance of a nuclear reactor). It could also be used, if necessary, for such tasks as disarming nuclear warheads and cleaning up after a nuclear accident.

nucleon A proton or neutron in the nucleus of an atom.

nucleonics The branch of physics concerned with nucleons and nuclear phenomena. The name is an acronym for *nuclear electronics.*

nucleon number See MASS NUMBER.

nucleus The center or core of an atom. Contains neutrons, protons, and other particles. The net electric charge of the nucleus is positive, and is equal to the sum of the negative charges of the orbital electrons of the atom.

null **1.** The condition of zero output current or voltage resulting from adjusting or balancing a circuit, such as a bridge. **2.** A local minimum in an interference pattern or directivity pattern.

null balance In potentiometric-measuring circuits for comparing one voltage to another, the balance condition in which no current flows through the galvanometer.

null current In potentiometric-measuring circuits for comparing one voltage to another, the galvanometer current remaining at null when the null point is not fully zero.

null detection Direction finding by means of an antenna with a bidirectional or unidirectional null response.

null detector See BRIDGE DETECTOR.

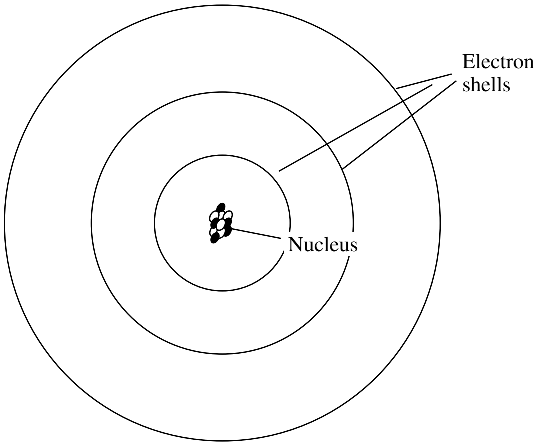

nucleus

null frequency The frequency at which a frequency-sensitive circuit, such as a Wien bridge or twin-tee network, can be balanced.

null meter See BRIDGE DETECTOR.

null method See ZERO METHOD.

null point In a balanced circuit, such as a bridge or potentiometer, the point of zero output voltage (or current) or minimum output voltage (or current).

null potentiometer **1.** The variable resistor that constitutes one arm of a four-arm bridge and is used to balance the bridge. **2.** A potentiometric circuit using the null method to compare one voltage with another. Also see POTENTIOMETER, **2**.

null setting **1.** The setting of a bridge circuit or other null device that balances the circuit. **2.** The electrical zero setting of an electronic voltmeter.

null voltage **1.** In a conventional bridge, the output voltage remaining when the bridge is set for its best null. **2.** For a voltage-sensitive bridge, the input voltage that will produce zero output voltage.

number The mathematical representation of a quantity. It is generally used in electronics to denote coefficients, magnitudes, component values, frequencies, etc.

number cruncher A computer with great computational power, but one not necessarily able to process large amounts of data (such as payroll information).

number system A systematic sequence of numbers based on a radix and a logical arrangement. See, for example, BINARY NUMBER SYSTEM and DECIMAL NUMBER SYSTEM.

numeral A member of a digit set in a number system.

numerical analysis A mathematical approach to solving problems numerically, including finding the limits of error in the results.

numerical code A code having a character set restricted to digits.

numerical control A method of programming computer-controlled mechanical devices, used in some early robots. An automated system in which number sequences fed to a digital computer cause it to control machines or processes in a manufacturing operation.

nV Abbreviation of NANOVOLT.

nW Abbreviation of NANOWATT.

nybble A piece of digital information that is larger than a bit and smaller than a byte. Compare GULP.

Nylon The trade name for a synthetic fiber-forming polyamide, useful for electrical insulation.

Nyquist criterion of stability With reference to a NYQUIST DIAGRAM for a feedback amplifier, the amplifier is stable if the polar plot of loop amplification for all frequencies from zero to infinity is a closed curve that neither passes through nor encloses the point $1 + j0$.

Nyquist diagram A graph of the performance of a reactive feedback system (such as a degenerative amplifier) that depicts the variation of amplitude and phase of the feedback factor with frequency. The plot is polar and accounts for the real and imaginary components.

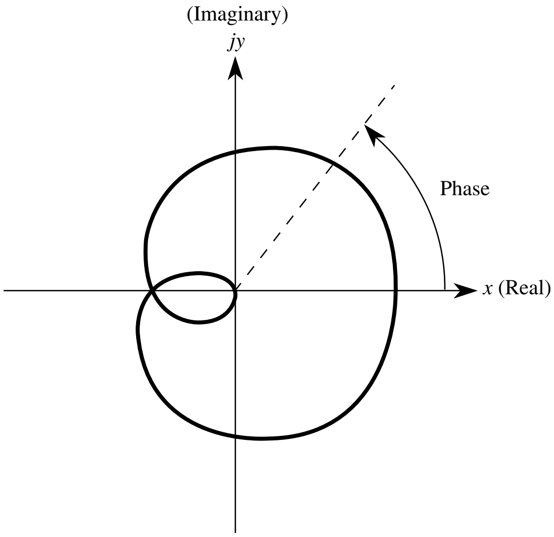

Nyquist diagram

Nyquist noise rule The power dissipated in a resistor because of thermal noise at a given frequency. The derivative of frequency, with respect to power, is equal to the absolute temperature times the Boltzmann constant.

N zone See N LAYER.

O

O **1.** Symbol for OXYGEN. **2.** Abbreviation of OUTPUT.

o **1.** Symbol for OUTPUT. (Also, OUT; both are used as subscripts.) **2.** Symbol for ORIGIN.

O_2 Symbol for OXYGEN.

O_3 Symbol for OZONE.

OAT Abbreviation of OPERATING AMBIENT TEMPERATURE.

object code **1.** In a computer system, the machine-language output of the compiler, directed to the computer. **2.** In a computer, the high-level output of the compiler, directed to the operator. **3.** The assembly language, directed to the compiler for translation between machine language and high-level language and vice versa.

object language The computer language that a compiler derives from a high-level (source) language (such as C++, LISP, etc.); it is usually machine language, but it can be an intermediate code that requires further conversion.

object-oriented graphics Also called *vector graphics.* In computer graphics, the use of equations to represent curves in a coordinate plane to define shapes, rather than defining the shapes pixel-by-pixel.

object-oriented language A computer program that uses on-screen objects, called icons, to represent commands. A movable device, such as a mouse or trackball, is used to move an arrow or other pointer to the icon; then a button is pressed to carry out the function indicated by the icon.

object-oriented programming Abbreviation, OOP. A computer programming language that builds sophisticated programs from basic programs called modules.

object program A machine or high-level language version of a user's computer program, as produced by a compiler.

object recognition In robotic systems, any method used to identify specific objects, according to characteristics, such as shape, texture, weight, etc. Common schemes use bar-code labels and machine vision. More-complex methods make use of pattern-recognition programs.

oblique-incidence transmission Transmission of radio signals via ionospheric reflection.

oblique mode In acoustics, a resonance within a room that involves all four walls, the floor, and the ceiling.

oboe A system of radar navigation in which a pair of ground stations measures the distance to an airborne transponder beacon and then transmits the information to the aircraft.

obsolescence-free Pertaining to a design or process that is not likely to become obsolete in the near future. Compare OBSOLESCENCE-PRONE.

obsolescence-prone Pertaining to a design or process subject to being soon outdated. Compare OBSOLESCENCE-FREE.

OBWO Abbreviation of O-TYPE BACKWARD-WAVE OSCILLATOR.

o/c Abbreviation of OPEN CIRCUIT.

occluded gas Gas that has been absorbed or adsorbed by solid material, such as glass or metals, and that must be eliminated during the evacuation of an electronic device, such as a vacuum tube. Also see OUTGASSING.

occultation **1.** The passage of the moon or other planetary body in front of a more distant celestial object, resulting in the cutting off of electromag-

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

netic radiation from that object. **2.** The eclipsing of one object by another.

occupied band A frequency band used by at least one transmitting station regularly.

occupied bandwidth For a given emission, the continuous band of frequencies ($f_2 - f_1$) for which the mean (average) radiated power above f_2 and below f_1 is 0.5 percent of the total mean radiated power.

occupied orbit In an atom, an orbit in which an electron is present.

Ocean Phonetic alphabet code word for the letter O.

OCR Abbreviation of OPTICAL CHARACTER RECOGNITION.

oct Abbreviation of OCTAL.

octal Abbreviation, oct. Based on the number eight. See, for example, OCTAL NUMBER SYSTEM.

octal digit One of the figures in the group 0 through 7 used in the OCTAL NUMBER SYSTEM.

octal notation See OCTAL NUMBER SYSTEM.

octal number system The base-eight system of number notation. It uses digits 0 through 7. Compare BINARY NUMBER SYSTEM and DECIMAL NUMBER SYSTEM. The octal system is often used as shorthand for otherwise-cumbersome binary numbers. The binary number is separated into groups of three digits from right to left; each such group is then converted into its decimal equivalent, with the result being the octal form of the binary number [e.g., binary 111 001 011 = octal 713 (111 = decimal 7; 001 = decimal 1; and 011 = decimal 3)].

Decimal	*Octal*
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	10
9	11
10	12
11	13
12	14
13	15
14	16
15	17
16	20

octal number system

octal-to-decimal conversion Conversion of numbers in the octal number system to numbers in the more-familiar decimal number system. This is done by expressing the octal number serially in powers of eight. Thus, octal 12 = $(1 \times 8^1) + (2 \times 8^0) = 8 + 2 =$ decimal 10.

octant One-eighth of a circle; therefore, 45 degrees, or 0.5 quadrant.

octave **1.** The range of frequencies between a given frequency (f) and twice that frequency ($2f$). **2.** The range of frequencies between a given frequency (f) and half that frequency ($f/2$).

octave band A band of frequencies one octave wide (see OCTAVE).

octave-band noise analyzer A noise analyzer having bandpass-filter channels whose center frequencies are one octave apart.

octave-band pressure level Sound pressure level within an OCTAVE BAND.

octave pressure level See OCTAVE-BAND PRESSURE LEVEL.

OCTL Abbreviation of OPEN-CIRCUITED TRANSMISSION LINE.

octonary signal An eight-level signaling code.

octonary system See OCTAL NUMBER SYSTEM.

OD Abbreviation of OUTSIDE DIAMETER.

odd-even check A method of checking the integrity of data transferred in a computer system, in which each word carries an extra digit to show whether the sum of ones in the word is odd or even.

odd harmonic In a complex waveform, a harmonic that is an odd-numbered multiple of the fundamental frequency (e.g., third harmonic, fifth harmonic, etc.). Compare EVEN HARMONIC.

odd-harmonic intensification In a complex waveform, emphasis of the amplitude of odd harmonics, with respect to that of even harmonics, a property of some multivibrators and nonsinusoidal waves.

odd line In a conventional television picture, one of the 262.5 odd-numbered horizontal lines scanned by the electron beam in developing the odd-line field. Compare EVEN LINE.

odd-line field On a conventional television screen, the complete field obtained when the electron beam has traced all the odd-numbered lines. Compare EVEN-LINE FIELD.

odd-line interlace See ODD-LINE FIELD.

odd parity check A computer check for an odd number of ones or zeros in digital data.

Odex Trade name for a series of autonomous robots developed by Odetics, Inc. They use legs for locomotion. Noted for their ability to maneuver in places that most robots cannot reach.

odograph An electromechanical or electronic plotter that traces the path of a vehicle on a map, or on the image of a map as portrayed on a display screen.

odometer An electromechanical device that indicates the speed of, and distance covered by, a moving vehicle or robot. Some such devices give a constant position indication, via mathematical integration of the measured velocity (speed and direction), relative to time. It can function in one, two, or three dimensions.

odometry A method of speed, velocity, and/or position sensing. It is commonly used in mobile vehicles and mobile robots. The most-sophisticated systems give a constant indication of position on the basis of a starting point (origin) and mathematical integration of the velocity (speed and direction), with respect to time. It can function in one, two, or three dimensions.

Oe Symbol for OERSTED.

oersted Symbol, Oe. The cgs unit of magnetic field intensity; 1 Oe = 79.58 A/m. Formerly the unit of reluctance.

off-air alarm A device that gives a visible or audible indication when the carrier of a transmitter is lost. In its most rudimentary form, the device consists of a radio-frequency relay that actuates a bell, buzzer, horn, or lamp.

off-center display A radar display whose center point does not correspond to the location of the antenna.

off-center-fed antenna An antenna in which a feeder is attached to one side of the center point of the radiator. See, for example, WINDOM ANTENNA.

off-center feed The connection of a feed line to an antenna radiator at some point other than the physical center of the radiator.

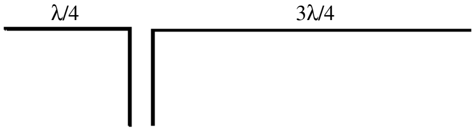

off-center feed

off delay The interval during which a circuit remains on after the control signal has been switched off. Compare ON DELAY.

off-ground **1.** An above- or below-ground operating voltage. **2.** The ground return for a voltage, as defined in **1**, not common with the system ground.

offhook In a telephone system, the condition in which the receiver is removed from its receptacle, or in which the line is otherwise engaged (e.g., a fax machine or modem is in use).

off isolation A measure of the extent to which an open switch isolates the output from the input, specified as the output voltage divided by the input voltage. For percentage, multiply this quantity by 100. The ideal value is zero (or zero percent).

off-limit In a stepping relay, a condition in which the armature has gone past the limit of its travel.

offline **1.** In a computer installation, equipment that is not controlled by the central processor. **2.** In computers and data processing, operations that are not under the direct real-time control of a central processor. **3.** A computer memory facility or device not connected to a central processor. **4.** In personal computing, the use of a computer by itself (i.e., not connected in a network). Compare ONLINE.

offline editing In video production, a process in which copies are made of the original recording; these copies are edited in a trial-and-error scheme to develop the final presentation. Compare ONLINE EDITING.

offloading The use of computers and/or robots to perform repetitive, mundane tasks, leaving people free to do more interesting things. Example: a personal robot that mows the lawn.

off-on operation See ON-OFF OPERATION.

off period **1.** The interval during which an on-off circuit or device is off. **2.** The time during which an equipment is shut down.

offset An imbalance between the halves of a normally symmetrical circuit, such as that of a differential amplifier. Also see OFFSET CURRENT and OFFSET VOLTAGE.

offset adjustment range The change in offset voltage, in millivolts or in microvolts, that can be affected by means of the external offset adjustment circuit(s).

offset current For an operational amplifier, the input current when the offset voltage is zero.

offset voltage For an operational amplifier, the particular value of direct-current bias voltage required at the input to produce zero output voltage.

off state **1.** The condition of an on-off circuit or device, such as a flip-flop, that is off. Compare ON STATE, **1**. **2.** The condition in which a circuit or device is shut down. Compare ON STATE, **2**.

off-state voltage The voltage drop across a semiconductor device, such as a diode, rectifier, or thyristor, when the device is in its normal off (nonconducting) state. Compare ON-STATE VOLTAGE.

off-target jamming In radio or radar jamming, the use of a remote jamming transmitter that will not betray the location of the base station.

off time The period during which no useful work is being performed, as of equipment when it is not functioning because of a circuit breakdown.

OGL Abbreviation of OUTGOING LINE.

OGO Abbreviation of *orbiting geophysical observatory.*

ohm Symbol, Ω. The basic unit of resistance, reactance, or impedance. A resistance of 1 ohm passes a current of 1 ampere in response to an applied emf of 1 volt.

ohmage Electrical resistance or resistivity expressed in ohms.

ohm-centimeter A unit of volume resistivity (see RESISTIVITY): the resistance of a centimeter cube of the material under measurement. Also see MICROHM-CENTIMETER.

ohmic component A resistive or reactive circuit or device exhibiting an OHMIC RESPONSE.

ohmic contact A usually very-low-resistance connection between two materials that provides bilateral linear conduction. It exhibits none of the properties of a rectifying junction or a nonlinear resistance.

ohmic heating **1.** Heating caused by current passing through a resistive material (i.e., I^{2R} losses in the material). **2.** In an electric field, heat generated by charged particles when they collide with other particles.

ohmic loss Loss resulting from the direct-current resistance in a circuit or transmission line.

ohmic region The portion of the response curve of a negative-resistance device that exhibits positive (ohmic) resistance. The *E-I* curve of a tunnel diode, for example, has two such positive-slope regions with a negative-slope (negative-resistance) region between them.

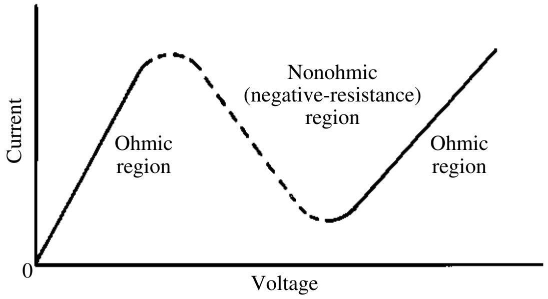

ohmic region

ohmic resistance A resistance exhibiting OHMIC RESPONSE.

ohmic response Response that follows OHM'S LAW: $I = E/R$. In strictly ohmic devices, neither resistance nor reactance changes with current or voltage. Compare NONOHMIC RESPONSE.

ohmic value Electrical resistance expressed in ohms, or in multiples or fractions of ohms (kilohms, megohms, milliohms, etc.)

ohmmeter An instrument for the direct measurement of electrical resistance. It usually consists of a milliammeter or microammeter, a battery, and several switchable resistors having very close tolerances. The scale is calibrated in ohms; the switch selects multiplier factors (e.g., ×100, ×10,000, ×1,000,000). The scale is usually reversed (i.e., 0 is at the extreme right and "infinity" is at the extreme left).

ohmmeter zero **1.** The condition of proper adjustment of an ohmmeter, indicating zero resistance for a direct short circuit. **2.** The potentiometer, or other control, used for adjusting an ohmmeter to obtain a reading of zero with a short circuit.

ohm-mile A rating meaning 1 mile of wire having a resistance of 1 ohm.

ohms adjust The rheostat or potentiometer used to set the pointer of an ohmmeter before it is used to take resistance readings.

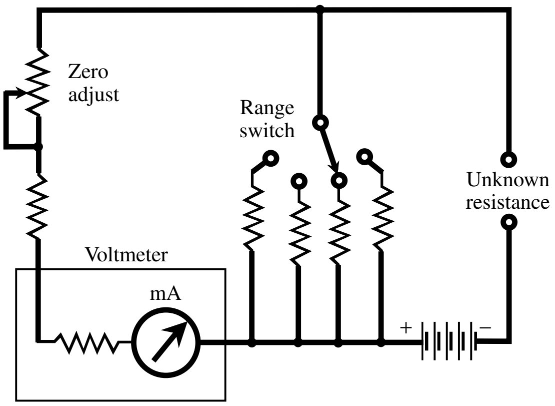

ohmmeter

Ohm's law The statement of the relationship among current, voltage, and resistance. In a direct-current circuit, current varies directly with voltage and inversely with resistance: $I = E/R$, where I is the current in amperes, E is the voltage in volts, and R is the resistance in ohms. For alternating current, Ohm's law states that $I = E/X = E/Z$, where X is reactance and Z is impedance.

ohms per square The resistance (in ohms) between two parallel edges of a square of thin-film resistance material.

ohms-per-volt A specification that indicates the sensitivity and impedance of a voltmeter. In general, the higher the rating, the better. When measuring voltage in high-impedance circuits, the rating should be as high as possible. Field-effect-transistor (FET) voltmeters and vacuum-tube voltmeters have the highest ratings.

oil-burner control An electronic system for starting and stopping the operation of an oil burner to prevent puffback and to interrupt the supply when the flame becomes erratic.

oil calorimeter A calorimeter used to measure power in terms of the rise in temperature of oil heated by the electrical energy of interest.

oil capacitor A capacitor impregnated or filled with oil, such as high-grade castor or mineral oil. Also see OIL DIELECTRIC.

oil circuit breaker A circuit breaker filled with a high-grade insulating oil for cooling and arc elimination.

oil-cooled transformer A heavy-duty transformer, through which oil is circulated for heat removal and arc prevention.

oil dielectric A highly refined oil used as an electrical insulating material (e.g., between the plates of a capacitor). Familiar examples are *castor oil*, *mineral oil*, and the synthetic oil *chlorinated diphenyl.*

oil diffusion pump See OIL PUMP.

oiled paper Insulating paper impregnated with oil for waterproofing and to increase its dielectric strength.

oil-filled cable A cable whose insulation is impregnated with oil that can be maintained at a constant pressure.

oil-filled capacitor See OIL CAPACITOR.

oil-filled circuit breaker See OIL CIRCUIT BREAKER.

oil-filled transformer A transformer whose case is filled with an insulating oil.

oil fuse cutout A fuse cutout that is filled with an insulating oil. Compare OPEN-FUSE CUTOUT.

oil-immersed transformer See OIL-FILLED TRANSFORMER.

oil-impregnated capacitor See OIL CAPACITOR.

oil pump A vacuum diffusion pump using oil instead of mercury. Also see DIFFUSION PUMP.

oil switch A switch enveloped by insulating oil.

OLRT Abbreviation of *online real time* (operation).

OM Amateur radio jargon for OLD MAN: chief (male) operator, or husband of female operator.

omega A phase-dependent radionavigation system using single-frequency, time-shared, ICW transmissions from two or more locations.

omnibearing A navigational bearing obtained by means of OMNIRANGE.

omnibearing converter An electromechanical device in which an OMNIRANGE signal and vehicle heading information are combined, the output being a signal that is fed to a meter.

omnibearing indicator Abbreviation, OBI. An omnibearing converter with a dial and pointer.

omnibearing line In an OMNIRANGE system, one of the imaginary lines extending from the geographic center of the omnirange.

omnibearing selector A device that can be set manually to a selected omnibearing.

omniconstant calculator A calculator that adds or multiplies numbers in succession in such a manner as to arithmetically or geometrically increase the exponent as a single button is repeatedly pressed.

omnidirectional **1.** Pertaining to a device that responds equally well to acoustic or electromagnetic energy from any direction in three dimensions. **2.** Pertaining to a device that radiates acoustic or electromagnetic energy equally well in any direction in three dimensions. **3.** Also, NONDIRECTIONAL. Pertaining to an antenna that intercepts or radiates equally well in any azimuth (horizontal) direction.

omnidirectional antenna See NONDIRECTIONAL ANTENNA.

omnidirectional hydrophone A hydrophone that picks up underwater sounds coming from any direction.

omnidirectional microphone A microphone that picks up sounds coming from any direction.

omnidirectional radio range See OMNIRANGE.

omnidirectional range See OMNIRANGE.

omnidirectional range station Abbreviation, ORS. A radionavigation station for OMNIRANGE service.

omnigraph A Morse-code generator that operates via marked or perforated paper tape.

omnirange A radionavigation system in which each station in a chain broadcasts a beam in all directions. It usually operates at very-high frequencies (VHF) or ultra-high frequencies (UHF). Pilots of aircraft home on a particular station by tuning it in and noting its bearings.

OMR Abbreviation for OPTICAL MARK RECOGNITION.

on air See ON THE AIR.

on-call channel An assigned radio channel of which exclusive, full time use is not demanded.

on-course curvature In navigation, the rate at which the course of a vehicle deviates with reference to the distance along the true course.

on-course signal A single-tone-modulated signal indicating to an aircraft pilot following a radio beam that the flight is substantially on course.

on current See ON-STATE CURRENT.

on delay An interval during which a circuit remains off after an actuating signal has been supplied. Compare OFF DELAY.

on-demand system A system, especially in computer and data-processing operations, that delivers information or service immediately upon request.

ondograph An electromechanical device that graphically draws alternating-current waveforms on paper.

ondoscope A radio-frequency (RF) energy detector that consists of a neon bulb attached to the end of an insulating rod. When a bulb is held in an intense RF field, the field energy ionizes the gas in the bulb, causing it to glow without direct connection to the RF circuit.

one-address code In computer programming, a code in which the address in an instruction refers to only one memory location.

one condition See ONE STATE.

one-digit adder See HALF ADDER.

one-element rotary antenna A directional antenna consisting of a radiator only (no directors or reflectors) that can be rotated. A straight, rigid, half-wave rotatable dipole is the most common configuration.

one-for-one compiler A compiler that generates one machine language instruction from one source language instruction.

one-input terminal In a flip-flop, the input terminal that must be energized to switch the circuit to its logic 1 output.

one-level address See ABSOLUTE ADDRESS.

one-level subroutine In a computer program, a subroutine in which no reference is made to other subroutines.

one-lunger Colloquialism for a radio transmitter consisting entirely of a one-transistor oscillator.

one output See ONE STATE.

one-output signal The signal that results from reading a computer memory unit that is in the logic 1 state.

one-output terminal In a flip-flop, the output terminal that is energized when the circuit is in its logic 1 state.

one-plus-one address A method of computer programming in which instructions contain two addresses and an operation, the addresses referring to the location of the next instruction and the location of the data to be used.

ones complement Binary notation in the *radix-minus-one-complement form.*

one shot See MONOSTABLE MULTIVIBRATOR.

one-shot circuit See MONOSTABLE MULTIVIBRATOR.

one-shot multivibrator See MONOSTABLE MULTIVIBRATOR.

one-sided wave A waveform consisting of only negative or positive half-cycles. Example: a rectified alternating-current signal.

one state Also, *logic 1 state.* The high, on, or true logic state of a bistable device, such as a flip-flop. Compare ZERO STATE. In binary notation, the one state is represented by the digit 1.

one-third-octave band A frequency band that is ⅓ octave wide; that is, the difference between the upper-frequency limit (f_2) and the lower-frequency limit (f_1) is one-third of an octave. Also see OCTAVE BAND.

one-to-one assembler An assembler computer program that produces a machine language instruction as a result of translating a source-language statement.

one-to-one correspondence A mapping between two sets *A* and *B* so that every element in set *A* has exactly one correspondent in *B*, and every element in *B* has exactly one correspondent in *A.*

O network A four-impedance network containing two series (upper and lower) arms and two shunt (input and output) arms.

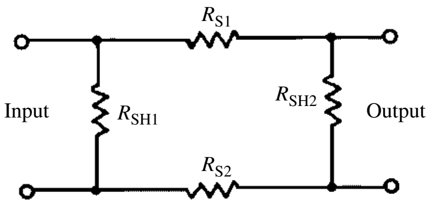

O network

one-way communication **1.** The transmission of a message to one or more stations that receive only. Compare TWO-WAY COMMUNICATION. **2.** See BROADCASTING.

one-way conduction See UNILATERAL CONDUCTIVITY.

one-way radio See ONE-WAY COMMUNICATION.

one-way repeater In wire telephony, a device that amplifies and retransmits a signal in the direction the signal was traveling when it arrived at the repeater. Compare TWO-WAY REPEATER.

one-way valve A diode or rectifier (British variation).

on interval See ON TIME.

online **1.** Descriptive of equipment under the control of a central processor. **2.** Pertaining to operations being controlled by a central processor. **3.** Relating to a computer storage device independent of a central processor. **4.** In personal computing, the use of a computer in a network (e.g., with a modem for operation on the Internet). Compare OFFLINE.

online data reduction Processing data as it enters a computer system.

online editing In video production, the final phase of editing. The original recording (not copies) is used to prepare the presentation, based on the trials that were performed using the copies. Compare OFFLINE EDITING.

online service A computer network accessible via telephone lines, television cable systems, and/or radio. It is generally available to owners of personal computers. It provides such services as electronic mail, special-interest forums, news, weather, sports, shopping, and general information.

on-off keying Keying, as in radiotelegraphy and wire telegraphy, by switching a signal source on and off to form dots and dashes or a binary code, rather than alternately changing the amplitude or frequency of the signal.

on-off operation A switching operation, especially that performed by nonmechanical (fully electronic) circuits.

on-off ratio **1.** For a circuit or device, the ratio of off time to on time. **2.** For a pulse, the ratio of pulse duration to the dead space between pulses.

on-off switch **1.** In electronic equipment, the switch by which the equipment can be started or stopped. It can be, but is not necessarily, the power-line or B-plus switch. **2.** An electronic circuit or stage that is designed to operate as a conventional switching element when triggered by an appropriate signal.

on period See ON TIME.

ONR Abbreviation of *Office of Naval Research.*

on resistance See ON-STATE RESISTANCE.

on state **1.** For a switch or switching device (such as a flip-flop), the condition of the device when it conducts current or delivers an output voltage. Compare OFF STATE, **1**. **2.** The condition of a circuit or device that is activated for operation. Compare OFF STATE, **2**.

on-state current The current flowing through a semiconductor device (such as a diode, rectifier, or thyristor) when it is conducting. Also see ON-STATE VOLTAGE.

on-state resistance The resistance of a voltage-dependent resistor that is conducting current.

on-state voltage The voltage drop across a semiconductor device (such as a diode, rectifier, or thyristor) when the device is conducting current. Also see ON-STATE CURRENT.

on the air **1.** The state of a radio station that is transmitting. **2.** In broadcasting, the condition in which the transmitter is operational and the sound and/or video from the studio is being disseminated.

on time The length of time a switch or switching device (such as a flip-flop) remains on.

on voltage See ON-STATE VOLTAGE.

OOP Abbreviation of OBJECT-ORIENTED PROGRAMMING.

op **1.** Abbreviation of OPERATE. **2.** Abbreviation of *operator* (see OPERATOR, **1**). **3.** Abbreviation of OPERATION. **4.** Abbreviation of *operational.*

opacimeter **1.** An instrument for measuring the extent to which a material blocks light. **2.** An instrument, such as a field-strength meter, used to measure the effectiveness of an electrical shielding material in blocking radio waves, X rays, or other radiation.

opacity The condition of being opaque. This applies to all forms of radiation. For example, a material can be opaque to light rays, but be transparent to radio waves, or it can be transparent to gamma rays while being opaque to alpha particles.

op amp See OPERATIONAL AMPLIFIER.

op code Abbreviation of OPERATION CODE.

open-air line See OPEN-WIRE LINE.

open-back cabinet A loudspeaker enclosure in which the space behind the speaker is open to the room.

open capacitance The value of a variable capacitor whose rotor plates have been rotated completely out of mesh with the stator plates. Compare CLOSED CAPACITANCE.

open-center display A radar display in which a ring around the center indicates zero range.

open-chassis construction A method of assembling electronic equipment by mounting components and wiring them on an unenclosed chassis, often without a front panel. Similar to breadboard construction.

open circuit **1.** A discontinuous circuit (i.e., one that is broken at one or more points and, consequently, cannot conduct current nor present a voltage at its extremities). Compare CLOSED CIRCUIT. **2.** Pertaining to no-load conditions, for example, the open-circuit voltage of a battery. **3.** For a bipolar transistor, the operating characteristics under independent input and output conditions.

open circuit breaker A circuit breaker whose contacts are open.

open-circuit current Current flowing in the primary winding of an unloaded transformer.

open-circuited line See OPEN-CIRCUITED TRANSMISSION LINE.

open-circuited transmission line Abbreviation, OCTL. An unterminated transmission line in which the conductors at the far end are not connected together.

open-circuit impedance For a transmission line or a four-terminal network, the input or driving-point impedance when the output end of the line or network is open-circuited.

open-circuit jack A telephone jack that introduces a break in a circuit until a plug connected to a closed external circuit is inserted.

open-circuit plug See OPEN PLUG.

open-circuit resistance For a four-terminal network, the input or driving-point resistance when the output end of the network is unterminated.

open-circuit signaling A system of signaling in which no current flows until the signal circuit is in active operation. In a simple telegraph circuit, for example, current flows only when the key is pressed (to form a dot or dash).

open-circuit alarm system A security system in which all the actuating sensors are normally open and connected in parallel. When one of the sensors is actuated, it closes, causing a short circuit that triggers the alarm.

open-circuit voltage See NO-LOAD VOLTAGE.

open-collector configuration In an integrated circuit, an output scheme utilizing no internal pull-up resistor. Wired-OR outputs can thus have opposite states without risk of damage to the device.

open component An open-circuit component (e.g., an open diode, coil, resistor, etc.).

open core A magnetic core having a cylindrical shape. A disadvantage of this core configuration, in some applications, is that much of the magnetic flux extends outside the core. Compare CLOSED CORE.

open-core choke A choke coil wound on an open core. Also called OPEN-CORE INDUCTOR. Compare CLOSED-CORE CHOKE.

open-core transformer A transformer wound on an open core. Compare CLOSED-CORE TRANSFORMER.

open-delta connection See VEE-CONNECTION OF TRANSFORMERS.

open-ended Pertaining to a circuit or device that can be built upon without modifying its original configuration.

open-end stub A stub that is neither short-circuited nor terminated at its far end.

open-end stub tuning Adjustment of an OPEN-END STUB, by pruning its length, for optimum operation at a given frequency.

open-entry contact In a connector, an unprotected, opening contact of the female type.

open feeder See OPEN-WIRE LINE.

open-frame machine See OPEN GENERATOR and OPEN MOTOR.

open-fuse cutout A type of enclosed-fuse cutout having an exposed fuse holder and support. Compare OIL FUSE CUTOUT.

open generator An unsealed generator (i.e., one that has openings in its housing for air circulation).

open line **1.** See OPEN-WIRE LINE. **2.** See OPEN-CIRCUITED TRANSMISSION LINE.

open loop **1.** In a control system, a feedthrough path having no feedback and that is not self-regulating. **2.** In an operational amplifier, the configuration in which there is no feedback, and in which the device operates at maximum gain. **3.** A loop within a program from which the computer automatically exits after a specific number of iterations.

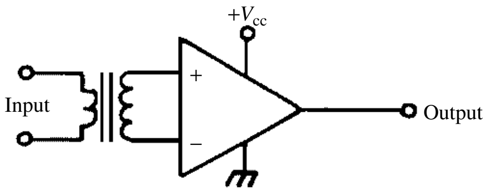

open loop, 2.

open-loop bandwidth The bandwidth of an open-loop device, such as an amplifier, without feedback. Also see OPEN LOOP.

open-loop control system A control system having no feedback and that is not self-regulating. An example is a fluid-level gauge that indicates the height of fluid in a tank, but that cannot correct the level automatically. Compare CLOSED-LOOP CONTROL SYSTEM.

open-loop differential voltage gain For a differential amplifier, the overall voltage gain (when either input is used) when the amplifier has no feedback.

open-loop gain The overall gain (ratio of output to input) of an open-loop device, such as an amplifier without feedback. Also see OPEN LOOP.

open-loop input impedance The input impedance of an open-loop device, such as an amplifier without feedback. Also see OPEN LOOP.

open-loop output impedance The output impedance of an open-loop device, such as an amplifier without feedback. Also see OPEN LOOP.

open-loop output resistance The output resistance of an open-loop device, such as an amplifier without feedback. Also see OPEN LOOP and OPEN-LOOP IMPEDANCE.

open-loop system **1.** A circuit in which the input and output currents are independent. **2.** A robot that does not use a servo system. It depends entirely on alignment and precision for positioning accuracy. **3.** An electromechanical device that does not use corrective feedback.

open-loop voltage gain The overall voltage gain of an open-loop amplifier (i.e., one having no feedback). Also see OPEN LOOP.

open magnetic circuit A magnetic circuit in which a complete path is not provided for magnetic flux. See, for example, OPEN CORE. Compare CLOSED MAGNETIC CIRCUIT.

open motor An unsealed motor (i.e., one that has openings in its housing for air circulation).

open-phase protection Use of an automatic device, such as an open-phase relay, to interrupt the power to a polyphase system when one or more phases are open-circulated.

open-phase relay In a polyphase system, a protective relay that opens when one or more phases are open-circuited. Also see OPEN-PHASE PROTECTION.

open plug A phone plug to which no external connections are made; it is used to hold the blades of a jack as if they were plugged in.

open-reel A tape-recording system in which the tape, during record or playback condition, is wound into a take-up reel that is physically separate from the tape reel. Also called *reel-to-reel arrangement.*

open relay **1.** A relay in its open-contact state. **2.** A relay having an open-circuited coil. **3.** An unenclosed relay.

open routine In computer operations, a routine that can be inserted directly into a larger routine and requires no link to the main program.

open subroutine In computer operations, a subroutine that can be inserted into a larger instructional sequence and must be recopied whenever it is required. Also called *direct-insert subroutine.*

Open Systems Interconnection Reference Model Abbreviation, OSI-RM. A standard set of protocols for computer network communication. It consists of seven levels, also called layers: *physical layer*, *link layer*, *network layer*, *transport layer*, *session layer*, *presentation layer*, and *application layer.*

open temperature pickup A temperature transducer that must be placed directly in contact with the monitored medium.

open volume Pertaining to the maximum-gain operation of a sound-reproducing system (i.e., operation at full volume).

open wire **1.** An unterminated wire. **2.** A wire supported above the surface of the earth and often ungrounded. **3.** See OPEN-WIRE LINE.

open-wire feeder See OPEN-WIRE LINE.

open-wire line A transmission line or feeder usually consisting of two straight, parallel wires held apart by bars of low-loss insulating material at regular intervals along the line.

open-wire loop A branch line connected to a main open-wire line.

open-wire transmission line See OPEN-WIRE LINE.

open-wire wavemeter See LECHER WIRES.

operand In computer operations, a quantity that enters into or results from an operation.

operate **1.** To manipulate according to an established procedure (e.g., to operate an instrument). **2.** To perform according to specifications, in the sense that an electronic circuit functions.

operate current A signal current or trigger current required to actuate a device. Compare OPERATE VOLTAGE.

operate delay See OPERATE TIME, **1**.

operate interval See OPERATE TIME, **2**.

operate time **1.** The interval starting after the application of an operate current or voltage to a device, and ending when the device operates. **2.** The period during which an electronic equipment is in operation. Also see OPERATING TIME, **1**.

operate voltage The signal voltage or trigger voltage required to actuate a device. Compare OPERATE CURRENT.

operating ambient temperature Abbreviation, OAT. The maximum or recommended temperature in the space immediately surrounding an equipment in operation.

operating angle In an amplifier circuit, the excitation-signal cycle, in degrees, during which drain, collector, or plate current flows. Class-A amplifiers operate for 360 degrees of the input signal cycle; class-AB amplifiers operate for more than 180 degrees, but less than 360 degrees of the input signal cycle; class-B amplifiers operate for 180 degrees of the input signal cycle; class-C amplifiers operate for less than 180 degrees of the input signal cycle.

operating bias In a circuit containing transistors, diodes, or vacuum tubes, the value(s) of direct-current bias required for normal operation.

operating code The code used by the operator in a computer or data-processing system.

operating conditions The environment in which a circuit or system functions in normal use.

operating current The current required by a device during its operation. Compare IDLING CURRENT.

operating cycle The sequence of events in the operation of a device. For example, the repetitive operation of a neon-bulb relaxation oscillator is a sequence of three events: (1) slow charge of capacitor, (2) firing of bulb, and (3) abrupt discharge of capacitor.

operating frequency **1.** The fundamental frequency at which a circuit or device is operated. **2.** The frequency of the current, voltage, or power delivered by a generator.

operating frequency range **1.** The range of operating frequencies, expressed as a minimum and maximum, for a communications receiver, transmitter, or transceiver. **2.** For an integrated-circuit oscillator, the minimum and maximum frequency, and all frequencies in between, at which the device is guaranteed to operate.

operating life The maximum period (from seconds to years) over which a device will operate before failure (from which it usually cannot recover). Compare SHELF LIFE.

operating line A line drawn across a family of curves depicting the performance of a device. It intersects each curve at a single point and graphically displays the performance of the device for a given condition. Thus, an operating line on a family of output curves for a transistor might depict operation with a given load resistor.

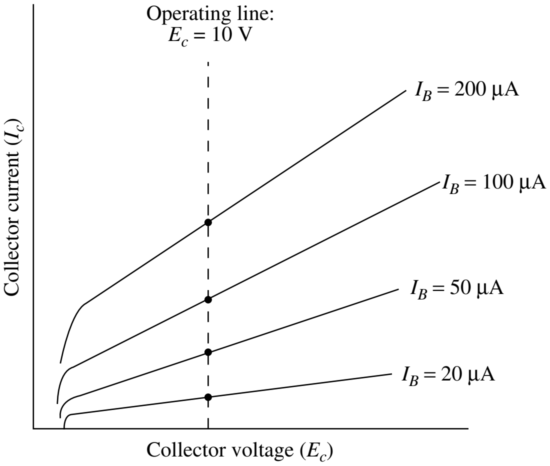

operating line

operating overload The extent and/or duration of overload to which an equipment can be exposed during customary operation, and still continue to properly function.

operating point On the response curve for a device, the point indicating the quiescent level of operation (such as determined by a fixed bias voltage or current). An alternating-current signal applied at this point oscillates above and below the point as a mean.

operating-point shift A movement of an operating point due to faulty operation of a circuit or device, or to a value change in some critical component.

operating position **1.** The control point in a system, [i.e., the place where an operator (see OPERATOR, **1**) normally functions]. **2.** The actual or recommended physical orientation of a device during its operation (e.g., a vertical operating position for a power vacuum tube).

operating power **1.** The power actually used by a device during its operation. **2.** The antenna power of a radio station.

operating ratio For a given period, the ratio t_o/t, where t_o is the time during which an equipment is operating correctly, and t is the duration of the period.

operating station In a computer installation, one or more consoles for the control of a data-processing system by an operator.

operating system Abbreviation, OS. In a computer, a set of programs that oversee the functioning of the hardware and application software. It works with the basic input/output system (BIOS).

operating temperature The actual or recommended temperature of a device during its operation.

operating-temperature range For a given device, the spread between maximum and minimum values of operating temperature.

operating time **1.** The interval during which an equipment is in operation. **2.** The period corresponding to OPERATING ANGLE.

operating-time characteristic For a coil-type relay, the relationship between operating time and operating power.

operating voltage The voltage required by a device, or measured at the device, during its operation. Compare IDLING VOLTAGE.

operation **1.** The working of a circuit or device (i.e., its performance). **2.** A process usually involving a sequence of steps (e.g., a mathematical operation).

operational amplifier Also called an *op amp.* A specialized linear integrated circuit (IC) that consists of several transistors, resistors, diodes, and capacitors, interconnected to produce gain over a wide range of frequencies. A single amplifier can comprise an entire IC, or an IC can contain several amplifiers. The *dual op amp* and the *quad op amp* are common variants. Some ICs contain one or more amplifiers in addition to other circuits. The devices can be used with resistance-capacitance (RC) network combinations to build active filters for use at audio frequencies.

operational differential amplifier An OPERATIONAL AMPLIFIER preceded by a DIFFERENTIAL AMPLIFIER.

operational readiness In statistical analysis, the probability that a system will, at a certain time, be correctly operating or ready to operate.

operational reliability Reliability determined empirically from a study of the actual operation of a device or system under controlled conditions. Also called ACHIEVED RELIABILITY.

operational transconductance amplifier Abbreviation, OTA. An integrated-circuit amplifier that differs from the conventional operational amplifier in that its output current is proportional to its input-signal voltage.

operation code In computer operations, the part of an instruction that specifies an operation.

operation decoder In a digital computer, the circuit that reads an OPERATION CODE and directs other circuitry in the execution of the code.

operation envelope See WORK ENVELOPE.

operation number In computer programming, a number that indicates the position in the program of a particular operation or subroutine.

operation part In a computer program, the part of an instruction containing the OPERATION CODE.

operation register In a digital computer, the register that stores the operation code of an instruction.

operations research A branch of computer engineering, devoted to the solution and/or optimization of functions of many variables.

operation time The interval between the instant of application of all voltages to a circuit, and the instant when the current reaches a specified percentage of its final steady value.

operator **1.** A person who performs an operation (see OPERATE, **1**). **2.** In mathematics, a symbol indicating an operation (e.g., *j*, +, –, ×, etc.).

operator j See J OPERATOR.

opposition **1.** The state of two quantities that are 180 degrees out of phase with each other. **2.** A state in which two effects or quantities operate against each other in some manner (e.g., physically, mathematically, electrically, etc.).

opt Abbreviation of OPTICAL.

Optacon An electronic aid for the blind. It has a camera that scans printed matter and a device that forms corresponding raised letters that can be read, as would Braille, with the fingertips. The name is a contraction of *optical to tactile converter.*

optical **1.** Generating or sensing visible light. **2.** Visible electromagnetic radiation in the range of approximately 390 to 750 nanometers wavelength.

optical ammeter A type of OPTICAL PYROMETER that measures the current flowing through the filament of an incandescent lamp.

optical antenna A photosensitive device intended for the reception of modulated-light or modulated-infrared signals. Usually consists of a convex lens or a concave reflector and a semiconductor photodetector.

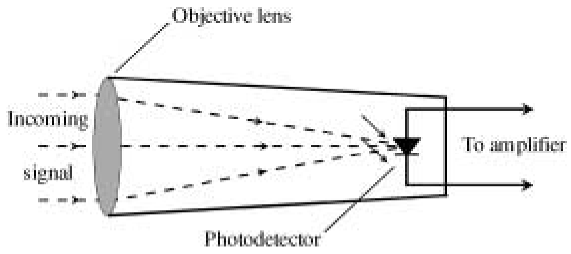

optical antenna

optical axis The z-axis of a quartz crystal (see Z-AXIS, **4**).

optical character reader A device that uses OPTICAL CHARACTER RECOGNITION to discern printed characters.

optical character recognition Abbreviation, OCR. In computer and data-processing operations, the reading of alphabetical, numerical, and other characters from hard copy (usually printed matter) by photoelectric methods. It converts the characters into digital data that can be stored in computer memory, on disks and tapes, and transmitted via digital communications networks. It can also allow a computer or robot to read signs, maps, etc.

optical communications One-way or two-way communications via modulated visible light. It can be conducted through optical fibers, the atmosphere, water, or any other transparent medium.

optical coupler A coupling device consisting essentially of a light source (actuated by an input signal) mounted in an opaque housing with a light-sensitive device (that delivers the output signal). In its simplest form, the arrangement consists of a light-emitting diode (LED) and a photodiode.

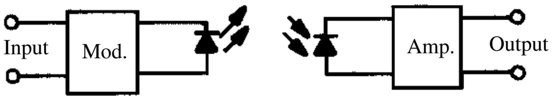

optical coupler

optical detector An integrated-circuit (IC) that provides light-to-voltage conversion. Its direct-current output voltage is proportional to the intensity of light impinging on its sensor.

optical encoder An electronic device that measures the extent to which a mechanical shaft has rotated. It can also measure the rate of rotation (angular speed). It consists of a light-emitting diode (LED), a photodetector, and a disk with alternate light and dark bands (chopping wheel). It is commonly used in digital controls and jointed robot arms.

optical fiber A glass or plastic medium through which light is propagated for optical-communications purposes. The refractive characteristics of the fiber keep the visible light inside.

optical lever A device for amplifying the effect of a small rotation. The rotating member carries a small mirror that reflects a light beam over a curved scale, the distance through which the light spot travels on the scale being proportional to the distance between the scale and the rotating mirror. In this way, the deflection on the scale is several times the length of the arc described by the mirror, the rotation being thus amplified.

optical link See OPTICAL COUPLER.

optical mark recognition A method of data transfer that involves the use of optical techniques.

optical maser See LASER.

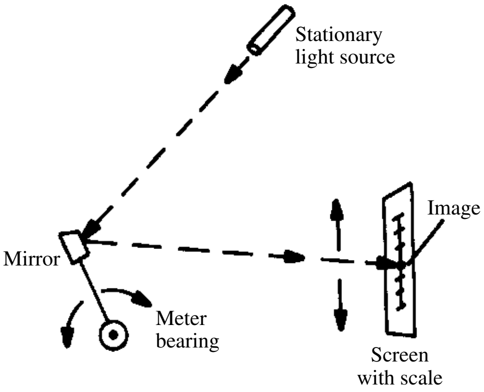

optical lever

optical mode The vibration mode that produces an oscillating dipole in a crystal lattice.

optical pattern A Christmas-tree-like pattern produced by parallel rays of light striking a phono disk. Also see DIFFRACTION.

optical playback See OPTICAL SOUND REPRODUCER.

optical pyrometer A pyrometer for measuring the temperature of a hot body in terms of the intensity and color of light it emits.

optical scanner An electronic device that converts hard-copy text and graphics into digital form suitable for processing and storage in a computer. There are several types; the most common resembles a small photocopy machine and is known as a *flatbed scanner*. See OPTICAL SCANNING.

optical scanning An electronic process that converts hard-copy text and graphics into digital form suitable for processing and storage in a computer. Text scanners work in conjunction with specialized software. The optical scanner and software together form the equivalent of audio speech recognition and sound interpretation. Usually, scanners are employed to convert printed matter to digital data. However, scanning technology can also be used at a distance, for example, by a robot for the purpose of reading road signs.

optical sound recorder A photoelectric machine for recording sound on photographic film. Also see OPTICAL SOUND RECORDING.

optical sound recording A system for recording sound on photographic film, as in sound motion pictures. The sound is picked up by a microphone and amplified to vary the intensity of a light source. The film passing this modulated beam becomes exposed to a variable-width or variable-density track corresponding to the modulation (see VARIABLE-DENSITY SOUND RECORD and VARIABLE-WIDTH SOUND RECORD). When developed film is played back, its sound track modulates a light beam in the reproducer by actuating a photocell or phototube to produce the audio signal, which is then amplified.

optical sound reproducer A photoelectric machine for reproducing sound on film. A light beam in the device is modulated by the passing sound track, and it, in turn, modulates photocell or phototube current, which is amplified to drive a loudspeaker. Also see OPTICAL SOUND RECORDING.

optical system Collectively, the functional arrangement of lenses, mirrors, prisms, and related devices in optoelectronic apparatus.

optical tachometer An optoelectronic instrument for measuring (by means of reflected-light variations) the speed of a body, such as a rotating shaft, without electrical or mechanical attachments to the latter.

optical thermometer See OPTICAL PYROMETER.

optical twinning A kind of defect in which two types of quartz occur in the same crystal. Compare ELECTRICAL TWINNING.

optical type font A special type (printing) style designed for use with OPTICAL CHARACTER RECOGNITION equipment.

optical wand A pencil-like optical probe, used to read bar codes from a printed page and translate the codes into information that is then loaded into a computer or calculator.

optic axis See OPTICAL AXIS.

optic flow See EPIPOLAR NAVIGATION.

optics **1.** The science of light, its measurement, application, and control. **2.** A system of lenses, prisms, filters, or mirrors used in electronics to direct, control, or otherwise modify light rays.

optimization The adjustment or manipulation of the elements of a process or system for the best operation or end result.

optimize **1.** To manipulate a set of variables or parameters for the best possible performance of a circuit or system. **2.** To maximize the value of a multivariable function.

optimum angle of radiation For a given altitude of the ionosphere, the angle, with respect to the horizon, at which a radio signal should be transmitted from a specific geographic location to provide optimum reception at some other specific location. This angle varies, depending on the altitude of the ionospheric layer from which signals are returned, the distance between the transmitting and receiving stations, and the number of hops taken by the signal between the earth and the ionosphere.

optimum bunching In a velocity-modulated tube, such as a klystron, the bunching associated with maximum output.

optimum collector load The ideal load impedance (see OPTIMUM LOAD) for a particular transistor operated in a specified manner.

optimum coupling The degree of coupling between two circuits tuned to the same frequency that results in maximum energy transfer. Also called CRITICAL COUPLING.

optimum current The value of current that produces the most-effective performance of a circuit or device.

optimum frequency See OPTIMUM WORKING FREQUENCY.

optimum load The value of load impedance that produces maximum transfer of power from a generator or amplifier.

optimum plate load The ideal load impedance (see OPTIMUM LOAD) for a particular electron tube operated in a specified manner.

optimum *Q* The most effective figure of merit for a capacitor, inductor, or tuned circuit at a specified frequency.

optimum reliability The value of reliability that ensures minimum project cost.

optimum voltage The value of voltage that results in the most effective performance of a circuit or device.

optimum working frequency In radio transmission involving reflection from the ionosphere, the frequency of use that results in the most reliable communication between two points.

optoelectronic coupler An assembly consisting of an LED and a phototransistor. An input signal causes the diode to glow, and the light activates the transistor, which in turn delivers an output signal of higher amplitude than that of the input signal.

optoisolator A light-emitting diode (LED) or infrared-emitting diode (IRED) and a photodiode combined in a single package with a transparent gap between them. The LED or IRED converts an electrical signal to visible light or IR; the photodiode changes the visible light or IR back into an electrical signal. This device eliminates interstage impedance reflections that can plague systems that employ electronic coupling. Even the most drastic changes in load impedance have no effect on the impedance “seen” by the signal source. The devices can also protect expensive equipment against electrical transients; for example, in telephone-line modems used with personal computers.

optoelectronics A branch of electronics that involves the use of visible light for communications or data-transfer purposes.

optoelectronic transistor A transistor having an electroluminescent emitter, transparent base, and photoelectric collector.

optophone A photoelectric device for converting light into sounds of proportionate pitch to enable blind persons to “see” by ear.

Orange Book Trade name (Sony and Philips) for a specialized format for *compact-disc recordable (CD-R)* computer data storage media. It allows recording of data onto, as well as retrieval of data from, an optical compact disc. See also GREEN BOOK, RED BOOK, and YELLOW BOOK.

orange peel On a phonograph disc, a mottled surface that produces high background noise; so called from its resemblance to an orange peel.

orbit **1.** Also called *shell.* The median path of an electron around the nucleus of an atom. **2.** The circular or elliptical path of an artificial satellite (e.g., a communications satellite) around the earth. **3.** The elliptical path of the moon around the earth. **4.** The elliptical path of the earth or another planet around the sun. **5.** Any closed, usually circular or elliptical path, that an object follows around another object.

orbital-beam multiplier tube An electron-multiplying UHF oscillator or amplifier tube in which a positively charged electrode focuses electrons in a circular path.

orbital electron An electron in orbit around the nucleus of an atom.

orbital period The length of time required by a satellite to go around a celestial body, usually the earth. It can vary from about 90 minutes for very low orbits to months or even years for orbits beyond that of the moon. The moon's period is about a month.

OR circuit Also called *inclusive-OR circuit.* In digital systems and other switching circuits, a logic gate whose output is high (logic 1) when any of the input signals is high. The output is low (logic 0) only when all of the inputs are low. Compare AND CIRCUIT.

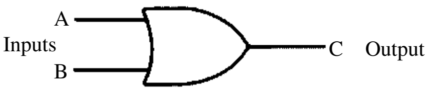

A	B	C
0	0	0
0	1	1
1	0	1
1	1	1

2-input OR circuit

order **1.** An instruction to a digital computer. **2.** Sequence.

OR element See OR CIRCUIT.

ordered pair A set of two variables or numbers that can be mapped one-to-one onto a set of points in a plane. The most familiar example is (x,y), used to depict the Cartesian (rectangular) coordinate system on which characteristic curves and other functions are plotted.

order of interference The position of an interference fringe that depends on whether the interference arises from one, two, three, or more wavelength differences. Generally, the first fringe results from a difference of one wavelength, the second fringe from a difference of two wavelengths, etc. The center fringe indicates that the two signals are exactly in phase; that is, the difference is zero wavelengths.

order of logic **1.** A mathematical expression for the complexity of a system of logic. **2.** The relative speed with which logical information is processed in a system.

order tone A warning signal to receiving operators in the form of a tone transmitted over a trunk preceding the transmission of an order.

ordinary ray As a result of the double refraction of electromagnetic waves, the member of a pair of rays that follows the usual laws of refraction. Also see O WAVE. Compare EXTRAORDINARY RAY.

ordinary wave See O WAVE.

ordinate In the rectangular coordinate system, a point located on the vertical axis.

organ **1.** A computer subsystem. **2.** An electronic device used for the purpose of generating music.

organic decay See EXPONENTIAL DECREASE.

organic electricity Electricity in the living tissues of animals and plants.

organic growth See EXPONENTIAL INCREASE.

organic semiconductor A semiconductor material consisting of, or combined with, some compound of carbon.

OR gate See OR CIRCUIT.

orientation **1.** The direction or position of an object in space, expressed in terms of coordinate values. **2.** In a teleprinter, the calibration or alignment that determines the speed of response to a received character.

orientation of quartz plates See CRYSTAL AXES and CRYSTAL CUTS.

orifice An opening or window, such as that in a loudspeaker enclosure or waveguide, through which energy is transmitted.

origin **1.** The starting point in a coordinate system. **2.** Relative to address modification in computer operations, an address to which a modifier is added to derive a variable operand address.

original lacquer In disc-recording operations, an original recording made on a lacquer-surfaced disc that is subsequently used to make a master.

original master In disk recording, the master disc produced from a wax or lacquer disc (see ORIGINAL LACQUER) by means of electroforming.

origin distortion The change in the shape of a wave as it swings through zero (polarity).

OR operation See OR CIRCUIT.

ORS Abbreviation of OMNIDIRECTIONAL RANGE STATION.

orthicon A television camera tube somewhat similar to the iconoscope, but which provides internal amplification of light and, accordingly, can be used in dimmer places than the iconoscope. Light amplification is provided by an arrangement similar to that of a photomultiplier tube.

orthiconoscope See ORTHICON.

orthoacoustic recording **1.** A system of disc recording in which the inherent differences between high-frequency recording and low-frequency recording are compensated to provide reproduction that more closely resembles the actual sound. **2.** A disc made by the method defined in **1**.

orthogonal axes Perpendicular axes [e.g., those in a Cartesian (rectangular) coordinate system].

Os Symbol for OSMIUM.

OS Abbreviation of OPERATING SYSTEM.

osc Abbreviation of OSCILLATOR.

OSCAR Abbreviation for *Orbiting Satellite Carrying Amateur Radio.* A satellite with a transponder that has an uplink in one amateur band and a downlink in another amateur band.

osciducer See OSCILLATING TRANSDUCER.

oscillate **1.** To fluctuate in amplitude in a uniform manner. **2.** To vary above and below a specified value at a constant rate.

oscillating arc A small arc, especially one produced by slow-opening relay contacts, that generates high-frequency oscillations.

oscillating circuit A closed circuit containing inductance, capacitance, and inherent resistance, in which energy passes back and forth between inductor and capacitor at a frequency determined by the inductance (L) and capacitance (C) values.

oscillating crystal **1.** A piezoelectric plate maintained in a state of oscillation in a circuit. See, for example, CRYSTAL OSCILLATOR and QUARTZ CRYSTAL. **2.** An oscillating semiconductor diode (see NEGATIVE-RESISTANCE DIODE, **1**, **2**).

oscillating current See OSCILLATORY CURRENT.

oscillating detector A detector provided with positive feedback; therefore, it is capable of generating a signal of its own. Compare NON-OSCILLATING DETECTOR.

oscillating diode **1.** A semiconductor diode biased into its negative-resistance region so that it oscillates in a suitable circuit. **2.** An oscillating tunnel diode. **3.** Any of several microwave diodes, such as the IMPATT diode, which will oscillate in a suitable system. **4.** See MAGNETRON. Also see DIODE OSCILLATOR.

oscillating field An alternating electric or magnetic field.

oscillating rod A rod of magnetostrictive metal maintained in a state of oscillation in a circuit. See, for example, MAGNETOSTRICTION and MAGNETOSTRICTION OSCILLATOR.

oscillating transducer A transducer in which an input quantity varies a frequency proportionately from its center value.

oscillating wire A wire of magnetostrictive metal maintained in a state of oscillation in a circuit. See, for example, MAGNETOSTRICTION and MAGNETOSTRICTION OSCILLATOR.

oscillation The periodic change of a body or quantity in amplitude or position (e.g., oscillation of a pendulum, voltage, or crystal plate).

oscillation constant For an oscillating inductance-capacitance (LC) circuit, the expression $(LC)^{1/2}$, where L is the inductance in henrys and C is the capacitance in farads. The reciprocal of the OSCILLATION NUMBER.

oscillation control A manual or automatic device for adjusting the frequency or amplitude of the signal generated by an oscillator.

oscillation efficiency The ratio, as a percentage, of the alternating-current (signal) power output of an oscillator (P_{out}) to the corresponding direct-current power input (P_{in}). $Efficiency = 100P_{out}/P_{in}$.

oscillation number For an oscillating circuit, the number of complete oscillation cycles that occur in 6.28 (2 pi) seconds.

oscillation test **1.** A test of an oscillator to determine if a signal is being generated. **2.** A test for transistors wherein the transistor is used as an oscillator to give a rough indication of its condition in terms of oscillation amplitude.

oscillation transformer A tank coil of a radio transformer—especially one that includes an output coupling coil.

oscillator A device that produces an alternating or pulsating current or voltage electronically. The term is sometimes used to describe any alternating-current-producing device other than an electromechanical generator.

oscillator circuit The specific manner in which the components of an OSCILLATOR are interconnected. The three general types are: positive-feedback, negative-resistance, and relaxation.

oscillator coil A tapped coil that provides the input and output windings required for an oscillator circuit. Such coils are used in signal generators, oscillators, and superheterodyne receivers.

oscillator-doubler A circuit consisting of an oscillator and a frequency doubler (e.g., a crystal oscillator whose output frequency is twice the crystal frequency).

oscillator drift A usually gradual change in frequency of an oscillator caused by such factors as warmup time, voltage variations, capacitance change, inductance change, or change in transistor characteristics.

oscillator frequency The fundamental frequency at which an oscillator operates. It can be determined by a tuned circuit, crystal, cavity, section of waveguide or transmission line, or by a resistance-capacitance circuit.

oscillator harmonic interference In a superheterodyne receiver, interference that is the beat product of local oscillator harmonics and received signals.

oscillator interference Radio-frequency interference caused by signals from the high-frequency oscillator of a receiver.

oscillator keying Keying by making and breaking the signal output, direct-current (dc) power, or dc bias of the oscillator stage of a radiotelegraph transmitter.

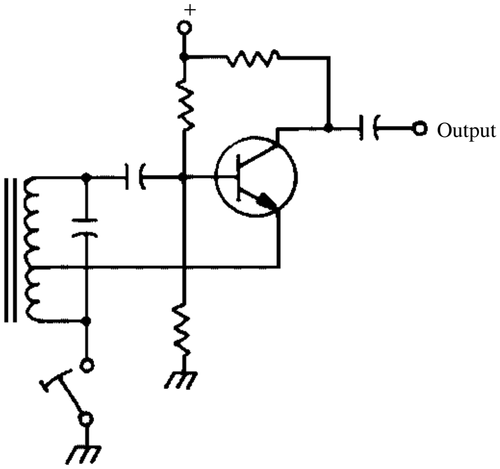

oscillator keying

oscillator-mixer **1.** A combination stage in which a transistor functions as a local oscillator and mixer in a receiver or test instrument. **2.** A device designed specifically to function as a local oscillator and mixer.

oscillator-mixer-detector **1.** In a superheterodyne receiver, a stage in which the functions of high-frequency oscillator, mixer, and first detector are performed by a single transistor. **2.** A device designed specifically to function as a local oscillator, mixer, and detector.

oscillator-multiplier A single circuit that serves simultaneously as an oscillator and frequency multiplier. See, for example, OSCILLATOR-DOUBLER.

oscillator padder In a superheterodyne receiver, a small, limited-range variable capacitor connected in series with the oscillator coil for tracking oscillator tuning at the low end of a band. Compare OSCILLATOR TRIMMER.

oscillator power supply **1.** The direct-current or alternating-current power supply for an oscillator. **2.** See OSCILLATOR-TYPE POWER SUPPLY.

oscillator radiation The emission of radio-frequency energy by the oscillator stage of a superheterodyne receiver. Also see OSCILLATOR INTERFERENCE.

oscillator-radiation voltage The radio-frequency voltage at the antenna terminals of a superheterodyne receiver that results from signal emission by the oscillator stage.

oscillator stabilization **1.** The automatic compensation of an oscillator circuit for the frequency drift resulting from changes in temperature, current, voltage, or component parameters. **2.** The automatic stabilization of the operating point of an oscillator circuit against variations resulting from changes in temperature, supply current or voltage, or component parameters.

oscillator synchronization The locking of an oscillator in step with another signal source, such as a frequency-standard generator.

oscillator tracking In a superheterodyne receiver, the constant separation of the oscillator frequency from the signal frequency by an amount equal to the intermediate frequency at all settings of the tuning control.

oscillator transmitter A radio transmitter consisting only of a radio-frequency oscillator and its power supply. The oscillator can be modulated in various ways [e.g., on-off keying, frequency-shift keying, voice amplitude modulation (AM), voice frequency modulation (FM)].

oscillator trimmer In a superheterodyne receiver, a small, limited-range capacitor connected in parallel with the oscillator coil for tracking oscillator tuning at the high end of a band. Compare OSCILLATOR PADDER.

oscillator tuning The separate, often ganged, tuning of the oscillator stage in a circuit.

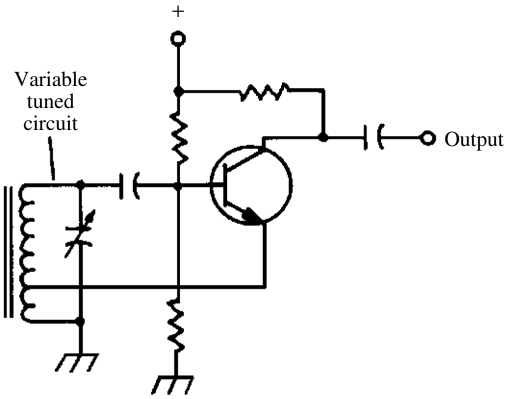

oscillator tuning

oscillator-type power supply A high-voltage, direct-current (dc) power supply in which a radio-frequency (RF) oscillator generates a low-voltage alternating current (ac). This ac voltage is stepped up by an RF transformer, and is finally rectified to obtain high-voltage dc.

oscillator-type transmitter See OSCILLATOR TRANSMITTER.

oscillator wavelength The fundamental wavelength at which an oscillator operates. It is usually expressed in meters, centimeters, or millimeters. It can be determined by a tuned circuit, crystal, cavity, section of waveguide or transmission line, or by a resistance-capacitance circuit.

oscillatory current A current that alternates periodically, particularly the current in an inductance-capacitance (LC) tank circuit that results from the oscillation of energy back and forth between the inductor and capacitor.

oscillatory discharge An electrical discharge, such as that of a capacitor, that sets up an OSCILLATORY CURRENT.

oscillatory surge A current or voltage surge that includes both positive and negative excursions.

oscillatory transient See OSCILLATORY SURGE.

oscillistor A device consisting essentially of a bar of semiconductor material positioned in a magnetic field; it will produce oscillations under certain conditions.

oscillogram **1.** The image produced on the screen of an oscilloscope. **2.** A permanent, usually photographic, record made from the screen of an oscilloscope.

oscillograph **1.** An instrument that makes a permanent record (photograph or pen recording) of a rapidly varying electrical quantity. Also called *recorder* (see RECORDER, **2**). Compare OSCILLOSCOPE. **2.** An obsolete term for OSCILLOSCOPE.

oscillograph recorder A direct-writing recorder (see RECORDER, **2**).

oscillography The use of a graphic oscillation recorder (OSCILLOGRAPH).

oscillometer A device used for determining the peak amplitude of an oscillation.

oscilloscope An instrument that presents for visual inspection the pattern representing variations in an electrical quantity. Also see CATHODE-RAY OSCILLOSCOPE. Compare OSCILLOGRAPH.

oscilloscope camera A special high-speed, short-focus camera with fixtures for attachment to an oscilloscope to record images from the screen. Standard and instant-film types are available.

oscilloscope differential amplifier An amplifier that processes the difference between two signals, for the purpose of displaying on an oscilloscope or oscillograph.

oscilloscope tube A cathode-ray tube for use in an oscilloscope. It contains an electron gun, accelerating electrode, horizontal and vertical deflecting plates, and a fluorescent screen.

Os-Ir Symbol for OSMIRIDIUM.

OSI-RM Abbreviation of OPEN SYSTEMS INTERCONNECTION REFERENCE MODEL.

OSL Abbreviation of *orbiting space laboratory.*

osmiridium Symbol, Os-Ir. A natural alloy of osmium and iridium.

osmium Symbol, Os. A metallic element of the platinum group. Atomic number, 76. Atomic weight, 190.2.

osmotic pressure The force that causes the positive ions to pass out of a solution toward a metal body immersed in an electrolyte. Also see HELMHOLTZ DOUBLE LAYER.

OSO Abbreviation of *orbiting solar observatory.*

osteophone A bone-conduction hearing aid.

OTA Abbreviation of OPERATIONAL TRANSCONDUCTANCE AMPLIFIER.

OTL Abbreviation of OUTPUT-TRANSFORMERLESS.

O-type backward-wave oscillator Abbreviation, OBWO. A backward-wave oscillator using harmonics having opposing phases.

ounce Abbreviation, oz. A unit of weight equal to $\frac{1}{16}$ pound or 28.35 grams.

ounce-inch Abbreviation, oz-in. A unit of torque equal to the product of a force of 1 ounce and a moment arm of 1 inch. Compare POUND-FOOT.

outage **1.** Loss of power to a system. **2.** Loss of a received signal.

outboard components **1.** Discrete components (capacitors, coils, resistors, or transformers) connected externally to an integrated circuit. **2.** Discrete components connected externally to any existing electronic device.

outcome In statistical analysis, the result of an experiment or test. An outcome can be numerical or nonnumerical.

outdoor antenna An antenna erected outside, usually high above the surface of the earth clear of obstacles. It generally provides superior performance compared with an INDOOR ANTENNA. Also reduces the probability of radio-frequency interference (RFI) when used for transmitting.

outdoor booster A signal preamplifier mounted on an outdoor television receiving antenna for improved reception.

outdoor transformer A weatherproof distribution transformer installed outside the building it services.

outer conductor The outer metal cylinder or jacket of a coaxial cable or coaxial tank. Compare INNER CONDUCTOR.

outgassing **1.** In the evacuation of electronic devices, such as vacuum tubes, the removal of occluded gases from glass, ceramic, and metal by means of slow baking and by flashing an internal metal getter (such as one of magnesium). **2.** The production of gases in certain electrochemical cells and batteries during the final stage of charging.

outgoing line A power or signal line that leaves a device, facility, or stage. Compare INCOMING LINE.

outlet A female receptacle that delivers a signal or operating power to equipment plugged into it.

outline flowchart In computer operations, a preliminary flowchart showing how a program will be divided into routines and segments, input and output functions, program entry points, etc.

out-of-line coding Instructions for a computer program routine stored in an area of memory other than that in which the routine's program is stored.

out of phase Pertaining to the condition in which the alternations or pulsations of two or more separate waves or wave phenomena, having identical frequencies, are out of step with each other. Compare IN PHASE.

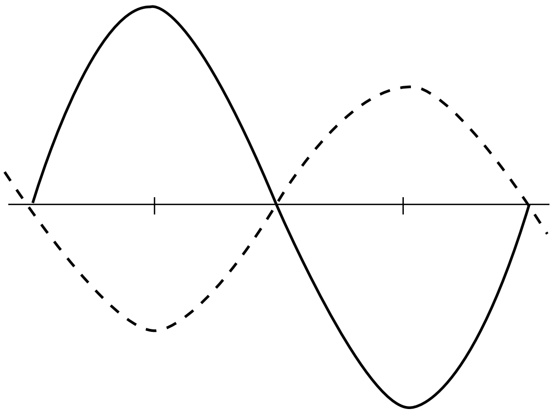

out of phase

out-of-phase current Reactive current in an alternating-current circuit (i.e., current that is out of phase with voltage. Also see QUADRATURE CURRENT).

out-of-phase voltage Voltage across a reactance; so called because it is out of phase with the current.

outphaser A device that converts a sawtooth wave to a square wave. It is used in electronic organs and synthesizers.

outphasing modulation A system of modulation in which the sideband frequencies are shifted 90 degrees from the phase position in an amplitude-modulated wave. The resulting constant-envelope wave is then amplified with high efficiency and low distortion by a class-C stage; then the signal is reconverted to an amplitude-modulated one by phase shifting the carrier, with respect to the sidebands.

out-plant system A data-processing system in which a central computer receives data from remote terminals.

output **1.** Energy or information delivered by a circuit, device, or system. Compare INPUT, **1**. **2.** The terminals at which energy or information is taken from a circuit, device, or system. Compare INPUT, **2**.

output admittance Symbol, Y_o. The internal admittance of a circuit or device, as "seen" at the output terminals; the reciprocal of OUTPUT IMPEDANCE. Compare INPUT ADMITTANCE.

output amplifier See FINAL AMPLIFIER.

output area In a computer system, the portion of storage holding information for delivery to an output device. Also called *output block.*

output axis For a gyroscope that has received an input signal, the axis around which the spinning wheel precesses.

output block See OUTPUT AREA.

output buffer **1.** A circuit that follows an oscillator and reduces the effects of variable load impedance on the oscillator frequency or signal amplitude. **2.** An amplifier, usually with a voltage gain of 6 dB, that follows a video multiplexer. The amplifier drives a coaxial transmission line.

output bus driver In a computer, a device that amplifies output signals sufficiently to provide signals to other devices without undue loading of the supply line (bus).

output capability The maximum power or voltage output that a circuit can deliver without distortion or other improper operating conditions.

output capacitance Symbol, C_o. The internal capacitance of a circuit or device, as seen at the output terminals. Compare INPUT CAPACITANCE.

output capacitive loading For an operational amplifier at unity gain, the maximum capacitance that can be connected to the output of the amplifier before phase shift increases to the point of oscillation.

output capacitor **1.** In a capacitance-coupled circuit, the output coupling capacitor. Compare INPUT CAPACITOR. **2.** The last capacitor in a power-supply filter circuit.

output capacity The maximum output capability of a device or system expressed in appropriate units, such as current, voltage, power, torque, horsepower, etc.

output choke The last choke (inductor) in a power-supply filter circuit.

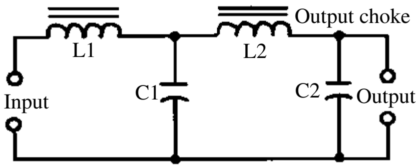

output choke

output circuit The circuit or subcircuit that constitutes the output portion of a network or device. Also see OUTPUT and OUTPUT TERMINALS. Compare INPUT CIRCUIT.

output-circuit distortion Distortion in the output portion of a circuit or device (such as a transistor or transformer), usually caused by an overload or nonlinear response.

output conductance Symbol, G_o. The internal conductance of a circuit or device, as "seen" at the output terminals. It is the reciprocal of OUTPUT RESISTANCE. Compare INPUT CONDUCTANCE.

output control **1.** The gain control of an amplifier. **2.** The level control of a variable power supply.

output coupling capacitor See OUTPUT CAPACITOR.

output coupling transformer See OUTPUT TRANSFORMER.

output current **1.** Symbol, I_o. The current delivered by a source, such as a battery, generator, or amplifier. Compare INPUT CURRENT, **1**. **2.** Symbol, I_o. Current flowing in the output leg or electrode of a circuit or device. Compare INPUT CURRENT, **2**.

output device **1.** A load device, such as a resistor, loudspeaker, lamp, relay, motor, etc., that utilizes the output energy delivered by a generator, amplifier, or network. **2.** A device, such as an output transformer, that serves to transfer energy or information from a circuit or device. Compare INPUT DEVICE. **3.** In computer operations, a device that presents the results of computer operation in a comprehensible form. Examples: printer, monitor, disk drive, tape drive, modem, etc.

output efficiency The efficiency of a device, such as a generator or amplifier, in delivering an output signal. For an amplifier, the efficiency ($Eff_{\%}$) is given as a percentage by the formula $Eff_{\%} = 100P_o/P_i$, where P_i is the direct-current power input, and P_o is the alternating-current (signal) power output.

output equipment See OUTPUT DEVICE, **3**.

output filter The direct-current filter of a power supply operating from alternating current. Also see CAPACITOR-INPUT FILTER and CHOKE-INPUT FILTER.

output gap A device via which current or power is intercepted from an electron beam in a beam-power tube.

output impedance Symbol, Z_o. The impedance “looking” into the output terminals of an amplifier, generator, or network. Compare INPUT IMPEDANCE.

output indicator A device, such as an analog meter, digital meter, or bar-graph meter, that provides a visual indication of the output-signal amplitude of an equipment.

output leakage current In an open-collector integrated circuit, the current from collector to emitter with the output in the “off” condition and a certain specified voltage applied to the device. It can be expressed in milliamperes or microamperes.

output limiting A process for automatically maintaining the amplitude of the signal delivered by a generator or amplifier. See, for example, AUTOMATIC GAIN CONTROL, AUTOMATIC MODULATION CONTROL, VOLUME COMPRESSION, and VOLUME LIMITER.

output load See OUTPUT DEVICE, **1**.

output load current **1.** The current through the output load of an amplifier. Generally, this current is expressed in root-mean-square (rms) form. **2.** The highest rms current that an amplifier can deliver to a load of a specified impedance.

output meter A meter that gives a quantitative or qualitative indication of the output of an amplifier or generator. See, for example, OUTPUT-POWER METER.

output offset In an integrated circuit, the voltage at the output when the inputs are grounded.

output port The output terminal of a logic device.

output power Symbol, P_o. The power deliverable by an amplifier, generator, or circuit. Also called *power output*. Compare INPUT POWER.

output-power meter A type of direct-reading wattmeter for measuring the power output of an amplifier or generator.

output regulator A circuit or device that automatically maintains the output of a power supply or signal source at a constant amplitude.

output resistance Symbol, R_o. The internal resistance of a circuit or device, as “seen” at the output terminals. Compare INPUT RESISTANCE.

output routine In computer operations, a routine (program segment) that performs the work involved in moving data to an output device, often including intermediate transferals and modifying the data as necessary.

output section See OUTPUT AREA.

output sink current In an integrated circuit, for a specified set of conditions at the input and output, current into the output as measured in milliamperes or microamperes.

output source current In an integrated circuit, for a specified set of conditions at the input and output, the current out of the output, as measured in milliamperes or microamperes.

output stage The last stage of an amplifier. Delivers the signal to the load.

output terminals Terminals (usually a pair) associated with the output of a circuit or device (see OUTPUT, **1**, **2**). Compare INPUT TERMINALS.

output tank In a transmitter or power generator, a parallel-tuned combination of inductance and capacitance in the collector, drain, or plate circuit, that is generally tuned to resonance at the operating frequency. IT optimizes efficiency and couples the signal to the load. Compare INPUT TANK.

output transformer The output-coupling transformer that delivers signal voltage or power from an amplifier, generator, or network to a load or to another circuit. Compare INPUT TRANSFORMER.

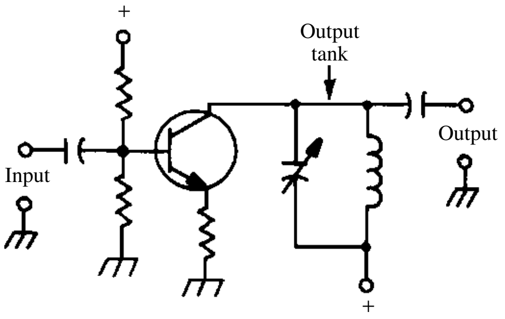

output tank

output-transformerless Abbreviation, OTL. Pertaining to an oscillator, amplifier, or generator that requires no output coupling transformer.

output transistor A transistor in the final stage of an amplifier or generator; usually, it is a power transistor.

output tube A vacuum tube in the final stage of an amplifier or generator; usually, it is a power tube.

output unit See OUTPUT DEVICE.

output voltage **1.** Symbol, E_o or V_o. The voltage delivered by a source, such as a battery, generator, or amplifier. Compare INPUT VOLTAGE, **2**. **2.** Symbol, E_o or V_o. The voltage across the output leg or electrode of a circuit or device. Compare INPUT VOLTAGE, **2**.

output voltage compliance In an integrated circuit, the voltage range over which the output can be made to swing, while keeping the operation of the circuit within a certain maximum allowable nonlinearity limit. It is measured in volts or millivolts.

output voltage noise In an integrated circuit, the output noise over a given range of frequencies, as measured in peak-to-peak millivolts or microvolts. It can also be measured as the root-mean-square (rms) value.

output voltage swing In an integrated circuit with a specified load, the output-voltage change measured as a difference between maximum and minimum in volts or millivolts.

output voltage tracking For an integrated-circuit dual regulator, the difference between the absolute values of the output voltages of a dual regulator. It can be expressed as a specific voltage or as a percentage of the specified output voltage of the device.

output winding The secondary coil of an output transformer.

outside antenna See OUTDOOR ANTENNA.

outside booster See OUTDOOR BOOSTER.

outside diameter Abbreviation, OD. The outermost diameter of a body or figure having two concentric diameters (e.g., tubing or conduit). Compare INSIDE DIAMETER.

outside lead See FINISH LEAD.

outside transformer **1.** See OUTDOOR TRANSFORMER. **2.** A transformer mounted outside of an equipment in whose circuit it is included. External mounting can eliminate hum in the equipment circuit, and can help to prevent overheating.

oven **1.** Also called *crystal oven.* A chamber providing a closely controlled operating temperature for an electronic component, such as a quartz crystal. **2.** An enclosure in which electronic equipment can be tested at selected, precise high temperatures. Compare COLD CHAMBER.

overall feedback Positive or negative feedback around an entire system (such as a public-address system), as opposed to feedback confined to one stage or a few stages within the system.

overall gain The total gain of an entire system (such as a multistage amplifier), as opposed to that of one or several stages.

overall loudness The apparent intensity of an acoustic disturbance, generally measured with respect to the threshold of hearing, and expressed in decibels, relative to the threshold level.

overbiased unit A component, such as a transistor or vacuum tube, whose bias current or voltage is higher than the correct value for a given mode of operation. Compare UNDERBIASED UNIT.

overbunching In a velocity-modulated tube, such as a Klystron, the condition in which the buncher voltage exceeds the value required for optimum bunching.

overcharging In a secondary cell or battery, the application of charging current longer than necessary to obtain full charge. This can sometimes cause problems, such as cell heating.

overcompounded generator A dynamo-type generator having a compound field winding in which the series-field winding increases the field intensity beyond the point needed to maintain the output voltage. Compare UNDERCOMPOUNDED GENERATOR.

overcompounding A characteristic of electromechanical motors, resulting in increased running speed with decreasing load resistance.

overcoupled transformer A transformer having greater than critical coupling between its primary and secondary windings. In tuned circuits, such as intermediate-frequency (IF) transformers, this produces a double-peak response.

overcoupling Extremely close coupling (see CLOSE COUPLING).

overcurrent A current greater than the specified, nominal, or desired level. Compare UNDERCURRENT.

overcurrent circuit breaker A circuit breaker that opens when current exceeds a predetermined value.

overcurrent protection The use of a circuit breaker, relay, or other device to protect a circuit or system from damage resulting from an excessive flow of current.

overcurrent relay A protective relay that opens a circuit when current exceeds a predetermined value. Compare UNDERCURRENT RELAY.

overcutting In disc recording, the condition in which an excessively high amplitude signal causes the stylus to cut through the wall between adjacent grooves. Compare UNDERCUTTING.

overdamping Damping greater than the critical value (see DAMPING ACTION, **2**). Compare UNDERDAMPING.

overdesign Also called *overengineering.* **1.** To use an unnecessarily high safety factor in the design of equipment. **2.** To design equipment for performance superior to that which is required in the intended application. **3.** A design that results from operations defined in **1** and **2**.

overdrive In an analog amplifier, an undesirable operating condition that occurs with the application of excessive input signal voltage or power. This results in increased distortion, nonlinearity, excessive harmonic generation, and excessive output signal bandwidth. In a high-fidelity audio system, such a condition grossly degrades the quality of the sound. The amplifying device (transistor or vacuum tube) is in or near saturation during part of the signal cycle. This reduces the efficiency of the circuit, can cause excessive collector, drain, or plate current, and can overheat the base–collector (B-C) junction of a bipolar transistor. In the worst case, it can destroy the component.

overdriven amplifier See OVERDRIVEN UNIT.

overdriven unit An amplifier, oscillator, or transducer whose driving signal (current, voltage, power, or other quantity) is higher than that which the device can properly or efficiently handle for correct or intended operational performance.

overdub In audio recording, a method of combining two or more signals onto a single tape track. For example, a live voice can be recorded on a tape containing pre-recorded music.

overexcited Receiving higher than normal excitation, as in radio-frequency amplifiers or alternating-current generators.

overflow **1.** In computer or calculator operation, the condition in which an arithmetic operation yields a result exceeding the capacity of the location or display for a result. **2.** The carry digit that results from the condition described in (1).

overflow indicator **1.** In a digital calculator, a display that indicates that a numerical value is too large or too small to be shown with the available number of decimal places. **2.** In data processing, a display that indicates the presence of too many bits or characters for the available storage capacity.

overflow position In a digital computer, an auxiliary register position for developing the overflow digit (see OVERFLOW, **1**, **2**).

overflow record In data processing, a record that will not fit the storage area allotted for it, and that must be kept where it can be retrieved, according to some reference stored in its place.

overflow storage In a calculator or computer, extra storage space, allowing a small amount of overflow without loss of accuracy.

overhanging turns The turns in the unused portion(s) of a tapped coil.

overhead line A power or transmission line suspended above the ground between poles or towers.

over-horizon radar A form of radar used at high frequencies, in which pulses are transmitted and received. The signals are returned to earth via the ionosphere, both in the forward and reflected directions, making it possible to detect such things as missile launchings from thousands of miles away.

over-horizon transmission See FORWARD SCATTER.

overinsulation Use of excessive insulation for a particular application. Compare UNDERINSULATION.

over insulation The insulation (usually a strip of tape) laid over a wire brought up from the center of a coil. Compare UNDER INSULATION.

overlap **1.** The time during which two successive operations are performed simultaneously. **2.** In a facsimile or television system, a condition in which the scanning line is wider than the center-to-center separation between adjacent scanning lines.

overlap radar A long-range radar situated in one sector and covering part of another sector.

overlay **1.** A sheet of transparent or translucent material laid over a schematic diagram for the purpose of tracing connections that have been made in wiring an equipment from the diagram. **2.** In computer operations, a method whereby the same internal storage locations are used for different parts of a program during a program run. It is used when the total storage requirements for instructions exceed the available main storage capacity.

overlay transistor A double-diffused epitaxial transistor having separate emitters connected together by means of diffusion and metallizing to increase the edge-to-area ratio of the emitters. This design raises the current-handling ability of the transistor. Also see DIFFUSED TRANSISTOR and EPITAXIAL TRANSISTOR.

overload **1.** Current or power drain in excess of the rated output of a circuit or device. **2.** An excessive driving signal.

overload circuit breaker See CIRCUIT BREAKER.

overloaded amplifier A power amplifier delivering excessive output power. Compare UNDERLOADED AMPLIFIER, **2**.

overloaded oscillator An oscillator from which excessive power is drawn, causing instability, frequency shift, lowered output voltage, and overheating.

overload indication Any attention-catching method, such as an audible or visual alarm, for warning that a prescribed signal or power level has been exceeded.

overload level The amount of overload that can safely be applied to an equipment (see OVERLOAD, **1**).

overload protection The use of circuit breakers, relays, automatic limiters, and similar devices to protect equipment from overload damage by reducing current or voltage, disconnecting the power supply, or both.

overload recovery time Following an overdrive at the input of an integrated-circuit device, the time required for the output to resume its normal characteristics.

overload relay A relay actuated when circuit current exceeds a predetermined value. Compare UNDERLOAD RELAY.

overload time The maximum length of time that an equipment can safely be subjected to an overload level of current.

overmodulation Modulation in excess of a prescribed level—especially amplitude modulation greater than 100%. Compare COMPLETE MODULATION and UNDERMODULATION.

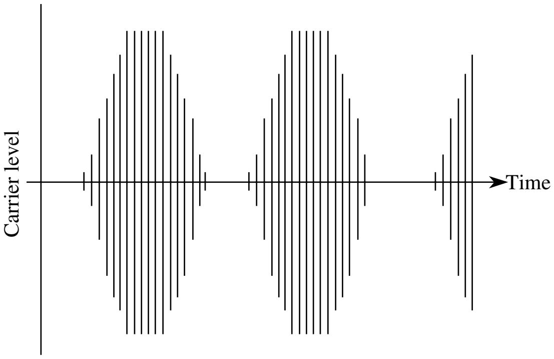

overmodulation

overmodulation alarm See OVERMODULATION INDICATOR.

overmodulation indicator A device, such as a neon bulb, incandescent lamp, light-emitting diode, analog meter, or digital meter, adapted to give an alarm when the modulation percentage of a signal exceeds a predetermined value.

overpotential See OVERVOLTAGE.

overpower relay A relay actuated by a rise in power above a predetermined level. Compare UNDERPOWER RELAY.

overpressure For a pressure transducer, pressure in excess of the maximum rating of the device.

override **1.** To intentionally circumvent an automatic control system. **2.** To bridge a functional stage of a system.

overscanning The deflection of the beam of a cathode-ray tube beyond the edges of the screen.

overshoot **1.** The momentary increase of a quantity beyond its normal maximum value (e.g., the spike sometimes seen on a square wave because of the overswing of a rising voltage). **2.** Momentary overtravel of the pointer of an analog meter.

overswing See OVERSHOOT, **2**.

overtemperature protection The use of an automatic device, such as a thermal relay or thermostat, to disconnect a device from the power supply when the device's temperature becomes excessive.

overthrow See OVERSHOOT, **2**.

overtone See HARMONIC.

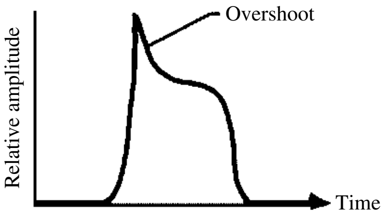

overshoot

overtone crystal A piezoelectric quartz crystal that oscillates at odd multiples of the frequency for which it was cut. This allows crystal operation at frequencies otherwise obtainable only from a fundamental-frequency crystal ground so thin as to be prohibitively fragile.

overtone oscillator A crystal oscillator using an OVERTONE CRYSTAL.

overtravel See OVERSHOOT, **2**.

overvoltage A voltage higher than a specified or rated value. Compare UNDERVOLTAGE.

overvoltage circuit breaker A circuit breaker that opens when voltage exceeds a predetermined value.

overvoltage protection The use of a special circuit or device to protect equipment from excessive voltage. When voltage increases beyond the overvoltage limit, the protective circuit causes shutdown.

overvoltage relay A relay actuated when voltage rises above a predetermined value. Compare UNDERVOLTAGE RELAY.

overwrite In computer operations, to record new data over existing data (e.g., to update the files on a magnetic disk or tape).

Ovshinsky effect In thin-film solid-state devices, the tendency for switches to have the same characteristics for currents in either direction.

O wave One (the ordinary) of the pair of components into which an ionospheric radio wave is divided by Earth's magnetic field. Compare X WAVE.

Owen bridge A wide-range four-arm bridge that measures inductance in terms of a standard capacitance and bridge-arm resistances.

own coding Additional program steps added to vendor-supplied software so that it can be modified to fit special needs.

ox Abbreviation of OXYGEN.

oxidation **1.** The combination of a substance with oxygen. Generally a slow process, such as the corrosion of iron or aluminum in the atmosphere. The process is accelerated by the presence of moisture and/or high temperatures. **2.** The loss of electrons from a cell or battery during discharge.

oxidation-reduction potential The potential at which oxidation occurs at the anode of an electrolytic cell, and at which reduction occurs at the cathode.

oxide-coated cathode See OXIDE-COATED EMITTER.

oxide-coated emitter An electron-tube cathode or filament coated with a material, such as thorium oxide, for increased electron emission at low emitter temperatures.

oxide-coated filament See OXIDE-COATED EMITTER.

oxide film **1.** The thin film of iron oxide that constitutes the recording surface of a magnetic disk or tape. **2.** The layer of copper oxide formed on the copper plate of a copper-oxide rectifier.

oxide-film capacitor An electrolytic capacitor, so called because the dielectric is a thin oxide film.

oxide rectifier A solid-state rectifier using a junction between copper and copper oxide. Also called COPPER-OXIDE RECTIFIER.

oximeter A photoelectric instrument for measuring the oxygen content of the blood. It operates by passing visible light through the earlobe, and analyzing the color and intensity of the emerging beam. Also called ANOXEMIA TOXIMETER.

oxygen Symbol, O. Abbreviation, O_2. A gaseous element. Atomic number, 8. Atomic weight, 15.999. Constitutes 21% of Earth's atmosphere. It readily combines with various elements to form compounds (see OXIDATION).

oxygen analyzer An electronic gas analyzer designed especially to measure oxygen content. The operation of this instrument is based on the paramagnetic properties of oxygen.

oxygen recombination In nickel-cadmium (NICAD) cells and batteries, a process in which oxygen is generated in the vicinity of the positive electrode, and is reduced with water in the vicinity of the negative electrode. This produces battery heating.

oz **1.** Abbreviation of OUNCE. **2.** Abbreviation of OZONE.

oz-in Abbreviation of OUNCE-INCH.

ozocerite An insulating mineral wax. Dielectric constant, 2.2. Dielectric strength, 4 to 6 kV/mm. Also spelled OZOKERITE.

ozone Symbol, O_3. An allotropic form of oxygen. Its formula indicates that each molecule has three atoms. Produced by the action of ultraviolet rays (or electrical discharge) on oxygen, its characteristic odor (somewhat like weak chlorine) can often be detected around sparking contacts or in the air after a thunderstorm.

ozone layer In the earth's atmosphere, a layer of ozone gas in the upper troposphere and lower stratosphere. It is produced by ultraviolet radiation from space, mainly from the sun. The ozone layer tends to block ultraviolet radiation, reducing the amount that reaches the surface of the earth.

ozone monitor An instrument for measuring the concentration of ozone in the atmosphere. One version measures the extent to which ultraviolet radiation is absorbed by a sample of air; the greater the absorption, the higher the ozone concentration in the sample.

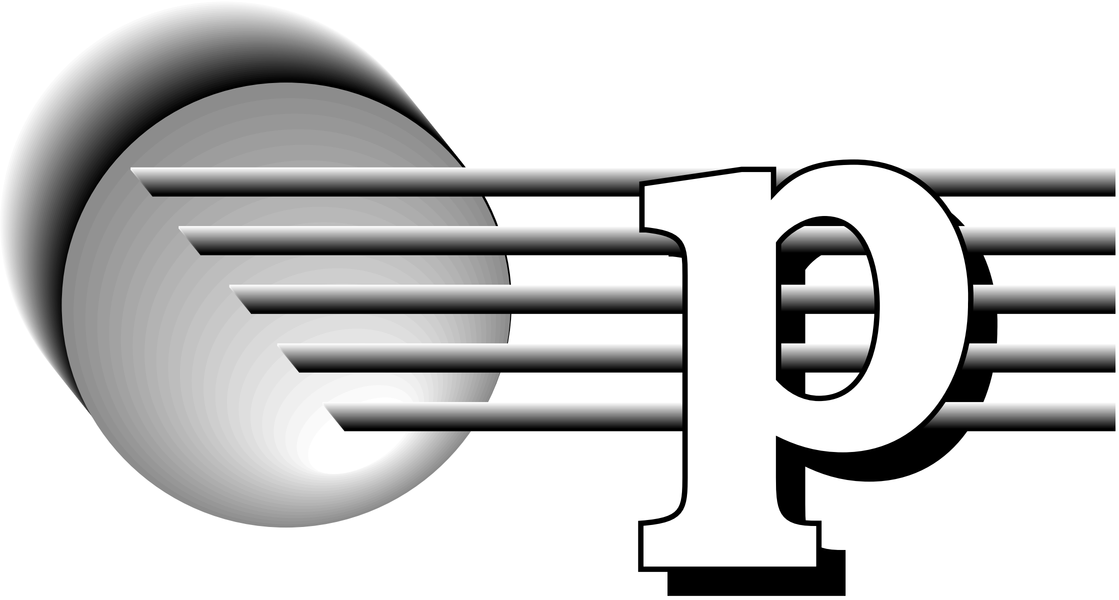

P **1.** Symbol for POWER. **2.** Symbol for PLATE (of a vacuum tube). **3.** Symbol for PHOSPHORUS. **4.** Abbreviation of PRESSURE. **5.** Symbol for PRIMARY. **6.** Abbreviation for prefix PETA-. **7.** Symbol for PERMEANCE. **8.** Abbreviation of POINT.

p **1.** Abbreviation of prefix PICO-. **2.** Subscript for PEAK. **3.** Abbreviation of POUND. **4.** Abbreviation of POINT (often capitalized). **5.** Subscript for PRIMARY. **6.** Subscript for PLATE (of a vacuum tube). **7.** Abbreviation of PITCH. **8.** Abbreviation of PER.

PA **1.** Abbreviation of POWER AMPLIFIER. **2.** Abbreviation of PULSE AMPLIFIER. **3.** Abbreviation of *particular average.* **4.** Abbreviation of *pilotless aircraft.* (Also, P/A.) **5.** Abbreviation of PUBLIC ADDRESS (as in PA system).

Pa **1.** Symbol for PROTACTINIUM. **2.** Symbol for PASCAL.

pA Abbreviation of PICOAMPERE.

pacemaker See CARDIAC STIMULATOR.

pacer See CARDIAC STIMULATOR.

Pacific Standard Time Abbreviation, PST. Local mean time at the 120th meridian west of Greenwich. Also see GREENWICH MEAN TIME, STANDARD TIME, TIME ZONE, and COORDINATED UNIVERSAL TIME.

pack A technique for maximizing a computer memory device's storage capacity, wherein more than one information item is stored in a single storage unit. Also called *crowd.*

package **1.** The enclosure for an electronic device or system. This includes a wide range of housings, from the simple encapsulation of miniature transistors to forced-air-cooled enclosures for heavy power units. **2.** To assemble and house an electronic equipment, or to design a housing for it, in accordance with good engineering techniques. **3.** A computer program of general use for an application (e.g., *payroll package*).

package count The number of discrete packaged circuits in a system.

packaging density **1.** See VOLUMETRIC EFFICIENCY. **2.** Computer storage capacity in terms of the number of information units that can be contained on a given segment of a magnetic medium. Also called PACKING DENSITY. **3.** Within a given integrated circuit, the capacity in terms of the number of active devices that can be contained on a single silicon chip.

packet **1.** A unit of digital information in PACKET COMMUNICATIONS. It consists of a header followed by a certain number of data bits or bytes. **2.** See WAVE PACKET. **3.** See PACKET COMMUNICATIONS. **4.** See PACKET RADIO.

packet communications A method via which data is exchanged through a network between or among people or computers. Information is sent and received in blocks of information called *packets.* Each packet is routed individually through the network according to the most efficient possible path at the time of its transit. At the destination, the packets are reassembled into the original signal. This scheme makes more efficient use of network resources than continuous-connection or single-path methods. However, when network usage is heavy, there can be a delay in the arrival of a sufficient number of packets to produce an intelligible received signal.

packet radio The transmission and reception of PACKET COMMUNICATIONS data via radio.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

packet switching In telephony, a method of connection in which data is exchanged between subscribers by splitting the data into units (packets). Each packet is sent over the optimum path at the time of transmission. The signal path can, and usually does, vary from packet to packet. At the destination, the packets are reassembled into the original singal. The connection is in effect nonexistent during periods of silence (no data transmitted by either subscriber). Compare CIRCUIT SWITCHING.

packing In the button of a carbon microphone, bunching and cohesion between the carbon granules.

packing density The number of discrete package circuits within a given surface area or volume.

packing factor **1.** See VOLUMETRIC EFFICIENCY. **2.** In computer operations, the number of bits that can be recorded in a given length of magnetic memory surface. Also called PACKING DENSITY.

pack transmitter A portable transmitter that can be strapped to the operator's back.

pack unit A portable transceiver that can be strapped to the operator's back or carried on an animal's back.

PACM Abbreviation of *pulse-amplitude code modulation.*

pad **1.** An attenuator network (usually a combination of resistors) that reduces the amplitude of a signal by a desired amount while maintaining constant input and output impedance. **2.** In computer operations, to make a record a fixed size by adding blanks or dummy characters to it. **3.** To lower the frequency of an inductance-capacitance (LC) circuit by adding capacitance to an already capacitively tuned network.

padder See OSCILLATOR PADDER.

padding capacitor See OSCILLATOR PADDER.

padding character In a digital communications system, a character that is inserted solely for the purpose of consuming time while no meaningful characters are sent. The insertion of such characters maintains the synchronization of the system.

paddle-handle switch A toggle switch the lever of which is a flattened rod. Compare BAT-HANDLE SWITCH, ROCKER SWITCH, and SLIDE SWITCH.

PADT Abbreviation of POST-ALLOY-DIFFUSED TRANSISTOR.

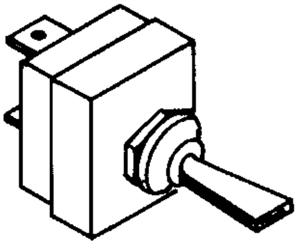

paddle-handle switch

page A display of text data on a computer display that completely fills the screen.

page printer A computer peripheral that prints a message in lines on a page, according to an established format, rather than in a single line.

pager **1.** A public-address system used for summoning purposes. **2.** See BEEPER, **2**.

page turning The successive display of pages (see PAGE).

pair **1.** Two wires, especially two insulated conductors in a cable. **2.** A set of two particles or charge carriers (e.g., *electron-hole pair*). **3.** A set of two transistors or vacuum tubes, operating together in push-pull or parallel in a power amplifier.

paired cable A cable consisting of separate twisted pairs of conducting wires.

paleomagnetism The study of certain rocks and minerals to determine the nature of the earth's magnetic field at the time the rocks were formed. When the age of the rock is determined by means of radioactive dating, and numerous rock samples are found covering many different eras, the nature of the earth's magnetic field can be graphed over time.

palladium Symbol, Pd. A metallic element of the platinum group. Atomic number, 46. Atomic weight, 106.42.

palletizing In industrial robots, the automatic placing of objects in a tray according to a computer program.

Palm See HANDHELD COMPUTER.

Palmer scan In radar, a method of simultaneously scanning the azimuth and the elevation.

PalmPilot See HANDHELD COMPUTER.

palmtop computer See HANDHELD COMPUTER.

PAM Abbreviation of PULSE-AMPLITUDE MODULATION.

Pan In radiotelephony, a spoken word indicating that an urgent message is to follow. It is equivalent to the XXX of radiotelegraphy.

pan **1.** To make a panoramic sweep [e.g., to sweep a wide area with a beam (as from an antenna), or to sweep a wide band of frequencies with a suitable tuning circuit]. **2.** A panoramic sweep made as defined in **1**. **3.** In audio engineering, to gradually shift from one audio channel to another or from one reproducer to another.

pan and tilt **1.** An azimuth-elevation mounting for a television camera. **2.** The simultaneous movement of a television camera in the vertical and horizontal directions.

pancake coil See DISK WINDING.

panel A flat surface on which are mounted the controls and indicators of an equipment, for easy access to the operator.

panel lamp **1.** See ELECTROLUMINESCENT PANEL. **2.** See PANEL LIGHT.

panel light A pilot light for illuminating the front panel of a piece of equipment.

panel meter A usually small meter for mounting on, or through an opening in, a panel.

panic button In a security system, a button or switch that immediately triggers an alarm when it is closed.

panoramic adapter An external device that can be connected to a receiver to sweep a frequency band and indicate carriers on the air as pips on a screen at the corresponding frequency points. Also called *pan adapter.*

panoramic display **1.** A wide-angle display. **2.** A spectrum-analyzer display that shows a wide range of frequencies, from zero to well above the maximum frequency in the monitored system.

panoramic radar An omnidirectional radar (i.e., one that transmits wide-beam signals in all directions without scanning).

panoramic receiver A receiver that displays pips on a screen to show carriers on the air in a given frequency band. All frequencies in the band are presented along the horizontal axis of the screen.

panpot A potentiometer with which panning can be achieved (see PAN, **3**).

pan-range A form of radar display in which target motion can be ascertained.

pantography The transmission of radar information to a distant location for observation or recording.

Papa Phonetic alphabet code word for the letter P.

paper advance mechanism In a data-processing system, the part of a printer that moves (sometimes by computer control) the paper through the printer.

paper capacitor A component that is made by placing paper, soaked with mineral oil, between two strips of foil. The assembly is rolled up, and wire leads are attached to the two pieces of foil. Finally, the rolled-up foil and paper are enclosed in a cylindrical case. These components are sometimes found in radio-frequency (RF) electronic equipment. They have values ranging from about 0.001 microfarads (μF) to 0.1 μF, and can handle low to moderate voltages, usually up to about 1000 volts. Compare CERAMIC CAPACITOR, ELECTROLYTIC CAPACITOR, MICA CAPACITOR, PLASTIC-FILM CAPACITOR, TANTALUM CAPACITOR.

PAR Abbreviation of PRECISION APPROACH RADAR.

par Abbreviation of PARALLEL.

parabola A plane curve that is the locus of points that are equidistant from a fixed point (the focus) and a fixed straight line (the directrix). In the Cartesian *xy*-plane, the general equation is $y = ax2 + b$, where a and b are constants.

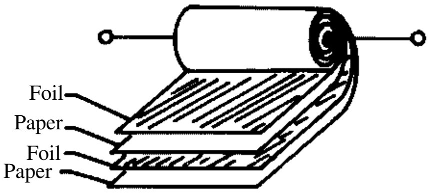

paper capacitor

parabola control See VERTICAL-AMPLITUDE CONTROL, **2**.

parabola generator A circuit for generating a parabolic-waveform signal.

parabolic microphone A directional microphone mounted at the principal focus of a parabolic sound reflector; the front of the microphone faces the inside of the parabola. It is useful for detecting sounds from great distances.

parabolic reflector Also called *paraboloidal reflector.* A reflector having the shape of a paraboloid. It is particularly useful for focusing or directing radiation. For example, if a radiator, such as an antenna rod, is placed at the focus of the paraboloid, a beam of parallel rays will be emitted by the reflector.

paraboloid The surface generated by a PARABOLA rotated about its axis of symmetry.

paraffin A relatively inexpensive, easily available, solid, white petroleum wax. At one time, it was used to impregnate capacitors and coils and to waterproof paper used for insulating purposes.

parallax The apparent shift in the position of a relatively nearby object when the observer moves or alternately blinks either eye. Thus, a pointer-type meter will seem to give different readings when viewed from different angles. Some meters have mirrored scales to eliminate this effect.

parallel **1.** Pertaining to the type of operation in a computer when all elements in an information item (e.g., bits in a word) are acted upon simultaneously, rather than serially (one at a time). **2.** The condition in which two comparably sized objects or figures are equidistant at all facing points. **3.** Pertaining to the shunt connection of components or circuits.

parallel access In computer operations, inputting or outputting data to or from storage in whole elements of information items (a word, rather than a bit at a time, for example).

parallel adder In a computer or calculator, an adder in which corresponding digits in multibit numbers are added simultaneously. Also see PARALLEL, **1**.

parallel antenna tuning Antenna-feeder tuning in which the tuning capacitor is connected in parallel with the two feeder wires. Compare SERIES ANTENNA TUNING.

parallel arithmetic unit See PARALLEL ADDER.

parallel capacitance **1.** A capacitance connected in parallel (shunt) with some other component. **2.** The capacitance between the turns of a coil. Also see DISTRIBUTED CAPACITANCE.

parallel capacitors Two or more capacitors connected in parallel (shunt) with each other. The total capacitance is equal to the sum of the individual capacitances. Also see PARALLEL CIRCUIT.

parallel circuit A circuit in which the components are connected across each other (i.e., so that the circuit segment could be drawn showing component leads bridging common conductors as rungs would across a ladder). Compare SERIES CIRCUIT.

parallel-component amplifier An amplifier stage in which the active devices (transistors or vacuum tubes) are connected in parallel with each other for increased power output. Also see PARALLEL CIRCUIT.

parallel-component oscillator An oscillator stage in which transistors are connected in parallel with each other for increased power output. Also see PARALLEL CIRCUIT.

parallel computer A computer equipped to handle more than one program at a time, but not through the use of multiple programming or time-sharing.

parallel-cut crystal See Y-CUT CRYSTAL.

parallel-diode half-wave rectifier See PARALLEL LIMITER.

parallel-fed amplifier An amplifier circuit in which the direct-current operating voltage is applied in parallel with the alternating-current output voltage. Also see PARALLEL FEED.

parallel-fed oscillator An oscillator circuit in which the direct-current operating voltage is applied in parallel with the alternating-current output voltage. Also see PARALLEL FEED.

parallel feed **1.** The presentation of parallel alternating-current (ac) and direct-current (dc) voltages to a device. **2.** The presentation of a dc operating voltage in parallel with the ac output voltage of a device (as in a parallel-fed amplifier or oscillator). Also see SHUNT FEED.

parallel gap welding A welding technique using two electrodes separated by a gap.

parallel gate circuit **1.** A gate circuit using two bipolar transistors with parallel-connected collectors and emitters, and a common collector resistor. The input signal is applied to one base, and the control signal to the other. **2.** A gate circuit using two field-effect transistors with parallel-connected drains and sources, and a common drain resistor. The input signal is applied to one gate, and the control signal to the other.

parallel inductance An inductance connected in parallel (shunt) with some other component.

parallel inductors Inductors connected in parallel and separated or oriented to minimize the effects of mutual inductance. Also see PARALLEL CIRCUIT.

parallel inverse feedback In a single-ended audio amplifier circuit, a simple system for obtaining negative feedback: A high resistance is connected from the output-transistor collector or drain to the driver-transistor collector or drain.

parallel limiter A limiter (clipper) circuit in which the diode is in parallel with the signal. Compare SERIES LIMITER.

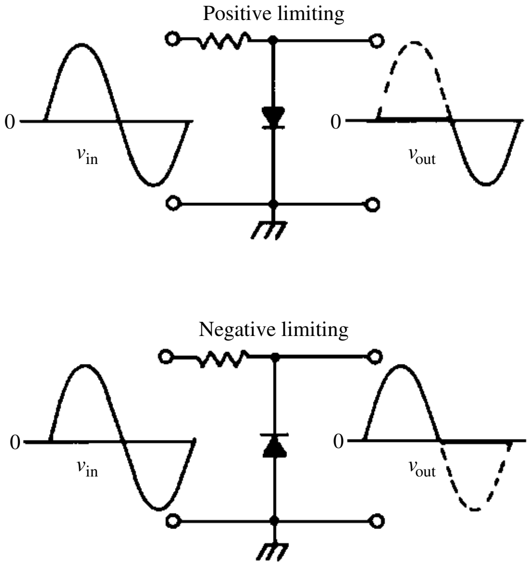

parallel limiter

parallel-line tuning At ultra-high frequencies (UHF) and microwave frequencies, the use of two parallel wires or rods for tuning. A straight short-circuiting bar is slid along the wires to accomplish tuning.

parallelogram A two-dimensional geometric figure that has four sides. Opposite pairs of sides are parallel. Opposite interior angles have equal measure.

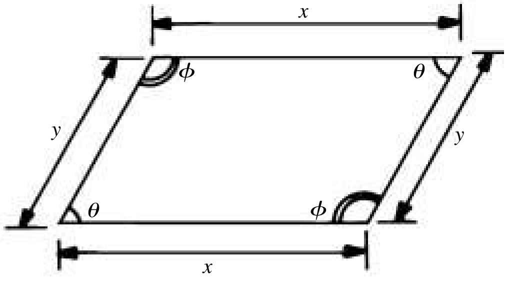

parallelogram

parallelogram of vectors A graphic device for finding the sum of two vectors. A parallelogram is constructed for which the two vectors are adjacent sides. The sum of the vectors is represented by the diagonal of the parallelogram.

parallel operation In computer operations, the simultaneous transmission of all bits in a multibit word over individual lines, as compared with the serial transmission of a word bit by bit.

parallel output A digital output consisting of two or more lines, all of which carry data at the same time.

parallel processing In computer operations, the simultaneous processing of several different programs through separate channels. Compare SERIAL PROCESSING.

parallel *Q* Symbol, Q_p. The figure of merit of a parallel circuit of inductance, capacitance, and resistance.

parallel resistance **1.** A resistance connected in parallel (shunt) with some other component. **2.** The resistance between the plates of a capacitor. **3.** The resistance across a coil.

parallel resistors Resistors connected in parallel. If the individual resistances are represented by R_1, R_2, R_3, . . ., Rn, then total resistance R_t is equal to $1/(1/R_1 + 1/R_2 + 1/R_3 + . . . + 1/R_n)$. Also see PARALLEL CIRCUIT.

parallel resonance Resonance in a circuit consisting of a capacitor, inductor, and alternating-current source connected in parallel. At the resonant frequency, the inductive reactance is equal in magnitude, but opposite in effect, to the capacitive reactance. The capacitor current and inductor current are maximum, the line current is minimum, and the circuit impedance is maximum. Compare SERIES RESONANCE.

parallel-resonant circuit A resonant circuit in which the capacitor, inductor, and alternating-current source are connected in parallel. Compare SERIES-RESONANT CIRCUIT.

parallel-resonant trap A wavetrap consisting of a parallel-resonant inductance-capacitance (LC) circuit. Compare SERIES-RESONANT TRAP.

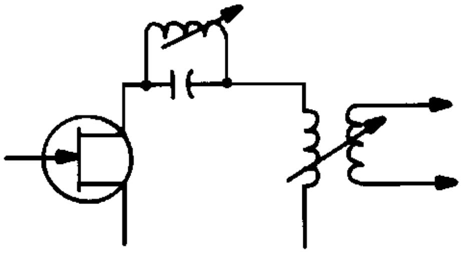

parallel-resonant trap

parallel-resonant wavetrap See PARALLEL-RESONANT TRAP.

parallel-rod oscillator An ultra-high-frequency (UHF) oscillator tuned by means of two straight, parallel quarter- or half-wave rods, one rod connected to the base or gate of a transistor, and the other rod connected to the collector or drain.

parallel-rod tuning Adjustment of the resonant frequency of a section of open-wire transmission line. A movable shorting bar allows quarter-wave resonance. The impedance at resonance is very high.

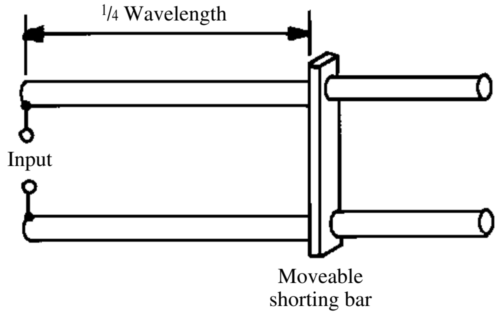

parallel-rod tuning

parallel-series Also called *series-parallel.* Pertaining to an arrangement of components, usually similar (e.g., resistances), consisting of parallel circuits connected in series with each other, or of series circuits connected in parallel with each other. Usually, the component values or ratings are all identical, so currents and/or voltages are equally shared among them. Also see PARALLEL CIRCUIT and SERIES CIRCUIT.

parallel-series capacitors Capacitors connected in PARALLEL-SERIES, usually to obtain higher voltage and/or current ratings than an individual capacitor can provide.

parallel-series inductors Inductors connected in PARALLEL-SERIES and separated or oriented to minimize the effects of mutual inductance.

parallel-series resistors Resistors connected in PARALLEL-SERIES, usually to obtain a higher power rating than an individual resistor can provide.

parallel storage In a computer, storage in which all information items can be made available in the same amount of time.

parallel-tee amplifier A bandpass amplifier having a parallel-tee network in its negative-feedback path. The null frequency of the network determines the pass frequency of the amplifier.

parallel-tee measuring circuit A parallel-tee network used for measuring circuit constants. Also called TWIN-TEE MEASURING CIRCUIT.

parallel-tee network A resistance-capacitance (RC) network containing two tee sections (with R and C elements opposite in the tees) connected in parallel. The network produces a null at one frequency. Also called TWIN-T NETWORK.

parallel-tee oscillator A resistance-capacitance tuned oscillator having a parallel-tee network in its negative-feedback path. The null frequency of the network determines the oscillator frequency.

parallel transfer A form of digital information transfer, consisting of two or more lines that carry data at the same time.

parallel-wire line A transmission line consisting of two parallel wires whose separation is kept constant by dielectric rods (open-wire line) or a solid dielectric web (ribbon line).

parallel-wire tank In an ultra-high-frequency (UHF) amplifier or oscillator, a resonant circuit consisting of two separate parallel wires connected to the transistor(s) or tube(s) at one end, and short-circuited or tuned at the other end.

paramagnet A paramagnetic substance (see PARAMAGNETISM). Compare DIAMAGNET.

paramagnetic Possessing PARAMAGNETISM. Compare DIAMAGNETISM.

paramagnetism The state of having a magnetic permeability slightly greater than **1**. Compare DIAMAGNETISM.

parameter **1.** An operating value, constant, or coefficient that can be either a dependent or an independent variable (e.g., a transistor-electrode current or voltage). **2.** The ratio of one coefficient to another, where both are either fixed or variable (e.g., transconductance of a vacuum tube).

parameter word In a computer memory, a place having a capacity of a word (bit group) in which is stored a parameter for a program.

parametric amplifier A radio-frequency power amplifier based on the action of a voltage-variable capacitor in a tuned circuit.

parametric amplifier diode See VARACTOR.

parametric converter A frequency converter in which a parametric device, such as a varactor, is used to change a signal of one frequency to a signal of another frequency. Also see PARAMETRIC DOWN-CONVERTER and PARAMETRIC UP-CONVERTER.

parametric diode A variable-capacitance diode (see VOLTAGE-VARIABLE CAPACITOR, **1**).

parametric down-converter A parametric converter in which the output signal is of a lower frequency than the input signal. Compare PARAMETRIC UP-CONVERTER.

parametric equalizer A set of audio filters similar to a GRAPHIC EQUALIZER, except that the center frequencies are adjustable, rather than fixed. The center frequencies are selected by the operator; then the attenuation level (in decibels) is set for each frequency. It is used in audio recording studios.

parametric modulation Modulation in which either the inductance or capacitance of a tank circuit or coupling device is varied at the modulation frequency.

parametric oscillator An oscillator that generates visible light energy by means of a parametric amplifier and a tunable cavity.

parametric up-converter A parametric converter in which the output signal is of a higher frequency than the input signal. Compare PARAMETRIC DOWN-CONVERTER.

parametron See PHASE-LOCKED OSCILLATOR.

paramistor A device consisting of several digital circuit elements that use parametric oscillators.

paramp Abbreviation of PARAMETRIC AMPLIFIER.

paraphase inverter A single-transistor phase inverter in which the two out-of-phase output signals are obtained by taking one output from the collector or drain, and the other output from the emitter or source. Thus, the 180-degree phase difference between collector/drain and emitter/source is exploited.

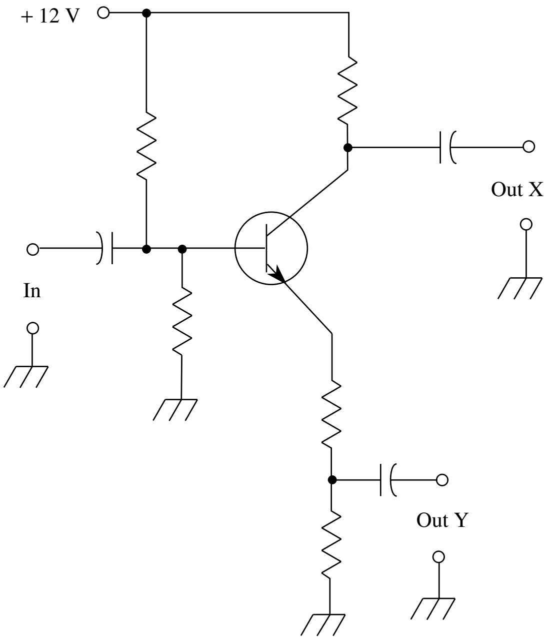

paraphase inverter

parasitic See PARASITIC OSCILLATION.

parasitic antenna See PARASITIC ARRAY, PARASITIC ELEMENT.

parasitic array Any of several types of directional antenna employing parasitic elements, sometimes in combination with phasing, to obtain directivity and gain. Common examples include the Yagi and quad. Arrays with numerous parasitic elements (usually directors) can produce upwards of 15 dBd forward gain. When several high-gain arrays are phased, the realizable gain becomes greater still. Two-element, three-element, and four-element arrays are common below 30 MHz. At very-high and ultra-high frequencies, especially above 100 MHz, bays of antennas of this type are used in satellite communications and radio astronomy. See PARASITIC ELEMENT, QUAD ANTENNA, YAGI ANTENNA.

parasitic capacitance Stray capacitance. It can be internal or external to a circuit and can introduce undesirable coupling or bypassing.

parasitic choke A small radio-frequency choke coil (with or without a shunting resistor) that suppresses or eliminates parasitic oscillation in a power amplifier.

parasitic director In a multielement directional antenna, a parasitic element acting as a director; usually, it is a few percent shorter than the driven element.

parasitic element An electrical conductor that comprises an important part of an antenna system, but that is not directly connected to the feed line. Such elements are used for the purpose of obtaining directivity and power gain. They operate via electromagnetic coupling to the driven element(s). The principle of operation was first discovered by the Japanese engineers Yagi and Uda, who observed that antenna elements parallel to a driven element but not connected to anything, at a specific distance from the driven element, and having a certain length, cause the radiation pattern to show gain in one direction and loss in the opposite direction. See DRIVEN ELEMENT, PARASITIC ARRAY, PARASITIC DIRECTOR, PARASITIC REFLECTOR.

parasitic-element directive antenna See PARASITIC ARRAY.

parasitic eliminator See PARASITIC SUPPRESSOR.

parasitic excitation Excitation of a beam-antenna element without a direct connection to the transmitter. Thus, a director or reflector element can be excited by the field of the radiator element.

parasitic inductance Stray inductance (e.g., the internal inductance of a wirewound resistor).

parasitic oscillation Extraneous, useless oscillation present as a fault in an electronic circuit, particularly a radio-frequency power amplifier.

parasitic reflector In a multielement beam antenna, a parasitic element acting as a reflector; usually, it is a few percent longer than the driven element.

parasitic resistance Stray resistance (e.g., the inherent, internal resistance of a multilayer coil).

parasitic suppressor A small resistor, coil, or parallel combination of the two, connected in series with the plate or collector of a vacuum tube or transistor to eliminate parasitic oscillations in a radio-frequency power amplifier.

PARD Abbreviation of PERIODIC AND RANDOM DEVIATION.

parity **1.** At par (with respect to the even-or-odd state of the characters in a group). **2.** Having the quality that the number of bits (or the number of similar bits) are even or odd, as intended.

parity bit **1.** In computer operations, a logic 1 added to a group of bits so that the number of 1s in the group is, according to specification, even or odd. **2.** In computer operations, a check bit that can be a logic 1 or 0, depending on the parity (see PARITY, **1**) of the total of 1s in the bit group being checked.

parity check A check of the integrity of data being transferred by adding the bits in, for example, a word, and then determining the parity bit needed and comparing that with the transmitted parity bit.

parity error An error disclosed by a parity check.

parity tree A digital device used to check parity.

parsec Abbreviation, pc. The distance at which the mean radius of the earth's orbit around the sun subtends an angle of 1 second of arc; 1 pc = 3.0857×10^{13} kilometers or 3.2616 light years.

part See CIRCUIT COMPONENT, **1**.

part failure The usually destructive breakdown of a circuit component.

partial One of the frequencies in a complex musical tone. It might be a harmonic of the fundamental frequency, although this is not always the case.

partial carry The temporary storage of some or all of the carry information in a digital calculation.

particle **1.** A tiny, discrete bit of matter. **2.** A unit of matter smaller and lighter than an atom. See, for example, ANTIPARTICLE, ELECTRON, MESON, NEUTRETTO, NEUTRINO, NEUTRON, NUCLEON, POSITRON, and PROTON.

particle accelerator See ACCELERATOR, **1**.

particle theory of radiation In physics, a model that explains the nature of electromagnetic radiation (radio waves, infrared, visible light, ultraviolet, X rays, and gamma rays) in terms of discrete particles. Each particle, called a PHOTON, carries a certain amount of energy that depends on the wavelength of the radiation.

particle velocity **1.** The speed and direction of the particles from a source of atomic radiation. **2.** The speed and direction of the molecules in the medium of an acoustic disturbance.

partitioning In computer operations, breaking down a large block of data into smaller blocks that can be better handled by the machine.

parton model A model for atomic nuclei, in which protons and neutrons are made up of smaller particles called *partons.* Subparticles have been found, commonly called *quarks.*

Pascal A high-level computer programming language, similar to BASIC or FORTRAN in structure. It is used in some schools to teach computer programming.

pascal Symbol, Pa. The SI (derived) unit of pressure; $1\ Pa = 1\ \mathrm{N/m^2} = 1.4503 \times 10^{-4}\ \mathrm{lb/in^2}$.

Paschen-Back effect See ZEEMAN EFFECT.

Paschen's law For a two-element, parallel-plate, gas-discharge tube, the plate-to-plate sparking potential is proportional to Pd, where P is the gas pressure, and d is the distance between plates.

pass amplifier A tuned amplifier having the response of a bandpass filter. Like the filter, the amplifier passes one frequency (or a narrow band of frequencies) readily while rejecting or attenuating others. Compare REJECT AMPLIFIER.

passband The continuous spectrum of frequencies transmitted by a filter, amplifier, or similar device. Compare STOPBAND.

passband ripple Multiple low-amplitude attenuation variations within the passband of a filter or tuner, resulting in a ripple pattern on the nose of the response curve.

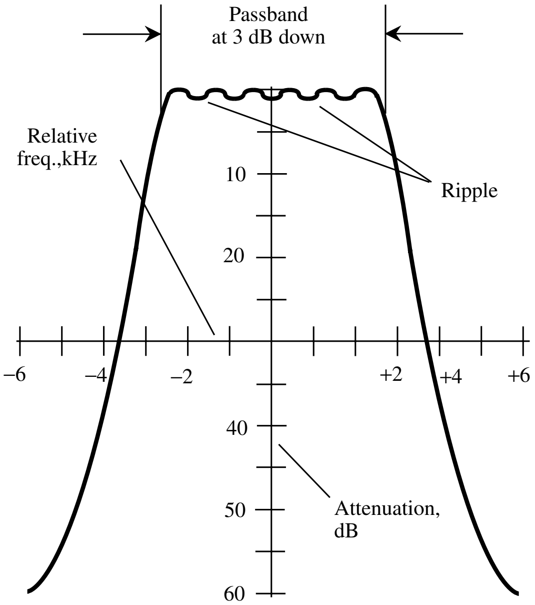

passband

passband ripple

passivation The process of growing a thin oxide film on the surface of a planar semiconductor device to protect the exposed junction(s) from contamination and short circuits. See, for example, PLANAR EPITAXIAL PASSIVATED TRANSISTOR and PLANAR TRANSISTOR.

passive absorber A substance that reflects minimal sound energy. Examples include acoustical ceiling tile and thick carpeting.

passive circuit A circuit consisting entirely of nonamplifying components, such as capacitors, resistors, inductors, and diodes.

passive communications satellite A communications satellite that reflects electromagnetic waves, but does not contain a transponder; that is, it does not receive and retransmit the signals. Also called *passive comsat.* Compare ACTIVE COMMUNICATIONS SATELLITE.

passive component A device that is basically static in operation (i.e., it is ordinarily incapable of amplification or oscillation and usually requires no power for its characteristic operation). Examples: conventional resistor, capacitor, inductor, diode, rectifier, and fuse. Compare ACTIVE COMPONENT, **1**.

passive comsat See PASSIVE COMMUNICATIONS SATELLITE.

passive decoder A decoder that responds to only one signal code, rejecting all others.

passive detection In reconnaissance, detecting a target without betraying the location of the detector.

passive electric network See PASSIVE NETWORK.

passive frequency multiplier A frequency multiplier that does not require a power supply, but operates only from the input signal energy. Usually consists of one or more semiconductor diodes, sometimes in conjunction with inductors and/or capacitors. The output signals appear at integral multiples of the input frequency.

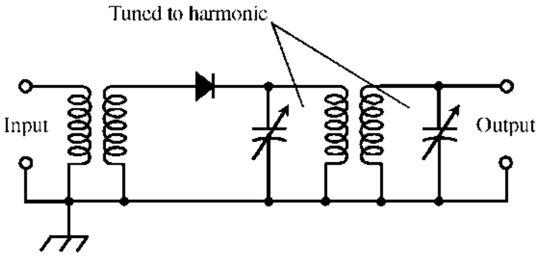

passive frequency multiplier

passive infrared sensor A device that detects infrared directly, such as that given off by humans because of their body heat. It does not generate energy of any kind. It is used in some intrusion detection systems.

passive mixer A signal mixer using only passive components (diodes, nonlinear resistors, and nonlinear reactances) (i.e., one without active components, such as transistors). Passive mixers introduce some loss. Compare ACTIVE MIXER.

passive modulator A modulator using only passive components (diodes, nonlinear resistors, and nonlinear reactances) (i.e., one without active components, such as transistors). Passive modulators introduce some loss. Compare ACTIVE MODULATOR.

passive network A network composed entirely of passive components (i.e., one containing no generators and providing no amplification).

passive radiator See DRONE CONE.

passive reflector A metal surface used to reflect electromagnetic energy at ultra-high and microwave frequencies.

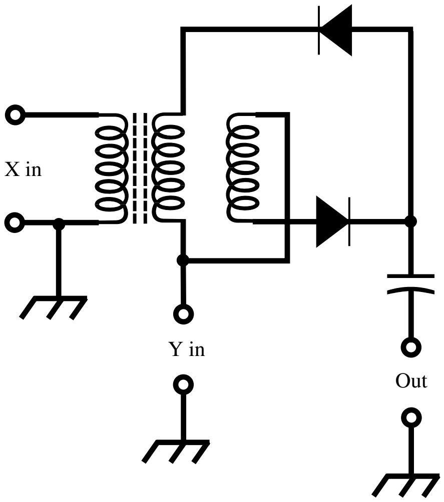

passive mixer

passive transponder A device that allows a machine, such as a computer or robot, to identify an object. A bar-code reader is a common example. Magnetic labels, such as those on credit cards and bank cash cards, are another example. It is so named because it does not transmit data; it requires a sensor for data detection.

password As a security device in computer operations, a group of characters upon whose presentation to the system via a terminal the user is allowed access to memory or control of information.

password retry limitation A security feature that prevents hackers from making repeated guesses at passwords in an attempt to break into a computer, network, or database. If more than three unsuccessful entries are made in succession, for example, the system will not accept further access attempts for a certain preprogrammed length of time, say 1 hour.

paste In “dry” batteries and electrolytic capacitors, a gelatinous electrolyte.

patch **1.** A temporary signal path, as between a radio receiver and a telephone or, conversely, between a telephone line and a radio transmitter. **2.** To make quick, usually temporary connections, as with a patch cord. **3.** Instructions entered by an unconditional branch to a computer program for the purpose of correction.

patch bay **1.** See PATCH PANEL. **2.** A set of patch panels.

patch cord A flexible line of one or more conductors with a jack or connector at each end, used to interconnect (patch) circuit points exposed for the purpose on a panel or breadboard.

Patchett tone control A dual tone-control circuit using a variable series resistance-capacitance (RC) filter for treble boost, and a variable shunt RC filter for bass boost. The input signal is applied in parallel to both filters. The outputs are combined in an audio mixer.

patching The interconnection of two or more signal media or lines.

patch panel A panel on which the terminals of a system are accessible for interconnection, tests, etc. It is used especially in high-fidelity audio recording systems. It was once commonly used in manual telephone-switching applications.

patch up **1.** To replace faulty or damaged parts in an electronic system with roughly appropriate surrogates to restore operation quickly (usually under emergency conditions). Also see DOCTOR. **2.** To wire a circuit quickly using patch cords for preliminary test and evaluation.

patent **1.** A document awarded by a government body, giving to an inventor the exclusive right to exploit an invention for a specified number of years. Formally called *letters patent.* **2.** The monopoly granted by a document, as defined in **1**.

path **1.** The route over which current flows. **2.** In radio and navigation, the imaginary line extending directly between transmitter and receiver (or target). **3.** In a computer program, the logical order of instructions.

pathometer A form of lie detector that indicates changes in the electrical resistance of the human body.

pattern **1.** An established sequence of steps in a process. **2.** An arrangement of terms in a matrix. **3.** The graphical representation of a varying quantity (e.g., an alternating-current wave pattern). **4.** The image on the screen of an oscilloscope, or the record traced by an oscillograph. **5.** The graphic polar representation of the radiation field of an antenna. **6.** The arrangement of bits in a word or field.

pattern recognition In machine-vision systems, a method of identifying an object or decoding data according to geometric shape. Optical character recognition (OCR) is an example. The machine recognizes combinations of shapes, and deduces their meanings via a computer program.

pause editing In the editing of audio tape recordings, the use of a “pause” switch to temporarily stop the tape when necessary.

PAV Abbreviation of PHASE-ANGLE VOLTMETER.

pawl In a mechanical stepping device, as in a nonelectric clock, a device made to engage the sloping sprockets on a wheel to ensure shaft rotation in one direction only.

PAX Abbreviation of PRIVATE AUTOMATIC EXCHANGE.

pay-per-view Abbreviation, PPV. Television service in which each subscriber pays only for individually selected programs.

pay TV See SUBSCRIPTION TV.

Pb Symbol for LEAD.

P band A radio-frequency band extending from 225 to 390 MHz.

PBX Abbreviation of PRIVATE BRANCH EXCHANGE.

PC **1.** Abbreviation of PERSONAL COMPUTER. **2.** Abbreviation of PRINTED CIRCUIT. **3.** Abbreviation of PHOTOCELL. **4.** Abbreviation of POSITIVE COLUMN. **5.** Abbreviation of POINT-CONTACT. **6.** Abbreviation of PERCENT (also, pct.). **7.** Abbreviation of PROGRAM COUNTER.

pc **1.** Abbreviation of PICOCOULOMB. Also, pC (preferred). **2.** Abbreviation of PICOCURIE. Also, pCi (preferred). **3.** Abbreviation of PARSEC.

pC Abbreviation for PICOCOULOMB.

PCB Abbreviation of PRINTED-CIRCUIT BOARD.

PC board See PRINTED-CIRCUIT BOARD.

PC diode See POINT-CONTACT DIODE.

p-channel JFET See P-CHANNEL JUNCTION FIELD-EFFECT TRANSISTOR.

p-channel junction field-effect transistor Abbreviation, PFET. A junction-type FET in which the gate junction has been formed on a bar or die of p-type semiconductor material. Compare N-CHANNEL JUNCTION FIELD-EFFECT TRANSISTOR.

p-channel MOSFET A metal-oxide semiconductor field-effect transistor in which the channel is composed of p-type silicon. Also see DEPLETION-TYPE MOSFET, DEPLETION-ENHANCEMENT-TYPE MOSFET, and ENHANCEMENT-TYPE MOSFET.

pCi Symbol for PICOCURIE.

PCL Abbreviation of PRINTED-CIRCUIT LAMP.

PCM Abbreviation of PULSE-CODE MODULATION.

PCM-FM Pertaining to a carrier that is frequency modulated by information that is pulse-code modulated. Also see FREQUENCY MODULATION and PULSE-CODE MODULATION.

PCM-FM-FM Pertaining to a carrier that is frequency modulated by one or more subcarriers that are frequency modulated by information that is pulse-code modulated. Also see FREQUENCY MODULATION and PULSE-CODE MODULATION.

PCM level In a pulse-code-modulated signal, one of several different possible signal conditions.

PCM-PM Pulse-code modulation that is accomplished by varying the phase of the carrier wave.

PC relay See PRINTED-CIRCUIT RELAY.

PCS Abbreviation for PERSONAL COMMUNICATION SERVICE.

PC transistor See POINT-CONTACT TRANSISTOR.

PD **1.** Abbreviation of PLATE DISSIPATION. **2.** Abbreviation of PULSE DURATION. **3.** Abbreviation of PROXIMITY DETECTOR. **4.** Abbreviation of POTENTIAL DIFFERENCE.

Pd Symbol for PALLADIUM.

PDA Abbreviation for *personal digital assistant.* See HANDHELD COMPUTER.

PDAS Abbreviation of *programmable data acquisition system.*

P display See PLAN POSITION INDICATOR.

PDM Abbreviation of PULSE-DURATION MODULATION.

PDM-FM Pertaining to a carrier that is frequency modulated by one or more subcarriers that are frequency modulated by pulses that are pulse-duration modulated. Also see FREQUENCY MODULATION and PULSE-DURATION MODULATION.

PDM-FM-FM Pertaining to a carrier that is frequency modulated by one or more subcarriers that are frequency modulated by pulses that are pulse-duration modulated. Also see FREQUENCY MODULATION and PULSE-DURATION MODULATION.

PDM-PM Pertaining to a carrier that is phase modulated by pulse-duration-modulated information. Also see PHASE MODULATION and PULSE-DURATION MODULATION.

PDT Abbreviation of PACIFIC DAYLIGHT TIME.

PDVM Abbreviation of PRINTING DIGITAL VOLTMETER.

PE **1.** Abbreviation of POTENTIAL ENERGY. **2.** Abbreviation of PROFESSIONAL ENGINEER. **3.** Abbreviation of PROBABLE ERROR.

peak **1.** The maximum value of a quantity. **2.** In an alternating-current cycle, the maximum positive or negative current or voltage point. **3.** The frequency at which the transmission by a bandpass circuit or device is maximum (attenuation is minimum), evidenced by a maximum in the frequency-response curve.

peak amplitude **1.** The maximum positive or negative current or voltage of a wave. **2.** The maximum instantaneous power of a signal.

peak anode (plate) current The maximum instantaneous current flowing in the anode (plate) circuit of a vacuum tube.

peak anode (plate) voltage The maximum instantaneous voltage applied to the anode (plate) of a vacuum tube.

peak chopper See PEAK CLIPPER.

peak current Abbreviation, I_p. The highest value reached by an alternating-current half-cycle or a current pulse. Also called MAXIMUM CURRENT.

peak detector See PEAK PROBE.

peak distortion **1.** The maximum instantaneous distortion in a signal, generally expressed as a percentage. **2.** Distortion of a modulated signal at envelope peaks.

peaked sawtooth A wave composed of a sawtooth and peaking-pulse components. The deflection voltage of a magnetic-deflection cathode-ray tube requires this waveform to produce a current sawtooth in the deflecting coils.

peaked waveform An alternating-current waveform having nearly pointed positive and negative half-cycles. Such a wave contains appreciable third-harmonic energy.

peak envelope power Abbreviation, PEP. For a linear radio-frequency (RF) power amplifier handling a modulated signal, the average RF output

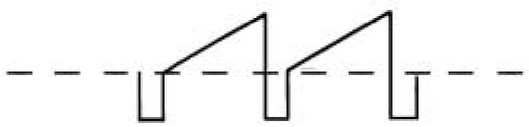

peaked sawtooth

power during a single RF cycle at the highest peak of the modulation envelope.

peaker **1.** See PEAK FILTER. **2.** See PEAKING TRANSFORMER.

peaker-notcher See NOTCHER-PEAKER.

peak factor For an alternating-current wave, the ratio E_m/E_{rms} or I_m/I_{rms}, where E_m is the maximum voltage, E_{rms} is the effective (root-mean-square) voltage, I_m is the maximum current, and I_{rms} is the effective current.

peak filter A frequency-selective circuit, such as a bandpass filter, for producing a peak response (see PEAK, **3**).

peak inductor current In a switching regulator, the maximum instantaneous current through the inductor when the device is switching at its fully rated duty cycle.

peaking The adjustment of a control or device for maximum indication on a meter or other display.

peaking coil A small inductor used to compensate the frequency response of a circuit, such as a video amplifier or video detector. Both series and shunt peaking coils are used.

peaking transformer A transformer whose output waveform is sharply peaked (of short duration, with respect to a cycle). The effect is obtained by means of a special core that, because it contains little iron, saturates easily.

peak inverse voltage Abbreviation, PIV. Often used interchangeably with the term PEAK REVERSE VOLTAGE. **1.** The peak value of the voltage applied to a rectifier diode in the reverse direction. **2.** The maximum value of reverse voltage that a rectifier diode will tolerate according to its specifications.

peak level lamp In audio recording and reproduction, a bulb or light-emitting diode (LED) that illuminates when sound peaks exceed a predetermined amplitude.

peak limiting **1.** A method of limiting the maximum amplitude of a signal. When the instantaneous peak amplitude, either positive or negative, exceeds a certain value, the output is clipped at that value. **2.** In pulse-code modulation, the effect resulting from the application of an input signal in excess of the virtual decision value.

peak modulated power In an amplitude-modulated wave, the maximum instantaneous signal power (including the carrier and sidebands). In 100-percent sinusoidal modulation, the peak modulated power is four times the unmodulated carrier power.

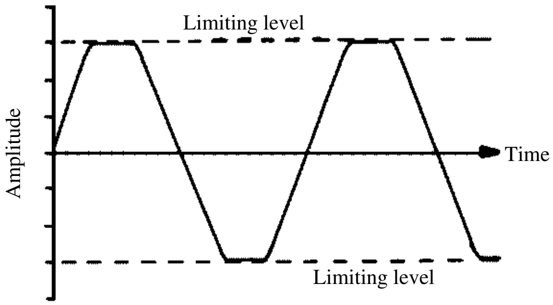

peak limiting

peak point The highest current point in the current-voltage response curve of a tunnel diode. Immediately beyond this point, the current decreases as the applied voltage is increased, indicating a negative-resistance region. Compare VALLEY POINT.

peak power **1.** Symbol, P_p. Unit, watt. Alternating-current power that is the product of the peak voltage (E_p) and the peak current (I_p). For E_p in volts and I_p in amperes, the peak power in watts is given by $P_p = E_p I_p$. **2.** The highest output power that an amplifier or device can produce without excessive distortion. **3.** The maximum instantaneous power that a speaker can handle without risk of damage.

peak probe A voltmeter test probe containing a diode circuit whose direct-current output voltage is close to the peak value of the applied alternating-current test voltage.

peak recurrent forward current For a semiconductor diode, the maximum repetitive instantaneous forward current as measured under specified conditions of operation.

peak reverse voltage Abbreviation, PRV. In semiconductor operations, the peak value of the voltage applied in reverse polarity across the junction. It is often used interchangeably with the term PEAK INVERSE VOLTAGE.

peak signal level The maximum instantaneous signal power or voltage specified for particular operating conditions.

peak-to-peak Abbreviations, p-p or pk-pk. For an alternating-current waveform, pertaining to the arithmetic difference between the positive peak and negative peak values of current or voltage.

peak-to-peak probe A voltmeter test probe containing a diode circuit whose direct-current output voltage is close to the peak-to-peak value of the applied alternating-current test voltage.

peak-to-peak voltage The arithmetic sum of positive and negative peak voltages in an alternating-current (ac) wave. Thus, a symmetrical sine-wave ac voltage of 115.0 V rms has a peak value of 162.63 V and a peak-to-peak value of 325.3 V. Also see PEAK VOLTAGE.

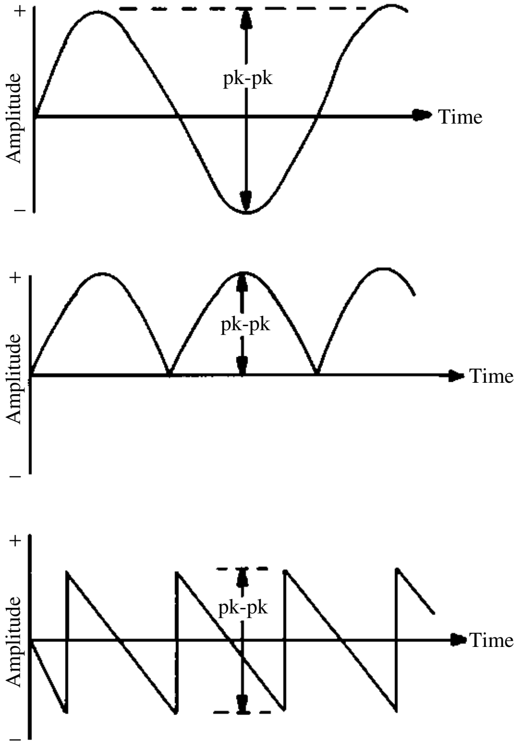

peak-to-peak voltage

peak torque Symbol, T_p. For a torque motor, the maximum useful torque at maximum recommended input current.

peak voltage Abbreviation, E_p. The highest value reached by an alternating-current voltage half cycle, or by a voltage pulse. Also called MAXIMUM VOLTAGE.

peak voltmeter **1.** An alternating-current (ac) voltmeter that responds to the peak value of the applied voltage. **2.** An ac voltmeter that responds to the average value of the applied voltage—even though its scale reads in peak volts.

pea lamp A miniature incandescent bulb, sometimes used as a control-panel or meter light.

PEC Abbreviation of PHOTOELECTRIC CELL.

pedestal See BLANKING PEDESTAL.

pedestal level See BLANKING LEVEL.

pel See PIXEL.

Peltier effect A drop below ambient temperature at the junction between two dissimilar metals when an electric current is passed through the junction.

PEM Abbreviation of *photoelectromagnetic.*

pen-and-ink recorder A graphic recorder in which a fountain-pen-type stylus inscribes an ink line on a paper chart. Also called *pen recorder.*

pencil **1.** A beam of electrons or other particles or rays that either converges to, or diverges from, a specific point. **2.** A pair of geometric entities sharing a property (e.g., lines intersecting at a single point).

pendulum switch A device that closes a circuit when subjected to physical shock. One type consists of a dangling element resembling a pendulum, with one or more nearby contacts.

penetrating frequency For a particular layer of the ionosphere, the lowest high frequency at which a vertically propagated wave penetrates the layer (i.e., it is not reflected back to earth). Also called CRITICAL FREQUENCY.

penetrating radiation Ionizing radiation that passes through otherwise opaque materials. A relative term; low-energy X rays are less penetrating than high-energy X rays, which, in turn, are less penetrating than gamma rays.

penetrating rays See COSMIC RAYS.

penetration depth See DEPTH OF PENETRATION.

pent Abbreviation of PENTODE.

pentavalent element An element whose atoms have five valence electrons (e.g., antimony or arsenic).

pentode A five-electrode vacuum tube in which the electrodes are an anode, cathode, control grid, screen grid, and suppressor grid.

pentode field-effect transistor A field-effect transistor with three separate gates.

pentode transistor A bipolar transistor with three emitters.

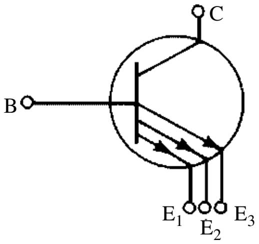

pentode transistor

penumbra **1.** That part of a shadow in which the light source is not fully obscured by the eclipsing object. **2.** In a sunspot, the outer part of the spot; it is less dark than the inner portion.

PEP **1.** Abbreviation of PLANAR EPITAXIAL PASSIVATED. **2.** Abbreviation of PEAK ENVELOPE POWER.

PEP diode See PLANAR EPITAXIAL PASSIVATED DIODE.

PEP reading wattmeter A wattmeter that shows the peak envelope power output of a transmitter.

PEP transistor See PLANAR EXPITAXIAL PASSIVATED TRANSISTOR.

perceived level The level of a disturbance, particularly sound, as sensed by a person. It is generally expressed in decibels, with respect to a certain threshold value. The threshold is assigned an intensity of 0 dB.

percent An expression of a fraction, in terms of hundredths. A quantity of x percent indicates a fraction of $x/100$. Percent is usually abbreviated by the symbol %.

percentage error The amount by which a measured value differs from the true value, expressed as a percentage (the number of parts per 100 that the measurement is in error).

percentage-modulation meter An instrument that provides direct readings of the modulation percentage of an amplitude-modulated signal. The meter scale or dial is graduated in increments from 0 to somewhat more than 100 percent.

percentage uncertainty The maximum possible error in a measurement, expressed as a percentage of the measured value. Also see UNCERTAINTY IN MEASUREMENT.

percent distortion Symbol, *%D.* In the determination of harmonic distortion, the total harmonic voltage expressed as a percentage of the fundamental voltage, plus total harmonic voltage; $\%D = 100E_h/E_t$, where E_h is the total voltage of the harmonic components, and E_t is the total signal voltage (fundamental plus harmonics).

percent modulation See MODULATION PERCENTAGE.

percent modulation meter See PERCENTAGE-MODULATION METER.

percent ripple The amount of ripple voltage in the direct-current (dc) output of a rectifier or generator, expressed as a percentage of the nominal dc output voltage.

perfect crystal A crystal without defects or impurities. The atoms are arranged in a regular pattern with no faults.

perforated board A plastic panel provided with a number of small holes in orderly columns and rows for the insertion of the pigtails of components, or of push-in terminals to facilitate quick assembly of prototype circuits. Also called *perfboard.*

performance curve A curve depicting the behavior of a component or circuit under specified conditions of operation. Such a curve, for example, might display the variation of output power with input power, the variation of frequency with voltage, etc. Compare CHARACTERISTIC CURVE.

performance test A test made primarily to ascertain how a system behaves. The test is concerned with normal operation, whereas a diagnostic test is a troubleshooting procedure. Compare TROUBLESHOOTING TEST.

perigee **1.** The point at which an earth-orbiting satellite attains its lowest altitude. It occurs once for every complete orbit. At this point, the satellite travels faster than at any other point in the orbit. **2.** The altitude, measured from the earth's surface or the earth's center, of an earth-orbiting satellite at its closest approach.

perihelion **1.** The point at which a solar-orbiting satellite attains its lowest altitude. It occurs once for every complete orbit. At this point, the satellite travels faster than at any other point in the orbit. **2.** The altitude, measured from the sun's surface or the sun's center, of a solar-orbiting satellite at its closest approach.

perimeter protection The use of a security system to restrict or prevent access to a designated area, using sensors and/or barriers around the boundaries of the area.

period Symbol, *T.* Unit, second. The duration of a complete alternating-current cycle or of any cyclic event; $T = 1/f$, where f is the frequency in Hertz. Also see CYCLE, FREQUENCY, and HERTZ.

periodic and random deviation Abbreviation, PARD. In the direct-current output of a rectifier, the combined PERIODIC DEVIATION, including ripple, noise, hum, and transient spikes.

periodic curve A curve that repeats its shape in each period (e.g., a sine curve).

periodic deviation Repetitive deviation of a quantity from its normal value (e.g., ripple in the direct-current output of a rectifier).

periodic function A mathematical function that is represented by a periodic curve (e.g., the sine function $y = \sin x$).

periodicity In a transmission line, the tendency for power to be reflected at a point or points where the diameter of the line changes.

periodic law The observation that when the chemical elements (see ELEMENT, **3**) are arranged in increasing order of atomic number, their physical and chemical properties recur periodically. Also see PERIODIC TABLE.

periodic table A table in which the chemical elements (see ELEMENT, **3**) are arranged according to the periodic law. The vertical columns in the table, labeled groups, contain elements possessing related properties (e.g., silicon and germanium in group IV). The rows, labeled periods, depict the periodic shift in the properties of the elements.

peripheral **1.** Pertaining to equipment accessory to a central system (e.g., peripheral input/output devices online or offline to computers, data recorders, and indicators). Also see ANCILLARY EQUIPMENT. **2.** Peripheral equipment in a computer system (e.g., printers, modems, external disk drives, tape drives, etc.).

peripheral buffer As part of a peripheral in a computer system, a storage unit in which data temporarily resides on its way to or from the central processing unit. Also called INPUT/OUTPUT BUFFER.

peripheral electron See VALENCE ELECTRON.

peripheral equipment See PERIPHERAL, **1**, **2**.

peripheral interface adapter Abbreviation, PIA. An integrated circuit that acts as an input/output port to interface a microprocessor with peripheral devices.

peripheral transfer In a computer system, the transfer of a unit of data between peripherals, or between a peripheral and the central processing unit.

Permalloy A high-permeability alloy of iron and nickel.

permamagnetic speaker See PERMANENT-MAGNET SPEAKER.

permanent magnet A body that is always magnetized (i.e., without the application of electricity and without requiring the presence of another magnet). Compare TEMPORARY MAGNET.

permanent-magnet erase Erasure of magnetic tape by the field of a permanent magnet. Typically, it is a two-step process: a magnet erases what it can of the signal, leaving any residual magnetization for erasure by a second magnet.

permanent-magnet focusing In a cathode-ray tube, the focusing of the electron beam by means of permanent magnets.

permanent-magnet generator An electromechanical generator in which the field (either stationary or rotating) is provided by a multipole permanent magnet. Also called *magneto.*

permanent-magnet loudspeaker See PERMANENT-MAGNET SPEAKER.

permanent-magnet magnetizer A magnetizer using a permanent magnet as the magnetic-field source.

permanent-magnet meter An indicating meter in which a movable coil rotates between the poles of a permanent magnet. Compare ELECTRODYNAMOMETER and IRON-VANE METER.

permanent-magnet motor A motor having a permanent-magnet field.

permanent-magnet relay A polarized relay using a permanent magnet.

permanent-magnet speaker An acoustic loudspeaker in which the core is a strong permanent magnet (as opposed to a direct-current electromagnet). Also see MAGNETIC SPEAKER.

permanent storage See NONVOLATILE MEMORY.

permeability Unit, H/m. A quantitative indicator of the extent to which a material concentrates magnetic flux: for a given constant magnetic-field intensity, the ratio of magnetic flux density in the material to the magnetic flux density in air.

permeability curve See B-H CURVE.

permeability-tuned oscillator A radio-frequency oscillator in which the frequency is varied or adjusted by moving a ferromagnetic core in and out of the coil of an inductance-capacitance (LC) tuned circuit.

permeability tuning Variation of the resonant frequency of an inductance-capacitance LC circuit by changing the position of a magnetic core within the inductor. This type of tuning is used in amplifiers, oscillators, filters, and wavetraps.

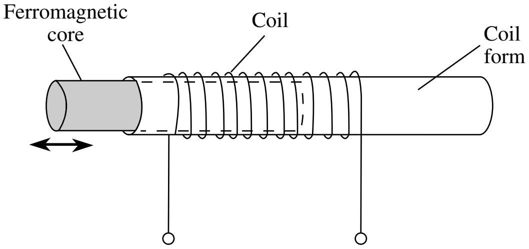

permeability tuning

permeameter An instrument for measuring permeability.

permeance Unit, Wb/A. In a magnetic circuit, the ease with which a magnetic field is established. The reciprocal of RELUCTANCE.

Permendur A high-permeability magnetic alloy containing equal parts of iron and cobalt. At saturation, the flux density of this material can be 2 teslas (20,000 gauss).

Perminvar A high-permeability magnetic alloy of cobalt, iron, and nickel. At saturation, the flux density of this material can approach 1.2 teslas (12,000 gauss).

permittivity See DIELECTRIC CONSTANT.

permutation A selection of several factors or objects from a group, in a specific ordered sequence. For example, one of the permutations of the elements of the set (1, 2, 3, 4, and 5) is the ordered sequence 4, 1, 3, 5, 2.

permutation modulation A method of modulation accomplished by varying the sequence of digital bits.

peroxide of lead In a lead-acid cell or battery, a compound of lead and oxygen that composes the positive electrode or electrodes.

persistence **1.** The effect whereby the retina of the eye continues to register a projected scene for approximately 0.05 second after the scene disappears. This allows perception of a sequence of video frames as a continuous moving image. **2.** The tendency of certain phosphors to glow after the excitation has been removed. Thus, after the electron beam in a cathode-ray tube has passed over the screen, the phosphor might continue to glow for a certain time along the path traced by the beam. Some phosphors, such as those used in high-speed oscilloscopes, have virtually no persistence, whereas others have long persistence.

persistent oscillations Successive oscillations of constant amplitude. Also called CONTINUOUS WAVE.

persistor A device used at low temperatures for temporary memory storage that operates between superconducting and normal conditions.

personal communications service Abbreviation, PCS. Also called *digital cellular.* An enhanced wireless network using digital modulation, cellular repeaters, and facilitating telephone and Internet connections. Emphasis is on maximizing user mobility and portability, and minimizing blind zones. Compare CELLULAR COMMUNICATIONS.

personal computer A small computer equipped with a keyboard, display, hard disk, diskette drive(s), a modem or fax/modem, one or more serial data ports, and one or more parallel data ports. They are used extensively by individuals and businesses for record keeping, data processing, communications, word processing, graphics, etc.; they are also used in schools as an educational aid.

personal digital assistant Abbreviation, PDA. See HANDHELD COMPUTER.

personal equation The value of systematic error for a person observing specific phenomena or making measurements.

personality Characteristics that make an intelligent computer or robot human-like. In general, the more powerful the computer, the more personality it can have, depending on the installed software. In some cases, certain malfunctions in a computer can produce personality quirks.

personal robot A usually autonomous robot intended for use by individuals. The most common examples are robot toys, programmable with a PERSONAL COMPUTER, intended for the education and entertainment of children. More sophisticated devices can perform domestic tasks, such as cleaning floors and mowing lawns.

peta- Abbreviation, P. A prefix meaning 10^{15}.

petagram Abbreviation, Pg. A large unit of mass or force, equal to 10^{15} grams or 10^{12} kilograms.

petameter Abbreviation, Pm. A large unit of (astronomical) distance, equal to 10^{15} meters or 10^{12} kilometers.

pF Abbreviation of PICOFARAD.

pf Symbol for POWER FACTOR.

PFET Abbreviation of P-CHANNEL JUNCTION FIELD-EFFECT TRANSISTOR.

PFM Abbreviation of PULSE-FREQUENCY MODULATION.

PG Abbreviation of POWER GAIN.

Pg Abbreviation of PETAGRAM.

pH **1.** Symbol for *hydrogen-ion concentration.* Numerically, pH is the negative logarithm of the effective hydrogen-ion concentration in gram equivalents per liter. The scale runs from zero to 14, on which 7 denotes neutrality relative to acidity vs. alkalinity; values between zero and 7 denote acidity, and values between 7 and 14 denote alkalinity. **2.** Abbreviation of PICOHENRY.

phantom Radio interference in the form of a beat note (heterodyne), resulting from interference between two strong carriers, often from local radio stations. When the phantom frequency lies within the tuning range of a receiver, the phantom can be tuned in as a separate signal. But when the phantom corresponds to the intermediate frequency (IF) of the receiver, it will ride into the IF amplifier and be present as an untunable interferential signal.

phantom channel In a properly phased high-fidelity stereo sound system, the apparent sound source centered between the left- and right-channel loudspeakers.

phantom circuit In wire telephony, a third circuit that has no wires; it results from a method (using repeating coils) of making two other circuits do the work of (this third) one.

phantom signal Also called *bogey.* In a radar system, a signal that does not correspond to an actual target. The origin of the phantom signal or echo cannot be readily determined.

phantom target See ECHO BOX.

phase angle Unit, degree or radian. In an alternating-current (ac) circuit, the lag or lead between the instant that one alternating quantity reaches its maximum value and the instant that another alternating quantity reaches its maximum value. It is usually given in degrees (a complete cycle being 360 degrees) along the horizontal axis of the time-versus-magnitude graph of the ac quantity.

phase-angle voltmeter An instrument that indicates both the magnitude and phase of a voltage.

phase coincidence For signals having the same frequency, the condition of their coinciding in terms of instantaneous amplitudes, so positive peaks of the first signal correspond to positive peaks of the second signal, and negative peaks of the first signal correspond to negative peaks of the second signal. For periodic waves that do not change their characteristics with time, this is the same thing as being shifted by an integral multiple of 360 degrees in phase. Compare PHASE OPPOSITION.

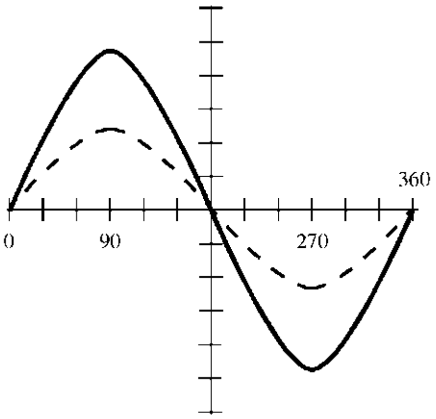

phase coincidence

phase compensation In an operational amplifier, compensation for excessive phase shift in the feedback.

phase compressor A push-pull phase-inverter circuit in which a capacitor is connected between each collector or drain and the opposite output terminal to attenuate in-phase components, such as even-numbered harmonics.

phase constant A figure providing the rate (in degrees of phase per unit length) at which the phase lag of the current or voltage field component in a traveling wave increases linearly in the propagation direction.

phase corrector A circuit that returns a signal to a certain phase after the signal has passed through a circuit or medium that has caused phase distortion.

phased antenna See PHASED ARRAY.

phased array Also called *phased antenna*. An antenna system having two or more driven elements fed with a certain relative phase, and spaced at a certain distance, resulting in a directivity pattern that exhibits gain in some directions and little or no radiation/response in other directions. Such an array can have two elements, producing a unidirectional cardioid or bidirectional figure-eight pattern. More complex arrays have several elements, usually vertical antennas, strategically positioned and fed with signals of specified phase to produce a highly tailored pattern. The most sophisticated systems have rotatable or steerable radiation/response patterns.

phase-delay equalizer See DELAY EQUALIZER.

phase detector See PHASE-SENSITIVE DETECTOR.

phase diagram A graphical representation of waves having equal frequency, but differing in phase. The phase difference for two identical waveforms is greater than or equal to zero degrees, but less than 360 degrees.

phase difference **1.** The difference (in time, angle, or fractional cycle) between the instants at which two alternating quantities reach a given value. **2.** For a dielectric, the complement of PHASE ANGLE; that is, 90 degrees minus the phase angle in degrees.

phase discriminator See DISCRIMINATOR, FOSTER-SEELEY DISCRIMINATOR, RATIO DETECTOR, and TRAVIS DISCRIMINATOR.

phase distortion Distortion characterized by input/output phase shift between various components of a signal passed by a circuit or device.

phase inverter A resistance-capacitance-coupled amplifier with a single-ended input and a push-pull output. This circuit enables a push-pull amplifier to be driven without an input transformer.

phase-locked loop Abbreviation, PLL. An oscillator that combines the flexibility of a conventional variable-frequency oscillator (VFO) with the stability of a crystal oscillator. The oscillator output is passed through a programmable divider that divides the frequency by a specific integral value *n* chosen by the operator. The output frequency of the divider is locked, by means of a phase comparator, to the signal from a crystal-controlled reference oscillator. As long as the output from the divider is exactly at the reference-oscillator frequency, the two signals are in phase, and the output of the phase comparator is zero volts dc. If the VCO frequency changes, the phase also changes, and the phase comparator produces a dc error voltage. The error voltage is applied to the VCO, causing the VCO frequency to correct itself. This maintains the VCO frequency at precisely *n* times the reference-oscillator frequency.

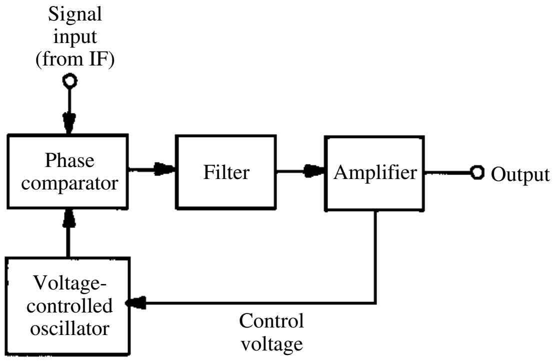

phase-locked loop

phase-locked oscillator An oscillator in which the inductance or the capacitance is varied periodically at half the driving frequency.

phase margin In an integrated-circuit amplifier, the extent to which the device shifts the phase of a signal more or less than one-half cycle (180 degrees) for a certain signal voltage.

phase modulation Abbreviation, PM. A method of modulation in which the phase of the carrier current is varied in accordance with the instantaneous modulating-signal voltage. The result is similar to FREQUENCY MODULATION.

phase modulator A circuit or stage that produces PHASE MODULATION.

phase multiplier A circuit used for the purpose of phase comparison between signals. The frequency of the measured signal is multiplied, resulting in multiplication of the phase difference. This improves the sensitivity of the measuring apparatus.

phase opposition For signals having the same frequency, the condition of their being inverted relative to each other in terms of instantaneous amplitudes, so positive peaks of the first signal correspond to negative peaks of the second signal, and negative peaks of the first signal correspond to positive peaks of the second signal. This is not the same thing as being shifted by an odd

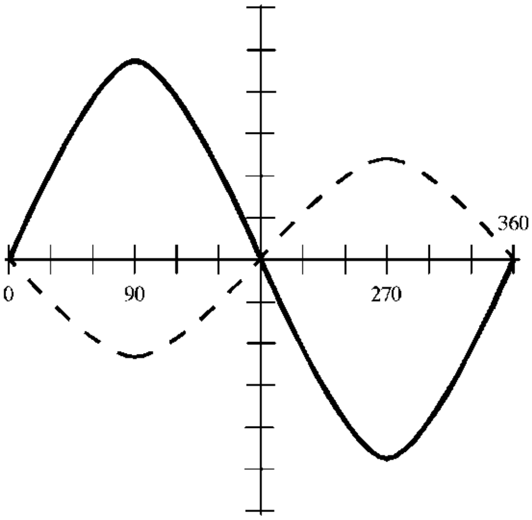

phase opposition

integral multiple of 180 degrees in phase, although in practice, with sine waves and square waves, the effect is the same. Compare PHASE COINCIDENCE.

phase resonance See VELOCITY RESONANCE.

phase reversal **1.** The inversion of an alternating-current (ac) signal. The instantaneous amplitude (current or voltage) is multiplied by a negative constant. Thus, the positive half-cycles become negative, and the negative half-cycles become positive. **2.** A phase shift of ±180 degrees (±½ cycle) in an ac signal.

phase-rotation relay See PHASE-SEQUENCE RELAY.

phase-rotation system A system for producing single-sideband signals without using selective filters. In one such system, two balanced modulators are used. One of these receives carrier and modulating voltages that are 90 degrees out of phase with voltages that are fed to the other balanced modulator.

phase-sensitive detector Abbreviation, PSD. A detector for frequency modulation (FM) and phase modulation (PM). It delivers a direct-current output voltage whose value is proportional to the difference in phase between a reference signal and the signal from a local oscillator.

phase-sequence relay In a polyphase system, a relay or relay circuit that is actuated by voltages reaching maximum positive amplitude in a predetermined phase sequence. Also called PHASE-ROTATION RELAY.

phase shift **1.** A change in the displacement, as a function of time, of a periodic disturbance having constant frequency. **2.** The magnitude of a change, as defined in **1**, measured in fractions of a wavelength or in electrical degrees.

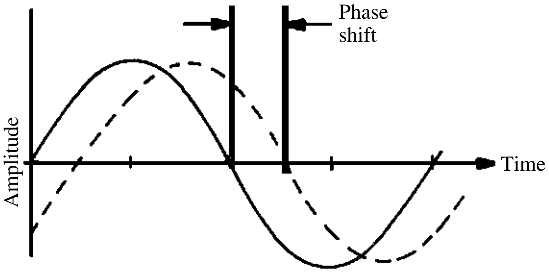

phase shift

phase-shift bridge A four-arm-bridge circuit for shifting the phase of an alternating-current signal. Such a circuit is often used (with one arm variable) to shift the phase of the firing voltage for a thyratron.

phase-shift discriminator See FOSTER-SEELEY DISCRIMINATOR.

phase shifter A circuit, such as an inductance-capacitance (LC) or resistance-capacitance (RC) network, or a device, such as a Helmholtz coil or phase-shifting capacitor, that introduces a phase shift between input and output signals.

phase-shifting capacitor A special four-stator, one-rotor variable capacitor that, with a transformer-coupled resistance-capacitance (RC) circuit, provides 360 degrees of continuously variable phase shift for one rotation of the rotor. The rotor plate turns like a cam under the stators because of the off-center insertion of the rotor shaft.

phase-shift oscillator A single-stage oscillator in which the required 180-degree phase shift in the signal (fed back from output to input) is obtained by passing the output through a phase-shifting network.

phase-shift-type distortion meter A distortion meter in which the output signal of a device under test is compared with a distortion-free input test signal. The output signal phase is shifted 180 degrees, with respect to the input, and the two amplitudes are made equal. If there is no distortion, the signals cancel each other, and the result is zero. Any remaining signal is proportional to the total harmonic distortion (THD).

phase-splitting circuit A circuit that produces, from a single input signal, two output signals differing in phase.

phase-splitting driver A PHASE INVERTER used as the driver of a push-pull amplifier.

phase velocity The velocity of a wave, provided by the product of the frequency and the wavelength.

phase windings In an alternating-current generator, windings that deliver voltages that differ in phase.

phasing capacitor In a crystal filter, a small variable capacitor that constitutes one arm of a four-arm bridge in which the crystal is another arm. Adjustment of this capacitor balances the bridge, thus preventing the undesirable passage of a signal through the capacitance of the crystal holder.

phenol-formaldehyde plastics A family of plastic insulating materials made with phenolic resin, and occasionally used as dielectrics and air-core coil forms. Some of the trade names for these materials include *Bakelite*, *Catalin*, *Durez*, *Durite*, *Formica*, and *Micarta.*

phenolic insulants See PHENOL-FORMALDEHYDE PLASTICS.

phenolic resin A synthetic resin made by condensing phenol (carbolic acid) with formaldehyde.

phenomenon An event or circumstance that can be verified by the senses, as opposed to one subject to theory or speculation (e.g., the phenomenon of magnetic attraction).

Phillips gate A device that allows measurement of the gas pressure in a confined chamber. A current is passed through the gas. The magnitude of the current, for a given gas, is a function of the gas pressure and temperature.

Phillips screw A screw with a pair of slots in its head. The slots are arranged like an x. Phillips screws are available in many different sizes, as are ordinary screws. The x-shaped pair of slots reduces the tendency for the screwdriver to slip out of the screw head as the screw is rotated.

Phi phenomenon The illusion of motion resulting from the rapid presentation to the eye of pictures showing objects in a succession of different positions. Television and motion pictures exploit this illusion. Also see PERSISTENCE.

pH meter An instrument used to measure the acidity or alkalinity of solutions. Also see PH, **1**.

phon A unit of apparent change in loudness discerned by a listener. Unlike the decibel, the phon includes compensation for the ear's nonlinear response to attendant frequency changes. At a frequency of 1 kHz, a change in loudness of 1 phon is the equivalent of 1 decibel.

phone **1.** Telephone (wire or radio). **2.** To establish communication via telephone. **3.** Colloquialism for voice communication (radiotelephone), particularly via amateur-radio single sideband on the high-frequency bands (160 through 10 meters). **4.** A minimal, unique speech sound. Also called SOUND UNIT.

phoneme An individual sound or syllable in the human voice, with a characteristic amplitude-vs. frequency spectral pattern. It is important in speech recognition and speech synthesis. Computers can be programmed to identify and transcribe these sounds; computers can also be programmed to generate the sounds from text data.

phone jack The female mating device for a PHONE PLUG.

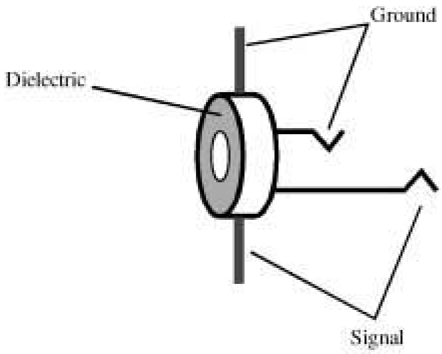

phone jack

phone monitor A simple device for listening to amplitude-modulated radio transmissions to test their quality. In its most rudimentary form, it consists of a pickup antenna, semiconductor-diode detector, and high-resistance headphones.

phone patch A device for establishing a connection (patch) between radio and wire-telephone facilities. Also see PATCH.

phone plug A type of plug originally designed for patching telephone circuits, now widely used in electronics and instrumentation. In its conventional form, it has a rod-shaped neck that serves as one contact, and a ball on the tip of the neck, but insulated from it, that serves as the other contact. Typical diameters are ⅛ inch and ¼ inch.

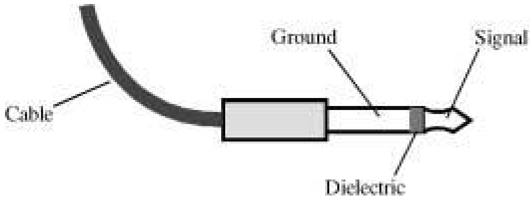

phone plug

phone test set An instrument for checking the performance of a radiotelephone transmitter. The set combines the functions of field-strength meter, modulation indicator, and aural monitor. Sometimes it includes a volt-ohm-milliammeter for troubleshooting the transmitter.

phonetic alphabet Words whose initial letters are used to identify the letters of the alphabet for which they stand. These words are spoken in radiotelephony to identify letters that, if spoken by themselves, might not be clearly heard.

Phonetic alphabet

Letter	Phonetic (Capitals indicate emphasis)
A	AL-fa
B	BRAH-vo
C	CHAR-lie
D	DEL-ta
E	ECK-o
F	FOX-trot
G	GOLF
H	ho-TEL
I	IN-dia
J	Ju-li-ETTE
K	KEE-low
L	LEE-ma
M	Mike
N	No-VEM-ber
O	OS-car
P	pa-PA
Q	Que-BECK
R	ROW-me-oh
S	see-AIR-ah
T	TANG-go
U	YOU-ni-form
V	VIC-tor
W	WHIS-key
X	X-ray
Y	YANK-key
Z	ZOO-loo

phonetic alphabet code word In radio and wire telephony, a word chosen for its easy recognition by ear to identify the letter of the alphabet with which it begins. For example: *Golf* for G, *Juliet* for J, *X-ray* for X.

phonics See ACOUSTICS, **1**.

phonocardiogram The record made by a PHONOCARDIOGRAPH.

phonocardiograph An instrument that makes a graphic record of heart sounds.

phono cartridge The vibration-to-electricity transducer (pickup) of a phonograph; it is actuated by the stylus (needle). Common types are ceramic, variable-inductance, and variable-reluctance. See PHONOGRAPH and PHONOGRAPH DISC.

phonocatheter A microphone that can be inserted into the body for the purpose of listening to the functions of internal organs.

phonograph A device for reproducing sound recorded on disc. It consists of a turntable, an amplifier, and one or more speakers.

phonograph cartridge See PHONO CARTRIDGE.

phonograph disc A thin, lightweight disc, usually made of vinyl or similar plastic, on which audio-frequency signals are recorded as irregularities in a spiral groove. In reproduction, these irregularities cause vibration in a PHONO CARTRIDGE as the turntable rotates.

phonograph oscillator See PHONO OSCILLATOR.

phono jack Also called *RCA jack.* A jack similar to a PHONE JACK, designed especially for the quick connection and disconnection of coaxial cables used with audio and low-frequency devices.

phonon A unit of energy resulting from vibration, as of a piezoelectric crystal.

phono oscillator A small radio-frequency (RF) oscillator modulated by the audio-frequency (AF) voltage from a phonograph. The modulated RF signal is picked up by a remote radio receiver (usually in the same room), and the sound is reproduced through a loudspeaker connected to the receiver.

phono plug Also called *RCA plug.* A plug similar to a PHONE PLUG, designed especially for the quick connection and disconnection of coaxial cables used with audio and low-frequency devices.

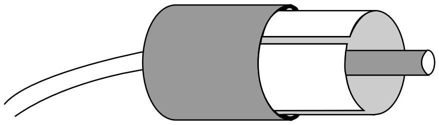

phono plug

phonoreception The hearing of high-frequency sounds.

phonorecord A PHONOGRAPH disc.

phonoselectroscope A special type of stethoscope, in which the main heartbeat is attenuated. This makes abnormal sounds more audible. The device can be adjusted in various ways to listen for abnormalities characteristic of various heart diseases.

phosphor A substance that glows when an electron beam strikes it. Such a substance is used as a coating on the screens of cathode-ray tubes. See also BEAT ZINC SILICATE; CADMIUM BORATE, SILICATE, and TUNGSTATE; CALCIUM PHOSPHATE, SILICATE, and TUNGSTATE; MAGNESIUM FLUORIDE, SILICATE, and TUNGSTATE; ZINC ALUMINATE; ZINC BERYLLIUM SILICATE; ZINC BERYLLIUM ZIRCONIUM SILICATE; ZINC BORATE; ZINC CADMIUM SULFIDE; ZINC GERMANATE; ZINC MAGNESIUM FLUORIDE; ZINC ORTHOSILICATE; ZINC OXIDE; ZINC SILICATE; and ZINC SULFIDE.

phosphor bronze A form of bronze whose elasticity, hardness, and toughness have been greatly improved by the addition of phosphorus. The metal is used for brushes, springs, switch blades, and contacts.

phosphor copper An alloy of copper and phosphorus used in the manufacture of PHOSPHOR BRONZE.

phosphorescence The property of some materials that ordinarily fluoresce to continue to glow after the stimulus (light or an electron beam) has been removed. Compare FLUORESCENCE.

phosphorescent screen A viewing screen coated with a phosphor (e.g., oscilloscope screen).

phosphorous Exhibiting the properties of phosphor (e.g., glowing after stimulation with light). Not to be confused with PHOSPHORUS.

phosphorus Symbol, P. A nonmetallic element of the nitrogen family. Atomic number, 15. Atomic weight, 30.974. It is used as a dopant in semiconductor processing.

phot The cgs unit of illumination: The direct illumination produced upon a one-centimeter-distant surface by a uniform point source of one international foot-candle. Equivalent to one lumen per square centimeter.

photocathode **1.** The photomosaic of a video camera tube. **2.** The light-sensitive cathode in a phototube.

PhotoCD Trade name for an image-recording system developed by Kodak, in which photographs can be stored on compact discs. Viewing is accomplished using personal computers.

photocell See PHOTOELECTRIC CELL.

photocell amplifier An amplifier used to boost the output of a photocell. With respect to the nature of the input signal, it can be an alternating-current (ac) or direct-current (dc) amplifier, depending on whether the output of the photocell is modulated dc or pure dc.

photochemical effect The phenomenon whereby certain substances undergo chemical change when exposed to light or other radiant energy. An example of such a substance is the silver bromide, silver chloride, or silver iodide on photographic film.

photoconductive cell A photoelectric cell, such as the cadmium-sulfide type, whose resistance is proportional to the intensity of light impinging upon it. The photoconductive cell acts as a light-sensitive variable resistor in a current path. Also see PHOTOCONDUCTIVE MATERIAL.

photoconductive effect The tendency for the electrical resistance of a substance to change when infrared radiation, visible light, or ultraviolet radiation strikes it. Different substances exhibit different degrees of this effect.

photoconductive material A substance that exhibits decreased electrical resistance when exposed to infrared rays, visible light, or ultraviolet. Some photoconductive substances are cadmium selenide, cadmium sulfide, germanium, lead sulfide, selenium, silicon, and thallous sulfide. Also see ACTINOELECTRIC EFFECT.

photoconductivity The phenomenon whereby the electrical resistance of certain materials (such as cadmium sulfide, cadmium selenide, germanium, selenium, and silicon) is lowered upon exposure to infrared rays, visible light, or ultraviolet. Also see PHOTOCONDUCTIVE MATERIAL.

photoconductor **1.** See PHOTOCONDUCTIVE MATERIAL. **2.** See PHOTOCONDUCTIVE CELL.

photocurrent See PHOTOELECTRIC CURRENT.

photo-Darlington Also, *photodarlington.* **1.** A phototransistor fabricated as a Darlington amplifier for high output current. **2.** A combination of photodiode (see LIGHT-SENSITIVE DIODE) and Darlington amplifier.

photodecomposition Chemical breakdown by the action of radiant energy. Also called *photolysis.*

photodetachment The removal of an electron from an atom or ion, resulting from the impact of a PHOTON.

photodetector **1.** An illumination meter that uses a PHOTOCELL. **2.** See OPTOELECTRONIC COUPLER.

photodielectric effect The tendency for the dielectric constant of a substance to change when infrared radiation, visible light, or ultraviolet radiation strikes it. Different substances exhibit different degrees of this effect.

photodiffusion effect See DEMBER EFFECT.

photodiode See LIGHT-SENSITIVE DIODE.

photodisintegration In the nucleus of an atom, disintegration resulting from PHOTON bombardment.

photoelasticity The tendency for the light-transmission characteristics of a substance to change with externally applied forces.

photoelectric alarm An alarm actuated when a light beam impinging on a photocell is interrupted.

photoelectric amplifier **1.** An amplifier for boosting the output of a photosensitive device. **2.** An OPTOELECTRONIC COUPLER possessing gain.

photoelectric cell A device that converts infrared, visible-light, or ultraviolet energy into electricity or electrical effects. It can function by producing a voltage (see PHOTOVOLTAIC CELL, SELENIUM CELL, SILICON CELL, SOLAR CELL, and SUN BATTERY) or by acting as a light-sensitive resistor (see LIGHT-SENSITIVE DIODE, PHOTOCONDUCTIVE CELL, and SELENIUM CELL).

photoelectric constant The quantity h/e, where h is Planck's constant and e is the unit electron charge.

photoelectric counter A counting device (electromechanical or fully electronic) that counts objects as they interrupt a light beam impinging upon a photocell.

photoelectric disintegration See PHOTODISINTEGRATION.

photoelectric effect The phenomenon whereby temporary changes occur in the atoms of certain substances under the influence of infrared, visible light, or ultraviolet radiation. Some of these materials undergo a change in their electrical resistance, whereas others generate electric current (see, for comparison, PHOTOCONDUCTIVE MATERIAL and PHOTOVOLTAIC MATERIAL).

photoelectric efficiency See QUANTUM YIELD.

photoelectric field-effect transistor See PHOTOFET.

photoelectricity Electricity produced by the action of light on certain materials, such as cesium, selenium, and silicon. Also see PHOTOEMISSION and PHOTOVOLTAIC CELL.

photoelectric material See PHOTOCONDUCTIVE MATERIAL and PHOTOEMISSIVE MATERIAL.

photoelectric multiplier A device that internally amplifies the current resulting from bombardment by infrared, visible light, or ultraviolet radiation. A PHOTOMULTIPLIER TUBE is an example of such a device.

photoelectric photometer An instrument that uses a photoelectric device for the purpose of measuring the intensity of infrared radiation, visible light, or ultraviolet radiation.

photoelectric proximity sensor A device that uses a light-beam generator, a photodetector, an amplifier, and a microprocessor to detect the presence of nearby objects. It is useful in robot guidance systems.

photoelectric pyrometer An optical pyrometer in which a photocell and appropriate filters act instead of the human eye.

photoelectric relay A relay actuated directly by a photocell or a photocell and amplifier. This type of relay is the basis of some PHOTOELECTRIC ALARM devices.

photoelectric sensor 1. See ELECTRIC EYE. **2.** See PHOTOELECTRIC PROXIMITY SENSOR. **3.** See PHOTOELECTRIC CELL.

photoelectric smoke alarm An alarm that is tripped by a PHOTOELECTRIC SMOKE DETECTOR when the density of smoke exceeds a safe level.

photoelectric smoke control A system for making automatic adjustments to a burning process when the smoke density exceeds a prescribed level. The initial element in the system is a PHOTOELECTRIC SMOKE DETECTOR.

photoelectric smoke detector A smoke detector in which a photocell, photodiode, phototransistor, or phototube is excited by a light beam passing through the air. The cell output current decreases when smoke fills the air. This current change trips an alarm or deflects an indicating meter when the density of the smoke exceeds a prescribed level.

photoelectric tape reader A punched-tape reader using a photocell, photodiode, phototransistor, or phototube to sense light passing through the holes.

photoelectric transducer A photocell, photodiode, phototransistor, or phototube used as a sensor.

photoelectric tube See PHOTOTUBE.

photoelectric wattmeter A power-measuring instrument useful for the approximate measurement of radio-frequency power. It consists of an incandescent lamp sharing an opaque enclosure with a photovoltaic cell. The power to be measured is applied to the lamp, which glows proportionately. The light excites the cell, causing it to deliver a direct current proportional to the power. This current deflects a milliammeter or microammeter. The meter can be calibrated to read directly in watts.

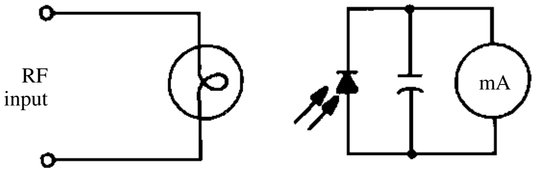

photoelectric wattmeter

photoelectromotive force The electromotive force (voltage) produced by a photovoltaic cell.

photoelectron An electron displaced within, or ejected from, an atom, as the result of infrared, visible light, or ultraviolet radiation striking the atom.

photo-emf See PHOTOELECTROMOTIVE FORCE.

photoemission The ejection of electrons from certain materials, such as cesium, when these materials are exposed to infrared, visible light, or ultraviolet radiation. Also see PHOTOEMISSIVE MATERIAL.

photoemissive material A substance that emits electrons when exposed to infrared, visible light, or ultraviolet radiation. A typical use of such a material is in the coating of the light-sensitive cathode of a phototube. The metals cesium, potassium, rubidium, and sodium are photoemissive.

photofabrication 1. A method of circuit-board manufacturing. The etching pattern is placed over the circuit-board material, the board is placed in a special solution, then the assembly is exposed to visible light. The light interacts with the solution to dissolve the metal in areas exposed to the light, but not in areas covered by the etching pattern. **2.** The technique in **1**, applied to the manufacture of integrated circuits.

photoFET A FIELD-EFFECT TRANSISTOR that exhibits properties similar to those of a bipolar PHOTOTRANSISTOR.

photoflash See ELECTRONIC FLASH, **1**.

photoglow tube See DISCHARGE LAMP.

photogram The permanent shadow produced by an object placed between a light source and photographic paper.

photographic exposure meter See EXPOSURE METER, **1**.

photographic recorder A graphic recorder that uses a light beam, deflected by galvanometer movement, that moves across photographic film or paper to produce a trace representing a varying quantity.

photographic sound recording See OPTICAL SOUND RECORDING.

photograph reception **1.** The use of FACSIMILE to print photographs transmitted in analog form via wire or radio. **2.** The use of a computer, equipped with a modem and graphics software, to display and/or store photographs transmitted in digital form via wire or radio.

photograph transmission **1.** The use of FACSIMILE to scan and send photographs in analog form via wire or radio. **2.** The use of a computer, equipped with a modem and video camera or optical scanner, to digitize and send photographs via wire or radio.

photoionization The ejection of electrons from atoms or molecules by the action of infrared, visible light, or ultraviolet radiation.

photoisolator See OPTOELECTRONIC COUPLER.

photojunction cell A photocell consisting of a semiconductor pn junction. The cell is useful mainly for its photoconductivity, although infrared, visible light, or ultraviolet energy striking the junction produces a small amount of photovoltaic action.

photokinesis Light-induced motion, as in a RADIOMETER.

photolithographic process A method of producing integrated circuits and printed circuits by photographing (often at considerable reduction) an enlarged pattern of the circuit on a suitable light-sensitized surface of metal or semiconductor, and chemically etching away unwanted portions of the surface.

photolysis See PHOTODECOMPOSITION.

photomagnetic effect Light-sensitive magnetic susceptibility in some materials.

photomap A photo taken of terrain from a high altitude and usually overlaid with a reference grid.

photomask In PHOTOFABRICATION, the transparent film or template on which the etching pattern is drawn.

photometer An instrument used to compare the luminous intensity of two light sources.

photometric measurement of power See PHOTOELECTRIC WATTMETER.

photometry The science of visible-light measurement. The response of the human eye is used as the basis for preferred sensors (those used with photometric instruments, which have spectral sensitivity curves resembling those of the eye). Compare RADIOMETRY.

photomosaic In a television camera tube, the flat photocathode screen on which the image is projected by the lens system and scanning electron beam. The surface of the screen is covered with tiny light-sensitive droplets. Also see DISSECTOR TUBE, ICONOSCOPE, and ORTHICON.

photomultiplier tube A type of PHOTOTUBE that delivers high output current for a given light intensity by utilizing the secondary emission of electrons. The initial light-sensitive cathode emits electrons; these strike a specially placed metal plate with a force that dislodges more electrons. These electrons, together with the initial emission, are reflected to a second plate, where they dislodge still more electrons. This process continues from deflection plate to deflection plate through the tube. The final plate deflects the accumulated electrons to the anode (collector electrode).

photon A quantum of radiant energy whose energy constant W (in joules) is equal to hf, where h is the PLANCK CONSTANT and f is the frequency in Hertz.

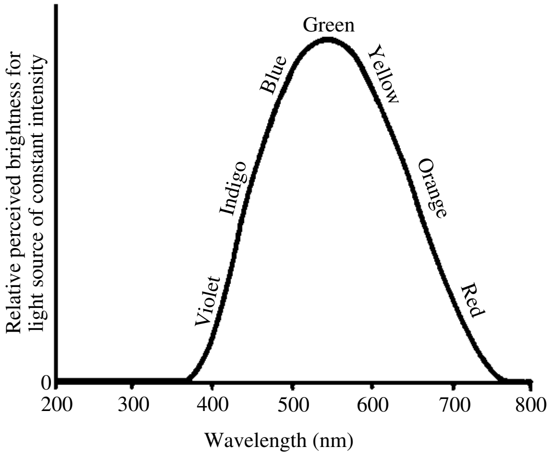

photometry

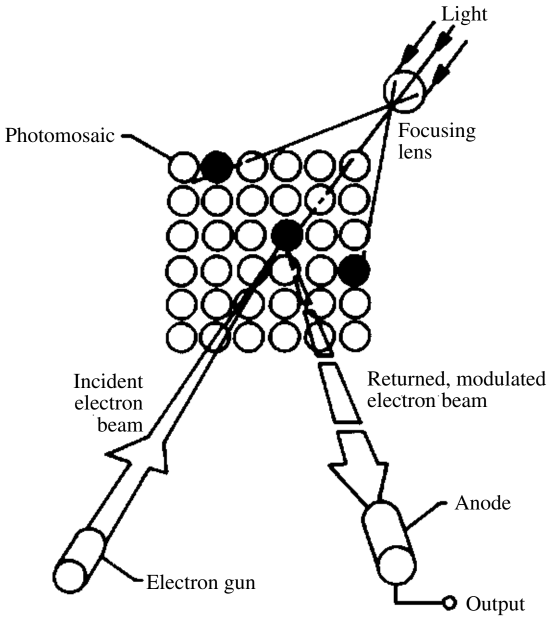

photomosaic

photoneutron A neutron released by PHOTODISINTEGRATION.

photophone **1.** A telephone-type communication system using a modulated light beam transmitted between stations. **2.** A process for recording sound on motion-picture film (see OPTICAL SOUND RECORDING).

photorelay See PHOTOELECTRIC RELAY.

photoresistive cell See PHOTOCONDUCTIVE CELL.

photoresistive material See PHOTOCONDUCTIVE MATERIAL.

photoresistivity See PHOTOCONDUCTIVITY.

photoresistor See PHOTOCONDUCTOR, **1**, **2**.

photosensitive device A light-sensitive electronic device. See, for example, PHOTOCONDUCTIVE CELL, PHOTODIODE, PHOTOFET, PHOTOMULTIPLIER TUBE, PHOTOTRANSISTOR, PHOTOTUBE, and PHOTOVOLTAIC CELL.

photosphere The luminous layer at the surface of a star.

photoswitch A light-activated switch. Some photoswitches contain an electromechanical relay; others, such as the light-activated silicon-controlled switch, have no moving parts.

phototimer An electronic timer for timing photographic processes.

phototransistor A transistor in which current carriers emitted as a result of illumination constitute an input-signal current. This current is amplified by the transistor. The output signal delivered by the transistor, accordingly, is larger than the output of an equivalent photodiode.

phototube An electron tube that converts light energy into electrical energy by acting as a light-sensitive resistor. Characteristically, the tube contains an illuminated cathode coated with a photoemissive material, and an anode wire situated nearby. Light energy causes electrons to be emitted from the cathode in amounts proportional to light intensity; the electrons are attracted by the anode, which is connected externally to a positive direct-current voltage.

photovoltaic cell Also called *solar cell.* A semiconductor diode, usually made from silicon, that converts visible light, infrared, and/or ultraviolet directly into electric current. The device consists of a flat P-N junction; the assembly is transparent so that radiant energy can fall directly on the P-type silicon. Metal ribbing, forming the positive electrode, is interconnected with tiny wires. The negative electrode is a metal backing, placed in contact with the N-type material. The component produces about 0.5 to 0.6 volts in direct sunlight under no-load conditions.

photovoltaic material A substance that generates a voltage when exposed to light. The principal substances exhibiting this effect are silicon, selenium, and germanium. Also see ACTINOELECTRIC EFFECT.

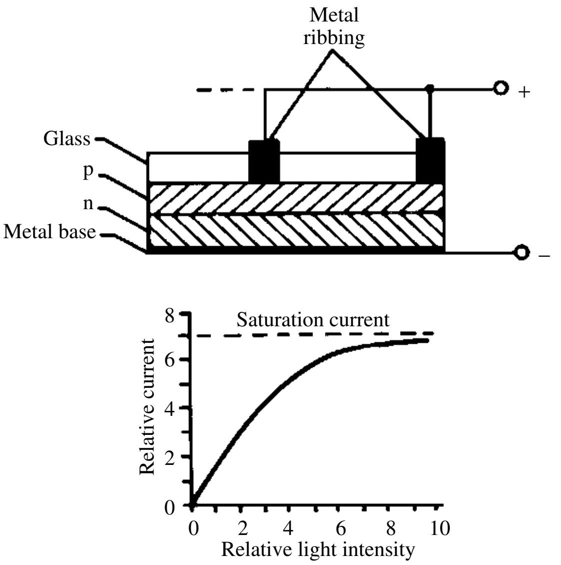

photovoltaic cell

photran A light-sensitive, four-layer semiconductor device, used for switching purposes.

physical properties The distinguishing characteristics of matter, apart from its chemical properties. Included are *boiling point*, *density*, *ductility*, *elasticity*, *electrical conductivity*, *hardness*, *heat conductivity*, *index of refraction*, *malleability*, *melting point*, *specific heat*, and *state* (solid, liquid, gaseous, or plasma).

physical quantity A quantity expressing the actual number of physical units under consideration, as compared with a dimensionless number. Examples: 50 volts, 39 kilometers, and 30 picofarads. Compare DIMENSIONLESS QUANTITY.

physics The science of energy and matter and their interactions. Physics is subdivided into several fields, including *mechanics*, *thermodynamics*, *acoustics*, *optics*, and *electricity/magnetism.* Many subdivisions are within the traditional fields.

P_i Symbol for INPUT POWER.

picket fencing An effect often observed at very-high frequencies (VHF) and ultra-high frequencies (UHF), in which movement of the transmitting station antenna, the receiving station antenna, or both antennas causes rapid fading. The fading is the result of phase effects between the direct wave and indirect wave(s). These effects are most pronounced with vertically polarized antennas.

PIA Abbreviation of PERIPHERAL INTERFACE ADAPTER.

pickoff **1.** To monitor a voltage, current, or other characteristic in an active circuit, without disturbing the operation of the circuit. **2.** A device for electronically monitoring linear or angular displacement.

pickup **1.** A device that serves as a sensor of a signal or quantity. This covers a wide variety of items, including *temperature sensors*, *vibration detectors*, *microphones*, *phonograph pickups*, etc. **2.** Collectively, energy or information that is received (e.g., *sound pickup*).

pickup arm The pivoted arm that holds the cartridge and stylus of a phonograph.

pickup cartridge See PHONO CARTRIDGE.

pickup current **1.** The current required to close a relay. **2.** Current flowing through, or generated by, a pickup.

pickup pattern The directional pattern of a microphone or other transducer that converts acoustic energy into electrical signals.

pickup voltage **1.** The voltage required to close a relay or circuit breaker. **2.** The voltage delivered by a pickup.

pico- **1.** Abbreviation, p. A prefix meaning 10^{-12}. **2.** A prefix meaning extremely small.

picoammeter A usually direct-reading instrument used to measure current in the picoampere range. Also see CURRENT METER.

picoampere Abbreviation, pA. A small unit of current equal to 10^{-12} ampere.

picocoulomb Abbreviation, pC. A small unit of electrical quantity equal to 10^{-12} coulomb.

picocurie Abbreviation, pCi. A small unit of radioactivity equal to 10^{-12} curie.

picofarad Abbreviation, pF. A small unit of capacitance equal to 10^{-12} farad.

picohenry Abbreviation, pH. A small unit of inductance equal to 10^{-12} henry.

picosecond Abbreviation, ps or psec. A small unit of time equal to 10^{-12} second.

pi coupler See COLLINS COUPLER.

picovolt Abbreviation, pV. A small unit of voltage equal to 10^{-12} volt.

picovoltmeter A usually direct-reading electronic instrument used to measure electromotive force in the picovolt range.

picowatt Abbreviation, pW. A small unit of power equal to 10^{-12} watt.

pictorial See PICTORIAL WIRING DIAGRAM.

pictorial diagram See PICTORIAL WIRING DIAGRAM.

pictorial wiring diagram A wiring diagram in the form of a drawing or photograph of the components, as opposed to one of circuit symbols. The components are shown in their positions in the finished equipment, and the wiring as lines running between them.

picture black In facsimile or television, the signal condition resulting from the scanning of a black portion of the image.

picture detector See VIDEO DETECTOR.

picture diagram See PICTORIAL WIRING DIAGRAM.

picture element See PIXEL.

picture information In a television signal, the variable-amplitude component (i.e., the one carrying energy corresponding to the picture elements) that fills the space between blanking pulses.

picture-in-picture Abbreviation, PIP. In some television receivers, a feature that allows simultaneous viewing of two programs. One program occupies the full screen, and another program appears in a small portion of the screen.

picture reception **1.** See PHOTOGRAPH RECEPTION. **2.** The reception of television signals.

picture transmission **1.** See PHOTOGRAPH TRANSMISSION. **2.** The transmission or broadcasting of television signals.

picture tube The cathode-ray tube used in a television receiver to display the image. Also called KINESCOPE.

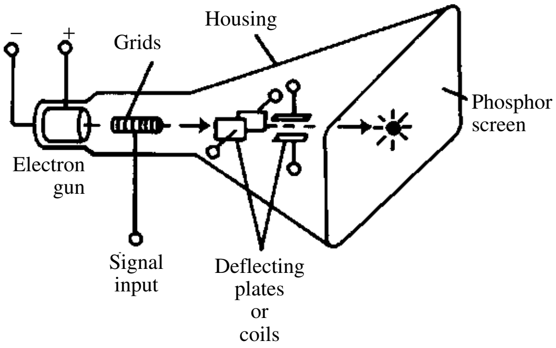

picture tube

pie chart See CIRCLE GRAPH.

Pierce oscillator A simple crystal oscillator in which the crystal is connected directly between the input and output terminals of the active device (usually a bipolar or field-effect transistor). A tuned inductance-capacitance (LC) circuit might be included, but is not required.

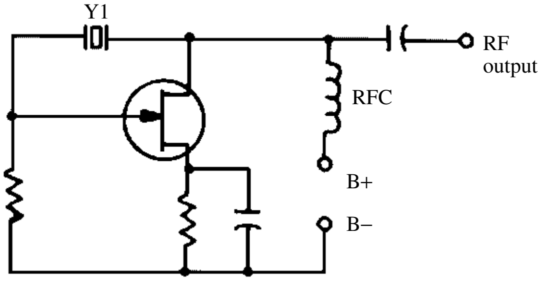

Pierce oscillator

pie winding A method of coil winding in which two or more separate, multilayer coils are connected in series and placed along a common axis. It is sometimes used in radio-frequency chokes to minimize capacitance among the windings.

piezo- A prefix meaning *pressure* (see PRESSURE, **2**).

piezodielectric A substance that, when stretched or compressed, exhibits a change in dielectric constant.

piezoelectric accelerometer An accelerometer using a piezoelectric crystal, whose voltage output is proportional to acceleration.

piezoelectric ceramic A ceramic material that delivers a voltage when deformed, or that changes in shape when a voltage is applied to it.

piezoelectric crystal A crystal (such as quartz, Rochelle salt, tourmaline, or various synthetics) that delivers a voltage when mechanical force is applied between its faces, or that changes its shape when a voltage is applied between its faces.

piezoelectric earphone See CRYSTAL EARPHONE.

piezoelectric filter See CRYSTAL FILTER and CRYSTAL RESONATOR.

piezoelectricity Electricity produced by deforming (squeezing, stretching, bending, or twisting) certain crystals, such as those of quartz, Rochelle salt, or tourmaline.

piezoelectric loudspeaker See CRYSTAL LOUDSPEAKER.

piezoelectric microphone See CERAMIC MICROPHONE and CRYSTAL MICROPHONE.

piezoelectric oscillator See CRYSTAL OSCILLATOR.

piezoelectric pickup See CRYSTAL PICKUP.

piezoelectric resonator See CRYSTAL FILTER and CRYSTAL RESONATOR.

piezoelectric sensor See CRYSTAL TRANSDUCER.

piezoelectric transducer See CRYSTAL TRANSDUCER.

piezoid A complete piezoelectric crystal device.

piezoresistance In certain substances, the tendency of the resistance to change with stretching or compression.

piezo tweeter A tweeter of the piezoelectric type (see CRYSTAL LOUDSPEAKER).

pi filter An unbalanced filter section having one series arm and two shunt arms; its schematic representation has the general shape of the uppercase Greek letter pi.

piggyback component See OUTBOARD COMPONENT.

piggyback control See CASCADE CONTROL.

piggyback tuner A separate ultra-high-frequency (UHF) television tuner operated in conjunction with the very-high-frequency (VHF) tuner of the receiver.

pigtail **1.** A usually long and sometimes flexible lead, such as the pigtail of a fixed capacitor. **2.** Descriptive of a device containing a long lead or leads, and usually mounted by such leads.

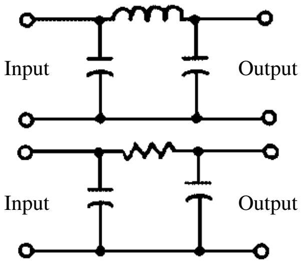

pi filters

pile **1.** See VOLTAIC PILE. **2.** See NUCLEAR REACTOR. **3.** A battery of electrochemical cells. **4.** Any packed group of particles or granules.

pillow speaker A small, flat loudspeaker intended for use under a pillow.

PILOT Acronym for *programmed inquiry learning or teaching.* A straightforward high-level computer programming language, used in computer-assisted instruction (CAI).

pilot lamp See PILOT LIGHT.

pilot light A usually small, incandescent or neon lamp. When glowing, it serves as a signal that a piece of equipment is in operation.

pilot model A preliminary model of a circuit or device constructed primarily to test the efficacy of a production process. The pilot model usually follows the PROTOTYPE.

pilot production The often small-scale production of a device in a special assembly line apart from the main line in a factory.

pilot regulator A variable-gain circuit that maintains a constant output—even if the input amplitude changes.

PIM Abbreviation of PULSE-INTERVAL MODULATION.

pi mode In a vane-anode magnetron, the mode of operation in which adjacent vanes have radio-frequency voltages of opposite polarity.

pin **1.** A semiconductor junction consisting of a layer of instrinsic semiconductor material situated between n and p layers. **2.** A slender, straight, stiff prong used as a terminal or locking device (see, for example, BASE PIN and BAYONET BASE).

pinchoff In a junction field-effect transistor, the condition in which the gate voltage causes the two depletion regions to meet and close the channel to obstruct drain-current flow.

pinchoff voltage In a junction field-effect transistor, the lowest value of gate voltage that will produce pinchoff.

pinch roller In a tape recorder, a rubber-tired, rotating cylinder that helps to pull the tape past the recording and/or playback heads.

pincushion A type of television picture distortion in which each side of the raster sags toward center screen. Also see ANTIPINCUSHIONING MAGNETS.

pincushion-correction generator A circuit for generating a deflection signal to correct pincushion distortion (see PINCUSHION). One form consists of a parabola generator and op-amp-type differentiator.

pin diode A silicon junction diode having a lightly doped intrinsic layer serving as a dielectric barrier between p and n layers.

pi network See COLLINS COUPLER.

ping An acoustic pulse; it can be audible sound or ultrasound.

pinhole **1.** A tiny hole present as a defect in a film of dielectric, semiconductor, or metal. **2.** A tiny aperture that acts as a universal lens by permitting the passage of a very small bundle of light rays. The smaller the aperture, the greater the depth of field.

pinhole detector An electronic device for finding pinholes in materials. Also see PINHOLE, **1**.

pin jack A jack into which a pin plug is inserted for quick connection.

pink noise Acoustical noise whose amplitude is inversely proportional to the frequency within a limited frequency spectrum. In the extreme, it creates a hissing sound. Compare WHITE NOISE.

pinout A diagram of an integrated circuit depicting the locations of the pins for various functions. It generally takes the form of a rectangle for the circuit itself, and short lines for the pins with designators printed next to the lines.

pin plug A plug consisting of a slender metal pin inserted between the blades of a PIN JACK for a quick connection. The plug usually has a small insulated back for convenient handling.

pin straightener A device for straightening the pins of a transistor, integrated circuit, or other electronic component.

pin switch A switch that changes state when a small pin is pushed or pulled.

pin-usage factor For an integrated circuit, the number of gate equivalents per package pin. Also see GATE EQUIVALENT.

PIO Abbreviation of *parallel input/output.*

pion A subatomic particle consisting of one quark and one antiquark.

PIP Abbreviation of PICTURE-IN-PICTURE.

pip See BLIP.

pi pad A resistance-type attenuator having a series arm, a shunt input arm, and shunt output arm; its name is derived from its resemblance to the Greek letter pi. Also see PAD.

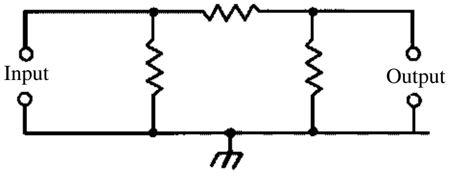

pi pad

pipe radiator A waveguide having an open end from which microwave energy is radiated.

Pirani gauge A type of vacuum gauge in which a heated filament, composing one arm of a four-arm resistance bridge, is sealed into the vacuum system. The bridge is balanced before evacuation starts. As evacuation progresses, the heat removed from the filament becomes proportional to the pressure in the system, and the resistance of the filament changes accordingly. The bridge is then rebalanced, and the difference between initial and subsequent null conditions indicate the extent of the vacuum when the bridge has been appropriately calibrated.

pi section An unbalanced filter or tuner section whose schematic representation has the general shape of the uppercase Greek letter pi.

pi-section coupling Use of a PI SECTION for coupling a radio transmitter to an antenna. Also see COLLINS COUPLER and PI-SECTION TANK.

pi-section filter A pi section used as either a low-pass or high-pass filter, depending on the position of the capacitors in the circuit.

pi-section tank A pi section used as the collector, drain, or plate tank circuit of a radio-frequency power amplifier, and also serving as an antenna coupler.

piston **1.** The movable element (cone) of a loudspeaker. **2.** The movable, solid plunger of a trimmer capacitor that consists of a plug within a cylinder.

piston directivity Directivity of sound emitted by the piston of a loudspeaker (see PISTON, **1**). As the frequency of the audio signal increases, radiation from a loudspeaker tends to be concentrated along the axis of the piston.

pit **1.** A microscopic depression in a compact disc; scatters and/or absorbs light from the laser, rather than reflecting it. Compare LAND, **1**. **2.** In a printed-circuit board, a pockmark in a component or foil run. **3.** A pockmark in a metallic substance, resulting from corrosion.

pitch **1.** The frequency of a sound, either in general terms (e.g., low, midrange, and high) or as a specific quantity (e.g., 2450 Hz). **2.** The distance between the peaks of adjacent grooves on a phonograph disc. **3.** The distance between adjacent threads of a screw. **4.** The distance between centers of turns in a coil (see PITCH OF WINDING) **5.** The number of teeth or threads per unit length. **6.** The distance along its axis a propeller moves in a revolution. **7.** Up-and-down motion of a robot end effector or other electromechanical device. **8.** The extent or range of movement, as defined in **7**.

pitch of winding In a coil, the distance from the center of one turn to the center of the adjacent turn in a single layer of winding.

PIV Abbreviation of PEAK INVERSE VOLTAGE.

pivot The sometimes jeweled, stationary member of the bearing in an analog meter movement.

pi-wound choke A choke coil consisting of several series-connected sections, mounted on a single core and separated to reduce internal capacitance.

pix Abbreviation of PICTURE.

pixel Contraction of *picture element.* The smallest bit of data in a video image. Also called *pel.* The smaller the size of the pixels in an image, the greater the resolution for a given image area.

pixel aspect ratio In a video image, the ratio of PIXEL height to pixel width.

pix tube See PICTURE TUBE.

PL Abbreviation for PRIVATE LINE.

place effect An apparent change in the perceived pitch of a sound, caused by variations in the way the waves interact inside the human ear.

planar diffusion In the production of a semiconductor device, the diffusion of all the elements into one face of a wafer. Consequently, connections to the elements all lie in one plane. Also see EPITAXIAL PLANAR TRANSISTOR and PLANAR TRANSISTOR.

planar diode A semiconductor diode, having a pn junction that lies entirely within a single plane.

planar epitaxial passivated diode A junction diode that, like the planar epitaxial transistor, has been manufactured by planar diffusion, then passivated to protect the junction. Also see EPITAXIAL GROWTH, EPITAXY, PASSIVATION, and PLANAR DIFFUSION.

PL/1 A computer programming language that is a hybrid of scientific and commercial types (like ALGOL and COBOL), the combined features being powerful problem-solving and mass-data-handling abilities.

planar epitaxial passivated transistor A planar epitaxial transistor that has been passivated to protect the exposed junctions. Also see EPITAXIAL PLANAR TRANSISTOR, PLANAR TRANSISTOR, EPITAXIAL TRANSISTOR, PASSIVATION, and PLANAR TRANSISTOR.

planar transistor A transistor in which the emitter, base, and collector elements terminate on the same face (plane) of the silicon wafer. A thin film of silicon dioxide is grown on top of the wafer to insulate the exposed junctions after the leads have been attached (i.e., the transistor is passivated).

Planck constant Symbol, *h.* Unit, joule-second. The constant of proportionality in the fundamental law of the quantum theory, stating that radiant energy is composed of quanta proportional to the frequency of the radiation; $h = q/f = 6.62608 \times 10^{-34}$ J•s, where q is the value of the quantum and f is the frequency in Hz.

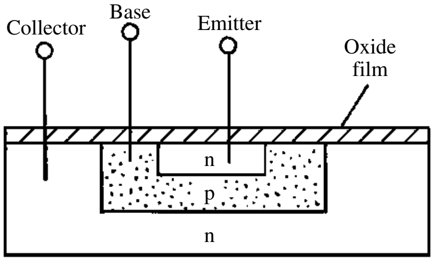

planar transistor

plane of polarization The plane containing the direction of propagation and the electric field vector of a plane-polarized wave (see POLARIZATION, **3** and POLARIZED LIGHT).

plane-polarized light See POLARIZATION, **3** and POLARIZED LIGHT.

plane-reflector antenna A directive antenna in which the reflector is a sheet of metal or a metal screen. In a corner-reflector antenna, the reflector is a folded sheet, or two sheets joined along one edge.

planetary electron See ORBITAL ELECTRON.

planimeter A mechanical instrument for measuring the area of a closed figure. The outline of the figure is traced with the pointer of the device, and the area is read from a pair of dials. In this application, the planimeter does the work of integral calculus.

PLANNER A high-level computer programming language sometimes used in artificial intelligence. It is a "goal-oriented" language in that it seeks a solution to a problem using various schemes, as necessary.

plan position indicator Abbreviation, PPI. A radar display on whose screen small spots of light reconstruct the scanned vicinity, revealing objects, such as buildings, boats, aircraft, etc. The distance from the center of the screen to a spot depicts the range of an object, and the radial angle reveals its bearing.

plant In computer operations, to put the result of an operation specified by a routine in a storage location from which it will be taken for implementation of an instruction further on in the program.

plaque resistor A flat, noninductive, power resistor, often used as a dummy load during high-frequency power measurements.

plasma A usually high-temperature gas that is so highly ionized that it is electrically conductive and susceptible to magnetic fields; it is recognized as one of the states of matter. Also see PHYSICAL PROPERTIES.

plasma diode A diode in which a plasma substance produces conduction in one direction, but not in the other.

plasma length See DEBYE LENGTH.

plasma oscillation In a plasma, a form of electric-field oscillation of the rapidly moving electrons.

plasma torch A torch, used for such high-heat applications as melting metal, in which a gas is heated by electricity to the high temperature at which it becomes a plasma.

plasmatron A form of amplifier tube sometimes used at ultra-high and microwave frequencies. It is similar to a thyratron. An inert gas is excited until it becomes a plasma, producing amplification under certain operating conditions.

Plastacele See CELLULOSE ACETATE.

plastic A synthetic material usually made from various organic compounds through polymerization (see POLYMERIZE). Plastics can be molded into solid shapes and are available as films. Examples: *celluloid*, *cellulose acetate*, *cellulose nitrate*, *polyethylene*, and *polystyrene.* Also see THERMOPLASTIC MATERIAL and THERMOSETTING MATERIAL.

plastic-film capacitor A capacitor made using polyester, polyethylene, or polystyrene. The method of manufacture is similar to that for paper capacitors when the plastic is flexible. Stacking methods can be used if the plastic is more rigid. The geometries can vary, and these capacitors are therefore found in several different shapes. Capacitance values for plastic-film units range from about 50 pF to several tens of microfarads. Most often they are in the range of 0.001 µF to 10 µF. Plastic-film capacitors are employed in audio equipment, and also in wireless transmitters and receivers. The efficiency is good, although not as high as that for mica-dielectric units. Compare CERAMIC CAPACITOR, ELECTROLYTIC CAPACITOR, MICA CAPACITOR, PAPER CAPACITOR, TANTALUM CAPACITOR.

plasticizer A substance added to a plastic to make it softer or more flexible.

plastic-leaded chip carrier Abbreviated PLCC. A surface-mounted package for an integrated circuit. It is small in size and has high electrical and mechanical reliability.

plate **1.** The anode of an electron tube. **2.** One of the electrodes of a primary or secondary battery cell. **3.** One of the electrodes of a capacitor.

plateau In a response curve, a region in which an increase in the independent variable produces no further change in the dependent variable. Example: the saturation region in a common-base transistor collector-current curve.

plate blocking capacitor A capacitor connected between the plate of an electron tube and the plate tank. It allows the direct-current supply voltage to be applied directly to the plate without it passing through the tank coil, while at the same time preventing the tank coil from short-circuiting the plate power supply. The capacitor freely transmits alternating-current signal energy to the tank.

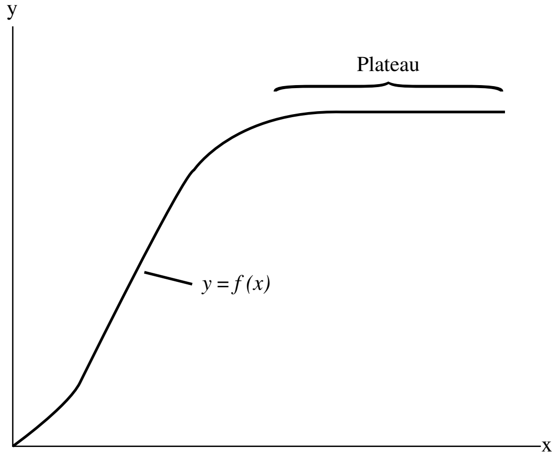

plateau

plate capacitance See PLATE-CATHODE CAPACITANCE.

plate-cathode capacitance Symbol, C_{PK}. Unit, pF. The internal capacitance between the plate and cathode of an electron tube. Also called OUTPUT CAPACITANCE.

plate characteristic For an electron tube, the family of plate current-vs-plate voltage curves for various grid-bias voltages.

plate circuit The external circuit associated with the plate of an electron tube.

plate-circuit relay A direct-current relay operated in series with the plate of an electron tube.

plate conductance Symbol, g_P. Unit, siemens. Conductance of the internal plate circuit of an electron tube. The value of static g_P is equal to the plate current divided by the plate voltage (I_P/E_P). The value of dynamic g_P is equal to the derivative of the static g_P: d_{IP}/d_{EP}. Plate conductance is the reciprocal of PLATE RESISTANCE.

plate current Symbol, I_P. Direct current flowing in the plate circuit of an electron tube.

plate-current shift A change in the plate current of a radio-frequency power amplifier during amplitude modulation. The action discloses faulty operation because the average plate current should remain constant during modulation.

plate dissipation Abbreviation, *PD.* Unit, watt. Power expended in the plate of an electron tube. For an unloaded tube, PD = $E_P I_P$, where E_P is the direct-current (dc) plate voltage in volts, and I_P is the dc plate current in amperes. For a loaded tube, $PD = P_o - P_i$, where P_o is the alternating-current (ac) power output of an amplifier or oscillator in which the tube operates, and P_i is the dc plate power input.

plated magnetic wire A wire with a ferromagnetic outer coating on a core that is not magnetic.

plated-wire memory See WIRE MEMORY.

plate-grid capacitance Symbol, C_{PG} or C_{GP}. Unit, pF. The internal capacitance between the plate and control grid of an electron tube. Also called INTERELECTRODE CAPACITANCE and FEEDBACK CAPACITANCE.

plate load The power-consuming load into which the plate circuit of an electron tube operates. In an intermediate stage of a multistage amplifier, this load is the grid circuit of the following tube.

plate load impedance Symbol, Z_{LP}. Unit, ohm. In a tube circuit, the (output) impedance that is ac-connected between the plate and ground, or dc-connected between the plate electrode and dc plate power supply.

plate meter A direct-current ammeter or milliammeter that indicates the plate current of an electron tube.

plate modulation A method of AMPLITUDE MODULATION in which a modulating voltage is superimposed on the direct-current plate voltage of a higher-frequency amplifier or oscillator.

platen The "roller" in a teletypewriter or impact printer. It supports the paper against impact by the print head; it also helps to move the paper through the machine.

plate-neutralized amplifier A vacuum-tube radio-frequency power amplifier in which a neutralizing capacitor is connected between the control grid and the free end of a center-tapped plate-tank coil.

plate power Symbol, P_P. Unit, watt. Power in the plate circuit of an electron tube; $P_P = E_P I_P$, where E_P is the plate voltage in volts and I_P is the plate current in amperes.

plate power input See PLATE POWER.

plate power output The output signal power delivered by the plate circuit of an electron tube. Compare PLATE POWER INPUT.

plate power supply The (usually direct current) power supply that furnishes energy to the plate of an electron tube.

plate pulse modulation A method of obtaining pulse modulation by injecting high-voltage pulses into the plate circuit of a vacuum-tube amplifier.

plate relay A relay operated in series with the plate of an electron tube.

plate resistance Symbol, r_P. Unit, ohm. Resistance of the internal plate circuit of an electron tube. The static value of r_P is equal to E_P/I_P, where E_P is the plate voltage in volts, and I_P is the plate current in amperes. The dynamic value is dE_P/dI_P.

plate saturation In an electron tube, the point at which, while plate voltage is increasing, the plate attracts all the electrons emitted by the cathode (i.e., the point beyond which no further significant increase in plate current results from a further increase in plate voltage).

plate-screen capacitance Symbol, C_{PS}. Unit, pF. The internal capacitance between the plate and screen grid of an electron tube.

plate-screen modulation A method of AMPLITUDE MODULATION in which the modulating voltage is superimposed simultaneously on the direct-current plate and screen voltages of a higher-frequency amplifier or oscillator.

plate series compensation In an audio amplifier, the use of a plate decoupling circuit to obtain a fixed amount of bass boost.

plate shunt compensation The addition of a network to the plate-output circuit of a tube to boost the bass response of an amplifier.

plate spacing **1.** The distance between plates in a fixed capacitor. This dimension is usually the same as dielectric thickness. **2.** The distance between plates in a variable capacitor. Also called *capacitor air gap.*

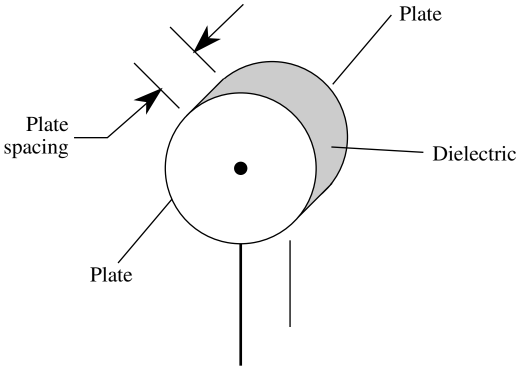

plate spacing, 1

plate supply voltage Symbol, E_{BB}. The output voltage of a plate power supply.

plate tank A resonant inductance-capacitance (LC) circuit operated from the plate of an electron tube.

plate tank capacitance The capacitance required to tune a PLATE TANK to resonance.

plate tank inductance The inductance of the coil in a PLATE TANK.

plate tank *Q* Figure of merit (see *Q*) of a plate tank, a function of the load resistance and the tank inductance-to-capacitance (L/C) ratio.

plate tank voltage The audio-frequency or radio-frequency voltage developed across the plate tank of an electron-tube circuit.

plate tuning Tuning an electron-tube circuit by varying the capacitance, inductance, or both in the plate tank.

plate tuning capacitance See PLATE TANK CAPACITANCE.

plate tuning inductance See PLATE TANK INDUCTANCE.

plate-type capacitor A capacitor having flat metal plates, rather than concentric cylinders, a cylinder and rod, etc.

plate voltage Symbol, E_P. The direct-current voltage applied to the plate of an electron tube.

plate winding **1.** An inductor connected in series between the plate of a vacuum tube and the positive power-supply voltage. **2.** The primary winding of a plate-circuit output transformer.

platiniridium A natural alloy of PLATINUM and IRIDIUM.

platinotron A form of traveling-wave vacuum tube used as an amplifier at ultra-high and microwave frequencies. There are two output connections.

platinum Symbol, Pt. A precious metallic element. Atomic number, 78. Atomic weight, 195.08. It is sometimes used for plating of relay and switch contacts, and for certain parts of vacuum tubes.

platinum metals The rare metals IRIDIUM, OSMIUM, PALLADIUM, PLATINUM, RHODIUM, and RUTHENIUM. They do not react readily with other elements.

platinum-tellurium thermocouple A thermocouple using the junction between platinum and tellurium wires; it is used in thermocouple-type meters.

platter **1.** One of the individual disks in a computer HARD DISK drive. **2.** The rotating turntable in a PHONOGRAPH.

playback The reproduction of recorded material in audio-tape, audio-disc, video-tape, or video-disc systems.

playback computer system A personal computer and associated peripherals, equipped for reproducing multimedia that has been recorded on CD-ROM.

playback head In a magnetic recorder/reproducer, the head that picks up the signal from the tape or disc for reproduction. Also called *read head* and *play head.*

playback loss In disc recording, the difference (at a particular point on the disc) between the recorded level and the reproduced level.

player A semiconductor layer that is doped to provide current carriers that are predominantly holes. Compare N LAYER.

player A device or system that reproduces (plays back) data from a tape or disc, but cannot be used to record data onto the tape or disc.

play head See PLAYBACK HEAD.

playthrough The condition in which an amplifier delivers a small output signal when the gain control is set to zero.

PLC Abbreviation of POWER-LINE COMMUNICATION.

PLCC Abbreviation of PLASTIC-LEADED CHIP CARRIER.

plethysmograph A medical-electronic device that allows the monitoring of the amount of blood in different parts of the body.

PLL Abbreviation of PHASE-LOCKED LOOP.

PLM Abbreviation of *pulse-length modulation* (see PULSE-DURATION MODULATION).

PLO Abbreviation of PHASE-LOCKED OSCILLATOR.

plot **1.** A curve depicting the variations of one quantity, with respect to another. **2.** To generate, print, or display a curve of the type defined in **1**.

plotter A machine that plots (see PLOT, **2**) automatically, often by the direction of a computer.

PL tone encoder An audio oscillator and modulator that cause a subaudible-tone modulation of a signal for use in restricted communications systems.

plug A usually male quick-connect device that can be inserted into a JACK to make a circuit connection, or be pulled out of the jack to break the connection. See, for example, MALE PLUG, PHONE PLUG, POWER PLUG, and POLARIZED POWER PLUG.

plug-and-jack connection A connection made by inserting a PLUG into a JACK.

plug fuse A fuse provided with an Edison base for screwing into a socket.

pluggable Capable of being completely removed from the rest of the system without the need for removing any wiring. Pluggable components and circuit boards simplify the servicing of electronic equipment.

plug-in See PLUG-IN COMPONENT and PLUG-IN UNIT.

plug-in capacitor A capacitor with pins or ferrules that can be quickly inserted into, or removed from, a socket.

plug-in coil A coil wound on a form having pins that can be quickly inserted into, or removed from, a socket.

plug-in coil form An insulating form with base pins that mate with socket terminals so that a coil wound on the form can be quickly inserted into, or removed from, a circuit.

plug-in component A component or module, such as a transistor, capacitor, coil, lamp, etc., provided with pins, clips, or contacts for easy insertion into, or removal from, a circuit. See, for example, PLUG-IN CAPACITOR, PLUG-IN COIL, PLUG-IN FUSE, PLUG-IN LAMP, PLUG-IN METER, PLUG-IN RESISTOR, PLUG-IN TRANSFORMER, and PLUG-IN UNIT.

plug-in fuse A cartridge fuse having a metal ferrule on each end for insertion into a matching clip for easy installation and removal.

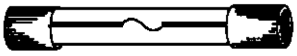

plug-in fuse

plug-in lamp A lamp with base pins for quick insertion into, or removal from, a socket.

plug-in meter A meter with pins or banana plugs for quick insertion into, or removal from, a circuit.

plug-in resistor A resistor with pins or ferrules for quick insertion into, or removal from, a socket.

plug-in transformer A small transformer with pins for quick insertion into, or removal from, a socket.

plug-in unit A unit, such as a tuned circuit, amplifier, or meter, that has pins or contacts for easy insertion into, or removal from, a larger piece of equipment.

plumber's delight An antenna whose construction, including that of the mast, is entirely of metal rods or tubing, with no insulating parts. Short circuits and grounds are prevented by making all attachments and joints at points that are at zero voltage, with respect to the standing-wave pattern.

Plumbicon A television camera tube, similar to the VIDICON, with a lead-oxide target. It is noted for high sensitivity. The image lag time is shorter than in the conventional vidicon.

plumbing Collectively, the waveguides, tees, elbows, and similar pipelike devices and fixtures used in microwave setups.

plunger-type meter A meter in which an iron or steel plunger is pulled into a coil by the magnetism produced by a current flowing in the coil. The plunger is attached to a pointer that moves over the scale.

plutonium Symbol, Pu. A radioactive metallic element that is artificially produced. Atomic number, 94. Atomic weight, approximately 244.

PM **1.** Abbreviation of PERMANENT MAGNET. **2.** Abbreviation of PULSE(D) MODULATOR. **3.** Abbreviation of *post meridian.* **4.** Abbreviation of PHASE MODULATION.

Pm **1.** Symbol for PROMETHIUM. **2.** Abbreviation of PETAMETER.

Pm Symbol for MAXIMUM POWER.

PME Abbreviation of *photomagnetoelectric.*

PMG Abbreviation of PERMANENT-MAGNET GENERATOR.

PMM Abbreviation of PERMANENT-MAGNET MAGNETIZER.

PMOS Abbreviation of P-CHANNEL METAL-OXIDE SEMICONDUCTOR.

PMU Abbreviation of *portable memory unit.*

PN **1.** Abbreviation of POLISH NOTATION. **2.** Abbreviation of POSITIVE-NEGATIVE (often lowercase).

pn Abbreviation of POSITIVE-NEGATIVE.

pn boundary See PN JUNCTION.

pnip transistor A junction transistor having an intrinsic layer between an n-type semiconductor base and one of the p-type semiconductor layers.

pneumatic computer A computer that uses fluid logic [i.e., one in which information is stored and transferred by the flow of a fluid (gas or liquid) and pressure variations therein].

pn junction The boundary between p-type and n-type semiconductor materials in a single block or wafer of the materials. The junction cannot be duplicated by merely touching two pieces of material (one n-type and one p-type) together, however smooth their mating faces.

pn-junction diode A diode consisting of the junction between p-type and n-type regions in the same wafer of semiconductor material.

PNM Abbreviation of PULSE-NUMBERS MODULATION.

pnpn device See NPNP DEVICE.

pnp transistor A bipolar junction transistor in which the emitter and collector layers are p-type semiconductor material, and the base layer is n-type semiconductor material. Compare NPN TRANSISTOR.

Po Symbol for POLONIUM.

Po Symbol for OUTPUT POWER or POWER OUTPUT.

POGO Abbreviation of *polar orbiting geophysical observatory.*

point **1.** A dot indicating the place of separation between the integral and fractional parts of a number (e.g., *decimal point*). Also called RADIX POINT. **2.** A precisely defined location in three-dimensional space that has theoretically zero length, zero width, and zero depth (e.g., *focal point*). **3.** The place on a graph in any number of dimensions, at which two or more curves or coordinates intersect. **4.** A set of operating conditions for a component, device, or system (e.g., *cutoff point, operating point*). **5.** A defined condition at which some specific physical phenomenon occurs (e.g., *melting point*).

point charge An electric charge imagined to occupy a single point in space; thus, it has neither area nor volume.

point contact The point at which the sharply pointed tip of a wire or rod conductor touches a second conductor (e.g., the contact between a "cat whisker" and a semiconductor wafer).

point-contact diode A semiconductor diode having a fine wire ("cat whisker"), whose point is in permanent contact with the surface of a wafer of semiconductor material, such as germanium or silicon.

point-contact junction The pn junction electroformed under the point at which the "cat whisker" touches the semiconductor wafer in a point-contact diode or transistor.

point-contact transistor A transistor having two fine wires ("cat whiskers") that serve as the emitter and collector electrodes. The pointed tips of the wires are nearly in contact with (a few mils apart from) the surface of a wafer of semiconductor material, such as germanium. The semiconductor serves as the base electrode. This device was a predecessor of the JUNCTION TRANSISTOR.

point counter A Geiger counter tube in which the central electrode is a pointed, fine wire. Also see PROPORTIONAL COUNTER.

point defect **1.** In a semiconductor substance or piezoelectric crystal, the absence of an atom from its place in the lattice structure. **2.** The presence of an extra atom in the lattice structure.

point effect The tendency of an electrical discharge to occur more readily at a sharp point than at a blunt surface (as of an electrode).

pointer A pointed blade, stiff wire, or inscribed line on a transparent blade; it moves over a scale to indicate a setting or the value of a quantity. Also called NEEDLE.

pointer-type meter An analog meter in which a pointer moves over a calibrated scale.

point impedance **1.** The impedance observed at a given point in a circuit. **2.** In a transmission line, the intensity of the electric field divided by the intensity of the magnetic field at a given point.

points of saturation For a magnetic core, saturation as evidenced by a leveling-off of the positive and negative halves of the magnetization curve.

point source A source from which electromagnetic radiation emanates, and that appears as a geometric point from a great distance.

point mode Descriptive of cathode-ray-tube display operation (in a computer system), in which data is portrayed as plotted dots.

point-to-point communication Communication between two stations whose location can be precisely specified.

point-to-point motion A method of robot arm movement in which the device can attain only certain positions. The coordinates of each stopping point are stored in the robot controller (computer) memory.

point-to-point station A radio station that provides POINT-TO-POINT COMMUNICATION.

point-to-point wiring A method of wiring an electronic circuit in which wires are run directly between the terminals or components, usually by the shortest practicable route. It is used mainly in high-voltage circuits, such as power amplifiers. Compare CABLED WIRING.

poise The cgs unit of absolute viscosity; 1 poise is the absolute viscosity of a fluid that requires a shearing force of 1 dyne to move a 1-sq-cm area of one of two parallel layers of the fluid (1 cm apart) with a velocity of 1 cm per second, with respect to the other layer. The comparable SI unit is the *newton-second per meter squared* ($N \cdot s/m^2$); 1 poise = $0.1\ N \cdot s/m^2$.

polar axis **1.** In a crystal, the axis of rotation not perpendicular to a reflection plane. **2.** The straight line connecting epicenters of electric, magnetic, or gravitational poles in a system. **3.** The axis about which the earth or another planet rotates.

polar coordinate conversion See POLAR COORDINATE TRANSFORMATION.

polar-coordinate geometry A two-dimensional system for movement of industrial-robot arms. It is based on a system of POLAR COORDINATES, in which a radius and a direction angle are assigned to each point in the working plane.

polar coordinates The magnitude and direction of a vector in a defined plane, listed as a radius (magnitude) in combination with an angle (direction) between the vector and the polar axis.

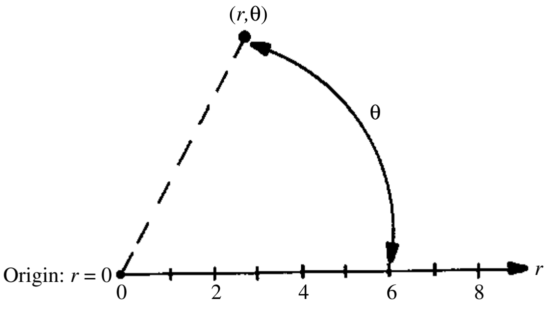

polar coordinates

polarimeter An instrument for measuring the amount of polarized light in a ray that is only partially polarized.

polariscope An instrument used in the observation or testing of materials under POLARIZED LIGHT.

polarity **1.** The condition of being electrically positive or negative. **2.** The condition of being magnetically north or south. **3.** The orientation of the positive and negative poles in a battery or power supply relative to a circuit. **4.** The orientation of a magnetic field, relative to the surrounding environment.

polarity blanking See POLARITY INHIBIT.

polarity inhibit In some instruments, especially those having automatic polarity, the automatic blanking of the polarity sign.

polarity-sensitive relay A direct-current relay actuated only when coil current flows in one direction. One of the simplest versions is a relay having a semiconductor diode connected in series with its coil.

polarity shifter A potentiometer connected to two direct-current sources so that a pair of output terminals has plus and minus polarities at one extreme of potentiometer adjustment, and minus and plus at the other extreme. At the center of the range, the output voltage is zero.

polarity switch A double-pole, double-throw switch connected between a pair of direct-current input terminals so that the polarity of a pair of output terminals can be interchanged.

polarization **1.** In a radio wave, the orientation of the electric lines of flux, with respect to the surrounding environment (e.g., *horizontal polarization* and *vertical polarization*). **2.** The disabling of a battery cell by the formation of insulating gas on one of the plates. **3.** The condition in which transverse waves of light are confined to a specific (e.g., horizontal or vertical) plane.

polarization diversity A form of reception in which two separate receivers, tuned to the same signal, are connected to independent antennas. One antenna is vertically polarized and the other is horizontally polarized. The result is a reduction in fading caused by ionospheric effects on the polarization of the incoming signal.

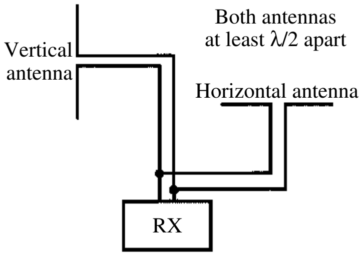

polarization diversity

polarization error In the operation of a loop antenna (e.g., that of a direction finder), null error caused by waves arriving with polarization opposite that of the loop (thus, vertically polarized waves at a horizontal-plane loop, and vice versa).

polarization fading In radio reception, a form of fading that results from changes in the polarization of the arriving signal with respect to the receiving antenna. When the polarization of the arriving signal coincides with that of the receiving antenna, the received signal strength is maximum. When the received-signal polarization is at right angles to the receiving antenna, the signal strength is minimum.

polarization modulation A method of impressing information on a signal by changing the polarization of the radiated electromagnetic field.

polarization selectivity For a photoemissive surface, the condition in which the ratio of photocurrents for two different angles of plane polarization of the light incident to the surface differs from the ratio of the corresponding amounts of light absorbed by the surface. Also see PHOTOEMISSION; PHOTOEMISSIVE MATERIAL; and POLARIZATION, **3**.

polarized capacitor A conventional electrolytic capacitor, so called because one particular terminal must be connected to the more positive of the two connection points. Compare NONPOLARIZED ELECTROLYTIC CAPACITOR.

polarized light Visible light waves whose electric lines of flux are confined more or less to a single plane. This effect can be obtained via filtering; it also occurs naturally under certain conditions. Scattered sunlight is polarized to some extent by the atmosphere. Light is polarized to some extent when it reflects from a plane surface. Also see POLARIZING FILTER.

polarized plug A plug that can be inserted into a socket or receptacle in only one way to ensure safe and foolproof operation.

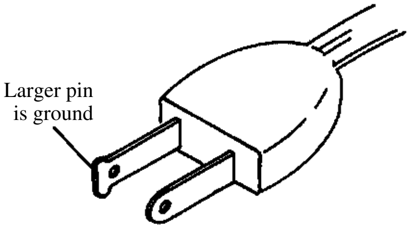

polarized plug

polarized power plug A polarized plug for connection of equipment to an alternating-current utility power source.

polarized reactor A saturable reactor in which the lines of flux produced in the three-leg core by the coils on the two outer legs are added in the center leg. Consequently, the flux can reinforce or oppose the controlling current in the coil on the center leg, depending on the direction of the current. Compare NONPOLARIZED REACTOR.

polarized receptacle A receptacle constructed so that it can receive a plug in only one way, thus preventing incorrect connections.

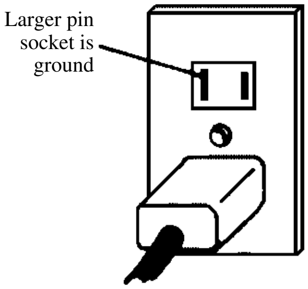
Larger pin socket is ground

polarized receptacle

polarized relay A relay actuated by one polarity of direct current, or by one particular phase of alternating current. Such a relay sometimes contains an armature-centering permanent magnet.

polarized socket See POLARIZED RECEPTACLE.

polarized X rays X-ray waves whose electric lines of flux are more or less oriented in a specific plane, as when they are scattered by carbon blocks. See, for illustration, POLARIZATION, **3** and POLARIZED LIGHT.

polarizing filter A light filter that consists essentially of microscopic heraphathite plastic plate. The plate can be rotated to cause light passing through it to be polarized in any particular plane between horizontal and vertical, depending on the degree of rotation. Also see POLARIZED LIGHT.

polarography In chemistry, a form of qualitative or quantitative analysis utilizing IE curves obtained when the voltage is gradually increased across a solution in ELECTROLYSIS.

polar orbit An orbit that carries a satellite over the geographic polar regions of the earth. This type of orbit is oriented at, or nearly at, 90 degrees with respect to the equator, and can have a period ranging from about 90 minutes to several weeks or even months. Low-earth-orbit (LEO) satellites generally have such orbits.

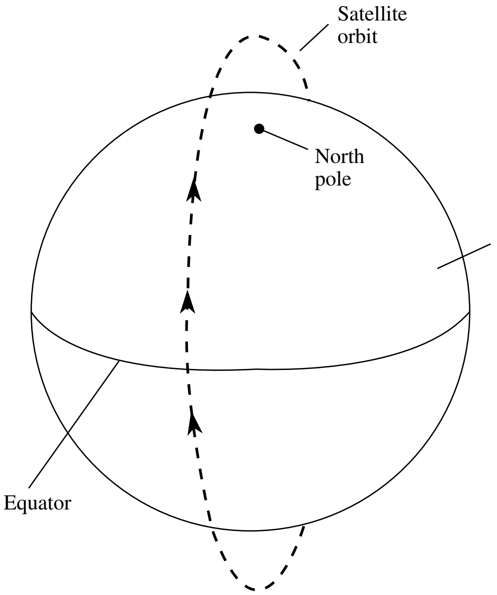

polar orbit

polar-orbiting satellite Any satellite in a POLAR ORBIT.

polar planimeter See PLANIMETER.

polar relay See POLARIZED RELAY.

polar response The horizontal-plane directional response of an antenna or other transducer.

pole **1.** An extremity or terminus that possesses POLARITY. Examples: *magnetic pole* and *electric pole.* **2.** The movable member of a switch. **3.** One of the frequencies at which a transfer function becomes infinite.

pole face The smooth end surface of a pole piece.

pole piece **1.** A section of specially shaped iron or steel that is attached to a magnetic core. **2.** Half of a two-piece magnetic core that terminates in a pole.

pole shoe In an electric motor, the section of the field pole nearest the armature.

poles of impedance For a reactive network, the frequencies at which the impedance is infinite. Compare ZEROS OF IMPEDANCE.

poles of network function The values at which a network function is infinite. Compare ZEROS OF NETWORK FUNCTION.

poles of transfer function The frequencies at which a transfer function becomes infinite. Compare ZEROS OF TRANSFER FUNCTION.

police robot A proposed robotic machine, either autonomous or remotely controlled (teleoperated), that would be used to assist in law enforcement, particularly in dangerous operations.

poling The deliberate adjustment of electromagnetic-field polarity.

Polish notation In Boolean algebra, a form of notation wherein the variables in a statement are preceded by the operators.

polling In data transmission, a technique in which channels being shared by more than one terminal are tested to find one over which data is coming in, or to ascertain which is free for transmission.

polonium Symbol, Po. A radioactive metallic element. Atomic number, 84. Atomic weight, approximately 209.

polychromatic radiation **1.** Visible-light radiation having more than one wavelength; in particular, covering a broad range of wavelengths. **2.** Electromagnetic radiation over a broad band of wavelengths.

polycrystalline material A substance, such as a semiconductor, of which even a very small sample consists of a number of separate crystals bound tightly together. Compare SINGLE-CRYSTAL MATERIAL.

polydirectional microphone See OMNIDIRECTIONAL MICROPHONE.

polyelectrolyte An ELECTROLYTE having high molecular weight.

polyester A resin made by reacting a dihydroxy alcohol with a dibasic acid.

polyester backing A polyester tape on the surface of which iron oxide is deposited to yield a magnetic recording tape.

polyethylene A plastic insulating material. Dielectric constant, 2.2. Dielectric strength, 585 V/mil.

polyethylene disc A phonograph disc made of polyethylene.

polygonal coil A coil wound on a form having a polygonal, rather than circular cross section. Some polygonal forms have as many as 12 sides in cross section.

polygraph An instrument for measuring and recording electrical signals proportional to blood pressure, skin resistance, breathing rate, and other reactions that vary under emotional stress. Also called LIE DETECTOR. Compare PATHOMETER.

polymer A compound that is the product of *polymerization*, resulting from the chemical union of monomers. Also see POLYMERIZE.

polymerize To unite monomers or polymers of the same kind to form a molecule having a higher molecular weight.

polyphase In alternating-current circuits, pertaining to the existence or generation of two or more specific electrical phases. Compare SINGLE PHASE.

polyphase antenna An antenna consisting of two dipole radiators mounted perpendicular to each other at their midpoint and excited 90 degrees out of phase. The radiation pattern is approximately circular in the plane of the elements. Also called TURNSTILE ANTENNA.

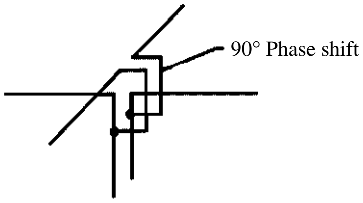

polyphase antenna

polyphase generator **1.** A dynamo-type generator of polyphase power (two-phase, three-phase, etc.). **2.** See POLYPHASE OSCILLATOR.

polyphase oscillator An oscillator circuit that generates polyphase alternating current. The circuit contains separate oscillators for each phase. A three-phase circuit, for example, has three symmetrical oscillators with matched inductance and capacitance values.

polyphase power **1.** The total dissipated power in a polyphase alternating-current circuit. **2.** Polyphase alternating current provided for utility purposes.

polyphase rectifier A rectifier of polyphase alternating current generally obtained from a three-phase power line through a transformer. The several common circuits usually contain a diode for each phase. Such rectifiers offer the advantage of higher ripple frequency than is obtainable by single-phase operation. For a three-phase rectifier, for example, the ripple frequency is three times the line frequency; for a six-phase rectifier, it is six times the line frequency.

polyphase system An alternating-current circuit in which voltages or currents are normally out of phase with each other by some fixed amount. Familiar types are two-phase and three-phase.

polyphase transformer An alternating-current transformer specifically designed for use in circuits that have two or more simultaneous current phases.

polypropylene A plastic material commonly used as an electrical insulator. Dielectric constant, 2.0. Dielectric strength, 600 V/mil.

polyrod antenna A tapered dielectric antenna, usually made of polystyrene, for directional microwave transmission.

polysilicon A polycrystalline form of silicon (see POLYCRYSTALLINE MATERIAL).

polystyrene A clear, colorless thermosetting-type plastic. It is widely used as an insulating material in radio-frequency circuits and, to some extent, as a dielectric film in fixed capacitors. Dielectric constant, 2.4 to 2.9. Dielectric strength, 20 to 28 kV/mm.

polystyrene capacitor A high-Q capacitor in which the dielectric film is polystyrene.

polyvinyl chloride Abbreviation, PVC. A plastic insulating material. Dielectric constant, 3.6 to 4.0. Dielectric strength, 800 V/mil.

pool cathode In an industrial electron tube, a cathode consisting of a pool of mercury.

pool-cathode tube An industrial electron tube using a pool cathode. Examples: excitron, ignitron, and mercury-arc rectifier.

popcorn noise A temperature-dependent, random-shot electrical noise. In a radio receiver or audio circuit, this noise resembles the sound of popping corn. It occurs in some operational amplifiers.

population In statistical analysis, the total group of items, quantities, or values under consideration. Sometimes called *universe.*

porcelain A hard, white, usually glazed ceramic used as a dielectric and insulant. Dielectric constant, 6 to 7.5. Dielectric strength, 40 to 100 V/mil. Also called *china.*

porcelain capacitor A ceramic-dielectric capacitor in which the dielectric is composed of porcelain or a related substance.

porcelain insulator An electric insulator fabricated from porcelain.

porch See BACK PORCH and FRONT PORCH.

port **1.** In a circuit, device, or system, a point at which energy or signals can be introduced or extracted in a particular manner (e.g., *two-port circulator* and *I/O port*). **2.** An aperture in a loudspeaker enclosure.

portable-mobile station See MOBILE STATION.

portable station A communications station that can be carried from one location to another. A portable station differs from a mobile station in that a portable station does not usually operate while in motion, whereas a mobile station does.

ported reflex enclosure A loudspeaker cabinet with openings that facilitate bass (low-frequency) sound reproduction.

pos **1.** Abbreviation of POSITIVE. **2.** Abbreviation of POSITION.

position **1.** The location of a point or object with respect to one or more (usually fixed) references. **2.** The setting of an adjustable device, such as a potentiometer, rotary switch, or variable capacitor.

positional notation A method of representing numbers in which the number is indicated by the positions and value of the component digits. The decimal number system belongs in this category (e.g., the decimal number 1284.67 is equal to $1 \times 10^3 + 2 \times 10^2 + 8 \times 10^1 + 4 \times 10^0 + 6 \times 10^{-1} + 7 \times 10^{-2}$).

positional number system See POSITIONAL NOTATION.

positional representation See POSITIONAL NOTATION.

position-control potentiometers In an oscilloscope, potentiometers used to control the voltage applied to the horizontal and vertical deflecting plates to position the spot on the screen. Also see CENTERING CONTROL.

position controls See POSITION-CONTROL POTENTIOMETERS.

position feedback In a servo or other control system, feedback current, or voltage that is proportional to the position assumed by a member.

position fixing The determination of a position from the intersection on a map of two lines derived from the direction-finding pickups of two transmitting stations. Also see DIRECTION FINDER.

position indicator In a tape recorder, a counter whose numbered wheels revolve when the reels do, thus aiding in locating a desired spot on the tape. Also called *tape counter.*

positioning circuit The circuit associated with a horizontal or vertical centering control (see CENTERING CONTROL).

position sensing **1.** In robotics and navigation systems, a method of determining location, relative to the surrounding environment. **2.** Any method via which a robot can accurately determine the location(s) of its end effector(s).

position sensor An electronic circuit that detects physical displacement, and transmits a signal proportional to the displacement.

positioning control See CENTERING CONTROL.

positive **1.** Possessing positive (plus) direct-current electrical polarity. **2.** Pertaining to real numbers greater than zero. **3.** A photographic image whose shadings are the same as those in the scene.

positive angle **1.** In a system of rectangular coordinates, an angle in the first or second quadrant. **2.** In rectangular coordinates, an angle measured counterclockwise from the positive *x*-axis. Compare NEGATIVE ANGLE.

positive bias A positive voltage or current applied continuously to an electrode of a device (as to a transistor base) to maintain the device's operating point. Compare NEGATIVE BIAS.

positive bus See POSITIVE CONDUCTOR.

positive charge An electrical charge characterized by having relatively fewer electrons than a negative charge. Also see CHARGE, **1**; ELECTRIC CHARGE; and UNIT ELECTROSTATIC CHARGE. Compare NEGATIVE CHARGE.

positive conductor The conductor or line connected to the positive terminal of a current, voltage, or power source. Compare NEGATIVE CONDUCTOR.

positive electricity See POSITIVE CHARGE and POSITIVE ELECTRIFICATION.

positive electrification Electrification characterized by a deficiency of electrons. For example, when a glass rod is rubbed with a silk cloth, the rod becomes positively charged because electrons are rubbed off the glass onto the silk. Similarly, when an atom loses an electron, it becomes electrified positively because it has a deficiency of electrons. Compare NEGATIVE ELECTRIFICATION.

positive electrode **1.** An electrode connected to the positive terminal of a current, voltage, or power source. **2.** The positive terminal of a current, voltage, or power source, such as a battery or generator.

positive electron See POSITRON.

positive element See POSITIVE ELECTRODE, **1**.

positive error of measurement An error of measurement in which the difference between a measured value and the true or most probable value is positive. Compare NEGATIVE ERROR OF MEASUREMENT.

positive exponent A positive superscript indicating that a number (*x*) is to be raised to the positive *n*th power. Thus, in the expression *xn*, the value of *n* is greater than zero. Compare NEGATIVE EXPONENT.

positive feedback Feedback that is in phase with an input signal. Also called REGENERATION and REGENERATIVE FEEDBACK. Compare NEGATIVE FEEDBACK.

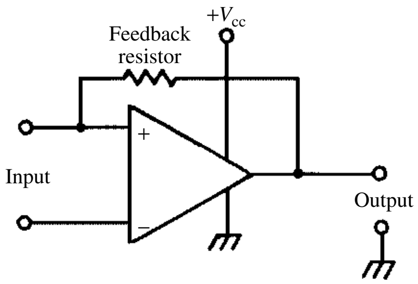

positive feedback

positive function A function having the positive sign. In the rectangular coordinate system, the trigonometric sine function is positive in the first and second quadrants, the cosine in the first and fourth, and the tangent in the first and third. Compare NEGATIVE FUNCTION.

positive ghost In a television picture, a ghost with positive shading (see POSITIVE, **3**). Also see GHOST.

positive-going Pertaining to a signal whose value is changing in a positive direction. This is not restricted to signals of actual positive polarity; a decreasing negative voltage, for example, is positive-going as it falls in the direction of zero—even if it never crosses the zero line.

positive grid In an electron tube, a control grid whose bias or signal voltage is positive, with respect to the cathode.

positive-grid oscillator A microwave oscillator circuit in which the control grid of a triode tube is operated at a positive direct-current potential, and the plate at a negative potential. Electrons move back and forth between cathode and plate, through the grid, and thus give rise to an oscillating current.

positive ground A direct-current electrical system in which the positive power-supply terminal is connected to the common ground. It is not generally used in North America.

positive half-alternation See POSITIVE HALF-CYCLE.

positive half-cycle That half of an alternating-current cycle in which the current or voltage increases from zero to maximum positive and returns to zero.

positive image **1.** A picture in which the blacks, whites, and grays correspond to those in the actual scene (see POSITIVE, **3**). **2.** A normal television picture (i.e., one that has the shading described in **1**).

positive ion An atom that has a deficiency of electrons and, consequently, exhibits a net positive charge. Also called CATION.

positive lead See POSITIVE CONDUCTOR.

positive light modulation In television transmission, the condition in which transmitted power increases as the light intensity increases. Compare NEGATIVE LIGHT MODULATION.

positive line See POSITIVE CONDUCTOR.

positive logic **1.** Binary logic in which a low positive state represents logic 0, and a high positive state represents logic 1. **2.** Binary logic in which a high negative state represents logic 0, and a low negative state represents logic 1. Compare NEGATIVE LOGIC.

positive magnetostriction A form of MAGNETOSTRICTION in which the physical size of a substance is directly proportional to the intensity of the surrounding magnetic field.

positive measurement error See POSITIVE ERROR OF MEASUREMENT.

positive modulation In amplitude-modulated television transmission, the increase in transmitted power when the brightness of the scene increases. Compare NEGATIVE MODULATION.

positive modulation factor For an amplitude-modulated wave having unequal positive and negative peaks, a ratio expressing the maximum positive deviation (increase) from the average value of the envelope. Compare NEGATIVE MODULATION FACTOR.

positive number A real number, whose value is greater than zero. Compare NEGATIVE NUMBER.

positive peak The maximum amplitude of a positive half-cycle or positive pulse.

positive-peak clipper A peak clipper that levels off the positive half-cycle of an alternating-current wave to a predetermined level.

positive-peak modulation Amplitude modulation of the positive peaks of a carrier wave.

positive-peak voltmeter An electronic voltmeter for measuring the amplitude of the positive peak of an alternating-current (ac) wave. In its simplest form, it consists essentially of a direct-current (dc) microammeter with a diode oriented to pass the positive half-cycle. A series capacitor in the circuit is charged to approximately the peak value of the applied ac voltage. Compare NEGATIVE-PEAK VOLTMETER.

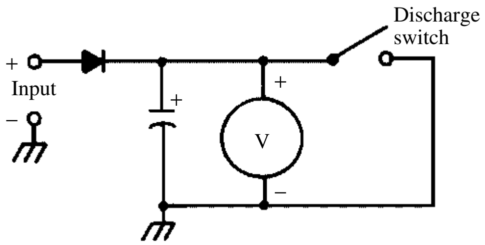

positive-peak voltmeter

positive phase-sequence relay A phase-sequence relay that responds to the positive phase sequence in a polyphase circuit. Compare NEGATIVE PHASE-SEQUENCE RELAY.

positive picture modulation See POSITIVE MODULATION.

positive picture phase In a television signal, the swinging of the picture-signal voltage from zero to positive, in response to an increase in brightness in the scene. Compare NEGATIVE PICTURE PHASE.

positive plate **1.** The positive member of an electrochemical cell or battery. Electrons flow to this plate from the negative plate, through the external circuit. **2.** A vacuum-tube plate that is biased positively, as in a conventional tube circuit.

positive pole See POSITIVE ELECTRODE, **1**, **2**.

positive potential **1.** The voltage at a positive electrode (with respect to the negative electrode). **2.** Voltage greater than that at ground as a reference.

positive power See POSITIVE EXPONENT.

positive resistance Ohmic resistance (see OHMIC RESPONSE). Compare NEGATIVE RESISTANCE.

positive resistor A resistor whose value does not change with current or voltage changes. Compare NEGATIVE RESISTOR.

positive temperature coefficient Abbreviation, PTC. A number expressing the amount by which a quantity (such as the value of a component) increases when temperature is increased. The coefficient is stated as a percentage of the rated value per degree, or in parts per million per degree. Compare NEGATIVE TEMPERATURE COEFFICIENT and ZERO TEMPERATURE COEFFICIENT.

positive transmission In facsimile or television, a form of amplitude modulation in which the picture brightness is directly proportional to the signal strength at any given instant of time.

positive valence The valence of a positive ion.

positron A positively charged particle having the same mass as that of the electron, and the same magnitude of electric charge, but positive (instead of negative). Sometimes called *positive electron.*

post See BINDING POST.

post- Prefix meaning “following,” “subsequent to,” or “behind.”

post-accelerating electrode In a cathode-ray tube, the high-voltage electrode that produces POST-DEFLECTION ACCELERATION of the electron beam. Also called INTENSIFIER ELECTRODE.

postacceleration See POST-DEFLECTION ACCELERATION.

post-alloy-diffused transistor Abbreviation, PADT. A transistor in which electrodes are diffused into the semiconductor wafer after other electrodes have been alloyed.

post-conversion bandwidth The bandwidth of a signal after it has been converted from one frequency to another.

post-deflection accelerating electrode See POST-ACCELERATING ELECTRODE.

post-deflection acceleration In a cathode-ray tube, the intensification of the electron beam following beam deflection. Also see POST-DEFLECTION CRT.

post-deflection CRT An oscilloscope tube provided with a high-voltage intensifier electrode in the form of a ring encircling the inside flare of the tube, between the deflecting plates and the screen. The deflected electron beam is accelerated by this electrode. This arrangement allows the beam to be deflected at low velocity and high sensitivity, then to be accelerated for a brighter image.

post edit The editing of data in a computer output.

postemphasis See DEEMPHASIS.

post-equalization **1.** In sound recording and reproduction, equalization during playback. Compare PREEQUALIZATION. **2.** See DEEMPHASIS.

postmortem An investigation into the cause of failure of a circuit, device, or system.

postmortem dump At the end of a computer program run, a dump to supply information for debugging purposes.

post office box A type of wheatstone bridge that contains resistance coils in a special box. The coils are connected so that they can be replaced by shorting connectors.

pot **1.** See POTENTIOMETER. **2.** See DASHPOT. **3.** Abbreviation of POTENTIAL. **4.** To encapsulate a circuit in a potting compound, such as epoxy resin.

potassium Symbol, K. A metallic element of the alkali-metal group. Atomic number, 19. Atomic weight, 39.098.

potassium chloride Formula, KCl. A compound used as a phosphor coating on the screen of a nearly permanent-persistence cathode-ray tube. The fluorescence is magenta or white, as is the phosphorescence.

potassium cyanide Formula, KCN. A highly toxic salt that is an electrolyte in some forms of electroplating.

potassium dihydrogen phosphate Abbreviation, KDP. An inorganic ferroelectric material.

pot core A magnetic core for a coil, made of ferrite or of powdered iron, consisting of a central rod, a surrounding potlike enclosure, and a lid. The rod passes through the center of the coil, and the pot and lid completely enclose the coil. This arrangement provides a completely closed magnetic circuit and coil shield.

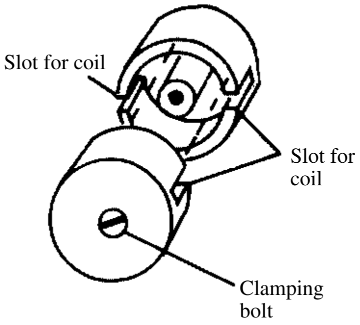

pot core

potential See ELECTROMOTIVE FORCE.

potential barrier The electric field produced on each side of a semiconductor junction by minority carriers (i.e., by holes in the n-layer and electrons in the p-layer) that face each other across the junction, but cannot diffuse across the junction and recombine.

potential coil The shunt coil in a conventional wattmeter.

potential difference See ELECTROMOTIVE FORCE and VOLTAGE.

potential divider See VOLTAGE DIVIDER.

potential drop **1.** A voltage difference between two points in a circuit. **2.** The voltage across a resistor in a direct-current circuit.

potential energy Energy resulting from the position of a body or particle (e.g., the energy stored in something lifted against gravity and held in its new position) or from the position of charges (e.g., the energy stored in a charged capacitor). Compare KINETIC ENERGY.

potential gradient See VOLTAGE GRADIENT.

potential profile A rectangular-coordinate display of the VOLTAGE GRADIENT across a body (e.g., the cross section of a transistor).

potential transformer A small step-up transformer for increasing the range of an alternating-current voltmeter.

potentiometer **1.** A variable resistor used as a voltage divider. The input voltage is applied across the entire resistance element and the output voltage is taken from the wiper, relative to one end of the element. One end is usually grounded (at zero potential). **2.** A null device whose operation is based on a variable resistor, and is used for precise voltage measurements. The unknown voltage is applied to the input of a variable resistor whose settings are known with great accuracy; the resistance is adjusted for an output voltage that exactly equals the voltage of a standard cell (as indicated by a null between the two voltages). The unknown voltage is then determined from the resistance and the standard-cell voltage.

potentiometer noise In a current-carrying potentiometer, electrical noise generated when the wiper blade rubs against the resistance element, or by contact between the blade and element.

potentiometric recorder A type of graphic recorder. It consists essentially of a resistance-calibrated potentiometer, a standard cell, and a galvanometer. When an unknown voltage (E_x) is applied to the input terminals of the potentiometer and the potentiometer is set for null, $E_x = E_s (R_2/R_1)$, where Es is the voltage of the standard cell, $R1$ is the input resistance of the potentiometer, and $R2$ is the output resistance of the potentiometer.

Potier diagram An illustration of the phase relationship between current and voltage in an alternating-current circuit that contains reactance.

POTS Acronym for *plain old telephone service*, meaning basic service without optional features (such as call waiting, conference calling, call forwarding, etc.).

potted circuit A circuit embedded in plastic or wax to protect it against the environment, and/or to minimize the effects of physical vibration (see POTTING).

potted component An electronic part embedded in a suitable plastic or wax to protect it against the environment, and/or to minimize the effects of physical vibration (see POTTING).

potting A process of embedding a component or circuit in a solid mass of plastic or wax held in a container. The process is similar to encapsulation, except that in potting, the container (envelope) remains as part of the assembly. Compare ENCAPSULATION.

potting material A substance, such as a resin or wax, used for potting electronic gear. Also called *potting compound.*

pound **1.** Abbreviations, lb, p. A unit of weight equal to 16 avoirdupois ounces. **2.** Abbreviation, lbf. A unit of force approximately equal to 4.448 newtons. **3.** Abbreviation, lbm. A unit of mass approximately equal to 0.4536 kilogram.

poundal A unit of force equal to approximately 13825.5 dynes or 0.138255 newton. One poundal is the force that, when acting for one second, will impart a speed of one foot per second to a one-pound mass.

pound-foot Abbreviation, lb-ft. A unit of torque equal to the product of a force of one pound and a moment arm of one foot. Compare OUNCE-INCH.

pounds per square inch absolute Abbreviation, psia. Absolute pressure (i.e., the sum of atmospheric pressure and the pressure indicated by a gauge). Compare POUNDS PER SQUARE INCH GAUGE.

pounds per square inch gauge Abbreviation, PSIG. The value of pressure indicated by a gauge, without correction for atmospheric pressure. Compare POUNDS PER SQUARE INCH ABSOLUTE.

powdered-iron core A magnetic core consisting of minute particles of iron, each coated with a film to insulate it from others, molded into a solid mass. Because of its low eddy-current loss, this type of core is usable in radio-frequency transformers and coils, where it increases the inductance of the winding.

power **1.** Symbol, P. Unit, watt. The rate of doing work, or producing or transmitting energy. In direct-current circuits, and in alternating-current circuits containing no reactance, power is the product of the root-mean-square current and voltage. See, for example, AC POWER, APPARENT POWER, DC POWER, KILOVOLT-AMPERE, POWER FACTOR, REACTIVE KILOVOLT-AMPERE, REACTIVE VOLT-AMPERE, TRUE POWER, VOLT-AMPERE, WATT, and WATTLESS POWER. **2.** The product obtained by multiplying a quantity x by itself n times, written xn. For example, $2^4 = 2 \times 2 \times 2 \times 2 = 8$; here, 8 is the *fourth power* of 2. Compare ROOT, **1**. **3.** The exponent in an expression, as defined in **2**.

power amplification **1.** The amplification of a signal having a certain power (wattage) to produce a signal having greater power. **2.** The signal power increase, expressed as a ratio or as a figure in decibels, resulting from the process defined in **1**. Also called POWER GAIN.

power amplification ratio See POWER AMPLIFICATION, **2** and POWER GAIN.

power amplifier An amplifier that delivers useful amounts of power to a load, such as one or more speakers. Compare CURRENT AMPLIFIER and VOLTAGE AMPLIFIER.

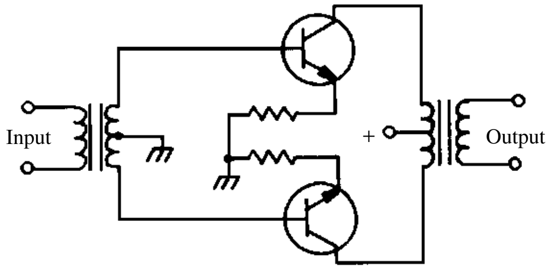

power amplifier

power-amplifier device A high-current tube or transistor designed especially for high power output. Such a device does not always provide significant voltage amplification, but always provides power amplification. Compare VOLTAGE-AMPLIFIER DEVICE.

power at peak torque Symbol, *Pp*. For a torque motor, the input power in watts needed for peak torque at stall at a winding temperature of 25 degrees Celsius.

power attenuation **1.** A reduction of power level. **2.** See POWER LOSS.

power bandwidth For a high-fidelity audio amplifier, the difference between the maximum and minimum frequencies at which the amplifier can produce at least 50 percent of its maximum power output, with less than a certain amount of total harmonic distortion (usually 10 percent).

power blackout A situation in which all electric power is lost to utility subscribers in a defined region.

power consumption **1.** For a direct-current device, the normal operating voltage multiplied by the normal drawn current. **2.** For an alternating-current circuit, the root-mean-square voltage multiplied by the root-mean-square current.

power control The adjustment of the output voltage of a power supply, usually by means of a variable autotransformer, silicon-controlled rectifier, thyratron, or similar device.

power cutoff frequency Symbol, f_{co}. The frequency at which the power gain of a transistor drops 3 dB below its low-frequency value.

power derating For a temperature higher than the specified ambient temperature, a deliberate reduction of the power dissipated by a component or device. This is done to prevent failure of the component or device. Also see DERATING, DERATING CURVE, and DERATING FACTOR.

power difference An expression of the power lost in a circuit when power is absorbed by a dielectric material.

power diode A heavy-duty diode that is usually used in power-supply service. Also called *rectifier diode.*

power dissipation Abbreviation, PD. The power consumed by a device during normal operation. This power is not available in the electrical output of the device. An example is the direct-current power dissipated in the collector circuit of a high-fidelity audio amplifier.

power divider A circuit that distributes power, in a predetermined manner, among various loads.

power drain The amount of power drawn by a device. It can be operating power or standby power.

power dump See DUMP, **2**.

power equations Variations of the basic power equation: $P = EI = E^2/R = I^2R$, where P is the power in watts, E is the voltage in volts, I is the current in amperes, and R is the resistance in ohms.

power factor Abbreviation, PF. In an alternating-current circuit, the ratio (expressed either as a decimal or a percentage) of true power (power actually consumed) to apparent power (simple product of voltage and current). The power factor is equal to the cosine of the phase angle. Also see AC POWER.

power-factor balance In a capacitance bridge, a separate null adjustment for the internal resistance component of a capacitor under measurement. The dial of the variable component for this adjustment reads directly in percent power factor in some bridges.

power-factor correction To raise the power factor of an inductive circuit by inserting a parallel capacitance. In power circuits, this affords improved economy of operation because the current drain is brought more in line with that of a resistive circuit.

power-factor meter An instrument that gives direct readings of power factor (lead or lag). One such meter uses a dynamometer-type movement (see ELECTRODYNAMOMETER) in which the rotating element consists of two coils fastened together at right angles.

power-factor regulator A device that regulates the power factor of an alternating-current line.

power-factor relay An alternating-current relay actuated by a rise or fall in power factor, with respect to a predetermined value.

power frequency **1.** See POWER-LINE FREQUENCY. **2.** The frequency of an alternating-current generator. **3.** The output frequency of a power inverter (see INVERTER, **1**).

power-frequency meter An instrument for measuring power-line frequency. It can use electromechanical devices, or can directly count the number of alternations per second.

power gain Abbreviations, PG or PG_{dB}. The extent to which power is increased by a power amplifier. It can be expressed as the ratio of power output to power input as $PG = P_o/P_i$, or in decibels as $PG_{dB} = 10 \log_{10}(P_o/P_i)$.

power grid An aggregation of power-generating stations, transmission lines, and associated equipment, usually extending over hundreds of miles and embracing several communities, so operated that individual members can deliver power to the system or draw power from it, according to local demand.

power ground The power-supply ground for a circuit or system.

power-handling capacity **1.** The amount of power that a device can dissipate, either continuously or intermittently, without suffering damage. **2.** The maximum input power that can be tolerated by an amplifier transistor or tube without overheating.

power hyperbola For a semiconductor device or vacuum tube, a curve plotted from the device's current and voltage values, which provide the power value when multiplied (e.g., a 2-watt curve for the direct-current collector input of a power transistor).

power input See INPUT POWER.

power-input control The adjustment of the output of a power supply by varying the alternating-current input to the power transformer. Usually, a variable autotransformer is operated ahead of the power transformer. See, for example, VARIABLE TRANSFORMER and VARIAC.

power inverter See CHOPPER POWER SUPPLY.

power-level indicator **1.** See DB METER. **2.** See OUTPUT POWER METER.

power line The line through which electrical energy is received by a subscriber.

power-line communication Abbreviation, PLC. Carrier-current telephony or telegraphy over power lines that are common to transmitting and receiving stations. Also see WIRED WIRELESS.

power-line filter **1.** A heavy-duty radio-frequency (RF) filter inserted in the power line close to a device that generates RF energy, such as a radio transmitter. It prevents transmission of RF energy via the power line. **2.** An RF filter inserted in the power line, where it enters the power supply of a sensitive electronic device, such as a computer or high-fidelity audio amplifier. It prevents RF energy on the power line from entering the device via the power supply.

power-line frequency The frequency of the alternating current and voltage available over commercial power lines. In the United States, it is 60 Hz; in some countries, it is 50 Hz.

power-line monitor An expanded-scale alternating-current voltmeter for the continuous monitoring of power-line voltage.

power-line pickup The interception of radio-frequency energy by utility power lines acting as receiving antennas. This energy can enter a sensitive electronic device, such as a computer or a high-fidelity audio amplifier, via the power supply.

power loss The power dissipated in a component. It generates heat while doing no useful work. Represents energy loss, except when the generation of heat is the end purpose.

power-loss factor Symbol, F_p. In interstage coupling, the ratio of available power (with the coupling network in place) to the available power when the network is disconnected.

power meter See WATTMETER.

power modulation factor In amplitude modulation, the ratio of the peak power to the average power.

power oscillator A heavy-duty oscillator delivering useful power output.

power output See OUTPUT POWER.

power-output meter See OUTPUT POWER METER.

power pack An external power-line-operated unit supplying alternating or direct current for the operation of electronic equipment.

power pentode A heavy-duty pentode vacuum tube designed to deliver relatively high output power.

power plug A plug for insertion into a power-line outlet.

power programmer A device that adjusts radar output power, in accordance with the target distance.

power rating **1.** The specified power required by an equipment for normal operation. **2.** The specified power output of a generator or amplifier.

power reactive See REACTIVE VOLT-AMPERE.

power rectifier A heavy-duty semiconductor diode used to rectify alternating current for power-supply purposes.

power relay A heavy-duty relay designed to switch significant amounts of power. The heavy contacts and armature require high actuating current; this necessitates a larger coil than is used in lighter-duty relays.

power resistor A heavy-duty resistor (i.e., one designed to carry large currents without overheating).

power stack A selenium rectifier consisting of a number of rectifier plates stacked in series for higher voltage handling.

power supply **1.** A device, such as a generator or a transformer-rectifier-filter arrangement, that produces the power needed to operate an electronic equipment. **2.** A reserve of available power (e.g., the power line, an installation of batteries, etc.).

power-supply filter A low-pass filter that is used to remove the ripple from the output of a power-supply rectifier. See, for example, BRUTE-FORCE FILTER.

power-supply rejection ratio The ratio of the output-voltage change for an amplifier, oscillator, or other circuit, to the change in power-supply voltage. It is determined on an instantaneous basis.

power-supply sensitivity In an operational amplifier, sensitivity of the offset to variations in the power-supply voltage.

power surge **1.** A momentary increase in the voltage on a utility line. **2.** An abnormally high voltage that sometimes exists for the first several milliseconds after utility power is restored following a blackout.

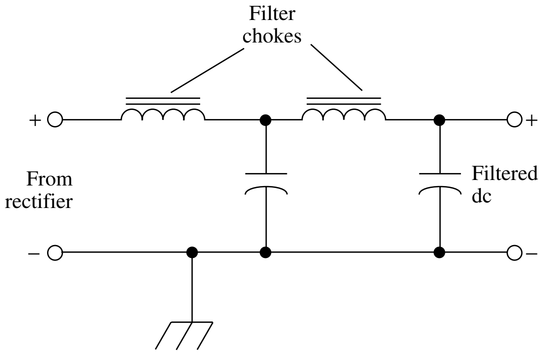

power-supply filter

power switch The switch for controlling power to a piece of equipment. Also see ON-OFF SWITCH.

power switching Switching operating power on and off. There are two principal methods: One involves making and breaking the connections between equipment and the power line; the other involves making and breaking the output of a line-operated or battery-type power supply.

power tetrode A heavy-duty tetrode vacuum tube designed to deliver relatively high output power.

power-to-decibel conversion Abbreviation, P_{dB}. Power expressed in decibels, with respect to a reference power level. Determined by the formula $P_{dB} = 10 \log_{10}(P_x/P_{ref})$, where P_x is the given power level and P_{ref} is the reference power level.

power transfer **1.** The passage of power from a generator to a load. **2.** The passage of power from one circuit to another.

power transfer theorem See MAXIMUM POWER TRANSFER THEOREM.

power transformer A transformer designed solely to supply operating power to electronic equipment—either directly or through a rectifier-filter circuit. Because a power transformer is used at low (power-line) frequencies, its core does not require the high-grade iron used in audio transformers, nor are special winding techniques needed to reduce the leakage inductance and interwinding capacitance.

power transistor A heavy-duty transistor designed for power-amplifier and power-control service.

power triode A heavy-duty triode vacuum tube designed to deliver relatively high output power.

power tube A heavy-duty electron tube designed to deliver useful amounts of power. See, for example, POWER PENTODE, POWER TETRODE, and POWER TRIODE.

power unit **1.** A power supply (see POWER SUPPLY, **1**). **2.** A unit of power measurement. See, for example, KILOWATT, MEGAWATT, MICROWATT, MILLIWATT, PICOWATT, and WATT.

power user In personal or business computing, a serious user who has extensive, up-to-date knowledge of hardware and software.

power winding In a magnetic amplifier or saturable reactor, the output winding (i.e., the winding through which the controlled current flows).

Poynting vector In an electromagnetic wave, the vector product of instantaneous electric intensity and magnetic intensity.

PP Abbreviation of *peripheral processor.*

PP **1.** Symbol for PLATE POWER. **2.** Symbol for PEAK POWER.

ppb Abbreviation of *parts per billion.*

PPI Abbreviation of PLAN POSITION INDICATOR.

pp junction In a semiconductor wafer, the boundary between two p-type regions that have somewhat different properties.

PPM Abbreviation of PULSE-POSITION MODULATION.

ppm **1.** Abbreviation of *parts per million.* **2.** Abbreviation of *pulses per minute.*

pps Abbreviation of *pulses per second.*

ppt Abbreviation of *parts per thousand.*

PPV Abbreviation of PAY PER VIEW.

Pr Symbol for PRASEODYMIUM.

practical component A circuit component considered in proper combination with the stray components inherent in it. Thus, a resistor has residual inductance and capacitance, an inductor has residual capacitance and resistance. Compare IDEAL COMPONENT.

practical units A set of physical/electrical units especially suited to a particular application. For example, in direct-current electrical applications, practical units are the AMPERE, OHM, VOLT, and WATT.

praetersonics See ACOUSTOELECTRONICS; ACOUSTIC DELAY LINE; SURFACE-WAVE AMPLIFIER; SURFACE-WAVE FILTER.

pragilbert The unit of MAGNETOMOTIVE FORCE in the absolute mks (Giorgi) system.

pragilbert per weber The unit of RELUCTANCE in the absolute mks (Giorgi) system.

praoersted The unit of MAGNETIZING FORCE in the absolute mks (Giorgi) system.

praseodymium Symbol, Pr. A metallic element of the rare-earth group. Atomic number, 59. Atomic weight, 140.908.

preaccelerating electrode In the electron gun of a cathode-ray tube, the high-voltage electrode that provides initial acceleration to the electron beam.

pre-alarm signal An audio and/or visual indicator that an alarm will sound if an area is not cleared within a short time. An example is a voice recording in a car alarm system (e.g., “Stand back!”).

preamplifier **1.** A high-sensitivity, low-noise amplifier that usually uses a high-gain field-effect transistor (FET) and is used to enhance the sensitivity of a radio communications receiver. They are especially used at frequencies above

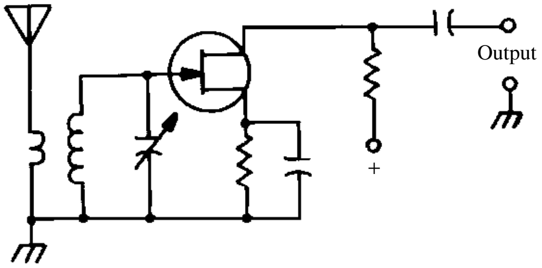

preamplifier

approximately 15 MHz, where most of the noise comes from the receiver, rather than from outside sources. **2.** A low-noise amplifier used for boosting weak signals for television reception. **3.** A low-noise, low-level amplifier used for boosting signal levels from transducers, such as microphones or photocells.

prebiased relay A relay through which is maintained a steady current that is just lower than that needed to close the relay. The actuating signal, then, need only be a small amount of additional current.

precedence effect See FUSION, **1**.

precipitation **1.** Water falling from the atmosphere in some form (rain, snow, hail, or sleet). See PRECIPITATION STATIC. **2.** The amount of precipitation occurring in a given period of time at a specific location. **3.** Separation of a solid material from a solution, as a result of a chemical or physical action.

precipitation static **1.** Radio noise that sometimes occurs when it rains or snows. It can be mistaken for artificially generated noise. **2.** Radio noise caused by atmospheric electricity arising from rain, snow, ice crystals, hail, or dust clouds, through which an aircraft carrying the radio flies.

precipitator See DUST PRECIPITATOR.

precipitron See DUST PRECIPITATOR.

precision **1.** Pertaining to electronic hardware, especially test instruments and measuring devices, designed and built to function with a high degree of accuracy. **2.** The relative accuracy of a meter or other indicating device. **3.** The accuracy of the results of an experiment, test, or measurement.

precision approach radar A radar aimed along the approach path to guide an aircraft during approach.

precision instrument An instrument possessing high accuracy and stability (i.e., one capable of reproducing readings or settings for various trials under set circumstances).

precision potentiometer **1.** A POTENTIOMETER possessing highly accurate resistance calibration, linearity, and repeatability of settings. **2.** A potentiometer-type voltage-measuring instrument.

preconduction current **1.** The cutoff current in a transistor. **2.** In a thyratron, the small (anode) current flowing before the tube is fired.

predetermined counter A counter programmed to count to a desired number and stop.

predistortion See PREEMPHASIS.

pre-Dolby **1.** To record a tape with DOLBY compression. **2.** A tape that has been recorded with Dolby compression.

preemphasis In frequency modulation, the introduction of a rising-response characteristic (response rises as modulation frequency increases). Compare DEEMPHASIS.

preequalization **1.** In sound recording and reproduction, equalization during recording. **2.** See PREEMPHASIS.

preferred values of components A number system used by the Electronics Industries Association (EIA) for establishing the values of composition resistors and small fixed capacitors.

prefix multiplier See MULTIPLIER PREFIX.

prefix notation As used with complex expressions involving many operators and operands, a type of notation in which the expressions, rather than containing brackets, are given a value, according to the relative positions of operators and operands.

preform **1.** A small wafer, usually dry-pressed from powdered plastic, from which the body of a component, such as a capacitor or resistor, is heat-molded. Also called a *pill* or *biscuit.* **2.** The preformed slab used in molding a phonograph disc. **3.** To shape a moldable circuit before fixing the final configuration or package.

preliminary information For manufactured electronic components, data that is released prior to the actual availability of the device. Subject to change when units are produced.

premix A molding compound of reinforced plastic.

prerecorded disc A phonograph disc on which a recording has been made (i.e., a recorded disc).

prerecorded tape Magnetic tape on which a program or data has been recorded. Also called RECORDED TAPE.

p region See P LAYER.

prescaler A device operated ahead of a counter to establish a new, usually higher-frequency, range over which frequency measurements can be made.

preselector A tuned or untuned radio-frequency amplifier operated ahead of a radio or television receiver to boost the sensitivity of the receiver.

presence **1.** In sound reproduction, the quality of being true to life. **2.** The effect of boosted upper-midrange frequencies in music.

preset counter A pulse counter that delivers one output pulse for a number of successive input pulses, determined by the settings of counter-circuit controls. Thus, a preset counter might give an output pulse for each train of 125 input pulses.

preset element In automation and control, an element that can be preset to a given level or value, and to which other elements can then be referred.

preset switch In the circuit of a PRESET COUNTER, a multiposition rotary switch that can be set to determine the number of input pulses that must be received for the circuit to deliver one output pulse.

preshoot A downward-moving transient pip that sometimes precedes the rise of a pulse.

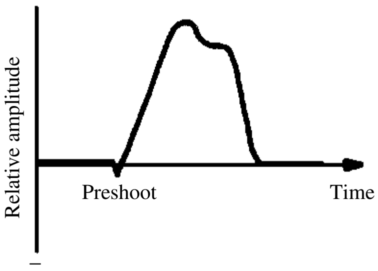

preshoot

preshoot amplitude The peak voltage of a PRESHOOT, measured from the zero line to the valley of the preshoot.

preshoot time The width of a PRESHOOT, measured along the horizontal base line (time axis).

pressing **1.** A process by which phonographic discs are fabricated from plastic. **2.** A disc pressed from plastic.

press-to-talk microphone A microphone that uses a PRESS-TO-TALK SWITCH for actuation.

press-to-talk switch A switch in a microphone or on the end of a control cord. It is used to actuate a transmitter, telephone, or recorder when the operator wishes to speak.

pressure Abbreviation, P or p. **1.** Force per unit area. It can be expressed in any appropriate units of force and area (e.g., newtons per square meter, pounds per square inch, grams per square centimeter, etc.). **2.** The application of force over part or all of a surface. **3.** Compression. **4.** See TENSION, **1**.

pressure amplitude The pressure caused by an acoustic disturbance. It is usually measured in dynes per square centimeter.

pressure capacitor An enclosed fixed or variable capacitor, whose breakdown voltage increases when the air pressure rises inside the container.

pressure contact **1.** Electrical contact made by pressing two conducting surfaces together (to complete a circuit). **2.** A contact for obtaining the condition described in **1**.

pressure-gradient microphone See PRESSURE MICROPHONE.

pressure microphone A microphone that receives sound waves at only one side of its diaphragm. This one-sided exposure results in the displacement of the diaphragm by an amount proportional to the instantaneous pressure of the sound waves.

pressure pad In a tape recorder, a small pad that holds the tape against one of the heads.

pressure pickup See PRESSURE TRANSDUCER.

pressure roller In a tape recorder, a rubber-tired roller that presses the tape against the capstan.

pressure sensor A device that detects the presence of, and/or measures, physical force within a specific area. One simple device uses two metal plates separated by electrically resistive foam. Pressure compresses the foam and reduces the resistance between the plates. This resistance change can be detected and measured.

pressure switch A switch that is opened or closed by a change in pressure within a system.

pressure transducer A sensor for converting pressure into proportionate current or voltage. Some use strain gauges; others use piezoelectric crystals, potentiometers, and other variable elements.

pressure zone A region of high air pressure that is immediately adjacent to a surface reflecting an acoustic (sound) wave.

pressure-zone microphone A microphone equipped with a deflector that helps to guide acoustic energy toward the diaphragm.

prestore To place data in memory before it is intended for use.

pretuned stage A stage, such as one in an intermediate-frequency amplifier or single-frequency receiver, that is preset to a frequency, rather than being continuously tuned.

prf Abbreviation of PULSE REPETITION FREQUENCY.

pri Abbreviation of PRIMARY.

primaries See PRIMARY COLORS.

primary **1.** See PRIMARY WINDING. **2.** See PRIMARY STANDARD.

primary battery A battery composed of primary cells.

primary block A fundamental group of channels in pulse-code modulation, combined by means of time-division multiplexing.

primary capacitance **1.** The distributed capacitance of the primary winding of a transformer whose secondary winding is unloaded. Compare SECONDARY CAPACITANCE, **1**. **2.** A series or shunt capacitance used to tune the primary coil of a radio-frequency transformer. Compare SECONDARY CAPACITANCE, **2**.

primary cell An electrochemical cell that does not require, and generally will not accept, an electrical charge in order to function. Once it has been discharged, the cell must usually be thrown away. Compare STORAGE CELL. Also see CELL, DRY CELL, and STANDARD CELL.

primary circuit **1.** The circuit associated with the primary winding of a transformer. **2.** The circuitry associated with the input to a device or system.

primary coil See PRIMARY WINDING.

primary colors See COLOR PRIMARY.

primary current The current flowing in the primary winding of a transformer. Also called TRANSFORMER INPUT CURRENT. Compare SECONDARY CURRENT.

primary electron The electron possessing the greater energy after a collision between two electrons. Compare SECONDARY ELECTRON.

primary emission Emission arising directly from a source, such as the cathode of an electron tube. Compare SECONDARY EMISSION.

primary frequency standard A device that generates unmodulated signals at precise frequencies. It generally uses a highly stable crystal oscillator that can be referred to a time standard and periodically corrected. A string of multivibrators, together with harmonic amplifiers and buffers, divide, and multiply the fundamental crystal frequency. The resulting signals provide markers for calibrating receivers and test equipment. Compare SECONDARY FREQUENCY STANDARD. Also see PRIMARY STANDARD.

primary impedance **1.** The impedance of the primary winding of a transformer whose secondary winding is unloaded. Compare SECONDARY IMPEDANCE, **1**. **2.** An external impedance presented to the primary winding of a transformer. Compare SECONDARY IMPEDANCE, **2**.

primary inductance The inductance of the primary winding of a transformer whose secondary winding is unloaded. Compare SECONDARY INDUCTANCE.

primary kVA The kilovolt-amperes in the primary circuit of a transformer. Compare SECONDARY KVA.

primary measuring element A detector, sensor, or transducer that performs the initial conversion in a measurement or control system. Such an element converts a phenomenon into a signal that can be transmitted to appropriate instruments for translation and evaluation.

primary power Power in the primary circuit of a transformer. Also see PRIMARY KVA and PRIMARY VA. Compare SECONDARY POWER.

primary radiator **1.** The driven element of a directive antenna system that incorporates parasitic elements. **2.** The driven element of a directive antenna that uses a reflector, such as a screen or dish.

primary resistance The direct-current resistance of the primary winding of a transformer. Compare SECONDARY RESISTANCE.

primary standard A usually stationary source of a quantity (e.g., capacitance, frequency, time, inductance, resistance, etc.). This source is so precise, and is maintained with such care, that it can be used as a universal reference. Compare SECONDARY STANDARD.

primary turns Symbol, N_p. The number of turns in the primary winding of a transformer. Compare SECONDARY TURNS.

primary utilization factor Abbreviation, UF_p. For a transformer in a rectifier circuit, the ratio of direct-current power output to primary volt-amperes. Numerically, the primary utilization factor is higher than the secondary utilization factor, but is less than 1. Also see SECONDARY UTILIZATION FACTOR and UTILITY FACTOR.

primary VA The volt-amperes in the input circuit of a transformer. Compare SECONDARY VA.

primary voltage The voltage across the primary winding of a transformer. Also called *transformer input voltage.* Compare SECONDARY VOLTAGE.

primary winding The normal or usual input winding of a transformer. Also called *primary coil.* Compare SECONDARY WINDING.

prime meridian See ZERO MERIDIAN.

prime mover A machine, such as a gas engine, steam engine, or water turbine, that converts a natural force or material into mechanical power.

primitive oscillation period In a complex oscillation waveform, the shortest period for which a definite repetition occurs; the highest fundamental frequency.

principal axis The line passing through the center of the spherical part of a lens, mirror, or dish reflector.

principal focus The focal point of rays arriving parallel to the principal axis of a lens, mirror, or dish reflector.

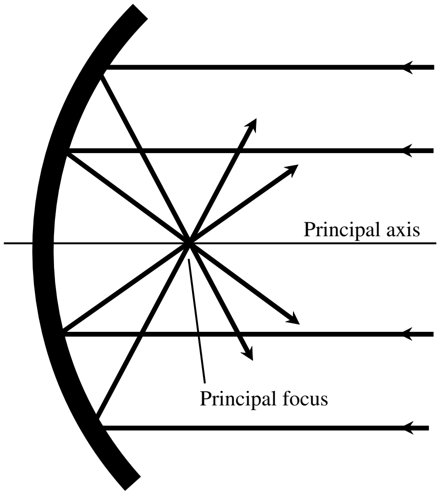

principal focus

principal mode See DOMINANT MODE.

principal ray The path described by an electron entering an electron lens parallel to the lens' axis, or by an electron leaving this lens parallel to the axis.

print **1.** The material transferred from a typewriter onto paper. **2.** The command, in a computer system, that causes data to be placed on paper or onto the output screen. **3.** The alphanumeric output of a computer or data terminal.

printed capacitor A two-plate capacitor formed on a printed circuit.

printed circuit A pattern of conductors (corresponding to the wiring of an electronic circuit) formed on a board of insulating material, such as a phenolic, by photo-etching, silk-screening of metallic paint, or by the use of pressure-sensitive preforms. The leads or pins of discrete components are soldered to the printed metal lines at the proper places in the circuit, or the components can be formed along with the conductors. Also see ETCHED CIRCUIT.

printed-circuit board A usually copper-clad plastic board used to make a printed circuit.

printed-circuit lamp A baseless lamp having flexible leads for easy soldering or welding to a printed circuit.

printed-circuit relay A usually small relay provided with pins or lugs for easy solder connection to a printed circuit.

printed-circuit switch A rotary switch whose contacts and contact leads are printed on a substrate.

printed-circuit template Also called *etching pattern.* A drawing for the purpose of making printed-circuit boards by photographic means.

printed coil A flat, spiral coil formed on a printed circuit.

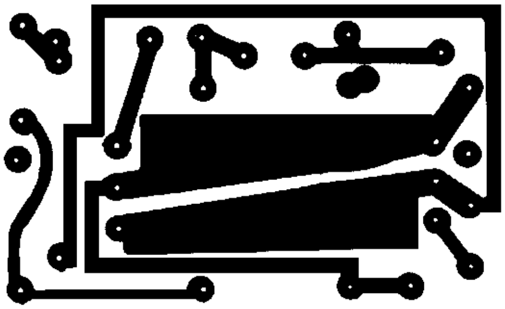

printed-circuit template

printed component A component formed on the substrate of a printed circuit. See, for example, PRINTED CAPACITOR, PRINTED COIL, and PRINTED RESISTOR.

printed display See DATA PRINTOUT, **2**.

printed element See PRINTED COMPONENT.

printed inductor See PRINTED COIL.

printed resistor A resistor printed or painted on a printed circuit.

printed wiring The printed or etched metal lines that serve as the conductors in a printed circuit.

printer In computer and calculator operations and in measurement procedures, a readout device that prints a permanent record of output data. There are several types, including the *dot-matrix printer*, the *daisy-wheel printer*, the *inkjet printer*, and the *laser printer.* Some printers, such as the daisy-wheel type, can render only text data; others, such as the laser type, can print high-resolution graphic images, sometimes in color.

printing calculator **1.** An electronic calculator that supplies a printed record of the results of a calculation. **2.** For a programmable calculator, the results, a record of program steps, and plots of curves.

printing digital voltmeter Abbreviation, PDVM. A digital voltmeter that delivers a printed record of a voltage reading, in addition to the usual digital readout of the voltage.

printing telegraph **1.** A telegraph that prints the received message on a tape or page. **2.** See TELETYPEWRITER.

printing wheel See PRINT WHEEL.

print format The form of data transmitted by a computer program to a printer (e.g., plain text, graphics, color graphics, etc.).

printout See DATA PRINTOUT, **1**, **2**.

print-through In prerecorded magnetic tape on a reel or cassette, the transfer of magnetism between layers of the rolled-up tape.

print wheel In a daisy-wheel printer, the rotatable wheel on whose rim the letters, numbers, and other symbols are inscribed in relief.

priority indicator In data transmission, a code that specifies the order of importance of a message in a group of messages to be sent.

priority processing In multiple programming operations, a system for ascertaining the order of processing for different programs.

privacy code **1.** A subaudible tone used in cordless telephone systems to reduce the chances of interference between phones operating on the same channel in close proximity. **2.** A subaudible tone used in radio transmissions, especially in conjunction with repeaters, to allow only those stations with the proper code to be received. **3.** A tone-burst sequence at the beginning of a transmission that actuates a receiver, allowing only those stations with the proper code to be received.

privacy equipment Devices, such as speech scramblers and digital encryption programs, that provide some measure of secrecy in communications.

privacy switch In a telephone amplifier, a switch (usually a pushbutton) for muting outgoing messages.

private automatic exchange Abbreviation, PAX. A dial telephone system for use within an organization and having no connection to the central office. Compare PRIVATE BRANCH EXCHANGE.

private branch exchange Abbreviation, PBX. A telephone system, complete with a private manually operated switchboard and individual telephone sets, installed and operated on private premises but having trunk-line connection to the central office. Compare PRIVATE AUTOMATIC EXCHANGE.

private line **1.** A communication circuit in which the use is limited, by electronic means, to certain subscribers. **2.** A subaudible-tone system used to restrict access to a communications system. The tone frequency is predetermined. For access to the system, a transmitted signal must contain the tone of the appropriate subaudible frequency, in addition to the voice or other information.

probability **1.** The branch of mathematics concerned with the likelihood of an event's occurrence. It has many applications in quality control and physics. **2.** The mathematical likelihood that an event will occur.

probable error Abbreviation, PE. The value of error above and below which all other error values are equally likely to occur.

probe **1.** A usually slender pencil-like implement with a pointed metal tip and a flexible, insulated lead. It is used to contact live points in a circuit under test (e.g., *voltmeter probe* and *oscilloscope probe*). **2.** A device used to sample a radio-frequency voltage or current at a desired point (e.g., WAVEGUIDE PROBE). **3.** A pickup device shaped like a probe for insertion into close quarters (e.g., PROBE THERMISTOR).

probe meter See PROBE-TYPE VOLTMETER.

probe thermistor A thermistor of slender construction for insertion into an area in which the temperature is to be monitored or controlled.

probe thermocouple A thermocouple in the form of a slender probe for insertion into close quarters for temperature sensing or temperature control.

probe tip See PROD.

probe-type voltmeter A voltmeter installed in a long probe or wand. Kilovoltmeters are sometimes constructed in this fashion, with a long multiplier resistor housed in the probe.

probing A process for locating, or determining the existence of, external artificial interference (e.g., power-line noise) in a radio communications circuit.

problem-oriented language Any high-level computer programming language that allows the user to write programs as statements in terms applicable to the field of interest (e.g., COBOL's statements in English for problems relating to business).

problem reduction In artificial intelligence, a process in which problems are made easier by breaking them down into smaller logical parts.

process control The control of a process, such as one of manufacturing, by means of computers.

processor **1.** A circuit or device used to modify a signal in response to certain requirements (e.g., *clipper* and *waveshaper*). **2.** See DATA-PROCESSING MACHINE. **3.** See CENTRAL PROCESSING UNIT. **4.** See MICROPROCESSOR.

prod The metal tip of a probe (see PROBE, **1**).

product **1.** The result of mixing or heterodyning of two or more signals. **2.** The result of modulating one signal with another. **3.** The result of combining or processing a signal or signals in a specified manner. **4.** The saleable end result of a manufacturing process. **5.** The result obtained when two or more quantities are multiplied by each other.

product detector A detector circuit whose output is the product of two signals applied simultaneously to the circuit. In a single-sideband receiver, for example, one of the signals is the incoming signal; the other, the signal from the local beat-frequency oscillator.

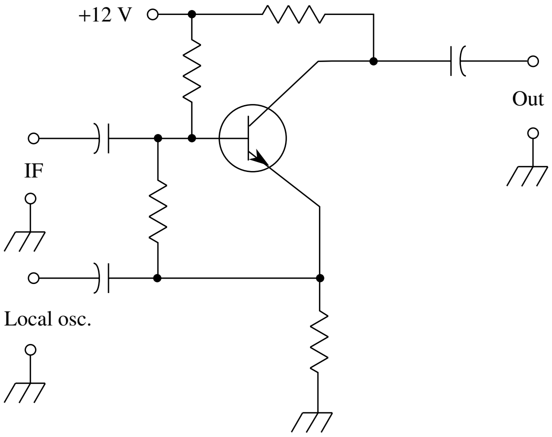

product detector

production lot A manufactured set of components, circuits, or systems, intended for sale. All of the units in the lot are identical. The finished product is suitable (presumably) for consumer use.

production unit One unit in a production lot; a finished unit, ready for use by a consumer.

product modulator A modulator whose output is equal or proportional to the product of carrier voltage and modulating voltage.

product of sine waves The result of multiplying one sine wave by another with attention being paid to the power factor. In the case of a resistive circuit, where the power factor is equal to 1, all voltage-current (*EI*) products are positive, and are equal to the true power. A product wave has negative half-cycles when the circuit contains reactance.

professional engineer A person licensed by a state board of examiners to work independently as an engineering. Also see PE and REGISTERED PROFESSIONAL ENGINEER.

program **1.** In computer operations, a detailed sequence of instructions representing an algorithm (the necessary steps in solving a problem) that can be implemented by a computer. **2.** The content of a radio or television broadcast during a specified period of time. **3.** In audio recording, the composite output from the mixer, used to make the master tape or disc.

program address counter See INSTRUCTION REGISTER.

program amplifier A broadcast preamplifier used at the studio or a remote location.

programatics The study of computer programming.

program circuit In wire telephony, a circuit capable of handling music and other audio data that covers a wide band of frequencies.

program compatibility The condition in which a program written for one computer can be used with another computer having a different architecture.

program controller In a central processor, a unit that controls the sequence and execution of program instructions.

program counter See CONTROL REGISTER.

program file A flexible reference system for software library maintenance.

program flowchart A representation of a computer program in the form of a flowchart. Each function and transition point is indicated by a box in the chart. A user can follow the flowchart and determine the outcome of the program for any given set of input parameters.

program library A collection of computer or programmable-calculator programs. Usually, it means the collection of programs used in a given computer system, often a software package supplied by the hardware vendor. It might also be a catalog of programs with instructions for their use.

programmable calculator A calculator that can be programmed to perform a chain of operations in a given order repetitively.

programmable read-only memory In a computer, a read-only memory (ROM) that can store a program.

program maintenance The ongoing correcting, updating, and modification of computer programs belonging to a system.

programmed dump A dump that occurs during a program run, according to a program instruction.

programmed halt During a computer program run, a temporary cessation resulting from an interrupt or halt instruction.

programmed instruction See MACRO INSTRUCTION.

programmed timer See CYCLE TIMER.

programmer A person who writes computer programs.

program modification **1.** In computer programming operations, a change in the effect of instructions and addresses during a program run by performing arithmetic and logical operations on them. **2.** Rewriting, or adding a patch to, a computer program. Also see PATCH, **3**.

program register See CONTROL REGISTER.

program segment A unit within a computer program that is stored with others in memory at the time of the program's execution, or sometimes, as overlays loaded individually when the entire program exceeds memory capacity.

program specification A description of the steps involved in the solution of a problem, from which a programmer devises a computer program.

program step An instruction in a computer program.

program tape In computer operations, a magnetic or paper tape that contains programs for a system or application.

program timer **1.** A programmed timer (see CYCLE TIMER). **2.** A timing unit that controls the duration of a program.

progressive scanning Non-interlaced television raster scanning, in which the lines are traced from top to bottom in succession. Conventional television broadcasting uses INTERLACED SCANNING.

progressive wave A wave disturbance that travels through a theoretically perfect homogeneous medium. This can be a compression (longitudinal) wave, a transverse wave, or an electromagnetic wave.

projected cutoff For an amplifier circuit, the operating point at which crossover distortion vanishes. The direct-current bias voltage (grid or gate) required for projected cutoff is somewhat lower than the value corresponding to conventional cutoff of plate or drain current.

project engineering A field of engineering dealing with the coordination of a complete project.

projection television Large-screen television for viewing by a relatively large group, usually accomplished via a projection tube and optical system.

projection tube A cathode-ray tube, especially a television picture tube, capable of producing a bright image that can be projected onto a large screen by means of a lens system.

projector **1.** A device that transmits a visible image onto a surface for reproduction. **2.** In general, any device that transmits a signal into space.

PROLOG Acronym for *programming in logic.* A high-level computer programming language, similar to LISP, used in artificial intelligence. The operator inputs facts and rules; the computer, in effect, derives theorems from the facts by following the logical rules.

PROM Abbreviation of PROGRAMMABLE READ-ONLY MEMORY.

promethium Symbol, Pm. A metallic element of the rare-earth group, produced artificially. Atomic number, 61. Atomic weight, approximately 145. Formerly called *illinium.*

promethium cell A radioactive battery cell using an isotope of promethium. Radioactive particles from this substance strike a phosphor, causing it to glow. Self-generating photocells then convert this light into electricity.

PROM programmer An electronic device that can store a computer program in a PROGRAMMABLE READ-ONLY MEMORY (PROM). It uses a built-in keyboard.

prompt In computer operations, a message received by an operator from an operating system or an individual program. For example, in disk operating system (DOS), it could be the statement "Bad command or file name."

prompting In computer or programmed-calculator operations, the entry of a special, required variable when the machine halts and awaits such entry.

prong See PIN.

prony brake An arrangement for measuring the mechanical power output of a rotating machine. It is a special form of friction brake consisting of a band passed around a pulley on the rotating shaft of the machine under test and held at each end by a spring balance.

propagation **1.** The extension of energy into and through space. Thus, radiant energy is *propagated* from and by its source. **2.** A phenomenon resulting from the extension of energy into and through space. Thus, radio waves can be spoken of as a *propagation.*

propagation constant For waves transmitted along a line, a number showing the effect the line has on the wave. This is a complex figure [i.e., one containing a real-number component (the attenuation constant) and an imaginary-number component (the phase constant)].

propagation delay **1.** Symbol, t_{pd}. In an integrated-circuit logic gate, the time taken for a logic signal to be propagated across the gate. **2.** In digital-circuit operation, the time required for a logic-level change to be transmitted through one or more elements.

propagation delay-power product See DELAY-POWER PRODUCT.

propagation factor The ratio E/E_o, where E is the complex electric-field strength at a point to which a wave has been propagated, and E_o is the complex electric-field strength at the point of origin. Also called *propagation ratio.*

propagation loss The path loss of an electromagnetic disturbance between the transmitting and receiving antennas.

propagation mode See WAVEGUIDE MODE.

propagation ratio See PROPAGATION FACTOR.

propagation time In digital-circuit operation, the time required for a binary bit to be transferred from one point to another in the system.

propagation velocity See VELOCITY OF PROPAGATION.

proportional action An action, such as amplification or conversion, that produces an output signal proportional to the input signal.

proportional amplifier An amplifier in which the instantaneous output amplitude is proportional to the instantaneous input amplitude.

proportional control A voltage-regulation system in which the feedback correction voltage is proportional to the output-voltage error.

proportional counter A Geiger tube having a pointed-wire (or ball-tipped-wire) anode. The voltage developed across the load resistor is proportional to the number of ions created by the radioactive particles entering the tube.

proprioceptor A set of transducers and associated circuitry that allows a computerized robot to constantly sense the positions of its end effectors, and use this data in carrying out programmed tasks.

prosodic features Variations in voice tone and emphasis that lend meaning and implication to spoken statements. It is important in advanced computer speech recognition and speech synthesis systems. Two sentences with identical wording can have greatly different meanings, depending on these factors (e.g., "*You* are!" versus "You *are?*").

prosthesis An electromechanical artificial human limb or body part. Examples: artificial legs, artificial hands, and artificial respirators. Some such devices are computer-controlled; others can be manipulated by nerve impulses.

protactinium Symbol, Pa. A relatively short-lived radioactive metallic element. Atomic number, 91. Atomic weight, 231.04 Formerly called *protoactinium.*

protected area A region to which access is restricted, and that is secured by an alarm system, surveillance cameras, or other intrusion-prevention systems.

protected location In computer storage, a location whose contents are protected from mutilation or erasure by making the location usable only by following a special procedure (e.g., using a password).

protection In a multiple processing computer system, preventing interference between data or programs.

protective bias In the final power amplifier of a radio transmitter, external direct-current bias applied to the base, gate, or grid. Prevents runaway in collector, drain, or plate current when the bias caused by the driving signal is lost.

protective capacitor A power-line bypass capacitor.

protective device **1.** A component that breaks a circuit in the event of excessive voltage or current from the power supply. **2.** A device that prevents excessive power from being delivered to a load by a driving circuit.

protective gap **1.** A spark gap connected in parallel with a component, or between a line and ground as protection against high-voltage transients and surges. **2.** A spark-gap-type lightning arrester.

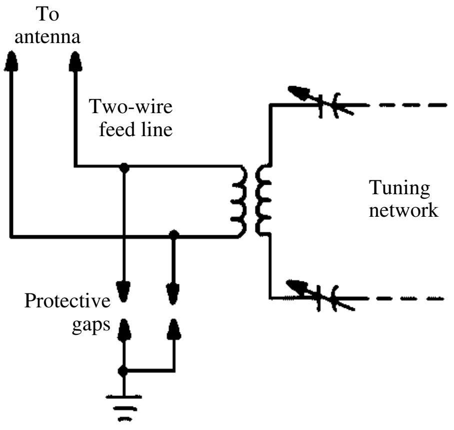

protective gap

protective resistor **1.** A bleeder resistor connected in parallel with a filter capacitor in a high-voltage direct-current power supply to discharge the capacitor automatically, thus preventing electric shock. **2.** A series resistor that limits the current going through a device.

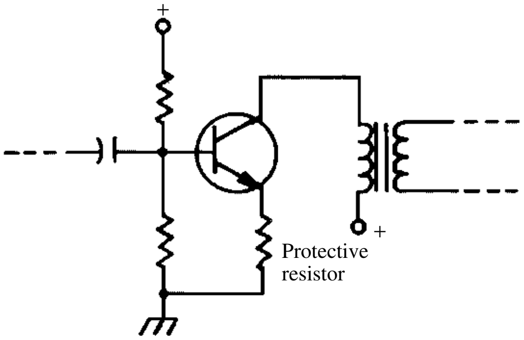

protective resistor, 2.

protector **1.** A fast-acting power-disconnect device, such as a circuit breaker or fuse, that acts to protect electronic equipment. **2.** A device or connection, such as a safety ground or ground-fault interrupter, that protects an operator from electric shock. **3.** See CONTACT PROTECTOR.

protium The light isotope of hydrogen, having an atomic mass of 1.

protoactinium See PROTACTINIUM.

protocol **1.** A set of parameters for a digital communications signal. **2.** The method by which a procedure is followed; a uniform set of governing regulations. It ensures proper operation of a system or network.

proton A positively charged particle in the nucleus of an atom. The mass of a proton is approximately 1840 times the mass of an electron.

proton rest mass See MASS OF PROTON AT REST.

proton-synchrotron A synchrotron that uses frequency modulation of the radio-frequency accelerating voltage. It can accelerate protons to energies of several billion electronvolts.

prototype The preliminary design or model of a device or system. It is often modified numerous times before the final design is attained. Compare PILOT MODEL.

proustite Crystalline silver arsenide trisulfide. Artificial crystals of this compound are used in tunable infrared-ray instruments.

proximity alarm A capacitance relay used to actuate an alerting-signal device when an area is intruded upon or a person is too close to a protected object. Also called INTRUSION ALARM.

proximity detector See PROXIMITY SENSOR.

proximity effect **1.** The influence of high-frequency current flowing in one conductor on the distribution of current flowing in an adjacent conductor. **2.** In an audio system, the result of placing a microphone too close to a person's mouth. Under these conditions, some spoken consonants (e.g., B, F, P, and T) produce clapping or booming sounds.

proximity fuse An electronic device situated in the nose of a missile. When the missile is near the target, the fuse transmits a signal that is reflected back from the target; this reflected signal detonates the missile.

proximity relay See CAPACITANCE RELAY.

proximity sensing The ability of a machine, especially a robot, to detect when an object is near. This is an aid in robot navigation because it prevents collisions. Some devices can measure the distance from a robot, or from a robotic end effector, to a nearby object.

proximity sensor A device that indicates the presence of a nearby body. Such a device uses some form of circuit, such as that of a CAPACITANCE RELAY, that changes its operating characteristics when an object enters its field.

proximity switch See CAPACITANCE RELAY.

PRR Abbreviation of PULSE REPETITION RATE.

PRV Abbreviation of PEAK REVERSE VOLTAGE.

PS Abbreviation of POWER SUPPLY.

ps Abbreviation of PICOSECOND. (Also, psec.)

PSD Abbreviation of PHASE-SENSITIVE DETECTOR.

psec Abbreviation of PICOSECOND. (Also, ps.)

pseudocode In a computer system, an instruction or code symbol that affects the operation of the programming in an indirect manner.

pseudo-instruction In computer programming operations, data representing an instruction and requiring translation by a compiler or assembler.

pseudo-offlining During input/output operations in a computer system, maximizing hardware by disconnecting slow devices from the process in question.

pseudo-operation In computer operations, an operation that, rather than being performed by hardware, is carried out by special software or by macroinstruction.

pseudo-random numbers Numbers that, although produced by a computer operating on an algorithm for their generation, are useful for an application requiring random numbers.

pseudo-stereophonic effect A somewhat heightened binaural effect obtained when two loudspeakers are situated, relative to the listener, so that a transit-time difference of 1 to 30 milliseconds results.

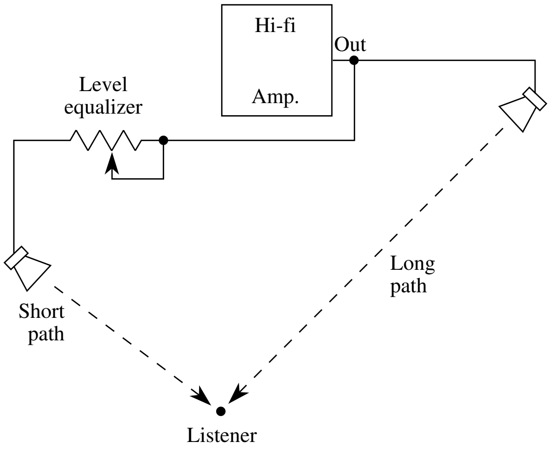

pseudo-stereophonic effect

psf Abbreviation of *pounds per square foot.* (Also, lb per sq ft, lb/ft^2.)

psi **1.** Abbreviation of *pounds per square inch.* (Also, lb per sq in, lb/in^2.)

psia Abbreviation of POUNDS PER SQUARE INCH ABSOLUTE.

psig Abbreviation of POUNDS PER SQUARE INCH GAUGE.

psi particle A massive elementary particle that represents a resonance in an electron-positron interaction.

PSK Abbreviation of PHASE-SHIFT KEYING.

PSM Abbreviation of *pulse-spacing modulation,* more commonly called PULSE-INTERVAL MODULATION.

psophometer A device used to measure noise in a wire communications system. It provides quantitative readings based on typical human observations.

psvm Abbreviation of *phase-sensitive voltmeter.*

PSWR Abbreviation of *power standing-wave ratio.*

psychoacoustics A field of acoustics, overlapping with psychology, concerned with the effects of sounds on human beings.

PT Abbreviation of *Pacific Time.*

Pt Symbol for PLATINUM.

PTC Abbreviation of POSITIVE TEMPERATURE COEFFICIENT.

PtIr Symbol for PLATINIRIDIUM.

PTM Abbreviation of PULSE-TIME MODULATION.

PTO Abbreviation of PERMEABILITY-TUNED OSCILLATOR.

PTT Abbreviation for *press-to-talk.* See PRESS-TO-TALK MICROPHONE; PRESS-TO-TALK SWITCH.

PTV Abbreviation of *public television.*

p-type conduction In a semiconductor, current flow consisting of the movement of holes. Compare N-TYPE CONDUCTION.

p-type material Semiconductor material that has been doped with an acceptor-type impurity and, consequently, conducts current mainly via hole migration. Germanium, for example, when doped with indium, becomes p-type. Compare N-TYPE MATERIAL.

p-type semiconductor An acceptor-type semiconductor (i.e., one containing an excess of holes in its crystal lattice).

PU Abbreviation of PICKUP.

Pu Symbol for PLUTONIUM.

public-address amplifier A high-gain, high-power audio amplifier designed especially for the reproduction of speech and music at large gatherings.

public-address system A system of sound reproduction especially designed for use at large gatherings indoors or outdoors. The system includes microphones, a public-address amplifier, loudspeakers, and sometimes recorders and playback devices. Also called *PA system.*

puck drive In a tape recorder, a speed-reduction system for driving the flywheel from the shaft of the (high-speed) motor. In some machines, a rubber tire mounted on the flywheel is driven, through friction, by the motor shaft. In others, an intermediate rubber-tired wheel is placed between the motor shaft and the rim of the flywheel.

puffer A meter or bridge for measuring small values of capacitance. The name comes from the spoken sound of *pF,* the abbreviation of PICOFARAD.

pulldown Descriptive of a circuit, device, or individual component used to lower the value (e.g., impedance) of a circuit to which it is connected.

pull-in current The current required to close a relay.

pulling **1.** The abnormal tendency of one circuit to cause another to slip into tune with it. This often results from coupling (intended or accidental) that is too tight. Thus, when two oscillators feed a common circuit, such as a mixer, one might pull the other into tune with itself. **2.** Lowering of a crystal frequency by an external reactance.

pull-in voltage The voltage required to close a relay.

pull switch A mechanical switch actuated by a pulling action.

pullup Descriptive of a circuit or component used to raise the value (e.g., impedance) of a circuit to which it is connected.

pulsar An extremely dense, rapidly rotating collapsed star that produces radio signals at regular intervals. The pulse frequency varies from less than one hertz to several tens or hundreds of hertz.

pulsating direct current A direct current that periodically rises and falls between zero and a maximum value (or between two positive or negative values) without changing polarity. Thus, it is possible to have either a pulsating positive current or a pulsating negative current. Also see DIRECT CURRENT.

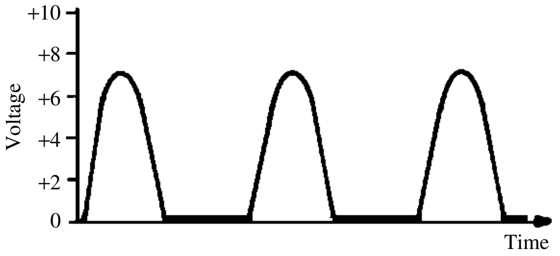

pulsating direct current

pulsating wave See PULSATING DIRECT CURRENT.

pulse A transient signal that is usually of short duration, constant amplitude, and one polarity. A typical example is a narrow positive or negative spike.

pulse amplifier An amplifier having wide frequency response and low distortion, used for amplifying steep-sided pulses of short duration.

pulse-amplitude modulation Abbreviation, PAM. A method of conveying information in wireless communications. A train of pulses is transmitted. The strength of each individual pulse varies according to the modulating waveform. Normally, the pulse amplitude increases as the instantaneous modulating-signal level increases (positive modulation). However, this can be reversed so that higher audio levels cause the pulse amplitude to go down (negative modulation).

pulse bandwidth For an amplitude pulse, the minimum bandwidth occupied. The faster the rise and/or decay times of a pulse, the greater the bandwidth. The greater the pulse frequency, the greater the bandwidth.

pulse code A code in which groups of pulses represent digits.

pulse-code modulation Abbreviation, PCM. A method of conveying information in wireless communications. A train of pulses is transmitted. The intelligence-carrying signal is sampled periodically and the amplitude is converted into binary code. The code might allow for eight levels (000 to 111), 16 levels (0000 to 1111), 32 levels (00000 to 11111), or 64 levels (000000 to 111111).

pulse-code-modulation binary code A pulse code used in communications not in the form of line transmission. Individual values are denoted by binary numbers.

pulse-code modulation multiplex equipment A multiplexer/demultiplexer for signal conversion between a single signal and multiple-channel signals. It uses both pulse-code modulation and time-division multiplexing.

pulse-count divider A circuit or device that receives an input of a certain number of pulses (or pulses per second) and delivers an output that is a function of that quantity. See, for example, DIVIDE-BY-SEVEN CIRCUIT and DIVIDE-BY-TWO CIRCUIT.

pulse counter A circuit or device that indicates the number of pulses presented to it in a given time interval.

pulse counting Counting pulses in a sequence. At low speed (pulse repetition rate), this can be done with an electromechanical dial-type counter. At high speed, a fully electronic circuit is required.

pulse delay circuit A monostable multivibrator adapted to deliver its single output pulse a predetermined time after the input pulse has been applied.

pulse dialing A form of telephone dialing in which each digit is formed by a series of pulses, usually at 10 to 20 Hz. The pulses are the equivalent of disconnecting the line for a few milliseconds. Each digit has the corresponding number of pulses, except digit 0, which is formed by 10 pulses.

pulse droop Distortion observable as a downward-sloping top on the oscilloscope trace of a pulse. It can be quantified in volts, millivolts, microvolts, amperes, milliamperes, or microamperes.

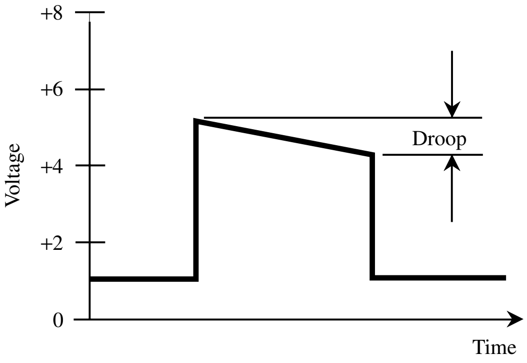

pulse droop

pulsed laser A laser in which flashes (pulses) of high-intensity light excite the lasing medium.

pulse duration The time period during which a pulse exists (i.e., its width on an oscilloscope display).

pulse-duration modulation Abbreviation, PDM. See PULSE WIDTH MODULATION.

pulse equalizer A MONOSTABLE MULTIVIBRATOR adapted to deliver pulses of equal amplitude, shape, and width—even when it receives trigger pulses of different kinds.

pulse fall time The time required for the trailing edge of a pulse to fall from 90 to 10 percent of its peak amplitude. Compare PULSE RISE TIME.

pulse-forming line A circuit used in radar for producing high-intensity pulses. Inductances and capacitances are combined in a long string, and the effect is to generate high-amplitude radio-frequency pulses.

pulse-frequency modulation Abbreviation, PFM. See PULSE-INTERVAL MODULATION.

pulse generator A signal generator that produces pulses. A general-purpose generator of this sort will produce pulses of adjustable amplitude, duration, shape, and repetition rate.

pulse-height discriminator A circuit or device that passes only pulses whose amplitudes exceed a predetermined level.

pulse interval The interval between successive pulses.

pulse-interval modulation Abbreviation, PIM. Also called *pulse-frequency modulation* (PFM) or *pulse-numbers modulation* (PNM). A method of conveying information in wireless communications. A train of pulses is transmitted. Every pulse has the same amplitude and the same duration, but their rate fluctuates with the modulating waveform. When there is no modulation, the pulses are evenly spaced with respect to time. An increase in the instantaneous data amplitude might cause pulses to be sent more often (positive modulation) or less often (negative modulation).

pulse inverter A single-stage, wideband, low-distortion, common-emitter, or common-source amplifier. The output-pulse waveforms are therefore inverted, with respect to the input-pulse waveforms.

pulse jitter In a pulse train, a disturbance characterized by random changes in the spacing between pulses.

pulse-length modulation See PULSE-DURATION MODULATION.

pulse load The load impedance for a pulse generator.

pulse mode See PULSE MODULATION.

pulse modulation See PULSE-AMPLITUDE MODULATION, PULSE-CODE MODULATION, PULSE-INTERVAL MODULATION, PULSE-POSITION MODULATION, PULSE-WIDTH MODULATION.

pulse modulator **1.** A modulator that delivers power or voltage pulses for modulating a carrier. **2.** A device that modulates pulses (see PULSE MODULATION, **2**).

pulse narrower A circuit or device that reduces the duration (width) of a pulse.

pulse-numbers modulation Abbreviation, PNM. See PULSE-INTERVAL MODULATION.

pulse operation Intermittent operation of a circuit, in the form of discrete pulses.

pulse oscillator Any oscillator with an output that consists of a series of pulses.

pulse-position modulation Abbreviation, PPM. A method of conveying information in wireless communications. A train of pulses is transmitted. The timing of each individual pulse varies according to the modulating waveform. The pulses occur earlier or later than the nominal (zero-modulation) time, depending on the instantaneous amplitude of the modulating signal.

pulse rate See PULSE REPETITION RATE.

pulse ratio The ratio of pulse height (amplitude) to pulse width (duration).

pulse regeneration Restoration of the original waveform and frequency to a pulse. It eliminates distortion caused by circuits or propagation conditions.

pulse repetition frequency Abbreviation, PRF. See PULSE-REPETITION RATE.

pulse-repetition rate Abbreviation, PRR. The number of pulses per unit time; usually pulses per second (pps).

pulse rise time The time required for the leading edge of a pulse to rise from 10 to 90 percent of its maximum amplitude. Compare PULSE FALL TIME.

pulse scaler A circuit actuated by the reception of a definite, predetermined number of input pulses.

pulse-shaping circuit **1.** A circuit for producing a pulse from a wave of some other shape (e.g., sine wave). **2.** A circuit for tailoring a pulse to desired shape, amplitude, and duration.

pulse spacing The interval between successive pulses.

pulse-spacing modulation Abbreviation, PSM. See PULSE-INTERVAL MODULATION.

pulse-steering diode In a flip-flop circuit, a diode through which the trigger pulse must pass to switch the circuit. Because of the unidirectional conductivity of a diode, pulses of only one polarity are passed.

pulse stretcher **1.** A shaping circuit that widens a pulse (i.e., increases its duration). **2.** A circuit, such as a special monostable multivibrator, that generates a pulse that is wider than the trigger pulse.

pulse stuffing See JUSTIFICATION, **2**.

pulse tilt The sloping of the normally flat top of a pulse either up or down. Also see PULSE DROOP.

pulse time See PULSE DURATION.

pulse-time modulation See PULSE-POSITION MODULATION.

pulse train A series of successive pulses of usually one kind.

pulse transformer A transformer designed to accommodate the fast rise and fall times of pulses and similar nonsinusoidal waveforms. Such transformers often use special core materials and are made using special winding techniques.

pulse transmitter **1.** A device that transmits a series of pulses. **2.** A pulse-modulated transmitter. **3.** See PULSE MODULATOR.

pulse waveform The general shape of a pulse as it appears on an oscilloscope display. The various forms range from positive or negative half-sinusoids, through rectangles, to thin-line spikes.

pulse width The horizontal dimension of a pulse (i.e., its duration).

pulse-width modulation Abbreviation, PWM. Also called *pulse-duration modulation* (PDM). A method of conveying information in wireless communications. A train of pulses is transmitted. The width (duration) of each individual pulse varies according to the modulating waveform. Normally, the pulse width increases as the instantaneous modulating-signal level increases (positive modulation). However, this can be reversed so that higher audio levels cause the pulse width to decrease (negative modulation).

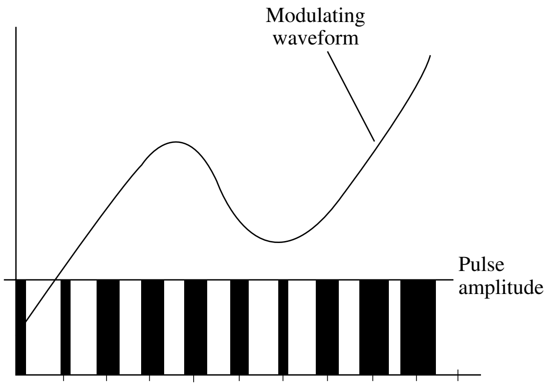

pulse-width modulation

pump **1.** In a parametric amplifier, the oscillator that supplies the signal that periodically varies the reactance of the varactor. **2.** The pumping signal in **1**. **3.** To perform the operation (pumping) described in **1**. **4.** To increase the energy level of an atom or molecule (by exposing it to electromagnetic radiation) to such an extent that oscillation or amplification occurs. A ruby laser, for example, produces its intense, coherent beam as a result of pumping. **5.** The radiation used to pump an atom or molecule. **6.** The device producing the radiation required to pump an atom or molecule.

pump frequency The frequency of a PUMP voltage.

pumping A method of laser actuation. A series of pulses, at the resonant frequency of the lasing material, is injected to cause laser output.

pump oscillator An oscillator for producing a pump voltage.

pump voltage The voltage of a pumping signal. Also see PARAMETRIC AMPLIFIER and PUMP, **1**, **2**.

punch **1.** A tool for cutting holes in metal chassis, panels, and boxes for electronic equipment. **2.** High signal strength.

punch-in editing In audio recording, a feature that allows convenient insertion of new material on a tape. The tape recorder can be switched instantly from Play to Record, and back again, whenever the operator wants to add material.

punchthrough In a bipolar transistor, the potentially damaging condition resulting when the reverse bias of the collector is increased to a voltage high enough to spread the depletion layer entirely through the base. This tends to effectively connect the emitter to the collector.

punchthrough region The conduction region associated with higher-than-punchthrough voltage, in which bipolar-transistor current is excessive. Also see PUNCHTHROUGH.

punchthrough voltage The voltage that causes PUNCHTHROUGH in a given bipolar transistor.

puncture voltage See BREAKDOWN VOLTAGE, **1**.

Pupin coil One of several loading coils that can be inserted at intervals in series with a telephone line to cancel line-capacitance effects and, thus, improve the clarity of speech.

pure tone An audio-frequency (AF) tone having essentially no harmonic content; a sine-wave AF tone.

pure wave A wave containing no distortion products.

purging The removal of an undesired gas or other substance from a system by introducing a material to displace it.

purifier A power-line operated alternating-current electromagnet that can be manually rotated in front of a color-television picture tube to demagnetize the tube. Also called a DEGAUSSER.

purity **1.** In color television, complete saturation of a hue. **2.** In a waveform, complete freedom from distortion. **3.** The extent to which spurious signals are attenuated in the output of a radio or television transmitter. Also called *spectral purity.*

purity adjustment In a color-television picture tube, adjustment of each purity control for pure color.

purity coil A variable-current coil around the neck of a color-television picture tube that is used to adjust color purity.

purity control For a purity coil, the variable resistor that controls the current for color correction.

purity magnet A ring-magnet collar around the neck of a color-television picture tube to adjust, by rotation, color purity.

purple plague Corrosion that occurs when aluminum and gold are placed in contact.

pushbutton tuner A radio or television tuner utilizing pushbutton tuning.

pushbutton tuning The tuning of a circuit to various frequencies in single steps by means of pushbutton switches.

pushdown list In data processing, a method of amending a list. A new item entered at the top moves each existing item one position down.

pushdown stack Also called *first-in/last-out.* A digital read-write memory in which data bits emerge in reverse sequence from the order they go in. If data bit *x* enters the pushdown stack before data bit *y*, then *x* will come out after *y*. Compare FIRST-IN/FIRST-OUT.

push-in terminal A circuit contact or tie point, usually of thin, springy material, that can be inserted into a hole in a perforated board.

push-pull Pertaining to a circuit in which two active devices are used, with the inputs and outputs both placed in phase opposition. In the output circuit, even harmonics are canceled, and odd harmonics are reinforced.

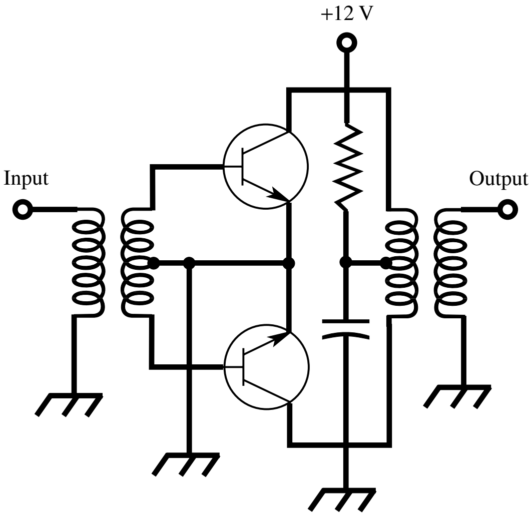

push-pull

push-pull amplifier An amplifier stage in which, for increased power output, two active devices are operated 180 degrees out of phase with each other in opposite halves of a symmetrical circuit. Also see PUSH-PULL CIRCUIT.

push-pull circuit A symmetrical circuit in which two active devices operate on separate halves of the input-signal cycle and deliver a combined output signal.

push-pull deflection In an oscilloscope, the application of deflection voltage to a pair of deflecting plates 180 degrees out of phase with each other. For this purpose, the output amplifier in the horizontal or vertical deflection channel is a push-pull stage.

push-pull doubler See PUSH-PULL MULTIPLIER.

push-pull microphone A set of two microphones, in which the audio-frequency outputs are in phase opposition.

push-pull multiplier A push-pull amplifier with its output circuit tuned to an odd-numbered multiple of the input frequency. This circuit is unsuitable for even-harmonic operation, but has some merit as an odd-harmonic multiplier (e.g., a *tripler* or *quintupler*). Also see PUSH-PUSH MULTIPLIER.

push-pull oscillator An oscillator stage in which, for increased power output, two active devices are operated 180 degrees out of phase with each other in opposite halves of a symmetrical circuit. Also see PUSH-PULL CIRCUIT.

push-pull/parallel amplifier An amplifier stage in which tubes or transistors are connected in push-pull/parallel for increased power output. Also see PARALLEL-COMPONENT AMPLIFIER, PUSH-PULL AMPLIFIER, and PUSH-PULL/PARALLEL CIRCUIT.

push-pull/parallel circuit A push-pull circuit in which two or more active devices are connected in parallel on each side of the circuit. This arrangement gives increased power output over that of the conventional push-pull circuit. See, for example, PUSH-PULL/PARALLEL AMPLIFIER and PUSH-PULL/PARALLEL OSCILLATOR.

push-pull/parallel oscillator An oscillator stage in which active devices are connected in push-pull/parallel for increased power output. Also see PARALLEL-COMPONENT OSCILLATOR, PUSH-PULL OSCILLATOR, and PUSH-PULL/PARALLEL CIRCUIT.

push-pull recording A type of film sound track consisting of two side-by-side images 180 degrees out of phase with each other.

push-pull transformer A transformer having a center-tapped winding for operation in a push-pull circuit.

push-push Pertaining to a circuit in which two active devices are used, with the inputs connected in phase opposition, and the outputs connected in parallel. The result is reinforcement of the even harmonics, and cancellation of the fundamental frequency and all odd harmonics.

push-push circuit See PUSH-PUSH MULTIPLIER.

push-push multiplier An amplifier circuit containing two active devices with their inputs connected in phase opposition and their outputs connected in parallel. This circuit is unsuitable for fundamental-frequency and odd-harmonic operation,

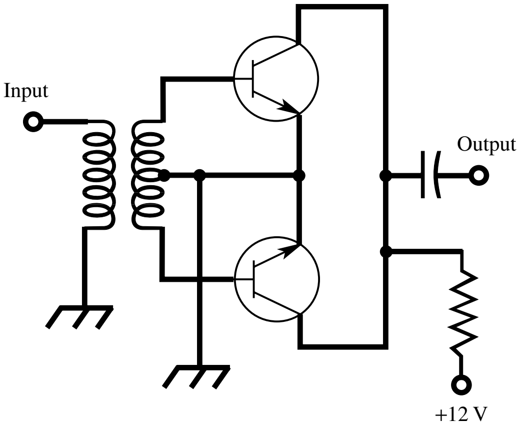

push-push

but has some merit as an even-harmonic multiplier (e.g., a *doubler* or *quadrupler*). Also see PUSH-PULL MULTIPLIER.

push-to-talk switch See PRESS-TO-TALK SWITCH.

pushup list In data processing, a method of amending a list, whereby new items are added at the end of the list; all other items retain their original positions. Compare PUSHDOWN LIST.

pV Abbreviation of PICOVOLT.

PVC Abbreviation of POLYVINYL CHLORIDE.

pW Abbreviation of PICOWATT.

PWM **1.** Abbreviation of PULSE-WIDTH MODULATION. **2.** Abbreviation of PLATED-WIRE MEMORY.

pwr Abbreviation of POWER.

Pyralin See CELLULOSE NITRATE.

pyramidal horn antenna A rectangular horn antenna that is flared in two dimensions. The horn width and height both increase linearly with increasing distance (in the direction of maximum radiation/response) from the feed point.

pyramidal wave See BACK-TO-BACK SAWTOOTH.

Pyrex A heat-resistant glass having numerous applications in electronics and chemistry.

pyrheliometer An instrument used to measure infrared radiation.

pyroelectricity In certain crystals, electricity generated by temperature change, and in particular, by the application of heat.

pyroelectric lamp See NERNST LAMP.

pyroelectric material A crystalline material that generates an output voltage when it is heated.

pyrolysis The process whereby heat changes a substance into one of several different substances by rearranging its atoms.

pyromagnetic effect In a material or circuit, the combined effect of heat and magnetism.

pyrometer An instrument, other than a thermometer, used for the measurement of temperature. See, for example, OPTICAL PYROMETER.

Pythagorean scale A sound scale defining a specific type of relationship among audio tones. If x and y are related by the Pythagorean scale and are adjacent in frequency, then a specific frequency (f) exists, so $x = f^2$ and $y = f^3$.

Pythagorean theorem A theorem of plane geometry. For a right triangle, with sides of lengths a, b, and c, where c is the side opposite the right angle, it is always true that $a^2 + b^2 = c^2$.

p-zone See P LAYER.

PZM Abbreviation of PRESSURE-ZONE MICROPHONE.

PZT Abbreviation of LEAD ZIRCONATE TITANATE.

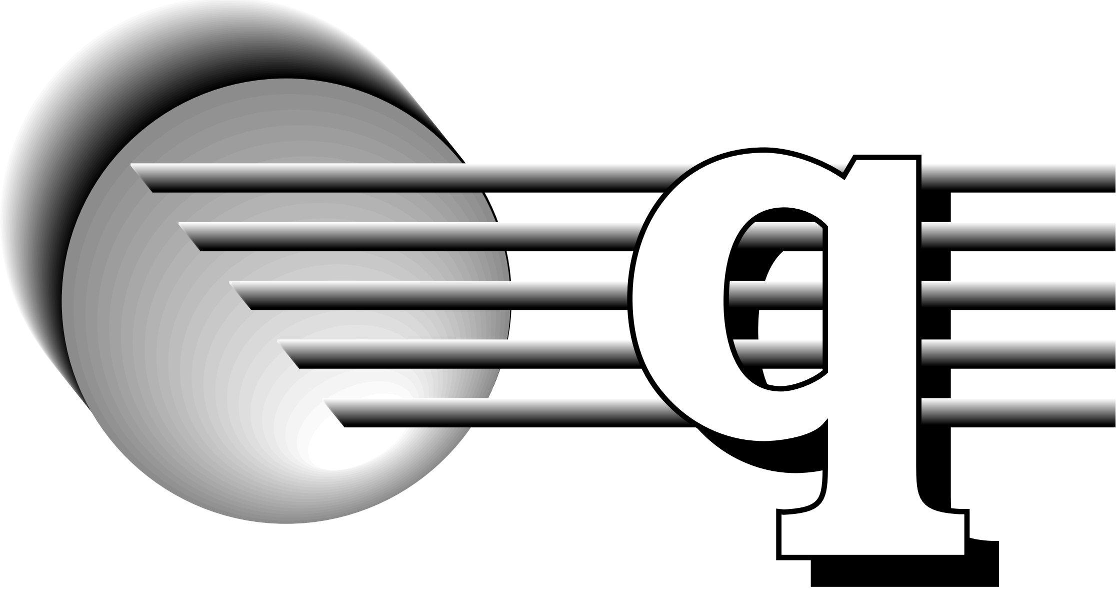

Q **1.** The figure of merit of a capacitor, inductor, or inductance-capacitance (LC) circuit, equal to the reactance divided by the resistance. **2.** Symbol for electrical charge. **3.** Occasional symbol for SELECTIVITY. **4.** See Q BAND. **5.** See *Q* OUTPUT.

q **1.** Symbol for quantity of electricity (in coulombs). **2.** Symbol for the charge carried by an electron (the charge carried by a hole is represented by –q). **3.** Symbol for the value of a quantum.

QA Abbreviation of QUALITY ASSURANCE.

***Q* adjustment** The separate null adjustment for the *Q* value of a component being tested in an impedance bridge having separate resistive and reactive balances.

***Q*-antenna** An antenna in which the transmission line (feeder) is matched in impedance to the center of the radiator by means of a *Q*-matching section.

QAVC Abbreviation of QUIET AUTOMATIC VOLUME CONTROL.

Q band The radio-frequency band 36 to 46 GHz. It is subdivided as follows: Q_a, 36 to 38 GHz; Q_b, 38 to 40 GHz; Q_c, 40 to 42 GHz, Q_d, 42 to 44 GHz; and Q_e, 44 to 46 GHz.

***Q* bar** One of the parallel metal tubes in a *Q*-matching section. Also see *Q* ANTENNA.

***Q* booster** See *Q* MULTIPLIER.

***Q* bridge** An alternating-current bridge used principally to determine the *Q* of capacitors and inductors. Bridges are usually used for audio-frequency *Q* determinations; resonant-type *Q* meters are generally used for measurement of radio-frequency *Q*.

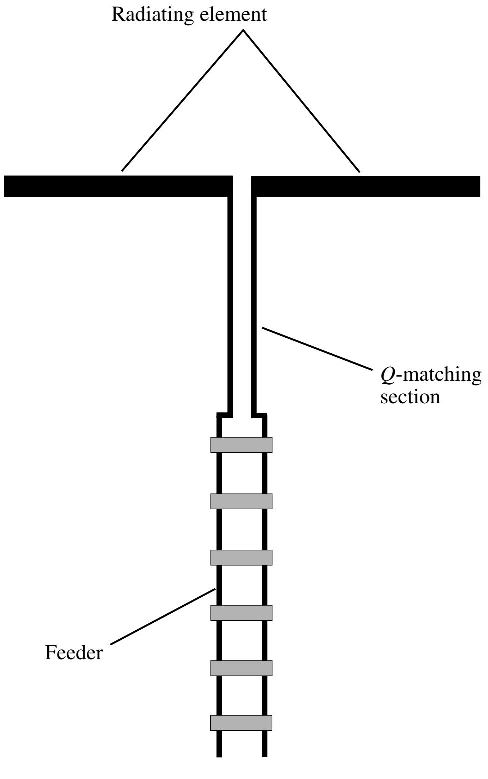

***Q*-antenna**

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

QC Abbreviation of QUALITY CONTROL.

QCE Abbreviation of QUALITY-CONTROL ENGINEERING or QUALITY-CONTROL ENGINEER.

Q channel In American (NTSC) color television, the 508-kHz-wide green-magenta color information transmission band.

QCT Abbreviation of QUALITY-CONTROL TECHNICIAN.

QCW In the local oscillator and associated circuitry of a color television receiver, a 3.85-MHz CW signal of Q PHASE.

QCW signal In a color television receiver, the component of the chrominance signal that is 90 degrees out of phase with the in-phase component.

Q demodulator In a color television receiver, the demodulator that combines the chrominance signal and the color-burst oscillator signal to recover the Q signal (see Q SIGNAL, **2**).

QED **1.** Abbreviation of QUANTUM ELECTRODYNAMICS. **2.** Abbreviation of *quod erat demonstrandum*, Latin for "which was to be demonstrated." Also, Q.E.D. Often written at the conclusions of valid logical proofs and derivations.

***Q* factor** See *Q*.

QFM Abbreviation of QUADRATURE MODULATION.

QM Abbreviation of QUADRATURE MODULATION.

***Q*-matching section** A linear radio-frequency impedance-matching transformer consisting of two straight, rigid, parallel metal conductors that are used to match a feeder to an antenna. The section is ¼ wavelength long at the operating frequency. The diameters and center-to-center spacing of the conductors are such that the characteristic impedance of the matching section is equal to the geometric mean of the feeder characteristic impedance and the radiation resistance of the radiator. Also see *Q* ANTENNA and QUARTER-WAVELENGTH MATCHING STUB.

***Q* meter** A usually direct-reading instrument for determining the *Q* of a capacitor, inductor, or inductance-capacitance (LC) circuit. Most *Q* meters are operated at radio frequencies, but audio-frequency instruments are available.

***Q* modulation** Amplitude modulation obtained by varying the effective *Q* of a radio-frequency tank circuit in step with a modulating component. See ABSORPTION MODULATION.

QMQB Abbreviation of *quick make/quick break*.

***Q* multiplier** A positive-feedback (regenerative) amplifier that increases the effective *Q* of a tuned circuit, and thereby sharpens its response, when its input is connected across the tuned circuit.

Q output The reference output of a flip-flop.

Q phase A color-television carrier signal that is 147 degrees out of phase with the color subcarrier.

Q point The point or points at which a load line intersects a device characteristic (such as the collector curve of a transistor or plate curve of a tube) and that identifies the quiescent operating point.

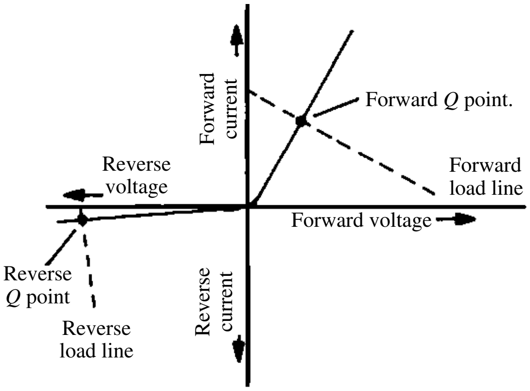

***Q* points on diode characteristic**

QRD Abbreviation of QUADRATIC RESIDUE DIFFUSOR.

***Q*-section transformer** See *Q*-MATCHING SECTION.

Q signal In color television, the quadrature component of the chrominance signal, equal to +0.48(R-Y) and to +0.41(B-Y), where B is the blue camera signal, R is the red camera signal, and Y is the luminance signal.

Q signals A set of three-letter groups, each beginning with the letter Q, used for simplified telegraph and radiotelegraph communication, and sometimes rapid voice communication (in radiotelephony). Each signal represents a commonly used phrase or message.

QSL card A card verifying communication with, or the reception of signals from, the station sending the card. Such verification is common in the amateur radio service and with some shortwave broadcast and CB stations.

QSO Amateur radiotelegraph abbreviation for TWO-WAY COMMUNICATION.

Q spoiler A device or circuit that produces Q SPOILING in a laser.

Q spoiling The technique of inhibiting laser action during an interval when an ion population excess is pumped up. When the laser is subsequently triggered by Q switching, a more powerful pulse of light results than would be otherwise obtained.

Q switching A laser-switching action obtainable with Kerr cells or rotating reflecting prisms, which consists of holding the Q of the laser cavity to a low value during an ion-population buildup, then abruptly switching the Q to a higher value.

***Q* transformer** See *Q*-MATCHING SECTION.

qty Abbreviation of QUANTITY.

quad **1.** A combination of four components, such as diodes, transistors, etc., in a single housing. The components are usually carefully matched. **2.** In a cable, a combination of four separately insulated conductors (sometimes, two twisted pairs) twisted together. **3.** Abbreviation of QUADRANT. **4.** See QUAD ANTENNA. **5.** See QUADROPHONIC.

quad antenna A directional antenna similar to the Yagi, except that full-wavelength loops are used instead of straight elements. A two-element array can consist of a driven element and a reflector, or it can have a driven element and a director. A three-element system has one driven element, one director, and one reflector. The director has a perimeter of about 0.97 electrical wavelength, the driven element measures exactly 1 electrical wavelength around, and the reflector has a perimeter of about 1.03 electrical wavelength. Additional director elements can be added to the basic three-element quad design to form arrays having any desired numbers of elements. The gain and directivity increase as the number of elements increases. Compare YAGI ANTENNA.

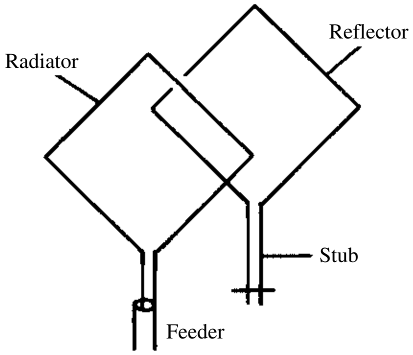

quad antenna

quadded cable See QUAD, **2**.

quadding Redundancy obtained by connecting components in series-parallel for enhanced reliability and/or increased power-handling capacity.

quad latch A set of four interconnected flip-flops that is used for digital data storage.

quadrant **1.** A specific 90-degree arc of a circle. **2.** One of the four parts formed on a plane surface by rectangular coordinates and designated I, II, III, and IV in a counterclockwise direction, starting with the upper-right quadrant. **3.** An altitude-measuring instrument.

quadrantal deviation The part of magnetic-compass deviation caused by the induction of transient magnetism into the horizontal soft iron of a vessel by the horizontal component of terrestrial magnetism.

quadrantal error See QUADRANTAL DEVIATION.

quadrant electrometer An electrometer whose principal parts are quadrants (a pillbox-shaped brass chamber split into four parts) and a needle (a flat, bowtie-shaped aluminum vane) suspended by a platinized quartz fiber between the quadrants.

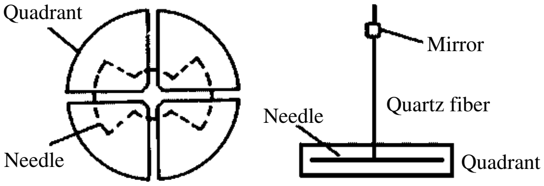

quadrant electrometer

quadraphonic sound Sound recording and reproduction involving four channels.

quadraphony Music recording or playback in which four distinct information channels are used. Also called *four-channel stereo.*

quadratic equation A second-degree equation [i.e., one in which the highest exponent is 2 (the square of an unknown) (e.g., $ax2 + bx + c = 0$)].

quadratic residue diffusor Abbreviation, QRD. In acoustics, a sound-reflection grating that scatters (diffuses) sound waves almost uniformly in all directions. The depths of the indentations in the grating are determined according to a QUADRATIC EQUATION.

quadrature The state of (cyclic events or points) being 90 degrees out of phase.

quadrature amplifier An amplifier circuit that introduces a 90-degree phase shift. Such amplifiers are used in control devices, test instruments, transmitters, and color television receivers.

quadrature axes The vertical axes in the complex-number plane (i.e., the $+j$ and $-j$ axes).

quadrature carrier See Q PHASE.

quadrature component **1.** The reactive component of an alternating current or voltage. **2.** A vector perpendicular to a reference vector. **3.** The imaginary-number component in a complex-number expression.

quadrature current Reactive current in an alternating-current circuit.

quadrature modulation In-phase modulation of two carrier components having a 90-degree phase difference.

quadrature number See IMAGINARY NUMBER.

quadrature-phase subcarrier signal See QCW SIGNAL.

quadrature portion In color television, the portion of the chrominance signal having the same (or opposite) phase as that of the Q-signal-modulated subcarrier, and that is 90 degrees out of phase with the in-phase portion.

quadrature sensitivity The sensitivity of a transducer to motions in a direction that is perpendicular to the normal axis of response.

quadrature voltage A voltage 90 degrees out of phase with another (reference) voltage.

quadrilateral **1.** Pertaining to an object having four sides. **2.** A four-sided plane polygon.

quadrillion The number 1,000,000,000,000,000 (10^{15}).

quadripartite Having four parts.

quadripole network A four-terminal network, usually with input- and output-terminal pairs.

quadrivalent Having a valence of 4. Tin, for example, is quadrivalent. Also called TETRAVALENT.

quadruped robot A robot with four legs that can move independently. It offers better stability than three-legged designs. Functions well in mobile machines that must navigate irregular terrain.

quadrupler **1.** A rectifier circuit that delivers a direct-current output voltage approximately equal to four times the peak value of the alternating-current input voltage. **2.** An amplifier or other circuit that delivers an output signal of four times the frequency of the input signal.

quadruplex circuit A data circuit in which two messages are carried in each direction simultaneously.

quadrupole **1.** A combination of two dipoles, producing a force that varies inversely with the fourth power of distance. **2.** A four-pole magnet used in some synchrotrons and linear accelerators to focus and bend a particle beam. **3.** A system consisting of two dipoles of equal and opposite direct moment.

qualification The quality-control or quality-assurance scheme used in the production of components, circuits, or systems. Certain minimum requirements must be met for a device to obtain qualification.

qualitative test A test performed to determine the general mode of operation or the presence of certain factors, without regard to numerical values. Compare QUANTITATIVE TEST.

quality **1.** In audio-frequency applications, fidelity of transmission or reproduction. **2.** The degree of conformity of a product to specifications.

quality assurance The outcome of measures taken to bring performance into conformity with specifications. See QUALITY, **2**.

quality control The surveillance of selection, manufacturing, and testing operations to ensure conformity of a product to specifications.

quality-control engineering The branch of engineering concerned principally with the technical methods of quality control and statistical methods of assessing quality (see QUALITY, **2**).

quality-control technician A technician whose principal duty is the performance of operations in the areas of incoming inspection, manufacturing support, and product testing. Sometimes statistical evaluations are required.

quality engineering A field of engineering that deals with quality assurance and quality control in the production of components, circuits, and systems.

quality factor See *Q*.

quality-factor bridge See *Q* BRIDGE.

quality-factor meter See *Q* METER.

quanta Plural of QUANTUM.

quantimeter An instrument used to measure the quantity of X rays to which a body has been exposed.

quantitative test A test performed to determine the numerical values (and their relationships) connected with observable phenomena. Compare QUALITATIVE TEST.

quantity **1.** A parameter (e.g., collector current, grid voltage, etc.). **2.** In calculations, a positive or negative real number. **3.** Electrical charge, usually specified or measured in coulombs (see COULOMB). Also called *electrical quantity.*

quantization The conversion of a quantity having infinitely many possible values or levels (such as an analog signal) into one that can attain only a finite number of defined values or levels (such as a digital signal). The number of levels is usually some integral power of 2 (i.e., 2, 4, 8, 16, 32, etc.). This allows the levels to be represented as binary numbers.

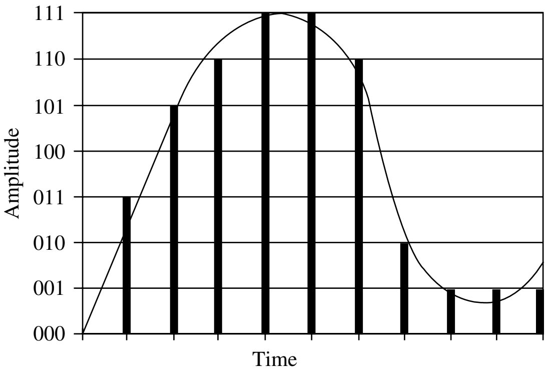

quantization

quantization distortion Distortion introduced by the process of QUANTIZATION in a communications or broadcast signal.

quantization distortion power The level of the distortion in a signal resulting from QUANTIZATION. It is expressed in microwatts, milliwatts, or watts. It can also be expressed as a percentage, or as a level in decibels, relative to the power level of the input signal.

quantization error The difference between the actual values of quantities and their quantized values.

quantization noise Noise introduced by the process of QUANTIZATION in a communications or broadcast signal.

quantize **1.** To perform the process of QUANTIZATION. **2.** To split a quantitative commodity, such as energy into its smallest measurable incremental units.

quantized pulse modulation Pulse modulation involving QUANTIZATION. Examples are PULSE-CODE MODULATION and PULSE-NUMBERS MODULATION.

quantizer A circuit or device that selects the digital subdivision into which an analog quantity is placed (e.g., an analog-to-digital converter).

quantizing See QUANTIZATION.

quantometer An instrument for measuring magnetic flux.

quantum **1.** Abbreviation, Q. Plural, *quanta.* In physics, the elemental unit or particle of electromagnetic energy. The energy contained in one such particle is directly proportional to the frequency, and inversely proportional to the wavelength. **2.** See PHOTON. **3.** Any discrete unit derived by QUANTIZATION.

quantum chromodynamics A term coined by Professor Murray Gell-Mann for the theory of *quarks* and *gluons.*

quantum counter A radiation-counter tube with a window for the admission of light to the cathode.

quantum efficiency See QUANTUM YIELD.

quantum electrodynamics A branch of quantum mechanics that involves the motions of electrons, photons, and muons caused by electromagnetic action. Quantum electrodynamics takes relativistic effects into account.

quantum electronics The branch of electronics concerned with energy states in matter.

quantum equivalence The principle that one electron is emitted for each photon absorbed by a material (when the photon has the necessary energy).

quantum jump The abrupt movement of a particle from one discrete energy state to another.

quantum level The orbit or ring occupied by an electron in an atom.

quantum mechanics A branch of physics concerned with the behavior of matter and energy, on the basis of observable data.

quantum noise A noise signal arising from random variations in the average rate at which quanta impinge upon a detector.

quantum number A number that describes the energy level, or change in energy level, for a particle.

quantum statistics A branch of QUANTUM MECHANICS concerned with the distribution of elementary particles through various quantized energy levels.

quantum theory The theory that the emission or absorption of energy by atoms or molecules occurs in discrete packages or units, rather than continuously. Each unit is the emission or absorption of an energy packet called a QUANTUM. Thus, radiant energy is thought to be divided into *quanta.*

quantum transition The movement of an electron from one energy level to another within an atom.

quantum yield The photoelectric efficiency of a light-sensitive surface in terms of the number of electrons emitted for each absorbed quantum of light.

quark A hypothetical particle having a fractional electrical charge; quarks are thought to be constituents of other subatomic particles.

quarter-deflection method A method of measuring high-frequency resistance, involving the use of a sine-wave signal source, a standard noninductive variable resistor, and a square-law radio-frequency ammeter.

quarter-phase See TWO-PHASE.

quarter wave **1.** The length of time corresponding to 90 electrical degrees in a wave disturbance. **2.** The distance in space, or along a wire or feed line that corresponds to 90 electrical degrees in a wave disturbance.

quarter-wave antenna An antenna in which the radiator is an electrical quarter-wavelength long at the operating frequency.

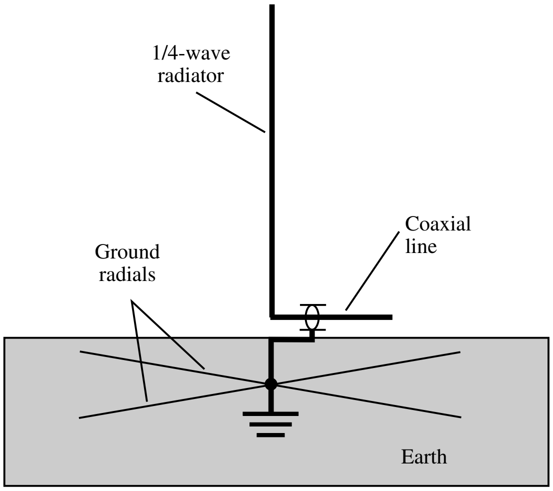

quarter-wave antenna

quarter-wave attenuator In a transmission line or waveguide, two energy-absorbing structures separated by an odd number of quarter wavelengths so that the reflection from one structure is canceled by that from the other.

quarter-wave balun A balun using quarter-wave elements. One form of this device consists of a grounded quarter-wavelength-long cylinder closed at one end and open at the other, for matching an unbalanced low-impedance line to a balanced high-impedance line.

quarter wavelength Symbol, $\lambda/4$. The distance that corresponds to 90 degrees of phase as an electromagnetic (EM) field is propagated. In free space, it is related to the frequency by a simple equation

$$L_{ft} = 246/f$$

where L_{ft} represents $\lambda/4$ in feet, and f represents the frequency in megahertz. If $\lambda/4$ is expressed in meters, then the formula is

$$L_m = 75/f$$

where L_m represents the displacement in meters. In general, if v is the velocity factor (expressed as a ratio) in a given medium, then:

$$L_{ft} = 246v/f$$

and

$$L_m = 75v/f$$

Compare FULL WAVELENGTH, HALF WAVELENGTH.

quarter-wavelength line A transmission line or feeder that is a quarter-wavelength long at the operating frequency. Also called *quarter-wave line.*

quarter-wavelength matching stub An arrangement consisting of a quarter-wavelength-long parallel-wire section of transmission line, used for matching the impedance of a nonresonant feeder to that of an antenna. It is similar to a *Q*-MATCHING SECTION, except that flexible transmission line (e.g., ladder line) is used, rather than rigid metal rods or tubing.

quarter-wave monopole A nondirectional UHF vertical antenna requiring no ground. The radiator is ¼ wavelength long, and is an enlarged-diameter outer sleeve connected to the outer conductor of the coaxial feeder. The two sections simulate a half-wave antenna.

quarter-wave plate A plate of double-refracting crystalline material whose thickness allows the introduction of a quarter-cycle phase difference between the ordinary and extraordinary components of light transmitted by it.

quarter-wave radiator An antenna consisting of a single, usually straight, active element that measures an electrical quarter wavelength from end to end. When operated against electrical ground, the element exhibits resonance. A simple quarter-wavelength ($\lambda/4$) conductor with a high length-to-diameter ratio measures approximately 95 percent of $\lambda/4$ in free space. The element can be much shorter than free-space $\lambda/4$ yet remain resonant when operated against electrical ground if inductance is inserted in series with the radiator. The element can be much longer than free-space $\lambda/4$ yet remain resonant when operated against electrical ground if distributed capacitances are inserted in series with the radiator.

quarter-wave resonance Resonance at the operating frequency in a quarter-wave antenna.

quarter-wave resonant line A section of transmission line (such as open-wire line or coaxial cable) that is a quarter-wavelength long at the operating frequency. Such a section is useful in impedance matching and in various radio-frequency tests and measurements.

quarter-wave stub See QUARTER-WAVE TRANSFORMER.

quarter-wave support In a coaxial line, a quarter-wave metal stub that can be used, instead of an insulator, to separate the inner and outer conductors.

quarter-wave termination In a waveguide, a set of two metal barriers separated by 90 electrical degrees. One barrier totally reflects the energy striking it. The other barrier allows some energy to pass through. Resonance occurs in the space between the barriers.

quarter-wave transformer A quarter-wave resonant line short-circuited at one end by an adjustable slider. This arrangement is useful for radio-frequency impedance matching.

quarter-wave transmission line See QUARTER-WAVE LINE.

quartic equation A fourth-degree equation of the form $ax^4 + bx^3 + cx^2 + dx + e = 0$, where a, b, c, d, and e are constants. Also called BIQUADRATIC EQUATION.

quartz A mineral that is a variety of natural silicon dioxide, or an artificially grown material of the same sort. In the natural state, quartz occurs in hexagonal crystals having pyramidal ends. It has various uses in electronics; one of the most common is the manufacture of piezoelectric crystals.

quartz bar A comparatively large, thick piezoelectric quartz plate used in standard-frequency oscillators and in sharply tuned low-frequency filters. Common resonant frequencies are 50 kHz, 100 kHz, and 1000 kHz.

quartz crystal A natural or artificial piece of quartz cut to specific dimensions, usually self-contained in a solder-in or plug-in enclosure. The device acts as a highly stable selective circuit. It exhibits a sharp resonance at the frequency for which it is cut, and at harmonics of this frequency. It is used as the frequency-determining element in precision oscillators.

quartz-crystal oscillator See CRYSTAL OSCILLATOR.

quartz-crystal resonator See CRYSTAL RESONATOR.

quartz delay line An acoustic delay line using quartz to transmit the sound waves.

quartz-fiber electroscope An electroscope using a gold-plated quartz fiber, instead of gold leaves.

quartz-halogen lamp An incandescent, usually low-voltage lamp used in automotive headlights, and in some home and office lighting appliances. It provides greater efficiency than conventional incandescent lamps.

quartz lamp A mercury-vapor lamp with a transparent quartz (instead of glass) envelope. Unlike glass, quartz readily passes the ultraviolet rays generated by the mercury discharge.

quartz lock A circuit that uses a CRYSTAL OSCILLATOR to regulate frequency, timing, or speed. It is used in electronic clocks and watches, television receivers, synthesized radio receivers, transmitters, transceivers, high-fidelity turntables, etc.

quartz oscillator See CRYSTAL OSCILLATOR.

quartz plate A piezoelectric plate cut from a quartz crystal. The plate is itself often called a *crystal.* Also see CRYSTAL AXES and CRYSTAL CUTS.

quartz resonator See CRYSTAL RESONATOR.

quartz timepiece A watch or clock having as its control element a time-determining quartz crystal.

quasi- A prefix meaning "to some extent" or "similar to," as in *quasi-optical radio wave* (a radio wave that behaves like a light ray).

quasi-bistable circuit A trigger-operated multivibrator. It operates as a flip-flop when the trigger frequency is sufficiently high.

quasi-instruction In a computer program, a data item appearing as an encoded instruction, but that is not acted upon.

quasi-linear feedback system A system in which the feedback elements are nearly linear, but not entirely linear.

quasi-negative Pertaining to a voltage that is negative, with respect to some other voltage, but whose absolute polarity is positive. For example, +0.5 volt is quasi-negative, with respect to +5.5 volts.

quasi-optical Behaving like light. The term is used to describe certain extremely short radio waves or other radiations that, like light rays, follow line-of-sight paths and can be directed, reflected, refracted, or diffused.

quasi-optical path A line-of-sight path followed by very short radio waves, such as microwaves.

quasi-positive Pertaining to a voltage that is positive, with respect to some other voltage, but whose absolute polarity is negative. For example, –0.5 volt is quasi-positive, with respect to –5.5 volts.

quasi-random A set of numbers considered to be random, but chosen according to an algorithm.

quasi-rectangular wave A wave whose shape approaches that of a rectangular wave, but that possesses a small amount of tilt and/or curvature.

quasi-scientific A term that is sometimes applied to the design of electronic systems or to the appraisal of circuit behavior, using an intuitive, rather than analytical approach.

quasi-sine wave A waveform that is not a perfect sine curve, but is close enough to be considered sinusoidal, for all practical purposes.

quasi-single sideband A modulated waveform that somewhat resembles single sideband, in which parts of both sidebands are present.

quasi-square wave A waveform that is not a perfect square, but is close enough to be considered square for all practical purposes. It is sometimes applied to a rectangular wave when a square wave is desired.

quasi-technical A term sometimes applied to qualitative tests, as opposed to quantitative tests.

quaternary **1.** Pertaining to a base-4 number system. **2.** Of an atom, joined to carbon atoms for four bonds. **3.** The fourth member of a 4-unit set.

Quebec Phonetic alphabet code word for the letter Q.

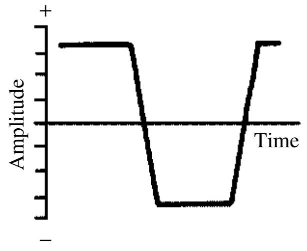

quasi square wave

quench **1.** To suddenly bring to an end (e.g., to quench an oscillation). **2.** To cool quickly, as in the quenching of a heated metal object. **3.** To extinguish the discharge in a gas tube.

quench capacitor A capacitor that prevents a spark from arcing across an inductor when current flow abruptly stops.

quench frequency See QUENCHING FREQUENCY.

quenching action Typical operation of a superregenerative circuit, in which regeneration is increased to nearly the point of oscillation and then reduced; this action is repeated at an ultrasonic frequency and results in very high sensitivity. Also see QUENCHING FREQUENCY, QUENCH OSCILLATOR, and SUPERREGENERATIVE CIRCUIT.

quenching frequency The frequency at which regeneration in a superregenerative circuit is increased and decreased.

quench oscillator In some superregenerative circuits, the separate ultrasonic oscillator that produces the required quenching action.

quench resistor A resistor in a quenching network that prevents a spark from occurring across an inductor when current flow stops.

queue A list of data, steps in a process, or commands awaiting execution in a specific order.

queuing theory A branch of mathematical electronics, dealing with the optimum order in which steps should be executed to obtain a particular end result.

quibinary code In computer operations, a binary-coded decimal system in which each decimal digit is represented by seven bits occupying places whose values are 8, 6, 4, 2, 0, 1, and 0.

quibinary decade circuit A decade counter consisting of a ring-of-5, followed by a single binary stage.

quick break An operating characteristic of a switch, relay, or circuit breaker whereby the contacts open rapidly—even when the actuating current or mechanical force is slow-acting.

quick-break fuse A fuse in which the wire melts and breaks almost instantly when the current rating is exceeded. Also called *quick-blow fuse.* Compare SLOW-BLOW FUSE.

quick-break switch A switch that opens rapidly—even if its handle or lever is moved slowly by the operator. This action minimizes arcing and prevents chatter. Compare QUICK-MAKE SWITCH.

quick charge The process of charging a battery, such as a nickel-cadmium (NICAD) or nickel-metal-hydride (NiMH) type, at a relatively rapid rate, at high charging current. It is sometimes used to charge a battery from a state of almost total discharge. Compare TRICKLE CHARGE.

quick-disconnect The characteristic of a connector that enables its mating halves to be separated quickly and simply, to break the circuit in which it is situated.

quickening liquid A solution of mercuric cyanide or mercuric nitrate, into which objects can be dipped prior to electroplating with silver. The process ensures good adhesion of the silver to the object.

quick make An operating characteristic of a switch, relay, or circuit breaker, whereby the contacts close rapidly—even when the actuating current or mechanical force is slow acting.

quick-make switch A switch that closes rapidly—even if its handle or lever is moved slowly by the operator. Compare QUICK-BREAK SWITCH.

quick printer A high-speed printer, used with a data terminal or computer. A relative term, depending on the user and the application.

quicksilver See MERCURY.

quick-stop control A control on tape recorders and some dictating machines that allows the operator to stop the tape, but keep the machine in the play or record mode. Also called *pause control.*

QuickTime Trade name (Apple Computer, Inc.) for system software commonly used in MULTIMEDIA applications with personal computers.

QUICKTRAN For multiaccess computer systems, a computer programming language based on FORTRAN and offering facilities, through the use of remote terminals, for running, testing, debugging, and compiling programs.

quiescent carrier operation A modulation system in which the carrier is present only during modulation (i.e., it is suppressed at all other times). Also called *controlled-carrier transmission.*

quiescent-carrier telephony A carrier-current (wired-wireless) telephone system in which the carrier is suppressed when there is no voice or alerting signal.

quiescent component In an electronic device, a component that is momentarily nonfunctional.

quiescent current Operating current (usually a direct current) flowing in a circuit or component during zero-signal or no-drain intervals. Also called IDLING CURRENT.

quiescent operation Zero-signal operation of a device, such as a transistor, diode, magnetic amplifier, or similar component.

quiescent period The no-signal interval during which equipment is not operating—even though it is energized.

quiescent point The point on the characteristic curve of a transistor, diode, or similar device, denoting the zero-signal operating conditions.

quiescent push-pull Denoting a push-pull stage, especially an audio power-output amplifier, in which the direct-current signal is essentially zero.

quiescent state The inactive, or resting, state of an active component, such as a transistor or vacuum tube.

quiescent value The zero-signal value of current or voltage for any component supplied with operating power.

quiet AGC See DELAYED AUTOMATIC GAIN CONTROL.

quiet automatic gain control See DELAYED AUTOMATIC GAIN CONTROL.

quiet automatic volume control See DELAYED AUTOMATIC GAIN CONTROL.

quiet AVC See DELAYED AUTOMATIC GAIN CONTROL.

quiet battery A direct-current source specially designed and filtered to minimize noise components in its output.

quieting Noise-voltage reduction in the output of a frequency-modulation (FM) receiver when an unmodulated carrier is received. Also called *noise quieting.*

quieting level In a frequency-modulation (FM) receiver, the limiter threshold point.

quieting sensitivity In a frequency-modulation (FM) receiver, the lowest input-signal amplitude at which the output signal-to-noise ratio is below the specified limit.

quiet tuning A system of tuning in which the output of a receiver is muted until a station is tuned in properly.

quinary code See BIQUINARY CODE.

quinary counter A decade counter consisting of a five-stage ring.

quinhydrone electrode A pH meter electrode consisting of a platinum wire in a solution of quinhydrone ($C_{12}H_{10}O_4$). Also see PH METER.

quintillion The number 1,000,000,000,000,000,000 (10^{18}).

quintupler **1.** A rectifier circuit that delivers a direct-current output voltage equal to about five times the peak value of the alternating-current input voltage. **2.** A circuit that delivers an output signal at the fifth harmonic of the input signal.

QWERTY The standard typewriter and computer keyboard layout. The name is derived from the first several letters in the top letter row: Q, W, E, R, T, and Y.

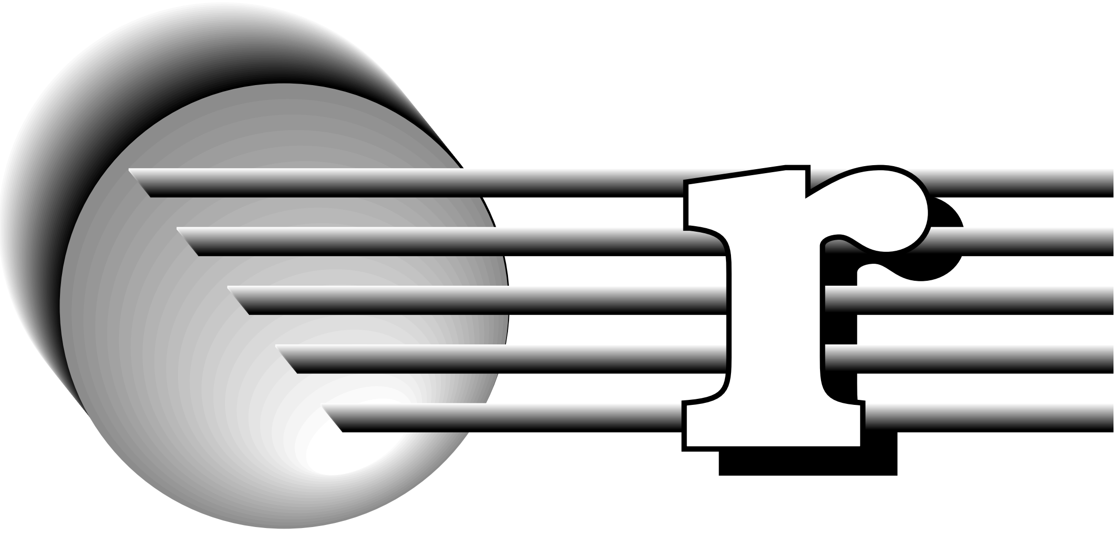

R **1.** Symbol for RESISTANCE. (Also, r.) **2.** Radiotelegraph abbreviation for ROGER. **3.** Symbol for RELUCTANCE. **4.** Abbreviation of RADIUS.

r **1.** Symbol for ROENTGEN. **2.** Symbol for *correlation coefficient.* **3.** Abbreviation of RADIUS. (Also, abbreviated R.)

r_e Symbol for CLASSICAL ELECTRON RADIUS.

RA **1.** Abbreviation of *right ascension.* **2.** Abbreviation of RANDOM ACCESS.

rabbit ears An indoor antenna, sometimes used with a television receiver, consisting of two vertical whips (usually telescoping), the angle between which is adjustable.

RAC Abbreviation of RECTIFIED ALTERNATING CURRENT.

R_{ac} Symbol for AC RESISTANCE. (Also, r_{ac}.)

race Incorrect interpretation of the clock pulses by a digital circuit. Also called *racing.* The circuit improperly attempts to do many operations during one clock pulse, rather than a single operation.

RACES Abbreviation of *Radio Amateur Civil Emergency System.*

raceway See WIRE DUCT and WIREWAYS.

rack An upright frame for holding equipment of RACK-AND-PANEL CONSTRUCTION.

rack-and-panel construction A method of building electronic equipment on a chassis attached horizontally or vertically to a vertical panel. After completion of a unit, the panel is fastened in place on a RACK. Several such panels fill the rack.

rack and pinion A device used for mechanical adjustment of a control, such as the tuning control in a radio receiver. A gear engages a serrated rod. As the gear is turned, the rod moves lengthwise.

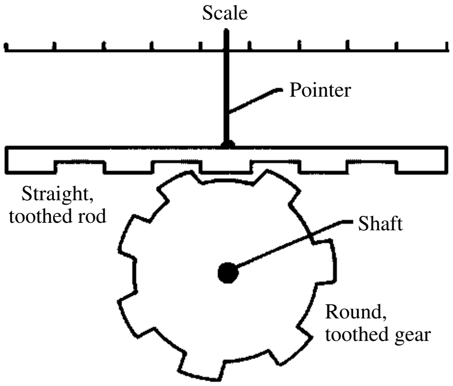

rack and pinion

rack up In computer operations, a way of displaying data, in which a new line added to the already completely occupied screen bumps up what has forgone, thus eliminating the top line.

racon Acronym for *radar beacon.*

rad **1.** A unit of ionizing radiation received by a body (dose) equal to 0.001 J/kg. **2.** Abbreviation of RADIAC. **3.** Abbreviation of RADIAN. **4.** Abbreviation of RADIO. **5.** Abbreviation of RADIX.

radar **1.** A microwave system for detecting objects and determining their distance, direction, heading, speed, and other characteristics. Signals from the transmitter are reflected back to the transmitter site by the object, and the reflection (sometimes along with the transmission) is displayed on a cathode-ray screen. The name is an

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

acronym for *radio detection and ranging.* **2.** The theory and application of radio detection and ranging systems as defined in **1**.

radar altitude The distance of an aircraft above the surface of the earth, as determined by radar. This value varies with the terrain over which the aircraft passes.

radar antenna Any antenna used for transmitting and/or receiving radar signals.

radar astronomy The use of radar equipment to observe and map planets, moons, and asteroids, and to measure their distance from the earth or from a spacecraft.

radar beacon A radar transceiver that, on receipt of radar signals, transmits encoded signals from which the operator can take a bearing.

radar beam The cone-shaped main lobe of energy emitted by a radar antenna. The narrower the beam, the greater the resolution of the radar system.

radar clutter Visual interference on a radar screen caused by reflections from ground or sea.

radar countermeasures Abbreviations, RCM and rad CM. In wartime, any method of interfering with enemy radar, such as jamming or use of decoys.

radar detector **1.** A device used in automobiles and trucks to detect the proximity of police or highway-patrol radar. **2.** A device used in military applications, especially aviation, to indicate the presence of radar.

radar display The scheme via which a radar set portrays the relative positions of the objects that produce echoes. The most common is the azimuth/range display, showing compass bearings (usually in degrees clockwise from geographic north) and radius from the transmitter (usually in miles or kilometers). Altitude above mean sea level can be displayed for individual echoes.

radar homing A method of missile homing in which radar is used to track a target.

radar speed trap A radar system used by traffic police to spot speeding vehicles.

radar telescope The transmission and reception unit used in radar astronomy. Compare RADIO TELESCOPE.

RadCM Abbreviation of RADAR COUNTERMEASURES. (Also, RCM.)

radial **1.** One of several conductors used to enhance the performance of an unbalanced, vertical antenna. These can be constructed from wire or metal tubing, and generally measure one-quarter wavelength or more. When a vertical antenna is mounted at the earth's surface, the ground conductivity is improved by these conductors, which run outward from the base of the radiator, and are connected to the shield of a coaxial feed line. The greater the number of radials of a given length, the more the ground loss is reduced. Also, the longer the radials for a given number, the better. If a vertical radiator is mounted with its base more than one-quarter wavelength above the earth's surface, there need only be three or four conductors measuring exactly one-quarter wavelength. See GROUND-MOUNTED VERTICAL ANTENNA, GROUND-PLANE ANTENNA. **2.** Pertaining to the distance from the center of a circle to its periphery. **3.** Pertaining to the distance from the center of a sphere to its surface. **4.** Extending or emanating outward in a straight line from a defined point.

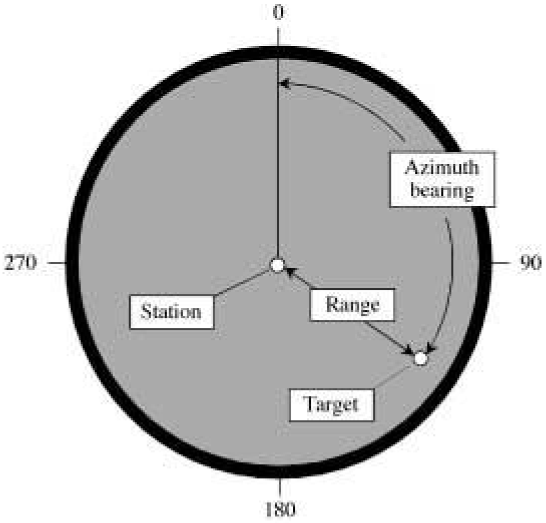

radar display

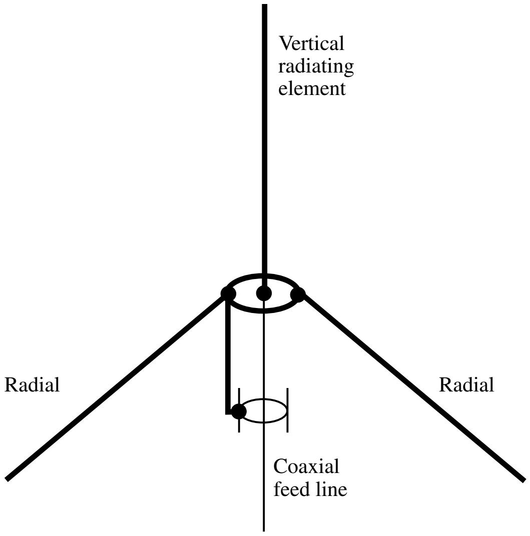

radials

radial ground An earth connection composed of radials buried in the ground.

radial lead A lead (pigtail) attached perpendicular to the axis of a component, such as a resistor or capacitor.

radian Abbreviation, rad. The angle at the center of a circle subtended by an arc whose length is equal to the radius. Equal to approximately 57.2958 degrees.

radiance The radiant flux emitted by an object. Radiance is measured in terms of the amount of energy contained in a unit solid angle (steradian) with the source at the apex.

radians-to-degrees conversion The conversion of radian angular measure into degrees. To change radians to degrees, multiply the number of degrees by 57.2958. Thus, 0.7854 radian = 45 degrees. Compare DEGREES-TO-RADIANS CONVERSION.

radiant efficiency The ratio of the radiant energy emitted by a source to the energy consumed by the source. The radiant energy is generally specified within a certain range of wavelengths. An example is the incandescent light bulb, which has relatively low radiant efficiency in the visible spectrum between about 750 and 390 nanometers.

radiant energy **1.** Any form of energy emitted by a source and propagated through space as an electromagnetic disturbance. Included are radio waves, infrared, visible light, ultraviolet, X rays, and gamma rays. **2.** Electromagnetic disturbances at infrared and shorter wavelengths.

radiant flux The rate at which radiant energy is emitted.

radiation **1.** The emission of energy or particles (e.g., waves from an antenna, X rays from an X-ray tube, energy from a radioactive material, heat from a body, etc.). **2.** Radio waves, infrared, visible light, ultraviolet, X rays, or gamma rays. **3.** Ionizing emissions from radioactive substances (e.g., alpha particles, beta particles, neutrons, gamma rays, etc.).

radiation angle The horizontal or vertical angle at which electromagnetic waves are radiated from an antenna. Measured between the central axis of the main lobe and the horizon, or between the central axis of the main lobe and geographic north.

radiation belts See VAN ALLEN RADIATION BELTS.

radiation counter An instrument used for determining the intensity of atomic-particle radiation, X rays, or gamma rays. It operates by means of ionization of a gas in a sealed tube.

radiation field The portion of the electromagnetic field that is propagated by a radiator, as opposed to the induction field.

radiation intensity For a directional radio transmitting antenna, the radiated power per steradian in a given direction.

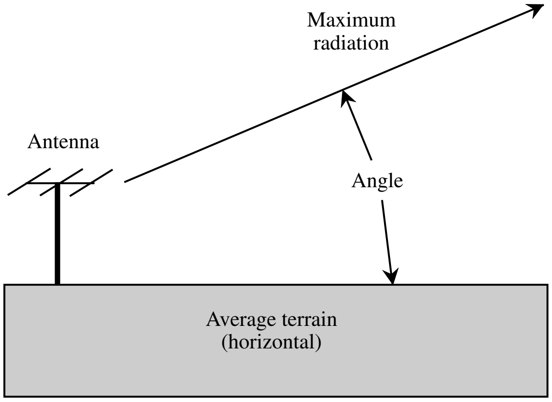

radiation angle

radiation loss Loss of energy through radiation from a conductor. Also see LOSS.

radiation pattern A graphical representation of the intensity of the electromagnetic field in various directions from a radiator, such as a transmitting antenna. It is usually shown in either the horizontal plane or a specific vertical plane containing the antenna. Also see LOBE.

radiation physics The study of radiation and its effects on matter. Radiation physics is especially concerned with ionizing radiation, but it can involve any kind of particle or electromagnetic energy.

radiation pressure Pressure exerted on a surface by impinging electromagnetic radiation.

radiation resistance The inherent resistance at the feed point of a resonant radio antenna.

radiation sickness General physiological symptoms resulting from a short-term overdose of X rays, gamma rays, or atomic-particle radiation.

radiator **1.** The element of an antenna from which radio energy is directly radiated, as opposed to the transmission line, lead-in, reflector, or director. **2.** See LOUDSPEAKER.

radio **1.** Wireless electrical communication, i.e., by means of electromagnetic waves. **2.** See RADIO RECEIVER. **3.** See RADIO TRANSCEIVER. **4.** See RADIO TRANSMITTER. **5.** To communicate by radio.

radio- **1.** A prefix meaning "pertaining to wireless electrical communication." Examples: *radiotelephone* and *radiotelegraph.* **2.** A prefix meaning "using radio waves." Examples: *radiosonde*, *radiolocator*, and *radiothermics.* **3.** A prefix meaning "pertaining to using or possessing radioactivity," or "pertaining to X rays." Examples: *radiograph*, *radioisotope*, and *radiologist.*

radioactive Having the property of emitting alpha, beta, and (sometimes) gamma rays as the result of nuclear disintegration. Also see HALF-LIFE.

radioactive element A chemical element that is RADIOACTIVE (e.g., uranium). Also called *radioelement.*

radioactive isotope See RADIOISOTOPE.

radioactive tracer A quantity of radioactive material put into a system so that its path can be monitored by means of a radiation detector. An example is the introduction of radioactive barium into the large intestine. The flow and concentration of the barium gives an indication of the functioning of the lower intestine.

radioactive transducer A pickup device for detecting and measuring radioactivity (e.g., *Geiger-Mueller tube*).

radioactivity counter See GEIGER COUNTER and SCINTILLATION COUNTER.

radio altitude See RADAR ALTITUDE.

radio amateur An electronics hobbyist licensed to operate two-way wireless communications stations in various assigned frequency bands, without receiving payment for services rendered.

Radio Amateur Civil Emergency System Abbreviation, RACES. A civil-defense organization of licensed amateur radio stations. Also see RADIO AMATEUR.

radio astronomy The observation, study, and analysis of radio-frequency electromagnetic emissions from bodies or points in space, and the study of these bodies through their radiations.

radioautograph See AUTORADIOGRAPH.

radio beacon **1.** A radio transmitter of direction-finding or guidance signals. **2.** Also called *radio beam.* The signals transmitted by a radio beacon, as defined in **1**.

radio beam **1.** Antenna radiation focused in one direction. **2.** See RADIO BEACON, **2**.

radiobiology A field of biology concerned with the influence of radiant energy or radioactivity on living organisms.

radio broadcast A radio transmission directed to numerous, nonspecific receivers—especially by a station in the *broadcast service.* Also called RADIOCAST. Also see BROADCAST SERVICE, **1**, **2**.

radio car An automobile equipped with a two-way radio.

radio carbon Radioactive carbon (i.e., carbon 14).

radiocast See RADIO BROADCAST.

radio channel A single, usually narrow radio-frequency band within a larger band, in which stations are authorized to transmit signals of a specified type. Also see CHANNEL, **1**; CHANNEL SEPARATION; and CHANNEL WIDTH.

radiochemistry The chemistry of radioactive substances.

radio communication Wireless communication carried on by means of radio-frequency electromagnetic waves.

radio compass See DIRECTION FINDER.

radioconductor A substance or body whose electrical conductivity is affected by radio waves, and that can be used as a sensor of such waves.

radio control See REMOTE CONTROL.

radio direction finder See DIRECTION FINDER.

radio Doppler **1.** A change in the frequency of a radio signal emitted by a source having radial motion, with respect to the receiver. **2.** An electronic device used to measure radial speed by means of the Doppler effect at radio frequencies.

radio-electronics The branch of electronics specifically involved with wireless communications.

radioelement See RADIOACTIVE ELEMENT.

radio engineer A trained professional skilled in the physics and mathematics of radio communications, and in the theory and application of basic electronics engineering and related subjects. Also see RADIO ENGINEERING.

radio engineering The branch of electronics engineering devoted to the theory and operations of radio communication.

radio field strength The intensity of radio waves at a given point. Also see FIELD INTENSITY, **2** and RADIO MAP.

radio frequency Abbreviation, RF. **1.** Consisting of, or pertaining to, alternating currents at frequencies above about 9 kHz (the lowest allocated radio communications frequency). **2.** Consisting of, or pertaining to, electromagnetic fields whose wavelengths are longer than those of infrared, but shorter than about 33 kilometers (corresponding to a frequency of 9 kHz). Also see RADIO SPECTRUM.

radio-frequency amplifier **1.** In a superheterodyne circuit, the channel in which the incoming signal is amplified. Compare INTERMEDIATE-FREQUENCY AMPLIFIER. **2.** Broadly, an amplifier of radio-frequency signals.

radio-frequency choke Abbreviation, RFC. A low-inductance coil used to block radio-frequency (RF) alternating currents. Many RF chokes have air cores; some have cores of ferrite or powdered iron.

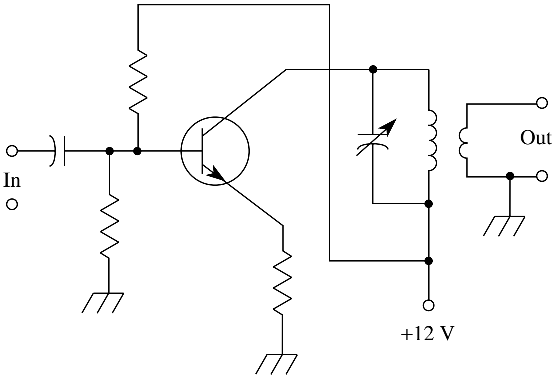

(General bipolar circuit)

radio-frequency amplifier, 2

radio-frequency current Symbol, I_{RF}. **2.** The intensity of a generated radio-frequency (RF) signal, usually expressed in microamperes. **2.** Loosely, any measurable RF signal.

radio-frequency heating The generation of heat in an object by an intense radio-frequency electromagnetic field. See, for example, DIATHERMY, **1**; DIELECTRIC HEATING; and INDUCTION HEATING.

radio-frequency interference Abbreviation, RFI. **1.** Annoying electrical noise in radio-frequency (RF) amplifiers, detectors, and instruments. **2.** Undesired RF signals that compete with desired ones in amplifiers, receivers, and instruments. **3.** The unwanted interception and demodulation of a strong RF signal by an audio-frequency (AF) device, such as a telephone set or high-fidelity stereo amplifier.

radio-frequency meter An instrument for measuring signals of RADIO FREQUENCY (9 kHz and above).

radio-frequency oscillator Abbreviation, RFO. An oscillator (self-excited or crystal-controlled) for operation at radio frequencies. In such an oscillator, stray components, efficiency, and general losses are of primary concern. Also see RADIO FREQUENCY.

radio-frequency power Symbol, P_{RF}. Alternating-current power at radio frequencies.

radio-frequency resistance The total in-phase resistance exhibited by a conductor at radio frequencies. This opposition to current includes direct-current resistance and the in-phase components caused by skin effect, shielding, and the presence of dielectrics.

radio-frequency selectivity The selectivity of a radio-frequency (RF) channel, such as the RF amplifier and first detector of a superheterodyne circuit.

radio-frequency transformer Abbreviation, RF transformer. A device used for the purpose of impedance matching, antenna tuning, or interstage coupling at frequencies above the range of human hearing (approximately 20 kHz and up). The device might consist of solenoidal windings with an air core, solenoidal windings with a powdered-iron or ferrite core, or toroidal windings with a powdered-iron or ferrite core. The windings are designed to minimize distributed capacitance and direct-current resistance. When no reactance is present, the impedance-transfer ratio is equal to the square of the turns ratio. Compare AUDIO-FREQUENCY TRANSFORMER.

radio-frequency transistor **1.** A transistor capable of providing significant amplification at radio frequencies. **2.** A transistor operable at frequencies above 100 kHz.

radiogenic Produced by radioactivity.

radiogoniometer A radio compass (see DIRECTION FINDER and GONIOMETER, **1**).

radiogram A (usually printed out) message transmitted and received via radiotelegraphy or radioteletype. The term is an acronym for *radio telegram.*

radiograph **1.** To contact by sending a RADIOGRAM. **2.** An X-ray photograph.

radio homing **1.** A method of homing that uses the tracking of a target on the basis of a radio signal emitted by that target. **2.** A method of keeping a missile on track via radio remote control.

radio interference **1.** Interference to radio communication, from whatever cause. **2.** See RADIO-FREQUENCY INTERFERENCE.

radioisotope A radioactive isotope (natural or artificial) of a normally nonradioactive chemical element (e.g., radioactive carbon). Also see ISOTOPE.

radio jamming See JAMMING.

radio knife A surgical instrument consisting essentially of a needle that forms a high-frequency arc. The arc simultaneously cuts and cauterizes tissue.

radiolocation A process whereby the position of a vehicle, aircraft, or ocean-going vessel is determined. The simplest method is the directional method. Two or three fixed receiving stations are used. Radio direction-finding (RDF) equipment is employed at each station, in conjunction with a transmitter aboard the vessel, to establish the bearings of the vessel with respect to each station. Radar can also be used to locate vessels. In wartime, enemy craft can sometimes be located by visual or infrared apparatus. Satellites can locate enemy ships and missiles, in some cases with an error smaller than the length of the vessel itself. Compare RADIONAVIGATION.

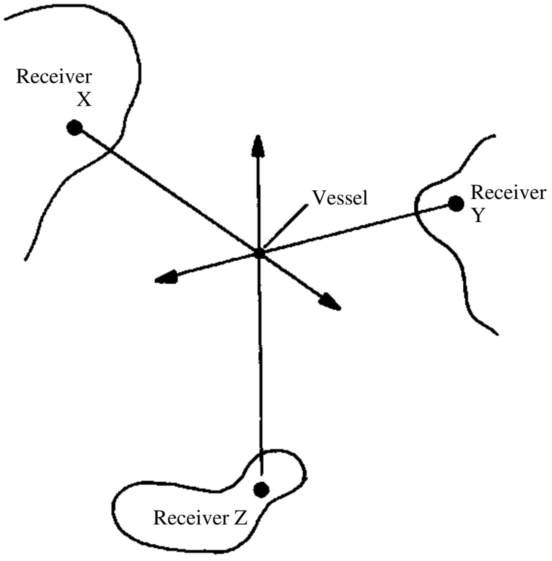

radiolocation

radiolocator See RADAR.

radiological system See X-RAY THERAPY SYSTEM.

radiologist A specialist skilled in RADIOLOGY.

radiology The science embracing the theory and use of X rays and radioactive substances in the diagnosis and treatment of diseases and ailments.

radiolucency **1.** The property of a material that allows ionizing radiation to pass through it with little or no absorption. **2.** The extent to which a material transmits ionizing radiation.

radioluminescence Visible light emitted from a radioactive material. A good example is radium; it was once used on wristwatch dials so that they could be seen in the dark.

radiolysis Chemical decomposition brought about by radiation.

radioman A radio technician or operator.

radio map A map of a geographic area, on which lines are drawn connecting measured points of equal field strength for signals from a radio station at the approximate center of the area.

radiometeorograph See RADIOSONDE.

radiometer A device for detecting and measuring the strength of radiant energy. One form consists of a set of vanes blackened on one side and mounted on pivots in a partially evacuated glass bulb. Visible light or infrared causes the vane assembly to rotate, the speed being proportional to the intensity of the light.

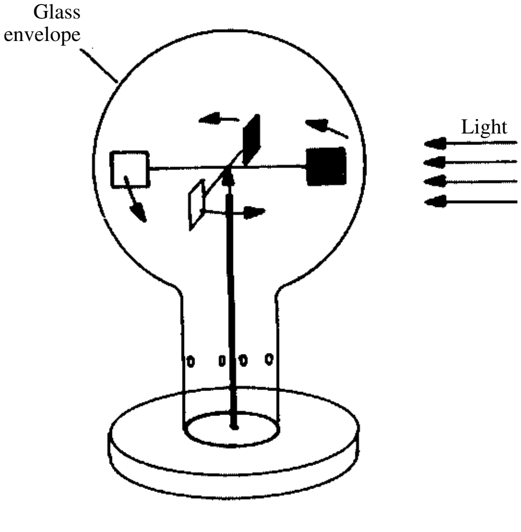

radiometer

radiometry The science and art of measuring radiation in the infrared, visible, and ultraviolet regions of the electromagnetic frequency spectrum. Compare PHOTOMETRY.

radio micrometer See MICRORADIOMETER.

radionavigation The use of radio apparatus, usually in conjunction with computers, by personnel aboard moving vessels, for the purpose of plotting, correcting, and maintaining a course. The intersecting-line method is simplest. Two or three land-based transmitters are needed. Their locations must be accurately known. A direction-finding device on the vessel is used to determine the bearings of each of the transmitters. Aircraft radionavigation can be performed with the aid of radar. The most sophisticated radionavigation techniques employ the Global Positioning System (GPS). Computers are used to project the course of a craft based on its current position, its speed, and the direction of its movement. Course corrections are made by choosing the desired course and having the computer calculate speed and/or direction changes. Compare RADIOLOCATION.

radio net A group of radio stations operating together in an organization, often on or near the same frequency.

radio network See RADIO NET.

radio operator A technician licensed to operate a transmitter in the radio, television, or radar services.

radiopaque Opaque to X rays or other ionizing radiation. Compare RADIOPARENT.

radioparent Transparent to X rays or other ionizing radiation. Compare RADIOPAQUE.

radiophone See RADIOTELEPHONE.

radiophoto A photograph transmitted and received by radio. Also see FACSIMILE.

radio pill See ENDORADIOSONDE.

radio prospecting The use of radio-frequency devices to locate underground or underwater metals and mineral deposits. Also see METAL LOCATOR.

radio range A radio station providing navigational aid to airplanes.

radio receiver The complete apparatus that selects, amplifies, demodulates, and reproduces a radio signal for purposes of communication, as distinct from *facsimile receiver*, *remote-control receiver*, *telemetry receiver*, *television receiver*, etc.

radiosensitivity **1.** The property of being sensitive to ionizing radiation. Most photographic films have this property. **2.** The extent to which a substance or device is sensitive to ionizing radiation.

radio service technician An electronics technician skilled in the repair and maintenance of radio equipment—especially receivers.

radiosonde A balloon-carried combination of radio transmitter and transducers, for sending to a ground monitoring station signals revealing such atmospheric conditions as temperature, humidity, and pressure. It is used mainly for gathering meteorological data at high altitudes.

radiosonobuoy See SONOBUOY.

radio spectroscope A device used by radio astronomers to obtain the radio-frequency profile of a distant star or galaxy. It generally consists of a graph, obtained by scanning the radio spectrum and plotting signal intensity as a function of frequency or wavelength.

radio spectrum The continuum of frequencies useful for radio communication and control. Classified in the following manner: *Very low frequency (VLF)*, 9 to 30 kHz; *low frequency (LF)*, 30 to 300 kHz; *medium frequency (MF)*, 300 to 3000 kHz; *high frequency (HF)*, 3 to 30 MHz; *very high frequency (VHF)*, 30 to 300 MHz; *ultrahigh frequency (UHF)*, 300 to 3000 MHz; *super high frequency (SHF)*, 3 to 30 GHz; *extremely high frequency (EHF)*, 30 to 300 GHz.

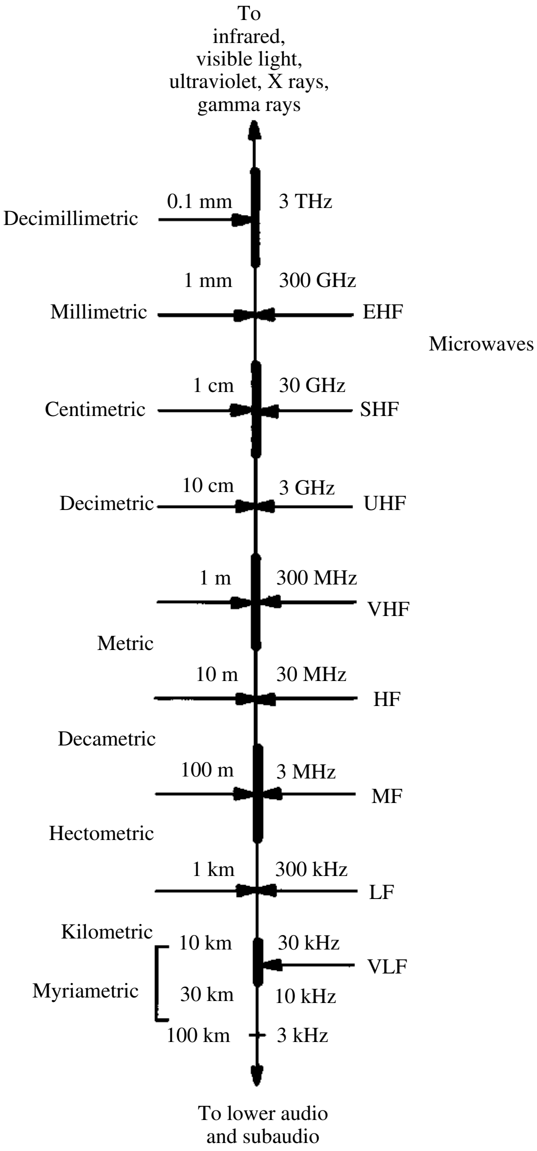

radio spectrum

radiostat See CRYSTAL FILTER.

radio station **1.** The location at which a radio transmitter and/or receiver is/are installed. **2.** The complete set of equipment for a radio receiving and/or transmitting installation, including the studio, linking apparatus, and antennas. **3.** A standard broadcast station.

radio technician A professional skilled in the construction, testing, repair, and maintenance of radio equipment, and sometimes in its design, and who usually works under the supervision of a radio engineer. Also see RADIO SERVICE TECHNICIAN.

radiotelegram See RADIOGRAM.

radiotelegraph **1.** Pertaining to the theory and application of, and the equipment for, Morse code transmission and/or reception via radio. **2.** An installation for Morse code transmission and/or reception via radio. **3.** The transmission and/or reception of Morse code signals via radio.

radiotelegraph code See CONTINENTAL CODE.

radiotelegraph distress signal See SOS.

radiotelegraph monitor See KEYING MONITOR.

radiotelegraphy The transmission and/or reception of telegraphic communications, usually Morse code, by means of radio.

radiotelephone **1.** Pertaining to the theory and application of, and the equipment for, voice transmission and/or reception via radio. **2.** An installation for voice transmission and/or reception via radio. **3.** The transmission and/or reception of voice signals via radio.

radiotelephone distress signal See MAYDAY.

radio/telephone patch See PHONE PATCH.

radiotelephony The transmission and/or reception of audio signals, usually human voices, by means of radio.

radio telescope A directional antenna and associated equipment for receiving and evaluating the radio-frequency electromagnetic radiation from space—especially from celestial objects (such as the sun, planets, stars, nebulae, galaxies, quasars, etc.). See RADIO ASTRONOMY.

radioteletype **1.** Pertaining to the theory and application of, and the equipment for, text data transmission and/or reception via radio. **2.** An installation for text data transmission and/or reception via radio. **3.** The transmission and/or reception of text data signals via radio.

radioteletypewriter A teletypewriter adapted to radio, rather than wire service; it is used in some RADIOTELETYPE installations. In recent years, personal computers and terminals have largely replaced adapted teletypewriters for this purpose.

radiotherapy The use of X rays and/or radioactive substances in the treatment of disease and disorders.

radiothermics The science of the generation of heat by means of radio-frequency current.

radiothermy See DIATHERMY.

radio thorium Radioactive THORIUM.

radiotracer See TRACER.

radio transceiver A RADIO RECEIVER and RADIO TRANSMITTER built into a single unit, and generally intended for use in two-way communication.

radio transmitter The complete apparatus that generates radio-frequency power, modifies it with the data needed for communication, and delivers the product to an antenna for radiation into space. Here, the radio transmitter is distinguished from similar equipment: *facsimile transmitter*, *remote-control transmitter*, *telemetry transmitter*, *television transmitter*, etc.

radio-transparent material 1. A substance through which radio waves pass with little or no attenuation. **2.** A substance through which X rays, gamma rays, or high-speed subatomic particles can pass with little or no attenuation.

radiotrician Acronym for *radio electrician.* See RADIO SERVICE TECHNICIAN.

radio tube 1. A VACUUM TUBE used at radio frequencies. **2.** A vacuum tube used as an amplifier, local oscillator, detector, or mixer in an early radio receiver.

radiovision See TELEVISION.

radio watch See WATCH.

radio waves Electromagnetic waves in the RADIO SPECTRUM.

radio window That portion of the radio-frequency electromagnetic spectrum that passes through the atmosphere, rather than being refracted or absorbed. The wavelength range is about 20 meters to 5 millimeters, or 15 MHz to 60 GHz. The lower limit of this range is affected by ionospheric conditions. The upper frequency limit depends on various factors, including relative humidity and dust content of the air.

radium Symbol, Ra. A rare radioactive metallic element. Atomic number, 88. Atomic weight, 226.025.

radius The straight-line distance from the center of a circle or sphere to its periphery.

radius vector In spherical or polar coordinates, a line segment drawn from the pole, or origin, and representing the vector magnitude.

radix The number indicating the number of symbols in a system of numerical notation, and the powers of which give the place values of the system. Thus, 10 is the radix of the decimal system, and 2 is the radix of the binary system. Also called BASE.

radix point In a number, the point (dot or period) separating the integral and fractional digits. Its specific name depends on the system of notation involved: *binary point*, *decimal point*, etc.

radome A plastic shell housing a radar antenna—especially aboard an aircraft.

radon Symbol Rn. A gaseous radioactive element that results from the disintegration of radium. Atomic number, 86. Atomic weight, 222.

rad/s Abbreviation of *radians per second*, the SI unit of angular velocity.

rad/s^2 Abbreviation of *radians per second squared*, the SI unit of angular acceleration.

radux A continuous-wave, low-frequency radionavigation system. Position is determined by comparing the phase of two signals sent from different locations.

RAID Acronym for REDUNDANT ARRAY OF INDEPENDENT DISKS.

rainbow generator A test-signal generator that produces a full color spectrum, a pattern resembling the successive coloration of a rainbow, on the screen of a color-television receiver. Also see RAINBOW PATTERN.

rainbow pattern A test pattern for servicing a color-television receiver. It consists of a full color spectrum, thus taking its name from its resemblance to a rainbow. Also see RAINBOW GENERATOR.

RAM Abbreviation of RANDOM-ACCESS MEMORY.

ramp A sawtooth wave with a linear rise and a practically instantaneous decay; its name was derived from its resemblance to an incline.

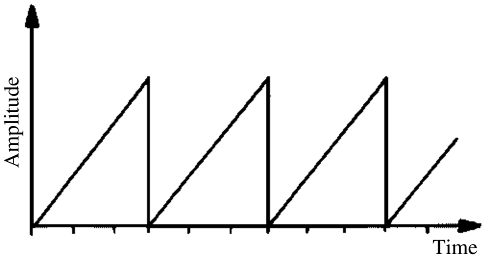

ramp

ramp generator A test-signal generator that produces sawtooth-wave signals. Also see RAMP.

R and D Abbreviation of research and development or research and design. (Also, R&D.)

random access Abbreviation, RA. In computer and data-processing operations, pertaining to storage or memory in which data can be recovered in any order.

random-access memory In computer and data-processing systems, a memory providing access time that is independent of the address.

random deviation Irregular RIPPLE.

random-fed antenna An antenna that uses RANDOM FEED.

random feed A method of connecting a transmission line to an antenna, wherein the feed point is not necessarily at the center and not necessarily at a current loop or voltage loop. This technique is rarely used; it generally results in some radiation from the feed line.

random noise Electrical noise in which the pulses or fluctuations have no discernible pattern of occurrence (i.e., they are haphazard in frequency and amplitude).

random number A number derived by chance. It is used in statistical analysis for various purposes.

random number generator Hardware or software that provides a sequence of numbers or digits that are random for the purpose of a given statistical application.

random occurrence See CHANCE OCCURRENCE.

random variable In statistics, a variable that can have a number of values, each of the same probability.

random winding A coil winding in which the turns are wound haphazardly to reduce distributed capacitance.

range **1.** The limits within which a circuit or device operates (i.e., the territory defined by such limits). Examples: *current range*, *frequency range*, and *voltage range*. **2.** The difference between the upper and lower limits of deflection of a meter. **3.** The distance over which a transmitter operates reliably. **4.** A clear area for testing antennas. **5.** The distance between a radar station and a target. **6.** The possible values for a quantity or function that lie between given limits.

range capacitor See TRIMMER CAPACITOR.

range-height indicator Abbreviation, RHI. A radar display in which the horizontal axis shows distance to the target, and the vertical axis shows elevation of the target.

range mark See DISTANCE MARK.

range plotting The creation of a graph of the distance (range) to objects, as a function of direction or orientation in two or three dimensions. Commonly used in robot guidance systems.

range resistor See TRIMMER RESISTOR.

range sensing The measurement of distances to objects via electronic methods such as radar, sonar, vision systems, etc. Commonly used in robot guidance systems.

ranging **1.** Any means of determining the distance from a station or vehicle to an object or objects. **2.** Any of several methods for a vehicle, vessel, aircraft, spacecraft, or robot to navigate in its environment by measuring, and keeping track of changes in, the distance between itself and one or more objects or beacons.

rank **1.** To arrange in a specific sequence according to significance. **2.** A place in such a sequence.

Rankine scale A temperature scale on which the freezing point of water is 491.69 degrees, and the boiling point 671.69 degrees. Absolute zero is represented by 0 degrees. For conversion to kelvins, multiply degrees Rankine by 5/9.

rapid drift A fast change of a quantity or setting (usually in one direction) with time.

rapid printer See QUICK PRINTER.

raser A device that produces coherent electromagnetic waves at radio frequencies; the radio-frequency equivalent of a LASER.

raster The rectangle of light (composed of unmodulated lines) seen on the screen of a television picture tube when no signal is present.

ratchet circuit See COMMUTATOR, **2** and ELECTRONIC RATCHET.

rate action See DERIVATIVE ACTION.

rate effect In a four-layer semiconductor device, the tendency for the switch to conduct undesirably as a result of a transient spike.

rate-grown transistor See GRADED-JUNCTION TRANSISTOR.

rate gyro A special gyroscope for measuring angular rates.

rate of change **1.** The extent to which the value of a dependent variable changes in accordance with a specified change in an independent variable (usually time). **2.** A quantitative expression of the speed with which a dependent variable changes, with respect to an independent variable (usually time).

rate signal A signal whose amplitude is proportional to the derivative of a variable, with respect to time.

rate time In automatic-control operations, the time over which the addition of DERIVATIVE ACTION advances PROPORTIONAL ACTION.

ratio-arm bridge A simple four-arm bridge in which the balancing potentiometer supplies the two arms, one on each side of the slider at all settings.

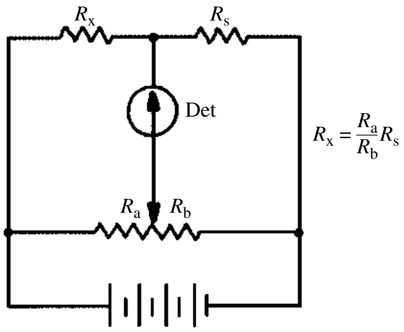

ratio-arm bridge

ratio arms Two impedance arms serving to establish the numerical ratio of a bridge circuit.

ratio control In automatic-control operations, a system in which the controlled variable is in a prescribed ratio to another variable.

ratio detector A frequency-modulation (FM) second detector resembling the FOSTER-SEELEY DISCRIMINATOR, except that one of the two diodes is reversed and the junction point of the load resistors is grounded. In an FM circuit using a ratio detector, no limiter is required. The ratio of the direct-current outputs is proportional to the ratio of the instantaneous signal voltages applied to the two diodes.

ratio meter An instrument that compares two different signals (and indirectly their sources) and delivers a reading of their ratio.

rational number A number that can be expressed in the form a/b, where a and b are integers and b is not equal to zero.

rational operation Any of the conventional arithmetic operations: multiplication, division, addition, or subtraction.

ratio of geometric progression In a geometric progression, the ratio of one value to the next.

ratio of similitude The ratio of the lengths of corresponding sides in similar geometric figures.

rat race See HYBRID RING.

raven red A variety of red oxide of iron, a commercial red paint used as the magnetic coating of early recording tapes.

raw ac Unrectified alternating current (ac) or voltage.

raw data Data that has not been processed in any way.

rawinsonde A RADIOSONDE tracked by a radio direction finder to determine wind velocity. The name is an acronym from *radar wind radiosonde.*

raw tape See BLANK TAPE.

ray **1.** A line of radiant energy. Such a line (e.g., the path of a single photon of visible light) is imagined to arise from a point source and have zero width. **2.** A thin beam of radiant energy (e.g., the beam of electrons in a cathode-ray tube). **3.** A quantity of radiant energy or ionizing radiation (e.g, *gamma ray*). **4.** One of numerous lines converging toward, or emanating from, a specific point. **5.** A vector representing the direction in which an electromagnetic field or acoustic disturbance travels. **6.** Also called *half line.* The set of points on a line consisting of a defined origin and all the points on one side of the origin. Example: the positive reactance axis in an ARGAND DIAGRAM.

Raydist A continuous-wave, medium-frequency radionavigation system. The position is determined according to the phase difference between two signals transmitted from different locations.

Rayleigh-Carson theorem An expression of the reciprocal relationship between the transmitting and receiving properties of an antenna. If voltage E applied to antenna A causes current I to flow at a given point in antenna B, then the same voltage (E) applied at that point in antenna B will produce identical current I (same magnitude and phase) at the point in antenna A, where voltage E originally was applied. Also see RECIPROCITY THEOREM.

Rayleigh distribution A probability-density function, used to describe the behavior of sky-wave electromagnetic signals.

Rayleigh's law The hysteresis loss in a magnetic material varies in proportion to the cube of the magnetic induction.

Rb Symbol for RUBIDIUM.

R_B Symbol for BASE RESISTANCE. (Also, r_B.)

RC **1.** Abbreviation of RESISTANCE-CAPACITANCE. **2.** Abbreviation of RADIO-CONTROLLED. **3.** Abbreviation of REMOTE CONTROL.

R_c **1.** Symbol for COLLECTOR RESISTANCE. (Also, r_C.) **2.** Symbol for COLD RESISTANCE.

RCA jack See PHONO JACK.

RCA plug See PHONO PLUG.

RC circuit See RESISTANCE-CAPACITANCE CIRCUIT.

RC-coupled amplifier See RESISTANCE-CAPACITANCE-COUPLED AMPLIFIER.

RC coupling See RESISTANCE-CAPACITANCE COUPLING.

RC filter See RESISTANCE-CAPACITANCE FILTER.

RCL **1.** Abbreviation of RECALL. **2.** Abbreviation of RESISTANCE-CAPACITANCE-INDUCTANCE.

RCM Abbreviation of RADAR COUNTERMEASURES. (Also, radCM.)

RC phase shifter See RESISTANCE-CAPACITANCE PHASE SHIFTER.

RC time constant See RESISTANCE-CAPACITANCE TIME CONSTANT.

RCTL Abbreviation of RESISTOR-CAPACITOR-TRANSISTOR LOGIC.

RC tuning See RESISTANCE-CAPACITANCE TUNING.

RCV Abbreviation for *receive.* (Also, rcv.)

RCVR Abbreviation for RECEIVER. (Also, rcvr, rx.)

rd Abbreviation for *rutherford.*

R & D See R AND D.

R_D Symbol for DRAIN RESISTANCE.

R_d **1.** Symbol for DIODE RESISTANCE. (Also, r_d.) **2.** Symbol for DISTRIBUTED RESISTANCE.

R-DAT Abbreviation of ROTARY DIGITAL AUDIO TAPE.

R_{dc} Symbol for DC RESISTANCE. (Also, r_{dc}.)

RDF Abbreviation of RADIO DIRECTION FINDER.

Re Symbol for RHENIUM.

R_e Symbol for EMITTER RESISTANCE. (Also, r_E.)

REA Abbreviation of *Rural Electrification Administration.*

reachthrough See PUNCHTHROUGH.

reachthrough region See PUNCHTHROUGH REGION.

reachthrough voltage See PUNCHTHROUGH VOLTAGE.

reactance Symbol, X. Unit, ohm. The opposition offered to the flow of alternating current by pure capacitance, pure inductance, or a combination of the two. Reactance introduces phase shift. Also see CAPACITIVE REACTANCE and INDUCTIVE REACTANCE. Compare RESISTANCE.

reactance chart A nomograph for capacitance, inductance, and frequency.

reactance factor The ratio of the alternating-current resistance of a conductor to the direct-current resistance. The reactance factor generally increases as the frequency increases because of skin effect and because the length of the conductor might be a sizable part of the wavelength of the transmitted energy.

reactance modulator A frequency modulator using a variable reactance, usually a varactor diode in the oscillator.

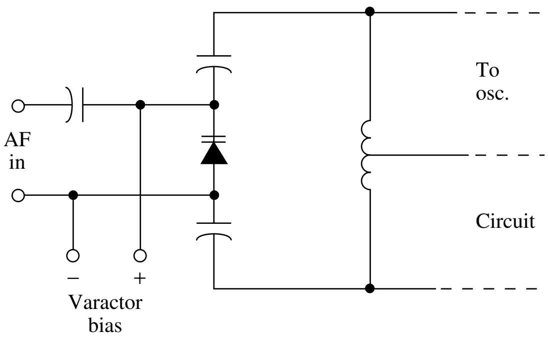

reactance modulator

reactance transistor A transistor used as a REACTANCE MODULATOR.

reaction-time meter See NEOMATACHOGRAPH and NEOMATACHOMETER.

reactive absorber In acoustics, a device that dissipates impinging sound waves by means of reflection, resonance, and other effects, besides dissipation in the form of heat.

reactive attenuator An attenuator that functions by means of reactance, rather than by means of resistance.

reactive current The component of alternating current that is not in phase with the voltage. Compare RESISTIVE CURRENT.

reactive kilovolt-ampere Abbreviation, kVAR. A unit of high apparent power; it is the product of kilovolts and amperes in a reactive component of a circuit. Also see APPARENT POWER, KILOVOLT-AMPERE, REACTIVE VOLT-AMPERE, and VOLT-AMPERE.

reactive load **1.** A load device that is capacitive or inductive, rather than resistive. **2.** A load device that contains reactance as well as resistance.

reactive power See REACTIVE KILOVOLTAMPERE and REACTIVE VOLT-AMPERE.

reactive volt-ampere Abbreviation, VAR. A unit of apparent power; it is the product of volts and amperes in a reactive component of a circuit. Also see APPARENT POWER, KILOVOLT-AMPERE, REACTIVE KILOVOLT-AMPERE, and VOLT-AMPERE.

reactor **1.** An inductor, especially one having very low internal resistance, used principally for its inductive reactance. **2.** A chamber in which the nuclei of atoms are split to provide atomic energy. Also see NUCLEAR REACTOR. **3.** In industrial chemistry, a vat in which reactions take place.

read **1.** In computer operations, to extract data from memory or a storage medium and (usually) transfer it to another area of memory or other medium. Compare WRITE. **2.** In digital communications, to transcribe data into printed form. **3.** In radiotelegraphy, to listen to Morse-code signals and comprehend the text without necessarily writing it down. **4.** To observe and note the indication of an instrument, such as a meter.

readability In electronic communications, the degree to which a desired signal can be recognized and interpreted in a given context.

readback In a multiplexer, a feature that facilitates inspection of the contents of the control latch.

reader A device that transcribes digital signals or markings into meaningful data. Examples: *Morse-code reader* and *bar-code reader.*

read head In a magnetic memory or in a tape recorder or wire recorder used for data recording, the head that picks up the magnetic pulses from the drum, tape, disk, or wire. Compare WRITE HEAD.

reading rate The number of input characters per second that a computer or other data-processing device handles.

read-only memory Abbreviation, ROM. In a computer or calculator, a memory unit in which instructions or data are permanently stored for use by the machine or for reference by the user. The stored information is read out nondestructively.

readout lamp An electron tube containing several cathodes, filled with a gas (such as neon), and used as a numeric or alphanumeric display device. Each cathode is connected to a separate pin on the base. A single anode is common to all cathodes. The cathode(s) to which a voltage is applied glow(s), showing the shape of a numeral, letter of the alphabet, or other symbol. In recent years, this type of display has been replaced by light-emitting diodes (LEDs) and liquid-crystal displays (LCDs).

readout pulse In a random-access memory (RAM), a pulse applied to the word line, facilitating readout of the information in a certain storage slot.

read pulse In computer operations, a pulse that activates the read function (see READ). Compare WRITE PULSE.

read rate The number of data units an input read device can transcribe per unit of time [e.g., *bits per second (bps)* and *words per minute (wpm)*].

readthrough **1.** The reception of signals between transmitted pulses at the same frequency. **2.** The continuous monitoring of a signal being jammed. Any change in the frequency, modulation, or other characteristics of the signal can then be detected, and the jamming signal adjusted accordingly.

read time The period during which data is being transferred from a computer storage unit.

read-write channel In computer operations, a channel over which activity between a central processing unit and a specific peripheral occurs.

read-write head An electromagnetic transducer used for both reading and writing data. See READ and WRITE.

read-write memory **1.** A small data storage bank for short-term use. The contents of the memory are easily changed. **2.** See RANDOM ACCESS MEMORY.

real address See ABSOLUTE ADDRESS.

real axis The axis of the real-number component of a COMPLEX NUMBER (i.e., the horizontal axis in an ARGAND DIAGRAM).

real component The real-number part of a COMPLEX NUMBER.

real image The image formed on a screen when rays from the object converge on passing through a lens. Compare VIRTUAL IMAGE.

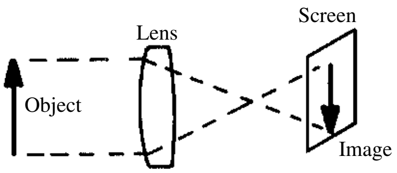

real image

real number A number in the category that includes zero, all rational numbers, and all irrational numbers. Also see COMPLEX NUMBER, IMAGINARY NUMBER, IRRATIONAL NUMBER, and RATIONAL NUMBER.

real power The apparent power multiplied by the power factor in an alternating-current circuit containing reactance. Real power is the difference between the apparent power and the reactive power. Actual radiated or dissipated power cannot exceed the real power.

real time Pertaining to the operation of a computer, communications, or data processing system in which events are represented or acted on as they occur. Data is processed as it becomes available, usually through the use of time-sharing, direct-access storage devices, and remote terminals.

real-time clock A device that produces periodic signals that reflect the interval between events. It is sometimes used to indicate time of day.

rear end The low-frequency portion of a superheterodyne receiver (i.e., the intermediate-frequency amplifier, second detector, and audio-frequency amplifier). Compare FRONT END.

rear projection A method of big-screen television picture reproduction. The image is focused onto a translucent, flat surface. The viewer is positioned on the side of the screen opposite to the projecting beam.

Reaumur scale A thermometer scale on which zero is the freezing point of water and 80 degrees is the boiling point of water. Compare ABSOLUTE SCALE, CELSIUS SCALE, FAHRENHEIT SCALE, and RANKINE SCALE.

rebecca The airborne interrogator in the British REBECCA-EUREKA SYSTEM.

rebecca-eureka system A British 90-mile-hovering radar navigation system that consists of an airborne interrogator (rebecca) and a ground transponder beacon (eureka).

rebroadcast The retransmission of a radio broadcast simultaneously by a station other than the originator. Also see AUTOMATIC RELAY STATION.

rebroadcast station See AUTOMATIC RELAY STATION.

recalescence During the cooling of a metal, the sudden release of heat. Also see RECALESCENT POINT. Compare DECALESCENCE.

recalescent point In a metal whose temperature is being lowered from a higher value, the temperature at which heat is suddenly released. Compare DECALESCENT POINT.

recall Abbreviation, RCL. In computers and calculators, an instruction that brings material from the memory for examination or use. The opposite instruction is STORE.

receiver **1.** A device or system operated at the destination end of a communication link; it accepts a signal and processes or converts it for local use. Also see specific entries for various types of receiver. **2.** The earpiece of a telephone. **3.** A radio broadcast-band tuner integrated with a general-purpose preamplifier and power amplifier, and containing standard jacks for input and output of audio signals to and from peripheral equipment.

receiver detector In a wireless communications or broadcast receiver, a circuit that extracts the information from the signal. The design depends on the type of emission to be received. See also DISCRIMINATOR, ENVELOPE DETECTOR, PHASE-LOCKED LOOP, PRODUCT DETECTOR, RATIO DETECTOR.

receiver dynamic range A quantitative measure of the ability of a wireless receiver to maintain a fairly constant output, and yet to maintain its rated sensitivity, in the presence of signals ranging from very weak to extremely strong. This figure is specified in decibels. It is typically 100 dB or more in a well-engineered communications receiver.

receiver front end The portion of a wireless communications or broadcast receiver consisting of the first radio-frequency (RF) amplifier, and often also including bandpass filters between this amplifier and the antenna. The dynamic range and sensitivity of a receiver are determined by the performance of this stage. These two characteristics are among the most important for any receiver.

Low-noise, high-gain amplifiers are the rule. Field-effect transistors are commonly used.

receiver IF chain In a wireless communications or broadcast receiver, the series of radio-frequency (RF) amplifier stages in which most of the gain takes place. These stages are also where the best possible RF selectivity is obtained. The intermediate frequency (IF) on which the amplifiers works is a constant frequency. This simplifies the design of the amplifiers to produce optimum gain and selectivity. Crystal-lattice filters or mechanical filters are commonly used in these stages to obtain the desired bandwidth and response.

receiver mixer In a superheterodyne wireless communications or broadcast receiver, a stage that converts the variable input signal frequency to a constant intermediate frequency (IF), or a stage that converts the first IF to a second IF usually having a lower frequency. This type of circuit is nonlinear, and combines the signal with a carrier from a local oscillator (LO). The output is either the sum or the difference of the signal frequency and the LO frequency.

receiver muting See MUTING, **1**.

receiver noise figure A quantitative measure of the ability of a wireless receiver to respond to desired signals while rejecting unwanted noise. This can be quantified in various ways. In general, the lower the noise figure, the better the sensitivity. Gallium–arsenide field-effect transistors (GaAsFETs) are well known for the low levels of noise they generate, even at quite high frequencies. Other types of FETs can be used at lower frequencies. Bipolar transistors tend to be noisy. See also NOISE FACTOR, NOISE FIGURE.

receiver post-detector stages In a wireless communications or broadcast receiver, one or more stages of amplification and/or filtering employed to refine the detector output to a form suitable for feeding a speaker, headset, printer, fax machine, slow-scan television picture tube, computer, electromechanical device, or other peripheral equipment.

receiver primaries See DISPLAY PRIMARIES.

receiver selectivity A quantitative measure of the ability of a wireless receiver to respond to a desired signal, but not to undesired ones. The frequency window is established by a preselector in the early RF amplification stages, and is honed to precision by bandpass filters in later amplifier stages. The preselector passes energy within a range of about plus-or-minus 10 percent of the signal frequency; other frequencies are attenuated. This reduces the chance for strong, out-of-band signals to impair the performance of the receiver. The narrowband filter in the final intermediate-frequency (IF) stage responds only to energy within the actual signal band. This minimizes adjacent-channel interference. In some receivers, yet another bandpass filter is used in the audio-amplifier stages.

receiver sensitivity A quantitative measure of the ability of a wireless receiver to recover weak signals and process them into readable data. The most common expression is the number of signal microvolts that must exist at the antenna terminals to produce a certain signal-to-noise ratio (S/N). Sometimes, the signal-plus-noise-to-noise ratio (abbreviated S+N/N) is given. The front end, or first RF amplifier stage, of a receiver is the most important stage with regard to sensitivity. Sensitivity is directly related to the gain of this stage, but the amount of noise the stage generates is even more significant. A good front end should produce the best possible S/N or S+N/N ratio at its output. All subsequent stages amplify the front-end noise output as well as the front-end signal output.

receiving set RADIO RECEIVER.

receiving station A station that ordinarily only receives signals (i.e., it makes no type of transmission). Compare TRANSMITTING STATION.

receptacle **1.** See SOCKET. **2.** The half of a connector that is mounted on a support, such as a panel, and that is therefore stationary.

recharge In certain cells and batteries, the restoration of chemical energy following use so that the device is ready to deliver its full rated electric current. Also see RECHARGEABLE.

rechargeable Pertaining to a secondary cell or battery that can accept a restoration of chemical energy following use, and thus can be completely charged and discharged numerous times. Examples: *nickel-metal-hydride (NiMH) battery* and *lead-acid battery.*

reciprocal impedances See INVERSE IMPEDANCES.

reciprocal ohm See SIEMENS and MHO.

reciprocation **1.** The determination of a mathematical reciprocal value from a given value. **2.** The transmission of a message in response to a received message.

reciprocity in antennas See RAYLEIGH-CARSON THEOREM.

reciprocity theorem When a voltage E across branch A of a network causes a current I to flow in branch B of the network, the voltage can be applied across branch B to cause the same value of current to flow in branch A. Compare COMPENSATION THEOREM, MAXIMUM POWER TRANSFER THEOREM, NORTON'S THEOREM, SUPERPOSITION THEOREM, and THEVENIN'S THEOREM.

recombination The refilling of holes by electrons in a semiconductor.

recombination current In a transistor circuit, base current resulting from recombination.

recombination rate In a semiconductor material, the speed at which the electrons and holes recombine. It can be expressed as the time required for a certain proportion of charge carriers to recombine.

recompile In computer operations, to COMPILE again, usually according to program amendments following debugging, or to create a different form of a program so that it will be compatible with other hardware.

record **1.** See PHONOGRAPH DISC. **2.** A chart delivered by a graphic recorder. **3.** To make one of the foregoing. **4.** In data processing, a constituent of a file. **5.** In data processing, a data unit portraying a specific transaction.

record blocking In data processing operations, making data blocks from groups of records so that the blocks can, in a single operation, be transferred to a nonvolatile storage medium, such as diskette or tape.

record count A usually running total of a file's records.

recorded disc A phonograph disc on which a recording has been made. Also called PRERECORDED DISC.

recorded tape Magnetic tape containing recorded material. Also called PRERECORDED TAPE. Compare BLANK TAPE.

recorder **1.** A machine for preserving sound, video, or data signals in the sequence in which they occur (e.g., DISC RECORDER, TAPE RECORDER, and WIRE RECORDER). **2.** A machine for making a permanent visual record (photographically or by stylus) of an electrical phenomenon. Examples: DRUM RECORDER and OSCILLOGRAPH.

record head See RECORDING HEAD.

recording density In a magnetic storage medium, the number of information units (bits, bytes, etc.) represented by magnetized areas, per unit area or length.

recording disc A phonograph record on which material has not been recorded, or from which recorded material has been removed. Compare PRERECORDED DISC.

recording head In a magnetic recorder/reproducer, the head that magnetizes the medium in accordance with sounds or other signals. Also called RECORD HEAD and WRITE HEAD. Compare PLAYBACK HEAD.

Recording Industry Association of America Abbreviation, RIAA. An organization that sets standards for audio recording and reproduction in the United States.

recording instrument A measuring instrument, such as a voltmeter or ammeter, that makes a permanent record of its deflections. Also see RECORDER, **2**.

recording loss **1.** Loss of data during a recording process. **2.** Loss resulting from recording efficiency of less than 100 percent; audio power loss.

recording tape Magnetic tape on which nothing has been recorded, or from which all data has been erased. Compare PRERECORDED TAPE.

recovery time **1.** Symbol, t_r. The time required for a semiconductor pn junction to attain its high-resistance state when the bias voltage is suddenly switched from forward to reverse. **2.** The time required for a circuit to recover from momentary overdrive. **3.** The time required for a computer system to stabilize following a degenerative operation. **4.** The time required for switching a memory from the write to the read mode. It is measured as the length of time from switching out of the write mode until meaningful signals occur at the output. **5.** In a transceiver, the time required from the completion of a transmitted signal until the receiver is activated.

rect Abbreviation of RECTIFIER.

rectangular coordinates See CARTESIAN COORDINATES.

rectangular scan **1.** A method of beam scanning in a cathode-ray tube, in which the beam moves sequentially in parallel lines to cover a rectangular region. Used in television. **2.** In radar, a two-dimensional scan, covering a specific rectangular region.

rectangular wave An alternating or pulsating current or voltage whose rise and decay times are essentially zero, and whose maxima and minima are essentially flat, but not necessarily of equal duration. The SQUARE WAVE is a special type of rectangular wave.

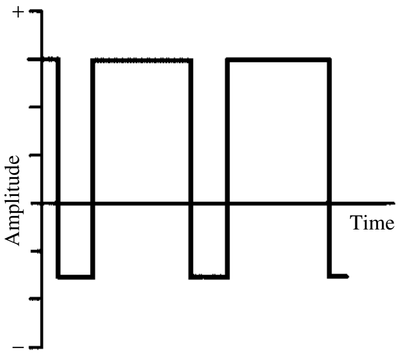

rectangular wave

rectangular waveguide A waveguide having a rectangular cross section.

rectification The conversion of alternating current into pulsating direct current by any means other than the use of a motor-generator. Also see RECTIFIER.

rectification efficiency The ratio (expressed as a percentage) of the direct-current output voltage to the peak alternating-current input voltage of a rectifier.

rectified alternating current The unfiltered, pulsating direct-current output of a rectifier. It consists of the unidirectional half-cycles passed by the rectifier (one per cycle for half-wave rectification, and two per cycle for full-wave rectification).

rectifier Abbreviation, rect. An electronic or electromechanical device that converts alternating current into pulsating direct current.

rectifier diode A heavy-duty tube or semiconductor diode designed primarily to change alternating current to pulsating direct current in power supplies.

rectifier filter A circuit containing parallel capacitance, sometimes in combination with series inductance, intended for smoothing out the ripple in the output of a power-supply rectifier.

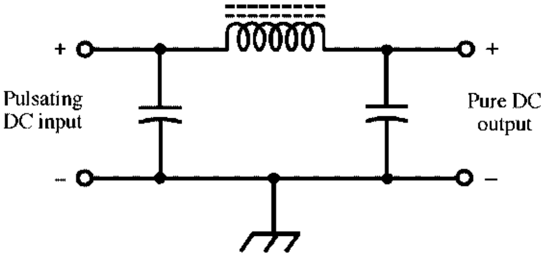

rectifier filter

rectifier-filter system The rectifier plus power-supply-filter combination for converting alternating current into direct current.

rectifier photocell A photovoltaic cell consisting of two layers of material with a semiconductor junction between them. The device produces direct current when exposed to visible light, infrared, or ultraviolet radiation.

rectifier probe A diode-type probe used with a direct-current (dc) voltmeter to measure radio-frequency (RF) voltage. The diode rectifies the RF signal and presents to the meter a dc voltage proportional to the peak RF voltage.

rectifier stack An assembly of separate rectifier disks or plates in series on a central bolt, as in most selenium rectifiers.

rectifier tube A two-element electron tube, once commonly used for converting alternating current into pulsating direct current in high-voltage, high-current power supplies.

rectifier-type meter See DIODE-TYPE METER.

rectilinear chart A graphic-recorder chart in which the crossing coordinates are arcs, rather than straight lines, to correspond to the swing of the pen. Also see STRIP CHART.

rectilinear scan See RECTANGULAR SCAN, **1**.

recurrent network A circuit in which several sections of identical configuration (e.g., L sections) are cascaded.

recurrent phenomenon A phenomenon that repeats itself periodically.

recurrent sweep In an oscilloscope, a repetitive horizontal sweep of the beam occurring at a frequency determined by the settings of the sweep controls. Also called REPETITIVE SWEEP. Compare NONREPETITIVE SWEEP.

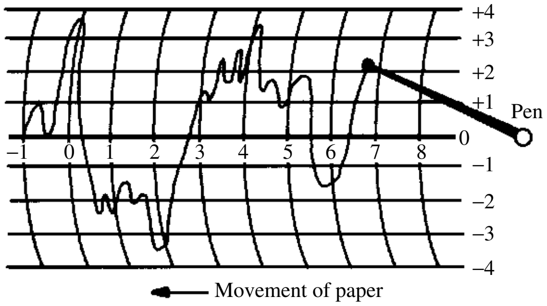

rectilinear chart

recursion **1.** Generating a complete sequence of functions or numbers by applying an algorithm to initial values in the sequence. **2.** In computer programming and artificial intelligence, a logical process containing loops in calculation or reasoning.

recursive Relating to a procedure or set of steps that repeat endlessly.

Red Book The first format developed for compact-disc data storage media, developed by Sony and Philips. It is commonly used in digital audio systems. See also CD-ROM, GREEN BOOK, ORANGE BOOK, and YELLOW BOOK.

red-green-blue Abbreviation, RGB. In video applications, the three primary colors from which all other colors are derived. Also see COLOR TELEVISION.

red gun In a three-gun color-television picture tube, the electron gun whose (correctly adjusted) beam strikes only the red phosphor dots on the screen.

red oxide of iron An iron oxide of the general formula Fe_2O_3, used as the magnetic coating of recording tape. Also see IRON OXIDE.

red oxide of zinc See ZINCITE.

red-tape operation An operation or function needed for organizational purposes, but that does not directly contribute to the completion of the task at hand.

reduced instruction set computer Abbreviation, RISC. A computer architecture in which program instructions are simplified to obtain enhanced processing speed. It is useful especially in complex graphics, animation, multimedia, and scientific work requiring many calculations.

reductio ad absurdum A method of obtaining a conclusion by proving that its negation results in a contradiction. It is sometimes used in computer programming involving mathematical proofs.

reduction In an electrochemical cell or battery, a transfer of electrons to the active chemical.

reductionism The theory that all human thought processes, including emotion and intuition, can be reduced to digital logic, and thus can be duplicated by a sufficiently powerful computer. It is of interest to researchers in artificial intelligence.

reductionist A person who subscribes to the theory of REDUCTIONISM.

redundancy **1.** The repetition of components in a circuit (e.g., series or parallel connection of them) so that one will be available for circuit operation if the other fails. **2.** Having available more than one method for performing a function. **3.** Having on hand several copies of data as a safeguard against data loss.

redundancy check A check for the integrity of digitized data to which extra bits have been added for the purpose (e.g., *parity check*).

redundant **1.** Pertaining to any two units of data that resemble each other in such a manner that if either unit is removed, no information is lost from the system. **2.** A unit of data that contains information already present in the system.

redundant array of independent disks Acronym, RAID. A set of data storage media used to store video programs.

red video voltage In a three-gun color-television circuit, the red-signal voltage that actuates the red gun.

reed A usually thin metal blade, leaf, or strip used in vibrators, reed-type relays, reed-type oscillators, and similar devices.

reed oscillator See REED-TYPE OSCILLATOR.

reed relay See DRY-REED SWITCH and MERCURY-WETTED REED RELAY.

reed-relay logic Logic circuits using reed relays. Also see RELAY LOGIC.

reed switch **1.** A frequency-sensitive switch in which the movable contact is mounted on the tip of a thin, metal strip (reed). The reed is actuated by an alternating-current (ac) coil. The reed closes the contacts when the ac excitation is at its natural frequency. **2.** See DRY-REED SWITCH.

reed-type oscillator An electromechanical audio-frequency oscillator whose frequency is controlled by a vibrating metal strip (reed) instead of a tuning fork. Also see HUMMER.

reed-type switch See REED SWITCH.

reel **1.** The spool around which a magnetic tape or video film is wound. **2.** A spool containing magnetic tape or video film.

reentrant cavity A resonant cavity in which one or more sections are directed inward to confine the electric field to a small volume.

reentrant winding A winding of wire that returns to its starting point—especially in a motor armature.

ref Abbreviation of REFERENCE.

reference address As a point of reference, an address for instructions having relative addresses.

reference amplifier A voltage-regulation device consisting of a transistor and Zener diode in the same envelope.

reference angle In radar, the angle of incidence of the beam against a target surface, measured with respect to the normal (perpendicular) line at the surface.

reference antenna A standard antenna, such as an isotropic radiator or a half-wave dipole, used to establish a reference for determining the relative gain of another antenna.

reference bias current In a reference amplifier, the input current that subtracts from the reference current. It is generally measured in microamperes.

reference current range In a digital-to-analog converter, the difference between the maximum and minimum reference current for which the device is within specifications for resolution.

reference diode A Zener diode whose constant voltage drop is used as a direct-current reference potential in calibrator circuits and voltage regulators.

reference dipole See REFERENCE ANTENNA.

reference electrode For use with a pH meter, an electrode that provides a reference potential.

reference input slew rate In a digital-to-analog converter, the average rate of change in output for a given change in the reference input. It is expressed in milliamperes or microamperes per microsecond.

reference level A specific value of a quantity (e.g., current, frequency, power, or voltage) to which other values of the same quantity are referred.

reference time The point at which a trigger pulse attains 10 percent of its maximum amplitude.

reference tone A standard audible tone of known frequency [e.g., 440 Hz (representing A below middle C)]. Sometimes the intensity as well, as the frequency, is specified.

reference white level The television picture signal value representing the uppermost limit for peak white signals.

R_{eff} Symbol for EFFECTIVE RESISTANCE.

reflectance **1.** See MISMATCH FACTOR. **2.** The reflected part of the radiant flux striking a surface. It is expressed as a fraction of the total incident radiation.

reflected binary code See CYCLIC CODE.

reflected electromagnetic field In a transmission line, the electromagnetic energy not absorbed by the load when an impedance mismatch exists between the load and the line. See INCIDENT POWER and REFLECTED POWER.

reflected impedance In a coupled circuit, the impedance in the secondary that appears in the primary circuit, or vice-versa, as if it were reflected through the coupling transformer.

reflected power In a transmission line not perfectly matched to a load at the feed point, an expression of the amount of electromagnetic field reflected from the feed point, rather than absorbed by the load. In general, this can be expressed in watts or as a percentage of the incident power. It is not a true indicator of the loss caused by the mismatch because the reflected field is usually all returned again when it arrives back at the transmitter.

reflected-power meter A radio-frequency instrument, connected between a source and a load, that can measure INCIDENT POWER and REFLECTED POWER.

reflected ray The ray that is reflected by the surface of a body or region it strikes. Compare INCIDENT RAY and REFRACTED RAY.

reflected resistance **1.** In a transformer, the effective resistance across the primary winding when a resistive load is connected to the secondary. **2.** In a transmission line, the resistance at the input end when a load is connected to the output end.

reflected wave **1.** An electromagnetic wave reflected by the ionosphere or by the surface of the earth. Compare INCIDENT WAVE and REFRACTED WAVE. Also see IONOSPHERE and IONOSPHERIC PROPAGATION. **2.** A wave that is bounced off an obstruction, such as a building or mountain.

reflecting galvanometer A galvanometer having a light-beam pointer.

reflecting shell See IONOSPHERE.

reflection **1.** The turning back of a ray by a surface it strikes. Examples of reflecting media are the surface of the earth, the polished surface of a material, and a layer of the ionosphere. Compare REFRACTION. **2.** The return of energy to the source by the mismatched end of a transmission line or by the end of a radiator.

reflection error In a radar, radionavigation, or radiolocation system, an error in the reading caused by reflections of the signal from objects other than the intended signal source or object.

reflection factor See MISMATCH FACTOR.

reflection law When a ray strikes a smooth reflecting surface, the angle of incidence is equal to the angle of reflection.

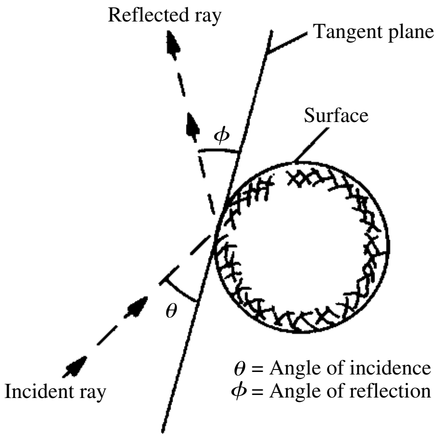

reflection law

reflection loss **1.** Loss caused by the reflection of an electromagnetic field at a discontinuity in a transmission line. **2.** Loss that occurs when an electromagnetic wave is reflected from a surface or object.

reflection phase grating A device that diffuses sound waves by diffraction effects. The acoustic equivalent of an electromagnetic-wave DIFFRACTION GRATING.

reflective code See GRAY CODE.

reflectivity **1.** See MISMATCH FACTOR. **2.** The degree to which a point, plane, or surface reflects the radiation (light, for example) that strikes it.

reflectometer **1.** See REFLECTED-POWER METER. **2.** A type of photometer used to measure reflection.

reflector **1.** A smooth, metal surface or wire screen for reflecting radio waves. See, for example, PARABOLIC REFLECTOR. **2.** A length of wire, rod, or tubing used in a parasitic antenna to reflect radio waves. Compare DIRECTOR and RADIATOR. **3.** A polished surface for reflecting visible light or infrared rays (i.e., a mirror). **4.** See REPELLER.

reflector element See REFLECTOR, **2**.

reflector satellite A satellite whose skin reflects radio waves.

reflector voltage In a reflex Klystron, the reflector-to-cathode voltage.

reflex baffle A loudspeaker BAFFLE constructed so that some of the sound radiated to the rear of the diaphragm is transmitted forward (after phase shift) to boost acoustic radiation at some frequencies.

reflex bunching In a Klystron, electron bunching following direct-current-field-induced reversal of the velocity-modulated electrons. Also see REFLEX KLYSTRON.

reflex circuit A radio receiver circuit in which a single transistor is used successively for different functions. For example, one active device can act as a mixer and as a radio-frequency amplifier.

reflex Klystron A Klystron having only one cavity. This cavity serves first as the buncher and then, as the electrons are turned around and caused to pass through again, as the catcher.

refracted ray The ray that is refracted by a body or region through which it passes. Compare INCIDENT RAY and REFLECTED RAY.

refracted wave An electromagnetic wave that is refracted by the ionosphere. Compare INCIDENT WAVE and REFLECTED WAVE. Also see IONOSPHERE and IONOSPHERIC PROPAGATION.

refraction The bending of an energy ray as it passes through media that cause a change in the speed of propagation. It can occur with radio waves, infrared, visible light, ultraviolet, X rays, gamma rays, and sound waves.

refractive index See INDEX OF REFRACTION.

refractivity The extent of the ability to refract, given as the quantity $(v_1/v_2) - 1$, where v_1 is the phase velocity in free space, and v_2 is the phase velocity in the medium through which a wave passes.

refractory A heat-resistant, nonmetallic ceramic material.

refrigerator A chamber used to maintain a circuit or component at a constant temperature that is lower than the ambient temperature. This device is analogous to the oven, which maintains a higher temperature than the surrounding medium. A refrigerator can be used to maintain precise frequency for a reference oscillator.

regeneration **1.** The processing of a distorted signal so that it has its original characteristics. **2.** Positive feedback generally used for the purpose of causing oscillation, or for detection in a regenerative receiver. See POSITIVE FEEDBACK.

regeneration period The period during which the electron beam scans a cathode-ray tube screen to restore changes to the screen surface.

regenerative amplifier An amplifier that uses regeneration to increase its gain and/or selectivity.

regenerative detector A detector provided with regenerative feedback. Although such a detector is sensitive, it can be unstable. Compare NONREGENERATIVE DETECTOR.

regenerative feedback Feedback producing regeneration (i.e., positive feedback). Compare DEGENERATIVE FEEDBACK.

regenerative IF amplifier An intermediate-frequency amplifier in which regeneration is introduced to boost sensitivity and, sometimes, selectivity.

regenerative reading A method of reading data (see READ) so that it is automatically restored, by writing, to locations from which it came.

register In computer systems, an arrangement of several storage devices, such as flip-flops, for storing a certain number of digits (a two-bit register, for example, requires two flip-flops).

register capacity The range of values for quantities that can be handled by a register.

registered professional engineer A title granted by a state board of examiners to a person licensed to work as an engineer.

register length The number of characters or bits that can be held in a register, according to its capacity.

registration The accurate alignment of terminals or other points on different components or on opposite sides of a board so that when the surfaces containing those points are overlaid, all points mate precisely.

regulated power supply A power supply whose output is held automatically to a constant level or within a narrow range, regardless of loading variations.

regulating transformer See VOLTAGE-REGULATING TRANSFORMER.

regulation **1.** In general, the adjustment or control of a component, device, or system. **2.** Automatic control. See, for example, SELF-REGULATION. **3.** See CURRENT REGULATION. **4.** See VOLTAGE REGULATION.

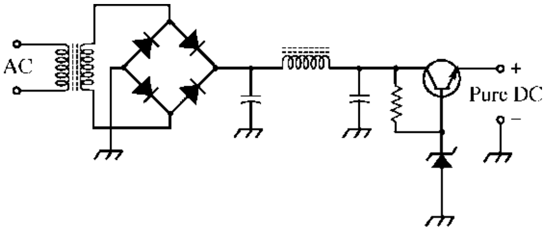

regulated power supply

regulator **1.** A device that automatically holds a quantity to a constant value (e.g., a voltage regulator). **2.** A device via which a quantity can be varied (e.g., *potentiometer*, *rheostat*, and *variable autotransformer*).

regulator diode A semiconductor diode—especially a Zener diode used as a two-terminal voltage regulator.

reinitialization The setting of all lines in a microcomputer or microprocessor to logic zero automatically when power is removed, then reapplied.

reject amplifier A tuned amplifier having the response of a band-suppression filter. Like the filter, the amplifier rejects or severely attenuates one frequency (or band of frequencies) while readily passing lower and higher frequencies. Compare PASS AMPLIFIER.

reject filter See REJECTION FILTER.

rejection circuit A circuit performing the function of a REJECTION FILTER.

rejection filter A filter that suppresses one frequency (or band of frequencies) while passing all other frequencies.

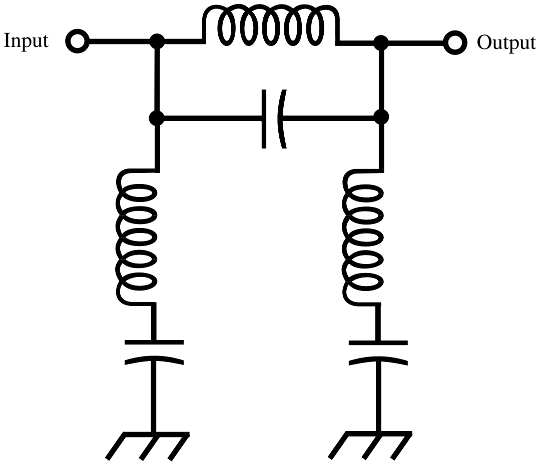

rejection filter

rejection notch A sharp dip in the transmission characteristic of a crystal filter. It provides rejection-filter action at the notch frequency. Also see CRYSTAL RESONATOR and REJECTION FILTER.

rejectivity The degree to which a selective circuit rejects an unwanted signal. Compare TRANSMITTIVITY.

rejuvenation See REACTIVATION.

rel Symbol, R. The cgs unit of reluctance, equivalent to gilberts per maxwell.

relative accuracy In a measuring instrument, the error determined as a percentage of the actual value; the difference between the actual and measured values, divided by the actual value, then multiplied by 100.

relative address In the address part of a computer program instruction, a number specifying a location relative to a BASE ADDRESS. When the base address is added to the relative address, it yields the ABSOLUTE ADDRESS.

relative error The ratio of the absolute error to the exact value of a quantity.

relative gain The current, voltage, or power gain, measured, with respect to a reference standard.

relative humidity Abbreviation, rh. The ratio, as a percentage, of the amount of moisture in the air to the amount the air could contain at a given temperature. Compare ABSOLUTE HUMIDITY.

relative luminosity Luminosity measured with respect to a reference level.

relative permeability The ratio of the permeability of a given material to the permeability of another material (or of the same material under different conditions).

relative power Power level specified with respect to another (often reference) power level.

relative uncertainty The uncertainty of a measurement divided by the measured value. The maximum value that this quotient can have is 1. Also see UNCERTAINTY IN MEASUREMENT.

relative visibility Response of the human eye to light. This is relative because the eye does not see equally well throughout the visible spectrum. The peak response of the human eye is around 5.4×10^{14} Hz; this represents yellow-green light. Photoelectric devices have peak responses that can differ considerably from this value.

relativity theory See EINSTEIN'S THEORY.

relaxation A delayed change in circuit conditions, as a result of change in the input.

relaxation inverter An inverter circuit in which the direct-to-alternating-current conversion device is a RELAXATION OSCILLATOR.

relaxation oscillator An oscillator whose operation results from the buildup of a charge in a capacitor, followed by sudden discharge of the capacitor, the sequence being repeated periodically. In one circuit, a capacitor is connected in series with a resistor and a direct-current power supply, and a neon bulb is connected in parallel with the capacitor. The output is a sawtooth wave.

relaxation time **1.** The time required for an exponentially decreasing variable to fall to $1/e$ (approximately 36.8 percent) of its initial value, where e is the natural-logarithm base (approximately 2.71828). **2.** For a gas, the time required for it to return to its original state after having been disturbed.

relay **1.** A signal-actuated switching device. In most instances, a relatively weak current or voltage is used to make the relay switch a higher current or voltage. A relay can be electromechanical or fully electronic (no moving parts). See, for example, ELECTROMECHANICAL RELAY and ELECTRONIC RELAY. **2.** A repeater station. **3.** In communications, to receive a message and retransmit it en route from a source to a destination.

relay amplifier See RELAY DRIVER.

relay booster See RELAY DRIVER.

relay driver A direct-current amplifier (usually one stage) used to actuate an electromechanical relay in response to a low-powered signal.

relay flip-flop See BISTABLE RELAY.

relay logic Abbreviation, RL. In computer and industrial-control operations, a logic system using electromechanical relays as flip-flops (see BISTABLE RELAY).

relay transmitter See AUTOMATIC REPEATER STATION.

release time **1.** The interval between the instant power is removed from a relay and the instant the armature is released sufficiently to operate the contacts. **2.** The time between one control input becoming inactive and another becoming active. **3.** The time required for reception to resume in a transceiver, once transmission has stopped. **4.** An expression for the rapidity with which an automatic gain control reverts to maximum sensitivity following the reception of a strong signal.

reliability **1.** The dependability of operation of a device or circuit under specified conditions. **2.** The proportion of units that still work after a set of units has been in use for a specified length of time.

reliability engineering The branch of engineering devoted to the theory and application of reliability; based on fundamental engineering and advanced statistical concepts.

reluctance Symbol, R. SI unit, A/Wb; cgs unit, rel. In a magnetic circuit, the opposition to the establishment of a magnetic field; it is analogous to resistance in electric circuits.

reluctance motor An electric motor having a squirrel-cage rotor with some of its teeth ground down, and a shaded-pole or split-phase type of stator that supplies a rotating magnetic field. When starting, this motor comes up to speed like an induction motor, but the protruding teeth of the rotor then follow the field in the manner of the poles of a hysteresis motor.

reluctivity Specific reluctance (i.e., the reluctance of a sample of magnetic material one centimeter long and one square centimeter in cross section). Reluctivity is the reciprocal of permeability.

rem Acronym for *roentgen equivalent man*, an amount of ionizing radiation having the same effect on the body as a one-roentgen dose of gamma or X radiation.

remagnetizer A magnetizer used principally to restore weakened permanent magnets.

remainder **1.** The result of subtracting one quantity (the subtrahend) from another (the minuend). Also called DIFFERENCE. **2.** In division, the numerical value left over after the integral part of the quotient has been determined; it becomes the fractional part when divided by the divisor. For example, in 25/3, the remainder is 1.

remanence See RESIDUAL MAGNETISM.

remanent flux density See REMANENCE.

remodulator Any device that changes the modulation of a signal from one form to another, such as from frequency modulation to amplitude modulation, without loss of intelligence.

remote alarm In security systems, an alarm that occurs at a location different from where an intrusion occurs (e.g., at the headquarters of a security company).

remote control Control of distant devices by mechanical means or by radio-frequency signals sent from a transmitter especially designed for the purpose; in the latter case, it is sometimes called *radio control.*

remote-control receiver The complete device that selects, amplifies, and demodulates or rectifies a radio signal for control of a circuit or mechanism at a distance from the transmitter of the control signal. Some receivers have self-contained antennas.

remote-control transmitter The complete device that generates radio-frequency power, adds to it the signals needed for remote control, and radiates the modified power.

remote-control system The complete set of hardware units and software programs facilitating the operation of a computer or robot from a distance.

remote data terminal In a computer system, a terminal connected to the central processor by a telephone line or radio link. It is used for the transfer of data without providing control of the system. Also called *remote data station.*

remote error sensing A method of regulation used in some power supplies. The voltage across the load, or the current through the load, is determined by remote control. The power-supply output is adjusted to compensate for losses in the system.

remote job entry In computer operations, the keying-in of input data at a site physically distant from the central processor.

remote tuning The electrical or radio tuning of a circuit or device from a distance.

rendering In three-dimensional computer animation, the software process in which all the aspects of the model are combined to obtain the presentation.

rep **1.** Acronym for *roentgen equivalent physical*, an amount of ionizing radiation that, upon absorption by body tissue, will develop the energy of a one-roentgen dose of gamma or X radiation. **2.** Colloquial abbreviation for *repetition*, as in *rep rate.* **3.** Colloquial abbreviation for *representative*, as in *service representative.*

repeatability The ability of an instrument, system, or method to give identical performance or results in successive instances.

repeater A receiver/transmitter device that retransmits a signal it receives from another source, often simultaneously. In this way, a signal can be transmitted on several frequencies, or the service area of the original station can be extended. Also see ONE-WAY REPEATER and TWO-WAY REPEATER.

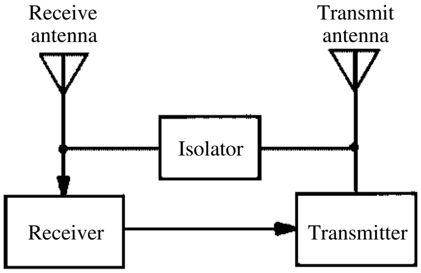

repeater

repeater station See AUTOMATIC REPEATER STATION and REPEATER.

repeating decimal A decimal fraction in which groups of digits recur endlessly (e.g., 25/99 = 0.252525. . .).

repeller An electrode, especially in a velocity-modulated tube, for reversing the direction of an electron beam.

repertoire The instruction set for a particular object or source computer programming language.

repetition instruction In a loop in a computer program, an instruction that causes the repetitive implementation of one or more instructions.

repetitive phenomenon See RECURRENT PHENOMENON.

repetitive sweep See RECURRENT SWEEP.

replication In an electronic system, especially in a computer, the redundancy of hardware units to provide standby facilities in case of failure.

replacement A component or circuit that can be substituted directly for another; it fits exactly into place and functions exactly like the component it replaces, without modification to the equipment.

report **1.** The results of testing and evaluation of a device, organized into a written document. **2.** The output of a computer, printed on paper for permanent reference.

report program generator Abbreviation, RPG. A computer programming language with which programs can be produced for the generation of business reports.

reproduce head See PLAYBACK HEAD.

reproducing stylus A stylus for the playback of material from a phonograph disc.

reproduction **1.** The recovery of data from storage, and its presentation in original form. **2.** Data obtained by the process defined in **1**. **3.** See PLAYBACK.

reproduction loss See PLAYBACK LOSS.

repulsion A force that pushes objects away from each other, as between similar electric charges or similar magnetic poles. Compare ATTRACTION. Also see LAW OF REPULSION.

repulsion-induction motor An alternating-current motor arranged to start as a REPULSION MOTOR and run as an INDUCTION MOTOR, but with better regulation than that of the latter.

repulsion motor An alternating-current motor having an armature and commutator similar to those of a direct-current motor, and a stator similar to that of a split-phase motor, without the auxiliary starting winding. Repulsion caused by the negative half-cycle of torque is utilized to drive the armature, by placing the brushes in such a way that they close the coils only when the latter are in position to receive this repulsive action.

repulsion-start motor An alternating-current motor that starts as a REPLUSION MOTOR but at approximately 75 percent of full speed. Its commutator is automatically short-circuited and the motor runs as an INDUCTION MOTOR. Also see REPULSION INDUCTION MOTOR.

request slip In computer operations, peripheral and memory needs for a program given in a written statement.

reradiation Radiation of energy by a body that has been exposed to radiation, as when a receiving antenna retransmits a signal.

rerecording A recording of played-back material.

reroute **1.** In computer operations, to establish new channels between peripherals and main memory. **2.** To establish new circuit paths, physically (as by changing conductor orientation) or electronically (as by selecting an alternate signal bus).

rerun See ROLLBACK.

res **1.** Abbreviation of RESISTANCE or RESISTOR. (Also, R and r.) **2.** Abbreviation of RESEARCH. **3.** Abbreviation of RESOLUTION.

reset **1.** The clearing of a flip-flop of data in storage (i.e., the setting of the flip-flop to its zero state). **2.** In a computer program, an instruction to initialize the value of a variable. **3.** In a security system, a function that terminates an alarm signal following an intrusion, and renders the system operational again so that it can detect subsequent intrusions should they occur.

reset action **1.** The return of a circuit or device to its normal operating condition. **2.** A method of adjusting a circuit to compensate for the severity of an abnormal condition. The extent of readjustment is determined by the extent of the departure from normal conditions.

reset generator A circuit or device that generates a pulse for resetting a flip-flop or counter. Also see RESET and RESET TERMINAL.

reset pulse A pulse that resets (see RESET, **1**) a storage cell in a computer memory.

reset terminal In a flip-flop, the zero-input terminal. Compare SET TERMINAL.

reset time The elapsed time between a malfunction and the completion of the reset action.

reset timer A device that returns a circuit or device to its initial state after a specified time delay.

reserve In multiple programming computer operations, to allocate memory areas and peripherals for a program.

reserve battery A battery in which the electrolyte is in a special standby chamber outside of the interelectrode section while the battery is on the shelf. When the battery is readied for service, the electrolyte is caused to flow into position between the electrodes, either by heating the battery, shocking it mechanically, or inverting it.

residual amplitude modulation See INCIDENTAL AM.

residual charge The electric charge remaining in a capacitor after it has been initially discharged. It results from dielectric absorption.

residual current A current that continues to flow in a circuit after removing power. The duration is measured in nanoseconds or microseconds.

residual frequency modulation **1.** See INCIDENTAL FM. **2.** Frequency modulation of the fundamental frequency of a Klystron by noise or alternating-current heater voltage.

residual gas Minute quantities of gas remaining in a vacuum tube after evacuation.

residual magnetism Magnetism remaining in a material, such as iron, after the magnetizing force has been removed.

residual modulation **1.** Modulation of a signal by hum or noise. **2.** See INCIDENTAL AM. **3.** See INCIDENTAL FM.

residual voltage In the output of a null device, such as a bridge, a usually small voltage still present at null and preventing zero balance.

residue check In computer operations, the verification of the result of an arithmetic operation using the remainders generated when each operand is divided by a special number; the remainder is transmitted along with the operand as a check digit.

resilience Also called *fault resilience.* The ability of an electronic device or system, especially a computer, to keep functioning after part of it has failed.

resin A natural or synthetic organic substance that is polymeric in structure and largely amorphous. Various plastics are made from synthetic resins.

resistance **1.** Symbol, R or r. Unit, ohm. In a device, component, or circuit, the simple opposition to current flow. Resistance by itself causes no phase shift. In a purely resistive circuit, $R = E/I$, where R is the resistance in ohms, E is the voltage in volts, and I is the current in amperes. **2.** A property of circuits, devices, or substances that causes impinging energy to be dissipated by conversion to heat. Compare REACTANCE.

resistance

Per kilometer of solid copper wire for American Wire Gauge (AWG) 1 through 40.

AWG	*ohms/km*	*AWG*	*ohms/km*
1	0.42	21	43
2	0.52	22	54
3	0.66	23	68
4	0.83	24	86
5	1.0	25	110
6	1.3	26	140
7	1.7	27	170
8	2.1	28	220
9	2.7	29	270
10	3.3	30	350
11	4.2	31	440
12	5.3	32	550
13	6.7	33	690
14	8.4	34	870
15	11	35	1100
16	13	36	1400
17	17	37	1700
18	21	38	2200
19	27	39	2800
20	34	40	3500

resistance alloys Metallic alloys used in the manufacture of resistance wire and resistance elements. Such alloys include CONSTANTAN, GERMAN SILVER, MANGANIN, MONEL METAL, and NICHROME.

resistance balance A device used to balance a circuit, by means of the insertion of resistances.

resistance brazing A method of brazing in which metal is heated by passing a current through it. The I^2R loss, or dissipated power, occurs in the form of heat.

resistance bridge A bridge (see BRIDGE, **2**) for measuring resistance only.

resistance-capacitance Abbreviation, *RC*. Pertaining to a combination of resistance and capacitance (e.g., RESISTANCE-CAPACITANCE CIRCUIT).

resistance-capacitance bridge **1.** A four-arm null circuit containing only resistors and capacitors. Also see BRIDGE, **1**. **2.** An alternating-current bridge (see BRIDGE, **2**) for measuring resistance and capacitance.

resistance-capacitance circuit A circuit containing only resistors and capacitors. There are no inductors.

resistance-capacitance-coupled amplifier A multistage amplifier circuit in which RESISTANCE-CAPACITANCE COUPLING is used between stages and at the input and output points of the circuit.

resistance-capacitance coupling Coupling, especially between stages in a circuit, using blocking capacitors and supply-path resistors.

resistance-capacitance filter A power-supply filter or wave filter containing only resistors and capacitors. The resistors are in the positions occupied by inductors in inductance-capacitance filters.

resistance-capacitance-inductance Abbreviation, *RCL*. Pertaining to a combination of resistance, capacitance, and inductance.

resistance-capacitance phase shifter A phase shifter containing only resistors and capacitors to obtain the desired shift.

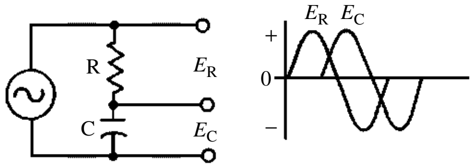

resistance-capacitance phase shifter

resistance-capacitance time constant Symbol, t. The time constant (see ELECTRICAL TIME CONSTANT) of a circuit containing (ideally) only resistance and capacitance; $t = RC$, where t is in seconds, R is in ohms, and C is in farads. Compare RESISTANCE-INDUCTANCE TIME CONSTANT.

resistance-capacitance tuning Tuning of a circuit, such as that of an amplifier or oscillator, by means of a variable resistor or ganged units of this type. See, for example, PARALLEL-TEE AMPLIFIER, PARALLEL-TEE OSCILLATOR, and WIEN-BRIDGE OSCILLATOR.

resistance-coupled amplifier See RESISTANCE-CAPACITANCE-COUPLED AMPLIFIER.

resistance drop The voltage drop across a resistor, or across the inherent resistance of a device.

resistance-inductance Abbreviation, *RL*. Pertaining to a combination of resistance and inductance (e.g., RESISTANCE-INDUCTANCE CIRCUIT).

resistance-inductance bridge **1.** A four-arm null circuit containing only resistors and inductors. Also see BRIDGE, **1**. **2.** An alternating-current bridge (see BRIDGE, **2**) for measuring resistance and inductance only.

resistance-inductance circuit A circuit containing only resistors and inductors. There are no capacitors.

resistance-inductance phase shifter A phase shifter containing only resistors and inductors to obtain the desired phase shift.

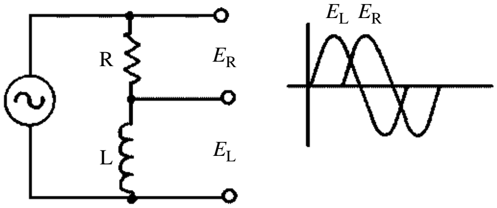

resistance-inductance phase shifter

resistance lamp An incandescent bulb inserted in series with a circuit to provide a dropping resistance. Such a lamp is capable of dissipating a large amount of power, shows very little reactance at low frequencies, and is inexpensive.

resistance magnetometer A magnetometer whose operation is based upon the change of electrical resistance of a material (such as bismuth wire) placed in the magnetic field under test.

resistance material A substance, such as carbon or German silver, whose resistivity is high enough to enable its use as a lumped resistor. See, for example, RESISTANCE ALLOYS and RESISTANCE METAL.

resistance metal A metal, such as iron, whose resistivity is high enough to enable its use as a lumped resistor. Also see RESISTANCE ALLOYS.

resistance pad An attenuator composed of noninductive resistors.

resistance standard A highly accurate and stable resistor used in precision measurements of resistance. Also see PRIMARY STANDARD and SECONDARY STANDARD.

resistance strain gauge An electrical strain gauge in which the stressed element is a thin resistance wire.

resistance strip A strip of metallic or nonmetallic resistance material. Also see RESISTANCE ALLOYS and RESISTANCE METAL.

resistance temperature detector A transducer consisting of a specially made resistor whose resistance varies linearly with temperature.

resistance thermometer An electronic thermometer whose operation is based on the change of resistance of a wire as it is heated or cooled.

resistance transducer See RESISTIVE TRANSDUCER.

resistance tuning See VARIABLE-RESISTANCE TUNING.

resistance welding An electrical or electronic welding process in which the workpieces are heated by current flowing through the inherent resistance of their junction.

resistance wire Wire made of a metal or alloy that exhibits significant resistivity. See, for example, RESISTANCE ALLOYS and RESISTANCE METAL.

resistance-wire sensor A specific length of resistance wire, properly mounted, whose resistance is proportional to a sensed phenomenon (such as strain, temperature, presence of gas, pressure, etc.). See, for example, ELECTRICAL STRAIN GAUGE, GAS DETECTOR, and PRESSURE TRANSDUCER.

resistive current The component of alternating current that is in phase with voltage. Also called WATT CURRENT. Compare REACTIVE CURRENT.

resistive cutoff frequency Symbol, f_{rco}. The frequency beyond which a tunnel diode ceases to exhibit negative resistance.

resistive load A load device that is essentially a pure resistance.

resistive losses Losses resulting from the resistance of a circuit or device, and usually appearing as heat.

resistive transducer A transducer in which the sensed phenomenon causes a change in resistance, which in turn produces a corresponding change in output current or voltage. Compare CAPACITIVE TRANSDUCER, CRYSTAL TRANSDUCER, INDUCTIVE TRANSDUCER, MAGNETIC TRANSDUCER, and PHOTOELECTRIC TRANSDUCER.

resistive trimmer See TRIMMER RESISTOR.

resistive voltage The voltage across the resistance component in a circuit. In an alternating-current circuit, the resistive voltage is in phase with the current.

resistivity Symbol, r. Resistance per unit volume or per unit area. It can be expressed in terms of ohms per cubic meter or ohms per square meter. Also see MICROHM-CENTIMETER and OHM-CENTIMETER.

resistor A device having resistance concentrated in lumped form. Also see RESISTANCE and RESISTIVITY.

resistor-capacitor-transistor logic Abbreviation, RCTL. A form of RESISTOR-TRANSISTOR LOGIC in which capacitors are used to enhance switching speed.

resistor color code See COLOR CODE.

resistor core A form around which a resistance wire can be wound for the purpose of constructing a high-power resistor.

resistor decade See DECADE RESISTOR.

resistor diode A usually forward-biased semiconductor diode that acts as a VOLTAGE-DEPENDENT RESISTOR.

resistor FET See ELECTRONIC RESISTOR.

resistor fuse See FUSIBLE RESISTOR.

resistors in parallel See PARALLEL RESISTORS.

resistors in parallel-series See PARALLEL-SERIES RESISTORS.

resistors in series See SERIES RESISTORS.

resistors in series-parallel See SERIES-PARALLEL RESISTORS.

resistor substitution box A self-contained assortment of common-value resistors arranged to be switched one at a time to a pair of terminals. In troubleshooting and circuit development, any of several useful fixed resistance values can thus be obtained.

resistor transistor See ELECTRONIC RESISTOR.

resistor-transistor logic Abbreviation, RTL. A circuit in which the logic function is performed by resistors, and an inverted output is provided by transistors.

resnatron A form of vacuum tube that is used as an oscillator and amplifier at ultra-high and microwave frequencies. It is essentially a cavity resonator.

resolution **1.** The degree to which closely adjacent parts of an image can be differentiated. **2.** The reduction of a problem by means of logical analysis. **3.** The ability of a vision or ranging system to distinguish between objects that are close together in terms of radial distance, direction, or absolute separation.

resolution ratio In a television image, the ratio of horizontal resolution to vertical resolution.

resonance **1.** The state in which the natural response frequency of a circuit coincides with the frequency of an applied signal, or vice versa, yielding intensified response. **2.** The state in which the natural vibration frequency of a body coincides with an applied vibration force, or vice versa, yielding reinforced vibration of the body.

resonance bridge An alternating-current bridge (see BRIDGE, **2**) in which one or two arms are series-resonant or parallel-resonant, the other arms being resistances. Also see SERIES-TYPE RESONANCE BRIDGE and SHUNT-TYPE RESONANCE BRIDGE.

resonance curve A graph of the insertion gain or loss of a component, device, circuit, or system to variations in the frequency of an applied sound or signal. Such curves are almost always plotted in rectangular coordinates with frequency as the independent variable on the horizontal axis. The dependent variable, plotted on the vertical axis, can be any characteristic that displays a peak or dip at the resonant frequency or frequencies. In radio-frequency circuits, such parameters include current, voltage, attenuation, gain, and impedance.

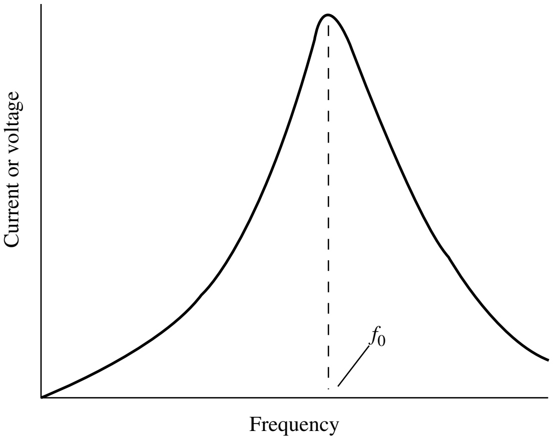

resonance curve

resonance theory of hearing The theory that sound waves pass down the *auditory canal* and cause the *eardrum* to vibrate. Behind the eardrum is a system of three bones leading to the *cochlea.* The cochlea consists of fibers that resonate. Therefore, they vary in length and tension. Various groups of fibers are activated by different sounds, and the vibrations are transmitted to nerves leading to the brain.

resonance radiation Electromagnetic radiation from an energized substance, resulting from movement of electrons from a higher to lower energy level. When an electron moves from a higher to a lower orbit, a photon, having a definite wavelength, is emitted.

resonant cavity A chamber whose size reinforces energy injected into it at a natural frequency, which is determined by the chamber's dimensions. Such cavities can be used with acoustic or electromagnetic waves.

resonant circuit A circuit whose constants are chosen for maximum circuit response at a given frequency. Examples: *parallel-resonant circuit* and *series-resonant circuit.* Also see RESONANCE and RESONANT FREQUENCY.

resonant current Current flowing in a tuned circuit at resonance.

resonant feeder An antenna feeder that is resonant at the operating frequency.

resonant filter A filter containing at least one series- or parallel-resonant arm for sharp response. Thus, a power-supply filter of this kind might have a parallel-resonant arm acting as a wavetrap at the ripple frequency.

resonant frequency Symbol, f_r or f_o. The natural frequency at which a circuit oscillates or a device vibrates. In an inductance-capacitance circuit (series-resonant or parallel-resonant), the reactances cancel at the resonant frequency, leaving only resistance.

resonant-gate transistor A transistor embodying a tiny tuning fork for resonance at low frequencies, thus eliminating bulky coils and capacitors.

resonant line A transmission line that is resonant at the operating frequency.

resonant-line amplifier See LINE-TYPE AMPLIFIER.

resonant-line circuit A circuit using resonant lines as a tank. See, for example, LINE-TYPE AMPLIFIER and LINE-TYPE OSCILLATOR.

resonant-line oscillator See LINE-TYPE OSCILLATOR.

resonant-line wavemeter See LECHER WIRES.

resonant rise See VOLTAGE RISE.

resonant-slope amplifier See DIELECTRIC AMPLIFIER.

resonant-slope detector See SLOPE DETECTOR.

resonant suckout The drawing of radio-frequency energy out of the energized part of a coil or transmission line by the part not intended to be energized, when the latter resonates at the same frequency.

resonant-voltage rise See VOLTAGE RISE.

resonant-voltage stepup See VOLTAGE RISE.

resonate **1.** To exhibit RESONANCE—either electrically or acoustically. **2.** To adjust a variable-frequency electrical or acoustical device so that it exhibits RESONANCE at a specific frequency.

resonator A device that produces or undergoes resonance. See, for example, HELMHOLTZ RESONATOR and RESONANT CAVITY.

resource A part of a computer system that can be used for a specific application as a unit (e.g., printer, PCMCIA standard adapter card, tape drive, etc.).

responder The transmitting section of a transponder.

response The behavior of a circuit or device (especially in terms of its dependent variables), in accordance with an applied signal (e.g., *frequency response* and *current-vs.-voltage response*).

response curve A graph depicting the performance of a circuit or device. Examples: *attenuation-vs.-frequency curve* and *current-vs.-voltage curve.*

response time The interval between the instant a signal is applied to or removed from a circuit or device and the instant that the circuit acts accordingly.

restart Following a malfunction or error occurring during a computer program run, to go back to an earlier point in the program.

resting state See QUIESCENT STATE.

restore See RESET.

resultant **1.** The vector that results from the addition of two or more vectors. **2.** A quantity that results from mathematical operations performed on other quantities.

retarding magnet See DRAG MAGNET.

retentivity **1.** The property whereby a material retains magnetism imparted to it. **2.** A quantitative measure of the extent to which a material retains magnetism imparted to it.

retention period In computer operations, the time during which the data on a magnetic medium (disk or tape) must be kept intact.

retrace **1.** In a cathode-ray tube, the return of the scanning beam to its starting point. **2.** In a cathode-ray tube, a line traced on the screen by the scanning beam as it returns to its starting point, if RETRACE BLANKING is not used.

retrace blanking Obliteration of the RETRACE of the electron beam in a cathode-ray tube. It renders the retrace line invisible on the screen so that it will not interfere with the display.

retrace line See RETRACE, **2**.

retrace ratio For the swept beam in a cathode-ray tube, the ratio of the scanning velocity in the trace direction to the scanning velocity in the RETRACE direction.

retrace time In a cathode-ray tube, the amount of time required for the scanning beam to move from the end of one trace or line to the beginning of the next.

retrofit To supply something with specially designed or adapted parts that were not available when it was made.

retrograde orbit For an artificial earth satellite, an orbit whose direction is east-to-west, relative to the earth's surface.

return **1.** See RETRACE. **2.** See RETURN CIRCUIT. **3.** See RETURN POINT. **4.** In an electronic circuit, the electrical ground and ground current path.

return circuit The circuit through which current returns to a generator.

return instruction In a computer program, an instruction in a subroutine directing operation back to a specific point in the main program.

return interval See RETRACE TIME.

return line See RETRACE, **2**.

return point **1.** The point to which circuits are returned (e.g., a common ground point). **2.** The terminal point of a return circuit.

return ratio See FEEDBACK FACTOR.

return time See RETRACE TIME.

return to zero **1.** Abbreviation RZ or RTZ. In the magnetic recording of data, a method in which the write current returns to zero following the write pulse. Compare NONRETURN-TO-ZERO. **2.** A logic system in which the zero and one states are represented by zero voltage and a discrete voltage.

return trace See RETRACE, **1**, **2**.

REV **1.** Abbreviation of REENTRY VEHICLE. **2.** Abbreviation of REVERSE.

rev **1.** Abbreviation of REVOLUTION. **2.** To quickly and substantially increase the angular velocity of a motor.

reverberation **1.** Multiple reflections of sound waves from the inside surfaces of an enclosed chamber. **2.** The dying-out of sound waves in an enclosed chamber as the waves reflect repeatedly from the inside surfaces. **3.** In sound recording and/or reproduction, an electronically produced echo. It is used for special effects—especially in electronic music systems. **4.** See RESONANCE, **2**.

reverberation chamber A room in which the walls, floor, and ceiling absorb very little sound, resulting in echoes. To avoid standing waves, the room is designed so that no two surfaces are exactly parallel.

reverberation system A system of devices operated with an electronic organ to simulate the effect of reverberation in a large room, such as a church auditorium.

reverberation time In an enclosed chamber, the time required for a sound to die down to a specified level (usually -60 dB) after the disturbance has stopped.

reverberation unit A device for producing artificial echoes—especially in the operation of electronic music systems.

reverse **1.** To alter the direction of a current or process or motion of an object so that the new direction is exactly opposite the previous direction. **2.** In a directional wattmeter, the reflected-power indication or switch position.

reverse AGC Automatic gain control in which a signal-dependent voltage is fed back to an earlier stage to adjust its gain automatically. Compare FORWARD AGC.

reverse bias Reverse voltage or current in a transistor or a semiconductor diode. Compare FORWARD BIAS.

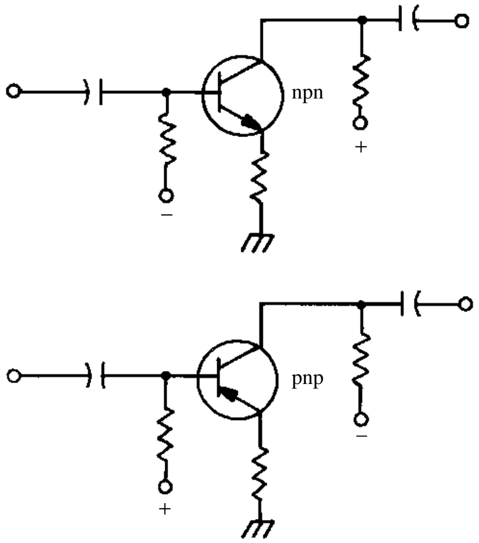

reverse bias

reverse breakdown See AVALANCHE.

reverse breakdown voltage See AVALANCHE VOLTAGE.

reverse characteristic The current-vs.-voltage response of a semiconductor junction that is biased in the reverse (low-conduction) direction. Compare FORWARD CHARACTERISTIC.

reverse conduction The very small current conduction through a pn junction when it is reverse-biased. Compare FORWARD CONDUCTION.

reverse current Symbol, I_r. The current that flows through a pn junction when it is reverse-biased. Also called *back current.* Compare FORWARD CURRENT.

reverse engineering A design process in which a specific device or system is copied functionally, but not literally.

reverse Polish notation Abbreviation, RPN. A system of notation for expressing mathematical operations in which the operators follow the operands being manipulated. It is a mode of entry for some calculators (e.g., the operation 7 × 2 might be entered by pressing keys in this order: 7, ENTER, 2, ×).

reverse recovery time See RECOVERY TIME, **1**.

reverse resistance Symbol, R_r. The resistance of a reverse-biased pn junction. Also called BACK RESISTANCE. Compare FORWARD RESISTANCE.

reverse voltage Symbol, E_r or V_r. Direct-current voltage applied to a pn junction so that the p-type material is electrically more negative than the n-type material. Also called BACK VOLTAGE. Compare VOLTAGE.

reverse-voltage capacitance The internal capacitance of a reverse-biased semiconductor pn junction.

reverse voltage drop The voltage drop across a semiconductor pn junction that is biased in the reverse (low-conduction) direction.

reversible counter A counter that, by a control signal, can have the value it is holding either increased or decreased.

reversible permeability The permeability of a ferromagnetic substance when the magnitude of the alternating-current field is arbitrarily small.

reversing switch **1.** A switch that reverses the polarity of a direct-current voltage. **2.** A switch that reverses the direction of motor rotation.

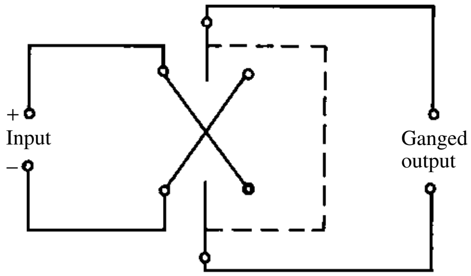

reversing switch

revolute geometry A method by which a robot arm can move freely in three dimensions. The entire assembly rotates from the base in a horizontal plane through a complete circle (360 degrees). An elevation joint at the base moves the arm from horizontal to vertical (90 degrees). A joint in the middle of the arm can bend through about 180 degrees.

revolution Abbreviation, r or rev. One complete rotation (i.e., 360 degrees of circular travel).

revolving field See ROTATING FIELD.

rewind To run a magnetic tape on a transport at a high speed, in the direction opposite to that associated with the play mode.

rewrite In computer operations, to return information read from a storage location to that location by recording.

R_F Symbol for FILAMENT RESISTANCE.

RF Abbreviation of RADIO FREQUENCY.

RF amplifier See RADIO-FREQUENCY AMPLIFIER.

RFC Abbreviation of RADIO-FREQUENCY CHOKE.

RF heating See RADIO-FREQUENCY HEATING.

RFI Abbreviation of RADIO-FREQUENCY INTERFERENCE.

RF inverse feedback A negative-feedback system for radiophone transmitters, in which a portion of the modulated radio-frequency (RF) signal is rectified, and the resulting direct-current voltage is filtered and applied as bias to one of the audio stages in the proper polarity for degeneration.

RF lamp A lighting lamp, used with radio-frequency (RF) alternating current, rather than the conventional 60-Hz utility current. This results in better efficiency, and allows much more light to be generated with a given filament lamp, as compared with 60-Hz current.

RF motion detector In security systems, an intrusion detection and alarm system that senses Doppler-effect-induced changes in the frequency or phase of a radio-frequency (RF) electromagnetic field. The Doppler effect results from motion of objects in the secured area.

RFO Abbreviation of *radio-frequency oscillator.*

RF power supply See OSCILLATOR-TYPE POWER SUPPLY.

RF preamplifier A sensitive, radio-frequency amplifier circuit intended for improving the signal-to-noise (S/N) ratio in a wireless receiver. Generally placed between the receiver and the antenna or feed line. Some such devices are tunable; others are broadbanded. See also PREAMPLIFIER.

RF probe See RECTIFIER PROBE.

RF resistance See RADIO-FREQUENCY RESISTANCE.

RF selectivity See RADIO-FREQUENCY SELECTIVITY.

RF transformer See RADIO-FREQUENCY TRANSFORMER.

RF transistor See RADIO-FREQUENCY TRANSISTOR.

R_g Symbol for GRID RESISTANCE.

R_G Symbol for GATE RESISTANCE.

RGB Abbreviation of RED-GREEN-BLUE.

RGT Abbreviation of RESONANT-GATE TRANSISTOR.

Rh Symbol for RHODIUM.

R/h Abbreviation of ROENTGENS PER HOUR.

R_H **1.** Symbol for HEATER RESISTANCE. **2.** Symbol for HOT RESISTANCE.

rh Abbreviation of RELATIVE HUMIDITY.

rhenium Symbol, Re. A metallic element. Atomic number, 75. Atomic weight, 186.207. It is used in some thermocouples.

rheostat A wirewound variable dropping resistor of the rotary type or slider type.

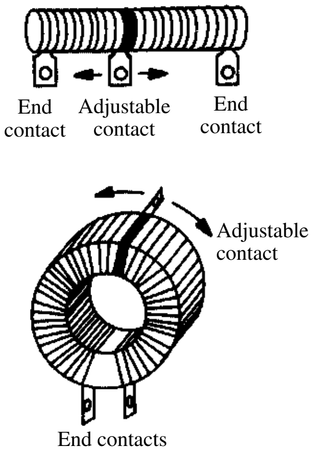

rheostat

R_{HF} Symbol for *high-frequency resistance* (see RADIO-FREQUENCY RESISTANCE).

RHI Abbreviation of RANGE-HEIGHT INDICATOR.

rhodium Symbol, Rh. A metallic element. Atomic number, 45. Atomic weight, 102.906.

rhombus A four-sided geometric plane figure, in which all four sides have equal length, and opposite angles have equal measure.

rhombic antenna See DIAMOND ANTENNA.

rho-theta A radio-navigation system in which a single transmitting station is used, and the position is determined, according to polar coordinates (distance and direction).

rhumbatron A RESONANT CAVITY—especially one in a KLYSTRON.

RI Abbreviation of RADIO INTERFERENCE.

R_i Symbol for INPUT RESISTANCE. (Also, R_{in}.)

RIAA Abbreviation of RECORDING INDUSTRY ASSOCIATION OF AMERICA.

RIAA curve The amplitude-versus-frequency function used in recording and reproduction of long-playing (33.3 rpm) phonograph discs, and specified by the Recording Industry Association of America (RIAA). The RIAA curve takes advantage of the sensitivity of the human ear at various frequencies to reduce the level of audible noise.

ribbon microphone See VELOCITY MICROPHONE.

ride gain In broadcasting, the operations of constantly adjusting the audio modulation of the transmitter for optimum operation.

Rieke chart A visual aid, similar to the SMITH CHART, used with traveling-wave tubes in the ultra-high-frequency (UHF) and microwave frequency bands for determining the optimum load impedance.

rig Colloquialism for a radio communications installation—especially a transmitter or transceiver. It is commonly used among amateur radio operators.

right-angle line section See ELL.

right-hand lay See DIRECTION OF LAY.

right-hand polarized wave See CLOCKWISE-POLARIZED WAVE.

right-hand rule for induced emf See FLEMING'S RIGHT-HAND RULE.

right-hand rule for wire A simple rule for indicating the direction of the magnetic field surrounding a straight wire that carries current. When the wire is grasped in the right hand with the thumb pointing in the direction of current flow, the fingers curl in the direction of the magnetic field.

right-hand taper Potentiometer or rheostat taper in which most of the resistance is in the clockwise half of rotation, as viewed from the front. Compare LEFT-HAND TAPER.

right justified In a computer memory location, a data item that takes up consecutive bit positions, from right to left.

Right-Leduc effect A phenomenon somewhat analogous to the Hall effect. When a metal strip conducting heat is placed in a magnetic field perpendicular to the plane of the strip, a temperature difference develops across the strip.

right shift In computer operation, a shift whereby word bits are displayed to the right; the effect is division in a right arithmetic shift.

rim drive **1.** In a tape recorder, a driving method in which the motor shaft is provided with a smooth pulley that transfers motion directly to the rubber-tired rim of the flywheel. **2.** A driving method for a phonograph turntable in which a rotating wheel contacts the outer edge of the platter.

R_{in} Symbol for INPUT RESISTANCE. (Also, R_i.)

ring **1.** The core of a toroidal coil. **2.** See HYBRID RING. **3.** See RING MODULATOR. **4.** See RING INDUCTOR. **5.** See RING MAGNET. **6.** See RINGING.

ring armature A motor or generator armature having a ring winding.

ringback In a telephone system, the signal sent from the receiving (destination) telephone set back to the sending (source) set, indicating that the signal is being received. This consists of a tone, interrupted by pulses at intervals of about 0.05 second. The signal stops when the destination set is taken off the hook. Also see RINGDOWN.

ring circuit **1.** See RING MODULATOR. **2.** A waveguide hybrid-tee resembling a ring having radial branches. **3.** In amateur radio operations, a circuit connected to a telephone line and radio transmitter. The radio transmitter is energized and modulated with an identifiable signal each time the telephone rings, so the telephone can be answered via remote control.

ring counter An electronic counter in which successive cascaded stages form a ring (i.e., the last stage in the chain is connected to the first stage so that the counter advances through the cycle, stage by stage, repetitively).

ringdown In a telephone system, the signal sent from the transmitting (source) set to the receiving (destination) set, causing the destination set to ring. Also see RINGBACK.

ring head In tape recorders, a recording and playback head that consists essentially of a metal ring with a gap at one point, and on which the coils are wound.

ring inductor **1.** An inductor consisting of a single turn of wire, or of a conductor bent into a loop. **2.** See SHADING COIL.

ringing Self oscillation in a pulsed inductance-capacitance circuit, sustained by the circuit's flywheel action (hysteresis), and usually producing a damped wave.

ringing coil In the horizontal oscillator in a television circuit, a small, adjustable coil (shunted by a capacitor) used to produce a sharp rise in input-signal voltage.

ringing current In wire telephony, an alternating current superimposed on the direct operating current. Produces RINGDOWN.

ringing time See RING TIME, **1**.

ring magnet A permanent magnet in the shape of a ring or donut.

ring main An electric power main that is closed to form a ring. This results in two independent electrical paths between any two points in the circuit. If one path is interrupted, power can still be transmitted to any other point in the circuit from a power station.

ring modulator A double-balanced diode-type modulator circuit; its name is derived from the ring-like arrangement of the four diodes.

ring oscillator A self-excited oscillator in which two sets of two transistors are operated in push-pull/parallel.

ring shift In computer operation, the cyclic shifting of digits from one end of a register to the other.

ring time **1.** The period of a damped oscillation—especially one set up in an inductance-capacitance circuit by a pulse. **2.** The time required for an ECHO BOX signal to decay below the display level.

ring winding A winding in which the turns of the coil are laid on the outside of a ring-shaped core and passed through its center, resulting in a donut coil with a core.

ripple **1.** A small alternating-current component in the output of a direct-current power supply with inadequate filtering. **2.** In computer and data-processing operations, the serial transmission of data.

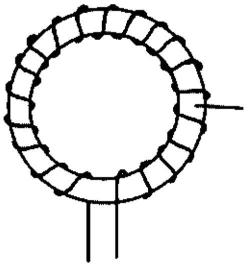

ring winding

ripple amplitude The peak or peak-to-peak value of ripple in a power supply (see RIPPLE, **1**).

ripple counter A binary counter consisting of flip-flops cascaded in series. A pulse must pass sequentially from the input, through each stage, to the output of the chain.

ripple current Current flowing in a circuit as the result of ripple voltage (see RIPPLE AMPLITUDE).

ripple factor The ratio of the RIPPLE AMPLITUDE to the direct-current voltage output of a power supply.

ripple frequency The frequency of a ripple component (see RIPPLE, **1**). In power supplies using half-wave rectification, this frequency is normally 60 Hz (the line frequency); in full-wave supplies, it is normally 120 Hz (twice the line frequency).

ripple percentage See PERCENT RIPPLE.

ripple torque Symbol, T_R. In a torque motor, the small fluctuation in torque resulting from commutator switching action.

ripple voltage See RIPPLE AMPLITUDE.

RISC Abbreviation of REDUCED INSTRUCTION SET COMPUTER.

rise **1.** See VOLTAGE RISE. **2.** See RISE TIME. **3.** An increase in the amplitude of a pulse or wave.

rise cable **1.** A vertical feeder cable. **2.** A vertical section of a feeder cable.

rise time The time required for a pulse to rise from 10 percent to 90 percent of its peak amplitude. Compare FALL TIME.

RJE Abbreviation of REMOTE JOB ENTRY.

R_k Symbol for cathode resistance.

RL **1.** Abbreviation of RESISTANCE-INDUCTANCE. **2.** Abbreviation of RELAY LOGIC.

R_L Abbreviation of LOAD RESISTANCE.

RL bridge See RESISTANCE-INDUCTANCE BRIDGE.

RL circuit See RESISTANCE-INDUCTANCE CIRCUIT.

R_{LF} Symbol for low-frequency resistance.

RL phase shifter See RESISTANCE-INDUCTANCE PHASE SHIFTER.

R_m Symbol for METER RESISTANCE.

r_m Symbol for *emitter-collector transresistance* of a bipolar transistor.

rms Abbreviation of ROOT MEAN SQUARE.

rms converter See ROOT-MEAN-SQUARE CONVERTER.

rms current See EFFECTIVE CURRENT.

rms meter A current meter or voltmeter whose deflection is proportional to the root-mean-square (rms) value of current or voltage. In most meters, the deflection is proportional to either the peak value or the average value, but the scale of an rms unit is graduated on the basis of sine-wave input.

rms value See EFFECTIVE VALUE.

rms voltage See EFFECTIVE VOLTAGE.

Rn Symbol for RADON.

R_n **1.** Symbol for NEGATIVE RESISTANCE. (Also, –R.) **2.** Symbol for *null resistance.*

R_o Symbol for OUTPUT RESISTANCE. (Also, R_{out}.)

roaming **1.** In cellular communications, an ideal condition in which the user of a mobile or portable wireless telephone set can travel throughout a specified geographic region with no "dead zones" or interruption in service. **2.** For a mobile or portable cellular telephone subscriber, the act of traveling throughout a large geographic region and, in particular, among different states or countries.

robot **1.** An electromechanical device or system capable of reliably performing complex and/or repetitive tasks. It can be controlled by a human operator or by a computer. **2.** A usually autonomous device, as defined in **1**, built to physically resemble a human being, with a head, two arms, and some form of locomotion.

robot generations Agreed-on milestones or criteria in the evolution of robots and smart machines. *First generation:* Before 1980. Mainly mechanical, stationary, physically rugged, no external sensors, no artificial intelligence. *Second generation:* 1980–1990. Tactile sensors, vision systems, position sensors, pressure sensors, microcomputer control, programmable. *Third generation:* After 1990. Mobile, autonomy or group control, artificial intelligence, speech recognition/synthesis, teleoperation, navigation systems. *Fourth generation:* In conceptual stages. Highly intelligent, capable of building other robots, capable of doing many human tasks.

robot gripper A robotic END EFFECTOR designed specifically to grasp objects. The two basic designs are: hand-like and specialized. Hand-like grippers are engineered according to the notion that the human hand has evolved to near perfection, and should be mimicked in robots. Specialized grippers are built by trial-and-error methods and often bear little resemblance to human hands.

robotics The science and art of robot design, construction, operation, and maintenance.

Rochelle salt Sodium potassium tartrate, a substance whose crystals are piezoelectric. Such crystals are used in some microphones, loudspeakers, and transducers. Also called SEIGNETE SALT.

rock Slang for QUARTZ CRYSTAL.

rockbound Pertaining to an oscillator or radio transmitter whose frequency is determined by a quartz crystal, and is, therefore, not variable over a significant range.

rocker switch A toggle switch whose lever is a specially shaped bar. The bar is rocked back and forth to operate the switch. Compare BAT-HANDLE SWITCH, PADDLE SWITCH, and SLIDE SWITCH.

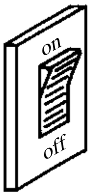

rocker switch

rod **1.** A unit of length or distance; 1 rod = 5.029 meters. **2.** A bar of material with special properties.

rod antenna See FERRITE-ROD ANTENNA.

rod magnet A permanent magnet in the shape of a rod with circular or elliptical cross section.

roentgen Abbreviation, r. A unit of ionizing radiation; 1 r is the quantity of radiation that produces 1 esu of electricity (positive or negative) per cubic centimeter of air at standard temperature and pressure. In average tissue, 1 r produces ionization equivalent to an energy concentration of 2.58×10^{-4} C/kg (93 ergs per gram). Also see MILLIROENTGEN.

roentgen equivalent man See REM.

roentgen equivalent physical See REP.

roentgen ray See X RAY.

roger A communications signal meaning "Acknowledged."

Roget spiral A spring-like wire device that contracts in proportion to the magnitude of the current flowing through it.

role indicator In computer operations, a code classifying a keyword as a part of the speech (e.g., noun).

roll **1.** In a display terminal having a line length of less than the standard 80 characters, an operating feature that allows the operator to follow the text along. The cursor remains fixed near the center of the displayed line, while the text moves from right to left. **2.** Vertical movement of a television picture, resulting from lack of vertical synchronization.

rollback In computer operations, the running again of a computer program or portion of the program. Also called RERUN.

roller inductor A variable inductor, usually of the air-core type, with a shaft attached that allows the whole assembly to be rotated; a wheel contact provides a movable tap. Such an indicator is continuously variable and is often used in such devices as antenna-tuning networks.

rolling In television, the apparent continuous upward or downward movement of the picture, resulting from lack of vertical synchronization between the transmitter and receiver.

rolloff **1.** The rate at which a dependent variable (e.g., output amplitude) diminishes above or below a certain critical value of the independent variable (e.g., frequency). It pertains especially to frequency response in audio devices and systems. **2.** Attenuation of the bass and/or treble response or output in an audio system.

ROM Abbreviation of READ-ONLY MEMORY.

Romeo Phonetic alphabet code word for the letter R.

Romex cable A form of wire cable with a covering that is highly resistant to the environment.

roof mount A metal bracket for fastening an antenna mast to a roof.

room noise Ambient acoustic noise in a room.

room resonance Acoustic resonance caused by the geometry and contents of a room.

room temperature Abbreviation, RT. **1.** The temperature of the chamber in which a test or fabrication is carried out. It is commonly used to distinguish between operations that can be performed at the ambient temperature and those that require an oven or a cold chamber. **2.** A temperature typical of an indoor environment, approximately 21 degrees Celsius (70 degrees Fahrenheit).

room tone A qualitative expression for the suitability or behavior of an enclosed area for a given acoustic application. It affected by factors (such as resonances or lack thereof, echoes or lack thereof, overall room size and proportions, and background noise).

root mean square Abbreviation, rms. The square root of the arithmetic mean (average) of the squares of a set of values.

root-mean-square converter A device that converts an input signal of any waveform into a direct-current signal of the same value as the EFFECTIVE VALUE of the input signal.

root-mean-square current See EFFECTIVE CURRENT.

root-mean-square value See EFFECTIVE VALUE.

root-mean-square voltage See EFFECTIVE VOLTAGE.

ROP Abbreviation of *record of production.*

rosin A substance derived chemically from an extract of pine wood, and used in some solders.

rotameter A fluid flow gauge consisting of a float within a glass tube having incremental markings.

rotary amplifier See AMPLIDYNE.

rotary antenna See ROTATABLE ANTENNA.

rotary beam A beam antenna, such as a Yagi, that can be rotated in a (usually horizontal) plane to allow transmission and/or reception in various directions. Also see ROTATABLE ANTENNA.

rotary-beam antenna See ROTARY BEAM.

rotary converter A dynamo (electric machine) having a direct-current armature connected to a commutator on one end of the shaft and to slip rings on the other end. When the machine is operated as a direct-current motor, it delivers an alternating-current output, and vice versa. Also called DOUBLE-CURRENT GENERATOR.

rotary dialing An older style of telephone dialing in which a rotary pulse generator is used to dial the digits.

rotary digital audio tape Digital audio tape used with a recording/playback system that uses a rotating head or heads.

rotary encoder An optical analog-to-digital (A/D) converter that operates by passing a light beam through a rotating disk. The amplitude of the analog input signal at any moment causes a certain angular rotation of the disk, cutting off the light beam, according to the nature of the pattern on the disk. The resulting modulated light beam has digital characteristics and can be detected using photocells.

rotary inverter A motor-generator used to change a direct-current input voltage into an alternating-current output voltage.

rotary-motion sensor A transducer that delivers an output voltage proportional to the arc over which its shaft has been turned.

rotary power amplifier See DC GENERATOR AMPLIFIER.

rotary relay An electromechanical relay in which a pivoted armature rotates to open or close the contacts. The *meter relay* is an example.

rotary selector switch See ROTARY SWITCH.

rotary stepping relay See STEPPING SWITCH.

rotary stepping switch See STEPPING SWITCH.

rotary switch A switch in which a blade moves in a circle or in arcs over the contacts.

rotary transformer A motor-generator used to change an input voltage into a lower or higher output voltage.

rotatable antenna An antenna that can be turned to change the orientation of its main lobe (direction of greatest forward gain) in a specific plane (usually horizontal).

rotating amplifier See DC GENERATOR AMPLIFIER.

rotating antenna An antenna that constantly turns to scan a given area (e.g., a RADAR ANTENNA).

rotating-capacitor modulator A frequency modulator consisting of a motor-driven variable capacitor. Also see WOBBULATOR.

rotating field An alternating-current electric or magnetic field, such as that generated by the stators of some motors, that revolves between poles.

rotating interrupter A commutator (see COMMUTATOR, **1**).

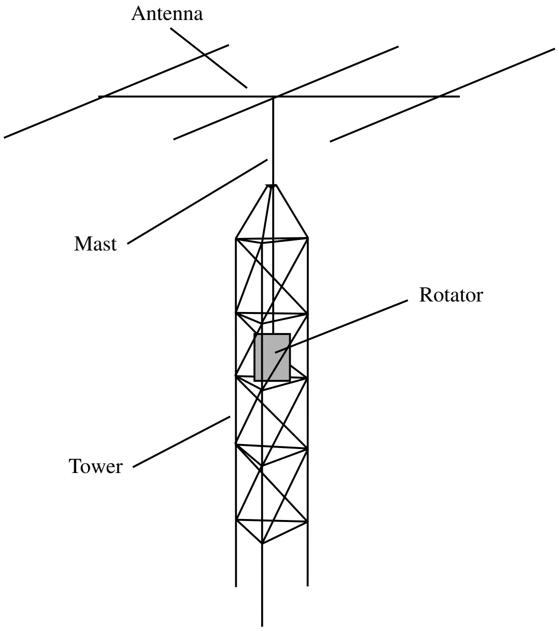

rotatable antenna

rotating machines Electromechanical devices (such as motors, generators, amplidynes, rotary converters, etc.) that utilize magnetic flux to convert angular motion or displacement into electrical energy, or vice-versa.

rotating memory See DISK MEMORY and DRUM MEMORY.

rotating voltmeter See GENERATING VOLTMETER.

rotator A motor-driven, remotely controlled mechanism for turning a directional antenna in a specific plane.

rotor **1.** A rotatable coil—especially in a motor or generator. Compare STATOR, **1**. **2.** The rotating member of a motor or generator. Compare STATOR, **2**. **3.** The rotating-plate assembly of a variable capacitor. Compare STATOR, **3**.

rotor blade The wiper arm of a rheostat or potentiometer.

rotor coil See ROTOR, **1**.

rotor plate The rotating plate(s) of a variable capacitor. Compare STATOR PLATE.

roulette pattern A circular pattern for frequency identification with an oscilloscope, consisting of loops around the screen's circumference. Compare GEAR-WHEEL PATTERN, LISSAJOUS FIGURE, and SPOT-WHEEL PATTERN.

rounding **1.** A method of approximating a quantity by reducing the number of significant digits. For example, rounding 3.44 to two significant digits yields 3.4; rounding 3.46 to two significant digits yields 3.5. Compare TRUNCATION. **2.** The approximation of a value to a specified number of decimal places or significant digits. **3.** Smoothing of the corners of a square wave or sawtooth wave, resulting in lengthening of the transition time from one state to another.

rounding error The error resulting from the rounding of a number (see ROUNDING, **1**, **2**).

round off To shorten an otherwise lengthy number by replacing numerical digits with zeros and increasing the final nonreplaced digit by 1 if the leftmost replaced digit is 5 or greater. Thus, 3.141592653 can be rounded off to 3.1416 or 3.14.

round-up A form of numerical approximation, in which a number with a value of $n.5$ or greater is assigned the value $n + 1$. This is a feature of many calculators using scientific notation or a fixed number of decimal places.

R_{out} Symbol for OUTPUT RESISTANCE. (Also, R_o.)

route **1.** To physically position wires or conducting circuit paths by planning and deliberation. **2.** The path over which conductors are positioned. **3.** A path over which signals or information can be carried.

routine **1.** In computer operations, the complete sequence of instructions for performing an operation (i.e., a program or program segment). **2.** A test or measurement procedure. **3.** An assembly or manufacturing procedure. **4.** A standard troubleshooting procedure.

row In a matrix, a horizontal arrangement or set of values.

R_P **1.** Symbol for PLATE RESISTANCE. (Also, r_P.) **2.** Symbol for POSITIVE RESISTANCE. **3.** Symbol for PARALLEL RESISTANCE. (Also, R_{par}.) **4.** Symbol for PRIMARY RESISTANCE. (Also, R_{pri}.)

R_{par} Symbol for PARALLEL RESISTANCE. (Also, R_P.)

r parameters **1.** Device or network parameters expressed as resistances. **2.** Transistor parameters in terms of resistance values in the equivalent tee network. Compare G PARAMETERS and H PARAMETERS.

RPG Abbreviation of REPORT PROGRAM GENERATOR.

rpm Abbreviation of *revolutions per minute.*

rpm meter See ELECTRONIC TACHOMETER and STROBOSCOPE.

RPN Abbreviation of REVERSE POLISH NOTATION.

R_{pri} Symbol for PRIMARY RESISTANCE. (Also, R_P.)

rps Abbreviation of *revolutions per second.*

RPT Radiotelegraphic abbreviation of *repeat.*

R_{req} **1.** Symbol for REQUIRED RESISTANCE. **2.** Symbol for SCREEN RESISTANCE.

R_{RF} Symbol for RADIO-FREQUENCY RESISTANCE.

R_S Symbol for SOURCE RESISTANCE in a field-effect transistor.

R_s **1.** Symbol for SERIES RESISTANCE. (Also, R_{ser}.) **2.** Symbol for SECONDARY RESISTANCE. (Also, R_{sec}.)

R_{sec} Symbol for SECONDARY RESISTANCE. (Also, R_s.)

R_{ser} Symbol for SERIES RESISTANCE. (Also, R_s.)

RST flip-flop A conventional flip-flop subject to the operations of reset, set, and trigger.

RST system In the amateur radio service, a method of reporting signal quality in terms of *readability*, *strength*, and *tone.*

R sweep In an oscilloscope or spectrum analyzer, an expanded portion of the trace produced by a long triggered sweep. It permits detailed analysis of a small portion of a displayed waveform or frequency band.

RT **1.** Abbreviation of RADIOTELEPHONE. **2.** Abbreviation of ROOM TEMPERATURE.

rt Abbreviation of RIGHT.

R_T **1.** Symbol for THERMAL RESISTANCE. **2.** Symbol for *total resistance.* (Also, R_t.)

R_t Symbol for *total resistance.*

RTD Abbreviation of RESISTANCE TEMPERATURE DETECTOR.

RTL Abbreviation of RESISTOR-TRANSISTOR LOGIC.

RTTY Abbreviation of RADIOTELETYPE.

RTZ Abbreviation of RETURN TO ZERO. (Also, RZ.)

Ru Symbol for RUTHENIUM.

rubber A natural insulating material; an elastomer exhibiting rapid elastic recovery. Dielectric constant, 2 to 3.5. Dielectric strength, 16 to 50 kV/mm. Also called *India rubber.* Compare HARD RUBBER.

rubber-covered wire Wire insulated with a jacket of rubber.

rubidium Symbol, Rb. A metallic element. Atomic number, 37. Atomic weight, 85.468.

ruby laser A laser using a ruby rod as the resonant element.

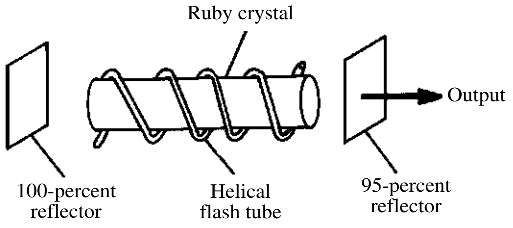

ruby laser

ruby maser A maser in which the resonant material is ruby.

Ruhmkorff coil See INDUCTION COIL.

rumble **1.** Low-frequency acoustical noise of a frequency below about 50 Hz. **2.** Low-frequency electrical noise in an audio circuit of a frequency below about 50 Hz. **3.** Vibrations that can occur in a poorly designed or malfunctioning phonograph turntable.

rumble filter An audio high-pass filter having sharp cutoff below 50 Hz, for eliminating rumble arising from irregularities in the rotation of a phonograph turntable. Also see RUMBLE, **3**.

run **1.** The execution of a computer routine or program. **2.** To execute a routine or program. **3.** To cause a routine or program to be executed. **4.** A command that causes a routine or program to be executed.

runaway In a current-carrying circuit or device, especially a semiconductor, a rapid increase in current that causes the temperature to rise, in turn resulting in a further increase in current. Unless preventive measures are taken, this will ultimately damage or destroy the component.

run chart In computer operations, a flowchart showing the organization and order of pertinent programs to be run.

running accumulator A computer storage unit having registers linked so that data is transferred unidirectionally from one to the other, and in which only one register is accessible from the outside.

running open **1.** The condition of a mechanical teleprinter running continuously in the absence of a signal. The teleprinter operates, but nothing is printed; this keeps the machine in synchronization. **2.** Operation of a transmitter at the maximum rated level of input or output power.

running-time meter See ELAPSED-TIME METER.

run time **1.** The period of time during which a computer program is executed. **2.** The length of time required for a computer program to be executed.

rupture **1.** The usually rapid and violent tearing apart, or breaking through, of an insulating material subjected to excessive voltage. **2.** The clean opening of relay, circuit-breaker, or switch contacts to interrupt a current-carrying circuit.

rush Broadband audio background noise, such as that arising from superheterodyne receivers and high-gain amplifiers. Its name is derived from resemblance to the gentle rushing of wind. Compare HISS, **1**, **2**.

ruthenium Symbol, Ru. A rare metallic element. Atomic number, 44. Atomic weight, 101.07.

rutherford Abbreviation, rd. A unit of radioactivity equal to 10^6 disintegrations per second (2.7×10^{-5} curie). Also see KILORUTHERFORD, MEGARUTHERFORD, MICRORUTHERFORD, and MILLIRUTHERFORD.

Rutherford atom An early concept of the nature of the atom, proposed by Rutherford in 1912. In this model, negatively charged electrons orbit a central, positively charged nucleus in a manner similar to the way planets orbit the sun. Compare BOHR ATOM.

rutherfordium Symbol, Rf. Also called *unnilquadium* (Unq) and *kurchatovium* (Ku). Atomic number, 104. The most common isotope has atomic weight 261. Classified as a transition metal. It has a half-life on the order of a few seconds to a few tenths of a second (depending on the isotope), is human-made, and is not known to occur in nature.

RW Abbreviation of *radiological warfare.*

R_x Symbol for unknown resistance.

RY Abbreviation of RELAY.

ryotron A form of inductive semiconductor switch, operated at cold temperatures to maximize conductivity.

R-Y signal In a color-television circuit, the signal representing primary red (R) minus luminance (Y). A primary red signal is obtained when the R-Y signal is combined with the luminance (Y) signal. Compare B-Y SIGNAL and G-Y SIGNAL.

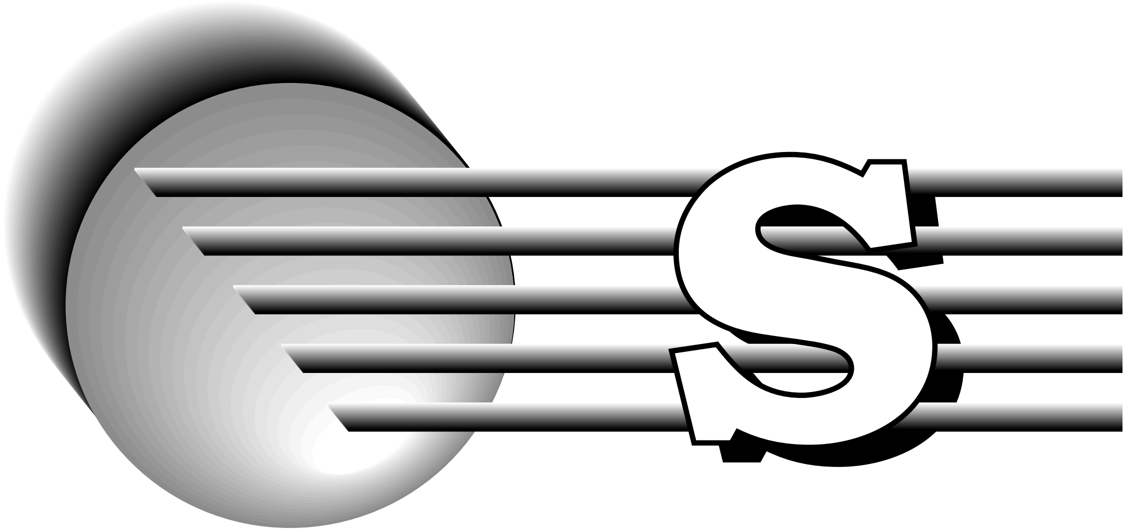

S **1.** Symbol for SCREEN GRID of a vacuum tube. **2.** Symbol for SHELL of a tube or semiconductor device. **3.** Symbol for SULFUR. **4.** Symbol for DEFLECTION SENSITIVITY. **5.** Symbol for SWITCH. **6.** Symbol for ELASTANCE. **7.** Abbreviation of SYNC. **8.** Symbol for SECONDARY. **9.** Abbreviation for SIEMENS. **10.** Abbreviation of SINE. **11.** Symbol for ENTROPY.

s **1.** Symbol for distance or DISPLACEMENT. **2.** Symbol for SCREEN GRID of a vacuum tube. **3.** Symbol for STANDARD DEVIATION. **4.** Abbreviation for SECOND.

SA Abbreviation of SUBJECT TO APPROVAL.

S_A band A section of the S band, extending from 3100 to 3400 MHz.

sabin A unit of sound absorption; 1 sabin represents a surface that can absorb sound at the same rate as 1 square foot of a perfectly absorbent surface.

SADT Abbreviation of *surface alloy diffused-base transistor.*

SAE **1.** Abbreviation of SHAFT-ANGLE ENCODER. **2.** Abbreviation of *Society of Automotive Engineers.*

safe noise level A level of acoustic intensity equal to 85 dB above the threshold of hearing. At sound levels higher than this, eardrum damage is possible.

safety factor A figure denoting the extent of overload that a device can withstand before breaking down.

safety ground A connection made between equipment (usually the metal chassis, panel, case, or B-minus circuit) and the earth as a protective measure against fire and electric shock.

safety switch See ELECTRICAL INTERLOCK.

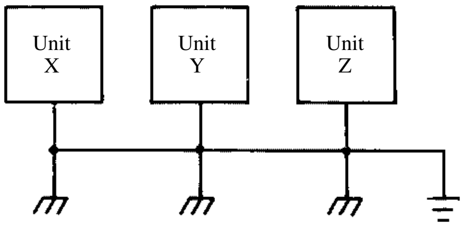

safety ground

sal ammoniac Formula, NH_4Cl. Ammonium chloride, the principal ingredient in the gelatinous electrolyte of some dry cells and batteries.

salient pole A pole, such as the polepiece of a motor or generator, that projects from the rest of the structure (rotor assembly or motor frame).

Salisbury chamber A radio-frequency test chamber in which the walls are non-reflective at various frequencies, thus simulating free space.

salt-spray test A test to assess the life and performance of electronic equipment in a marine environment. The equipment is sprayed, usually with a saltwater mist, and various electrical parameters are measured at prescribed time intervals.

samarium Symbol, Sm. A metallic element of the rare-earth group. Atomic number, 62. Atomic weight, 150.36.

sample **1.** A selection of quantities, events, or objects, taken at a specific time or time interval for analysis or testing. **2.** To take a sample, as defined in **1**. **3.** To test a quantity (current, voltage, temperature, pressure, etc.) or a material

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

(electrolyte, insulant, corrosion, rust, etc.) taken from a larger group or body.

sample and hold A method of storing a variable signal for detailed examination.

sampled data system A system that can be either analog or digital, and that works from samples of the input signals.

sampler In audio systems, a device that digitizes and stores sound for a brief interval of time.

sample size In statistics, the number of items in the sample space chosen for analysis.

sample space In statistics, the set of events, numbers, or other items chosen for analysis.

sampling **1.** Observation of a signal at various points in a circuit, without affecting the operation of the circuit. **2.** The conversion of analog signals to binary signals—especially for use in digital communications systems and in computers. **3.** In statistics and probability, a set of function values corresponding to specifically chosen points in the domain.

sampling rate The frequency with which samples are taken [e.g., 1/hr (one sample per hour) or 10/min (10 samples per minute)].

sampling window See WINDOW, **2**.

sand load A microwave power dissipator in which the absorptive material is a mixture of sand and carbon.

sapphire needle See SAPPHIRE STYLUS.

sapphire stylus A jewel-tipped stylus for disc recording and playback. It is noted for durability.

sat **1.** Abbreviation of *saturate.* **2.** Abbreviation of SATURATION. **3.** Abbreviation of SATELLITE.

satd Abbreviation of *saturated.*

satellite An artificial object sent into orbit around the earth or another planet. See, for example, ACTIVE COMMUNICATIONS SATELLITE and PASSIVE COMMUNICATIONS SATELLITE.

satellite communication Communication via one or more satellite transponders. Usually both stations are on the ground, although sometimes one or both stations are airborne or in space.

satellite processor **1.** In a computer, a microprocessor that is subsidiary to the central processing unit (CPU). **2.** In a data-processing system, a CPU used to handle the running of programs of secondary importance to the system's main application.

satellite television The broadcasting and reception of television (TV) signals via earth-orbiting satellite. Usually, the satellite is in a geostationary orbit. The receiving station employs a dish antenna, a tuner, a digital-to-analog (D/A) converter (if the signals are digital), and a TV receiver.

satisfy To make a statement of inequality or an equation true (e.g., $x = 2$ satisfies the equation $2x = x^2$).

saturable capacitor A voltage-variable ceramic or semiconductor capacitor in which variations in capacitance stabilize at a reasonably constant value after a particular voltage level is reached.

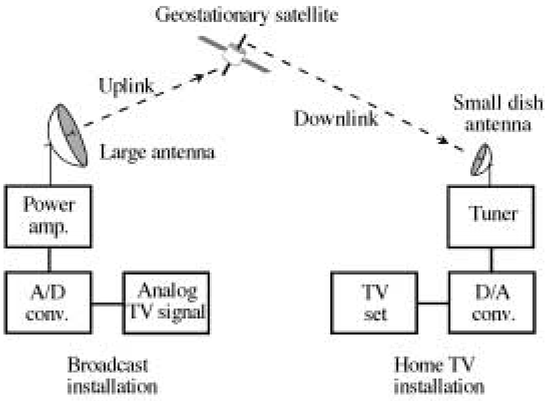

satellite television

saturable-core magnetometer A MAGNETOMETER in which the sensor is a saturable magnetic core with a winding. The readout is proportional to the change in permeability of the core produced by the magnetic field under test.

saturable reactor An inductor consisting essentially of a coil wound on a core of magnetic material whose magnetic flux can easily reach saturation level. The inductance and, accordingly, the reactance of the device can be varied by passing a direct current through the coil simultaneously with the alternating current to be controlled.

saturable transformer A transformer having a saturable core that permits automatic regulation of an alternating-current voltage.

saturated color A visible color whose energy is concentrated at a single wavelength or in a narrow band of wavelengths. Also called *pure color.*

saturated logic Any form of digital-logic circuit in which the transistors are either completely cut off or completely saturated. It is characterized by relative immunity to noise, high speed, and high input-level requirements.

saturated operation **1.** The operation of a magnetic core at or beyond its saturation point (i.e., in the region where an increase in coil current produces no change in core magnetization). **2.** The operation of a transistor or vacuum tube beyond its saturation point (i.e., in the region where an increase in voltage produces no change in current, or vice versa). Compare UNSATURATED OPERATION.

saturated solution A solution, such as an electrolyte, that contains all of the solute that it ordinarily will hold at a given temperature and pressure. Compare SUPERSATURATED SOLUTION. Also see SOLUTE; SOLUTION, **1** and SOLVENT, **2**.

saturating current See SATURATION CURRENT.

saturation **1.** See SATURATION POINT. **2.** The state of purity of a color. In general, the greater the saturation, the narrower the bandwidth of the electromagnetic energy. The highest possible saturation is represented by energy at a single wavelength [e.g., light of 700 nanometers (nm) appears as highly saturated red, and light of 400 nm appears as highly saturated violet]. Compare HUE. **3.** The condition of a ferromagnetic material in which it can accommodate no additional magnetic flux. **4.** The condition of a dielectric material in which it can accommodate no additional electric flux.

saturation current In a device, the current flowing at and beyond the SATURATION POINT.

saturation flux **1.** The magnetic flux density that will saturate a given sample of magnetic material. **2.** The electric flux density that will saturate a given sample of dielectric material.

saturation flux density See SATURATION INDUCTION.

saturation induction For a magnetic material, the maximum possible flux density.

saturation limiting Output peak clipping that occurs when a transistor or vacuum tube is driven into saturation during part of the input cycle. Compare CUTOFF LIMITING.

saturation point On a voltage-current conduction curve, the point beyond which a further increase in voltage produces no appreciable increase in current.

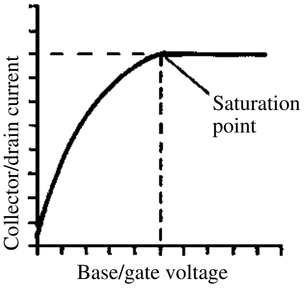

saturation point

saturation resistance The voltage-to-current ratio for a saturated semiconductor.

saturation switching The on/off switching operation in which a transistor is in its saturated state when conducting.

saturation value **1.** In a transistor, field-effect transistor, or vacuum tube, the lowest level of the input current, voltage, or power that results in saturation. **2.** The maximum obtainable output level for a given circuit. **3.** In a magnetic material, the smallest level of magnetizing force that results in the maximum possible flux density.

saturation voltage The (usually direct-current output) voltage appearing across a device operating in its saturation region (e.g., the collector voltage of a switching transistor in its switched-on state).

SAVOR Abbreviation of SIGNAL-ACTUATED VOICE RECORDER.

sawtooth An alternating or pulsating wave of current or voltage that is characterized by a gradual rise in amplitude followed by a rapid fall, or vice versa; its name is derived from its graphic resemblance to the teeth of a saw.

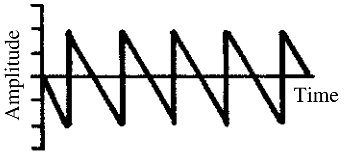

sawtooth

SB **1.** Abbreviation of SIDEBAND. **2.** Abbreviation of SIMULTANEOUS BROADCAST.

Sb Symbol for ANTIMONY.

S band A radio-frequency band extending from 1550 to 5200 MHz. For subdivisions of this band, see S_A BAND, S_C BAND, S_D BAND, S_F BAND, S_G BAND, S_H BAND, S_Q BAND, S_S BAND, S_T BAND, S_W BAND, S_Y BAND, and S_Z BAND.

SBC Abbreviation of SINGLE-BOARD COMPUTER.

SBDT Abbreviation of SURFACE-BARRIER DIFFUSED TRANSISTOR.

SBT Abbreviation of SURFACE-BARRIER TRANSISTOR.

SC **1.** Abbreviation of SUPPRESSED CARRIER. **2.** Abbreviation of SHORT CIRCUIT. **3.** On drawings, the abbreviation for *silk-covered.*

Sc Symbol for SCANDIUM.

sc **1.** Abbreviation of SINGLE CRYSTAL. **2.** Abbreviation of SCALE. **3.** Abbreviation of *sine-cosine.* **4.** Abbreviation of *science.*

SCA Abbreviation of SUBSIDIARY COMMUNICATION AUTHORIZATION.

SCA adapter An auxiliary tuner unit for separating the SCA subcarrier from a main frequency-modulated signal on which it is superimposed. Also see SCA SUBCARRIER and SUBSIDIARY COMMUNICATIONS AUTHORIZATION.

scalar quantity A quantity having magnitude, but for which direction is not specified. Compare VECTOR QUANTITY.

scale **1.** A graduated line or curve for indicating values of a quantity. **2.** An ordered set of values. **3.** An ordered series of quantities, such as tones, frequencies, voltages, etc. (e.g., *musical scale*).

scale division The space between consecutive graduations on a scale (see SCALE, **1**).

scale down In computer operations, to adjust a group of quantities according to a fixed factor so that it can be accommodated by hardware or software.

scale expansion Spreading out the divisions in part of a scale (see SCALE, **1**).

scale factor **1.** A figure by which the readings from a particular scale must be multiplied or divided to give the true values of measured quantities. **2.** A figure via which values in one system of notation are converted to those in another system. **3.** In scaling down (see SCALE DOWN), the factor by which a group of quantities is adjusted. **4.** The ratio of output frequency to input voltage for a voltage-to-frequency converter.

scale-factor adjustment In some meters, an adjustment that allows full-scale deflection to be set at any desired value (within certain limits) of applied-signal amplitude.

scale-factor error The absolute value of the difference between the actual scale factor and the ideal scale factor for a multiplier circuit.

scale-factor tolerance The extent to which a measured value for the scale factor differs from the computed value. It is generally given as a percentage.

scale length The end-to-end dimension of a scale (see SCALE, **1**), in inches, centimeters, geometric degrees, or number of divisions.

scale multiplier See SCALE FACTOR, **1**.

scale-of-two counter A circuit that delivers one output pulse for two successive input pulses.

scale-of-10 counter A circuit that delivers one output pulse for 10 successive input pulses.

scale-of-ten scaler See SCALE-OF-10 COUNTER.

scaler A circuit or device for extending the frequency range of another device (e.g., a circuit that extends the range of a 1-MHz counter to 100 MHz).

scale range The difference between the lowest and highest values on a scale.

scale span See SCALE RANGE.

scaling The fact, and the implications of the fact, that the mechanical strength of a structure increases in proportion to the square of linear dimension while the weight increases, according to the cube of linear dimension. Thus, weight increases more rapidly than strength as a structure, composed of a given material, is made larger. It is important in the design of large antennas and support structures.

scaling adder An inverting OPERATIONAL AMPLIFIER used to weight and sum multiple voltages.

scaling circuit A circuit, such as one or more flip-flops, that will deliver one output pulse after a predetermined number of input pulses have been received; therefore, it will provide pulse or frequency division. See, for example, SCALE-OF-TWO COUNTER.

scaling factor For a scaler, the number of input pulses required for one output pulse.

scaling ratio See SCALING FACTOR.

scan **1.** To traverse a range, field, or dimension. **2.** The amount of traversal in **1**. **3.** See SWEEP. **4.** To sample or reproduce an image in a single-line element, as in facsimile or television. **5.** A single line resulting from 4. **6.** In information retrieval operations, to inspect each record in a file or constituent of a list. **7.** To check communications or data channels for availability.

scan conversion In television reception, the scanning of each line twice to convert a conventional image into one that can be displayed on a high-definition picture tube.

scan-converter tube A face-to-face assembly of a cathode-ray tube and a vidicon in one envelope.

scandium Symbol, Sc. A metallic element. Atomic number, 21. Atomic weight, 44.956.

scan frequency See SCANNING FREQUENCY.

scanner A device, especially a radio receiver, equipped with a circuit that searches communications or data channels for signals.

scanner amplifier An amplifier for boosting a scanning signal. Also see SCAN, **1**, **3**, **4**.

scanning **1.** In a cathode-ray tube or camera tube, the synchronized movement of the electron beam (or other marker) from right to left and/or from top to bottom. **2.** The intermittent, but repetitive, monitoring of two or more communications channels in rotating sequence. **3.** The movement of a radar beam for the purpose of obtaining coverage over a specified area.

scanning antenna A transmitting or receiving antenna (such as a rotating one) that covers a generally circular region.

scanning beam The deflected electron beam in a cathode-ray tube. Also see SCAN, **1**, **2**, **3**, **4**, **5**.

scanning circuit A circuit for producing a scan (see SCAN, **2**, **4**).

scanning frequency The number of scans per unit time, usually expressed in lines, sweeps, or channels per second or per minute. Also called SCANNING RATE.

scanning line A single line sampled or produced by a scanning process, as in facsimile, television, and graphic recording.

scanning line frequency See SCANNING FREQUENCY.

scanning linearity Uniformity of scanning rate. In a linear scan, for example, scan speed is the same at all points along a line.

scanning loss The effective reduction in radar sensitivity that occurs as the beam scans a given area, rather than remaining in a fixed orientation.

scanning rate See SCANNING FREQUENCY.

scanning receiver A receiver whose tuning is automatically and continuously swept through a frequency band to detect all signals in the band.

scanning sonar A form of distance-measuring or depth-finding sonar, in which the receiving transducer scans to find the direction of the echo or echoes.

scanning speed The rate at which a line, region, or quantity is scanned or at which samples are taken.

scanning yoke See YOKE, **2**.

scan rate **1.** The rate at which a controlled quantity is checked periodically by a control computer. **2.** See SCANNING FREQUENCY.

scan tuning Repetitive, automatic sweeping of a frequency band by a tuned circuit containing a varactor, whose capacitance is periodically varied by a sawtooth voltage.

SCARA Abbreviation of SELECTIVE COMPLIANCE ASSEMBLY ROBOT ARM.

SCA subcarrier An auxiliary carrier (commonly 67 kHz) superimposed on a main frequency-modulated carrier to convey subsidiary communications, such as music without commercials. Also see SUBSIDIARY COMMUNICATION AUTHORIZATION.

scatter To disperse or diffuse transmitted electromagnetic radiation.

scattering **1.** The tendency of a concentrated beam of energy to be spread out when it passes through a given medium or substance. **2.** The spreading out of radio waves as they pass through the ionosphere or troposphere.

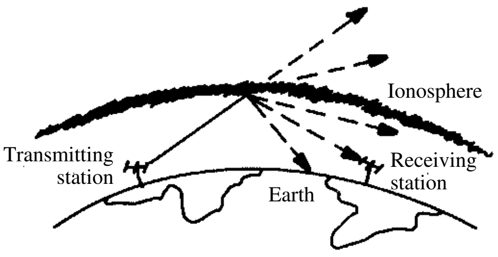

scattering

scatter read In data processing, to distribute data from an input record to several storage areas.

scatter transmission See BACK SCATTER and FORWARD SCATTER.

S_C band A section of the S BAND extending from 2000 to 2400 MHz.

scc Abbreviation of SINGLE-COTTON-COVERED (WIRE).

SCDSB Abbreviation of SUPPRESSED CARRIER DOUBLE-SIDEBAND (see DOUBLE-SIDEBAND SUPPRESSED CARRIER).

SCE Abbreviation of *saturated calomel electrode.*

sce Abbreviation of SINGLE-COTTON-ENAMELED (WIRE).

SCEPTRON Acronym for SPECTRAL COMPARATIVE PATTERN RECOGNIZER.

schedule In computer operations, to establish the order of importance of jobs to be run, and assign the necessary resources for those jobs.

schedule method A method of waveform analysis involving the evaluation of instantaneous amplitudes at numerous points in time. The values are obtained at specific intervals from the image of one complete wave cycle, as displayed on an oscilloscope or plotted on a graph.

schematic diagram An electronic circuit diagram. Also called WIRING DIAGRAM and *schematic.*

schematic symbol A graphic symbol used to represent electronic components in a circuit diagram.

Schering bridge A four-arm capacitance bridge in which the unknown capacitance is compared with a standard capacitance. This bridge is frequently used in testing electrolytic capacitors, to which a direct-current polarizing voltage is applied during the measurement.

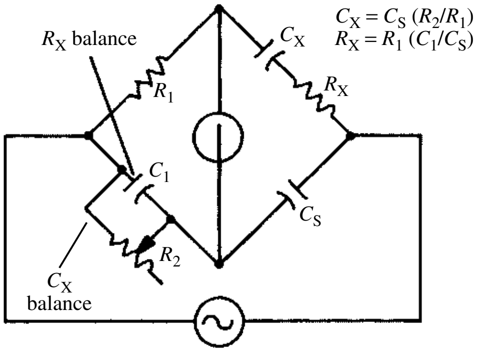

Schering bridge

Schmidt optical system In projection television, a lens system used between the bright-image picture tube and the screen.

Schmitt limiter See SCHMITT TRIGGER.

Schmitt trigger A multivibrator that produces uniform-amplitude output pulses from a variable-amplitude input signal. This circuit has many applications, one being the conversion of a sine wave into a square wave.

Schottky diode A solid-state diode in which a metal and a semiconductor form the pn junction. Electrons injected into the metal have a higher energy level than the charge carriers in a semiconductor, and energy storage at the junction is low because current flow is not accompanied by hole movement. Also known as HOT-CARRIER DIODE.

Schottky logic A form of integrated-injection logic with enhanced operating speed.

Schottky noise The random noise resulting from the emission of charged particles, usually electrons or holes, from an electrode in an amplifying device. This noise is usually of a wideband nature.

scientific language Any computer programming language used primarily for mathematical or scientific applications.

scientific notation The expression of very large and very small numbers as a fixed-point part (mantissa) and a power of the radix (usually 10).

Generally, the mantissa is greater than or equal to 1, but less than 10; the power of 10 is adjusted accordingly. Thus, for example, 203,700 = 2.037×10^5; 0.000533 = 5.33×10^{-4}. See also SIGNIFICANT FIGURES.

scintillating crystal A crystal, such as one of sodium iodide, that sparkles or flashes when exposed to radioactive particles or rays.

scintillation **1.** In radar operations, the apparent rapid displacement of a target from its mean position. **2.** A momentary flash of light produced in a phosphor or scintillating crystal when a high-velocity particle strikes it. **3.** A small fluctuation in radio field intensity at a receiving point.

scintillation counter A radiation counter consisting essentially of a scintillating crystal in combination with a photomultiplier tube. Flashes from the excited crystal cause the tube to deliver output pulses that are totaled and indicated.

scintillator material A substance, such as crystalline sodium iodide, that scintillates under certain stimuli.

scissoring A method of interrupting the electron beam in a cathode-ray tube when the beam would not land on the phosphor screen.

SCLC Abbreviation of SPACE-CHARGE-LIMITED CURRENT.

sco Abbreviation of SUBCARRIER OSCILLATOR.

scope Colloquialism for OSCILLOSCOPE.

Scott oscillator See PARALLEL-TEE OSCILLATOR.

scp Abbreviation of SPHERICAL CANDLEPOWER.

SCR Abbreviation of SILICON-CONTROLLED RECTIFIER.

scrambled signal Any signal in which (for secrecy or exclusivity) the elements are disarranged according to an encryption algorithm. Thus, intelligent reception is possible only if the signal is processed via a suitable decryption algorithm. Example: SCRAMBLED SPEECH.

scrambled speech Voice transmission in which the frequencies have been inverted to prevent eavesdropping. It is automatically unscrambled (by reinversion) at the receiver to restore intelligibility.

scrambler circuit A circuit containing filters and frequency inverters for scrambling speech.

scratch filter An audio-frequency low-pass filter that suppresses high-frequency noise caused by friction between a phonograph disc and the needle.

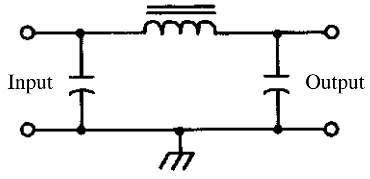

scratch filter

scratchpad memory In computers, a low-capacity memory that stores an intermediate result of a calculation.

scratch tape Magnetic data tape that can be overwritten for any purpose.

screen **1.** See SCREEN GRID. **2.** See SHIELD. **3.** The front surface of a cathode-ray tube. **4.** The surface of a computer or terminal display.

screen angle In radar, the angular difference between the actual horizon and the plane perpendicular to the line connecting the radar set with the center of the earth.

screen current Symbol, I_S or I_{SG}. The current flowing in the screen circuit of an electron tube.

screen grid In a vacuum tube, a grid element between the control grid and plate. It reduces the internal grid-plate capacitance, and consequently prevents self-oscillation when the tube is used in a straight-through amplifier.

screen-grid neutralization Neutralization of an amplifier that uses a tetrode vacuum tube. Such circuits require smaller neutralizing capacitances than those used in triode amplifiers because of the lower interelectrode capacitance of the screen-grid tube.

screen illumination Edge lighting of the transparent screen of an oscilloscope, to make the lines of the graticule more clearly visible.

screen material See PHOSPHOR.

screen resistance Symbol, R_S or R_{SG}. The internal resistance presented by the screen-grid/cathode path of an electron tube.

screen room See CAGE.

screen voltage Symbol, E_S or E_{SG}. The voltage at the screen grid of an electron tube.

scribing The etching of a semiconductor wafer to facilitate breaking the wafer into smaller pieces.

ScriptX A high-level programming language used in the writing of software for personal computers—especially in multimedia. Developed by Kaleida Laboratories.

SCS Abbreviation of SILICON-CONTROLLED SWITCH.

SCT Abbreviation of SURFACE-CHARGE TRANSISTOR.

S curve **1.** The voltage-versus-current curve for a negative-resistance device. Compare N CURVE. **2.** The response curve for a frequency-modulation discriminator or ratio detector.

SD Abbreviation of STANDARD DEVIATION.

S_D band A section of the S BAND extending from 4200 to 5200 MHz.

SDF Abbreviation of static direction finder.

SDR See SOFTWARE-DEFINED RADIO.

Se Symbol for SELENIUM.

seaborgium Symbol, Sg. Also called *unnilhexium* (Unh). Atomic number, 106. The most common isotope has atomic weight 263. Classified as a transition metal. It has a half-life on the order of 1 second or less, is human-made, and is not known to occur in nature.

sea clutter Collectively, the radar echoes that the sea reflects.

seal **1.** The point at which a lead or electrode enters or leaves and is secured to an envelope, case, or housing. Such a point is often tightly closed against the passage of air in or out of the envelope. **2.** To close off a circuit or component from tampering.

sealed dry battery A set of electrochemical dry cells that can be installed without concern for orientation or position. Example: 9-volt "transistor battery."

sealed meter **1.** A meter that is tightly closed against the entry of moisture and foreign materials. **2.** A meter that is locked or otherwise protected against tampering.

sealing compound A substance (such as wax, pitch, or plastic) used to enclose and protect electronic devices.

search **1.** To scan or sweep through a range of quantities or through a region of interest. **2.** To examine (usually in some prescribed order) items of information in a computer memory to find those satisfying a given criterion.

search coil An inductive probe (exploring coil) used to sample magnetic fields.

search oscillator A variable-frequency oscillator used to locate and identify signals by the heterodyne method.

search probe **1.** See SEARCH COIL. **2.** A capacitive probe used to sample electric fields.

search radar A radar that displays a target almost immediately after that target enters a scanned area.

search time The time needed to test items during a search (see SEARCH, **2**).

sea return See SEA CLUTTER.

seasonal effects In ionospheric propagation, the changes produced as a result of the revolution of the earth around the sun. The path of the sun across the sky, and the length of the day, are primarily responsible for such effects.

seasonal static Atmospheric electrical interference, most prevalent during the summer.

S_E band A section of the S BAND extending from 1550 to 1650 MHz.

sec **1.** Abbreviation of SECOND. **2.** Abbreviation of SECTION. (Also, sect.) **3.** Abbreviation of SECONDARY. **4.** Abbreviation of SECANT.

secant Abbreviation, sec. The trigonometric function representing the ratio of the hypotenuse of a right triangle to the adjacent side (c/b). The secant is the reciprocal of the cosine; $\sec x = 1/\cos x$.

second **1.** Abbreviation, s and sec. A unit of time. The mean solar second is 1/86,400 of a mean solar day, and is 1/60 minute or 1/3600 hour. **2.** Symbol ("). A unit of arc measure. 1" = 1/3600 geometric degree.

secondaries See SECONDARY COLORS.

secondary **1.** See SECONDARY WINDING. **2.** See SECONDARY STANDARD.

secondary battery See STORAGE BATTERY.

secondary calibration The calibration of an instrument, based on a reference instrument calibrated against an absolute source.

secondary capacitance **1.** The distributed capacitance of the secondary winding of a transformer whose primary winding is unloaded. Compare PRIMARY CAPACITANCE, **1**. **2.** A series or shunt capacitance used to tune the secondary coil of a radio-frequency transformer. Compare PRIMARY CAPACITANCE, **2**.

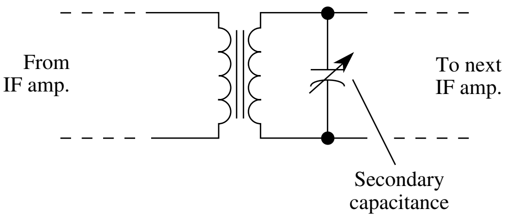

secondary capacitance, 2

secondary cell See STORAGE CELL.

secondary circuit **1.** The circuit associated with the secondary winding of a transformer. **2.** See OUTPUT CIRCUIT.

secondary coil See SECONDARY WINDING.

secondary color **1.** A color prepared by mixing two primary colors. **2.** In television operations, any displayed color composed of two or more color primaries.

secondary current The current flowing in the secondary winding of a transformer. Also called TRANSFORMER OUTPUT CURRENT. Compare PRIMARY CURRENT.

secondary electron **1.** The electron possessing the lesser energy after a collision between two electrons. Compare PRIMARY ELECTRON. **2.** An electron ejected by secondary emission.

secondary emission The action whereby electrons in the atoms at the surface of a target are ejected as a result of bombardment by a beam of (primary) electrons. Thus, in an electron tube, electrons from the cathode strike the plate with a force that drives secondary electrons out of the plate, into the surrounding space.

secondary emitter A source of secondary electrons (e.g., the plate of an electron tube or a dynode in a photomultiplier tube).

secondary failure The failure of a component or circuit, resulting from the failure of some other component. For example, the pass transistor in a power supply might burn out, causing the output voltage to increase; this increased voltage can damage equipment connected to the supply.

secondary frequency standard A device for generating signals of accurate frequency, but that does not possess the very high stability and extreme accuracy of a primary frequency standard. The secondary standard is periodically checked against a PRIMARY FREQUENCY STANDARD and appropriately corrected.

secondary impedance **1.** The impedance of the secondary winding of a transformer whose primary winding is unloaded. Compare PRIMARY IMPEDANCE, **1**. **2.** An external impedance presented to the secondary winding of a transformer. Compare PRIMARY IMPEDANCE, **2**.

secondary inductance The inductance of the secondary winding of a transformer whose primary winding is unloaded. Compare PRIMARY INDUCTANCE.

secondary kVA The kilovolt-amperes in the secondary circuit of an operating transformer. Compare PRIMARY KVA.

secondary power The power in the secondary circuit of a transformer. Also see SECONDARY KVA and SECONDARY VA. Compare PRIMARY POWER.

secondary radiation **1.** The (sometimes random) reradiation of electromagnetic waves, as from a receiving antenna. **2.** Rays emitted by atoms or molecules when the latter are struck by other radiation.

secondary rays Rays emitted by atoms or molecules that have been bombarded by other rays of the same general nature. Examples: *secondary X rays* and *secondary beta rays.*

secondary resistance The direct-current resistance of the secondary winding of a transformer. Compare PRIMARY RESISTANCE.

secondary standard An accurate source of a quantity (capacitance, frequency, inductance, resistance, etc.), that is referred periodically to a primary standard for correction.

secondary storage In computer and data-processing operations, storage that is auxiliary to the main storage. Also called *backing storage.*

secondary turns The number of turns in the secondary winding of a transformer. Compare PRIMARY TURNS.

secondary utilization factor For a transformer in a rectifier circuit, the utility factor of the secondary winding (ratio of direct-current power output to secondary volt-amperes).

secondary VA The volt-amperes in the secondary circuit of a transformer. Compare PRIMARY VA.

secondary voltage The voltage across the secondary winding of a transformer. Also called TRANSFORMER OUTPUT VOLTAGE. Compare PRIMARY VOLTAGE.

secondary winding The normal output winding of a transformer. Also called SECONDARY COIL. Compare PRIMARY WINDING.

second breakdown In a large-area power transistor, a destructive breakdown caused by thermal runaway.

second-breakdown voltage The collector voltage at which second breakdown occurs in a transistor.

second-channel attenuation See SELECTANCE, **2**.

second-channel interference In a given channel, interference arising from authorized signals two channels removed.

second detector In a superheterodyne receiver, the intermediate-frequency detector. Compare FIRST DETECTOR.

second-level address In a computer program instruction, an address giving the location of the address of a required operand. Also called *indirect address.*

sect Abbreviation of SECTION. (Also, sec.)

section **1.** A subcircuit or stage of a larger circuit (e.g., the *oscillator section* of a receiver). **2.** The smaller unit described in **1**, when self-contained and operated independently (e.g., *filter section*). **3.** See PROGRAM SEGMENT.

sectionalized antenna A set of collinear radiating elements, placed end-to-end with reactances between them, for the purpose of modifying the radiation pattern.

sectionalized winding **1.** A method of winding a coil in complete, multilayer sections that are stacked side by side or top to bottom, a technique that reduces distributed capacitance. **2.** A coil wound, as defined in **1**.

sector On a magnetic disk, a specific portion of a track.

sectoral horn antenna An (usually sheetmetal) antenna with the shape of a horn of rectangular cross section. It is flared in one dimension only.

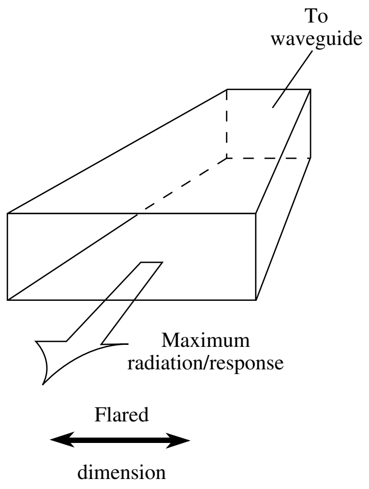

sectoral horn antenna

sector display In a radar, a display that allows the continuous observation of a portion of the scanned area.

secular equilibrium The state in which a radioactive substance changing into another substance is decaying as fast as the second substance is being formed.

secular variation A slow change in the intensity of the terrestrial magnetic field.

secure mode In a security system, the condition of being fully activated.

securite In radiotelephony, a spoken word (pronounced say-kyoor-ee-tay) identifying a transmission concerning safety. Equivalent to TT in radiotelegraphy.

security code **1.** A set of alphanumeric characters (letters and/or numbers) or switch settings that activates or deactivates a security system. **2.** See PASSWORD.

security robot In a smart home, a robot that assists in protection of a home or business, and its occupants, from intrusion, burglary, or attack. If an intruder enters the property, the security robot might drive the offender away or detain the offender until police arrive. The robot might employ iris-print detection or facial feature detection to identify an intruder from a distance. The robot would be linked to the home computer via a broad-bandwidth, high-speed wireless system. Compare SENTRY ROBOT. Also see SMART HOME OR BUSINESS.

security system A set of electronic (and sometimes also electromechanical) devices designed to do one or more of the following: restrict access to a premises or computer system, detect abnormal conditions, detect unauthorized entry, alert human operators of abnormal conditions, alert owners and/or authorities of unauthorized entry to a premises, and (in some cases) provide physical protection of property.

Seebeck effect The development of an electromotive force in a junction of two dissimilar metals (a thermocouple) when the temperature of the junction is different from that of the rest of the metal.

Seebeck emf The electromotive force resulting from the Seebeck effect.

seed crystal A small single crystal from which a larger single crystal (e.g., germanium or silicon) is grown. Also see CZOCHRALSKI METHOD.

seeing-eye robot A robot that serves in the capacity of a seeing-eye dog, to help visually impaired people in their daily activities.

seek See SEARCH.

seek area An area of direct-access storage to which are assigned specific records and from which the records can be accessed quickly.

segment **1.** The portion of a line or curve lying between two points. **2.** See PROGRAM SEGMENT.

segmental meter An expanded-scale meter (see SCALE EXPANSION).

segmented encoding law An approximation of a smooth encoding law, done by means of linear partitions or segments. The greater the number of segments, the more accurate the approximation.

segmenting See PARTITIONING.

segment mark A character indicating the division between tape file sections.

segue (Pronounced SEG-way.) A smooth transition from one sound or channel to another (e.g., between musical selections in a radio broadcast). The first selection decreases in volume as the second selection increases. There is some overlap.

Seignette salt See ROCHELLE SALT.

seismic detector A vibration sensor used in some types of intrusion-detection systems.

seismogram A record produced by a SEISMOGRAPH.

seismograph An instrument for detecting and recording earth tremors. Indicates the direction, magnitude, and time of a quake.

seismometer See SEISMOGRAPH.

seismoscope An instrument that shows the occurrence and time of an earthquake. Compare SEISMOGRAPH.

select To accept or separate a unit, quantity, or course of action from all those available (in a group, mixture, or series).

selectance **1.** For a resonant circuit, the ratio Er/Ex, where Er is the voltage at resonance, and Ex is the voltage at a specified nonresonant frequency. **2.** For a receiver, the figure $S2/S1$, where $S1$ is the sensitivity of the receiver in a given frequency channel, and $S2$ is the sensitivity in another specified channel.

selected mode In an encoder, a mode in which one output is read and the others are ignored.

selective Pertaining to a device or system, such as a radio receiver, with a high degree of SELECTIVITY.

selective absorption The attenuation or absorption of some frequencies or bands of frequencies, with little or no attenuation at other frequencies or bands of frequencies.

selective amplifier An amplifier that can be tuned, with the desired degree of sharpness, to a single frequency or band of frequencies. Radio-frequency amplifiers are generally tuned by means of inductance-capacitance (LC) circuits; audio-frequency amplifiers are usually tuned by means of resistance-capacitance (RC) circuits.

selective calling The calling, alerting, or alarming of a desired station without interfering with other stations.

selective compliance assembly robot arm Abbreviation, SCARA. An electromechanical device designed especially for use in assembly lines. It uses cylindrical coordinate geometry to allow precise, programmable movements in three dimensions.

selective dump In computer operations, a dump (see DUMP) affecting a small, specific memory area.

selective fading Fading caused by propagation conditions, whose effects differ at slightly different frequencies. In an amplitude-modulated signal, this effect causes the sidebands and carrier to arrive in various phase relationships, with resulting distortion in the received signal.

selective interference Interference confined to a narrow band of frequencies.

selective polarization See POLARIZATION SELECTIVITY.

selective reflection In the reflection of electrons directed into a crystal by means of an electron gun, the tendency of the electrons to be reflected more readily when they strike the crystal at certain speeds.

selective relay **1.** A relay or relay circuit tuned to open or close at one signal frequency. **2.** A relay or relay circuit adjusted to open or close at one value of current or voltage.

selective trace In computer operations, a diagnostic program used to analyze certain areas of memory or specific kinds of program instructions, for debugging purposes.

selectivity **1.** The ability of a circuit or device to pass signals of one frequency and reject signals at other frequencies. **2.** The degree to which a circuit or device passes signals of one frequency and rejects signals at other frequencies.

selectivity control In some equipment (such as receivers, crystal filters, wave analyzers, and vibration meters) an adjustment that permits variation of selectivity.

Selectoject A fully electronic, continuously tunable, notcher-peaker that is resistance-capacitance tuned. The name is an acronym for *select or reject.*

selector **1.** A channel switch in a radio or television receiver. **2.** See SELECTOR SWITCH.

selector channel In data-processing and computer systems, a data transmission channel controlling the information flow between peripherals and a central processing unit.

selector pulse In digital communications, an identifying pulse that represents a certain group of bits or data.

selector relay A device, such as a stepping switch, that actuates one of a number of available circuits on receipt of a predetermined number of pulses.

selector switch A (usually rotary) multiposition switch that allows an operator to select from among several options (such as frequency channels, frequency bands, or selective filters).

selenium Symbol, Se. A nonmetallic element. Atomic number, 34. Atomic weight, 78.96. It is used in the manufacture of some diodes, rectifiers, and photocells.

selenium cell A photoelectric cell that uses specially processed selenium as the light-sensitive material. It can be operated as a photoconductive cell or a photovoltaic cell.

selenium diode A junction diode in which the semiconductor material is specially processed selenium. Also see JUNCTION DIODE.

selenium photocell See SELENIUM CELL.

selenium rectifier A disk- or plate-type power rectifier utilizing the junction between selenium and aluminum or selenium and iron.

self-bias For a transistor or vacuum tube, input-electrode bias voltage resulting from the flow of output-electrode current through a resistor common to both circuits. Also called AUTOMATIC BIAS.

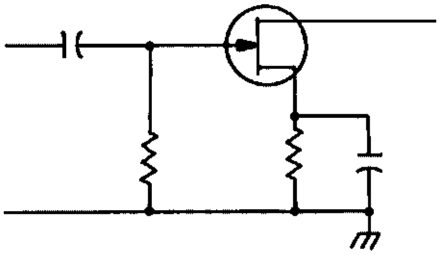

self-bias

self-capacitance The inherent internal capacitance of a device other than a capacitor.

self-checking number A number whose digits have a value that determines the check digit attached to it; thus, it can be verified following its transfer between storage locations or peripherals.

self-cleaning contacts Switch or relay contacts that clean themselves automatically by means of wiping action.

self-contained device A device containing all parts and sections (e.g., main circuit, power supply, meter, loudspeaker, etc.) needed for full operation (i.e., no auxiliary equipment is needed).

self-controlled oscillator See SELF-EXCITED OSCILLATOR.

self discharge The tendency of an electrochemical cell or battery to gradually lose stored energy when not in use.

self-discharge rate A quantitative expression of the speed with which SELF DISCHARGE occurs in an electrochemical cell or battery when it is stored without being used.

self-energy Symbol, *E.* The energy mc^2, in joules, of a particle traveling at the speed of light c (2.9979×10^8 meters per second) and whose mass is m kilograms.

self erasing In a magnetic tape, the unwanted erasing of data near a highly magnetized region.

self-excited generator An alternating-current generator in which the field coils are supplied with direct current produced by the machine itself. Compare SEPARATELY EXCITED GENERATOR.

self-excited oscillator An oscillator consisting of an amplifier that supplies its own input signal through positive feedback, and whose oscillation frequency depends entirely on circuit constants, such as the capacitance and inductance in a tank circuit. Compare CRYSTAL OSCILLATOR, FORK OSCILLATOR, HUMMER, and MAGNETOSTRICTION OSCILLATOR.

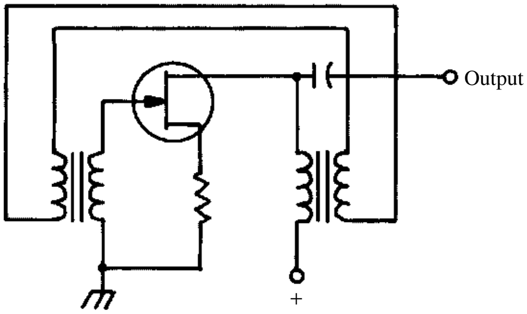

self-excited oscillator

self-focus picture tube A television picture tube in which the electron gun has an automatic, electrostatic focusing arrangement.

self-generating photocell See PHOTOVOLTAIC CELL.

self-generating transducer A voltage-producing transducer, such as a piezoelectric pickup or dynamic microphone.

self-healing capacitor A capacitor, such as a wet electrolytic unit, in which the dielectric is restored to its normal condition after a high-voltage breakdown.

self-heated thermocouple A thermocouple in which the passage of an applied current produces the heat necessary for the activation of the thermocouple.

self-heating thermistor A thermistor heated to above ambient temperature by the current passing through it. Also called DIRECTLY HEATED THERMISTOR. Compare INDIRECTLY HEATED THERMISTOR.

self impedance The effective or measured impedance at a circuit point.

self-inductance **1.** The inductance of an inductor. **2.** The inherent internal inductance of a device other than an inductor.

self-induction Induction that occurs in a single circuit. An instance is the generation of an opposing voltage across a coil by an alternating current flowing through it. Compare INDUCTION. Also see INDUCTANCE.

self-latching relay A relay that remains in the state that it has been switched (i.e., locked open or closed) until a subsequent signal is received.

self-modulated oscillator A circuit, such as a blocking oscillator, in which oscillation occurs simultaneously at two frequencies, one modulating the other.

self-organizing Pertaining to a computer or system that can change the arrangement of data files for particular purposes.

self-powered device A device that requires no external source of power (i.e., it is equipped with a self-contained battery or a generator).

self-pulsing blocking oscillator A blocking oscillator that produces a train of radio-frequency pulses. Compare SINGLE-SWING BLOCKING OSCILLATOR.

self-quenching detector A super-regenerative detector (see SUPER-REGENERATIVE CIRCUIT) in which the low-frequency quenching voltage is supplied by the regenerative detector itself. Also see QUENCHING ACTION and QUENCH OSCILLATOR. Compare SEPARATELY QUENCHED DETECTOR.

self-quenching oscillator A circuit, such as a blocking oscillator, in which oscillation is periodically switched off automatically, resulting in a self-interrupted wave train.

self-rectifying vibrator A vibrator-type power supply in which one vibrator reed chops the direct-current input to the primary winding of the transformer, and a second vibrator reed rectifies the alternating-current output delivered by the secondary winding. Also see VIBRATOR RECTIFIER.

self-rectifying X-ray tube An X-ray tube operated with alternating-current anode voltage.

self-regulation The ability of a circuit or device to control its output automatically, according to some predetermined plan, by using output error to correct operation or to vary the input.

self reset **1.** The action of a circuit breaker to reapply power after a certain elapsed time. **2.** The action of any device, returning a circuit or system to normal automatically.

self-resetting loop In a computer program, a loop in which instructions cause locations used in the loop to assume their condition prior to the loop's execution.

self-resistance The inherent internal resistance of a device other than a resistor.

self-resonant frequency The frequency at which a device will resonate naturally (without external tuning). Thus, an inductor will self-resonate with its distributed capacitance; similarly, a capacitor will resonate with its stray inductance.

self-saturation In a magnetic amplifier, saturation resulting from rectification of the saturable-reactor output current.

self-starting motor An alternating-current motor that starts running as soon as voltage is applied (i.e., no external mechanical force is needed). Also see SHADING COIL.

self-sustained oscillations Oscillations maintained by means of positive feedback (inductive or capacitive) from the output to the input of a circuit. See, for example, SELF-EXCITED OSCILLATOR.

self-test Any arrangement whereby a device or system determines, without the aid of an external operator, whether or not it is operating correctly.

self-ventilated motor See OPEN MOTOR.

self-wiping contacts See SELF-CLEANING CONTACTS.

selsyn See AUTOSYN and SYNCHRO.

SEM **1.** Abbreviation of SINGLE-ELECTRON MEMORY. **2.** Abbreviation of SCANNING ELECTRON MICROSCOPE.

semantic network A reasoning scheme sometimes used in artificial intelligence. Logical state-ments or sentences are broken down into nodes (generally nouns) and relationships (generally verbs and modifiers). This allows statements to be mapped in a way that is easy for computers to store and modify.

semiautomatic key Also called *bug.* A telegraph key that mechanically produces a string of Morse-code dots (short pulses) when its lever is pressed to one side (usually toward the right), and continuous circuit closure when the lever is pressed to the opposite side (usually toward the left). Dashes are manually sent by the operator.

semiconductor A material whose natural resistivity lies between that of conductors and insulators (e.g., GERMANIUM, SILICON, SELENIUM, and GALLIUM ARSENIDE).

semiconductor capacitor A miniaturized component that takes advantage of reverse biasing in a semiconductor P-N junction. When a voltage source is connected across a diode so that it does not conduct, the diode acts as a capacitor. The capacitance varies depending on how much reverse voltage is applied to the diode. The greater the reverse voltage, the smaller the capacitance. Some diodes are especially manufactured to serve this function. This phenomenon can be useful in the fabrication of integrated circuits (ICs). See also INTEGRATED CIRCUIT, VARACTOR.

semiconductor counter A device for measuring the intensity of ionizing radiation (such as alpha particles, beta particles, or gamma rays) using a photodiode and sensing circuit.

semiconductor device A component (such as a diode, photocell, rectifier, or transistor) that exploits the properties of a semiconductor.

semiconductor diode A solid-state diode, as opposed to a vacuum-tube diode or gas-tube diode. Examples: *germanium diode*, *selenium diode*, and *silicon diode.*

semiconductor junction Within a body of semiconductor material, the area of physical contact between two regions (usually n and p) having opposite electrical properties.

semiconductor laser See LASER DIODE.

semiconductor material See SEMICONDUCTOR.

semiconductor-metal junction The area of physical contact between a metal and a semiconductor.

semiconductor photosensor A semiconductor photodiode or phototransistor, as opposed to a phototube.

semiconductor rectifier A heavy-duty semiconductor diode (or assembly of such diodes) designed primarily to change alternating current to direct current in power-supply units. Rectifiers commonly are made from copper oxide, germanium, magnesium-copper sulfide, selenium, or silicon. Also see JUNCTION DIODE and METER RECTIFIER.

semiconductor resistor A tiny resistor manufactured from semiconductor material, especially one that is etched onto the chip of an integrated circuit (IC). The thickness, and the types and concentrations of impurities added, determine the resistance of the component. Such resistors can handle only a tiny amount of power because of their small size. But because IC circuits are designed to consume minimal power, this is not a problem. The small signals produced by ICs can be amplified, using circuits made from discrete components, if it is necessary to obtain higher signal power. See also INTEGRATED CIRCUIT.

semidirectional Pertaining to a transducer that exhibits different directional characteristics at different frequencies.

semiduplex operation A two-frequency communication system that operates in duplex at one end of the link, and in simplex at the other end. Also see DUPLEX OPERATION and SIMPLEX TELEGRAPHY.

semilogarithmic graph Also called *semilog graph.* A graph in which one axis is logarithmic and the other axis is linear.

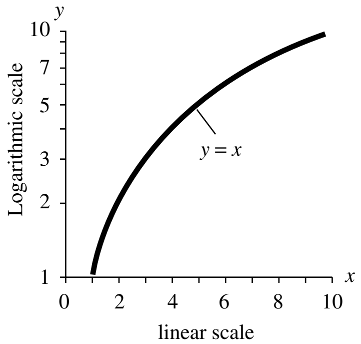

semilogarithmic graph

semimetal An elemental substance that exhibits some, but not all, of the properties of a metal (e.g., antimony and arsenic). Also called METALLOID.

semiresonant line An open-wire transmission line cut approximately to resonant length at the frequency of operation.

semitone See HALF STEP.

sender See TRANSMITTER, **1**, **3**.

sending-end impedance See DRIVING-POINT IMPEDANCE.

sending set **1.** See RADIO TRANSMITTER. **2.** An equipment for transmitting electromagnetic waves. Also see TRANSMITTER, **1**.

sensation level The level of sound that produces a tingling or noticeable sensation in the ear.

sense **1.** To check the condition of a switching device, such as a gate. **2.** See READ.

sense amplifier A device that produces a control signal when some characteristic of the input signal changes.

sense determination In a direction finder that provides an ambiguous indication (two readings 180 degrees apart), the process of deducing, or the ability of the apparatus to deduce, the true direction from which the signal is coming.

sense resistor A (usually low-value) resistor used to sense current in a circuit without introducing a significant loss. The voltage drop across this resistor is proportional to the current and can be applied to a voltmeter, oscilloscope, or other instrument for measurement or observation.

sensing circuit **1.** A circuit that samples a quantity. **2.** In a voltage regulator, the circuit that monitors the output voltage and delivers a control voltage proportional to the output-voltage error.

sensing window See WINDOW, **2**.

sensitive communications Two-way communications of an emergency or priority nature, or involving the security of government operations.

sensitive device A device that responds to a signal of low amplitude.

sensitivity **1.** The ability of a circuit or device to respond to a low-level applied stimulus. **2.** For a receiver, the input-signal (in microvolts or millivolts) required for a specified output level. **3.** For a galvanometer, microamperes or milliamperes per scale division. **4.** The ohms-per-volt rating of a voltmeter. Also see VOLTMETER SENSITIVITY.

sensitivity adjustment **1.** An input gain control in an amplifier circuit. **2.** The radio-frequency gain control of a receiver. **3.** A control or switch that is used to select the range or threshold of a piece of test equipment.

sensitivity control A manual or automatic device for adjusting the sensitivity of a circuit or device.

sensitometer An instrument used to measure the sensitivity of certain materials to light.

sensor **1.** A device that samples a phenomenon, and delivers a proportionate current or voltage in terms of which the intensity of the phenomenon can be measured, or with which control action can be initiated. **2.** An electronic device that detects abnormal conditions (e.g., smoke and heat) and delivers a warning signal to human operators and/or computers. **3.** An electronic device that detects intrusion to a premises and delivers a warning signal to human operators and/or computers.

sentry robot In a smart home or business, a robot that alerts the owner to abnormal conditions. It can detect fire, burglars, or water in places it should not be. It might also detect abnormal temperature, barometric pressure, wind speed, humidity, or air pollution. A wireless link alerts the owner via a device similar to a common beeper. Compare SECURITY ROBOT. Also see SMART HOME OR BUSINESS.

separately excited generator An alternating-current generator whose field coils are supplied with direct current from another generator or from a battery. Compare SELF-EXCITED GENERATOR.

separately quenched detector A superregenerative detector (see SUPERREGENERATIVE CIRCUIT) in which the quenching voltage is supplied by a separate low-frequency oscillator. Also see QUENCHING ACTION and QUENCH OSCILLATOR. Compare SELF-QUENCHED OSCILLATOR.

separation See CHANNEL SEPARATION.

separation energy The energy required to remove a proton or neutron from the nucleus of an atom. The separation energy depends on the atomic number.

separator **1.** See FILTER, **1**. **2.** A perforated or porous plate of insulating material (usually plastic or wood) for holding active plates apart in a storage cell. **3.** In computer operations, a character marking the division between logical data units. Also called *data delimiter.*

septate cavity A coaxial cavity containing a SEPTUM between the inner and outer conductors.

septate waveguide A waveguide containing one or more septa (see SEPTUM) to control power transmission.

septum A thin metal vane used as a reflector in a waveguide or cavity.

sequence **1.** A succession of objects, parameters, or numbers. **2.** An ordered set of numbers—each of which is related to its predecessor by a specific mathematical function.

sequence checking routine In computer operations, a routine that verifies the order of items of data.

sequence control register In a computer memory, a register whose contents determine the instruction to be implemented next.

sequence programmer A timing device that can be preset to start or stop various operations at predetermined times.

sequencer A device that initiates or terminates events in a desired sequence.

sequence relay A relay whose several contacts close in a predetermined order.

sequence timer A timer in which separate delay circuits are actuated in a predetermined sequence.

sequential In computer operations, a term denoting operations on data items in which the items (e.g., records in a file) are taken in an order determined by key values, rather than in the order in which the items are physically arranged.

sequential access memory Any semiconductor memory in which data can be recalled or addressed only in a certain specified order.

sequential analysis In statistics, using an unspecified number of observations as samples from which a result is derived. Each observation is accepted or rejected, or another observation is made.

sequential color television The successive transmission of the three primary colors in a television system, and their reproduction at the receiver in the same order. Also see DOT-SEQUENTIAL SYSTEM, FIELD-SEQUENTIAL SYSTEM, and LINE-SEQUENTIAL SYSTEM.

sequential control Computer operation in which the order of instruction implementation is the same as the order of instruction storage.

sequential relay See SEQUENCE RELAY.

sequential scanning Rectilinear television scanning in which the center-to-center distance between successive lines is the same as the nominal line width.

sequential switch **1.** A switch that provides selection of two or more ports in a rotating succession. **2.** In a television system, a switch that allows the monitoring technician to select any of the cameras for viewing.

sequential timer See SEQUENCE TIMER.

ser **1.** Abbreviation of SERIES. **2.** Abbreviation of SERIAL.

serial **1.** Pertaining to the performance of steps, or the occurrence of elements (such as data items on magnetic tape), in succession. **2.** An order, row, or sequence in which one item follows another (as opposed to parallel).

serial access Access to data file records in their order in a storage medium.

serial adder See SERIAL ARITHMETIC UNIT.

serial arithmetic unit In computer operations, an arithmetic unit in which digits are handled in order. Compare PARALLEL ADDER.

serial bit Data in which the bits of each byte or word are sent or received one at a time.

serial memory A register in which the input and output data is stored and retrieved one bit at a time.

serial-parallel **1.** Pertaining to data transfer that is serial in one sense and parallel in another sense. For example, entire words might be serially transmitted within a system, but the constituent bits of each word might be transferred in parallel. **2.** See PARALLEL-SERIES.

serial processing In computer operations, the sequential processing of several different programs through a single channel. Compare PARALLEL PROCESSING.

serial storage In computer operation, storage in which elements are entered in order and are available only in the same order. Compare PARALLEL STORAGE.

serial transfer The propagation of information along a single path, in which data bits are sent one after the other.

series **1.** The sum of a mathematical sequence (see SEQUENCE, **2**). **2.** Pertaining to the connection of elements or components end-to-end (see SERIES CIRCUIT).

series addition See SERIES-AIDING.

series-aiding The condition in which two series voltages or magnetic fields are added together. Compare SERIES-BUCKING.

series antenna tuning Antenna-feeder tuning in which a separate tuning capacitor is connected in series with each wire. Compare PARALLEL ANTENNA TUNING.

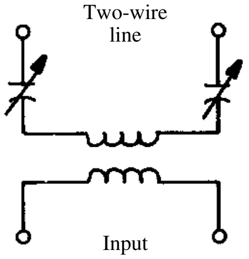

series antenna tuning

series bucking The condition in which two series voltages or magnetic fields oppose each other. Compare SERIES-AIDING.

series capacitance Capacitance acting, or connected, in series with another capacitance or other quantity.

series capacitors Capacitors connected in series. If the individual capacitors have values $C1$, $C2$, . . . , Cn, then the total capacitance Ct is equal to $1/(1/C1 + 1/C2 + \dots + 1/Cn)$. Also see SERIES CIRCUIT.

series circuit A circuit whose components are, in effect, connected in a string (i.e., end-to-end). Compare PARALLEL CIRCUIT.

series-diode half-wave rectifier See SERIES-DIODE RECTIFIER.

series-diode rectifier A rectifier circuit in which the diode is connected in series with the source and load. Compare SHUNT-DIODE RECTIFIER.

series dropping resistor See DROPPING RESISTOR.

series equivalent impedance A series impedance that will draw the same current (magnitude and phase) drawn by a given parallel circuit connected across the same single-phase source.

series equivalent of parallel circuit See SERIES EQUIVALENT IMPEDANCE.

series-fed amplifier An amplifier circuit in which the operating voltages are applied in series with the alternating-current signal voltages. Also see SERIES FEED.

series-fed oscillator An oscillator circuit in which the direct-current operating voltage is applied in series with the alternating-current output voltage. Also see SERIES FEED.

series feed The application of alternating-current (ac) and direct-current (dc) voltages in series to a device. Example: the presentation of the dc operating voltages for an amplifier in series with the ac signal voltages (see SERIES-FED AMPLIFIER).

series feedback A feedback system in which the feedback signal is presented to the input point in series with the input signal. Compare SHUNT FEEDBACK.

series field A magnetic field produced by a series winding in a motor or generator.

series generator An electric generator in which the armature and field windings are connected in series. Compare SHUNT GENERATOR.

series inductance **1.** Inductance acting, effectively, in series with some other quantity (e.g., the inherent inductance of a wirewound resistor). **2.** An inductance connected in series with other inductances, or with some other quantity (e.g., capacitance and resistance).

series inductors Inductors connected in series, and separated or oriented in a way that minimizes the effects of mutual inductance. Assuming zero mutual inductance, if the individual inductors have values $L1$, $L2$, . . . , Ln, then the total inductance Lt is equal to $L1 + L2 + \ldots + Ln$. Also see SERIES CIRCUIT.

series limiter A limiter (clipper) circuit in which the diode is essentially in series with the signal. Compare PARALLEL LIMITER.

series loading The series insertion of reactances in a circuit for the purpose of impedance matching.

series magnetic circuits A combination of several magnetic paths in line so that flux extends through each path in sequence; this is analogous to the passage of electric current successively through series-connected resistors.

series motor An electric motor whose armature and field windings are connected in series. Compare SHUNT MOTOR.

series operation The operation of units in succession, necessitating sequential current flow through each. Also see SERIES CIRCUIT.

series opposition See SERIES BUCKING.

series-parallel See PARALLEL-SERIES.

series-parallel capacitors See PARALLEL-SERIES CAPACITORS.

series-parallel inductors See PARALLEL-SERIES INDUCTORS.

series-parallel resistors See PARALLEL-SERIES RESISTORS.

series regulator A voltage regulator circuit in which the controlled device is in series with the load. Compare SHUNT REGULATOR.

series resistance **1.** Resistance acting in series with another resistance or with another quantity (e.g., capacitance and inductance). **2.** The inherent resistance that acts effectively in series with the plates of a capacitor. **3.** The resistance of the wire in a coil, acting effectively in series with the inductance.

series resistors Resistors connected in series with each other. If the individual resistors have values $R1$, $R2$, . . . , Rn, then the total resistance Rt is equal to $R1 + R2 + \ldots + Rn$. Also see SERIES CIRCUIT.

series resonance Resonance in a circuit consisting of a capacitor, inductor, and an alternating-current generator in series. At the resonant frequency, the inductive reactance and the capacitive reactance cancel, so the net reactance is zero. The capacitor current and inductor current are maximum and equal, and the circuit impedance is minimum. Compare PARALLEL RESONANCE.

series-resonant circuit A resonant circuit in which the capacitor, inductor, and generator are connected in series. Also see SERIES RESONANCE. Compare PARALLEL-RESONANT CIRCUIT.

series-resonant trap A wavetrap consisting of a series-resonant inductance-capacitance (LC) circuit. Compare PARALLEL-RESONANT TRAP.

series-resonant wavetrap See SERIES-RESONANT TRAP.

series-shunt circuit See PARALLEL-SERIES.

series tee junction See E-PLANE TEE JUNCTION.

series tracking capacitor See OSCILLATOR PADDER.

series-type frequency multiplier A varactor frequency-multiplier circuit in which the varactor is in series with the input and output. Compare SHUNT-TYPE FREQUENCY MULTIPLIER.

series-type resonance bridge A resonance bridge in which the impedance arm is a series-resonant circuit. Compare SHUNT-TYPE RESONANCE BRIDGE.

series winding **1.** In a motor or generator, a winding connected in series with the armature. **2.** A method of motor or generator construction in which the field winding is connected in series with the armature.

series-wound generator See SERIES GENERATOR.

series-wound motor See SERIES MOTOR.

serrated pulse A pulse having a notched or slotted top. An example is the vertical sync pulse in television.

serrated rotor plate In a variable capacitor, an external rotor plate that is slotted radially. This alters the capacitance-variation curve of the capacitor to allow alignment of sensitive apparatus (e.g., the tracking of radio-frequency tuned circuits in a radio receiver).

serrated vertical sync pulse In television, the vertical sync pulse notched at twice the horizontal sweep frequency.

service **1.** To maintain or repair electronic equipment. **2.** To provide maintenance or repair of electronic equipment.

serviceability ratio For a device or system, the ratio *ts*/(*ts* + *td*), where *ts* is serviceable (operational) time, and *td* is downtime (non-operational time).

serviceable time The cumulative time during which an operator-monitored (but not necessarily operated) device or system is capable of normal operation.

service area For a broadcast or communications station, the useful coverage area.

service band **1.** For a communications system, the band of frequencies in which operation is normally carried out. **2.** A band of frequencies specifically assigned, by government regulation, to a certain communications service or services.

service channel The band of frequencies that a particular broadcast or communications station occupies, when the carrier frequency is held constant.

service charge The amount charged by a technician for installation, maintenance, or repair of equipment. It is often performed on a per-hour basis.

service maintenance For a cell or battery, the relative amount of energy capacity (percentage of full-charge capacity) available at a given time, or after a certain length of time in normal use.

service meter **1.** An energy ("power") meter. Also see KILOWATT-HOUR METER. **2.** A rugged multimeter used by a service technician.

service oscillator A signal generator designed expressly for troubleshooting and repair service.

service switch **1.** The main switch controlling the electric service to a building or other place of installation. **2.** In television repair, a switch on the rear of a chassis. The switch facilitates adjustment of screen controls by removing vertical deflection temporarily.

service-type instrument An instrument having reasonable accuracy and a degree of ruggedness so that it is suitable for field or shop use. Examples: SERVICE METER and SERVICE OSCILLATOR. Compare LABORATORY-GRADE INSTRUMENT.

servo amplifier A highly stable amplifier designed expressly for use in a SERVOMECHANISM.

servo loop In a control system (particularly a servo amplifier), the output-to-input feedback loop, through which automatic control is performed.

servomechanism Also called *servo.* A self-correcting, closed-loop control system. It usually uses an electromechanical device, such as a motor, that controls some electronic device. Error signals, supplied by the controlled electronic device, cause the motor to run in such a way as to optimize or stabilize the system.

servomotor A motor operated by the output signal of a servo amplifier. Depending on the end application of the servo system, the motor signal might or might not be corrected.

servo oscillation In a servo system, a back-and-forth movement or fluctuation, relative to the optimum setting or position. It Results from improper system adjustment. Sometimes the system stabilizes at the optimum after a short period of oscillation; in some cases, the oscillation continues indefinitely.

servo robot A (usually industrial) robot whose motion sequence is programmed into a computer. The robot follows the instructions given by the computer, and makes precise, timed movements on that basis. Different computer programs allow different motion sequences, so a single robot can be used for various tasks.

servo system An automatic control system using one or more servomechanisms.

set **1.** A piece of equipment or a system (e.g., *radio set*). **2.** In a flip-flop circuit, an input that is not controlled by the clock. **3.** To adjust a circuit or device, such as a flip-flop, to a desired operating point or condition. **4.** A class of numbers, things, or events. **5.** In computer programming, to initialize a variable (i.e., to assign a label to a location).

set analyzer A combination test instrument designed originally for troubleshooting radio receivers. It consists of a multimeter and transistor tester or vacuum-tube tester.

set noise Electrical noise arising inside a radio or television receiver, as opposed to that picked up from the outside.

set pulse A pulse used to adjust a device to a certain state (see SET, **3**).

set terminal In a flip-flop, the one-input terminal. Compare RESET TERMINAL.

setting The position or value to which an adjustable device is set for a particular purpose.

settling time **1.** In a digital voltmeter, the time required between the application of a test voltage and the final display of an accurate readout. **2.** In a digital-to-analog converter, the time between half of the level change over all inputs and the arrival of the output to a level within a certain tolerance of its specified final level. It is defined for either full-scale to zero or zero to full-scale.

set up To arrange and prepare equipment for operation.

setup See SET, **1**.

set-up time **1.** The time required to install and test an electronic system, and to ready the system for operation. **2.** In a digital gate, the length of time that a pulse must be held to produce a change of state.

sexadecimal number system See HEXADECIMAL NUMBER SYSTEM.

sexagesimal number system A number system whose radix is 60.

SF **1.** Abbreviation of SAFETY FACTOR. **2.** Abbreviation of SINGLE FREQUENCY. **3.** Abbreviation of STANDARD FREQUENCY. **4.** Abbreviation of STABILITY FACTOR.

SFA Abbreviation of SINGLE-FREQUENCY AMPLIFIER.

S_F band A section of the S BAND extending from 1650 to 1850 MHz.

sferics In wireless broadcast or communications reception, random electromagnetic noise generated by the earth's atmosphere. Some of this noise is thermal in origin; some originates in thunderstorms. Sometimes this noise is called *static.*

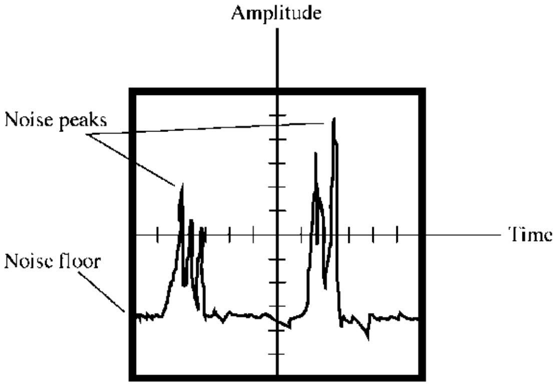

sferics

SFO **1.** Abbreviation of SINGLE-FREQUENCY OSCILLATOR. **2.** Abbreviation of STANDARD-FREQUENCY OSCILLATOR.

SFR Abbreviation of SINGLE-FREQUENCY RECEIVER.

SFR-Chireix-Mesny antenna See CHIREIX-MESNY ANTENNA.

SG Abbreviation of SCREEN GRID.

S_G band A section of the S BAND extending from 2700 to 2900 MHz.

SGCS Abbreviation of *silicon gate-controlled switch* (see SILICON-CONTROLLED SWITCH).

shaded-pole motor An induction-type alternating-current motor using shading coils on the field poles for self-starting with a single-phase supply.

shading Electronic enhancement of a television picture, resulting in a different brightness over various portions of the background, as compared with the actual situation. It can be used, for example, to make certain subjects stand out from the background.

shading coil **1.** A single, short-circuited turn (copper ring) encircling the tip of the core of a coil that carries alternating current (ac), such as the field pole of a motor. Current induced in the coil causes a momentary flux shift that approximates a rotating field that self-starts a simple single-phase induction motor. **2.** A coil used in a simple ac relay to prevent chatter.

shading ring See SHADING COIL.

shading signal In a television camera, a signal that increases the gain of the amplifier while the beam scans a dark part of the image.

shadow area A region in which signal attenuation or the absence of a signal results from the SHADOW EFFECT.

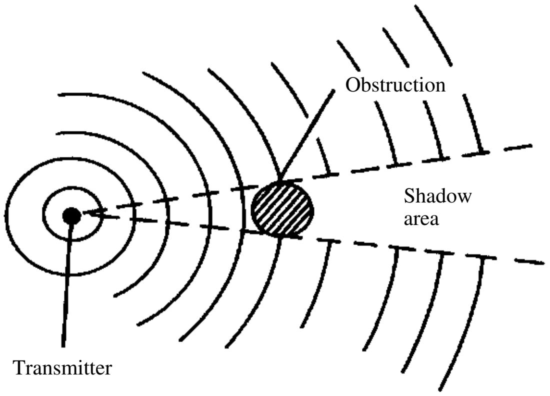

shadow area

shadow attenuation **1.** The attenuation of electromagnetic energy caused by an obstacle. It is generally measured in decibels. **2.** The attenuation of electromagnetic energy produced by the curvature of the earth.

shadow effect The obstruction of radio waves by objects in their path.

shadow mask See APERTURE MASK.

shadow meter See SHADOW TUNING INDICATOR.

shadow region See SHADOW AREA.

shadow tuning indicator A tuning meter in which the indicating medium is a shadow whose width is proportional to meter current.

shaft The rodlike part of an adjustable component (such as a potentiometer or variable capacitor) to which a rotating (or turning) member is attached.

shaft-angle encoder An electronic system for converting shaft rotation into direct binary or decimal readings.

shaft lock A device for fastening the shaft of an adjustable component (such as a potentiometer, rotary switch, or variable capacitor) in position at a particular setting.

shaft-position encoder See SHAFT-ANGLE ENCODER.

shaft-position indicator A device that delivers an analog or digital output signal proportional to the arc of rotation of a shaft.

shaker See VIBRATOR, **2**.

shake table A platform, actuated by a vibrator, on which components can be mounted for a vibration test.

shallow-diffused junction A pn junction made by diffusing the impurity material for a short distance into the semiconductor wafer. Compare DEEP-DIFFUSED JUNCTION.

shape factor **1.** For a tuned circuit, the ratio of the 60-dB bandwidth to the 6-dB bandwidth. **2.** For a filter, the ratio of bandwidth at high attenuation to that at low attenuation.

shaping network A combination of components for changing the natural response of a circuit to a desired response (i.e., a curve-changing circuit).

shared file A data file that is available for use by more than one system simultaneously.

shared files system A data-processing system having one direct-access storage device from which information can be accessed by more than one computer.

sharpener **1.** A circuit or device for increasing the selectivity of another circuit or device. **2.** A circuit or device for decreasing the rise or fall time of a pulse or square wave. **3.** A circuit or device for steepening the response of a filter.

sharpness See SELECTIVITY.

sharp pulse A pulse having extremely fast rise and fall times and narrow width (i.e., a spike).

shaving The physical modification of a phonograph disc, or other permanent recording surface, in preparation for rerecording.

S_H band A section of the S BAND extending from 3700 to 3900 MHz.

sheath See POSITIVE-ION SHEATH.

shelf corrosion In a dry cell in storage, deterioration of the negative electrode because of local action in the zinc.

shelf life **1.** The longest period of time that electronic equipment can be continuously kept in storage before deterioration of materials or degradation of performance occurs. **2.** The longest period of time that a battery can be stored without use before it must be discarded or recharged.

shell **1.** An electronic orbit (imaginary shell) in an atom. **2.** The envelope of a component (e.g., the outer casing of a power transistor or the housing of a plug). **3.** The rigid case in an audio or video tape cassette.

shell-type choke See SHELL-TYPE INDUCTOR.

shell-type core A core that completely surrounds the coil(s) of a choke or transformer.

shell-type inductor An inductor in which the core completely surrounds the coil.

shell-type transformer A transformer in which the core completely surrounds the coils.

shf Abbreviation of SUPERHIGH FREQUENCY.

shield A metallic partition or box for confining an electric or magnetic field.

shield baffle A sheet-type shield. Also see BAFFLE, **2** and SHIELD PARTITION.

shield box A shield having a general box shape, and which is usually enclosed on all sides.

shield braid Tubing woven from wire, through which an insulated wire is passed and thus shielded.

shield can A cylindrical shield, usually enclosed on all sides.

shield disk A flat shield having a disk shape. Also see BAFFLE, **2**; SHIELD BAFFLE; and SHIELD PARTITION.

shielded cable Cable completely enclosed within a metal sheath that is either flexible or rigid.

shielded wire A single strand of insulated wire completely enclosed in a flexible or rigid shield.

shield partition A wall-type shield usually consisting of a single, flat sheet of metal, sometimes bent into an angle. Also called BAFFLE SHIELD (see BAFFLE, **2**).

shield plate See BAFFLE, **2**; SHIELD BAFFLE; SHIELD DISK; and SHIELD PARTITION.

shield room See CAGE.

shield wire A (usually grounded) wire, near and parallel to another wire that it shields.

shift **1.** To move from one operating point to another in a characteristic curve, or in the operation of an equipment. **2.** To transfer data from one point to another in a system, or move it left or right in a register.

shift flip-flop circuit A flip-flop designed especially as a stage in a shift register.

shift pulse In a shift register, a drive pulse that initiates the shifting of characters.

shift register In computers, calculators, and storage systems, a circuit (usually composed of flip-flops in cascade) in which pulses can be shifted from stage to stage and finally out of the circuit.

shingle-type photocell A device in which several separate photocells are series connected by slightly overlapping the ends of adjacent cells.

ship station A radio or radar station installed aboard a ship that is not in port.

ship-to-shore communication Radio communication between a ship at sea and a land-based station.

shock **1.** See ELECTRIC SHOCK. **2.** A signal applied momentarily to a circuit, as in the *shock excitation* of a tank. **3.** A sudden, dramatic change in an environmental variable (such as temperature). **4.** Physical blows or vibration.

shock absorber Any object or device intended for reducing physical vibration of a component, set of components, circuit, or system.

shock device **1.** A device for administering shock therapy (see ELECTROSHOCK, **1**). **2.** An induction coil and associated primary supply for applying high voltage to a wire fence.

shock excitation Driving an inductance-capacitance (LC) tuned circuit into damped oscillation by momentarily applying a pulse.

shock-excited oscillator A type of self-excited oscillator in which the transistor is suddenly cut off

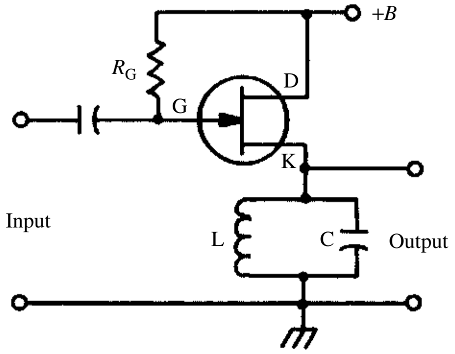

shock-excited oscillator

by applying a cutoff voltage to the gate or base electrode. This abrupt interruption of steady drain or collector current shocks the tank into damped oscillations.

shock hazard **1.** Any situation that presents the danger of electric shock to attendant personnel. **2.** The existence of a potential difference that will cause a current of at least 5 mA to flow through a resistance of 500 ohms or more, for a prolonged period of time.

Shockley diode See FOUR-LAYER DIODE.

shock mount A structure that secures a microphone while minimizing the pickup of vibrations through the table, floor, or other surface on which the microphone is placed.

shock therapy See ELECTROSHOCK, **1**.

shoran Contraction of SHORT-RANGE NAVIGATION.

shore effect The tendency of radio waves traveling along a shore to be bent either toward or away from the shore. It can occur because of differences in surface conductivity and/or atmospheric temperature over land, as compared with water.

shore station A fixed, land-based radio station that communicates with ships at sea.

shore-to-ship communication See SHIP-TO-SHORE COMMUNICATION.

short circuit An often unintended low-resistance path through which current bypasses a component or circuit.

short-circuit current In a power supply, the current that flows when the output is directly shorted. Many power supplies have shutdown devices that cause the current to stop flowing when the output terminals are short-circuited; other supplies effectively insert resistance in series with the load, if necessary, to limit the current.

short-circuiting switch A rotary selector switch in which unused contacts are automatically short-circuited.

short-circuit parameter A parameter for which zero resistance is assumed in the part of the circuit under consideration. The current amplification factor (alpha) of a common-base-connected transistor is such a parameter because its collector load resistance is assumed to be zero.

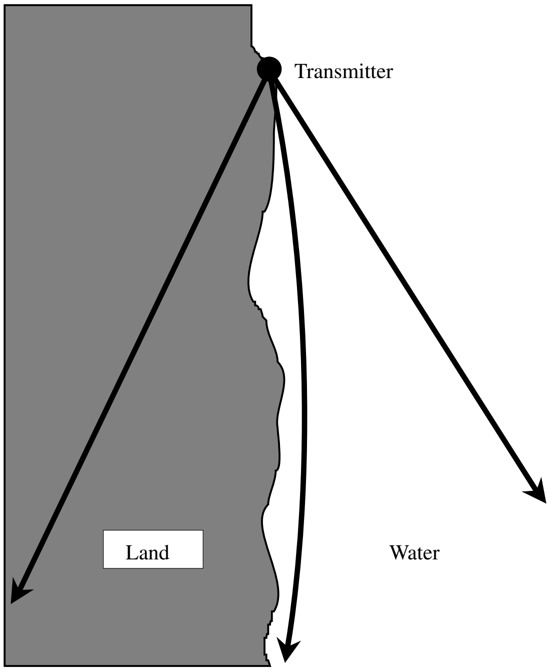

shore effect

shorted-stub tuning Tuning a stub to match a feeder to an antenna by sliding a short-circuiting bar along the two wires of the feeder.

shorting bar A thick, metal strap for short-circuiting two binding posts.

shorting link A sheet-metal strip for connecting together two binding posts.

shorting loop In a telephone system, a device that short-circuits two specified points for the purpose of testing or line fault location.

shorting stick A metal rod with an insulating handle, used to short-circuit a charged capacitor to remove the shock hazard.

shorting switch See SHORT-CIRCUITING SWITCH.

short-line tuning Use of a parallel capacitance to tune a transmission line that is less than a quarter-wave long.

short-range navigation Contraction, shoran. Navigation by means of SHORT-RANGE RADAR.

short-range radar A radar having a 50- to 150-mile maximum line-of-sight range for a 1-square-meter reflecting target that is perpendicular to the radar beam.

short skip Skip of only a few hundred miles range. Also see SHORT-SKIP COMMUNICATION.

short-skip communication Radio communication via the ionosphere over relatively short distances (400 to 1300 miles). See, for example, SPORADIC-E SKIP.

short-term drift The gradual change in the value of a quantity, such as frequency or voltage, observed over a comparatively brief interval, as opposed to change occurring over a long period. Compare LONG-TERM DRIFT.

short-term effect The variation of any electrical parameter over a relatively brief time interval. Example: frequency drift over a short time period. Also called *short-time effect.*

short-term stability Stability reckoned over a comparatively brief time interval, as opposed to stability for a long period. Compare LONG-TERM STABILITY.

short-time effect See SHORT-TERM EFFECT.

shortwave **1.** Pertaining to wavelengths shorter than 200 meters (i.e., frequencies higher than 1.50 MHz). **2.** Pertaining to the frequencies above the standard amplitude-modulation broadcast band (above 1.605 MHz), but below 30 MHz.

shortwave converter A superheterodyne converter for adapting a longwave receiver (such as a broadcast receiver) for shortwave reception.

shortwave listener Abbreviation, SWL. A radio hobbyist who receives, but does not transmit, shortwave signals.

shortwave receiver Any radio receiver capable of intercepting and demodulating signals in the range 1.705 MHz to 30 MHz.

shortwave transmitter Any radio transmitter capable of producing energy in the range 1.705 MHz to 30 MHz.

shot-effect noise Electrical noise caused by random fluctuations in a current, as in a diode or transistor. Also see EQUIVALENT NOISE RESISTANCE. Compare THERMAL NOISE.

shotgun microphone A highly directional microphone sensitive only to sounds coming from a specific direction; the response pattern has a narrow main lobe. It name results from its long, cylindrical configuration.

shot noise Electrical noise arising from intermittent impulses, such as those produced by spark discharges, make-and-break contacts, etc. Its name results from its resemblance to pistol shots.

shrink The amount by which a material being measured with an electronic instrument decreases in surface dimension. Compare STRETCH.

shrink tubing Plastic sleeving placed over a conductor or at a conductor/connector joint, and made to shrink tightly with the application of heat.

shunt Synonym, *parallel.* **1.** Pertaining to the connection of one component across (in parallel with) another (e.g., *shunt resistor*). **2.** Pertaining to connection of components in such a manner that they each (or all) are subjected to identical voltages. **3.** A deliberately produced short circuit between two specific points in a device or system. **4.** To deliberately bypass some part of a system by means of a short circuit.

shunt circuit See PARALLEL CIRCUIT.

shunt-diode rectifier A rectifier circuit in which the diode is connected in parallel with the source and load. Compare SERIES-DIODE RECTIFIER.

shunt-fed **1.** Pertaining to a circuit or device in which the direct-current operating voltage and alternating-current signal voltage are applied in parallel to an electrode. **2.** Pertaining to a base-grounded vertical antenna excited at some point above ground.

shunt feed See PARALLEL FEED.

shunt feedback A feedback system in which the fed-back signal is presented to the input of the network in parallel with the input signal. Compare SERIES FEEDBACK.

shunt generator An electric generator in which the armature and field windings are connected in parallel. Compare SERIES GENERATOR.

shunting effect The condition in which a quantity, such as stray capacitance or resistance, acts in parallel with another quantity. Example: the shunting (parallel) resistance of an electrolytic capacitor.

shunt leads Interconnecting wires used for the purpose of attaching a shunting component to a test instrument.

shunt limiter See PARALLEL LIMITER.

shunt loading The parallel insertion of reactance in a circuit, for the purpose of impedance matching.

shunt motor An electric motor whose armature and field windings are connected in parallel. Compare SERIES MOTOR.

shunt regulator A voltage-regulator circuit in which the controlled transistor or vacuum tube is in parallel with the output (load) terminals. Compare SERIES REGULATOR.

shunt resistor **1.** A resistor connected in parallel with a meter or recorder to increase its current range. **2.** A resistor connected in parallel with a voltmeter to convert it into a current meter. Compare MULTIPLIER RESISTOR.

shunt-series circuit See PARALLEL-SERIES CIRCUIT.

shunt tee junction A waveguide H-PLANE TEE JUNCTION.

shunt-type frequency multiplier A varactor frequency multiplier circuit in which the varactor is in parallel with the input and output. Compare SERIES-TYPE FREQUENCY MULTIPLIER.

shunt-type resonance bridge A resonance bridge in which the impedance arm is a parallel-resonant circuit. Compare SERIES-TYPE RESONANCE BRIDGE.

shunt-wound generator See SHUNT GENERATOR.

shunt-wound motor See SHUNT MOTOR.

SI Abbreviation of (Standard) INTERNATIONAL SYSTEM OF UNITS.

S/I Abbreviation of *signal-to-intermodulation ratio.*

Si Symbol for SILICON.

sibilants **1.** High-frequency (hissing) components of speech. **2.** High-frequency sounds or audio signals.

SIC Abbreviation of *specific inductive capacity* (see DIELECTRIC CONSTANT).

SiC Formula for SILICON CARBIDE.

sideband **1.** With respect to a carrier, one of the additional frequencies generated by the modulation process. In simple amplitude modulation, the two sidebands are $fc + fm$ and $fc - fm$, where fc is the carrier frequency, and fm is the modulation frequency. **2.** Pertaining to sidebands as defined in **1**.

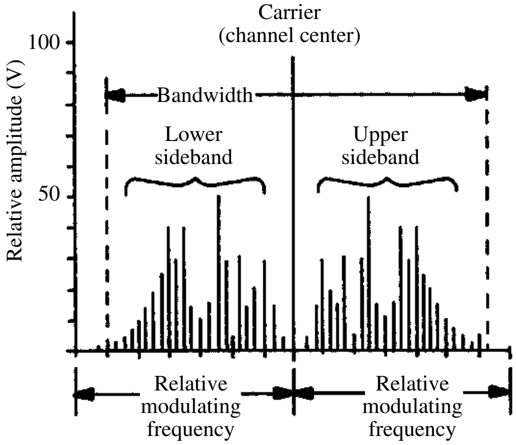

sideband

sideband attenuation See SIDEBAND CUTTING.

sideband cutting Elimination or attenuation of the sidebands of a modulated signal by a circuit having insufficient bandwidth.

sideband frequency The frequency of the modulation-generated signal accompanying a carrier. One sideband frequency is that of the carrier minus that of the modulating signal; another is the sum of the carrier and the modulation frequency. See also SIDEBAND, **1**.

sideband interference **1.** Interference arising from one or both of the normal sidebands of a modulated signal. **2.** Interference caused by spurious sidebands, resulting from overmodulation.

sideband power The power contained in the sideband(s) of a signal.

sideband slicing See SIDEBAND CUTTING.

sideband splatter In an amplitude-modulated or single-sideband signal, the emission of sideband energy at frequencies other than within the designated channel. Also simply called splatter.

sideband technique A method of using, for communications or other purposes, one or both of the sidebands of a modulated signal without the carrier.

side-chain amplifier An auxiliary amplifier that is external to a main amplifier. Such an amplifier might be used, for example, in a feedback channel or in a volume-compression or volume-expansion channel.

side frequency See SIDEBAND.

side lobe In certain directional antenna systems, a minor lobe in the horizontal-plane directivity pattern that appears at right angles, or nearly at right angles, to the main lobe. Such a lobe represents reduced sensitivity and/or power gain relative to the main lobe. Also see MAIN LOBE, MINOR LOBE.

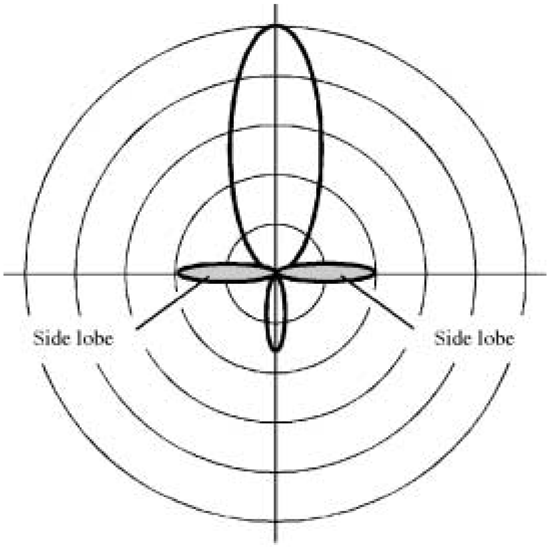

side lobe

sidelobe suppression Elimination of the sidelobe(s) from the radiation pattern of an antenna.

sidestacked antennas Antennas mounted in a horizontal line, parallel to each other, and connected by a common coupler to a transmitter or receiver.

sideswiper A manual telegraph key operated by moving the lever sideways, rather than up and down.

sidetone **1.** In wire telephony, the reproduction by the receiver of sounds picked up by the transmitter of the same telephone. **2.** In radiotelegraphy, an audible tone actuated when the carrier is transmitted. It allows the sending operator to hear Morse code elements as they are sent.

sidetone telephone A telephone set with no provision for canceling the sidetone.

siemens Symbol, S. The SI unit of conductance. The conductance of a component or medium in siemens is equal to the reciprocal of the resistance in ohms.

Siemen's electrodynamometer A spring-tension meter that operates by means of torque, with zero current through the device representing zero torque. It can be used for measurements of current, voltage, or power.

Sierra Phonetic alphabet code word for the letter S.

sig Abbreviation of SIGNAL.

sign **1.** Any indicator denoting whether a value is positive or negative. **2.** A graphic device indicating an operation. Examples: + (addition), × (multiplication). **3.** Any symbol. An ampersand, for example, is an "and" sign. **4.** A characteristic symptom of malfunction or improper operation (e.g., a high standing-wave ratio in an antenna system is a sign of an impedance mismatch).

signal An electrical quantity, such as a current or voltage, that can be used to convey information for communication, control, calculation, etc.

signal-actuated voice recorder Abbreviation, SAVOR. A recorder that goes into operation automatically when the speaker starts talking and stops when the speaker finishes.

signal amplitude The intensity of a signal quantity (see SIGNAL).

signal booster See PREAMPLIFIER.

signal channel In a system, a channel through which only signals flow, control and modifying impulses being accommodated by other channels.

signal circuit A circuit handling signal currents and voltages to the exclusion of control and operating currents and voltages.

signal conditioner Any accessory device (such as a peak probe, demodulator probe, current shunt, etc.) used to modify or change the function of a basic instrument (such as an electronic voltmeter).

signal converter See CONVERTER, **1**.

signal current The current component of a signal, as opposed to operating current in a system.

signal diode A diode designed primarily for light-duty signal applications (detection, demodulation, modulation, curve changing), as opposed to the heavy-duty applications of power diodes and rectifiers.

signal distance In two words (bit groups) of the same length, the number of corresponding bit positions whose states differ. For example, the signal distance between 01001 and 10011 is 3.

signal-flow analysis A graphic method of analyzing circuits, particularly those using feedback, through the use of diagrams in which straight arrows represent transmission paths, dots represent nodes, and curved arrows represent feedback paths.

signal-flow diagram The transmission-path diagram used in SIGNAL-FLOW ANALYSIS.

signal gain The gain of an amplifier circuit—especially if used in small-signal applications. See also AMPLIFICATION, and GAIN.

signal generator An instrument that produces signals of precise frequency and amplitude, usually over a wide range.

signal ground **1.** Any circuit point that remains at zero signal potential. **2.** A connection to a point that is deliberately maintained at zero signal potential.

signal/image ratio See SIGNAL-TO-IMAGE RATIO.

signaling In a communications system, the exchange of data in electrical form, either analog or digital.

signaling rate In data communications, the speed at which data is transmitted. It is commonly expressed in bits per second (bps). Also, it is sometimes expressed in baud or in words per minute (wpm).

signaling time slot In a communications signal, a specified interval of time, starting at a certain instant in each signal frame. This interval is used exclusively for the purpose of signaling.

signal injection **1.** The introduction of a signal into a circuit. **2.** A method of troubleshooting in communications receivers. A signal generator is used to introduce a test signal into each stage, starting with the output and proceeding stage-by-stage toward the input, until the defective stage or component is located.

signal injector A simple (usually single-frequency) signal generator used in troubleshooting to introduce a test signal at selected points in a circuit, to locate malfunctioning stages or components. Also see SIGNAL INJECTION, **2**.

signal intensity See SIGNAL STRENGTH.

signal inversion Phase reversal of a signal passing through a circuit, device or medium.

signal level At a given point in a circuit, the strength of a signal, with respect to a reference amplitude.

signal loss **1.** A reduction in the amplitude of a signal as it passes through a system. **2.** The complete disappearance of a signal. **3.** See FRACTIONAL GAIN.

signal mixer See MIXER.

signal/noise ratio See SIGNAL-TO-NOISE RATIO.

signal notcher See NOTCH FILTER.

signal peaker See PEAK FILTER.

signal power The amplitude of a signal expressed in watts, milliwatts, or microwatts, as opposed to amplitude expressed as a current or voltage.

signal processor Any device, (e.g., preamplifier, expander, amplitude limiter, delay network) inserted into or added onto a system to modify an input or output signal.

signal rectification The conversion of an alternating-current signal into a proportionate direct-current signal, usually by means of a diode circuit.

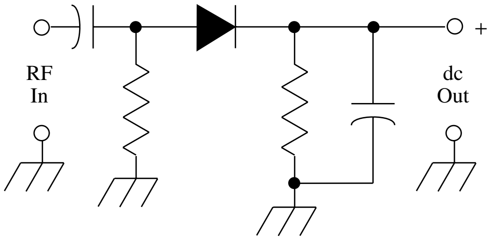

signal rectification

signal rectifier See SIGNAL DIODE.

signal regeneration See SIGNAL RESHAPING.

signal reshaping **1.** The processing of a signal so that it acquires its original waveform. Also called *signal regeneration.* **2.** Passing a digital signal of any type through a circuit that delivers a uniform output pulse on a real-time one-to-one basis.

signal shifter **1.** A device used for quickly changing the frequency of a transmitted signal. **2.** A device that automatically causes a transmitted signal to be sent on a frequency that differs from the receiver frequency by a known and predetermined amount. **3.** See MIXER, **1**. **4.** See CONVERTER, **1**.

signal squirter See SIGNAL INJECTOR.

signal strength The amplitude of a signal, usually in terms of voltage. Current or power is specified in some applications.

signal-strength meter **1.** See FIELD-STRENGTH METER. **2.** See S-METER.

signal synthesizer A special signal generator delivering signals whose frequency, amplitude, and waveshape can be adjusted at will.

signal time delay The time required for an element of a signal to be transmitted through a circuit or network. This delay results in phase shift in an amplifier.

signal-to-distortion ratio In a receiver, the ratio of the desired signal to the level of distortion other than the specified signal. Usually expressed in decibels (dB).

signal-to-image ratio Abbreviation, S/I. In a receiver, the ratio of signal amplitude to image amplitude, both being measured in the same units. It is usually expressed in decibels (dB).

signal-to-noise-plus-noise ratio Abbreviation, (S+N)/N. In a receiver, the ratio of the combined signal and noise amplitude to the amplitude of the noise alone. It is usually expressed in decibels (dB). Peak voltages are used to determine this ratio in the case of pulse noise; root-mean-square (rms) voltages are used in the case of random noise.

signal-to-noise ratio Abbreviations: S/N, SNR. In a receiver, the ratio of signal amplitude to noise amplitude. It is usually expressed in decibels (dB). Peak voltages are used to determine this ratio in the case of pulse noise; root-mean-square (rms) voltages are used in the case of random noise.

signal-to-noise-and-distortion ratio Abbreviation, SINAD. In a receiver, the ratio of the desired signal to the level of noise and distortion, other than the specified signal. It is usually expressed in decibels (dB).

signal tracer A tuned or untuned detector/amplifier having an input probe and an output indicator (meter, loudspeaker, or both), for following a test signal through a circuit.

signal voltage The voltage component of a signal, as opposed to the operating voltage of the circuit generating or passing the signal.

signal wave **1.** Any electromagnetic disturbance of a periodic nature that is modulated to convey information. **2.** The visual illustration or rendition of an electromagnetic disturbance that is modulated to convey information.

signal winding In a magnetic amplifier or saturable reactor, the coil that receives the control current.

signal window See WINDOW, **2**.

signal wobbulator A frequency modulator used with an unmodulated signal generator to provide sweep signals for visual alignment. Also see WOBBULATOR.

sign bit A one-bit SIGN DIGIT.

sign digit A character indicating the sign (positive or negative) of the value of the field or word to which it is attached (usually at the end).

signed field In a computer record, a field having a number whose sign is indicated by a SIGN DIGIT.

significant digits See SIGNIFICANT FIGURES.

significant figures In a numerical quantity, especially one expressed in scientific (power of 10) notation, those figures (digits) that depict a quantity to a required, relevant, or justifiable degree of precision. For example, 173,201 expressed to three significant figures is 1.73×10^5; 0.00477583 expressed to four significant figures is 4.776×10^{-3}. See also SCIENTIFIC NOTATION.

silencer See AUTOMATIC NOISE LIMITER.

silent alarm In security systems, the transmission of a warning signal to attendant human operators and/or computers, without producing an audible or visible warning to intruders.

silent alignment See VISUAL ALIGNMENT.

silent piano See ELECTRONIC PIANO.

silica pencil A rod of silicon dioxide heated to emit infrared rays.

silicon Symbol, Si. A metalloidal element. Atomic number, 14. Atomic weight, 28.086. Silicon is abundant in the earth's crust. It is used in many semiconductor devices, including integrated circuits, diodes, photocells, rectifiers, and transistors.

silicon capacitor See VOLTAGE-VARIABLE CAPACITOR.

silicon carbide Formula, SiC. A compound of silicon and carbon valued as a semiconductor, an abrasive material, and a refractory substance. The commercial product is made by heating carbon and sand to a high temperature in an electric resistance furnace. Also called CARBORUNDUM.

silicon cell A type of photovoltaic cell using specially processed silicon as the light-sensitive material. This cell has a comparatively high voltage output.

silicon-controlled rectifier Abbreviation, SCR. A four-layer semiconductor device commonly used in power control applications (e.g., light dimmers and motor-speed controls). The electrodes are called the anode, the cathode, and the gate. The control signal is applied to the gate.

silicon-controlled switch Abbreviation, SCS. A four-terminal semiconductor switching device similar to the SILICON-CONTROLLED RECTIFIER. It is used for light-duty switching.

silicon crystal detector **1.** See SILICON JUNCTION DIODE. **2.** See SILICON POINT-CONTACT DIODE. **3.** A point-contact diode in which a lump of silicon is contacted by either a fine wire (cat whisker) or a blunt-tipped steel screw under pressure.

silicon detector See SILICON CRYSTAL DETECTOR.

silicon-diffused transistor A silicon bipolar transistor fabricated by diffusion techniques. It is characterized by high power-dissipation tolerance.

silicon diode A semiconductor diode in which the semiconductor material is specially processed silicon. Also see SILICON JUNCTION DIODE and SILICON POINT-CONTACT DIODE.

silicon dioxide Formula, SiO_2. A compound of silicon and oxygen. In the passivation of transistors and integrated circuits, a thin layer of silicon dioxide is grown on the surface of the wafer to protect the otherwise exposed junctions.

silicone A polymeric material characterized by a recurring chemical group containing oxygen and silicon atoms in the main chain as links. Various silicone compounds have numerous uses in electronics.

silicon junction diode A semiconductor diode using a pn junction in a silicon wafer. Compare SILICON POINT-CONTACT DIODE.

silicon on sapphire Abbreviation, SOS. Pertaining to integrated-circuit fabrication in which a silicon epitaxial layer is grown on a sapphire substrate.

silicon oxide A compound containing both silicon monoxide and silicon dioxide, and having dielectric properties. It is used in the manufacture of metal-oxide-semiconductor (MOS) devices.

silicon photocell A photocell using a silicon pn junction as the light-sensitive medium.

silicon photodiode A silicon diode constructed so that radiant energy can strike the barrier between the P- and N-type materials. A reverse bias is applied. When light falls on the P-N junction, current flows. The current is proportional to the intensity of the impinging energy, within certain limits. The greatest sensitivity is in the near infrared (IR). When energy of variable brightness falls on the P-N junction under conditions of reverse bias, the output current follows the intensity variations. This makes the device useful in fiberoptic communications systems.

silicon point contact The contact between a pointed metal wire (cat whisker) and a silicon wafer.

silicon point-contact diode A diode in which a tungsten wire (cat whisker) contacts a wafer of single-crystal silicon. It is useful at ultra high frequencies (UHF). Compare SILICON JUNCTION DIODE.

silicon rectifier A semiconductor rectifier consisting essentially of a junction between n- and p-type silicon inside a specially processed wafer or plate of single-crystal silicon.

silicon resistor See CRYSTAL RESISTOR.

silicon solar cell A relatively heavy-duty photovoltaic cell using specially processed silicon as the light-sensitive material.

silicon steel A high-permeability, high-resistance steel containing 2 to 3 percent silicon. It is used as core material in transformers and other electromagnetic devices.

silicon transistor A transistor in which the semiconductor material is single-crystal silicon.

silk-enameled wire Wire whose insulation is a layer of silk on top of an enamel coating.

silver Symbol, Ag. A precious metallic element. Atomic number, 47. Atomic weight, 107.87. It is used in circuits where low resistance and high Q are mandatory.

silver arsenide trisulfide See PROUSTITE.

silver-dollar construction Printed-circuit assembly on a disk-shaped board, about the size of a U.S. silver dollar.

silver-mica capacitor A fixed capacitor made by painting or depositing a silver layer (capacitor plate) on both faces of a thin mica film (dielectric separator).

silver migration The undesirable tendency of silver to be removed from one location and deposited in another under adverse environmental conditions.

silver-oxide battery A set of two or more silver-oxide cells stacked one atop the other, electrically connected in series. The resulting battery has a cylindrical shape. A set of four cells provides approximately 6 volts under no-load conditions; a battery of six cells provides 9 volts; a battery of eight cells provides 12 volts. See also SILVER-OXIDE CELL.

silver-oxide cell An electrochemical cell having a button-like shape, small enough to fit inside a wristwatch. There are several available sizes and thicknesses, all with similar appearance. The generated potential difference under no-load conditions is 1.5 volts, with a high ratio of stored energy per unit mass. The cell has a flat discharge curve; the voltage remains essentially constant until the charge is almost depleted, and then the voltage drops rapidly.

silver solder A solder consisting of an alloy of silver, copper, and zinc. It has a comparatively high melting temperature. Also see HARD SOLDER.

silverstat A multiconductor device used to adjust the balance of a resistance or reactance bridge.

similar decimals Two or more decimal numbers that have the same number of digits to the right of the radix point (e.g., 3.14 and 6.39, or 1.234 and 1.000).

simple quad A combination of two parallel paths—each containing two elements in series.

simple tone A pure sine-wave tone (i.e., one having negligible harmonic content).

simplex channel An information channel for unidirectional transmission.

simplex system **1.** In data communications, a system that transmits data in only one direction. Compare FULL DUPLEX SYSTEM and HALF DUPLEX SYSTEM. **2.** In voice communications via radio, a direct path over a single channel, used alternately for transmitting and receiving at each station.

simplex telegraphy Wire telegraphy in which only one message at a time can be sent over a line.

simplification of circuits See CIRCUIT SIMPLIFICATION.

simulation **1.** Imitation of the performance of a process, device, or system. **2.** The use of a mathematical model to represent a physical process, device, or system. **3.** The use of a computer, sometimes with virtual reality hardware and software, to mimic a real-life situation.

simulator **1.** A software or hardware system capable of simulation (see SIMULATION, **2**). **2.** A computer program whose implementation allows programs written for one computer to be compatible with another computer. **3.** A system of equipment for simulation (see SIMULATION, **1**).

simulcast **1.** To broadcast a program over two or more different channels at the same time. **2.** To broadcast a program over two or more different types of mode, for example, television and radio, at the same time. **3.** A program broadcast over two or more channels or modes at the same time.

simultaneous access See PARALLEL ACCESS.

simultaneous broadcasting See SIMULTANEOUS TRANSMISSION.

simultaneous computer See PARALLEL COMPUTER.

simultaneous transmission The transmission of the same information in two or more channels, or by means of two or more processes, at the same time.

sin Abbreviation of SINE.

$\sin^{-1}$ Symbols for the inverse of the sine function, also called the *arc sine.*

sine Abbreviation, sin. The trigonometric function a/c, the ratio of the opposite side of a right triangle to the hypotenuse.

SINAD See SIGNAL-TO-NOISE-AND-DISTORTION RATIO.

sine galvanometer A galvanometer in which the sine of the angle of deflection is proportional to the current. Compare TANGENT GALVANOMETER.

sine law The variation in radiation intensity in any direction from a linear source is proportional to the sine of the angle between the axis of the source and the direction of interest.

sine potentiometer A POTENTIOMETER whose output is proportional to the sine of the angle through which the shaft has rotated.

sine wave A periodic wave that can be represented by a sine curve (i.e., its amplitude is directly proportional to the sine of a linear quantity, such as displacement or time). Compare COSINE WAVE.

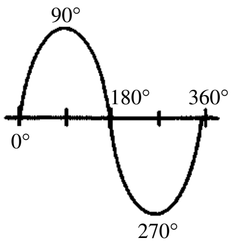

sine wave

singing Audible oscillation in a circuit or device, such as the low-level buzz emanating from the filament of a lamp dimmed with a phase-control circuit.

single-address coding In computer programming, the use of instruction words that contain the address for the location of the data to be operated on, and no other addresses.

single-board computer Abbreviation, SBC. A computer built entirely on one circuit board.

single-button microphone A carbon microphone having only one button attached to the diaphragm. Also see BUTTON MICROPHONE.

single-channel codec A form of CODEC intended for operation on a single signal source, rather than in a multiplexed system.

single-chip codec An integrated circuit contained entirely on one chip and in one package that can accomplish all CODEC functions. It can be a single-channel device or multiplexed.

single-cotton-covered wire Wire insulated with one layer of cotton.

single cotton enameled wire Wire insulated with one layer of cotton on top of an enamel coating.

single-crystal Pertaining to the internal structure of a crystalline material, in which the characteristic lattice is continuous throughout any size piece of the material. Also called MONOCRYSTALLINE.

single-crystal material A substance, such as a semiconductor, of which a sample, regardless of size, consists of only one crystal (i.e., there are no grain boundaries). Also see SINGLE-CRYSTAL. Compare POLYCRYSTALLINE MATERIAL.

single-crystal pulling See CZOCHRALSKI METHOD.

single-dial control Adjustment of a multistage system via one rotatable, calibrated control attached to a ganged arrangement that tunes all stages simultaneously.

single-diffused transistor A transistor in which only one diffusion of an impurity substance is made. Thus, in a diffused-base transistor, a single diffusion provides the base region and at the same time creates the emitter-base and collector-base junctions. Compare DOUBLE-DIFFUSED TRANSISTOR.

single-electron memory Abbreviation, SEM. A computer memory in which the movement of one electron can change a logic bit from 1 (high) to 0 (low) or vice versa.

single-element rotary antenna See ONE-ELEMENT ROTARY ANTENNA.

single-ended circuit A circuit that has one end grounded, as opposed to a double-ended circuit and push-pull circuit.

single-ended deflection In an oscilloscope or similar device, horizontal or vertical deflection provided by a single-ended deflection channel. Compare PUSH-PULL DEFLECTION.

single-ended input An input circuit with one terminal grounded (or the equivalent ungrounded input circuit). Also called *unbalanced input.* Compare BALANCED INPUT.

single-ended multiplexer A group of analog switches that selects from several analog signals.

single-ended output An output circuit with one terminal grounded (or the equivalent ungrounded output). Also called *unbalanced output.* Compare BALANCED OUTPUT.

single-ended push-pull circuit An arrangement, such as a complementary symmetry circuit, that provides push-pull output with single-ended input, but does not require transformers.

single-frequency Also called *fixed-frequency.* Pertaining to circuits or devices that normally operate at one frequency only (e.g., *single-frequency oscillator*).

single-frequency amplifier An amplifier that normally operates at only one frequency (or within a very narrow band of frequencies) (e.g., an intermediate-frequency amplifier, or a selective audio-frequency amplifier used for harmonic analysis or bridge balancing).

single-frequency duplex Two-way communications over one medium or frequency. Voice-actuated (VOX) or break-in devices are used at both ends of the circuit.

single-frequency oscillator An oscillator that normally delivers a signal at only one frequency until it is switched to another frequency (e.g., *crystal-controlled oscillator*).

single-frequency receiver A radio or television receiver that normally operates at one carrier frequency, rather than being tunable. Such receivers are used in monitoring specific programs, picking up standard-frequency signals, and in similar applications.

single-gun color picture tube A color-television picture tube in which the image is produced by a single beam that scans the red, green, and blue color-phosphor dots sequentially.

single-hop propagation Long-distance radio-wave propagation involving only one encounter with the ionosphere, and involving no intermediate reflections from the earth's surface.

single-hop return distance The return distance, as a function of the angle of departure from a radio transmission, from a layer of the ionosphere. The illustration shows the maximum possible distance under average conditions and assuming an angle of departure of zero degrees.

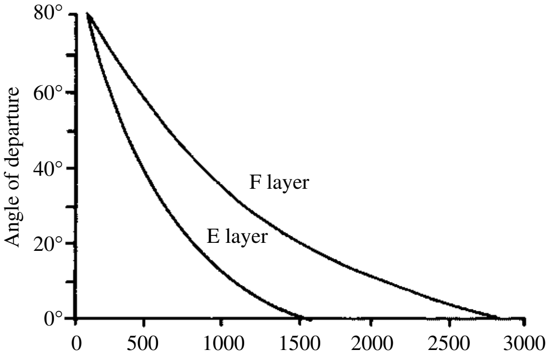

single-hop return distance (miles)

single-image response In an oscilloscope presentation, a single pattern, as opposed to a double-trace pattern.

single-inline package Abbreviation, SIP. A flat, molded component package having terminal pins along one edge. All the pins lie along a common line.

single-junction transistor See UNIJUNCTION TRANSISTOR.

single-layer coil A coil whose turns are wound side by side in one layer.

single-layer solenoid See SINGLE-LAYER COIL.

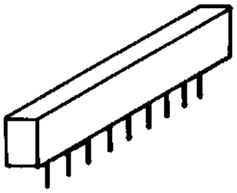

single-inline package

single-line tap In a telephone system, a connection that provides or designates a separate line (e.g., to serve a single household).

single-loop feedback Feedback through only one path.

single phase Pertaining to the presence or generation of one alternating-current phase only. Compare POLYPHASE.

single-phase full-wave Pertaining to a rectifier operated from a single-phase alternating-current (ac) power line, and rectifying both half-cycles of ac voltage. Compare SINGLE-PHASE/HALF WAVE.

single-phase/full-wave bridge A bridge rectifier operated from a single-phase alternating-current supply, usually from the untapped secondary winding of a transformer. Compare SINGLE-PHASE/FULL-WAVE CIRCUIT and SINGLE-PHASE/HALF-WAVE CIRCUIT.

single-phase/full-wave circuit A rectifier circuit in which each half-cycle of single-phase alternating current is rectified by a separate diode supplied from the ends of a center-tapped winding of a transformer. Compare SINGLE-PHASE/FULL-WAVE BRIDGE and SINGLE-PHASE/HALF-WAVE CIRCUIT.

single-phase/half-wave Pertaining to a rectifier operated from a single-phase alternating-current (ac) power line, and rectifying only one half-cycle of ac voltage.

single-phase/half-wave circuit A rectifier circuit in which a diode, output load, and single-phase alternating-current supply are connected in series, only one half-cycle of the cycle being passed by the diode. Compare SINGLE-PHASE/FULL-WAVE CIRCUIT and SINGLE-PHASE/FULL-WAVE BRIDGE.

single-phase rectifier See SINGLE-PHASE/FULL-WAVE BRIDGE, SINGLE-PHASE/FULL-WAVE CIRCUIT, and SINGLE-PHASE/HALF-WAVE CIRCUIT.

single-point ground One ground connection to which all channels of a circuit are returned. Such a common connection eliminates or greatly minimizes the common coupling often encountered when separate ground points are used.

single-pole double-throw Abbreviation, SPDT. Descriptive of an electrical, electronic, or mechanical switch with a pole that can be connected to either of two adjacent poles, but not to both.

single-pole single-throw Abbreviation, SPST. Descriptive of an electrical, electronic, or mechanical switch with a pole that can be connected to an adjacent pole (or disconnected from it) at will. It is used to provide the make and break function in a single circuit.

single rail **1.** A one-conductor communications medium with a ground return. **2.** A one-conductor data line, with a ground return.

single shot Also called *one shot.* Pertaining to circuit operation in which a single input pulse applied to a switching device (such as a multivibrator) causes it to deliver a single output pulse, rather than switch to a stable “on” state. A MONOSTABLE MULTIVIBRATOR operates in this mode.

single-shot multivibrator See MONOSTABLE MULTIVIBRATOR.

single sideband Abbreviation, SSB. Pertaining to a system of modulation in which one of the sidebands from an amplitude-modulated signal is attenuated or canceled out, leaving only one sideband. The carrier is generally suppressed also.

single-sideband suppressed-carrier Abbreviation, SSSC or SSBSC. Pertaining to a system of modulation in which the carrier and one sideband from an amplitude-modulated signal are suppressed; only the remaining sideband is transmitted. Sometimes this mode is simply called SINGLE SIDEBAND.

single signal Pertaining to a mode of reception in which signals appear on only one side of zero beat, enhancing selectivity and reducing interference among received signals. Most superheterodyne receivers have this feature; most direct-conversion receivers do not.

single-signal receiver A superheterodyne receiver that achieves high selectivity via a selective filter in the intermediate-frequency amplifier chain. Signals appear on only one side of zero beat, in a band whose width can be adjusted or selected for various values from about 200 Hz to 3 kHz.

single silk-covered wire Wire insulated with one layer of silk.

single-skip propagation See SINGLE-HOP PROPAGATION.

single-step operation See STEP-THROUGH OPERATION.

single sweep In an oscilloscope, a single time-axis deflection of the electron beam. Also see SWEEP, **1**, **2**. Compare RECURRENT SWEEP.

single-sweep blocking oscillator A blocking oscillator that cuts off after generating a single cycle or pulse.

single-throw switch A single-action switch with two or more poles.

single-tone keying Modulated continuous-wave keying. A single audio-frequency tone is used to amplitude-modulate or frequency-modulate the carrier.

single-track recorder A recorder, such as a magnetic-tape recorder or a graphic recorder, that permits recording along only one track.

single-trip multivibrator See MONOSTABLE MULTIVIBRATOR.

single-tuned circuit A circuit tuned by varying only one of its components [e.g., an intermediate-frequency transformer in which only the secondary coil (rather than both primary and secondary) is tuned].

single-turn coil **1.** A coil consisting of a single turn of wire, tubing, or strip. **2.** See RING INDUCTOR. **3.** See SHADING COIL.

single-turn potentiometer A potentiometer that can be adjusted through its entire resistance range by no more than one full rotation of the shaft. Usually, the turning range is somewhat less than a full circle (e.g., 300 degrees).

single-wire-fed antenna See WINDOM ANTENNA.

single-wire line **1.** See SINGLE-WIRE TRANSMISSION LINE. **2.** A single wire used for communication or control purposes. The earth furnishes the return path.

single-wire transmission line An antenna transmission line or feeder consisting of one wire only (see, for example, WINDOM ANTENNA).

sink **1.** A device or circuit into which current drains. **2.** See HEATSINK.

sink circuit The circuit associated with a load or other sink. Compare SOURCE CIRCUIT, **2**.

sinker A piece of semiconductor material used to reduce the base-collector junction resistance in a bipolar transistor.

sintering A process in which various solid bodies are formed from fusible powders at temperatures below their melting points. Example: *sintered magnetic core.*

sinusoidal Having the shape and properties of a SINE WAVE.

SIO Abbreviation of *serial input/output.*

SIP Abbreviation of SINGLE-INLINE PACKAGE.

six-phase rectifier A polyphase rectifier circuit operated from a three-phase supply. The output ripple frequency is six times the supply frequency.

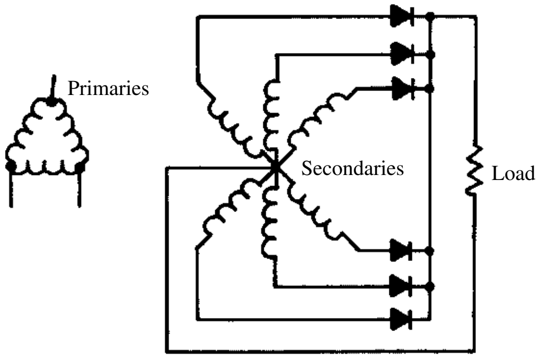

six-phase rectifier

SJD Abbreviation of SILICON JUNCTION DIODE.

skating In a phonograph turntable, the tendency of the tone arm to swing toward the spindle during record play, independent of the action produced by the stylus following the groove.

skeletal code A generalized computer routine needing only certain parameters to be usable for a specific application.

skeleton bridge A bridge consisting of an adjustable arm (potentiometer) and a pair of binding posts for each of the other three arms. Suitable resistors, capacitors, or inductors are connected to the binding posts to set up the bridge circuit desired.

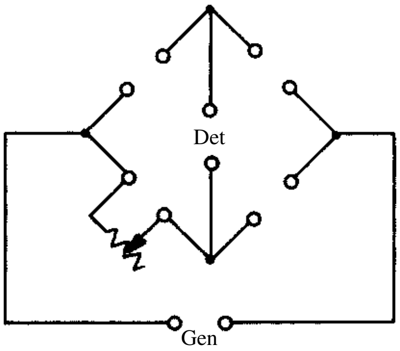

skeleton bridge

skeleton-type assembly **1.** A method of electronic-equipment construction in which a minimum of supporting members is used. An example is the use of an open framework, instead of a chassis, to support components. **2.** An assembly of electronic equipment, consisting essentially of a foundation unit (containing the basic circuitry) and plug-in units for setting up various complete equipments.

skew **1.** A condition resulting from failure of the horizontal synchronization in facsimile or television. The picture appears distorted, and appears as a non-rectangular parallelogram. **2.** In a print display, nonalignment of columns resulting from an incorrect number of line spaces in each line. **3.** In a probability function, an accumulation of values toward either side of center.

skewing **1.** The bending of a curve away from its normal shape. **2.** In a differential amplifier, the offset between two signals. Also see OFFSET.

Skiatron A special form of cathode-ray tube, with the fluorescent coating replaced by a screen of halide crystals that darken, instead of glow, when exposed to the electron stream.

skin depth The depth to which current penetrates below the surface of a conductor, as a result of the SKIN EFFECT.

skin effect The tendency of high-frequency alternating current to travel along the surface of a conductor; the high-frequency reactance is lower along the outside than at the center of a conductor. This tends to increase the resistivity of solid conductors at high alternating-current frequencies, as compared with low frequencies and direct current.

skip **1.** Ionosphere-reflected radio transmissions. **2.** In a computer program, an instruction whose sole function is that of causing a jump to the next instruction.

skip distance For a signal propagated via the ionosphere, the distance from the transmitter to the point at which the returned skywave strikes the earth.

skip fading For a signal propagated via the ionosphere, changes in signal strength caused by fluctuations in the altitude and/or contour of the ionized layer(s).

skip zone See ZONE OF SILENCE.

skirt selectivity **1.** The bandwidth between points of high attenuation (usually 30 dB or 60 dB) on the selectivity curve in a communications receiver. **2.** The relative steepness of the attenuation-vs.-frequency curve in a communications receiver.

SKM Abbreviation of *sine-cosine multiplier.*

skyhook **1.** Colloquialism for ANTENNA. **2.** A wire antenna supported by a captive balloon or kite.

sky noise Radio noise originating in outer space.

skywave A radio wave propagated by ionospheric reflections and/or refractions. Compare GROUND WAVE.

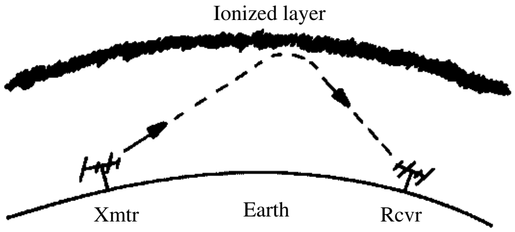

skywave

skywave correction A factor applied to long-range radionavigation signals to account for the time delay resulting from ionospheric propagation.

skywire See OUTSIDE ANTENNA.

slab **1.** A relatively thick body of quartz, ceramic, semiconductor, or dielectric. **2.** See SUBSTRATE.

slap back The return of sounds by an acoustically reflective object or surface a short distance away, resulting in almost immediate echoes.

slashed-field-gun CRT A straight-gun television picture tube (see STRAIGHT-GUN CRT). Because the gap between the anodes in this tube is slanted, the electrostatic field is diagonal, causing the electron and ion beams to be diverted at an angle.

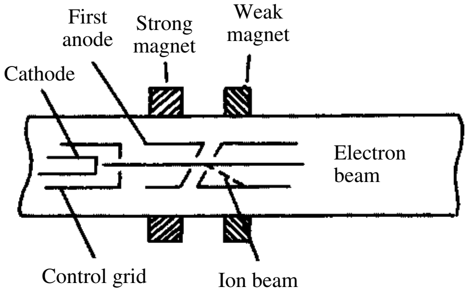

slashed-field-gun CRT

slave flash A photoflash operated by the light flash from another such unit.

slave relay A relay operated by, and whose action follows, a MASTER RELAY.

SLC Abbreviation of STRAIGHT-LINE CAPACITANCE.

sleeping sickness A gradual increase in transistor leakage current.

sleep machine An electronic device sometimes used as an aid for relaxation or sleeping. It consists of a wideband audio-frequency noise generator that produces low-level sounds similar to the noise of waves on a beach or a light wind through trees.

sleeve antenna A vertical antenna in which the upper half is a quarter-wave rod connected to the inner conductor of a coaxial feeder, and the lower half is a quarter-wave metal sleeve connected to the outer conductor of the feeder. Also called COAXIAL ANTENNA.

sleeving A material in tubular form that can be slipped over another material [e.g., insulating sleeving for wires (spaghetti)].

slew rate In an operational amplifier, the rate at which the output can be driven between its limits.

SLF Abbreviation of STRAIGHT-LINE FREQUENCY.

slice A semiconductor wafer cut from a single-crystal ingot.

slicer See CLIPPER-LIMITER.

slide-back meter An electronic voltmeter in which an unknown alternating-current signal voltage applied to the input of an amplifier stage is bucked by an internal, adjustable, accurately known signal voltage. The internal voltage is adjusted until a null occurs, indicating that its magnitude is equal to that of the unknown voltage.

slider A flat-spring contact that slides along the turns of a resistance or inductance coil to vary the coil's resistance or inductance. Also called a WIPER.

slide-rule dial A dial mechanism having a straight scale that resembles a slide rule.

slide switch A switch actuated by sliding a block-shaped button. Compare BAT-HANDLE SWITCH, PADDLE SWITCH, and ROCKER SWITCH.

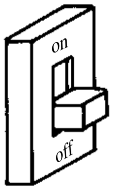

slide switch

slide wire A simple potentiometer consisting of a single, straight piece of resistance wire with a sliding contact. Also see SLIDE-WIRE RESISTOR.

slide-wire bridge A simple four-arm bridge in which the adjustable element is a single, straight resistance wire along which a clip or slider is moved, and that supplies two arms of the bridge (one on each side of the slider).

slide-wire resistor A variable resistor consisting of a single wire (straight or coiled) along whose length a slider is moved to vary the resistance.

sliding contact A contact that mates with another contact, or moves along a contacted surface, with a sliding motion. Also called SELF-CLEANING CONTACT and WIPING CONTACT.

slip **1.** In an eddy-current brake, coupling, or drive, the difference in speed between the field magnets and the iron eddy-current ring. **2.** In a synchronous motor, the difference between rotor speed and stator speed.

slip clutch In a gear or rack-and-pinion drive system, a device that releases the load if the torque becomes excessive. The gears then slip instead of being damaged.

slip ring See COLLECTOR RING, **1**.

slip speed See SLIP, **2**.

slope **1.** The slant of a line (graph) in rectangular coordinates, depicted as the ratio of the change in the dependent variable y to the change in the independent variable x. If (x_1,y_1) and (x_2,y_2) are two points on the line, then slope m is determined by $m = (y_2 - y_1)/(x_2 - x_1)$. **2.** The slant of a line in rectangular coordinates as defined in **1**, when the line is tangent to a curve (graph) at a specified point. **3.** The skirt(s) of a selectivity curve, particularly in a communications receiver, where a small change in frequency results in a significant change in gain or attenuation. **4.** The ratio of the extent of change in a quantity to the extent of change in some other quantity, when a causal relation exists between the magnitudes of the quantities. Example: See SLOPE RESISTANCE.

slope detector An amplitude-modulation (AM) receiving circuit detuned to one side of resonance (i.e., to a point along the skirt of the selectivity curve) to detect a frequency-modulated (FM) signal. The FM swing occurs along the slope of the resonance curve. Slope detection is useful in narrowband FM when conventional FM circuitry is not available.

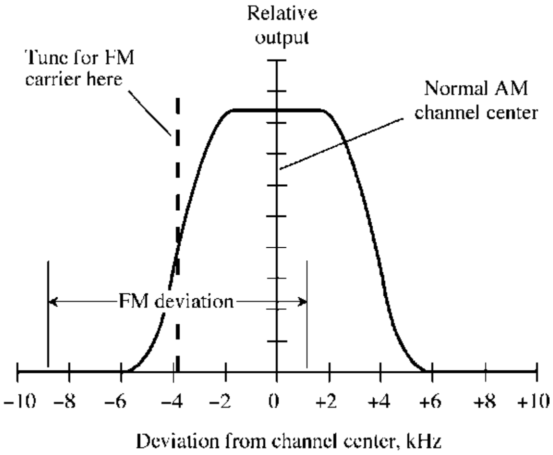

slope detector

slope resistance The ratio of a small change in voltage to a small change in current at an electrode or in a component.

slop-jar capacitor See WATER CAPACITOR.

slop-jar rectifier See ELECTROLYTIC RECTIFIER.

slot **1.** In the armature of a motor or generator, a groove in which the windings are laid. **2.** The notch in the response curve of a crystal filter.

slot antenna A microwave antenna that radiates energy through a slot cut in a surface, such as the metal skin of an aircraft.

slot cell A reinforcing, dielectric material (such as plastic) placed in the slot of a ferromagnetic core.

slot coupling Coupling microwave energy between a waveguide and a coaxial cable by means of two slots, one in the waveguide and the other in the outer conductor of the cable.

slot-discharge resistance See CORONA RESISTANCE.

slot insulation **1.** Insulation of wires in the slots of the armature of a motor or generator (see SLOT, **1**). **2.** A material in the form of tape or sheets, used for the purpose defined in **1**.

slot radiator See SLOT ANTENNA.

slotted line A device consisting of a section of air-dielectric coaxial line arranged for microwave measurements. The outer conductor is a metal cylinder, and the inner conductor is a concentric metal rod. The cylinder is provided with a lengthwise slot through which a small pickup probe extends for sampling the electromagnetic field inside the device. The probe is attached to a carriage that slides along a graduated scale on the outside of the cylinder. Radio-frequency energy is injected into one end of the line through a coaxial cable; as the probe moves along, response points are indicated by an external detector connected to the probe. The scale is read at these points to determine frequency, standing-wave ratio, impedance, and power. An alternate form of slotted line uses a section of slotted waveguide, instead of a section of coaxial line.

slotted rotor See SERRATED ROTOR PLATE.

slotted section See SLOTTED LINE.

slotted waveguide See SLOTTED LINE.

slot width **1.** The width of a slot in the armature of a motor or generator (see SLOT, **1**). **2.** The bandwidth of the notch in the response curve of a band-suppression filter of any kind. See, for example, NOTCH FILTER.

slow-acting relay Any relay designed to operate at some finite period following the application of the actuation voltage.

slow-blow fuse A fuse in which the melting wire breaks apart slowly. The time delay allows the fuse to withstand momentary current surges that do not damage the protected equipment, but that would cause a fast-blow fuse to break the circuit needlessly.

slow-break, fast-make relay A relay that opens slowly and closes rapidly.

slow-break, slow-make relay A relay that opens slowly and closes slowly.

slow charge Storage-battery charging in which a low current is passed through the battery over a long period of time. It ensures that the rated ampere-hour capacity will be restored to the battery.

slow death **1.** The gradual deterioration of transistor performance. **2.** The gradual deterioration in the performance of a component, circuit, device, or system.

slow drift The gradual change of a quantity or setting (usually in one direction). Compare FAST DRIFT.

slow-make, fast-break relay A relay that closes slowly and opens rapidly.

slow-make, slow-break relay A relay that closes slowly and opens slowly.

slow-operate, fast-release relay See SLOW-MAKE-FAST-BREAK RELAY.

slow-operate, slow-release relay See SLOW-MAKE-SLOW-BREAK RELAY.

slow-release, fast-operate relay See SLOW-BREAK-FAST-MAKE RELAY.

slow-release, slow-operate relay See SLOW-BREAK-SLOW-MAKE RELAY.

slow storage A form of memory with long storage and recovery time.

slow time scale An extended time scale, i.e., one larger than the time unit of the system under consideration.

SLS Abbreviation of SIDELOBE SUPPRESSION.

slug **1.** A movable core of ferromagnetic material, used to tune (varying the inductance of) coil by changing its position along the axis of the coil. Also see SLUG-TUNED COIL. **2.** A copper ring attached to the core of a relay for time-delay purposes (see SLUG-TYPE DELAY RELAY).

slug-tuned coil A coil whose inductance is varied by means of a ferromagnetic slug that slides in and out of the coil.

slug tuner A tuner for a radio or television receiver or test instrument, using slug-tuned coils.

slug-type delay relay A delayed-response relay that achieves time delay through the action of a heavy copper slug on the core. The slug forms a low-resistance, short-circuited single turn in which a current is induced by the magnetic flux, resulting from energizing the relay. The resulting flux of the slug opposes the buildup of relay-coil flux.

slumber switch An alarm-reset switch on an electronic clock radio. If the alarm activates, the slumber switch (usually a pushbutton device) can be pressed to turn off the alarm for a predetermined length of time. Also called *snooze button.*

SLW Abbreviation of STRAIGHT-LINE WAVELENGTH.

Sm Symbol for SAMARIUM.

small-current amplifier **1.** A direct-current (dc) amplifier for low-level input currents (i.e., currents of 1 milliampere or less). **2.** An amplifier (such as a silicon-transistor unit) requiring very low dc operating current.

small-scale frequency response For an analog circuit, the output frequency at which the level is –3 dB, relative to the maximum level, with a small signal at the input, normally 1 volt peak-to-peak.

small loop antenna A closed loop antenna with one to several turns and a circumference less than 0.1 wavelength at the highest operating frequency. This antenna is suitable for wireless reception, but generally not for transmission because the radiation resistance is extremely low. The antenna is least responsive along its axis, and is most responsive in the plane defined by its turn(s). The null along the axis is sharp and deep. A capacitor can be connected in series or parallel with the loop to provide a resonant response. This type of antenna can reduce interference caused by human-made noise or strong local signals. It is also useful for radio direction finding (RDF) at frequencies up to approximately 20 MHz. Compare FERRITE-ROD ANTENNA, LARGE LOOP ANTENNA.

small signal A low-amplitude signal. Such a signal covers so small a part of the operating characteristic of a device that operation is essentially linear. Compare LARGE SIGNAL.

small-signal analysis Analysis of circuit or component operation in which it is assumed that the signals deviate from (fluctuate to either side of) the steady bias levels by only a small amount. Also see SMALL SIGNAL.

small-signal bandwidth The frequency at which the output signal of an analog circuit decreases to –3 dB, relative to the value for direct current. The output voltage is generally set at 0.1 volt peak-to-peak for testing this value so that the circuit is not overdriven.

small-signal component **1.** A coefficient or parameter (such as amplification, transconductance, dynamic resistance, etc.) calculated or measured under conditions of small-signal operation. Also see SMALL SIGNAL and SMALL-SIGNAL EQUIVALENT CIRCUIT. **2.** A device designed for operation at low signal levels.

small-signal diode See SIGNAL DIODE.

small-signal equivalent circuit For a given transistor circuit, the equivalent circuit for low signal levels (i.e., at amplitudes lower than saturation and cutoff levels). Also see EQUIVALENT CIRCUIT.

small-signal operation Operation at low signal amplitudes (i.e., at signal levels that do not extend into the saturation or cutoff levels of a transistor, diode, or other component).

small-signal transistor A transistor designed for low-level applications, such as the amplification of small voltages and currents and low-voltage switching. Compare POWER TRANSISTOR.

SmallTalk A high-level computer programming language that uses a graphical user interface (GUI). It is used in complex design and research, and in robotics.

smart home or business An electronically controlled home or business, in which computers and robots take care of cooking, dish washing, floor scrubbing, waste removal, laundry, yard maintenance, snow removal, and other mundane chores. Intrusion detection and fire prevention are constantly maintained. In some cases, intruders can actually be identified or detained; fires can be extinguished by controlled sprinkler systems or remotely controlled robots. Some tasks can be monitored and controlled directly by the owner from remote locations, via telephone lines or wireless.

smartness The ability of an electronic system, especially a computer or control system, to perform a complete series of operations, substituting alternative steps, where necessary—all with a minimum of instructions from, and supervision by, human operators.

smearing In television or facsimile, a form of picture distortion caused by an excessively narrow receiving bandpass. The image appears fattened and horizontally blurred. Contrast might also be lost.

smectic crystal A liquid crystal in which the molecules are arranged in parallel layers and cannot slide past each other.

S meter In a radio communications receiver, a meter graduated in S units and/or decibels to indicate the strength of a received signal.

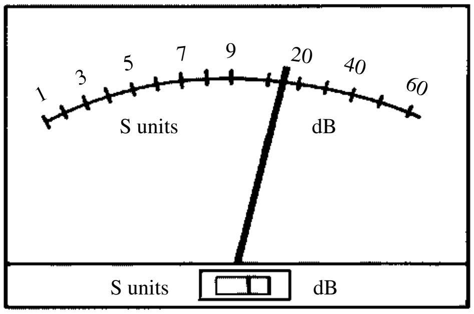

S meter

Smith chart A curvilinear graph on which complex-number impedance values can be plotted. It is useful in evaluating the behavior of radio-frequency circuits, transmission lines, and antenna systems—especially with regard to impedance mismatches and standing-wave ratio.

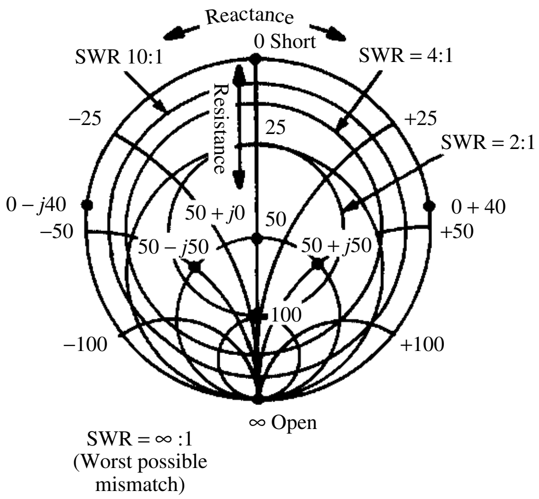

Smith chart

smoke alarm A device that produces audible and/or visible signals in the presence of smoke or unusual gases in the air. Also see PHOTOELECTRIC SMOKE ALARM.

smoke control See PHOTOELECTRIC SMOKE CONTROL.

smoke detector Any circuit or device used to sense the presence of smoke or noxious gases.

Some types detect changes in the ionization potential of the air; others sense changes in the dielectric constant of the air. Also see PHOTOELECTRIC SMOKE DETECTOR.

smoke sensor See SMOKE DETECTOR.

smooth **1.** Relatively free from surface irregularity. **2.** To reduce or eliminate irregularities in the voltage or current from a direct-current power source. **3.** To reduce or eliminate irregularities in data or signal amplitude.

smoothing choke A power-supply filter choke having a core with an air gap that prevents saturation at maximum rated direct current. Compare SWINGING CHOKE.

smoothing factor For a power-supply filter, a quantity approximately equal to $6.28fRC$, where f is the alternating-current frequency in Hertz, R is the filter resistance (in an RC filter) or the series reactance of the choke (in an LC filter) in ohms, and C is the filter capacitance in farads.

smoothing filter **1.** A filter for smoothing the alternating-current ripple component of a direct-current power supply following rectification. It can consist of one or more parallel capacitors of large value, and one or more series chokes of large inductance. **2.** A low-pass filter used at the output of a digital-to-analog (D/A) converter for eliminating high-frequency components generated by sampling.

SMPTE Abbreviation of SOCIETY OF MOTION PICTURE AND TELEVISION ENGINEERS.

smudge See SQUEEZEOUT.

SN Abbreviation of semiconductor network.

S/N Abbreviation of SIGNAL-TO-NOISE RATIO.

Sn Symbol for TIN.

snake **1.** A long, strong, flexible wire or strip used to pull other wires through electrical conduits. **2.** To route wires or cables through a group of circuits, components, or boards.

snap-action switch A switch that snaps quickly into the on or off position to prevent arcing and consequent premature contact deterioration.

snap diode A semiconductor diode in which switch-off time after carrier storage is extremely short (e.g., under 1 nanosecond).

snap magnet A magnet that reduces the tendency for arcing in relay-control instruments, thereby minimizing electromagnetic interference and prolonging contact life.

snap-on connector An electrical connector that locks in place, reducing the chance that it will detach unless it is deliberately removed.

snapshot dump During a computer program run, a dump, for debugging purposes, of certain storage areas.

snap switch See SNAP-ACTION SWITCH.

sneak current Unintended current flow through a path that is auxiliary to a main circuit.

sneak path A path through which current is accidentally detoured; it is usually a leakage path.

Snell's law A rule of physics that applies to visible light passing from air (or a vacuum) to some medium with an index of refraction c. If the light ray strikes the medium at an angle u, relative to a line normal (perpendicular) to the surface, and passes into the medium at angle v relative to the normal, then $(\sin u)/(\sin v) = c$.

sniffer See EXPLORING COIL.

sniperscope A telescope, snooperscope, or starlight scope for a carbine or rifle.

(S+N)/N Abbreviation of SIGNAL-PLUS-NOISE-TO-NOISE RATIO.

SNOBOL Acronym for *string-oriented symbolic language*, a computer-programming language for manipulating character strings.

snooperscope **1.** An infrared-sensitive device that permits viewing objects and surroundings in total darkness. It presents the image on a fluorescent screen. **2.** A rifle-mounted starlight scope.

snow A type of television picture interference that typically occurs when the signal-to-noise ratio is low (the reception is marginal or poor). Characterized by countless tiny out-of-focus light spots, whose rapid, random motion mimics the appearance of falling snow.

SNR Abbreviation of SIGNAL-TO-NOISE-RATIO.

soak value The smallest value of current that will cause saturation of a relay core.

Society of Motion Picture and Television Engineers Abbreviation, SMPTE. A group that decides on various procedures in video recording and reproduction, both on magnetic media (tape or disk) and on film.

socket A (usually female) fixture into which a plug, integrated circuit, or other component is inserted for easy installation in, or removal from, a circuit.

socket punch See PUNCH, **2**.

sodium Symbol, Na. A metallic element of the alkali-metal group. Atomic number, 11. Atomic weight, 22.9898.

sodium silicate See WATER GLASS.

sodium-vapor lamp A gas-discharge lamp containing neon and a small amount of sodium. After the filaments of the lamp are lighted for a short time, the heat vaporizes the sodium, and the filaments are disconnected by an automatic switch. Under the influence of the voltage across the lamp, the sodium vapor glows with a characteristic yellow light.

sofar A system for pinpointing the source of underwater sounds (coming from as far away as 2000 miles) through triangulation. The name is an acronym for *sound fixing and ranging.*

soft-drawn wire Wire that is highly malleable, and is therefore easily bent and unbent. Compare HARD-DRAWN WIRE.

soft iron A grade of iron, used in some cores, that is easily demagnetized.

soft solder A low-melting-point solder.

software **1.** Vendor-supplied or user-generated programs or groups of programs for a computer or computer system. **2.** The detailed instructions for performing a particular operation with a calculator or a computer.

software-defined radio Abbreviation, SDR. **1.** Wireless communications in which the modulation waveforms and protocols are generated and decoded by computer software. This allows a single, microcomputer-controlled radio receiver, transmitter, or transceiver to operate in a variety of services that use different protocols. Changing the modulation waveform or protocol requires only a change in the program run by a microcomputer that controls the radio. **2.** A receiver, transmitter, or transceiver that employs the technology defined in (1). **3.** The use of wireless equipment that employs the technology defined in (1).

soft X rays Low-frequency (long-wavelength) X rays. Such radiation has relatively poor penetrating power. Compare HARD X RAYS.

sol **1.** Abbreviation of SOLUTION. **2.** Abbreviation of SOLUBLE.

solar access For a specific property or location, the availability of direct exposure to the sun's rays as a source of energy.

solar absorption index A quantitative measure of the effect of the sun on the ionospheric absorption of radio waves.

solar activity See SOLAR RADIATION and SUNSPOT cycle.

solar battery A battery composed of solar cells connected in series and/or parallel for increased output.

solar cell A photovoltaic power transducer that converts visible light to electricity. It is called a cell because its output is a low direct-current voltage. Such cells can be connected in series and/or parallel to provide useful electric power output.

solar cycle See SUNSPOT CYCLE.

solar energy **1.** The total energy arriving from the sun, over a given period of time and within a specific surface area, at a given location on the surface of the earth. **2.** Any energy derived entirely from the sun.

solar-energy conversion Any process that changes solar radiant energy into another useful form.

solar flare A violent storm on the surface of the sun. These events tend to occur near the peak of the 11-year sunspot cycle. They cause an increase in the level of radio noise that comes from the sun, and they emit high-speed subatomic particles that reach earth a few hours after the first appearance of the flare. Because the particles are electrically charged, they are accelerated by the geomagnetic field. Sometimes a geomagnetic storm results, producing aurora near the poles and deterioration of ionospheric radio-propagation conditions.

solar flux An indicator of general solar activity. The solar noise level is measured at a particular frequency, such as 2800 MHz. The solar flux tends to be highest during periods of the greatest sunspot activity and immediately following a solar flare.

solar laser See SUNLIGHT-POWERED LASER.

solar noise Broadband electromagnetic noise generated by the sun. In particular, noise that occurs at radio frequencies and affects satellite, moonbounce, and space communications. This noise varies in intensity with the 11-year sunspot cycle, being generally higher at and near sunspot maxima. A solar flare can cause a sudden and dramatic increase in the noise at all wavelengths. See SOLAR FLARE.

solar panel An array consisting of a number of series-connected or series-parallel-connected solar cells mounted on a flat plate.

solar power Useful amounts of electricity obtained from suitable arrays of solar cells.

solar radiation Electromagnetic energy of various wavelengths originating in the sun. Such radiation, after passing through the earth's atmosphere, consists mostly of infrared rays and visible light. Some ultraviolet rays also reach the earth's surface.

solar relay See SUN SWITCH.

solar switch See SUN SWITCH.

solar wind Continuous emission of high-speed subatomic particles by the sun. It causes distortion of the lines of flux in the earth's magnetic field. It becomes more intense following a solar flare.

solder **1.** A metal alloy (usually of tin and lead) that is melted to electrically and mechanically join pieces of other metals. Also see HARD SOLDER and SOFT SOLDER. **2.** To join metals with solder.

solder

Solder type	*Melting point (°F/°C)*	*Principal uses*
Tin-lead 50 : 50, rosin-core	430/220	Electronic circuits
Tin-lead 60 : 40, rosin-core	370/190	Electronic circuits, low-heat
Tin-lead 63 : 37, rosin-core	360/180	Electronic circuits, low-heat
Tin-lead 50 : 50, acid-core	430/220	Non-electronic metal bonding
Silver	600/320	High-current, high-heat

soldering Joining (usually nonferrous) metal parts with solder, a lead-alloy substance. Compare BRAZING.

soldering gun An electric soldering iron having the general shape of a handgun. The element heats and cools more rapidly than the element in a typical soldering iron. The element is heated by pressing a device similar to the trigger on a pistol.

soldering iron An electric or nonelectric tool having a heated tip for melting solder.

solderless breadboard A foundation (see BREADBOARD, **1**) on which a circuit can be assembled by plugging components into tiny jacks without the use of solder.

solderless connection A connection between leads, or between leads and terminals, accomplished entirely through crimping, pinching, splicing, or wire wrapping. Solder is not used. Also see WIRE-WRAP CONNECTION.

solderless terminal A terminal to which a solderless connection can be made. Also see WRAP POST.

solenoid **1.** A coil of wire having a single layer, wound on a cylindrical form. **2.** A multilayer coil used as an electromagnet, and usually having a straight, iron core.

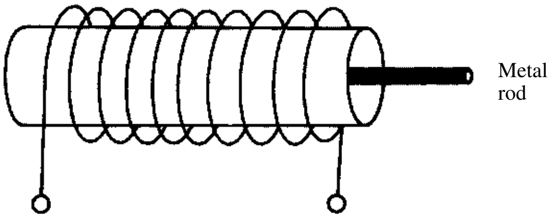

solenoid

solenoid switch A switch consisting of a solenoid coil (see SOLENOID, **2**) into which a core is pulled by the magnetic field to close a pair of contacts.

solid **1.** One of the states of matter. It is characterized by a definite shape and volume, and by atoms that maintain a fixed position, relative to each other. Compare GAS, LIQUID, and PLASMA. **2.** An enclosed, defined volume of three-dimensional space. **3.** In communications, descriptive of error-free reception of a series of coded signals. **4.** In printing and data transmission, a large print area whose entire surface is of equal and maximum intensity (of ink, light, or darkness).

solid angle Unit, steradian. The angle within the apex of a cone formed by all line segments between the center of a sphere and a defined circle on the surface of the sphere.

solid circuit Any circuit consisting of a single piece of hardware that is not normally separated into smaller parts.

solid conductor See SOLID WIRE.

solid electrolyte A solid substance affording ionic action similar to that in a liquid electrolyte.

solid electrolytic capacitor A capacitor using a solid electrolyte.

solid ground See DIRECT GROUND.

solid-state Pertaining to devices and circuits in which the flow of charge carriers (electrons and holes) is controlled in specially prepared blocks, wafers, rods, or disks of solid materials. Semiconductor devices, such as transistors and integrated circuits, are solid-state components.

solid-state battery An atomic battery consisting essentially of a photovoltaic cell in combination with a quantity of radioactive material, whose radiation causes the cell to generate electricity.

solid-state camera A video camera device that makes use of solid-state technology. The target is a matrix of charge-coupled devices (CCDs). When light strikes these devices, charge carriers are separated in a manner similar to that in a photovoltaic cell. The matrix is scanned according to a particular scheme, and the voltages developed in each CCD combine to produce the video signal output.

solid-state capacitor See SOLID ELECTROLYTIC CAPACITOR.

solid-state chronometer Any semiconductor device to indicate or measure time.

solid-state circuit See MONOLITHIC INTEGRATED CIRCUIT.

solid-state lamp **1.** See LIGHT-EMITTING DIODE. **2.** See LASER DIODE. **3.** See ELECTROLUMINESCENT CELL.

solid-state maser A device, such as the ruby maser, in which the stimulated medium is a solid material.

solid-state photosensor A semiconductor photodiode or phototransistor, as opposed to a phototube.

solid-state physics The branch of physics concerned with the nature and applications of such solids as electronic semiconductors.

solid-state relay **1.** A sensitive relay consisting of a conventional electromagnetic relay preceded by a transistorized amplifier. **2.** A completely electronic relay (i.e., one without moving parts) in which switching transistors provide the on and off states. **3.** See THYRISTOR.

solid-state thermometer An electronic thermometer utilizing one or more solid-state components, such as transistors, integrated circuits, or thermistors.

solid-state thyratron See SILICON-CONTROL-LED RECTIFIER and SILICON-CONTROLLED SWITCH.

solid-state tube A semiconductor device (diode, rectifier, transistor, SCR, etc.) whose housing and base allow it to replace directly an electron tube.

solid tantalum capacitor A capacitor using tantalum as a solid electrolyte.

solid wire Wire consisting of a single strand of metal. Compare STRANDED WIRE.

solute A substance that is dissolved in some other substance. Also see SOLUTION, **1**.

solution **1.** A well-diffused mixture of two or more substances. It can consist of a gas in a liquid, a gas in a solid, a gas in a gas, a liquid in a solid, or a solid in a solid. A solution, typically, is molecular (i.e., there is no chemical reaction between its constituents). Also see SATURATED SOLUTION; SOLUTE; SOLVENT, **1**; and SUPERSATURATED SOLUTION. **2.** The result of solving a problem or making a calculation. Also called *answer* or *result.*

solution conductivity The electrical conductivity of a solution, such as an electrolyte. The conductivity (and conversely, the resistance) depends on the number and mobility of ions in the solution.

solution-conductivity bridge A direct-current bridge specially designed and calibrated to measure the conductivity of chemical solutions.

solution pressure In an electrolyte into which a metal body is immersed, the force that causes the metal to tend to pass into solution as positive ions and to form a Helmholtz double layer.

solvent **1.** A fluid that dissolves other materials. **2.** The constituent of a solution that dissolves one or more other constituents. Thus, in a saltwater solution, water is the solvent and salt the solute. Also see SOLUTION, **1**.

SOM Abbreviation of *start of message.*

Sonalert Tradename for a small, but loud, sound reproducer used with solid-state circuits for alarm purposes.

sonar A system of detection and ranging by means of sonic and ultrasonic signals. In this system of echo ranging, the distance to an underwater object is determined from the time it takes a sound signal to reach the object and be reflected back to the transmitter. The name is an acronym for *sound navigation and ranging.*

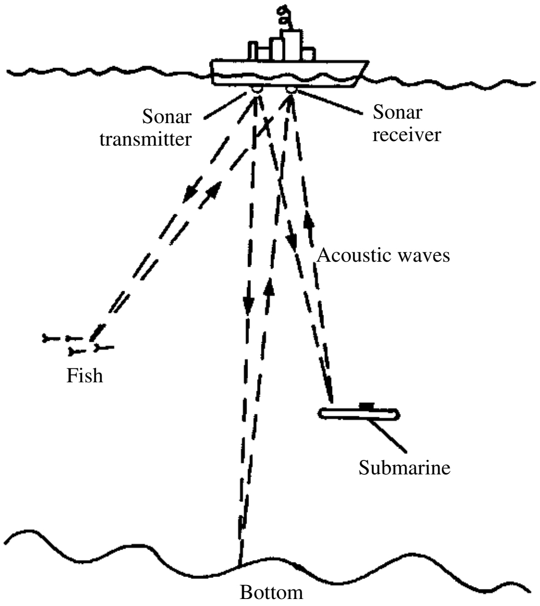

sonar

sonde A device for automatically gathering meterological data at high altitudes. An example is the *radiosonde.*

sone A unit of loudness for an individual listener. The level of 1 sone is the loudness of a 1000-Hz tone that is 40 dB above the particular listener's threshold of hearing.

sonic altimeter An altimeter (see ABSOLUTE ALTIMETER) using sound waves. The time required for a transmitted wave to reach a target and be reflected back to the transmitter is proportional to the distance between the transmitter and the target.

sonic boom An explosive sound occurring when the shock wave produced by an aircraft flying at supersonic speed strikes the earth.

sonic delay line A delay line using electroacoustic transducers and an intervening medium through which a sound wave is transmitted.

sonic depth finder See ACOUSTIC DEPTH FINDER.

sonic thermocouple A thermocouple whose heat-absorbing properties are enhanced by subjecting it to acoustic vibrations.

sonobuoy A buoy equipped with an acoustic receiver and radio transmitter. The device is parachuted into the water, where it receives submarine sounds and transmits them to a monitoring station. Several sonobuoys communicating with a computer will track the path of a submarine.

sonovox An electronic device used to produce special sound effects when it is held against the throat of the operator. The special sounds are formed into words by the operator's mouth.

SOP Abbreviation of *standard operating procedure.*

sophisticated electronics Advanced electronics theory and operations, usually dealing with complex devices or systems and requiring rigorous analysis to describe their operation and devise applications. Compare UNSOPHISTICATED ELECTRONICS.

sorption processes Processes whereby certain substances (e.g., activated charcoal) occlude and retain gases and vapors. A chamber containing such a substance is often useful in the production of a vacuum. Sorption includes both absorption and adsorption.

sort **1.** To group information items using their keys. Also see KEY. **2.** To group information items according to some system of classification, as to print an alphabetical list of words stored in a random sequence.

sorting routine A computer program for sequencing data items according to key words (values in specific fields) of the different records.

SOS **1.** The international radiotelegraph distress signal; equivalent to *mayday* in radiotelephony. It consists of three dits ("dots"), followed by three dahs ("dashes"), followed by three more dits. **2.** Abbreviation of SILICON ON SAPPHIRE.

sound The vibratory or wave phenomenon to which the sense of hearing is responsive. Conducted by waves in solids, liquids, and gases; not propagated through a vacuum.

sound absorption The nonreflection and nontransmission of acoustic energy by a body or medium, and the attendant conversion of the acoustic energy into another form of energy (usually heat).

sound absorption coefficient A quantitative expression of the extent to which a surface absorbs acoustic energy (as opposed to reflecting, transmitting, diffusing, or scattering it).

sound amplifier **1.** An audio amplifier—especially the sound channel of a television system. **2.** A device, such as a horn or reflector, that directly boosts the intensity of sound at a given listening point.

sound analyzer An instrument, often a wave analyzer equipped with a microphone, for measuring such characteristics of sound as amplitude, frequency (pitch), and harmonic content (timbre).

sound articulation See ARTICULATION.

sound bars In a television picture, horizontal bars resulting from interference between the audio and video channels of the receiver.

sound carrier In a television signal, the frequency-modulated carrier that transmits the audio part of the program. Compare VIDEO CARRIER.

sound chamber An air enclosure, usually a box or can, for modifying the acoustic qualities of sound or of an audio signal.

sound detector The discriminator or ratio detector that demodulates the sound signal in a television receiver circuit.

sound-energy density Sound energy per unit volume, expressed in joules per cubic meter or ergs per cubic centimeter.

sound-energy flux The average rate of flow of sound energy through a specified area, as expressed in ergs or joules per second.

sound field A volume of space or material containing sound waves.

sound film Motion-picture film on which a sound track is recorded. Also see OPTICAL SOUND RECORDING.

sound flux The rate of flow of sound energy, usually expressed in terms of sound pressure at a point or over a unit area normal to the direction of sound propagation.

sound gate An optical device used to convert the sound track of a movie film into electrical impulses.

sound generator Any combination of oscillator, amplifier, and transducer (loudspeaker or headphones) for producing sound waves.

sound-hazard integrator An instrument used to measure cumulative noise exposure received by persons in noisy environments. One such instrument provides direct readings in percent of permissible exposure.

sound IF amplifier In a television receiver circuit, the separate amplifier for the sound intermediate frequency. See, for illustration, INTERCARRIER RECEIVER and SPLIT-SOUND RECEIVER.

sound-level meter See SOUND SURVEY METER.

sound marker A marker indicating the sound-carrier point on a television alignment curve displayed on an oscilloscope screen.

sound-marker generator A special radio-frequency signal generator (or a special circuit in a television-alignment generator) for the production of a sound marker.

sound mirage See ACOUSTIC MIRAGE.

sound mix In sound recording or reproduction, the composite output from an audio mixer circuit.

sound-on-film recording See OPTICAL SOUND RECORDING.

sound-on-sound recording The simultaneous recording (on a single track on magnetic tape) of new material with previously recorded material. The old recording is not erased.

sound-operated relay **1.** A relay operated indirectly from sound, through the medium of a pickup microphone and amplifier. **2.** A relay having a delicately poised armature that operates directly from sound vibrations.

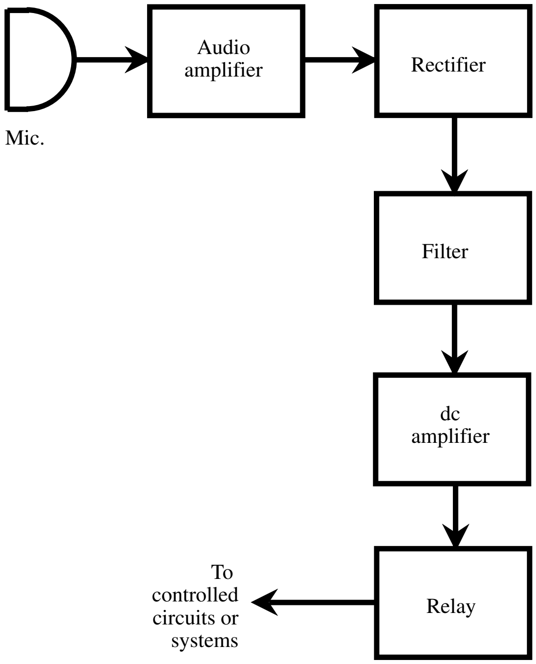

sound-operated relay

sound power The total sound energy per unit time produced by a sound source, as expressed in ergs per second or in watts.

sound power level The extent, in decibels, by which SOUND POWER exceeds one picowatt (10^{-12} watt).

sound pressure **1.** The force exerted by sound waves on a surface area, expressed in dynes per square centimeter (as an rms value over 1 cycle). The sound pressure is proportional to the square root of sound-energy density. **2.** The instantaneous difference between actual air pressure and average air pressure at a given point.

sound pressure level The extent, in decibels, via which SOUND PRESSURE exceeds 20 micropascals (2.0×10^{-5} pascal).

sound probe A transducer used to receive acoustic vibrations for detection or measurement purposes.

sound recording The electrical recording of sound, using cylinder, disc, tape, wire, or other comparable storage medium.

sound-recording system A complete, integrated array of equipment for recording sound, including such components as microphones, amplifiers, pickups, filters and other shaping networks, attenuators, level indicators, and recording mechanisms. Compare SOUND-REPRODUCTION SYSTEM.

sound reinforcement Intensification of sound by horns, resonant chambers, or other acoustical devices.

sound relay See SOUND-OPERATED RELAY.

sound reproduction The electrical reproduction of sound from recordings on vinyl discs, magnetic tapes, magnetic discs, compact optical disbs, etc.

sound-reproduction system A complete, integrated array of equipment for the playback of recorded sound, including such components as tape or record players, amplifiers, filters and other shaping networks, attenuators, level indicators, loudspeakers, and headphones. Compare SOUND-RECORDING SYSTEM.

sound spectrograph A device that produces a display of sound amplitude vs. frequency. Similar to a SPECTRUM ANALYZER, except that it operates at audio frequencies (about 20 Hz to 20 kHz) and is actuated by acoustic disturbances, rather than by electromagnetic signals.

sound spectrum The continuous band of frequencies (about 20 Hz to 20 kHz) constituting audible sounds, and sometimes the immediately adjacent (subaudible and superaudible) frequencies.

sound stage The apparent dimensions of a sound source.

sound survey meter A portable instrument for measuring the intensity and other characteristics of sound.

sound sweetening In audio recording or reproduction, the modification of the sound to achieve some desired effect.

sound system A sound-recording system, sound-reproduction system, or a combination of the two.

sound takeoff In a television receiver circuit, the point at which the frequency-modulated sound signal is extracted from the complex signal.

sound track The variable-density or variable-width recording on one side of the film in sound-on-film recording and reproduction. Also see OPTICAL SOUND RECORDING.

sound transducer See ACOUSTIC TRANSDUCER.

sound-transmission coefficient See ACOUSTICAL TRANSMITTIVITY.

sound trap In a television receiver circuit, a wavetrap that prevents the sound signal from entering the picture channels.

sound unit See PHONE, **2**.

sound wave The vibratory phenomenon produced in a medium by acoustic energy. A sound wave in air consists of alternate compressions and rarefactions of the air. Also see ACOUSTIC WAVE.

source **1.** The origin of a signal or electrical energy (e.g., a transmitting station). **2.** In a field-effect transistor, the electrode that is equivalent to the emitter of a common-emitter-connected bipolar transistor, or the cathode of a vacuum tube. **3.** That which is being transcribed to magnetic tape. **4.** Manufacturer, wholesaler, or retailer.

source circuit **1.** The circuit associated with the source electrode of a field-effect transistor. **2.** A generator circuit. Compare SINK CIRCUIT.

source code See SOURCE LANGUAGE.

source computer A computer for compiling a source program.

source-coupled multivibrator A multivibrator circuit using field-effect transistors, in which feedback coupling is achieved with a common source resistor for the two FETs. This circuit is equivalent to the emitter-coupled bipolar-transistor-type multivibrator.

source data automation A means of storing a master data file for easy duplication, whenever necessary.

source deck An audio or video tape player that reproduces original recordings in the editing process.

source follower A field-effect-transistor circuit in which the output is taken across a resistor between source and ground. This circuit is equivalent to the emitter follower, and is a unity-gain stage whose impedance-transformation characteristics make it ideal in signal conditioning, buffering, and impedance-matching applications.

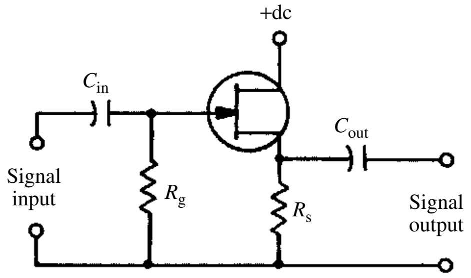

source follower

source impedance **1.** The impedance of a generator in a circuit. **2.** The impedance of the source electrode of a field-effect transistor.

source language A computer programming language from which is derived (by a compiler) the machine (object) language on which the computer operates.

source program A computer program written in a source language.

south magnetic pole The south pole of the equivalent bar magnet constituted by earth's magnetic field (see EARTH'S MAGNETIC FIELD). The south magnetic pole lies close to the geographic south pole. Compare NORTH MAGNETIC POLE.

south-seeking pole Symbol, S. The so-called south pole of a magnet. When a bar magnet is suspended horizontally, this pole points to earth's south magnetic pole. Compare NORTH-SEEKING POLE.

SP Abbreviation of STACK POINTER.

sp **1.** Abbreviation of *single pole.* **2.** Abbreviation of *special.*

space-charge field **1.** The electric field existing within a group of charged particles. **2.** The electric field existing in a plasma.

space diversity See DIVERSITY RECEPTION.

space-diversity reception See DIVERSITY RECEPTION.

space division A method of data transfer in which different paths are used for the transmission of different signals.

space-division switch A switch having two or more ports and different paths connecting the ports.

space lattice The three-dimensional, redundant pattern formed by atoms and molecules in a crystal and having a shape that is characteristic of a particular crystalline material.

spacer An insulating rod or bar that serves to hold apart the conductors of a two-wire, four-wire, or coaxial air-dielectric transmission line.

space suppression Following the printing of a line by a printer, the prevention of normal paper travel.

space-time-space switch Abbreviated STS switch or STSS. A large switching array with two space switch blocks and a time switch block between them.

space wave One of the components of an electromagnetic ground wave. The space wave, unlike the *surface wave*, is not earth-guided. It has two components: the *direct wave* and the *ground-reflected wave.*

spaghetti Slender, varnished-cambric tubing used as slipover insulation for wires and busbars.

span On an instrument scale, the difference between the highest value and the lowest value.

spark See ELECTRIC SPARK.

spark absorber **1.** See SPARK SUPPRESSOR. **2.** See KEY-CLICK FILTER.

spark coil A small induction coil. Its name is derived from its initial purpose of supplying the high voltage for spark plugs in gas engines.

spark energy The energy dissipated by an electric arc or spark.

spark gap A device consisting essentially of two metal points, tips, or balls that are separated by a small air gap. A high voltage applied to the electrodes causes a spark to jump across the gap.

sparking distance The maximum separation of the electrodes of a spark gap at which a given voltage will produce a spark.

sparking voltage The lowest voltage that will cause a spark to jump across a gap of a given width.

spark killer See SPARK SUPPRESSOR.

sparkover A discharge in air, a vacuum, or a dielectric. It is characterized by sparking between electrodes in the medium.

spark plate In some automobile radios, a noise-interference-eliminating bypass capacitor in which the chassis is one plate.

spark-plug suppressor A small resistive device connected in series with a spark plug to suppress electrical noise arising from the ignition in an internal combustion engine.

spark quencher See SPARK SUPPRESSOR.

spark suppressor A resistor, capacitor, and/or diode used to eliminate or minimize sparking between make-and-break contacts.

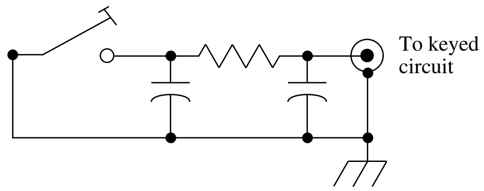

spark suppressor

spark-suppressor diode See SUPPRESSOR DIODE.

spatial distribution The three-dimensional directional pattern of a transducer (such as an antenna, microphone, or speaker).

spc **1.** Abbreviation of *silicon point contact.* **2.** Abbreviation of *silver-plated copper.*

SPDT Abbreviation of SINGLE-POLE/DOUBLE-THROW (switch or relay).

speaker See LOUDSPEAKER.

speaker damping See DAMPED LOUDSPEAKER.

speaker-level audio In a sound reproduction system, radio receiver, or other audio circuit, an audio-frequency signal of sufficient amplitude to drive a speaker or speaker system directly, without the need for additional amplification.

speaking arc A method of modulated-light transmission. An electric arc is modulated by audio-frequency signals.

special character A printed, displayed, or encoded character other than a numeral or letter, such as an ampersand (&) or a pound sign (#). Also called SYMBOL.

special effects Various techniques used in filmmaking, computer animation, and videotape recording for achieving certain visual scenes or images.

special-purpose computer A computer designed to handle problems or be suitable for applications of a specific category; a dedicated computer.

special-purpose calculator An electronic calculator intended for essentially "nonmathematical" purposes, such as biorhythm data, astrological information, metric conversions, musical composing, etc.

specific address See ABSOLUTE ADDRESS.

specification **1.** For an electronic device, a statement of performance over specific parameters. Example: for a high-fidelity stereo amplifier, 50 watts per channel over a frequency range of 10 Hz to 30 kHz, with less than 1 percent total harmonic distortion. **2.** A precise listing of requirements or expectations.

specific conductivity Conductance per unit volume. In SI units, this is expressed in siemens per cubic centimeter (S/cm^3).

specific dielectric strength For an insulant, the dielectric strength per millimeter of thickness.

specific gravity Abbreviation, sp gr. The ratio of the density of a material to the density of a substance accepted as a standard (usually water at 4 degrees Celsius or 39.2 degrees Fahrenheit).

specific inductive capacity See DIELECTRIC CONSTANT.

specific resistance See RESISTIVITY.

specific sound-energy flux See SOUND INTENSITY.

spectral comparative pattern recognizer Acronym, SCEPTRON. Equipment used to classify automatically complex signals obtained from information that has been converted into electrical signals.

spectral density For a complex signal, the amount of energy contained within a given band of frequencies.

spectral energy distribution The occurrence of different amounts of energy in different areas of a spectrum (as in a visible-light spectrum, sound spectrum, or radio-frequency spectrum).

spectral response The characteristic of a device, such as a photocell or the human eye, that describes the device's sensitivity to radiations of various frequencies in a given spectrum.

spectral sensitivity The color response of a light-sensitive device.

spectrograph A recording SPECTROMETER.

spectrometer **1.** An instrument used to measure spectral wavelengths. **2.** An instrument used to measure the index of refraction. **3.** See MASS SPECTROMETER.

spectrophotometer A photoelectric instrument for chemical analysis. In the device, light passing through the material under analysis is broken up into a spectrum that is examined with a photoelectric circuit, which, in turn, plots a spectrogram.

spectroscope An instrument that resolves a radiation into its various frequency components and permits measurement of each.

spectrum A band of frequencies or wavelengths (e.g., *radio spectrum*, *visible-light spectrum*, and *sound spectrum*).

spectrum analyzer **1.** An automatic wave analyzer with a visual display (oscilloscope). **2.** A scanning receiver with a screen that shows a plot of signals and their bandwidths over a specific frequency band.

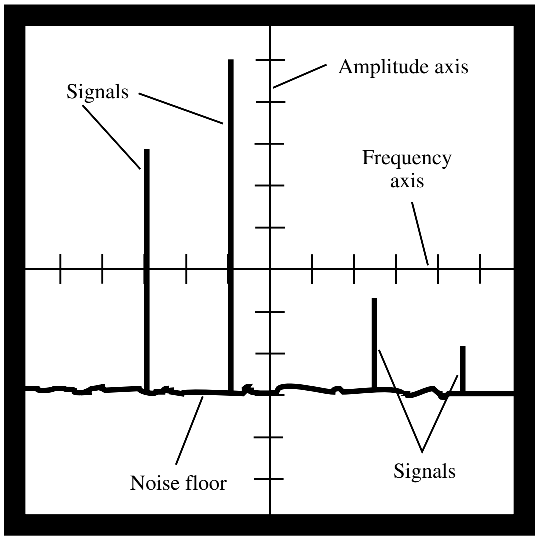

spectrum analyzer, 2.

specularity A qualitative or experimentally derived expression of the efficiency with which a DIFFRACTION GRATING works.

specular reflection Reflection in which the reflected ray is in the same plane as the incident ray, as in reflection from an extremely smooth surface.

speech amplifier A (usually low-level) audio amplifier designed especially for speech frequencies. It is generally used to amplify the signals from a microphone.

speech clipper A device, such as a biased diode, (see POSITIVE PEAK CLIPPER) for holding speech signals to a constant amplitude. Compare SPEECH COMPRESSOR. Also see SPEECH CLIPPING.

speech clipping The use of a limiting circuit to maintain the output-signal amplitude of a speech amplifier against fluctuations in the intensity of speech input. The resulting signal requires filtering to remove harmonics generated by the process. Compare SPEECH COMPRESSION.

speech compression Automatic regulation of the gain of a speech amplifier to maintain its output-signal amplitude against speech-input fluctuations. Compare SPEECH CLIPPING.

speech compressor A circuit or device, such as an automatic-gain-control (agc) system, for performing speech compression. Compare SPEECH CLIPPER.

speech digit signaling A method of digital signaling, where time slots generally used for encoded audio or video are used alternately for signaling.

speech frequencies See VOICE FREQUENCIES.

speech intelligibility The quality of reproduced speech that makes it easily understood by a reasonably proficient user of the language. For good speech intelligibility, a circuit should transmit frequencies between 300 Hz and 3000 Hz with minimal distortion. Increased bandwidth improves fidelity, but does not provide a significant increase in intelligibility for normal speech.

speech inverter See SCRAMBLER CIRCUIT.

speech power **1.** The alternating-current power in an electric wave corresponding to speech, as opposed to that in a sine wave. **2.** Sound power in a speech transmission.

speech recognition The ability of a device to translate audible spoken words, phrases, or sentences into binary digital signals that can be used by machines, such as computers and robots.

speech recognizer An electronic device that translates audible spoken words, phrases, or sentences into binary digital signals that can be used by machines, such as computers and robots.

speech scrambler See SCRAMBLER CIRCUIT.

speech synthesis The ability of a device to translate binary digital signals from a machine, such as a computer or robot, into audible, coherent spoken words, phrases, or sentences.

speech synthesizer An electronic device that translates binary digital signals from a machine, such as a computer or robot, into audible, coherent spoken words, phrases, or sentences.

speed key **1.** A semiautomatic key for manual generation of Morse code characters. **2.** A similar, fully automatic key used especially for high-speed telegraphy.

speed of light Symbol, *c.* The speed at which electromagnetic waves propagate in a vacuum; approximately 299,792 kilometers per second (186,282 miles per second).

speed of sound The speed at which acoustic waves propagate. It depends on the nature of the medium. In air, at ordinary temperatures, it is approximately 344 meters (1129 feet) per second. In fresh water, it is approximately 1463 meters (4800 feet) per second.

speed of transmission The amount of data sent in a given unit of time. It is generally measured in

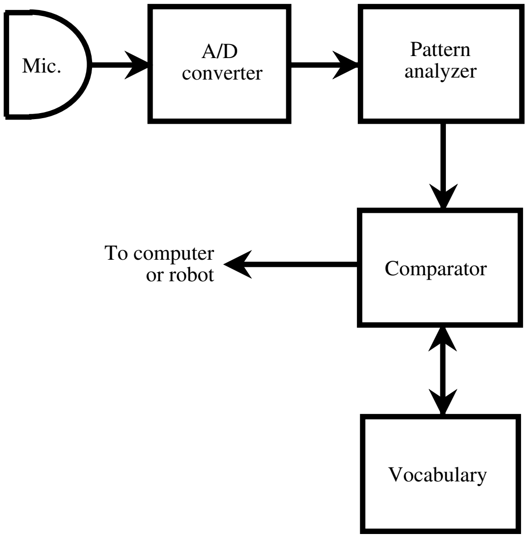

speech recognizer

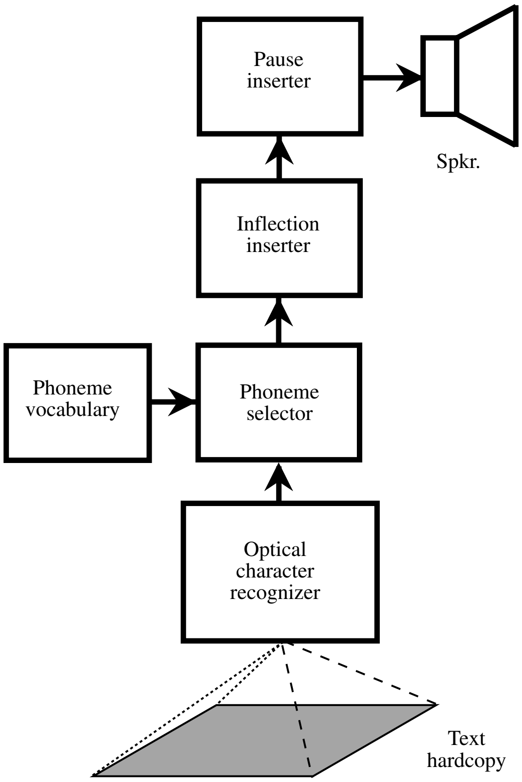

speech synthesizer

bits per second (bps), characters per second (cps), characters per minute (cpm), or words per minute (wpm). It is used primarily for digital codes.

speedup capacitor See COMMUTATING CAPACITOR.

SPFW Abbreviation of SINGLE-PHASE FULL-WAVE.

sp gr Abbreviation of SPECIFIC GRAVITY.

sphere **1.** A closed surface in three-dimensional space, represented by the set of all points equidistant from a specified center point. **2.** A solid in three-dimensional space, represented by the set of all points on or within a closed surface, as defined in **1**.

sphere gap A spark gap in which the spark passes between two polished metal spheres. When the air gap is adjustable, unknown high voltages can be measured in terms of the largest gap width at which sparking occurs. Compare NEEDLE GAP.

sphere-gap voltmeter See SPHERE VOLTMETER.

sphere voltmeter A gap voltmeter using a SPHERE GAP.

spherical aberration In a spherical lens, mirror, or reflecting dish, distortion as a result of the spherical (as opposed to paraboloidal) shape of the surface. This causes the focus to be elongated into a short line segment along the principal axis.

spherical angle An angle formed by the intersection of two arcs on the surface of a sphere.

spherical coordinate geometry A scheme for guiding a robot arm in three dimensions via SPHERICAL COORDINATES. The length (radius) of the arm can be varied, as can the elevation (latitude) and azimuth (longitude).

spherical coordinates A method of defining a point (*P*) in three-space using two angles (latitude and longitude) and a radial distance (*r*) from the origin.

spherical degree A unit equal to $\frac{1}{720}$ of the surface area of a sphere.

spherical distance The length of the shortest arc (lying on a great circle) connecting two specified points on a sphere.

spherical divergence The manner in which energy normally propagates from a fixed point source in three dimensions. Wavefronts expand from the source in the form of spheres, whose centers are at the point source.

spherical reflector A microwave reflector (dish) whose contour is that of a sphere, rather than that of a paraboloid.

spherical wave A wave characterized by wavefronts that are concentric spheres.

spheroidal antenna A sheetmetal or wire-mesh antenna having the cross section of a sphere that is flattened at the ends of one axis.

SPHW Abbreviation of SINGLE-PHASE HALF-WAVE.

spider **1.** The flat, round, springy part that holds the apex of the vibrating cone of a dynamic speaker. **2.** A quickly assembled, chassisless hookup in which the pigtails of components form the wiring and the mechanical support for a circuit.

spiderweb antenna A set of dipole antennas for different frequencies, arranged in a common (usually horizontal) plane. The result is a broadband antenna.

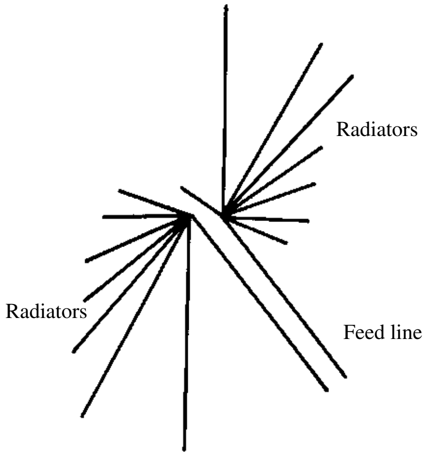

spiderweb antenna

spiderweb coil A flat, single-layer coil in which a strand of wire is woven over and under the spokes of a wheel-like form, and having the general appearance of a spiderweb. The criss-cross winding reduces distributed capacitance by breaking up the parallelism of adjacent turns.

spike **1.** A current or voltage pulse of extremely short duration. **2.** A sharp transient, such as an overshoot on a pulse or square wave.

spike suppressor A clipper or similar device for removing a spike from a signal voltage.

spin A quantity of angular momentum possessed by a subatomic particle. It can be positive, negative, or zero, but it always exists in integral multiples of ½.

spindle **1.** The pivoted shaft that carries the movable element and rotates between the pivots in a meter movement. **2.** The rotor of an alternator—especially when the rotor is a permanent magnet. **3.** The rotating shaft in a motor, generator, or similar electric machine.

spinning electron An electron having nonzero angular momentum.

spinthariscope An optical device for observing the effect of alpha particles emitted by a radioactive substance, from the scintillations they produce upon striking a small, fluorescent screen.

spiral coil See DISK WINDING.

spiral distortion A form of television camera-tube distortion caused by spiraling of the electrons as they are emitted from the photocathode. The result is a twisted image.

spiral-rod oscillator A parallel-line-type oscillator in which the lines are rods that are rolled up into spirals to conserve space. Also see LINE-TYPE OSCILLATOR.

spiral sweep **1.** A means of sweeping the electron beam in a cathode-ray tube to produce a spiral trace on the screen. **2.** The circuit for producing such a sweep.

spiral trace See SPIRAL SWEEP.

spiral winding See DISK WINDING.

spkr Abbreviation of *speaker* (see LOUDSPEAKER).

splatter See SIDEBAND SPLATTER.

splatter-suppression filter In an amplitude-modulated (AM) or single-sideband (SSB) radio transmitter, a low-pass filter inserted between the output of the audio amplifier and the audio input of the modulator. It suppresses high-frequency audio components that would otherwise cause sideband splatter.

splaying The construction of a room or auditorium so that the walls do not meet at right angles. It is useful in optimizing the acoustic characteristics of the enclosure because it tends to reduce acoustic resonant effects.

splice **1.** A physical or electrical connection between two wires, made by tightly winding the ends together. Solder is often applied for additional mechanical strength and electrical continuity. **2.** A physical connection between two lengths of magnetic tape, made in such a way as to cause minimal disturbance in reproduced audio, video, or data. **3.** To prepare a joint, as defined in **1** or **2**.

splicer A device for making a SPLICE.

splicing block A device specifically designed to facilitate easy splicing of audio, video, or digital magnetic tape.

splicing tape A durable, flexible adhesive designed to hold a magnetic-tape splice together.

spline-based modeling In video animation and advanced computer graphics, a scheme that uses curve-generating formulas (e.g., Bezier curves) to create lifelike images.

split-anode magnetron A magnetron in which the plate (anode) consists of two semicylinders with the cathode at their axis.

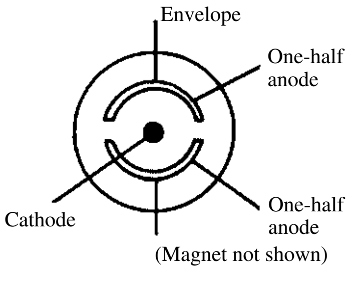

split-anode magnetron

split-load phase inverter See PARAPHASE INVERTER.

split-phase motor A fractional-horsepower, alternating-current motor that starts like a two-phase motor and runs like a single-phase motor. After the machine approaches approximately 75 percent of full speed, a centrifugal switch cuts out the starting winding.

split projector An acoustic transmission device with several independently operated transducers.

split-reed vibrator See SELF-RECTIFYING VIBRATOR.

split-rotor plate See SERRATED ROTOR PLATE.

split-stator capacitor A variable capacitor having two separate stator sections and a common rotor section.

split-sound receiver A television receiver circuit in which the sound signal is separated from the composite signal shortly after pickup by the antenna, and is amplified separately from the video signal. Compare INTERCARRIER RECEIVER.

splitter A device used to couple two or more television receivers to a single antenna.

spool See REEL.

sporadic-E layer ionization Occasional, scattered ionization in the E-layer of the ionosphere.

sporadic-E propagation At certain radio frequencies, the long-distance propagation of electromagnetic (EM) waves via the E layer of the ionosphere. This layer exists at an altitude of approximately 50 miles (80 km) above sea level. This mode of propagation tends to be intermittent, and conditions can change rapidly. It is most likely to occur between 20 MHz and 150 MHz. Occasionally, it is observed at frequencies as high as 200 MHz. The propagation range is normally several hundred miles, but occasionally can occur over distances of 1,000 to 1,500 miles. The standard frequency modulation (FM) broadcast band is sometimes affected by this propagation. The same is true of the lowest television (TV) broadcast channels, especially channels 2 and 3. See IONOSPHERE.

SPOT Abbreviation of *satellite position and tracking.*

spot brightness In a cathode-ray tube, the relative brilliance of the glowing dot or trace produced on the screen by the electron beam.

spot check **1.** To take a random sample or to make a random test by arbitrarily selecting a single item from a run of similar items and subjecting it to analysis, examination, performance, or parametric evaluation, etc. **2.** A random sample or test.

spot frequency **1.** A single frequency or signal to which other frequencies can be referred. **2.** A frequency or signal that acts as a limit marker (e.g., to define the edges of an allocated frequency band).

spot jamming Deliberate interference to a radio signal at some frequency and at some specific time.

spot modulation In a cathode-ray tube, modulation of the brightness of the spot (and, accordingly, of the image) produced on the screen by the electron beam. Also see INTENSITY MODULATION.

spot welding A method of electrical welding in which the parts to be joined are held together, overlapping, between the points of two electrodes, between which a current is passed to heat the parts at the spot of contact.

spot-wheel pattern A frequency-identifying wheel pattern produced on an oscilloscope screen by intensity-modulating a circular trace. The circular trace is produced by applying a standard-frequency signal to the horizontal and vertical input terminals 90 degrees out of phase. A square-wave signal of unknown frequency is applied to the intensity-modulation (z-axis) input terminals. The square wave chops the circle into a number of bright sectors or spots. The unknown frequency (fx) equals Nfs, where N is the number of spots in the circle, and fs is the standard frequency. Compare GEAR-WHEEL PATTERN.

sprat Acronym for *small portable radar torch.* A portable radar unit that uses a Gunn diode to generate microwave energy. The range is about 500 meters (⅛ mile).

spray coating **1.** Applying a protective coat of insulating material to a conductor or component by spraying it with a liquid substance and allowing it to dry. Compare DIP COATING. **2.** The coat applied, as defined in **1**.

spreader **1.** An insulator used to separate the wires of an air-spaced transmission line. **2.** Any of the rods composing the supporting structure in a cubical quad antenna.

spreading current In a semiconductor, current caused by the movement of charge carriers by circuitous routes, that is, in paths that deviate significantly from straight lines.

spreading loss Energy lost during the transmission of radiation.

spreading resistance In a semiconductor device, the resistance that is a consequence of electrical paths through material that is not along straight lines between electrodes.

spread spectrum **1.** A method of transmission in which the occupied bandwidth of the signal is deliberately increased, or spread out, over a much wider range than it would normally occupy with conventional modulation. **2.** A signal transmitted, as defined in **1**.

spring coil See SOLENOID, **1**.

spring contact See FLEXIBLE CONTACT.

sprite In video and animated computer graphics, a brief insert, such as a little insect that scurries across the screen, or a face that pops in and smiles. It is used primarily for effect.

SPST Abbreviation of SINGLE-POLE/SINGLE-THROW (switch or relay).

spurious emission From a radio or television transmitter, an unintended and unwanted output signal on a frequency other than the fundamental signal frequency. It can be generated by faulty modulation, amplification, and/or oscillation.

spurious oscillation **1.** Oscillation in a normally nonoscillatory circuit. **2.** In an oscillator, simultaneous oscillation at a frequency other than the normal one.

spurious response In a communications receiver, a signal that appears to be on a certain frequency, when, in fact, the received signal is not on that frequency. It often results from inadequate image rejection.

spurious-response ratio The ratio of the transmission (or gain) of a circuit of a desired signal to its transmission (or gain) for a spurious signal at the same setting of the circuit (e.g., *signal-to-image ratio*).

spurious sidebands In an amplitude-modulated (AM) or single-sideband (SSB) radio signal, sideband energy at frequencies outside the nominal signal band, usually resulting from improper modulation, inadequate filtering, improper envelope clipping, or nonlinear amplification.

Sputnik The first orbiting artificial earth satellite. It was launched by Russia (then known as the Union of Soviet Socialist Republics) in 1957.

sputter **1.** A layer of material obtained by sputtering (see SPUTTERING, **1**). **2.** To carry out the process of sputtering (see SPUTTERING, **1**).

sputtering **1.** A technique for electrically depositing a film of metal on a metallic or nonmetallic surface. In a vacuum chamber, the piece of metal to be deposited is made the cathode of a high-voltage circuit, with respect to a nearby anode plate. The high voltage causes atoms to be ejected from the surface of the cathode and strike the surface of an object placed in their path, becoming deposited on it as a film of cathode metal. Compare EVAPORATION, **1**. **2.** The disintegration of a vacuum-tube cathode through ejection of surface atoms from the cathode by impinging positive ions.

sq Abbreviation of SQUARE.

S$_Q$ band A section of the S BAND, from 2400 to 2600 MHz.

SQC Abbreviation of STATISTICAL QUALITY CONTROL.

SQR **1.** Abbreviation of SQUARE ROOTER. **2.** In the BASIC computer-programming language, a function that computes the square root of a positive number.

S quad See SIMPLE QUAD.

square-law demodulator See SQUARE-LAW DETECTOR.

square-law detector A detector whose output is proportional to the square of the root-mean-square (rms) value of the input. Also called WEAK-SIGNAL DETECTOR.

square-law meter **1.** A meter whose deflection is proportional to the square of the quantity applied to it. Also see CURRENT-SQUARED METER. **2.** A high-impedance electronic voltmeter, whose deflection is proportional to the square of the applied voltage.

square-law modulator A circuit or device that accomplishes amplitude modulation of one signal current by another, by simultaneously passing the two currents through a component, such as a nonlinear resistor, having a square-law response.

square-law response Circuit or component operation that results in an output signal, proportional to the square of the input.

square rooter An analog or digital device used to extract the square root of a number.

square wave An alternating or pulsating current or voltage whose rise and decay times are essentially zero, and whose maxima and minima are essentially flat. The duration of the maxima is equal to the duration of the minima. A special form of RECTANGULAR WAVE.

square-wave amplifier An amplifier designed to operate with square waves.

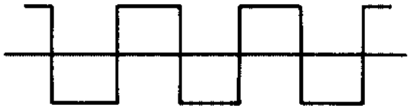

square wave

square-wave converter See SQUARING CIRCUIT.

square-wave generator A signal generator delivering an output signal that has a square waveform. Compare SQUARING CIRCUIT, **1**.

square-wave testing Testing the response of a circuit or device, such as an amplifier, by observing the extent to which it distorts a square-wave signal passing through it (a measure of high-frequency response).

squaring circuit **1.** A circuit (such as a twin-diode clipper, overdriven amplifier, or Schmitt trigger) that converts a sine wave or pulse into a square wave. **2.** A circuit whose instantaneous output-signal amplitude is equal to the square of the instantaneous input-signal amplitude.

squarish wave **1.** A signal whose oscilloscope trace is nearly, but not exactly, the same as that of a square wave. **2.** A rectangular wave that is not a square wave; that is, whose maxima and minima are not of the same duration. See RECTANGULAR WAVE and SQUARE WAVE.

squawker **1.** In a three-way speaker system, the midrange speaker. **2.** Any slave station in a multistation intercom network.

squeal A high-pitched interferential sound, such as that encountered in spuriously oscillating systems.

squeezeout In optical character recognition (OCR), a condition in which errors occur because the printed characters have excessive ink at the edges. Also called *smudge.*

squegging A choking-type cutoff action in a circuit caused by an excessively strong signal.

squegging oscillator An oscillator that starts and stops oscillating intermittently as a result of SQUEGGING.

squelch See SQUELCH CIRCUIT.

squelch amplifier An amplifier that can be controlled by a squelch signal. Also see SQUELCH CIRCUIT.

squelch circuit One of several circuits that automatically disable a receiver or amplifier, except when incoming signals exceed a predetermined threshold amplitude. This action mutes the equipment, eliminating annoying background noise and unwanted signals. Also called MUTING CIRCUIT.

squelch signal The activating or deactivating signal delivered by a SQUELCH CIRCUIT.

squiggle See BLIP, **2**.

squint **1.** The angular resolution of a radar antenna. **2.** The angular difference between the antenna axis and the major lobe of a radar transmitter.

squirrel-cage induction motor See SQUIRREL-CAGE MOTOR.

squirrel-cage motor An induction-type alternating-current motor using a squirrel-cage rotor.

squirrel-cage rotor In an alternating-current motor, a rotor composed of straight copper bars embedded in a laminated soft-iron core and short-circuited at the ends by rings. Its name is derived from its resemblance to a revolving squirrel cage.

squirter See SIGNAL INJECTOR.

SR **1.** Abbreviation of SILICON RECTIFIER. **2.** Abbreviation of *silicone rubber* (see SILICONE). **3**. Abbreviation of SHIFT REGISTER.

S-R Abbreviation of SEND-RECEIVE.

Sr Symbol for STRONTIUM.

sr Abbreviation of STERADIAN.

SRAM Abbreviation of *static random-access memory.*

S-rays See SECONDARY RAYS.

SRF Abbreviation of SELF-RESONANT FREQUENCY.

S-RF meter A dual-function meter in a radio transceiver. In the receiving mode, the meter indicates S units. In the transmit mode, the meter indicates relative output power.

SRR Abbreviation of SHORT-RANGE RADAR.

SS **1.** Abbreviation of SOLID STATE. **2.** Abbreviation of SINGLE SHOT. **3.** Abbreviation of *small signal.* **4.** Abbreviation of SINGLE SIGNAL. **5.** Abbreviation of *same size.* **6.** Abbreviation of *stainless steel.*

SSB Abbreviation of SINGLE SIDEBAND.

S_S band A section of the S BAND extending from 2900 to 3100 MHz.

SSBSC Abbreviation of SINGLE-SIDEBAND SUPPRESSED CARRIER.

ssc Abbreviation of *single-silk-covered* (wire).

S scale A scale of numbers used in radio communications, and especially in amateur radio, to report the approximate strength of signals: S1, faint signals; S2, very weak signals; S3, weak signals; S4, fair signals; S5, fairly good signals; S6, good signals; S7, moderately strong signals; S8, strong signals; S9, extremely strong signals. Also see S METER.

sse Abbreviation of *single-silk-enameled* (wire).

SSI Abbreviation of SMALL-SCALE INTEGRATION.

SSL Abbreviation of SOLID-STATE LAMP.

SSSC Abbreviation of SINGLE-SIDEBAND SUPPRESSED CARRIER. (Also, SSBSC.)

ST Abbreviation of SINGLE THROW.

sta **1.** Abbreviation of STATION. (Also *stn.*) **2.** Abbreviation of STATIONARY.

stab Abbreviated form of stabilizer (see STABILIZER, 4).

stability **1.** The condition in which an equipment or device is able to maintain a particular mode of operation without deviation. **2.** The condition in which the setting or adjustment of a device remains at a particular point without movement. **3.** The condition in which a quantity remains constant, with respect to time, temperature, or another variable. **4.** The ability of inks used in optical character recognition to retain their color after exposure to light or heat.

stability factor Abbreviation, SF. For a bipolar transistor, the derivative dI_c/dI_{co}, where I_c is the steady-state collector current and I_{co} is the collector cutoff current.

stabilized platform See STABLE PLATFORM.

stabilizer **1.** See DAMPING DIODE. **2.** See DAMPING RESISTOR. **3.** A device or circuit for the self-regulation of current or voltage. **4.** A chemical used to control or arrest a reaction.

stabilizing windings Auxiliary field windings used to prevent speed runaway in shunt motors.

Stabistor Trade name for a type of voltage-regulating semiconductor diode.

stable device A device whose characteristics and performance remain substantially unchanged with time or variations in temperature, applied power, or other quantities.

stable element **1.** A component that maintains its value or ratings, despite widely variable environmental conditions. **2.** A navigational instrument that maintains its orientation at all times.

stable platform A gyro-type device used to stabilize objects in space, and to provide accurate information regarding attitude (pitch, roll, and yaw).

stable state A stable condition, such as the high and low states of a flip-flop. The flip-flop has two stable states and will remain in one until it is switched to the other, whereupon it will then remain in that latter state until switched back to the former. Compare UNSTABLE STATE.

stack **1.** A piled assembly of capacitor plates and separating dielectric films. **2.** An assembly of selenium rectifier plates (see POWER STACK). **3.** To assemble a stacked array. **4.** A temporary storage area consisting of a small group of registers in a computer memory.

stacked array An antenna system in which two or more identical antennas, such as dipoles, Yagis, or halos, are placed one above the other or side-by-side. It provides additional forward gain, and, in some cases, enhances the front-to-back ratio and/or front-to-side ratio.

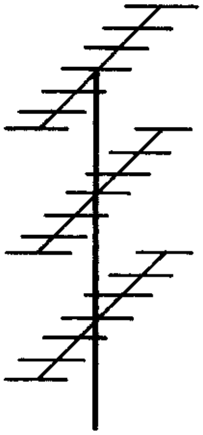

stacked array

stacked-dipole antenna A stacked array of half-wave dipole antennas.

stacking The combination of two or more identical antennas, such as dipoles, Yagis, or halos, in a STACKED ARRAY to provide enhanced forward gain. It can also enhance the front-to-back ratio and/or front-to-side ratio. Stacking can be done vertically or horizontally.

stack pointer Abbreviation, SP. A register indicating the last data item to be entered in a stack (see STACK, 4).

stage A complete functional unit of a system (e.g., *amplifier stage*, *oscillator stage*, *modulator stage*, etc.).

stage-by-stage elimination See SIGNAL INJECTION.

stage gain The amplification provided by a single stage in a system.

stage loss The loss introduced by a single stage in a system.

stagger **1.** An error in facsimile reception, occurring as a constant discrepancy in the position of the received dot. **2.** To deliberately tune a set of resonant circuits, especially in a bandpass filter, to one side or the other of the center frequency.

staggered tuning The tuning of the input and output circuits of a single stage, or the tuning of successive stages to slightly different frequencies to obtain flat-top response.

stagger tuning See STAGGERED TUNING.

stagger-wound coil See BASKET-WEAVE COIL and SPIDERWEB COIL.

staircase circuit See STAIR-STEP CIRCUIT.

staircase generator A circuit or device for generating a STAIR-STEP WAVE.

staircase wave See STAIR-STEP WAVE.

stair-step circuit A circuit that converts a series of equal-amplitude pulses into a stair-step wave.

stair-step wave A nonsinusoidal wave characterized by a multistep rise and a steep fall. It is so called from its resemblance to the cross section of a staircase.

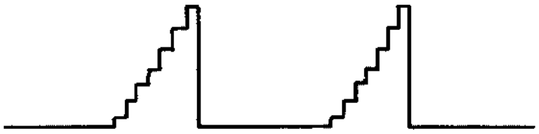

stair-step wave

stall torque The torque produced when a motor shaft is prevented from turning.

stalo Acronym for *stabilized oscillator.*

stand-alone photovoltaic system A solar-power plant that uses large banks of rechargeable electrochemical batteries, such as the lead–acid type, to store electric energy as it is supplied by photovoltaics during hours of bright sunshine. The energy is released by the batteries at night or in gloomy daytime weather. This system does not depend on the electric utility companies. Although this scheme offers independence from the utility companies, a blackout will occur if the system goes down. Compare INTERACTIVE PHOTOVOLTAIC SYSTEM.

standard **1.** A precise specification governing the dimensions and characteristics of a device or system (e.g., *military standard*). **2.** A highly accurate physical or electrical quantity to which similar quantities can be compared (e.g., *standard frequency*). **3.** The device or system that produces a standard quantity as defined in **2** (e.g., *frequency standard*). **4.** Having conventional and widely accepted characteristics.

standard atmosphere Abbreviation, atm. Air pressure at sea level (1.013 Pascals, or about 14.7 pounds per square inch). Also called ATMOSPHERE.

standard broadcast band Any of numerous frequency bands allocated to conventional broadcast stations. In the United States, the amplitude-modulation (AM) radio broadcast band extends from 535 to 1705 kHz, and the frequency-modulation (FM) radio broadcast band extends from 88 to 108 MHz. The television (TV) broadcast bands range from 54 MHz to 806 MHz in several sections, designated in channels from 2 through 69. Also see BROADCAST SERVICE, **1**.

standard candle See CANDELA.

standard cell A highly refined primary cell used to supply a precise direct-current voltage for electronic measurements. The *Weston standard cell* contains a mercury positive electrode, cadmium amalgam negative electrode, and cadmium sulfate electrolyte, and delivers 1.0183 volts at 20 degrees Celsius. Also see ZINC STANDARD CELL.

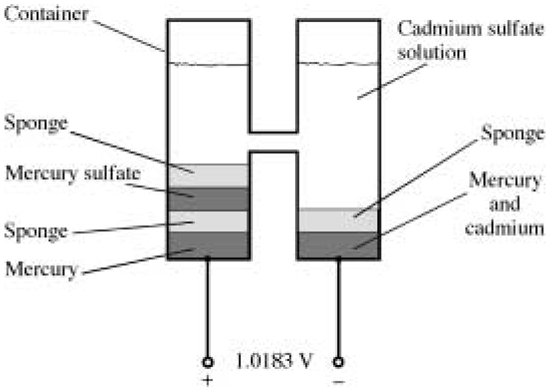

standard cell

standard deviation In statistical analysis, the square root of the mean of squares of deviation from the mean.

standard frequency A highly precise frequency to which other frequencies can be compared for identification or measurement.

standard-frequency oscillator A stable, precise oscillator that delivers a standard frequency. Also see PRIMARY FREQUENCY STANDARD and SECONDARY FREQUENCY STANDARD.

standard-frequency signal A calibration and reference signal that is broadcast on a standard frequency, such as those transmitted on 2.5, 5, 10, and 15 MHz by the National Bureau of Standards.

standard pitch The tone corresponding to the frequency 440 Hz (in music, the note A above middle C).

standard signal generator A (usually continuously variable) high-grade generator of modulated and unmodulated radio-frequency test signals. A general-purpose instrument of this type usually covers a wide range (e.g., 15 kHz to 100 MHz) in several tuning bands. For calibration, a standard signal generator is referred to a primary frequency standard or secondary frequency standard.

standard subroutine A usually vendor-supplied computer program segment applicable to more than one program and used as needed as a subroutine.

standard temperature and pressure The condition where the temperature is zero degrees Celsius and the pressure is one atmosphere. Abbreviated STP.

standard time Official civil time in a particular region. See TIME ZONE.

standby **1.** The state in which equipment is out of operation, but can be immediately activated. Also called IDLING. **2.** A state of readiness on the part of personnel, equipment, or systems.

standby battery An emergency power source for a battery-powered installation.

standby current The CURRENT DRAIN of a circuit, device, or system when in the standby condition.

standby equipment See EMERGENCY EQUIPMENT.

standby operation Keep-alive operation during a standby interval (see STANDBY).

standby power The power drawn by an equipment connected to the power supply, but out of operation.

standby power supply A circuit containing a battery, an automatic switch, and sometimes a power inverter. When utility power fails, the switch actuates the supply, and the battery supplies power to essential devices or systems. Similar to an UNINTERRUPTIBLE POWER SUPPLY.

standing wave A stationary distribution of current or voltage along a line because of the interactions between a wave transmitted down the line and a wave reflected back; it is characterized by maximum-amplitude points (loops) and minimum-amplitude or zero points (nodes).

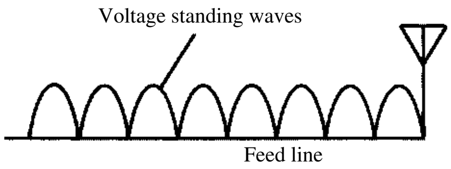

standing wave

standing-wave distortion Distortion of current or voltage caused by standing waves on a transmission line terminated in an impedance that contains reactance, and/or that differs from the characteristic impedance of the line.

standing-wave indicator **1.** A device, such as a lamp or meter, used to detect standing waves. **2.** Standing-wave meter (see SWR BRIDGE).

standing-wave loss The additional loss, over the matched-line loss, that occurs in a transmission line when the standing-wave ratio (SWR) is not 1.

standing-wave meter See SWR BRIDGE.

standing-wave ratio Abbreviation, SWR. **1.** The ratio between the maximum and minimum voltage along a transmission line. This quantity is sometimes specifically called *voltage standing-wave ratio (VSWR).* **2.** The ratio between the maximum and minimum current along a transmission line. **3.** The ratio of load impedance to feed-line characteristic impedance or vice versa, whichever is greater than or equal to 1. Ideally, the SWR is equal to 1 or 1:1, representing a load impedance that is purely resistive and has the same value as the characteristic impedance of the feed line. A high standing-wave ratio causes increased loss in the line and can also result in excessive conductor heating or dielectric breakdown.

standoff insulator An insulator (usually of the post type) that is used to hold a wire or component away from a chassis or base.

star **1.** In a gravity-battery cell, the copper electrode. The name is derived from its star shape. **2.** A star-shaped circuit of three-phase components. Also see WYE CONNECTION.

star connection See WYE CONNECTION.

Stark effect The influence of a strong transverse electric field on the spectrum lines of a gas.

starlight scope A device capable of viewing in apparent total darkness. Its operation depends on its ability to provide high amplification of extremely low light levels, such as that of objects reflecting the light from a moonless, but starlit, sky.

star rectifier See WYE RECTIFIER.

starter **1.** An ignitor electrode in an ignitron (see IGNITOR). **2.** See STARTING BOX.

starting box A special rheostat for starting a motor gradually in steps. The device is provided with an electromagnet for holding the arm in the maximum-speed position and releasing it when power is interrupted.

starting rod An ignitor electrode in an ignitron (see IGNITOR).

starting voltage **1.** For a gas tube, the minimum voltage that will initiate the glow discharge. **2.** In appropriate solid-state devices (e.g., a diac), the voltage at which conduction between electrodes occurs.

start lead The lead attached to the first turn of a coil. Also called *inside lead.* Compare FINISH LEAD.

start/stop multivibrator See MONOSTABLE MULTIVIBRATOR.

stat- A prefix denoting ELECTROSTATIC.

statampere The cgs electrostatic unit of current; 1 statampere = 3.335640×10^{-10} ampere.

statcoulomb The cgs electrostatic unit of charge; 1 statcoulomb = 3.335640×10^{-10} coulomb.

state **1.** The present condition (i.e., on or off, true or false, 1 or 0, high or low) of a bistable device, such as a flip-flop. **2.** The physical or electrial condition or status of a component, device, circuit, or system.

statement The contents of a line in a source-language computer program.

state of charge The amount of charge, measured in coulombs or ampere hours, in a storage cell or battery at a given time. A measure of the available remaining energy in the cell or battery.

statfarad The cgs electrostatic unit of capacitance; 1 statfarad = 1.112650×10^{-12} farad.

stathenry The cgs electrostatic unit of inductance; 1 stathenry = 8.987554×10^{11} henry.

static **1.** Pertaining to that which is constant in quantity (e.g., *static collector current* of a transistor). **2.** Pertaining to that which is at rest (e.g., *static electricity*). **3.** The radio noise (sferics) produced by electric discharges in the atmosphere, usually lightning. **4.** Pertaining to a test-and-measurement mode for a unit or device, without subjecting the unit or device to regular operation. Compare DYNAMIC.

static base current See DC BASE CURRENT.

static base resistance See DC BASE RESISTANCE.

static base voltage See DC BASE VOLTAGE.

static cathode current See DC CATHODE CURRENT.

static cathode resistance See DC CATHODE RESISTANCE.

static cathode voltage See DC CATHODE VOLTAGE.

static characteristic An operating characteristic determined from constant, rather than alternating or fluctuating, values of independent and dependent variables. Examples: the direct-current (dc) characteristics of diodes, transistors, and integrated circuits. Compare DYNAMIC CHARACTERISTIC.

static charge Energy stored in a stationary electric field; electricity at rest.

static collector A device that grounds the rotating wheels of a motor vehicle, thereby removing the static electricity generated by the friction of the tires on the roadway.

static collector current See DC COLLECTOR CURRENT.

static collector resistance See DC COLLECTOR RESISTANCE.

static collector voltage See DC COLLECTOR VOLTAGE.

static convergence In a color-television picture tube, the convergence of the three undeflected electron beams at the center of the aperture mask.

static device A device with no moving parts.

static discharge resistor A fixed resistor connected between the earth and the high side of the power line in a television receiver to drain off atmospheric electric charge.

static drain current See DC DRAIN CURRENT.

static drain resistance See DC DRAIN RESISTANCE.

static drain voltage See DC DRAIN VOLTAGE.

static dump In computer operations, a dump occurring at a predetermined point in a program run or at the end of the run.

static electricity Energy in the form of a stationary electric charge, such as that stored in capacitors or produced by friction or induction.

static emitter current See DC EMITTER CURRENT.

static emitter resistance See DC EMITTER RESISTANCE.

static emitter voltage See DC EMITTER VOLTAGE.

static flip-flop A flip-flop (see BISTABLE MULTIVIBRATOR) using direct-current operating voltages. A single pulse switches the unit from on to off, and vice versa. Compare DYNAMIC FLIP-FLOP.

static forward current transfer ratio Symbol, H_{FE}. An expression of gain in a bipolar transistor. It can range from a factor of just a few times up to hundreds of times. Mathematically,

$$H_{FE} = I_C / I_B$$

where I_C is the collector current and I_B is the base current. The H_{FE} rating is important because it gives engineers an indication of the greatest current amplification that can be obtained with a particular transistor.

static frequency multiplier A magnetic-core device, similar to a magnetic amplifier or peaking transformer, that provides harmonics by distorting a sine-wave signal.

static gate current See DC GATE CURRENT.

static gate resistance See DC GATE RESISTANCE.

static gate voltage See DC GATE VOLTAGE.

static grid current See DC GRID CURRENT.

static grid voltage See DC GRID VOLTAGE.

static hysteresis The condition in which the magnetization of a material (when it has the same intensity as the magnetizing force) is different when the force is increasing than when the force is decreasing, regardless of the time lag. Compare VISCOUS HYSTERESIS.

static induction See ELECTROSTATIC INDUCTION.

static machine See ELECTROSTATIC GENERATOR.

static memory Also called *nonvolatile memory*. In a computer, a data memory medium (such as programmable read-only memory, or PROM) in which information is stored until it is altered or erased. It does not require a source of power to maintain the integrity of the data. Compare VOLATILE MEMORY.

static mutual conductance See STATIC TRANSCONDUCTANCE.

static plate current See DC PLATE CURRENT.

static plate resistance See DC PLATE RESISTANCE.

static plate voltage See DC PLATE VOLTAGE.

statics The study of forces, bodies, poles, charges, or fields at rest or in equilibrium. Compare DYNAMICS.

static skew In magnetic tape recording or playback, the amount of lead or lag time of one track, with respect to another. Ideally, the static skew should be zero or practically zero.

static source current See DC SOURCE CURRENT.

static source resistance See DC SOURCE RESISTANCE.

static source voltage See DC SOURCE VOLTAGE.

static storage Also called *nonvolatile storage.* In a computer, a data storage medium (such as magnetic or optical disk) in which information is stored until it is altered or erased. It does not require a source of power to maintain the integrity of the data. Virtually all data storage media are of this type, as contrasted with memory, which is often volatile (see STATIC MEMORY and VOLATILE MEMORY).

static stability The ability of a robot to maintain its balance while standing still. A robot with two legs is generally poor in this respect. This is one of the reasons why humanoid robots (androids) are difficult to engineer. A minimum of three legs is necessary for good static stability.

static subroutine In computer programming, a subroutine that always serves the same purpose [i.e., it does not need to be tailored (according to parameters) for a specific application].

station **1.** An installation consisting of a transmitter, receiver, or both. **2.** A test-equipment installation or position. **3.** A computer installation including peripherals.

stationary battery A (usually wet storage) battery not normally moved when in use.

stationary state A particular energy state for an atom represented by its electrons being in shells at specific energy levels.

stationary wave See STANDING WAVE.

station authorization The legal privilege assigned to a broadcast or communications station, allowing that station to be used for the purpose of transmitting electromagnetic signals.

station license See STATION AUTHORIZATION.

statistical quality control Quality control based upon the techniques of probability and statistics in analyzing findings, making predictions, and formulating procedures for sampling.

statmho The cgs electrostatic unit of direct-current conductance; 1 statmho = 1.112650 × 10^{-12} siemens.

statoersted The cgs electrostatic unit of magnetizing force; 1 statoersted = 265.458 A/m (3.33585 × 10^{-11} oersted).

statohm The cgs electrostatic unit of direct-current resistance; 1 statohm = 8.987554 × 10^{11} ohm.

stator **1.** A stationary coil. Compare ROTOR, **1**. **2.** The stationary member of a motor or generator. Compare ROTOR, **2**. **3.** The stationary-plate section of a variable capacitor. Compare ROTOR, **3**.

stator coil A stationary coil (see STATOR, **1**, **2**).

stator plate(s) The stationary plate(s) of a variable capacitor. Compare ROTOR PLATE.

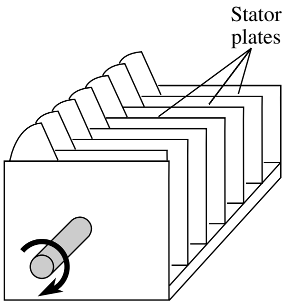

stator plate

stator section See STATOR, **3**.

statoscope An aircraft altimeter that shows small changes in altitude.

statvolt The cgs electrostatic unit of electromotive force; 1 statvolt = 299.7925 volts.

statweber The cgs electrostatic unit of magnetic flux; 1 statweber = 299.7925 webers (2.997925 × 10^{10} maxwells).

S_T band A section of the S BAND, extending from 1850 to 2000 MHz.

std Abbreviation of STANDARD.

steady-state component A quantity whose value remains constant during normal operation of a circuit or device, as opposed to an alternating, fluctuating, or transient component.

steerable antenna A directional antenna having a rotatable major lobe.

steering diode See DIRECTIONAL DIODE.

Stefan-Boltzmann constant Value, 5.67051 × 10^{-8} $\mathrm{Wm^{-2}K^{-4}}$.

Stefan-Boltzmann law The thermal-radiation law that shows the total emissive power of a blackbody to be proportional to the fourth power of the absolute temperature of the body.

stenode See CRYSTAL FILTER.

step **1.** A computer program instruction. **2.** A single action in the operation, maintenance, or troubleshooting of equipment. **3.** A specific increment in a quantity (such as frequency, voltage, current, etc.). **4.** A sharp or rapid change in the value of a quantity.

step-by-step operation See STEPTHROUGH OPERATION.

step change A single-increment change in a value.

step circuit A circuit that produces a step (sharp change of slope) in the response curve of an amplifier.

step counter **1.** A stair-step circuit arranged to count input pulses. The output capacitor of the circuit discharges when a predetermined number of input pulses has raised the capacitor voltage to the level required to trigger a counter. **2.** In a computer or calculator, a circuit or device that counts the steps in an operation (such as division, multiplication, or shifting) called for by an instruction.

step-down ratio In a circuit or device, such as a step-down transformer or cathode follower, the ratio of the low output voltage to the high input voltage. Compare STEP-UP RATIO.

step-down transformer A transformer delivering an output voltage that is lower than the input voltage. In such a transformer, the secondary (output) winding contains fewer turns than the primary (input) winding. Compare STEP-UP TRANSFORMER.

step function See UNIT FUNCTION.

step generator **1.** A signal generator that delivers a step function (see UNIT FUNCTION). **2.** A circuit or device that generates a STAIR-STEP WAVE.

stepped leader The probing flow of electrons through the atmosphere preceding a lightning stroke. Once the path has been established, the discharge takes occurs along the ionized path determined by the stepped leader. It is so called because the electrons move in hesitations, jumping several meters with each advance or step.

stepper motor A motor in which the shaft advances in uniform angular steps, instead of rotating continuously. These motors are extensively used in robotic devices. When such a motor is stopped and its coils are carrying current, the shaft resists turning.

stepping relay See STEPPING SWITCH.

stepping switch A multiposition rotary switch in which an electromechanical ratchet mechanism advances to the next contact position each time that a pulse of current is received.

step-through operation A way of operating a computer, usually during a debugging operation, in which program instructions are executed one at a time by direction of the user. Also called *single-step operation* and *step-by-step operation.*

step-up ratio In a circuit or device, such as a step-up transformer or voltage amplifier, the ratio of the high output voltage to the low input voltage. Compare STEP-DOWN RATIO.

step-up transformer A transformer delivering an output voltage that is higher than the input voltage. In such a transformer, the secondary (output) winding contains more turns than the primary (input) winding. Compare STEP-DOWN TRANSFORMER.

steradian A unit of solid-angle measure. A cone-shaped solid angle that has a vertex at the center of a sphere (of radius r), that cuts off a portion of the sphere's surface whose outer perimeter is a circle, and that has an area (as measured on the sphere's surface) of r^2. Also see SOLID ANGLE.

Sterba array See BARRAGE ARRAY.

Sterba curtain See BARRAGE ARRAY.

stereo **1.** Contraction of STEREOPHONIC. **2.** General term for a two-channel high-fidelity audio reproduction system.

stereo-adaptable Pertaining to a television receiver or videocassette recorder (VCR) that requires a special circuit to obtain stereo sound from Multichannel Television Sound signals. Compare STEREO-READY.

stereo amplifier A two-channel amplifier for binaural reproduction (see BINAURAL).

stereophonic Pertaining to equipment or techniques for producing a (somewhat) three-dimensional perspective of sound reproduction.

stereophonic sound system See STEREO SYSTEM.

stereo-irrelevant Pertaining to sound components in a stereo system that are of equal magnitude in both (or all) channels. Thus, these components sound the same whether the system is reproducing stereophonic sound or monaural sound.

stereo phono cartridge A phono cartridge capable of reproducing sound from stereo discs.

stereo-ready Pertaining to a television receiver or videocassette recorder (VCR) that can deliver stereo sound from Multichannel Television Sound signals without the need for a decoding circuit. Compare STEREO-ADAPTABLE.

stereo recording A method of recording in which two independent sound channels are transferred to some medium simultaneously, with the intention that the two channels be reproduced at the same time.

stereoscopic television Television in which the reproduced image appears three-dimensional.

stereo system A multichannel, high-fidelity sound reproduction system including an amplifier and various other components, such as a radio receiver, compact-disc (CD) player, tape player, turntable, and speakers.

stereotape Magnetic tape bearing more than one channel (usually two channels) for the recording and reproduction of stereophonic sound.

sterilizer Any electronic device, such as an ultraviolet generator, used to kill germs.

stethoscope An electronic or nonelectronic instrument used by physicians to listen to the heartbeat and other body sounds, and by technicians to listen to mechanical sounds.

still **1.** A stationary picture on television. **2.** A picture transmitted or received by means of facsimile. **3.** A print on photographic paper of a negative.

still television See FACSIMILE.

stinger A brief, loud sound burst, such as a musical chord, sometimes used for effect in recorded audio or audio-visual presentations.

stn Abbreviation of STATION. (Also, sta.)

STO Abbreviation of STORAGE FUNCTION.

stochastic The condition in which, at any instant, a given variable can assume a state dependent on previous states, as well as chance elements (e.g., words uttered extemporaneously by a speaker are random in that they cannot be predicted), while also being dependent, by the application of grammar, on previously spoken words (i.e., previous states).

stop amplifier See REJECT AMPLIFIER.

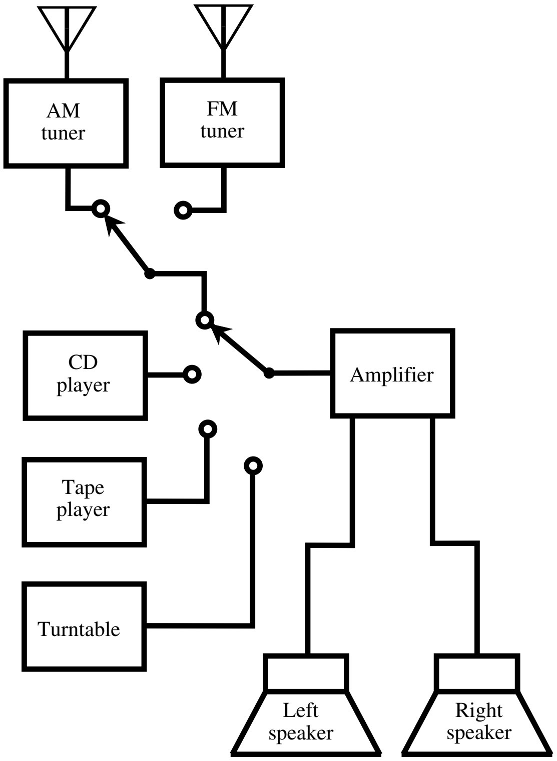

stereo system

stopband The continuous spectrum of frequencies rejected by a filter, selective amplifier, or other band-suppression device.

stopband ripple Single or multiple attenuation peaks within the stopband of an elliptic filter, where the frequency of the attenuation peak(s) is/are associated with the resonant circuit(s) in the filter.

stop code See HALT INSTRUCTION.

stop element In digital transmission, a bit or set of bits indicating the end of a character, and serving to inform the receiving device of the end of the character.

stop filter See BAND-SUPPRESSION FILTER.

stopper resistor See STOPPING RESISTOR.

stopping capacitor See BLOCKING CAPACITOR.

stopping coil See CHOKE COIL.

stopping resistor A parasitic-suppressing resistor, usually inserted in series with the input and/or output of a power amplifying device. Also called *stopper resistor.*

storage **1.** In computer operations, a medium on which data can be kept for an extended period of time. Examples: magnetic disk, magnetic tape, optical disk. **2.** The transferring of data from memory to a more permanent medium. Compare MEMORY. **3.** The retention of data of any kind, such as an oscilloscope image or video display, for use at a later time. **4.** The retention of electric energy or charge, as in a capacitor or electrochemical cell. **5.** The retention of energy in the form of a magnetic field, as in an inductor. **6.** The retention of potential energy in any form.

storage allocation The assignment of computer memory areas to certain kinds of information, as outlined in a source program and implemented by a compiler.

storage battery A rechargeable battery; the technical term is *secondary battery.* Also see STORAGE CELL.

storage capacity **1.** The amount of data that can be stored in a specific medium, such as a hard disk, diskette, or tape. Generally measured in bytes, kilobytes, megabytes, gigabytes, or terabytes. **2.** The energy-delivering capability of a storage battery in terms of ampere-hours, milliampere-hours, or other current-time units for a specific rated voltage.

storage cell **1.** An electrochemical cell whose potential can be restored by charging it with electricity. Compare PRIMARY CELL. Also see CELL. **2.** The smallest part of a computer storage medium. **3.** In computer data storage, the part that can hold a data unit (e.g., a bit).

storage CRT See STORAGE TUBE.

storage cycle In computer operation, the period during which a location of a cyclic storage device cannot be accessed.

storage density The number of data units (e.g., bytes, kilobytes, or megabytes) that can be stored in a given length or area of a storage medium.

storage device A medium into which data can be placed and kept for later use. Examples: magnetic disk, magneto-optical disk, optical disk, and magnetic tape.

storage dump See DUMP.

storage function Abbreviation, STO. The so-called “memory” function of a microcomputer chip. It causes data to be inserted into memory or storage for later use. It is commonly used in programmable calculators, automatic-dialing telephones, and radio receivers. Also called *memory function.*

storage laser A laser that stores intense energy before flashing.

storage life See SHELF LIFE.

storage mesh In the cathode-ray tube (CRT) of a storage oscilloscope, a fine metal mesh electrode that serves as the target on which the image is electrostatically stored. Also see STORAGE TUBE.

storage oscilloscope An oscilloscope that retains a displayed image until the display is erased. Also see STORAGE TUBE.

storage register In computers and calculators, a storage unit composed of flip-flops. In computers, it is independent of the central processing unit (CPU).

storage temperature **1.** The recommended temperature for storing specified electronic components. **2.** The particular temperature at which electronic components have been stored.

storage time **1.** The interval during which carriers remain in a semiconductor-junction device after the bias has been removed. Also see DIODE RECOVERY TIME. **2.** For a switching semiconductor device, the time required for the amplitude of the output pulse to fall from maximum to 90% of maximum after the input pulse has fallen to zero. **3.** In a computer, the time required for data to be transferred from random-access memory (RAM) to nonvolatile storage (e.g., hard disk).

storage tube A cathode-ray tube that retains information in the form of images on a special electrode until erased by a signal.

store **1.** To place data in a nonvolatile medium (such as a hard disk, diskette, optical disk, magnetic tape, etc.). **2.** To place in the memory of a calculator or computer. **3.** In computing, a command that causes data to be placed in a nonvolatile medium. In some applications, this is called *save.* **4.** A nonvolatile medium on which data has been placed for future use or for archival purposes.

stored base charge The carriers that remain in the base layer of a bipolar transistor immediately after the forward bias has been interrupted. This charge maintains collector current momentarily.

stored-energy welding A method of electric welding in which electrical energy is stored slowly, then released at the rate required for the welding.

STP See STANDARD TEMPERATURE AND PRESSURE.

straight adapter An inline coaxial fitting for joining two fixture-terminated coaxial lines in series.

straight angle An angle measuring 180 degrees.

straight dipole A (usually center-fed) dipole antenna having only one radiator. Also see DIPOLE ANTENNA.

straightforward Pertaining to data transmission in one direction only.

straight-gun CRT A cathode-ray tube (CRT) in which the electron gun projects the beam in a straight line through the deflecting elements to the screen. Compare BENT-GUN CRT.

straight-line capacitance Abbreviation, SLC. Pertaining to a variable capacitor for which the setting-vs.-capacitance curve is a straight line; the capacitance variation is linear. Compare STRAIGHT-LINE FREQUENCY and STRAIGHT-LINE WAVELENGTH.

straight-line coding See STRAIGHT-LINE PROGRAMMING.

straight-line frequency Abbreviation, SLF. Pertaining to a variable capacitor in a tuned circuit for which the setting-vs.-frequency curve is a straight line; the frequency variation is linear. Compare STRAIGHT-LINE CAPACITANCE and STRAIGHT-LINE WAVELENGTH.

straight-line programming During the writing of a computer program, avoiding the creation of loops by repeating a series of instructions to reduce execution time.

straight-line tracking In a phonograph turntable, linear lateral stylus movement (as opposed to motion along an arc) as the disc is played. This ensures that the stylus is always at the optimum angle in the disc groove. The result is improved sound reproduction, and longer disc and stylus life because of minimal friction between the stylus and groove.

straight-line wavelength Abbreviation, SLW. Pertaining to a variable capacitor in a tuned circuit for which the setting-vs.-wavelength curve is a straight line; the wavelength variation is linear. Compare STRAIGHTLINE CAPACITANCE and STRAIGHT-LINE FREQUENCY.

straight-through amplifier An amplifier in which the input and output circuits are tuned to the same frequency. Compare MULTIPLIER AMPLIFIER.

strain A force that compresses or squeezes a body. Compare TENSION, **1**.

strain gauge See ELECTRIC STRAIN GAUGE.

strain-gauge bridge A four-arm resistance bridge in which an ELECTRIC STRAIN GAUGE forms one arm. The resistance of the gauge changes because of strain. The amount of strain can be determined by balancing the bridge.

strain-gauge transducer A transducer, other than a strain sensor, that uses strain gauges to convert values of pressure into their electrical analogs (e.g., *pressure transducer* and *strain-gauge phonograph pickup*).

strain pickup A phonograph pickup using a strain gauge to convert sound vibrations into a varying electric current.

strand A single solid conductor in a STRANDED WIRE.

stranded wire A conductor composed of several non-insulated wires twisted together to provide mechanical flexibility. Compare SOLID WIRE.

stratosphere The portion of earth's upper atmosphere beginning at a height of approximately 10 miles and extending to the ionosphere.

stray capacitance Inherent capacitance in a place where it can be detrimental, such as that between the turns of a coil or between adjacent areas in a circuit. Also see STRAY COMPONENT.

stray component An electrical property that exists as an inherent, and usually undesirable, side effect in a circuit or device. Thus, for example, STRAY CAPACITANCE unavoidably exists between parallel conductors, and STRAY INDUCTANCE is present in all wiring.

stray field The portion of an electric or magnetic field that extends beyond the immediate vicinity of the circuit with which it is associated, and which is, therefore, capable of interfering with other circuits or devices.

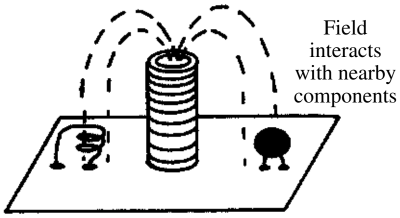

stray field

stray inductance Inherent inductance in a place where it can be detrimental (e.g., inductance in the coil of a wirewound resistor). Also see STRAY COMPONENT.

stray resistance Inherent resistance in a place where it can be detrimental, such as leakage resistance in a dielectric, and wire resistance in an inductor.

streaking In television or facsimile, a form of distortion in which the image appears enlarged in the horizontal, but not the vertical.

stress **1.** See STRAIN. **2.** See TENSION, **1**. **3.** The force per unit area that produces STRAIN or TENSION on a body.

stretch The amount by which a material being measured with an electronic device increases in surface dimensions. Compare SHRINK.

stretched string A long, thin wire or string that vibrates at a certain frequency, causing standing waves. It generally exhibits a specific fundamental frequency and integral harmonics of this frequency. As the wavelength is cut in half, the frequency doubles.

strike To initiate a discharge, as in striking a gas tube.

striking voltage See STARTING VOLTAGE.

string **1.** In computer operations, a set of items in a sequence determined by the order of keys. **2.** In a computer memory, a sequence of bits or characters. **3.** Any group of series-connected components or circuits.

string electrometer See BIFILAR ELECTROMETER.

string variable A string of characters, usually forming a word or phrase, represented by a variable name and character string symbol (BASIC's $, for example) in a computer program.

strip chart A longitudinal, as opposed to circular, chart for graphic recording. In a rectilinear chart, both coordinates are straight; in a curvilinear chart, the crosswise coordinates are arcs.

strip core A ferromagnetic core material made from a strip of the substance. The method of manufacture results in superior ferromagnetic qualities, but also imparts a polarization to the material.

strip fuse A fuse in which the fusible element is a flat strip of low-melting-point metal. Compare WIRE FUSE.

strobe **1.** See ELECTRONIC FLASH, **1**. **2.** See STROBOSCOPE.

strobe light **1.** See ELECTRONIC FLASH, **1**. **2.** See STROBOSCOPE.

Strobolume Trade name for a type of high-output stroboscope.

stroboscope **1.** An instrument that emits bright, adjustable-rate flashes of light. When this light illuminates an object that is rotating or vibrating at a fixed period, and the flash rate is made to match that period, the object seems to stand still and can be examined for flaws or faulty operation (and its speed can be measured). **2.** A rotatable, slotted disk for producing the effect defined in **1**.

stroboscopic disk A rotatable disk with alternating white and black radial regions, used in conjunction with a strobe light for precise measurement of the speed of a phonograph turntable.

strobotron A gas tetrode tube used as the flashing light source in a stroboscope.

stroke speed See SCANNING FREQUENCY.

strong coupling See CLOSE COUPLING.

strontium Symbol, Sr. A metallic element of the alkaline-earth group. Atomic number, 38. Atomic weight, 87.62. It is used in some ceramic dielectrics, such as *barium-strontium titanate.*

Strowger exchange A telephone system incorporating Strowger switches.

Strowger switch A switch with one input and 100 individually selectable outputs. It is used with telephone switching networks. The telephone dial code causes a contact to move vertically and horizontally in such a way that a particular output is connected to the input. Each output has a unique dial code and each dial code has a unique output.

structured programming Computer programming using a limited number of procedural sets, while minimizing branches to make the program as forward-going as possible. This allows it to be easily modified or debugged.

STS switch See SPACE-TIME-SPACE SWITCH.

stub A (usually short) section of transmission line that is patched onto a longer line for tuning or impedance matching.

Stubs gauge See BIRMINGHAM WIRE GAUGE.

stub tuner A tuning unit consisting of a stub with a short circuit that can be moved along the stub.

stub trap See INTERFERENCE STUB.

stub-type wavetrap See INTERFERENCE STUB.

stuffing bits In a digital communications system, extra bits inserted into some words so that all the words are the same length.

styli Plural of STYLUS.

stylus **1.** The "needle" that conveys vibrations to or from the disk in phonograph-disc recording or playback. **2.** One of the pins in the print head of a dot-matrix printer.

stylus drag See NEEDLE DRAG.

stylus friction Rubbing of the stylus against the record groove in phonograph-disc playback.

stylus pressure See VERTICAL STYLUS FORCE.

stylus printer See WIRE PRINTER.

stylus scratch See NEEDLE SCRATCH.

sub Abbreviation of *subtract.*

sub- Prefix denoting *under*, *below*, *less than*, or *lower than*, with respect to size, value, or rank. Compare SUPER-.

subassembly A completely fabricated unit that forms part of a larger unit into which it easily fits.

subatomic particle **1.** Any of various particles that comprise atoms of matter. **2.** A particle smaller than an atom. See, for example, ANTI-PARTICLE, ELECTRON, MESON, NEUTRETTO, NEUTRINO, NEUTRON, NUCLEON, POSITRON, and PROTON.

subaudible **1.** Pertaining to any frequency falling below the limit of human hearing, that is, less than about 20 Hz. **2.** Any sound that is too low in amplitude to be heard.

subaudible tone A signal, usually a steady, sine-wave tone, sent along with a radio signal. The tone frequency varies from about 20 Hz to 200 Hz, below the audio cutoff frequency of most voice communications systems. Subaudible tones are used mainly for privacy. The receiver is programmed to receive only signals having the correct subaudible tone frequency.

suballocation A portion of a radio-frequency broadcast or communications band that is legally set aside for specific purposes or users, e.g., the Extra-class segment of the 40-meter amateur-radio band.

subband **1.** A portion of a frequency band with specific characteristics. **2.** A portion of a radio-frequency broadcast or communications band that is set aside, legally or by convention, for specific purposes [e.g., the single-sideband (SSB) portion of the 20-meter amateur-radio band].

subcarrier A modulated carrier wave that composes the modulating signal for another carrier wave.

subcarrier band The band of frequencies in which a subcarrier signal is transmitted.

subcarrier frequency modulation In a system in which a carrier frequency is obtained by beating a low-frequency radio-frequency (RF) signal with a high-frequency RF signal, the application of frequency modulation to the low-frequency component. The technique is sometimes used in sweep-frequency signal generators.

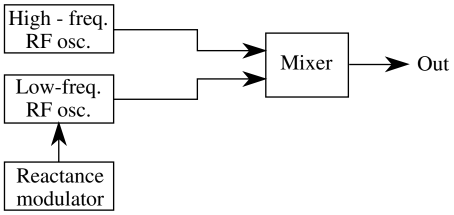

subcarrier frequency modulation

subcarrier oscillator In a color-television receiver, the oscillator operating at the burst (chrominance-subcarrier) frequency of 3.579545 MHz.

subchassis An auxiliary chassis on which one section of a larger piece of equipment is completely assembled and wired.

subfrequency See SUBHARMONIC.

subharmonic An integral submultiple of a fundamental frequency. Thus, for example, the 10th subharmonic of 15 MHz is 1.5 MHz.

submarine cable An underwater cable designed to withstand continuous immersion.

submarine robot A robot designed for underwater operation. It can be operated via telepresence or by simple remote control, usually using a conventional or fiberoptic cable. Some underwater robots have manipulators attached; others are equipped only with cameras, lights, and propulsion devices.

subminiature jack A female connector with an inside diameter of 3⁄32 inch.

subminiature plug A male connector with an outside diameter of 3⁄32 inch.

submultiple A fractional multiple, usually in reference to a frequency. For example, 7.2 MHz is a submultiple of 14.4 MHz. See SUBHARMONIC.

subpanel The front panel of a removable unit or module that forms a part of a larger unit.

subroutine In a computer program, a sequence of instructions for carrying out a section of the program's function. It is usually entered (led to) by a conditional branch (jump) instruction in the main program.

subscriber An individual user of a communications network or service.

subscript A small number or letter written to the lower right (and occasionally to the lower left) of another number or letter to identify the latter from others of the same designation (e.g., a_5, S_n). Compare SUPERSCRIPT.

subscription TV A television (TV) service paid for by subscribing viewers. The signals are scrambled so as to be useless to nonsubscribers, and legitimate subscribers are provided with a decoder to unscramble the telecasts.

subset **1.** In statistics and set theory, a set whose members are all contained in a larger set. **2.** A telephone handset or deskset (subscriber's set). **3.** A modulator/demodulator for making business machines compatible with telephone circuits.

subsidiary communication authorization Abbreviation, SCA. An authorization provided by the Federal Communications Commission (FCC) for a frequency-modulation (FM) broadcast station to transmit, in addition to its main program, a second program within the assigned bandwidth usually consisting of commercial-free background music on a subcarrier that can be detected only with a special receiver or with a special adapter attached to a standard FM receiver.

subsidiary feedback Feedback other than the main feedback in a system.

subsonic frequencies Frequencies below the range of hearing, that is, less than about 20 Hz. Also called ULTRALOW FREQUENCIES and SUBAUDIBLE FREQUENCIES.

substation An intermediate electricity-distributing location from which electrical energy is transformed and transmitted to users within a given geographical area.

substitution capacitor A capacitor used temporarily in place of another of usually the same value, as in troubleshooting. Also see CAPACITOR SUBSTITUTION BOX.

substitution inductor An inductor used temporarily in place of another of usually the same value, as in troubleshooting. Also see INDUCTOR SUBSTITUTION BOX.

substitution method **1.** A method of measuring a quantity (such as capacitance, inductance, or resistance) in which the value of the unknown quantity is determined in terms of the amount of a standard quantity that must be removed to restore the test circuit to its original state of balance. **2.** A method of troubleshooting in which good components are substituted for bad ones in a circuit (see, for example, SUBSTITUTION CAPACITOR, SUBSTITUTION INDUCTOR, SUBSTITUTION RESISTOR, SUBSTITUTION SPEAKER, and SUBSTITUTION TRANSFORMER).

substitution resistor A resistor used temporarily in place of another of usually the same value, as in troubleshooting. Also see RESISTOR SUBSTITUTION BOX.

substitution speaker A loudspeaker used temporarily in place of another, as in troubleshooting.

substitution transformer A transformer used temporarily in place of another having the same characteristics, as in troubleshooting.

substrate A plate, wafer, panel, or disk of suitable material on (or in) which the components of a unit, such as an integrated or printed circuit, are deposited or formed.

subterranean **1.** Pertaining to components, systems, or devices installed underground. It is applicable especially to cables. **2.** Pertaining to a phenomenon, such as the propagation of electric currents or acoustical waves, that occurs underground.

subterranean acoustical communication A method of communication that uses low-frequency sound waves, such as SONAR, to communicate via conduction through earth or water.

subtracter See ELECTRONIC SUBTRACTER.

subtractive color A color formed by mixing subtractive primary pigments.

subtractive primaries Broad-spectrum pigments used in printing to produce a wide variety of colors through filtering. These primaries are cyan (blue-green), magenta (pink-red), yellow, and sometimes black. They are used to print images that have been filtered through additive-primary lenses.

subtrahend In the process of subtraction, the quantity that is subtracted from another (the minuend) to produce the remainder or difference.

subwoofer A speaker designed to effectively reproduce extremely low audio frequencies, in some cases, subaudible (below 20 Hz).

successive derivatives Successive repetition of the operation of differentiating a function, which yields the first derivative, second derivative, and so on to the *n*th derivative.

successive integration The operation of double or triple integration.

Suhl effect A reduction in hole life that occurs in a semiconductor material in the presence of a magnetic field.

suite **1.** A group of computer programs run successively as a job. **2.** A bundled, high-end software package used especially in business computing.

sulfate Contraction of lead sulfate.

sulfation In a lead-acid storage cell, the formation of disabling lead sulfate during discharge of the cell.

sulfur Symbol, S. A nonmetallic element. Atomic number, 16. Atomic weight 32.06.

sulfur hexafluoride A gas used as a coolant and insulant in some power transformers.

sulfuric acid Formula, H_2SO_4. An acid used in dilute solution as the electrolyte of a lead-acid battery. This highly corrosive fluid also has many industrial uses.

sum The result obtained by adding two or more terms. Compare SUMMATION.

sumcheck See SUMMATION CHECK.

sum frequency **1.** In an amplitude-modulated carrier, the upper sideband frequency (i.e., the sideband equal to the carrier frequency plus the modulating frequency). Compare DIFFERENCE FREQUENCY. **2.** In superheterodyne operation, an intermediate frequency equal to the signal frequency plus the local-oscillator frequency.

summation **1.** The sum of a finite number of terms. Thus, the total resistance of n resistors connected in series is the summation of all R (resistance) terms. **2.** A frequency equal to the sum of two other frequencies.

summation check In computer operations, a check carried out on a group of digits. The result of adding the digits, and disregarding any overflow, is a check digit that can be compared with a standard value for the operation to verify the integrity of data.

summer **1.** See ADDER. **2.** See SUMMING AMPLIFIER.

summing amplifier An operational amplifier whose output is the sum, or is proportional to the sum, of several inputs.

sun battery A set of photovoltaic cells connected in series, parallel, or series-parallel to produce useful output voltages and currents.

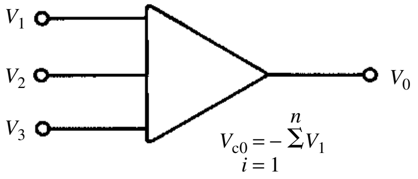

summing amplifier

S units In radio (especially amateur radio), gradations reflecting the strength of received signals. Typically, a value of S9 (9 s-units), representing "extremely strong signals," is equal to a strength of 50 microvolts. The next lower S unit (S8) is 6 dB lower in voltage (i.e., 25 microvolts); S7 is 12 dB below S9 (i.e., 12.5 microvolts), etc.

sun lamp An incandescent or fluorescent lamp with high ultraviolet output, used in medicine for the treatment of certain skin disorders. It can also be used for skin tanning; most physicians discourage habitual use of the lamps for this purpose.

sunlight lamp A lamp that produces visible light whose spectral distribution is similar to that of typical daylight. It generally produces more blue and violet light than a conventional lamp. It is sometimes used for indoor lighting in winter at high latitudes, and/or to enhance plant growth.

sunlight-powered laser A laser whose action is stimulated by sunlight collected by a system of mirrors and lenses. The life of the device is long, compared with that of conventional lasers.

sun-pumped laser See SUNLIGHT-POWERED LASER.

sun relay See SUN SWITCH.

sunspot An area on the sun's surface that is visible as a dark, irregular region of variable size, generally several thousand miles across. Sunspots are believed to be comparatively cool regions associated with solar magnetic disturbances. The number of sunspots is correlated with the frequency and intensity of solar flares (see SOLAR FLARE).

sunspot cycle Regular periodic variation of sunspot activity. The time between peaks in activity is approximately 11 years.

sun switch A photoelectric switch or relay actuated by sunlight and used for various domestic and industrial purposes, such as switching lights, operating window shades, etc.

sup Abbreviation of SUPPRESSOR.

super Contraction of *supersonic.*

super- Prefix denoting *over*, *above*, *greater than*, or *higher than*, with respect to size, value, or rank. Compare SUB-.

superaudible frequency See ULTRASONIC FREQUENCY.

superbeta transistor A transistor or transistor combination, such as a Darlington pair (see COMPOUND CONNECTION), that provides a very high current amplification factor (beta).

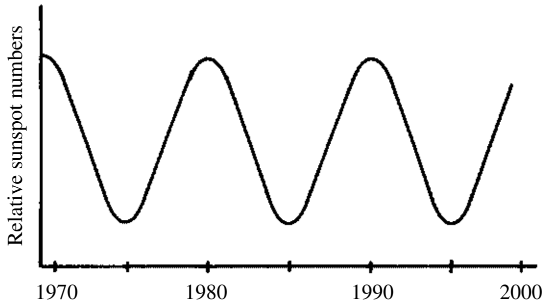

sunspot cycle

supercardioid microphone A microphone that is highly sensitive in one direction and insensitive in all other directions. The directional pattern is similar to a CARDIOID PATTERN, but is exaggerated along the axis of optimum response.

superconducting cable A cable in which superconductivity is achieved by surrounding the cable with liquid helium to lower its temperature to near absolute zero.

superconductivity The virtual disappearance of resistance in some metals cooled to temperatures in the vicinity of absolute zero. Also see CRYOGENICS, CRYOSTAT, and CRYOTRON.

superconductor A material or device that displays superconductivity.

super flatpack An integrated-circuit package of the flatpack type having considerably more components and leads than those in the conventional flatpack.

superhet Contraction of *superheterodyne.*

superheterodyne circuit A circuit in which the incoming signal in a first detector (or mixer) beats with the signal of a local oscillator, resulting in a lower (intermediate) frequency, which then is amplified by an intermediate-frequency (IF) amplifier. This IF signal is detected by a second detector whose output is amplified by an audio-frequency (AF) amplifier. Because the IF amplifier operates at a single (fixed) frequency, it can be adjusted for optimum selectivity and gain. Also called *superhet circuit.*

superheterodyne receiver A radio or television receiver using a SUPERHETERODYNE CIRCUIT.

superhigh frequency See RADIO SPECTRUM.

Supermalloy An alloy having a maximum permeability of 10^6.

supermodulation A type of amplitude modulation (AM) in which one radio-frequency (RF) power stage continuously generates the carrier, and a second (usually identical) RF power stage is gated into full operation at the proper instant by the audio modulation to add additional RF power (corresponding to 100% modulation) to the signal. At the same time, the carrier amplitude is decreased by the proper amount to fulfill the conditions of a signal amplitude swing between zero and twice maximum for 100% modulation.

superposition In a complex wave, the manner in which the constituent waves combine. The instantaneous value of the complex wave is equal to the vector sum of the instantaneous values of all the constituent waves.

superposition theorem In a network of linear elements, if a voltage E_1 in branch A causes a current I_1 to flow through branch C, and if a voltage E_2 in branch B (which might be identical with branch A) causes a current I_2 to flow through branch C, then E_1 in branch A and E_2 in branch B applied simultaneously will cause a current equal to $I_1 + I_2$ to flow through branch C. Compare COMPENSATION THEOREM, MAXIMUM POWER TRANSFER THEOREM, NORTON'S THEOREM, RECIPROCITY THEOREM, and THEVENIN'S THEOREM.

superpower An arbitrary term denoting very high power. In the rating of standard broadcast stations, it has come to signify 1,000,000 watts (one megawatt) radio-frequency (RF) power output.

superradiance In a laser, a rapid increase in intensity of fluorescent-line emission with increasing excitation power.

superregenerative circuit A regenerative detector circuit in which regeneration is periodically increased almost to the point of oscillation, then decreased. This quenching action takes place at a supersonic rate (typically at 50 or 100 kHz) so that the quenching is inaudible. The result is that much more regeneration is afforded, without the detector going into oscillation, than is possible by simply increasing the regeneration manually. An extremely sensitive detector is the result.

supersaturated solution A solution that contains more solute than it normally would hold. Supersaturated solutions are obtained through special techniques and are extremely unstable. Compare SATURATED SOLUTION. Also see SOLUTE; SOLUTION, **1**; and SOLVENT, **2**.

superscript A small number or letter written to the upper right of another number or letter, the BASE, to indicate the power to which the base must be raised. Example: 10^5, e^x, y^2. Also called EXPONENT. Compare SUBSCRIPT.

supersensitive relay A relay that operates with a current of less than one milliampere, or with a voltage of less than one millivolt.

supersonic flow In a gas or liquid, movement of the medium at a speed greater than the speed of sound in that medium. Supersonic flow results in a greatly increased resistance or drag because of shock waves that form in the medium.

supersonic frequency See ULTRASONIC FREQUENCY.

supersonics See ULTRASONICS.

supervised line In a security system, a wire or foil strip that carries electrical current. If the current changes in such a line, an alarm is actuated.

supervisor **1.** In a computer, a set of routines that oversees the operation of the system. The supervisor routines are coordinated by the central processing unit. **2.** The execution of such a set of routines. **3.** A microcomputer that oversees the operation of a security system.

supervisory circuit In a security system, a link between a sensor and the central computer or control device. This link can be via electric current through a wire or cable, but other methods can be used, such as fiberoptics, line-of-sight optics, infrared, ultrasonic, or radio.

supply **1.** See CURRENT SUPPLY. **2.** See POWER SUPPLY. **3.** See VOLTAGE SUPPLY.

supply current Alternating or direct current available for operating a circuit, device, or system.

supply frequency The frequency of an alternating-current power supply.

supply power The maximum power that can be reliably delivered by an alternating-current or direct-current power supply.

supply reel In a reel-to-reel tape recorder or player, the reel that is initially full, and that gradually empties as the tape moves through the machine.

supply voltage The voltage of an alternating-current or direct-current power supply.

suppressed carrier A carrier that has been canceled or filtered out of a carrier/sideband combination.

suppressed-carrier double sideband See DOUBLE-SIDEBAND and SUPPRESSED CARRIER.

suppressed-zero instrument A meter or graphic recorder in which the zero point is off-scale or upscale, but has been brought to scale-zero by means of mechanical adjustment or use of a bucking voltage.

suppressor **1.** A filter used to suppress radio interference. **2.** See AUTOMATIC NOISE LIMITER. **3.** See SPARK SUPPRESSOR. **4.** In a pentode vacuum tube, a gridlike element between the screen grid and the plate, used to suppress secondary emission. Also see GRID, **2** and PENTODE.

suppressor circuit The circuit associated with the suppressor electrode of a vacuum tube.

suppressor diode A semiconductor diode used to prevent inductive kickback in circuits, to eliminate or reduce transients, or to prevent arcing between make-and-break contacts.

suppressor grid See SUPPRESSOR, **4**.

suppressor modulation A method of modulation in which a modulating voltage is superimposed on the suppressor voltage of a pentode radio-frequency power amplifier tube.

suppressor pulse A pulse that prevents electron flow.

surface analyzer A device designed for the measurement of surface flatness or uniformity.

surface-barrier diffused transistor See MICRO-ALLOY DIFFUSED TRANSISTOR.

surface-barrier transistor Abbreviation, SBT. A pnp transistor made by means of electrolysis and electroplating: Two fine streams of indium sulfate

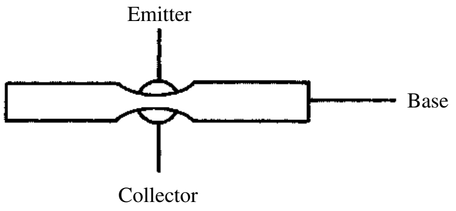

surface-barrier transistor

solution are placed on axially opposite points on the faces of an n-type wafer. At the same time, a direct current is passed through the wafer and solution in such a direction as to remove semiconductor material electrolytically from the faces; the tiny sprayed areas are etched away. When the desired wafer thickness is reached at the points of impact, the etching process is arrested by reversing the direction of current flow. This reversal causes an indium dot to be plated on each opposite face in the etched-out pit. Leads are attached to the collector and emitter dots and to the wafer (base).

surface-charge transistor A semiconductor device consisting essentially of two narrowly separated plates (source electrode and receiver electrode) deposited on the film-insulated surface of a silicon chip, and a third, overlapping electrode (the transfer gate) deposited on, but insulated from, the other electrodes. An input signal stores a charge in the capacitor formed by the source electrode and chip. A subsequent trigger signal applied to the transfer gate transfers the charge to the receiver electrode, where it becomes an output signal (often amplified, with respect to the input signal).

surface effect An effect (such as current, resistance, or resistivity) observed on the surface of a sample of material, rather than throughout the body of the material. Compare BULK EFFECT.

surface insulation A coating applied to the surfaces of core laminations to prevent the passage of currents between laminations.

surface leakage Leakage of current over the surface of a dielectric material, as opposed to leakage through the interior of the material.

surface noise See NEEDLE SCRATCH.

surface recombination rate For a semiconductor, the rate at which electrons and holes recombine at the surface. Compare VOLUME RECOMBINATION RATE.

surface resistivity The resistance of a unit area of a material, measured between opposite edges. Compare VOLUME RESISTIVITY. Also see RESISTIVITY.

surface tension The tendency of the surface of a liquid to “shrink.” This property varies with different liquids and is caused by a net molecular force directed inward from the surface.

surface wave **1.** The earth-guided component of a ground wave. (The other component is the SPACE WAVE.) **2.** An acoustic wave traveling along the surface of a plate in a surface-wave amplifier or surface-wave filter.

surface-wave amplifier An amplifying device consisting essentially of a surface-wave filter to which has been added a direct-current-biased n-type silicon electrode, which is separated from the crystal substrate of the filter by a very thin oxide layer. Amplification is produced by interaction between the electron current in the silicon and the piezoelectric field of the filter. Also see ACOUSTOELECTRONICS.

surface-wave filter An acoustoelectronic device consisting essentially of a crystal plate having electrodes at each end. An alternating-current (ac) input signal applied to one electrode sets up acoustic waves that travel along the surface of the plate to the other electrode, where they generate an ac output voltage by piezoelectric action. The resonant frequency of the device is governed by the dimensions of the crystal plate. Also see ACOUSTOELECTRONICS.

surge A sudden rise or flow of current or voltage.

surge absorber See SURGE SUPPRESSOR.

surge arrester See SURGE SUPPRESSOR.

surge current A heavy current that flows initially into a capacitor when a charging voltage is applied.

surge impedance Symbol, Z_o. The impedance seen by a pulse applied to a transmission line; $Z_o = L/C$ (approximately), where L and C are the inductance and capacitance, in microhenrys and microfarads, per unit length of the line. Also called CHARACTERISTIC IMPEDANCE.

surge protector Misnomer for SURGE SUPPRESSOR.

surge suppressor A semiconductor device used to absorb potentially destructive transients or overvoltages on a utility power line. It has a three-wire cord for plugging into a 117-volt outlet, a power switch, and several three-wire outlets for connection to sensitive electronic equipment (such as personal computers, videocassette recorders, television sets, hi-fi amplifiers, etc.).

surround See SUSPENSION, **1**.

Surround Sound The trade name for a multichannel sound system for use with television receivers and videocassette players. Some televised movies, especially on cable and satellite networks, deliver multichannel sound through receiving/recording systems equipped with special decoders.

surveillance **1.** A method of monitoring a specific area or volume for intrusion or other disturbance. **2.** A means of monitoring a specified portion of the electromagnetic spectrum for unauthorized signals.

surveillance radar An air-traffic-control radar that supplies continuous information regarding the azimuth and distance of aircraft inside a selected radius around an airport.

susceptance Symbol, *B.* Unit, siemens. The reactive component of admittance, as distinguished from conductance.

susceptibility The capacity of a substance to become magnetized, expressed as the ratio of magnetization to the strength of the magnetizing force.

suspension **1.** In a speaker, the flexible, circular or elliptical structure via which the cone is attached to the frame. **2.** The wire or metallized fiber supporting the movable coil of a galvanometer. **3.** Particles of a substance and the liquid in which it is mixed, but not dissolved. **4.** The substance, as defined in **3**.

suspension galvanometer A meter with a light-beam apparatus for lengthening the arc through which the pointer travels. When the beam of light is cast a long distance, a tiny movement of the coil will cause considerable movement of the image.

sustained oscillations Oscillations that continue as long as power is supplied to the oscillation generator. Also see CONTINUOUS WAVE. Compare DAMPED OSCILLATIONS.

sustaining voltage The voltage at which second-collector breakdown occurs in a transistor (see SECOND BREAKDOWN).

S video In animation, a scheme that separates brightness and color. It can enhance the video in some applications.

SW Abbreviation of SHORTWAVE.

sw Abbreviation of SWITCH. (Also, S or s.)

swamping resistor **1.** A noninductive resistor connected in parallel with the input circuit of a class-B linear amplifier for automatic regulation of the excitation. **2.** A resistor connected in series with the emitter of a bipolar transistor to minimize the effects of temperature-induced variations in junction resistance.

swarf The string of material that threads off a disc during sound recording.

S_W band A section of the S BAND, extending from 3400 to 3700 MHz.

sweep **1.** To deflect the electron beam in a cathode-ray tube, usually horizontally, to provide a time base. **2.** The circuit for achieving the particular deflection described in **1**.

sweep circuit A circuit, such as a deflection generator (e.g., a sawtooth oscillator), for producing a sweep signal. Also see SWEEP.

sweep delay In an oscilloscope, the process of initiating the sweep of the electron beam at some selected instant after the signal has started.

sweep-delay circuit In an oscilloscope or radar, the circuit for delaying the sweep until the start of the signal. Also see DELAYED SWEEP.

sweeper **1.** See SWEEP GENERATOR. **2**. See SWEEP-SIGNAL GENERATOR.

sweep frequency **1.** The frequency at which the electron beam in a cathode-ray tube is deflected along the reference axis. **2.** The frequency at which the carrier frequency is increased and decreased by a sweep-signal generator. **3.** In an oscilloscope, the number of times that the trace moves across the screen in one second. It is equal to the reciprocal of the SWEEP PERIOD.

sweep generator **1.** A device that causes the electron beam in a cathode-ray tube to scan at a known speed. **2.** An oscillator that generates a signal that rapidly varies in frequency. It is used for the testing and adjustment of bandpass filters and other selective circuits.

sweeping receiver See SCANNING RECEIVER.

sweep magnification In an oscilloscope, increasing or multiplying the sweep frequency, thus reducing the time per horizontal division. This increases the maximum frequency of waveforms that can be analyzed, and allows closer inspection of high-frequency signal components.

sweep magnifier In an oscilloscope, a circuit for achieving sweep magnification.

sweep oscillator See SWEEP GENERATOR.

sweep period The duration, in seconds, of one complete cycle of sweep signal in an oscilloscope. It is equal to the reciprocal of the SWEEP FREQUENCY.

sweep signal The (usually linear, sawtooth) signal used to sweep the beam of an oscilloscope tube. Also see SWEEP, **1**, **2**.

sweep-signal generator A signal generator that supplies a signal whose frequency varies automatically and periodically throughout a given band.

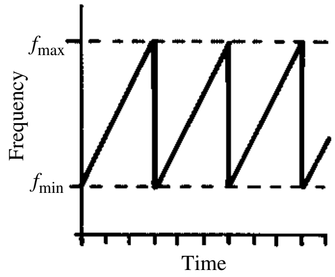

sweep-signal generator

sweep test A method of testing the attenuation-vs.-frequency characteristics of a selective circuit, using a radio-frequency sweep generator.

sweep time The actual time required for a single sweep by a deflecting signal; $t = 1/f$, where t is sweep time in seconds, and f is sweep frequency in hertz.

sweep voltage The peak voltage amplitude of the sweep signal.

SWG Abbreviation of *standard wire gauge.*

swing The maximum change exhibited by a varying quantity (e.g., *amplitude swing* and *frequency swing*).

swinging choke A filter choke that exhibits relatively high inductance when low current flows through it, and lower inductance when high current flows through it. This inductance, which swings under conditions of varying load current, permits the use of a high-resistance bleeder resistor. Compare SMOOTHING CHOKE.

Swiss-cheese packaging A method of packaging an electronic circuit, in which components are inserted into the assembly through holes drilled or punched in parallel, stacked printed-circuit boards.

switch **1.** A circuit or device (electronic, electromechanical, or mechanical) for opening and closing a circuit or for connecting a line to one of several different lines (e.g., *rotary selector switch*). **2.** To change the logic state of a circuit or device. **3.** In a computer program, a branch instruction directing the program to a line number dependent on the value of a variable or result (e.g., BASIC's GOTO). **4.** To cause an electrical circuit to change state, as from low to high or vice versa.

switch current **1.** The current flowing through a switch. **2.** The current flowing through a switching diode or transistor. **3.** The minimum current necessary to produce switching of a transistor, specified in milliamperes or microamperes.

switchgear Collectively, devices and systems for making and breaking circuits—either automatically or manually.

switchhook In a telephone set, the spring-and-switch device that engages the line when the receiver is lifted.

switching characteristics Technical data describing the performance and capabilities of switching devices and circuits.

switching circuit An on-off type of circuit containing electronic or mechanical switches.

switching diode See COMPUTER DIODE.

switching frequency The frequency at which a repetitive switch operates. Also see SWITCHING RATE.

switching mode Operation in which a device, such as a transistor or diode functions as a binary digital device, rather than as an analog device. The current is generally either zero (cutoff or pinchoff) or some value that depends on the bias and on the applied voltage.

switching rate The rate (e.g., closures per second) at which a repetitive switch operates. Also see SWITCHING FREQUENCY.

switching speed The time required for a switch to open or close or for a switching device to change states (as from cutoff to saturation). Also see SWITCHING TIME.

switching time The time required, after the application of a pulse, for an electronic switch to change state. Also see SWITCHING SPEED.

switching transistor A transistor designed especially for on-off operation. Such units exhibit short recovery time and low capacitance.

switching voltage The largest voltage that a switching device can handle without malfunctioning.

switch leakage current **1.** The current flowing through a switching device when it is supposed to be nonconducting. **2.** In a switching transistor, for a given voltage, the leakage current between the emitter and collector when the device is supposed to be nonconducting.

SWL Abbreviation of SHORTWAVE LISTENER.

SWR Abbreviation of STANDING-WAVE RATI0.

SWR bridge A four-arm resistance bridge for measuring voltage standing-wave ratio. This radio-frequency bridge has noninductive resistors in three of its arms and the device under test in the fourth arm. The bridge is balanced first with an equivalent noninductive resistor that replaces the device, and the output voltage is noted. Then the device is substituted for the test resistor, and the change in voltage is noted. The standing-wave ratio is determined from the voltage ratio.

SWR meter See SWR BRIDGE.

S_Y band A section of the S BAND, extending from 2600 to 2700 MHz.

syllable compandor A device that compresses or expands the amplitude of an audio signal. The time constant is fast enough to allow response to individual syllables. Compression is generally used at the transmitting station, and expansion at the receiving station.

sym **1.** Abbreviation of *symmetrical.* **2.** Abbreviation of SYMBOL.

symbol **1.** A letter or graphic device representing a quantity or term [e.g., *I* (current), *f* (frequency), etc.]. **2.** A conventional device denoting a mathematical operation (e.g., +, /). **3.** In a circuit diagram, a pictorial device representing a component.

symbolic address An address in a source-language computer program (i.e., the arbitrary label used by the programmer).

symbolic language See SOURCE LANGUAGE.

symbolic logic A system for representing logical relationships, such as those acted upon by computer and switching circuits, by means of symbols that are usually nonnumerical. Also see BOOLEAN ALGEBRA.

symmetrical circuit A circuit having identical configurations on each side of a dividing line, such as the ground bus. A push-pull circuit is an example.

symmetrical communications **1.** Two-way communications in which the volume of transmitted data is the same, or nearly the same, in both directions. **2.** Two-way communications in which the speed of transmitted data is the same, or nearly the same, in both directions. Compare ASYMMETRICAL COMMUNICATIONS.

symmetrical conductivity Identical conductivity for both positive and negative electricity. Compare ASYMMETRICAL CONDUCTIVITY.

symmetrical FET See SYMMETRICAL FIELD-EFFECT TRANSISTOR.

symmetrical field-effect transistor A field-effect transistor whose source and drain terminals can be interchanged without affecting circuit operation. Also called BIDIRECTIONAL TRANSISTOR. Compare UNILATERAL FIELD-EFFECT TRANSISTOR.

symmetrical input See BALANCED INPUT.

symmetrical output See BALANCED OUTPUT.

symmetrical transistor See BIDIRECTIONAL TRANSISTOR.

symmetrical wave A wave whose positive and negative half-cycles are identical in shape and peak amplitude.

symmetry **1.** The condition of having the same shape on each side of an axis. **2.** The condition of conducting positive and negative currents equally well. **3.** The condition in which a circuit is identical on both sides of a reference line, such as the ground line.

sympathetic vibration Resonant vibration of one body in response to the vibration of another body.

sync **1.** Contraction of SYNCHRONIZATION. **2.** Contraction of SYNCHRONISM.

sync amplifier In a television circuit, the amplifier used to increase the amplitude of the sync pulses after they are separated from the composite video signal.

sync generator A circuit that produces the synchronization pulses in a television transmitter.

synchro A dynamo-electric-control device that, when connected to a similar device and the alternating-current power line, permits remote control. Thus, when the rotor of one synchro is turned to a certain position, the rotor of the other assumes the same position. Also see AUTOSYN and SELSYN.

synchrocyclotron A type of cyclotron in which the variation in mass, because of increased velocity, of the charged particles is compensated, resulting in higher energy for the particles.

synchro differential A synchro that receives two input signals and delivers a single output signal. The inputs can be two electrical signals, or one electrical signal and one mechanical signal.

synchrodyne receiver A direct-conversion receiver in which the local oscillator frequency or phase is locked into synchronism with the incoming signal carrier frequency or phase.

synchroflash A flash that is synchronized with the shutter of a camera.

synchro generator The transmitting member of a synchro system.

synchro motor The receiving member of a synchro system.

synchronism **1.** The condition of being in step, as when two motors are running in synchronism with each other and the power frequency, or when two relays open and close in step. **2.** The condition of being in phase, as when two pulses occur simultaneously.

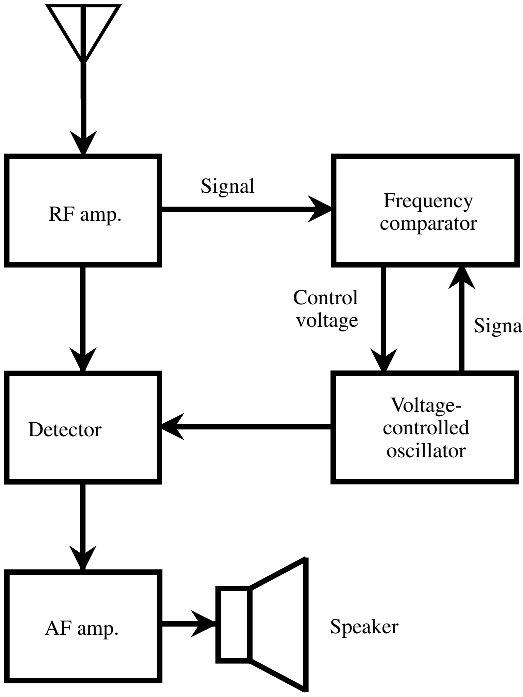

synchrodyne receiver

synchronization The coincidence of one process or operation with another, as in the synchronization of an oscillator frequency by means of an applied standard-frequency voltage, in which case the oscillator frequency becomes that of the standard signal.

synchronized clamping A type of clamping in which an output voltage is maintained at a predetermined fixed value until a synchronizing pulse is applied, whereupon the output follows the input.

synchronized multivibrator See DRIVEN MULTIVIBRATOR.

synchronizer A computer storage device used between two devices transmitting data at different speeds, to counteract this differential (as a buffer).

synchronizing signal A signal used to synchronize another signal, usually in frequency.

synchronous The condition of operating in step (phase) with some reference signal.

synchronous clock **1.** An alternating-current clock driven by a synchronous motor. Although 60-Hz models are common, such clocks are not restricted to low-frequency ac operation; 1-kHz types, for example, are used in some primary frequency standards. **2.** The timing source in a synchronous computer.

synchronous computer A computer whose operations are timed by single-frequency clock signals.

synchronous contacts The rectifying contacts of a synchronous vibrator (see VIBRATOR-TYPE RECTIFIER).

synchronous converter A synchronous machine that can run on alternating current and generate direct current, or vice versa. Also called ROTARY CONVERTER.

synchronous gate A gate whose output is synchronized, according to the input signal.

synchronous generator An alternator operating in synchronism with one or more other alternators. Also see SYNCHROSCOPE, **2**.

synchronous induction motor An alternating-current motor that is intermediate between the fractional-horsepower reluctance motor and the multiple-horsepower three-phase, synchronous motor. The synchronous induction machine starts like an induction motor and runs like a synchronous motor.

synchronous inputs In a computer flip-flop, inputs that accept pulses only at the command of the clock.

synchronous machine See SYNCHRONOUS INDUCTION MOTOR.

synchronous motor See SYNCHRONOUS INDUCTION MOTOR.

synchronous network A communications network in which all clocks are set so that they run at the same rate, their increments are identical in duration, and transitions occur simultaneously or with a specified phase difference. Such a system allows for greatly enhanced signal-to-noise ratio for a given amount of transmitter power, and also reduces the bandwidth necessary for a single signal so that many more signals can be placed in a given frequency band.

synchronous orbit See GEOSTATIONARY ORBIT.

synchronous satellite See GEOSTATIONARY SATELLITE.

synchronous speed For an alternating-current (ac) machine, the speed corresponding to the ac frequency.

synchronous transmission A method of signal transmission in which individual symbols are sent at a specified rate, according to a clock that also governs the receiver.

synchronous vibrator See VIBRATOR-TYPE RECTIFIER.

synchroscope **1.** An oscilloscope having a high-speed sweep triggered by a synchronizing signal. Such an instrument is valuable for viewing high-speed pulses. **2.** A pointer-type instrument used to indicate the synchronism between two power alternators.

synchro system A circuit or system using synchros for the transmission and reception of positioning signals. Also see SYNCHRO.

synchrotron A particle accelerator that uses a high-frequency electrostatic field and a low-frequency magnetic field to impart very high velocity to the particles.

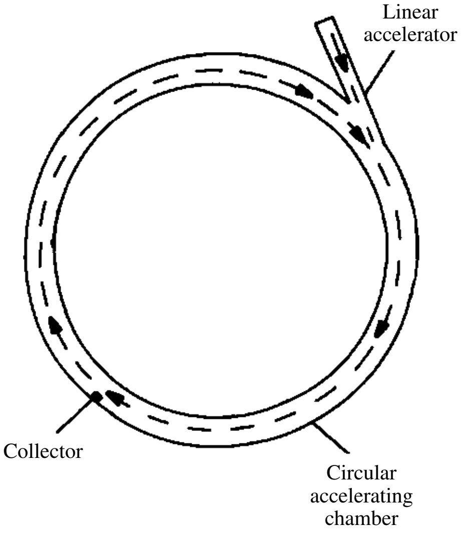

synchrotron

sync pulse **1.** A pulse used to control the frequency or repetition rate of an oscillator or other generator. **2.** In a television system, a pulse transmitted as part of the composite video signal to control scanning. Also see HORIZONTAL SYNC PULSE and VERTICAL SYNC PULSE.

sync separator In a television receiver circuit, a stage used to separate and deliver the sync pulses from the composite video signal. See, for example, DIODE SYNC SEPARATOR.

sync signal See SYNCHRONIZING SIGNAL.

sync takeoff The point in the video amplifier circuit of a television receiver at which the composite video signal is sampled to extract the sync pulses.

syntax **1.** The rules by which computer program statements are structured. **2.** The way that a written or spoken sentence is constructed. It is important in speech recognition and speech synthesis.

synthesis The rigorous (usually mathematical) design of an electronic circuit or device and the accurate prediction of its performance. Compare ANALYSIS.

synthesizer **1.** See SIGNAL SYNTHESIZER. **2.** A circuit synthesizer (i.e., a device that allows a wide variety of circuits to be set up temporarily or simulated, for testing and evaluating). Sometimes, a specially programmed computer serves this purpose. **3.** A keyboard on which music can be played, and whose output can be adjusted to simulate the sounds of various musical instruments. **4.** See MOOG SYNTHESIZER.

synthetic bass An apparent accentuation of bass notes resulting from intermodulation distortion in an amplifier.

synthetic crystal An artificially produced crystal, such as synthetic quartz.

synthetic resin An artificially produced resin. Also see THERMOPLASTIC MATERIAL and THERMOSETTING MATERIAL.

syntony See RESONANCE.

syst Abbreviation of SYSTEM.

system **1.** An integrated assemblage of hardware and/or software elements operating together to accomplish a prescribed end purpose (e.g., *servo system*, *operating system*, and *communications system*). **2.** A methodology incorporating fixed and ordered procedures for accomplishing an end purpose. **3.** A self-contained computer workstation.

systematic error See CUMULATIVE ERROR.

system engineering See SYSTEMS ENGINEERING.

system of units A set of fundamental units for defining the magnitudes of all physical variables. The most common system of units is the standard international (SI) system.

systems analysis In computer system operation, analyzing the way something is done and devising a better alternative by isolating the problem area, scrutinizing the system as it stands, studying what is thereby disclosed, devising the alternate application of software and/or hardware, disseminating the revised operational procedure, and overseeing the implementation of the new method.

systems engineering The branch of engineering devoted to the design, development, and application of complete systems. The approach takes into consideration all elements in a system or process and their integration.

systems flowchart A flowchart showing the interrelationship of activities in a system.

S_Z band A section of the S BAND, extending from 3900 to 4200 MHz.

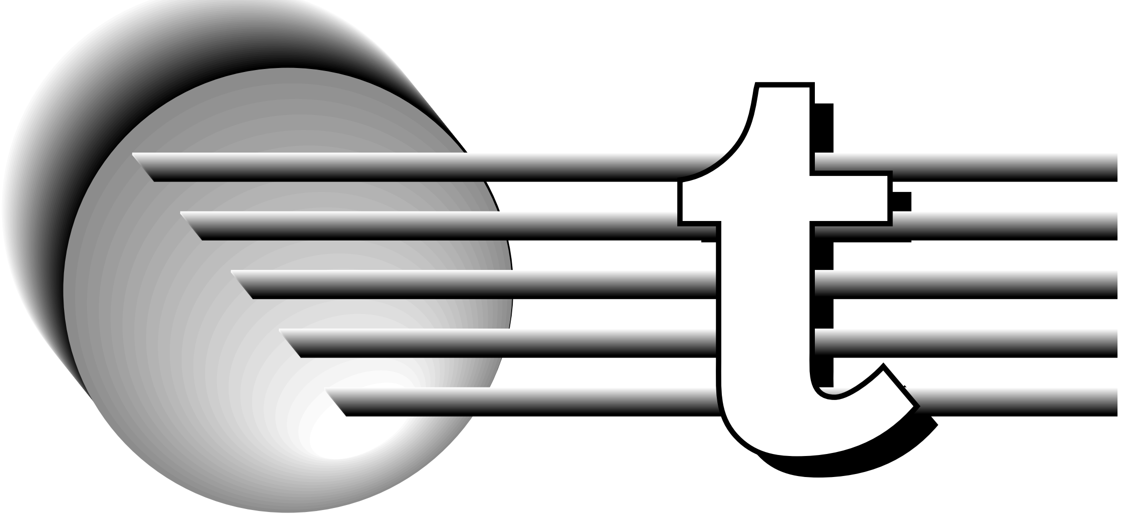

T **1.** Symbol for TRANSFORMER. **2.** Abbreviation of prefix TERA-. **3.** Symbol for *thermodynamic temperature.* **4.** Symbol for TRITIUM. **5.** Abbreviation of TON. (Also, t and tn.) **6.** Abbreviation of TESLA. **7.** Symbol for KINETIC ENERGY. **8.** Symbol for PERIOD. **9.** Symbol for *true.*

t **1.** Symbol for TIME. **2.** Abbreviation of TON. (Also, T and tn.) **3.** Symbol for CELSIUS TEMPERATURE. (Also, T.) **4.** Abbreviation of TARGET. **5.** Abbreviation of *technical.* **6.** Abbreviation of TENSION.

Ta Symbol for TANTALUM.

tab On the keyboard for a computer, typewriter, terminal, or word-processing system, a key that moves the cursor a specified number of spaces toward the right. It also performs various other functions in menu-driven or graphical computer interfaces.

table In an internal or external computer memory, an array (i.e., a list or matrix) of data that can be recalled using keys (e.g., single- or double-subscripted variables).

table look-at Abbreviation, TLA. In computer operations, finding the position of a data item in a table by implementing an algorithm.

table look-up Abbreviation, TLU. In computer operations, locating items in a table by inspecting what is in the table, key by key.

tabulate In data processing, to combine the totals for data item groups having the same key.

tabulation **1.** The printout of what has been tabulated (see TABULATE). **2.** The computer-program-directed movement of the cursor on a cathode-ray-tube display, or of a typewriter carriage, to certain positions in a line.

tabulator See TAB.

tacan A pulse-type UHF air navigation system in which a station is interrogated by signals from an aircraft to provide bearing and range information. The name is an acronym for *tactical air navigation.*

tach Abbreviated form of *tachometer.*

tachometer See ELECTRONIC TACHOMETER.

tachometer generator A small, dynamo-type electric generator that delivers a voltage proportional to the rotational speed of a shaft to which it is attached.

tachyon A high-speed subatomic particle thought to move faster than the speed of light.

tactical air navigation See TACAN.

tactical radar A radar system used in military operations.

tactile sensor A device that provides an intelligent machine with a sense of “touch”: temperature, pressure, force, texture, and torque. It is important in robotics, and also in some computer applications, such as virtual reality (VR).

T-adapter See TEE-JUNCTION.

tag **1.** In data-processing and computer operations, the identification of digits or characters forming part of a record. **2.** An encoded price tag (i.e., a passive transponder or barcode strip). It is commonly used in retail stores.

tag converter A device that senses the information on tags (see TAG, **2**) and transfers it to a computer system.

tail **1.** The decay of a waveform from maximum amplitude to zero amplitude. **2.** Any pulse that follows a main pulse as a result of the main pulse.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

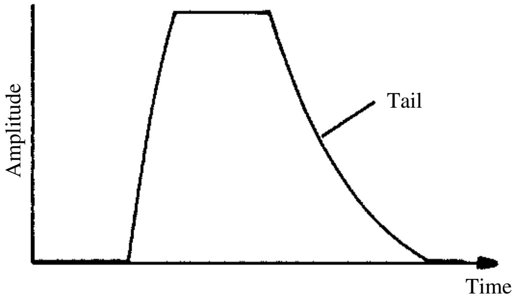

tail

taillight monitor An electronic device for warning a motorist of taillight failure.

tail pulse **1.** A pulse with a fast rise time, but a slow decay time. **2.** See TAIL, **2**.

takedown In computer operations, the process of clearing output peripherals for upcoming program runs.

take-down time The duration of the TAKEDOWN process.

take-up reel In a reel-to-reel tape recorder/reproducer, the reel on which the tape accumulates during recording or reproduction.

talk-back circuit See INTERPHONE.

talk-listen switch A transmit-receive switch in an intercommunication system (see INTERCOM).

talk power See SPEECH POWER.

tally **1.** To obtain a sum or total. **2.** The rows of operands, subtotals, and totals that an adding machine prints.

tally reader A device that, via optical character recognition (OCR), can read the digits and symbols on a tally (see TALLY, **2**).

tamper switch A device that closes a circuit or actuates an alarm when a certain set of conditions is altered.

tan Abbreviation of TANGENT.

$\tan^{-1}$ Arc tan (inverse tangent function).

tandem transistor An assembly of two series-connected transistors in the same envelope.

tangent **1.** Abbreviation, tan. The ratio of the side opposite to the side adjacent to an acute angle in a right triangle. **2.** A line that intersects a curve at a single point without crossing the curve. **3.** A plane that intersects a curved surface at a single point.

tangent galvanometer A galvanometer in which the current is proportional to the tangent of the angle of deflection. Compare SINE GALVANOMETER.

tangential mode In acoustics, the reflection of sound waves from four surfaces in a room having six interior surfaces.

tanh Abbreviation of HYPERBOLIC TANGENT.

tank **1.** A parallel-resonant inductance-capacitance circuit in the output of a radio-frequency amplifier. **2.** A circuit in which is stored electrical energy of frequencies in a range whose midpoint is resonance for the circuit. **3.** See MERCURY DELAY LINE.

tantalum Symbol, Ta. A metallic element of the vanadium family. Atomic number, 73. Atomic weight, 180.95. Tantalum is used in the elements of some electron tubes and in some electrolytic capacitors.

tantalum capacitor A type of electrolytic capacitor that uses tantalum rather than aluminum. The tantalum can be foil, as is the aluminum in a conventional electrolytic capacitor. It might also take the form of a porous pellet, the irregular surface of which provides a large area in a small volume. An extremely thin oxide layer forms on the tantalum. These capacitors have high reliability and excellent efficiency, and are used in military applications because they have a low failure rate. They can be used in audio and digital circuits in place of aluminum electrolytic capacitors. The main disadvantage of tantalum capacitors is that they are comparatively expensive. Compare CERAMIC CAPACITOR, ELECTROLYTIC CAPACITOR, MICA CAPACITOR, PAPER CAPACITOR, PLASTIC-FILM CAPACITOR.

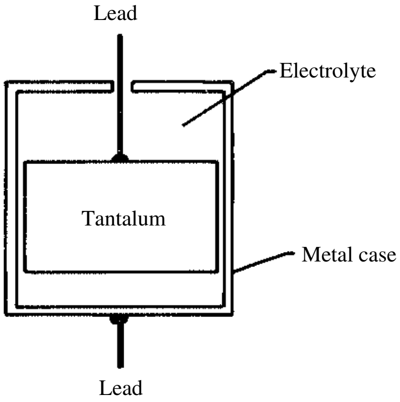

tantalum capacitor

tantalum detector A radio-frequency contact detector consisting essentially of a fine tantalum wire whose point lightly touches the surface of a small pool of mercury.

tantalum-nitride resistor A resistor consisting of a thin film of tantalum nitride deposited on a substrate. Also see THIN FILM.

T-antenna See TEE-ANTENNA.

tap A connection made to an intermediate point on a coil, resistor, or other device. See, for example, CENTER TAP and TAPPED COMPONENT.

tap changer A device that facilitates adjustment of the turns ratio of a transformer by changing the tap on the primary winding, the secondary winding, or both. It also facilitates the adjustment of the impedance-transfer ratio. It can be used for power supplies or audio- or radio-frequency transformers.

tape **1.** Magnetic tape. **2.** Insulating tape. **3.** To make a magnetic tape recording (audio, video, and/or digital data).

tape cable **1.** A form of cable in which all of the conductor centers lie in the same plane. **2.** A flat cable, commonly used in situations, where repeated flexing occurs.

tape cartridge A holder and the reel of blank or prerecorded magnetic (⅛-inch audio or wider video) tape it contains, which can be inserted into a recorder/reproducer without having to thread or otherwise handle the tape for either playing or rewinding.

tape comparator In a data-processing or computer system, a machine that compares tapes generated from the same input, for differences in the data thereon; it is a character-by-character process.

tape core A ring-type magnetic core made by tightly winding metal tape in several layers for the desired thickness.

tape counter See POSITION INDICATOR.

tape deck In a tape recorder/reproducer, the complete tape-transport mechanism (drive, heads, equalization circuitry, and preamplifiers).

tape drive **1.** See TAPE TRANSPORT. **2.** See TAPE DECK. **3.** A tape recorder/reproducer for computer data, used for backup and/or archiving.

tape file A data file recorded on magnetic tape.

tape group An assembly of two or more tape decks.

tape label On a reel or cassette of magnetic tape containing a data file, a record at the beginning or end that contains information about the file.

tape loop An endless loop of magnetic tape.

tape magazine See TAPE CARTRIDGE.

tape mark **1.** A character that subdivides the magnetic tape file on which it is recorded. Also called CONTROL MARK. **2.** A character marking the end of a length of magnetic tape on a reel. Also called END-OF-TAPE MARK.

tape plotting system In computer operations, a system for operating a digital incremental plotter using the information on magnetic tape.

taper In a potentiometer or rheostat, the rate of change in resistance during uniform rotation of the shaft. See, for example, LINEAR TAPER and LOG TAPER.

tape recorder A machine for recording audio, video, or data signals on magnetic tape; it can usually also play back the recorded material.

tapered potentiometer A potentiometer having a tapered resistance winding (see TAPERED WINDING).

tapered winding A resistance winding (in a rheostat or potentiometer) in which the resistance change per unit length of winding is nonuniform (see, for example, LOG TAPER). However, a winding having uniform resistance change is often called LINEAR TAPER.

tape skew A condition in which a magnetic tape passes between the recording or playback heads in an irregular way. The result is that the various channels, or tracks, are not always perfectly aligned.

tape sort The (computer) operation of sorting data in a magnetic tape file into a record-key-determined sequence.

tape splicer A mechanism that aligns and secures the overlapping ends of broken magnetic tape so that they can be cut (often with an integral cutter) to form butted ends, and taped into a splice.

tape station See TAPE DECK.

tape-to-head contact See HEAD-TO-TAPE CONTACT.

tape transmitter A transmitter that receives its signal input from a recorded tape.

tape transport In a tape deck or reproducer, the device that moves the tape past the heads.

tape unit **1.** See TAPE DECK. **2.** See TAPE GROUP.

tape verifier In computer operations, a device that checks the integrity of data on paper tape through comparison with an original document.

tape width In magnetic tape, the dimension perpendicular to tape travel; in general, the greater the tape width, the more tracks the tape can contain.

tape-wound core See TAPE CORE.

tapped coil An inductor to which one or more intermediate connections (taps) are made at appropriate turns to provide intermediate values of inductance.

tapped component A coil, transformer, choke, resistor, or other component in which an intermediate connection is made. See, for example TAPPED COIL.

tapped inductor See TAPPED COIL.

tapped resistor A resistor in which one or more intermediate connections (taps) are made to appropriate parts of the resistance element to provide steps of resistance.

tapped transformer A transformer having one or more tapped windings.

tapped winding A transformer or choke winding with one or more taps. Also see TAP and TAPPED COIL.

tap switch A multiposition switch used to connect an external circuit to various taps on a component. Also see SELECTOR SWITCH and TAP.

target **1.** The bombarded electrode in an X-ray tube. **2.** The scanned storage element in a television camera tube. **3.** In radar operations, the scanned object. **4.** An object intended for nuclear-particle bombardment. **5.** A goal—especially in a production process (deadline, desired number of units, etc.).

target acquisition **1.** The moment at which a target comes within the range of a radar system. **2.** The observation of a new target on a radar screen.

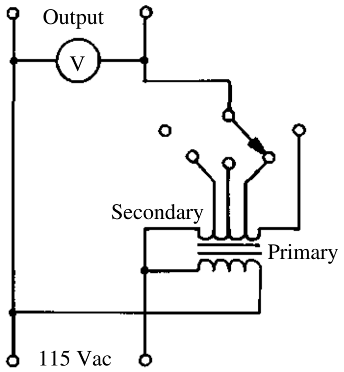

tapped transformer

target discrimination The extent to which a radar system can distinguish between two targets that are close together. Also called *target resolution.* It is specified in linear units (such as meters, kilometers, or miles).

target identification Any method by which the identity of a radar target is determined.

target voltage In a television camera tube, the backplate-to-cathode voltage.

task environment The characteristics of the space in which an autonomous robot works. It depends on such factors as the intended application(s), the required speed at which the robot(s) must work, and the human/robotic or computer/robotic interface.

task-level programming In robotics, the writing of programs to perform sequences of actions. A complex process, it is a primitive level of artificial intelligence (AI). An example is a program that directs a robot to prepare and serve a meal.

taut-band meter A movable-coil meter in which the conventional spiral springs of the coil are replaced by two tightly stretched, thin, straight metal ribbons, whose twist provides torque that returns the pointer to zero after a deflection.

Tb Symbol for TERBIUM.

BS Abbreviation of *talk between ships.*

Tc Symbol for TECHNETIUM.

TCCO Abbreviation of TEMPERATURE-CONTROLLED CRYSTAL OSCILLATOR.

T circuit See TEE NETWORK.

T circuit parameters See R PARAMETERS.

T circulator See TEE CIRCULATOR.

TCL Abbreviation of TRANSISTOR-COUPLED LOGIC.

TCM Abbreviation of *thermocouple meter* (see THERMOCOUPLE-TYPE METER).

TDM Abbreviation of TIME-DIVISION MULTIPLEX.

TDR **1.** Abbreviation of TIME-DELAY RELAY (see DELAY RELAY). **2.** Abbreviation of TIME-DOMAIN REFLECTOMETRY.

TDS Abbreviation of TIME-DELAY SPECTROMETRY.

TE Abbreviation of *transverse electric* (see, for example, TRANSVERSE ELECTRIC MODE).

Te Symbol for TELLURIUM.

teach box A robot-control device via which an operator can program the machine to carry out specific movements. The operator manipulates controls, and the robot's computer stores the data. When the data is accessed, the robot reproduces the motions.

tearing In a television picture, the abnormal condition in which poor synchronization causes the horizontal lines to be irregularly displaced. The effect resembles cloth being torn.

technetium Symbol, Tc. A metallic element produced artificially. Atomic number, 43. Atomic weight, 98. Formerly called *masurium.*

technician **1.** A person who repairs faulty electronic equipment. **2.** A person who assists with the design and debugging of a system prototype. **3.** A person who operates an electronic system.

technocentrism Overdependence on, and/or obsession with, the products of technology, particularly computers.

tee Pertaining to a network, connection, or configuration whose geometric shape or schematic representation resembles an uppercase letter T.

tee adapter See TEE JUNCTION.

tee antenna An antenna consisting of a horizontal radiator with a vertical lead-in or feeder connected to its center point.

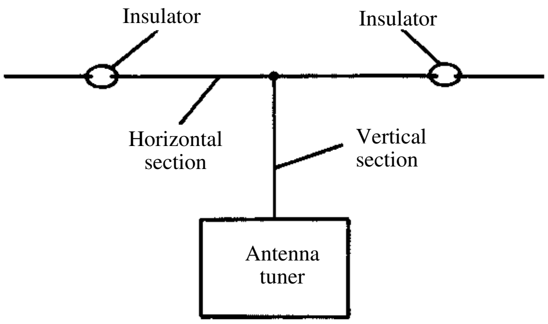

tee antenna

tee circuit See TEE NETWORK.

tee circulator In microwave systems, a tee-shaped junction of three waveguides with a ferrite post at the junction.

tee-equivalent circuit An equivalent circuit in which the components are arranged in the form of a tee. See, for example, TEE NETWORK and R PARAMETERS.

tee junction **1.** A tee-shaped splice between two wires. **2.** A tee-shaped fixture for splicing one coaxial line perpendicularly to another. **3.** A tee-shaped section for joining one waveguide perpendicularly to another. Also called WAVE GUIDE TEE.

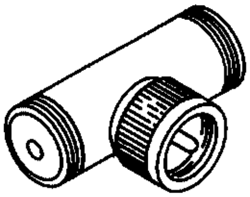

tee junction

tee network A three-terminal network resembling a tee.

tee pad A three-resistor pad in which two series resistors and a shunt resistor are arranged to form a tee.

tee switch A combination of three switches arranged to form a tee; two switches are in series with a shunt switch in between. If the series switches are open and the shunt switch is closed, isolation is greatly improved, compared to a single series switch.

Teflon FEP A plastic insulating material. Dielectric constant, 2.1. Dielectric strength, 2800 V/mil.

Teflon TFE A plastic insulating material. Dielectric constant, 2.2. Dielectric strength, 600 V/mil.

tel **1.** Abbreviation of TELEPHONE. **2.** Abbreviation of TELEGRAPH. **3.** Abbreviation of TELEGRAM.

telautograph A device that transmits and receives handwriting, drawings, and similar material. At the receiver, a pen follows the movements of a similar pen at the transmitter.

teleammeter A TELEMETER for measuring current generated at a remote point.

telecamera See TELEVISION CAMERA.

telecast A television program for general reception. The term a contraction of *television broadcast.*

telechir A remotely controlled, autonomous robot. See TELEOPERATION and TELEPRESENCE.

telecommunication Communication, usually between widely separated points, by electrical or electronic means.

telecontrol See REMOTE CONTROL and TELEOPERATION.

telefacsimile See FACSIMILE.

telegram Abbreviation, tel. A (usually printed-out) message transmitted and received via telegraph or teletypewriter. Compare CABLEGRAM and RADIOGRAM.

telegraph Abbreviation, tel. An instrument for transmitting and receiving messages by means of telegraphy. In its simplest form, it consists of a key and a sounder powered by a battery. Also see PRINTING TELEGRAPH, **1**, **2**.

telegraph channel **1.** The frequency band assigned to a particular telegraph station. **2.** The frequency band occupied by a telegraph signal.

telegraph code Broadly, Morse code. Wire telegraphy often uses a special version, such as the American Morse code.

telegraph key See KEY, **1**.

telegraph sounder See SOUNDER.

telegraph system A complete, integrated, coordinated arrangement of equipment for communication by means of telegraphy. Included are telegraph keys or keyers, sounders or printers, relays, switchboards, wire lines and cables, and power supplies.

telegraphy The branch of electrical communications that deals with the transmission and reception of messages by means of prearranged codes—especially over wires. Also see MORSE, **1**, **2**, **3**; MORSE CODE; and WIRE TELEGRAPHY.

telemeter **1.** An indicating instrument that measures the value of a quantity generated at a distant point or measures and transmits the value. **2.** The action afforded by the device in **1**.

telemetering See TELEMETRY.

telemeter receiver See TELEMETRIC RECEIVER.

telemeter transmitter See TELEMETRIC TRANSMITTER.

telemetric receiver A system that selects, amplifies, and demodulates or rectifies a radio signal, and actuates indicating instruments, recorders, or data processors.

telemetric transmitter A specialized transmitter that generates radio-frequency (RF) power, adds to it signals delivered by data transducers, and delivers the modulated power to an antenna for transmission to a distant telemetric receiver.

telemetry The transmission of data signals over a distance, either by radio or wire, and the reception and application of the signals to indicating instruments, recorders, etc.

teleoperation The remote control of autonomous robots. A human operator can control the speed, direction, and other movements of a robot from some distance away. See also TELEPRESENCE.

telephone Abbreviation, tel. An instrument for transmitting and receiving messages by means of telephony. In its simplest form, it consists of a microphone, earphone, switching and ringing devices, wire line or cable, and power supply. Also see HANDSET.

telephone accessories Devices (such as answering machines, speaker phones, facsimile machines, etc.) used in conjunction with a telephone set.

telephone amplifier A small audio amplifier, usually with a self-contained loudspeaker, for increasing the sound volume of a telephone output. Some amplifiers of this kind are connected to the telephone line, and others receive sound from the telephone receiver.

telephone-answering machine See ANSWERING MACHINE.

telephone bypass capacitors A set of fixed capacitors installed between each wire in a telephone line and an electrical ground to bypass radio-frequency energy. It can reduce or prevent radio-frequency interference (RFI) to telephone sets and data terminals.

telephone data set A device that converts signals from a data terminal for passage over a telephone circuit to a data-processing center.

telephone dialer In a security system, a circuit that automatically dials one or more telephone numbers, alerting the recipient(s) that an emergency exists at a given location.

telephone induction coil A small telephone-to-line impedance-matching transformer used in telephone systems.

telephone patch See PHONE PATCH.

telephone pickup A device for receiving conversations from a telephone to which it isn't directly connected.

telephone plug See PHONE PLUG.

telephone-radio patch See PHONE PATCH.

telephone receiver The handheld part of a telephone set, containing the microphone and earphone, and, in some cases, the dialing keypad.

telephone repeater An amplifier and associated equipment used to boost the amplitude of a telephone signal at an appropriate point along the line.

telephone service entrance The point at which the telephone wiring in a house or building connects to the outside telephone line.

telephone silencer A device for muting a telephone or its bell.

telephone system A complete, integrated, and coordinated arrangement of equipment for communication by means of telephony. Included are telephones, switchboards and associated equipment, wire lines and cables, and power supplies. Also see DIAL TELEPHONE SYSTEM, HANDSET, PRIVATE AUTOMATIC EXCHANGE (PAX), PRIVATE BRANCH EXCHANGE (PBX), and TELEPHONE.

telephone test set See PHONE TEST SET.

telephone transmitter The sound pickup unit (microphone) of a telephone. Also see TRANSMITTER, **2**.

telephony The branch of electrical communication dealing with the transmission and reception of sounds—especially over wires. Also see WIRE TELEPHONY.

Telephoto **1.** The transmission and/or reception of photographs by means of FACSIMILE. **2.** A photograph transmitted and/or received by means of FACSIMILE.

telephoto lens A camera lens providing a telescopic effect.

Telephoto receiver See FACSIMILE RECEIVER.

Telephoto transmitter See FACSIMILE TRANSMITTER.

telepresence An advanced form of robotic TELEOPERATION, in which a human operator has the impression of being at the robot's location. It includes vision systems, pressure sensors, sound sensors, tactile sensors, and electromechanical control devices. The operator uses, or wears, equipment similar to that used for VIRTUAL REALITY (VR).

teleprinter A terminal telegraph printing machine. Also see PRINTING TELEGRAPH, **1**, **2**.

TelePrompTer A device that presents a running display on a screen before a television announcer, performer, or speaker, of dialogue.

teleran A ground-to-air communications system. Ground-based radar pictures are transmitted, via television, to aircraft.

telescoping antenna A vertical antenna consisting of separate lengths of metal tubing of progressively smaller diameter so that one can slide into another. The antenna can be pulled out to full length or compressed to the length of the largest-diameter section.

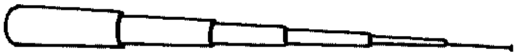

telescoping antenna

teletext A method of communication in which a video image is sent of a page of printed material.

telethermoscope A device for measuring the temperature of distant objects.

Teletorque See AUTOSYN and SYNCHRO.

Teletype **1.** The transmission and/or reception of text messages in digital form, displayed or printed at a distant location. **2.** See TELETYPEWRITER.

teletype grade A term descriptive of a circuit having the quality necessary for communication via telegraphy.

Teletypesetter An electronic system for operating a distant Linotype.

teletypewriter A variety of printing telegraph using electric typewriters and associated equipment. The message is typed on the keyboard at the transmitting station and is typed out in corresponding letters at the receiving station. The same typewriter is able to send and receive messages. Also see RADIOTELETYPEWRITER.

teletypewriter exchange Abbreviation, TWX. A center for switching and routing teletypewriter communications. Also see TELETYPEWRITER.

teleview To observe a scene or program via television.

televise **1.** To convert a scene into a television signal. **2.** To broadcast a scene or program via television.

television **1.** Abbreviation, TV. The transmission and/or reception, via electromagnetic fields, wire cable, and/or fiberoptic cable, of images, usually with sound. **2.** A system for receiving signals, as defined in **1**. **3.** Video programs or data, with or without sound, transmitted via electromagnetic fields, wire cable, and/or fiberoptic cable.

television band See UHF TELEVISION CHANNELS and VHF TELEVISION CHANNELS.

television camera The pickup device that scans a scene and delivers a series of electrical signals that can be used to reconstruct the image on the screen of a picture tube.

television-camera tube See CAMERA TUBE, ICONOSCOPE, and ORTHICON.

television channel A radio-frequency band allocated exclusively for the transmission of a television signal. In conventional broadcasting, this band is usually 6 MHz wide. See, specifically, UHF TELEVISION CHANNELS and VHF TELEVISION CHANNELS.

television engineer A trained professional skilled in video electronics engineering, as well as in basic engineering and associated subjects.

television engineering The branch of electronics engineering devoted to the theory and application of television.

television interference Abbreviation, TVI. Interference to the reception of television signals, usually occasioned by signals from radio services or computers, or by electrical noise.

televisor **1.** A television transmitter or receiver. **2.** A person or entity responsible for the broadcast of television programs or signals.

televoltmeter A TELEMETER for measuring voltage generated at a remote point.

telewriter See TELAUTOGRAPH.

Telex **1.** A teleprinter system that operates via the telephone lines, and was once commonly used by businesses for sending and receiving short messages. It has largely been supplanted by computerized data communications systems. **2.** A hard-copy message sent or received by such a system.

telluric currents Also called *terrestrial currents.* A flow of electrical charge carriers, primarily electrons, in the earth.

tellurium Symbol, Te. A rare, metalloidal element related to selenium. Atomic number, 52. Atomic weight, 127.60.

TE mode See TRANSVERSE ELECTRIC MODE.

temp **1.** Abbreviation of *temperature.* (Also, T.) **2.** Abbreviation of *temporary.* **3.** Abbreviation of TEMPLATE.

temperament The tuning of a keyboard-type musical instrument to produce a nearly perfect diatonic scale.

temperature Symbol, T. A quantitative measure of the heat exhibited by an object or phenomenon. Also see THERMOMETER SCALE.

temperature coefficient A figure that states the extent to which a quantity drifts or varies under the influence of changing temperature. It is generally expressed in percent per degree (%/°C) or in parts per million per degree (ppm/°C).

temperature-compensated crystal oscillation Oscillation in a crystal oscillator, in which the crystal and/or circuit is automatically compensated against frequency drift caused by temperature change.

temperature-compensating component A circuit component, such as a capacitor or resistor, whose temperature coefficient is equal in magnitude and opposite in sign to that of a conventional component to which it is connected to cancel temperature-induced variation in the latter's value.

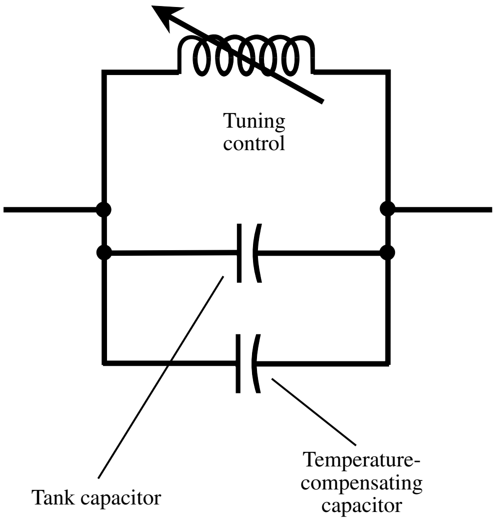

temperature-compensating component

temperature compensation The use of a device, such as a temperature-compensating component, to correct a temperature-induced deviation in the value of a conventional component.

temperature control **1.** The adjustment of temperature. **2.** The automatic maintenance of temperature at a desired level, as in a temperature-controlled oven. **3.** A device for controlling temperature, as defined in **1** or **2**.

temperature-controlled crystal oscillator Abbreviation, TCCO. A high-precision crystal oscillator in which the crystal plate (and sometimes the circuitry, as well) is held at constant temperature.

temperature degree See DEGREE, **2** and THERMOMETER SCALE.

temperature derating Deliberate reduction of operating current and/or voltage of a device to a specific temperature to ensure proper operation. Also see DERATING, DERATING CURVE, and DERATING FACTOR.

temperature gradient A range of temperature variation, such as the rate of change of temperature, with respect to change of power dissipation, or the rate of change of temperature with spatial displacement.

temperature inversion See INVERSION, **1**.

temperature meter An indicator (usually a direct-current voltmeter or millivoltmeter) whose scale reads directly in degrees.

temperature scale See THERMOMETER SCALE.

temperature-sensitive resistor See THERMISTOR.

temperature shock See THERMAL SHOCK.

temperature-to-voltage converter A circuit or device, such as a thermocouple, that delivers an output voltage proportional to an applied temperature.

template **1.** A diagram, usually drawn on paper, to show the locations at which components should be placed or tasks should be performed. It is taped or cemented temporarily to the work, and the points are transferred to the latter by prick-punching. **2.** A stencil-like plate with alphanumeric and circuit symbols, used as a drafting aid. Sometimes called "drafting stencil."

temporary magnet **1.** A body that exhibits magnetism only briefly after it has been exposed to another magnet. Compare PERMANENT MAGNET. **2.** See ELECTROMAGNET.

temporary storage **1.** In computer and data-processing operations, the storage of data or instructions only until they are needed. Also called INTERIM STORAGE. **2.** Locations in a computer memory set aside during a program run for holding intermediate results of operations.

TEM wave See TRANSVERSE ELECTROMAGNETIC WAVE.

ten code A set of abbreviations used by two-way radio operators. Each "ten signal" represents a specific statement or query.

tension **1.** A force that tends to stretch, pull tight, or pull apart. Compare STRAIN. **2.** A term referring to VOLTAGE—especially in a utility power transmission line.

ten-turn potentiometer A precision potentiometer whose shaft must be turned through 10 complete revolutions to cover the entire resistance range. Also see HELICAL POTENTIOMETER and MULTITURN POTENTIOMETER.

T-equivalent circuit See TEE-EQUIVALENT CIRCUIT.

tera- Abbreviation, T. **1.** A prefix meaning trillion(s), (i.e., 10^{12}). **2.** A prefix meaning 2^{40}.

tera-electronvolt Abbreviation, TeV. A large unit of energy; 1 TeV = 10^{12} electronvolts. Also see ELECTRONVOLT.

terahertz Abbreviation, THz. A unit of extremely high frequency equal to 10^{12} Hz. Also called *Fresnel.*

teraohm A unit of high resistance, reactance, or impedance equal to 10^{12} ohms.

terawatt Abbreviation, TW. A large unit of power; 1 TW = 10^{12} W.

terbium Symbol, Tb. A metallic element of the rare-earth group. Atomic number, 65. Atomic weight, 158.93.

terbium metals A group of rare-earth metals, including *europium*, *gadolinium*, *terbium*, and occasionally *dysprosium.*

term In an algebraic expression, constants, variables, or combinations of these, separated by operation signs (e.g., the expression $4xy + z$ has two terms).

terminal **1.** A connection point at the input, output, or an intermediate point of a device, or a point at which a voltage is to be applied. **2.** A metal tab or lug attached to the end of a lead for connection purposes. **3.** Pertaining to the end of a series of events, etc. (e.g., terminal tests). **4.** In a data-communications system, a point of data input or output. Also called *data terminal.* If it does not have computing capability of its own, it is often called a *dumb terminal.*

terminal block A group of several terminals, intended for interconnection of circuits, mounted on a solid insulating block.

terminal board An insulating board carrying several lugs, tabs, or screws as terminals (see TERMINAL, **2**). Also see TERMINAL STRIP.

terminal guidance The navigation of a missile or aircraft to help it reach its target or destination.

terminal impedance The internal impedance of a device measured at the input or output terminals.

terminal point of degradation The point at which degradation of a circuit or component is complete. Also see DEGRADATION FAILURE.

terminal repeater A telephone repeater operated at the end of a line.

terminal strip A strip of insulating material, such as plastic or ceramic, on which are mounted one or more screws, lugs, or other terminals. Also see TERMINAL, **2**.

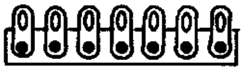

terminal strip (lug-type)

terminal voltage The voltage at the output terminals of an unloaded battery or generator.

ternary code See TRINARY NUMBER SYSTEM.

ternary fission The splitting of an atomic nucleus into three nuclear pieces. Also see FISSION.

ternary number system See TRINARY NUMBER SYSTEM.

terrain echoes Radar images caused by reflections from hills, mountains, and other natural terrestrial surface features. Also see GROUND CLUTTER.

terrestrial magnetism See EARTH'S MAGNETIC FIELD.

tertiary coil A third winding (see TERTIARY WINDING).

tertiary winding A third winding on a transformer or magnetic amplifier.

tesla Symbol, T. A unit of magnetic flux density; 1 T = 1 weber per square meter = 10^4 gauss.

Tesla coil A special type of air-core step-up transformer for developing high voltage at radio frequencies. It consists essentially of a low-turn primary coil, through which radio-frequency (RF) current flows, and a multiturn secondary coil, across which the high voltage, is developed.

test A procedure consisting of one or several steps, in which (1) the mode of operation of a circuit or device is established, (2) the value of a component is ascertained, or (3) the behavior of a circuit or device is observed.

test bench An equipment installation intended for the testing, repair, or debugging of electronic devices by a technician.

test data Data used to test a computer program, including samples within limits that might be encountered during the program's implementation.

tester **1.** See TEST INSTRUMENT. **2.** A technician who primarily makes tests and measurements.

testing window See WINDOW, **2**.

test instrument A device for checking the operation of a circuit or the value of a component. This class of instrument is usually less accurate than measurement instruments. Also see TEST SET.

test lead The flexible, insulated wire attached to a test prod.

test pattern A picture-and-line display on the screen of a television picture tube, used to check such features as aspect ratio, linearity, contrast, etc.

test point A terminal intended for connection of test equipment in the repair or debugging of a circuit. Often, test points are labeled by the letters TP followed by numerals (such as TP1, TP2, etc.).

test probe See PROBE, **1**.

test prod A stick-type probe (see PROBE, **1**) with a flexible, insulated lead terminating in a plug or lug for attachment to an instrument.

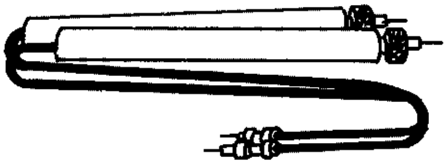

test prods

test program In computer operations, a program devised to check the functioning of hardware. Also called *utility* and *test routine.*

test run In computer operations, using test data to check the operation of a program, by comparing the results obtained thereby with what should ideally result.

test set A combination of instruments assembled on a single panel, and usually enclosed in a carrying case, for convenience in making tests.

test signal **1.** A signal used for conducting a test of a component, circuit, or system. **2.** In radiotelegraphy, a special signal signifying that the transmitting station is testing equipment. Also see VEE SIGNAL.

test-signal generator A device, such as an oscillator, for producing a signal for testing equipment (see TEST SIGNAL, **1**).

test tape A magnetic tape containing signals for testing equalization, frequency response, head adjustment, stereo balance, etc. in an audiotape recorder, high-fidelity sound system, computer tape drive, or videocassette recorder.

tetravalent See QUADRIVALENT.

tetrode An electron tube in which the principal electrodes are cathode, control grid, screen, and plate.

tetrode transistor **1.** A bipolar transistor with two emitters. **2.** A dual-gate field-effect transistor.

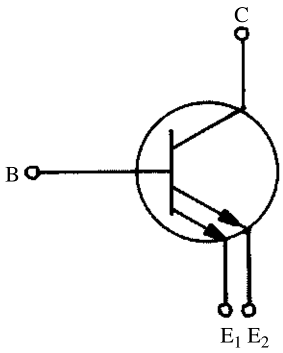

tetrode transistor

TeV Abbreviation of TERA-ELECTRONVOLT.

T_e value The temperature at which a centimeter cube of glass or ceramic exhibits 1 megohm of resistance.

TE wave See TRANSVERSE ELECTRIC MODE.

text editor A computer program for finding and changing data in a file.

texture map In computer graphics, the topographical qualities of the surface of a three-dimensional rendition. For example, the image of an orange would have a "bumpy" texture map; the image of a peach would have a "fuzzy" texture map.

texture sensing The ability of a robotic end effector to determine the relative smoothness or roughness of a surface. One common scheme uses reflected light or infrared. Scattered reflections indicate a rough or matte surface.

tgt Abbreviation of TARGET.

TH Abbreviation of *true heading.*

Th Symbol for THEORIUM.

TH Symbol for *heater temperature.*

thallium Symbol, Tl. A metallic element. Atomic number, 81. Atomic weight, 204.38.

thalofide cell An evacuated photoconductive cell using thallium oxysulfide as the light-sensitive material.

THD Abbreviation of TOTAL HARMONIC DISTORTION.

theory A reasonable proposition put forth to account for the behavior of, or the relationships between, bodies and forces, or to explain concepts and their relations. When a theory has stood up under exhaustive tests, it might reveal a scientific law.

therm A gas heating unit. 1 therm = 100,000 British thermal units (10^5 Btu).

thermal agitation Random movement of particles (such as electrons) in a substance, because of heat.

thermal-agitation noise See THERMAL NOISE.

thermal ammeter See HOT-WIRE AMMETER.

thermal anemometer See HOT-WIRE ANEMOMETER.

thermal conductivity The heat-conducting ability of a material. Compare electrical conductivity (see CONDUCTIVITY).

thermal conductivity

Element	*Thermal conductivity (mW/m/°C)*
Aluminum	22.0
Carbon	2.4
Chrominum	6.9
Cooper	39.0
Gold	30.0
Iron	7.9
Lead	3.5
Magnesium	16.0
Mercury	0.8
Nickel	8.9
Platinum	6.9
Silicon	8.4
Silver	41.0
Thorium	4.1
Tin	6.4
Tungsten	20.0
Zinc	11.0

thermal-conductivity device An instrument or control unit using a heated filament whose temperature and, accordingly, conductivity is varied by some sensed phenomenon. See, for example, GAS DETECTOR, HEATED-WIRE SENSOR, HOT-WIRE ANEMOMETER, HOT-WIRE FLOW-METER, and HOT-WIRE MICROPHONE.

thermal-conductivity gasmeter See GAS DETECTOR.

thermal detector **1.** See BOLOMETER. **2.** In a security or fire-protection system, a device that closes a circuit or actuates an alarm if the temperature rises to a specific level.

thermal emf See SEEBECK EFFECT.

thermal gasmeter See GAS DETECTOR.

thermal imaging See THERMOGRAPHY.

thermal instrument See HOT-WIRE METER and THERMOCOUPLE-TYPE METER.

thermally sensitive resistor See THERMISTOR.

thermal meter See HOT-WIRE METER and THERMOCOUPLE-TYPE METER.

thermal neutron A neutron that is essentially in thermal equilibrium with the surrounding medium or environment.

thermal noise Frequency-independent electrical noise caused by the agitation of particles (e.g., atoms and electrons) in a material by heat. Thermal noise is proportional to bandwidth, resistance, and absolute temperature.

thermal radiation See HEAT.

thermal recorder A graphic recorder in which a strip of paper coated with a thin layer of opaque wax is drawn between a knife-edge platen and a heated writing stylus that melts the wax beneath its point, exposing the underlying black paper as a fine line.

thermal resistance Symbol, RT. For a semiconductor device, the rate of change of junction temperature, with respect to power dissipation; $RT = \mathrm{d}T/\mathrm{d}P$, where RT is in degrees Celsius per milliwatt, T is the temperature in degrees Celsius, and P is the power in milliwatts.

thermal resistor A resistor that is sufficiently temperature-sensitive to be used as a heat sensor. Examples: thermistor and germanium diode.

thermal response time For a power-dissipating component, the elapsed time between the initial change in power dissipation and the moment at which the temperature has changed by a specified percentage (usually 90%) of the total value.

thermal runaway A destructive process resulting from cumulative temperature effects. In bipolar transistors, this can occur if the collector current increases as the temperature rises. As the unit gets hotter, the collector-base junction dissipation increases, generating still more heat. The ultimate result, if the process continues unchecked, is destruction of the component. The process can also occur in certain batteries when they are charged too rapidly or at excessively high temperatures.

thermal shock The effect of applying heat or cold to a device so rapidly that abnormal reactions occur, such as rapid (often catastrophic) expansions and contractions.

thermal switch A switch actuated by a temperature change. Types vary from the simple thermostat to complex servosystem switches with a temperature-transducer input.

thermal time-delay relay A delay relay utilizing the slow-heating and slow-cooling property of one of its components.

thermion An ion or electron emitted by a hot body, such as the heated cathode of a vacuum tube.

thermionic Pertaining to thermions and their applications.

thermionic cathode A heated cathode, in contrast to a cold cathode, used as an emitter of electrons or ions. Also see THERMION.

thermionic current Current caused by thermionic emission—especially in an electron tube.

thermionic detector A vacuum-tube detector. Also see THERMION and THERMIONIC EMISSION.

thermionic diode A hot-cathode diode tube.

thermionic emission The emission of electrons by a hot body, such as the filament or cathode of a vacuum tube. Also see THERMION and HOT CATHODE.

thermionics The study of electron emission from objects or materials at high temperature.

thermionic tube An electron tube (i.e., one in which electrons or ions are emitted by a heated cathode). Also see THERMION, THERMIONIC CURRENT, and THERMIONIC EMISSION.

thermionic work function The energy required to force an electron from inside a heated cathode into the surrounding space (in thermionic emission). Also see WORK FUNCTION.

thermistor A temperature-sensitive resistor, usually made from specially processed oxides of cobalt, magnesium, manganese, nickel, uranium, or mixtures of such substances. Thermistors are available with either a positive or negative temperature coefficient of resistance. The name is a contraction of *thermally sensitive resistor.*

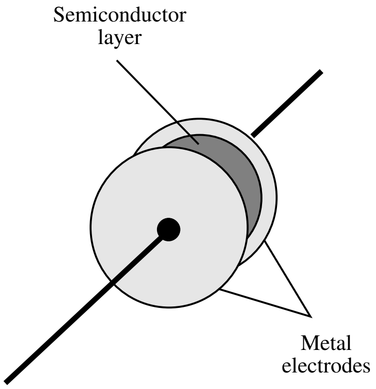

thermistor

thermistor bridge A four-arm bridge (see BRIDGE, **1**) in which one arm is a thermistor and, therefore, is temperature sensitive.

thermistor power meter A radio-frequency power-measuring instrument based on a thermistor bridge.

thermistor probe A temperature probe containing a thermistor as the sensing element.

thermistor thermometer An electronic thermometer in which the temperature-sensitive element is a thermistor.

thermoammeter See THERMOCOUPLE-TYPE METER.

thermocouple A device consisting essentially of a bond between two wires or strips of dissimilar metals (such as antimony and bismuth). When the bond is heated, a direct-current voltage appears across it.

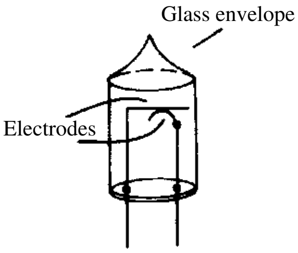

thermocouple

thermocouple bridge A four-arm bridge (see BRIDGE, **1**) in which one arm is a thermocouple; therefore, it is temperature sensitive, as well as being voltage-productive.

thermocouple meter See THERMOCOUPLE-TYPE METER.

thermocouple-type meter A radio-frequency (RF) meter consisting of a thermocouple and direct-current (dc) ammeter, milliammeter, or microammeter, connected in series. The thermocouple is heated, directly or indirectly, by an applied RF current, and the resulting dc output deflects the meter pointer.

thermodynamics The science dealing with the relationships between heat and mechanical energy and their interconversion. Also see CARNOT THEOREM, FIRST LAW OF THERMODYNAMICS, SECOND LAW OF THERMODYNAMICS, and THIRD LAW OF THERMODYNAMICS.

thermoelectric cooler A cooling device based on the Peltier effect.

thermoelectric effect The production of thermoelectricity by certain materials.

thermoelectricity Heat-produced electricity, as in thermocouple operation.

thermoelectric junction A junction between two conductors that exhibits variable characteristics under conditions of changing temperature.

thermoelectric series A series of conducting metals, sequenced so that if a thermocouple is made from two of the metals, current flows at the hot junction from the metal occurring earlier in the series.

thermoelectron An electron emitted by a hot cathode. Also see THERMION and THERMIONIC EMISSION.

thermoelement A thermocouple, especially a miniature one, used in a THERMOCOUPLE-TYPE METER.

thermogalvanometer See THERMOCOUPLE TYPE METER.

thermography A means of locating or mapping by detecting infrared images. It can be in the form of live video, similar to television, or in the form of photographs, using infrared film. Many common cameras can be used with infrared film to make infrared photographs (thermographs).

thermojunction The junction of the two metals in a thermocouple. A voltage appears when the dissimilar metals are heated.

thermoluminescence Luminescence resulting from the moderate heating of certain materials.

thermomagnetic effect The effects of temperature on the magnetism of a body, or vice versa.

thermometer A (usually direct reading) device for measuring temperature. Also see ELECTRONIC THERMOMETER and THERMOMETER SCALE.

thermometer scale The scale on a thermometer, graduated in degrees, from which temperature is read. For a description of different scales, see ABSOLUTE SCALE, CENTIGRADE SCALE, FAHRENHEIT SCALE, REAUMUR SCALE, and RANKINE SCALE. The Kelvin scale is the same as the absolute scale, and the Celsius scale is the same as the centigrade scale.

thermonuclear reaction A nuclear reaction in which energy is released when lighter atoms are converted into heavier atoms at temperatures in the millions of degrees Celsius. Also see FISSION, FUSION, and NUCLEAR REACTOR.

thermopile A device consisting of two or more thermocouples connected in series for increased voltage output.

thermoplastic material A plastic that can be resoftened by applying heat after having been molded into a desired shape. Example: polystyrene. Compare THERMOSETTING MATERIAL.

thermorelay See THERMOSTAT.

thermosensitivity Sensitivity of a circuit or device to heat.

thermosetting material A plastic that cures chemically (will not ordinarily soften again when heat is applied) after having been heat-molded into a desired shape. Example: Bakelite. Compare THERMOPLASTIC MATERIAL.

thermostat A temperature-sensitive switch. In one common form, a movable contact is carried by a strip of bimetal and the stationary contact is mounted nearby.

thermostatic switch See THERMOSTAT.

thermoswitch See THERMOSTAT.

theta wave A form of brain wave that occurs at extremely low frequencies, and is associated with sleep or mental incoherence.

Thevenin's theorem The proposition that, with reference to a particular set of terminals, a network containing a number of generators and constant impedances can be simplified to a single generator in series with a single impedance. The equivalent circuit will deliver to a given load, the same current, voltage, and power delivered by the original network. Compare COMPENSATION THEOREM, MAXIMUM POWER TRANSFER THEOREM, NORTON'S THEOREM, RECIPROCITY THEOREM, and SUPERPOSITION THEOREM.

thick film A film of selected material (conductive, resistive, dielectric, etc.) applied to a substrate by painting, photography, or similar process. See, for example, PRINTED CIRCUIT. Compare THIN FILM. Typically, thick films are 1 mil or more in thickness.

thick-film component A unit, such as a capacitor or resistor, fabricated by thick-film techniques. See, for example, PRINTED COMPONENT. Also see THICK FILM. Compare THIN-FILM COMPONENT.

thick-film resistor A resistor fabricated by thick-film techniques. See, for example, PRINTED RESISTOR. Also see THICK FILM. Compare THIN-FILM COMPONENT.

thick magnetic film See MAGNETIC THICK FILM.

thin film An extremely thin layer (less than 1 mil) of a selected material (conductive, resistive, semiconductive, dielectric, etc.) electrodeposited or grown on a substrate. Compare THICK FILM.

thin-film capacitor A capacitor made by electrodepositing a thin film of metal on each side of a grown layer of oxide, as in an integrated circuit. Also see THIN FILM.

thin-film component A unit (such as a capacitor, resistor, diode, or transistor), fabricated by thin-film techniques. Also see THIN FILM. Compare THICK-FILM COMPONENT.

thin-film integrated circuit An integrated circuit in which the components and "wiring" are produced by depositing (or growing) and processing materials on a semiconductor slab or wafer (the substrate). Compare HYBRID INTEGRATED CIRCUIT and MONOLITHIC INTEGRATED CIRCUIT.

thin-film memory In a computer, a storage medium that is a magnetic thin film (see THIN FILM) on a nonmagnetic substrate (often glass) and that can be magnetized to represent digital data.

thin-film microelectronic circuit An integrated circuit that occupies (essentially) two dimensions; that is, a very thin integrated circuit.

thin-film resistor A resistor fabricated by thin-film techniques (see THIN FILM) (e.g., TANTALUM-NITRIDE RESISTOR). Compare THICK-FILM RESISTOR.

thin-film semiconductor A very shallow layer of semiconductor material, such as single-crystal silicon, electrodeposited on a substrate. Also see THIN FILM.

thin-film transistor A transistor fabricated by thin-film techniques. Also see THIN FILM and THIN-FILM COMPONENT.

thin magnetic film See MAGNETIC THIN FILM.

third-generation computer A computer in which the active discrete components are integrated circuits.

Third Law of Thermodynamics As the temperature of absolute zero is approached in an isothermal process involving a solid or liquid, the change in entropy approaches zero; and the entropy of a substance is zero at absolute zero.

third-octave band See ONE-THIRD-OCTAVE BAND.

thirty-channel multiplex A form of pulse-code-modulated (PCM) multiplex using eight digits and the A-law. There are 30 speech channels and two utility channels.

Thomson bridge See KELVIN BRIDGE.

Thomson effect The liberation or absorption of heat (depending on the material of interest) when an electric current flows from a warmer to a cooler part of a conductor.

Thomson heat The amount of thermal energy transferred because of Thomson effect.

Thomson voltage The voltage drop between two points on a conductor that are at different temperatures.

thoriated-tungsten filament In a vacuum tube, a filament made of tungsten to which thorium oxide has been added to increase the emission of electrons. Also see THERMION and THERMIONIC EMISSION.

thorium Symbol, Th. A radioactive metallic element. Atomic number, 90. Atomic weight, 232.04. Thorium, when heated, is a copious emitter of electrons, so the filaments or cathode cylinders of some electron tubes are coated with one of its compounds.

thoron Symbol, Tn. A radioactive isotope of RADON.

three-address instruction A computer program instruction having three addresses, two for operands and one for the result (of the operation called for).

three-channel stereo A form of stereophonic sound recording and reproduction in which three distinct channels are used; these are usually designated the left, right, and center channels.

three-coil meter See ELECTRODYNAMOMETER.

three-conductor jack A female connector in which two separate conductors are provided, in addition to the ground conductor.

three-conductor plug A male connector in which two separate conductors are provided, in addition to the ground conductor.

three-dimensional television See STEREOSCOPIC TELEVISION.

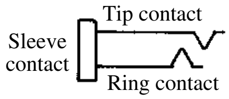

three-conductor jack

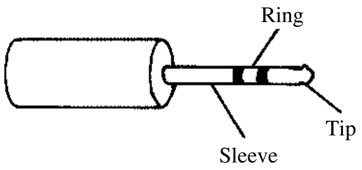

three-conductor plug

three-electrode tube See TRIODE.

three-element tube See TRIODE.

three-gun picture tube A color-television picture tube having a separate gun for each primary color (red, green, and blue).

three-halves-power law See CHILD'S LAW.

three-junction transistor A pnpn or npnp transistor. Also see NPNP DEVICE.

three-phase bridge rectifier A bridge-rectifier circuit for three-phase alternating current. Two diodes are provided for each phase. The ripple frequency is six times the line frequency. Also see BRIDGE RECTIFIER, POLYPHASE RECTIFIER, and THREE-PHASE CIRCUIT.

three-phase circuit The circuit of a three-phase system. See THREE-PHASE SYSTEM and, specifically, DELTA CONNECTION and WYE CONNECTION.

three-phase current Current in a three-phase circuit. The currents in the three legs differ in phase by 120°.

three-phase four-wire system See FOUR-WIRE WYE SYSTEM.

three-phase generator A (usually dynamo-type) generator of three-phase current or voltage. See THREE-PHASE SYSTEM.

three-phase half-wave rectifier A half-wave rectifier circuit for three-phase alternating current. One diode is provided for each phase. The ripple frequency is three times the line frequency. Also see HALF-WAVE RECTIFIER, POLYPHASE RECTIFIER, and THREE-PHASE CIRCUIT. Compare THREE-PHASE BRIDGE RECTIFIER.

three-phase motor An alternating-current motor operating on three-phase power. Above fractional-horsepower size, the three-phase motor is smoother running and more simply structured than the single-phase counterpart.

three-phase power The total power dissipated or delivered in a three-phase alternating-current circuit.

three-phase rectifier A rectifier for three-phase alternating current. At least one diode is included for each phase. Also see POLYPHASE RECTIFIER, THREE-PHASE BRIDGE RECTIFIER, THREE-PHASE CIRCUIT, and THREE-PHASE HALF-WAVE RECTIFIER.

three-phase system An alternating-current electrical system in which three currents or voltages exist simultaneously. They are of equal amplitude, but are 120° out of phase with each other.

three-phase, three-wire system An electrical system having three conductors, with an alternating-current phase difference of 120° between conductor pairs.

three-phase-to-single-phase transformer An alternating-current transformer with three-phase input and single-phase output (one terminal is grounded).

three-phase-to-two-phase transformer An alternating-current transformer with three-phase input and two-phase output, the output currents are 180° out of phase with each other.

three-phase voltage Voltage in a three-phase alternating-current circuit. The voltages across the three legs differ by 120°.

three-quarter bridge A bridge rectifier having diodes in three arms and a resistor in the fourth.

three-space A mathematical continuum in which each point is uniquely defined by three variables in an ordered triple, such as (x,y,z), and each ordered triple corresponds to exactly one point in the space. A common coordinate system is the Cartesian system consisting of three axes, x, y, and z—each mutually perpendicular and intersecting at the origin (0,0,0).

three-space coordinates Any set of coordinates used for locating points or plotting graphs in three dimensions.

three-state logic See TRI-STATE LOGIC.

three-way speaker A set of three individual speakers contained in a single cabinet: a woofer (low-frequency speaker), a midrange speaker, and a tweeter (high-frequency speaker). It is common in high-fidelity music reproduction systems. Compare TRIAXIAL SPEAKER.

three-wire system **1.** An electric-power feed system using three wires, the center one (neutral) being at a potential midway between the potential across the other (outer) two. **2.** See THREE-PHASE and THREE-WIRE SYSTEM. **3.** See TWO-PHASE and THREE-WIRE CIRCUIT.

threshold **1.** The initial (observable) point of an effect (e.g., *threshold of hearing*). **2.** A predetermined point, such as of minimum current or voltage, for the start of operation of a circuit or device.

threshold component A value of current, voltage, sound intensity, etc., selected as the minimum level at which a circuit or device is to operate in some prescribed manner, or beyond which a certain condition will prevail. Also see THRESHOLD, **1**, **2**.

threshold current **1.** The minimum value of current at which a certain effect takes place. **2.** The smallest amount of forward current that flows through a diode. **3.** In a gas, the smallest level of current for which a discharge will maintain itself under variable conditions.

threshold detector A device that prevents a signal from passing until its peak amplitude reaches a certain value.

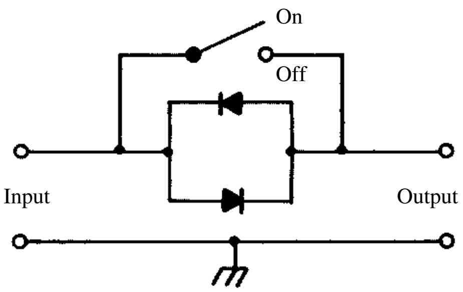

threshold detector

threshold frequency **1.** The cutoff frequency or frequencies in a band-pass, band-rejection, low-pass, or high-pass filter. **2.** The lowest frequency for a metal or semiconductor junction at which incident radiation gives rise to the photoelectric effect.

threshold of hearing The minimum intensity level at which sounds are audible in an environment containing essentially no background acoustic noise.

threshold of pain The intensity level at which hearing a sound causes physical discomfort. This is approximately 120 dB above the THRESHOLD OF HEARING for most people.

threshold signal The weakest signal that can be detected in a receiving system.

throat See HORN THROAT.

throat microphone A small microphone operated in contact with the user's throat.

throttle The feedback control device in a regenerative detector or amplifier.

throughput In computer operations, an overall quantitative indicator of processing power, expressed in terms of the amount of data processed in a given period of time.

throw-out spiral On a phonograph disc, a lead-out groove.

thulium Symbol, Tm. A metallic element of the rare-earth group. Atomic number, 69. Atomic weight, 168.93.

thumbwheel potentiometer A potentiometer operated by means of a knurled knob (usually protruding perpendicularly through a panel) that is turned with the thumb or with a finger.

thumbwheel switch A switch operated by means of a knurled knob (usually protruding perpendicularly through a panel) that is turned with the thumb or with a finger.

thy Abbreviation of THYRATRON.

thyratron A gas triode or gas tetrode used principally for switching and control purposes. Thyratron action differs from that of the vacuum tube in the unique behavior of the thyratron control electrode (grid): Anode current starts to flow abruptly when grid voltage reaches a particular value, at which point the grid provides no further control; anode current continues to flow until the anode voltage is either interrupted or reversed.

thyratron inverter An inverter circuit (see INVERTER, **1**) using thyratrons as the switching devices. Also see THYRATRON.

Thyrector A silicon diode exhibiting very high resistance (approaching that of an insulator) up to a certain voltage, beyond which the unit switches to a low-resistance conducting state.

thyristor **1.** A pnpn-type bistable semiconductor device having anode, cathode, and gate terminals that is used as an electronic switch. **2.** The generic term for all thyratronlike solid-state devices, such as the silicon-controlled rectifier.

Thyrite Ceramic silicon carbide, a nonlinear resistance material, or a resistor made of this material. The resistance of Thyrite decreases sharply as the applied voltage is increased. Thyrite resistors are used in voltage regulators, equipment protectors, lightning arresters, curve changers, and similar devices.

THz Abbreviation of TERAHERTZ.

Ti Symbol for TITANIUM.

tickler A (usually small) coil, through which energy is inductively fed back from the output to the input of a circuit to induce oscillation.

tickler coil See TICKLER.

tickler-coil regeneration Positive feedback obtained via inductive coupling between a small coil (tickler) in the output circuit of an amplifier, and a (usually larger) coil in the input circuit. Also see TICKLER.

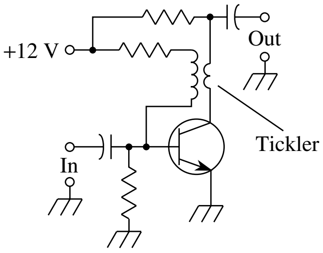

tickler-coil regeneration

tickler oscillator An oscillator circuit in which positive feedback is obtained through inductive coupling between an output (tickler) coil and an input coil.

tie A bracket, clamp, clip, ring, or strip for holding several wires tightly as a cable or bundle.

tie cable **1.** A cable that connects two distributing points in a telephone system. **2.** Any cable that interconnects two circuits.

tie point A lug, screw, or other terminal to which wires are connected at a junction.

tie-point strip A terminal strip with lugs to which conductors can be soldered.

tight coupling See CLOSE COUPLING.

tilt switch A device, such as a mercury switch, that is actuated by tilting it to a certain angle.

timbre The quality that distinguishes the sound of one voice or instrument from that of another, largely because of harmonic content.

time Symbol, *t*. **1.** The instant at which an event occurs. **2.** The instant at which a time-base variable reaches a given value. **3.** The interval between two instants, commonly called *duration* or *length of time*. Also see STANDARD TIME, TIME BASE, TIME ZONE, and ZERO TIME.

time base Time as the independent variable in a physical relation or function. It appears in expressions such as *pulses per second*, *feet per minute*, *watts per hour*, etc.

time compressor In audio recording and reproduction, a device that speeds up or slows down the tempo without changing the audio frequencies. It is used for special effects.

time constant See ELECTRICAL TIME CONSTANT and MECHANICAL TIME CONSTANT.

time delay See DELAY, **1**, **2**.

time-delay relay See DELAY RELAY.

time-delay spectrometry In acoustics, a method of simulating an echo-free environment within an enclosure that is actually not echo-free.

time-division multiplexing Abbreviation, TDM. In data and computer communications, a time-sharing technique in which several terminals use the same channel by transmitting data at regular, staggered intervals (i.e., one is active while the others are idle). This gives the appearance of simultaneous real-time operation.

time-division-multiplex switch A switch with multiple ports—each port corresponding to a certain time slot in a time-division-multiplex scheme. The input and output (send and receive) ports are connected by selecting the same TIME SLOT for each.

time-domain reflectometry Abbreviation, TDR. Measuring the reflective characteristics of a device or system by superimposing the direct and reflected components of a step-formed test signal on a time-calibrated oscilloscope screen.

time duration See TIME, **3**.

time-duration modulation See PULSE-DURATION MODULATION.

time factor The ratio *ta*/*tr*, where *ta* is analog time (the relativistic duration of an event as simulated by a computer), and *tr* is real time (the actual duration of the event). Also called TIME SCALE.

time-interval mode In computer operations, operation that allows a number of events to be counted between two points on a waveform.

time modulation Any form of modulation in which the instantaneous characteristics of a signal are varied.

timeout The expiration of an allotted time period for a given operation.

timer **1.** A device for automatically controlling the duration of an operation. See ELECTRONIC TIMER. **2.** A device for measuring the duration of an operation.

time sharing In computer operations, a method for interlaced (i.e., nearly simultaneous) use of a machine or facility by two or more persons or agencies. As the cost of computers diminishes, so does the need for time sharing. Also see TIME-DIVISION MULTIPLEXING.

time-sharing dynamic allocator In computer storage, a program that allocates memory areas and peripherals to programs being entered into a time-sharing system; it also controls program execution.

timeshift **1.** To receive a message at a significantly later time than when it was sent. **2.** To use a videocassette or videodisc recorder to view a television program at a later time than when it was transmitted.

time-shifting communications Any form of communications in which the recipient reads or views the message(s) at a significantly later time than the message is sent from the source. A common example is electronic mail (e-mail) using online computer networks.

time signals Special radio transmissions made under the auspices of the National Bureau of Standards, for indicating Coordinated Universal Time (UTC).

time slot A specifically defined time interval in a data signal. It is of importance primarily in digital communications, where a given time interval can be high or low or at some discrete value.

time-space-time switch A large switch consisting of a space block between two time blocks.

time zone One of the 24 zones into which the global map is divided for the purpose of standardizing time. Within these zones, mean solar time is determined in terms of distance east or west of the zero meridian at Greenwich (near London, England). Each zone is equal to 15 degrees of longitude, or 1 hour. Four zones fall within the continental United States: Eastern Standard Time (zone of the 75th meridian), Central Standard Time (zone of the 90th meridian), Mountain Standard Time (zone of the 105th meridian), and Pacific Standard Time (zone of the 120th meridian). Also see MERIDIAN, ZERO MERIDIAN, and ZONE TIME.

timing extraction The retrieval of a timing signal from incoming data.

timing pin current Measured in microamperes. The current that generates a timing waveform in an integrated-circuit voltage regulator.

timing signal A repeating signal sent along with data to control the synchronization of transmitter and receiver.

tin **1.** Symbol, Sn. A metallic element. Atomic number, 50. Atomic weight, 118.71. Tin is widely used in electronics as a structural material, a constituent of solder, and (in foil form) as the plates of some fixed capacitors. **2.** To prepare the tip of a soldering gun or iron, or the stripped end of a wire or cable, by applying a coat of solder.

tin-oxide resistor See METAL-OXIDE RESISTOR.

tinsel Metal film strips interwoven with fabric threads to provide a flexible cord, particularly for headphones.

tint control In a color-television receiver, the control for changing color hue.

tip jack The mating connector for a tip plug.

tip plug **1.** A prod terminating in a phone tip. **2.** A plug-type connector terminating in a phone tip.

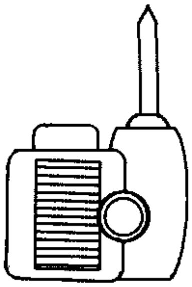

tip plug, 2

titanium Symbol, Ti. A metallic element. Atomic number, 22. Atomic weight, 47.88. Titanium enters into some dielectric compounds (e.g., titanium dioxide).

titanium dioxide Formula, TiO_2. A ceramic dielectric material. Dielectric constant, 90 to 170. Dielectric strength, 100 to 210 V/mil.

T junction See TEE JUNCTION.

Tl Symbol for THALLIUM.

T^2L Abbreviation of TRANSISTOR-TRANSISTOR LOGIC. (Also, TTL.)

TLA Abbreviation of TABLE LOOK-AT.

TLC Abbreviation of *thin-layer chromatography.*

TLU Abbreviation of TABLE LOOK-UP.

TM **1.** Abbreviation of TRANSVERSE MAGNETIC. **2.** Abbreviation of *technical manual.*

Tm Symbol for THULIUM.

TM mode See TRANSVERSE MAGNETIC MODE.

TM wave See TRANSVERSE MAGNETIC MODE.

T network See TEE NETWORK.

TNS Abbreviation of TRANSCUTANEOUS NERVE STIMULATOR.

toggle A bistable device.

toggle switch A switch having a mechanism that snaps into the on or off position at the opposite extremes to which a lever is moved.

tolerance The amount by which error is allowed in a value, rating, dimension, etc. It is usually

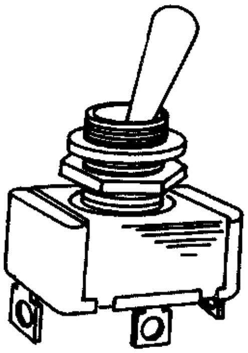

toggle switch

expressed as a percent of nominal value, plus and minus so many units of measurement (or parts per million).

toll call In a telephone system, a call that is charged on a per-minute or per-second basis.

ton Abbreviation, T, t, or tn. A unit of avoirdupois weight; in the United States, it is usually taken to mean *short ton,* a unit equal to 907.20 kilograms (2000 lb). Compare METRIC TON.

tone **1.** The pitch (frequency) and timbre (relative harmonic content) of a sound other than noise. **2.** A sound consisting of a periodic waveform having constant frequency; also called *note.*

tone arm See PICKUP ARM.

tone burst A test signal consisting of a single-frequency sine wave sustained for a brief period of time, usually having a rectangular envelope (rapid rise and decay).

tone-burst entry In repeater systems, a technique whereby a short tone signal is used at the start of a transmission to trigger a particular repeater so that all repeaters in the system will not go into operation simultaneously.

tone-burst generator An oscillator and associated circuitry for producing a tone burst.

tone control An adjustable device or circuit for modifying the frequency response of an amplifier (i.e., for emphasizing bass, treble, or midrange pitches).

tone dialing A method of telephone dialing that uses standard tone pairs actuated via a keypad with 12 keys representing digits 0 through 9 and symbols # and *. Some keypads have four additional keys: A, B, C, and D.

tone generator An oscillator for producing simple audio-frequency signals for communications, control, or testing.

tone keying In wire and radio telegraphy, the representation of code characters by audio-frequency tones. Also see MODULATED CONTINUOUS WAVE.

tone localizer A localizer that provides lateral guidance for an aircraft by comparing the amplitudes of two modulating frequencies.

tone modulation **1.** The transmission of Morse, Baudot, or ASCII signals by audio-frequency modulation of a radio-frequency carrier. **2.** Any rapid variation of the amplitude or frequency of an audio tone.

top cap A small metal cap on top of some electron tubes, serving as a direct, low-capacitance connection to one of the internal elements—usually the control grid, but sometimes the plate.

top loading A method of feeding a vertical antenna at or near the top.

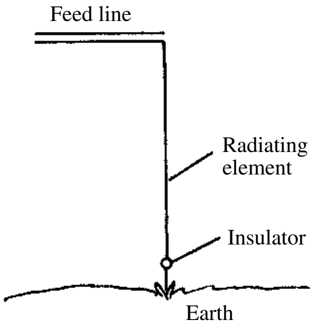

top loading

topology **1.** A branch of mathematics concerned with the properties of surfaces and spaces. **2.** The details of layout of an integrated circuit. **3.** The characteristics of a surface.

toroid See TOROIDAL COIL.

toroidal coil A coil wound on a form that is shaped like a donut. The form is made of powdered-iron or ferrite. Toroidal coils have certain advantages over solenoidal coils: greater inductance for a given physical size, better isolation properties, and higher Q factor. A disadvantage is that an air core is not practical.

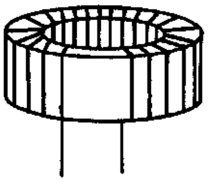

toroidal coil

torque **1.** The force that tends to produce a rotating motion. **2.** Rotation of the plane of polarization of light by some crystals.

torque amplifier A device having rotating input and output shafts and that delivers greater torque at the output shaft than that required to turn the input shaft.

torque sensitivity Symbol, *KT*. For a torque motor, torque output per ampere of input current.

torr A unit of pressure equal to the pressure required to support a column of mercury 1 millimeter high at 0°C and standard gravity; 1 torr = 133.322 Pa (1333.2 microbars).

torsion The effect on an object by torque applied to one end while the other is being held fast or torqued in the opposite direction.

torsion delay line A delay line in which the delay is manifested in a material that is torqued by mechanical vibrations.

torsion waves Waves that travel by means of torque, instead of displacement or compression. The velocity of propagation depends on the modulus of rigidity and the density of the propagating medium.

tot **1.** Abbreviation of *total.* **2.** To derive a sum or total.

total breakaway torque For a torque motor, the sum of magnetic retarding torque and brush-commutator friction.

total harmonic distortion Abbreviation, THD. The distortion caused by the combined action of all the harmonics present in a complex waveform. An important specification in high-fidelity audio amplifiers.

total distortion See TOTAL HARMONIC DISTORTION.

total internal reflection **1.** The reflection of visible light from a boundary between two substances having different indices of refraction. When the angle of incidence, relative to the tangent to the boundary, is smaller than a certain value, as light travels through the more dense medium total reflection occurs at the boundary. **2.** Reflection of electromagnetic waves from an ionized layer in the atmosphere. This occurs at angles smaller than a certain angle, relative to the tangent of the plane of the ionizing layer. In some cases, total internal reflection does not occur. Actually, most ionospheric reflection is really refraction; the electromagnetic waves are bent by the ionized layer rather than reflected.

total reflection Full return of a ray by a reflector, none of the energy being transmitted by or absorbed in the reflecting material.

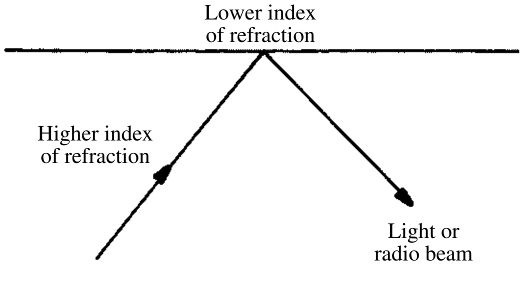

total internal reflection

touchplate relay A capacitance relay in which the pickup element is a small, metal plate that actuates the relay when it is touched.

tough dog A malfunctioning circuit that seemingly defies all attempts to diagnose and correct its trouble. Also see DOG.

tourmaline crystal A piezoelectric and pyroelectric crystal; a transparent plate of tourmaline has the ability to polarize light passing through it. Also see POLARIZATION, **3**; POLARIZED LIGHT; and POLARIZER.

tower **1.** A usually tall and self-supporting openwork structure used to support an antenna, and usually having three or four sides. **2.** A metal structure, as defined in **1**, used as a vertical antenna.

Townsend discharge In a glow-discharge tube, the discharge that begins after the applied voltage reaches a given level. It is a low-current, non-self-sustaining discharge. Compare ABNORMAL GLOW and NORMAL GLOW DISCHARGE.

touch screen A special cathode-ray tube (CRT) or video display unit that allows input of data via physical contact with the screen surface. Items are selected by simply touching the appropriate spot (“button” or icon) on the screen.

TP Abbreviation of TRANSACTION PROCESSING.

T pad See TEE PAD.

TR Abbreviation of transmit-receive.

tr **1.** Symbol for RECOVERY TIME. **2.** Symbol for RISE TIME.

trace **1.** A tiny or insignificant quantity. **2.** The movement of the electron beam across the face of a cathode-ray tube. **3.** A routine used for testing of, or for locating a fault in, a circuit or computer program. **4.** The process of implementing such a routine.

trace element See MICROELEMENT.

tracer A suitable substance or object introduced into the human body and whose progress through the body can be followed (or its state monitored) by means of electronic equipment. Tracers are sometimes used also in nonbiological systems, such as pipelines.

track **1.** A discrete information band on a magnetic disk or tape. **2.** To follow, as by a stylus, a phono disc groove. **3.** To follow, as by radar, a target.

trackability An expression of the accuracy with which a phonograph stylus follows the irregularities in a disc.

trackball A device for guiding the cursor or pointer in a computer. It is often used in laptop (notebook) computers. It resembles a ball bearing; the operator moves the cursor or pointer by pushing on the bearing with a finger.

track-drive locomotion A method of robotic locomotion using two or more wheels that drive a belt (track). It uses the same principle as a military tank.

tracking Following in step, as when ganged circuits resonate at the same frequency (or some frequency difference) at all settings, or when a missile closely follows its guiding signal.

tracking force See VERTICAL STYLUS FORCE.

track label On a magnetic storage medium, a record that identifies a track.

tracking mode In tracking supplies, the usual manner of operation in which the output of each of the separate supplies automatically follows that of the one being adjusted. Compare INDEPENDENT MODE.

tracking supplies Adjustable power supplies, packaged two or more to the unit, in which the output of each will automatically follow adjustment made to one of them.

track pitch The distance between tracks (see TRACK, **2**).

traffic **1.** Collectively, messages handled by a communications station. **2.** Collectively, data and instructions handled by a computer system in continual use.

trailer label At the end of a magnetic-tape or floppy-disk file, a record signaling the end of the file and often giving such information as the number of records in the file.

trailer record At the end of a group of computer records, a record containing information relevant to the group's processing.

trailing edge The falling edge of a pulse. Compare LEADING EDGE.

trans Abbreviation of TRANSVERSE.

transaction The exchange of activity that occurs between a computer, via a terminal, and the user, including any processing required (e.g., that involved in adding records to, or deleting them from, a file).

transaction file In data processing, a group of records used to update a master file.

transaction processing In computer operations, the use of a central processor for handling, modifying, or otherwise acting on information by transactions.

transaction tape Magnetic tape on which a transaction file has been recorded.

transadmittance For an active device, the ratio $dI2/dE1$, where $I2$ is the alternating-current (ac) component of the current in a second electrode (such as the drain), and $E1$ is the ac component of voltage on a first electrode (such as the gate), with constant direct-current operating voltages.

transceiver **1.** A combination transmitter and receiver, housed in a single enclosure, with frequency control and some ancillary stages common to both units. This design is economical because it eliminates redundancy. Such a system is more easily tuned than a separate transmitter and receiver if the operating frequency must be changed often, but it can be difficult to carry out split-frequency or split-band communication on two frequencies that differ greatly. Compare TRANSMITTER–RECEIVER. **2.** In computer practice, a read/write data terminal capable of transmitting and receiving information to and from a channel.

transcendental functions Nonalgebraic functions. These include logarithmic functions, exponential functions, trigonometric functions, and inverse trigonometric functions.

transconductance **1.** Symbol, gm. Unit, micromho. In an electron tube, the extent to which plate current (IP) changes in response to a change in grid voltage (EG); $gm = dIP/dEG$. **2.** Symbol, gfs. Unit, micromho. In a field-effect transistor, the extent to which drain current (ID) changes in response to a change in gate-to-source voltage (VGS); $gfs = dID/dVGS$.

transconductance amplifier An amplifier in which the output current is a linear function of the input voltage.

transcribe **1.** To record material, such as a radio program, for future transmission. **2.** In computer operations, the intermedia transfer of data, as from tape to disk.

transcriber A device used for intermedia data transfer (e.g., one that can move the data on magnetic tape to magnetic or optical disc).

transcription A recording of material, such as a record or tape of a radio program, for later use in a transmitter. See ELECTRICAL TRANSCRIPTION.

transcutaneous nerve stimulator Abbreviation, TNS. An electronic device for the temporary relief of pain. In its use, electrodes taped to the skin over the painful area are connected to a portable generator of suitable pulse energy.

transducer A device that converts one quantity into another quantity, specifically when one of the quantities is electrical. Thus, a loudspeaker converts electrical impulses into sound, a microphone converts sound into electrical impulses, a photocell converts light into electricity, a thermocouple converts heat into electricity, etc.

transducer amplifier An amplifier used expressly to boost the output of a transducer.

transducer efficiency For a transducer, the ratio of the output power to the input power.

transductor See MAGNETIC AMPLIFIER.

transfer **1.** To move a signal from one point to another—especially through a modifying circuit or device. **2.** To transmit information or data from one point or device to another, inside or outside a system.

transfer characteristic A figure or plot expressing the output-input signal relationship in a circuit or device. Also see TRANSFER, **1**.

transfer efficiency In a charge-coupled device, the proportion of charge that is transferred under given conditions.

transfer function **1.** See TRANSFER CHARACTERISTIC. **2.** An expression that mathematically shows how two entities or events occurring in different places or at different times are related.

transfer rate The speed at which data can be moved between a computer's internal memory and a peripheral.

transferred charge In a circuit containing a capacitor, the net electric charge that moves around the external circuit from one plate of the capacitor to the other.

transform **1.** See LAPLACE TRANSFORM. **2.** To change the voltage or nature of an electrical parameter (e.g., high to low voltage and alternating to direct current). **3.** To change the form, but not the content, of data.

transformation constant See DISINTEGRATION CONSTANT.

transformer **1.** A device using electromagnetic induction to transfer electrical energy from one circuit to another (i.e., without direct connection between them). In its simplest form, a transformer consists of separate primary and secondary coils wound on a common core of ferromagnetic material, such as iron. When an alternating current flows through the primary coil, the resulting magnetic flux in the core induces an alternating voltage across the secondary coil; the induced voltage can cause a current to flow in an external circuit. Also see AIR-CORE TRANSFORMER, INDUCTION, INDUCTIVE COUPLING, IRON-CORE TRANSFORMER, and TURNS RATIO. **2.** A section of radio-frequency (RF) transmission line used to match impedances. Also see LINEAR TRANSFORMER.

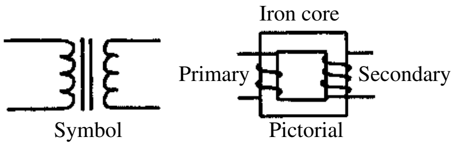

transformer, 1

transformer-coupled amplifier An amplifier using coupling transformers between its stages or at its input and output points. Compare RESISTANCE-CAPACITANCE-COUPLED AMPLIFIER.

transformer coupling The inductive coupling of circuits through a transformer. Also see COEFFICIENT OF COUPLING; INDUCTIVE COUPLING; MUTUAL INDUCTANCE; and TRANSFORMER, **1**.

transformer equivalent circuit An equivalent circuit depicting the various parameters of a transformer (such as primary and secondary resistances, primary and secondary reactances, core losses, etc.) and their relationship to each other.

transformer feedback Inductively coupled feedback (positive or negative) through a transformer.

transformer input current See PRIMARY CURRENT.

transformer input voltage See PRIMARY VOLTAGE.

transformer iron See SILICON STEEL.

transformer loss The expression (in decibels) 10 $\log_{10}(P1/P2)$, where $P1$ is the calculated power that a given transformer should deliver to a particular load impedance, and $P2$ is the actual power delivered.

transformer oil A petroleum product in which high-voltage, high-current transformers are sometimes immersed. The oil protects the windings from environmental damage.

transformer output current See SECONDARY CURRENT.

transformer output voltage See SECONDARY VOLTAGE.

transformer-type voltage regulator See VOLTAGE-REGULATING TRANSFORMER.

transformer utilization factor See UTILITY FACTOR.

transient **1.** A sudden, high-voltage spike in an alternating-current system, caused by arcing or lightning. **2.** A spurious signal in a hardcopy-receiving system. **3.** Any short pulse attributable to external causes. **4.** Existing for a short period of time, or intermittently for short periods of time.

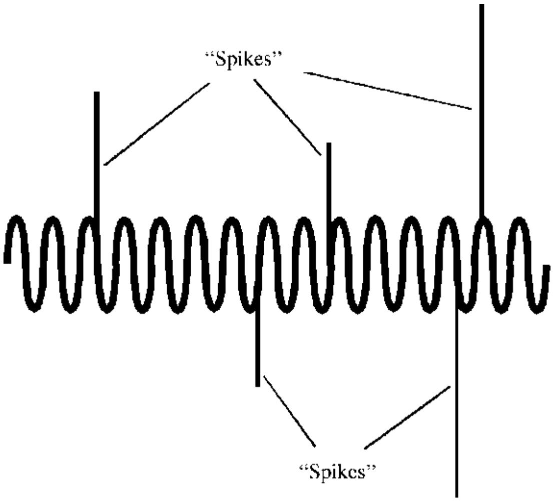

transient, 1.

transient absorber See SURGE ARRESTER.

transient arrester See SURGE ARRESTER.

transient-based amplifier See CRYSTAL AMPLIFIER, **2**.

transient current A momentary pulse of current. Also see TRANSIENT.

transient response The response of a circuit to a transient, as opposed to its steady-state response, usually evaluated in terms of its ability to reproduce a square wave.

transient suppressor See SURGE ARRESTOR.

transient voltage A momentary pulse of voltage. Also see TRANSIENT.

transient voltmeter An instrument that indicates the voltage of momentary signals. Usually, a peak-reading meter in the instrument displays the highest positive or negative value the transient attains (sometimes holding the deflection for reading later) and can respond to pulses of 1 microsecond in duration.

transistor An active (commonly three-terminal) semiconductor device capable of amplification, oscillation, and switching action. The name is a contraction of *transfer resistor.* Also see ALLOY-DIFFUSED TRANSISTOR, ALLOY TRANSISTOR, BIPOLAR TRANSISTOR, DIFFUSED-BASE TRANSISTOR, DIFFUSED EMITTER-AND-BASE TRANSISTOR, DIFFUSED-JUNCTION TRANSISTOR, DIFFUSED MESA TRANSISTOR, DIFFUSED PLANAR TRANSISTOR, DIFFUSED TRANSISTOR, DIFFUSION TRANSISTOR, DOUBLE-BASE JUNCTION TRANSISTOR, FIELD-EFFECT TRANSISTOR, FIELDISTOR, GERMANIUM TRANSISTOR, GROWN-DIFFUSED TRANSISTOR, GROWN-JUNCTION TRANSISTOR, JUNCTION TRANSISTOR, MESA TRANSISTOR, METAL-OXIDE-SEMICONDUCTOR FIELD-EFFECT TRANSISTOR, MICROALLOY DIFFUSED TRANSISTOR, MICROALLOY TRANSISTOR, PHOTOTRANSISTOR, PLANAR EPITAXIAL PASSIVATED TRANSISTOR, PLANAR TRANSISTOR, POINT-CONTACT TRANSISTOR, POWER TRANSISTOR, SILICON TRANSISTOR, SURFACE-BARRIER TRANSISTOR, SURFACE-CHARGE TRANSISTOR, TANDEM TRANSISTOR, THIN-FILM TRANSISTOR, THREE-JUNCTION TRANSISTOR, and UNIJUNCTION TRANSISTOR.

transistor amplifier An amplifier containing only transistors as the active components. Also called *transistorized amplifier.*

transistor analyzer An instrument for measuring the electrical characteristics of transistors. Compare TRANSISTOR TESTER.

transistor battery A 9-volt electrochemical battery consisting of six tiny zinc–carbon or alkaline cells in series. Each of the six cells supplies 1.5 volts. The ampere-hour capacity is small. The battery has a characteristic box shape with adjacent clip-on terminals at one end. These batteries are used in low-current electronic devices such as remote-control garage-door openers, television channel changers, remote videocassette recorder (VCR) controls, electronic calculators, and smoke detectors.

transistor-coupled logic Abbreviation, TCL. In computer and automatic-control operations, logic circuitry and systems using multi-emitter-coupled transistors. Also see RESISTOR-TRANSISTOR LOGIC and TRANSISTOR-TRANSISTOR LOGIC.

transistor current meter An ammeter, milliammeter, or microammeter circuit containing an amplifier using only transistors. Also see ELECTRONIC CURRENT METER.

transistorized voltmeter See TRANSISTOR VOLTMETER.

transistorized voltohmmeter An electronic voltmeter-ohmmeter combination using a transistorized circuit similar to that of a vacuum-tube voltohmmeter.

transistor keyer A power transistor acting as a keying device.

transistor oscillator An oscillator using only transistors as the active components. Also called *transistorized oscillator.*

transistor power supply A high-current, well-filtered, direct-current power supply for operating transistor circuits.

transistor radio Any small, portable, battery-powered radio receiver whose active components are transistors and/or other semiconductor devices.

transistor-resistor logic See RESISTOR-TRANSISTOR LOGIC.

transistor tester An instrument for checking the condition of transistors (i.e., whether good or bad). Compare TRANSISTOR ANALYZER.

transistor tetrode See DOUBLE-BASE JUNCTION TRANSISTOR.

transistor thyratron See SOLID-STATE THYRATRON.

transistor-transistor logic Abbreviation, TTL or T^2L. In computer operations, a circuit in which the multiple-diode cluster of the diode-transistor logic circuit has been replaced by a multiple-emitter transistor.

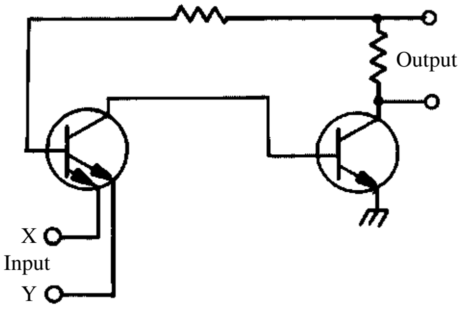

transistor-transistor logic

transistor voltmeter Abbreviation, TVM. A voltmeter containing an amplifier that uses only bipolar transistors. Also called *transistorized voltmeter.* Also see ELECTRONIC VOLTMETER. Compare FET VOLTMETER and VACUUM-TUBE VOLTMETER.

transistor VOM See TRANSISTORIZED VOLTOHMMETER.

transit angle For an electron, the angular frequency multiplied by the time required to travel from one point to another.

transit time The time taken by an electron to travel from one electrode to another—especially from the cathode to the plate in a vacuum tube.

transistor element A metallic element whose atoms have valence electrons in two shells. Examples: chromium, iron, and nickel.

transition factor See MISMATCH FACTOR.

transition region See BARRIER, **1**.

translation loss See PLAYBACK LOSS.

translator **1.** See COMPILER. **2.** See ASSEMBLER.

transliteration **1.** To change the characters in one alphabet or code system to the characters in a different system. **2.** The function that maps the characters in one alphabet or code system to those in another.

translucence **1.** The transmission of radiation, especially visible light, through a material. **2.** The extent to which a substance can transmit radiation—especially visible light. **3.** Pertaining to a material that partially or totally transmits radiation—especially visible light.

transmission gain Current amplification, power amplification, or voltage amplification.

transmission line **1.** A single conductor or group of conductors for carrying electrical energy from one point to another. **2.** A correctly dimensioned conductor or pair of conductors for carrying radio-frequency energy from a transmitter to an antenna or coupling device.

transmission mode **1.** In a transceiver, the condition in which the transmitter is enabled and the receiver is disabled. **2.** In a waveguide, propagation via transverse waves.

transmission speed The number of information elements (words, code groups, data symbols, bits, and bytes) that can be generated or received per unit time (second or minute) by a system or operator.

transmission wavemeter A (usually simple) inductance-capacitance-tuned wavemeter that provides peak response when tuned to the frequency of a signal passing through it; also, the comparable microwave device. Compare ABSORPTION WAVEMETER.

transmit-receive switch A manually or electrically operated switch for transferring a single antenna between a transmitter and receiver.

transmittancy The relative ability of a substance to transmit radiation. Transmittancy depends on the frequency of the radiation, as well as the substance.

transmitter **1.** An equipment for producing and sending signals or data. **2.** See MICROPHONE. **3.** One who originates signals or data.

transmitter–receiver A transmitter and receiver usually contained in a single enclosure, but with separate and independent oscillators, mixers, and amplifiers. The main asset of this scheme is the fact that it is possible to conduct split-frequency communication on bands whose wavelengths differ greatly. Another asset is the ability of the system, with proper design, to operate in full-duplex mode. Installations of this type are often used in satellite communications. Compare TRANSCEIVER, **1.**

transmitting antenna An antenna designed expressly for the efficient radiation of electromagnetic energy into space.

transmitting station A station that only transmits signals (i.e., it engages in no official form of reception). Compare RECEIVING STATION.

transmittivity The degree to which a selective circuit transmits a desired signal. Compare REJECTIVITY.

transmultiplexer A device that changes a signal from one multiplexed form to another while maintaining all of the information contained in the signal. For example, a transmultiplexer might convert time-division-multiplex data to frequency-division multiplex or vice-versa.

transonic Equal to, or approximating, the speed of sound in air (approximately 1100 feet per second).

transparence The practically unimpeded transmission of radiation, such as light, through a material. Compare OPACITY and TRANSLUCENCE.

transponder An acronym for transmitter and responder. **1.** A combination transmitter–receiver that automatically transmits an identification signal whenever it receives an interrogating signal. **2.** In a communications satellite, a broadband repeater that receives signals sent up from the earth, converts them to another frequency, and retransmits them back to the earth. Some satellites have units that employ multiplexing, data storage and retrieval systems, and other schemes to maximize the amount and variety of data they can handle.

transponder overload A condition that occurs when a satellite transponder is heavily used or when an excessively strong uplink signal is received. During heavy usage, all downlink signals are consistently weaker than they are when the transponder is not dealing with many signals. If an overly powerful uplink signal is received, all downlink signals are attenuated intermittently; the greater the instantaneous power of the offending uplink signal, the greater the instantaneous attenuation of all the downlink signals. See TRANSPONDER, **2.**

transport See TAPE TRANSPORT.

transportable equipment Portable electronic equipment. See, for example, PORTABLE TRANSMITTER.

transuranium An element whose atomic number is higher than that of uranium.

transverse Occurring in a direction or directions perpendicular to the direction of propagation.

transverse electric mode In a waveguide, the condition in which the electric lines of flux are perpendicular to the direction of propagation. Compare TRANSVERSE MAGNETIC MODE. Also see WAVEGUIDE MODE.

transverse electromagnetic wave An electromagnetic wave having electric-field vectors and magnetic-field vectors perpendicular to the direction of propagation.

transverse magnetic mode In a waveguide, the condition in which the magnetic lines of flux are perpendicular to the direction of propagation. Compare TRANSVERSE ELECTRIC MODE. Also see WAVEGUIDE MODE.

trap **1.** See WAVETRAP. **2.** In a semiconductor crystal, an imperfection capable of trapping current carriers.

trapezoid **1.** A polygon having four sides, of which only two are parallel. **2.** See TRAPEZOIDAL PATTERN. **3.** See TRAPEZOIDAL WAVE.

trapezoidal distortion In television or facsimile, a form of distortion in which the frame is wider at the top than at the bottom, or vice versa.

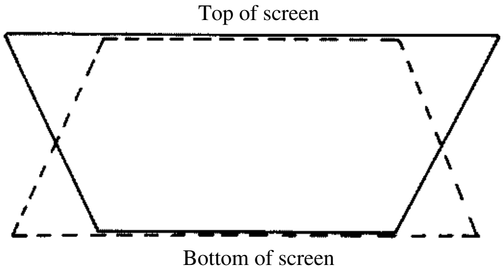

trapezoidal distortion

trapezoidal pattern An oscilloscope pattern used to check the percentage of modulation of an amplitude-modulated wave. Its name is derived from its trapezoidal shape.

trapezoidal wave A nonsinusoidal wave that is a combination of a rectangular component and a sawtooth component. It is the required waveform of the voltage applied to a magnetic deflecting coil (oscilloscope or television) to ensure a sawtooth wave of current in the coil.

traveling-wave amplifier Abbreviation, TWA. An amplifier based on the unique operation of the traveling-wave tube.

traveling-wave tube Abbreviation, TWT. A microwave tube containing an electron gun, helical transmission line, collector, and input and output couplers. A microwave signal is coupled into the helix, through which it travels while the gun projects an electron beam through the helix. When wave and electron velocities are equal, a power gain is obtained in the signal coupled out of the helix. Also see BACKWARD-WAVE OSCILLATOR.

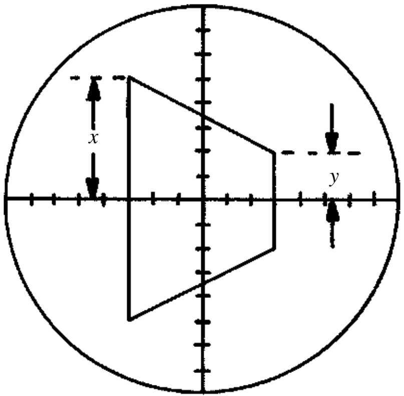

$$\text{Modulation (\%)} = 100\left(\frac{x-y}{x+y}\right)$$

trapezoidal pattern

Travis discriminator A discriminator circuit in which the diodes are operated from separately tuned halves of the secondary winding of the input transformer. Compare FOSTER-SEELEY DISCRIMINATOR.

treasure locator See METAL LOCATOR.

treble The higher portion of the sound spectrum—especially the upper end of the musical scale (middle C and above). Compare BASS.

treble boost **1.** Amplification of high audio frequencies (the treble notes) in an audio system—especially in high-fidelity music reproduction. **2.** The extent of amplification of high audio frequencies in audio applications—especially in music reproduction. **3.** A control that allows adjustment of the relative treble gain in an audio system.

treble control See TREBLE BOOST, **3**.

tree **1.** A cause-and-effect chain with two or more independent branches. **2.** A circuit with two or more branches but no meshes.

TRF Abbreviation of *tuned radio frequency.*

TRF amplifier See TUNED RADIO-FREQUENCY AMPLIFIER.

TRF receiver See TUNED RADIO-FREQUENCY RECEIVER.

triac A three-terminal, gate-controlled, semiconductor switching device. Compare DIAC.

triangular wave See BACK-TO-BACK SAWTOOTH.

triangulation A method of radiolocation in which directional bearings are obtained for the signals from two transmitters whose positions are accurately known. The position of the receiver is indicated by the intersection of two lines on a

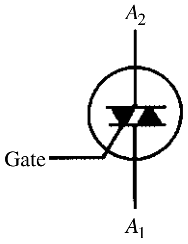

triac

map, drawn emanating from the transmitter locations 180 degrees opposite their respective bearings.

Triax A double-shielded coaxial cable. A center conductor is surrounded by two concentric, independent shield conductors.

triaxial connector A male or female connector having three concentric contacting surfaces. By contrast, the coaxial jack and coaxial plug have only two such surfaces.

triaxial speaker A dynamic loudspeaker containing three coils and three cones, for high-, low-, and middle-frequency ranges. It is, in effect, three speakers in one. It is commonly used in high-fidelity music reproduction systems. Also called *triaxial driver.* Compare THREE-WAY SPEAKER.

triboelectric Pertaining to frictional electricity.

triboelectric series See ELECTROSTATIC SERIES.

triboluminescence Luminescence produced by means of friction.

trickle charge A continuous slow charge of a storage battery, in which the charging rate is just sufficient to compensate for internal losses or normal discharge.

trickle charger A light-duty unit for charging a battery gradually at low current.

tricorner reflector A device that reflects a ray of incident energy at a 180-degree angle, regardless of the direction from which the ray approaches. Thus, the ray will return to its source, unless some obstruction intervenes. It consists of three mutually perpendicular reflecting plane surfaces that intersect in a common point. The geometry is identical to that in a typical room, where the ceiling meets two walls. It is useful at microwave, infrared, visible, and ultraviolet wavelengths.

triethanolamine An amino alcohol that precipitates metallic silver from a silver-nitrate solution in the deposition of a silver surface on a substrate, such as glass or a ceramic.

trig Abbreviation of TRIGONOMETRY.

trigatron A form of electrically operated switch. The circuit is closed by the breakdown of an electrical gap.

trigger **1.** A pulse used to start or stop the operation of a circuit or device, such as a flip-flop. **2.** To place a circuit or device into or out of operation with a start or stop pulse, as defined in **1**.

trigger diode See DIAC and FOUR-LAYER DIODE.

triggered multivibrator See DRIVEN MULTIVIBRATOR.

triggered sweep In an oscilloscope or similar device, a driven sweep.

triggering point The voltage level at which the action of an electronic switching device is initiated. Also see BREAKOVER POINT.

trigistor A three-junction semiconductor device that exhibits two-state operation, and is useful as a flip-flop or switch. Also see NPNP DEVICE.

trigonometric functions See CIRCULAR FUNCTIONS and HYPERBOLIC FUNCTIONS.

trigonometry The branch of mathematics devoted to the application of CIRCULAR FUNCTIONS in a plane. It is useful in electronics for determining impedance and phase.

trim To make a fine adjustment, as of a tuning control, balance control, output adjustment, etc.

trimmer A low-valued variable capacitor, inductor, or resistor operated in conjunction with a main unit (usually of the same sort) for vernier adjustment or range setting. See, for example, HIGH-FREQUENCY TRIMMER and OSCILLATOR TRIMMER.

trimmer capacitor A variable capacitor used as a trimmer.

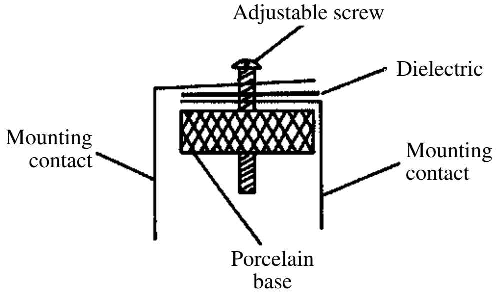

trimmer capacitor

trimmer coil See TRIMMER INDUCTOR.

trimmer inductor A variable inductor used as a trimmer.

trimmer resistor A variable resistor used as a trimmer.

trinary number system The base-3 system of notation. This system uses the digits 0, 1, and 2, the positional values being successive powers of 3 (e.g., decimal 14 equals trinary 112). Also called *ternary number system.*

Trinitron **1.** A single-gun color-television picture tube developed by Sony. **2.** A television set using this tube. The gun has three cathodes that modulate the three color beams (red, green, and blue). The beams are accelerated by common grids and are focused at different angles by convergence plates.

Trinoscope See TRINITRON.

triode A three-electrode tube or transistor, embodying an anode, cathode, and a control electrode as the principal elements.

triple-diffused transistor A diffused transistor in which the base and emitter are diffused into the top face of the chip, and the collector into the bottom face.

triple diode An assembly of three (often closely matched) semiconductor diodes in a single housing.

tripler **1.** A rectifier that delivers a direct-current output voltage at approximately triple the peak value of the alternating-current input voltage. **2.** An amplifier or other circuit that delivers an output signal at triple the frequency of the input signal.

triplexer In radar, a device that facilitates the use of two receivers at the same time.

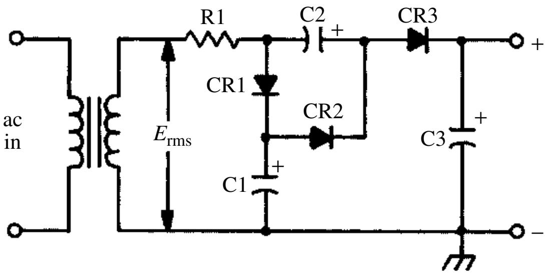

tripler, 1

Tri-state logic Digital logic in which there are three possible states, rather than the usual two. The conditions are defined as 0, 1, and undecided. Trade name of National Semiconductor.

tritium Symbol, T or H_3. An isotope of hydrogen whose nucleus contains two neutrons and one proton. Compare DEUTERIUM.

TRL Abbreviation of *transistor-resistor logic* (see RESISTOR-TRANSISTOR LOGIC).

Trojan horse A program written with the malicious intent of sabotaging the operating system and/or files in a computer. It is somewhat similar to a VIRUS; it is occasionally spread via software whose origin is questionable or unknown.

troposcatter See TROPOSPHERIC-SCATTER PROPAGATION.

troposphere The portion of the atmosphere in which virtually all weather phenomena occur. It extends from the surface to an altitude of 8 to 12 miles above sea level.

tropospheric bending A common type of radio-wave propagation that takes place when radio waves are refracted in the lower atmosphere. This effect is most dramatic near weather fronts, where warm, relatively light air lies above cool, heavy air. The cooler air has a higher index of refraction than the warm air, causing radio-frequency electromagnetic fields to be bent downward at a considerable distance from the transmitter. It is often responsible for anomalies in reception of FM and TV broadcast signals. Compare TROPOSPHERIC DUCTING, TROPOSPHERIC-SCATTER PROPAGATION.

tropospheric ducting Also called *duct effect.* A form of tropospheric propagation that occurs close to the earth's surface at very-high frequencies (VHF) and ultra-high frequencies (UHF). A duct forms when a layer of cool air becomes sandwiched between two layers of warmer air. Total internal reflection of the electromagnetic (EM) field takes place inside the duct. For this phenomenon to provide communications, both the transmitting and receiving antennas must be located within the same duct, and this duct must be unbroken and unobstructed between the two locations. A duct might measure only a few feet from top to bottom, but cover thousands of square miles parallel to the surface. Compare TROPOSPHERIC BENDING, TROPOSPHERIC-SCATTER PROPAGATION.

tropospheric propagation Also called *tropo.* Over-the-horizon propagation of radio-frequency electromagnetic waves that occurs as a result of reflection, scattering, and/or refraction in the troposphere, the lowest 10 or 12 miles of the earth's atmosphere. At wavelengths shorter than about 15 m (frequencies above 20 MHz), refraction and reflection can take place within and between air masses of different density. The air also produces some scattering of EM energy at wavelengths shorter than about 3 m (frequencies above 100 MHz). "Tropo" can often allow

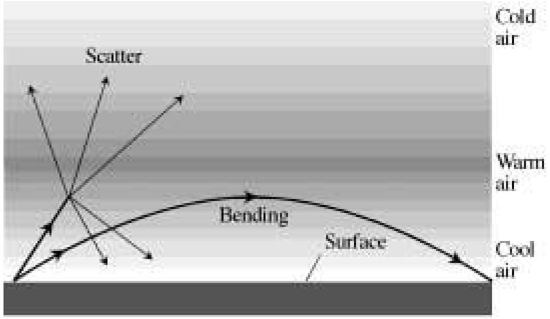

tropospheric propagation

communications over distances of several hundred miles. Also see TROPOSPHERIC BENDING, TROPOSPHERIC DUCTING, and TROPOSPHERIC-SCATTER PROPAGATION.

tropospheric-scatter propagation Also called *troposcatter.* A form of tropospheric propagation that occurs at frequencies above about 100 MHz, where the atmosphere has a scattering effect on radio waves. Dust in the air increases the scattering effect, but some troposcatter occurs regardless of the weather. Troposcatter takes place mostly at low altitudes where the air is the most dense. This mode can provide reliable communication over distances of several hundred miles when the appropriate equipment is used. Communication via troposcatter requires the use of high-gain antennas. Scatter is often observed along with other modes of tropospheric propagation. Compare TROPOSPHERIC BENDING, TROPOSPHERIC DUCTING.

troubleshoot **1.** To look for the cause of equipment failure. **2.** To look for flaws in computer software or to debug a program.

troubleshooting test A test that is part of the procedure for finding the cause of faulty electronic equipment operation. Also see DIAGNOSTIC TEST. Compare PERFORMANCE TEST.

trough value The minimum amplitude of a composite current or voltage.

true complement See RADIX COMPLEMENT.

true ground The earth, as opposed to an artificial ground, such as that provided by the radials of a ground-plane antenna or an equipment chassis.

true power In an alternating-current (ac) circuit, the power actually dissipated in the resistive component; the reactive component consumes no power. In ac circuits containing reactance, the true power is less than the product of the voltage and current. Also see AC POWER and POWER FACTOR.

truncation A method of approximating a quantity by cutting off its digits beyond a certain point. For example, 3.44 and 3.46, when truncated to two significant digits, both become 3.4. Compare ROUNDING.

trunk **1.** A communications link between two points, one usually being external. **2.** The interface between a central processing unit and a peripheral device.

trunk link In computer systems, an interface permitting access to main storage via a peripheral.

truth table In logic analysis and logic circuit design, a table in which are listed all combinations of input values and the corresponding output values for a given logic function.

ts Abbreviation of tensile strength.

Tschebyscheff filter See CHEBYSHEV FILTER.

TSS Abbreviation of TIME-SHARING SYSTEM.

T switch See TEE SWITCH.

TST switch See TIME-SPACE-TIME SWITCH.

truth table

$(x + y)' + xz$

x	y	z	$x + y$	$(x + y)'$	xz	$(x + y)' + xz$
0	0	0	0	1	0	1
0	0	1	0	1	0	1
0	1	0	1	0	0	0
0	1	1	1	0	0	0
1	0	0	1	0	0	0
1	0	1	1	0	1	1
1	1	0	1	0	0	0
1	1	1	1	0	1	1

TTL Abbreviation of TRANSISTOR-TRANSISTOR LOGIC. (Also, T^2L.)

TTT In radiotelegraphy, a signal indicating that a message concerning safety is to follow (equivalent to *securite* in radiotelephony).

TTY Abbreviation of TELETYPE.

T-type antenna See TEE ANTENNA.

T-type attenuator See TEE PAD.

tube **1.** Generic term for any electron tube (e.g., vacuum tube, gas tube, cathode-ray tube, X-ray tube, etc.). **2.** Glow lamp: argon bulb, neon bulb, mercury-vapor lamp, etc.

tube capacitances The internal capacitances between the elements of an electron tube.

tube diode A two-element (cathode and anode) electron tube for current rectification. Also see DIODE.

tube of flux A group of flux lines within a circular cross section. It can vary in diameter as the density of the flux lines changes.

tube parameters Operating coefficients of electron tubes (e.g., plate current, grid voltage, screen current, transconductance, amplification factor, etc.).

tube tester An instrument for checking one or more of the parameters of an electron tube.

tubular capacitor A fixed capacitor consisting of a wound section enclosed in a cylindrical can.

tune **1.** To adjust a selective circuit to accept or reject a signal. **2.** To correct the natural frequency of vibration of a body. **3.** To adjust a radio transmitter for optimum output. **4.** To adjust the frequency of an oscillator—especially in a radio transmitter or receiver. **5.** To adjust the resonant frequency of an antenna, antenna coupler, or antenna system.

tuned AF amplifier **1.** An audio amplifier that is continuously tunable over a band of frequencies. **2.** An audio oscillator that is set to a fixed, precise frequency. See, for example, PARALLEL-TEE AMPLIFIER.

tuned-base oscillator A self-excited, common-emitter connected, bipolar-transistor oscillator in which the tuned tank is in the base circuit.

tuned circuit A (usually series-resonant or parallel-resonant) circuit adjusted to accept or reject a signal. Also see RESONANCE.

tuned-collector oscillator A self-excited, common-emitter connected, bipolar-transistor oscillator in which the tuned tank is in the collector circuit.

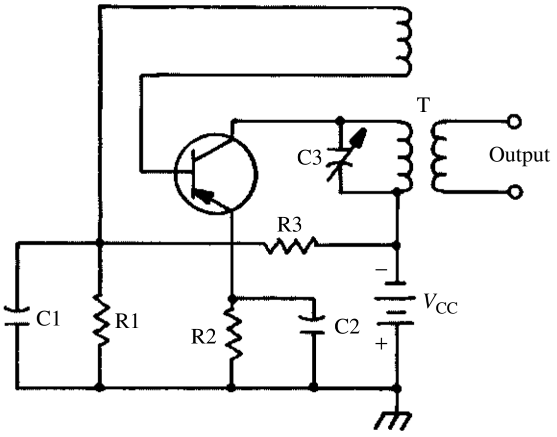

tuned-collector oscillator

tuned coupler An antenna coupler (transmatch) that can be adjusted independently of the transmitter or receiver with which it is used.

tuned dipole A half-wave, center-fed, resonant antenna.

tuned feeders An antenna feed line that is adjusted or trimmed so that the entire system (feeders and radiating element) is resonant at the transmitted-signal frequency.

tuned headphones Headphones used in radio-telegraphy that are fix-tuned to a single audio frequency (e.g., 1 kHz) by means of a small parallel capacitor.

tuned line An antenna wire or transmission line that provides a resistive load at a specific resonant frequency.

tuned pickup A pickup circuit or device (such as a radio-frequency sampling coil) that is tuned to the signal frequency.

tuned radio-frequency amplifier An amplifier circuit that is continuously tunable over a specific band of radio frequencies.

tuned radio-frequency receiver A radio receiver consisting only of a tuned-radio-frequency amplifier, detector, and audio amplifier. Compare SUPERHETERODYNE RECEIVER.

tuned reed A vibrating reed whose length, width, and/or thickness have been adjusted so that it vibrates naturally at a desired frequency.

tuned-reed frequency meter An audio frequency meter using tuned metal reeds as the indicators. Also see POWER FREQUENCY METER, **2**.

tuned relay An electronic or electromechanical relay that closes at one frequency. See, for example, REED SWITCH, **1**.

tuned signal tracer A signal tracer that can be tuned sharply to the frequency of the test signal being traced.

tuned transformer An audio-frequency or radio-frequency transformer tuned via a capacitor in parallel with its primary winding and/or a capacitor in parallel with it secondary winding.

tuner A circuit or device that can be set to select one signal from a number of signals in a frequency band.

tungsten Symbol, W. A metallic element. Atomic number, 74. Atomic weight, 183.85. Also called *wolfram.* It is used in switch and relay contacts, in the elements of some electron tubes, and in incandescent-lamp filaments.

tuning **1.** Adjustment of the frequency of a receiver to intercept a signal on a given frequency. **2.** Adjustment of a transmitter oscillator to a desired frequency. **3.** Adjustment of an inductance-capacitance circuit for resonance on a desired frequency. **4.** Adjustment of an antenna or antenna system to a desired frequency. **5.** Adjustment of a radio-frequency amplifier for optimum performance. **6.** Alignment of a musical instrument for correct tone frequency.

tuning capacitor A variable capacitor used to tune an inductance-capacitance circuit (series-resonant or parallel-resonant).

tuning coil A variable inductor used to tune an inductance-capacitance circuit (series-resonant or parallel-resonant).

tuning core See TUNING SLUG.

tuning diode See VOLTAGE-VARIABLE CAPACITOR, **1**.

tuning fork A metal device that vibrates at a precise audio frequency when struck physically. It usually has two prongs and looks something like a fork.

tuning-fork oscillator See FORK OSCILLATOR.

tuning indicator **1.** A meter or display that indicates when a receiver is properly tuned to the frequency of an incoming signal. **2.** Sometimes, a bridge null detector.

tuning meter A meter-type resonance indicator.

tuning potentiometer A single or ganged potentiometer (depending on the circuit in which it is used) used to vary the frequency response of a resistance-capacitance-tuned circuit, such as the bridge-tee, parallel-tee or Wien-bridge circuit.

tuning slug A powdered-iron core that slides or screws in and out of a coil to vary the inductance.

tuning voltage An adjustable direct-current voltage used to vary the capacitance of a varactor diode (serving as the capacitor in a tuned circuit).

tuning wand See NEUTRALIZING TOOL.

tunnel diode A specially processed junction diode whose forward conduction characteristic displays negative resistance.

tunnel-diode amplifier A circuit that uses the negative-resistance properties of a tunnel diode to produce gain. It is used primarily at ultra-high and microwave frequencies.

tunnel-diode oscillator A circuit that uses the negative-resistance properties of a tunnel diode to produce oscillation at a specific frequency. It is used primarily at ultra-high and microwave frequencies.

turbidimeter A device used to measure the turbidity (relative opaqueness) of liquids.

turbidity meter See TURBIDIMETER.

Turing machine A hypothetical model, conceived by Alan Turing in the 1930s, devised as an exercise in problem solving. In effect, it was a computer programming language with a limited instruction set, operating in a computer with infinite memory and storage capacity.

Turing test A scheme devised by Alan Turing, designed to test computers for artificial intelligence (AI). If the machine can fool a person into thinking it is another person most of the time, then the machine in effect "passes" the test, indicating that it has some AI.

turn One complete loop of a coil.

turnaround time **1.** The length of time required for a repair to be performed. **2.** The length of time required for a computer program to be completed. **3.** The time required for a trans-ceiver to switch from the full-transmit to the full-receive condition.

turn factor For a given inductor core, the number of turns (in a standard configuration) that results in an inductance of 1 H.

turn-off time The time required for the operation of a circuit or device to cease completely after a signal or operating power has been removed. Compare TURN-ON TIME.

turn-on time The time required for a circuit or device to come up to full (normal) operation after application of a trigger or operating power. Compare TURN-OFF TIME.

turnover pickup See DUAL PICKUP.

turns ratio For a transformer, the ratio of the number of turns in the primary winding to the number of turns in the secondary winding.

turnstile antenna A polyphase antenna that resembles a turnstile gate.

turntable **1.** The motor-driven, rotating platter on which a phonograph disc rests during recording or reproduction. **2.** In a high-fidelity phonograph system, the assembly containing the platter, as defined in **1**.

turret tuner A television front-end tuner having a separate tuned-circuit section for each channel. Turning a knob rotates a desired section into position against switch contacts. Also called DETENT TUNER.

TV **1.** Abbreviation of TELEVISION. **2.** Abbreviation of *terminal velocity.*

T/V Abbreviation of *temperature to voltage.*

TV decoder A device for unscrambling a television broadcast that has been intentionally encrypted to prevent its use by unauthorized viewers. See SUBSCRIPTION TV.

TV ghost See GHOST.

TVI Abbreviation of TELEVISION INTERFERENCE.

TVL Amateur radio abbreviation of *television listener* (looker).

TVM Abbreviation of TRANSISTORIZED VOLTMETER.

TVO Abbreviation of TRANSISTORIZED VOLT-OHMMETER. (Also, TVOM.)

TVOM Abbreviation of TRANSISTORIZED VOLT-OHMMETER. (Also, TVO.)

TV projector A combination electronic and optical device for projecting television images on a large screen.

TV sound marker See SOUND MARKER.

TVT Abbreviation of *television terminal* (computer peripheral).

TW Abbreviation of TERAWATT.

TWA Abbreviation of TRAVELING-WAVE AMPLIFIER.

tweeter A loudspeaker that favors the extreme treble (high) audio frequencies, ranging from 4000 or 5000 Hz to more than 20,000 Hz. Compare WOOFER.

twin diode See DUAL DIODE.

twinlead See FLAT-RIBBON LINE.

twin line See FLAT-RIBBON LINE.

twin meter See DUAL METER.

twin-T measuring circuit See PARALLEL-TEE MEASURING CIRCUIT.

twin-T network See PARALLEL-TEE NETWORK.

twisted pair A simple wire line used for communications. It consists of two wires twisted together to form a helix.

twister A piezoelectric crystal, such as one of Rochelle salt, that generates a voltage when torqued.

twistor A magnetic-memory element consisting of a winding of magnetic wire on a length of nonmagnetic wire.

two-channel amplifier See DUAL-CHANNEL AMPLIFIER.

two-channel recorder See DOUBLE-CHANNEL RECORDER.

two-electrode amplifier See DIODE AMPLIFIER.

two-electrode tube See TUBE DIODE.

two-element amplifier See DIODE AMPLIFIER.

two-element beam A Yagi antenna consisting of a driven element and one parasitic element. The parasitic element can be either a reflector, about 5% longer than the resonant half wavelength at the operating frequency, or it can be a director, about 5% shorter than a half wavelength at the operating frequency.

two-element tube An electron tube having two principal electrodes (i.e., a tube diode).

two-five code See BIQUINARY CODE.

two-phase Pertaining to circuits or devices in which two components (two voltages, two

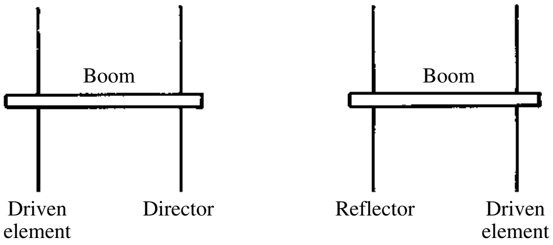

two-element beam

currents, or a current and a voltage) are 90° out of phase with each other.

two-phase system See QUARTER-PHASE SYSTEM.

two-phase, three-wire circuit A circuit using three conductors, one of which (the common return) is 90° out of phase with the other (outer) two.

two-pincher gripper A simple robotic end effector using two tongs, and resembling large tweezers or needle-nosed pliers. It grasps objects by "pinching" them.

two-point tuning See DOUBLE-SPOT TUNING.

two-space A two-dimensional mathematical continuum in which each point is uniquely defined by two variables in an ordered pair, such as (x,y), and each ordered pair corresponds to exactly one point in the continuum. A common scheme is the Cartesian coordinate system; another scheme is the polar coordinate system. See CARTESIAN COORDINATES and POLAR COORDINATES.

two-space coordinates A coordinate system for defining points in two-space. See CARTESIAN COORDINATES and POLAR COORDINATES.

two-state device A device having two stable states (e.g., a flip-flop).

two-state logic Digital logic in which there are two possible conditions (called high and low, true and false, or 1 and 0).

two-tone keying See FREQUENCY-SHIFT KEYING.

two-track recording Recording on two adjacent tracks on magnetic tape. The separate recordings can be in the same direction (as in stereo) or in opposite directions.

two-way amplifier An amplifier whose input and output terminals can be used interchangeably.

two-way communication The exchange of messages between two or more stations that transmit as well as receive. Compare ONE-WAY COMMUNICATION, **1**.

two-way radio **1.** Any form of TWO-WAY COMMUNICATION using electromagnetic waves. **2.** A radio transceiver—especially one used for voice communication at very-high or ultra-high frequencies.

two-way repeater In telephony, a device that amplifies and retransmits a signal in either direction. Also see TWO-WAY AMPLIFIER.

two-way speaker A woofer/midrange speaker and a tweeter occupying the same enclosure, and interconnected by a crossover network for wideband frequency response.

TWT Abbreviation of TRAVELING-WAVE TUBE.

TWX Abbreviation of TELETYPEWRITER EXCHANGE.

Twystron A form of microwave-oscillator tube, similar to the Klystron.

Type A telephone line A telephone line serving one subscriber (as opposed to a party line).

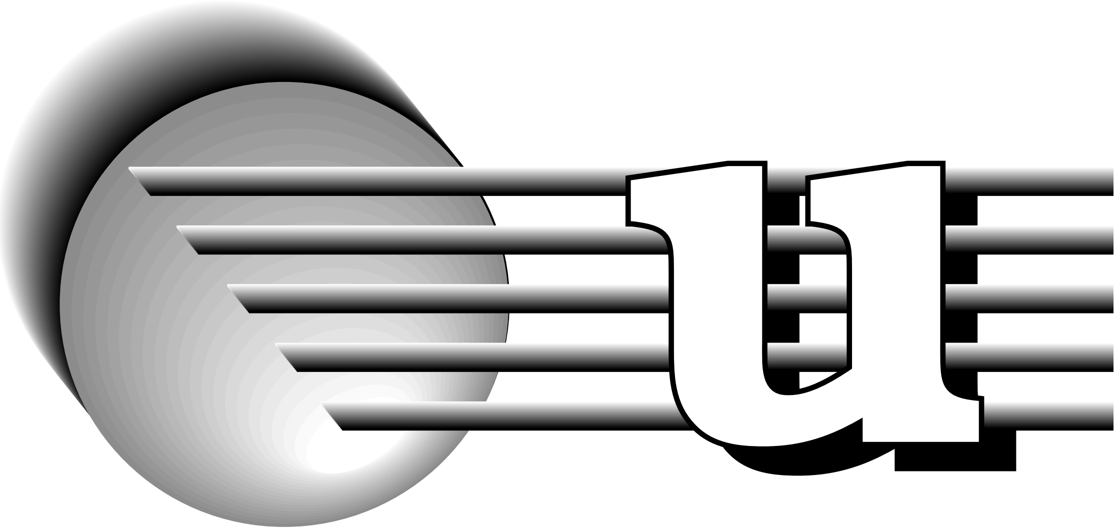

U **1.** Symbol for URANIUM. **2.** Abbreviation of UNIT.

u Symbol (ital.) for UNIFIED ATOMIC MASS UNIT.

Ua Abbreviation of *unit of activity* (electroencephalography).

UAM Abbreviation of *underwater-to-air missile.*

ubiquitous carrier A current carrier (especially an electron) whose velocity is so high, or mechanism of transfer so subtle, that it appears to be in two places simultaneously. Thus, in the tunnel diode, during the interval of tunneling a single electron, appears to be on both sides of the barrier at the same time.

Ubitron A tube in which a periodic magnetic field causes an electron beam to undulate. Through its transverse velocity component, the beam interacts with a radio-frequency wave, and its kinetic energy is converted into radio-frequency energy. The ubitron can be used as an amplifier or oscillator.

UCT Abbreviation of UNIVERSAL COORDINATED TIME.

UDC Abbreviation of *universal decimal classification.*

UDOP Abbreviation of *ultra-high-frequency Doppler.* A 440-MHz phase-coherent tracking system used to determine the velocity and position of missiles and space vehicles.

UEP Abbreviation of *underwater electric potential.*

UFET Abbreviation of unipolar field-effect transistor. (Also, UNIFET.) See UNIPOLAR TRANSISTOR.

UFO See UNIDENTIFIED FLYING OBJECT.

UG Abbreviation of underground.

UHF Abbreviation of ULTRAHIGH FREQUENCY.

UHF adaptor See UHF CONVERTER, **1**.

UHF capacitor A button-type capacitor. Because of its unique design, it is an efficient bypass capacitor at ultrahigh frequencies.

UHF converter **1.** A circuit, usually consisting of a radio-frequency amplifier and mixer, for converting ultra-high-frequency (UHF) signals to a lower band of frequencies. **2.** A circuit for converting UHF television signals to very-high-frequency (VHF) signals so that they can be accommodated by an older (VHF only) television receiver. See UHF TELEVISION CHANNELS and VHF TELEVISION CHANNELS.

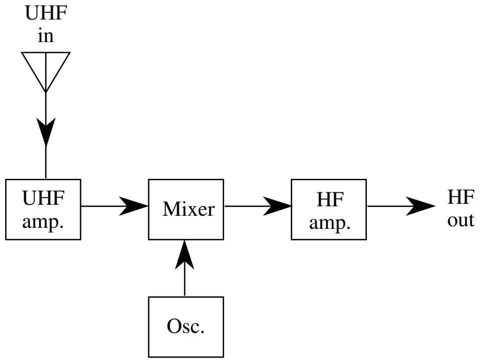

UHF converter, 1

UHF diode A semiconductor diode whose rectification efficiency is good at ultra-high frequencies.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

UHF generator **1.** An oscillating device (such as a transistor, tunnel diode, Klystron, or magnetron) used to produce radio-frequency energy at ultra-high frequencies (UHF). **2.** The equipment in which such a device is used. **3.** An ultra-high-frequency test-signal generator.

UHF loop A (usually single-turn) loop antenna with a toroidal radiation pattern perpendicular to the loop; the antenna has a natural wavelength in the ultra-high-frequency (UHF) range.

UHF receiver A receiver tunable to frequencies in any band in the range 300 MHz to 3000 MHz.

UHF television channels Television channels 14 through 69, whose frequencies lie in the ultra-high range.

UHF transistor A transistor specially designed and fabricated for ultra-high-frequency (UHF) operation. It is characterized by extended beta cutoff frequency, low junction capacitance, and fast recovery time.

UHF translator A television broadcast translator station transmitting in an ultra-high-frequency (UHF) channel.

UHF transmitter A transmitter that is specially designed for operation at ultra-high frequencies. In such a transmitter, stray parameters are minimized and special tubes or transistors are required.

UHF tube A vacuum tube specially designed for ultra-high-frequency operation. It is characterized by low interelectrode capacitance, low input and output capacitances, short electron transit time, and low lead inductance.

UHF tuner **1.** A superheterodyne receiver front end tunable in the ultra-high-frequency (UHF) range. It is usually arranged to deliver the mixer output to a lower-frequency receiver. **2.** A television receiver front end tunable to the UHF channels.

UHR Abbreviation of ULTRA-HIGH RESISTANCE.

U index The difference between consecutive daily mean values of the horizontal component of the geomagnetic field.

UJT Abbreviation of UNIJUNCTION TRANSISTOR.

UL **1.** Abbreviation of UNDERWRITERS LABORATORIES. **2.** Abbreviation of ULTRALINEAR.

ULD Abbreviation of ULTRA-LOW DISTORTION.

ULF Abbreviation of ULTRA-LOW FREQUENCY.

U-line section A short section of coaxial cable or hardline having the shape of a squared U.

ULSI See ULTRA-LARGE-SCALE INTEGRATION.

ultimate attenuation The maximum attenuation in a circuit or device when it is operated outside its passband, or within its stopband.

ultimately controlled variable In an automated system, the quantity or operation (e.g., liquid level, temperature, pressure, precision drilling, etc.), whose control is the end purpose of the system.

ultimate-range ballistic missile A ballistic missile whose range is greater than half the earth's circumference (approximately 12,000 miles).

ultimate ratio The limiting value of a ratio (i.e., the value approached by a ratio).

ultimate sensitivity **1.** In an instrument or system, the maximum degree of perception of, or response to, a quantity or condition. **2.** In a graphic recorder, half the deadband.

ultimate threshold See ULTIMATE SENSITIVITY, **2**.

ultimate trip current For a specified set of environmental conditions, the lowest value of current that will trip a circuit breaker.

ultimate trip limits For a circuit breaker, the maximum and minimum current at which the breaker trips and drops out.

ultor anode The second anode of a television picture tube or oscilloscope tube.

ultor element In a television picture tube, the element that receives the highest direct-current voltage. Also called *ultor electrode.* Also see ULTOR ANODE.

ultor voltage The high direct-current voltage applied to the second anode of an oscilloscope tube or television picture tube. Also see ULTOR ANODE.

ultra- A prefix meaning *above*, *larger than*, *greater than*, or *higher than.* See the following several definitions for examples.

ultra-atomic Pertaining to particles smaller than atoms (i.e., SUBATOMIC).

ultra-audible frequency A frequency higher than the highest audible frequency. Also called ULTRASONIC FREQUENCY and *supersonic frequency.*

ultra-audion oscillator See ULTRAUDION OSCILLATOR.

ultracentrifuge A centrifuge that spins at a very high speed.

ultrafast switch An electronic switch capable of microsecond operation.

Ultrafax A system for high-speed transmission of printed matter by facsimile, radio, or television.

ultra-high frequency A radio frequency in the range 300 MHz to 3000 MHz, corresponding to free-space wavelengths between 1 meter and 10 centimeters.

ultra-high-frequency band See UHF BAND.

ultra-high resistance Resistance of 100 megohms or higher.

ultra-large-scale integration An integrated-circuit process that goes several orders of magnitude beyond VERY-LARGE-SCALE INTEGRATION. More than 4×10^6 transistors could be on a single chip.

ultra-linear Pertaining to the most linear operation obtainable with state-of-the-art electronic equipment.

ultra-linear amplifier A high-fidelity audio-frequency power amplifier having wide frequency response and very low distortion.

ultra-low distortion Total harmonic distortion (THD) of less than 0.1 percent.

ultra-low frequencies **1.** Radio frequencies between 300 Hz and 3 kHz, pertaining to wavelengths between 1000 kilometers and 100 kilometers. **2.** Audio frequencies lower than those in the human hearing range.

ultramicrometer An electronic instrument for extremely small linear measurements.

ultramicroscope An optical microscope using refracted light to illuminate minute particles.

ultramicrowave Pertaining to wavelengths in the range 10^{-3} to 10^{-6} meter (300 GHz to 300 THz).

ultraminiature See SUBMINIATURE.

ultraphotic rays Collectively, infrared and ultraviolet rays.

ultrared rays See INFRARED RAYS.

ultra-short waves Waves whose length correspond to ultra-high frequencies (i.e., wavelengths between 0.1 and 1 meter).

ultrasonic Of higher frequency than those that are audible (supersonic). See, for example, ULTRASONIC FREQUENCY.

ultrasonic bonding A method of bonding metal by means of physical vibration at frequencies above the human hearing range.

ultrasonic brazing Forming a nonporous bond between metal parts through the use of ultrasonic energy and a second, different metal (or alloy) having a lower melting point.

ultrasonic cleaning A method of cleaning delicate or intricately structured items (such as dentures, contact lenses, or jewelry) in which the soiled items are immersed in a fluid that is agitated by ultrasonic transducers; the foreign particles are shaken away.

ultrasonic cleaning tank A thick-walled stainless-steel tank with ultrasonic transducers mounted within its walls and used for ultrasonic cleaning.

ultrasonic coagulation The coagulation of a substance through ultrasonic agitation.

ultrasonic communication Underwater telegraphic communication between ships and/or submarines by keying the echo-ranging sonar equipment.

ultrasonic delay line A delay line, such as the mercury type, through which an ultrasonic signal is propagated. The delay results from the relatively slow propagation of the ultrasonic wave through the substance in the line.

ultrasonic densitometer A density-measuring apparatus whose operation is based on the time required for an ultrasonic signal to penetrate the material under test, or for the signal to penetrate the material and be reflected back to the transmitter.

ultrasonic depth finder See ACOUSTIC DEPTH FINDER.

ultrasonic detector A device that responds to ultrasonic waves by indicating their presence, intensity, and/or frequency.

ultrasonic diagnosis The medical determination of the condition of tissues or structures within the body, in terms of reflection of ultrasonic waves by those tissues or structures. The technique is useful in situations where other techniques, such as radiology or the use of catheter probes, are risky. Also see ULTRASOUND.

ultrasonic disintegrator In biology and related fields, a device that uses ultrasonic energy to rupture or shatter cells, tissues, or foreign bodies, such as kidney stones.

ultrasonic drill A drill running at speeds corresponding to ultrasonic frequencies. The tool is valuable in drilling hard or brittle materials and in dentistry.

ultrasonic filter **1.** A bandpass filter whose operation is based on the natural (ultrasonic) vibration frequency of small disks or rods of magnetostrictive metal. See MAGNETOSTRICTION. Also called *mechanical filter.* **2.** Generally, a filter operating at ultrasonic frequencies.

ultrasonic flaw detector A system analogous to radar, in which an ultrasonic wave is transmitted through a solid material and is reflected back to a detector and display device to reveal flaws, cracks, and strain in the material.

ultrasonic frequency For an acoustic disturbance, any frequency above the limit of human hearing (higher than 20 kHz).

ultrasonic generator **1.** An oscillator that operates at frequencies above the range of human hearing, and the output of which is intended for coupling to an electroacoustic transducer. **2.** Any device that produces ultrasonic waves.

ultrasonic grating constant For a sound wave producing a diffraction spectrum, a figure indicating the distance between diffracting centers of the wave.

ultrasonic heating The production of heat in a specimen by means of ultrasonic energy directed into or through it.

ultrasonic image converter A device that makes visible acoustic field patterns.

ultrasonic inspection The use of ultrasonic waves to detect internal flaws in solid materials, such as metals. It is valued particularly because it is nondestructive. Also see ULTRASONIC FLAW DETECTOR.

ultrasonic intrusion alarm A security system activated when an intruder disturbs a pattern of ultrasonic waves in the protected area; a sensitive relay closes, setting off an alarm.

ultrasonic level detector A level-monitoring system in which an ultrasonic transmitter and detector are mounted together on one wall of a tank or chamber. When the tank is empty, the transmitted signal is reflected by the opposite wall back to the detector. When the liquid rises, the reflection time is reduced; this change is used to operate a device that indicates that the tank has been filled to a desired level.

ultrasonic light diffraction Diffraction resulting from the periodic variation of light refraction when a beam of light passes through a longitudinal sound-wave field.

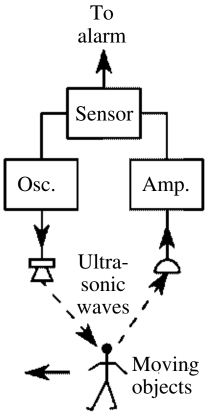

ultrasonic intrusion alarm

ultrasonic light modulator A device that modulates a light beam passing transversely through a fluid agitated by sound waves.

ultrasonic material dispersion The use of the agitative action of high-intensity ultrasonic waves to disperse or emulsify one substance in another.

ultrasonic plating The use of ultrasonic energy to deposit or bond one material onto the surface of another.

ultrasonic probe A rodlike director of ultrasonic energy.

ultrasonic relay An electronic relay actuated by ultrasound.

ultrasonics **1.** The use of high-frequency acoustic energy in industry, medicine, research, and the home. See ULTRASONIC FREQUENCY. **2.** The use of high-frequency acoustic energy for medical diagnosis and treatment. **3.** The branch of physics dealing with the effects and behavior of acoustic disturbances at frequencies above the range of human hearing.

ultrasonic soldering See ULTRASONIC BRAZING.

ultrasonic sounding Determining the depth of a body of water in terms of the time taken for a transmitted ultrasonic signal to be reflected back to a transmitting point on the surface of the water.

ultrasonic space grating A periodic variation in the index of refraction when acoustic waves are present in a light-transmitting medium. Also see ULTRASONIC LIGHT DIFFRACTION.

ultrasonic storage cell See ULTRASONIC DELAY LINE.

ultrasonic stroboscope A stroboscope that uses an ultrasonically modulated beam of light.

ultrasonic switch An electronic switch actuated by ultrasound.

ultrasonic therapy The use of ultrasonic energy in medicine for treatment of certain disorders.

ultrasonic thickness gauge An instrument that determines the thickness of a specimen in terms of the propagation time (or echo time) of ultrasonic waves through the specimen.

ultrasonic transducer A transducer that converts electrical energy into ultrasonic energy, or vice-versa. Common types are the quartz crystal, ceramic crystal, and magnetostrictive disk.

ultrasonic waves Acoustic waves whose length corresponds to ultrasonic frequencies (i.e., frequencies above 20 kHz).

ultrasonic welding A below-melting-point technique of joining two metallic bodies by clamping them tightly together and applying ultrasonic energy, rather than heat, in the plane of the desired weld. Compare ULTRASONIC BRAZING.

ultrasonic whistle A whistle whose pitch is beyond the range of human hearing. Although some are of the simile (blown) type, others are (usually miniature) electronic sound generators. These devices can be used for the remote control of television receivers, garage-door openers, and other equipment.

ultrasonography A method of examining human tissues or organs by transmitting ultrasonic waves into the body and receiving the echoes.

ultrasound **1.** Acoustic disturbances at frequencies above 20 kHz. **2.** See ULTRASONIC FREQUENCY. **3.** See ULTRASONOGRAPHY.

ultraviolet **1.** Electromagnetic radiation at wavelengths somewhat shorter than those of visible light. Longwave or *near ultraviolet* extends from approximately 390 nanometers (nm) down to 50 nm. Shortwave or *far ultraviolet* extends from 50 nm down to 4 nm. **2.** Pertaining to the behavior and effects of electromagnetic radiation in the range of approximately 390 nm to 4 nm.

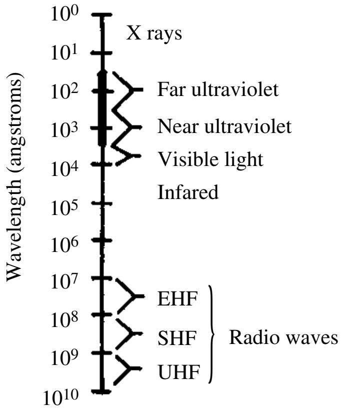

ultraviolet

ultraviolet altimeter An altimeter that uses ultraviolet rays instead of radio waves to determine the altitude of an aircraft or spacecraft.

ultraviolet lamp A lamp that delivers ultraviolet rays. Common types are arc lamps and mercury-vapor lamps.

ultraviolet light See ULTRAVIOLET RAYS.

ultraviolet power The power (in watts) of ultraviolet radiation.

ultraviolet rays Radiation at frequencies in the ultraviolet region (i.e., between the highest visible-light frequencies and the lowest X-ray frequencies).

ultraviolet therapy The use of ultraviolet rays by physicians and other health personnel in the treatment of certain disorders.

ultraviolet wavelength The wavelength range of ultraviolet radiation [e.g., 390 nanometers (nm) to 4 nm].

ultraviolet waves See ULTRAVIOLET RAYS.

umbilical cord **1.** A cord through which missiles and rockets are controlled and powered until they are launched. **2.** A cord connecting an astronaut to his space vehicle during extravehicular activity (e.g., during a space “walk”).

umbilical tower The vertical tower supporting the umbilical cords extending to a rocket in the launch position.

umbra **1.** The region of total (conical) shadow behind an object situated in the path of a radiation. **2.** The comparatively dark central region in a sunspot.

umbrella antenna An antenna consisting of a number of wires, extending from the top of a vertical mast to points on a circle below, at the center of which the mast is mounted. The wires are usually fed in parallel from the top.

Umklapp process A process responsible for thermal resistance in nonconducting materials. It results from collision between phonons, or phonons and electrons.

UMW Abbreviation of ULTRAMICROWAVE.

unabsorbed field strength The field strength with no absorption between transmitter and detector. The unabsorbed field strength is an ideal quantity; the actual field strength over a given distance is always less. It is expressed in volts per meter.

unamplified back bias A negative-feedback voltage developed across a fast-time-constant circuit within a single amplifier stage.

unamplified feedback A positive or negative current or voltage taken from the output of a system and presented to the input without being boosted through auxiliary amplification.

unamplified ALC Automatic level control (ALC) in which the control-signal voltage is taken from a point in the circuit and fed to the controlled point without being boosted by an auxiliary amplifier. Compare AMPLIFIED ALC.

unattended operation Operation, as of an electronically programmed machine, with minimal (or no) human supervision.

unattended time The period, excluding down time for repair or checkout, during which a computer is unpowered.

unattenuated Not reduced in intensity of amplitude. Thus, an unattenuated signal is one that has retained its original strength during its transmission through a system.

unavailable energy The difference between the quantity of heat energy supplied to a system and the available energy of the system.

unbalance **1.** Lack of balance or symmetry in a circuit, line, or system. **2.** The condition in which a bridge (or the equivalent) is not nulled.

unbalanced circuit See SINGLE-ENDED CIRCUIT.

unbalanced delta system A three-phase circuit in which the elements are connected in a triangular (delta) configuration, but are of unequal impedance. In an unbalanced delta system, no definite relationship exists between line and phase currents. Compare BALANCED DELTA SYSTEM.

unbalanced input See SINGLE-ENDED INPUT.

unbalanced line **1.** A transmission line in which one side (conductor) is grounded (e.g., a coaxial line). **2.** A normally balanced transmission line in which the currents in the two halves are out of balance (i.e., they are not equal in amplitude and opposite in phase).

unbalanced multivibrator A multivibrator in which the time constant of one base or gate circuit differs from that of the other base or gate circuit. The asymmetry results in short-duration output pulses separated by long time periods.

unbalanced output See SINGLE-ENDED OUTPUT.

unbalanced two-phase system A two-phase circuit in which the load elements are of unequal impedance.

unbalanced wire circuit A wire line whose (two) sides are dissimilar.

unbalanced-wye system A three-phase circuit in which the elements are connected in the familiar wye, or star, configuration, but are of unequal impedance.

unbalanced-Y system See UNBALANCED-WYE SYSTEM.

unbalance-to-balance transformer A transformer for matching an unbalanced transmission line to a balanced line (e.g., a coaxial line to a parallel-wire line).

unbiased limiting Peak limiting that results from overdriving an unbiased active device.

unbiased unit A device or circuit operated without bias [e.g., a transistor without base bias (common emitter and emitter follower) or emitter bias (common base)].

unblanking Removal of the blanking pulse in a cathode-ray-tube circuit (i.e., turning on the beam).

unblanking generator A (usually pulse-type) signal source for turning on the beam of a blanked cathode-ray tube.

unblanking interval The period during which the beam of a cathode-ray tube is turned on. Compare BLANKING INTERVAL.

unblanking pulse A pulse that turns on the beam of a blanked cathode-ray tube. Compare BLANKING PULSE.

unblanking time **1.** See UNBLANKING INTERVAL. **2.** The instant at which unblanking begins.

unbonded strain gauge A strain gauge that is not directly attached (by cement, for example) to the strained surface.

unbound electron A free electron (i.e., an electron not confined to a shell within an atom).

unbuffered output An output (signal, power, etc.) that is delivered directly from the generating device without the benefit of an isolating stage, such as a buffer amplifier. Compare BUFFERED OUTPUT.

unbypassed emitter resistor In a common-emitter transistor circuit, an emitter resistor without a bypass capacitor. The flow of output-signal current through the resistor produces negative feedback within a single stage.

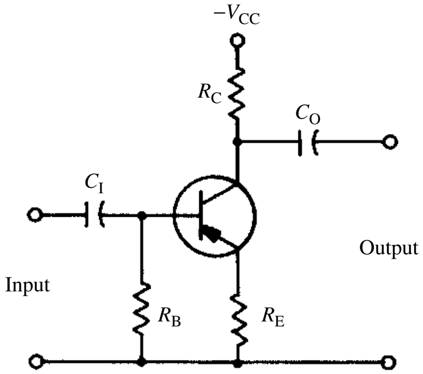

unbypassed emitter resistor (R_E)

unbypassed source resistor In a common-source field-effect transistor circuit, a source resistor without a bypass capacitor. The flow of output-signal current through the resistor produces negative feedback within a single stage.

uncage In a displacement gyroscope system, to disconnect the erection circuit.

uncalibrated unit A component, circuit, or instrument that has never been calibrated or has not recently been calibrated and, hence, is of questionable accuracy.

uncanny valley According to a theory put forth by certain Japanese roboticists, a phenomenon that can occur among people as technology advances. As machines rapidly get more sophisticated, a point is reached at which their human users become unnerved, partially offsetting the advantages the technology has to offer.

uncased choke See UNSHIELDED CHOKE.

uncased transformer See UNSHIELDED TRANSFORMER.

uncertainty in computation The degree of doubt concerning the exactness of computations. This uncertainty is always greater than that of the roughest measurement used in obtaining data for the computation. Uncertainty is closely related to the tolerance of the instruments and formulas used.

uncertainty in measurement The estimated maximum amount whereby the numerical value of a measured quantity can differ from the true value.

uncertainty principle The observation that high precision in the location of an electron is obtainable only at a sacrifice in the accuracy with which the momentum of the electron can be determined, and vice versa.

uncharged Without an electric charge, as opposed to discharged (depleted of a charge).

unclamp To switch off clamping action in a circuit. See CLAMPER and CLAMPING.

uncoated filament **1.** A plain filament (i.e., one without a coating of electron-increasing material). **2.** A stripped filament (i.e., one from which the electron-increasing coating has been burned off).

uncompensated Not modified to produce a desired type of performance, such as increased bandwidth or reduced temperature sensitivity. Example: an uncompensated amplifier put into service in the (wide) video band. Compare COMPENSATED AMPLIFIER, COMPENSATED CAPACITOR, and COMPENSATED DIODE DETECTOR.

uncompensated amplifier An amplifier without a provision for the automatic or manual modification or correction of its response. Compare COMPENSATED AMPLIFIER.

uncompensated capacitor A fixed or variable capacitor without a provision for the automatic or manual correction or modification of its capacitance or range or the improvement of its temperature coefficient. Compare COMPENSATED CAPACITOR.

uncompensated inductor A fixed or variable inductor without a provision for the automatic or manual correction or modification of its inductance or range or the improvement of its temperature coefficient.

uncompensated resistor A fixed or variable resistor without a provision for the automatic or manual correction or modification of its resistance or range or the improvement of its temperature coefficient.

unconditional branch See UNCONDITIONAL JUMP.

unconditional jump In computer operations, a program instruction that interrupts, without a relational test being made, the normal, ordered sequence of instructions and gives the address of (usually) the first instruction in a subroutine [e.g., BASIC's GOTO (line number)]. Also called *unconditional branch.* Compare CONDITIONAL JUMP.

unconditional stability Stability of a system at all frequencies and ages. Also called *absolute stability.*

unconditional transfer See UNCONDITIONAL JUMP.

uncontrolled multivibrator A multivibrator that is not synchronized with a signal source. Also called ASTABLE MULTIVIBRATOR and FREE-RUNNING MULTIVIBRATOR.

uncontrolled oscillator See SELF-EXCITED OSCILLATOR.

uncorrected power In an alternating-current circuit, a power value calculated without regard for power factor (i.e., the volt-ampere product).

uncorrected time Local standard time that has not been corrected in terms of the distance of the locality from the nearest standard-time meridian.

uncorrected tube A television picture tube operated without antipincushioning magnets.

uncoupled mode A mode of vibration existing in a system concurrent with (but independent of) other modes.

undamped galvanometer A galvanometer for which no provision has been made to limit overswing or prevent oscillation.

undamped meter **1.** A meter for which no provision has been made to limit overswing or prevent oscillation. **2.** A meter that is unprotected by a short circuit for limiting vibratory movement of the pointer during transport.

undamped natural frequency In the absence of damping, the oscillation frequency of a system having one degree of freedom when a transient force displaces the system momentarily from its position of rest.

undamped oscillation Continuous-wave oscillation. Also see CONTINUOUS WAVE.

undamped output circuit An audio power-amplifier output circuit that has not been designed so that overswing of the loudspeaker cone is prevented.

undamped speaker enclosure A loudspeaker cabinet for which provision has not been made to deaden resonances and other undesired vibration.

undamped wave See CONTINUOUS WAVE.

undeflected beam A cathode-ray beam or a light beam in its normal position of rest (quiescence).

underbiased unit A component, such as a transistor or vacuum tube, whose bias voltage or current is lower than the prescribed value. Compare OVERBIASED UNIT.

underbunching Less-than-optimum bunching of electrons in a velocity-modulated tube because of lowered buncher voltage.

undercompounded generator A generator in which the output voltage varies inversely with load resistance.

undercoupling Loose coupling, usually of an amount insufficient for optimum signal transfer.

undercurrent A current of lower than specified strength. Compare OVERCURRENT.

undercurrent relay A relay that is actuated when coil current falls below a predetermined level.

undercut The removal of the metal under the edge of the resist in a printed circuit by the etchant; thus the cross section of the conductor is reduced.

undercutting A phonograph-disc groove that is cut too shallow or with reduced internal movement of the recording stylus.

underdamping Insufficient damping of a system (i.e., not enough to prevent output oscillation following application of a transient force).

underdriven unit An amplifier, oscillator, or transducer whose driving signal (current, voltage, power, or other quantity) is lower than the prescribed value. Compare OVERDRIVEN UNIT.

underexcited Receiving lower-than-normal excitation, as in an underdriven final amplifier of a radio transmitter. Compare OVEREXCITED.

underflow In computer operations, the condition in which a quantity entered into storage is shorter than the space provided for it (e.g., a 12-digit quantity in a 16-position register).

underground antenna A transmitting or receiving antenna installed and operated below ground. Included are buried antennas and antennas on equipment operated in tunnels, cellars, and similar areas.

underground cable A cable that is buried in the earth.

underground communication Communication between a transmitter and receiver, both below the surface of the earth.

underground image The below-ground mirror image of an antenna; it combines with the actual antenna to form the complete radiation pattern.

underground line A power line laid below the surface of the earth.

underground receiver **1.** A receiver situated completely below the surface of the earth. **2.** A clandestine receiver.

underground reception Reception of an above-ground station's transmissions by an underground receiver. Compare UNDERGROUND COMMUNICATION.

underground transmitter **1.** A transmitter situated completely below the surface of the earth. **2.** A clandestine transmitter.

under insulation The insulation (usually a strip of tape) laid under a wire brought up from the center of a coil.

underinsulation Inadequate or insufficient insulation.

underlap In a facsimile or television picture, a crowding of the scanned lines.

underload circuit breaker See CIRCUIT BREAKER.

underloaded amplifier **1.** An amplifier whose load resistance (impedance) is less than the prescribed value. **2.** A power amplifier delivering less than its rated output power. Compare OVERLOADED AMPLIFIER.

underload relay A relay actuated when circuit current drops below a predetermined value. Compare OVERLOAD RELAY.

undermodulation Incomplete modulation of a carrier wave. Compare COMPLETE MODULATION and OVERMODULATION.

underpass The crossing of two conductors on a semiconductor wafer, without an electrical connection.

underpower relay A relay actuated when power drops below a predetermined level. Compare OVERPOWER RELAY.

underrate To assign a rating (e.g., current or power) lower than the quantity of the rating an equipment can handle, or tolerate. For safety or reliability, apparatus sometimes is deliberately underrated—especially in power output and maximum current or voltage.

undershoot On an oscilloscope screen or graph, a momentary swing of a current or voltage below the reference axis. Compare OVERSHOOT.

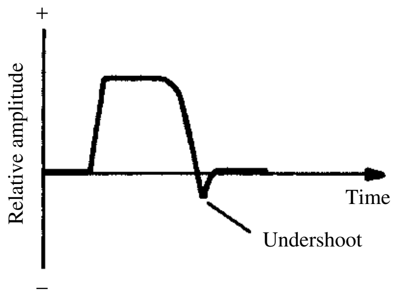

undershoot

undershoot distortion Distortion caused by reduction of the maximum amplitude of a signal wavefront below the steady-state amplitude that would be reached by a prolonged signal wave.

understudy See BACKUP.

underthrow A form of signal distortion that occurs when the modulating-waveform frequency is too high in proportion to the frequency of the wave itself.

undervoltage A voltage of lower than specified value. Compare OVERVOLTAGE.

undervoltage protection The automatic disconnection of a load device from its driving source when the driving voltage falls below a predetermined (threshold) level. This action is sometimes accomplished with a Zener diode (whose breakdown voltage is equal to the threshold voltage) in series with a disconnect relay.

undervoltage relay A relay actuated when voltage drops below a predetermined level. Compare OVERVOLTAGE RELAY.

underwater antenna A transmitting or receiving antenna normally operated in a body of water. It is usually associated with a submarine.

underwater sound projector A hydroacoustic transducer that converts audio-frequency power into sound waves, which it radiates through a body of water to a receiver.

Underwriters Code The advisory National Electrical Code adopted by the National Fire Protection Association; it is often enforced by courts of law and inspection agencies.

Underwriters Laboratories Abbreviation, UL. A private corporation that issues standards of safety for electrical components and equipment, and for their operation.

undistorted power output For active amplifier devices, a specified maximum audio power output level at which the total distortion does not exceed a specified low value (i.e., at which operation is practically distortionless).

undistorted wave **1.** A sine wave that contains essentially no harmonic energy. **2.** A nonsinusoidal wave whose shape corresponds exactly to the equation for the wave.

undisturbed-one output In digital-memory operations, the one-output of a magnetic cell or other memory unit that has received a full, rather than partial, read pulse. Compare UNDISTURBED OUTPUT SIGNAL.

undisturbed output signal In digital-memory operations, the output of a magnetic cell or other memory unit previously set to one or zero, and that has received a full, rather than partial, read pulse. Also called *undisturbed response voltage.*

undisturbed response voltage See UNDISTURBED OUTPUT SIGNAL.

undisturbed-zero output In digital-memory operations, the zero-output of a magnetic cell or other memory unit that has received a full, rather than partial, read pulse. Compare UNDISTURBED OUTPUT SIGNAL.

Universal Truth Machine (UTM) A hypothetical computer capable of proving any logically true statement. According to the Incompleteness Theorem proved in 1930, such a machine cannot exist for a first-order logical system.

undoped Pertaining to a pure semiconductor material (i.e., one containing no lattice-altering additives).

undulating current A current, such as an alternating or composite current, whose value oscillates in the manner of a wave.

undulation A wavelike alternation.

undulatory theory The theory that light is transmitted by means of an undulating (wavelike) movement between the luminous object and the eye.

unequal alternation In an alternating-current waveform, a form of asymmetry in which the positive and negative half-cycles are not exact mirror images; they might vary in peak amplitude, duration, shape, etc.

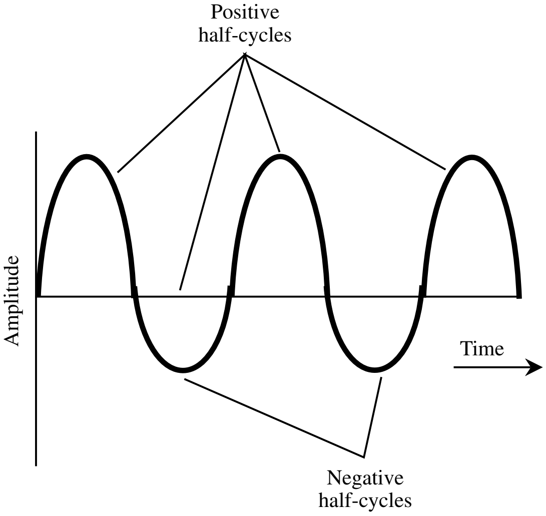

unequal alternation

unexpected halt During a computer program run, an undirected (unplanned) halt (e.g., one caused by a machine fault or program bug).

unfiltered Not having been subjected to filtering action. An example is the pulsating-direct-current output of an alternating-current-operated rectifier circuit.

unfocused light source In a photoelectric system, a light source that delivers diffused light.

unformed rectifier A newly fabricated semiconductor rectifier (especially selenium) or an electrolytic rectifier before it has been electroformed for improved characteristics.

unfurlable antenna An antenna that can be unrolled to increase its length, thereby lowering its fundamental resonant frequency; it can also be rolled up to decrease its length and increase its fundamental resonant frequency.

ungrounded item A component, circuit, or circuit point having no connection to ground, except an inadvertent one through common impedances or leakage paths.

ungrounded system A system operated entirely above ground, any path to ground being accidental.

unguided Without electronic guidance (pertaining to missiles, rockets, satellites, etc.).

uni- A prefix (combining form) meaning *one* or *single* and appearing in a number of electronic terms. See the following several definitions for examples.

uniaxial **1.** Having one axis. **2.** Referred to one axis.

uniaxial crystal A crystal having one optical axis.

unibivalent electrolyte An electrolyte, such as sodium carbonate (Na_2CO_3), that dissociates into two univalent ions and one bivalent ion.

Unicode Trade name for a 16- or 32-bit binary digital code that allows for the representation of characters and glyphs in all the written languages of the world. The 16-bit code contains 65,536 code points; the 32-bit version contains more than 1 million code points.

uniconductor cable A cable having a single conductor, usually of braided or twisted wires.

uniconductor waveguide A waveguide consisting of a metal sheath deposited on a solid dielectric of cylindrical or rectangular cross section.

unidentified flying object Abbreviation, UFO. An object claimed to have been seen in flight, but not identified by reliable authorities as any known type of vehicle.

unidirect To commutate (see COMMUTATOR).

unidirectional **1.** Flowing or acting in one direction only. **2.** Having a radiation or response sensitivity that is maximum in primarily one direction in space.

unidirectional antenna See UNIDIRECTIONAL ARRAY.

unidirectional array A beam antenna that radiates in one direction only, or principally in one direction, unless rotated.

unidirectional conductivity Electrical conductivity in only one direction; it characterizes an ideal diode.

unidirectional coupler A directional coupler sampling only one direction of transmission.

unidirectional current A current that flows always in the same direction, although it might fluctuate in intensity.

unidirectional elements Circuit elements, such as diodes or transistors, that transmit energy in only (or best in) one direction.

unidirectional field-effect transistor See UNILATERAL FIELD-EFFECT TRANSISTOR.

unidirectional hydrophone An underwater unidirectional microphone.

unidirectional loudspeaker A loudspeaker that radiates sound substantially in one direction.

unidirectional microphone A microphone that receives sound waves from one direction, usually from the front, minimum response usually being from the sides and back.

unidirectional network A network that transmits signals in only one direction (i.e., the input and output terminals are not interchangeable).

unidirectional pattern For a transducer (such as an antenna, speaker, or microphone), a radiation or response pattern that shows a pronounced maximum in one direction only.

unidirectional pulse A single-polarity pulse.

unidirectional pulse train A series of unidirectional pulses.

unidirectional response For a receiving transducer, such as an antenna or microphone, a response pattern that shows a pronounced maximum in one direction only.

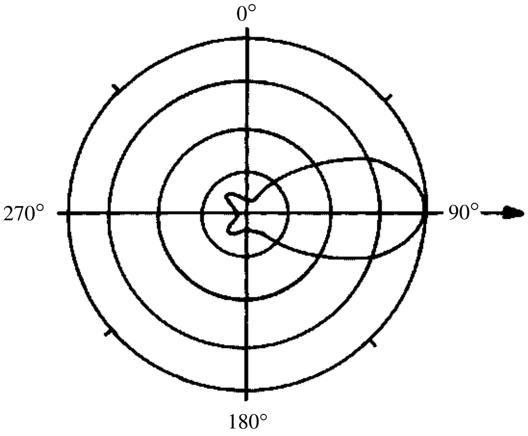

unidirectional pattern

unidirectional speaker See UNIDIRECTIONAL LOUDSPEAKER.

unidirectional transducer A transducer that operates in one direction only (i.e., its input and output cannot be interchanged).

unidirectional voltage A voltage that never changes polarity, although it might fluctuate in value.

UNIFET Abbreviation of UNIPOLAR FIELD-EFFECT TRANSISTOR. (Also, UFET.)

unified-field theory A theory, as yet unproved, that all forces in nature are interrelated and are controlled by the same causes and factors. Albert Einstein worked on this theory after he successfully developed the theories of relativity.

unifilar Having (or wound as) one fiber, wire, filament, or thread.

unifilar magnetometer A magnetometer in which a bar is suspended at its center of gravity by a single thread.

unifilar suspension A method of meter construction in which the moving part is suspended by a single filament. The torque might be provided by this filament or by a spring or other resisting device; the greater the displacement, the greater the torque.

uniform circular motion Motion at a uniform rate and describing a circle (e.g., a motor armature rotating at a constant speed).

uniform electric field An electric field in which all the lines of flux are straight and parallel, and in which the electrostatic force has the same value at all points (e.g., the field between two oppositely charged, flat, parallel plates).

uniform field See UNIFORM ELECTRIC FIELD and UNIFORM MAGNETIC FIELD.

uniform frequency response Frequency response that is flat throughout a specified range. Such response is characterized by the transmission of a signal with no introduced amplitude or phase distortion.

uniform line A transmission line having identical electrical properties over its entire length.

uniform magnetic field A magnetic field in which all the lines of flux are straight and parallel, and in which the magnetic force has the same value at all points.

uniform plane wave A free-space plane wave at an infinite distance from the generator, having constant-amplitude electric and magnetic field vectors over the equiphase surfaces.

uniform precession In regions of the uniform magnetic field of a sample of material, the state in which the magnetic moments of all atoms are parallel and precess in phase around the magnetic field.

uniform waveguide A waveguide having constant electrical and physical characteristics along its axis.

uniground **1.** The grounding of a circuit at one point to reduce susceptibility to hum and noise. **2.** The point at which such a connection is made.

unijunction transistor A semiconductor device consisting of a thin silicon bar on which a single pn junction acting as an emitter is formed near one end. Two bases are provided—each an ohmic connection made to one end of the bar. Also called *double-base diode.*

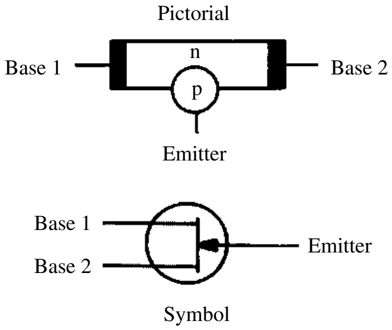

unijunction transistor

unilateral area track A film sound track having modulation on only one edge of the opaque area.

unilateral bearing In radio direction finding, a bearing obtained with a finder that produces unilateral response and thereby eliminates 180-degree error.

unilateral conductivity See UNIDIRECTIONAL CONDUCTIVITY.

unilateral element See UNIDIRECTIONAL ELEMENTS.

unilateral field-effect transistor A field-effect transistor whose source and drain terminals cannot be interchanged. Also called ASYMMETRICAL FIELD-EFFECT TRANSISTOR. Compare SYMMETRICAL FIELD-EFFECT TRANSISTOR.

unilateralization A system of neutralization used in transistor amplifiers at very-high and ultra-high frequencies, in which the internal capacitive and resistive components are compensated by the neutralization feedback.

unilateralized amplifier A transistor amplifier in which both the internal resistive and capacitive components are compensated by the neutralizing circuit. Also see UNILATERALIZATION.

unilateral network See UNIDIRECTIONAL NETWORK.

unilateral transducer See UNIDIRECTIONAL TRANSDUCER.

uninterruptible power supply Abbreviation, UPS. A device that provides a continuous utility current to electronic devices in the event of a utility power dip or blackout. It contains a battery and a power inverter. These devices are used extensively in personal computing to prevent loss of data that could otherwise result from a utility power failure.

union **1.** The logical inclusive-OR operation. **2.** See BOND, **1**.

union catalog In computer operations, the compiled list of the contents of two or more tape or disk libraries.

unionic material A material having no ions (i.e., one in which all atoms are neutral).

uniphase antenna See COLLINEAR ANTENNA.

UNIPOL Acronym for *universal problem-oriented language,* a high-level computer-programming language.

unipolar **1.** Having or using a single pole or polarity. **2.** Operating with one class of current carrier.

unipolar armature An electric-motor armature that maintains its polarity throughout a complete revolution.

unipolar field-effect transistor See UNIPOLAR TRANSISTOR.

unipolar induction Induction by only one pole of a magnet.

unipolar input The input circuit of an instrument or device designed for input signals of one polarity only.

unipolar pulse A pulse in which the current flows in only one direction, or in which the voltage occurs with only one polarity.

unipolar transistor A field-effect transistor (FET). It utilizes only one kind of carrier (electrons in the n-channel FET, holes in the p-channel FET). Compare BIPOLAR TRANSISTOR.

unipolar winding The winding of a UNIPOLAR ARMATURE.

unipole **1.** An all-pass filter having one pole and for which there is one zero. Also see ALL-PASS FILTER, POLES OF IMPEDANCE, and ZEROS OF IMPEDANCE. **2.** A hypothetical, omnidirectional antenna.

unipotential cathode An indirectly heated tube cathode. Also called *equipotential cathode.* See INDIRECTLY HEATED CATHODE.

unit **1.** A named, single magnitude adopted as a standard by physical measurement. Thus, the unit of current is the ampere, the unit of frequency is the hertz, the unit of capacitance is the farad, etc. Also see ABSOLUTE SYSTEM OF UNITS, INTERNATIONAL SYSTEM OF UNITS, and CENTIMETER-GRAM-SECOND. **2.** A single piece or assemblage of equipment, such as amplifier, converter, power supply, etc. **3.** A quantity of 1 implied when unit is the adjective describing a quantity (e.g., *unit length* means a distance of 1 meter, 1 foot, etc.).

unitary code In computer operations, a code based on one digit. The number of times the digit is repeated indicates a given number.

unit cell In crystals, the simplest polyhedron exhibiting all the structural features and which is repeated to form the crystal lattice.

unit charge See UNIT ELECTROSTATIC CHARGE.

unit electric charge See UNIT ELECTROSTATIC CHARGE.

unit electrostatic charge An electrostatic point charge that will attract or repel a point charge of equal strength 1 centimeter away in a vacuum, with a force of 1 dyne (10^{-5} newton).

unit function Symbol, *H* or *1.* A time-dependent quantity that is zero before the start of a period (when time *t* is zero) and 1 for all values of *t* greater than zero. It is approximated by a square wave, and is useful in solving problems involving transients.

unit length **1.** A fundamental unit of distance or time, used for reference in a measuring system. For example, in the mks (meter-kilogram-second) system, the distance unit length is the meter, and the time unit length is the second. **2.** The duration of a fundamental element, or bit, in a binary-code transmission system.

unitized construction The fabrication of an electronic equipment in subassemblies (such as modules) that can be tested separately, and that can be easily replaced (plugged in) in the event of individual failure.

unit line A line of electric or magnetic flux.

unit magnetic pole See UNIT POLE.

unit matrix A matrix whose main diagonal terms (all) are all at unity, the other terms being zero.

unitooth Pertaining to the use in electrical machinery of one slot per pole per phase.

unitor See OR GATE.

unit pole The strength (magnetic flux) of a (hypothetical) magnetic pole that will attract or repel a pole of equal strength 1 centimeter away in a vacuum, with a force of 1 dyne; 1 unit pole = 1.257×10^{-7} weber.

unit-ramp function A function whose value is zero before time *t*, and becomes equal to the time measured from *t* at all other instants. The integral of the UNIT FUNCTION.

unit record In computer operations, a complete record consisting of many data elements and contained within one storage medium, such as a magnetic disk.

unitrivalent electrolyte An electrolyte, such as sodium phosphate (Na_3PO_4), that dissociates into three univalent ions and one trivalent ion.

units position In a numbering system, the rightmost position in a multidigit whole number; or, if the number contains a radix point, the digit immediately to the left of the radix point.

unit-step function See UNIT FUNCTION.

unitunnel diode A special form of semiconductor diode, used as an oscillator at ultra-high and microwave frequencies.

unit vector A vector that is 1 unit long in terms of the scale representing a factor of interest, and has the same direction as the vector of interest. For example, if *u* is a unit vector, then 5*u* represents a vector having the same direction and five times the magnitude of *u*.

unity **1.** The figure 1 implied. **2.** A ratio of 1:1. **3.** A gain of 0 dB.

unity coupling Tight coupling between two coils, the turns ratio often being 1 to 1 and the coils always being closely interwound.

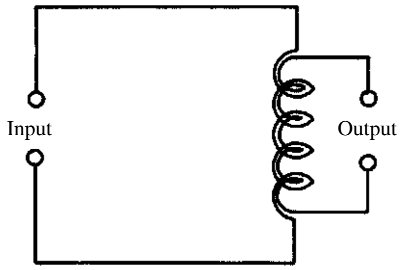

unity coupling

unity gain The condition in which the output amplitude is the same as the input amplitude; that is, a gain factor of 1, or 0 dB.

unity-gain bandwidth **1.** For an active filter, the bandwidth between the frequencies at which the gain is 0 dB. **2.** For an operational amplifier having a rolloff of 6 dB per octave, the frequency at which the open-loop gain is 0 dB.

uni-univalent electrolyte An electrolyte that dissociates into two univalent ions [e.g., sodium chloride (NaCl)].

univ Abbreviation of *universal.*

UNIVAC The name given to a family of digital computers and data-processing systems. The name is an acronym for *Universal Automatic Computer.*

univalent Having a valence of 1. For example, in the compound sodium chloride (NaCl), the sodium ion (Na+) and the chlorine ion (Cl–) are univalent.

universal bridge A bridge for the measurement of capacitance (*C*), inductance (*L*), and resistance (*R*). Such a bridge often is a skeletal foundation circuit with a provision for plugging L, C, or R standards in the various arms, as required. Also called *general-purpose bridge.*

universal coordinated time Abbreviation, UTC. Greenwich mean time coordinated by the International Time Bureau for broadcast as signals (on WWV, for example).

universal coupler **1.** An arrangement of one or more inductors and variable capacitors for matching a transmitter to virtually any antenna. One such device is the Collins coupler. **2.** A device for matching numerous generator output impedances to numerous load impedances.

universal filter An active filter that is continuously tunable over a wide frequency range and that offers low-pass, high-pass, bandpass, and band-suppression functions.

universal frequency counter A digital frequency meter usable at radio and audio frequencies.

universal motor A small series-wound motor that runs on direct current or single-phase alternating current. This type of motor is used in many household appliances and in motor-driven tools, such as portable electric drills, polishers, etc.

universal output transformer An output transformer having a number of taps on its primary and secondary windings for use with a wide variety of impedances.

universal product code Abbreviation, UPC. The variable-width-bar code appearing on price tags or product labels, and yielding such information as cost, size, and color when read by supermarket (or other retailer) optical character-recognition equipment. Also called *universal vendor marking (UVM).*

universal receiver A radio receiver that can be operated either from alternating current or direct current (i.e., from utility power or battery power).

universal shunt See AYRTON-MATHER GALVANOMETER SHUNT.

universal time See GREENWICH MEAN TIME.

universal transformer See UNIVERSAL OUTPUT TRANSFORMER.

universal transmitter A radio transmitter that can be operated either from alternating current or direct current (i.e., from utility power or battery power).

universal Turing machine In computer theory, a TURING MACHINE capable of simulating other Turing machines.

universal vendor marking See UNIVERSAL PRODUCT CODE.

universal winding A zigzag winding used to reduce the distributed capacitance of multilayer coils. Also called *lattice winding* and *honeycomb winding.*

universal-wound coil A coil using a universal winding. Such coils are common in intermediate-frequency (IF) transformers and in radio-frequency (RF) chokes.

univibrator See MONOSTABLE MULTIVIBRATOR.

unknown quantity Any variable quantity sought in calculations; it is usually represented by lowercase italic letters from the second half of the alphabet (e.g., *x*, *y*, *z*).

unlike charges Dissimilar electric charges (i.e., positive and negative).

unlike poles Dissimilar poles, such as the north and south poles of a magnet.

unload **1.** To remove data from a file. **2.** To disconnect the load from a circuit.

unloaded amplifier See UNTERMINATED AMPLIFIER.

unloaded antenna An antenna operated without added inductance or capacitance.

unloaded-applicator impedance In dielectric-heating operations, the unloaded impedance of applicator electrodes placed in their normal working position without the dielectric-material load in place.

unloaded battery **1.** A battery in the standby condition. **2.** A battery tested for open-circuit terminal voltage (i.e., with no load, except the voltmeter).

unloaded generator See UNTERMINATED GENERATOR.

unloaded line See UNTERMINATED LINE.

unloaded potentiometer A potentiometer or voltage divider with an open-circuited output.

unloaded *Q* The *Q* factor (degree of selectivity) of a coil or tuned circuit that is activated by a signal, but that delivers no output to a load.

unloaded transmitter A transmitter operated under open-circuit conditions (i.e., without an external load, such as an antenna or dummy resistor).

unmanned factory A manufacturing plant run largely by robots and computers. There must generally be at least one human being, whose job(s) is/are to oversee and maintain the operation of the machines.

unmatched elements **1.** Components (such as resistors, capacitors, semiconductors, etc.) having different values. **2.** Mating elements or devices not having the same impedance. Also called MISMATCH.

unmodulated carrier See CONTINUOUS WAVE.

unmodulated current A current whose characteristics (amplitude, phase, or frequency) are not varied by a signal or by noise.

unmodulated voltage A voltage whose characteristics (amplitude, phase, or frequency) are not varied by a signal or by noise.

unmodulated wave See CONTINUOUS WAVE.

unnilennium See MEITNERIUM.

unnilhexium See SEABORGIUM.

unniloctium See HASSIUM.

unnilpentium See DUBNIUM.

unnilquadium See RUTHERFORDIUM.

unnilseptium See BOHRIUM.

unode See MONODE.

unpack In computer operations, to remove data from a storage location into which, with other data, it has been packed (as a memory-conservation measure).

unpolarized light Light in which the wave orientation is random around the axis of the beam.

unpolarized plug A plug that can be inserted into a socket in two or more different ways, thus increasing the likelihood of a wrong connection. Compare POLARIZED PLUG.

unpolarized receptacle A receptacle into which a plug can be inserted in two or more different ways, thus increasing the likelihood of a wrong connection. Also called UNPOLARIZED SOCKET. Compare POLARIZED RECEPTACLE.

unpolarized relay A relay that always closes in the same direction, regardless of the direction of current in its coil. Also called *nonpolarized relay.*

unpolarized socket A socket into which a component or plug can be inserted in two or more different ways, thus increasing the likelihood of wrong connections. Also called UNPOLARIZED RECEPTACLE.

unprotected antenna An outside antenna operated without a lightning arrester or grounding switch. Increases the risk of damage to equipment, such as radio receivers and transmitters, connected to the antenna. It also increases the danger to personnel using the equipment.

unrationalized systems of units Collectively, the absolute centimeter-gram-second (cgs) electrostatic system of units, the absolute cgs electromagnetic system of units, and the absolute meter-kilogram-second (mks) system of units.

unreflected ray See UNREFLECTED WAVE.

unreflected wave **1.** See DIRECT WAVE. **2.** See GROUND WAVE. **3.** An electromagnetic wave that penetrates the ionosphere and passes into space.

unregulated Not held to a constant value. For example, an unregulated voltage is free to fluctuate in value.

unregulated power supply A power supply whose output current or voltage is not automatically held to a constant value. Compare CONSTANT-CURRENT SOURCE and CONSTANT-VOLTAGE SOURCE.

uns Abbreviation of UNSYMMETRICAL.

unsaturated core A magnetic core operated below the point of saturation (i.e., below the point at which an increase in coil current produces no further increase in magnetization of the core).

unsaturated logic A digital-logic scheme in which the switching devices operate between the saturated and cut-off conditions during either or both of the high and low states.

unsaturated operation Operation of a device at a point below saturation [i.e., below the point at which (1) an increase in voltage produces no further increase in current, or vice versa, or (2) an increase in coil current produces no further increase in magnetization of a core]. Compare SATURATED OPERATION.

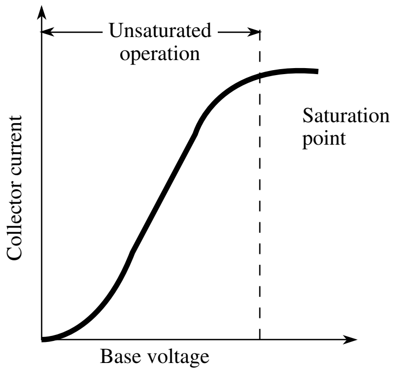

unsaturated operation

unsaturated standard cell See STANDARD CELL.

unshielded cable A cable, such as a twisted pair or a multiwire twist, lacking a shielding jacket. Unless special precautions are taken, such as transposing conductors, such a cable is susceptible to stray pickup and is capable of radiation.

unshielded choke An uncased choke (i.e., one without a protective and shielding metal housing).

unshielded coil An inductor without a field-confining enclosure.

unshielded probe An instrument probe that has no shielding enclosure. Such a probe is desirable for some tests, but it is subject to antenna-pickup effect, body capacitance, and stray-field pickup.

unshielded stage In electronic equipment, a stage operating entirely in the open (i.e., without electrostatic or magnetic shielding enclosures). It is, therefore, susceptible to stray pickup and is capable of radiation.

unshielded transformer **1.** An uncased transformer (i.e., one without a protective and shielding metal housing). **2.** A transformer having no electrostatic shield between its windings.

unshift The mechanical action in a teletypewriter when the carriage moves from the figures position to the letters position.

unshunted current meter A single-range D'Arsonval milliammeter or microammeter that has no shunt resistor. The resistance of the instrument is, therefore, equal to the resistance of the movable coil.

unsolder To separate wires, contacts, or fastenings by melting the solder that holds them together. Also called DESOLDER.

unsophisticated electronics Electronics theory and operations of the simplest kind, involving little or no mathematics. It is generally intended for a hobbyist audience. Compare SOPHISTICATED ELECTRONICS.

unstable oscillation **1.** Relaxation oscillation. **2.** The periodically interrupted oscillation characteristic of a blocking oscillation.

unstable region The negative-resistance portion of an N- or S-shaped response curve.

unstable servo A servo having unlimited output drift.

unstable state A condition, such as a negative-resistance region, that is difficult to maintain and often results in oscillation. The negative-resistance region of the tunnel diode forward-conduction curve is evidence of such a state. Compare STABLE STATE.

untapped winding A winding having two terminals, one at either end, and no intermediate terminals or connections. Also called *untapped coil.*

unterminated amplifier An amplifier operated without a load device (i.e., under open-circuit output conditions).

unterminated generator A generator operated without a load device (i.e., under open-circuit output conditions).

unterminated line A transmission line that is not terminated with an impedance (i.e., an open-ended line).

untriggered flip-flop A flip-flop that, at a particular instant, receives no actuating pulse, and that, therefore, does not change state.

untuned amplifier A radio-frequency amplifier that is not tuned to a single frequency, but has useful gain over a wide band of frequencies. Examples are the *distributed amplifier* and the *video amplifier.*

untuned filter See ALL-PASS FILTER.

untuned line An aperiodic transmission line (i.e., one that is only tuned to a particular frequency by its own distributed characteristics).

untuned transformer A transformer having simple primary and secondary windings and no tuning devices (such as capacitors in series or parallel with the windings), and designed so that its natural resonant frequency (because of distributed capacitance) lies outside the specified range of operation.

ununbium Symbol, Uub. Atomic number, 112. The most common isotope has atomic weight 277. Classified as a transition metal. It is human-made and not known to occur in nature.

ununhexium Symbol, Uuh. Atomic number, 116. The most common isotope has atomic weight 289. First reported in January 1999. It is a decomposition product of ununoctium, and it in turn decomposes into ununquadium. It is not known to occur in nature.

ununnilium Symbol, Uun. Atomic number, 110. The most common isotope has atomic weight 269. Classified as a transition metal. It is human-made and not known to occur in nature.

ununoctium Symbol, Uuo. Atomic number, 118. The most common isotope has atomic weight 293. It is the result of the fusion of krypton and lead, and decomposes into ununhexium. It is not known to occur in nature.

ununquadium Symbol, Uuq. Atomic number, 114. The most common isotope has atomic weight 285. First reported in January 1999. It is human-made, and not known to occur in nature.

unununium Symbol, Uuu. Atomic number, 111. The most common isotope has atomic weight 272. Classified as a transition metal. It is human-made, and not known to occur in nature.

unused time See UNATTENDED TIME.

unweighted average In the statistical analysis of data, an average calculated from terms not weighted by a factor. Compare WEIGHTED AVERAGE. Also see UNWEIGHTED TERM.

unweighted term In the statistical analysis of data, a term not operated upon by a weighting factor. Therefore, it has no extraordinary influence on a final result, such as an average. Compare WEIGHTED TERM.

unwind To eliminate all redundant or unnecessary operations in a computer system.

up-and-downer amplifier In broadcasting and sound amplification, an auxiliary amplifier that allows the level of sound effects and background music to be controlled automatically without disturbing the dialogue levels.

UPC Abbreviation of UNIVERSAL PRODUCT CODE.

upconv Abbreviation of UP CONVERTER.

up convert In superheterodyne conversion, to heterodyne a signal to an intermediate frequency higher than the signal frequency. Compare DOWN CONVERT.

update In data-processing operations, to make a file reflect the current status of pertinent information by using transactions for its modification.

up-down counter See BIDIRECTIONAL COUNTER.

uplink **1.** The signal sent up to a satellite transponder. **2.** The frequency of the signal sent up to a satellite transponder.

uplink frequency **1.** The specific frequency of a given signal sent up to a satellite transponder. **2.** The band of frequencies sent up to a satellite transponder. There can be many signals in an uplink; there can also be many different stations sending uplink signals within the band.

uplink power **1.** The power output of the transmitter sending the signal up to a satellite transponder. **2.** The effective radiated power of the station sending the uplink signal.

uplink station Any station that transmits a signal, or group of signals, up to a satellite transponder.

upper atmosphere See IONOSPHERE, STRATOSPHERE, and TROPOSPHERE.

upper sideband Abbreviation, USB. In an amplitude-modulated wave, the component whose frequency is the sum of the carrier frequency and the modulation frequency. Compare LOW-ER SIDEBAND.

upper sideband, suppressed carrier Abbreviation, USSC. A single-sideband transmission technique in which the upper sideband is transmitted, the lower sideband is suppressed, and the carrier is suppressed. Compare DOUBLE SIDEBAND, SUPPRESSED CARRIER, LOWER SIDEBAND, and SUPPRESSED CARRIER.

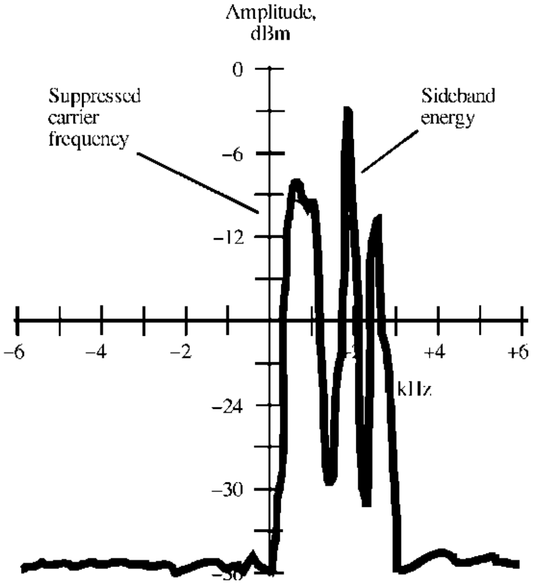

upper sideband

UPS Abbreviation for UNINTERRUPTIBLE POWER SUPPLY.

upside down **1.** In radioteletype, a signal in which the MARK and SPACE signals are reversed in frequency. **2.** In single sideband, reception of a signal on the wrong sideband, resulting in “monkey chatter.”

up time See SERVICEABLE TIME.

upturn A usually sudden rise in a performance curve. Compare DOWNTURN.

upward modulation Modulation in which the output current of the modulated circuit increases during modulation. Compare DOWNWARD MODULATION.

UR Abbreviation of ULTRARED.

uranium Symbol, U. A radioactive metallic element. Atomic number, 92. Atomic weight, 238.03.

uranium 235 Symbol, U-235. An isotope of uranium of mass number 235, used in atomic bombs and atomic power plants.

uranium 238 Symbol, U-238. An isotope of uranium of mass number 238, which, under fast-neutron bombardment, decays to become neptunium and then plutonium (of mass number 239).

uranium metals The group of heavy, radioactive elements of atomic numbers 89 through 103. Also called the *actinide series.*

uranium rays The ionizing radiation emitted by uranium. In 1896, Antoine Henri Becquerel observed that uranium salts emit rays (later identified as alpha, beta, and gamma) that can pass through bodies opaque to light and are capable of exposing a photographic plate. Also called *Becquerel rays.*

urea formaldehyde resin A thermosetting, synthetic resin used in the manufacture of a number of plastic dielectric bodies and formed by the reaction of urea with formaldehyde.

urea plastic Any of several thermosetting plastics having a urea base and used as a dielectric and as a molding material in electronics (e.g., *urea formaldehyde resin*).

URSI Abbreviation of *Union Radio Scientifique Internationale (International Radio Scientific Union).*

ursigram A broadcast message giving information on sunspots, radio propagation, terrestrial magnetism, and related subjects. Also see URSI.

usable frequency **1.** Any frequency at which communications can be maintained between two points via ionospheric propagation. **2.** Any frequency at which a communications system is operational.

usable sample **1.** The portion of an oscilloscope or monitor display that is visible on the screen. **2.** In statistics, a sample that is considered valid for calculation purposes.

USB Abbreviation of UPPER SIDEBAND.

user friendliness **1.** For electronic equipment and systems, the quality of being easy for people to operate. **2.** The relative ease with which a machine, especially a computer, can be operated by people.

useful life The elapsed time between the installation of an expendable component, circuit, or system, and the time it must be replaced.

useful line In television, the portion of a scanning line that can be used for picture signals. Also called AVAILABLE LINE.

user group A group of hobbyists or company representatives having in common the ownership and/or use of a specific brand of computer, and who meet or otherwise interact to share their expertise with, or programs, for, the machine. Also applicable to similar groups of programmable-calculator owners.

user's library A compilation of programs supplied to (by the vendor) or generated by a user group.

USM Abbreviation of *underwater-to-surface missile.*

USSC Abbreviation of UPPER SIDEBAND, SUPPRESSED CARRIER.

UTC Abbreviation of UNIVERSAL COORDINATED TIME.

utility **1.** An organization providing a public service, such as electric power or electronic communications. **2.** In computer operations, a program or set of programs intended for diagnosing and/or correcting hardware and software problems.

utility box A general-purpose aluminum or steel box, supplied in various convenient sizes (painted or unpainted), used as a housing or shield for electronic equipment.

utility factor For a transformer used in a direct-current (dc) power supply, the ratio of dc output to required kilovolt-ampere (kVA) capacity.

utilization factor See UTILITY FACTOR.

UTL Abbreviation of *unit transmission loss.*

UUM Abbreviation of *underwater-to-underwater missile.*

UV **1.** Abbreviation of ULTRAVIOLET. **2.** Abbreviation of UNDERVOLTAGE.

Uvicon A television camera tube in which a conventional vidicon scanning section is preceded by an ultraviolet-sensitive photocathode, an electron-accelerating section, and a special target.

Uviol lamp An ultraviolet lamp using uviol glass—a special glass that is transparent to ultraviolet rays.

UVM Abbreviation of UNIVERSAL VENDOR MARKING.

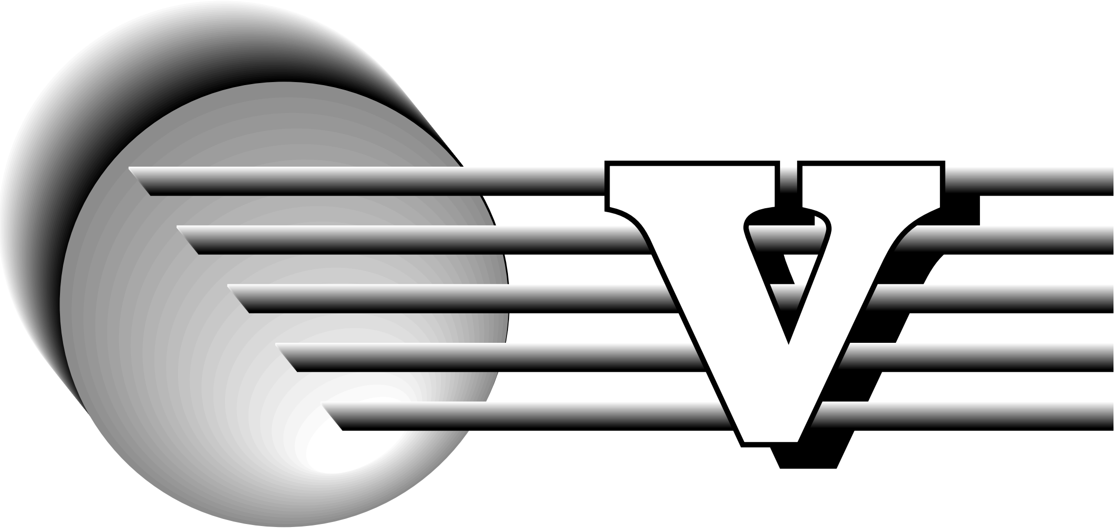

V **1.** Abbreviation of VOLT. **2.** Symbol (ital) for VOLTAGE or POTENTIAL. **3.** Symbol for VANADIUM. **4.** Symbol for VOLUME. **5.** Symbol (ital) for RELUCTIVITY. **6.** Abbreviation of VERTICAL. (Also, vert.)

v **1.** Abbreviation of VELOCITY. **2.** Abbreviation of VECTOR.

VA Abbreviation of VOLT-AMPERE.

V/A Abbreviation of *volts per ampere* (ohms).

VAC Abbreviation of *vector analog computer.*

vac Abbreviation of VACUUM.

V_{ac} Symbol for AC VOLTS. (Also, Vac.)

vacancy In a crystal lattice, a position not occupied by a nucleus.

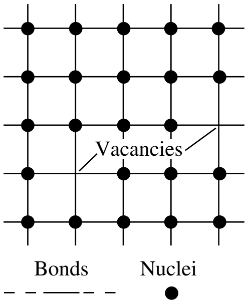

vacancy

vacuum A space from which all air and other gases have been removed to the greatest practicable extent. Some electronic parts, such as the elements of an electron tube, are housed in an evacuated space to protect them from the deterioration that would result from open-air operation.

vacuum capacitor A plate- or concentric-cylinder-type capacitor sealed in an evacuated glass tube or bulb. The vacuum acts as the dielectric and provides a dielectric constant of 1 and very high voltage breakdown.

vacuum deposition The electrical application of a layer of one material (such as a metal) to the surface of another (the substrate), carried out in a vacuum chamber (e.g., evaporation and sputtering). Also see DEPOSITION and EVAPORATION, **1**.

vacuum envelope The shell or tube of an electron tube, X-ray tube, or other electron device that requires a vacuum.

vacuum evaporation A method of manufacturing thin-film circuits by vaporizing a substance, and letting it accumulate or condense on a base.

vacuum gauge An instrument for measuring the vacuum in a device being evacuated. One of the several varieties of this gauge uses the elements of a triode, which are sealed in part of the vacuum system, some characteristic of the tube being continuously monitored as the vacuum pumping progresses. Another uses a thermistor sealed in part of the system; the resistance of the thermistor changes proportionately during the vacuum pumping.

vacuum impregnation The impregnation of a device (such as a capacitor, transformer, or choke coil) in a vacuum chamber. The process causes the pores in the device and its insulating materials to be completely filled with the impregnant.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

vacuum level **1.** The pressure, in millimeters of mercury (mm/Hg). Normal atmospheric pressure is 760 mm/Hg. A perfect vacuum would be represented by 0 mm/Hg. **2.** The proportion, or percentage, of normal atmospheric pressure in a given environment.

vacuum phototube A phototube enclosed in an evacuated envelope, as opposed to one that is gas-filled.

vacuum pump A pump for removing air and gases from electron tubes, X-ray tubes, lamp bulbs, etc. Such pumps include mechanical types, diffusion types, and combinations of these. Also see DIFFUSION PUMP.

vacuum range The range of a communications system if propagation takes place through a perfect-vacuum (theoretical) medium.

vacuum seal An airtight seal between adjoining parts in an evacuation system.

vacuum switch A switch that is enclosed in a vacuum bulb or tube to reduce contact sparking.

vacuum thermocouple A thermocouple enclosed in a vacuum bulb with a small heater element. Radio-frequency current passed through the heater raises its temperature and causes the thermocouple to generate a proportional direct-current voltage.

vacuum tube An electron tube from which virtually all air and gases are removed. Also see ELECTRON TUBE.

vacuum-tube amplifier An amplifier using one or more vacuum tubes, rather than semiconductor devices.

vacuum-tube bridge A special bridge for determining vacuum-tube characteristics by null methods.

vacuum-tube characteristics The operating parameters of a vacuum tube, such as plate current, grid voltage, input resistance, interelectrode capacitances, amplification factor, transconductance, etc., that describe tube performance.

vacuum-tube coefficients See VACUUM-TUBE CHARACTERISTICS.

vacuum-tube current meter An ammeter, milliammeter, microammeter, or nanoammeter embodying an amplifier that uses one or more vacuum tubes. Also see ELECTRONIC CURRENT METER.

vacuum-tube modulator A circuit using one or more vacuum tubes to impress a modulating signal on a carrier.

vacuum-tube rectifier **1.** An alternating-to-direct-current converter circuit using one or more vacuum tubes, rather than gas tubes or semiconductor devices. **2.** A rectifier tube.

vacuum-tube sweep A sweep oscillator using a vacuum tube, rather than a gas tube. Also called HARD-TUBE SWEEP.

vacuum-tube voltmeter Abbreviation, vtvm. A voltmeter using a tube-type amplifier. Also see ELECTRONIC VOLTMETER.

val Abbreviation of VALUE.

valence A unit showing the degree to which elements or radicals (replaceable atoms or groups of atoms) will combine to form compounds.

valence band In the energy diagram for a semiconductor, the band of lowest energy. This band lies below the forbidden band (energy gap), which is below the conduction band. Also see ENERGY BAND DIAGRAM.

valence bond In a semiconductor material, an interatomic path over which shared electrons travel.

valence electrons Electrons in the outermost orbits of an atom. These electrons determine the chemical and physical properties of a material. Also see FREE ELECTRON.

valence shells See ELECTRON SHELLS.

validate To check data for correctness.

validity **1.** Correctness or accuracy of data. **2.** The logical truth of a derivation or statement, based on a given set of propositions.

valley A dip between adjacent peaks in a curve or wave.

valley current In a tunnel diode, the current at the valley point.

valley point The lowest point of finite current on the current-voltage response curve of a tunnel diode. Immediately before this point, current decreases with increasing applied voltage (an indication of negative resistance). Beyond this point, however, the current again increases with increasing voltage. Compare PEAK POINT.

valley voltage In a tunnel diode, the voltage at the valley point.

value **1.** The level or magnitude of a quantity (e.g., voltage value). **2.** The worth of a system, procedure, device, etc., in terms of goal fulfillment or other criterion.

valve Variation (Brit.) of ELECTRON TUBE. The term was applied to the first tube, a diode (the Fleming valve), and is descriptive of the action of a tube (a controller of electric current), rather than its appearance.

vanadium Symbol, V. A metallic element. Atomic number, 23. Atomic weight, 50.94.

Van Allen radiation belts Two high-altitude zones that surround the earth and consist of high-energy subatomic particles. These belts were once thought to preclude space travel by humans beyond the immediate vicinity of the earth.

Van de Graaff generator See BELT GENERATOR.

vane-anode magnetron A magnetron having plane parallel walls between adjacent cavities.

vane attenuator A waveguide attenuator consisting essentially of a slab (vane) of resistive material that slides laterally through the waveguide.

vane instrument See IRON-VANE METER.

vane-type magnetron See VANE-ANODE MAGNETRON.

vane-type meter See IRON-VANE METER.

V antenna See VEE ANTENNA.

vapor lamp A discharge lamp consisting essentially of a glass tube filled with a small amount of gas under low pressure and some element, such as mercury or sodium. A high voltage is applied between electrodes sealed into each end of the tube. The voltage causes the element to vaporize. This, in turn, ionizes the gas, causing it to glow.

vapor pressure In a confined medium, the pressure of a gas, measured in atmospheres, pounds per square inch, or millimeters of mercury.

VAR Abbreviation of VOLT-AMPERES REACTIVE.

var Abbreviation of VARIABLE.

varactor A semiconductor-type voltage-variable capacitor. Sometimes called a *varactor diode.*

varactor amplifier A dielectric amplifier using a varactor as the voltage-variable capacitor.

varactor flip-flop A bistable multivibrator based on the nonlinear performance of one or two varactors.

varactor frequency multiplier A frequency multiplier (doubler, tripler, etc.) in which multiples of a fundamental frequency result from the nonlinear action of a varactor.

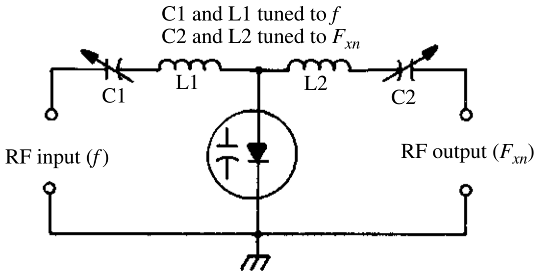

varactor frequency multiplier

varactor tuning A method of tuning a circuit or adjusting the frequency of an oscillator, using an inductor and a varactor to obtain the desired frequency. The varactor acts as a variable capacitor.

VAR-hour Abbreviation, VARh. Short for *volt-ampere-reactive-hour*, a unit of reactive energy; 1 VARh can be represented by 1 VAR for 1 hour, 2 VAR for 0.5 hour, 0.5 VAR for 2 hours, etc.

VAR-hour meter An instrument that measures reactive energy.

variable Abbreviation, var. A quantity whose value changes at some stated or calculable rate, and is given names in expressions or equations, such as *x*, *y*, or *z*. Compare CONSTANT. Also see DEPENDENT VARIABLE and INDEPENDENT VARIABLE.

variable-area sound track See VARIABLE-WIDTH SOUND RECORD.

variable block In computer operations, a unit of data, such as a group of records, whose size is dependent on data requirements (i.e., it is not fixed).

variable-capacitance diode See VARACTOR.

variable-capacitance transducer See CAPACITIVE TRANSDUCER.

variable capacitor A capacitor that can be adjusted from a low value (practically zero) to some maximum value. A step-type unit contains a number of fixed capacitors that can be switched in parallel with each other until, at the last step, all are in parallel. A continuously variable unit has a provision for moving one plate or set of plates, relative to another plate or set of plates; or one plate might be moved, with respect to another, so that the distance between them is changed. In a voltage-variable capacitor (such as a varactor), capacitance varies in accordance with an applied direct-current voltage.

variable-carrier modulation See QUIESCENT-CARRIER OPERATION.

variable connector In a flowchart, a path leading from a box in which a decision is made, to a number of other points on the diagram.

variable coupling Coupling that is adjustable. In an inductively coupled circuit, the distance or angle between the coils is usually the adjustable factor. In a capacitively coupled circuit, a variable capacitor is generally used.

variable-density sound record In photographic sound-on-film recording, a sound track made by varying, in sympathy with the sound frequency, the amount of light reaching the film. Compare VARIABLE-WIDTH SOUND RECORD.

variable-depth sonar A sonar device that can be placed far below the surface, for the purpose of detecting objects or terrain at greater depths than would be possible with a surface-located sonar device.

variable-efficiency modulation See EFFICIENCY MODULATION.

variable-erase recording In magnetic-tape operations, recording a signal by selectively erasing a previously recorded signal.

variable field **1.** Any field with an intensity that changes with time. An electromagnetic field is a common example. **2.** In a computer record, a field that is a variable block with a terminal symbol at the end.

variable-frequency oscillator Abbreviation, VFO. An oscillator (usually self-excited) whose frequency is continuously variable.

variable-inductance pickup A phonograph pickup in which vibration of the stylus causes the inductance of a small coil to vary in sympathy with the sound frequency.

variable-inductance transducer A transducer in which a monitored quantity causes the inductance of a coil to vary proportionately. The coil thereby offers a varying impedance to an alternating-current supply voltage.

variable inductor An inductor whose value can be adjusted from zero to a certain maximum. The variability might be in steps, provided by taps on the inductor, or a sliding contact can be used.

Several different inductors can be connected in series and parallel combinations, with switches to facilitate variability. One type of variable inductor uses a roller contact to allow continuous adjustment. See ROLLER INDUCTOR. Another method of varying inductance is to change the permeability of the core, such as by moving a rod of ferrite or powdered iron in and out of the coil. See VARICOUPLER and VARIOMETER.

variable-pitch indication An audible signal used in lieu of, or in conjunction with, a meter to indicate voltage, current, logic state, etc. The higher the value of the measured quantity, the higher the pitch.

variable-reluctance microphone A microphone in which the impinging sound waves cause corresponding variations in the reluctance of an internal magnetic circuit.

variable-reluctance pickup A phonograph pickup in which the stylus causes an armature to vibrate in a magnetic field, and consequently the reluctance of the magnetic circuit varies in sympathy with the audio frequency.

variable-reluctance transducer A transducer in which the monitored quantity causes the reluctance of an internal magnetic circuit to vary proportionately.

variable-resistance pickup A phonograph pickup in which the vibration of the stylus causes the resistance of an internal resistive element to vary in sympathy with the sound frequency.

variable-resistance transducer A transducer in which a monitored quantity causes the resistance of an internal resistive element to vary proportionately.

variable-resistance tuning Tuning of a selective resistance-capacitance (RC) circuit, such as a Wien bridge or parallel-tee network, by varying one or more of its resistance arms.

variable resistor A resistor whose value can be varied either continuously or in steps. Also see POTENTIOMETER and RHEOSTAT.

variable selectivity In a circuit or device, selectivity that is adjustable. Example: a variable-selectivity intermediate-frequency (IF) amplifier.

variable-speed motor Any motor whose speed is adjustable, smoothly or in steps, under load.

variable transformer A transformer (often an autotransformer) whose output voltage is adjustable from zero (or some minimum value) to maximum. For this purpose, one winding (usually the secondary in a two-winding transformer) can have a number of taps. Smooth variation can be provided by a wiper arm that slides over the turns of the winding.

variable-width wound record In photographic sound-on-film recording, a sound track made by varying the width of the track in sympathy with the sound variations. Compare VARIABLE-DENSITY SOUND RECORD.

Variac See VARIABLE TRANSFORMER.

Varicap A voltage-variable capacitor of the semiconductor-diode type. Also see VOLTAGE-VARIABLE CAPACITOR.

varicoupler An adjustable radio-frequency transformer consisting of a primary coil (usually the rotor) and a secondary coil (usually the stator), the former being rotatable to vary the coupling between the coils. The inductance of the stator is varied via a set of taps.

varindor A coil having a special core and whose inductance varies with the amount of direct current flowing through the winding. Also see SATURABLE REACTOR.

variocoupler A radio-frequency transformer in which one winding is rotatable, for the purpose of adjusting the mutual inductance between the primary and secondary. It is used in certain antenna-coupling applications.

variolosser A variable attenuator.

variometer A continuously variable inductor consisting of two coils connected in series or parallel and mounted concentrically—one rotating inside the other. Inductance is maximum when the coils are perpendicular to each other, and minimum when they are parallel.

varioplex A time-sharing method of transmitting and receiving wire telegraph signals. It allows the optimum usage of available lines.

varistor See VOLTAGE-DEPENDENT RESISTOR.

Varley-loop bridge A four-arm, direct-current bridge circuit in which one of the arms, a two-wire line, is accidentally grounded at a distant point. By adjustment of the bridge, the resulting resistance indicates the distance to the fault.

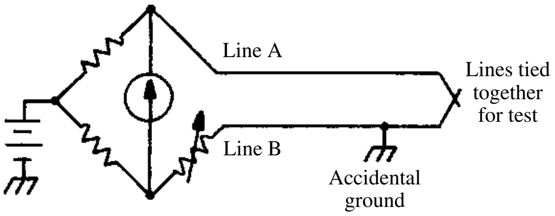

Varley loop bridge

VAR meter An instrument for measuring the apparent (VAR) power in a reactive circuit.

varnished cambric Cotton or linen fabric in sheets, tape or tubes, that has been varnish-impregnated and baked. A common form of this insulating material, the slender tubing known as *spaghetti*, has largely been replaced by plastic tubing.

varnished-cambric tubing Slender tubes of varnished cambric slipped over bare wires and busbars to insulate them. Also called *spaghetti.*

VATE Abbreviation of VERSATILE AUTOMATIC TEST EQUIPMENT.

V_B Symbol for BASE VOLTAGE of a bipolar transistor.

V_{BB} Symbol for *base-voltage supply* for a bipolar transistor.

V beam system See VEE BEAM SYSTEM.

V_C Symbol for COLLECTOR VOLTAGE of a bipolar transistor.

VCA Abbreviation of VOLTAGE-CONTROLLED AMPLIFIER.

V_{CC} Symbol for *collector-voltage supply* of a bipolar transistor.

VCCO Abbreviation of VOLTAGE-CONTROLLED CRYSTAL OSCILLATOR. (Also, lowercase.)

VCD Abbreviation of VARIABLE-CAPACITANCE DIODE.

VCG Abbreviation of VOLTAGE-CONTROLLED GENERATOR.

VCO Abbreviation of VOLTAGE-CONTROLLED OSCILLATOR.

V connection of transformers See VEE CONNECTION OF TRANSFORMERS.

VCR Abbreviation of VIDEO CASSETTE RECORDER.

VCSR Abbreviation of *voltage-controlled shift register.*

V-cut crystal See VEE-CUT CRYSTAL.

VCXO Abbreviation of VOLTAGE-CONTROLLED CRYSTAL OSCILLATOR. (Also, lowercase.)

VD **1.** Abbreviation of VOLTAGE DROP. **2.** Abbreviation of *vapor density.*

V_D Symbol for drain voltage of a field-effect transistor.

Vdc Symbol for DC VOLTS.

VDU Abbreviation of VISUAL DISPLAY UNIT.

vdcw Symbol for DC WORKING VOLTAGE. (Also, dcwv.)

VDIG Abbreviation of VIDEO DIGITIZER.

VDR **1.** Abbreviation of VOLTAGE-DEPENDENT RESISTOR. **2.** Abbreviation of VIDEODISC RECORDER.

V_{drive} Abbreviation for DRIVING VOLTAGE or OUTPUT VOLTAGE.

VE Abbreviation of value engineering.

V_E Symbol for EMITTER VOLTAGE of a bipolar transistor.

vectograph A graphic three-dimensional representation of a scene composed of superimposed images photographed through polarizing filters of different orientation, and reconstructed by a similar viewing technique.

vector **1.** A quantity having direction (in two or three dimensions) and magnitude. **2.** A graphical representation of a quantity, as defined in **1**. It consists of a straight arrow indicating the direction, and whose length is proportional to the magnitude.

vector addition The summation of two vectors, determined by adding their corresponding component values. In two dimensions, for example, suppose **A** = (x_1,y_1) and **B** = (x_2,y_2). Then **A** + **B** = $[(x_1 + x_2),(y_1 + y_2)]$. The sum can be found geometrically by constructing a parallelogram from the vectors, corresponding to the two known sides, as determined by the vectors. The sum is the vector originating at the point where the two vectors originate, and extending the diagonal of the parallelogram.

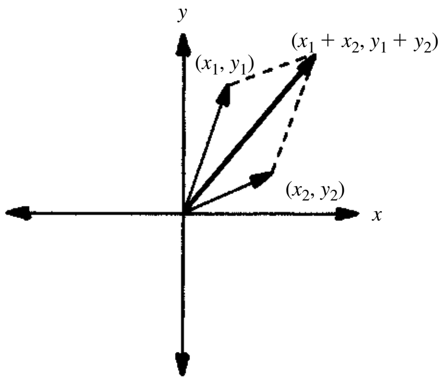

vector addition

vector admittance The reciprocal of VECTOR IMPEDANCE.

vector cardiograph An electrocardiograph that indicates the magnitude and the direction of a heart signal.

vector components **1.** The Cartesian coordinates of a vector, or its component angle and radius (polar coordinates). **2.** Quantities that can be represented by means of vectors (e.g., impedance and velocity).

vector diagram A graphical representation of vector quantities.

vector function A function having both magnitude and direction. Also see VECTOR; VECTOR COMPONENTS, **2**; and VECTOR QUANTITY.

vector generator A device that graphically illustrates vectors.

vectorial angle **1.** The angle between a vector and the horizontal axis. **2.** The angle between two vectors.

vector impedance Complex impedance (i.e., an impedance of the form $R + jX$, where R is resistance, X is reactance, and j is an operator indicating that the reactance is 90 degrees out of phase, relative to the resistance).

vector-mode display On a cathode-ray-tube display, the data representation in which points on the screen are connected by straight lines.

vector power The vector quantity $(P_a^2 + P_x^2)^{1/2}$, where P_a is the active power and Px is the reactive power.

vector power factor The ratio $P_a/(P_a^2 + P_x^2)^{1/2}$, where P_a is the active power and Px is the reactive power. In sine-wave situations, this ratio (of active power to vector power) produces a figure identical to the conventional power factor.

vector quantity A quantity having both magnitude and direction, and that can, therefore, be represented by a vector.

vectorscope A special oscilloscope for visual adjustment of a color-television receiver by means of a color-phase diagram.

vector sum The resultant of two nonparallel vectors; therefore, the resultant of the forces or quantities represented by them. For example, reactance and resistance can be represented by two perpendicular vectors.

vector voltmeter A voltmeter that indicates the phase as well as the amplitude of an alternating-current voltage.

V_{EE} Symbol for *emitter-voltage supply* of a bipolar transistor.

vee antenna A center-fed antenna in which the two halves of the radiator form an angle considerably less than 180 degrees.

vee-beam system An elevation-measuring radar system in which fan-shaped vertical and inclined beams, intersecting at ground and rotating continuously around the vertical axis, are radiated by one antenna. Target elevation is determined from the interval between successive echoes from the target.

vee connection of transformers In a three-phase system, a method of connecting two transformers so that the line current and voltage equal the phase current and voltage. Also called *open-delta connection.*

vee-cut crystal A piezoelectric plate cut from a quartz crystal so that its faces are not parallel to the *x*, *y*, or *z* axis of the crystal. Also see CRYSTAL AXES and CRYSTAL CUTS.

vee particle A short-lived elementary particle that results when high-energy neutrons or protons collide with nuclei. The particle can be positive, negative, or neutral, and gets its name from its cloud-chamber track.

vee signal In radiotelegraphy, the letter V (di-di-di-dah) transmitted as a test signal, usually three times in rapid succession. It is used during on-the-air transmitter tests, followed by the station's call letters.

vehicle An inert substance, usually a liquid, that acts as a solvent, carrier, or binder for some other, more active substance. Thus, shellac can be the vehicle for the metallic powder in a silver paint used in silk-screening electronic circuits.

vehicle smog-control device See COMPUTER-CONTROLLED CATALYTIC CONVERTER and EXHAUST ANALYZER.

vel Abbreviation of VELOCITY.

velocimeter **1.** An instrument for measuring the velocity of sound in various materials. **2.** An electronic velocity meter—especially a radial-velocity meter using Doppler radar. **3.** An electronic flow meter.

velocity Symbol, v. Abbreviation, vel. Unit, distance per unit time [e.g., meters per second (m/s)]. **1.** The change in the position of a body over a specified period of time. **2.** The derivative of displacement with respect to time. Compare ACCELERATION.

velocity constant See VELOCITY FACTOR.

velocity error For a servomechanism in which the input and output shafts rotate at the same speed, the angular displacement between them.

velocity factor Abbreviation, v. For a transmission line, the ratio of the speed of electromagnetic wave propagation in the line to the speed of electromagnetic waves in a vacuum (299,792 kilometers per second). It can be expressed as a number between 0 and 1, or as a percentage between 0 and 100. In practical feed lines, v ranges from about 0.66 (for coaxial line with a solid polyethylene dielectric) to about 0.95 (for open-wire line with widely separated spacers).

velocity hydrophone A hydrophone whose output, like that of the velocity microphone, is proportional to the instantaneous particle velocity in the sound wave impinging on the device.

velocity-lag error A lag (proportional to the input-variation rate) between the input and output of a device, such as a servomechanism.

velocity level For a sound, the ratio v_0, expressed in decibels:

$$v_0 = 20 \log_{10}(v_s/v_r)$$

where v_s is the particle velocity of the sound, and v_r is a reference particle velocity.

velocity microphone A microphone in which the vibratory element is a thin, aluminum or Duralumin ribbon suspended loosely between the poles of a strong permanent magnet. Vibration of the corrugated ribbon in the magnetic field causes an audio-frequency voltage to be induced across the ribbon. The microphone is so called because its output is proportional to the instantaneous particle velocity in the sound wave impinging on the ribbon. Also called *ribbon microphone.*

velocity-modulated amplifier A circuit in which radio-frequency amplification is obtained by velocity modulation.

velocity-modulated oscillator A vacuum-tube device in which an electron stream is velocity-modulated (see VELOCITY MODULATION) as it passes through a resonant cavity (the buncher); the subsequent energy, of a higher intensity, is extracted from the bunched stream as it passes through another resonant cavity (the catcher). Feedback from catcher to buncher sustains oscillations. See, for example, KLYSTRON OSCILLATOR.

velocity-modulated tube A vacuum tube utilizing velocity modulation. See, for example, KLYSTRON.

velocity modulation The process in which the input signal of a vacuum tube varies the velocity of the electrons in a constant-current electron beam in sympathy with the input signal.

velocity resonance A state of resonance in which a 90-degree phase difference is between the fundamental frequency of oscillation of a system and the fundamental frequency of the applied signal.

velocity sorting Selecting and separating electrons, according to their velocity.

velocity spectrograph A device that uses electric or magnetic deflection to separate charged particles into various streams, according to their velocity.

velocity variation See VELOCITY MODULATION.

velocity-variation amplifier See VELOCITY-MODULATED AMPLIFIER.

velocity-variation oscillator See VELOCITY-MODULATED OSCILLATOR.

velometer An instrument used to measure the velocity of air (e.g., to determine wind speed).

venetian-blind effect A television-display malfunction in which the picture seems to appear through horizontal slits.

Venn diagram A method of illustrating the relationship among various subsets within a specified universal set. The subsets are shown as geometric figures, usually circles.

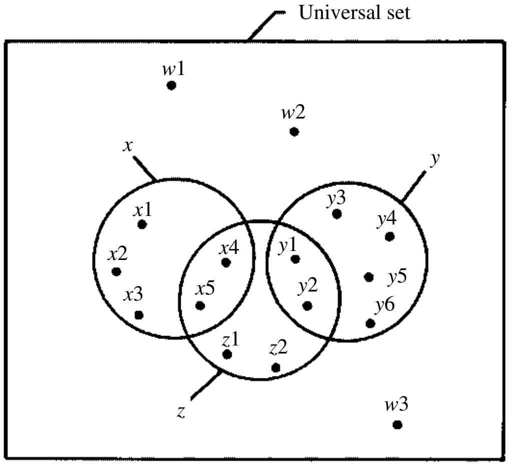

Venn diagram

vent **1.** An opening in an equipment enclosure for the entry of cool air or the escape of warm air. **2.** In an electrolytic capacitor, a lightly covered blowout hole for the relief of gas pressure in the event of severe overload. **3.** A small opening for relieving gas pressure in an automobile battery. **4.** In a loudspeaker enclosure, such as the bass-reflex type, an auxiliary opening that extends the low-frequency range of the speaker.

vented baffle A loudspeaker enclosure having the proper auxiliary opening(s) for coupling the speaker to the surrounding air.

vented-baffle loudspeaker See ACOUSTICAL PHASE INVERTER and BASS-REFLEX LOUDSPEAKER.

vent mount A metal bracket for fastening an antenna mast to a (plumbing) vent pipe on a roof.

ventricular fibrillation See FIBRILLATION.

verification **1.** The process of ensuring that two sets of data are identical. **2.** The process of validating (see VALIDATE).

vernier **1.** An auxiliary scale along which a regular, linear scale (such as that of a tuning dial) slides. The vernier scale is graduated so that when the main scale is set to an unmarked point between two of its graduations, and the "0" point on the vernier scale is used as the index, "1" on the vernier scale will exactly coincide with "1" on the main scale. The corresponding number on the vernier scale indicates the exact number of subdivisions between two points on the main scale. Vernier arrangements are provided with the dials of some precision analog instruments. **2.** See VERNIER CAPACITOR. **3.** See VERNIER DIAL. **4.** See VERNIER RESISTOR.

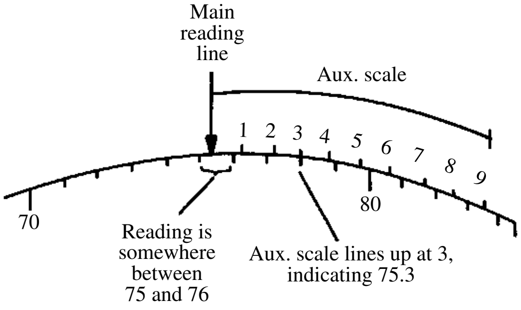
Main reading line
Aux. scale
1
2
3
4
5
6
7
8
9
70
80
Reading is somewhere between 75 and 76
Aux. scale lines up at 3, indicating 75.3

vernier, 1

vernier capacitor A low-capacitance variable capacitor connected in parallel with a higher-capacitance fixed or variable capacitor for precise adjustment of the total capacitance.

vernier dial **1.** A slow moving dial for fine tuning an adjustable device. The required reduction ratio is obtained with gears, friction wheels, or a belt-and-pulley combination. **2.** A dial provided with an accessory vernier scale.

vernier resistor A low-resistance variable resistor connected in series with a higher-resistance fixed or variable resistor to precisely adjust the total resistance.

vernier rheostat See VERNIER RESISTOR.

vers Abbreviation of VERSED SINE.

versatile automatic test equipment Abbreviation, VATE. For troubleshooting the electronic systems of missiles, a computer-controlled tester that isolates faults through logical operations.

versed sine Abbreviation, vers. A trigonometric function equal to the difference between the cosine of an angle and one; vers $x = 1 - \cos x$.

vert Abbreviation of VERTICAL.

vertex **1.** The terminal point at which two or more branches of a network meet. Also see NODE, **1**. **2.** The point of intersection of two lines that form an angle.

vertex plate In the reflector of a microwave antenna, a matching plate mounted at the vertex.

vertical **1.** The direction indicated by a line perpendicular to the plane of the horizon (e.g., a vertical axis). **2.** In the 45-45 recording process, the signal resulting from sound that arrives simultaneously and 180 degrees out of phase at the two microphones and causes up-and-down movement of the cutting stylus.

vertical amplification Gain provided by the vertical channel of a device, such as an oscilloscope, cathode-ray electrocardiograph, or television receiver. Compare HORIZONTAL AMPLIFICATION.

vertical amplifier The circuit or device that provides gain in the vertical channel of a device. Compare HORIZONTAL AMPLIFIER.

vertical-amplitude control **1.** See VERTICAL-GAIN CONTROL. **2.** In a color television receiver, one of the three controls by which the amplitude of the parabolic voltages applied to the coils of the magnetic-convergence assembly are adjusted.

vertical angle of radiation The angle, with respect to the plane of the horizon, at which radio-frequency energy is propagated from a transmitting antenna.

vertical antenna **1.** An antenna that consists essentially of a single, straight, vertical conductor. **2.** An antenna that is mounted vertically, instead of horizontally.

vertical blanking See VERTICAL RETRACE BLANKING.

vertical-blanking pulse In a television signal, the pulse at the end of each field that accomplishes vertical retrace blanking. Compare HORIZONTAL-BLANKING PULSE.

vertical-centering control See CENTERING CONTROL.

vertical channel The system of amplifiers, controls, and terminations that constitutes the path of the vertical signal applied to a device, such as an oscilloscope. Compare HORIZONTAL CHANNEL.

vertical compliance In disc sound reproduction, the ease with which the stylus can move vertically while it is in position on the disc.

vertical coordinates See CARTESIAN COORDINATES.

vertical deflection **1.** In an oscilloscope or television receiver, the movement of the spot up and down on the screen. Compare HORIZONTAL DEFLECTION. **2.** In a direct-writing recorder, up or down deflection of the pen.

vertical deflection coils In a cathode-ray tube, the pair of coils in a deflection yoke that provides the electromagnetic field for the vertical deflection of the electron beam. Compare HORIZONTAL DEFLECTION COILS.

vertical deflection electrodes See VERTICAL DEFLECTION COILS and VERTICAL DEFLECTION PLATES.

vertical deflection plates In an oscilloscope (and in some early television picture tubes), a pair of plates that provides the electrostatic field for vertical deflection. Compare HORIZONTAL DEFLECTION PLATES.

vertical dipole An antenna consisting of a straight, center-fed, half-wave conductor oriented vertically.

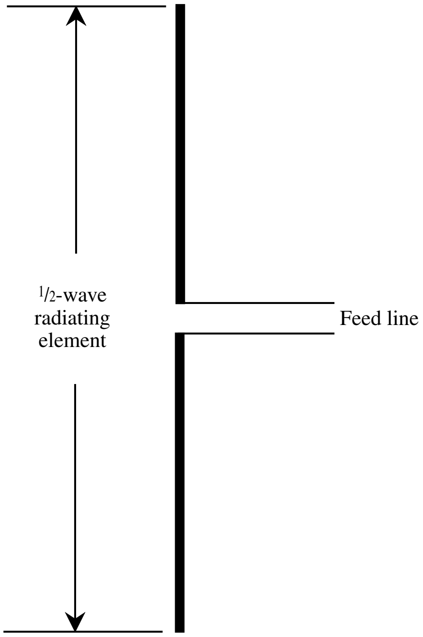

vertical dipole

vertical dynamic convergence During the scanning of points along a vertical line through the center of a color-television picture tube, the convergence of the electron beams at the aperture mask. Compare HORIZONTAL DYNAMIC CONVERGENCE.

vertical field strength The field strength of signals in a vertical plane passing through an antenna. Compare HORIZONTAL FIELD STRENGTH.

vertical-field-strength diagram A plot of vertical field strength.

vertical frequency response The gain-frequency characteristic of the vertical channel of an instrument, such as an oscilloscope. Compare HORIZONTAL FREQUENCY RESPONSE.

vertical gain The overall amplification (gain) provided by the vertical channel of an instrument, such as an oscilloscope, graphical recorder, or television receiver. Compare HORIZONTAL GAIN.

vertical-gain control A control, such as a potentiometer, for adjusting vertical amplification. Compare HORIZONTAL-GAIN CONTROL.

vertical-hold control In a television receiver, the control for adjusting the frequency of the vertical oscillator so that the picture can be locked to prevent vertical roll. Compare HORIZONTAL-HOLD CONTROL.

vertical-incidence transmission Transmission of a wave vertically to the ionosphere and the subsequent reflection of the wave to the earth.

vertical-lateral recording In stereophonic disc recording, the recording of one signal vertically (see VERTICAL RECORDING) and the other laterally (see LATERAL RECORDING).

vertical linearity Linearity of response (gain and deflection) of the vertical channel of an oscilloscope or television receiver. A linear picture is neither contracted nor expanded vertically in any part of the screen. Compare HORIZONTAL LINEARITY.

vertical-linearity control In an oscilloscope or television receiver, the control whereby vertical linearity is adjusted. Compare HORIZONTAL-LINEARITY CONTROL.

vertically polarized wave An electromagnetic wave whose electric lines of flux are perpendicular to the plane of the horizon. Compare HORIZONTALLY POLARIZED WAVE.

vertical-metal-oxide semiconductor field-effect transistor Abbreviation, VMOSFET. A metal-oxide semiconductor field-effect transistor fabricated so that the current flow within the device is vertical, instead of the usual horizontal, affording several advantages over the conventional MOSFET.

vertical oscillator In a television receiver, the oscillator that generates the vertical sweep signal. Compare HORIZONTAL OSCILLATOR.

vertical-output regulator In a television receiver, a voltage-dependent resistor used to stabilize the sweep voltage across the horizontal deflection coils—especially during warmup.

vertical output stage In a television receiver, an output amplifier following the vertical oscillator. Compare HORIZONTAL OUTPUT STAGE.

vertical polarization Pertaining to an electromagnetic wave whose electric lines of flux are perpendicular to the plane of the horizon. In general, when the radiating element of an antenna is vertical, the electric lines of flux in the transmitted waves are vertical, and the antenna is most sensitive to incoming signals whose electric lines of flux are vertical. Compare HORIZONTAL POLARIZATION.

vertical positioning control See CENTERING CONTROL.

vertical quantity The quantity measured along the *y*-axis of a graph, or represented by the vertical deflection of an electron beam. Compare HORIZONTAL QUANTITY.

vertical radiator See VERTICAL ANTENNA.

vertical recording Disc recording in which the depth of the groove is varied in sympathy with the sound. Also called HILL-AND-DALE RECORDING. Compare LATERAL RECORDING.

vertical redundance In computer operation, the error state when a character has an odd number of bits in an even-parity system, or vice versa.

vertical resolution The number of horizontal wedge-lines that can be easily seen in a television test pattern before they blend. Compare HORIZONTAL RESOLUTION.

vertical retrace In a cathode-ray device, such as an oscilloscope or television receiver, the rapid return of the beam to its starting point at the top of the screen after completely traversing the screen from top to bottom. Compare HORIZONTAL RETRACE.

vertical retrace blanking In a television receiver, the automatic shutoff of the electron beam during a vertical retrace period, to prevent an extraneous line being traced on the screen during this period. Also see BLACKOUT, **2**; BLANKING INTERVAL, **1**, **2**; BLANKING PEDESTAL; BLANKING TIME; VERTICAL RETRACE; and VERTICAL RETRACE PERIOD. Compare HORIZONTAL RETRACE BLANKING.

vertical retrace period In a television receiver, the interval during which the spot returns from the bottom to the top of the screen after having traced all horizontal lines from top to bottom. Compare HORIZONTAL RETRACE PERIOD.

vertical sensitivity The signal voltage required at the input of a vertical channel to produce full vertical deflection. Also see VERTICAL GAIN. Compare HORIZONTAL SENSITIVITY.

vertical signal A signal serving as a vertical quantity (e.g., one that deflects the beam in a cathode-ray tube upward and/or downward). Compare HORIZONTAL SIGNAL.

vertical-speed transducer A transducer whose electrical output is proportional to the vertical speed of an aircraft or missile carrying the transducer.

vertical stylus force In disc sound reproduction, the downward force (in grams or ounces) that the stylus exerts on the disc.

vertical sweep **1.** In a cathode-ray tube, especially a television picture tube, the movement of the spot up or down on the screen. Compare HORIZONTAL SWEEP. **2.** The circuit that produces this sweep.

vertical sweep frequency The frequency at which vertical sweep occurs. In a television receiver, it is 60 Hz. Also called *vertical sweep rate.* Compare HORIZONTAL SWEEP FREQUENCY.

vertical sweep rate See VERTICAL SWEEP FREQUENCY.

vertical synchronization In a television receiver, synchronization of the vertical component of scanning with that of the camera. Also see VERTICAL SYNC PULSE. Compare HORIZONTAL SYNCHRONIZATION.

vertical sync pulse In a video signal, the pulse that synchronizes the vertical component of scanning in a television receiver with that of the camera, and that triggers vertical retrace and blanking. Compare HORIZONTAL SYNC PULSE.

vertical wave See VERTICALLY POLARIZED WAVE.

vertical width control See WIDTH CONTROL, **1**, **2**.

very high frequency Abbreviation, VHF. Radio frequencies in the range 30 MHz to 300 MHz, corresponding to free-space wavelengths of 10 meters to 1 meter. It is sometimes divided into *VHF low band* (30 MHz to 50 MHz) and *VHF high band* (50 MHz to 300 MHz). Also see RADIO SPECTRUM, **1**.

very high resistance Abbreviation, VHR. Large values of resistance; usually expressed in megohms, gigohms, or teraohms.

very-high-resistance voltmeter Abbreviation, VHRVM. A voltmeter using a low-range microammeter or picoammeter and a very high value of multiplier resistance (see VOLTMETER MULTIPLIER).

very-high-speed integrated circuit An integrated circuit used for switching or other digital functions at thousands, millions, or billions of changes of state per second.

very-large-scale integration Abbreviation, VLSI. The inclusion of several complete systems, such as computers, on a single integrated-circuit chip. This can extend several orders of magnitude beyond large-scale integration (LSI).

very long range Abbreviation, VLR. Pertaining to ground radar sets having a maximum slant range of over 250 miles. Compare VERY SHORT RANGE.

very low frequency Abbreviation VLF. A radio frequency in the range 10 kHz to 30 kHz, corresponding to wavelengths between 30 kilometers and 10 kilometers. Also see RADIO SPECTRUM, **1**.

very low resistance Abbreviation, VLR. Values of resistance less than 1 ohm, usually expressed in milliohms or microhms.

very short range Abbreviation, VSR. Pertaining to ground radar sets having a maximum slant range of less than 25 miles. Compare VERY LONG RANGE.

vestigial **1.** An effect that remains as a by-product, but that serves no directly applicable purpose. See, for example, VESTIGIAL SIDEBAND. **2.** Unnecessary; extraneous.

vestigial sideband **1.** A portion of one sideband in an amplitude-modulated signal, remaining after passage through a selective filter. **2.** An amplitude-modulated signal in which one sideband has been partially or largely suppressed. **3.** The small amount of energy emitted in the unused sideband in a single-sideband transmitter.

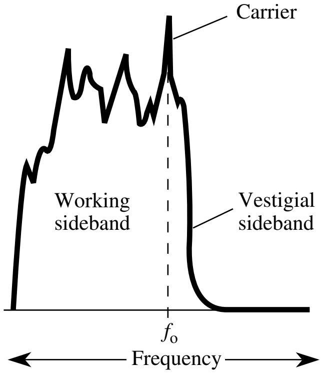

vestigial sideband, 2

vestigial-sideband filter A filter operated between an amplitude-modulated transmitter and an antenna to obtain a vestigial-sideband signal.

vestigial-sideband signal An amplitude-modulated signal in which one of the sidebands has been partially suppressed.

vestigial-sideband transmission Transmission of a signal containing a vestigial sideband. In television, for example, the upper sideband is transmitted fully, while the lower sideband is almost completely suppressed. The lower sideband is, therefore, a vestigial sideband.

vestigial-sideband transmitter An amplitude-modulated transmitter equipped with the filters or other subcircuits necessary for emitting a vestigial-sideband signal.

VF Abbreviation of VIDEO FREQUENCY.

V_{FB} Abbreviation for *feedback voltage* in an integrated circuit device. The term applies especially to operational amplifiers.

VFO Abbreviation of VARIABLE-FREQUENCY OSCILLATOR.

V_g Abbreviation of GENERATOR VOLTAGE. (Also, *Eg.*)

VGA Abbreviation of VARIABLE-GAIN AMPLIFIER.

V_{GD} Symbol for GATE-DRAIN VOLTAGE of a field-effect transistor.

V_{GS} Symbol for GATE-SOURCE VOLTAGE of a field-effect transistor.

VHF Abbreviation of VERY HIGH FREQUENCY.

VHF high band See VERY HIGH FREQUENCY.

VHF low band See VERY HIGH FREQUENCY.

VHF omnirange Abbreviation, VOR. A very-high-frequency radionavigation system that provides radial lines of position.

VHF oscillator A radio-frequency oscillator specially designed to operate in the range 30 MHz to 300 MHz.

VHF receiver **1.** A radio receiver specially designed to operate in the range of 30 MHz to 300 MHz. **2.** A television receiver for the VHF channels (see VHF TELEVISION CHANNELS).

VHF television channels Twelve 6-MHz-wide channels (television channels 2 through 13) between 54 and 216 MHz.

Channel	*Frequency (MHz)*
2	54.0 – 60.0
3	60.0 – 66.0
4	66.0 – 72.0
5	76.0 – 82.0
6	82.0 – 88.0
7	174 – 180
8	180 – 186
9	186 – 192
10	192 – 198
11	198 – 204
12	204 – 210
13	210 – 216

VHF television channels (United States)

VHF transmitter A radio or television transmitter specially designed to operate in the range 30 MHz to 300 MHz.

VHR Abbreviation of VERY HIGH RESISTANCE.

VHRVM Abbreviation of VERY-HIGH-RESISTANCE VOLTMETER.

VHS videocassette recorder A popular scheme for videocassette recording, used in millions of households in the United States alone. Typical VHS cassette play times range from about two hours to more than eight hours.

VHSIC Abbreviation for VERY HIGH SPEED INTEGRATED CIRCUIT.

VI **1.** Abbreviation of VOLUME INDICATOR. **2.** Abbreviation of VISCOSITY INDEX.

V_i Symbol for INPUT VOLTAGE. (Also, V_{in}.)

vibrating bell A bell with a striking mechanism that oscillates, causing a continuous ringing sound.

vibrating-reed frequency meter See POWER-FREQUENCY METER, **2**.

vibrating-reed oscillator An audio-frequency oscillator in which an iron, steel, or alloy reed vibrating in a magnetic field acts like a tuning fork to control the oscillation frequency.

vibrating-reed relay An electromechanical relay in which the movable contact is carried by the end of a thin, metal strip (reed) of iron or magnetic alloy. The strip is supported within the magnetic field of a coil carrying an alternating control current; when the frequency of this current corresponds to the natural (resonant) frequency of the reed, it vibrates vigorously enough to close the contacts. Such a relay, consequently, is frequency selective.

vibrating-wire oscillator An oscillator similar to the vibrating-reed oscillator, but using a wire, instead of a reed.

vibrating-wire transducer A transducer in which the fluctuating tension of a thin wire suspended in a magnetic field frequency-modulates the operating voltage.

vibration **1.** The changing of the position or dimensions of a body back and forth, usually at a rapid rate, an action seen in the repetitive movement of a musical string, headphone diaphragm, loudspeaker cone, loose machine, etc. **2.** See OSCILLATION.

vibration analyzer See VIBRATION METER.

vibration calibrator A device that generates a standard vibration, usually at a fixed frequency, for calibrating vibration meters, pickups, transducers, etc.

vibration galvanometer A type of alternating-current galvanometer. The natural frequency of the movable element of the instrument is made equal to that of the alternating current under test, to obtain a reading.

vibration isolator In an electronic-equipment assembly, a cushioning support that protects the equipment from vibration.

vibration machine See VIBRATOR, **2**.

vibration meter An instrument for measuring the amplitude and frequency of vibration (see VIBRATION, **1**). It consists essentially of a vibration pickup followed by a selective amplifier and an indicating voltmeter graduated in vibration units.

vibration pickup A transducer that senses mechanical vibration and converts it into a proportionate output voltage, or changes resistance in sympathy with the vibration.

vibrato In electronic musical instruments, a circuit or device that modulates a note by varying its frequency, amplitude, or both, at an extremely low frequency (a few hertz).

vibrator **1.** See INTERRUPTER. **2.** See VIBRATING-REED RELAY. **3.** A device for shaking something under test at selected frequencies and amplitudes.

vibrator-type power supply A battery-operated, high-voltage power supply in which one vibrator (see INTERRUPTER) makes and breaks direct current flowing from the battery into the primary winding of a step-up transformer, and another vibrator rectifies the high voltage delivered by the secondary winding.

vibrator-type rectifier A vibrator (see INTERRUPTER) that connects one terminal of the secondary winding of a transformer to an output terminal each time that terminal is positive. When the negative half-cycle appears at the transformer terminal, the vibrator is open. In this way, the alternating-current output of the transformer is rectified.

vibrocardiography A method of monitoring and recording the movement of the chest cavity as a result of the beating of the heart.

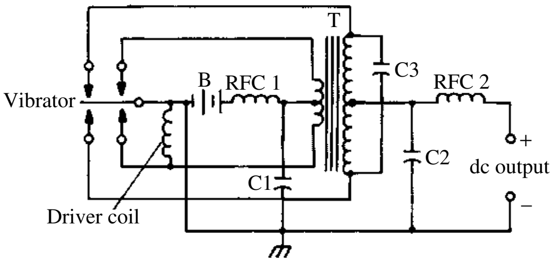

vibrator-type power supply

vibrograph A device for observing and recording vibration.

vibrometer See VIBRATION METER.

vibroscope An instrument for observing (and sometimes recording) vibration.

Victor Phonetic alphabet communications code word for the letter V.

Victron A brand of polystyrene, a low-loss plastic insulant that is especially useful at high radio frequencies.

video **1.** Pertaining to television—especially to its picture circuitry or to circuits and devices related to television, but used for other purposes. **2.** The picture portion of a television broadcast, as opposed to the sound (audio) portion. **3.** Sometimes, a cathode-ray-tube terminal or display. **4.** Generally, television in adjective sense (e.g., *video play*). **5.** The images on a computer display or monitor.

video amplifier **1.** In television, the wideband stage (or stages) that amplifies the picture signal and presents it to the picture tube. **2.** A similar wideband amplifier, such as an instrument amplifier or preamplifier having at least a 4-MHz bandwidth.

video animation Movement of graphic images on a cathode-ray-tube display by the use of computer-manipulation through special software, under the direction of an operator using a keyboard, joystick, or digitizer.

video buffer A unity-gain circuit that minimizes loading effects caused by having several video multiplexer inputs connected to the same signal source.

video carrier The television carrier amplitude-modulated by the picture information, synchronization pulses, and blanking pulses.

videocassette recorder Abbreviation, VCR. A machine for recording television programs on magnetic-tape cassettes (see CASSETTE, **1**).

videocast **1.** Television broadcast. Also called TELECAST. **2.** To make a television broadcast.

videocaster See TELECASTER.

video CD-ROM A digital optical VIDEODISC, resembling an audio or computer compact disc. A typical disc can store more than an hour's worth of conventional television programming.

video correlator A device used with radar to locate and identify a target with high precision.

video detector In a television receiver, the demodulator for the video signal. This detector follows the video intermediate-frequency amplifier and precedes the video amplifier. It usually embodies a relatively simple, wideband diode circuit.

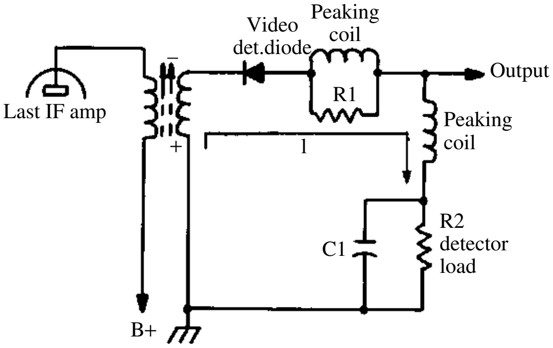

video detector

Video Digitizer Abbreviation, VDIG. A program developed for multimedia video in personal computer systems.

videodisc A flat magnetic or optical disc on which audio-visual programs are recorded, and from which they can be played back into a television receiver or multimedia personal computer.

videodisc recorder Abbreviation, VDR. A machine for recording television programs on discs comparable to those used for sound or data recording.

video discrimination In radar, the use of a circuit or device to decrease the width of the video-amplifier passband.

video frequency **1.** Pertaining to signals in the wide passband of a video amplifier: 30 Hz to 4 MHz. **2.** Pertaining to a device capable of operating over the wide passband of a video amplifier (e.g., a video-frequency alternating-current voltmeter).

video-frequency amplifier An amplifier capable of handling signals of a wide frequency range (e.g., direct current to 4 MHz). Also see VIDEO FREQUENCY, **1**.

video gain control In a television receiver, a gain control for adjusting the video signal amplitude. In a color-television receiver, a pair of these controls permits adjustment of the three color signals to the proper amplitude ratio.

video game A game, such as football, basketball, slot machine, tic-tac-toe, etc., played on the screen of a television receiver or on a computer.

videogenic Suitable for television.

video IF amplifier In a television receiver, the broadband intermediate-frequency (IF) amplifier of the video signal. In modern receivers, this IF amplifier also handles the sound signal.

video mapping A system of surveillance or mapping in which, for reference, the outlines of the area being surveyed are superimposed electronically on the radar display of the area.

video masking A form of radar signal processing in which ground clutter and other unwanted echoes are removed, making the desired targets more readily visible.

video mixer A circuit or device for mixing the signals from two or more television cameras.

video modulation In television transmission, amplitude modulation of the carrier wave with pulses and waves corresponding to the picture elements.

video on demand Television service in which subscribers can choose the programs they want to watch, and also the specific times for viewing.

video random-access memory Abbreviation, VRAM. In computer systems, RANDOM-ACCESS MEMORY used primarily for enhancing the performance of the display.

video recording **1.** Recording wideband material (such as a video signal) on a tape or disc. **2.** The recording made of a telecast. Also see VIDEOTAPE RECORDER.

video signal **1.** In television, the amplitude-modulated signal containing picture information and pulses. Also see VIDEO, **2** and VIDEO MODULATION. **2.** Broadly, a telecast signal, including sound.

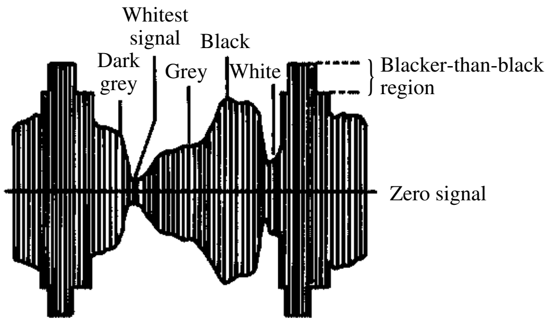

video signal

video stretching In some electronic navigation systems, increasing the duration of a video pulse.

video synthesizer A computerized device that produces graphical renditions of objects or circuits in three dimensions.

videotape **1.** A special magnetic tape for video recording (see VIDEO RECORDING, **2**). **2.** To make a video recording.

videotape recorder Abbreviation, VTR. A wideband, magnetic-tape recorder for producing video recordings with a camera, or for making a record of television programs, for subsequent reproduction (playback).

videotape recording **1.** Abbreviation, VTR. The technique of recording video signals on magnetic tape. Also see VIDEO RECORDING, **1**, **2** and VIDEOTAPE. **2.** A tape on which a video signal has been recorded.

videotext A system that allows television viewers to dial up special material, such as stock market quotations, weather data, sports scores, etc.

vidicon A television camera tube in which the electron beam scans a charged-density pattern that has been formed and stored on the surface of the photoconductor. It is commonly used in camcorders, closed-circuit television systems, and robot vision systems. Its assets include compactness and sensitivity.

viewfinder An accessory or integral device in which an image (formed optically or electronically) corresponds to the image viewed by the camera with which it is used.

viewing mesh In a cathode-ray storage tube, the mesh on which the image is presented for viewing by the operator. Also see CHARGE-STORAGE TUBE.

viewing mirror In an oscilloscope-camera assembly, a flat, slant-mounted mirror that reflects the image on the oscilloscope screen to the viewer's eye.

viewing screen In a cathode-ray device, such as an oscilloscope tube or television picture tube, the face on which the image appears.

viewing time In a storage type cathode-ray tube, the length of time for which the image persists.

viewing window See WINDOW, **2**.

Villard circuit **1.** A voltage-doubler circuit using only one diode and two capacitors. The open-circuit direct-current output voltage is approximately twice the peak value of the alternating-current input voltage. Also see VOLTAGE DOUBLER. **2.** See SELECTOJECT.

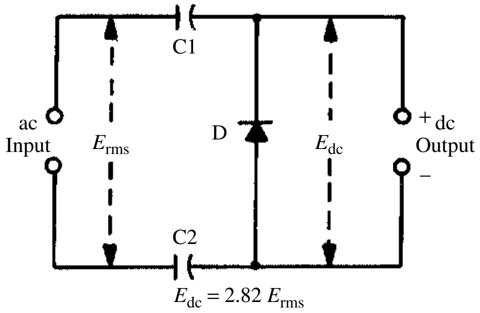

Villard circuit, 1

Villari effect In certain magnetostrictive materials, the change in magnetic induction that accompanies the application of a mechanical stress in a given direction. Also see MAGNETOSTRICTION.

vinyl A general name for *vinyl copolymer resins.*

vinylidene chloride A plastic insulant. Dielectric constant, 3 to 5. Dielectric strength, 20 kV/mm.

Vinylite-A A brand of the plastic *polyvinyl acetate.*

Vinylite-Q A brand of the plastic *polyvinyl chloride.*

vinyl resin A synthetic resin resulting from the polymerization of compounds that contain the group $CH_2 = CH–$.

violation In a security system, a condition that results in the actuation of an alarm. Example: an intruder breaks the light beam in an electric eye.

virgin magnetic material Core or shield material that has never been magnetized. When virgin material is first subjected to magnetization, the hysteresis loop starts at zero, but never returns to zero (see BOX-SHAPED LOOP).

virgin neutron Any neutron that has not been involved in a collision with another particle subsequent to its initial generation.

virgin record See BLANK RECORD.

virgin tape See BLANK TAPE.

virtual address In computer operations, an address that must be modified to refer to a location in main memory.

virtual decision value In quantizing or encoding, a method of expressing the maximum input amplitude. Two values are extrapolated from actual decision values—one at either end of the working range used.

virtual height The altitude that a vertically propagated electromagnetic wave would reach before reflection if its path in the ionosphere were a straight line. The actual distance at which the wave penetrates the ionosphere before reflection is less than the virtual height.

virtual image The image formed when rays from a scene diverge after passing beyond the focal point of a convex lens. The scene appears inverted to an observer. Compare REAL IMAGE.

virtual memory **1.** In a computer system, a means of using two or more memory stores simultaneously. **2.** Auxiliary memory used in conjunction with the main, or core, memory.

virtual ppi reflectoscope A device used to superimpose a virtual image of a chart or map onto a plan-position indicator (ppi) radar display. Also see VIDEO MAPPING.

virtual reality Abbreviation, VR. A general term for any of various high-level computer simulation or remote-control programs. The user often wears a head-mounted display that provides vivid, three-dimensional imagery and binaural sound. It can be used in robotic telepresence systems for precision remote control.

virus See COMPUTER VIRUS.

vis **1.** Abbreviation of *visibility.* **2.** Abbreviation of *visual.*

viscometer An instrument for measuring viscosity. There are several electronic versions of this device. In one, a steel ball falls through a material (such as an oil) under test and distorts the magnetic field of a pickup coil, causing the deflection of a meter by an amount proportional to the speed of the ball and, consequently, to the viscosity of the fluid.

viscosimeter See VISCOMETER.

viscosity The resistance offered by a fluid (liquid or gas) to objects passing through it. The viscosity of pure water is low; that of heavy oil is high. Expressed in newton-seconds per meter squared.

viscosity index Abbreviation, VI. A number indicating how well an oil retains its viscosity with temperature changes; larger indexes are assigned to oils that are little influenced by variations in temperature.

viscous-damped arm A phonograph pickup arm with an oil dashpot to prevent arm resonance and to slow the descent of the arm to the disc.

viscous damping The use of a viscous fluid in the dashpot of a device (such as a relay, timer, or pickup arm) to provide damping. See, for example, DASHPOT RELAY.

viscous hysteresis A slow, slight increase in the magnetization of a material when the magnetizing field is constant. Compare STATIC HYSTERESIS.

visibility factor See DISPLAY LOSS.

visible radiation Electromagnetic radiation that is perceptible to the eye. Also see VISIBLE SPECTRUM.

visible spectrum The band of electromagnetic wavelengths that the human eye perceives as visible light. For most people, this band extends from approximately 750 nanometers (nm), representing red light, down to 390 nm, representing violet light. The visibility curve peaks in the yellow-green region at about 560 nm; 1 nm = 0.000

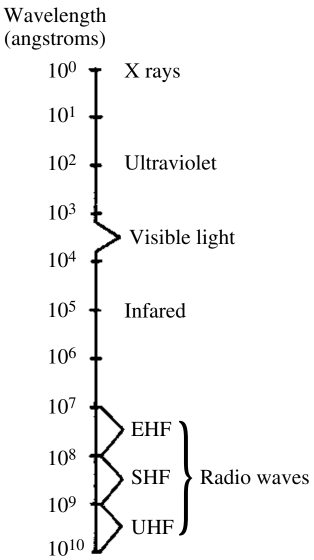

visible spectrum

000 001 meter = 10^{-9} m. Also see ELECTROMAGNETIC THEORY OF LIGHT.

vision system A set of devices that allows a machine (computer or robot) to observe and respond to visual stimuli. Typically, it includes a camera [such as a vidicon or charge-coupled device (CCD)], a cable, electromagnetic or fiberoptic link, a receiver, and software for interpreting the images.

visual alignment Optimization of a circuit (such as a radio receiver) with the aid of meter deflections, rather than with audible indications. Also called *silent alignment.*

visual-aural radio range A radio aid to air navigation (characterized by an audible signal, meter deflection, or both) used by the pilot of an aircraft to determine if the course is being followed.

visual-aural range See VISUAL-AURAL RADIO RANGE.

visual-aural signal tracer A troubleshooting instrument for following a signal through a circuit by sensing it at successive points in the circuit. The signal strength is indicated by a meter and a loudspeaker.

visual carrier frequency See VIDEO CARRIER and VIDEO-FREQUENCY, **1**.

visual communication **1.** Transmission and reception of signals by visible-light means, such as modulated beams of flashing lights. **2.** Transmission and reception of messages by direct visual observation, such as signal lights.

visual display unit Abbreviation, VDU. **1.** A computer peripheral consisting of a cathode-ray tube, a keyboard, and often a pointing device. **2.** A dumb terminal for operation of a computer from a remote site.

visual horizon **1.** The distance from a given point to the farthest visible point on the surface of the earth in a particular direction. **2.** The enclosed geometric plane figure on the surface of the earth, representing the set of farthest-visible points for a particular location.

visual telegraphy Telegraphy in which the received signals are read from a visual device, such as a blinking light, meter, or swinging pen.

visual transmitter The equipment used to transmit the picture portion of a television broadcast.

visual transmitter power The peak power output of a visual transmitter during normal operation.

vitreous Pertaining to a material or surface resembling, or related to, glass (e.g., the *vitreous enamel* on certain types of resistors).

VLF Abbreviation of VERY LOW FREQUENCY.

VLR **1.** Abbreviation of VERY LOW RESISTANCE. **2.** Abbreviation of VERY LONG RANGE.

VLSI Abbreviation of VERY-LARGE-SCALE INTEGRATION.

V/m Abbreviation of VOLTS PER METER (a unit of electric field strength).

VMOS Abbreviation of *vertical-metal-oxide semiconductor.*

VMOSFET Abbreviation of VERTICAL-METAL-OXIDE SEMICONDUCTOR FIELD-EFFECT TRANSISTOR.

V_o Symbol for OUTPUT VOLTAGE. Also, V_{out}.

VOA Abbreviation of VOLT-OHM-AMMETER.

vocabulary In computer-programming operations, a list of available operating codes and instructions for the computer. Also called INSTRUCTION SET.

Vocoder A device for reducing speech to frequencies low enough for efficient transmission through a limited-bandwidth channel. The term is a contraction of *voice coder.*

Voda In a telephone system that utilizes a radio link (using the same frequency for transmission in both directions) and land-line links, an automatic, voice-operated switching system for enabling the transmitter and disabling the receiver, and vice versa. The name is an acronym for *voice-operated device, anti-sing.*

Vodacom Contraction of *voice data communications.*

Voder An electronic device that synthesizes speech. The term is an acronym for *voice operation demonstrator.*

vogad A type of automatic gain control for audio amplifiers and modulators. In a radio transmitter, it maintains 100% modulation of the carrier—even when the speaker's voice level varies widely. The name is an acronym for *voice-operated gain adjusting device.*

voice actuation See VOICE-OPERATED CONTROL.

voice analyzer A circuit that evaluates the characteristics of human voices, such as the amplitude-vs.-frequency function or the amplitude-vs.-time function.

voice band See VOICE FREQUENCIES.

voice coil The moving coil of a dynamic microphone or dynamic speaker.

voice-controlled break-in A type of break-in operation for radiotelephony, in which the transmitter automatically switches on and the receiver switches off immediately when the operator starts talking; the transmitter switches off and the receiver switches on a moment after the operator stops talking.

voice-controlled relay An electronic relay that is actuated by the human voice.

voice filter **1.** A filter for passing, suppressing, or modifying voice frequencies. **2.** A parallel-resonant circuit inserted into a line that feeds several loudspeakers, to trap certain frequencies and thereby improve the sound of the reproduced voice.

voice frequencies **1.** The audio-frequency range of human speech, from about 60 Hz to 8000 Hz. **2.** The frequencies within the audio passband of a typical single-sideband voice transmitter or receiver. The lower limit is generally 200 Hz to 300 Hz; the upper limit is about 2500 Hz to 3000 Hz.

voice-frequency carrier telegraphy A type of carrier-current telegraphy (see WIRED WIRELESS) in which the modulated carrier can be transmitted over a telephone line having a voice-frequency bandwidth.

voice-frequency dialing A system of telephone dialing involving the conversion of direct-current pulses into voice-frequency alternating-current pulses.

voice-frequency telephony Wire telephone communication in which the frequencies of the electric waves are identical to the frequencies of the sound waves (or very nearly so).

voice grade Pertaining to a communications system having a bandpass capable of transferring a human voice with reasonable intelligibility. See VOICE FREQUENCIES, **2**.

voice-grade channel **1.** A telephone line and attendant equipment suitable for the transmission of speech and certain other information, such as control signals, digital data, etc. **2.** In a radiotelephone transmitter, a speech amplifier-modulator channel suitable only for voice frequencies.

voice-operated control Abbreviation, vox. Pertaining to a device (such as a relay, automatic modulation control, transmit-receive switch, etc.) that is actuated when the operator speaks into a microphone.

voice-operated device, anti-sing See VODA.

voice-operated gain-adjusting device See VOGAD.

voice-operated loss control and suppressor In wire telephony, a device that switches the loss from the transmitting line to the receiving line when the subscriber speaks, and vice versa.

voice-over The simultaneous recording of a human voice (such as that of a narrator) along with other sounds (such as music, children playing, wind in trees, ocean waves, chirping birds, etc.). The voice is generally louder than the other sounds.

voice print A graphic recording of the speech frequencies produced by an individual and used as a means of identifying that individual.

voice-recognition device See SPEECH RECOGNIZER.

voice security In voice communications system, the use of encryption and decryption, usually in the form of digital algorithms, to “scramble” speech at the source and “unscramble” it at the destination.

voice-stress analyzer Abbreviation, VSA. An instrument that samples the spoken voice and produces a display from which the relative amount of stress experienced by the speaker can be determined, and from which, in turn, probable truth or falsity of statements or answers can be inferred.

voice synthesis device See SPEECH SYNTHESIZER.

void space See VACUUM.

vol Abbreviation of VOLUME.

volatile **1.** Capable of evaporating (e.g., *volatile solvents* used in the encapsulation of electronic equipment). **2.** Explosive (noun and verb). **3.** Pertaining to a state which is difficult to maintain (e.g., a VOLATILE MEMORY).

volatile memory Memory, usually of the random-access type, whose data vanishes when power is removed unless some provision is made for memory backup. The most common means of memory backup is the use of an electrochemical cell or battery. Modern memory chips need so little current to store their data that a backup battery lasts almost as long in the circuit as it would on the shelf. The main advantage of this type of memory is the fact that it can easily be erased if desired. Compare NONVOLATILE MEMORY.

volatile storage See VOLATILE MEMORY.

volatile store See VOLATILE MEMORY.

Voldicon A form of semiconductor logic device used for analysis of analog signals. Trade name of Adage, Inc.

volt Abbreviation, V. The basic practical unit of difference of potential (i.e., of electrical pressure); 1 volt is the difference of potential produced across a resistance of 1 ohm by a current of 1 ampere. Also see KILOVOLT, MEGAVOLT, MICROVOLT, MILLIVOLT, NANOVOLT, and PICOVOLT.

Volta effect See VOLTA’S PRINCIPLE.

voltage Symbols, E, e, V, v. Electromotive force, or difference of potential; $E = IR$, where I is current and R is resistance. Also see VOLT.

voltage-actuated device An electronic device, such as a field-effect transistor, that amplifies a voltage signal or is controlled by a voltage, and draws virtually no signal current or control current. The opposite is a current-actuated device, such as a bipolar transistor.

voltage amplification **1.** Abbreviation, A_v. Amplification of an input-signal voltage to provide a higher output-signal voltage. **2.** Abbreviation, A_v. The signal increase (V_{out}/V_{in}) resulting from this process. Also called *voltage gain.*

voltage-amplification device A low-current device designed especially for voltage amplification. It provides little or no power amplification.

voltage amplifier An amplifier operated primarily to increase a signal voltage. Compare CURRENT AMPLIFIER and POWER AMPLIFIER.

voltage at peak torque Symbol, V_p. For a torque motor operated at 25°C, the voltage required to produce peak torque at standstill.

voltage attenuation **1.** The reduction in voltage at a given point in a circuit. **2.** For a device, the ratio of input voltage to output voltage when the output voltage is the lower quantity.

voltage-balance relay A relay actuated by a voltage differential.

voltage breakdown See BREAKDOWN VOLTAGE.

voltage-breakdown test **1.** A test in which the measured voltage applied to an insulating material is continuously increased until the breakdown point is reached. **2.** A test in which the measured reverse voltage applied to a semiconductor junction is continuously increased until the reverse breakdown point is reached (see AVALANCHE BREAKDOWN).

voltage burden The voltage drop across a CURRENT SHUNT.

voltage calibrator A device used to calibrate, in terms of voltage, a meter scale, oscilloscope screen, graphic-recorder chart, etc.

voltage-capacitance curve A plot depicting the variation of capacitance with applied voltage for a voltage-variable capacitor. For a varactor (variable-capacitance diode), capacitance varies inversely with reverse direct-current voltage.

voltage coefficient A figure that shows the extent to which a quantity drifts under the influence of voltage. It is generally expressed in percent per volt or in parts per million per volt (ppm/V).

voltage coefficient of capacitance For a voltage-dependent capacitor, the capacitance change per unit change in applied voltage.

voltage coefficient of resistance For a voltage-dependent resistor, the resistance change per unit change in applied voltage.

voltage coil See POTENTIAL COIL.

voltage comparator A device for comparing (usually only two) voltages. The various types range from simple, manually balanced potentiometers to analog or digital devices that automatically compare the applied voltages and provide a direct readout of either their difference or the percent of unbalance.

voltage control **1.** A component or circuit that allows the adjustment of the output voltage of a power supply within a given range. **2.** The adjustment of the output voltage of a power supply to optimize the performance of a circuit connected to the supply. **3.** Any form of circuit control that is accomplished by the adjustment of the voltage at a given circuit point.

voltage-controlled amplifier Abbreviation, VCA. An amplifier in which gain is controlled by means of a voltage applied to a control terminal.

voltage-controlled attenuator An attenuator circuit in which a transistor serves as a voltage-variable resistor. The output resistance of the transistor varies inversely with the direct-current bias voltage applied to the input electrode (base, emitter, or gate).

voltage-controlled capacitor See VOLTAGE-DEPENDENT CAPACITOR.

voltage-controlled crystal oscillator Abbreviation, VCCO or VCXO. A voltage-controlled oscillator of the crystal-stabilized type.

voltage-controlled generator Abbreviation, VCG. Any signal-generating device whose output frequency is varied by changing one of the direct-current operating voltages of the device.

voltage-controlled oscillator Abbreviation, VCO. An oscillator of the inductance-capacitance (LC) type that employs a varactor diode in the capacitive portion of the tuned circuit. The varactor is placed in series or parallel with the tuning capacitor and is isolated for direct current by blocking capacitors. The frequency can be adjusted, within certain limits, by nonmechanical means. The signal can also be frequency modulated by applying the modulating voltage across the varactor.

voltage-controlled resistor See VOLTAGE-DEPENDENT RESISTOR.

voltage corrector See VOLTAGE REGULATOR.

voltage crest See VOLTAGE PEAK.

voltage-current characteristic See VOLTAGE-CURRENT CURVE.

voltage-current crossover The point at which a voltage-regulated supply becomes a current-regulated supply.

voltage-current curve For a circuit or device, the plot of the interrelationship of current and voltage, with voltage as the independent variable.

voltage-current feedback See CURRENT-VOLTAGE FEEDBACK.

voltage decay The exponential decrease of voltage across a discharging capacitor. Also see EXPONENTIAL DECREASE.

voltage-dependent capacitor A capacitor (such as a varactor) whose capacitance varies with applied voltage.

voltage-dependent resistor A nonlinear resistor whose value varies inversely with the voltage drop across it. Also called *varistor.*

voltage detector A circuit or device that delivers an output voltage only when the input voltage is of a prescribed value. Compare VOLTAGE DISCRIMINATOR.

voltage-determined property A property (capacitance, current, frequency, or resistance) whose magnitude depends on the value of an applied voltage. See, for example, VOLTAGE-DEPENDENT CAPACITOR, VOLTAGE-DEPENDENT RESISTOR, and VOLTAGE-CONTROLLED OSCILLATOR.

voltage differential Symbol, dE or dV. An infinitesimal change in voltage. Compare VOLTAGE INCREMENT.

voltage-directional relay **1.** A relay that is actuated when voltage exceeds a certain value in a given direction. **2.** A relay that closes only when the applied voltage is in a specific direction.

voltage discriminator A circuit or device whose output voltage is zero when the input voltage is of a prescribed value. When the input voltage is greater than this value, the output is positive; when it is less, the output is negative.

voltage distribution **1.** The delivery of operating voltage throughout a circuit (e.g., high and low direct-current voltages in the various stages of a control circuit). **2.** Sometimes, the distribution of electrical energy ("power" distribution).

voltage divider A resistive or capacitive potentiometer or network used to divide an applied voltage by a desired amount.

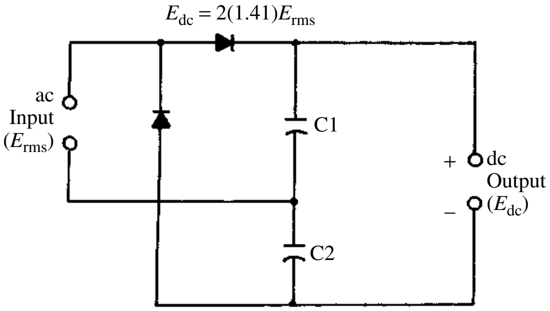

voltage doubler

voltage doubler A power-supply circuit that supplies a direct-current output voltage of about twice the peak value of the alternating-current input voltage.

voltage drift See DRIFT VOLTAGE.

voltage drop Abbreviation, VD. The voltage (*E*) set up across a resistance (*R*) carrying a current (*I*); $E = IR$. For alternating current, reactance *X* and impedance *Z* can be substituted for resistance, where applicable.

voltage-equalizing resistors In a power-supply filter, resistors connected across each capacitor in a string of electrolytics connected in series to withstand a high voltage. The resistors protect the capacitors by equalizing the voltage across them.

voltage-fed antenna An antenna in which the feeder is attached to the radiator at a voltage loop (current node). Compare CURRENT-FED ANTENNA.

voltage feed The delivery of voltage to a device or circuit at a point where voltage, rather than current, is dominant. Compare CURRENT FEED.

voltage feedback A feedback signal consisting of voltage fed from the output to the input circuit of an amplifier or other device. Compare CURRENT FEEDBACK and CURRENT-VOLTAGE FEEDBACK.

voltage-frequency transducer **1.** A device whose output voltage is a function of the frequency of a signal at the input. **2.** A device whose output frequency is a function of the voltage at the input.

voltage gain See VOLTAGE AMPLIFICATION.

voltage generator See VOLTAGE SUPPLY.

voltage gradient The voltage per unit length along a defined path.

voltage increment A change in voltage represented by $E_2 - E_1$, where E_2 and E_1 are the voltages at two defined points on a curve, and the difference is finite (nonzero). Compare VOLTAGE DIFFERENTIAL.

voltage input encoder An analog-to-digital encoder for which the input is an analog voltage. See, for example, VOLTAGE-TO-SHAFT-POSITION ENCODER.

voltage inverter A circuit or device whose voltage output has the sign opposite that of the input voltage. The device might or might not provide amplification.

voltage jump **1.** An upward transient in voltage. **2.** In a glow-discharge tube, a sudden break or increase in voltage drop across the tube.

voltage lag The condition in which voltage changes occur after corresponding current changes. Compare CURRENT LAG and VOLTAGE LEAD.

voltage lead The condition in which voltage changes precede corresponding current changes. Compare CURRENT LEAD and VOLTAGE LAG.

voltage level **1.** A prescribed reference value of voltage (e.g., the black level in a television picture signal). **2.** The discrete value of a steady voltage, or the average value of a fluctuating voltage, as observed or measured in a circuit.

voltage limit The maximum or minimum level in a voltage range.

voltage loop In a standing-wave system, such as an antenna or transmission line, a maximum-voltage point. Compare VOLTAGE NODE.

voltage loss **1.** Reduction of a voltage across a load, occurring because of a series resistance. **2.** The ratio, in decibels, between the input voltage to a transmission line and the output voltage at the load end of the line, assuming that the impedance is the same at both points.

voltage maximum See VOLTAGE PEAK.

voltage minimum See VOLTAGE TROUGH.

voltage-mode switching circuit A resistor-transistor-logic (RTL) NAND or NOR circuit in which (in the off state) the transistor is cut off by the V_{BB} bias voltage. The output is then approximately equal to the collector supply voltage, V_{CC}. The proper combination of input pulses overrides the cutoff bias, and the transistor switches on. The output then drops to a level equal to V_{CC} minus the voltage drop across the external collector resistor.

voltage multiplier A special type of rectifier circuit that delivers a direct-current output voltage that is a multiple of the peak value of the alternating-current input voltage, thus affording voltage step-up without a transformer. See, for example, VOLTAGE DOUBLER; VOLTAGE QUADRUPLER; QUINTUPLER, **1**; and VOLTAGE TRIPLER.

voltage node In a standing-wave system, such as an antenna or transmission line, a minimum-voltage point. Compare VOLTAGE LOOP.

voltage of self-induction The voltage drop across an inductor, resulting from the flow of alternating current through the inductor; it is caused by self-induction.

voltage peak The highest value attained by a voltage during an excursion. Also called *voltage crest* or *voltage maximum.* Compare VOLTAGE TROUGH.

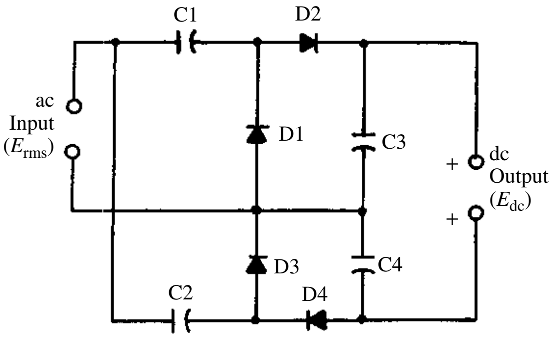

$E_{dc} = 4(1.41)E_{rms}$

voltage quadrupler

voltage-power directional relay A relay system in which two circuits are connected when their voltage difference exceeds a predetermined value in one direction, and are disconnected when the voltage in the opposite direction exceeds a predetermined level.

voltage quadrupler A special rectifier circuit that supplies a direct-current output voltage of approximately four times the peak alternating-current input voltage.

voltage quintupler See QUINTUPLER, **1**.

voltage-range multiplier **1.** A multiplier resistor (see VOLTMETER MULTIPLIER) connected in series with a voltmeter that has an internal multiplier, to increase the range of the instrument. **2.** For an alternating-current voltmeter, an input amplifier used to increase the sensitivity of the instrument.

voltage rating **1.** For a circuit or device, the recommended maximum voltage that can be applied, or the recommended working voltage, as specified. **2.** For a generator, the specified output voltage.

voltage ratio The quotient of the voltages, E_1/E_2, at two specific points in a circuit, device, or system. Examples: ratio of input voltage to output voltage and ratio of primary voltage to secondary voltage.

voltage-ratio box See VOLT BOX.

voltage reference See STANDARD CELL and ZENER-DIODE VOLTAGE REFERENCE.

voltage-reference cell See STANDARD CELL.

voltage-reference diode A Zener diode biased into its Zener region. The voltage drop across the diode is comparatively constant. Under proper conditions, it can be used for reference purposes. Also see ZENER-DIODE VOLTAGE REFERENCE.

voltage reflection coefficient In a reflected-wave situation, the ratio E_r/E_i, where E_i is the field-strength voltage of the incident wave, and Er is the field-strength voltage of the reflected wave.

voltage-regulated supply See CONSTANT-VOLTAGE SOURCE.

voltage-regulating transformer A special transformer in which a resonant circuit and core saturation (see SATURATED OPERATION, **1**) are used to provide a constant output voltage.

voltage regulation The stabilization of a voltage against fluctuations in source or load.

voltage-regulation constant Symbol, K_v. For a voltage-regulated power supply, the ratio dE_L/dE_i, where dE_i is a change in input voltage, and dE_L is the corresponding change in load voltage.

voltage regulator A circuit or device that holds an output voltage constant during variations in the output load or input voltage.

voltage-regulator diode See ZENER-DIODE VOLTAGE REGULATOR.

voltage-regulator transformer See VOLTAGE-REGULATING TRANSFORMER.

voltage relay A relay or relay circuit that is actuated by a discrete voltage, rather than by current or power.

voltage-responsive device See VOLTAGE-ACTUATED DEVICE.

voltage rise The normal condition in a series-resonant circuit, in which the voltage across the coil or capacitor is higher than the voltage applied to the circuit.

voltage saturation In a current-actuated device, such as a bipolar transistor, the situation in which an increase in current provides no increase in voltage drop beyond a certain point (the saturation point).

voltage-sensitive bridge A bridge having a nonlinear element (such as a voltage-dependent resistor) as one of the arms. Because of this element, the bridge can be balanced (with a given set of other arms) at only one value of applied voltage. At lower voltages, the bridge becomes unbalanced in one direction; at higher voltage, in the opposite direction.

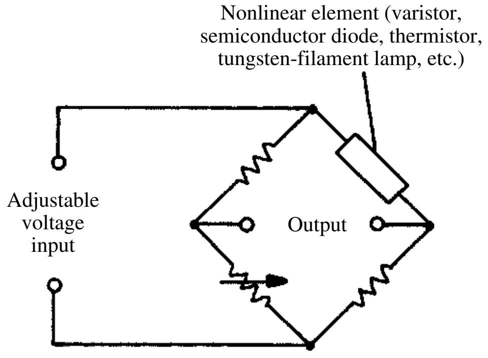
Nonlinear element (varistor, semiconductor diode, thermistor, tungsten-filament lamp, etc.)
Adjustable voltage input
Output

voltage-sensitive bridge

voltage-sensitive capacitor See VOLTAGE-DEPENDENT CAPACITOR.

voltage-sensitive resistor See VOLTAGE-DEPENDENT RESISTOR.

voltage sensitivity **1.** Particular responsiveness of a circuit or device to voltage, rather than current or power. See, for example, VOLTAGE-DEPENDENT CAPACITOR, VOLTAGE-DEPENDENT RESISTOR, and VOLTAGE RELAY. **2.** See VOLTMETER SENSITIVITY.

voltage-spectrum function The voltage-vs.-frequency curve at the output of a circuit or transducer.

voltage-stabilized supply See CONSTANT-VOLTAGE SOURCE.

voltage stabilizer See VOLTAGE REGULATOR.

voltage-stabilizing diode See ZENER-DIODE VOLTAGE REGULATOR.

voltage standard A device that delivers a voltage of accuracy and stability so that it can be used to calibrate other voltage generators and test instruments. See, for example, STANDARD CELL and ZENER-DIODE VOLTAGE REFERENCE.

voltage standing-wave ratio Abbreviation, VSWR. In a standing-wave system, the ratio of the maximum voltage to the minimum voltage.

voltage supply A power supply whose usable output is voltage, rather than current or power. When such a supply is not voltage-regulated, it can only be used reliably with a very light load.

voltage to chassis In electronic equipment mounted on a metal chassis, the voltage between the chassis and a given point in the circuit.

voltage-to-frequency converter A device or circuit that delivers an output frequency proportional to an input voltage (usually direct current). Compare FREQUENCY-TO-VOLTAGE CONVERTER.

voltage to ground **1.** In a circuit, the voltage measured between a given point and the ground point. **2.** The voltage measured between the earth and a line or piece of equipment.

voltage to panel In electronic equipment mounted on a metal panel, the voltage between the panel and a given point in the circuit.

voltage-to-shaft-position encoder An encoder for which the output is the rotation of a motor shaft over an arc proportional to an input voltage.

voltage transformer **1.** A transformer used primarily to supply voltage with little or no current. **2.** A small step-up transformer for increasing the sensitivity of an alternating-current voltmeter. Also called POTENTIAL TRANSFORMER.

voltage tripler A rectifier circuit that, without a transformer, supplies a direct-current output voltage of approximately three times the peak value of the alternating-current input voltage.

voltage trough The lowest value reached by a voltage during an excursion. Compare VOLTAGE PEAK.

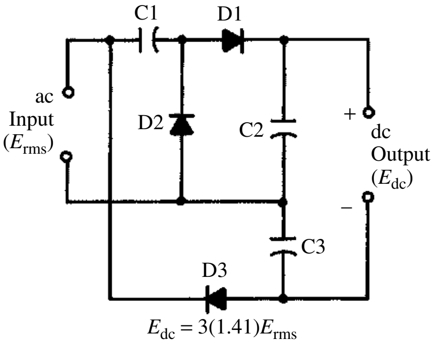

voltage tripler

voltage tuning A method of adjusting the frequency of an oscillator or resonant circuit by means of a variable, direct-current voltage.

voltage-tunable magnetron A magnetron oscillator in which the output frequency is varied by adjusting the direct-current voltage on the tube.

voltage-tunable oscillator See VOLTAGE-CONTROLLED OSCILLATOR.

voltage-turns ratio The turns ratio of a transformer, indicated by the ratio of primary voltage to secondary voltage, or vice-versa.

voltage-type telemetry Telemetry based upon the variation of a single voltage in sympathy with the changes in a sensed phenomenon.

voltage-variable capacitor **1.** A specially processed semiconductor diode of which the voltage-variable capacitance of the junction is utilized. Also called VARACTOR. **2.** A capacitor having a specially processed nonlinear dielectric, such as barium strontium titanate, whose capacitance varies inversely with the applied direct-current bias voltage.

voltage-variable resistor See VOLTAGE-DEPENDENT RESISTOR.

voltage vector In a vector diagram, a vector showing the magnitude and phase of a voltage. Compare CURRENT VECTOR.

voltaic Pertaining to chemically produced direct current. Sometimes interchangeable with *galvanic.*

voltaic cell For the generation of a direct-current voltage, a primary cell consisting of two electrodes made of different metals and immersed in an electrolyte. Also called GALVANIC CELL. Also see CELL, ELECTROMOTIVE SERIES, and PRIMARY CELL.

voltaic couple A pair of dissimilar metals (or other substances) that generate a direct-current voltage when they contact an electrolyte.

voltaic pile A rudimentary primary battery consisting of a series of disks made of two different metals, stacked alternately with electrolyte-soaked cloth or paper disks.

voltaic series See ELECTROMOTIVE SERIES.

voltameter An electrolytic cell for determining the value of an unknown current, or of an unknown quantity of electricity, from the weight of metal deposited out of an electrolyte onto the cathode by the passage of current over an accurately timed interval.

volt-ammeter A combination meter for measuring electrical voltages and currents.

volt-ampere Symbol, VA. Unit, watt. The simple product of voltage and current in volts and amperes, yielding the true power in a direct-current circuit and the apparent power in an alternating-current circuit. Also see APPARENT POWER and TRUE POWER.

volt-ampere-hour meter See VAR-HOUR METER.

volt-ampere meter See VAR METER.

volt-amperes reactive Abbreviation, VAR. The product of volts and amperes for a purely reactive circuit. This product does not produce the true power because the power factor is neglected; it produces only the apparent power. A true reactance absorbs power during one half-cycle of alternating current, and returns it to the generator during the next half-cycle.

Volta's law See VOLTA'S PRINCIPLE.

Volta's pile See VOLTAIC PILE.

Volta's principle Two dissimilar metals brought into contact (even in air) will generate a difference of potential whose value is characteristic of the metals. Also see ELECTROMOTIVE SERIES.

volt box A precision, potentiometer-type voltage divider used in the calibration of meters and other instruments. The device is usually provided with a set of terminal posts for selecting various ratios of output voltage to input voltage.

volt-electron See ELECTRONVOLT.

voltmeter A usually direct-reading instrument used to measure voltage. Also see ELECTRONIC VOLTMETER, FET VOLTMETER, TRANSISTOR VOLTMETER, and VACUUM-TUBE VOLTMETER.

voltmeter-ammeter See VOLT-AMMETER.

voltmeter amplifier A wideband, flat-frequency-response, low-distortion preamplifier used to boost the sensitivity of an alternating-current voltmeter.

voltmeter-millivoltmeter A voltmeter that provides several low ranges, as well as several high ones. A familiar example is an alternating-current voltmeter, which has various full-scale ranges from about 1 millivolt to about 1000 volts.

voltmeter multiplier A resistor connected in series with a current meter (usually a milliammeter or microammeter) to convert it into a voltmeter.

voltmeter sensitivity Unit, ohm per volt. For a voltmeter, the total resistance of the instrument (multiplier resistance plus the resistance of the meter movement) divided by the full-scale deflection of the meter. Thus, a 0-to-10 direct-current voltmeter with an input resistance of 100 kilohms has a sensitivity of 10,000 ohms per volt.

volt-milliammeter A combination meter for measuring volts and milliamperes.

volt-ohm-ammeter A multimeter for measuring voltage, resistance, and current (in amperes).

volt-ohmmeter A combination meter for measuring voltage and resistance.

volt-ohm-milliammeter Abbreviation VOM. A multimeter for measuring voltage, resistance, and current (in milliamperes and microamperes).

voltsensor See VOLTAGE DETECTOR.

volts per meter Abbreviation, V/m or Vpm. The unit of electric field strength.

volume **1.** Intensity of sound. Also called LOUDNESS. **2.** A circumscribed portion of space, either imaginary or actually occupied, and described by three dimensions (e.g., sphere, ellipsoid, cube, pyramid, etc.). **3.** In a computer system, a unit of magnetic storage medium (e.g., a diskette).

volume compression The automatic reduction of the gain of an audio amplifier. The process differs from clipping (which "slices off" the tops of waves) in that compression (ideally) reduces the amplitude while preserving the waveform. Compare VOLUME EXPANSION.

volume compressor A circuit or device (such as an automatic-gain-control amplifier) for achieving volume compression. Compare VOLUME EXPANDER.

volume conductivity The reciprocal of VOLUME RESISTIVITY.

volume control A variable resistor (usually a potentiometer) or a network of resistors (such as a pad) for adjusting the gain, and, consequently, the output-signal loudness, of an amplifier.

volume equivalent In wire telephony, speech loudness throughout the system, expressed in terms of the trunk loss in a reference system and adjusted for equal loudness.

volume expander A circuit or device for automatically boosting the volume of an audio-frequency signal. Also see EXPANDER. Compare VOLUME COMPRESSOR.

volume expansion The technique of automatically increasing the gain, and consequently the output-signal volume, of an audio amplifier. Also see VOLUME EXPANDER. Compare VOLUME COMPRESSION.

volume indicator A fast-acting alternating-current meter used to monitor the volume level in an audio channel in which the signal level is fluctuating. The scale is graduated in VOLUME UNITS. Also called *VU meter.*

volume lifetime In a homogeneous semiconductor, the interval between minority-carrier generation and recombination.

volume limiter A circuit or device that automatically holds the volume level of an audio channel to a predetermined maximum. Also see VOLUME COMPRESSION and VOLUME COMPRESSOR.

volume-limiting amplifier An amplifier that functions as a volume limiter through the action of volume-limiting subcircuits.

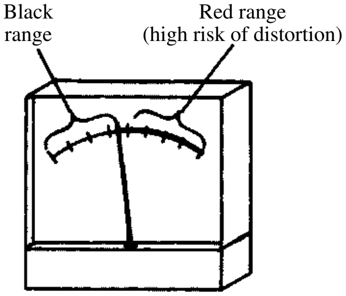

volume indicator

volume magnetostriction **1.** The decrease in the total volume of certain ferromagnetic substances under the influence of a magnetic field. **2.** The increase in the total volume of certain ferromagnetic substances under the influence of a magnetic field.

volume range Unit, decibel (dB). The difference between the maximum and minimum volume levels that a device or system can accommodate.

volume recombination rate Within the volume of a semiconductor, the rate at which electrons and holes recombine.

volume resistance The effective resistance, through a given medium, between two electrodes placed within that medium.

volume resistivity The resistance of a specific volume of a material (e.g., the resistance between opposite faces of a 1-centimeter cube of the material). Also see MICROHM-CENTIMETER; OHM-CENTIMETER; RESISTIVITY.

volumetric efficiency In an electronic assembly, the ratio of the volume of parts to the total volume of the assembly.

volumetric radar A radar providing a three-dimensional display.

volumetric sensor In security systems, a device that detects effects in a specifically defined three-dimensional region, such as an entire room (from floor to ceiling).

volume unit Abbreviation, VU. The unit of measurement of fluctuating alternating-current power, such as that of speech or music. Zero VU corresponds to a reference power of 2.51 milliwatts, or +4 dBm (four decibels above 1 milliwatt). Volume units are measured with a VOLUME INDICATOR.

volume-unit indicator See VOLUME INDICATOR.

volunteer examiner In amateur radio, a person who works with the Federal Communications Commission (FCC) to administer license examinations. Such a person is not actually an employee of the FCC, but serves on an independent, volunteer basis.

VOM Abbreviation of VOLT-OHM-MILLIAMMETER.

von Hippel breakdown theory The theory that, assuming no electron-energy distribution, breakdown occurs at field intensities for which the recombination rate of electrons and holes is lower than the rate of ionization by collision.

von Neumann architecture In computer systems, a design scheme in which the bits are transmitted serially (one by one) along a single line or bus.

VOR Abbreviation of *very-high-frequency omnirange* (see VHF OMNIRANGE).

voter See MAJORITY LOGIC.

vox Abbreviated form of VOICE-OPERATED CONTROL.

V_P **1.** Symbol for GATE-SOURCE PINCHOFF VOLTAGE of a field-effect transistor. **2.** Symbol for PLATE VOLTAGE.

V_p Symbol for VOLTAGE AT PEAK TORQUE.

V particle See VEE PARTICLE.

Vpm Abbreviation of VOLTS PER METER. (Also, V/m.)

VR **1.** Abbreviation of VIRTUAL REALITY. **2.** Abbreviation of VOLTAGE REGULATOR.

VRAM Abbreviation of VIDEO RANDOM-ACCESS MEMORY.

V_{ref} Abbreviation of *reference voltage.*

vrr Abbreviation of *visual radio range.*

V•s Abbreviation of *volt-seconds* (webers).

V•s/A Abbreviation of *volt-seconds per ampere* (henrys).

VSA Abbreviation of VOICE-STRESS ANALYZER.

VSB Abbreviation of VESTIGIAL SIDEBAND.

VSF Abbreviation of VESTIGIAL-SIDEBAND FILTER.

V signal See VEE SIGNAL.

vsr Abbreviation of VERY SHORT RANGE.

VSWR Abbreviation of VOLTAGE STANDING-WAVE RATIO.

vt **1.** Abbreviation of VACUUM TUBE. **2.** Abbreviation of *variable time.*

VT fuse See PROXIMITY FUSE.

VTL Abbreviation of VARIABLE-THRESHOLD LOGIC.

vtm Abbreviation of VOLTAGE-TUNABLE MAGNETRON.

VTO Abbreviation of VOLTAGE-TUNABLE OSCILLATOR.

VTR Abbreviation of VIDEOTAPE RECORDER or VIDEOTAPE RECORDING.

VTVM Abbreviation of VACUUM-TUBE VOLTMETER.

vt voltmeter See VACUUM-TUBE VOLTMETER.

VU Abbreviation of VOLUME UNIT.

vulcanized fiber A tough insulating material derived from cellulose. Dielectric constant, 5 to 8.

VU meter See VOLUME INDICATOR.

VVCD Abbreviation of *voltage-variable capacitor diode* (see VOLTAGE-VARIABLE CAPACITOR, **1**).

VVV signal See VEE SIGNAL.

VW Abbreviation of *volts working* (see WORKING VOLTAGE).

vy Abbreviation for *very.*

W

W **1.** Symbol for WORK. **2.** Abbreviation of WATT. **3.** Symbol for TUNGSTEN. **4.** Symbol for ENERGY. **5.** Abbreviation of *west.* **6.** Abbreviation of WIDTH.

w **1.** Abbreviation of WEIGHT. (Also, wt.) **2.** Abbreviation of *week.*

WAC Amateur radio abbreviation of *Worked All Continents*, an award given to operators who have engaged in verified two-way communication with stations on all continents.

wafer **1.** Semiconductor die. **2.** A thin, flat disk, ring, or plate around which the contacts of a rotary switch are spaced. **3.** A thin square or rectangle of dielectric material used as the dielectric member in a fixed capacitor. **4.** A plate cut from a crystal (e.g., a quartz wafer).

wafer fabrication The various processes used in the manufacture of semiconductor integrated circuits.

wafer slicing Cutting plates from a mother crystal, as when piezoelectric plates are cut from a quartz crystal.

wafer socket A component socket consisting of a plastic or ceramic wafer with spring-type contacts for gripping the pins.

wafer switch A rotary switch whose contacts are arranged around the periphery of a plastic or ceramic wafer.

Wagner ground A circuit (often a single potentiometer) that facilitates cancellation of stray reactance in an alternating-current bridge. The bridge is balanced alternately with the bridge-balance control and the Wagner control, until there is no further shift of null point when changing from one to the other.

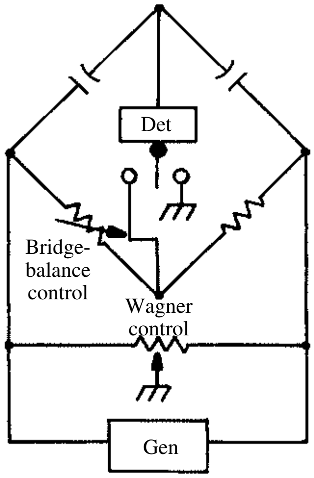

Wagner ground (with capacitance bridge)

waiting time **1.** See WARMUP TIME. **2.** The delay between the time a data transfer to or from a computer memory is called for and the actual transfer of the data. Also called LATENCY.

walkie-lookie A portable combination camera and transmitter for remote television pickup. At sports events and other gatherings, the unit is strapped to the shoulder of the camera operator.

walkie-talkie A portable, compact transceiver (or transmitter-receiver).

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

wall absorption Diminished radioactive emission (especially of beta and gamma rays) because of absorption by the radiating substance.

wall box A wall-mounted (usually metal) box enclosing circuit breakers, fuses, switches, etc.

wall effect In an ionization chamber, the rise in ionization because of electrons being released by the walls of the chamber.

wall energy In a ferromagnetic substance, the energy per unit area stored in a domain wall between two regions of opposite magnetization.

Wallman amplifier A cascode amplifier (see CASCODE).

wall mount A metal bracket for fastening an antenna to a wall.

wall outlet A plug or socket, usually mounted in a protective box or can and recessed in a wall that can be accessed from the front. It is commonly used for supplying alternating-current power to appliances at 117 volts or 234 volts. It is also used in telephone and cable-television systems.

wall plaque A loudspeaker so thin that, mounted in a frame (sometimes behind grill cloth), it can be hung on a wall.

wall plate A (usually rectangular) plate of metal or plastic for holding a wall outlet or wall switch.

wall plug A male or female plug usually mounted in a protective box or can and recessed in a wall. Such a device can provide easy access to an antenna, telephone line, or load; or it can deliver alternating-current, direct-current, or radio-frequency power.

walls The sides of the groove cut into a record disc.

wall socket A male or female socket usually mounted in a protective box or can and recessed in a wall. Such a device can provide easy access to an antenna, telephone line, or load; or it can deliver alternating-current, direct-current, or radio-frequency power.

wall speaker See WALL PLAQUE.

wall switch A switch usually mounted in a protective box or can and recessed in a wall.

wall telephone A wall-mounted telephone set. It generally fits over the jack, so the only cord is between the main unit and the receiver.

wall-through tube See LEAD-IN TUBE.

Walmsley antenna A phased array consisting of several full-wavelength loops.

wamoscope A radar display tube that performs several microwave-receiver functions (detection, oscillation, amplification, etc.) as well as displaying an image. The name is an acronym for *wave-modulation oscilloscope.*

wander See SCINTILLATION, **1**.

WAP Acronym for WIRELESS ACCESS PROTOCOL.

warble **1.** A periodic rise and fall in the pitch of a musical tone or combination of tones. **2.** To rise and fall in pitch with a definite period.

warble-tone generator An audio-frequency oscillator whose frequency is varied at a subaudible rate over a fixed frequency range.

warm junction The heated junction in a two-junction thermocouple circuit. Also called HOT JUNCTION. Compare COLD JUNCTION.

warmup The process of stabilizing an electronic equipment by allowing its temperature to rise to the optimum level.

warmup time The time required for an electronic circuit to become fully operational, or to stabilize, after the power has been switched on.

warning bell (buzzer) A bell (or buzzer) used as an audible alarm in a WARNING DEVICE.

warning device A device, such as an electronic signaler, for alerting a person to an emergency (intrusion on premises, danger to life and safety, etc.) or to the existence of an intrusion into a secured area.

warning lamp See WARNING LIGHT.

warning light A lamp used as a visual alarm in a warning device.

warpage Distortion of the normally straight sides of a triangular wave.

warping In multimedia computer graphics, a change in the shape of an object that occurs smoothly over a period of time.

WAS Amateur radio abbreviation of *Worked All States*, an award given to operators who have engaged in verified two-way communication with stations in all states of the United States.

washer capacitor A very thin donut capacitor.

washer resistor A resistor made in the general shape of a washer or ring and having a center hole for a mounting screw or stacking rod.

washer thermistor A thermistor made in the general shape of a washer and having a center hole for a mounting screw or stacking rod.

washer varistor A varistor made in the general shape of a washer and having a center hole for a mounting screw or stacking rod.

washout process A method of fabricating bipolar transistors. The contact metal is deposited in the diffusion hole.

waste instruction In a computer program, an instruction not meant to be acted upon (e.g., one used to take up space in the listing for some reason). Also called *null instruction* and *dummy instruction.*

watch A work shift, as of electronic personnel (e.g., radio station operators).

watchcase receiver An earphone enclosed in a small, round case with a screw-on cap. It is derived from its resemblance to a large pocket watch.

water absorption For a solid material, such as a dielectric, the ratio of the weight of water absorbed by the material to the weight of the material.

water-activated battery A battery that contains all the ingredients of its electrolyte, except water, which must be added when the battery is put into service.

water adsorption The formation of a thin layer of water molecules on the surface of normally dry material, but not by absorption. Also see ADSORPTION.

water analogy The useful, but not wholly accurate, teaching device of comparing an electric current with the flow of water. In such a comparison, voltage is shown equivalent to water pressure, and current to the quantity (e.g., gallons) of water flowing in unit time.

water battery A primary battery or cell using water as the electrolyte.

water calorimeter A calorimeter used to measure power in terms of the increase in temperature of water heated by the electrical energy.

water capacitor An emergency capacitor made by setting one glass of water in another larger jar of water so that the two bodies of water are separated by the walls of the smaller jar. The bodies of water, in which an electrolyte has been dissolved, form the "plates" of the capacitor, and the wall of the smaller jar serves as the dielectric between them.

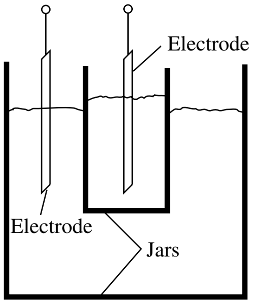

water capacitor

water-cooled tube A power tube, such as a large radio transmitting tube, cooled by the circulation of water in the space between the outside of the tube envelope and a surrounding jacket.

water cooling A method of cooling components by pumping water through pipes surrounding them.

water-flow alarm An electronic circuit that actuates an alarm when the flow of water through pipes or other channels changes from a predetermined rate.

water-flow control A servo system for automatically maintaining or adjusting the flow of water through pipes or other channels.

water-flow gauge See WATER-FLOW METER.

water-flow indicator See WATER-FLOW METER.

water-flow meter An instrument used to monitor the flow of water through pipes or other channels, sometimes showing its direction as well as its rate.

water-flow switch In water-cooled systems (e.g., water-cooled tubes), a switch that actuates an alarm when the water slows or stops.

water glass Sodium silicate, a substance used as a fireproofing agent and protective coating.

water ground An earth connection made by dropping a weighted wire into a body of water.

water jacket In a water-cooled device, the outer jacket that, along with the outer wall of the cooled component, forms a space through which the cooling water flows.

water-level control A servo system that automatically maintains or adjusts the level of water inside a tank or other container.

water-level gauge An electronic system that gives direct water level readings inside a tank or other container.

water-level indicator See WATER-LEVEL GAUGE.

water load **1.** A makeshift, power-dissipating resistive load (see DUMMY LOAD) consisting of a container of tap water or saltwater into which two wires are immersed. **2.** A waveguide termination containing water heated by the microwave energy. It is usable as a water calorimeter.

water monitor A sensitive electronic instrument for checking radioactivity in a water supply.

water-pipe ground An earth connection made by running a wire to the nearest cold-water pipe. This scheme works only with metal pipes devoid of insulating joints or splices.

water power Hydroelectric power (i.e., electrical energy produced by generators driven by water).

water-pressure alarm An electronic circuit that actuates an alarm when water in pipes or other channels changes from a predetermined level.

water-pressure control A servo system for automatically maintaining or adjusting water pressure in pipes or other channels.

water-pressure gauge See WATER-PRESSURE METER.

water-pressure indicator See WATER-PRESSURE METER.

water-pressure meter An instrument that directly indicates water pressure in a pipe or tank.

water-pressure switch A switch that actuates an alarm when water pressure rises or falls.

water pump In a water-cooled electronic system, the (usually rotary) pump that circulates the water.

water resistor An electrolytic resistor in which the electrolyte is tap water or diluted saltwater.

water rheostat A variable water resistor. Usually, the resistance is varied by moving the immersed electrodes closer together or farther apart.

water tester An instrument for checking pH, electrical resistance, and other properties of water.

water witching Locating underground water by electronic methods.

WATS Acronym for *Wide Area Telephone Service.* A form of long-distance telephone service. Rates are charged on a different basis than normal long-distance service. The system is especially favored by businesses because it saves money for subscribers making a large number of calls in each billing period.

watt Abbreviation, W. The practical unit of electric and other power. One watt is dissipated by a resistance of 1 ohm through which a current of 1 ampere flows. See also KILOWATT, MEGAWATT, MICROWATT, and MILLIWATT.

wattage Electrical power, especially when expressed in watts.

wattage rating **1.** The recommended output power of a device. **2.** The recommended power dissipation of a device.

watt current The component of alternating current that is in phase with voltage. Also called RESISTIVE CURRENT. Compare REACTIVE CURRENT.

watt-decibel conversion The conversion of a power level, such as the power output of an amplifier, in watts to the corresponding power level in decibels, with respect to a reference level. Thus, n dB = 10 $\log_{10}(P/P_{ref})$, where P is the power of interest (watts), and P_{ref} is the reference level (e.g., one milliwatt).

watt-hour Abbreviation, WH. The unit of electrical energy or work; 1 WH = 3600 joules = 10^{-3} kWH. Also see ENERGY, KILOWATT-HOURS, POWER, and WATT-SECOND.

watt-hour capacity The number of watt-hours that a storage battery can deliver reliably and safely under specified operating conditions.

watt-hour-demand meter A combination watt-hour meter and demand meter.

watt-hour efficiency For a storage battery, the ratio of watt-hours output to watt-hours of recharge.

watt-hour constant In an electric-energy meter, the number of watt-hours in one revolution of the indicating disk.

watt-hour meter An instrument for measuring electrical energy in watt-hours. One well-known type consists essentially of a small motor geared to a row of four dial indicators. An eddy-current disk keeps the motor speed proportional to the watt-hours consumed by a load, a value which is the sum of the readings of the dials. Also called SERVICE METER and KILOWATT-HOUR METER.

wattless current The component of alternating current that is out of phase with voltage. Also called REACTIVE CURRENT. Compare RESISTIVE CURRENT.

wattless power The apparent power in a reactive circuit, indicated by the product of volts and amperes. There is no actual power consumption because the power taken by a reactance during a half-cycle is returned to the generator during the next half-cycle. Also see AC POWER, APPARENT POWER, REACTIVE KILOVOLT-AMPERES, and REACTIVE VOLT-AMPERES.

wattless volt-amperes See WATTLESS POWER.

wattless watts See WATTLESS POWER.

wattmeter Abbreviation, WM. An instrument used to measure electrical power. The scale usually reads directly in watts, kilowatts, milliwatts, or microwatts. Also see ELECTRONIC WATTMETER.

watt-second Abbreviation, Ws. A small unit of electrical energy or work; 1 watt-second = 1 joule = 1/3600 watt-hour. Also see ENERGY and POWER.

watt-second constant In an electric-energy meter, the number of watt seconds in one revolution of the indicating disk.

wave **1.** A single oscillation of some property of matter or space, such as density, displacement, or field strength. It moves outward from a point of disturbance and grows weaker as it travels farther. Wave motion is associated with mechanical vibration, sound, radio, heat, light, X rays, gamma rays, and cosmic rays. See SOUND and WAVELENGTH. **2.** A single cycle of alternating or pulsating current or voltage. Also see AC VOLTAGE, ALTERNATING CURRENT, PULSATING DC VOLTAGE, and PULSATING DIRECT CURRENT.

wave absorption The removal of energy from electromagnetic waves as they pass through certain media, such as solid bodies, water, and the atmosphere. Compare POLARIZATION, WAVE REFLECTION, and WAVE REFRACTION.

wave amplitude The peak value of a wave. Also see WAVE CREST and WAVE TROUGH.

wave analyzer An instrument consisting essentially of a continuously tunable bandpass filter and an electronic alternating-current voltmeter. As the filter is tuned successively to the fundamental frequency of a complex wave and to its various harmonics, the voltmeter shows the amplitude of each of the components. Also see HETERODYNE WAVE ANALYZER.

wave angle The angle, measured with respect to the horizon, at which a radio wave is transmitted or received.

wave antenna See BEVERAGE ANTENNA.

wave attenuation The reduction of wave amplitude, with respect to distance from the source.

waveband A band of radio frequencies. Also called FREQUENCY BAND.

waveband switch See BANDSWITCH.

wave beam Unidirectional radiation from a directive antenna.

wave bounce See WAVE REFLECTION.

wave clutter Radar interference caused by waves on a body of water, particularly large swells on the ocean.

wave converter A waveguide part, such as a baffle-plate or grating, that changes a wave pattern from one type to another.

wave crest The maximum value of a wave envelope. Compare WAVE TROUGH.

wave cycle A complete single alternation of a wave.

wave direction The direction in which an electromagnetic wave travels. It is perpendicular to the wave front and depends (whether it is forward or backward) upon the direction of the electric and magnetic components. If either is reversed, wave direction reverses; if both are reversed, the direction remains unchanged.

wave duct **1.** See DUCT, **1**. **2.** A tubular waveguide in which wave propagation is concentrated.

wave envelope The outline described by the various amplitude peaks of the cycles in an amplitude-modulated wave. The envelope frequency is equal to the modulating frequency.

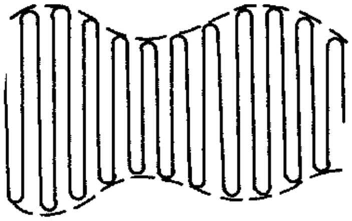

wave envelope

wave equation A second-degree partial differential equation whose solution describes wave phenomena.

wave filter A circuit or device that offers different amounts of attenuation to signals of different frequencies. See BANDPASS FILTER, BANDSUPPRESSION FILTER, HIGH-PASS FILTER, and LOW-PASS FILTER.

waveform The shape of a wave described in terms of its resemblance to some well-known figure or to its conformity to the curve of the applicable wave equation (e.g., sinusoidal, square, sawtooth, cosine, rectangular, and triangular).

waveform-amplitude distortion See AMPLITUDE DISTORTION.

waveform analyzer See WAVE ANALYZER.

waveform converter A circuit or device for changing a signal of one waveform (such as a sine wave) into one of another waveform (such as a pulse or square wave).

waveform distortion The malfunction evidenced by a change of the waveshape of a signal passing through a circuit.

waveform error In a quantity displayed by an alternating-current test instrument, an error caused by the waveform of the measured signal. Thus, a voltmeter calibrated with a sine-wave voltage is subject to error when a measured signal is nonsinusoidal. Also called *waveform effect.*

waveform generator See FUNCTION GENERATOR.

waveform influence See WAVEFORM ERROR.

waveform monitor In television operations, an oscilloscope that continuously displays the video waveform.

waveform synthesizer A variable-frequency signal generator that allows the tailoring of waveshape to suit individual applications. A *function generator* (see FUNCTION GENERATOR, **1**) is such an instrument, but it usually provides only a choice of common waveshapes.

wave front For a radio wave, the plane that is parallel to the perpendicular electric and magnetic lines of flux. The wave propagates at right angles to this plane.

wave function A point function in a wave equation, specifying wave amplitude.

wave group The resultant of several different-frequency waves traveling over a common path.

waveguide A transmission line used at ultra-high and microwave radio frequencies. It is a hollow metal pipe with a rectangular or circular cross section. A rectangular waveguide must have sides measuring at least 0.5 wavelength and preferably more than 0.7 wavelength. A circular waveguide should be at least 0.6 wavelength in diameter, and preferably 0.7 wavelength or more. The characteristic impedance (Z_0) varies with frequency. In this sense, it differs from coaxial or parallel-wire lines, whose Z_0 values are independent of frequency.

waveguide apparatus See WAVEGUIDE COMPONENTS.

waveguide attenuator A device, such as an interposed energy-absorbing plate, for signal attenuation in a waveguide.

waveguide choke flange A waveguide flange that presents no impedance to the signal, and which need not be metallic for continuity.

waveguide component A device adapted for connection to, or insertion into, a waveguide system. Such components include waveguide parts and accessories (e.g., splicing hardware, attenuators, loads, wavemeters, etc.).

waveguide connector A fitting for joining waveguides for the efficient propagation of a signal.

waveguide coupling See WAVEGUIDE CONNECTOR.

waveguide critical dimension The cross-sectional dimension that determines the cutoff frequency for a waveguide.

waveguide cutoff In a waveguide, the highest or lowest frequency that can be propagated with less than a specified amount of attenuation per unit length.

waveguide directional coupler A directional coupler made of two parallel waveguides with a common wall. Two slots cut in the wall allow part of the microwave energy propagated in one direction in the main waveguide to be extracted, and energy traveling in the opposite direction to be rejected.

waveguide dummy load A waveguide section that dissipates the microwave energy entering it.

waveguide elbow **1.** A curved bend in a waveguide. **2.** A waveguide connector with a bend.

waveguide flange A flat, liplike fitting at the end of the pipe of a waveguide. It fastens waveguide sections together or attaches a waveguide

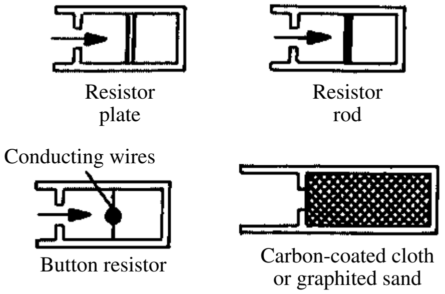

waveguide dummy load

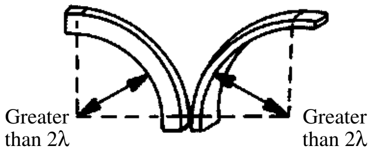

waveguide elbow

component, equipped with an identical flange, to the end of a waveguide.

waveguide frequency meter See WAVEGUIDE WAVEMETER.

waveguide gasket A gasket that provides electrical continuity between mating waveguide sections.

waveguide grating An array of wires mounted inside a waveguide, and that passes signals at some frequencies while obstructing others.

waveguide impedance Where the power P is known, and voltage E and current I are defined, with respect to a type of wave and waveguide, the value is either E^2/P or P/I^2.

waveguide junction A fitting that allows one waveguide section to be joined at an angle to another section. See, for example, WAVEGUIDE TEE and WAVEGUIDE WYE.

waveguide lens A microwave lens consisting of waveguide sections that provide the required phase shifts.

waveguide load See WAVEGUIDE DUMMY LOAD.

waveguide mode The form of propagation indicated by the field pattern in a plane transverse to the direction in which energy is propagated through a waveguide. Common modes are TRANSVERSE ELECTRIC MODE, also called *TE mode*, and TRANSVERSE MAGNETIC MODE, also called *TM mode.*

waveguide mode suppressor A filter that suppresses undesired propagation modes in a waveguide.

waveguide phase shifter A shifter for adjusting the phase of waveguide output energy, with respect to input energy.

waveguide plunger A plunger-like device that reflects incident microwave energy in a waveguide.

waveguide post A transverse rod inside a waveguide that acts as a parallel susceptance.

waveguide probe A pickup probe (tip or loop, as required) for sampling the field inside a waveguide, or a similar injection probe for introducing energy into a waveguide. Also see WAVEGUIDE SLOTTED LINE.

waveguide propagation **1.** The transmission of microwave energy through a waveguide by successive reflections between the inner walls. **2.** Propagation of very-high-frequency (VHF), ultra-high-frequency (UHF), or microwave electromagnetic fields through an atmospheric duct (see DUCT, **1**), as if through a waveguide. **3.** Propagation of very-low-frequency (VLF) electromagnetic fields through a waveguide-like duct between the ionosphere and the earth's surface.

waveguide radiator An antenna consisting of an open-ended waveguide with or without a horn. It radiates microwave energy into space or to a reflector.

waveguide resonator A waveguide section used as a cavity resonator (see RESONANT CAVITY).

waveguide seal A protective cover for the end of a waveguide. The seal introduces very little microwave attenuation, while preventing entry of moisture and debris.

waveguide shim A thin, pliable metal sheet inserted between mating waveguide components for electrical continuity. Also see WAVEGUIDE GASKET.

waveguide shutter An adjustable mechanical barrier, such as a rotatable vane, inserted into a waveguide to block or divert microwave energy.

waveguide slotted line A section of waveguide having a slot that accommodates a movable probe or coupling element.

waveguide slug tuner A quarter-wave dielectric slug inserted into a waveguide so that its amount of penetration and position can be adjusted for tuning purposes.

waveguide stub A stub consisting of a waveguide section joined to a main waveguide at an angle and provided with a nondissipative termination.

waveguide stub tuner An adjustable piston in a waveguide stub for tuning purposes.

waveguide switch A switch consisting of a movable section of waveguide that can be positioned for coupling to one of several other waveguide sections; it thus passes the energy it receives to any of the other sections.

waveguide system Microwave "plumbing" consisting of waveguides, their fittings and accessories, and associated components (such as attenuators, loads, wavemeters, etc.).

waveguide taper A connector that is flared to allow coupling between two waveguide sections having different cross-sectional sizes.

waveguide tee In a waveguide assembly, a tee-shaped junction used to connect a section of waveguide in series or parallel with another section.

waveguide transformer A waveguide component that functions as an impedance transformer.

waveguide tuner In a waveguide system, an adjustable tuner providing impedance transformation.

waveguide twist A length of waveguide whose cross section is rotated around the longitudinal axis (e.g., from vertical to horizontal).

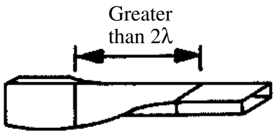

waveguide twist

waveguide wavelength In a uniform waveguide operating at a given frequency and in a particular mode, the distance between similar points for 360° phase shift.

waveguide wavemeter A waveguide component that acts as an absorption wavemeter or transmission wavemeter for identifying microwave frequencies.

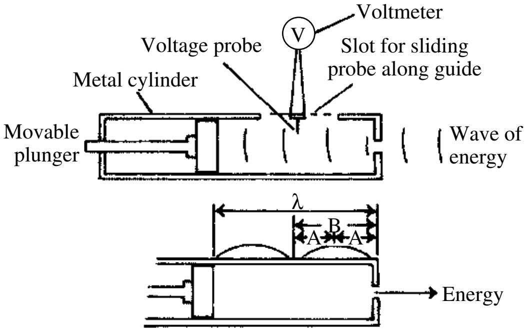

waveguide wavemeter

waveguide wedge See WEDGE, **1**.

waveguide window A thin metal opening mounted transversely inside a waveguide for impedance-matching purposes. The edges of the slit in a capacitive window are perpendicular to the electric field; in an inductive window, they are parallel to the electric field.

waveguide wye In a waveguide assembly, a wye-shaped junction for joining three waveguide sections.

wave heating Heating a material by energy absorbed from traveling electromagnetic waves.

wave interference Interaction between two or more waves, resulting in reinforcements and cancellations of energy.

wavelength Unit, meter. The displacement in one complete wave of an alternating or vibrating phenomenon, generally measured from crest to crest or from trough to trough of successive waves. For electromagnetic waves in free space, the wavelength in meters is equal to 3×10^8 divided by the frequency in Hertz. Also see WAVELENGTH-PERIOD-FREQUENCY RELATIONSHIPS.

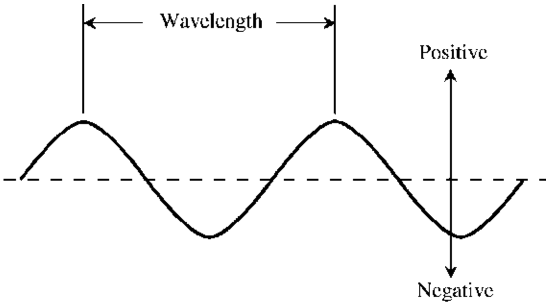

wavelength

wavelength constant The imaginary-number component of the propagation constant.

wavelength shifter **1.** A frequency shifter whose performance is indicated in units of wavelength, rather than in units of frequency. **2.** In certain photosensitive cells and tubes, a photofluorescent substance that raises the efficiency of the device by absorbing photons and then releasing ones of longer wavelength.

wave mechanics A theory of matter that views subatomic particles as complex wave patterns, and attempts to account for all physical processes in terms of wave phenomena.

wavemeter An instrument for measuring the wavelength or frequency of radio waves. One form consists of a series-resonant circuit containing an inductor, variable capacitor, and diode-type meter. The dial of the capacitor is calibrated to read in MHz. The inductor picks up energy from the radio-frequency source of unknown frequency, the capacitor is tuned for peak deflection of the meter, and the unknown frequency is read from the dial. This instrument is often called an *absorption wavemeter* because it absorbs a certain amount of power from the signal source under test. See also CAVITY WAVEMETER, COAXIAL WAVEMETER, LECHER WIRES, and SLOTTED LINE.

wave motion Undulating motion (e.g., up and down, and side to side). An electromagnetic wave has undulating electric and magnetic components that are both in phase and perpendicular to each other and to the direction of propagation of the wave.

wave normal **1.** The direction of propagation of an electromagnetic wave. **2.** A unit vector directed at a right angle to both the electric and magnetic lines of flux in an electromagnetic wave.

wave number The reciprocal of wavelength. This number denotes the number of waves per unit distance.

wave packet A short pulse composed of waves.

wave packets Radiant energy resulting from a number of wave trains of different wavelength.

wave path The line along which a WAVE TRAIN is propagated.

wave polarization The direction (horizontal or vertical) of wave undulations (i.e., the plane of the undulations, with respect to the direction of propagation). In general, a vertical antenna radiates a vertically polarized wave, and a horizontal antenna radiates a horizontally polarized wave.

wave propagation The movement of waves through space or through some medium. Electromagnetic waves travel through space at the speed of light (approximately 3×10^8 meters, or 186,000 miles, per second) and, like light, can be reflected and refracted.

wave reflection The reflection of electromagnetic waves by an obstruction, such as a solid body or a layer of the ionosphere. Compare WAVE ABSORPTION, WAVE POLARIZATION, and WAVE REFRACTION.

wave refraction Bending of the line of propagation of electromagnetic waves as they pass through various media, such as the troposphere or the ionosphere. Compare WAVE ABSORPTION, WAVE POLARIZATION, and WAVE REFLECTION.

waveshape The overall contour of a wave—especially as revealed by a curve plotted for the particular wave equation. Also see WAVEFORM.

waveshaping circuit A circuit that receives an input signal having a certain waveshape, and delivers an output signal having a different waveshape. For example, a squaring circuit converts a sine wave into a square wave at the same frequency.

wave surface See WAVE FRONT.

wave tail In a decaying pulse or signal envelope, the interval between the beginning of the decay and the point at which the amplitude reaches zero.

wave telegraphy See RADIOTELEGRAPHY.

wave telephony See RADIOTELEPHONY.

wave theory of matter A physical theory that the charge of an electron is distributed in space, rather than being focused at a point. Also see WAVE MECHANICS.

wave tilt A slight forward tilt of the electric flux lines in a radio wave radiated at the surface of the earth by a vertical antenna.

wave train A series of identical electromagnetic wave cycles propagated at equal intervals; an electromagnetic energy burst lasting at least several cycles.

wavetrap A resonant circuit consisting of an inductor and capacitor, either or both of which can be adjustable for tuning, used to remove (trap) a signal at the resonant frequency from a signal mixture.

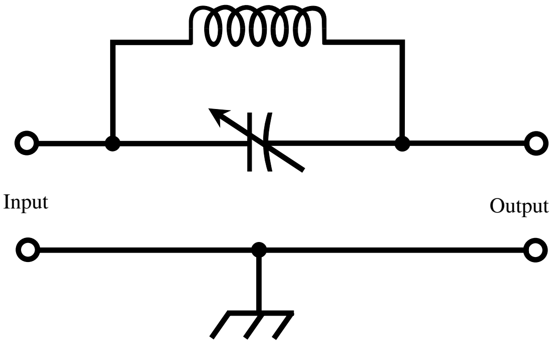

wavetrap

wave trough The minimum value of a wave envelope. Compare WAVE CREST.

wave velocity The distance per unit time traversed by a wave passing through a given medium.

wave winding See DRUM WINDING.

wa-wa pedal A foot-operated device used with an electronic musical instrument to produce a "wah-wah" sound fluctuation.

wax **1.** Any of a series of organic materials having important uses as dielectrics, impregnants, sealers, and lubricants in electronics. They are usually solid or semisolid, waterproof, and easily melted. **2.** In certain phonograph record discs, a blend of wax (see **1**, above) and metallic soaps. Also see WAX MASTER, **2**.

wax cake See WAX MASTER, **1**, **2**.

wax capacitor A fixed capacitor that has been dipped in or impregnated with a wax, such as halowax.

wax-dipped capacitor A fixed capacitor that has been dipped in a wax for sealing against moisture.

waxed paper See WAX PAPER.

wax-filled capacitor A fixed capacitor impregnated with a wax for enhancing the properties of its dielectric (usually paper) and sealing the capacitor unit.

wax master **1.** In disc-recording operations, the original recording made on a wax-surface disc. **2.** To make an original recording on a wax-surface disc.

wax original See WAX MASTER.

wax paper Wax-saturated paper used as a dielectric film in fixed capacitors and as an insulator.

way-operated circuit A single or duplex circuit shared by three or more party stations.

way point An important point selected on a radionavigational course line.

way station A station consisting of a teletypewriter connected at an intermediate point in a line (i.e., between, and in series with, other teletypewriter stations).

WAZ Amateur radio abbreviation of *Worked All Zones*, an award given to operators who have carried on verified two-way communication with stations in all communications zones of the world.

WB Abbreviation of *weather bureau.*

W_B Abbreviation of *base-region width* (in a transistor).

Wb Symbol for WEBER.

Wb/m^2 Abbreviation of *Webers per square meter* (see TESLA).

W_C Abbreviation of *collector-region width* (in a transistor).

WCEMA Abbreviation of *West Coast Electronic Manufacturers' Association.*

W/cm^2 Abbreviation of *watts per square centimeter.*

WE Abbreviation of *write enable.*

W_E Abbreviation of *emitter-region width* (in a transistor).

weak battery **1.** A battery that has been depleted to the point that its output (no-load or full-load) is too low to be useful. **2.** A battery specially designed for low-voltage output.

weak color Lack of color vividness or poor contrast between colors in a color-television picture. The condition is often caused by some malfunction in the chroma demodulator(s).

weak contrast In a television picture, poor differentiation of adjacent tonal areas.

weak coupling See LOOSE COUPLING.

weak current An extremely small electric current. The term is relative; generally, it refers to currents of a few microamperes or less.

weak magnet **1.** A magnet whose power has deteriorated considerably below a prescribed level. **2.** A body that normally is only slightly magnetic.

weak signal A signal whose amplitude is very low compared with that of signals considered satisfactory in a given application. Although the term is relative, it usually implies a signal that is noncompetitive with other signals in a given environment.

weak-signal detector A detector in which, at low input-signal amplitudes (weak-signal levels), the direct-current output is proportional to the square of the root-mean-square (rms) value of the input-signal voltage.

wearout The complete deterioration of a component or system (i.e., beyond restoration to useful service).

wearout failure Failure because of wearout, which can be predicted on the basis of known lifetime and the deterioration characteristics of components and equipment.

wearout point The instant of wearout, in terms of power output, watt-hour capacity, or some other specification.

weather antenna An antenna dimensioned for reception exclusively in the 162.4- to 162.55-MHz weather band. See WEATHER TRANSMISSION.

weathering Deterioration of electronic equipment as a result of exposure to outdoor heat, cold, moisture, wind, and similar conditions.

weather-protected machine A machine (usually a generator or motor) whose vent holes are designed to prevent entry of dust, water, and debris.

weather protection The coating, sealing, or treating of electronic equipment for protection against corrosion, humidity, and temperature changes in outdoor use.

weather satellite A satellite designed to photograph weather systems in infrared and/or visible light, and relay the pictures to earth via facsimile or television.

weather sonde See RADIOSONDE.

weather transmission The radio transmission of meteorological reports. Sometimes the transmissions are combined with guidance transmissions, from which they can be separated by means of a filter in the receiver.

weber Abbreviation, Wb. The SI unit of magnetic flux and of the magnetic flux quantum; $1 \text{ Wb} = 10^8$ maxwells $= 1.257 \times 10^{-7}$ unit pole.

Weber-Fechner law The law expressing the relationship between a stimulus and the physiological reaction it produces: The sensation is proportional to the logarithm of the stimulus.

weber per square meter Symbol, Wb/m^2. See TESLA.

weber turn A unit of magnetic flux linkage equal to 10^8 maxwell turns.

wedge **1.** In a waveguide, a termination consisting of a tapered block or plate of carbon (or other dissipative material). **2.** In a television test pattern, convergent, equally spaced lines for checking resolution.

wedge bonding In integrated-circuit fabrication, a method of bonding in which a thermocompression bond (see COLD-COMPRESSION WELDING) is obtained through pressure from a wedge-shaped tool.

Wehnelt cathode An oxide-coated cathode in an electron tube.

Wehnelt cylinder In a cathode-ray tube, the cathode-enclosing cylinder that concentrates the electrons emitted by the cathode.

weight **1.** The amount of gravitational pull on a body or particle. **2.** Extra significance given to a term or value. See, for example, WEIGHTED TERM. **3.** The dot-to-space ratio in a Morse-code signal.

weight-density Symbol, *d.* Unit, kg/m^3. The weight per unit volume of a liquid, such as an electrolyte or insulating oil; also called *density.*

weighted distortion factor In the measurement of harmonic distortion, a factor whose use allows the harmonics in the complex waveform to be weighted in proportion to their relationship.

weighted noise level Unit, dBm. The noise level weighted with respect to the 70-dB equal-loudness contour of hearing.

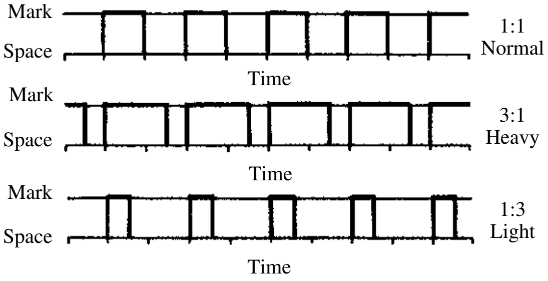

weight, 3

weighting **1.** Adjustment of a parameter to compensate for some imbalance in a system. **2.** Adjustment of the dot-to-space ratio in a Morse code signal. **3.** Adjustment of the mark-to-space ratio in a digital communications signal.

weighting filter A filter used in a communications network to represent the characteristics of the transmission passband.

weighting network A network that weighs differently (in a prescribed ratio) the frequency components appearing in an output signal by offering unequal attenuation to those frequencies.

weightlessness switch See ZERO-GRAVITY SWITCH.

Weir circuit In frequency-modulated signal transmission, a circuit used to stabilize the carrier wave. It compares the average carrier frequency with the frequency of a standard crystal oscillator, obtaining a direct-current compensating voltage (proportional to frequency deviation) that is applied to the frequency modulator. Also called *Weir stabilization circuit.*

Weiss constant In paramagnetism, a constant also known as *paramagnetic Curie temperature.* It can be positive or negative, depending on the particular paramagnetic material. It is important in defining the behavior of certain paramagnetic substances.

weld A strong bond of materials (usually metals) obtained by applying heat to areas to be joined while they are held or pressed together. No foreign metal is used, as is the case in brazing and soldering. The required heat is sometimes obtained by passing a high electric current through the materials.

welder **1.** An electrical device, often electronically controlled, for welding materials. **2.** A person who operates a device, as defined in **1**.

weldgate pulse In a welding device, the pulse that affects the arc current; therefore, it also affects the intensity of the heat produced by the device.

welding control An electronic system for controlling the interval during which current is passed through a workpiece in spot welding or seam welding. In this system, an electronic timer circuit determines the conduction time of thyratrons, ignitrons, or silicon-controlled rectifiers in the welding circuit. The control system can also regulate the welding current.

welding current The high electric current passed through a workpiece to produce the heat required for welding.

welding cycle The required sequence of steps (and the time required) in making a weld electronically.

welding time See WELD TIME.

welding transformer For electronic welding, a special very-high-current step-down transformer.

weld junction See WELD.

weld polarity The polarity of welding current. Some materials require a certain direction of current flow for a good weld.

weld time The interval during which welding current flows through the bodies to be bonded together.

well counter A radiation-counter setup in which a radioactive sample and detector are enclosed together in a thick-walled (usually lead) cylinder to minimize background count.

well-structured language An advanced form of high-level computer programming language. It is used in graphical and control applications.

Wenner element An adjustable, dual-slidewire balancing resistor used in constant-current, laboratory potentiometers to eliminate the necessity for sliding contacts in the measuring circuit.

Wenner winding A low-capacitance, low-inductance winding for high-frequency wirewound resistors in which the direction of the wire is reversed by looping alternate turns along the form.

Wertheim effect The tendency for a potential difference to develop between opposite ends of a length of wire, when the wire is placed parallel to magnetic lines of flux and rotated.

Western Union joint A strong splice of two wires made by tightly twisting a short portion of the tip of each wire along the body of the other. For increased ruggedness, the joint is often soldered. Also called *Western Union splice.*

Weston cell See STANDARD CELL.

Westrex system A system of sound recording in which signals from two separate microphone channels are recorded on opposite walls of a groove on a disc.

wet battery A battery of cells having a liquid electrolyte.

wet Liquid, especially pertaining to the electrolyte material in an electrochemical cell.

wet cell A battery cell having a liquid electrolyte. Compare DRY CELL.

wet-charged stand The length of time that a fully charged, wet storage cell can stand idle before its capacity drops by a specified amount.

wet electrolytic capacitor An electrolytic capacitor in which the electrolyte is a liquid. The leakage current in this type is higher than in the dry electrolytic, but it is self-healing after momentary voltage breakdown. Compare DRY ELECTROLYTIC CAPACITOR.

wet grab The adherence of a pressure-sensitive tape or sheet to a surface when very little pressure is used.

wet rectifier See ELECTROLYTIC RECTIFIER.

wet shelf life The specified shelf life of a discharged, wet storage cell. Compare DRY SHELF LIFE.

wetted-contact relay See MERCURY-WETTED REED RELAY.

wetting Applying a mercury coating to a contact surface.

wetting agent A substance (such as an alcohol or ester) that promotes the spreading and adhesion of a liquid or its absorption by a porous material.

WG Abbreviation of WIRE GAUGE.

WH Abbreviation of WATT-HOUR.

wh Abbreviation of WHITE.

Wheatstone bridge A four-arm balancing circuit (see BRIDGE), having resistors in each arm and used to measure an unknown resistance in terms of a standard resistance. The bridge supply is usually direct current, but alternating current can be used if all four resistances are nonreactive.

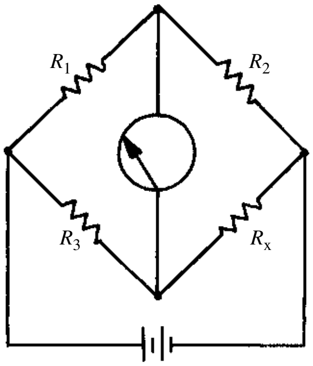

$$R_x = R_3(R_2/R_1)$$

Wheatstone bridge

wheel-drive locomotion The use of wheels for moving mobile apparatus such as robots. Offers simplicity and low cost. The main disadvantage is the inability to negotiate irregular terrain.

Wheeler's formula A formula for calculating the inductance of a multilayer air-core coil:

$$L = (0.8a^2n^2)/(6a + 9l + 10b)$$

where L is in microhenrys, a is the mean radius (inches) of the winding (axis to center of cross section), b is the depth of winding (inches), l is the length of winding (inches), and n is the number of turns.

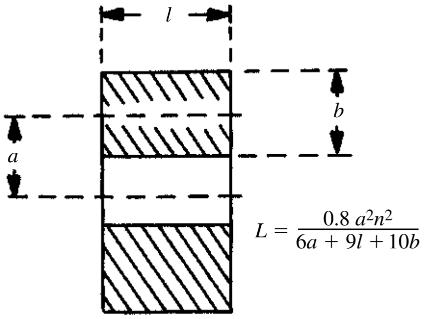

Wheeler's formula (dimensions illustrated)

wheel pattern A frequency test pattern produced on an oscilloscope screen by *z*-axis modulation of a circular trace. A sinusoidal axis signal produces a *gear-wheel pattern*, and a square-wave or pulse *z*-axis signal produces a *spot-wheel pattern.*

wheel printer A printout device for computers and calculators. It consists essentially of a rotating metal wheel around whose rim letters and numbers stand in relief. When the desired character comes into position, a hammer strikes it through the recording paper and carbon paper, printing the character on the recording paper. Also called *daisy-wheel printer.*

wheel static Static electricity (and the resulting radio interference) generated by friction between automobile tires and the road.

whiffletree switch In computer operations, a multiposition electronic switching circuit, so called from its circuit configuration, which resembles the contrivance used between a wagon and the horse team pulling it.

whip antenna A small-diameter, vertical rod (often telescoping) used as an antenna—especially in mobile communications, portable radio and television receivers, field-strength meters, etc.

whirl One of the circular, magnetic lines of flux around a straight wire carrying current.

whisker **1.** The pointer-wire electrode of a point-contact diode, point-contact transistor, or crystal detector. **2.** A slender filament of metal or ceramic, having high purity and high tensile strength.

whisker resistance In a semiconductor material, the resistance of a whisker component (see WHISKER, **1**).

Whiskey Phonetic alphabet code word for the letter W.

whistle A high-pitched tone (e.g., a beat note or acoustic feedback).

whistle filter A notch filter used to eliminate a whistle in an amplifier or other audio-frequency circuit. This filter can be of several versions, ranging from a simple resistance-capacitance

(RC) null circuit tuned to remove the offending frequency to a feedback-type band-suppression amplifier.

whistle interference The appearance of extraneous whistles in audio-frequency circuits. In some equipment, whistles result from oscillating amplifiers; in radio receivers, they are usually audible beat notes (heterodynes) produced by interfering carriers.

whistler A type of very-low-frequency (VLF) radio noise. They are thought to be caused by electromagnetic fields from distant lightning strokes, circulating in the earth's magnetic field. The name is derived from the peculiar sound the noise makes in a VLF radio receiver.

whistler mode propagation Radio transmission from the northern hemisphere to the southern hemisphere along the flux lines of earth's magnetic field.

whistlestop See WHISTLE FILTER.

whistling atmospheric See WHISTLER.

white The color that results from mixing all the additive primary colors: red, green, and blue.

white acid Hydrofluoric acid. Formula, HF. Used as an etchant, especially of glass.

white brass Brass that is more than 49% zinc.

white compression In television transmission, reduction of gain at highlight levels in the picture.

white-dot pattern In color-television tests with a dot generator, the one-white-dot pattern obtained when beam convergence has been secured.

white lamp See DAYLIGHT LAMP.

white level The lower-voltage point in a video signal, corresponding to full brilliance of the line on the screen (i.e., to the condition of whiteness in the picture).

white light See WHITE RADIATION, **2**.

white noise Random noise (acoustic or electric) equally distributed over a given frequency band, an example being the noise resulting from the random motion of free electrons in conductors and semiconductors.

white-noise generator A test device that generates electrical noise over a wide frequency spectrum. A simple type uses a reverse-biased silicon diode, the output of which is useful for testing audio amplifiers and radio receivers.

white-noise record A phonograph record containing recorded bands of white noise, each accompanied by voice announcements (e.g., instructions), and used to test the frequency response of a sound system.

white object A body that reflects and diffuses light of all wavelengths equally well.

white peak In a television picture signal, the maximum excursion in the white direction.

white radiation **1.** See WHITE NOISE. **2.** Visible light radiated with more or less equal intensity throughout the visible spectrum; seen as gray or white.

white raster See CHROMA-CLEAR RASTER.

white recording **1.** In amplitude modulation, recording characterized by a correspondence of maximum received power to minimum recording-medium density. **2.** In frequency modulation, recording characterized by a correspondence of lowest received frequency to minimum recording-medium density.

white room See CLEAN ROOM.

white saturation See WHITE COMPRESSION.

white signal The facsimile signal (see FACSIMILE) corresponding to scanning the copy area having maximum density.

white-to-black amplitude range **1.** At a point in a positive amplitude-modulated (AM) facsimile system (see FACSIMILE), the ratio (in dB) of signal current or voltage corresponding to white in the picture to that for black in the picture. **2.** At a point in a negative AM facsimile system, the ratio (in dB) of signal current or voltage corresponding to black in the picture to that for white in the picture.

white-to-black frequency swing At a point in a frequency-modulated facsimile system, the frequency difference $fw - fb$, where fb is the frequency corresponding to black in the picture, and fw is the frequency corresponding to white.

white transmission A system of picture or facsimile transmission in which the maximum copy darkness corresponds to the smallest amplitude (in an amplitude-modulated transmitter) or the highest instantaneous frequency (in a frequency-modulated transmitter). The opposite of BLACK TRANSMISSION.

white X radiation X rays of the continuous, or general, type.

whizzer An attachment that can improve the high-frequency reproduction in some audio loudspeakers.

whole-number division Arithmetic division (as in the division of binary numbers) in which the quotient is a whole number (i.e., division in which the divisor is contained in the dividend an integral number of times).

whorl See WHIRL.

WHP Abbreviation of *water horsepower.*

wick action **1.** The absorption of a liquid by, and its flow through, a cloth or thread, such as a lamp wick or lubricating wick. **2.** The flow of molten solder along and under the insulation of a wire.

wicking See WICK ACTION.

wide-angle diffusion A form of diffusion characterized by the wide-angle scattering of light; causes the source to have the same brightness at all viewing angles.

wide-area network A group of computers linked together, but separated by large geographic distances. Links can be made via telephone lines or radio.

wide-area service A teletype network that operates over long-distance wire lines.

Wide Area Telephone Service See WATS.

wideband **1.** Having a bandwidth greater than the minimum necessary to transmit a signal with acceptable intelligibility. **2.** For a voice signal, having a bandwidth greater than 6 kHz. **3.** Having the capability to operate, without adjustment, over a broad and continuous range of frequencies or wavelengths. **4.** In a digital network, a channel having a bandwidth of more than 64 kbps but less than 2 Mbps.

wideband amplifier An amplifier that exhibits reasonably flat response to a broad band of frequencies. The term is relative, depending on the application.

wideband antenna An antenna that transmits or receives signals over a broad frequency range, usually without the need for tuning.

wideband axis In a color-television signal, the direction of the fine chrominance primary phasor.

wideband communications **1.** Communications carried out over a band of frequencies wider than the minimum necessary for effective transfer of the information. **2.** A method of transmitting and receiving signals by deliberately varying the channel frequency over a wide range. Also called *spread-spectrum communications.*

wideband generator A signal generator covering a wide frequency range. Typical coverage in a laboratory-type instrument is 10 kHz to 1000 MHz.

wideband oscilloscope An oscilloscope whose horizontal, vertical, and sweep channels operate over a wide band of frequencies. Although the term wideband is relative, a wideband oscilloscope is usually assumed to be capable of displaying both radio and audio frequencies.

wideband ratio The ratio $B1/B2$, where $B1$ is frequency bandwidth and $B2$ is intelligence bandwidth.

wideband receiver A radio receiver that can tune in signals over a broad range of frequencies. An example is a communications receiver that can cover 10 kHz to 30 MHz continuously.

wideband repeater A repeater capable of operating over a wide range of input and output frequencies. Such repeaters are used in active communications satellites handling many different channels at the same time.

wideband signal generator See WIDEBAND GENERATOR.

wideband sweep **1.** In the operation of an oscilloscope, a repetitive sweep of the electron beam, the frequency of which is adjustable to any desired point within a wide range. The basic sweep rate in simple oscilloscopes is restricted to the audio-frequency spectrum (up to about 20 kHz), but a wideband sweep extends to much higher frequencies, typically several tens of MHz. **2.** A sweep circuit that produces the wideband sweep action described in **1**.

wideband test meter An alternating-current (ac) meter that can measure quantities over a wide frequency range in its basic form (i.e., without special external probes or converters). An example is an electronic ac voltmeter with a range extending from 10 Hz to 2.5 MHz.

wide-base diode A junction diode in which the p region is considerably wider than the n region.

wide-open **1.** Pertaining to wideband, untuned response. **2.** Pertaining to maximum-gain operation (e.g., a wide-open amplifier or receiver).

wide-range ammeter An ammeter that employs one or more shunt resistances to increase the full-scale deflection, usually by a power of 10 (10, 100, 1000, etc.). The resistor must be capable of carrying the current without burning out. Shunts are used when it is necessary to measure very large currents, such as hundreds of amperes. Shunts also allow a microammeter or milliammeter to be used as a multimeter with many current ranges.

wide-range reproduction High-fidelity audio-frequency reproduction.

width **1.** The horizontal dimension of a pulse, usually corresponding to its effective duration; also called PULSE DURATION. **2.** The horizontal dimension of an image, such as a television picture.

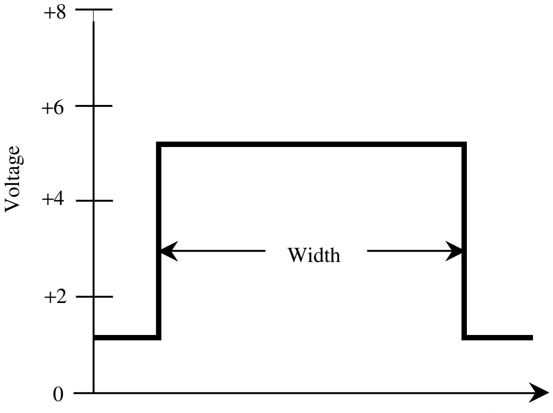

width, 1.

width coding The modification of pulse duration according to a code.

width coil See WIDTH CONTROL, **1**.

width control **1.** In a television receiver, the variable component for adjusting the swing of the horizontal deflection voltage and, therefore, the width of the picture. It is often a slug-tuned coil connected in parallel with a portion of the secondary winding of the horizontal output transformer. **2.** Sometimes, the horizontal gain control in an oscilloscope.

width mode In electronic-counter operations, a time-interval mode in which the signal is measured.

width modulation Modulation of the interval during which a gate circuit is open.

Wiedemann effect See DIRECT WIEDEMANN EFFECT.

Wiedemann-Franz law For metals that are good electrical conductors, the ratio of thermal conductivity to electrical conductivity is nearly constant, and is proportional to the absolute temperature.

Wien bridge A frequency-sensitive bridge in which two adjacent arms are resistances and the other two adjacent arms are resistance-capacitance (RC) combinations. One of the latter contains resistance and capacitance in series; the other contains resistance and capacitance in parallel. Because the bridge can be balanced at only one frequency at a time, it is useful as a simple audio frequency meter (see BRIDGE-TYPE AF METER). It is used also for capacitance and resistance measurements. When inductors are substituted for the capacitors, the Wien bridge can be used for inductance measurements (see WIEN INDUCTANCE BRIDGE).

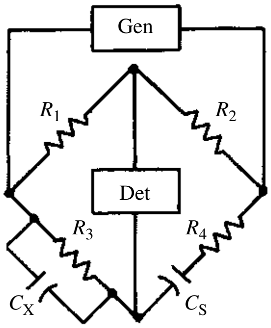

Wien bridge

Wien-bridge audio frequency meter See WIEN-BRIDGE FREQUENCY METER.

Wien-bridge distortion meter A distortion meter in which a Wien bridge circuit is used to remove the fundamental frequency from the complex waveform. The bridge is inserted between amplifier stages, and its notch response is sharpened by means of overall negative feedback.

Wien-bridge equivalent A resistance-capacitance null circuit, such as the parallel-tee (twin-tee) network, which has the same balance equation as the Wien bridge.

Wien-bridge filter A Wien bridge used as a band-suppression filter (notch circuit).

Wien-bridge frequency meter A bridge-type audio frequency meter in which the continuously variable frequency-selective circuit is a Wien bridge. Also see BRIDGE-TYPE AF METER.

Wien-bridge heterodyne eliminator A notch filter composed of a continuously variable Wien-bridge circuit (see WIEN BRIDGE). The device is connected in the audio-frequency channel of a radio receiver, and is tuned to remove a troublesome heterodyne whistle.

Wien-bridge oscillator An oscillator in which the frequency-selective feedback network is a Wien-bridge circuit (see WIEN BRIDGE). Also see BRIDGE-TYPE OSCILLATOR.

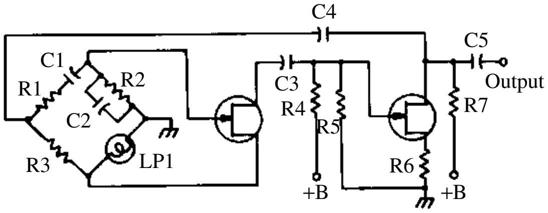

Wien-bridge oscillator

Wien capacitance bridge A Wien bridge arranged for the measurement of unknown capacitance.

Wien effect For an electrolytic substance, an increase in conductivity (decrease in resistivity) that occurs when a very large voltage is placed across the material. This voltage must be greater than approximately 1 megavolt per 50 cm.

Wien inductance bridge A Wien bridge containing inductors in place of capacitors and used for the measurement of unknown inductance.

Wien's displacement law The wavelength of maximum radiation of a black body is inversely proportional to the absolute temperature.

Wien's first law See WIEN'S DISPLACEMENT LAW.

Wien's laws See WIEN'S FIRST LAW, WIEN'S SECOND LAW, and WIEN'S THIRD LAW.

Wien's radiation law See WIEN'S SECOND LAW.

Wien's second law The emissive power of a black body is proportional to the fifth power of the absolute temperature.

Wien's third law An empirical law for the spectral distribution of energy radiated from a black body at a specified temperature. The distribution is a curve with a peak (maximum) at a wavelength that depends on the temperature. The higher the temperature, the shorter the wavelength at which maximum radiation occurs.

willemite See ZINC ORTHOSILICATE PHOSPHOR and ZINC SILICATE PHOSPHOR.

Williams tube A type of electrostatic cathode-ray storage tube.

Wilson chamber See WILSON CLOUD CHAMBER.

Wilson cloud chamber An airtight chamber containing water vapor or alcohol vapor at low pressure, and provided with a viewing window.

Radioactive particles penetrating the chamber are made visible as droplets form trails when the vapor condenses on them.

Wilson effect Electric polarization of a dielectric material being moved in a magnetic field.

Wilson electroscope A leaf-type electroscope (see ELECTROSCOPE) using a single leaf hanging vertically.

Wilson experiment An experiment demonstrating the WILSON EFFECT. It consists of rotating about its axis a hollow dielectric cylinder (whose inner and outer surfaces are metallized) in a magnetic field parallel to the axis. The result: An alternating voltage appears between the metallized surfaces.

Wimshurst machine A rotating machine used to produce high-voltage static electricity. The machine contains two glass disks, each having separate sectors of metal foil spaced around its face. The disks rotate in opposite directions, and the foil sectors passing each other form variable capacitors. Metal brushes pick up charges from the sectors passing under them and deliver this energy to one or more Leyden jars for storage.

wind charger A wind-driven generator used specifically to supply direct current for charging storage batteries.

wind-driven generator A dynamo-type generator, either stationary or mobile, powered by a windmill-like device.

wind gauge See ANEMOMETER.

winding **1.** A coil in an inductor or transformer (e.g., *primary winding*, *secondary winding*, *output winding*, etc.). **2.** A coil in a motor or generator.

winding arc The winding length of a coil, expressed in degrees.

winding cross section The cross section of a multilayer coil.

winding depth The depth of a multilayer coil, measured from the outermost layer to the innermost layer.

winding factor For a transformer or choke coil (or for a toroidal coil), the ratio of the total area of wire in the window to the window area.

winding length The length of a coil from the first turn to the last in a single-layer coil, or from the first turn to the last in one layer of a multilayer coil when all layers are identical.

winding space The window area of the core of a transformer or choke (see WINDOW, **3**).

wind loading **1.** The total wind pressure on an antenna system, generally measured in pounds or kilograms. The greater the wind speed, the greater the wind loading for a given antenna. The greater the exposed surface of the antenna, the greater the wind loading for a given wind speed. **2.** The highest wind speed, in miles per hour or meters per second, that an antenna system can safely withstand, assuming no accumulation of ice on the antenna structure.

Windom antenna **1.** A multiband antenna in which a single-wire feeder is attached to a horizontal half-wave radiator wire somewhat off center, at a point about ⅙ wavelength from one end of the radiator. The antenna operates with reasonable efficiency at the first several harmonics of the frequency at which it measures ½ wavelength. **2.** A similar antenna using a parallel-wire feeder.

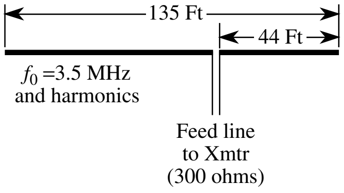

Windom antenna, 2

window **1.** Radar-interference material (see CHAFF). **2.** An interval during which a circuit is gated open to permit signal sampling. During this interval, a window is figuratively open to the signal. **3.** The open spaces between the legs of an iron core for a transformer or choke coil. **4.** An electromagnetic frequency band easily transmitted by earth's atmosphere. **5.** The period during which conditions are ideal for a complex operation, such as a rocket or spacecraft launch. **6.** An application space in a computer program that uses a graphical interface.

window area See WINDOW, **3**.

window comparator A comparator that detects voltage levels within a certain range of values, rather than simply indicating whether a voltage is more or less than a certain specified value.

window corridor An area where window (see WINDOW, **1**) has been dispersed.

window jamming The disturbance of electronic communications, especially radar, by dumping reflective material, such as metal foil, from an aircraft. Also see CHAFF; JAMMING; and WINDOW, **1**.

Windows Trade name (Microsoft) for a personal-computer interface scheme. The several versions all use selectable icons and menus.

window strip An insulated, flat, metal strip for bringing an antenna lead-in through a window. The window can be closed on the strip.

wind screen A foam covering that can minimize the roar caused by wind blowing against or across a microphone.

wind shield A radio-transparent cover placed over a radar antenna to protect it against damage from high winds. It is used especially in radar systems aboard aircraft.

wing In an antenna or other radiator, a (usually flat) member attached to, and sticking out from, another member. Its name is derived from its characteristic shape.

winterization The protective treatment and preparation of electronic equipment for winter storage or for operation in frigid regions.

wipeout Severe interference that obliterates all desired signals.

wipeout area An area in which WIPEOUT occurs.

wiper **1.** A thin metal blade or strip in sliding contact with a coil or other element over which it is turned to vary some quantity (such as resistance, inductance, voltage, current, etc.). **2.** A brush in a motor or generator.

wiper arm See WIPER, **1**.

wiper blade See WIPER, **1**.

wiping action The movement of contacts against each other when they slide as they mate or withdraw.

wiping contact A contact that makes and breaks with a sliding motion.

wire **1.** A metal strand or thread serving as a conductor of electricity. **2.** To connect wires between points in a circuit. **3.** See TELEGRAM. **4.** To send a TELEGRAM.

wire bonding **1.** The interconnection of components within a discrete package, by means of fine wire conductors welded to the individual components. **2.** A method of temporarily splicing the outer conductors of two coaxial cables. **3.** In general, any solderless method of splicing between two conductors.

wire broadcasting Broadcasting by means of carrier-current communication (see WIRED RADIO).

wire cloth A net of fine wire, used as an electrical shield when air circulation is necessary.

wire communication Communication carried on by means of signals transmitted over wire lines, as opposed to wireless (radio) communication.

wire control The control of remote devices by means of signals transmitted over wire lines, as opposed to radio control.

wire core A magnetic core consisting of a bundle of iron (or magnetic alloy) wires.

wired AND See DOT AND.

wired-in component A component to which wires are permanently connected in a circuit, as opposed to a plug-in component.

wired OR See DOT OR.

wired-program computer A computer in which instructions are wired-in by making appropriate connections to a panel with patch cords.

wired radio The transmission and reception of voice and other sounds, telegraph signals, pictures, control signals, telemetry signals, and the like by means of radio-frequency energy conducted by wire lines. Usually, the lines are used primarily for some other purpose (e.g., power lines or telephone lines).

wired-radio receiver The complete device or system that selects, amplifies, and demodulates or rectifies a radio signal picked up from wires that conduct it from a transmitting station.

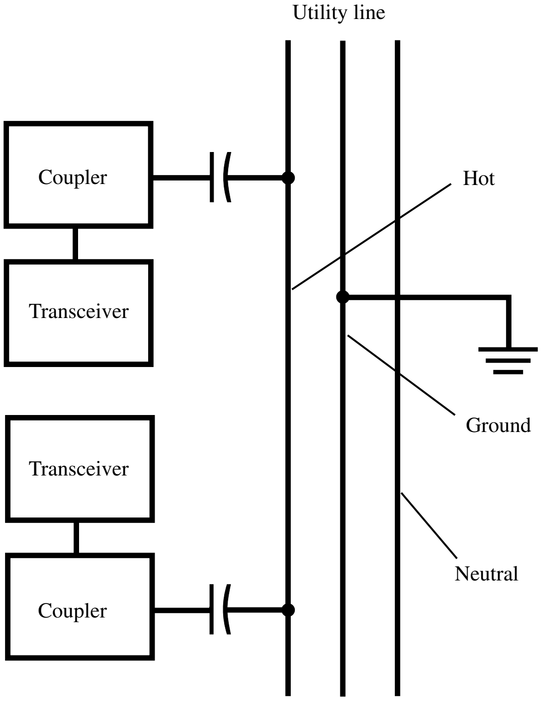

wired radio

wired-radio transmitter The complete device or system that generates radio-frequency power, adds signals (for communication, remote control, telemetry, etc.), and delivers the power to wires or to a cable for reception at a distant point.

wire drawing In the manufacture of wire, pulling metal through special dies to form wire of selected diameter.

wire dress The careful arrangement of wires in a chassis to ensure optimum performance. Also called *lead dress* (see DRESS).

wire duct A conduit through which wires are run. Such ducts have several shapes, but generally are rectangular.

wired wireless See WIRED RADIO.

wire-equivalent security See LEVEL-1 SECURITY.

wire fuse A fuse in which the fusible element is a wire of low-melting-point metal. Compare STRIP FUSE.

wire gauge **1.** A system for specifying the characteristics of wire. See AMERICAN WIRE GAUGE and BIRMINGHAM WIRE GAUGE. **2.** A wire number governed by diameter (e.g., 18-gauge wire for #18 wire). **3.** A tool or instrument for measuring wire diameter or for determining wire number.

wire gauze A fine screen of thin wires.

wire grating See WAVEGUIDE GRATING.

wire-guided Pertaining to guiding a machine, such as a robot, by means of signals sent by wire from the control point.

wire leads Leads that are (usually thin, flexible) wires, rather than pins, bars, strips, etc. An example is the wire leads of some transistors (as opposed to terminal pins).

wireless **1.** Pertaining to data communications and control systems that operate without wires (e.g., an infrared link between a notebook computer and a desktop computer). **2.** An early name for radio; it is still used in some countries. Sometimes for specificity, radio is referred to as *wireless telegraphy* or *wireless telephony*. See RADIOTELEPHONY.

wireless access protocol Acronym, WAP. A communications standard used on a worldwide basis with portable cellular telephone sets incorporating electronic mail, online service access, and/or Internet access.

wireless broadcaster See WIRELESS MICROPHONE.

wireless compass A radio compass (see DIRECTION FINDER).

wireless device A device that operates over a distance, without the use of interconnecting wires.

wireless e-mail **1.** The use of electronic mail (e-mail) in conjunction with a wireless Internet service. **2.** The use of a specialized portable wireless communications device, resembling a stripped-down computer or an enhanced cell phone, to send and receive e-mail messages.

wireless intercom An intercom using wired-radio for transmission and reception over the power line from which it is operated. Also see WIRED RADIO.

wireless Internet **1.** The use of a personal computer, usually a notebook or portable, in conjunction with a wireless modem connected to the Internet. Services of this type are provided with high-end cellular telephone subscriptions. Assets and limitations are similar to those associated with cell phone services. **2.** The use of a specialized portable or mobile wireless communications device, resembling a stripped-down computer or an enhanced cell phone, for basic Internet communications and information retrieval.

wireless microphone A device consisting of a small radio-frequency oscillator modulated by the microphone to which it is attached, and provided with a tiny antenna. The modulated signal is picked up and reproduced by a radio receiver.

wireless modem **1.** A modem that facilitates connection of a computer to the Internet using a cell phone set. **2.** A wireless transceiver that facilitates connection of a computer to the Internet. A cellular telephone connection can be included as an option. **3.** Also called *cordless modem*. A pair of short-range (on the order of a few feet) radio or infrared transceivers. One unit connects to a telephone jack or cable port, and the other unit connects to a computer. Can protect a portable computer against damage by lightning, assuming the computer uses batteries and is not connected to utility mains.

wireless tap A method of eavesdropping on communications in which a portion of the circuit makes use of a radio-frequency wireless link. An eavesdropping receiver can be positioned within range of the wireless transmitting antenna, and the signals intercepted. The existence of such a tap causes no change in the electronic characteristics of any equipment in the system, so its deployment is difficult to detect even if eavesdropping is anticipated.

wireless telegraphy See RADIOTELEGRAPHY.

wireless telephone **1.** A radio transceiver interconnected with the telephone lines. **2.** A telephone in which a short-range radio link replaces the cord between the receiver and the base unit. It normally has a range of several hundred feet.

wireless telephony See RADIOTELEPHONY.

wire line One or more wires or cables conducting currents for communication, control, or measuring purposes.

wire link A line for wire communication or wire control.

wire-link telemetry Telemetry carried on over a wire link.

wireman **1.** An electrician who specializes in the installation and servicing of electrical wiring. **2.** See LINEMAN.

wire mile A unit of measure equal to the product *ns*, where *n* is the number of separate, equal-length conductors in a line, and *s* is the length of a conductor in miles.

Wirephoto **1.** A system for transmitting and receiving photographs over wire lines. Also called *Telephoto.* See FACSIMILE. **2.** A photograph transmitted and/or received, as defined in **1**. **3.** A trademark for a photograph transmitted and received over telephone lines.

wire printer A printout device consisting of a wire array in which each wire is activated by an electromagnet. When a wire is selected (electrically), it is pushed against, and makes a dot on the recording paper. The letters and figures, then, are dot patterns. Compare WHEEL PRINTER.

wire radio See WIRED RADIO.

wire recorder In audio applications and data recording, a machine that records sounds or data pulses in the form of magnetized spots on a thin wire.

wiresonde A system for transmitting weather information data through a cable that holds a captive balloon. Compare RADIOSONDE.

wire splice A strong, low-resistance joint between (usually two) wires. See, for example, WESTERN UNION JOINT.

wire stripper A hand tool or power machine for cutting the insulating jacket on a wire and removing it without cutting or nicking the wire.

wiretap **1.** An instance of wiretapping. **2.** A wiretapping device.

wiretapper **1.** A device for wiretapping. **2.** A person who practices wiretapping.

wiretapping The act of making direct or indirect connections to a communications line to overhear or record a conversation.

wire telegraphy Telegraphic communication carried on by means of signals transmitted over wire lines, as opposed to wireless (radio) telegraphy.

wire telephony Voice communication carried on by means of signals transmitted over wire lines, as opposed to wireless (radio) telephony.

wire tie See TIE.

wire wave communication See WIRED RADIO.

wireways Hinged-cover metal troughs for containing and protecting wires and cables. Also see WIRE DUCT.

wirewound potentiometer A potentiometer in which the resistance element is a coil of resistance wire wound on a cord bent into a cylinder, or on a rigid, circular core.

wirewound resistor A resistor made from a coil of wire that is a poor conductor. The wire is usually wound around a cylindrical form; some components use toroidal forms. The resistance is determined by how well the wire metal conducts, by its diameter (gauge), and by its length. This type of resistor can be manufactured to precision tolerance. If the wire gauge is heavy, it can dissipate large amounts of power. The component has inductive reactance, making it unsuitable for use at radio frequencies. Compare CARBON-COMPOSITION RESISTOR, FILM RESISTOR.

wirewound rheostat A rheostat in which the resistance element is a coil of resistance wire wound on a card bent into a cylinder, or on a rigid, circular core.

wire wrap A method of circuit-board wiring in which components are interconnected by individual wire conductors, the ends of which are wrapped around terminal posts using a special tool.

wire-wrap connection An electrical connection made by tightly wrapping a bare wire around a special terminal.

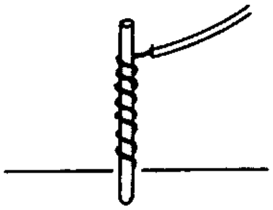

wire wrap

wire-wrapping tool A device used to make a wire-wrap connection.

wiring **1.** The system of wires or similar conductors connected between circuit components. **2.** Connecting and dressing such wires, as in wiring a circuit. **3.** Collectively, the connections (either actual wire or printed metal lines or processed semiconductor paths) between terminals and components in an electronic circuit. **4.** The process of installing or making such connections.

wiring board The control panel of a computer (i.e., a plugboard).

wiring capacitance Unavoidable capacitance between wires in a circuit, or between the wires and nearby metal bodies.

wiring connector A small (usually metal) fitting used to tie wires together.

wiring diagram See CIRCUIT DIAGRAM.

wiring impedance Unavoidable impedance in the wires in a circuit, and in associated terminals and hardware conducting alternating current. This impedance is the vector sum of WIRING REACTANCE and WIRING RESISTANCE.

wiring inductance Unavoidable inductance in the wires in a circuit, and in associated terminals and other hardware, through which alternating currents flow.

wiring reactance Unavoidable reactance in the wires in a circuit, and in associated terminals and other hardware, conducting alternating current. This reactance is caused by WIRING CAPACITANCE and/or WIRING INDUCTANCE.

wiring resistance Unavoidable resistance in the wires in a circuit and in associated terminals and other hardware.

Witka circuit A voltage-tripler circuit using only two diodes and two capacitors. Its no-load direct-current output voltage is approximately three times the peak value of the alternating-current input voltage. Also see VOLTAGE TRIPLER.

wk **1.** Abbreviation of WORK. **2.** Abbreviation of *week.*

WL **1.** Abbreviation of WAVELENGTH. **2.** Abbreviation of *waterline.*

WM Abbreviation of WATTMETER.

$\mathbf{Wm^2}$ Symbol for *watt square meter* (unit of the first radiation constant).

W/(MK) Symbol for *watts per meter kelvin* (unit of thermal conductivity).

WMO Abbreviation of *World Meteorological Organization.*

wobbulator A device that sweeps the frequency of an oscillator. There are several types, from a motor-driven, rotating, tank capacitor to sophisticated, fully electronic variable capacitors or variable inductors.

wolfram See TUNGSTEN.

Wolf's equation A sunspot-number equation used to forecast maximum usable frequency: $R = k(10G + N)$, where R is relative sunspot number, k is a constant for the telescope used, G is the number of sunspot groups observed, and *N* is the total number of sunspots.

Wollaston wire An extremely fine wire made by coating platinum with silver, drawing it thinner, and dissolving the silver coat.

womp A sudden surge in amplitude that causes a flare up of brightness in a television picture.

woodpecker A slang term used to describe the sound of the signal from over-the-horizon radar. Operating in the high-frequency part of the spectrum and changing in wavelength, roughly following the maximum usable frequency, the signal consists of repeated "spikes" that can interfere with routine communications.

woodpecker filter A special type of blanking filter used to quiet a shortwave receiver during strong, brief "spikes." It is similar to a noise blanker, but is designed to filter out the longer peaks characteristic of the over-the-horizon radar, known as the WOODPECKER.

Wood's alloy See WOOD'S METAL.

Wood's metal A low-melting-point (159.8°F) alloy containing bismuth (50%), cadmium (12.5%), lead (25%), and tin (12.5%). The alloy looks like lead and is used to mount rectifying crystals (such as galena) whose electrical sensitivity would be destroyed by the high temperatures required to melt most soft metals.

woofer A large loudspeaker, often a foot or more in diameter, designed specifically to reproduce very low (bass) audio frequencies. It is commonly used in high-fidelity stereo sound systems. Compare TWEETER.

word **1.** In computer operations, a group of bits or characters treated as a unit. **2.** In telegraphy, a data unit consisting of five characters plus one space.

word-address format A method of addressing a word by means of a single character, such as the first letter of the word.

word code **1.** A cipher in which word meanings are interchanged, rather than letter symbols. **2.** A word having an altered meaning in a cipher system, as defined in **1**. For example, "word" might mean "code."

word format The specific sequence of characters that forms a word of data.

word generator A special signal generator that delivers pulses in selected combinations corresponding to digital words (see WORD, **1**). It is used in testing computers and digital systems.

word length **1.** In computer operations, the number of bits in a word. **2.** In telegraphic communications (radio or wire), the average number of letters in a word.

word pattern The shortest meaningful word (see WORD, **1**, **2**) that can be recognized by a machine.

word processor **1.** An electronic device, similar to a typewriter, used for writing. Words, phrases, sentences, or paragraphs can be changed, replaced, or deleted prior to the final printout. **2.** A computer used for writing, as defined in **1**. **3.** Software that allows a computer to be used for writing, as defined in **1**.

word rate In a communications or computer system, the number of words per unit time (e.g., WORDS PER MINUTE).

word size See WORD LENGTH.

words per minute In telegraphy, a measure of data speed. It is approximately equal to the number of characters (including spaces) per minute divided by 6.

word time In computer operations, the time required to process one word that is in storage.

work **1.** Symbol, *W*. Units: joule, erg, foot-pound, kilogram-meter. That which is accomplished by the transfer of energy from one body to another, as when an exerted force causes a displacement. The amount of work performed is equal to force times distance: $W = Fd$. **2.** An amateur radio term meaning to engage in two-way communication with another station.

work area In computer operations, a temporary area of memory for data items being processed. Also called INTERMEDIATE STORAGE, WORKING MEMORY, and WORKING STORAGE.

work coil The alternating-current-carrying coil that induces energy in the workpiece in induction heating.

work envelope The range of motion over which a robot arm can move. It can be two-dimensional or three-dimensional.

work function Unit, eV. The energy required to bring an internal particle to the surface of a material and out into space, as when an electron is emitted by the hot cathode of a vacuum tube. The work function is the voltage required to extract 1 electrostatic unit of electricity from the material.

working data file A temporary accumulation of data sets that is erased or otherwise discarded after its transferal to another medium.

working life The expected or guaranteed lifetime of a material, device, or system in actual operation or use. Compare SHELF LIFE.

working memory See WORK AREA.

working point The operating point of an active device (i.e., the point along one of the characteristic curves around which operation is fixed).

working *Q* The *Q* of a loaded circuit or device (e.g., tank-circuit *Q* of a radio transmitter loaded with an antenna or dummy load).

working range **1.** The usable maximum distance between a transmitter and receiver in a wireless communications circuit. **2.** The allowed range over which specified parameters can vary in a particular system while facilitating normal operation, or operation within rated specifications.

working storage See WORK AREA.

working voltage The (usually maximum) voltage at which a circuit or device can be operated continuously with safety and reliability.

workout **1.** A dry run; a test. Example: checking a computer's processing power using a set of benchmark programs. **2.** An exacting test of equipment; a burn-in procedure.

workpiece The object heated by an induction or dielectric heater.

worm A cylindrical gear whose spiral "teeth" resemble a screw thread.

worm drive A mechanism for transferring motion from a tuning-knob shaft through a right angle to the shaft of an adjustable component, such as a potentiometer or variable capacitor. The knob turns a threaded shaft that mates with a gear wheel.

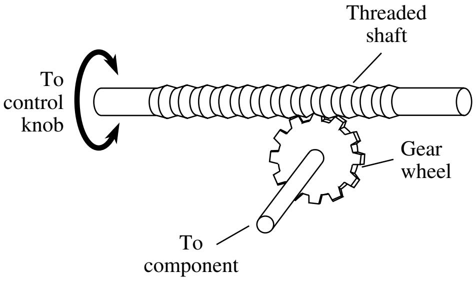

worm drive

worst-case circuit analysis Analysis that seeks the worst possible effects of variations in circuit parameters on circuit performance—especially variations in component characteristics.

worst-case noise pattern See DOUBLE-CHECKERBOARD PATTERN.

worst-case design The design of electronic equipment in such a way that normal operation is obtained—even though the characteristics of circuit components might vary widely.

woven resistor A resistance element made of strands of resistance material woven in the form of gauze.

wow Slow, periodic variations in the pitch of reproduced sound because of variations in the speed of the drive mechanism. Its name is derived from its characteristic sound. Compare FLUTTER.

wow meter An instrument that indicates the amount of wow produced by a turntable or other moving part.

Wpc Abbreviation of *watts per candle.*

wpm Abbreviation of *words per minute.*

wrap See WIRE-WRAP CONNECTION.

wrap-around **1.** The extent of curvature in magnetic tape passing over the heads during recording or playback. **2.** An enclosure that resembles a sheet (wrapped around a piece of electronic equipment). **3.** In computer and data-processing operations, a technique in which the display is cleared of data once it is filled (all available lines are occupied) and additional data is traced from top to bottom.

wrap joint See WIRE-WRAP CONNECTION.

wrapper A paper or tape wound around a component, such as a coil or capacitor, for insulation and protection.

wrapping **1.** See WRAPPER. **2.** Insulating a wire or other conductor by wrapping insulating tape around it.

wrap post A terminal post, tip, pin, or lug used in a wire-wrap connection.

wrap-up The (usually successful) completion of a design, fabrication, test, or investigation.

Wratten filter A light filter for separating colors. It is available in transparent sheets of various colors and is useful in photography and in several phases of electronics, including the operation of color meters and color matchers.

wrinkle finish A pattern of fine wrinkles created by special paint when it dries on a surface, such as that of a metal cabinet for a piece of electronic equipment.

wrist-force sensor In robotics, a set of strain gauges that detects the various forces in the joint connecting an end effector to an arm, and sends signals back to the robot controller. The controller can use the signals to direct the movements of the arm and end effector.

write **1.** In computer operations, to transfer data from one form of memory or storage to another form. Example: To transfer data in a computer from random-access memory (RAM) to the hard disk. Compare READ. **2.** To produce an image on the storage mesh in the cathode-ray tube of a storage oscilloscope.

write enable ring See WRITE PERMIT RING.

write gun See WRITING GUN.

write head In a magnetic memory or in a tape recorder or wire recorder used for recording data, the head that magnetizes the drum, tape, disk, or wire. Compare READ HEAD.

write pulse In computer operation, the pulse that causes information to be recorded in a magnetic cell, or sets it to the one-state. Compare READ PULSE.

write time The time taken to write data to a storage device.

writing gun In a storage oscilloscope, the electron gun that produces the image electronically on the storage mesh. It is mounted in the rear of the cathode-ray tube. Compare FLOOD GUN.

writing head See WRITE HEAD.

writing rate In photographing the image on a cathode-ray tube screen, the highest spot speed at which an acceptable picture can be made.

writing speed See WRITING RATE.

writing telegraph See TELAUTOGRAPH.

wrong color The instance of an undesired color (e.g., purple instead of red) in a color-television picture. The condition is often caused by a malfunction of the chroma demodulator(s).

W/(sr•m^2) Abbreviation of *watts per steradian square meter.*

WT **1.** Abbreviation of wireless telegraphy. **2.** Abbreviation of WATERTIGHT.

wt Abbreviation of WEIGHT.

WUI Abbreviation of *Western Union International.*

Wulf electrometer See BIFILAR ELECTROMETER.

Wullenweber antenna An electronically steerable antenna composed of two concentric circular arrays of masts connected to the steering circuitry.

WUX Abbreviation of *Western Union telegram.*

WVdc Abbreviation of *direct-current working voltage:* the maximum continuous direct-current voltage that can safely be placed across a component.

ww Abbreviation of *wirewound.*

W/sr Abbreviation of *watts per steradian* (unit of radiant intensity).

WWV The call letters of a standard-frequency/ standard-time broadcasting station operated by the National Bureau of Standards and located in the continental United States. Also see WWVH.

WWVH The call letters of a standard-frequency/ standard-time broadcasting station operated by the National Bureau of Standards and located in Hawaii. Also see WWV.

wx Radiotelegraph abbreviation of weather.

WXD International Telecommunications Union symbol for *meteorological radar station.*

WXR International Telecommunications Union symbol for *radiosonde station.*

wye adapter A connector that provides two outputs for a single input, or vice versa. It is commonly used in audio applications.

wye box In a three-phase power-measuring setup, a special arrangement of two impedances, each of which is equal to the impedance of the potential element of the wattmeter used in the setup. The box permits a single wattmeter to be used, which indicates one-third the total power. Without the box, three wattmeters would be needed.

wye connection A method of connecting three windings in a three-phase system so that one terminal of each winding is connected to the neutral point. It is shaped like the letter Y. Also called *star connection.*

wye current Current through one of the branches of a three-phase star (wye) connection.

wye delta starter A starter circuit for a three-phase squirrel-cage induction motor. When the starter switch is thrown in one direction, the stator of the motor is wye-connected for starting; thrown in the opposite direction, the stator is delta-connected for continued operation.

wye equivalent circuit A wye-connected three-phase circuit equivalent to a delta-connected circuit when the impedance of any pair of corresponding lines for both wye and delta are the same; the third line is unconnected.

wye junction See WAVEGUIDE WYE.

wye-matched-impedance antenna An antenna in which the impedance of a nonresonant open-wire feed line is matched to that of the center of the radiator by spreading the end of the feeder wires out into a Y-shaped or delta-shaped matching section. Also called DELTA-MATCHED-IMPEDANCE ANTENNA.

wye point The star point in a three-phase system. Also see WYE CONNECTION.

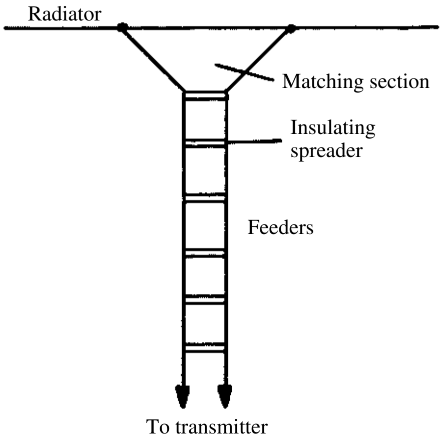

wye-matched-impedance antenna

wye potential In a three-phase armature, the voltage between a terminal and the neutral point.

wye rectifier A three-phase rectifier circuit in which the transformer or generator windings are arranged in a wye connection.

wye-wye circuit A circuit consisting a wye-connected generator and a wye-connected load.

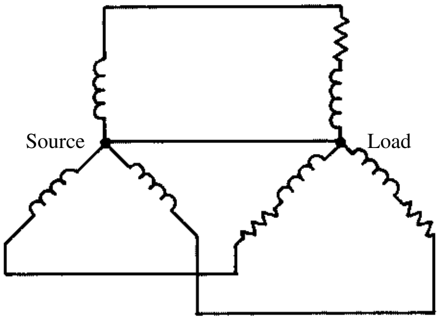

wye-wye circuit

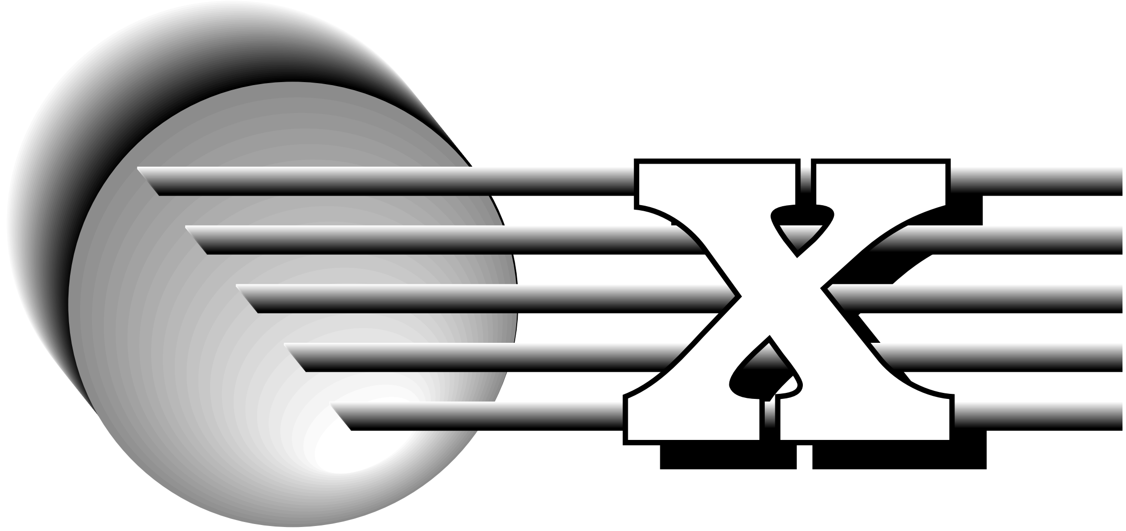

X **1.** Symbol for REACTANCE. **2.** Symbol for *no connection.* (Also, NC.) **3.** Symbol for the HORIZONTAL AXIS of a graph or screen in the rectangular (Cartesian) coordinate system. **4.** Symbol for an unknown quantity.

x **1.** Symbol for number of carriers drawn from collector to base of a transistor, for each carrier collected. **2.** Symbol for an unknown quantity. **3.** Symbol for the HORIZONTAL AXIS of a graph or screen in the rectangular (Cartesian) coordinate system.

x amplifier The horizontal amplifier of an oscilloscope or recorder. Compare Y AMPLIFIER and Z AMPLIFIER.

X and Z demodulation Color television demodulation based on the 60-degree difference between the two reinserted 3.58-MHz subcarrier signals. The R-Y, B-Y, and G-Y voltages derived from the demodulated signals control the three guns of the picture tube. Compare Q MODULATION.

x-axis **1.** The horizontal axis of a chart, graph, or screen in the rectangular (Cartesian) coordinate system. **2.** In a quartz crystal, the axis drawn through the corners of the hexagon.

x-axis amplifier See X AMPLIFIER.

XB Abbreviation of *crossbar.*

X balance The reactance balance (either the variable component or the adjustment of it) in an impedance bridge in which separate resistance (*R*) and reactance (*X*) balancing is provided.

X band The frequency band extending from 5.2 to 11 GHz.

X bar A rectangular, piezoelectric quartz bar cut from a Z-section, whose faces are parallel to the X-axis and whose edges are parallel to the X-, Y-, and Z-axes. See CRYSTAL CUTS.

X bridge An alternating-current bridge for measuring reactance.

X_C Symbol for CAPACITIVE REACTANCE.

X channel The horizontal channel of an oscilloscope or recorder. Compare Y CHANNEL and Z CHANNEL.

X-channel gain See X GAIN.

X component In a complex impedance, the reactive component (either inductive or capacitive).

x coordinate See ABSCISSA.

X-cut crystal A piezoelectric plate cut with its faces perpendicular to the X-axis of a quartz crystal. See CRYSTAL CUTS.

xcvr Abbreviation of TRANSCEIVER. Often capitalized.

X deflection Horizontal deflection of the spot on the screen of a cathode-ray tube. Compare Y DEFLECTION.

X diode The decoding diode in each of the X lines of a memory matrix. Compare Y DIODE.

X direction The horizontal direction in deflections and in graphical presentations of data.

X drive The driving source or energy for the X lines of a computer memory matrix.

xducer Abbreviation of TRANSDUCER.

Xe Symbol for XENON.

xenon Symbol, Xe. An inert gaseous element. Atomic number, 54. Atomic weight, 131.29. Xenon is present in trace amounts in the earth's atmosphere and is used in some thyratrons, electric lamps, and lasers.

xenon tube A flash tube filled with xenon.

xerography A system for the reproduction of printed matter, drawings, and other graphic matter. It is an electrostatic process in which a charged photoconductive surface is exposed to an

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

image of the material to be copied. A latent image is formed on the surface and is developed (by dusting with black powder attracted to the charged image) to make the image visible.

xeroradiography Xerography using X RAYS.

xerothermic A condition characterized by both heat and dryness.

Xerox **1.** Name of a company that manufactures a wide variety of office and computer equipment. **2.** Trade name for a XEROGRAPHY machine. **3.** The reproduction obtained by means of XEROGRAPHY (generic term). **4.** To make a xerographic reproduction (generic term).

X factor **1.** An unknown or unidentifiable quantity or parameter. **2.** See X COMPONENT.

xformer Abbreviation of TRANSFORMER.

X gain The gain (or gain control) of the horizontal channel of an oscilloscope or X-Y recorder. Compare Y GAIN and Z GAIN.

X-H array See LAZY-H ANTENNA.

XHV Abbreviation of EXTREMELY HIGH VACUUM.

x intercept The *x* coordinate of the point at which a line or plane intersects the X-AXIS.

x irradiate To expose to X RAYS.

xistor Abbreviation of TRANSISTOR.

X_L Symbol for INDUCTIVE REACTANCE.

X line A horizontal line in a memory matrix. Compare Y LINE.

XLR connector A microphone connector with a locking device to prevent unintentional disconnection. It has three pins and is commonly used with balanced audio systems.

X meter **1.** An instrument for measuring reactance and phase angle. **2.** A form of simple capacitance meter whose dial is direct-reading in microfarads, although the deflection is actually proportional to the reactance of the capacitor under measurement.

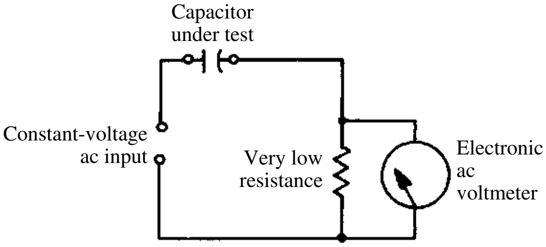

X meter, 2

xmission Abbreviation of *transmission.*

xmit Abbreviation of *transmit.* Also, xmt.

xmitter Abbreviation of TRANSMITTER. Also, xmtr.

XMODEM In data transmission, an error-correcting mode in which data is sent in blocks of 128 kilobytes (128K), or 131,072 bytes. The source and destination tally up the bytes in each block. If they agree, the next data block is sent. If they do not agree, the current data block is retransmitted. Compare YMODEM and ZMODEM.

xmt Abbreviation of *transmit.* Also, xmit.

xmtr Abbreviation of TRANSMITTER. Also, xmitter.

XOR Abbreviation of EXCLUSIVE-OR.

xover Abbreviation of CROSSOVER.

X particle See MESON.

xponder Abbreviation of TRANSPONDER.

X position **1.** The alignment of a cathode-ray beam along the horizontal axis of an oscilloscope screen. **2.** The position of a point, with respect to the horizontal axis of a graph in the rectangular (Cartesian) coordinate system.

XR Abbreviation of INDEX REGISTER.

X radiation Electromagnetic energy in the form of X rays.

X-ray **1.** Phonetic alphabet communications code word for the letter X. **2.** Pertaining to, or consisting of, X rays.

X-ray astronomy The science of observing celestial objects and the sky in the X-ray band of wavelengths. Generally, this must be done from above the earth's atmosphere because the atmosphere absorbs X rays.

X-ray crystallography The science of observing atomic patterns in a crystal by means of X rays.

X-ray detecting device **1.** An X-ray instrument for spotting flaws in solid bodies. **2.** A device, such as a fluoroscope, for showing the presence of X rays.

X-ray diagnosis The use of X rays in the diagnosis of disease and in the observation of internal parts of the body.

X-ray diffraction The diffraction of X rays by a material into or onto which they are directed. Also see X-RAY DIFFRACTION CAMERA and X-RAY DIFFRACTION PATTERN.

X-ray diffraction camera A special camera that furnishes a photograph of the pattern created by X rays diffracted by a material. See X-RAY DIFFRACTION and X-RAY DIFFRACTION PATTERN.

X-ray diffraction pattern The pattern produced on film exposed to X rays diffracted by a material. Also see X-RAY DIFFRACTION and X-RAY DIFFRACTION CAMERA.

X-ray goniometer An instrument used to find the position of the axes of a quartz crystal. It uses X rays reflected from the atomic planes of the crystal.

X-ray inspection The use of X rays in the examination and study of the internal features of materials and devices.

X-ray laser A laser designed to emit coherent X rays in a narrow beam.

X-ray load **1.** An X-ray tube that is the terminal member of an electronic system. **2.** A body, mass, or material exposed to X rays.

X-ray machine A device that generates X rays for a specific purpose, such as medical diagnosis.

X-ray photograph A photograph made by exposure to X rays, especially an *X-ray shadowgram,* a picture made without a camera by interposing an object (such as a part of the human body) between an X-ray tube and photographic film.

X-ray radiation See X RADIATION.

X rays Invisible, electromagnetic radiation having wavelengths ranging from approximately 0.01 nanometer (10^{-11} meter) to 0.15 nanometer (1.5×10^{-10} meter). These waves are shorter than ultraviolet, but longer than gamma rays. They can be produced by bombarding a target of heavy metal (such as tungsten) with a stream of high-speed electrons in a vacuum tube. X rays have high penetrating power, can expose photographic film, and cause some substances to fluoresce. Also see X-RAY DIAGNOSIS, X-RAY INSPECTION, and X-RAY THERAPY.

X-ray spectra The continuous band of ionizing electromagnetic radiation having wavelengths ranging from approximately 0.01 nanometers to 0.15 nanometers.

X-ray spectrograph A spectrograph used to disperse and measure the wavelength of X rays.

X-ray spectrometer A spectrometer with which the diffraction angle of X rays reflected from the surface of a crystal can be measured. The device allows the characteristics and composition of almost any material to be studied.

X-ray star A collapsed star that emits high-intensity energy concentrated mainly in the X-ray region of the electromagnetic spectrum.

X-ray system Collectively, an X-ray tube and the associated equipment required for a specific application, such as crystallography, irradiation, or medical therapy.

X-ray technician **1.** A professional skilled in the operation and maintenance of X-ray systems, who usually works under the supervision of an engineer. **2.** A professional skilled in the medical use of X rays, who works under the supervision of a medical doctor.

X-ray therapy The use of X rays in the treatment of certain physiological diseases.

X-ray therapy system An X-ray system designed for X-ray therapy.

X-ray thickness gauge An instrument used to measure the thickness of metal, such as a continuously moving sheet of steel. The measurement is in terms of the amount of absorption of an applied X-ray beam by the metal.

X-ray tube A specialized, very-high-voltage diode in which a high-speed electron beam bombards an anode (target) of heavy metal, such as tungsten, causing it to emit X rays. The target is tilted to reflect the X rays through the glass wall of the tube.

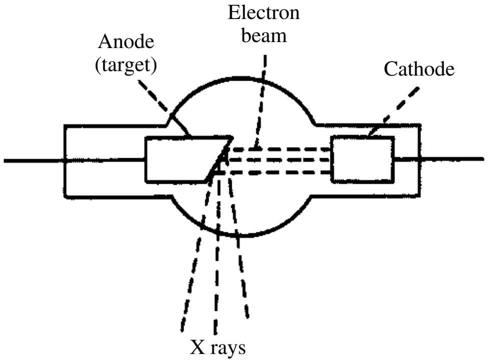

X-ray tube

X section Abbreviation of *cross section.*

X sink The circuit or device into which the X lines of a memory matrix feed. Compare Y SINK.

X_T Symbol for total reactance.

xtal Abbreviation of CRYSTAL.

xtalk Abbreviation of CROSSTALK.

x unit A small unit of length, equal to approximately 10^{-13} meter or 10^{-4} nanometer. Now seldom used, the x unit was once often used in the expression of ultraviolet and X-ray wavelengths.

X wave One (the extraordinary) of the pair of components into which an ionospheric radio wave is divided by the earth's magnetic field. Compare O-WAVE.

XXX In radiotelegraphy, a signal meaning *urgent.*

X-Y counting A technique used with electronic counters to determine the ratio of one frequency (X) to another (Y). When X is the higher frequency, the counter readout shows the number of pulses of X producing one cycle of Y.

XY-cut crystal A piezoelectric plate cut from a quartz crystal at such an angle that its electrical characteristics fall between those of the X-cut and Y-cut crystal.

x-y plotter An output device similar to the x-y recorder. It allows a digital computer to plot graphs. Also called *data plotter.*

x-y recorder An instrument that produces a permanent record (photographic print or directly inked paper) of a variable quantity on a chart having Cartesian (rectangular) coordinates.

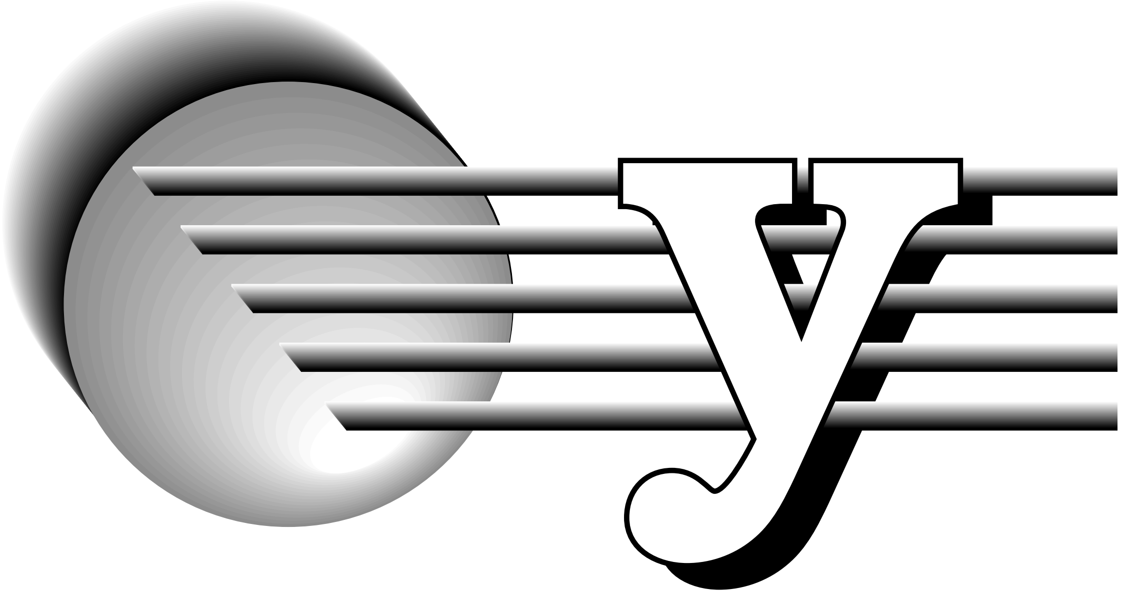

Y **1.** Symbol for ADMITTANCE. **2.** Symbol for YTTRIUM. **3.** Symbol for YOUNG'S MODULUS.

y **1.** Abbreviation of YEAR. (Also, yr.) **2.** Abbreviation of YARD. (Also, yd.) **3.** Symbol for the vertical axis of a graph or screen in the rectangular (Cartesian) coordinate system.

Y adapter See WYE ADAPTER.

YAG Abbreviation of *yttrium-aluminum-garnet*, the stimulated substance in some lasers.

Yagi antenna Also called *Yagi-Uda array* or *beam.* An antenna consisting of two or more parallel, straight elements, including at least one parasitic element and at least one driven element. The elements all lie in the same plane. Each driven element is connected to the feed line, half-wave resonant, and center-fed. A two-element array can be formed by adding a director or a reflector alongside a single driven element. An array with one director and one reflector, along with the driven element, increases the gain and directivity compared with a two-element system. The gain and directivity increase further as elements are added. This is usually done by placing additional directors in front of a three-element array. Compare QUAD ANTENNA.

y amplifier The vertical amplifier of an oscilloscope or recorder. Compare X AMPLIFIER and **Z** AMPLIFIER.

Yankee Phonetic alphabet communications code word for the letter Y.

yard Abbreviation, yd. A unit of linear measure in the English system. 1 yd = 3 ft = 36 inches = 0.9144 meter.

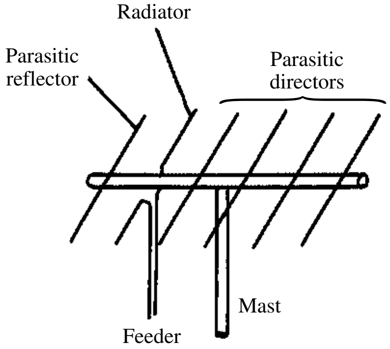

yagi antenna

yaw Side-to-side movement in a vehicle or robotic end effector. Essentially horizontal displacement about the vertical axis.

yaw amplifier In an aircraft servo system, the unit that amplifies the yaw signal (the signal proportional to the deviation of the aircraft from the line of flight).

yaw meter An instrument for measuring the angle of yaw of an aircraft.

y-axis **1.** The vertical axis of a chart, graph, or screen in the rectangular coordinate system.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

2. In a quartz crystal, the axis drawn perpendicular to the faces of the hexagon.

y-axis amplifier See Y AMPLIFIER.

Yb Symbol for YTTERBIUM.

Y box See WYE BOX.

Y channel See Y AMPLIFIER.

Y-channel gain See Y GAIN.

Y circulator An interconnection among three waveguides. When power is applied to the junction through one waveguide, the wave is transferred to the adjacent waveguide immediately to the right or left.

Y connection See WYE CONNECTION.

y coordinate See ORDINATE.

Y current See WYE CURRENT.

Y-cut crystal A piezoelectric plate cut from a quartz crystal in such a way that the plane of the plate is perpendicular to the Y-axis of the crystal. See CRYSTAL AXES and Y-AXIS, **2**.

yd Abbreviation of YARD.

Y deflection Vertical deflection of the spot on the screen of a cathode-ray tube. Compare X DEFLECTION.

Y diode The decoding diode in each of the Y lines of a memory matrix. Compare X DIODE.

y direction The vertical direction in deflections and in graphical presentations of data.

Y drive The driving source or energy for the y lines of a computer memory matrix. Compare X DRIVE.

year Abbreviation, y or yr. The period of the earth's revolution around the sun, with respect to distant stars: approximately 365.25 days or 3.15576×10^7 seconds.

Yellow Book A specialized format for *compact-disk read-only memory (CD-ROM)* computer data storage media, developed by Sony and Philips. It is similar to RED BOOK, but it includes superior error correction and storage capacity. See also CD-ROM, GREEN BOOK, ORANGE BOOK, and RED BOOK.

yellow copper ore See CHALCOPYRITE.

yellow metal A copper-zinc alloy that is 60% copper.

yellow pyrites See CHALCOPYRITE.

Y-equivalent circuit See WYE-EQUIVALENT CIRCUIT.

Y gain The gain (or gain control) of the vertical channel of an oscilloscope or x-y recorder. Compare X GAIN and Z GAIN.

yield In a phototube or photocell, the photoelectric current per unit intensity of light.

yield map A diagram of an integrated circuit showing the locations of faulty components.

yield strength The lowest stress for plastic deformation of a material, below which the material is elastic and above which it is viscous.

yield value The amount of physical stress that causes a substance to become stretched permanently out of shape.

YIG Abbreviation of *yttrium-iron-garnet*, a crystalline material used in certain types of acoustic delay lines, parametric amplifiers, and filters.

YIG crystal A crystal of yttrium-iron-garnet. See YIG.

YIG filter A filter using a YIG crystal and tuned electromagnetically.

YIG-tuned parametric amplifier A parametric amplifier in which tuning is accomplished by continuously varying direct current through the solenoid of a YIG filter.

y intercept The y coordinate of the point at which a line or plane intersects the y-axis.

ylem The material from which the primordial fireball is thought to have been made, and from which the entire known universe is believed to have originated approximately 10 billion (10^{10}) years ago.

Y line A vertical line of a memory matrix. Compare X LINE.

Y-matched-impedance antenna See WYE-MATCHED-IMPEDANCE ANTENNA.

YMODEM In data transmission, an error-correcting mode in which data is sent in blocks of one megabyte (1MB), or 1,048,576 bytes. The source and destination tally up the bytes in each block. If they agree, the next data block is sent. If they do not agree, the current data block is retransmitted. Compare XMODEM and ZMODEM.

yoke **1.** The ferromagnetic ring or cylinder that holds the pole pieces of a dynamo-type generator and acts as part of the magnetic circuit. **2.** The system of coils used for magnetic deflection of the electron beam in cathode-ray tubes.

yoke method A method of measuring the permeability of a sample of magnetic material. It involves completing the magnetic circuit with a heavy yoke of soft iron.

Young's modulus Symbol, Y. The ratio of tensile stress to tensile strain.

Y parameters The admittance parameters of a four-terminal network or device.

Y point See WYE POINT.

Y position **1.** On a cathode-ray screen, the vertical position of the beam spot. **2.** On a graph, the position of a point along the vertical axis.

Y potential See WYE POTENTIAL.

yr Abbreviation of YEAR. Also, y.

Y rectifier See WYE RECTIFIER.

Y signal In color television, the monochromatic signal conveying brightness information.

Y sink The circuit or device into which the Y lines of a memory matrix feed. Compare X SINK.

ytterbium Symbol, Yb. A metallic element of the rare-earth group. Atomic number, 70. Atomic weight, 173.04.

ytterbium metals The rare-earth metals dysprosium, erbium, lutetium, holmium, thulium, and ytterbium.

Yttralox A transparent polycrystalline ceramic composed primarily of yttrium oxide, which has many applications in electro-optics.

yttria Formula, Y_2O_3. Yttrium oxide, a white powder used in Nernst lamps.

yttrium Symbol, Y. A metallic element of the rare-earth-metals group. Atomic number, 39. Atomic weight, 88.906.

yttrium-iron-garnet See YIG.

yttrium metal Any of the group of metals, including yttrium, erbium, holmium, lutetium, thulium, ytterbium, and sometimes dysprosium, gadolinium, and terbium.

Yukon Standard Time Standard time of the ninth time zone west of Greenwich, embracing the Yukon Territory and a portion of southern Alaska.

Y winding See WYE CONNECTION.

Y-Y circuit See WYE-WYE CIRCUIT.

Z

Z **1.** Symbol for IMPEDANCE. **2.** Symbol for ATOMIC NUMBER. **3.** Symbol for *zenith distance* (astronomy).

z **1.** Abbreviation of ZERO. **2.** Symbol for ELECTROCHEMICAL EQUIVALENT. **3.** Abbreviation of ZONE. **4.** Symbol for the vertical axis of a three-dimensional Cartesian graph.

zag The short, straight deflection of a point or particle, or of waves along a jagged path in a direction opposite that of a ZIG; one of the components of ZIGZAG DEFLECTION.

Zamboni pile A high-voltage electrochemical cell consisting of an aluminum anode, a manganese-dioxide cathode, and an aluminum-chloride electrolyte.

Z amplifier The intensity-modulation amplifier in an oscilloscope. Compare X AMPLIFIER and Y AMPLIFIER.

Z-angle meter An instrument, commonly of the null-balance type, that indicates the impedance and phase angle of capacitors, inductors, and sometimes of inductance-capacitance-resistance (ICR) combinations.

z-axis **1.** The intensity-modulation input of an oscilloscope, including the associated circuit. **2.** The video-signal input of a television picture tube, including the associated circuit. **3.** The third axis in a three-dimensional coordinate system. **4.** The lengthwise axis of a quartz crystal.

z-axis amplifier See Z AMPLIFIER.

z-axis modulation See INTENSITY MODULATION.

Z bar See Z-CUT CRYSTAL.

z channel See Z AMPLIFIER.

z-channel gain See Z GAIN.

Z-cut crystal A plate cut from a quartz crystal so that the plane of the plate is perpendicular to the Z-axis of the crystal and parallel to the X-axis. See CRYSTAL AXES.

ZD Abbreviation of *zero defects.*

Z deflection Deflection of a cathode-ray beam beyond the (ordinarily defined) deflection area on the tube screen.

Zeeman effect The splitting of a spectral line of a gas into closely spaced, polarized frequency components by an applied magnetic field.

Zenely electroscope A sensitive alpha-ray electroscope (see ELECTROSCOPE) whose leaf is attracted by a metal plate biased at 50 to 200 volts. On touching the plate, the leaf becomes charged and is repelled. But the gas ions around the leaf neutralize the charge, and the leaf returns to the plate to repeat the action. This sequence causes the leaf to oscillate, the number of oscillations per second being proportional to the strength of the ionizing source.

Zener See ZENER DIODE.

Zener breakdown See AVALANCHE BREAKDOWN.

Zener current See AVALANCHE CURRENT.

Zener diode Also sometimes called *avalanche diode* or *Zener.* A semiconductor diode (usually silicon) specially fabricated to take advantage of the effects of avalanche breakdown. It is available in a wide variety of configurations and ratings; it is commonly used as a voltage regulating device in low-voltage power supplies.

Zener-diode clipper A signal-amplitude clipper that does not require direct-current bias because it uses a Zener diode. Clipping takes place at the avalanche voltage. Zener diodes can be connected back to back for alternating-current clipping.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

Zener-diode coupling element A Zener diode used as a direct coupling element between amplifier stages in an electronic system. When no signal is present, the resistance of the diode is extremely high because the direct-current voltages of the diode-coupled amplifier stages reverse-bias the diode to below the avalanche point, and no current flows from one stage to the next. The signal raises the voltage enough to cause avalanche breakdown; the signal is thus readily transmitted from one stage to the following one.

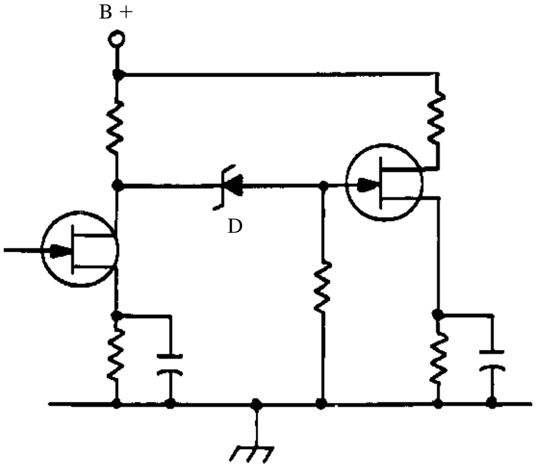

zener-diode coupling element

Zener-diode voltage reference A device that utilizes the constant voltage drop across a Zener diode operated in its breakdown region, to provide a standard voltage for reference purposes. It consists essentially of a direct-current (dc) voltage source, limiting resistor, and Zener diode. For an alternating-current (ac) reference voltage, an ac voltage source, limiting resistor, and two Zener diodes (connected in parallel, back to back) are required. Zener diodes can be connected in series to supply a higher reference voltage than can be delivered by a single diode.

Zener-diode voltage regulator A simple voltage regulator whose output is the constant voltage drop developed across a Zener diode conducting in the reverse breakdown region. The simple regulator circuit consists of a Zener diode and an appropriate limiting resistor connected in series across the output terminals of an unregulated direct-current power supply. For alternating-current voltage regulation, two Zener diodes can be connected in parallel, back to back. Zener diodes can be connected in series to regulate a higher voltage than can be accommodated by a single diode.

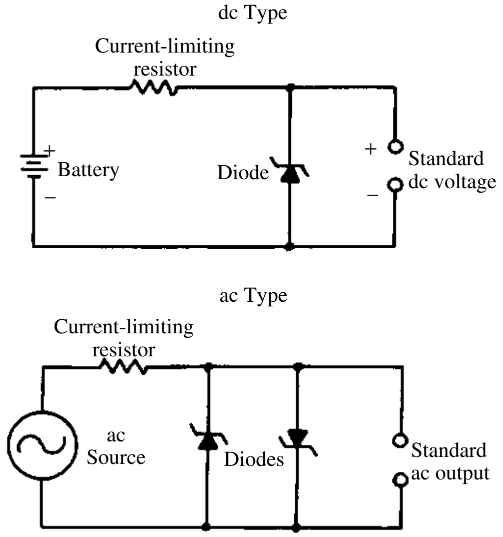
dc Type
Current-limiting resistor
+
Battery
−
Diode
+
−
Standard dc voltage
ac Type
Current-limiting resistor
ac Source
Diodes
Standard ac output

zener-diode voltage reference

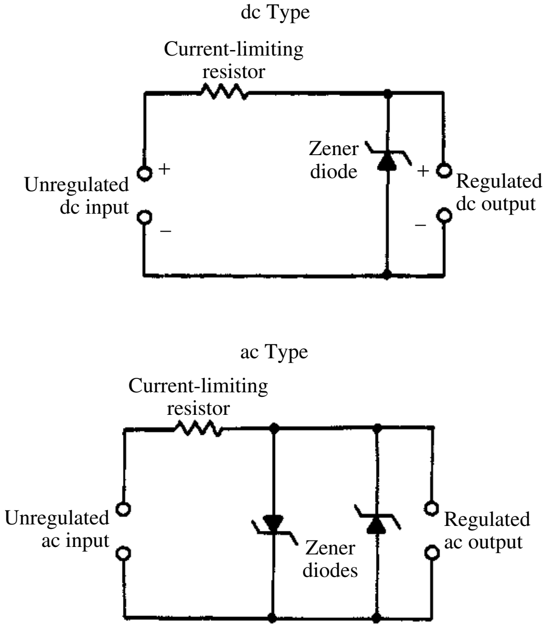
dc Type
Current-limiting resistor
Unregulated dc input
+
−
Zener diode
+
−
Regulated dc output
ac Type
Current-limiting resistor
Unregulated ac input
Zener diodes
Regulated ac output

zener-diode voltage regulator

Zener effect See AVALANCHE BREAKDOWN.

Zener knee In the response of a reverse-biased Zener diode, the point of abrupt transition from low current (high resistance) to high current (low resistance). For voltage regulation and voltage standardization, the knee should be as sharp as possible.

Zener knee current In a Zener diode, the current that flows when the reverse bias reaches the avalanche voltage.

Zener-protected MOSFET See GATE-PROTECTED MOSFET.

Zener voltage The particular value of reverse voltage at which a Zener diode or other reverse-biased pn junction abruptly exhibits the avalanche effect. Depending on the Zener diode, this potential can be less than 10 volts, or as much as several hundred volts.

zenith In the sky, the point directly overhead, exactly 180 degrees opposite the direction of the earth's center.

Zeolite process A method using certain artificial Zeolites (hydrous silicates) to soften water used in some electronic manufacturing and testing operations.

zeppelin antenna Also called *zepp.* **1.** A half-wavelength radiator, fed at one end with a quarter-wavelength section of open-wire line. The feed line can come away from the radiator at any angle, usually 90 degrees or more. The antenna was originally used aboard zeppelins; the entire system was dangled in flight and the feed line was collinear with the radiating element. The impedance at the feed point is high; the impedance at the transmitter end of the feed line is low. The antenna will operate satisfactorily at all harmonics of the design frequency. **2.** A radiator that is an integral multiple of a half wavelength, fed with open-wire line of any length. A transmatch is employed at the transmitter and of the feed line. This arrangement is popular among radio amateurs. The primary advantage of this antenna is ease of installation; the main drawback is the fact that the feed line radiates to some extent because the system is not perfectly balanced.

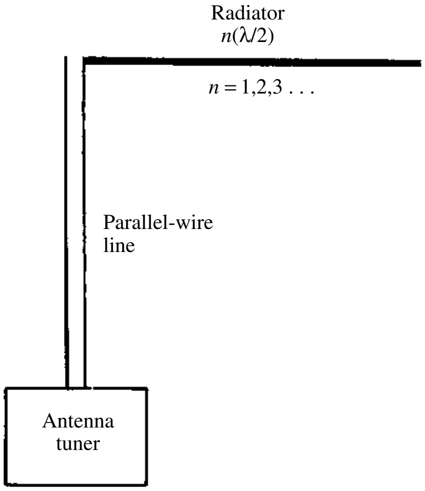

zeppelin antenna

zero **1.** The number represented by the cipher symbol (0). **2.** To set a meter or other instrument to its zero reading or condition. **3.** To align one element of a system precisely with another. **4.** In computer operations, to set a register to zero. **5.** In computer operations, using zero pulses to replace what is in a storage area.

zero-access memory See ZERO-ACCESS STORAGE.

zero-access storage Computer storage requiring negligible waiting time (latency).

zero-address instruction A computer instruction requiring no address; the operation specified by the instruction defines the locations of the operands. Also called *addressless instruction.*

zero adjust In an analog metering device, a mechanical or electrical control that allows precise setting of the reading to zero, when the parameter to be measured is actually zero.

zero adjustment **1.** The act of setting an instrument or circuit to its zero reading or zero-operating condition. **2.** A device or subcircuit used to set a meter to its zero reading. Also see ZERO SET, **1**, **2**.

zero-angle cut An alternate term for *X cut*, as applied to quartz crystal plates. Also see CRYSTAL CUTS.

zero axis In a plot of a variable quantity, the (usually horizontal) reference line that indicates the zero value of the quantity.

zero-axis symmetry The condition in which a waveshape is symmetrical about a zero axis.

zero beat Complete absence of a beat note (i.e., silence), a condition occurring when the frequencies of the beating signals (or their harmonics) are equal. See BEAT NOTE.

zero-beat detector A device or circuit used to sense and indicate the condition of zero beat.

zero-beat indicator See ZERO-BEAT DETECTOR.

zero-beat method A means of adjusting the frequency of one signal exactly to that of another (usually standard) signal by setting the first signal frequency to zero beat with the second signal frequency.

zero-beat reception A system of reception entailing zero beating an incoming signal with the signal from a local oscillator. Also called *homodyne reception.*

zero-bias operation Operation of a transistor or vacuum tube without direct-current base, gate, or grid bias.

zero-bias tube A vacuum tube designed for operation (particularly in a class-B power amplifier) without direct-current grid bias. In such a tube, the zero-signal plate current is very low because of the large amplification factor.

zero capacitance **1.** Absence of capacitance (a theoretically ideal condition). **2.** In some circuits,

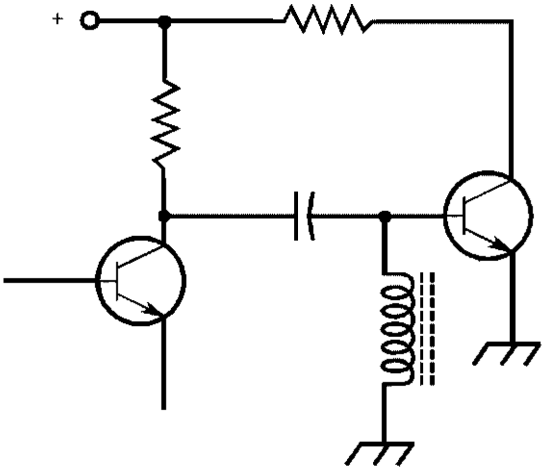

zero-bias operation

the lowest capacitance point, to which all other capacitances are referred.

zero compensation The elimination or minimizing of the zero-signal output of a transducer or similar device.

zero compression See ZERO SUPPRESSION, **1**.

zero condition See ZERO STATE.

zero current **1.** Absence of movement of electrical charge carriers. **2.** In some circuits, the lowest current level, to which all other currents are referred.

zero-cut crystal A piezoelectric plate cut from a quartz crystal in such a way that the frequency-temperature coefficient is zero.

zero-dB reference level An agreed-upon zero level for decibel ratings (which are by nature relative). Zero dBm, for example, corresponds to 1 milliwatt. Compare VOLUME UNIT.

zero drift **1.** The (usually gradual) drift of a zero point, such as the zero setting of an electronic voltmeter. **2.** The condition of no change in the value of a quantity.

zero elimination See ZERO SUPPRESSION, **1**.

zero energy The condition in which energy is neither generated, consumed, nor dissipated.

zero error **1.** In instruments and measurements, an error so small that it can be considered inconsequential. **2.** In a radar system, the inherent delay in the transmitter and receiver circuits.

zero field emission Thermionic emission from a cathode or hot conductor within a uniform electrostatic field.

zero fill See ZERO, **5**.

zero gravity The condition of weightlessness (i.e., the state in which gravitational pull by a celestial body is completely absent or has been nullified).

zero-gravity switch A switch actuated automatically when the condition of zero gravity is reached.

zero impedance **1.** Absence of impedance (a theoretically ideal condition). **2.** In some circuits, the lowest impedance level, to which all other impedances are referred.

zero inductance **1.** Absence of inductance (a theoretically ideal condition). **2.** In some circuits, the lowest inductance level, to which all other inductances are referred.

zero input **1.** Complete absence of input (signal or noise). **2.** Absence of operating voltage or power to a system. **3.** In a flip-flop, the input terminal that is not receiving a trigger signal.

zero-input terminal In a flip-flop, the input-signal terminal at which a trigger signal will switch the flip-flop output to zero. Also called *zero terminal* and RESET TERMINAL.

zero instant See ZERO TIME.

zero level The reference level for the comparison of quantities. For example, it might be a voltage or current level; in audio measurements, it is the zero-dB reference level.

zero-line stability In an analog metering device, the ability of the instrument to maintain proper zero adjustment over a period of time.

zero magnet A permanent magnet used to set the pointer of a meter to zero.

zero magnitude **1.** For a quantity, the state of being valueless (i.e., a complete absence of the quantity). **2.** In some tests, measurements, or calculations, the lowest value of a quantity, to which all other values are referred.

zero meridian The meridian at Greenwich (near London), England, from which longitude and standard time are reckoned. Also see TIME ZONE and ZONE TIME.

zero method A method of measurement entailing adjustment of a circuit or device (such as a bridge, tee network, or potentiometer) for zero response of the detector. Also called *null method.*

zero modulation The momentary lack of modulation in a communications or broadcast system, as during a pause in speech.

zero-modulation noise Noise produced by previously erased tape that is run under specified operating conditions.

zero output **1.** Complete absence of output signal or output power, sometimes disregarding noise output. **2.** In a flip-flop, the normal condition of no signal pulse at a particular output terminal.

zero-output terminal In a flip-flop, the output terminal that is not delivering an output pulse.

zero phase In an alternating-current circuit, the condition in which the current and voltage are in step. That is, the current peaks occur at exactly the same times as the voltage peaks; the phase difference between the current and voltage is zero degrees.

zero pole In a multiconductor line, the common or reference conductor.

zero potential **1.** Complete absence of voltage. **2.** In some circuits, the lowest voltage, to which all other voltages are referred. **3.** The potential of the earth as a reference point.

zero power **1.** Complete absence of dissipated power. **2.** In some circuits and systems, the lowest power level, to which all other power values are referred.

zero-power resistance In a thermistor, the resistance at which power dissipation is zero.

zero-power resistance-temperature characteristic For a thermistor, a figure that reveals the extent to which zero-power resistance varies with the temperature of the thermistor body.

zero-power temperature coefficient of resistance A temperature coefficient that reveals the extent to which the temperature of the thermistor body causes the zero-power resistance to change (expressed in ohms per ohm per degree Celsius).

zero reactance **1.** Absence of reactance (a theoretically ideal condition in alternating-current circuits). **2.** In some circuits, the lowest reactance, to which all other reactances are referred.

zero resistance **1.** Absence of resistance (a theoretically ideal condition). **2.** In some circuits, the lowest resistance, to which all other resistances are referred.

zero scale current In a digital-to-analog (D/A) converter, the current into the output when all logic inputs are low (off) and the output is at a certain predetermined value, in microamperes or milliamperes.

zero screw The mechanical zero adjuster of a meter.

zero set **1.** A (usually screwdriver-adjusted) mechanism for setting the pointer of a meter to zero. **2.** An electrical circuit consisting of a resistance bridge or adjustable bucking voltage for setting a meter to read zero under specific conditions.

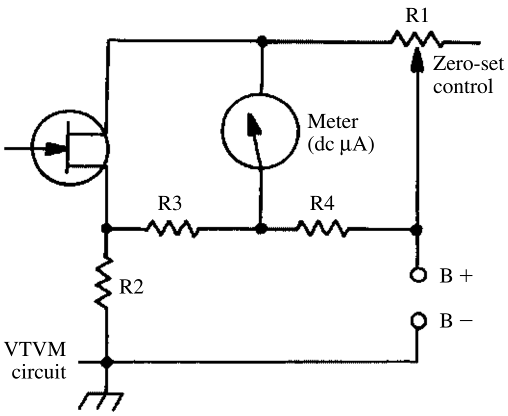

zero set, 2

zero shift See ZERO DRIFT.

zero signal **1.** The condition of complete absence of signal. **2.** A finite signal level used as a reference point against which all other signal levels are measured.

zeros of impedance For an alternating-current network, the frequencies at which the impedance is zero. Also called *zeros.*

zeros of network function The real or complex values at which the network function is zero. Compare POLES OF NETWORK FUNCTION.

zeros of transfer function The frequencies at which a transfer function becomes zero. Compare POLES OF TRANSFER FUNCTION.

zero stability Constancy of the zero condition in an instrument or system (i.e., absence of zero drift).

zero state The low, zero, off, or false logic state of a bistable device, such as a flip-flop or magnetic cell. It might be actual zero output or a low-voltage output. Compare ONE STATE. In binary notation and calculation, the zero state is represented by a cipher.

zero suppression **1.** In computer operation, automatic nonsignificant leading-zero cancellation. **2.** Absence or removal of a restraining medium, such as a noise-suppression voltage. **3.** In an audio recording system, the introduction of a voltage to cancel the steady-state component of the input signal.

zero temperature **1.** The point from which temperatures are reckoned on a thermometer scale. On the Celsius (centigrade) scale, zero degrees corresponds to the freezing point of pure water at standard atmospheric pressure. On the Fahrenheit scale, zero degrees is 32 degrees colder than the freezing point of pure water at standard atmospheric pressure. On the Kelvin or Rankine scales, zero degrees corresponds to the complete absence of thermal energy; it is the coldest possible temperature. **2.** A temperature point relative to which all other temperatures are reckoned.

zero temperature coefficient A temperature coefficient having the value zero (i.e., one that indicates there is no temperature-dependent drift of a given quantity).

zero terminal See ZERO-INPUT TERMINAL and ZERO-OUTPUT TERMINAL.

zero time **1.** In some measurements, the first instant of time, to which all other instants are referred. **2.** See ZERO PHASE.

zero time reference During one cycle of radar operation, the time reference of the schedule of events.

zero vector A vector whose magnitude is zero.

zero voltage See ZERO POTENTIAL.

zero voltage coefficient A voltage coefficient having the value of zero (i.e., one that indicates there is no voltage-dependent drift of a given quantity).

zero-zero The atmospheric condition in which the ceiling and visibility both are zero (i.e., extremely dense fog).

Z gain The gain (or gain control) of the intensity channel of an oscilloscope. Compare X GAIN and Y GAIN.

zig The short, straight deflection of a point or particle, or of a wave along a jagged path in a direction opposite that of ZAG; a component of ZIGZAG DEFLECTION.

zigzag deflection Deflection of a particle, point, or object in a path that contains side-to-side motion, as well as forward motion. Also see ZAG and ZIG.

zigzag rectifier A special version of the three-phase star (three-phase, half-wave) rectifier circuit. Direct-current (dc) saturation of the transformer secondary is avoided by winding half the turns of each secondary on a separate core (i.e., each core carries two half-windings). The opposing flux resulting from different phases in the half-windings causes cancellation of the dc component of flux in each core.

zigzag reflections Multihop reflections of waves along a zigzag path, resulting from repeated reflections within the ionosphere.

zinc Symbol, Zn. A metallic element. Atomic number, 30. Atomic weight, 65.39. It is used as the negative-electrode material in dry cells and as a protective coating for some metals used in electronics.

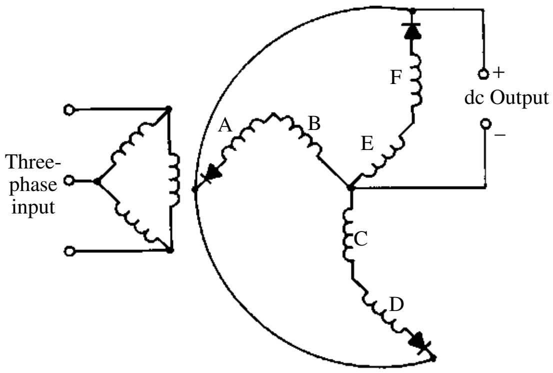

zigzag rectifier

zinc aluminate phosphor Either of two similar substances used as a phosphor coating for cathode-ray tube screens. One form glows blue; the other form glows red.

zinc beryllium silicate phosphor A substance used as a phosphor coating for cathode-ray-tube screens. It glows yellow.

zinc beryllium zirconium silicate phosphor A substance used as a phosphor coating for cathode-ray-tube screens. It glows white.

zinc borate phosphor A substance used as a phosphor coating for cathode-ray-tube screens. It glows yellow-orange.

zinc cadmium sulfide phosphor Either of two similar substances used as a phosphor coating for cathode-ray-tube screens. One form glows blue; the other form glows red.

zinc–carbon cell A common non-rechargeable electrochemical cell. Produces approximately 1.5 volts under no-load conditions. Zinc forms the outer case or shell, and is the negative electrode. A carbon rod serves as the positive electrode. The electrolyte is a paste of manganese dioxide and carbon. The cell is inexpensive, and is commercially available in various sizes. The shelf life is fairly long. Cells of this type work well at moderate temperatures, and in applications where the current drain is moderate to high. They function poorly at low temperatures. Compare ALKALINE CELL.

zinc germanate phosphor A substance used as a phosphor coating for cathode-ray-tube screens. It glows yellow-green.

zinc magnesium fluoride phosphor A substance used as a phosphor coating for cathode-ray-tube screens. It glows orange.

zincolysis Electrolysis in a cell having a zinc anode.

zinc orthoscilicate phosphor Also called *Willemite.* A substance used as a phosphor coating for cathode-ray-tube screens. It glows yellow-green.

zinc oxide A substance used as a phosphor coating for cathode-ray-tube screens. It glows blue-green. It is also used in the manufacture of certain electronic components, such as resistors.

zinc-oxide resistor A voltage-dependent resistor whose active ingredient is zinc oxide.

zinc silicate phosphor A substance used as a phosphor coating for cathode-ray-tube screens. It glows blue.

zinc standard cell A standard cell using zinc and mercury electrodes, and a mercurous sulfate excitant and depolarizer. Produces 1.4322 volts at 15°C. Also called *Clark cell.* Compare WESTON CELL.

zinc sulfide phosphor A substance used as a phosphor coating for cathode-ray-tube screens. Glows blue-green or yellow-green.

ZIP Abbreviation of *zinc-impurity photodetector.*

zip cord A simple two-conductor, flexible power cord.

zirconia Preparations of zirconium (especially $Zr0_2$), valued for their high-temperature dielectric properties.

zirconium Symbol, Zr. A metallic element. Atomic number, 40. Atomic weight, 91.22.

Z marker A vertically radiating marker beacon defining the zone above a radio-range station.

Z meter A device for measuring impedances. Instruments of this kind take four principal forms: **(1)** a direct-reading meter resembling an ohmmeter; **(2)** an adjustable circuit manipulated somewhat like a bridge and that compares an unknown impedance with a standard resistance;

(3) an impedance bridge for evaluating the reactive and resistive components of an unknown impedance; **(4)** a section of transmission line used with a signal source and voltmeter for measuring impedance in terms of a standard resistor and/or standing waves.

ZMODEM In data communications, an error-correction mode similar to XMODEM, except that when an error is found during transmission of a block of data, the source retransmits only that portion of the block following the error. This improves data transmission speed because, when an error occurs, the number of bytes retransmitted is generally fewer than the 128K block size standard in XMODEM. Compare XMODEM and YMODEM.

Zn **1.** Symbol for ZINC. **2.** Symbol for AZIMUTH.

Z_0 Symbol for CHARACTERISTIC IMPEDANCE.

zone **1.** On a magnetic disk, a group of tracks whose associated transfer rate is not the same as that for the rest of the disk. **2.** In computer operations, the area of a storage medium containing digits. **3.** In a computer system, a main memory area set aside for a particular purpose. **4.** In a security system, a specified area or region under surveillance.

zone blanking In a radar system, a method of extinguishing the cathode-ray-tube beam during a portion of the antenna sweep.

zone candle power In a given zone, the luminous flux per steradian, emitted by a light source under test.

zoned circuit In a security system, a circuit in which some areas are protected at all times, while the protection in other areas can be temporarily disabled for entry or exit.

zone leveling See ZONE REFINING.

zone marker See Z MARKER.

zone melting See ZONE REFINING.

zone of silence The region between alternate reflections of a radio wave, in which no signal is detectable. Also called *skip zone.*

zone plate antenna A rapid-scan radar antenna having a reflector composed of confocal parabolas arranged in a circle.

zone-position indicator A radar that reveals the position of an object to a second radar having a restricted field.

zone purification See ZONE REFINING.

zone refining A method of purifying semiconductor materials (such as germanium and silicon) in which a molten zone in an ingot of the material moves along the length of the ingot, dissolving impurities as it travels, eventually depositing them at the end of the ingot, which is sawed off. This concentrated and segregated melting is accomplished by means of radio-frequency heating.

zone time In a given time zone, standard time in terms of the number of hours that must be added to local time to equal the time at the zero (Greenwich) meridian.

zoning **1.** A method of fabricating a microwave reflector in concentric, flat regions, producing the same practical results as a continuous paraboloid. **2.** In a communications system, the division of the coverage area into different geographic regions for a specific purpose.

zoom **1.** To rapidly change the focal length of a television or motion-picture camera lens so that the object seems to advance toward or recede from the viewer, remaining in focus as it does so. **2.** See ZOOM LENS. **3.** To magnify the image in a computer graphical user interface. A user can enlarge a specific portion of the display, at the expense of other portions. It is generally measured in percent (e.g., 200% zoom represents a magnification factor of 2).

zoom lens A continuously adjustable lens system that allows zoom effects to be obtained with a television or motion-picture camera, or a similar arrangement for still cameras that obviates the need for lens interchange when different focal lengths are needed. The lens system, which can be operated electronically, allows either narrow- or wide-angle views to be obtained without losing focus at any time.

Z parameters Device or network parameters expressed as impedances.

ZPI Abbreviation of ZONE-POSITION INDICATOR.

Z-plunger In a waveguide, a combination choke and bucket plunger for radiation leakage reduction.

Zr Symbol for ZIRCONIUM.

Z signals A collection of letter groups, each starting with the letter Z, used for simplified telegraphy and radiotelegraphy by the military services.

Zulu Phonetic alphabet communications code word for the letter Z.

Zulu time Greenwich mean time. Also see ZEBRA TIME.

zwitterion An ion that carries both a positive and a negative charge.

This page intentionally left blank.

Appendix A

Schematic symbols

Ammeter	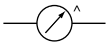
Amplifier general	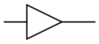
Amplifier, inverting	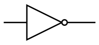
Amplifier, operational	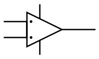
AND gate	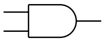
Antenna, balanced	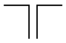
Antenna, general	
Antenna, loop	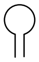
Antenna, loop, multiturn	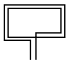
Battery	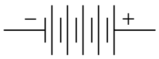
Capacitor, feedthrough	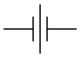
Capacitor, fixed	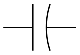
Capacitor, variable	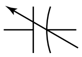
Capacitor, variable, split-rotor	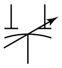
Capacitor, variable, split-stator	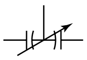

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

Cathode, electron-tube, cold

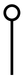

Cathode, electron-tube, directly heated

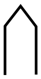

Cathode, electron-tube indirectly heated

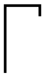

Cavity resonator

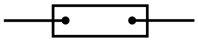

Cell, electrochemical

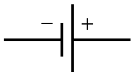

Circuit breaker

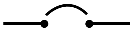

Coaxial cable

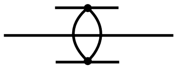

Crystal, piezoelectric

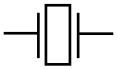

Delay line

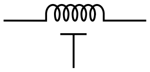

Diac

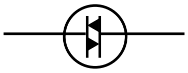

Diode, field-effect

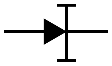

Diode, general

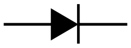

Diode, Gunn

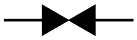

Diode, light-emitting

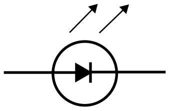

Diode, photosensitive

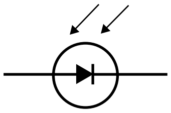

Diode, PIN	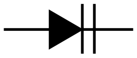
Diode, Schottky	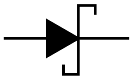
Diode, tunnel	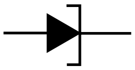
Diode, varactor	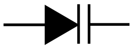
Diode, zener	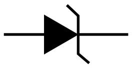
Directional coupler	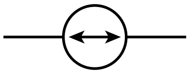
Directional wattmeter	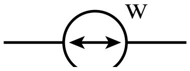
Exclusive-OR gate	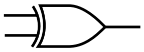
Female contact, general	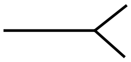
Ferrite bead	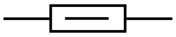
Filament, electron-tube	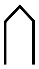
Fuse	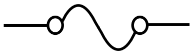
Galvanometer	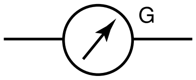
Grid, electron-tube	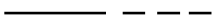
Ground, chassis	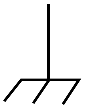

Ground, earth

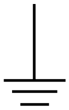

Headset

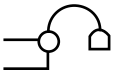

Handset, double

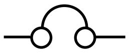

Headset, single

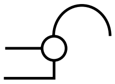

Headset, stereo

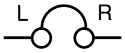

Inductor, air core

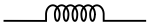

Inductor, air core, bifilar

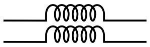

Inductor, air core, tapped

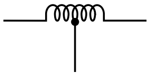

Inductor, air core, variable

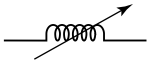

Inductor, iron core

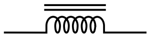

Inductor, iron core, bifilar

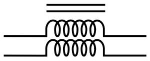

Inductor, iron core, tapped

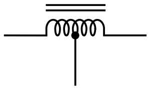

Inductor iron core, variable

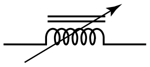

Inductor, powdered-iron core

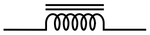

Inductor, powdered-iron core, bifilar

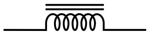

Inductor, powdered-iron core, tapped

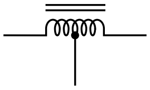

Inductor, powdered-iron core, variable

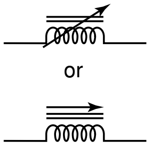

Integrated circuit, general

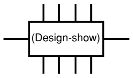

Jack, coaxial or phono

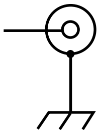

Jack, phone, two-conductor

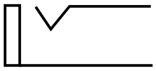

Jack, phone, three-conductor

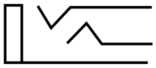

Key, telegraph

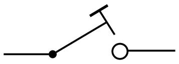

Lamp, incandescent

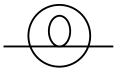

Lamp, neon

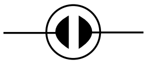

Male contact, general

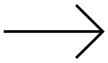

Meter, general

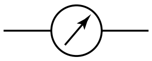

Microammeter

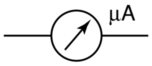

Microphone

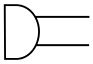

Microphone, directional

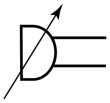

Milliammeter

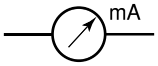

NAND gate

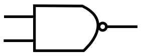

Negative voltage connection

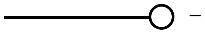

NOR gate

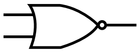

NOT gate

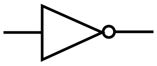

Optoisolator

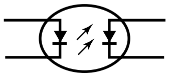

OR gate

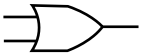

Outlet, two-wire, nonpolarized

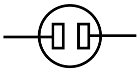

Outlet, two-wire, polarized

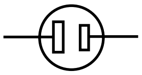

Outlet, three-wire

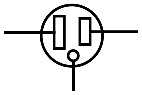

Outlet, 234-V

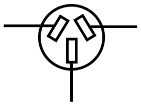

Plate, electron-tube

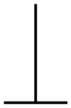

Plug, two-wire, nonpolarized

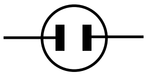

Plug, two-wire, polarized	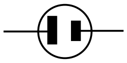
Plug, three-wire	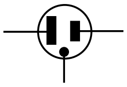
Plug, 234-V	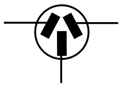
Plug, coaxial or phono	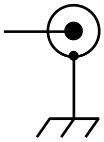
Plug, phone, two-conductor	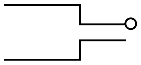
Plug, phone, three-conductor	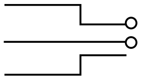
Positive voltage connection	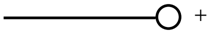
Potentiometer	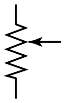
Probe, radio-frequency	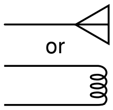
Rectifier, gas-filled	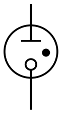
Rectifier, high-vacuum	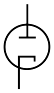
Rectifier, semiconductor	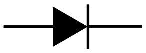
Rectifier, silicon-controlled	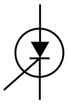

Relay, double-pole, double-throw

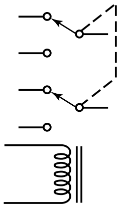

Relay, double-pole, single-throw

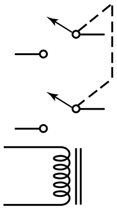

Relay, single-pole, double-throw

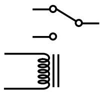

Relay, single-pole, single-throw

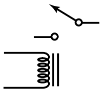

Resistor, fixed

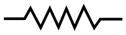

Resistor, preset

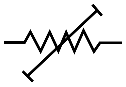

Resistor, tapped

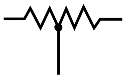

Resonator

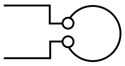

Rheostat

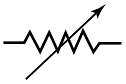

Saturable reactor

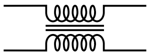

Signal generator

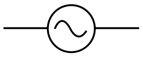

Solar battery

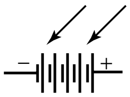

Solar cell

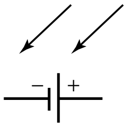

Source, constant-current

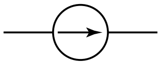

Source, constant-voltage

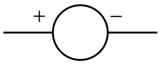

Speaker

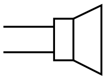

Switch, double-pole, double-throw

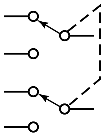

Switch, double-pole, rotary

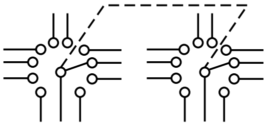

Switch, double-pole, single-throw

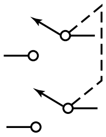

Switch, momentary-contact

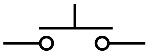

Switch, silicon-controlled

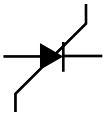

Switch, single-pole, rotary

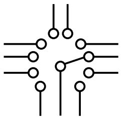

Switch, single-pole, double-throw

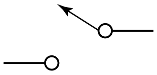

Switch, single-pole, single-throw

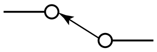

Terminals, general, balanced

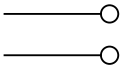

Terminals, general, unbalanced

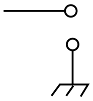

Test point

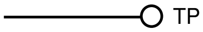

Thermocouple

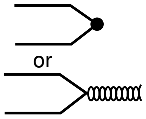

Transformer, air core

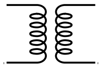

Transformer, air core, step-down

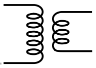

Transformer, air core, step-up

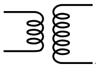

Transformer, air core, tapped primary

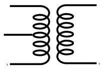

Transformer, air core, tapped secondary

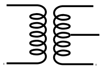

Transformer, iron core

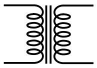

Transformer, iron core, step-down

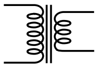

Transformer, iron core, step-up

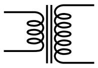

Transformer, iron core, tapped primary

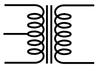

Transformer, iron core, tapped secondary

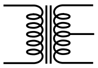

Transformer, powdered-iron core

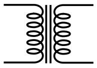

Transformer, powdered-iron core, step-down

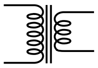

Transformer, powdered-iron core, step-up

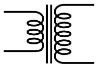

Transformer, powdered-iron core, tapped primary

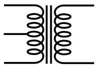

Transformer, powdered-iron core, tapped secondary

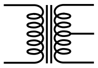

Transistor, bipolar, *NPN*

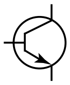

Transistor, bipolar, *PNP*

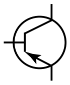

Transistor, field-effect, *N*-channel

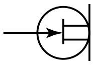

Transistor, field-effect, *P*-channel

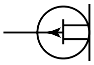

Transistor, MOS field-effect, *N*-channel

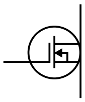

Transistor, MOS field-effect, *P*-channel

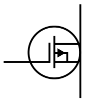

Transistor, photosensitive, *NPN*

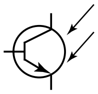

Transistor, photosensitive, *PNP*

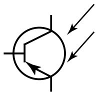

Transistor, photosensitive, field-effect, *N*-channel

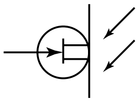

Transistor, photosensitive, field-effect, *P*-channel

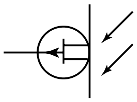

Transistor, unijunction

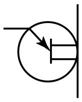

Triac

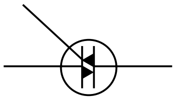

Tube, diode

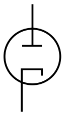

Tube, heptode

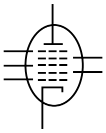

Tube, hexode

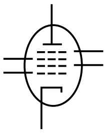

Tube, pentode

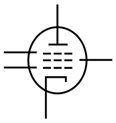

Tube, photosensitive

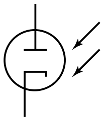

Tube, tetrode

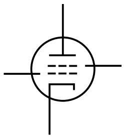

Tube, triode

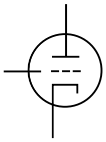

Voltmeter

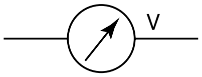

Wattmeter

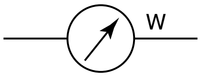

Waveguide, circular

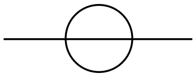

Waveguide, flexible

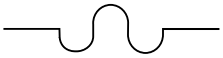

Waveguide, rectangular

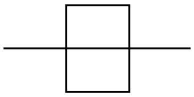

Waveguide, twisted

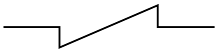

Wires, crossing, connected

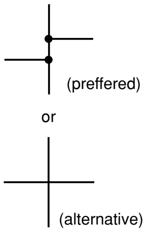
(preffered)
or
(alternative)

Wires, crossing, not connected

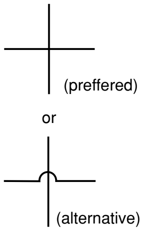
(preffered)
or
(alternative)

Appendix B
Tables and data

Conversion between Electrical Systems

		System	
Property	**mks**	**cgs electromagnetic**	**cgs electrostatic**
Capacitance	1 farad	10^{-9} abfarad	9×10^{11} statfarad
	10^{9} F	1 abF	9×10^{20} statF
	$10^{-11}/9$ F	$10^{-20}/9$ abF	1 statF
Charge	1 coulomb	0.1 abcoulomb	3×10^{9} statC
	10 C	1 abC	3×10^{10} statC
	$10^{-9}/3$ C	$10^{-10}/3$aC	1 statC
Charge density	1 coulomb/m^{3}	10^{-7} abcoulomb/cm^{3}	3×10^{3} statcoulomb/cm^{3}
	10^{7} C/m^{3}	1 abC/cm^{3}	3×10^{10} statCcm^{3}
	$10^{-3}/3$ C/m^{3}	$10^{-10}/3$ aC/cm^{3}	1 statC/cm^{3}
Conductivity	1 siemens/m	10^{-11} absiemens/cm	9×10^{9} statSiemens/cm
	10^{11} S/m	1 abS/cm	9×10^{20} statS/cm
	$10^{-9}/9$ S/m	$10^{-20}/9$ abS/cm	1 statS/cm
Current	1 ampere	10^{-1} abampere	3×10^{9} statampere
	10 a	1 abA	3×10^{10} statA
	$10^{-9}/3$ a	$10^{-10}/3$ abA	1 statA
Current density	1 ampere/m^{2}	10^{-5} abampere/cm^{2}	3×10^{5} statampere/cm^{2}
	10^{5} A/m^{2}	1 abA/cm^{2}	3×10^{10} statA/cm^{2}
	$10^{-5}/3$ A/m^{2}	$10^{-10}/3$ aA/cm^{2}	1 statA/cm^{2}
Electric field intensity	1 volt/m	10^{6} abvolt/cm	$10^{-4}/3$ statvolt/cm
	10^{-6} V/m	1 abV/cm	$10^{-10}/3$ statV/cm
	3×10^{4} V/m	3×10^{10} aV/cm	1 statV/cm
Electric potential	1 volt	10^{8} abvolt	$10^{-2}/3$ statvolt
	10^{-8} V	1 abV	$10^{-10}/3$ statV
	3×10^{2} V	3×10^{10} aV	1 statV
Electric dipole moment	1 coulomb-m	10 abC-cm	3×10^{11} statC-cm
	0.1 C-m	1 abC-cm	3×10^{10} statC-cm
	$10^{-11}/3$ C-m	$10^{-10}/3$ abC-cm	1 statC-cm

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

	System		
Property	**mks**	**cgs electromagnetic**	**cgs electrostatic**
Energy	1 joule	10^7 erg	10^7 erg
	10^{-7} J	1 e	1 e
	10^{-7} J	1 e	1 e
Force	1 newton	10^5 dyne	10^5 dyne
	10^{-5} N	1 d	1 d
	10^{-5} N	1 d	1 d
Flux density	1 Weber/m^2	10^4 gauss (or abtesla)	$10^{-6}/3$ electrostatic unit
	10^{-4} Wb/m^2	1 G	$10^{-10}/3$ esu
	3×10^6 Wb/m^2	3×10^{10} G	1 esu
Inductance	1 henry	10^9 abhenry	$10^{-11}/9$ stathenry
	10^{-9} H	1 abH	$10^{-20}/9$ statH
	9×10^{11} H	9×10^{20} abH	1 statH
Inductive capacity	1 farad/m	10^{-11} abfarad/cm	9×10^9 statfarad/cm
	10^{11} F/m	1 abF/cm	9×10^{20} statF/cm
	$10^{-9}/9$ F/m	$10^{-20}/9$ abF/cm	1 statF/cm
Magnetic flux	1 weber	10^8 Maxwell	$10^{-2}/3$ electrostatic unit
	10^{-8} W	1 Mx	$10^{-10}/3$ esu
	3×10^2 W	3×10^{10} Mx	1 esu
Magnetic dipole moment	1 ampere-m^2	10^3 abampere-cm^2	3×10^{13} statampere-cm^2
	10^{-3} A-m^2	1 abA-cm^2	3×10^{10} statA-cm^2
	$10^{-13}/3$ A-m^2	$10^{-10}/3$ abA-cm^2	1 statA-cm^2
Permeability	1 henry/m	10^7 abhenry/cm	$10^{-13}/9$ stathenry/cm
	10^{-7} H/m	1 abH/cm	$10^{-20}/9$ statH/cm
	9×10^{13} H/m	9×10^{20} abH/cm	1 statH/cm
Power	1 watt	10^7 erg/s	10^7 erg/s
	10^{-7} W	1 e/s	1e/s
	10^{-7} W	1 e/s	1e/s
Resistance	1 ohm	10^9 abohm	$10^{-11}/9$ statohm
	10^{-9} ohm	1 abohm	$10^{-20}/9$ statohm
	9×10^{11} ohm	9×10^{20} abohm	1 statohm

GREEK ALPHABET

Capital	Lowercase	Name
Α	α	alpha
Β	β	beta
Γ	γ	gamma
Δ	δ or ∂	delta
Ε	ε	epsilon
Ζ	ζ	zeta
Η	η	eta
Θ	θ or ϑ	theta
Ι	ι	iota
Κ	κ	kappa
Λ	λ	lambda
Μ	μ	mu
Ν	ν	nu
Ξ	ξ	xi
Ο	ο	omicron
Π	π	pi
Ρ	ρ	rho
Σ	σ or ζ	sigma
Τ	τ	tau
ϒ	υ	upsilon
Φ	ϕ or φ	phi
Χ	χ	chi
Ψ	ψ	psi
Ω	ω	omega

Mathematical Functions

Signs and symbols

·	radix (base) point	≠	not equal to
•	multiplication symbol; logic AND function	~	similar to
∞	infinity	<	less than
+	plus; positive; logic OR function	≮	not less than
−	minus; negative	<<	much less than
±	plus or minus; positive or negative	>	greater than
∓	minus or plus; negative or positive	≯	not greater than
×	times	>>	much greater than
÷	divided by	≤	equal to or less than
/	divided by (expressive of a ratio)	≥	equal to or greater than
=	equal to	∝	proportional to; varies directly as
≡	identical to; is defined by	≐ or →	approaches
≃ or ≅	approximately equal to, congruent to	:	is to; proportional to
≈	approximately equal to	∴	therefore

@	at the rate of; at cost of
e	natural number = 2.71828
π	pi ≈ 3.14159 . . .
( )	parentheses (use to enclose a group of terms)
[ ]	brackets (use to enclose a group of terms that includes one or more groups in parentheses)
{ }	braces (use to enclose a group of terms that includes one or more groups in brackets.
$\angle$	angle
$^\circ$	degrees (arc or temperature)
$'$	minutes; prime
$''$	seconds; double prime
$\parallel$	parallel to
$\perp$	perpendicular to
. . .	and beyond
∇	(del or nabla) vector differential operator

Operations

$x + y$	x added to y; x OR y
$x - y$	y subtracted from x
$x \cdot y$, $x \times y$, or xy	multiplied by y; x AND y
$x \div y$	x divided by y
x/y or $\frac{x}{y}$	x divided by y
$1/x$	reciprocal of x
x^n	x raised to the indicated power of n
$\sqrt[n]{x}$	indicated root n of x
$x : y$	x is to y
$\|x\|$	absolute value of x, magnitude of x
$\bar{X}$, $\dot{X}$, or X	vector X

$\bar{x}$	average value of x; x not
$f(x)$ or $F(x)$	function of x
i	$\sqrt{-1}$
j	operator, equal to $\sqrt{-1}$
Δx	increment of x
dx	differential of x
∂x	partial differential of x
$\frac{\Delta x}{\Delta y}$	change in x with respect to y
$\frac{dx}{dy}$	derivative of x with respect to y
$\frac{d}{dy}(x)$	derivative of x with respect to y
$d_y x$	derivative of x with respect to y
$\frac{\partial x}{\partial y}$	partial derivative of x with respect to y
$\sum$	summation
$\sum_{b}^{a}$	summation between limits (from a to b)
$\prod$	product
$\prod_{b}^{a}$	product between limits (from a to b)
$\int$	integral
$\int_a^b$	integral between limits (from a to b)
$\int x\,dy$	integral of x with respect to y
$\|a$	evaluated at a
$\|_a^b$	evaluated between limits (from a to b)
$n!$	factorial of n

Prefixes

Prefix	Quantity	Symbol
atto-	10^{-18}	a
femto-	10^{-15}	f
pico-	10^{-12}	p
nano-	10^{-9}	n
micro-	10^{-6}	μ
milli-	10^{-3}	m
centi-	10^{-2}	c
deci-	10^{-1}	d
deka-	10	da
hecto-	10^{2}	h
kilo-	10^{3}	k
mega-	10^{6}	M
giga-	10^{9}	G
tera-	10^{12}	T
peta-	10^{15}	P
exa-	10^{18}	E

RESISTOR COLOR CODE

The first three color bands supply the total resistance. The fourth color (if any) gives the tolerance. Example: A 5,600-ohm resistor would have green (first color, 5), blue (second color, 6) and red

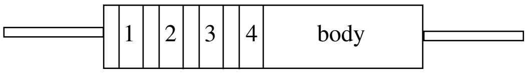

1

black = 0
brown = 1
red = 2
orange = 3
yellow = 4
green = 5
blue = 6
violet = 7
gray = 8
white = 9

2

black = 0
brown = 1
red = 2
orange = 3
yellow = 4
green = 5
blue = 6
violet = 7
gray = 8
white = 9

3

black =
brown = 0
red = 00
orange = 000
yellow = 0000
green = 00000
blue = 000000
gold = multiply by 0.1
silver = multiply by 0.01

4

gold ± 5%
silver ± 10%
no band ± 20%

This page intentionally left blank.

About the author

Stan Gibilisco has authored or coauthored dozens of nonfiction books about electronics and science. He first attracted attention with *Understanding Einstein's Theories of Relativity* (TAB Books, 1983). His *Encyclopedia of Electronics* (TAB Professional and Reference Books, 1985) and *Encyclopedia of Personal Computing* (McGraw-Hill, 1996) were annotated by the American Library Association as among the best reference volumes published in those years. Stan's work has gained reading audiences in the Far East, Europe, and South America.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

This page intentionally left blank.

SOFTWARE AND INFORMATION LICENSE

The software and information on this diskette (collectively referred to as the "Product") are the property of The McGraw-Hill Companies, Inc. ("McGraw-Hill") and are protected by both United States copyright law and international copyright treaty provision. You must treat this Product just like a book, except that you may copy it into a computer to be used and you may make archival copies of the Products for the sole purpose of backing up our software and protecting your investment from loss.

By saying "just like a book," McGraw-Hill means, for example, that the Product may be used by any number of people and may be freely moved from one computer location to another, so long as there is no possibility of the Product (or any part of the Product) being used at one location or on one computer while it is being used at another. Just as a book cannot be read by two different people in two different places at the same time, neither can the Product be used by two different people in two different places at the same time (unless, of course, McGraw-Hill's rights are being violated).

McGraw-Hill reserves the right to alter or modify the contents of the Product at any time.

This agreement is effective until terminated. The Agreement will terminate automatically without notice if you fail to comply with any provisions of this Agreement. In the event of termination by reason of your breach, you will destroy or erase all copies of the Product installed on any computer system or made for backup purposes and shall expunge the Product from your data storage facilities.

LIMITED WARRANTY

McGraw-Hill warrants the physical diskette(s) enclosed herein to be free of defects in materials and workmanship for a period of sixty days from the purchase date. If McGraw-Hill receives written notification within the warranty period of defects in materials or workmanship, and such notification is determined by McGraw-Hill to be correct, McGraw-Hill will replace the defective diskette(s). Send request to:

Customer Service
McGraw-Hill
Gahanna Industrial Park
860 Taylor Station Road
Blacklick, OH 43004-9615

The entire and exclusive liability and remedy for breach of this Limited Warranty shall be limited to replacement of defective diskette(s) and shall not include or extend to any claim for or right to cover any other damages, including but not limited to, loss of profit, data, or use of the software, or special, incidental, or consequential damages or other similar claims, even if McGraw-Hill has been specifically advised as to the possibility of such damages. In no event will McGraw-Hill's liability for any damages to you or any other person ever exceed the lower of suggested list price or actual price paid for the license to use the Product, regardless of any form of the claim.

THE McGRAW-HILL COMPANIES, INC. SPECIFICALLY DISCLAIMS ALL OTHER WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO, ANY IMPLIED WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Specifically, McGraw-Hill makes no representation or warranty that the Product is fit for any particular purpose and any implied warranty of merchantability is limited to the sixty day duration of the Limited Warranty covering the physical diskette(s) only (and not the software or in-formation) and is otherwise expressly and specifically disclaimed.

This Limited Warranty gives you specific legal rights; you may have others which may vary from state to state. Some states do not allow the exclusion of incidental or consequential damages, or the limitation on how long an implied warranty lasts, so some of the above may not apply to you.

This Agreement constitutes the entire agreement between the parties relating to use of the Product. The terms of any purchase order shall have no effect on the terms of this Agreement. Failure of McGraw-Hill to insist at any time on strict compliance with this Agreement shall not constitute a waiver of any rights under this Agreement. This Agreement shall be construed and governed in accordance with the laws of New York. If any provision of this Agreement is held to be contrary to law, that provision will be enforced to the maximum extent permissible and the remaining provisions will remain in force and effect.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

Suggested additional references

Crowhurst, N. and Gibilisco, S., *Mastering Technical Mathematics—2nd Edition.* New York: McGraw-Hill, 1999.

Dorf, R., *Electrical Engineering Handbook—2nd Edition.* Boca Raton, FL: CRC Press, 1997.

Gibilisco, S., *Electronics Portable Handbook.* New York: McGraw-Hill, 1999.

Gibilisco, S., *Electronics Formulas Pocket Reference.* New York: McGraw-Hill, 1999

Gibilisco, S., *Mathematical and Physical Data, Equations, and Rules of Thumb.* New York: McGraw-Hill, 2001.

Van Valkenburg, M., *Reference Data for Engineers: Radio, Electronics, Computer and Communications.* Indianapolis: Howard W. Sams & Co., 1998.

Veley, V., *The Benchtop Electronics Reference Manual.* New York: McGraw-Hill, 1994.

Copyright 2001 The McGraw-Hill Companies, Inc. Click Here for Terms of Use

Amateur Radio MØMTJ

MDS975.co.uk HOME PAGE

My Home Page

Amateur Radio Operating Conditions
Antennas
Accessories
Projects & Kits
Data Modes
Information
Portable
QSL
RSGB Ltd
Links
QRM
Bits 'n' Bobs
Contact

WSPR
Weak Signal Propagation Reporter

PORTABLE

Site Map

MDS975 Home

Feedback

PLT QRM

Site Search

Find
powered by FreeFind

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits | Useful Information
Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites | RSGB | QSL | The Amateur Radio Mini Site Map
News, Developments, Events, Photographs and other 'Bits 'n' Bobs' | WSPR Weak Signal Propagation Reporter

AERIALS (ANTENNAS) 5

Antennas 1 | **Antennas 2** | **Antennas 3** | **Antennas 4** | **Antennas 6** | **Antennas 7**

Index To Other Antenna Pages:
Antennas 1 : Aerials used by M0MTJ
Antennas 2 : Including ideas for compact antennas for Top Band /160 metres
Antennas 3 : Felix Scerri VK4FUQ discusses Loop Antennas, baluns, masts & other antenna related topics
Antennas 4 : Many antenna ideas from various sources particularly for multi-band operation & also gives information about
antenna trimming, knots for wire antennas and useful antenna rigging accessory ideas.
Antennas 5 : Half Wave End Fed antennas for 144 MHz VHF / 430 MHz UHF and 50 MHz 6 Metre band & J-Pole Aerials
Antennas 6 : Simple and effective H.F. Antenna ideas - Ground Plane and All Band Doublet
Antennas 7 : Omni-Directional - Circularly (Mixed) Polarized Antenna for VHF / 2 Meters.

The VK2ZOI "Flowerpot Antenna" - A physically end fed Half Wave "Coaxial Dipole" for 2 metres and 70 centimetres

For 2m and 70cm FM I use a mounted on a lightweight aluminium telescopic pole on the apex of the hose. The base of the antenna (the bottom of its radiating element) is approximately 11 metres above ground level. This antenna is based on the Controlled Feeder Radiation principle (CFR) and is described by VK2ZOI on his website. Also known as a "Coaxial Dipole". My version is described below.

Also seen in the photograph are the ropes that support the H.F. wire aerials.

G Whip Antenna Products

Get on the air with TOP QUALITY BRITISH G-WHIP ANTENNAS

Home brew dual band vertical antenna for 2 metres and 70 cms
(Coaxial Dipole / Controlled Feeder Radiation Antenna / 'Flowerpot Antenna')

VK2ZOI has produced some extremely interesting and potentially very useful dipole antenna designs. The designs could form the basis for a great home-brew antenna project since it is physically end fed and can also be made into a dual band aerial for 2 metres and 70 centimetres, so forming the basis of a viable alternative to buying an expensive commercially manufactured 'white stick' antenna.

The final dual band version works very well and can form the basis of a viable alternative to commercially made 'white stick' antennas, because there's nothing better than using your own home-brew antenna!

Physically, the feeder cable enters the antenna at the bottom end, so it looks like an end fed aerial. VK3TWO / VK6TWO describes it as a "Coaxial Dipole". 'Electrically' it is a simple dipole. The RF is travelling 'inside' the bottom 'element' and doesn't 'feed' the antenna until where the coax is cut - in the centre of the antenna, as shown in the diagram below. Where the outer braid is cut (electrically the centre feed point), the RF then radiates like a simple dipole, via the top radiator (coax core), and via the outside of the coax - the bottom half of the dipole. The top radiator is thinner than the lower radiator (hence why the lower radiator is slightly shorter than the upper radiator).

The coiling of the coax is simply forming an RF choke (high impedance point), to stop the RF continuing down the outside of the braid, thus electrically it 'appears' to be the end of the radiating element.

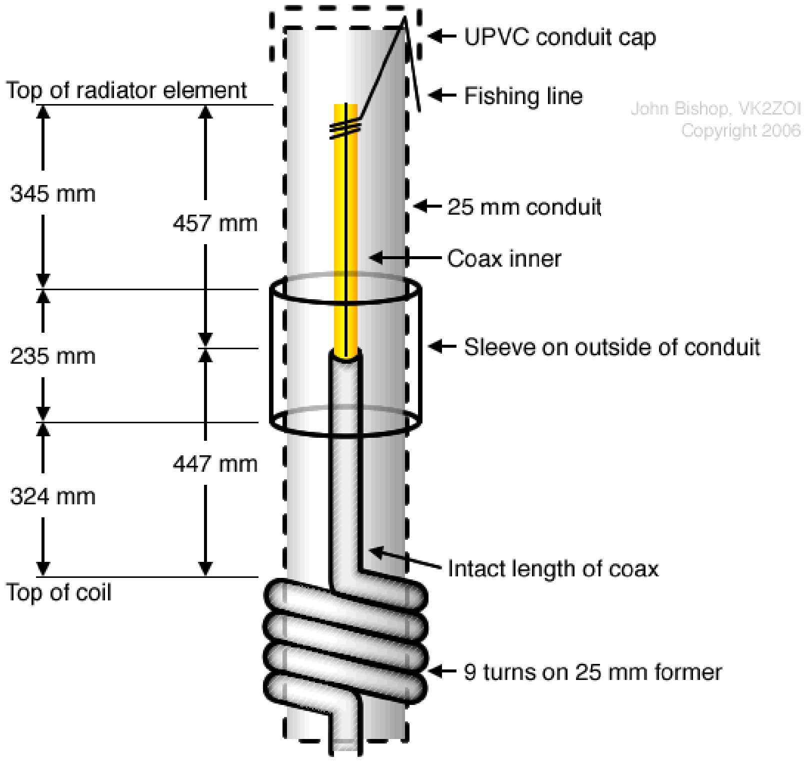

http://vk2zoi.com/articles/half-wave-flower-pot/

I purchased a 3 metre length of 25mm diameter conduit from B&Q, our local DIY centre and ordered some 25mm end caps and heatshrink from ebay. I already had some good quality RG58 for the feeder and main 2 metre radiating element and some aluminium foil for the 70cm sleeve dipole.

First of all I cut the RG58 cable to form the 2m radiating section and choke coil. Because the UK's 2m allocation of 144 to 146 MHz is narrower than the 144 to 148 MHz available in Australia, I varied slightly from the design shown. The centre of the UK's 2m band is about 1% longer in wavelength, so I decided to make both the top and bottom measurements 1% longer.

I therefore stripped 460 mm of the outer sheath and braid from the cable to form the top 1/4 wave element of the dipole. I then measured down 450 mm and marked the point where the lower 1/4 wave element would finish and the choke coil would start.

Next I attached a thin nylon cord to the top of the top radiator, the coax inner.

I then cut the 3 meter length of 25 mm conduit down to about 2.3 metres and drilled a hole where the coil would start, wound 9 turns of RG58 cable from that hole and marked the position of the lower hole. I then removed the coaxial cable and drilled the second, lower, hole.

I then pushed the radiating section of RG58 into the top hole and fed it up towards the top of the tube, stopping when the marker tape reached the hole. I then wound the coil and pushed the remainder of the RG58 through to lower hole and fitted a PL259 plug on the end.

I pulled the top of the radiator wire tight using the nylon cord and pushed the end cap on. The antenna was then ready to be tested on the 2 metre band. I found that the resonant frequency was rather too high, just above 146 MHz, so I pushed an additional 10 mm of coaxial cable into the upper section of the tube - therefore making the lower 1/4 wave section of the dipole 460 mm long - the same as the top section. I tightened up the choke coil winding again and performed another test.

This time the resonant point was just over 145.000 MHz - near enough the centre if the UK's 2 metre band. That was perfect, so the 70 cm sleeve element was then added - this is a 235 mm long tube of kitchen foil positioned exactly at the centre (feed) point of the 2 meter dipole within the tube.

The SWR was tested and found to be acceptable across both the 2m and 70cm bands.

I then fitted the end cap, applied the heatshrink to the coil and to the aluminium foil sleeve. I noticed that when the antenna tube was moved around the cable inside rattled around making a noise that may be rather annoying to anyone near its final location.

To help hold prevent the cable from rattling I pushed up 4 or 5 small pieces of foam material up the tube from the bottom to rest at various positions along its length. With these pieces in place the cable was certainly silenced, however it may have had a deleterious effect on the SWR.

With the antenna now in its finished physical state I naturally checked the SWR again to compare against the performance in its semi-complete state. I was pleased to find that the SWR was still fine across the 2m band, in fact the SWR was a little lower. However the SWR at the edges of the 70cm band was considerably higher - 1.8 at 440 MHz and about 2.2 at 430 MHz.

I conjectured that the heatshrink covering the foil sleeve dipole may have caused the change in response so I removed it, but the SWR was little different and the bandwidth on 70cm was now disappointingly narrower than expected and hoped for.

Although I cannot say for certain, because they cannot now be easily removed, but it may be possible that the pieces of foam may be the culprits for the difference.

While the bandwidth could not be improved, I decided to move the centre point of resonance down a little by increasing the length of the sleeve element from 235 mm to 245 mm. With that adjustment the SWR was now approximately 1.6 at 430 MHz but rising to 2.0 at 440 MHz. (Unfortunately I forgot to note the exact figures down in all cases).

When the antenna was connected to the 20 meter length of Westflex-103 back to the shack, the SWR reading were as follows:

2 Metres - SWR		**70cms - SWR**	
144.00	1.3	430.00	1.2
144.50	1.2	431.00	1.4
145.00	1.2	432.00	1.4
145.50	1.2	433.40	1.0
146.00	1.2	434.00	1.3
		435.00	1.8
		436.00	1.6
		437.00	1.0
		438.00	1.5
		439.00	2.0
		440.00	2.1

The SWR readings in the shack for 2 meters are lower than at the feed point, which is presumably due the losses in the feeder. The SWR readings for 70 cms look rather erratic, with a strange peak at 435 MHz, while the 430 MHz figure is lower than at the feed point of the antenna, and the 439 and 440 MHz figures are disappointingly higher than hoped for. The peculiar readings are likely due to feeder effects.

However the SWR at 433.4, in the FM simplex portion of the band, is very low.

The completed antenna was mounted to the aluminium mast by utilising brackets of the Watson W50 antenna. The brackets had to be reversed so that the narrower diameter of the 25mm tube could be held in place by the V Bolt, while the circular section that previously fitted over the base of the W50 now slid over the mast, which was coincidentally a similar diameter.

I made a small addition to the design in the form of a second small 150mm length of of the 25 mm conduit glued to a coupler section. This is slid into place at the bottom of the antenna to provide additional weather protection to the joint between the W-103 feeder and the RG58 of the antenna - which itself is covered in self amalgamating tape.

Shown in the table below are some signal comparisons with the Watson W-50 antenna; both were mounted on the same mast in the same position and at a height of approximately 7 metres above ground level. Becuase the S-Meter of the transceiver is not calibrated in absolute values, the figres are for relative comparison only - also bear in mind that a typical S Point represents 6dB - therefore the accuracy of these readings will be coarse and might be considered to be +/- 3dB - that's a rather wide variation.

Despite the relative crudeness of these comparisons, the results do seem to indiacte that the VK2ZOI antenna is marginally or slightly better than the W-50 on 2 metres and marginally worse on 70cms. I am quite pleased with this result and beleive that this antenna really could replace the need to buy an expensive commercially manufactured antenna.

My only concern with this type of antenna is that there is no path to ground from the top element, as there would be with a folded dipole or a J-Pole type antenna. This may be a cause for concern as far as static build up is concerned.

Signal comparisons		
	Watson W-50	**"Flowerpot Antenna"**
2 Metres		
Station A	S7	S9
Station B	S5	S6
Station C	S4	S4
Station D	S9	S9
Station E	S6	S6
Station F	S0	S1
Station G	S2	S2
Station H	S2	S3
Station I	S6	S6
70 cms		
Station J	S0	S1
Station K	S5	S5
Station L	S5	S4
Station M	S0	S0
Station N	S9	S9
Station O	S6	S6
Station P	S7	S6

Please see the photographs below for a visual explanation of this project. Mike, MØMTJ. 05/03/2013

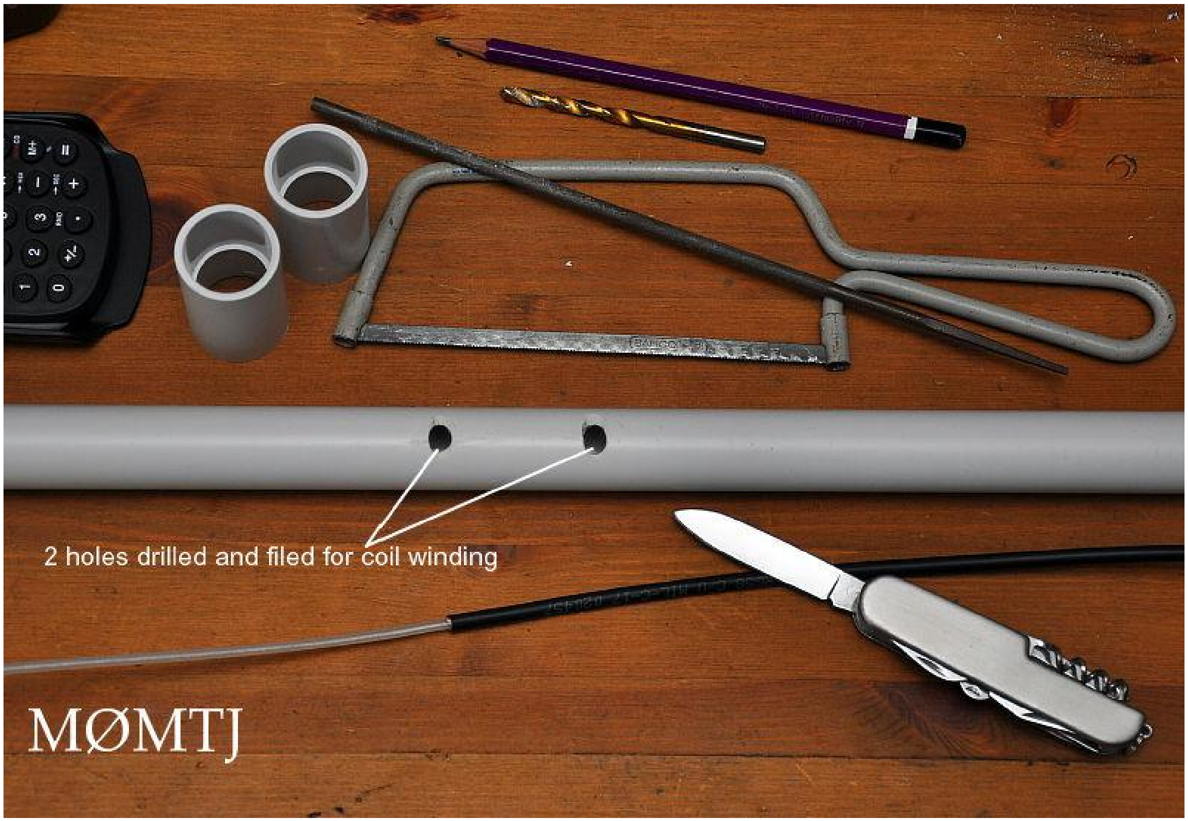

First stage of construction of the 2m / 70cm dual band antenna
Cutting the 25mm diameter conduit to the desired length and drilling
the two holes allowing the coil to be wound. MØMTJ

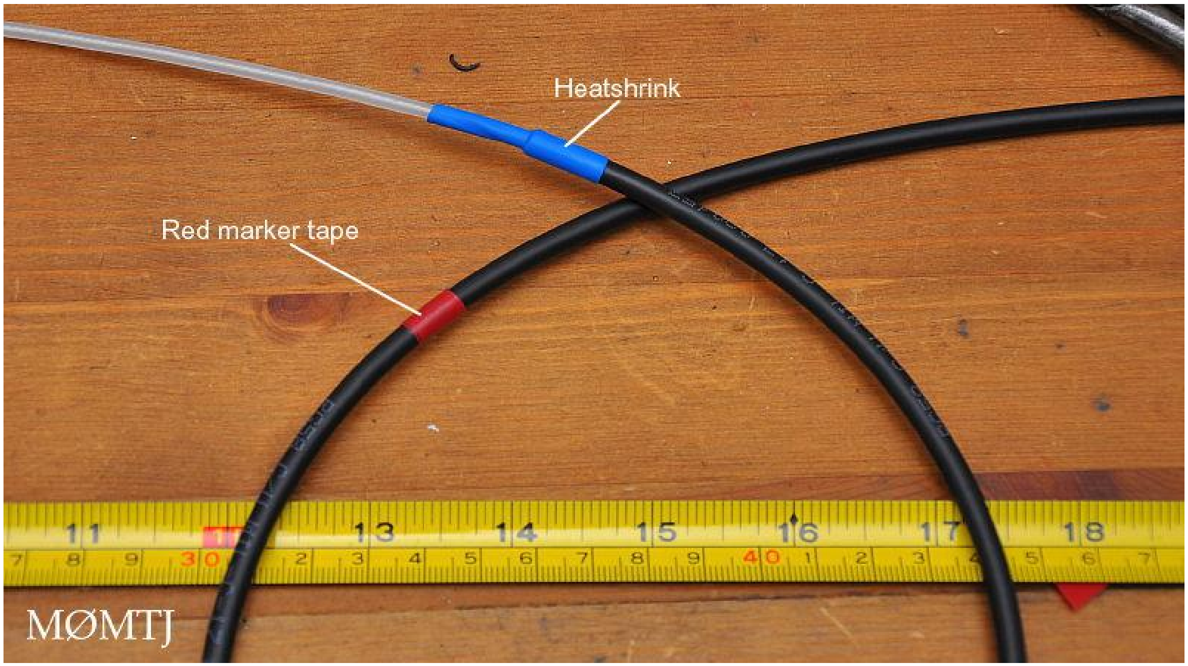
Heatshrink
Red marker tape
MØMTJ

Heatshrink applied to the centre of the dipole section and red insulation tape added to mark the bottom of the dipole where the coil starts. MØMTJ

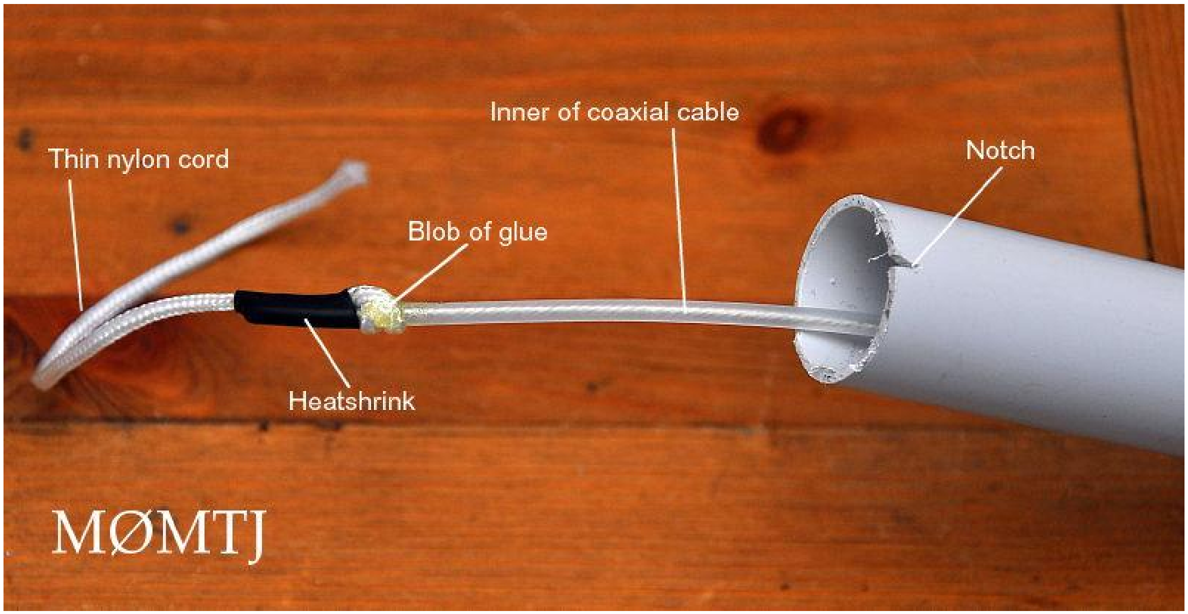
Inner of coaxial cable
Thin nylon cord
Notch
Blob of glue
Heatshrink
MØMTJ

Thin cord attached to top of the inner conductor of the coaxial cable, which forms the top 1/4 wave section of the dipole with a piece of heatshrink and a blob of glue. MØMTJ

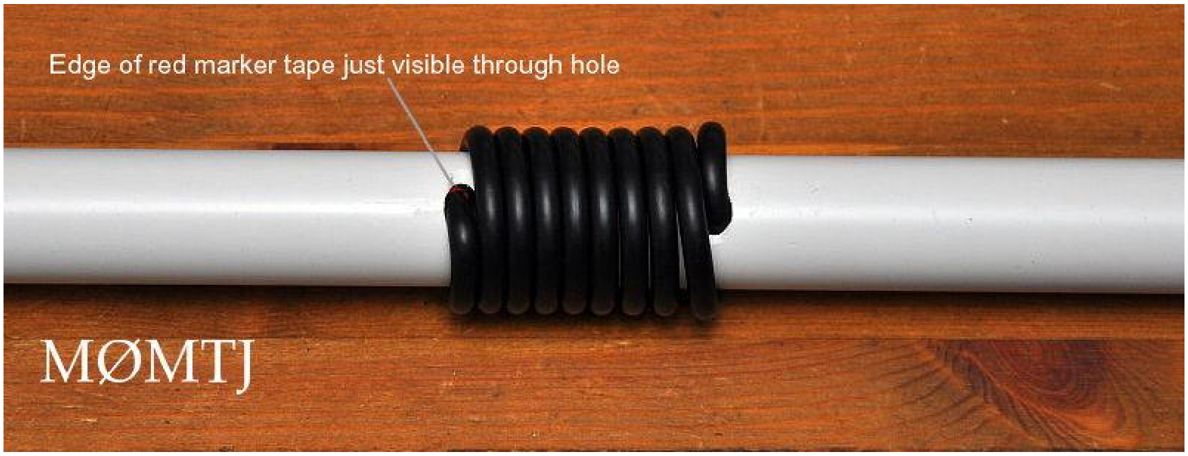
Edge of red marker tape just visible through hole
MØMTJ

The line isolator choke is formed by winding 9 turns of the RG58 coaxial cable around the 25mm diameter conduit tubing. MØMTJ

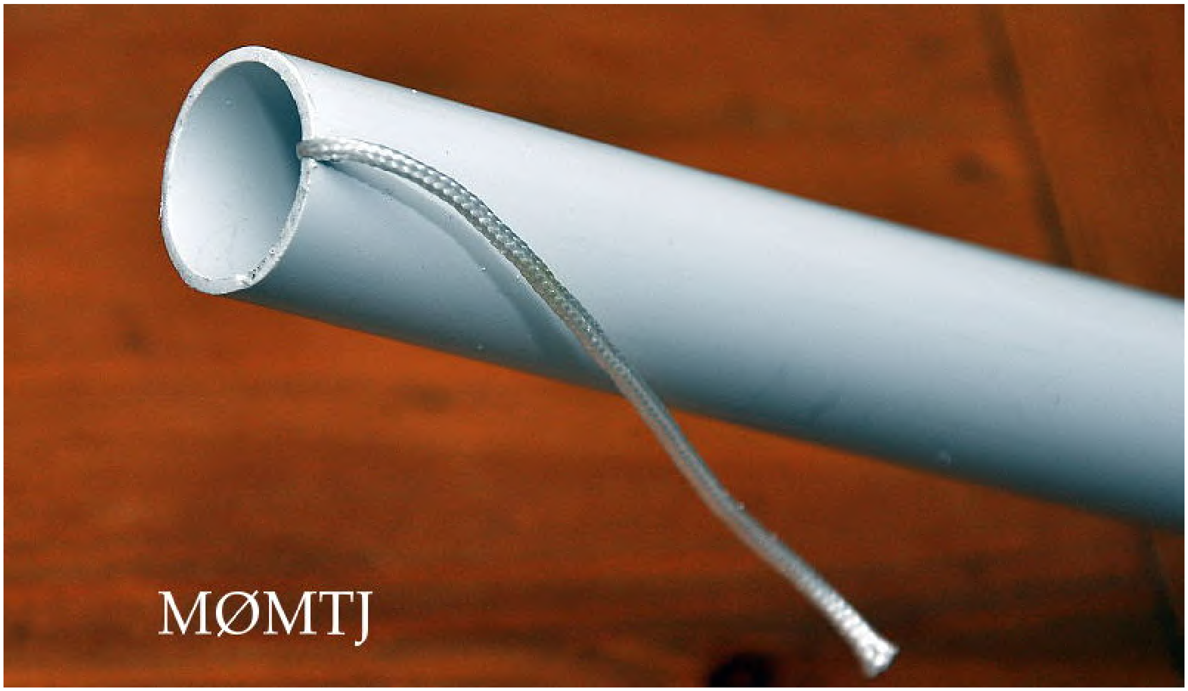

The thin cord that holds the 1/4 wave radiator in place is located in the notch and will be trapped in place when the end cap is fitted. MØMTJ

MØMTJ - www.mds975.co.uk

After testing, the coil was covered in heatshrink to prevent water entering the plastic tube.

MØMTJ - www.mds975.co.uk

The foil sleeve dipole for 70cms is covered in heatshrink.

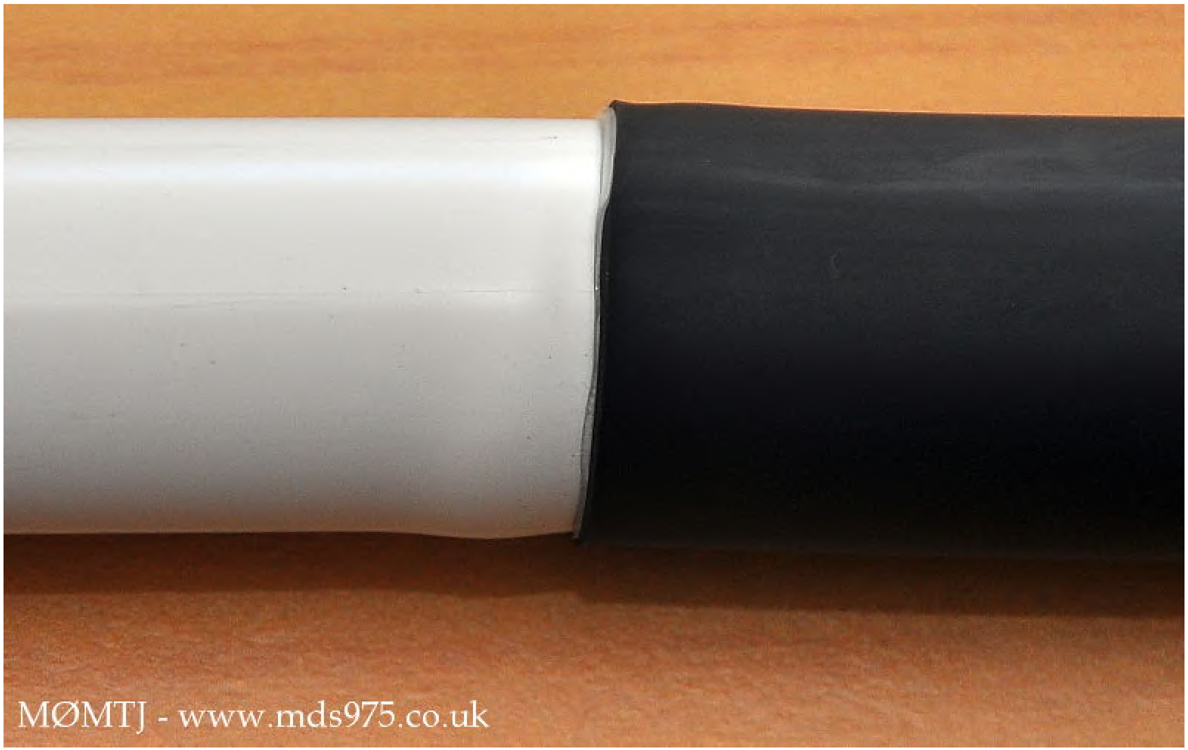

Oops. Lesson learned. When applying heat to the heatshrink I held the tube above the ground and the plastic of the tube started to go soft and bend out of shape. The buckle in the tube can be seen in this photograph.
Lesson: When applying heat, keep the tube flat on the ground or work bench and roll the tube along as the heatshrink shrinks into place ensuring that the tube does not distort or bend.

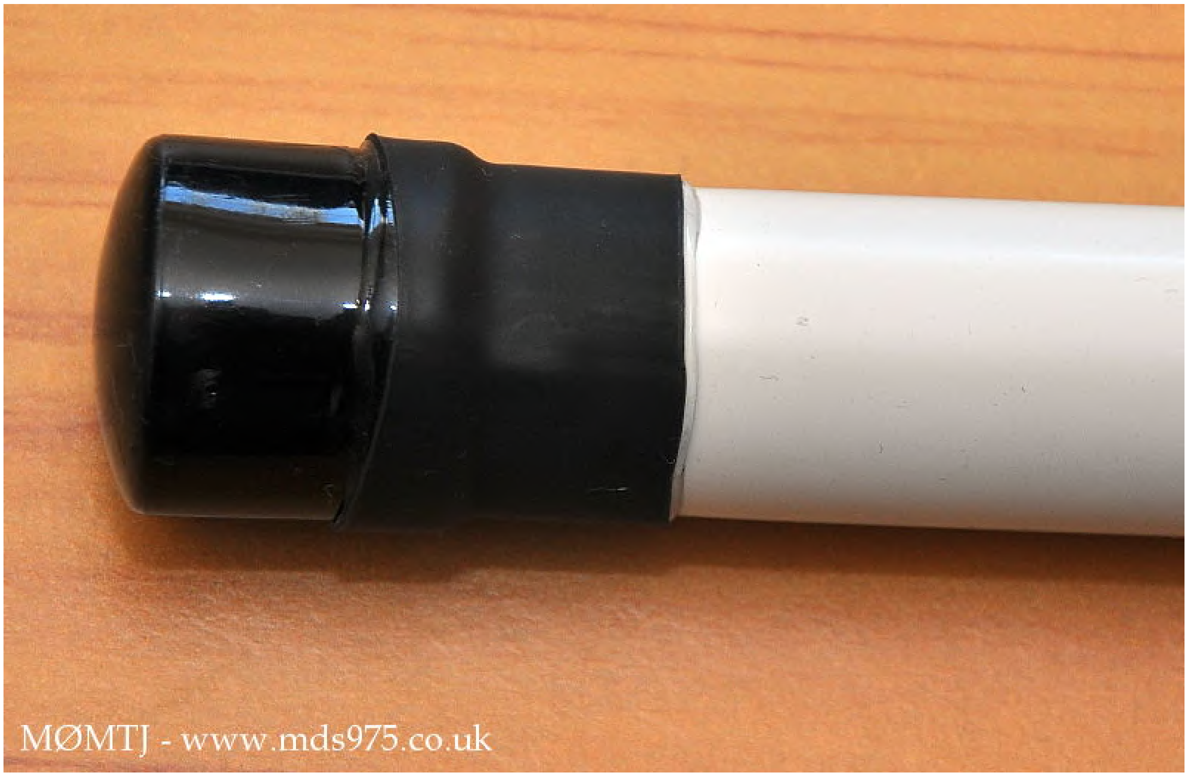

25 mm end cap sealed in place by heatshrink.

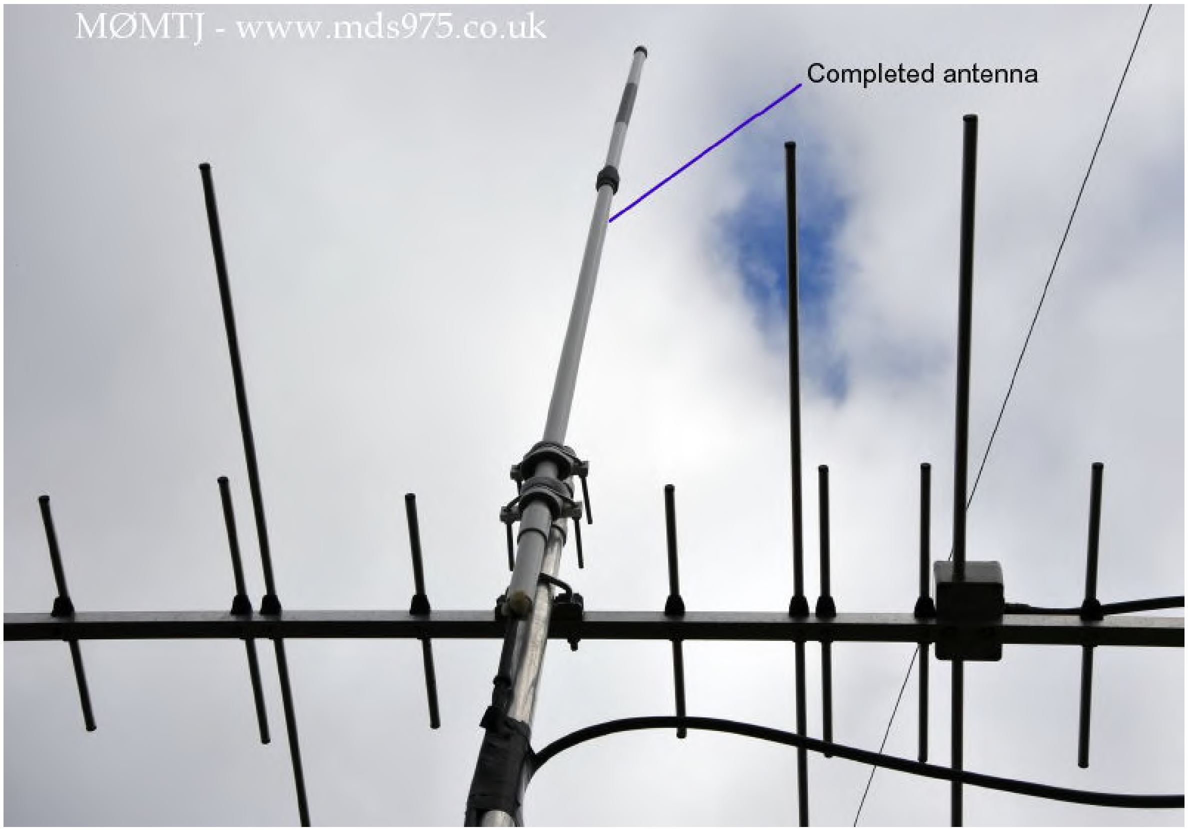
M0MTJ - www.mds975.co.uk
Completed antenna

The completed antenna in place, mounted at the top of my aluminium push-up mast. The fixings used are the brackets from the Watson W-50 antenna which have been reversed so that the smaller diameter PVC tube is held in place by the V Bolts.

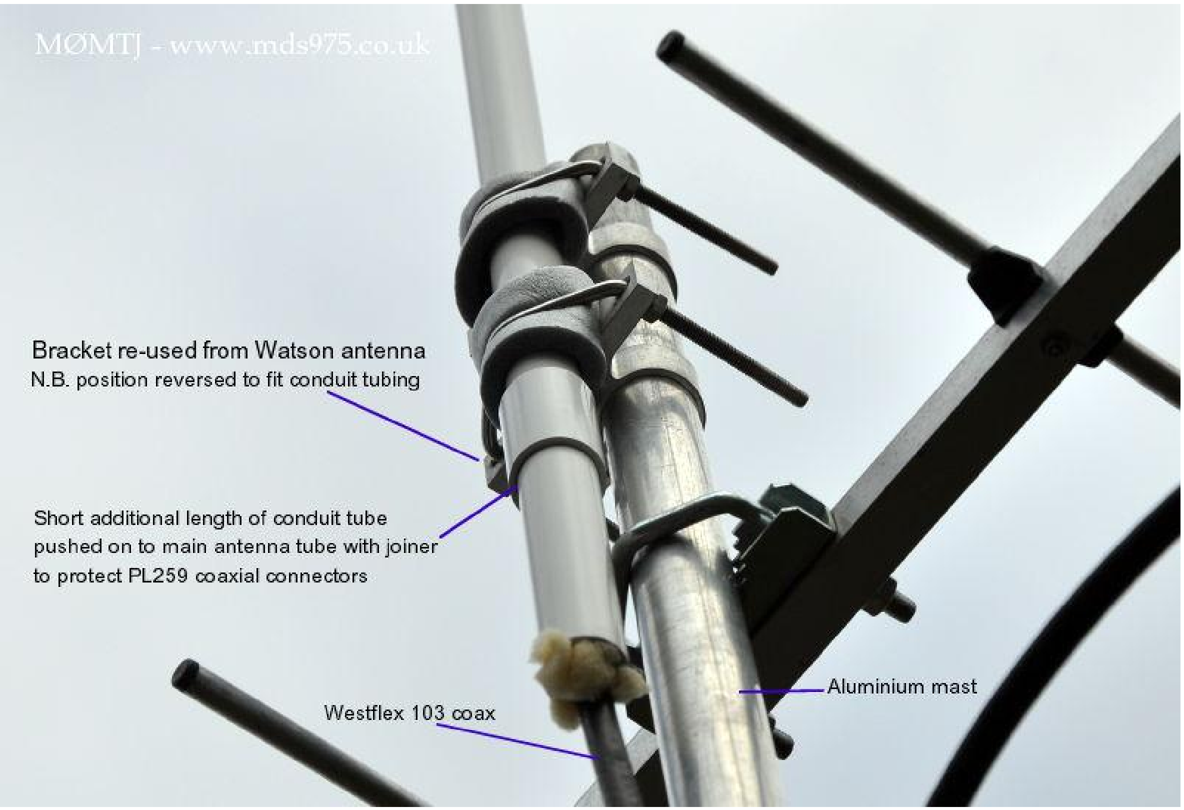
M0MTJ - www.mds975.co.uk
Bracket re-used from Watson antenna
N.B. position reversed to fit conduit tubing
Short additional length of conduit tube
pushed on to main antenna tube with joiner
to protect PL259 coaxial connectors
Aluminium mast
Westflex 103 coax

Photograph detailing the the fixings. The brackets are brackets are from the Watson W-50 antenna which have been reversed so that the smaller diameter PVC tube is held in place by the V Bolts.

For further detailed information and reading, please visit the excellent website of John Bishop VK2ZOI here: http://vk2zoi.com/articles/half-wave-flower-pot/

6 Metre Half Wave Coaxial Dipole - An end fed CFR Dipole antenna supported by a 3 metre long fibreglass fishing pole for 50 MHz

In 2014 I decided to remove the 4 Metre J-Pole antenna from my push-up mast due to the fact that the band is relatively quiet and that I only have the 5 watt Wouxon handheld transceiver for 70 MHz.

I decided that having a good, full size antenna, for the 6 Metre band would be more useful and potentially more rewarding since it can be used with a 100watt HF radio that covers 50 MHz.

The choice of antenna was an easy one. With the great success of the "flowerpot antenna", I decided to build a version for 6 Metres.

Similar designs of Coaxial Dipole antennas had also been featured in recent editions of the RSGB publication RadCom during 2012 and 2013. The antennas in RadCom are described as Controlled Feeder Radiation (C.F.R.) Dipole antennas - the tuned choke at the feed-point controlling, or choking off, the common mode current that would otherwise flow down the outside of the coaxial feeder cable causing E.M.C. issues on transmit, high S.W.R. reduced efficiency and noise on receive. See references below

The advantage of this design is that it can be physically end fed, so there is no feeder cable to route away from the centre of the dipole. Electrically, however, the feed point is at the centre of this dipole aerial, as explained earlier.

Rather than fit the antenna inside a PVC pipe, as with the previous dual band 2m / 70cm antenna described above, I decided to use a lighter weight and less conspicuous 3 metre long fibreglass fishing pole as the support. The completed radiating element simply being taped to the fishing pole.

I cut a length of MIL spec RG58 cable, about 4.5 metres long, to form the bottom half of the radiating section and the choke coil, leaving enough to form a short length (about 30 cm) of cable below the choke coil on to which is soldered a PL259 plug.

The choke consists of 15 turns of the RG58 coaxial cable wound on a 50mm diameter plastic former cut from the empty tube of a cartridge gun that previously contained silicone sealant - allowing a 30cm tail on to which the PL259 plug is fixed on one side and and about 1.31 metre length on the other side that will form part of the radiating element.

The half wave radiator therefore consists of a quarter wave bottom section of the RG58 cable and a quarter wave top section consisting of a length of multi-strand (single conductor) P.V.C. covered antenna wire.

A quarter wavelength at the mid point of the 6 Metre band is: 300 ÷ 51 MHz = 5.88 metres ÷ 4 = 1.47 metres

Due to velocity factor the actual length of the quarter wave sections will be shorter. With the materials that I used, I found that a factor of about 87% was about right, the 1/4 wave length being 1.29 metres.

The top tip of the bottom 1/4 wave section of the coaxial cable is stripped of about 2 cm of outer sheath and braid leaving the length of braided section, measured from where it exits the coil, 129 cm long. The inner conductor is then stripped of 1cm of insulation. This is effectively the centre point of the dipole. To this point is soldered the 129 cm length of the P.V.C. covered aerial wire to form the top half of the antenna. In practice, use a slightly longer length of wire, and then fold over the excess to for the 129 cm length - this can then be used to adjust for lowest SWR at 51 MHz.

The length of the radiating section was therefore about 260 centimetres, plus about 13 centimeters for the coil former giving a total length of 273 cm. This allows about 27 cm of the bottom section of a 3 metre fishing pole to be used to fix to a supporting pole or mount - e.g. to the top of an aluminium mast.

Drill four small holes in the choke former so that when the fishing pole is placed through the centre of the former it can be fixed to the pole using two cable ties.

The radiating section (coax and PVC covered wire) is fixed to the top section of the fishing pole with good quality insulating tape.

The aerial can now be temporarily fixed to the mounting pole using suitable brackets - taking care not to crush the delicate fibreglass! Connect the PL259 plug to the antenna feeder cable using an SO239 back-to-back coupler and test the SWR with an antenna analyzer or SWR bridge. The lowest SWR should be centred on 51MHz and be low - less than 1.5. My reading was 1.2.

If the point of lowest SWR is significantly away from 51 MHz and/or the SWR at the band edges is too high (i.e. over 2) then length of the radiator will need to be adjusted. If the point is too low in frequency, the antenna is too long and will need to be shortened. If the point is too high in frequency, the antenna is too short and will need to be lengthened.

Adjustment can be achieved by pulling the coax through the coil to make it longer, or pushing the coaxial cable back into the coil to make it shorter. Also ensure that the coil winding are adjusted to that they remain tight and neat. The top PVC covered wire section will also need to be lengthened or shortened accordingly by adjusting the folded over section.

Note that in practice, to obtain the very lowest SWR, the top PVC covered wire section may need to be slightly longer by perhaps 1 or 2 centimetres. This is probably due to the fact that the velocity factor of the PVC covered antenna wire is a little greater than the coaxial cable.

Once the antenna is adjusted correctly, ensure that the wires are securely taped to the fishing pole. Connect the permanent antenna feeder to the aerial using the SO239 coupler and weatherproof the joint thoroughly using self amalgamating tape. Use the very best quality coaxial cable possible to ensure lowest loss. I use Westflex 103, but consider MIL Spec RG8 or RG213 as the minimum standard.

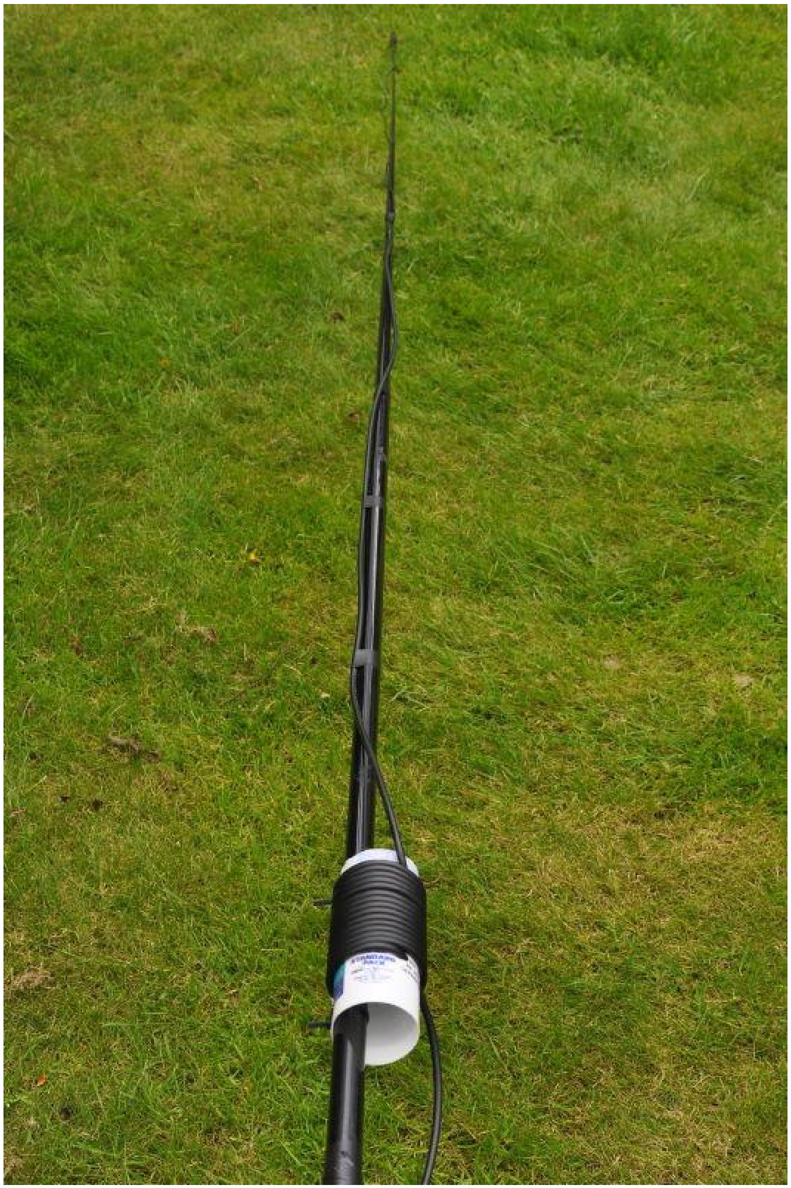

50 MHz Coaxial Dipole / Controlled Feeder Radiation Antenna / 'Flowerpot Antenna'
A physically end fed fed half wave dipole antenna for 6 Metres

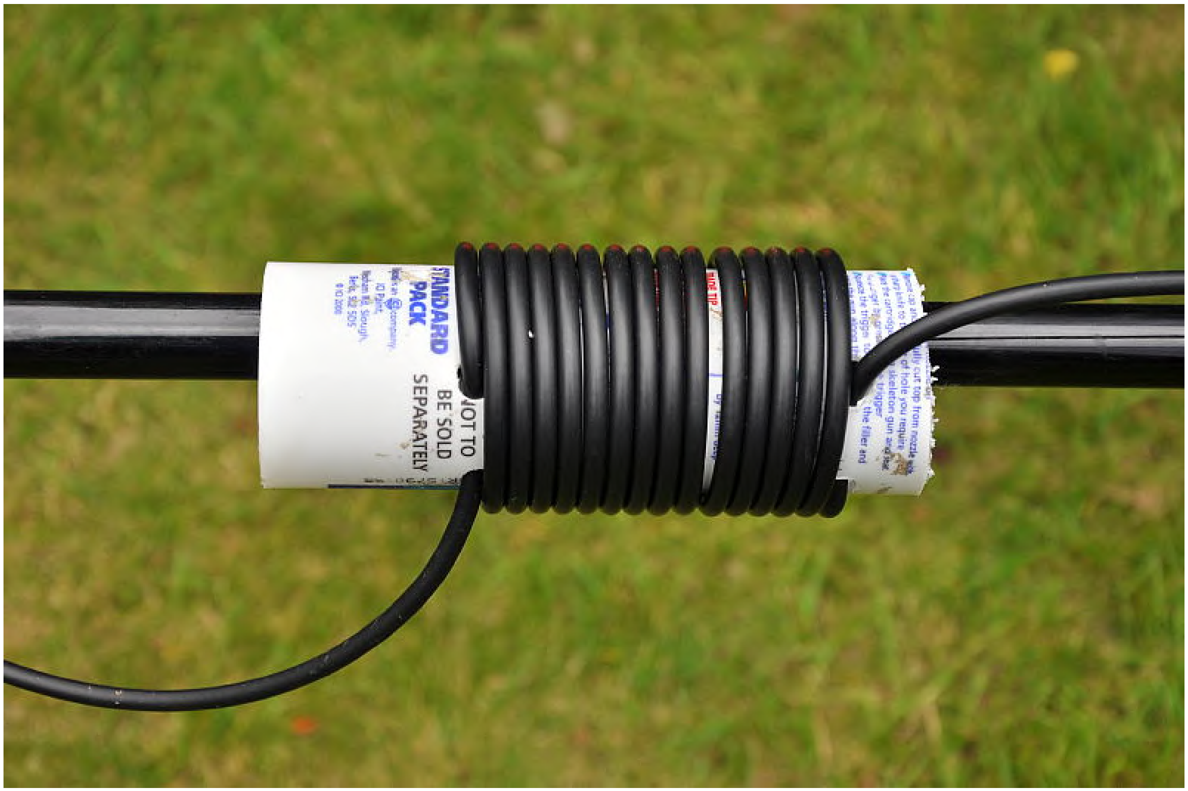

Choke Coil - 15 turns of the RG58 coaxial cable on a 50mm diameter former

VHF Band II Broadcast Band antenna for 88 to 108MHz -
Physically end fed, Coaxial Dipole / Controlled Feeder Radiation Antenna (CFR Dipole) for VHF broadcasts

I used a 2 metre long length of 25mm white plastic pipe, strengthened with with a 2 metre length of 21.5mm overflow pipe pushed up the inside.

The radiating element is made from good quality 75 Ohm coaxial cable. Since VHF/FM broadcast tuners are designed to be fed with 75 Ohm coaxial cable, use high quality, low loss double shielded satellite grade coaxial cable for the feed between the aerial and the radio tuner. Use satellite F type connectors and joiners for lowest loss.

The top half of the radiating section is 670mm of the centre conductor (or a length of multi-strand PVC covered wire). The bottom half of the dipole is 650mm of the complete coaxial cable - choked off at the bottom by the coil section. The top part is held in place by a a short length of thin nylon cord, trapped in place by the top PVC cap. The cap itself is sealed on the outside by some self amalgamating tape.

The choke coil is 22 turns of the 75 Ohm coax wound around the 25mm pipe. Tightly spacing the windings of the coil will minimize the bandwidth covered but provide the lowest SWR at the centre point. A looser winding of the coil will widen the bandwidth covered, but lowest SWR achived will be a little higher.

Once the tuning of the radiating elements, coil winding and band coverage has been checked with an Antenna Analyzer, the coil section should be covered with heat shrink, taking great care not to overheat and deform the plastic pipe.

The actual final dimensions (as shown below) may well need some adjustment in length due to differences in cable and type of pipe used. However, with the dimensions shown, I achieved a minimum SWR of 1.3 at 97.4 MHz. The band edges at 88MHz and 108MHz were at an SWR of around 3.8 to 4.0 - which is probably OK for broadcast band reception. The frequency of lowest SWR can can be changed, if desired, by changing the lengths of the radiating sections - slightly longer for a lower frequency and slightly shorter for a higher frequency.

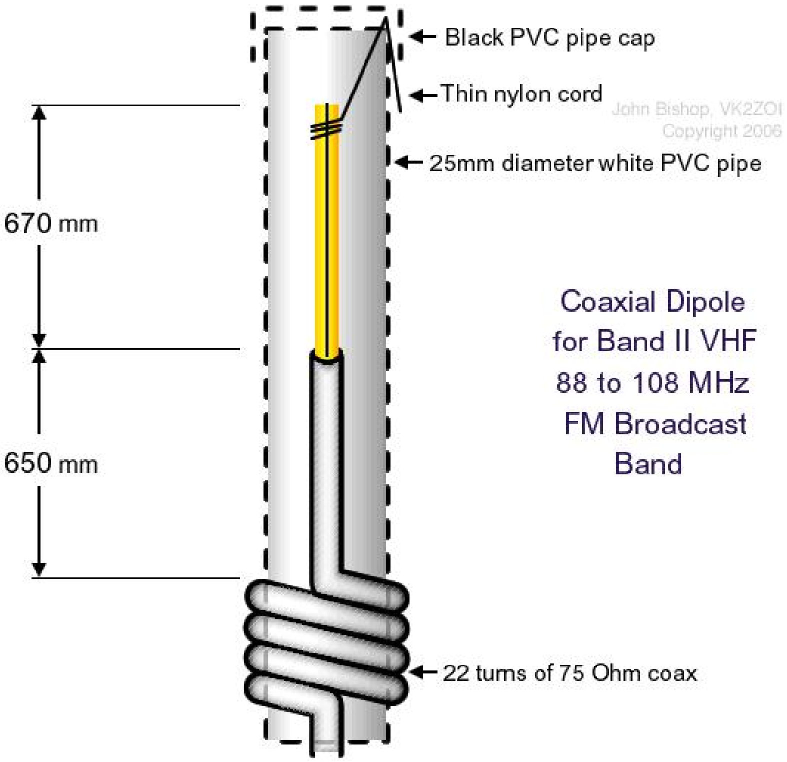

Coaxial Dipole for the 88 to 108 MHz Band II Broadcast Band
An effective, cheap and simple vertical antenna that is fed at its base

RSGB RadCom Articles:

The Controlled Feeder Radiation Dipole. Peter Dodd. RadCom, September 2012, page 22.

More On The CFR Dipole and Coax Chokes. Peter Dodd. RadCom, October 2012, page 54.

VHF CFR Dipoles and More on Common Modes Chokes. Peter Dodd. RadCom, January 2013, page 24.

The HAK Chokes Coaxial Dipole. Encouraging results from 2m to 20m. Peter Grant. RadCom, April 2013, page 22.

Join the RSGB to receive your monthly RadCom magazine: http://rsgb.org/main/publications-archives/radcom/

VK3TWO / VK6TWO Comments:

Many years ago when I was working for a Service Centre, by accident we broke a commercial white stick antenna whilst using it for a task it wasn't intended for. I found that inside the fairly expensive commercial antenna the basis for the design was very similar to the coaxial dipole (as I call them).

Our local repeater club calls them "pogo sticks", which I can only assume is due to the coax coil resembling the spring of a pogo stick. This design had RG213 being fed inside an aluminium tube (the bottom radiator), and the outer braid was then terminated to this. The inner of the coax, then terminated to an identical aluminium tube which was of course the top radiator. Where this design largely differs, is that there was also a 1/4 wave stub of coax running parrallel with the bottom radiator (note that this was for single band operation, not dual band). The whole lot then slid inside the typical white tapered fibreglass housing. I didn't cut open the bottom mounting section to see how it was choking the RF, but I assume it had a handful of ferrites inside the metal base (an alternative way of making the RF choke).

The repeater group has built probably hundreds of these and sells them for $40 at local hamfests. We had a 'Jig' made so that all of the measurements were 'pre-marked' etc, and allowed us to mass manufacture them. Last year, we had planned the typical 'working bee' to make about 30 of them, but this time we had a very expensive Anritsu Sitemaster at our disposal. We discovered that the whilst the design we'd used for decades had a good SWR, it actually was far from optimal. With a heap of 'trial and error' (and with excellent visibility of what was really going on via the Anritsu - not just SWR) we were able to fine tune the design.

From memory, we actually needed more turns of the coax than we had been using (10.5 turns from memory), and our cutting measurements altered slightly. As to be expected, the number of turns was largely dependent on the size of the conduit being used, even changing from 25mm to 30mm etc.

We took several screenshots of the final SWR plots etc. These showed that the bandwidth of this antenna as VERY wide from an SWR point of view. I'll see if I can dig some of them up for you if you'd like.

73, Heath, VK3TWO / VK6TWO
MEngSc, GDipCompSc, DipEE
www.spooktech.net

http://www.warg.org.au - The West Australian Repeater Group Inc (WARG) is the largest amateur radio club in Western Australia (VK6).

D.I.Y. J-Pole Antennas - A really simple, quick and very cheap 'home brew' project J-Pole Antennas

J-Pole antennas for 2 meters, 4 metres, 6 metres and 10 metres :

While experimenting with antennas in the garden in the summer of 2012 I thought that it would be good to have a hand-held radio in the shed to do some monitoring and make a few contacts. To improve upon the performance of the 'rubber duck' antenna I quickly made a J-Pole antenna for the 2 metre band.

It is made from a 47cm length of 450 ohm Wireman ladder line as the 1/4 wave matching section, plus a 97cm length of stranded wire as the 1/2 wave radiator. It is fed with 3 metres of Mil spec RG58 c/u coaxial cable that is soldered to the 1/4 wave matching section's impedance matching point at 3.5 cm from the bottom. The coax feeder is wound around some PVC tube to form a choke. The completed antenna is taped to a 2.2 metre long fibreglass fishing pole that I purchased from Poundland (for £1.00). It took about 20 minutes to make followed by some testing and adjustment with the antenna analyser. The fishing pole is lashed to the shed with some cable ties.

This simple antenna works pretty well, but being so low down signal strengths are not huge, but it's pleasing to get on the air with something so simple and cheap!

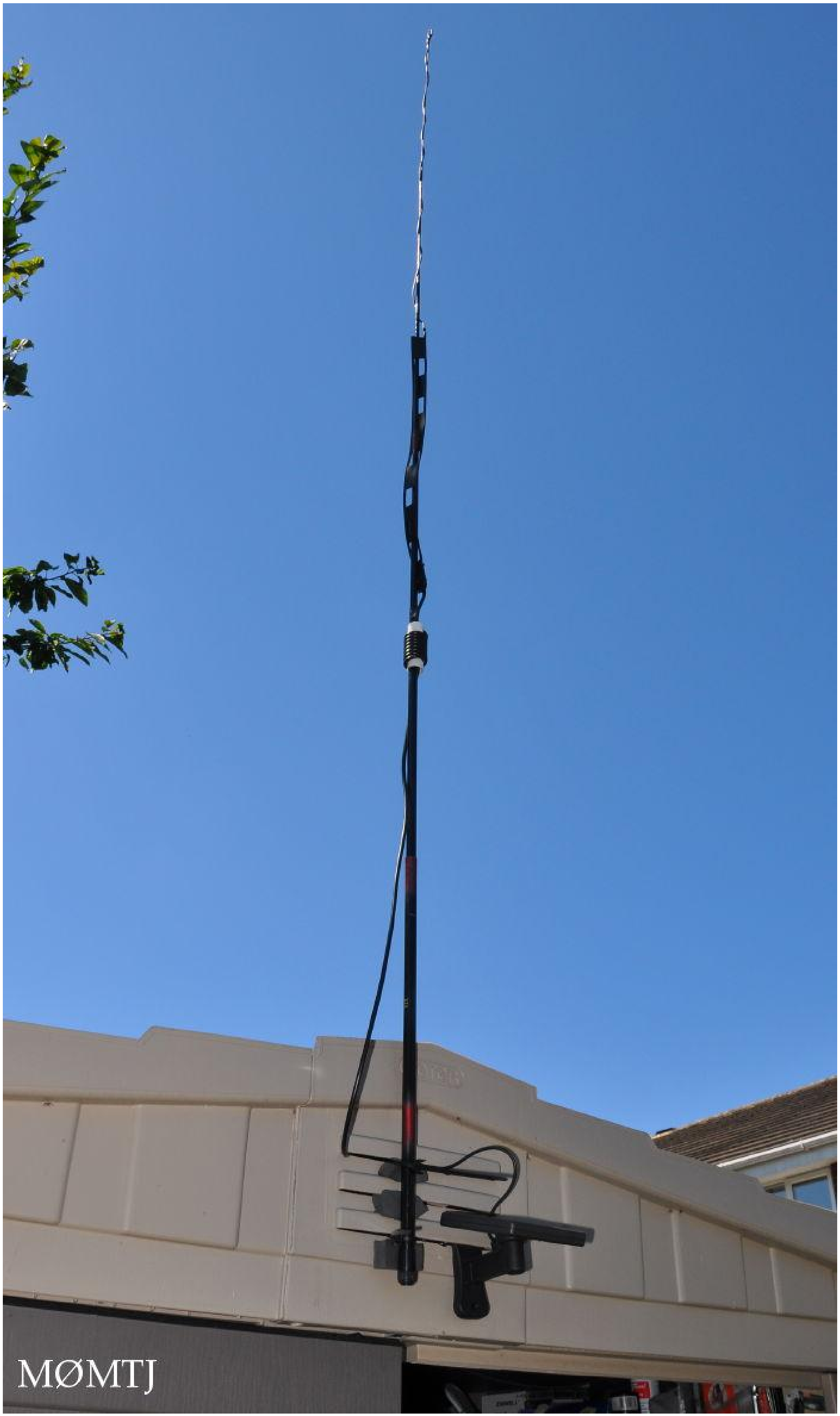

The Shed Antenna - a 2m J-Pole by M0MTJ
Note the simple choke balun at its base made by winding 8 turns
of the coaxial cable around a small off cut of white PVC water pipe.

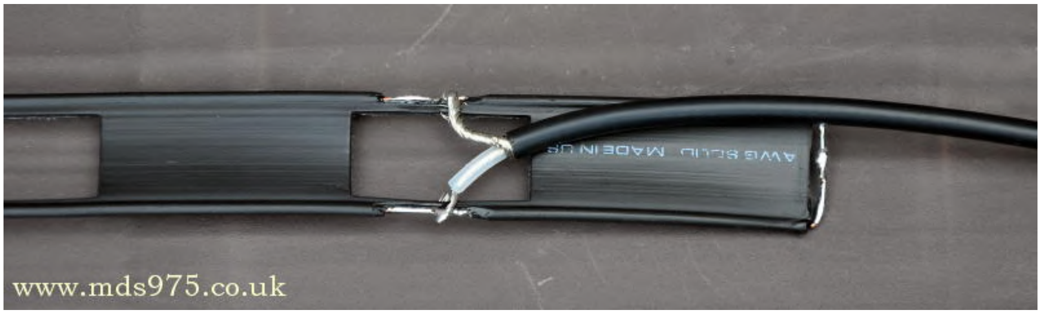

The feed point of a J-Pole antenna made from Wireman 450 ohm ladder line.
For the 145 MHz antenna this feed point is 3.5 cm from the bottom of the ladder line section which is on the right hand side in this photograph. The coaxial cable used in this case was Mil spec RG58 c/u. But any good quality, low loss 50 ohm coaxial cable could be used. The wire radiator section is connected to the same conductor of the ladder line as the coaxial cable's centre conductor. For my antenna, fixed to a fibreglass fishing pole, the radiator wire was 97cm in length.

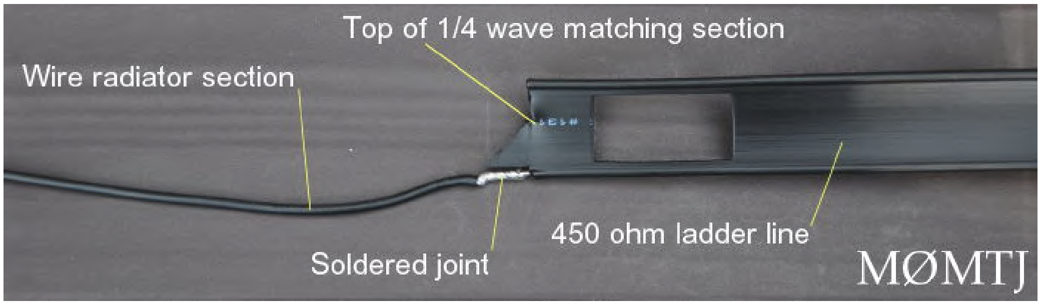

Photograph showing the point where the PVC covered wire that forms the half wave radiator section is soldered onto the top of the 450 ohm ladder line that forms the quarter wave matching section.

Inspired by DK7ZB. The J-Pole is a very effective antenna and being made of wire it is very light weight making it quite easy to fix in different positions. If you have problems installing a permanent antenna then making a wire antenna that can be easily supported on a lightweight push up telescopic fishing pole can make an ideal alternative.

The formulas to make a J-Pole antenna from 450 Ohm Wireman ladder line in this way are:

Length of 1/4 wave impedance matching section (450 ohm ladder line) Wavelength x 0.223
Length of 1/2 wave radiator (any reasonably strong PVC covered stranded wire) Wavelength x 0.471
The point at which the coax is connected to the 450 ohm ladder line will be about 5 to 10% of the length of the ladder line section up from the bottom.

The wavelength at mid point of the 2 metre band (145.00 MHz) is found by the quick calculation 300 ÷ 145 = 2.068 metres

So, to make a practical antenna:

The 1/4 wave section of 450 ladder line will be 2.07 x 0.223 = 0.47 metres long
The 1/2 wave wire radiator will be worked out as 2.07 x .471 = 0.975 metres long
The connecting point of the coax will be about 3.5 cm from the bottom of the 1/4 wave section. The optimal point may have to be found by some experimentation - as will the best length for the wire radiator.

The length of the wire radiator will be affected by surroundings. For example I fixed the wire to a fishing pole. The proximity of the fishing pole has the effect of electrically lengthening the wire; so using a 97.5cm length of wire fixed to a pole I found that it resonated (as expected) at a lower frequency, it therefore had to be shortened until the point of resonance (indicated by lowest SWR) was around 145.00 MHz. This should be done in the antenna's expected final position since the J-Pole is quite sensitive to its surroundings, so if these checks are done near the ground, once it is raised into its final position the SWR will have changed and the adjustments will have to be done again.

I found that 3.5 cm was good for the 2 metre band antenna, but for the 10 metre band version of the antenna a little more experimentation was required:

The VSWR reading may not be especially low, even though the point of resonance for the wire radiator may have been found. For the 10 metre band antenna at this this stage was about 1.7 indicating that the connection point of the coaxial cable to the 450 ladder line needs to be adjusted. The ladder line is used as an impedance transformer, transforming the very high impedance (hundreds of ohms) of the half wave wire radiator down to the 50 ohms required by the transceiver and the coaxial feeder cable. This connection point therefore affects the impedance of the antenna, the higher up the matching section it is the higher the impedance will be, and visa versa.

Once the length of the wire radiator has been set, the connection point can be moved up and down the ladder line until lowest SWR is achieved. A few centimetres of the PVC insulation has to be carefully scraped away from the copper conductor on each side of the ladder line using a craft knife. The inner conductor of the coaxial cable is quickly tack soldered on the side that is connected to the 1/2 wave wire radiator. The coaxial cable's braid is quickly tack soldered to the opposite side of the ladder line at this point, ensuring that both points are equal distance from the bottom. At this point temporary croc clips could be used, but I preferred a quick solder joint.

With radiator trimmed for resonance, the connection point of the coaxial cable can then be moved up or down the ladder line little by little; un-soldering and re-soldering the coax to the ladder line until a lowest possible SWR is achieved, indicating that the antenna is near the ideal 50 ohm impedance.

Once the ideal point is found the coaxial cable can be properly and permanently soldered to the ladder line.

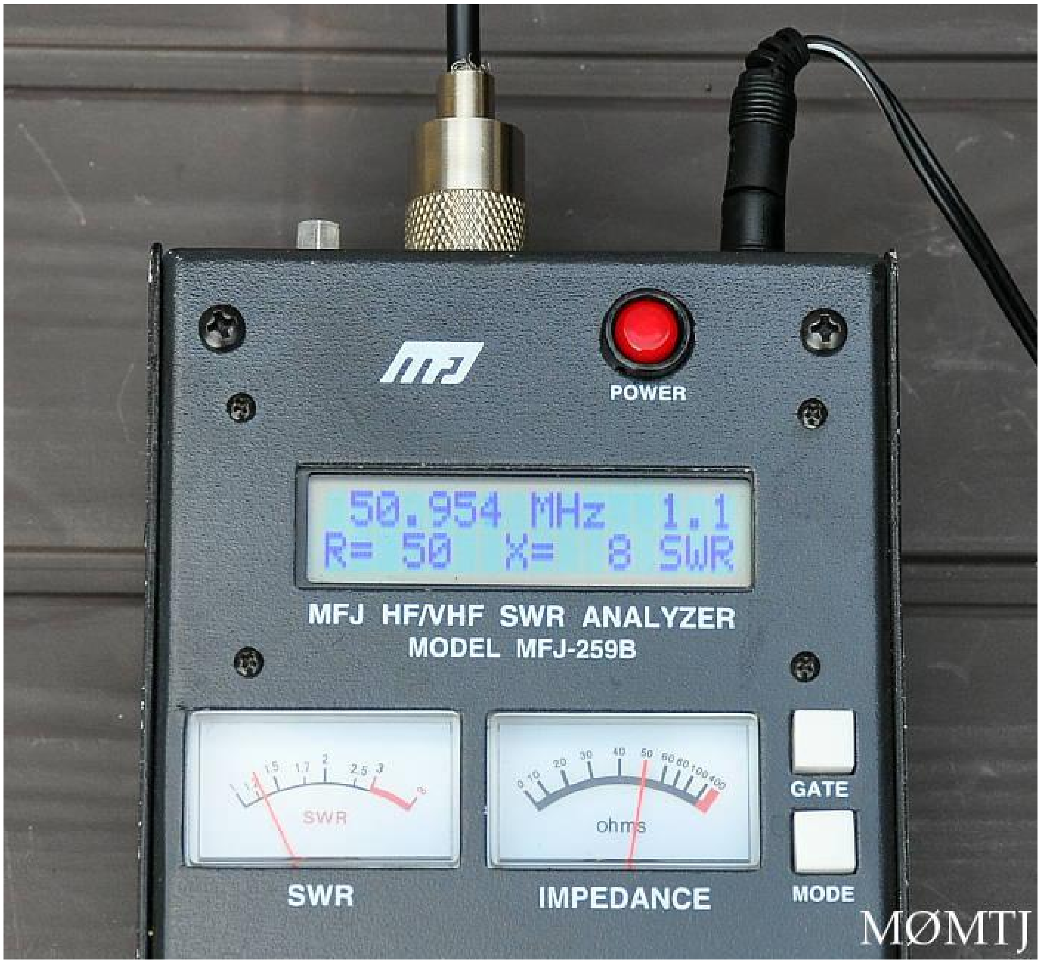

6 Metre Band J Pole on the antenna analyser - it's getting close!

Each J-Pole took about 20 minutes to physically make out of the wire components. However the testing and adjusting took a bit more time. I used an antenna analyser which saved having to key the mike every time when using a basic VSWR bridge and causing unnecessary QRM, but even so, hoisting the fishing pole up and down numerous times took a little more time:

10 Meter J-Pole. For the 10 metre band J-Pole antenna this took perhaps another 20 or 30 minutes until I was satisfied with the adjustments. It may take a little longer if using an SWR meter.

6 Meter J-Pole. For the 6 metre band antenna the radiator wire had to be trimmed a little and the feed point adjusted to 6 cm, taking about 10 additional minutes to complete.

4 Metre J-Pole. For the 4 metre band, centred on 70.37 MHz

2 Meter J-Pole. For the 2 metre band antenna the wire radiator took a couple of attempts to get it to the correct length when attached to a fishing pole, but the feed point was spot on first time at 3.0 cm, again taking about 10 additional minutes to complete.

Here are some suggested dimensions for the 2 metre, 6 metre and 10 metre band versions, when supported by a fibreglass fishing pole:

Wire J-Pole Antennas	1/2 Wave Radiator	1/4 Wave Section	Feed Point
2 Metre Band Antenna	**0.975 m**	**0.47 m**	**3.0 cm**

4 Metre Band Antenna	**1.90 m**	**0.95 m**	**6.1 cm**
6 Metre Band Antenna	**2.815 m**	**1.33 m**	**6 cm**
10 Metre Band Antenna	**4.96 m**	**2.45 m**	**15 cm**

N.B. The 1/2 wave wire radiator section will be shorter than calculated when fixed to a fibreglass pole or other object.

To re-cap, the 1/2 wave section should be adjusted for resonance and the feed point position adjusted for minimum VSWR.

Sealing and waterproofing. Once the antenna is complete and has been checked and tested all the bare joints should be sealed against the weather with liquid electrical tape and self amalgamating tape. The coax should also be secured against the ladder line with a nylon cable tie as a strain relief to prevent the soldered feed point joints from breaking.

These J-Pole Antennas were inspired by DK7ZB - http://www.qsl.net/dk7zb/J_Pole/wiremanjpole.htm

---

Antennas 1 | Antennas 2 | Antennas 3 | Antennas 4 | Antennas 5 | Antennas 6 | Antennas 7

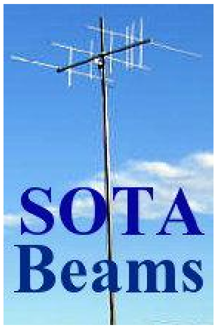

SOTA Beams
Lightweight 2 metre &
70cms Yagis, Dipole,
Accessories & Poles
www.sotabeams.co.uk

G Whip Antenna Products

G-Whip Antenna Products
Get on the air with
TOP QUALITY
BRITISH
G-WHIP
ANTENNAS
End Fed Zepps and
many more designs
and products

PLA-PLT Latest News
Keep up-to-date on this important issue

Radio Interference?:
PLT & PLA: Radio Interference
British Telecom, BT Home Plug
P C World, Powerline Adapters
YouView, Belkin Linksys
Netgear Devolo, BT Vision - BPL
Radio Interference
PLT Page 1

Antennas 1 | Antennas 2 | Antennas 3 | Antennas 4 | Antennas 5 | Antennas 6 | Antennas 7

Top of page ^

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits | Useful Information

Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites | RSGB | QSL | The Amateur Radio Mini Site Map

WSPR Weak Signal Propagation Reporter

News, Developments, Events, Photographs and other 'Bits 'n' Bobs'

MDS975 Home

Feedback

MAIL

Mike Smith - MDS975.co.uk © 2003 - 2016

M0MTJ

Amateur Radio; Ham Radio; Radio; Transceivers; HF; VHF; UHF; Data Modes; Morse Code; RTTY; PSK31; SSTV; FSTV; Amtor; Sitor Antennas; Aerials; Cable; Coaxial Cable; Twin Lead; Masts; Poles; Propagation; Computer; PC; USB Computer Interface; Microphone Loudspeaker; Filters; Noise Reduction; DSP; Digital Signal Processing; Morse Key; SWR ; Inverted L; Inverted V; Dipole; Doublet.

My Home Page

Amateur Radio
Operating Conditions
Antennas
Accessories
Projects & Kits
Data Modes
Information
Portable
QSL
RSGB Ltd
Links
QRM
Bits 'n' Bobs
Contact

WSPR
Weak Signal Propagation Reporter

PORTABLE

Site Map

MDS975 Home

Feedback

PLT QRM

Site Search
Find
powered by FreeFind

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits
Useful Information | Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites
RSGB | QSL | The Amateur Radio Mini Site Map
News, Developments, Events, Photographs and other 'Bits 'n' Bobs' | WSPR Weak Signal Propagation Reporter

AERIALS (ANTENNAS) 7

Antennas 1 | Antennas 2 | Antennas 3 | Antennas 4 | Antennas 5 | Antennas 6 | Antennas 7

Index To Other Antenna Pages:
Antennas 1 : Aerials used by M0MTJ
Antennas 2 : Including ideas for compact antennas for Top Band /160 metres
Antennas 3 : Felix Scerri VK4FUQ discusses Loop Antennas, baluns, masts & other antenna related topics
Antennas 4 : Many antenna ideas from various sources particularly for multi-band operation & also gives information about
antenna trimming, knots for wire antennas and useful antenna rigging accessory ideas.
Antennas 5 : Half Wave End Fed antennas for 144 MHz VHF / 430 MHz UHF and 50 MHz 6 Metre band & J-Pole Aerials
Antennas 6 : Simple and effective H.F. Antenna ideas - Ground Plane and the All Band Doublet Aerial
Antennas 7 : Experimental Omni-Directional - Circularly (Mixed) Polarized Antenna for VHF / 2 Meters.

"Everything should be made as simple as possible, but not simpler." - *Albert Einstein*

MØMTJ - Experimental (Almost) Omni-Directional - Circularly (Mixed) Polarized Antenna for 144 MHz / VHF / 2 Meters.

Introduction and Background

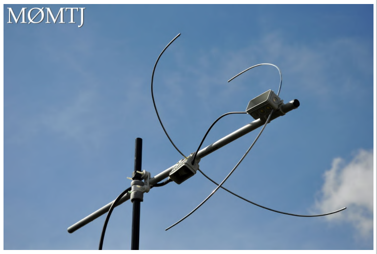

This is an experimental project that may lead to further development in the future, but even as presented below, it does work and may prove very useful.

I have therefore included this page of information so that you may can construct the aerial and develop these ideas further - either just out of interest and for the sake good old traditional Amateur Radio experimentation or for more permanent and practical uses!

The idea was sparked by the fact that I had been toying with the idea of removing the rotator and dual-band horizontal Yagi from my lightweight push-up aluminium mast to reduce mass and visual impact. I certainly did not want to lose access to the horizontally polarized portions of the 2 Metre (144 MHz) or 70cms (432 MHz) bands, so decided upon trying an Omni-Directional, horizontally polarized aerial.

The question was which antenna would I use to replace the rotatable dual-band Yagi?

Due to the weight of the rotator, the lightweight aluminium mast could not safely be pushed up higher than about 6 meters. Therefore, the Yagi on the stub mast was never much higher than 6 or 7 metres above the ground. A lighter weight omnidirectional aerial, without the heavy rotator, would allow the past to be pushed up to, perhaps, 8 or 9 meters. Maybe even to its full 10 meters on a fine day!

My thinking is that this additional height may compensate slightly for the loss of gain from the Yagi. The additional height would also benefit my vertical Coaxial Dipole for 6 Meters (50 MHz) which sits at the top of this mast. A small bonus.

I decided that I would try simple Halo antennas (HAlf wave LOop) and Big Wheel antennas for 2 Meters and 70cms.

Circularly Polarized, Omnidirectional Antenna for VHF 2 Meters / 144 MHz

.
However, before embarking on those particular experiments and possible future changes, I wanted to experiment with mixed polarization, or more correctly in this case, Circular Polarization - the subject of this project and article.

Why? - Well, when the IBA (Independent Broadcasting Authority) began building transmitters for their Independent Local Radio (ILR) in 1973, they specified that mixed polarized aerials should be used used at each of their VHF/FM

G Whip Antenna Products
[Get on the air with TOP QUALITY BRITISH G-WHIP ANTENNAS]

transmitter sites, starting with Croydon for LBC and Capital Radio in London, then Black Hill for Radio Clyde in Glasgow and Lichfield for BRMB in Birmingham and the many other sites that followed. The aerial systems were either described as "Mixed", "Slant" or "Circular".

The reason that mixed polarization was chosen was to improve reception on car radios and portable radios that generally used a vertical or slanted telescopic rod aerial. The BBC, at the time, used horizontal polarization for VHF/FM radio which was good for permanently installed roof-top receiving aerials, but provided poorer reception for listeners with portable sets and car radios. Some years later the BBC followed the IBA and began converting that majority of their transmitter sites to mixed polarization to improve reception and reduce mobile 'flutter' effects.

I thought that I would like to experiment with mixed polarisation and took some inspiration from some of the more recent, and simpler, antennas that have been installed at commercial radio ("ILR") transmitter sites. Two examples are shown below, both from the Radio Wave transmitter site at Blackpool Tower. The original aerial and its more recent replacement.

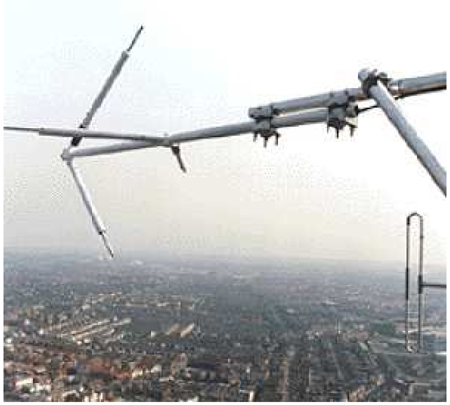

Original SBS transmitting aerial for
Radio Wave Blackpool 96.5 MHz
Photo from MB21.co.uk

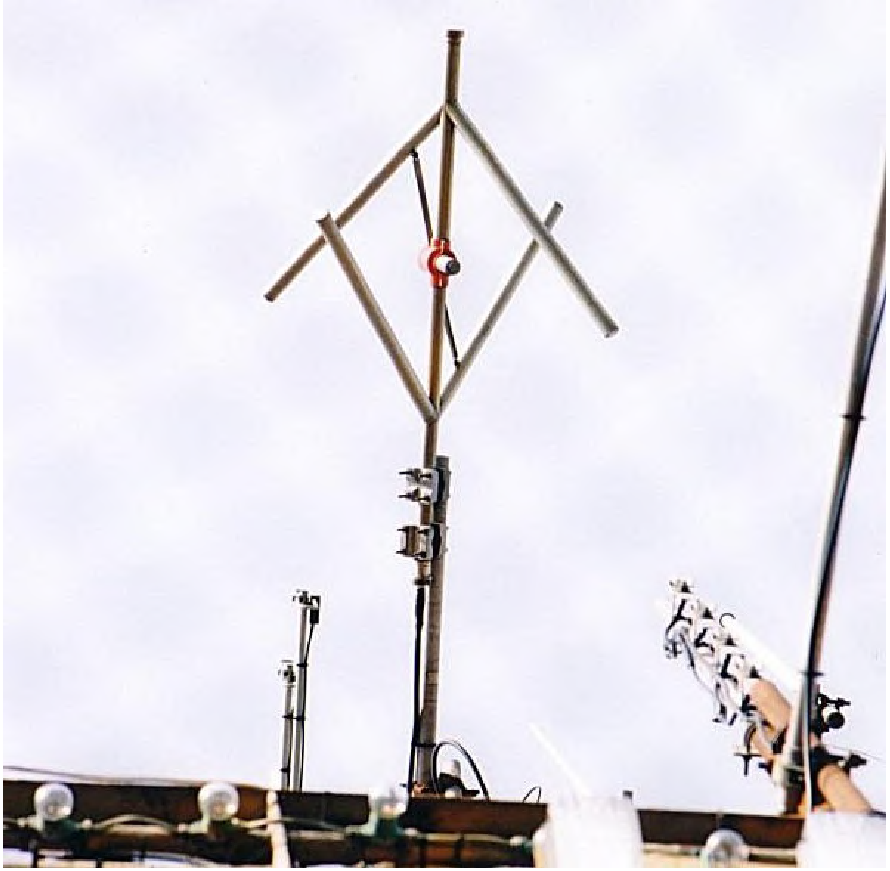

The newer antenna viewed from below
Photo from MB21.co.uk

What's Needed

4 x 500mm lengths of 4mm diameter aluminium rod
2 x Dipole Centres

6 x Solder Tags with 3mm diameter hole
2 x Dipole Mounting Boxes - e.g. IP54 85 x 45 x 40mm
2 x Stainless Steel Nuts, Bolts and Washers M4 x 20mm
2 x Stainless Steel Nuts, Bolts and Washers M4 x 14mm
4 x Stainless Steel Self Tapping Screws No. 4 (2.9mm) diameter by 9.5mm long
1 meter of 20mm or 25mm Diameter PVC Pipe
8 x Pipe Clamps to suit diameter of boom and diameter of supporting mast
4 x Stainless Steel Nuts, Bolts and Washers
2 x End Caps to suit diameter of PVC boom (i.e. either 20mm or 25mm)
1 x SO239 Chassis Mount Socket
500mm length of small diameter coaxial cable, e.g. RG188 A / U - 50 Ohms PTFE coaxial cable
1 x piece of PVC tube for choke balun - 16mm diameter x 30mm long
1 x length of coaxial cable for phasing section
Some small self tapping screws to fix the pipe clamps to the boom

These items will be available from many and various sources, but Nuxcom is a good supplier - http://shop.nuxcom.de/

The antennas essentially comprise of two Crossed Dipoles, set a 90 degrees from each other. In the first photo, the antenna can be seen to have one dipole positioned vertically and the other horizontally. The second photograph shows that the to dipoles are again set at right angles, but one slopes 45 / 225 degrees and the other at 315 / 135 degrees.

The two dipoles are connected to each other, but have to be fed with the current to each dipole 90 degrees out of phase. This is called phase quadrature. It is a very common method used for other types of antenna, such as the Turnstile which is often used for communication with satellites where circular polarization must be used.

One dipole is connected directly to the coaxial cable. (I used Westflex 103 coaxial cable, for lowest loss, to feed the antenna). The second dipole has to be fed via a 1/4 wave (90 degrees) matching section. I had an off-cut of RG58, so used this for my initial experiment.

A quarter wave at 145.00 MHz is 300 / 145 = 2.068m / 4 = 51.7 cm. However the Velocity Factor (VF) of the cable has to be taken into account because radio waves travel more slowly through transmission lines than the do through free space.

Ideally I should have measured the VF with my antenna analyzer, however I took the typical VF of RG58 as being 0.66. So 51.7cm x 0.66 = 34.1cm. I cut the cable to 34.1 cm and used this to connect the two dipoles together.

Initially I cut the 4mm diameter aluminium rod into four 50cm long pieces and mounted them into the two dipole mounting boxes using the Nuxcom Dipole Centres. A small hole has to be very carefully drilled into the end of each rod to align with the fixing hole of the dipole centre. Then very carefully screw the 2.9mm stainless steel self tapping screws into each hole being sure not break off the head of the screws. When tapping these holes it is best to ease back and forth until the hole is successfully tapped.

The dipole centres and pipe clamps are fixed to the box with the 20mm diameter M4 hex bolts, washers and nuts.

The second pipe clamp is fixed to the box with the 14mm long M4 stainless steel hex bolt, washer and nut.

Each mounting box is fixed to the boom by clipping on the pipe clamp. The two dipoles are mounted at right angles (90 degrees) to each other, approximately 335mm apart. Once the final adjustments have been made and the positions along the boom are decided, the clamps are held firmly in place by screwing them to the boom, as can be seen in the photographs below.

Each dipole is then carefully bent into a semi-circle.

The SO239 socket is then fitted to the box and the choke balun is wound and one end of the coaxial cable soldered to the socket and the other end fitted with the solder tags. The small 2.9mm x 9.5mm stainless steel self tapping screws are then used to fix the tags to the dipole centre.

Now connect this dipole to an antenna analyzer (or SWR meter) and check the point of resonance. I decided to centre the aerial on 144.3 MHz. Almost certainly the dipole will be too long. I had to cut off 1mm of rod from each end.

Once complete, do the same for the other dipole so that they are exactly the same size.

Next, the phasing cable is prepared, as mentioned above. A suitable diameter hole is drilled in the end of the box opposite the SO239 socket and the phasing cable passed through this and soldered on to the two tags, as can be seen in the photograph below.

A similar small hole is then drilled into the other dipole mounting box and the other end of the phasing coaxial cable is passed through this, two solder tags are soldered on to the end and then fixed into place on the dipole centre using the small self tapping screws.

Now connect the antenna analyzer again, using a few metres of low loss coaxial cable, and mount the antenna in some free space, preferably outside on a pole about two or three metres above the ground a. It will almost certainly be found that the point of resonance is now two low, so each dipole will now need to be carefully trimmed, cutting small and equal amounts off each of the four ends until the point of resonance is around 144.3 MHz.

It should be possible to obtain an SWR of 1.5:1 or less across the 2 Meter band - to achieve this it may be necessary to move one of the dipole mounting boxes further away, or closer together to the other. Also, it may be necessary to bend the dipole rod of each dipole element inwards, or outwards slightly. I was able to achieve an SWR of 1:1 with an impedance of 50 Ohms according to the antenna analyzer.

Once the adjustment were complete, I used 20 meters of Westflex 103 to connect the antenna to the Icom IC-706 mk2g. A similar length of the same cable is used to connect the 5 element horizontal Yagi to the radio. A low loss coaxial switch was used to make instand comparisons between each antenna.

I raised the experimental antenna to 8 meters above ground level on the telescopic pole and began to make some comparisons.

My guess was that the experimental omni-directional circular polarized antenna would be noticeably inferior to the much higher gain, horizontal, directional Yagi.

The Yagi has a quoted gain of 8.4dDBd (i.e. gain with reference to a standard dipole - not dBi).

A standard straight dipole has a gain of 0 dBd. However if it used horizontally and is bent into a circle to produce a roughly omni-directional radiation pattern, then the gain must be reduced, so I assumed that the 'gain' must now be negative, perhaps -1 or -2 dBd.

Additionally, being as this is a circularly (mixed) polarised antenna, the power is essentially divided in half across the horizontal plane and the vertical plane. i.e. another 'loss' of 3dBd.

Therefore the gain of the antenna may be assumed to be -5dBd (minus 5 dBd), i.e. about 1 S point down on a simple straight linearly polarized aerial and about 13 dBd down compared to the Yagi - i.e. about 2 or 3 S points worse.

Here, at my QTH in Staffordshire, the Kent beacon on 144.430MHz on my 5 element Yagi is just audible, although there is no S Meter reading. I therefore did not expect to be able to hear it on the experimental antenna as it is so weak, but to my surprise, the beacon was still there - obviously a bit weaker, but still audible. That was an encouraging start!

I therefore went ahead and made some QSO's with other stations who were using both vertical and horizontal polarization and also made some comparisons purely on receive.

I therefore went ahead and made some QSO's with other stations who were using both vertical and horizontal polarization and also made some comparisons purely on receive.

I haven't recorded enough data to draw truly meaningful conclusions. It's not scientific either. However, my impression from what I have recorded and from other general use is that the performance is broadly in line with my expectations - i.e. about 2 or 3 S points, or around 12dB, down on the Yagi.

Station	**Location**	**QSO or RX**	**Mode**	**Other station's Aerial**	**Rprt Sent (Yagi)**	**Rprt Rec'd (Yagi)**	**Rprt Sent (Omni)**	**Rprt Rec'd (Omni)**
G3XEV	Sedgley	QSO	SSB	V2000 Vertical	S2-3	S7	S6-7	S7
2W0JYN	Rhos	QSO	SSB	3 Ele Horiz Yagi	S6	S9	S5	S7
2W0JYN	Rhos	QSO	SSB	Vertical Colinear	S1	S1	S0	S0
G8WEG	SW B'Ham	RX	FM	Unknown (Vert?)	N/A	N/A	S9+20	N/A
MW0WML/p	GWNW012	QSO	FM	Vertical Dipole	N/A	N/A	S9+10	S8
G7HEM	Stourbridge	QSO	SSB	6 Ele Horiz Yagi	S7	S9	S2	S3
N/K	Burton	RX	SSB	Unknown (Horiz?)	S9	N/A	S5	N/A

With particular thanks for John G3XEV and Steven 2W0JYN for their assistance with these initial tests!

I hope this information is useful and may encourage you to experiment further and make some much better scientific comparisons!

If you have any comments, please do email them to me via the contact page here.

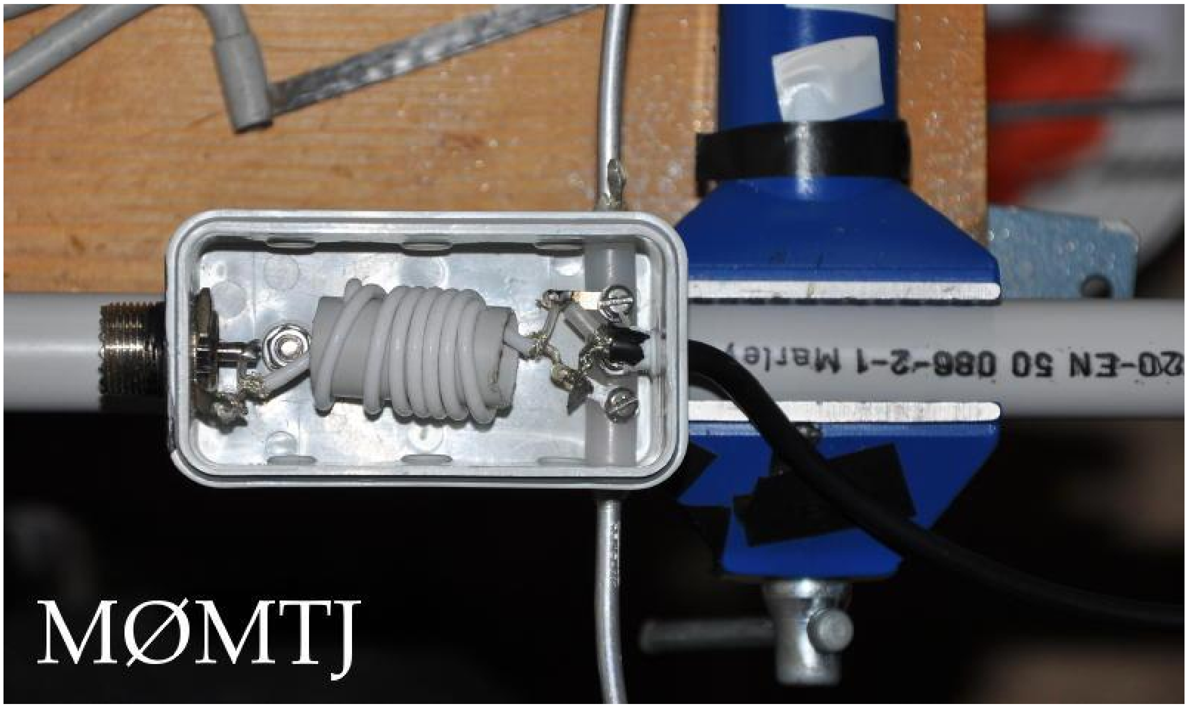

Detail showing the inside of the first dipole mounting box with SO239 socket and choke balun

Nuxcom Dipole Centre for 4 mm dipole rods
10 mm space between dipole rods
Three holes: One in the middle for fixing the box on the boom. The other two for connecting the cable to the dipole rods.
Length: 36 mm
Space in the between rods: 10 mm
Outer diameter: 8.4 mm
Inner diameter: 4,1 mm
Material: Polyamid
http://shop.nuxcom.de/

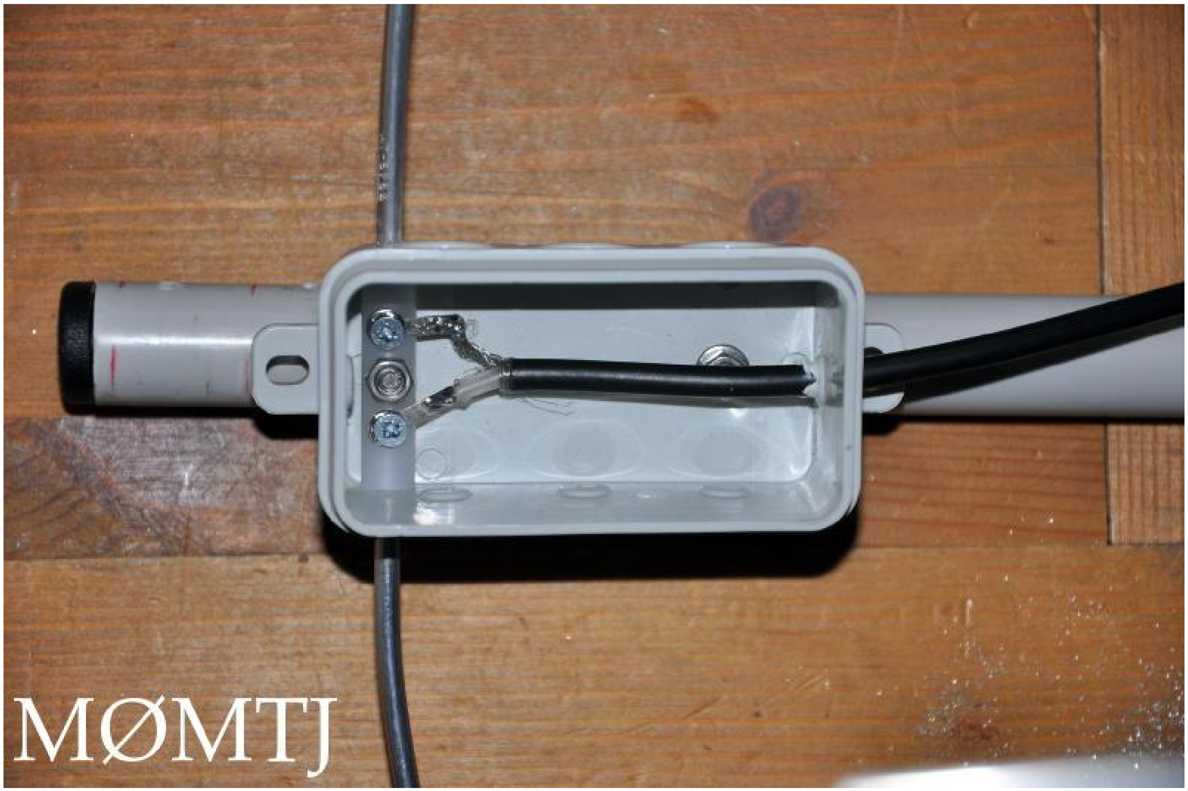

Detail showing the inside of the second dipole mounting box

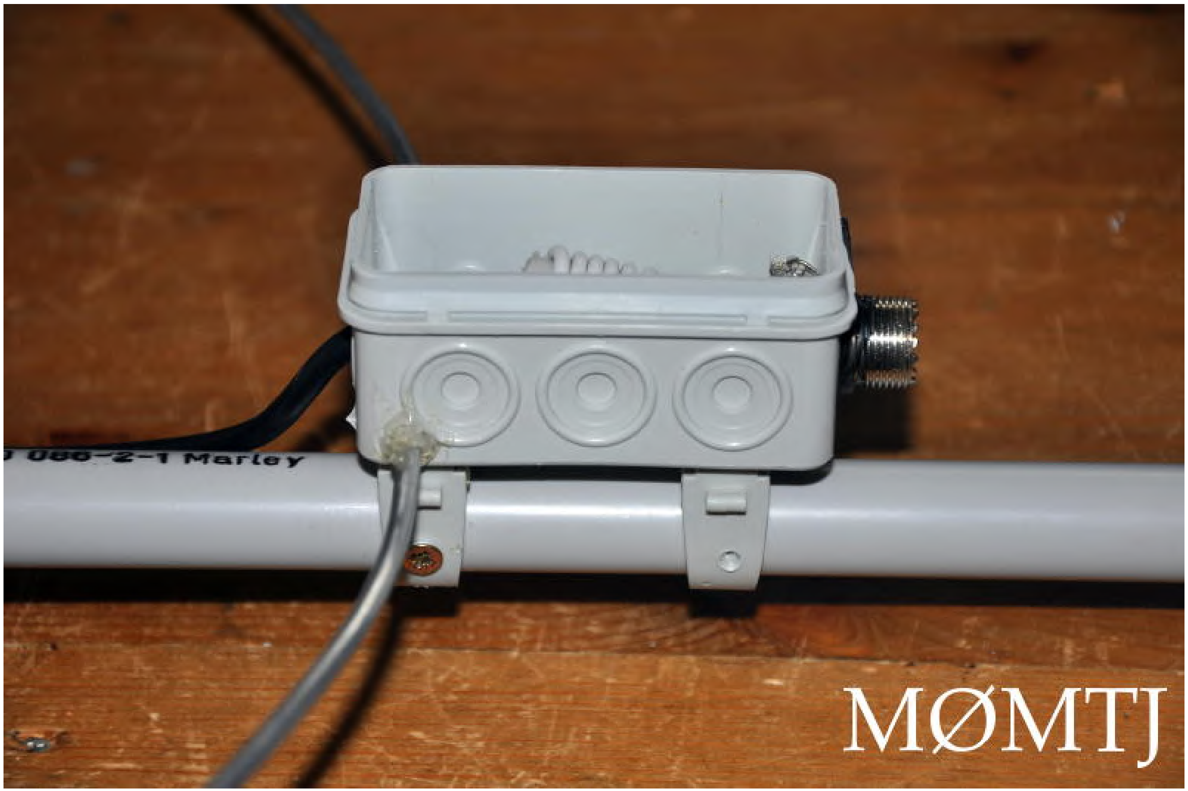

Detail showing how the dipole mounting box is fixed to the plastic pipe with pipe clamps and firmly secured in place with small self tapping screws.

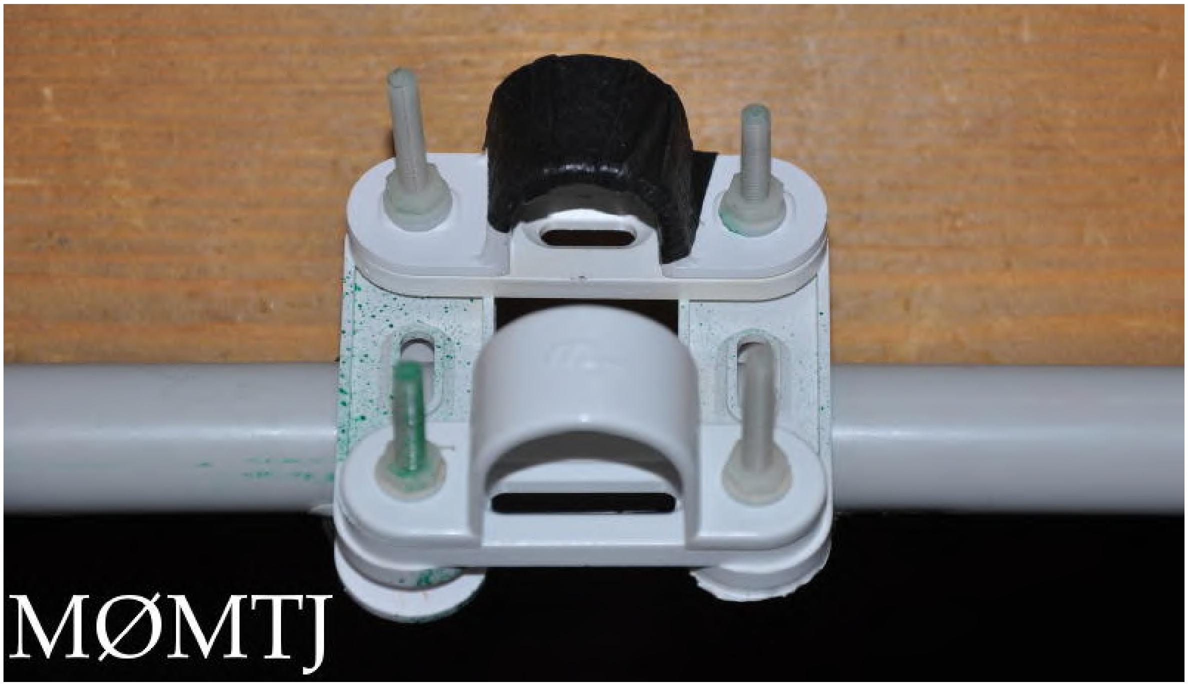

Right angled mast to boom clamp made from four plastic pipe clamps

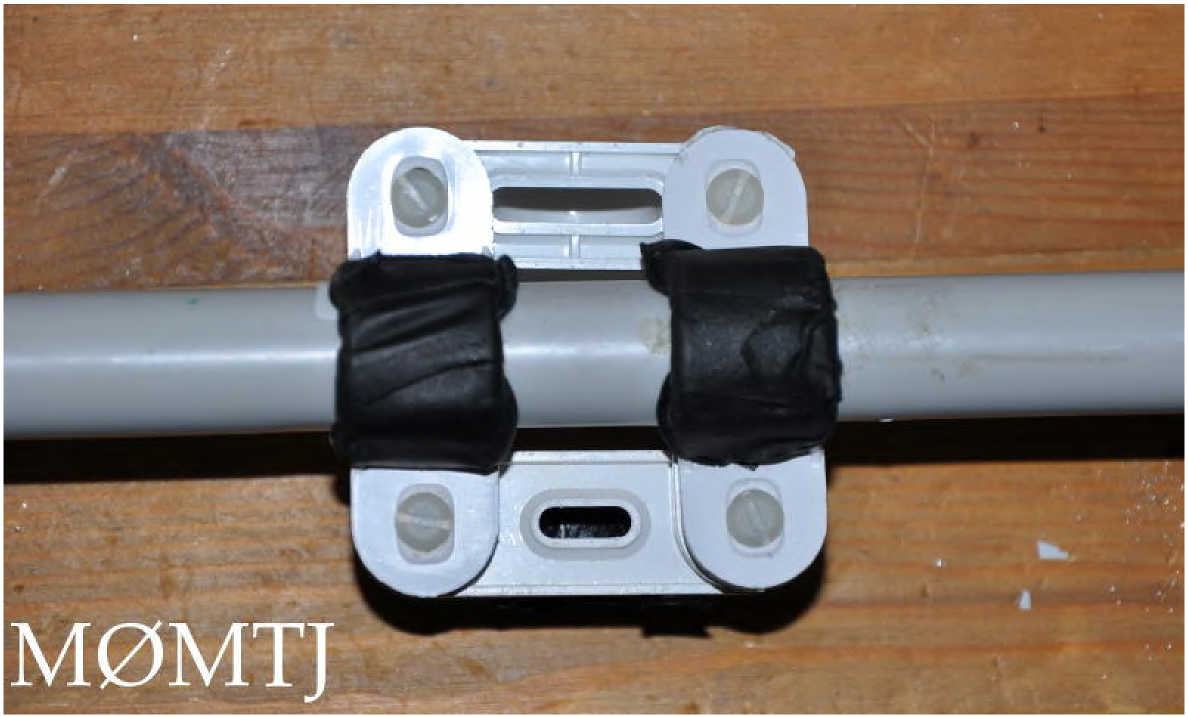

Right angled mast to boom clamp made from four plastic pipe clamps

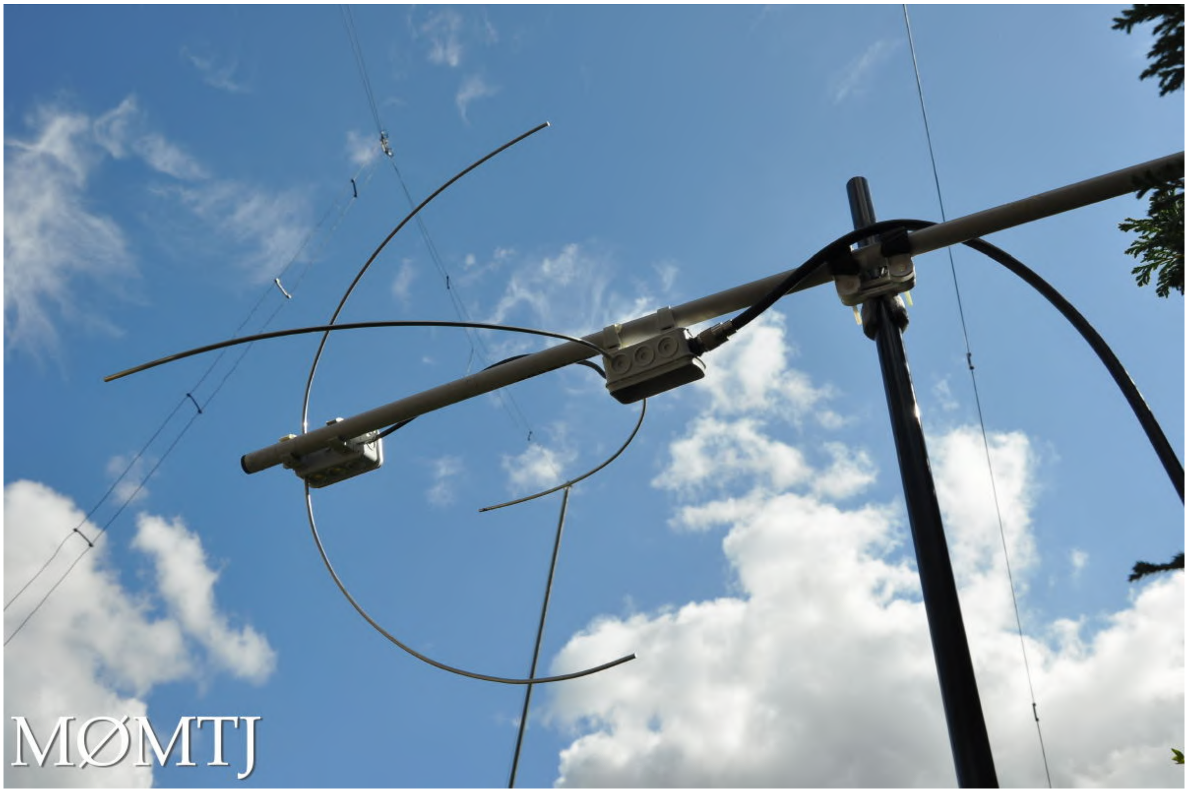

Antenna under initial test and adjustment at approximately 2 metres a.g.l.
The coaxial cable is low loss Westflex 103
(My HF Doublet antenna, fed with open wire balanced line can be seen in the background on the left)

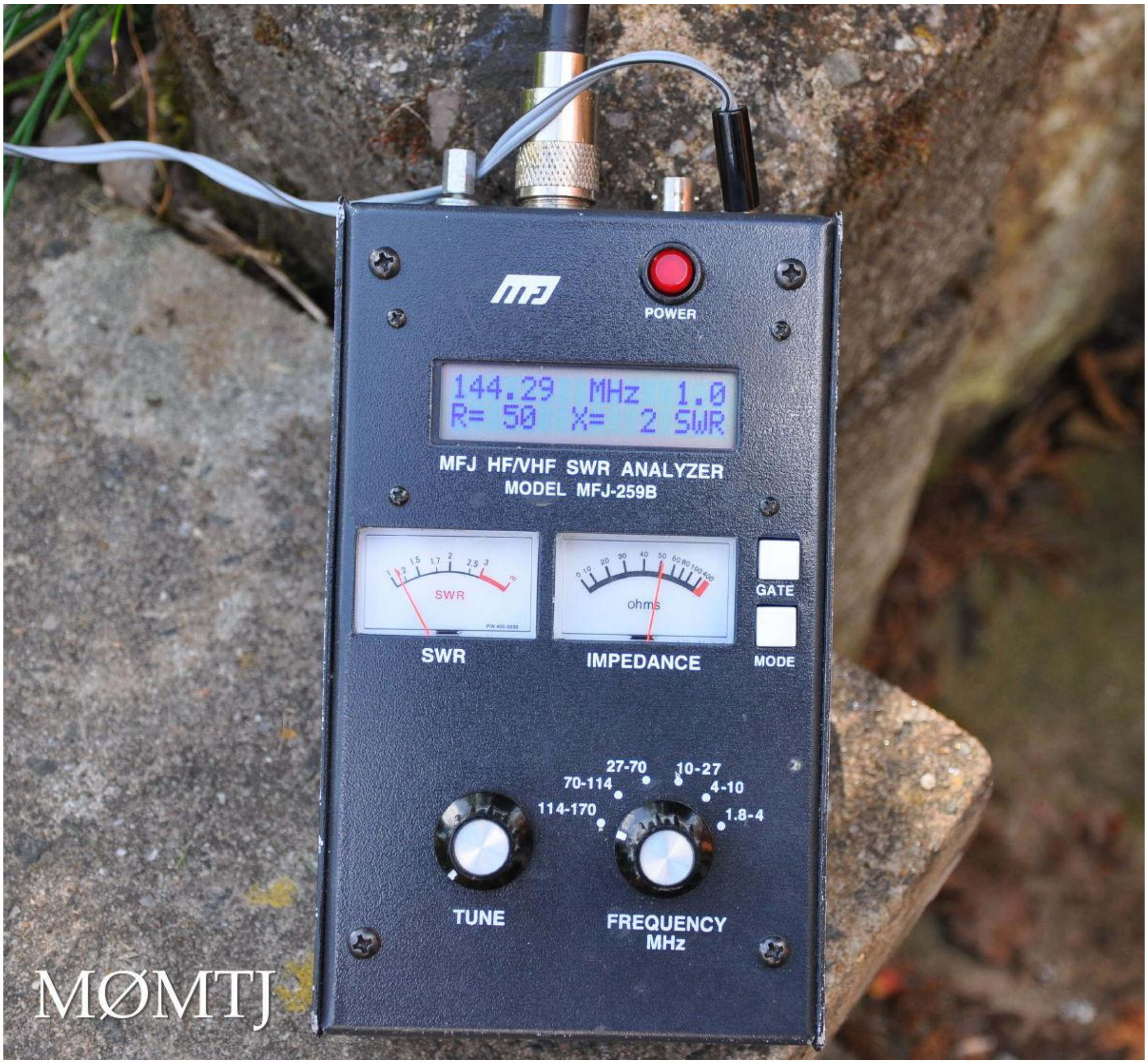
MFJ
POWER
144.29 MHz 1.0
R= 50 X= 2 SWR
MFJ HF/VHF SWR ANALYZER
MODEL MFJ-259B
SWR
IMPEDANCE
GATE
MODE
27-70
10-27
70-114
4-10
114-170
1.8-4
TUNE
FREQUENCY
MHz
MØMTJ

After adjustment to element lengths and dipole positions along the boom arm, the antenna shows a good match.

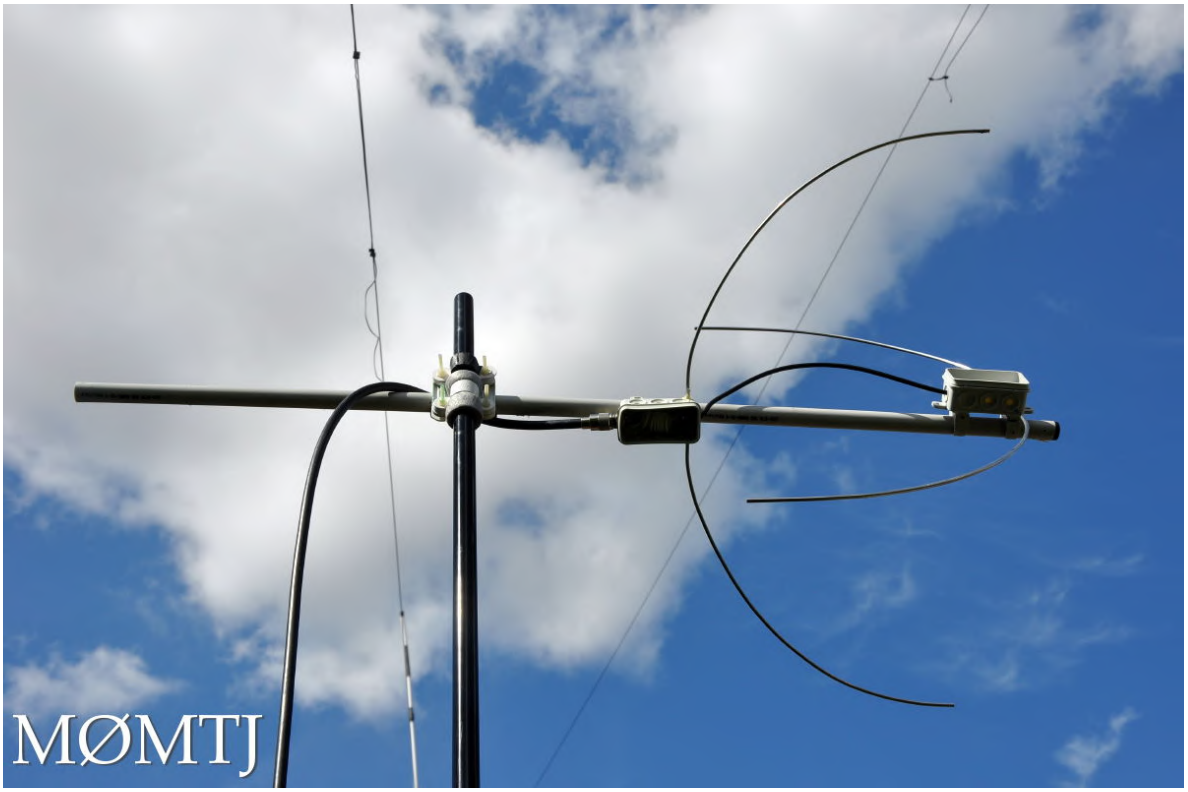

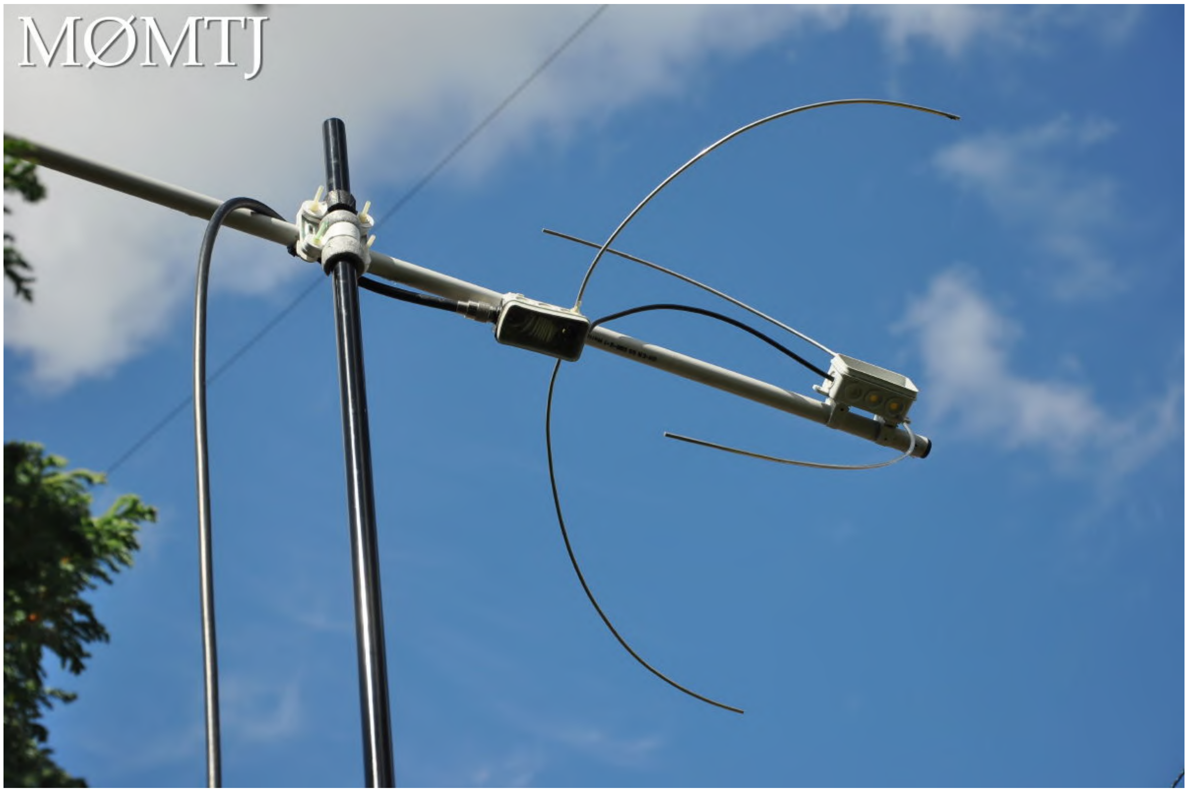
MØMTJ

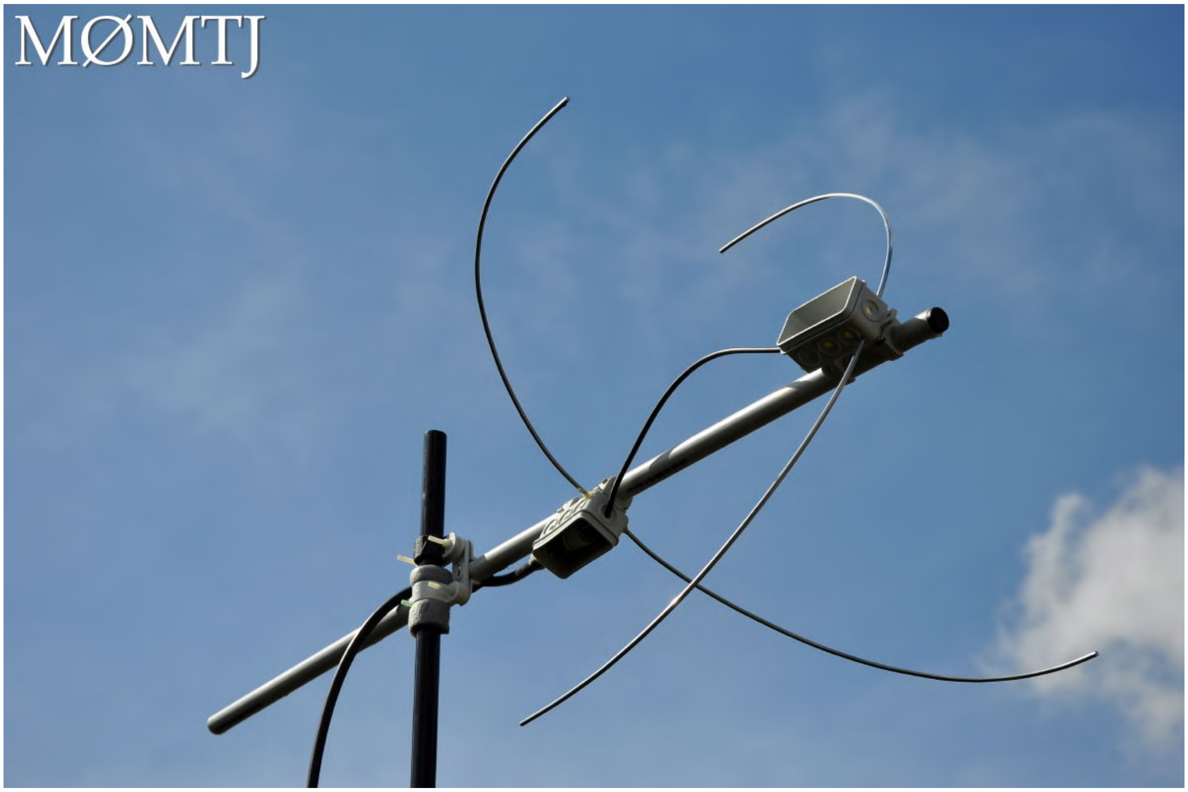
MØMTJ

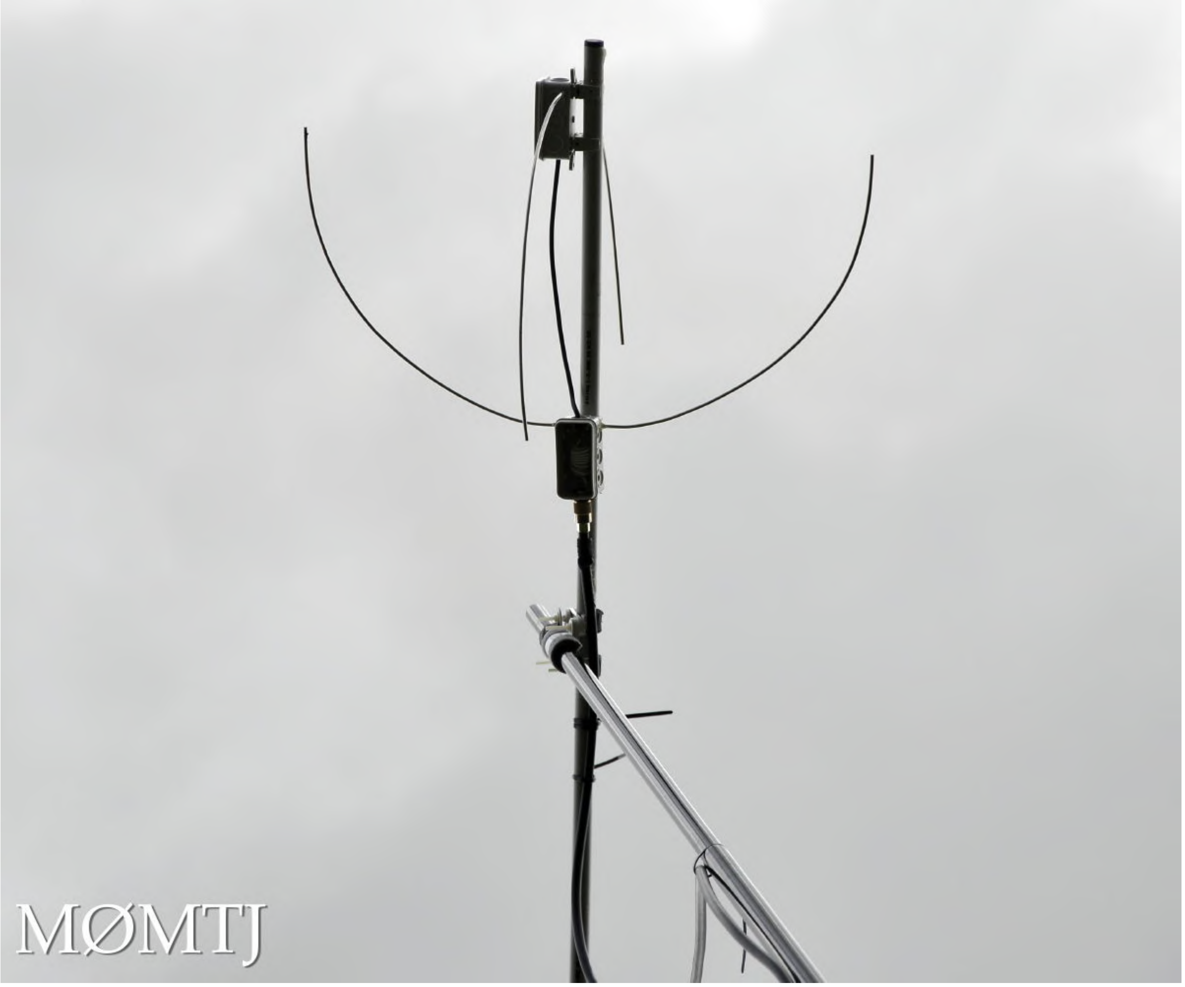
MØMTJ

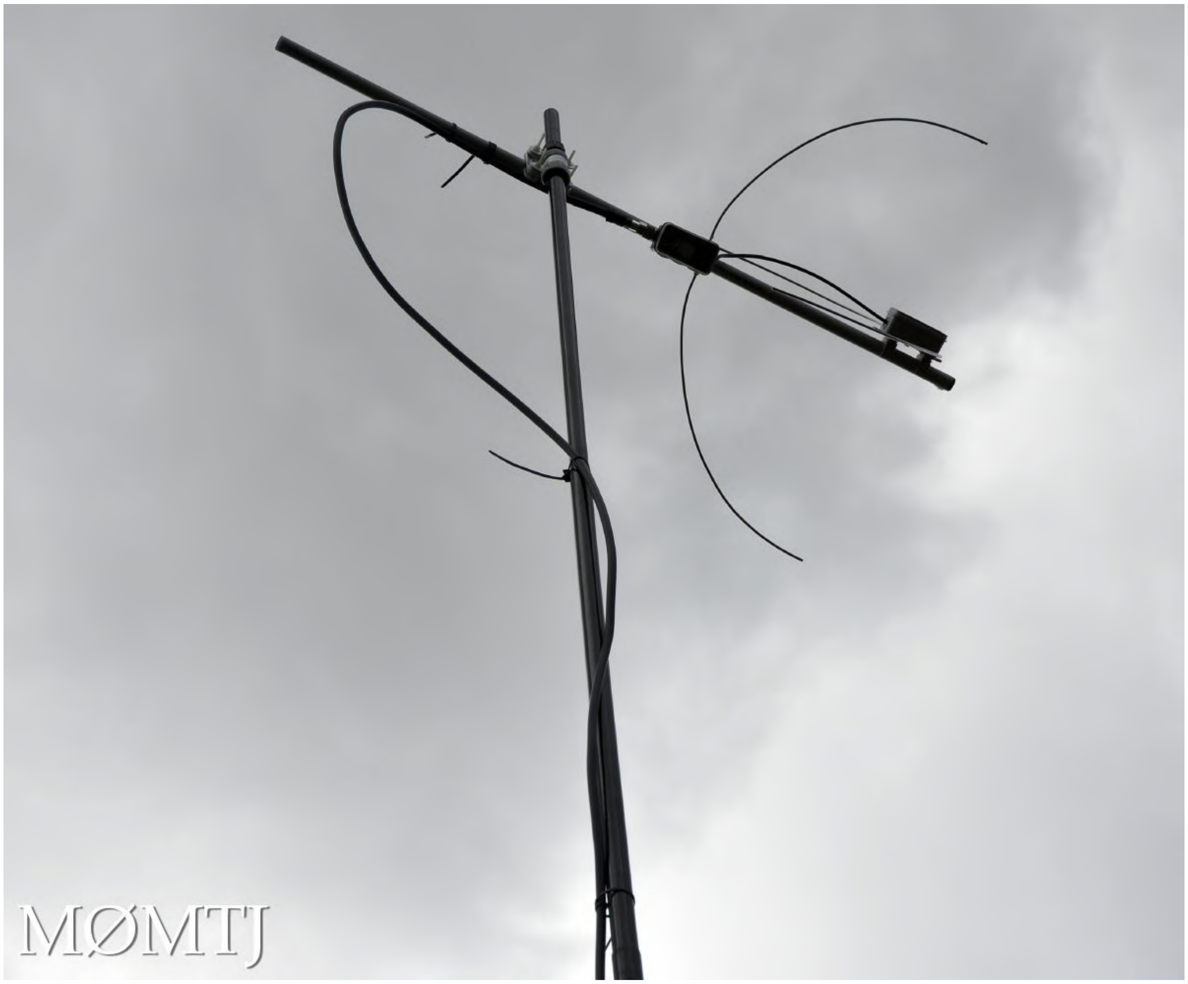
MØMTJ

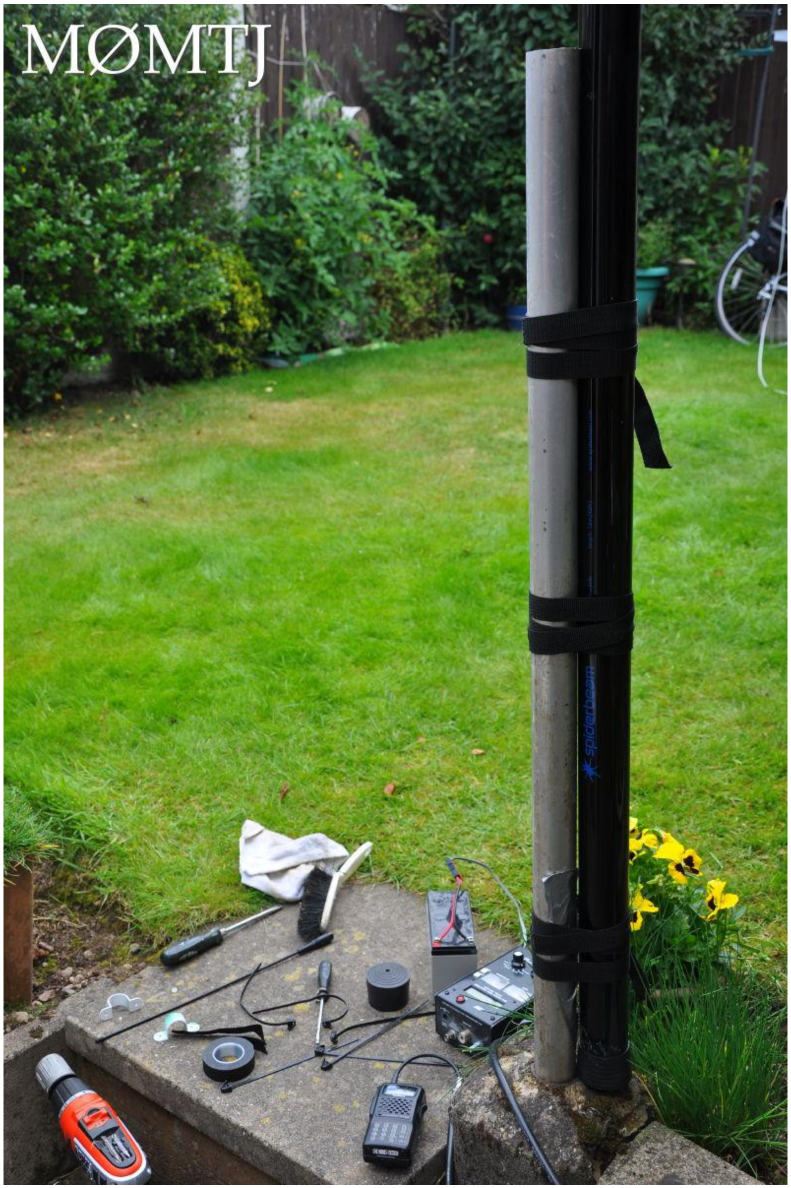

The Test 'Rig' comprised of a heavy duty Spiderbeam 12m fibreglass telescopic pole with the antenna mounted approximately 8 meters a.g.l.

If you have any comments, please do email them to me via the contact page here.

Antennas 1 | Antennas 2 | Antennas 3 | Antennas 4 | Antennas 5 | Antennas 6 | Antennas 7

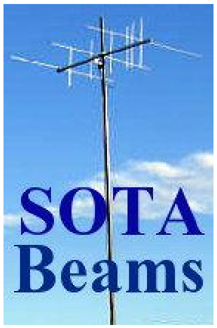
SOTA Beams
Lightweight 2 metre &
70cms Yagis, Dipole,

G Whip
Antenna
Products

G-Whip Antenna Products
Get on the air with
TOP QUALITY
BRITISH
G-WHIP
ANTENNAS
End Fed Zepps and
many more designs
and products

Accessories & Poles
www.sotabeams.co.uk

PLA-PLT Latest News
Keep up-to-date on this important issue

Radio Interference?:
PLT & PLA: Radio Interference
British Telecom, BT Home Plug
P C World, Powerline Adapters
YouView, Belkin Linksys
Netgear Devolo, BT Vision - BPL
Radio Interference
PLT Page 1

Antennas 1 | Antennas 2 | Antennas 3 | Antennas 4 | Antennas 5 | Antennas 6

Top of page ^

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits | Useful Information

Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites | RSGB | QSL | The Amateur Radio Mini Site Map

WSPR Weak Signal Propagation Reporter

News, Developments, Events, Photographs and other 'Bits 'n' Bobs'

MDS975 Home

Feedback

MAIL

Mike Smith - MDS975.co.uk © 2003 - 2016

M0MTJ

Amateur Radio; Ham Radio; Radio; Ground Plane Vertical; Sloper; HF Antenna; Dipole;
Doublet; Antenna - The Best Multi-Band Antenna for Easiest Construction, Quickest Deployment and Lowest Cost

9:1 Magnetic Longwire Balun / Unun

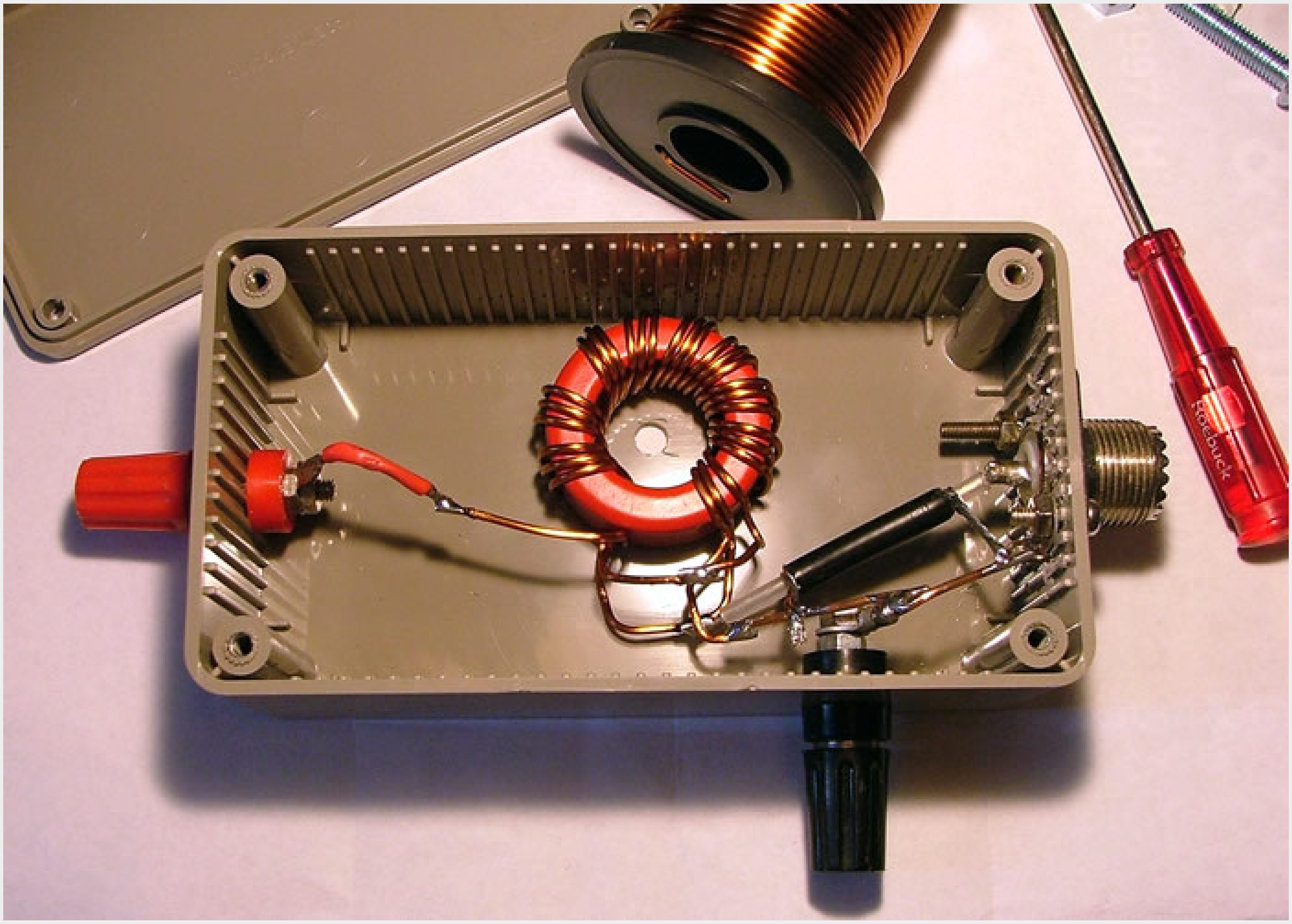

Completed 9:1 UnUn

These are sometimes known as a *Magnetic Longwire Balun*. Its really an impedance transformer (9:1) to feed a high impedance, end fed (unbalanced) random wire which is likely to be a few hundred ohms, and transform it into something closer to a 50Ω (unbalanced) coaxial input, hence UnUn. It can however also be used as a balun by not grounding the other half of the balanced output. The toroid is a T130-2 Iron Powder core, with 3 x 9 turns of 18SWG enammeled copper wire, and the connections can be seen below.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

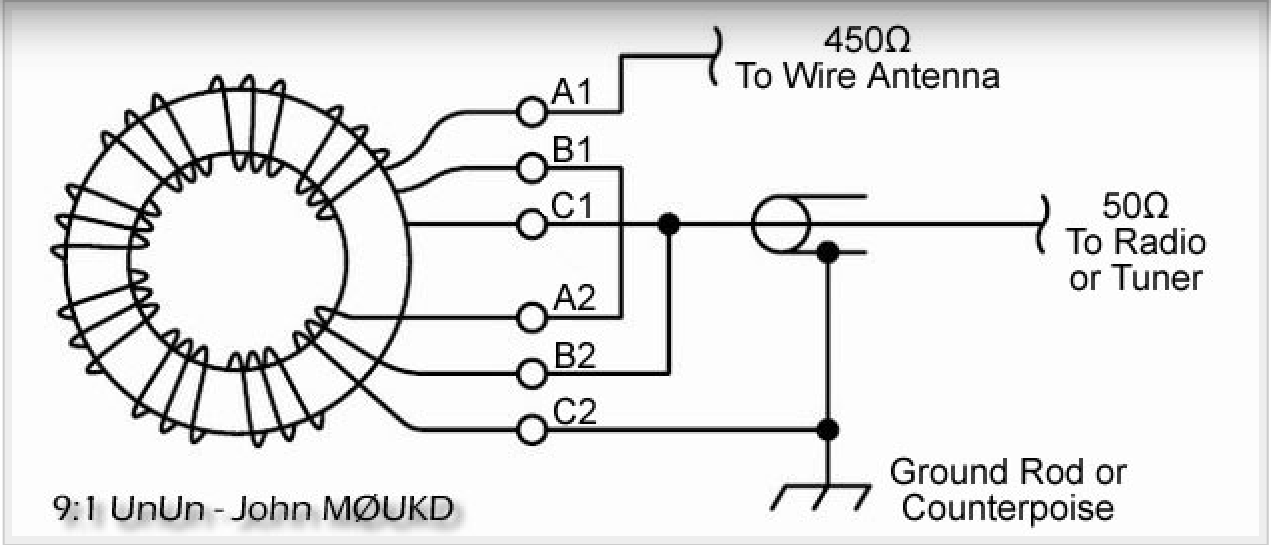
450Ω
To Wire Antenna
A1
B1
C1
A2
B2
C2
50Ω
To Radio
or Tuner
Ground Rod or
Counterpoise
9:1 UnUn - John M0UKD

The triple winding connections. 8 turns is shown here, but more turns is preferred. 10 would be ideal.

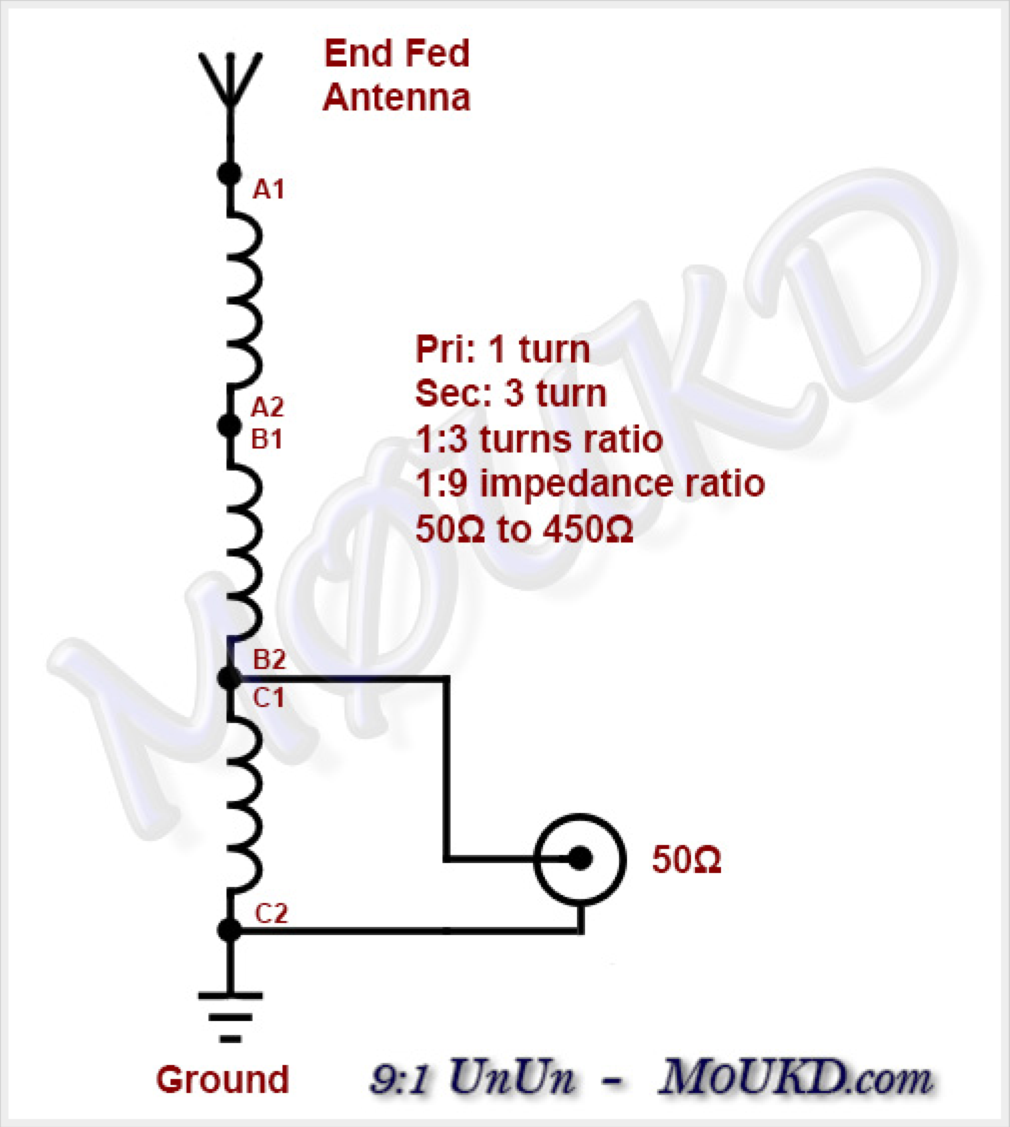
End Fed
Antenna
A1
A2
B1
Pri: 1 turn
Sec: 3 turn
1:3 turns ratio
1:9 impedance ratio
50Ω to 450Ω
B2
C1
50Ω
C2
Ground
9:1 UnUn - M0UKD.com
M0UKD

9:1 UnUn Schematic

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

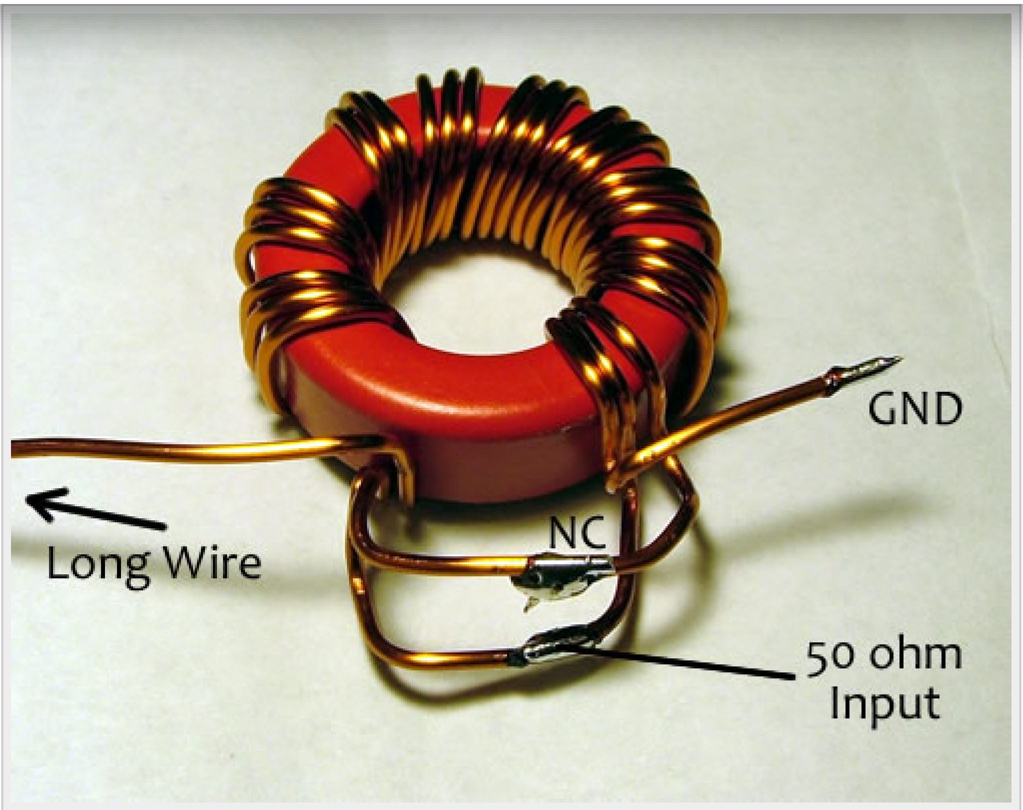
GND
Long Wire
NC
50 ohm Input

Completed windings

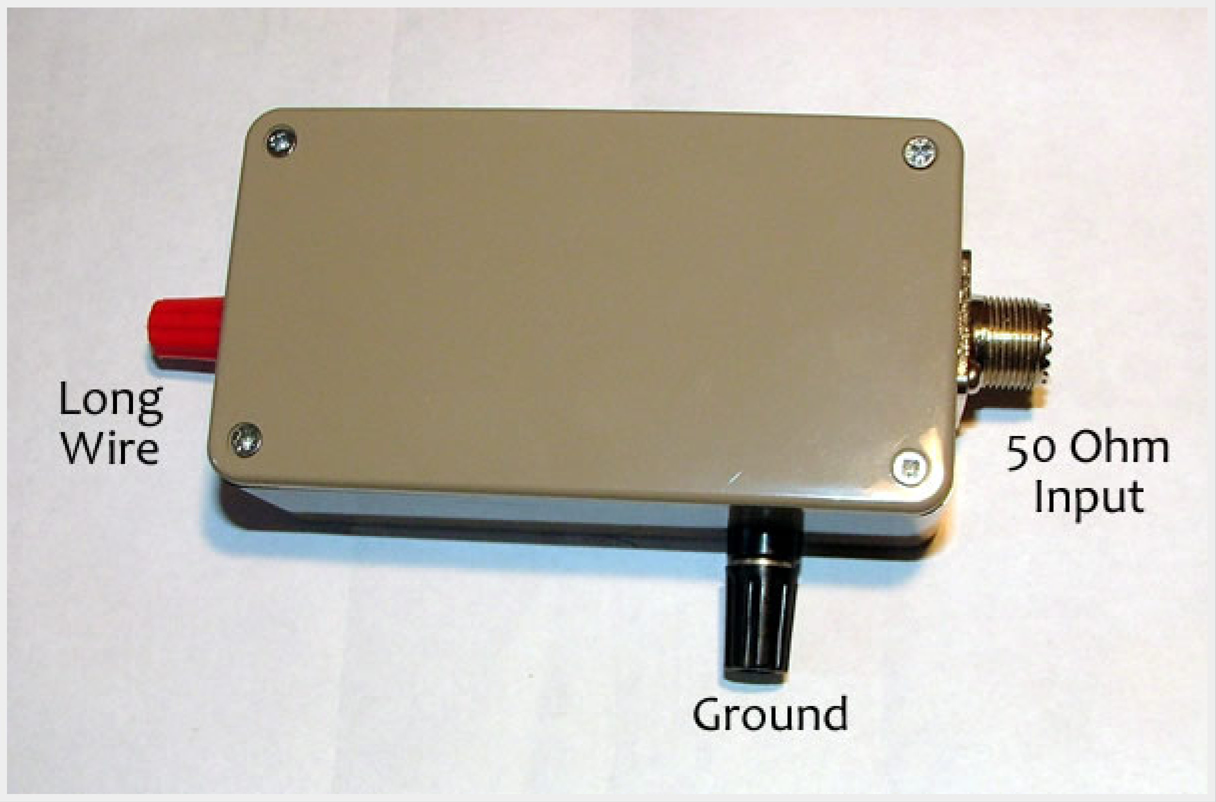
Long Wire
50 Ohm Input
Ground

The completed unit.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

1:1 Voltage Balun for HF wire dipoles

Heres a neat 1:1 50 ohm balun for use on HF horizontal wire dipoles. It uses an AM radio ferrite rod, with 3×14 turns of wire. 10-14 turns should be good for 2-30MHz. I used 18SWG enammeled copper wire. It all fits in a small project box. Great for setting up portable. For 20 metres, I started with 6 metres of wire per element, and trimmed each down until it has minimum S.W.R. I managed to get a 1:1.1 S.W.R. on my desired frequency. Bandwidth from the resonant frequency is about 200KHz either way, any more needs an ATU. The wire and ferrite core I had already, the box, terminal plugs and SO-239 came to about £6. The balun can also be built on a ferrite toroid such as an FT140-43 or FT240-43.

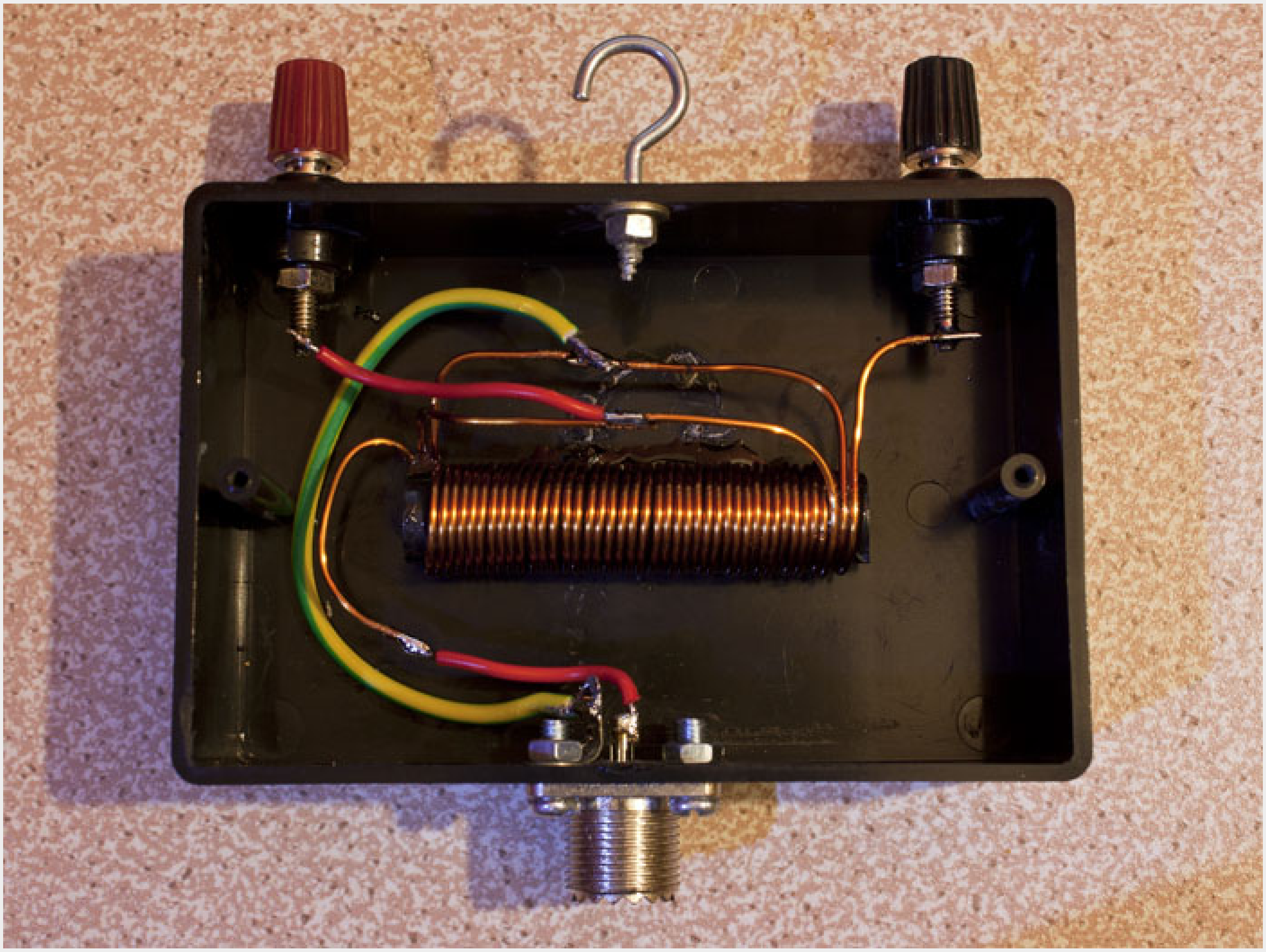

1:1 ferrite rod voltage balun

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Accept Read More

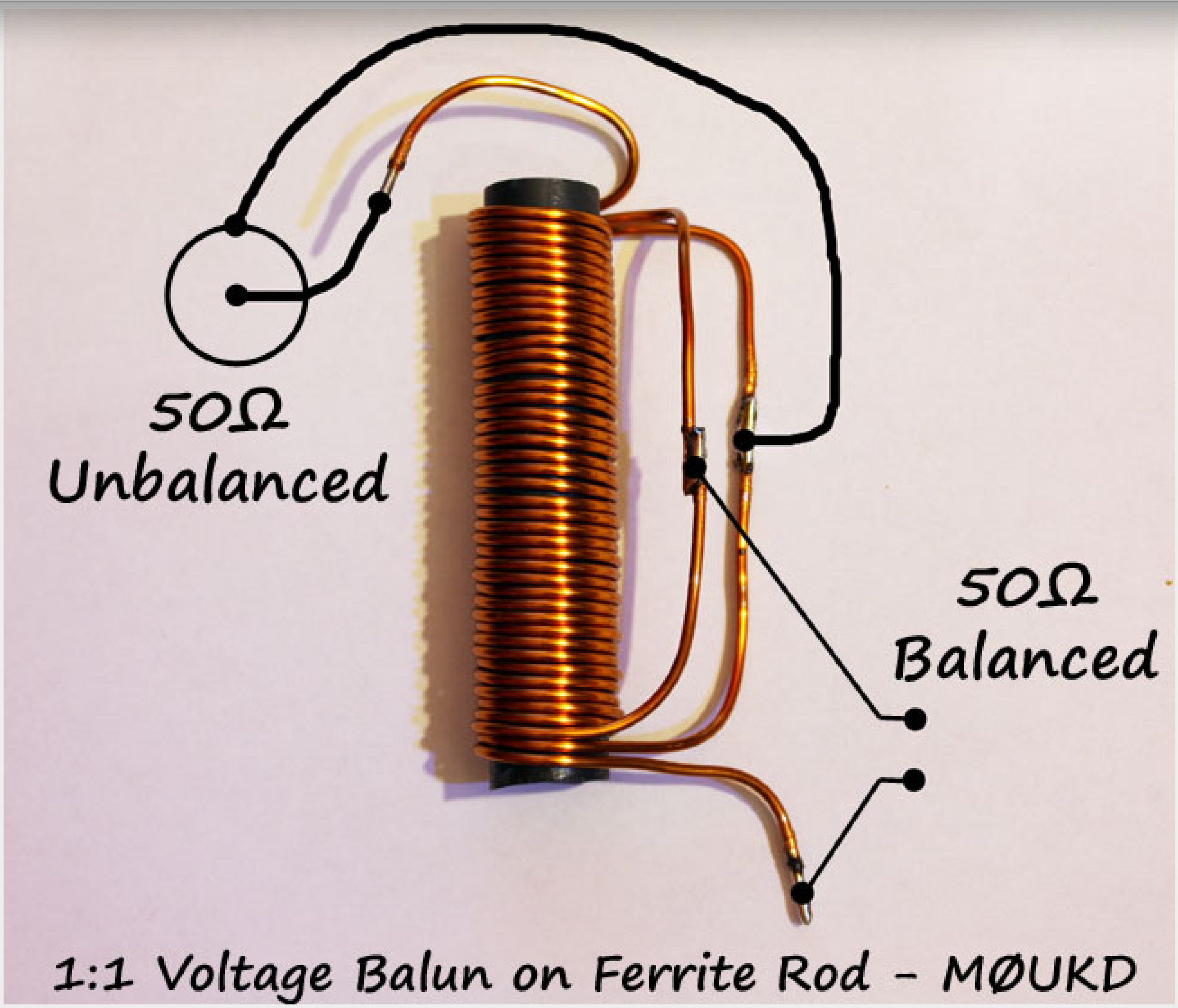
50Ω
Unbalanced
50Ω
Balanced
1:1 Voltage Balun on Ferrite Rod - MØUKD

Bandwidth may be increased by more spacing between each trifilar turn. I didn't have room on this tiny ferrite rod, but it performs OK.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

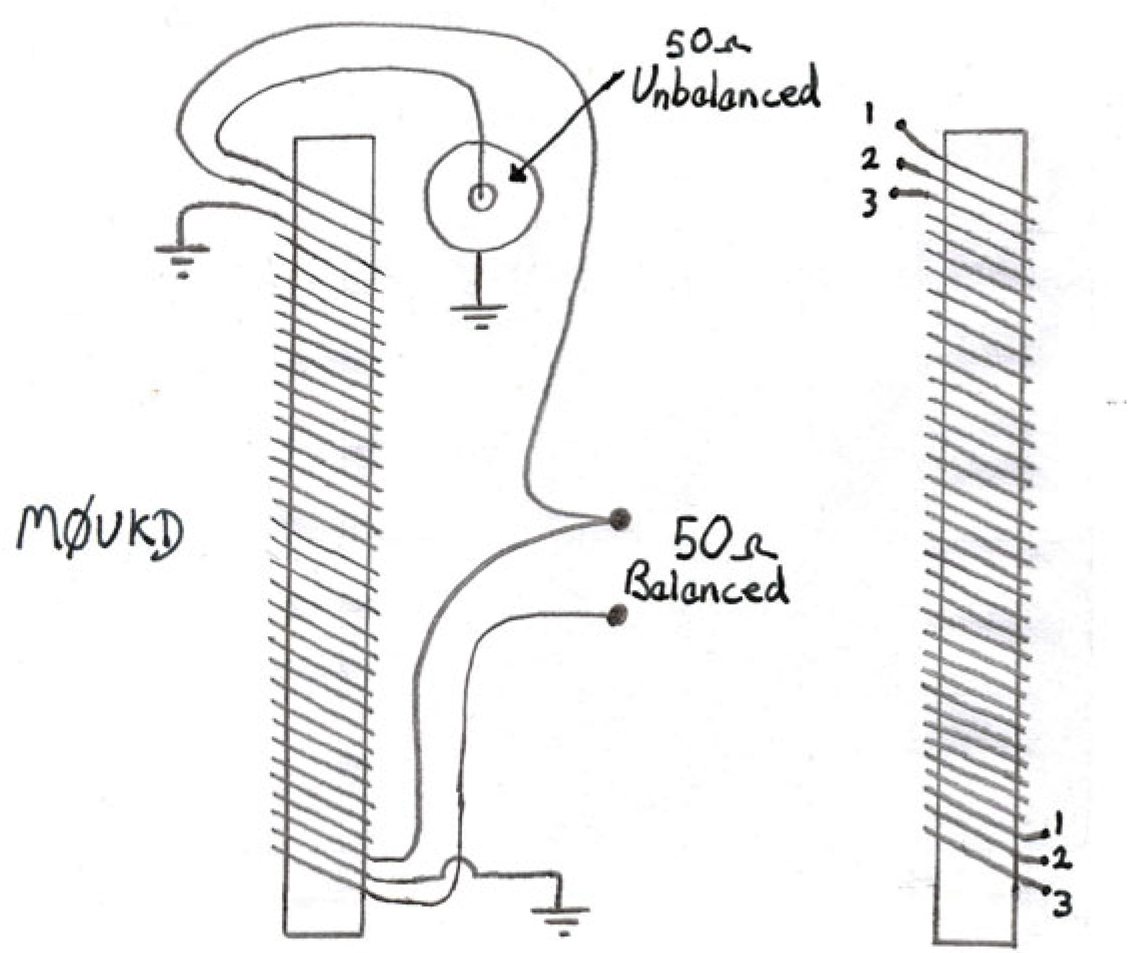
1:1 Voltage balun ("Ruthroff")

50Ω
Unbalanced
M0UKD
50Ω
Balanced
1
2
3
1
2
3

10 - 15 trifilar turns on ferrite rod.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

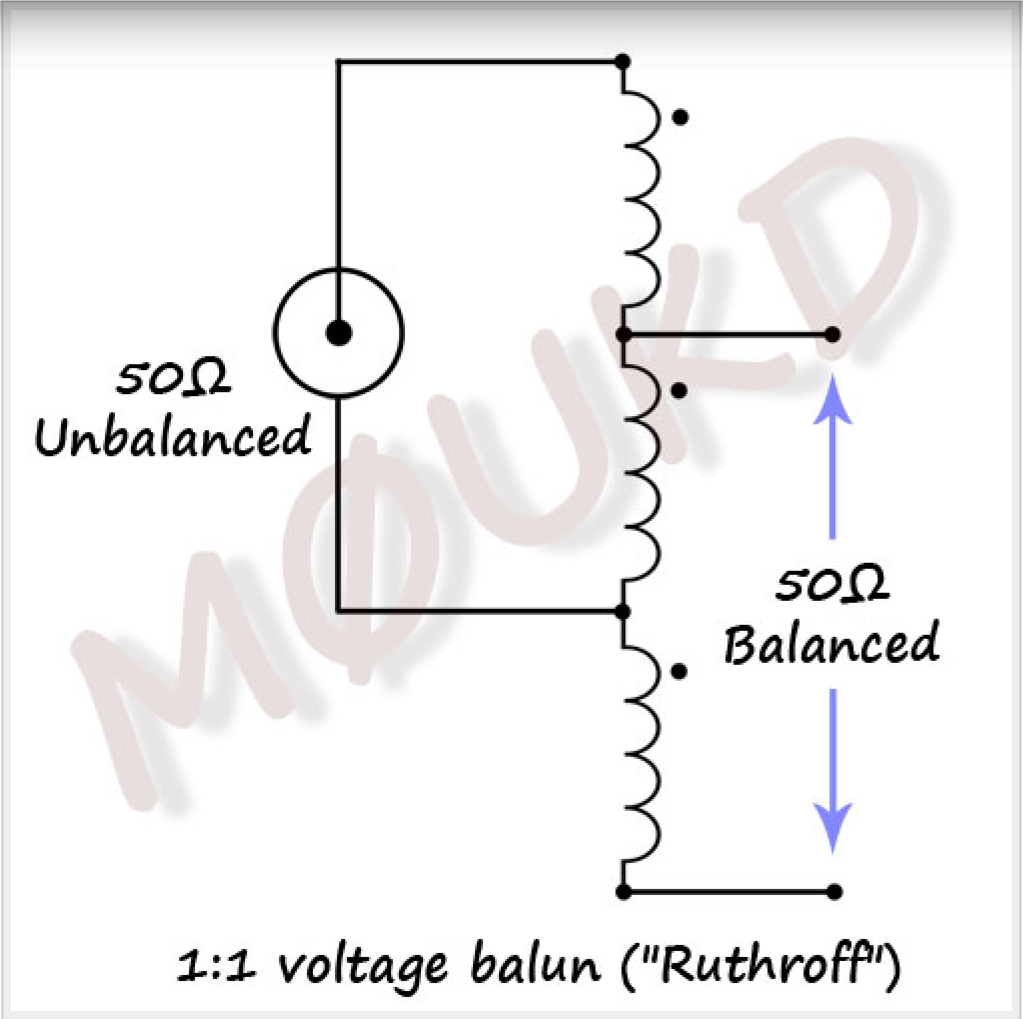
50Ω
Unbalanced
50Ω
Balanced
1:1 voltage balun ("Ruthroff")
M0UKD

Schematic Diagram

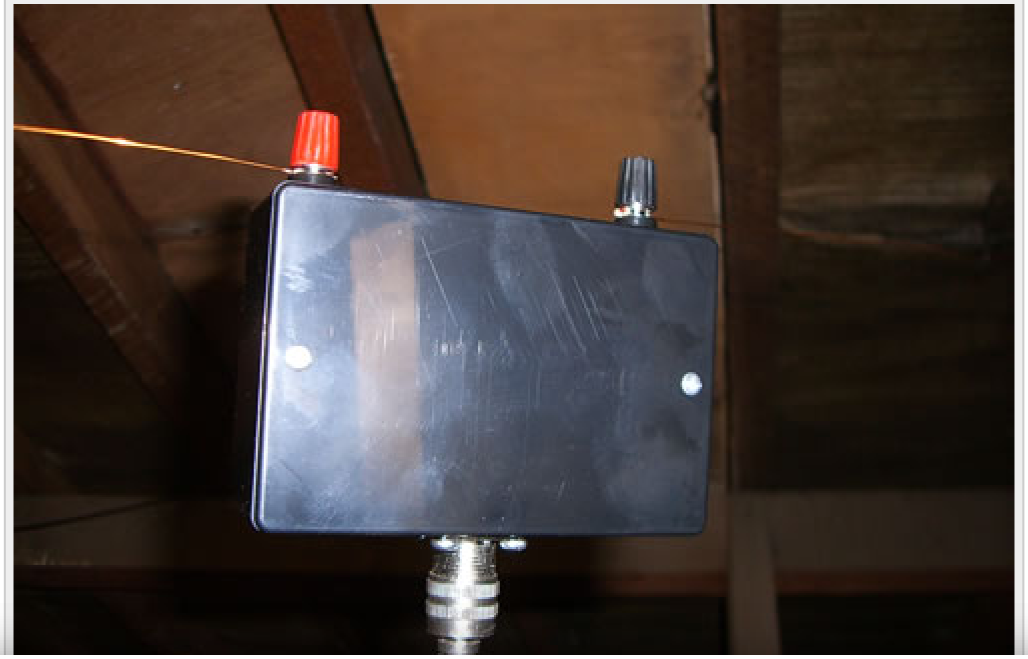

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

End Fed Half Wave Antenna Coupler (EFHW)

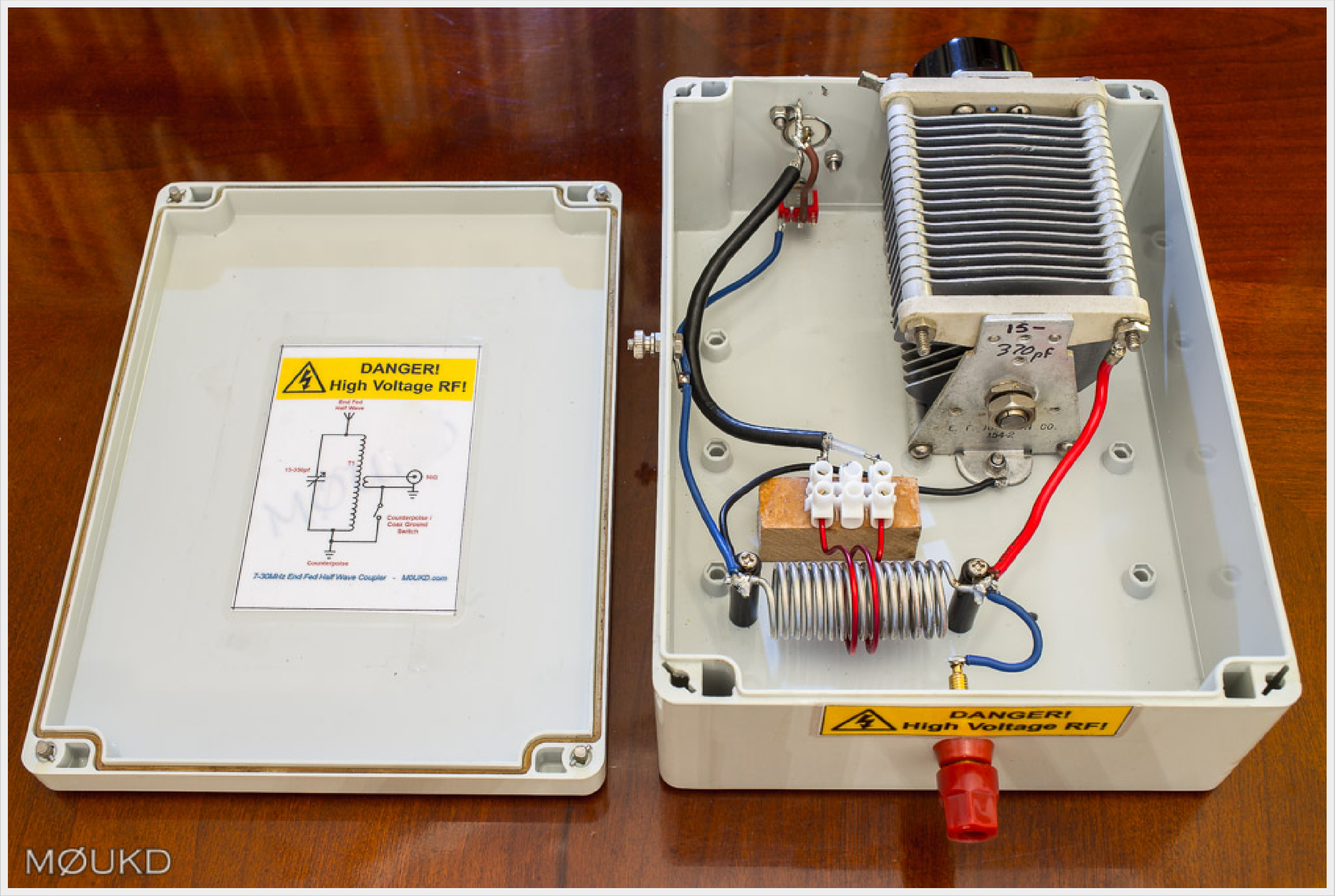

The finished End Fed Half Wave antenna coupler.

Centre fed half wave dipoles make great, simple and effective antennas for the HF bands. Sometimes however, the centre feed is not ideal, for example when you want to use it as a vertical. Being able to feed the dipole from one end gives you more options on how to erect an antenna and makes portable operation easier. A vertical, a sloper, a piece of wire hung in a hedge are all good examples. A ground mounted half wave vertical has a peak radiation angle of 20°, so it makes a good choice for DX.

I have been experimenting with feeding end fed half wave antennas matched by a parallel tuned circuit coupler. This article will explain my findings and reasons for constructing it the way I did.

A dipole can be fed anywhere along its length. A centre feed gives around 70Ω. A 'Windom' type off centre fed dipole (fed 38% along its length) provides around a 200Ω feed. The feed impedance at the very end of a half wave is thousands of ohms, usually somewhere between 2000Ω to 5000Ω, which we need to match to our 50Ω transceiver. The problem with end feeding a half wave is also its advantage. The high impedance means that the feedpoint has a very high voltage but low current, therefore very little ground is required. A very small counterpoise should be adequate from 7 to 30MHz, or you can even use the coax and transceiver

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Accept Read More

Below is the schematic of the final build.

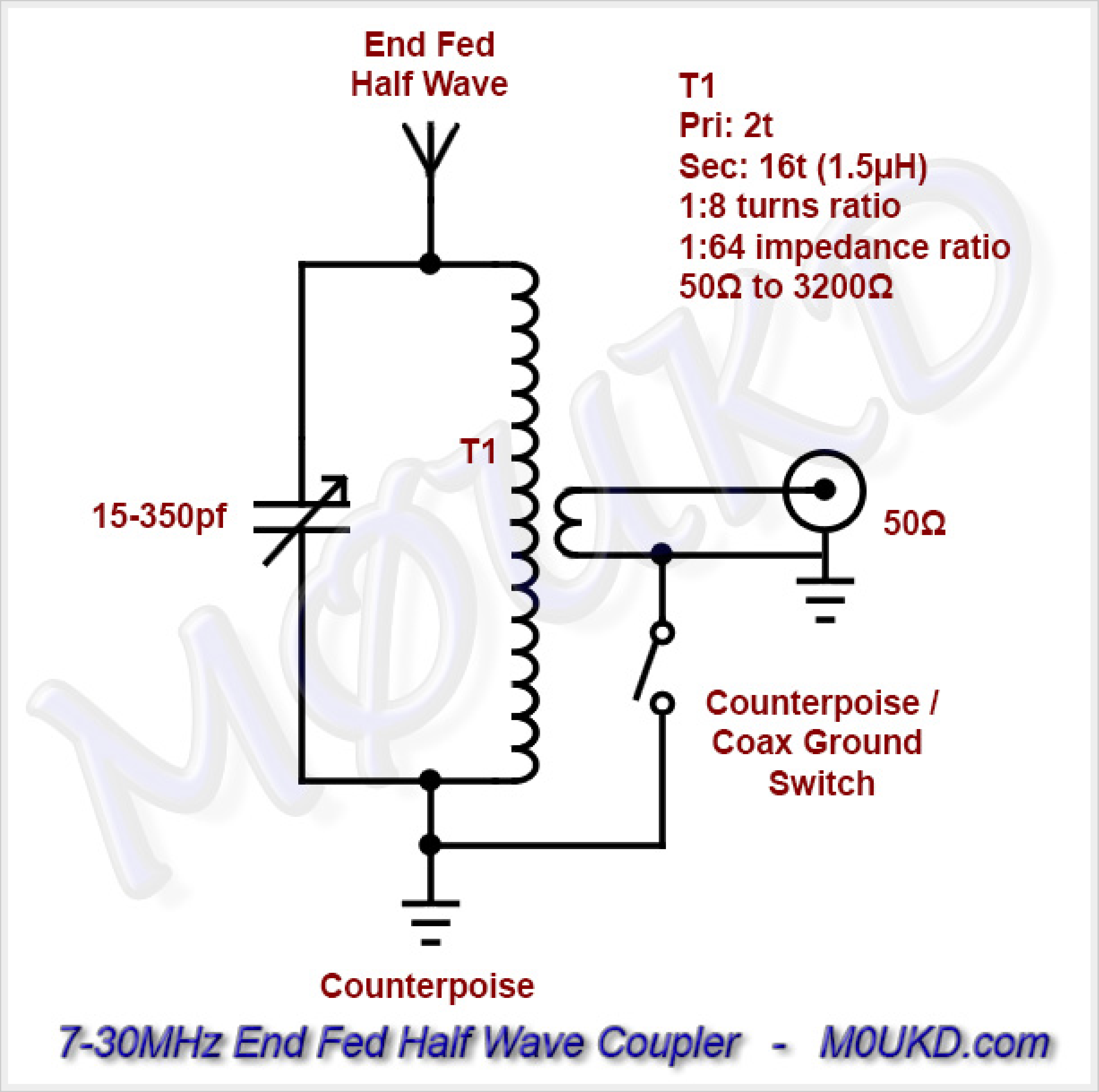
End Fed
Half Wave
T1
Pri: 2t
Sec: 16t (1.5μH)
1:8 turns ratio
1:64 impedance ratio
50Ω to 3200Ω
T1
15-350pf
50Ω
Counterpoise /
Coax Ground
Switch
Counterpoise
7-30MHz End Fed Half Wave Coupler - M0UKD.com
MØUKD

End Fed Half Wave Antenna Coupler Schematic – 7-30MHz

For calculating the length of a half wave in metres, I use 141÷f(MHz) for wire elements. These dimensions were derived from modelling the antenna in EZNEC. I have found this calculation to work well, however it depends on many factors such as wire used, location etc. I have made a javascript calculator below for simplicity with a 15m half wave ready to go!

Enter the frequency: 21.25 MHz

Calculate

Half Wave antenna is: Metres

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

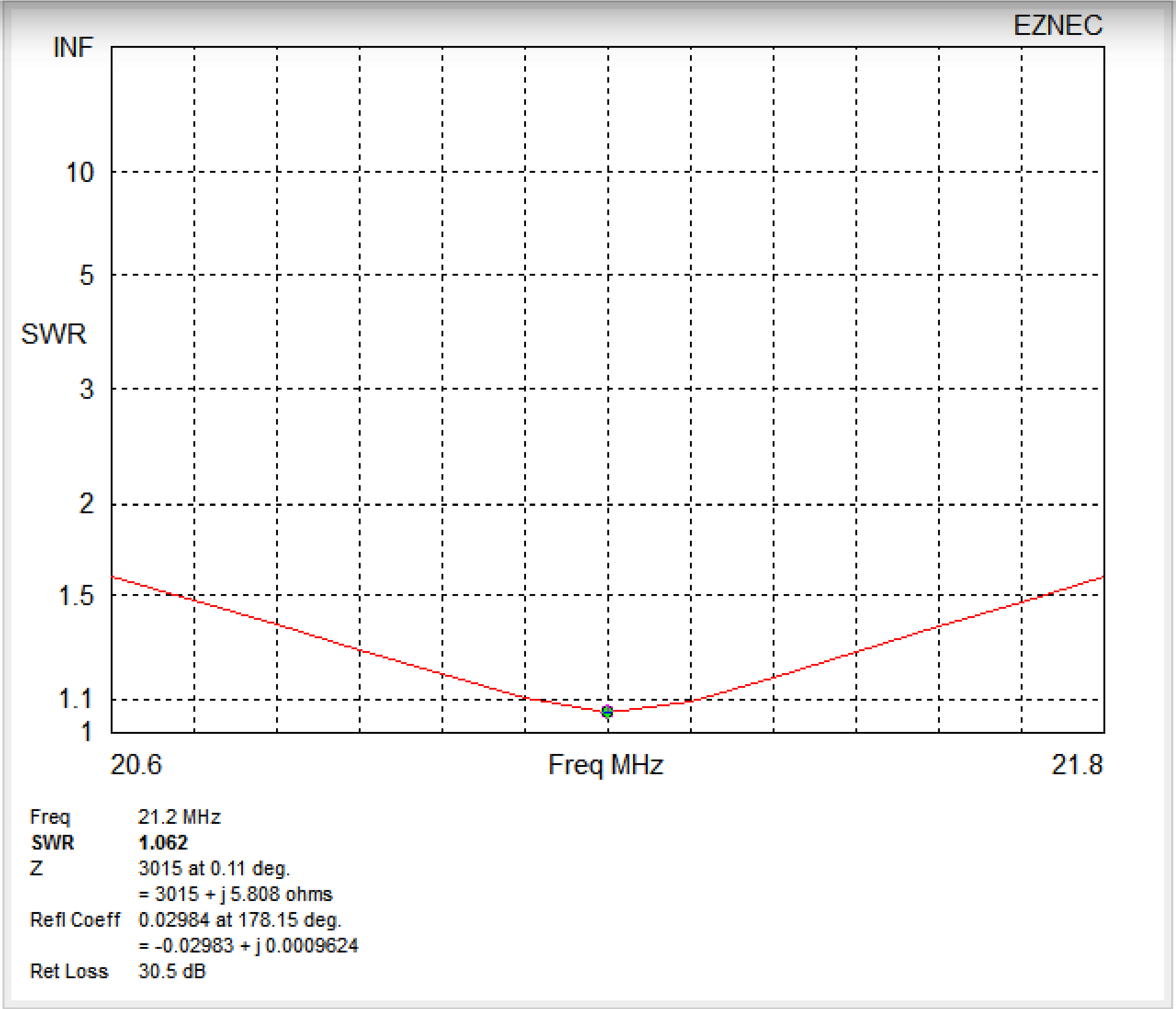

```
Freq        21.2 MHz
SWR         1.062
Z           3015 at 0.11 deg.
            = 3015 + j 5.808 ohms
Refl Coeff  0.02984 at 178.15 deg.
            = -0.02983 + j 0.0009624
Ret Loss    30.5 dB
```

End Fed Half Wave (6.65m @ 21.2MHz, ground mounted) SWR – Actual bandwidth on the 50Ω side of the coupler will be much narrower, due to the tuned circuit we are using. This is showing the feed point SWR at 3200Ω

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

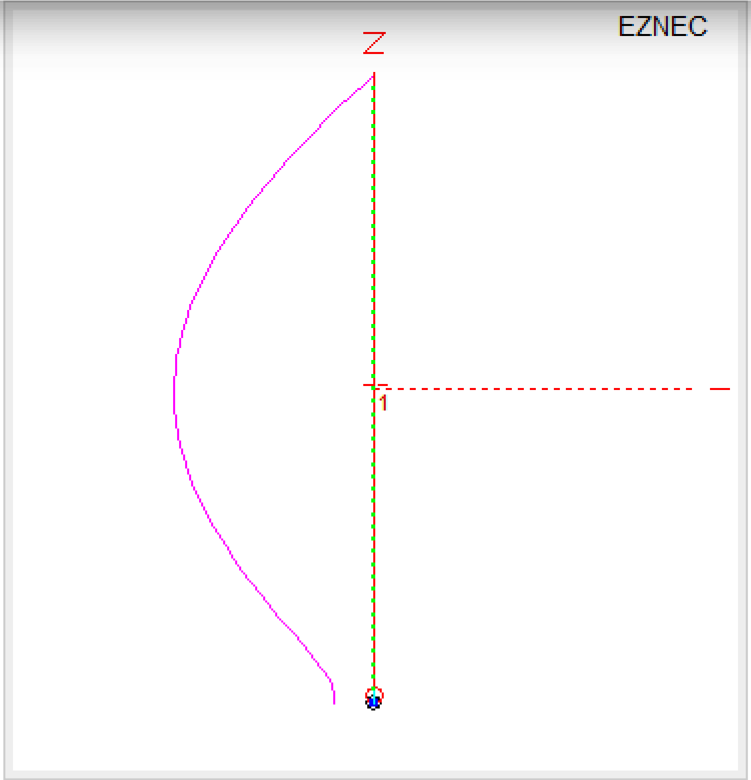

Current along the half wave. Notice how the feedpoint is not quite as high impedance as the top of the antenna.

First off, construction will be a juggle between available components and the desired band coverage. I had a capacitor that I wanted to use, which is a Johnson 154-2 air variable capacitor, with a range of 15-353pf. I would like to cover 7 to 30MHz with my coupler if possibe, so the next thing I needed to design was an inductor that would resonate in a parallel LC circuit at just above 30MHz when the capacitor is at minimum. Using this **Resonant LC Calculator**, I worked out that an inductance of 1.5µH will resonate at about 33MHz when the capacitor is at 15pf and 6.9MHz when the capacitor is at its maximum of 353pf. Sounds perfect to me! If you have a smaller value capacitor, for example 200pf maximum, it should still cover 10MHZ to 30MHz. A switch could be added to add a further capacitance of 150-200pf to include 7MHz operation.

So, lets build the secondary transformer inductor first. I wanted to transform 50Ω to around 3000Ω. This would require a 1:8 turns ratio. Impedance transformation is calculated by squaring the secondary turns ratio (note ratio, not actual turns) 8^2 is 64. 50×64=3200Ω (our input impedance x 8^2). I looked at using a **T200-6** iron powder toroid, but the problem was 1.5µH required *only* 12 turns. As I wanted a 1:8 turns ratio (for a 1:64 impedance ratio), I needed a primary of 1.33 turns, which was going to be impossible. I could have made it 16 turns and a 2 turn primary, but then the inductance would be too large for 30MHz.

So, I decided to use an air wound transformer. This way, I can build it with a 16 turn secondary, *and* at the inductance I wanted by altering the diameter and/or length of the coil. Using this **Air Inductance Calculator**, I worked out that a 19mm diameter, 52mm long, 16 turn inductor should give an inductance of 1.5uH, so this is what I built

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

drywipe marker
orizontally

Winding the secondary inductor

Once it was wound, I tested it on an LC meter, which confirmed it was around 1.5µH. You can expand or contract the inductor to fine tune the impedance to suit. The importance is that it remains 16 turns to match with the 2 turn primary. I then wound the primary coil, which had to be a slightly larger diameter so that it could fit over the secondary coil and provide inductive coupling. You can see the finished 2 turn primary and 16 turn secondary below.

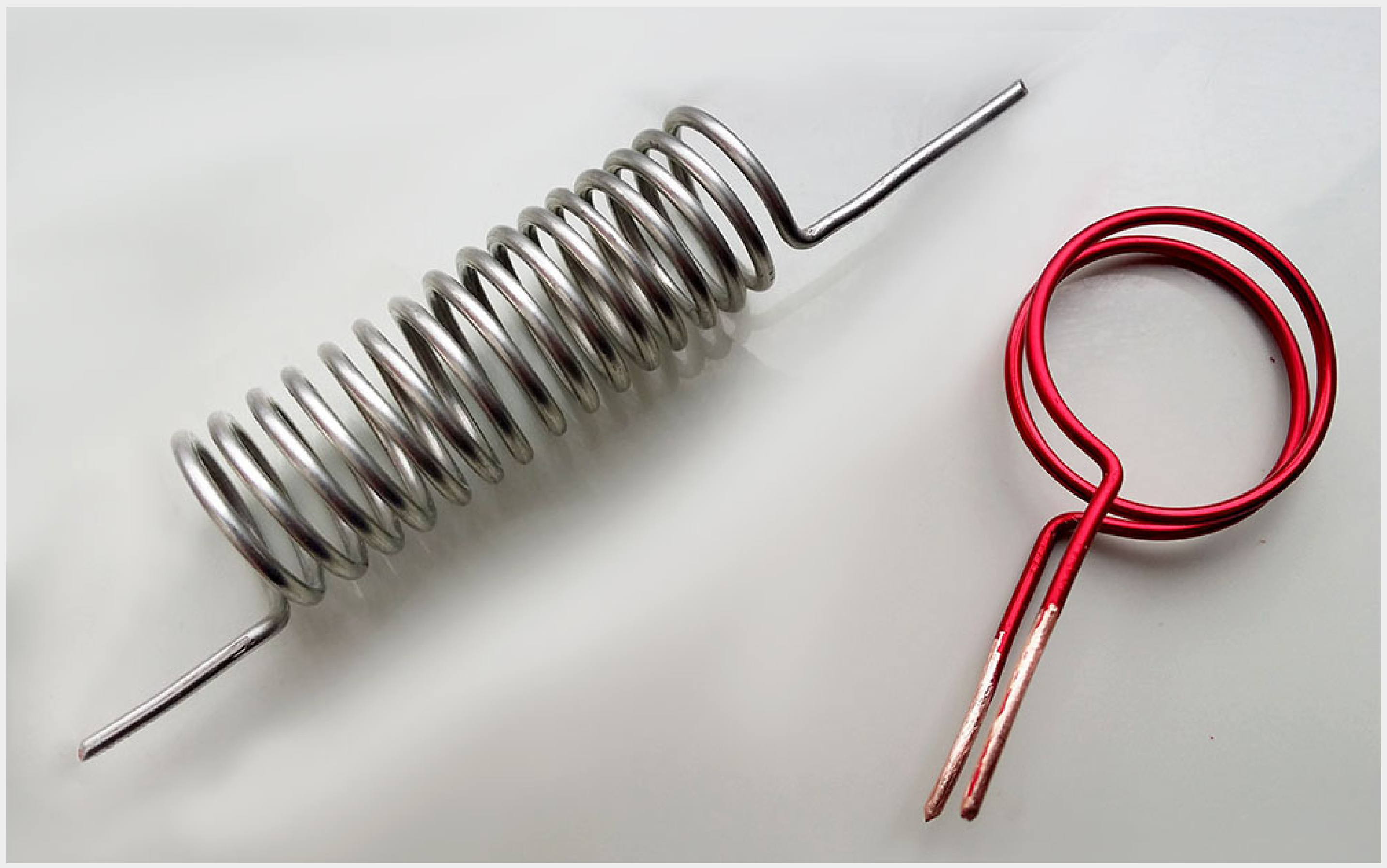

The completed inductors. They have their differences, but I will make them a happy 'couple'!

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

secondary and a terminal block glued to a small block of MDF for the primary. The coax is soldered to the primary, the terminal block is just to hold it in place. There is a counterpoise connection on the side, the antenna connection on the front and the capacitor, SO239 and counterpoise/coax ground switch on the rear. Some photos of the completed unit are below.

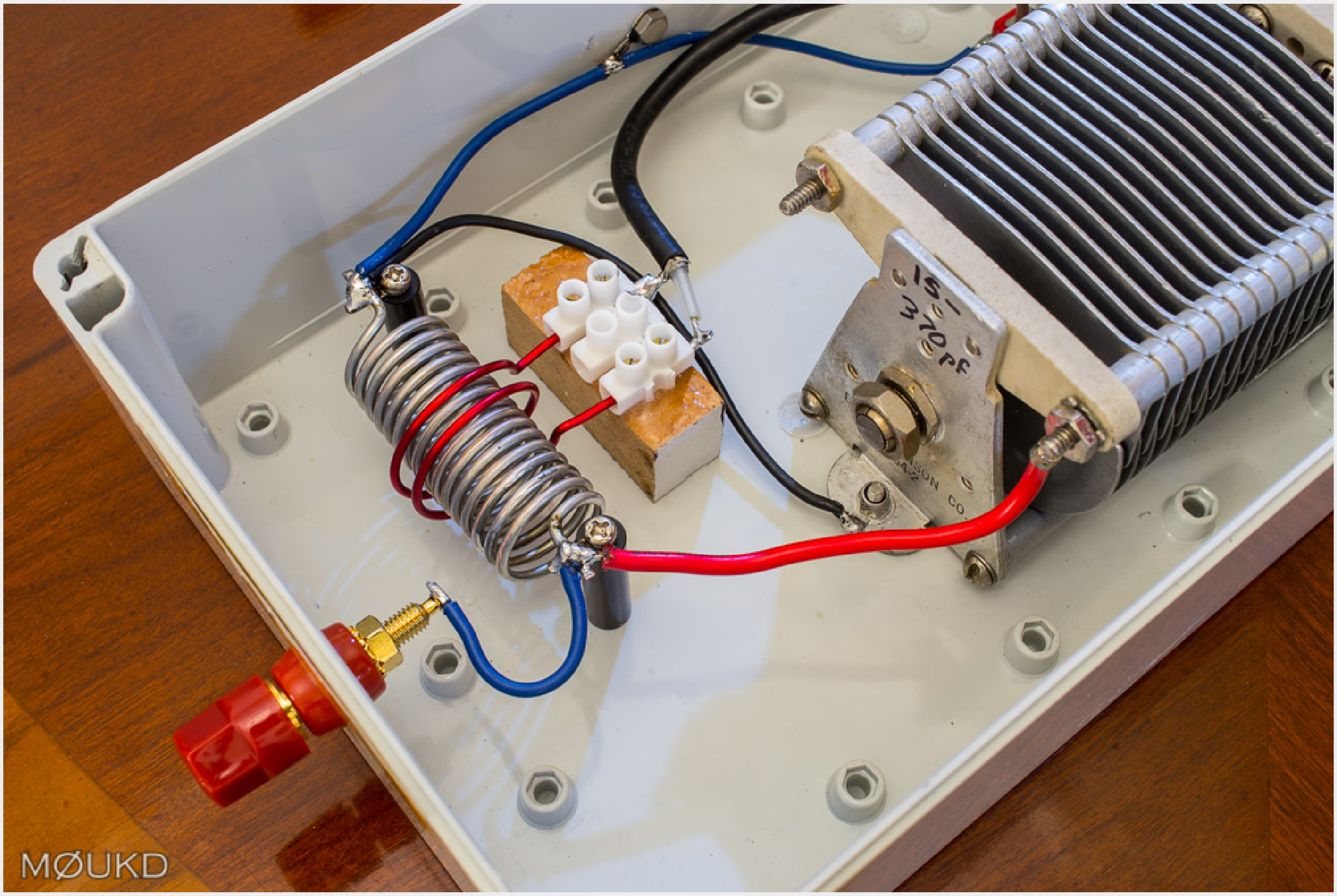

The completed End Fed Half Wave Coupler

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

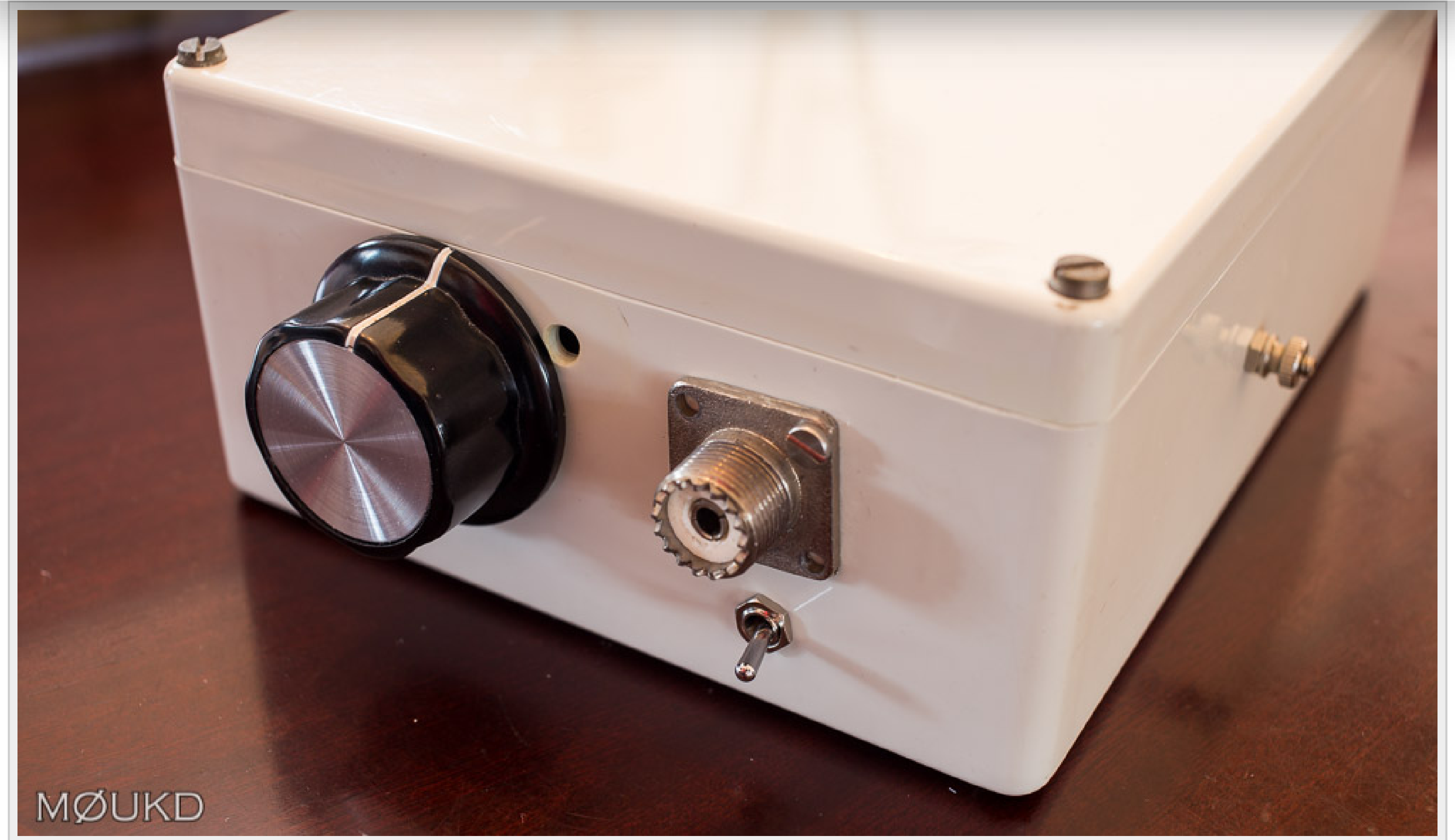

The tuning control, SO239, ground switch and on the right, the counterpoise connection.

DANGER!
High Voltage RF!
MØUKD

The antenna connection. The second hand box had a few holes in it, so I made this sticker to cover them up!

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

this is due to extra stray capacitance. I tuned the coupler to resonance into a resistor (in my case 3.2kΩ) so a 1:1 SWR was obtained on my chosen frequency of 21.25MHz, then I removed the resistor and attached the half wave antenna (6.65m) and counterpoise (66cm) from the above calculator. Without adjusting the variable capacitor, I had a 1:1 SWR. This confirmed that I had a resonant half wave and that I was feeding it at or very close to the peak voltage point and the feed impedance was around 3200Ω. I found that a small counterpoise was required. Sometimes, the stray capacitance to ground was good enough on its own, especially on the higher bands and if the coupler was laying on the ground. If its all away from ground, I get consistent results using the above setup. I have had good success also by using the coax as the return, by connecting the 'bottom' of the primary and secondary together via a switch to use the coax shield and transceiver as a counterpoise.

You may be able to adjust the capacitor to match the antenna if its not a perfect half wave, but then it will be a higher current feed and the simple ground system will be inadequate. The coupler will also not be working as it was designed and will be inefficient, generating heat. The coupler is a tuned impedance transformer, not an antenna 'tuning' unit. Stick to the above, and it should be a sure thing!

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

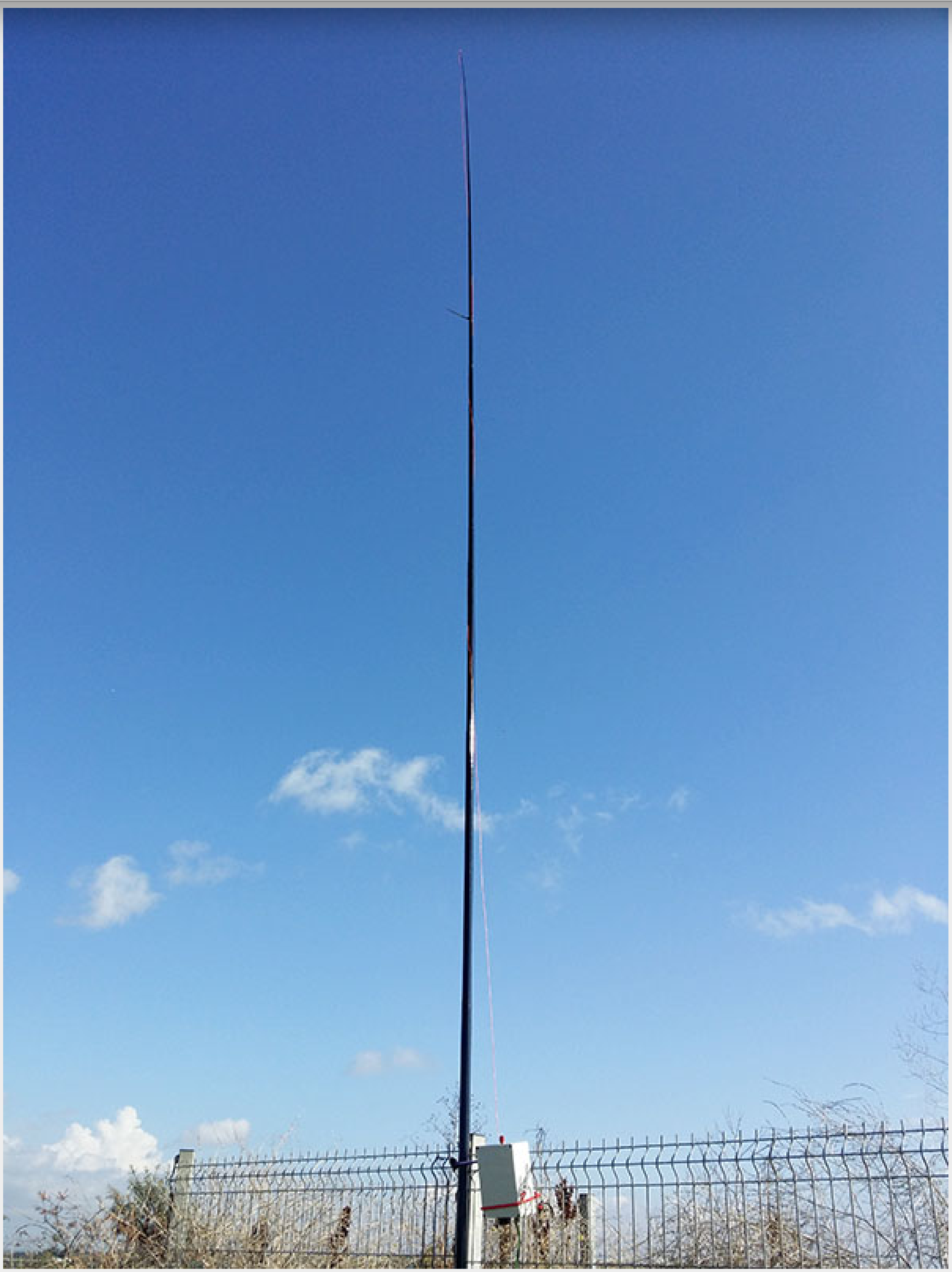

A half wave vertical for 21MHz. 6.65 metres of wire on a 9 metre fibreglass pole. The coupler is mounted 2 metres up and attached to the top of the fence. The 67cm counterpoise just hangs down in free air.

Below is a short video from my second test with the coupler, here on 17m.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

EARC
EMERGENCY
Amateur Radio Club

END FED 6–40 Meter Multiband HF Antenna

Introduction

This EARC project produces an inexpensive, multiband, end fed HF antenna matchbox that is quick and easy to setup and use. The end fed feature adds portable convenience, but does present another issue. The problem with an end fed half wave antenna is that the antenna presents high impedance, creating a significant miss match with the usual transceiver impedance of 50 ohms. This miss match is significantly greater than typical tuners can accommodate without a matching transformer.

This project creates a trifilar wound, 9:1 UNUN (unbalance to unbalance) toroid matching transformer that will match the high input impedance of an end fed antenna into the range where most antenna tuners can produce good performance. The matchbox handles 100 watts of power. ***This project requires an antenna tuner to achieve satisfactory SWR***. If you need an auto tuner, we recommend the LDG line of auto tuners, shown to work well with these matchbox antennas, and highly rated by users. They are reasonably priced and available at most ham stores.

The matchbox project uses readily available common hardware and materials listed below.

Matchbox Parts List

End Fed HF
Matchbox Antenna

1	small plastic enclosure (shown right)
1	powdered iron toroid T130-2
3	20” pieces of 18 AWG solid insulated copper wire in red, green, and black
2	# 8-32 x 3/4” hex head machine screws
2	# 8 lock washers
2	# 8 ring wire lugs
2	# 8 flat washers
2	# 8 lock washer/nut combination
2	# 8 wing nut
1	SO-239 panel mount connector to fit keyed enclosure opening

30’ # 18 AWG insulated stranded wire antenna with ring lug attached Small amount of clear silicone caulk to secure toroid in place

Preparing the MATCHBOX Plastic Enclosure

The enclosure needs one 5/8” hole for the SO-239 connector, and one 11/64” hole for the counterpoise wing nut on the lower side of the enclosure. An 11/64” hole on the upper right side of the box facilitates antenna connector mounting.

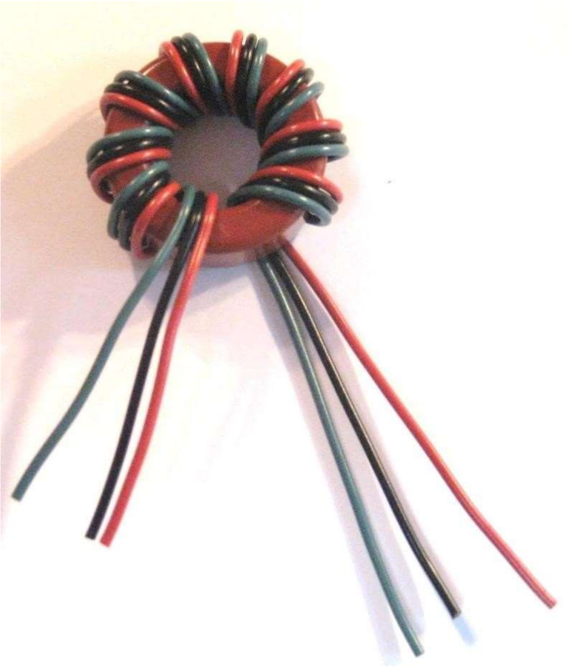

TOROID WINDING

First, wind the three 20” pieces of insulated solid wire onto the toroid. Place the wires as shown green-black-red, and wrap the toroid 9 turns so that it looks like the photo on the right. Notice there are three wires extending from the left winding and three wires extending from the right winding. As the connections are completed, the steps refer to the specific wires by left or right and color.

NOTE: Count turns by counting the number of times the wire goes thru the toroid center.

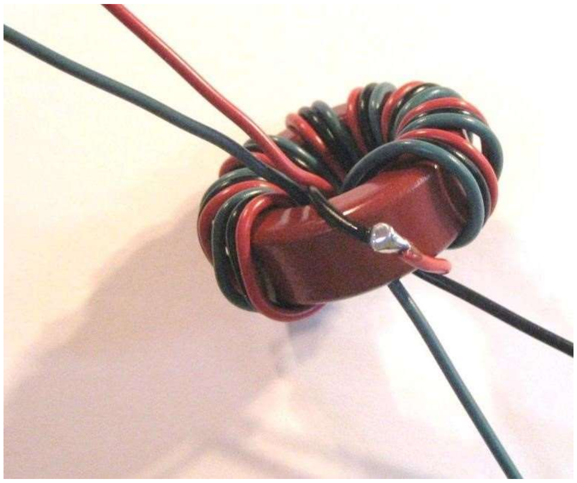

Crimp together and solder ***the left black wire*** with the ***right red wire*** When the step is completed, it will look like the right photo.

The next three steps should appear as shown in the first photo on page 3. Crimp and solder a # 8 lug to the ***left red*** wire about 2” from the toroid. The completed lug will later connect to the antenna connection bolt on the upper inside of the enclosure.

Twist the **left green** wire with the ***right black*** wire. Strip the ends of the two wires and twist together at about 2”. This twisted pair will solder to the center connection of the SO-239 connector in a later step.

Trim and strip the remaining ***right green*** wire at about 2”. Cut an additional 2” piece of green wire, and crimp and solder both wires to a # 8 lug. The 2” green wire will connect to the ground connector on the SO-239 already installed in the enclosure. Strip remaining green wire end 3/8" and bend into a hook for connection to the SO-239 ground connector.

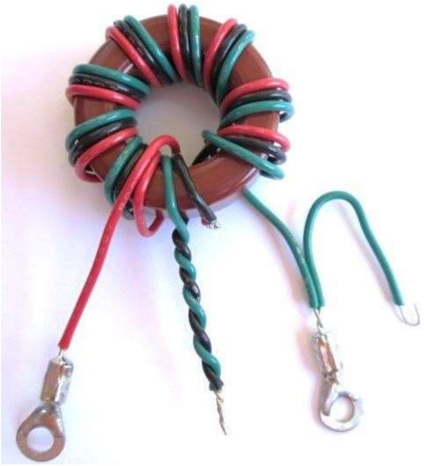

The SO-239 connector and ground lug should be installed through the keyed hole in the lower right with the mounting nut securely tightened.

Solder the green and black twisted pair to the SO-239 center connector. Solder the green wire hook to the ground lug on the SO-239 connector.

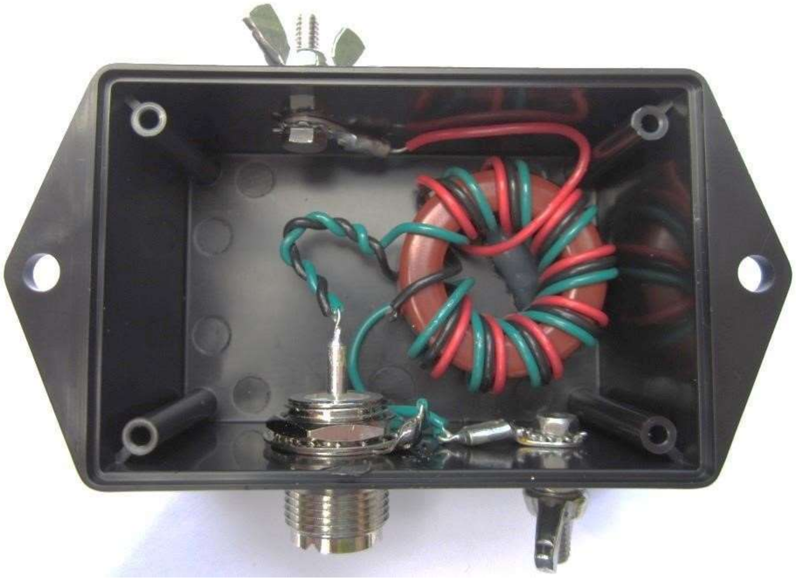

From inside the box, place an 8-32 machine screw through a lock washer, the # 8 lug on the green wire, then through the lower 11/64'" mounting hole. Place a flat washer on the outside protruding machine screw followed by # 8 lock washer/nut and tighten securely. Place a # 8 wing nut on the machine screw to finish the **counterpoise connector**.

Position the toroid inside the box to allow connection of the red antenna wire lug to an 832 machine screw and lock- washer on the upper box side. Place a flat washer on the outside of the box followed with a # 8 lock washer/nut. Tighten the nut securely. Next, place the wing nut on the **antenna connector** and your project appears as shown above.

A small spot of clear silicone compound is used to secure the toroid from movement in the enclosure. The final assembly step is to secure the box cover in place with four screws.

Preparing the Antenna Wire

Matchbox performance will be determined by two factors: The length of the antenna wire, and the capability of the tuner. The length of the antenna wire should be between 24 and 60 feet for best performance. Additional counterpoise is not normally required, as the coax shield provides the counterpoise function.   Wires longer than 60 feet may have excessive impedance for some tuners to properly match. Wires shorter than 24 feet may not radiate as effectively. A 30 foot insulated 18 gauge stranded wire antenna and connecting lug is included with the project and should meet most requirements without need for any counterpoise.

Experience has shown that most external tuners and some internal tuners will tune 80–6 meters with an antenna length of 24’ to 30’. If a longer antenna is desired, the provided antenna can be lengthened.

Some tuners, in particular internal tuners, may not tune the full 80-6 meter range. You may need to try different wire lengths to optimize your antenna configuration. If you are having difficulty getting your rig to tune, start with a 26’ wire. This should produce good results on at least 40-6 meters using the narrowest performance range of internal tuners.

Best performance is achieved with a coax of 16’ or longer, with the coax shield providing the counterpoise function. Additional counterpoise is usually not required in this design, although the lower wing nut provides a convenient counterpoise connector if needed. The end fed antenna system works well in horizontal, sloper, and vertical configurations.

Observe established safety practices when working with antennas, and avoid proximity to power or utility wires. Permanent installations should be equipped with appropriate static and lighting protection.

Keep amateur radio safe and fun!

End Fed HF
Matchbox Antenna

Toroid Wiring Schematic

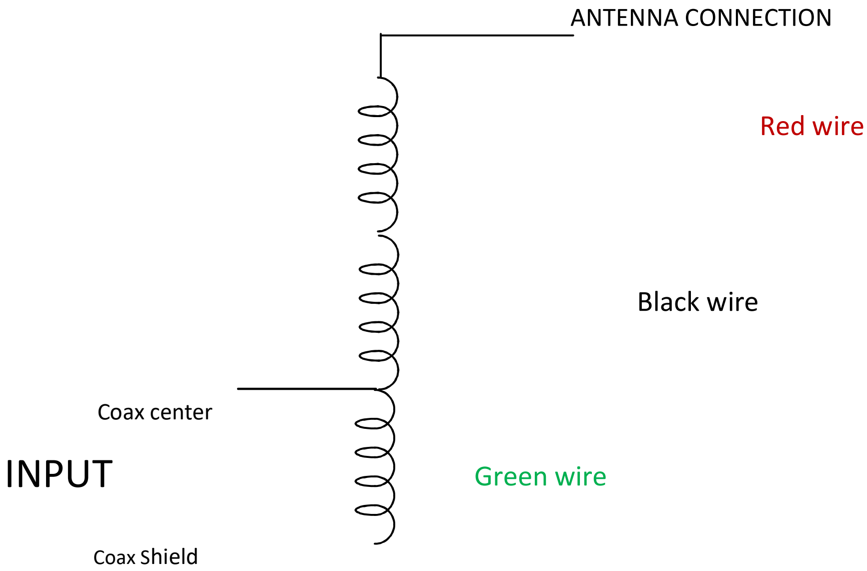
ANTENNA CONNECTION
Red wire
Black wire
Coax center
INPUT
Green wire
Coax Shield

Feedback on 6-40 End Fed Matchbox Antennas

The antenna is amazing. I receive fantastic signal report from DX stations as well as hams around the country. I previously used a short version GR5V that was in an inverted V configuration on the roof of our house. This antenna was noisy but worked. Your antenna out performs the prior antenna and is much quieter. Thanks again for the fine product that your club offers. ***May 2012***

It took me 10 minutes to install and WOW - within 10 minutes I made my first HF contact (from CT to GA). And in the span of 4 hours made 3 more !! I was transmitting just 5 Watts from Yaesu 857ND using this antenna. This antenna rocks ! Thanks a ton ! 73 ***May 2012***

The 6 - 40 Matchbox antenna works great! Easily tunes with the Z100 tuner - very low SWR's on 20m thru 10m. Antenna was up about 45 feet on tree branch - wire length was about 52.5 feet, slight slope, 55 feet of coax, line isolator at radio. ***Work Hawaii for the very first time. Guess the antenna wanted to call home :-) 57 report, 100 watts, 15 Meter band, Icom 718.*** A great portable antenna that I will use often at Montrose Harbor along the lakefront in Chicago. Thank you for a fine, well made product. ***April 2012***

It arrived today in fine shape. I got is up in the air this afternoon using a 31 foot radiator. Top is at 45 feet. Loads on all bands 80-10M with my LDG auto-tuner absolutely no problem, My first QSO was KH7X in the ARRL SS contest with 100W. Amazing. This thing rocks! It's a great antenna! Nice job and a great buy. ***November 2011***

I set-up my jackite pole today and tried a 68' wire with the 9:1 unun matchbox. I set it up as an inverted L with about 26 feet vertical and 42 feet horizontal. With my IC-703 it tuned 80 through 6 all well below 1.5:1. I am going to leave the antenna up a few days and make a few contacts. I think the matchbox is terrific. ***October 2011***

My 6 – 40 meter end fed arrived through the UK holiday mail and I departed from my norm of a sloper and mounted it vertically on a 10m fishing pole. It is surprisingly effective and rewarded me with a surprise contact on 17m into the Falkland Islands at 20:30UT last night. ***January 2012***
When I attached it to a 10m fiberglass fishing pole and went vertical - wow! Easy 5/9 contacts out to 6,000+ miles when propagation was anything better than the worst. My simple ATU easily matches the antenna with SWR never greater than about 1.4 on any band it is built for. ***December 2012***

You can dither and procrastinate, but with this antenna, you'll be talking to the world in no time on any band that happens to be hot - and with your hard-earned cash hardly touched! ***January 2012***

Just wanted to report back that I have tossed the antenna into a tree about 20' up in an inverted vee shape and I am pleased to report that I can tune anywhere I need on 40-10M using the TS-590 internal tuner!! My first contact was on 40M within 10 minutes of erecting the antenna and was all the way to TX from my condo here in MI using just 10 watts. This is by far the best HAM related investment I have made in a long time. Living in a condo with strict association rules, I am able to conceal this wire in a tree out my bedroom window and talk all over. I'm very happy and would recommend this antenna to anyone in a similar situation. ***March 2013***

I have had your multiband end fed antenna installed here in Singapore for a few weeks now and am very pleased. I used the wire from my previous antenna which is 66 feet long. It works beautifully! I have it hanging vertically outside my 15th story apartment window on a 5 foot horizontal pole (to get it away from the reinforced concrete building). A couple of fish weights on the end keep it from blowing around too much. I would recommend this antenna to anyone who is facing a challenging QTH. ***April 2013*** I received my antenna yesterday, a quick 4 days since I ordered it. I attached it to a 31 foot pole and my IC-706MK2. Within minutes, I made my first contact with 9A4KW in Zagreb, Croatia on 20 meters. Being able to reach out over 4,000 miles gives me real confidence that I have found my portable antenna. Thanks for your rapid response and fine product. ***May 2013***

In only 2 weeks of casual on-air time. If you could see how I had to zig-zag it through a big tree above the roof of my RV, you'd be saying like I am, "how the heck can that work? I worked 90% of the stations I called, and just added Mexico City on 15m that gave me a 5x9. ***September 2013***

Just wanted to let you know how pleased I am with my ERAC matchbox antenna. I put the EARC antenna in the backyard with the feed point at 4 feet and the wire sloping up to 35 feet. I have 125 feet of coax feed line. My transceiver's built-in antenna tuner will adjust the SWR to below 1.2 to 1 on all bands 80-6 meters. Last weekend I worked 65 countries on 5 continents. I live in an antenna restricted community and this antenna has allowed me to enjoy ham radio again. ***February 2014***

I used the EARC Matchbox today. It works GREAT on 20m with a 30 ft radiator in vertical positioning, fed with 25ft of RG175/u coax. I've been talking up and down the east coast with 3W on SSB all afternoon! People ask me to repeat my power output again and again. I can't wait until I get the chance to use it at night, where I think it'll really shine! This has a permanent place in my portable kit. Thanks for the great product! ***April 2014***

I performed a simple install in Virginia making it vertical by using a 28 foot fiber glass kite pole. Within 30 minutes (NO KIDDING) I was receiving Lebanon on 17 meters! I also worked Russia and Eastern Europe that night on 20 meters. I love this antenna and I give it a 100% (5 STAR) rating. ***May 2014***

This little box is magic. NY, IA and NM from GA with ease on 100w ***June 2014***

Just wanted to let you guys know, this matchbox is wonderful. I built the matchbox, connected it to 55 feet of solid #14 about 20 feet in the air. My rig is a Yaesu FT-817ND connected to a MFJ-971 tuner with coax running between the tuner and the matchbox. So far I have worked Slovenia twice, Morocco, St. Helena Island, Cuba, Cost Rica, Vermont, Wyoming, Florida, North Dakota, Brazil, and Colombia. This setup works really well on the higher bands. Thank you for putting this design on the internet. I like it so much its become permanent here. ***September 2014***

This antenna continues to amaze me. It worked well U-shaped "indoors" with 50 watts, and now outdoors as a sloper. It easily loads 10-40m with my TS-570s internal tuner. I'm sold on it and own two of them now. I'm working at least 90% of the DX I chase. I have no RFI in the shack either. ***November 2014***

EARC Antennas are operating in virtually every state, these countries and more!

Argentina Australia Belgium Brazil Canada Chile Denmark Greece

Hong Kong India Ireland Israel Italy Mexico Mozambique New Zealand Russia Singapore
Sweden Tasmania United Kingdom

ORDERING A MATCHBOX ANTENNA FROM EARC

Volunteers of the Honolulu Emergency Amateur Radio Club (EARC) carefully assemble the 6-40 End Fed matchbox antennas for those who prefer not to build one. For a donation of **$56** (Including USPS priority mail shipping) club volunteers will build and deliver a 6-40 matchbox with a 30 foot antenna wire anywhere in the U.S., we do not ship internationally. Note that coax cable is not included.

Payment by PayPal

To pay via **PayPal**, make your payment to paypal@earchi.org using any major credit card. Include your name, mailing address, and email address with the order.

We will ship promptly and you will likely receive the order in the US in 7 days or less. If you have any questions about your order, please email webmaster@earchi.org

Thank you for your interest in the activities of Honolulu Emergency Amateur Radio Club and amateur radio.

GPS coordinates for 20 of the Top HAARP facilities

GPS coordinates for 20 of the Top HAARP facilities:

1- High Frequency Active Auroral Research Programn (HAARP)
Gakona Alaska
+62° 23’ 30.00″, -145° 8’ 48.00″

2- High Power Auroral Stimulation Observatory (HIPAS)
Fairbanks Alaska
64° 52’ 19″ N 146° 50’ 33″ W

3- Poker Flat Research Range
Near Chatanika, Alaska
+65° 7’ 23.90″, -147° 28’ 7.05″

4- Platteville Atmospheric Observatory
Platteville, Colorado
+40° 10’ 54″, -104° 43’ 30″

5- Millstone Hill Radio Observatory
Westford, Massachusetts
+42° 37’ 09.25″, -71° 29’ 28.49″

6- The Arecibo Observatory radio telescope
Arecibo Puerto Rico.
18° 20’38.97″N 66° 45’9.77″W

7- Jicamarca Radio Observatory
Lima, Peru
11° 57’ 08.25″ S 76° 52’ 30.67″ W

8- São Luiz Space Observatory
Cruzeiro Santa Bárbara, Sao Luis-MA, Brasil
-2° 35’ 40.47″, -44° 12’ 35.90″

9- Nerc MST Radar Facility
Capel Dewi, Carmarthenshire, Wales, United Kingdom
+52° 25’ 28.26″, -4° 00’ 19.59″

10- European Incoherent Scatter Scientific Association EISCAT
Ramfjordmoen, Near Tromsø, Norway

+69° 35’ 10.94″, +19° 13’ 20.89″

11- Sura Facility
Nizhniy Novgorod, Russia
+56° 7’ 9.70″, +46° 2’ 3.66″

12- The Russian Woodpecker
Duga Radar Array, Chenobyl, Ukraine
51° 18’20.17″N, 30° 04’02.60″E

13- National MST Radar Facility NMRF
Gadanki, India
+13° 27’ 26.68″, +79° 10’ 30.74″

14- China Research Institute of Radiowave Propagation (CRIRP)
Xinjiang (Sinkiang) Region
40° 24’15.91″N, 93° 38’09.74″E

15- Sheshan, Shanghai,
China
31° 5′ 41.98″ N, 121° 11′ 29.72″ E

16- Mu Radar
in Japan
34° 51’14.80″N 136° 6’19.45″E

17- Jindalee Operational Radar Network (JORN)
Laverton, West Australia
-28° 19’ 36.29″, +122° 0’ 18.84″

18- Jindalee Operational Radar Network (JORN)
Longreach, Queensland, Australia
-23° 39’ 29.53″, +144° 8’ 49.58″

19- Australia Naval Communications Station Harold E Holt (NCS HEH)
Exmouth, Western Australia. 21° 48’ 59.47″s, 114° 09’ 55.60″

20- Zhong Shan Antarctic Polar Station (China)
69º 22’ 43.69″ S 76º 23’ 15.07″ E

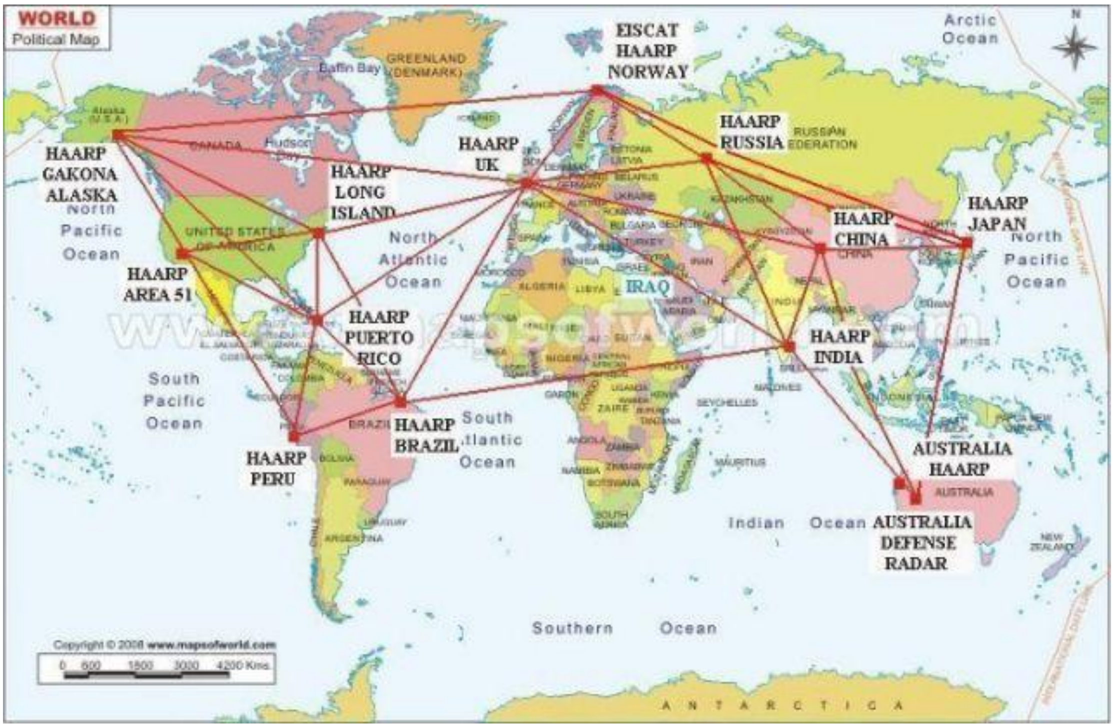
WORLD
Political Map
EISCAT
HAARP
NORWAY
Arctic Ocean
GREENLAND (DENMARK)
Baffin Bay
HAARP
GAKONA
ALASKA
CANADA
Hudson Bay
HAARP
LONG
ISLAND
HAARP
UK
HAARP
RUSSIA
RUSSIAN FEDERATION
HAARP
JAPAN
HAARP
CHINA
North Pacific Ocean
UNITED STATES OF AMERICA
HAARP
AREA 51
North Atlantic Ocean
IRAQ
HAARP
PUERTO
RICO
HAARP
INDIA
South Pacific Ocean
HAARP
BRAZIL
South Atlantic Ocean
HAARP
PERU
AUSTRALIA
HAARP
AUSTRALIA
DEFENSE
RADAR
Indian Ocean
Southern Ocean
ANTARCTICA
Copyright © 2008 www.mapsofworld.com
0 600 1800 3000 4200 Kms.

The Future of the High frequency Active Auroral Research Program (HAARP)

Robert P. McCoy
Geophysical Institute, University of Alaska Fairbanks
903 Koyukuk Drive, P. O. Box 757320
Fairbanks, Alaska 99775

ABSTRACT

The HAARP facility, located in Gakona Alaska, is the world's premier laboratory for active experimentation in the ionosphere and upper atmosphere. Understanding the ionosphere is particularly important because it affects long-distance communication, navigation, and radar applications relying upon radio propagation through this ionized atmospheric region. The primary component of HAARP, the Ionospheric Research Instrument (IRI), is a phased array of 180 HF tower antennas spread across 33 acres and capable of radiating 3.6 MW into the upper atmosphere and ionosphere. The array is fed by five 2500 kW generators, each driven by a 3600 hp diesel engine (4 + 1 spare). Transmit frequencies are selectable in the range 2.8 to 10 MHz. HAARP is owned by the Space Vehicles Directorate of the Air Force Research Laboratory (AFRL/RV) in Albuquerque, NM. In 2013 AFRL/RV decided that they no longer needed the facility, and by prior arrangement, were prepared to remediate the site. In response to a letter writing campaign and petition by the scientific community, the Secretary of the Air Force directed that the demolition be postponed until May, 2015 to allow time for another federal agency, a consortium of academic or research institutions, or the University of Alaska, to build a plan to sustain the facility and preserve its capability to enable fundamental ionospheric research. For the past year the Geophysical Institute of the University of Alaska Fairbanks has been in active negotiations with the Air Force and a sustainment plan for HAARP, using a pay per use model may soon be approved and implemented preserving HAARP for future upper atmospheric investigations.

OVERVIEW

The HAARP facility, located in Gakona Alaska, is the world's premier laboratory for active experimentation in the ionosphere and upper atmosphere. Scientists at HAARP use high frequency (HF) radio transmitters to heat selected regions of the ionosphere and observe the effects. With a facility like HAARP, the ionosphere becomes a laboratory to study a range of physics phenomena. For traditional space research using ground-based observations or experiments on sounding rockets, it can take an extremely long time (days, weeks, years) to get the desired natural overhead conditions. Satellites can amass much larger databases but it is difficult to coordinate the satellite with the desired phenomena. With a facility like HAARP, it is possible to perform an experiment at

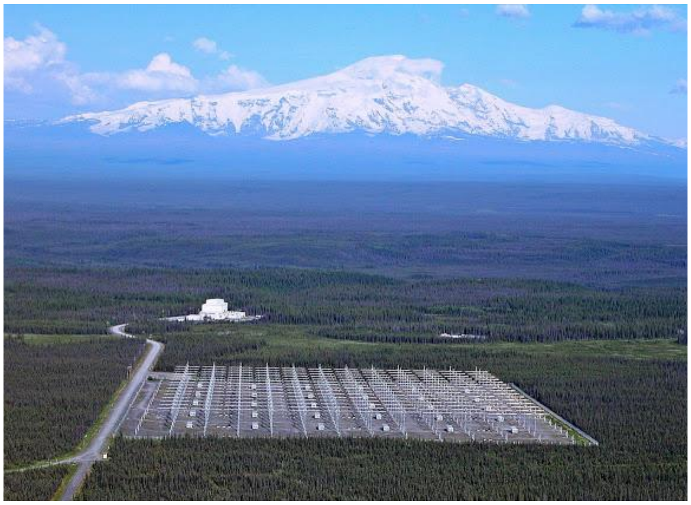

Figure 1 Overhead Photo of HAARP Courtesy of M. McCarrick

will to create plasma structures and irregularities; use the ionosphere like an antenna to excite low frequency waves; create weak luminous aurora-like glows; and a variety of other experiments. Ground-based instruments or satellite overflights are used to collect the necessary data for the investigation.

The primary component of HAARP is the Ionospheric Research Instrument (IRI), a phased array of 180 HF tower antennas spread across 33 acres and capable of radiating 3.6 MW into the upper atmosphere and ionosphere. The array is fed by five 2500 kW generators, each driven by a 3600 hp diesel engine (4 + 1 spare). Transmit frequencies can be chosen anywhere in the range 2.8 to 10 MHz and since the antennas are part of an array, the transmitted beam can take many shapes, can be scanned over a wide angular range (down to 30° deg. elevation angle) and multiple beams can be formed and steered. The facility employs 30 transmitter shelters each with six pairs of 10 KW transmitters (6 x 2 x 30 x 10) to achieve the 3.6 MW transmit power.

The facility is located at 62.39° N, 145.15° W in the subarctic which is an ideal location for a wide range of scientific experiments. The location is just south of the auroral region which allows active experiments outside of the aurora. During substorms HAARP's location is under the aurora or inside the auroral oval. The high latitude location is also ideal for experiments generating low frequency waves which propagate along magnetic field lines high into the magnetosphere.

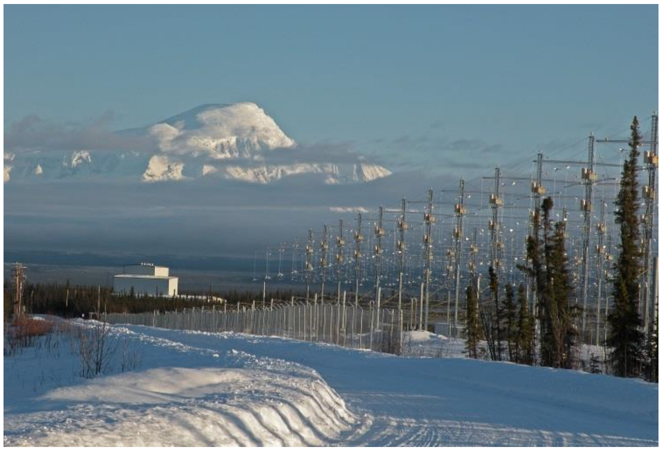

Figure 2 HAARP Antennas Courtesy USRA

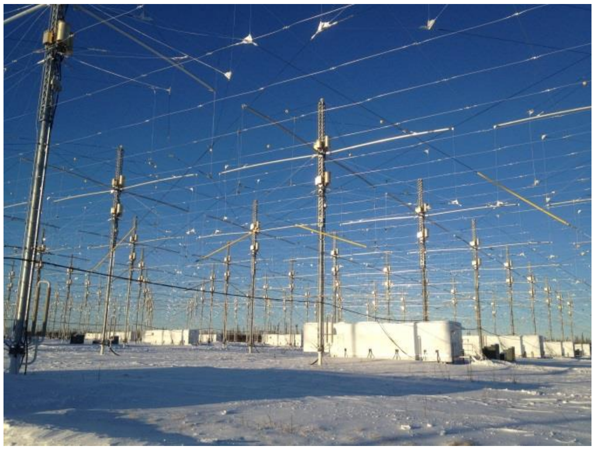

Figure 3 HAARP Antennas and Transmitter Shelters Courtesy USRA

Initiated in 1989 with congressional earmarks (~$180M), HAARP was completed around 2009 with additional investments of $110M from the Air Force Research Laboratory (AFRL), the Office of Naval Research (ONR) and Defense Advanced Research Program Agency (DARPA) and represents a total investment of ~$290M. HAARP is currently owned by the AFRL Space Vehicles Directorate (AFRL/RV) in Albuquerque NM.

There are two other ionospheric heaters in the world, EISCAT in Norway and Sura in Russia. The National Science Foundation is currently building a fourth hearter in Arecibo Puerto Rico. HAARP is by far the most powerful and capable heater in the world.

HAARP SCIENCE

A full discussion of the science capabilities, accomplishments and future science potential of HAARP is not practical here but the 2013 National Research Council Workshop Report: ***Opportunities for High-Power, High-Frequency Transmitters to Advance Ionospheric/Thermospheric Research*** (see below) is an excellent resource. As a scientific tool for active experimentation in the ionosphere HAARP has a number of characteristics that make it unsurpassed in the world including:

- Effective radiated power (ERP) = power x antenna gain;
- Steerability of the beam and ability to create multiple beams;
- Frequency range and ability to select frequency;
- Polarization control;
- Modulation control;
- Location – relative to the aurora and on high latitude magnetic field lines;
- Accessibility of the site from either Anchorage or Fairbanks Alaska.

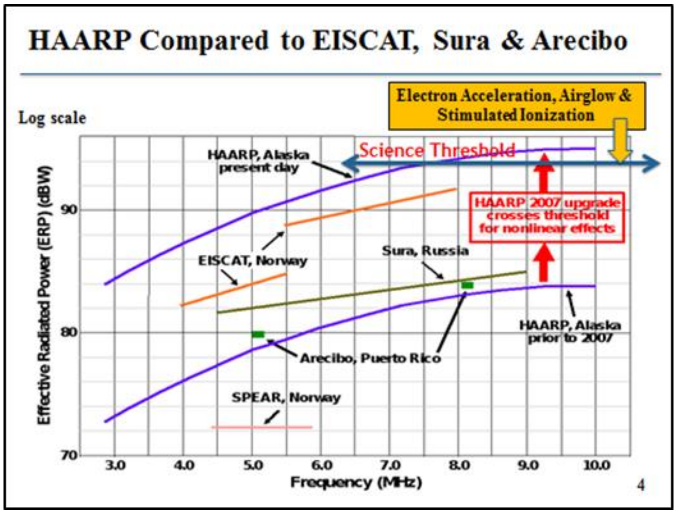

Figure 1 Comparison of Existing Ionospheric Heaters (Kennedy, 1998)

A short summary of HAARP science highlights in a few topic areas includes:

RADIO SCIENCE: Production of artificial plasma layers to study radio frequency propagation at a range of frequencies from a few Hz to a few GHz (including GPS frequencies). Studies include guidance, redirection, enhancement and degradation and loss of signals.

MESOSPHERE-THERMOSPHERE DIAGNOSTICS: Production of artificial periodic irregularities; neutral density and temperature measurements and generation of optical emissions. Upper atmospheric heating has been used to change density, composition, neutral diffusion, thermospheric winds and potentially – drag on orbiting debris or satellites. HAARP can be used in studies of polar mesospheric clouds.

SPACE WEATHER: The use of HAARP to make comparisons quiet ionospheric conditions to substorms and mapping up into the plasmasphere. Use of the facility to study a variety of transient phenomena including subauroral polarization stream (SAPS); subauroral ion drift (SAID), high electron temperature regions and troughs, and ion-outflow-created atmospheric gravity waves.

MAGNETOSPHERE-RADIATION BELTS: This line of investigation primarily involves using the ionosphere in a variety of modes as a "virtual antenna" to inject a variety of waves including whistler waves, shear Alfvén, and magnetosonic waves and waves at frequencies down to a few Hz and study resonance and amplification effects. There are a wide range investigations of wave/particle interactions in the magnetosphere enabled by HAARP (Rietveld et al., 1989; Barr, 1998; Papadopoulos et al., 1990, 2005).

FUTURE SCIENCE POSSIBILITIES

Recent experiments at HAARP (Fallen et al., 2011) demonstrated experimental techniques to continuously generate artificial plasma ducts that extend from the upper ionosphere into the earth's magnetosphere. These plasma structures facilitate the ducting of VLF waves generated artificially either by, HAARP signals interacting with the ionosphere, from other ground based sources, or generated in space. This new discovery is relevant to the efforts to reduce high energy electrons in the radiation belts (see below).

Continued basic science efforts on this topic should reveal how this technique may be optimized for a variety of geophysical conditions. Measurements of duct geometry may then be used in conjunction with numerical models and simulation to estimate levels of VLF power density that may be deposited in the equatorial region of the earth's magnetosphere within the radiation belts. This efficient injection of artificially created VLF waves into the radiation belts acts to scatter and remove high levels of energetic electrons injected from high-altitude nuclear burst.

DOD APPLICATIONS

Interest by the DoD (AFRL, ONR and DARPA) in HAARP is obvious by the $105M they invested to complete and upgrade the facility. Historically, most of the applications of interest stem from the ability of HAARP to generate low frequency waves that propagate both within the Earth – atmosphere waveguide around the globe and which propagate upward along magnetic field lines into the magnetosphere. More recently the focus has been on using HAARP to stimulate artificial glows in the upper atmosphere and creating artificial plasma clouds.

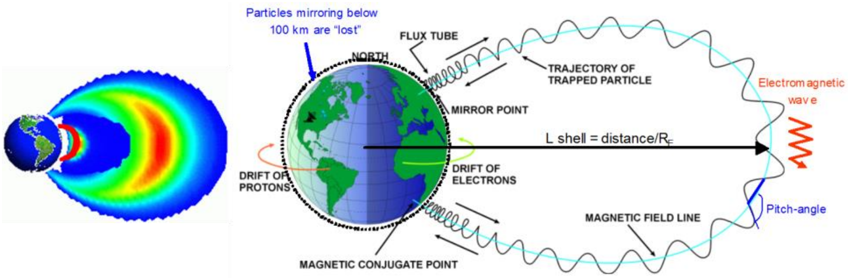

Figure 2 Radiation Belts and Trapped Electrons in Magnetic Flux Tubes

REDUCTION OF ELECTRONS IN THE SPACE RADIATION BELTS: The highest priority application for HAARP and the one stimulating most of the investment involves using low frequency waves to cause high energy million-electron-volt "killer" electrons in the magnetosphere to precipitate out. The Earth's natural radiation belts contain high energy electrons trapped in magnetic bottles (flux tubes) and propagating for extended periods from pole to pole spiraling along magnetic field lines. The density of these trapped particles can go up dramatically from a solar coronal mass ejection (SME) or from a manmade event like a high altitude nuclear burst. A 1000 fold electron density increase can easily be obtained from the latter and these electrons can persist for months or years in the radiation belts. These killer electrons can damage and destroy electronics in low Earth orbiting satellites and there is an estimated $100B of non-DoD satellites at risk.

Experiments with HAARP and other radio frequency transmitters have shown that a variety of low frequency waves can be generated in the ionosphere which propagate along magnetic field lines to interact with these trapped electrons. Wave-particle interaction occur which modify the angle between the trapped electron and the magnetic field line, the pitch angle. On subsequent reflections near either pole these electrons penetrate deeper into the upper atmosphere colliding with neutral atoms and molecules and are lost. This process reduces the density of electrons trapped in the radiation belts.

There have been a number of HAARP investigations on this topic sponsored by several agencies. Based on results observations and models, AFRL/RV built the DSX satellite to investigate the effect of emitting low frequency waves directly in space and will soon launch this vehicle soon to test the effect. Neither HAARP nor DSX provide a comprehensive solution for this problem but both provide valuable insights into the science of wave-particle interactions on the radiation belts.

SUBMARINE COMMUNICATION: The Navy's interest in HAARP was frequently related to using the ionosphere as a large antenna to excite extremely long wavelength (ELF) waves to propagate long distances and penetrate the oceans to communicate with submarines.

DETECTION OF CAVITIES IN THE EARTH: The same ELF waves that penetrate the oceans can also penetrate into the Earth's lithosphere and experiments with HAARP have been used to detect cavities; mines and caves underground.

OVER THE HORIZON RADAR: The current HAARP site was originally an over the horizon (OTH) radar site. OTH radar depends on reflecting HF waves off the ionosphere. Solar and geomagnetic activity produce problems for HF radar propagation in the Arctic. Auroral activity creates irregularities and bubbles in the ionosphere causing scintillation and loss of HF signals. Recent HAARP experiments (Pedersen et al., 2010) demonstrate the ability to create sustainable artificial plasma clouds or plasma balls at altitudes below the natural ionosphere which may be useable as an artificial mirror to reflect HF radar and other RF communication signals.

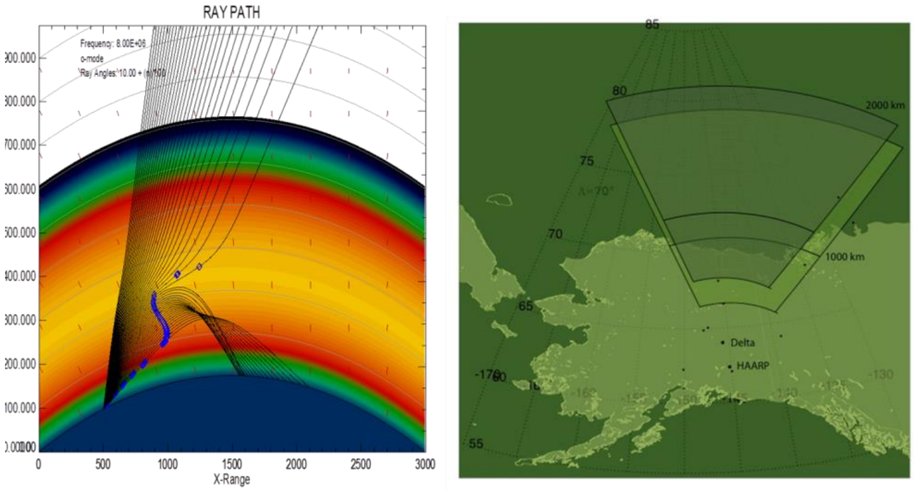

Figure 3 Ionospheric propagation of 8MHz rays and potential OTH footprint from HAARP Courtesy of Bill Bristow

As the sea ice in the Arctic continues to melt opening up polar navigation routes, there is increasing interest by the US Northern Command and NORAD in the possibility of deploying a system of OTH radars across North America for increased Arctic maritime domain awareness. HAARP may play an important role as an OTH test bed and demonstration platform for this new capability.

TRANSIONOSPHERIC RF PROPAGATION: As HAARP heats the ionosphere it can create bubbles and irregularities. These irregularities can then interfere with a wide range of space to ground RF propagation at frequencies up to 10 GHz. A facility like HAARP could be used to cause interference with communication signals, GPS navigation or blur space-based synthetic aperture radar.

LASER FUSION APPLICATIONS: Another application of an ionospheric heater is to use the natural ionosphere to perform nonlinear plasma experiments in unbounded plasma. The power levels of HAARP are small compared to those in fusion experiments but the lack of walls in the ionosphere, combined with HAARP's tunability and phase array transmissions allow a wide range of experimentation in parametric instabilities and nonlinear plasma physics.

THERMOSPHERIC HEATING EXPERIMENTS: Through coupling between the ionosphere and thermosphere it is possible to heat the neutral atmosphere above HAARP. This heating can be used to create density plumes and drive neutral waves over HAARP to test applications to increase drag on selected orbiting satellites or orbital debris.

HAARP GROUND-BASED DIAGNOSTICS

The IRI at HAARP creates the effects in the ionosphere but it is the diagnostic ground-based instruments which allow scientific investigations. The state of the diagnostic instruments at HAARP is in flux and many of the sensors have been temporarily removed. Most of the sensors are slated to be returned and restored to service during the spring/summer 2015. The discussion that follows lists the instruments that were traditionally resident at HAARP and used to support science and may represent the set of diagnostics later in 2015.

The primary diagnostic tool at HAARP is the digisonde. This instrument, basically a swept frequency HF radar, measures the bottom side of the electron layer in the ionosphere. Additional sensors at the facility include:

- All-sky Riometer
- Imaging Riometer (8 X 8 Array)
- Fluxgate Magnetometer
- Induction Magnetometer
- Optics
  - All-sky imager
  - Telescopic imager
  - Photometers
  - 14 ft Optical Dome
- Tomography Chain (Cordova -> Kaktovik)
- VHF Radar (139 MHz)
- MUIR (Modular UHF Ionospheric Radar
- Ionospheric Scintillation Receivers
  - SATSIN (offsite)
  - GPS-NOVATEL
  - Total Electron Content
- HF 2-30 MHz High Angle Receiving Antenna
- Scanning Doppler Interferometer (SDI)

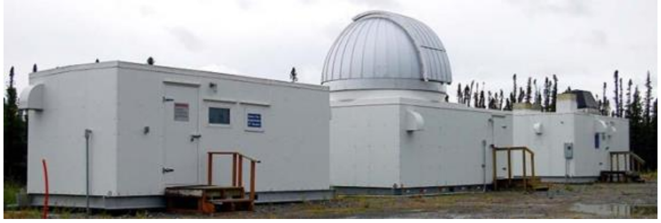

Figure 4 HAARP Optics Facility

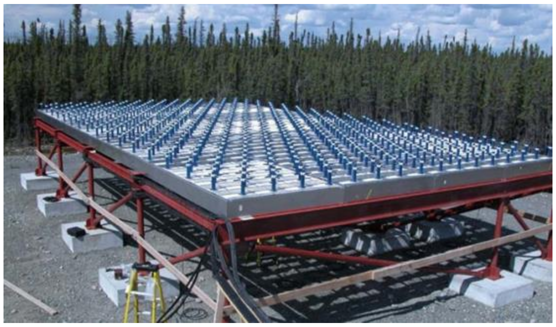

Figure 5 MUIR Radar

- Radio Background Receivers
    - Broadband ELF / VLF Receiver network.
    - SEE Receiver string.
    - HF to UHF Spectrum Monitor

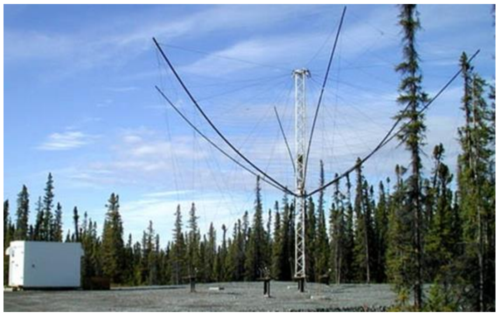

Figure 6 HF High Angle Receive Array

NATIONAL ACADEMY OF SCIENCE INPUT

The National Research Council (NRC), the operating arm of the National Academy of Sciences has taken up the issue of HAARP twice over the past two years. In 2013 the NRC completed a Decadal Survey in Solar and Space Physics. One of the priority recommendations of that study included:

> **Priority -** ***Fully realize the potential of ionospheric modification techniques through collocation of modern heating facilities with a full complement of diagnostic instruments including incoherent scatter radars. This effort requires coordination between NSF and DOD agencies in planning and operation of existing and future ionospheric modification facilities.***

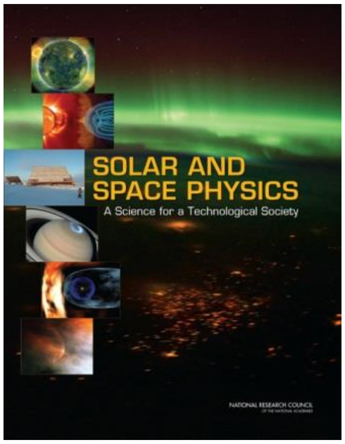
SOLAR AND SPACE PHYSICS
A Science for a Technological Society
NATIONAL RESEARCH COUNCIL

Figure 7 2013 Space Physics Decadal Study

In March 2013 in response to a request by AFRL and NSF the NRC conducted a workshop entitled: **Opportunities for High-Power, High-Frequency Transmitters to Advance Ionospheric/Thermospheric Research** focused primarily on HAARP. Topics addressed in that three day workshop included:

- Geospace and space weather;
- Stimulated emission and radiation belts;
- Radio science; and
- Communications.

NRC Workshops do not produce recommendations, but the NRC published a report containing 72 pages by 40+ scientists documenting past science results from HAARP and a wealth of prospects for future science from the facility.

The report reiterated the importance of co-locating an incoherent scatter radar to HAARP to maximize the science potential of the facility. The NRC report can be obtained at:

http://www.nap.edu/booksearch.php?booksearch=1&term=sale&record_id=18620

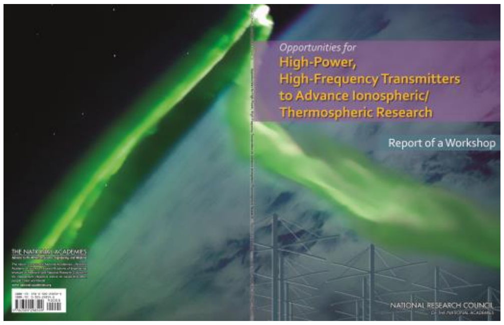
Opportunities for
High-Power,
High-Frequency Transmitters
to Advance Ionospheric/
Thermospheric Research
Report of a Workshop
THE NATIONAL ACADEMIES
NATIONAL RESEARCH COUNCIL

Figure 8 2013 NRC Workshop Report

NEED FOR AN INCOHERENT SCATTER RADAR

A diagnostic instrument noticeably missing from HAARP is an incoherent scatter radar (ISR). An ISR is one of the most powerful tools for studying the ionosphere from the ground. An ISR can routinely measure vertical profiles of electron density, electron and ion temperatures, bulk ion motion, ion masses, ion-neutral collision frequencies. Over the years there have been several plans to build an ISR at HAARP but because of funding variability for the facility it was never completed. There have been several studies, workshops and panels which made recommendations to install an ISR at HAARP, most notably the two NRC studies mentioned above but also a DoD sponsored panel in 2001 entitled HAARP "Future Directions" but usually called the "Tether Panel" after the chairman Dr. Tony Tether.

The NSF owns and operates a chain of ISRs from near the north magnetic pole to the magnetic equator at such sites as: Resolute Bay Canada; Sondrestrom Greenland; Millstone Hill Massachusetts; Arecibo Puerto Rico and Jicamarca Peru. Additionally, they own one more outside of that chain at Poker Flat Alaska about 300 miles north of HAARP. Older ISRs have a conventional radiating transmitter and dish reflector design, but two of the newer ISRs in Canada and Alaska are built with phased array transmitters and are far more flexible and capable. Both of these ISRs are called Advanced Modular Incoherent Scatter Radar (AMISR) and for distinction the Resolute Bay is commonly referred to as RISR and the Alaskan ISR is called the PFISR.

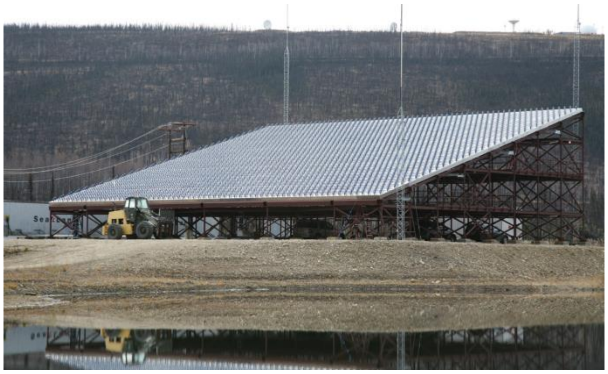

Figure 9 PFISR Incoherent Scatter Radar

The NSF PFISR is located on the property of the University of Alaska owned Poker Flat Research Range (PFRR) which is used by NASA to launch sounding rockets to study space physics in the arctic. The PFISR is employed for routine and campaign ionospheric science and frequently supports NASA sounding rocket launches. The radar consists of 128 identical panels which constitute the phased array. The MUIR radar at HAARP (Figure 9) is built with the same technology but with only 16 panels. That radar has some capability to sound the ionosphere but it does not have ISR capability. Expanding the MUIR to a total of 96 panels should give it ISR capability. There have been discussions of temporarily relocating PFISR to HAARP for a year, or two, to test out the heater/ISR combination to enable space research. Other options which have been discussed include reducing PFISR to 96 panels by transferring panels to HAARP and expanding the (adding panels) to the MURI array with a potential of ending up with two ISRs in Alaska.

EDUCATIONAL OPPORTUNITIES AT HAARP

With sponsorship from NSF and AFRL HAARP has provided numerous opportunities for graduate and undergraduate research projects and there are several ionospheric scientists working in the field who got degree in heater science with experiments at HAARP. For several years NSF and AFRL sponsored the Polar Aeronomy and Radio Science (PARS) summer school which gave undergraduates and graduate students opportunities to work with their advisors to perform science experiments at HAARP. This summer school also included tours and classes at the Geophysical Institute at the University of Alaska Fairbanks (UAF-GI) and tours of PFRR.

RECENT DEVELOPMENTS WITH HAARP

In 2012, anticipating budget cuts, AFRL/RV contacted the University Space Research Associates (USRA) and the UAF-GI to see if we could help them find a way to lower operating costs and/or divest themselves of management/operation of HAARP. USRA and UAF-GI formed a team and submitted a proposal to operate HAARP under an extended use lease arrangement which was not funded.

In 2013, after the release of the NRC workshop report and after a workshop sponsored by the Office of Science Technology Policy (OSTP), the UAF-GI proposed the creation of a Federal Facility Partnership for HAARP spreading the operating costs across five federal agencies (AFRL, ONR, DARPA, NSF and DOE), selling each a two-week campaign with an operating cost similar to what EISCAT charges customers. That proposal was not successful.

In July 2013, in response to a letter writing campaign by the scientific community; letters of support from leadership of the American Meteorological Society (AMS) and the Space Physics and Aeronomy Section of the American Geophysical Union (AGU/SPA); and a letter from U. S. Senator Lisa Murkowski; Air Force Secretary Deborah Lee James decided to postpone decommissioning plans for HAARP which had been scheduled for June 2014. She allowed time (May, 2015) for the “University of Alaska or a consortium of academic or research institutions” to develop a plan to take ownership of HAARP for the purpose of keeping the facility open covering operating costs from the user base.

In June 2014, DARPA completed a phase of the Basic Research on Ionospheric Characteristics and Effects (BRIOCHE) campaign and soon after the facility was placed in cold-storage. Transmitter tubes were removed from the transmitter shelters and moved into the heated main building to minimize heating costs for the winter. The diesel generators were winterized and many of the diagnostics were transported to AFRL at Kirtland AFB.

Since then negotiations have progressed well and the UAF-GI is working with AFRL/RV on a cooperative research and development agreement (CRADA) to transition ownership of the facility to the University of Alaska to allow the UAF-GI to bring it back on line for continued ionospheric, thermospheric and magnetospheric research. Current projections are that the UAF-GI will take ownership in the fall of 2015. Plans are being finalized to return the diagnostics to HAARP in the spring 2015. The UAF-GI is working to develop a plan to reduce the operating costs of the facility and is actively seeking customers, both domestically and internationally, to come to HAARP to do active experiments at the world’s premier upper atmospheric active research facility.

REFERENCES

Barr, R. 1998. The generation of ELF and VLF radio waves in the ionosphere using powerful HF transmitters. Advances in Space Research 21:677-687, doi =10.1016/S0273-1177(97)01003-X.

Fallen, C. T., J. A. Secan, and B. J. Watkins. 2011. In-situ measurements of topside ionosphere electron density enhancements during an HF-modification experiment, Geophys. Res. Lett., VOL. 38, L08101, doi:10.1029/2011GL046887.

This is Google's cache of http://lucianarchy.proboards.com/index.cgi?board=zorgon&action=print&thread=8872. It is a snapshot of the page as it appeared on 18 Apr 2011 10:28:22 GMT. The current page could have changed in the meantime. Learn more

These search terms are highlighted: **vlf haarp** These terms only appear in links pointing to this page: **thompson manitoba**

'Special Guest' Boards » The Zorgon Zone, with Special Guest Ron Schmidt » **Every One Speaks about HAARP, but did you know...**
http://lucianarchy.proboards.com/index.cgi?board=zorgon&action=display&thread=8872

Every One Speaks about HAARP, but did you know...
Post by zorgon on Sept 21, 2010, 6:46pm

Every One Speaks about **HAARP**...

It gets blames for everything, it has magical abilities etc...

While it is true that it does indeed have some interesting abilities, there are many threads and documents on **HAARP**. I have been collecting material for years BUT, though its been hinted at from time to time that other countries also have a version of **HAARP** and that incoherent scatter sites have a similar function as **HAARP** and EISCAT has been mentioned...

Very few people have taken the time to follow the leads in the various documents and threads and put together the GLOBAL NETWORK

While **HAARP** itself may not have the power to create a planetary shield (for missile defense and solar radiation) the world wide network certainly does.

So for this thread I will leave the much discussed Alaskan **HAARP** in the side lines for now and lets do a tour of the other facilities around the world.

First a little bit on connections to other facilities. In reports like the NAVY Press Release of Dr Paul A. Bernhard... we see some terms already...

Dr. Bernhardt used the high-power HF facility near Arecibo, Puerto Rico, and Nizhny Novgorod, Russia, from 1985 to 1999...

Dr. Bernhardt has conducted over eight experiments from 1985 to the present, using the Space Shuttle Orbital Maneuver Subsystem (OMS) engines, to modify the ionosphere with high-speed exhaust injections into the upper atmosphere. The results of these dedicated engine burns have been recorded using the VHF and UHF radars at Arecibo, Puerto Rico; Kwajalein, Marshall Islands; Millstone Hill, Massachusetts; and Jicamarca, Peru. Currently, Dr. Bernhardt is the principal investigator for the Charged Aerosol Release Experiment (CARE) program, designed to study the scattering of radar from electrons in the vicinity of charged particulates that form artificial "dusty plasmas."

...Coherent Electromagnetic Radio Tomography (CERTO)
...Computerized Ionospheric Tomography Receiver in Space (CITRIS)

http://www.nrl.navy.mil/pao/pressRelease.php?Y=2006&R=17-06r

So from this single press release we have three search terms. Just ignore the facts about the chemical releases and the shuttle spewing fumes, we already have a thread on that 😛)

...Incoherent Scatter Radar (ISR) (Like EISCAT for European Incoherent Scatter Radar)
...Coherent Electromagnetic Radio Tomography (CERTO)
...Computerized Ionospheric Tomography Receiver in Space (CITRIS)

and these locations mentioned directly

Arecibo, Puerto Rico
Nizhny Novgorod, Russia
Kwajalein, Marshall Islands
Millstone Hill, Massachusetts
Jicamarca, Peru

From our research collected at Pegasus we can add these...

HIPAS - High Power Auroral Stimulation Observatory
IRIS - Imaging Riometer for Ionospheric Studies

The problem is I have gathered so much material on these sites and still have more to do... that it may get lost in a single thread...

So lets start with the best **HAARP** like arrays and Incoherent Scatter facilities.

Now this will take me a while to pull up all my pages so I won't be answering any posts for a while 😉

continued...

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 6:50pm

Sura Facility

HAARP Like Facility
Nizhniy Novgorod, Russia
+56° 7' 9.70", +46° 2' 3.66"

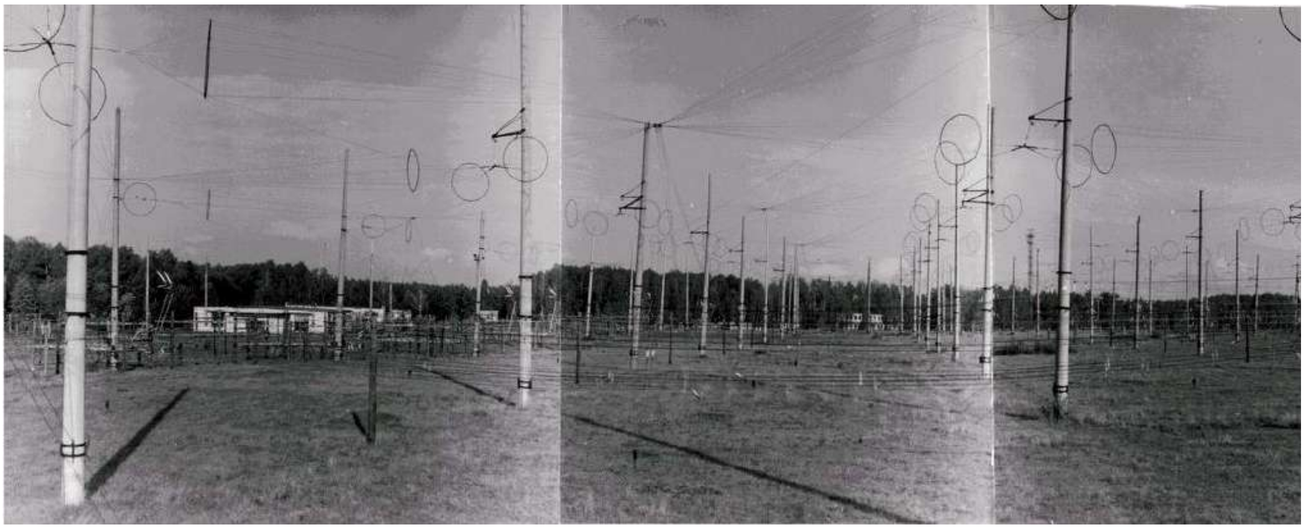

courtesy http://sura.nirfi.sci-nnov.ru/Photogalery/Photogalery.html

SURA Official Site Russian

The Sura Ionospheric Heating Facility, located near the small town of Vasilsursk about 100 km eastward from Nizhniy Novgorod in Russia, is a laboratory for ionosphere research. Sura is capable of radiating about 190 MW, effective radiated power (ERP) on short waves. This facility is operated by the radiophysical research institute NIRFI in Nizhny Novgorod. The Sura facility was commissioned in 1981. Using this facility, Russian researchers achieved extremely interesting results regarding the ionosphere behavior and discovered the effect of generation of low-frequency emission at the modulation of ionosphere current[1]. At the beginning, Soviet Defense Department mostly footed the bill. The American **HAARP** ionospheric heater is similar to the Sura facility. The **HAARP** project began in 1993.

Technical information

The frequency range of the heating facility is from 4.5 to 9.3 MHz. The facility consists of three 250 kW broadcasting transmitters and a 144 crossed dipole antenna-array with dimensions of 300 m x 300 m. At the middle of the operating frequency range (4.5 � 9.3 MHz) a maximum zenith gain of about 260 (~24 dB) is reached, the ERP of the facility is 190 MW (~83 dbW).

http://en.wikipedia.org/wiki/Sura_Ionospheric_Heating_Facility

пос. Васильсурск
Vasil'sursk

http://www.thelivingmoon.com/45jack_files/03files/SURA_Radar_Facility.html

Artificial Ionospheric Turbulence and Radio Wave Propagation (Sura/**HAARP**)

PDF File

Russian US cooperation

In March and May 2006 two
sessions of measurements in the framework of the SURA–**HAARP** experiment were carried out.

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 6:52pm

São Luiz Space Observatory

HAARP Like Facility
Cruzeiro Santa Bárbara, Sao Luis-MA, Brasil
-2° 35' 40.47", -44° 12' 35.90"

Coherent Back-Scatter Radar 50 MHz (RESCO)

The Coherent Back-Scatter Radar of 50 MHz (RESCO) was installed at the Space Observatory of São Luís / INPE, whose operation begun in August of 1998, is capable to accomplish measures of dynamics of the plasma of the electrojet and of bubbles equatorial ionospheric. This radar was projected to map the turbulence and the electromagnetic drift of the irregularities of short length scale (3 meters) in a height range that extends from ~90 km to ~1000 km of the equatorial ionosphere. Such plasma irregularities have big influence in the propagation trans-ionosphere of waves in a great frequency range, VHF to UHF, and, therefore, it influence all of the activities of space communications of the Brazilian tropical area. The formation, the development and the space distribution of these irregularities are highly sensitive to the space climatic change (in other words, "Space Weather") besides the convection processes and of the storms of the troposphere.

This radar resulted of the development and construction begun in INPE there are several years. It transmits signs pulse of high potency through an network antenna with 768 dipoles which allow to concentrate all the energy transmitted in only narrow beam radiation. The same antenna also captures the echo signs spread by the irregularities ionosphere. The transmitted maximum power (120 kW) it is reached through the use of a modular system of 8 transmitters in phase to maximize the transmitted energy. The operational control of the radar is made by a computer, which also accomplishes the acquisition, the treatment and processing 'on line' of the received data of the ionosphere. The data recorded are available also for the processing and analyze subsequent. This radar was already operated in several campaigns since 1998 and now it is collecting in a routine way data of the dynamics of the equatorial electrojet.

This radar, with the FCI radar of 30 MHz together offers great opportunities to the researchers of studying the peculiar phenomenon of the equatorial area. These, beside the radars of Peru (Jicamarca), of the India (Thumba) and of the Indonesia, are some of the few radars of this type that exist in the world around of the magnetic equator. Due to the peculiar configuration of the geomagnetic field, the Brazilian equatorial area have characteristics very different from the other areas. It was for this reason that NASA of the USA, in collaboration of INPE, accomplished in Alcântara in 1994 the campaign GUARÁ when 26 rockets were thrown (in the period of September-October) to study the equatorial electrojet and the bubbles ionosphere. In this campaign were used another radar similar to the radar RESCO (which was brought of the USA), Digissonde (which provided the diagnoses of the ionosphere) and of the magnetometers operated by INPE in the Space Observatory of São Luís. The radar RESCO, that is now in a phase of technological improvement, offers great potential to promote researches of the environment of the Brazilian equatorial area.

Rua Horto Florestal, 100, Cruzeiro Santa Bárbara, Sao Luis-MA - Brasil

SOURCE: São Luiz Space Observatory
http://www.dae.inpe.br/slz/english/resco.html

http://www.thelivingmoon.com/45jack_file....bservatory.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 6:54pm

Millstone Hill Radio Observatory

EISCAT Like Facility
Westford, Massachusetts
+42° 37' 09.25", -71° 29' 28.49"

Millstone Hill

Millstone Hill, Westford, Massachusetts, USA (Longitude: W Latitude: N)

The Millstone Hill Steerable Antenna, or MISA, is a fully steerable dish antenna, 46 meters in diameter, designed by the Stanford Research Institute (SRI) in 1959. It is currently located at MIT Haystack Observatory in Westford, Massachusetts. It was originally installed at the Sagamore Hill Radio Observatory in Hamilton, Massachusetts in 1963. The antenna operated at that location until 1978, at which time it was relocated to Millstone Hill. Since that time it has been primarily used as a UHF radar antenna to provide measurements of the near space environment using the incoherent scatter radar technique. It is one of two surviving dish antennas of this type in the world with the other antenna being located at the Stanford University radio science field site in Stanford, California. MISA is used to provide wide radar coverage in latitude and longitude.

MISA is a broad-based observatory capable of addressing a wide range of atmospheric science investigations. The incoherent scatter radar facility at Millstone Hill has been supported by the National Science Foundation since 1974 for studies of the earth's upper atmosphere and ionosphere. During this time the facility has evolved from a part-time research operation sharing radar cooling and power supply elements with the M.I.T. Lincoln Laboratory Millstone satellite tracking radar, to a separately funded, operationally independent system dedicated to upper atmospheric research. The scientific capability of the Millstone Hill facility was greatly expanded in 1978 with the installation of a fully-steerable 46 meter antenna to complement the 67 meter fixed zenith pointing dish.

The favorable location of Millstone Hill at sub-auroral latitudes combined with the great operational range afforded by the steerable antenna permit observations over a latitude span encompassing the region between the polar cap and the near-equatorial ionosphere. Since 1982 the Haystack Observatory Atmospheric Sciences Group has been supported for operating the Millstone Hill research radar as a part of the incoherent scatter radar chain and for associated studies of the auroral and sub-auroral ionosphere and thermosphere. The meridional radar chain extends from Sondre Stromfjord, Greenland through Millstone Hill at mid-latitudes, beyond Arecibo at low latitudes, to the Jicamarca facility at the magnetic equator in Peru. The radar chain forms an integral part of the NSF-supported CEDAR (Coupling, Energetics, and Dynamics of Atmospheric Regions) observing network and Millstone Hill observations and analysis have contributed extensively to the successes of the CEDAR initiative.

The Millstone Hill Radar uses Thomson backscatter from ionospheric electrons to deduce height- and time-resolved plasma drift velocities, electron and ion temperatures, electron densities, ion composition, and ion-neutral collision frequencies. These parameters provide further information about the neutral gas, neutral temperatures and winds, and electric fields present in the medium. The incoherent scatter technique provides observations of many of these parameters over an altitude range extending from less than 100 km to a thousand kilometers or more. Methods have been developed that allow these measurements to be made with an altitude resolution of hundreds of meters. The complete steerability of the radar allows horizontal gradients and structure to be examined along with vertical variations.

The Observatory is the premier scientific facility in the world for studying the equatorial ionosphere. It has a 2-MW transmitter and a main antenna with 18,432 dipoles covering an area of nearly 85,000 square meters.

SOURCE: Wikipedia Millstone Hill
http://en.wikipedia.org/wiki/Millstone_Hill

Related Links:
MIT Haystack Observatory
http://www.haystack.mit.edu/

Credit: Haystack Observatory

http://www.thelivingmoon.com/45jack_file....bservatory.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 6:56pm

The first American '**HAARP**'

Platteville Atmospheric Observatory

HAARP Like Facility
+40° 10' 54", -104° 43' 30"

History
of the
Platteville Atmospheric Observatory

Image courtesy Colorado University

The Platteville Atmospheric Observatory was envisioned in 1962 by what is now the Institute of Telecommunication Sciences (ITS), a part of the National Telecommunications and Information Administration (NTIA), as a site for high-powered radio experiments. While the initial experiment, that took place in 1968, studied over-the-horizon radar, the majority of later experiments used high power radio waves to modify the ionosphere in a process that is sometimes called ionospheric heating because it raises the electron temperature in the ionosphere.

The ionospheric heater was still used until 1984, when the last ionospheric experiments were performed. In the same year, the transmitters were loaned to the Office of Naval Research and sent to HIPAS in Alaska where they are still used.

With the removal of the transmitters, the focus of the facility changed to smaller-powered observation of the atmosphere rather than modifying it. In 1988 the 404 MHz RASS was installed and the ownership of the facility was transferred from NTIA to NOAA-ERL.

SOURCE: Platteville Atmospheric Observatory
http://grison.colorado.edu/Radar_Stations/Platteville/Platteville.html

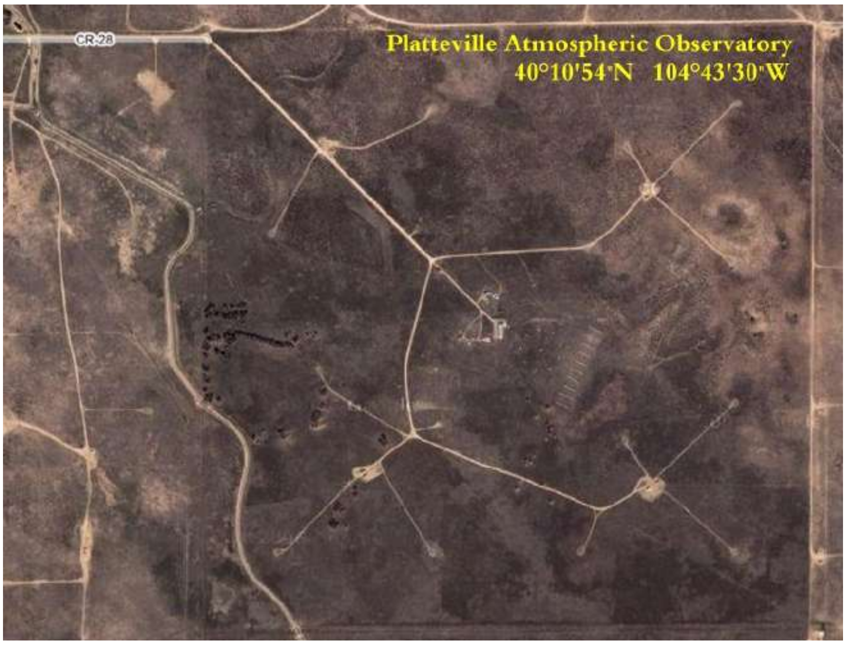
CR-28
Platteville Atmospheric Observatory
40°10'54"N 104°43'30"W

http://www.thelivingmoon.com/45jack_file....bservatory.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 6:58pm

High Power Auroral Stimulation Observatory

HIPAS
Fairbanks Alaska
64° 52' 19" N 146° 50' 33" W

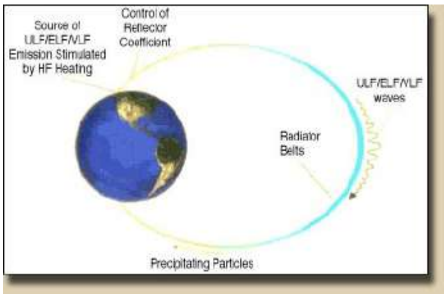
Source of ULF/ELF/VLF Emission Stimulated by HF Heating
Control of Reflector Coefficient
ULF/ELF/VLF waves
Radiator Belts
Precipitating Particles

HIPAS is located 30 miles Northeast of Fairbanks Alaska; in the small community of Two Rivers. It occupies 120 acres of land and has six buildings. The facility is located at: latitude and longitude. The facility operates year-round.

The HIPAS facility is engaged in the study of the Ionosphere through the use of high power radio transmission as well as a state-of-the-art LIDAR (LIght Detection And Ranging) facility.

The Heater system consists of 8 transmitters capable of conducting amplitude modulation of 100 Hz - 20 kHz and phase modulation of 0 -20 kHz. Each transmitter can transmit up to 150kW at 2.85 or 4.53 MHz on CW mode.

The Heater antenna system consists of a circular array of 8 crossed dipoles, copper wire ground-planes and resonant triaxial baluns.

The LIDAR facility consists of a 2.7 meter LMT (Liquid Mirror Telescope) with a 4.5 meter focal length as well as 6 state-of-the-art lasers.

SOURCE: HIPAS Home Page
http://www.hipas.alaska.edu/hipasweb/hipas.htm
Link invalid... Internet Archive Page
http://web.archive.org/web/20080212193205/http://www.hipas.alaska.edu/hipasweb/hipas.htm

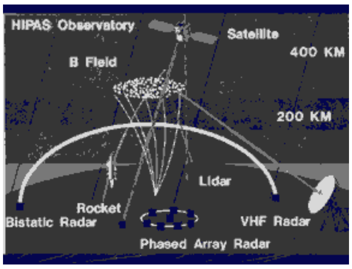
HIPAS Observatory
Satellite
400 KM
B Field
200 KM
Lidar
Rocket
Bistatic Radar
VHF Radar
Phased Array Radar

Courtesy HIPAS

http://www.thelivingmoon.com/45jack_file....bservatory.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 6:59pm

European Incoherent Scatter Scientific Association

EISCAT
Ramfjordmoen, Near Tromsø, Norway
+69° 35' 10.94", +19° 13' 20.89"

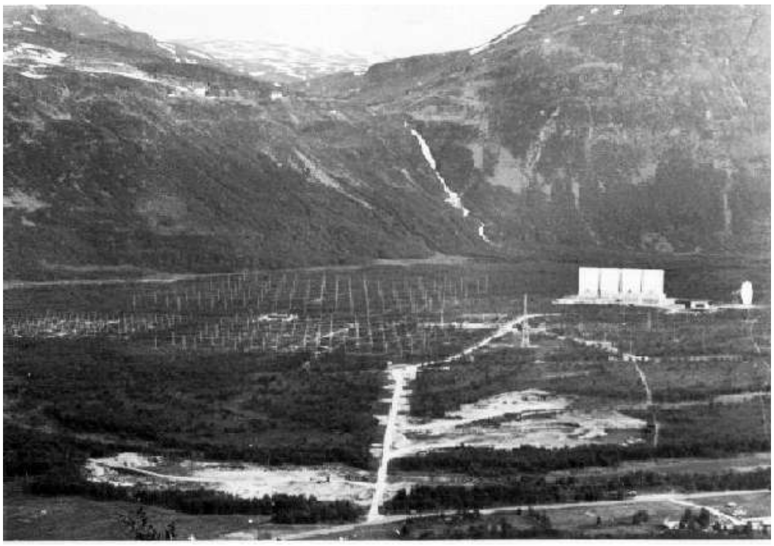

The EISCAT Scientific Association is an international research organisation operating three incoherent scatter radar systems, at 931 MHz, 224 MHz and 500 MHz, in Northern Scandinavia. It is funded and operated by the research councils of Norway, Sweden, Finland, Japan, France, the United Kingdom and Germany (collectively, the EISCAT Associates).

EISCAT (European Incoherent Scattter) studies the interaction between the Sun and the Earth as revealed by disturbances in the magnetosphere and the ionised parts of the atmosphere (these interactions also give rise to the spectacular aurora, or Northern Lights). The radars are operated in both Common and Special Programme modes, depending on the particular research objective, and Special Programme time is accounted and distributed between the Associates according to rules which are published from time to time.

One EISCAT transmitter site is located close to the city of Tromsø, in Norway, and additional receiver stations are located in Sodankylä, Finland, and Kiruna, Sweden. See an animation that shows the basic operation. The EISCAT Headquarters are also located in Kiruna. 1996 the EISCAT Scientific Association constructed a second incoherent scatter radar facility, the EISCAT Svalbard Radar, near Longyearbyen on the island of Spitsbergen, far to the North of the Norwegian mainland.

The Incoherent Scatter Radar technique requires sophisticated technology and EISCAT engineers are constantly involved in upgrading the systems.

In addition to the incoherent scatter radars, EISCAT also operates an Ionospheric Heater facility at Ramfjordmoen (including a Dynasonde) to support various active plasma physics experiments in the high latitude ionosphere.

SOURCE ARTICLE
http://www.eiscat.se:8080/about.html

Tromsø 931 MHz Ultra High Frequency Radar [facts]
These signals are also received at Kiruna and Sodankylä

Tromsø 224 MHz Very High Frequency Radar [facts]
We have also a heating facility and a Dynasonde

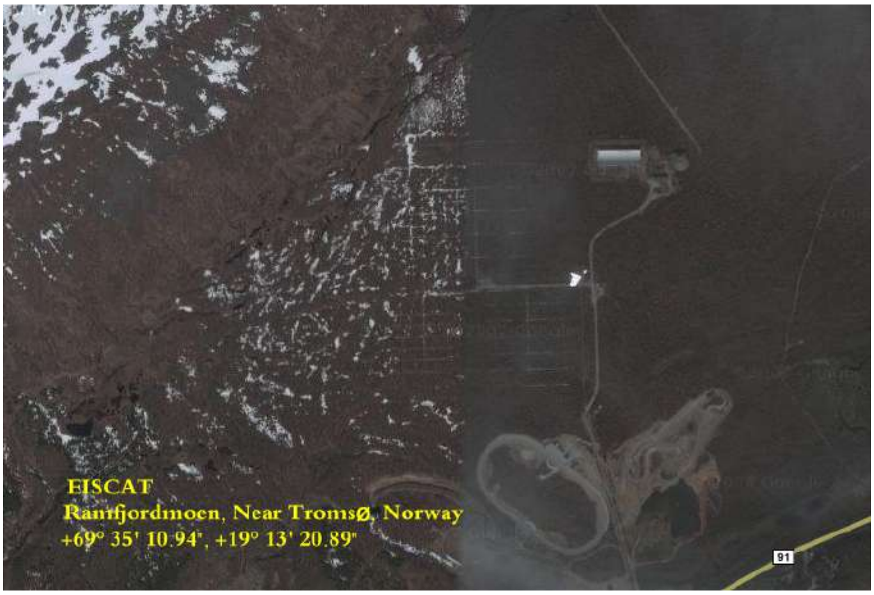

http://www.thelivingmoon.com/45jack_file....mso_Norway.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:01pm

Nerc MST Radar Facility

NMRF
HAARP Like Facility
+52° 25' 28.26", -4° 00' 19.59"
Capel Dewi, Carmarthenshire, Wales, United Kingdom
near Aberystwyth, Wales, UK

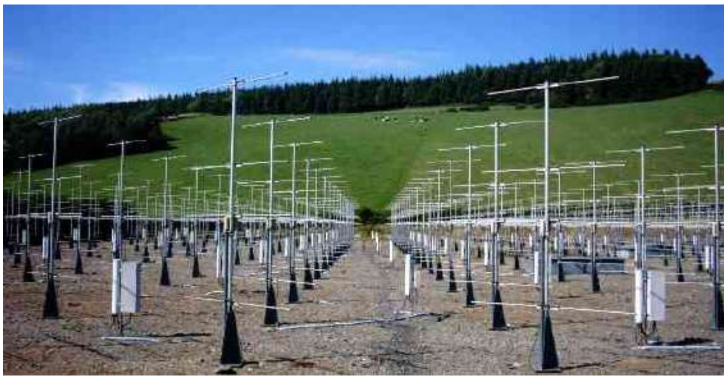

This one is different but included anyway

The Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar at Aberystwyth

SOURCE: NERC
http://mst.nerc.ac.uk/index.html

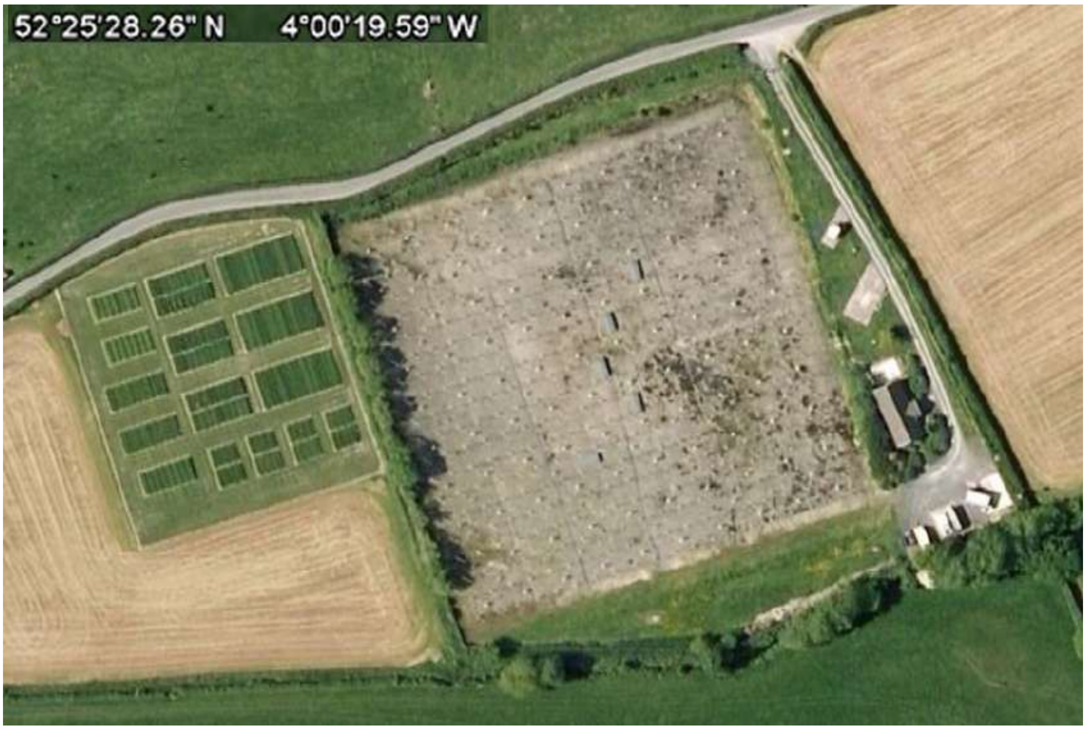

http://www.thelivingmoon.com/45jack_files/03files/NERC_MST_Radar_Facility.html

---

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:03pm

National MST Radar Facility

NMRF
HAARP Like Facility
+13° 27' 26.68", +79° 10' 30.74"
Gadanki, near Tirupati, in southern Andra Pradesh, India

I found this one on Google Earth so Pegasus has first find on that one 😁

Visiting the National MST Radar Facility in Gadanki, near Tirupati, in southern Andra Pradesh, India

ABOUT NATIONAL MST RADAR FACILITY

Indian scientists have carried out pioneering research work in the fields of astronomy and astrophysics, solar/interplanetary medium, earth's upper atmosphere/ionosphere, aeronomy/middle atmosphere and weather/climate phenomena. The nationally coordinated Indian Middle Atmosphere Programme (IMAP) was implemented during the period 1982-89 with well focussed campaign experiments with ground based, balloon, rocket and satellite based techniques. The IMAP programme led to the decision to conduct in-depth studies of atmosphericdynamical phenomena by developing a versatile ground based radar technique...

...The MST Radar is a state of the art instrument capable of providing estimates of atmospheric parameters with very high resolution on a continuous basis which are essential in the study of different dynamical processes in the atmosphere. It is an important research tool in the investigation of prevailing winds, waves (including gravity waves) turbulence, and atmospheric stability & other mesoscale phenomena . A reliable three dimensional model of the atmosphere over the low latitudes improves our understanding of the climatic and weather variations...

Establishment of NMRF

Attaching great importance to the scientific utilisation of the Indian MST Radar, the Government of India decided to create an autonomous Scientific Society called the National MST Radar Facility (NMRF). This society is affliated to the Department Of Space. The NMRF was registered on January 11, 1993 under the Indian Societies Act 1860.
This society is administered by a Governing Council under the chairmanship of Dr. K. Kasturirangan, Secretary DOS with, Director , NMRF as the member secretary . The present Director of NMRF is Prof D. Narayan Rao .The Governing Council consists of other eminent scientists, representatives of the National Laboratories and some of the funding agencies.The Governing Council sets broad policy guidelines to ensure the effective scientific utilisation of the facility, supported by a Scientific Advisory Commitee & a Finance Commitee.

Location Of NMRF

The scientific requirements dictated that the Indian MST Radar should be located preferably below 15 degrees North latitude. Hence after careful consideration of the various constraints, a site at Gadanki Village, near the temple town of Tirupati in the Chitoor district of Andhra Pradesh was selected for locating the Indian MST Radar . NMRF is located off the Chitoor -Tirupati main road in a picturesque landscape spreading over an area of about 42 acres. Regular train and bus services are operated between Tirupati and Bangalore/Madras. On request NMRF may provide transport between Tirupati and Gadanki.

SOURCE: The National MST Radar Facility (NMRF)
http://www.isro.org/nmrf/default.htm

http://www.thelivingmoon.com/45jack_file....r_Facility.html

Cherenkov Radiation around Antennas

Image courtesy ISRO

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:05pm

Jicamarca Radio Observatory

HAARP Like Facility
Lima, Peru
11° 57' 04.82" S 78° 52' 27.43" W

Image courtesy Colorado University

Introduction

The Jicamarca Radio Observatory is the equatorial anchor of the Western Hemisphere chain of incoherent scatter radar (ISR) observatories extending from Lima, Perú, to Søndre Strømfjord, Greenland. It is part of the Geophysical Institute of Peru (Instituto Geofísico del Perú, or IGP) and receives the majority of its financial support from the National Science Foundation of the U.S. through a Cooperative Agreement with Cornell University.

The Observatory is the premier scientific facility in the world for studying the equatorial ionosphere. It has a 2-MW transmitter and a main antenna with 18,432 dipoles covering an area of nearly 85,000 square meters.

SOURCE: Jicamarca Radio Observatory
http://jicamarca.ece.cornell.edu/overview.html

Piura facility: Jicamarca Radio Observatory

The Observatory is about a half-hour drive inland (east) from Lima, Peru at a geographic latitude of 11.95° south and a longitude of 76.87° west. The altitude of the Observatory is about 500 m ASL.

It is about 10 km from the Carretera Central, the main highway east in Peru.

The Jicamarca Radio Observatory was built in 1960-61 by the Central Radio Propagation Laboratory (CRPL) of the National Bureau of Standards (NBS).

Station Database
http://jro.igp.gob.pe/english/database/database_en.htm

Image Courtesy Cornell University

The Jicamarca Radio Observatory - Cornell University
http://jicamarca.ece.cornell.edu/

The 49.92 MHz ISR is the principal facility of the Observatory. The radar antenna consists of a large square array of 18,432 half-wave dipoles arranged into 64 separate modules of 12 x 12 crossed dipoles. Each linear polarization of each module can be separately phased (by hand, changing cable lengths), and the modules can be fed separately or connected in almost any desired fashion. There is great flexibility, but changes cannot be made rapidly. The individual modules have a beam width of about 7 deg, and the array can be steered within this region by proper phasing. The one way half power beam width of the full array is about 1.1 deg; the two way (radar) half power width is about 0.8 deg. The frequency bandwidth is about 1 MHz. The isolation between the linear polarizations is very good, at least 50 dB, which is important for certain measurements. Since the array is on the ground and the Observatory is the only sign of man in a desert region completely surrounded by mountains, there is no RF interference. The transmitter consists of four completely independent modules which can be operated together or separately. Only two of these modules are currently in operation. Each can deliver ~1.5 MW peak power, with a maximum duty cycle of 6%, and pulses as short as 0.8-1.0 µs. Pulses as long as 2 ms show little power droop; considerably longer pulses are probably possible. Two additional modules with the same properties will eventually be available. The third is more than 90% complete; the fourth is well advanced but its completion will require additional funding. The drivers of the main transmitter can also be used as transmitters for applications requiring only 50- 100 KW of peak power...

SOURCE: Jicamarca Radio Observatoryhttp://www.kurasc.kyoto-u.ac.jp/radar-group/local/isr/jicamarca/overview.html

MUCH BIGGER THAN **HAARP**
Main Array

Visit the Jicamarca Radio Observatory homepage
http://jro.igp.gob.pe/

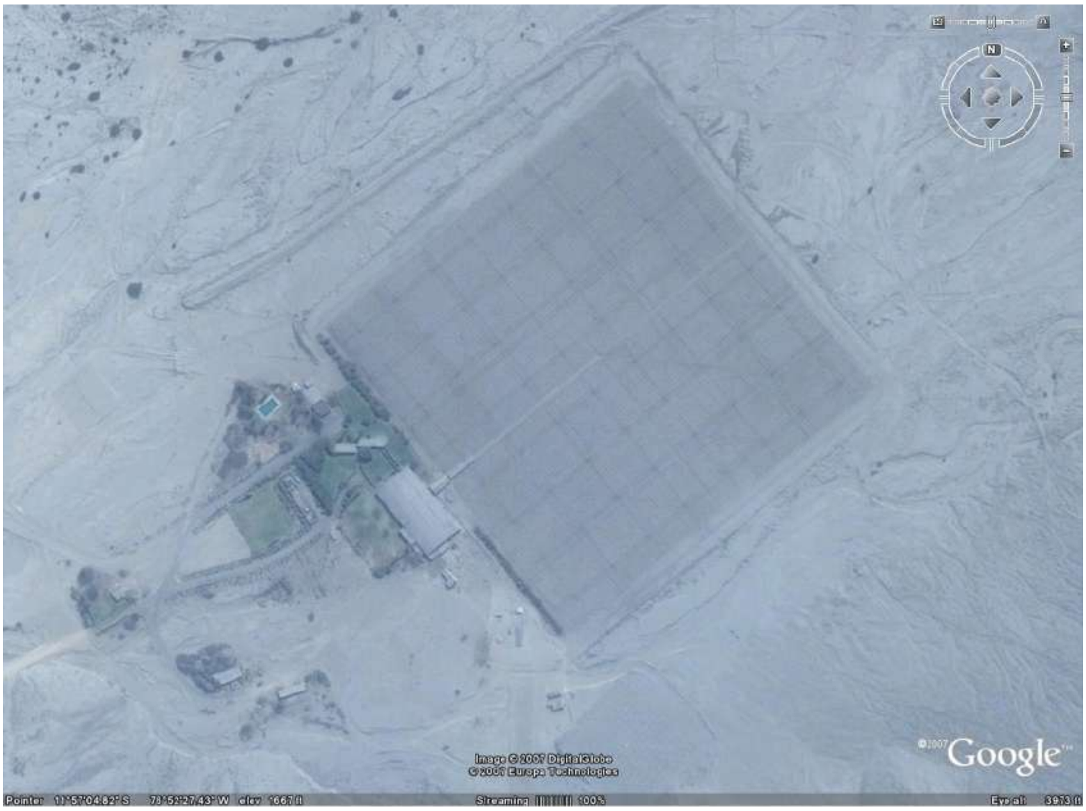
Image © 2007 DigitalGlobe
© 2007 Europa Technologies
©2007 Google™
Pointer 11°57'04.82" S 78°52'27.43" W elev 1667 ft
Streaming 100%
Eye alt 3973 ft

http://www.thelivingmoon.com/45jack_file....bservatory.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:07pm

Jindalee Operational Radar Network

JORN
JP 2025
Laverton, West Australia
-28° 19' 36.29", +122° 0' 18.84"

Two Part System

CAND98/0153-03
JORN PROJECT RECEIVER SITE, LAVERTON W.A.
PIC BY CPL DAVE BROOS, DEFENCE PUBLIC AFFAIRS.

Projects
JP 2025 - Jindalee Operational Radar Network (JORN)

The JORN project arose out of extensive research undertaken by the Defence Science and Technology Organisation (DSTO) into over-the horizon radar (OTHR) beginning in the early 1970s. As part of the 1987 Defence White Paper, the Government placed a high priority on wide area surveillance of the north and north western approaches to Australia and OTHR was seen to be the most cost effective solution. As a consequence, in December 1990, the Government approved the design and construction of JORN.

The Jindalee Operational Radar Network (JORN) consists of two OTHR, one near Longreach, Qld. and the other near Laverton, WA, jointly operated from the JORN Coordination Centre (JCC) at RAAF Base Edinburgh, SA by No 1 Radar Surveillance Unit. The radars are an advanced development of the Australian designed Jindalee radar at Alice Receiver Site, Laverton WA - click on image to enlargeSprings which is in operational use as well as being a research and development facility used by DSTO for ongoing OTHR improvement. JORN radars are capable of all weather detection of air and surface targets inside an arc of up to 3,000 km range extending from Geraldton in the west around to Cairns in the east. JORN makes a crucial contribution to broad area surveillance of Australia's strategically important northern approaches.

SOURCE:
Australian Department of Defence
Defence Materiel Organization
http://www.defence.gov.au/dmo/esd/jp2025/jp2025.cfm

Electronic Warfare & Radar Division - DSTO, Australia
http://www.dsto.defence.gov.au/page/3603/

CAND98/0153-25
JORN PROJECT RECEIVER SITE ANTENNA ARRAY, LAVERTON W.A.
PIC BY CPL DAVE BROOS, DEFENCE PUBLIC AFFAIRS.

Courtesy Australian Department of Defence

Courtesy Australian Department of Defence

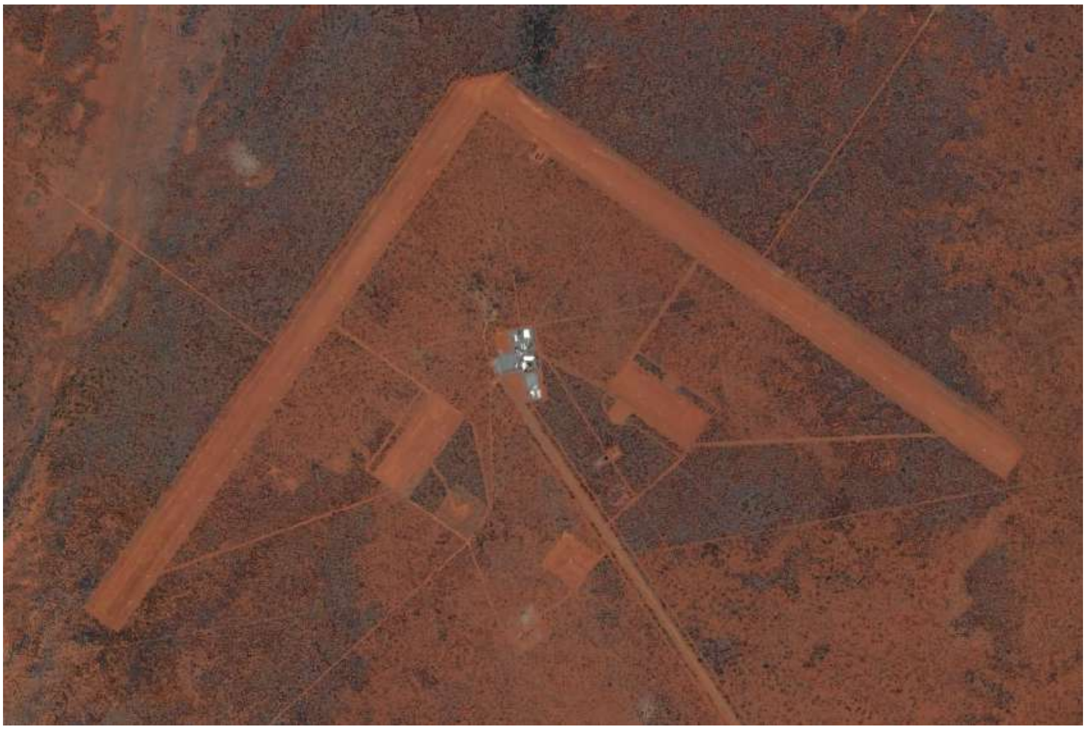

http://www.thelivingmoon.com/45jack_file....k_Laverton.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:09pm

Jindalee Operational Radar Network

JORN
Longreach, Queensland, Australia
-23° 39' 29.53", +144° 8' 49.58"

JORN Part Two

Communications '92: Communications Technology, Services and Systems; Getting It All Together

The Jindalee Operational Radar Network Communications and the Operational Concept
Nicholson, P1; Cameron, A2

Abstract: The operational concept of the Jindalee over-the-horizon operational radar network (JORN) is the centralised control and co-ordination of remote sensors. The radar sites are in Laverton, WA and Longreach, Queensland, while the co-ordination centre is situated in Adelaide, South Australia. An extensive communications network is needed to control the radars and their associated frequency management systems, transfer partly processed data for final analysis at the co-ordination centre, and pass track information to the command support systems of the Australian Defence Force and other users. The principle of operation, configuration and concept of the Jindalee project are briefly outlined to provide the context of the communications requirement. The communications infrastructure to support this operational concept is then described together with the main factors which have influenced the design of JORN communications.

Papers Available Here
http://search.informit.com.au/documentSummary;dn=563982181256363;res=IELENG

Consultancy Projects

Jindalee Operational Radar Network

RayTec Consulting has since its inception offered sub-contract services on the JORN project across a broad spectrum of Systems Engineering disciplines from Requirements Analysis and Design, Verification and Validation to
Integration and Test.

SOURCE: http://www.raytec.com.au/wp/projects

Electronic Warfare & Radar Division

High frequency surface wave radar. Credit DSTO

Electronic Warfare & Radar Division

Electronic Warfare & Radar Division provides scientific leadership and support to the Australian Defence Organisation on the exploitation of the electromagnetic spectrum to enhance the performance of our own sensors, weapons, platforms and command systems, together with the ability to destroy the effectiveness of adversarial systems.

Weapons Systems Division

Weapons Systems Division provides scientific leadership and support covering all aspects of weapon systems - including sensors, guidance, propulsion and warheads, and their integration into combat platforms and command and control systems.

Command, Control, Communications & Intelligence Division

Command, Control, Communications & Intelligence Division provides scientific leadership and support for Defence command, intelligence, communications, and business processes, at both the operational and theatre levels of command. Support to the Australian Defence Organisation includes Information Operations with special capabilities in Information Security and Digital Forensics; Communications with special capabilities in Satellite Communications, Mobile Networks and Network Management; Intelligence Processing and Analysis with special capabilities in signals analysis, communications analysis, automated fact extraction, and speech processing. The Division has organised its work program to have a strong emphasis on support achieving the goals as outlined in the Network Centric Warfare roadmap.

Intelligence, Surveillance & Reconnaissance Division

Intelligence, Surveillance & Reconnaissance Division provides scientific leadership and support for strategic intelligence, surveillance and reconnaissance systems, with a focus on the needs of the intelligence community.

Land Operations Division

Land Operations Division provides scientific leadership and support to the Land Force through structured and analytical approaches to capability development.

SOURCE: DSTO
http://www.dsto.defence.gov.au/page/3603/

http://www.thelivingmoon.com/45jack_file.... Longreach.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:12pm

[quote]*Originally posted by zorgon*

Poker Flat Research Range

Near Chatanika, Alaska
+65° 7' 23.90", -147° 28' 7.05"

Still looking for the actual array in this area but Poker Flats covers a wide terrain and the array is most likely hidden in the trees

Gate sign says NASA is lurking about and they mention the Aurora studies...

POKER
FLAT
RESEARCH RANGE
30 MI STEESE HWY FAIRBANKS, ALASKA
ADVANCE NOTICE REQUIRED FOR TOURS
POKER FLAT
RESEARCH RANGE
GEOPHYSICAL INSTITUTE UNIVERSITY OF ALASKA
FAIRBANKS, ALASKA 99775-0800
(907) 474-7015
THE BEST EQUIPPED SOUNDING ROCKET LAUNCH FACILITY IN THE NORTH, ESTABLISHED BY THE GEOPHYSICAL INSTITUTE OF THE UNIVERSITY OF ALASKA, FAIRBANKS. THIS FACILITY IS UNIQUELY DEDICATED TO STUDIES OF THE AURORA BOREALIS AND OTHER ATMOSPHERIC RESEARCH STUDIES FOR THE PAYING CUSTOMER SUCH AS THE NATIONAL AERONAUTICS AND SPACE ADMINISTRATION, THE UNITED STATES AIR FORCE, THE UNITED STATES ARMY, THE NATIONAL SCIENCE FOUNDATION, AND OTHERS.
FIRST LAUNCH FROM THIS FACILITY MARCH 1969
LATITUDE 65° 07' NORTH LONGITUDE 147° 28' WEST

HAARP likely not Primary Ionospheric array in Alaska by Guy Cramer
http://www.superforce.com/**haarp**/index.htm

How to hide a **HAARP** by Guy Cramer

This first photo shows **HAARP**

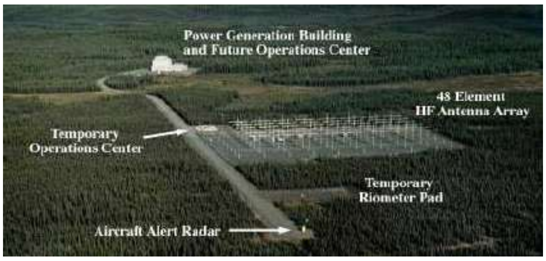
Power Generation Building
and Future Operations Center
48 Element
HF Antenna Array
Temporary
Operations Center
Temporary
Riometer Pad
Aircraft Alert Radar

The second photo shows **HAARP** with a camouflaged Array

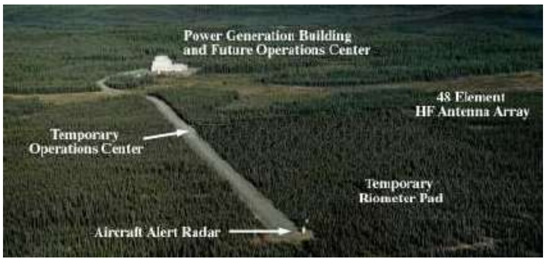
Power Generation Building
and Future Operations Center
48 Element
HF Antenna Array
Temporary
Operations Center
Temporary
Riometer Pad
Aircraft Alert Radar

The third image shows a completely concealed array, road and power generation center, to show how easy it is to hide a large facility in a location like Alaska.

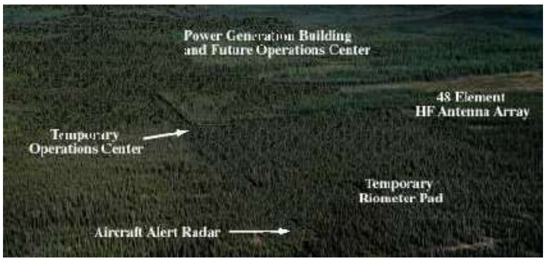
Power Generation Building
and Future Operations Center
48 Element
HF Antenna Array
Temporary
Operations Center
Temporary
Riometer Pad
Aircraft Alert Radar

"The above illustrations demonstrate the capability to effectively conceal large land areas containing hardware and buildings. In a similar initiative comprising the the most comprehensive infrastructure concealment program since World War II, the design team of Dr. Resnick, Lt. Col. Timothy R. O'Neill, PhD (U.S. Army, Ret.) and Mr. Guy Cramer have produced remarkable results. Using specially designed "Fractal Based" camouflage patterns in projects related to concealment of critical infrastructure under the auspices of the US Department of the Interior's Bureau of Land Management, the team continues to achieve desired objectives such as those shown below."

SOURCE: Guy Cramer Hyperstealth Tech
http://www.superforce.com/**haarp**/index.htm

We have found this one...

Chatanika Incoherent Scatter Facility

EISCAT Like Facility
Poker Flats
+65° 7' 1.34", -147° 27' 37.23"

Also found one small Array...
+65° 7' 55.61", -147° 27' 14.98"

And several 'suspicious' areas like this one

But Poker Flats is still a work in progress

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:20pm

Baikal-1 and IGR sites
Semipalatinsk-21, Kazakhstan

This one is a direct tie in with Remote Viewing success and Tesla technology.

It is said Russia has several **HAARP** like facilities, bt so far I have only found SURA

Back in the scalar weapon and Tesla research days there are several objects left in Russia that are very impressive. As they deal with antennas and as they are cool I will add them here as well...

You will also have to visit my pages at the bottom of each post as I am only using a few of the images and a little of the text... the full collection is on each page as I get new data...

First one is...

URDF-3 (Unidentified Research and Development Facility-3)

Baikal-1, Semipalatinsk, Kazakhstan
50°10'12.69"N, 78°22'36.84"W

This first appeared as a CIA satellite image

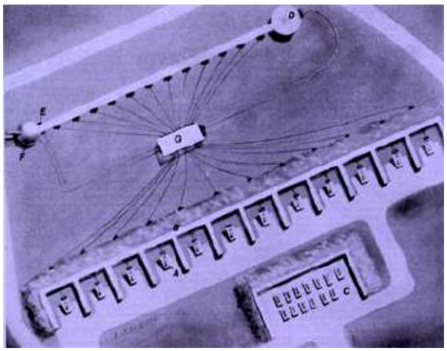

U.S. satellite reconnaissance photo of suspected Soviet beam weapon installation near Semipalatinsk.
Published July 28, 1980. (Courtesy Aviation Week & Space Technology)

Much has been written about the CIA search for Russia's Tesla Beam Weapon facility and it is even one of the biggest cases for the Remote Viewing Project Stargate that the CIA ran. One of the things that was a positive hit from the RV Team was the huge crane at the site...

Well that crane is STILL THERE...

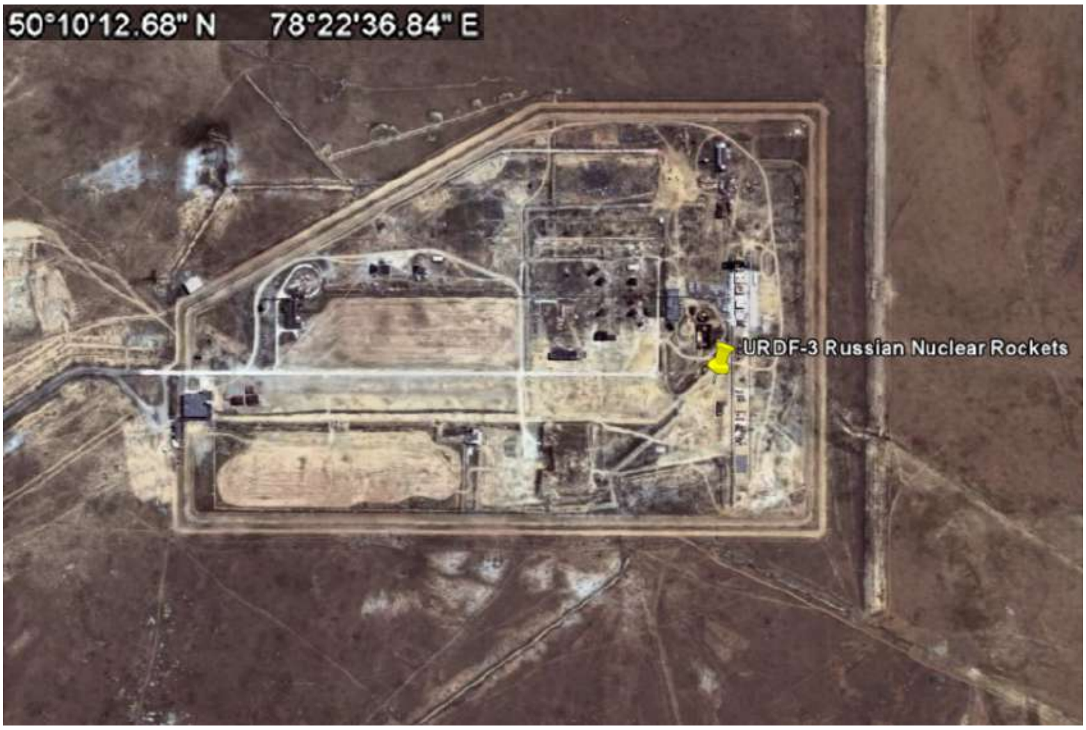

Space Nuclear Facility test capability at the Baikal-1 and IGR sites
Semipalatinsk-21, Kazakhstan

Hill, T. J.; Stanley, M. L.; Martinell, J. S.
Presented at the Nuclear Power Engineering in Space Nuclear Rocket Engines, Kazakhstan, Russia, 22-26 Sep. 1992

The International Space Technology Assessment Program was established 1/19/92 to take advantage of the availability of Russian space technology and hardware. DOE had two delegations visit CIS and assess its space nuclear power and propulsion technologies. The visit coincided with the Conference on Nuclear Power Engineering in Space Nuclear Rocket Engines at Semipalatinsk-21 (Kurchatov, Kazakhstan) on Sept. 22-25, 1992. Reactor facilities assessed in Semipalatinski-21 included the IVG-1 reactor (a nuclear furnace, which has been modified and now called IVG-1M), the RA reactor, and the Impulse Graphite Reactor (IGR), the CIS version of TREAT. Although the reactor facilities are being maintained satisfactorily, the support infrastructure appears to be degrading. The group assessment is based on two half-day tours of the Baikals-1 test facility and a brief (2 hr) tour of IGR; because of limited time and the large size of the tour group, it was impossible to obtain answers to all prepared questions. Potential benefit is that CIS fuels and facilities may permit USA to conduct a lower priced space nuclear propulsion program while achieving higher performance capability faster, and immediate access to test facilities that cannot be available in this country for 5 years. Information needs to be obtained about available data acquisition capability, accuracy, frequency response, and number of channels. Potential areas of interest with broad application in the U.S. nuclear industry are listed.

SOURCE: Harvard Abstracts
http://adsabs.harvard.edu/abs/1993STIN...9416773H

Here is the crane sketch from the CIA RV Team members

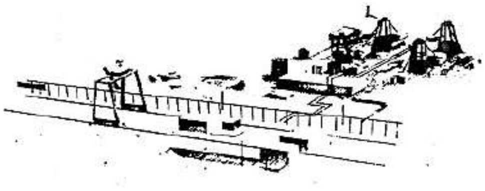

(a) TARGET SITE

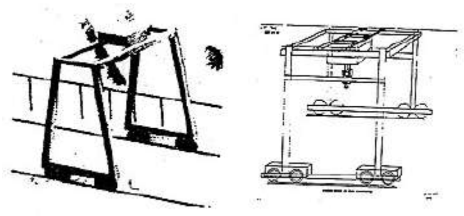

(b) CRANE COMPARISON

Figure 1: Comparison of Target Site and Drawing by Remote Viewer

573310 3-78 CIA

SECRET
NOFORN
WNINTEL

~~SECRET~~

Here is the large crane on Google Earth

Large Gantry Crane

Credit for the find and info gathering goes to
admiraltogo
Fort Smith, AR
Google Earth poster

This base was also the spot where the Russians tested their version of a Nuclear Rocket

http://www.thelivingmoon.com/45jack_file....Kazakhstan.html [/quote]

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:26pm

Strange Towers in a Russian Forest
Tesla Generators

Sychëvka, Moskovskaya Oblast' (Russia)
+55°55'26.15"N,36°49'10.97"

Yet another old Russian project tied into Tesla research... I have included these because they are related, and just plain cool 😎

Update 2007-12-02 09:59:49

"These were part of the experiments do by the SU with Teslas work towards power transmission and communication. Pictures all over the place on the internet. Nothing mysterious or new about it. Or perhaps its a secret installation for taking over the world. Take you pick."

Below is the entry gate from Google Earth images... the caption translates to...
"Isled. the center of the high energies"

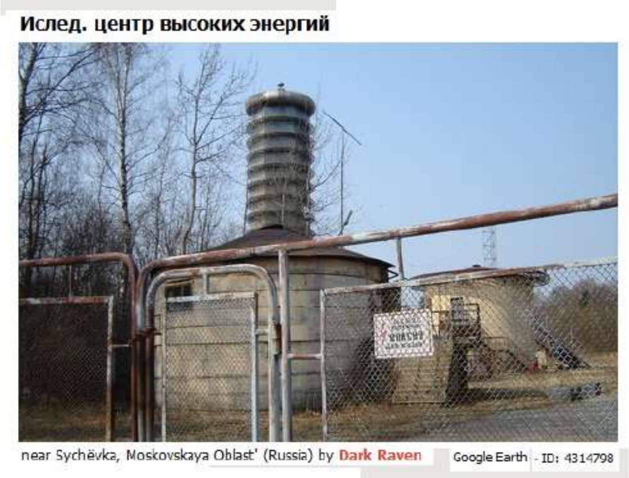
Иcлед. центр высоких энергий
near Sychëvka, Moskovskaya Oblast' (Russia) by Dark Raven
Google Earth - ID: 4314798

Test benches VNITS - ALL-UNION SCIENTIFIC RESEARCH INSTITUTE OF CEMENT VEI
(high-voltage scientific research center of All-Russian electrotechnical institute).

Satellite View of the Installation near Sychëvka, Moskovskaya Oblast' (Russia)

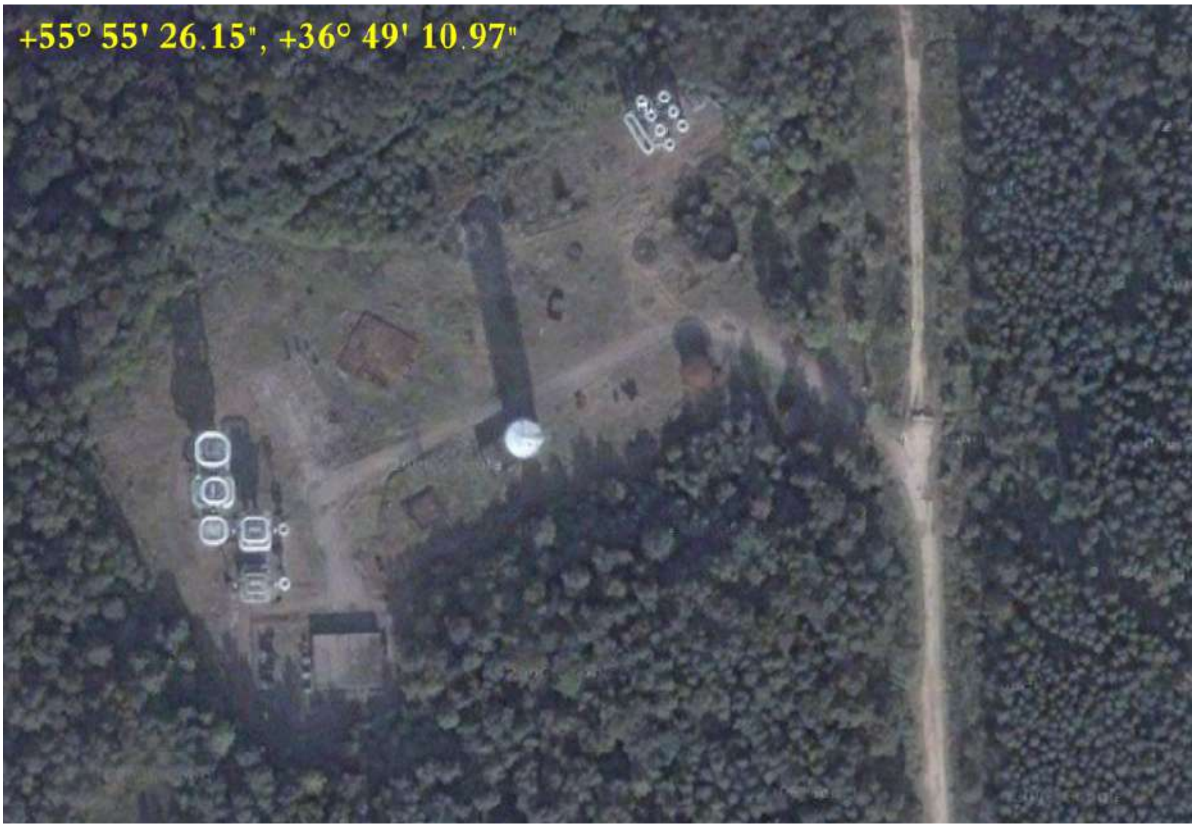
+55° 55' 26.15", +36° 49' 10.97"

http://www.thelivingmoon.com/47john_lear....ian_Forest.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:29pm

The Russian Woodpecker

Duga Radar Array, Chenobyl, Ukraine
51°18'20.17"N, 30°04'02.60"E

"Woodpecker" Duga radar array, Chenobyl, Ukraine by Necator

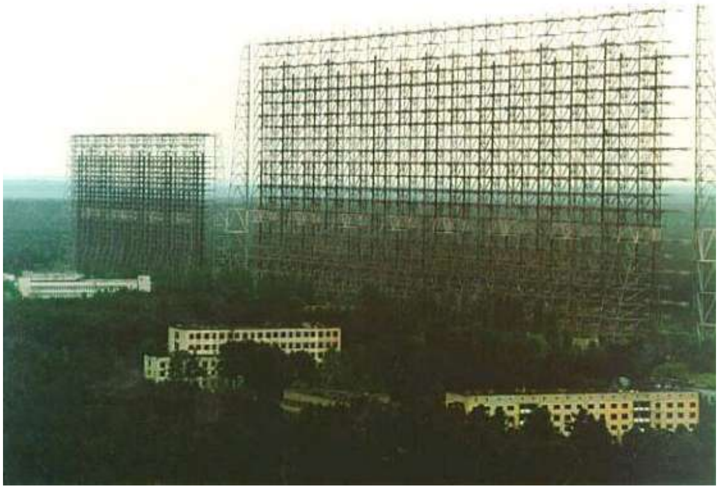

РЛС Дуга-1 Чернобыль-2 (Radar arch-1 Chernobyl-2) by DM.HANTER

History

The Soviets had been working on early warning radars for their anti-ballistic missile systems through the 1960s, but most of these had been line-of-sight systems that were useful for raid analysis and interception only. None of these systems had the capability to provide early-warning of a launch, which would give the defenses time to study the attack and plan a response. At the time the Soviet early-warning satellite network was not well developed, so work started on over-the-horizon radar systems for this associated role in the late 1960s.

The first experimental system, Duga-1, was built outside Mykolaiv in the Ukraine, successfully detecting rocket launches from Baikonur Cosmodrome at 2,500 kilometers. This was followed by the prototype Duga-2, built on the same site, which was able to track launches from the far east and submarines in the Pacific Ocean as the missiles flew towards Novaya Zemlya. Both of these radars were aimed east and were fairly low power, but with the concept proven work began on an operational system. The new Duga-3 systems used a transmitter and receiver separated by about 60 km.[2]

http://en.wikipedia.org/wiki/Russian_Woodpecker

Чорнобиль-2. Приймальні антени (Chernobyl-2. Receiving Antenna) by djcrok

51°18'20.17" N 30°04'02.60" E

MANY more pictures

http://www.thelivingmoon.com/45jack_file....ar_Ukraine.html

Main Stream Media coverage..

The Missing Secrets Of Nikola Tesla - Woodpecker Signal
Russia makes Aurora's with Woodpecker Array

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:33pm

China Research Institute of Radiowave Propagation (CRIRP)
HAARP Like Facility
Ionospheric Laboratory

Ionospheric Laboratory, Xinjiang (Sinkiang) Region
40°24'15.91"N, 93°38'09.74"E

So far I have only found ONE of China's Arrays and one EISCAT type dish.

Not much data to go on either but the name...
CRIRP
China Research Institute of Radiowave Propagation

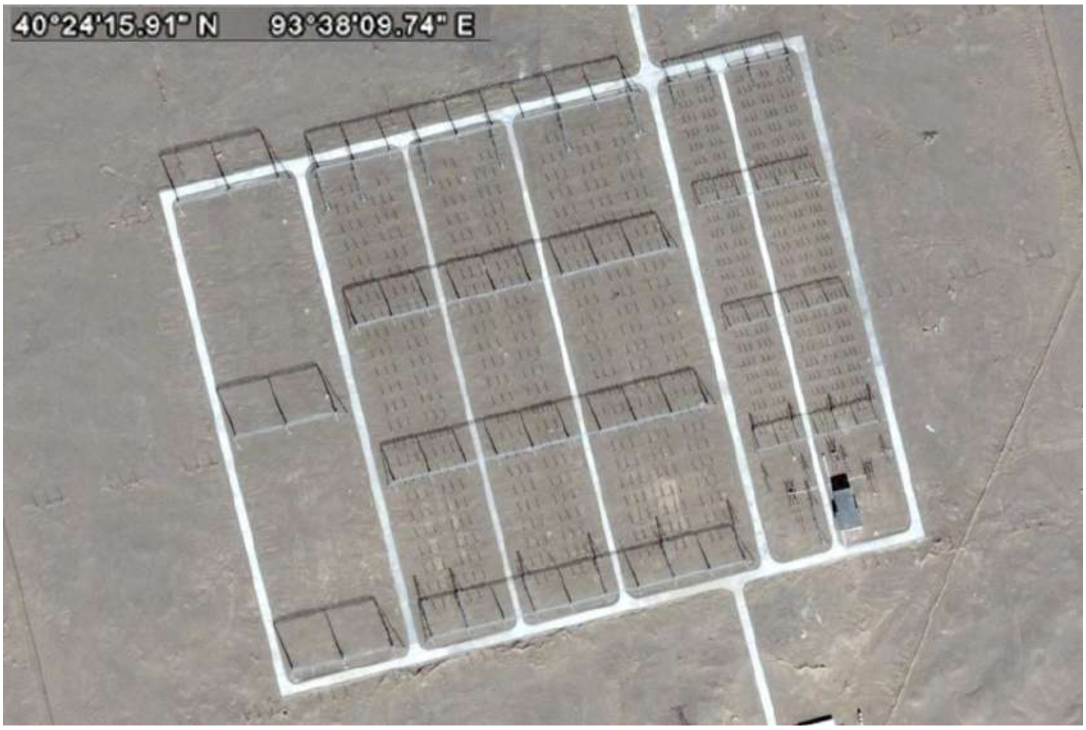

This facility is closest to Urumchi (Urumqi), Xinjiang Region, China. Actually the whole region is filled with odd markings and will be a whole thread on its own but not related to radio waves...

Ionospheric sounding network and data in China

Wu Jian Jiao Peinan Xiao Zuo Wan Weixing Liu Ruiyuan Zhao Zhengyu
LEME, China Res. inst. of Radiowave Propagation, Beijing;

This paper appears in: Antennas, Propagation and EM Theory, 2000. Proceedings. ISAPE 2000. 5th International Symposium on
Publication Date: 2000
On page(s): 688-691
Meeting Date: 08/15/2000 - 08/18/2000
Location: Beijing, China
ISBN: 0-7803-6377-9
References Cited: 10
INSPEC Accession Number: 6963766
Digital Object Identifier: 10.1109/ISAPE.2000.894880
Current Version Published: 2002-08-06

Abstract
Ionospheric sounding has been conducted routinely for more than 60 years in China. A complete network of ground-based sounding sites covers the Chinese subcontinent, including vertical and oblique sounding, GPS measurement of ionospheric TEC and scintillation, **VLF** receivers measuring the lower ionosphere etc. In this paper, we give a picture of the sounding network, equipment situation and data acquired with emphasis on vertical ionosonde network

SOURCE: IEEE
http://ieeexplore.ieee.org/xpl/freeabs_a....&isnumber=19345

The Ionospheric Sounding in China

The ionospheric sounding in China has a long history and has a well spread network, which is still keeping routine operation, providing a good background to do the ionospheric long-term prediction and short-term forecasting. The ionospheric sounding in China started in early 1940s (Wu et al., 2002). Fig.1 shows the ionospheric sounding network in China. The sounding equipments and operation periods are listed in Table 1. Among them 11 ionosonde stations are still in operation in China mainland. The data at integer UT hours are sent to forecasting center in Beijing twice a day through Internet.

There is also a daily exchange of ionospheric data with Russian (for 4 stations) and with Australia (for 4 stations) respectively. A method of predicting the ionospheric F2 layer in the Asia and Oceania Region (AOR Method [Sun X.R., 1987]) was adopted as a regional ionospheric long-term prediction method in China and its surrounding area. Then this was cooperated with the International Reference Ionosphere and became the Reference Ionosphere in China (CRI) [Liu et al., 1994].

http://www.irf.se/~christer/China_IONO.doc

Preliminary studies on ionospheric forecasting in China and its surrounding area

R. Liua, Corresponding Author Contact Information
E-mail The Corresponding Author, Z. Xua, J. Wub, S. Liua, B. Zhanga and G. Wangb

Polar Research Institute of China, 451 Jinqiao Road., Shanghai 200136,
China bChina Research Institute of Radiowave Propagation, Xinxiang 453003, China

Abstract

The ionospheric sounding in China has a long history and has a well spread network, which is still keeping routine operation. The autocorrelation method is adopted for the short-term forecasting of ionospheric characteristics. The performances of the forecasts at Chongqing have been examined for different combination of parameters and algorithms by estimating the prediction errors. Preliminary results show that for predictions of more than 10 h ahead the “at once” method with f0F2 is preferable. For predictions of less than 10 h ahead the “iterations” method with View the MathML source is the best. A corrected method of the International Reference Ionosphere used in China region (the CRI model) is described in this paper. By introducing an effective ionospheric index Ice into the CRI model the regional forecasting could be realized.

SOURCE: Science Direst

Home Address:
National Key Lab of Electromagnetic Environment, China Research Institute of Radiowave Propagation (CRIRP), P.O. Box 6301, Beijing 102206, China,

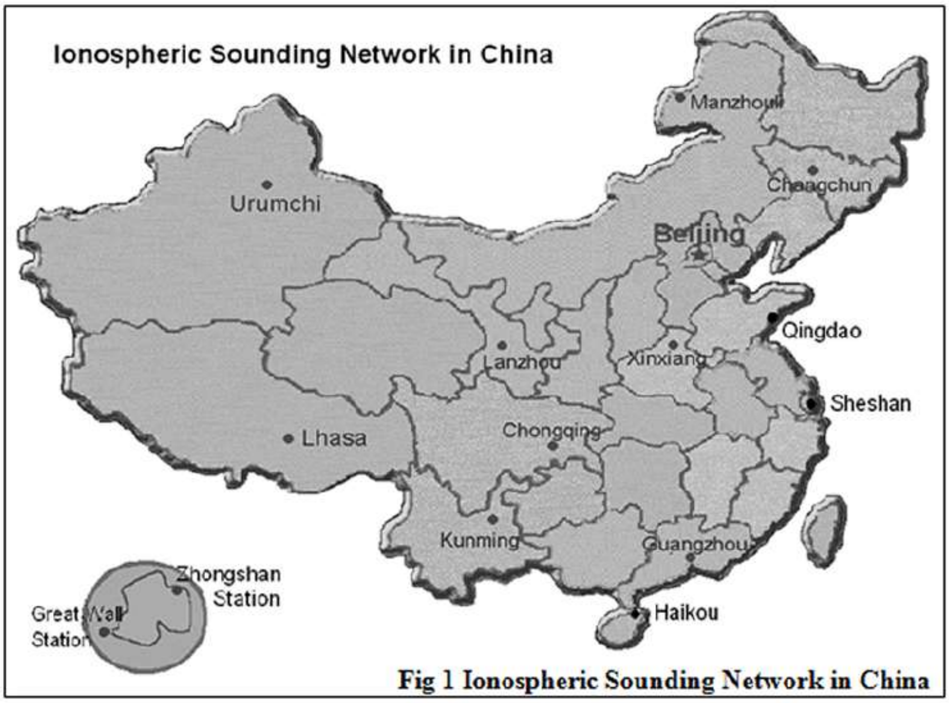

Fig 1 Ionospheric Sounding Network in China

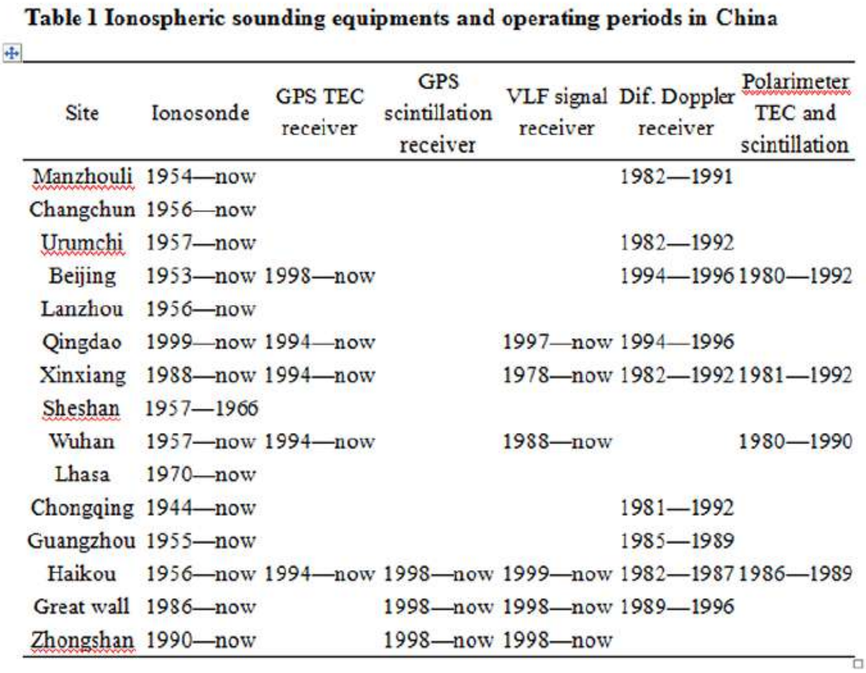

Table 1 Ionospheric sounding equipments and operating periods in China

Site	Ionosonde	GPS TEC receiver	GPS scintillation receiver	VLF signal receiver	Dif. Doppler receiver	Polarimeter TEC and scintillation
Manzhouli	1954—now				1982—1991	
Changchun	1956—now					
Urumchi	1957—now				1982—1992	
Beijing	1953—now	1998—now			1994—1996	1980—1992
Lanzhou	1956—now					
Qingdao	1999—now	1994—now		1997—now	1994—1996	
Xinxiang	1988—now	1994—now		1978—now	1982—1992	1981—1992
Sheshan	1957—1966					
Wuhan	1957—now	1994—now		1988—now		1980—1990
Lhasa	1970—now					
Chongqing	1944—now				1981—1992	
Guangzhou	1955—now				1985—1989	
Haikou	1956—now	1994—now	1998—now	1999—now	1982—1987	1986—1989
Great wall	1986—now		1998—now	1998—now	1989—1996	
Zhongshan	1990—now		1998—now	1998—now		

http://www.thelivingmoon.com/45jack_file....ropagation.html

Zhong Shan Antarctic Polar Station (China)

69° 22' 23.63" S 76° 22' 19.11" E

Introduction

Of the two Chinese scientific stations in Antarctic, the one in Zhong Shan is closely conjugated with the Svalbard area in the northern hemisphere. This makes it a valuable place for coordinated measurements both with the EISCAT Svalbard Radar (ESR) and the Polar satellite. In 1991 a cooperation agreement was signed between the Ionospheric Laboratory of the China Research Institute of Radiowave Propagation (CRIRP) in Xinxiang, China, and the Department of Physical Sciences, University of Oulu, Finland, to build an auroral photometer system for the Zhong Shan station (Kaila et al, 1997). The multichannel scanning photometer agreed upon was constructed in Oulu by April 1995, and delivered to China in May. Measurements were supposed to start in March 1997, to be continued during the local winter time until October (the same schedule of measurements is planned for every year). However, due to problems mentioned above, the current status of the system is unknown!

The Zhong Shan Station, Antarctica
http://www.thelivingmoon.com/45jack_file....Zhong_Shan.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:36pm

Sheshan, Shanghai, China

EISCAT Like Facility
31°5′41.98″N, 121°11′29.72″E

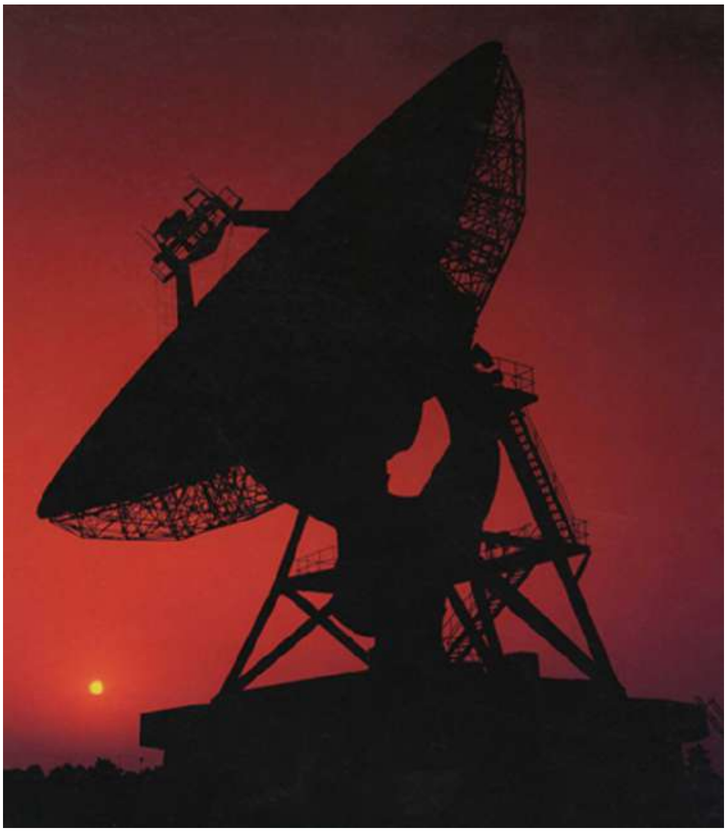

The Sheshan 25m radio telescope is an alt-az antenna run by SHAO.
The telescope is located in the Sheshan area, about 40km west of Shanghai.

This one is in the middle of the Old Jesuit Cathedral lot. They have an observatory there as well

Observatory and Church

http://www.thelivingmoon.com/45jack_files/03files/China_Sheshan.html

Been a lot of rumors around that the Jesuits still have a lot of control in the world...

Look up Wheeling Jesuit University... search for NASA grants and ask why is NASA giving money to a Church organization?
http://www.wju.edu/

Look up the National Technology Transfer Center (NTTC) These guys are in charge of who gets new technology

The National Technology Transfer Center (NTTC) has entered its 20th year of providing technology transfer services for federal facilities of our nation. Our work has provided clients with technical and market research expertise dedicated to meeting the complex needs of technology transfer organizations, while providing close support focusing on customer service. NTTC is inspired by the technologies developed at our federal R&D institutions, and is proud of our work commercializing and disseminating federal innovation for the betterment of our nation's economy.

Ask why do the Jesuits control technology transfer?

©2009 National Technology Transfer Center, Wheeling Jesuit University
http://www.nttc.edu/

But that is a whole new story

Back to **HAARP**/EISCAT like stuff...

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 7:51pm

HAARP High Frequency Active Auroral Research Program
Gacona, Alaska

There are more but that is all I have cataloged so far

But each one of those sites links to documents that link to other cogs in the wheel...

I am sure I have missed a few

I will add more as I have time to create the webpages. Feel free to add anything that you find from those links...

I will finish with some of the original **HAARP** images. You can also go to Jack Arnesons page and look under **HAARP** and Similar Facilities

HARRP Alaska
+62° 23' 30.00", -145° 8' 48.00"

HAARP cam image...

2009/09/21 19:40 AKDT

http://www.thelivingmoon.com/45jack_file....ona_Alaska.html

This is the **HAARP** 'home page'
http://www.**haarp**.alaska.edu/

But remember that the real **HAARP** project is run by the US NAVY NRL (Naval Research Laboratory

High Frequency Active Auroral Research Program
http://www.nrl.navy.mil/content.php?P=04REVIEW106

HAARP Photo Gallery and web cam access links

http://www.thelivingmoon.com/45jack_file....Gallery_01.html

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 8:00pm

Going through my old thread for relevant posts... here is one...

Quote:
> *Originally posted by TheWretched*
> Many of the ones you have listed are nothing but observatories for looking into the stars. Some are nothing but radar stations.

If you believe THAT then you haven't read anything about them. So much for your 'expertise' Most of the ones listed came from EISCAT and **HAARP** OFFICIAL websites

I haven't even listed all yet..but the antennas at
Kiruna, Sweden
Longyearbyen, Svalbard (Spitsbergen), Norway
Sodankylä, Finland
Søndre Strømfjord, Greenland
Ramfjordmoen, near Tromsø, Norway

Are most certainly used for Ionosphere manipulation and research which is after all THE POINT of these facilities and hence the thread

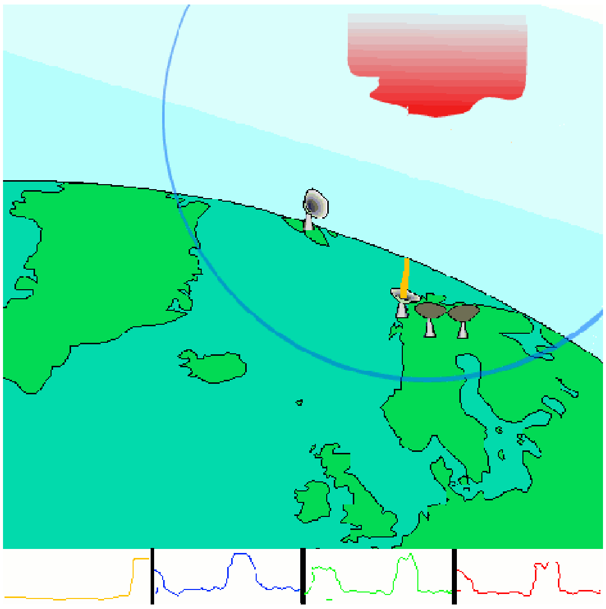

Why do these skeptics always have nasty names?

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 21, 2010, 8:10pm

Lot of weird stuff happening lately... Ocker has noticed it too, he said so in an ATS thread...

Stuff I have been looking for for years is suddenly posted on Wikipedia, Buzz Aldrin now posts the confirming data on his site about Aquila and the very day I made that **HAARP** thread at ATS...

Quote:

> *Originally posted by autowrench*
> Great post, starred and flagged. Did anyone happen to watch the History Channel last night? They had as show called "That's Impossible," and the show as all about **HAARP**, and the different facilities around the world, and how they use it as a weather weapon. I too have researched this thing, but the show had all of the tech, plus computer animation, and was fabulously presented.

Now how is that for coincidence? Same dang thing happened with PBS NOVA... A week before I was going to post all the original Secret Astronaut material we had collected, NOVA does :Astrospies"

{{{insert score to 'Twilight Zone' here}}}

Re: Every One Speaks about HAARP, but did you know
Post by dragar on Sept 21, 2010, 10:09pm

Zorgon im going to add this in please feel free to delete this post.
Take alook at it see if this is a bit strange.
HAARP= High Frequency Active Auroral Research Program
The Prince Albert Radar Laboratory (PARL):The original study that led to PARL came about as the side effect of questions about the effects of the aurora on radar systems.
Work on the **HAARP** Station began in 1993
Prince Albert Radar Laboratory limited operations started in 1959.
Could this be what sparked **HAARP**?

The Prince Albert Radar Laboratory

A satellite is any smaller object traveling around a larger one. This means that the moon is a satellite to the Earth. However, the term satellite often refers to a human-made spacecraft placed in space to orbit another body.

The first "communicatons satellite" was the moon. The USA started using the moon for communications trials in the early 1950's. Signals beamed towards the moon were reflected and received back on Earth. It was a very inefficient means of communication requiring high power and large antennas to produce a detectable system. In addition, communication was only possible while the moon was above the horizon.

DRTE staff participated in the trials starting in 1959 using a large radar station at the newly-built Prince Albert Radar Laboratory (PARL) in Saskatchewan. The first demonstration was given when Prime Minister John Diefenbaker inaugurated the station on June 6, 1959. Here are some photos of the inauguration. A pre-taped greeting* from President Eisenhower was transmitted from a station at Millstone Hill near Boston and received by the PARL station. Each station was equipped with a 26-m dish antenna. Diefenbaker used the occasion to announce that Canada would initiate its own satellite program. Here* is part of Mr. Diefenbaker's speech.

http://www.friendsofcrc.ca/Projects/PARL/PARL.html

As a side note :MY family and I have been there twice. It is very active many upgrades including a new security fence around it.
You may have to dig around in the link I gave for further info . ty

At the beginning, the work of the laboratory was focused on high frequency radio and the problem of forecasting the ionospheric conditions which determined the availability of any given frequency and transmission path. To this end, a new ionospheric sounder was designed by a team led by Bill McLeish. These sounders were constructed at RPL and were installed in a number of stations established across Canada. Most of the components were those which had been removed from surplus military equipment. The stations were manned by DOT personnel who were trained in their operation by Bob Stevens in a small building below the main laboratory by the shore of the Rideau Canal.

A second project arose from some of the HFDF work during the war. It had been noted that transmissions passing through the auroral zone had been deflected by as much as 20 degrees. To investigate this phenomenon further. Jack Meek proposed what became known as the "Mobile" experiment. A railroad car was equipped with a laboratory, diesel generators, accommodation, and cooking and eating facilities. Ionospheric sounders were installed, a crew was chosen and the car was hauled up to The Pas in Manitoba. From there, it was attached to the train which ran to Churchill and was shunted onto sidings at various points along the route. Antennas would be erected, and a series of ionospheric soundings taken. The car would then be hauled to the next siding and the process repeated. A line of soundings was thus obtained crossing the auroral zone. Jack Meek was in charge of the operation for the first period from the autumn of 1947 to January 1948 when Claire McKerrow took over. With Claire were Harold Serson, Cam Baker, and Sgt. Len Hagg on loan from the Canadian Army. The cook, provided by CNR, had an unfortunate habit of disappearing whenever the expedition arrived in The Pas. He was usually found in one of the local bars shortly before the train left. On one occasion, the town was searched from end to end to no avail. He was finally located in the local hospital, and they had to carry on without him. Claire called for volunteers to do the cooking with no success until Harold Serson finally admitted that he could cook a little. (In fact, Harold could have competed with most professional chefs.) Claire and Harold then shared the task until the cook's return. Food was supplied by the CNR from Winnipeg. Both Claire and Harold were inordinately fond of Winnipeg goldeye, and they ordered several cases of this delicacy.. The rest of the crew ate smoked goldeye, like it or not.

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Sept 22, 2010, 6:14am

Nice addition Did not have that one yet

Location, Location, Location

The requirements for the radar's site were twofold.

1. The site should permit the radar beam to illuminate the ionosphere over the Churchill, Manitoba region at the height at which the Aurora Borealis (Northern Lights) normally occurs. This would enable the comparison of auroral radar returns with simultaneous geophysical measurements made from Churchill.
2. The site should be accessible to scientists from the Physics department of the University of Saskatchewan, Saskatoon, where considerable expertise in auroral studies existed.

So it seems the location was specifically chosen to mess with the Aurora's.. so yeah... not only the predecessor for HAARP but proof that all big EISCAT and other huge radar dishes can effect the ionosphere

Nice catch 😮

Re: Every One Speaks about HAARP, but did you know
Post by dragar on Sept 22, 2010, 3:58pm

University of Saskatchewan
Auroral Radar - Saskatchewan

Early 1948
Peter Forsyth was preparing for medical school when Balfour Currie, returning from meeting in Ottawa, asked if he would help modify a radar to be borrowed from the Navy to seek radar echoes from aurora.
This turned out to be a 3000 MHz surveillance radar. It was modified and operated, but no auroral echoes were detected.

Auroral Radar - Saskatchewan

1948-49
With the aid of Nate Gerson, the USAF provided two huge 106 MHz early-warning radars. One went to Saskatoon, the other to NRC (McKinley), where it was to be modified to 56 MHz
First Canadian auroral echoes were detected with this radar in 1949 (paper in Nature, April, 1950).
As a result, Peter Forsyth decided to do his PhD on auroral echoes, rather than go into medicine.

Auroral Radar - Saskatchewan

1948-49 - Postscript
Around 1990 a phone call came to the Department of Physics at the University of Saskatchewan asking if they had an SCR-270 radar with a particular serial number
They did! It was the original radar which detected the incoming Japanese planes arriving to bomb Pearl Harbor (that information was ignored).
For over 40 years that radar had been sitting “buried” in Saskatoon. It is now at the Historical Electronics Museum near Baltimore.

Auroral Radar - Saskatchewan

1951-52
The modified radar came to Saskatoon from Ottawa, and a combined dual-frequency antenna was constructed.
Simultaneous observations of auroral echoes were observed at 56 and 106 MHz (1953 paper)
In 1951, Peter Forsyth finished his PhD and was hired to work at the Defence Research Telecommunications Establishment(DRTE) at Shirley Bay, near Ottawa, but since the facility didn't exist yet, he continued at Saskatoon for another year. It was a year later when he moved to DRTE, where he remained until 1958, rising to the position of Superintendent of the Radio Physics Lab.

Pre-History

1931 - NRC Radio Section established
1932 - Dr. John Henderson (PhD with Appleton)
1935 - Maj.-Gen. A.G.L. McNaughton new NRC President helped invent a cathode-ray tube direction-finder in the 1920s
1936-38 - Henderson developed direction-finder for marine and air use, installed in Nova Scotia in 1938
1938 - McNaughton was consulting with DND re detection of aircraft by electrical means
1939 - Henderson sent to UK to be briefed on radar developments

World War II

McNaughton contacted by Sir Charles Wright
you have expertise, get into this radar development
send physics graduates to UK Navy - badly needed
Peter Forsyth developed air-borne radar in UK, 1942-45
NRC became centre of radar development in Canada
McGill, Queens, Toronto, Western also involved in radar research
REL (Crown corporation) established to manufacture optical equipment, radar
During course of war, NRC designed more than a dozen different radars at many frequencies: 75, 95, 150, 200, 215, 2,800, 3,000, 10,000 and 24,000 MHz.
REL manufactured over 2000 radar sets, mostly for Navy, Army

Post-war

NRC
microwave experience --> Covington's 10.7cm solar radio-telescope (1946)
McKinley --> meteor radar program (1947)
DRB
established 1947 to take over defence research role which NRC refused to continue with after the war.
took over labs from services (e.g. Naval Research Establishment, CARDE), and started new ones (DRTE, DRNL, Ft. Churchill)
Weather radar
started in 1943 as a Canadian Army project in Ottawa, led by Stewart Marshall.
Post-war, at DRB & McGill, until 1950, then McGill alone. (Stormy Weather)

Re: Every One Speaks about HAARP, but did you know
Post by dragar on Sept 22, 2010, 4:09pm

EISCAT

In addition to the incoherent scatter radars, EISCAT also operates an Ionospheric Heater facility at Ramfjordmoen (including a Dynasonde) to support various active plasma physics experiments in the high latitude ionosphere.

Description of the EISCAT heating facility:

The Heating facility is situated next to the UHF and VHF incoherent scatter radars.
The Heater is used for ionospheric modification experiments applying high-power transmissions of high-frequency electro-magnetic waves to study plasma parameters in the ionosphere. The name Heating stems from the fact that these high power electromagnetic waves, which are transmitted into the ionosphre with high-gain antennas, heat the electrons and thus modify the plasma state. To create plasma turbulence, the transmitted frequencies have to be close to the plasma resonances, which are 4 to 8 MHz.

Re: Every One Speaks about HAARP, but did you know
Post by gigas on Sept 23, 2010, 9:32am

Here in wisconsin up north of me, we have clam lake and they run an ELF facility there to communicate with subs deep under the sea. This ELF transmitter is connected by another facility 128 miles east in michigan. I would guess this also holds true for the underground communications of other secret facilities regular radio communications cannot be linked up with.

This facility was said to be shut down some years ago but according to the dot mil link below, it still operates.

http://enterprise.spawar.navy.mil/UploadedFiles/fs_clam_lake_elf2003.pdf

---

Re: Every One Speaks about HAARP, but did you know
Post by pman35 on Sept 27, 2010, 11:46am

Hi guys have'nt read any of this but i have to say i'm pretty impressed zorgon with what u've posted to me it looks like if all these were connected they would indeed create a shield around the globe and because there are so many if one was to fail it might not affect the rest.

I will read this post as i have an interest in haarp.

Respect And May Disclosure Keep Moving Forwards

Re: Every One Speaks about HAARP, but did you know
Post by magooollo on Oct 5, 2010, 11:25am

buzz aldrin and edgar mitchell (apollo astronauts) are involved in the UFO community, perhaps as information conduits tho that isnt offically acknolwedged , would make good intermediatries in the disclosure race

---

Re: Every One Speaks about HAARP, but did you know
Post by somamech on Oct 5, 2010, 4:14pm

Interesting pdf Gigas!

EDIT

For anyone curious as to when Zorgon mentioned that sometimes documents coming from us mil sites cannot be accessed by folk abroad, this is what you may see showing in your browser. *Obviously I'm in Australia... and this is how our allie treats us?*

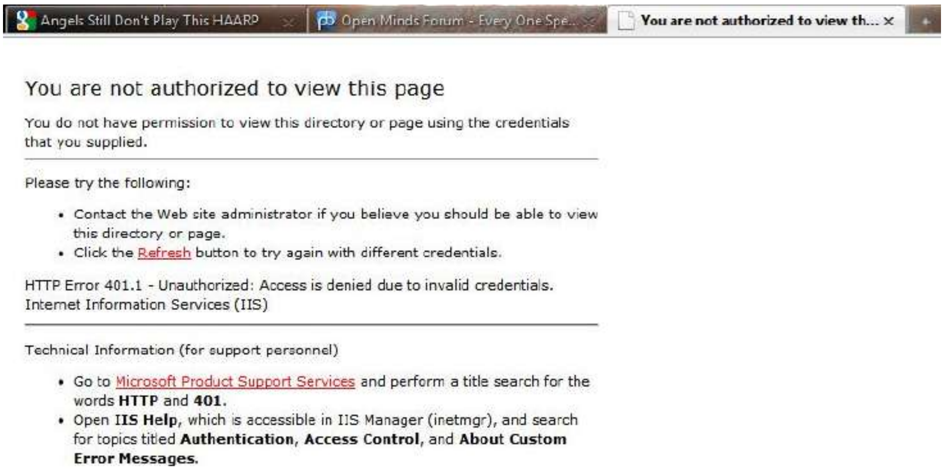
Angels Still Don't Play This HAARP
Open Minds Forum - Every One Spe...
You are not authorized to view th...

You are not authorized to view this page

You do not have permission to view this directory or page using the credentials that you supplied.

Please try the following:

- Contact the Web site administrator if you believe you should be able to view this directory or page.
- Click the Refresh button to try again with different credentials.

HTTP Error 401.1 - Unauthorized: Access is denied due to invalid credentials.
Internet Information Services (IIS)

Technical Information (for support personnel)

- Go to Microsoft Product Support Services and perform a title search for the words **HTTP** and **401**.
- Open **IIS Help**, which is accessible in IIS Manager (inetmgr), and search for topics titled **Authentication**, **Access Control**, and **About Custom Error Messages**.

EDIT: PDF in question which Gigas posted is this link below:

http://enterprise.spawar.navy.mil/UploadedFiles/fs_clam_lake_elf2003.pdf

---

Re: Every One Speaks about HAARP, but did you know
Post by winniethepooh4102 on Oct 5, 2010, 4:47pm

I found a torrent for it, but the downloading is very slow:

http://extratorrent.com/torrent_download....s+HAARP.torrent

But there is something more to HAARP if one
could believe in some claims.

Re: Every One Speaks about HAARP, but did you know
Post by gigas on Oct 5, 2010, 5:37pm

Oct 5, 2010, 4:14pm, somamech wrote:

Interesting pdf Gigas!

EDIT

For anyone curious as to when Zorgon mentioned that sometimes documents coming from us mil sites cannot be accessed by folk abroad, this is what you may see showing in your browser. *Obviously I'm in Australia... and this is how our allie treats us?*

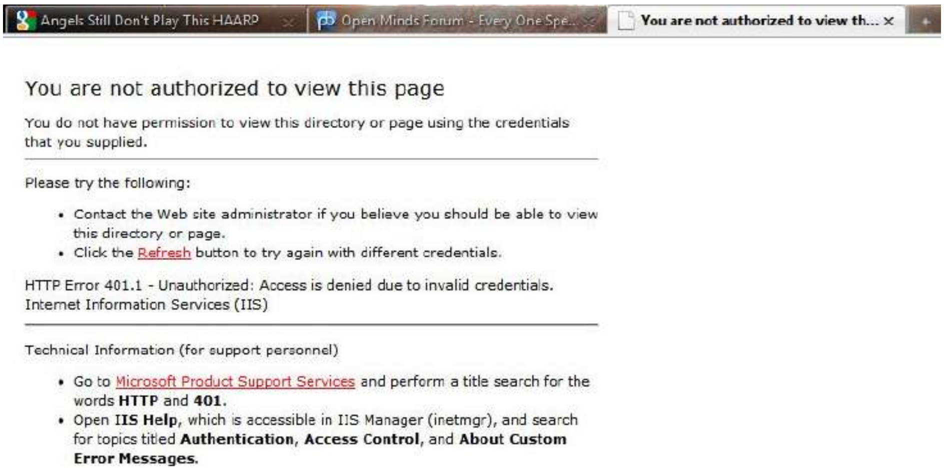
Angels Still Don't Play This HAARP
Open Minds Forum - Every One Spe...
You are not authorized to view th...

You are not authorized to view this page

You do not have permission to view this directory or page using the credentials that you supplied.

Please try the following:

- Contact the Web site administrator if you believe you should be able to view this directory or page.
- Click the Refresh button to try again with different credentials.

HTTP Error 401.1 - Unauthorized: Access is denied due to invalid credentials.
Internet Information Services (IIS)

Technical Information (for support personnel)

- Go to Microsoft Product Support Services and perform a title search for the words **HTTP** and **401.**
- Open **IIS Help**, which is accessible in IIS Manager (inetmgr), and search for topics titled **Authentication**, **Access Control**, and **About Custom Error Messages.**

EDIT: PDF in question which Gigas posted is this link below:

http://enterprise.spawar.navy.mil/UploadedFiles/fs_clam_lake_elf2003.pdf

I can no longer access that pdf as well with I am not authorized to access that page. It appears they locked it from web access. I tried to navigate to the pdf with a proxy web service and it says I am not authorized to see the pdf.

Looks like they don't like this information being out there for all to see.

Re: Every One Speaks about HAARP, but did you know
Post by somamech on Oct 5, 2010, 5:41pm

I saved a copy

EDIT

Re-reading that PDF I can clearly see why they don't like that info in the public domain

[image]

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Oct 5, 2010, 5:56pm

Oct 5, 2010, 4:14pm, somamech wrote:

> For anyone curious as to when Zorgon mentioned that sometimes documents coming from us mil sites cannot be accessed by folk abroad, this is what you may see showing in your browser.

FAS.org has a copy

http://www.fas.org/nuke/guide/usa/c3i/fs_clam_lake_elf2003.pdf

---

Re: Every One Speaks about HAARP, but did you know
Post by winniethepooh4102 on Oct 5, 2010, 6:02pm

The ELF cannot conquer the Elves.

Re: Every One Speaks about HAARP, but did you know
Post by gigas on Oct 5, 2010, 6:04pm

Oct 5, 2010, 5:41pm, somamech wrote:

> I saved a copy
>
> EDIT
>
> Re-reading that PDF I can clearly see why they don't like that info in the public domain
>
> [image]

Reading that pdf, it looks like the facility is still operating when it was supposed to be closed sept 30 2004.

Read the following link to see whats happening in the wisc land of dot mil.

http://www.zcommunications.org/project-elf-rip-by-site-administrator

---

Re: Every One Speaks about HAARP, but did you know
Post by gigas on Oct 5, 2010, 6:13pm

Thanks zorgon and somamech.

I thought I had saved it when found but maybe not since I looked and could not find it on this pc. Maybe on my other machine. I did save it from zorgons link though. really I did.

I tried to use the cache way back machines and it is scrubbed from the web as is the link to http://enterprise.spawar.navy.mil. I can't access that link as well.

Re: Every One Speaks about HAARP, but did you know
Post by winniethepooh4102 on Oct 5, 2010, 6:13pm

My favourite movie was "The Mad Max", the day after
tomorrow.

Re: Every One Speaks about HAARP, but did you know
Post by somamech on Oct 5, 2010, 6:31pm

Oct 5, 2010, 6:13pm, gigas wrote:

> Thanks zorgon and somamech.

I thought I had saved it when found but maybe not since I looked and could not find it on this pc. Maybe on my other machine. I did save it from zorgons link though. really I did.

I tried to use the cache way back machines and it is scrubbed from the web as is the link to http://enterprise.spawar.navy.mil. I can't access that link as well.

Me neither

I was always able to view that previously

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Oct 5, 2010, 6:34pm

One thing I need to mention....

Normally I will NEVER post a direct link to a file at a .mil site...

Why?

1) Many cannot get access from overseas
2) Many .mil sites that I visit require a certificate acceptance, which means you give them permission to monitor you. Some are several levels deep. This scares off the looky loos
3) Some ,mil sites like Maxwell AFB Archives will 'ping' back on the IP... happened once at ATS... everyone simply opening the page on that thread with the direct link got a warning pop up... I believe that contributed to ATS getting the rep of being government owned
4) This is the most important FILES GET MOVED because of the sudden traffic

Look at it this way An obscure document from 2003 has been idle for years... then suddenly I post a link say at ATS where several thousand viewers suddenly look at that file... well that sends up a red flag real fast

I have over 16,000 files I have saved... haven't even read them all yet but a friendly spook at ATS once told me that my posts generate unique hits on sites that they would rather remain quiet

So best bet is download the file and save it then post it somewhere else to share

One time I found a link to NASA Apollo photos... the FULL RS TIFF copies... I d/led a few posted something on a thread and while ArMaP was in the middle of a d/l the directory vanished. I never got all of them and I never found them again yet...

Another case I had a link to all the Rover images in full true color... that directory also moved... but fortunately I found that one again...

Because there are 40,000 pics on each Rover, I cannot d/l them all so I keep that link to myself now

Keith Laney has some of them on his site
http://www.keithlaney.net

Re: Every One Speaks about HAARP, but did you know
Post by somamech on Oct 5, 2010, 6:48pm

Good advice!

But in my case I don't mind clicking on those sites that require accepting a security certificate. In this case the whole site is offline...

Code:

```
http://enterprise.spawar.navy.mil./
```

Re: Every One Speaks about HAARP, but did you know
Post by somamech on Oct 5, 2010, 6:50pm

BTW..the moon boggy better not look like that POS LOL

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Oct 5, 2010, 6:51pm

Oct 5, 2010, 6:31pm, somamech wrote:

> Me neither
> I was always able to view that previously

Well see? You have just witnessed the system in action... I posted that link mainly because the word 'enterprise' was in that url. A similar one before was 'hyperspace.nrl.navy.mil/TiPS/' That was the info on the other tether satellite that flew at the same time the NASA one did (STS75)

It is still available on wayback machine

LINK

But this is what happens and that is why I have stopped posting direct links to sites like that because it means that I lose my sources when they plug the 'leaks'

Grab this one quick... it offers a 'public' access portal to the enterprise link, though I doubt the really cool stuff will be found there... but you never know

http://nesipublic.spawar.navy.mil/docs/misc/PEO_C4I_NESI_Memo.pdf

---

Re: Every One Speaks about HAARP, but did you know
Post by somamech on Oct 5, 2010, 7:13pm

Thanks Zorg!

That pesky system... Arrgh. When I get time after the weeks break i will show you my own instance of posting a link in an email that had way too many coincindes to be called so.

I don't mind posting DSTO links though... USA gets way too much attention... and lets face it, a good idea is as simple as placing some magnets and concrete together ala Ed. It does not need wads of cash thrown at it. Heck any country can come up with THE right idea

---

Re: Every One Speaks about HAARP, but did you know
Post by gigas on Oct 5, 2010, 8:30pm

Oct 5, 2010, 6:34pm, zorgon wrote:

> One thing I need to mention....
>
> Normally I will NEVER post a direct link to a file at a .mil site...
>
> Why?
>
> 1) Many cannot get access from overseas
> 2) Many .mil sites that I visit require a certificate acceptance, which means you give them permission to monitor you. Some are several levels deep. This scares off the looky loos
> 3) Some ,mil sites like Maxwell AFB Archives will 'ping' back on the IP... happened once at ATS... everyone simply opening the page on that thread with the direct link got a warning pop up... I believe that contributed to ATS getting the rep of being government owned
> 4) This is the most important FILES GET MOVED because of the sudden traffic

Who knew, cept master zorgon. Ok my friend, the lesson has surely been learned and in the future, I will follow your proven advice and keep dot mil links on the low.

Thanks again exalted one for sharing that important bit of information on dot mil. We have seen it in action and know it to be true .

Now, all hail zorgon .

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Oct 5, 2010, 8:38pm

Oct 5, 2010, 6:02pm, winniethepooh4102 wrote:

> The ELF cannot conquer the Elves.

I hope not I like Elves... have a few myself

Re: Every One Speaks about HAARP, but did you know
Post by winniethepooh4102 on Oct 6, 2010, 7:38am

Observing those very strange clouds and chemtrails
in the sky above me I'm inclined to believe, as though
the atmosphere is growing wiser and overripe. Do you
admit the claims about the HAARP and the elves?

P.S. I meant exactly this and some claims surrounding it:

http://www.educate-yourself.org/cn/sylphandchemtrailindex.shtml

http://educate-yourself.org/zsl/sylphessayindex.shtml

If the elves and the sylphs really exist they have to
be in one category. Considering the UFO problem
every possible information has to be processed.

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Oct 7, 2010, 9:00am

Elves and Sprites Play above huge thunderstorms

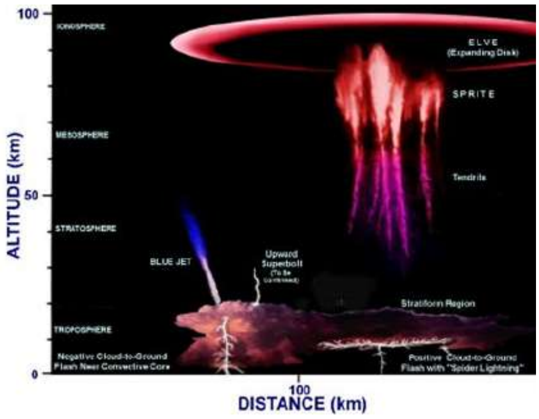
100
IONOSPHERE
ELVE
(Expanding Disk)
SPRITE
MESOSPHERE
Tendrils
ALTITUDE (km)
50
STRATOSPHERE
BLUE JET
Upward
Superbolt
(To be
confirmed)
Stratiform Region
TROPOSPHERE
Negative Cloud-to-Ground
Flash Near Convective Core
Positive Cloud-to-Ground
Flash with "Spider Lightning"
0
100
DISTANCE (km)

The Sylphs on the other hand only reveal themselves to those who can see them. I have a magnificent collections of Dragons... perhaps I may do a thread

They have little to do with HAARP as the Sylphs play in the atmosphere while HAARP messes several miles higher in the ionosphere, home of the Aurora...

The Dragons will come though... all you need to do is ask them to appear

---

Re: Every One Speaks about HAARP, but did you know
Post by winniethepooh4102 on Oct 7, 2010, 11:21am

I would leave the theme in this ouroboros style
not knowing how exactly the matrix is structured.
Anyway those creatures from under the ground
and under the water flying also everywhere in
the air and into the space cannot be unknown
for The Big Brother. The question is why they are
so big secret and what is their relation to us
indepedently of those massive space ships
and the ETs.

Re: Every One Speaks about HAARP, but did you know
Post by dragar on Nov 4, 2010, 5:51pm

Naughty! Naughty!
Hotlinking of images
is prohibited

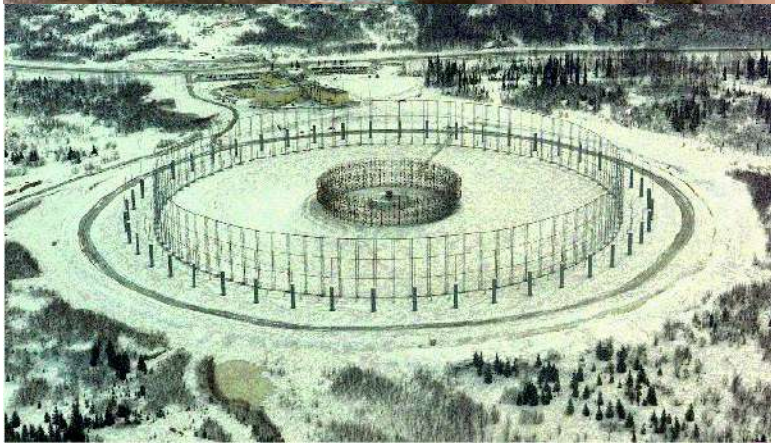

The AN/FLR-9, operated by the Air Intelligence Agency, is part of the DOD World Wide High Frequency Direction Finding System. High-frequency radio communication signals travel to receivers over the horizon by bouncing off the ionosphere. The shell of ionised particles which surrounds the earth refracts the signals so that they return to earth rather than disappear out into space. The clarity of the signal received depends on atmospheric and topographical conditions. HF-DF stations detect radio signals from aircraft or ships, and calculate the direction, or line of bearing, of the radio transmitter from the direction finding antenna. When the same signal is received by two or more antennae, the intersection of the lines of bearing mark the transmitter's location, using either precision single station location (SSL) capability, or in a network of DF stations using both multi-station azimuth triangulation and SSL. High Frequency Acquisition (AQ) and Direction Finding (DF) operations are performed with the Narrowband System (NBS) and Wideband Direction Finding (WBDF) Subsystem in support of normal and degraded communications modes, using both adaptive reception and super-resolution direction finding techniques.

Re: Every One Speaks about HAARP, but did you know
Post by ivalen on Nov 9, 2010, 7:50pm

Sept 21, 2010, 7:03pm, zorgon wrote:

http://www.thelivingmoon.com/45jack_file....r_Facility.html

Cherenkov Radiation around Antennas

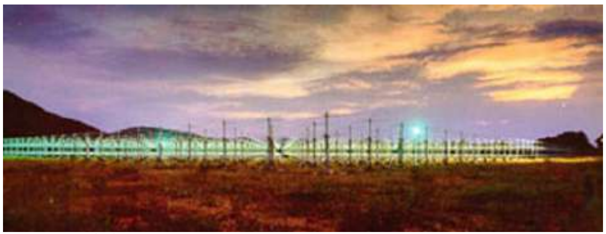

Image courtesy ISRO

http://en.wikipedia.org/wiki/Cherenkov_radiation

> Cherenkov radiation (also spelled Cerenkov or Čerenkov) is electromagnetic radiation emitted when a charged particle (such as an electron) passes through an insulator at a constant speed greater than the speed of light in that medium. The charged particles polarize the molecules of that medium, which then turn back rapidly to their ground state, emitting radiation in the process.

For this to be the case in the above photograph - what is the insulating medium?

Re: Every One Speaks about HAARP, but did you know
Post by sarentack on Nov 11, 2010, 4:10am

one would think some of these facilities was like ginormous computer chip.

or signals or landmarks for ET like that Triangle look from above.

Wonder if their any relation to what we see in Peru?

Re: Every One Speaks about HAARP, but did you know
Post by somamech on Dec 5, 2010, 5:26pm

Hey HAARP is old news to the mountain people

The uncanny resemblance to a mil array

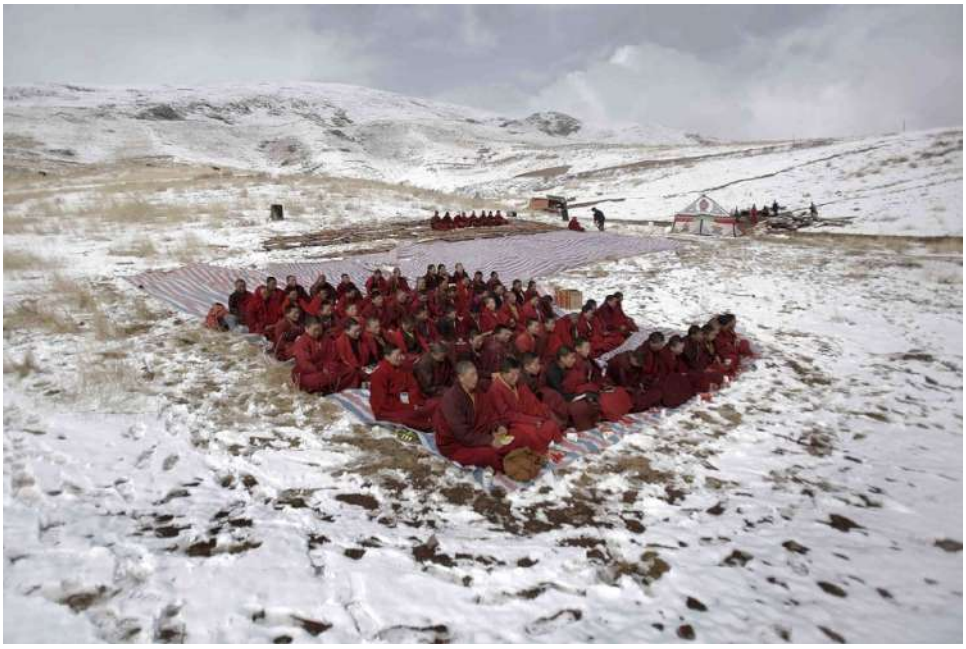

^compare that...

To this:

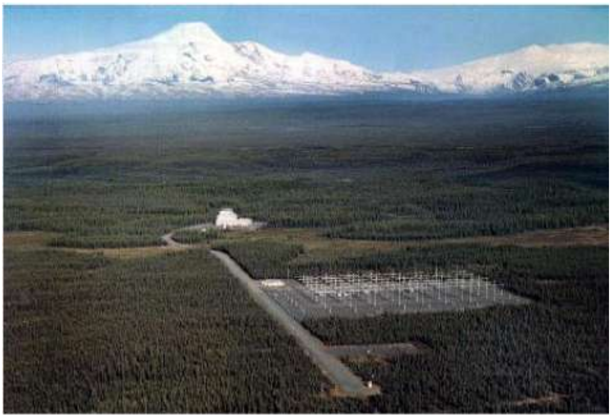

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Dec 7, 2010, 8:55am

Nov 9, 2010, 7:50pm, ivalen wrote:

> For this to be the case in the above photograph - what is the insulating medium?

Good question... I might have said 'plasma' but that would require even more explanation. I do need to pinpoint exactly what these glows are... related to another glowing object on the Moon

St. Elmo's fire (also St. Elmo's light) is an electrical weather phenomenon in which luminous plasma is created by a coronal discharge originating from a grounded object in an atmospheric electric field

Physically, St. Elmo's fire is a bright blue or violet glow, appearing like fire in some circumstances, from tall, sharply pointed structures

Working on it

Re: Every One Speaks about HAARP, but did you know
Post by zorgon on Dec 7, 2010, 9:12am

Nov 4, 2010, 5:51pm, dragar wrote:

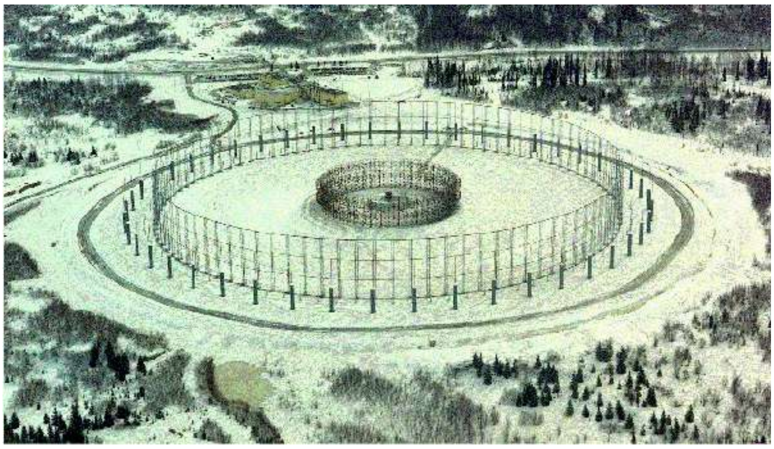

We have one of those in Japan too

Misawa, Japan
Misawa NSA Echelon Station
Code Name: Ladylove
+40° 43' 22.91", +141° 19' 43.57"

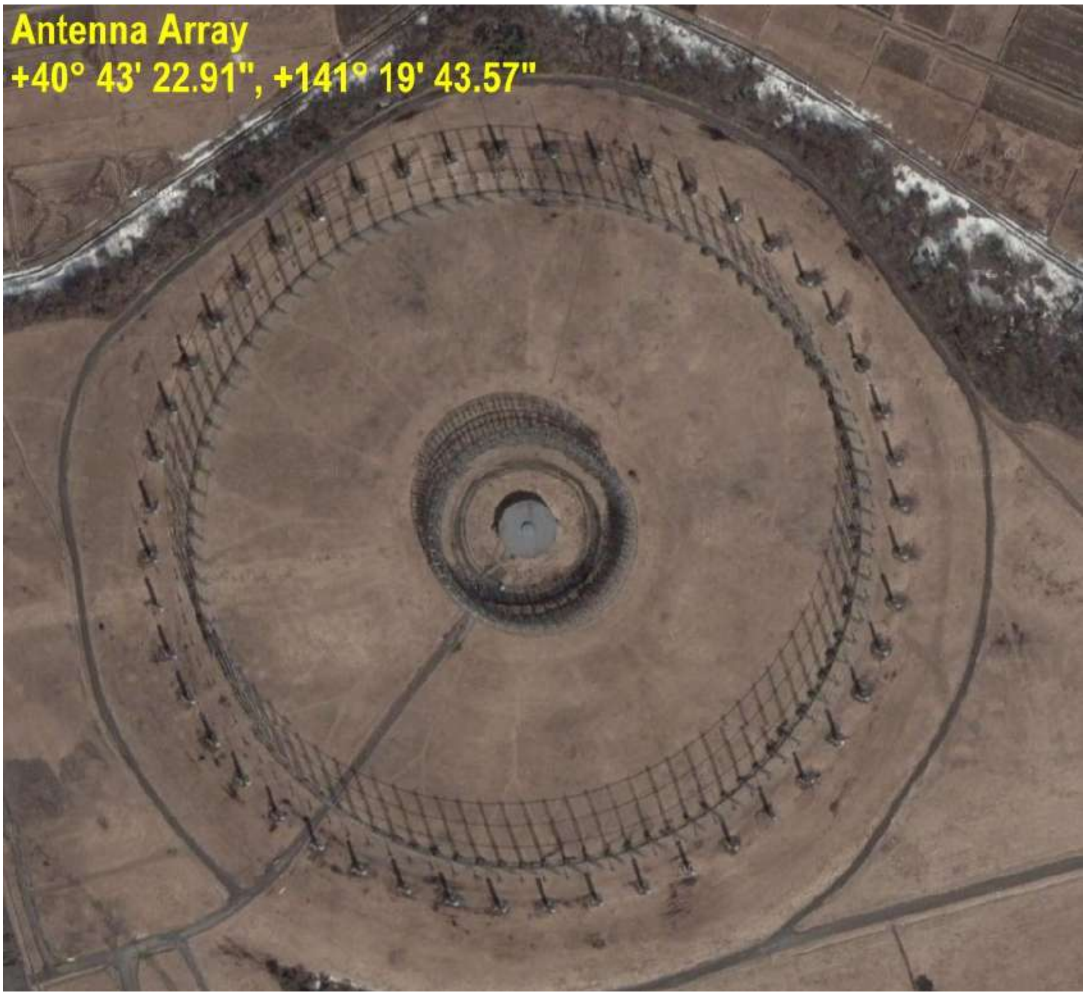
Antenna Array
+40° 43' 22.91", +141° 19' 43.57"

Quite a NAVY and AF presence there as well as an ECHELON station. We have nukes in storage in that area as well

http://www.thelivingmoon.com/45jack_files/03files/ECHELON_Misawa.html

NAVY INFORMATION OPERATIONS COMMAND MISAWA
NETWARCOM
三沢

Re: Every One Speaks about HAARP, but did you know
Post by zaggan on Dec 7, 2010, 3:31pm

I'm late in the game on this discussion, but it surfaced in June of last year in one of Dan's threads (think Reality 101). Someone had brought up one of the antenna arrays, along with a picture, and I commented back in late June. Dan got snippy and deleted all references to it, but want to add it back in here:

Received a reply to an inquiry to POA Elmendorf AFB, listed as follows:

Good Afternoon Sir,

My name is SrA Xxxxxx Xxxxxxxxx from the 3rd Wing Public Affairs on Elmendorf and I contacting you in regards to the question about the structure you saw on Google Earth.

It is a HF Antenna. If you would like more information you can go to the following link. http://www.fas.org/irp/program/collect/an-flr-9.htm

If you have any other questions, comments or concerns please do not hesitate to contact me at 552-8915 or by email.

Hope this helps!

Thank You and have a nice day!

XXXXXX XXXXXXXXX, SRA, USAF
COMMUNITY RELATIONS

3WG PUBLIC AFFAIRS
DSN: 552-8915 COMM: 552-8915

The link is pretty interesting.

Re: Every One Speaks about HAARP, but did you know
Post by gigas on Dec 7, 2010, 8:22pm

Dec 7, 2010, 9:12am, zorgon wrote:

Nov 4, 2010, 5:51pm, dragar wrote:

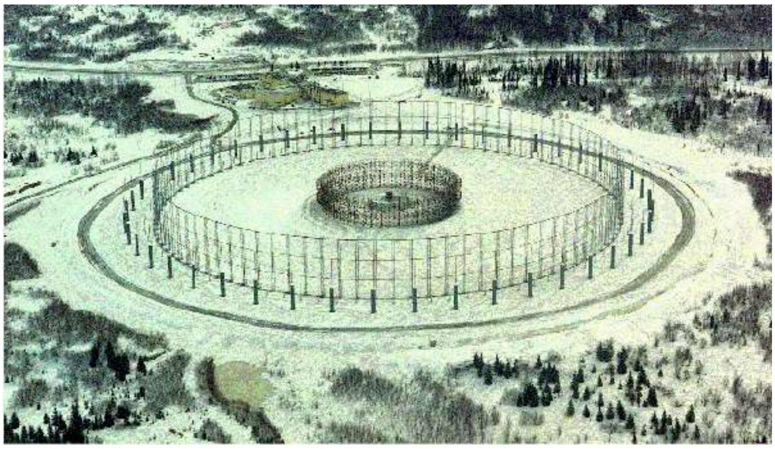

We have one of those in Japan too

Misawa, Japan
Misawa NSA Echelon Station
Code Name: Ladylove
+40° 43' 22.91", +141° 19' 43.57"

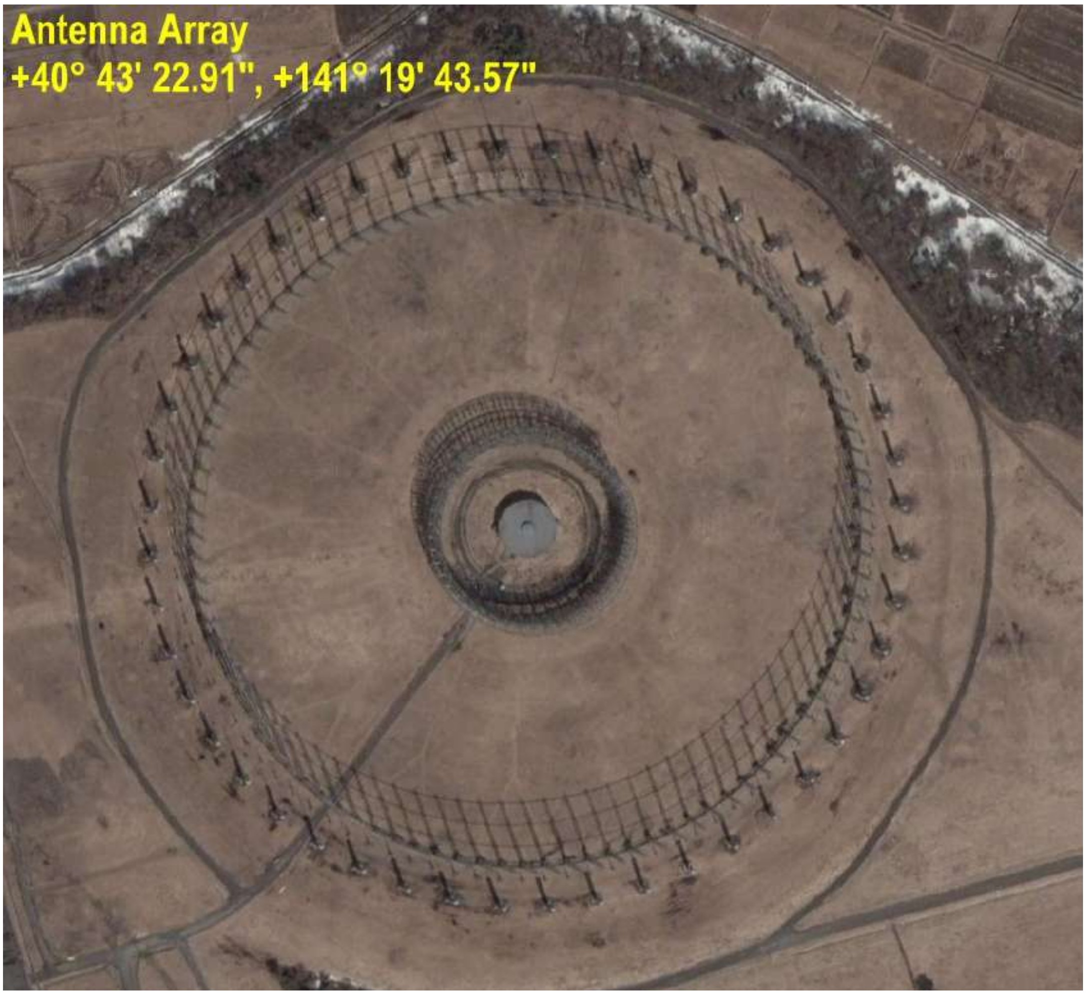

Looks like they had one back 5,000 years ago in the UK around salisbury called stonehenge as well.

THE CHEMTRAIL DIARY (/)

HOME (/) FOTO'S (/FOTOS.HTML) DOCUMENTEN (/DOCUMENTEN.HTML)

VIDEO'S (/VIDEOS.HTML) LINKS (/LINKS.HTML) TOP 10 (/TOP-10.HTML)

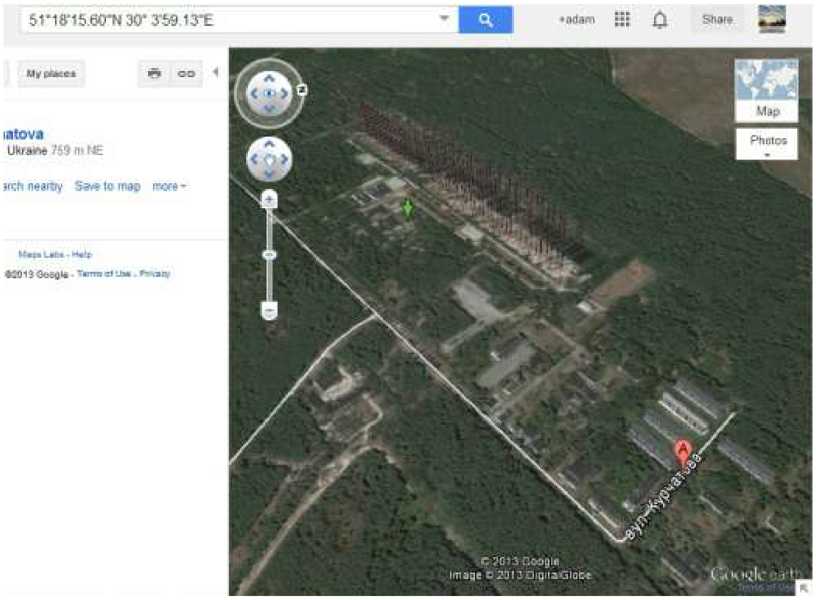
51°18'15.60"N 30° 3'59.13"E

+adam
Share
My places
atova
Ukraine 759 m NE
arch nearby Save to map more
Maps Labs - Help
©2013 Google - Terms of Use - Privacy
Map
Photos
вул. Курчатова
© 2013 Google
Image © 2013 DigitalGlobe
Google earth
Terms of Use

HAARP LOCATIES

- GPS Locaties -

(/haarp.html)

- GEBRUIK / USE GOOGLE MAPS -

1. High Frequency Active Auroral Research Programn (HAARP)

Gakona Alaska

+62° 23' 30.00", -145° 8' 48.00"

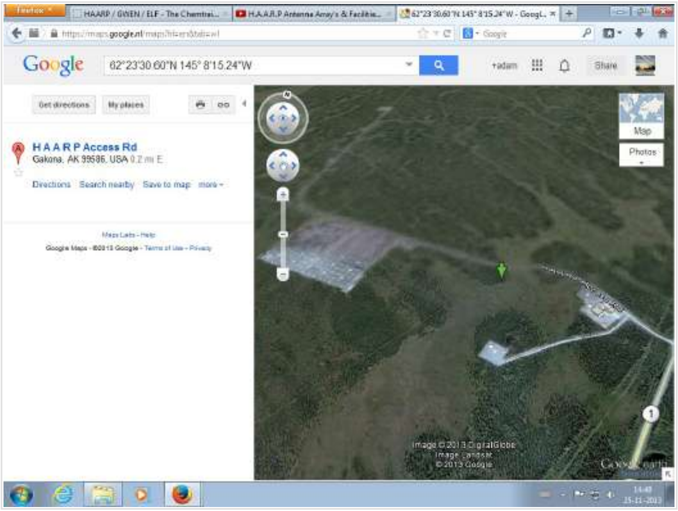
Google
62°23'30.60"N 145° 8'15.24"W
H A A R P Access Rd
Gakona, AK 99586, USA 0.2 mi E
Directions Search nearby Save to map more
Map
Photos
Image © 2013 DigitalGlobe
Image Landsat
© 2013 Google

(/uploads/2/3/1/7/23175228/4799260_orig.jpg)

2. High Power Auroral Stimulation Observatory (HIPAS)

Fairbanks Alaska

64° 52' 19" N 146° 50' 33" W

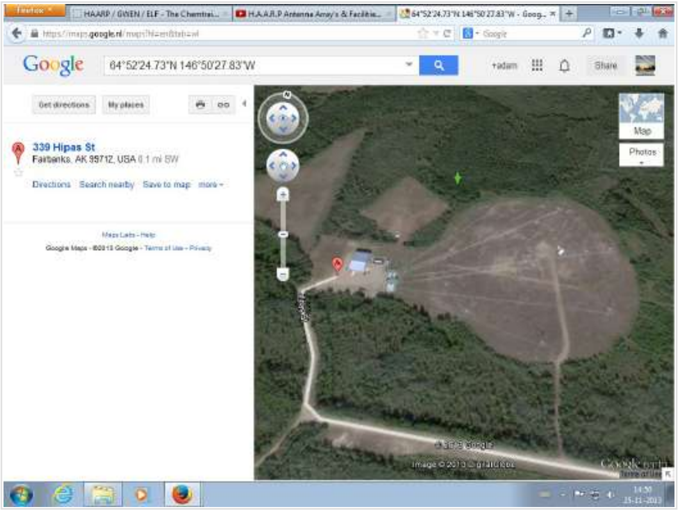
Google
64°52'24.73"N 146°50'27.83"W
339 Hipas St
Fairbanks, AK 99712, USA 0.1 mi SW
Directions Search nearby Save to map more
Map
Photos
© 2013 Google
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/7260176_orig.jpg)

3. Poker Flat Research Range

Near Chatanika, Alaska

+65° 7' 23.90", -147° 28' 7.05"

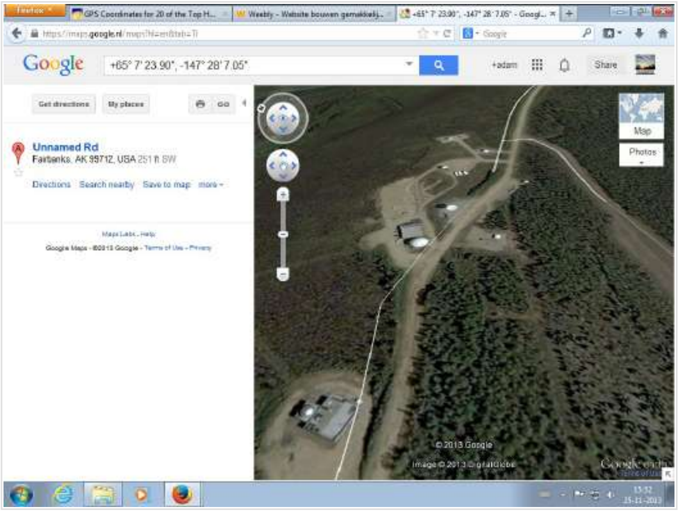
Google
+65° 7' 23.90", -147° 28' 7.05"
Unnamed Rd
Fairbanks, AK 99712, USA 251 ft SW
Directions Search nearby Save to map more
Map
Photos
© 2013 Google
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/7191263_orig.jpg)

4. Platteville Atmospheric Observatory

Platteville, Colorado

+40° 10' 54", -104° 43' 30"

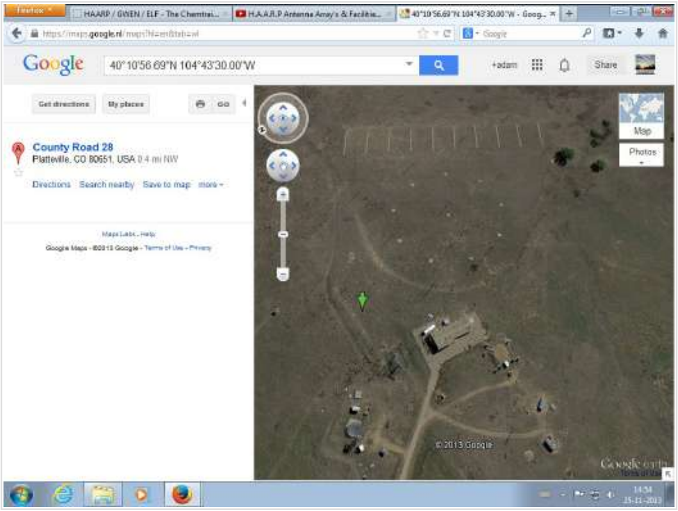
Google
40°10'56.69"N 104°43'30.00"W
County Road 28
Platteville, CO 80651, USA 0.4 mi NW
Directions Search nearby Save to map more
Map
Photos
© 2013 Google

(/uploads/2/3/1/7/23175228/2019020_orig.jpg)

5. Haarp - Westford, Massachusetts, USA (EISCAT-like)
42°37'6.89"N 71°29'25.72"W

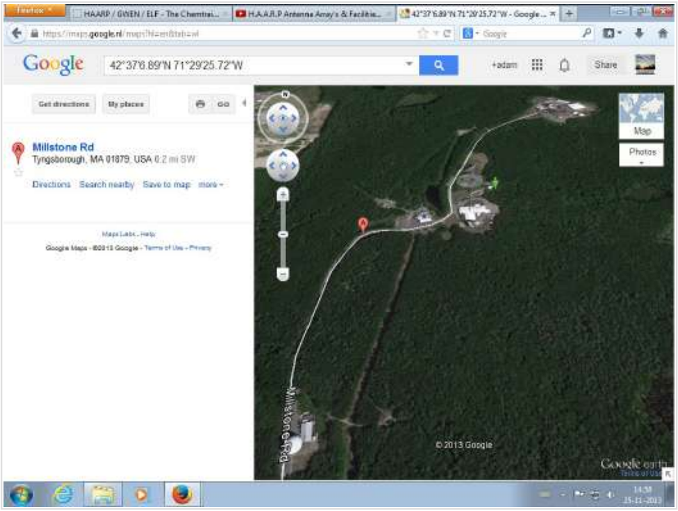
Google
42°37'6.89"N 71°29'25.72"W
Get directions
My places
Millstone Rd
Tyngsborough, MA 01879, USA 0.2 mi SW
Directions Search nearby Save to map more
Map
Photos
Millstone Rd
© 2013 Google

(/uploads/2/3/1/7/23175228/7575768_orig.jpg)

6. Haarp - Westford, Massachusetts, USA (EISCAT-like)
Construction
42°37'26.36"N 71°29'16.10"W

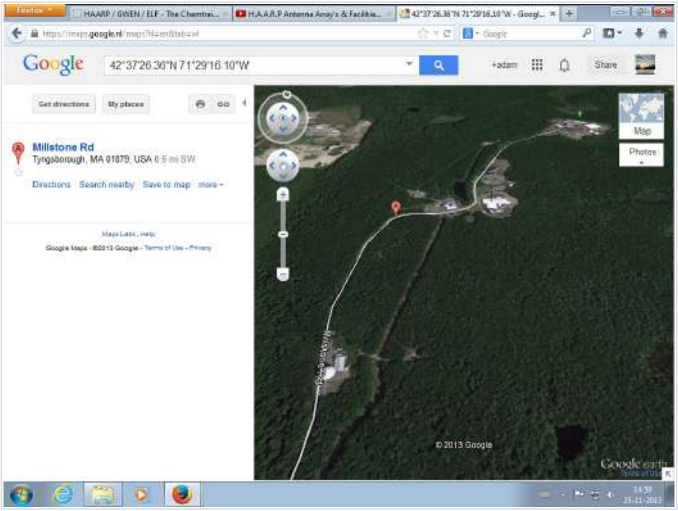
Google
42°37'26.36"N 71°29'16.10"W
Get directions
My places
Millstone Rd
Tyngsborough, MA 01879, USA 0.5 mi SW
Directions Search nearby Save to map more
Map
Photos
© 2013 Google

(/uploads/2/3/1/7/23175228/1981341_orig.jpg)

7. Nerc MST Radar Facility
Capel Dewi, Carmarthenshire, Wales, United Kingdom
52°25'25.95"N 4° 0'19.48"W

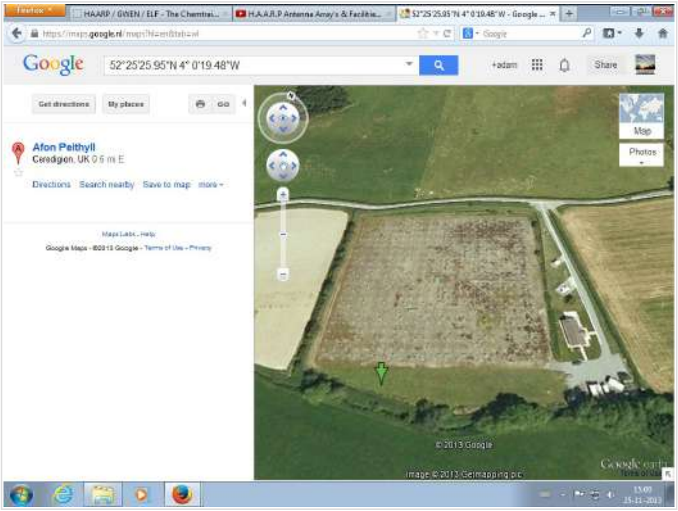
Google
52°25'25.95"N 4° 0'19.48"W
Get directions
My places
Afon Peithyll
Ceredigion, UK 0.6 mi E
Directions Search nearby Save to map more
Map
Photos
© 2013 Google
Image © 2013 Getmapping plc

(/uploads/2/3/1/7/23175228/5483133_orig.jpg)

8. The Russian Woodpecker
Duga Radar Array, Chenobyl, Ukraine
51°18'20.17"N, 30°04'02.60"E

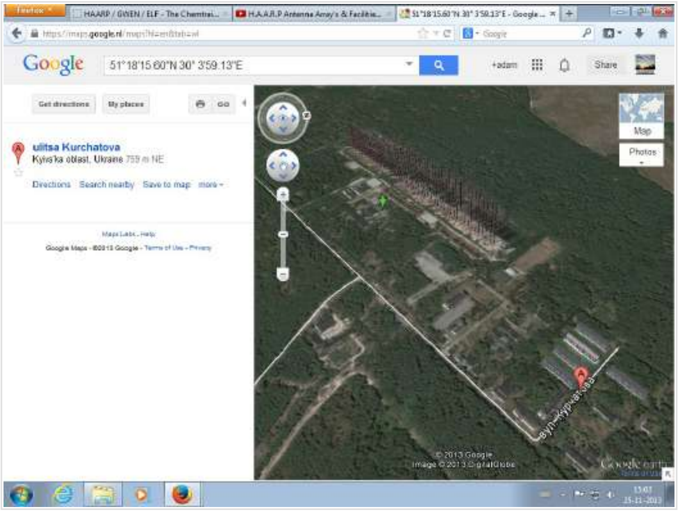
Google
51°18'15.60"N 30° 3'59.13"E
Get directions
My places
ulitsa Kurchatova
Kyivs'ka oblast, Ukraine 759 m NE
Directions Search nearby Save to map more
Map
Photos
© 2013 Google
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/4301759_orig.jpg)

9. The Arecibo Observatory radio telescope
Arecibo Puerto Rico.
18°20’38.97”N 66°45’9.77”W

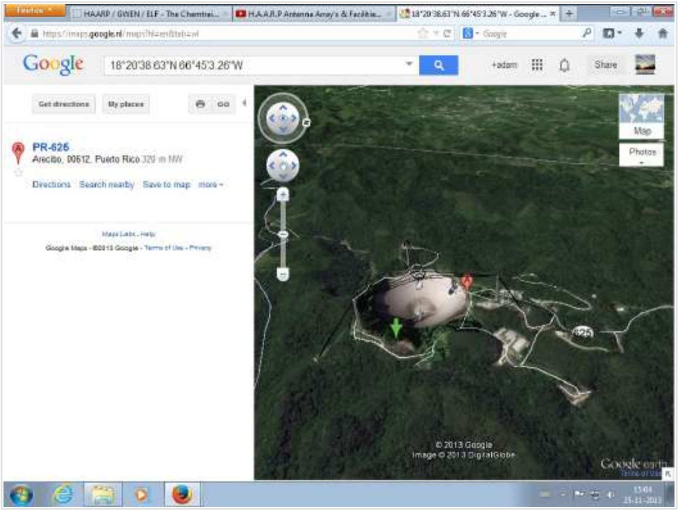

18°20'38.63"N 66°45'3.26"W

PR-625
Arecibo, 00612, Puerto Rico 320 m NW

(/uploads/2/3/1/7/23175228/7398577_orig.jpg)

10. National MST Radar Facility NMRF
Gadanki, India
+13° 27’ 26.68”, +79° 10’ 30.74”

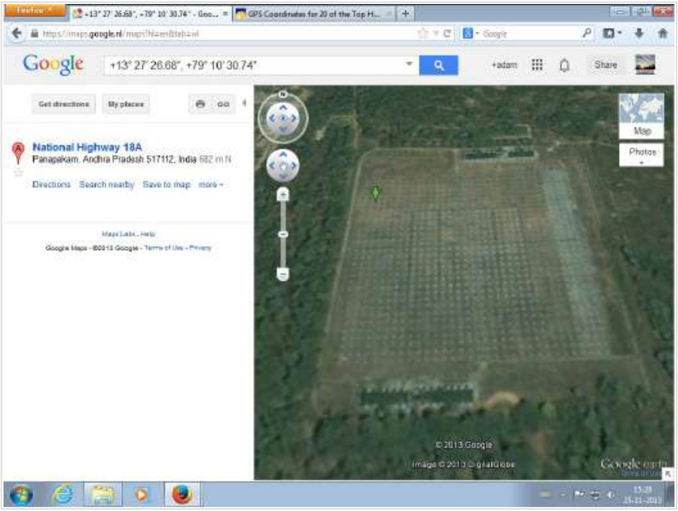

+13° 27' 26.68", +79° 10' 30.74"

National Highway 18A
Panapakam, Andhra Pradesh 517112, India 682 m N

(/uploads/2/3/1/7/23175228/6048676_orig.jpg)

11. Jicamarca Radio Observatory
Lima, Peru
11° 57’ 08.25” S 76° 52’ 30.67” W

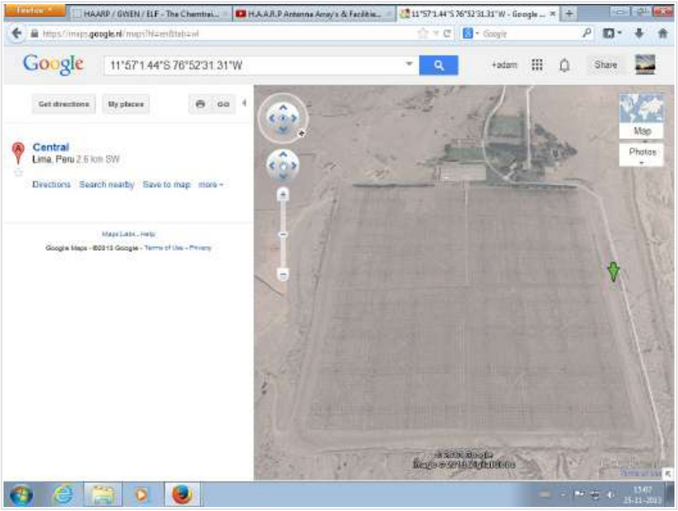

11°57'1.44"S 76°52'31.31"W

Central
Lima, Peru 2.6 km SW

(/uploads/2/3/1/7/23175228/6266894_orig.jpg)

12. São Luiz Space Observatory
Cruzeiro Santa Bárbara, Sao Luis-MA, Brasil
-2° 35’ 40.47”, -44° 12’ 35.90”

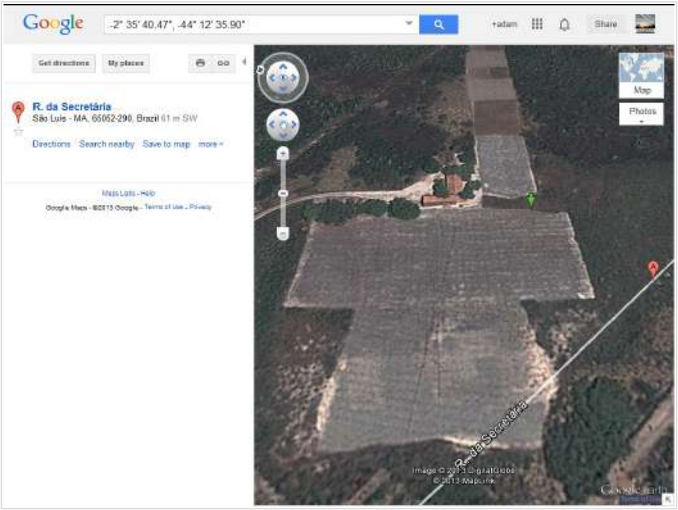

-2° 35' 40.47", -44° 12' 35.90"

R. da Secretária
São Luís - MA, 65052-290, Brazil 61 m SW

(/uploads/2/3/1/7/23175228/7926199_orig.jpg)

13. European Incoherent Scatter Scientific Association EISCAT
Ramfjordmoen, Near Tromsø, Norway
+69° 35’ 10.94”, +19° 13’ 20.89”

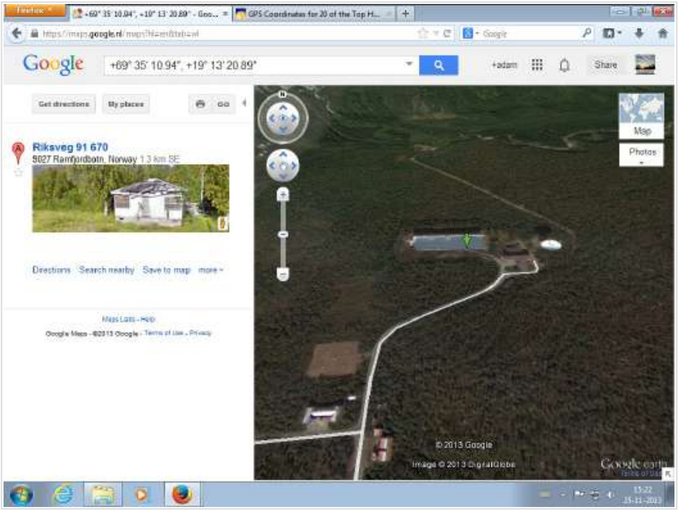
Google
+69° 35' 10.94", +19° 13' 20.89"
Get directions
My places
Riksveg 91 670
9027 Ramfjordbotn, Norway 1.3 km SE
Directions Search nearby Save to map more
Map
Photos
© 2013 Google
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/3199053_orig.jpg)

14. Sura Facility
Nizhniy Novgorod, Russia
+56° 7’ 9.70”, +46° 2’ 3.66”

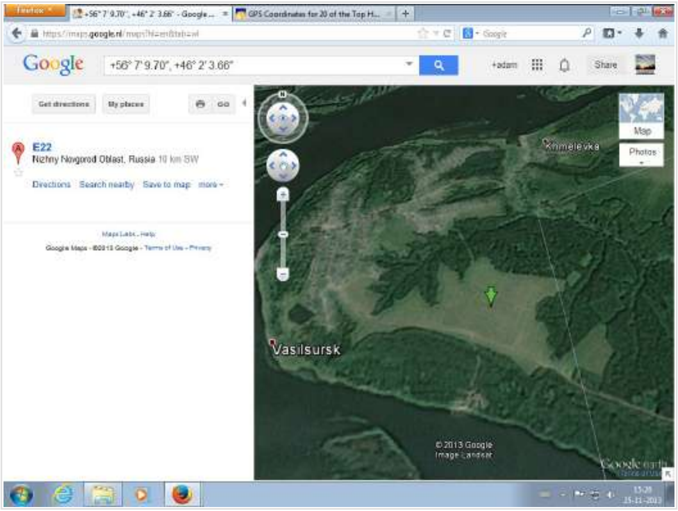
Google
+56° 7' 9.70", +46° 2' 3.66"
Get directions
My places
E22
Nizhny Novgorod Oblast, Russia 10 km SW
Directions Search nearby Save to map more
Map
Photos
Khmelevka
Vasilsursk
© 2013 Google
Image Landsat

(/uploads/2/3/1/7/23175228/5474744_orig.jpg)

15. China Research Institute of Radiowave Propagation (CRIRP)
Xinjiang (Sinkiang) Region
40°24’15.91”N, 93°38’09.74”E

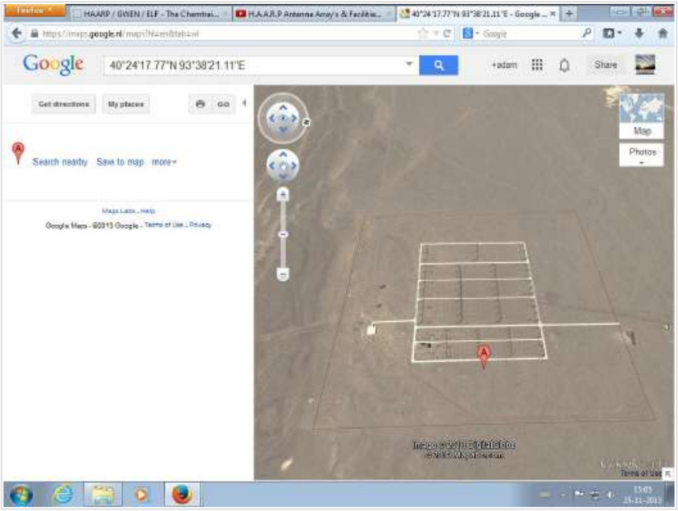
HAARP / GWEN / ELF - The Chemtrai...
H.A.A.R.P Antenna Array's & Faciliti...
Google
40°24'17.77"N 93°38'21.11"E
Get directions
My places
Search nearby Save to map more
Map
Photos

(/uploads/2/3/1/7/23175228/7563411_orig.jpg)

16. Sheshan, Shanghai,
China
31°5’41.98”N, 121°11’29.72”E

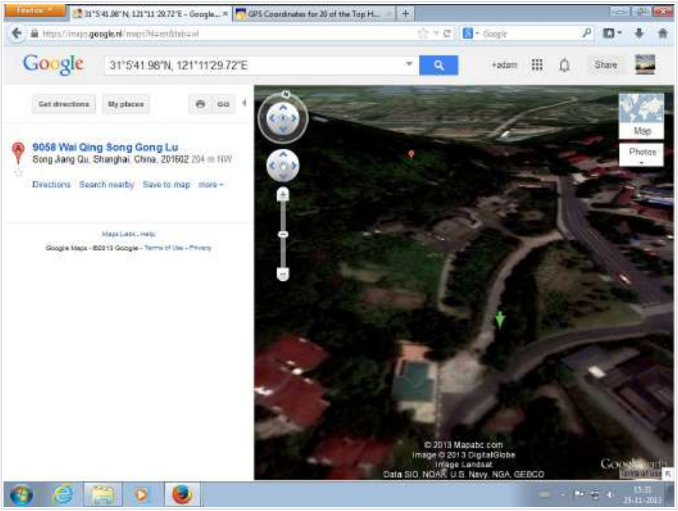
Google
31°5'41.98"N, 121°11'29.72"E
Get directions
My places
9058 Wai Qing Song Gong Lu
Song Jiang Qu, Shanghai, China, 201602 204 m NW
Directions Search nearby Save to map more
Map
Photos
© 2013 Mapabc.com
Image © 2013 DigitalGlobe
Image Landsat
Data SIO, NOAA, U.S. Navy, NGA, GEBCO

(/uploads/2/3/1/7/23175228/6051246_orig.jpg)

17. Mu Radar
in Japan
34°51’14.80”N 136° 6’19.45”E

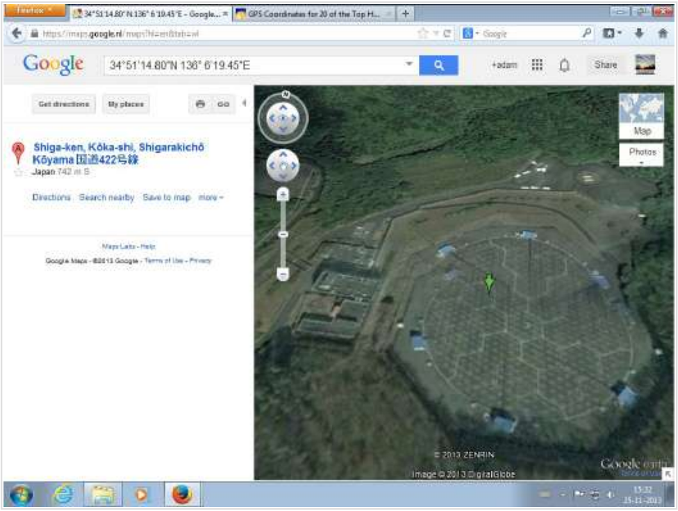
34°51'14.80"N 136° 6'19.45"E
Shiga-ken, Kōka-shi, Shigarakichō Kōyama 国道422号線
Japan
Directions Search nearby Save to map more
© 2013 ZENRIN
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/4904711_orig.jpg)

18. Jindalee Operational Radar Network (JORN)
Laverton, West Australia
-28° 19’ 36.29”, +122° 0’ 18.84”

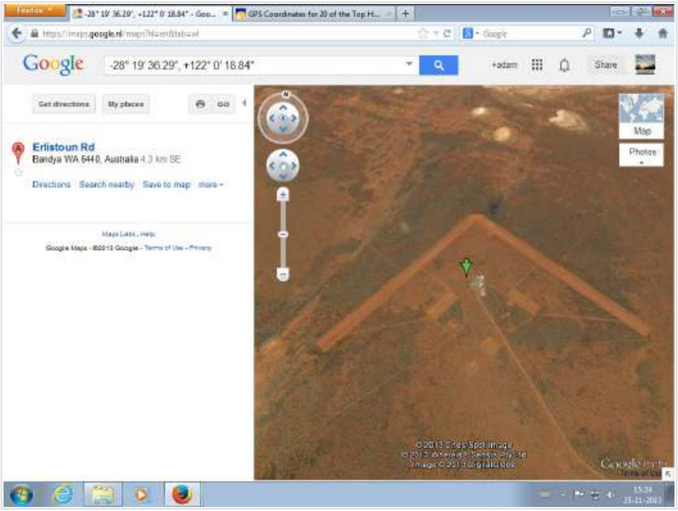
-28° 19' 36.29", +122° 0' 18.84"
Erlistoun Rd
Bandya WA 6440, Australia
Directions Search nearby Save to map more

(/uploads/2/3/1/7/23175228/97514_orig.jpg)

19. Jindalee Operational Radar Network (JORN)
Longreach, Queensland, Australia
-23° 39’ 29.53”, +144° 8’ 49.58”

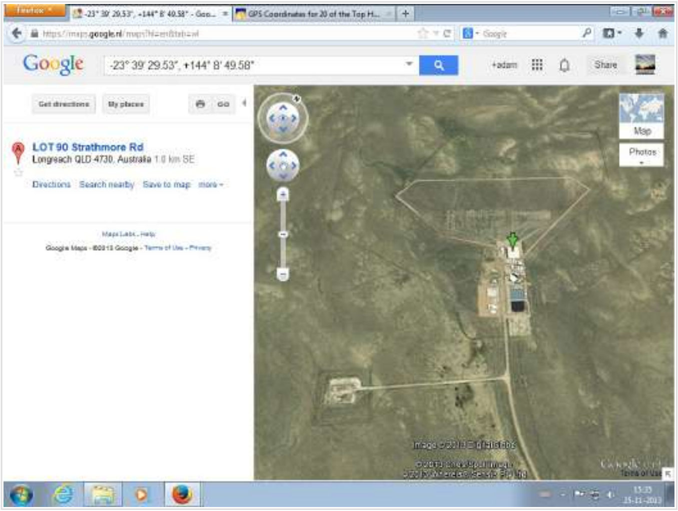
-23° 39' 29.53", +144° 8' 49.58"
LOT 90 Strathmore Rd
Longreach QLD 4730, Australia
Directions Search nearby Save to map more
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/908329_orig.jpg)

20. Australia Naval Communications Station Harold E Holt (NCS HEH)
Exmouth, Western Australia.

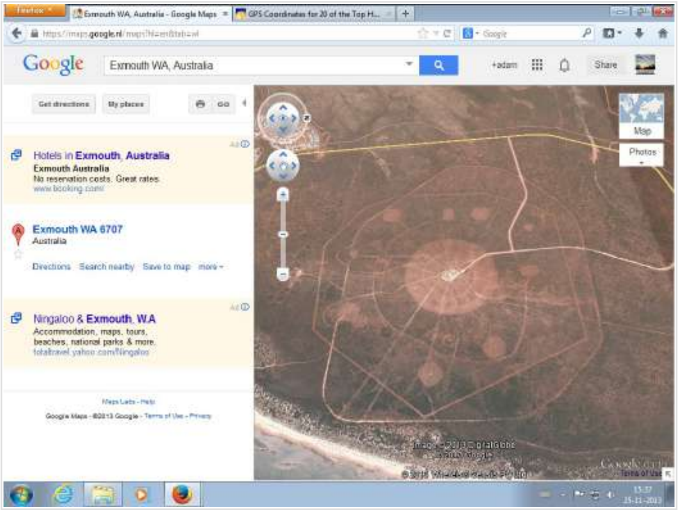
Exmouth WA, Australia
Hotels in Exmouth, Australia
Exmouth WA 6707
Australia
Ningaloo & Exmouth, WA
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/3863477_orig.jpg)

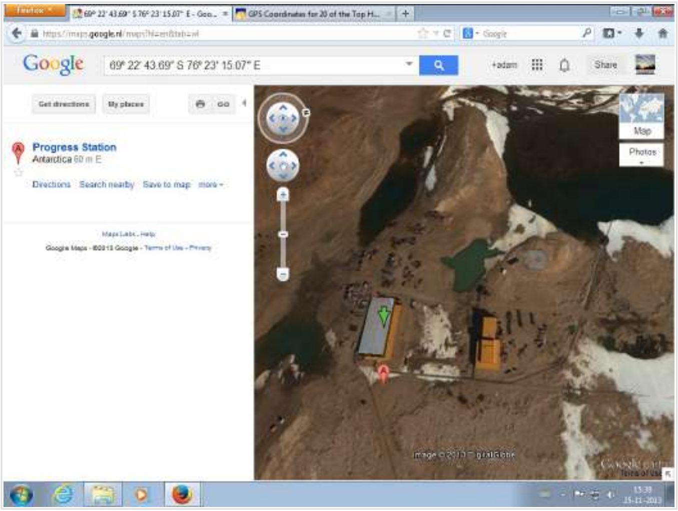
Google
69° 22' 43.69" S 76° 23' 15.07" E
Get directions
My places
Progress Station
Antarctica 60 m E
Directions
Search nearby
Save to map
more
Map
Photos
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/8315126_orig.jpg)

21. Zhong Shan Antarctic Polar Station (China)
69° 22’ 43.69” S 76° 23’ 15.07” E

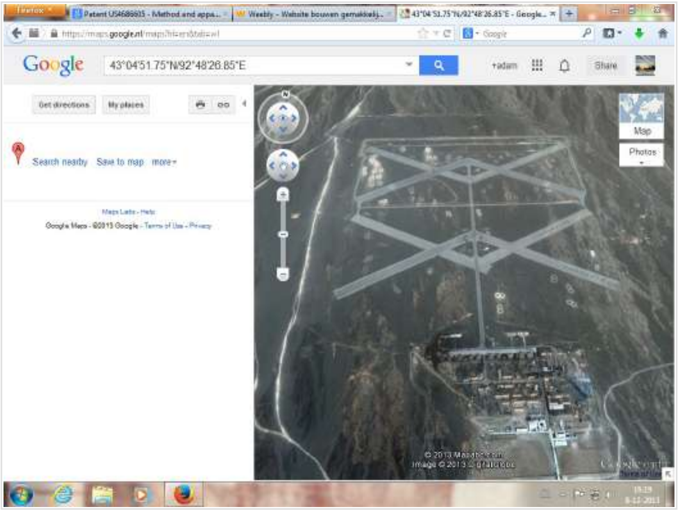
Google
43°04'51.75"N/92°48'26.85"E
Get directions
My places
Search nearby
Save to map
more
Map
Photos
© 2013 Mapabc.com
Image © 2013 DigitalGlobe

(/uploads/2/3/1/7/23175228/4903180_orig.jpg)

22. 43°04'51.75"N/92°48'26.85"E

(/microwaves.html)

(/)

- Chemtrails in Nederland en meer -

CREATE A FREE WEBSITE

POWERED BY

M0UKD

MFJ-971 Portable Antenna Tuner Modifications

Ads by Google Mods Guitar Tuner Tuner Tuner Antenna

The MFJ-971 has the potential to be a great portable tuner. It is a small, compact tuner that has a sensitive SWR circuit allowing precise tuning with low power thanks to its 6w FSD setting and it has outputs for coax, wire and balanced feeder (although I won't talk about the MFJ 'heavy duty' 1:4 voltage baluns here!). To make it the almost perfect portable / QRP tuner, there were a couple of things that needed to be solved...

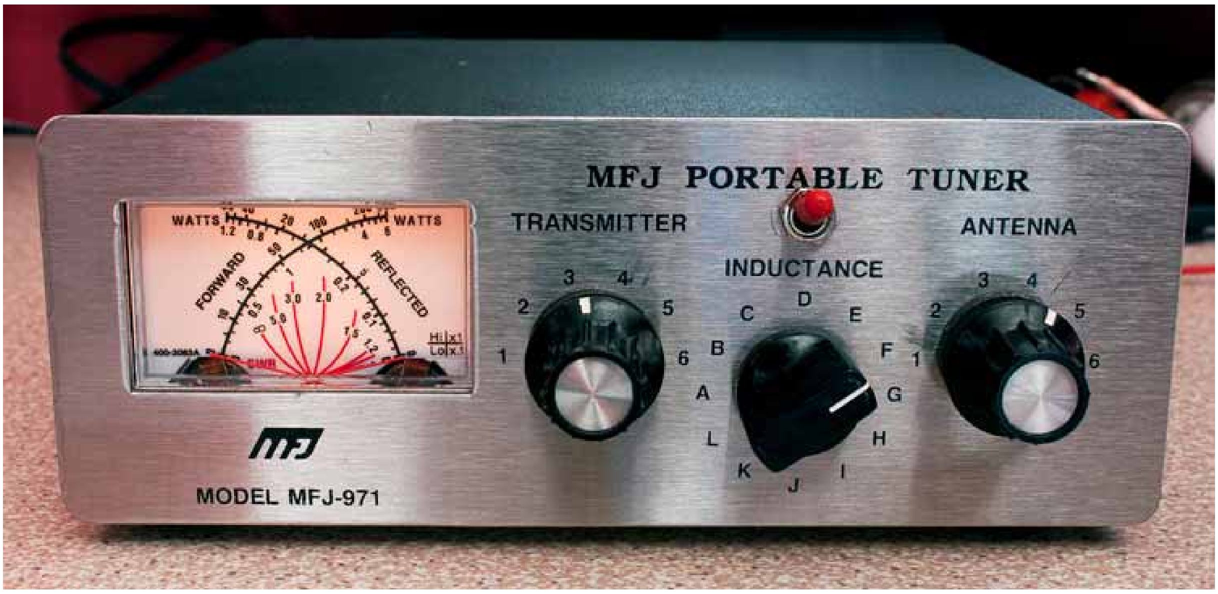
MFJ PORTABLE TUNER
TRANSMITTER
INDUCTANCE
ANTENNA
WATTS
FORWARD
REFLECTED
SWR
MODEL MFJ-971

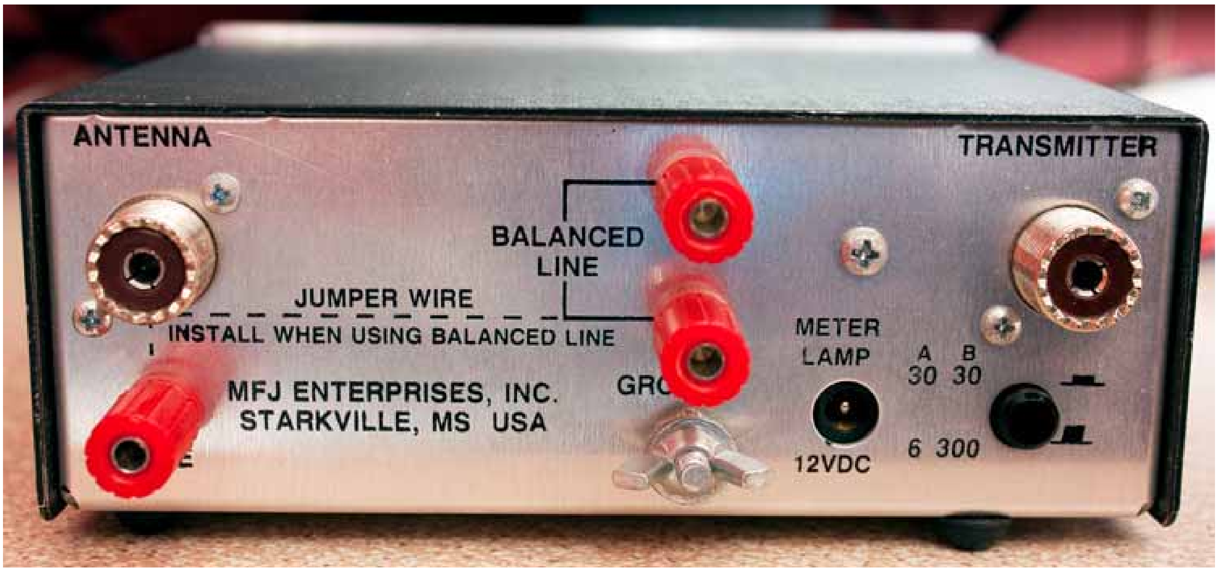
ANTENNA
TRANSMITTER
BALANCED LINE
JUMPER WIRE
INSTALL WHEN USING BALANCED LINE
MFJ ENTERPRISES, INC.
STARKVILLE, MS USA
METER LAMP
12VDC
A B
30 30
6 300

The problems I had with this tuner are:

1) No bypass switch

2) Not enough inductance to tune on 1.8MHz

The first thing that I would want in any tuner, is a bypass switch. Its nice to flick a switch and be able to receive on all bands and the lack of a bypass switch is frustrating. This was easily added by mounting a DPDT switch to the front panel, and connecting it appropriately, as seen in the pictures below.

Full SiC Modules ROHM
Constructed by Original SiC-MOS/SBD
Suited for replacemet of IGBT
www.rohm.com

Coil Bobbins
Inch and Metric coil bobbins 5000+ standard range available
www.milesplatts.co.uk

Isolators and Circulators
300MHz to 20GHz: Coax, Drop-in & WG
Fast Delivery, Manufactured in UK.
www.mcsmicrowave.co.uk

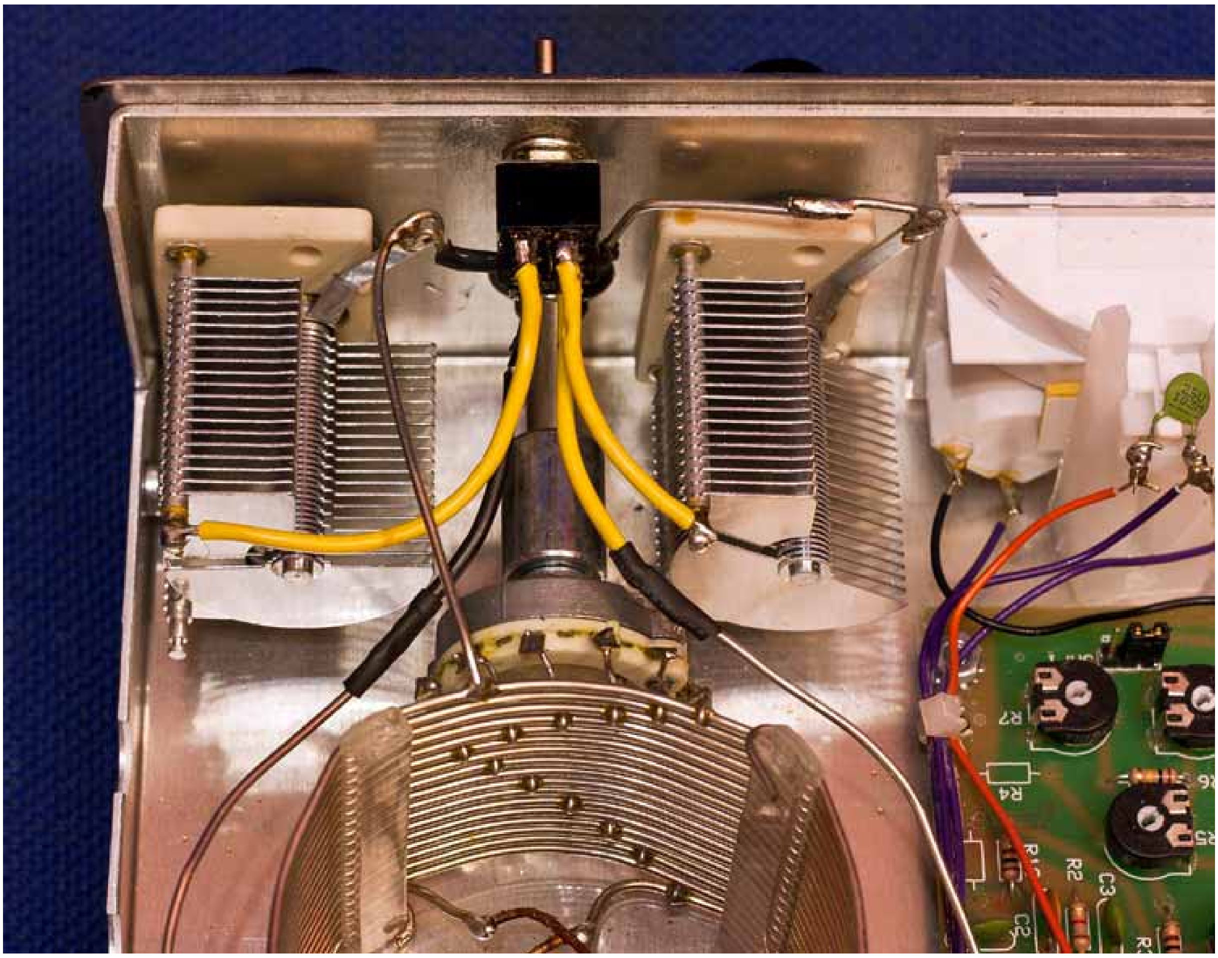

Usually, the signal comes off the PCB and straight into the 'transmitter' capacitor and the output of the 'antenna' capacitor goes to the coax/wire connection on the rear. The switch has been added to provide a way to bypass the tuner section. A drawing of the DPDT switch wiring is below.

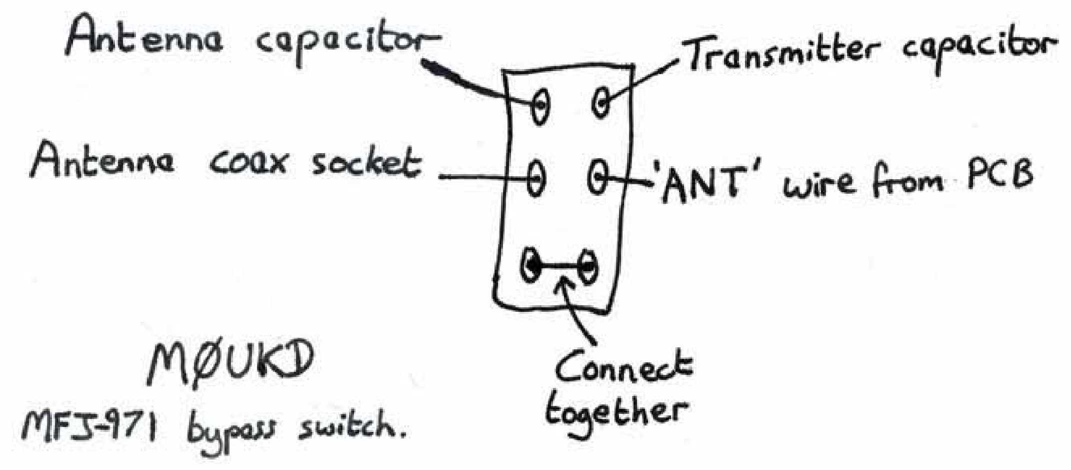

OK, now that annoyance is out of the way, lets fix the top band issue. MFJ say *'The [MFJ-971] T-match tuner covers*

1.8-30 MHz'. It may well do, if you want to match a perfect 50Ω load on 1.8MHz! The inductor inside this tuner, although adequate for 80-10m, does not have enough inductance to be useful on 160m.

To overcome this problem, I included extra inductance by winding 29 turns of enamelled copper wire (any suitable insulated wire will do) on a T130-2 iron powder toroid, so when the tuner is on the 'L' inductance setting, the whole original inductor is used *plus* the extra inductance provided by the toroid. This does not affect any of the other settings from A-K.

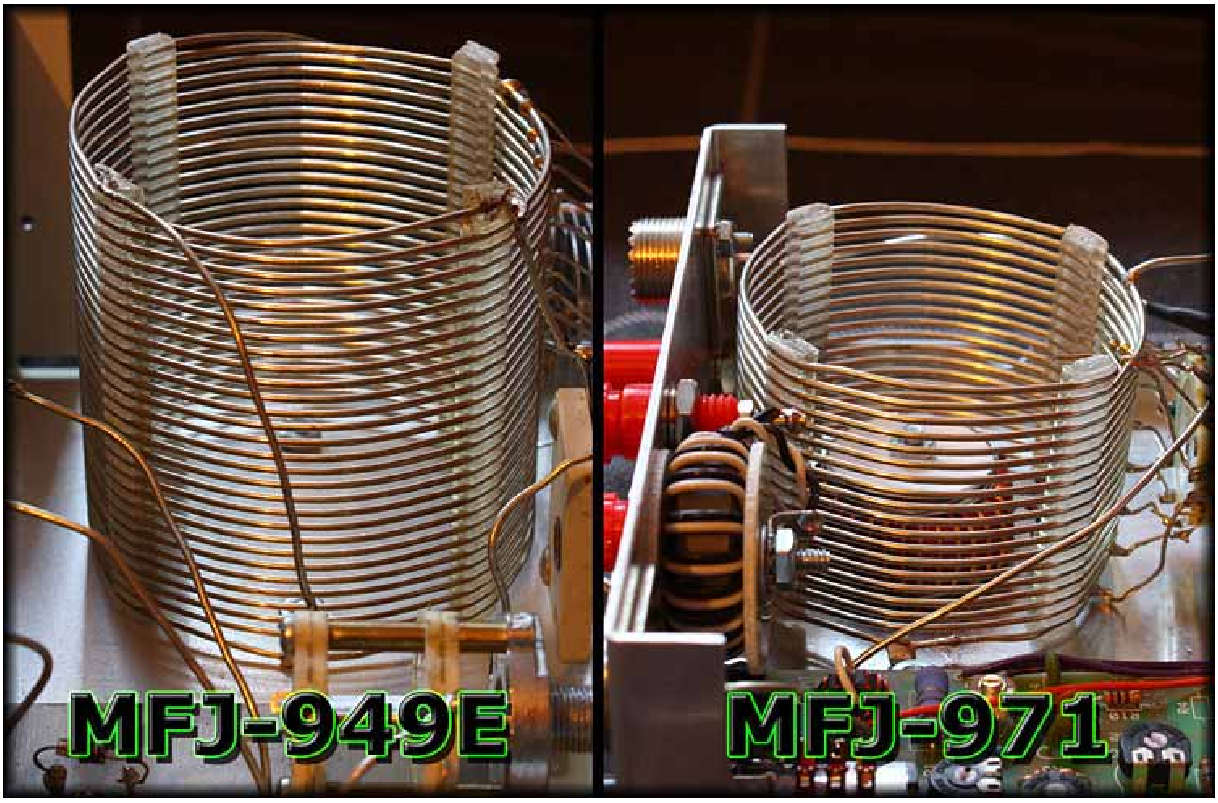

The MFJ-949E inductor has 27 turns, 75mm long and 75mm diameter which is about 37µH.
The MFJ-971 inductor has 17 turns, 43mm long and 66mm diameter which is about 17µH.
Adding on the extra inductance means we have a max inductance of around 27µH which allows matching on 1.8MHz.
Air inductance coils can be calculated **here** and the T130-2 toroid **here**.

High Power Fiber Laser 15W single frequency with low noise 1064 and 1550 nm ranges www.nktphotonics.com/high_power

1553 Box/Flange Couplers Mfr of flight worthy, 1 to 8 stub, 4-hole mount bus couplers. www.databusproducts.com

Coil Winding Specialist Design and Manufacturing of Chokes, Inductors, Transformers, Coils, CTs www.coilws

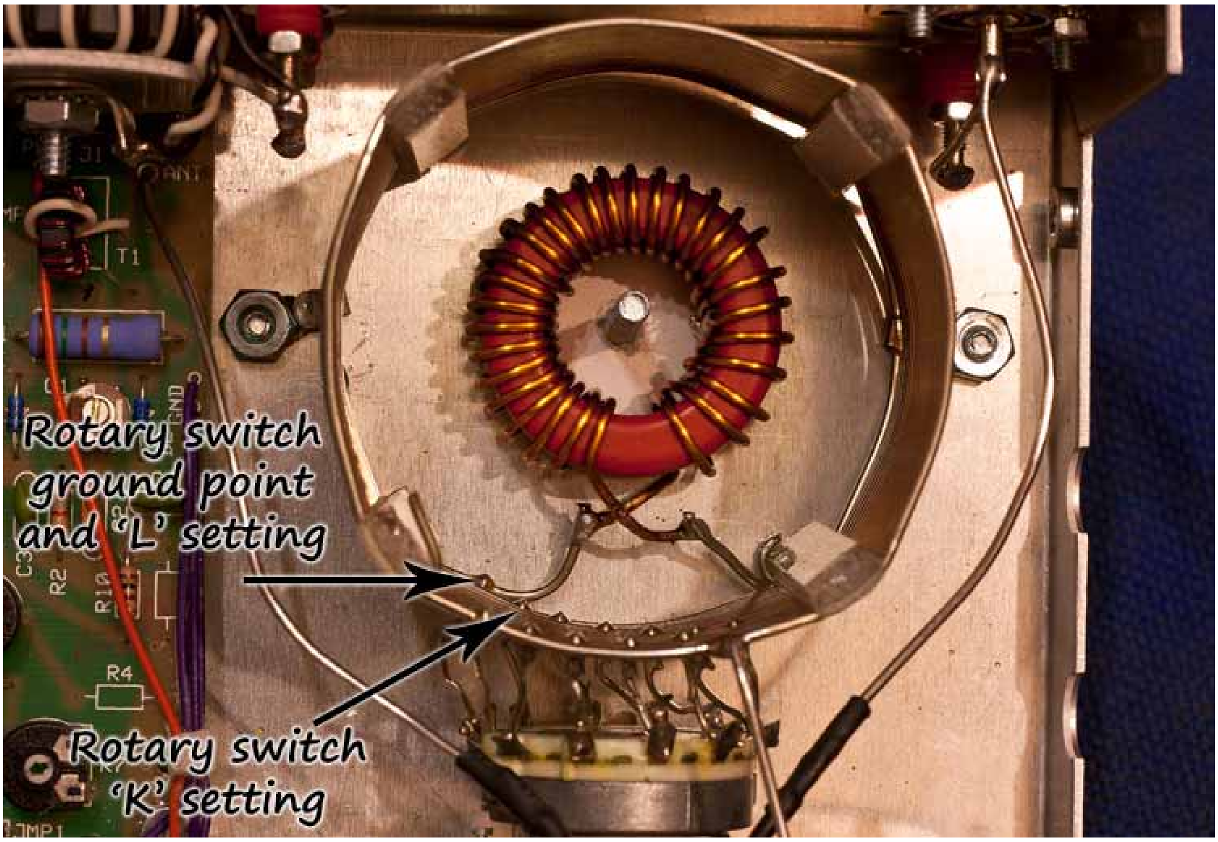
Rotary switch ground point and 'L' setting
Rotary switch 'K' setting

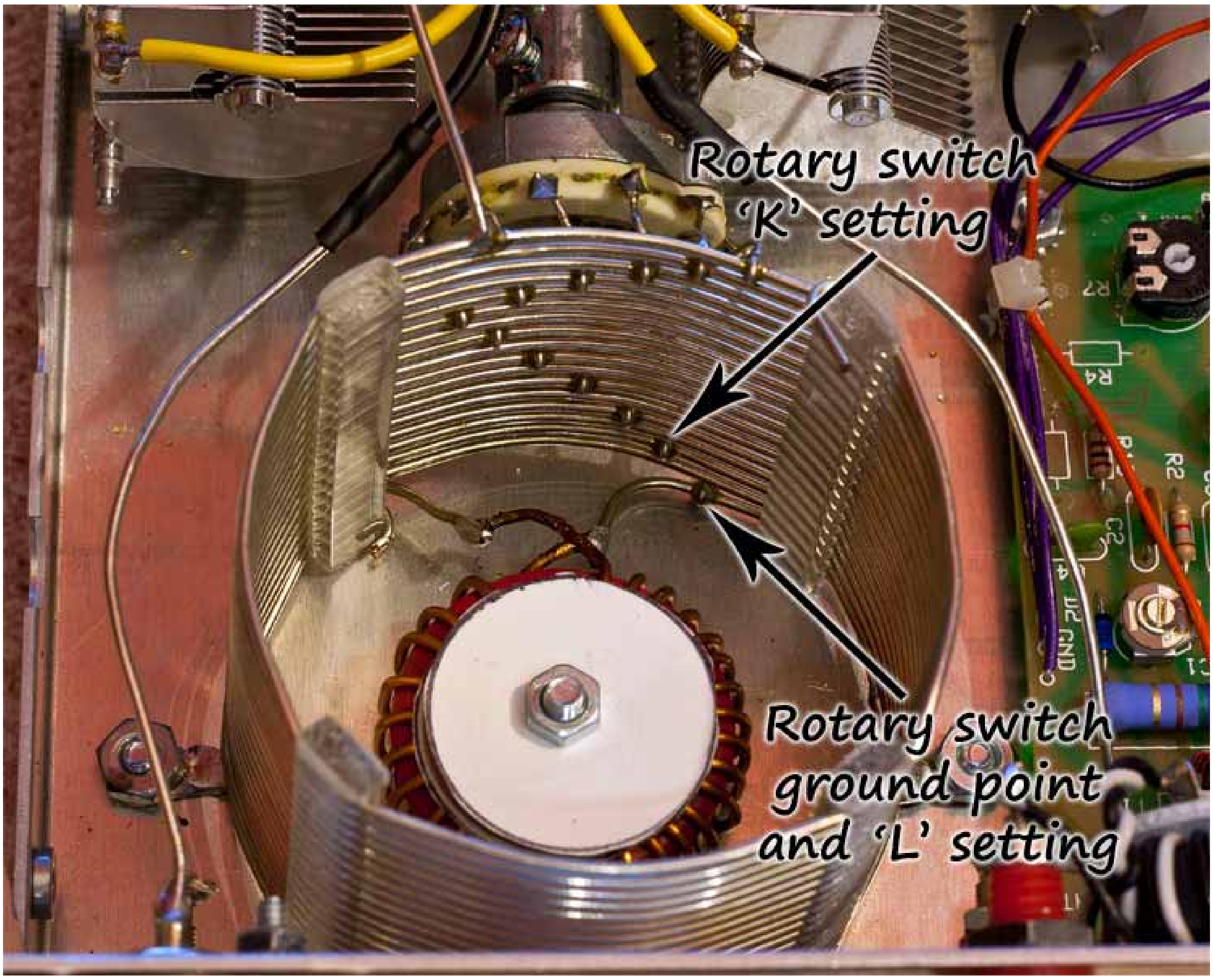
Rotary switch 'K' setting
Rotary switch ground point and 'L' setting

The inductor switch picks up its ground from the bottom of the inductor and this is also the 'L' setting, meaning the whole coil is in use when 'L' is selected. The 'K' inductance setting is 2 turns up from this point. I cut the wire just after the 'L' tap and inserted the extra inductance. Now, when 'L' is selected, the whole original inductor is in use, plus the extra inductance of the toroid, making it suitable for use on 160m! Of course, the 'K' setting still uses the whole original coil (less 2 turns) so nothing is really lost. The wires cross to keep any magnetic fields circulating the same way between the 2 inductors. The toroid is then fixed with a nut & bolt and suitable insulation washers either side of the toroid.

DANGER
MFJ-971
May give you
RF Burns!

The completed, modified, much improved MFJ-971 is seen below. Please let me know if you perform any of these modifications, I'd be interested to know what you think.

There is one small problem remaining and that is RF BURNS! The nuts that hold the variable capacitors in place on many MFJ tuners are at RF potential. Most of the time they are hard to come into contact with being tucked away behind the control knobs, however, on the MFJ-971 the knobs must be pulled forward a little otherwise the metal inside the knob comes into contact with the metal shaft causing the grub screws that hold the two control knobs on to be at a high RF voltage. This can give you RF burns whilst you are tuning, even at 5 watts. I hate to imagine what the result would be if you put the 'rated' 200w into this thing and touched a control! One problem with pulling the knobs forward a little is that then the nuts themselves are more visible, although anything is better than an RF burn! 73 de M0UKD.

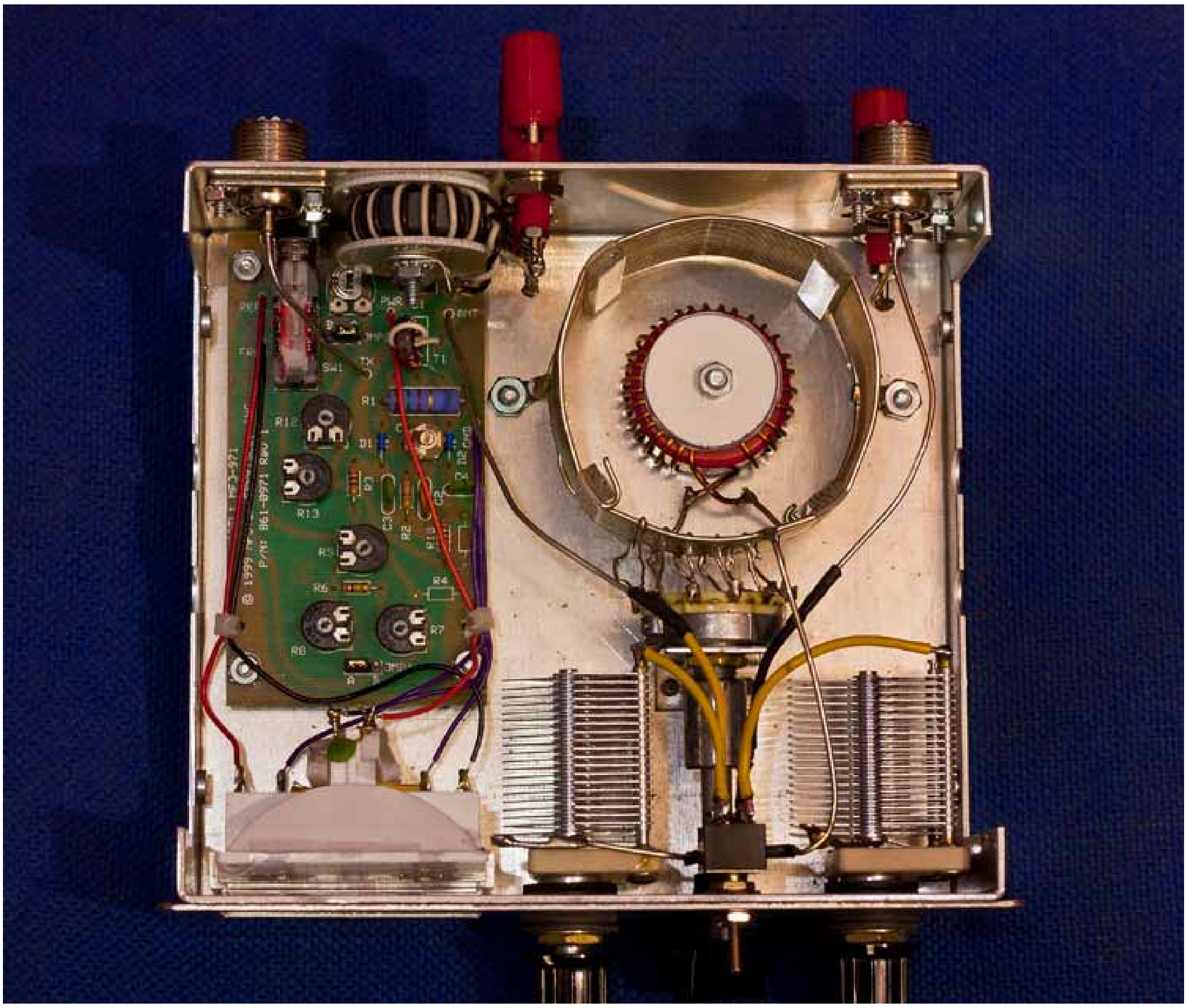

Ads by Google Mods Guitar Tuner Tuner Tuner Antenna

Ugly/Ground Plane Construction

The Method I Use in Building My Simple Receivers

Oct. 7, 2006 by Rick Andersen, KE3IJ

Occasionally I get emails from readers who want to know how I build these simple radio receiver circuits on an "L-shaped chassis" without cutting and scoring the copper, what I mean by "L-shaped chassis," or "what's all this about 10 megohm resistors?", etc. I tend to forget that not everybody has done this before, and I usually take it for granted that the reader is an accomplished "basement-tinkerer" and should therefore be able to just look at a schematic diagram and then come up with his own layout and method of construction. Sorry if I have annoyed anyone -- that wasn't my intention.

When I first design a circuit, I make what a friend of mine used to refer to as a "spider" (or, "spider-web")-- basically just construct the circuit by soldering the components together, lead by lead, into a reasonable facsimile of the schematic (no chassis or pc board necessary-- the stiffness of the soldered leads themselves hold the "spider" circuit together). This method is quick, and allows me to modify things fairly easily. It also grows rather quickly into a "haywire monstrosity" if I allow it.

Sometimes I do use a piece of copper-clad pc board as a ground plane, even during the "spider" stage, as the photo below illustrates:

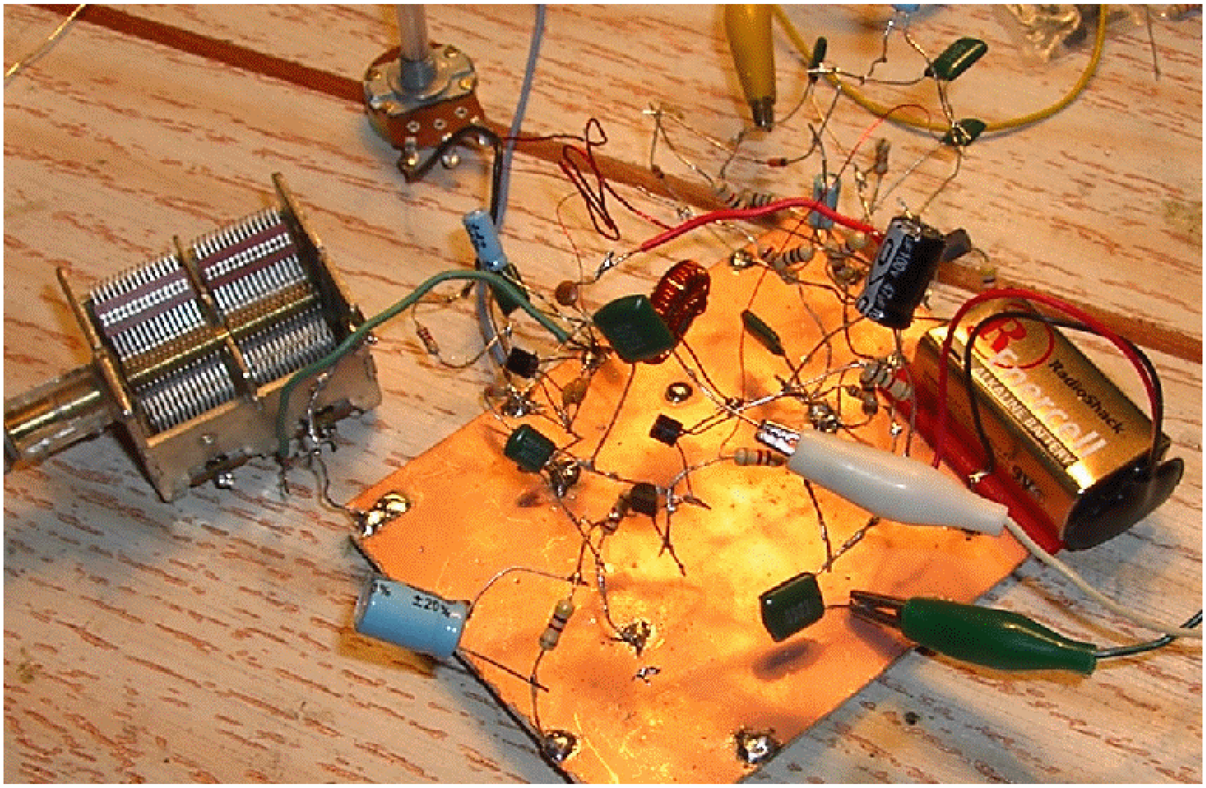

A "haywire" mess of my "AGC-80" Regen receiver in its early, experimental stage

To make things a bit clearer to you who have never built projects in this way before, I have attempted a free-hand diagram using Microsoft "Paint", which I converted to a .gif file and which you see below.

In the top-center of the diagram is the schematic for a simple common-emitter audio amplifier stage. The left side of the diagram illustrates how one might translate that schematic into an actual circuit on the copper-clad pc board. At right is a diagram of the "L-shaped Chassis." This is the style in which I build most of my circuits.

Basically, I use 2 pieces of 5x7 inch copper-clad pc board, available in the USA at **Radio Shack** [free plug]. One piece lies flat (the bottom piece), which is what I have shown in the diagram. I build my circuit on this bottom piece, using the copper floor as my ground connections ("ground plane"). Any resistors, caps, etc., that go to ground in the schematic, are literally soldered to this bottom ground plane and stood upright. At their top ends, the rest of the circuit is soldered, suspended up above the ground plane. Anywhere that I think I need some mechanical stability, I use a vertical 10 Megohm resistor, which acts to prop up and support the rest of the wiring above the ground plane. For most of the circuits I build, 10 Megohms is much higher a resistance value than any of the resistors in the circuit itself; therefore, the circuit pretty much ignores the 10M "posts", electrically, and they remain simply a mechanical "standoff insulator". The 2nd piece of copper clad board is drilled for switches, potentiometers, tuning caps, and earphone jacks, etc., and then is soldered at right angles to the flat bottom piece, making an "L-shaped chassis". It also helps to solder triangular pieces of copper-clad board as mechanical braces, at the corners where the front panel is soldered to the bottom ground plane.

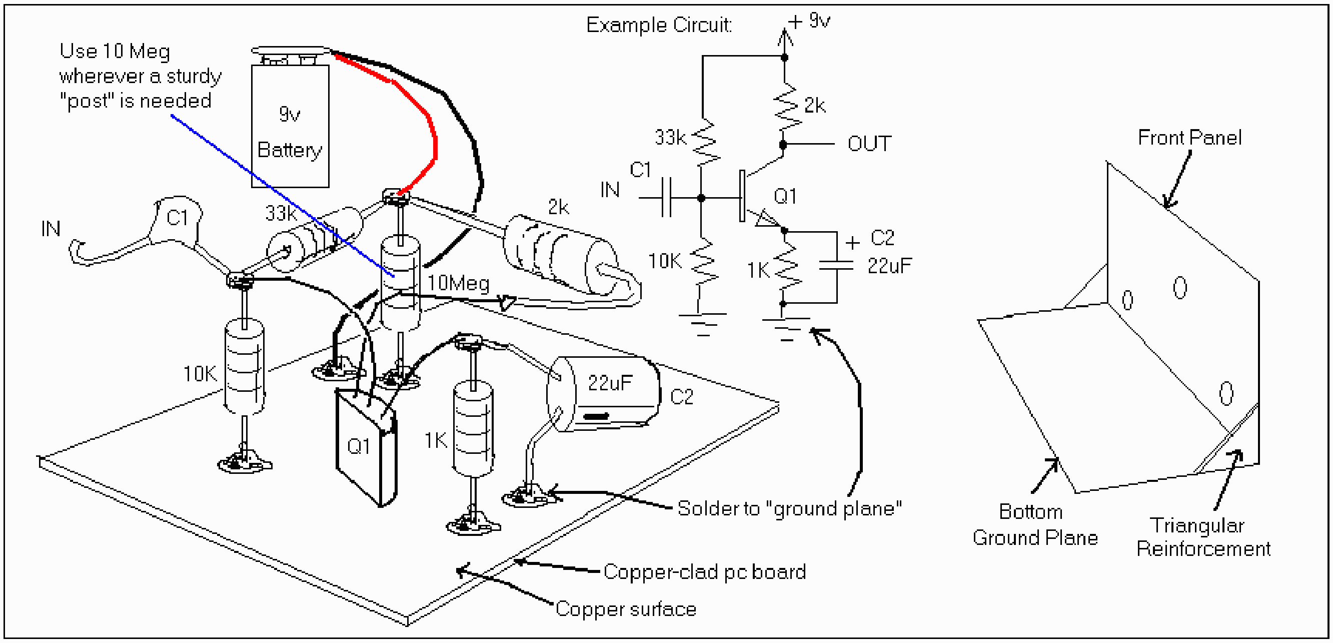
Use 10 Meg
wherever a sturdy
"post" is needed
9v
Battery
IN
C1
33k
2k
10Meg
10K
Q1
1K
22uF
C2
Solder to "ground plane"
Copper-clad pc board
Copper surface
Example Circuit:
+ 9v
2k
OUT
33k
C1
IN
Q1
+ C2
22uF
10K
1K
Front Panel
Bottom
Ground Plane
Triangular
Reinforcement

Below is a photo of my 2006 creation, the "AGC-80", a Regen receiver for the 80 Meter ham band with audio-derived AGC. There's an article on this receiver elsewhere on this website; meanwhile, notice that it is built in the same way outlined above-- an "L-shaped" chassis made of a bottom piece of copper-clad board, soldered to a vertical piece of the same kind of board which is the front panel. [I also added a back wall to hold the antenna jack, and a right wall to hold the audio output jack, as well as a triangular reinforcement piece of copper-clad, at bottom-left between the front panel and bottom piece, although you can't see it in this photo.]

To the right of the AGC-80 is the infamous Radio Shack/Archer little beige Amplified Speaker that I'm always mentioning. I may knock Radio Shack for other reasons [for example, that they have pretty much left us hobbyists behind, and have become just another consumer electronics store] but I will always say good things about their little $12.95 beige-colored Amplified Speaker. It works like a charm for the kind of projects I build. I also run an audio cable out of the Radio Shack amp to a bigger speaker, for better-quality sound-- in the photo you can see a white "Optimus" computer speaker pressed into this service. While the Radio Shack's speaker does a fine job, the Optimus sounds quite a bit louder and way nicer. Of course, you may argue that a *real* Electronics Geek rolls his *own* audio amp rather than rely on a commercially-made, external one. Well, I'm often too impatient to build the audio power amp myself, after having spent hours or days getting the RF part of the radio to work the way I want. Also, homemade audio amps tend to oscillate (squeal, motorboat, howl) in mysterious ways once lovingly installed on the same circuit board as the RF circuitry.... After a while one grows tired of trying 50 ways to make it happy, and would rather just hook up an Amplified Speaker and be done with it.

To each his own.

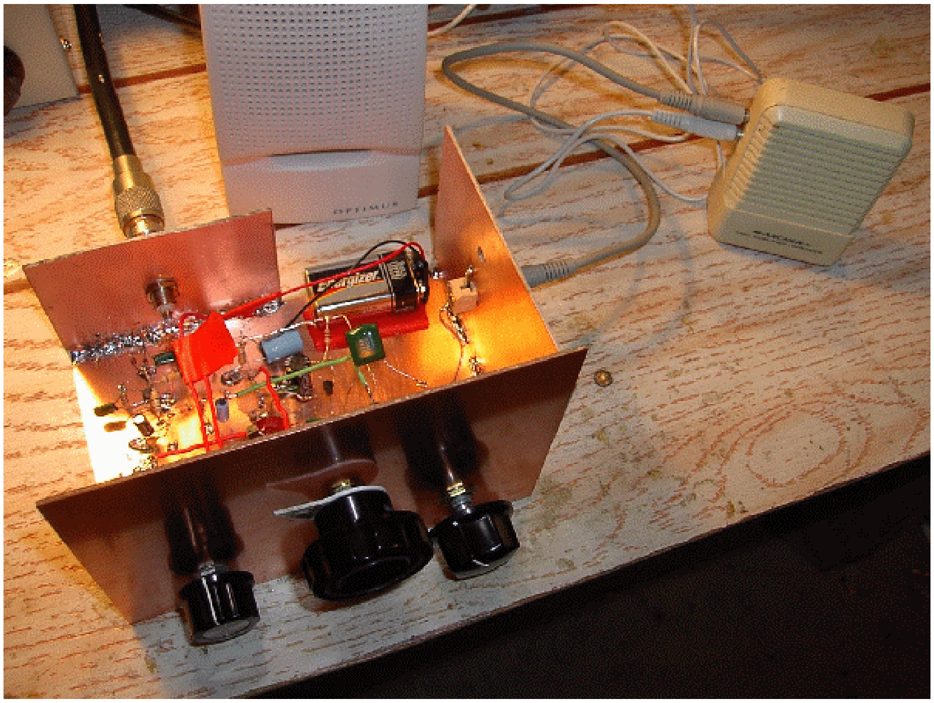

Please realize that you don't have to follow this method of construction (usually called "Ugly Construction" by hams), but I would advise you to make sure there is some equivalent of a copper ground plane in most of your radio projects... it helps to shield the circuit from outside influence (particularly when connected to earth ground) and contributes to stability.

I hope this has helped to give you a little better idea of the methods I use!
73 de Rick, KE3IJ

Back

ON6MU's "Vipormutant" Antenna
RE-AHFV14P

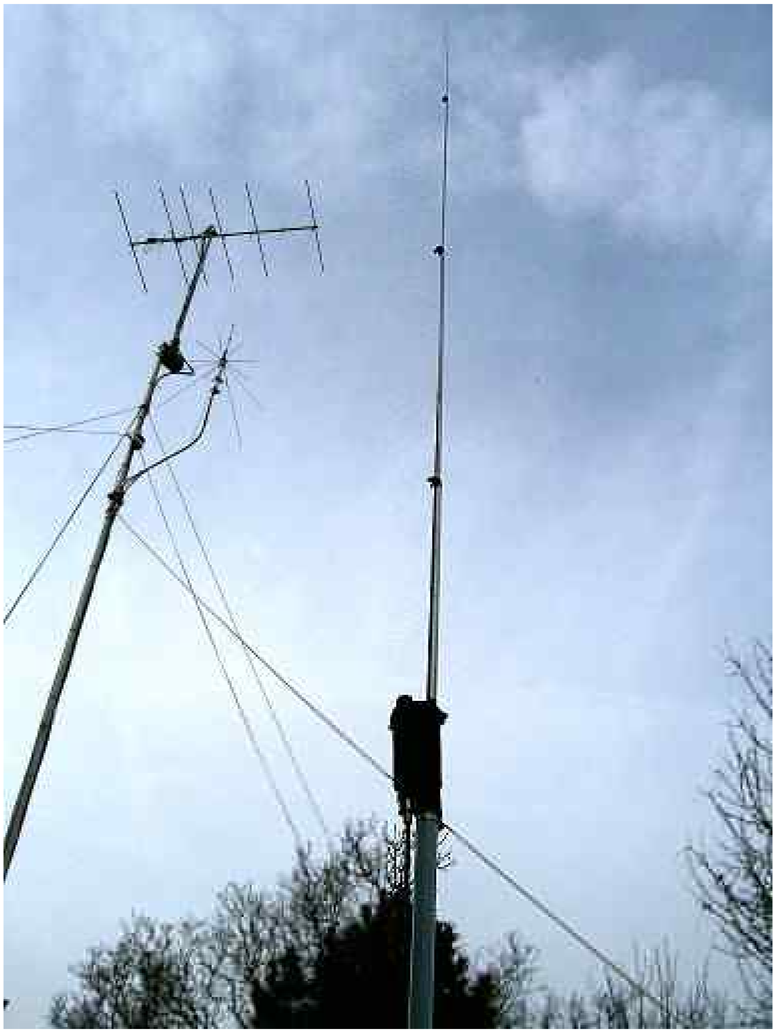

Versatile Inexpensive Portable Multi-band Tunable Antenna

de ON6MU
revision 2

AdChoices RF Antenna Design Antenna 2M 70CM Antenna

Features

- Only 3 meters high fully extended (effective radiating element height)
- Less then one meter inserted (no element is larger then one meter including the tunable section)
- Tunable without the use of an extra tuner (just switch till you get the best SWR)
- Covers all frequencies from UHF to 7 MHz without a tuner
- UHF (tuned by de- or increasing the length of the antenna) 3/4 2*5/8 4*5/8
- VHF (tuned by de- or increasing the length of the antenna) 1/4 5/8 2*5/8
- HF (tuned by switching)
  6 meters (1/4 1/2)
  10 meters (1/4)
  12 meters
  15 meters

  17 meters
  20 meters
  30 meters
  40 meters: with large counterpoise and/or with longer radiating element, or extra tuner
  80 meters: if radiating element is > 5 meters, or with large counterpoise and/or with extra tuner
- Works with or without cointerpoise
- Ideal as portable or balcony antenna
- Compact and extremly portable
- Not too critical on the material or sizes of the elements
- +- 50 watt input
- SWL's Note: tunable on all frequencies between the bands mentioned above

What you need to build the "Vipormutant"

- 5 (or more depending on how high you want your boom) alu tubes
- piece of hard insulating tube (+- 7 cm), examples: plastic, nylon, bamboo...
- some low loss RG174 50 Ohm coax
- carbon/ferrite bead or toroid (to act as a choke)
- a few meters of 0,75mm enamelled copper wire to make the coil
- SO239 (PL259 female)
- Paint, silicon, glue etc. to seal things up
- Plastic box to mount over the coil and where we'll put the switch and SO239 connector
- 12 position rotary switch
- a few innox hose clamps

About the "Vipormutant" antenna:

Well yes, one must have a name HI...It is nothing more then a base coil loaded antenna, but with a selector direct on the base to tune the antenna.
Most of us don't have the luxury of building a 1/4, 1/2 or even a 5/8 wavelength vertical antenna for HF. We have to settle for something a little shorter. (A lot shorter, in the case of people following the FCC's Part 15 rules, which limit them to 3 meters in size.) Shorter vertical antennas can give acceptable (but not spectacular) performance.
I needed a highly (HF) portable antenna to use with my FT-817 which should have the highest possible frequency range (also VHF) and still compact enough to take along almost anywhere! The antenna should be versatile enough to allow further experimenting, to allow being mounted on a balcony, caravan, outdoors etc... So I came up with a compact vertical (dismounted no higher then 1 meter) with a "tuner" directly connected to the antenna radiating element (the best possible place a tuner can be).
The "Vipormutant" tuning principle gets energy into the antenna on a wide range of frequencies, but the design of an antenna is what controls what happens to the RF energy from there. For some antennas, the antenna is simply not complete without a radial system, or at least a counterpoise. Other types of antennas need no RF ground system at all. Most reference books on antennas provide solid guidance on radials and counterpoises, but only for antennas cut to a specific frequency. When using the "Vipormutant" it will also act like tuner and at the best place a tuner should be: directly beneeth the antenna! So the rules have to change somewhat because the "Vipormutant" almost operates across the full range of HF frequencies unto UHF. It doesn't need a counterpoise to work, but the efficiency will increase when you do use it.

Considerations:

- If the length of the conductor is very short compared to a wavelength (< wave/4), the electric and magnetic fields will decrease dramatically within a distance of one or two wavelengths.
- It is impossible to make a small antenna to radiate as efficient like a big antenna.
- Ground losses affect radiation patterns and cause high signal losses for some frequencies. Such losses can be greatly reduced if a good conducting ground is provided in the vicinity of the antenna.

The coil/tuner

Wind 0,8mm enamelled copper wire around the isolator (+- 16mm diamter) and make a tap every xx turns (see fig. 2 and 3)

Fig.1

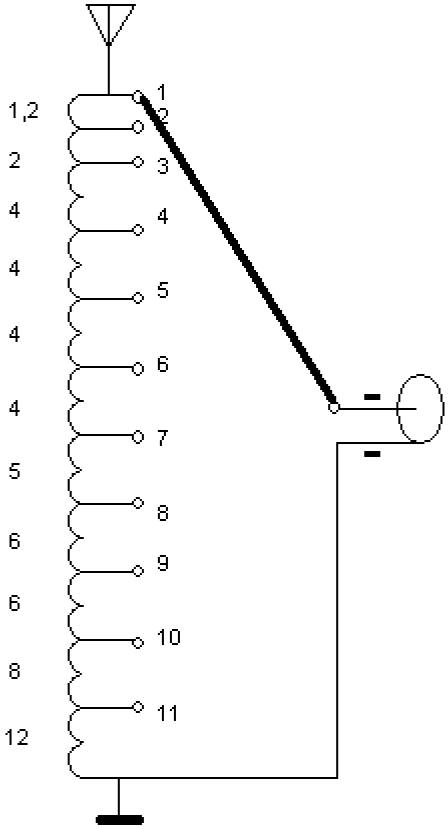

Fig.2

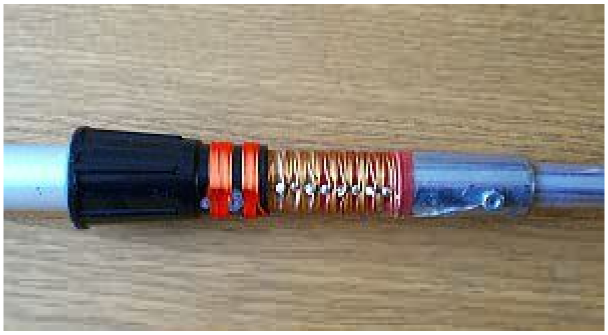

The coil dimensions isn't too critical.

Relatively short antennas behave like lossy capacitors and present a high impedance load to the transmitter due to the large amount of capacitive reactance that is present. The loading coil helps to tune out that reactance. Tuning out the reactance is important because a tuned antenna will accept and radiate much more power than a mismatched antenna.

When the loading coil is installed at the bottom of the vertical radiator, we call it a "base loaded" antenna. Base loading requires the smallest amount of inductance to achieve resonance.

The shoke

Is made out of miniature 50 Ohm coax (rg174) that goes a few turns through the carbon/ferrite bead or toroid. You can also use a Snap-Together Ferrite Choke Core.
If a ferrite is put over a cable which includes both signal and return lines, it will have no effect on the signal (differential-mode) current but it will increase the impedance to common-mode currents. This is because the differential currents, by definition, sum to zero in each wire pair and therefore there is no net magnetic field. If there is no field, the ferrite is invisible. But the common mode currents do produce a net magnetic flux and this flux is concentrated in the bulk of the ferrite, leading to an increased impedance for these currents only. The choke should prevent any mantle currents flowing and should decrease RFI.

The effectiveness can be increased by looping the cable several times through the core, but the benefit is limited at higher frequencies by the stray capacitance between the turns of the cable.

Fig 3.

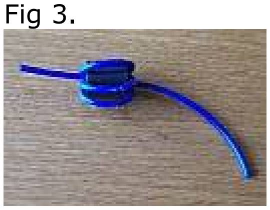

4 turns

Highlighted

Start Download - View PDF

Convert From Doc to PDF, PDF to Doc Simply With The Free Online App! download.fromdoctopdf.com

The base insulator

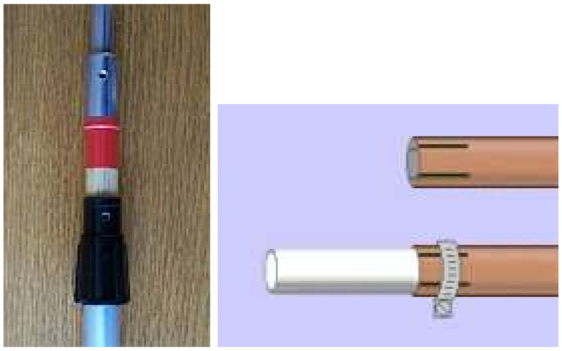

The vertical radiator "driver" element and tuning box

I used a plastic box of 130x70x40mm. On top and bottom I drilled a hole to fit the driver element (radiator of +/- 40cm length)and boom (also +/- 40cm length).
On the side I drilled a hole for the rotary switch and the SO239 connector, whilest on the opposite side I drilled a hole to fit a "ground" bolt where I can easily connect the counterpoise and/or ground to if needed.

Fig 4

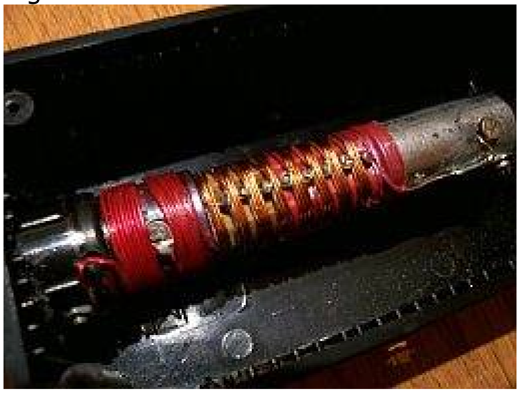

Fig 5

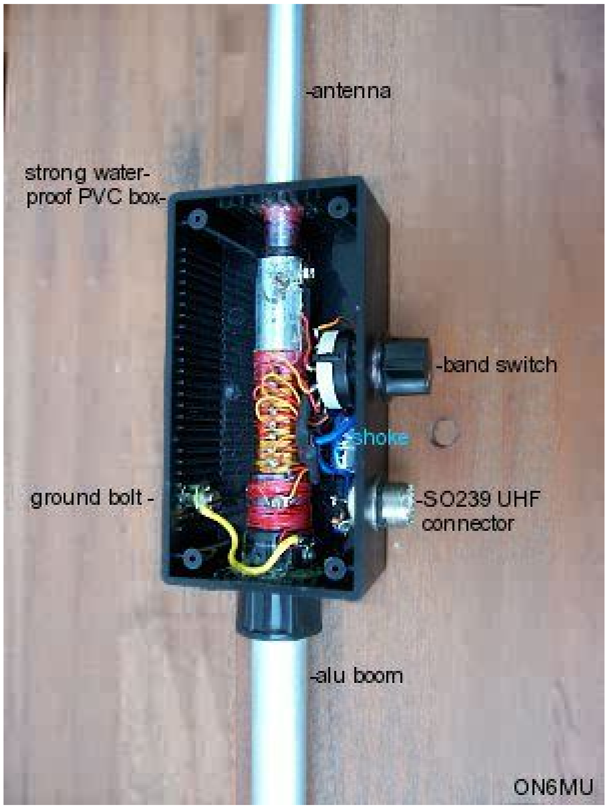
-antenna
strong water-proof PVC box-
-band switch
shoke
ground bolt -
-SO239 UHF connector
-alu boom
ON6MU

This is how things are connected inside the box:

Fig. 6

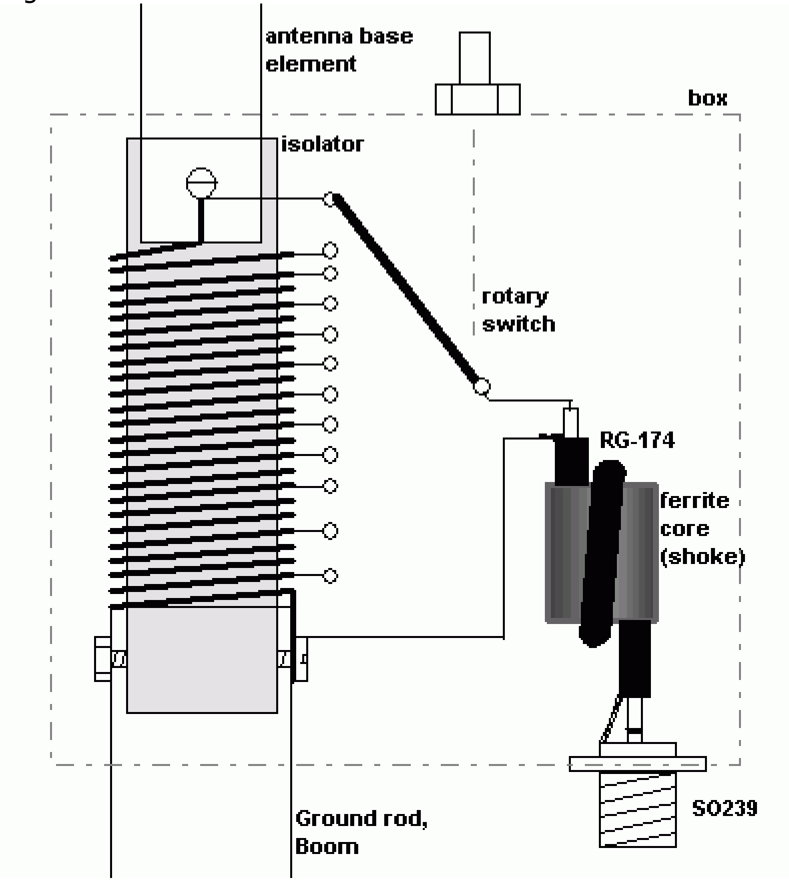
antenna base element
box
isolator
rotary switch
RG-174
ferrite core (shoke)
Ground rod, Boom
SO239

Example

The rotary switch is used for tuning the antenna on each band. The first position allows UHF/VHF ranges.
Tuning is done by sliding in/out of the elements.
The "lower" the switch (higher inductance) the lower the resonance frequency of the antenna.

Remember, and this is important too, to seal everything up so no moisture can penetrate the antenna!
Because the radiating vertical antenna elements are made out of separate pieces that fits inside each other and are tightened by hose clamps, the construction isn't waterproof.
If you use a hollow isolating piece you need to to prevent moisture from getting inside the box (via the places where the elements are hold together). I've used a rubber "stopper" that fits snugly on the bottom of the driver element and glued tight.

Black paint finishes the job:

fig. 7

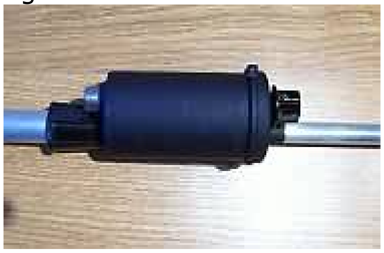

using a round box
In my first design I used a plastic box of 50mm diameter and 9 cm heigh. On top I drilled two holes: one for the driver element (radiator of 40cm) and a hole for the rotary switch (as was used in the first prototype).

fig.8: The main driver element and tuning unit finished

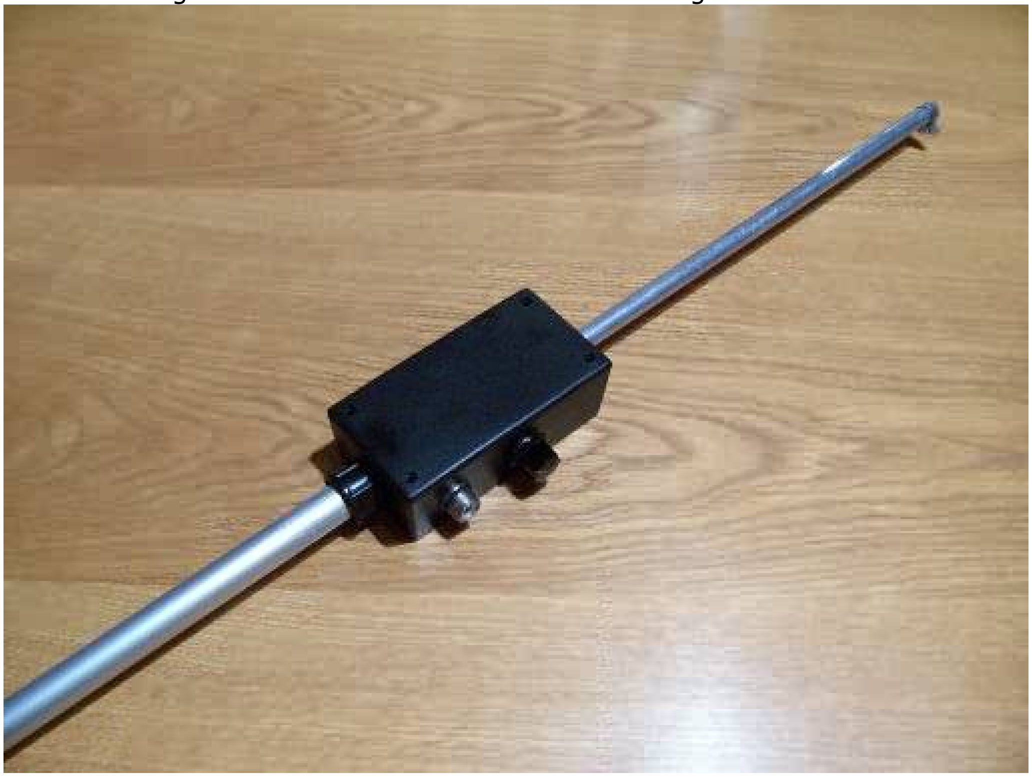

This allows it to be used on almost any boom or can be extended to use with or without vertical elements!
Ideal for experimenting!

Featuring Today

The antenna construction specs

All elements are made out of aluminum.
Fig. 9

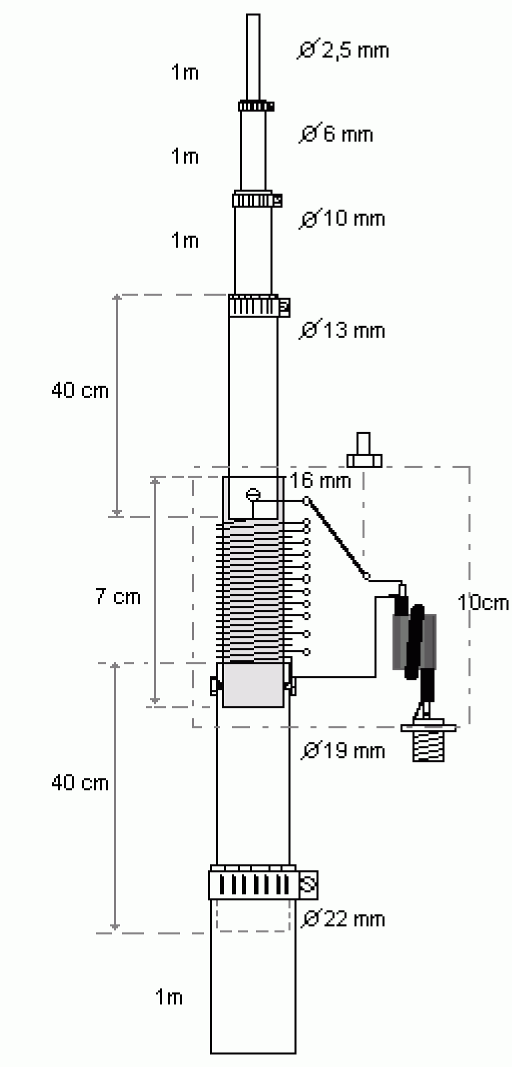
Ø2,5 mm
1m
Ø6 mm
1m
Ø10 mm
1m
Ø13 mm
40 cm
16 mm
7 cm
10cm
Ø19 mm
40 cm
Ø22 mm
1m

This makes the antenna effective radiating elements a total length of 3 meters. The boom elements can be chosen freely and on your needs. A short one (one element of a meter), a medium sized one of several 1 meter tubes or none at all! The bottom piece where the boom is "connected" too is 40 cm and can/could be put directly in the ground (if made pointed for sure). Or you could fix it in a umbrella stand. Use your imagination HI.

Examples of "Vipormutant's" utilization

Used it on a balcony:

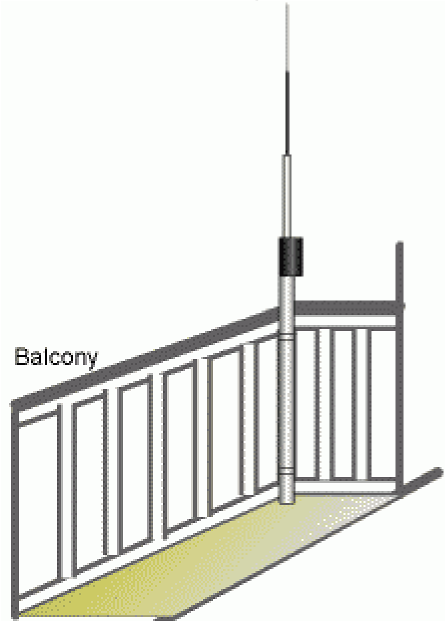

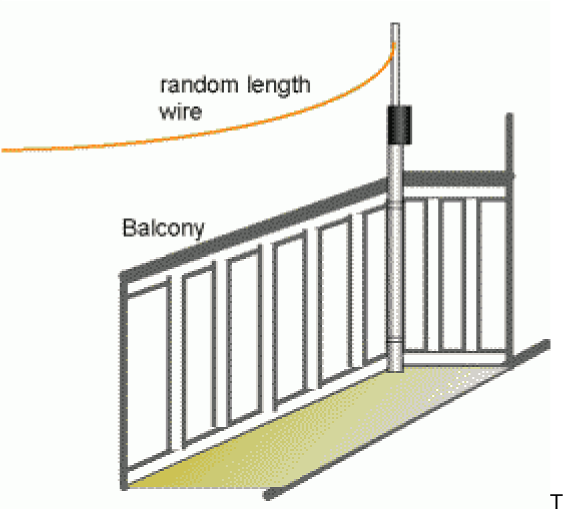
random length
wire
Balcony

This could cover even the lowest HF-bands!

Use it outdoors without grounding:

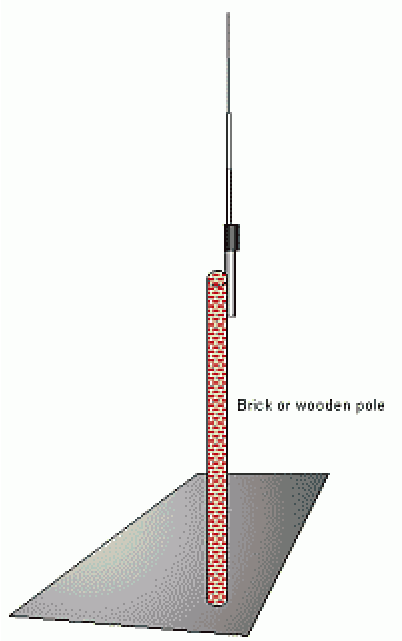
Brick or wooden pole

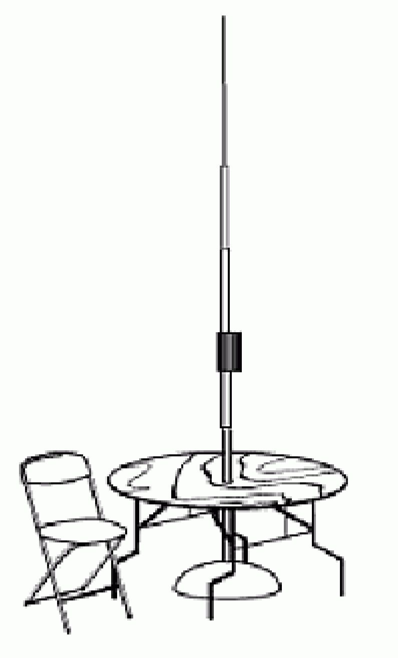

Use it outdoors as a shortened "dipole" balanced antenna (rotary switch set approx. in the middle):

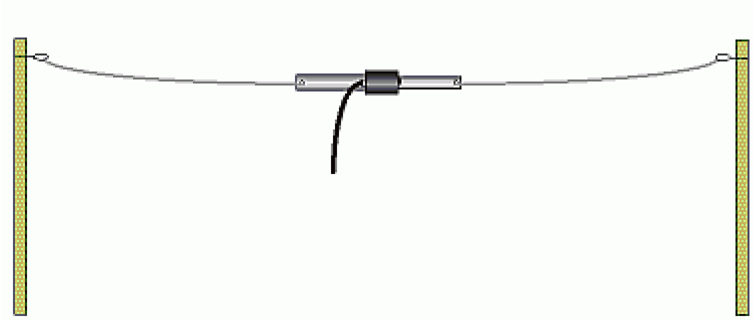

Frequencies below 7 MHz could easily be match 1:1 SWR if total length of the "dipole" > 4 meters
Further tuning can be done with selecting a different impedance using the rotary switch.
Different lengths of wire can be used (an example: one part 3 meter, the other 5 meter or more)

Use it outdoors with ground or counterpoise:

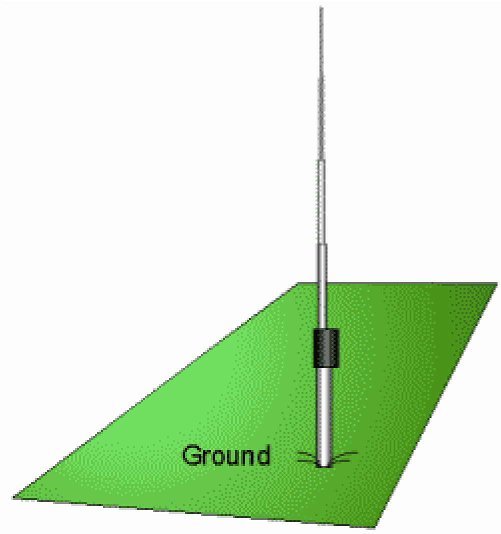

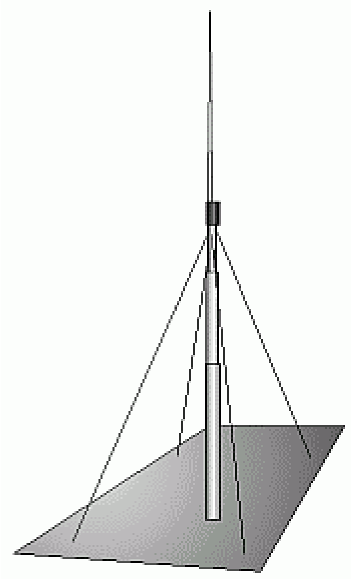

Radials and Counterpoises basic purposes:

1. To improve the RF ground conductivity for the ground current return path. Unless you live in a salt-water swamp, your ground conductivity makes a very poor path for the return of ground currents. This increases the ground losses and reduces the efficiency of an antenna that needs a good RF ground.

2. To provide a counterbalance for the feed point of the antenna to reduce RF radiation back to the radio room. The "Vipormutant" changes the rules because there is no single frequency that you will be operating on, so all of the thumb rules for 1/4 and 1/2 wavelength radials don’t apply. It is possible to be either a purist or a pragmatist in deciding what radials to put in place.

3. Number of radials: More is better, up to a point. In carefully controlled experiments, it has been proved that increasing the number of radials from 2 to 15, or from 4 to 16, produces significant increases in signal strength. Further increasing the number of radials to 60 only produces 1 to 2 dB of increase in field strength. Follow this link to see some of the empirical data.

4. Where to put the radials: For a semi-permanent installation, it is customary to bury the radials a few inches down in the soil. This makes it much easier to mow and walk in the area around the antenna. However, some experimenters have gotten an improvement in performance by raising the radials and the antenna base a few inches above the soil. Raising the antenna and ground system several meters above the earth, for example by installing the base of the antenna on a roof-top, can improve the antenna's performance by reducing capacitive earth losses.

While the "Vipormutant" will provide a good match with a poor RF ground system which will will able you to transmit, your antenna efficiency will be low. Nevertheless, by using a tuning circuit directly at the antenna radiating element losses are kept to a miminum. Getting the greatest efficiency out of your antenna system needs a proper RF ground unless you’re using a balanced antenna system

The efficiency of the antenna increases by using a counterpoise. However, the antenna can be tuned perfectly without!

Don't forget to check these out:

ON6MU Homebrew projects
Radioamateur related projects

ON6MU Ham mods
Modifications of transceivers

ALLIED ELECTRONICS & AUTOMATION Allied Doesn't Take Lighting Lightly MW MEAN WELL SHOP NOW

73"

Have fun and my best 73"

Guy, ON6MU
http://www.qsl.net/on6mu

Comments, pictures or experiences with my antenna are always welcome!

ON6MU

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG IOP Institute of Physics

The Large Hadron Collider

To cite this article: Lyndon Evans 2007 *New J. Phys.* **9** 335

View the article online for updates and enhancements.

Related content

- Protection of the CERN Large Hadron Collider
  R Schmidt, R Assmann, E Carlier et al.
- Accelerator physics at LEP
  D Brandt, H Burkhardt, M Lamont et al.
- LHC as nucleus--nucleus collider
  John M Jowett

Recent citations

- Alexandru Iosup *et al*
- Design of distributed JT (Joule-Thomson) effect heat exchanger for superfluid 2 K cooling device
  S Jeong *et al*
- LHC optics measurement with proton tracks detected by the Roman pots of the TOTEM experiment
  The TOTEM Collaboration *et al*

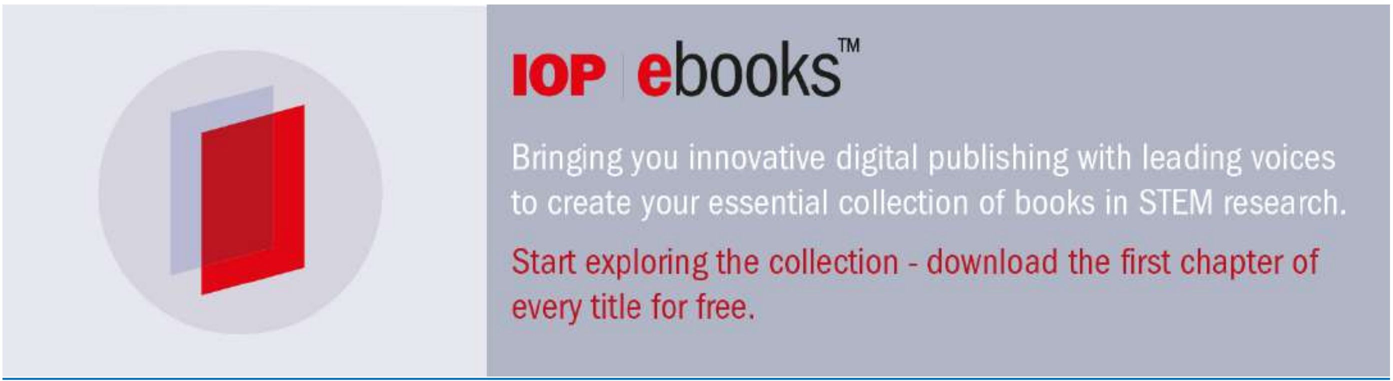

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 172.0.74.23 on 15/01/2019 at 06:21

New Journal of Physics
The open-access journal for physics

The Large Hadron Collider

Lyndon Evans
CERN – European Organization for Nuclear Research, Geneva, Switzerland

New Journal of Physics **9** (2007) 335
Received 8 May 2007
Published 19 September 2007
Online at http://www.njp.org/
doi:10.1088/1367-2630/9/9/335

Abstract. The Large Hadron Collider (LHC), now close to completion at CERN will provide proton–proton collisions with unprecedented luminosity and energy. It will allow the Standard Model of physics to be explored in an energy range where new phenomena can be studied. This includes the validity of the Higgs mechanism, supersymmetry and CP violation. The machine presents a number of novel features discussed in detail below.

Contents

1. Introduction **2**
2. Machine design **2**
  2.1. Machine layout . . . 4
  2.2. Main hardware systems . . . 6
  2.3. Magnets . . . 6
  2.4. Cryogenics . . . 8
  2.5. The radiofrequency acceleration system . . . 11
  2.6. The vacuum system . . . 14
3. Accelerator physics issues **15**
  3.1. Dynamic aperture . . . 15
  3.2. The beam–beam interaction . . . 16
  3.3. Intrabeam scattering . . . 16
  3.4. Coherent instabilities . . . 19
  3.5. Electron cloud effects . . . 20
4. Conclusions **21**
References **22**

New Journal of Physics **9** (2007) 335
PII: S1367-2630(07)50015-7
1367-2630/07/010335+22$30.00
© IOP Publishing Ltd and Deutsche Physikalische Gesellschaft

1. Introduction

The Large Hadron Collider (LHC) at CERN is now in its final installation and commissioning phase. It is a two-ring superconducting proton–proton collider housed in the 27 km tunnel previously constructed for the large electron positron collider (LEP). It is designed to provide proton–proton collisions with unprecedented luminosity ($10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) and a centre-of-mass energy of 14 TeV for the study of rare events such as the production of the Higgs particle if it exists. In order to reach the required energy in the existing tunnel, the dipoles must operate at 1.9 K in superfluid helium. In addition to p–p operation, the LHC will be able to collide heavy nuclei (Pb–Pb) with a centre-of-mass energy of 1150 TeV (2.76 TeV $\mathrm{u}^{-1}$and 7 TeV per charge). By modifying the existing obsolete antiproton ring (LEAR) into an ion accumulator (LEIR) in which electron cooling is applied, the luminosity can reach $10^{27}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

The LHC presents many innovative features and a number of challenges which push the art of safely manipulating intense proton beams to extreme limits. The beams are injected into the LHC from the existing super proton synchrotron (SPS) at an energy of 450 GeV. After the two rings are filled, the machine is ramped to its nominal energy of 7 TeV over about 28 min. In order to reach this energy, the dipole field must reach the unprecedented level for accelerator magnets of 8.3 T. This high field can only be achieved using ‘conventional’ and affordable superconducting material (NbTi), by cooling the magnets in superfluid helium at 1.9 K. The cryogenic equipment needed to produce the 100 tons or so of superfluid helium is unprecedented in scale and complexity.

The tunnel diameter in the regular arc is only 3.8 m, insufficient for the installation of two separate rings. The two rings are therefore incorporated into a single magnetic structure with two sets of coils in a common yoke and cryostat.

2. Machine design

At nominal luminosity, the energy stored in each beam is more than 350 MJ (equivalent to 80 kg of TNT). This is more than two orders of magnitude in stored energy, and three orders of magnitude in energy density (due to the very small beam emittance) than in any other previous machine (figure 1). It imposes unprecedented conditions on the reliability of the safety systems which must abort the beams cleanly if necessary as well as on the collimation systems which protect the machine and detectors from halo particles.

The number of events per second generated by beam–beam collisions for a given process is given by:

$$N = L\sigma,$$

where $\sigma$is the cross-section for the process in question and L is the luminosity. For the study of very rare events the luminosity must be as high as possible. The luminosity depends only on beam parameters and can be written for a Gaussian beam profile as:

$$L = \frac{N_\mathrm{b}^2\, n\, f_\mathrm{r}\, \gamma}{4\pi\ \varepsilon_\mathrm{n}\, \beta*},$$

where N_b is the number of particles per bunch, n the number of bunches per beam, f_r the revolution frequency (11.245 kHz), γ the relativistic γ factor, ε_n the normalized transverse emittance and β^* the β function at the collision point. As well as the rare hard collisions between

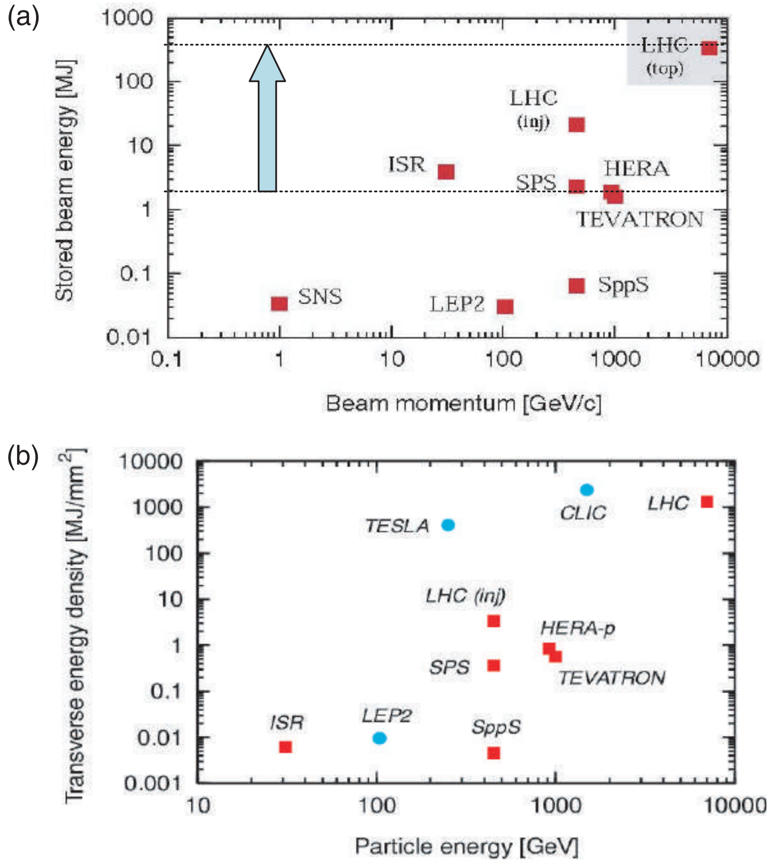

Figure 1. (a) Energy stored in the accelerator beam, as a function of beam momentum. At less than 1% of nominal intensity LHC enters new territory. (b) Stored energy density as a function of beam momentum. Transverse energy density is a measure of damage potential and is proportional to luminosity.

particles, the particles in one beam experience the global electromagnetic field of the other beam (the beam–beam interaction). This force is very nonlinear and gives rise to unwanted effects if not limited to as small a value as possible. Therefore, the bunches must have a crossing angle in order to limit the beam–beam force to the colliding bunches. This produces a geometrical luminosity reduction factor F given by:

$$F = 1/\sqrt{1 + \left(\frac{\theta_c \sigma_z}{2\sigma^*}\right)},$$

where θ_c is the full crossing angle at the interaction point (IP), σ_z the rms bunch length and σ^* the transverse rms beam size at the crossing point.

Table 1 shows the main parameters required to reach a peak luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ for proton–proton collisions at 14 TeV centre-of-mass. It can be seen from the table that for the first time in a hadron machine, the synchrotron radiation at top energy is not negligible. At 3.6 kW per beam it is still very small compared with lepton storage rings and has no influence on the design of the radio frequency system, but this power is radiated into a cryogenic environment and strongly influences the design of the vacuum and cryogenic systems.

Table 1. Performance parameters.

Circumference	26.7 km
Beam energy at collision	7 TeV
Beam energy at injection	0.45 TeV
Dipole field at 7 TeV	8.33 T
Luminosity	$10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$
Beam current	0.56 A
Protons per bunch	1.1×10^{11}
Number of bunches	2808
Nominal bunch spacing	24.95 ns
Normalized emittance	3.75 μm
Total crossing angle	300 μrad
Energy loss per turn	6.7 keV
Critical synchrotron energy	44.1 eV
Radiated power per beam	3.8 kW
Stored energy per beam	350 MJ
Stored energy in magnets	11 GJ
Operating temperature	1.9 K

2.1. *Machine layout*

The basic layout of the LHC follows the LEP tunnel geometry and is shown in figure 2. The machine has eight arcs and straight sections. Each straight section is approximately 528 m long. Four of the straight sections house the LHC detectors whilst the other four are used for machine utilities, radio frequency, collimation and beam abort. The two high luminosity detectors are located at diametrically opposite straight sections. The ATLAS detector is located at point 1 and CMS at point 5, which also incorporates the small angle scattering experiment TOTEM. Two more detectors are located at point 2 (ALICE) and at point 8 (LHCb), which also contain the injection systems for the two rings. The beams only cross from one ring to the other at these four locations.

In order to leave enough space for the LHCb spectrometer magnet in the already existing experimental cavern at point 8, the IP at this location is shifted by 11.25 m towards point 7. Bunches with the nominal 25 ns separation collide at all four crossing points simultaneously. However, because of this displacement, the only other bunch separation with simultaneous collisions in all four detectors is 75 ns. For early commissioning, this bunch separation is interesting because there is no possibility of electron cloud buildup (see below) and the number of long-range beam–beam interactions is reduced, simplifying early machine operation. The 25 GeV proton synchrotron (the first circular machine in the injector chain), where the LHC bunch structure is generated, has therefore been equipped with the means to provide either of these two bunch separations.

Straight sections at points 3 and 7 contain two collimation systems for capturing stray particles. Point 3 is designed to capture off-momentum particles (momentum collimation) and point 7 for removing the beam halo (betatron collimation). Point 4 contains the two radio frequency systems, one independent system for each beam operating at 400 MHz, twice the

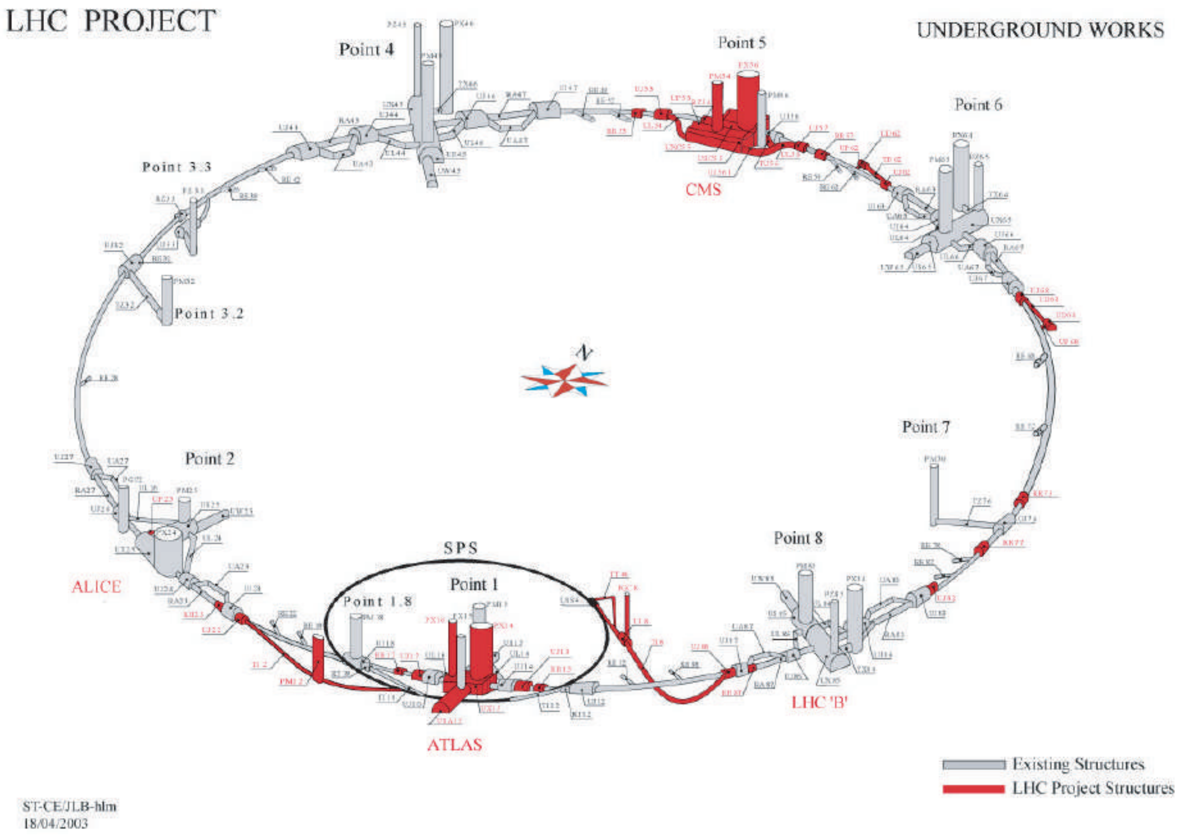
LHC PROJECT
UNDERGROUND WORKS
Point 4
Point 5
Point 6
Point 3.3
Point 3.2
CMS
Point 2
Point 7
ALICE
SPS
Point 1
Point 1.8
Point 8
LHC 'B'
ATLAS
Existing Structures
LHC Project Structures
ST-CE/JLB-hlm
18/04/2003

Figure 2. LHC layout.

frequency of the LHC injector. Finally, point 6 contains the two beam abort systems which will allow the beams to be extracted safely and dumped onto external absorbers.

The regular LHC lattice was designed to maximize the amount of bending power in the arc by making the dipoles as long as reasonably possible. This minimizes the amount of dead space between interconnects as well as the number of dipoles to be manufactured, tested and interconnected. After careful optimization, the dipole length was chosen to be 14.2 m (magnetic), 15 m overall with 23 regular lattice periods per arc. Each period is 106.9 m long and is made up of six dipoles and two short straight sections (SSSs) each of 6.6 m length containing the main quadrupoles and lattice correctors. The two apertures of rings 1 and 2 are separated by 194 mm. Both dipole apertures are connected in series whereas the quadrupoles are powered in two families, all focusing quadrupoles of rings one and two in series and likewise for the defocusing quadrupoles.

The transition from the regular arc contains a dispersion suppressor consisting of two perturbed lattice periods. The purpose of the dispersion suppressor is threefold:

1. adapt the LHC reference orbit to the geometry of the tunnel,
2. cancel the horizontal dispersion generated in the arc and by the separation and recombination dipoles,
3. help with the matching of the beams between the arcs and straight sections.

A generic design of a dispersion suppressor can be made using standard arc cells with missing dipoles and dipoles of different length to those in the arc. However, due to the constraints imposed by the geometry of the existing tunnel and the economy of having only one standard dipole length, the dispersion can only be fully cancelled by individual powering of the quadrupoles in the dispersion suppressor cells. These quadrupoles also provide additional parameters for matching the insertion optics.

The SSSs in the arcs contain the main quadrupoles and also the correction sextupoles for chromaticity control and the orbit correction dipoles. Depending on their location, they can also contain trim normal or skew quadrupoles or Landau damping octupoles.

The optics of the long straight sections differ according to their functionality. At points 1 and 5 in the high luminosity insertions the small β (0.5 m) at the collision point is generated with the help of a quadrupole triplet assembly. At points 2 and 8, the optics at the 450 GeV injection level must allow beam injection whereas at top field the beams must be focused to a moderate β. At the utility insertions the optics is tailored to their functionality. Details of the optics in the various insertions can be found elsewhere [1].

2.2. Main hardware systems

The main hardware systems include magnets, vacuum, cryogenics, radio frequency, power converters, injection and extraction equipment, beam instrumentation, collimation and controls, far too numerous for an article of this size. Details can be found in the LHC design report [1]. Only a few selected systems of general interest will be treated here.

2.3. Magnets

The LHC contains more than 7000 superconducting magnets ranging from the 15 m long main dipoles to the 10 cm long octupole/decapole correctors inside the dipole cold masses as well as more than 100 conventional warm magnets not counting the 500 or so conventional magnets in the two 2.6 km long transfer lines between the SPS and the LHC. The most challenging are the superconducting dipoles and the quadrupoles in the arcs, dispersion suppressor and matching regions.

The three large superconducting accelerators operating today, the Tevatron (FNAL), HERA (DESY) and RHIC (BNL) all use magnets made with classical Nb–Ti superconductor cooled with supercritical helium at a temperature slightly above 4.2 K. In each case, the nominal field is below or around 5 T. In order to increase the field to above 8 T, two possibilities exist today.

Figure 3 shows the critical current density of two superconductors available commercially today. The Nb–Ti conductor at 4.5 K has a critical current density that is too low to reach the LHC objective. The other commercially available conductor made from Nb_3Sn gives a shift of about 3 T upwards at 4.5 K and could be used for such high field magnets. However, Nb_3Sn superconductor is a very brittle material and cannot be used to wind magnet coils. Instead, the coils have to be wound before the heat treatment needed to produce the superconducting state, which requires them to be heated to over 600° for many tens of hours. Whilst this may be reasonable for a small number of specialized magnets, it is clearly not feasible for a large series.

The other way to high fields is to cool conventional superconductor to lower temperature. It can be seen from figure 3 that Nb–Ti cooled to 1.8 K gives the same 3 T shift as Nb_3Sn. This is a much more economical way to getting the high fields needed. However, it does not come without difficulties. First of all, there is a phase transition in helium at 2.17 K when the liquid

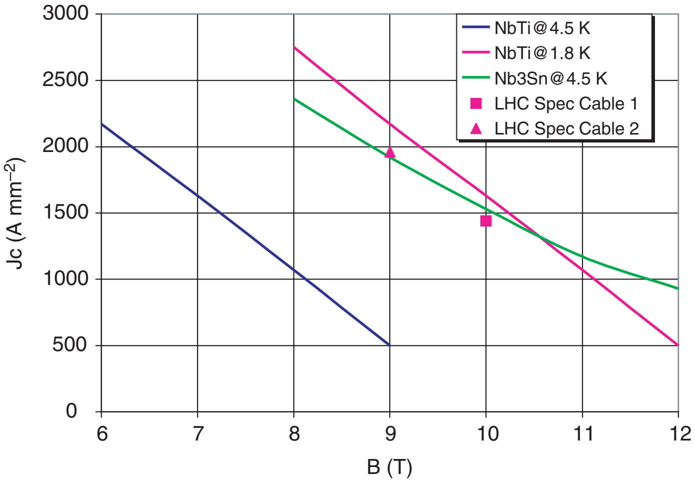

Figure 3. Critical current density of superconductors.

becomes superfluid. The advantages and disadvantages of this are discussed below. In addition, the heat capacity of the superconductor decreases by an order of magnitude between 4.5 and 1.8 K, so for a given heat inleak into the cable, for example due to small movement under the very high electromagnetic force or heating due to beam loss, the temperature rise is much higher, making the magnets much more sensitive to quenches. However, the superfluid helium can help if the coils are properly engineered.

A twin aperture dipole consists of two dipoles in a common iron yoke (figure 4). The two coils are clamped with austenitic steel collars with very low permeability surrounded by a yoke of low carbon steel which carries the magnetic flux. Figure 5 shows the flux plot as computed for the whole structure of the main dipole.

The requirement that the field must be in the opposite directions in the two apertures (for a proton–proton collider) ensures that there is no saturation of the central part of the yoke. The stored energy of 500 kJ m^{-1} in the magnet at nominal field requires active quench protection. If a quench is detected, the whole coil is made resistive by firing a capacitor bank into resistive strips (quench heaters) built into the coil. The current is then diverted through a diode until the power supply can be switched off.

The cross-section of a coil is shown in figure 6. The coil is wound in two layers in six blocks separated with copper wedges. The geometry of the conductor distribution has been carefully optimized to achieve as pure a dipole field as possible. The optimum geometry, which minimizes the higher harmonics of the field distribution, has been computed using a genetic algorithm [2]. Once the coil geometry has been fixed, then it is very important to keep it constant during the whole series production.

Equipment has been provided to the magnet manufacturers to allow the measurement of the field harmonics in the collared coil before assembly into the yoke. In this way, a tight quality control can be kept throughout the manufacturing run. Figure 7 shows the sextupole component of the field in about 1000 produced magnets. At the beginning of production it was necessary to make two small iterations on the geometry of the copper wedges in order to bring the harmonics inside the required control limits.

Apart from the field quality, it is also important to keep the integrated transfer function (effectively the bending angle for a given current) inside tight limits and identical for all three

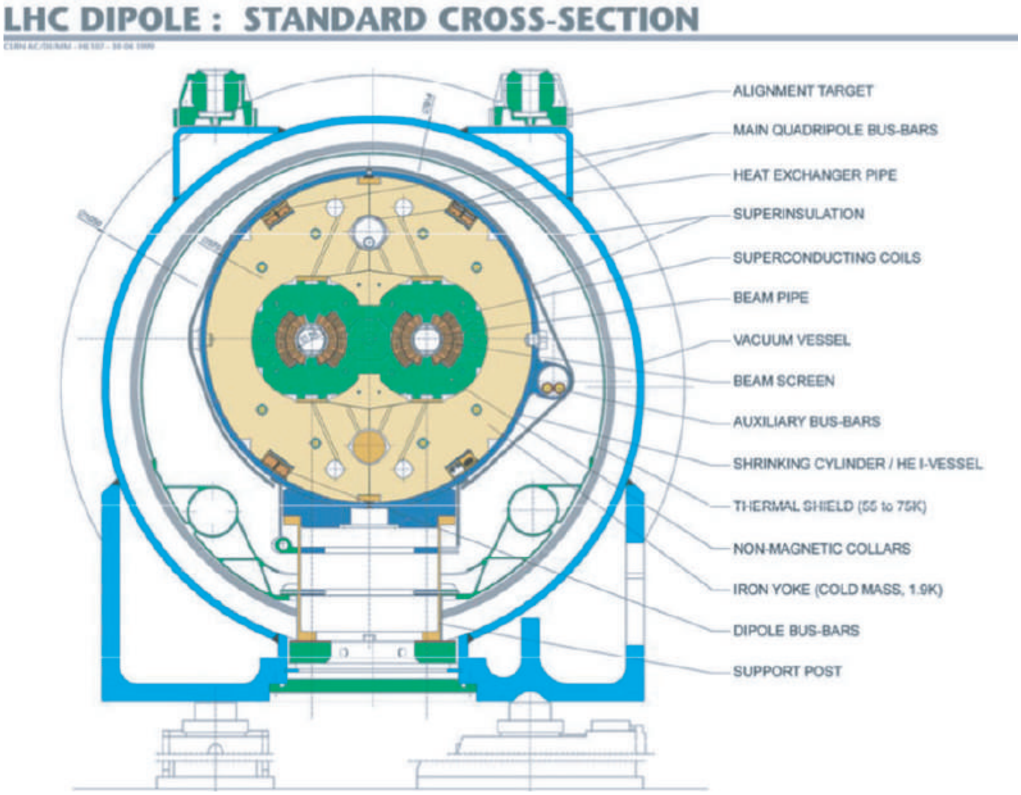
LHC DIPOLE : STANDARD CROSS-SECTION
ALIGNMENT TARGET
MAIN QUADRIPOLE BUS-BARS
HEAT EXCHANGER PIPE
SUPERINSULATION
SUPERCONDUCTING COILS
BEAM PIPE
VACUUM VESSEL
BEAM SCREEN
AUXILIARY BUS-BARS
SHRINKING CYLINDER / HE I-VESSEL
THERMAL SHIELD (55 to 75K)
NON-MAGNETIC COLLARS
IRON YOKE (COLD MASS, 1.9K)
DIPOLE BUS-BARS
SUPPORT POST

Figure 4. Dipole cross-section.

manufacturers. Too large dispersion in bending angle would result in very large orbit distortion which would be difficult to correct and variation in the mean for the three manufacturers would require that they are sorted into different octants depending on the manufacturer. Figure 8 shows what has been achieved for 1000 dipoles produced. The dispersion is much smaller than specified and the variation in the mean is so small that magnets from the three manufacturers can freely be mixed, simplifying enormously the logistics.

The main arc quadrupoles, 3.25 m long, are made with the same superconducting cable as the outer layer of the dipoles. Since the electromagnetic forces are much less and the geometry is more suited, the coils have separate austenitic steel collars instead of the combined collar structure of the dipoles. Each quadrupole is integrated into a SSS, each containing a sextupole for chromaticity correction and a closed orbit correction dipole. Depending on its position in the arc, a SSS can also contain a trim quadrupole or a Landau octupole.

In addition to the main arc magnets, the LHC contains many more elements for correction of dipole imperfections, matching of the optics and in the final focus. Table 2 gives a full list of all superconducting magnets, their number and function.

2.4. Cryogenics

The LHC magnets are cooled with pressurized superfluid helium, which has some interesting properties that make it a unique engineering material. Best known is its very low bulk viscosity which allows it to permeate the smallest cracks. This is used to advantage in the magnet

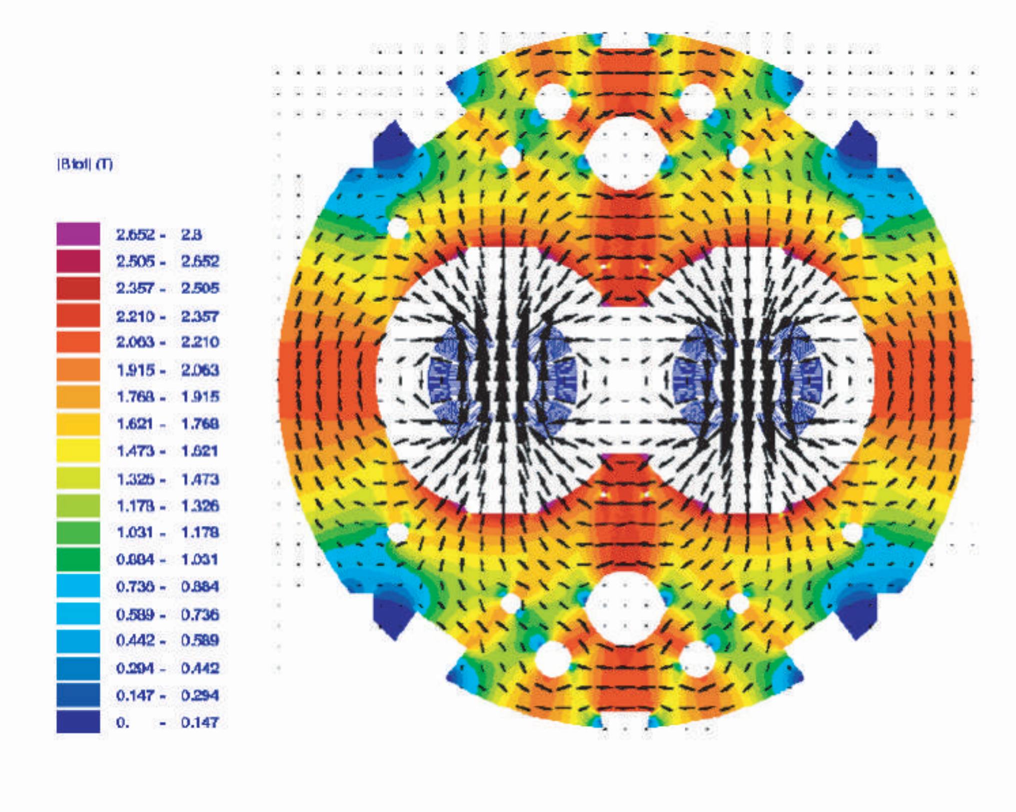

Figure 5. Flux plot in dipole.

design by making the coil insulation porous and enabling the fluid to be in contact with the strands of the superconductor. It also has a very large specific heat, 100 000 times that of the superconductor per unit mass and 2000 times per unit volume. Finally the thermal conductivity peaks at 1.9 K (figure 9) and is approximately 1000 times higher than that of cryogenic grade OFHC copper. The helium in the coil can therefore make up for the very low specific heat of the conductor at this temperature by helping to absorb unwanted thermal loads and transporting them efficiently outside the coil.

The most usual method of making superfluid helium in the laboratory is by pumping on the helium bath. Figure 10 shows the phase diagram of helium in the region of the point of the phase transition (the lambda point). Pumping on the bath reduces the pressure and at around 50 mbar and at a temperature of 2.17 K it crosses the lambda point. Further reduction of the pressure to 15 mbar lowers the temperature to 1.9 K.

The magnets themselves are cooled in a bath of superfluid at atmospheric pressure. This is achieved through a linear heat exchanger traversing each 107 m period of magnets (figure 11) containing superfluid at 15 mbar pressure, cooling the helium in the magnets at 1 bar to the same temperature as the saturated superfluid inside the heat exchanger. One big advantage of this system is that there is no problem with the bad dielectric strength of gaseous helium since the coil is permanently in liquid.

The machine is cooled using 8 cryogenic plants, each of 18 kW capacity at 4.5 K, located in pairs at the even points except for one singularity at point 2. Figure 12 shows the cryogenic

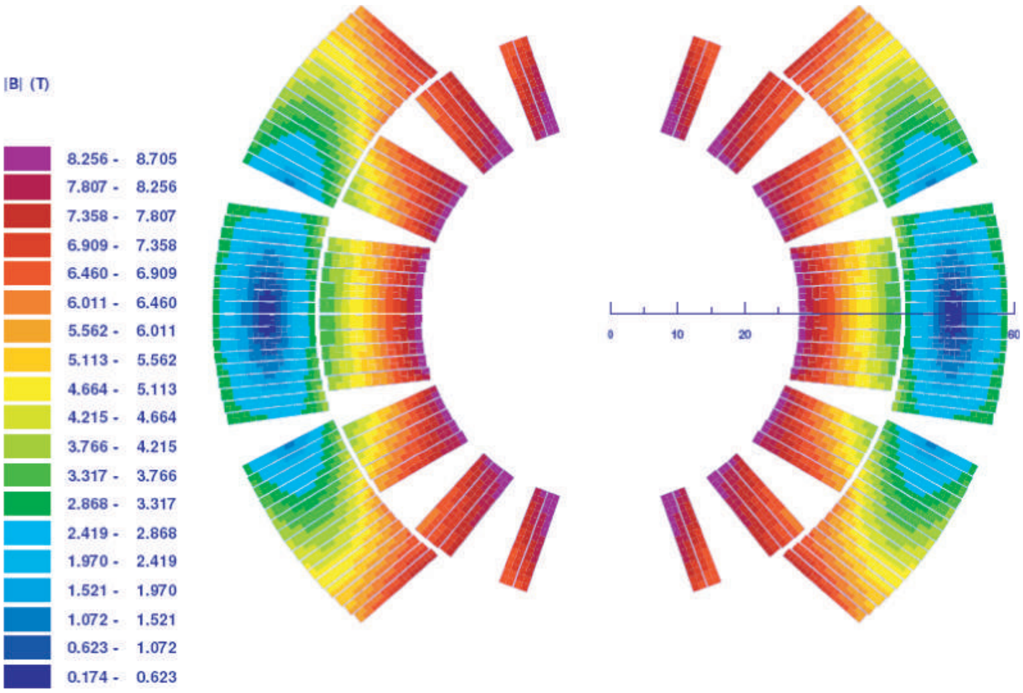
|B| (T)
8.256 - 8.705
7.807 - 8.256
7.358 - 7.807
6.909 - 7.358
6.460 - 6.909
6.011 - 6.460
5.562 - 6.011
5.113 - 5.562
4.664 - 5.113
4.215 - 4.664
3.766 - 4.215
3.317 - 3.766
2.868 - 3.317
2.419 - 2.868
1.970 - 2.419
1.521 - 1.970
1.072 - 1.521
0.623 - 1.072
0.174 - 0.623
0
10
20
60

Figure 6. Distribution of conductors in dipole coil. The coil blocks are separated by copper wedges.

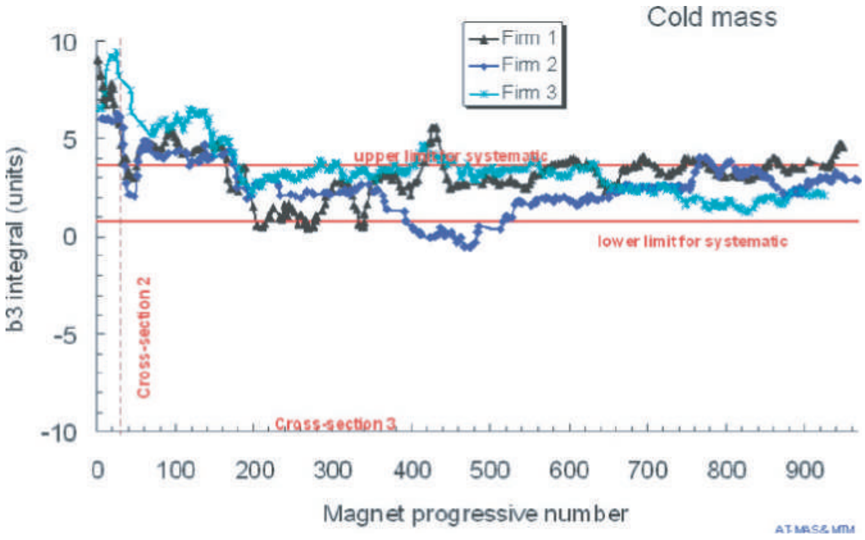

Figure 7. Sextupole component (b_3) of magnetic field in 1000 of the 1232 dipoles.

architecture at an even point. Four of the refrigerators are recuperated and upgraded from LEP, the other four are new.

The LEP refrigerators have their cold box split between surface and tunnel level whereas the new plants are located on the surface. At the tunnel level, the conventional refrigerators are supplemented by cold compressors. These multi-stage axial centrifugal compressors (figure 13)

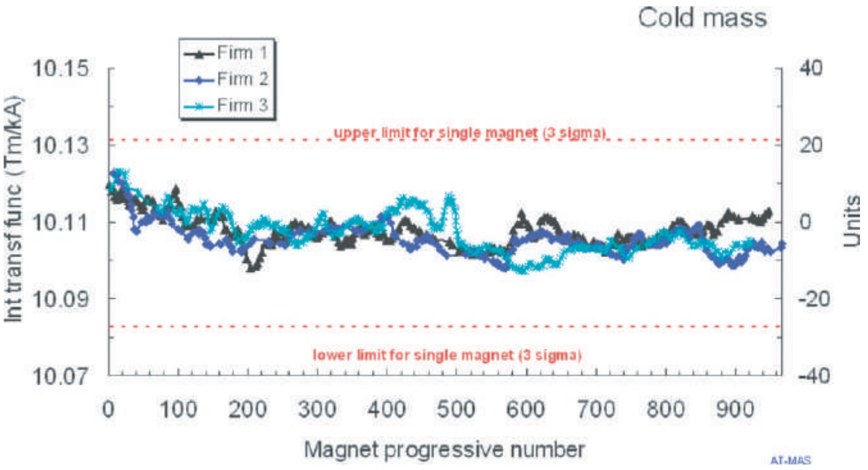

Figure 8. Integrated bending strength measured over 1000 dipoles from three manufacturers.

Table 2. List of superconducting magnets and their function.

Type	Number	Function
MB	1232	Main dipoles
MQ	392	Arc quadrupoles
MBX/MBR	16	Separation and recombination dipoles
MSCB	376	Combined chromaticity and closed orbit correctors
MCS	2464	Sextupole correctors for persistent currents at injection
MCDO	1232	Octupole/decapole correctors for persistent currents at injection
MO	336	Landau damping octupoles
MQT/MQTL	248	Tuning quadrupoles
MCB	190	Orbit correction dipoles
MQM	86	Dispersion suppressor and matching section quadrupoles
MQY	24	Enlarged-aperture quadrupoles in insertions
MQX	32	Low-β insertion quadrupoles

pump the cold helium gas, producing the 15 mbar pressure in the linear heat exchangers inside the magnets in order to produce the primary superfluid. The connection to the magnets is made through a cryogenic distribution line running in the tunnel parallel to the machine. An interconnect box allows the plants to be used in the arc either side of the even point or if needed, for the full power of both plants to be used in one arc.

Using a single cryogenic plant boosted with a liquid nitrogen precooler, a sector can be cooled down in less than 15 days. By coupling two plants to the same octant, this time can be reduced by a factor of two. Warming of a sector takes about the same time.

2.5. *The radiofrequency acceleration system*

The RF system is located at point 4. Two independent sets of cavities operating at 400 MHz (twice the frequency of the SPS injector) allow independent control of the two beams. The

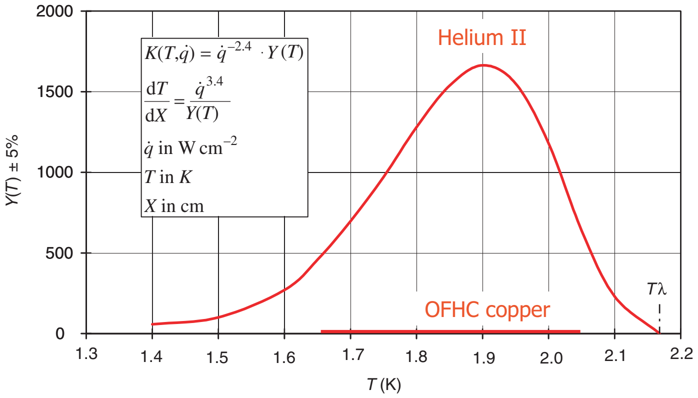

Figure 9. Thermal conductivity of He II.

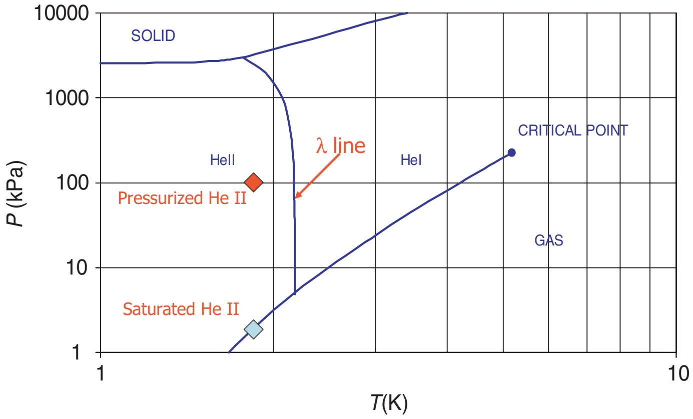

Figure 10. Phase diagram of helium.

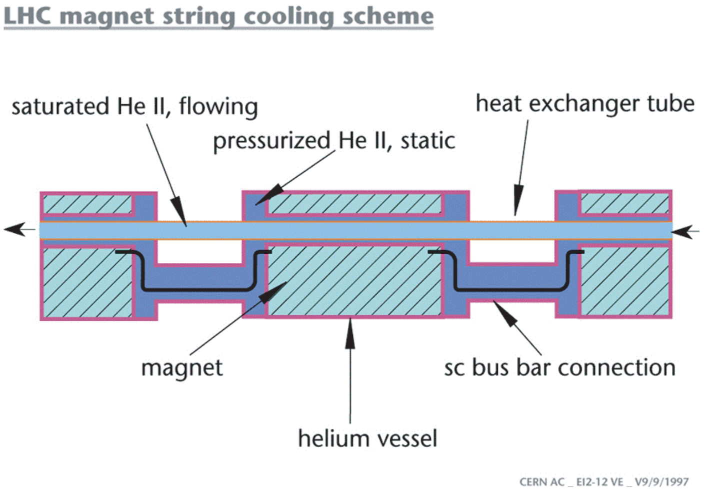
LHC magnet string cooling scheme
saturated He II, flowing
heat exchanger tube
pressurized He II, static
magnet
sc bus bar connection
helium vessel
CERN AC _ EI2-12 VE _ V9/9/1997

Figure 11. Linear heat exchanger.

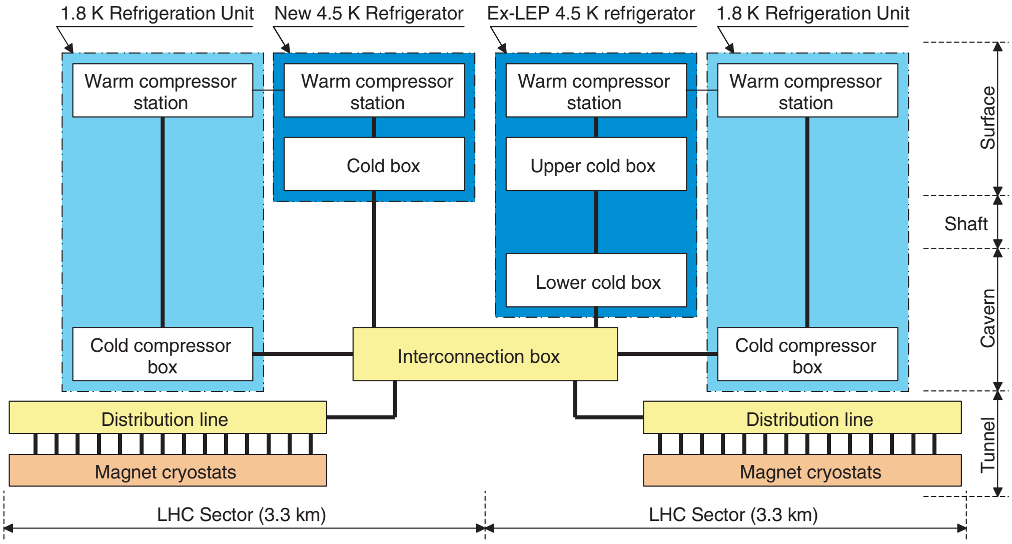
1.8 K Refrigeration Unit
New 4.5 K Refrigerator
Ex-LEP 4.5 K refrigerator
1.8 K Refrigeration Unit
Warm compressor station
Warm compressor station
Cold box
Warm compressor station
Upper cold box
Lower cold box
Warm compressor station
Cold compressor box
Interconnection box
Cold compressor box
Distribution line
Distribution line
Magnet cryostats
Magnet cryostats
LHC Sector (3.3 km)
LHC Sector (3.3 km)
Surface
Shaft
Cavern
Tunnel

Figure 12. Cryoplant architecture at an even point.

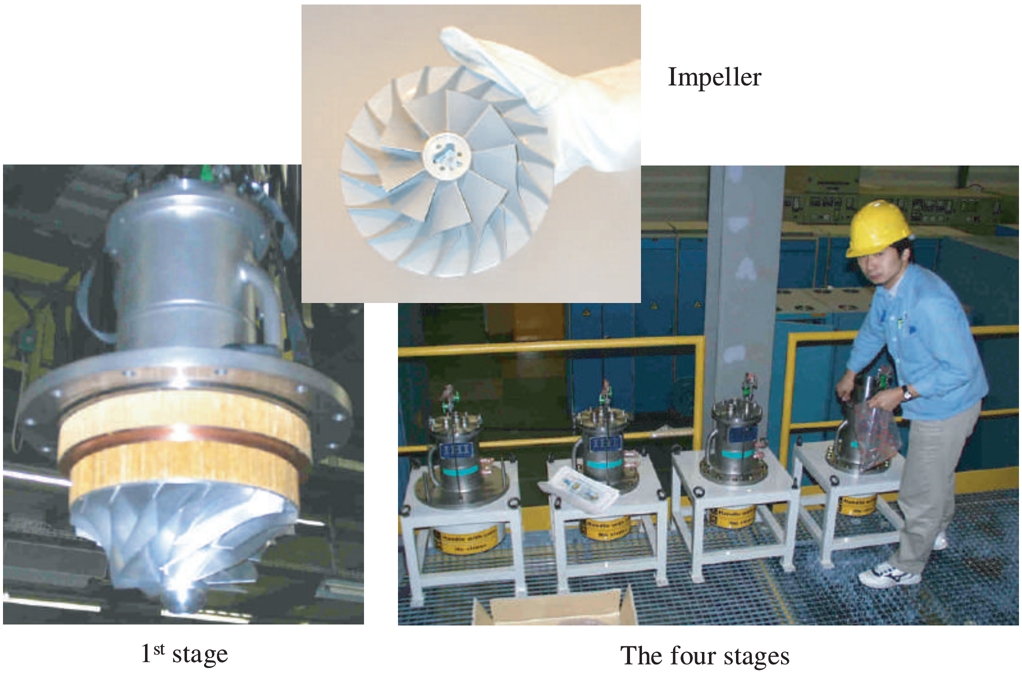
Impeller
1st stage
The four stages

Figure 13. Cold compressor impeller stages.

superconducting cavities are made from copper on which a thin film of a few microns of niobium is sputtered on to the internal surface. In order to allow for the lateral space, the beam separation must be increased from 194 mm in the arcs to 420 mm. In order to combat intrabeam scattering (see below), each RF system must provide 16 MV during coast while at injection 8 MV is needed. For each beam there are 8 single cell cavities, each providing 2 MV, with a conservative gradient of 5.5 MV m^{-1}. The cavities are grouped into two modules per beam, each containing four cells (figure 14). Each cavity is driven by an independent RF system, with

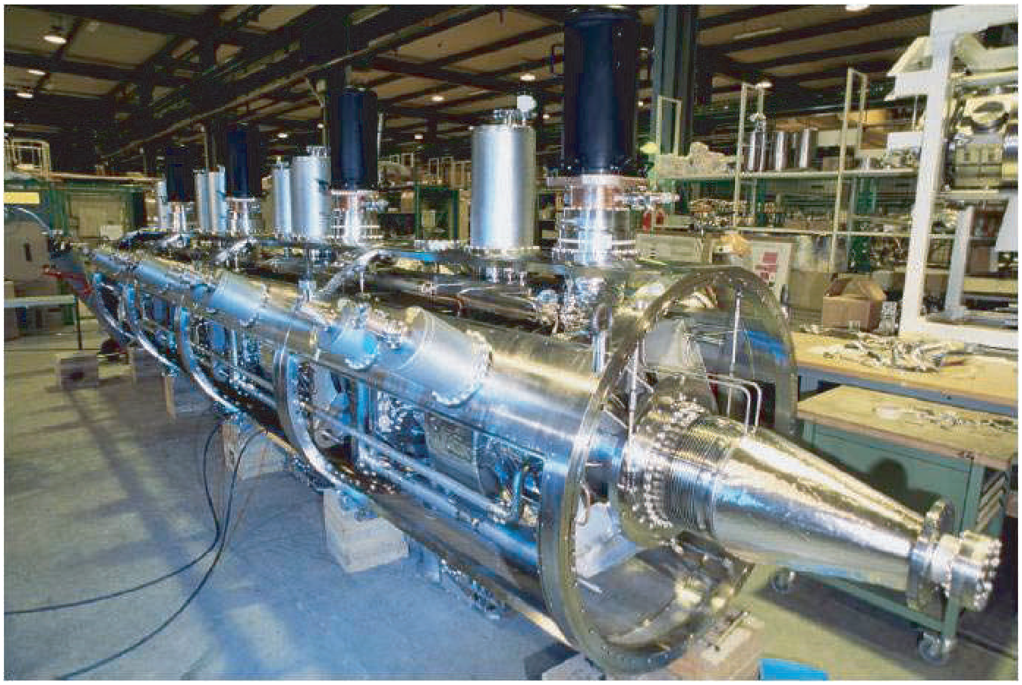

Figure 14. Four-cavity module during assembly.

independent klystron, circulator and load. Although the RF hardware required is much smaller than LEP due to the very small synchrotron radiation power loss, the real challenges are in controlling beam loading and RF noise.

2.6. The vacuum system

The LHC presents several original requirements compared with classical vacuum systems. It has to ensure adequate beam lifetime in a cryogenic system where heat input to the 1.9 K helium circuit must be minimized and where significant quantities of gas can be condensed on the vacuum chamber. The main heat sources are:

1. synchrotron light radiated by the beam at high energy (0.2 W m^{-1}per beam), with a critical energy of about 44 eV;
2. image currents (0.2 W m^{-1} per beam);
3. energy dissipated by the development of electron clouds (see below);
4. energy loss by nuclear scattering (30 mW m^{-1}per beam).

In order to remove the heat from all these processes but the last with high thermodynamic efficiency, the 1.9 K cold bore of the magnets is shielded with a beam screen cooled to between 5 and 20 K (figure 15). This beam screen is perforated with about 4% of the surface area to allow the cold bore of the magnets at 1.9 K to act as a distributed cryopump, allowing gas to be condensed on the cold bore surface protected against desorption by bombardment with synchrotron radiation photons.

For roughly 3 km of the 27 km circumference, mainly in the long straight sections, the vacuum chambers are at room temperature, requiring a low residual pressure without the benefit

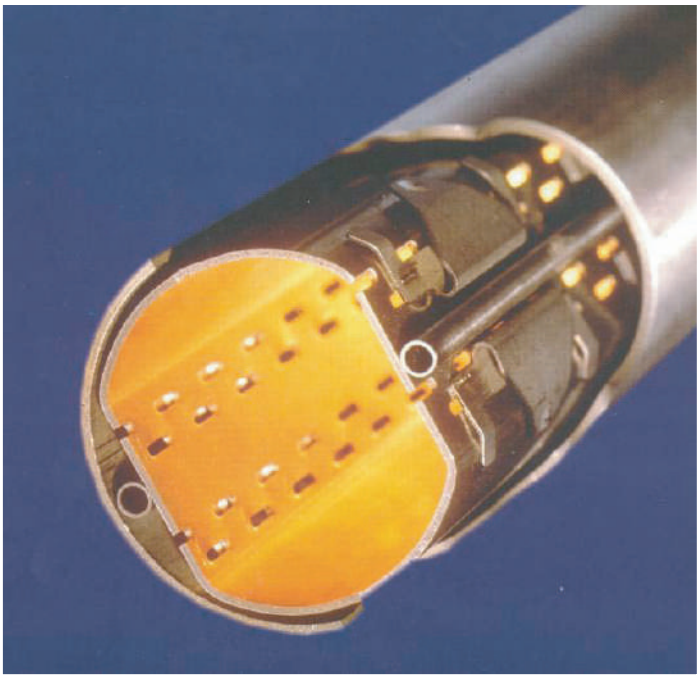

Figure 15. LHC beam screen.

of the distributed cryopumping. As a spin-off of the development of sputtering technology for superconducting cavities a new getter material (TiZrV) has been developed which can be sputtered on the internal surface of the copper vacuum chambers and can be activated at the very low temperature of 200° (conventional getters require activation at 600°). When activated, the chamber wall itself becomes a distributed pump, producing very low residual pressure and at the same time a very low secondary emission yield (SEY), preventing the buildup of an electron cloud. All warm chambers, including those inside the detectors, are treated in this way.

3. Accelerator physics issues

More than 30 years of accumulated experience from the first hadron collider, the intersecting storage rings as well as the proton–antiproton colliders at CERN and FNAL and the superconducting storage rings at BNL and DESY has gone into the LHC design. In the following chapters, the different effects that could limit machine performance and the remedies adopted are described.

3.1. Dynamic aperture

In superconducting magnets of the type used in the LHC, the field quality is determined by the precision of the positioning of the superconductor and not by the geometry of the iron yoke, so it can never be as good as in conventional magnets. It has been shown by experience in the different superconducting machines and by particle tracking that the aperture inside which particle orbits are stable is much smaller than the physical aperture of the beam pipe. This is called the dynamic aperture and is limited by a complex interplay between the unwanted higher field harmonics due to magnet imperfections. Sophisticated computer codes have been developed to track particle orbits around virtual machines with distributed random and systematic imperfections [3]. And these results are used to define maximum systematic and random deviations of each field multipole. It was from this work that the control limits for the sextupole component of the dipole field shown in figure 7 were derived. Even with present day computers it is not possible to perform full scale simulation over a large number of virtual machines over 4×10^7 turns, which corresponds to 1 h of storage time. The dynamic aperture obtained from tracking of existing machines is always too optimistic when compared to experiment by 20% or more. For the LHC, in order to insure a dynamic aperture of 6 sigmas it has been decided that the tracked dynamic aperture over 10^6 turns should be a factor of 2 larger. These results have been used to supply the tables of allowed multipole errors to the magnet builders.

Since the dynamic aperture depends strongly on the horizontal and vertical tunes, the tracking studies are also used to find the best working points.

3.2. *The beam–beam interaction*

When the beams are brought into collision, a much stronger nonlinearity than the magnet imperfections comes into play. It is called the beam–beam interaction and is caused by the force due to the electromagnetic field of one beam on the particles in the other beam. It produces two main effects.

The first is to cause a variation of the tune with amplitude. This means that the beam does not occupy a point on the Qh, Qv tune diagram but produces an extended ‘footprint’. The second effect is that because of the periodic nature of the force (particles experience a delta function kick on each revolution) it excites nonlinear resonances which can strongly limit the beam lifetime. Figure 16 is an example of a tune scan made many years ago on the SPS proton–antiproton collider [4]. The lower figure shows the tune diagram with nests of 10th, 7th and 11th order resonances. The beams are scanned across these resonances in steps, the lines corresponding to each scan position being intended to indicate approximately the size of the footprint. The upper diagram shows a chart recorder output of the intensity decay of the beams for each position in tune space. In this experiment, the antiproton intensity was very low so that the effect of the beam–beam interaction of the antiprotons on the protons is very small. Consequently, as observed, the decay rate of the protons is completely insensitive to the tune. On the other hand, the strong proton beam affects the antiprotons much more, producing a clear effect as the beams are scanned across the resonances. Experiments like this indicate that the total tune spread due to the sum of the tune spreads from each IP should not exceed 0.015. With three proton detectors requiring almost head-on collisions this implies that the tune shift per experiment should not exceed 0.005, a value routinely achieved in previous and existing colliders.

The long-range beam–beam interaction between successive bunches must also be avoided by colliding the beams with a small crossing angle of about 400 μrad.

3.3. *Intrabeam scattering*

As particles perform their betatron and synchrotron oscillations, they exchange energy due to multiple Coulomb scattering. The correct frame of reference to understand the phenomenon

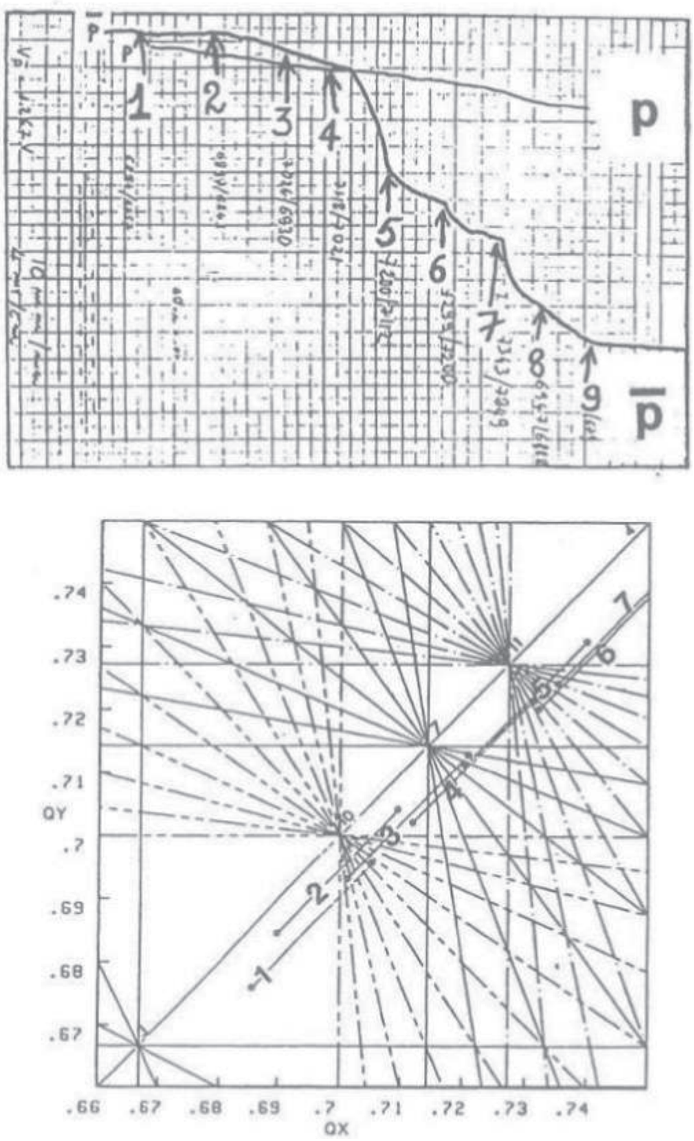

Figure 16. A beam–beam resonance scan at the SPS collider.

is the rest frame of the beam [5]. The transverse rms momenta $\sigma'_{x,y}$ are unchanged by this transformation whereas the longitudinal momentum σ_p is transformed into σ_p/γ. In a highly relativistic beam like the LHC, the longitudinal plane is therefore very 'cold' compared with the transverse planes and one would expect a damping of the transverse dimensions and an increase in the energy spread, which would be good for luminosity preservation. This indeed does occur in the vertical plane although the damping time is very long. Unfortunately, in the regions where the dispersion is not zero (most of the machine), a particle changes its energy by Coulomb scattering but does not change its position and therefore finds itself on the wrong orbit for its momentum. It can only make a betatron oscillation around its new equilibrium orbit, adding a heating term that completely swamps the slow damping in the radial plane.

Intrabeam scattering has been clearly observed in the SPS proton–antiproton collider (figure 17) [6] and the growth rate is in quite good agreement with the theory. In the LHC,

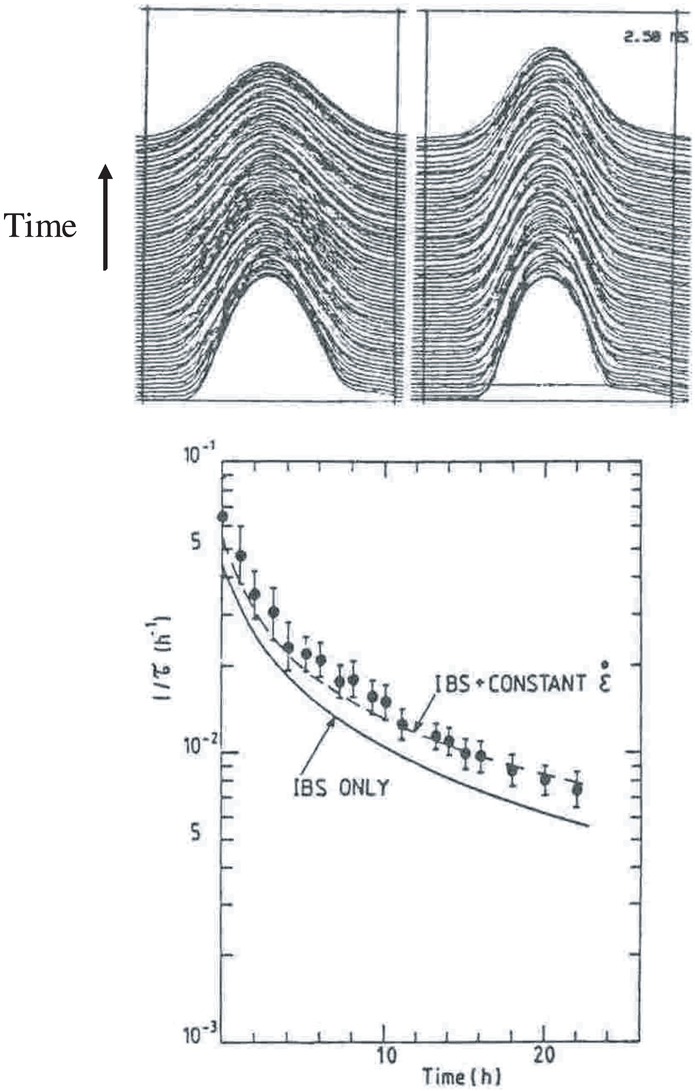

Figure 17. Intrabeam scattering in the SPS. Top bunch lengthening with time for a strong proton bunch (left) and a weak antiproton bunch (right) bottom. IBS growth rate compared with theory.

if no action is taken, the beam would blow up so fast that the luminosity lifetime would be reduced to a few hours.

The growth rate depends strongly on the six-dimensional phase space density. The longitudinal emittance of the beam arriving from the SPS is about 1 eV s. During the acceleration, the emittance is increased to 2.5 eV s using RF noise to blow up the energy spread and bunch length. This reduction in phase space density is sufficient to increase the transverse emittance growth time to about 80 h, making its effect on luminosity lifetime negligible compared to other processes. A consequence of this is that the RF voltage during coast must be increased to 16 MV to provide a big enough bucket for the large emittance beam.

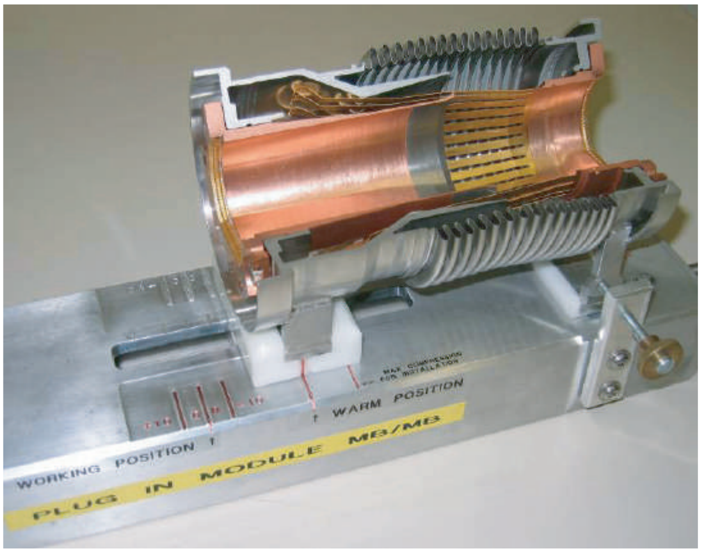

Figure 18. A vacuum interconnect.

3.4. *Coherent instabilities*

The interaction of the beam with its environment generates electromagnetic fields which can react back on it and drive it unstable. The first remedial action is to design the vacuum chamber to reduce this coupling as much as possible. The chamber should be as smooth as possible without discontinuities. As an example, figure 18 shows an interconnect module between two dipoles. When the magnets are cold, the bellows is stretched and the RF ‘fingers’ ensure a smooth transition from one beam screen to the other. The material of the beam screen is stainless steel to give it structural strength in case of quench co-laminated with a 75 μm layer of copper on its inner surface. The resistivity of the copper layer is reduced by cooling it to between 5 and 20 K. In the room temperature regions the vacuum chamber is made of 2 mm thick high conductivity copper.

Reducing the impedance of the environment can reduce the growth rate of the instabilities but cannot eliminate them all together. In the LHC, the two instabilities that must be controlled are the transverse coupled bunch instability (resistive wall) and the single bunch head-tail instability.

The resistive wall instability is driven mainly by the long-range wake fields due to image currents in the beam screen. The instability can occur at the frequencies of the so-called slow waves:

$$f_n = (n - Q)\, f_\mathrm{r},$$

where f_r is the revolution frequency (11.245 kHz), Q is the tune and n is the mode number. Only modes with $n > Q$ are unstable. The other modes drive backward waves which are stable.

The lowest and fastest growing mode is around 8 kHz ($Q = 59.3$) with a rise time of around 300 turns (26 ms). It must be damped with active feedback. A signal from a transverse pickup is delayed by one turn, amplified and fed back into a pair of electrostatic deflectors placed at a betatron phase of 90° with respect to the pickup. The bandwidth of the system is around 20 MHz, allowing many modes to be damped simultaneously. Very high frequency modes have such slow

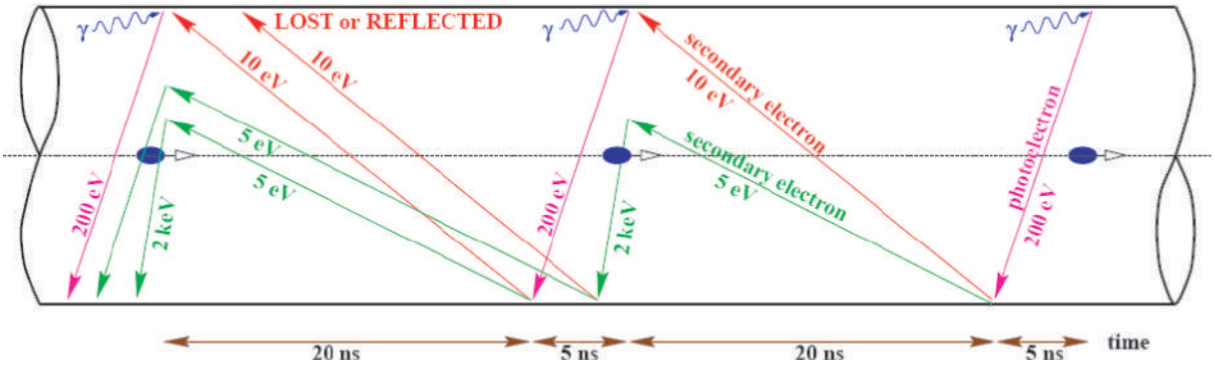

Figure 19. The electron cloud effect.

growth rates that they will be damped by Landau damping (see below). The transverse feedback system is also useful for damping injection errors to avoid filamentation and emittance growth.

The head–tail instability is a well-known single bunch effect driven by short-range wakefields where the tail of the bunch is driven by the head and half a synchrotron period later the roles are reversed, driving the bunch unstable. For the lowest head-tail mode this can only happen if the chromaticity is negative. The chromatic aberrations must anyway be corrected by the sextupoles integrated into the short straight sections. To control the head–tail instability it is sufficient to keep the chromaticity slightly positive at all times. This will be particularly critical at the start of acceleration where persistent currents in the superconducting magnets can produce a large and rapid change in the chromaticity.

The ultimate panacea for beam instabilities is Landau damping where the tune spread in the beam is large enough to stop it from oscillating coherently. To provide Landau damping, two families of strong octupoles are integrated into the lattice in selected short straight sections. Octupoles give a variation of tune with betatron amplitude and will allow instability control at high energy without active feedback. This will be particularly important if the transverse feedback system has noise problems. During collisions, the tune spread due to the beam–beam footprint should be enough to keep the beams stable.

3.5. Electron cloud effects

A significant number of electrons can accumulate in the LHC vacuum chamber through ionization of residual gas molecules or by the impact of synchrotron radiation on the beam screen. When a proton bunch passes, these electrons will be given an impulse and can hit the beam screen with energies of several hundred electron volts. The primary electrons produce secondaries (figure 19) and if the transit time of the electrons across the chamber is resonant with the 25 ns bunch separation, the electron cloud can grow exponentially. This process is called beam-induced multipacting and is known to limit the performance of storage rings with small bunch separation, notably the two B factories, PEP-II and KEKB.

The threshold for the buildup of the electron cloud depends on the bunch current, the geometry of the beam screen and the SEY of the surface of the beam screen. It has been shown that it can occur in the LHC under some conditions with the nominal bunch separation and has already been observed in the SPS with LHC-like beams. The main effect is an additional heat load in the cryogenic system and can also lead to instabilities.

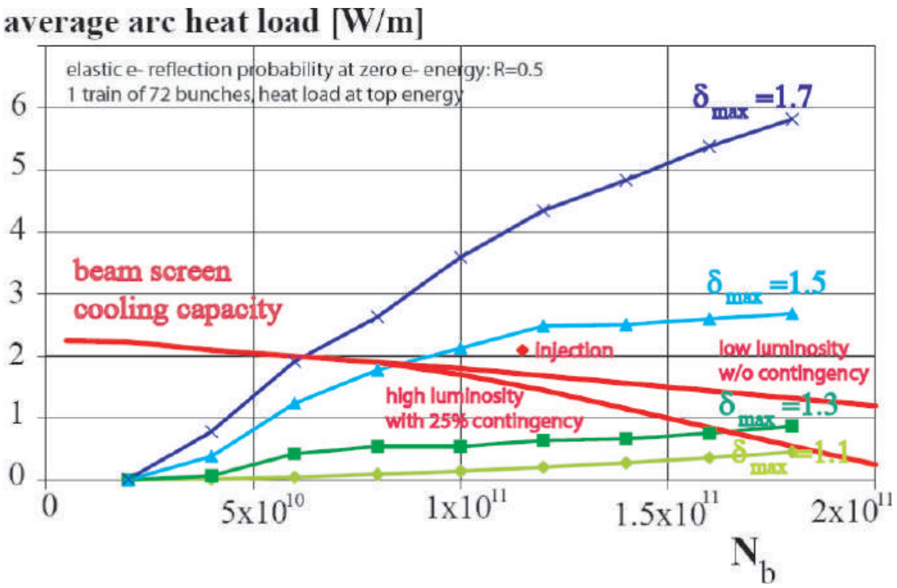

Figure 20. Heat load on the beam screen for different values of SEY.

A considerable amount of work has been done in order to understand the phenomenon, both experimentally and through the development of simulation codes which have been calibrated on experimental data from the SPS [7]. Figure 20 shows the predicted heat load on the beam screen as a function of bunch intensity for the nominal 25 ns bunch separation with different values of the SEY of the beam screen surface. The available cooling capacity of the cryogenic system is also shown. At zero intensity, all of the capacity is available for the electron cloud but as the intensity increases, part of the capacity is used for image currents and synchrotron radiation. The main conclusions are the following.

For the raw surface of the beam screen with a SEY of around 1.7, the electron cloud will limit the intensity to about half nominal. It has been shown experimentally that the presence of the cloud ‘scrubs’ the surface and the SEY reduces quickly to below 1.3, where this effect should no longer be a problem up to and beyond nominal intensity. Simulations also show that for larger bunch separations, above 50 ns, the threshold is well above nominal intensity. For this reason, the LHC injector chain has been equipped to provide beam a beam of 75 ns as well as the nominal one, so that early commissioning can be done without the need for a scrubbing run.

4. Conclusions

The LHC is now in its final stage of installation and commissioning. A beam has already been extracted from the SPS with the parameters needed to achieve design luminosity and has been transported along the 2.6 km tunnel TI8 to the LHC injection point. The stored energy in this beam at 7 TeV is two orders of magnitude higher than in any previous machine and one of the biggest challenges will be to ensure that the machine protection systems work with an extremely high degree of reliability in order to protect both machine and detectors.

The accelerator physics is well understood and the appropriate means to combat the various effects have been integrated into the machine design. The one new effect not seen in hadron machines before is the electron cloud. Simulations and experiments on the SPS have shown that a moderate scrubbing of the surface of the beam screen by electron bombardment will quickly reduce the SEY to a low enough value to allow design luminosity to be reached.

References

[1] 2004 *LHC Design Report* CERN-2004-03
[2] Ramberger S and Nussenschuck S 1998 *IEEE Trans. Magn.* **34** 2944
[3] Koutchouk J P 1999 *Proc. 1999 Particle Accelerator Conference, LHC Project Report 296*
[4] Evans L *AIP Conf. Proc.* **153** 1722
[5] Piwinski A 1974 *Proc. 9th Int. Conf. on High Energy Accelerators* (*Stanford*)
[6] Evans L 1984 *Proc. 12th Int. Conf. on High Energy Accelerators* (*Batavia*)
[7] Jimenez J M *et al* 2004 *AIP Conf. Proc.* **773** 211

Amateur Radio MØMTJ

MDS975.co.uk HOME PAGE

My Home Page

Amateur Radio Operating Conditions
Antennas
Accessories
Projects & Kits
Data Modes
Information
Portable
QSL
RSGB Ltd
Links
QRM
Bits 'n' Bobs
Contact

WSPR
Weak Signal Propagation Reporter

Site Map

MDS975 Home

Feedback

Site Search

Find

powered by FreeFind

G Whip Antenna Products
Get on the air with TOP QUALITY BRITISH

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits | Useful Information
Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites | RSGB | QSL | The Amateur Radio Mini Site Map
News, Developments, Events, Photographs and other 'Bits 'n' Bobs' | WSPR Weak Signal Propagation Reporter

AERIALS (ANTENNAS) 2

Antennas 1 | **Antennas 3** | **Antennas 4** | **Antennas 5** | **Antennas 6** | **Antennas 7**

Index To Other Antenna Pages:
Antennas 1 : Aerials used at M0MTJ
Antennas 2 : Including ideas for compact aerkals for Top Band /160 metres
Antennas 3 : Felix Scerri VK4FUQ discusses Loop Antennas, baluns, masts & other antenna related topics
Antennas 4 : Many antenna ideas from various sources particularly for multi-band operation & also gives information about
antenna trimming, knots for wire antennas and useful antenna rigging accessory ideas.
Antennas 5 : Half Wave End Fed aerials for 144 MHz VHF / 430 MHz UHF and 50 MHz 6 Metre band & J-Pole Aerials
Antennas 6 : Simple and effective H.F. Antenna ideas - Ground Plane and All Band Doublet
Antennas 7 : Omni-Directional - Circularly (Mixed) Polarized Aerial for VHF / 2 Meters.

Compact Top Band Antenna - An Inverted L for stealth with low(ish) visual impact

With a small back garden there is no way that I could accommodate an aerial for the 160m band that would be anything approaching full size. A full size dipole would be about 65 metres long and would need to be mounted at a very good height to be at its most effective. A full size vertical 1/4 wave would be about 37.5 meters tall. Impossible! Bending a 1/4 wave wire into an inverted L would still result in a very long wire - say 10 metres vertically and 27.5 metres horizontally. Still too large.

Full size top band antennas are big, far too big for my small plot, so I have tried a few different shortened 160m aerials. I really would prefer to use a balanced dipole not only for the radiation efficiency, but just as importantly for the lower noise on receive - like a ground mounted vertical aerial an inverted L can be rather noisy on RX. However I have to settle for a compromise, so shown in the drawings and photographs below is my current top band aerial, along with some previous experiments and ideas further down the page.

This incarnation of my Top Band aerial takes two forms. A compact Inverted L and, in its lower position, a less conspicuous sloping wire, shown below:

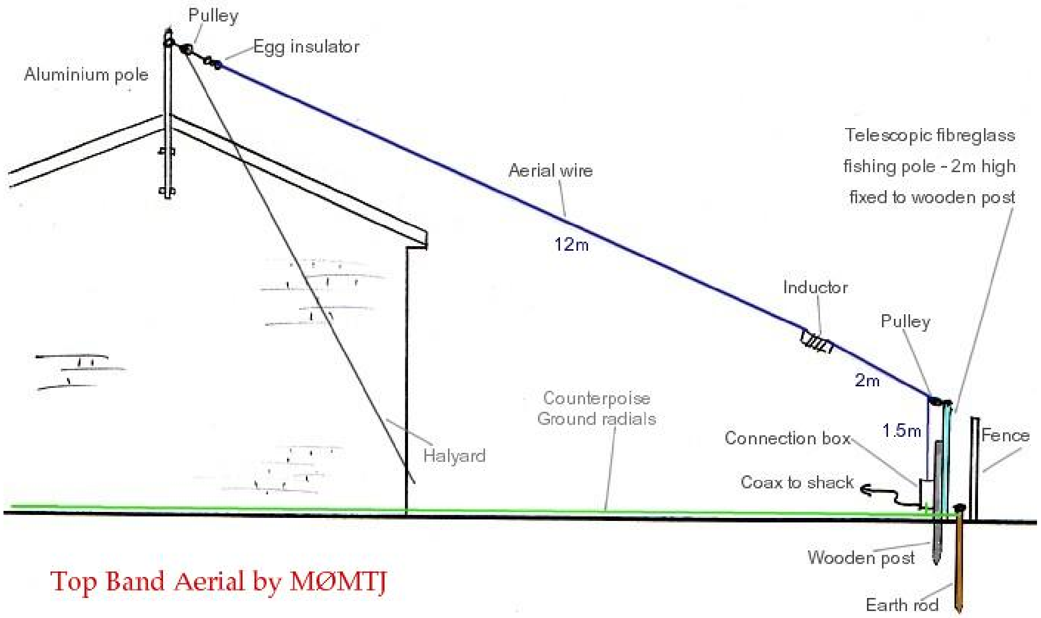

General layout of Top Band Aerial with fibreglass pole retracted to a height of 2 metres
Wire lengths are approximate: Inductor 5cm dia with approx 40 turns of 0.9mm e.c.w.

G-WHIP ANTENNAS

My first experiments were with a base loaded sloper which worked ok-ish for local ground wave, but could have been better. The first weak link that I wanted to change was moving the base loading coil further up the aerial wire so that it would be positioned on the more horizontal section.

It was also very low, so could surely benefit from some additional vertical height. However I could not realistically erect another permanent pole in the garden for fear of upsetting the XYL!

Therefore I decided that I needed a design that could be semi-permanent and offer two slightly different configurations: The idea is that, as a sloping wire, the aerial can be left in place at all times for RX and still be operable on TX. Since it is essentially just a wire sloping from fence height (2m) up to the roof apex it has fairly minimal visual impact. Alternatively, when required - at night - the telescopic pole can be extended to transform the aerial into the larger Inverted L.

The Earth and Radials: The aerial is fed against ground, so the ground needs to be as good as possible. At the base of the aerial a copper earth rod is driven into the ground. Additional ground radial wires are also added to provide a counterpoise and help reduce earth loss. The more the merrier.

The Support: A 10 metre long telescopic fibreglass fishing pole is used as part of the support. A wooden post was driven into the ground, adjacent to the earth rod. The connector box is screwed on to the post and the fishing pole was fixed to the post using nylon straps.

Since the top sections of a fishing pole are too thin to support the aerial wire, the last four sections were removed, leaving the pole 6 metres in length, which is a respectable height in a small back garden, though being even taller would be better.

A small lightweight pulley is fixed to the top of the fibreglass pole using two nylon ties - as seen in the photograph. As a sloping wire, the fishing pole is collapsed to 2m in height while being extended to 6 metres when being used as the 'full' Inverted L. The pole is quite flexible and so will tend to bend quite noticeably when 6 metres long with the wire attached to the top. This could be remedied by adding a back guy rope of required.

The far end of the aerial wire is tied to a egg type insulator and nylon halyard which is suspended through a pulley on a pole attached near the apex of the roof. The nylon halyard is tied off on a cleat hook near ground level to enable easy raising and lowering of the aerial wire.

The Wire: Aerial wire is PVC covered stranded wire of about 16/18 swg http://www.whwestlake.co.uk. The wires were cut for the maximum length that could be fitted into the available space - so as much wire was in the air as possible. This was about 3.5 metres from the feed point to inductor and around 12 to 13 metres from the inductor to the far end.

(If you make one of these get the vertical portion as high as possible and the horizontal section as long as possible and reduce the coil's inductance accordingly)**.**

The Inductor and Resonance:

The resonance of the aerial system was adjusted not by altering the length of the wires, but by changing the inductance of the loading coil.

I had some 0.9mm (20 s.w.g.) enamelled copper wire on a reel, so I used that to wind the coil. The coil former is 5cm in diameter and about 15 cm long, cut from an empty tube of silicone bath sealant (the 'gun' type). Two holes are drilled in the former for the fixing bolts and wing nuts that secure the aerial and coil wires.

I used an MFJ antenna analyzer to check resonance and adjust the number of turns. Initially the coil was wound with 60 turns, but the resonance was well below 1800 kHz. Turns were gradually removed until resonance was around 1940 kHz, leaving 39 turns on the coil. (Thicker e.c.wire might have been better, I would have preferred to use something like 18.swg 1.2mm diameter and this may possibly get changed in the future)

Enamelled copper wire:
http://www.esr.co.uk/electronics/cable-copper.htm

When the adjustments were finished the aerial wires and coil wires are terminated with lugs and fixed in place with the bolts and wing nuts. The winding was also covered in duck tape. All exposed connections should be suitably protected against the weather for a permanent installation.

When setting up any aerial an antenna analyzer is extremely useful - allowing all the measurements to be done in the garden or back yard thereby simplifying and speeding up the process considerably. These simple measurements can be made using the transceiver and VSWR meter, though that is obviously less convenient and will involve transmitting regular short burst of carrier, so it's very important to ensure that the power used is the absolute minimum and that the frequency being used is clear.

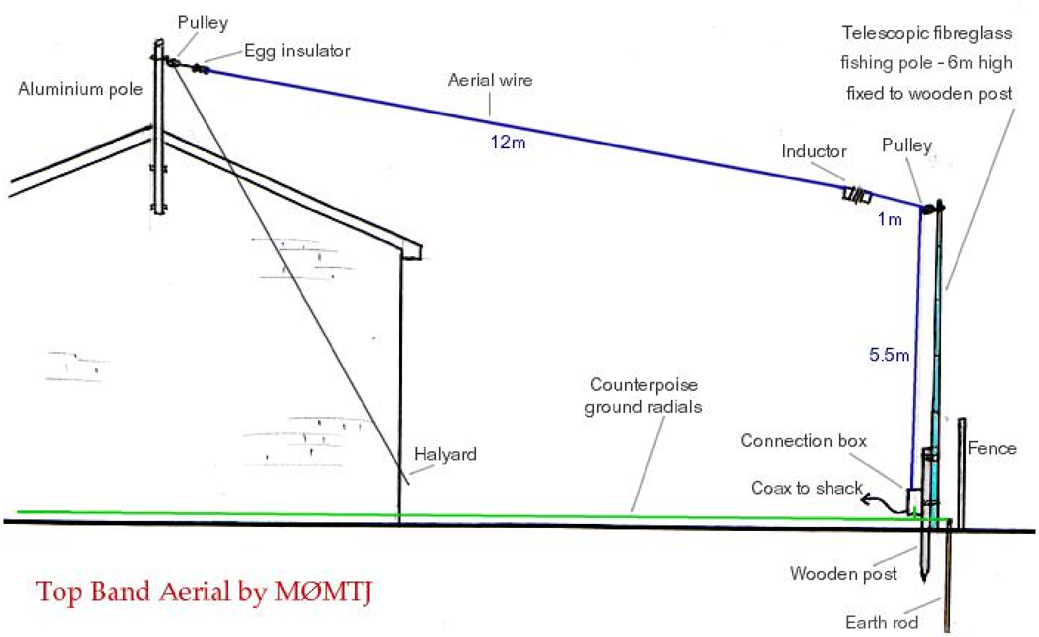

General layout of Top Band Aerial with a fibreglass pole extended to a height of 6 metres
Wire lengths are approximate: Inductor 5cm dia with approx 39 turns of 0.9mm e.c.w.

To transform the aerial to the Inverted L format (shown in the above diagram) I envisaged that changing the first length of wire, from a 3.5 metre length to 6.5 metre length, and the inductor would be necessary. I had assumed that a coil with less inductance (fewer turns) would be required due to the longer length of aerial wire used in the Inverted L arrangement.

I was, however, proved wrong. On first attempt I left the original coil in place, just adding the longer wire section (from connection box to coil) and raising the pole to its full 6 metres. The point of resonance was only about 20 kHz away from where it was as a sloper.

This, I think, can be explained by the fact that in the sloper configuration, the wire is close to the ground, other objects and vegetation which will in effect add loading to the aerial, Consequently a loading coil with less inductance than might otherwise be expected is required. So, by happy coincidence, the same coil is also suitable for the longer wire that is at a greater height (and therefore is influenced less by ground loading).

This, quite by chance, made the swap from Sloper to Inverted L a little easier since there was no need to wind a second coil - the change being made by merely adding the longer first wire section and pushing up the pole to its full 6 metre height.

If there are no concerns or objections to installing a tall pole at the bottom of the garden, then the Inverted L arrangement could be left as a permanent installation. In that case the pole need not be a fibreglass fishing pole, although it is still an attractive lightweight method, instead a couple of 3 metre long sections of treated timber could be joined together to form a 6 metre high post - the higher and longer the better though!

These links may be useful in helping to calculate loading coils:

Antenna Loading Coil Calculator : http://eweb.chemeng.ed.ac.uk/jack/radio/software/loading.html

Coil winding design program : http://ecosse.org/jack/radio/software/newcoil3.html

Loaded dipole calculator by K7MEM : http://www.k7mem.com/Electronic_Notebook/antennas/shortant.html

Ring Core Calculato r: http://www.dl5swb.de/html/mini_ring_core_calculator.htm

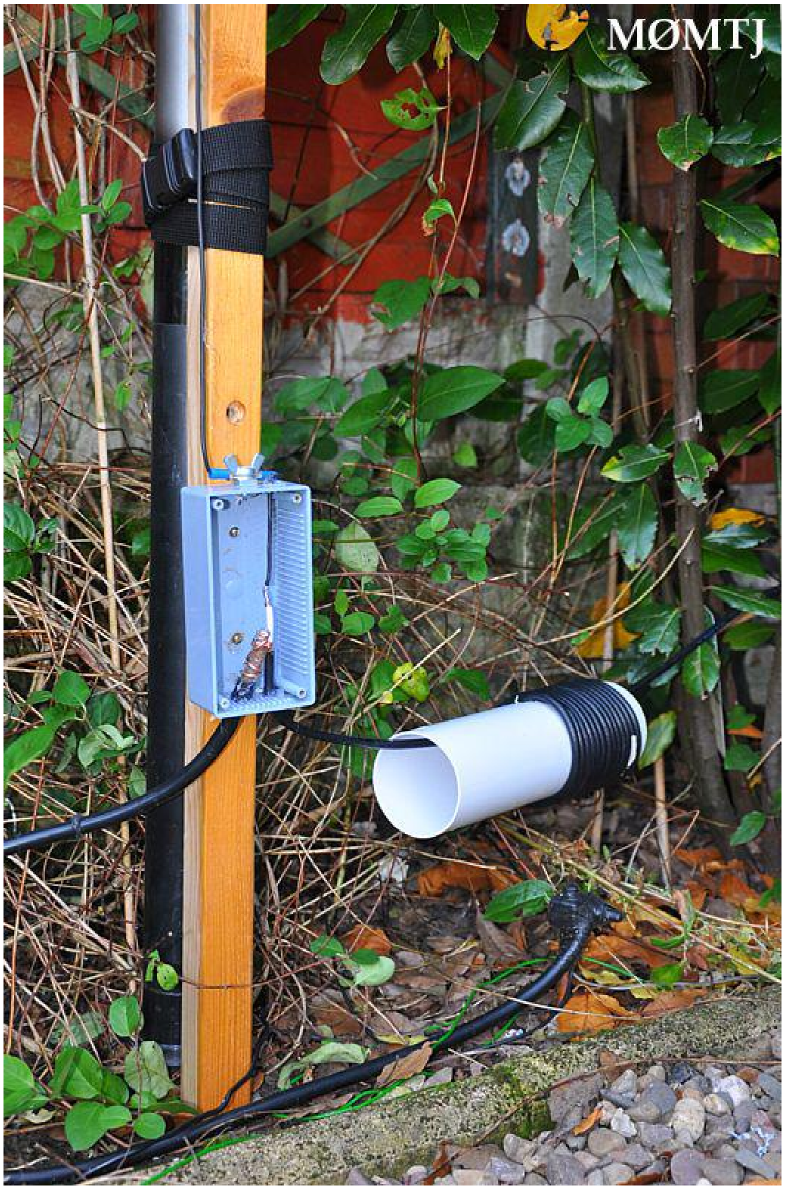

Top Band Inverted L - photograph showing the wooden support post with telescopic pole attached with connection box, cables and earth rod. The cable entering from the left is the earthing strap which is connected to ground rods and radial wires. The coaxial feeder cable exits from the bottom right of the box and is wound around a length of white plastic pipe to form a simple RF Choke.

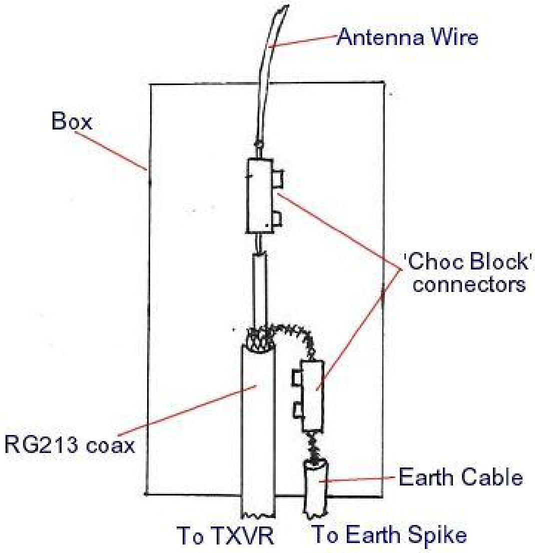
Antenna Wire
Box
'Choc Block' connectors
RG213 coax
Earth Cable
To TXVR
To Earth Spike

My original implementation which used 'choc block' connectors. The final design has all soldered joints with a stainless steel bolt and wing not at the top of the connecting box.

Top Band Inverted L - pulley on top of the 6m telescopic pole

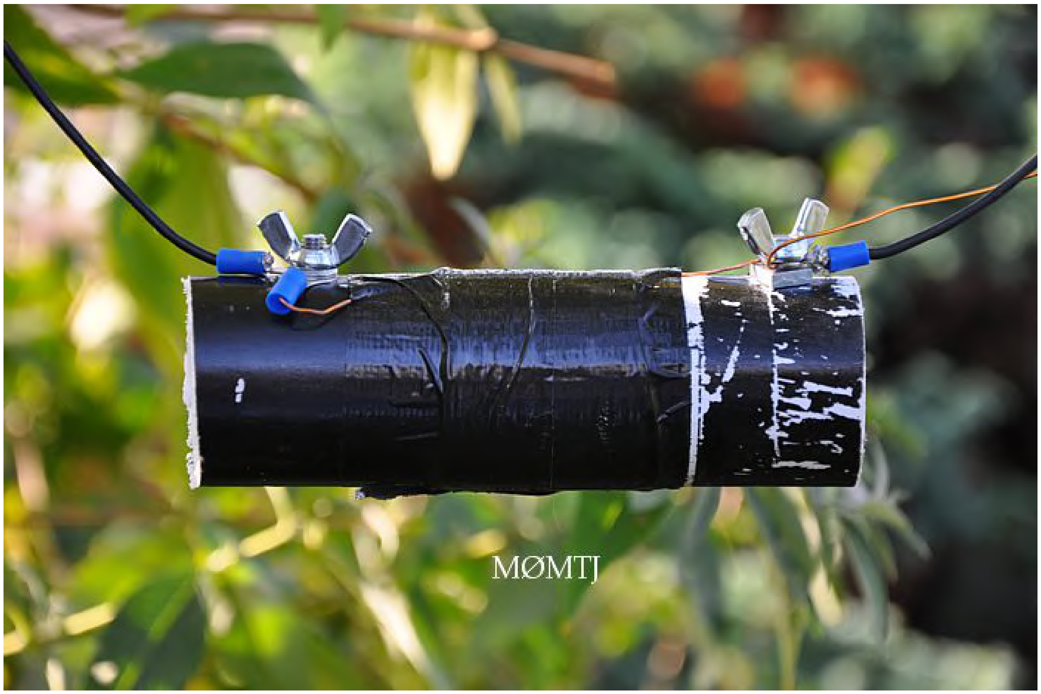

Top Band Inverted L - loading inductor.

This is the unfinished coil, the connections on the left are properly finished off using lugs, but the connections on the right are unfinished to allow adjustment of the number of turns. When that is settled the final lug is soldered to the end of the enamelled copper wire and held in place with the bolt and wing nut. The coil is 39 turns of 0.9mm e.c.w., wound on a 5cm diameter former made out of an empty silicone bathroom sealant tube. (I would have preferred to use slightly thicker e.c.w. of about 18 swg 1.2mm diameter, so that may get changed in the future)

Enamelled copper wire:
http://www.sycomcomp.co.uk
http://www.esr.co.uk/electronics/cable-copper.htm

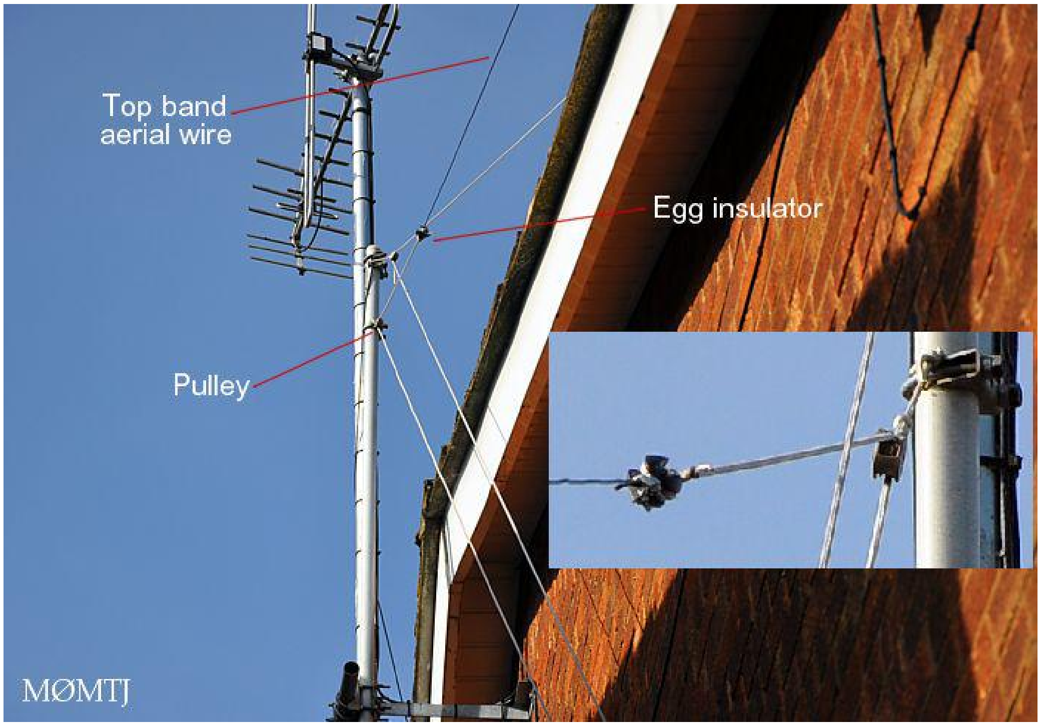

Top Band Inverted L - pulley, rope and insulator supporting the wire at the far end

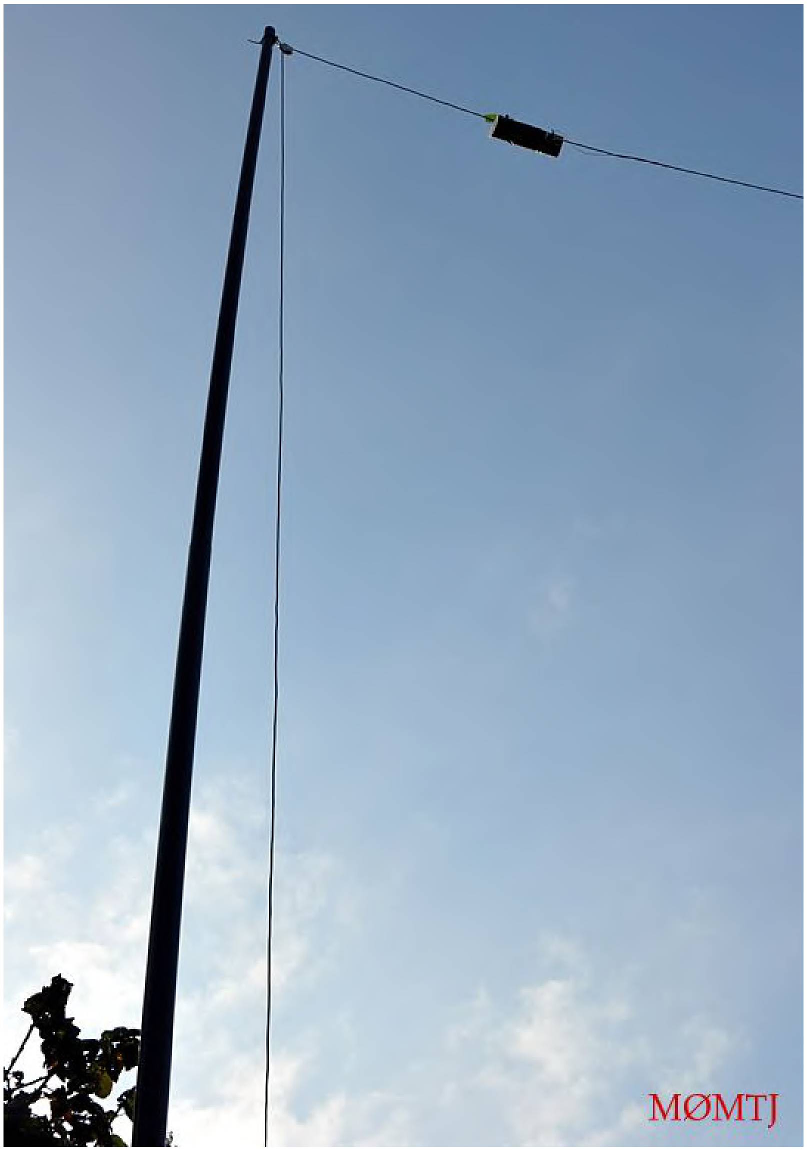

Top Band Inverted L - supported by a 6m long fibrelass (roach) pole

Top Band Inverted L - looking up!

The photo exaggerates the bend a little, but it's not enough to worry about. The bend could easily be corrected by adding a back guy to pull the pole back towards the vertical position, but for this job it's fine as it is.

DOES IT WORK?

Yes! This aerial does work, I am glad to say!

It works better than my initial short sloping wire shown further down this page, which is what one would expect!

Rather than merely settling for a local ground wave signal I was very pleased to contact Top Band enthusiast Steve G7KLJ on my first attempt. Steve is 170 miles distant from me and the night time sky wave provided a pretty good result. G7KLJ gave me a 5 / 8 with my 100 watts on 1843kcs using my small aerial. I gave G7KLJ 5 / 9+10 with Steve's large doublet and 400 watts (lovely audio too Steve!).

Admittedly that's potentially a 16dB difference, which is pretty huge, but G7KLJ had the benefit of 300 more watts - that's a 6dB advantage. It also depends on how the individual S meters compare. Steve seemed to indicate that his S meter was a bit on the lazy side - 1 S Point pessimistic? But maybe my S meter similarly pessimistic?

Compared to a full size inverted L my loaded aerial is less than half the size and is obviously going to lose out, but by how much? Half the size, half the effective power - 3dB down? Probably a lot more - 6dB? There's one S point lost already (Who wants to bet that it's a lot more than that?). Since I have a ground system that is far from ideal there's another few dB's lost - another 6dB? That would be 2 S Points lost before taking into account anything else. This is all pure guesswork and speculation admittedly and probably quite groundless (almost literally in my case), and might best be described as pointless ramblings! (Where did the delete key go?) The pessimist? The realist?

The only thing proved by this single event is that my aerial radiates at least something useful - and that G7KLJ has a great station! www.g7klj.com

Droitwich

OK so it's not Droitwich, but it does work and gets me on Top Band when otherwise there would be no chance.

I am happy - but as with anything improvements can always be made. The aerial could be taller perhaps, though I cannot make it any longer. The earth system could be far far better, though I would struggle to accommodate any improvements at the present QTH. I could use better coaxial cable and also (importantly) attempt to lower the received noise by experimenting with chokes or burying the feeder to reduce common mode currents.

Plenty to think about and plenty to play with!

Important notes on effective Grounding by Jim K8OZ

Mike - I was reading about your work on the 160 meter Inverted L, and it makes me want to go out and build some more antennas! Congratulations. Your story is fascinating, and very well documented.

The only thing I can offer as a suggestion is to get as much radial wire along the edge of your property as possible (assuming your XYL will not allow you to bury radial wire all over your yard). Even if you can only run multiple wires 1/8th of a meter apart from each other, and parallel to each other, your losses will be reduced. The ground losses have quite an impact on your transmitted signal, so any wire you can "hide" along the edge of your property will help improve your signal strength - little, by little...! { It may also affect your resonant frequency slightly, but that's easy to deal with by adjusting with an antenna tuner or slightly changing the loading coil. }

Good luck OM, and keep up the refinements on your antenna system. You're doing great! 73,

Jim, K8OZ
Albuquerque, NM

Top Band Antenna by Mark, G0MGX and Vince G0ORC

Mike, I've been reading with interest your musings on top band antennas and have tried to build a replica of yours today with a fellow ham G0ORC. Thanks very much for the information and the link. You can see the results of my attempts here:

Construction: http://g0mgx.blogspot.co.uk/2012/08/top-band-what-happens-there-then.html

It Works!: http://g0mgx.blogspot.co.uk/2012/08/well-this-top-band-twig-does-it-work.html

Mark. G0MGX

Mike,

I know you have been in contact with my good friend Mark G0MGX but I felt I needed to say thanks and acknowledge your work on a 160m sloper.

It works very well here and, thanks to Mark who built several inductors until one gave us an SWR of 1.1 on the CW end of the band and also braised several copper rods together for form a reasonable earth under a large pine tree. He has passed on to me details of the website of K7MEM which may well inspire me to try a sloper for 80m as well.

I've always wanted a top band antenna but felt that I didn't have the room - but thanks to your idea and Mark's enthusiasm for the project I now have what I wanted - I'm being heard (and can hear) into European Russia with it so it certainly works!

Thanks again - I enjoyed your website very much. (Just heard A65BP as well!)

73, Vince G0ORC http://www.qrz.com/db/G0ORC

Further Developments of this antenna...

Another Top Band Aerial by MØMTJ

After reading the article "Top Band in a Small Garden" in the August 2012 edition of Practical Wireless Magazine ('PW'), I investigated further development of my Top Band Sloper / Inverted L Antenna (shown above).

The article by Stuart Craigen G4GTX described a compact wire antenna for top band that could be accommodated in small gardens. The antenna is shunt fed at the base, the coax being connected across a 7 turn coil (wound on an off-cut of an empty silicone sealant cartridge), with the braid connected to the earth stake and radial ground wires. The centre conductor is connected to the aerial wire.

I could not quite accommodate the shape that Stuart suggested for the aerial wire, but it is very similar: For my version the feed point is at the bottom of the garden near the fence immediately adjacent the earth stake. The first section of aerial is about 9.14 metres long - the wire rises 3.5 metres vertically up the 6.5 metre tall telescopic pole, at the 3.5 metre point the wire is held in place on the mast by a small bungee and then folds over almost horizontally and runs for a further 5.64 metres where it connects to the loading inductor.

The inductor consists of 36 turns of 18 swg (1.2mm diameter) enamelled copper wire on a 50mm diameter, lightweight plastic former (again made from an off-cut of an empty silicone sealant cartridge). After the inductor there is a further 6 metres (approx) of wire rising to a dog-bone insulator, fixed to the end of a length of para cord that runs up to the pulley shown in a previous photograph. From the dog-bone insulator the aerial wire runs back, in the opposite direction, for about 12 metres, returning to the top of the 6.5 metre tall mast at the bottom of the garden, tied off to an egg insulator which is itself fixed to the top of the mast by a very short length of para-cord fixed to the mast by a thick nylon cable tie.

If the wire dimensions are changed, the number of turns on the inductor will have to be changed accordingly. Similarly my antenna is located quite near the 80m/40m Inverted L and the interaction of the antenna affects the tuning and hence the number of turns required on the loading coil.

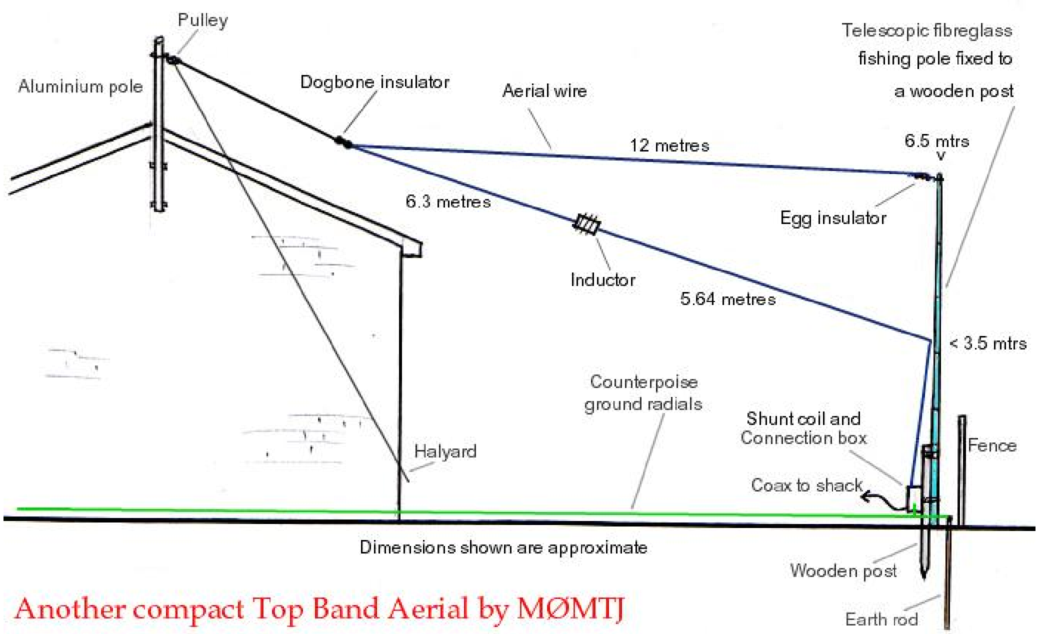
Pulley
Aluminium pole
Dogbone insulator
Aerial wire
Telescopic fibreglass fishing pole fixed to a wooden post
12 metres
6.5 mtrs
6.3 metres
Egg insulator
Inductor
5.64 metres
< 3.5 mtrs
Counterpoise ground radials
Shunt coil and Connection box
Fence
Halyard
Coax to shack
Dimensions shown are approximate
Wooden post
Another compact Top Band Aerial by MØMTJ
Earth rod

Another Top Band Aerial (drawing not to scale)
The inductor is 36 turns of 1.2mm diameter (18 swg) enamelled copper wire on a 2 inch former

With that arrangement there is about 27 metres of wire in the air. The point of resonance is 1900 kHz, but this can be adjusted by altering the number of turns on the inductor. As expected bandwidth is quite narrow. As with any antenna fed against ground the earthing needs to as extensive as possible - as noted previously, above. I suspect that it may be better that the top wire should be horizontal, or sloping slightly upwards, however with the available space and materials available the best that I could achieve had the top wire sloping slightly downwards to the post. Lowering the position of the dog-bone insulator or using a taller post could overcome this.

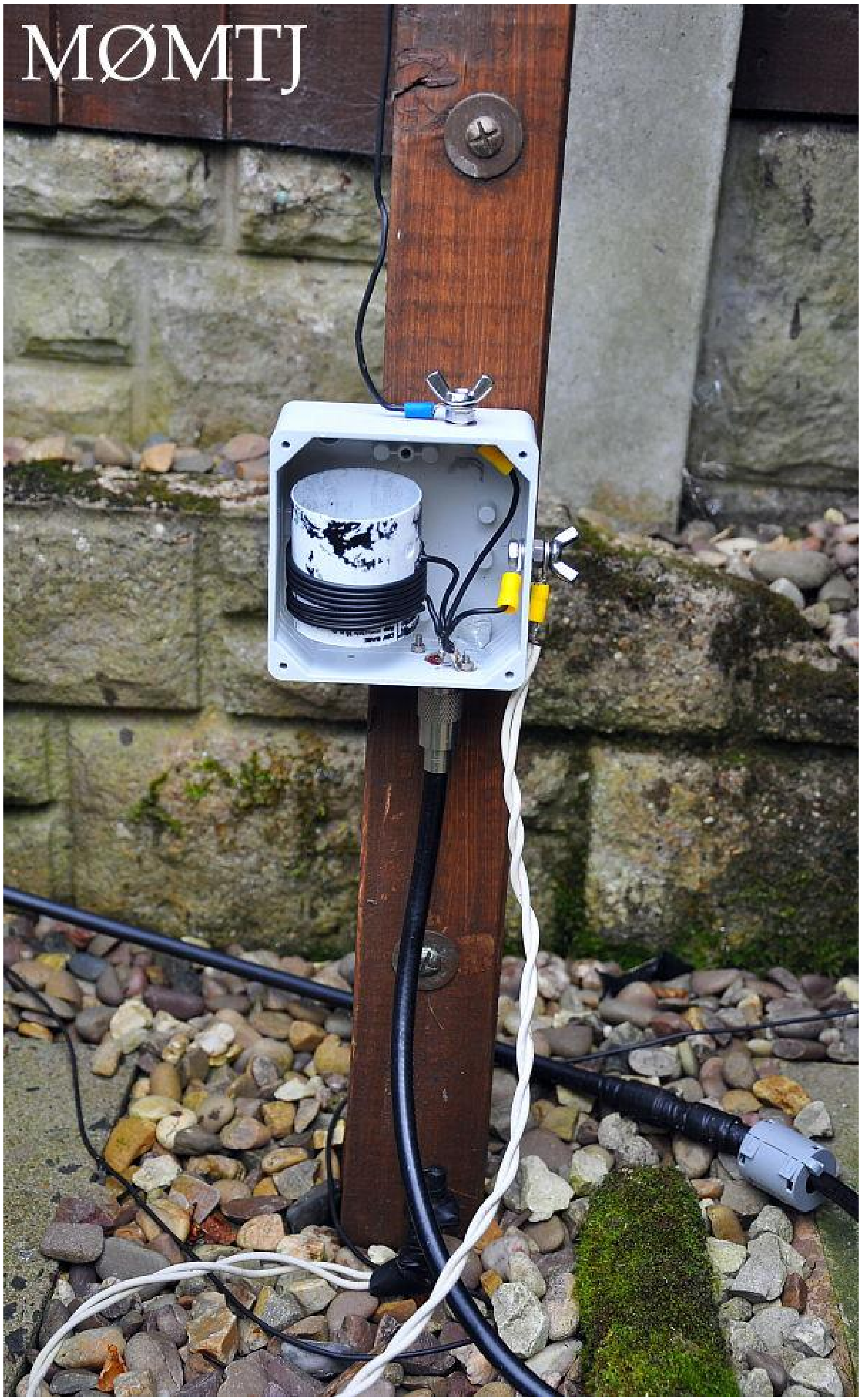

Photograph of the feed-point of the compact, folded, Top Band antenna showing the 7 turn shunt fed coil in the weatherproof housing. Nuts and bolts are stainless steel.

First Test: A brief on air test on 8th September 2012 between 2100 and 2130 UTC on 1.933 MHz brought in two prizes: Peter EI7JM near Malin Head and Tony M3LTD in Naseby, Leicestershire. EI7JM was a true 5/9 to me and he gave me a report of a true 5/9. M3LTD was 5/7. Many thanks to both stations for the useful contacts.

For different sized gardens the best idea might be to get as much aerial wire into the air and then make adjustments to the point of resonance by simply adjusting the number of turns on the inductor.

Practical Wireless Magazine
Obtain the article "Top Band in a Small Garden" by Stuart Craigen G4GTX from Practical Wireless Back Issues department

Stuart Craigen G4GTX comments:

Hi Mike, I have just been looking at your fascinating antenna page on top band antennas! Glad to see that you had a go at building my shunt fed wire vertical that I did for Practical Wireless! I found it tremendous. A few weeks ago I was talking to G3YFN in Newcastle at about 4pm when I was called by G4BIM in the Isle of Wight! I also got a QSL 5/9 from Luxembourg!. Have fun.with it!

Thanks for the nice comments about my shunt fed 160m aerial. I enclose some QSL cards which you might find interesting to look at! Note the times of the QSO's and the reports! At night I have been told that I sound like the BBC from some European stations!!

Note G4BIM Isle of Wight 5/7 from Sunderland at 4.30 pm and Aberdeen at night. It certainly works for me. I think that the shunt feeding is the answer as it taps into the aerial at the 50 ohm point. Putting the loading coil away from the base should increase efficiency hence approx midway. A chap near Newcastle built one of these a few years ago and his groundwave signal was 5/9+20 here in Sunderland 12 miles away!!!
73
Stuart G4GTX

More ideas....

Adding Top Band to the 40m / 80m Inverted L antenna using a switchable loading coil:

Due to an aborted house move in 2010 I had removed all the antennas. While re-establishing the aerials in 2011, and considering space limitations, I decided to experiment with adding a loading coil to the 40m / 80m Inverted L aerial. The loading coil has a link wire to short it out of circuit when using the aerial for 80 metres and 40 metres. The link wire is removed when Top Band is required. The coil consists of approximately 37 turns wound on a piece of 68mm (2.8 inch approx) diameter PVC pipe:

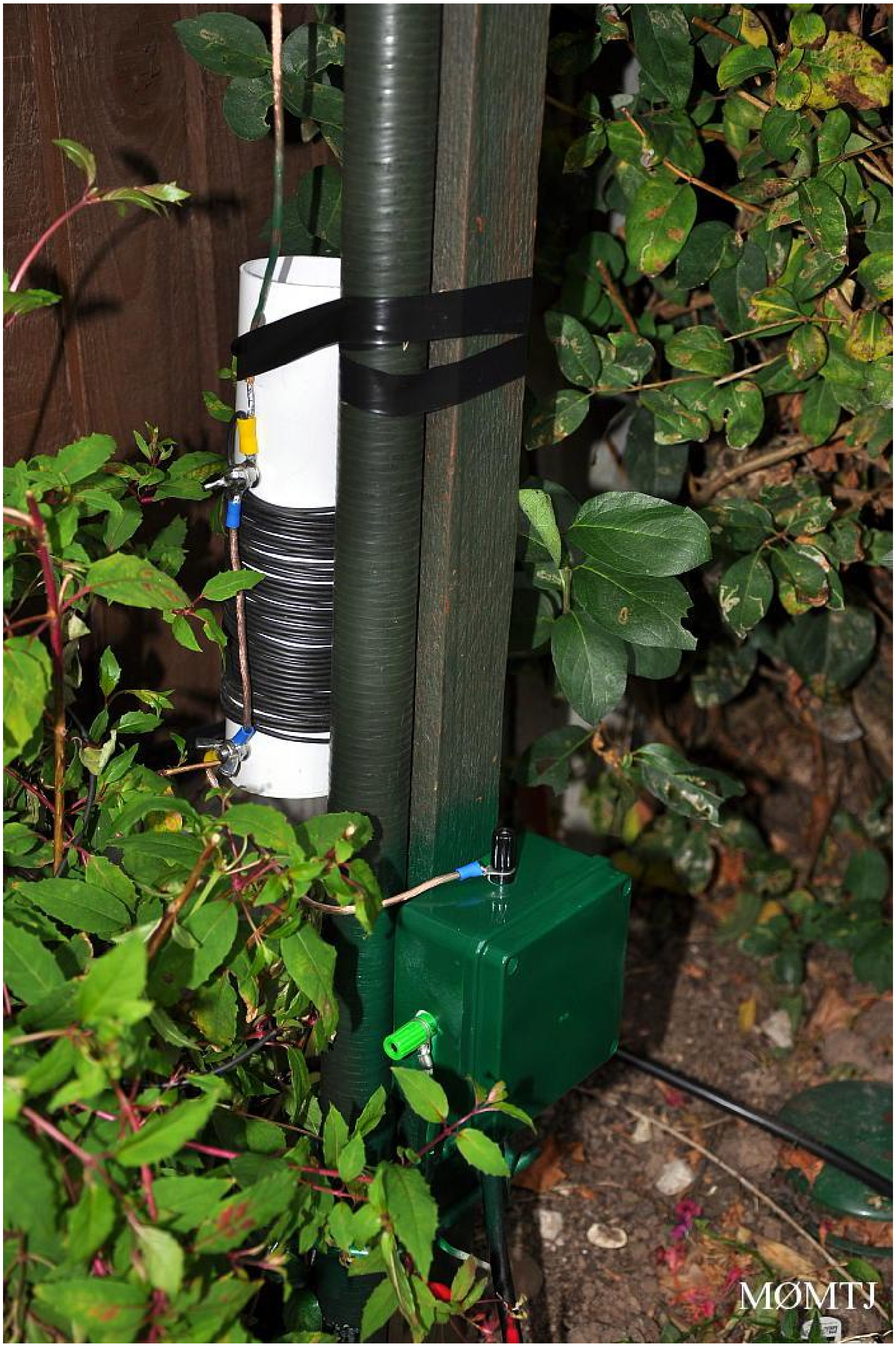

Work In Progress! - Adding 160 metre loading coil to the 80m / 40m Inverted L Aerial.
The loading coil has a link wire to short it out of circuit when using the aerial for 80 metres and 40 metres.
The link wire is removed when Top Band is required.
The coil consists of approximately 37 turns wound on a piece of 68mm (2.7 inch approx) diameter PVC pipe.

More about the 80 / 40 metre Inverted L Aerial on Antennas Page 1 >

Simple Vertical Top Band Aerial Using A Fibreglass Telescopic Fishing Pole (Roach Pole)?

If there really is no space to accommodate a top band dipole, inverted V or even the inverted L above, then another option could be a simple vertical. Again based on a 10 metre long fibreglass roach pole, this could be installed as a permanent aerial, or only deployed as and when required on a temporary and stealthy basis.

The 10 metre long fishing pole would be erected by standing it on the ground, using three or four nylon guy ropes as support, or tying it to a wooden ground stake driven into the ground. The vertical radiator would consist of about 9.5 metres of PVC covered wire fixed at the top of the pole and with a loading coil at the centre. As with the design above, the loading coil could be wound on a 5 cm diameter using 1.2 mm diameter enamelled copper wire. To make the former as light weight as possible, once again, the tube from an empty bathroom sealant (or silicone sealer etc) gun could be used. They are, conveniently, 50mm in diameter.

I have not experimented with the number of turns required for the centre loading coil, but from experience with the above aerial, the figure may lie somewhere between 50 and 100 turns.

The aerial would be fed with coaxial cable and would need a good earth rod at its base and a good number of ground radials, as along as possible.

Previous Top Band Aerial Experiments:

Top Band Inverted V

This is a considerably shortened dipole for 160 metres, using loading coils to reduce the length of each leg to about 7.5 metres length. Each coil is wound on a 5 cm diameter former and consists of approximately 120 turns of 1.2mm diameter enamelled copper wire. The dipole was arranged in an inverted V configuration.

Being as a dipole is a balanced aerial, it was fed at its centre using balanced twin feeder for lowest loss and, being balanced, lowest noise.

In this configuration it was very quiet, but did also appear to receive very well. At my QTH the noise on top band generally ranges from S7 to S9, on this inverted V the noise was S4 to S6. However it did not transmit well at all - I believe this to be not only because the dipole is very short (only 15m) but because the angle at which the wires at the top of the V was far too acute - only about 45 degrees. The angle at the top of the V really should be over well 90 degrees - maybe something like 110 degrees would be best. Clearly this seems to be a problem.

I had not got the space to separate the wires any further, but it was an interesting experiment that appeared to demonstrate that a balanced dipole fed with balanced twin feeder could help reduce noise. If more space was available this is an antenna that I would wish to pursue further.

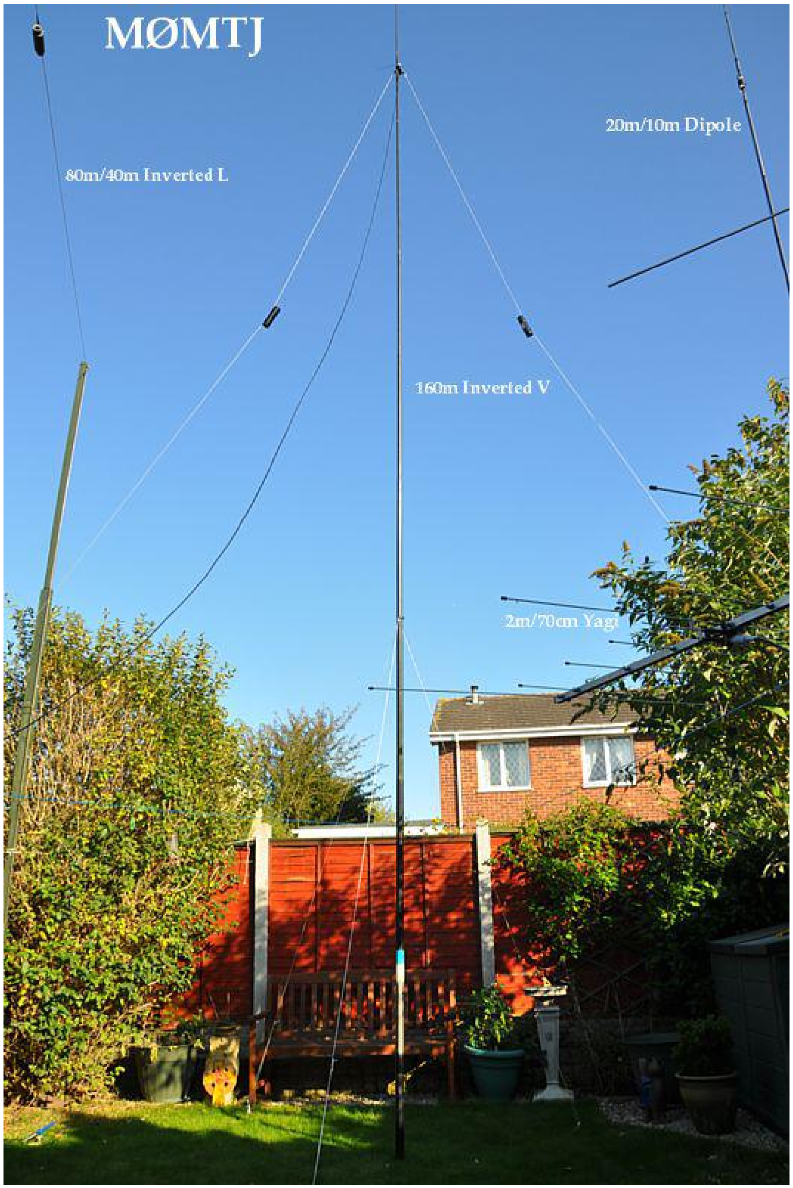
MØMTJ
20m/10m Dipole
80m/40m Inverted L
160m Inverted V
2m/70cm Yagi

Top Band Shortened (loaded) Inverted V

Short Base Loaded Sloper - a first attempt at a top band antenna

160m metres is probably the noisiest band, but I wanted to give it a go anyway even though space at my QTH is at a premium and fitting in a suitable antenna is quite challenging. The most promising candidate that I initially found, apart from small commercially available antennas, was the "Practical Antenna For 160 metres" described by Frank G3YCC and featured on the website of IW5EDI and linked to on the pages of www.dxzone.com

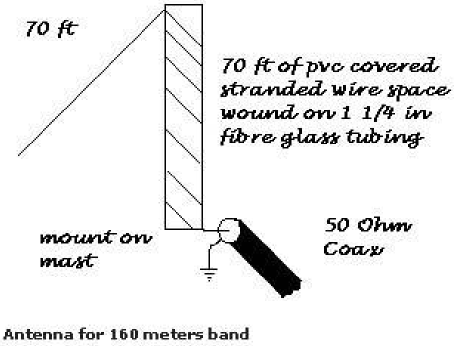
70 ft
70 ft of pvc covered stranded wire space wound on 1 1/4 in fibre glass tubing
mount on mast
50 Ohm Coax
Antenna for 160 meters band

G3YCC Top Band Antenna

This Top Band antenna consists of 140 feet of wire, 70 feet space wound around a non-conductive (fibreglass) pole or tube about 6 feet long, with remainder forming a sloping wire falling back to near ground level. The antenna is UN-Balanced and is therefore fed with unbalanced coaxial cable, with the braid connected to a ground stake at the base of the aerial.

The above design is intended to have the coil section mounted on top of a mast or pole with the wire running downwards. I could not accommodate that arrangement, so I adapted the design to produce the antenna shown below:

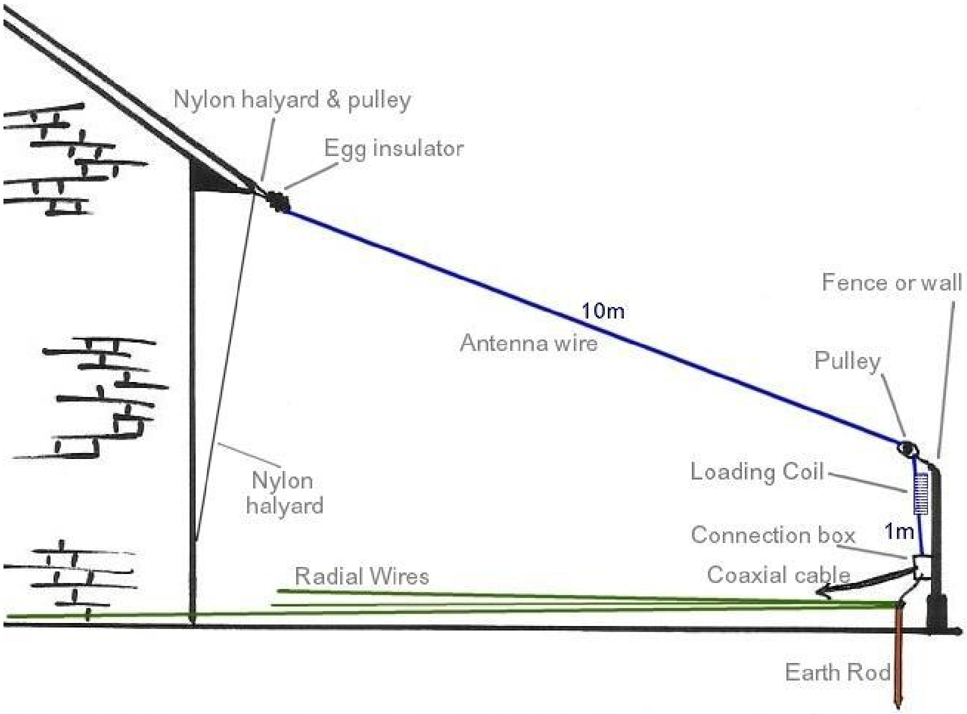
Nylon halyard & pulley
Egg insulator
Fence or wall
10m
Antenna wire
Pulley
Loading Coil
Nylon halyard
Connection box
1m
Radial Wires
Coaxial cable
Earth Rod

The drawing above shows the loaded Top-Band Antenna - a first attempt at an aerial for 160m - consisting of the radiating top wire, a base loading coil, earth stake and additional earth radials that act as a counterpoise: The top wire is about 10 metres long and slopes down to the loading coil which consists of 64 turns of the aerial cable wound around a 4" diameter PVC pipe and then falls vertically for about 1 metre where it enters the connection box. The halyard arrangement allows the antenna wire to be quickly dropped for adjustment and then raised back into position. There is an earth rod at the base of the antenna connected to several ground radials and a second earth rod a few metres away.After performing some tests I find that it does work reasonably well for local contacts ground wave contacts around the town, but it could be better and is certainly very disappointing for longer distance work.

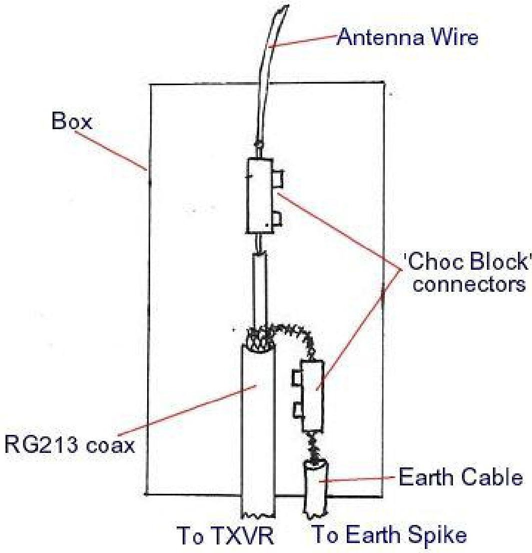
Antenna Wire
Box
'Choc Block' connectors
RG213 coax
Earth Cable
To TXVR
To Earth Spike

Above: The Connection Box

Read more developments here>

Efficiency

The aerial shown above is obviously very much shortened and therefore is very inefficient when compared to a full size resonant quater wave aerial. A shortened aerial still needs to make best use of the space - so the antenna wire needs to be as long as possible. Resonance is adjusted by varying the inductance of the loading coil. Also it is necessary to use the largest earth system as is practicable to obtain the best possible efficiency.

Performance

In my configuration described above it is admittedly a very low efficiency antenna - but at least it gets one on Top Band! It's ok for local working around a town or a city, but it's certainly not a DX aerial by any stretch of the imagination. I could hear many stations 100 to 200 miles away, but they could not read my signals. For local ground wave work it was adequate.

I suspect that if the original design by G3YCC was employed, i.e. the coil mounted at the top of a pole and the loading wire running down, that it may work rather better than my implementation. The problem that my implementation has is that the loading coil is at the bottom of the aerial, near the ground. Better efficiency may be acheived with centre loading. The longer the radiating wire the better too.

More Details

The antenna was fed by coaxial cable from the shack in the bedroom at the front of the house to the feed point at the base of the fence at the bottom of the back garden. The coaxial cable enters a small plastic box, screwed to the fence post, where it is terminated: The outer conductor of the coax is connected to the earth cable with a 'chock block' connecter, the earth cable going to a four foot long earth rod hammered into the ground near by. The centre conductor of the coax is connected to the antenna cable also using a 'choc block' connector.

The antenna wire runs vertically up the fence for about 1 metre to the loading coil which consists of 64 turns of the antenna wire wound around a piece of 4 inch diameter PVC pipe. About 32 turns a close wound and the other 32 turns space wound; The inductance, and therefore the point of resonance could be easily adjusted by altering the spacing of the windings. Once set at the required point of resonance the windings can be held in place with duck tape.

The antenna wire exits the coil at the top and passes through a pully secured to the top of the fence and then slopes upwards for about 10 metres to the fixing point at a convenient point on the house, in this case the eaves. The end of the wire is tied off to a plastic 'egg insulator'. The egg insulator is tied to a nylon halyard wich passes through a second pully at the fixing point. This allows the antenna wire to be quickly dropped for adjustment and then raised back into position.

The earth system should be as extensive as possible for better efficiency; e.g. earth stake and a number earth radials, as long as possible around the garden to provide a counterpoise.

Link:

Visit www.g7klj.com - the website of 160 metre entusiast Steve G7KLJ

OTHER THINGS

Antennas for HF

There are dozens, if not hundreds, of antenna designs from which to choose. All have their own proponents. Some are genuinely good while others might be considered as nothing more than glorified dummy loads! The key thing for me was to choose the most effective an antenna that could fit within the constrains of my small back garden and also one that would not be too ugly.

I decided that wire antennas, made from PVC covered stranded wire, would be least objectionable on visual grounds. After all they look like glorified washing lines and almost everyone has a washing line**!**

Knots to use when fixing Wire Antennas

It's important to use the correct knot for the job. I find that the Bowline is very useful for fixing end, egg and dog-bone insulators to the ends of the wire and/or ropes. The Round Turn & Two Hitches, Anchor Bend and Buntline Hitch knots are very good for tying a rope to a pole or a mast. A Double Sheet Bend can join two pieces of rope together - even if they are of unequal size. 'Animated Knots' will show you how to do them: http://www.animatedknots.com

The antenna or antennas really are the key to an effective amateur radio station. The most expensive radio will be of little value unless it is used with the best possible and most effective antenna; so I had to set about finding the best antennas that I could accommodate for HF and VHF work.

The dilemma facing most radio amateurs is that antennas for HF are often large and difficult to accommodate in modern small back gardens. Additionally there may be objections on the grounds of visual impact. Some people do not regard antennas as the beautiful creations that radio enthusiasts do!

Cables and Feeders

As I have learned by experience and from the tutorials I have studied as part of the licensing exam process, a very important consideration concerning the antenna system is the loss incurred in the cable that connects the radio to the antenna. A loss of 3dB sounds like a small amount when you say it quickly, but that equates to only HALF of your transmitter power reaching the aerial! We'd soon complain if our new 100 watt transceiver only produced 50 watts - so why stand the chance of losing this much power in the feed line?

Starting off as an M3 licensee I could only use 10dBw (10 watts) power at the antenna termination, so if I had a feeder with a loss of 3dB I would have to run the transmitter at 20 Watts to get 10 watts to the aerial!

Coaxial Cables

For HF using a single band resonant antenna, a dipole for example, Mini 8 or even RG58 is probably fine since the SWR should be between 1:1 and 2:1. However if one was to consider using the antenna on non resonant frequencies the VSWR will be higher, maybe very much higher, and the consequential losses in coaxial cable will be greatly magnified and power losses very significant. I did not want to run such an inefficient station.

For this reason I chose to use low loss RG213 coaxial cable for my Inverted L since I may need to use it on bands other than the 40 metre and 80 metres that it is intended for and where the SWR may be 4:1 or 5:1. RG213 will help keep feedline losses to a minimum. No lossy RG58 on Mini8 for me! Westflex 103 would be even better, of course.

For VHF and UHF I now consider Westflex103 as the minimum standard of cable to use, it's half the loss of even RG213. Losses at VHF are much higher than HF and at UHF they are even higher, so Westflex 103 helps preserve as much of that precious transmitter power as possible.

See more about Coaxial Cable Losses here >

Balanced Feeders = Low Noise

It's difficult to feed a centre fed wire dipole with RG213 or Westflex 103, both cables are really too thick and heavy. RG58 or Mini 8 are lighter and therefore more suitable for suspended wire dipoles, but to match the unbalanced coaxial cable to the balanced dipole would need a 1:1 'choke balun' at its centre, as I have learned. However, as mentioned above, when attempting to use the dipole at non-resonant frequencies the VSWR will be higher and the losses in the coaxial cable will be very much larger too, and hence much less of the precious transmitter power will be actually radiated by the antenna. Not an ideal situation

I have learned that the only really sensible way to feed dipole type antennas, which are 'balanced', is to use a balanced feeder such as 300 ohm ribbon cable, 450 ohm ladder line or the best option, it seems, might be 75 ohm Twin Feeder. All such twin feeders are extremely low loss, much much lower than any coaxial cable, so low that it could almost be considered lossless by comparison. Of course nothing is lossless, but twin feeder is as near as you'll get, so that's how I will feed my dipoles from now on! See the note on Doublet antennas below.

Tony Nailer of Spectrum Communications notes that: "75 ohm twin feeder is lower loss than coax. It allows the aerial to be properly balanced and the very close spacing of the wires prevents pickup or radiation from the feeder. It does not need to be spaced off, unlike ribbon feeder. Use of twin feeder makes this aerial much lower noise than one fed with coax. Also importantly it generates less TVI !! Note that the [trapped dipole] aerial is generally 72 ohms, and will need to be used with an ATU with transistorised rigs which are unforgiving about SWR mismatches."

Spectrum Communications are now supplying a new design of top quality, very low loss, twin feeder. It is rated at 2kW at 100 ohms. It can be supplied by 100 metre reel or by the meter, or in various lengths with the ends expertly terminated and made off. I can confirm the superior quality of this product and its low noise properties.

Spectrum Communications also supply a very high quality, well made 1:1 Balun that is perfect for connecting the twin feeder to the unbalanced input of an A.T.U. of transceiver.

Doublet or Dipole (Horizontal or Sloper)

I have also experimented with a second HF antenna - the 'doublet' style antenna. I fed this with low loss twin feeder and cut it for the lowest frequency of operation, in my case 20 metres (14 MHz).

Some initial experiments have shown that it is very effective on 20 metres, better than my Inverted L - which is designed essentially for 80m and 40m.

In theory a doublet antenna, when fed by twin feeder (NOT coaxial cable) and matched at the transmitter end via a

1:1 balun and an ATU (or to use a more correct term, Antenna Matching Unit) at the transceiver should be able to work on all bands with a higher frequency than the band that it is cut for.

On other bands such as 17m (18 MHz), 15m (21 MHz), 12m (24 MHz) and 10m 28 MHz) the doublet showed generally lower VSWR than the Inverted L .... however (and here is a nice lesson) it is actually not quite as effective as the Inverted L! I think that size may matter here - the overall length of the Inverted L is much longer.

This short 14 MHz doublet is only about 10 metres in total length and was suspended at about 25 feet high at one end from the same fibreglass pole fixed at the apex of the roof and taken down to another 16 foot high wooden pole installed on the other side of the back garden in somewhat of a 'sloper' style.

Trapped Horizontal Dipole or Vertical Antenna for 20m / 10m

I also experimented with a trapped wire dipole cut to be resonant for 20 metres and 10 metres, again fed with twin feeder and matched at the radio end via a 1:1 balun and the ATU (AMU !).

I may also try an antenna for 20m / 15m and 10m in a Vertical 'ground plane' arrangement which should give a lower angle of radiation than a horizontal dipole - therefore better for longer distance DX.

At the moment I am using the 20m / 10m trapped dipole in a rather unorthodox arrangement. I have installed a fibreglass fishing pole, mounted vertically on a stake on one side of the garden. Most of the wire dipole is fixed to the vertical pole, but being as the dipole is somewhat longer than 7 metres, a portion of the wire runs away from the pole horizontally along a fence panel about 1.5 meters above ground level. It's vaguely L shaped, but mostly vertical which should help with the angle of radiation. It seems to work very well indeed - I like the mostly vertical idea.
The antenna can be seen on this page >

Other Dipoles:

Spectrum Communications can supply a full size W3DZZ style antenna with a 7MHz trap for use on 40m and 80m and usable on 20m, 15m and 10m. A half size version with a 14 MHz trap for use on 40m and 20 metres, plus 15m and 10m is also available. As a special order an even shorter version using a 28 MHz trap is available for 20 metres and 10 metres. http://www.spectrumcomms.co.uk/G2DYM.htm Al versions designed to be used with balanced line feeder - lince they are balanced aerials - for low noise and lowest loss.

PDF Document - The W3DZZ Antenna -
http://www.users.icscotland.net/~len.paget/GM0ONX%20trap%20dipole.pdf (But rather than using coaxial cable with a 'choke balun' at the centre of the dipole, try using twin feeder with the Choke Balun at the radio end for less power loss!)

Links

Enamelled copper wire supply: http://www.esr.co.uk/electronics/cable-copper.htm

Loading coil calculators: http://www.k7mem.com/Electronic_Notebook/antennas/shortant.html

Ring Core Calculator: http://www.dl5swb.de/html/mini_ring_core_calculator.htm

Vertical Antenna For 50 MHz, 144 MHz and 430 MHz

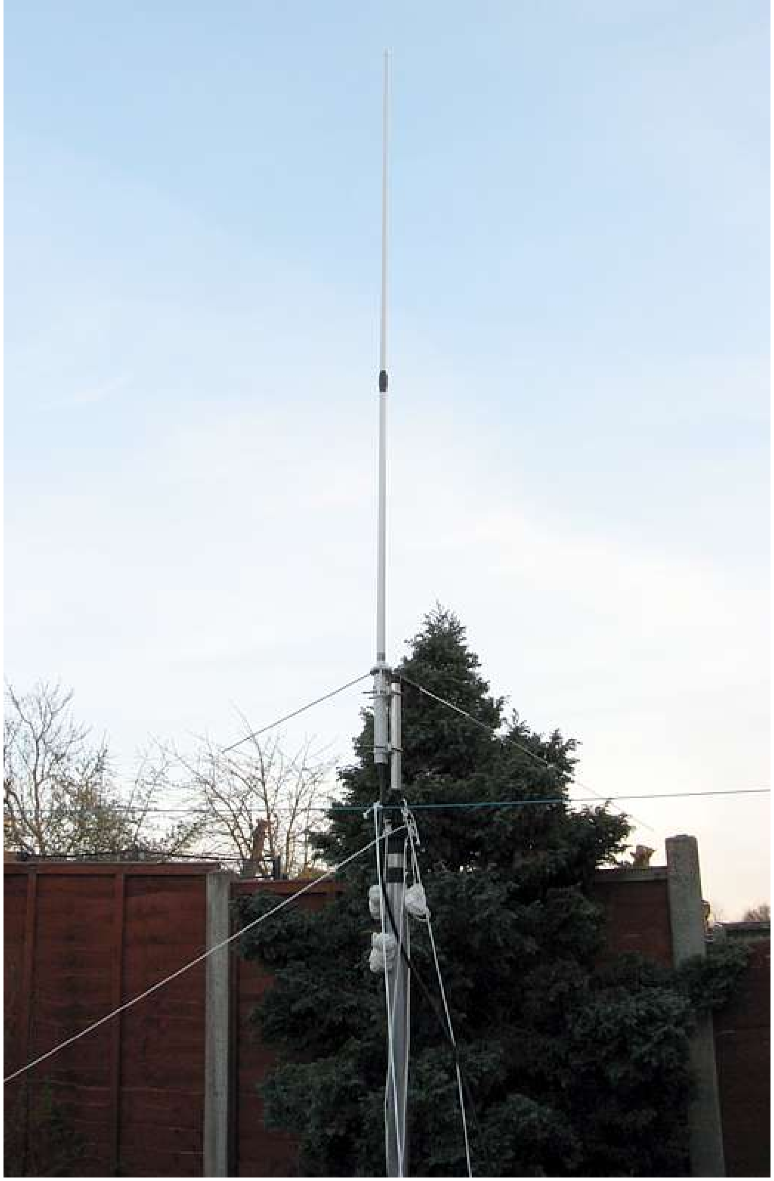

Watson W-2000 on Telescopic Mast at the lowest position
more on this page

OTHER ANTENNAS

Our good friend in Australia Felix Scerri, VK4FUQ, uses Inverted V antennas but also highly recommends the Quad Loop style antenna for HF work. These are well worth investigating, and you can read more here: Antennas 3 see the antennas at the MØMTJ QTH here with many more antenna ideas on Antennas 4 and the Links Page here and here

Antennas 1 | Antennas 3 | Antennas 4 | Antennas 5 | Antennas 6 | Antennas 7

G Whip Antenna Products
G-Whip Antenna Products
Geoff G4ICD offers a multitude of high quality solutions for portable, mobile and permenant base installations

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits | Useful Information

Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites | RSGB | QSL | The Amateur Radio Mini Site Map

WSPR Weak Signal Propagation Reporter

News, Developments, Events, Photographs and other 'Bits 'n' Bobs'

MDS975 Home

Feedback

MAIL

Mike Smith - MDS975.co.uk © 2003 - 2016

M0MTJ

Amateur Radio; Ham Radio; Radio; Transceivers; HF; VHF; UHF; Data Modes; Morse Code; RTTY; PSK31; SSTV; FSTV; Amtor; Sitor Antennas; Aerials; Cable; Coaxial Cable; Twin Lead; Masts; Poles; Propagation; Computer; PC; USB Computer Interface; Microphone Loudspeaker; Filters; Noise Reduction; DSP; Digital Signal Processing; Morse Key; SWR ; Inverted L; Inverted V; Dipole; Doublet.

My Home Page

Amateur Radio Operating Conditions
Antennas
Accessories
Projects & Kits
Data Modes
Information
Portable
QSL
RSGB Ltd
Links
QRM
Bits 'n' Bobs
Contact

WSPR
Weak Signal Propagation Reporter

PORTABLE

Site Map

MDS975 Home

Feedback

QRM from PLT

Site Search
Find
powered by FreeFind

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits | Useful Information
Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites | RSGB | QSL | The Amateur Radio Mini Site Map
News, Developments, Events, Photographs and other 'Bits 'n' Bobs' | WSPR Weak Signal Propagation Reporter

"Everything should be made as simple as possible, but not simpler." - *Albert Einstein*

AERIALS (ANTENNAS) 4 - Notes and Queries

Antennas 1 | **Antennas 2** | **Antennas 3** | **Antennas 5** | **Antennas 6** | **Antennas 7**

Index To Other Antenna Pages:
Antennas 1 : Aerials used by M0MTJ
Antennas 2 : Including ideas for compact aerials for Top Band /160 metres
Antennas 3 : Felix Scerri VK4FUQ discusses Loop Antennas, baluns, masts & other antenna related topics
Antennas 4 : Many antenna ideas from various sources particularly for Multi-Band operation & also gives information about...
antenna trimming, knots for wire antennas and useful antenna rigging accessory ideas.
Antennas 5 : Half Wave End Fed antennas for 144 MHz VHF / 430 MHz UHF and 50 MHz 6 Metre band & J-Pole Aerials
Antennas 6 : Simple and effective H.F. Antenna ideas - Ground Plane and All Band Doublet
Antennas 7 : Omni-Directional - Circularly (Mixed) Polarized Antenna for VHF / 2 Meters.

Multi Band Aerial Options | Useful Aerial Rigging Accessories | Aerial Trimming Chart | Useful Knots | A few antenna related Links

As G4ILO notes: An antenna may have two of the attributes: Small; Efficient or Broadband (works over a wide frequency range without retuning) but never all three.

Marconi spins in his grave every time a ham buys an aerial instead of building it ! *(W1GFH)*

THE QUEST FOR MULTI - BAND OPERATION - In A Limited Space

The aerial is arguably the most important part of any station. No matter what transceiver is being used it is the aerial that is the last and most vital link in the chain and needs to be efficient and effective to radiate the signal to best best effect. Some amateurs are quite content to operate on one or two bands while others might want to be able to operate on many or even all of the amateur HF bands. When I gained my licence I definitely fell into the latter camp!

It seems, then, that the holy grail of many amateurs is the perfect multi-band aerial!

Luckily N4UJW has designed a new limited space 160m through to 70cm marvel antenna the plans of which can be found here:
http://www.hamuniverse.com/antwish.html

Having experimented with various types of antenna I am of the opinion that, perhaps along with many other amateurs, for simplicity a resonant dipole is the most efficient and effective of aerial. A resonant dipole it is only a single band aerial of course, but it is extremely cheap and very simple to make - and it's a very efficient radiator. So one could make a dipole for every band of interest and simply swap aerials to work different individual bands. Unfortunately the aerial described by N4UJW does not exist and compromises, such as lack of bandwidth or poor radiation efficiency, have to be made.

The principal of lowering a dipole cut for one band, removing it and hoisting another dipole cut for a different band in to place sounds pretty straightforward, but would the process become frustrating after a while. I think it could, so what about an antenna that will allow operation on several bands?

G Whip Antenna Products

G-WHIP
Antenna Products
Top Quality
British Manufacture

The Quest For A Multi - Band Aerial

Here is a collection of commercial and 'home brew' (DIY) antenna ideas that will allow multi-band operation, many of which could also be used in a location that offers limited space. Perhaps this only scratches the surface, but hopefully will provide a good starting point and fuel the mind in a quest for a good multi-band HF aerial. Do check out the manufacturer and supplier websites given on this page for lots more options and details.

Home Brew! It has been said that no radio amateur should ever buy an aerial - especially a wire aerial! Joe Tyburczy WB1GFH also comments about suitable antenna installing weather on his web pages:

"When you put up your antenna is also crucial. I must mention here the importance of what many early hams called "antenna weather". That is, snow, sleet, freezing rain, or combination of all the above. It has been proven time and time again that any antenna installed in conditions better than abysmal will not function worth a darn. Or, put another way, it takes bad weather to put up a decent antenna. Dark and cold New England winter days are ideal for this activity. Any antenna erected on such a day will inevitably produce miracles." Joe Tyburczy WB1GFH

Most of what is available commercially could be 'home-brewed' if one has the time and a few suitable mechanical skills.

It's worth bearing in mind that Joe Tyburczy's $4.00 Special may well be more effective than a commercial antenna costing $400.00. Perhaps the equally inexpensive Zepp Antenna could produce far greater Value For Money than shelling out £$hundreds on a commercial antenna?

I give no particular recommendation here, but a good rule of thumb is not to believe the marketing hype of any commercial company supplying antennas.

A good place to start is, I think, the classic All Band Doublet Antenna, which is mentioned several times below. Plenty of links are also provide further down this page.

As important as not believing any of the marketing hype is remembering the words of the Star Ship Enterprise's Chief Engineer Mongomery Scott; (paraphrasing) *You cannae change the laws of physics!* - Size matters. A coil loaded 5 foot long aerial probably won't be that much good on the H.F. bands. Indeed, a a 5 meter long antenna may be perfect for the 10 metre band but it won't be a great antenna on 20m, and certainly even less effective on 40m and 80m

So, before falling for all the marketing hyperbole its wise to read up on a little antenna theory and do some lengthy research into the antennas being considered before making an expensive decision.

Below are a few clues and ideas listed in no particular order - some of the antennas will be great while some of the commercial designs shown below may promise great things but fail to meet high expectations.

There is no "miracle antenna". If one understands the theory and the compromises made with shortened, under-sized, multi-band (and often very expensive) antennas there may not be so much disappointment. However most will be compromised in one way or another, so if one is happy to live with those compromises, understand what they mean, and are happy to have a lighter bank balance then that's fine!

Low SWR across all the bands does not indicate a good antenna. My dummy load has very low SWR across all the bands but it is most certainly not a good antenna!

Someone telling you that the antenna that they're using is great and that they've worked the world doesn't mean it really is a good antenna - it's just 'hearsay'. One could quite possibly work around the globe on a coat hanger in the right conditions - it doesn't suddenly make a coat hanger the great antenna that should be widely heralded!

The best way to learn about antennas is to go out and make some - well actually go out and make many antennas.

Here are a few ideas. Some that are easily 'home brewed' and some commercial ideas that could be adapted to 'home brewing'. I don't have any particular recommendation, but I hope that this list will provide some good food for thought. (Some of the commercial antennas may be okay, some may be awful).

Remember; ***Marconi spins in his grave every time a radio amateur buys an aerial instead of building it !***

The Doublet Antenna - the classic all band / multi-band Aerial

The Doublet Antenna is my favourite Multi-Band aerial. The main benefits of a Doublet Antenna are that it can be used on whatever frequency it is cut for, and higher frequencies. Usefully, there can be some useful gain on higher frequencies, although there will be some petal shaped lobes on the higher frequency bands.

In fact my 20 meter long Doublet for the 40meter band to the 6 metre band also works quite well on 80 meters! Not only a very good starter antenna for those looking to work as many bands as possible, but also an aerial that would be useful and efficient for years and years to come.

Ladder line or open wire feeder must be used (NOT coaxial cable). Ladder Line or open wire ensures low loss at High Frequencies so that as much of your precious transmitter power as possible will be radiated. Likewise receive efficiency should be maximized. Assuming that the aerial is successfully balanced, the feeder should not radiate, even when there is (inevitably) high SWR

High SWR is not an especially bad thing and will not reduce the aerial's performance, but note that high SWR on the balanced feeder does increase feeder loss compared to a when matched - but the losses will be a lot less significant then if coaxial cable were to be used. This ensures that efficiency should be better so that all the power that reaches the antenna wires will be radiated - save for any losses in the antenna tuning unit (AMU) or the balanced feeder cable.

Remember: Ladder Line = Lower Loss - Read more about my M0MTJ Doublet Antenna HERE >> and more notes on this page here >>

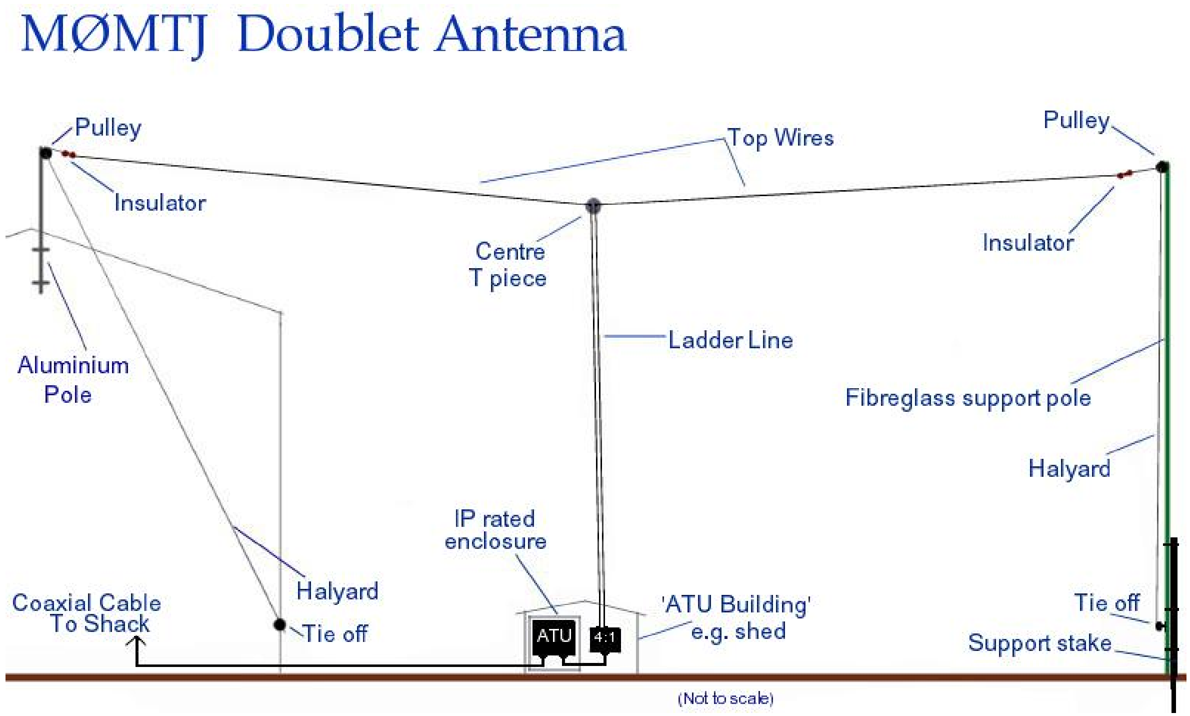

The Classic All Band DOUBLET ANTENNA - Read More about my M0MTJ All Band Doublet Aerial Here >>>

The G5RV and more about Doublets & Dipoles

The G5RV - (Even Louis Varney himself said that a Doublet would be better!)

Why the G5RV continues to remain popular is a mystery. The 'legendary' name, perhaps? That it can be 'thrown up' without too much thought? The fact that it' cheap and cheerful? Well, cheap anyway. There has been much written about the G5RV and the overall conclusion is that it's not a very good multi-band antenna. Certainly it will work quite well on, perhaps, a couple of bands, but there are better alternatives, especially when one considers the wide availability of remote automatic antenna matching units (remote auto 'a.t.u.'s) that will greatly assist in minimizing antenna feeder losses.

Perhaps the best thing that can be done with a G5RV is cut off the coax, re-size the top wires and turn it into a true Doublet Antenna - or, better than that, don't buy a G5RV in the first place and simply make (home brew) a true All Band HF Balanced Doublet Antenna.

The G5RV and its derivatives such as the Western HF-10 and the ZS6BKW are a bodge. In fact, Louis Varney G5VA himseld stated that a much better alternative to any of these 'G5RV' designs is to run the open-wire line from the center of a dipole all the way to a balanced antenna tuner!! This aerial arrangement will work on ANY band from 80 metres through to 10 meters.

The G5RV is, therefore, not the most efficient aerial and, as users report, may not always be the easiest to use, but because it is coax fed, it remains an ever popular choice because it's easy to feed the coax cable from the antenna's location back to the shack. Unfortunately this easy (lazy?) approach means lots of loss due to high SWR on the coaxial cable.

One may say, Oh well, there's perhaps only 2 or 3 S points in it. (Cough!). A difference of 3 S points is the difference between radiating 100 watts or only radiating 1.6 watts. Um, - I'm sure we'd soon complain if our shiny, new 100 watt transceiver only produced 1.6 watts!!

QRP and lower power portable operators should be especially keen to minimize antenna system losses so that as much power as possible is radiated. Losing 3 S points from a 5 watt FT-817 results in only 0.08 watts. QRP operators may likely favour resonant, single band antennas for greatest efficiency.

Coaxial Cable is only 'low loss' when the antenna impedance is similar to the cable's designed impedance of 50 Ohms, or near 1:1 SWR (or at least less than 2:1). Coax cable is therefore fine for a Dipole antenna that is designed to be resonant on one band of choice, which will have a reasonably low SWR across that band, but coax' is entirely unsuitable when one wants to use the aerial on several bands. The G5RV will be used an multiple bands and will therefore present widely varying impedance values (and SWR) depending on the band of operation and so the power losses will be significant.

Note: The antenna's true impedance (SWR) must be measured at the antenna's feed point (i.e outside and up in the air) not at the radio end of the cable, in the shack, where the coaxial cable's losses will mask the true SWR reading. Unfortunately 50 Ohms, or a near 1:1 match, will not be encountered on any band on a G5RV. The best band will be 20 metres at typically 2:1 (100 Ohms), but on any other band the impedance will be very much higher. One might expect rersults like these: 80 metres 3:1 to 5:1, 40 meters 5:1, 30 metres 20:1 or more, 17 metres 15:1 or more, 15 metres 6:1 to 12:1, 12 meters 3:1, 10 meters 10:1 to 20:1.

1000 Ohms into 50 Ohm coaxial cable will result in a massive dumping of your power - lost as heat in the coaxial cable instead of being radiated as a useful signal. However, high SWR is not in itself a bad thing - it's just that the correct type of feeder and impedance transformation method needs to be used. The *incorrect* feeder is coaxial cable - the correct type of feeder is balanced Line or Ladder Line feeder. This can be the typical slotted 300 Ohm type or the 450 Ohm 'window line' which are commercially available. Even better, make your own open wire, balanced line feeder. Spacers can be bought commercially or 'home brewed'. The spacing for this parallel line should be about 2 to 3 inches (50 to 80 mm) - the width of the spacing is not particularly critical, but whatever width is chosen should be maintained as accurately as possible along the length of the feeder - i.e. it should be parallel.

The Doublet can take more time and effort to tune, prune and adjust. It will also need a good antenna matching unit (so-called "ATU") and low loss balun, but the reward will be more effective performance due to the much reduced loss in the balanced line feeder system.

Doublet Aerials - a better alternative to a G5RV

Check out the The classic All Band Doublet and the NorCal Doublet for very simple, effective and versatile antennas for multi band operation:

The ALL BAND Doublet Antenna

This could very well be the first antenna that one should try in order to get on at least one band very efficiently but also several other bands with the use of a good ATU.

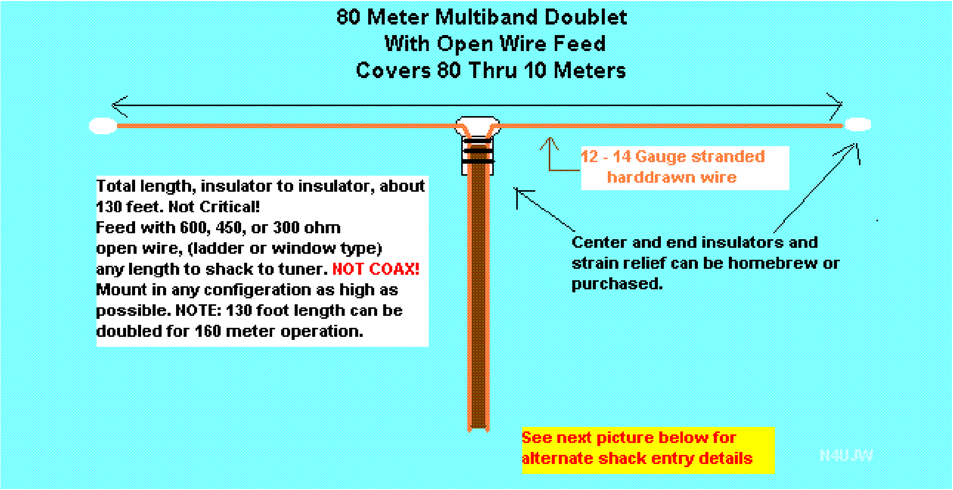
80 Meter Multiband Doublet
With Open Wire Feed
Covers 80 Thru 10 Meters
12 - 14 Gauge stranded harddrawn wire
Total length, insulator to insulator, about 130 feet. Not Critical!
Feed with 600, 450, or 300 ohm open wire, (ladder or window type) any length to shack to tuner. NOT COAX!
Mount in any configeration as high as possible. NOTE: 130 foot length can be doubled for 160 meter operation.
Center and end insulators and strain relief can be homebrew or purchased.
See next picture below for alternate shack entry details
N4UJW

All Band Doublet - http://www.hamuniverse.com
See My All Band Doublet Antenna here

The all band Doublet antenna is nothing more than a 1/2 wave dipole cut for your lowest operating frequency and fed with twin lead, ladder line, open wire, etc to a tuner that will accept a balanced line connection. IT MUST NOT BE FED WITH COAXIAL CABLE !

It can be designed for use from 160 through to 10 meters very easily using the standard 1/2 wave dipole formula:

468/freq MHz = total length (ft)
The exact length is not critical!

If there is insufficient room for a lower frequency version (160m or 80m), then the Doublet can be designed to the shorter wavelength of the 40 metre band and used up to the 10 metre band. (Do not attempt to operate on a lower frequency than 7 MHz in that case since this could damage the a.t.u.) It may be possible to connect the ends together and tune it against earth - if you have a good enough earth - and use lower frequency bands. For best results a doublet should be mounted as high as possible (as with many aerials) and can be erected as a flat top or inverted V.

A Doublet Antenna needs a good antenna matching unit with a wide impedance matching range (obviously not the one in the radio!). Preferably this should be a balanced antenna matching unit, but an unbalanced matching unit can also be used together with a good low loss current / Guanella balun, such as the extremely high quality items available from G-Whip Antennas. I use a 4:1 G-Whip guanella current balun with my Doublet, but many users recommend using a 1:1 ratio current balun - again, this is a case for individual experimentation.

A more detailed description of the Doublet Antenna can be found here >>

External Links

The All Band HF Doublet can also be found on Ham Universe: http://www.hamuniverse.com/hfdoublet.html

Introducing The All Band Doublet by the late L.B. Cebik W4RNL: http://www.cebik.com/content/edu/edu6.html
N.B. Create a free account at http://www.cebik.com

The Norcal Doublet

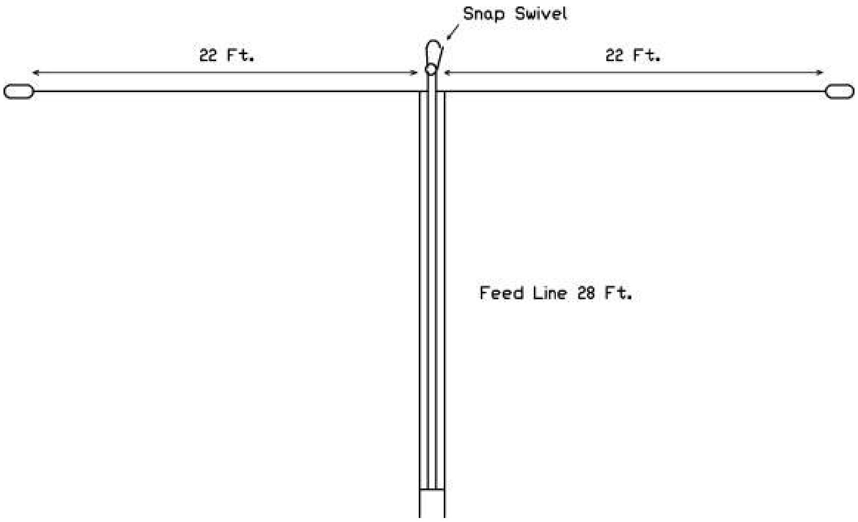

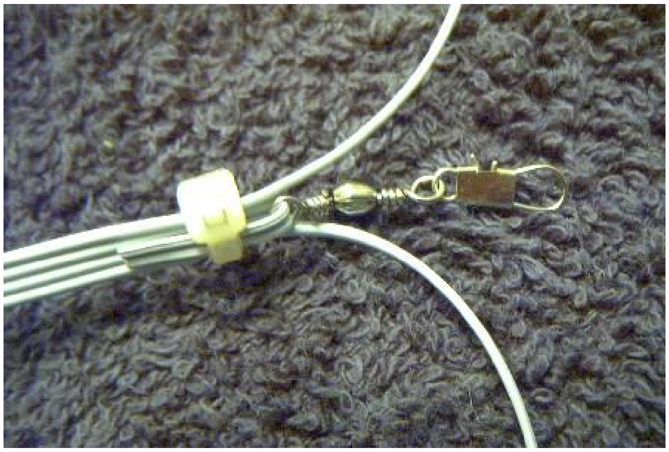

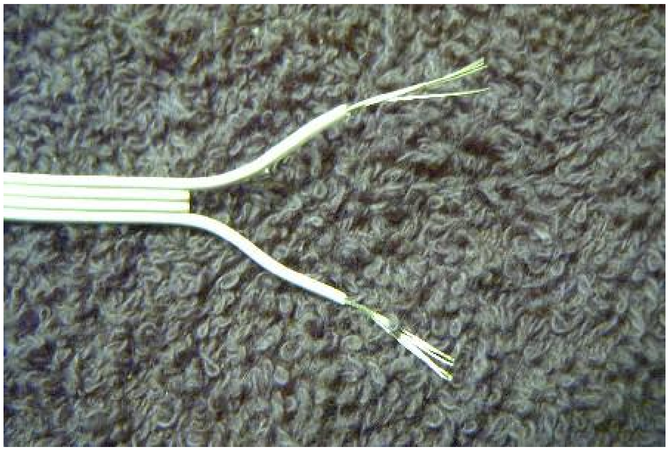

The Norcal Doublet Antenna: http://www.norcalqrp.org/norcaldoublet.htm

The Norcal Doublet is a simple antenna that is 44 feet (13.4 metres) long, 22 feet (6.7 metres) per side. The Norcal pages report "...that the antenna would have basically the same radiation patterns on all bands from 40 - 10 meters. This would be very handy to have for field operation..... You will need the following materials: 50 feet of 4 core stranded computer cable; 1 #0 Fishing Swivel; 1 Cable tie; 2 pieces of fishing cord."

The antenna can be hung from trees or cheap telescopic 'roach' / Sota poles. Doubling the size would allow operation on 80 metres and even 160 metres by shorting the twin feed together at the transmitter end and feeding it against a good earth as a 'Marconi' type antenna.

An effective multi-band "4 Dollar Special" by Joe Tyburcy - W1GFH (An Inverted Vee Doublet Antenna - super cheap - super effective - multi-band)

Joe Tyburczy, W1GFH provides some sensible insight and advice, he writes: *"I am a big fan of "balanced line" (twin lead, open wire line, etc.) vs. coax. By using balanced line and a tuner you can have one, single-element antenna that works well on all bands. You can't do that as easily with coax. The basic "W1GFH $4 SPECIAL" shown below is a variation on the type of versatile skyhook I've been using for years......Now at this point, some of you may be looking at the diagram and muttering, "Jeez Joe, that's just a dipole fed with twin lead and used with a tuner". Well of course it is. Virtually all antennas are "di-poles" (i.e. "two sides") in some form or another. This one just happens to be made from low-cost materials......I won't go into the theory here, but trust me: balanced feed line, properly used, does not "leak" RF and is less lossy than coax. I've tried the commercial 450-ohm ladder line, but prefer 300-ohm TV twin lead, and the cheaper the better. Radio Shack TV twin lead is ideal. Home Depot has some good stuff, too. Forget all the obsessive junk about standing waves, impedance and velocity factor. What you really need to concentrate on is getting an interesting set of antenna insulators."*

Read Joe's excellent article in its entirety here: http://www.qsl.net/wb1gfh/antenna.html

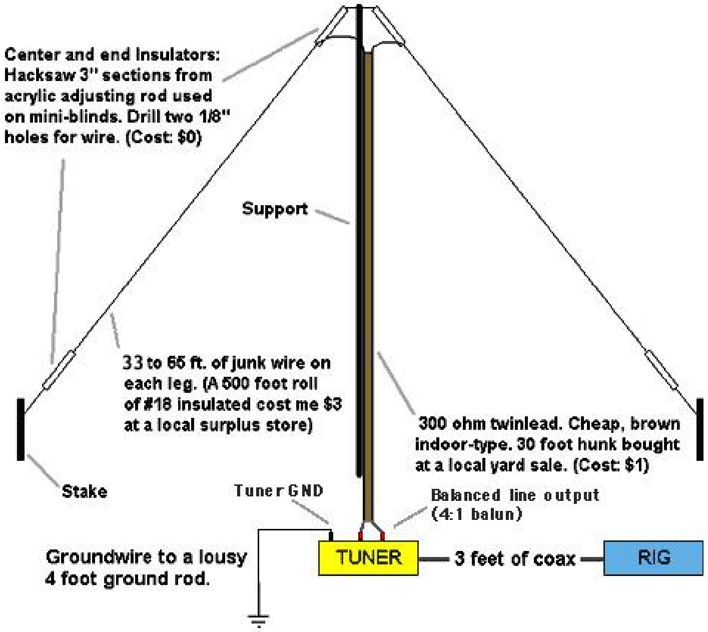
Center and end Insulators: Hacksaw 3" sections from acrylic adjusting rod used on mini-blinds. Drill two 1/8" holes for wire. (Cost: $0)
Support
33 to 65 ft. of junk wire on each leg. (A 500 foot roll of #18 insulated cost me $3 at a local surplus store)
300 ohm twinlead. Cheap, brown indoor-type. 30 foot hunk bought at a local yard sale. (Cost: $1)
Stake
Tuner GND
Balanced line output (4:1 balun)
Groundwire to a lousy 4 foot ground rod.
TUNER
3 feet of coax
RIG

4 Dollar Special by W1GFH
http://www.qsl.net/wb1gfh/antenna.html

Inverted L - 80 metres to 10 metres

A typical Inverted L antenna will be trapped for 40m/80m using a 7.1 MHz trap. It is essentially one half of a W3DZZ dipole so can be accommodated very much more easily into a small plot or garden - especially as part of the antenna is running vertically up a wooden or fibreglass (non conductive) pole. This should allow it to be fitted into quite a small garden such as mine. It's a very useful antenna in this respect, and because there is a good length of wire in the air, it radiates quite well.

The Inverted L is also a very effective aerial because it has the benefit of both vertical and horizontal radiation. While Inverted L's might make good TX aerials, like ground mounted vertical aerials they do have the potential to be a little noisy on RX. This will depend upon local circumstances and noise sources. However the fact that the feed point will generally at the bottom of the garden, well away from the house may help to keep QRM to a minimum.

The Inverted L is extremely easy to 'home brew'. Spectrum Communications can also supply the complete aerial as shown below. It should give excellent performance on 80m and 40 metres, with 20 metres also being good but also allowing use on 15m and 10m and possibly one or two of the WARC bands.

If you fancy home-brewing the complete antenna, except for the 7 MHz trap, traps are available to buy commercially from suppliers such as Spectrum Communications and Sotabeams who supply very neat and lightweight devices,

Use a Remote Automatic Antenna Coupler : The use of a good Automatic Antenna Coupler, such as the CG Antennas CG-3000, at the feed-point with a simple inverted L wire will provide a very good Multi-Band aerial that could be used on all bands between 160m and 10m and should be very simple to deploy.

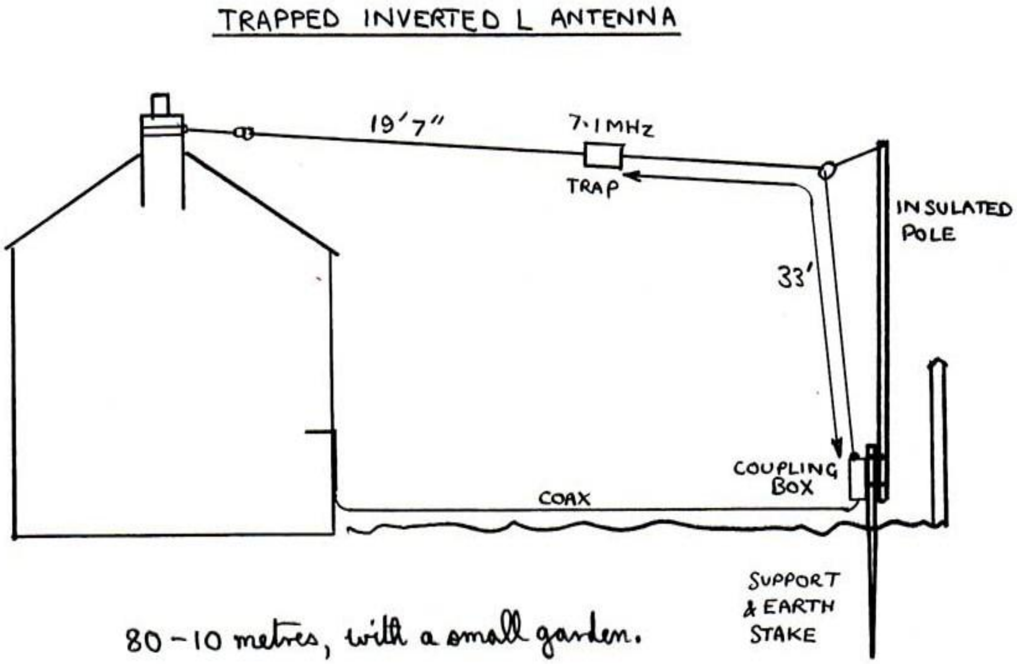
TRAPPED INVERTED L ANTENNA
19'7"
7·1 MHz
TRAP
INSULATED POLE
33'
COUPLING BOX
COAX
SUPPORT & EARTH STAKE
80 – 10 metres, with a small garden.

Spectrum Communications Inverted L - benefits from both vertical and horizontal radiation
http://www.spectrumcomms.co.uk

More about Inverted L Aerials:

The Inverted L for 40m/80m is shown below is essentially one half of a W3DZZ dipole fed against ground using one 7.1 MHz trap. It's a very compact antenna and is simple to construct. It is most efficient, of course, on 80 metres and 40 metres, but can also be used, with an a.t.u., on 20m, 15m and 10m.

Find out how to make one here: http://www.users.icscotland.net/~len.paget/5%20band%20Inverted%20L.pdf

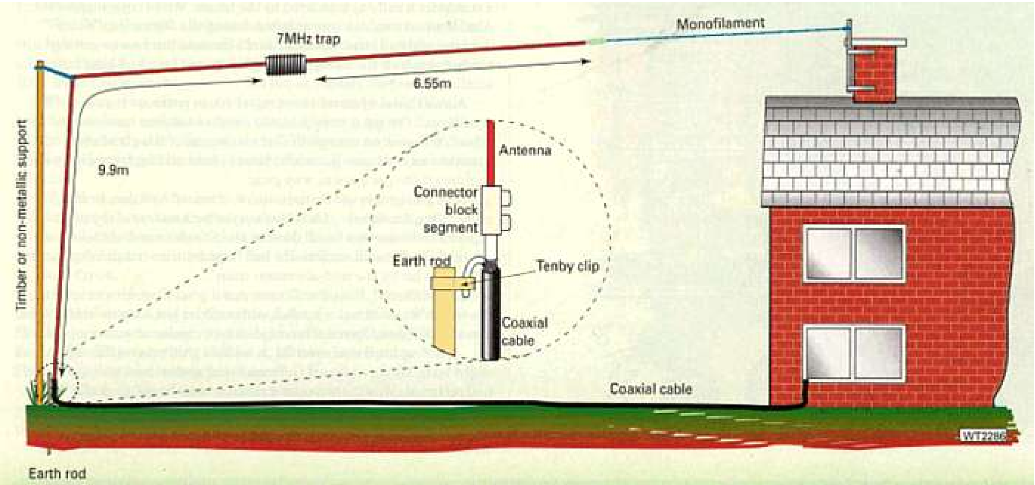
Monofilament
7MHz trap
6.55m
9.9m
Timber or non-metallic support
Antenna
Connector block segment
Earth rod
Tenby clip
Coaxial cable
Coaxial cable
WT2286
Earth rod

The basic layout of the Inverted L Antenna by Len Paget GM0ONX (Practical Wireless)

Adding 160m / Top Band to an Inverted L

The 160 metre Top Band can be added to this aerial by connecting a 3.5 MHz trap at the end of the 80 metre wire (where to monofilament joins the 6.55m section of wire below) with another length of wire on the other side, increasing the overall length of the antenna.

Find out how to do it here: http://www.users.icscotland.net/~len.paget/Inverted%20L%20adding%20top%20band.pdf

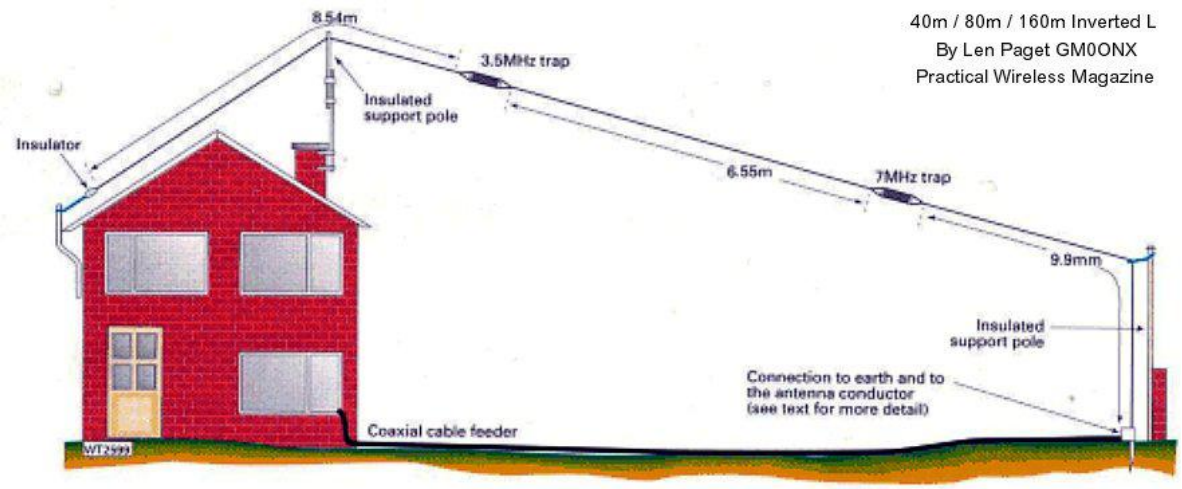
8.54m
3.5MHz trap
Insulated support pole
Insulator
6.55m
7MHz trap
9.9mm
Insulated support pole
Connection to earth and to the antenna conductor (see text for more detail)
Coaxial cable feeder
WT2599
40m / 80m / 160m Inverted L
By Len Paget GM0ONX
Practical Wireless Magazine

Adding Top Band to an Inverted L by Len Paget GM0ONX (Practical Wireless magazine)

160m Top Band 'Inverted L' Wire Antenna

At the time of writing I am using a 40m/80m Inverted L aerial and in an attempt to get on Top Band I have also been experimenting with a shortened 160m antenna in an inverted L configuration that uses a wire about 19 metres long - about half the size of a full size inverted L for 160 Metres.

A Top Band aerial of this type also needs a very good ground / counterpoise and can suffer the disadvantage, like ground mounted vertical aerials, of being rather noisy on RX . The drawing below shows the general idea. Read more here >

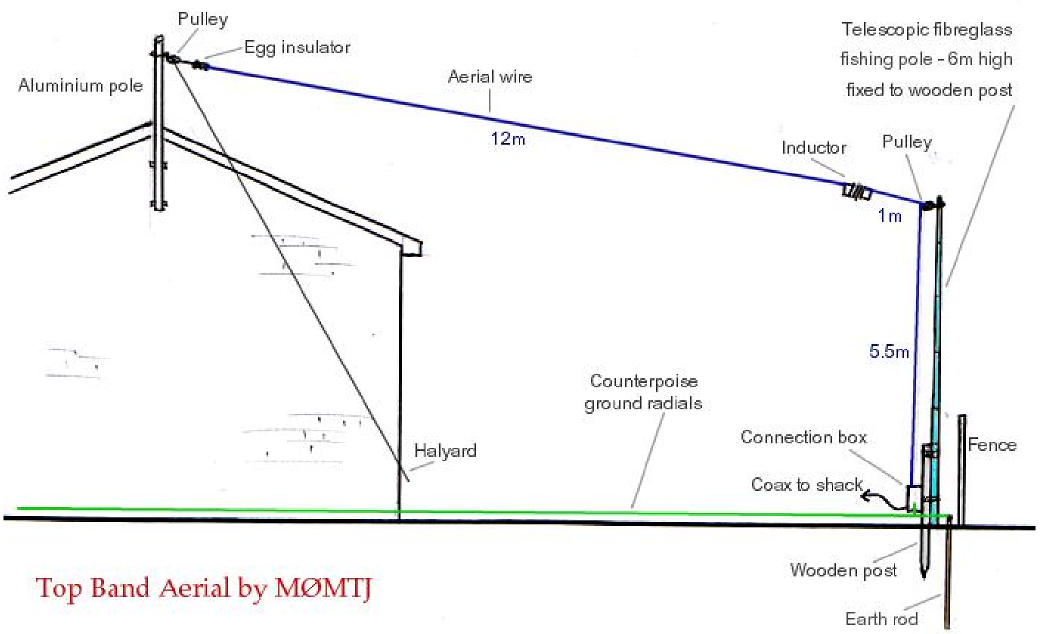
Pulley
Egg insulator
Aluminium pole
Aerial wire
12m
Telescopic fibreglass fishing pole - 6m high fixed to wooden post
Inductor
Pulley
1m
5.5m
Counterpoise ground radials
Connection box
Fence
Halyard
Coax to shack
Wooden post
Earth rod
Top Band Aerial by M0MTJ

General layout of Top Band Aerial with fibreglass pole retracted to a height of 2 metres
Wire lengths are approximate: Inductor 5cm dia with approx 40 turns of enamelled copper wire

Full Wave Loops and Delta Loops - An easy to install and effective antenna for multi band operation

Easy and low visual impact - a full wave loop: With a larger garden a full wave loop could be easily accommodated horizontally without the neighbours even noticing. A garden with a perimeter of 40 metres would easily accommodate a full wave loop for the 40 metre band and would work up to 10 metres or even 6 metres with an 'ATU'. A garden with a perimeter of 80 meters would accommodate a full wave loop for the 80 metre band and would work up to 10 metres.

Small Loop for 20 metres to 10 metres: A loop for 20 metres or 17 meters is relatively compact and could easily be installed in small 'postage stamp' sized gardens. A loop antenna could be triangular, square (Quad) or circular, but a square loop (and indeed a circular loop) would need more supporting points than a delta (triangular) loop, so a Delta loop is likely to be the easier option.

The loop is really a single band antenna cut for one wavelength on the band of interest, however it can also work quite well as a cheap and easy to install multi-band H.F. aerial. A loop consisting of a 17 metre length of thin antenna wire, for example, will work well on 17 metres but may also give 15m, 12m and 10m with an ATU. My own loop is made from an 16 metre length of wire, tuned for the 17m band, but can work on higher bands. A 40 metre loop will be considerably larger, but it might still possible to accommodate in many fairly compact gardens. Performance will depend on height and orientation.

Feeding the loop at the top or bottom will give horizontal polarisation, while placing the feed point on the side will give vertical polarisation. The apex can be at the top or the bottom, but performance should be better with the apex at the bottom with the flat wire across the top - however for ease it may be more convenient to support a Delta Loop on a single pole, meaning that the apex would be at the top.

Ideally a loop should be fed with balanced line back to the shack, connected to a balanced line ATU or other ATU via a 4:1 balun. Alternatively use a 4:1 balun at the antenna end and run 50 ohm coax back to the ATU / txvr - though losses will be greater doing it by this method if the coaxial cable is quite long.

If one can install a separate antenna for the lower frequency bands of say 160m, 80m and 40m, then a Loop Antenna could be a good partner to allow operation on the higher bands of 20 metres to 10 meters or even 6 metres.

A loop should be really very easy to install using a single support pole and very cheap too! All that's needed is the supporting pole, some cheap wire, a 4:1 balun which can be 'home brewed' and some thin cord and insulators which should not be an eyesore either.

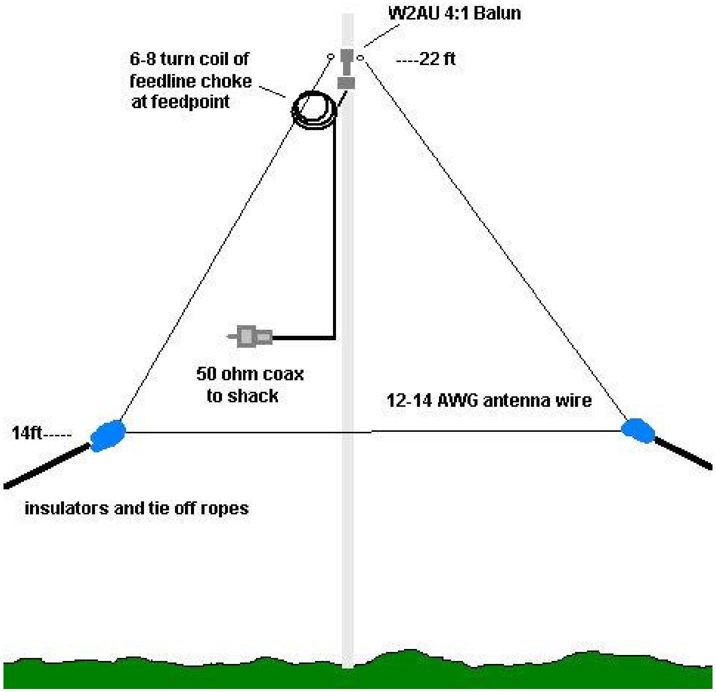

Diagram from the excellent article by W5SDC
http://w5sdc.net/delta_loop_for_hf.htm

Band	Length of antenna wire at mid band in metres
10	10.617
12	12.282
15	14.432
17	16.90

20	21.502
30	29.622
40	42.842
80	82.234
160	161.223

Delta or Quad Loop Antenna - An ideal multi band antenna solution?

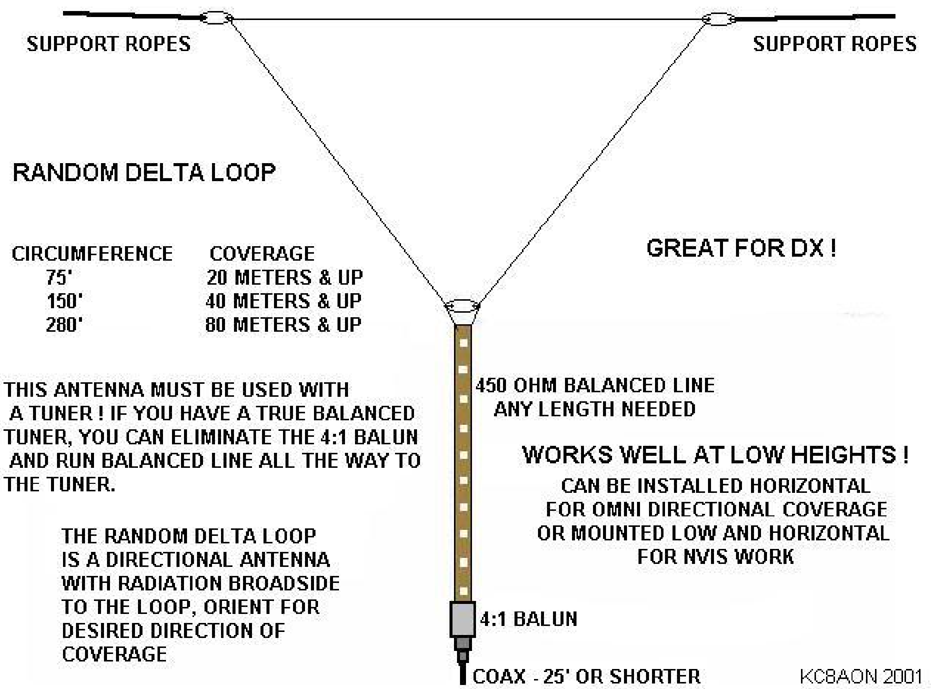
SUPPORT ROPES
SUPPORT ROPES
RANDOM DELTA LOOP
CIRCUMFERENCE COVERAGE
75' 20 METERS & UP
150' 40 METERS & UP
280' 80 METERS & UP
GREAT FOR DX !
THIS ANTENNA MUST BE USED WITH A TUNER ! IF YOU HAVE A TRUE BALANCED TUNER, YOU CAN ELIMINATE THE 4:1 BALUN AND RUN BALANCED LINE ALL THE WAY TO THE TUNER.
450 OHM BALANCED LINE ANY LENGTH NEEDED
WORKS WELL AT LOW HEIGHTS !
CAN BE INSTALLED HORIZONTAL FOR OMNI DIRECTIONAL COVERAGE OR MOUNTED LOW AND HORIZONTAL FOR NVIS WORK
THE RANDOM DELTA LOOP IS A DIRECTIONAL ANTENNA WITH RADIATION BROADSIDE TO THE LOOP, ORIENT FOR DESIRED DIRECTION OF COVERAGE
4:1 BALUN
COAX - 25' OR SHORTER
KC8AON 2001

A Multi Band Wire DX Loop Antenna by KC8AON
http://www.i1wqrlinkradio.com/antype/ch10/chiave1827.htm

Delta Loop by Arthur M0PLK (SQ2PLK)
Details at http://pdxa.one.pl/articles.php?article_id=17
Available at http://ham-radio.urbasket.eu and http://www.vpa-systems.pl/

Using fibreglass fishing poles (Sota Poles) two 7 metre long poles can be erected in an inverted V shape and used to support a 20 metre delta loop which will be usable on 20m to 10m and also adaptable for use on the 40 metre band.

The two aerial wires used are connected directly to a 4:1 balun which is, in turn, connected to an ATU such as the Z-11 Pro or Z-100 via coaxial cable. See this page which shows the W6ZO delta loop to get for the general idea of what will be achieved. The finished aerial will be very much like the commercially available ProAntennas DMV-Pro.

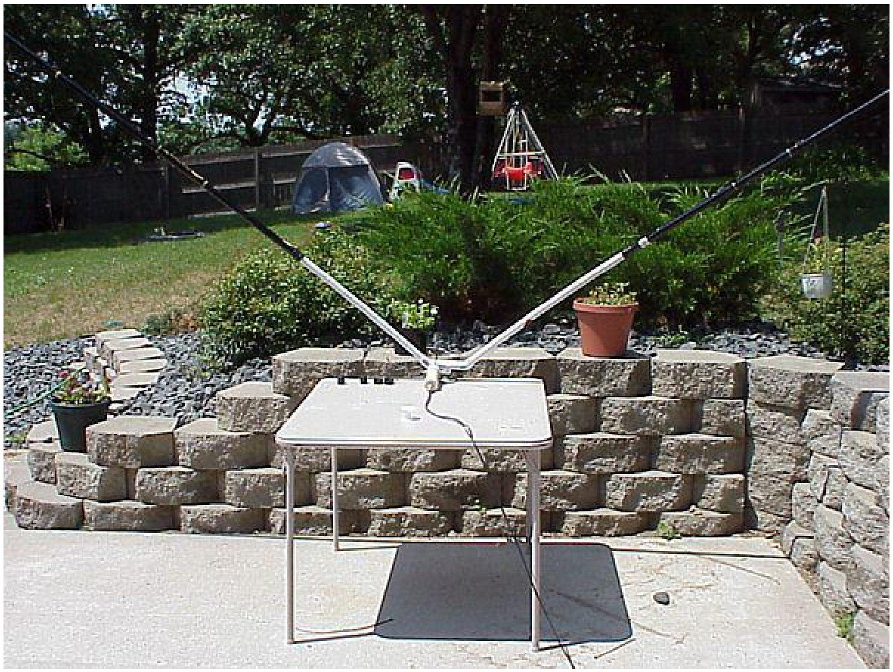

W6ZO Delta Loop - fed with 4:1 Balun - 40m to 10m
http://www.fros.com/KI0GU/w6zodelta.htm

The Loop Antenna. Ideal - cheap, easy, multi-band, simples! -
Lots of links to other Loop Antenna information on the links page here

Omni Directional Multi-Band Horizontally Polarized Delta Loop

If all that can be erected is a single pole, masts or telescopic pole, here is an option that will provide a horizontally polarized signal on 20 / 17 / 15 / 12 / 10 metres - the Sandpiper Aerial Technology GM3 (designed by GM3VLB). A similar idea, providing omni-directional horizontally polarized signals, is the G3TPW Cobwebb Antenna from Steve Webb.

These look like good ideas, though I have yet to make one to test.

G-WHIP
Antenna Products
Top Quality
British Manufacture

Site Search
Find
powered by FreeFind

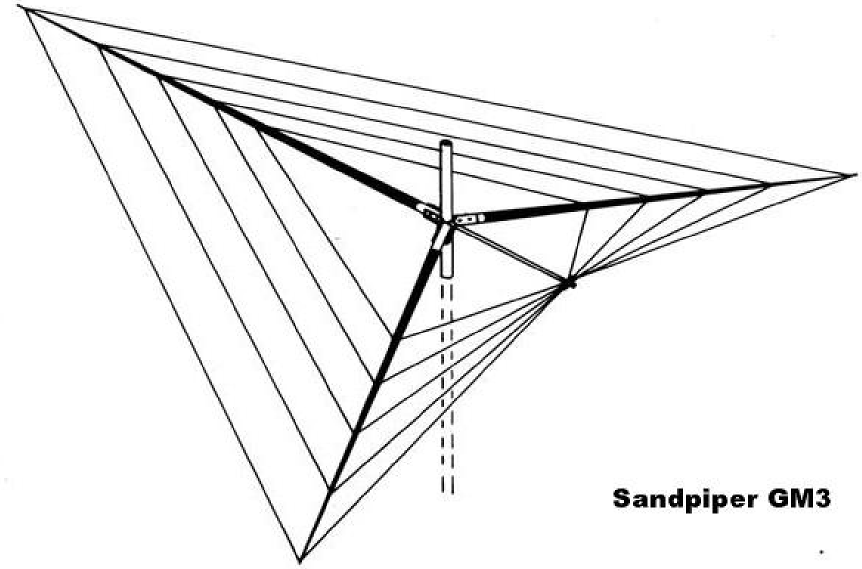

Sandpiper GM3VLB Antenna
http://www.sandpiperaerials.co.uk/

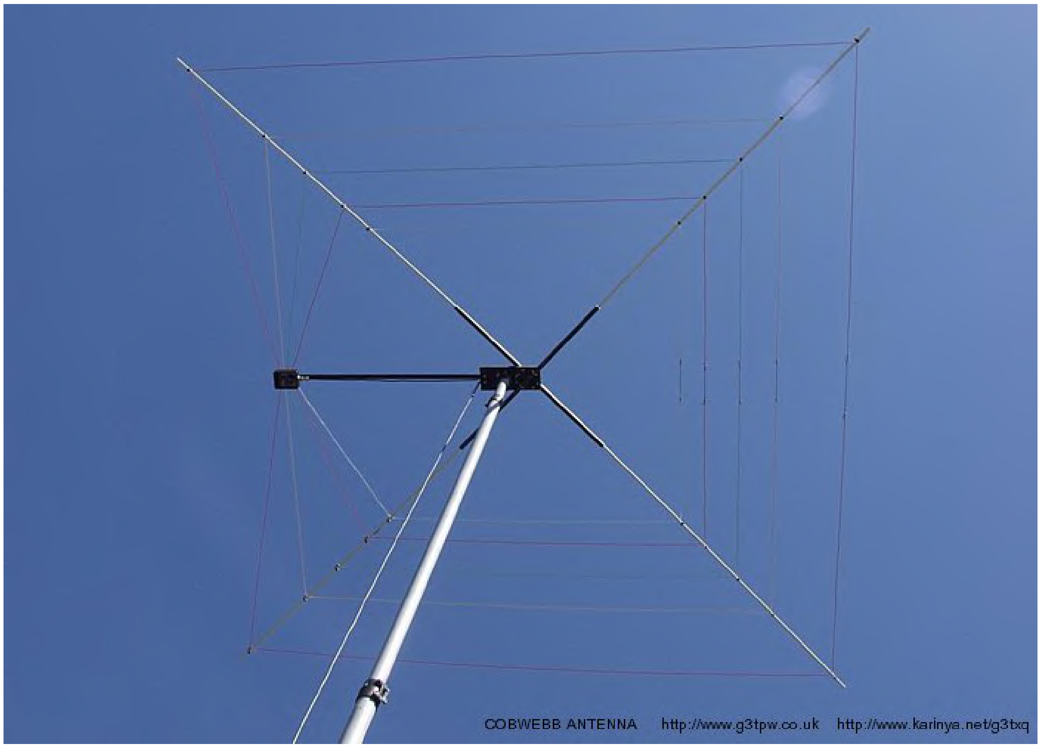

G3TPW Cobwebb Antenna from Steve Webb - photo by G3TXQ
http://www.g3tpw.co.uk http://www.karinya.net/g3txq

Resonant Vertical Antennas

Low SWR: Having a low v.s.w.r. across the whole of HF may seem, at first glance, to be a good thing - but my dummy load has a very low v.s.w.r. from HF to UHF, it does not make it a good antenna! As far as aerial systems are concerned, having a low v.s.w.r. across the whole of HF is probably not the best way to judge an antenna - a wide band low v.s.w.r. could indicate a fault with the aerial or feeder system - or just that lossy matching transformer that is gently heated up by the power applied from the transmitter!

A low v.s.w.r. is a good thing in a resonant antenna. It will help demonstrate the antenna's point of resonance - but the v.s.w.r. will rise either side of resonance. So:

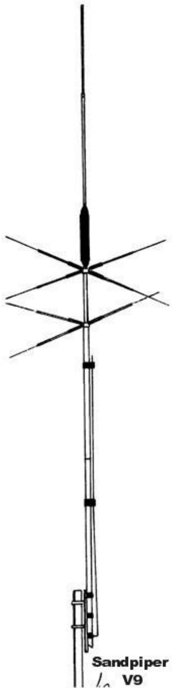

The next commercial option is an antenna that is truly resonant on a number, but perhaps not all, bands. The Hustler and Butternut varieties are very well known and offer well reported good performance.

Some vertical antennas use traps to achieve multi-band performance and as such are essentially one half of a trap dipole, fed against ground. A consideration is that the traps themselves, particularly if multiple traps are used, will introduce losses to the antenna system. It should also be noted as an additional consequence of using traps on a ground mounted vertical antenna, is that the highest frequency element will inevitably be positioned at the lowest position on the aerial - not a good position to be, especially for local ground wave radiation where signals will more easily be screened by local structures. Ground conductivity also needs to be good for verticals to operate efficiently.

The Hustler 4-BTV, 5-BTV & 6-BTV aerials are examples of multi-band verticals that use traps; many amateurs report great success with Hustler aerials - and it is very low profile too, indeed an amateur friend of mine uses a Hustler with great success and has even modified an additional top resonator so that the aerial can be used on 160 metres..

Other manufacturers of resonant vertical antennas, using varying design methods, include GAP, Cushcraft, Diamond, HyGain, and the well known British names Moonraker and Sandpiper.

Advisory: These are perhaps a good idea for limited space situations, but the shorter versions will not work well on the lower HF bands. A 6 metre tall antenna cannot possibly work terribly well on 40m or 80m, but may work well on the 20 metre band and above. However one could easily 'home brew' an equally effective antenna for a fraction of the cost of a commercial antenna.

Sandpiper Aerials

Sandpiper Aerial Technology offer an enormous range of amateur radio antennas from HF to UHF. For HF working Sandpiper offer numerous options including simple multi band resonant antennas supported on fibreglass telescopic poles using either changeable or plug-in coils for different bands to the famous V range and shorter and more compact MV range and the Portable MV antenna on its own tripod base. The V and MV ranges use a rather innovative design, seen above right. The V and MV antennas are available in versions that cover all the HF bands - or as cheaper versions that just cover the particular bands of interest. http://www.sandpiperaerials.co.uk/

Vertical antennas will be quite short when compared to the wavelengths of some of the longer wavelength bands (particularly 40m, 80m and 160m) and so will not radiate as efficiently as a full size resonant aerial. The bandwidth will also be quite narrow. Setting up a multi band antenna to be resonant at the required portion of each band can sometimes be a little time consuming, but for the convenience it might be a price worth paying. A vertical antenna will generally have a low angle of radiation which is a good thing for long distance DX but verticals can be very noisy on RX compared to a balanced dipole and as previously alluded to, radiation efficiency will be very low when compared to a full size single band resonant antenna. Attractive options?

Alternatives: My favourite way of experimenting with aerials is using a 10 metre long fibreglass telescopic fishing pole as the support. These fibreglass poles are lightweight and easy to carry, put up and take down, ideal for supporting lightweight v.h.f. and u.h.f. yagis, wire dipoles and doublets and also for supporting vertical wire aerials.

The telescopic pole must be made from fibreglass, not carbon fibre which is electrically conductive.

One great design is by Dave G4AON who writes on his web page: "There seems to be a myth among many newly licensed radio amateurs that an antenna works better if it costs a lot of money..... The antenna shown here costs around one tenth the price of a commercial vertical, yet it will perform as well as (and in many cases better than) a trapped vertical antenna. This antenna is based on a 10 Metre long fibreglass fishing pole.....the poles will collapse inside the sections unless each joint is secured with PVC tape, for more permanent installations glue could be used.

...The wire lengths are calculated from the formula L = 234/F, where F is the frequency in MHz and L is the wire length in feet. These lengths work out to around 33' 3", 23' 2" and 16' 7" for the 7, 10 and 14 MHz bands. The lengths for 7 and 10 MHz were more or less correct, however probably due to interaction between the wires the 14 MHz wire needed lengthening by around 4" for minimum SWR. Wire size is not critical, but it is probably better to avoid the thinnest "hookup" wire. Note, ground conductivity/loss and elevated/buried radials make a significant impact on both the performance and tuning of a ground mounted vertical. In the case of buried radials the vertical may resonate significantly lower in frequency than expected."

The antenna shown on the right is made for triple band operation on 7 MHz, 10 MHz and 14 MHz by the use of parallel wires, but an aerial based on a fibreglass pole could be single, dual, triple or even - at a push - quadruple band.

The more bands included the more difficult it will be able to trim to tune for resonance (as with a fan dipole) so to keep interaction to a minimum the wires should be quite well spaced. Like all quarter wave verticals aerials a good ground plane will be needed.

See G4AON's excellent full article here: http://www.qsl.net/g4aon/vertical/

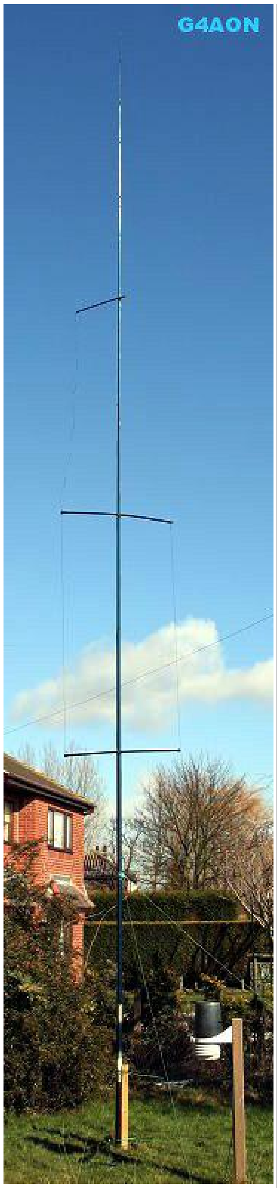

Horizontal Resonant Wire Aerials

The great advantage of a vertical antenna is that they have a very small footprint, i.e. they can be installed in the corner of many small plots and gardens. True resonant verticals can properly cover many, if not all, the HF bands. However to be reasonably effective a vertical needs a very good ground and also must be quite tall, in the order of 6 to 9 meters in many cases (about 19 to 30 feet). This may cause objections from the XYL. Another disadvantage might be that a vertical has little near vertical incidence skywave radiation (NVIS), a consideration for the lower HF bands, so after the local ground wave coverage there will be little or no signal until after about 500 miles, not good for inter G working. (ref. G8JNJ) - So try some Dipole Antennas:

Dipole Antennas

The dipole antenna is possibly the simplest and cheapest antenna to make. It it cut for single band operation where it should make a very efficient radiator. The simple wire dipole should be quite discrete, though not entirely invisible, but should not raise too many objections from the XYL or neighbours.

For some dipole ideas see this page:
http://www.qsl.net/ta1dx/amator/practical_dipole_antenna.htm

How to make a basic dipole by Marshall N1FN : http://www.morsex.com/dipole/index.htm

Also see this detailed and useful page: http://www.dxzone.com/cgi-bin/dir/jump2.cgi?ID=7499

Calculator:
http://www.sean1226.pwp.blueyonder.co.uk/design_your_own_antennas%201.htm

TRAP Dipoles:

The next type of antenna to consider might, then, be a horizontal or sloping wire antenna. Perhaps the most familiar multi band wire aerial is the trap dipole. The traps, in simplest terms, divide a wire dipole into two or more resonant sections so that two or more bands can be covered.

As far as commercial options go then there are quite a number. Tony Nailer of Spectrum Communications produces a great deal of useful equipment and accessories including some well designed and well made trapped dipole aerials based on the very effective W3DZZ design. In particular the Full Size G4CFY resonant on 80m and 40m and also allowing operation on 20m, 15m and 10m, also the Half Size G4CFY resonant on 40m and 20m and additionally allowing operation on 15m and 10m.

Visit http://www.spectrumcomms.co.uk/G2DYM.htm for more information.

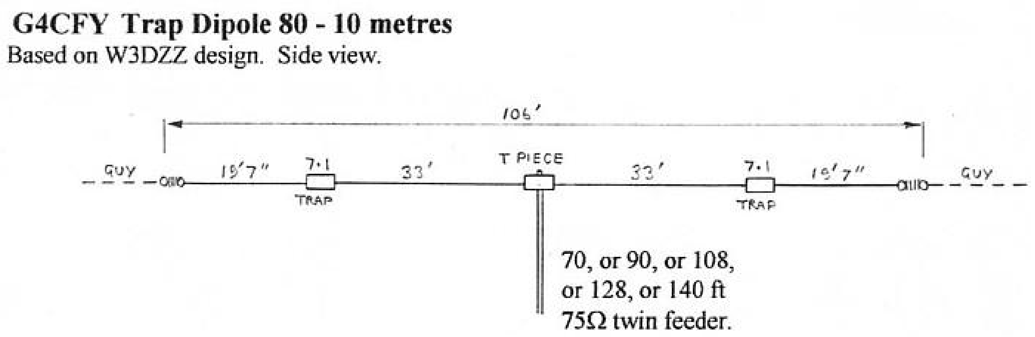
G4CFY Trap Dipole 80 - 10 metres
Based on W3DZZ design. Side view.
106'
GUY
19'7"
7·1
TRAP
33'
T PIECE
33'
7·1
TRAP
19'7"
GUY
70, or 90, or 108, or 128, or 140 ft
75Ω twin feeder.

Spectrum Communications G4CFY Trapped Dipole
http://www.spectrumcomms.co.uk

Spectrum Communications G4CFY Trapped Dipole
http://www.spectrumcomms.co.uk

More about Coaxial Trapped Dipoles

A trapped dipole for 40m and 80m offers the advantage of being somewhat shorter than a full size single band 80m resonant dipole plus it offers 40m as a resonant band plus the possibility of working on 20m, 15m and 10m. There are several designs available on the web for this type of aerial so Google W3DZZ. One of the most comprehensive sets of instructions is by Len Paget G0ONX. Fine out more here:
http://www.users.icscotland.net/~len.paget/GM0ONX%20trap%20dipole.pdf

This would be my choice if I had the space, though since a dipole is a balanced aerial it would make more sense to use balanced twin feeder (as in the Spectrum Communications implementation of this design) rather than coaxial cable which is an un-balanced and more lossy feeder.

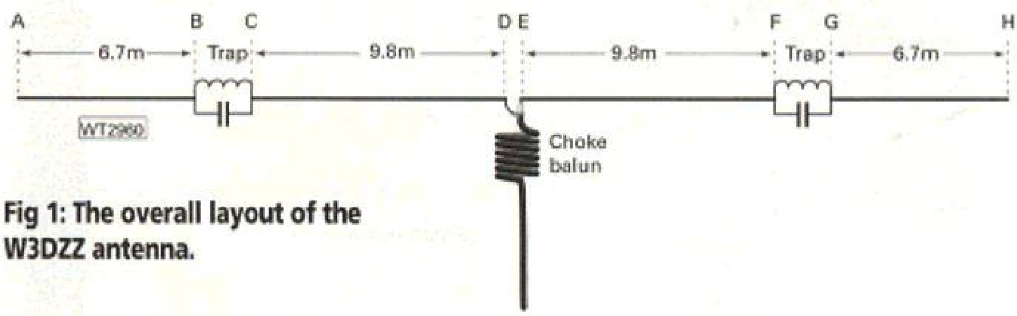

Fig 1: The overall layout of the W3DZZ antenna.

The W3DZZ Trapped Dipole - a balanced aerial, so use balanced twin feeder!

Here is a variation on the W3DZZ antenna by the Maidstone Amateur Radio Society that adds a dedicated 10 meter (28MHz) resonant element as a 'fan'.

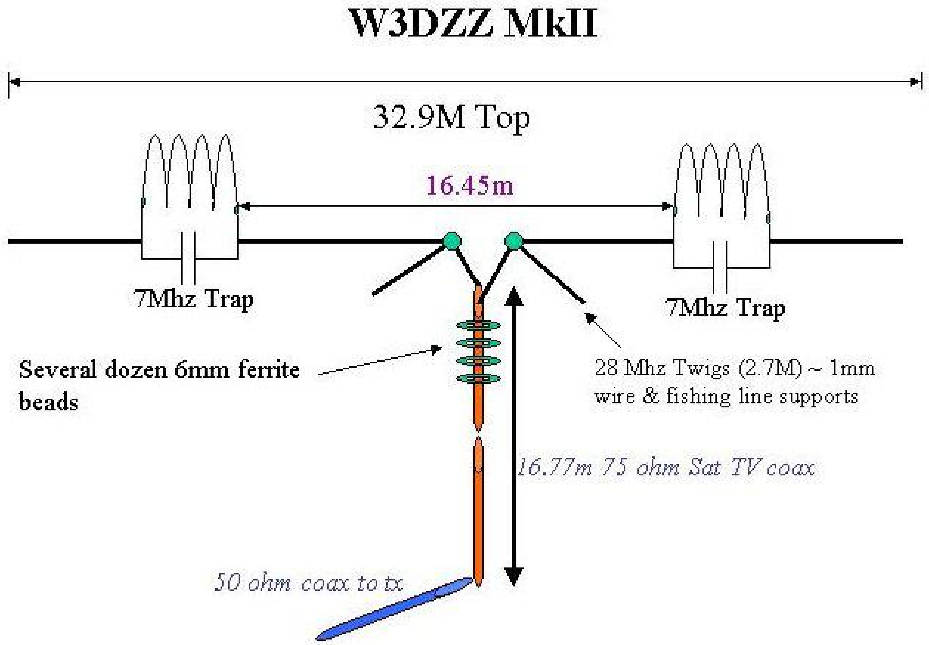
W3DZZ MkII
32.9M Top
16.45m
7Mhz Trap
7Mhz Trap
Several dozen 6mm ferrite beads
28 Mhz Twigs (2.7M) ~ 1mm wire & fishing line supports
16.77m 75 ohm Sat TV coax
50 ohm coax to tx

W3DZZ Dipole Aerial design by the Maidstone Amateur Radio Society
http://www.btinternet.com/~shaun.scannell/club/w3dzz.htm

Moonraker supply a whole range of wire trap dipoles covering from 2 to 5 HF bands (MTD1; MTD2; MTD3; MTD4; MTD5; MTD6). Diamond also produce trapped wire antennas, the W-721, W-728 and W735. Comet and Diamond each produce similar interesting 5 band wire dipoles that utilize both traps and a fan arrangement - the Diamond W8010 and the Comet CWA-1000. If space really is limited then look out for KZJ Communications (dongo1950 on ebay) - he produces 'Limited Space Inductive Dipoles'. These are inductively loaded and shortened dipoles so they will have reduced efficiency, of course, but are very nicely made, so might be very useful in a tight spot.

Alpha Delta Communications produce a couple of substantial parallel dipole models: http://www.alphadeltacom.com

To obtain good efficiency and achieve a low angle of radiation, desirable for longer distance DX, a horizontal dipole needs to be installed at a good height - over 20 feet would be desirable and it is quite common to install horizontal dipoles at around 30 to 40 feet above ground level. This might be a problem at some QTH's, it certainly is at mine!

Allan Copland, GM1SXX comments: "The dipole will operate well on the band it has been sized for , if placed at a suitable height, but will also operate as a’ three-half-wave’ aerial at three times the frequency and so on, so it’s not strictly a single band aerial. An 80M dipole (132 feet typical) will work nicely on 30 metres (three half waves) but not on 40m (two half waves)… because on 40M the feed-point is at a voltage node and not at a current node, for easy feeding. Most aerials are current fed.

The radiation pattern changes when a dipole is not used on its design frequency. The pattern will break up into multiple ‘petals’. This can be either a disadvantage or an advantage depending on what you expect from it. Since most of us use co-ax, an UN-BAL should really be used to connect the unbalanced feeder to the balanced aerial, but how many people actually bother? Not many I suspect. It’s possible of course to use a balanced feed-line system instead with a dipole and just have a delta match (no centre insulator… none needed). There are many choices and permutations, but in general, dipoles are centre fed at a point of current maximum (and minimum voltage).

A normal dipole is current fed but of course can be voltage fed instead. This is what’s done in the EFHWA or Fuchs aerial where a resonant half wave wire is fed at one end (max voltage / min current) from an L/C tank, against a very short counterpoise wire.

Fan Dipoles (a.k.a. Parallel) Dipoles:

Another design of multi band wire dipole is the fan dipole, or parallel dipole. A fan (or parallel) dipole will have, perhaps, two, three or four individual resonant dipoles with the arms arranged in a fan shape.

A fan dipole is a very handy way of using a dipole that will be resonant on several bands - typically three or four. The fan dipole (a.k.a. Parallel Dipole)

See M0WYM's page for a QRP Fan Dipole design: http://www.radiowymsey.org/FanDipole/fandiploe.htm

See this page for construction details: http://www.hamuniverse.com/multidipole.html

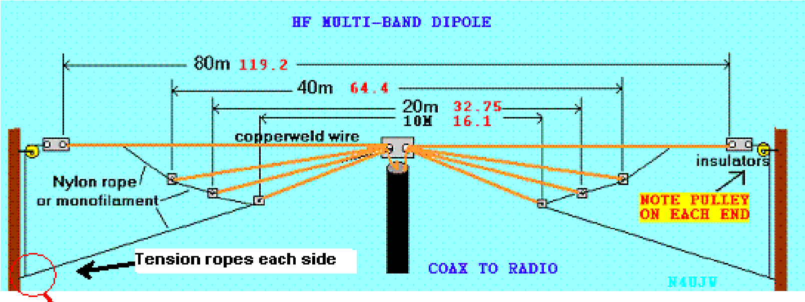
HF MULTI-BAND DIPOLE
80m 119.2
40m 64.4
20m 32.75
10M 16.1
copperweld wire
insulators
Nylon rope or monofilament
NOTE PULLEY ON EACH END
Tension ropes each side
COAX TO RADIO
N4UJW

Tension rope is not tied to pully rope in picture. It is tied near location of pully rope down on supports within easy reach. It is tied last after final SWR adjustment and the antenna is in it's final position.

Suggested total lengths:
80 meters - 120 feet
40 meters - 65 to 66 feet
20 meters - 34 feet
10 meters - 17 feet
These lengths are not exact. Some tuning may be required. Use the standard formula 468 / freqmhz for total feet for each band (freq) of interest. Adjust each length longer or shorter as needed.

Fan Dipole shown on Ham Universe

KI4NRH built a really neat fan dipole shown in the photograph below:

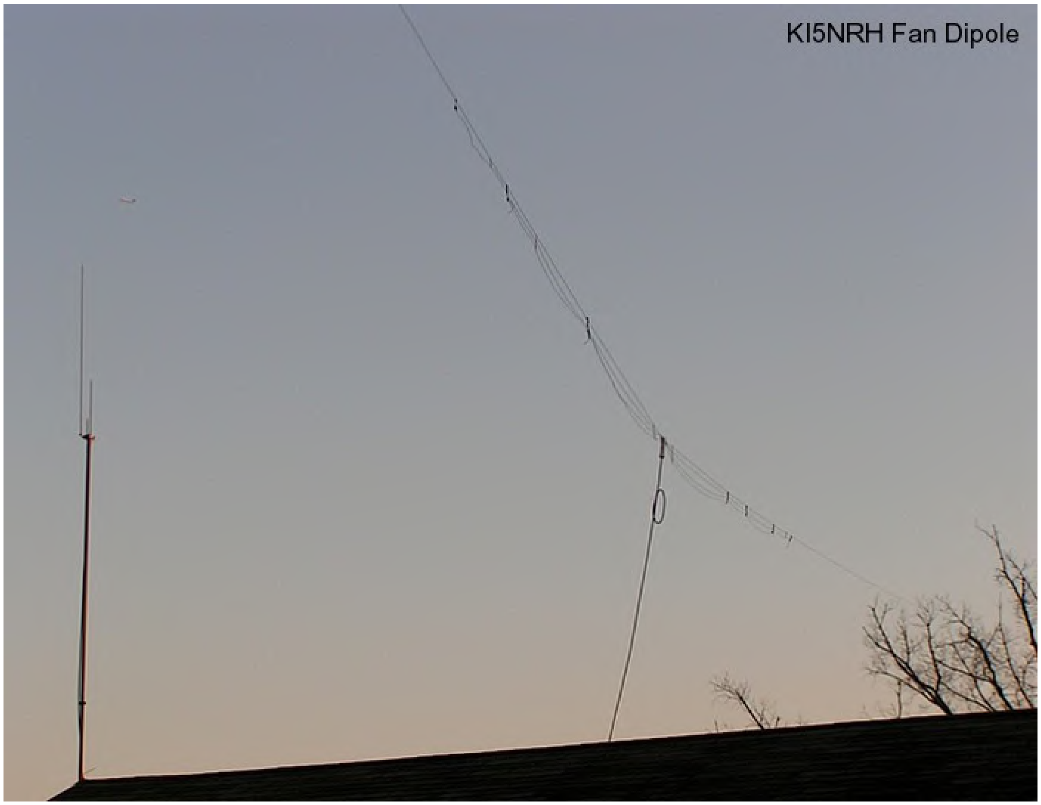
KI5NRH Fan Dipole

Fan Dipole by ki4nrh

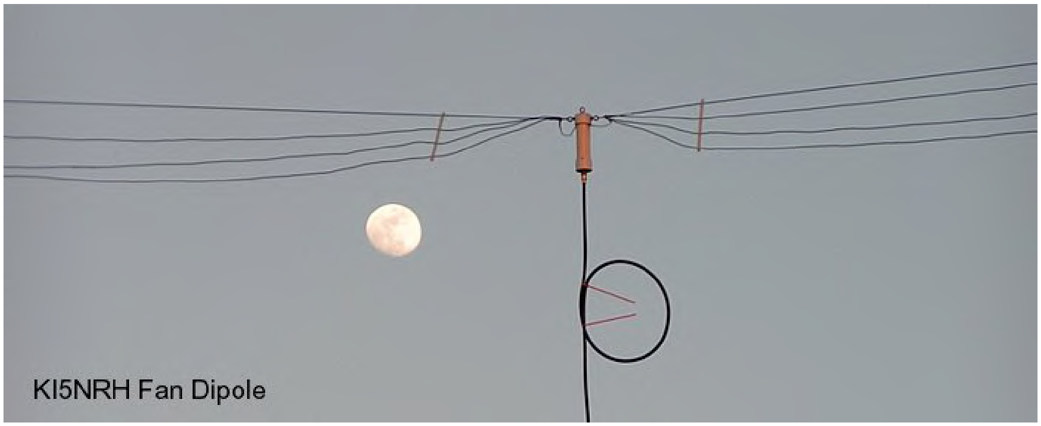
KI5NRH Fan Dipole

Fan Dipole by ki4nrh
http://forums.qrz.com/showthread.php?t=159953

Link Dipoles

Link Dipoles (a.k.a Jumpered Dipoles) facilitate multi band operation by simply connecting the jumpers (one on each side of the aerial) to achieve the desired resonant band. Perhaps a bit bothersome for frequent band changes, but a very simple and effective aerial and very handy for portable operation, not to mention efficient for QRP.

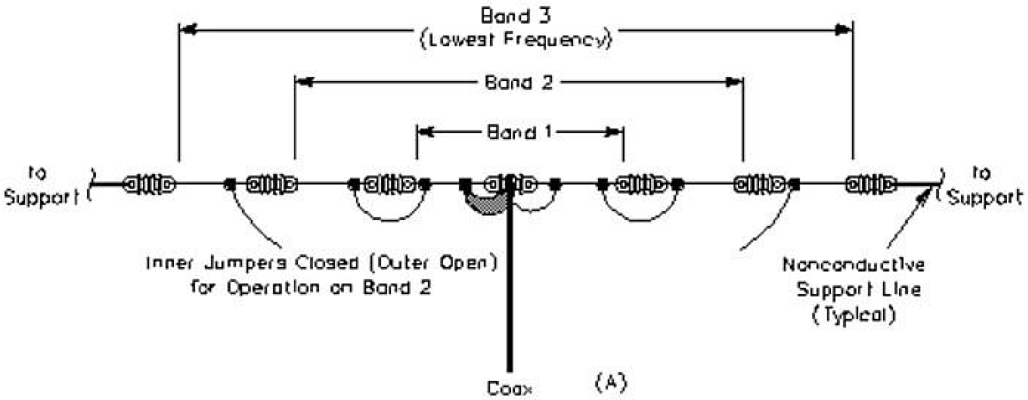
Band 3
(Lowest Frequency)
Band 2
Band 1
to Support
to Support
Inner Jumpers Closed (Outer Open) for Operation on Band 2
Nonconductive Support Line (Typical)
Coax
(A)

Link Dipole - ref ARRL

Sloper Aerials

Alternatively a dipole can be installed as a sloper; one end fixed to a high point on the house or building, or a tall post maybe 8 to 10 metres high, with the other end attached to a lower point such as a post maybe 3 or 4 metres high. This will give the aerial some directivity.

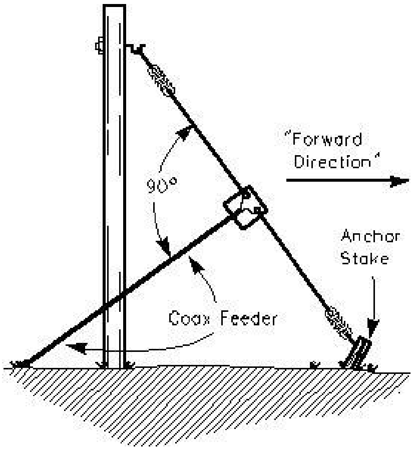
"Forward Direction"
90°
Anchor Stake
Coax Feeder

Sloper configuration of a wire dipole

Inverted V Aerials

Another option maybe to support the dipole at its centre on a tall pole, or roof apex, with each end sloping downwards to lower fixing points. This will give the aerial an upside down V shape. As with a sloper, the Inverted V arrangement will give the aerial some directivity - a different radiation pattern compared to a straight horizontal dipole.

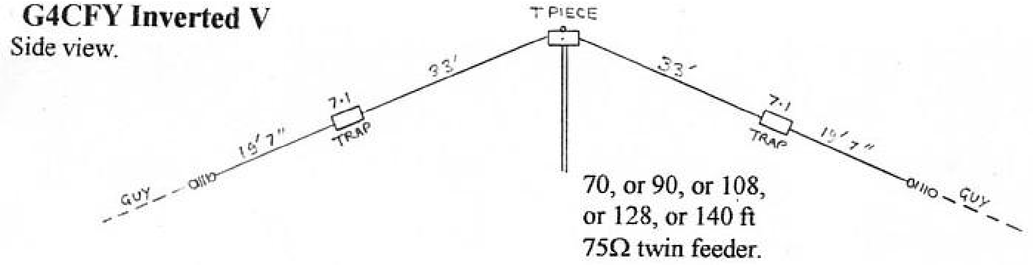
G4CFY Inverted V
Side view.
T PIECE
33'
33'
7·1 TRAP
7·1 TRAP
19' 7"
19' 7"
GUY
GUY
70, or 90, or 108, or 128, or 140 ft 75Ω twin feeder.

Spectrum Communications G4CFY Trapped Dipole in "Inverted V" configuration
http://www.spectrumcomms.co.uk

The Classic Doublet Antenna Again. This time in an Inverted V formation:

Using an Inverted V can help fit a dipole into a slightly restricted space. The Inverted V arrangement can be used for single band resonant dipoles, trapped dipoles and fan dipoles. The Doublet must be fed with Ladder Line or Open Wire balanced feeder for efficient Multi-Band operation.

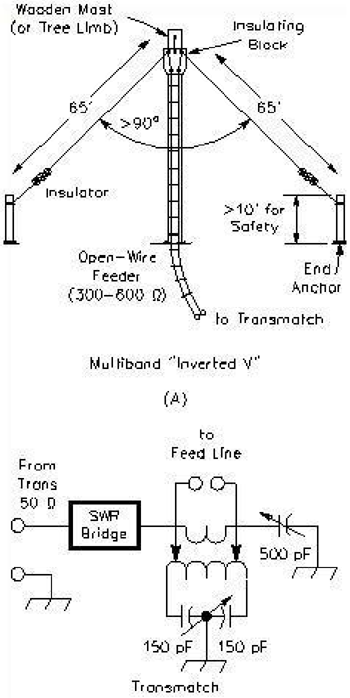
Wooden Mast
(or Tree Limb)
Insulating
Block
65'
>90°
65'
Insulator
>10' for
Safety
Open-Wire
Feeder
(300-600 Ω)
End
Anchor
to Transmatch
Multiband "Inverted V"
(A)
to
Feed Line
From
Trans
50 Ω
SWR
Bridge
500 pF
150 pF
150 pF
Transmatch

At A, details for an inverted V fed with open-wire line for multi-band HF operation. A Transmatch is shown at B, suitable for matching the antenna to the transmitter over a wide frequency range. The included angle between the two legs should be greater than 90° for best performance. [ref: QSL.net]

Vee Configuration

Comet and Diamond produce Vee antennas that can be mounted on the side of a building at roof height, or on a pole, telescopic pole or other suitable support. These are trapped dipoles in an upright V configuration, not made of wire but of aluminium tubing for solid construction. Typically covering 40m, 20m, 15m and 10m. The Comet model is H-422V. The Diamond Model is HFV5 which also covers 6m.

These are shortened antennas so cannot be expected to have high performance on the lower HF bands, but if space is restricted the RF compromise may one that has to be taken.

Comet H-422V
http://www.cometantenna.com

Diamond HFV5 Antenna

Diamond HFV5 Antenna
http://www.diamond-ant.co.jp/english/amateur/antenna/ante_2base/ante_base7.html

Thoughts from Martin G8JNJ on multi-band wire antennas:

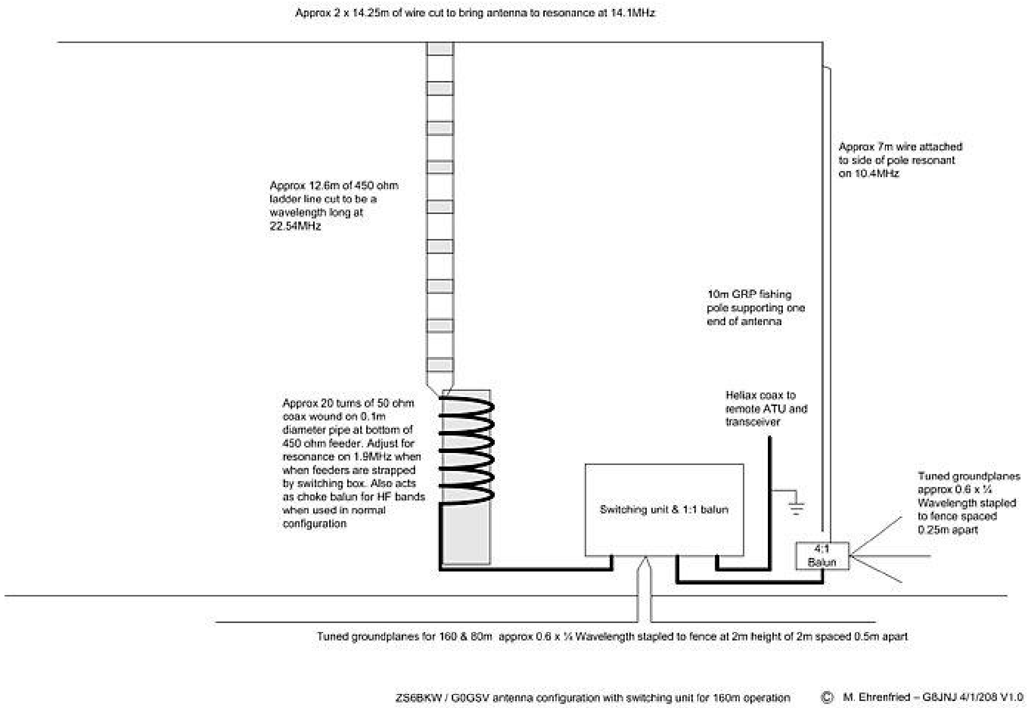
Approx 2 x 14.25m of wire cut to bring antenna to resonance at 14.1MHz
Approx 12.6m of 450 ohm ladder line cut to be a wavelength long at 22.54MHz
Approx 7m wire attached to side of pole resonant on 10.4MHz
10m GRP fishing pole supporting one end of antenna
Approx 20 turns of 50 ohm coax wound on 0.1m diameter pipe at bottom of 450 ohm feeder. Adjust for resonance on 1.9MHz when when feeders are strapped by switching box. Also acts as choke balun for HF bands when used in normal configuration
Heliax coax to remote ATU and transceiver
Switching unit & 1:1 balun
Tuned groundplanes approx 0.6 x ¼ Wavelength stapled to fence spaced 0.25m apart
4:1 Balun
Tuned groundplanes for 160 & 80m approx 0.6 x ¼ Wavelength stapled to fence at 2m height of 2m spaced 0.5m apart
ZS6BKW / G0GSV antenna configuration with switching unit for 160m operation
© M. Ehrenfried – G8JNJ 4/1/208 V1.0

Multiband Wire Antenna by Martin G8JNJ
http://g8jnj.webs.com/

Linear Loading

Linear loading a dipole can reduce the length to help fit a ling dipole into a shorted space by essentially folding back some of the dipole elements. Here is a design by K4VX for a 7MHz Linear Loaded Dipole: http://www.arrl.org/files/file/Technology/tis/info/pdf/0207040.pdf

End Loaded Dipole

End loading can also help reduce the size of antennas, particularly useful for dipoles used on the 80m and 160m bands.

An end loaded dipole will produce an antenna that is H shaped. There are several commercial designs available produced in designs that cover a single band and others that cover multiple bands. The version shown below is only 3 metres tall so will be suitable for very unobtrusive, low profile use. It is the ProAntennas Multi-band I-PRO: 20m 17m 15m 12m 11m & 10m which uses a capacity hat with some loading at the centre. http://www.proantennas.co.uk/

Other similar antennas were available from Force12 Antennas in the form of, amongst others, the Sigma 5 and Sigma GT5. The Sigma design used T-bars at each end of the vertical dipole for loading technique and off-center loading coils. http://www.force12inc.com This was supplied supplied in the UK by Vine Antennas at one time http://www.vinecom.co.uk . Transworld Antennas also have produce antennaa using a similar concept - the TW2010 Adventurer and Backpacker http://transworldantennas.com

K9AY Notes that: "I have come to the conclusion from my experiments, readings and observations, that a capacity hatted vertical dipole, a few feet over ground, is less compromised than a 1/4 w/l vertical of the same height fed against a less than perfect ground. Let's face it, most amateur's ground systems are mediocre at best. Also, the dipole is easier and cheaper to rig, and is two dimensional..Very important in my situation, as I cannot run out radials on my neighbours property. Or, to quote W4RNL.."Since only a handful of hams can ever have 160-meter antennas high enough to yield a low angle DX signal, more practical are vertical arrays such as yours. Vertical dipoles with hats (or Tees) save a plethora of wire needed by monopoles." http://www.dxzone.com/cgi-bin/dir/jump2.cgi?ID=7466

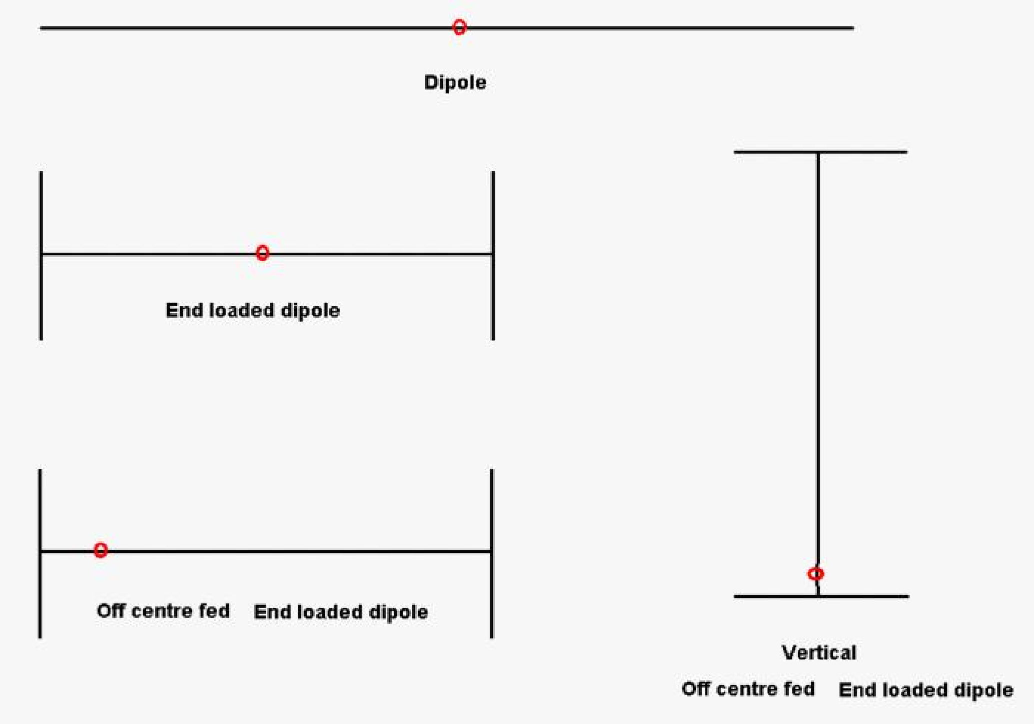
Dipole
End loaded dipole
Off centre fed End loaded dipole
Vertical
Off centre fed End loaded dipole

Information by K9AY

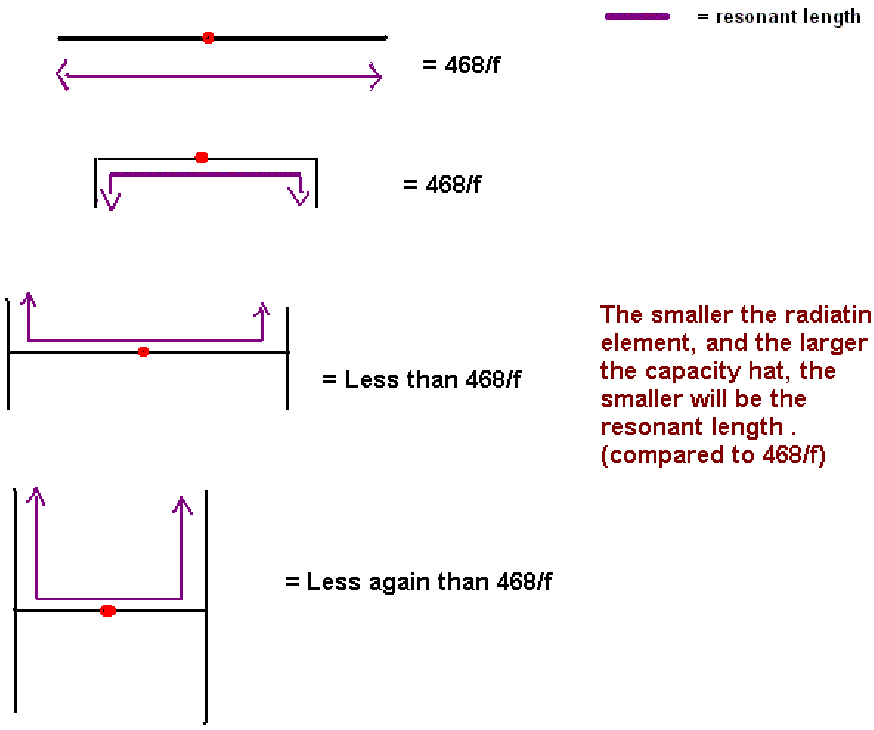
= resonant length
= 468/f
= 468/f
= Less than 468/f
= Less again than 468/f

The smaller the radiating element, and the larger the capacity hat, the smaller will be the resonant length . (compared to 468/f)

Information by K9AY

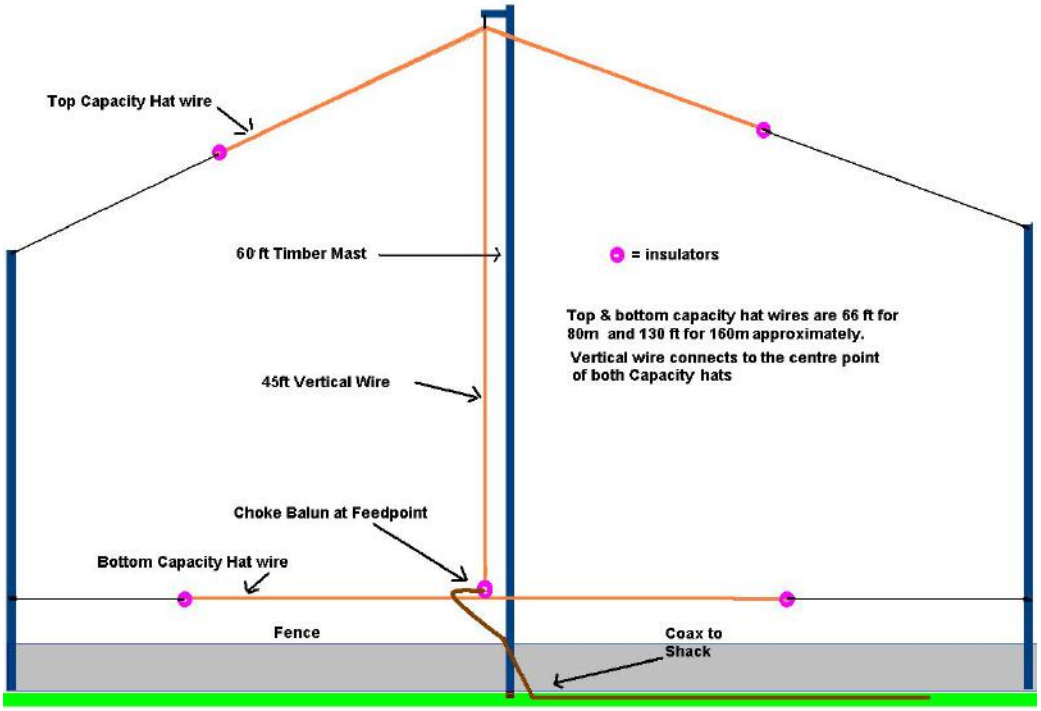
Top Capacity Hat wire
60 ft Timber Mast
= insulators
Top & bottom capacity hat wires are 66 ft for 80m and 130 ft for 160m approximately.
Vertical wire connects to the centre point of both Capacity hats
45ft Vertical Wire
Choke Balun at Feedpoint
Bottom Capacity Hat wire
Fence
Coax to Shack

Interesting concepts from K9AY

End Fed Half Wave Antenna

The End Fed Half Wave Antenna (EFHWA) is fed at a voltage node via a parallel resonant circuit against a ‘short counterpoise’, it is a favourite of backpackers and outdoor types. It can be considered as a half wave dipole that’s end-fed at a voltage node rather than the current node, as is more usual. This is a very handy arrangement for portable QRP work.

EFHWA Link: http://www.aa5tb.com/efha.html

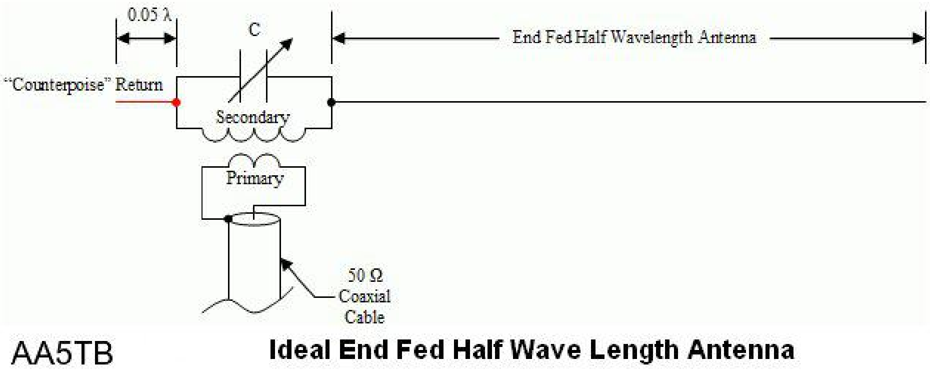
0.05 λ
C
End Fed Half Wavelength Antenna
"Counterpoise" Return
Secondary
Primary
50 Ω
Coaxial
Cable
AA5TB
Ideal End Fed Half Wave Length Antenna

End Fed Half Wave Antenna by AA5TB
http://www.aa5tb.com/efha.html

"I suspect that nothing new or radical has happened in the field of radio aerials in a VERY long time, like at least many tens of decades. Most of the new wonder aerials are really a con. Choke off the feed-line and then see how good they really are. Prime among the baddies is the CFA. It doesn't really work, at least if you place a choke in the feed-line. With any real aerial, there should be minimal radiation from the feed system… so a choke should really make no difference at all, but for the CFA it does! The CFA is not alone, there are others. The popular G5RV is another design with a radiating feed, deliberately so, but of course G5RV planned it that way. He wasn't cheating… merely being a bit devious, to make it multi-band".

"Lots of stuff to pass on to my fellow radio club members, most of whom are of the 'if it's not expensive, it can't be any good' school of thought when it comes to aerials. Nothing of course could be further from the truth! Aerials are one area where it makes a lot of sense to build our own." : Website of GM1SXX - www.observations.biz

Thanks for your email Allan. It's a good idea to point out that an antenna could be pressed into use on odd multiples of its resonant frequency, hence a 3.6MHz antenna for 80m could be useful near the 30 metre, 10.1MHz, band - near to the third harmonic of 3.5 MHz although, as you observe, the radiation pattern may be quite distorted from the traditionally expected dipole pattern and be more petal shaped. The same goes for a 7.1 MHz antenna for 40m being usable on its third harmonic of 21.3 MHz for the 15m band - a 40m dipole being three half waves an the 21 MHz band.

I have not experimented with a full size 80m dipole, but I would guess that it might be useful at 5 times 3.6Mhz in the 18 MHz / 17m band?

The point made about feeding a familiar dipole at the current node rather than the voltage node is obviously very important and, I imagine, sometimes overlooked.

PLANS: Download the pdf plans produced by G0KYA here > More from G0KYA here: http://g0kya.blogspot.com

W3EDP Antenna

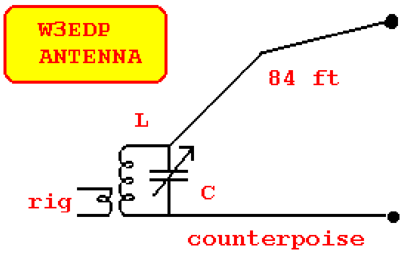

Frank, G3YCC comments on his website: The W3EDP needs a simple matching unit is needed to couple the wire to the rig and a counterpoise is required for some bands, however there is room for experimentation. It has been shown that different lengths or removal of the counterpoise altogether, can improve performance, as described in RadCom, August 1996 by G3LCK.

The Tuning capacitor in the AMU can be a 365 - 500pF broadcast type or a miniature version is OK for QRP use.

Counterpoise lengths: 3.5 & 7.0Mhz - 17ft ; 14Mhz - 6.5ft ; 28Mhz - none

Tuning Unit: Values for coils in the unit, based on a 2 inch former and 16 swg wire:
3.5Mhz 21 turns ; 7.0Mhz 7 turns ; 14.0Mhz - 5 turns.

K3HRN Notes: "Some folks have told me the modifications below make the antenna something other than a W3EDP. I can tell you that it works very well with 5 watts. Create a "bundle" of counterpoise wires, 1/4 wave length for each band you will use. Attach the bundle to the tuner in place of the counterpoise pictured above. Be cautious, 1/4 wave length elements can have high RF voltages present, even at QRP power levels. I've been able to work 160-10, including WARC bands with this type of antenna".

W3EDP or Zepp ?

It's the antenna favoured by VA3QV for all band Portable QRP operating!

VA3QV uses this home-brewed antenna with a small LDG Z100 antenna tuner for portable QRP work.

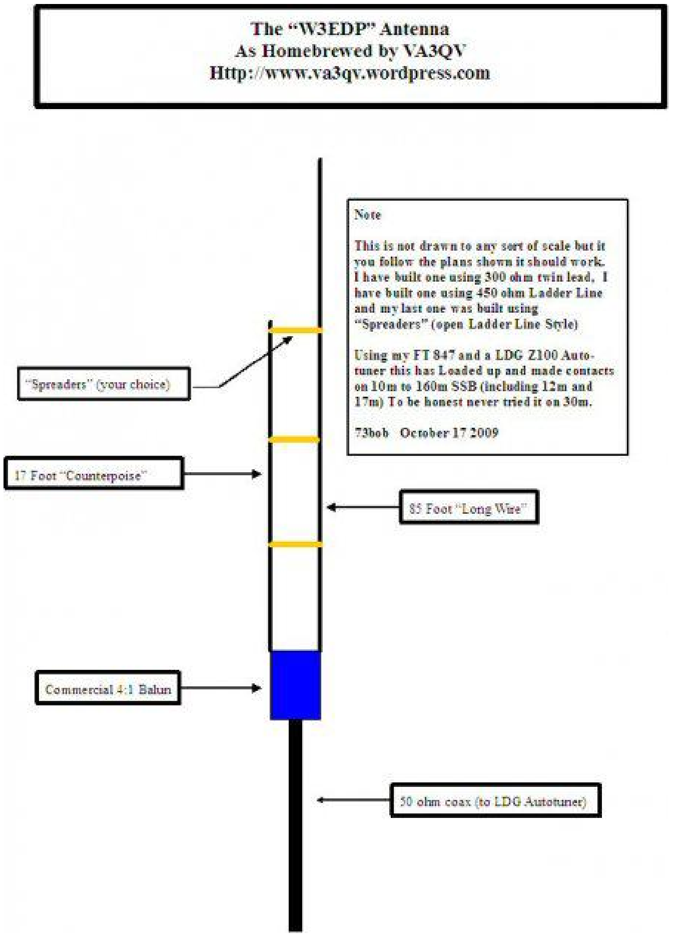

The "W3EDP" Antenna
As Homebrewed by VA3QV
Http://www.va3qv.wordpress.com

Note

This is not drawn to any sort of scale but it you follow the plans shown it should work. I have built one using 300 ohm twin lead. I have built one using 450 ohm Ladder Line and my last one was built using "Spreaders" (open Ladder Line Style)

Using my FT 847 and a LDG Z100 Auto-tuner this has Loaded up and made contacts on 10m to 160m SSB (including 12m and 17m) To be honest never tried it on 30m.

73bob October 17 2009

Homebrew antenna by VA3QV
http://va3qv.wordpress.com

Ye Olde Zepp

Marconi spins in his grave every time a ham buys an aerial instead of building it ! (W1GFH)

Here is a wonderful olde worlde style cartoon from WB1GFH that certainly inspires antenna experimentation with antenna designs:

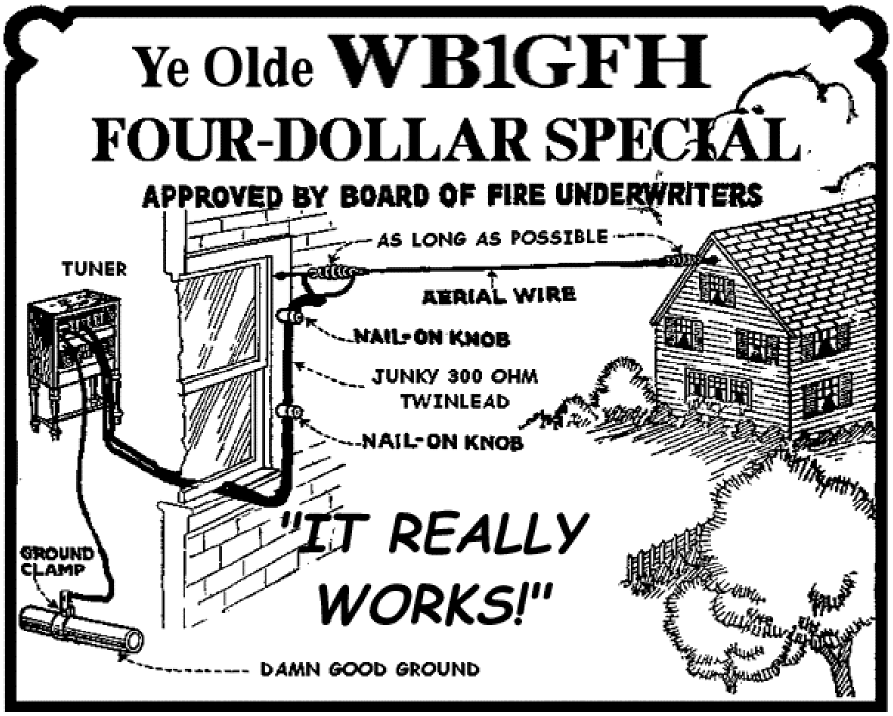
Ye Olde WB1GFH
FOUR-DOLLAR SPECIAL
APPROVED BY BOARD OF FIRE UNDERWRITERS
AS LONG AS POSSIBLE
TUNER
AERIAL WIRE
NAIL-ON KNOB
JUNKY 300 OHM TWINLEAD
NAIL-ON KNOB
GROUND CLAMP
"IT REALLY WORKS!"
DAMN GOOD GROUND

Superb. I love it!

See more inspiration from Joe Tyburczy, W1GFH, here: http://www.hamuniverse.com/fourdollarspecialw1gfh.html

End Fed Zepp "Zepp"

The End Fed Zepp consists of a 1/2 wavelength horizontal radiator wire connected to one conductor of a length of parallel open wire feeder, 300 ohm or 450 ohm twin feeder. The feeder is often quoted as being 1/4 wavelength long.

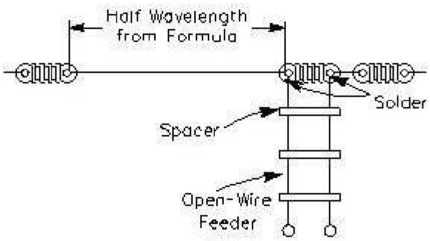
Half Wavelength from Formula
Solder
Spacer
Open-Wire Feeder

Basic design of an end fed Zepp

G Whip Antenna Products manufacture and supply a version of the Zepp antenna. Geoff G4ICD explains: "The end fed Zepp shown below has no counterpoise, just a tuned circuit in the feedpoint plus a half wave radiator. This is a most interesting antenna and can be used on other bands with the use of an Antenna Matching Unit."

G Whip Antenna News
New Products.
WWW.GWHIP.CO.UK
G Whip Antenna Products
MADE IN BRITAIN
G Whip Antenna Products

A high quality End Fed Zepp style antenna: This variation uses a tuned circuit rather than a tuned twin feeder.
Supplied by G Whip Antenna Products of the UK
www.gwhip.co.uk

The end fed zepp is a popular antenna often used to save space and gets its name from the fact that it was used as an end fed wire trailing out from the rear of Zeppelin airships. It consists of a 1/2 wavelength horizontal radiator wire connected to one conductor of a length of parallel 300 ohm or 450 ohm twin feeder, often quoted as being 1/4 wavelength long.

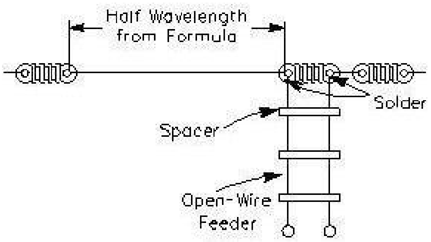
Half Wavelength from Formula
Solder
Spacer
Open-Wire Feeder

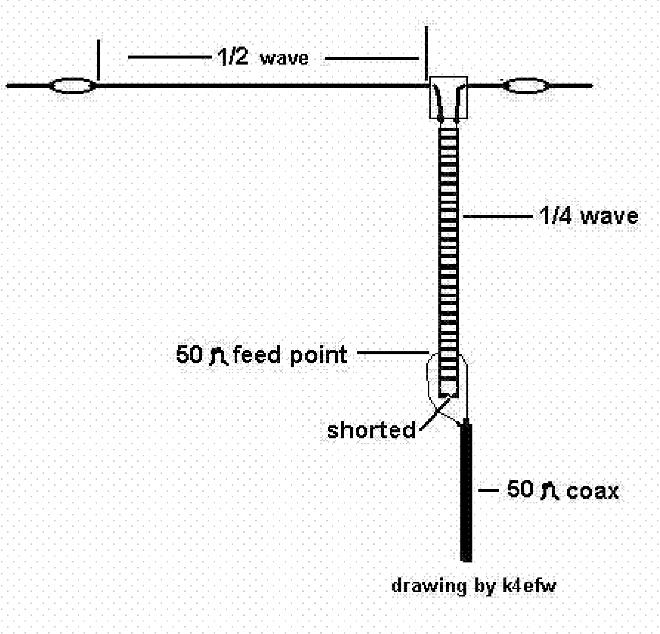

Zepp Antenna by K4EFW
http://www.hamuniverse.com/n4jaantennabook.html

K4EFW notes: "...A half-wave resonant antenna can be fed from its end. When fed this way, it is also known as an end-fed zepp. An end-fed zepp will work on its fundamental frequency and on odd and even harmonic frequencies. The end of a half-wave antenna has very high impedance, and an antenna fed this way is said to be voltage fed. Feeding a half-wave resonant dipole in the center means it is current fed. The normal way of feeding the end-fed antenna is with ladder-line. One side of the ladder-line is connected to one end of the antenna and the other side of the ladder-line is connected to nothing. To secure the unconnected side of the ladder-line, it is connected to a short wire running between two insulators. Since the antenna is connected at its high impedance point, no current flows into an antenna, but there will be a large current in the center of this antenna. No current flows from the open side of the feed-line because it is at a zero current point. The end-fed zepp can be matched by cutting the ladder-line to a quarter wavelength with the bottom end of the ladder-line shorted. A certain distance above the short is a 50-ohm feet-point and it can be fed directly with coax. You will have to find the 50-ohm point by trial and error. This method of feed makes it a single band antenna". Quoted from K4EFW.

Here is a commercial product made by G-Whip Antennas of the UK offering their version of a Zepp antenna design:
http://www.gwhip.co.uk/

Martin G8JNJ highlights a very interesting antenna designed by Mike G7FEK here:-

The G7FEK antenna goes several stages further than the simple single band End Fed Zepp. G7FEK has produced a design for Multi-Band operation claimed to offer much improved performance over a half sized G5RV or 'Windom' antenna while additionally providing access to the 80 Metre Band.

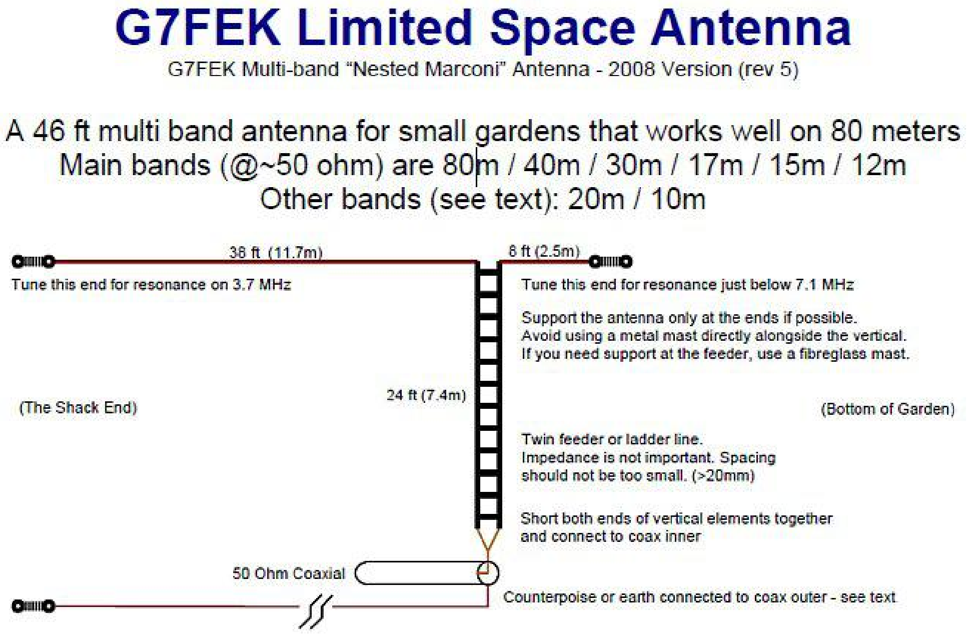
G7FEK Limited Space Antenna
G7FEK Multi-band "Nested Marconi" Antenna - 2008 Version (rev 5)
A 46 ft multi band antenna for small gardens that works well on 80 meters
Main bands (@~50 ohm) are 80m / 40m / 30m / 17m / 15m / 12m
Other bands (see text): 20m / 10m
38 ft (11.7m)
8 ft (2.5m)
Tune this end for resonance on 3.7 MHz
Tune this end for resonance just below 7.1 MHz
Support the antenna only at the ends if possible.
Avoid using a metal mast directly alongside the vertical.
If you need support at the feeder, use a fibreglass mast.
24 ft (7.4m)
(The Shack End)
(Bottom of Garden)
Twin feeder or ladder line.
Impedance is not important. Spacing should not be too small. (>20mm)
Short both ends of vertical elements together and connect to coax inner
50 Ohm Coaxial
Counterpoise or earth connected to coax outer - see text

G7FEK Limited Space Multi-Band Antenna

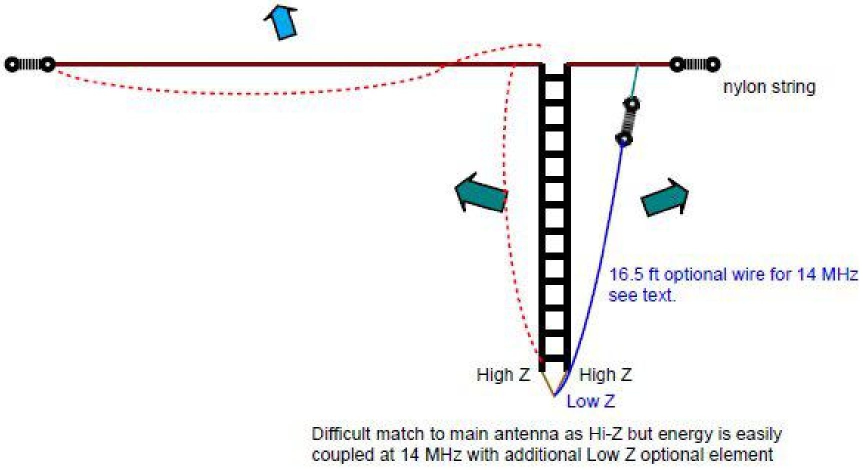
nylon string
16.5 ft optional wire for 14 MHz see text.
High Z
High Z
Low Z
Difficult match to main antenna as Hi-Z but energy is easily coupled at 14 MHz with additional Low Z optional element

Adding 14MHz to the G7FEK Limited Space Multi-Band Antenna

The G7FEK design will allow operation on 80m / 40m / 30m / 17m / 15m / 12m with the possibility to add the 20m band.

G7FEK Plans: Download the plans for the G7FEK antenna from G7FEK's website here or from this website here

Tuner Design For Half Wave Vertical and Similar Length End Fed Antennas by G4FGQ

http://www.radioaficion.com/HamNews/articles/8218-tuner-design-for-half-wave-vertical-a-similar-length-end-fed-antennas.html

Off Centre Fed Dipole (OCFD) - so called 'Windom' Antenna

The "Windom Antenna" was described by Loren G. Windom W8GZ. It could be an ideal wire aerial for use in restricted spaces for multi-band operation. It may also be an good candidate for portable work.

It is a wire antenna, similar to a dipole, but unlike a dipole or doublet which is fed at the exact centre, a Windom or Off Centre Fed Dipole, as the name suggests, has the feed point off center.

Current versions of the Windom are not actually Windom antennas at all - instead they are fed with coaxial cable and have a balun placed at the feed point. These are Off Centre Fed Dipole antennas - OCFD. They are similar in appearance to a true wire Windom, but they are different.

As with all aerials the aerial should be as high as possible. With the feed point at between 20 and 40 feet above ground the typical claimed impedance will be somewhere in the region of 200 Ohms so a 4:1 balun will typically be required. At greater heights, and depending upon the exact position of the feed point, the impedance may be higher and a 5:1 or 6:1 balun might be a better choice although balun losses will be greater.

The point at which a Windom is fed in the original design, which used an open wire to feed the aerial, was 15 percent off-centre. The current designs, which are fed with coaxial cable, are typically fed about 33 percent off centre, so one leg is 67 percent of the total length and the other leg is 33 percent of the overall length of the aerial.

The bands that are covered depends upon the overall length of the aerial:

11 metres long (approx) should cover 20m, 15m and 10m and the WARC bands with a tuner.

21 metres long (approx) should cover 40m, 20m, 15m and the 10m bands and WARC with a tuner.

41 metres long (approx) should cover 80m, 40m, 20m, 15m and 10m and WARC with a tuner.

80 metres long (approx) should cover 160m, 80m, 40m, 20m, 15m and 10m and WARC with a tuner.

Cut the aerial for the lowest band to be used. In imperial measurements using a familiar formula:
The longer leg will be 468 divided by the frequency and multiplied by .67 = length in feet
The shorter leg will be 468 divided by the frequency and multiplied by .33 = length in feet

OCFD Formulas:

The offset proportions differ according to which sources one refers. Some sources suggest 33% / 67% but other dimensions are also to be found:

62.2% for one side and 37.8% for the other leg. So:
The longer leg will be 468 divided by the frequency and multiplied by .622 = length in feet
The shorter leg will be 468 divided by the frequency and multiplied by .378 = length in feet
[Source: New Caroline Windom - http://www.hamuniverse.com/k4iwlnewwindom.html]

Other ideas:

The proportions of 69% / 37% are used by Buxcomm who say that "One third plus two thirds will not work. Use the formula below, as is: Do not be concerned with the off-set of the feed point, as this formula takes into consideration, the correct off-set for feeding the (BUXCOMM) Windom.for the other leg." **So:**
The longer leg will be 468 divided by the frequency and multiplied by .69 = length in feet
The shorter leg will be 468 divided by the frequency and multiplied by .37 = length in feet
[Source Buxcom: http://www.buxcomm.com/windom_files/WINDOM.htm]

Given the fairly simple formula it should be quite easy to make an OCFD Windom - however a Windom can be purchased at very reasonable cost commercially, for example from M0CVO at http://m0cvoantennas.webs.com alternatively G-Whip Antennas of the UK supply extremely high quality, high efficiency 4:1 baluns (and other baluns) that could be used at the centre of any Off Centre Fed Dipole - just add the correct wire lengths to each side.

Geoff G4ICD / GJ4ICD of GWhip highlighted his website which has an interesting page with comments concerning the quality and construction of balun products. This feature on the G-Whip site can be seen here: http://www.g4icd.co.uk/baluns.htm

Here is a photograph of the very high quality G-Whip OCFD antenna product www.gwhip.co.uk :

An 'Off Set Centre Dipole' - OSCD - Fed made by G Whip Antennas
G-Whip Antennas - www.gwhip.co.uk

M0CVO ANTENNAS
HW-20HP Multi-band Antenna, 20 - 6m,
400W, 50Ω
Designed and manufactured by M0CVO
Made in England

HW-20HP Off Centre Fed Dipole - produced by M0CVO
6.76 metres long one side and 3.38 metres on the other side.
http://m0cvoantennas.webs.com

M0CVO produces a couple of off centre fed dipoles, the HW-40HP and the HW-20HP. Both antennas handle 400 watts - the HP designation refers to high power.

The M0CVO HW-20HP is 10.14m in length and covers 6 bands - 20, 17, 15, 12, 10 & 6m, no ATU, and is said to also work on 30m & 40m with an ATU. Can be mounted as a horizontal, inverted vee or a sloper. VSWR is said to be 1.4 on 20m, 2.8 on 17m, 1.1 on 15m, 1.0 on 12m, 1.0 on 10m and 1.5 on 6m. The antenna is6.76 metres long on one side and 3.38 metres long on the other side. This uses a 66.6% / 33.3% formula.

The HW-20P was reviewed by Steve Nichols, G0KYA, in the January 2012 edition of RadCom.

The HW-40HP is 20.28m in length and will operate on 40, 20 and 10m without an ATU and 80, 60, 15, 6 and WARC Bands with an ATU. (Presuably one leg is 13.52 metres long and the other 6.76 metres if it follows the same 66.6% / 33.3% formula as the HW-20HP).

Here is a graphic of an Off Centre Fed Dipole typical of those available in the USA:

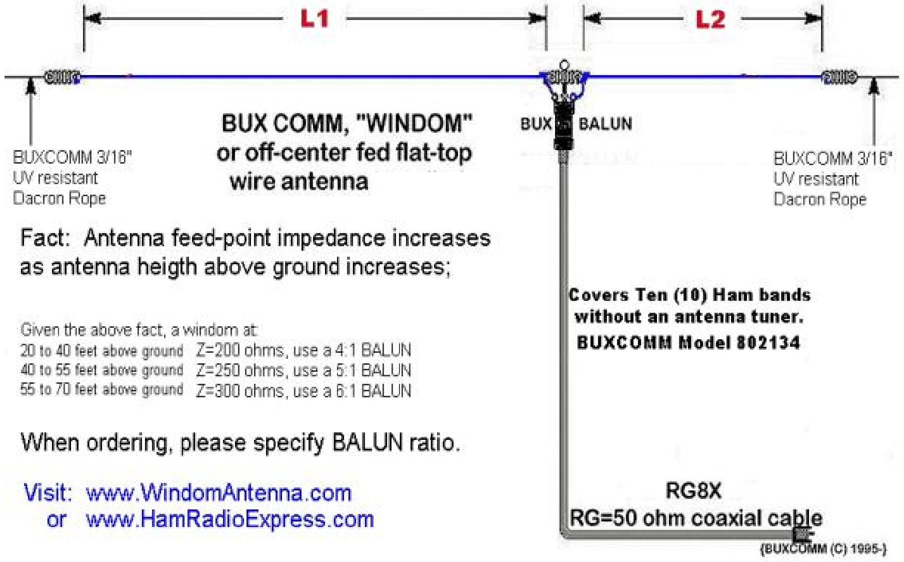
L1
L2
BUXCOMM 3/16" UV resistant Dacron Rope
BUX COMM, "WINDOM" or off-center fed flat-top wire antenna
BUX BALUN
BUXCOMM 3/16" UV resistant Dacron Rope
Fact: Antenna feed-point impedance increases as antenna heigth above ground increases;
Given the above fact, a windom at
20 to 40 feet above ground Z=200 ohms, use a 4:1 BALUN
40 to 55 feet above ground Z=250 ohms, use a 5:1 BALUN
55 to 70 feet above ground Z=300 ohms, use a 6:1 BALUN
When ordering, please specify BALUN ratio.
Visit: www.WindomAntenna.com
or www.HamRadioExpress.com
Covers Ten (10) Ham bands without an antenna tuner.
BUXCOMM Model 802134
RG8X
RG=50 ohm coaxial cable
(BUXCOMM (C) 1995-)

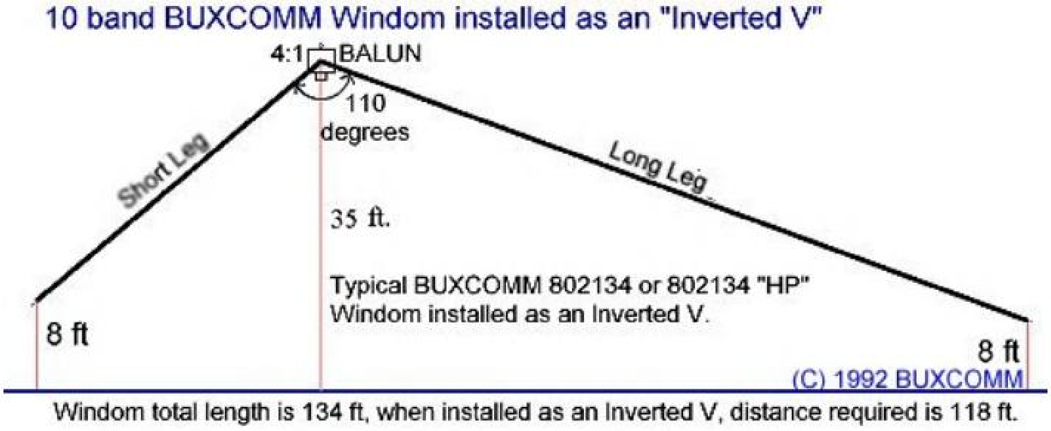
10 band BUXCOMM Windom installed as an "Inverted V"
4:1 BALUN
110 degrees
Short Leg
Long Leg
35 ft.
Typical BUXCOMM 802134 or 802134 "HP" Windom installed as an Inverted V.
8 ft
8 ft
(C) 1992 BUXCOMM
Windom total length is 134 ft, when installed as an Inverted V, distance required is 118 ft.

http://www.buxcomm.com

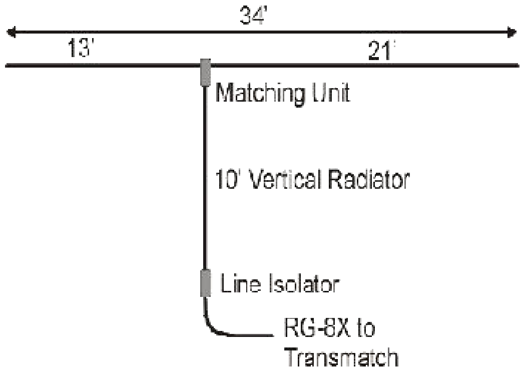
34'
13'
21'
Matching Unit
10' Vertical Radiator
Line Isolator
RG-8X to Transmatch

Carolina Windom for 20 metres to 10 metres
http://www.radioworks.com/ccwcover.html

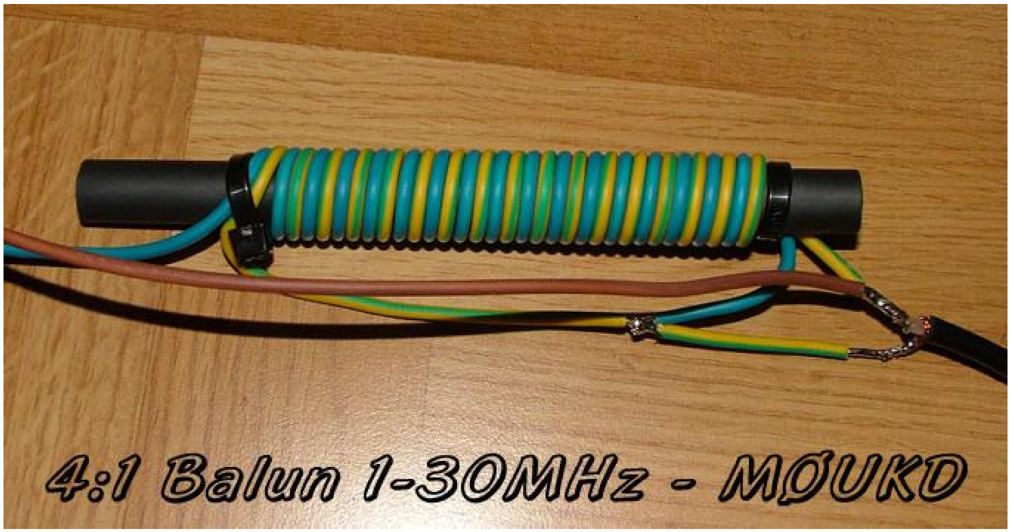
4:1 Balun 1-30MHz - M0UKD

M0UKD - 4:1 balun. It is 17 bifilar turns on a half inch ferrite rod. 50Ω - 200Ω, 1-30MHz
http://www.m0ukd.com/Carolina_Windom/index.php

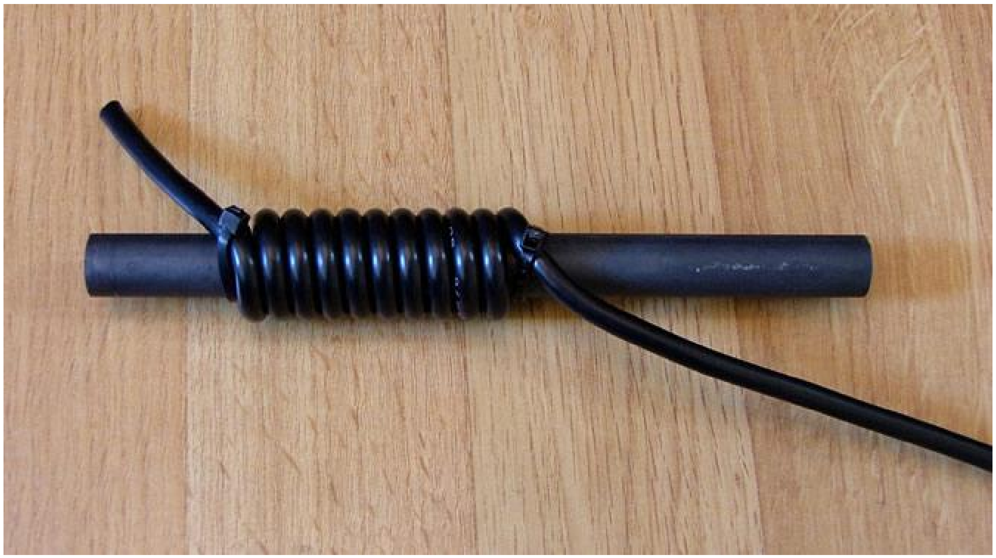

M0UKD Line isolator - 10 turns RG8 on a half inch ferrite rod
http://www.m0ukd.com/Carolina_Windom/index.php

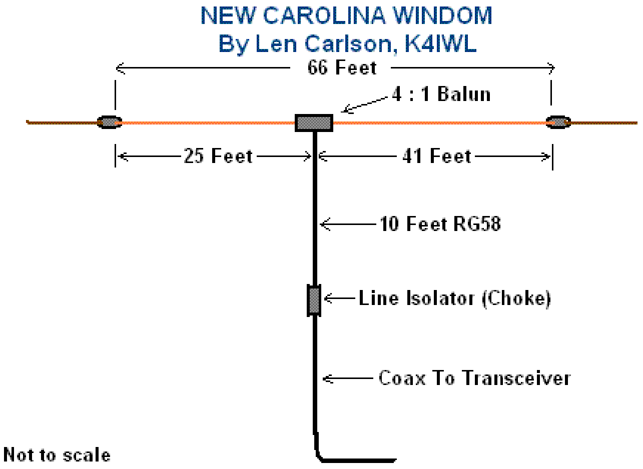
NEW CAROLINA WINDOM
By Len Carlson, K4IWL
66 Feet
4 : 1 Balun
25 Feet
41 Feet
10 Feet RG58
Line Isolator (Choke)
Coax To Transceiver
Not to scale

Windom design for 40m 20m 15m and 10m by K4IWL
http://www.hamuniverse.com/k4iwlnewwindom.html

More information on this general subject at BucksCom: http://www.packetradio.com/windom.htm **or** http://www.buckscom.com/pdfzips/windom.pdf

Commercial Suppliers:

G-WHIP Antennas (UK) : http://www.gwhip.co.uk
M0CVO OCFD Antennas (UK) : http://m0cvoantennas.webs.com

In the U.S.A:
Bux Comm Windom Antennas : http://www.buxcomm.com
Buck Master OFC Dipole Antennas : http://hamcall.net/
Alpha Delta : http://www.alphadeltacom.com
Carolina Windom : http://www.radioworks.com/ccwcover.html

See some Windom - Off Centre Fed Dipole - designs at these links:
http://users.erols.com/k3mt/windom/windom.htm
http://www.dxzone.com/cgi-bin/dir/jump2.cgi?ID=7478
http://www.radioelectronicschool.net/files/downloads/ocfdipole.pdf
http://www.hamuniverse.com/k4iwlnewwindom.html
http://www.g4nsj.co.uk/windom.shtml
http://www.m0ukd.com/Carolina_Windom/index.php

Semi-Permanent Antenna Installations

If it impossible to install a permanent aerial, then another option is to use an antenna designed for portable or mobile work deploying it only as and when necessary in the back yard or garden - perhaps supported with a portable tripod and/or guy ropes.

From the ideas above it should be possible to rig up a semi-permanent or removable antenna for low visual impact.

There are also very many portable antennas produced commercially that might be very useful to utilize on a semi-permanent basis. Commercially bought antennas can be very expensive indeed, especially when compared to 'home brew' aerials, but examples that immediately spring to mind for consideration include: The DMV-Pro, I-Pro, G Whip or G Whip Backpacker, the TW2010 from Transworld Antennas, the Sigma5 from Force 12 and aerials from SuperAntennas. Sandpiper Aerial Technology offer a very good choice of aerials such as the MV and MV-Portable, Buttie or Walkabout mk11 at very attractive prices. http://www.sandpiperaerials.co.uk There are very many other compact and portable antenna systems that are widely available. Check out all the amateur radio dealers for more ideas.

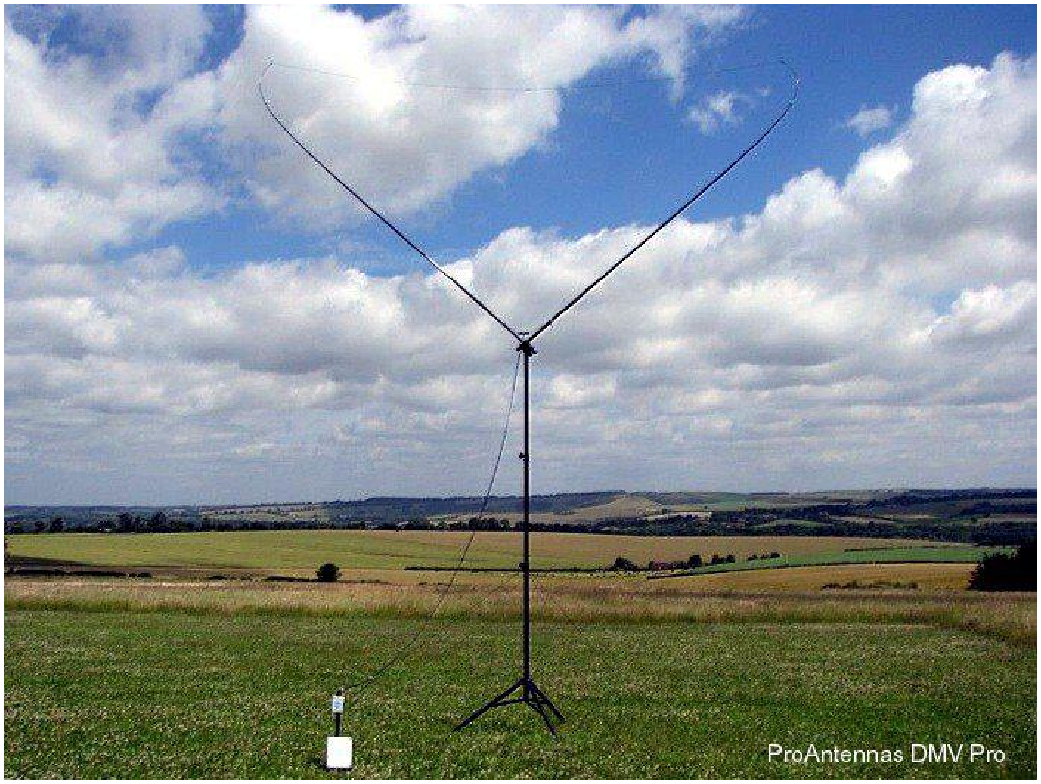

The DMV-Pro Antenna from ProAntennas (shown above) could be used in a back garden whenever required as could the company's I-Pro antenna. The DMV-Pro uses two fibreglass arms that allow the wire aerial to be supported in a number of different configurations such a "V" shaped, "L" shaped, "M" shaped and Delta. The aerial elements are fed to a 4:1 balun using low loss twin feeder, the balun is then connected to an auto ATU, such as the LDG Z-11 Pro, which is connected to the transceiver using coaxial cable. A versatile idea that could also be implemented on a DIY basis with a little experimentation! http://www.proantennas.co.uk

Geoff G4ICD / GJ4ICD mentions the original design, the JJ1VKL published in CQ ham radio Sep' 2000 in Japan. "This one goes back to 2000 and is now copied by several antenna manufacturers in the UK" It is an HF multi-band Delta loop antenna for 3.5-50MHz
http://www.geocities.jp/yoshiki_ja/deltae.htm

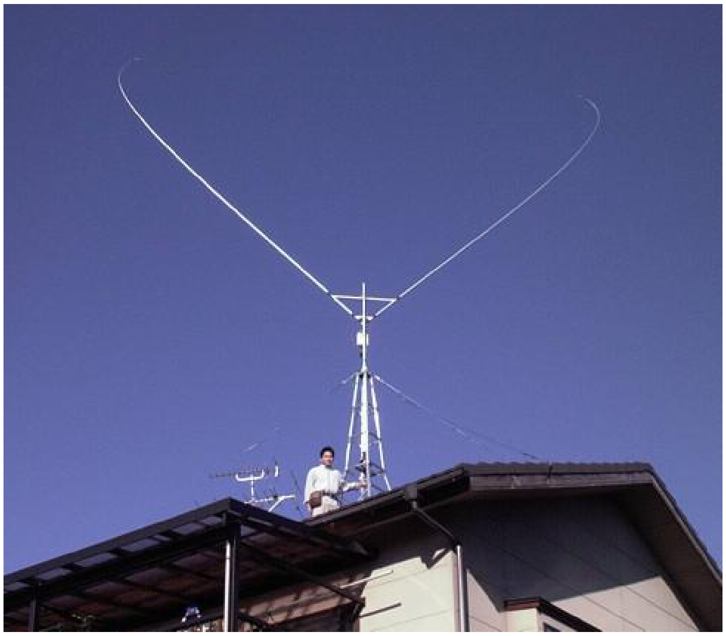

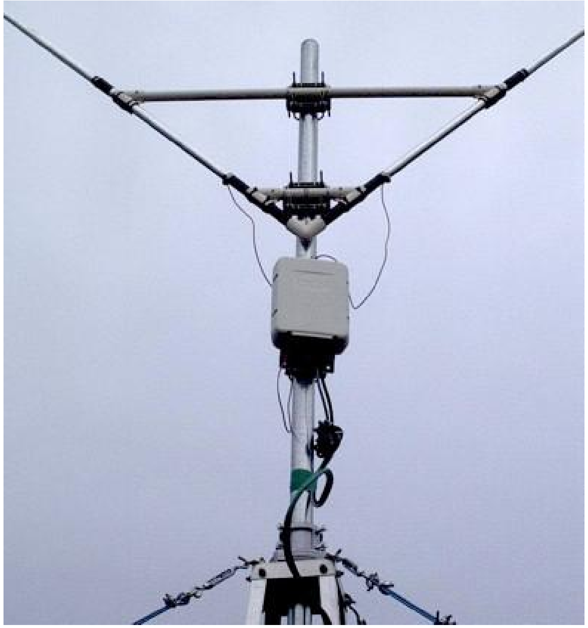

An Original Delta Loop design by JJ1VKL
Read more here: http://www.geocities.jp/yoshiki_ja/deltae.htm

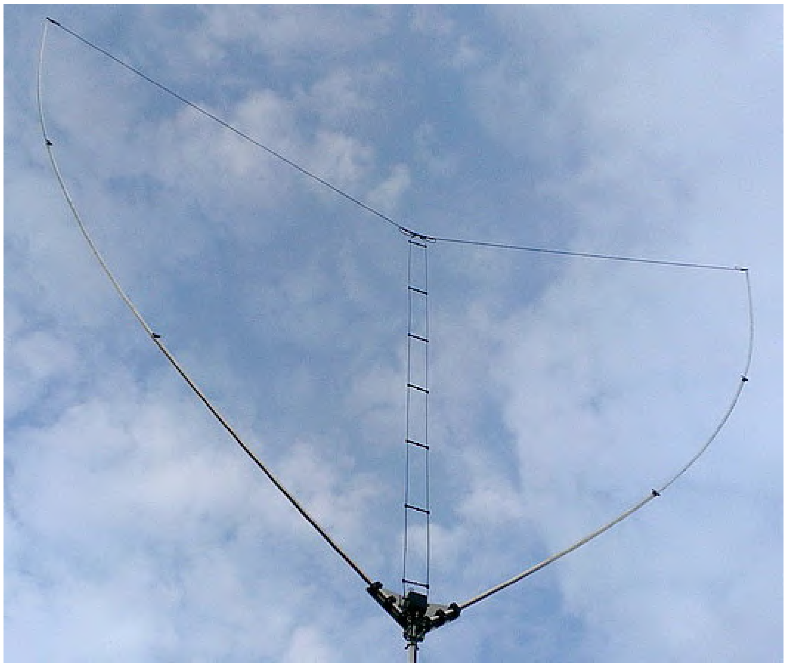

Delta Loop by Arthur M0PLK (SQ2PLK)
Details at http://pdxa.one.pl/articles.php?article_id=17
Available at http://ham-radio.urbasket.eu and http://www.vpa-systems.pl/

STEALTH / COVERT / HIDDEN or DISGUISED ANTENNAS:

Ideas from **G4ILO** - Stealth Antennas: http://www.g4ilo.com/stealth.html

VK2ZOI - "Flowerpot" Antennas

Some ideas by VK2ZOI about producing inconspicuous antennas - perhaps disguised as a plant in a flower pot! There are ideas for 6 metre, 2metre and 70cm antennas including a dual band 2,/70cm design.

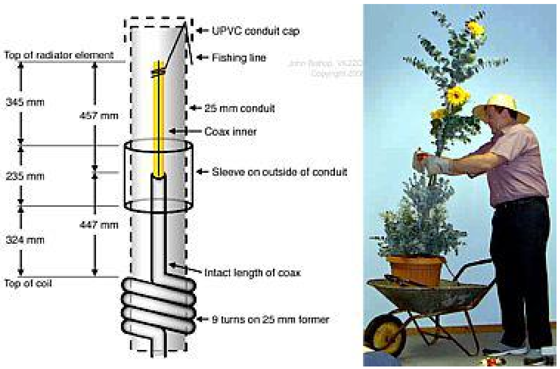
UPVC conduit cap
Fishing line
Top of radiator element
345 mm
457 mm
25 mm conduit
Coax inner
Sleeve on outside of conduit
235 mm
447 mm
324 mm
Top of coil
Intact length of coax
9 turns on 25 mm former

VK2ZOI Flowerpot Antenna projects
http://vk2zoi.com

All Band HF Vertical Antennas (non resonant) - 'Untennas'

So this is where the search for a multi-band antenna begins. It's a difficult task especially if space is limited. First considerations might lie with the commercially available options that are available. Commercially manufactured aerials are available at widely varying price points - perhaps from under £100 to many many £100's

One of the first commercial multi-band antennas that many keen new amateurs come across is something like the Comet CHA250B, or the Diamond BB7V or Moonraker GP2500 (pictured right). These are broad-band antennas and look like large CB antennas with a matching network at the base. Such antennas claim to allow operation of all bands between 80 metres and 6 metres with acceptably low v.s.w.r. Sounds like the perfect multi-band solution, especially as Comet and Moonraker are well known names that make excellent products.

These multi-band antennas have their critics though: Sure enough, they exhibit a seemingly acceptably low v.s.w.r. across the whole of HF, but low v.s.w.r. isn't everything. Ctitics do, in fact, call these types of broadband antennas glorified dummy loads - a bit unfair possibly, but maybe they have a point when most of your transmitter's precious power is wasted as heat rather than radiated as a useful RF signal!

The only way a simple, single vertical radiator can be made to work on across such a wide range of frequencies is by having a broad band matching transformer at the base of the radiator. This will inevitably result in the absorption of some - or much of the transmitter's power - the power loss represented by the heating up of the coils/transformer rather than actually being radiated as a useful signal by the antenna's vertical element.

Such antennas could present a loss of around 6 - 12 dB compared to a resonant antenna - how do do fancy putting all 100 of your precious watts in to the antenna and only getting 6.31 watts of effective power radiated?

Maybe that's a bit simplistic, so Martin G8JNJ has many superb articles analyzing the CHA250B and similar antenna designs here: http://g8jnj.webs.com/cometcha250b.htm

The article Anatomy Of The Comet CHA250B by VK5ZBD can be found here:
http://www.radiomanual.info/schemi/ACC_antenna/Comet_CHA-250BX2_anatomy.pdf
(Formerly found at this site http://www.vk5zdb.com/CHA250BXII.htm)

G8JNJ is also developing a better version of this type of antenna here:
http://g8jnj.webs.com/broadbandhfvertical.htm

I admit that, due to limited space, I considered this type of antenna when first starting out - but in the end dismissed them due to the extreme inefficiency and power loss problems. They should not be entirely discounted however, because if this really is all that can be accommodated at one's QTH then at least such an aerial will at get you on the air - and on all bands - at least in some sort of fashion. Many amateurs use these aerials with success, so they do have a place. Have a look and decide for yourself.

Other similar types of broad-band antenna:

There are a number of very similar designs (i.e. longish vertical radiator, with a transformer / unun at the base) available from some other British suppliers:

The G Pro-Whip 'Widebander' antenna (see http://www.gwhip.co.uk/);
The ProWhip Portable Antenna (see http://www.prowhipantennas.co.uk/);
Snowdonia Radio Company (SRC) - various types of wideband antennas (see http://www.snowdonia-radio-company.co.uk)

All these antennas appear to be based around an UNUN (typically 9:1) matching transformer at the base of the aerial. These aerials cost considerably less than those previously mentioned. The G Pro Whip and Pro Whip Portable offer particularly convenient portable operating opportunities as they are based on one of my favourite methods of antenna support - a long telescopic fibreglass (fishing) pole. Really neat!

For the 'fishing pole' types, essentially there is a vertical radiating wire of about 7 to 10 meters long, a 10m long horizontal counterpoise wire and the 9:1 unun at the base. This makes for a simple and attractive installation proposition (but remember the penalty of power losses) - all these aerials will be easy to install for permanent, semi-permanent use and easily removable or portable operating.

Considering the 10 meter vertical type, the performance on 40 meters (1/4 wave) should be quite reasonable, with reduced performance on other bands.

G-WHIP ANTENNA PRODUCTS
PRO

G-WHIP G Pro Whip antennas (now discontinued)
by Geoff Brown G4ICD
2011/2 - See G-WHIP'S WideBander Antenna as an alternative
http://www.gwhip.co.uk/

Buy (or build) a 9:1 UNUN and Make Your Own:

If you already have a 10m telescopic fibreglass fishing pole and some wire, then you could easily wind a 9:1 unun, or even buy one from suppliers such as G WHip Antenna products for a reasonable cost. So, you could make your own aerial with 10m vertical radiator working against the 10m counterpoise and fed to the ATU via the 9:1 unun at the aerial's base - just for fun, for experimentation, analysis or for permanent installation or portable work. (The telescopic pole must be fibreglass not carbon fibre)

Martin G8JNJ, suggests that a slightly better way to home-brew a broadband HF aerial might be to cut a vertical aerial for about 8.5 MHz, i.e. not a resonant 1/4 wave on any amateur band, but optimised to present a moderate impedance on as many bands as possible. In which case the vertical wire would be about 8.8 metres long, working against the counterpoise, and fed to the a.t.u. via an unun - perhaps 6:1 or 9:1 - this is all open to further research and experimentation! See http://g8jnj.webs.com/currentprojects.htm

G0KYA has also written lots of interesting articles about antennas and several pieces about using a 9:1 unun and a length of wire. He found that a wire length of 19.8 metres offered a good compromise for a multi band aerial. Read G0KYA's blog here: http://g0kya.blogspot.com/search/label/antennas

More : http://g0kya.blogspot.co.uk/search/label/antennas

Interestingly 2W0SAK of Snowdonia Radio Company recommended an antenna wire length of 7.13 metres with their 9:1 unun - or for better efficiency a wire that is 19.8 metres long which would be run out horizontally. Both the 7.13 m and 19.8 metre lengths should cover the 80m to 10m bands.

Freq Coverage	Wire Length	or try
7 - 29 (50) MHz	2.74 m	6 m
3.5 - 29 (50) Mhz	7.13 m	9.8 m or 16.1 m
3.5 - 29 (50) MHz	19.8 m	21.9 m or 26.8 m
1.8 - 29 MHz	29.0 m	29.9 m
1.8 - 29 (50) MHz	39.6 m	36.6 m

High Quality Baluns and UnUns Available From www.gwhip.co.uk - either boxed and ready to use or just the core and quality windings suitable to put into own box or project.

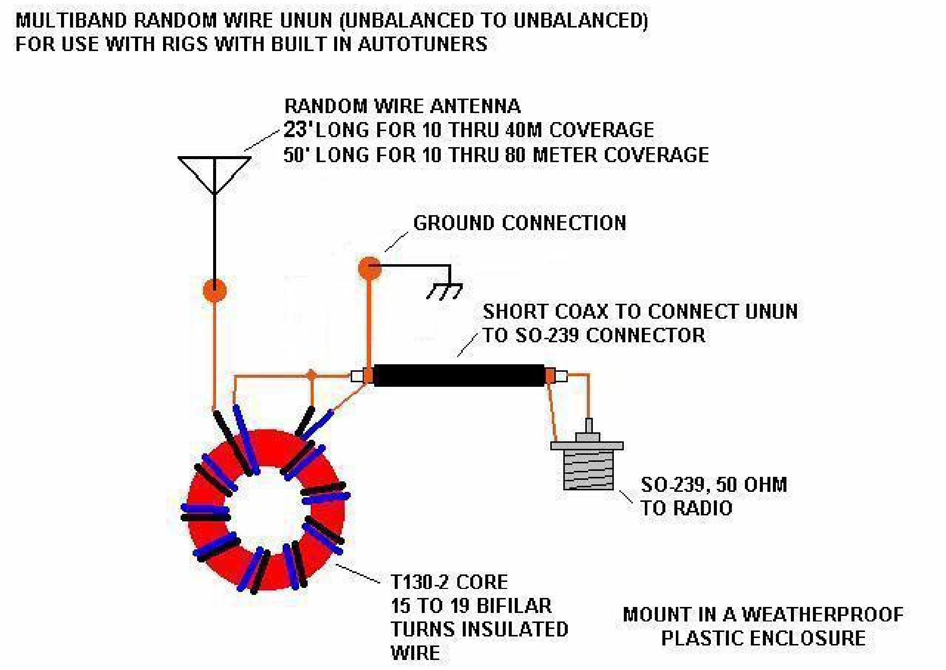
MULTIBAND RANDOM WIRE UNUN (UNBALANCED TO UNBALANCED)
FOR USE WITH RIGS WITH BUILT IN AUTOTUNERS
RANDOM WIRE ANTENNA
23' LONG FOR 10 THRU 40M COVERAGE
50' LONG FOR 10 THRU 80 METER COVERAGE
GROUND CONNECTION
SHORT COAX TO CONNECT UNUN TO SO-239 CONNECTOR
SO-239, 50 OHM TO RADIO
T130-2 CORE
15 TO 19 BIFILAR TURNS INSULATED WIRE
MOUNT IN A WEATHERPROOF PLASTIC ENCLOSURE

Above is a neat design for an "Untenna" KC8AON - Link: http://www.angelfire.com/electronic2/qrp/unun.html

For a jack of all trades try a high quality GWhip "Widebander Antenna"

G0KYA writes a very useful piece in his blog:

http://g0kya.blogspot.com/search/label/antennas

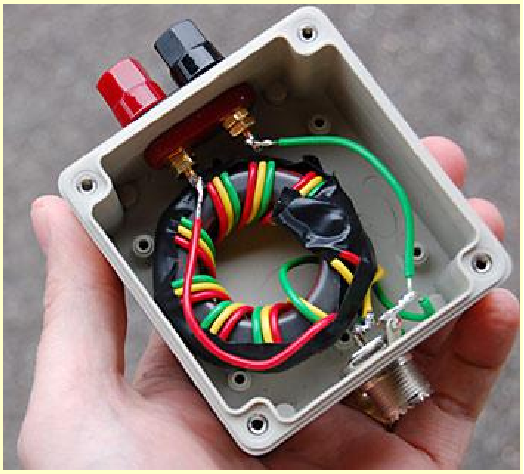

In the next round of tests I used the same 9:1 Un-Un wound using PVC covered cable and a T200-2 toroid. Note in the photograph that the PVC tape is only used to keep the turns neatly arranged.

If you want to build your own follow these instructions:

Building a 9:1 unun
To understand how to construct an unun lets build a 9:1 version. You will need a T-200 (red) toroid and three pieces of wire, each 24 inches (60cm) long . It will also help if you a small plastic box with an SO239 socket mounted at one end and with two wing nuts or mounting posts at the other. In the UK you can buy a small plastic box from Maplin which is watertight with a rubber seal, yet inexpensive.
It will help if the wires are different colours, although that isn't critical if you have a multimeter available. It just makes it a lot easier to follow these instructions.
For the sake of this explanation I'll assume that you are using green, red and black pieces of wire.
Put the three pieces of wire together and wind them carefully onto the T130-2 toroid. Place the wires (left to right) green-black-red, and wrap nine turns on to the toroid.
Try not to let the wires overlap.
You should end up with a toroid with three wires extending from the left winding and three wires extending from the right.
Now twist and solder the left black wire with the right red wire. This can be covered with PVC tape once complete.
Now twist the left green wire with the right black wire. Strip the ends of the two wires, twist and solder them together leaving the length about 2” long from the toroid.
Finally trim and strip the remaining right green wire and solder another 5” piece of green solid wire to it.
Now take the left green wire and right black wires that you twisted together and connect them to the centre pin of the SO239 socket – this is the input side and will connect to your radio via a length of coax.
One of the green wires is now soldered to the ground connection of the SO239 socket. The other end of the wire you soldered on (which is connected to it) becomes the earth connection for the unun and typically goes to a ground stake and ground radials.
This leaves the remaining red wire which connects to the other wingnut and will become the connection for the antenna.
If you are worried about the wires unravelling you can either use PVC tape to hold them in place or plastic cable ties.
So how do we use an unun? Lets look at a typical example.

This time I erected a 10m high fishing pole and attached a 65ft quarter wave antenna for 80m in an inverted L fashion. That is, 10m up and then 9.8m out to the nearby summerhouse.

This was arranged away from the house and fed with 12m of RG8 coax, a single earth stake and two 20ft radials at the feed point..

Here are the SWR readings at the end of the coax:
3.5MHz – SWR 3:1
3.65MHz – SWR 4.2
3.8MHz – SWR 5.9
7.10MHz - SWR 13.6:1
10.1MHz – SWR 2.5:1
14.2MHz – SWR 3.3:1

18.14MHz – SWR 1.8:1
21.2MHz – SWR 2.4:1
24.9MHz – SWR 1.9:1
28.5MHz – SWR 1.2:1

From this you can see that by shortening the wire to 65ft from the original 85ft you gain 80m, but lose 40m. The rig (FT2000) would quite happily tune seven bands with its internal ATU. Here are the quick comparison results against my 80m Windom and parallel-fed dipoles in the loft for 40m, 20m, 17, and 10m.

80m
Not as good around the UK as the Windom - probably due to the maximum current being in the vertical section. Modelling shows the antenna to be down about 10dB on a low dipole.
30m
Lithuania similar. Other EU and Italy similar. Bulgaria down 2 S points
17m
Similar – inverted L has slight edge at times. Slightly noisier
15m
Better than Windom by about 1 S point.
10m
Much better than Windom, dipole and mag loop around Europe via Es, by about 2 S points. Slightly more noise (+ 1 S point).

From this I can see that I need to do more tests, especially on 20m, but for an all-in cost for the antenna of about £15-£20 it shows promise. If you have a tree then the up and out idea with a 65ft wire looks quite good. A way to get 40m back would be to put a 40m trap in the wire at the 10m mark. If you don't fancy making your own UnUn you can buy the whole antenna from the Snowdonia Radio Company for £35 inc P&P – see http://www.snowdonia-radio-company.co.uk/srcproducts.html [From a collection of excellent articles from G0KYA] Link to SRC: http://www.snowdonia-radio-company.co.uk

GWhip Antennas, and ProWhip Antennas all supply this type of antenna as a commercial item.

CONCLUSIONS?

Arguably the most effective, simplest and, indeed, cheapest way to attain multi-band operation is by using a full size resonant dipole for each band of interest - perhaps having a couple suspended at any one time and swapping aerials when other bands are needed.

As mentioned in the introduction this is a little inconvenient which is why the holy grail of so many amateurs is one antenna that that will do everything - perfectly. As we have seen such an aerial does not exist, and never will due to those pesky laws of physics. Compromises will always have to be made; compromises of efficiency, size, number of bands and bandwidth per band etc - nevertheless there are enough options available to be able to choose a configuration of antenna or antennas that should be able to make the best use of precious transmitter power for a particular circumstance.

My two key criteria are that the antenna should be truly resonant on the band(s) of interest and that the radiating elements should be as near to full size as possible, relative to the wavelength(s) being used, to ensure the best possible efficiency (i.e. lowest loss of power). This means full size quarter wave vertical or full size half wavelength long trapped dipole.

I don't especially like aerials that are shortened by using a loading coil, but accept that such an arrangement is sometimes necessary for the longer wavelength bands. Top Band is a real problem in average size gardens. Where the is a coil, a trap, or transformer there will be some loss or reduced efficiency introduced into the antenna system. I do find that using a trap is an excellent compromise - the 80m / 40m Inverted L and the 20m / 10m trapped dipole work especially well. If I could not use the Inverted L, my next favourite option is the Loop or a fan (Parallel) dipole.

All the pros and cons have to be weighed up to find the best compromise for particular operating circumstances. I hope that this page has given new operators some ideas to take away and mull over, but remember those words from Joe Tyburczy WB1GFH that this page started with:

"When you put up your antenna is also crucial. I must mention here the importance of what many early hams called "antenna weather". That is, snow, sleet, freezing rain, or combination of all the above. It has been proven time and time again that any antenna installed in conditions better than abysmal will not function worth a darn. Or, put another way, it takes bad weather to put up a decent antenna. Dark and cold New England winter days are ideal for this activity. Any antenna erected on such a day will inevitably produce miracles."

Some Further Reading:

Understanding Antennas For The Non-Technical Ham
A Book By Jim Abercrombie, N4JA
basicantennas.pdf

More Antenna Ideas by other amateurs (.doc)

ARRL document Multi Band Dipoles Compared (.pdf)

High Quality Baluns and UnUns Available From www.gwhip.co.uk

Useful Aerial Rigging Accessories

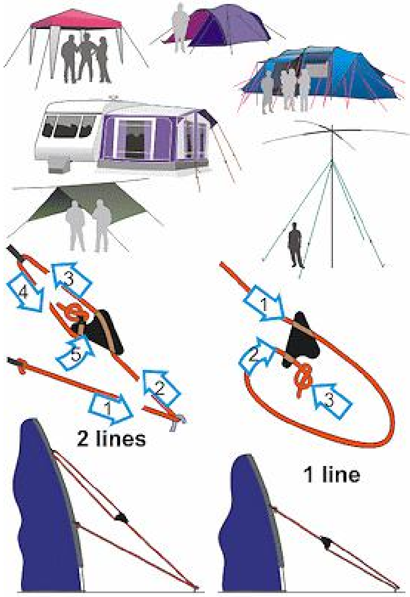

Line-Lok guy runners from ClamCleats - fantastic for guying antenna masts quickly and successfully
http://www.cleats.co.uk http://www.clamcleat.com

Reusable Nylon Hose Clip / Reusable Circular Clamps
Useful for securing telescopic fibreglass poles - e.g. holding each section in place in windy weather or to use as a guying ring (spider)
(Herbie Clips) (Kaf-flex Nylon Clamps)
http://www.malpasonline.co.uk
http://news.thomasnet.com
http://www.hclfasteners.co.uk/acatalog/Herbie-Clip.html
http://www.cheapham.com/products/S9V31-Replacement-Clamps.html
http://rotocon.homestead.com/shoponline2.html

OTHER THINGS THAT MAY BE NEEDED:

POLE(S) Aluminium, fibreglass or wood ; POLE TO POLE CLAMPS ; MET POST(S) ; NYLON CORD or PARA CORD ; SPIDERS / 3 or 4 WAY GUY RINGS ; PULLEYS ; SNAP HOOKS ; DEE SHACKELS ; GROUND STAKES for anchoring guy ropes ; DOG BONE or EGG INSULATORS ; DIPOLE CENTRES ; EARTHING STAKE ; V BOLTS ; ROPE GRIPS ; THIMBLES ; SLEEVE JOINER(S) ; T&K BRACKETS ; TRIPOD or other GROUND MOUNTING HARDWARE ; STAINLESS STEEL M6 Nuts Bolts and Washers ; SPADE and / LUG TERMINALS ; SELF AMALGAMATING TAPE ; HEATSHRINK ; WEATHERPROOFING SEALANT ; STAINLESS STEEL JUBILEE CLIPS.

TRIMMING AERIALS

Antenna Trimming Chart

This following information below could be very useful indeed when constructing aerials and is compiled by DX Zone / Radio Works from the web page at: http://www.dxzone.com/cgi-bin/dir/jump2.cgi?ID=13444

Use this chart as an aid in trimming the length of your antenna. It gives you an idea of the change in wire length needed to move antenna resonance a specific number of KHz.

* Dimensions are for each leg of a half-wave dipole
* For quarter-wave antennas (i.e. verticals) use the dimensions directly from this chart
* Full-wavelength antennas (loops) - multiply the chart dimensions by four (4) and change the overall length of the antenna by that amount.

Lengths are estimates. Many factors will affect their exact value.

To move	80/75 m	40 m	20 m	15 m	10 m
-500 kHz	+8' 4"	+2'	+8"	+3"	+1.5"
-400 kHz	+6' 8"	+1' 9"	+6.5"	+2.5"	+1.25'
-300 kHz	+5'	+1' 4"	+5"	+1.75"	+1"
-200 kHz	+3' 4"	+10"	+3.25"	+1.25	+5/8"
-100 kHz	+1' 7"	+5"	+1.5"	+1/2"	+3/8
00 kHz	0	0	0	0	0
+100 kHz	-1' 7"	-5"	-1.5"	-1/2"	-3/8"
+200 kHz	-3' 4"	-10"	-3.25"	-1.25	-5/8"
+300 kHz	-5'	-1' 4"	-5"	-1.75"	-1"
+400 kHz	-6' 8"	-1' 9"	-6.5"	-2.5"	-1.25'
+500 kHz	-8' 4"	-2'	-8"	-3"	-1.5"

Example:

You have measured the SWR of your 40 meter dipole at various frequencies across the band. You have determined that the SWR is lowest at 7.00 MHz. You actually want the lowest SWR to occur up in the sideband portion of the band, so you need to move resonance up in frequency about 200 KHz.

According to the chart, to move +200 KHz on 40 meters, you will have to shorten each leg of the dipole 10" (-10"). The overall length of the antenna is shortened a total of 20 inches.

Lengthening or shortening the antenna is done at the end insulators. To shorten the antenna, unwind the antenna wire as it wraps around itself at the end insulator. Move the insulator several inches toward the center of the antenna. Re wrap the antenna wire to secure the end insulator. Do not cut the wire. Wrap it back around the antenna wire. You may need to increase the antenna later. [From Radio Works / DXZone]

KNOTS FOR SECURING WIRE ANTENNAS

I have found the Bowline to be one of the most useful, it is strong and easy to tie. A Bowline will not slip in any circumstances and, usefully, the more load that is put on it, the tighter it gets.

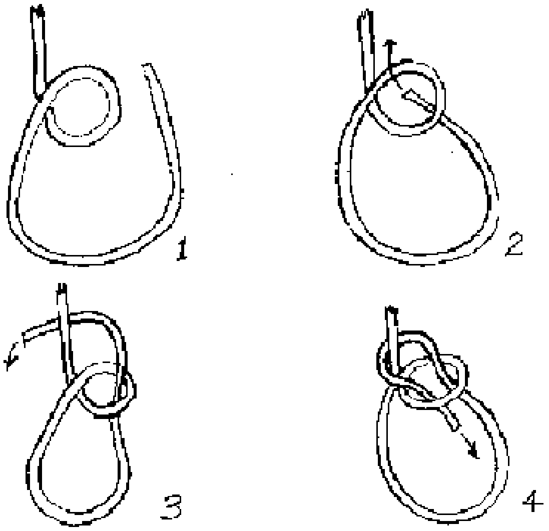

The Bowline Knot

A Bowline can be used to tie two ropes together and should be used to tie a support rope to a pulley, dipole centre and other antenna items.

It's important to use the correct knot for the job when fixing up wire antennas. I find the Bowline is a very useful for fixing end, egg and dog-bone insulators to the ends of the wire and/or ropes.

The Buntline Hitch is an excellent knot as is the Round Turn & Two Hitches, Anchor Bend (Anchor Hitch) knots which are very good for tying a rope to a pole or a mast.

The Bowline is most useful for fixing end, egg and dog-bone insulators to the ends of the wire and/or ropes.

A Double Sheet Bend can join two pieces of rope together - even if they are of unequal size.

'Animated Knots' will show you how to do them. Visit http://www.animatedknots.com

More websites with knot information : http://www.netknots.com/ http://www.southee.com/Knots/Index.htm

The correct knot will ensure that the antenna will be as strong as possible.

LINKS

Understanding Antennas For The Non-Technical Ham - A Book By Jim Abercrombie, N4JA :
PDF Book: http://www.hamuniverse.com/basicantennas.pdf
HTML page: http://www.hamuniverse.com/n4jaantennabook.html

G4ILO - Stealth Antennas: http://www.g4ilo.com/stealth.html

All Band Doublet Antenna by M0MTJ :
http://www.mds975.co.uk/Content/amateur_radio_antennas_06.html#All_Band_Doublet_Antenna

All Band Doublet Antenna : http://www.hamuniverse.com/hfdoublet.html

All Band Doublet Antenna by AI4JI : http://www.ai4ji.com/Projects/antennas/doublet.htm

My Home Page

Amateur Radio Operating Conditions
Antennas
Accessories
Projects & Kits
Data Modes
Information
Portable
QSL
RSGB Ltd
Links
QRM
Bits 'n' Bobs
Contact

WSPR
Weak Signal Propagation Reporting

PORTABLE

Site Map

MDS975 Home

Feedback

Site Search
Find
powered by FreeFind

QRM from PLT

G Whip Antenna Products

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits | Useful Information
Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites | RSGB | QSL | The Amateur Radio Mini Site Map
News, Developments, Events, Photographs and other 'Bits 'n' Bobs' | WSPR Weak Signal Propagation Reporter

AERIALS (ANTENNAS) 1

Antennas 2 | Antennas 3 | Antennas 4 | Antennas 5 | Antennas 6 | Antennas 7

"Success is 90% antenna and 10% rig. Hobby is 90% listening 10% transmitting" *- MM0HDW.*

AERIALS used by MØMTJ

This page shows some the antennas that I have used over the course of time.

Index To Other Antenna Pages:
Antennas 1 : Aerials used by M0MTJ
Antennas 2 : Including ideas for compact aerials for Top Band /160 metres
Antennas 3 : Felix Scerri VK4FUQ discusses Loop Aerials, baluns, masts & other antenna related topics
Antennas 4 : Many antenna ideas from various sources particularly for multi-band operation & also gives information about
antenna trimming, knots for wire antennas and useful aerial rigging accessory ideas.
Antennas 5 : Half Wave (physically end fed) aerials for 144 MHz VHF / 430 MHz UHF and 50 MHz 6 Metre band & J-Pole Aerials
Antennas 6 : Simple and effective H.F. Aerial ideas: The All Band Doublet, an All Band Sloper & a Ground Plane Aerial
Antennas 7 : Omni-Directional - Circularly (Mixed) Polarized Aerial for VHF / 2 Meters.

2014 : Current Set Up - as at September 2014:

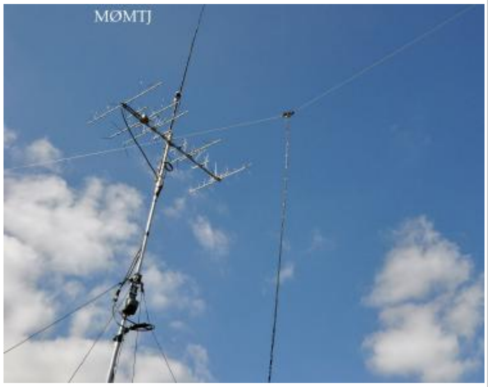

The Summer of 2014 has seen a few changes of aerials. The 80m/40m Inverted L was removed and replaced with a newly constructed Doublet Antenna fed with ladder line via a 4:1 G-Whip current balun and an LDG "ATU". The Tecadi support pole originally used for the Inverted L is now used to support one end of the Doublet. You have probably read this elsewhere, but I can confirm that the All Band Doublet (fed with ladder line *not coax*) is a superb all round HF aerial. Highly recommended.

The SGC-230 Automatic Antenna Coupler was moved from the original feedpoint of the Inverted L antenna to the second feed-point on the other side of the garden. This now feeds a sloping wire of approximately 20 meters length to ensure that the 80 Meter Band and Top Band (160 meters) remain available. This antenna will also 'tune' on most other bands, so remains a very useful second antenna, though the All Band Doublet is often considerably better, particularly on the higher frequencies. The old SGC-230 was also faulty and I replaced it with a new CG-3000 auto coupler from Martin Lynch and Sons.

The J-Pole for 4 Meters (70 MHz) was removed since it wasn't used much, being only connected to a 4m hand-held radio - it seemed to be a waste of precious antenna space. This was replaced with a newly constructed Half Wave "CFR" Antenna (Coaxial Dipole) for 6 Meters (50 MHz) connected to the main HF radio via Westflex 103 coax. The DK7ZB dual band Yagi antenna remains, as does the 10m / 6m Fan Dipole in the loft space. Also remaining in place at the apex of the house is the home-brew 2m / 70cm "CFR" Antenna (Coaxial Dipole) for VHF/UHF FM operation - this has been a particularly effective antenna.

To sumarise:

1) A 20 meter long doublet antenna fed with balanced ladder line for 40 meters to 6 meters. The 'old timers' really knew their stuff, this an excellent all round antenna that is easy and cheap to make and should be quite easy to accommodate and install. See photograph below and read more here

2) A Sloper Antenna consisting of two parallel wires - one wire being 20 meters long with a second parallel wire

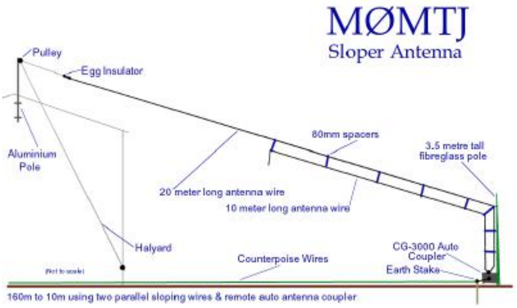
MØMTJ
Sloper Antenna
Pulley
Egg Insulator
Aluminium Pole
80mm spacers
3.5 metre tall fibreglass pole
20 meter long antenna wire
10 meter long antenna wire
Halyard
Counterpoise Wires
CG-3000 Auto Coupler
Earth Stake
(Not to scale)
160m to 10m using two parallel sloping wires & remote auto antenna coupler

that is 10 meters in length. The first 2.5 meters (approx) of wire runs vertically up a non metallic post, then the remainder slopes back towards the house being finished off with an insulator which is attached to a Para-Cord lanyard that runs through a pulley on a pole at the apex of the roof. The antenna wire can therefore easily be let down for maintenance or adjustment whenever required. For lowest possible loss, the antenna is fed via a GC3000 automatic antenna coupler located at the bottom of the garden. The antenna is primarily for lower HF bands including 160 meters, 80 meters and and 40 meters, but it also works very well on some higher bands. The GC3000 itself is mounted in a waterproof IP56 rated enclosure near ground level at the bottom of the garden. Grounding is achieved from two 4 foot long copper ground stakes and several radial grounding wires.

Read more about the Sloper Antenna and the CG3000 here.

3) A 'home-brew' omnidirectional, vertical dual band, end fed antenna for 2 metres and 70cm. This is of the Controlled Feeder Radiation design (CFR) by VK2ZOI; effectively an end fed half wave dipole on 2m with an aluminium sleeve dipole section to achieve 70cms with a few extra dB's of gain. It is mounted on an aluminium mast. It's feed-point is about 11 metres a.g.l. read more here

4) A DK7ZB design dual band Yagi antenna, with 5 elements for 2 metres and 8 elements for 70cms, mounted horizontally for SSB. A lightweight antenna rotator is employed and uses a push-up telescopic mast. Height above ground level is again approximately 7 metres. The DK7ZB is an excellent twin band Yagi antenna. read more here

5) A Half Wave End Fed / Controlled Feeder Radiation (CFR) antenna for 6 meters / 50 MHz. Supported by a 3 meter long telescopic fibreglass fishing pole attached to the top of the aluminium push up mast that supports the DK7ZB dual band yagi and rotator.

6) Dual Band Fan Dipole, made from thick loudspeaker wire, mounted horizontally in the loft space for 10 meters an 6 metres. Cheap & quite effective.

Other Options that can be deployed on an 'as required basis' :

7) A half wave Wire J-Pole fixed to a telescopic fibreglass fishing pole for 10m. Cheap and effective. more about J-Poles here and also here on Antenna page 5

8) Compact Loaded Top Band Antenna, based on a design idea by Stuart Craigen G4GTX more

9 & 10) G Whip End Fed Zepps (EFZ's) for either 20m, 15m or 17m or the G-Whip "WideBander" which is an 'UnTenna' style antenna that can be used for 20m through to 10m using good quality G Whip 9:1 UnUn; useful additions for antenna flexibility. more

11) N9TAX Dual Band Slim Jim (J-Pole) antenna mounted in the loft as a back-up antenna for 2m and 70cms. Very good. more

12) Delta Loop Antenna - 16 metre loop of wire in triangular Delta shape, hung from the top of the pole supporting the inverted L antenna and fed via RG213 coaxial cable via a 4:1 balun. The loop is really a single band antenna cut for one wavelength on the band of interest, however it also can be pressed into service for some higher bands - a good, cheap and easy to install aerial; Often works better than the inverted L on the higher bands, but on 10 metres the tuned 10 metre dipole in the loft is sometimes better. more

Knots: Knots for securing wire aerials and other things more here

H.F. ANTENNAS used by M0MTJ

All Band Doublet Antenna

The Doublet Antenna consists of two 10 meter long top wires to form the 20 meter long 'dipole' section. The centre is fed with Ladder Line rather than coaxial cable. A dipole fed with coaxial cable is essentially a single band (mono band) antenna. Feeding such an aerial with ladder line, or open wire twin feeder makes a much more effective multi-band antenna.

The ladder line runs down to a high efficiency 4:1 Current Balun (G-Whip) which is connected to an LDG AT-200 automatic antenna matching unit via a very short RG213 patch lead to ensure lowest losses. The Antenna tuner and balun are housed in a box which is itself contained in a small garden shed to protect it from the weather. The LDG antenna matching unit is then connected back to the 'shack' via a run of RG-213 coaxial cable. Read more

Get on the air with TOP QUALITY BRITISH G-WHIP ANTENNAS

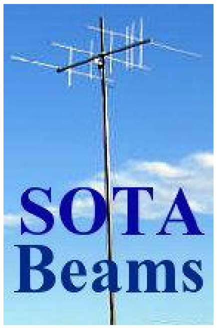

SOTA Beams Lightweight 2 metre & 70cms Yagis, Dipole, Accessories & Poles www.sotabeams.co.uk

about the All Band Doublet Antenna here.

A view of part of the MØMTJ All Band Doublet Antenna
Read more here

ANTENNAS FOR VHF and UHF - 2 m & 70 cms

The main antennas are as follows:

Home-Brew dual band end (physically end fed) half wave "Coaxial Dipole" for 2m & 70cm

For 2m and 70cm FM I use a mounted on a lightweight aluminium telescopic pole on the apex of the hose. The base of the antenna (the bottom of the radiating element) is approximately 11 metres above ground level. This antenna is based on the Controlled Feeder Radiation principle (CFR) and is described by VK2ZOI. Read more about its construction here. Also seen in the photograph below are the ropes that support the H.F. wire aerials.

Home brew dual band vertical antenna for 2 metres and 70 cms
Read more about its construction here

Dual Band Yagi for 2m & 70cm

For 2 metres and 70cms SSB there is a horizontal DK7ZB design dual band Yagi antenna. This has 5 elements for 2 metres and 8 elements for 70 cms. A lightweight antenna rotator is employed and uses the same push-up telescopic mast that the Home-Brew 70MHz J-Pole is mounted on. Height above ground level is again approximately 7 metres. The DK7ZB is an excellent twin band Yagi antenna**.**

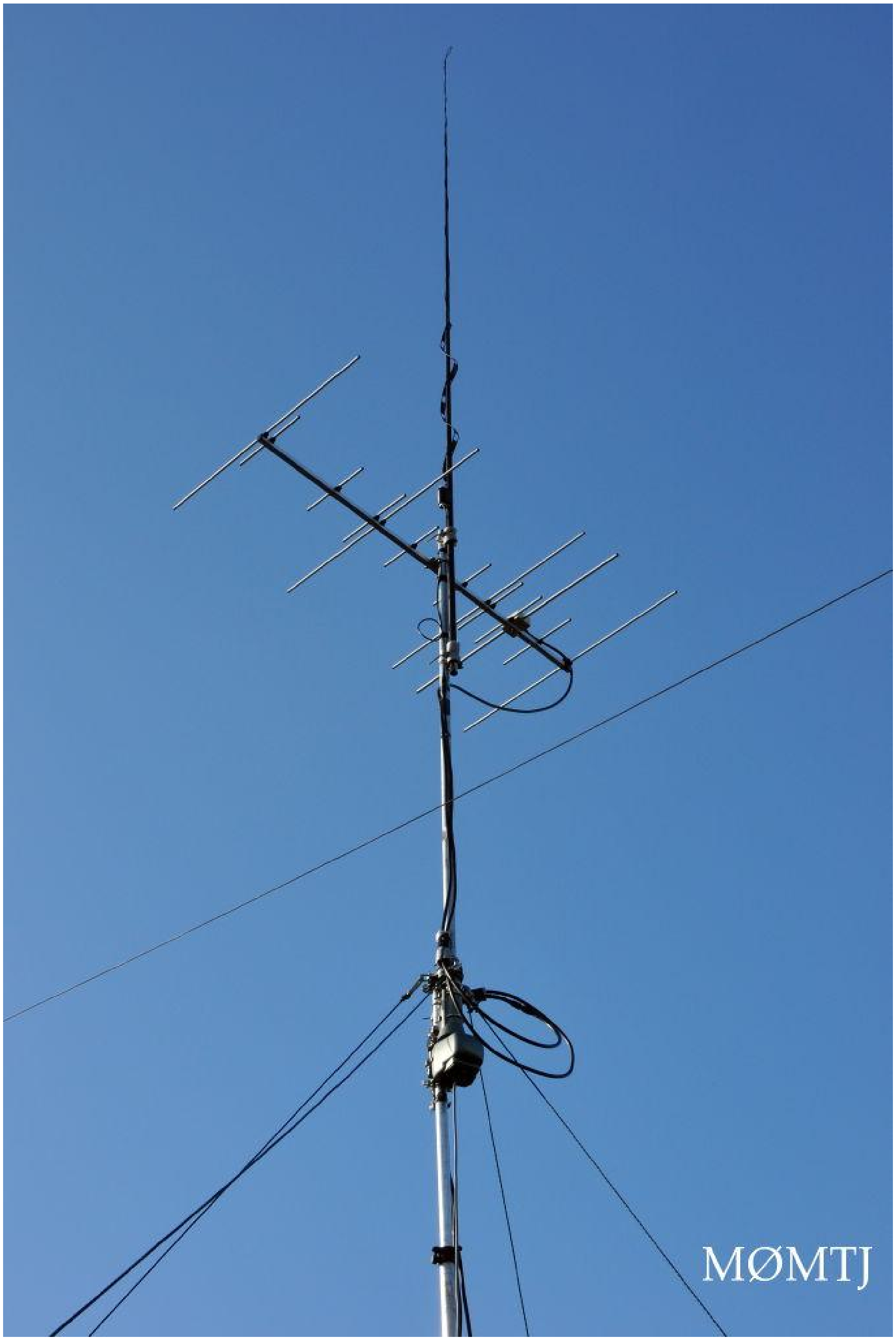

Push up aluminium mast with rotator, 2m / 70cm DK7ZB Yagi and a 4m J-Pole at the top

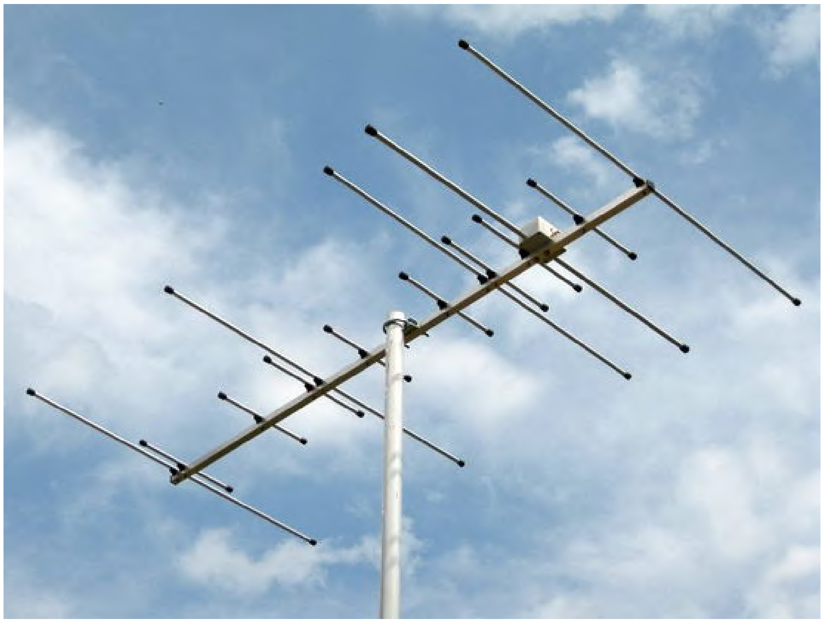

The DK7ZB 5 + 8 element dual band yagi for 2m and 70 cm - designed by Martin Steyer DK7ZB
Available from Arthur M0PLK (SQ2PLK) at Ham Radio Shop:
http://stores.ebay.co.uk/urbasket-eu
http://ham-radio.urbasket.eu
Also available from VPA SYSTEMS by SQ9VPA http://www.vpa-systems.pl
Kits available from NUXCOM.de : http://shop.nuxcom.de

2 Meter J-Pole Antenna for the garden shed - and other J-Pole antennas for 6 metres and 10 metres :

While experimenting with antennas in the garden in the summer of 2012 I thought that it would be good to have a hand-held radio in the shed to do some monitoring and make a few contacts. To improve upon the performance of the 'rubber duck' antenna I quickly made a J-Pole antenna for the 2 metre band.

It is made from a 47cm length of 450 ohm Wireman ladder line as the 1/4 wave matching section, plus a 97cm length of stranded wire as the 1/2 wave radiator. It is fed with 3 metres of Mil spec RG58 c/u coaxial cable that is soldered to the 1/4 wave matching section's impedance matching point at 3.5 cm from the bottom. The coax feeder is wound around some PVC tube to form a choke. The completed antenna is taped to a 2.2 metre long fibreglass fishing pole that I purchased from Poundland (for £1.00). It took about 20 minutes to make followed by some testing and adjustment with the antenna analyser. The fishing pole is lashed to the shed with some cable ties.

This simple antenna works pretty well, but being so low down signal strengths are not huge, but it's pleasing to get on the air with something so simple and cheap!

Find out how easy it is to construct J-Pole Antennas here

Now, if it was at the top of my 10 metre long fishing pole. . . . . (!)

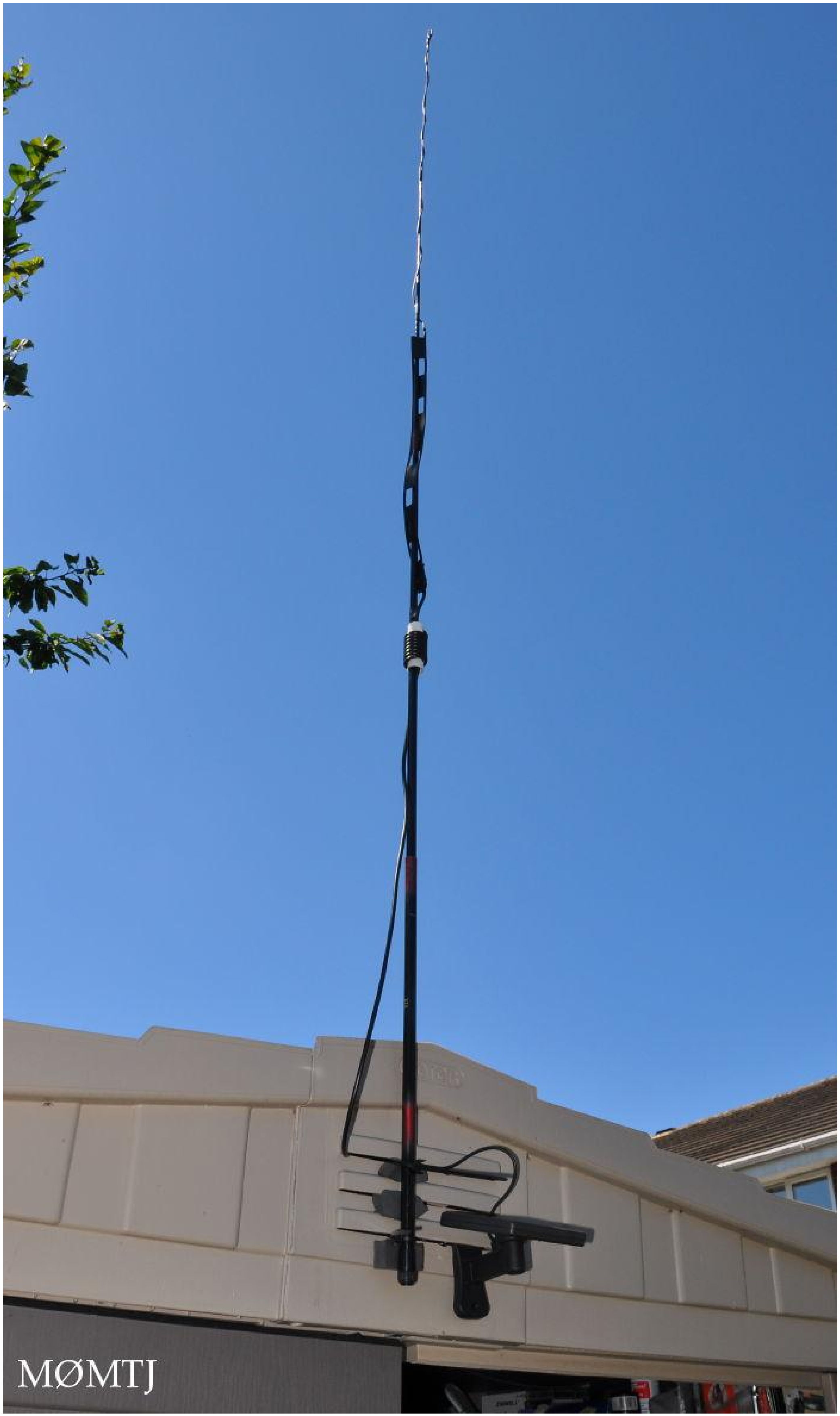

The Shed Antenna - a 2m J-Pole by M0MTJ
Note the simple choke balun at its base made by winding 8 turns
of the coaxial cable around a small off cut of white PVC water pipe.
Find out how easy it is to construct J-Pole Antennas here

www.mds975.co.uk

The feed point of a J-Pole antenna made from Wireman 450 ohm ladder line.
More here

More Antennas . . .

Inverted L Antenna for 80m and 40m (and some other HF Bands from 80m to 10m)

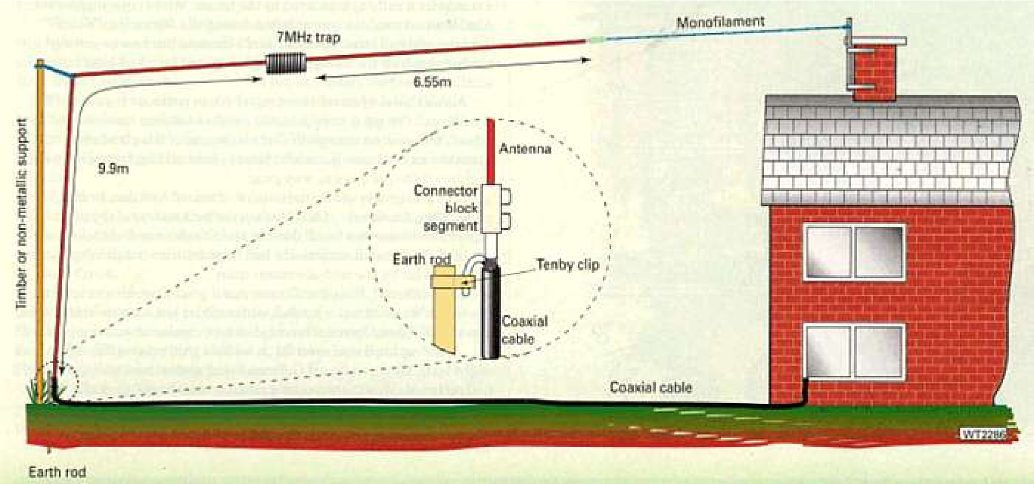

The basic layout of the Inverted L Antenna (Practical Wireless)

The first antenna that I installed was for HF. I decided on an Inverted L that incorporates a 7MHz trap so that it can be used on both 7MHz (40 metres) and 3.5 MHz (80 metres).

The design of this Inverted L is well known and a good design has been published previously in Practical Wireless by Len Paget GM0ONX. It is based on one half of the famous W3DZZ trapped dipole antenna.

It can be made entirely from scratch as a DIY project, or the 7MHz trap could be purchased commercially as a ready made item, or whole antenna can bought as a complete kit from Tony Nailer, G4CFY, at Spectrum Communications. I opted to buy the 7MHz trap from Spectrum Communications, as I already had most of the other materials required - rope, egg insulator, plastic box, and some good aerial wire. The Spectrum Communications trap is solid and well made and 'potted' to protect against the elements.

This antenna is tuned for 40 metres and 80 metres, but the VSWR is acceptable on several other bands being in the region of 2:1 to 5:1. The designer anticipated that this antenna would be usable on five of the H.F. bands between 80m and 10m.

I have found that with the use of the Antenna Tuning Unit it can be used on all of the H.F. bands. However the polar radiation pattern may very well be less predictable on bands other than the intended 40 and 80 metres, and it may well be less effective than might be desirable - but it does work!

The antenna is in the back garden, while the shack (radio room) is in a bedroom at the front of the house. It is fed by a 30 metre length of RG213 coaxial cable (it is not possible to use twin feeder for this type of antenna as the Inverted L is an UN-Balanced aerial, whereas twin feeder is balanced). With this length of cable I estimate the loss in the feeder alone to be about 1dB at 7MHz. The feed point of the aerial is located at the base of a 16 foot high wooden pole near the bottom of the garden. The horizontal top wire returns to a fibreglass pole installed at the apex of the roof.

+ 17 + 160: I have since added a separate sloping wire element for the 17 Metre Band and a switchable loading coil for Top Band - see notes below...

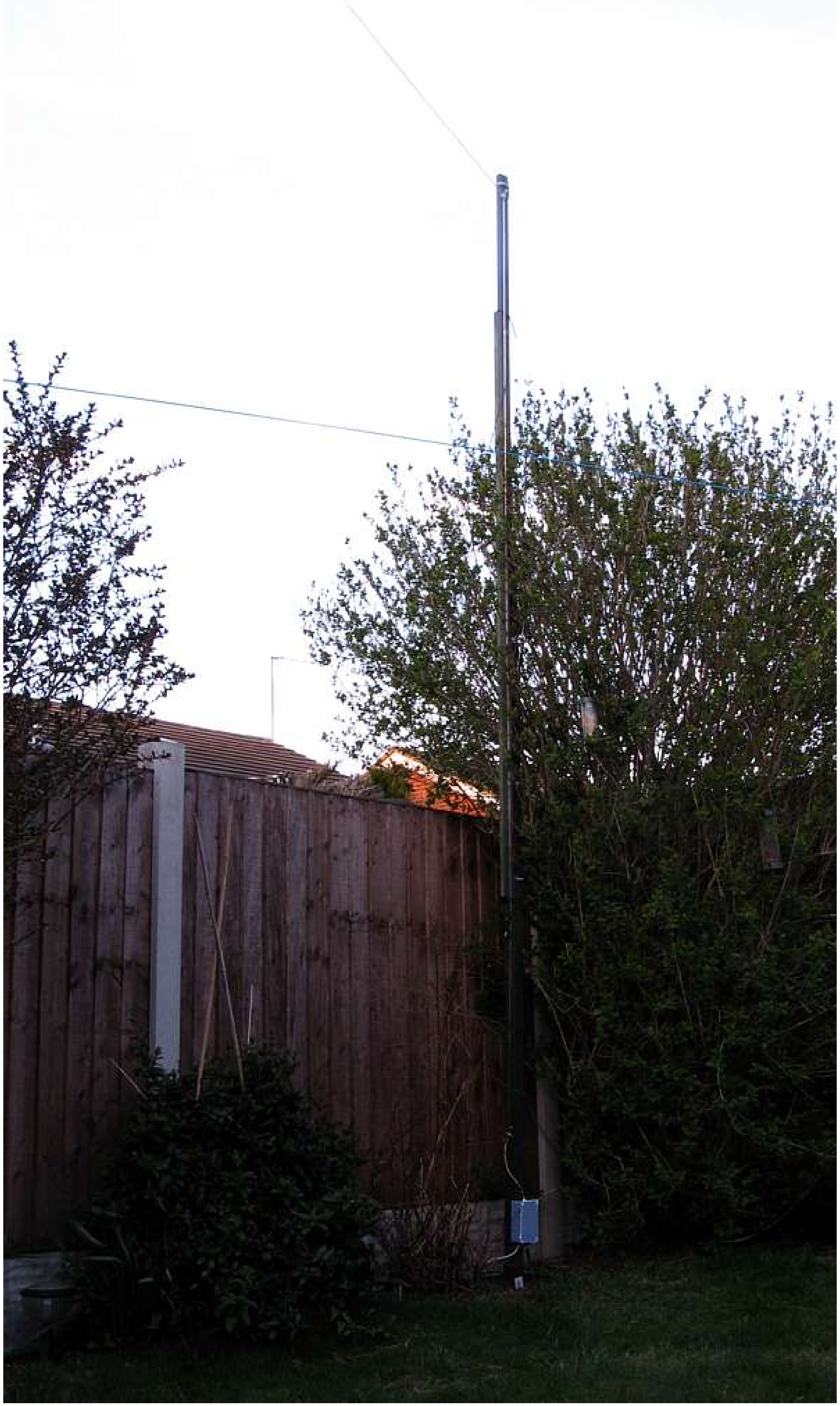

Photograph showing the wooden support post and termination point of the Inverted L antenna The post is coloured green with green fence treatment to mast it somewhat. I have also placed it near the bush to provide further camouflage.

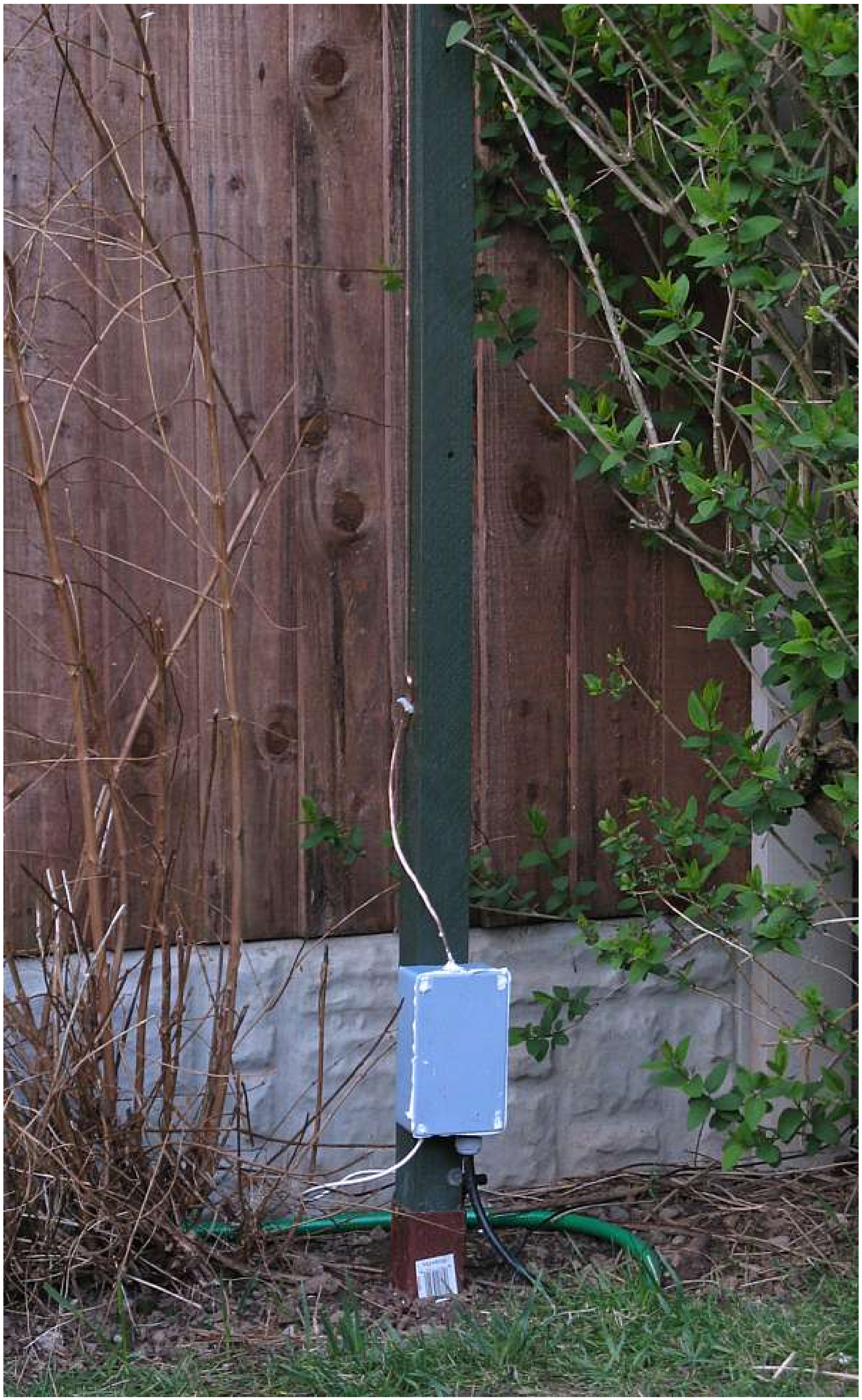

The low loss RG213 coaxial cable runs from the shack at the front of the house up into the loft and exits into the back garden down the rear wall, through some garden hose to offer some protection along the flower bed to the bottom of the garden where it is connected to the base of the Inverted L antenna.

The suspended Inverted L aerial with 7 MHz trap.

The Inverted L antenna - lower section now nicely camouflaged. The wooden support post is some 6 metres long.

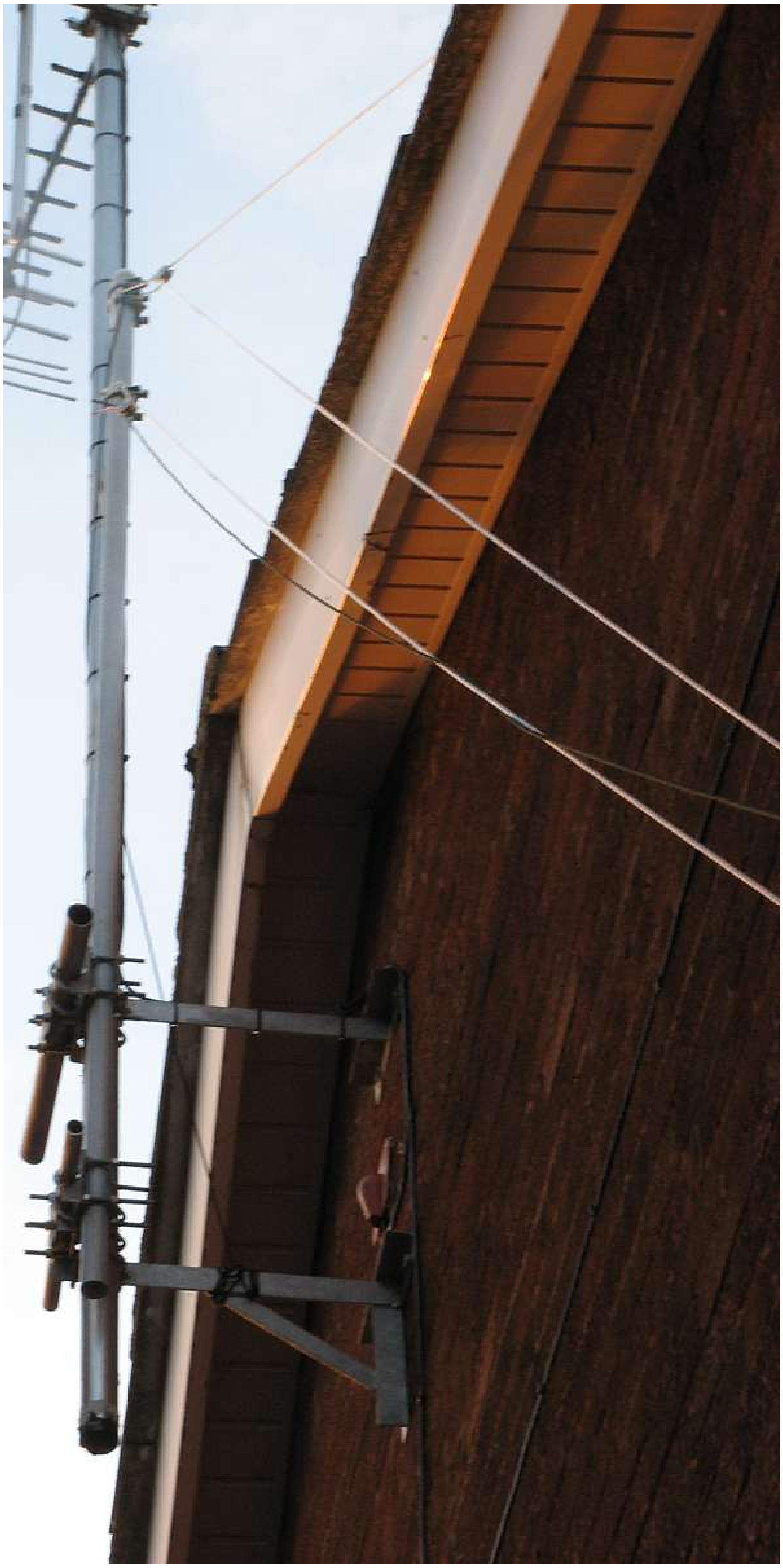

The photograph above shows the Dacron ropes supporting the ends of the Inverted L and Dipole antennas are held in place at the top of the fibreglass support mast by a pulley - one pulley for each support rope. This facilitates rapid lowering of either antenna for adjustment or replacement. This photograph also shows a second rope and pulley system that was originally used to support the 20m dipole and is now used for the top band inverted L wire aerial.

I needed a good reliable knot for securing ropes when installing wire antennas and have found the Bowline to be one of the most useful, it is strong and easy to tie. A Bowline will not slip in any circumstances and, usefully, the more load that is put on it, the tighter it gets.

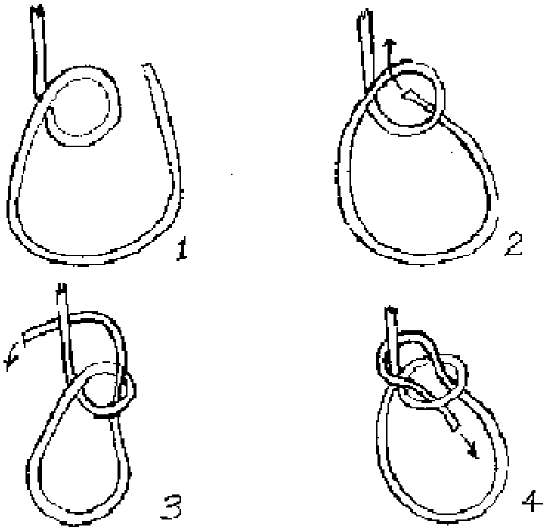

The Bowline Knot

A Bowline can be used to tie two ropes together and should be used to tie a support rope to a pulley, dipole centre and other antenna items.

It's important to use the correct knot for the job when fixing up wire antennas. I find the Bowline is a very useful for fixing end, egg and dog-bone insulators to the ends of the wire and/or ropes. The Round Turn & Two Hitches, Anchor Bend (Anchor Hitch) and Buntline Hitch knots are very good for tying a rope to a pole or a mast. A Double Sheet Bend can join two pieces of rope together - even if they are of unequal size. 'Animated Knots' will show you how to do them: http://www.animatedknots.com

Photo showing pulley fixed to the top of the wooden support post and the aerial support rope that it holds in place.

The Spectrum Communications Trap

View of trap showing that the joints have been thoroughly sealed against the weather with self amalgamating tape and silicone sealant.

Just for good measure I sealed the internal side of the machine screw that forms the connection terminal against the weather with Watson sealant putty.

Adding Top Band to the 80 / 40 metre Inverted L Antenna:

Due to an aborted house move in 2010 I had removed all the antennas. While re-establishing the aerials in 2011, and considering space limitations, I decided to experiment with adding a loading coil to the 40m / 80m Inverted L aerial. The loading coil has a link wire to short it out of circuit when using the aerial for 80 metres and 40 metres. The link wire is removed when Top Band is required.

I also took the opportunity to replace the original wooden post with a strong 6 metre tall fibreglass pole.

The coil consists of approximately 37 turns of PVC covered antenna wire wound on a short piece of PVC pipe. Once the required points of resonance were set for 40 metres and 80 metres, the link wire was removed and number of turns on the coil were adjusted until the required point of resonance was found on the 160 metre band. I set it to around 1900 kHz - the bandwidth is quite narrow.

Once the work was done, the joints and connections were sealed with either Liquid Electrical Tape or self amalgamating tape, then the connecting box, V bolts and white PVC pipe were sprayed with green paint to help it all blend in with the surroundings a little better.

Adding the 17 Metre (18 MHz) Band to the 80 / 40 metre Inverted L Antenna:

The Inverted L is not too good for the 'WARC' bands so to obtain better performance on the 17 Metre band I added a single slightly sloping wire element cut for that band. The lower end of the wire is permanently connected to the feed terminal on the junction box, the other end is tied to a small dog bone insulator. This dog bone is then supported by a length of thin para-cord which is tied to the horizontal wire of the main Inverted L element. (N.B. The 17 metre modification is not currently shown in the photographs below.)

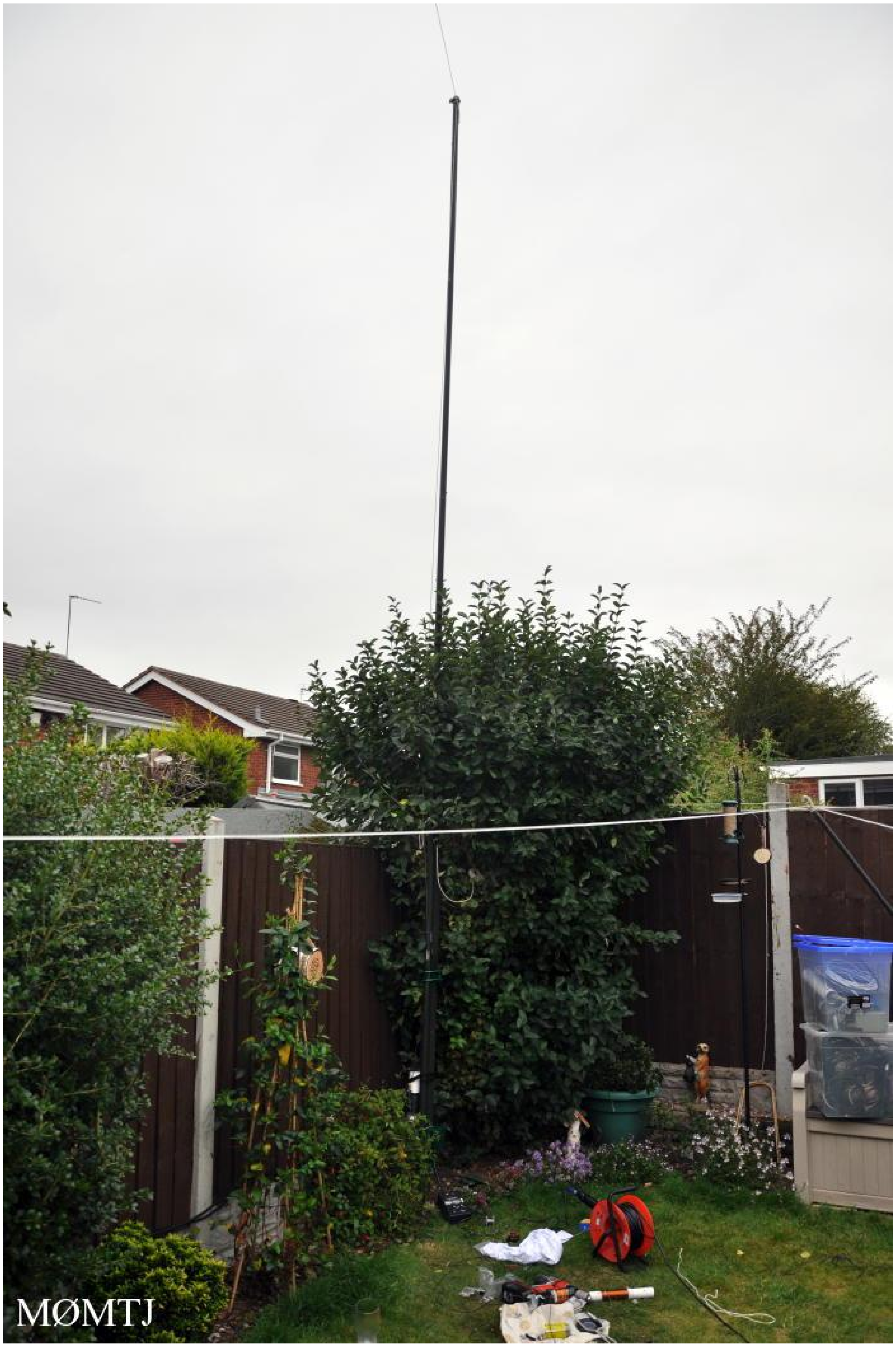

Work In Progress! - September 2011
Reconfiguring Inverted L with additional Top Band Loading Coil for 160 metres.
A strong 6 metre tall fibreglass pole replaces the original heavy wooden post.

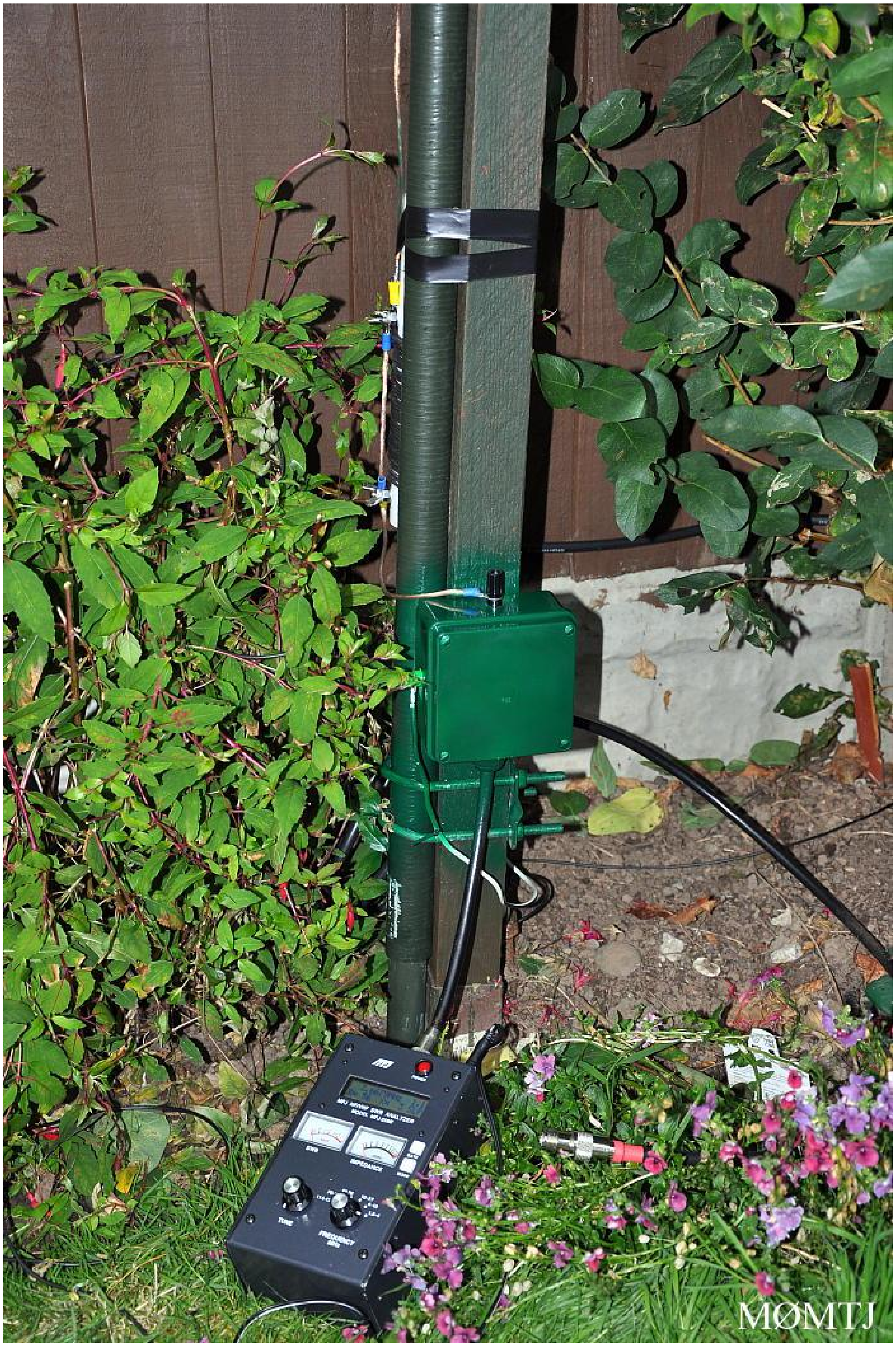

Adding 160 metre loading coil to the 80m / 40m Inverted L Aerial.
The loading coil has a link wire to short it out of circuit when using the aerial for 80 metres and 40 metres.
The link wire is removed when Top Band is required.
The coil consists of approximately 37 turns wound on a piece of PVC pipe.

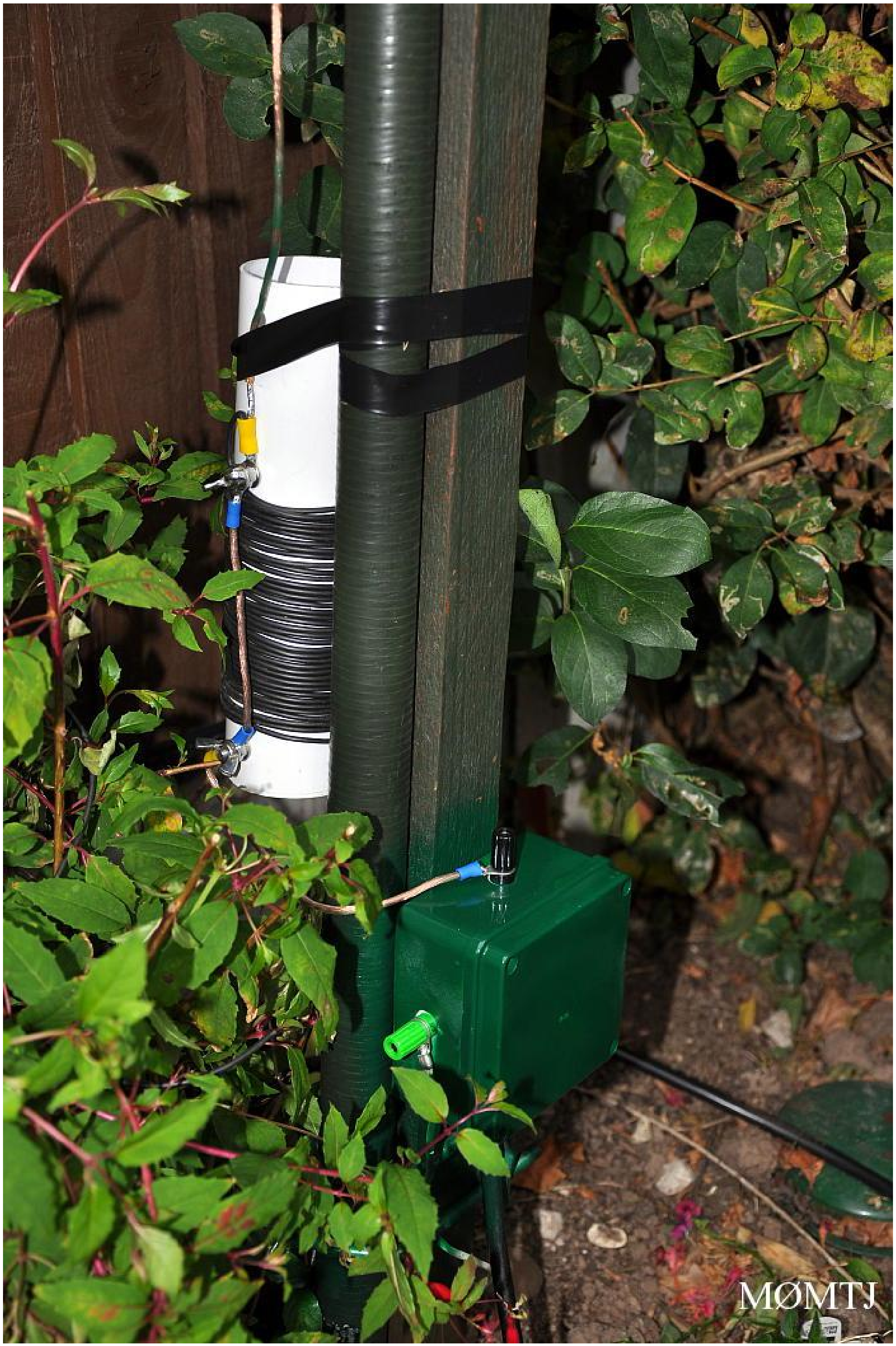

Adding 160 metre loading coil to the 80m / 40m Inverted L Aerial.
The loading coil has a link wire to short it out of circuit when using the aerial for 80 metres and 40 metres.
The link wire is removed when Top Band is required.
The coil consists of approximately 37 turns wound on a piece of PVC pipe.

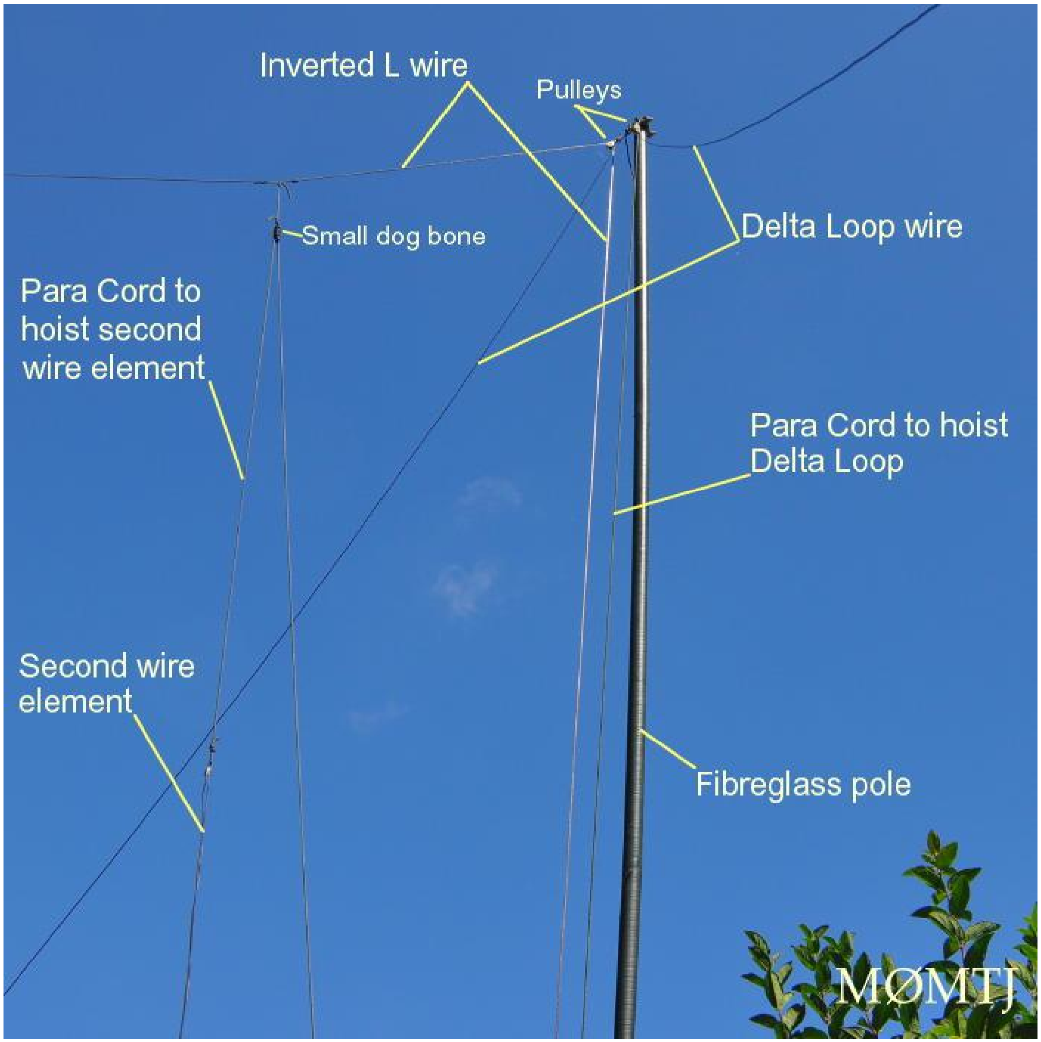
Inverted L wire
Pulleys
Small dog bone
Delta Loop wire
Para Cord to hoist second wire element
Para Cord to hoist Delta Loop
Second wire element
Fibreglass pole
MØMTJ

Photograph showing the Inverted L antenna with additional vertical wire element and position of Delta Loop

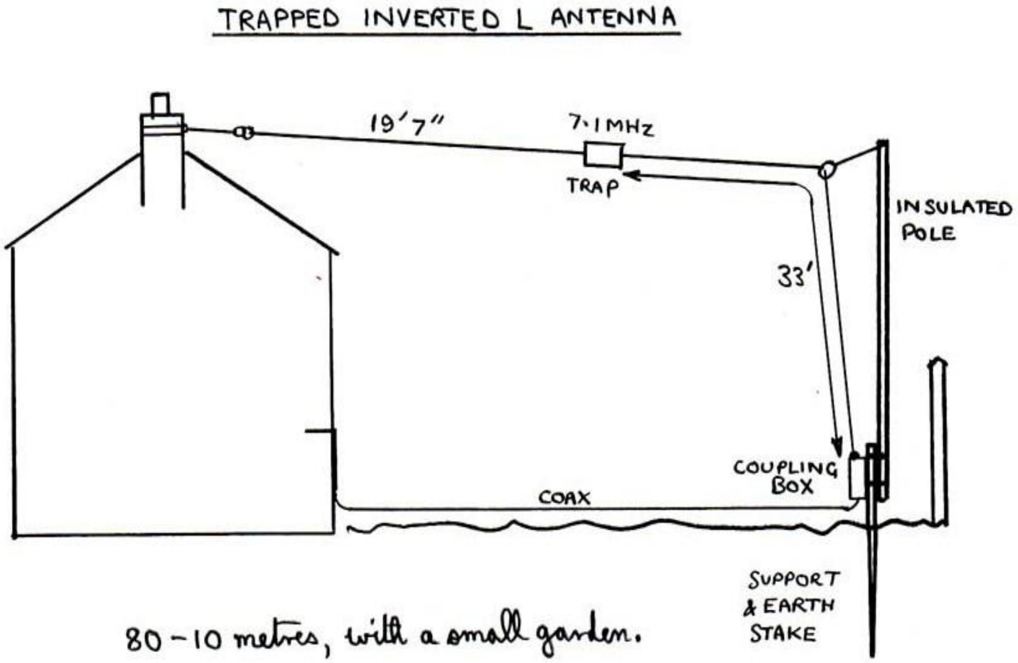
TRAPPED INVERTED L ANTENNA
19'7"
7·1MHz
TRAP
INSULATED POLE
33'
COUPLING BOX
COAX
SUPPORT & EARTH STAKE
80 – 10 metres, with a small garden.

The commercial version of the basic 80m - 10m Inverted L is available from Tony Nailer at [Spectrum Communications]

Important notes on effective Grounding by Jim K8OZ

Mike - I was reading about your work on the 160 meter Inverted L, and it makes me want to go out and build some more antennas! Congratulations. Your story is fascinating, and very well documented.

The only thing I can offer as a suggestion is to get as much radial wire along the edge of your property as possible (assuming your XYL will not allow you to bury radial wire all over your yard). Even if you can only run multiple wires 1/8th of a meter apart from each other, and parallel to each other, your losses will be reduced. The ground lossses have quite an impact on your transmitted signal, so any wire you can "hide" along the edge of your property will help improve your signal strentgh - little, by little...! { It may also affect your resonant frequency slightly, but that's easy to deal with by adjusting with an antenna tuner or slightly changing the loading coil. }

Good luck OM, and keep up the refinements on your antenna system. You're doing great! 73,

Jim, K8OZ
Albuquerque, NM

Delta Loop Antenna - Tuned for the 17 metre band but also usable as a multi-band operation

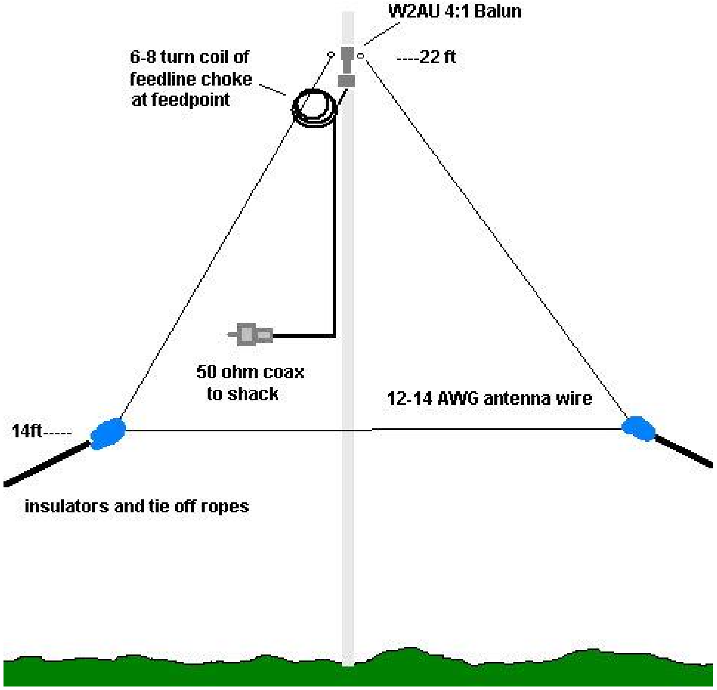

A typical Delta Loop antenna - diagram by W5SDC - gives multi-band operation with minimal cost.
My Delta Loop is fed near the bottom corner - it cannot be fed at the top, as in the diagram above, due to unwanted interaction with the antenna wire of the Inverted L antenna which is supported on the same pole.

My Delta Loop is fed near the bottom at one corner - it cannot be fed at the top, as in the diagram above, due to unwanted interaction with the antenna wire of the Inverted L antenna which is supported on the same pole.

It consists of a 16 metre long loop of wire in triangular Delta shape, hung from the top of the pole supporting the inverted L antenna and fed via RG213 coaxial cable via a home-brew 4:1 balun. A loop is really a single band antenna cut for one wavelength on the band of interest, however it can also work on higher bands with an ATU - as a good, cheap and easy to install multi-band H.F. aerial. Performance is a little better than the inverted L on the higher bands, but on 10 metres the tuned 10 / 6 metre wire 'fan' dipole in the loft can still be better.

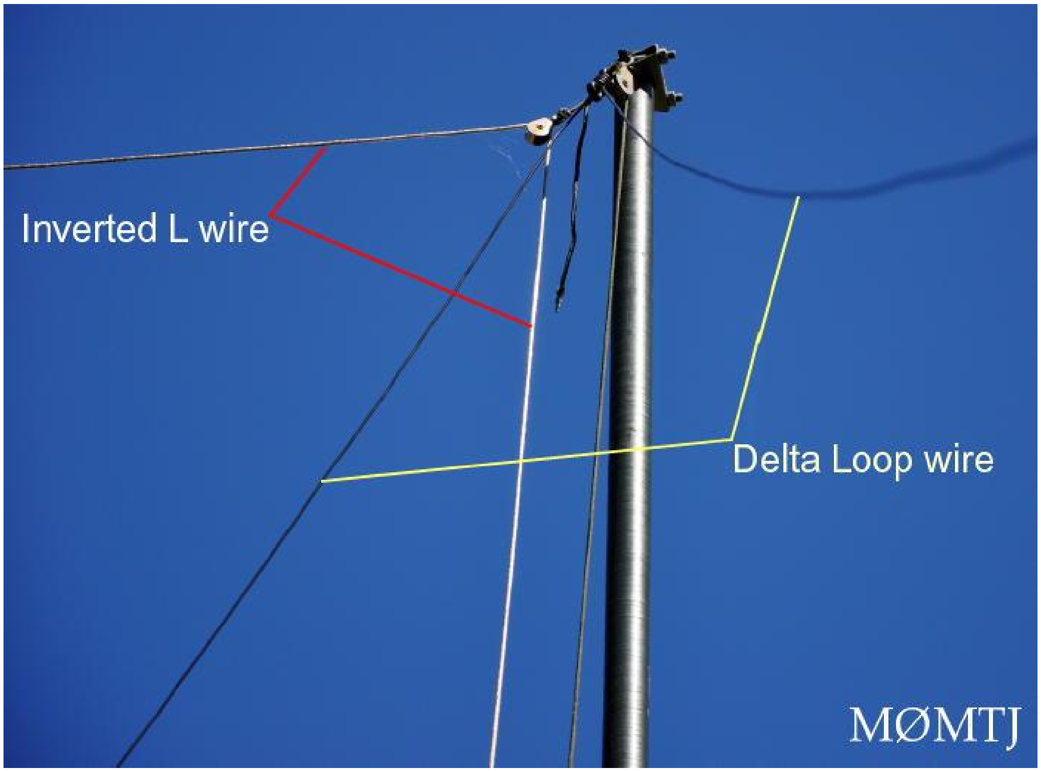
Inverted L wire
Delta Loop wire
MØMTJ

Apex of Delta Loop by M0MTJ

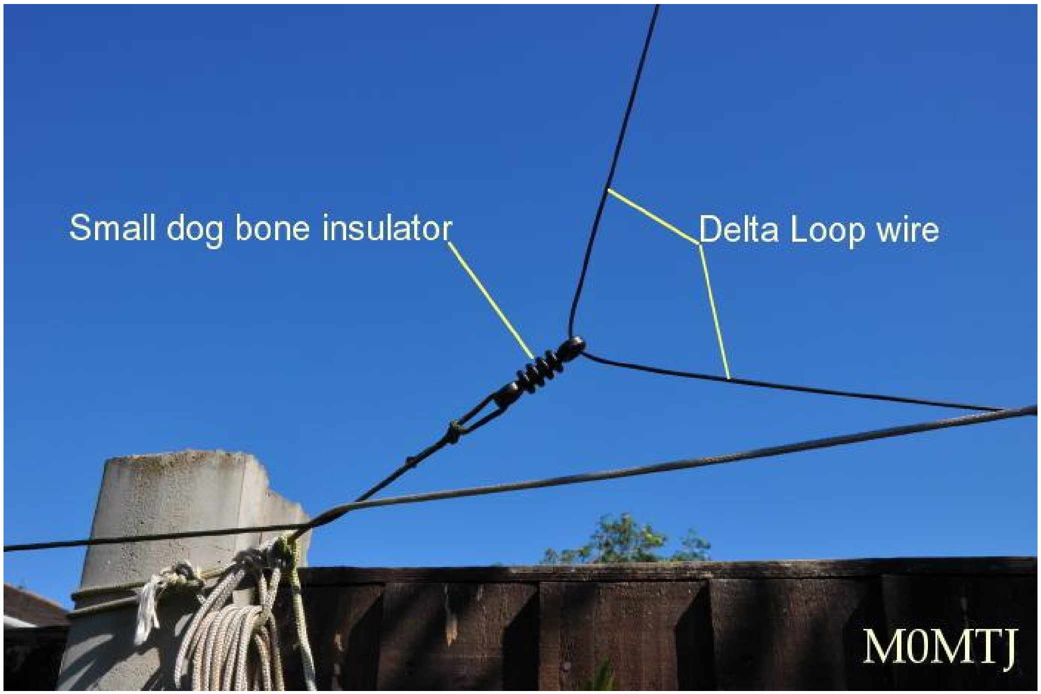
Small dog bone insulator
Delta Loop wire
M0MTJ

Bottom left corner of Delta Loop antenna by M0MTJ

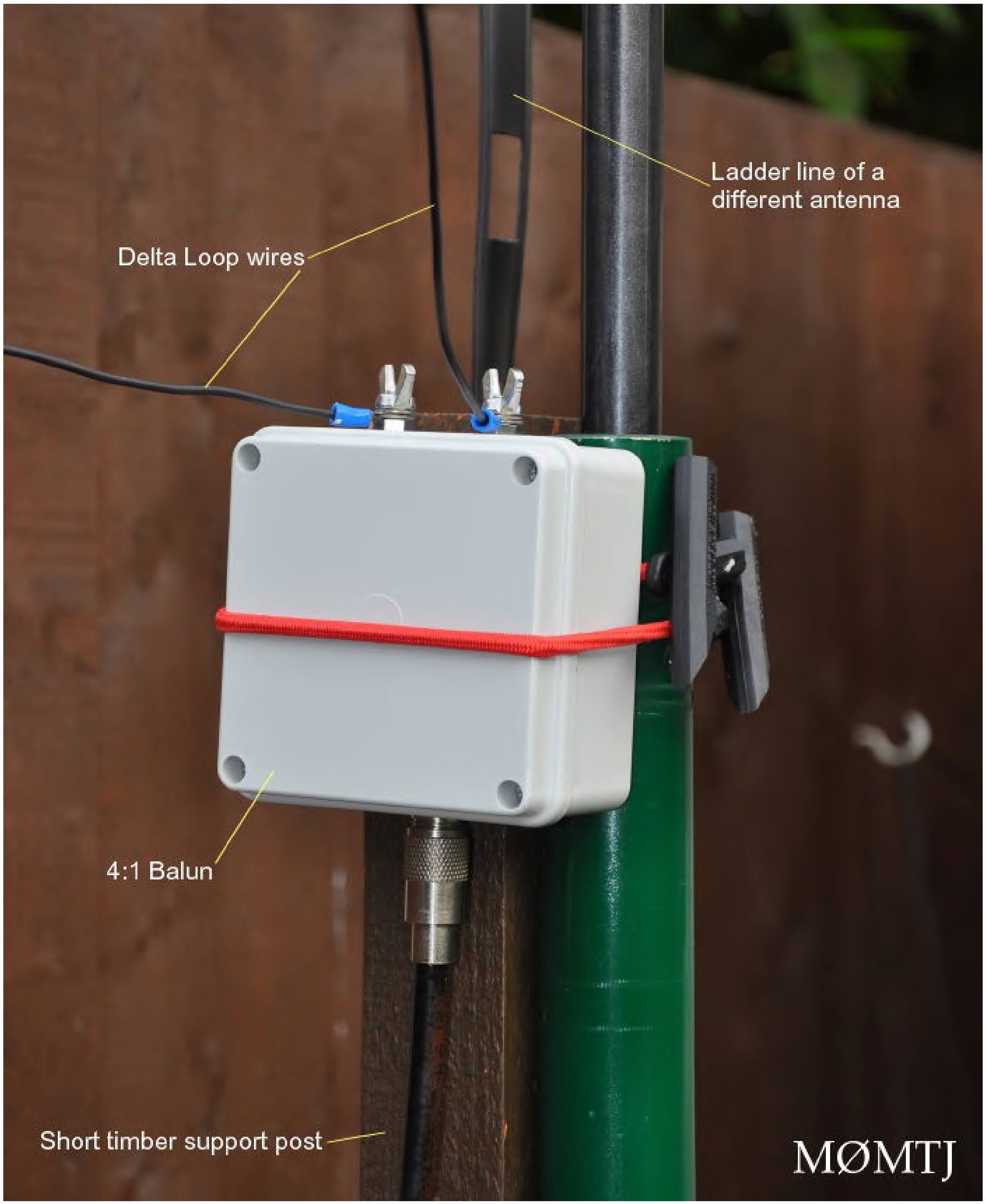
Ladder line of a different antenna
Delta Loop wires
4:1 Balun
Short timber support post
MØMTJ

Feed point of the Delta Loop at the bottom right hand corner

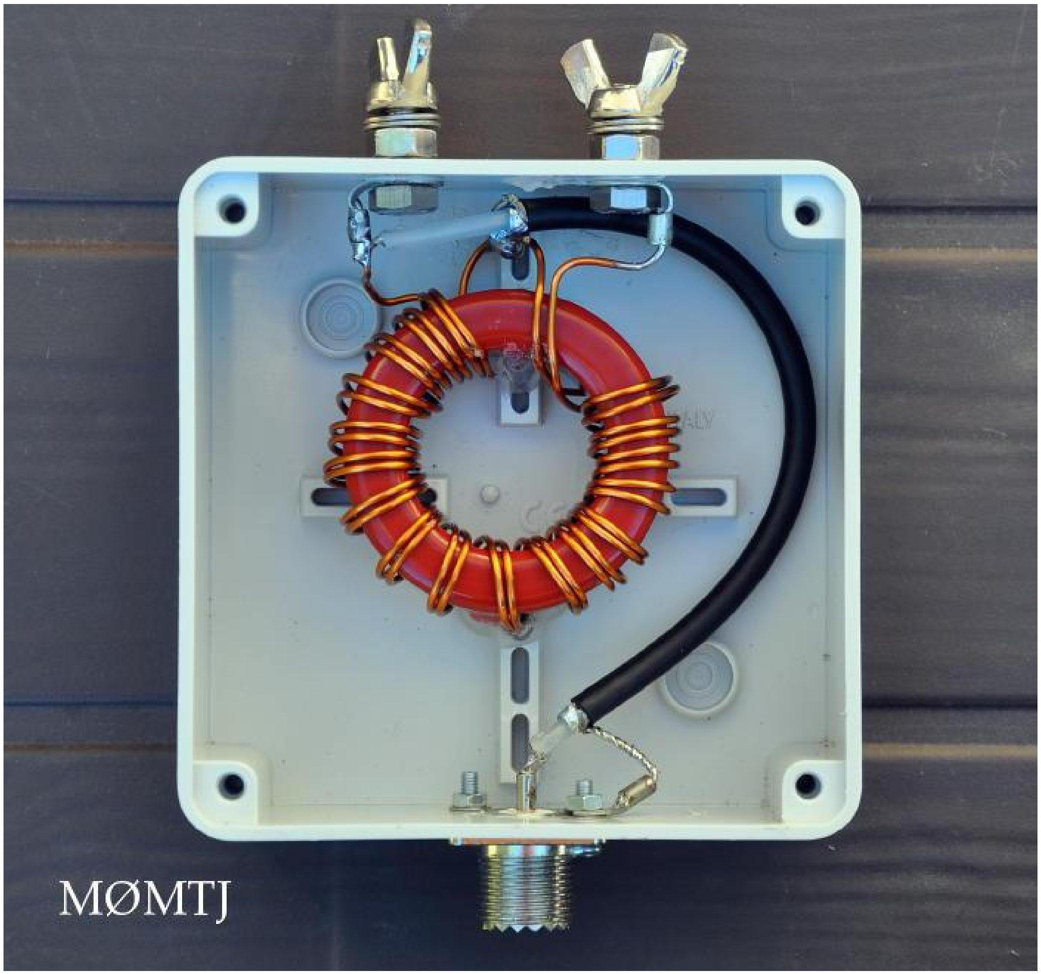

The Feed Point of the Delta Loop Antenna is at the bottom right hand corner
The Antenna is fed via a 'home brew' 4:1 balun by M0MTJ

Below are the VSWR measurements for the 16 metre long loop which has been measured an cut for resonance in the 17 metre band. For comparison are the measurements for the 12 metre long loop (which has not been optimised) and an 18 metre long loop which is of arbitrary length:

16 metre long loop of wire for the 17 Metre Band (optimised for 17m band)		
BAND	VSWR	VSWR
20m	14.0 MHz = 6.5	14.35 MHz = 4.9
17m	18.07 MHz = 1.2	18.16 MHz = 1.2
15m	21.0 MHz = 3.1	21.45 MHz = 3.7
12m	24.8 MHz = 5.9	25.9 MHz = 6.1
10m	28.0 MHz = 4.1	29.5 MHz = 4.4

12 metre long loop of wire for the 12 Metre Band (NOT optimised)		
BAND	VSWR	VSWR
20m	14.0 MHz = 22.1	14.35 MHz = 21.1
17m	18.07 MHz = 8.4	18.16 MHz = 8.1
15m	21.0 MHz = 5.0	21.45 MHz = 4.6
12m	24.8 MHz = 1.9	25.9 MHz = 2.0
10m	28.0 MHz = 5.0	29.5 MHz = 6.1

18 metre long loop of wire (An arbitary length between 20m & 17m)		
BAND	VSWR	VSWR
20m	14.0 MHz = 2.1	14.35 MHz = 1.4

17m	18.07 MHz = 2.6	18.16 MHz = 2.5
15m	21.0 MHz = 6.8	21.45 MHz = 8.4
12m	24.8 MHz = 6.3	25.9 MHz = 6.5
10m	28.0 MHz = 2.9	29.5 MHz = 2.5

Many users claim that loop aerials are quieter than typical vertical antennas. There are many plans available in the internet and given a suitable support or pole and a 4:1 balun it can take only a few minutes to install a loop antenna.

Arthur M0PLK (SQ2PLK) at Ham Radio Shop supplies an interesting lightweight self supporting Delta Loop antenna: http://ham-radio.urbasket.eu - see the review on the Polonia DX Award here: http://pdxa.one.pl/articles.php?article_id=17

LOOP ANTENNA LINKS: See lots more links to Loop Antennas on my links page here

Dipole or Doublet Antenna for 20m and 10m

This is an antenna trapped dipole for 20m and 10m. Currently it is fed by 75 ohm twin feeder to a 1:1 balun in the loft, then on to the ATU (AMU) via a short RG213 coaxial cable. Although it was initially installed horizontally, it is now installed with one leg supported vertically on a 7 metre fibreglass 'Sota' pole with the other leg supported horizontally about 2 metres above the ground. This is a rather unorthodox arrangement for a balanced dipole, but it seems to work ok and was inspired by another radio amateur's idea - although I don't recommend balanced feeder for this arrangement!

It looks much neater than the horizontally strung dipole and offers a more omnidirectional radiation pattern too.

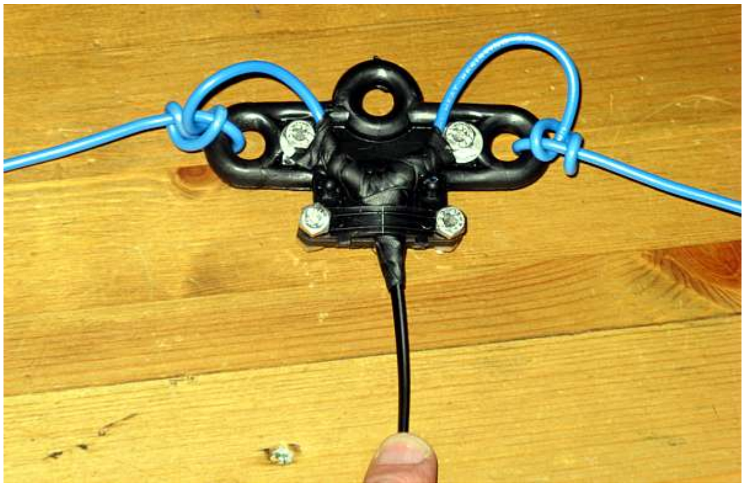

Dipole Centre with PVC covered wire and 75 ohm twin feeder attached.

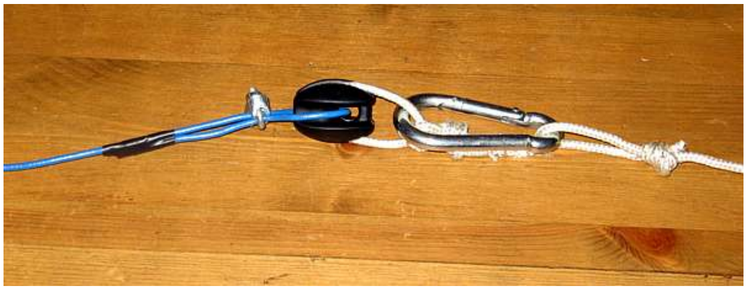

Removable end support method for wire dipole using a plastic antenna insulator, snap-hook and Dacron rope.

Photo showing how the wooden support posts are held in the ground by the steel Met Post.
This Met Post and wooden pole now supports the 7 metre high fibreglass Sota Pole (fishing pole).

The 'deformed dipole'.

A Dipole for 20m and 10m.

One leg is vertical, giving a more omnidirectional pattern and supported by the 7 metre long fibreglass fishing pole, while the other leg runs off horizontally at about 2 metres above the ground.

This antenna is fed by 75 ohm twin feeder.

The Deformed Dipole

Compact Antenna for Top Band

A Shortened Inverted L for 160 Metres

Despite the dreadful noise on top band caused by modern electronic gadgets and the difficulty in accommodating a necessarily large aerial in a small garden, I was keen to try to get on to top band. I experimented with some different ideas during 2009, some of which are shown on this page.

Eventually I settled on the design shown below. It is an Inverted L type aerial, shortened by the use of a loading coil. It uses a fibreglass telescopic fishing pole to allow it to be easily lowered out of sight when not in use. Read more on Antennas page 2 here>

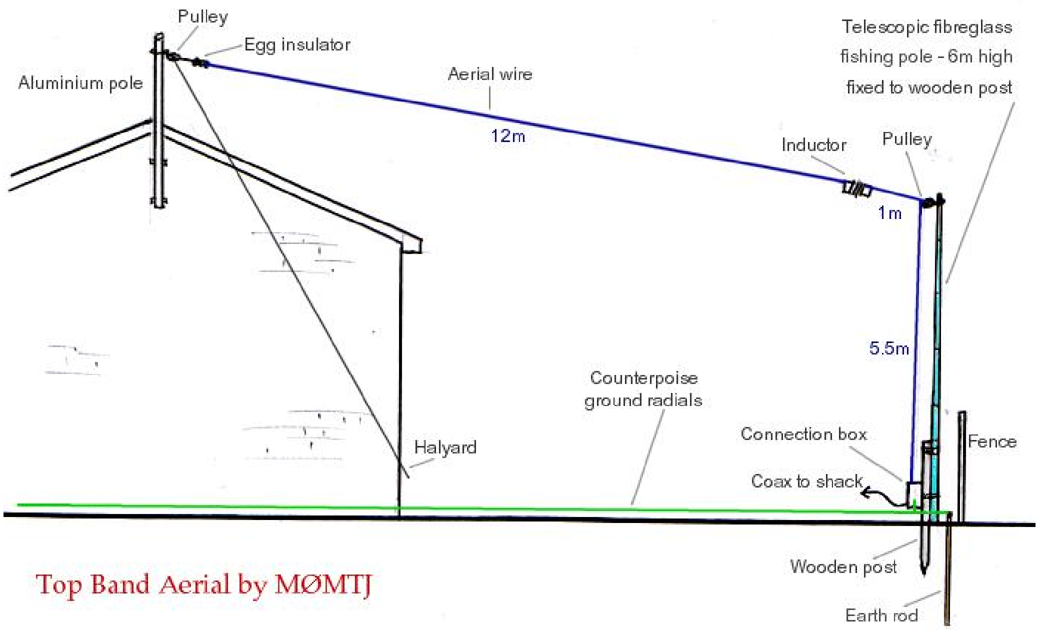
Pulley
Egg insulator
Aluminium pole
Aerial wire
12m
Telescopic fibreglass fishing pole - 6m high fixed to wooden post
Inductor
Pulley
1m
5.5m
Counterpoise ground radials
Connection box
Fence
Halyard
Coax to shack
Wooden post
Earth rod
Top Band Aerial by MØMTJ

Shortened Base Loaded Top Band Antenna For Small Gardens
uses a fibreglass telescopic fishing pole to allow it to be easily lowered out of sight when not in use.

Read more about Top Band Antennas on Antennas page 2 >

Other Antennas:

End Fed Zepp Antennas for 20m / 17m and 15m :

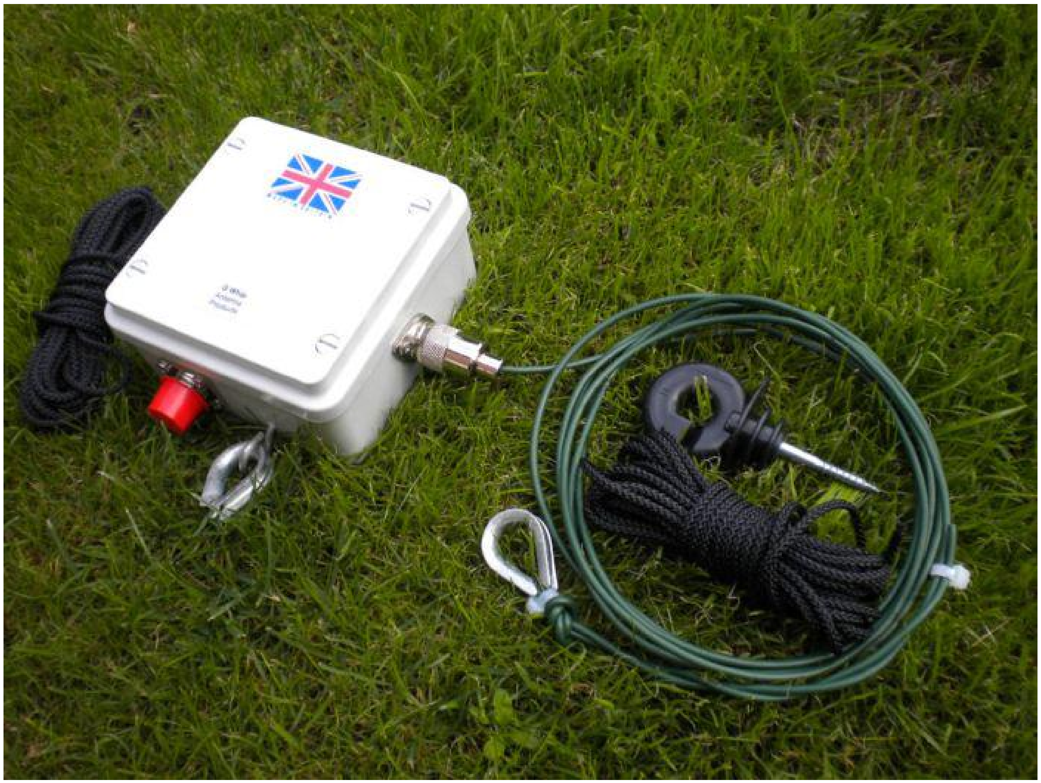

High quality commercially built Zepp antenna from G Whip Antenna Products.
http://www.gwhip.co.uk/

Above is a high efficiency, high quality commercially built Zepp style antenna using a half wave radiator. However there is a difference - unlike the traditional Zepp antennas, G4ICD's design dispenses with the long trailing 1/4 wave twin feeder matching section and counterpoise and, instead, uses a G-Whip's helical tuned unit.

The end fed Zepp is extremely versatile - it can simply be hung from the fascia board or guttering just outside one's window: The 1/2 wave wire radiator made of high strength Kevlar is easily sloped down the garden and is a camouflage green in colour so as to be almost invisible. The G-Whip Zepp is supplied with a fascia board insulator, a throwing line with plus cable strain relief and fittings, the IP65 waterproof resin potted box fitted with UHF connectors (SO239) for coax feeder. The Zepp can be mounted vertically (e.g. using a telescopic fibreglass pole), horizontally or sloping and can be ready in a matter of minutes.

I then decided to try an excellent new design developed by Geoff G4ICD - an End Fed Zepp antenna with a difference. The G-Whip End Fed Zepps are high efficiency, resonant mono band antennas utilising a half wave radiator, however this new G-Whip design by G4ICD dispenses with the long trailing 1/4 wave twin feeder matching section and counterpoise and, instead, uses a helical tuned unit. Using versions for 20m, 17m and 15m will enable me to get on the air with the minimum of fuss since the G-Whip end fed Zepp can simply be hung from the facia board or guttering just outside one's window. The high-strength Kevlar, camouflage green 1/2 wave wire radiator is easily sloped down the garden to be almost invisible.

G-Whip Widebander :

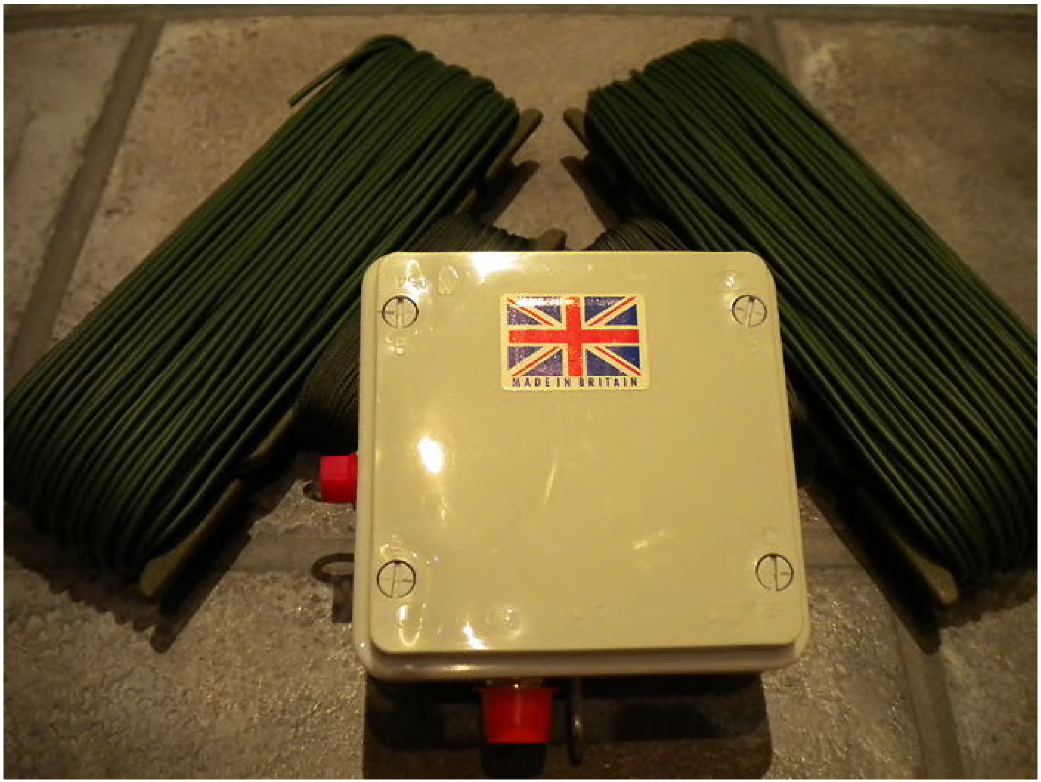

GWhip Widebander antenna.

Using the popular idea of feeding a large but non resonant antenna with an impedance converting 9:1 unun.

The G Whip wideband antenna consists of a 9:1 unun of GWhip's customary high quality for best efficiency, two 20metre lengths of kevlar wire for the radiator and counterpoise which provide operation from 3.6MHz to 50MHz. Feed with good quality low loss coax and use an ATU to match. The 20 metre radiator wire can be used as a sloper, or supported by convenient supports such as poles or trees in a straight line or 'dog legged'. I use a shorter radiator wire run up a telescopic fibreglass pole for operation on 20 metres an above.

Dual Band J-Pole :

There is a dual band vertical J-Pole antenna in the loft as a back up for the 2m and 70cm bands.

DUAL BAND SLIM JIM
BY N9TAX 2M-70CM

The excellent Dual Band N9TAX Slim Jim antenna that I use in the loft as a reserve antenna.
More information from Joe N9TAX at: www.n9tax.com
www.ebay.co.uk/itm/VHF-UHF-Slim-Jim-J-Pole-Dual-Band-2m-70cm-Antenna-jpole

Previous Antenna installations:

(2011 - 2013)

1) A trapped Inverted L for 80m and 40m with an SGC-230 auto antenna coupler at its feed-point at the bottom of the garden. RG213 coaxial cable is used to feed the output of the auto-coupler back to the shack. This can be used on all bands from 160 through to 10 metres. The support post is installed at the bottom of the garden with the end of the antenna wire being supported by Dacron rope that is attached to a pulley on a pole at the apex of the roof. A simple, single sloping wire element provides use on the 17 metre band. Although too short to be really effective on 160 metres, Top Band can be tuned by the SGC coupler. A pretty good all-round antenna. more

2) A half wave Wire J-Pole fixed to a telescopic fibreglass fishing pole for 10m. Cheap and effective. more about J-Poles

3) A 'home-brew' omnidirectional, vertical dual band, end fed antenna for 2 metres and 70cm. This is of the Controlled Feeder Radiation design (CFR) by VK2ZOI; effectively an end fed half wave dipole on 2m with an aluminium sleeve to achieve 70cms with a few extra dB's of gain. It is mounted on an aluminium mast 10 metres a.g.l. more

4) A DK7ZB design dual band Yagi antenna, with 5 elements for 2 metres and 8 elements for 70cms, mounted horizontally for SSB. A lightweight antenna rotator is employed and uses a push-up telescopic mast. Height above ground level is again approximately 7 metres. The DK7ZB is an excellent twin band Yagi antenna. more

5) Dual Band Fan Dipole, made from thick loudspeaker wire, mounted horizontally in the loft space for 10 meters an 6 metres. Cheap & effective.

6) (Installed late September 2013) Wire J-Pole antenna for 4 Metres (70 MHz) supported on a 3 metre long telescopic fibreglass pole to be attached to the top of the aluminium push up mast that supports the DK7ZB dual band yagi and rotator,

Other Options that can be deployed on an 'as required basis':

7) Compact Loaded Top Band Antenna, based on a design idea by Stuart Craigen G4GTX more

8 & 9) G Whip End Fed Zepps (EFZ's) for either 20m, 15m or 17m or the G-Whip "WideBander" which is an 'UnTenna' style antenna that can be used for 20m through to 10m using good quality G Whip 9:1 UnUn; useful additions for antenna flexibility. more

10) N9TAX Dual Band Slim Jim (J-Pole) antenna mounted in the loft as a back-up antenna for 2m and 70cms. Very good. more

11) Delta Loop Antenna - 16 metre loop of wire in triangular Delta shape, hung from the top of the pole supporting the inverted L antenna and fed via RG213 coaxial cable via a 4:1 balun. The loop is really a single band antenna cut for one wavelength on the band of interest, however it also can be pressed into service for some higher bands - a good, cheap and easy to install aerial; Often works better than the inverted L on the higher bands, but on 10 metres the tuned 10 metre dipole in the loft is sometimes better. more

(2011)

In mid 2011 I experimented with an excellent N9TAX designed dual band Slim-Jim (J-Pole) antenna for 2m and 70cms. This is made from lightweight 450 Ohm ladder line which can be fixed to the top of a 10m tall fibreglass, telescopic, fishing pole. The N9TAX works extremely well indeed. More information from Joe N9TAX at: www.n9tax.com and buy at: www.ebay.co.uk/itm/VHF-UHF-Slim-Jim-J-Pole-Dual-Band-2m-70cm-Antenna-jpole

N.B. I tried to home-brew the DJB-1 dual band J-Pole antenna using plans published by the ARRL in QST magazine. I wanted a neat antenna that could be enclosed in a protective tube to minimise weathering effects. However trying to tune this antenna at UHF frequencies proved to be frustratingly difficult to do and after two full days work I could not get the thing resonate accurately at the correct frequency. Sadly, for this reason, I cannot recommend the Dual Band J-Pole as a home-brew project.

The N9TAX antenna on the other hand works very well. However it cannot be enclosed in a tube due to the velocity factor effect de-tuning the antenna's resonant frequencies.

(Late 2011)

Due to difficulties with the stability of a lightweight fishing pole as a support I moved back to using the lightweight aluminium telescopic mast, with stays, to support a Watson W-50 vertical dual band collinear for 2 metres and 70 cms FM.

The excellent N9TAX dual band Slim Jim is now installed in the loft.

(Antennas used up until 2010)

1) A trapped Inverted L for 80m and 40m fed by RG213 coaxial cable to the LDG Z-11 Pro antenna matching unit in the shack. This can be used on all bands from 80 through to 10 metres. The support post is installed at the bottom of the garden with the end of the antenna wire being supported by Dacron rope that is attached to a pulley on a pole at the apex of the roof. This excellent antenna is still use. *more*

2) A trapped dipole for 20m and 10m. This was fed by 75 ohm twin feeder to a 1:1 balun then on to the AMU via RG213 coaxial cable. It was initially installed horizontally, but more latterly installed with one leg supported vertically on a 7 metre fibreglass 'Sota' pole with the other leg supported horizontally about 2 metres above the ground. A rather unorthodox arrangement for a balanced dipole, but it seemed to work ok, it looked much neater than a horizontally slung dipole and also offered a more omnidirectional radiation pattern. *more*

4) A compact Inverted L for the 160 metre band - Top Band - shortened with a loading coil. *more*

5) As N9TAX Dual Band Slim Jim (J-Pole) antenna for 2m and 70cms. This was fixed near the top of a 10m telescopic fibreglass fishing pole that I pushed up whenever it was required. *more*

(2008)

W-2000 - Vertical Collinear mounted on a temporary 10m telescopic pole:

I no longer have the Watson W-2000 but this is how it was used previously:

The Watson W-2000 covered VHF (2 metres / 144 MHz) and UHF (70 cms / 430 MHz) and also, rather usefully, 6 metres (50 MHz) too. The W-2000 is 2.5 metres long and enclosed in white fibreglass with three radial elements at the base.

Unfortunately I had nowhere practical to install a separate mast for the VHF / UHF antenna, so this was mounted on top of a 30 foot (10 metre) high telescopic aluminium mast in the back garden. The base of the mast was placed in a handy metal sleeved hole that was already present in a small wall in the garden. Very fortunate indeed!

The antenna is connected to the radio via the very low loss Westflex 103 coaxial cable. The cable was left in place permanently, running from the shack in the front bedroom, up into the loft and out of a small hole in the back of the house, down a drain pipe into the back garden. From there the aerial can be connected as an when required:

When VHF or UHF operation is required I have to connect the coaxial cable to the Watson W-2000, fix it to the top of the telescopic mast, which is very quick using two V bolts and 4 wing nuts, put the mast in the hole and raise it to a good height. I tend to extend it so that the bottom of the antenna is at about 24 feet in the air, the height of the apex of the house, so it is in fairly clear space.

A VHF and UHF aerial needs to be as high as possible since at these frequencies communication is essentially local and 'line of sight' - unless heightened propagation conditions, such as Sporadic E or a Temperature Inversion is prevailing at the time.

Even at 24 feet the mast is rather wobbly, so it was tied down using three nylon guy ropes.

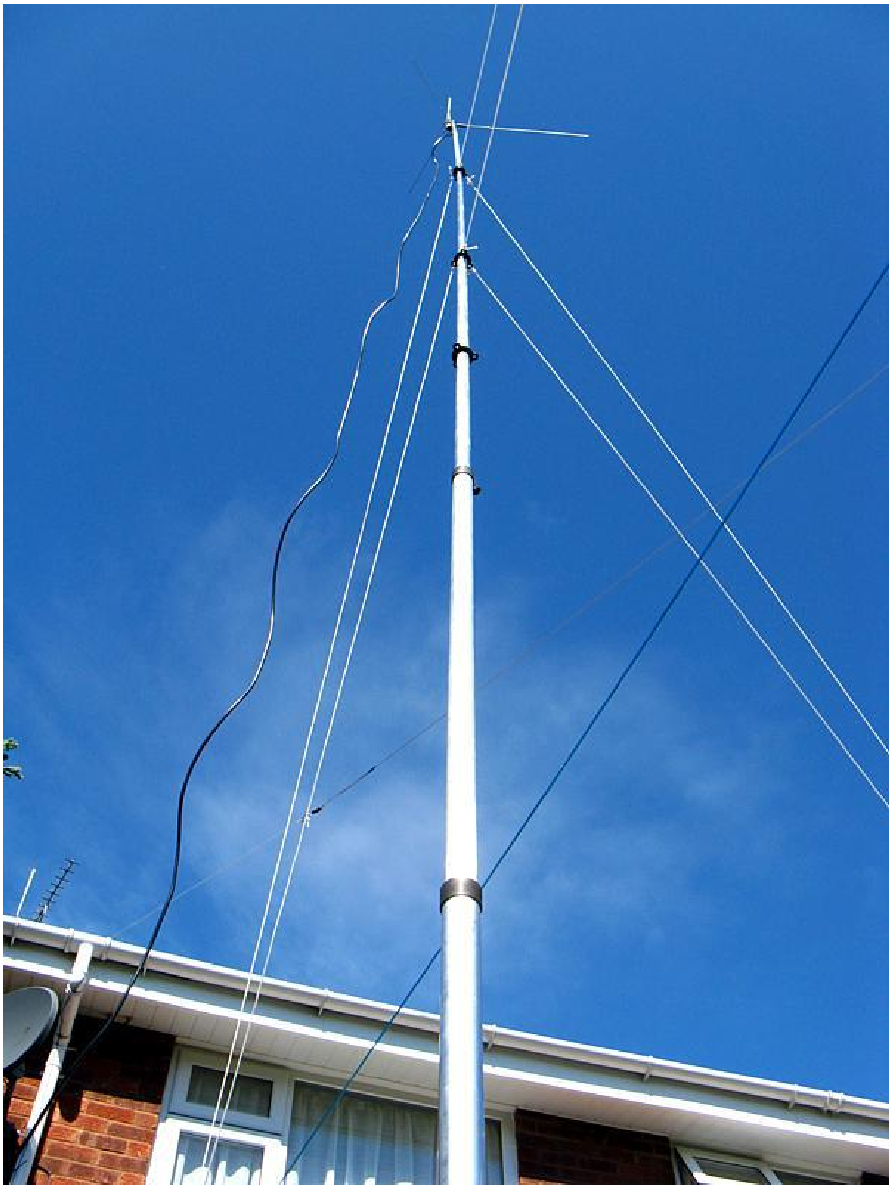

The Watson W-2000 on to of the extended telescopic pole - about 8 or 9 metres high.

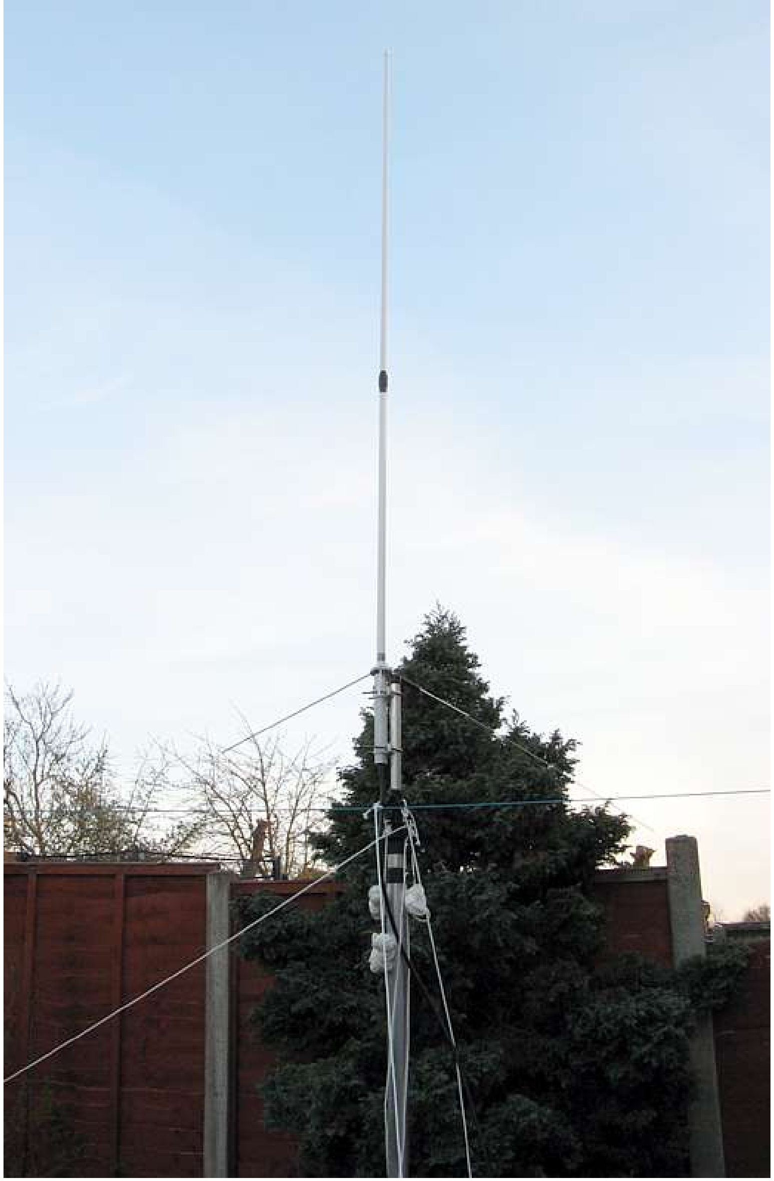

Photograph of a Watson W-2000 on Telescopic Mast at the lowest position.

MORE ANTENNAS

Our good friend in Australia Felix Scerri, VK4FUQ, uses Inverted V antennas but also highly recommends the Quad Loop style antenna for HF work. These are well worth investigating, and you can read more here: Antennas 3 with more antenna ideas on Antennas 2 and Antennas 4 and the Links Page here and here

KNOTS FOR SECURING WIRE ANTENNAS

I have found the Bowline to be one of the most useful, it is strong and easy to tie. A Bowline will not slip in any circumstances and, usefully, the more load that is put on it, the tighter it gets. Read more about good knots for amateur radio aerials here...

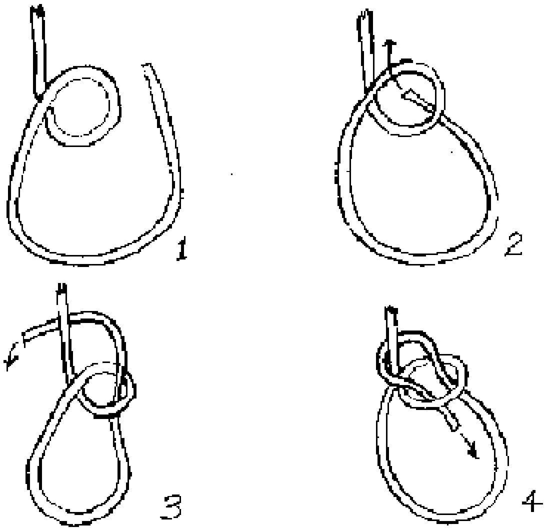

The Bowline Knot - Read more about knots here ...

Antenna Trimming Chart and useful Antenna Rigging Accessory ideas

On Antennas 4 I have included a helpful Antenna Trimming Chart and some useful ideas for Antenna Rigging Accessories

More project ideas here>

73

Mike, M0MTJ 2011 / 2012

Antennas 2 | **Antennas 3** | **Antennas 4** | **Antennas 5** | **Antennas 6**

Links to further reading:

Introducing The All Band Doublet: http://www.cebik.com/content/edu/edu6.html ***N.B. Create a free account at*** http://www.cebik.com

The All Band Doublet - http://www.cebik.com/wire/abd.html

The ALL Band HF Doublet on Ham Universe - http://www.hamuniverse.com/hfdoublet.html

Multi Band Dipoles Compared - by ARRL on QST and DX Zone:
http://www.arrl.org/tis/info/pdf/9611073.pdf http://www.dxzone.com/cgi-bin/dir/jump2.cgi?ID=7499

PDF Document - The W3DZZ Antenna -
http://www.users.icscotland.net/~len.paget/GM0ONX%20trap%20dipole.pdf (**!!! But don't use coaxial cable with a 'choke balun' at the centre of the dipole! Use twin feeder with the Choke Balun at the other end. Less power loss.!!!***)

See Practical Dipole Antennas Compared: http://www.qsl.net/ta1dx/amator/practical_dipole_antenna.htm

Practical Antenna For 160 Metres - http://www.iw5edi.com/ham-radio/?a-practical-antenna-for-160-metres,32
http://www.ik1mnj.net/id202.htm

More:

The website of GM0ONX http://www.gm0onx.co.uk/

The Inverted L - PDF document: http://www.users.icscotland.net/~len.paget/5%20band%20Inverted%20L.pdf

Adding Top Band To The Inverted L - PDF Document:
http://www.users.icscotland.net/~len.paget/Inverted%20L%20adding%20top%20band.pdf

The All Band Inverted L - http://www.antennex.com/preview/archive3/ltv.htm

Q.T.H. Move in 2010 !%*?"*!?!

We decided that we'd like to move house in 2009, we found a new property in early 2010. My amateur radio aerials were duly taken down and the ham shack packed away. However after months of delays we still had not moved by August 2010, but we were still hoping to move. However after months of messing about and stringing us along our buyer pulled out the very day before we were to exchange contracts later in August 2010.

This cost us a lot of time and a great deal of wasted money. Thank you Mrs xxxxxx :-(

After a wasted year we decided to stay where we were and take the house of the market. Instead we spent the next eight months remodelling and redecorating. No time for playing radio of course and besides everything was still all packed away in boxes!

In mid 2011 I was getting frustrated that I had no radio. So I suppose it's time to think about re-establishing the station and to start planning the installation of some antennas. Of course Jules, my XYL, understandably questions my antennas and experiments!

Due to time constraints I will probably start again with a somewhat temporary antenna. I was thinking along the lines of an "Untenna" - so I installed a 7.2 ish long wire supported on a vertical fibreglass pole with a horizontal 'counterpoise' connected via a 9:1 balun to the coax back to the shack. Of course it is a fairly low efficiency multi-band (wideband) antenna, but easy to get going quickly. The GWhip Widebander antenna by Geoff Brown G4ICD is possibly the highest quality antenna of this type available, using a very high efficiency, top quality 9:1 UnUn with a 17 meter wire radiator and 10 metre long counterpoise - a very useful, versatile 'all situations' antenna.

As time progressed I re-established my full size trapped Inverted L antenna for 80metres and 40 metres and added a switchable loading coil at its base for use on 160 meters, as described above.

Then I gradually re-established the 2 metres and 70 cms antennas with the vertical W-50 and horizontal dual band DK7ZB Yagi - as detailed above.

Antennas 2 | Antennas 3 | Antennas 4 | Antennas 5 | Antennas 6

Index To Other Antenna Pages:
Antennas 1 : Aerials used at M0MTJ
Antennas 2 : Including ..... Ideas for compact aerials for Top Band /160 metres
Antennas 3 : Felix Scerri VK4FUQ discusses Loop Antennas, baluns, masts & other antenna related topics
Antennas 4 : Including ...... Many antenna ideas from various sources particularly for multi-band operation & also gives information about
antenna trimming, knots for wire antennas and useful antenna rigging accessory ideas.
Antennas 5 : Including ..... Half Wave End Fed aerials for 144 MHz VHF / 430 MHz UHF and 50 MHz 6 Metre band & J-Pole Aerials
Antennas 6 : Including .... Simple and effective H.F. Antenna ideas - Ground Plane and All Band Doublet

G-Whip Antenna Products
Geoff Brown G4ICD offers a multitude of high quality solutions for portable, mobile and permanent base installations

G-WHIP ANTENNA PRODUCTS
PRO

G-WHIP G Pro Whip antennas
http://www.gwhip.co.uk/

G Whip Antenna News
HF BASE ANTENNAS, PORTABLE ANTENNAS, MOBILE ANTENNAE, ZEPP ANTENNAS, BALUNS
WWW.GWHIP.CO.UK
G Whip Antenna Products
MADE IN BRITAIN
G Whip Antenna Products

New Products.

More products have been added to the range.. The Zepp range of antennas, these end fed antennas are absolutely superb, I personally took samples of them to GI in December 2009 and had massive pileups of stations. At present there are models for 40/20/17/15/10/6 Mtrs, note that these are dedicated antennas for each band.

After many requests we now have a range of baluns, these are just wound components or top quality finished units. Several models are available for unbal to bal, or un to un. Most ratios are available like 1/1, 1/4, 6/1, 9/1 with power levels up to 1.5 kw

http://www.gwhip.co.uk/

Just For Fun:

A tower that we may like to have to attach our antennas to - but I don't think that Health And Safety was taken into account here:

M0MTJ | Operating Conditions | Antennas / Aerials | /P Portable Operating | Accessories | Projects & Kits

Useful Information | Contact MØMTJ | Contact M6ORS | Links to Amateur Radio Sites

RSGB | QSL | The Amateur Radio Mini Site Map

WSPR Weak Signal Propagation Reporter

News, Developments, Events, Photographs and other 'Bits 'n' Bobs'

MDS975 Home

Feedback

MAIL

Mike Smith - MDS975.co.uk © 2003 - 2016

M0MTJ

Amateur Radio; Ham Radio; Radio; Transceivers; HF; VHF; UHF; Data Modes; Morse Code; RTTY; PSK31; SSTV; FSTV; Amtor; Sitor Antennas; Aerials; Cable; Coaxial Cable; Twin Lead; Masts; Poles; Propagation; Computer; PC; USB Computer Interface; Microphone Loudspeaker; Filters; Noise Reduction; DSP; Digital Signal Processing; Morse Key; SWR ; Inverted L; Inverted V; Dipole; Doublet.

http://www.freeradio.co.uk/

Wire Antennas for Ham Radio

Iulian Rosu YO3DAC / VA3IUL,

http://www.qsl.net/va3iul

Note: Dimensions of the antennas presented below are either in: feet ('), inch ("), meters (m), or centimeters (cm)

01 - Tee Antenna
02 - Half-Lamda Tee Antenna
03 - Twin-Led Marconi Antenna
04 - Swallow-Tail Antenna
05 - Random Length Radiator Wire Antenna
06 - Windom Antenna
07 - Windom Antenna - Feed with coax cable
08 - Quarter Wavelength Vertical Antenna
09 - Folded Marconi Tee Antenna
10 - Zeppelin Antenna
11 - EWE Antenna
12 - Dipole Antenna - Balun
13 - Multiband Dipole Antenna
14 - Inverted-Vee Antenna
15 - Sloping Dipole Antenna
16 - Vertical Dipole
17 - Delta Fed Dipole Antenna
18 - Bow-Tie Dipole Antenna
19 - Bow-Tie Folded Dipole Antenna for RX
20 - Multiband Tuned Doublet Antenna
21 - G5RV Antenna
22 - Wideband Dipole Antenna
23 - Wideband Dipole for Receiving
24 - Tilted Folded Dipole Antenna
25 - Right Angle Marconi Antenna
26 - Linearly Loaded Tee Antenna
27 - Reduced Size Dipole Antenna
28 - Doublet Dipole Antenna
29 - Delta Loop Antenna
30 - Half Delta Loop Antenna
31 - Collinear Franklin Antenna
32 - Four Element Broadside Antenna
33 - The Lazy-H Array Antenna
34 - Sterba Curtain Array Antenna
35 - T-L DX Antenna
36 - 1.9 MHz Full-wave Loop Antenna
37 - Multi-Band Portable Antenna
38 - Off-center-fed Full-wave Doublet Antenna
39 - Terminated Sloper Antenna
40 - Double Extended Zepp Antenna
41 - TCFTFD Dipole Antenna
42 - Vee-Sloper Antenna
43 - Rhombic Inverted-Vee Antenna
44 - Counterpoise Longwire
45 - Bisquare Loop Antenna
46 - Piggyback Antenna for 10m
47 - Vertical Sleeve Antenna for 10m
48 - Double Windom Antenna
49 - Double Windom for 9 Bands
50 - Collinear Trap Antenna
51 - Short Dipole Antenna for 40m - 80m - 160m
52 - Center Fed-Zepp Antenna for 80m - 40m
53 - All Bands Antenna
54 - All-Bands Dipole Antenna
147 - KE4PT OCEF All-Band Dipole
148 - Wire Quad Antenna for 40m
149 - Inclined Dipole Antenna for 80m, 40m
150 - Pyramidal Wire Antenna for 80m
151 - Random Wire Antenna All Bands
152 - Multiband Dipole Antenna for 80m, 40m,15m, 10m
153 - Slim Jim Wire Antenna for 4m
154 - Delta Loop for 6m
155 - Re-Configurable Antenna for 160m and 80m
156 - Very Low Frequency Inverted-L Antenna
157 - Reduced Size Half Sloper Antenna for 160m
158 - Tree-Mounted HF Antenna
159 - Multiband Vertical Antenna for 80m, 40m, 20m
160 - Marconi Antenna for 136 kHz
161 - Simple Killer Antenna for 40m
162 - Stub-Directed V Antenna for 80m
163 - KT0NY Over-and-Under DX Antenna for 20m
164 - Horizontal Loop Antennas
165 - Ribbon J-Pole for 2m
166 - Dual Band Ribbon J-Pole for 2m, 70cm
167 - Square Vertical Loop Antenna for 40m
168 - Tri-Band Quad Antenna for 20m, 15m, 10m
169 - 3D Quad Antenna for 80m, 40m, 20m, 15m, 10m
170 - Sloping Wire Antenna for 30m, 20m, 17m, 15m, 12m, 10m
171 - Broadband Dipole Antenna with Coaxial Resonator for 80m
172 - Resonant Feedline Dipole Antenna for 80m
173 - Dual-Band Loading Wire Antenna for 80m, 30m
174 - Stub Matching Antennas
175 - J-Style Vertical Wire Antenna for 10m
176 - Dual Band Vertical with Zepp Feeders for 40m, 20m
177 - RCA Double Doublet for 40m to 12m
178 - RCA Spiderweb Antenna for 40m to 6m
179 - Folded Dipole with Shorted Straps
180 - Twin-Lead Marconi Antenna for 160m, 80m
181 - Broadband Antenna Quarter-Wave Balun for 80m
182 - Three-Quarter-Wave Folded Doublet for Dual-Band
183 - Three-Quarter-Wave Folded Doublet No-Switch for Dual-Band
184 - Wideband Omni-Directional Discone Antenna
185 - Wideband Rhombic Antenna for 40m to 10m
186 - Pre-Cut Linear Array Antenna 3dB-Gain for 40m
187 - X-Array Antenna 6dB-Gain for 20m, 15m, 10m
188 - Double-Bruce Array Antenna 5dB-Gain for 20m, 10m
189 - Bi-Square Broadside Array 4dB-Gain for 20m, 15m, 10m
190 - Six-Shooter Broadside Array 7.5dB-Gain for 20m, 15m, 10m
191 - Triplex Flat-Top Beam 4.5dB-Gain for 20m, 15m, 10m
192 - Dual-Band Tilt Antenna for 20m, 10m
193 - Super Space Multiband Dipole Antenna for 80m to 10m
194 - Bi-Square Beam Antenna Gain=5.5dB for 10m
195 - Cousin of G5RV Multiband Antenna for 40m to 10m
196 - Cayman Quad Antenna for 20m
197 - Hentenna Gain=3dB with Bazooka Match for 6m
198 - X-Beam Antenna Gain=3dB for 20m

55 - Multiband Z Antenna
56 - Multiband Dipole Antenna
57 - Five-Bands No-Tuner Antenna
58 - Dualband Full-wave Loop Antenna for 80m-40m
59 - Loop Antenna for 10m
60 - Lazy Quad Antenna for 10m
61 - Tri-band Delta Loop Antenna for 80m - 40m - 30m
62 - Dual-band Loop Antenna for 30m - 40m
63 - Wire-Beam Antenna for 80m
64 - Dual-Band Sloper Antenna
65 - Inverted-V Beam Antenna for 30m
66 - ZL-Special Beam Antenna for 15m
67 - Half-Sloper Antenna for 160m
68 - Two-Bands Half Sloper for 80m - 40m
69 - Linear Loaded Sloper Antenna for 160m
70 - Super-Sloper Antenna
71 - Tower Pole as a Vertical Antenna for 80m
72 - Clothesline Antenna
73 - Curtain Zepp Antenna for 160m, 80m, 40m
74 - Collinear Array Antenna for 40m, 30m, 20m
75 - 160m Inverted Delta Loop
76 - Half Rhombic Unidirectional Vertical for 20m to 6m
77 - Capacitance Loaded Vertical Antenna for 160m
78 - Fan Dipole Antenna for 80m to 6m
79 - Wire Ground Plane Antenna
80 - Inverted Delta Loop Antenna for 160m
81 - Inverted-L for 160m
82 - 300ohm-Ribbon Dual Band Dipole
83 - Tri-Band Beam for 20m, 15m, 10m
84 - Mini-Horse Yagi Antenna
85 - Backpack J-Pole Antenna for 10m, 6m, 2m
86 - Fan-Dipole Antenna for 80m, 40m, 20m
87 - Capacity Tuned Folded Loop Antenna for 20m
88 - Indoor Loop Antenna for 80m to 30m
89 - Indoor Loop Antenna for 80m
90 - Double-Delta Antenna 80m and 40m
91 - Inductance-Loaded Shortened Dipole for 160m
92 - V-Beam Antenna for 15m
93 - Picnic Vertical Wire Antenna
94 - Laid-Back Quad Antenna for 80m
95 - Phased Loop Antenna
96 - Loop Antenna for TX for 160m
97 - Morgain-Dipole Antenna for 160m and 80m
98 - ZL-Special for 20m, 15m, 10m
99 - Biconical Antenna
100 - Directive Delta-Birdcage Antenna for 20m to 10m
101 - Dual Polarization Antenna for 80m and 40m
102 - Directive 300-ohm-Ribbon Folded Dipole for 15m
103 - Miniature Directive Antenna for 10m
104 - Biquad Antenna 12dBi-Gain for 2.4GHz
105 - Dual-Rhomboid Antenna for 435MHz to 870MHz
106 - Double-Bazooka Antenna for 80m
107 - J-Style Antenna
108 - Vertical / Horizontal / Circular Polarization Antenna
109 - Coax Inverted-L Antenna for 80m
110 - Indoor Compact Loop Antenna for 80m
111 - Helix Antenna
112 - Novice Vertical Antenna for 80m, 40m, 15m, 10m
113 - Stub-Loaded Shortened Dipole for 80m
114 - Six-Band Wire-Stub Trap Antenna for 40m-10m
115 - Multiband Half-Wave Delta-Loop Antenna
199 - Twin Delta Loop Antenna Gain=6dB for 160m, 80m, 40m
200 - Inverted-V Beam Antenna for 30m
201 - FD4 Windom Antenna for 80m, 40m, 20m, 10m
202 - Two Elements Phased Delta Loop for 40m
203 - Three Elements Wire Yagi for 10m
204 - Low Radiation Angle Full-Wave Loop Antenna for 80m
205 - Collinear Antenna Gain=3dB for 2m
206 - X-Ray Monoband Antenna Gain=6.5dBd
207 - Broadside Collinear Curtain Array for 20m
208 - Shrunk Loop Antenna for 80m, 40m
209 - λ/3 Multiband Dipole for 40m, 20m, 17m, 15m, 13m, 10m
210 - K3LR Sloper Antenna Gain=3dB for 160m
211 - Mini Folded Vertical Monopole Antenna for 160m, 80m, 40m
212 - Three Band Dipole for 80m, 40m, 20m
213 - Stub Matched Dipole Antenna
214 - Inverted-Vee Antenna with Bazooka Match
215 - Multiband Long Wire Antenna for 40m, 20m, 15m, 10m
216 - Converted Vee Antenna for 80m, 40m
217 - Swiss Quad Antenna for 20m, 15m, 10m
218 - Japanese Quad Antenna for 20m, 15m, 10m
219 - Seven-Elements Triangle Beam Antenna for 20m
220 - Log Periodic Wire Antenna for 40m
221 - Three-Elements 90-degrees Wire Beam for 20m
222 - Two-Elements Wire Beam for 20m
223 - Suitcase-Fit Dipole Antenna for 80m, 40m, 20m
224 - Multiband Coaxial Dipole for 80m, 40m
225 - Vertical Log Periodic Antenna for 80m, 40m
226 - Inverted Vertical Log Periodic Antenna for 40m
227 - Double Extended Zepp Antenna Gain=7dB for 15m
228 - Double Lazy-H Antenna G=10dB for 70cm
229 - Wire Log Periodic Antenna G=12dB for 15m, 20m
230 - Five-Elements Vertical Log Periodic for 80m, 40m, 20m
231 - Sloping Delta Loop Antenna for 40m
232 - Doublet Antenna for 160m
233 - Log Periodic Wire Beam for 40m
234 - G3LDO Wire Beam Antenna for 20m, 15m, 10m
235 - Phased Delta Loop Array for 20m
236 - Dual Band V-dipole for 80m, 40m
237 - Collinear Array 5dB-Gain for 15m
238 - Two-Band Quad Loop Antenna 80m, 40m
239 - K7CW Quad Antenna
240 - Broadband Short Dipole for 80m
241 - Off-Center Fed Dipole for 40m, 20m, 10m
242 - Broadband Sloper Antenna for 80m
243 - Two-Frequency Dipole Antenna for 80m
244 - Top Loaded Vertical Antenna for 160m
245 - Two Loop Beam Antenna for 20m, 15m
246 - Corner-Fed Delta Loop Antenna for 80m, 40m, 20m
247 - Loaded Delta Loop Antenna for 40m, 20m
248 - Compact Size Dipole for 80m
249 - JAWS 3dB Gain Antenna for 40m
250 - Inverted-V Antenna for 160m
251 - Lazy-U Antenna for 160m
252 - AF3V Compact Antenna for 160m
253 - Hot Beam Wire Antenna for 40m
254 - Sloping Delta Loop for 40m
255 - K6FD Multiband Antenna with T-tuner
256 - Dual Band Dipole Antenna for 17m, 10m
257 - Compact Travel Dipole Antenna for 10m, 15m, 20m
258 - End-Fire Antenna for 160m, 80m, 40m, 20m
259 - Four Band Loop for 40m, 20m, 15m, 10m

116 - Hybrid Vee for 20m, 17m
117 - Six-Shooter Array Antenna, Gain=7.5dB
118 - Multiband Ground-Plane Antenna for 10m, 15m, 20m
119 - Wire Superbeam Antenna for 10m, 15m, 20m
120 - Two Elements Delta-Loop Antenna
121 - Sterba Curtain Antenna
122 - Half-Wave Vertical Zepp Antenna
123 - Lazy-Loop Antenna for 40m
124 - Terminated Folded Dipole for 80m, 40m
125 - Short-Fat Antenna for 15m
126 - Cobra Antenna for 80m
127 - Log-Periodic Wire Antenna for 20m, 15m, 10m
128 - 5-Element Log-Periodic Vertical Antenna for 80m, 40m, 20m
129 - 2m Vertical Wire Antennas
130 - Earth-Mover Inverted-V Antenna for 40m
131 - Coax-Cable Collinear Antennas
132 - Double Bobtail Antenna for 20m
133 - Collinear Zepp Antenna
134 - Taylor Vee Antenna for 20m
135 - Collinear Vertical Antenna 6dB-Gain for 2m, 1.3m, 73cm
136 - Bi-Loop Antenna for 20m
137 - Wire Beam 6dBd-Gain for 10m
138 - Sloping Diamond Antenna 4dB-Gain for 40m
139 - Twisted Loop Antenna for 160m
140 - DX RX Loop Antenna for 160m
141 - Hentenna 3dB-Gain for 10m, 6m, 2m
142 - VK2AAR Wire Antenna for 20m
143 - 2-Elements Quad Antenna for 6m
144 - Hula-Loop Bidirectional 6dB-Gain Antenna for 17m
145 - Moxon Rectangle Beam for 15m-10m
146 - Double-D Beam 4dB-Gain
260 - Carolina Windom Antenna for 80 to 10m
261 - Shortened Loaded Dipole for 40m
262 - Shortened Loaded Dipole for 80m
263 - Low Radiation Angle Inverted Delta Loop for 80m
264 - Compact Dipole Antenna for 40m, 15m
265 - Vertical Delta Loop for 40m to 10m
266 - Delta Loop Antenna for 17m
267 - Bi-Square Beam Gain=3dBd for 17m
268 -Trap Dipole Antenna for 17m,13m
269 - Hanging Unipole Antenna for 160m, 80m
270 - Wideband Sloping Vee Antenna - 30m to 6m
271 - 5/8 Wavelength Antenna
272 - C-Pole Vertical Antenna
273 - Bobtail Antenna for 30m
274 - Bobtail Antenna for 40m
275 - Two Elements Vertical Loop Antenna for 40m
276 - Double Extended Zepp Antenna for 40m
277 - Reversible Moxon Flip Beam for 40m
278 - Spitwire Antenna for 40m
279 - Fork Antenna for 40m
280 - Veebeam Antenna 3dB-Gain for All Bands
281 - Vertical Moxon Antenna for 10m
282 - Half-Square Antenna for 30m
283 - Rotatable Vertical Delta Loop for 20m
284 - Mini-Horse Antenna 10dB-Gain for 40m-to-2m
285 - Compact Moxon Antenna 5dB-Gain for 40m
286 - Two Elements Sloping Delta Loop 8dB-Gain for 10m
287 - Two Elements Phased Vertical Antenna 3dB-Gain for 40m

---

01 - Tee Antenna

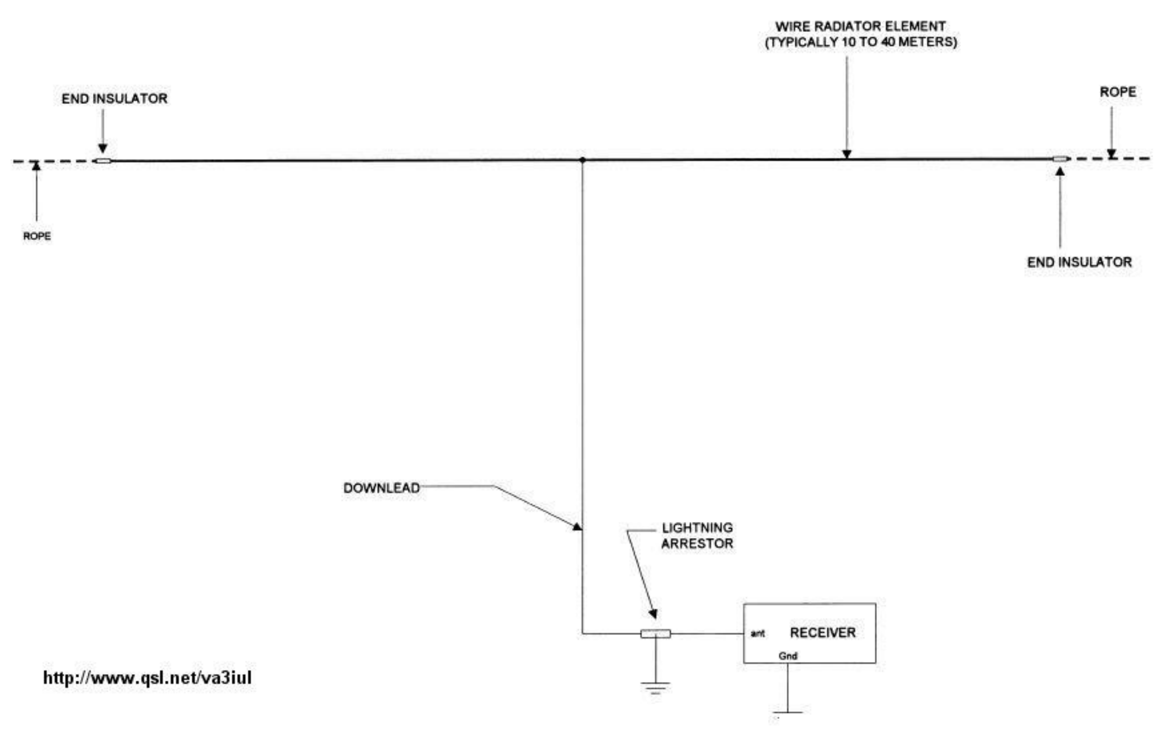
WIRE RADIATOR ELEMENT
(TYPICALLY 10 TO 40 METERS)
END INSULATOR
ROPE
ROPE
END INSULATOR
DOWNLEAD
LIGHTNING
ARRESTOR
ant
RECEIVER
Gnd
http://www.qsl.net/va3iul

02 - Half-Lamda Tee Antenna

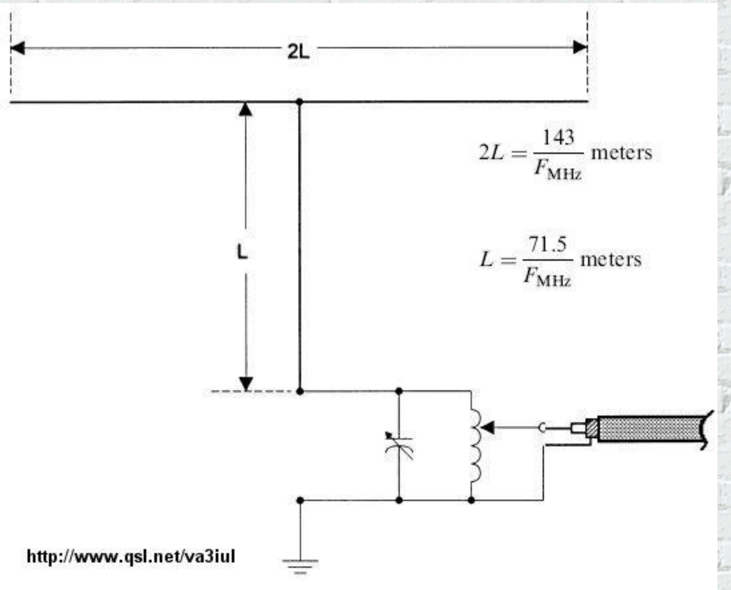
2L
L
http://www.qsl.net/va3iul

$$2L = \frac{143}{F_{\mathrm{MHz}}} \text{ meters}$$

$$L = \frac{71.5}{F_{\mathrm{MHz}}} \text{ meters}$$

03 - Twin-Led Marconi Antenna

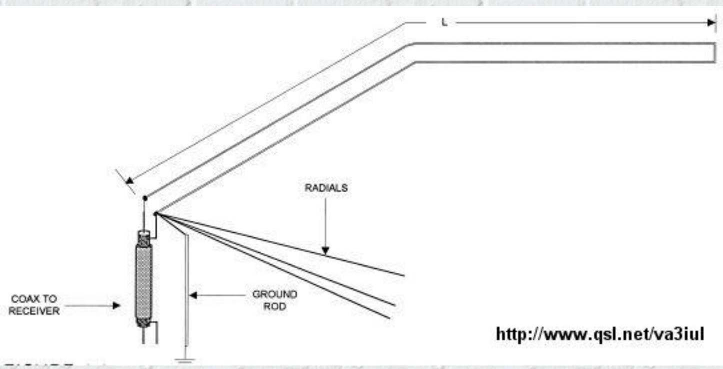
L
RADIALS
COAX TO
RECEIVER
GROUND
ROD
http://www.qsl.net/va3iul

04 - Swallow-Tail Antenna

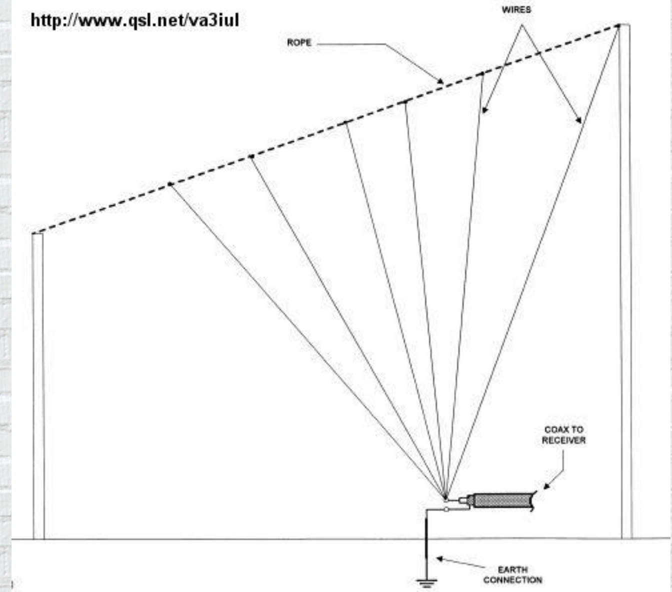
http://www.qsl.net/va3iul
WIRES
ROPE
COAX TO RECEIVER
EARTH CONNECTION

05 - Random Length Radiator Wire Antenna

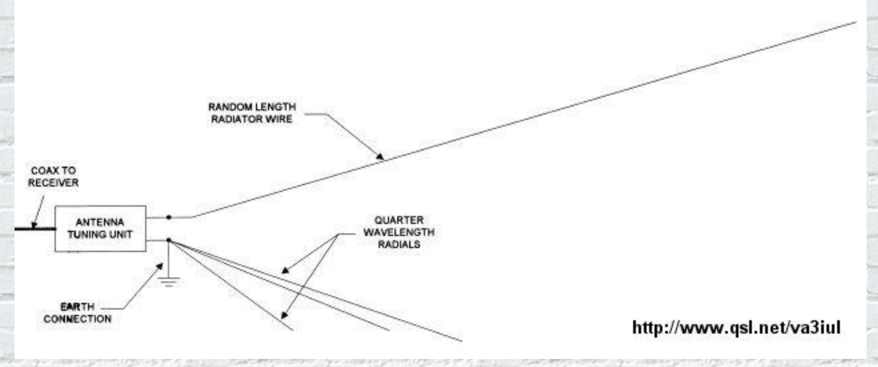
RANDOM LENGTH RADIATOR WIRE
COAX TO RECEIVER
ANTENNA TUNING UNIT
QUARTER WAVELENGTH RADIALS
EARTH CONNECTION
http://www.qsl.net/va3iul

06 - Windom Antenna

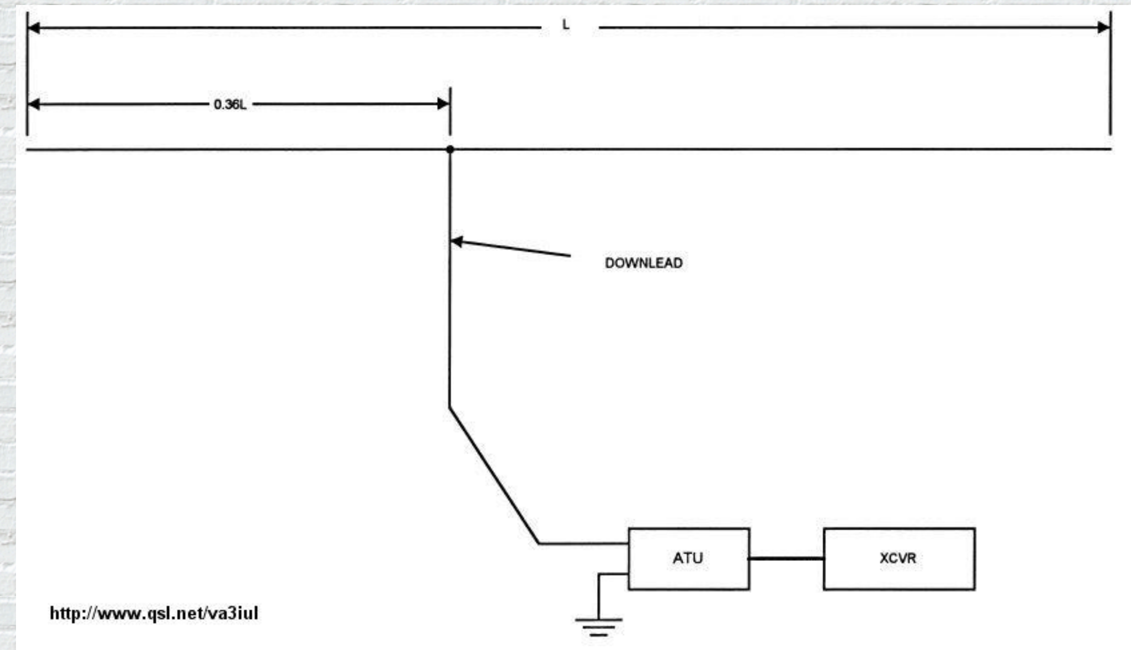
L
0.36L
DOWNLEAD
ATU
XCVR
http://www.qsl.net/va3iul

07 - Windom Antenna - Feed with coax cable

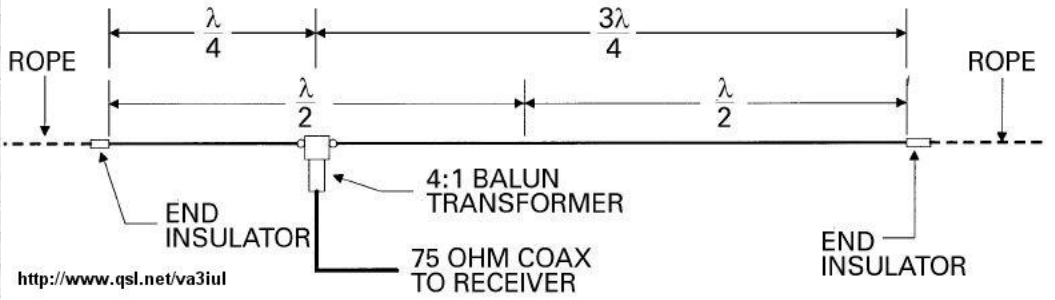
ROPE
$\frac{\lambda}{4}$
$\frac{3\lambda}{4}$
ROPE
$\frac{\lambda}{2}$
$\frac{\lambda}{2}$
4:1 BALUN TRANSFORMER
END INSULATOR
75 OHM COAX TO RECEIVER
END INSULATOR
http://www.qsl.net/va3iul

08 - Quarter Wavelength Vertical Antenna

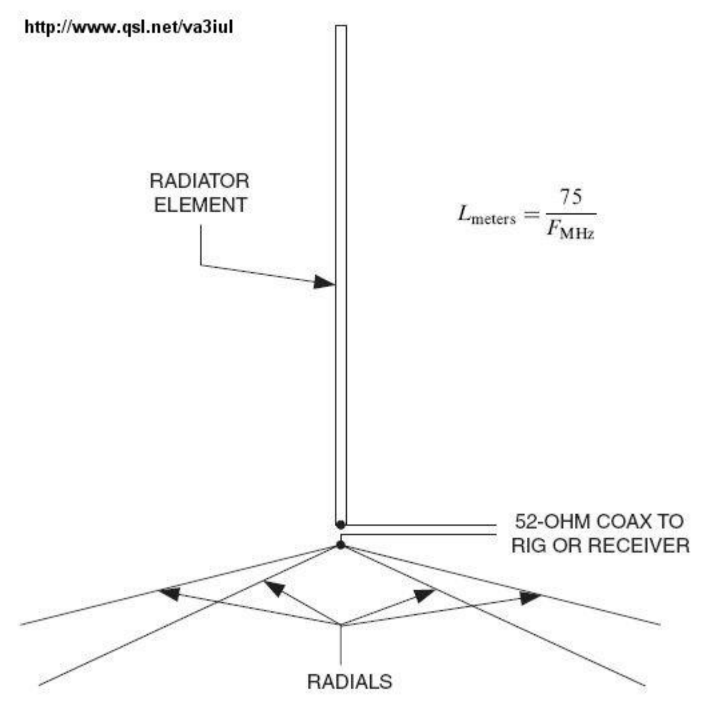
http://www.qsl.net/va3iul
RADIATOR ELEMENT
$L_{\text{meters}} = \frac{75}{F_{\text{MHz}}}$
52-OHM COAX TO RIG OR RECEIVER
RADIALS

09 - Folded Marconi Tee Antenna

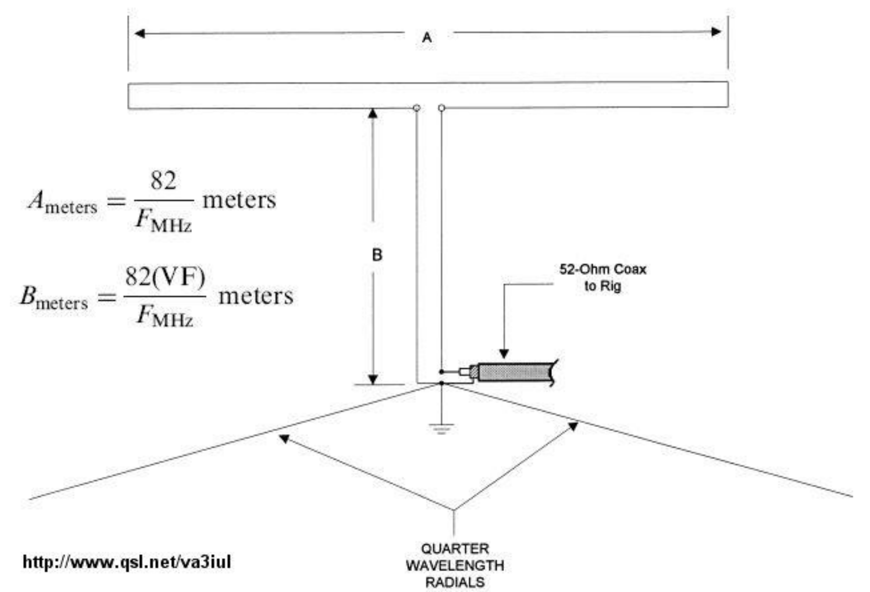
A
B
$A_{\text{meters}} = \frac{82}{F_{\text{MHz}}}$ meters
$B_{\text{meters}} = \frac{82(VF)}{F_{\text{MHz}}}$ meters
52-Ohm Coax to Rig
http://www.qsl.net/va3iul
QUARTER WAVELENGTH RADIALS

10 - Zeppelin Antenna

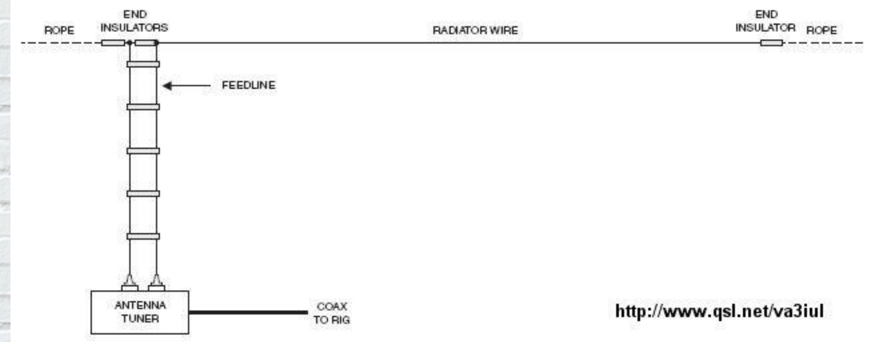
ROPE
END INSULATORS
RADIATOR WIRE
END INSULATOR
ROPE
FEEDLINE
ANTENNA TUNER
COAX TO RIG
http://www.qsl.net/va3iul

11 - EWE Antenna

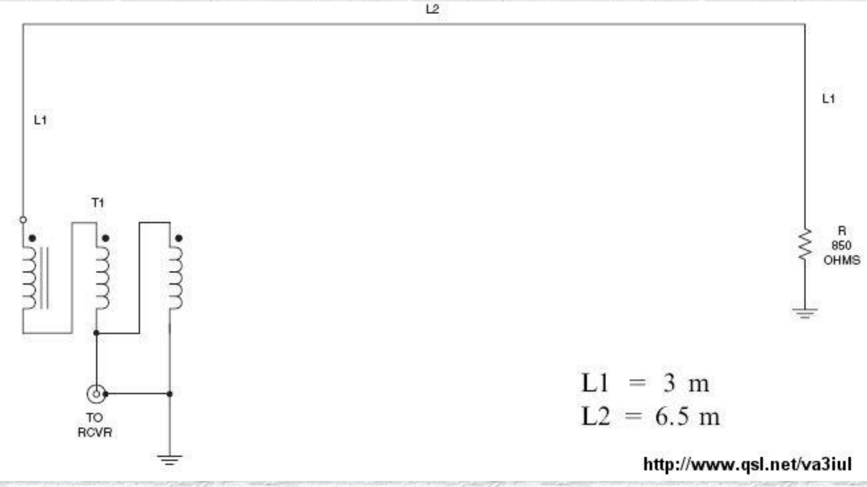
L2
L1
L1
T1
R 850 OHMS
TO RCVR
L1 = 3 m
L2 = 6.5 m
http://www.qsl.net/va3iul

12 - Dipole Antenna - Balun

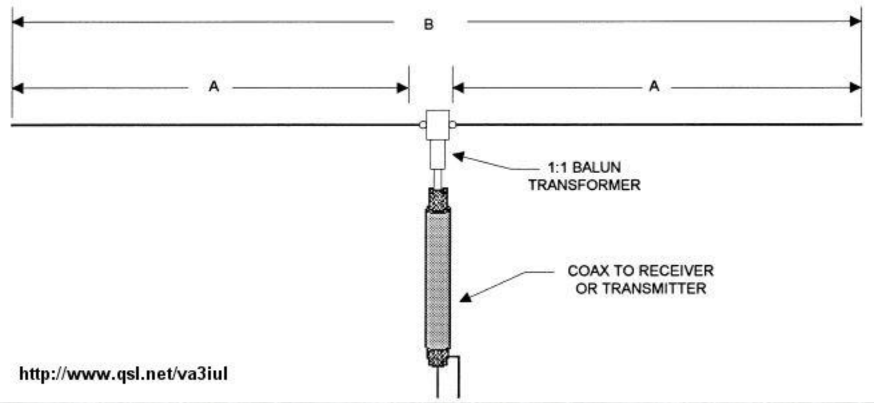
B
A
A
1:1 BALUN TRANSFORMER
COAX TO RECEIVER OR TRANSMITTER
http://www.qsl.net/va3iul

13 - Multiband Dipole Antenna

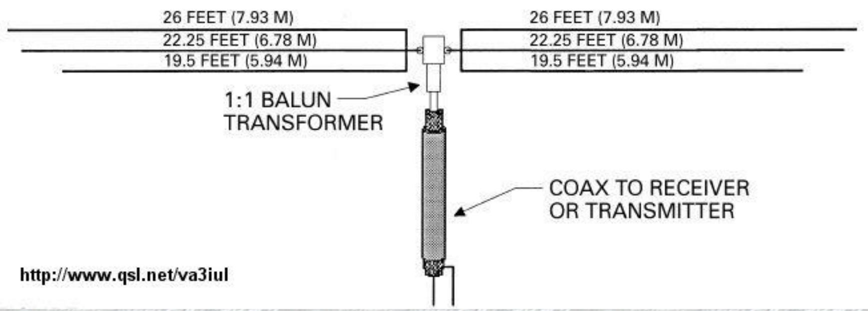
26 FEET (7.93 M)
22.25 FEET (6.78 M)
19.5 FEET (5.94 M)
26 FEET (7.93 M)
22.25 FEET (6.78 M)
19.5 FEET (5.94 M)
1:1 BALUN TRANSFORMER
COAX TO RECEIVER OR TRANSMITTER
http://www.qsl.net/va3iul

14 - Inverted-Vee Antenna

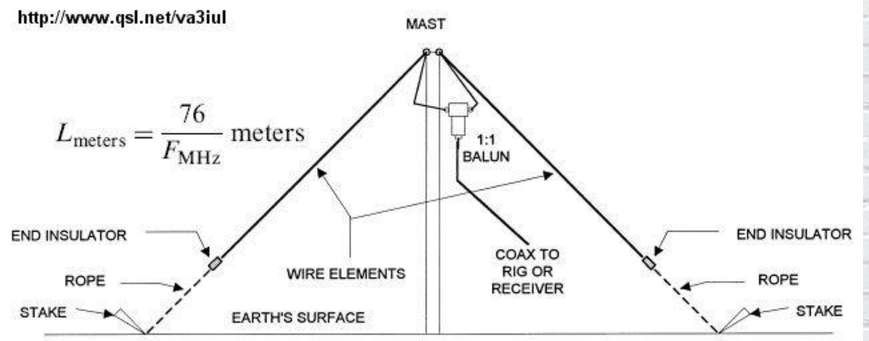
http://www.qsl.net/va3iul
$L_{meters} = \frac{76}{F_{MHz}}$ meters
MAST
1:1
BALUN
END INSULATOR
ROPE
STAKE
WIRE ELEMENTS
COAX TO
RIG OR
RECEIVER
EARTH'S SURFACE
END INSULATOR
ROPE
STAKE

---

15 - Sloping Dipole Antenna

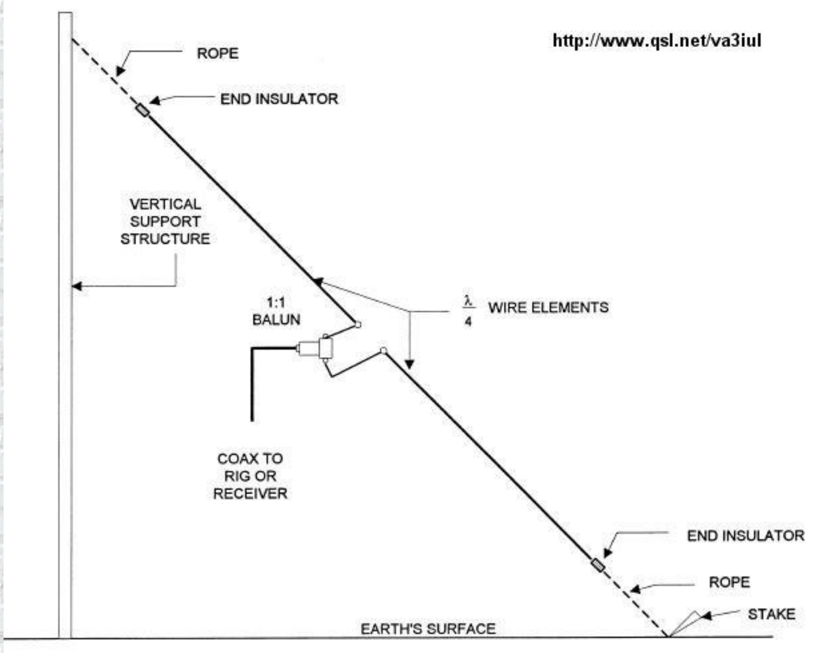
ROPE
END INSULATOR
http://www.qsl.net/va3iul
VERTICAL
SUPPORT
STRUCTURE
1:1
BALUN
$\frac{\lambda}{4}$ WIRE ELEMENTS
COAX TO
RIG OR
RECEIVER
END INSULATOR
ROPE
STAKE
EARTH'S SURFACE

---

16 - Vertical Dipole

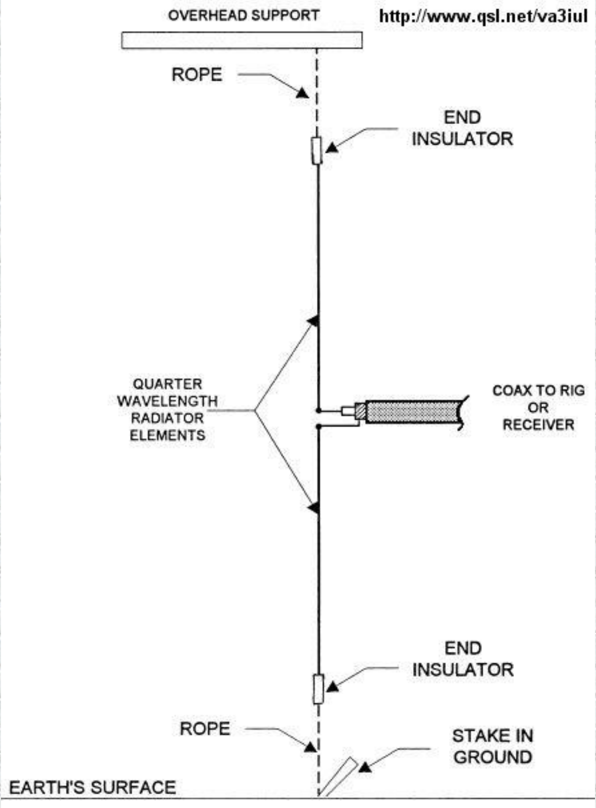
OVERHEAD SUPPORT
http://www.qsl.net/va3iul
ROPE
END INSULATOR
QUARTER WAVELENGTH RADIATOR ELEMENTS
COAX TO RIG OR RECEIVER
END INSULATOR
ROPE
STAKE IN GROUND
EARTH'S SURFACE

17 - Delta Fed Dipole Antenna

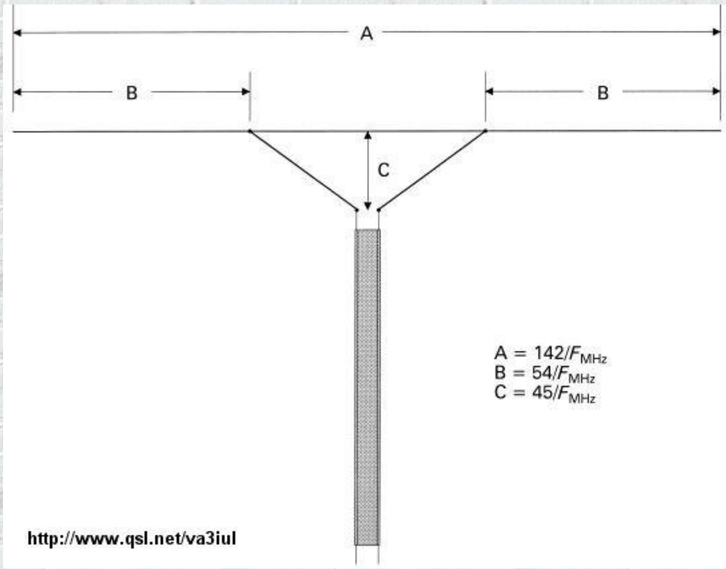
A
B
B
C
$A = 142/F_{MHz}$
$B = 54/F_{MHz}$
$C = 45/F_{MHz}$
http://www.qsl.net/va3iul

18 - Bow-Tie Dipole Antenna

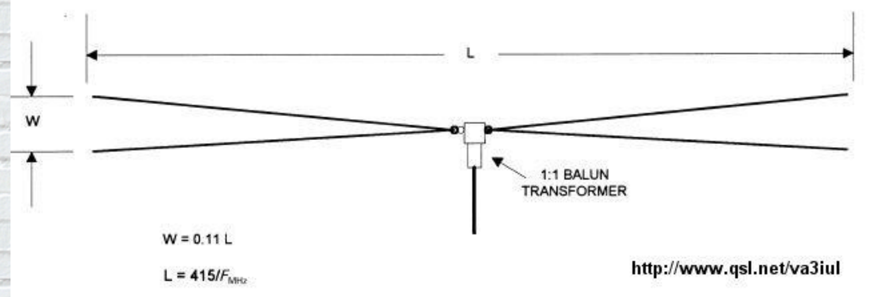
L
W
1:1 BALUN TRANSFORMER
$W = 0.11\ L$
$L = 415/F_{MHz}$
http://www.qsl.net/va3iul

19 - Bow-Tie Folded Dipole Antenna for RX

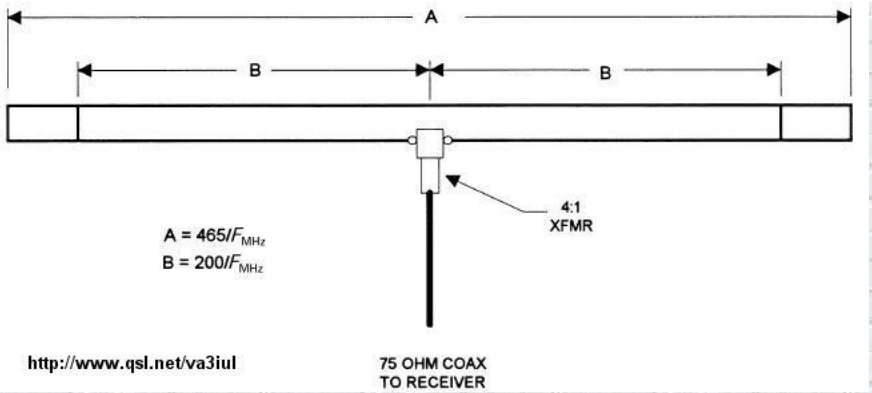
A
B
B
4:1
XFMR
A = 465/F_{MHz}
B = 200/F_{MHz}
http://www.qsl.net/va3iul
75 OHM COAX
TO RECEIVER

20 - Multiband Tuned Doublet Antenna

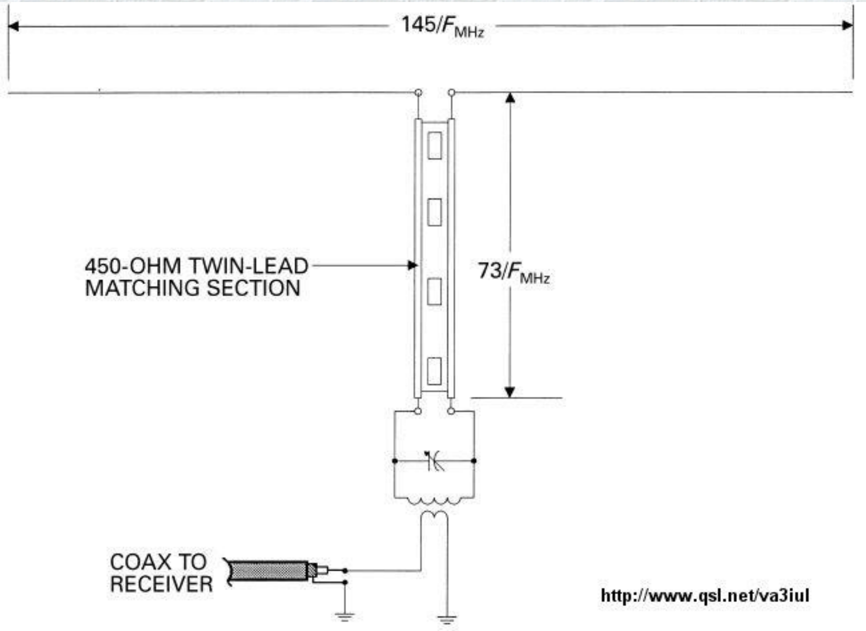
145/F_{MHz}
450-OHM TWIN-LEAD
MATCHING SECTION
73/F_{MHz}
COAX TO
RECEIVER
http://www.qsl.net/va3iul

21 - G5RV Antenna

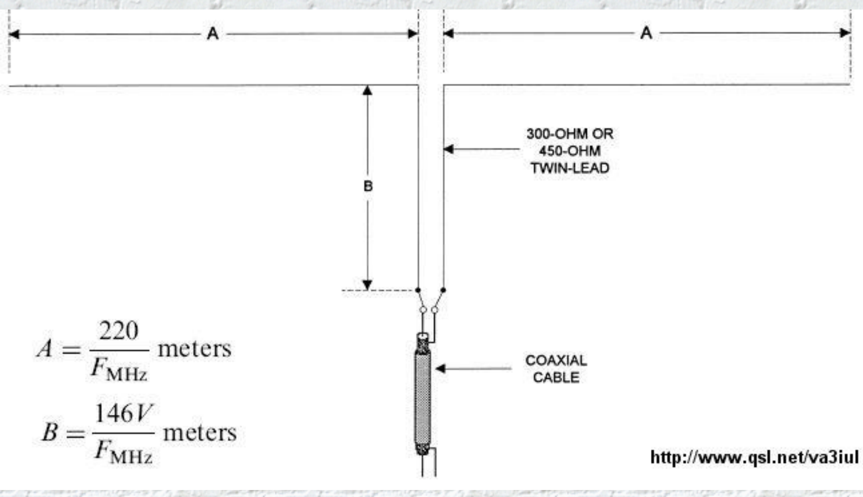
A
A
B
300-OHM OR
450-OHM
TWIN-LEAD
COAXIAL
CABLE
http://www.qsl.net/va3iul

$$A = \frac{220}{F_{\text{MHz}}} \text{ meters}$$

$$B = \frac{146V}{F_{\text{MHz}}} \text{ meters}$$

22 - Wideband Dipole Antenna

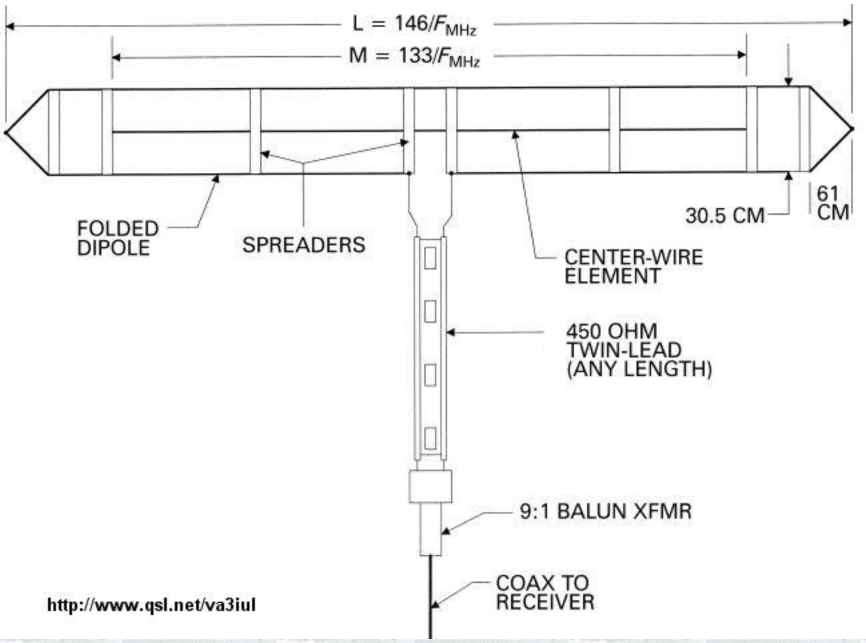
L = 146/FMHz
M = 133/FMHz
FOLDED DIPOLE
SPREADERS
30.5 CM
61 CM
CENTER-WIRE ELEMENT
450 OHM TWIN-LEAD (ANY LENGTH)
9:1 BALUN XFMR
COAX TO RECEIVER
http://www.qsl.net/va3iul

23 - Wideband Dipole for Receiving

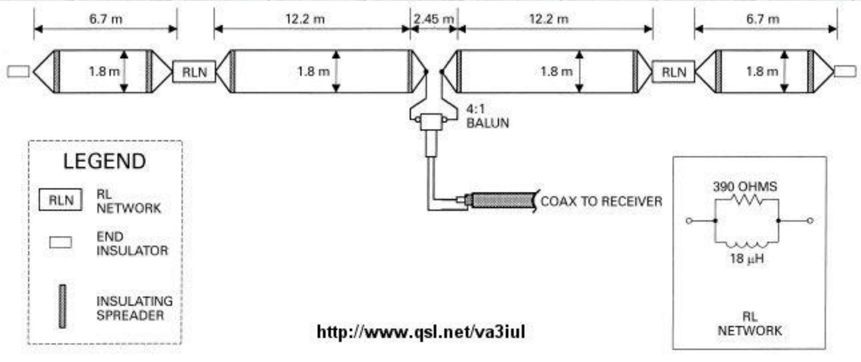
6.7 m
12.2 m
2.45 m
12.2 m
6.7 m
1.8 m
RLN
1.8 m
1.8 m
RLN
1.8 m
4:1 BALUN
LEGEND
RLN
RL NETWORK
END INSULATOR
INSULATING SPREADER
COAX TO RECEIVER
390 OHMS
18 µH
RL NETWORK
http://www.qsl.net/va3iul

24 - Tilted Folded Dipole Antenna

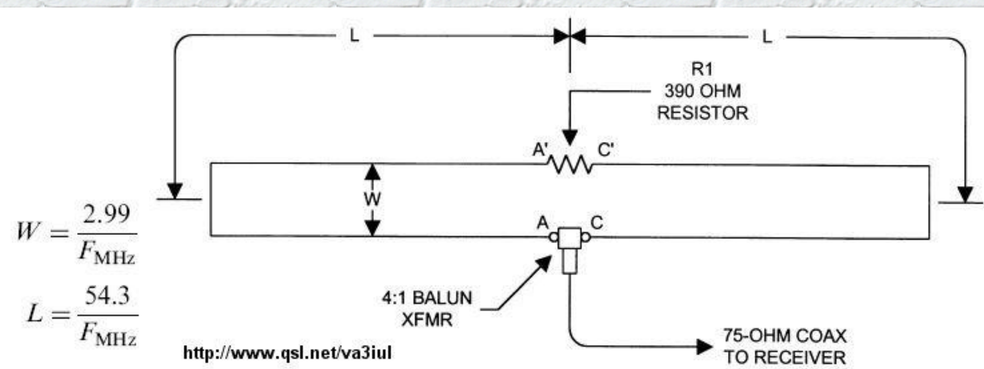
L
L
R1
390 OHM
RESISTOR
A'
C'
W
A
C
W = 2.99/FMHz
L = 54.3/FMHz
4:1 BALUN XFMR
75-OHM COAX TO RECEIVER
http://www.qsl.net/va3iul

25 - Right Angle Marconi Antenna

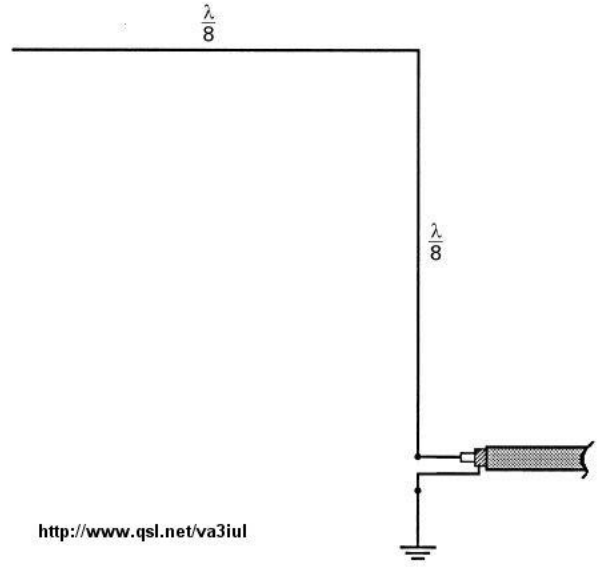
λ/8
λ/8
http://www.qsl.net/va3iul

26 - Linearly Loaded Tee Antenna

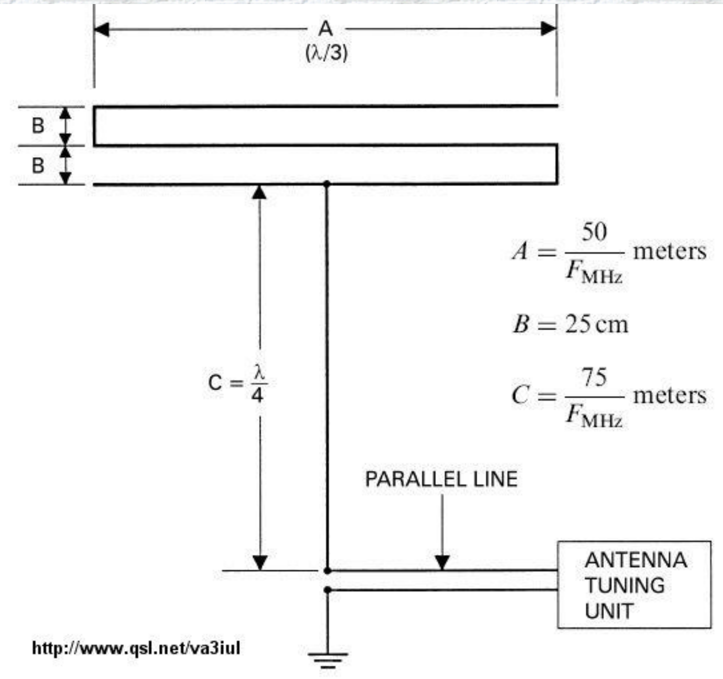
A
(λ/3)
B
B
C = λ/4
PARALLEL LINE
ANTENNA TUNING UNIT
http://www.qsl.net/va3iul

$$A = \frac{50}{F_{\text{MHz}}} \text{ meters}$$

$$B = 25\,\text{cm}$$

$$C = \frac{75}{F_{\text{MHz}}} \text{ meters}$$

27 - Reduced Size Dipole Antenna

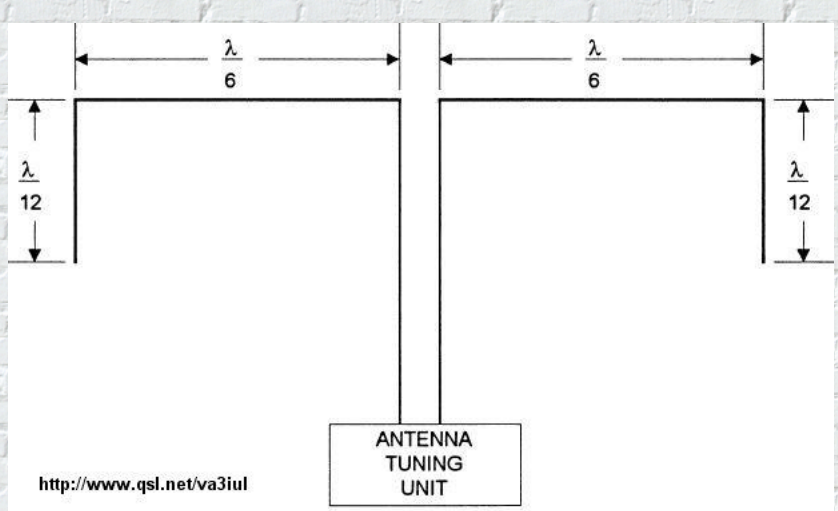
λ/6
λ/6
λ/12
λ/12
ANTENNA TUNING UNIT
http://www.qsl.net/va3iul

28 - Doublet Dipole Antenna

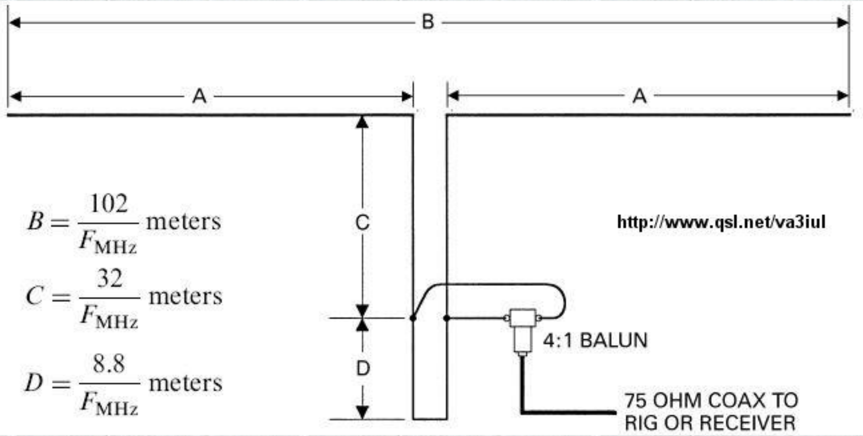
B
A
A
$B = \frac{102}{F_{\mathrm{MHz}}}$ meters
$C = \frac{32}{F_{\mathrm{MHz}}}$ meters
$D = \frac{8.8}{F_{\mathrm{MHz}}}$ meters
C
D
http://www.qsl.net/va3iul
4:1 BALUN
75 OHM COAX TO RIG OR RECEIVER

---

29 - Delta Loop Antenna

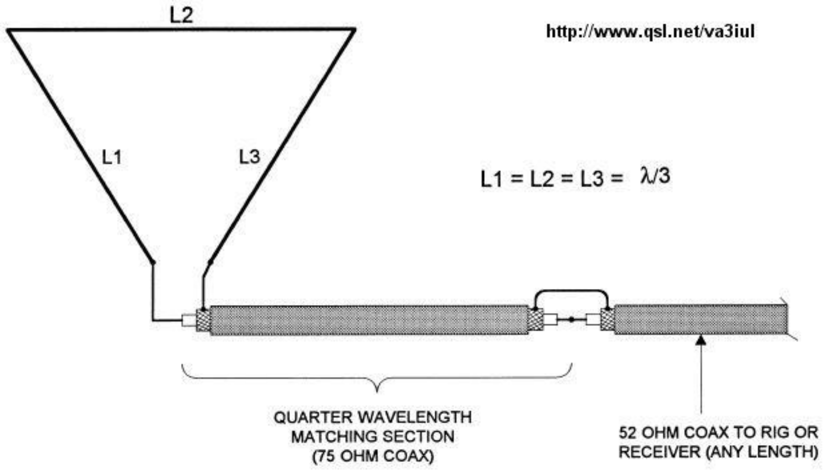
L2
http://www.qsl.net/va3iul
L1
L3
L1 = L2 = L3 = λ/3
QUARTER WAVELENGTH MATCHING SECTION (75 OHM COAX)
52 OHM COAX TO RIG OR RECEIVER (ANY LENGTH)

---

30 - Half Delta Loop Antenna

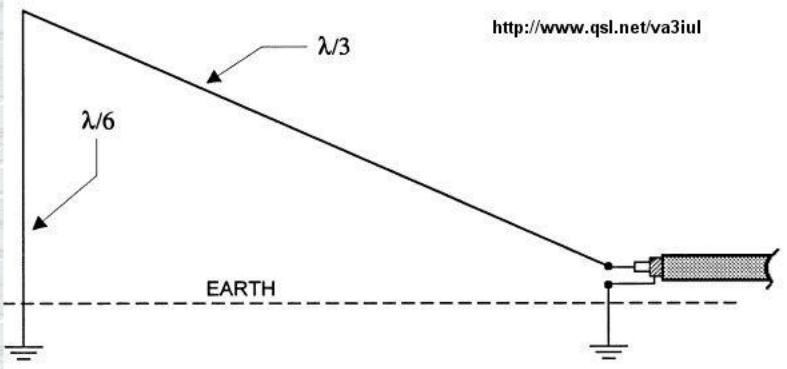
http://www.qsl.net/va3iul
λ/3
λ/6
EARTH

---

31 - Collinear Franklin Antenna

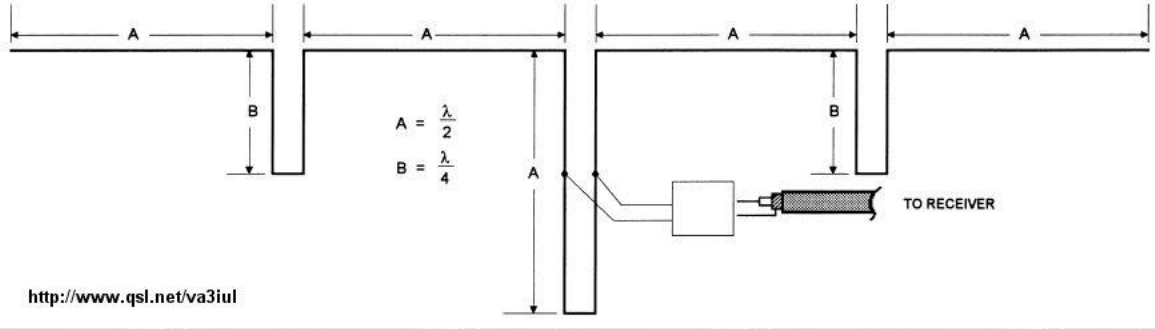
A
A
A
A
B
$A = \frac{\lambda}{2}$
$B = \frac{\lambda}{4}$
A
B
TO RECEIVER
http://www.qsl.net/va3iul

32 - Four Element Broadside Antenna

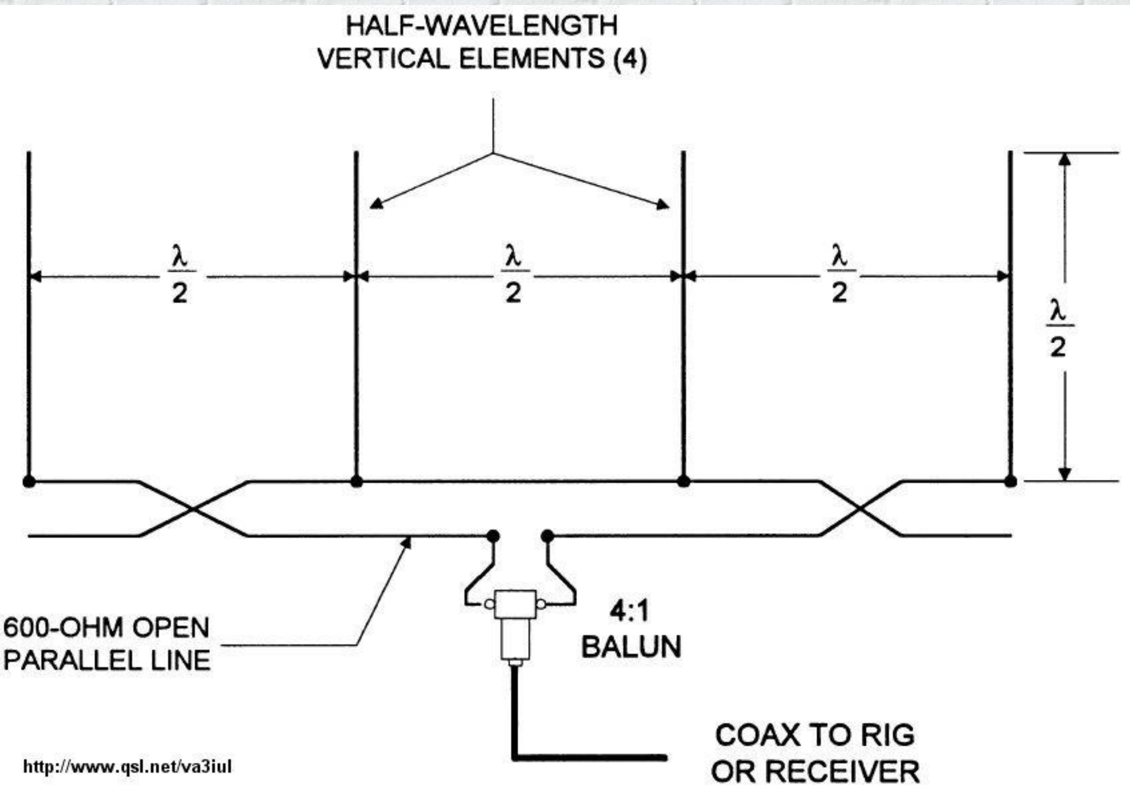
HALF-WAVELENGTH
VERTICAL ELEMENTS (4)
λ/2
λ/2
λ/2
λ/2
600-OHM OPEN
PARALLEL LINE
4:1
BALUN
COAX TO RIG
OR RECEIVER
http://www.qsl.net/va3iul

33 - The Lazy-H Array Antenna

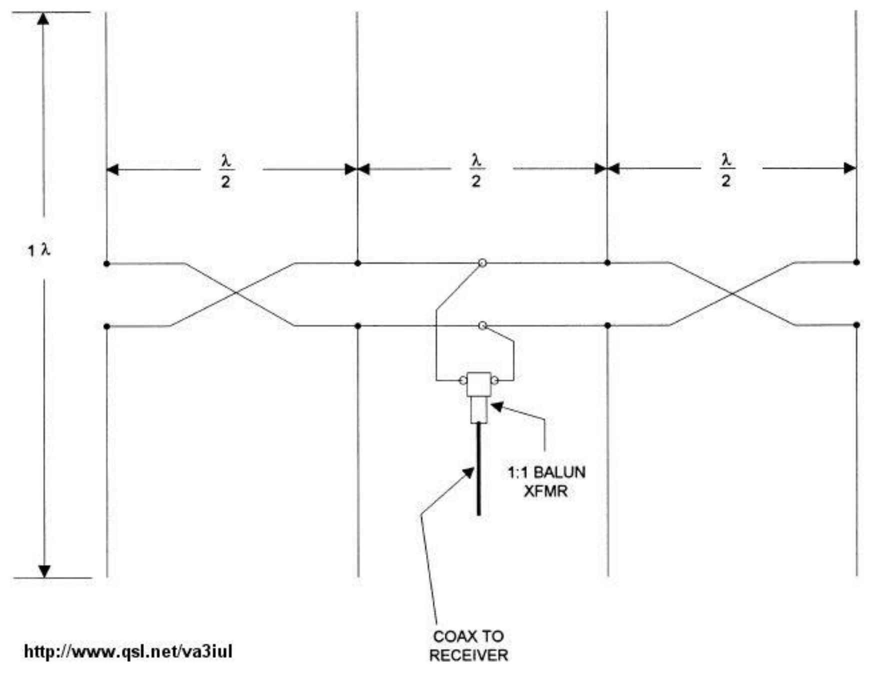
1 λ
λ/2
λ/2
λ/2
1:1 BALUN
XFMR
COAX TO
RECEIVER
http://www.qsl.net/va3iul

34 - Sterba Curtain Array Antenna

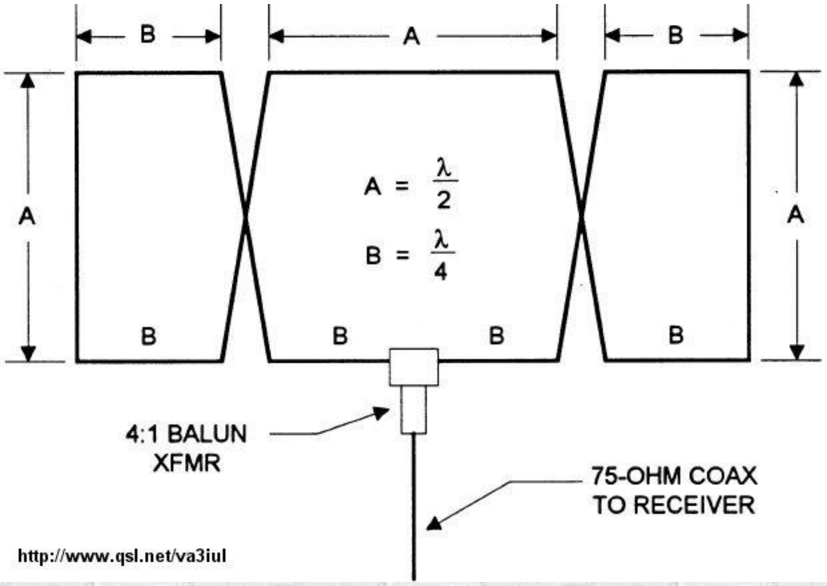
B
A
B
A
A = λ/2
B = λ/4
A
B
B
B
B
4:1 BALUN XFMR
75-OHM COAX TO RECEIVER
http://www.qsl.net/va3iul

---

35 - T-L DX Antenna

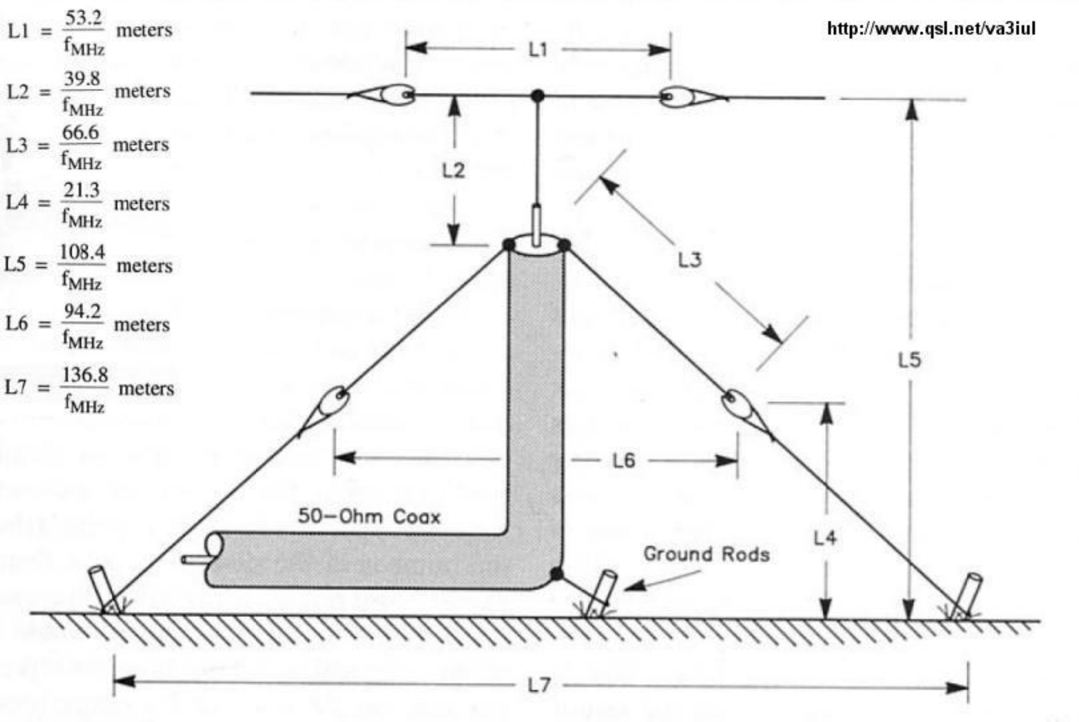
L1 = 53.2 / fMHz meters
L2 = 39.8 / fMHz meters
L3 = 66.6 / fMHz meters
L4 = 21.3 / fMHz meters
L5 = 108.4 / fMHz meters
L6 = 94.2 / fMHz meters
L7 = 136.8 / fMHz meters
http://www.qsl.net/va3iul
L1
L2
L3
L5
L6
L4
50-Ohm Coax
Ground Rods
L7

---

36 - 1.9 MHz Full-wave Loop Antenna

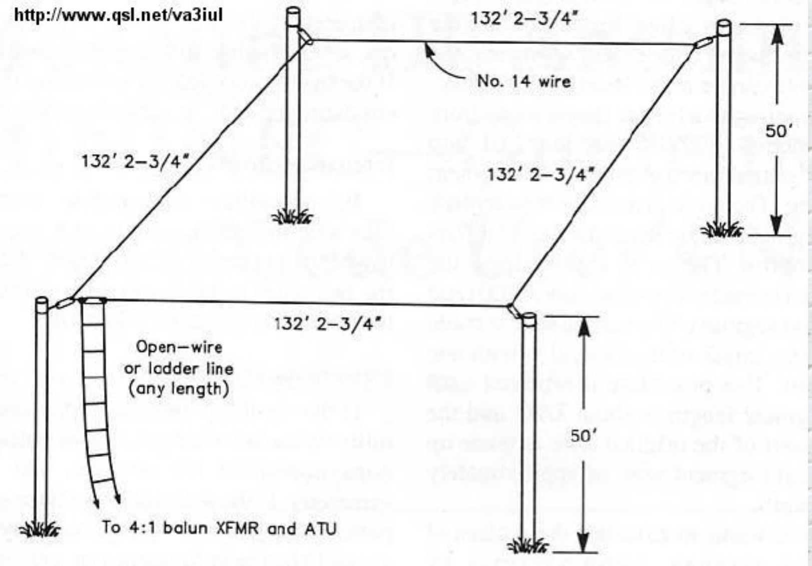
http://www.qsl.net/va3iul
132′ 2–3/4″
No. 14 wire
50′
132′ 2–3/4″
132′ 2–3/4″
132′ 2–3/4″
Open–wire
or ladder line
(any length)
50′
To 4:1 balun XFMR and ATU

---

37 - Multi-Band Portable Antenna

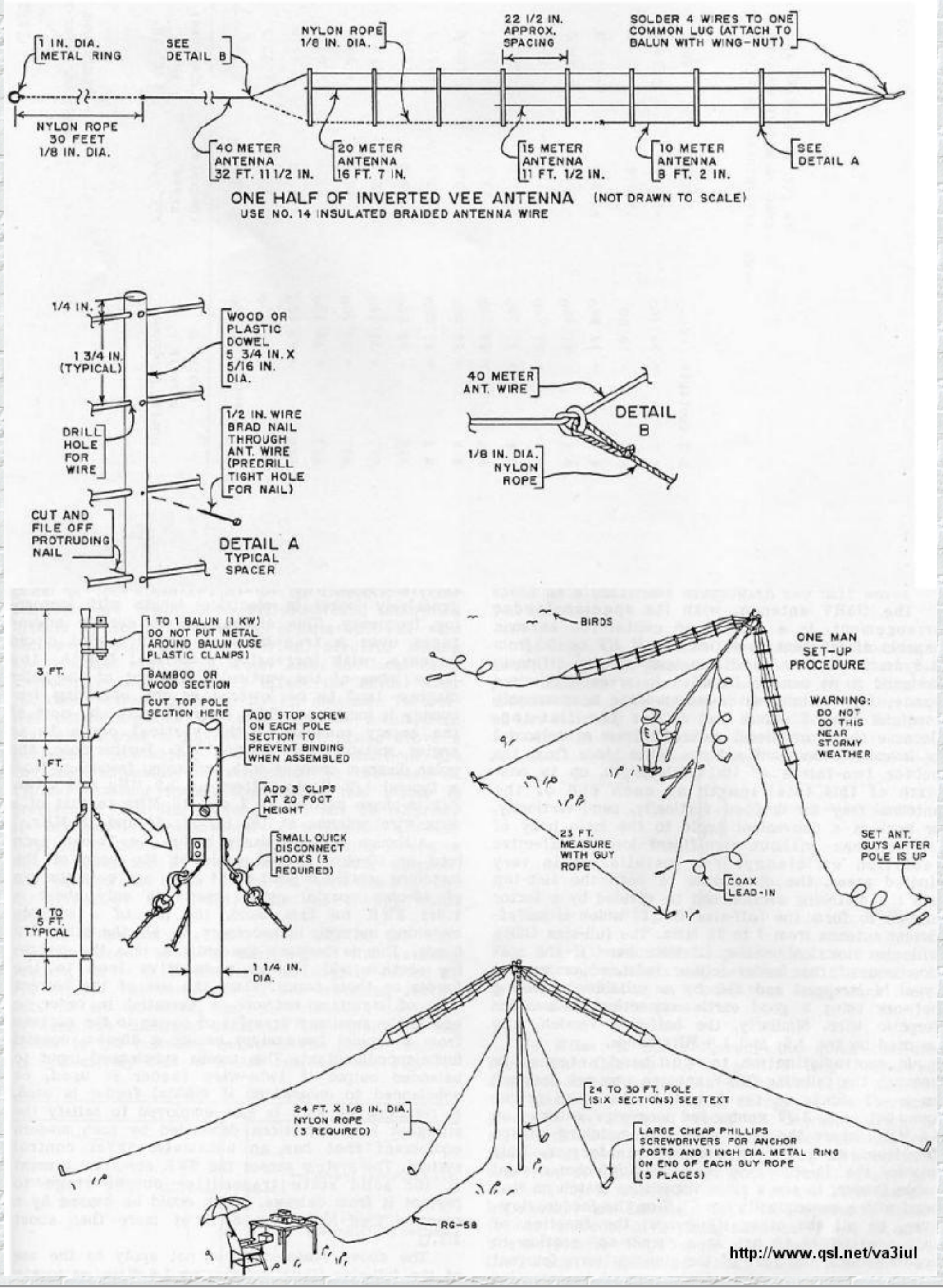
1 IN. DIA. METAL RING
SEE DETAIL B
NYLON ROPE 1/8 IN. DIA.
22 1/2 IN. APPROX. SPACING
SOLDER 4 WIRES TO ONE COMMON LUG (ATTACH TO BALUN WITH WING-NUT)
NYLON ROPE 30 FEET 1/8 IN. DIA.
40 METER ANTENNA 32 FT. 11 1/2 IN.
20 METER ANTENNA 16 FT. 7 IN.
15 METER ANTENNA 11 FT. 1/2 IN.
10 METER ANTENNA 8 FT. 2 IN.
SEE DETAIL A
ONE HALF OF INVERTED VEE ANTENNA (NOT DRAWN TO SCALE)
USE NO. 14 INSULATED BRAIDED ANTENNA WIRE
1/4 IN.
1 3/4 IN. (TYPICAL)
WOOD OR PLASTIC DOWEL 5 3/4 IN. X 5/16 IN. DIA.
DRILL HOLE FOR WIRE
1/2 IN. WIRE BRAD NAIL THROUGH ANT. WIRE (PREDRILL TIGHT HOLE FOR NAIL)
CUT AND FILE OFF PROTRUDING NAIL
DETAIL A TYPICAL SPACER
40 METER ANT. WIRE
DETAIL B
1/8 IN. DIA. NYLON ROPE
1 TO 1 BALUN (1 KW) DO NOT PUT METAL AROUND BALUN (USE PLASTIC CLAMPS)
3 FT.
BAMBOO OR WOOD SECTION
CUT TOP POLE SECTION HERE
ADD STOP SCREW ON EACH POLE SECTION TO PREVENT BINDING WHEN ASSEMBLED
1 FT.
ADD 3 CLIPS AT 20 FOOT HEIGHT
SMALL QUICK DISCONNECT HOOKS (3 REQUIRED)
4 TO 5 FT. TYPICAL
1 1/4 IN. DIA.
BIRDS
ONE MAN SET-UP PROCEDURE
WARNING: DO NOT DO THIS NEAR STORMY WEATHER
23 FT. MEASURE WITH GUY ROPE
SET ANT. GUYS AFTER POLE IS UP
COAX LEAD-IN
24 TO 30 FT. POLE (SIX SECTIONS) SEE TEXT
24 FT. X 1/8 IN. DIA. NYLON ROPE (3 REQUIRED)
LARGE CHEAP PHILLIPS SCREWDRIVERS FOR ANCHOR POSTS AND 1 INCH DIA. METAL RING ON END OF EACH GUY ROPE (3 PLACES)
RG-58
http://www.qsl.net/va3iul

38 - Off-center-fed Full-wave Doublet Antenna

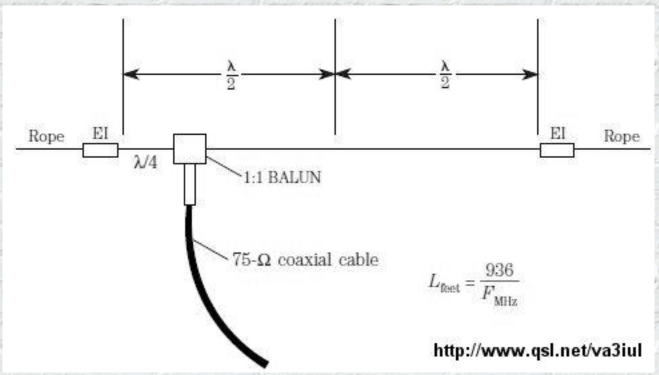
$\frac{\lambda}{2}$
$\frac{\lambda}{2}$
Rope
EI
EI
Rope
λ/4
1:1 BALUN
75-Ω coaxial cable
$L_{\text{feet}} = \frac{936}{F_{\text{MHz}}}$
http://www.qsl.net/va3iul

39 - Terminated Sloper Antenna

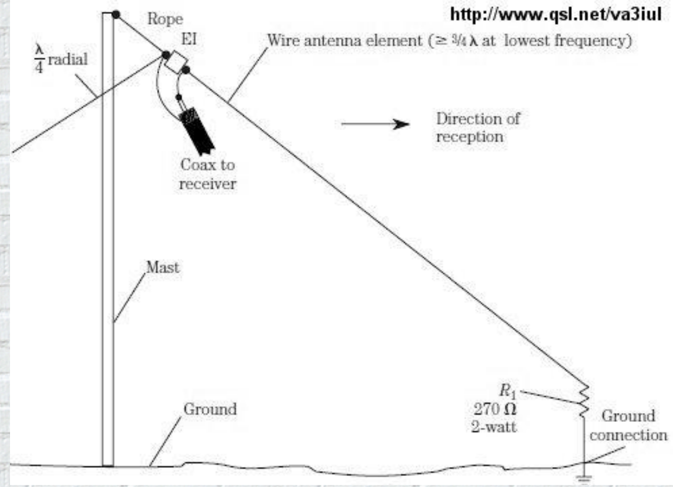
http://www.qsl.net/va3iul
Rope
EI
$\frac{\lambda}{4}$ radial
Wire antenna element (≥ ¾ λ at lowest frequency)
Direction of reception
Coax to receiver
Mast
Ground
R_1
270 Ω
2-watt
Ground connection

40 - Double Extended Zepp Antenna

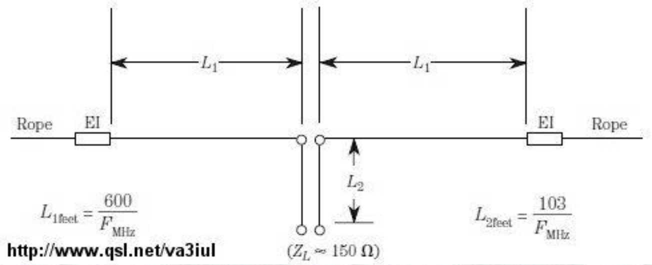
L_1
L_1
Rope
EI
EI
Rope
L_2
$L_{1\text{feet}} = \frac{600}{F_{\text{MHz}}}$
$L_{2\text{feet}} = \frac{103}{F_{\text{MHz}}}$
http://www.qsl.net/va3iul
$(Z_L \approx 150\ \Omega)$

41 - TCFTFD Dipole Antenna

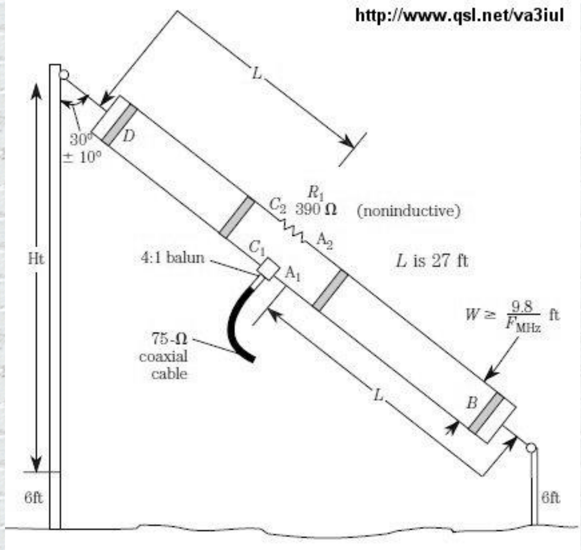
http://www.qsl.net/va3iul
L
30°
± 10°
D
C_2
R_1
390 Ω
(noninductive)
A_2
C_1
A_1
4:1 balun
Ht
L is 27 ft
$W \geq \frac{9.8}{F_{\text{MHz}}}$ ft
75-Ω coaxial cable
L
B
6ft
6ft

42 - Vee-Sloper Antenna

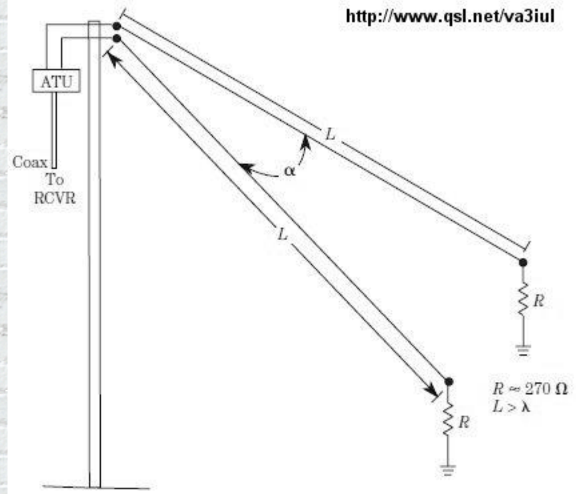
http://www.qsl.net/va3iul
ATU
Coax
To
RCVR
L
α
L
R
R
R ≈ 270 Ω
L > λ

---

43 - Rhombic Inverted-Vee Antenna

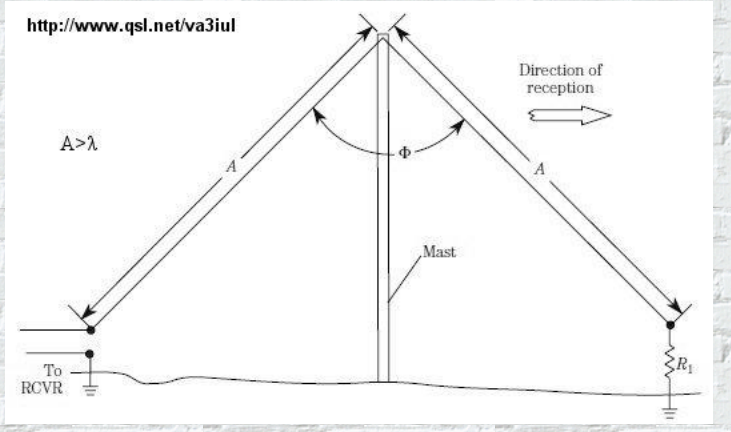
http://www.qsl.net/va3iul
A>λ
A
Φ
A
Direction of reception
Mast
To
RCVR
R1

---

44 - Counterpoise Longwire

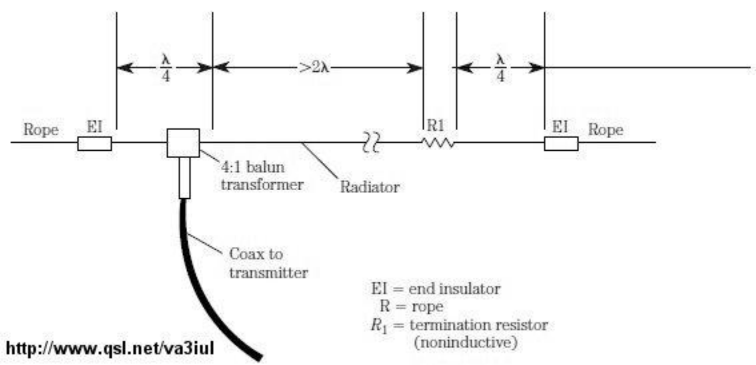
λ/4
>2λ
λ/4
Rope
EI
R1
EI
Rope
4:1 balun transformer
Radiator
Coax to transmitter
EI = end insulator
R = rope
R1 = termination resistor (noninductive)
http://www.qsl.net/va3iul

---

45 - Bisquare Loop Antenna

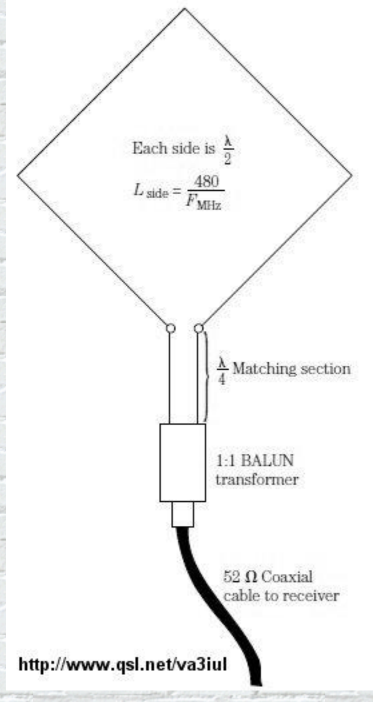
Each side is $\frac{\lambda}{2}$
$L_{side} = \frac{480}{F_{MHz}}$
$\frac{\lambda}{4}$ Matching section
1:1 BALUN transformer
52 Ω Coaxial cable to receiver
http://www.qsl.net/va3iul

---

46 - Piggyback Antenna for 10m

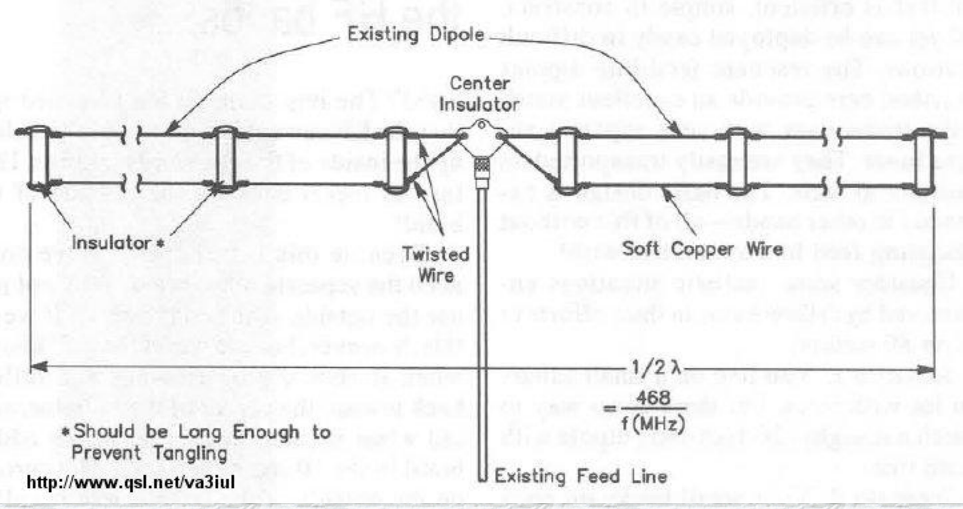
Existing Dipole
Center Insulator
Insulator *
Twisted Wire
Soft Copper Wire
1/2 λ
$= \frac{468}{f(MHz)}$
* Should be Long Enough to Prevent Tangling
Existing Feed Line
http://www.qsl.net/va3iul

---

47 - Vertical Sleeve Antenna for 10m

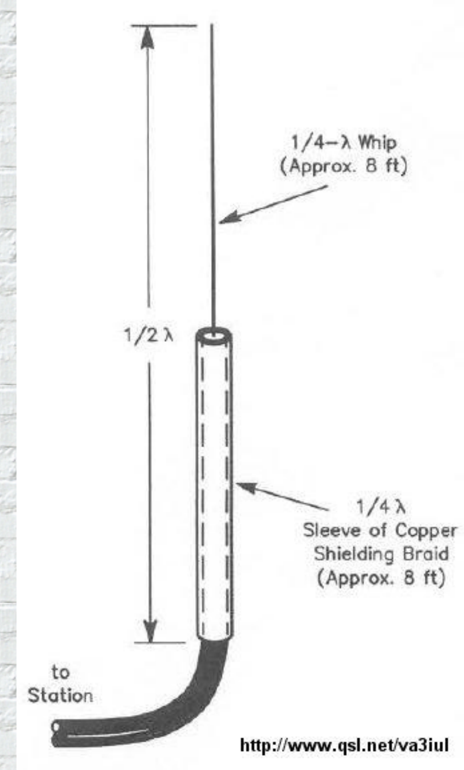
1/4-λ Whip
(Approx. 8 ft)
1/2 λ
1/4 λ
Sleeve of Copper
Shielding Braid
(Approx. 8 ft)
to
Station
http://www.qsl.net/va3iul

---

48 - Double Windom Antenna

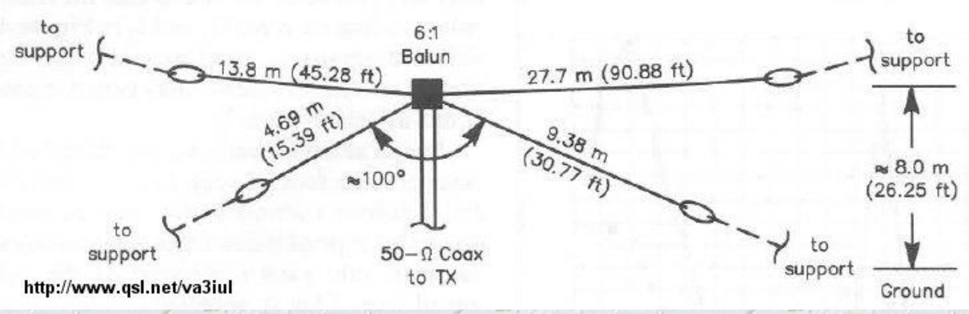
to
support
13.8 m (45.28 ft)
6:1
Balun
27.7 m (90.88 ft)
to
support
4.69 m
(15.39 ft)
≈100°
9.38 m
(30.77 ft)
≈ 8.0 m
(26.25 ft)
to
support
50-Ω Coax
to TX
to
support
Ground
http://www.qsl.net/va3iul

---

49 - Double Windom for 9 Bands

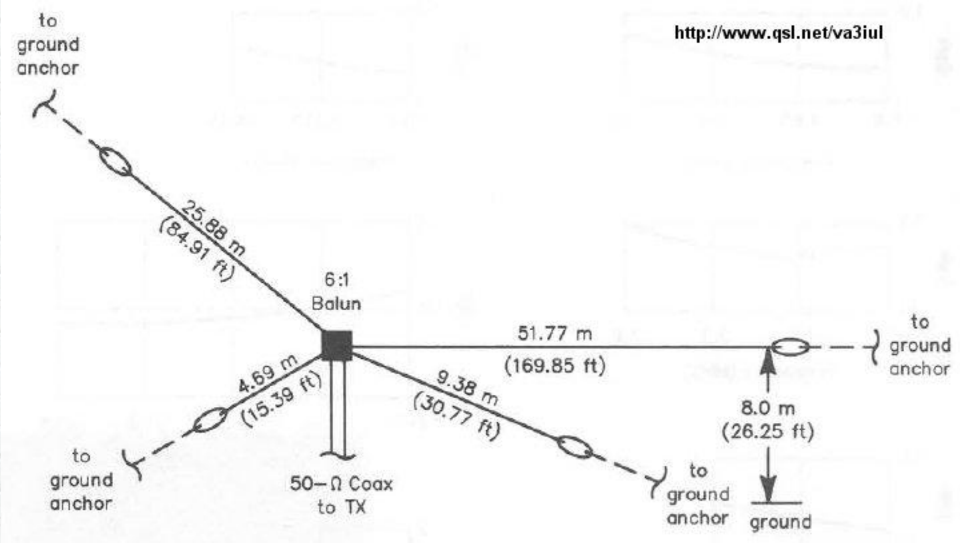
to
ground
anchor
http://www.qsl.net/va3iul
25.88 m
(84.91 ft)
6:1
Balun
51.77 m
(169.85 ft)
to
ground
anchor
4.69 m
(15.39 ft)
9.38 m
(30.77 ft)
8.0 m
(26.25 ft)
to
ground
anchor
50-Ω Coax
to TX
to
ground
anchor
ground

50 - Collinear Trap Antenna

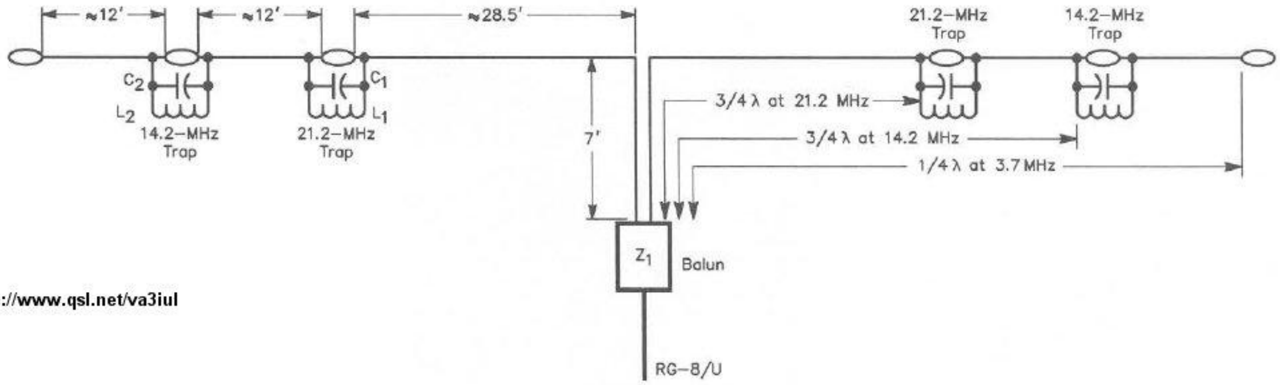
≈12'
≈12'
≈28.5'
C2
L2
14.2–MHz Trap
21.2–MHz Trap
C1
L1
7'
21.2–MHz Trap
14.2–MHz Trap
3/4 λ at 21.2 MHz
3/4 λ at 14.2 MHz
1/4 λ at 3.7 MHz
Z1
Balun
RG–8/U

http://www.qsl.net/va3iul

C1, C2—25-pf. 6000-volt disk ceramic. See text.
L_1—Approx. 2 µH—$4\frac{3}{4}$ turns No. 18, $2\frac{1}{4}$-inches diam., $\frac{3}{8}$ inch long, or 5 turns No. 18, 2-inch diam., 16 tpi. See text.

L_2—Approx. 5 µH.—8 turns No. 18, $2\frac{1}{4}$-inch dia., $\frac{5}{8}$ inch long, or 9 turns No. 18, 2-inch diam. 16 tpi. See text.
Z_1—1 to 1 balun.

---

51 - Short Dipole Antenna for 40m - 80m - 160m

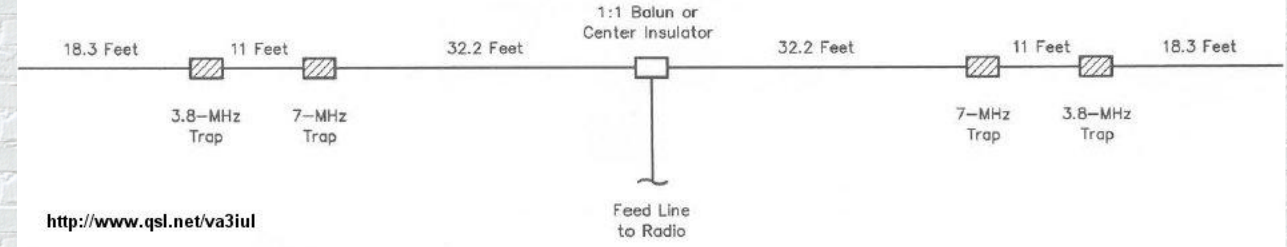
1:1 Balun or Center Insulator
18.3 Feet
11 Feet
32.2 Feet
32.2 Feet
11 Feet
18.3 Feet
3.8–MHz Trap
7–MHz Trap
7–MHz Trap
3.8–MHz Trap
Feed Line to Radio

http://www.qsl.net/va3iul

---

52 - Center Fed-Zepp Antenna for 80m - 40m

http://www.qsl.net/va3iul

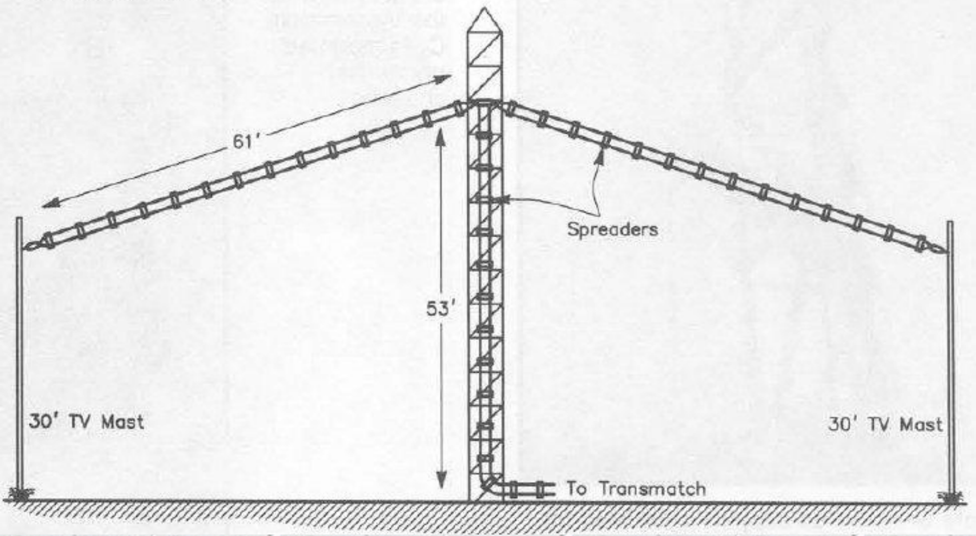
61'
Spreaders
53'
30' TV Mast
30' TV Mast
To Transmatch

---

53 - All-Bands Antenna

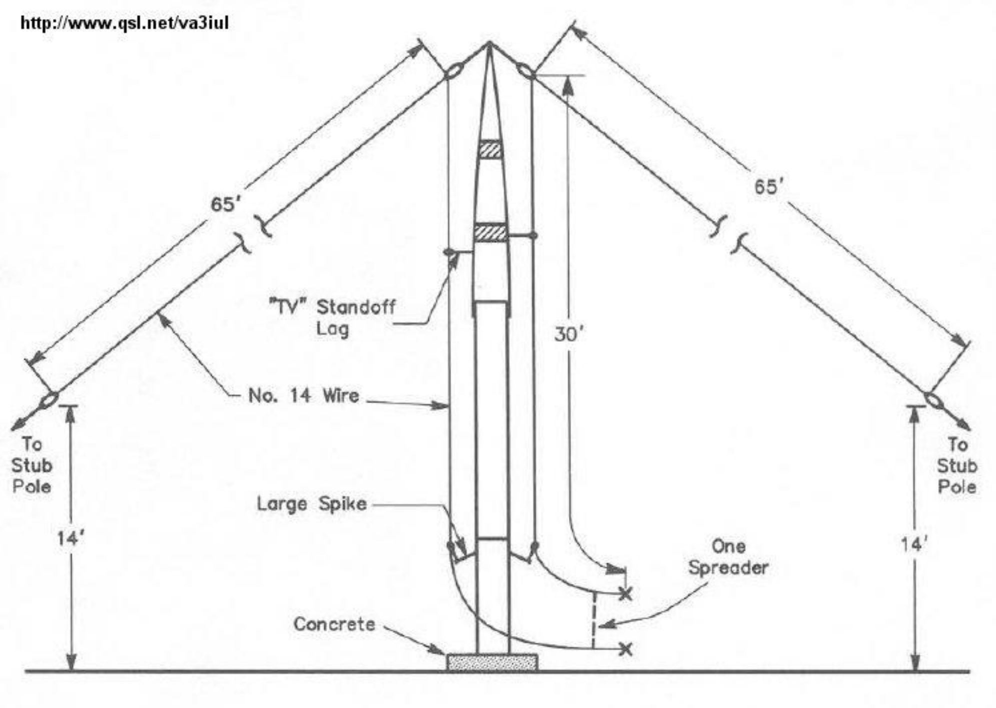
http://www.qsl.net/va3iul
65'
65'
"TV" Standoff Lag
30'
No. 14 Wire
To Stub Pole
To Stub Pole
14'
14'
Large Spike
One Spreader
Concrete

---

54 - All-Bands Dipole Antenna

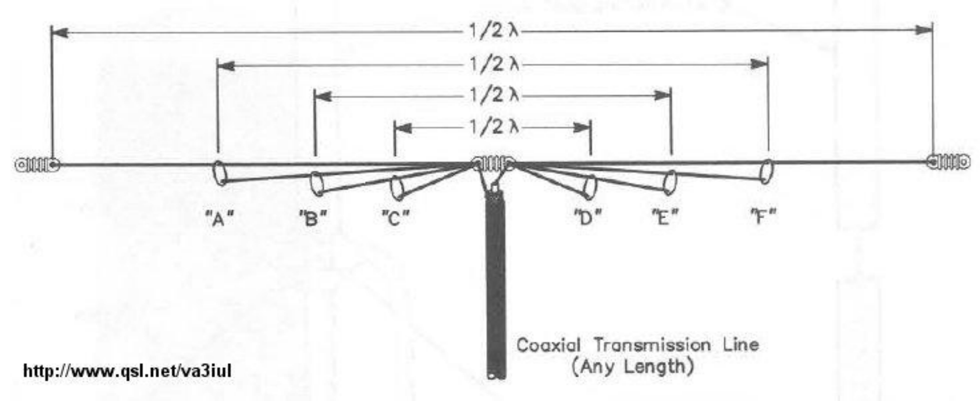
1/2 λ
1/2 λ
1/2 λ
1/2 λ
"A"
"B"
"C"
"D"
"E"
"F"
Coaxial Transmission Line (Any Length)
http://www.qsl.net/va3iul

---

55 - Multiband Z Antenna

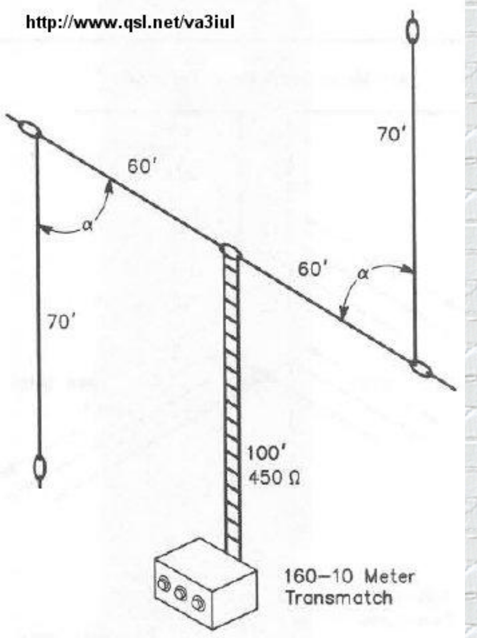
http://www.qsl.net/va3iul
70'
60'
α
60'
α
70'
100'
450 Ω
160–10 Meter
Transmatch

---

56 - Multiband Dipole Antenna

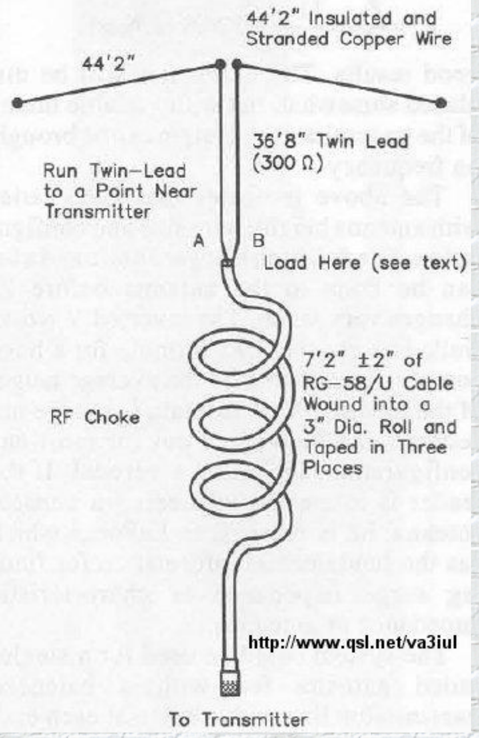
44'2" Insulated and
Stranded Copper Wire
44'2"
36'8" Twin Lead
(300 Ω)
Run Twin–Lead
to a Point Near
Transmitter
A
B
Load Here (see text)
7'2" ±2" of
RG–58/U Cable
Wound into a
3" Dia. Roll and
Taped in Three
Places
RF Choke
http://www.qsl.net/va3iul
To Transmitter

---

57 - Five-Bands No-Tuner Antenna

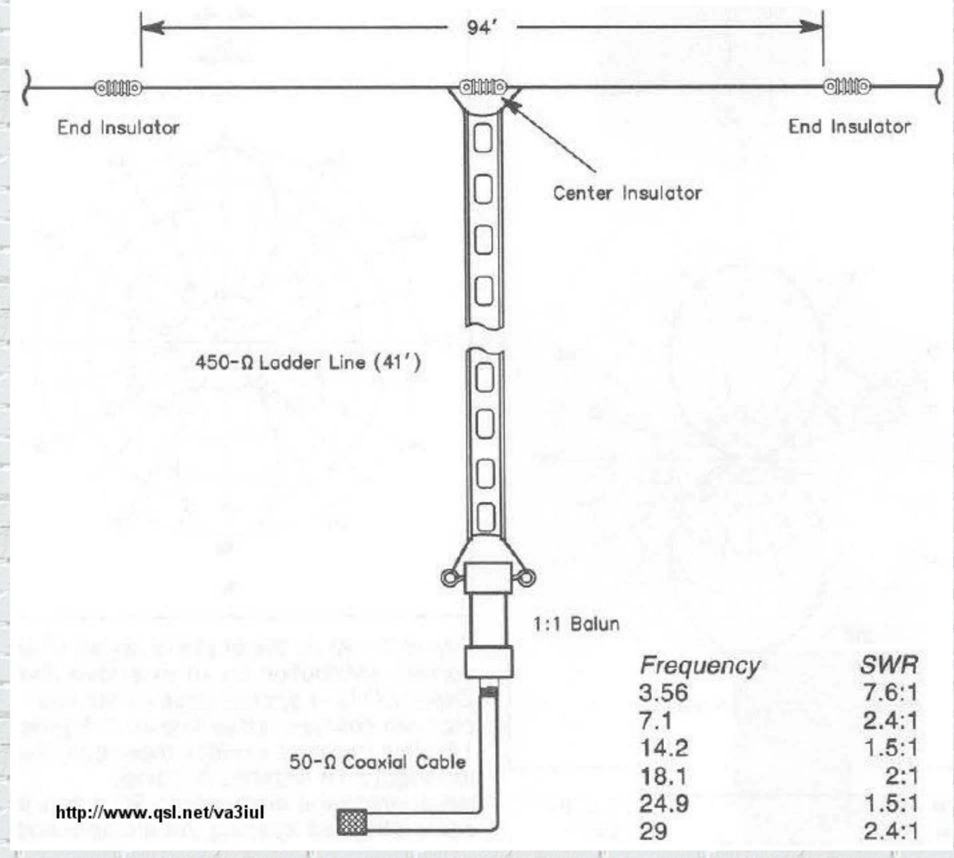
94′
End Insulator
End Insulator
Center Insulator
450-Ω Ladder Line (41′)
1:1 Balun
50-Ω Coaxial Cable
http://www.qsl.net/va3iul

Frequency	*SWR*
3.56	7.6:1
7.1	2.4:1
14.2	1.5:1
18.1	2:1
24.9	1.5:1
29	2.4:1

58 - Dualband Full-wave Loop Antenna for 80m-40m

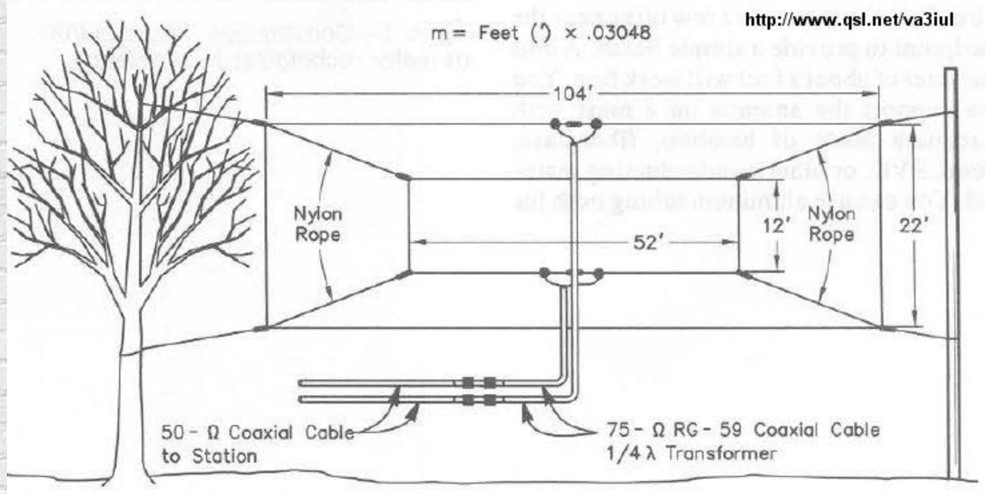
http://www.qsl.net/va3iul
m = Feet (′) × .03048
104′
Nylon Rope
52′
12′
Nylon Rope
22′
50 - Ω Coaxial Cable to Station
75 - Ω RG - 59 Coaxial Cable 1/4 λ Transformer

59 - Loop Antenna for 10m

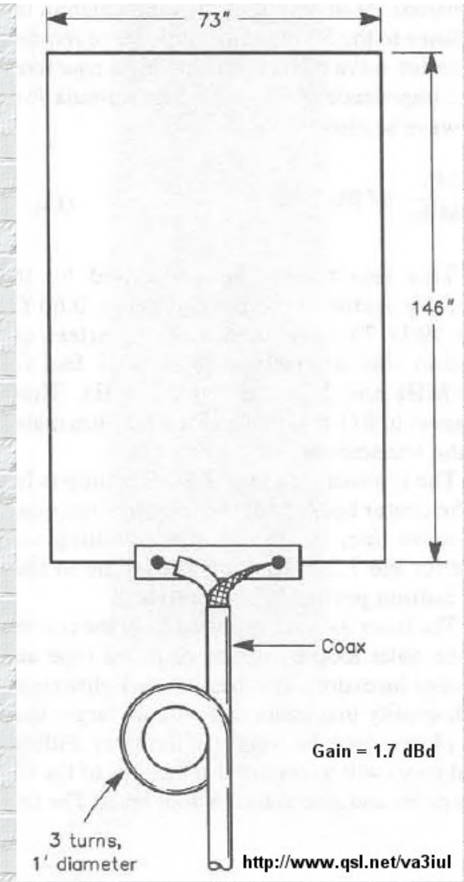
73"
146"
Coax
Gain = 1.7 dBd
3 turns,
1' diameter
http://www.qsl.net/va3iul

---

60 - Lazy Quad Antenna for 10m

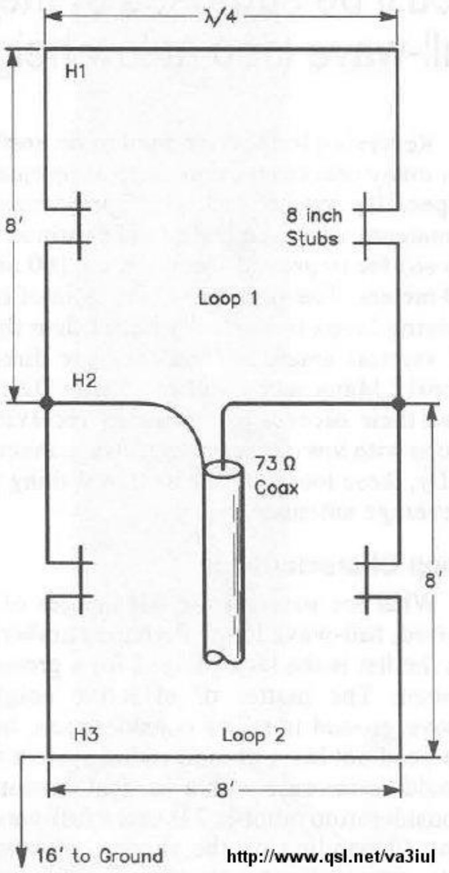
λ/4
H1
8'
8 inch
Stubs
Loop 1
H2
73 Ω
Coax
8'
H3
Loop 2
8'
16' to Ground
http://www.qsl.net/va3iul

61 - Tri-band Delta Loop Antenna for 80m - 40m - 30m

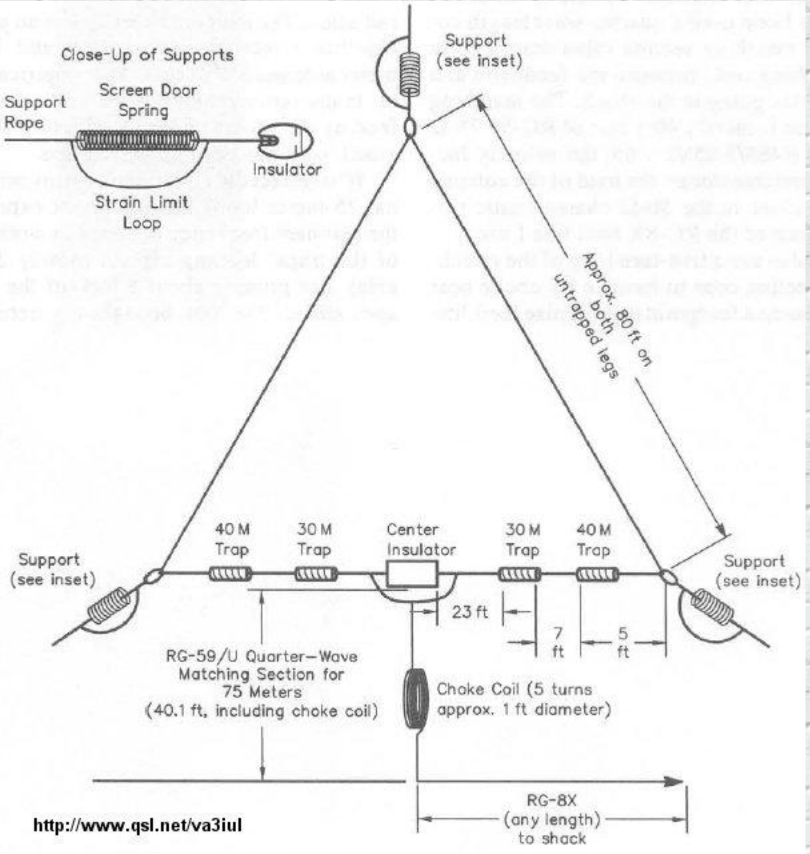
Close-Up of Supports
Support Rope
Screen Door Spring
Insulator
Strain Limit Loop
Support (see inset)
Approx. 80 ft on both untrapped legs
40 M Trap
30 M Trap
Center Insulator
30 M Trap
40 M Trap
Support (see inset)
Support (see inset)
23 ft
7 ft
5 ft
RG-59/U Quarter-Wave Matching Section for 75 Meters (40.1 ft, including choke coil)
Choke Coil (5 turns approx. 1 ft diameter)
RG-8X (any length) to shack
http://www.qsl.net/va3iul

---

62 - Dual-band Loop Antenna for 30m - 40m

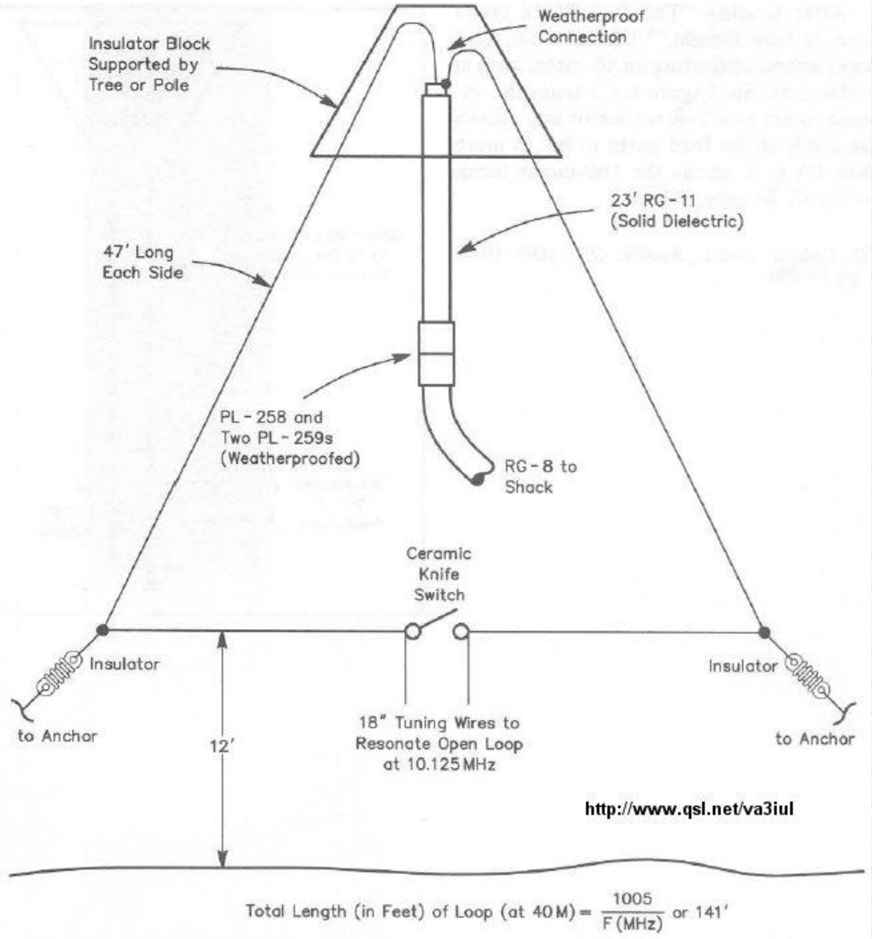
Insulator Block
Supported by
Tree or Pole
Weatherproof
Connection
23′ RG-11
(Solid Dielectric)
47′ Long
Each Side
PL-258 and
Two PL-259s
(Weatherproofed)
RG-8 to
Shack
Ceramic
Knife
Switch
Insulator
Insulator
to Anchor
to Anchor
12′
18″ Tuning Wires to
Resonate Open Loop
at 10.125 MHz
http://www.qsl.net/va3iul
Total Length (in Feet) of Loop (at 40 M) = 1005 / F (MHz) or 141′

63 - Wire-Beam Antenna for 80m

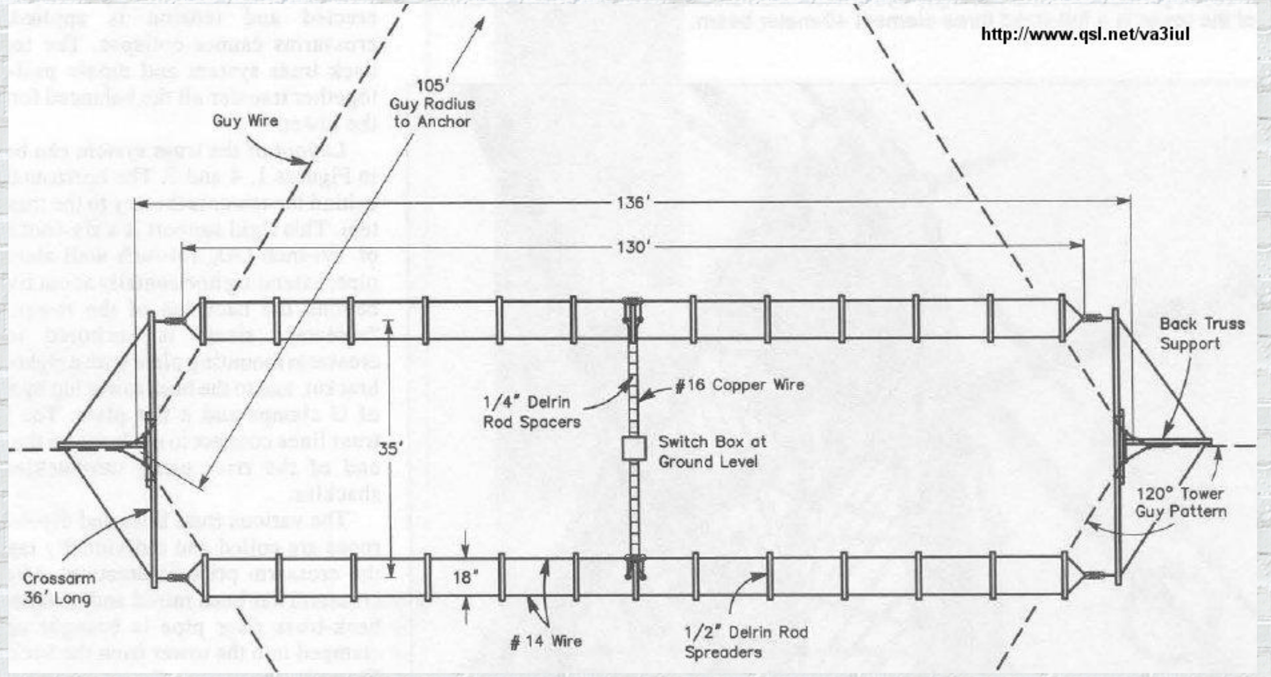
http://www.qsl.net/va3iul
105′
Guy Radius
to Anchor
Guy Wire
136′
130′
Back Truss
Support
#16 Copper Wire
1/4″ Delrin
Rod Spacers
35′
Switch Box at
Ground Level
120° Tower
Guy Pattern
Crossarm
36′ Long
18″
# 14 Wire
1/2″ Delrin Rod
Spreaders

64 - Dual-Band Sloper Antenna

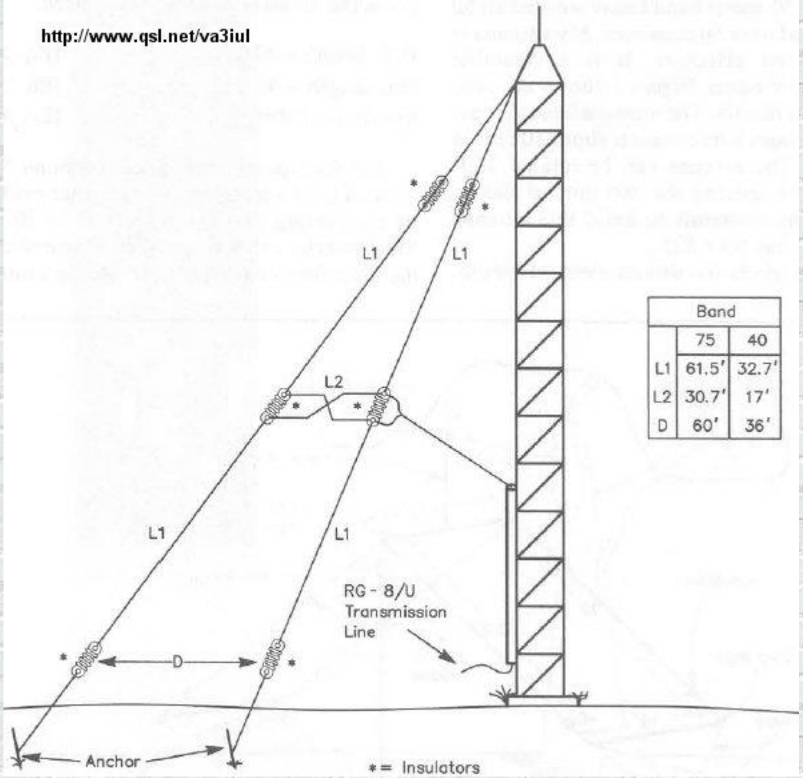
http://www.qsl.net/va3iul
L1
L1
L2
L1
L1
D
RG - 8/U Transmission Line
Anchor
* = Insulators

	Band	
	75	40
L1	61.5'	32.7'
L2	30.7'	17'
D	60'	36'

65 - Inverted-V Beam Antenna for 30m

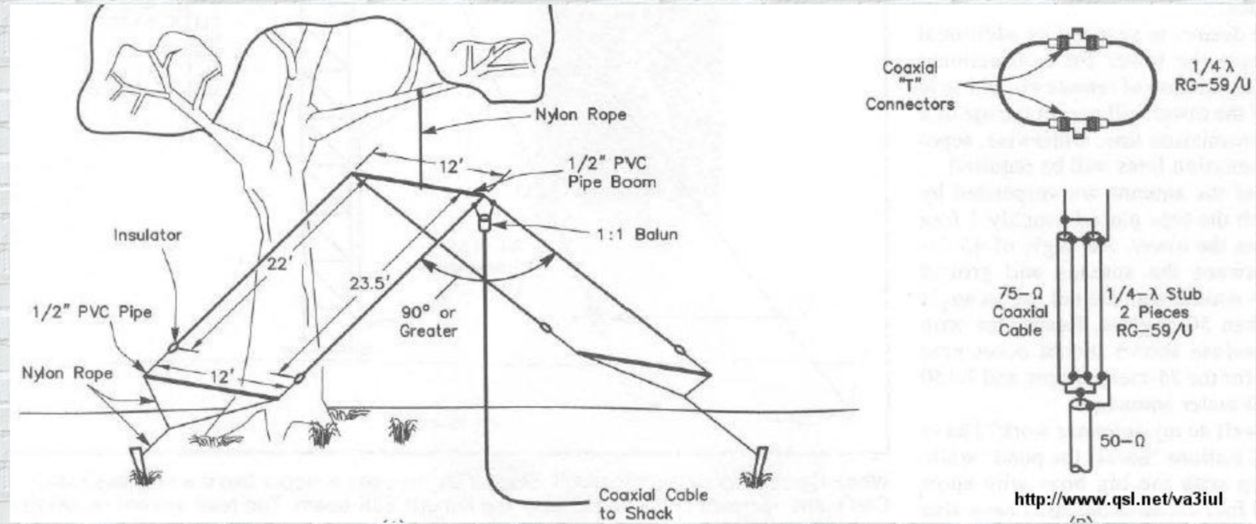
Nylon Rope
12'
1/2" PVC Pipe Boom
1:1 Balun
Insulator
22'
23.5'
1/2" PVC Pipe
90° or Greater
Nylon Rope
12'
Coaxial Cable to Shack
Coaxial "T" Connectors
1/4 λ RG-59/U
75-Ω Coaxial Cable
1/4-λ Stub 2 Pieces RG-59/U
50-Ω
http://www.qsl.net/va3iul

66 - ZL-Special Beam Antenna for 15m

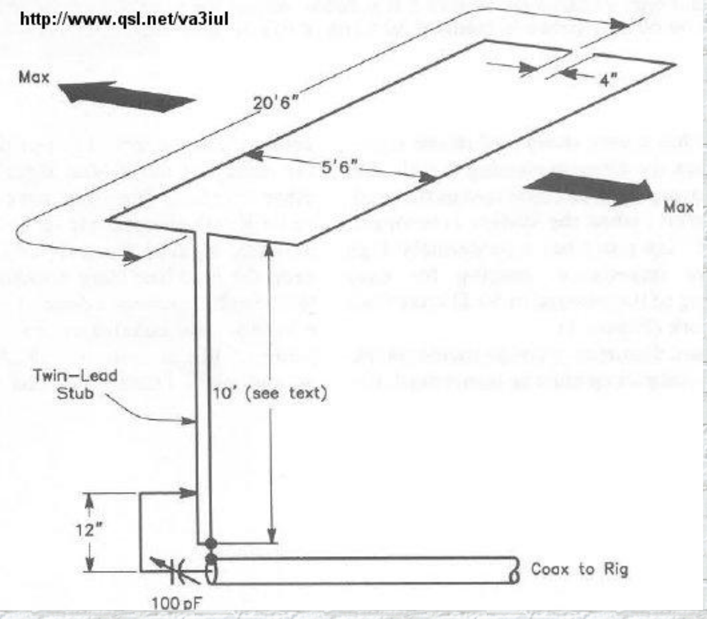
http://www.qsl.net/va3iul
Max
20'6"
4"
5'6"
Max
Twin–Lead Stub
10' (see text)
12"
Coax to Rig
100 pF

67 - Half-Sloper Antenna for 160m

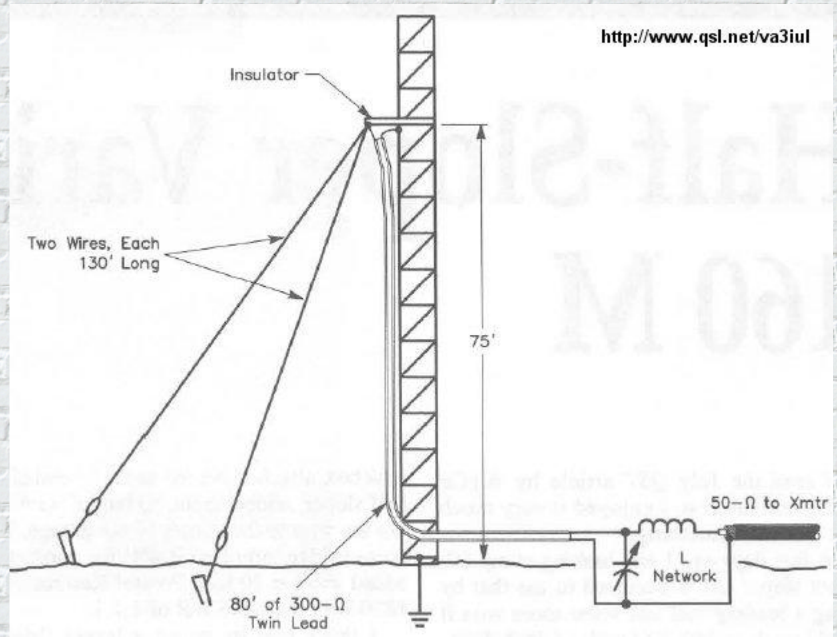
http://www.qsl.net/va3iul
Insulator
Two Wires, Each 130' Long
75'
50–Ω to Xmtr
L Network
80' of 300–Ω Twin Lead

68 - Two-Bands Half Sloper for 80m - 40m

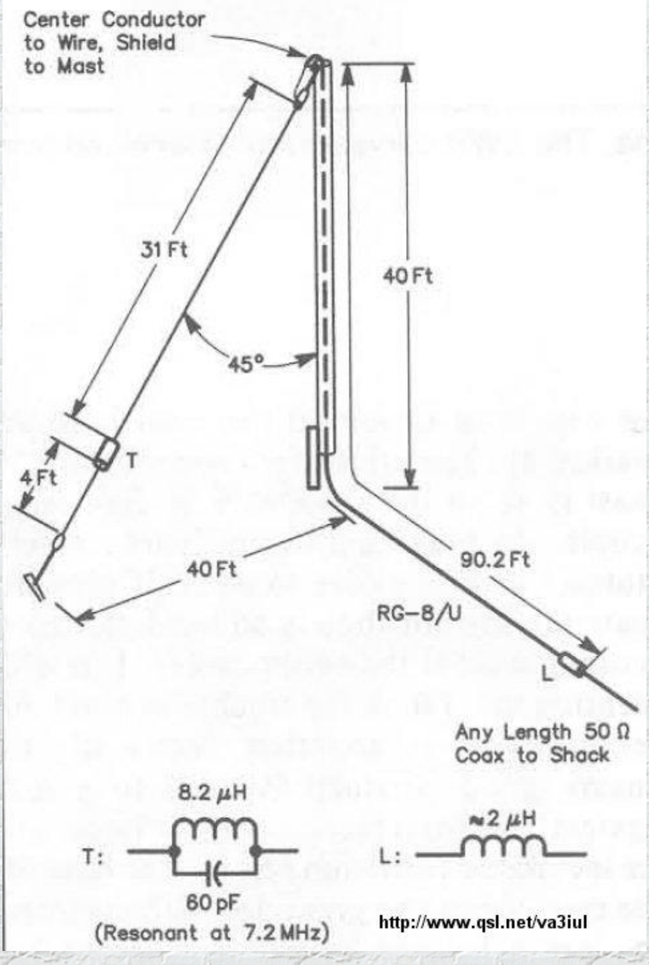
Center Conductor to Wire, Shield to Mast
31 Ft
40 Ft
45°
T
4 Ft
40 Ft
90.2 Ft
RG–8/U
L
Any Length 50 Ω Coax to Shack
8.2 µH
T:
60 pF
(Resonant at 7.2 MHz)
≈2 µH
L:
http://www.qsl.net/va3iul

69 - Linear Loaded Sloper Antenna for 160m

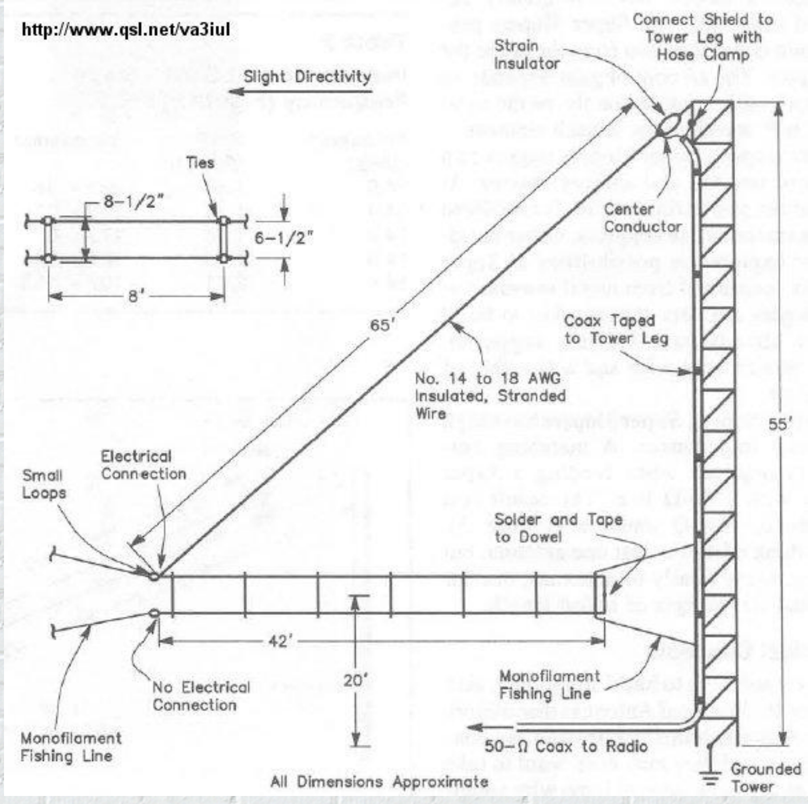
http://www.qsl.net/va3iul
Slight Directivity
Strain Insulator
Connect Shield to Tower Leg with Hose Clamp
Ties
8–1/2″
6–1/2″
8′
Center Conductor
65′
Coax Taped to Tower Leg
No. 14 to 18 AWG Insulated, Stranded Wire
55′
Small Loops
Electrical Connection
Solder and Tape to Dowel
42′
20′
No Electrical Connection
Monofilament Fishing Line
Monofilament Fishing Line
50–Ω Coax to Radio
Grounded Tower
All Dimensions Approximate

70 - Super-Sloper Antenna

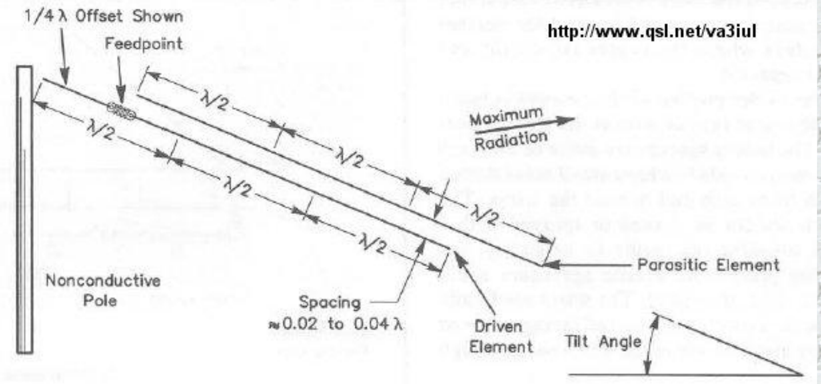
1/4 λ Offset Shown
Feedpoint
http://www.qsl.net/va3iul
λ/2
λ/2
λ/2
λ/2
λ/2
λ/2
Maximum Radiation
Parasitic Element
Nonconductive Pole
Spacing ≈0.02 to 0.04 λ
Driven Element
Tilt Angle

---

71 - Tower Pole as a Vertical Antenna for 80m

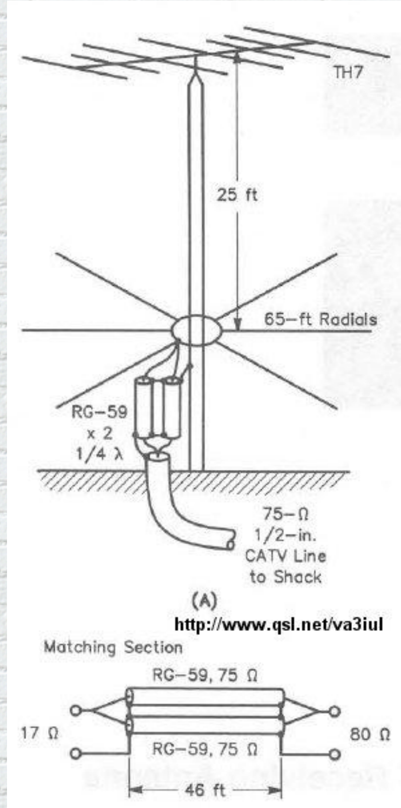
TH7
25 ft
65–ft Radials
RG–59 x 2 1/4 λ
75–Ω 1/2–in. CATV Line to Shack
(A)
http://www.qsl.net/va3iul
Matching Section
RG–59, 75 Ω
17 Ω
80 Ω
RG–59, 75 Ω
46 ft

---

72 - Clothesline Antenna

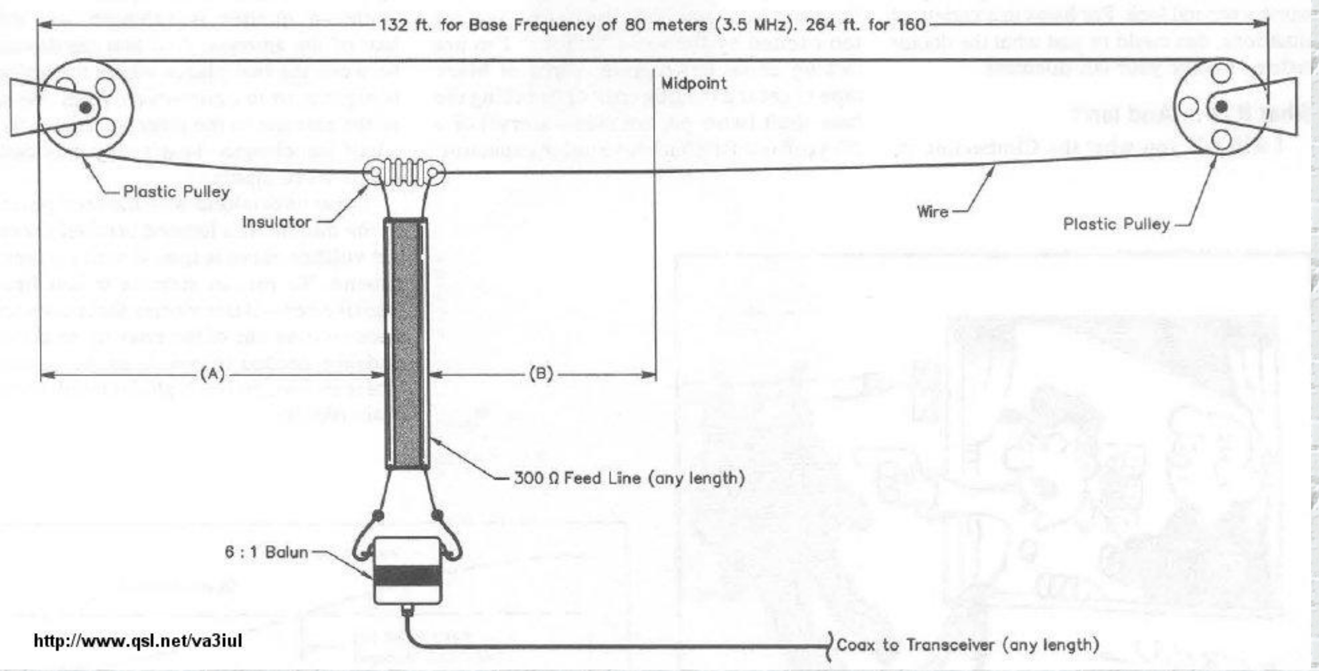
132 ft. for Base Frequency of 80 meters (3.5 MHz). 264 ft. for 160
Midpoint
Plastic Pulley
Insulator
Wire
Plastic Pulley
(A)
(B)
300 Ω Feed Line (any length)
6 : 1 Balun
Coax to Transceiver (any length)
http://www.qsl.net/va3iul

73 - Curtain Zepp Antenna for 160m, 80m, 40m

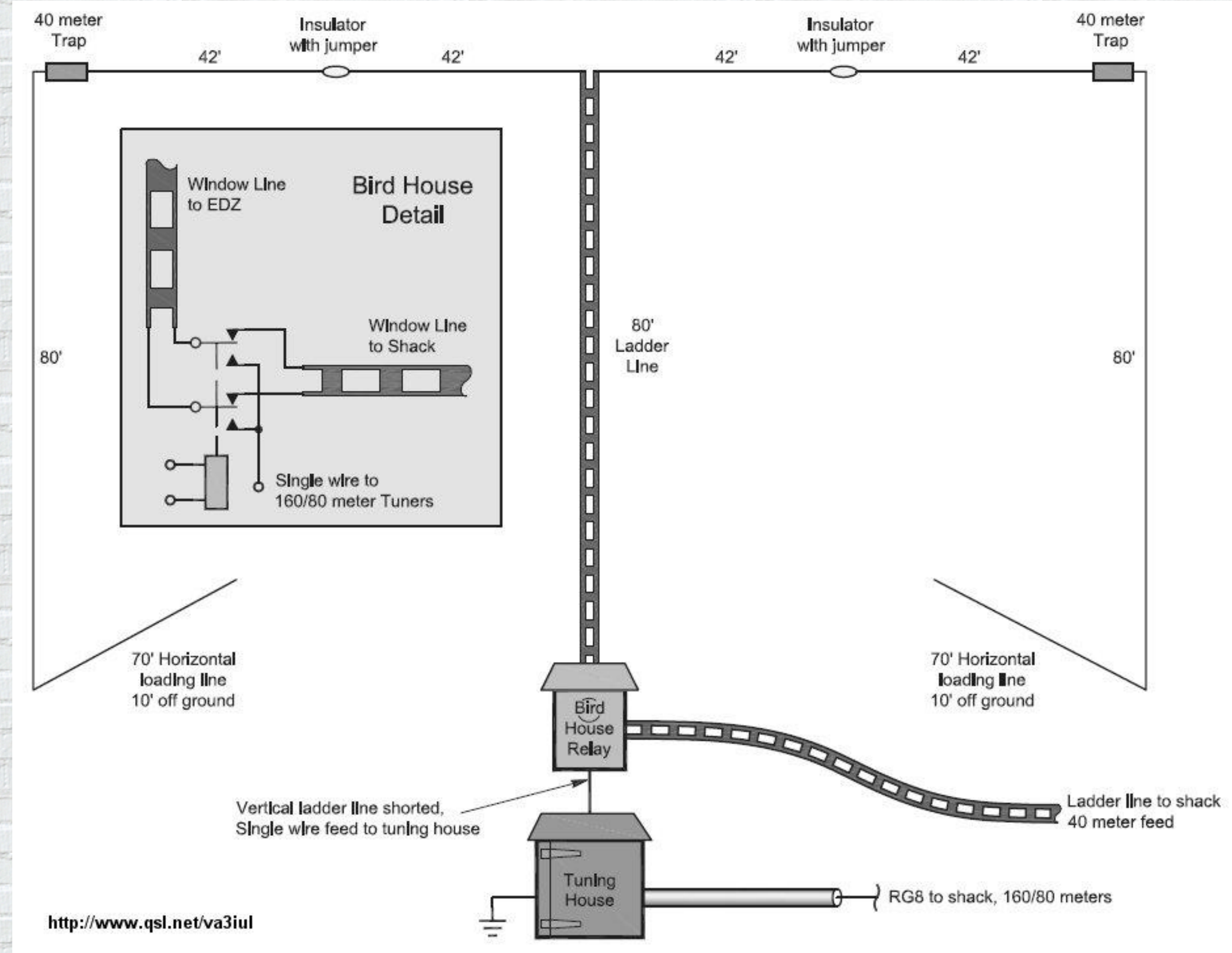
40 meter Trap
Insulator with jumper
42'
42'
42'
Insulator with jumper
42'
40 meter Trap
Window Line to EDZ
Bird House Detail
Window Line to Shack
Single wire to 160/80 meter Tuners
80'
80' Ladder Line
80'
70' Horizontal loading line 10' off ground
70' Horizontal loading line 10' off ground
Bird House Relay
Vertical ladder line shorted, Single wire feed to tuning house
Ladder line to shack 40 meter feed
Tuning House
RG8 to shack, 160/80 meters
http://www.qsl.net/va3iul

74 - Collinear Array Antenna for 40m, 30m, 20m

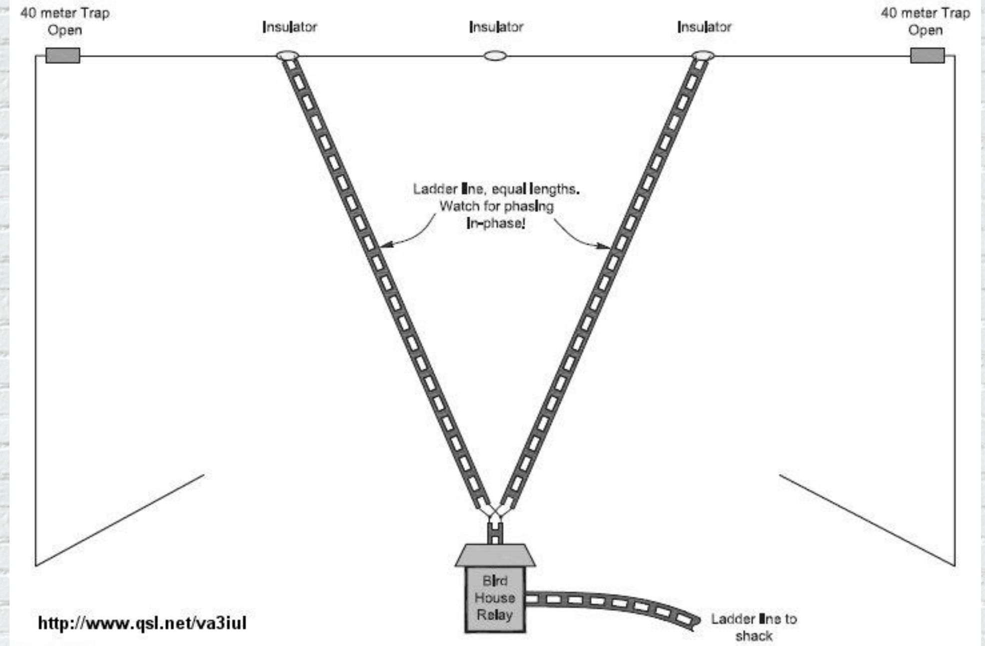
40 meter Trap
Open
Insulator
Insulator
Insulator
40 meter Trap
Open
Ladder line, equal lengths.
Watch for phasing
In-phase!
Bird
House
Relay
Ladder line to
shack
http://www.qsl.net/va3iul

75 - 160m Inverted Delta Loop

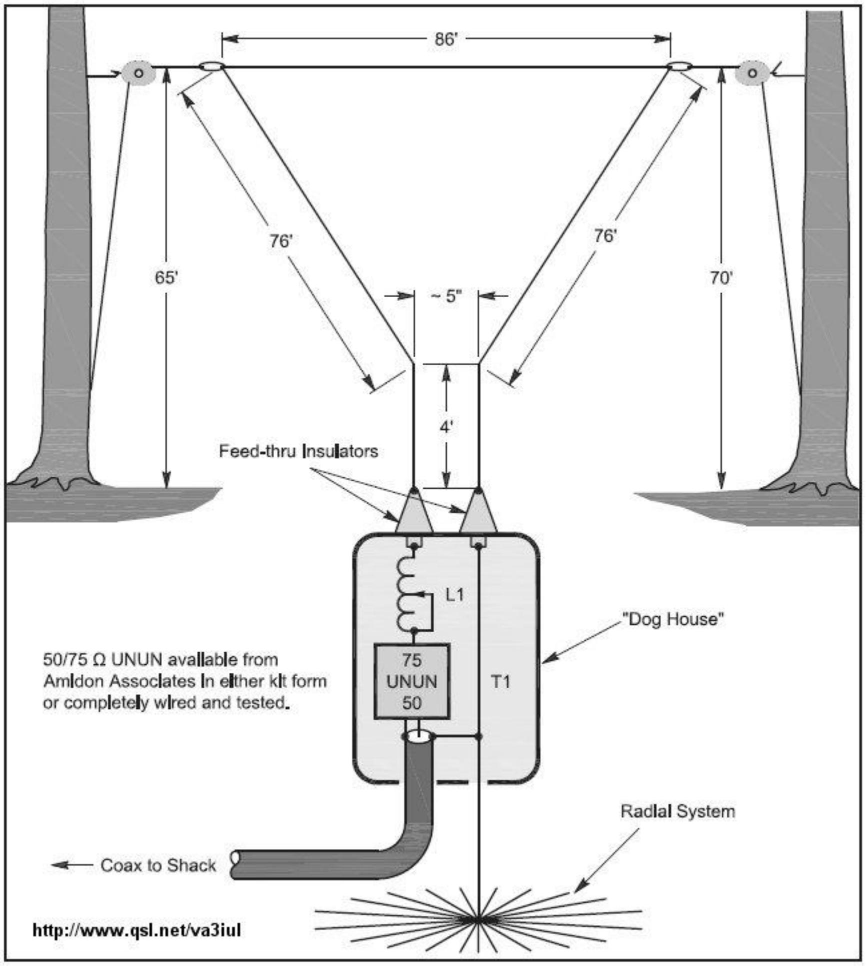
86'
76'
76'
65'
70'
~ 5"
4'
Feed-thru Insulators
L1
75
UNUN
50
T1
"Dog House"
50/75 Ω UNUN available from
Amidon Associates in either kit form
or completely wired and tested.
Radial System
Coax to Shack
http://www.qsl.net/va3iul

76 - Half Rhombic Unidirectional Vertical for 20m to 6m

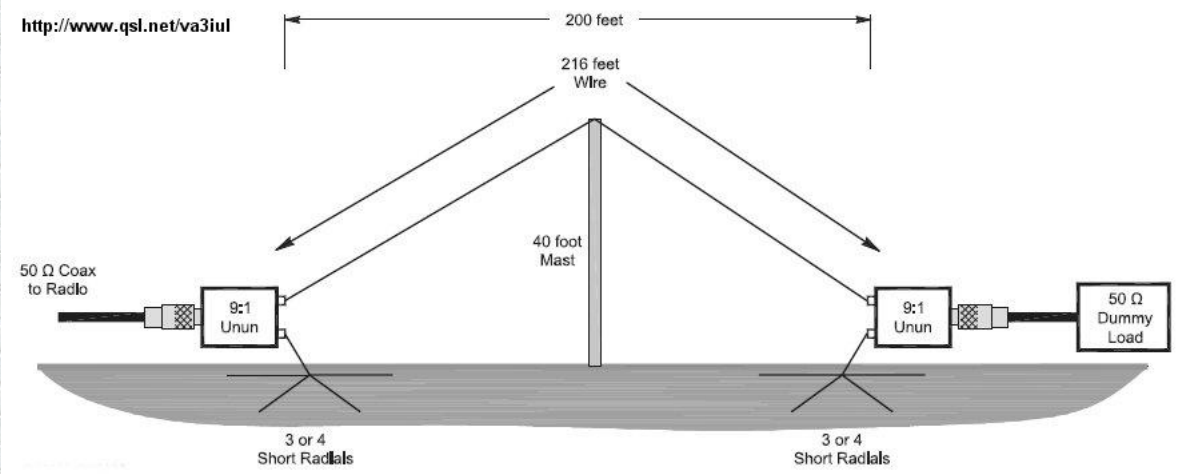
http://www.qsl.net/va3iul
200 feet
216 feet
Wire
40 foot
Mast
50 Ω Coax
to Radio
9:1
Unun
9:1
Unun
50 Ω
Dummy
Load
3 or 4
Short Radials
3 or 4
Short Radials

---

77 - Capacitance Loaded Vertical Antenna for 160m

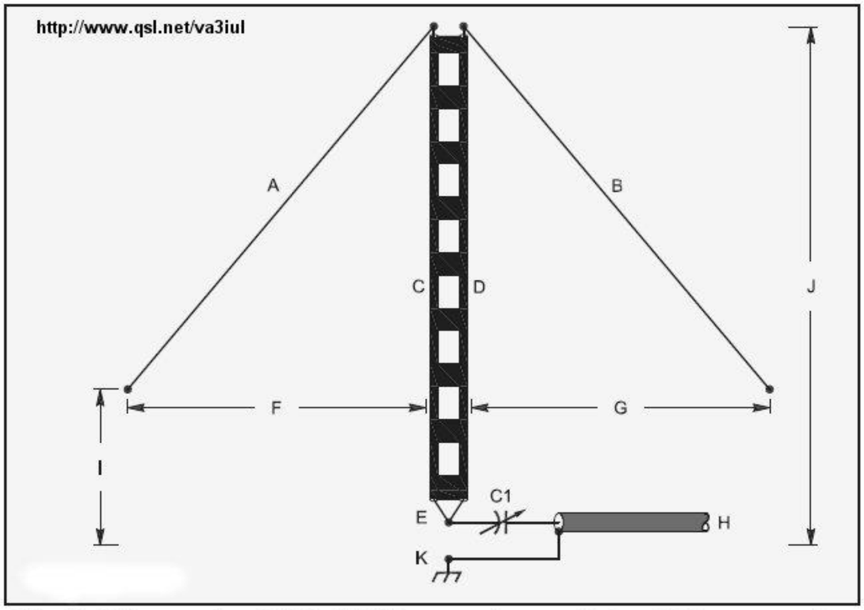
http://www.qsl.net/va3iul
A
B
C
D
J
F
G
I
C1
E
H
K

C1 — Variable capacitor, 0-1500 pF, 7 kV.

A, B — Sloping segments of 66 foot, 4 inch pieces of #16 AWG enameled copper wire soldered to the top of segments C and D and forming a 45° angle.

C, D — 59 feet of 450 Ω open wire or window line or, alternatively, two parallel lengths of #16 AWG enameled copper wire with a 1 inch separation.

E — 1 inch triangle formed by the bottom ends of the open wire line. Solder them tightly at the vertex and connect to C1.

F, G — Distance between the lower ends of segments A and B from segments C and D, 51 feet.

H — Any length of 50 Ω coaxial cable (RG-213 in this case). The central conductor must be soldered to C1, and the braid to the ground radials.

I — Approximately 16 feet from ground to lower ends of sloping segments A and B.

J — 60 feet from antenna top to bottom.

K — Ground radials, 60 ¼ λ (the more the better — 32 will form a reasonable ground).

---

78 - Fan Dipole Antenna for 80m to 6m

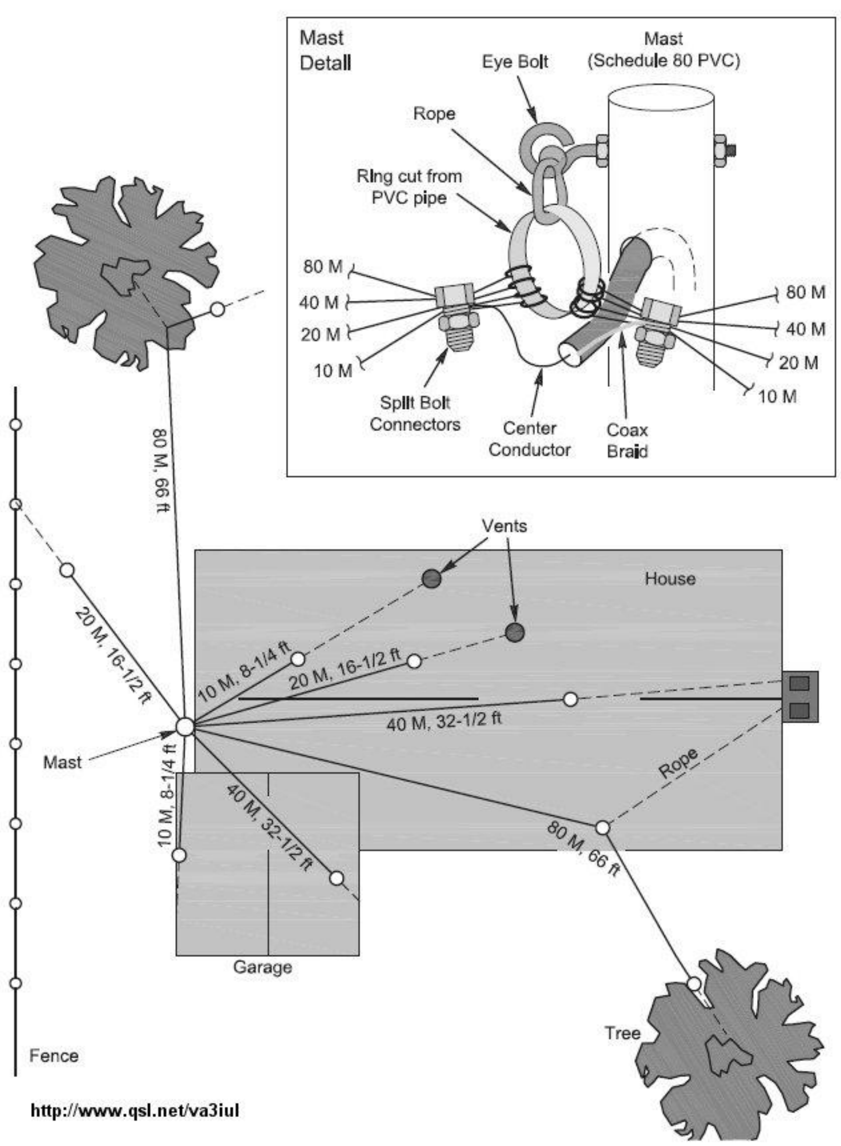
Mast Detail
Eye Bolt
Mast (Schedule 80 PVC)
Rope
Ring cut from PVC pipe
80 M
40 M
20 M
10 M
Split Bolt Connectors
Center Conductor
Coax Braid
80 M
40 M
20 M
10 M
80 M, 66 ft
Vents
House
20 M, 16-1/2 ft
10 M, 8-1/4 ft
20 M, 16-1/2 ft
40 M, 32-1/2 ft
Rope
Mast
10 M, 8-1/4 ft
40 M, 32-1/2 ft
80 M, 66 ft
Garage
Tree
Fence

http://www.qsl.net/va3iul

79 - Wire Ground Plane Antenna

The lengths of all 3 sides is determined by the formula of a 1/4 wave antenna

This antenna configuration will work for all frequencies from 160 meters to VHF.

$$\text{Length(in feet)} = \frac{234}{\text{frequency in MHz}}$$

The vertical side uses the inner conductor of the coax. On the Budwig the is the side the shows a "+" (plus sign). The outer (lower) legs use the outer braid or the "-" (minus sign).

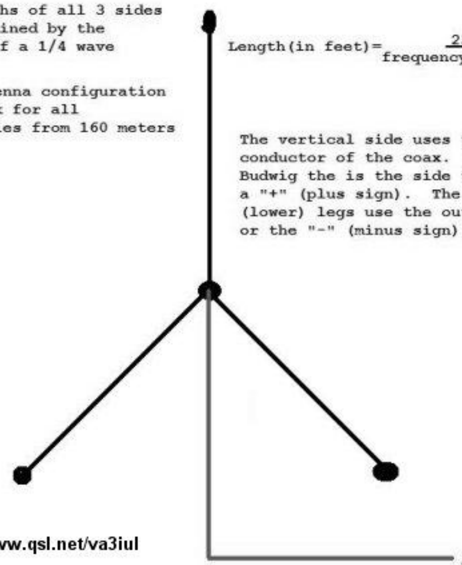
Coax to XCVR or XMTR

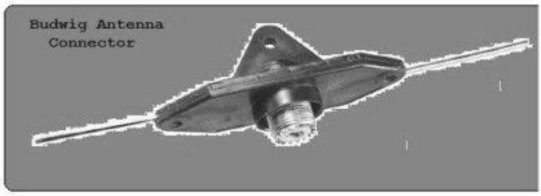
Budwig Antenna Connector

SO-239 Chassis mount, 4 hole

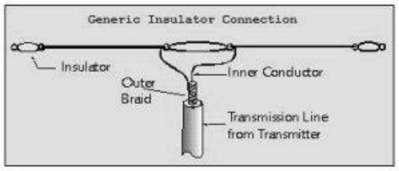
Generic Insulator Connection
Insulator
Outer Braid
Inner Conductor
Transmission Line from Transmitter

http://www.qsl.net/va3iul

80 - Inverted Delta Loop Antenna for 160m

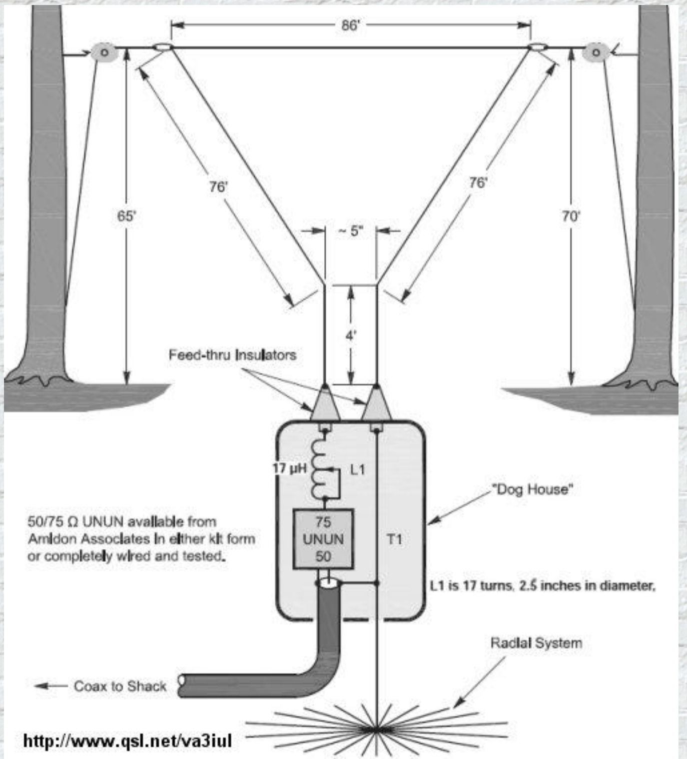
86'
76'
76'
65'
70'
~ 5"
4'
Feed-thru Insulators
17 µH
L1
"Dog House"
50/75 Ω UNUN available from Amidon Associates in either kit form or completely wired and tested.
75 UNUN 50
T1
L1 is 17 turns, 2.5 inches in diameter,
Radial System
Coax to Shack
http://www.qsl.net/va3iul

---

81 - Inverted-L for 160m

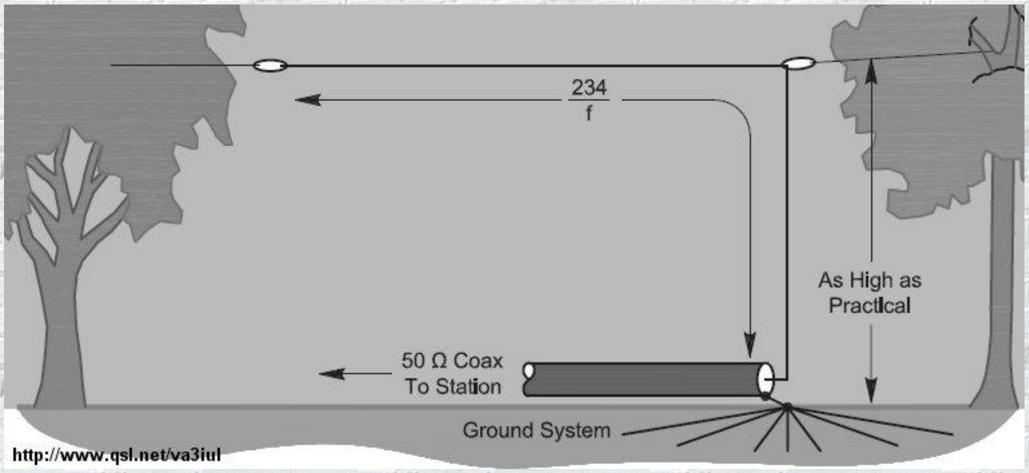
234 f
As High as Practical
50 Ω Coax To Station
Ground System
http://www.qsl.net/va3iul

82 - 300ohm-Ribbon Dual Band Dipole

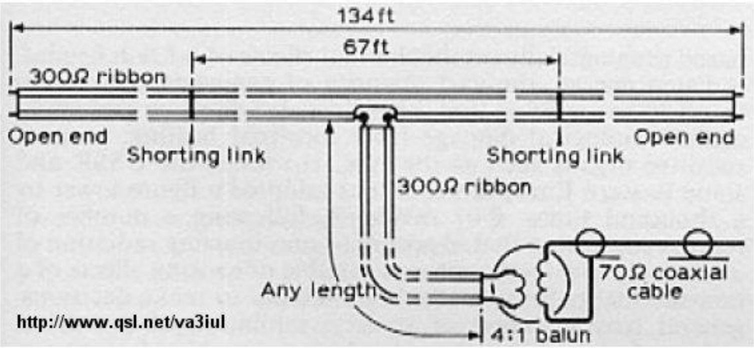
134 ft
67 ft
300Ω ribbon
Open end
Shorting link
Shorting link
Open end
300Ω ribbon
Any length
70Ω coaxial cable
4:1 balun
http://www.qsl.net/va3iul

83 - Tri-Band Beam for 20m, 15m, 10m

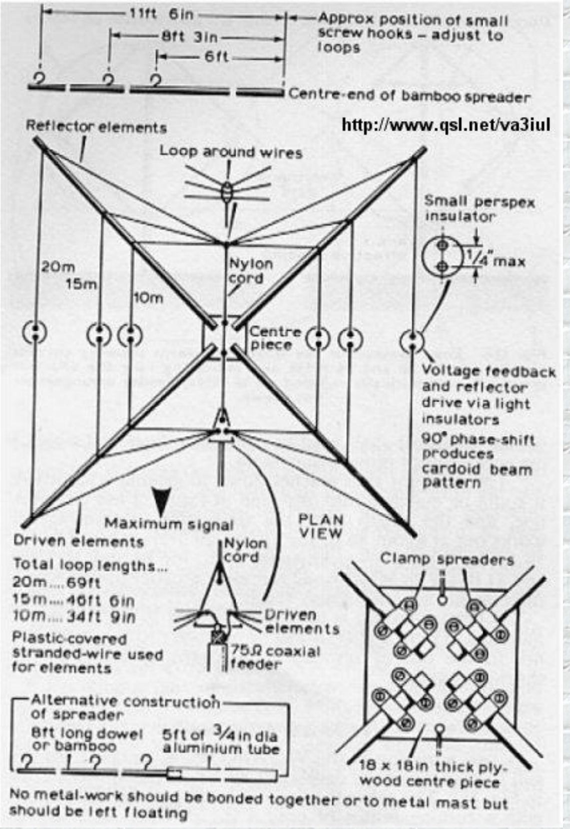
11ft 6in
8ft 3in
6ft
Approx position of small screw hooks – adjust to loops
Centre-end of bamboo spreader
Reflector elements
http://www.qsl.net/va3iul
Loop around wires
Small perspex insulator
1/4" max
20m
15m
10m
Nylon cord
Centre piece
Voltage feedback and reflector drive via light insulators
90° phase-shift produces cardoid beam pattern
Maximum signal
PLAN VIEW
Driven elements
Nylon cord
Total loop lengths...
20m.... 69ft
15m.... 46ft 6in
10m.... 34ft 9in
Driven elements
Plastic-covered stranded-wire used for elements
75Ω coaxial feeder
Clamp spreaders
Alternative construction of spreader
8ft long dowel or bamboo
5ft of 3/4 in dia aluminium tube
18 x 18in thick ply-wood centre piece
No metal-work should be bonded together or to metal mast but should be left floating

---

84 - Mini-Horse Yagi Antenna

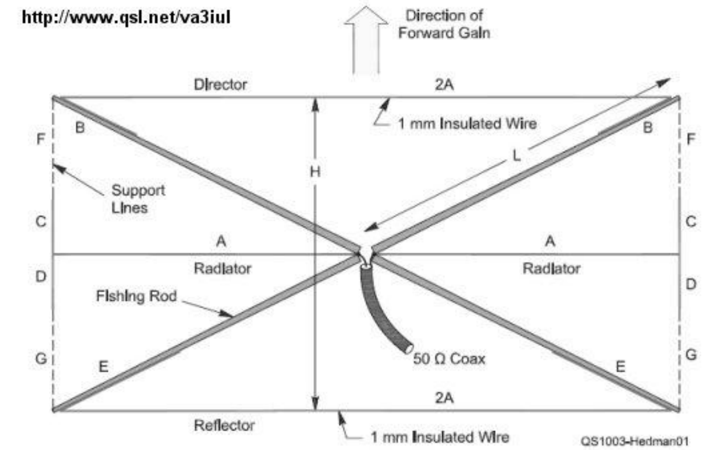
http://www.qsl.net/va3iul
Direction of Forward Gain
Director
2A
1 mm Insulated Wire
B
F
L
B
F
H
Support Lines
C
C
A
A
Radiator
Radiator
D
D
Fishing Rod
G
E
50 Ω Coax
E
G
2A
Reflector
1 mm Insulated Wire
QS1003-Hedman01

MHz	A	B	C	D	E	F	G	H	L
144	1' 2.2"	0' 4.7"	0' 3.5"	0' 1.8"	0' 7.1"	0' 4.7"	0' 6.5"	1' 4.5"	1' 4.3"
28	5' 11.6"	1' 10.8"	1' 6.9"	0' 9.4"	2' 11.8"	1' 10.8"	2' 8.3"	6' 11.4"	6' 11.0"
21	7' 11.6"	2' 6.3"	2' 1.2"	1' 0.6"	3' 11.6"	2' 6.3"	3' 6.9"	9' 3.0"	9' 2.6"
14	11' 11.3"	3' 9.7"	3' 1.8"	1' 6.9"	5' 11.2"	3' 9.7"	5' 4.6"	13' 10.9"	13' 9.7"
7.1	23' 10.5"	7' 7.3"	6' 3.6"	3' 1.8"	11' 10.5"	7' 7.3"	10' 9.1"	27' 9.8"	27' 7.4"

---

85 - Half-Rhombic Directive Antenna for 20m to 6m

http://www.qsl.net/va3iul

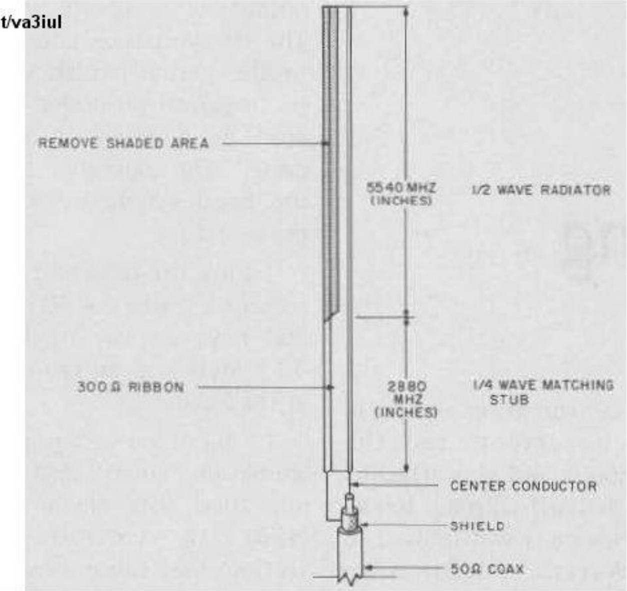
REMOVE SHADED AREA
5540 MHZ (INCHES)
1/2 WAVE RADIATOR
300 Ω RIBBON
2880 MHZ (INCHES)
1/4 WAVE MATCHING STUB
CENTER CONDUCTOR
SHIELD
50Ω COAX

Frequency (MHz)	28.0	28.5	29.0	50.0	51.0	52.0	144	145	146	147
Radiator (inches) ½ wave	197.8	194.3	191.0	110.8	108.6	106.5	38.4	38.2	37.9	37.7
Stub (inches) ¼ wave	102.8	101.0	99.3	57.6	56.4	55.3	20.0	19.8	19.7	19.6

86 - Fan-Dipole Antenna for 80m, 40m, 20m

http://www.qsl.net/va3iul

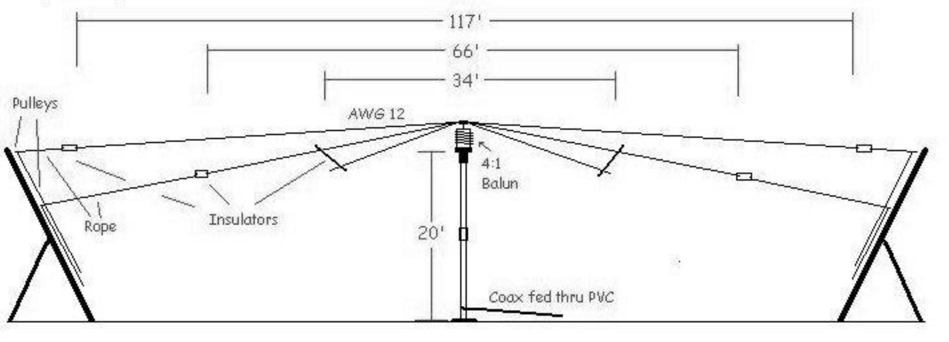
117'
66'
34'
Pulleys
AWG 12
4:1
Balun
Rope
Insulators
20'
Coax fed thru PVC

87 - Capacity Tuned Folded Loop Antenna for 20m

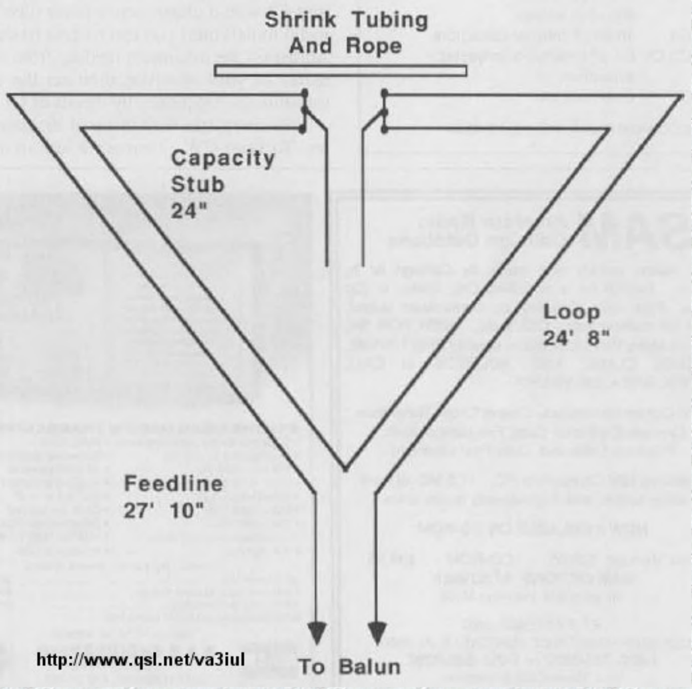
Shrink Tubing
And Rope
Capacity
Stub
24"
Loop
24' 8"
Feedline
27' 10"
To Balun
http://www.qsl.net/va3iul

88 - Indoor Loop Antenna for 80m to 30m

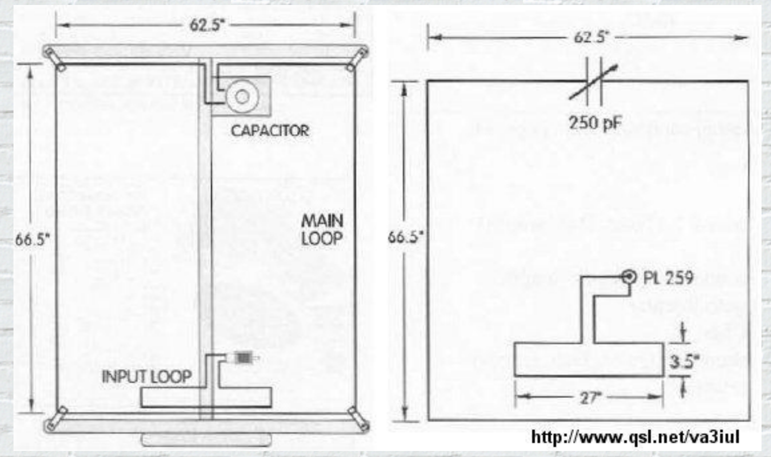
62.5"
CAPACITOR
MAIN
LOOP
66.5"
INPUT LOOP
62.5"
250 pF
66.5"
PL 259
3.5"
27"
http://www.qsl.net/va3iul

89 - Indoor Loop Antenna for 80m

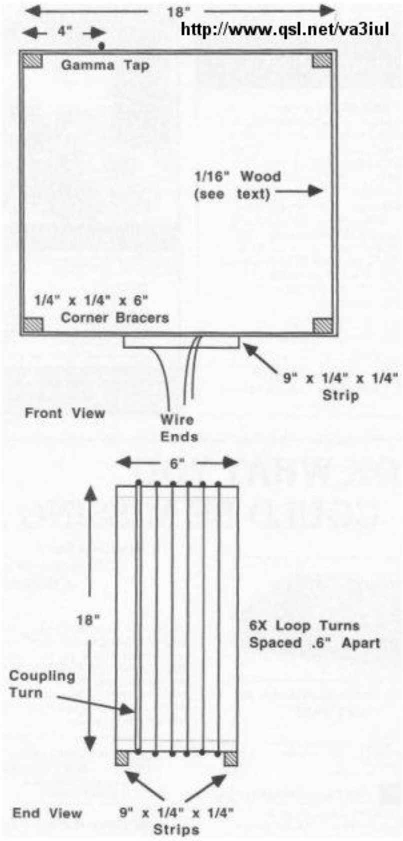
18"
4"
http://www.qsl.net/va3iul
Gamma Tap
1/16" Wood
(see text)
1/4" x 1/4" x 6"
Corner Bracers
9" x 1/4" x 1/4"
Strip
Front View
Wire
Ends
6"
18"
6X Loop Turns
Spaced .6" Apart
Coupling
Turn
End View
9" x 1/4" x 1/4"
Strips

---

90 - Double-Delta Antenna 80m and 40m

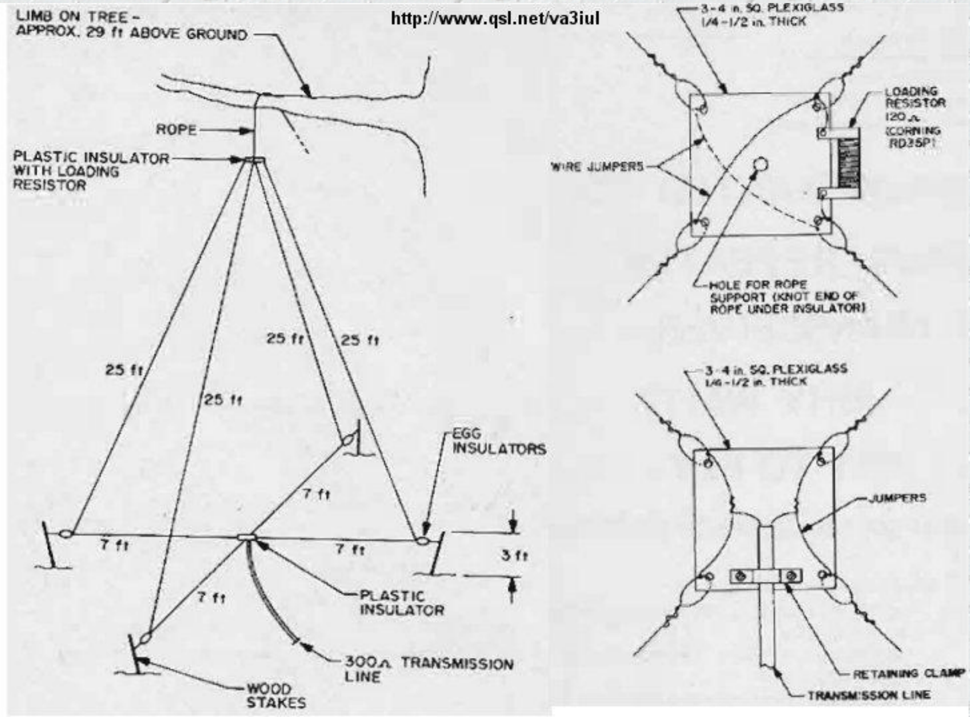
LIMB ON TREE -
APPROX. 29 ft ABOVE GROUND
http://www.qsl.net/va3iul
ROPE
PLASTIC INSULATOR
WITH LOADING
RESISTOR
25 ft
25 ft
25 ft
25 ft
EGG
INSULATORS
7 ft
7 ft
7 ft
7 ft
3 ft
PLASTIC
INSULATOR
300Ω TRANSMISSION
LINE
WOOD
STAKES
3-4 in. SQ. PLEXIGLASS
1/4-1/2 in. THICK
LOADING
RESISTOR
120 Ω
(CORNING
RD35P)
WIRE JUMPERS
HOLE FOR ROPE
SUPPORT (KNOT END OF
ROPE UNDER INSULATOR)
3-4 in. SQ. PLEXIGLASS
1/4-1/2 in. THICK
JUMPERS
RETAINING CLAMP
TRANSMISSION LINE

---

91 - Inductance-Loaded Shortened Dipole for 160m

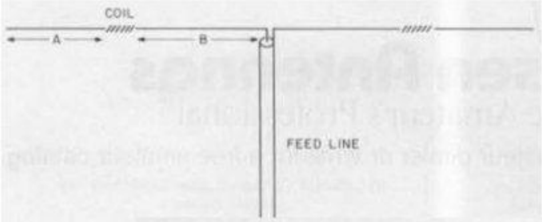
COIL
A
B
FEED LINE

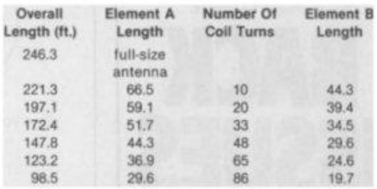

Overall Length (ft.)	Element A Length	Number Of Coil Turns	Element B Length
246.3	full-size antenna		
221.3	66.5	10	44.3
197.1	59.1	20	39.4
172.4	51.7	33	34.5
147.8	44.3	48	29.6
123.2	36.9	65	24.6
98.5	29.6	86	19.7

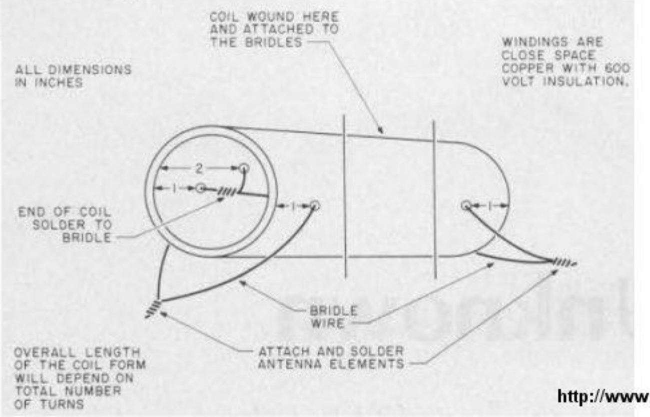
COIL WOUND HERE AND ATTACHED TO THE BRIDLES
ALL DIMENSIONS IN INCHES
WINDINGS ARE CLOSE SPACE COPPER WITH 600 VOLT INSULATION.
2
1
1
1
END OF COIL SOLDER TO BRIDLE
BRIDLE WIRE
OVERALL LENGTH OF THE COIL FORM WILL DEPEND ON TOTAL NUMBER OF TURNS
ATTACH AND SOLDER ANTENNA ELEMENTS
http://www.qsl.net/va3iul

---

92 - V-Beam Antenna for 15m

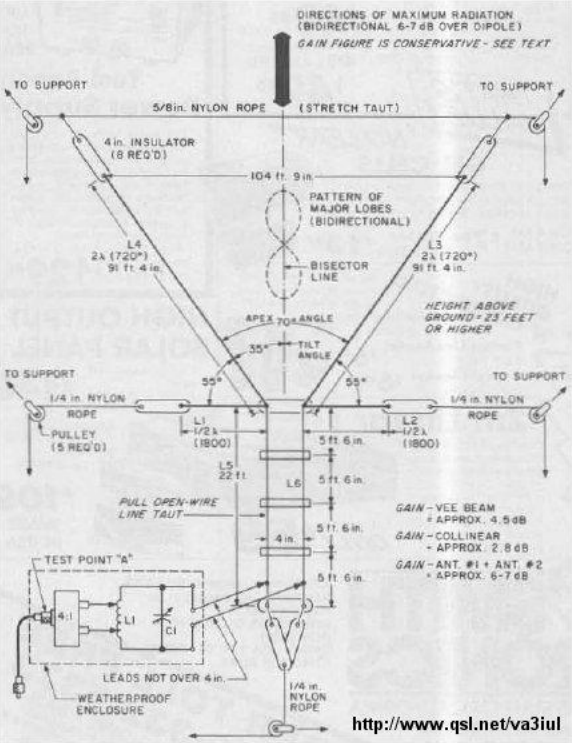
DIRECTIONS OF MAXIMUM RADIATION (BIDIRECTIONAL 6-7 dB OVER DIPOLE)
GAIN FIGURE IS CONSERVATIVE - SEE TEXT
TO SUPPORT
TO SUPPORT
5/8in. NYLON ROPE
(STRETCH TAUT)
4 in. INSULATOR (8 REQ'D)
104 ft. 9 in.
PATTERN OF MAJOR LOBES (BIDIRECTIONAL)
L4
2λ (720°)
91 ft. 4 in.
L3
2λ (720°)
91 ft. 4 in.
BISECTOR LINE
HEIGHT ABOVE GROUND = 23 FEET OR HIGHER
APEX 70° ANGLE
35°
TILT ANGLE
TO SUPPORT
TO SUPPORT
55°
55°
1/4 in. NYLON ROPE
1/4 in. NYLON ROPE
PULLEY (5 REQ'D)
L1
1/2 λ
(180°)
L2
1/2 λ
(180°)
L5
22 ft.
L6
5 ft. 6 in.
5 ft. 6 in.
5 ft. 6 in.
5 ft. 6 in.
4 in.
PULL OPEN-WIRE LINE TAUT
GAIN - VEE BEAM = APPROX. 4.5 dB
GAIN - COLLINEAR - APPROX. 2.8 dB
GAIN - ANT. #1 + ANT. #2 = APPROX. 6-7 dB
TEST POINT "A"
4:1
L1
C1
LEADS NOT OVER 4 in.
WEATHERPROOF ENCLOSURE
1/4 in. NYLON ROPE
http://www.qsl.net/va3iul

---

93 - Picnic Vertical Wire Antenna

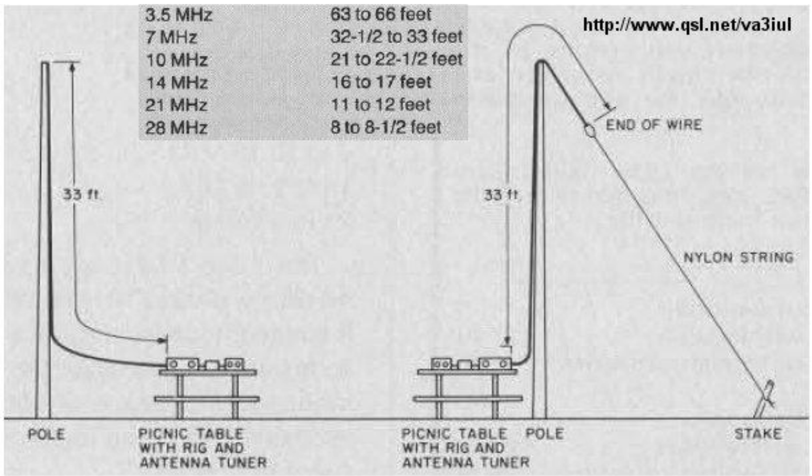
http://www.qsl.net/va3iul
33 ft.
33 ft.
END OF WIRE
NYLON STRING
POLE
PICNIC TABLE WITH RIG AND ANTENNA TUNER
PICNIC TABLE WITH RIG AND ANTENNA TUNER
POLE
STAKE

3.5 MHz	63 to 66 feet
7 MHz	32-1/2 to 33 feet
10 MHz	21 to 22-1/2 feet
14 MHz	16 to 17 feet
21 MHz	11 to 12 feet
28 MHz	8 to 8-1/2 feet

94 - Laid-Back Quad Antenna for 80m

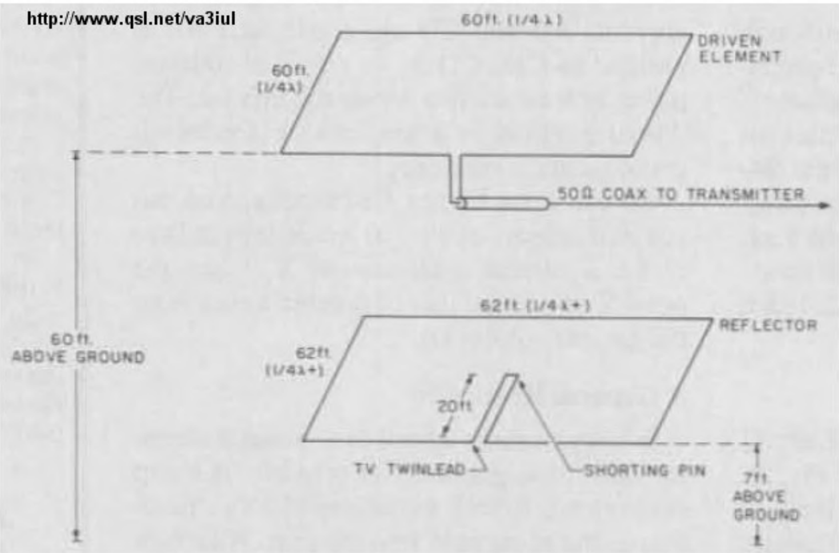
http://www.qsl.net/va3iul
60ft. (1/4λ)
60ft. (1/4λ)
DRIVEN ELEMENT
50Ω COAX TO TRANSMITTER
62ft. (1/4λ+)
REFLECTOR
60 ft. ABOVE GROUND
62ft. (1/4λ+)
20ft
TV TWINLEAD
SHORTING PIN
7ft. ABOVE GROUND

95 - Phased Loop Antenna

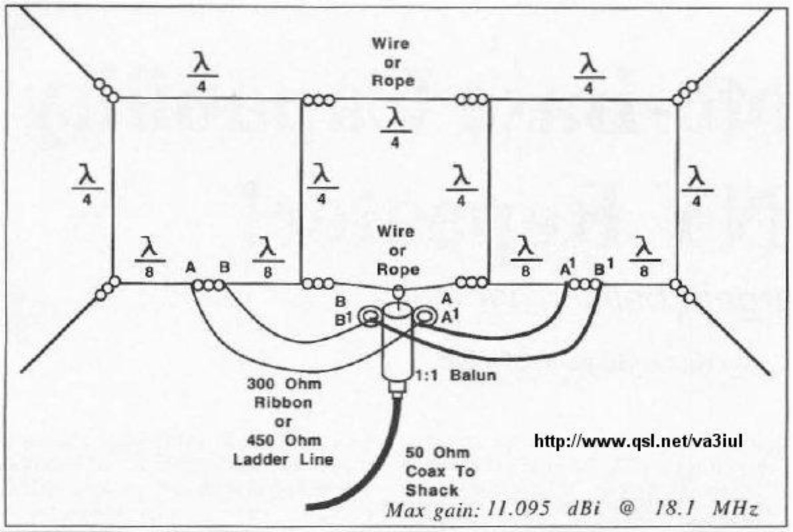
$\frac{\lambda}{4}$
Wire or Rope
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
Wire or Rope
$\frac{\lambda}{8}$
A
B
$\frac{\lambda}{8}$
$\frac{\lambda}{8}$
A¹
B¹
$\frac{\lambda}{8}$
B
B¹
A
A¹
1:1 Balun
300 Ohm Ribbon or 450 Ohm Ladder Line
50 Ohm Coax To Shack
http://www.qsl.net/va3iul
Max gain: 11.095 dBi @ 18.1 MHz

	λ/4	λ/8	300Ω Ribbon	450Ω Ladderline
40m	34.9 ft.	17.5 ft.	53.3 ft.	62.2 ft.
20m	17.6 ft.	8.8 ft.	27.0 ft.	31.2 ft.
17m	13.9 ft.	6.9 ft.	21.2 ft.	25.2 ft
15m	11.8 ft.	5.9 ft.	18.1 ft.	21.97 ft.
10m	8.67 ft.	4.33 ft.	13.2 ft.	15.33 ft.

96 - Loop Antenna for TX for 160m

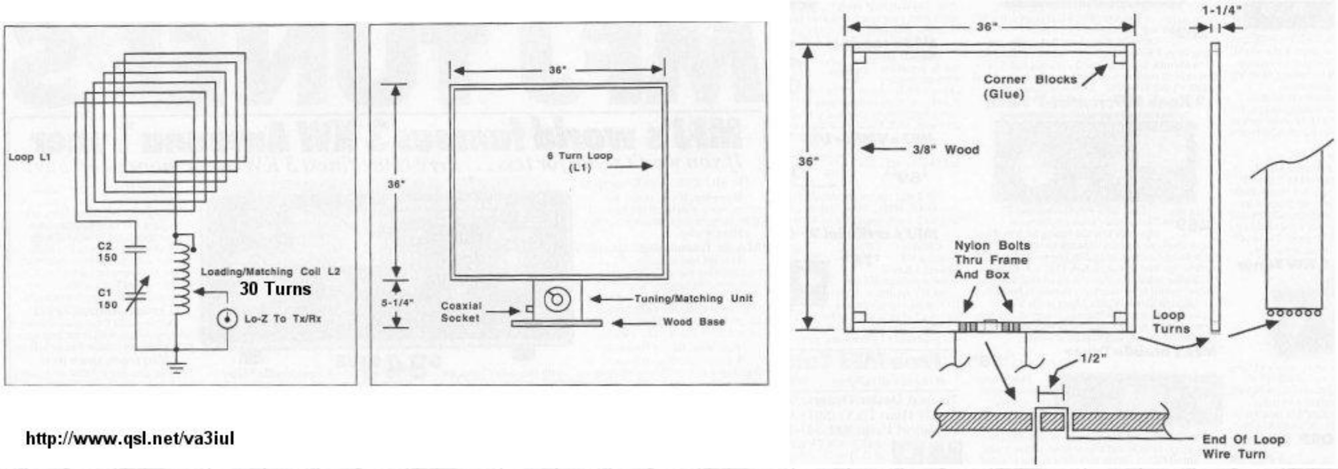
Loop L1
C2
150
C1
150
Loading/Matching Coil L2
30 Turns
Lo-Z To Tx/Rx
36"
36"
6 Turn Loop
(L1)
5-1/4"
Coaxial
Socket
Tuning/Matching Unit
Wood Base
36"
1-1/4"
Corner Blocks
(Glue)
36"
3/8" Wood
Nylon Bolts
Thru Frame
And Box
Loop
Turns
1/2"
End Of Loop
Wire Turn

http://www.qsl.net/va3iul

97 - Morgain-Dipole Antenna for 160m and 80m

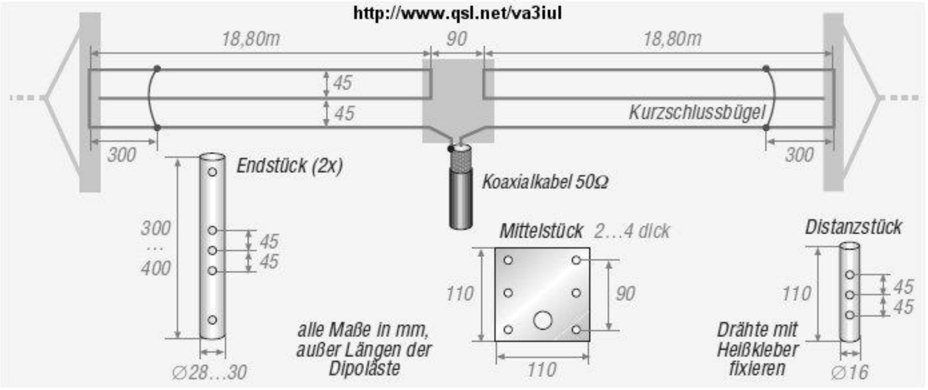
http://www.qsl.net/va3iul
18,80m
90
18,80m
45
45
Kurzschlussbügel
300
300
Endstück (2x)
Koaxialkabel 50Ω
300
...
400
45
45
Mittelstück 2…4 dick
110
90
110
Distanzstück
110
45
45
alle Maße in mm,
außer Längen der
Dipoläste
Drähte mit
Heißkleber
fixieren
∅28…30
∅16

98 - ZL-Special for 20m, 15m, 10m

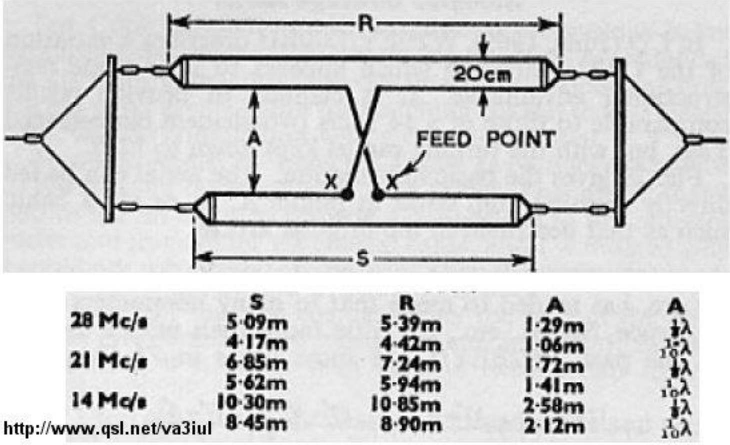
R
20cm
FEED POINT
A
X
X
S

	S	R	A	A
28 Mc/s	5·09m	5·39m	1·29m	$\frac{1}{8}\lambda$
	4·17m	4·42m	1·06m	$\frac{1}{10}\lambda$
21 Mc/s	6·85m	7·24m	1·72m	$\frac{1}{8}\lambda$
	5·62m	5·94m	1·41m	$\frac{1}{10}\lambda$
14 Mc/s	10·30m	10·85m	2·58m	$\frac{1}{8}\lambda$
	8·45m	8·90m	2·12m	$\frac{1}{10}\lambda$

http://www.qsl.net/va3iul

99 - Biconical Antenna

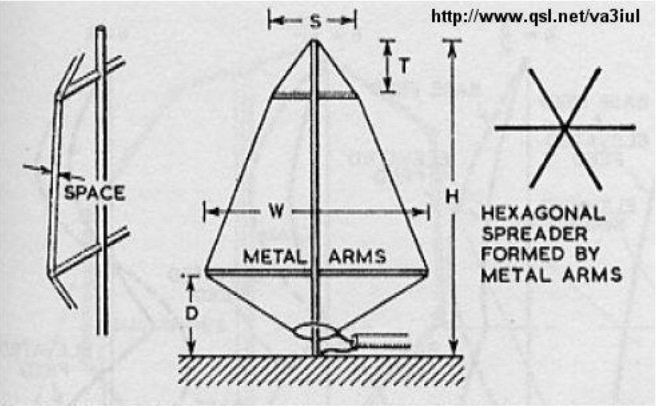
http://www.qsl.net/va3iul
S
T
H
SPACE
W
METAL ARMS
D
HEXAGONAL SPREADER FORMED BY METAL ARMS

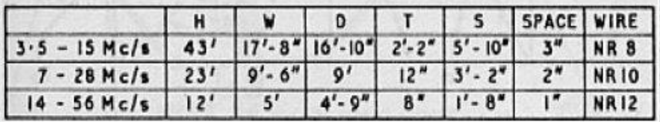

	H	W	D	T	S	SPACE	WIRE
3·5 – 15 Mc/s	43'	17'-8"	16'-10"	2'-2"	5'-10"	3"	NR 8
7 - 28 Mc/s	23'	9'-6"	9'	12"	3'-2"	2"	NR 10
14 - 56 Mc/s	12'	5'	4'-9"	8"	1'-8"	1"	NR 12

100 - Directive Delta-Birdcage Antenna for 20m to 10m

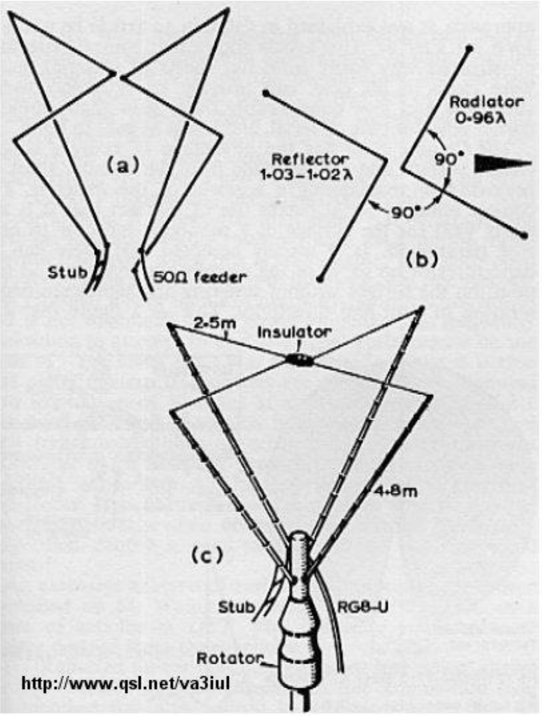
(a)
Stub
50Ω feeder
Radiator 0·96λ
Reflector 1·03–1·02λ
90°
90°
(b)
2·5m
Insulator
4·8m
(c)
Stub
RG8-U
Rotator
http://www.qsl.net/va3iul

101 - Dual Polarization Antenna for 80m and 40m

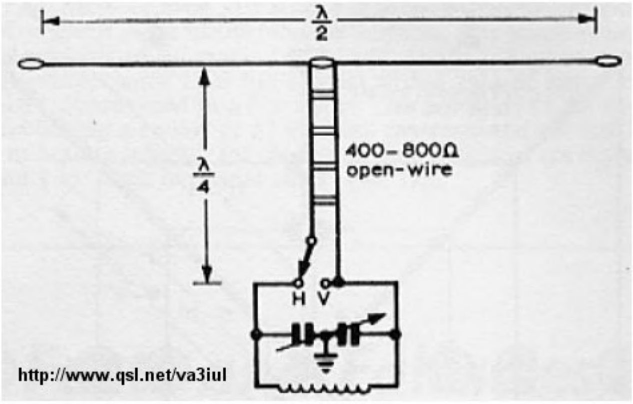
λ/2
λ/4
400–800Ω open-wire
H V
http://www.qsl.net/va3iul

102 - Directive 300-ohm-Ribbon Folded Dipole for 15m

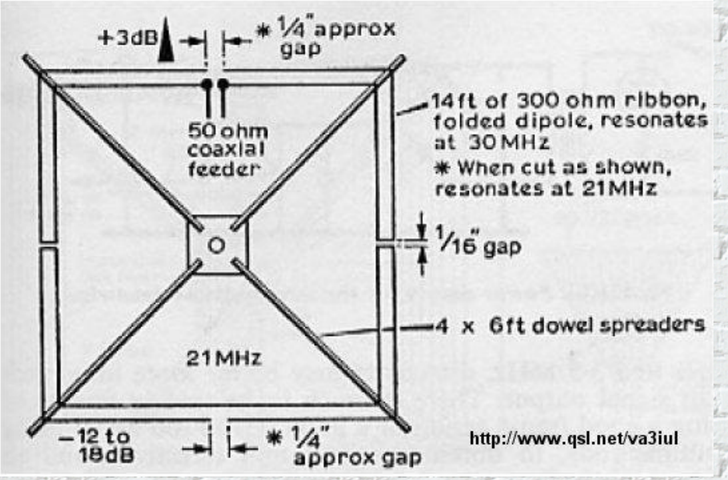
+3dB
* 1/4" approx gap
50 ohm coaxial feeder
14ft of 300 ohm ribbon, folded dipole, resonates at 30MHz
* When cut as shown, resonates at 21MHz
1/16" gap
4 x 6ft dowel spreaders
21MHz
-12 to 18dB
* 1/4" approx gap
http://www.qsl.net/va3iul

---

103 - Miniature Directive Antenna for 10m

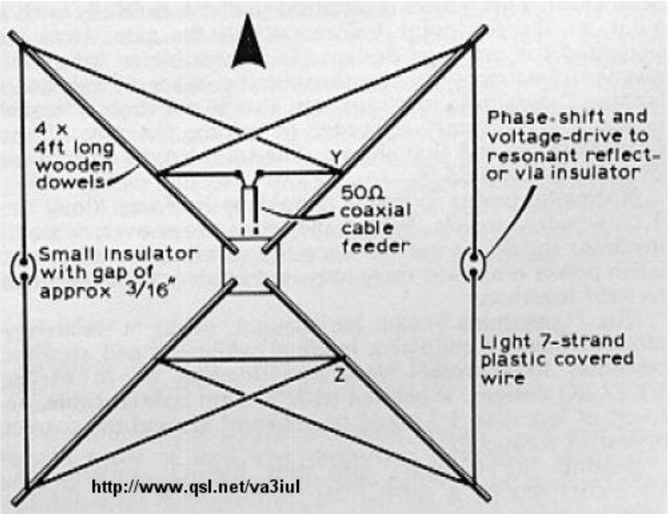
4 x 4ft long wooden dowels
Y
50Ω coaxial cable feeder
Phase-shift and voltage-drive to resonant reflector via insulator
Small insulator with gap of approx 3/16"
Light 7-strand plastic covered wire
Z
http://www.qsl.net/va3iul

---

104 - Biquad Antenna 12dBi-Gain for 2.4GHz

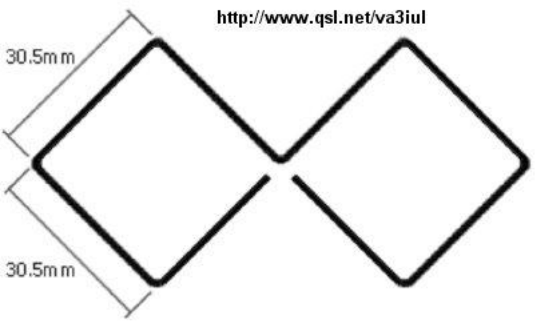
http://www.qsl.net/va3iul
30.5mm
30.5mm

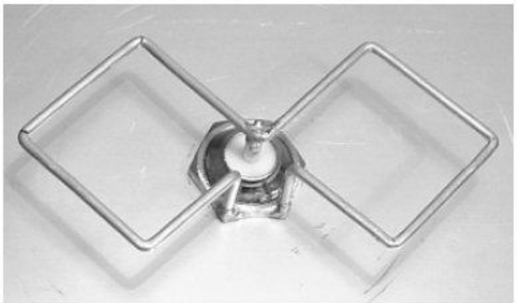

---

105 - Dual-Rhomboid Antenna for 435MHz to 870MHz

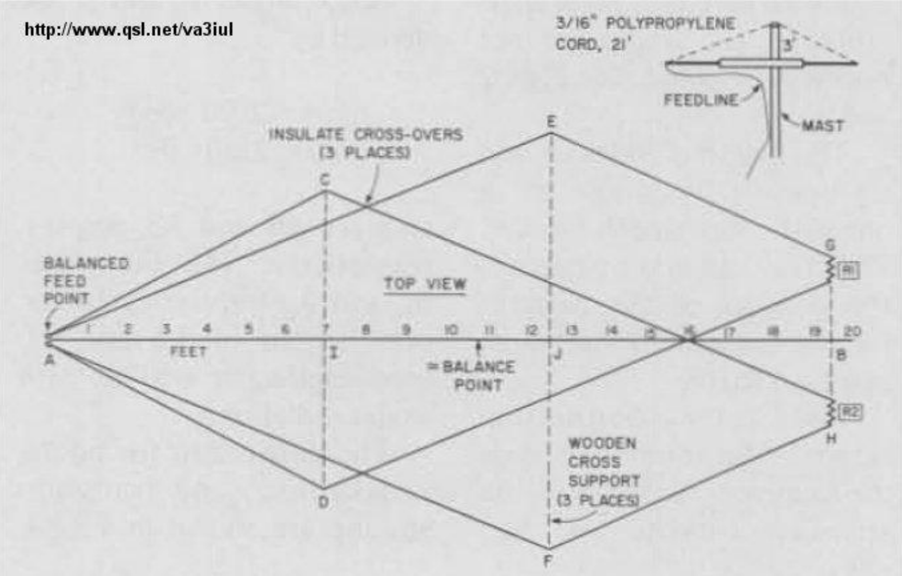
http://www.qsl.net/va3iul
3/16" POLYPROPYLENE CORD, 21'
3'
FEEDLINE
MAST
INSULATE CROSS-OVERS (3 PLACES)
E
C
BALANCED FEED POINT
TOP VIEW
G
R1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
A
FEET
I
≈ BALANCE POINT
J
B
R2
H
WOODEN CROSS SUPPORT (3 PLACES)
D
F

Dual rhomboid antenna for 435-870 MHz. Beamwidth ≈ 10° H x 6° V. Gain over DP ≈ 26 dB. Boom length: AB = 19'6". Support spacing : AI = 7'; IJ = 5'6"; JB = 7'. Support length: CD = 7'3"; EF = 10'3"; GH = 3'0". Rhomboid sides: AC, AD, EG, FH = 7'9.5"; AE, AF, CH, DG = 13'5.5". Feedline: Wire needed: 14 AWG formvar, ≈ 86'0". Boom material: AI, JB = wood; IJ = metal. Cross support: CD, EF, GH = wood. Terminators: R1, R2 = 600 Ohms; Watts

---

106 - Double-Bazooka Antenna for 80m

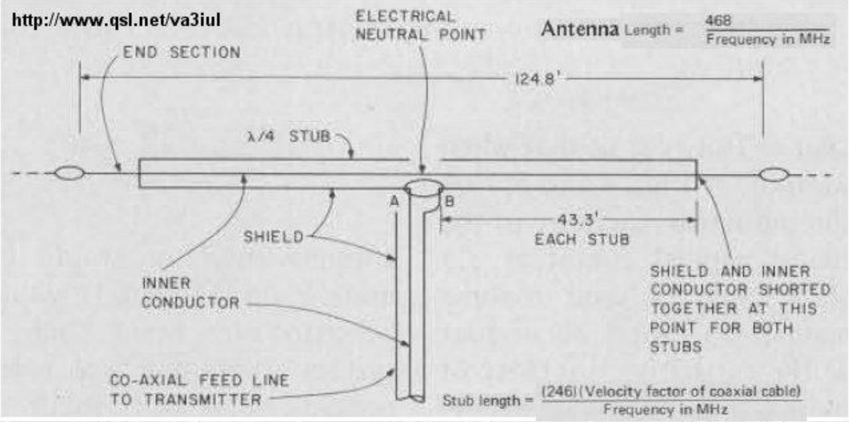
http://www.qsl.net/va3iul
ELECTRICAL NEUTRAL POINT
Antenna Length = 468 / Frequency in MHz
END SECTION
124.8'
λ/4 STUB
A
B
43.3' EACH STUB
SHIELD
INNER CONDUCTOR
SHIELD AND INNER CONDUCTOR SHORTED TOGETHER AT THIS POINT FOR BOTH STUBS
CO-AXIAL FEED LINE TO TRANSMITTER
Stub length = (246)(Velocity factor of coaxial cable) / Frequency in MHz

---

107 - J-Style Antenna

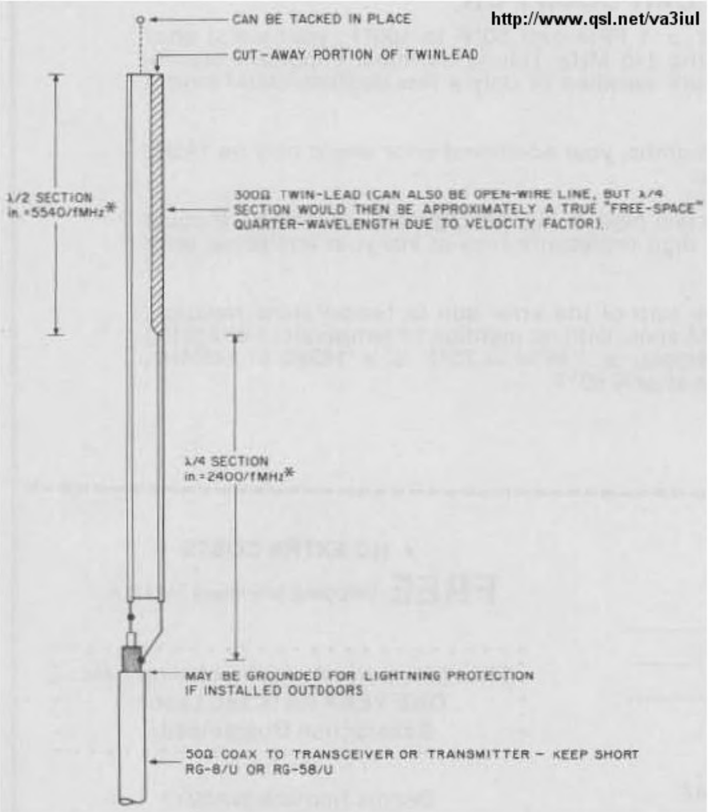
CAN BE TACKED IN PLACE
http://www.qsl.net/va3iul
CUT-AWAY PORTION OF TWINLEAD
λ/2 SECTION in.=5540/fMHz*
300Ω TWIN-LEAD (CAN ALSO BE OPEN-WIRE LINE, BUT λ/4 SECTION WOULD THEN BE APPROXIMATELY A TRUE "FREE-SPACE" QUARTER-WAVELENGTH DUE TO VELOCITY FACTOR).
λ/4 SECTION in.=2400/fMHz*
MAY BE GROUNDED FOR LIGHTNING PROTECTION IF INSTALLED OUTDOORS
50Ω COAX TO TRANSCEIVER OR TRANSMITTER – KEEP SHORT RG-8/U OR RG-58/U

108 - Vertical / Horizontal / Circular Polarization Antenna

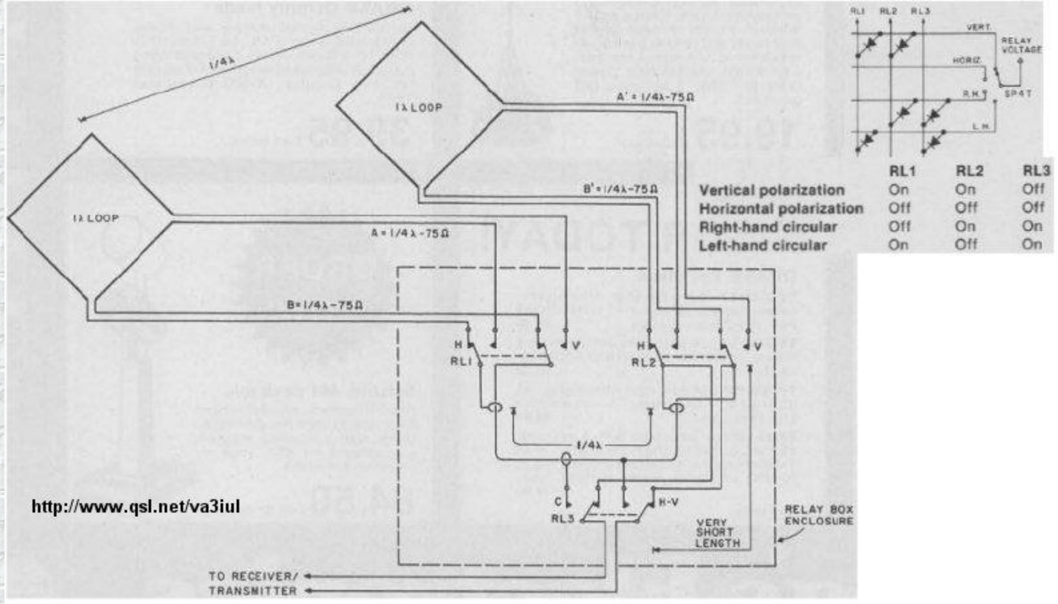
1/4λ
1λ LOOP
A' = 1/4λ-75Ω
B' = 1/4λ-75Ω
1λ LOOP
A = 1/4λ-75Ω
B = 1/4λ-75Ω
H
V
RL1
H
V
RL2
1/4λ
C
H-V
RL3
VERY SHORT LENGTH
RELAY BOX ENCLOSURE
TO RECEIVER/TRANSMITTER
http://www.qsl.net/va3iul
RL1
RL2
RL3
VERT.
HORIZ.
R.H.
L.H.
RELAY VOLTAGE
SP4T

	RL1	RL2	RL3
Vertical polarization	On	On	Off
Horizontal polarization	Off	Off	Off
Right-hand circular	Off	On	On
Left-hand circular	On	Off	On

109 - Coax Inverted-L Antenna for 80m

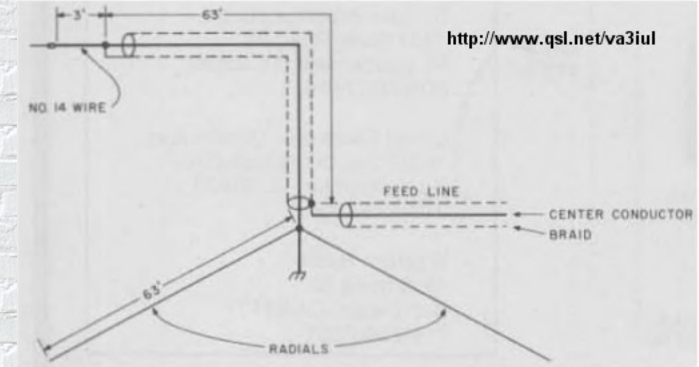
3'
63'
http://www.qsl.net/va3iul
NO. 14 WIRE
FEED LINE
CENTER CONDUCTOR
BRAID
63'
RADIALS

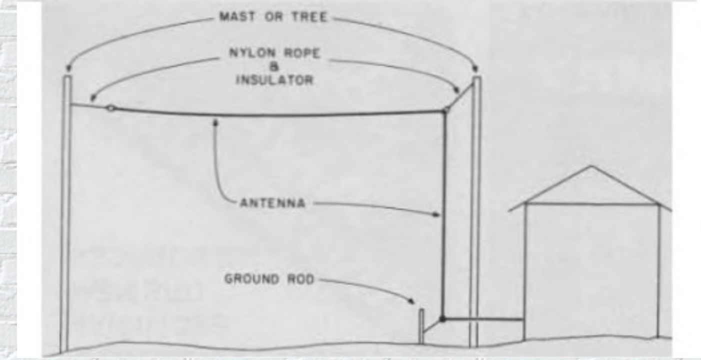
MAST OR TREE
NYLON ROPE & INSULATOR
ANTENNA
GROUND ROD

110 - Indoor Compact Loop Antenna for 80m

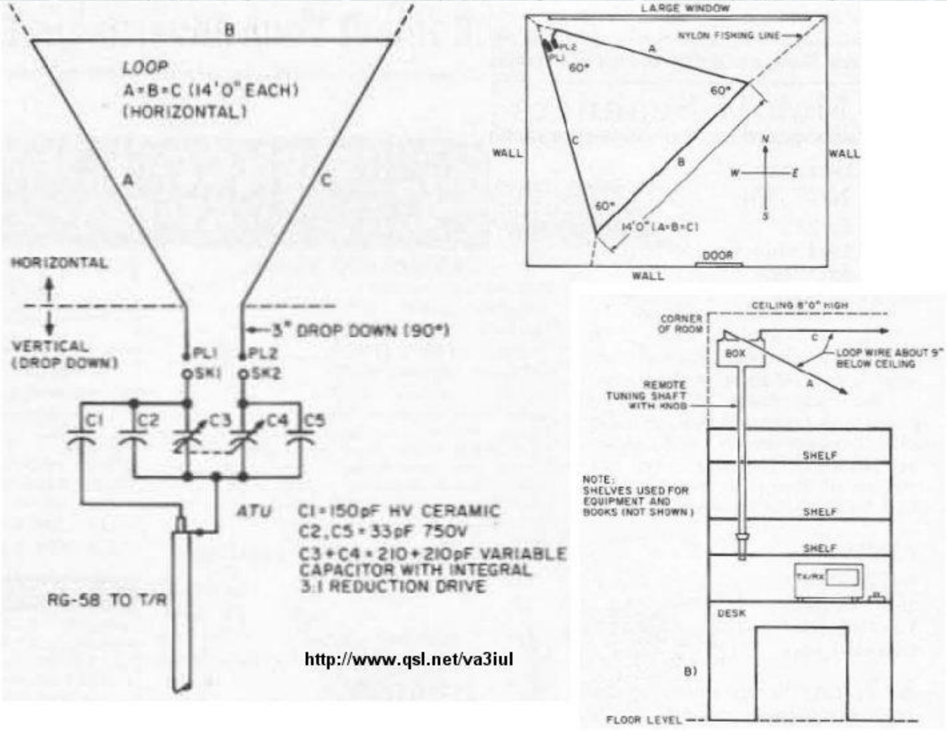
B
LOOP
A=B=C (14'0" EACH)
(HORIZONTAL)
A
C
HORIZONTAL
VERTICAL
(DROP DOWN)
3" DROP DOWN (90°)
PL1
PL2
SK1
SK2
C1
C2
C3
C4
C5
ATU
C1=150pF HV CERAMIC
C2,C5=33pF 750V
C3+C4=210+210pF VARIABLE
CAPACITOR WITH INTEGRAL
3:1 REDUCTION DRIVE
RG-58 TO T/R
http://www.qsl.net/va3iul
LARGE WINDOW
NYLON FISHING LINE
PL2
PL1
60°
60°
60°
WALL
WALL
WALL
N
W
E
S
14'0" (A=B=C)
DOOR
CEILING 8'0" HIGH
CORNER OF ROOM
BOX
LOOP WIRE ABOUT 9" BELOW CEILING
REMOTE TUNING SHAFT WITH KNOB
SHELF
SHELF
SHELF
NOTE: SHELVES USED FOR EQUIPMENT AND BOOKS (NOT SHOWN)
TX/RX
DESK
B)
FLOOR LEVEL

111 - Helix Antenna

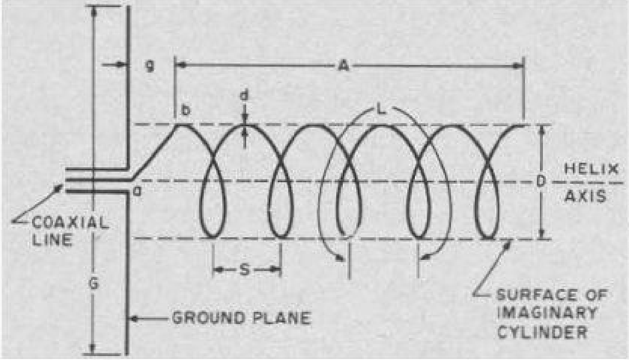
g
A
d
b
L
HELIX
AXIS
D
a
COAXIAL
LINE
S
G
GROUND PLANE
SURFACE OF
IMAGINARY
CYLINDER

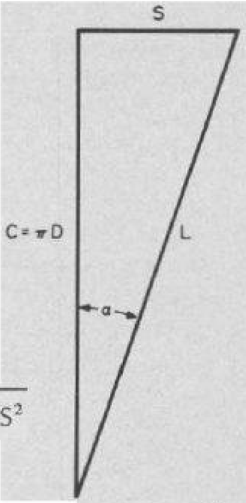

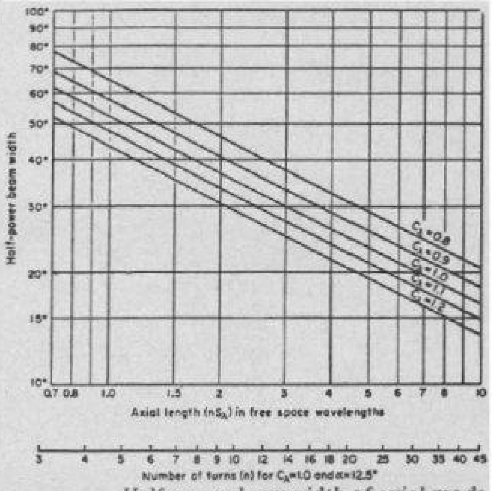

Half-power beamwidth of axial mode

D = Diameter of the helix
S = Spacing between turns
α = Pitch angle = Arctan $\left(\frac{S}{\pi D}\right)$
L = Length of 1 turn
n = Number of turns
A = Axial length = nS
d = Diameter of conductor
g = Distance of ground plane to first turn
G = Ground plane diameter

$$L = \sqrt{(\pi D)^2 + S^2}$$

$D = 0.32\lambda$
$S = 0.22\lambda$
$G \geqslant 0.80\lambda$

$$g = \frac{S}{2} = 0.12\lambda$$

$$\beta = \frac{52}{C_\lambda \sqrt{nS_\lambda}} \text{ degrees}$$

http://www.qsl.net/va3iul

---

112 - Novice Vertical Antenna for 80m, 40m, 15m, 10m

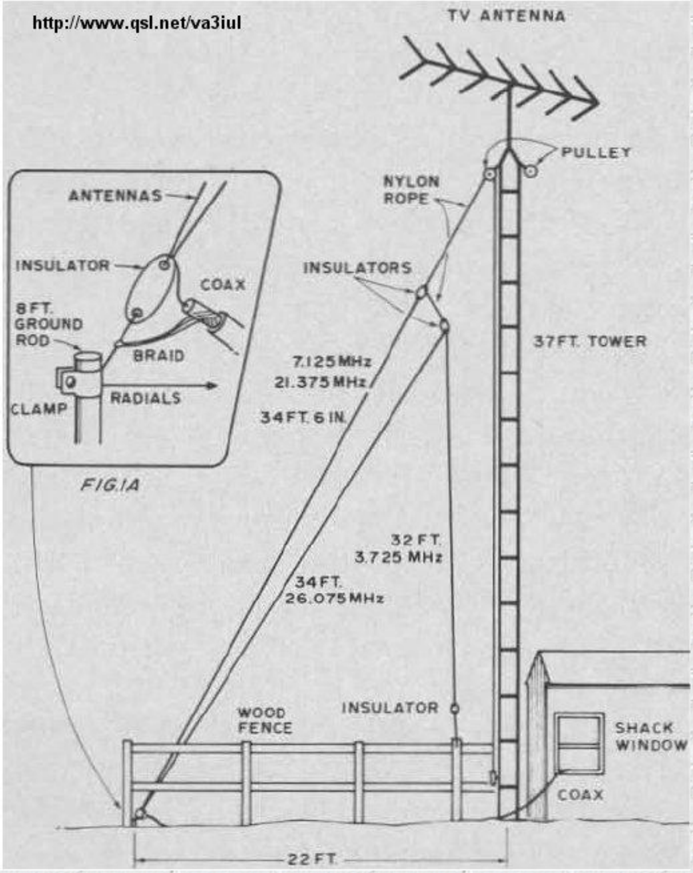
http://www.qsl.net/va3iul
TV ANTENNA
PULLEY
NYLON
ROPE
ANTENNAS
INSULATOR
COAX
8 FT.
GROUND
ROD
BRAID
CLAMP
RADIALS
FIG.1A
INSULATORS
37 FT. TOWER
7.125 MHz
21.375 MHz
34 FT. 6 IN.
32 FT.
3.725 MHz
34 FT.
26.075 MHz
WOOD
FENCE
INSULATOR
SHACK
WINDOW
COAX
22 FT.

---

113 - Stub-Loaded Shortened Dipole for 80m

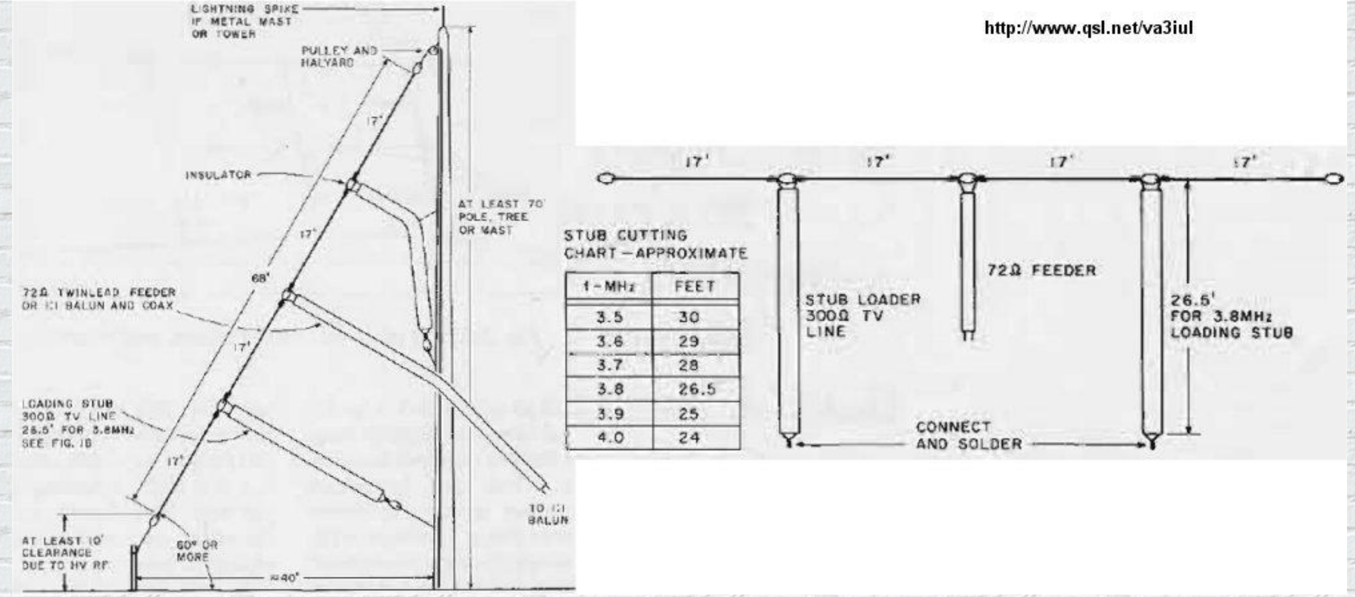
LIGHTNING SPIKE
IF METAL MAST
OR TOWER
PULLEY AND
HALYARD
17'
INSULATOR
AT LEAST 70'
POLE, TREE
OR MAST
17'
68'
72Ω TWINLEAD FEEDER
OR 1:1 BALUN AND COAX
17'
LOADING STUB
300Ω TV LINE
26.5' FOR 3.8MHz
SEE FIG. 18
17'
AT LEAST 10'
CLEARANCE
DUE TO HV RF
60° OR
MORE
≈40'
TO 1:1
BALUN
http://www.qsl.net/va3iul
17'
17'
17'
17'
72Ω FEEDER
STUB LOADER
300Ω TV
LINE
26.5'
FOR 3.8MHz
LOADING STUB
CONNECT
AND SOLDER

STUB CUTTING CHART – APPROXIMATE

f-MHz	FEET
3.5	30
3.6	29
3.7	28
3.8	26.5
3.9	25
4.0	24

114 - Six-Band Wire-Stub Trap Antenna for 40m-10m

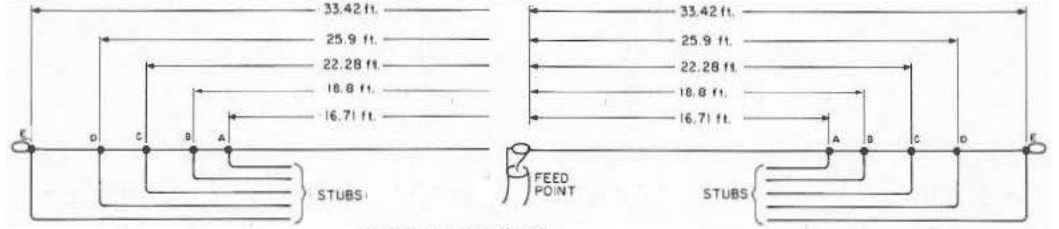
33.42 ft.
25.9 ft.
22.28 ft.
18.8 ft.
16.71 ft.
33.42 ft.
25.9 ft.
22.28 ft.
18.8 ft.
16.71 ft.
E
D
C
B
A
A
B
C
D
E
STUBS
FEED
POINT
STUBS

Point	Length (feet)	
A	8.37	40m
B	9.40	20m
C	11.14	17m
D	12.95	15m
E	16.72	12m
		10m

http://www.qsl.net/va3iul

115 - Multiband Half-Wave Delta-Loop Antenna

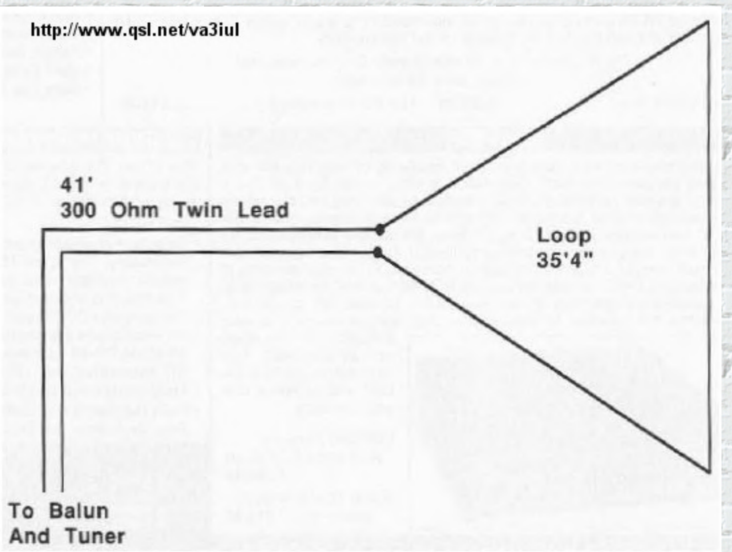
http://www.qsl.net/va3iul
41'
300 Ohm Twin Lead
Loop
35'4"
To Balun
And Tuner

116 - Hybrid Vee for 20m, 17m

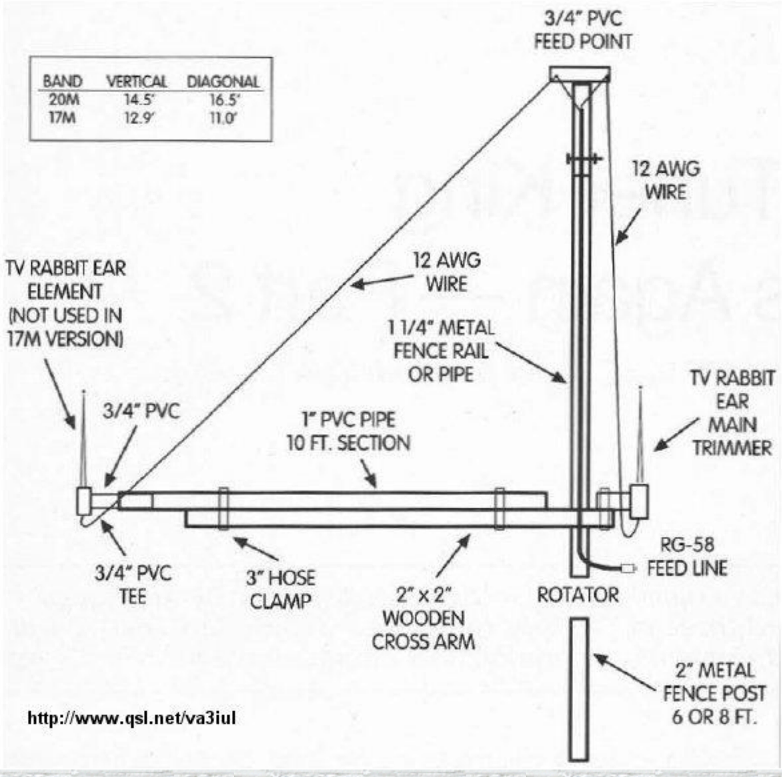
BAND VERTICAL DIAGONAL
20M 14.5' 16.5'
17M 12.9' 11.0'
3/4" PVC FEED POINT
12 AWG WIRE
12 AWG WIRE
TV RABBIT EAR ELEMENT (NOT USED IN 17M VERSION)
1 1/4" METAL FENCE RAIL OR PIPE
TV RABBIT EAR MAIN TRIMMER
3/4" PVC
1" PVC PIPE 10 FT. SECTION
RG-58 FEED LINE
3/4" PVC TEE
3" HOSE CLAMP
2" x 2" WOODEN CROSS ARM
ROTATOR
2" METAL FENCE POST 6 OR 8 FT.
http://www.qsl.net/va3iul

117 - Six-Shooter Array Antenna, Gain=7.5dB

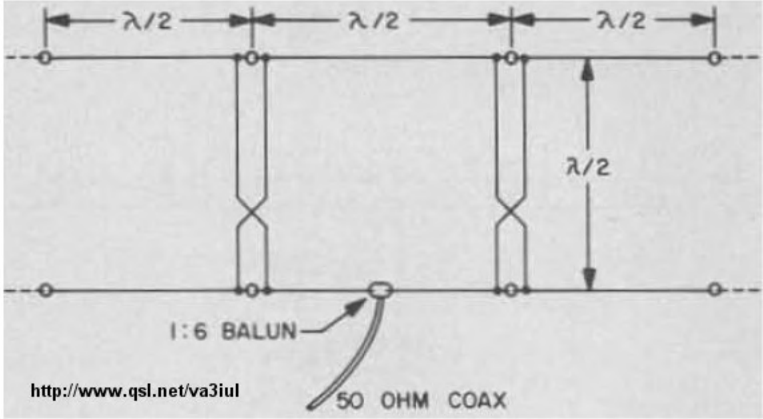
λ/2
λ/2
λ/2
λ/2
1:6 BALUN
50 OHM COAX
http://www.qsl.net/va3iul

118 - Multiband Ground-Plane Antenna for 10m, 15m, 20m

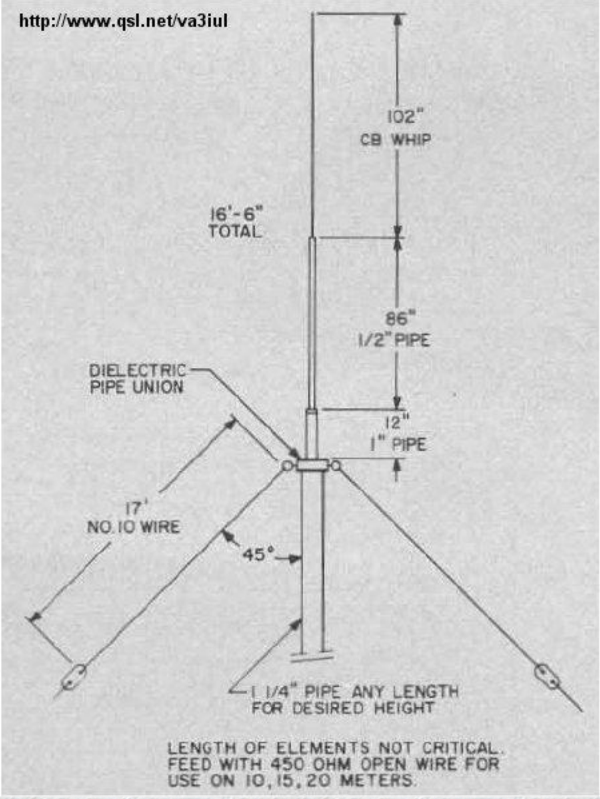
http://www.qsl.net/va3iul
102"
CB WHIP
16'-6"
TOTAL
86"
1/2" PIPE
DIELECTRIC
PIPE UNION
12"
1" PIPE
17'
NO.10 WIRE
45°
1 1/4" PIPE ANY LENGTH
FOR DESIRED HEIGHT
LENGTH OF ELEMENTS NOT CRITICAL.
FEED WITH 450 OHM OPEN WIRE FOR
USE ON 10, 15, 20 METERS.

119 - Wire Superbeam Antenna for 10m, 15m, 20m

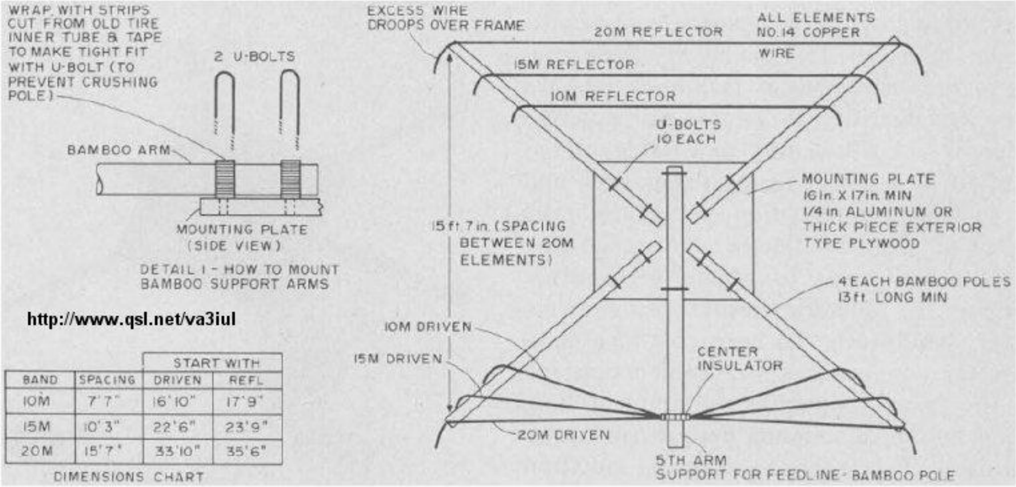
WRAP WITH STRIPS
CUT FROM OLD TIRE
INNER TUBE & TAPE
TO MAKE TIGHT FIT
WITH U-BOLT (TO
PREVENT CRUSHING
POLE)
2 U-BOLTS
BAMBOO ARM
MOUNTING PLATE
(SIDE VIEW)
DETAIL I - HOW TO MOUNT
BAMBOO SUPPORT ARMS
http://www.qsl.net/va3iul
EXCESS WIRE
DROOPS OVER FRAME
20M REFLECTOR
ALL ELEMENTS
NO.14 COPPER
WIRE
15M REFLECTOR
10M REFLECTOR
U-BOLTS
10 EACH
MOUNTING PLATE
16in. X 17in. MIN
1/4 in. ALUMINUM OR
THICK PIECE EXTERIOR
TYPE PLYWOOD
15 ft. 7 in. (SPACING
BETWEEN 20M
ELEMENTS)
4 EACH BAMBOO POLES
13 ft. LONG MIN
10M DRIVEN
15M DRIVEN
CENTER
INSULATOR
20M DRIVEN
5TH ARM
SUPPORT FOR FEEDLINE - BAMBOO POLE

		START WITH	
BAND	SPACING	DRIVEN	REFL
10M	7'7"	16'10"	17'9"
15M	10'3"	22'6"	23'9"
20M	15'7"	33'10"	35'6"

DIMENSIONS CHART

120 - Two Elements Delta-Loop Antenna

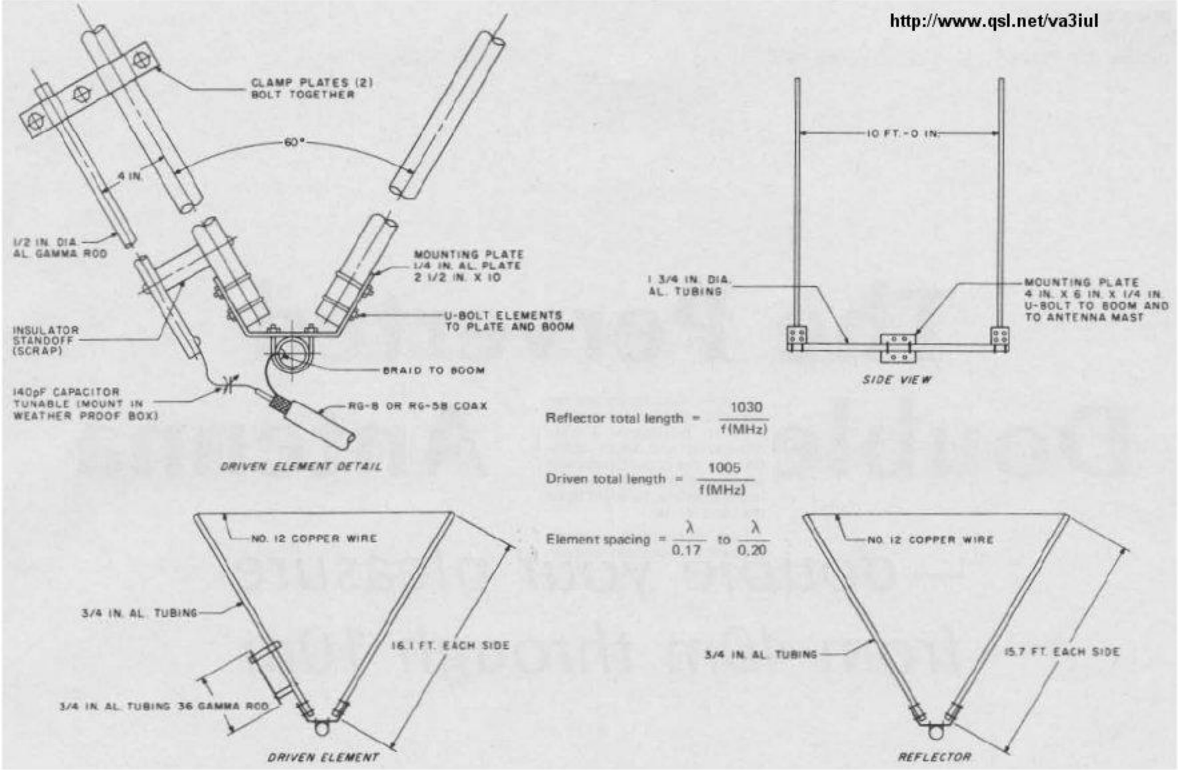
http://www.qsl.net/va3iul
CLAMP PLATES (2) BOLT TOGETHER
60°
4 IN.
1/2 IN. DIA AL. GAMMA ROD
MOUNTING PLATE 1/4 IN. AL. PLATE 2 1/2 IN. X 10
U-BOLT ELEMENTS TO PLATE AND BOOM
INSULATOR STANDOFF (SCRAP)
BRAID TO BOOM
140pF CAPACITOR TUNABLE (MOUNT IN WEATHER PROOF BOX)
RG-8 OR RG-58 COAX
DRIVEN ELEMENT DETAIL
10 FT.-0 IN.
1 3/4 IN. DIA. AL. TUBING
MOUNTING PLATE 4 IN. X 6 IN. X 1/4 IN. U-BOLT TO BOOM AND TO ANTENNA MAST
SIDE VIEW
Reflector total length = 1030 / f(MHz)
Driven total length = 1005 / f(MHz)
Element spacing = λ / 0.17 to λ / 0.20
NO. 12 COPPER WIRE
3/4 IN. AL. TUBING
16.1 FT. EACH SIDE
3/4 IN. AL. TUBING 36 GAMMA ROD
DRIVEN ELEMENT
NO. 12 COPPER WIRE
3/4 IN. AL. TUBING
15.7 FT. EACH SIDE
REFLECTOR

121 - Sterba Curtain Antenna

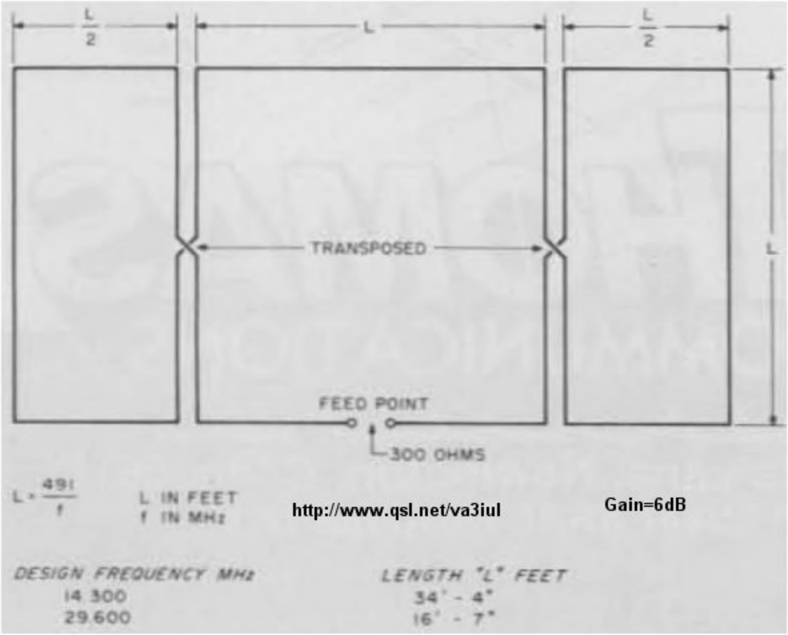
L/2
L
L/2
TRANSPOSED
L
FEED POINT
300 OHMS
L = 491/f
L IN FEET
f IN MHz
http://www.qsl.net/va3iul
Gain=6dB
DESIGN FREQUENCY MHz
14.300
29.600
LENGTH "L" FEET
34' - 4"
16' - 7"

122 - Half-Wave Vertical Zepp Antenna

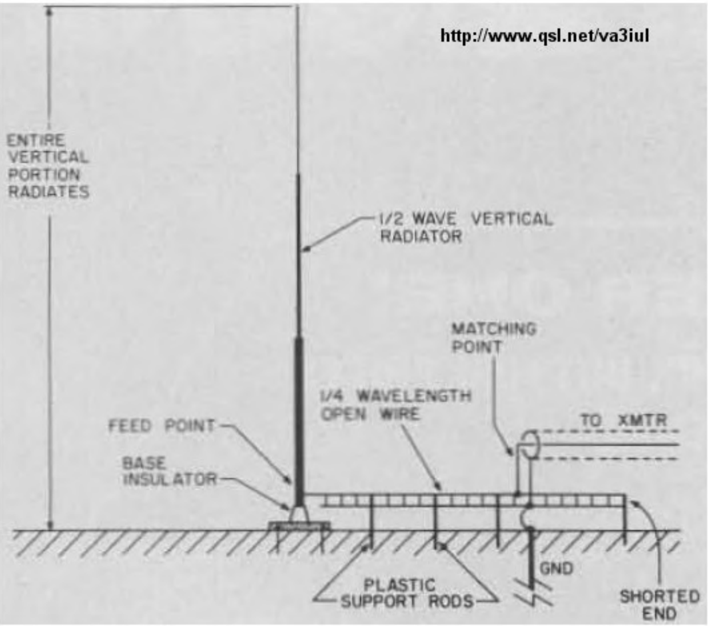
http://www.qsl.net/va3iul
ENTIRE VERTICAL PORTION RADIATES
1/2 WAVE VERTICAL RADIATOR
MATCHING POINT
1/4 WAVELENGTH OPEN WIRE
TO XMTR
FEED POINT
BASE INSULATOR
GND
PLASTIC SUPPORT RODS
SHORTED END

123 - Lazy-Loop Antenna for 40m

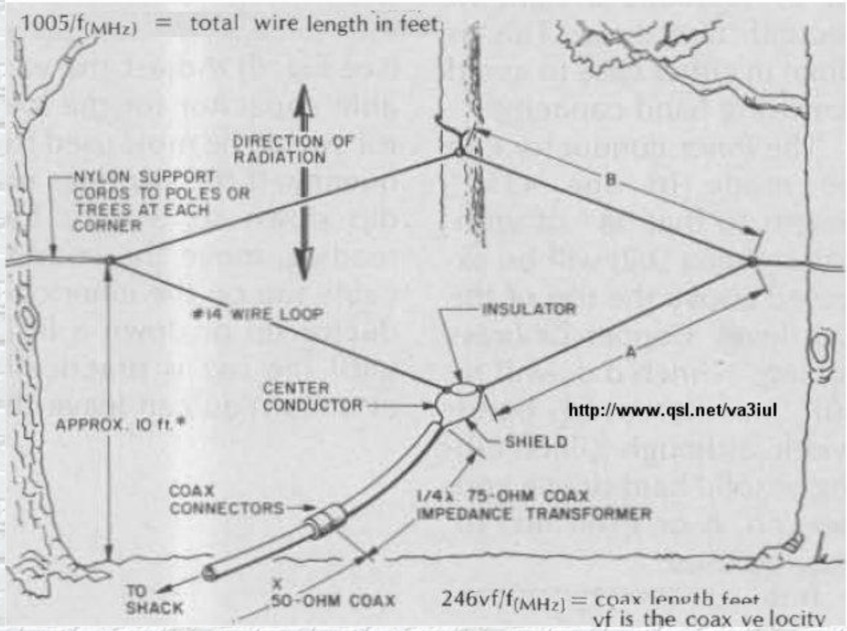
1005/f(MHz) = total wire length in feet
DIRECTION OF RADIATION
NYLON SUPPORT CORDS TO POLES OR TREES AT EACH CORNER
B
#14 WIRE LOOP
INSULATOR
A
CENTER CONDUCTOR
http://www.qsl.net/va3iul
SHIELD
APPROX. 10 ft.*
COAX CONNECTORS
1/4λ 75-OHM COAX IMPEDANCE TRANSFORMER
TO SHACK
X
50-OHM COAX
246vf/f(MHz) = coax length feet
vf is the coax velocity

124 - Terminated Folded Dipole for 80m, 40m

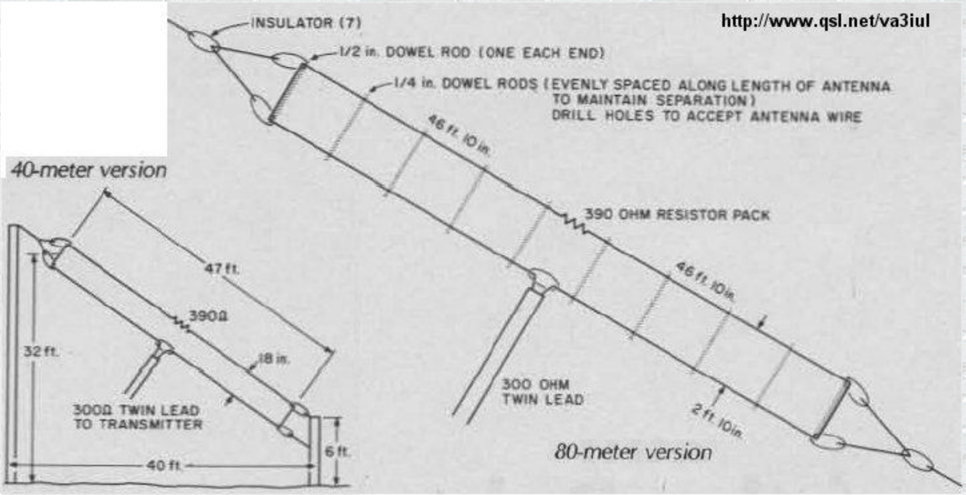
INSULATOR (7)
http://www.qsl.net/va3iul
1/2 in. DOWEL ROD (ONE EACH END)
1/4 in. DOWEL RODS (EVENLY SPACED ALONG LENGTH OF ANTENNA TO MAINTAIN SEPARATION) DRILL HOLES TO ACCEPT ANTENNA WIRE
46 ft. 10 in.
390 OHM RESISTOR PACK
46 ft. 10 in.
300 OHM TWIN LEAD
2 ft. 10 in.
80-meter version
40-meter version
47 ft.
390Ω
32 ft.
18 in.
300Ω TWIN LEAD TO TRANSMITTER
6 ft.
40 ft.

125 - Short-Fat Antenna for 15m

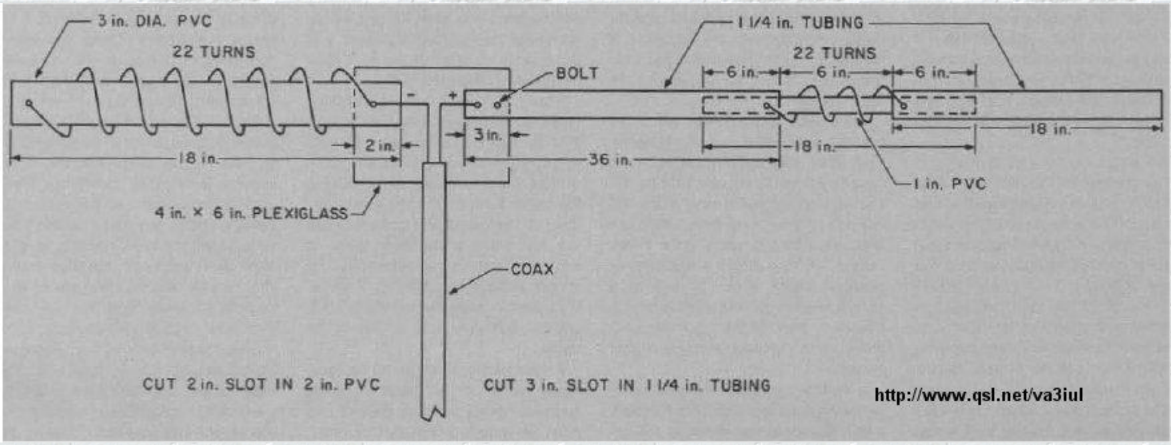
3 in. DIA. PVC
22 TURNS
BOLT
1 1/4 in. TUBING
22 TURNS
6 in.
6 in.
6 in.
-
+
2 in.
3 in.
18 in.
36 in.
18 in.
18 in.
1 in. PVC
4 in. × 6 in. PLEXIGLASS
COAX
CUT 2 in. SLOT IN 2 in. PVC
CUT 3 in. SLOT IN 1 1/4 in. TUBING
http://www.qsl.net/va3iul

126 - Cobra Antenna for 80m

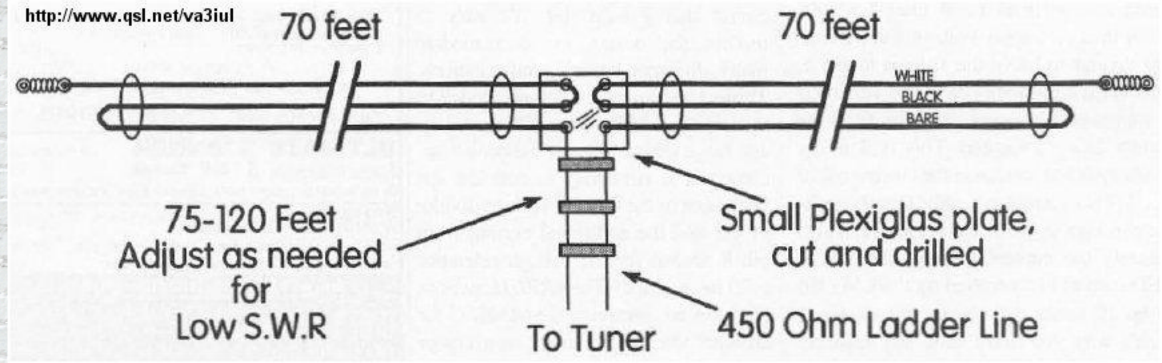
http://www.qsl.net/va3iul
70 feet
70 feet
WHITE
BLACK
BARE
75-120 Feet
Adjust as needed
for
Low S.W.R
To Tuner
Small Plexiglas plate,
cut and drilled
450 Ohm Ladder Line

127 - Log-Periodic Wire Antenna for 20m, 15m, 10m

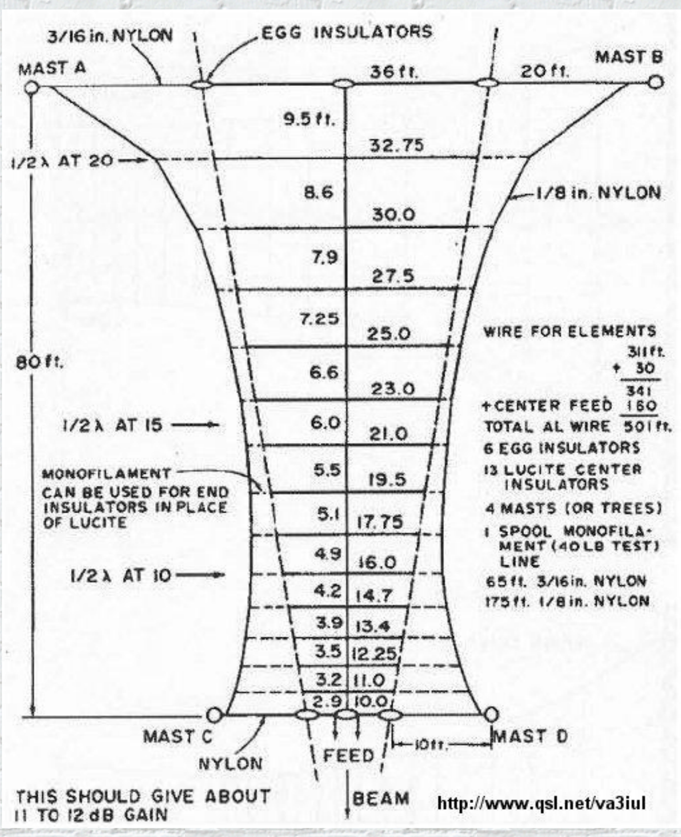
3/16 in. NYLON
EGG INSULATORS
MAST A
36 ft.
20 ft.
MAST B
9.5 ft.
32.75
1/2 λ AT 20
8.6
1/8 in. NYLON
30.0
7.9
27.5
7.25
25.0
80 ft.
6.6
23.0
1/2 λ AT 15
6.0
21.0
5.5
19.5
MONOFILAMENT
CAN BE USED FOR END
INSULATORS IN PLACE
OF LUCITE
5.1
17.75
4.9
16.0
1/2 λ AT 10
4.2
14.7
3.9
13.4
3.5
12.25
3.2
11.0
2.9
10.0
MAST C
NYLON
FEED
10 ft.
MAST D
BEAM
http://www.qsl.net/va3iul

WIRE FOR ELEMENTS 311 ft.
+ 30
341
\+CENTER FEED 160
TOTAL AL WIRE 501 ft.
6 EGG INSULATORS
13 LUCITE CENTER INSULATORS
4 MASTS (OR TREES)
1 SPOOL MONOFILAMENT (40 LB TEST) LINE
65 ft. 3/16 in. NYLON
175 ft. 1/8 in. NYLON

THIS SHOULD GIVE ABOUT
11 TO 12 dB GAIN

128 - 5-Element Log-Periodic Vertical Antenna for 80m, 40m, 20m

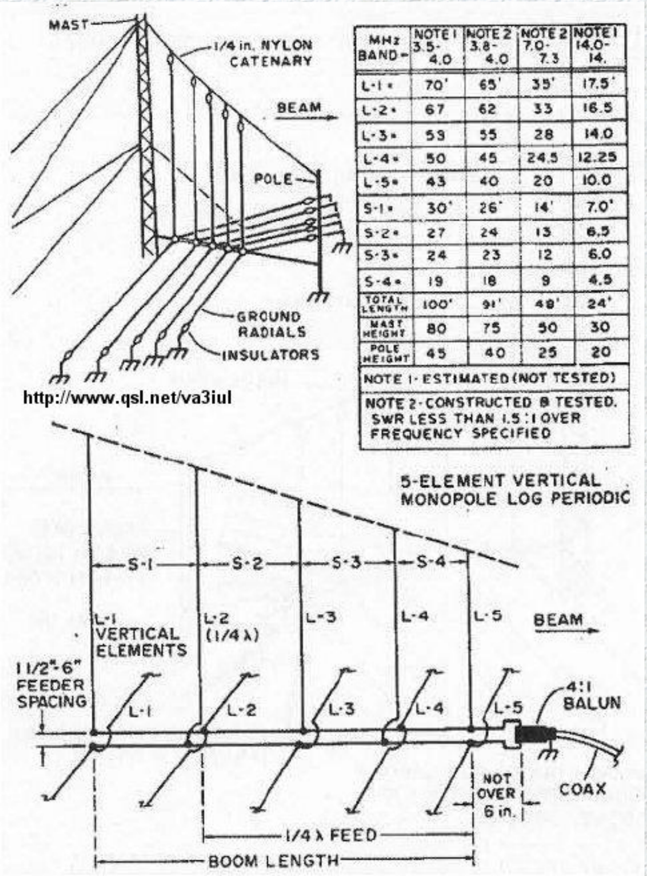
MAST
1/4 in. NYLON CATENARY
BEAM
POLE
GROUND RADIALS
INSULATORS
http://www.qsl.net/va3iul
5-ELEMENT VERTICAL MONOPOLE LOG PERIODIC
S-1
S-2
S-3
S-4
L-1 VERTICAL ELEMENTS
L-2 (1/4 λ)
L-3
L-4
L-5
BEAM
1 1/2"-6" FEEDER SPACING
L-1
L-2
L-3
L-4
L-5
4:1 BALUN
COAX
NOT OVER 6 in.
1/4 λ FEED
BOOM LENGTH

MHz BAND-	NOTE 1 3.5-4.0	NOTE 2 3.8-4.0	NOTE 2 7.0-7.3	NOTE 1 14.0-14.
L-1 =	70'	65'	35'	17.5'
L-2 =	67	62	33	16.5
L-3 =	53	55	28	14.0
L-4 =	50	45	24.5	12.25
L-5 =	43	40	20	10.0
S-1 =	30'	26'	14'	7.0'
S-2 =	27	24	13	6.5
S-3 =	24	23	12	6.0
S-4 =	19	18	9	4.5
TOTAL LENGTH	100'	91'	48'	24'
MAST HEIGHT	80	75	50	30
POLE HEIGHT	45	40	25	20

NOTE 1 - ESTIMATED (NOT TESTED)

NOTE 2 - CONSTRUCTED & TESTED. SWR LESS THAN 1.5:1 OVER FREQUENCY SPECIFIED

129 - 2m Vertical Wire Antennas

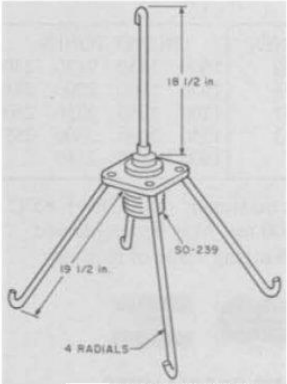
18 1/2 in.
SO-239
19 1/2 in.
4 RADIALS

Simple 1/4λ 2-meter ground-plane antenna.

http://www.qsl.net/va3iul

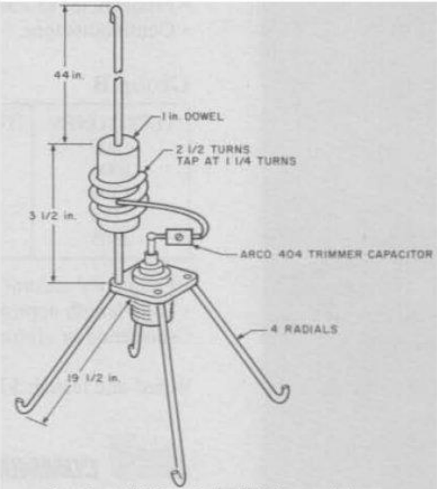
44 in.
1 in. DOWEL
2 1/2 TURNS TAP AT 1 1/4 TURNS
3 1/2 in.
ARCO 404 TRIMMER CAPACITOR
4 RADIALS
19 1/2 in.

Experimental 5/8λ vertical with 1/4λ ground plane.

130 - Earth-Mover Inverted-V Antenna for 40m

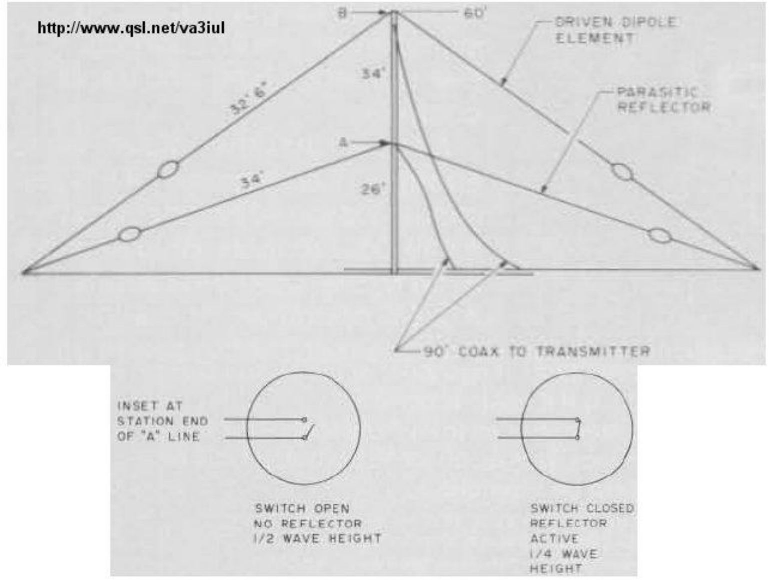
http://www.qsl.net/va3iul
B
60'
DRIVEN DIPOLE ELEMENT
34'
32' 6"
PARASITIC REFLECTOR
A
34'
26'
90' COAX TO TRANSMITTER
INSET AT STATION END OF "A" LINE
SWITCH OPEN
NO REFLECTOR
1/2 WAVE HEIGHT
SWITCH CLOSED
REFLECTOR ACTIVE
1/4 WAVE HEIGHT

131 - Coax-Cable Collinear Antennas

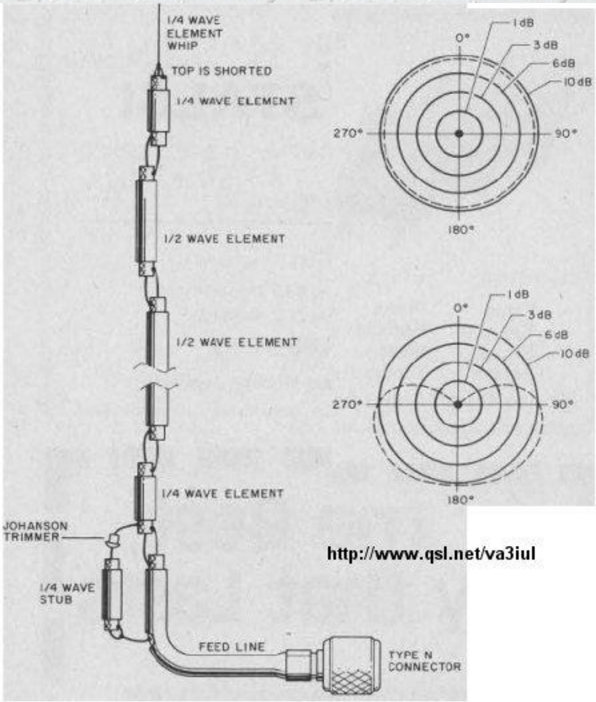
1/4 WAVE ELEMENT WHIP
TOP IS SHORTED
1/4 WAVE ELEMENT
1/2 WAVE ELEMENT
1/2 WAVE ELEMENT
1/4 WAVE ELEMENT
JOHANSON TRIMMER
1/4 WAVE STUB
FEED LINE
TYPE N CONNECTOR
0°
1 dB
3 dB
6 dB
10 dB
270°
90°
180°
0°
1 dB
3 dB
6 dB
10 dB
270°
90°
180°
http://www.qsl.net/va3iul

8 elements = 6 dB. 16 elements = 9 dB

132 - Double Bobtail Antenna for 20m

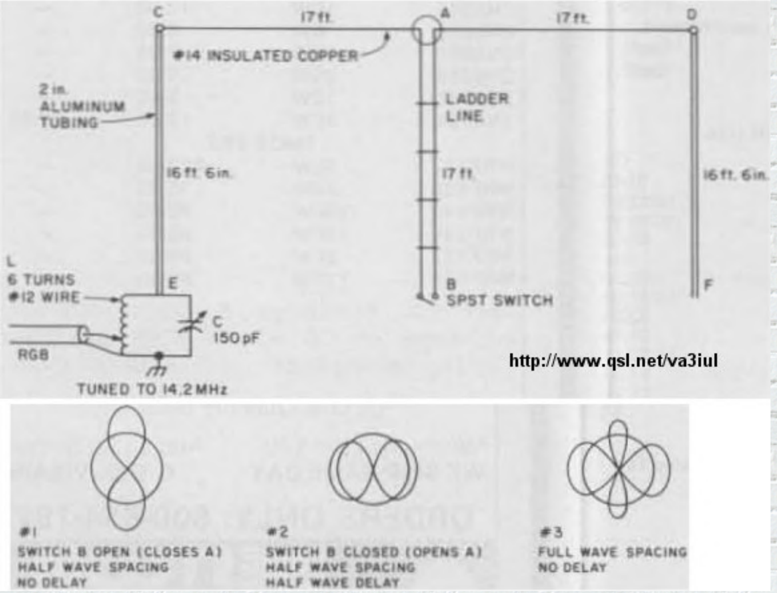
C
17 ft.
A
17 ft.
D
#14 INSULATED COPPER
2 in. ALUMINUM TUBING
LADDER LINE
16 ft. 6in.
17 ft.
16 ft. 6in.
L
6 TURNS #12 WIRE
E
B
SPST SWITCH
F
C
150 pF
RG8
TUNED TO 14.2 MHz
http://www.qsl.net/va3iul
#1
SWITCH B OPEN (CLOSES A)
HALF WAVE SPACING
NO DELAY
#2
SWITCH B CLOSED (OPENS A)
HALF WAVE SPACING
HALF WAVE DELAY
#3
FULL WAVE SPACING
NO DELAY

133 - Collinear Zepp Antenna

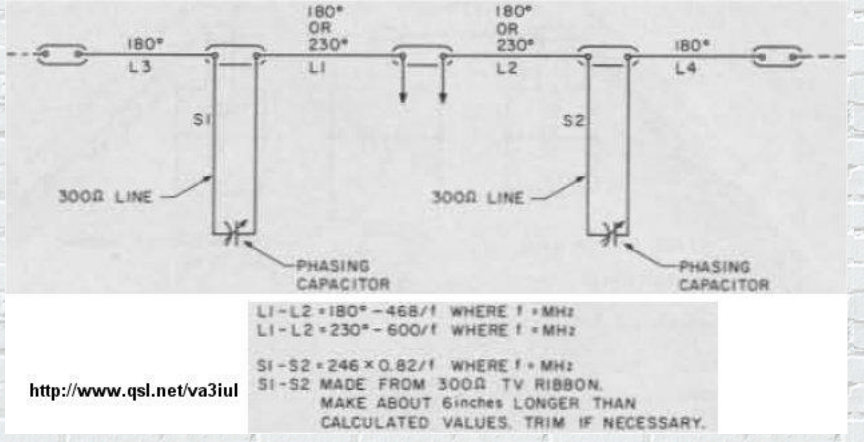
180°
L3
180° OR 230°
L1
180° OR 230°
L2
180°
L4
S1
S2
300Ω LINE
300Ω LINE
PHASING CAPACITOR
PHASING CAPACITOR
http://www.qsl.net/va3iul

L1-L2 = 180° - 468/f WHERE f = MHz
L1-L2 = 230° - 600/f WHERE f = MHz

S1-S2 = 246 × 0.82/f WHERE f = MHz
S1-S2 MADE FROM 300Ω TV RIBBON. MAKE ABOUT 6inches LONGER THAN CALCULATED VALUES. TRIM IF NECESSARY.

134 - Taylor Vee Antenna for 20m

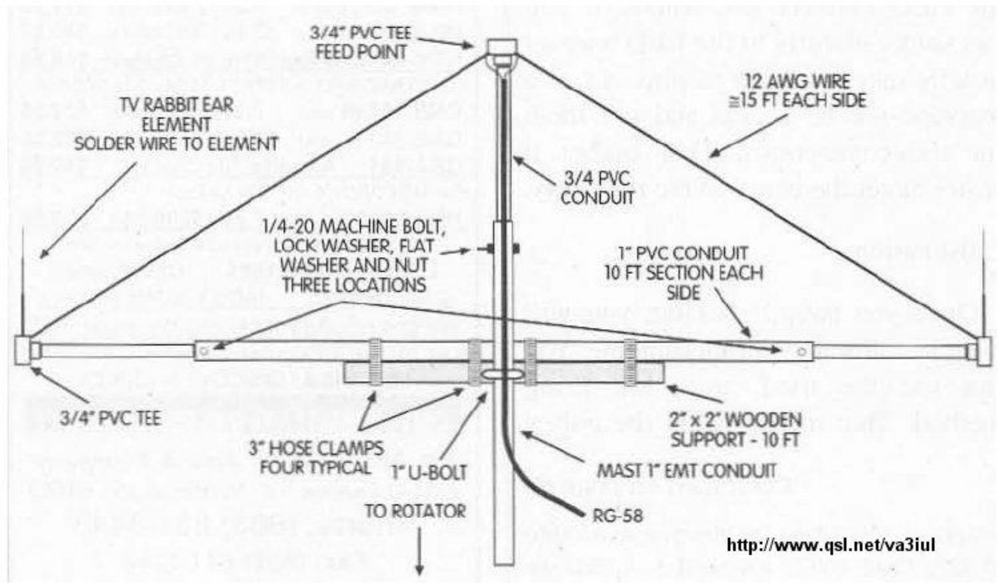
3/4" PVC TEE
FEED POINT
12 AWG WIRE
≥15 FT EACH SIDE
TV RABBIT EAR
ELEMENT
SOLDER WIRE TO ELEMENT
3/4 PVC
CONDUIT
1/4-20 MACHINE BOLT,
LOCK WASHER, FLAT
WASHER AND NUT
THREE LOCATIONS
1" PVC CONDUIT
10 FT SECTION EACH
SIDE
3/4" PVC TEE
2" x 2" WOODEN
SUPPORT - 10 FT
3" HOSE CLAMPS
FOUR TYPICAL
1" U-BOLT
MAST 1" EMT CONDUIT
TO ROTATOR
RG-58
http://www.qsl.net/va3iul

135 - Collinear Vertical Antenna 6dB-Gain for 2m, 1.3m, 73cm

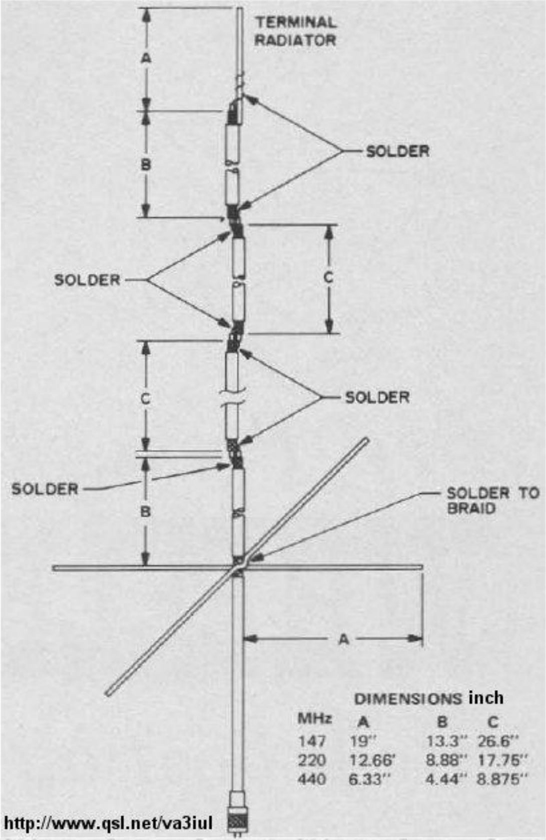
TERMINAL RADIATOR
A
B
SOLDER
C
SOLDER
C
SOLDER
SOLDER
B
SOLDER TO BRAID
A
http://www.qsl.net/va3iul

DIMENSIONS inch

MHz	A	B	C
147	19"	13.3"	26.6"
220	12.66'	8.88"	17.75"
440	6.33"	4.44"	8.875"

136 - Bi-Loop Antenna for 20m

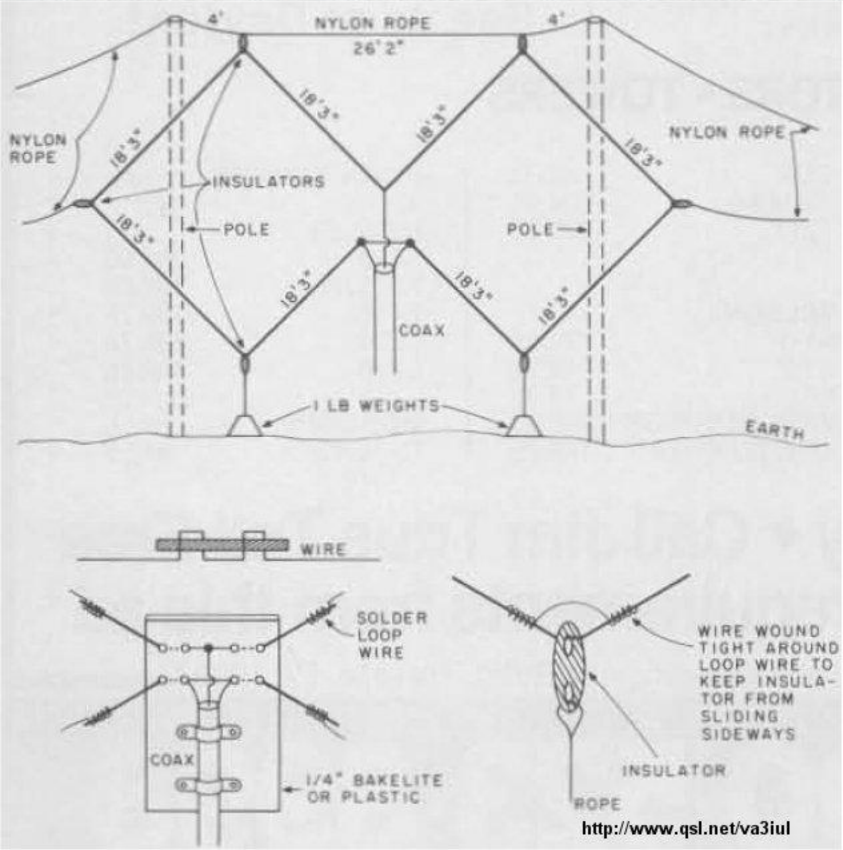
4'
NYLON ROPE
26'2"
4'
NYLON ROPE
18'3"
18'3"
18'3"
18'3"
NYLON ROPE
INSULATORS
POLE
POLE
18'3"
18'3"
18'3"
18'3"
COAX
1 LB WEIGHTS
EARTH
WIRE
SOLDER LOOP WIRE
COAX
1/4" BAKELITE OR PLASTIC
WIRE WOUND TIGHT AROUND LOOP WIRE TO KEEP INSULATOR FROM SLIDING SIDEWAYS
INSULATOR
ROPE
http://www.qsl.net/va3iul

137 - Wire Beam 6dBd-Gain for 10m

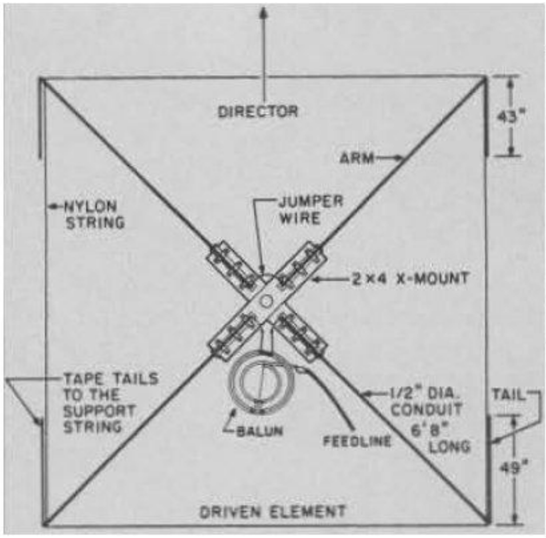
DIRECTOR
ARM
43"
NYLON STRING
JUMPER WIRE
2×4 X-MOUNT
TAPE TAILS TO THE SUPPORT STRING
BALUN
FEEDLINE
1/2" DIA. CONDUIT 6'8" LONG
TAIL
49"
DRIVEN ELEMENT

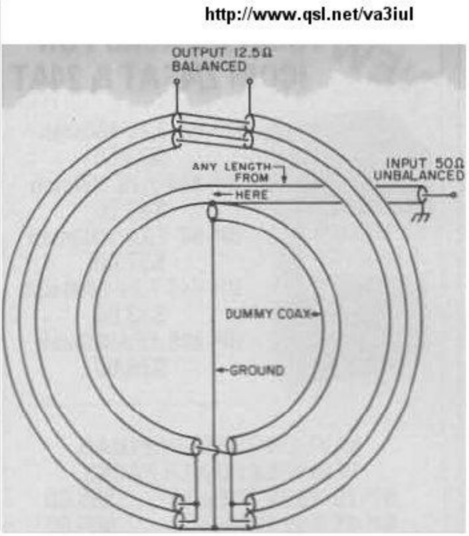
http://www.qsl.net/va3iul
OUTPUT 12.5Ω BALANCED
ANY LENGTH FROM HERE
INPUT 50Ω UNBALANCED
DUMMY COAX
GROUND

138 - Sloping Diamond Antenna 4dB-Gain for 40m

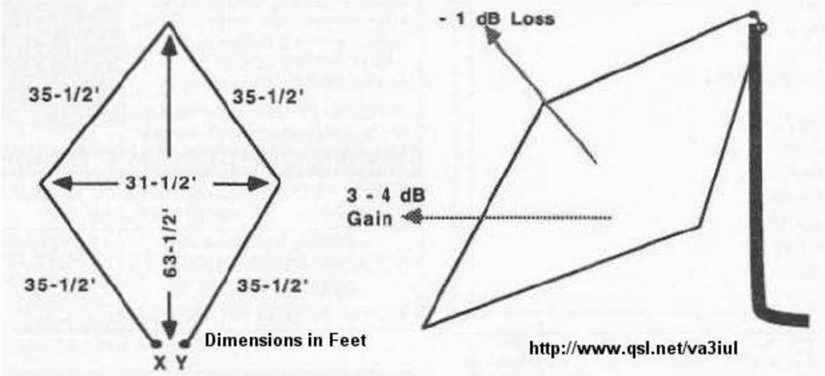
35-1/2'
35-1/2'
31-1/2'
63-1/2'
35-1/2'
35-1/2'
X Y
Dimensions in Feet
- 1 dB Loss
3 - 4 dB Gain
http://www.qsl.net/va3iul

139 - Twisted Loop Antenna for 160m

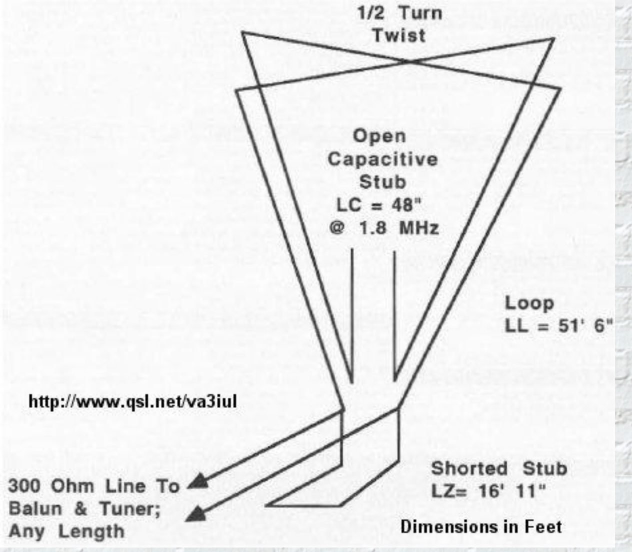
1/2 Turn Twist
Open Capacitive Stub LC = 48" @ 1.8 MHz
Loop LL = 51' 6"
300 Ohm Line To Balun & Tuner; Any Length
Shorted Stub LZ= 16' 11"
Dimensions in Feet
http://www.qsl.net/va3iul

140 - DX RX Loop Antenna for 160m

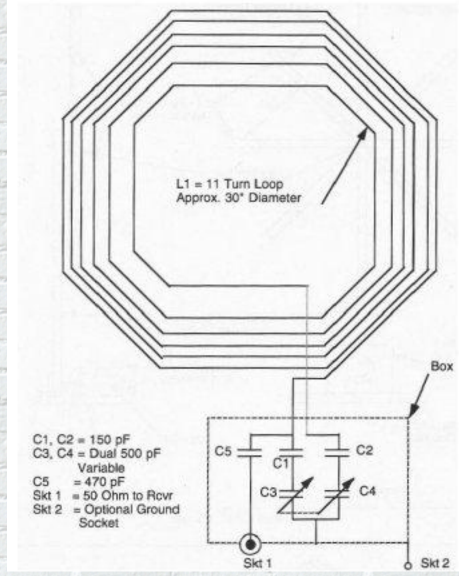
L1 = 11 Turn Loop
Approx. 30" Diameter
Box
C1, C2 = 150 pF
C3, C4 = Dual 500 pF
Variable
C5 = 470 pF
Skt 1 = 50 Ohm to Rcvr
Skt 2 = Optional Ground
Socket
C5
C1
C2
C3
C4
Skt 1
Skt 2

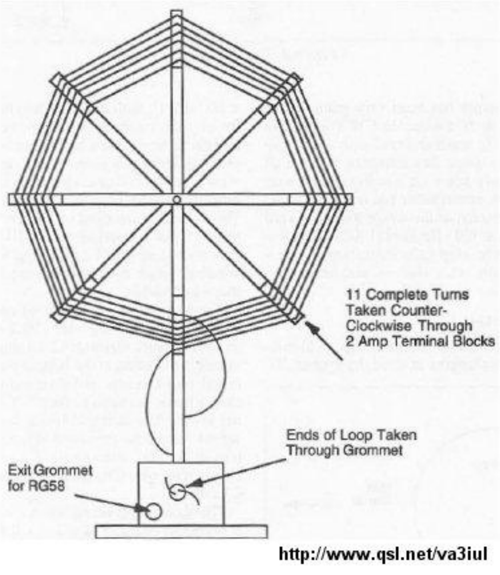
11 Complete Turns
Taken Counter-
Clockwise Through
2 Amp Terminal Blocks
Ends of Loop Taken
Through Grommet
Exit Grommet
for RG58
http://www.qsl.net/va3iul

141 - Hentenna 3dB-Gain for 10m, 6m, 2m

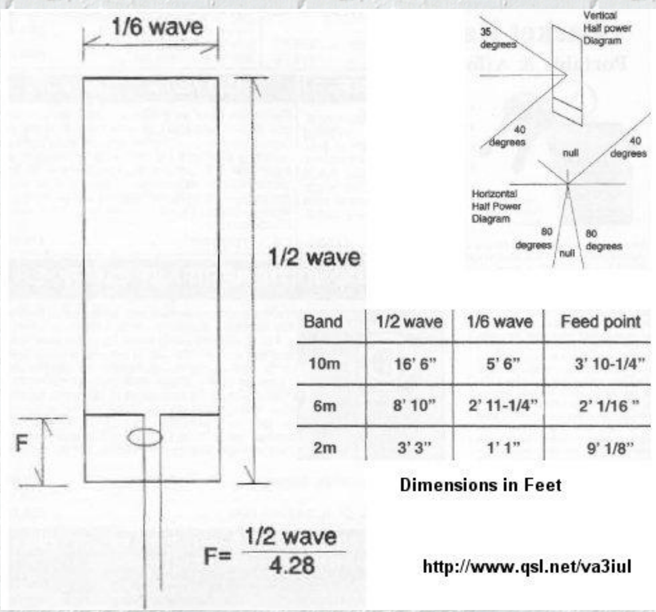
1/6 wave
1/2 wave
F
F= 1/2 wave / 4.28
35 degrees
Vertical
Half power
Diagram
40 degrees
null
40 degrees
Horizontal
Half Power
Diagram
80 degrees
null
80 degrees

Band	1/2 wave	1/6 wave	Feed point
10m	16' 6"	5' 6"	3' 10-1/4"
6m	8' 10"	2' 11-1/4"	2' 1/16 "
2m	3' 3"	1' 1"	9' 1/8"

Dimensions in Feet

http://www.qsl.net/va3iul

142 - VK2AAR Wire Antenna for 20m

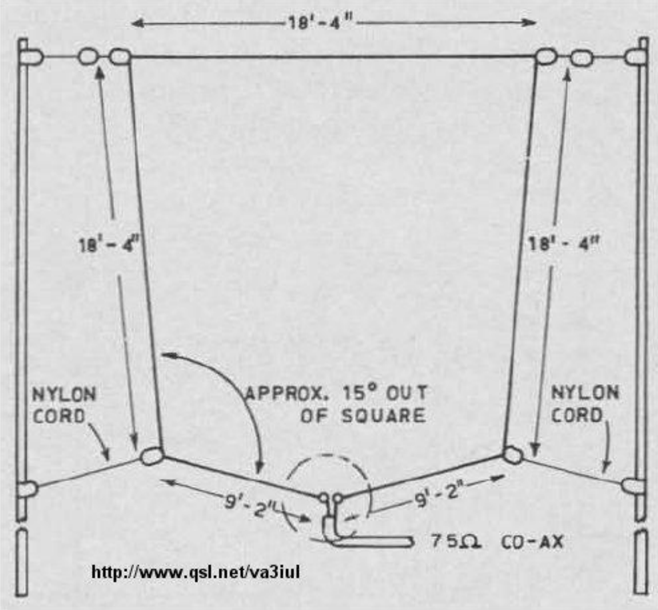
18'-4"
18'-4"
18'-4"
NYLON CORD
NYLON CORD
APPROX. 15° OUT OF SQUARE
9'-2"
9'-2"
75Ω CO-AX
http://www.qsl.net/va3iul

143 - 2-Elements Quad Antenna for 6m

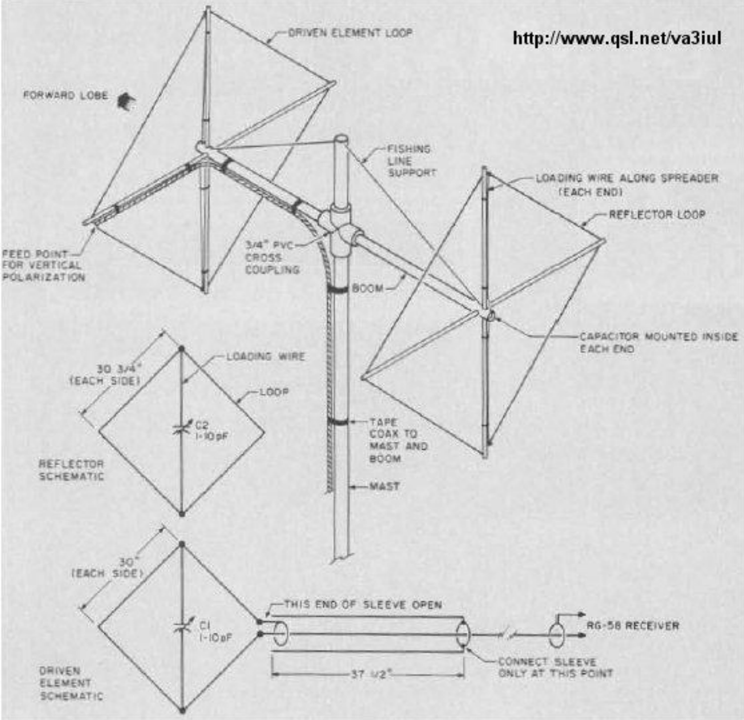
DRIVEN ELEMENT LOOP
http://www.qsl.net/va3iul
FORWARD LOBE
FISHING LINE SUPPORT
LOADING WIRE ALONG SPREADER (EACH END)
REFLECTOR LOOP
FEED POINT FOR VERTICAL POLARIZATION
3/4" PVC CROSS COUPLING
BOOM
CAPACITOR MOUNTED INSIDE EACH END
30 3/4" (EACH SIDE)
LOADING WIRE
LOOP
C2 1-10pF
REFLECTOR SCHEMATIC
TAPE COAX TO MAST AND BOOM
MAST
30" (EACH SIDE)
C1 1-10pF
THIS END OF SLEEVE OPEN
RG-58 RECEIVER
37 1/2"
CONNECT SLEEVE ONLY AT THIS POINT
DRIVEN ELEMENT SCHEMATIC

144 - Hula-Loop Bidirectional 6dB-Gain Antenna for 17m

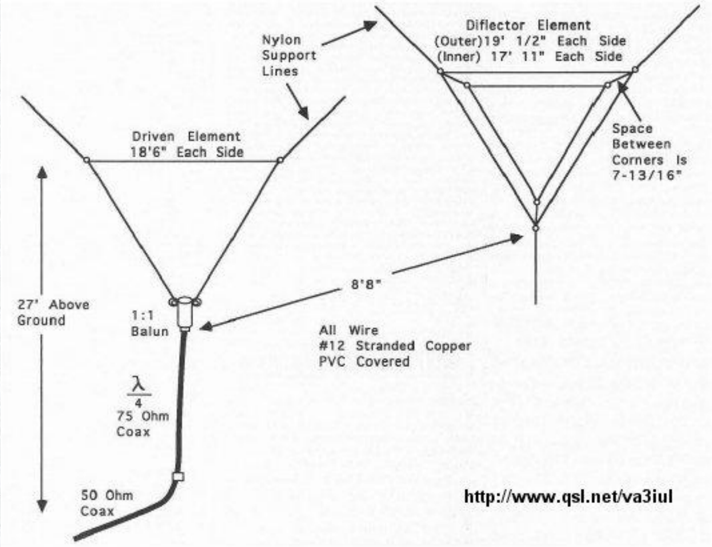
Nylon Support Lines
Diflector Element
(Outer)19' 1/2" Each Side
(Inner) 17' 11" Each Side
Driven Element
18'6" Each Side
Space Between Corners Is 7-13/16"
27' Above Ground
1:1 Balun
8'8"
All Wire #12 Stranded Copper PVC Covered
λ/4 75 Ohm Coax
50 Ohm Coax
http://www.qsl.net/va3iul

145 - Moxon Rectangle Beam for 15m-10m

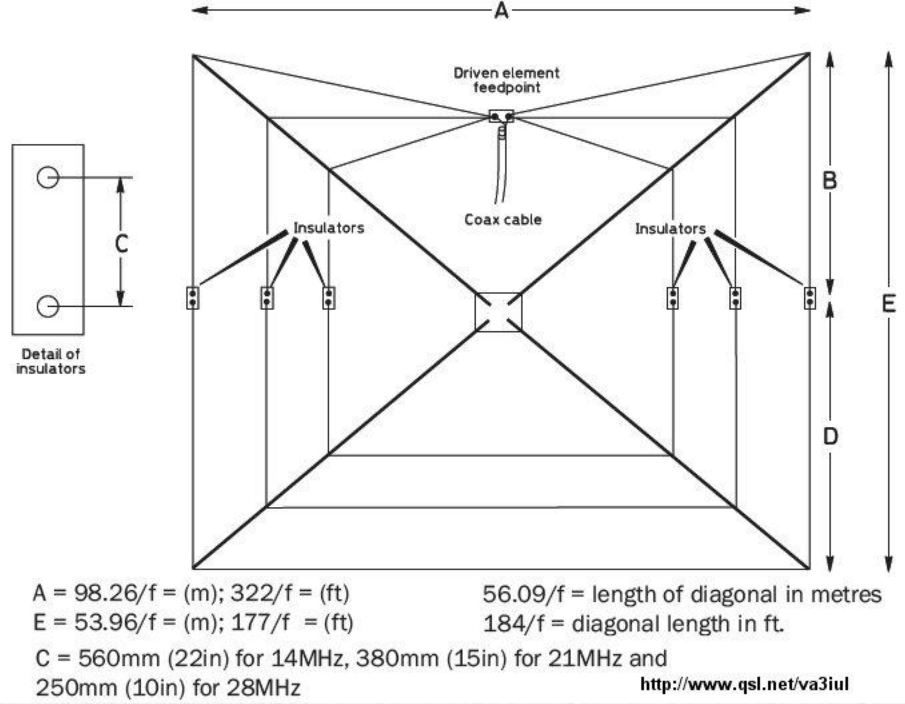
A
Driven element feedpoint
Coax cable
Insulators
Insulators
B
C
D
E
Detail of insulators

A = 98.26/f = (m); 322/f = (ft)
E = 53.96/f = (m); 177/f = (ft)
56.09/f = length of diagonal in metres
184/f = diagonal length in ft.
C = 560mm (22in) for 14MHz, 380mm (15in) for 21MHz and 250mm (10in) for 28MHz

http://www.qsl.net/va3iul

146 - Double-D Beam 4dB-Gain

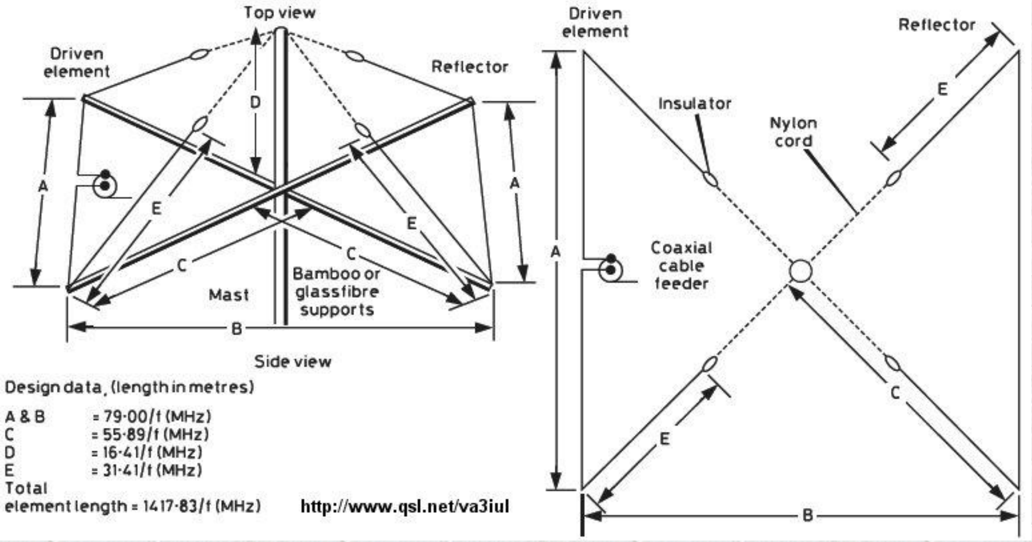
Top view
Driven element
Reflector
D
A
E
C
Mast
Bamboo or glassfibre supports
B
Side view
Driven element
Reflector
Insulator
Nylon cord
Coaxial cable feeder
http://www.qsl.net/va3iul

Design data, (length in metres)

A & B = 79·00/f (MHz)
C = 55·89/f (MHz)
D = 16·41/f (MHz)
E = 31·41/f (MHz)
Total element length = 1417·83/f (MHz)

---

147 - KE4PT OCEF All-Band Dipole

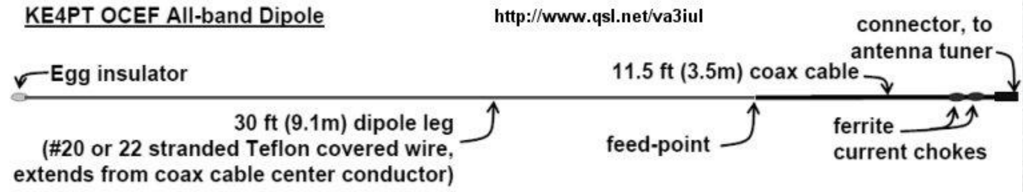
KE4PT OCEF All-band Dipole
http://www.qsl.net/va3iul
Egg insulator
connector, to antenna tuner
11.5 ft (3.5m) coax cable
30 ft (9.1m) dipole leg (#20 or 22 stranded Teflon covered wire, extends from coax cable center conductor)
feed-point
ferrite current chokes

KE4PT Off center end fed (OCEF) All-band Dipole

Lengths are not critical, as long as they are not near a half wavelength on any frequency used.
Dipole radiating length is 30+11.5=41.5 ft, or 12.6 m.
Coax cable may be 50 or 75 ohm on left side of connector, short 50 ohm coax to antenna tuner.
Use antenna tuner for operation on any ham band 40m – 6m.
May be hung by Egg insulator at one end, or draped and drooped over a fiberglass pole.

---

148 - Wire Quad Antenna for 40m

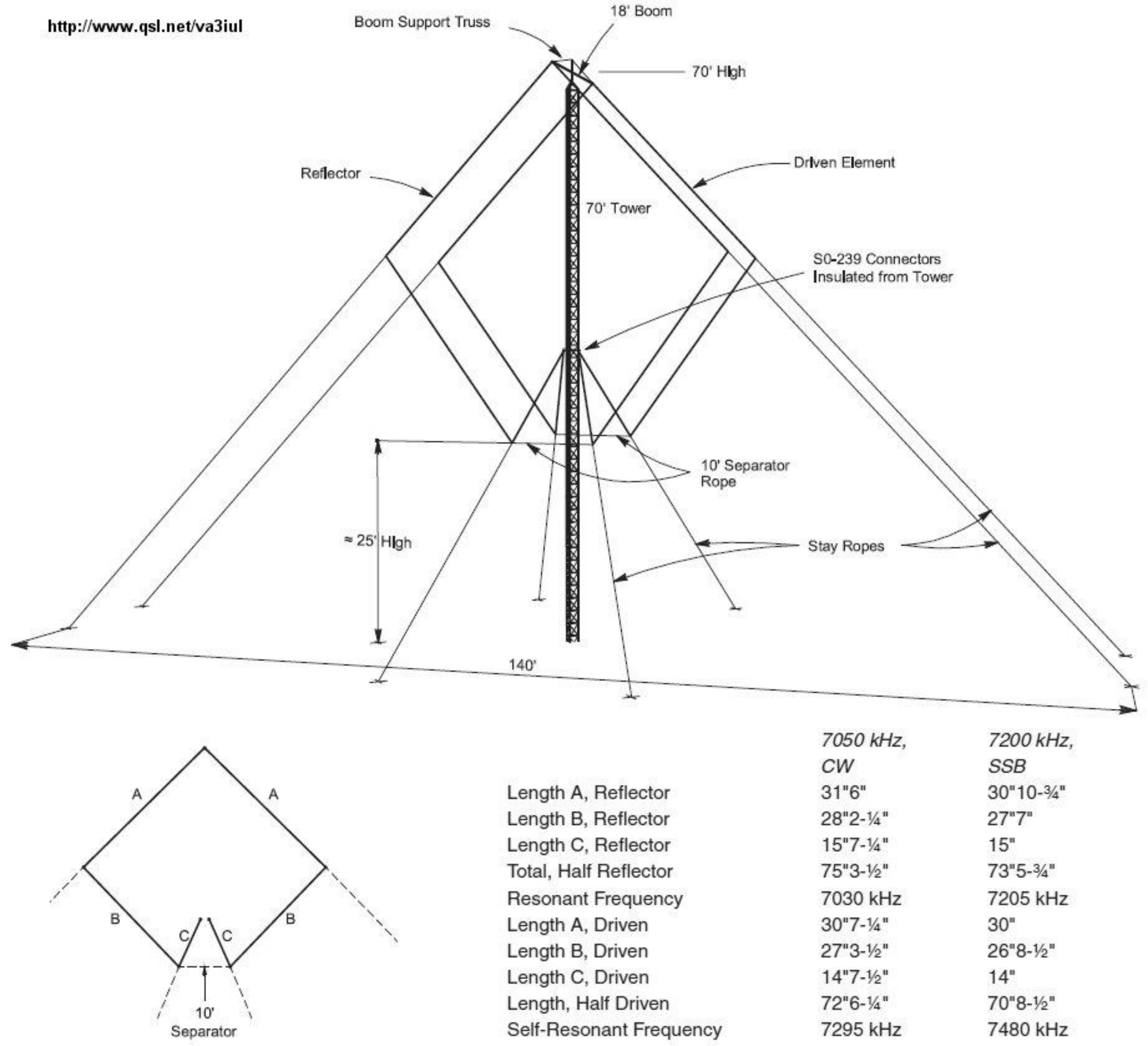
http://www.qsl.net/va3iul
Boom Support Truss
18' Boom
70' High
Reflector
Driven Element
70' Tower
S0-239 Connectors Insulated from Tower
10' Separator Rope
≈ 25' High
Stay Ropes
140'
A
A
B
B
C
C
10' Separator

	7050 kHz, CW	*7200 kHz, SSB*
Length A, Reflector	31"6"	30"10-¾"
Length B, Reflector	28"2-¼"	27"7"
Length C, Reflector	15"7-¼"	15"
Total, Half Reflector	75"3-½"	73"5-¾"
Resonant Frequency	7030 kHz	7205 kHz
Length A, Driven	30"7-¼"	30"
Length B, Driven	27"3-½"	26"8-½"
Length C, Driven	14"7-½"	14"
Length, Half Driven	72"6-¼"	70"8-½"
Self-Resonant Frequency	7295 kHz	7480 kHz

---

149 - Inclined Dipole Antenna for 80m, 40m

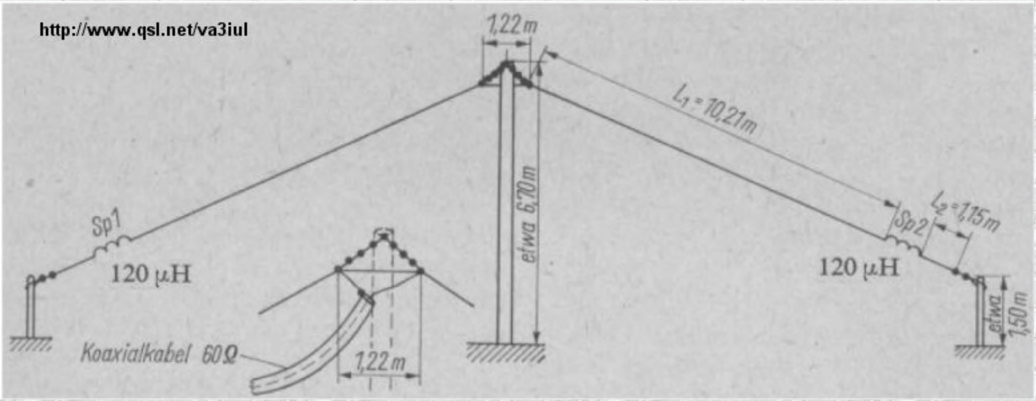
http://www.qsl.net/va3iul
1,22 m
L_1 = 10,21m
Sp1
120 µH
etwa 6,70m
Sp2
L_2 = 1,15m
120 µH
etwa 1,50 m
Koaxialkabel 60Ω
1,22 m

---

150 - Pyramidal Wire Antenna for 80m

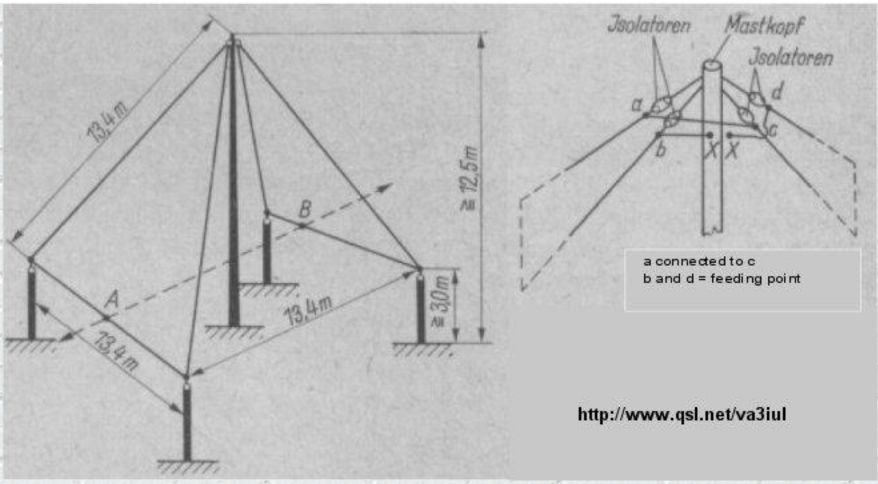
13,4 m
13,4 m
13,4 m
≈ 12,5 m
≈ 3,0 m
A
B
Jsolatoren
Mastkopf
Jsolatoren
a
b
c
d
X
X
a connected to c
b and d = feeding point
http://www.qsl.net/va3iul

151 - Random Wire Antenna All Bands

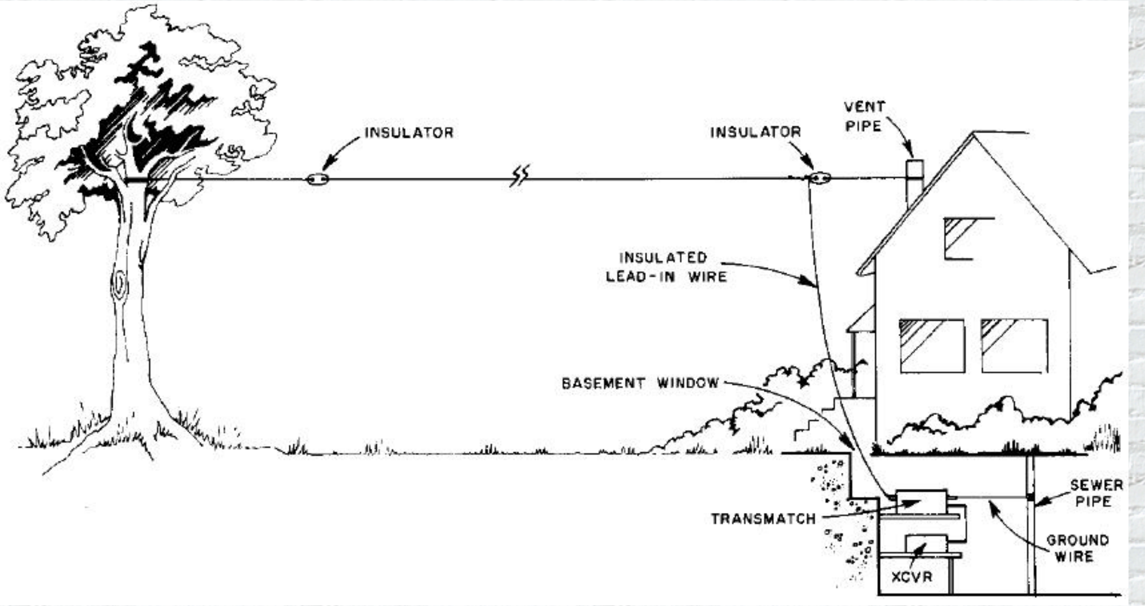
INSULATOR
INSULATOR
VENT PIPE
INSULATED LEAD-IN WIRE
BASEMENT WINDOW
TRANSMATCH
XCVR
SEWER PIPE
GROUND WIRE

152 - Multiband Dipole Antenna for 80m, 40m,15m, 10m

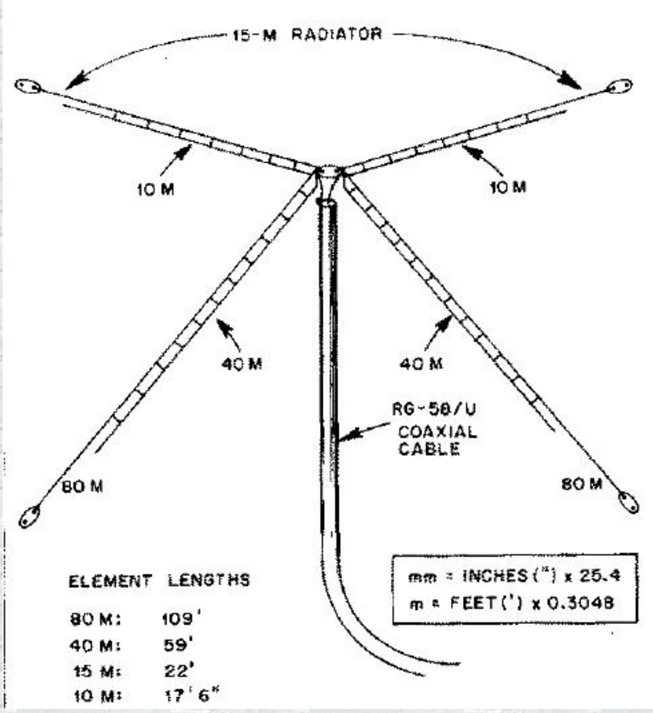
15-M RADIATOR
10 M
10 M
40 M
40 M
RG-58/U COAXIAL CABLE
80 M
80 M
ELEMENT LENGTHS
80 M: 109'
40 M: 59'
15 M: 22'
10 M: 17' 6"
mm = INCHES (") x 25.4
m = FEET (') x 0.3048

153 - Slim Jim Wire Antenna for 4m

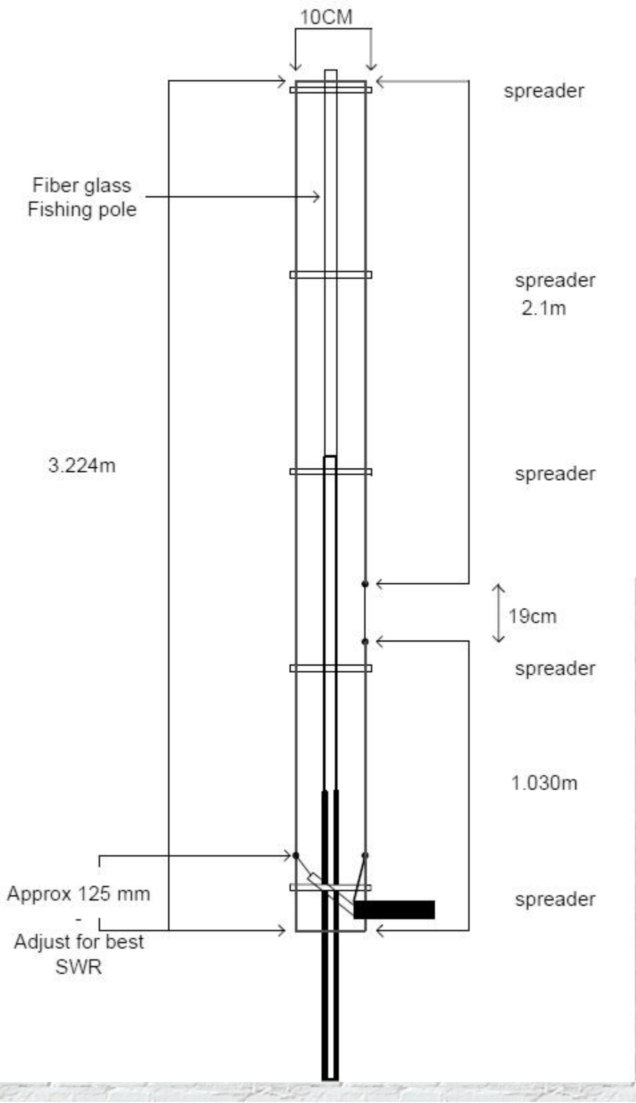
10CM
spreader
Fiber glass Fishing pole
spreader
2.1m
3.224m
spreader
19cm
spreader
1.030m
Approx 125 mm - Adjust for best SWR
spreader

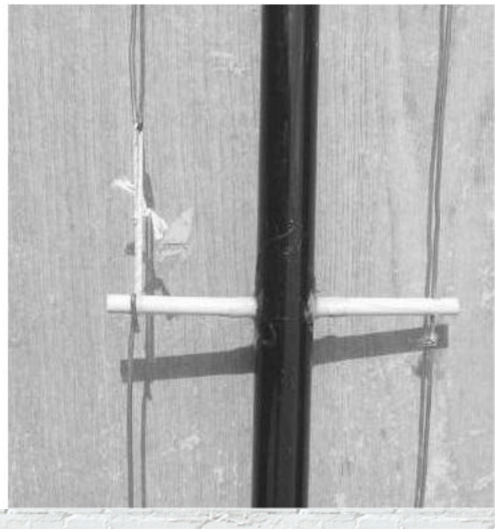

154 - Delta Loop for 6m

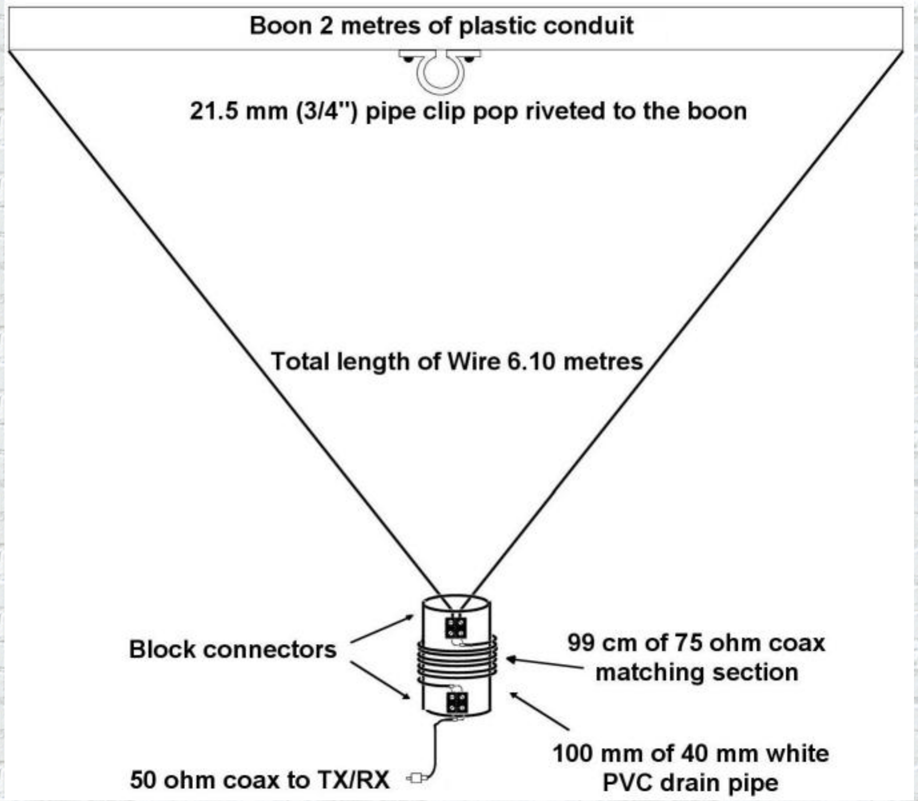
Boon 2 metres of plastic conduit
21.5 mm (3/4") pipe clip pop riveted to the boon
Total length of Wire 6.10 metres
Block connectors
99 cm of 75 ohm coax matching section
50 ohm coax to TX/RX
100 mm of 40 mm white PVC drain pipe

155 - Re-Configurable Antenna for 160m and 80m

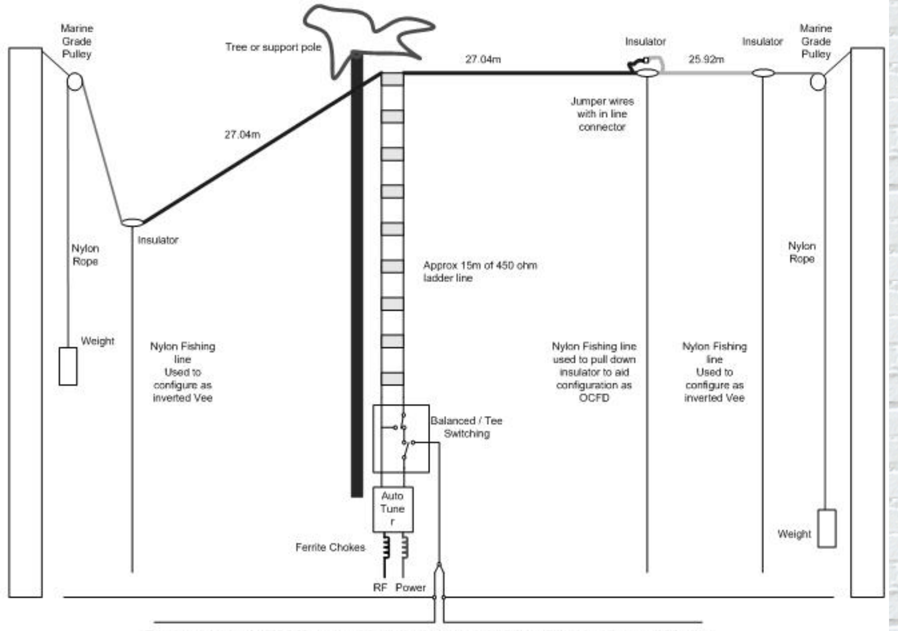
Marine Grade Pulley
Tree or support pole
27.04m
Insulator
25.92m
Insulator
Marine Grade Pulley
Jumper wires with in line connector
27.04m
Insulator
Nylon Rope
Approx 15m of 450 ohm ladder line
Nylon Rope
Weight
Nylon Fishing line Used to configure as inverted Vee
Nylon Fishing line used to pull down insulator to aid configuration as OCFD
Nylon Fishing line Used to configure as inverted Vee
Balanced / Tee Switching
Auto Tune r
Weight
Ferrite Chokes
RF
Power
Tuned ground planes for 160 & 80m approx 0.6 x ¼ Wavelength stapled to fence at 2m height of 2m spaced 0.5m apart

156 - Very Low Frequency Inverted-L Antenna

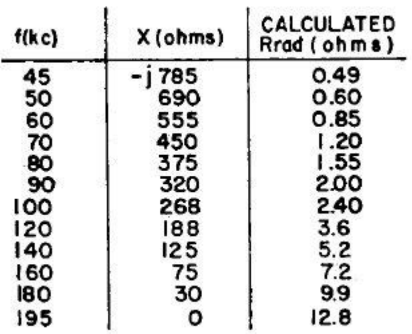

f(kc)	X (ohms)	CALCULATED Rrad (ohms)
45	-j 785	0.49
50	690	0.60
60	555	0.85
70	450	1.20
80	375	1.55
90	320	2.00
100	268	2.40
120	188	3.6
140	125	5.2
160	75	7.2
180	30	9.9
195	0	12.8

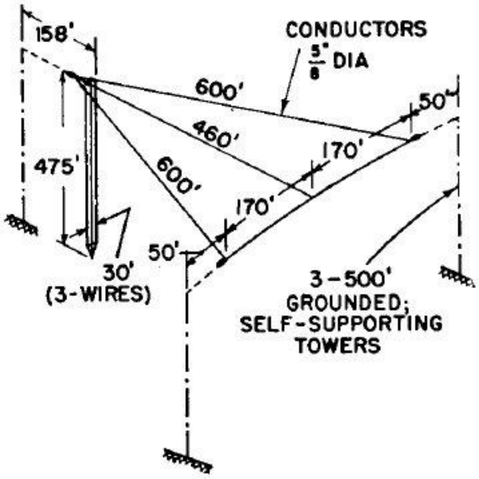
158'
CONDUCTORS
5/8" DIA
600'
460'
600'
50'
170'
170'
475'
30'
(3-WIRES)
50'
3-500' GROUNDED; SELF-SUPPORTING TOWERS

157 - Reduced Size Half Sloper Antenna for 160m

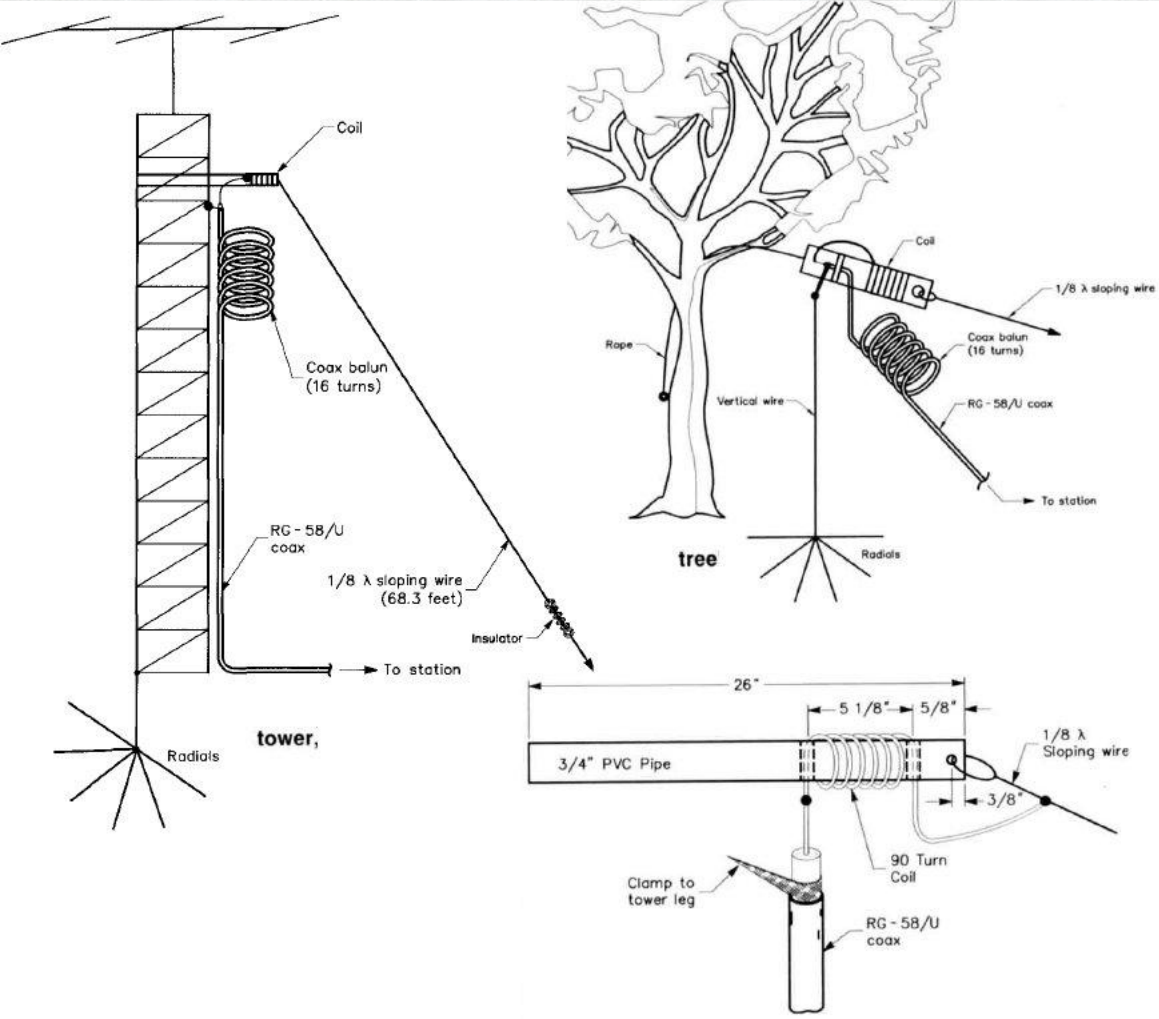
Coil
Coax balun (16 turns)
RG-58/U coax
1/8 λ sloping wire (68.3 feet)
Insulator
To station
Radials
tower,
Coil
1/8 λ sloping wire
Rope
Coax balun (16 turns)
Vertical wire
RG-58/U coax
To station
tree
Radials
26"
5 1/8"
5/8"
3/4" PVC Pipe
1/8 λ Sloping wire
3/8"
90 Turn Coil
Clamp to tower leg
RG-58/U coax

158 - Tree-Mounted HF Antenna

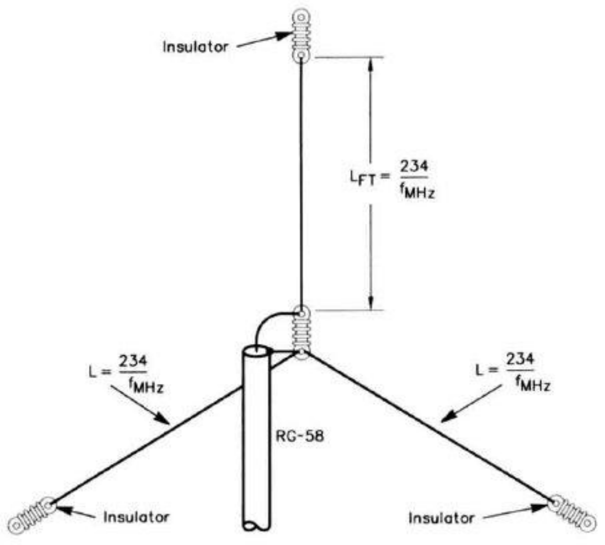
Insulator
$L_{FT} = \frac{234}{f_{MHz}}$
$L = \frac{234}{f_{MHz}}$
$L = \frac{234}{f_{MHz}}$
RG-58
Insulator
Insulator

159 - Multiband Vertical Antenna for 80m, 40m, 20m

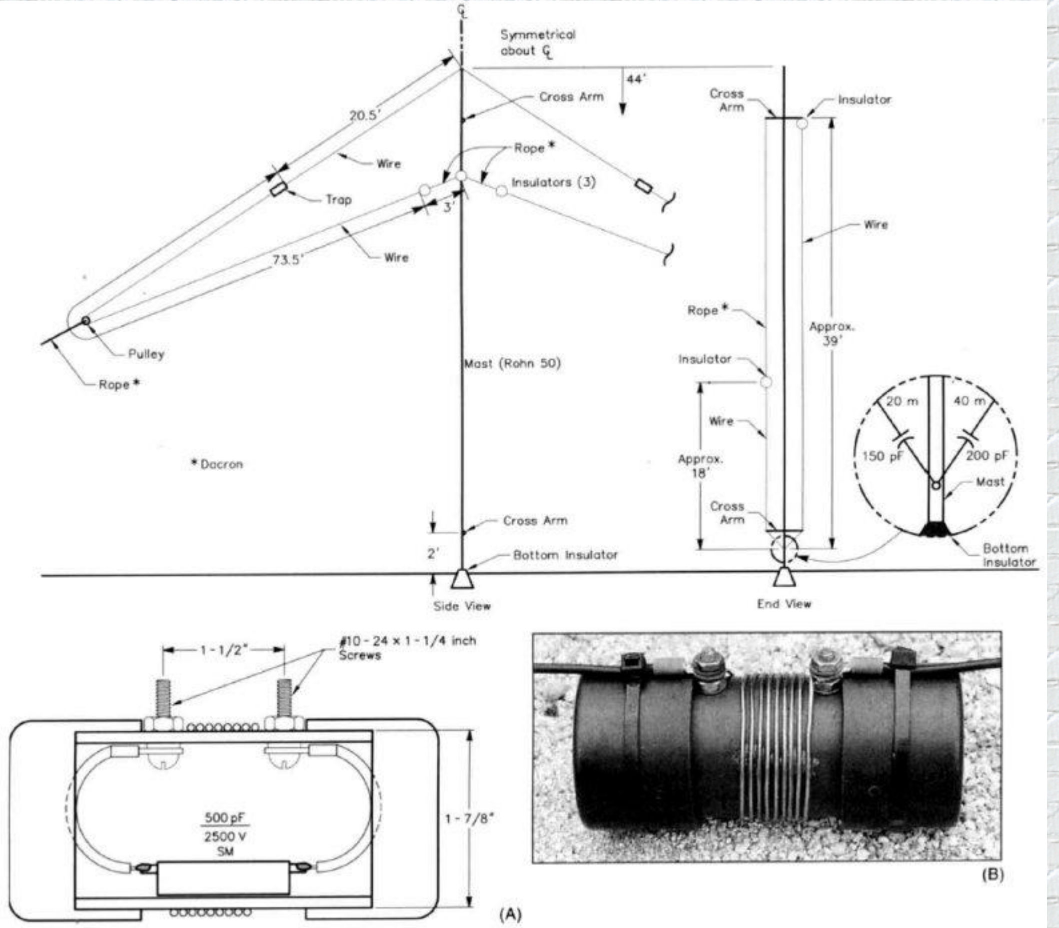
℄
Symmetrical about ℄
44'
Cross Arm
20.5'
Wire
Trap
Rope *
Insulators (3)
3'
73.5'
Wire
Pulley
Rope *
* Dacron
Mast (Rohn 50)
Cross Arm
2'
Bottom Insulator
Side View
Cross Arm
Insulator
Wire
Rope *
Approx. 39'
Insulator
Wire
Approx. 18'
Cross Arm
End View
20 m
40 m
150 pF
200 pF
Mast
Bottom Insulator
1-1/2"
#10-24 x 1-1/4 inch Screws
500 pF 2500 V SM
1-7/8"
(A)
(B)

160 - Marconi Antenna for 136 kHz

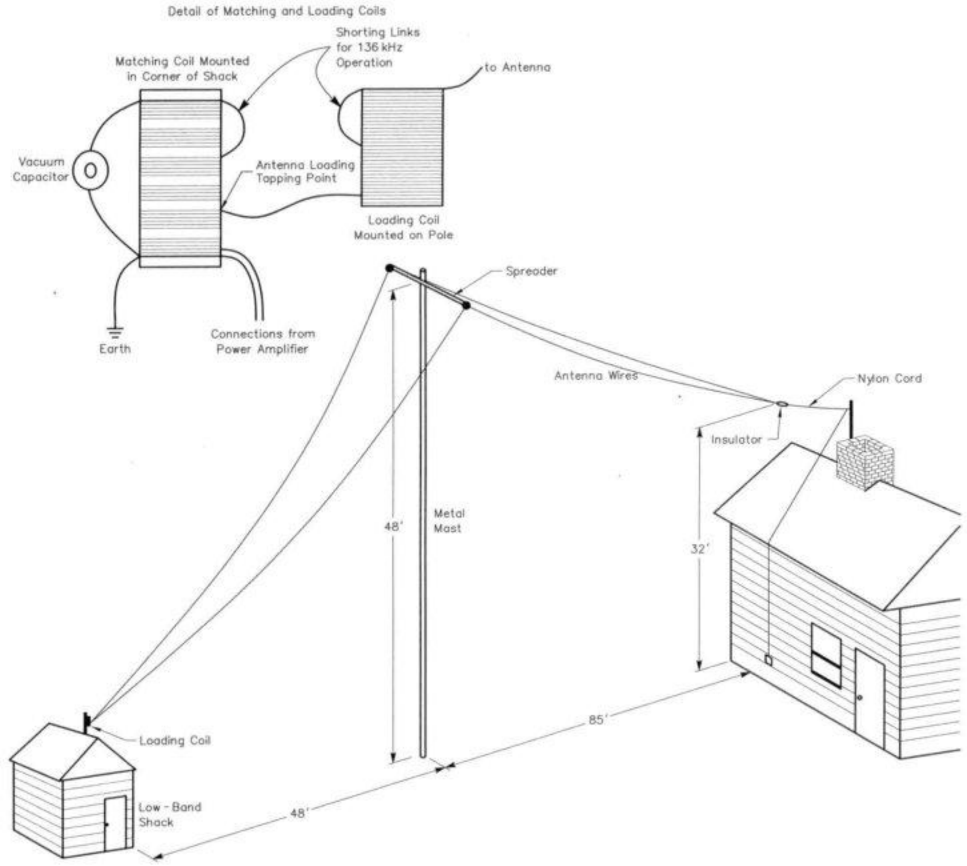
Detail of Matching and Loading Coils
Shorting Links for 136 kHz Operation
Matching Coil Mounted in Corner of Shack
to Antenna
Vacuum Capacitor
Antenna Loading Tapping Point
Loading Coil Mounted on Pole
Spreader
Earth
Connections from Power Amplifier
Antenna Wires
Nylon Cord
Insulator
48'
Metal Mast
32'
Loading Coil
85'
Low - Band Shack
48'

---

161 - Simple Killer Antenna for 40m

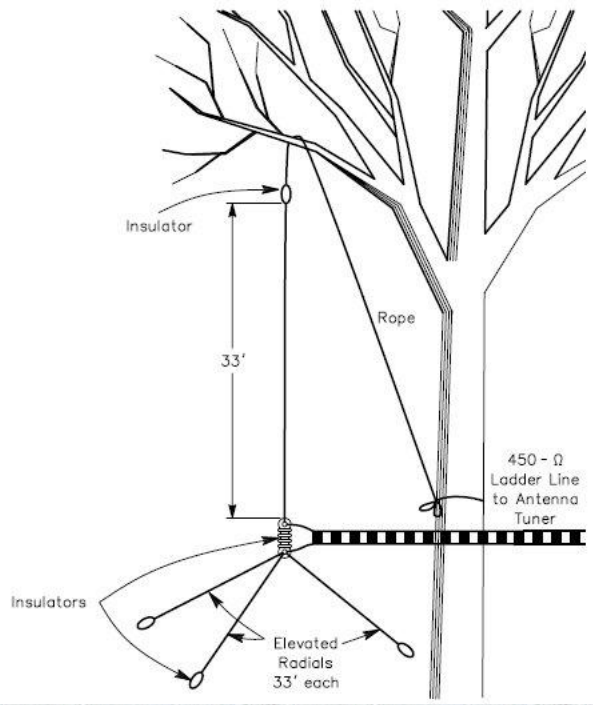
Insulator
33'
Rope
450 - Ω Ladder Line to Antenna Tuner
Insulators
Elevated Radials 33' each

---

162 - Stub-Directed V Antenna for 80m

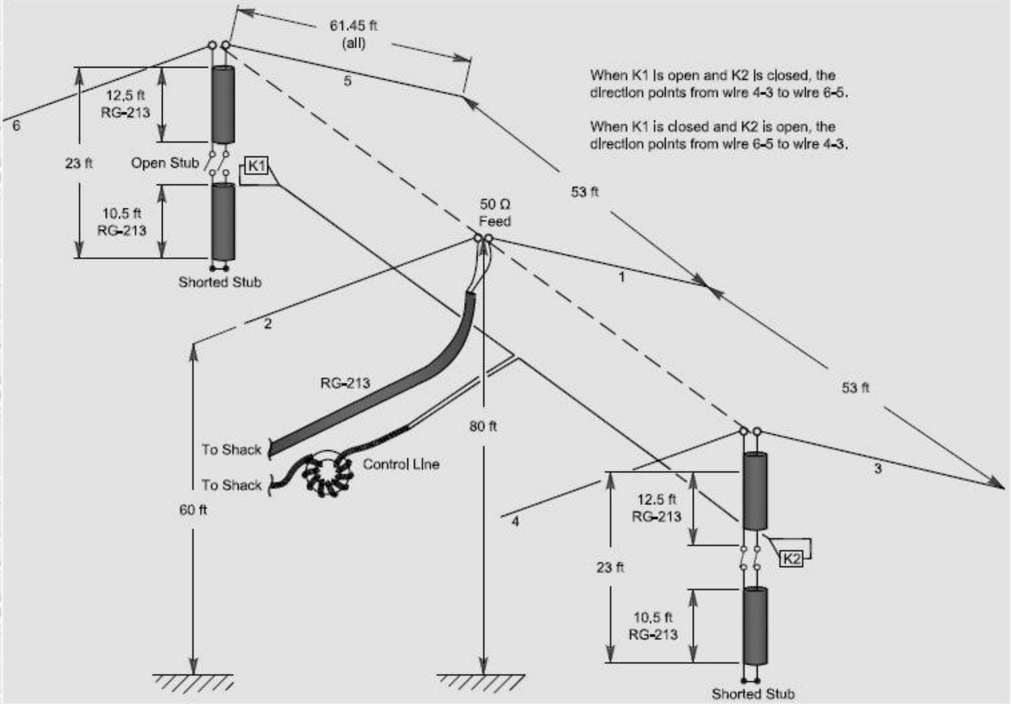
61.45 ft
(all)
6
5
12.5 ft
RG-213
23 ft
Open Stub
K1
10.5 ft
RG-213
Shorted Stub
50 Ω
Feed
53 ft
1
2
RG-213
80 ft
To Shack
Control Line
To Shack
60 ft
53 ft
3
4
12.5 ft
RG-213
23 ft
K2
10.5 ft
RG-213
Shorted Stub

When K1 is open and K2 is closed, the direction points from wire 4-3 to wire 6-5.

When K1 is closed and K2 is open, the direction points from wire 6-5 to wire 4-3.

163 - KT0NY Over-and-Under DX Antenna for 20m

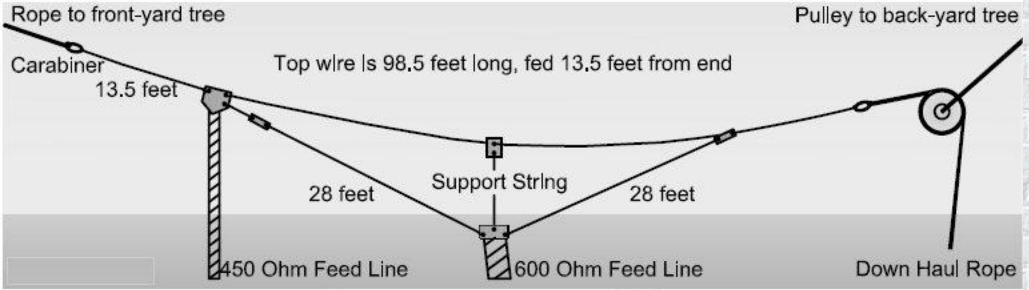
Rope to front-yard tree
Pulley to back-yard tree
Carabiner
13.5 feet
Top wire is 98.5 feet long, fed 13.5 feet from end
28 feet
Support String
28 feet
450 Ohm Feed Line
600 Ohm Feed Line
Down Haul Rope

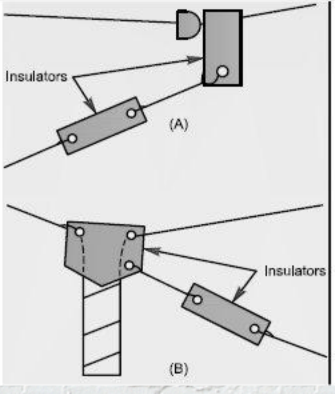
Insulators
(A)
Insulators
(B)

164 - Horizontal Loop Antennas

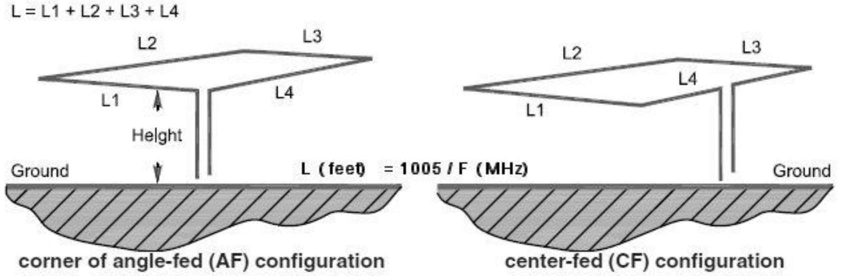
L = L1 + L2 + L3 + L4
L2
L3
L1
L4
Height
Ground
L (feet) = 1005 / F (MHz)
L2
L3
L4
L1
Ground
corner of angle-fed (AF) configuration
center-fed (CF) configuration

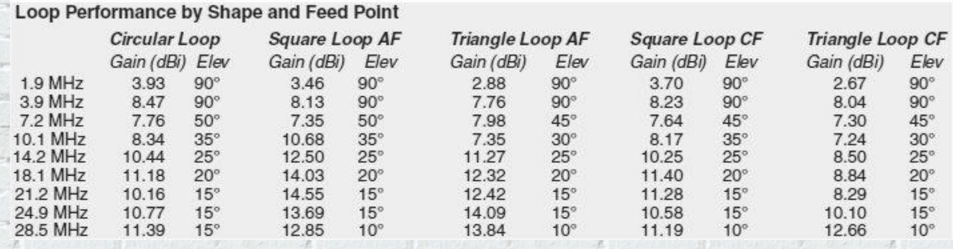

Loop Performance by Shape and Feed Point

	Circular Loop		*Square Loop AF*		*Triangle Loop AF*		*Square Loop CF*		*Triangle Loop CF*	
	Gain (dBi)	*Elev*	*Gain (dBi)*	*Elev*	*Gain (dBi)*	*Elev*	*Gain (dBi)*	*Elev*	*Gain (dBi)*	*Elev*
1.9 MHz	3.93	90°	3.46	90°	2.88	90°	3.70	90°	2.67	90°
3.9 MHz	8.47	90°	8.13	90°	7.76	90°	8.23	90°	8.04	90°
7.2 MHz	7.76	50°	7.35	50°	7.98	45°	7.64	45°	7.30	45°
10.1 MHz	8.34	35°	10.68	35°	7.35	30°	8.17	35°	7.24	30°
14.2 MHz	10.44	25°	12.50	25°	11.27	25°	10.25	25°	8.50	25°
18.1 MHz	11.18	20°	14.03	20°	12.32	20°	11.40	20°	8.84	20°
21.2 MHz	10.16	15°	14.55	15°	12.42	15°	11.28	15°	8.29	15°
24.9 MHz	10.77	15°	13.69	15°	14.09	15°	10.58	15°	10.10	15°
28.5 MHz	11.39	15°	12.85	10°	13.84	10°	11.19	10°	12.66	10°

---

165 - Ribbon J-Pole for 2m

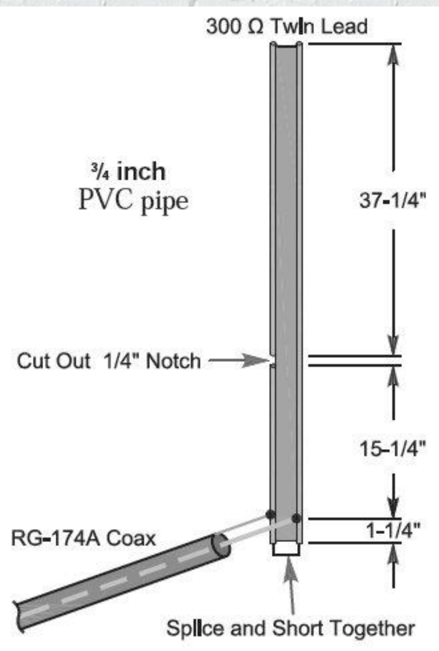
300 Ω Twin Lead
¾ inch
PVC pipe
37-1/4"
Cut Out 1/4" Notch
15-1/4"
1-1/4"
RG-174A Coax
Splice and Short Together

---

166 - Dual Band Ribbon J-Pole for 2m, 70cm

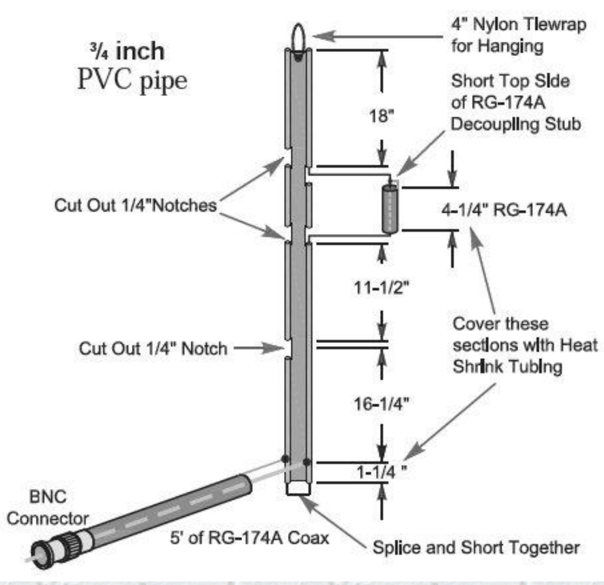
¾ inch
PVC pipe
4" Nylon Tiewrap for Hanging
Short Top Side of RG-174A Decoupling Stub
18"
Cut Out 1/4"Notches
4-1/4" RG-174A
11-1/2"
Cut Out 1/4" Notch
Cover these sections with Heat Shrink Tubing
16-1/4"
1-1/4 "
BNC Connector
5' of RG-174A Coax
Splice and Short Together

167 - Square Vertical Loop Antenna for 40m

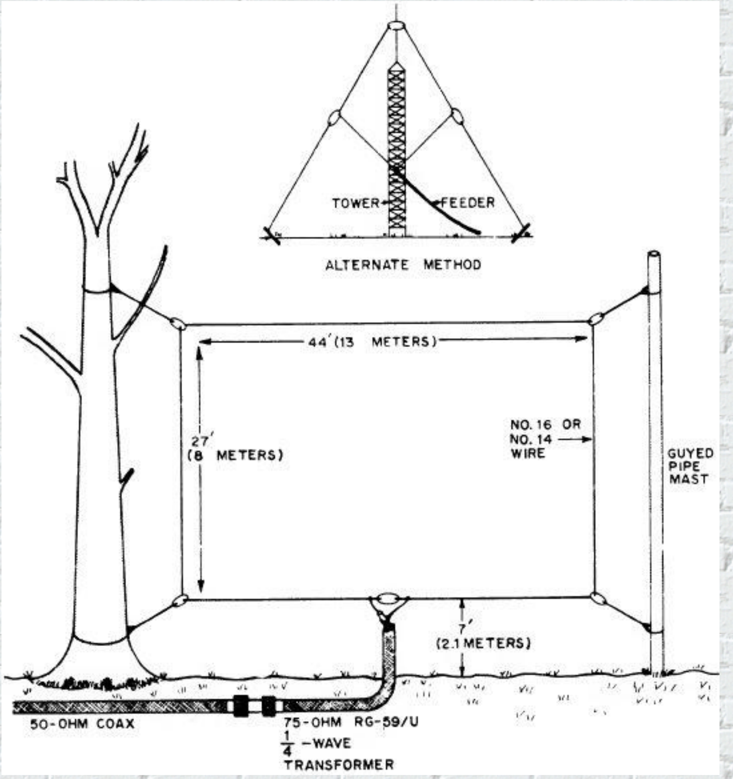
TOWER
FEEDER
ALTERNATE METHOD
44' (13 METERS)
27' (8 METERS)
NO. 16 OR NO. 14 WIRE
GUYED PIPE MAST
7' (2.1 METERS)
50-OHM COAX
75-OHM RG-59/U
$\frac{1}{4}$-WAVE TRANSFORMER

168 - Tri-Band Quad Antenna for 20m, 15m, 10m

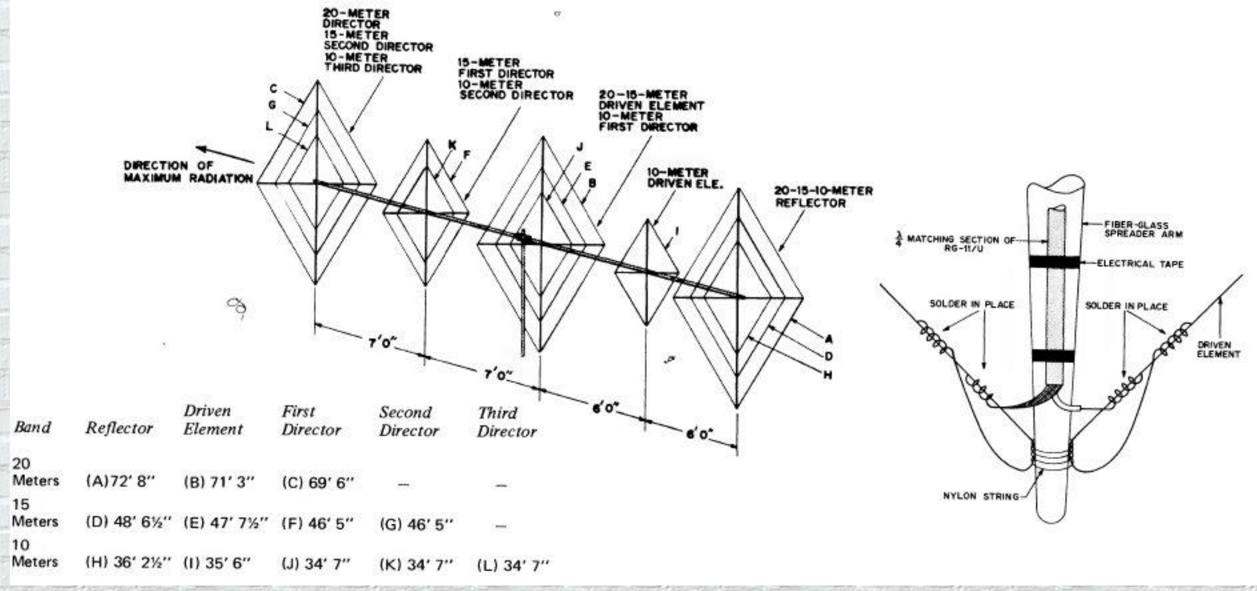
20-METER DIRECTOR
15-METER SECOND DIRECTOR
10-METER THIRD DIRECTOR
15-METER FIRST DIRECTOR
10-METER SECOND DIRECTOR
20-15-METER DRIVEN ELEMENT
10-METER FIRST DIRECTOR
10-METER DRIVEN ELE.
20-15-10-METER REFLECTOR
DIRECTION OF MAXIMUM RADIATION
C
G
L
K
F
J
E
B
I
A
D
H
7'0"
7'0"
6'0"
6'0"
¼ MATCHING SECTION OF RG-11/U
FIBER-GLASS SPREADER ARM
ELECTRICAL TAPE
SOLDER IN PLACE
SOLDER IN PLACE
DRIVEN ELEMENT
NYLON STRING

Band	*Reflector*	*Driven Element*	*First Director*	*Second Director*	*Third Director*
20 Meters	(A)72' 8"	(B) 71' 3"	(C) 69' 6"	–	–
15 Meters	(D) 48' 6½"	(E) 47' 7½"	(F) 46' 5"	(G) 46' 5"	–
10 Meters	(H) 36' 2½"	(I) 35' 6"	(J) 34' 7"	(K) 34' 7"	(L) 34' 7"

169 - 3D Quad Antenna for 80m, 40m, 20m, 15m, 10m

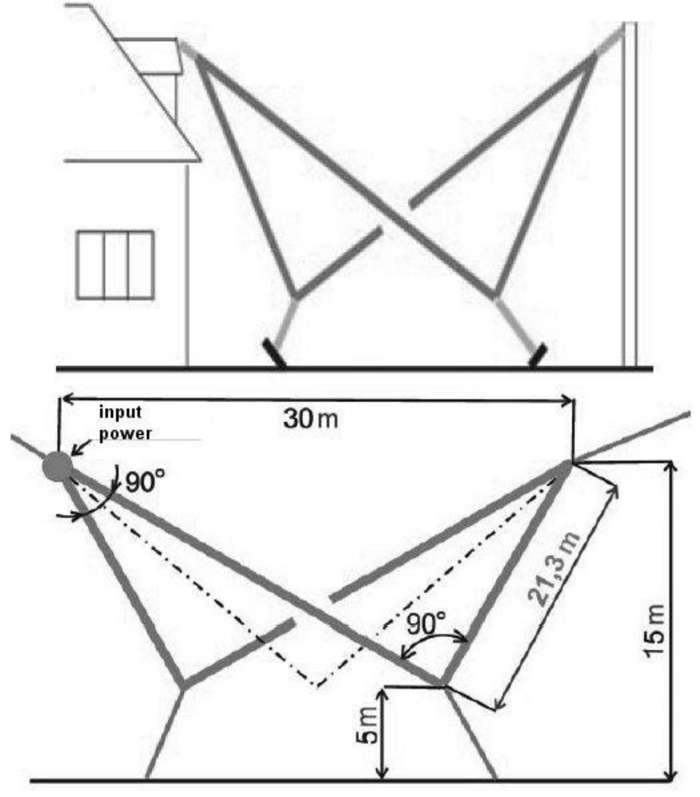
input power
30 m
90°
21,3 m
90°
15 m
5 m

170 - Sloping Wire Antenna for 30m, 20m, 17m, 15m, 12m, 10m

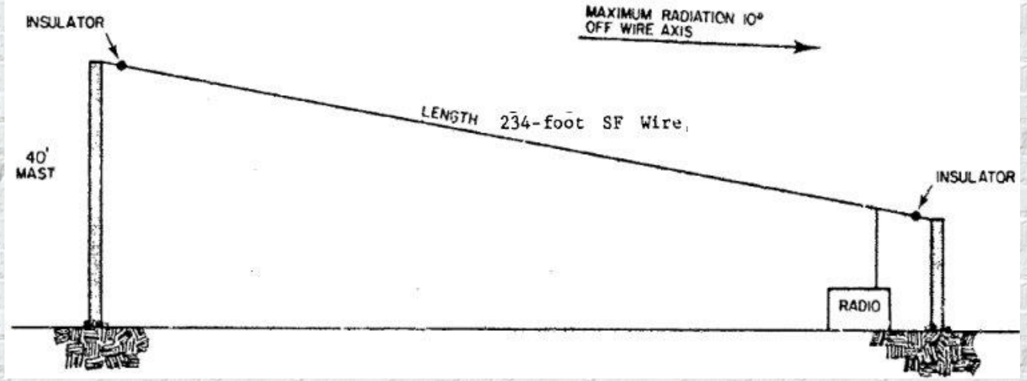
INSULATOR
MAXIMUM RADIATION 10° OFF WIRE AXIS
LENGTH 234-foot SF Wire,
40' MAST
INSULATOR
RADIO

171 - Broadband Dipole Antenna with Coaxial Resonator for 80m

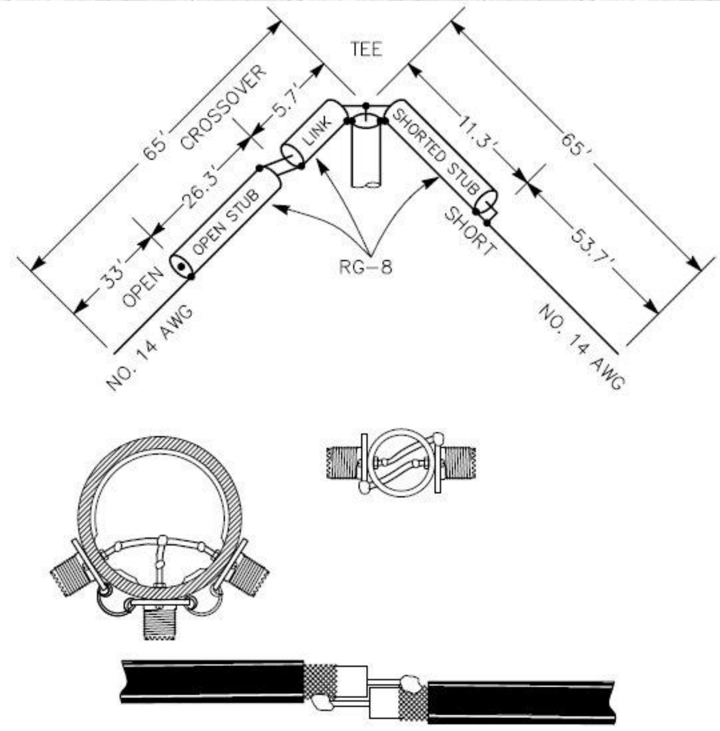
TEE
65' CROSSOVER
5.7'
LINK
26.3'
OPEN STUB
33'
OPEN
NO. 14 AWG
RG-8
SHORTED STUB
11.3'
65'
SHORT
53.7'
NO. 14 AWG

172 - Resonant Feedline Dipole Antenna for 80m

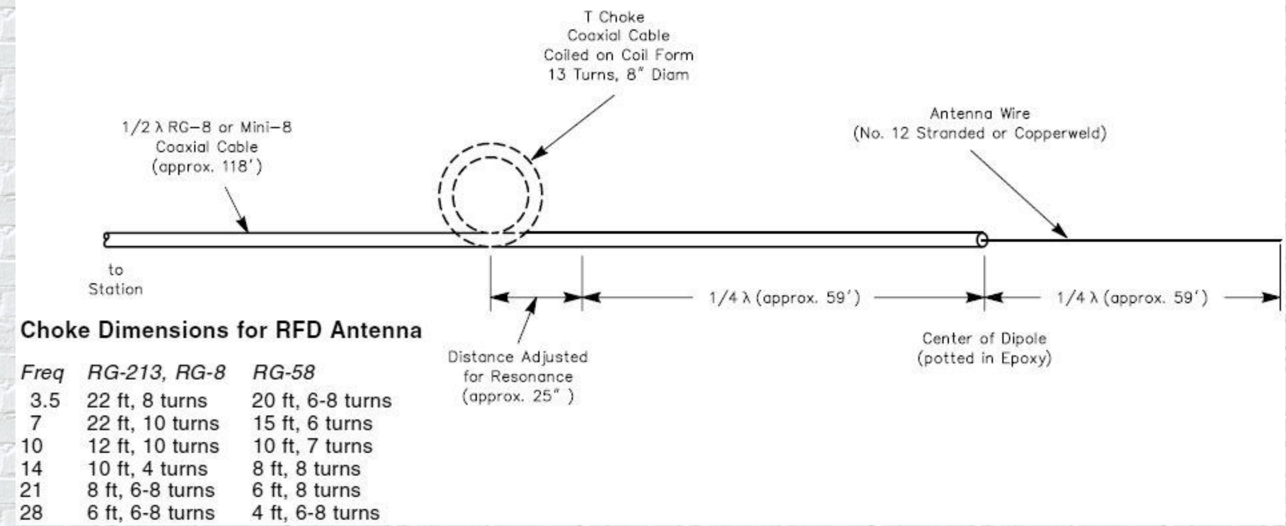
T Choke
Coaxial Cable
Coiled on Coil Form
13 Turns, 8" Diam
1/2 λ RG-8 or Mini-8
Coaxial Cable
(approx. 118')
Antenna Wire
(No. 12 Stranded or Copperweld)
to
Station
1/4 λ (approx. 59')
1/4 λ (approx. 59')
Distance Adjusted
for Resonance
(approx. 25")
Center of Dipole
(potted in Epoxy)

Choke Dimensions for RFD Antenna

Freq	*RG-213, RG-8*	*RG-58*
3.5	22 ft, 8 turns	20 ft, 6-8 turns
7	22 ft, 10 turns	15 ft, 6 turns
10	12 ft, 10 turns	10 ft, 7 turns
14	10 ft, 4 turns	8 ft, 8 turns
21	8 ft, 6-8 turns	6 ft, 8 turns
28	6 ft, 6-8 turns	4 ft, 6-8 turns

173 - Dual-Band Loading Wire Antenna for 80m, 30m

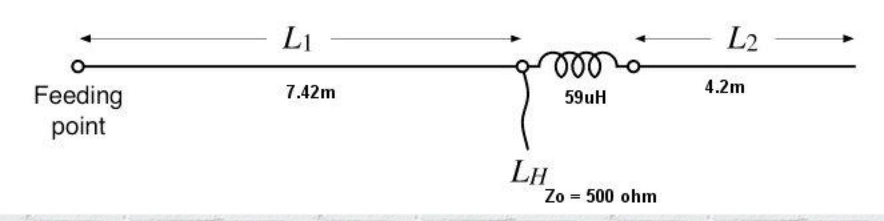
L_1
L_2
Feeding point
7.42m
59uH
4.2m
L_H
Zo = 500 ohm

---

174 - Stub Matching Antennas

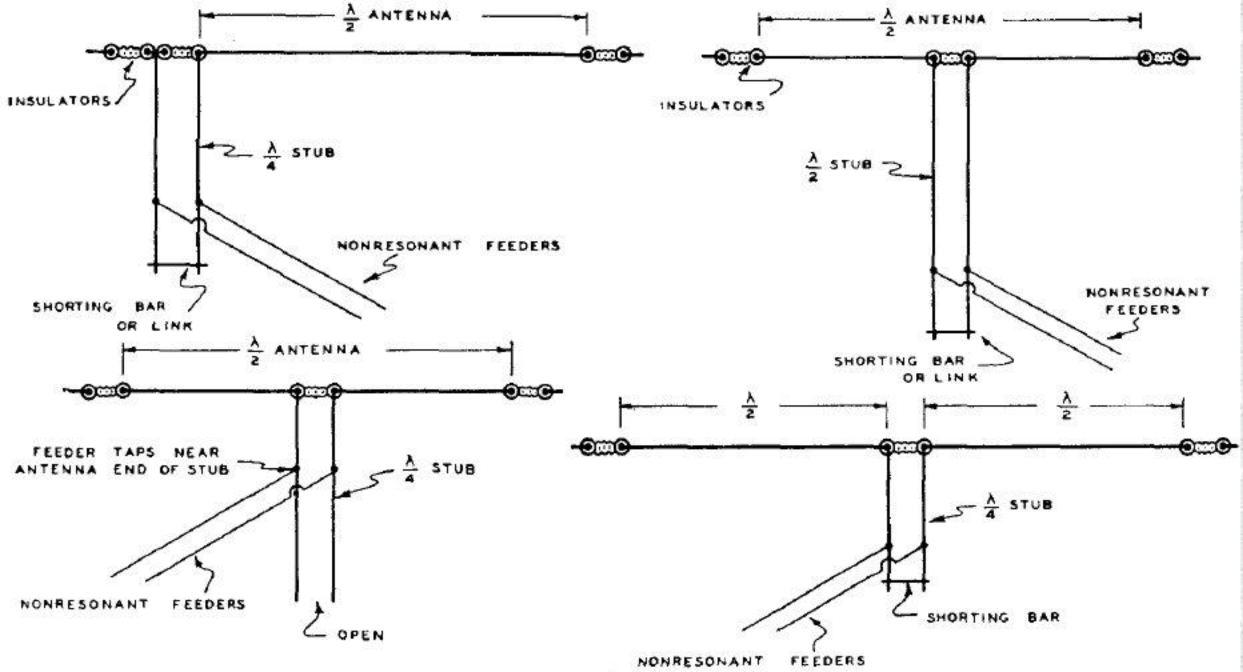
λ/2 ANTENNA
INSULATORS
λ/4 STUB
NONRESONANT FEEDERS
SHORTING BAR OR LINK
λ/2 ANTENNA
INSULATORS
λ/2 STUB
NONRESONANT FEEDERS
SHORTING BAR OR LINK
λ/2 ANTENNA
FEEDER TAPS NEAR ANTENNA END OF STUB
λ/4 STUB
NONRESONANT FEEDERS
OPEN
λ/2
λ/2
λ/4 STUB
SHORTING BAR
NONRESONANT FEEDERS

---

175 - J-Style Vertical Wire Antenna for 10m

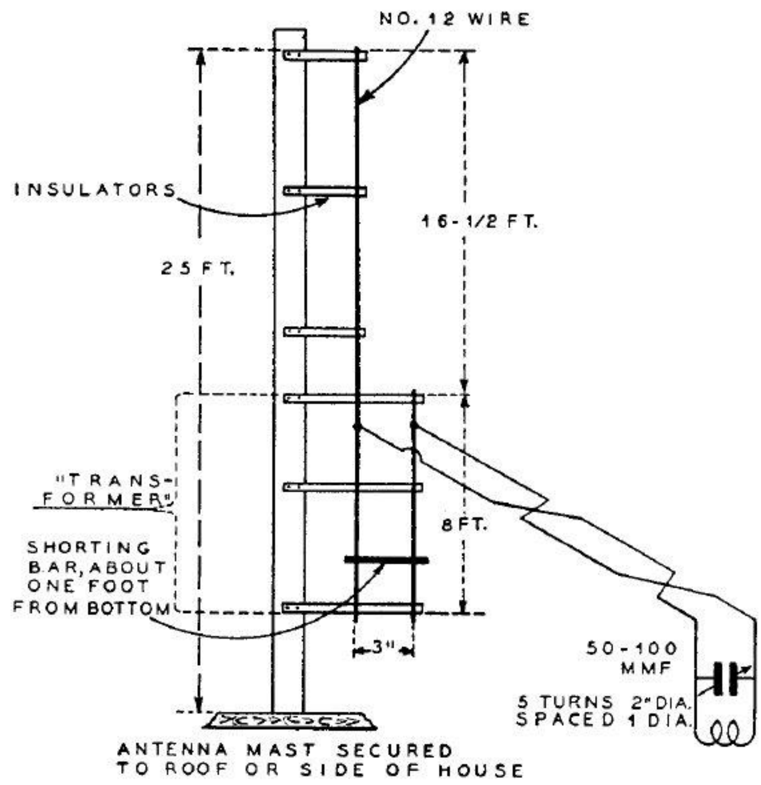
NO. 12 WIRE
INSULATORS
25 FT.
16-1/2 FT.
"TRANS-FORMER"
8 FT.
SHORTING BAR, ABOUT ONE FOOT FROM BOTTOM
3"
50-100 MMF
5 TURNS 2" DIA. SPACED 1 DIA.
ANTENNA MAST SECURED TO ROOF OR SIDE OF HOUSE

176 - Dual Band Vertical with Zepp Feeders for 40m, 20m

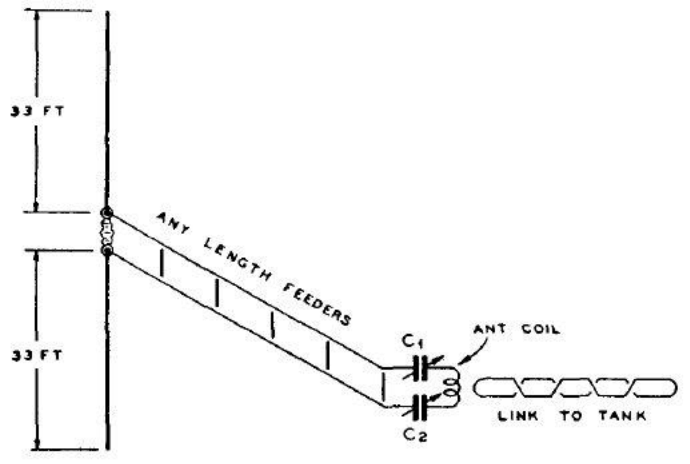
33 FT
33 FT
ANY LENGTH FEEDERS
C1
C2
ANT COIL
LINK TO TANK

177 - RCA Double Doublet for 40m to 12m

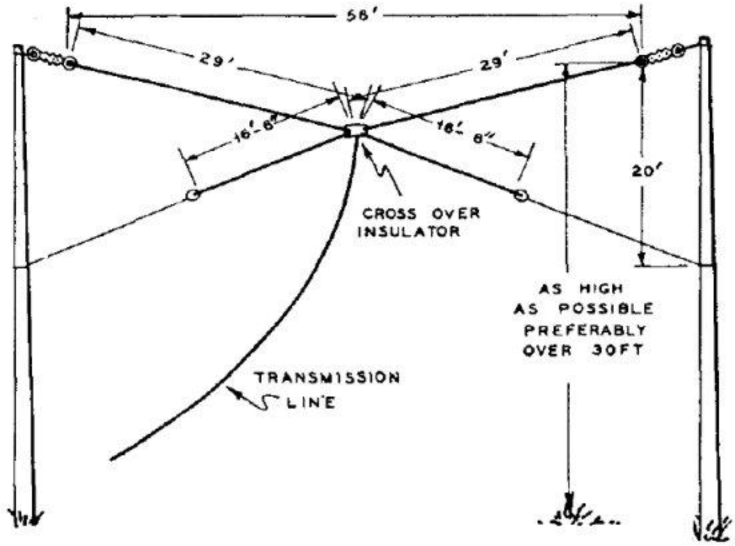
58'
29'
29'
16'-8"
18'- 8"
20'
CROSS OVER
INSULATOR
AS HIGH
AS POSSIBLE
PREFERABLY
OVER 30FT
TRANSMISSION
LINE

178 - RCA Spiderweb Antenna for 40m to 6m

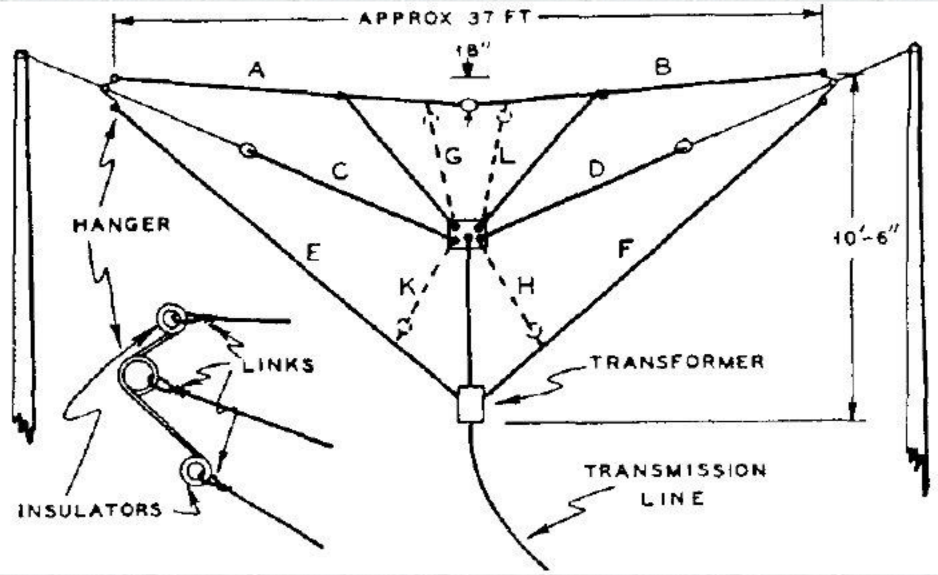
APPROX 37 FT
18"
A
B
C
D
G
L
E
F
K
H
HANGER
LINKS
INSULATORS
TRANSFORMER
10'-6"
TRANSMISSION
LINE

179 - Folded Dipole with Shorted Straps

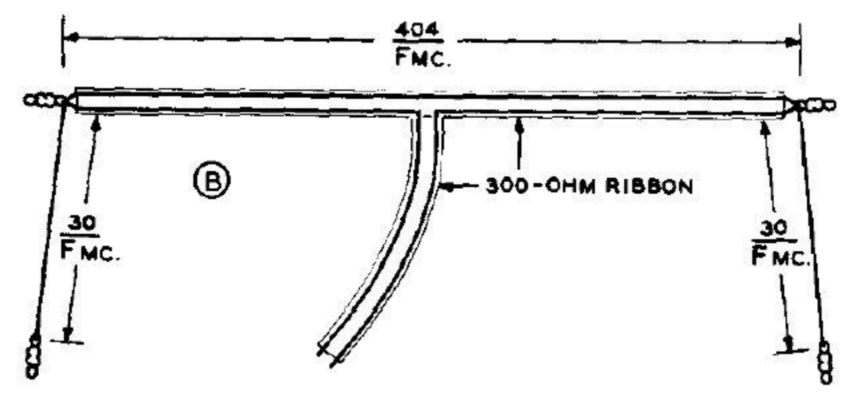
404
FMC.
B
300-OHM RIBBON
30
F MC.
30
F MC.

180 - Twin-Lead Marconi Antenna for 160m, 80m

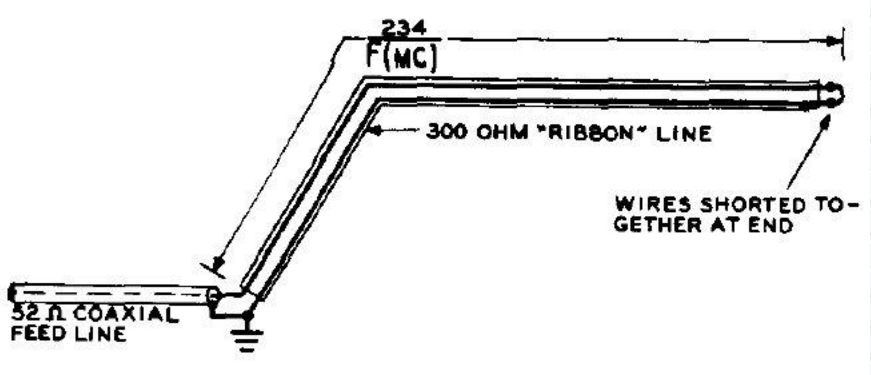
234
F(MC)
300 OHM "RIBBON" LINE
WIRES SHORTED TO-GETHER AT END
52 Ω COAXIAL FEED LINE

---

181 - Broadband Antenna Quarter-Wave Balun for 80m

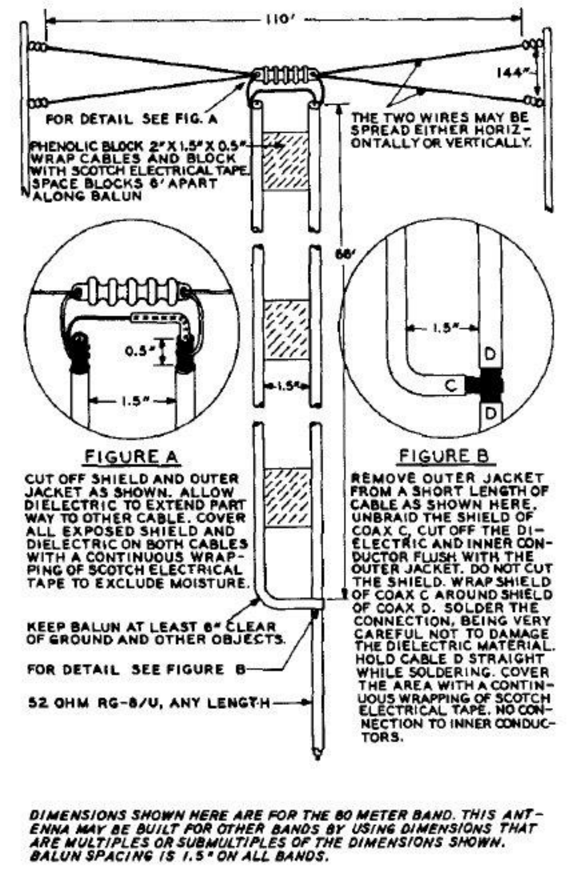
110'
144"
FOR DETAIL SEE FIG. A
PHENOLIC BLOCK 2" X 1.5" X 0.5" WRAP CABLES AND BLOCK WITH SCOTCH ELECTRICAL TAPE. SPACE BLOCKS 6' APART ALONG BALUN
THE TWO WIRES MAY BE SPREAD EITHER HORIZONTALLY OR VERTICALLY.
66'
0.5"
1.5"
1.5"
1.5"
C
D
D
KEEP BALUN AT LEAST 6" CLEAR OF GROUND AND OTHER OBJECTS.
FOR DETAIL SEE FIGURE B
52 OHM RG-8/U, ANY LENGTH

FIGURE A

CUT OFF SHIELD AND OUTER JACKET AS SHOWN. ALLOW DIELECTRIC TO EXTEND PART WAY TO OTHER CABLE. COVER ALL EXPOSED SHIELD AND DIELECTRIC ON BOTH CABLES WITH A CONTINUOUS WRAPPING OF SCOTCH ELECTRICAL TAPE TO EXCLUDE MOISTURE.

FIGURE B

REMOVE OUTER JACKET FROM A SHORT LENGTH OF CABLE AS SHOWN HERE. UNBRAID THE SHIELD OF COAX C, CUT OFF THE DIELECTRIC AND INNER CONDUCTOR FLUSH WITH THE OUTER JACKET. DO NOT CUT THE SHIELD. WRAP SHIELD OF COAX C AROUND SHIELD OF COAX D. SOLDER THE CONNECTION, BEING VERY CAREFUL NOT TO DAMAGE THE DIELECTRIC MATERIAL. HOLD CABLE D STRAIGHT WHILE SOLDERING. COVER THE AREA WITH A CONTINUOUS WRAPPING OF SCOTCH ELECTRICAL TAPE. NO CONNECTION TO INNER CONDUCTORS.

DIMENSIONS SHOWN HERE ARE FOR THE 80 METER BAND. THIS ANTENNA MAY BE BUILT FOR OTHER BANDS BY USING DIMENSIONS THAT ARE MULTIPLES OR SUBMULTIPLES OF THE DIMENSIONS SHOWN. BALUN SPACING IS 1.5" ON ALL BANDS.

---

182 - Three-Quarter-Wave Folded Doublet for Dual-Band

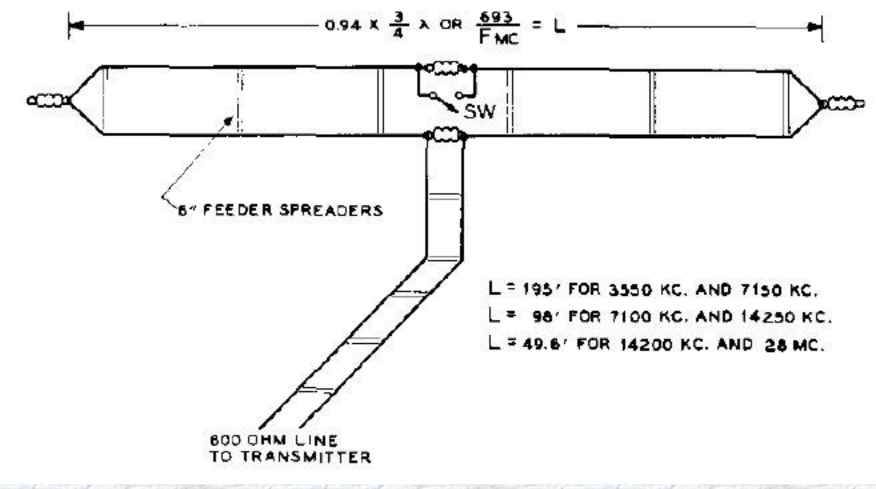
0.94 X $\frac{3}{4}$ λ OR $\frac{693}{F_{MC}}$ = L
SW
6" FEEDER SPREADERS
L = 195' FOR 3550 KC. AND 7150 KC.
L = 98' FOR 7100 KC. AND 14250 KC.
L = 49.6' FOR 14200 KC. AND 28 MC.
600 OHM LINE
TO TRANSMITTER

183 - Three-Quarter-Wave Folded Doublet No-Switch for Dual-Band

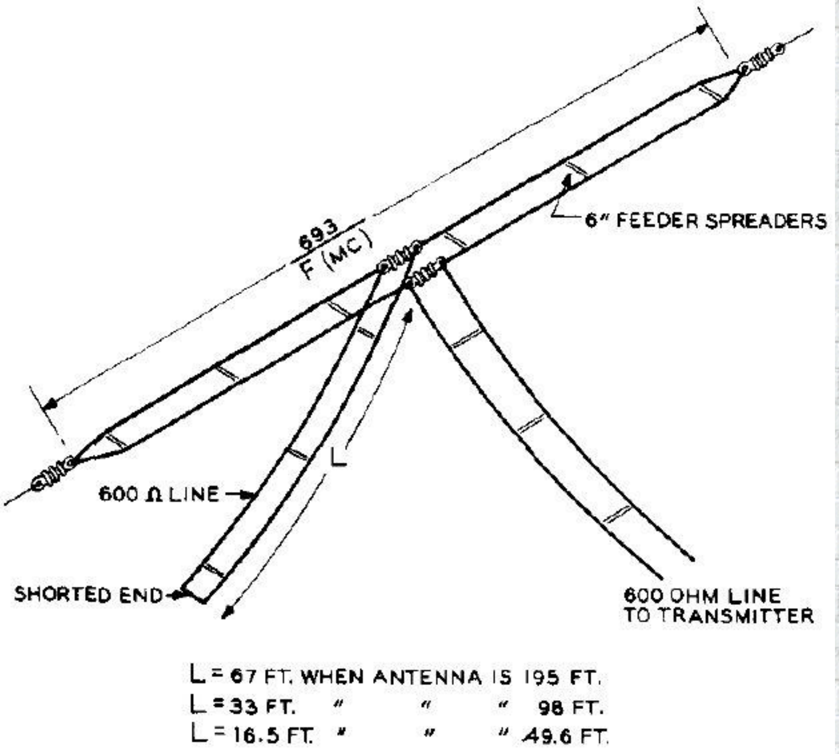
$\frac{693}{F(MC)}$
6" FEEDER SPREADERS
600 Ω LINE
L
SHORTED END
600 OHM LINE
TO TRANSMITTER
L = 67 FT. WHEN ANTENNA IS 195 FT.
L = 33 FT. " " " 98 FT.
L = 16.5 FT. " " " 49.6 FT.

184 - Wideband Omni-Directional Discone Antenna

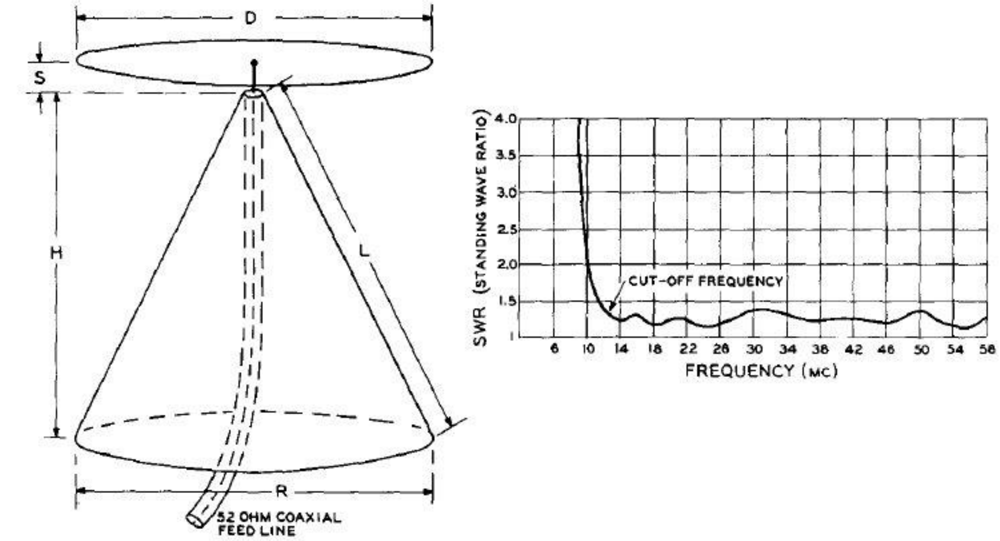
D
S
H
L
R
52 OHM COAXIAL FEED LINE
SWR (STANDING WAVE RATIO)
4.0
3.5
3.0
2.5
2.0
1.5
1
6 10 14 18 22 26 30 34 38 42 46 50 54 58
FREQUENCY (MC)
CUT-OFF FREQUENCY

DIMENSIONS

20, 15, 11, 10, 6 METERS	15, 11, 10, 6 METERS	11, 10, 6, 2 METERS
D=12′ L=18′	D=8′ L=12′	D=6′ L=9′6″
S=10″ R=18′	S=6″ R=12′	S=4″ R=9′6″
H=15′7″	H=10′5″	H=8′3″

185 - Wideband Rhombic Antenna for 40m to 10m

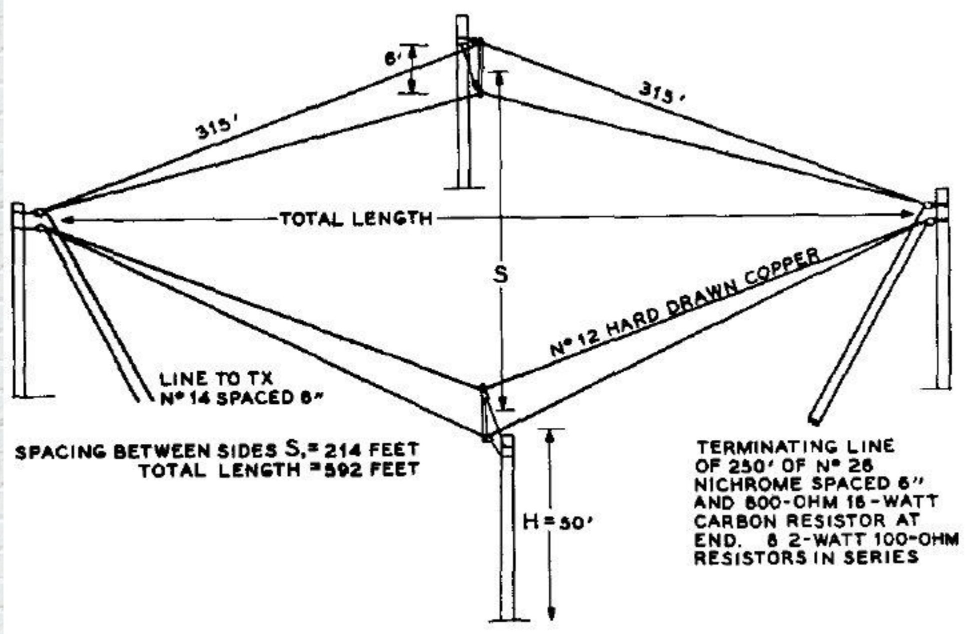
6′
315′
315′
TOTAL LENGTH
S
Nº 12 HARD DRAWN COPPER
LINE TO TX
Nº 14 SPACED 6″
SPACING BETWEEN SIDES S,=214 FEET
TOTAL LENGTH =592 FEET
H=50′
TERMINATING LINE
OF 250′ OF Nº 26
NICHROME SPACED 6″
AND 800-OHM 16-WATT
CARBON RESISTOR AT
END. 8 2-WATT 100-OHM
RESISTORS IN SERIES

186 - Pre-Cut Linear Array Antenna 3dB-Gain for 40m

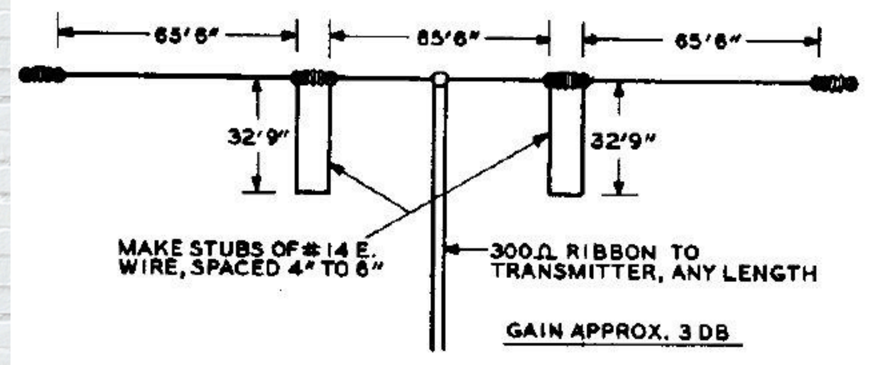
65′6″
65′6″
65′6″
32′9″
32′9″
MAKE STUBS OF # 14 E.
WIRE, SPACED 4″ TO 6″
300Ω RIBBON TO
TRANSMITTER, ANY LENGTH
GAIN APPROX. 3 DB

187 - X-Array Antenna 6dB-Gain for 20m, 15m, 10m

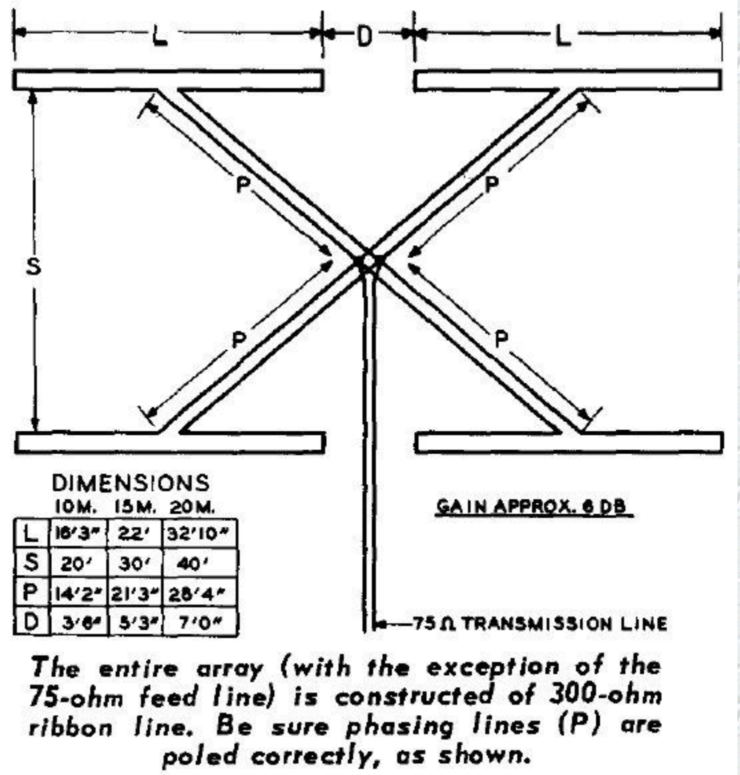
L
D
L
S
P
P
P
P
GAIN APPROX. 6 DB
75 Ω TRANSMISSION LINE

DIMENSIONS

	10M.	15M.	20M.
L	16'3"	22'	32'10"
S	20'	30'	40'
P	14'2"	21'3"	28'4"
D	3'6"	5'3"	7'0"

The entire array (with the exception of the 75-ohm feed line) is constructed of 300-ohm ribbon line. Be sure phasing lines (P) are poled correctly, as shown.

188 - Double-Bruce Array Antenna 5dB-Gain for 20m, 10m

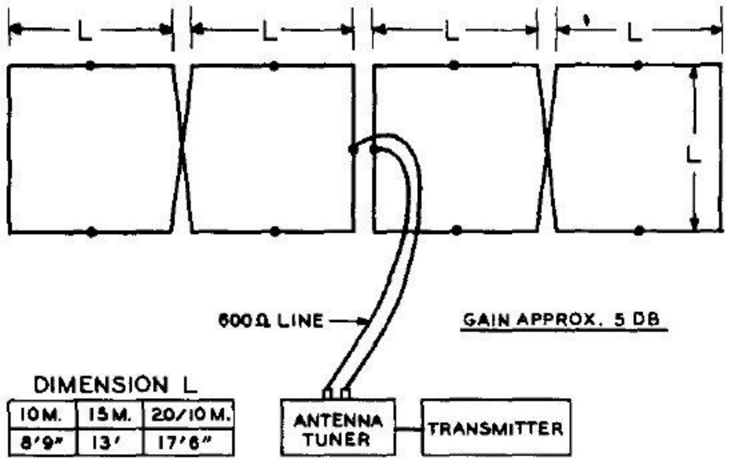
L
L
L
L
L
600 Ω LINE
GAIN APPROX. 5 DB
ANTENNA TUNER
TRANSMITTER

DIMENSION L

10M.	15M.	20/10 M.
8'9"	13'	17'6"

189 - Bi-Square Broadside Array 4dB-Gain for 20m, 15m, 10m

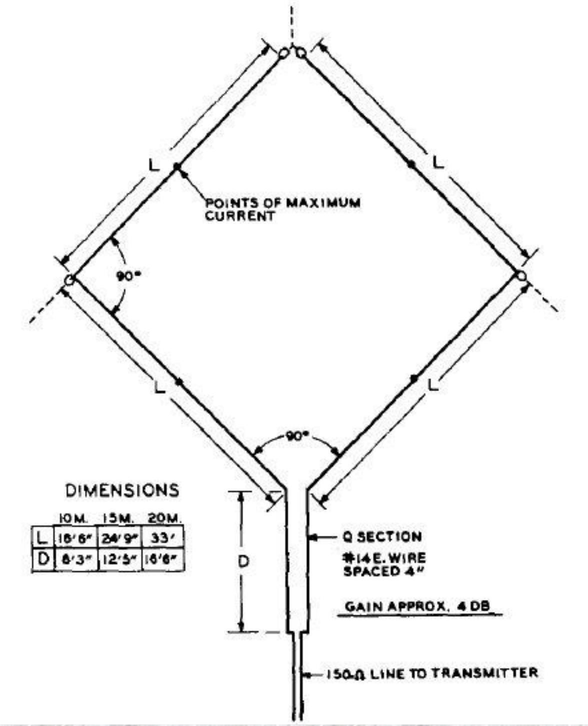
L
L
POINTS OF MAXIMUM CURRENT
90°
L
L
90°
D
Q SECTION
#14 E. WIRE SPACED 4"
GAIN APPROX. 4 DB
150 Ω LINE TO TRANSMITTER

DIMENSIONS

	10M.	15M.	20M.
L	16'6"	24'9"	33'
D	8'3"	12'5"	16'6"

---

190 - Six-Shooter Broadside Array 7.5dB-Gain for 20m, 15m, 10m

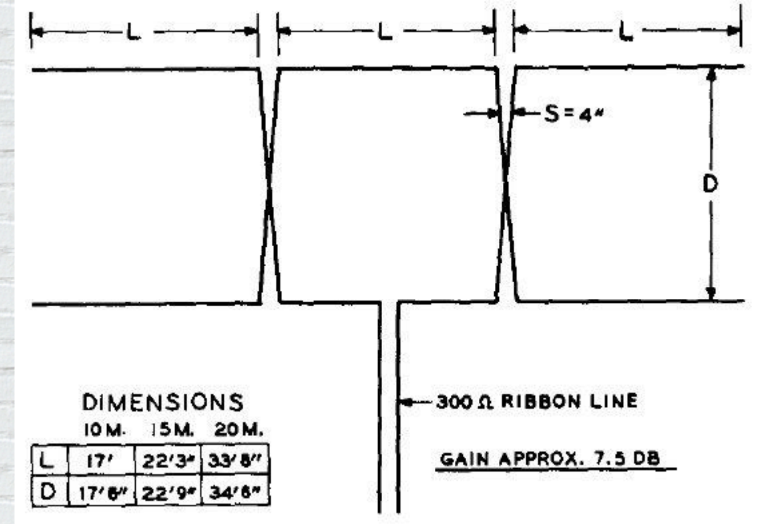
L
L
L
S = 4"
D
300 Ω RIBBON LINE
GAIN APPROX. 7.5 DB

DIMENSIONS

	10M.	15M.	20M.
L	17'	22'3"	33'8"
D	17'6"	22'9"	34'6"

---

191 - Triplex Flat-Top Beam 4.5dB-Gain for 20m, 15m, 10m

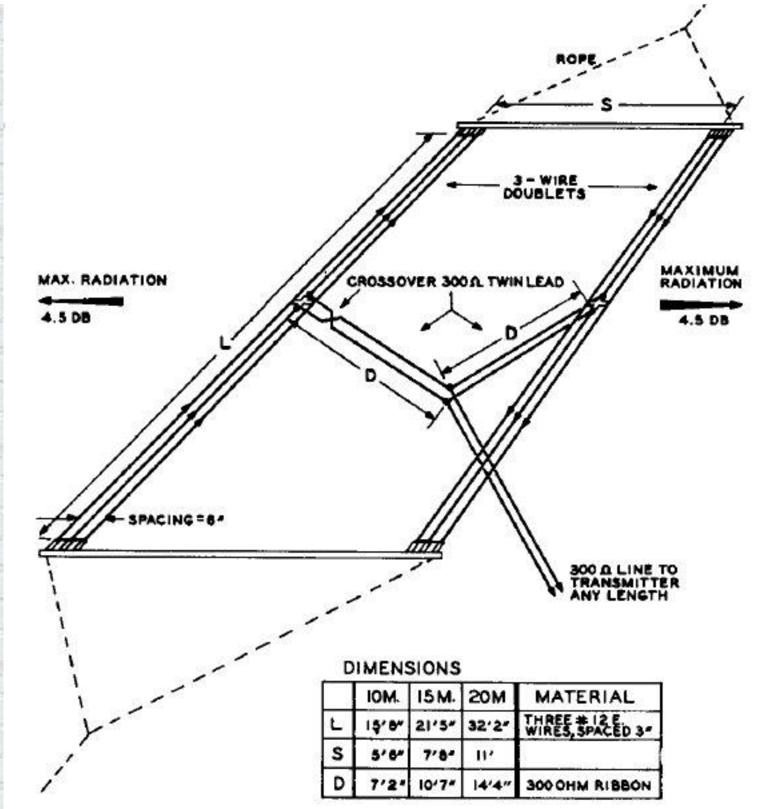
ROPE
S
3-WIRE DOUBLETS
MAX. RADIATION
4.5 DB
CROSSOVER 300Ω TWIN LEAD
MAXIMUM RADIATION
4.5 DB
L
D
D
SPACING = 6"
300 Ω LINE TO TRANSMITTER ANY LENGTH

DIMENSIONS

	10M.	15M.	20M	MATERIAL
L	15'8"	21'5"	32'2"	THREE # 12 E. WIRES, SPACED 3"
S	5'6"	7'8"	11'	
D	7'2"	10'7"	14'4"	300 OHM RIBBON

---

192 - Dual-Band Tilt Antenna for 20m, 10m

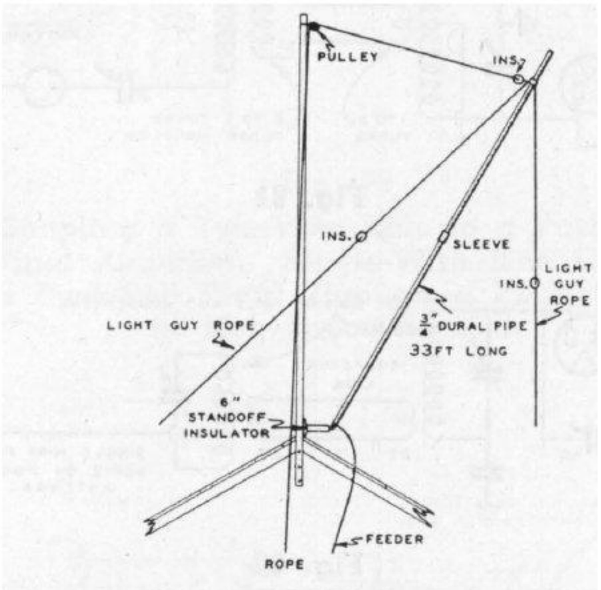
PULLEY
INS.
INS.
SLEEVE
LIGHT GUY ROPE
INS.
LIGHT GUY ROPE
3/4" DURAL PIPE 33FT LONG
6" STANDOFF INSULATOR
FEEDER
ROPE

TWO BAND (10-20 METER) TILT ANTENNA — VERTICAL FOR 20 METERS - TILT 54° FROM HORIZON FOR 10 METERS

---

193 - Super Space Multiband Dipole Antenna for 80m to 10m

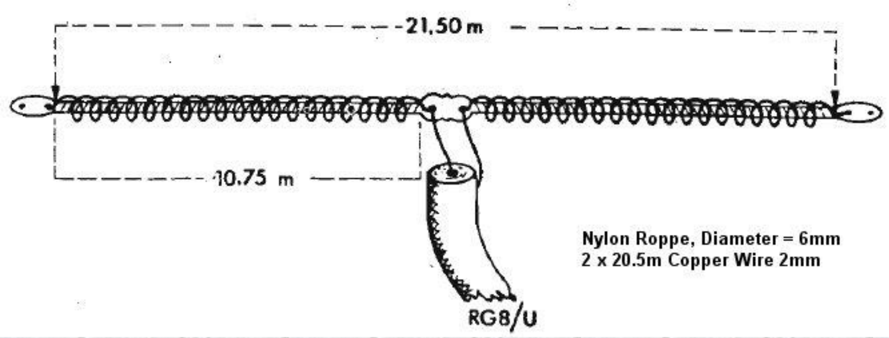
21,50 m
10.75 m
Nylon Roppe, Diameter = 6mm
2 x 20.5m Copper Wire 2mm
RG8/U

---

194 - Bi-Square Beam Antenna Gain=5.5dB for 10m

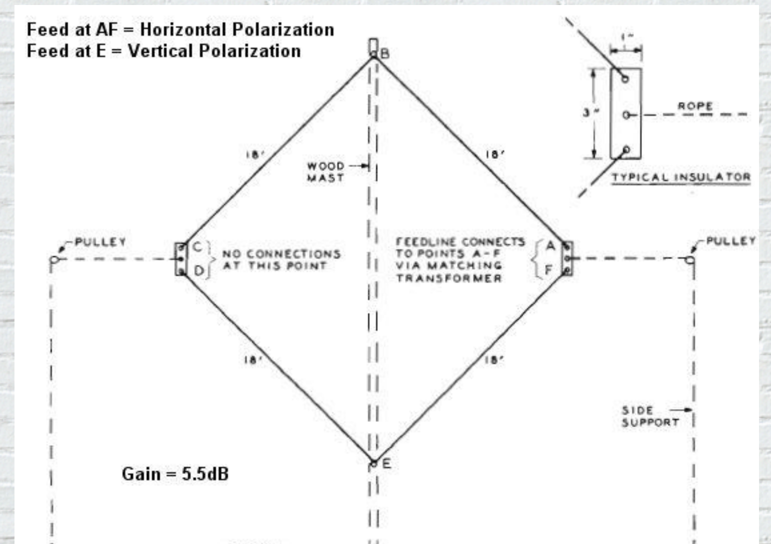
Feed at AF = Horizontal Polarization
Feed at E = Vertical Polarization
B
18'
18'
WOOD MAST
1"
3"
ROPE
TYPICAL INSULATOR
PULLEY
C
D
NO CONNECTIONS AT THIS POINT
FEEDLINE CONNECTS TO POINTS A-F VIA MATCHING TRANSFORMER
A
F
PULLEY
18'
18'
SIDE SUPPORT
E
Gain = 5.5dB
GROUND

---

195 - Cousin of G5RV Multiband Antenna for 40m to 10m

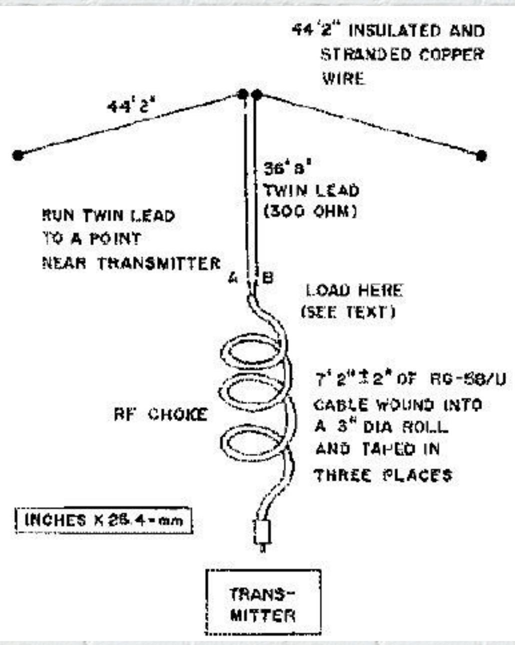
44'2" INSULATED AND STRANDED COPPER WIRE
44'2"
36'8" TWIN LEAD (300 OHM)
RUN TWIN LEAD TO A POINT NEAR TRANSMITTER
A
B
LOAD HERE (SEE TEXT)
7'2"±2" OF RG-58/U CABLE WOUND INTO A 3" DIA ROLL AND TAPED IN THREE PLACES
RF CHOKE
INCHES X 25.4=mm
TRANS-MITTER

---

196 - Cayman Quad Antenna for 20m

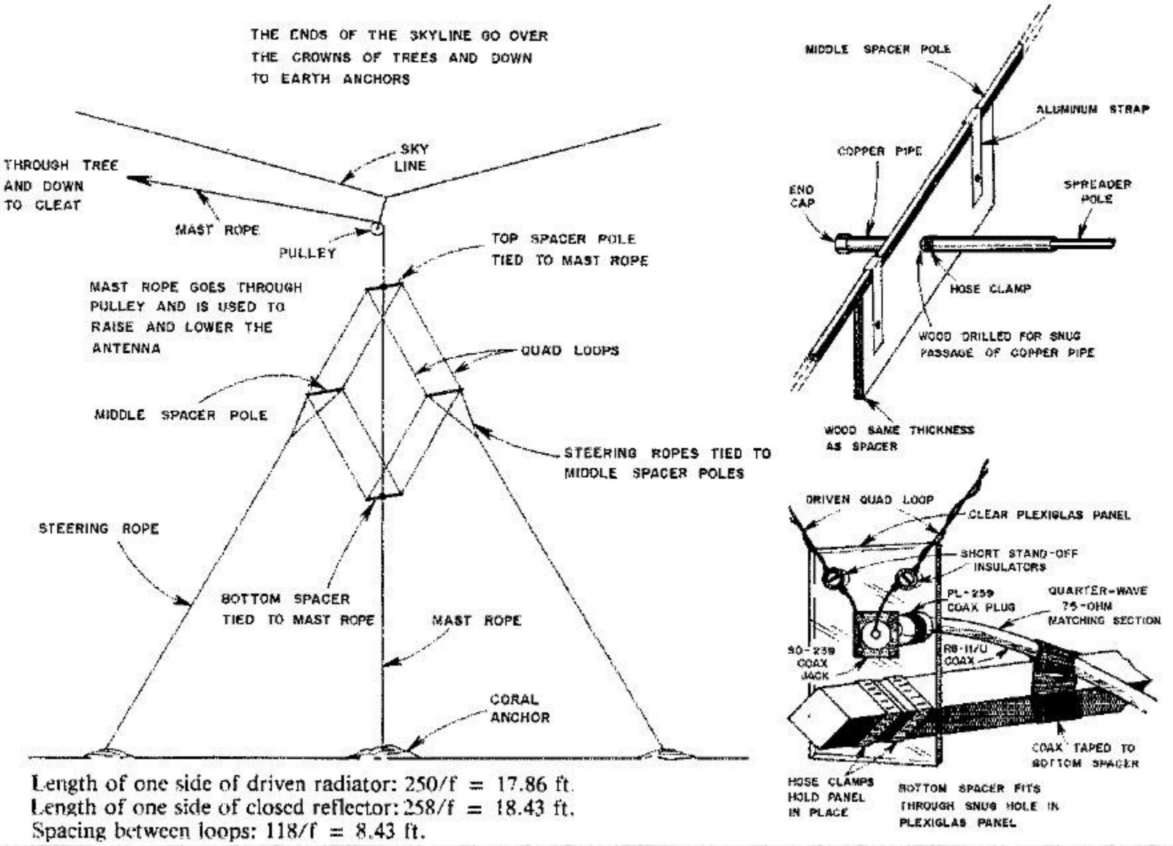
THE ENDS OF THE SKYLINE GO OVER THE CROWNS OF TREES AND DOWN TO EARTH ANCHORS
SKY LINE
THROUGH TREE AND DOWN TO CLEAT
MAST ROPE
PULLEY
TOP SPACER POLE TIED TO MAST ROPE
MAST ROPE GOES THROUGH PULLEY AND IS USED TO RAISE AND LOWER THE ANTENNA
QUAD LOOPS
MIDDLE SPACER POLE
STEERING ROPES TIED TO MIDDLE SPACER POLES
STEERING ROPE
BOTTOM SPACER TIED TO MAST ROPE
MAST ROPE
CORAL ANCHOR
Length of one side of driven radiator: 250/f = 17.86 ft.
Length of one side of closed reflector: 258/f = 18.43 ft.
Spacing between loops: 118/f = 8.43 ft.
MIDDLE SPACER POLE
ALUMINUM STRAP
COPPER PIPE
END CAP
SPREADER POLE
HOSE CLAMP
WOOD DRILLED FOR SNUG PASSAGE OF COPPER PIPE
WOOD SAME THICKNESS AS SPACER
DRIVEN QUAD LOOP
CLEAR PLEXIGLAS PANEL
SHORT STAND-OFF INSULATORS
PL-259 COAX PLUG
QUARTER-WAVE 75-OHM MATCHING SECTION
SO-239 COAX JACK
RG-11/U COAX
COAX TAPED TO BOTTOM SPACER
HOSE CLAMPS HOLD PANEL IN PLACE
BOTTOM SPACER FITS THROUGH SNUG HOLE IN PLEXIGLAS PANEL

197 - Hentenna Gain=3dB with Bazooka Match for 6m

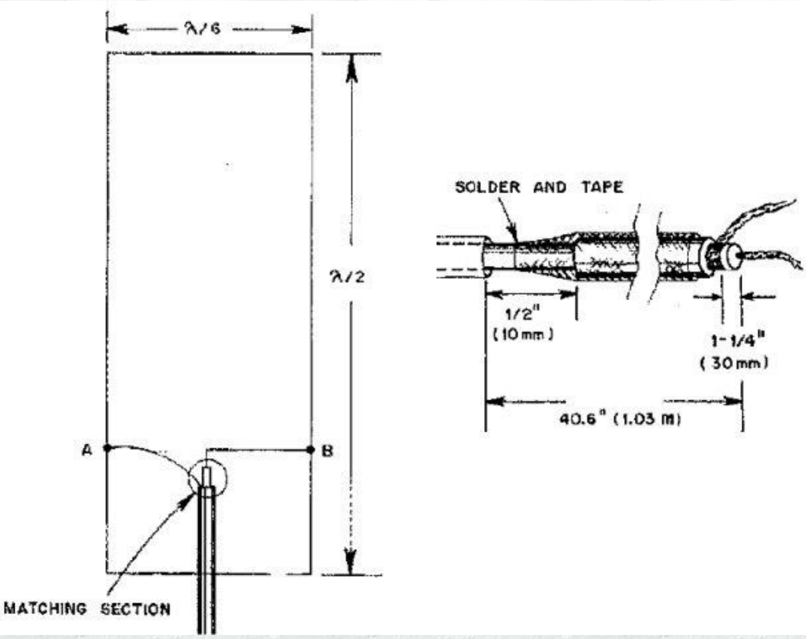
λ/6
λ/2
A
B
MATCHING SECTION
SOLDER AND TAPE
1/2" (10 mm)
1-1/4" (30 mm)
40.6" (1.03 m)

198 - X-Beam Antenna Gain=3dB for 20m

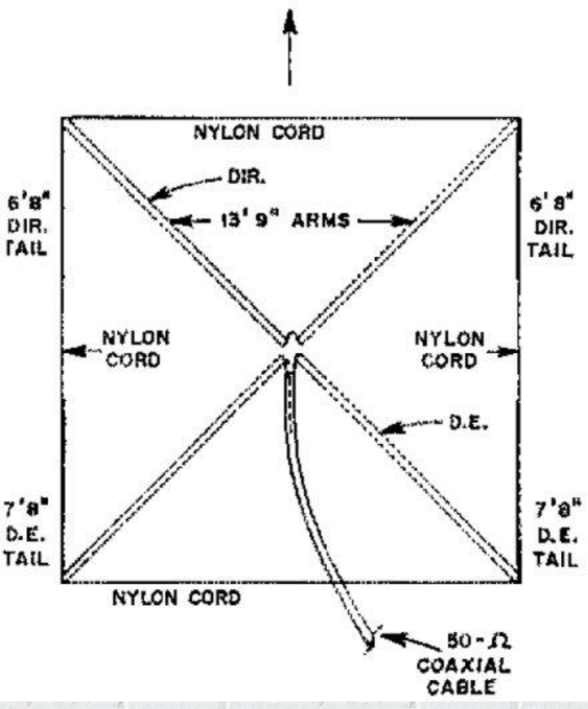
NYLON CORD
DIR.
13' 9" ARMS
6'8" DIR. TAIL
6'8" DIR. TAIL
NYLON CORD
NYLON CORD
D.E.
7'8" D.E. TAIL
7'8" D.E. TAIL
NYLON CORD
50-Ω COAXIAL CABLE

---

199 - Twin Delta Loop Antenna Gain=6dB for 160m, 80m, 40m

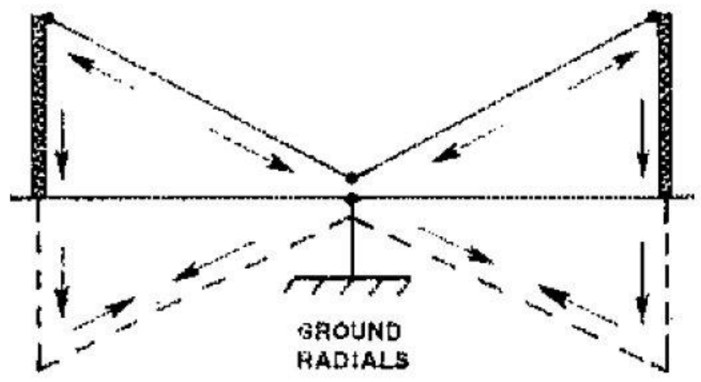
GROUND RADIALS

Twin- Delta Loop, with the image in the ground plane.

Band	*Tower Height (ft)*	*Length of Sloping Wires (ft)*
160 m†	102.03	204.06
80 m	49.86	99.72
40 m	25.92	51.84

---

200 - Inverted-V Beam Antenna for 30m

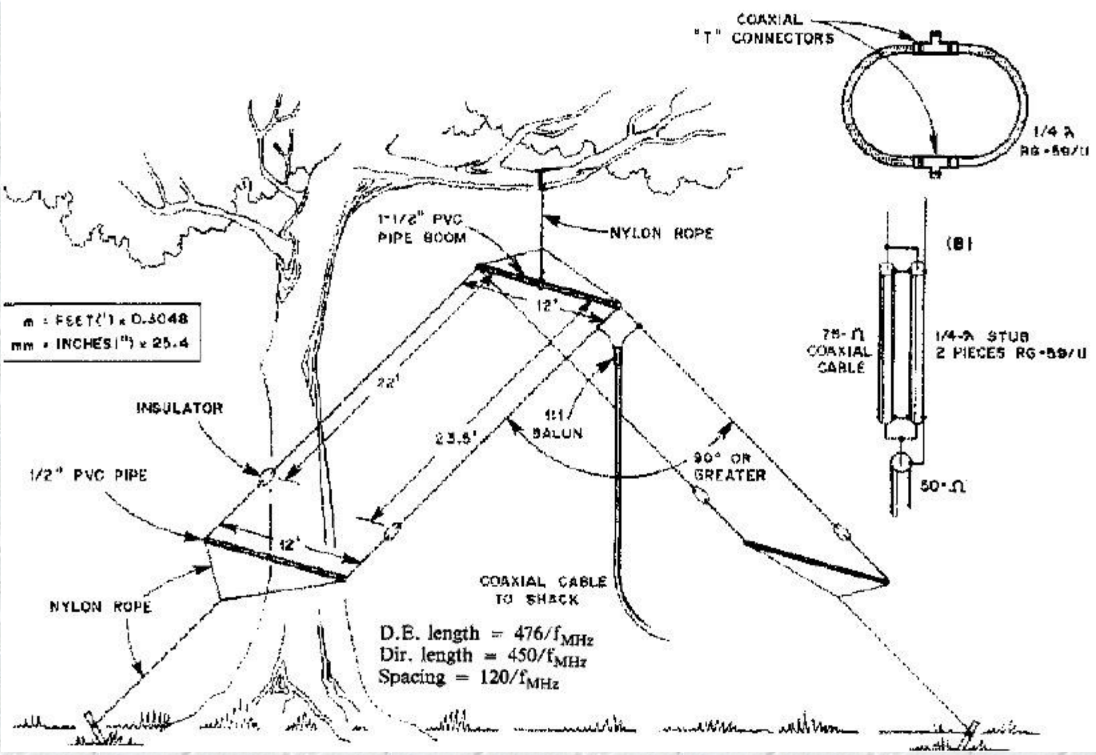
COAXIAL "T" CONNECTORS
1/4 λ RG-59/U
(B)
75-Ω COAXIAL CABLE
1/4-λ STUB 2 PIECES RG-59/U
50-Ω
1"1/2" PVC PIPE BOOM
NYLON ROPE
m = FEET(') x 0.3048
mm = INCHES(") x 25.4
12'
22'
23.5'
1:1 BALUN
INSULATOR
90° OR GREATER
1/2" PVC PIPE
12'
NYLON ROPE
COAXIAL CABLE TO SHACK
D.E. length = 476/f_{MHz}
Dir. length = 450/f_{MHz}
Spacing = 120/f_{MHz}

---

201 - FD4 Windom Antenna for 80m, 40m, 20m, 10m

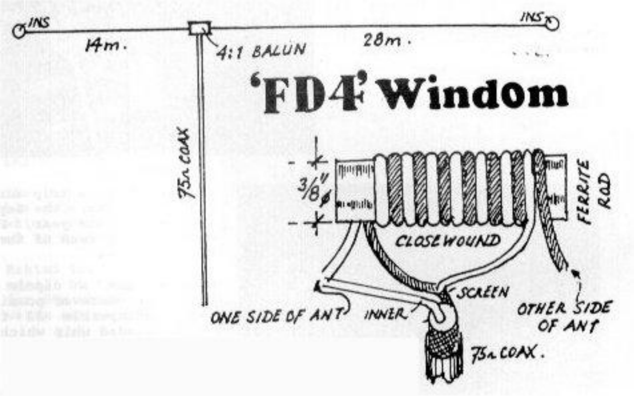
INS
INS
14m.
28m.
4:1 BALUN
'FD4' Windom
75Ω COAX
3/8" ϕ
FERRITE ROD
CLOSEWOUND
SCREEN
ONE SIDE OF ANT
INNER
OTHER SIDE OF ANT
75Ω COAX.

---

202 - Two Elements Phased Delta Loop for 40m

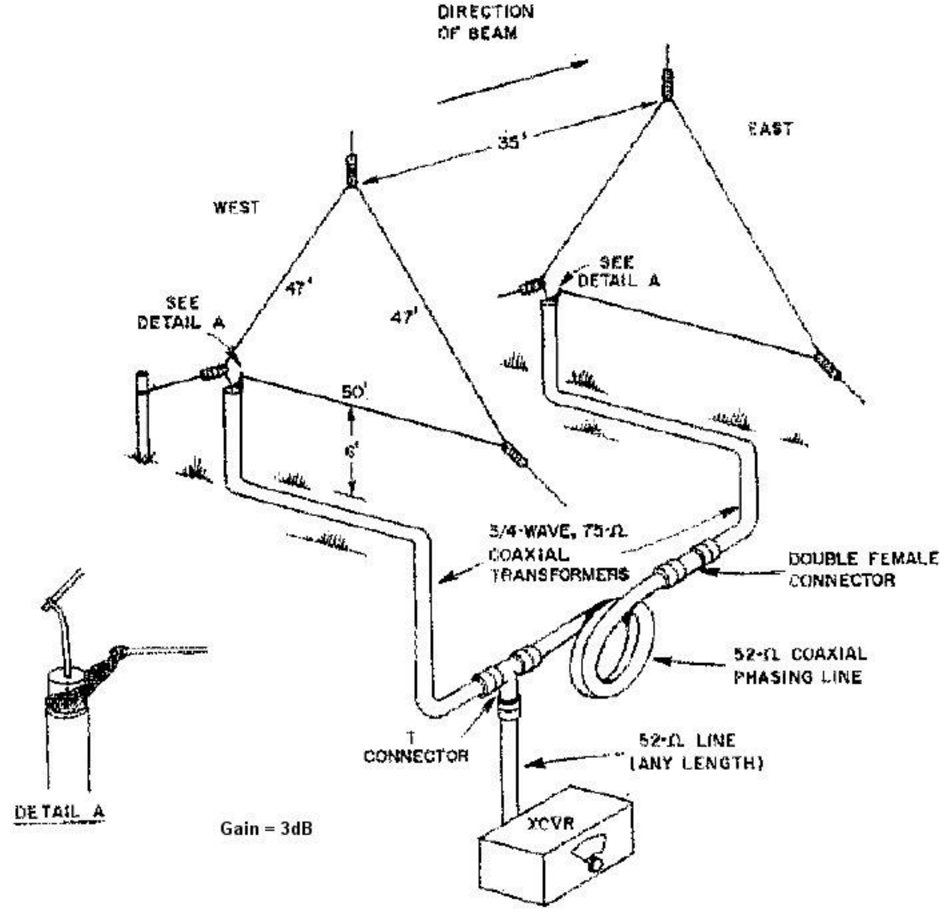
DIRECTION OF BEAM
35'
EAST
WEST
SEE DETAIL A
47'
47'
SEE DETAIL A
50'
6'
3/4-WAVE, 75-Ω COAXIAL TRANSFORMERS
DOUBLE FEMALE CONNECTOR
52-Ω COAXIAL PHASING LINE
T CONNECTOR
52-Ω LINE (ANY LENGTH)
XCVR
DETAIL A
Gain = 3dB

203 - Three Elements Wire Yagi for 10m

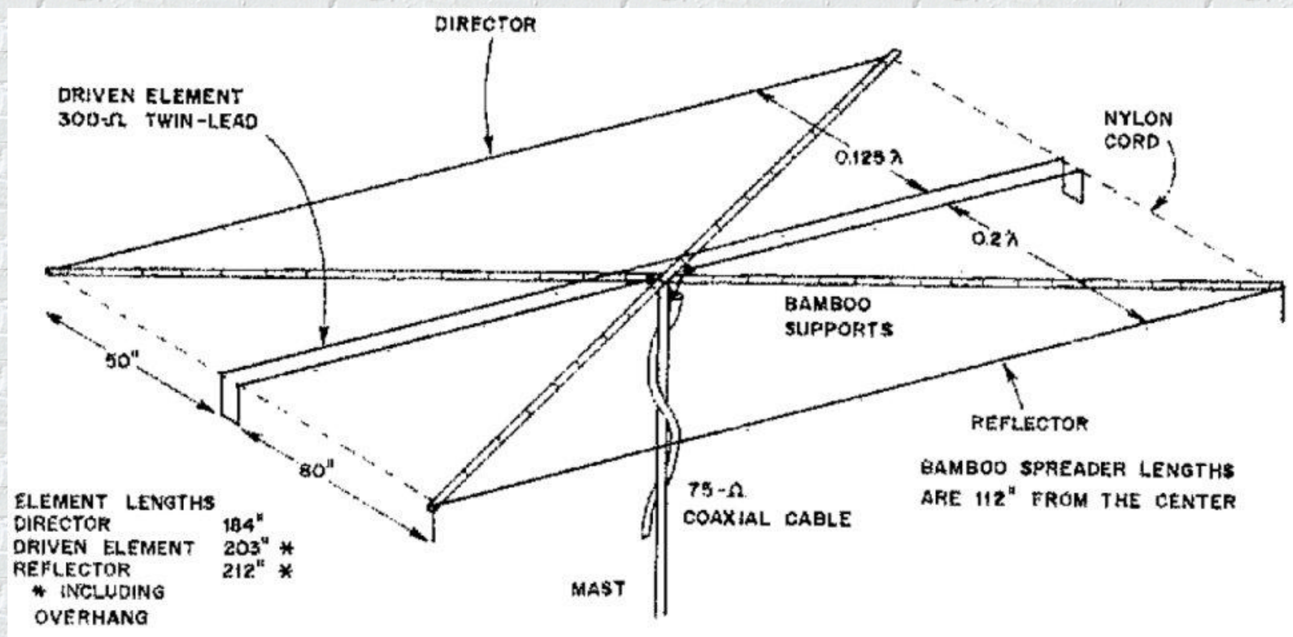
DIRECTOR
DRIVEN ELEMENT 300-Ω TWIN-LEAD
NYLON CORD
0.125 λ
0.2 λ
BAMBOO SUPPORTS
50"
80"
REFLECTOR
BAMBOO SPREADER LENGTHS ARE 112" FROM THE CENTER
75-Ω COAXIAL CABLE
MAST
ELEMENT LENGTHS
DIRECTOR 184"
DRIVEN ELEMENT 203" *
REFLECTOR 212" *
* INCLUDING OVERHANG

204 - Low Radiation Angle Full-Wave Loop Antenna for 80m

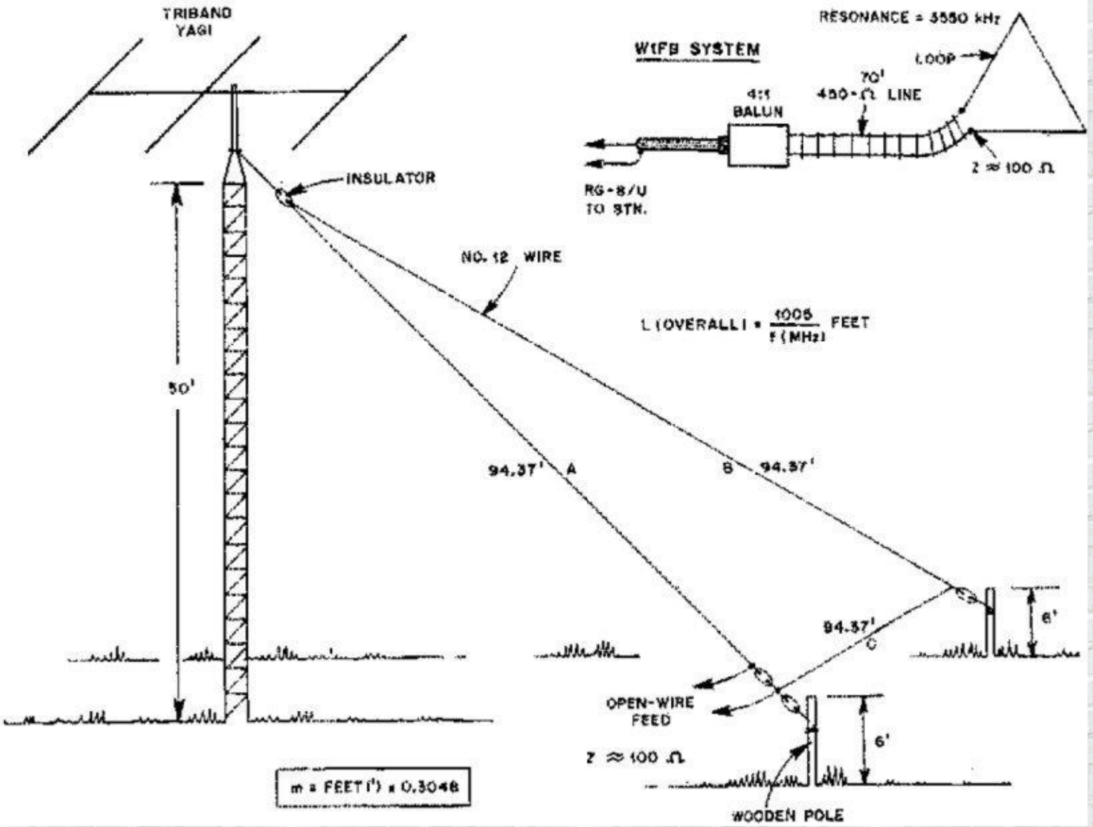
TRIBAND YAGI
W1FB SYSTEM
RESONANCE = 3550 kHz
LOOP
70'
4:1 BALUN
450-Ω LINE
Z ≈ 100 Ω
RG-8/U TO STN.
INSULATOR
NO. 12 WIRE
L (OVERALL) = 1005 / f (MHz) FEET
50'
94.37' A
B 94.37'
94.37' C
6'
OPEN-WIRE FEED
Z ≈ 100 Ω
6'
WOODEN POLE
m = FEET (') x 0.3048

205 - Collinear Antenna Gain=3dB for 2m

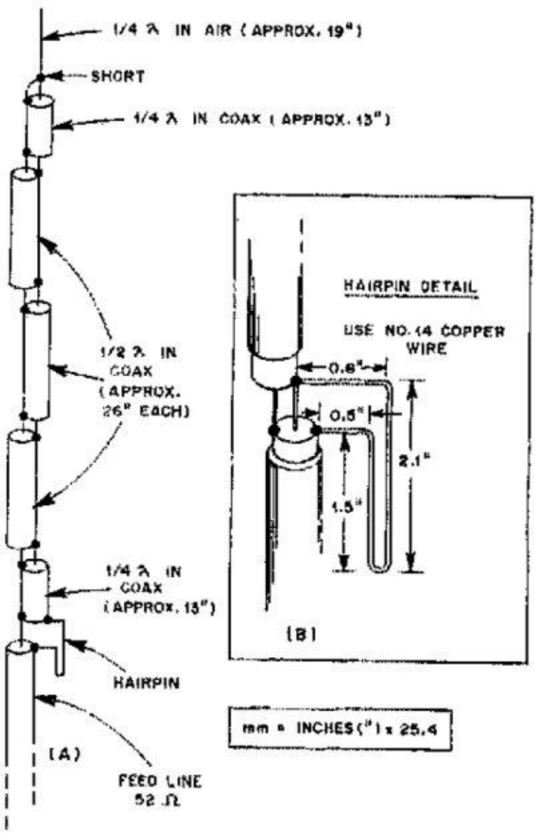
1/4 λ IN AIR (APPROX. 19")
SHORT
1/4 λ IN COAX (APPROX. 13")
1/2 λ IN COAX (APPROX. 26" EACH)
1/4 λ IN COAX (APPROX. 13")
HAIRPIN
(A)
FEED LINE 52 Ω
HAIRPIN DETAIL
USE NO. 14 COPPER WIRE
0.8"
0.5"
2.1"
1.5"
(B)
mm = INCHES (") x 25.4

206 - X-Ray Monoband Antenna Gain=6.5dBd

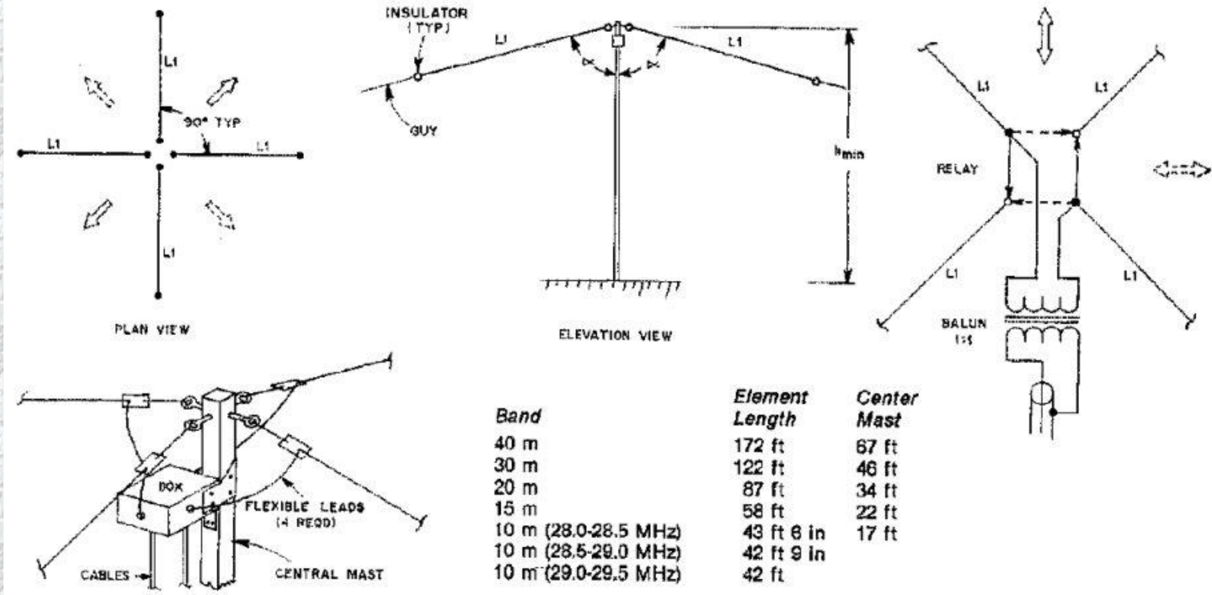
INSULATOR (TYP)
L1
GUY
α
h_min
90° TYP
PLAN VIEW
ELEVATION VIEW
RELAY
BALUN 1:1
BOX
FLEXIBLE LEADS (4 REQD)
CABLES
CENTRAL MAST

Band	*Element Length*	*Center Mast*
40 m	172 ft	67 ft
30 m	122 ft	46 ft
20 m	87 ft	34 ft
15 m	58 ft	22 ft
10 m (28.0-28.5 MHz)	43 ft 6 in	17 ft
10 m (28.5-29.0 MHz)	42 ft 9 in	
10 m (29.0-29.5 MHz)	42 ft	

---

207 - Broadside Collinear Curtain Array for 20m

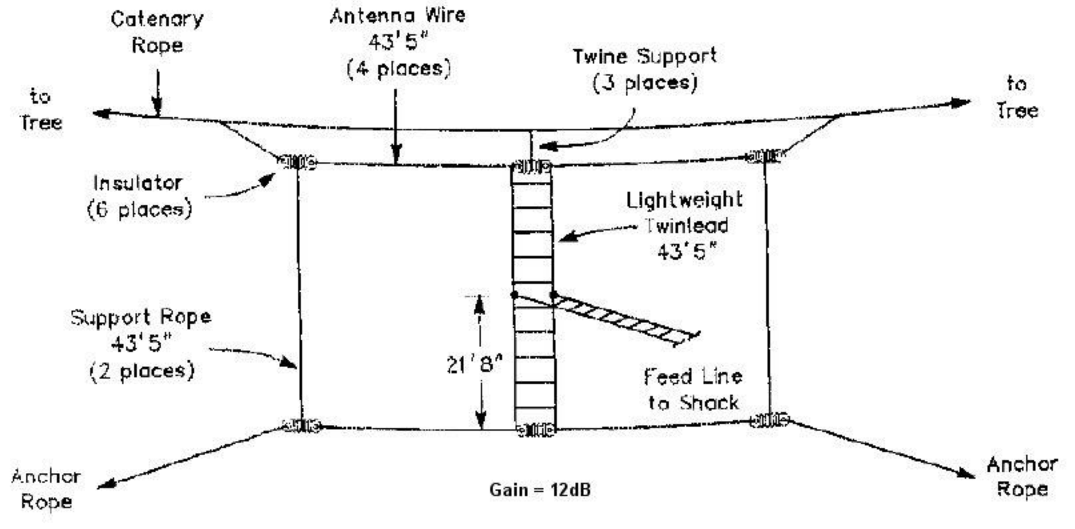
Catenary Rope
Antenna Wire 43' 5" (4 places)
Twine Support (3 places)
to Tree
to Tree
Insulator (6 places)
Lightweight Twinlead 43' 5"
Support Rope 43' 5" (2 places)
21' 8"
Feed Line to Shack
Anchor Rope
Anchor Rope
Gain = 12dB

---

208 - Shrunk Loop Antenna for 80m, 40m

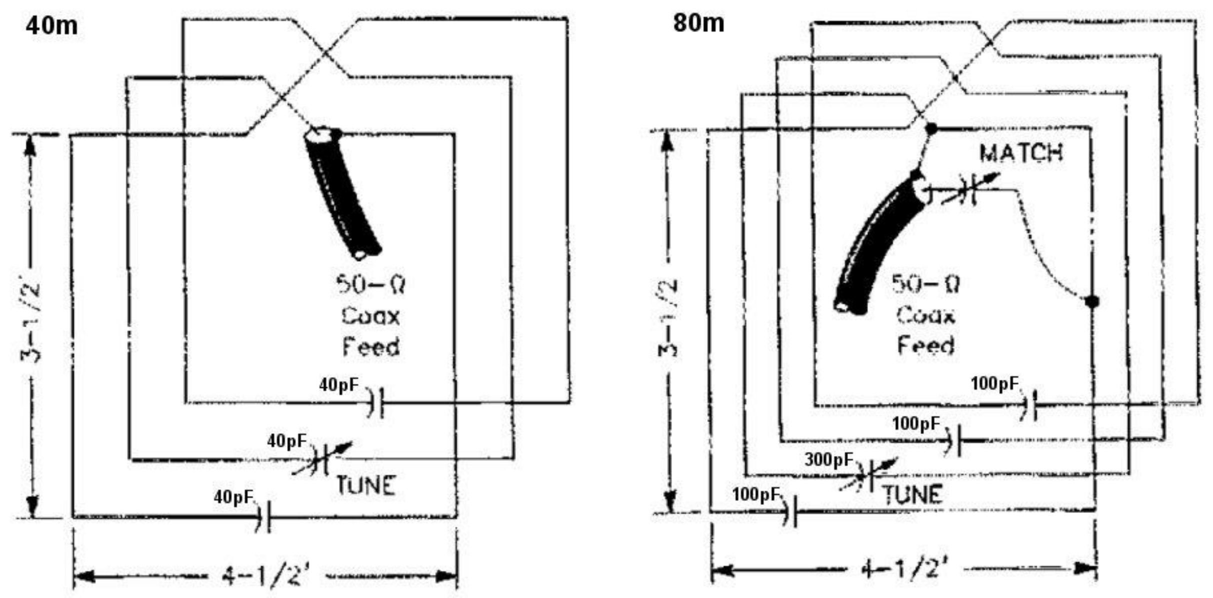
40m
50-Ω
Coax
Feed
40pF
40pF
TUNE
40pF
3-1/2'
4-1/2'
80m
MATCH
50-Ω
Coax
Feed
100pF
100pF
300pF
TUNE
100pF
3-1/2
4-1/2'

---

209 - λ/3 Multiband Dipole for 40m, 20m, 17m, 15m, 13m, 10m

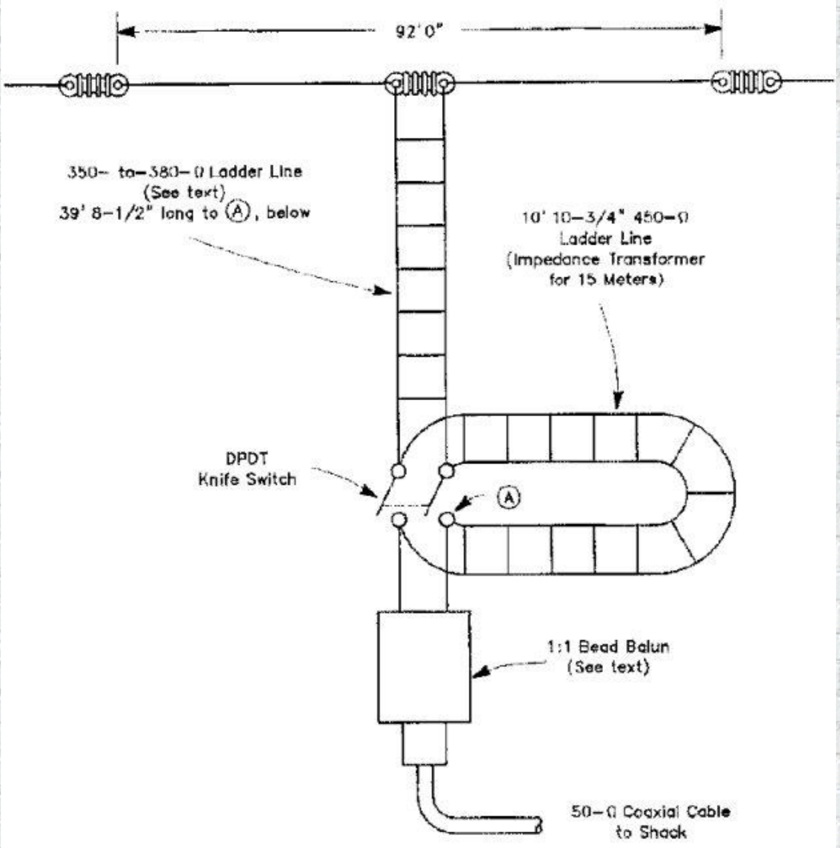
92'0"
350- to-380-Ω Ladder Line
(See text)
39' 8-1/2" long to Ⓐ, below
10' 10-3/4" 450-Ω
Ladder Line
(Impedance Transformer
for 15 Meters)
DPDT
Knife Switch
Ⓐ
1:1 Bead Balun
(See text)
50-Ω Coaxial Cable
to Shack

---

210 - K3LR Sloper Antenna Gain=3dB for 160m

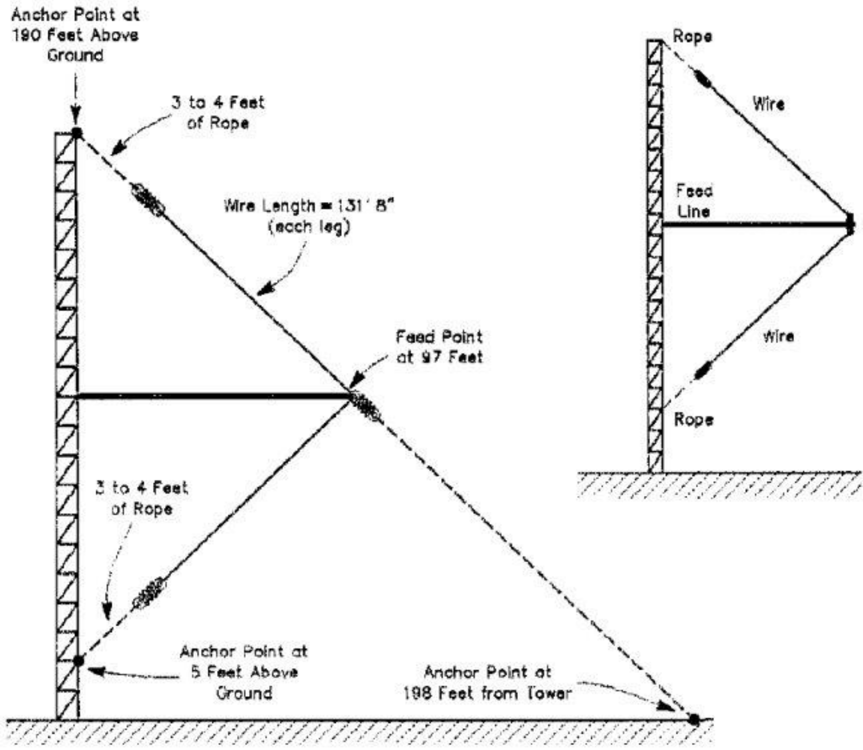
Anchor Point at 190 Feet Above Ground
3 to 4 Feet of Rope
Wire Length = 131' 8" (each leg)
Feed Point at 97 Feet
3 to 4 Feet of Rope
Anchor Point at 5 Feet Above Ground
Anchor Point at 198 Feet from Tower
Rope
Wire
Feed Line
Wire
Rope

---

211 - Mini Folded Vertical Monopole Antenna for 160m, 80m, 40m

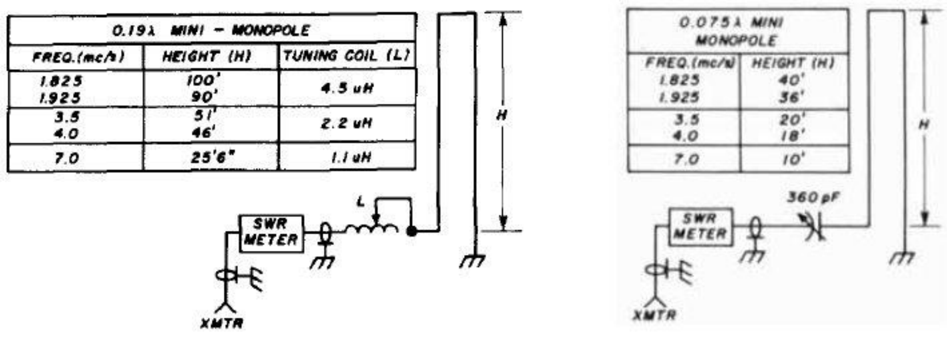

0.19λ MINI - MONOPOLE

FREQ.(mc/s)	HEIGHT (H)	TUNING COIL (L)
1.825 1.925	100' 90'	4.5 uH
3.5 4.0	51' 46'	2.2 uH
7.0	25'6"	1.1 uH

0.075λ MINI MONOPOLE

FREQ.(mc/s)	HEIGHT (H)
1.825 1.925	40' 36'
3.5 4.0	20' 18'
7.0	10'

H
L
SWR METER
XMTR
H
360 pF
SWR METER
XMTR

---

212 - Three Band Dipole for 80m, 40m, 20m

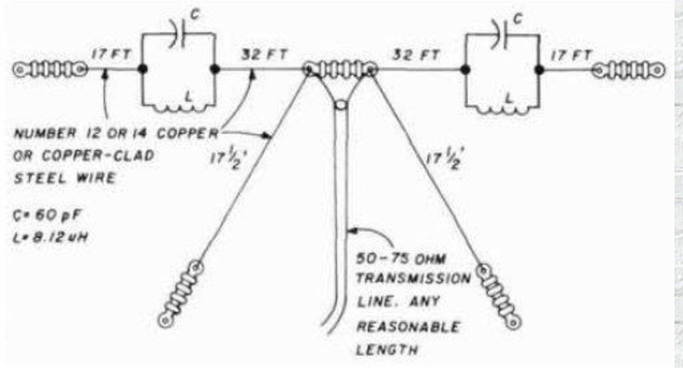
C
17 FT
32 FT
32 FT
C
17 FT
L
L
NUMBER 12 OR 14 COPPER OR COPPER-CLAD STEEL WIRE
17½'
17½'
C= 60 pF
L= 8.12 uH
50-75 OHM TRANSMISSION LINE. ANY REASONABLE LENGTH

---

213 - Stub Matched Dipole Antenna

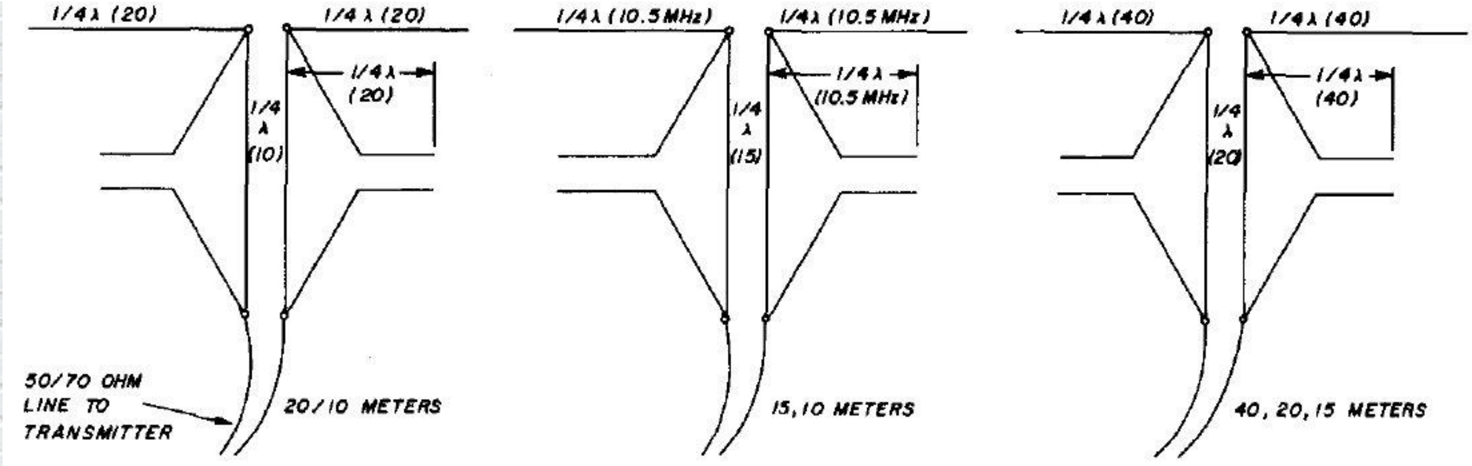
1/4 λ (20)
1/4 λ (20)
1/4 λ (20)
1/4 λ (10)
50/70 OHM LINE TO TRANSMITTER
20/10 METERS
1/4 λ (10.5 MHz)
1/4 λ (10.5 MHz)
1/4 λ (10.5 MHz)
1/4 λ (15)
15, 10 METERS
1/4 λ (40)
1/4 λ (40)
1/4 λ (40)
1/4 λ (20)
40, 20, 15 METERS

214 - Inverted-Vee Antenna with Bazooka Match

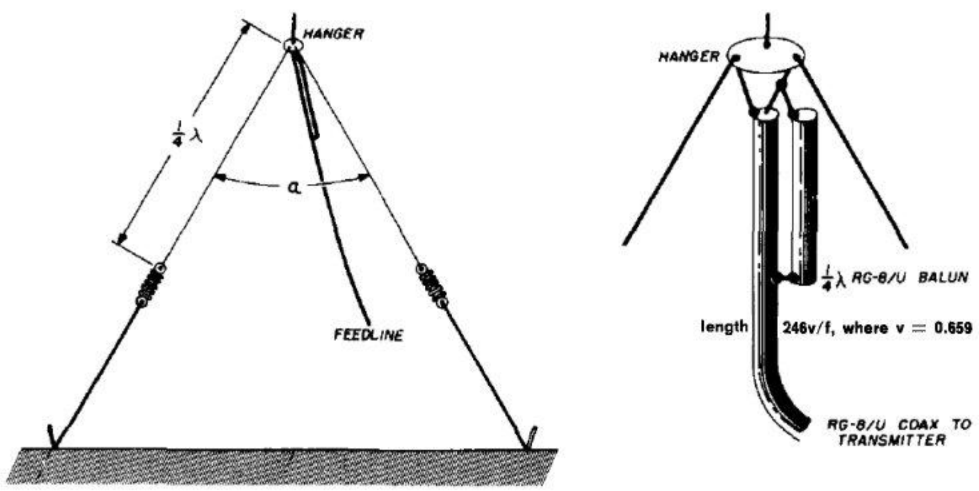
HANGER
$\frac{1}{4}\lambda$
a
FEEDLINE
HANGER
$\frac{1}{4}\lambda$ RG-8/U BALUN
length 246v/f, where v = 0.659
RG-8/U COAX TO TRANSMITTER

215 - Multiband Long Wire Antenna for 40m, 20m, 15m, 10m

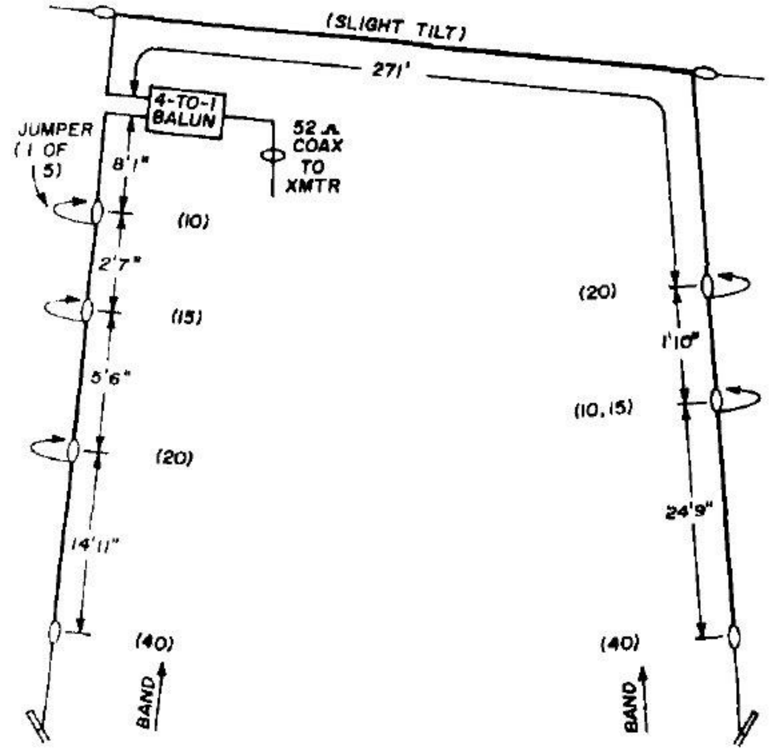
(SLIGHT TILT)
271'
4-TO-1 BALUN
52 Ω COAX TO XMTR
JUMPER (1 OF 5)
8'1"
(10)
2'7"
(15)
5'6"
(20)
14'11"
(40)
BAND
(20)
1'10"
(10, 15)
24'9"
(40)
BAND

216 - Converted Vee Antenna for 80m, 40m

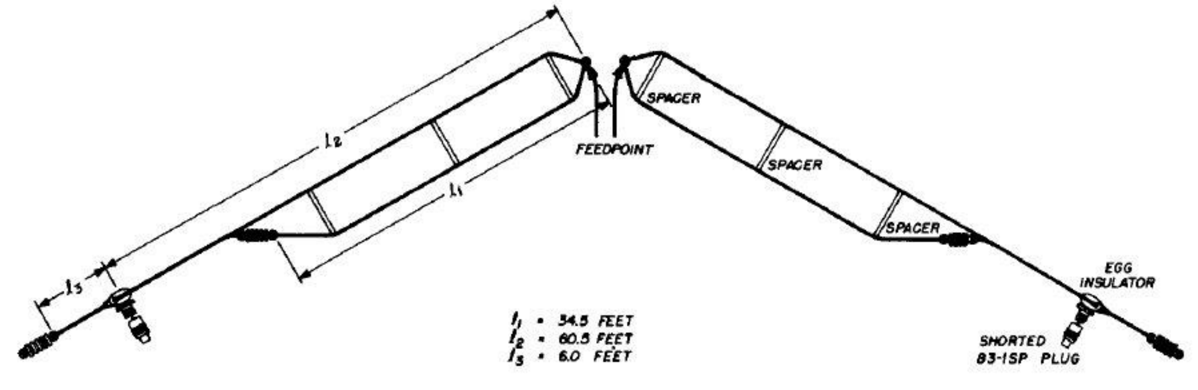
l_2
l_1
l_3
FEEDPOINT
SPACER
SPACER
SPACER
EGG INSULATOR
SHORTED 83-1SP PLUG
l_1 = 34.5 FEET
l_2 = 60.5 FEET
l_3 = 6.0 FEET

217 - Swiss Quad Antenna for 20m, 15m, 10m

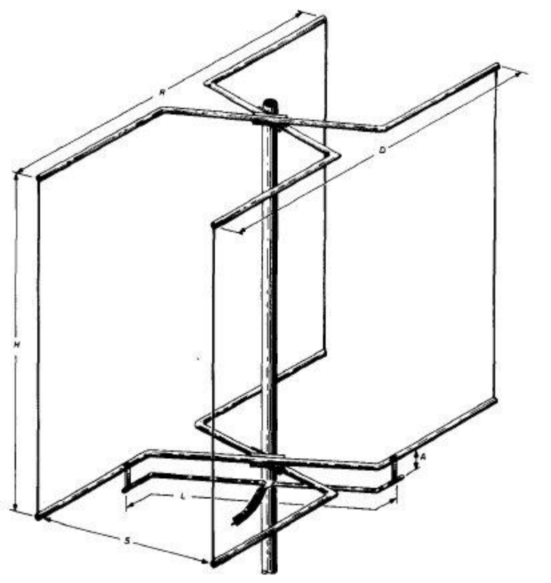

Dimensions for the Swiss quad (in inches).

frequency (MHz)	reflector horizontal (R)	radiator horizonal (D)	vertical (H)	spacing (S)	gamma length (L)	gamma spacing (A)
14.05	248	223½	236	84	188½	4¼
14.2	245	221	233	83	186½	4¼
21.05	165	149	157	56	126	2½
21.2	162½	148	155½	55½	124½	2½
28.05	124	112	118	42	94½	2
28.5	122	110	116	41½	93	2

218 - Japanese Quad Antenna for 20m, 15m, 10m

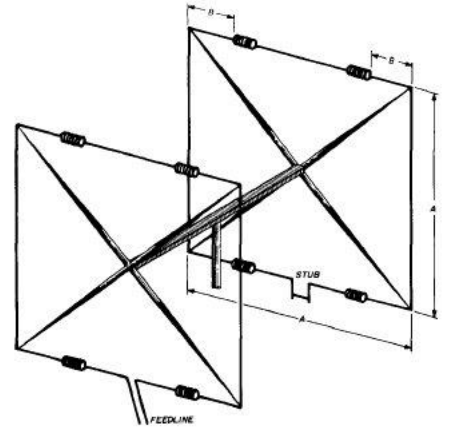
B
B
A
A
STUB
FEEDLINE

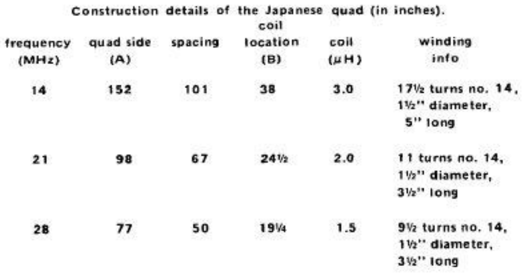

Construction details of the Japanese quad (in inches).

frequency (MHz)	quad side (A)	spacing	coil location (B)	coil (µH)	winding info
14	152	101	38	3.0	17½ turns no. 14, 1½" diameter, 5" long
21	98	67	24½	2.0	11 turns no. 14, 1½" diameter, 3½" long
28	77	50	19¼	1.5	9½ turns no. 14, 1½" diameter, 3½" long

219 - Seven-Elements Triangle Beam Antenna for 20m

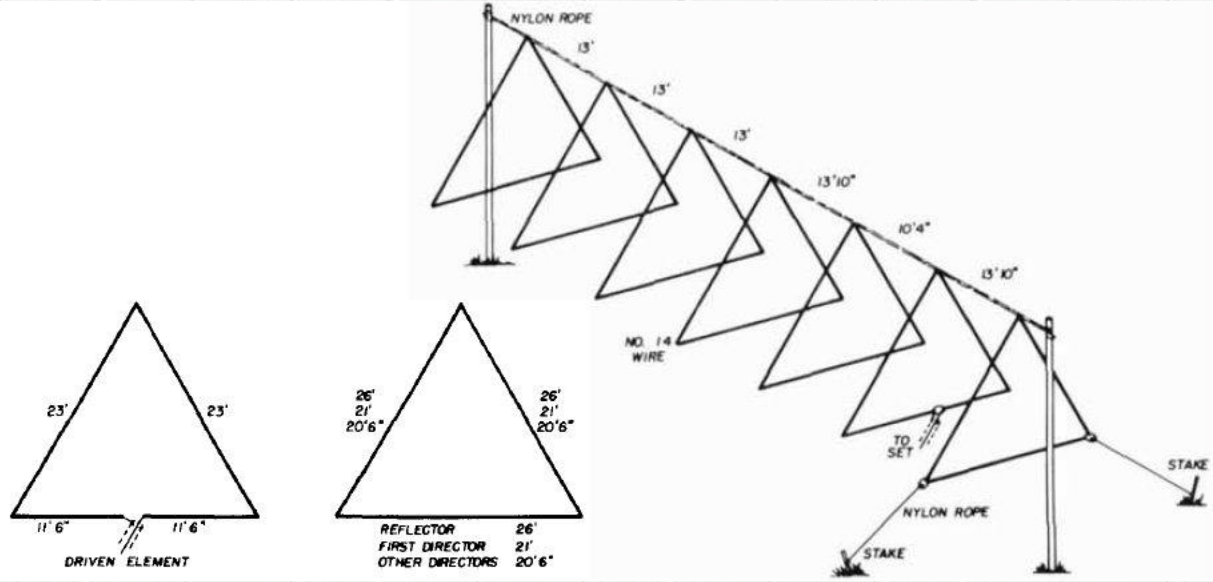
NYLON ROPE
13'
13'
13'
13'10"
10'4"
13'10"
NO. 14 WIRE
TO SET
STAKE
NYLON ROPE
STAKE
23'
23'
11'6"
11'6"
DRIVEN ELEMENT
26'
21'
20'6"
26'
21'
20'6"
REFLECTOR 26'
FIRST DIRECTOR 21'
OTHER DIRECTORS 20'6"

220 - Log Periodic Wire Antenna for 40m

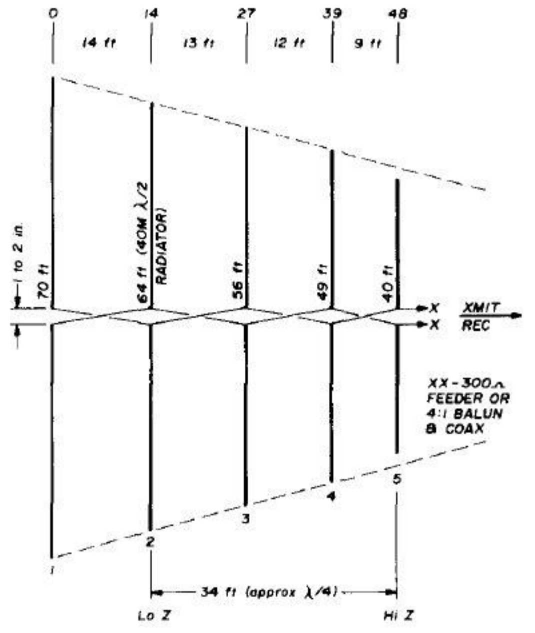
0
14
27
39
48
14 ft
13 ft
12 ft
9 ft
1 to 2 in.
70 ft
64 ft (40M λ/2 RADIATOR)
56 ft
49 ft
40 ft
X XMIT
X REC
XX-300Ω FEEDER OR 4:1 BALUN & COAX
1
2
3
4
5
34 ft (approx λ/4)
Lo Z
Hi Z

221 - Three-Elements 90-degrees Wire Beam for 20m

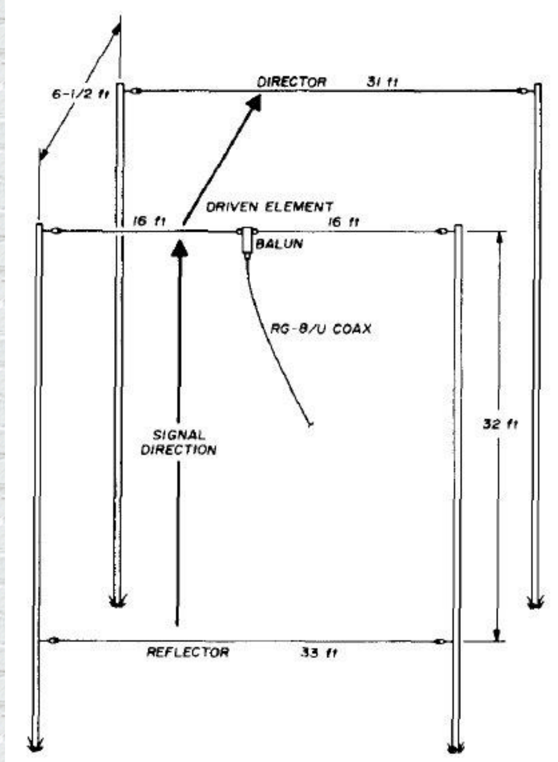
6-1/2 ft
DIRECTOR
31 ft
16 ft
DRIVEN ELEMENT
16 ft
BALUN
RG-8/U COAX
SIGNAL
DIRECTION
32 ft
REFLECTOR
33 ft

---

222 - Two-Elements Wire Beam for 20m

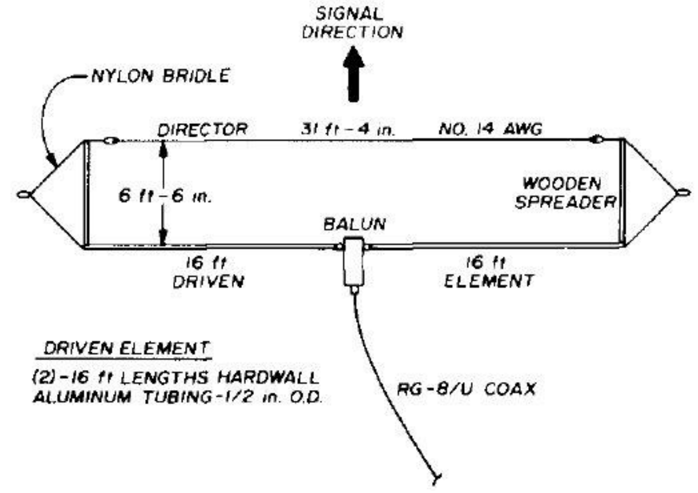
SIGNAL
DIRECTION
NYLON BRIDLE
DIRECTOR
31 ft - 4 in.
NO. 14 AWG
6 ft - 6 in.
WOODEN
SPREADER
BALUN
16 ft
DRIVEN
16 ft
ELEMENT
DRIVEN ELEMENT
(2)-16 ft LENGTHS HARDWALL
ALUMINUM TUBING-1/2 in. O.D.
RG-8/U COAX

---

223 - Suitcase-Fit Dipole Antenna for 80m, 40m, 20m

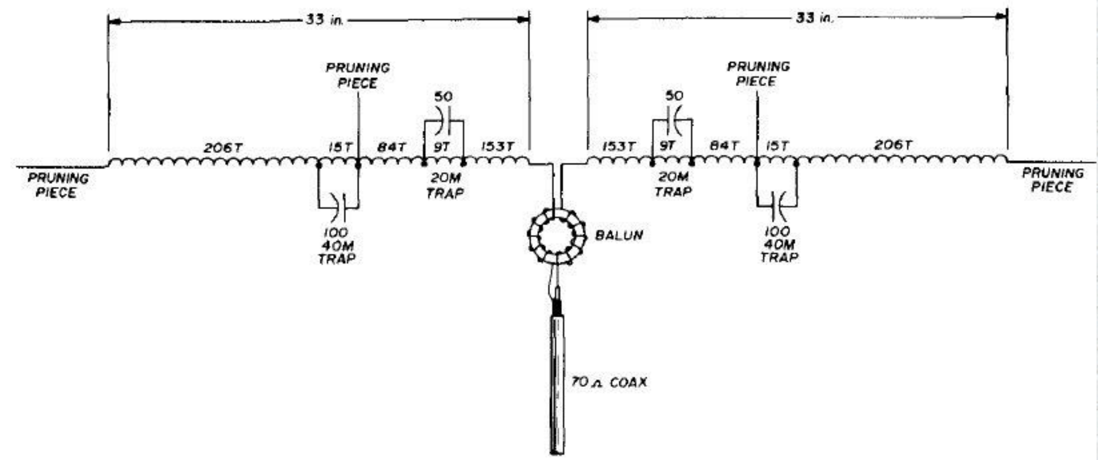
33 in.
33 in.
PRUNING PIECE
PRUNING PIECE
50
50
206T
15T
84T
9T
153T
153T
9T
84T
15T
206T
PRUNING PIECE
PRUNING PIECE
20M TRAP
20M TRAP
BALUN
100 40M TRAP
100 40M TRAP
70 Ω COAX

---

224 - Multiband Coaxial Dipole for 80m, 40m

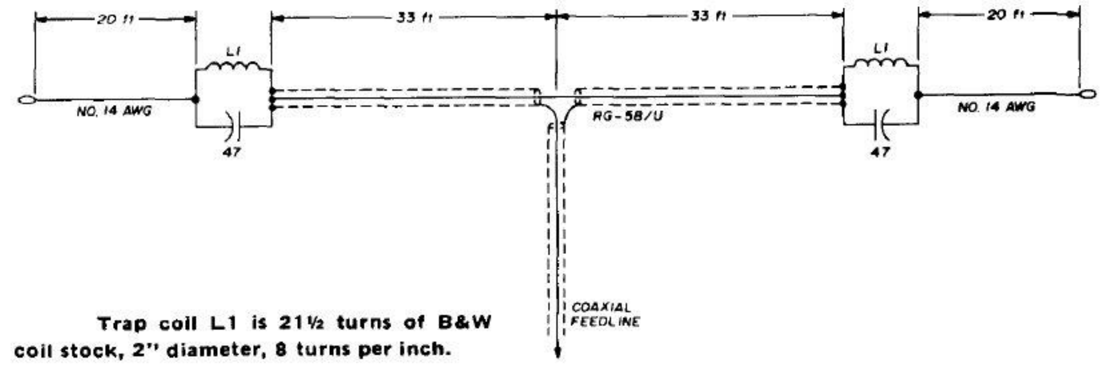
20 ft
33 ft
33 ft
20 ft
L1
L1
NO. 14 AWG
NO. 14 AWG
RG-58/U
47
47
COAXIAL FEEDLINE

Trap coil L1 is 21½ turns of B&W coil stock, 2" diameter, 8 turns per inch.

---

225 - Vertical Log Periodic Antenna for 80m, 40m

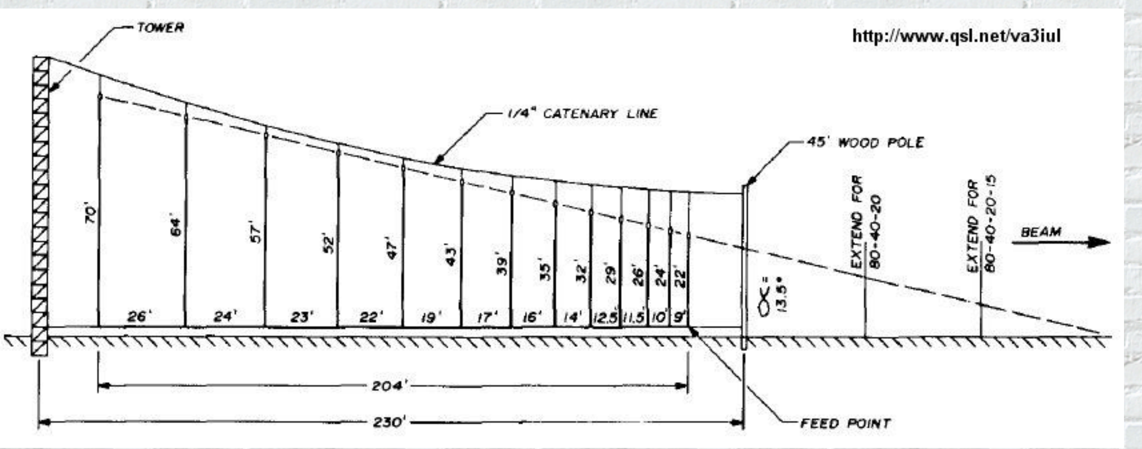
TOWER
http://www.qsl.net/va3iul
1/4" CATENARY LINE
45' WOOD POLE
70'
64'
57'
52'
47'
43'
39'
35'
32'
29'
26'
24'
22'
26'
24'
23'
22'
19'
17'
16'
14'
12.5
11.5
10'
9'
α = 13.5°
EXTEND FOR 80-40-20
EXTEND FOR 80-40-20-15
BEAM
204'
230'
FEED POINT

---

226 - Inverted Vertical Log Periodic Antenna for 40m

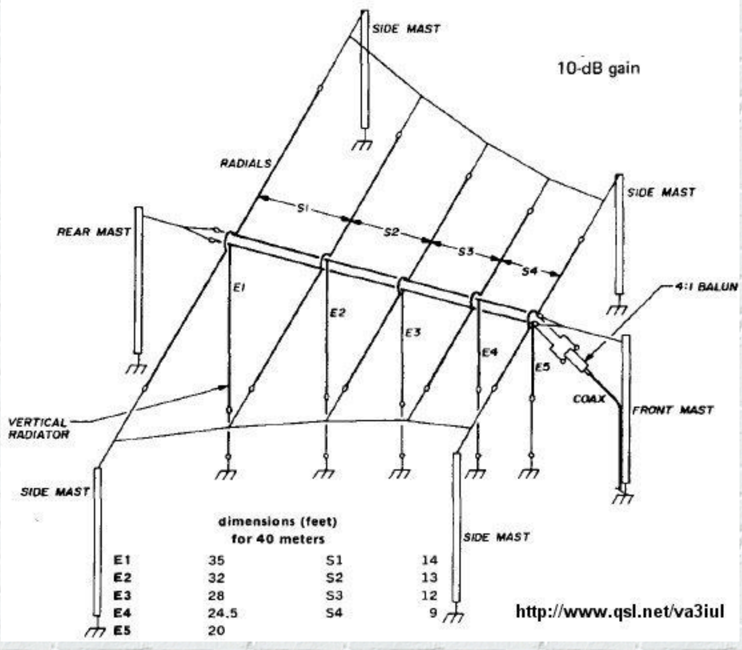
SIDE MAST
10-dB gain
RADIALS
REAR MAST
S1
S2
S3
S4
SIDE MAST
4:1 BALUN
E1
E2
E3
E4
E5
COAX
FRONT MAST
VERTICAL RADIATOR
SIDE MAST
SIDE MAST

dimensions (feet)
for 40 meters

E1	35	S1	14
E2	32	S2	13
E3	28	S3	12
E4	24.5	S4	9
E5	20		

http://www.qsl.net/va3iul

227 - Double Extended Zepp Antenna Gain=7dB for 15m

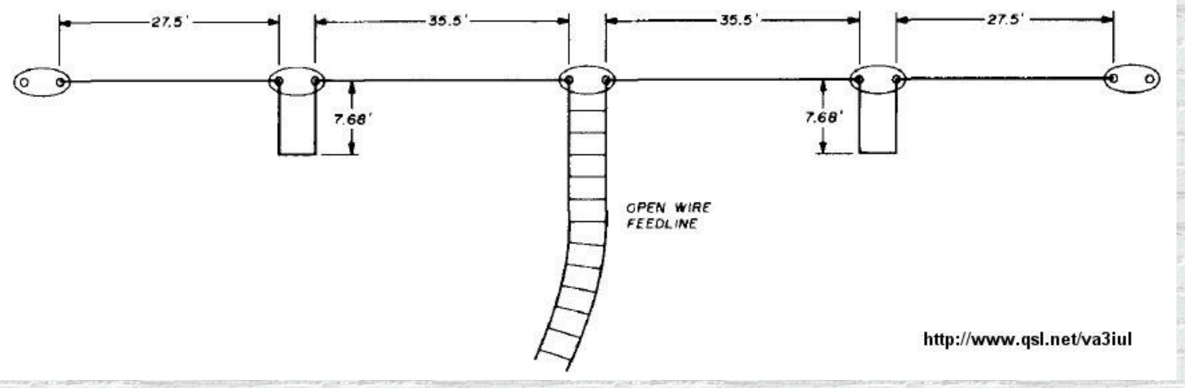
27.5'
35.5'
35.5'
27.5'
7.68'
7.68'
OPEN WIRE
FEEDLINE

http://www.qsl.net/va3iul

228 - Double Lazy-H Antenna G=10dB for 70cm

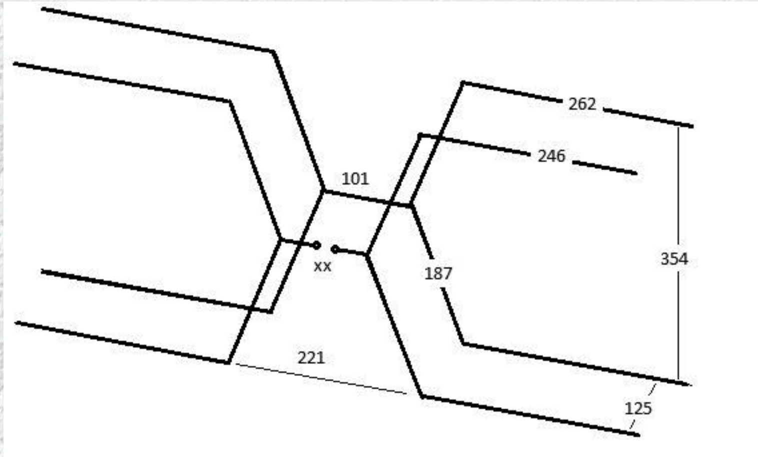
262
246
101
xx
187
354
221
125

10.8 dBd 21 dB F/B
xx - 50 Ohm, distance ~5mm
SWR = 1 @ 435 MHz
Material 3-5 mm dia aluminium

http://www.qsl.net/va3iul

229 - Wire Log Periodic Antenna G=12dB for 15m, 20m

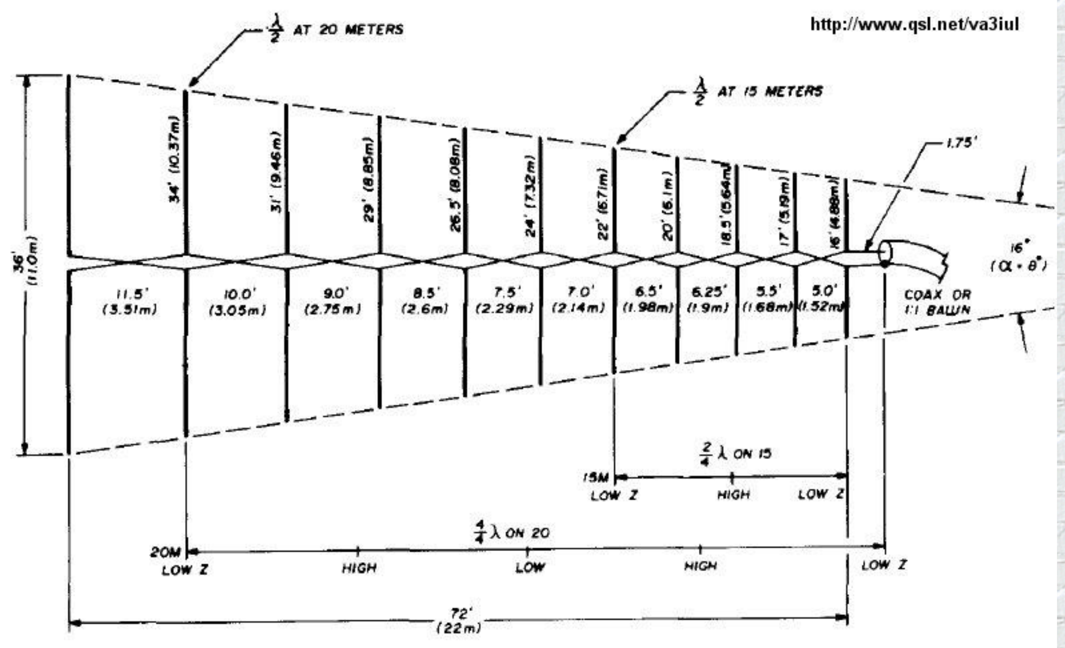
http://www.qsl.net/va3iul
λ/2 AT 20 METERS
λ/2 AT 15 METERS
1.75'
36' (11.0m)
34' (10.37m)
31' (9.46m)
29' (8.85m)
26.5' (8.08m)
24' (7.32m)
22' (6.71m)
20' (6.1m)
18.5' (5.64m)
17' (5.19m)
16' (4.88m)
11.5' (3.51m)
10.0' (3.05m)
9.0' (2.75m)
8.5' (2.6m)
7.5' (2.29m)
7.0' (2.14m)
6.5' (1.98m)
6.25' (1.9m)
5.5' (1.68m)
5.0' (1.52m)
COAX OR 1:1 BALUN
16° (α • 8°)
2/4 λ ON 15
15M LOW Z
HIGH
LOW Z
4/4 λ ON 20
20M LOW Z
HIGH
LOW
HIGH
LOW Z
72' (22m)

230 - Five-Elements Vertical Log Periodic for 80m, 40m, 20m

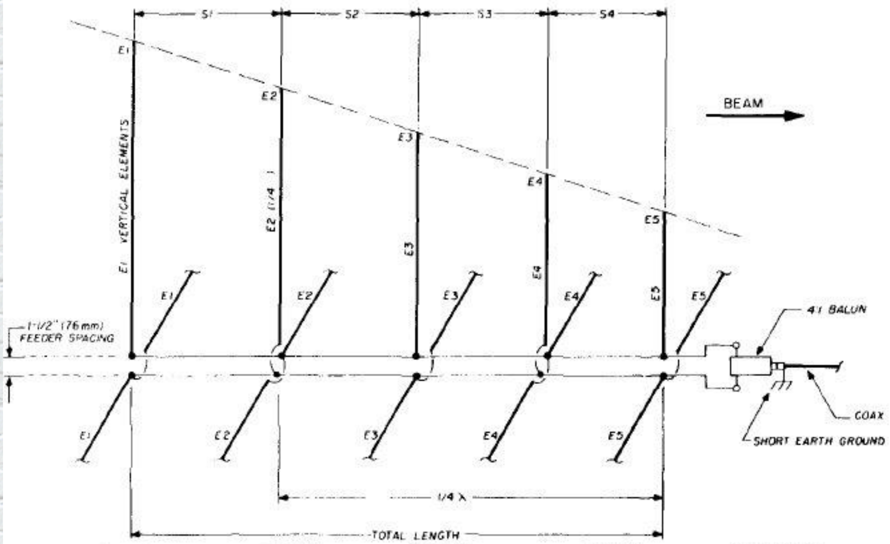
S1
S2
S3
S4
E1
E2
E3
E4
E5
BEAM
E1 VERTICAL ELEMENTS
E2 (1/4 λ)
E3
E4
E5
E1
E2
E3
E4
E5
1-1/2" (76mm) FEEDER SPACING
4:1 BALUN
COAX
SHORT EARTH GROUND
E1
E2
E3
E4
E5
1/4 λ
TOTAL LENGTH

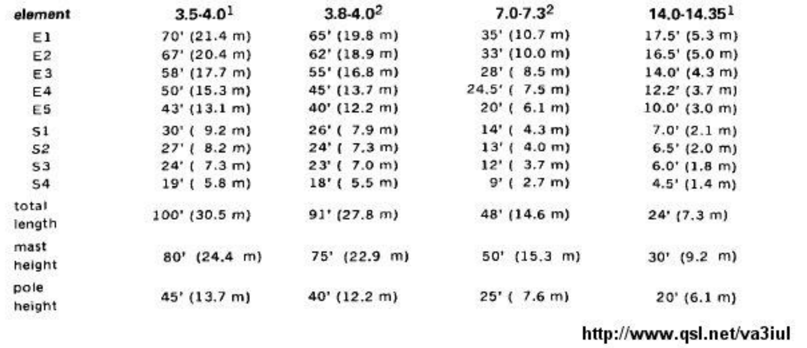

element	3.5-4.0[1]	3.8-4.0[2]	7.0-7.3[2]	14.0-14.35[1]
E1	70' (21.4 m)	65' (19.8 m)	35' (10.7 m)	17.5' (5.3 m)
E2	67' (20.4 m)	62' (18.9 m)	33' (10.0 m)	16.5' (5.0 m)
E3	58' (17.7 m)	55' (16.8 m)	28' (8.5 m)	14.0' (4.3 m)
E4	50' (15.3 m)	45' (13.7 m)	24.5' (7.5 m)	12.2' (3.7 m)
E5	43' (13.1 m)	40' (12.2 m)	20' (6.1 m)	10.0' (3.0 m)
S1	30' (9.2 m)	26' (7.9 m)	14' (4.3 m)	7.0' (2.1 m)
S2	27' (8.2 m)	24' (7.3 m)	13' (4.0 m)	6.5' (2.0 m)
S3	24' (7.3 m)	23' (7.0 m)	12' (3.7 m)	6.0' (1.8 m)
S4	19' (5.8 m)	18' (5.5 m)	9' (2.7 m)	4.5' (1.4 m)
total length	100' (30.5 m)	91' (27.8 m)	48' (14.6 m)	24' (7.3 m)
mast height	80' (24.4 m)	75' (22.9 m)	50' (15.3 m)	30' (9.2 m)
pole height	45' (13.7 m)	40' (12.2 m)	25' (7.6 m)	20' (6.1 m)

http://www.qsl.net/va3iul

231 - Sloping Delta Loop Antenna for 40m

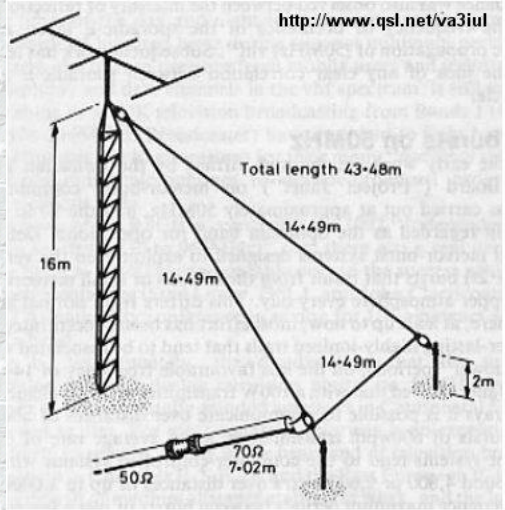
http://www.qsl.net/va3iul
Total length 43·48m
14·49m
16m
14·49m
14·49m
2m
70Ω
7·02m
50Ω

---

232 - Doublet Antenna for 160m

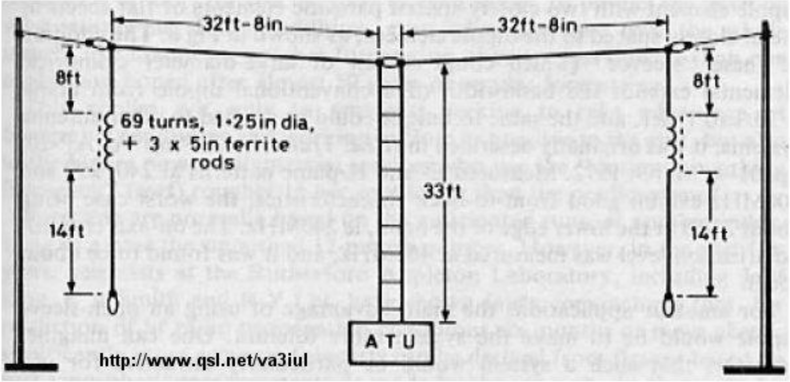
32ft-8in
32ft-8in
8ft
8ft
69 turns, 1·25in dia, + 3 x 5in ferrite rods
33ft
14ft
14ft
A T U
http://www.qsl.net/va3iul

---

233 - Log Periodic Wire Beam for 40m

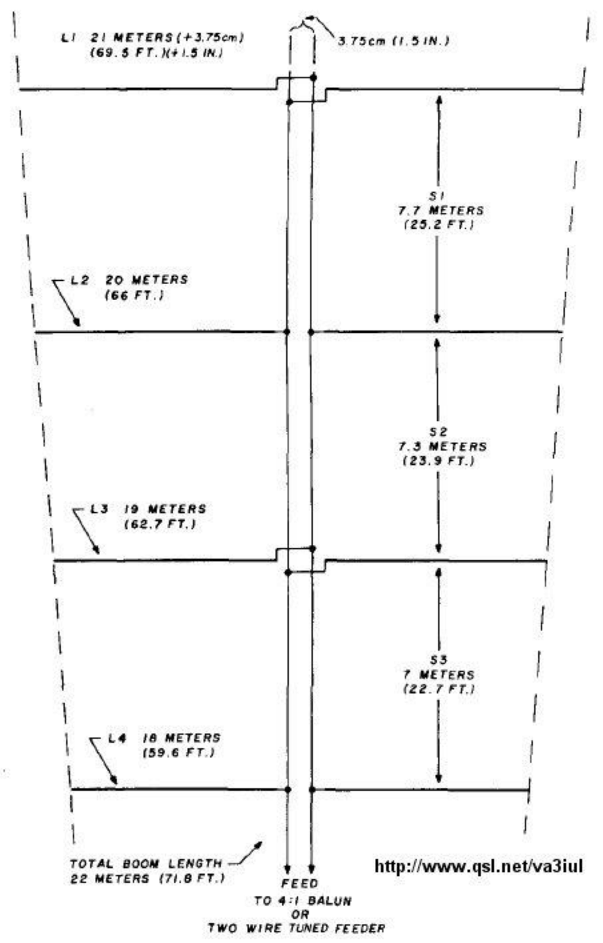
L1 21 METERS (+3.75cm) (69.5 FT.)(+1.5 IN.)
3.75cm (1.5 IN.)
S1 7.7 METERS (25.2 FT.)
L2 20 METERS (66 FT.)
S2 7.3 METERS (23.9 FT.)
L3 19 METERS (62.7 FT.)
S3 7 METERS (22.7 FT.)
L4 18 METERS (59.6 FT.)
TOTAL BOOM LENGTH 22 METERS (71.8 FT.)
FEED TO 4:1 BALUN OR TWO WIRE TUNED FEEDER
http://www.qsl.net/va3iul

234 - G3LDO Wire Beam Antenna for 20m, 15m, 10m

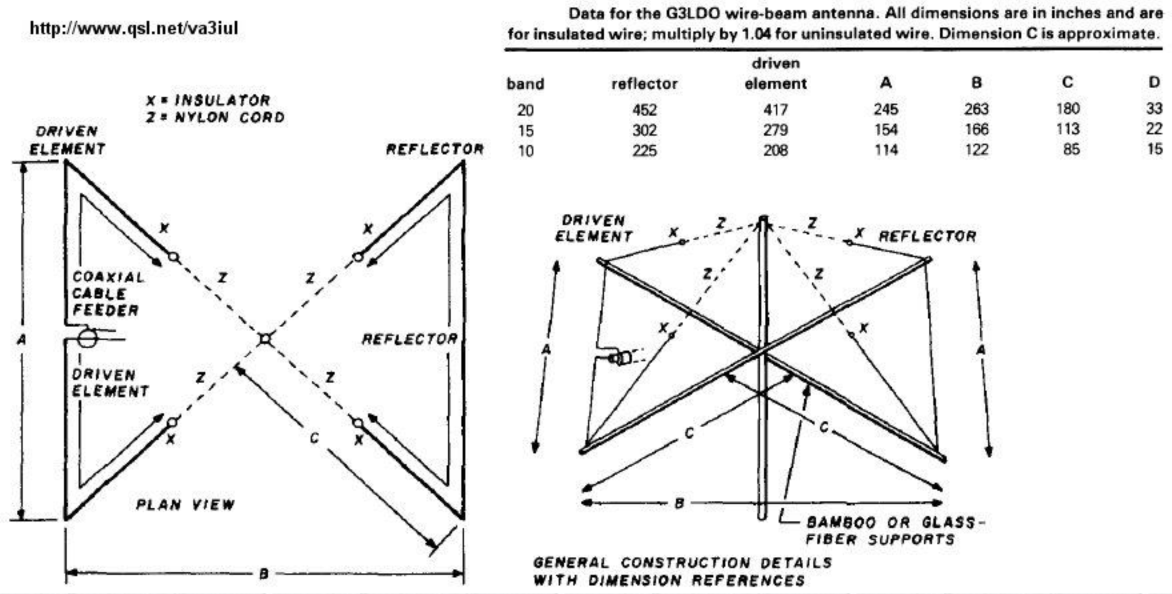

Data for the G3LDO wire-beam antenna. All dimensions are in inches and are for insulated wire; multiply by 1.04 for uninsulated wire. Dimension C is approximate.

band	reflector	driven element	A	B	C	D
20	452	417	245	263	180	33
15	302	279	154	166	113	22
10	225	208	114	122	85	15

http://www.qsl.net/va3iul
X = INSULATOR
Z = NYLON CORD
DRIVEN ELEMENT
REFLECTOR
COAXIAL CABLE FEEDER
DRIVEN ELEMENT
REFLECTOR
A
B
C
X
Z
PLAN VIEW
DRIVEN ELEMENT
REFLECTOR
BAMBOO OR GLASS-FIBER SUPPORTS
GENERAL CONSTRUCTION DETAILS WITH DIMENSION REFERENCES

235 - Phased Delta Loop Array for 20m

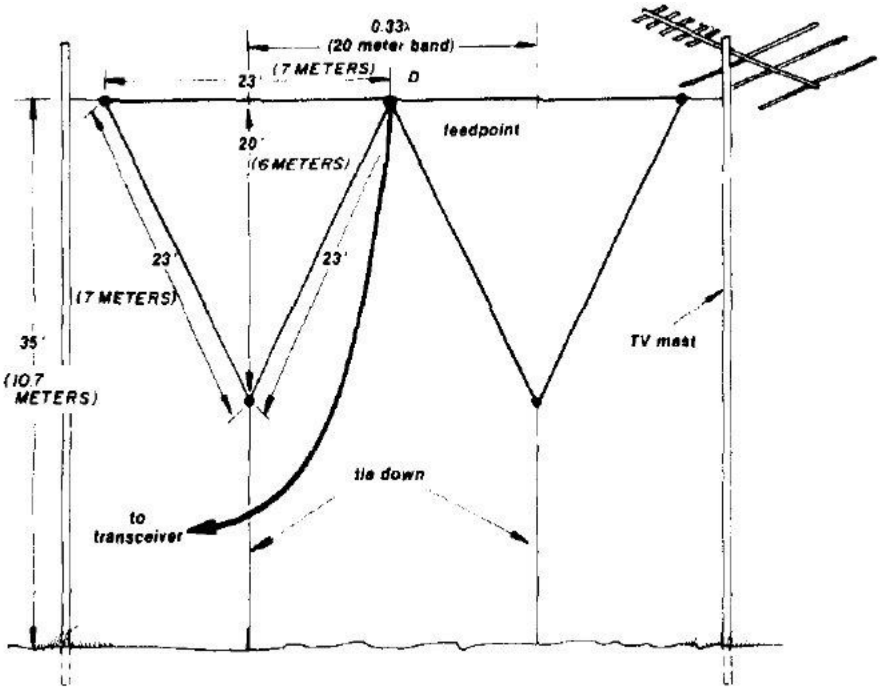
0.33λ
(20 meter band)
23'
(7 METERS)
D
20'
(6 METERS)
feedpoint
23'
23'
(7 METERS)
35'
(10.7 METERS)
TV mast
tie down
to transceiver

spacing wavelength	feet	(meters)	gain (dB)
0.33	23	(7)	2.0
0.40	27	(8)	2.8
0.50	34	(10)	4.0
0.65	44	(13)	4.8

http://www.qsl.net/va3iul

---

236 - Dual Band V-dipole for 80m, 40m

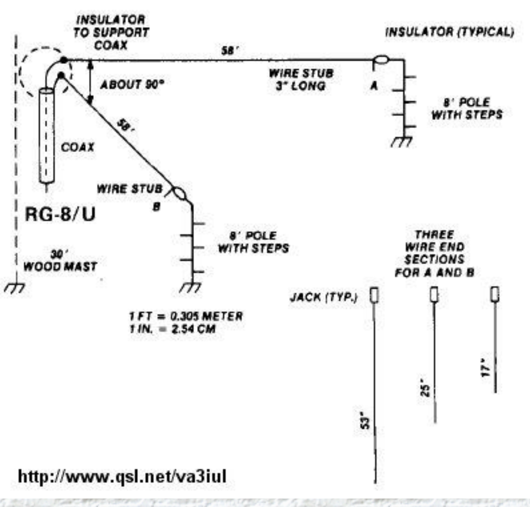
INSULATOR TO SUPPORT COAX
58'
INSULATOR (TYPICAL)
ABOUT 90°
WIRE STUB 3" LONG
A
8' POLE WITH STEPS
58'
COAX
WIRE STUB
B
RG-8/U
30' WOOD MAST
8' POLE WITH STEPS
THREE WIRE END SECTIONS FOR A AND B
JACK (TYP.)
53"
25"
17"
1 FT = 0.305 METER
1 IN. = 2.54 CM

http://www.qsl.net/va3iul

---

237 - Collinear Array 5dB-Gain for 15m

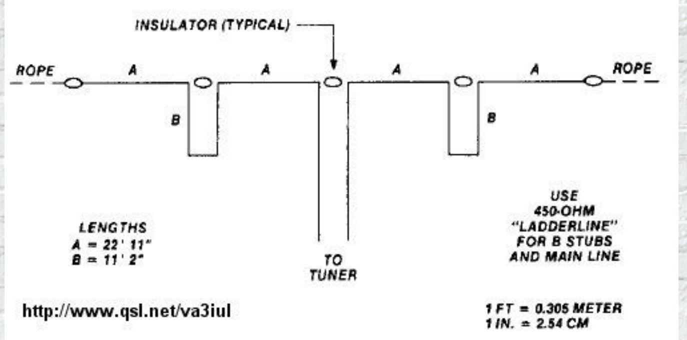
INSULATOR (TYPICAL)
ROPE
A
A
A
A
ROPE
B
B
TO TUNER
USE 450-OHM "LADDERLINE" FOR B STUBS AND MAIN LINE
1 FT = 0.305 METER
1 IN. = 2.54 CM

LENGTHS
A = 22' 11"
B = 11' 2"

http://www.qsl.net/va3iul

238 - Two-Band Quad Loop Antenna 80m, 40m

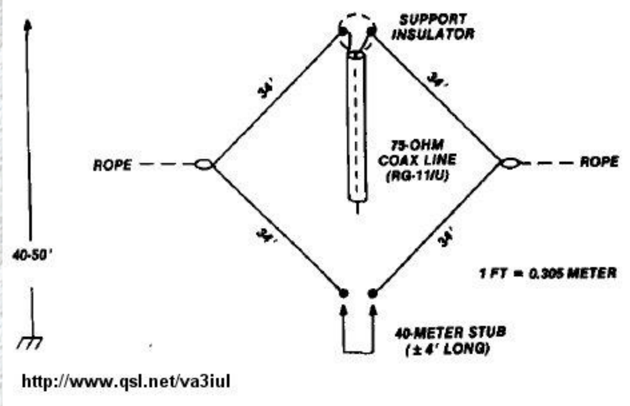
SUPPORT INSULATOR
34'
34'
ROPE
75-OHM COAX LINE (RG-11/U)
ROPE
40-50'
34'
34'
1 FT = 0.305 METER
40-METER STUB (±4' LONG)
http://www.qsl.net/va3iul

---

239 - K7CW Quad Antenna

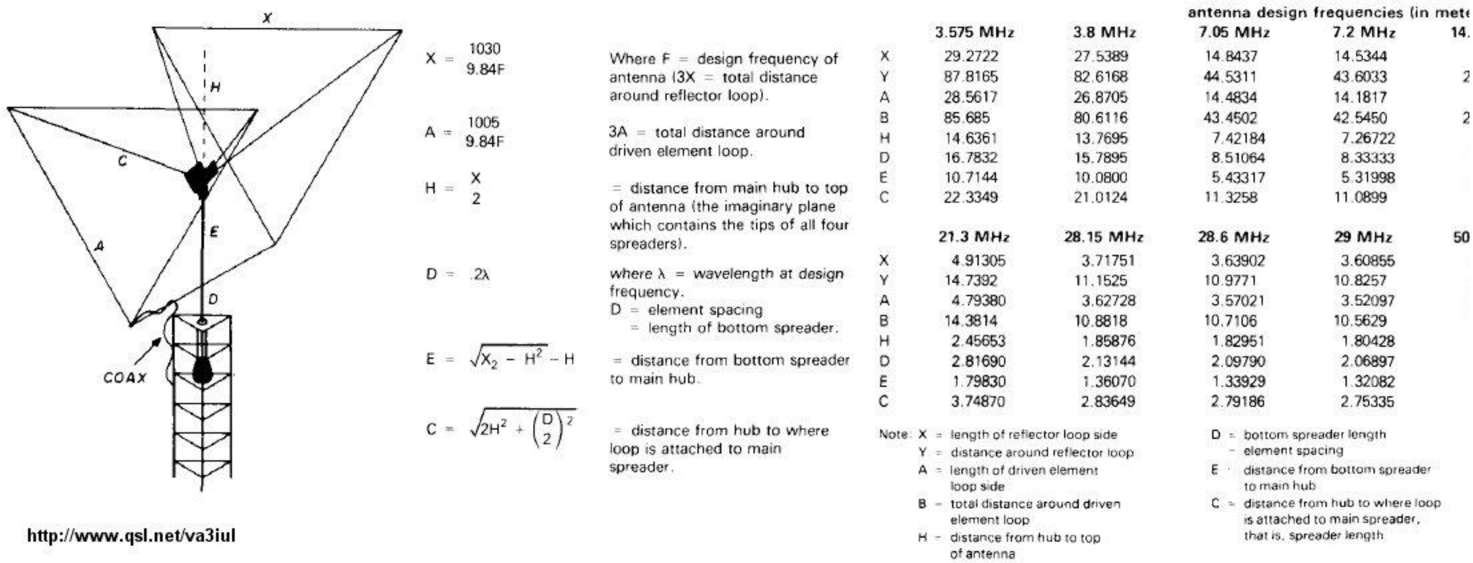
X
H
C
A
E
D
COAX
http://www.qsl.net/va3iul

$X = \frac{1030}{9.84F}$ Where F = design frequency of antenna (3X = total distance around reflector loop).

$A = \frac{1005}{9.84F}$ 3A = total distance around driven element loop.

$H = \frac{X}{2}$ = distance from main hub to top of antenna (the imaginary plane which contains the tips of all four spreaders).

$D = .2\lambda$ where λ = wavelength at design frequency.
D = element spacing
= length of bottom spreader.

$E = \sqrt{X_2 - H^2} - H$ = distance from bottom spreader to main hub.

$C = \sqrt{2H^2 + \left(\frac{D}{2}\right)^2}$ = distance from hub to where loop is attached to main spreader.

antenna design frequencies (in mete

	3.575 MHz	3.8 MHz	7.05 MHz	7.2 MHz	14.
X	29.2722	27.5389	14.8437	14.5344	
Y	87.8165	82.6168	44.5311	43.6033	2
A	28.5617	26.8705	14.4834	14.1817	
B	85.685	80.6116	43.4502	42.5450	2
H	14.6361	13.7695	7.42184	7.26722	
D	16.7832	15.7895	8.51064	8.33333	
E	10.7144	10.0800	5.43317	5.31998	
C	22.3349	21.0124	11.3258	11.0899	

	21.3 MHz	28.15 MHz	28.6 MHz	29 MHz	50
X	4.91305	3.71751	3.63902	3.60855	
Y	14.7392	11.1525	10.9771	10.8257	
A	4.79380	3.62728	3.57021	3.52097	
B	14.3814	10.8818	10.7106	10.5629	
H	2.45653	1.85876	1.82951	1.80428	
D	2.81690	2.13144	2.09790	2.06897	
E	1.79830	1.36070	1.33929	1.32082	
C	3.74870	2.83649	2.79186	2.75335	

Note: X = length of reflector loop side
Y = distance around reflector loop
A = length of driven element loop side
B = total distance around driven element loop
H = distance from hub to top of antenna
D = bottom spreader length
= element spacing
E = distance from bottom spreader to main hub
C = distance from hub to where loop is attached to main spreader, that is, spreader length

---

240 - Broadband Short Dipole for 80m

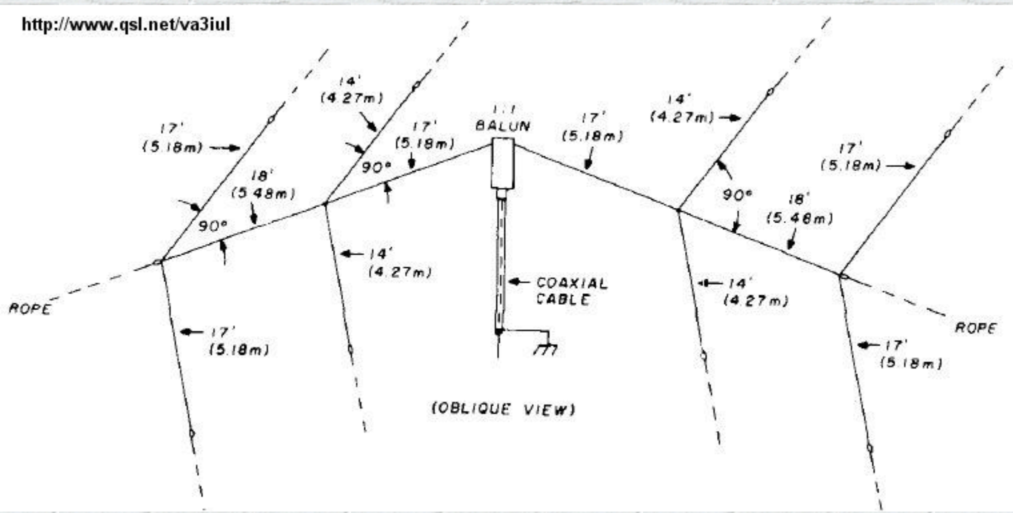
http://www.qsl.net/va3iul
14' (4.27m)
17' (5.18m)
18' (5.48m)
90°
17' (5.18m)
90°
1:1 BALUN
17' (5.18m)
14' (4.27m)
17' (5.18m)
90°
18' (5.48m)
14' (4.27m)
COAXIAL CABLE
14' (4.27m)
ROPE
17' (5.18m)
ROPE
17' (5.18m)
(OBLIQUE VIEW)

---

241 - Off-Center Fed Dipole for 40m, 20m, 10m

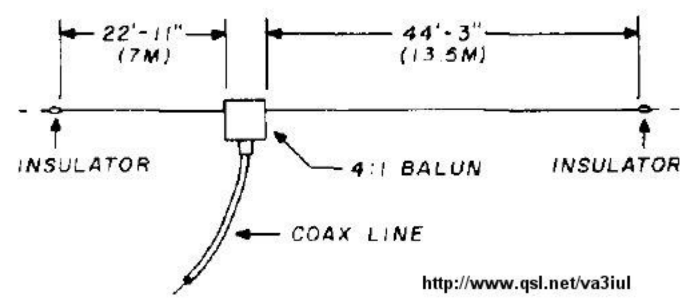
22'-11"
(7M)
44'-3"
(13.5M)
INSULATOR
4:1 BALUN
INSULATOR
COAX LINE
http://www.qsl.net/va3iul

---

242 - Broadband Sloper Antenna for 80m

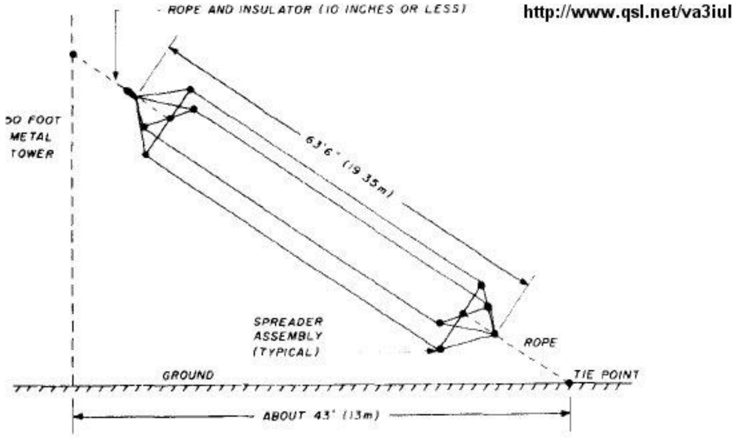
ROPE AND INSULATOR (10 INCHES OR LESS)
http://www.qsl.net/va3iul
50 FOOT
METAL
TOWER
63'6" (19.35m)
SPREADER
ASSEMBLY
(TYPICAL)
ROPE
GROUND
TIE POINT
ABOUT 43' (13m)

80-meter WD4FAB broadband sloper. Spreader assemblies are made of 36-inch long fiberglass tubes. The four antenna wires are tied together about 24 inches past each spreader assembly. A nylon line may be required from the crossover point of the spreader to the outer insulator to prevent bowing. The SWR plot of the antenna falls below 1.4-to-1 at 3.5 MHz; 1.25-to-1 at 3.8 MHz and 1.4-to-1 at 4.0 MHz. Feedpoint resistance of the antenna may be changed by altering the slope angle. The resonant frequency is adjusted by changing length — no climbing necessary!

---

243 - Two-Frequency Dipole Antenna for 80m

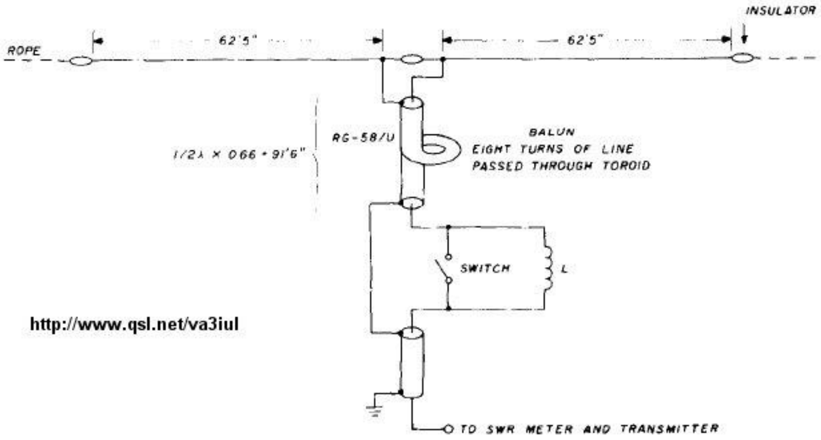
INSULATOR
ROPE
62'5"
62'5"
RG-58/U
BALUN
EIGHT TURNS OF LINE
PASSED THROUGH TOROID
1/2λ × 0.66 = 91'6"
SWITCH
L
http://www.qsl.net/va3iul
TO SWR METER AND TRANSMITTER

The W6TC two-frequency dipole for 80 meters. Antenna is cut for resonance at high frequency end of band. A single loading coil for operation at the low frequency end of the band is placed an electrical half-wavelength down the transmission line. The line is decoupled from the antenna by a simple balun placed directly at the antenna insulator. Toroid is powdered iron material T-200 (2″ OD) (1.25″ ID), 2-mix (red, μ = 10 (Amidon T-200-2), or equivalent.

---

244 - Top Loaded Vertical Antenna for 160m

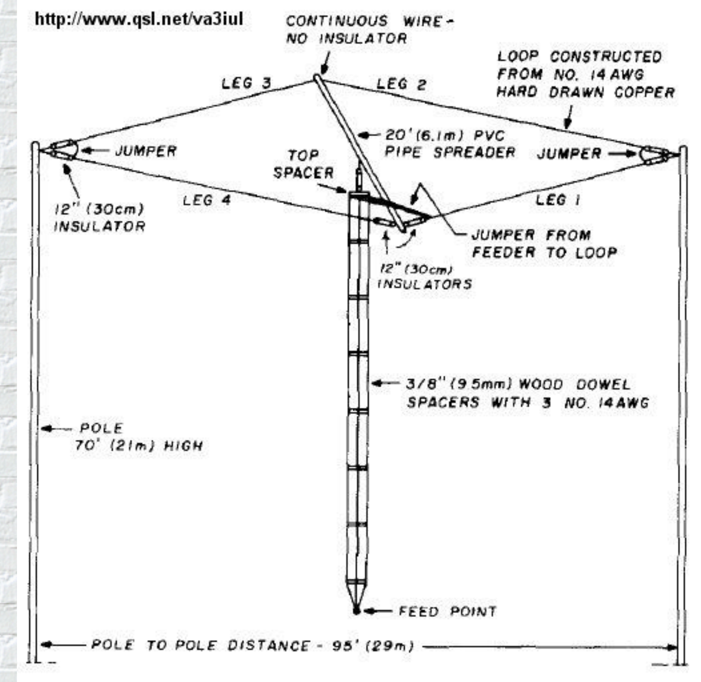
http://www.qsl.net/va3iul
CONTINUOUS WIRE - NO INSULATOR
LOOP CONSTRUCTED FROM NO. 14 AWG HARD DRAWN COPPER
LEG 3
LEG 2
20' (6.1m) PVC PIPE SPREADER
JUMPER
JUMPER
TOP SPACER
12" (30cm) INSULATOR
LEG 4
LEG 1
JUMPER FROM FEEDER TO LOOP
12" (30cm) INSULATORS
3/8" (9.5mm) WOOD DOWEL SPACERS WITH 3 NO. 14AWG
POLE 70' (21m) HIGH
FEED POINT
POLE TO POLE DISTANCE - 95' (29m)

---

245 - Two Loop Beam Antenna for 20m, 15m

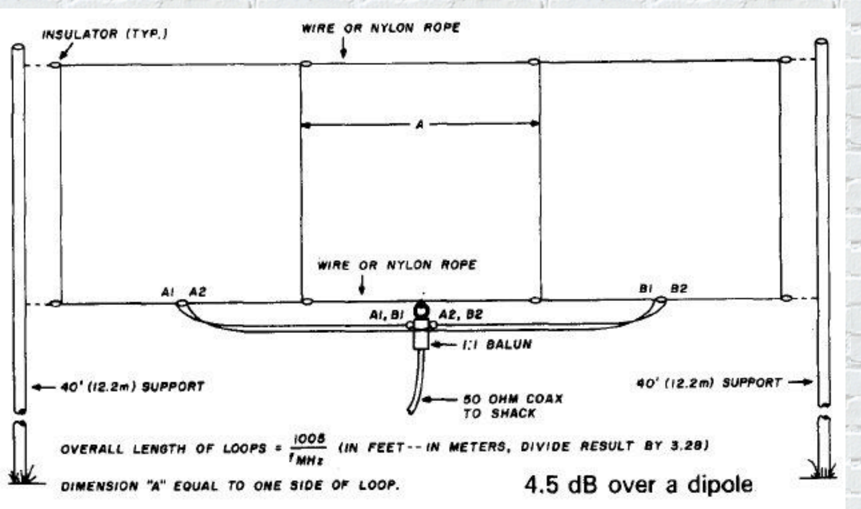
INSULATOR (TYP.)
WIRE OR NYLON ROPE
A
WIRE OR NYLON ROPE
A1 A2
B1 B2
A1, B1
A2, B2
1:1 BALUN
40' (12.2m) SUPPORT
40' (12.2m) SUPPORT
50 OHM COAX TO SHACK
OVERALL LENGTH OF LOOPS = 1005/fMHz (IN FEET--IN METERS, DIVIDE RESULT BY 3.28)
DIMENSION "A" EQUAL TO ONE SIDE OF LOOP.
4.5 dB over a dipole

The W2TBZ loop beam. The phasing lines are made of 300-ohm TV line. Length of phasing line = $\frac{468}{f_{MHz}}$ × velocity factor of 300-ohm line. Balun is supported at center of lower nylon rope.

http://www.qsl.net/va3iul

---

246 - Corner-Fed Delta Loop Antenna for 80m, 40m, 20m

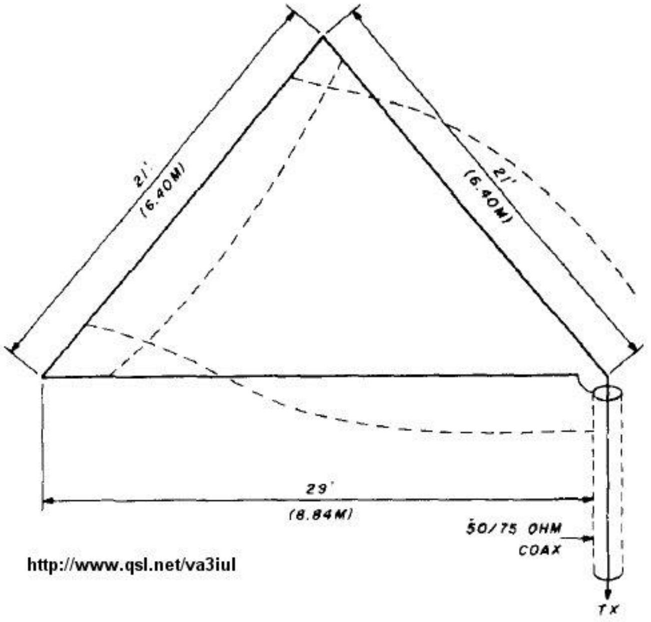
21'
(6.40M)
21'
(6.40M)
29'
(8.84M)
50/75 OHM
COAX
http://www.qsl.net/va3iul
TX

DIMENSIONS FOR FULL-SIZE CORNER-FED DELTA LOOPS

frequency (MHz)	total loop circumference (feet)	(meters)	length of each side (feet)	(meters)	length of base (feet)	(meters)
3.600	279 ft. 2 in.	(85.09)	82 ft.	(24.99)	115 ft. 2 in.	(35.10)
7.075	142 ft. 0 in.	(43.28)	42 ft.	(12.80)	58 ft. 0 in.	(17.68)
14.150	71 ft. 0 in.	(21.64)	21 ft.	(6.40)	29 ft. 0 in.	(8.84)

247 - Loaded Delta Loop Antenna for 40m, 20m

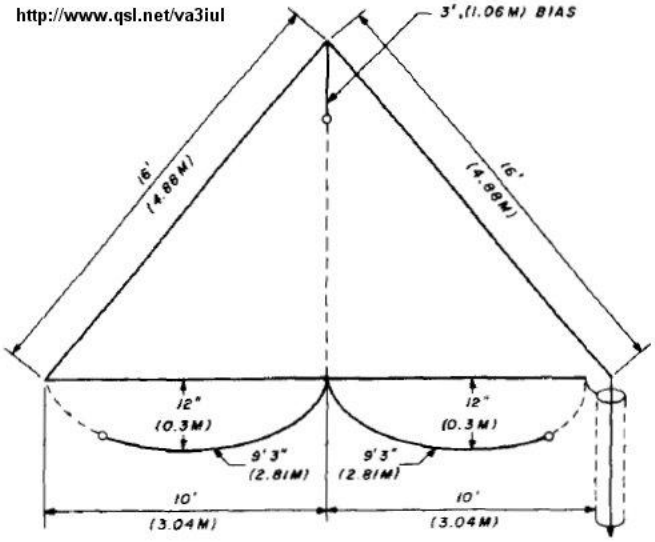
http://www.qsl.net/va3iul
3',(1.06M) BIAS
16'
(4.88M)
16'
(4.88M)
12"
(0.3M)
12"
(0.3M)
9'3"
(2.81M)
9'3"
(2.81M)
10'
(3.04M)
10'
(3.04M)

The final loaded loop for 14 MHz. All dimensions are doubled for operation on 7 MHz. Base loading wires are each 18 feet 4 inches (5.59 meters). The top loading wire is made taut by securing to base with thin light nylon cord.

248 - Compact Size Dipole for 80m

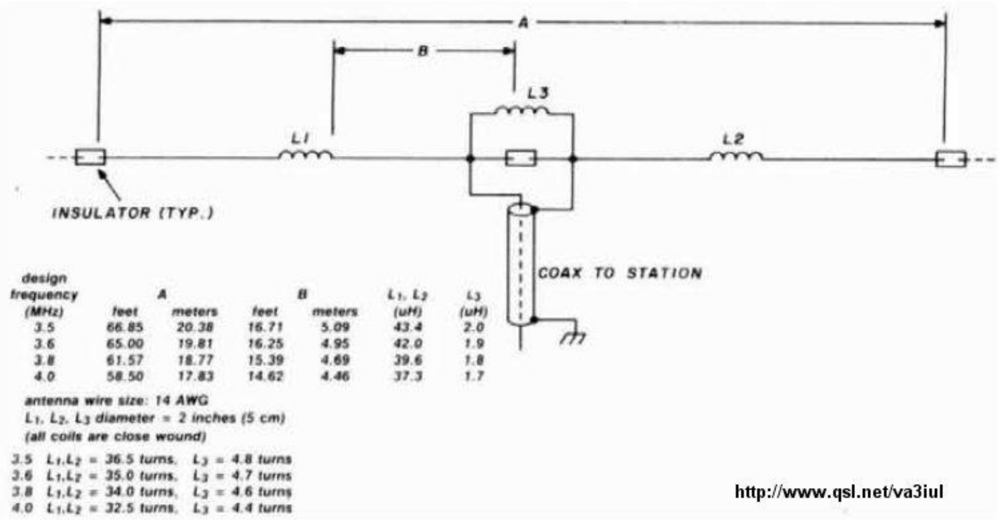
A
B
L3
L1
L2
INSULATOR (TYP.)
COAX TO STATION

design frequency (MHz)	A feet	A meters	B feet	B meters	L_1, L_2 (uH)	L_3 (uH)
3.5	66.85	20.38	16.71	5.09	43.4	2.0
3.6	65.00	19.81	16.25	4.95	42.0	1.9
3.8	61.57	18.77	15.39	4.69	39.6	1.8
4.0	58.50	17.83	14.62	4.46	37.3	1.7

antenna wire size: 14 AWG
L_1, L_2, L_3 diameter = 2 inches (5 cm)
(all coils are close wound)

3.5 L_1,L_2 = 36.5 turns, L_3 = 4.8 turns
3.6 L_1,L_2 = 35.0 turns, L_3 = 4.7 turns
3.8 L_1,L_2 = 34.0 turns, L_3 = 4.6 turns
4.0 L_1,L_2 = 32.5 turns, L_3 = 4.4 turns

http://www.qsl.net/va3iul

249 - JAWS 3dB Gain Antenna for 40m

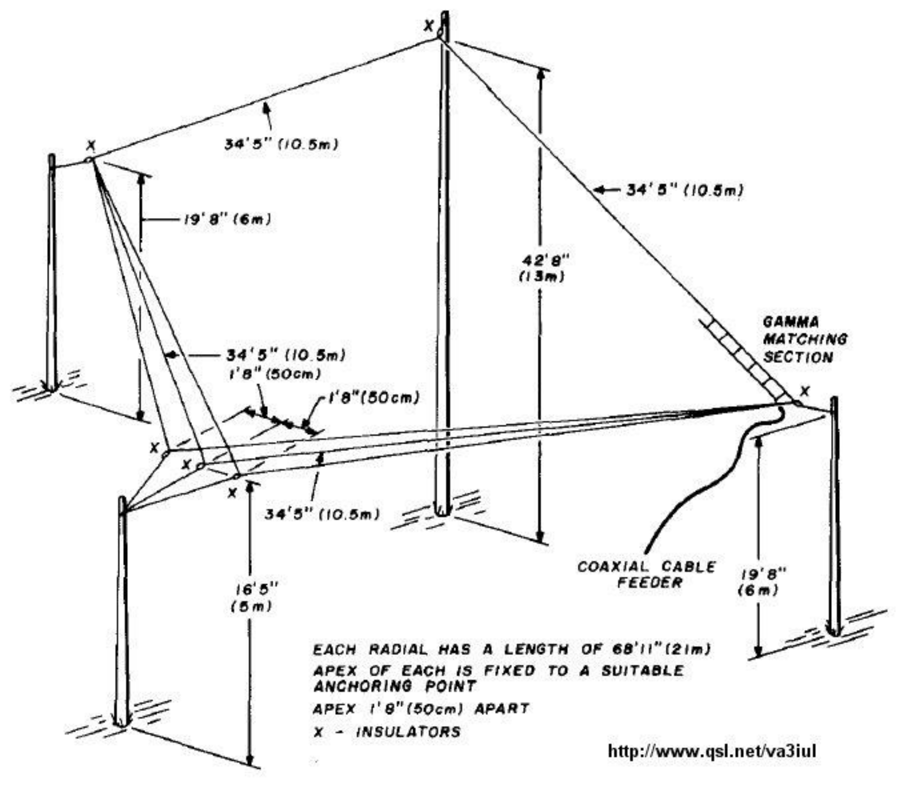
X
34'5" (10.5m)
34'5" (10.5m)
19'8" (6m)
42'8" (13m)
GAMMA MATCHING SECTION
34'5" (10.5m)
1'8" (50cm)
1'8" (50cm)
34'5" (10.5m)
16'5" (5m)
COAXIAL CABLE FEEDER
19'8" (6m)

EACH RADIAL HAS A LENGTH OF 68'11" (21m)
APEX OF EACH IS FIXED TO A SUITABLE ANCHORING POINT
APEX 1'8" (50cm) APART
X - INSULATORS

http://www.qsl.net/va3iul

250 - Inverted-V Antenna for 160m

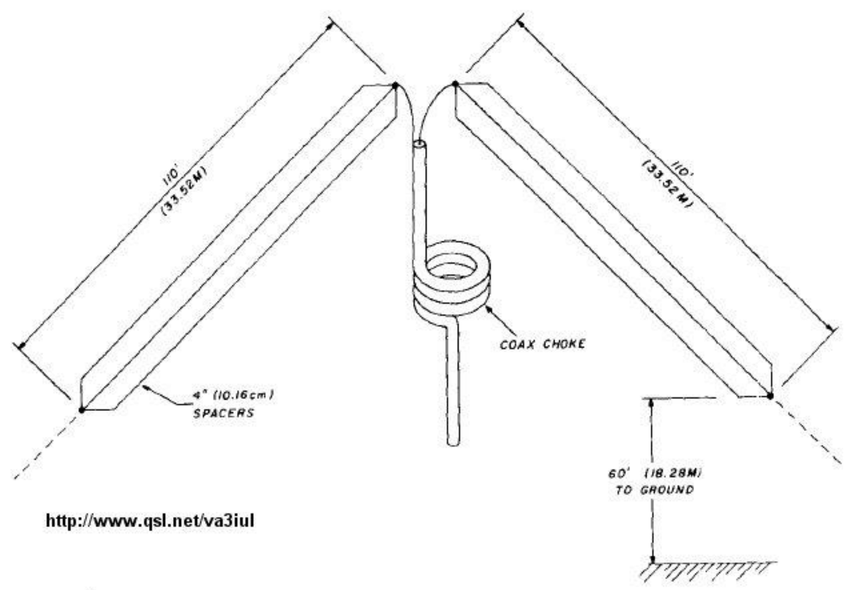
110'
(33.52M)
110'
(33.52M)
COAX CHOKE
4" (10.16cm)
SPACERS
60' (18.28M)
TO GROUND
http://www.qsl.net/va3iul

The inverted-V antenna for 160 meters at VE7BS. Wires all lie in the same plane. Apex is at 105 feet (32 meters).

---

251 - Lazy-U Antenna for 160m

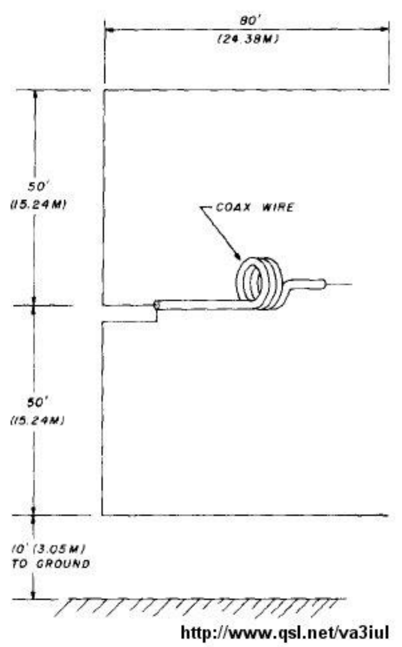
80'
(24.38M)
50'
(15.24M)
COAX WIRE
50'
(15.24M)
10' (3.05M)
TO GROUND
http://www.qsl.net/va3iul

---

252 - AF3V Compact Antenna for 160m

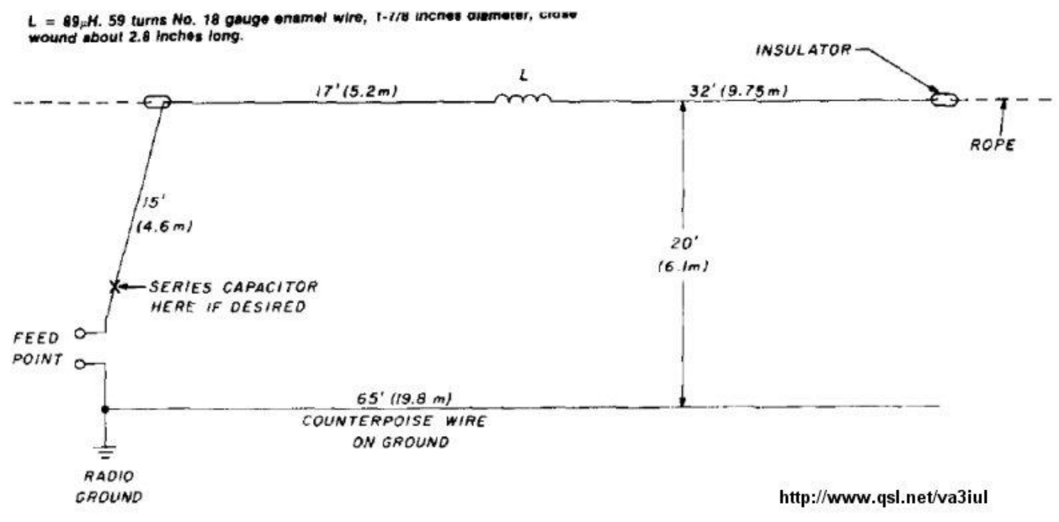

L = 89μH. 59 turns No. 18 gauge enamel wire, 1-7/8 inches diameter, close wound about 2.8 inches long.

INSULATOR
17' (5.2m)
L
32' (9.75m)
ROPE
15' (4.6m)
20' (6.1m)
SERIES CAPACITOR HERE IF DESIRED
FEED POINT
65' (19.8 m)
COUNTERPOISE WIRE ON GROUND
RADIO GROUND
http://www.qsl.net/va3iul

---

253 - Hot Beam Wire Antenna for 40m

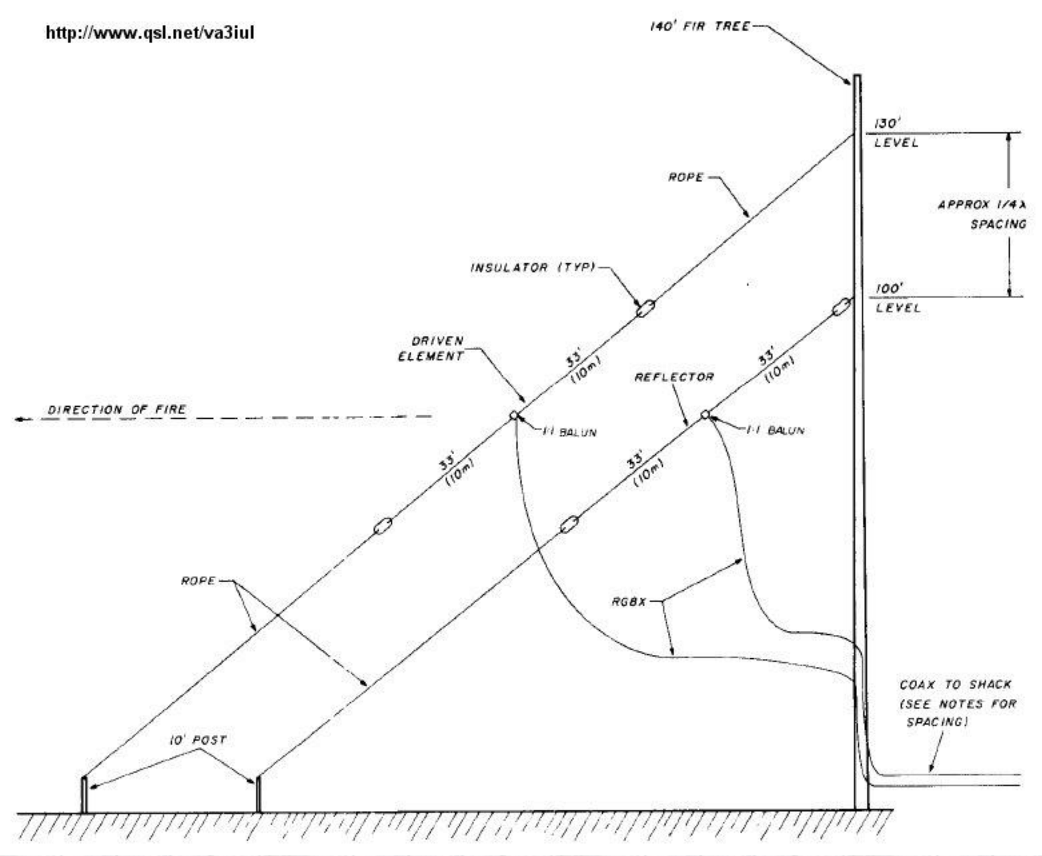

http://www.qsl.net/va3iul
140' FIR TREE
130' LEVEL
ROPE
APPROX 1/4λ SPACING
INSULATOR (TYP)
100' LEVEL
DRIVEN ELEMENT
33' (10m)
REFLECTOR
33' (10m)
DIRECTION OF FIRE
1:1 BALUN
1:1 BALUN
33' (10m)
33' (10m)
ROPE
RG8X
COAX TO SHACK (SEE NOTES FOR SPACING)
10' POST

---

254 - Sloping Delta Loop for 40m

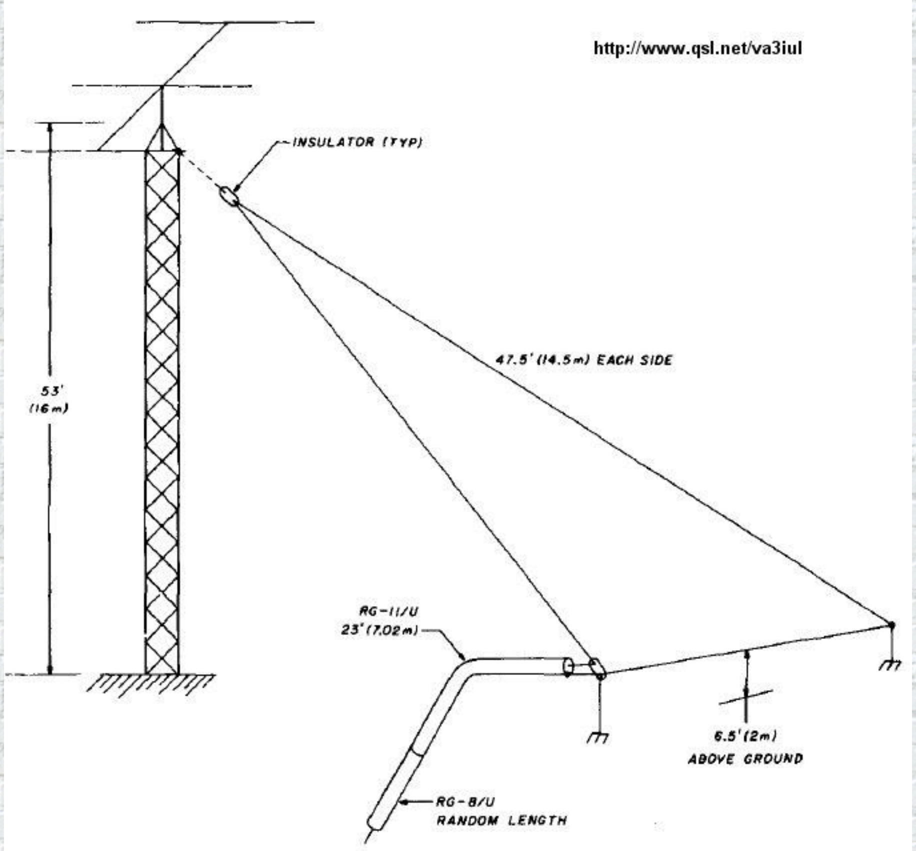
http://www.qsl.net/va3iul
INSULATOR (TYP)
53'
(16 m)
47.5' (14.5 m) EACH SIDE
RG-11/U
23' (7.02 m)
6.5' (2 m)
ABOVE GROUND
RG-8/U
RANDOM LENGTH

---

255 - K6FD Multiband Antenna with T-tuner

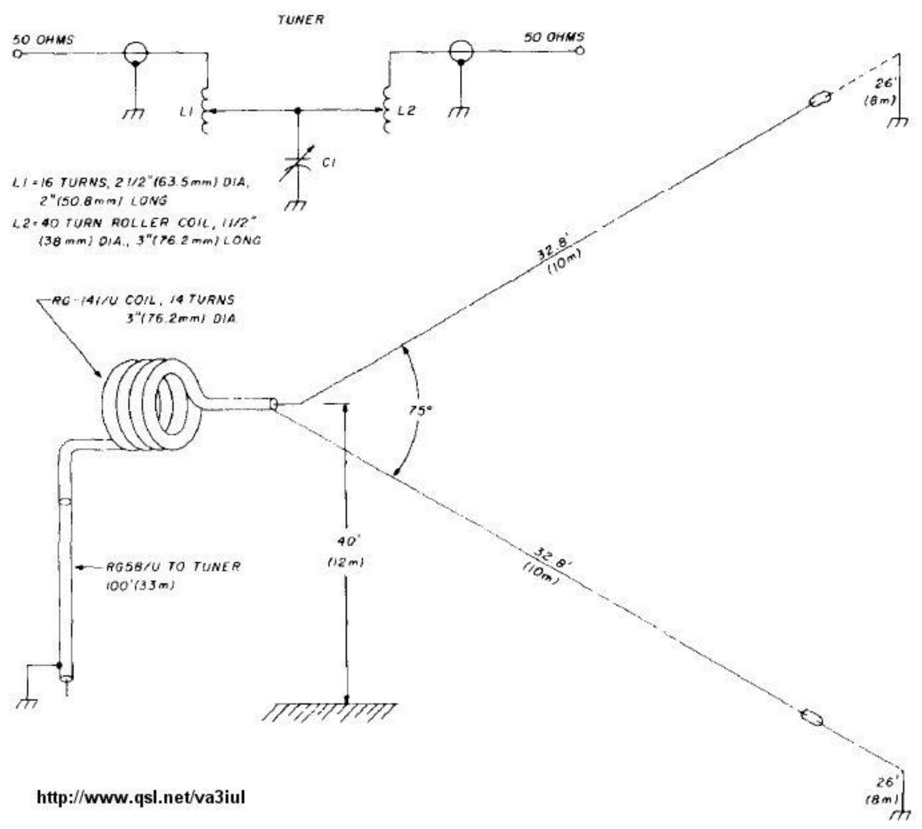
TUNER
50 OHMS
50 OHMS
L1
L2
C1
L1 = 16 TURNS, 2 1/2" (63.5 mm) DIA, 2" (50.8 mm) LONG
L2 = 40 TURN ROLLER COIL, 1 1/2" (38 mm) DIA., 3" (76.2 mm) LONG
RG-141/U COIL, 14 TURNS 3" (76.2 mm) DIA
26' (8 m)
32.8' (10 m)
75°
40' (12 m)
32.8' (10 m)
RG58/U TO TUNER 100' (33 m)
26' (8 m)
http://www.qsl.net/va3iul

---

256 - Dual Band Dipole Antenna for 17m, 10m

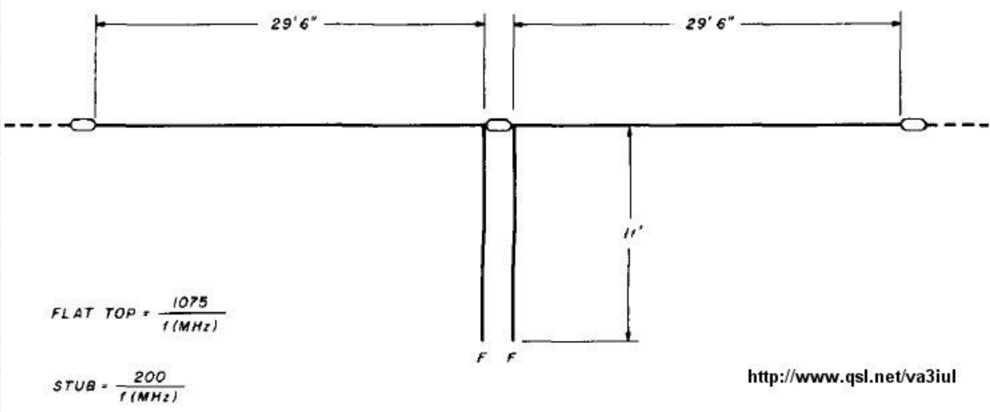
29'6"
29'6"
11'
FLAT TOP = 1075 / f (MHz)
STUB = 200 / f (MHz)
F F
http://www.qsl.net/va3iul

257 - Compact Travel Dipole Antenna for 10m, 15m, 20m

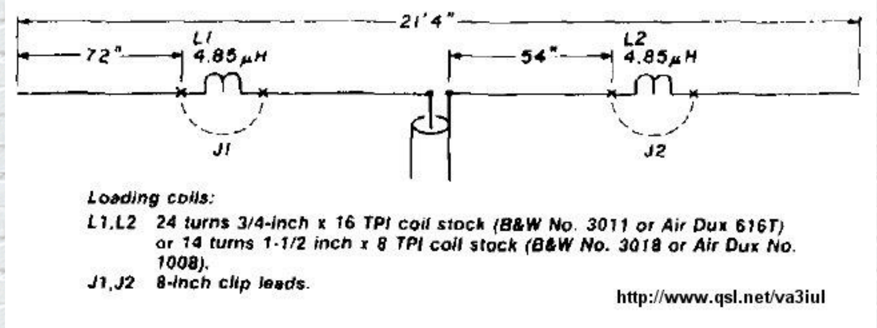
21'4"
72"
L1 4.85µH
54"
L2 4.85µH
J1
J2
http://www.qsl.net/va3iul

Loading coils:

L1,L2 24 turns 3/4-inch x 16 TPI coil stock (B&W No. 3011 or Air Dux 616T) or 14 turns 1-1/2 inch x 8 TPI coil stock (B&W No. 3018 or Air Dux No. 1008).

J1,J2 8-inch clip leads.

258 - End-Fire Antenna for 160m, 80m, 40m, 20m

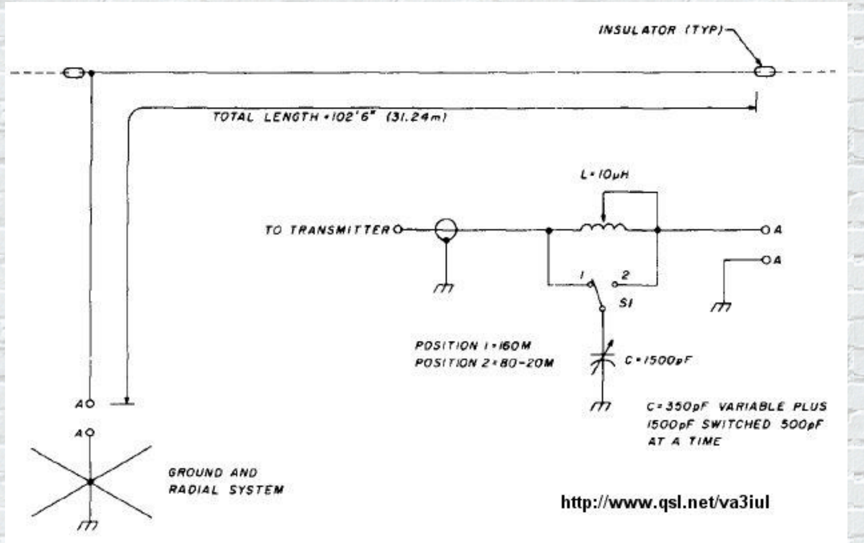
INSULATOR (TYP)
TOTAL LENGTH = 102'6" (31.24m)
L = 10µH
TO TRANSMITTER
A
A
1
2
S1
POSITION 1 = 160M
POSITION 2 = 80-20M
C = 1500pF
C = 350pF VARIABLE PLUS 1500pF SWITCHED 500pF AT A TIME
A
A
GROUND AND RADIAL SYSTEM
http://www.qsl.net/va3iul

259 - Four Band Loop for 40m, 20m, 15m, 10m

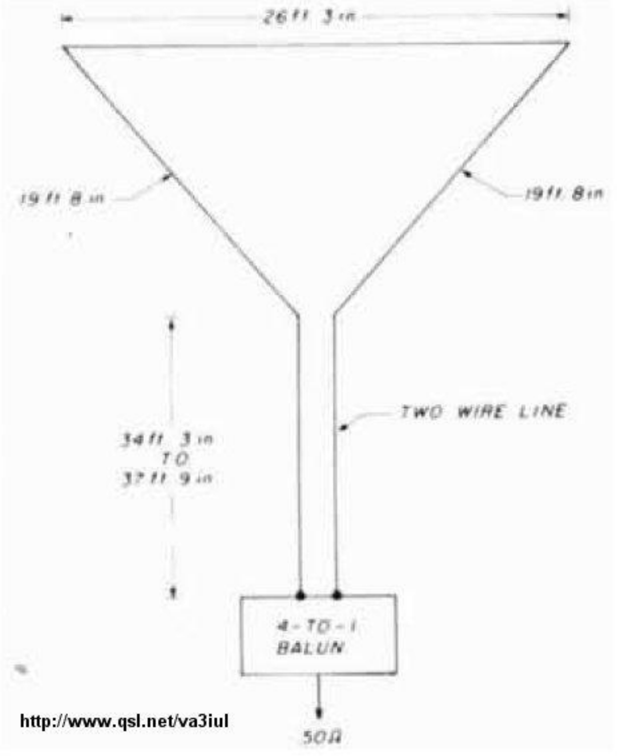
26 ft 3 in
19 ft 8 in
19 ft 8 in
34 ft 3 in
TO
37 ft 9 in
TWO WIRE LINE
4-TO-1
BALUN
50Ω

http://www.qsl.net/va3iul

---

260 - Carolina Windom Antenna for 80 to 10m

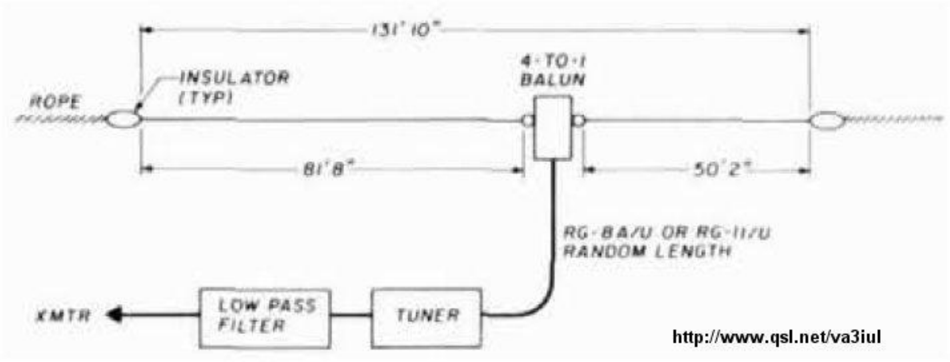
131' 10"
INSULATOR
(TYP)
4-TO-1
BALUN
ROPE
81' 8"
50' 2"
RG-8A/U OR RG-11/U
RANDOM LENGTH
XMTR
LOW PASS
FILTER
TUNER

http://www.qsl.net/va3iul

---

261 - Shortened Loaded Dipole for 40m

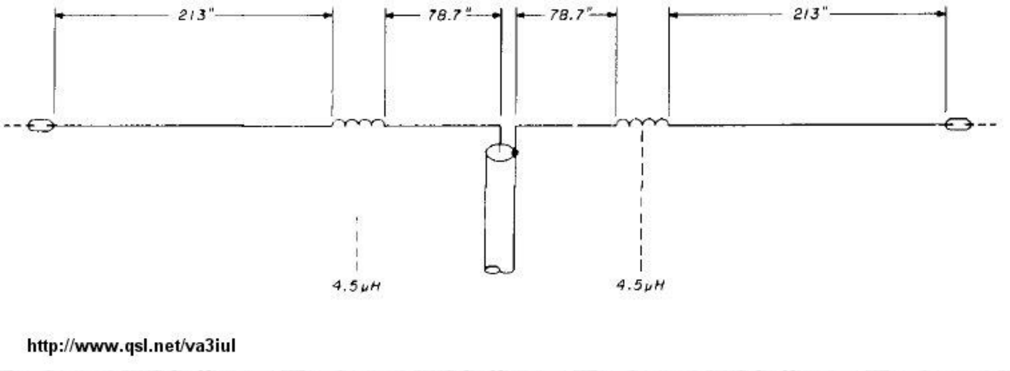
213"
78.7"
78.7"
213"
4.5μH
4.5μH

http://www.qsl.net/va3iul

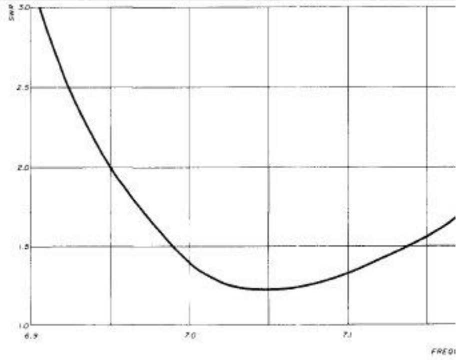

---

262 - Shortened Loaded Dipole for 80m

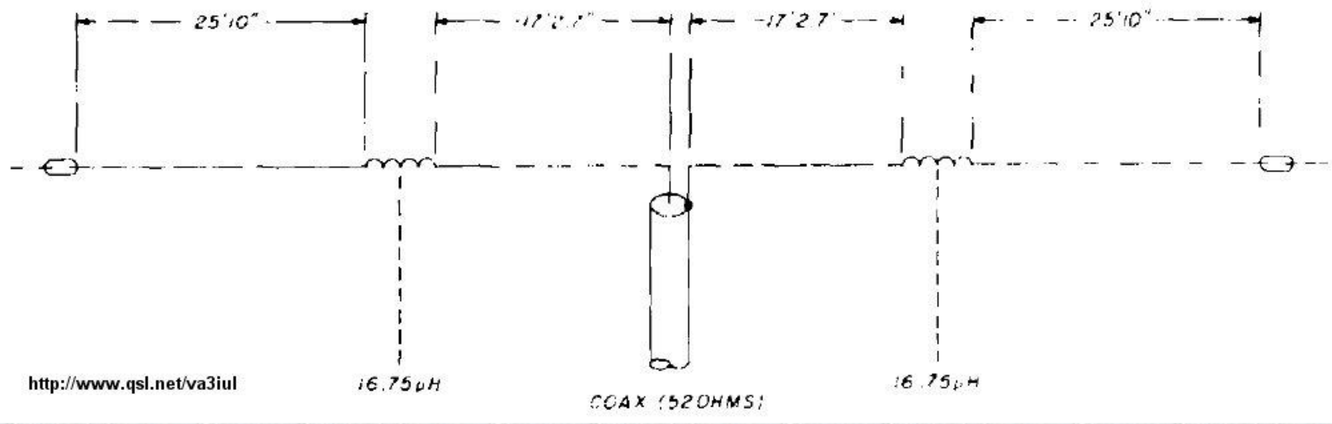
25'10"
17'2.7"
17'2.7"
25'10"
http://www.qsl.net/va3iul
16.75µH
COAX (52OHMS)
16.75µH

263 - Low Radiation Angle Inverted Delta Loop for 80m

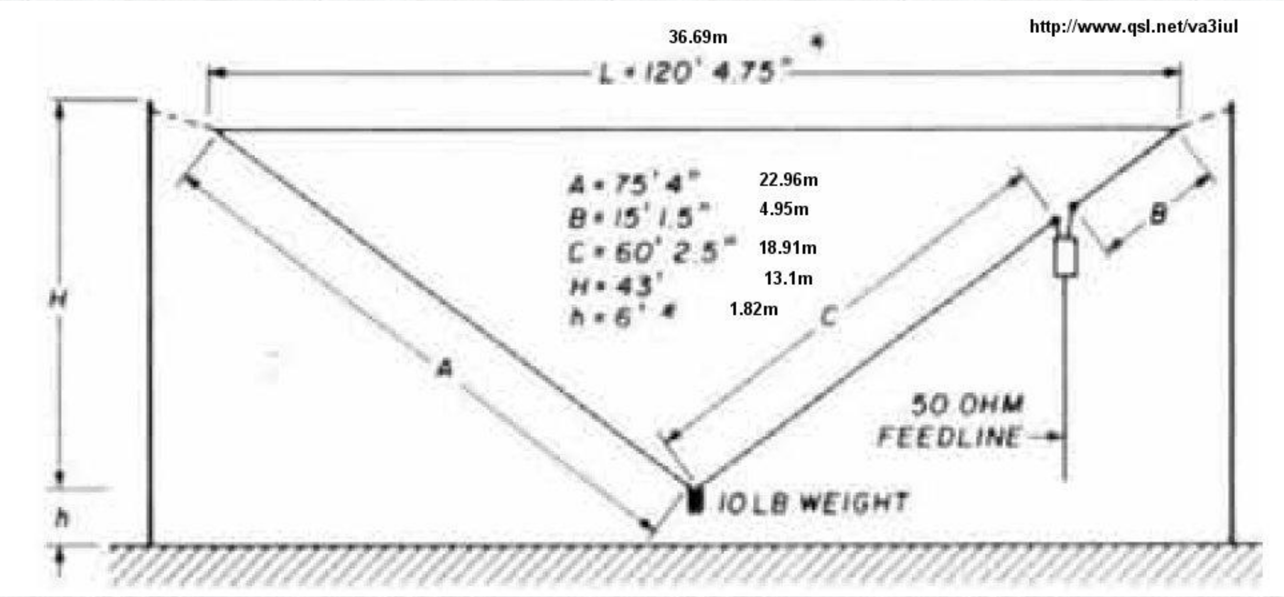
36.69m
L = 120' 4.75"
http://www.qsl.net/va3iul
A = 75' 4" 22.96m
B = 15' 1.5" 4.95m
C = 60' 2.5" 18.91m
H = 43' 13.1m
h = 6' * 1.82m
H
h
A
B
C
50 OHM FEEDLINE
10LB WEIGHT

264 - Compact Dipole Antenna for 40m, 15m

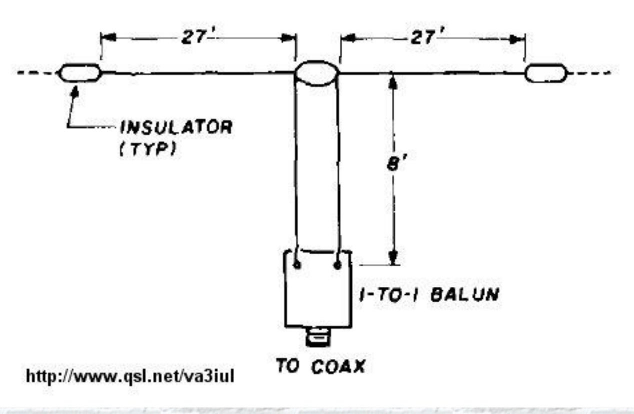
27'
27'
INSULATOR (TYP)
8'
1-TO-1 BALUN
TO COAX
http://www.qsl.net/va3iul

265 - Vertical Delta Loop for 40m to 10m

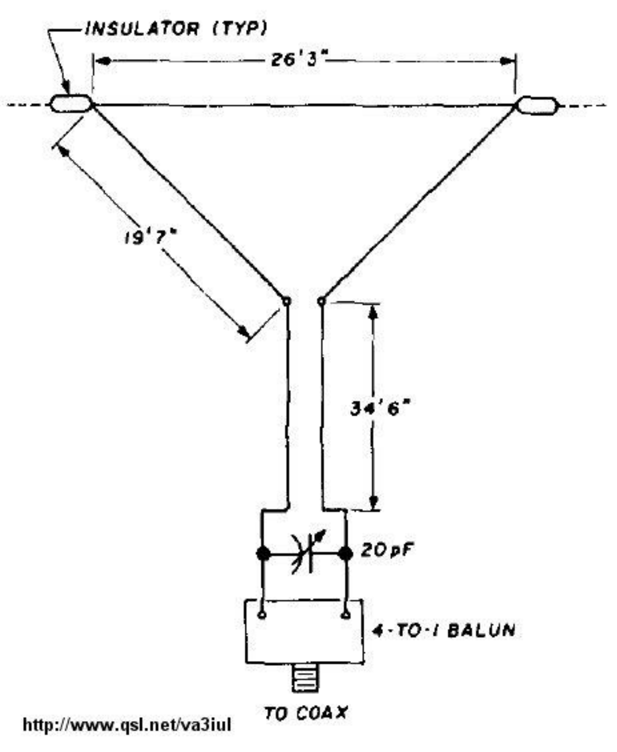
INSULATOR (TYP)
26'3"
19'7"
34'6"
20 pF
4-TO-1 BALUN
TO COAX
http://www.qsl.net/va3iul

---

266 - Delta Loop Antenna for 17m

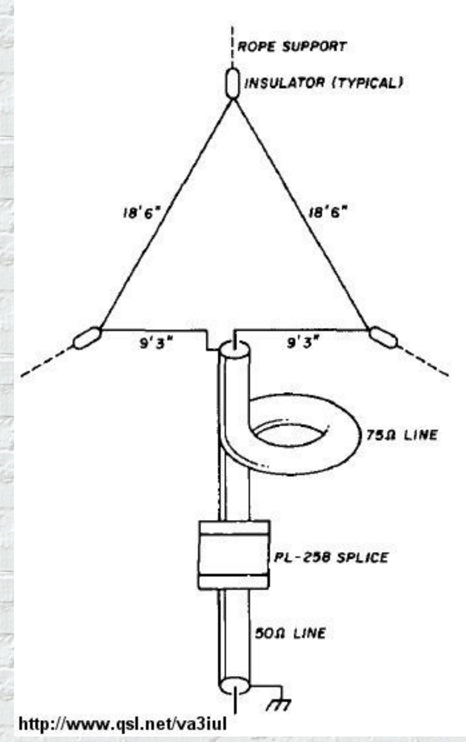
ROPE SUPPORT
INSULATOR (TYPICAL)
18'6"
18'6"
9'3"
9'3"
75Ω LINE
PL-258 SPLICE
50Ω LINE
http://www.qsl.net/va3iul

---

267 - Bi-Square Beam Gain=3dBd for 17m

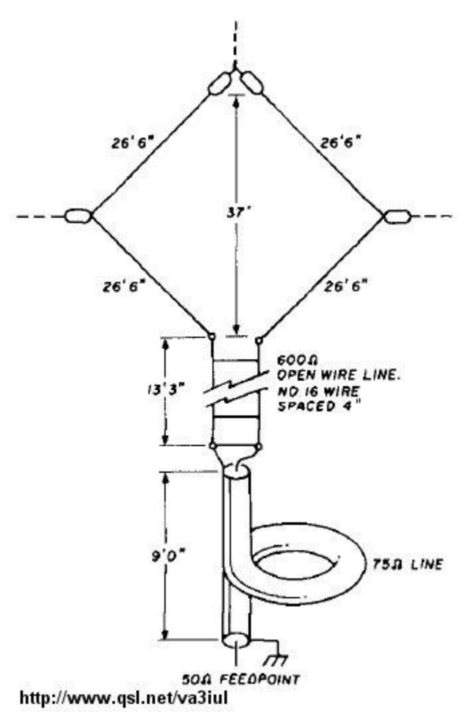
26'6"
26'6"
37'
26'6"
26'6"
13'3"
600Ω
OPEN WIRE LINE.
NO 16 WIRE
SPACED 4"
9'0"
75Ω LINE
50Ω FEEDPOINT

http://www.qsl.net/va3iul

268 -Trap Dipole Antenna for 17m,13m

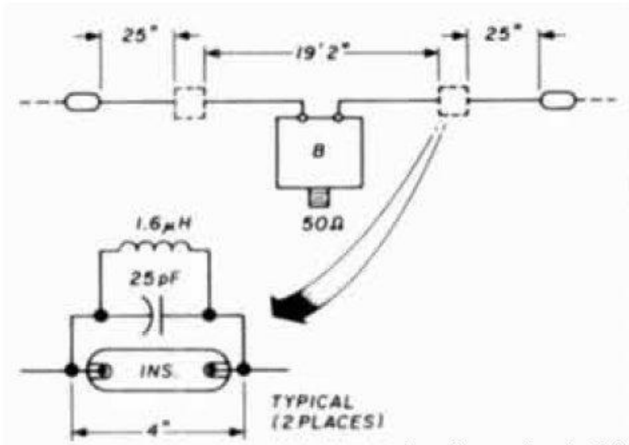
25"
19'2"
25"
B
50Ω
1.6μH
25pF
INS.
4"
TYPICAL
(2 PLACES)

http://www.qsl.net/va3iul

269 - Hanging Unipole Antenna for 160m, 80m

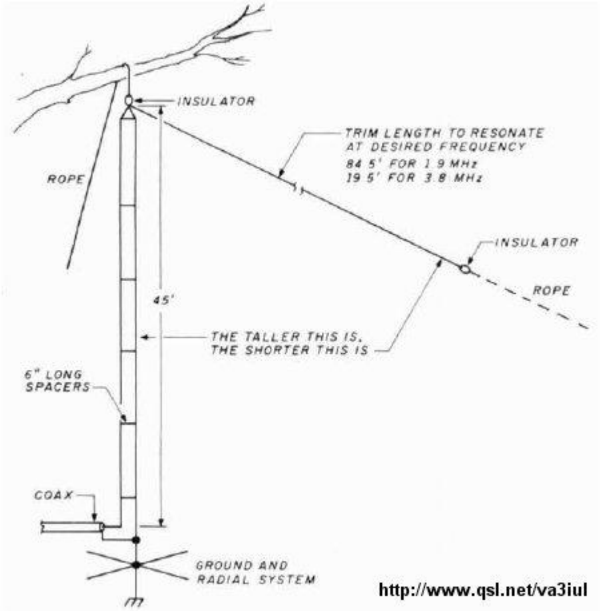
INSULATOR
TRIM LENGTH TO RESONATE
AT DESIRED FREQUENCY
84.5' FOR 1.9 MHz
19.5' FOR 3.8 MHz
ROPE
INSULATOR
ROPE
45'
THE TALLER THIS IS,
THE SHORTER THIS IS
6" LONG
SPACERS
COAX
GROUND AND
RADIAL SYSTEM
http://www.qsl.net/va3iul

---

270 - Wideband Sloping Vee Antenna - 30m to 6m

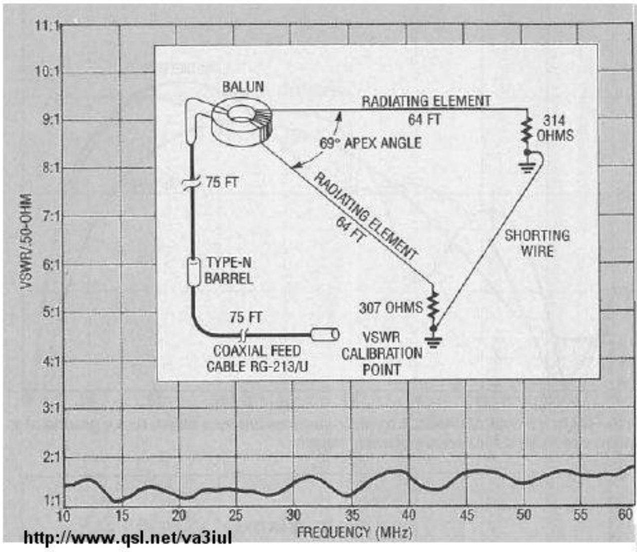

SOURCES OF MATERIALS

The following companies are sources for materials and computer software for this project:

- Toroidal ferrite cores (Part No. FT240-43)—Radio Kit, Inc., P.O. Box 973, Pelham, NH 03076, (603) 635-2235
- Film power resistors—Power Film Systems, Inc., Yellville, AR 72687, (501) 449-4091
- Antenna design software—Phadean Engineering Co., Inc., P.O. Box 611, Shrewsbury, MA 01545, (508) 869-6077
- Phosphor-bronze wire—Astro Industries, Inc., Dayton, OH 43432, (800) 543-5810
- Fiberglass tubing—J. T. Ryerson Co., P. O. Box 1111, Boston, MA 02103. (617) 782-6900

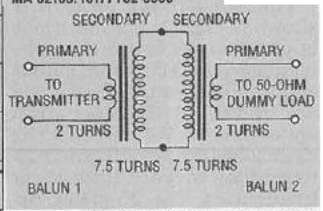
SECONDARY
SECONDARY
PRIMARY
PRIMARY
TO
TRANSMITTER
TO 50-OHM
DUMMY LOAD
2 TURNS
2 TURNS
7.5 TURNS
7.5 TURNS
BALUN 1
BALUN 2

BACK-TO-BACK INSERTION LOSS TEST of baluns for sloping-vee antenna.

---

271 - 5/8 Wavelength Antenna

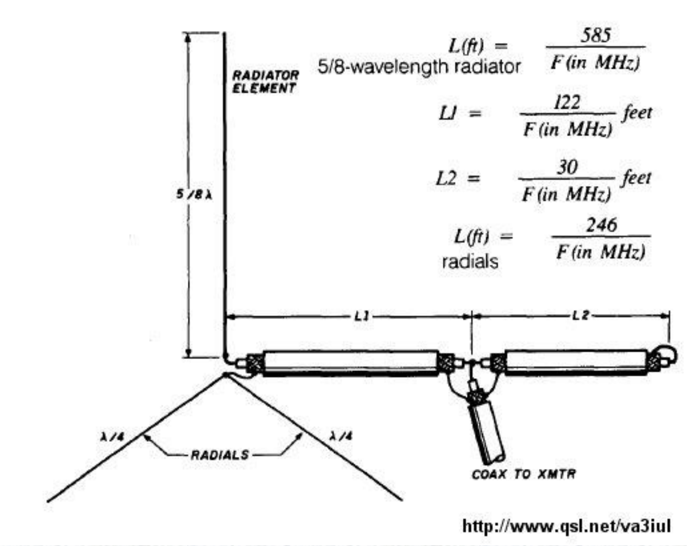
L(ft) = 585 / F(in MHz)
5/8-wavelength radiator
RADIATOR ELEMENT
L1 = 122 / F(in MHz) feet
L2 = 30 / F(in MHz) feet
L(ft) = 246 / F(in MHz)
radials
5/8 λ
L1
L2
λ/4
RADIALS
λ/4
COAX TO XMTR
http://www.qsl.net/va3iul

272 - C-Pole Vertical Antenna

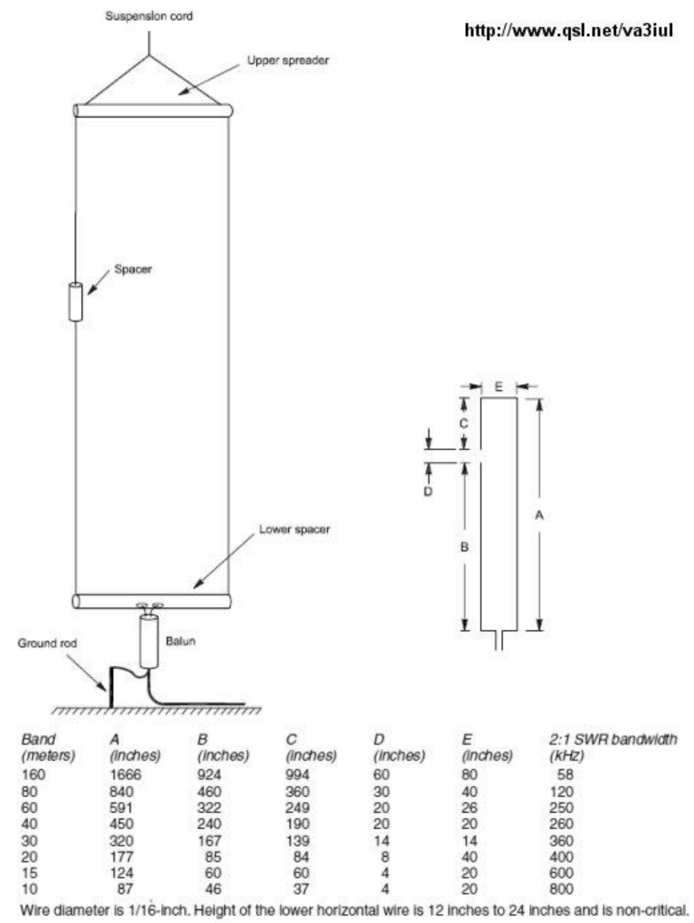
Suspension cord
http://www.qsl.net/va3iul
Upper spreader
Spacer
Lower spacer
Ground rod
Balun
E
C
D
A
B

Band (meters)	A (inches)	B (inches)	C (inches)	D (inches)	E (inches)	2:1 SWR bandwidth (kHz)
160	1666	924	994	60	80	58
80	840	460	360	30	40	120
60	591	322	249	20	26	250
40	450	240	190	20	20	260
30	320	167	139	14	14	360
20	177	85	84	8	40	400
15	124	60	60	4	20	600
10	87	46	37	4	20	800

Wire diameter is 1/16-inch. Height of the lower horizontal wire is 12 inches to 24 inches and is non-critical.

273 - Bobtail Antenna for 30m

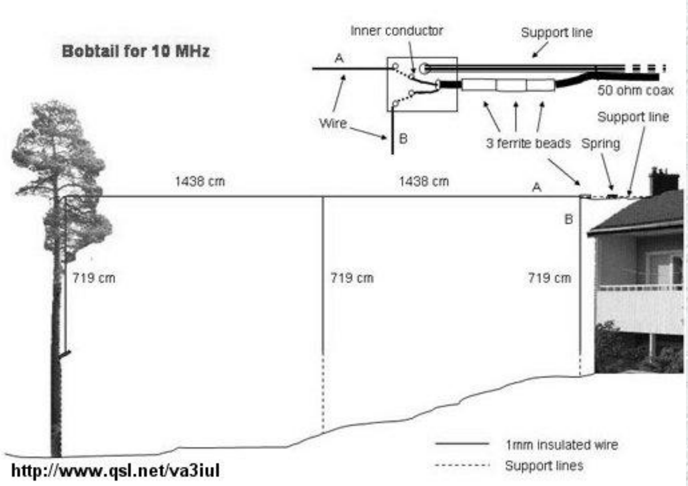
Bobtail for 10 MHz
Inner conductor
Support line
A
50 ohm coax
Wire
B
Support line
3 ferrite beads
Spring
1438 cm
1438 cm
A
B
719 cm
719 cm
719 cm
1mm insulated wire
Support lines
http://www.qsl.net/va3iul

---

274 - Bobtail Antenna for 40m

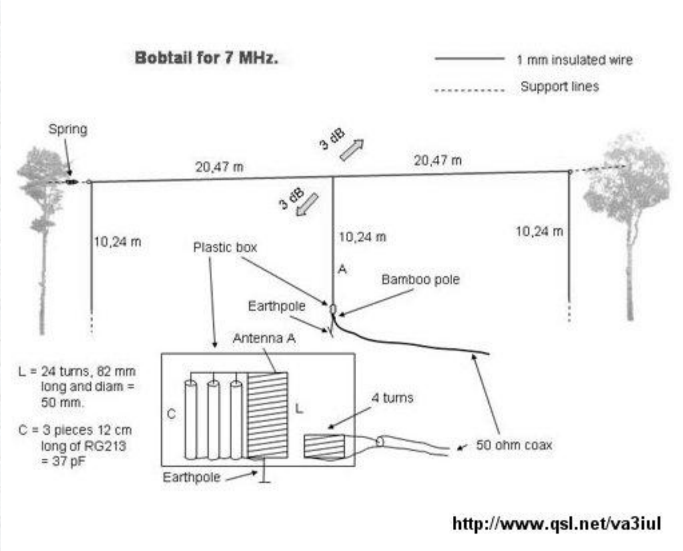
Bobtail for 7 MHz.
1 mm insulated wire
Support lines
Spring
3 dB
20,47 m
20,47 m
3 dB
10,24 m
10,24 m
10,24 m
Plastic box
A
Bamboo pole
Earthpole
Antenna A
L = 24 turns, 82 mm long and diam = 50 mm.
C = 3 pieces 12 cm long of RG213 = 37 pF
C
L
4 turns
50 ohm coax
Earthpole
http://www.qsl.net/va3iul

---

275 - Two Elements Vertical Loop Antenna for 40m

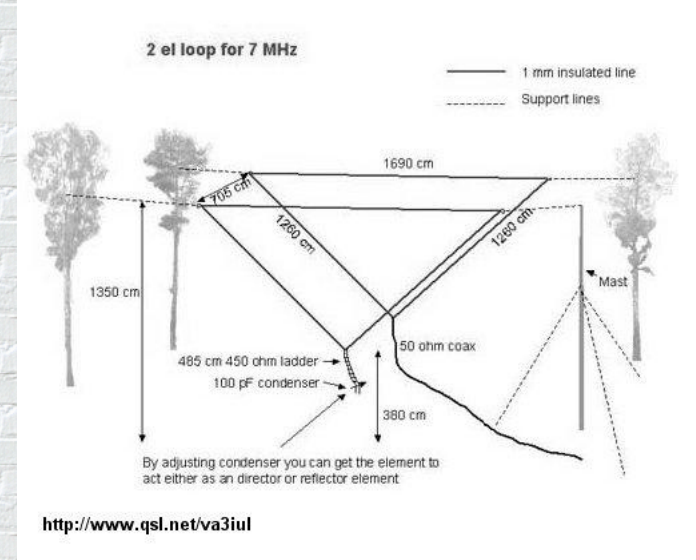
2 el loop for 7 MHz
1 mm insulated line
Support lines
1690 cm
705 cm
1260 cm
1260 cm
1350 cm
Mast
50 ohm coax
485 cm 450 ohm ladder
100 pF condenser
380 cm
By adjusting condenser you can get the element to act either as an director or reflector element
http://www.qsl.net/va3iul

---

276 - Double Extended Zepp Antenna for 40m

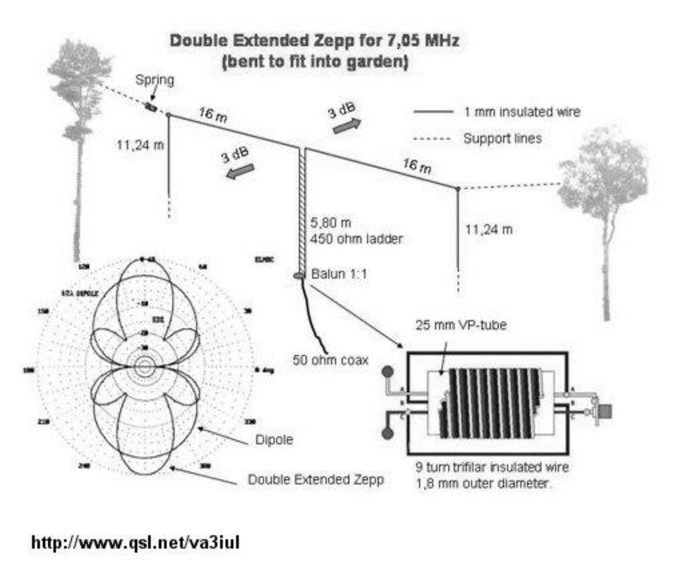
Double Extended Zepp for 7,05 MHz
(bent to fit into garden)
Spring
16 m
3 dB
11,24 m
3 dB
16 m
1 mm insulated wire
Support lines
5,80 m
450 ohm ladder
11,24 m
Balun 1:1
25 mm VP-tube
50 ohm coax
Dipole
Double Extended Zepp
9 turn trifilar insulated wire
1,8 mm outer diameter

http://www.qsl.net/va3iul

277 - Reversible Moxon Flip Beam for 40m

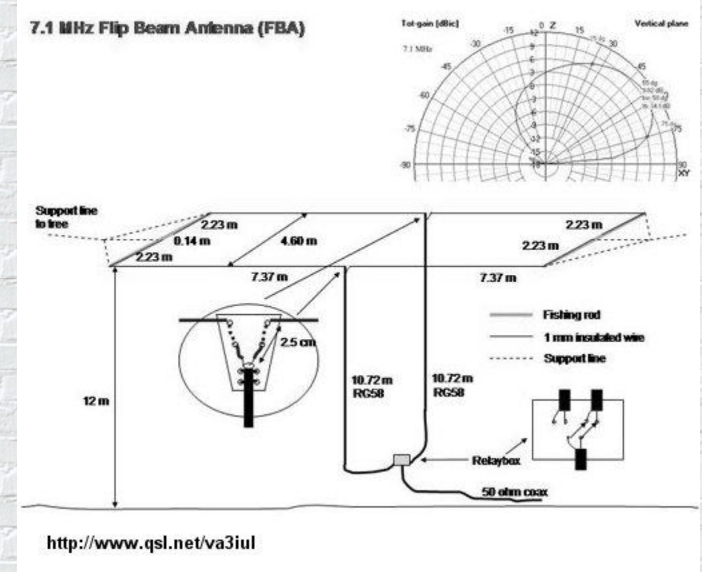
7.1 MHz Flip Beam Antenna (FBA)
Tot-gain [dBic]
7.1 MHz
Vertical plane
Support line to tree
2.23 m
0.14 m
2.23 m
4.60 m
2.23 m
2.23 m
7.37 m
7.37 m
2.5 cm
12 m
10.72 m RG58
10.72 m RG58
Fishing rod
1 mm insulated wire
Support line
Relaybox
50 ohm coax

http://www.qsl.net/va3iul

278 - Spitwire Antenna for 40m

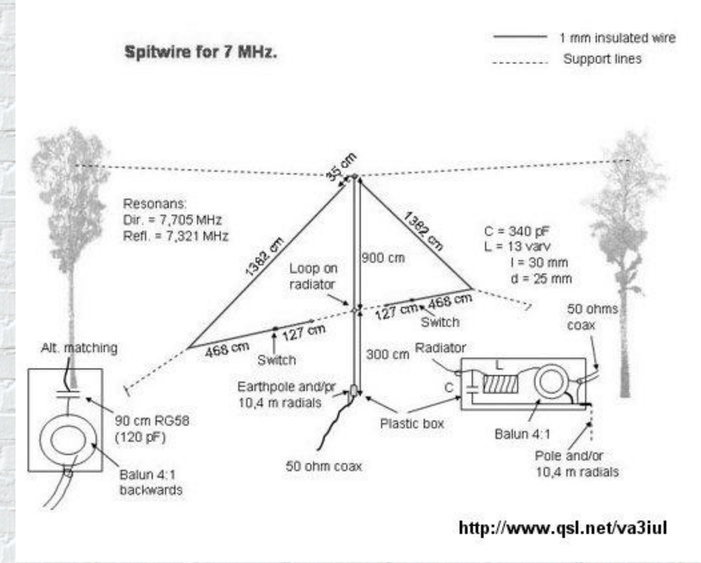
Spitwire for 7 MHz.
1 mm insulated wire
Support lines
35 cm
Resonans:
Dir. = 7,705 MHz
Refl. = 7,321 MHz
1382 cm
1382 cm
900 cm
C = 340 pF
L = 13 varv
l = 30 mm
d = 25 mm
Loop on radiator
127 cm
468 cm
127 cm
468 cm
Switch
Switch
50 ohms coax
Alt. matching
300 cm
Radiator
Earthpole and/pr 10,4 m radials
90 cm RG58 (120 pF)
Plastic box
C
L
Balun 4:1
Balun 4:1 backwards
Pole and/or 10,4 m radials
50 ohm coax
http://www.qsl.net/va3iul

279 - Fork Antenna for 40m

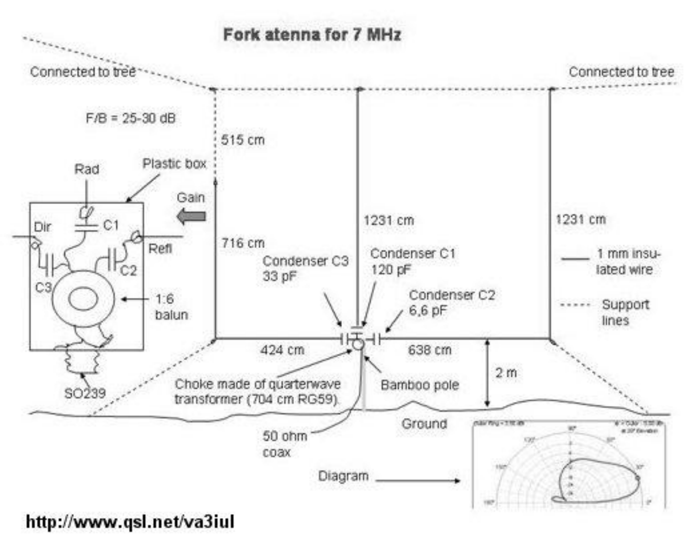
Fork atenna for 7 MHz
Connected to tree
Connected to tree
F/B = 25-30 dB
515 cm
Rad
Plastic box
Gain
Dir
C1
Refl
C2
C3
1:6 balun
716 cm
1231 cm
1231 cm
Condenser C3 33 pF
Condenser C1 120 pF
1 mm insulated wire
Condenser C2 6,6 pF
Support lines
424 cm
638 cm
2 m
SO239
Choke made of quarterwave transformer (704 cm RG59)
Bamboo pole
Ground
50 ohm coax
Diagram
http://www.qsl.net/va3iul

280 - Veebeam Antenna 3dB-Gain for All Bands

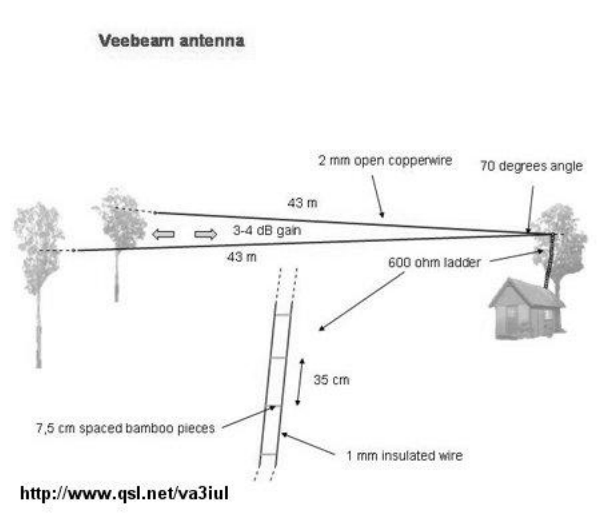
Veebeam antenna
2 mm open copperwire
70 degrees angle
43 m
3-4 dB gain
43 m
600 ohm ladder
35 cm
7,5 cm spaced bamboo pieces
1 mm insulated wire
http://www.qsl.net/va3iul

281 - Vertical Moxon Antenna for 10m

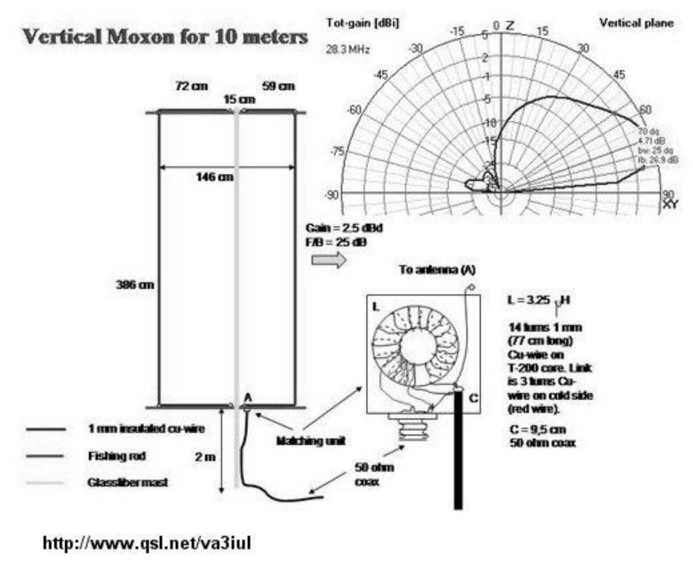
Vertical Moxon for 10 meters
Tot-gain [dBi]
28.3 MHz
Vertical plane
0 Z
-15
15
-30
30
-45
45
-60
60
-75
-90
90
XY
70 dg
4.71 dB
bw: 25 dg
fb: 26.9 dB
72 cm
59 cm
15 cm
146 cm
386 cm
A
Gain = 2.5 dBd
F/B = 25 dB
To antenna (A)
L
C
L = 3.25 μH
14 turns 1 mm
(77 cm long)
Cu-wire on
T-200 core. Link
is 3 turns Cu-
wire on cold side
(red wire).
C = 9,5 cm
50 ohm coax
1 mm insulated cu-wire
Fishing rod
Glassfibermast
2 m
Matching unit
50 ohm
coax
http://www.qsl.net/va3iul

---

282 - Half-Square Antenna for 30m

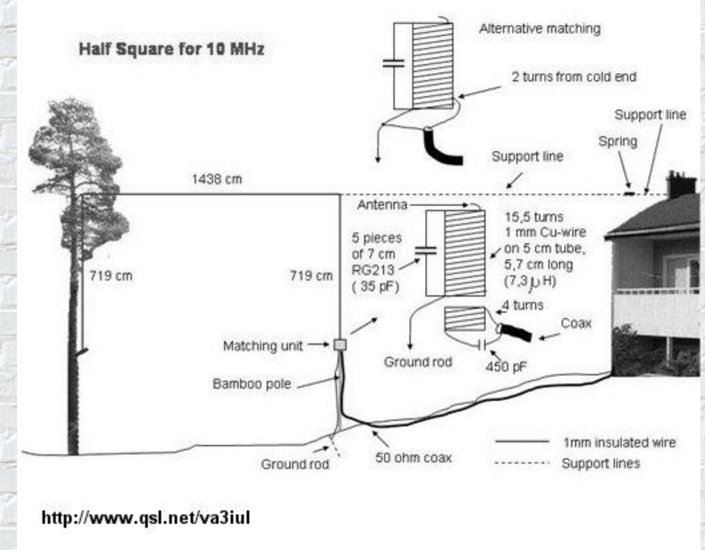
Half Square for 10 MHz
Alternative matching
2 turns from cold end
Support line
Spring
Support line
1438 cm
Antenna
5 pieces
of 7 cm
RG213
(35 pF)
15,5 turns
1 mm Cu-wire
on 5 cm tube,
5,7 cm long
(7,3 μH)
4 turns
Coax
719 cm
719 cm
Matching unit
Ground rod
450 pF
Bamboo pole
Ground rod
50 ohm coax
1mm insulated wire
Support lines
http://www.qsl.net/va3iul

---

283 - Rotatable Vertical Delta Loop for 20m

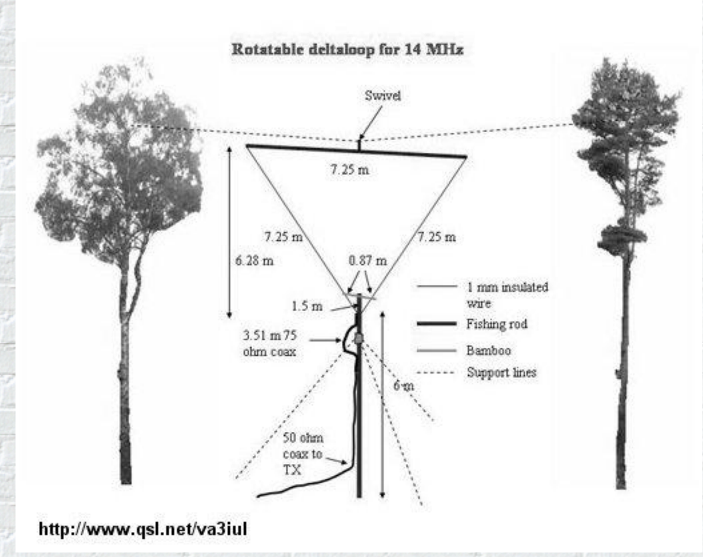
Rotatable deltaloop for 14 MHz
Swivel
7.25 m
7.25 m
7.25 m
6.28 m
0.87 m
1.5 m
3.51 m 75 ohm coax
6 m
50 ohm coax to TX
1 mm insulated wire
Fishing rod
Bamboo
Support lines
http://www.qsl.net/va3iul

284 - Mini-Horse Antenna 10dB-Gain for 40m-to-2m

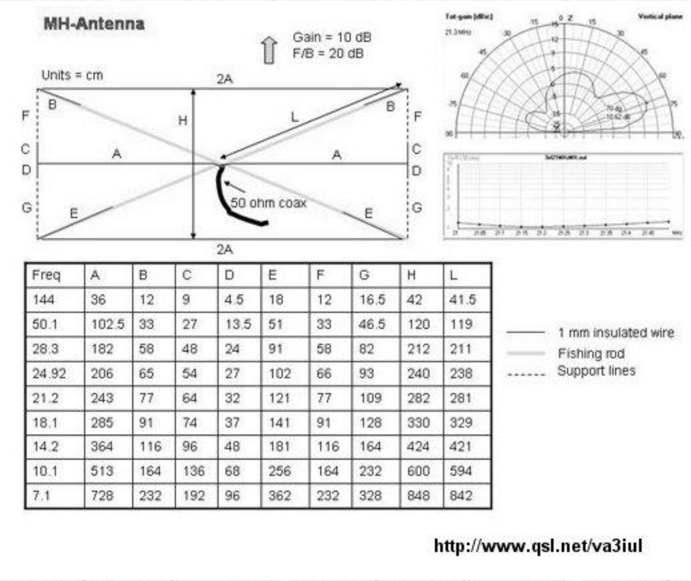
MH-Antenna
Gain = 10 dB
F/B = 20 dB
Units = cm
2A
B
B
F
F
H
L
C
C
D
D
A
A
G
G
E
E
50 ohm coax
2A
1 mm insulated wire
Fishing rod
Support lines
http://www.qsl.net/va3iul

Freq	A	B	C	D	E	F	G	H	L
144	36	12	9	4.5	18	12	16.5	42	41.5
50.1	102.5	33	27	13.5	51	33	46.5	120	119
28.3	182	58	48	24	91	58	82	212	211
24.92	206	65	54	27	102	66	93	240	238
21.2	243	77	64	32	121	77	109	282	281
18.1	285	91	74	37	141	91	128	330	329
14.2	364	116	96	48	181	116	164	424	421
10.1	513	164	136	68	256	164	232	600	594
7.1	728	232	192	96	362	232	328	848	842

285 - Compact Moxon Antenna 5dB-Gain for 40m

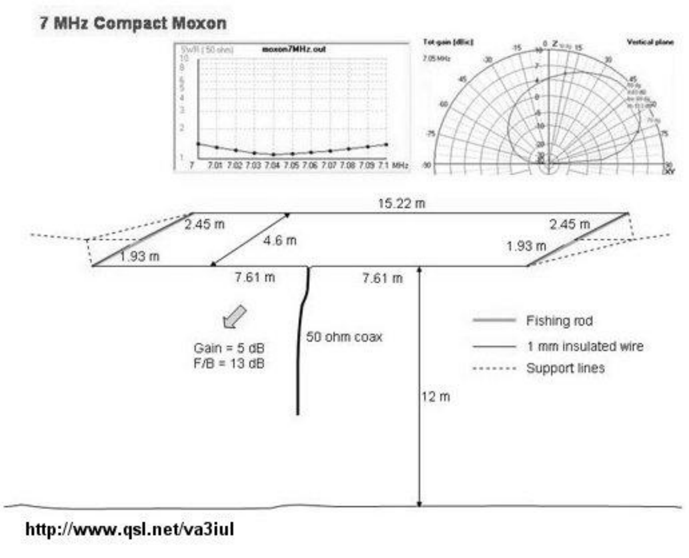
7 MHz Compact Moxon
moxon7MHz.out
Vertical plane
15.22 m
2.45 m
2.45 m
4.6 m
1.93 m
1.93 m
7.61 m
7.61 m
Gain = 5 dB
F/B = 13 dB
50 ohm coax
Fishing rod
1 mm insulated wire
Support lines
12 m
http://www.qsl.net/va3iul

286 - Two Elements Sloping Delta Loop 8dB-Gain for 10m

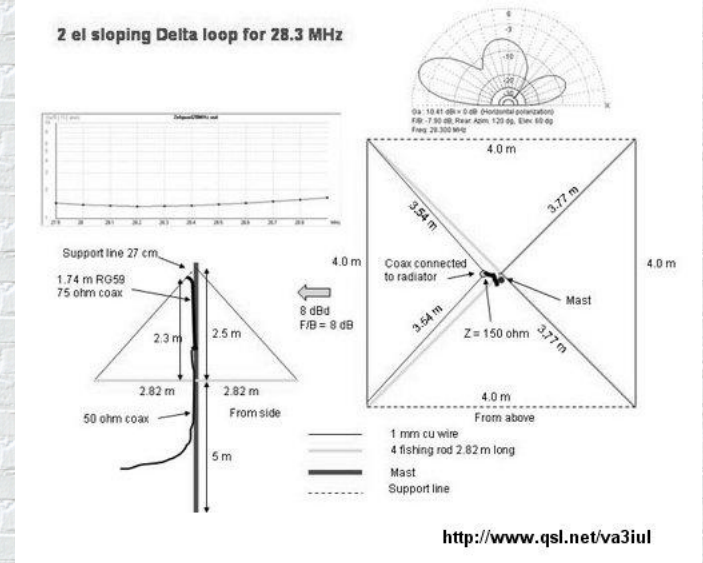
2 el sloping Delta loop for 28.3 MHz
Support line 27 cm
1.74 m RG59
75 ohm coax
2.3 m
2.5 m
2.82 m
2.82 m
50 ohm coax
From side
5 m
8 dBd
F/B = 8 dB
4.0 m
4.0 m
4.0 m
4.0 m
3.54 m
3.77 m
3.54 m
3.77 m
Coax connected
to radiator
Mast
Z = 150 ohm
From above
1 mm cu wire
4 fishing rod 2.82 m long
Mast
Support line
http://www.qsl.net/va3iul

287 - Two Elements Phased Vertical Antenna 3dB-Gain for 40m

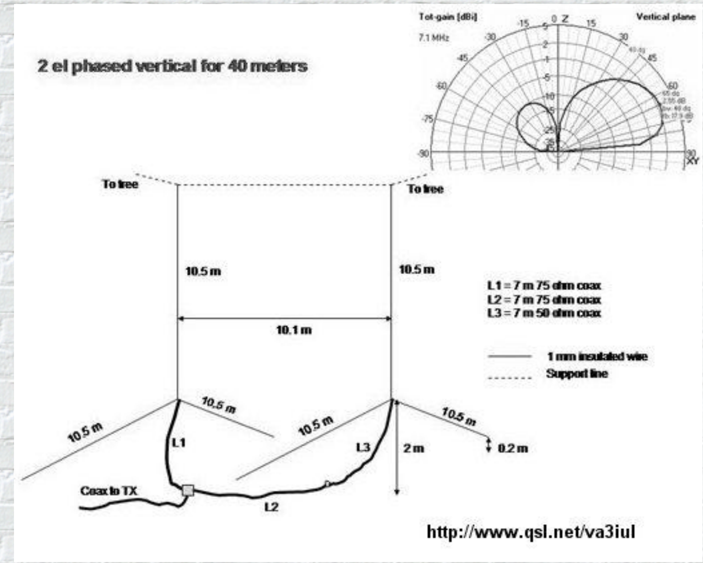
2 el phased vertical for 40 meters
Tot-gain [dBi]
7.1 MHz
Vertical plane
To tree
To tree
10.5 m
10.5 m
10.1 m
L1 = 7 m 75 ohm coax
L2 = 7 m 75 ohm coax
L3 = 7 m 50 ohm coax
1 mm insulated wire
Support line
10.5 m
10.5 m
10.5 m
10.5 m
L1
L3
2 m
0.2 m
Coax to TX
L2
http://www.qsl.net/va3iul

References:

Antenna Toolkit - J. Carr
Practical Antenna Handbook - J. Carr
More Antenna Classics - C. Hutchinson
Amateur Radio Techniques - P. Hawker
The Radio Handbook - W. Orr (editor), 15th edition
The Radio Antenna Handbook - Engineering Staff of "Radio"
Jones Antenna Handbook - Frank C. Jones
SM0DTK - http://www.sm0dtk.se/antennas.htm
ARRL Antenna Handbook 1990-2007
ARRL Handbook 1937-2009
ARRL Antenna Compendium
73 Magazine 1970-2002
Funkamateur Magazine 1990-2011
RSGB Handbook - 2007 - 2010
QST Magazine - 1980 - 2010
Ham Radio Magazine - 1969 - 1990
Radio-Electronics Magazine - 1986 - 1990

home: ***http://www.qsl.net/va3iul***

Live Traffic Feed

Watch Live | Get FEEDJIT

Richmond, California, United States arrived 0 secs ago.

United Kingdom arrived 2 mins ago.

Kundan, Kashmir, India arrived 3 mins ago.

San Francisco, California, United States arrived 9 mins ago.

Rostock, Mecklenburg-Vorpommern, Germany arrived 13 mins ago.

Russia arrived 16 mins ago.

Abu Dhabi, United Arab Emirates arrived 32 mins ago.

United States arrived 35 mins ago.

Menlo Park, California, United States arrived 58 mins ago.

Québec, Canada arrived 59 mins ago.

$44.95

The ARRL

ANTENNA BOOK

21st Edition

The *ultimate* reference
for Amateur Radio antennas,
transmission lines and
propagation

INSIDE: Includes the fully
searchable book on CD-ROM

The ARRL Antenna Book

All the information you need to design your own complete antenna system.

Since the first edition in September 1939, radio amateurs and professional engineers have turned to **The ARRL Antenna Book** as THE source of current antenna theory and a wealth of practical how-to construction projects. Use this book to discover even the most basic antenna designs— wire and loop antennas, verticals, and Yagis—and for advanced antenna theory and applications. Many of the antennas in this edition benefit directly from advances in sophisticated computer modeling.

This 21st edition has been extensively revised to include information you can use to build highly optimized or specialized antennas. The book includes new content on Near Vertical Incidence Skywave (NVIS) techniques, phased arrays, S-parameters as used in modern vector network analyzers (VNA), Beverage receiving antennas, mobile "screwdriver" antennas, ionospheric area-coverage maps, *and much…much more*.

Fully searchable CD-ROM included!
Bundled with this book is a CD-ROM containing **The ARRL Antenna Book** in its entirety, using the popular Adobe® Reader® software for Microsoft® Windows® and Macintosh® systems. View, search and print from the entire text, including images, photographs, drawings…**everything!**

The CD-ROM contains additional utility programs, including:

- **YW**—Yagi for Windows
- **TLW**—Transmission Line for Windows
- **HFTA**—HF Terrain Assessment for Windows
- **Range-Bearing**—compute range/bearing or latitude/longitude
- **Arrayfeed1**—designing phased-array feed systems
- **EZNEC-ARRL**—antenna modeling for Windows

PLUS propagation forecast tables for more than 170 locations around the world for all portions of the solar cycle!

CONTENTS

Safety First
Antenna Fundamentals
The Effects of Ground
Antenna Modeling and System Planning
Loop Antennas
Low-Frequency Antennas
Multiband Antennas
Multielement Arrays
Broadband Antenna Matching
Log Periodic Arrays
HF Yagi Arrays
Quad Arrays
Long-Wire and Traveling-Wave Antennas
Direction Finding Antennas
Portable Antennas
Mobile and Maritime Antennas
Repeater Antenna Systems
VHF and UHF Antenna Systems
Antenna Systems for Space Communications
Antenna Materials and Accessories
Antenna Products Suppliers
Antenna Supports
Radio Wave Propagation
Transmission Lines
Coupling the Transmitter to the Line
Coupling the Line to the Antenna
Antenna and Transmission-Line Measurements
Smith Chart Calculations

Includes a comprehensive glossary and index

Published by:
ARRL The national association for AMATEUR RADIO
225 Main Street, Newington, CT 06111-1494 USA
www.arrl.org

ISBN-13: 978-0-87259-987-1
ISBN-10: 0-87259-987-6
54495
EAN
9 780872 599871

ISBN 0-87259-987-6 Order No. 9876

The ARRL

Antenna Book

21st Edition

Editor

R. Dean Straw, N6BV

Contributing Editors

L. B. Cebik, W4RNL
Dave Hallidy, K2DH
Dick Jansson, WD4FAB
Roy Lewallen, W7EL
Rudy Severns, N6LF
Frank Witt, AI1H

Production

Michelle Bloom, WB1ENT
Sue Fagan, KB1OKW—Cover Art
Jodi Morin, KA1JPA
David Pingree, N1NAS
Kathy Ford

Software Beta Testers

Ahti Aintila, OH2RZ
Alan Bloom, N1AL
L. B. Cebik, W4RNL
Loney Duncan, W0GZV
Chuck Hutchinson, K8CH
Steve London, N2IC
Dan Maguire, AC6LA
Steve Miller, N8SM (SK)
Bob Neece, K0KR,
Danny Richardson, K6MHE
Rudy Severns, N6LF
Jim Tabor, KU5S
Bill Tippett, W4ZV
Don Toman, K2KQ
Pete Smith, N4ZR
Scott Tuthill, K7ZO
Frank Witt, AI1H.

The Covers:

Tower photo at sunset (front and back): Tower and beam at the station of Francisco R.F. Aragao, PT2TD, in Brasilia, Brazil.

Center image: Kurt Andress, K7NV, working on 40-meter Yagi at N6RO in Oakley, CA. Photo by Dean Straw, N6BV.

Upper right image: UHF antenna. Photo credit: ShutterStock.com.

Lower right image: Some of the members of the C5Z contest DXpedition near Banjul, The Gambia. Photo by Henryk Kotowski, SM0JHF.

ARRL The national association for AMATEUR RADIO
225 Main Street, Newington, CT 06111-1494 USA
www.arrl.org/

Contents

1 Safety First
2 Antenna Fundamentals
3 The Effects of Ground
4 Antenna Modeling & System Planning
5 Loop Antennas
6 Low-Frequency Antennas
7 Multiband Antennas
8 Multielement Arrays
9 Broadband Antenna Matching
10 Log Periodic Arrays
11 HF Yagi Arrays
12 Quad Arrays
13 Long-Wire and Traveling-Wave Antennas
14 Direction Finding Antennas
15 Portable Antennas
16 Mobile and Maritime Antennas

17 Repeater Antenna Systems
18 VHF and UHF Antenna Systems
19 Antenna Systems for Space Communications
20 Antenna Materials and Accessories
21 Antenna Products Suppliers
22 Antenna Supports
23 Radio Wave Propagation
24 Transmission Lines
25 Coupling the Transmitter to the Line
26 Coupling the Line to the Antenna
27 Antenna and Transmission-Line Measurements
28 Smith Chart Calculations
A-1 Appendix
AB-1 Index of Advertisers
931 Index

Schematic Symbols Used in Circuit Diagrams

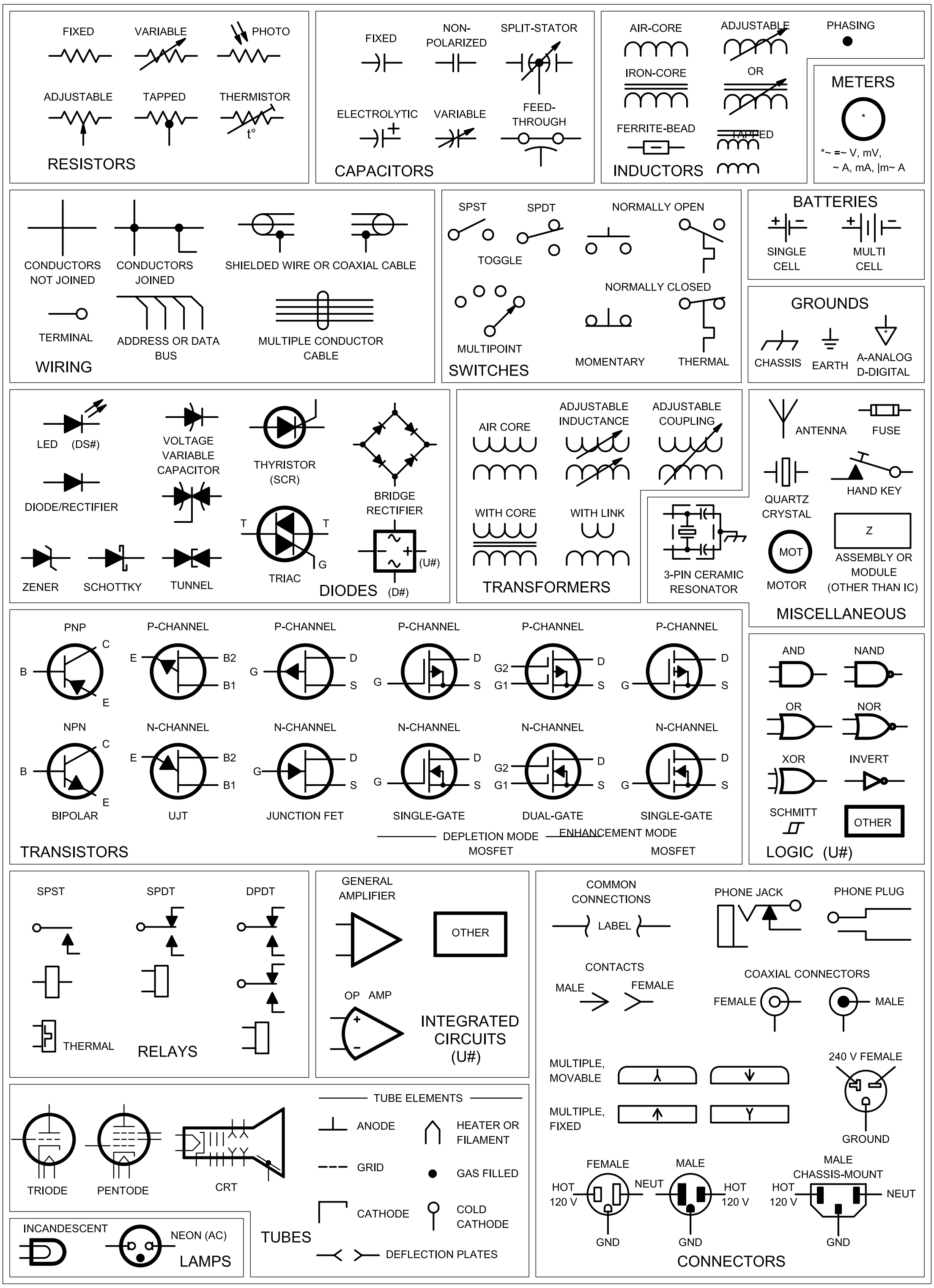
FIXED
VARIABLE
PHOTO
ADJUSTABLE
TAPPED
THERMISTOR
t°
RESISTORS
FIXED
NON-POLARIZED
SPLIT-STATOR
ELECTROLYTIC
VARIABLE
FEED-THROUGH
CAPACITORS
AIR-CORE
ADJUSTABLE
PHASING
IRON-CORE
OR
FERRITE-BEAD
TAPPED
INDUCTORS
METERS
*
*~ =~ V, mV,
~ A, mA, |m~ A
CONDUCTORS NOT JOINED
CONDUCTORS JOINED
SHIELDED WIRE OR COAXIAL CABLE
TERMINAL
ADDRESS OR DATA BUS
MULTIPLE CONDUCTOR CABLE
WIRING
SPST
SPDT
NORMALLY OPEN
TOGGLE
NORMALLY CLOSED
MULTIPOINT
MOMENTARY
THERMAL
SWITCHES
BATTERIES
SINGLE CELL
MULTI CELL
GROUNDS
CHASSIS
EARTH
A-ANALOG
D-DIGITAL
LED (DS#)
VOLTAGE VARIABLE CAPACITOR
THYRISTOR (SCR)
BRIDGE RECTIFIER
DIODE/RECTIFIER
T
T
G
TRIAC
(U#)
ZENER
SCHOTTKY
TUNNEL
DIODES (D#)
AIR CORE
ADJUSTABLE INDUCTANCE
ADJUSTABLE COUPLING
WITH CORE
WITH LINK
TRANSFORMERS
ANTENNA
FUSE
QUARTZ CRYSTAL
HAND KEY
Z
3-PIN CERAMIC RESONATOR
MOT
MOTOR
ASSEMBLY OR MODULE (OTHER THAN IC)
MISCELLANEOUS
PNP
P-CHANNEL
P-CHANNEL
P-CHANNEL
P-CHANNEL
P-CHANNEL
B
C
E
E
B2
B1
G
D
S
G
D
S
G2
G1
D
S
G
D
S
NPN
N-CHANNEL
N-CHANNEL
N-CHANNEL
N-CHANNEL
N-CHANNEL
BIPOLAR
UJT
JUNCTION FET
SINGLE-GATE
DUAL-GATE
SINGLE-GATE
DEPLETION MODE MOSFET
ENHANCEMENT MODE MOSFET
TRANSISTORS
AND
NAND
OR
NOR
XOR
INVERT
SCHMITT
OTHER
LOGIC (U#)
SPST
SPDT
DPDT
THERMAL
RELAYS
GENERAL AMPLIFIER
OTHER
OP AMP
+
-
INTEGRATED CIRCUITS (U#)
COMMON CONNECTIONS
LABEL
PHONE JACK
PHONE PLUG
CONTACTS
MALE
FEMALE
COAXIAL CONNECTORS
FEMALE
MALE
MULTIPLE, MOVABLE
240 V FEMALE
MULTIPLE, FIXED
GROUND
TRIODE
PENTODE
CRT
TUBE ELEMENTS
ANODE
HEATER OR FILAMENT
GRID
GAS FILLED
CATHODE
COLD CATHODE
DEFLECTION PLATES
TUBES
INCANDESCENT
NEON (AC)
LAMPS
FEMALE
MALE
MALE CHASSIS-MOUNT
HOT 120 V
NEUT
HOT 120 V
HOT 120 V
NEUT
GND
GND
GND
CONNECTORS

About the ARRL

The seed for Amateur Radio was planted in the 1890s, when Guglielmo Marconi began his experiments in wireless telegraphy. Soon he was joined by dozens, then hundreds, of others who were enthusiastic about sending and receiving messages through the air—some with a commercial interest, but others solely out of a love for this new communications medium. The United States government began licensing Amateur Radio operators in 1912.

By 1914, there were thousands of Amateur Radio operators—hams—in the United States. Hiram Percy Maxim, a leading Hartford, Connecticut inventor and industrialist, saw the need for an organization to band together this fledgling group of radio experimenters. In May 1914 he founded the American Radio Relay League (ARRL) to meet that need.

Today ARRL, with approximately 150,000 members, is the largest organization of radio amateurs in the United States. The ARRL is a not-for-profit organization that:

- promotes interest in Amateur Radio communications and experimentation
- represents US radio amateurs in legislative matters, and
- maintains fraternalism and a high standard of conduct among Amateur Radio operators.

At ARRL headquarters in the Hartford suburb of Newington, the staff helps serve the needs of members. ARRL is also International Secretariat for the International Amateur Radio Union, which is made up of similar societies in 150 countries around the world.

ARRL publishes the monthly journal *QST*, as well as newsletters and many publications covering all aspects of Amateur Radio. Its headquarters station, W1AW, transmits bulletins of interest to radio amateurs and Morse code practice sessions. The ARRL also coordinates an extensive field organization, which includes volunteers who provide technical information and other support services for radio amateurs as well as communications for public-service activities. In addition, ARRL represents US amateurs with the Federal Communications Commission and other government agencies in the US and abroad.

Membership in ARRL means much more than receiving *QST* each month. In addition to the services already described, ARRL offers membership services on a personal level, such as the ARRL Volunteer Examiner Coordinator Program and a QSL bureau.

Full ARRL membership (available only to licensed radio amateurs) gives you a voice in how the affairs of the organization are governed. ARRL policy is set by a Board of Directors (one from each of 15 Divisions). Each year, one-third of the ARRL Board of Directors stands for election by the full members they represent. The day-to-day operation of ARRL HQ is managed by a Chief Executive Officer.

No matter what aspect of Amateur Radio attracts you, ARRL membership is relevant and important. There would be no Amateur Radio as we know it today were it not for the ARRL. We would be happy to welcome you as a member! (An Amateur Radio license is not required for Associate Membership.) For more information about ARRL and answers to any questions you may have about Amateur Radio, write or call:

ARRL—The national association for Amateur Radio
225 Main Street
Newington CT 06111-1494
Voice: 860-594-0200
Fax: 860-594-0259
E-mail: **hq@arrl.org**
Internet: **www.arrl.org/**

Prospective new amateurs call (toll-free):
800-32-NEW HAM (800-326-3942)
You can also contact us via e-mail at **newham@arrl.org**
or check out *ARRLWeb* at **http://www.arrl.org/**

Chapter 1

Safety First

Safety begins with your attitude. If you make it a habit to plan your work carefully and to consider the safety aspects of a project before you begin the work, you will be much safer than "Careless Carl," who just jumps in, proceeding in a haphazard manner. Learn to have a positive attitude about safety. Think about the dangers involved with a job before you begin the work. Don't be the one to say, "I didn't think it could happen to me."

Having a good attitude about safety isn't enough, however. You must be knowledgeable about common safety guidelines and follow them faithfully. Safety guidelines can't possibly cover all the situations you might face, but if you approach a task with a measure of "common sense," you should be able to work safely.

This chapter offers some safety guidelines and protective measures for you and your Amateur Radio station. You should not consider it to be an all-inclusive discussion of safety practices, though. Safety considerations will affect your choice of materials and assembly procedures when building an antenna. Other chapters of this book will offer further suggestions on safe construction practices. For example, Chapter 22 includes some very important advice on a tower installation.

PUTTING UP SIMPLE WIRE ANTENNAS

No matter what type of antenna you choose to erect, you should remember a few key points about safety. If you are using a slingshot or bow and arrow to get a line over a tree, make sure you keep everyone away from the "downrange" area. Hitting one of your helpers with a rock or fishing sinker is considered not nice, and could end up causing a serious injury.

Make sure the ends of the antenna are high enough to be out of reach of passers-by. Even when you are transmitting with low power there may be enough voltage at the ends of your antenna to give someone nasty "RF burns." If you have a vertical antenna with its base at ground level, build a wooden safety fence around it at least 4 feet away from it. Do not use metal fence, as this will interfere with the proper operation of the antenna. Be especially certain that your antenna is not close to any power wires. That is the only way you can be sure it won't come in contact with them!

Antenna work often requires that one person climb up on a tower, into a tree or onto the roof of a house. Never work alone! Work slowly, thinking out each move before you make it. The person on the ladder, tower, tree or rooftop should wear a safety belt, and keep it securely anchored. It is helpful (and safe!) to tie strings or lightweight ropes to all tools. If your tools are tied on, you'll save time getting them back if you drop them, and you'll greatly reduce the risk of injuring a helper on the ground. (There are more safety tips for climbing and working on towers later in this chapter. Those tips apply to any work that you must do above the ground to install even the simplest antenna.)

Tower Safety

Working on towers and antennas is dangerous, and possibly fatal, if you do not know what you are doing. Your tower and antenna can cause serious property damage and personal injury if any part of the installation should fail. Always use the highest quality materials in your system. Follow the manufacturer's specifications, paying close attention to base pier and guying details. Do not overload the tower. If you have any doubts about your ability to work on your tower and antennas safely, contact another amateur with experience in this area or seek professional assistance.

Chapter 22, Antenna Supports, provides more detailed guidelines for constructing a tower base and putting up a tower. It also explains how to properly attach guy wires and install guy anchors in the ground. These are extremely important parts of a tower installation, and you should not take shortcuts or use second-rate materials. Otherwise the strength and safety of your entire antenna system may be compromised.

Any mechanical job is easier if you have the right tools. Tower work is no exception. In addition to a good assortment of wrenches, screwdrivers and pliers, you will need some specialized tools to work safely and efficiently on a tower. You may already own some of these tools. Others may be purchased or borrowed. Don't start a job until you have assembled all of the necessary tools. Shortcuts or improvised tools can be fatal if you gamble and lose at 70 feet in the air. The following sections describe in detail the tools you will need to work safely on a tower.

CLOTHING

The clothing you wear when working on towers and antennas should be selected for maximum comfort and safety. Wear clothing that will keep you warm, yet allow complete freedom of movement. Long denim pants and a long-sleeve shirt will protect you from scrapes and cuts. (A pull-on shirt, like a sweat shirt with no openings or buttons to snag on tower parts, is best.) Wear work shoes with heavy soles, or better yet, with steel shanks (steel inserts in the soles), to give your feet the support they need to stand on a narrow tower rung.

Gloves are necessary for both the tower climber and all ground-crew members. Good quality leather gloves will protect hands from injury and keep them warm. They also offer protection and a better grip when you are handling rope. In cooler weather, a pair of gloves with light insulation will help keep your hands warm. The insulation should not be so bulky as to inhibit movement, however.

Ground-crew members should have hard hats for protection in case something falls from the tower. It is not uncommon for the tower climber to drop tools and hardware. A wrench dropped from 100 feet will bury itself several inches in soft ground; imagine what it might do to an unprotected skull.

SAFETY BELT AND CLIMBING ACCESSORIES

Any amateur with a tower must own a high-quality safety belt, such as the one shown in **Fig 1**. *Do not attempt* to climb a tower, even a short distance, without a belt. The climbing belt is more than just a safety device for the experienced climber. It is a tool to free up both hands for work. The belt allows the climber to lean back away from the tower to reach bolts or connections. It also provides a solid surface to lean against to exert greater force when hoisting antennas into place.

A climber must trust his life to his safety belt. For this reason, nothing less than a professional quality, commercially made, tested and approved safety belt is acceptable. Check the suppliers' list in Chapter 21, Antenna Product Suppliers, and ads in *QST* for suppliers of climbing belts and accessories. Examine your belt for defects before each use. If the belt or lanyard (tower strap) are cracked, frayed or worn in any way, destroy the damaged piece and replace it with a new one. You should never have to wonder if your belt will hold.

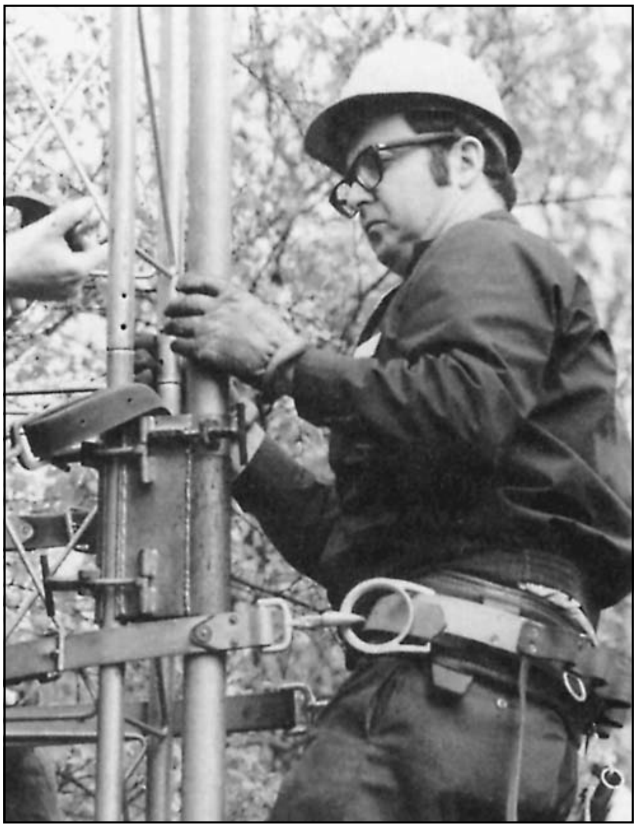

Fig 1—Bill Lowry, W1VV, uses a good quality safety belt, a requirement for working on a tower. The belt should contain large steel loops for the strap snaps. Leather loops at the rear of the belt are handy for holding tools. (*Photo by K1WA*)

Along with your climbing belt, you should seriously consider purchasing some climbing accessories. A canvas bucket is a great help for carrying tools and hardware up the tower. Two buckets, a large one for carrying tools and a smaller one for hardware, make it easier to find things when needed. A few extra snap hooks like those on the ends of your belt lanyard are useful for attaching tool bags and equipment to the tower at convenient spots. These hooks are better than using rope and tying knots because in many cases they can be hooked and unhooked with one hand.

Many hams use climbing belts such as shown in Fig 1. But fully integrated fall-arrest and positioning safety harnesses are preferable. The model ASL-301 in **Fig 2** has a D-ring on the back of the harness to which a safety lanyard is attached. These harnesses are available through Champion Radio Products, Box 572, Woodinville, WA 98072, **www.championradio.com**.

Rope and Pulley

Every amateur who owns a tower should also own a

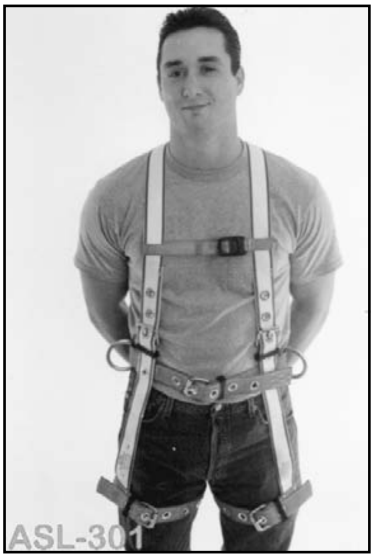

Fig 2—Fall-arrest safety harness integrated with positioning safety belt for tower climbing. ***(Courtesy of Champion Radio Products)***

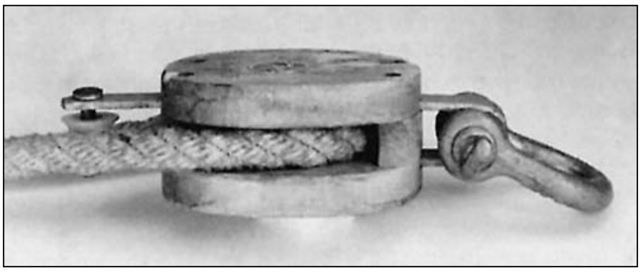

Fig 3—A good quality rope and pulley are essential for anyone working on towers and antennas. This pulley is encased in wood so the rope cannot jump out of the pulley wheel and jam.

good quality rope at least twice as long as the tower height. The rope is essential for safely erecting towers and installing antennas and cables. For most installations, a good quality $^{1}/_{2}$-inch diameter manila hemp rope will do the job, although a thicker rope is stronger and may be easier to handle. Some types of polypropylene rope are acceptable also; check the manufacturer's strength ratings. Nylon rope is not recommended because it tends to stretch and cannot be securely knotted without difficulty.

Check your rope before each use for tearing or chafing. Do not attempt to use damaged rope; if it breaks with a tower section or antenna in mid-air, property damage and personal injury are likely results. If your rope should get wet, let it air dry thoroughly before putting it away.

Another very worthwhile purchase is a pulley like the one shown in **Fig 3**. Use the right size pulley for your rope. Be sure that the pulley you purchase will not jam or bind as the rope passes through it.

THE GIN POLE

A gin pole, like the one shown in **Fig 4**, is a handy device for working with tower sections and masts. This gin pole is designed to clamp onto one leg of Rohn No. 25 or 45 tower. The tubing, which is about 12 feet long,

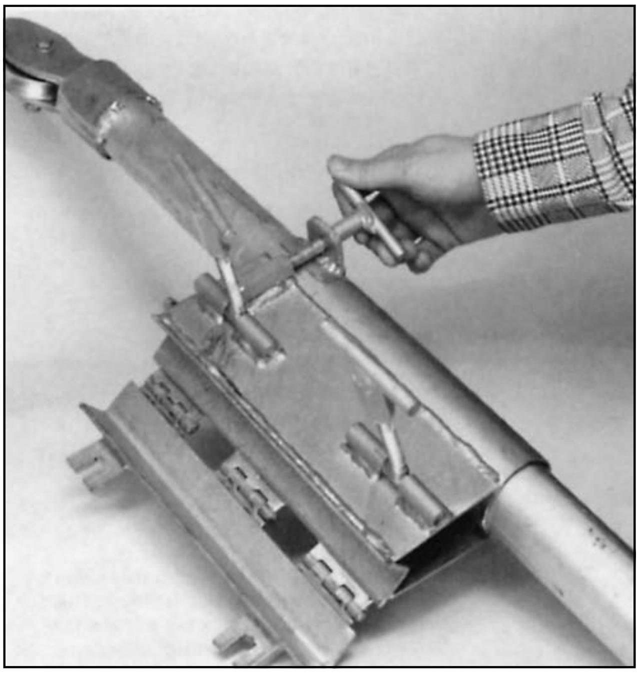

Fig 4—A gin pole is a mechanical device that can be clamped to a tower leg to aid in the assembly of sections as well as the installation of the mast. The aluminum tubing extends through the clamp and may be slipped into position before the tubing clamp is tightened. A rope should be routed through the tubing and over the pulley mounted at the top.

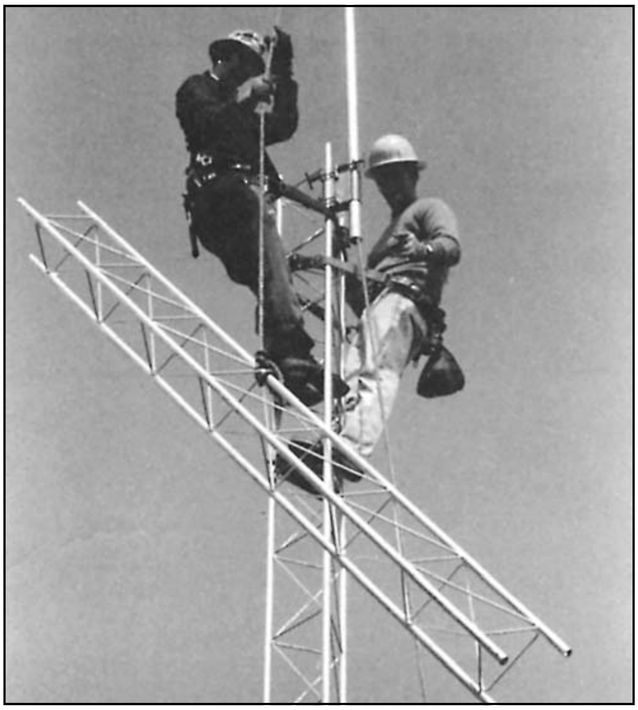

Fig 5—The assembly of tower sections is made simple when a gin pole is used to lift each one into position. Note that the safety belts of both climbers are fastened below the pole, thereby preventing the strap from slipping over the top section. (*Photo by K1WA*)

has a pulley on one end. The rope is routed through the tubing and over the pulley. When the gin pole is attached to the tower and the tubing extended into place, the rope may be used to haul tower sections or the mast into place. **Fig 5** shows the basic process. A gin pole can be expensive for an individual to buy, especially for a one-time tower installation. Some radio clubs own a gin pole for use by their members. Stores that sell tower sections to amateurs and commercial customers frequently will rent a gin pole to erect the tower. If you attempt to make your own gin pole, use materials heavy enough for the job. Provide a means for securely clamping the pole to the tower. There are many cases on record where homemade gin poles have failed, sending tower sections crashing down amidst the ground crew.

When you use a gin pole, make every effort to keep the load as vertical as possible. Although gin poles are strong, you are asking for trouble if you apply too much lateral force.

INSTALLING ANTENNAS ON THE TOWER

All antenna installations are different in some respects. Therefore, thorough planning is the most important first step in installing any antenna. At the beginning, before anyone climbs the tower, the whole process should be thought through. The procedure should be discussed to be sure each crew member understands what is to be done. Plan how to work out all bugs. Consider what tools and parts must be assembled and what items must be taken up the tower. Extra trips up and down the tower can be avoided by using forethought.

Getting ready to raise a beam requires planning. Done properly, the actual work of getting the antenna into position can be accomplished quite easily with only one person at the top of the tower. The trick is to let the ground crew do all the work and leave the person on the tower free to guide the antenna into position.

Before the antenna can be hoisted into position, the tower and the area around it must be prepared. The ground crew should clear the area around the base while someone climbs the tower to remove any wire antennas or other objects that might get in the way. The first person to climb the tower should also rig the rope and pulley that will be used to raise the antenna. The time to prepare the tower is before the antenna leaves the ground, not after it becomes hopelessly entwined with your 3.5-MHz dipole.

SOME TOWER CLIMBING TIPS

The following tower climbing safety tips were compiled by Tom Willeford, N8ETU. The most important safety factor in any kind of hazardous endeavor is the right attitude. Safety is important and worthy of careful consideration and implementation. The right attitude toward safety is a requirement for tower climbers. Lip service won't do, however; safety must be practiced.

The safe ham's safety attitude is simple: *Don't take any unnecessary chances*. There are no exceptions to this plain and simple rule. It is the first rule of safety and, of course, of climbing. The second rule is equally simple: *Don't be afraid to terminate an activity* (climbing, in this case) at any time if things don't seem to be going well.

Take time to plan your climb; this time is never wasted, and it's the first building block of safety. Talk the climb over with friends who will be helping you. Select the date and alternative dates to do the work. Choose someone to be responsible for all activities on the ground and for all communication with the climbers. Study the structure to be climbed and choose the best route to your objective. Plan emergency ascent and descent paths and methods.

Make a list of emergency phone numbers to keep by your phone, even though they may never be used. Develop a plan for rescuing climbers from the structure, should that become necessary.

Give careful thought to how much time you will need to complete the project. Allow enough time to go up, do the work, and then climb down during daylight hours. Include time for resting during the climb and for completing the work in a quality fashion. Remember that the

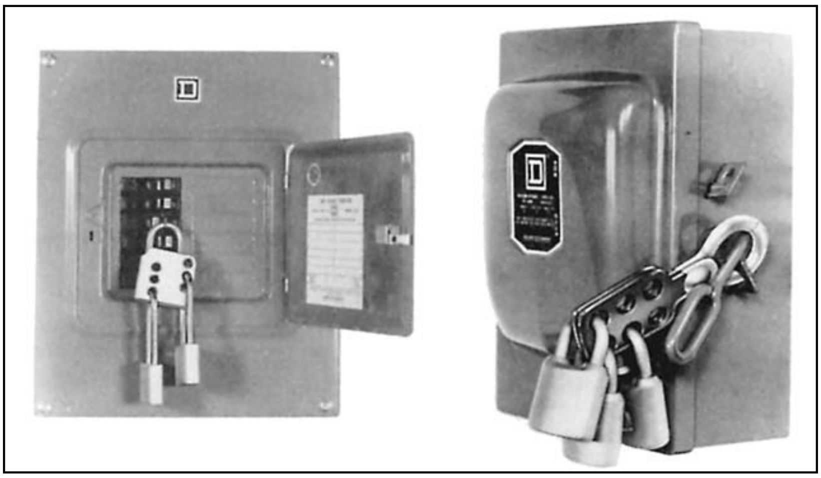

Fig 6—If the switch box feeding power to equipment on your tower is equipped with a lock-out hole, use it. With a lock through the hole on the box, the power cannot be accidentally turned back on. (*Photos courtesy of American ED CO®, at left, and Osborn Mfg Corp, at right.*)

temperature changes fast as the sun goes down. Climbing up or down a tower with cold hands and feet is very difficult—and dangerous.

Give careful consideration to the weather, and climb only in good weather. Investigate wind conditions, the temperature, and the weather forecast. The weather can change quickly, so if you're climbing a really tall tower, it may be a good idea to have a weather alert radio handy during the climb. *Never* climb a wet tower.

The person who is going to do the climbing should be the one to disconnect and tag all sources of power to the structure. All switches or circuit breakers should be labeled clearly with DO NOT TOUCH instructions. Use locks on any switches designed to accept them. (See **Fig 6**.) Only the climber should reconnect power sources.

An important part of the climbing plan is to review notes on the present installation and any previous work. It's a good idea to keep a notebook, listing every bolt and nut size on your tower/antenna installation. Then, when you have to go up to make repairs, you'll be able to take the minimum number of tools with you to do the job. If you take too many tools up the tower, there is a much greater chance of dropping something, risking injury to the ground crew and possibly damaging the tool.

It is also a good idea to review the instruction sheets and take them with you. In other words, plan carefully what you are going to do, and what you'll need to do it efficiently and safely.

It's better to use a rope and pulley to hoist tools. Climbing is hard work and there's no sense making it more difficult by carrying a big load of tools. Always rig the pulley and rope so the ground crew raises and lowers tools and equipment.

Climbing Equipment

Equipment is another important safety consideration. By equipment, we don't just mean tools. We mean safety equipment. Safety equipment should be selected and cared for as if your life depends on it—because it does!

The list of safety equipment essential to a safe climb and safe work on the tower should include:

1) A first class safety belt,
2) Safety glasses,
3) Hard hat,
4) Long-sleeved, pull-over shirt with no buttons or openings to snag (long sleeves are especially important for climbing wooden poles),
5) Long pants without cuffs,
6) Firm, comfortable, steel-shank shoes with no-slip soles and well-defined heels, and
7) Gloves that won't restrict finger movement (insulated gloves if you MUST work in cold weather).

Your safety belt should be approved for use on the structure you are climbing. Different structures may require different types of safety hooks or straps. The belt should be light weight, but strength should not be sacrificed to save weight. It should fit you comfortably. All moving parts, such as snap hooks, should work freely. You should inspect safety belts and harnesses carefully and thoroughly before each climb, paying particular attention to stitching, rivets and weight-bearing mechanical parts.

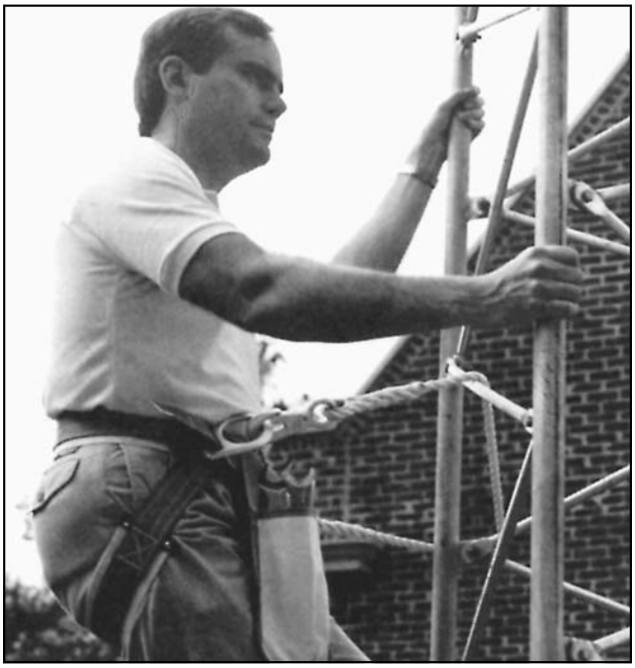

Fig 7—Mark Wilson, K1RO, shows the proper way to attach a safety hook, with the hook opening facing away from the tower. That way the hook can't be accidentally released by pressing it against a tower leg.

Support belt hooks should always be hooked to the D rings in an outward configuration. That is, the opening part of the hook should face away from the tower when engaged in the D rings (see **Fig 7**). Hooks engaged this way are easier to unhook deliberately but won't get squeezed open by a part of the tower or engage and snag a part of the tower while you are climbing. The engagement of these hooks should always be checked visually. A snapping hook makes the same sound whether it's engaged or not. Never check by sound—look to be sure the hook is engaged properly before trusting it.

Remember that the D rings on the safety belt are for support hooks *only*. No tools or lines should be attached to these hooks. Such tools or lines may prevent the proper engagement of support belt hooks, or they may foul the hooks. At best, they could prevent the release of the hooks in an emergency. No one should have to disconnect a support hook to get a tool and then have to reconnect the support hook before beginning to work again. That's foolish.

Equipment you purchase new is best. Homemade belts or home-spliced lines are dangerous. Used belts may have worn or defective stitching, or other faulty components. Be careful of so-called "bargains" that could cost you your life.

Straps, lanyards and lines should be as short as possible. Remember, in general knots reduce the load strength of a line by approximately 50%.

Before actually climbing, check the structure visually. Review the route. Check for obstacles, both natural (like wasp's nests) and man-made. Check the structure supports and add more if necessary. Guy wires can be obstacles to the climb, but it's better to have too many supports than not enough. Check your safety belt, support belts and hooks at the base of the tower. Really test them before you need them. Never leave the ground without a safety belt—even 5 or 10 feet. After all of this, the climb will be a "cake walk" if you are careful.

Climb slowly and surely. Don't overreach or overstep. Patience and watchfulness is rewarded with good hand and foot holds. Take a lesson from rock climbers. Hook on to the tower and rest periodically during the climb. Don't try to rest by wedging an arm or leg in some joint; to rest, hook on. Rests provide an opportunity to review the remainder of the route and to make sure that your safety equipment feels good and is working properly. Rest periods also help you conserve a margin of energy in case of difficulty.

Finally, keep in mind that the most dangerous part of working on a tower occurs when you are actually climbing. Your safety equipment is not hooked up at this time, so be extra careful during the ascent or descent.

You must climb the tower to install or work on an antenna. Nevertheless, any work that can be done on the ground should be done there. If you can do any assembly or make any adjustments on the ground, that's where you should do the work! The less time you have to spend on the tower, the better off you'll be.

When you arrive at the work area, hook on to the tower and review what you have to do. Determine the best position to do the work from, disconnect your safety strap and move to that position. Then reconnect your safety strap at a safe spot, away from joints and other obstacles. If you must move around an obstacle, try to do it while hooked on to the tower. Find a comfortable position and go to work. Don't overreach—move to the work.

Use the right tool for the task. If you don't have it, have the ground crew haul it up. Be patient. Lower tools, don't drop them, when you are finished with them. Dropped tools can bounce and cause injury or damage, or can be broken or lost. It's a good idea to tie a piece of string or light rope to the tools, and to tie the other end to the tower or some other point so if you do drop a tool, it won't fall all the way to the ground. Don't tie tools to the D ring or your safely belt, however!

Beware of situations where an antenna may be off balance. It's hard to obtain the extra leverage needed to handle even a small beam when you are holding it far from the balance point. Leverage can apply to the climber as well as the device being levered. Many slips and skinned knuckles result from such situations. A severely injured hand or finger can be a real problem to a climber.

Before descending, be sure to check all connections and the tightness of all the bolts and nuts that you have worked with. Have the ground crew use the rope and pulley to lower your tools. Lighten your load as much as possible. Remember, you're more tired coming down than going up. While still hooked on, wiggle your toes and move a little to get your senses working again. Check your downward route and begin to descend slowly and even more surely than you went up. Rest is even more important during the descent.

The ground captain is the director of all activities on the ground, and should be the only one to communicate with a person on a tower. Hand-held transceivers can be very helpful for this communication, but no one else should transmit to the workers on the tower. Even minor confusion or misunderstanding about a move to be made could be very dangerous.

"Antenna parties" can be lots of fun, but the joking and fooling around should wait until the job is done and everyone is down safely. Save the celebrating until after the work is completed, even for the ground crew.

These are just a few ideas on tower climbing safety; no list can include everything that you might run into. Check Chapter 22 for additional ideas. *Just remember—you can't be too careful when climbing*. Keep safety in mind while doing antenna work, and help ensure that after you have fallen *for* ham radio, you don't fall *from* ham radio.

THE TOWER SHIELD

A tower can be legally classified as an "attractive nuisance" that could cause injuries. You should take some precautions to ensure that "unauthorized climbers" can't get hurt on your tower. This tower shield was originally described by Baker Springfield, W4HYY, and Richard Ely, WA4VHM, in September 1976 *QST*, and should eliminate the worry.

Generally, the attractive-nuisance doctrine applies to your responsibility to trespassers on your property. (The law is much stricter with regard to your responsibility to an invited guest.) You should expect your tower to attract children, whether they are already technically trespassing

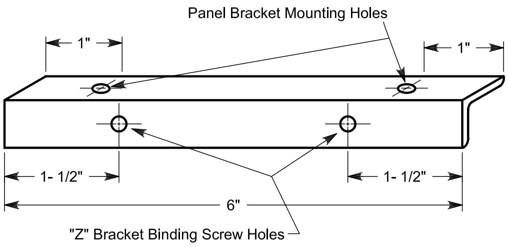

Fig 8—Z-bracket component pieces.

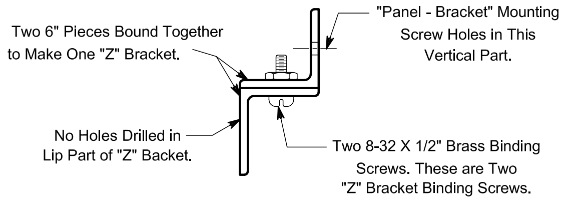
Two 6" Pieces Bound Together
to Make One "Z" Bracket.
"Panel - Bracket" Mounting
Screw Holes in This
Vertical Part.
No Holes Drilled in
Lip Part of "Z" Backet.
Two 8-32 X 1/2" Brass Binding
Screws. These are Two
"Z" Bracket Binding Screws.

Fig 9—Assembly of the Z bracket.

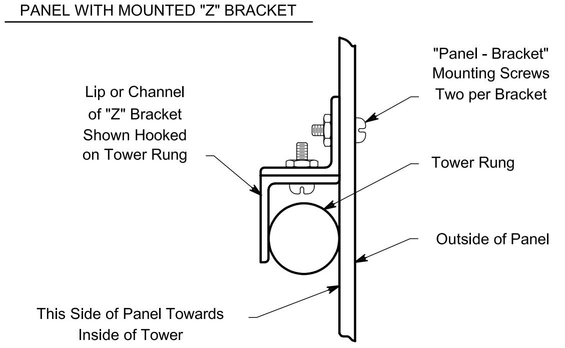
PANEL WITH MOUNTED "Z" BRACKET
"Panel - Bracket"
Mounting Screws
Two per Bracket
Lip or Channel
of "Z" Bracket
Shown Hooked
on Tower Rung
Tower Rung
Outside of Panel
This Side of Panel Towards
Inside of Tower

Fig 10—Installation of the shield on a tower rung.

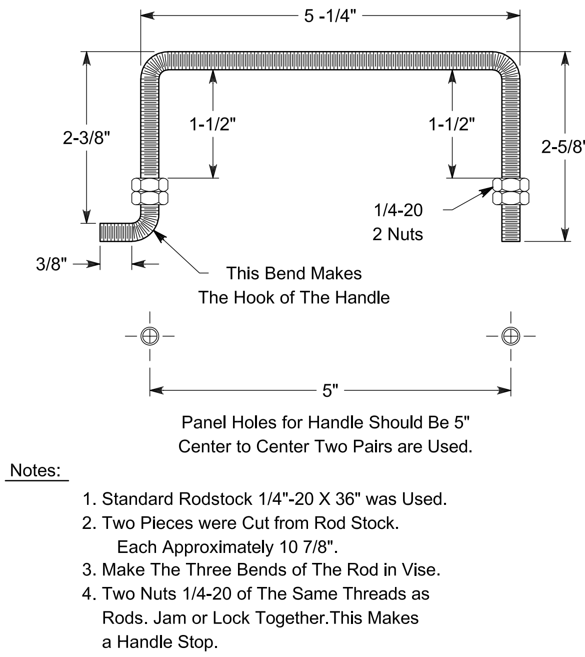
5 -1/4"
2-3/8"
1-1/2"
1-1/2"
2-5/8"
1/4-20
2 Nuts
3/8"
This Bend Makes
The Hook of The Handle
5"
Panel Holes for Handle Should Be 5"
Center to Center Two Pairs are Used.
Notes:
1. Standard Rodstock 1/4"-20 X 36" was Used.
2. Two Pieces were Cut from Rod Stock.
Each Approximately 10 7/8".
3. Make The Three Bends of The Rod in Vise.
4. Two Nuts 1/4-20 of The Same Threads as
Rods. Jam or Lock Together.This Makes
a Handle Stop.

Fig 11—Removable handle construction.

or whether the tower itself lures them onto your property. A tower is dangerous to children, especially because of their inability to appreciate danger. (What child could resist trying to climb a tower once they see one?) Because of this danger, you have a legal duty to exercise reasonable care to eliminate the danger or otherwise protect children against the perils of the attraction.

The tower shield is composed simply of panels that enclose the tower and make climbing practically impossible. These panels are 5 feet in height and are wide enough to fit snugly between the tower legs and flat against the rungs. A height of 5 feet is sufficient in almost every case. The panels are constructed from 18-gauge galvanized sheet metal obtained and cut to proper dimensions from a local sheet-metal shop. A lighter gauge could probably be used, but the extra physical weight of the heavier gauge is an advantage if no additional means of securing the panels to the tower rungs are used. The three types of metals used for the components of the shield are supposedly rust proof and nonreactive. The panels are galvanized sheet steel, the brackets aluminum, and the screws and nuts are brass. For a triangular tower, the shield consists of three panels, one for each of the three sides, supported by two brackets. Construct these brackets from 6-inch pieces of thin aluminum angle stock. Bolt two of the pieces together to form a Z bracket (see **Figs 8**, **9** and **10**). The Z brackets are bolted together with binding head brass machine screws.

Lay the panels flat for measuring, marking and drilling. First measure from the top of the upper mounting rung on the tower to the top of the bottom rung. (Mounting rungs are selected to position the panel on the tower.) Then mark this distance on the panels. Use the same size brass screws and nuts throughout the shield. Bolt the top vertical portion of each Z bracket to the panel. Drill the mounting-screw holes about 1 inch from the end of the Z

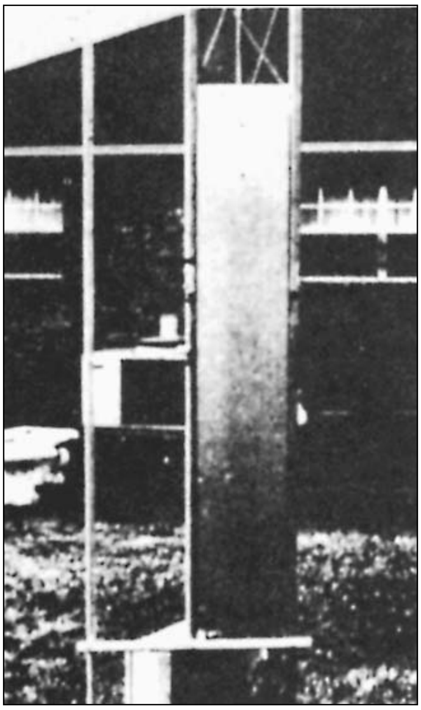

Fig 12—Installed tower shield. Note the holes for using the handles.

brackets so there is an offset clearance between the Z-bracket binding-screw holes and the panel-bracket mounting-screw holes. Drill holes in each panel to match the Z-bracket holes.

The panels are held on the tower by their own weight. They are not easy to grasp because they fit snugly between the tower legs. If you feel a need for added safety against deliberate removal of the panels, this can be accomplished by means of tie wires. Drill a small hole in the panel just above, just below, and in the center of each Z bracket. Run a piece of heavy galvanized wire through the top hole, around the Z bracket, and then back through the hole just below the Z bracket. Twist together the two ends of the wire. One tie wire should be sufficient for each panel, but use two if desired.

The completed panels are rather bulky and difficult to handle. A feature that is useful if the panels have to be removed often for tower climbing or accessibility is a pair of removable handles. The removable handles can be constructed from one threaded rod and eight nuts (see **Fig 11**). Drill two pair of handle holes in the panels a few inches below the top Z bracket and several inches above the bottom Z bracket. For panel placement or removal, you can hook the handles in these panel holes. The hook, on the top of the handle, fits into the top hole of each pair of the handle holes. The handle is optional, but for the effort required it certainly makes removal and replacement much safer and easier.

Fig 12 shows the shield installed on a tower. This relatively simple device could prevent an accident.

Electrical Safety

Although the RF, ac and dc voltages in most amateur stations pose a potentially grave threat to life and limb, common sense and knowledge of safety practices will help you avoid accidents. Building and operating an Amateur Radio station can be, and is for almost all amateurs, a perfectly safe pastime. However, carelessness can lead to severe injury, or even death. The ideas presented here are only guidelines; it would be impossible to cover all safety precautions. Remember, there is no substitute for common sense.

A fire extinguisher is a requirement for the well-equipped amateur station. The fire extinguisher should be of the carbon-dioxide type to be effective in electrical fires. Store it in an easy-to-reach spot and check it at recommended intervals.

Family members should know how to turn the power off in your station. They should also know how to apply artificial respiration. Many community groups offer courses on cardiopulmonary resuscitation (CPR).

AC AND DC SAFETY

The primary wiring for your station should be controlled by one master switch, and other members of your household should know how to kill the power in an emergency. All equipment should be connected to a good ground. All wires carrying power around the station should be of the proper size for the current to be drawn and should be insulated for the voltage level involved. Bare wire, open-chassis construction and exposed connections are an invitation to accidents. Remember that high-current, low-voltage power sources are just as dangerous as high-voltage, low-current sources. Possibly the most-dangerous voltage source in your station is the 120-V primary supply; it is a hazard often overlooked because it is a part of everyday life. Respect even the lowliest power supply in your station.

Whenever possible, kill the power and unplug equipment before working on it. Discharge capacitors with an insulated screwdriver; don’t assume the bleeder resistors are 100% reliable. In a power amplifier, always short the tube plate cap to ground just to be sure the supply is discharged. If you must work on live equipment, keep one hand in your pocket. Avoid bodily contact with any grounded object to prevent your body from becoming the return path from a voltage source to ground. Use insulated tools for adjusting or moving any circuitry. Never work alone. Have someone else present; it could save your life in an emergency.

National Electrical Code

The National Electrical Code® is a comprehensive document that details safety requirements for all types of electrical installations. In addition to setting safety standards for house wiring and grounding, the Code also contains a section on Radio and Television Equipment — Article 810. Sections C and D specifically cover Amateur Transmitting and Receiving Stations. Highlights of the section concerning Amateur Radio stations follow. If you are interested in learning more about electrical safety, you may purchase a copy of *The National Electrical Code* or *The National Electrical Code Handbook*, edited by Peter Schram, from the National Fire Protection Association, Batterymarch Park, Quincy, MA 02269.

Antenna installations are covered in some detail in the Code. It specifies minimum conductor sizes for different length wire antennas. For hard-drawn copper wire, the Code specifies #14 wire for open (unsupported) spans less than 150 feet, and #10 for longer spans. Copper-clad steel, bronze or other high-strength conductors may be #14 for spans less than 150 feet and #12 wire for longer runs. Lead-in conductors (for open-wire transmission line) should be at least as large as those specified for antennas.

The Code also says that antenna and lead-in conductors attached to buildings must be firmly mounted at least 3 inches clear of the surface of the building on nonabsorbent insulators. The only exception to this minimum distance is when the lead-in conductors are enclosed in a "permanently and effectively grounded" metallic shield. The exception covers coaxial cable.

According to the Code, lead-in conductors (except those covered by the exception) must enter a building through a rigid, noncombustible, nonabsorbent insulating tube or bushing, through an opening provided for the purpose that provides a clearance of at least 2 inches or through a drilled window pane. All lead-in conductors to transmitting equipment must be arranged so that accidental contact is difficult.

Transmitting stations are required to have a means of draining static charges from the antenna system. An antenna discharge unit (lightning arrester) must be installed on each lead-in conductor (except where the lead-in is protected by a continuous metallic shield that is permanently and effectively grounded, or the antenna is permanently and effectively grounded). An acceptable alternative to lightning arrester installation is a switch that connects the lead-in to ground when the transmitter is not in use.

Grounding conductors are described in detail in the Code. Grounding conductors may be made from copper, aluminum, copper-clad steel, bronze or similar erosion-resistant material. Insulation is not required. The "protective grounding conductor" (main conductor running to the ground rod) must be as large as the antenna lead-in, but not smaller than #10. The "operating grounding conductor" (to bond equipment chassis together) must be at least #14. Grounding conductors must be adequately supported and arranged so they are not easily damaged. They must run in as straight a line as practical between the mast or discharge unit and the ground rod.

The Code also includes some information on safety inside the station. All conductors inside the building must be at least 4 inches away from conductors of any lighting or signaling circuit except when they are separated from other conductors by conduit or a nonconducting material. Transmitters must be enclosed in metal cabinets, and the cabinets must be grounded. All metal handles and controls accessible by the operator must be grounded. Access doors must be fitted with interlocks that will disconnect all potentials above 350 V when the door is opened.

Ground

An effective ground system is necessary for every amateur station. The mission of the ground system is twofold. First, it reduces the possibility of electrical shock if something in a piece of equipment should fail and the chassis or cabinet becomes "hot." If connected properly, three-wire electrical systems ground the chassis, but older amateur equipment may use the ungrounded two-wire system. A ground system to prevent shock hazards is generally referred to as "dc ground."

The second job the ground system must perform is to provide a low-impedance path to ground for any stray RF current inside the station. Stray RF can cause equipment to malfunction and contributes to RFI problems. This low-impedance path is usually called "RF ground." In most stations, dc ground and RF ground are provided by the same system.

The first step in building a ground system is to bond together the chassis of all equipment in your station. Ordinary hookup wire will do for a dc ground, but for a good RF ground you need a low-impedance conductor. Copper strap, sold as "flashing copper," is excellent for this application, but it may be hard to find. Braid from coaxial cable is a popular choice; it is readily available, makes a low-impedance conductor, and is flexible.

Grounding straps can be run from equipment chassis to equipment chassis, but a more convenient approach is illustrated in **Fig 13**. In this installation, a $^1/_2$-inch diameter copper water pipe runs the entire length of the operating bench. A thick braid (from discarded RG-8 cable) runs from each piece of equipment to a clamp on the pipe. Copper water pipe is available at most hardware stores and home centers. Alternatively, a strip of flashing copper may be run along the rear of the operating bench.

After the equipment is bonded to a common ground bus, the ground bus must be wired to a good earth ground. This run should be made with a heavy conductor (braid is a popular choice, again) and should be as short and direct as possible. The earth ground usually takes one of two forms.

In most cases, the best approach is to drive one or more ground rods into the earth at the point where the conductor from the station ground bus leaves the house. The best ground rods to use are those available from an electrical supply house. These rods are 8 to 10 feet long and are made from steel with a heavy copper plating. Do not depend on shorter, thinly plated rods sold by some home electronics suppliers. These rods begin to rust almost immediately after they are driven into the soil, and they become worthless within a short time. Good ground rods, while more expensive initially, offer long-term protection.

If your soil is soft and contains few rocks, an acceptable alternative to "genuine" ground rods is $^1/_2$-inch diameter copper water pipe. A 6- to 8-foot length of this material offers a good ground, but it may bend while being driven into the earth. Some people have recommended that you make a connection to a water line and run water down through the copper pipe so that it forces its own

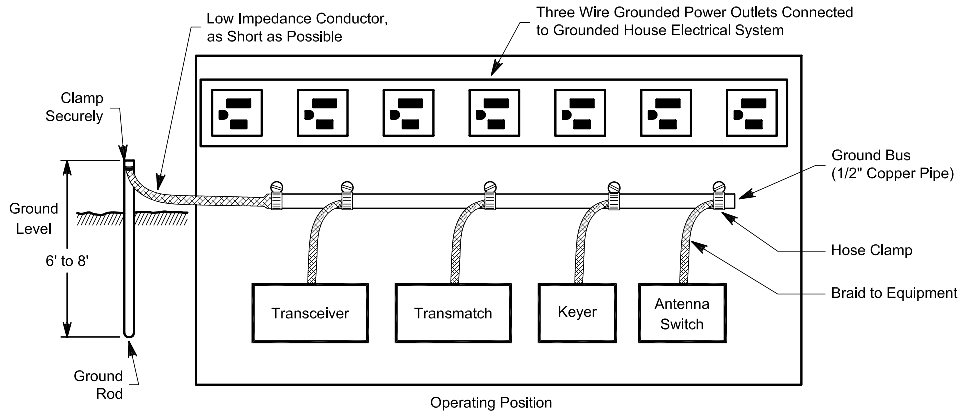
Low Impedance Conductor, as Short as Possible
Three Wire Grounded Power Outlets Connected to Grounded House Electrical System
Clamp Securely
Ground Bus (1/2" Copper Pipe)
Ground Level
6' to 8'
Hose Clamp
Braid to Equipment
Transceiver
Transmatch
Keyer
Antenna Switch
Ground Rod
Operating Position

Fig 13—An effective station ground bonds the chassis of all equipment together with low-impedance conductors and ties into a good earth ground.

hole in the ground. There may be a problem with this method, however. When the ground dries, it may shrink away from the pipe and not make proper contact with the ground rod. This would provide a rather poor ground.

Once the ground rod is installed, clamp the conductor from the station ground bus to it with a clamp that can be tightened securely and will not rust. Copper-plated clamps made especially for this purpose are available from electrical supply houses, but a stainless-steel hose clamp will work too. Alternatively, drill several holes through the pipe and bolt the conductor in place.

Another popular station ground is the cold water pipe system in the building. To take advantage of this ready-made ground system, run a low-impedance conductor from the station ground bus to a convenient cold water pipe, preferably somewhere near the point where the main water supply enters the house. Avoid hot water pipes; they do not run directly into the earth. The advent of PVC (plastic) plumbing makes it mandatory to inspect the cold water system from your intended ground connection to the main inlet. PVC is an excellent insulator, so any PVC pipe or fittings rule out your cold water system for use as a station ground.

For some installations, especially those located above the first floor, a conventional ground system such as that just described will make a fine dc ground but will not provide the necessary low-impedance path to ground for RF. The length of the conductor between the ground bus and the ultimate ground point becomes a problem. For example, the ground wire may be about $^1/_4$ λ (or an odd multiple of $^1/_4$ λ) long on some amateur band. A $^1/_4$-λ wire acts as an impedance inverter from one end to the other. Since the grounded end is at a very low impedance, the equipment end will be at a high impedance. The likely result is RF hot spots around the station while the transmitter is operating. A ground system like this may be worse than having no ground at all.

An alternative RF ground system is shown in **Fig 14**. Connect a system of $^1/_4$-λ radials to the station ground bus. Install at least one radial for each band used. You should still be sure to make a connection to earth ground for the ac power wiring. Try this system if you have problems with RF in the shack. It may just solve a number of problems for you. Be careful, however, to prevent contact with the ends of the radials, where there is high-voltage RF for powers greater than QRP level.

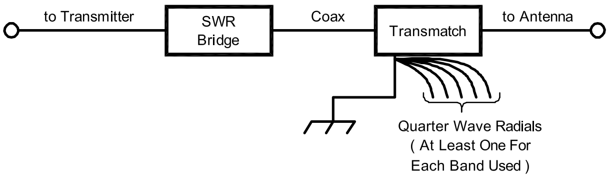
to Transmitter
SWR Bridge
Coax
Transmatch
to Antenna
Quarter Wave Radials (At Least One For Each Band Used)

Fig 14—Here is an alternative to earth ground if the station is located far from the ground point and RF in the station is a problem. Install at least one $^1/_4$-λ radial for each band used.

Ground Noise

Noise in ground systems can affect sensitive radio equipment. It is usually related to one of three problems:

1) Insufficient ground conductor size,
2) Loose ground connections, or
3) Ground loops.

These matters are treated in precise scientific research equipment and some industrial instruments by paying attention to certain rules. The ground conductor should be at least as large as the largest conductor in the primary power circuit. Ground conductors should provide a solid connection to both ground and to the equipment being grounded. Liberal use of lock washers and star washers is highly recommended. A loose ground connection is a tremendous source of noise, particularly in a sensitive receiving system.

Ground loops should be avoided at all costs. A short discussion of what a ground loop is and how to avoid them may lead you down the proper path. A ground loop is formed when more than one ground current is flowing in a single conductor. This commonly occurs when grounds are “daisy-chained” (series linked). The correct way to ground equipment is to bring all ground conductors out radially from a common point to either a good driven earth ground or to a cold water system.

Ground noise can affect transmitted as well as received signals. With the low audio levels required to drive amateur transmitters, and with the ever-increasing sensitivity of our receivers, correct grounding is critical.

Lightning and EMP Protection

The National Fire Protection Association (NFPA) publishes a booklet called *Lightning Protection Code* (NFPA no. 78-1983) that should be of interest to radio amateurs. For information about obtaining a copy of this booklet, write to the National Fire Protection Association, Batterymarch Park, Quincy, MA 02269. Two paragraphs of particular interest to amateurs are presented below:

“3-26 Antennas. Radio and television masts of metal, located on a protected building, shall be bonded to the lightning protection system with a main size conductor and fittings.

“3-27 Lightning arresters, protectors or antenna discharge units shall be installed on electric and telephone service entrances and on radio and television antenna lead-ins.”

The best protection from lightning is to disconnect all antennas from equipment and disconnect the equipment from the power lines. Ground antenna feed lines to safely bleed off static buildup. Eliminate the possible paths for lightning strokes. Rotator cables and other control cables from the antenna location should also be disconnected during severe electrical storms.

In some areas, the probability of lightning surges entering homes via the 120/240-V line may be high. Lightning produces both electrical and magnetic fields that vary with distance. These fields can be coupled into power lines and destroy electronic components in equipment that is miles from where the lightning occurred. Radio equipment can be protected from these surges to some extent by using transient-protective devices.

ELECTROMAGMETIC PULSE AND THE RADIO AMATEUR

The following material is based on a 4-part *QST* article by Dennis Bodson, W4PWF, that appeared in the August through November 1986 issues of *QST*. The series was condensed from the National Communications System report NCS TIB 85-10.

An equipment test program demonstrated that most Amateur Radio installations can be protected from lightning and electromagnetic pulse (EMP) transients with a basic protection scheme. Most of the equipment is not susceptible to damage when all external cabling is removed. You can duplicate this stand-alone configuration simply by unplugging the ac power cord from the outlet, disconnecting the antenna feed line at the rear of the radio, and isolating the radio gear from any other long metal conductors. Often it is not practical to completely disconnect the equipment whenever it is not being used. Also, there is the danger that a lightning strike several miles away could induce a large voltage transient on the power lines or antenna while the radio is in use. You can add two transient-protection devices to the interconnected system, however, that will also closely duplicate the safety of the stand-alone configuration.

The ac power line and antenna feed line are the two important points that should be outfitted with transient protection. This is the minimum basic protection scheme recommended for all Amateur Radio installations. (For fixed installations, consideration should also be given to the rotator connections—see **Fig 15**.) Hand-held radios equipped with a “rubber duck” require no protection at the antenna jack. If a larger antenna is used with the hand-held, however, a protection device should be installed.

General Considerations

Because of the unpredictable energy content of a nearby lightning strike or other large transient, it is possible for a metal-oxide varistor (MOV) to be subjected to an energy surge in excess of its rated capabilities. This may result in the destruction of the MOV and explosive rupture of the package. These fragments can cause damage to nearby components or operators and possibly ignite flammable material. Therefore, the MOV should be physically shielded.

A proper ground system is a key factor in achieving protection from lightning and EMP transients. A low-impedance ground system should be installed to eliminate transient paths through radio equipment and to provide a good physical ground for the transient-suppression devices. A single-point ground system is recommended (**Fig 16**) to help prevent lightning from getting into the shack on the shields of antenna feed-line coaxes. Many hams use a well-grounded radio-entrance panel mounted *outside* the shack to ground their coaxes before they enter the house. **Fig 17** shows an entrance panel at K8CH’s home in Michigan. All external conductors going to the radio equipment should enter and exit the station through this panel. Install all transient-suppression devices directly on the panel. Use

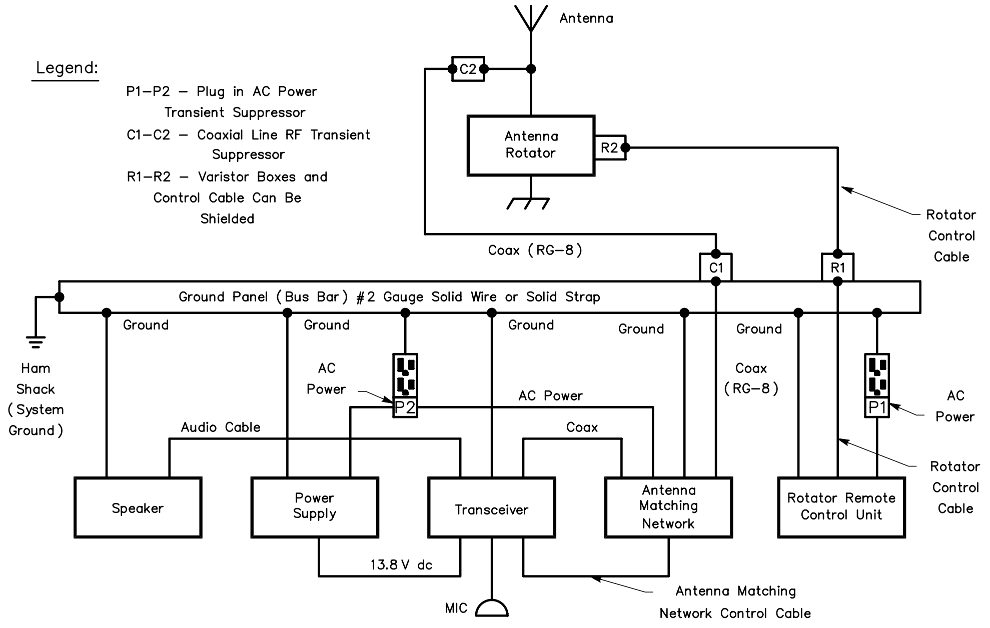
Antenna
Legend:
P1–P2 – Plug in AC Power Transient Suppressor
C1–C2 – Coaxial Line RF Transient Suppressor
R1–R2 – Varistor Boxes and Control Cable Can Be Shielded
C2
Antenna Rotator
R2
Rotator Control Cable
Coax (RG–8)
C1
R1
Ground Panel (Bus Bar) #2 Gauge Solid Wire or Solid Strap
Ground
Ground
Ground
Ground
Ground
Ham Shack (System Ground)
AC Power
P2
AC Power
Coax (RG–8)
P1
AC Power
Audio Cable
Coax
Rotator Control Cable
Speaker
Power Supply
Transceiver
Antenna Matching Network
Rotator Remote Control Unit
13.8 V dc
MIC
Antenna Matching Network Control Cable

Fig 15—Transient suppression techniques applied to an Amateur Radio station.

the shortest length(s) possible of #6 solid wire to connect the radio equipment case(s) to the ground bus.

Ac Power-Line Protection

Tests have indicated that household electrical wiring inherently limits the maximum transient current that it will pass to approximately 120 A. Therefore, if possible, the amateur station should be installed away from the house ac entrance panel and breaker box to take advantage of these limiting effects.

Ac power-line protection can be provided with easy-to-install, plug-in transient protectors. Ten such devices were tested for the article series in 1986. The plug-in-strip units are the best overall choice for a typical amateur installation. They provide the protection needed, they're simple to install and can be easily moved to other operating locations with the equipment.

In their tests, NCS found that the model that provided transient paths to ground from the hot and neutral lines (common mode) as well as the transient path between the hot and neutral lines (normal mode) performed best. The best model used three MOVs and a 3-electrode gas-discharge-tube arrester to provide fast operation and large power dissipation capabilities. This unit was tested repeatedly and operated without failure.

The flood of low-cost computers in the 1990s spawned a host of surge protection devices designed to limit transient voltage spikes coming from the ac line and also through the telephone line into a modem connected to a computer. Many of these devices are well-designed and can be relied upon to provide the protection they claim.

You can, however, easily find a variety of really low-cost bargain strips at flea markets and discount hardware stores. A bargain-brand $6 unit may prove to be a poor bargain indeed if it allows a spike to get through to damage your $2000 computer or $4000 transceiver.

You should be careful to find one that carries a sticker indicating that it meets Underwriters Laboratories safety standard UL 1449. This defines the minimum level of clamping voltage beyond which a surge protector will "fire" to protect the device connected to its output. The UL 1149 limit is 330 V ac. Prices for brand-name units from Tripp Lite, APC or Curtis vary from about $30 to $80, depending on how many ac sockets they have and the number of indicator lights and switched/unswitched sockets. A brand-name device is well worth the small additional cost over the bargain-basement units.

A transient suppressor requires a 3-wire outlet; the outlet should be tested to ensure all wires are properly connected. In older houses, an ac ground may have to be installed by a qualified electrician. The ac ground must be available for the plug-in transient suppressor to function properly. The ac ground of the receptacle should be attached to the station ground bus, and the plug-in receptacle should be installed on the ground panel behind the radio equipment.

Emergency Power Generators

Emergency power generators provide two major transient-protection advantages. First, the station is disconnected from the commercial ac power system. This isolates the radio equipment from a major source of damaging transients. Second, tests have shown that the emergency power generator may not be susceptible to EMP transients.

When the radio equipment is plugged directly into the generator outlets, transient protection may not be

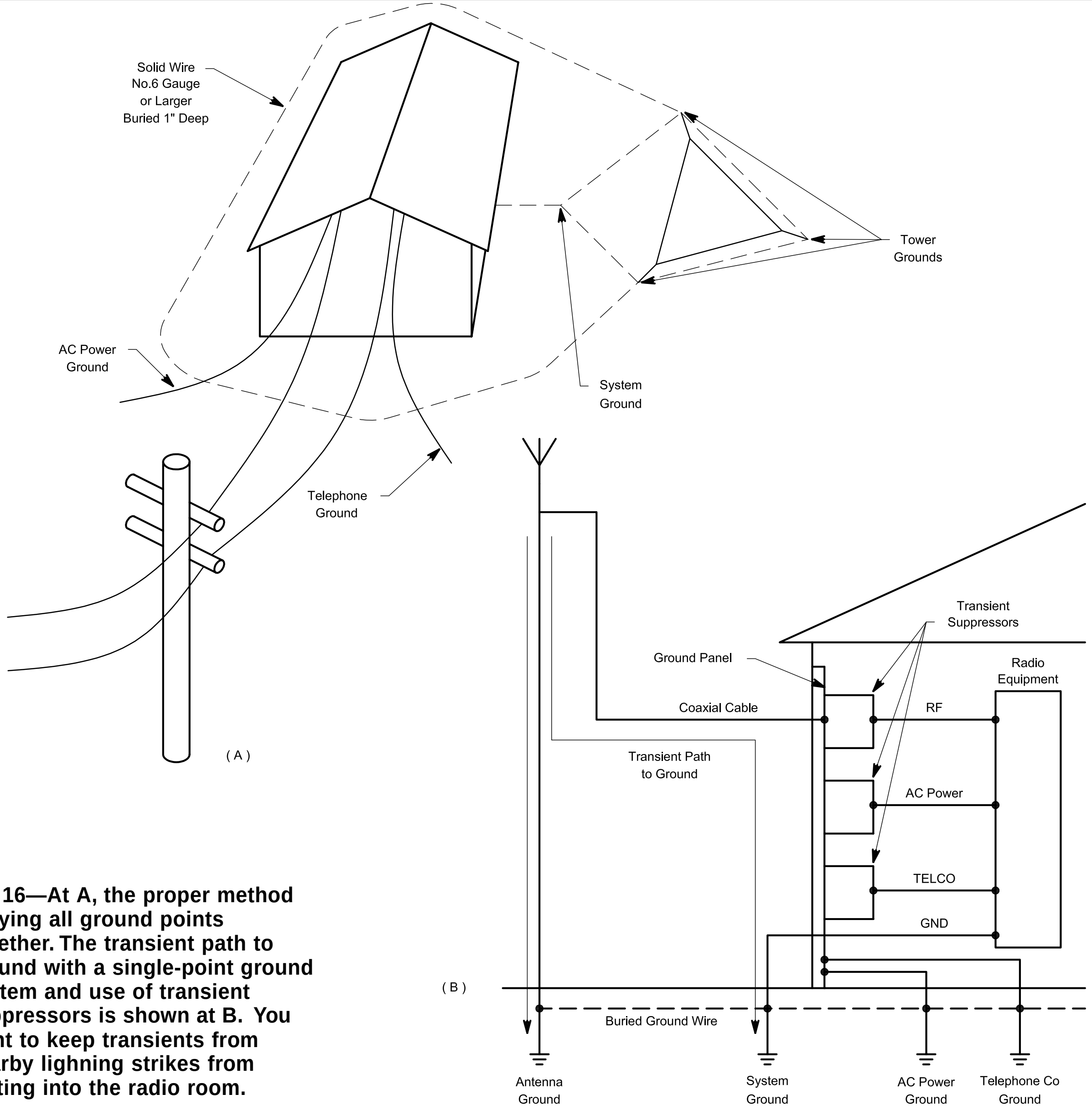
Solid Wire
No.6 Gauge
or Larger
Buried 1" Deep
Tower
Grounds
AC Power
Ground
System
Ground
Telephone
Ground
(A)
Transient
Suppressors
Ground Panel
Radio
Equipment
Coaxial Cable
RF
Transient Path
to Ground
AC Power
TELCO
GND
(B)
Buried Ground Wire
Antenna
Ground
System
Ground
AC Power
Ground
Telephone Co
Ground

Fig 16—At A, the proper method of tying all ground points together. The transient path to ground with a single-point ground system and use of transient suppressors is shown at B. You want to keep transients from nearby lightning strikes from getting into the radio room.

needed. If an extension cord or household wiring is used, transient protection should be employed.

An emergency power generator should be wired into the household circuit only by a qualified electrician. When properly connected, a switch is used to disconnect the commercial ac power source from the house lines before the generator is connected to them. This keeps the generator output from feeding back into the commercial power system. If this is not done, death or injury to unsuspecting linemen can result.

Feed-Line Protection

Coaxial cable is recommended for use as the transmission line because it provides a certain amount of transient surge protection for the equipment to which it is attached. The outer conductor shields the center conductor from the transient field. Also, the cable limits the maximum conducted transient voltage on the center by arcing the differential voltage from the center conductor to the grounded cable shield.

By providing a path to ground ahead of the radio equipment, the gear can be protected from the large currents impressed upon the antenna system by lightning and EMP. A single protection device installed at the radio antenna jack will protect the radio, but not the transmission line. To protect the transmission line, another tran-

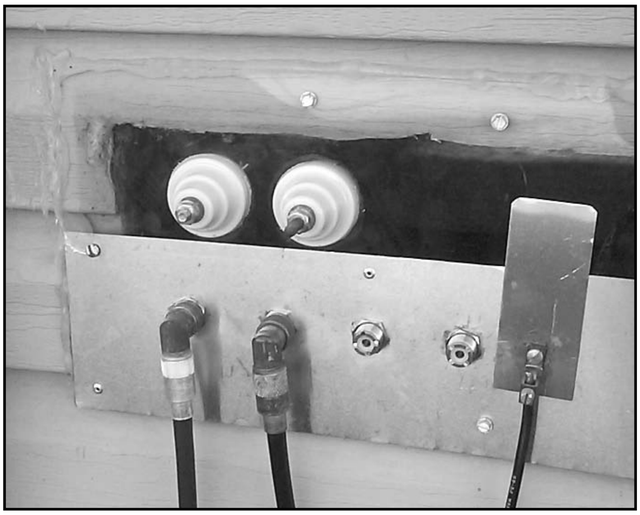

Fig 17—Radio-entrance panel at K8CH. A flat aluminum plate serves as a common grounding point for all coax cables to prevent transients from nearby lightning strikes from getting into the shack on the feed lines. "Spark-gap" protectors to kill transients on the coax inner conductors are located behind the panel. A spark-gap protector is also used on the two open-wire feeder insulators mounted on the Plexiglas sheet behind the aluminum panel.

sient protector must be installed between the antenna and the transmission line. (See Fig 15).

RF transient protection devices from several manufacturers were tested (see **Table 1**) using RG-8 cable equipped with UHF connectors. All of the devices shown can be installed in a coaxial transmission line. Recall that during the tests the RG-8 cable acted like a suppressor; damaging EMP energy arced from the center conductor to the cable shield when the voltage level approached 5.5 kV.

Low price and a low clamping-voltage rating must be considered in the selection of an RF transient-protection device. The lower cost devices have the higher clamping voltages, however, and the higher-cost devices have the lower clamping voltages. Because of this, medium-priced devices manufactured by Fischer Custom Communications were selected for testing. The Fischer Spikeguard Suppressors ($55 price class) for coaxial lines can be made to order to operate at a specific clamping voltage. The Fischer devices satisfactorily suppressed the damaging transient pulses, passed the transmitter RF output power without interfering with the signal, and operated effectively over a wide frequency range.

Polyphaser Corporation devices are also effective in providing the necessary transient protection. The devices available limited the transmitter RF output power to 100 W or less, however. These units cost approximately $83 each.

RF coaxial protectors should be mounted on the station ground bus bar. If the Fischer device is used, it should be attached to a grounded UHF receptacle that will serve as a hold-down bracket. This creates a conductive path between the outer shield of the protector and the bus bar. The Polyphaser device can be mounted directly to the bus bar with the bracket provided.

Table 1
RF Coaxial-Line Protectors

Manufacturer	*Device*	*Approximate Cost (US Dollars)*	*Measured High-Z Clamping Voltage (Volts)*
Fischer	FCC-250-300-UHF	55	393
Fischer	FCC-250-350-UHF	55	260
Fischer	FCC-250-150-UHF	55	220
Fischer	FCC-250-120-UHF	55	240
Fischer	FCC-450-120-UHF	55	120
Polyphaser	IS-NEMP	83	140
Polyphaser	IS-NEMP-1	83	150
Polyphaser	IS-NEMP-2	83	160

Note: The transmitter output power, frequency of operation, and transmission line SWR must be considered when selecting any of these devices.

Attach the transceiver or antenna matching network to the grounded protector with a short (6 foot or less) piece of coaxial cable. Although the cable provides a ground path to the bus bar from the radio equipment, it is not a satisfactory transient-protection ground path for the transceiver. Another ground should be installed between the transceiver case and the ground bus using solid #6 wire. The coaxial cable shield should be grounded to the antenna tower leg at the tower base. Each tower leg should have an earth ground connection and be connected to the single-point ground system as shown in Fig 16.

Antenna Rotators

Antenna rotators can be protected by plugging the control box into a protected ac power source and adding protection to the control lines to the antenna rotator. When the control lines are in a shielded cable, the shield must be grounded at both ends. MOVs of the proper size should be installed at both ends of the control cable. At the station end, terminate the control cable in a small metal box that is connected to the station ground bus. Attach MOVs from each conductor to ground inside the box. At the antenna end of the control cable, place the MOVs inside the rotator case or in a small metal box that is properly grounded.

For example, the Alliance HD73 antenna rotator uses a 6-conductor unshielded control cable with a maximum control voltage of 24.7 V dc. Select an MOV with a clamping voltage level 10% higher (27 V or more) so the MOV won't clamp the control signal to ground. If the control voltage is ac, be sure to convert the RMS voltage value to peak voltage when considering the clamping voltage level.

Mobile Power Supply Protection

The mobile amateur station environment exposes radio equipment to other transient hazards in addition to those of lightning and EMP. Currents as high as 300 A are switched when starting the engine, and this can produce voltage spikes of over 200 V on the vehicle's electrical system. Lightning and EMP are not likely to impact the vehicle's electrical system as much as they would that of a fixed installation because the automobile chassis is not normally grounded. This would not be the case if the vehicle is inadvertently grounded; for example, when the vehicle is parked against a grounded metal conductor. The mobile radio system has two advantages over a fixed installation: Lightning is almost never a problem, and the vehicle battery is a natural surge suppressor.

Mobile radio equipment should be installed in a way that takes advantage of the protection provided by the battery. See **Fig 18**. To do this, connect the positive power lead of the radio directly to the positive battery post, not to intermediate points in the electrical system such as the fuse box or the auxiliary contacts on the ignition switch. To prevent equipment damage or fire, in-line fuses should be installed in the positive lead where they are attached to the battery post.

An MOV should be installed between the two leads of the equipment power cord. A GE MOV (V36ZA80) is recommended for this application. This MOV provides the lowest measured clamping voltage (170 V) and is low in cost.

Mobile Antenna Installation

Although tests indicate that mobile radios can survive an EMP transient without protection for the antenna system, protection from lightning transients is still required. A coaxial-line transient suppressor should be installed on the vehicle chassis between the antenna and the radio's antenna connector.

A Fischer suppressor can be attached to a UHF receptacle that is mounted on, and grounded to, the vehicle chassis. The Polyphaser protector can be mounted on, and grounded to, the vehicle chassis with its flange. Use a short length of coaxial cable between the radio and the transient suppressor.

Clamping Voltage Calculation

When selecting any EMP-protection device to be used at the antenna port of a radio, several items must be considered. These include transmitter RF power output, the SWR, and the operating frequency. The protection device must allow the outgoing RF signal to pass without clamping. A clamping voltage calculation must be made for each amateur installation.

The RF-power input to a transmission line develops a corresponding voltage that becomes important when a voltage-surge arrester is in the line. SWR is important because of its influence on the voltage level. The maximum voltage developed for a given power input is determined by:

$$V = \sqrt{2 \times P \times Z \times SWR} \qquad \text{(Eq 1)}$$

where

P = peak power in W
Z = impedance of the coaxial cable (Ω)
V = peak voltage across the cable

Eq 1 should be used to determine the peak voltage present across the transmission line. Because the RF transient-protection devices use gas-discharge tubes, the voltage level at which they clamp is not fixed; a safety margin must be added to the calculated peak voltage. This is done

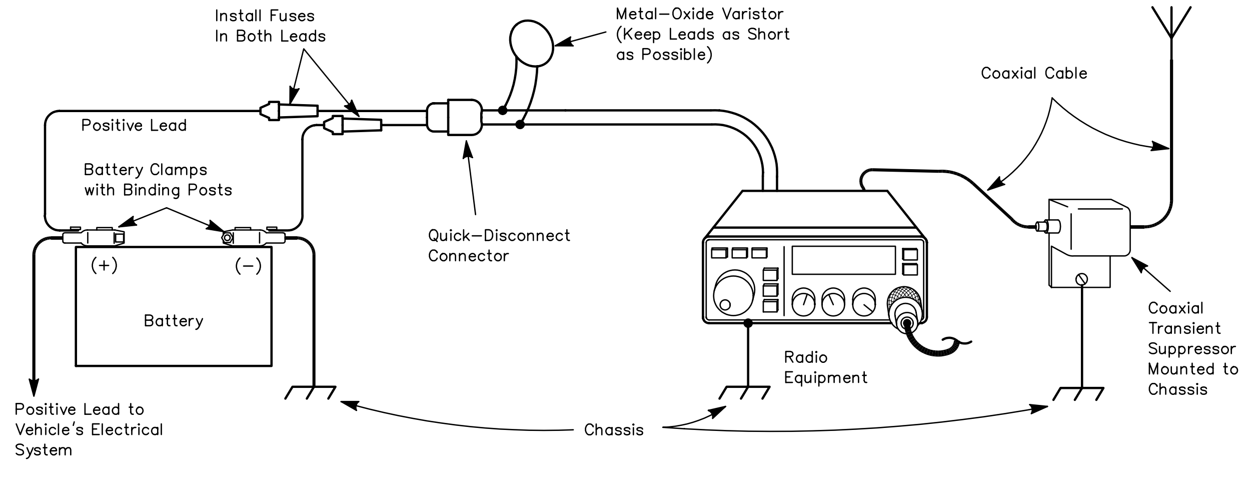

Fig 18—Recommended method of connecting mobile radio equipment to the vehicle battery and antenna.

by multiplying the calculated value by a factor of three. This added safety margin is required to ensure that the transmitter's RF output power will pass through the transient suppressor without causing the device to clamp the RF signal to ground. The final clamping voltage obtained is then high enough to allow normal operation of the transmitter while providing the lowest practical clamping voltage for the suppression device. This ensures the maximum possible protection for the radio system.

Here's how to determine the clamping voltage required. Let's assume the SWR is 1.5:1. The power output of the transceiver is 100 W PEP. RG/8 coaxial cable has an impedance of 52 Ω. Therefore:

P = 100 W
Z = 52 Ω
SWR = 1.5

Substituting these values into Eq 1:

$$V = \sqrt{2 \times 100 \times 52 \times 1.5} = 124.89 \text{ V peak}$$

Note that the voltage, V, is the peak value at the peak of the RF envelope. The final clamping voltage (FCV) is three times this value, or 374.7 V. Therefore, a coaxial-line transient suppressor that clamps at or above 375 V should be used.

The cost of a two-point basic protection scheme is estimated to be $100 for each fixed amateur station. This includes the cost of a good quality plug-in power-line protector ($45) and one Fischer coaxial-line protector ($55).

Inexpensive Transient-Protection Devices

Here are two low-cost protection devices you can assemble. They performed flawlessly in the tests.

The radio antenna connection can be protected by means of another simple device. As shown in **Fig 19**, two spark gaps (Siemens BI-A350) are installed in series at one end of a coaxial-cable T connector. Use the shortest practical lead length (about 1/4 inch) between the two spark gaps. One lead is bent forward and forced between the split sections of the inner coaxial connector until the spark gaps approach the body of the connector. A short length of insulating material (such as Mylar) is placed between the spark gaps and the connector shell. The other spark-gap lead is folded over the insulator, then conductive (metallic) tape is wrapped around the assembly. This construction method proved durable enough to allow many insertions and removals of the device during testing. Estimated cost of this assembly is $9. Similar devices can be built using components from Joslyn, General Electric, General Semiconductor or Siemens.

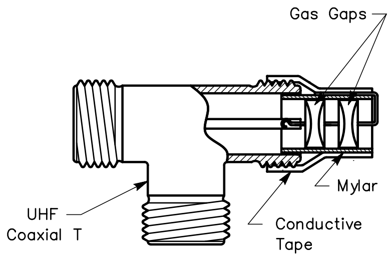

Fig 19—Pictorial diagram of an inexpensive, homemade RF coax transient protector.

Summary

Amateurs should be aware of which components in their radio system are most likely to be damaged by EMP. They should also know how to repair the damaged equipment. Amateurs should know how to reestablish communications after an EMP event, taking into consideration its adverse effects on the earth's atmosphere and radio equipment. One of the first things that would be noticed, providing the radio equipment is operative, is a sudden silence in radio transmissions across all frequencies below approximately 100 MHz. This silence would be caused in part by damage to unprotected radio gear from the EMP transient. Transmissions from one direction, the direction of the nuclear blast, would be completely out. RF signal loss by absorption and attenuation by the nuclear fireball are the reasons for this.

After an EMP event, the amateur should be prepared to operate CW. CW gives the most signal power under adverse conditions. It also provides a degree of message security from the general public.

Amateurs should develop the capability and flexibility to operate in more than one frequency band. The lower ground-wave frequencies should be useful for long-distance communications immediately after an EMP event. Line-of-sight VHF would be of value for local communications.

What can be done to increase the survivability of an Amateur Radio station? Here are some suggestions:

1) If you have spare equipment, keep it disconnected; use only the primary station gear. The spare equipment would then be available after an EMP event.
2) Keep equipment turned off and antenna and power lines disconnected when the equipment is not in use.
3) Connect only those external conductors necessary for the current mode of operation.
4) Tie all fixed equipment to a single-point earth ground to prevent closed loops through the ground.
5) Obtain schematic diagrams of your equipment and tools for repair of the equipment.
6) Have spare parts on hand for sensitive components of the radio equipment and antenna system.
7) Learn how to repair or replace the sensitive components of the radio equipment.
8) Use nonmetallic guy lines and antenna structural parts where possible.
9) Obtain an emergency power source and operate from

it during periods of increased world political tension. The power source should be completely isolated from the commercial power lines.

10) Equipment power cords should be disconnected when the gear is idle. Or the circuit breaker for the line feeding the equipment should be kept in the OFF position when the station is off the air.
11) Disconnect the antenna lead-in when the station is off the air. Or use a grounding antenna switch and keep it in the GROUND position when the equipment is not in use.
12) Have a spare antenna and transmission line on hand to replace a damaged antenna system.
13) Install EMP surge arresters and filters on all primary conductors attached to the equipment and antenna.
14) Retain tube type equipment and spare components; keep them in good working order.
15) Do not rely on a microprocessor to control the station after an EMP event. Be able to operate without microprocessor control.

The recommendations contained in this section were developed with low cost in mind; they are not intended to cover all possible combinations of equipment and installation methods found in the amateur community. Amateurs should examine their own requirements and use this report as a guideline in providing protection for the equipment.

RF Radiation and Electromagnetic Field Safety

Amateur Radio is basically a safe activity. In recent years, however, there has been considerable discussion and concern about the possible hazards of electromagnetic radiation (EMR), including both RF energy and power-frequency (50-60 Hz) electromagnetic (EM) fields. FCC regulations set limits on the maximum permissible exposure (MPE) allowed from the operation of radio transmitters. These regulations do not take the place of RF-safety practices, however. This section deals with the topic of RF safety.

This section was prepared by members of the ARRL RF Safety Committee and coordinated by Dr Robert E. Gold, WBØKIZ. It summarizes what is now known and offers safety precautions based on the research to date.

All life on Earth has adapted to survive in an environment of weak, natural, low-frequency electromagnetic fields (in addition to the Earth's static geomagnetic field). Natural low-frequency EM fields come from two main sources: the sun, and thunderstorm activity. But in the last 100 years, man-made fields at much higher intensities and with a very different spectral distribution have altered this natural EM background in ways that are not yet fully understood. Researchers continue to look at the effects of RF exposure over a wide range of frequencies and levels.

Both RF and 60-Hz fields are classified as *nonionizing radiation,* because the frequency is too low for there to be enough photon energy to ionize atoms. (*Ionizing radiation*, such as X-rays, gamma rays and even some ultraviolet radiation has enough energy to knock electrons loose from their atoms. When this happens, positive and negative ions are formed.) Still, at sufficiently high power densities, EMR poses certain health hazards. It has been known since the early days of radio that RF energy can cause injuries by heating body tissue. (Anyone who has ever touched an improperly grounded radio chassis or energized antenna and received an *RF burn* will agree that this type of injury can be quite painful.) In extreme cases, RF-induced heating in the eye can result in cataract formation, and can even cause blindness. Excessive RF heating of the reproductive organs can cause sterility. Other health problems also can result from RF heating. These heat-related health hazards are called *thermal effects.* A microwave oven is a positive application of this thermal effect.

There also have been observations of changes in physiological function in the presence of RF energy levels that are too low to cause heating. These functions return to normal when the field is removed. Although research is ongoing, no harmful health consequences have been linked to these changes.

In addition to the ongoing research, much else has been done to address this issue. For example, FCC regulations set limits on exposure from radio transmitters. The Institute of Electrical and Electronics Engineers, the American National Standards Institute and the National Council for Radiation Protection and Measurement, among others, have recommended voluntary guidelines to limit human exposure to RF energy. The ARRL has established the RF Safety Committee, consisting of concerned medical doctors and scientists, serving voluntarily to monitor scientific research in the fields and to recommend safe practices for radio amateurs.

THERMAL EFFECTS OF RF ENERGY

Body tissues that are subjected to *very high* levels of RF energy may suffer serious heat damage. These effects depend on the frequency of the energy, the power density of the RF field that strikes the body and factors such as the polarization of the wave.

At frequencies near the body's natural resonant frequency, RF energy is absorbed more efficiently, and an increase in heating occurs. In adults, this frequency usually is about 35 MHz if the person is grounded, and about 70 MHz if insulated from the ground. Individual body parts may be resonant at different frequencies. The adult head, for example, is resonant around 400 MHz, while a baby's smaller head resonates near 700 MHz. Body size thus determines the frequency at which most RF energy is absorbed. As the frequency is moved farther from resonance, less RF heating generally occurs. *Specific absorption rate (SAR)* is a term that describes the rate at which RF energy is absorbed in tissue.

Maximum permissible exposure (MPE) limits are based on whole-body SAR values, with additional safety factors included as part of the standards and regulations. This helps explain why these safe exposure limits vary with frequency. The MPE limits define the maximum electric and magnetic field strengths or the plane-wave equivalent power densities associated with these fields, that a person may be exposed to without harmful effect—and with an acceptable safety factor. The regulations assume that a person exposed to a specified (safe) MPE level also will experience a safe SAR.

Nevertheless, thermal effects of RF energy should not be a major concern for most radio amateurs, because of the power levels we normally use and the intermittent nature of most amateur transmissions. Amateurs spend more time listening than transmitting, and many amateur transmissions such as CW and SSB use low-duty-cycle modes. (With FM or RTTY, though, the RF is present continuously at its maximum level during each transmission.) In any event, it is rare for radio amateurs to be subjected to RF fields strong enough to produce thermal effects, unless they are close to an energized antenna or un- shielded power amplifier. Specific suggestions for avoiding excessive exposure are offered later in this chapter.

ATHERMAL EFFECTS OF EMR

Research about possible health effects resulting from exposure to the lower level energy fields, the athermal effects, has been of two basic types: epidemiological research and laboratory research.

Scientists conduct laboratory research into biological mechanisms by which EMR may affect animals including humans. Epidemiologists look at the health patterns of large groups of people using statistical methods. These epidemiological studies have been inconclusive. By their basic design, these studies do not demonstrate cause and effect, nor do they postulate mechanisms of disease. Instead, epidemiologists look for associations between an environmental factor and an observed pattern of illness. For example, in the earliest research on malaria, epidemiologists observed the association between populations with high prevalence of the disease and the proximity of mosquito infested swamplands. It was left to the biological and medical scientists to isolate the organism causing malaria in the blood of those with the disease, and identify the same organisms in the mosquito population.

In the case of athermal effects, some studies have identified a weak association between exposure to EMF at home or at work and various malignant conditions including leukemia and brain cancer. A larger number of equally well designed and performed studies, however, have found no association. A risk ratio of between 1.5 and 2.0 has been observed in positive studies (the number of observed cases of malignancy being 1.5 to 2.0 times the "expected" number in the population). Epidemiologists generally regard a risk ratio of 4.0 or greater to be indicative of a strong association between the cause and effect under study. For example, men who smoke one pack of cigarettes per day increase their risk for lung cancer tenfold compared to nonsmokers, and two packs per day increases the risk to more than 25 times the nonsmokers' risk.

Epidemiological research by itself is rarely conclusive, however. Epidemiology only identifies health patterns in groups—it does not ordinarily determine their cause. And there are often confounding factors: Most of us are exposed to many different environmental hazards that may affect our health in various ways. Moreover, not all studies of persons likely to be exposed to high levels of EMR have yielded the same results.

There also has been considerable laboratory research about the biological effects of EMR in recent years. For example, some separate studies have indicated that even fairly low levels of EMR might alter the human body's circadian rhythms, affect the manner in which T lymphocytes function in the immune system and alter the nature of the electrical and chemical signals communicated through the cell membrane and between cells, among other things. Although these studies are intriguing, they do not demonstrate any effect of these low-level fields on the overall organism.

Much of this research has focused on low-frequency magnetic fields, or on RF fields that are keyed, pulsed or modulated at a low audio frequency (often below 100 Hz). Several studies suggested that humans and animals can adapt to the presence of a steady RF carrier more readily than to an intermittent, keyed or modulated energy source.

The results of studies in this area, plus speculations concerning the effect of various types of modulation, were and have remained somewhat controversial. None of the research to date has demonstrated that low-level EMR causes adverse health effects.

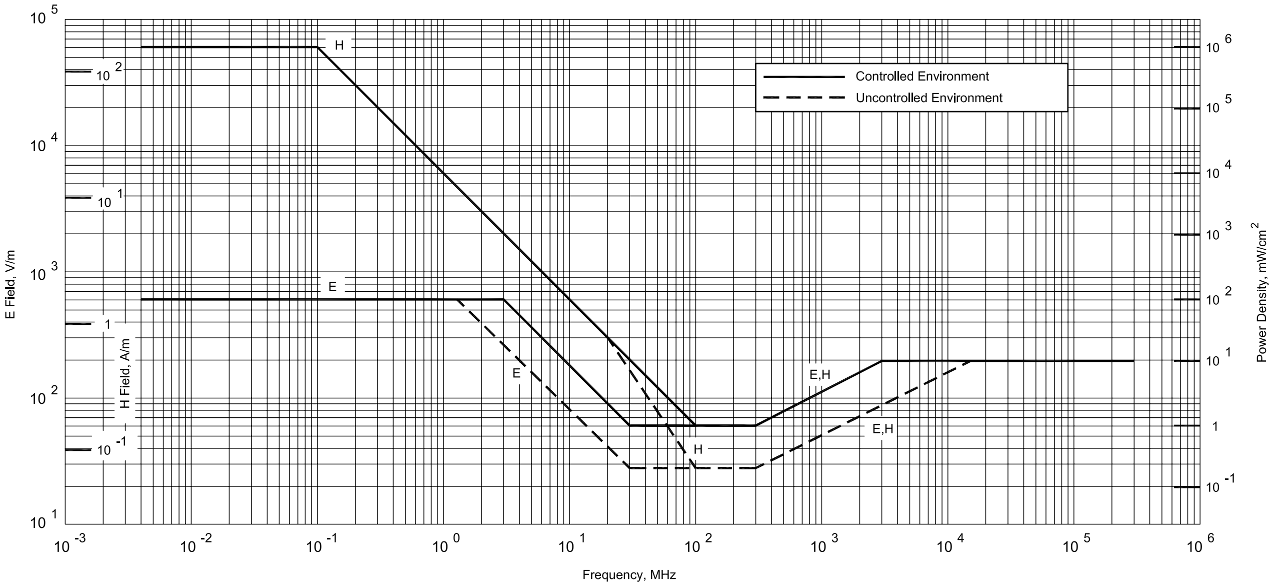

Fig 20—1991 RF protection guidelines for body exposure of humans. It is known officially as the "IEEE Standard for Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz."

Given the fact that there is a great deal of ongoing research to examine the health consequences of exposure to EMF, the American Physical Society (a national group of highly respected scientists) issued a statement in May 1995 based on its review of available data pertaining to the possible connections of cancer to 60-Hz EMF exposure. This report is exhaustive and should be reviewed by anyone with a serious interest in the field. Among its general conclusions were the following:

1. The scientific literature and the reports of reviews by other panels show no consistent, significant link between cancer and power line fields.
2. No plausible biophysical mechanisms for the systematic initiation or promotion of cancer by these extremely weak 60-Hz fields has been identified.
3. While it is impossible to prove that no deleterious health effects occur from exposure to any environmental factor, it is necessary to demonstrate a consistent, significant, and causal relationship before one can conclude that such effects do occur.

In a report dated October 31, 1996, a committee of the National Research Council of the National Academy of Sciences has concluded that no clear, convincing evidence exists to show that residential exposures to electric and magnetic fields (EMFs) are a threat to human health.

A National Cancer Institute epidemiological study of residential exposure to magnetic fields and acute lymphoblastic leukemia in children was published in the *New England Journal of Medicine* in July 1997. The exhaustive, seven-year study concludes that if there is any link at all, it is far too weak to be concerned about.

Readers may want to follow this topic as further studies are reported. Amateurs should be aware that exposure to RF and ELF (60 Hz) electromagnetic fields at all power levels and frequencies has not been fully studied under all circumstances. "Prudent avoidance" of any avoidable EMR is always a good idea. Prudent avoidance doesn't mean that amateurs should be fearful of using their equipment. Most amateur operations are well within the MPE limits. If any risk does exist, it will almost surely fall well down on the list of causes that may be harmful to your health (on the other end of the list from your automobile). It does mean, however, that hams should be aware of the potential for exposure from their stations, and take whatever reasonable steps they can take to minimize their own exposure and the exposure of those around them.

Safe Exposure Levels

How much EM energy is safe? Scientists and regulators have devoted a great deal of effort to deciding upon safe RF-exposure limits. This is a very complex problem, involving difficult public health and economic considerations. The recommended safe levels have been revised downward several times over the years —and not all scientific bodies agree on this question even today. An Institute of Electrical and Electronics Engineers (IEEE) standard for recommended EM exposure limits was published in 1991 (see Bibliography). It replaced a 1982 American National Standards Institute (ANSI) standard. In the new standard, most of the permitted exposure levels were revised downward (made more stringent), to better reflect the current research. The new IEEE stan-

FCC RF-Exposure Regulations

FCC regulations control the amount of RF exposure that can result from your station's operation (§§97.13, 97.503, 1.1307 (b)(c)(d), 1.1310 and 2.1093). The regulations set limits on the maximum permissible exposure (MPE) allowed from operation of transmitters in all radio services. They also require that certain types of stations be evaluated to determine if they are in compliance with the MPEs specified in the rules. The FCC has also required that five questions on RF environmental safety practices be added to Novice, Technician and General license examinations.

These rules went into effect on January 1, 1998 for new stations or stations that file a Form 605 application with the FCC. Other existing stations have until September 1, 2000 to be in compliance with the rules.

The Rules

Maximum Permissible Exposure (MPE)

All radio stations regulated by the FCC must comply with the requirements for MPEs, even QRP stations running only a few watts or less. The MPEs vary with frequency, as shown in **Table A**. MPE limits are specified in maximum electric and magnetic fields for frequencies below 30 MHz, in power density for frequencies above 300 MHz and all three ways for frequencies from 30 to 300 MHz. For compliance purposes, all of these limits must be considered *separately*. If any one is exceeded, the station is not in compliance.

The regulations control human exposure to RF fields, not the strength of RF fields. There is no limit to how strong a field can be as long as no one is being exposed to it, although FCC regulations require that amateurs use the minimum necessary power at all times (§97.311 [a]).

Environments

The FCC has defined two exposure environments — *controlled* and *uncontrolled*. A controlled environment is one in which the people who are being exposed are aware of that exposure and can take steps to minimize that exposure, if appropriate. In an uncontrolled environment, the people being exposed are not normally aware of the exposure. The uncontrolled environment limits are more stringent than the controlled environment limits.

Although the controlled environment is usually intended as an occupational environment, the FCC has determined that it generally applies to amateur operators and members of their immediate households. In most cases, controlled-environment limits can be applied to your home and property to which you can control physical access. The uncontrolled environment is intended for areas that are accessible by the general public, such as your neighbors' properties.

The MPE levels are based on average exposure. An averaging time of 6 minutes is used for controlled exposure; an averaging period of 30 minutes is used for uncontrolled exposure.

Table A—(From §1.1310) Limits for Maximum Permissible Exposure (MPE)

(A) Limits for Occupational/Controlled Exposure

Frequency Range (MHz)	*Electric Field Strength (V/m)*	*Magnetic Field Strength (A/m)*	*Power Density (mW/cm^2)*	*Averaging Time (minutes)*
0.3-3.0	614	1.63	(100)*	6
3.0-30	1842/f	4.89/f	(900/f^2)*	6
30-300	61.4	0.163	1.0	6
300-1500	—	—	f/300	6
1500-100,000	—	—	5	6

f = frequency in MHz

* = Plane-wave equivalent power density (see Note 1).

(B) Limits for General Population/Uncontrolled Exposure

Frequency Range (MHz)	*Electric Field Strength (V/m)*	*Magnetic Field Strength (A/m)*	*Power Density (mW/cm^2)*	*Averaging Time (minutes)*
0.3-1.34	614	1.63	(100)*	30
1.34-30	824/f	2.19/f	(180/f^2)*	30
30-300	27.5	0.073	0.2	30
300-1500	—	—	f/1500	30
1500-100,000	—	—	1.0	30

f = frequency in MHz

* = Plane-wave equivalent power density (see Note 1).

Note 1: This means the equivalent far-field strength that would have the E or H-field component calculated or measured. It does not apply well in the near field of an antenna. The equivalent far-field power density can be found in the near or far field regions from the relationships: $P_d = |E_{total}|^2 / 3770$ mW/cm^2 or from $P_d = |H_{total}|^2 \times 37.7$ mW/cm^2.

Station Evaluations

The FCC requires that certain amateur stations be evaluated for compliance with the MPEs. Although an amateur can have someone else do the evaluation, it is not difficult for hams to evaluate their own stations. The ARRL book *RF Exposure and You* contains extensive information about the regulations and a large chapter of tables that show compliance distances for specific antennas and power levels. Generally, hams will use these tables to evaluate their stations. Some of these tables have been included in the FCC's information — *OET Bulletin 65* and its *Supplement B*. If hams choose, however, they can do more extensive calculations, use a computer to model their antenna and exposure, or make actual measurements.

Table B—Power Thresholds for Routine Evaluation of Amateur Radio Stations

Wavelength Band	***Evaluation Required if Power* (watts) Exceeds:***
MF	
160 m	500
HF	
80 m	500
75 m	500
40 m	500
30 m	425
20 m	225
17 m	125
15 m	100
12 m	75
10 m	50
VHF (all bands)	50
UHF	
70 cm	70
33 cm	150
23 cm	200
13 cm	250
SHF (all bands)	250
EHF (all bands)	250
Repeater stations (all bands)	*non-building-mounted antennas*: height above ground level to lowest point of antenna < 10 m *and* power > 500 W ERP *building-mounted antennas*: power > 500 W ERP

*Transmitter power = Peak-envelope power input to antenna. For repeater stations ***only,*** power exclusion based on ERP (effective radiated power).

Categorical Exemptions

Some types of amateur stations do not need to be evaluated, but these stations must still comply with the MPE limits. The station licensee remains responsible for ensuring that the station meets these requirements.

The FCC has exempted these stations from the evaluation requirement because their output power, operating mode and frequency are such that they are presumed to be in compliance with the rules.

Stations using power equal to or less than the levels in **Table B** do not have to be evaluated. For the 100-W HF ham station, for example, an evaluation would be required *only* on 12 and 10 meters.

Hand-held radios and vehicle-mounted mobile radios that operate using a push-to-talk (PTT) button are also categorically exempt from performing the routine evaluation. Repeater stations that use less than 500 W ERP or those with antennas not mounted on buildings, if the antenna is at least 10 meters off the ground, also do not need to be evaluated.

Correcting Problems

Most hams are already in compliance with the MPE requirements. Some amateurs, especially those using indoor antennas or high-power, high-duty-cycle modes such as a RTTY bulletin station and specialized stations for moonbounce operations and the like may need to make adjustments to their station or operation to be in compliance.

The FCC permits amateurs considerable flexibility in complying with these regulations. As an example, hams can adjust their operating frequency, mode or power to comply with the MPE limits. They can also adjust their operating habits or control the direction their antenna is pointing.

More Information

This discussion offers only an overview of this topic; additional information can be found in *RF Exposure and You* and on *ARRLWeb* at **www.arrl.org/news/rfsafety/**. *ARRLWeb* has links to the FCC Web site, with *OET Bulletin 65* and *Supplement B* and links to software that hams can use to evaluate their stations.

dard was adopted by ANSI in 1992.

The IEEE standard recommends frequency-dependent and time-dependent maximum permissible exposure levels. Unlike earlier versions of the standard, the 1991 standard recommends different RF exposure limits in controlled environments (that is, where energy levels can be accurately determined and everyone on the premises is aware of the presence of EM fields) and in uncontrolled environments (where energy levels are not known or where people may not be aware of the presence of EM fields). FCC regulations also include controlled/occupational and uncontrolled/general population exposure environments.

The graph in **Fig 20** depicts the 1991 IEEE standard. It is necessarily a complex graph, because the standards differ not only for controlled and uncontrolled environments but also for electric (E) fields and magnetic (H) fields. Basically, the lowest E-field exposure limits occur at frequencies between 30 and 300 MHz. The lowest H-field exposure levels occur at 100-300 MHz. The ANSI standard sets the maximum E-field limits between 30 and 300 MHz at a power density of 1 mW/cm^2 (61.4 V/m) in controlled environments—but at one-fifth that level (0.2 mW/cm^2 or 27.5 V/m) in uncontrolled environments. The H-field limit drops to 1 mW/cm^2 (0.163 A/m) at 100-300 MHz in controlled environments and 0.2 mW/cm^2 (0.0728 A/m) in uncontrolled environments. Higher power densities are permitted at frequencies below 30 MHz (below 100 MHz for H fields) and above 300 MHz, based on the concept that the body will not be resonant at those frequencies and will therefore absorb less energy.

In general, the 1991 IEEE standard requires averaging the power level over time periods ranging from 6 to

30 minutes for power-density calculations, depending on the frequency and other variables. The ANSI exposure limits for uncontrolled environments are lower than those for controlled environments, but to compensate for that the standard allows exposure levels in those environments to be averaged over much longer time periods (generally 30 minutes). This long averaging time means that an intermittently operating RF source (such as an Amateur Radio transmitter) will show a much lower power density than a continuous-duty station—for a given power level and antenna configuration.

Time averaging is based on the concept that the human body can withstand a greater rate of body heating (and thus, a higher level of RF energy) for a short time than for a longer period. Time averaging may not be appropriate, however, when considering nonthermal effects of RF energy.

The IEEE standard excludes any transmitter with an output below 7 W because such low-power transmitters would not be able to produce significant whole-body heating. (Recent studies show that hand-held transceivers often produce power densities in excess of the IEEE standard within the head.)

There is disagreement within the scientific community about these RF exposure guidelines. The IEEE standard is still intended primarily to deal with thermal effects, not exposure to energy at lower levels. A small but significant number of researchers now believe athermal effects also should be taken into consideration. Several European countries and localities in the United States have adopted stricter standards than the recently updated IEEE standard.

Another national body in the United States, the National Council for Radiation Protection and Measurement (NCRP), also has adopted recommended exposure guidelines. NCRP urges a limit of 0.2 mW/cm^2 for non-occupational exposure in the 30-300 MHz range. The NCRP guideline differs from IEEE in two notable ways: It takes into account the effects of modulation on an RF carrier, and it does not exempt transmitters with outputs below 7 W.

The FCC MPE regulations are based on parts of the 1992 IEEE/ANSI standard and recommendations of the National Council for Radiation Protection and Measurement (NCRP). The MPE limits under the regulations are slightly different than the IEEE/ANSI limits. Note that the MPE levels apply to the FCC rules put into effect for radio amateurs on January 1, 1998. These MPE requirements do not reflect and include all the assumptions and exclusions of the IEEE/ANSI standard.

Cardiac Pacemakers and RF Safety

It is a widely held belief that cardiac pacemakers may be adversely affected in their function by exposure to electromagnetic fields. Amateurs with pacemakers may ask whether their operating might endanger themselves or visitors to their shacks who have a pacemaker. Because of this, and similar concerns regarding other sources of electromagnetic fields, pacemaker manufacturers apply design methods that for the most part shield the pacemaker circuitry from even relatively high EM field strengths.

It is recommended that any amateur who has a pacemaker, or is being considered for one, discuss this matter with his or her physician. The physician will probably put the amateur into contact with the technical representative of the pacemaker manufacturer. These representatives are generally excellent resources, and may have data from laboratory or “in the field” studies with specific model pacemakers.

One study examined the function of a modern (dual chamber) pacemaker in and around an Amateur Radio station. The pacemaker generator has circuits that receive and process electrical signals produced by the heart, and also generate electrical signals that stimulate (pace) the heart. In one series of experiments, the pacemaker was connected to a heart simulator. The system was placed on top of the cabinet of a 1-kW HF linear amplifier during SSB and CW operation. In another test, the system was placed in close proximity to several 1 to 5-W 2-meter hand-held transceivers. The test pacemaker was connected to the heart simulator in a third test, and then placed on the ground 9 meters below and 5 meters in front of a three-element Yagi HF antenna. No interference with pacemaker function was observed in these experiments.

Although the possibility of interference cannot be entirely ruled out by these few observations, these tests represent more severe exposure to EM fields than would ordinarily be encountered by an amateur—with an average amount of common sense. Of course, prudence dictates that amateurs with pacemakers, who use hand-held VHF transceivers, keep the antenna as far as possible from the site of the implanted pacemaker generator. They also should use the lowest transmitter output required for adequate communication. For high power HF transmission, the antenna should be as far as possible from the operating position, and all equipment should be properly grounded.

Low-Frequency Fields

Although the FCC doesn’t regulate 60-Hz fields, some recent concern about EMR has focused on low-frequency energy rather than RF. Amateur Radio equipment can be a significant source of low-frequency magnetic fields, although there are many other sources of this kind of energy in the typical home. Magnetic fields can be measured relatively accurately with inexpensive 60-Hz meters that are made by several manufacturers.

Table 2 shows typical magnetic field intensities of Amateur Radio equipment and various household items. Because these fields dissipate rapidly with distance, “prudent avoidance” would mean staying perhaps 12 to 18 inches away from most Amateur Radio equipment (and

24 inches from power supplies with 1-kW RF amplifiers).

Determining RF Power Density

Unfortunately, determining the power density of the RF fields generated by an amateur station is not as simple as measuring low-frequency magnetic fields. Although sophisticated instruments can be used to measure RF power densities quite accurately, they are costly and require frequent recalibration. Most amateurs don't have access to such equipment, and the inexpensive field-

Table 2

Typical 60-Hz Magnetic Fields Near Amateur Radio Equipment and AC-Powered Household Appliances

Values are in milligauss.

Item	*Field*	*Distance*
Electric blanket	30-90	Surface
Microwave oven	10-100 1-10	Surface 12″
IBM personal computer	5-10 0-1	Atop monitor 15″ from screen
Electric drill	500-2000	At handle
Hair dryer	200-2000	At handle
HF transceiver	10-100 1-5	Atop cabinet 15″ from front
1-kW RF amplifier	80-1000 1-25	Atop cabinet 15″ from front

(Source: measurements made by members of the ARRL RF Safety Committee)

Table 3

Typical RF Field Strengths Near Amateur Radio Antennas

A sampling of values as measured by the Federal Communications Commission and Environmental Protection Agency, 1990

Antenna Type	*Freq (MHz)*	*Power (W)*	*E Field (V/m)*	*Location*
Dipole in attic	14.15	100	7-100	In home
Discone in attic	146.5	250	10-27	In home
Half sloper	21.5	1000	50	1 m from base
Dipole at 7-13 ft	7.14	120	8-150	1-2 m from earth
Vertical	3.8	800	180	0.5 m from base
5-element Yagi at 60 ft	21.2	1000	10-20 14	In shack 12 m from base
3-element Yagi at 25 ft	28.5	425	8-12	12 m from base
Inverted V at 22-46 ft	7.23	1400	5-27	Below antenna
Vertical on roof	14.11	140	6-9 35-100	In house At antenna tuner
Whip on auto roof	146.5	100	22-75 15-30 90	2 m antenna In vehicle Rear seat
5-element Yagi at 20 ft	50.1	500	37-50	10 m antenna

Table 4

RF Awareness Guidelines

These guidelines were developed by the ARRL RF Safety Committee, based on the FCC/EPA measurements of Table 3 and other data.

- Although antennas on towers (well away from people) pose no exposure problem, make certain that the RF radiation is confined to the antennas' radiating elements themselves. Provide a single, good station ground (earth), and eliminate radiation from transmission lines. Use good coaxial cable or other feed line properly. Avoid serious imbalance in your antenna system and feed line. For high-powered installations, avoid end-fed antennas that come directly into the transmitter area near the operator.
- No person should ever be near any transmitting antenna while it is in use. This is especially true for mobile or ground-mounted vertical antennas. Avoid transmitting with more than 25 W in a VHF mobile installation unless it is possible to first measure the RF fields inside the vehicle. At the 1-kW level, both HF and VHF directional antennas should be at least 35 ft above inhabited areas. Avoid using indoor and attic-mounted antennas if at all possible. If open-wire feeders are used, ensure that it is not possible for people (or animals) to come into accidental contact with the feed line.
- Don't operate high-power amplifiers with the covers removed, especially at VHF/UHF.
- In the UHF/SHF region, never look into the open end of an activated length of waveguide or microwave feed-horn antenna or point it toward anyone. (If you do, you may be exposing your eyes to more than the maximum permissible exposure level of RF radiation.) Never point a high-gain, narrow-bandwidth antenna (a paraboloid, for instance) toward people. Use caution in aiming an EME (moonbounce) array toward the horizon; EME arrays may deliver an effective radiated power of 250,000 W or more.
- With hand-held transceivers, keep the antenna away from your head and use the lowest power possible to maintain communications. Use a separate microphone and hold the rig as far away from you as possible. This will reduce your exposure to the RF energy.
- Don't work on antennas that have RF power applied.
- Don't stand or sit close to a power supply or linear amplifier when the ac power is turned on. Stay at least 24 inches away from power transformers, electrical fans and other sources of high-level 60-Hz magnetic fields.

strength meters that we do have are not suitable for measuring RF power density.

Table 3 shows a sampling of measurements made at Amateur Radio stations by the Federal Communications Commission and the Environmental Protection Agency in 1990. As this table indicates, a good antenna well removed from inhabited areas poses no hazard under any of the IEEE/ANSI guidelines. However, the FCC/EPA survey also indicates that amateurs must be careful about using indoor or attic-mounted antennas, mobile antennas, low directional arrays or any other antenna that is close to inhabited areas, especially when moderate to high power is used.

Ideally, before using any antenna that is in close proximity to an inhabited area, you should measure the RF power density. If that is not feasible, the next best option is make the installation as safe as possible by observing the safety suggestions listed in **Table 4**.

It also is possible, of course, to calculate the probable power density near an antenna using simple equations. Such calculations have many pitfalls. For one, most of the situations where the power density would be high enough to be of concern are in the near field. In the near field, ground interactions and other variables produce power densities that cannot be determined by simple arithmetic. In the far field, conditions become easier to predict with simple calculations.

The boundary between the near field and the far field depends on the wavelength of the transmitted signal and the physical size and configuration of the antenna. The boundary between the near field and the far field of an antenna can be as much as several wavelengths from the antenna.

Computer antenna-modeling programs are another approach you can use. *MININEC* or other codes derived from *NEC* (Numerical Electromagnetics Code) are suitable for estimating RF magnetic and electric fields around amateur antenna systems.

These models have limitations. Ground interactions must be considered in estimating near-field power densities, and the "correct ground" must be modeled. Computer modeling is generally not sophisticated enough to predict "hot spots" in the near field—places where the field intensity may be far higher than would be expected, due to reflections from nearby objects. In addition, "nearby objects" often change or vary with weather or the season, so the model so laboriously crafted may not be representative of the actual situation, by the time it is running on the computer.

Intensely elevated but localized fields often can be detected by professional measuring instruments. These "hot spots" are often found near wiring in the shack, and metal objects such as antenna masts or equipment cabinets. But even with the best instrumentation, these measurements also may be misleading in the near field.

One need not make precise measurements or model the exact antenna system, however, to develop some idea of the relative fields around an antenna. Computer modeling using close approximations of the geometry and power input of the antenna will generally suffice. Those who are familiar with *MININEC* can estimate their power densities by computer modeling, and those who have access to professional power-density meters can make useful measurements.

While our primary concern is ordinarily the intensity of the signal radiated by an antenna, we also should remember that there are other potential energy sources to be considered. You also can be exposed to RF radiation directly from a power amplifier if it is operated without proper shielding. Transmission lines also may radiate a significant amount of energy under some conditions. Poor microwave waveguide joints or improperly assembled connectors are another source of incidental radiation.

Further RF Exposure Suggestions

Potential exposure situations should be taken seriously. Based on the FCC/EPA measurements and other data, the "RF awareness" guidelines of Table 4 were developed by the ARRL RF Safety Committee. A longer version of these guidelines, along with a complete list of references, appeared in a *QST* article by Ivan Shulman, MD, WC2S ("Is Amateur Radio Hazardous to Our Health?" *QST* , Oct 1989, pp 31-34). For more information or background, see the list of RF Safety References in the next section.

In addition, the ARRL has published a book, *RF Exposure and You*, that is helping hams comply with the FCC's RF-exposure regulations. The ARRL also maintains an RF-exposure news page on its Web site. See **www.arrl.org/news/rfsafety**. This site contains reprints of selected *QST* articles on RF exposure and links to the FCC and other useful sites.

Chapter 2

Antenna Fundamentals

Antennas belong to a class of devices called *transducers*. This term is derived from two Latin words, meaning literally "to lead across" or "to transfer." Thus, a transducer is a device that transfers, or converts, energy from one form to another. The purpose of an antenna is to convert radio-frequency electric current to *electromagnetic waves*, which are then *radiated* into space. [For more details on the properties of electromagnetic waves themselves, see Chapter 23, Radio Wave Propagation.]

We cannot directly see or hear, taste or touch electromagnetic waves, so it's not surprising that the process by which they are launched into space from our antennas can be a little mystifying, especially to a newcomer. In everyday life we come across many types of transducers, although we don't always recognize them as such. A comparison with a type of transducer that you can actually see and touch may be useful. You are no doubt familiar with a *loudspeaker*. It converts audio-frequency electric current from the output of your radio or stereo into acoustic pressure waves, also known as *sound waves*. The sound waves are propagated through the air to your ears, where they are converted into what you perceive as sound.

We normally think of a loudspeaker as something that converts electrical energy into sound energy, but we could just as well turn things around and apply sound energy to a loudspeaker, which will then convert it into electrical energy. When used in this manner, the loudspeaker has become a microphone. The loudspeaker/microphone thus exhibits the principle of *reciprocity*, derived from the Latin word meaning to move back and forth.

Now, let's look more closely at that special transducer we call an *antenna*. When fed by a transmitter with RF current (usually through a transmission line), the antenna launches electromagnetic waves, which are propagated through space. This is similar to the way sound waves are propagated through the air by a loudspeaker. In the next town, or perhaps on a distant continent, a similar transducer (that is, a receiving antenna) intercepts some of these electromagnetic waves and converts them into electrical current for a receiver to amplify and detect.

In the same fashion that a loudspeaker can act as a microphone, a radio antenna also follows the principle of reciprocity. In other words, an antenna can transmit as well as receive signals. However, unlike the loudspeaker, an antenna does not require a *medium*, such as air, through which it radiates electromagnetic waves. Electromagnetic waves can be propagated through air, the vacuum of outer space or the near vacuum of the upper ionosphere. This is the miracle of radio—electromagnetic waves can propagate without a physical medium.

Essential Characteristics of Antennas

What other things make an antenna different from an ordinary electronic circuit? In ordinary circuits, the dimensions of coils, capacitors and connections usually are small compared with the wavelength of the frequency in use. Here, we define wavelength as the distance in free space traveled during one complete cycle of a wave. The velocity of a wave in free space is the speed of light, and the wavelength is thus:

$$\lambda_{\text{meters}} = \frac{299.7925 \times 10^6 \text{ meters/sec}}{\text{f hertz}} = \frac{299.7925}{\text{f MHz}} \qquad \text{(Eq 1)}$$

where λ_{meters}, the Greek letter lambda, is the free-space wavelength in meters.

Expressed in feet, Eq 1 becomes:

$$\lambda_{feet} = \frac{983.5712}{f\ MHz} \approx \frac{983.6}{f\ MHz} \qquad \text{(Eq 2)}$$

When circuit dimensions are small compared to λ, most of the electromagnetic energy is confined to the circuit itself, and is used up either performing useful work or is converted into heat. However, when the dimensions of wiring or components become significant compared with the wavelength, some of the energy escapes by radiation in the form of electromagnetic waves.

Antennas come in an enormous, even bewildering, assortment of shapes and sizes. This chapter on fundamentals will deal with the theory of simple forms of antennas, usually in *free space*, away from the influence of ground. Subsequent chapters will concentrate on more exotic or specialized antenna types. Chapter 3 deals with the complicated subject of the effect of ground, including the effect of uneven local terrain. Ground has a profound influence on how an antenna performs in the real world.

No matter what form an antenna takes, simple or complex, its electrical performance can be characterized according to the following important properties:

1. Feed-point Impedance
2. Directivity, Gain and Efficiency
3. Polarization

FEED-POINT IMPEDANCE

The first major characteristic defining an antenna is its *feed-point impedance*. Since we amateurs are free to choose our operating frequencies within assigned bands, we need to consider how the feed-point impedance of a particular antenna varies with frequency, within a particular band, or even in several different bands if we intend to use one antenna on multiple bands.

There are two forms of impedance associated with any antenna: *self impedance* and *mutual impedance*. As you might expect, self impedance is what you measure at the feed-point terminals of an antenna located completely away from the influence of any other conductors.

Mutual, or coupled, impedance is due to the parasitic effect of nearby conductors; that is, conductors located within the antenna's reactive near field. (The subject of fields around an antenna will be discussed in detail later.) This includes the effect of ground, which is a lossy conductor, but a conductor nonetheless. Mutual impedance is defined using Ohm's Law, just like self impedance. However, mutual impedance is the ratio of voltage in one conductor, divided by the current in another (coupled) conductor. Mutually coupled conductors can distort the pattern of a highly directive antenna, as well as change the impedance seen at the feed point.

In this chapter on fundamentals, we won't directly deal with mutual impedance, considering it as a side effect of nearby conductors. Instead, here we'll concentrate on simple antennas in free space, away from ground and any other conductors. Mutual impedance will be considered in detail in Chapter 11, HF Yagi Arrays, where it is essential for proper operation of these beam antennas.

Self Impedance

The current that flows into an antenna's feed point must be supplied at a finite voltage. The self impedance of the antenna is simply equal to the voltage applied to its feed point divided by the current flowing into the feed point. Where the current and voltage are exactly in phase, the impedance is purely resistive, with zero reactive component. For this case the antenna is termed *resonant*. (Amateurs often use the term "resonant" rather loosely, usually meaning "nearly resonant" or "close-to resonant.")

Please recognize that an antenna *need not be resonant* in order to be an effective radiator. There is in fact nothing magic about having a resonant antenna, provided of course that you can devise some efficient means to feed the antenna. Many amateurs use non-resonant (even random-length) antennas fed with open-wire transmission lines and antenna tuners. They radiate signals just as well as those using coaxial cable and resonant antennas, and as a bonus they usually can use these antenna systems on multiple frequency bands. It is important to consider an antenna and its feed line as a *system*, in which all losses should be kept to a minimum. See Chapter 24, Transmission Lines, for details on transmission-line loss as a function of impedance mismatch.

Except at the one frequency where it is truly resonant, the current in an antenna is at a different phase compared to the applied voltage. In other words, the antenna exhibits a feed-point *impedance*, not just a pure resistance. The feed-point impedance is composed of either capacitive or inductive reactance in series with a resistance.

Radiation Resistance

The power supplied to an antenna is dissipated in two ways: radiation of electromagnetic waves, and heat losses in the wire and nearby dielectrics. The radiated power is what we want, the useful part, but it represents a form of "loss" just as much as the power used in heating the wire or nearby dielectrics is a loss. In either case, the dissipated power is equal to I^2R. In the case of heat losses, R is a real resistance. In the case of radiation, however, R is a "virtual" resistance, which, if replaced with an actual resistor of the same value, would dissipate the power actually radiated from the antenna. This resistance is called the *radiation resistance*. The total power in the antenna is therefore equal to $I^2(R_0+R)$, where R_0 is the radiation resistance and R represents the total of all the loss resistances.

In ordinary antennas operated at amateur frequencies, the power lost as heat in the conductor does not exceed a few percent of the total power supplied to the antenna. Expressed in decibels, the loss is less than 0.1 dB. The RF loss resistance of copper wire even as small as #14 is very low compared with the radiation resistance of an antenna that is reasonably clear of surrounding objects and is not too close to the ground. You can therefore assume that the ohmic loss in a reasonably well-located antenna is negligible, and that the total resistance shown by the antenna (the feed-point resistance) is radiation resistance. As a radiator of electromagnetic waves, such an antenna is a highly efficient device.

Impedance of a Center-Fed Dipole

A fundamental type of antenna is the *center-fed half-wave dipole*. Historically, the λ/2 dipole has been the most popular antenna used by amateurs worldwide, largely because it is very simple to construct and because it is an effective performer. It is also a basic building block for many other antenna systems, including beam antennas, such as Yagis.

A center-fed half-wave *dipole* consists of a straight wire, one-half wavelength long as defined in Eq 1, and fed in the center. The term "dipole" derives from Greek words meaning "two poles." See **Fig 1**. A λ/2-long dipole is just one form a dipole can take. Actually, a center-fed dipole can be any length electrically, as long as it is configured in a symmetrical fashion with two equal-length legs. There are also versions of dipoles that are not fed in the center. These are called *off-center-fed dipoles*, sometimes abbreviated as "OCF dipoles."

In free space—with the antenna remote from everything else—the theoretical impedance of a physically half-wave long antenna made of an infinitely thin conductor is 73 + j 42.5 Ω. This antenna exhibits both resistance and reactance. The positive sign in the + j 42.5-Ω reactive term indicates that the antenna exhibits an inductive reactance at its feed point. The antenna is slightly long electrically, compared to the length necessary for exact resonance, where the reactance is zero.

The feed-point impedance of any antenna is affected by the wavelength-to-diameter ratio (λ/dia) of the conductors used. Theoreticians like to specify an "infinitely thin" antenna because it is easier to handle mathematically.

What happens if we keep the physical length of an antenna constant, but change the thickness of the wire used in its construction? Further, what happens if we vary the frequency from well below to well above the half-wave resonance and measure the feed-point impedance? **Fig 2** graphs the impedance of a 100-foot long, center-fed dipole in free space, made with extremely thin wire—in this case, wire that is only 0.001 inches in diameter. There is nothing particularly significant about the choice here of 100 feet. This is simply a numerical example.

We could never actually build such a thin antenna (and neither could we install it in free space), but we can model how this antenna works using a very powerful piece of computer software called *NEC-4.1*. See Chapter 4, Antenna Modeling and System Planning, for details on antenna modeling.

The frequency applied to the antenna in Fig 2 is varied from 1 to 30 MHz. The x-axis has a logarithmic scale because of the wide range of feed-point resistance seen over the frequency range. The y-axis has a linear scale representing the reactive portion of the impedance. Inductive reactance is positive and capacitive reactance is negative on the y-axis. The bold figures centered on the spiraling line show the frequency in MHz.

At 1 MHz, the antenna is very short electrically, with a resistive component of about 2 Ω and a series capaci-

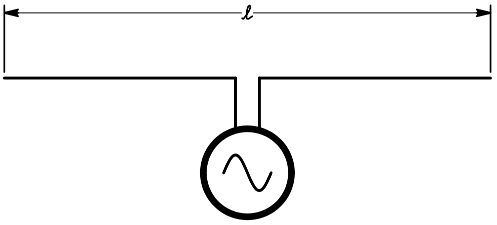

Fig 1—The center-fed dipole antenna. It is assumed that the source of power is directly at the antenna feed point, with no intervening transmission line. Most commonly in amateur applications, the overall length of the dipole is λ/2, but the antenna can in actuality be any length.

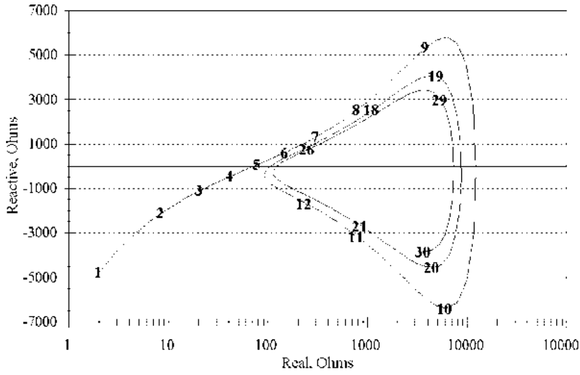

Fig 2—Feed-point impedance versus frequency for a theoretical 100-foot long dipole in free space, fed in the center and made of extremely thin 0.001-inch diameter wire. The y-axis is calibrated in positive (inductive) series reactance up from the zero line, and negative (capacitive) series reactance in the downward direction. The range of reactance goes from –6500 Ω to +6000 Ω. Note that the x-axis is logarithmic because of the wide range of the real, resistive component of the feed-point impedance, from roughly 2 Ω to 10,000 Ω. The numbers placed along the curve show the frequency in MHz.

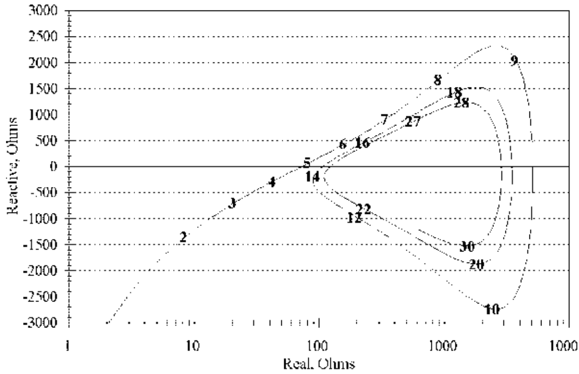

Fig 3—Feed-point impedance versus frequency for a theoretical 100-foot long dipole in free space, fed in the center and made of thin 0.1-inch (#10) diameter wire. Note that the range of change in reactance is less than that shown in Fig 2, ranging from –2700 Ω to +2300 Ω. At about 5,000 Ω, the maximum resistance is also less than that in Fig 2 for the thinner wire, where it is about 10,000 Ω.

tive reactance about –5000 Ω. Close to 5 MHz, the line crosses the zero-reactance line, meaning that the antenna goes through half-wave resonance there. Between 9 and 10 MHz the antenna exhibits a peak inductive reactance of about 6000 Ω. It goes through full-wave resonance (again crossing the zero-reactance line) between 9.5 and 9.6 MHz. At about 10 MHz, the reactance peaks at about –6500 Ω. Around 14 MHz, the line again crosses the zero-reactance line, meaning that the antenna has now gone through 3/2-wave resonance.

Between 19 and 20 MHz, the antenna goes through 4/2-wave resonance, which is twice the full-wave resonance or four times the half-wave frequency. If you allow your mind's eye to trace out the curve for frequencies beyond 30 MHz, it eventually spirals down to a resistive component somewhere between 200 and 3000 Ω. Thus, we have another way of looking at an antenna—as a sort of *transformer*, one that transforms the free-space impedance into the impedance seen at its feed point.

Now look at **Fig 3**, which shows the same kind of spiral curve, but for a thicker-diameter wire, one that is 0.1 inches in diameter. This diameter is close to #10 wire, a practical size we might actually use to build a real dipole. Note that the y-axis scale in Fig 3 is different from that in Fig 2. The range is from ±3000 Ω in Fig 3, while it was ±7000 Ω in Fig 2. The reactance for the thicker antenna ranges from +2300 to –2700 Ω over the whole frequency range from 1 to 30 MHz. Compare this with the range of +5800 to –6400 Ω for the very thin wire in Fig 2.

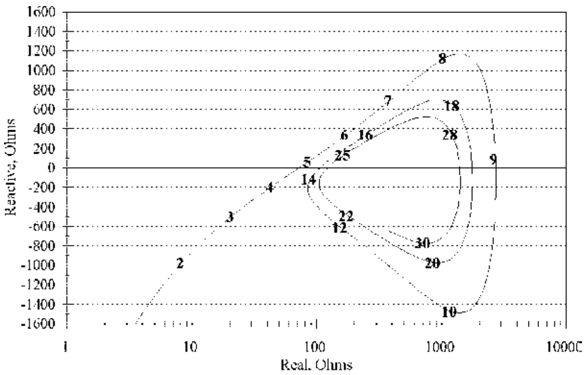

Fig 4—Feed-point impedance versus frequency for a theoretical 100-foot long dipole in free space, fed in the center and made of thick 1.0-inch diameter wire. Once again, the excursion in both reactance and resistance over the frequency range is less with this thick wire dipole than with thinner ones.

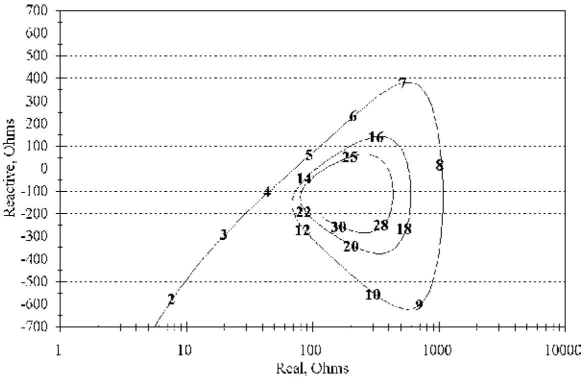

Fig 5—Feed-point impedance versus frequency for a theoretical 100-foot long dipole in free space, fed in the center and made of very thick 10.0-inch diameter wire. This ratio of length to diameter is about the same as a typical rod type of dipole element commonly used at 432 MHz. The maximum resistance is now about 1,000 Ω and the peak reactance range is from about –625 Ω to +380 Ω. This performance is also found in "cage" dipoles, where a number of paralleled wires are used to simulate a fat conductor.

Fig 4 shows the impedance for a 100-foot long dipole using really thick, 1.0-inch diameter wire. The reactance varies from +1000 to –1500 Ω, indicating once again that a larger diameter antenna exhibits less of an excursion in the reactive component with frequency. Note that at the half-wave resonance just below 5 MHz, the resistive component of the impedance is still about 70 Ω, just about what it is for a much thinner antenna. Unlike the reactance, the half-wave radiation resistance of an antenna doesn't radically change with wire diameter, although the maximum level of resistance at full-wave resonance is lower for thicker antennas.

Fig 5 shows the results for a very thick, 10-inch diameter wire. Here, the excursion in the reactive component is even less: about +400 to –600 Ω. Note that the full-wave resonant frequency is about 8 MHz for this

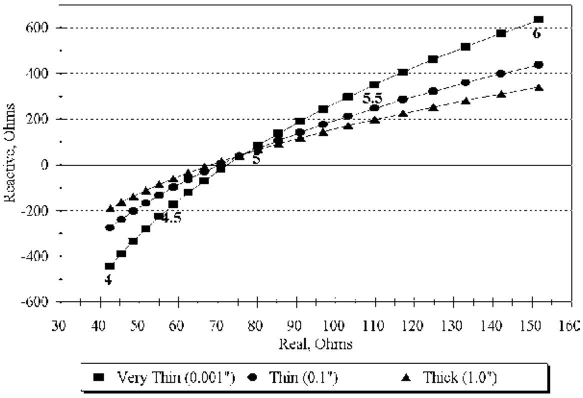

Fig 6—Expansion of frequency range around half-wave resonant point of three center-fed dipoles of three different thicknesses. The frequency is shown along the curves in MHz. The slope of change in series reactance versus series resistance is steeper for the thinner antennas than for the thick 1.0-inch antenna, indicating that the Q of the thinner antennas is higher.

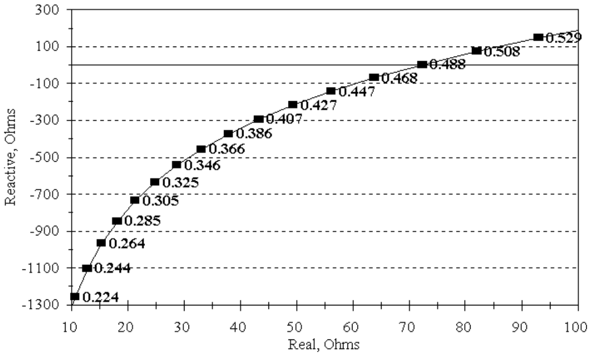

Fig 7—Another way of looking at the data for a 100-foot, center-fed dipole made of #14 wire in free space. The numbers along the curve represent the fractional wavelength, rather than frequency as shown in Fig 6. Note that this antenna goes through its half-wave resonance about 0.488 λ, rather than exactly at a half-wave physical length.

extremely thick antenna, while thinner antennas have full-wave resonances closer to 9 MHz. Note also that the full-wave resistance for this extremely thick antenna is only about 1,000 Ω, compared to the 10,000 Ω shown in Fig 2. All half-wave resonances shown in Figs 2 through 5 remain close to 5 MHz, regardless of the diameter of the antenna wire. Once again, the extremely thick, 10-inch diameter antenna has a resistive component at half-wave resonance close to 70 Ω. And once again, the change in reactance near this frequency is very much less for the extremely thick antenna than for thinner ones.

Now, we grant you that a 100-foot long antenna made with 10-inch diameter wire sounds a little odd! A length of 100 feet and a diameter of 10 inches represent a ratio of 120:1 in length to diameter. However, this is about the same length-to-diameter ratio as a 432-MHz half-wave dipole using 0.25-inch diameter elements, where the ratio is 109:1. In other words, the ratio of length-to-diameter for the 10-inch diameter, 100-foot long dipole is not that far removed from what might actually be used at UHF.

Another way of highlighting the changes in reactance and resistance is shown in **Fig 6**. This shows an expanded portion of the frequency range around the half-wave resonant frequency, from 4 to 6 MHz. In this region, the shape of each spiral curve is almost a straight line. The slope of the curve for the very thin antenna (0.001-inch diameter) is steeper than that for the thicker antennas (0.1 and 1.0-inch diameters). **Fig 7** illustrates another way of looking at the impedance data above and below the half-wave resonance. This is for a 100-foot dipole made of #14 wire. Instead of showing the frequency for each impedance point, the wavelength is shown, making the graph more universal in application.

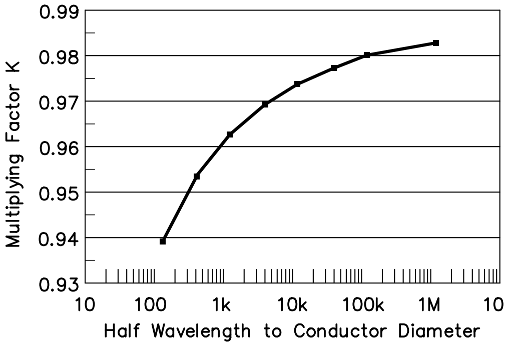

Fig 8—Effect of antenna diameter on length for half-wavelength resonance, shown as a multiplying factor, K, to be applied to the free-space, half-wavelength.

Just to show that there are lots of ways of looking at the same data, **Fig 8** graphs the constant "K" used to multiply the free-space half-wavelength as a function of the ratio between the half-wavelength and the conductor diameter. The curve approaches the value of 1.00 for an infinitely thin conductor, in other words an infinitely large ratio of half-wavelength to diameter.

The behavior of antennas with different λ/diameter ratios corresponds to the behavior of ordinary series-resonant circuits having different values of Q. When the Q of a circuit is low, the reactance is small and changes rather slowly as the applied frequency is varied on either side of resonance. If the Q is high, the converse is true. The response curve of the low-Q circuit is *broad*; that of the high-Q circuit *sharp*. So it is with antennas—the impedance of a thick antenna changes slowly over a comparatively wide band of frequencies, while a thin antenna has a faster change in impedance. Antenna Q is defined

$$Q = \frac{f_0 \Delta X}{2R_0 \Delta f} \qquad \text{(Eq 3)}$$

where f_0 is the center frequency, ΔX is the change in the reactance for a Δf change in frequency, and R_0 is the resistance the f_0. For the "Very Thin," 0.001-inch diameter dipole in Fig 2, a change of frequency from 5.0 to 5.5 MHz yields a reactance change from 86 to 351 Ω, with an R_0 of 95 Ω. The Q is thus 14.6. For the 1.0-inch-diameter "Thick" dipole in Fig 4, $\Delta X = 131\ \Omega$ and R_0 is still 95 Ω, making Q = 7.2 for the thicker antenna, roughly half that of the thinner antenna.

Let's recap. We have described an antenna first as a transducer, then as a sort of transformer to a range of free-space impedances. Now, we just compared the antenna to a series-tuned circuit. Near its half-wave resonant frequency, a center-fed λ/2 dipole exhibits much the same characteristics as a conventional series-resonant circuit. Exactly at resonance, the current at the input terminals is in phase with the applied voltage and the feed-point impedance is purely resistive. If the frequency is below resonance, the phase of the current leads the voltage; that is, the reactance of the antenna is capacitive. When the frequency is above resonance, the opposite occurs; the current lags the applied voltage and the antenna exhibits inductive reactance. Just like a conventional series-tuned circuit, the antenna's reactance and resistance determines its Q.

ANTENNA DIRECTIVITY AND GAIN

The Isotropic Radiator

Before we can fully describe practical antennas, we must first introduce a completely theoretical antenna, the *isotropic radiator*. Envision, if you will, an infinitely small antenna, a point located in outer space, completely removed from anything else around it. Then consider an infinitely small transmitter feeding this infinitely small, point antenna. You now have an isotropic radiator.

The uniquely useful property of this theoretical *point-source* antenna is that it radiates equally well in all directions. That is to say, an isotropic antenna favors no direction at the expense of any other—in other words, it has absolutely no *directivity*. The isotropic radiator is useful as a *measuring stick* for comparison with actual antenna systems.

You will find later that real, practical antennas all exhibit some degree of directivity, which is the property of radiating more strongly in some directions than in others. The radiation from a practical antenna never has the same intensity in all directions and may even have zero radiation in some directions. The fact that a practical antenna displays directivity (while an isotropic radiator does not) is not necessarily a bad thing. The directivity of a real antenna is often carefully tailored to emphasize radiation in particular directions. For example, a receiving antenna that favors certain directions can discriminate against interference or noise coming from other directions, thereby increasing the signal-to-noise ratio for desired signals coming from the favored direction.

Directivity and the Radiation Pattern—a Flashlight Analogy

The directivity of an antenna is directly related to the *pattern* of its radiated field intensity in free space. A graph showing the actual or relative field intensity at a fixed distance, as a function of the direction from the antenna system, is called a *radiation pattern*. Since we can't actually see electromagnetic waves making up the radiation pattern of an antenna, we can consider an analogous situation.

Fig 9 represents a flashlight shining in a totally darkened room. To quantify what our eyes are seeing, we might use a sensitive light meter like those used by photographers, with a scale graduated in units from 0 to 10. We place the meter directly in front of the flashlight and adjust the distance so the meter reads 10, exactly full scale. We also carefully note the distance. Then, always keeping the meter the same distance from the flashlight and keeping it at the same height above the floor, we move the light meter around the flashlight, as indicated by the

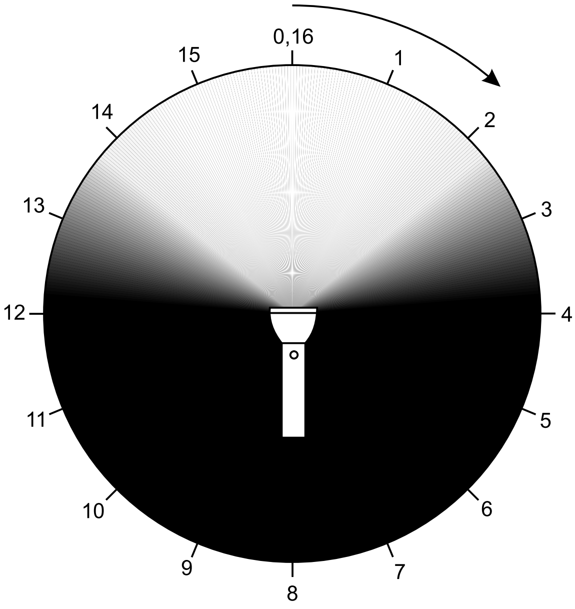

Fig 9—The beam from a flashlight illuminates a totally darkened area as shown here. Readings taken with a photographic light meter at the 16 points around the circle may be used to plot the radiation pattern of the flashlight.

arrow, and take light readings at a number of different positions.

After all the readings have been taken and recorded, we plot those values on a sheet of polar graph paper, like that shown in **Fig 10**. The values read on the meter are plotted at an angular position corresponding to that for which each meter reading was taken. Following this, we connect the plotted points with a smooth curve, also shown in Fig 10. When this is finished, we have completed a radiation pattern for the flashlight.

Antenna Pattern Measurements

Antenna radiation patterns can be constructed in a similar manner. Power is fed to the antenna under test, and a field-strength meter indicates the amount of signal. We might wish to rotate the antenna under test, rather than moving the measuring equipment to numerous positions about the antenna. Or we might make use of antenna reciprocity, since the pattern while receiving is the same as that while transmitting. A source antenna fed by a low-power transmitter illuminates the antenna under test, and the signal intercepted by the antenna under test is fed to a receiver and measuring equipment. Additional information on the mechanics of measuring antenna patterns is contained in Chapter 27, Antennas and Transmission-Line Measurements.

Some precautions must be taken to assure that the measurements are accurate and repeatable. In the case of the flashlight, let's assume that the separation between the light source and the light meter is 2 meters, about 6.5 feet. The wavelength of visible light is about one-half micron, where a micron is one-millionth of a meter.

For the flashlight, a separation of 2 meters between source and detector is $2.0/(0.5 \times 10^{-6}) = 4$ million λ, a very large number of wavelengths. Measurements of practical HF or even VHF antennas are made at much closer distances, in terms of wavelength. For example, at 3.5 MHz a full wavelength is 85.7 meters, or 281.0 feet. To duplicate the flashlight-to-light-meter spacing in wavelengths at 3.5 MHz, we would have to place the field-strength measuring instrument almost on the surface of the Moon, about a quarter-million miles away!

The Fields Around an Antenna

Why should we be concerned with the separation between the source antenna and the field-strength meter, which has its own receiving antenna? One important reason is that if you place a receiving antenna very close to an antenna whose pattern you wish to measure, mutual coupling between the two antennas may actually alter the pattern you are trying to measure.

This sort of mutual coupling can occur in the region very close to the antenna under test. This region is called the *reactive near-field* region. The term "reactive" refers to the fact that the mutual impedance between the transmitting and receiving antennas can be either capacitive or inductive in nature. The reactive near field is sometimes called the "induction field," meaning that the magnetic field usually is predominant over the electric field in this region. The antenna acts as though it were a rather large, lumped-constant inductor or capacitor, storing energy in the reactive near field rather than propagating it into space.

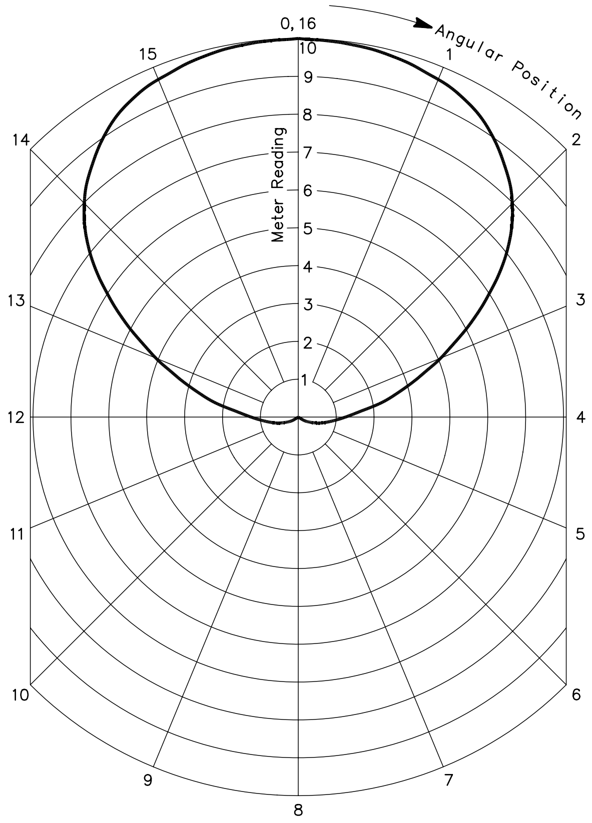

Fig 10—The radiation pattern of the flashlight in Fig 9. The measured values are plotted and connected with a smooth curve.

For simple wire antennas, the reactive near field is considered to be within about a half wavelength from an antenna's radiating center. Later on, in the chapters dealing with Yagi and quad antennas, you will find that mutual coupling between elements can be put to good use to purposely shape the radiated pattern. For making pattern measurements, however, we do not want to be too close to the antenna being measured.

The strength of the reactive near field decreases in a complicated fashion as you increase the distance from the antenna. Beyond the reactive near field, the antenna's radiated field is divided into two other regions: the *radiating near field* and the *radiating far field.* Historically, the terms *Fresnel* and *Fraunhöfer* fields have been used for the radiating near and far fields, but these terms have been largely supplanted by the more descriptive terminology used here. Even inside the reactive near-field

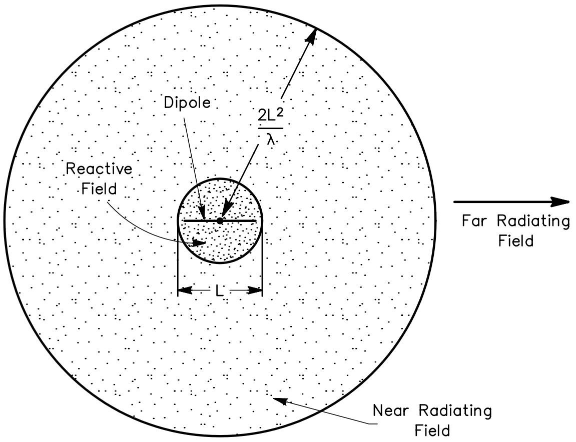

Fig 11—The fields around a radiating antenna. Very close to the antenna, the reactive field dominates. Within this area mutual impedances are observed between antenna and any other antennas used to measure response. Outside of the reactive field, the near radiating field dominates, up to a distance approximately equal to $2L^2/\lambda$, where L is the length of the largest dimension of the antenna. Beyond the near/far field boundary lies the far radiating field, where power density varies as the inverse square of radial distance.

region, both radiating and reactive fields coexist, although the reactive field predominates very close to the antenna.

Because the boundary between the fields is rather fuzzy, experts debate where one field begins and another leaves off, but the boundary between the radiating near and far fields is generally accepted as:

$$D \approx \frac{2L^2}{\lambda} \qquad \text{(Eq 4)}$$

where L is the largest dimension of the physical antenna, expressed in the same units of measurement as the wavelength λ. Remember, many specialized antennas do not follow the rule of thumb in Eq 4 exactly. **Fig 11** depicts the three fields in front of a simple wire antenna.

Throughout the rest of this book we will discuss mainly the radiating far-fields, those forming the traveling electromagnetic waves. Far-field radiation is distinguished by the fact that the intensity is inversely proportional to the distance, and that the electric and magnetic components, although perpendicular to each other in the wave front, are in time phase. The total energy is equally divided between the electric and magnetic fields. Beyond several wavelengths from the antenna, these are the only fields we need to consider. For accurate measurement of radiation patterns, we must place our measuring instrumentation at least several wavelengths away from the antenna under test.

Pattern Planes

Patterns obtained above represent the antenna radiation in just one plane. In the example of the flashlight, the plane of measurement was at one height above the floor. Actually, the pattern for any antenna is three dimensional, and therefore cannot be represented in a single-plane drawing. The *solid* radiation pattern of an antenna in free space would be found by measuring the field strength at every point on the surface of an imaginary sphere having the antenna at its center. The information so obtained would then be used to construct a solid figure, where the distance from a fixed point (representing the antenna) to the surface of the figure is proportional to the field strength from the antenna in any given direction. **Fig 12B** shows a three-dimensional wire-grid representation of the radiation pattern of a half-wave dipole.

For amateur work, *relative* values of field strength (rather than absolute) are quite adequate in pattern plotting. In other words, it is not necessary to know how many microvolts per meter a particular antenna will produce at a distance of 1 mile when excited with a specified power level. (This is the kind of specifications that AM broadcast stations must meet to certify their antenna systems to the FCC.)

For whatever data is collected (or calculated from theoretical equations), it is common to normalize the plotted values so the field strength in the direction of maximum radiation coincides with the outer edge of the chart. On a given system of polar coordinate scales, the *shape* of the pattern is not altered by proper normalization, only its size.

E and H-Plane Patterns

The solid 3-D pattern of an antenna in free space cannot adequately be shown with field-strength data on a flat sheet of paper. Cartographers making maps of a round Earth on flat pieces of paper face much the same kind of problem. As we discussed above, cross-sectional or plane diagrams are very useful for this purpose. Two such diagrams, one in the plane containing the straight wire of a dipole and one in the plane perpendicular to the wire, can convey a great deal of information. The pattern in the plane containing the axis of the antenna is called the *E-plane pattern*, and the one in the plane perpendicular to the axis is called the *H-plane pattern*. These designations are used because they represent the planes in which the electric (symbol E), and the magnetic (symbol H) lines of force lie, respectively.

The E lines represent the *polarization* of the antenna. Polarization will be covered in more detail later in this chapter. As an example, the electromagnetic field pictured in Fig 1 of Chapter 23, Radio Wave Propagation, is the field that would be radiated from a vertically polarized antenna; that is, an antenna in which the conductor is mounted perpendicular to the earth.

When a radiation pattern is shown for an antenna mounted over ground rather than in free space, we automatically gain two frames of reference—an *azimuth angle* and an *elevation angle*. The azimuth angle is usually ref-

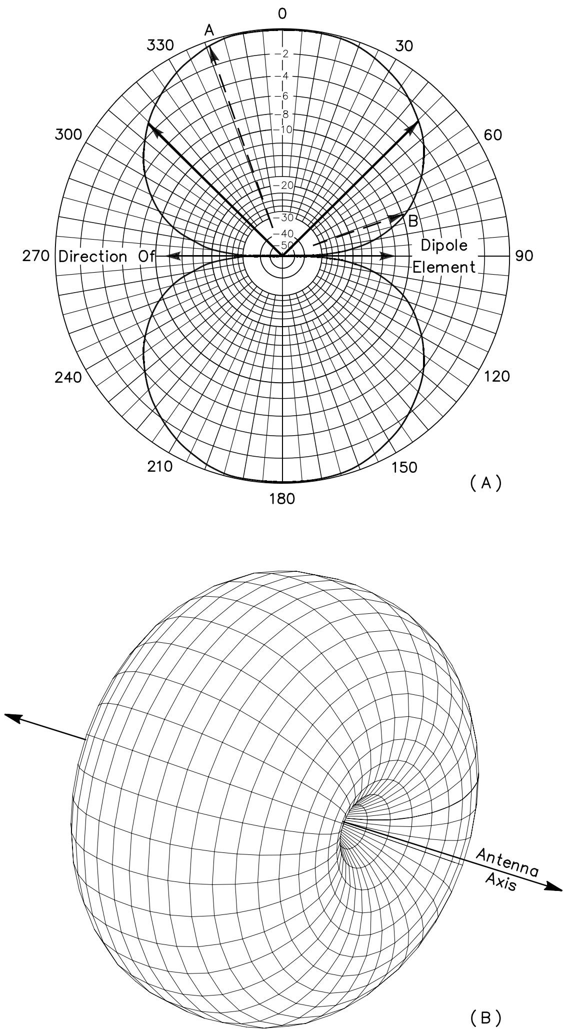

Fig 12—Directive diagram of a free-space dipole. At A, the pattern in the plane containing the wire axis. The length of each dashed-line arrow represents the relative field strength in that direction, referenced to the direction of maximum radiation, which is at right angles to the wire's axis. The arrows at approximately 45° and 315° are the half-power or –3 dB points. At B, a wire grid representation of the solid pattern for the same antenna. These same patterns apply to any center-fed dipole antenna less than a half wavelength long.

Introduction to the Decibel

The power gain of an antenna system is usually expressed in *decibels*. The decibel is a practical unit for measuring power ratios because it is more closely related to the actual effect produced at a distant receiver than the power ratio itself. One decibel represents a just-detectable change in signal strength, regardless of the actual value of the signal voltage. A 20-decibel (20-dB) increase in signal, for example, represents 20 observable steps in increased signal. The power ratio (100 to 1) corresponding to 20 dB gives an entirely exaggerated idea of the improvement in communication to be expected. The number of decibels corresponding to any power ratio is equal to 10 times the common logarithm of the power ratio, or

$$\mathrm{dB} = 10 \log_{10} \frac{P_1}{P_2}$$

If the voltage ratio is given, the number of decibels is equal to 20 times the common logarithm of the ratio. That is,

$$\mathrm{dB} = 20 \log_{10} \frac{V_1}{V_2}$$

When a *voltage* ratio is used, both voltages must be measured across the same value of impedance. Unless this is done the decibel figure is meaningless, because it is fundamentally a measure of a *power* ratio.

The main reason a decibel is used is that successive power gains expressed in decibels may simply be added together. Thus a gain of 3 dB followed by a gain of 6 dB gives a total gain of 9 dB. In ordinary power ratios, the ratios must be multiplied together to find the total gain.

A *reduction* in power is handled simply by subtracting the requisite number of decibels. Thus, reducing the power to ½ is the same as *subtracting* 3 decibels. For example, a power gain of 4 in one part of a system and a reduction to ½ in another part gives a total power gain of 4 × ½ = 2. In decibels, this is 6 – 3 = 3 dB. A power reduction or *loss* is simply indicated by including a negative sign in front of the appropriate number of decibels.

erenced to the maximum radiation lobe of the antenna, where the azimuth angle is defined at 0°, or it could be referenced to the Earth's True North direction for an antenna oriented in a particular compass direction. The E-plane pattern for an antenna over ground is now called the *azimuth pattern*.

The elevation angle is referenced to the horizon at the Earth's surface, where the elevation angle is 0°. Of course, the Earth is round but because its radius is so large, it can in this context be considered to be flat in the area directly under the antenna. An elevation angle of 90° is straight over the antenna, and a 180° elevation is toward the horizon directly behind the antenna.

Professional antenna engineers often describe an antenna's orientation with respect to the point directly overhead—using the *zenith angle*, rather than the elevation angle. The elevation angle is computed by subtracting the zenith angle from 90°.

Referenced to the horizon of the Earth, the H-plane pattern is now called the *elevation pattern*. Unlike the free-space H-plane pattern, the over-ground elevation pattern is drawn as a half-circle, representing only positive elevations above the Earth's surface. The ground reflects or blocks radiation at negative elevation angles, making below-surface radiation plots unnecessary.

After a little practice, and with the exercise of some imagination, the complete solid pattern can be visualized with fair accuracy from inspection of the two planar diagrams, provided of course that the solid pattern of the antenna is *smooth*, a condition that is true for simple antennas like $\lambda/2$ dipoles.

Plane diagrams are plotted on polar coordinate paper, as described earlier. The points on the pattern where the radiation is zero are called *nulls*. The curved section from one null to the next on the plane diagram, or the corresponding section on the solid pattern, is called a *lobe*. The strongest lobe is commonly called the *main lobe*. Fig 12A shows the E-plane pattern for a half-wave dipole. In Fig 12, the dipole is placed in free space. In addition to the labels showing the main lobe and nulls in the pattern, the so-called *half-power* points on the main lobe are shown. These are the points where the power is 3 dB down from the peak value in the main lobe.

Directivity and Gain

Let us now examine directivity more closely. As mentioned previously, all practical antennas, even the simplest types such as dipoles, exhibit directivity. Here's another picture that may help explain the concept of directivity. **Fig 13A** shows a balloon blown into its usual spherical shape. This represents a "reference" isotropic source. Squeezing the balloon in the middle in Fig 13B produces a dipole-like figure-8 pattern whose peak levels at top and bottom are larger than the reference sphere. Compare this with Fig 13C. Next, squeezing the bottom end of the balloon produces a pattern that gives even more "gain" compared to the reference.

Free-space directivity can be expressed quantitatively by comparing the three-dimensional pattern of the antenna under consideration with the perfectly spherical three-dimensional pattern of an isotropic antenna. The field strength (and thus power per unit area, or *power density*) is the same everywhere on the surface of an imaginary sphere having a radius of many wavelengths and having an isotropic antenna at its center. At the surface of the same imaginary sphere around an antenna radiating the same total power, the directive pattern results in greater power density at some points on this sphere and less at others. The ratio of the maximum power density to the average power density taken over the entire sphere (which is the same as from the isotropic antenna under the specified conditions) is the numerical measure of the directivity of the antenna. That is,

$$D = \frac{P}{P_{av}} \quad \text{(Eq 5)}$$

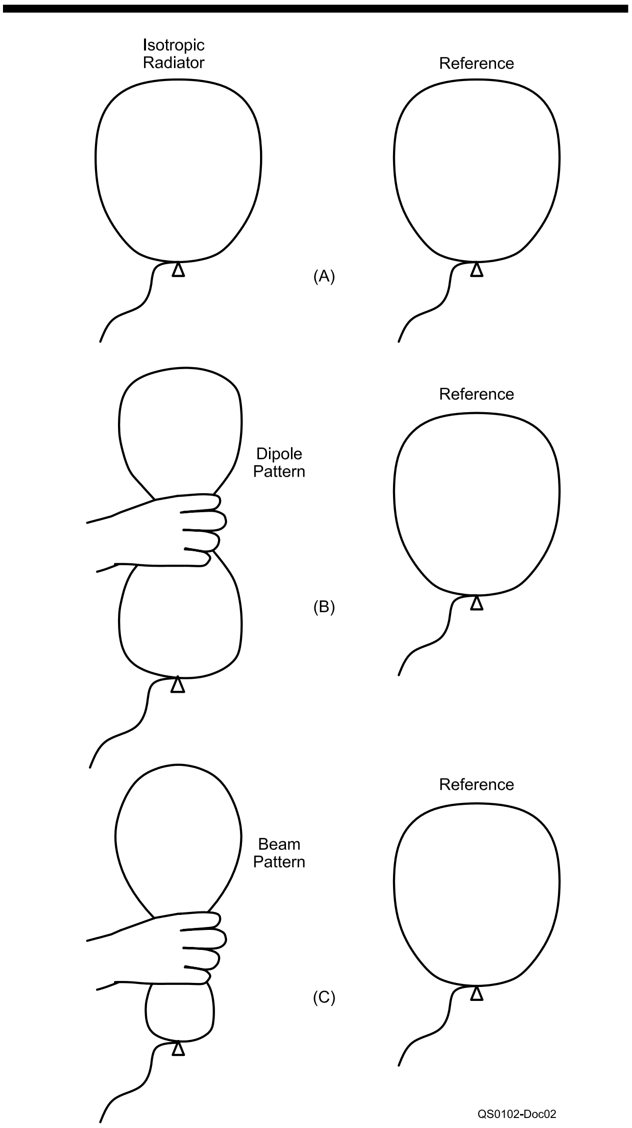

Fig 13—Demonstrating antenna pattern gain with balloons. Take a balloon, blow it up so that it is roughly circular in shape and then declare that this is a radiation pattern from an isotropic radiator. Next, blow up another balloon to the same size and shape and tell the audience that this will be the "reference" antenna (A). Then, squeeze the first balloon in the middle to form a sort of figure-8 shape and declare that this is a dipole and compare the maximum size to that of the reference "antenna" (B). The dipole can be seen to have some "gain" over the reference isotropic. Next, squeeze the end of the first balloon to come up with a sausage-like shape to demonstrate the sort of pattern a beam antenna creates (C).

where

D = directivity
P = power density at its maximum point on the surface of the sphere
P_{av} = average power density

The *gain* of an antenna is closely related to its directivity. Because directivity is based solely on the *shape* of the directive pattern, it does not take into account any power losses that may occur in an actual antenna system. To determine gain, these losses must be subtracted from the power supplied to the antenna. The loss is normally a constant percentage of the power input, so the antenna gain is

$$G = k\frac{P}{P_{av}} = kD \qquad \text{(Eq 6)}$$

where

G = gain (expressed as a power ratio)
D = directivity
k = efficiency (power radiated divided by power input) of the antenna
P and P_{av} are as above

For many of the antenna systems used by amateurs, the efficiency is quite high (the loss amounts to only a few percent of the total). In such cases the gain is essentially equal to the directivity. The more the directive diagram is compressed—or, in common terminology, the *sharper* the lobes—the greater the power gain of the antenna. This is a natural consequence of the fact that as power is taken away from a larger and larger portion of the sphere surrounding the radiator, it is added to the volume represented by the narrow lobes. Power is therefore concentrated in some directions, at the expense of others. In a general way, the smaller the volume of the solid radiation pattern, compared with the volume of a sphere having the same radius as the length of the largest lobe in the actual pattern, the greater the power gain.

As stated above, the gain of an antenna is related to its directivity, and directivity is related to the shape of the directive pattern. A commonly used index of directivity, and therefore the gain of an antenna, is a measure of the width of the major lobe (or lobes) of the plotted pattern. The width is expressed in degrees at the half-power or –3 dB points, and is often called the *beamwidth*.

This information provides only a general idea of relative gain, rather than an exact measure. This is because an absolute measure involves knowing the power density at every point on the surface of a sphere, while a single diagram shows the pattern shape in only one plane of that sphere. It is customary to examine at least the E-plane and the H-plane patterns before making any comparisons between antennas.

A simple approximation for gain over an isotropic radiator can be used, but only if the sidelobes in the antenna's pattern are small compared to the main lobe and if the resistive losses in the antenna are small. When the radiation pattern is complex, numerical integration is employed to give the actual gain.

$$G \approx \frac{41253}{H_{3dB} \times E_{3dB}} \qquad \text{(Eq 7)}$$

where H_{3dB} and E_{3dB} are the half-power points, in degrees, for the H and E-plane patterns.

Radiation Patterns for Center-Fed Dipoles at Different Frequencies

Earlier, we saw how the feed-point impedance of a fixed-length center-fed dipole in free space varies as the frequency is changed. What happens to the radiation pattern of such an antenna as the frequency is changed?

In general, the greater the length of a center-fed antenna, in terms of wavelength, the larger the number of lobes into which the pattern splits. A feature of all such patterns is the fact that the main lobe—the one that gives the largest field strength at a given distance—always is the one that makes the smallest angle with the antenna wire. Furthermore, this angle becomes smaller as the length of the antenna is increased.

Let's examine how the free-space radiation pattern changes for a 100-foot long wire made of #14 wire as the frequency is varied. (Varying the frequency effectively changes the wavelength for a fixed-length wire.) **Fig 14** shows the E-plane pattern at the λ/2 resonant frequency of 4.8 MHz. This is a classical dipole pattern, with a gain in free space of 2.14 dBi referenced to an isotropic radiator.

Fig 15 shows the free-space E-plane pattern for the same antenna, but now at the full-wave (2λ/2) resonant frequency of 9.55 MHz. Note how the pattern has been pinched in at the top and bottom of the figure. In other words, the two main lobes have become sharper at this frequency, making the gain 3.73 dBi, higher than at the λ/2 frequency.

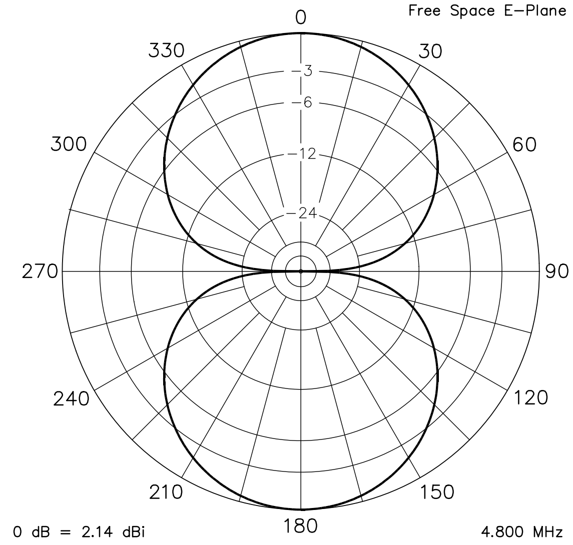

Fig 14—Free-space E-Plane radiation pattern for a 100-foot dipole at its half-wave resonant frequency of 4.80 MHz. This antenna has 2.14 dBi of gain. The dipole is located on the line from 90° to 270°.

Coordinate Scales for Radiation Patterns

A number of different systems of coordinate scales or *grids* are in use for plotting antenna patterns. Antenna patterns published for amateur audiences are sometimes placed on rectangular grids, but more often they are shown using polar coordinate systems. Polar coordinate systems may be divided generally into three classes: linear, logarithmic and modified logarithmic.

A very important point to remember is that the shape of a pattern (its general appearance) is highly dependent on the grid system used for the plotting. This is exemplified in **Fig A**, where the radiation pattern for a beam antenna is presented using three coordinate systems discussed in the paragraphs that follow.

Linear Coordinate Systems

The polar coordinate system for the flashlight radiation pattern, Fig 10, uses linear coordinates. The concentric circles are equally spaced, and are graduated from 0 to 10. Such a grid may be used to prepare a linear plot of the power contained in the signal. For ease of comparison, the equally spaced concentric circles have been replaced with appropriately placed circles representing the decibel response, referenced to 0 dB at the outer edge of the plot. In these plots the minor lobes are suppressed. Lobes with peaks more than 15 dB or so below the main lobe disappear completely because of their small size. This is a good way to show the pattern of an array having high directivity and small minor lobes.

Logarithmic Coordinate System

Another coordinate system used by antenna manufacturers is the logarithmic grid, where the concentric grid lines are spaced according to the logarithm of the voltage in the signal. If the logarithmically spaced concentric circles are replaced with appropriately placed circles representing the decibel response, the decibel circles are graduated linearly. In that sense, the logarithmic grid might be termed a linear-log grid, one having linear divisions calibrated in decibels.

This grid enhances the appearance of the minor lobes. If the intent is to show the radiation pattern of an array supposedly having an omnidirectional response, this grid enhances that appearance. An antenna having a difference of 8 or 10 dB in pattern response around the compass appears to be closer to omnidirectional on this grid than on any of the others. See Fig A-(B).

ARRL Log Coordinate System

The modified logarithmic grid used by the ARRL has a system of concentric grid lines spaced according to the logarithm of 0.89 times the value of the signal voltage. In this grid, minor lobes that are 30 and 40 dB down from the main lobe are distinguishable. Such lobes are of concern in VHF and UHF work. The spacing between plotted points at 0 dB and –3 dB is significantly greater than the spacing between –20 and –23 dB, which in turn is significantly greater than the spacing between –50 and –53 dB.

For example, the scale distance covered by 0 to –3 dB is about $^{1}/_{10}$ of the radius of the chart. The scale distance for the next 3-dB increment (to –6 dB) is slightly less, 89% of the first, to be exact. The scale distance for the next 3-dB increment (to –9 dB) is again 89% of the second. The scale is constructed so that the progression ends with –100 dB at chart center.

The periodicity of spacing thus corresponds generally to the relative significance of such changes in antenna performance. Antenna pattern plots in this publication are made on the modified-log grid similar to that shown in Fig A-(C).

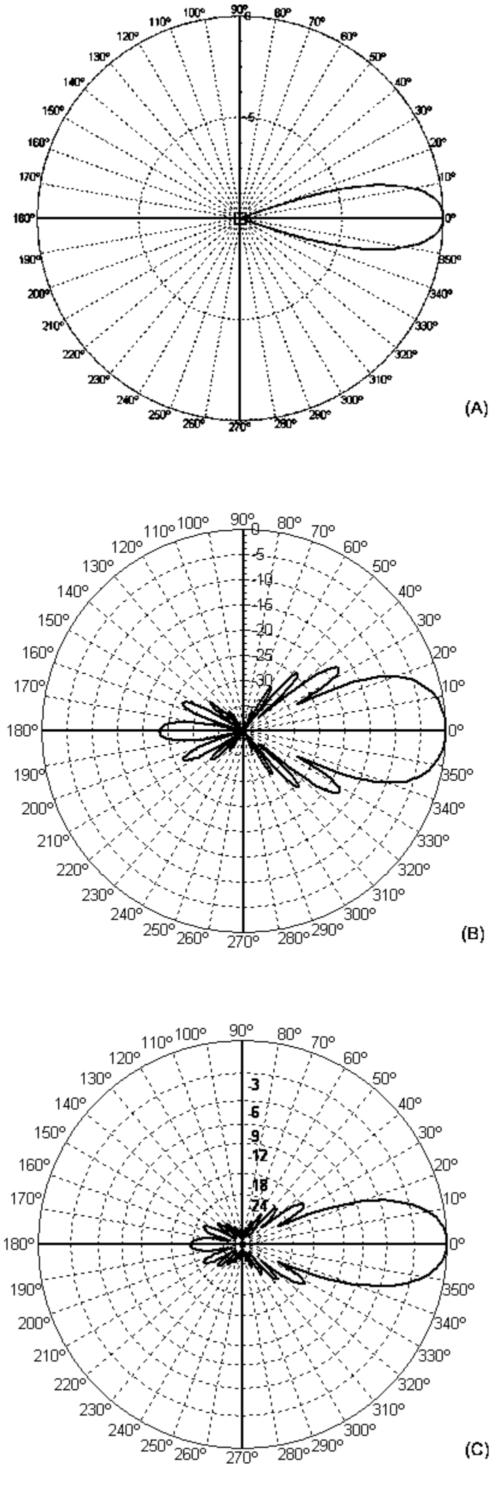

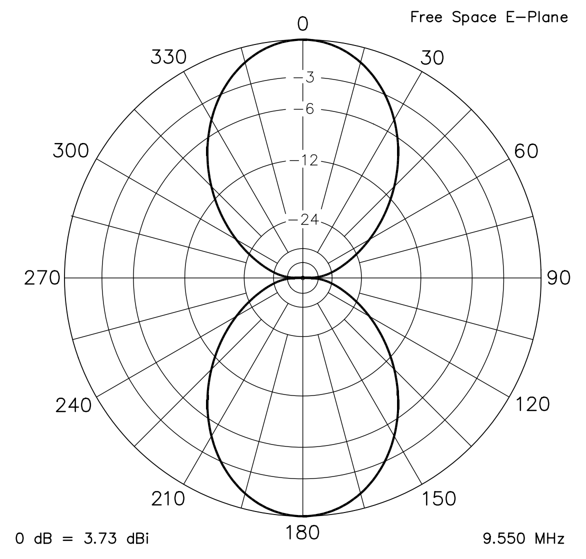

Fig 15—Free-space E-Plane radiation pattern for a 100-foot dipole at its full-wave resonant frequency of 9.55 MHz. The gain has increased to 3.73 dBi, because the main lobes have been focused and sharpened compared to Fig 13.

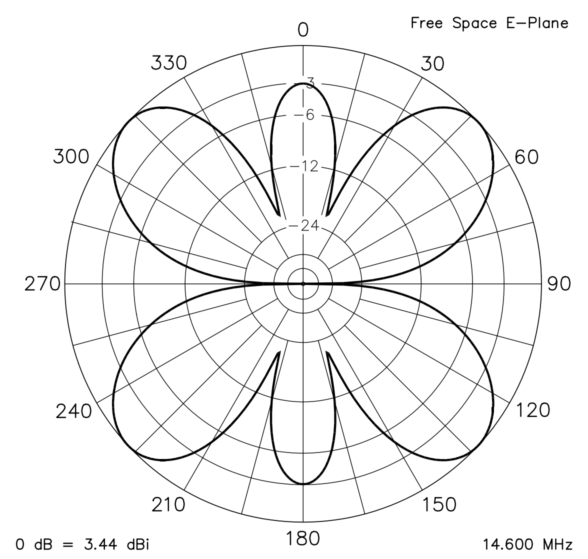

Fig 16—Free-space E-Plane radiation pattern for a 100-foot dipole at its 3/2λ resonant frequency of 14.60 MHz. The pattern has broken up into six lobes, and thus the peak gain has dropped to 3.44 dBi.

Fig 16 shows the pattern at the 3λ/2 frequency of 14.6 MHz. More lobes have developed compared to Fig 14. This means that the power has split up into more lobes and consequently the gain decreases a small amount, down to 3.44 dBi. This is still higher than the dipole at its λ/2 frequency, but lower than at its full-wave frequency. **Fig 17** shows the E-plane response at 19.45 MHz, the 4λ/2, or 2λ, resonant frequency. Now the pattern has reformed itself into only four lobes, and the gain has as a consequence risen to 3.96 dBi.

In **Fig 18** the response has become quite complex at the 5λ/2 resonance point of 24.45 MHz, with ten lobes showing. Despite the presence all these lobes, the main lobes now show a gain of 4.78 dBi. Finally, **Fig 19** shows the pattern at the 3λ (6λ/2) resonance at 29.45 MHz. Despite the fact that there are fewer lobes taking up power than at 24.45 MHz, the peak gain is slightly less at 29.45 MHz, at 4.70 dBi.

Fig A—Radiation pattern plots for a high-gain Yagi antenna on three different grid coordinate systems. At A, the pattern on a linear-power dB grid. Notice how details of sidelobe structure are lost with this grid. At B, the same pattern on a grid with constant 5 dB circles. The sidelobe level is exaggerated when this scale is employed. At B, the same pattern on the modified log grid used by ARRL. The side and rearward lobes are clearly visible on this grid. The concentric circles in all three grids are graduated in decibels referenced to 0 dB at the outer edge of the chart. The patterns look quite different, yet they all represent the same antenna response!

The pattern—and hence the gain—of a fixed-length antenna varies considerably as the frequency is changed. Of course, the pattern and gain change in the same fashion if the frequency is kept constant and the length of the wire is varied. In either case, the wavelength is changing. It is also evident that certain lengths reinforce the pattern to provide more peak gain. If an antenna is not rotated in azimuth when the frequency is changed, the peak gain may occur in a different direction than you might like. In other words, the main lobes change direction as the frequency is varied.

POLARIZATION

We've now examined the first two of the three major properties used to characterize antennas: the radiation pattern and the feed-point impedance. The third general property is *polarization*. An antenna's polarization is defined to be that of its electric field, in the direction where the field strength is maximum.

For example, if a λ/2 dipole is mounted horizontally over the Earth, the electric field is strongest perpendicular to its axis (that is, at right angle to the wire) and parallel to the earth. Thus, since the maximum electric field is horizontal, the polarization in this case is also considered to be *horizontal* with respect to the earth. If the dipole is mounted vertically, its polarization will be *vertical*. See **Fig 20**. Note that if an antenna is mounted in free space, there is no frame of reference and hence its

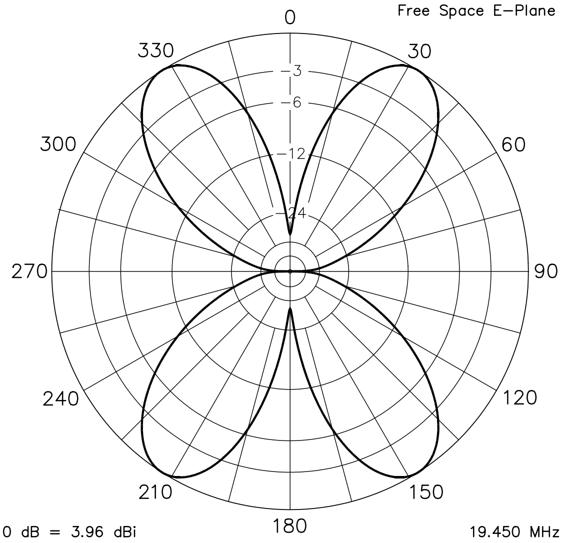

Fig 17—Free-space E-Plane radiation pattern for a 100-foot dipole at twice its full-wave resonant frequency of 19.45 MHz. The pattern has been refocused into four lobes, with a peak gain of 3.96 dBi.

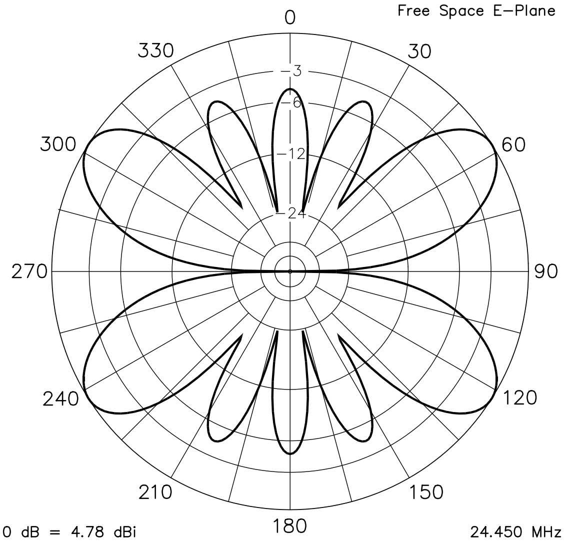

Fig 18—Free-space E-Plane radiation pattern for a 100-foot dipole at its 5/2λ resonant frequency of 24.45 MHz. The pattern has broken down into ten lobes, with a peak gain of 4.78 dBi.

polarization is indeterminate.

Antennas composed of a number of λ/2 elements arranged so that their axes lie in the same or parallel directions have the same polarization as that of any one of the elements. For example, a system composed of a group of horizontal dipoles is horizontally polarized. If both horizontal and vertical elements are used in the same plane and radiate in phase, however, the polarization is the *resultant* of the contributions made by each set of elements to the total electromagnetic field at a given point some distance from the antenna. In such a case the resultant polarization is still *linear,* but is tilted between horizontal and vertical.

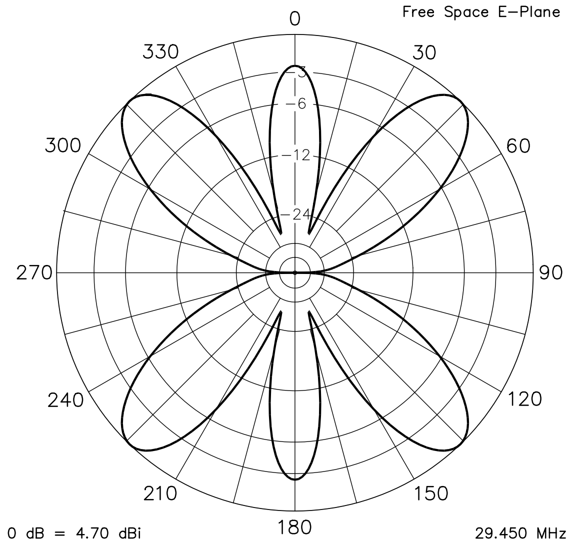

Fig 19—Free-space E-Plane radiation pattern for a 100-foot dipole at three times its full-wave resonant frequency of 29.45 MHz. The pattern has returned to six lobes, with a peak gain of 4.70 dBi.

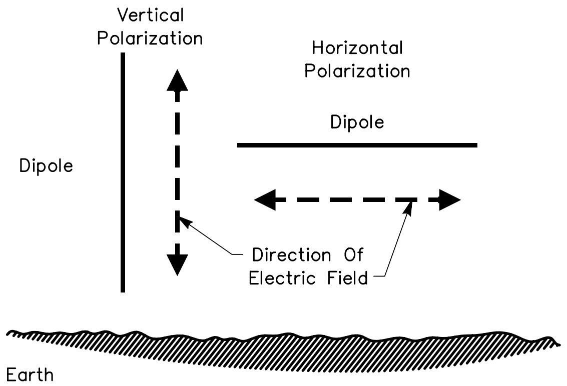

Fig 20—Vertical and horizontal polarization of a dipole above ground. The direction of polarization is the direction of the maximum electric field with respect to the earth.

In directions other than those where the radiation is maximum, the resultant wave even for a simple dipole is a combination of horizontally and vertically polarized components. The radiation off the ends of a horizontal dipole is actually vertically polarized, albeit at a greatly reduced amplitude compared to the broadside horizontally polarized radiation—the sense of polarization changes with compass direction.

Thus it is often helpful to consider the radiation pattern from an antenna in terms of polar coordinates, rather than trying to think in purely linear horizontal or vertical coordinates. See **Fig 21**. The reference axis in a polar system is vertical to the earth under the antenna. The zenith angle is usually referred to as θ (Greek letter theta), and the azimuth angle is referred to as ϕ (Greek letter phi). Instead of zenith angles, most amateurs are more familiar with *elevation angles*, where a zenith angle of 0° is the same as an elevation angle of 90°, straight overhead. Native *NEC* or *MININEC* computer programs use zenith angles rather than elevation angles, although most commercial versions automatically reduce these to elevation angles.

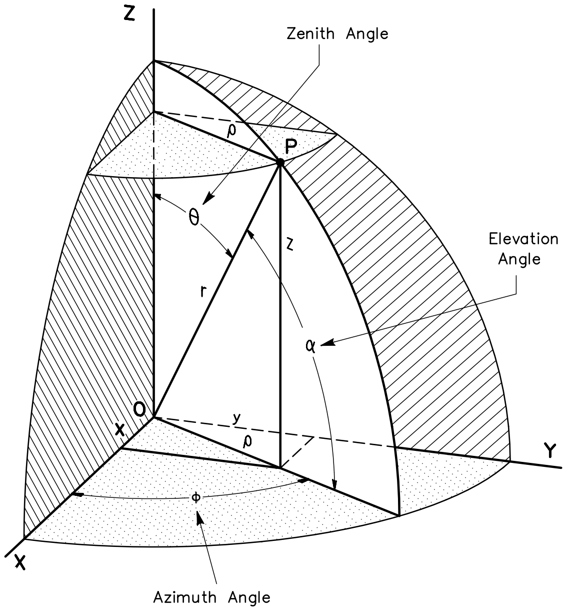

Fig 21—Diagram showing polar representation of a point P lying on an imaginary sphere surround a point-source antenna. The various angles associated with this coordinate system are shown referenced to the x, y and z-axes.

If vertical and horizontal elements in the same plane are fed out of phase (where the beginning of the RF period applied to the feed point of the vertical element is not in time phase with that applied to the horizontal), the resultant polarization is *elliptical*. Circular polarization is a special case of elliptical polarization. The wave front of a circularly polarized signal appears (in passing a fixed observer) to rotate every 90° between vertical and horizontal, making a complete 360° rotation once every period. Field intensities are *equal* at all instantaneous polarizations. Circular polarization is frequently used for space communications, and is discussed further in Chapter 19, Antenna Systems for Space Communications.

Sky-wave transmission usually changes the polarization of traveling waves. (This is discussed in Chapter 23, Radio Wave Propagation.) The polarization of receiving and transmitting antennas in the 3 to 30-MHz range, where almost all communication is by means of sky wave, need not be the same at both ends of a communication circuit (except for distances of a few miles). In this range the choice of polarization for the antenna is usually determined by factors such as the height of available antenna supports, polarization of man-made RF noise from nearby sources, probable energy losses in nearby objects, the likelihood of interfering with neighborhood broadcast or TV reception and general convenience.

Other Antenna Characteristics

Besides the three main characteristics of impedance, pattern (gain) and polarization, there are some other useful properties of antennas.

RECIPROCITY IN RECEIVING AND TRANSMITTING

Many of the properties of a resonant antenna used for reception are the same as its properties in transmission. It has the same directive pattern in both cases, and delivers maximum signal to the receiver when the signal comes from a direction in which the antenna has its best response. The impedance of the antenna is the same, at the same point of measurement, in receiving as in transmitting.

In the receiving case, the antenna is the source of power delivered to the receiver, rather than the load for a source of power (as in transmitting). Maximum possible output from the receiving antenna is obtained when the load to which the antenna is connected is the same as the impedance of the antenna. We say that the antenna is *matched* to its load.

The power gain in receiving is the same as the gain in transmitting, when certain conditions are met. One such condition is that both antennas (usually λ/2-long antennas) must work into load impedances matched to their own impedances, so that maximum power is transferred in both cases. In addition, the comparison antenna should be oriented so it gives maximum response to the signal used in the test. That is, it should have the same polarization as the incoming signal and should be placed so its direction of maximum gain is toward the signal source.

In long-distance transmission and reception via the ionosphere, the relationship between receiving and transmitting, however, may not be exactly reciprocal. This is because the waves do not always follow exactly the same paths at all times and so may show considerable variation in the time between alternations between transmitting and receiving. Also, when more than one ionospheric layer is involved in the wave travel (see Chapter 23, Radio Wave Propagation), it is sometimes possible for reception to be good in one direction and poor in the other, over the same path.

Wave polarization usually shifts in the ionosphere. The tendency is for the arriving wave to be elliptically polarized, regardless of the polarization of the transmitting antenna. Vertically polarized antennas can be expected to show no more difference between transmission and reception than horizontally polarized antennas. On the average, however, an antenna that transmits well in a certain direction also gives favorable reception from the same direction, despite ionospheric variations.

FREQUENCY SCALING

Any antenna design can be scaled in size for use on another frequency or on another amateur band. The dimensions of the antenna may be scaled with Eq 8 below.

$$D = \frac{f1}{f2} \times d \qquad \text{(Eq 8)}$$

where

D = scaled dimension
d = original design dimension
f1 = original design frequency
f2 = scaled frequency (frequency of intended operation)

From this equation, a published antenna design for, say, 14 MHz can be scaled in size and constructed for operation on 18 MHz, or any other desired band. Similarly, an antenna design could be developed experimentally at VHF or UHF and then scaled for operation in one of the HF bands. For example, from Eq 8, an element of 39.0 inches length at 144 MHz would be scaled to 14 MHz as follows: D = 144/14 × 39 = 401.1 inches, or 33.43 feet.

To scale an antenna properly, *all* physical dimensions must be scaled, including element lengths, element spacings, boom diameters and element diameters. Lengths and spacings may be scaled in a straightforward manner as in the above example, but element diameters are often not as conveniently scaled. For example, assume a 14-MHz antenna is modeled at 144 MHz and perfected with $^3/_8$-inch cylindrical elements. For proper scaling to 14 MHz, the elements should be cylindrical, of 144/14 × $^3/_8$ or 3.86 inches diameter. From a realistic standpoint, a 4-inch diameter might be acceptable, but cylindrical elements of 4-inch diameter in lengths of 33 feet or so would be quite unwieldy (and quite expensive, not to mention heavy). Choosing another, more suitable diameter is the only practical answer.

Diameter Scaling

Simply changing the diameter of dipole type elements during the scaling process is not satisfactory without making a corresponding element-length correction. This is because changing the diameter results in a change in the λ/dia ratio from the original design, and this alters the corresponding resonant frequency of the element. The element length must be corrected to compensate for the effect of the different diameter actually used.

To be more precise, however, the purpose of diameter scaling is not to maintain the same resonant frequency for the element, but to maintain the same ratio of self-resistance to self-reactance at the operating frequency—that is, the Q of the scaled element should be the same as that of the original element. This is not always possible to achieve exactly for elements that use several telescoping sections of tubing.

Tapered Elements

Rotatable beam antennas are usually constructed with elements made of metal tubing. The general practice at HF is to taper the elements with lengths of telescoping tubing. The center section has a large diameter, but the ends are relatively small. This reduces not only the weight, but also the cost of materials for the elements. Tapering of HF Yagi elements is discussed in detail in Chapter 11, HF Yagi Arrays.

Length Correction for Tapered Elements

The effect of tapering an element is to alter its electrical length. That is to say, two elements of the same length, one cylindrical and one tapered but with the same average diameter as the cylindrical element, will not be resonant at the same frequency. The tapered element must be made longer than the cylindrical element for the same resonant frequency.

A procedure for calculating the length for tapered elements has been worked out by Dave Leeson, W6NL (ex-W6QHS), from work done by Schelkunoff at Bell Labs and is presented in Leeson's book, *Physical Design of Yagi Antennas*. In the software accompanying this book is a subroutine called EFFLEN.FOR. It is written in Fortran and is used in the *SCALE* program to compute the effective length of a tapered element. The algorithm uses the W6NL-Schelkunoff algorithm and is commented step-by-step to show what is happening. Calculations are made for only one half of an element, assuming the element is symmetrical about the point of boom attachment.

Also, read the documentation SCALE.PDF for the *SCALE* program, which will automatically do the complex mathematics to scale a Yagi design from one frequency to another, or from one taper schedule to another.

The Vertical Monopole

So far in this discussion on Antenna Fundamentals, we have been using the free-space, center-fed dipole as our main example. Another simple form of antenna derived from a dipole is called a *monopole*. The name suggests that this is one half of a dipole, and so it is. The monopole is always used in conjunction with a *ground plane*, which acts as a sort of electrical mirror. See **Fig 22**, where a λ/2 dipole and a λ/4 monopole are compared. The *image antenna* for the monopole is the dotted line beneath the ground plane. The image forms the missing second half of the antenna, transforming a monopole into the functional equivalent of a dipole. From this explanation you can see where the term *image plane* is sometimes used instead of ground plane.

Although we have been focusing throughout this chapter on antennas in free space, practical monopoles are usually mounted vertically with respect to the surface of the ground. As such, they are called *vertical monopoles*, or simply *verticals*. A practical vertical is supplied power by feeding the radiator against a ground system, usually made up of a series of paralleled wires radiating from and laid out in a circular pattern around the base of the antenna. These wires are termed *radials*.

The term *ground plane* is also used to describe a vertical antenna employing a λ/4-long vertical radiator working against a *counterpoise* system, another name for the ground plane that supplies the missing half of the antenna. The counterpoise for a ground-plane antenna usually consists of four λ/4-long radials elevated well above the earth. See **Fig 23**.

Chapter 3, The Effects of Ground, devotes much attention to the requirements for an efficient grounding system for vertical monopole antennas, and Chapter 6, Low-Frequency Antennas, gives more information on ground-plane verticals.

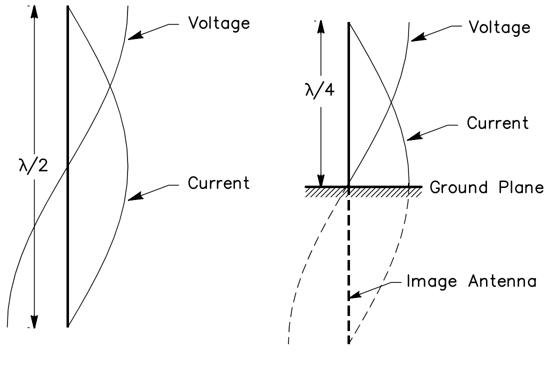

Fig 22—The λ/2 dipole antenna and its λ/4 ground-plane counterpart. The "missing" quarter wavelength is supplied as an image in "perfect" (that is, high-conductivity) ground.

Characteristics of a λ/4 Monopole

The free-space directional characteristics of a λ/4 monopole with its ground plane are very similar to that of a λ/2 antenna in free space. The gain for the λ/4 monopole is slightly less because the H-plane for the λ/2 antenna is compressed compared to the monopole. Like a λ/2 antenna, the λ/4 monopole has an *omnidirectional* radiation pattern in the plane perpendicular to the monopole.

The current in a λ/4 monopole varies practically sinusoidally (as is the case with a λ/2 wire), and is highest at the ground-plane connection. The RF voltage is highest at the open (top) end and minimum at the ground plane. The feed-point resistance close to λ/4 resonance of a vertical monopole over a perfect ground plane is one-half that for a λ/2 dipole at its λ/2 resonance. In this case, a "perfect ground plane" is an infinitely large, lossless conductor.

See **Fig 24**, which shows the feed-point impedance of a vertical antenna made of #14 wire, 50 feet long, located over perfect ground. This is over the whole HF range from 1 to 30 MHz. Again, there is nothing special about the choice of 50 feet for the length of the vertical radiator; it is simply a convenient length for evaluation. **Fig 25** shows an expanded portion of the frequency range above and below the λ/4 resonant point, but now calibrated in terms of wavelength. Note that this particular antenna goes through λ/4 resonance at a length of 0.244 λ, not at exactly 0.25 λ. The exact length for resonance varies with the diameter of the wire used, just as it does for the λ/2 dipole at its λ/2 resonance.

The word *height* is usually used for a vertical monopole antenna whose base is on or near the ground, and in this context, height has the same meaning as *length* when applied to λ/2 dipole antennas. Older texts often refer to heights in electrical degrees, referenced to a free-space wavelength of 360°, but here height is expressed in terms of the free-space wavelength. The range shown in Fig 24 is from 0.132 λ to 0.300 λ, corresponding to a frequency range of 2.0 to 5.9 MHz.

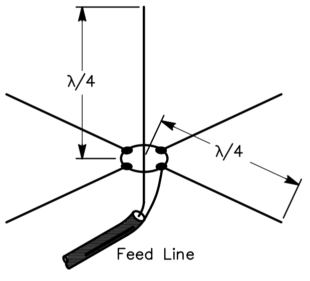

Fig 23—The ground-plane antenna. Power is applied between the base of the vertical radiator and the center of the four ground plane wires.

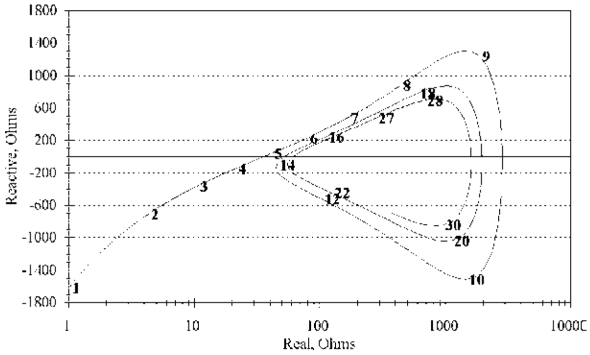

Fig 24—Feed-point impedance versus frequency for a theoretical 50-foot-high grounded vertical monopole made of #14 wire. The numbers along the curve show the frequency in MHz. This was computed using "perfect" ground. Real ground losses will add to the feed-point impedance shown in an actual antenna system.

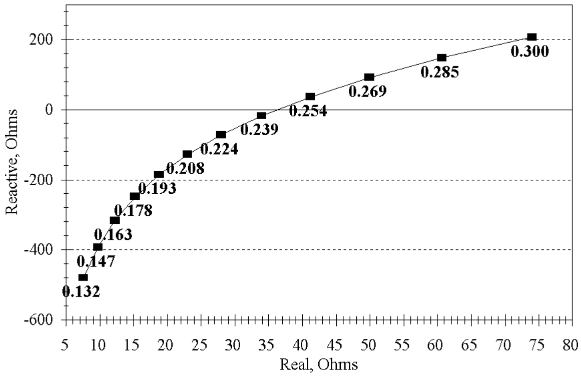

Fig 25—Feed-point impedance for the same antennas as in Fig 21, but calibrated in wavelength rather than frequency, over the range from 0.132 to 0.300λ, above and below the quarter-wave resonance.

The reactive portion of the feed-point impedance depends highly on the length/dia ratio of the conductor, as was discussed previously for a horizontal center-fed dipole. The impedance curve in Figs 24 and 25 is based on a #14 conductor having a length/dia ratio of about 800 to 1. As usual, thicker antennas can be expected to show less reactance at a given height, and thinner antennas will show more.

Efficiency of Vertical Monopoles

This topic of the efficiency of vertical monopole systems will be covered in detail in Chapter 3, The Effects of Ground, but it is worth noting at this point that the efficiency of a real vertical antenna over real earth often suffers dramatically compared with that of a λ/2 antenna. Without a fairly elaborate grounding system, the efficiency is not likely to exceed 50%, and it may be much less, particularly at monopole heights below λ/4.

BIBLIOGRAPHY

J. S. Belrose, "Short Antennas for Mobile Operation," *QST*, Sep 1953.

G. H. Brown, "The Phase and Magnitude of Earth Currents Near Radio Transmitting Antennas," *Proc IRE*, Feb 1935.

G. H. Brown, R. F. Lewis and J. Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proc IRE*, Jun 1937, pp 753-787.

G. H. Brown and O. M. Woodward, Jr., "Experimentally Determined Impedance Characteristics of Cylindrical Antennas," *Proc IRE*, April 1945.

A. Christman, "Elevated Vertical Antenna Systems," *QST*, Aug 1988, pp 35-42.

R. B. Dome, "Increased Radiating Efficiency for Short Antennas," *QST*, Sep 1934, pp 9-12.

A. C. Doty, Jr., J. A. Frey and H. J. Mills, "Characteristics of the Counterpoise and Elevated Ground Screen," Professional Program, Session 9, Southcon '83 (IEEE), Atlanta, GA, Jan 1983.

A. C. Doty, Jr., J. A. Frey and H. J. Mills, "Efficient Ground Systems for Vertical Antennas," *QST*, Feb 1983, pp 20-25.

A. C. Doty, Jr., technical paper presentation, "Capacitive Bottom Loading and Other Aspects of Vertical Antennas," Technical Symposium, Radio Club of America, New York City, Nov 20, 1987.

A. C. Doty, Jr., J. A. Frey and H. J. Mills, "Vertical Antennas: New Design and Construction Data," *The ARRL Antenna Compendium, Volume 2* (Newington: ARRL, 1989), pp 2-9.

R. Fosberg, "Some Notes on Ground Systems for 160 Meters," *QST*, Apr 1965, pp 65-67.

G. Grammer, "More on the Directivity of Horizontal Antennas; Harmonic Operation—Effects of Tilting," *QST*, Mar 1937, pp 38-40, 92, 94, 98.

H. E. Green, "Design Data for Short and Medium Length Yagi-Uda Arrays," *Trans IE Australia*, Vol EE-2, No. 1, Mar 1966.

H. J. Mills, technical paper presentation, "Impedance Transformation Provided by Folded Monopole Antennas," Technical Symposium, Radio Club of America, New York City, Nov 20, 1987.

B. Myers, "The W2PV Four-Element Yagi," *QST*, Oct 1986, pp 15-19.

L. Richard, "Parallel Dipoles of 300-Ohm Ribbon," *QST*, Mar 1957.

J. H. Richmond, "Monopole Antenna on Circular Disc," *IEEE Trans on Antennas and Propagation*, Vol. AP-32, No. 12, Dec 1984.

W. Schulz, "Designing a Vertical Antenna," *QST*, Sep 1978, pp 19- 21.

J. Sevick, "The Ground-Image Vertical Antenna," *QST*, Jul 1971, pp 16-17, 22.

J. Sevick, "The W2FMI 20-Meter Vertical Beam," *QST*, Jun 1972, pp 14-18.

J. Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, Mar 1973, pp. 13-18, 41.

J. Sevick, "A High Performance 20-, 40- and 80-Meter Vertical System," *QST*, Dec 1973.

J. Sevick, "Short Ground-Radial Systems for Short Verticals," *QST*, Apr 1978, pp 30-33.

J. Sevick, *Transmission Line Transformers*, 3rd ed. (Tucker, GA: Noble Publishing, 1996).

C. E. Smith and E. M. Johnson, "Performance of Short Antennas," *Proc IRE*, Oct 1947.

J. Stanley, "Optimum Ground Systems for Vertical Antennas," *QST*, Dec 1976, pp 13-15.

R. E. Stephens, "Admittance Matching the Ground-Plane Antenna to Coaxial Transmission Line," Technical Correspondence, *QST*, Apr 1973, pp 55-57.

D. Sumner, "Cushcraft 32-19 'Boomer' and 324-QK Stacking Kit," Product Review, *QST*, Nov 1980, pp 48-49.

W. van B. Roberts, "Input Impedance of a Folded Dipole," *RCA Review*, Jun 1947.

E. M. Williams, "Radiating Characteristics of Short-Wave Loop Aerials," *Proc IRE*, Oct 1940.

TEXTBOOKS ON ANTENNAS

C. A. Balanis, *Antenna Theory, Analysis and Design* (New York: Harper & Row, 1982).

D. S. Bond, *Radio Direction Finders*, 1st ed. (New York: McGraw-Hill Book Co).

W. N. Caron, *Antenna Impedance Matching* (Newington: ARRL, 1989).

K. Davies, *Ionospheric Radio Propagation*—National Bureau of Standards Monograph 80 (Washington, DC: U.S. Government Printing Office, Apr 1, 1965).

R. S. Elliott, *Antenna Theory and Design* (Englewood Cliffs, NJ: Prentice Hall, 1981).

A. E. Harper, *Rhombic Antenna Design* (New York: D. Van Nostrand Co, Inc, 1941).

K. Henney, *Principles of Radio* (New York: John Wiley and Sons, 1938), p 462.

H. Jasik, *Antenna Engineering Handbook*, 1st ed. (New York: McGraw-Hill, 1961).

W. C. Johnson, *Transmission Lines and Networks*, 1st ed. (New York: McGraw-Hill Book Co, 1950).

R. C. Johnson and H. Jasik, *Antenna Engineering Handbook*, 2nd ed. (New York: McGraw-Hill, 1984).

R. C. Johnson, *Antenna Engineering Handbook*, 3rd ed. (New York: McGraw-Hill, 1993).

E. C. Jordan and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall, Inc, 1968).

R. Keen, *Wireless Direction Finding*, 3rd ed. (London: Wireless World).

R. W. P. King, *Theory of Linear Antennas* (Cambridge, MA: Harvard Univ. Press, 1956).

R. W. P. King, H. R. Mimno and A. H. Wing, *Transmission Lines, Antennas and Waveguides* (New York: Dover Publications, Inc, 1965).

King, Mack and Sandler, *Arrays of Cylindrical Dipoles* (London: Cambridge Univ Press, 1968).

M. G. Knitter, Ed., *Loop Antennas—Design and Theory* (Cambridge, WI: National Radio Club, 1983).

M. G. Knitter, Ed., *Beverage and Long Wire Antennas*—Design and Theory (Cambridge, WI: National Radio Club, 1983).

J. D. Kraus, *Electromagnetics* (New York: McGraw-Hill Book Co).

J. D. Kraus, *Antennas*, 2nd ed. (New York: McGraw-Hill Book Co, 1988).

E. A. Laport, *Radio Antenna Engineering* (New York: McGraw-Hill Book Co, 1952).

J. L. Lawson, *Yagi-Antenna Design*, 1st ed. (Newington: ARRL, 1986).

P. H. Lee, *The Amateur Radio Vertical Antenna Handbook*, 2nd ed. (Port Washington, NY: Cowen Publishing Co., 1984).

A. W. Lowe, *Reflector Antennas* (New York: IEEE Press, 1978).

M. W. Maxwell, *Reflections—Transmission Lines and Antennas* (Newington: ARRL, 1990). Out of print.

M. W. Maxwell, *Reflections II—Transmission Lines and Antennas* (Sacramento: Worldradio Books, 2001).

G. M. Miller, *Modern Electronic Communication* (Englewood Cliffs, NJ: Prentice Hall, 1983).

V. A. Misek, *The Beverage Antenna Handbook* (Hudson, NH: V. A. Misek, 1977).

T. Moreno, *Microwave Transmission Design Data* (New York: McGraw-Hill, 1948).

L. A. Moxon, *HF Antennas for All Locations* (Potters Bar, Herts: Radio Society of Great Britain, 1982), pp 109-111.

Ramo and Whinnery, *Fields and Waves in Modern Radio* (New York: John Wiley & Sons).

V. H. Rumsey, *Frequency Independent Antennas* (New York: Academic Press, 1966).

P. N. Saveskie, *Radio Propagation Handbook* (Blue Ridge Summit, PA: Tab Books, Inc, 1980).

S. A. Schelkunoff, *Advanced Antenna Theory* (New York: John Wiley & Sons, Inc, 1952).

S. A. Schelkunoff and H. T. Friis, *Antennas Theory and Practice* (New York: John Wiley & Sons, Inc, 1952).

J. Sevick, *Transmission Line Transformers* (Atlanta: Noble Publishing, 1996).

H. H. Skilling, *Electric Transmission Lines* (New York: McGraw-Hill Book Co, Inc, 1951).

M. Slurzburg and W. Osterheld, *Electrical Essentials of Radio* (New York: McGraw-Hill Book Co, Inc, 1944).

G. Southworth, *Principles and Applications of Waveguide*

Transmission (New York: D. Van Nostrand Co, 1950).

F. E. Terman, *Radio Engineers' Handbook*, 1st ed. (New York, London: McGraw-Hill Book Co, 1943).

F. E. Terman, *Radio Engineering*, 3rd ed. (New York: McGraw-Hill, 1947).

S. Uda and Y. Mushiake, *Yagi-Uda Antenna* (Sendai, Japan: Sasaki Publishing Co, 1954). [Published in English—Ed.]

P. P. Viezbicke, "*Yagi Antenna Design*," NBS Technical Note 688 (U. S. Dept of Commerce/National Bureau of Standards, Boulder, CO), Dec 1976.

G. B. Welch, *Wave Propagation and Antennas* (New York: D. Van Nostrand Co, 1958), pp 180-182.

The GIANT Book of Amateur Radio Antennas (Blue Ridge Summit, PA: Tab Books, 1979), pp 55-85.

IEEE Standard Dictionary of Electrical and Electronics Terms, 3rd ed. (New York: IEEE, 1984).

Radio Broadcast Ground Systems, available from Smith Electronics, Inc, 8200 Snowville Rd, Cleveland, OH 44141.

Radio Communication Handbook, 5th ed. (London: RSGB, 1976).

Radio Direction Finding, published by the Happy Flyer, 1811 Hillman Ave, Belmont, CA 94002

Chapter 3

The Effects of Ground

The ground around and under an antenna is part of the environment in which any actual antenna must operate. Chapter 2, Antenna Fundamentals, dealt mainly with theoretical antennas in free space, completely removed from the influence of the ground. This chapter is devoted to exploring the interactions between antennas and the ground.

The interactions can be analyzed depending on where they occur relative to two areas surrounding the antenna: the *reactive near field* and the *radiating far field*. You will recall that the reactive near field only exists very close to the antenna itself. In this region the antenna acts as though it were a large lumped-constant inductor or capacitor, where energy is stored but very little is actually radiated. The interaction with the ground in this area creates mutual impedances between the antenna and its environment and these interactions not only modify the feed-point impedance of an antenna, but also often increase losses.

In the radiating far field, the presence of ground profoundly influences the radiation pattern of a real antenna. The interaction is different, depending on the antenna's polarization with respect to the ground. For horizontally polarized antennas, the *shape* of the radiated pattern in the elevation plane depends primarily on the antenna's height above ground. For vertically polarized antennas, both the *shape* and the *strength* of the radiated pattern in the elevation plane strongly depend on the nature of the ground itself (its dielectric constant and conductivity at the frequency of operation), as well as on the height of the antenna above ground.

The Effects of Ground in the Reactive Near Field

FEED-POINT IMPEDANCE VERSUS HEIGHT ABOVE GROUND

Waves radiated from the antenna directly downward reflect vertically from the ground and, in passing the antenna on their upward journey, induce a voltage in it. The magnitude and phase of the current resulting from this induced voltage depends on the height of the antenna above the reflecting surface.

The total current in the antenna consists of two components. The amplitude of the first is determined by the power supplied by the transmitter and the free-space feed-point resistance of the antenna. The second component is induced in the antenna by the wave reflected from the ground. This second component of current, while considerably smaller than the first at most useful antenna heights, is by no means insignificant. At some heights, the two components will be in phase, so the total current is larger than is indicated by the free-space feed-point resistance. At other heights, the two components are out of phase, and the total current is the difference between the two components.

Changing the height of the antenna above ground will change the amount of current flow, assuming that the power input to the antenna is constant. A higher current at the same power input means that the effective resistance of the antenna is lower, and vice versa. In other words, the feedpoint resistance of the antenna is affected by the height of the antenna above ground because of mutual coupling between the antenna and the ground beneath it.

The electrical characteristics of the ground affect both the amplitude and the phase of reflected signals. For this reason, the electrical characteristics of the ground under the antenna will have some effect on the impedance of that antenna, the reflected wave having been influenced by the ground. Different impedance values may be encountered when an antenna is erected at identical heights but over different types of earth.

Fig 1 shows the way in which the radiation resistance of horizontal and vertical half-wave antennas varies with height above ground (in λ, wavelengths). The height of the vertical half-wave is the distance from the bottom of the antenna to ground. For horizontally polarized half-wave antennas, the differences between the effects of perfect ground and real earth are negligible if the antenna height is greater than 0.2 λ. At lower heights, the feed-point resistance over perfect ground decreases rapidly as the antenna is brought closer to a theoretically perfect ground, but this does not occur so rapidly for actual ground. Over real earth, the resistance actually begins increasing at heights below about 0.08 λ. The reason for the increasing resistance at very low heights is that more and more of the reactive (induction) field of the antenna is absorbed by the lossy ground in close proximity. This results in increased loss that is reflected in the increased value of the feedpoint resistance.

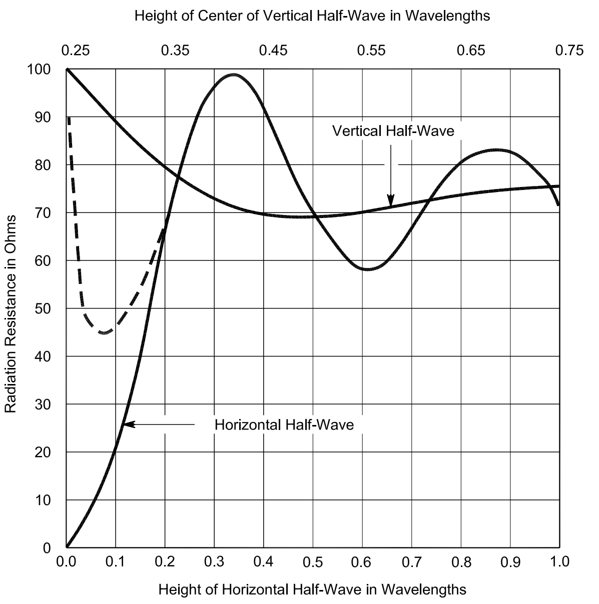

Fig 1—Variation in radiation resistance of vertical and horizontal half-wave antennas at various heights above flat ground. Solid lines are for perfectly conducting ground; the broken line is the radiation resistance of horizontal half-wave antennas at low height over real ground.

For a vertically polarized λ/2-long dipole, differences between the effects of perfect ground and real earth on the feed-point impedance is negligible, as seen in **Fig 1**. The theoretical half-wave antennas on which this chart is based are assumed to have infinitely thin conductors.

GROUND SYSTEMS FOR VERTICAL MONOPOLES

In this section, we'll look at vertical monopoles, which require some sort of ground system in order to make up for the "missing" second half of the antenna and reduce the power lost in the near field. Rudy Severns, N6LF, contributed much of the new material in this chapter.

In Chapter 2, Antenna Fundamentals, and up to this point in this chapter, the discussion about vertical monopoles has mainly been for antennas where *perfect ground* is available. We have also briefly looked at the ground-plane vertical in free space, where the four ground-plane radials form a built-in ground system.

Perfect ground makes a vertical monopole into the functional equivalent of a center-fed dipole, although the feedpoint resistance at resonance is half that of the center-fed dipole. But how can we manage to create that elusive perfect ground, or at least a reasonable approximation, for our real vertical antennas?

Simulating a Perfect Ground in the Reactive Near Field

The effect of a perfectly conducting ground (so far as feed-point resistance and losses are concerned) can be simulated under a real antenna by installing a very large metal screen or mesh, such as poultry netting (chicken wire) or hardware cloth, on or near the surface of the ground. The screen (also called a *counterpoise system*, especially if it is elevated off the ground) should extend at least a half wavelength in every direction from the antenna. The feed-point resistance of a quarter-wave long, thin vertical radiator over such a ground screen will approach the theoretical value of 36.6 Ω. Of course on the lower HF bands such a screen is not practical for most amateurs.

Based on the results of a study published in 1937 by Brown, Lewis and Epstein (see Bibliography), a grounding system consisting of 120 wires, each at least λ/2 long, extending radially from the base of the antenna and spaced equally around a circle, is also the practical equivalent of perfectly conducting ground for reactive-field currents. The wires can either be laid directly on the surface of the ground or buried a few inches below.

Another approach to simulating a perfect ground system is to utilize the ground-plane antenna, with its four ground-plane radials elevated well above lossy earth. Heights (between the bottom of the ground-plane and the surface of the ground) greater than λ/8 have proven to yield excellent results. See Chapter 6, Low-Frequency Antennas, for more details on practical ground-plane verticals.

For a vertical antenna, a large ground screen, either made of wire mesh or a multitude of radials, or an elevated system of ground-plane radials will reduce ground losses near the antenna. This is because the screen conductors are solidly bonded to each other and the resistance is much lower than that of the lossy, low-conductivity earth itself. If the ground screen or elevated ground plane were not present, RF currents would be forced to flow through the lossy, low-conductivity earth to return to the base of the radiator. The ground screen or elevated ground plane in effect shield ground-return currents from the lossy earth.

Less-Than-Ideal Ground Systems

Now, what happens when something less than an ideal ground screen is used as the ground plane for a vertical monopole? Typically this will take the form of an on-ground wire radial system. A great deal of mystery and lack of information seems to surround the vertical antenna ground system. In the case of ground-mounted vertical antennas, many general statements such as "the more radials the better" and "lots of short radials are better than a few long ones" have served as rules of thumb, but many questions as to relative performance differences and optimum number for a given length remain unanswered, as is the justification for the rules of thumb. Most of these questions boil down to one: namely, how many radials, and how long, should be used in a given vertical antenna installation?

A ground system with 120 $\lambda/2$ radials is not very practical for many amateur installations, which often must contend with limited space and funding. Unfortunately the ground resistance, R_g, increases rapidly when the number of radials is reduced. To minimize ground loss where a large, optimum ground system is not possible requires that we understand how ground losses occur and how to optimize the design of a ground system that can fit within the space and budget available.

E and H Fields

E and H fields were introduced in Chapter 2, Antenna Fundamentals, to explain some basic concepts concerning antennas. To understand the reasons for ground loss we need to look at the E and H fields in the near field, but we need to have some feeling for what E and H fields are. The following is a brief description of these fields. It is certainly not a rigorous description but should give at least an intuitive feeling for what is happening.

In 1820 Hans Oerstad discovered that a current flowing in a wire would deflect the needle of a nearby compass. We attribute this effect to a *magnetic* or *H-field*, which at any given location is denoted by the bold-faced letter **H**. **H** is a vector, with an amplitude expressed in A/m (Amperes/meter) and a direction. **Fig 2** shows a typical experimental arrangement. The shape of the magnetic field is roughly shown by the distribution of the iron filings. This field distribution is very similar to that for a vertical antenna.

A compass needle (a small magnet itself) will try to align itself parallel to **H.** As the compass is moved around the conductor, the orientation of the needle changes accordingly. The orientation of the needle gives the direction of **H**. If you attempt to turn the needle away from alignment you will discover a torque trying to restore the needle to its original position. The torque is proportional to the strength of the magnetic field at that point. This is called the *field intensity* or amplitude of **H** at that point. If a larger current flows in the conductor going through the piece of paper holding the iron filings, the amplitude of **H** will be larger. Currents flowing in the conductors of an antenna also generate a magnetic field, one component of the near field.

An antenna will also have an *electric* or *E-field*, which can be visualized using a parallel-plate capacitor, as shown in **Fig 3**. If we connect a battery with a potential V_{dc} across the capacitor plates there will be an electric field **E** established between the plates, as indicated by the lines and directional arrows between the plates. The magnitude of vector **E** is expressed in V/m (volts per meter), so for a potential of V volts and a spacing of d meters, E = V/d V/m. The amplitude of **E** will increase with greater voltage and/or a smaller distance (d). In an antenna, there will be ac potential differences between different parts of the antenna and from the antenna to ground. These ac potential differences establish the electric field associated with the antenna.

Conduction And Displacement Currents

If we replace the dc voltage source in Fig 3 with an ac source, a steady ac current will flow in the circuit. In the

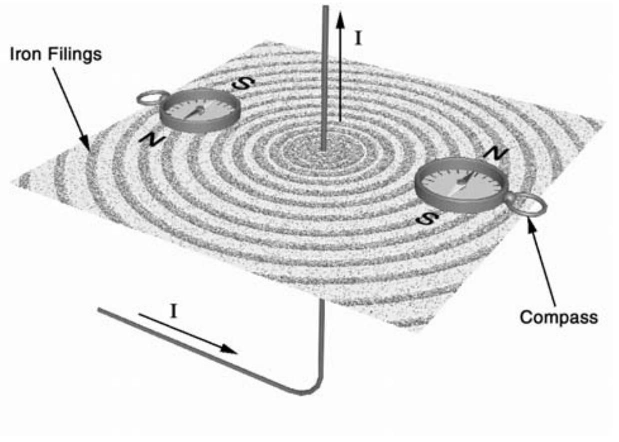

Fig 2—The magnetic lines of force that surround a conductor with an electric current flowing in it are shown by iron filings and small compass needles. The needles point in the direction of the magnetic or H-field. The filings give a general view of the field distribution in the plane perpendicular to the conductor.

conductors between the ac source and the capacitor plates, current (I_c) flows, because of the movement of charge, usually electrons. But in the space between the capacitor plates—particularly in a vacuum—there are no charge carriers available to carry a conduction current. Nonetheless, current still flows in the complete circuit, and we attribute this to a *displacement current* (I_d) flowing between the capacitor plates to account for the continuity of current in the circuit. Displacement and conduction currents are two different phenomena but they both represent current, just two different kinds. Some observers prefer to call conduction currents "currents" and displacement currents "imaginary currents." That terminology is OK, but to account for the current flow in a closed circuit with capacitance you have to keep track of both kinds of current, whatever you call them.

In an antenna over ground, the displacement current represents the current flow from the antenna surface through the air into the ground. The currents flowing in the ground are predominantly conduction currents, but there may also be displacement currents.

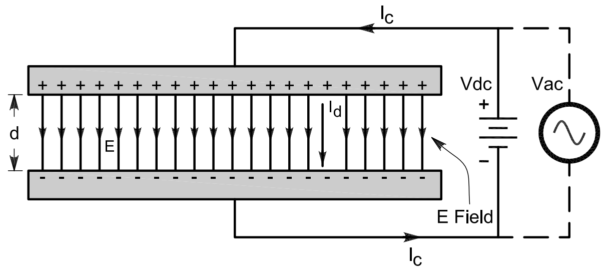

Fig 3—Example of an electric field, $E=V_{dc}/d$. When the dc source is replaced with an ac source there will be a displacement current (I_d) flowing between the capacitor plates.

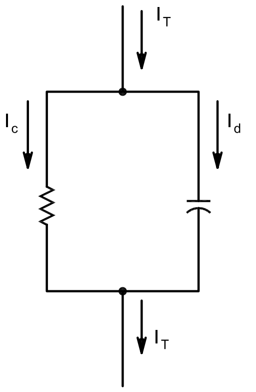

Fig 4—When the capacitor dielectric is less than perfect there will be a conduction current (I_c) in addition to the displacement current (I_d). Soil will typically have both resistive and capacitive components. Power loss in the soil is due to the current flowing through the resistive component.

Where the dielectric material between the capacitor plates is not a perfect insulator, both conduction and displacement currents can flow between the capacitor plates. A good example of this would be a soil dielectric, which has both resistive and capacitive characteristics. Soil can be represented in the circuit of **Fig 4**, where there is a resistor with a conduction current I_c in parallel with a capacitor with a displacement current I_d. The two currents add up (vectorially) as the total current I_T.

A Closer Look at Verticals

A vertical antenna has two field components that induce currents in the ground around the antenna. **Fig 5** shows in a general way the electric-field component (E_z, in V/m) and magnetic-field component (H_ϕ, in A/m) in the region near a vertical. Because the soil near the antenna usually has relatively high resistance, both of these field components will induce currents (I_V and I_H) in the ground, resulting in losses. While the worms may enjoy the heated ground, power dissipated in the ground is subtracted from the radiated power, weakening your signal.

As shown in Fig 5, the tangential component of the H-field (H_ϕ) induces horizontal currents (I_H) flowing radially. The normal component (perpendicular to the ground surface) of the E-field (E_z) induces vertically flowing currents (I_v). Actually, things are more complex than this but we don't need to thrash through that to understand conceptually what's going on.

When modeling an antenna we account for the radiated power (P_r) by assuming there is a resistor we call the *radiation resistance* (R_r) through which the antenna base current (I_o) flows. The radiated power is then $P_r = I_o^2R_r$. Similarly, we can account for the power dissipated in the ground (P_g) by adding a loss resistance (R_g) in series with R_r. The ground loss is then $P_g = I_o^2R_g$. Additional losses due to conductors, loading coils, etc can also be simulated

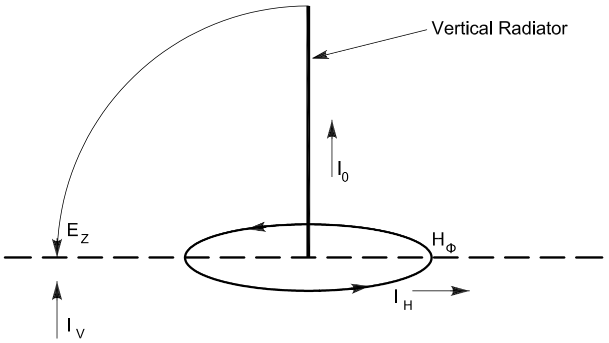

Fig 5—A general view of the fields and ground currents near the base of a vertical antenna. Note that the H-field distribution is equivalent to that shown in Fig 3.

by adding more series loss resistances. Putting aside for the moment these additional losses, the efficiency (η) of a vertical can be expressed as:

$$\eta = \frac{P_r}{P_r + P_g} \qquad \text{(Eq 1)}$$

This can be restated in terms of resistances as:

$$\eta = \frac{R_r}{R_r + R_g} = \frac{1}{1 + \frac{R_g}{R_r}} \qquad \text{(Eq 2)}$$

In essence, efficiency is the ratio of the radiated power to the total input power ($P_T = P_r + P_g$). Another way of saying this is that efficiency depends on the ratio of ground loss resistance (R_g) to radiation resistance (R_r), as Eq 2 shows. The smaller we make R_g the more power will be radiated for a given input power. Reducing R_g is the purpose of the ground system.

A sketch of current flow in the antenna and the surrounding ground (due to H-field), near the base of a vertical, is shown in **Fig 6**. I_z represents the total *zone current* flowing radially through a cylindrical zone at a given radius (r) due to the H field, while I_o is the current at the feed point at the base of the antenna. Technically speaking, the cylinder is infinitely deep, with I_z being the total current integrated over the surface of the cylinder at a given radius.

Fig 7 is a graph of the amplitude of I_z for several antenna heights in wavelengths (h) as we move away from the base of the antenna. Fig 7 shows the zone current that would flow in the ground returning to the base of the antenna, assuming a single ground rod is placed at the feed point for the vertical radiator. The heights indicated are the effective electrical heights. For example, if you use some top loading on the vertical, the effective electrical height will be greater than the physical height.

It is important to recognize that simply adding a top hat to a vertical of a given physical height may reduce ground losses. We can see this from the effect of h on ground current amplitude in Fig 7. Increasing h reduces the ground current. Even something as simple as moving a loading coil from the base up to the center of the antenna may reduce ground losses because it reduces ground current amplitude. But we do have to be careful that the loss introduced by the loading coil does not overcome the reduction in ground loss! Both loading-coil and top-hat schemes also increase the radiation resistance R_r, which further improves efficiency.

The currents in Fig 7 have been adjusted for constant radiated power at the base of the antenna by varying I_o to compensate for the change in R_r as we vary h. To maintain constant radiated power as R_r is falling, you must increase I_o. The base feed-point impedance is a strong function of h. For example, for h = 0.25 λ, R_r will be in the neighborhood of 36 Ω. However, for h = 0.1 λ, R_r will be less than 4 Ω. More information on short antennas can be found in Chapter 16, Mobile and Maritime Antennas.

Fig 7 clearly shows the high currents that flow in the ground near the base of a short antenna due to the antenna's H field. Compared to a 0.25-λ vertical, the 0.1-λ vertical has about three times the base current. As you shorten the antenna further, the zone current increases even more quickly. The ground loss is proportional to the square of the ground current ($P_g = I_g^2 R_g$), so the power loss in the immediate

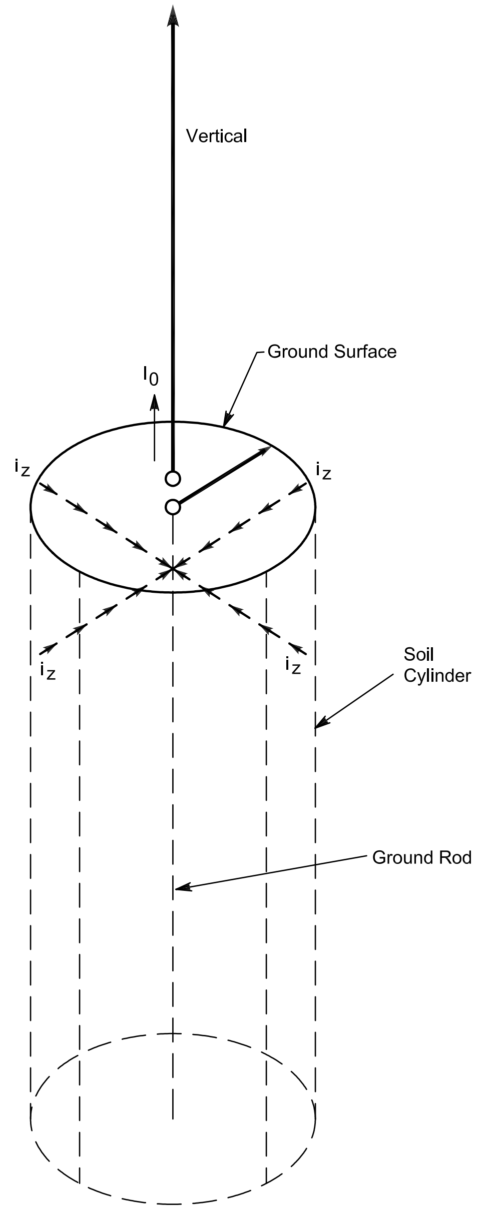

Fig 6—Representation of the zone current near the base of a vertical antenna. Individual i_z components of current flow into a cylinder of soil, with a radius r centered on the base of the vertical. The total current, I_z, thus represents the net current induced in the soil by the H-field for a given radius. (I_z was labeled I_H in Fig 5.)

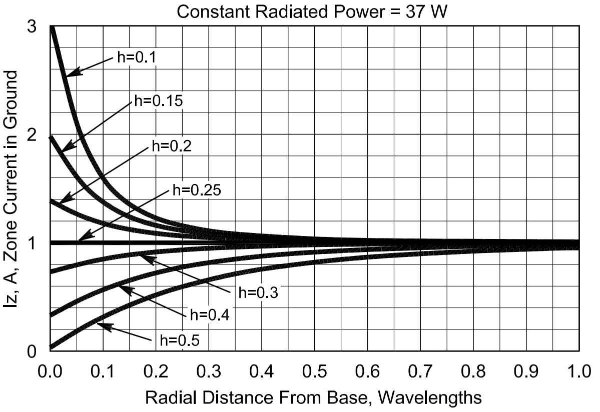

Fig 7—Plot of zone current (I_z) in amperes near the base of a vertical as a function of height (h) and radius (r) in wavelengths. The current in the base of the 0.25-λ antenna is assumed to be 1 A and the current for other values of h is adjusted to maintain the same radiated power ($P_r = I_o^2 R_r$ = 37 W) as the radiation resistance (R_r) changes with h.

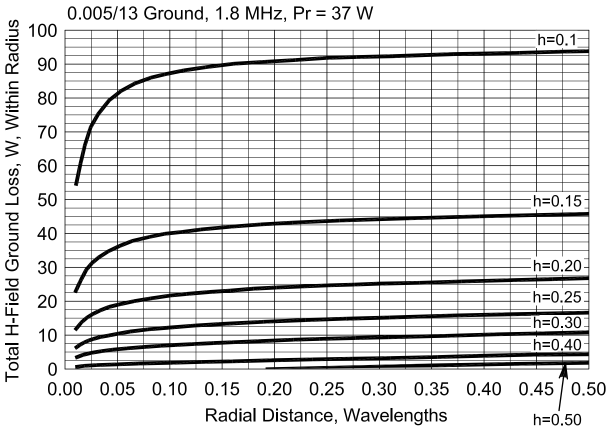

Fig 8—Total H-field induced ground loss within a circle of radius r around the base of a vertical for different values of h and constant P_r = 37 W. Note how the total loss increases rapidly near the base of the antenna indicating high loss. Beyond r = 0.15 λ, however, the additional loss is much lower and the curves flatten out. Note also how much higher the loss is for shorter antennas.

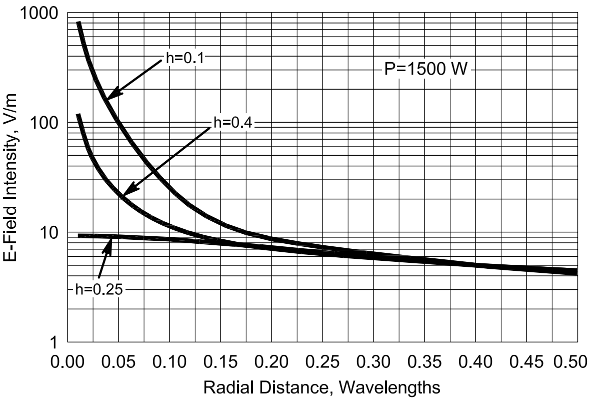

Fig 9—Electric field intensity near the base of a vertical for different values of h. P_r is held constant at 1500 W.

region of the base is much higher for a short antenna operating with the same input power as for a quarter-wave vertical.

We can calculate the losses induced in the soil by either the E- or H-field intensity. **Fig 8** shows an example of the H-field losses for several different antenna heights, given a constant radiated power of 37 W. Note that the total loss within 0.5 λ of the base for h = 0.25 λ is about 16 W (right side of the graph). This gives $\eta = 37/(37 + 16) = 70\%$. However, for h = 0.1 λ, the total loss is about 94 W. Taking into consideration only the H-field losses, $\eta = 37/(37 + 94) = 28\%$. Note that in both cases the majority of the loss is near (< 0.1 λ) the antenna, with the rate of increase of total loss decreasing rapidly as we move farther away from the base, where the lines are almost flat.

Fig 9 is a graph of the E-field intensity around a vertical with 1500 W radiated power, for three values of h. The E-field intensity doesn't depend on the exact type of ground system. (You can see this when you consider that the voltage across a capacitor doesn't depend on the size of the capacitor's plates.) Notice that close to the base, the E-field intensity for the 0.1-λ vertical is almost 100 times that for the 0.25-λ vertical. Because loss is proportional to the square of the voltage, the E-field losses close to the base will be ten thousand times larger in the 0.1-λ vertical! At a 1500-W power level the field intensity near the base of a short vertical is high enough to pose some risk of igniting grass and bushes that grow above any radial system close to the vertical's base. The grass should be kept mowed within 0.1 λ of the base.

Fig 10 shows a computation for the E-field losses, again for a constant radiated power of 37 W and several values of h. For the 0.25-λ vertical, the electric field intensity is quite low and so are the losses associated with it, at only 1.5 W. With any reasonable ground system, the E-field losses for a 0.25-λ vertical will be insignificant.

For shorter or longer verticals, however, the picture is different. This is why we see the very high losses (> 100 W) in Fig 10 for h = 0.1 λ. This loss, when added to the H-field loss, reduces the efficiency of the 0.1-λ vertical to 16% or less without a good ground system. In short antennas the E-field losses cannot be ignored, since they get worse exponentially as the antenna is shortened further.

The presence of a top-loading hat will also increase the E-field intensity in the area below the hat. However, most practical amateur hats will be quite small and the associated E-field loss small. The benefit, however, of reducing I_o because of the addition of the hat—which reduces the field

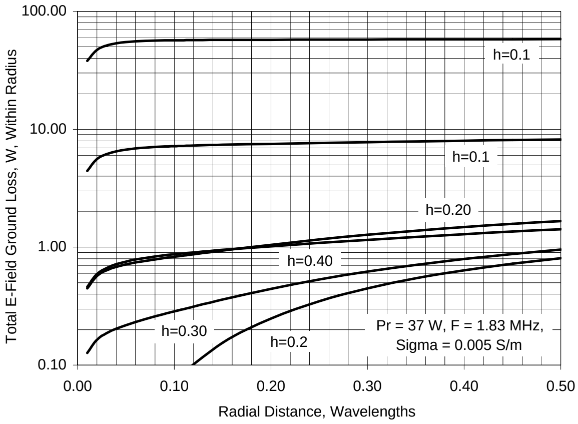

Fig 10—Total E-field induced ground loss within a circle of radius r around the base of a vertical for different values of h and constant P_r = 37 W. Note how the total loss increases rapidly near the base of the antenna, indicating high loss. Beyond r = 0.05 λ, however, the additional loss is much lower and the curves flatten out. Note again how much higher the loss is for shorter antennas. For h = 0.1 λ the E-field loss is greater than the H-field loss.

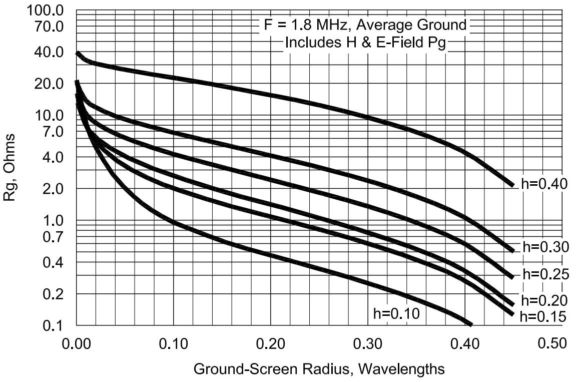

Fig 11—Effective ground resistance (R_g) at the base of the vertical as a function of the radius of a ground screen for several different antenna heights. Note how R_g falls as power dissipation in the soil is eliminated by the highly conducting ground screen.

from the vertical part of the antenna—more than compensates for the small additional E-field loss due to the hat.

Verticals taller than 0.25 λ also display increased E-field intensity, but not nearly so severe as short verticals. In verticals both shorter and longer than a 0.25 λ, the critical loss region is within a radius of about 0.05 λ. We can see this in Fig 10, where the power-loss curves for the shorter antennas flatten out by the time we reach a radius of 0.05 λ.

It's not widely known, but while radial wire systems reduce the H-field losses very effectively, Larsen (see Bibliography) has shown that the E-field losses with the same radial system do not fall in the same fashion as H-field losses. For h > 0.15 λ this doesn't matter much because the E-field loss is so low anyway. However, for short antennas it is very helpful to install either a ground screen or a dense radial system within 0.05 λ of the base.

We can take the data in Figs 8 and 10 and calculate the effective value of the ground resistance R_g. **Fig 11** shows the results of such a computation. Fig 11 assumes a perfect ground screen that varies in radius from 0.001 λ to 0.5 λ. As we would expect, when the ground screen is very small the ground losses are high, meaning R_g is high and the efficiency is low. As we increase the radius of the ground screen, forcing current out of the lossy soil and into the very low-loss screen, R_g drops rapidly and the efficiency increases.

Fig 11 demonstrates why it is desirable to have a good ground system out to at least 0.125 λ, and better yet, even farther. The shorter the antenna, the more important the ground system becomes, especially close to the base. In this example the ground system consists of a highly conductive, bonded ground screen, not always practical for amateur installations. A more typical ground system would consist of a number of individual radial wires. This kind of ground will be inferior to a screen but represents a practical compromise. We'll examine this in more detail shortly.

Note that as we reduce h in Fig 11, R_g actually goes down—even though the ground losses are higher. When h is made smaller the radiation resistance declines rapidly (see Chapter 16, Mobile and Maritime Antennas), so that for a given radiated power I_o must increase. If we measure R_g as we reduce h over a given ground system, we would see that the value for R_g goes down as shown in Fig 11. But because I_o^2 is rising more rapidly than R_g is falling, the power lost in the ground increases and efficiency decreases. The point here is that the value of R_g depends on the ground system, soil characteristics *and* the antenna configuration. You cannot assign an arbitrary value to R_g independent of the antenna system.

Wire Radial Systems

Fig 11 shows R_g for a dense, perfectly conducting ground screen, but what we really need to know is the effect of length and number of individual radials on R_g in a wire radial system. We can calculate the current division between a radial system and the soil and use this to determine R_g. A typical graph of the proportion of the zone current flowing in the radial system, as a function of radius and various numbers of radials (N), is shown in **Fig 12**.

The radial currents decrease as we move away from the base, and the lower the number of radials, the more rapidly the radial current decreases. This means that close to the base of the antenna most of the current is in the radial system, but as we move away from the base the current increasingly flows in the lossy ground. When only a few radials are used, the outer ends of the radials contribute little to reducing ground loss.

Why is this? The problem is that I_z does not go immediately to the nearest radial but may flow for some distance in the soil. This is illustrated in a general way in **Fig 13**. As we move away from the base of the antenna, adjacent radials are further apart from each other and the current must flow further in the soil before it reaches one of the radials. When we use more radials, the distance between radials is less and more of the total current will be in the radials and less in the soil. This reduces ground loss.

When we know the current distribution in the ground we can calculate the power loss and R_g. A typical example for h = 0.25 λ is given in **Fig 14**. We can learn a lot about radial system design from Fig 14 and similar graphs. If we use only a few radials, the radial current drops off very rapidly. Most of the current is flowing in the soil.

Such a ground system is by nature inefficient—that is, R_g is large. We can also see that if we have only 16 radials, R_g falls an ohm or two as we lengthen the radials, but is essentially flat by 0.1 λ. There is no point in making them longer because there is little current in the outer portions and R_g is essentially constant beyond 0.1 λ. As we increase the number of radials, we gather more current further out, making longer radials more useful. The result is cumulative—more radials allow longer radials to be effective and both together reduce ground loss. We can see this in Fig 14, where the initial value of R_g drops as N increases and flattens out at longer radial lengths. For 128 radials, for example, lengths of 0.25 λ or more are useful.

The example in Fig 14 uses #12 wire for the radials. Compared to soil, the resistance of the radial wires is very small, especially if many radials are used, and does not greatly affect overall losses no matter how small the wire. The effect of changing wire size is to slowly change the current division between ground and the radial system. Larger wire results in only a small decrease in R_g. In principle, very small wire could be used for radials but from a mechanical point of view, #18 or #20 wire is about as small as is practical. Any smaller wire breaks too easily to be buried, and also breaks easily when left on the ground surface and is walked on or driven over.

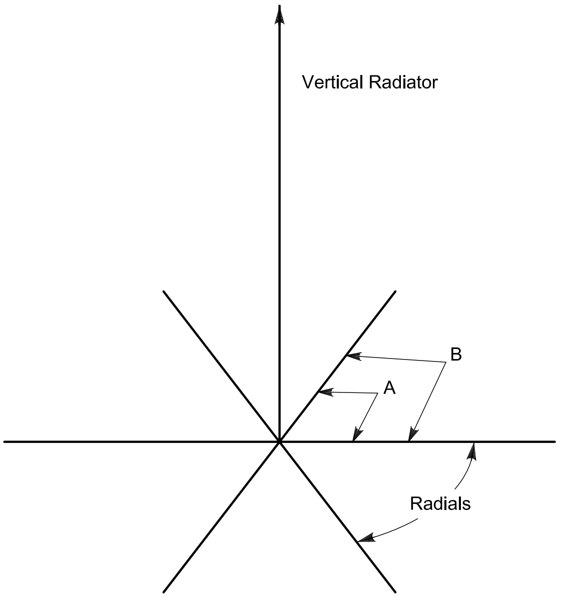

Fig 13—An example of current entering the ground between the radials and flowing for some distance before being picked up by a radial.

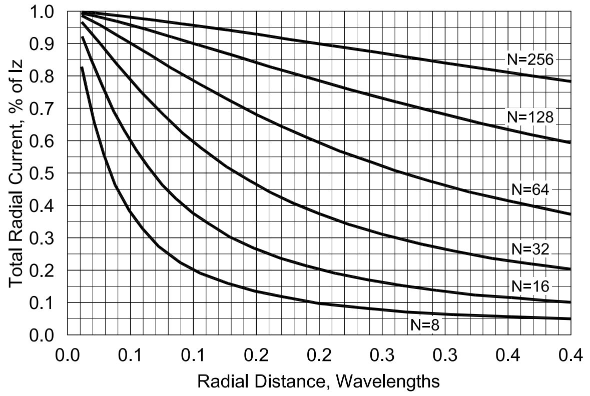

Fig 12—An example of the portion of the zone current flowing in the radial system as you move away from the base of a 0.25 λ vertical for different numbers of radials (N). Note that when more radials are used, more of the zone current flows in the radials and not in the ground, reducing ground loss. The proportion of current in the radial system falls rapidly when only a few radials are used. This leads to high ground loss because most of the zone current is flowing in the ground rather than in the radials.

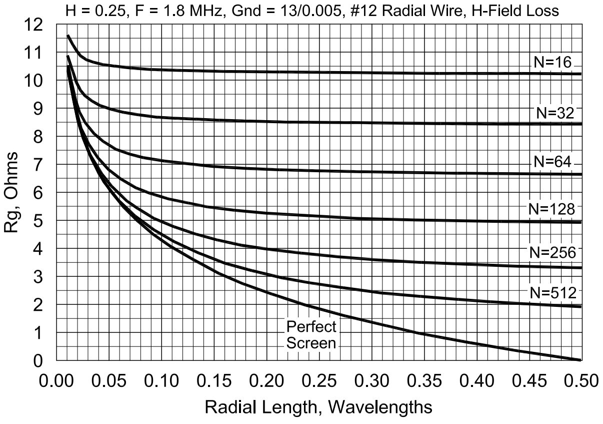

Fig 14—An example of the variation of R_g with radial length and number of radials (N) for h = 0.25 λ. When only a few radials are used there is little point in making them longer than 0.1 λ. Increasing N reduces R_g at a given radius and also makes longer radials useful, further reducing R_g. R_g for other values of h behave similarly.

On the other hand, wire larger than #12 is expensive. Thousands of feet of #8 wire may be affordable for broadcast stations but not for most hams. Increasing the wire size from #20 to #10 would result in only a small reduction in R_g. Of course, if you happen to have a few thousand feet of old RG-8 cable lying around (the diameter is comparable to #0000 wire) then that might indeed help to reduce R_g, as W9QQ has shown (see Bibliography). You are still better off, however, using many radials with small wire than a few radials with large wire. Radial wire size is usually a mechanical and financial issue, not an electrical one.

If the example in Fig 14 were changed for different ground characteristics, then the curves would have a similar shape but would be shifted either up or down. For example, poorer ground will result in higher R_g but the usable length for the radials for a given N would increase somewhat. For better-quality ground, with higher conductivity, R_g will be lower but the usable length of the radials for a given N will be shorter.

For short antennas, the initial drop in R_g will be more rapid and the curves flatten out sooner. This implies that somewhat shorter radials are useful with short antennas. However, given the high losses, it is still a very good idea to use lots of radials with short antennas. As in the above example, increasing N also increases the usable length.

As you go up in frequency from 160 meters, R_g generally rises slowly and then stabilizes around 7 MHz, depending on the ground characteristics. This effect is related to the change in skin depth with frequency, which is discussed in a later section of this chapter. There is also a small shift in current division between the radials and ground as the frequency increases.

A Word Of Caution

In the preceding discussion we presented a number of graphs and the CD-ROM accompanying this book contains some spreadsheets containing the equations from which these graphs were derived. From these graphs we extracted a number of observations on how to design radial systems. Basic to each graph is the assumption that we know the ground characteristics: conductivity and permitivity. In the real world, we amateurs very rarely have more than a rough idea of the ground characteristics under our antennas. Even when careful measurements are made, the characteristics will vary through the year with rainfall or the lack thereof.

Soils are always stratified vertically and can vary by factors or two or more horizontally over distances comparable to radial length, so that even good ground measurements are at best an average. In addition, there will frequently be constraints on the size and shape of the ground system. As a result, we use the calculated information and the previous graphs for general guidance and preliminary design, but when actually installing a ground system we try to measure—or at least estimate—R_g as we go along.

When R_g stops falling, or our patience and/or money run out, we stop adding ground radials. We can measure the feed-point resistance with an impedance bridge to estimate of R_g. The impedance seen at the feed point of the antenna is the sum of the loss and the radiation resistance. To determine R_g you have to estimate R_r (from the antenna height) and other losses due to loading or conductors, and then subtract that from the total measured input resistance. The remainder is R_g, and R_g should fall as we add radials. When R_g stops falling we probably have as many radials of a given length as will be useful. Further reduction in R_g would require more, longer radials.

Practical Suggestions For Vertical Ground Systems

At least 16 radials should be used if at all possible. Experimental measurements and calculations show that with this number, the loss resistance decreases the antenna efficiency by 30% to 50% for a 0.25 λ vertical, depending on soil characteristics. In general, a large number of radials (even though some or all of them must be short) is preferable to a few long radials for a vertical antenna mounted on the ground. The conductor size is relatively unimportant as mentioned before: #12 to #22 copper wire is suitable.

Table 1 summarizes these conclusions. John Stanley, K4ERO, first presented this material in December 1976 *QST*. Another source of information on ground-system design is *Radio Broadcast Ground Systems* (see the Bibliography at the end of this chapter). Most of the data presented in Table 1 is taken from that source, or derived from the interpolation of data contained therein.

Table 1 is based on the number of radials. For each configuration, there is a corresponding optimum radial length. Each configuration also includes the amount of wire used, expressed in wavelengths. Using radials considerably longer than suggested for a given N or using a lot more radials than suggested for a given length, while not adverse to performance, does not yield significant improvement either. That would represent a non-optimum use of wire and construction time. Each suggested configuration represents an optimum relationship between length and number for the given amount of total wire used. Table 1 leads to these conclusions:

- If you install only 16 radials (in configuration A), they need not be very long—0.1 λ is sufficient. The total length of wire will be 1.6 λ, which is about 875 feet at 1.8 MHz.
- If you have the wire, the space and the patience to lay down 120 radials (optimal configuration F), they should be 0.4 λ long. This radial system will gain about 3 dB over the 16-radial case and you'll use 48 λ of wire, or about 26,000 feet at 1.8 MHz.
- If you install 36 radials that are 0.15 λ long, you will lose 1.5 dB compared to optimal configuration F. You will use 5.4 λ of wire, or almost 3,000 feet at 1.8 MHz.

The loss figures in Table 1 assume h = 0.25 λ. A very rough approximation of loss when using shorter antennas can be obtained by doubling the loss in dB each

Table 1
Optimum Ground-System Configurations

	Configuration Designation					
	A	*B*	*C*	*D*	*E*	*F*
Number of radials	16	24	36	60	90	120
Length of each radial in wavelengths	0.1	0.125	0.15	0.2	0.25	0.4
Spacing of radials in degrees	22.5	15	10	6	4	3
Total length of radial wire installed, in wavelengths	1.6	3	5.4	12	22.5	48
Power loss in dB at low angles with a quarter-wave radiating element	3	2	1.5	1	0.5	0*
Feed-point impedance in ohms with a quarter-wave radiating element	52	46	43	40	37	35

Note: Configuration designations are indicated only for text reference.
*Reference: The loss of this configuration is negligible compared to a perfectly conducting ground.

time the antenna height is halved. For taller antennas the losses decrease, approaching 2 dB for configuration A of Table 1 for a half-wave radiator. Even longer antennas yield correspondingly better performance.

Table 1 is based on average ground conductivity. Variation of the loss values shown can be considerable, especially for configurations using fewer radials. Those building antennas over dry, sandy or rocky ground should expect more loss. On the other hand, higher than average soil conductivity and wet soils would make the compromise configurations (those with the fewest radials) even more attractive.

When antennas are combined into arrays, either parasitic or all-driven types, mutual impedances lower the radiation resistance of the elements. This drastically increases the effects of ground loss because I_0 will be higher for the same power level. For instance, an antenna with a 50-Ω feed-point impedance, of which 10 Ω is ground-loss resistance, will have an efficiency of approximately 83%. An array of two similar antennas in a driven array with similar ground losses may have an efficiency of 70% or less.

Special precautions must be taken in such cases to achieve satisfactory operation. Generally speaking, a wide-spaced broadside array presents little problem because R_r is high, but a close-spaced end-fire array should be avoided because R_r is much lower, unless low-loss radial system configurations are used or other precautions taken. Chapter 8, Multielement Arrays, covers the subject of vertical arrays in great detail.

In cases where directivity is desirable or real-estate limitations dictate, longer, more closely spaced radials can be installed in one direction, and shorter, more widely spaced radials in another. Multiband ground systems can be designed using different optimum configurations for different bands. Usually it is most convenient to start at the lowest frequency with fewer radials and add more short radials for better performance on the higher bands.

There is nothing sacred about the exact details of the configurations in Table 1, and small changes in the number of radials and lengths will not cause serious problems. Thus, a configuration with 32 or 40 radials of 0.14 λ or 0.16 λ will work as well as configuration C shown in the table.

If less than 90 radials are contemplated, there is no need to make them a quarter wavelength long. This differs rather dramatically from the case of a ground-plane antenna, where resonant radials are installed above ground. For a ground-mounted antenna, quarter-wave long radials may not be optimum. Because the radials of a ground-mounted vertical are actually on, if not slightly below the surface, they are coupled by capacitance or conduction to the ground, and thus resonance effects are not important. The basic function of radials is to provide a low-loss return path for ground currents.

Radio Broadcast Ground Systems states, "Experiments show that the ground system consisting of only 15 radial wires need not be more than 0.1 wavelength long, while the system consisting of 113 radials is still effective out to 0.5 wavelength." Many graphs in that publication confirm this statement. This is not to say that these two systems will perform equally well; they most certainly will not. However, if 0.1 λ is as long as the radials can be, there is little point in using more than 15 of them unless the vertical radiator's height is also small.

The antenna designer should:

1. Study the cost of various radial configurations versus the gain of each.
2. Compare alternative means of improving transmitted signal and their cost (more power, etc).
3. Consider increasing the physical antenna height (the electrical length) of the vertical radiator, instead of improving the ground system.
4. Use multi-element arrays for directivity and gain, observing the necessary precautions related to mutual impedances discussed in Chapter 8, Multielement Arrays.

The Effect of Ground in the Far Field

The properties of the ground in the far field of an antenna are very important, especially for a vertically polarized antenna, as discussed above. Even if the ground-radial system for a vertical has been optimized to reduce ground-return losses in the reactive near field to an insignificant level, the electrical properties of the ground may still diminish far-field performance to lower levels than "perfect-ground" analyses might lead you to expect. The key is that *ground reflections* from horizontally and vertically polarized waves behave very differently.

Reflections in General

First, let us consider the case of flat ground. Over flat ground, either horizontally or vertically polarized downgoing waves launched from an antenna into the far field strike the surface and are reflected by a process very similar to that by which light waves are reflected from a mirror. As is the case with light waves, the angle of reflection is the same as the angle of incidence, so a wave striking the surface at an angle of, say, 15° is reflected upward from the surface at 15°.

The reflected waves combine with direct waves (those radiated at angles above the horizon) in various ways. Some of the factors that influence this combining process are the height of the antenna, its length, the electrical characteristics of the ground, and as mentioned above, the polarization of the wave. At some elevation angles above the horizon the direct and reflected waves are exactly in phase—that is, the maximum field strengths of both waves are reached at the same time at the same point in space, and the directions of the fields are the same. In such a case, the resultant field strength for that angle is simply the sum of the direct and reflected fields. (This represents a theoretical increase in field strength of 6 dB over the free-space pattern at these angles.)

At other elevation angles the two waves are completely out of phase—that is, the field intensities are equal at the same instant and the directions are opposite. At such angles, the fields cancel each other. At still other angles, the resultant field will have intermediate values. Thus, the effect of the ground is to increase radiation intensity at some elevation angles and to decrease it at others. When you plot the results as an elevation pattern, you will see *lobes* and *nulls*, as described in Chapter 2, Antenna Fundamentals.

The concept of an *image antenna* is often useful to show the effect of reflection. As **Fig 15** shows, the reflected ray has the same path length (AD equals BD) that it would if it originated at a virtual second antenna with the same characteristics as the real antenna, but situated below the ground just as far as the actual antenna is above it.

Now, if we look at the antenna and its image over perfect ground from a remote point on the surface of the ground, we will see that the currents in a horizontally polarized antenna and its image are flowing in opposite directions, or in other words, are 180° out of phase. But the currents in a vertically polarized antenna and its image are flowing in the *same* direction—they are *in* phase. This 180° phase difference between the vertically and horizontally polarized reflections off ground is what makes the combinations with direct waves behave so very differently.

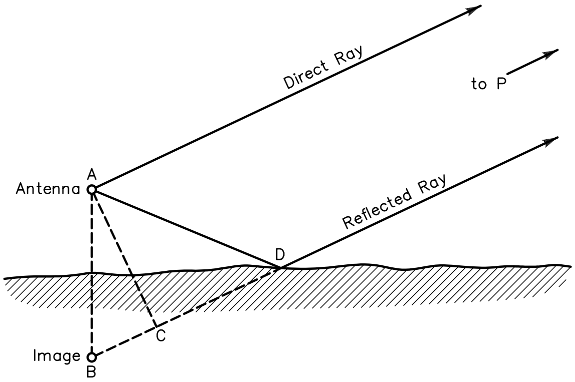

Fig 15—At any distant point, P, the field strength will be the vector sum of the direct ray and the reflected ray. The reflected ray travels farther than the direct ray by the distance BC, where the reflected ray is considered to originate at the *image* antenna.

FAR-FIELD GROUND REFLECTIONS AND THE VERTICAL ANTENNA

A vertical's azimuthal directivity is omnidirectional. A λ/2 vertical over ideal, perfectly conducting earth has the elevation-plane radiation pattern shown by the solid line in **Fig 16**. Over real earth, however, the pattern looks more like the shaded one in the same diagram. In this case, the low-angle radiation that might be hoped for because of perfect-ground performance is not realized in the real world.

Now look at **Fig 17A**, which compares the computed elevation-angle response for two half-wave dipoles at 14 MHz. One is oriented horizontally over ground at a height of λ/2 and the other is oriented vertically, with its center just over λ/2 high (so that the bottom end of the wire doesn't actually touch the ground). The ground is "average" in dielectric constant (13) and conductivity (0.005 S/m). At a 15° elevation angle, the horizontally polarized dipole has almost 7 dB more gain than its vertical brother. Contrast Fig 17A to the comparison in Fig 17B, where the peak gain of a vertically polarized half-wave dipole over seawater, which is virtually perfect for RF reflections, is quite comparable with the horizontal dipole's response at 15°, and exceeds the horizontally polarized antenna dramatically below 15° elevation.

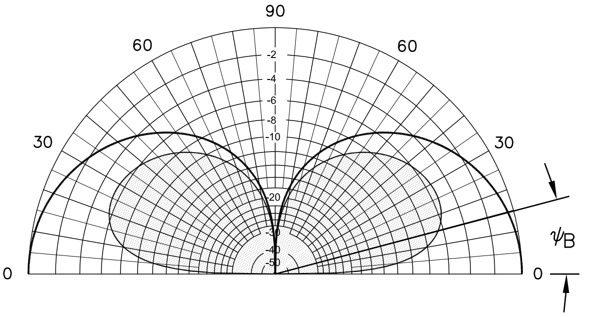

Fig 16—Vertical-plane radiation pattern for a ground-mounted quarter-wave vertical. The solid line is the pattern for perfect earth. The shaded pattern shows how the response is modified over average earth (k = 13, G = 0.005 S/m) at 14 MHz. ψ is the pseudo-Brewster angle (PBA), in this case 14.8°.

To understand in a qualitative fashion why the desired low-angle radiation from a vertical is not delivered when the ground isn't "perfect," examine **Fig 18A**. Radiation from each antenna segment reaches a point P in space by two paths; one directly from the antenna, path AP, and the other by reflection from the earth, path AGP. (Note that P is so far away that the slight difference in angles is insignificant—for practical purposes the waves are parallel to each other at point P.)

If the earth were a perfectly conducting surface, there would be no phase shift of the vertically polarized wave upon reflection at point G. The two waves would add together with some phase difference because of the different path lengths. This difference in path lengths of the two waves is why the free-space radiation pattern differs from the pattern of the same antenna over ground.

Now consider a point P that is close to the horizon, as in Fig 18B. The path lengths AP and AGP are almost the same, so the magnitudes of the two waves add together, producing a maximum at zero angle of radiation. The arrows on the waves point both ways since the process works similarly for transmitting and receiving.

With real earth, however, the reflected wave from a vertically polarized antenna undergoes a change in both *amplitude* and *phase* in the reflection process. Indeed, at a low-enough elevation angle, the phase of the reflected wave will actually change by 180° and its magnitude will then subtract from that of the direct wave. At a zero takeoff angle, it will be almost equal in amplitude, but 180° out of phase with the direct wave.

Note that this is very similar to what happens with horizontally polarized reflected and direct waves at low elevation angles. Virtually complete cancellation will result in a deep null, inhibiting any radiation or reception at 0°. For real-world soils, the vertical loses the theoretical advantage it has at low elevation angles over a horizontal antenna, as Fig 17A so clearly shows.

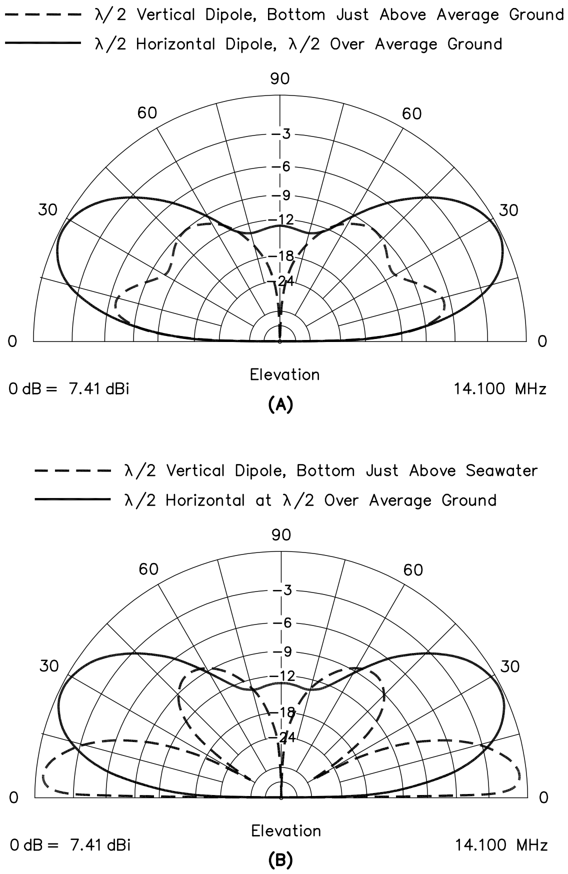

Fig 17—At A, comparison of horizontal and vertical λ/2 dipoles over average ground. Average ground has conductivity of 5 mS/m and dielectric constant of 13. Horizontal dipole is λ/2 high; vertical dipole's bottom wire is just above ground. Horizontal antenna is much less affected by far-field ground losses compared with its vertical counterpart. At B, comparison of 20-meter λ/2 vertical dipole whose bottom wire is just above seawater with λ/2-high horizontal dipole over average ground. Seawater is great for verticals!

The degree that a vertical works better than a horizontal antenna at low elevation angles is largely dependent on the characteristics of the ground around the vertical, as we'll next examine.

THE PSEUDO-BREWSTER ANGLE AND THE VERTICAL ANTENNA

Much of the material presented here regarding pseudo-Brewster angle was prepared by Charles J. Michaels, W7XC, and first appeared in July 1987 *QST*, with additional information in *The ARRL Antenna Compendium, Vol 3*. (See the Bibliography at the end of this chapter.)

Most fishermen have noticed that when the sun is low, its light is reflected from the water's surface as glare, obscuring the underwater view. When the sun is high, however, the sunlight penetrates the water and it is possible to see objects below the surface of the water. The angle at which this transition takes place is known as the *Brewster angle*,

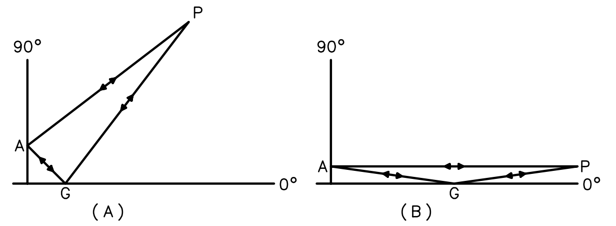

Fig 18—The direct wave and the reflected wave combine at point P to form the pattern (P is very far from the antenna). At A the two paths AP and AGP differ appreciably in length, while at B these two path lengths are nearly equal.

named for the Scottish physicist, Sir David Brewster (1781-1868).

A similar situation exists in the case of vertically polarized antennas; the RF energy behaves as the sunlight in the optical system, and the earth under the antenna acts as the water. The *pseudo-Brewster angle* (PBA) is the angle at which the reflected wave is 90° out of phase with respect to the direct wave. "Pseudo" is used here because the RF effect is similar to the optical effect from which the term gets its name. Below this angle, the reflected wave is between 90° and 180° out of phase with the direct wave, so some degree of cancellation takes place. The largest amount of cancellation occurs near 0°, and steadily less cancellation occurs as the PBA is approached from below.

The factors that determine the PBA for a particular location *are not related to the antenna itself, but to the ground around it*. The first of these factors is earth conductivity, G, which is a measure of the ability of the soil to conduct electricity. Conductivity is the inverse of resistance. The second factor is the dielectric constant, k, which is a unitless quantity that corresponds to the capacitive effect of the earth. For both of these quantities, the higher the number, the better is the ground (for vertical antenna purposes). The third factor determining the PBA for a given location is the frequency of operation. The PBA increases with increasing frequency, all other conditions being equal. **Table 2** gives typical values of conductivity and dielectric constant for different types of soil. The map of **Fig 19** shows the approximate conductivity values for different areas in the continental United States.

As the frequency is increased, the role of the dielectric constant in determining the PBA becomes more significant. **Table 3** shows how the PBA varies with changes in ground conductivity, dielectric constant and frequency. The table shows trends in PBA dependency on ground constants and frequency. The constants chosen are not necessarily typical of any geographical area; they are just examples.

At angles below the PBA, the reflected vertically polarized wave subtracts from the direct wave, causing the radiation intensity to fall off rapidly. Similarly, above the PBA, the reflected wave adds to the direct wave, and the radiated pattern approaches the perfect-earth pattern. Fig 16 shows the PBA, usually labeled ψ_B.

When plotting vertical-antenna radiation patterns over real earth, the reflected wave from an antenna segment is multiplied by a factor called the *vertical reflection coefficient,* and the product is then added vectorially to the direct wave to get the resultant. The reflection coefficient consists of an attenuation factor, A, and a phase angle, ϕ, and is usually expressed as $A\angle\phi$. (ϕ is always a negative angle, because the earth acts as a lossy capacitor in this situation.) The following equation can be used to calculate the reflection coefficient for vertically polarized waves, for earth of given conductivity and dielectric constant at any frequency and elevation angle (also called the wave angle in many texts).

Table 2

Conductivities and Dielectric Constants for Common Types of Earth

Surface Type	*Dielectric Constant*	*Conductivity (S/m)*	*Relative Quality*
Fresh water	80	0.001	
Salt water	81	5.0	
Pastoral, low hills, rich soil, typ Dallas, TX, to Lincoln, NE areas	20	0.0303	Very good
Pastoral, low hills, rich soil typ OH and IL	14	0.01	
Flat country, marshy, densely wooded, typ LA near Mississippi River	12	0.0075	
Pastoral, medium hills and forestation, typ MD, PA, NY, (exclusive of mountains and coastline)	13	0.006	
Pastoral, medium hills and forestation, heavy clay soil, typ central VA	13	0.005	Average
Rocky soil, steep hills, typ mountainous	12-14	0.002	Poor
Sandy, dry, flat, coastal	10	0.002	
Cities, industrial areas	5	0.001	Very Poor
Cities, heavy industrial areas, high buildings	3	0.001	Extremely poor

Table 3
Pseudo-Brewster Angle Variation with Frequency, Dielectric Constant, and Conductivity

Frequency (MHz)	*Dielectric Constant*	*Conductivity (S/m)*	*PBA (degrees)*
7	20	0.0303	6.4
	13	0.005	13.3
	13	0.002	15.0
	5	0.001	23.2
	3	0.001	27.8
14	20	0.0303	8.6
	13	0.005	14.8
	13	0.002	15.4
	5	0.001	23.8
	3	0.001	29.5
21	20	0.0303	10.0
	13	0.005	15.2
	13	0.002	15.4
	5	0.001	24.0
	3	0.001	29.8

$$A_{\mathrm{Vert}} \angle \phi = \frac{k' \sin\psi - \sqrt{k' - \cos^2 \psi}}{k' \sin\psi + \sqrt{k' - \cos^2 \psi}} \qquad \text{(Eq 3)}$$

where

$A_{\mathrm{Vert}} \angle \phi$ = vertical reflection coefficient

ψ = elevation angle

$$k' = k - j \left| \frac{1.8 \times 10^4 \times G}{f} \right|$$

k = dielectric constant of earth (k for air = 1)

G = conductivity of earth in S/m

f = frequency in MHz

j = complex operator ($\sqrt{-1}$)

Solving this equation for several points indicates what effect the earth has on vertically polarized signals at a particular location for a given frequency range. **Fig 20** shows the reflection coefficient as a function of elevation angle at 21 MHz over average earth (G = 0.005 S/m, and k = 13).

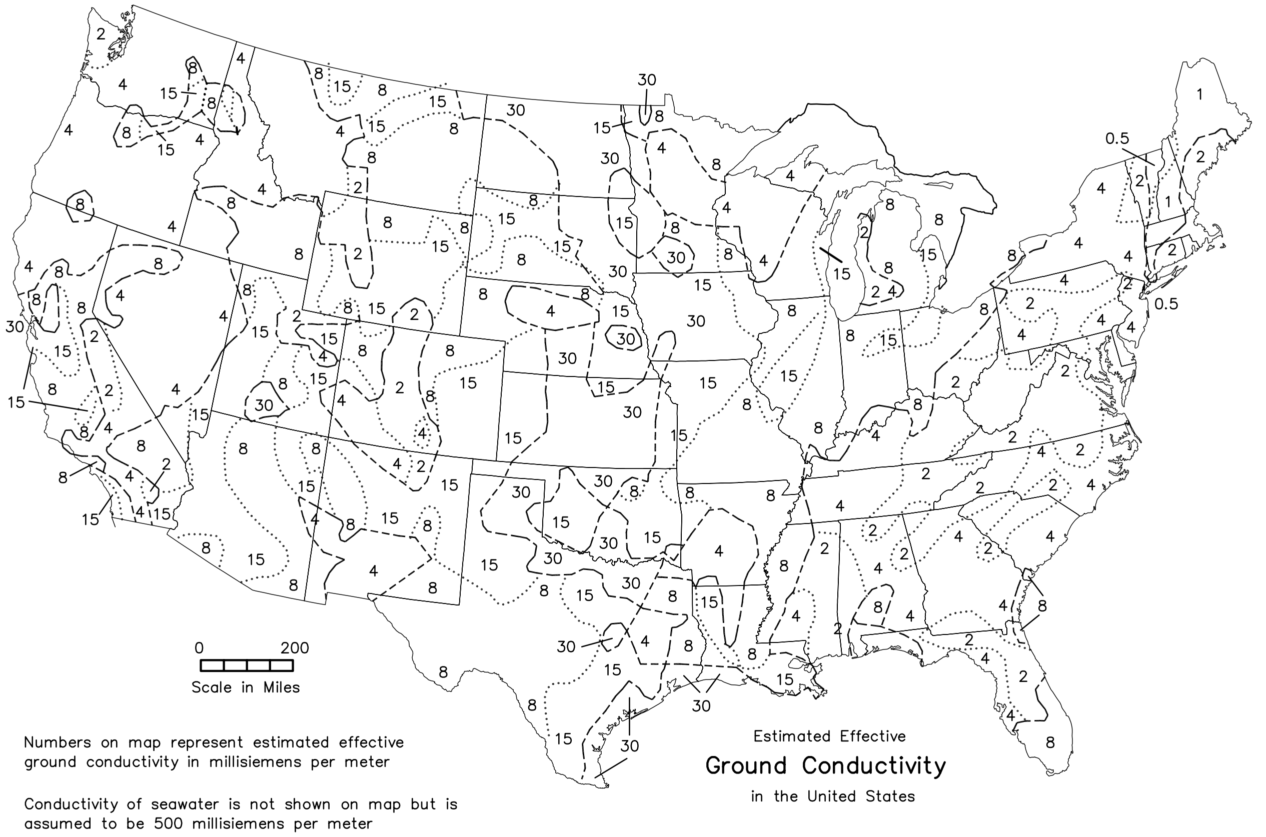

Numbers on map represent estimated effective ground conductivity in millisiemens per meter

Conductivity of seawater is not shown on map but is assumed to be 500 millisiemens per meter

Estimated Effective
Ground Conductivity
in the United States

Fig 19—Typical average soil conductivities for the continental United States. Numeric values indicate conductivities in millisiemens per meter (mS/m), where 1.0 mS/m = 0.001 S/m.

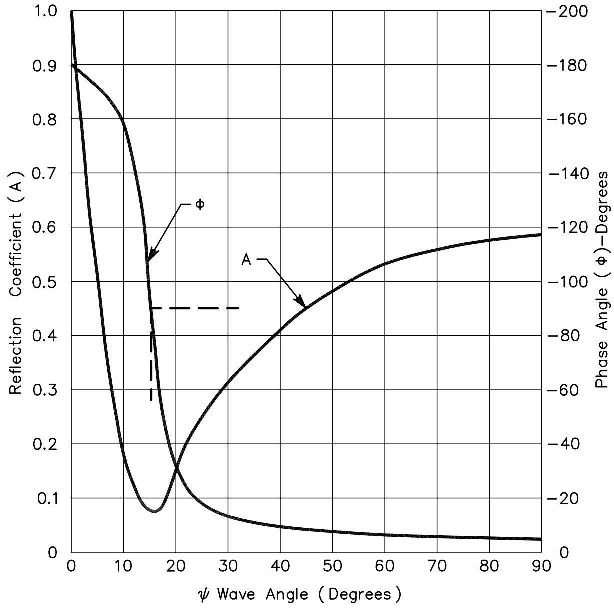

Fig 20—Reflection coefficient for vertically polarized waves. A and ϕ are magnitude and angle for wave angles ψ. This case is for average earth, (k = 13, G = 0.005 S/m), at 21 MHz.

Note that as the phase curve, ψ, passes through 90°, the attenuation curve (A) passes through a minimum at the same wave angle ψ. This is the PBA. At this angle, the reflected wave is not only at a phase angle of 90° with respect to the direct wave, but is so low in amplitude that it does not aid the direct wave by a significant amount. In the case illustrated in Fig 20 this elevation angle is about 15°.

Variations in PBA with Earth Quality

From Eq 3, it is quite a task to search for either the 90° phase point or the attenuation curve minimum for a wide variety of earth conditions. Instead, the PBA can be calculated directly from the following equation.

$$\psi_B = \arcsin\sqrt{\frac{k-1+\sqrt{(x^2+k^2)^2(k-1)^2+x^2[(x^2+k^2)^2-1]}}{(x^2+k^2)^2-1}} \tag{Eq 4}$$

where k, G and f are as defined for Eq 3, and

$$x = \frac{1.8 \times 10^4 \times G}{f}$$

Fig 21 shows curves calculated using Eq 4 for several different earth conditions, at frequencies between 1.8 and 30 MHz. As expected, poorer earths yield higher PBAs. Unfortunately, at the higher frequencies (where low-angle radiation is most important for DX work), the PBAs are highest. The PBA is the same for both transmitting and receiving.

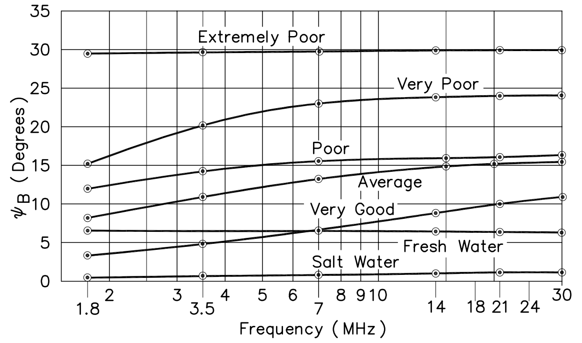

Fig 21—Pseudo-Brewster angle (ψ) for various qualities of earth over the 1.8 to 30-MHz frequency range. Note that the frequency scale is logarithmic. The constants used for each curve are given in Table 2.

Relating PBA to Location and Frequency

Table 2 lists the physical descriptions of various kinds of earth with their respective conductivities and dielectric constants, as mentioned earlier. Note that in general, the dielectric constants and conductivities are higher for better earths. This enables the labeling of the earth characteristics as extremely poor, very poor, poor, average, very good, and so on, without the complications that would result from treating the two parameters independently.

Fresh water and salt water are special cases; in spite of high resistivity, the fresh-water PBA is 6.4°, and is nearly independent of frequency below 30 MHz. Salt water, because of its extremely high conductivity, has a PBA that never exceeds 1° in this frequency range. The extremely low conductivity listed for cities (the last case) in Table 2 results more from the clutter of surrounding buildings and other obstructions than any actual earth characteristic. The PBA at any location can be found for a given frequency from the curves in Fig 21.

FLAT-GROUND REFLECTIONS AND HORIZONTALLY POLARIZED WAVES

The situation for horizontal antennas is different from that of verticals. **Fig 22** shows the reflection coefficient for horizontally polarized waves over average earth at 21 MHz. Note that in this case, the phase-angle departure from 0° never gets very large, and the attenuation factor that causes the most loss for high-angle signals approaches unity for low angles. Attenuation increases with progressively poorer earth types.

In calculating the broadside radiation pattern of a horizontal λ/2 dipole, the perfect-earth image current, equal to the true antenna current but 180° out of phase with it) is multiplied by the horizontal reflection coefficient given by Eq 5 below. The product is then added vectorially to the direct wave to get the resultant at that elevation angle. The reflection coefficient for horizontally polarized waves can be calculated using the following equation.

$$A_{\text{Horiz}}\angle\phi = \frac{\sqrt{k'-\cos^2\psi}-\sin\psi}{\sqrt{k'-\cos^2\psi}+\sin\psi} \qquad \text{(Eq 5)}$$

where

$A_{\text{Horiz}}\angle\phi$ = horizontal reflection coefficient

ψ = elevation angle

$$k' = k - j\left(\frac{1.8\times 10^4 \times G}{f}\right)$$

k = dielectric constant of earth

G = conductivity of earth in S/m

f = frequency in MHz

j = complex operator ($\sqrt{-1}$)

For a horizontal antenna near the earth, the resultant pattern is a modification of the free-space pattern of the antenna. **Fig 23** shows how this modification takes place for a horizontal λ/2 antenna over a perfectly conducting flat surface. The patterns at the left show the relative radiation when one views the antenna from the side; those at the right show the radiation pattern looking at the end of the antenna. Changing the height above ground from λ/4 to λ/2 makes a significant difference in the high-angle radiation, moving the main lobe down lower.

Note that for an antenna height of λ/2 (Fig 23, bottom), the out-of-phase reflection from a perfectly conducting surface creates a null in the pattern at the zenith (90° elevation angle). Over real earth, however, a *filling in* of this null occurs because of ground losses that prevent perfect reflection of high-angle radiation.

At a 0° elevation angle, horizontally polarized antennas also demonstrate a null, because out-of-phase reflection cancels the direct wave. As the elevation angle departs from 0°, however, there is a slight filling-in effect so that over other-than-perfect earth, radiation at lower angles is enhanced compared to a vertical. A horizontal antenna will often outperform a vertical for low-angle DX work, particularly over lossy types of earth at the higher frequencies.

Reflection coefficients for vertically and horizontally polarized radiation differ considerably at most angles above ground, as can be seen by comparison of Figs 20 and 22. (Both sets of curves were plotted for the same ground constants and at the same frequency, so they may be compared directly.) This is because, as mentioned earlier, the image of a horizontally polarized antenna is out-of-phase with the antenna itself, and the image of a vertical antenna is in-phase with the actual radiator.

The result is that the phase shifts and reflection magnitudes vary greatly at different angles for horizontal and vertical polarization. The magnitude of the reflection coefficient for vertically polarized waves is greatest (near unity) at very low angles, and the phase angle is close to 180°. As mentioned earlier, this cancels nearly all radiation at very low angles. For the same range of angles, the magnitude of the reflection coefficient for horizontally polarized waves is also near unity, but the phase angle is near 0° for the specific conditions shown in Figs 20 and 22. This causes reinforcement of low-angle horizontally polarized waves. At some relatively high angle, the reflection coefficients for

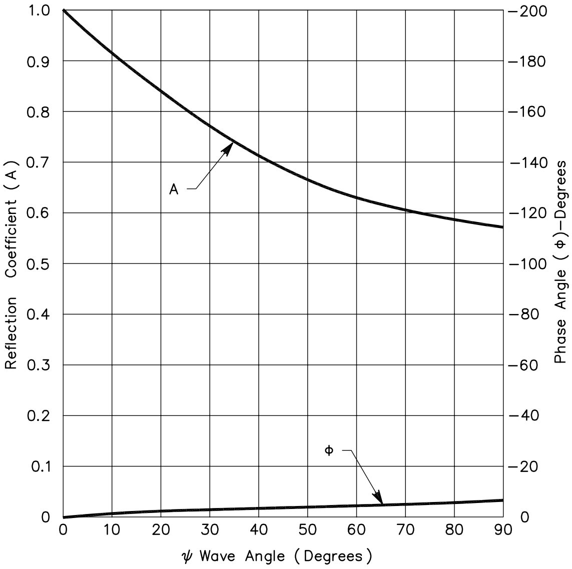

Fig 22—Reflection coefficient for horizontally polarized waves (magnitude A at angle ϕ), at 21 MHz over average earth (k = 13, G = 0.005 S/m).

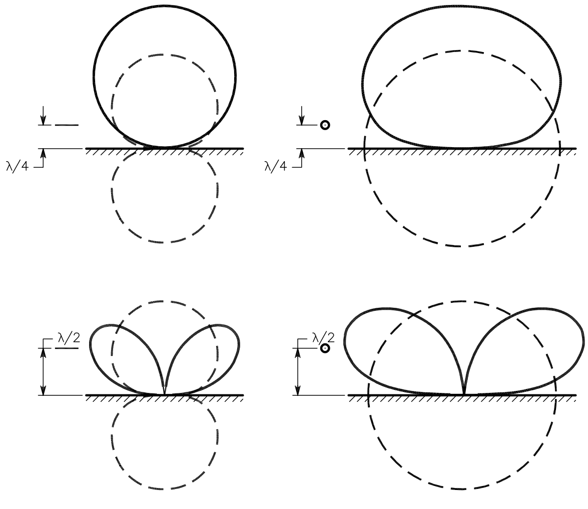

Fig 23—Effect of the ground on the radiation from a horizontal half-wave dipole antenna, for heights of one-fourth and one-half wavelength. Broken lines show what the pattern would be if there were no reflection from the ground (free space).

horizontally and vertically polarized waves are equal in magnitude and phase. At this angle (approximately 81° for the example case), the effect of ground reflection on vertically and horizontally polarized signals will be the same.

DEPTH OF RF CURRENT PENETRATION

When considering earth characteristics, questions about depth of RF current penetration often arise. For instance, if a given location consists of a 6-foot layer of soil overlying a highly resistive rock strata, which material dominates? The answer depends on the frequency, the soil and rock dielectric constants, and their respective conductivities. The following equation can be used to calculate the current density at any depth.

$$e^{-pd} = \frac{\text{Current Density at Depth d}}{\text{Current Density at Surface}} \qquad \text{(Eq 6)}$$

where

d = depth of penetration in cm

e = natural logarithm base (2.718)

$$p = \left(\frac{X \times B}{2} \times \left(\sqrt{1 + \frac{G^2 \times 10^{-4}}{B^2}} - 1 \right) \right)^{1/2}$$

$X = 0.008 \times \pi^2 \times f$

$B = 5.56 \times 10^{-7} \times k \times f$

k = dielectric constant of earth

f = frequency in MHz

G = conductivity of earth in S/m

After some manipulation of this equation, it can be used to calculate the depth at which the current density is some fraction of that at the surface. The depth at which the current density is 37% (1/*e*) of that at the surface (often referred to as *skin depth*) is the depth at which the current density would be zero if it were distributed uniformly instead of exponentially. (This 1/*e* factor appears in many physical situations. For instance, a capacitor charges to within 1/*e* of full charge within one RC time constant.) At this depth, since the power loss is proportional to the square of the current, approximately 91% of the total power loss has occurred, as has most of the phase shift, and current flow below this level is negligible.

Fig 24 shows the solutions to Eq 6 over the 1.8 to 30-MHz frequency range for various types of earth. For example, in very good earth, substantial RF currents flow down to about 3.3 feet at 14 MHz. This depth goes to 13 feet in average earth and as far as 40 feet in very poor earth. Thus, if the overlying soil is rich, moist loam, the underlying rock stratum is of little concern. However, if the soil is only average, the underlying rock may constitute a major consideration in determining the PBA and the depth to which the RF current will penetrate.

The depth in fresh water is about 156 feet and is nearly independent of frequency in the amateur bands below 30 MHz. In salt water, the depth is about seven inches at 1.8 MHz and decreases rather steadily to about two inches at 30 MHz. Dissolved minerals in moist earth increase its conductivity.

The depth-of-penetration curves in Fig 24 illustrate a noteworthy phenomenon. While skin effect confines RF current flow close to the surface of a conductor, the earth is so lossy that RF current penetrates to much greater depths than in most other media. The depth of RF current penetration is a function of frequency as well as earth type. Thus, the only cases in which most of the current flows near the surface are with very highly conductive media (such as salt water), and at frequencies above 30 MHz.

DIRECTIVE PATTERNS OVER REAL GROUND

As explained in Chapter 2, Antenna Fundamentals, because antenna radiation patterns are three-dimensional, it is helpful in understanding their operation to use a form of representation showing the elevation-plane directional characteristic for different heights. It is possible to show selected elevation-plane patterns oriented in various directions with respect to the antenna axis. In the case of the horizontal half-wave dipole, a plane running in a direction along the axis and another broadside to the antenna will give a good deal of information.

The effect of reflection from the ground can be expressed as a separate *pattern factor*, given in decibels. For any given elevation angle, adding this factor algebraically to the value for that angle from the free-space pattern for that antenna gives the resultant radiation value at that

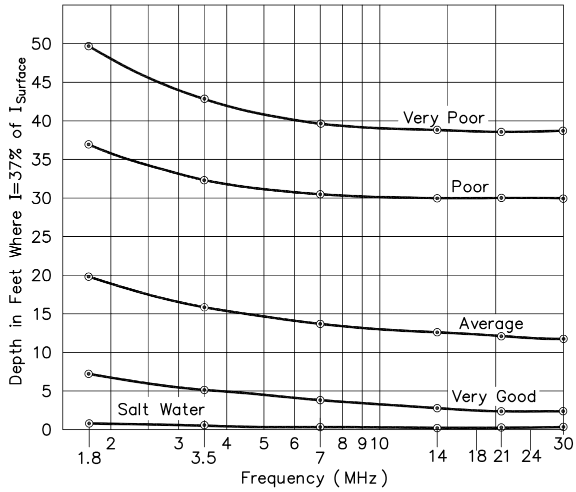

Fig 24—Depths at which the current density is 37% (1/*e*) of that at the surface for different qualities of earth over the 1.8- to 30-MHz frequency range. The depth for fresh water, not plotted, is 156 feet and almost independent of frequency below 30 MHz. See text and Table 2 for ground constants.

angle. The limiting conditions are those represented by the direct ray and the reflected ray being exactly in-phase and exactly out-of-phase, when both (assuming there are no ground losses) have equal amplitudes. Thus, the resultant field strength at a distant point may be either 6 dB greater than the free-space pattern (twice the field strength), or zero, in the limiting cases.

Horizontally Polarized Antennas

The way in which pattern factors vary with height for horizontal antennas over flat earth is shown graphically in the plots of **Fig 25**. The solid-line plots are based on perfectly conducting ground, while the shaded plots are based on typical real-earth conditions. These patterns apply to horizontal antennas of any length. While these graphs are, in fact, radiation patterns of horizontal single-wire antennas (dipoles) as viewed from the axis of the wire, it must be remembered that the plots merely represent pattern factors.

Fig 26 shows vertical-plane radiation patterns in the directions off the ends of a horizontal half-wave dipole for various antenna heights. These patterns are scaled so they may be compared directly to those for the appropriate heights in Fig 25. Note that the perfect-earth patterns in Figs 26A and 25B are the same as those in the upper part of Fig 23. Note also that the perfect-earth patterns of Figs 26B and 25D are the same as those in the lower section of Fig 23. The reduction in field strength off the ends of the wire at the lower angles, as compared with the broadside field strength, is quite apparent. It is also clear from Fig 26 that, at some heights, the high-angle radiation off the ends is nearly as great as the broadside radiation, making the antenna essentially an omnidirectional radiator.

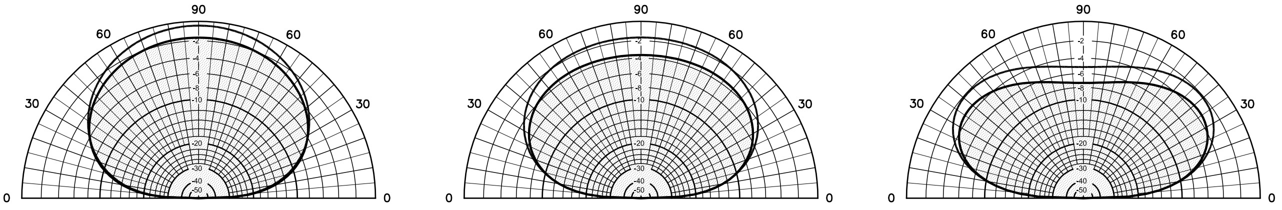

A — Antenna 1/8 λ high

B — Antenna 1/4 λ high

C — Antenna 3/8 λ high

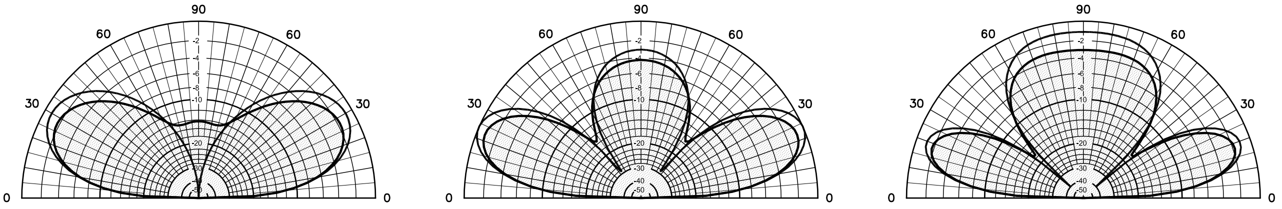

D — Antenna 1/2 λ high

E — Antenna 5/8 λ high

F — Antenna 3/4 λ high

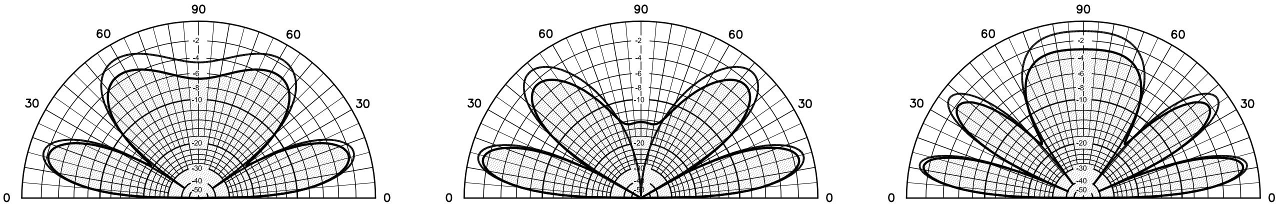

G — Antenna 7/8 λ high

H — Antenna 1 λ high

I — Antenna 1–1/4 λ high

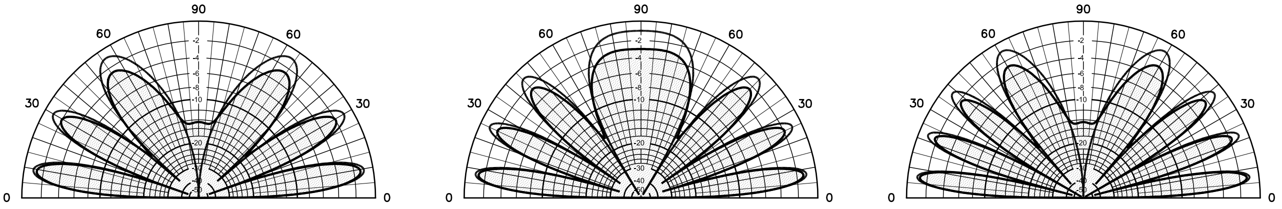

J — Antenna 1–1/2 λ high

K — Antenna 1–3/4 λ high

L — Antenna 2 λ high

Fig 25—Reflection factors for horizontal dipole antennas at various heights above flat ground. The solid-line curves are the perfect-earth patterns (broadside to the antenna wire); the shaded curves represent the effects of average earth (k = 13, G = 0.005 S/m) at 14 MHz. Add 7 dB to values shown for absolute gain in dBd referenced to dipole in free space, or 9.15 dB for gain in dBi. For example, peak gain over perfect earth at $^{5}/_{8}\lambda$ height is 7 dBd (or 9.15 dBi) at 25° elevation.

In vertical planes making some intermediate angle between 0° and 90° with the wire axis, the pattern will have a shape intermediate between the broadside and end-on patterns. By visualizing a smooth transition from the end-on pattern to the broadside pattern as the horizontal angle is varied from 0° to 90°, a fairly good mental picture of the actual solid pattern may be formed. An example is shown in **Fig 27**. At A, the elevation-plane pattern of a half-wave dipole at a height of λ/2 is shown through a plane 45° away from the favored direction of the antenna. At B and C, the pattern of the same antenna is shown at heights of 3λ/4 and 1λ (through the same 45° off-axis plane). These patterns are scaled so they may be compared directly with the broadside and end-on patterns for the same antenna (at the appropriate heights) in Figs 25 and 26.

The curves presented in **Fig 28** are useful for determining heights of horizontal antennas that give either maximum or minimum reinforcement at any desired wave angle. For instance, if you want to place an antenna at a height so that it will have a null at 30°, the antenna should be placed where a broken line crosses the 30° line on the horizontal scale. There are two heights (up to 2 λ) that will yield this null angle: 1 λ and 2 λ.

As a second example, you may want to have the ground reflection give maximum reinforcement of the direct ray from a horizontal antenna at a 20° elevation angle. The antenna height should be 0.75 λ. The same height will give a null at 42° and a second lobe at 90°.

Fig 28 is also useful for visualizing the vertical pattern of a horizontal antenna. For example, if an antenna is erected at 1.25 λ, it will have major lobes (solid-line crossings) at 12° and 37°, as well as at 90° (the zenith). The nulls in this pattern (dashed-line crossings) will appear at 24° and 53°.

The Y-axis in Fig 28 plots the wave angle versus the height in wavelength above flat ground on the X-axis. Fig 28 doesn't show the elevation angles required for actual

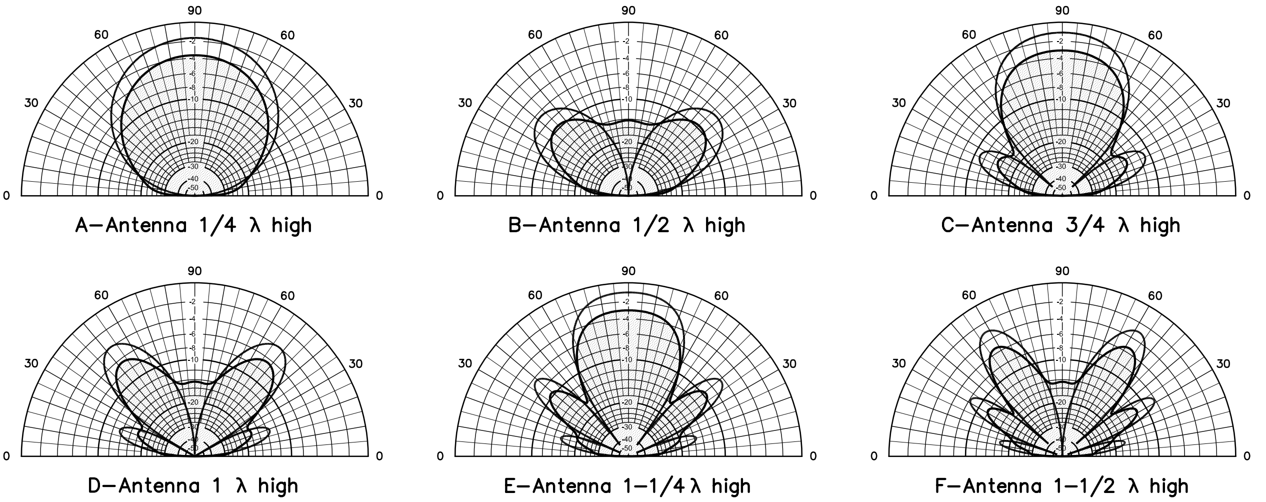

Fig 26—Vertical-plane radiation patterns of horizontal half-wave dipole antennas off the ends of the antenna wire. The solid-line curves are the flat, perfect-earth patterns, and the shaded curves represent the effects of average flat earth (k = 13, G = 0.005 S/m) at 14 MHz. The 0-dB reference in each plot corresponds to the peak of the main lobe in the favored direction of the antenna (the maximum gain). Add 7 dB to values shown for absolute gain in dBd referenced to dipole in free space, or 9.15 dB for gain in dBi.

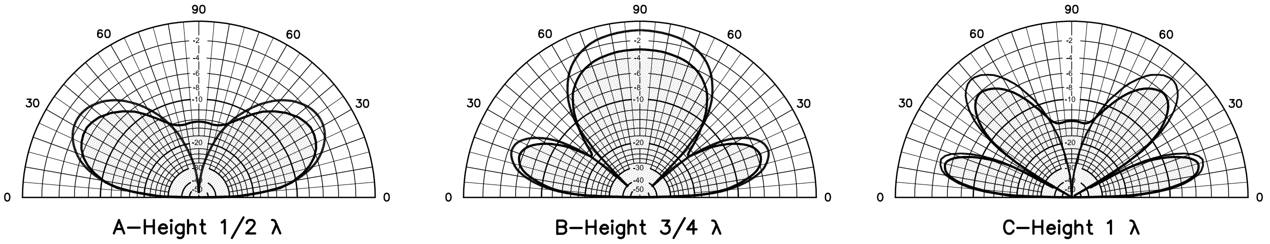

Fig 27—Vertical-plane radiation patterns of half-wave horizontal dipole antennas at 45° from the antenna wire over flat ground. The solid-line and shaded curves represent the same conditions as in Figs 25 and 26. These patterns are scaled so they may be compared directly with those of Figs 25 and 26.

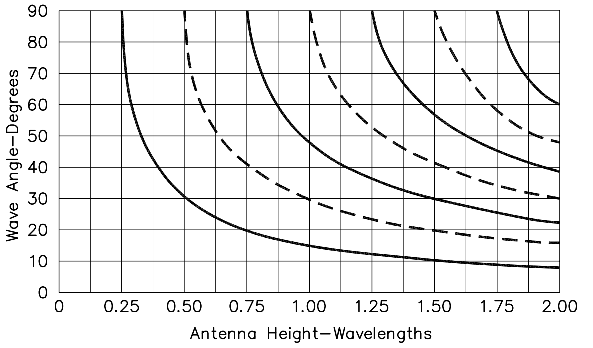

Fig 28—Angles at which nulls and maxima (factor = 6 dB) in the ground-reflection factor appear for antenna heights up to two wavelengths over flat ground. The solid lines are maxima, dashed lines nulls, for all horizontal antennas. See text for examples. Values may also be determined from the trigonometric relationship θ = arc sin (A/4h), where θ is the wave angle and h is the antenna height in wavelengths. For the first maximum, A has a value of 1; for the first null A has a value of 2, for the second maximum 3, for the second null 4, and so on.

communications to various target geographic locations of interest. Chapter 23, Radio Wave Propagation, and the CD-ROM in the back of this book give details about the range of angles required for target locations around the world. It is very useful to overlay plots of these angles together with the elevation pattern for horizontally polarized antennas at various heights above flat ground. This will be demonstrated in detail later in this chapter.

Vertically Polarized Antennas

In the case of a vertical λ/2 dipole or a ground-plane antenna, the horizontal directional pattern is simply a circle at any elevation angle (although the actual field strength will vary, at the different elevation angles, with the height above ground). Hence, one vertical pattern is sufficient to give complete information (for a given antenna height) about the antenna in any direction with respect to the wire. A series of such patterns for various heights is given in **Fig 29**. Rotating the plane pattern about the zenith axis of the graph forms the three-dimensional radiation pattern in each case.

The solid-line curves represent the radiation patterns of the λ/2 vertical dipole at different feed-point heights over perfectly conducting ground. The shaded curves in Fig 29 show the patterns produced by the same antennas at the same heights over average ground (G = 0.005 S/m, k = 13) at 14 MHz. The PBA in this case is 14.8°.

In short, far-field losses for vertically polarized antennas are highly dependent on the conductivity and dielectric constant of the earth around the antenna, extending far beyond the ends of any radials used to complete the ground return for the near field. Putting more radials out around the antenna may well decrease ground-return losses in the reactive near field for a vertical monopole, but will not increase radiation at low elevation launch angles in the far field, unless the radials can extend perhaps 100 wavelengths in all directions! Aside from moving to the fabled “salt water swamp on a high hill,” there is very little that someone can do to change the character of the ground that affects the far-field pattern of a real vertical. Classical texts

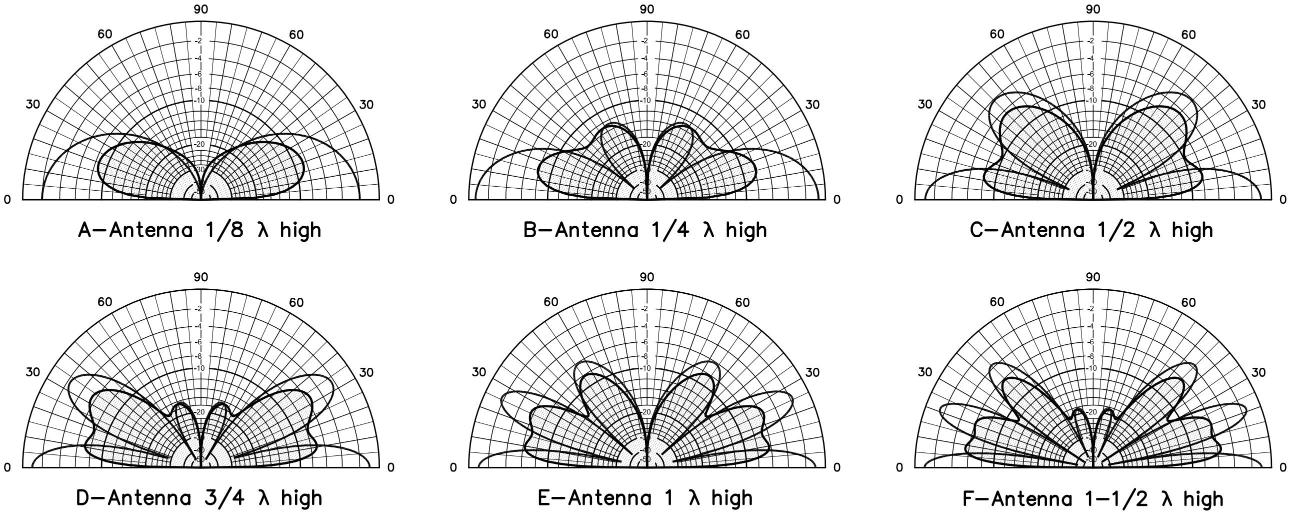

A—Antenna 1/8 λ high
B—Antenna 1/4 λ high
C—Antenna 1/2 λ high
D—Antenna 3/4 λ high
E—Antenna 1 λ high
F—Antenna 1-1/2 λ high

Fig 29—Vertical-plane radiation patterns of a ground-plane antenna above flat ground. The height is that of the ground plane, which consists of four radials in a horizontal plane. Solid lines are perfect-earth patterns; shaded curves show the effects of real earth. The patterns are scaled—that is, they may be directly compared to the solid-line ones for comparison of losses at any wave angle. These patterns were calculated for average ground (k = 13, G = 0.005 S/m) at 14 MHz. The PBA for these conditions is 14.8°. Add 6 dB to values shown for absolute gain in dBd over dipole in free space.

on verticals often show elevation patterns computed over an "infinitely wide, infinitely conducting ground plane." Real ground, with finite conductivity and less than perfect dielectric constant, can severely curtail the low-angle radiation at which verticals are supposed to excel.

While real verticals over real ground are not a surefire method to achieve low-angle radiation, cost versus performance and ease of installation are still attributes that can highly recommend verticals to knowledgeable builders. Practical installations for 160 and 80 meters rarely allow amateurs to put up horizontal antenna high enough to radiate effectively at low elevation angles. After all, a half-wave on 1.8 MHz is 273 feet high, and even at such a lofty height the peak radiation would be at a 30° elevation angle.

The Effects of Irregular Local Terrain in the Far Field

The following material is condensed and updated from an article by R. Dean Straw, N6BV, in July 1995 *QEX* magazine. *HFTA* (HF Terrain Assessment) and supporting data files are included on the CD-ROM at the back of this book. *HFTA* is the latest version of the *YT* program included with earlier editions of *The ARRL Antenna Book*.

Choosing a QTH for DXing

The subject of how to choose a QTH for working DX has fascinated hams since the beginning of amateur operations. No doubt, Marconi probably spent a lot of time wandering around Newfoundland looking for a great radio QTH before making the first transAtlantic transmission. Putting together a high-performance HF station for contesting or DXing has always followed some pretty simple rules. First, you need the perfect QTH, preferably on a rural mountaintop or at least on top of a hill. Even better yet, you need a mountaintop surrounded by seawater! Then, after you have found your dream QTH, you put up the biggest antennas you possibly can, on the highest towers you can afford. Then you work all sorts of DX—sunspots willing, of course.

The only trouble with this straightforward formula for success is that it doesn't always work. Hams fortunate enough to be located on mountain tops with really spectacular drop-offs often find that their highest antennas don't do very well, especially on 15 or 10 meters, but often even on 20 meters. When they compare their signals with nearby locals in the flatlands, they sometimes (but not always) come out on the losing end, especially when sunspot activity is high.

On the other hand, when the sunspots drop into the cellar, the high antennas on the mountaintop are usually the ones crunching the pileups—but again, not always. So, the really ambitious contest aficionados, the guys with lots of resources and infinite enthusiasm, have resorted to putting up antennas at all possible heights, on a multitude of towers.

There is a more scientific way to figure out where and how high to put your antennas to optimize your signal during all parts of the 11-year solar cycle. We advocate a *system approach* to HF station design, in which you need to know the following:

1. The range of elevation angles necessary to get from point A to point B
2. The elevation patterns for various types and configurations of antennas
3. The effect of local terrain on elevation patterns for horizontally polarized antennas.

WHAT IS THE RANGE OF ELEVATION ANGLES NEEDED?

Up until 1994, *The ARRL Antenna Book* contained only a limited amount of information about the elevation angles needed for communication throughout the world. In the 1974 edition, Table 1-1 in the Wave Propagation chapter was captioned: "Measured vertical angles of arrival of signals from England at receiving location in New Jersey."

What the caption didn't say was that Table 1-1 was derived from measurements made during 1934 by Bell Labs. The highest frequency data seemed pretty shaky, considering that 1934 was the low point of Cycle 17. Neither was this data applicable to any other path, other than the one from New Jersey to England. Nonetheless, many amateurs located throughout the US tried to use the sparse information in Table 1-1 as the only rational data they had for determining how high to mount their antennas. (If they lived on hills, they made estimates of the effect of the terrain, assuming that the hill was adequately represented by a long, unbroken slope. More on this later.)

In 1993 ARRL HQ embarked on a major project to tabulate the range of elevation angles from all regions of the US to important DX QTHs around the world. This was accomplished by running many thousands of computations using the *IONCAP* computer program. *IONCAP* has been under development for more than 25 years by various agencies of the US government and is considered the standard of comparison for propagation programs by many agencies, including the Voice of America, Radio Free Europe, and more than 100 foreign governments throughout the world. *IONCAP* is a real pain in the neck to use, but it is the standard of comparison.

The calculations were done for all levels of solar activity, for all months of the year, and for all 24 hours of the day. The results were gathered into some very large databases, from which special custom-written software extracted detailed statistics. The results appeared in summary form in Tables 4 through 13 printed in Chapter 23, Radio Wave Propagation, of the 17th Edition and in more detail on the diskette included with that book. (This book, the 20th Edition, contains even more statistical data, for more areas of the world, on the accompanying CD-ROM.)

Fig 30 shows the full range of elevation angles (represented as vertical bars) for the 20-meter path from New England (centered on Newington, Connecticut) to all of Europe. This is for all openings, in all months, over the entire 11-year solar cycle. The most likely elevation angle occurs at 5° for about 13% of the times when the 20-meter band is open to Europe from New England. From 4° to 6° the band is open a total of about 34% of the times the band is open. There is a secondary peak between 10° to 12°, occurring for a total of about 25% of the times the band is open.

Overlaid on Fig 30 along with the elevation-angle statistics are the elevation-plane responses for three different horizontally polarized Yagi beams, all over flat ground. The first is mounted 140 feet high, 2 λ in terms of wavelength. The second Yagi is mounted 70 feet high (at 1 λ) and the third is 35 feet (0.5 λ). The 140-foot high antenna has a deep null at 15°, but it also has the highest response (13.4 dBi) of the three at the statistical peak elevation angle of 5°. However, at 12°—where the band is open some 9% of the time—the 140-foot high Yagi is down 4 dB compared to the 70-foot antenna.

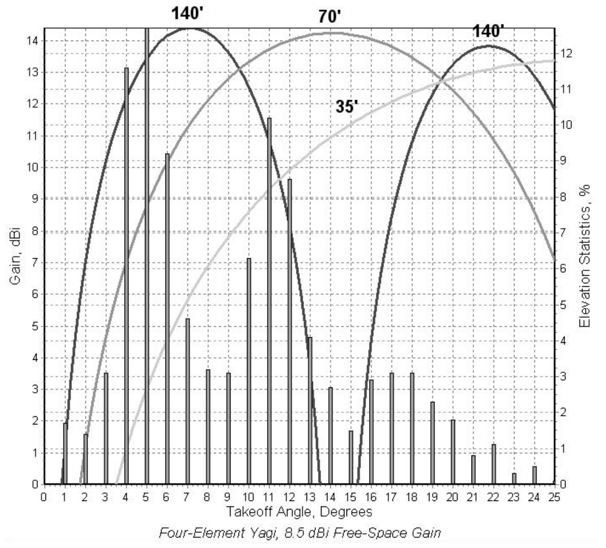

Fig 30—Graph showing 20-meter percentage of all openings from New England to Europe versus elevation angles, together with overlay of elevation patterns over flat ground for three 20-meter antenna systems. The most statistically likely angle at which the band will be open is 5°, although at any particular hour, day, month and year, the actual angle will likely be different. Note the deep null exhibited by the 140-foot high antenna centered at 14°.

The 70-foot high Yagi arguably covers the overall range best, since it has no disastrous nulls in the 1° to 25° range, where most of the action is occurring on 20 meters. At 5°, however, its response is only 8.8 dBi, 4.6 dB down from the 140-foot high antenna at that angle. The 35-foot antenna peaks above 26° in elevation angle, and is down some 10.4 dB compared to the 140-foot antenna at 5°. Obviously, no single antenna covers the complete range of elevation angles needed.

Note that the highest Yagi has a strong *second lobe* peaking at 22°. Let's say that you could select between two antennas, one at 140 and one at 70 feet, and that the incoming angle for a particular distant station is 22°. You might be fooled into thinking that the incoming angle is around 6°, favoring the first peak of the higher antenna, when in truth the angle is relatively high. The 70-foot antenna's response would be lower at 22° than the higher one, but only because the 140-foot antenna is operating on its second lobe. (What would clinch a determination of the correct incoming angle—6° or 22°—would be the response of the 35-foot high Yagi, which would be close to its peak at 22°, while it would be very far down at 6°.)

Now, we must emphasize that these elevation angles are *statistical entities*— in other words, just because 5° is the "statistically most likely angle" for the 20-meter path from New England to Europe doesn't mean that the band will be open at 11° at any particular hour, on a particular day, in a particular month, in any particular year. In fact, however, experience agrees with the *IONCAP* computations: the 20-meter path to Europe usually opens at a low angle in the New England morning hours, rising to about 11° during the afternoon, when the signals remain strongest throughout the afternoon until the evening in New England.

What would happen if we were to feed all three Yagi at 140, 70 and 35 feet in-phase as a stack? **Fig 31** shows this situation, along with a more highly optimized stack at 120, 80 and 40 feet that better covers the overall range of elevation angles from Connecticut to Europe.

Now see **Fig 32**, which uses the same 120/80/40-foot stack of 20-meter antennas as in Fig 31, but this time from Seattle, Washington, to Europe. For comparison, the response of a single 4-element Yagi at 100 feet over flat ground is also shown in Fig 32. Just because 5º is the statistically most prevalent angle (occurring some 13% of the time) from Seattle to Europe on 20 meters, this doesn't mean that the actual angle *at any particular moment in time* might not be 10º, or even 2º. The statistics for W7 to Europe say that 5º is the most likely angle, but 20-meter signals from Europe arrive at angles ranging from 1º to 18º. Note that this range of angles is quite a bit less than from W1 to Europe, which is much closer geographically to Europe than is the Pacific Northwest coast of the US. If you design an antenna system to cover all possible angles needed to talk to Europe from Seattle (or from Seattle to Europe) on 20 meters, you would need to cover the full range from 1º to 18º equally well.

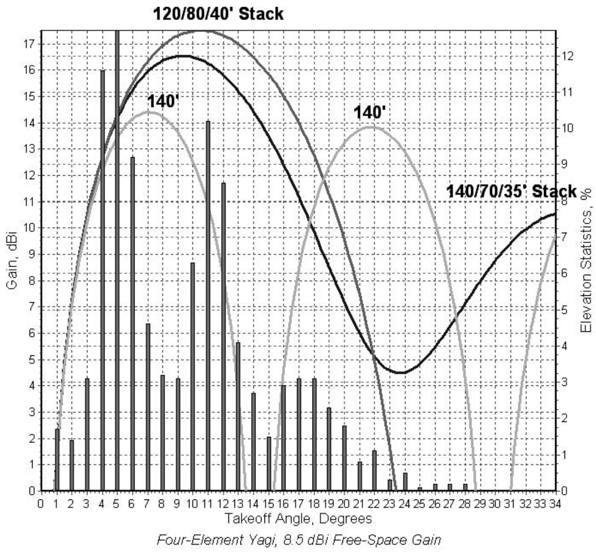

Fig 31—Graph showing results of stacking antennas at different heights on the same tower to cover a wider range of elevation angles, in this case for the path from Connecticut (W1) to all of Europe on 20 meters. The optimized stack at 120/80/40 feet covers the needed range of elevation angles better than the stack at 140/70/35 feet or the single Yagi at 140 feet.

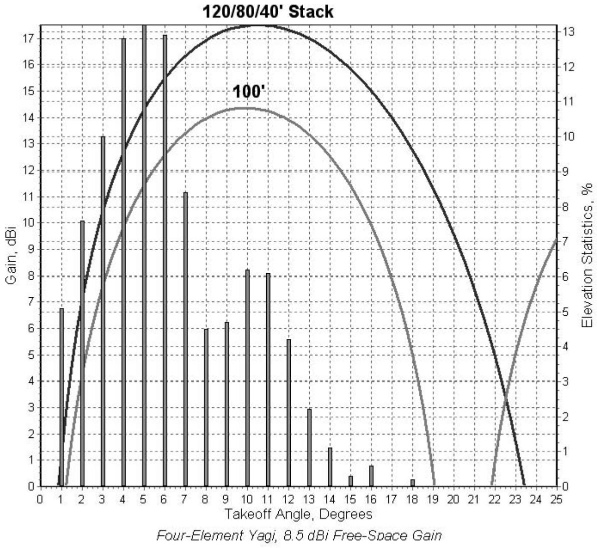

Fig 32—Graph showing 20-meter percentage of all openings, this time from Seattle, WA, to Europe, together with an overlay of elevation patterns over flat ground for two 20-meter antenna systems. The statistically most likely angle on this path is 5°, occurring about 13% of the time when the band is actually open. Higher antennas predominate on this low-angle path.

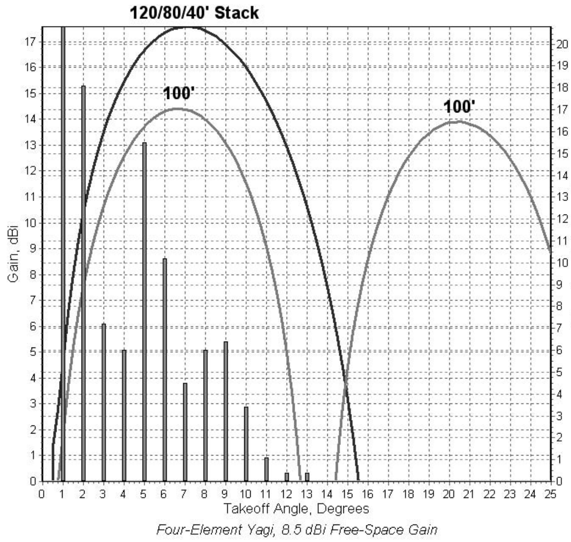

Fig 33—Graph showing 15-meter percentage of all openings from Chicago to Southern Africa, together with overlay of elevation patterns over flat ground for two 15-meter antenna systems. On this long-distance, low-angle path, higher antennas are again most effective.

Similarly, if you wish to cover the full range of elevation angles from Chicago to Southern Africa on 15 meters, you would need to cover 1º to 13º, even though the most statistically likely signals arrive at 1º, for 21% of the time when that the band is open for that path. See **Fig 33**.

It is important to recognize that Figs 30 through 33 are for flat ground. When the antennas are mounted over irregular local terrain, things get much more complicated. First, however, we'll discuss general-purpose antenna modeling programs as they try to model real terrain.

DRAWBACKS OF COMPUTER MODELS FOR ANTENNAS OVER REAL TERRAIN

Modern general-purpose antenna modeling programs such as *NEC* or *MININEC* (or their commercially upgraded equivalents, such as *NEC-Win Plus*, *EZNEC* and *EZNEC ARRL*) can accurately model almost any type of antenna commonly used by radio amateurs. In addition, there are specialized programs specifically designed to model Yagis efficiently, such as *YO* or *YW* (Yagi for Windows, bundled on the CD-ROM with this book) or *YagiMax*. These programs however are all unable to model antennas accurately over anything other than *purely flat ground*.

While both *NEC* and *MININEC* can simulate irregular ground terrain, they do so in a decidedly crude manner, employing step-like concentric rings of height around an antenna. The documentation for *NEC* and *MININEC* both clearly state that diffraction off these steps is not modeled.

Common experience among serious modelers is that the warnings in the manuals are worth heeding.

Although you can analyze and even optimize antenna designs using free-space or flat-earth ground models, it is *diffraction* that makes the real world a very, very complicated place. This should be clarified—diffraction is hard, even tortuous, to analyze properly, but it makes analysis of real world results far more believable than a flat-world reflection model does.

RAY-TRACING OVER UNEVEN LOCAL TERRAIN

The Raytracing Technique

First, let's look at a simple raytracing procedure involving only horizontally polarized reflections, with no diffractions. From a specified height on the tower, an antenna shoots "rays" (just as though they were bullets) in 0.25° increments from +35° above the horizon to –35° below the horizon. Each ray is traced over the foreground terrain to see if it hits the ground at any point on its travels in the direction of interest. If it does hit the ground, the ray is reflected following the classical *law of reflection*. That is, the outgoing angle equals the incoming angle, reflected through the normal to the slope of the surface. Once the rays exit into the ionosphere, the individual contributions are vector-summed to create the overall far-field elevation pattern.

The next step in terrain modeling involves adding *diffractions* as well as reflections. At the Dayton antenna forum in 1994, Jim Breakall, WA3FET, gave a fascinating and tantalizing lecture on the effect of foreground terrain. Later Breakall, Dick Adler, K3CXZ, Joel Young and a group of other researchers published an extremely interesting paper entitled "The Modeling and Measurement of HF Antenna Skywave Radiation Patterns in Irregular Terrain" in July 1994 *IEEE Transactions on Antennas and Propagation*. They described in rather general terms the modifications they made to the *NEC-BSC* program. They showed how the addition of a ray-tracing reflection and diffraction model to the simplistic stair-stepped reflection model in regular *NEC* gave far more realistic results. For validation, they compared actual pattern measurements made on a site in Utah (with an overflying helicopter) to computed patterns made using the modified *NEC* software. However, because the US Navy funded this work the software remained for a long time a military secret.

Thumbnail History of the Uniform Theory of Diffraction

It is instructive to look briefly at the history of how *Geometric Optics* (GO) evolved (and still continues to evolve) into the *Uniform Theory of Diffraction* (UTD). The following is summarized from the historical overview in one book found to be particularly useful and comprehensive on the subject of UTD: *Introduction to the Uniform Geometrical Theory of Diffraction*, by McNamara, Pistorius, and Malherbe.

Many years before the time of Christ, the ancient Greeks studied optics. Euclid is credited with deriving the law of reflection about 300 BC. Other Greeks, such as Ptolemy, were also fascinated with optical phenomena. In the 1600s, a Dutchman named Snell finally figured out the law of refraction, resulting in *Snell's law*. By the early 1800s, the basic world of classical optics was pretty well described from a mathematic point of view, based on the work of a number of individuals.

As its name implies, classical geometric optical theory deals strictly with geometric shapes. Of course, the importance of geometry in optics shouldn't be minimized—after all, we wouldn't have eyeglasses without geometric optics. Mathematical analysis of shapes utilizes a methodology that traces the paths of straight-line *rays* of light. (Note that the paths of rays can also be likened to the straight-line paths of particles.) In classical geometric optics, however, there is no mention of three important quantities: phase, intensity and polarization. Indeed, without phase, intensity or polarization, there is no way to deal properly with the phenomenon of *interference*, or its cousin, *diffraction*. These phenomena require theories that deal with *waves* rather than rays.

Wave theory has also been around for a long time, although not as long as geometry. Workers like Hooke and Grimaldi had recorded their observations of interference and diffraction in the mid 1600s. Huygens had used elements of wave theory in the late 1600s to help explain refraction. By the late 1800s, the work of Lord Rayleigh, Sommerfeld, Fresnel, Maxwell and many others led to the full mathematic characterization of all electromagnetic phenomena, light included.

Unfortunately, ray theory doesn't work for many problems, at least ray theory in the classical optical form. The real world is a lot more jagged, pointy and fuzzy in shape than can be described in a totally rigorous mathematic fashion. Some properties of the real world are most easily explained on the micro level using electrons and protons as conceptual objects, while other macro phenomena (like resonance, for example) are more easily explained in terms of waves. To get a handle on a typical real-world physical situation, a combination of classical ray theory and wave theory was needed.

The breakthrough in the combination of classical geometric optics and wave concepts came from J. B. Keller of Bell Labs in 1953, although he published his work in the early 1960s. In the very simplest of terms, Keller introduced the notion that shooting a ray at a diffraction *wedge* causes wave interference at the tip, with an infinite number of diffracted waves emanating from the diffraction point. Each diffracted wave can be considered to be a point source radiator at the place of generation, the diffraction point. Thereafter, the paths of individual waves can be traced as though they were individual classical optic rays again. What Keller came up with was a reasonable mathematical description of what happens at the tip of the diffraction wedge.

Fig 34 is a picture of a simple diffraction wedge, with an incoming ray launched at an angle of α_r, referenced to the horizon, impinging on it. The diffraction wedge here is considered to be perfectly conducting, and hence impenetrable by the ray. The wedge generates an infinite number of diffracted waves, going in all directions not blocked by the wedge itself. The amplitudes and phases of the diffracted waves are determined by the interaction at the wedge tip, and this in turn is governed by the various angles associated with the wedge. Shown in Fig 34 are the included angle a of the wedge, the angle ϕ' of the incoming ray (referenced to the incoming surface of the wedge), and the observed angle ϕ of one of the outgoing diffracted waves, also referenced to the wedge surface.

The so-called *shadow boundaries* are also shown in Fig 34. The Reflection-Shadow Boundary (RSB) is the angle beyond which no further reflections can take place for a given incoming angle. The Incident-Shadow Boundary (ISB) is that angle beyond which the wedge's face blocks any incident rays from illuminating the observation point.

Keller derived the amplitude and phase terms by comparing the classical Geometric Optics (GO) solution with the exact mathematical solution calculated by Sommerfeld for a particular case where the boundary conditions were well known—an infinitely long, perfectly conducting wedge illuminated by a plane wave. Simply speaking, whatever was left over had to be diffraction terms. Keller combined these diffraction terms with GO terms to yield the total field everywhere.

Keller's new theory became known as the *Geometric Theory of Diffraction* (abbreviated henceforth as GTD). The beauty of GTD was that in the regions where classical GO predicted zero fields, the GTD "filled in the blanks," so to speak. For example, see **Fig 35**, showing the terrain for a hypothetical case, where a 60-foot high 4-element 15-meter Yagi illuminates a wide, perfectly flat piece of ground. A 10-foot high rock has been placed 400 feet away from the tower base in the direction of outgoing rays. **Fig 36** shows the elevation pattern predicted using reflection-only GO techniques. Due to blockage of the direct wave (A) trying to shoot past the 10-foot high rock, and due to blockage of (B) reflections from the flat ground in front of the rock by the rock, there is a *hole* in the smooth elevation pattern.

Now, doesn't it defy common sense to imagine that a single 10-foot high rock will really have such an effect on a 15-meter signal? Keller's GTD took diffraction effects into account to show that waves do indeed sneak past and over the rock to fill in the pattern. The whole GTD scheme is very clever indeed.

However, GTD wasn't perfect. Keller's GTD predicts some big spikes in the pattern, even though the overall shape of the elevation pattern is much closer to reality than a simple GO reflection analysis would indicate. The region right at the RSB and ISB shadow boundaries is where problems are found. The GO terms go to zero at these points because of blockage by the wedge, while Keller's diffraction terms tend to go to infinity at these very spots. In mathematical terms this is referred to as a *caustic problem*. Nevertheless, despite these nasty problems at the ISB and RSB, the GTD provided a remarkably better solution to diffraction problems than did classical GO.

In the early 1970s, a group at Ohio State University under R. G. Kouyoumjian and P. H. Pathak did some pivotal work to resolve this caustic problem, introducing what amounts to a clever *fudge factor* to compensate for the tendency of the diffraction terms at the shadow boundaries to go to infinity. They introduced what is known as a *transition function*, using a form of Fresnel integral. Most importantly, the Ohio State researchers also created several FORTRAN computer programs to compute the amplitude and phase of diffraction components. Now computer hackers could get to work!

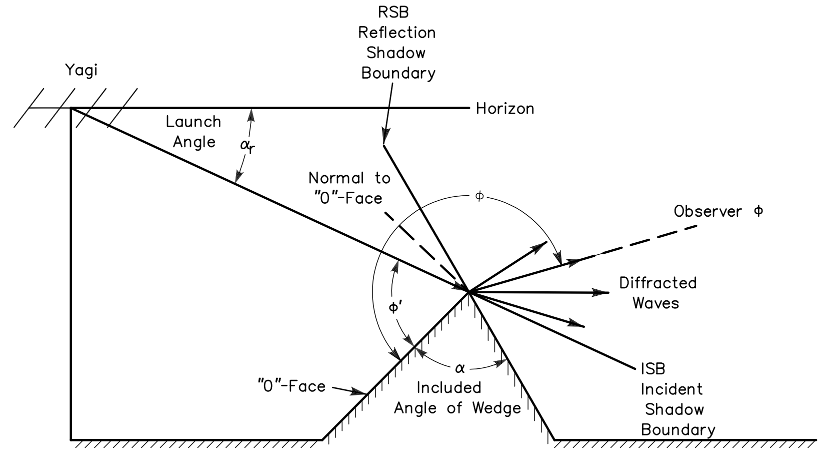

Fig 34—Diagram showing diffraction mechanism of ray launched at angle α_r below the horizon at a diffraction wedge, whose included angle is α. Referenced to the incident face (the *o-face* as it is called in UTD terminology), the incoming angle is ϕ' (phi prime). The wedge creates an infinite number of diffracted waves. Shown is one whose angle referenced to the o-face is ϕ, the so-called *observation angle* in UTD terminology.

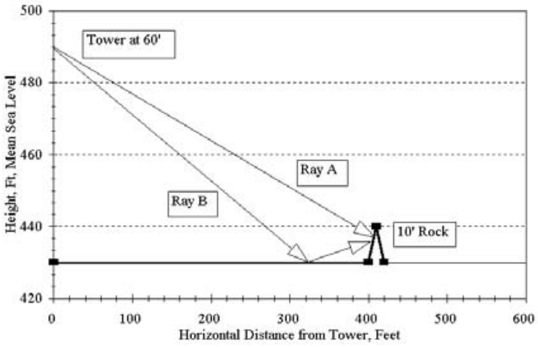

Fig 35—Hypothetical terrain exhibiting so-called "10-foot rock effect." The terrain is flat from the tower base out to 400 feet, where a 10-foot high rock is placed. Note that this forms a diffraction wedge, but that it also blocks direct waves trying to shoot through it to the flat surface beyond, as shown by Ray A. Ray B reflects off the flat surface before it reaches the 10-foot rock, but it is blocked by the rock from proceeding further. A simple Geometric Optics (GO) analysis of this terrain without taking diffraction into account will result in the elevation response shown in Fig 36.

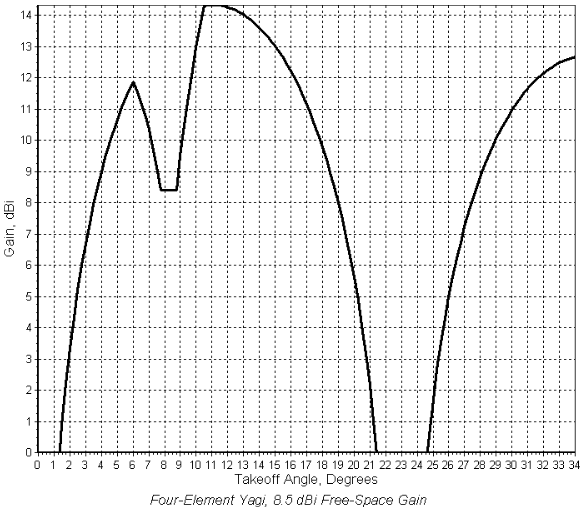

Fig 36—Elevation response for rays launched at terrain in Fig 35 from a height of 60 feet using a 4-element Yagi. This was computed using a simple Geometrical Optics (GO) reflection-only analysis. Note the *hole* in the response between 6° to 10° in elevation. It is not reasonable for a 10-foot high rock to create such a disturbance at 21 MHz!

The program that finally resulted is called *HFTA*, standing for "HF Terrain Assessment." (The DOS version of *HFTA* was known as *YT*, standing for "Yagi Terrain.") As the name suggests, *HFTA* analyzes the effect of local terrain on HF propagation through the ionosphere. It is designed for horizontally polarized Yagis, although it will model the effects of a simple flattop dipole also. The accurate appraisal of the effect of terrain on vertically polarized signals is a far more complex problem than for horizontally polarized waves, and *HFTA* doesn't do verticals.

SIMULATION OF REALITY—SOME SIMPLE EXAMPLES FIRST

We want to focus first on some simple results, to show that the computations do make some sense by presenting some simulations over simple terrains. We've already described the "10-foot rock at 400 feet" situation, and showed where a simple GO reflection analysis is inadequate to the task without taking diffraction effects into account.

Now look at the simple case shown in **Fig 37**, where a very long, continuous downslope from the tower base is shown. Note that the scales used for the X and Y axes are different: the Y-axis changes 300 feet in height (from 800 to 1100 feet), while the X-axis goes from 0 to 3000 feet. This exaggerates the apparent steepness of the downwards slope, which is actually a rather gentle slope, at $\tan^{-1}(1000 - 850) / (3000 - 0) = -2.86°$. In other words, the terrain falls 150 feet in height over a range of 3000 feet from the base of the tower.

Fig 38 shows the computed elevation response for this terrain profile, for a 4-element horizontally polarized Yagi on a 60-foot tower. The response is compared to that of an identical Yagi placed 60 feet above flat ground. Compared to the "flatland" antenna, the hilltop antenna has an elevation response shifted over by almost 3° towards the lower elevation angles. In fact, this shift is directly due to the –2.86° slope of the hill. Reflections off the slope are tilted by the slope. In this situation there is a single diffraction at the bottom of the gentle slope at 3000 feet, where the program assumes that the terrain becomes flat.

Look at **Fig 39**, which shows another simple terrain profile, called a "Hill-Valley" scenario. Here, the 60-foot high tower stands on the edge of a gentle hill overlooking a long valley. Once again the slope of the hill is exaggerated by the different X and Y-axes. **Fig 40** shows the computed elevation response at 21.2 MHz for a 4-element Yagi on a 60-foot high tower at the edge of the slope.

Once again, the pattern is overlaid with that of an identical 60-foot-high Yagi over flat ground. Compared to the flatland antenna, the hilltop antenna's response above 9° in elevation is shifted by almost 3° towards the lower elevation angles. Again, this is due to reflections off the downward slope. From 1° to 9°, the hilltop pattern is enhanced even more compared to the flatland antenna, this time by diffraction occurring at the bottom of the hill.

Now let's see what happens when there is a hill ahead in the direction of interest. **Fig 41** depicts such a situation, labeled "Hill-Ahead." Here, at a height of 400 feet above mean sea level, the land is flat in front of the tower, out to a distance 500 feet, where the hill begins. The hill then rises 100 feet over the range 500 to 1000 feet away

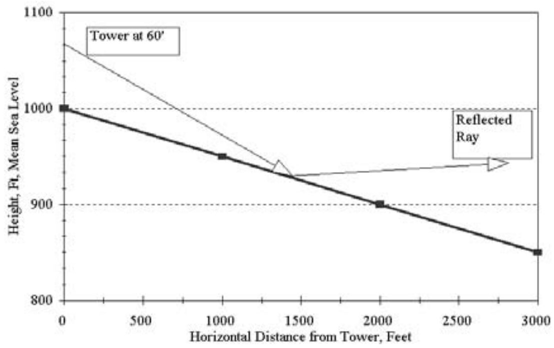

Fig 37—A long, gentle downwards-sloping terrain. This terrain has no explicit diffraction points and can be analyzed using simple GO reflection techniques.

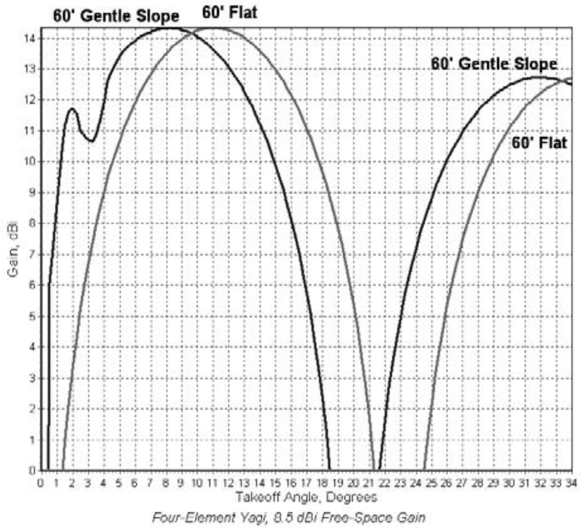

Fig 38—Elevation response for terrain shown in Fig 37, using a 4-element 15-meter Yagi, 60-foot high. Note that the shape of the response is essentially shifted towards the left, towards lower elevation angles, by the angle of the sloping ground. For reference, the response for an identical Yagi placed over flat ground is also shown.

from the tower base. After that, the terrain is a plateau, at a constant 500 feet elevation.

Fig 42 shows the computed elevation pattern for a 4-element 21-MHz Yagi 60-feet high on the tower, compared again with an overlay for an identical 60-foot high antenna over flat ground. The hill blocks low-angle waves directly radiated from the antenna from 0° to 2.3°. In addition, waves that would normally be reflected from the ground, and that would normally add in phase from about 2.3° to 12°, are blocked by the hill also. Thus the signal at 8° is down almost 5 dB from the signal over flat ground, all due to the effect of the hill. Diffracted waves start kicking in once the direct wave rises enough above the horizon to illuminate the top edge of the hill. These diffracted waves tend to augment elevation angles above about 12°, which reflected waves can't reach.

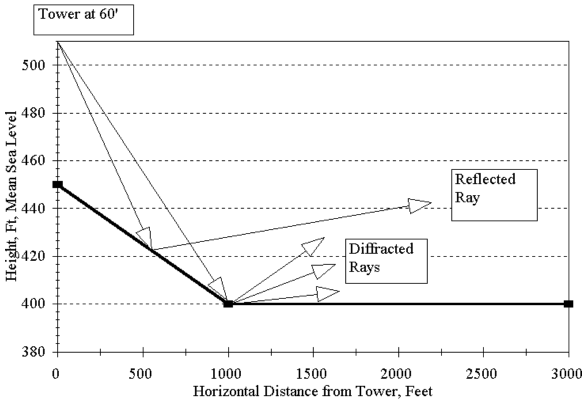

Fig 39—"Hill-Valley" terrain, with reflected and diffracted rays.

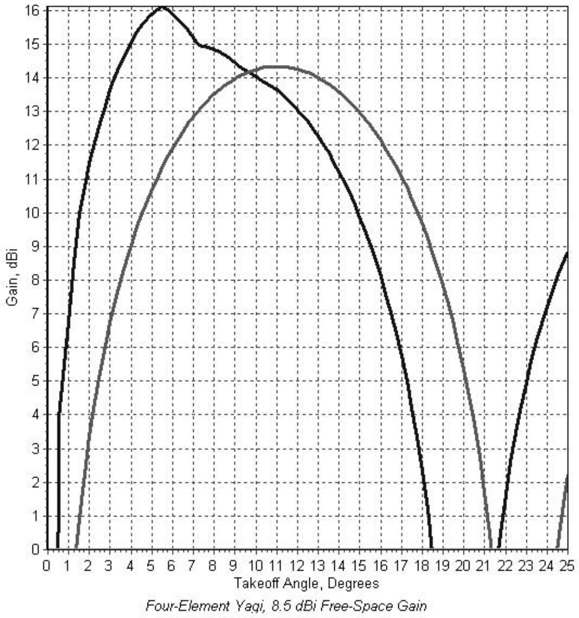

Fig 40—Elevation response computed by *HFTA* program for single 4-element 15-meter Yagi at 60 feet above "Hill-Valley" terrain shown in Fig 39. Note that the slope has caused the response in general to be shifted towards lower elevation angles. At 5° elevation, the diffraction components add up to increase the gain slightly above the amount a GO-only analysis would indicate.

Is there is any hope for someone in such a lousy QTH for DXing? **Fig 43** shows the elevation response for a truly heroic solution. This involves a stack of four 4element Yagis, mounted at 120, 90, 60 and 30 feet on the tower. Now, the total gain at low angles is just about comparable to that from a single 4-element Yagi mounted over flat ground. Where there's a ham, there is a way!

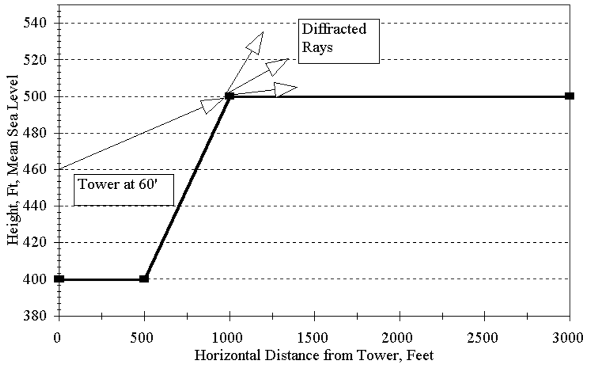

Fig 41—"Hill-Ahead" terrain, shown with diffracted rays created by illumination of the edge of the plateau at the top of the hill.

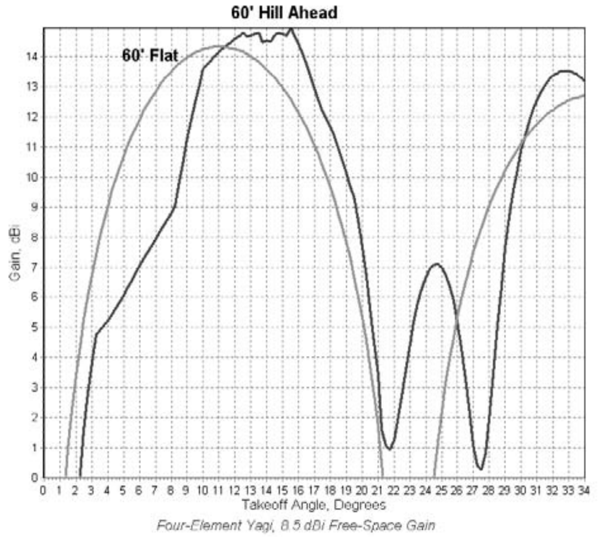

Fig 42—Elevation response computed by *HFTA* for "Hill-Ahead" terrain shown in Fig 41. Now the hill blocks direct rays and also precludes possibility of any constructive reflections. Above 10°, diffraction components add up together with direct rays to create the response shown.

At 5° elevation, four diffraction components add up (there are zero reflection components) to achieve the far-field pattern. This seems reasonable, because each of the four antennas is illuminating the diffraction point separately and we know that none of the four antennas can *see over* the hill directly to produce a reflection at a low launch angle.

At an elevation angle of 5°, 15-meter signals arrive from Europe from New England about 13% of the total time when the band is actually open. We can look at this another way. For about two-thirds of the times when the band is open on this path, the incoming angle is between 3° to 12°. For about one-third of the time, signals arrive above 10°, where the "heroic" four-stack is really beginning to come into its own.

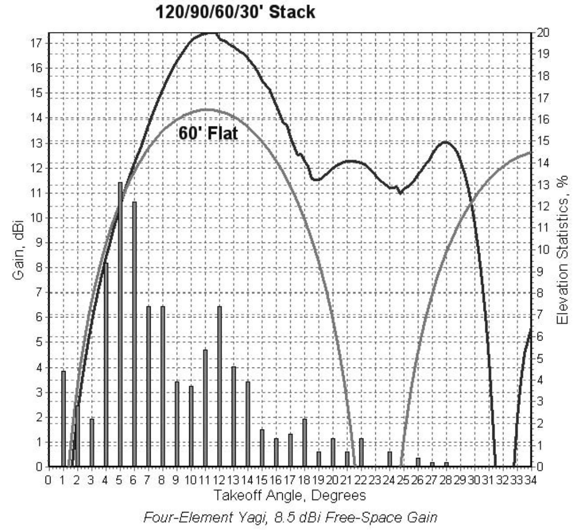

Fig 43—Elevation response of "heroic effort" to surmount the difficulties imposed by hill in Fig 41. This effort involves a stack of four 4-element Yagis in a stack starting at 120 feet and spaced at 30-foot increments on the tower. The response is roughly equivalent to a single 4-element Yagi at 60 feet above flat ground, hence the characterization as being a "heroic effort." The elevation-angle statistics from New England to Europe are overlaid on the graph for reference.

A More Complex Terrain

The results for simple terrains look reasonable; let's try a more complicated real-world situation. **Fig 44** shows the terrain from the New Hampshire N6BV/1 QTH towards Japan. The terrain was complex, with 52 different points *HFTA* identifies as diffraction points. **Fig 45** shows a labeled *HFTA* output for three different types of antennas on 20 meters: a stack at 120 and 60 feet, the 120-foot antenna by itself, and then a 120/60-foot stack over flat ground, for reference. The elevation-angle statistics for New England to Japan are overlaid on the graph also, making for a very complicated looking picture—it is a *lot* easier to decipher the lines on the color CRT, by the way than on a black-and-white printer.

Comparison of the same 120/60-foot stacks over irregular terrain and flat ground is useful to show where the terrain itself is affecting the elevation response. The flatland stack has more gain in the region of 3° to 7° than the same stack over the N6BV/1 local terrain towards Japan. On the other hand, the N6BV/1 local terrain boosts signals in the range of 8° to about 12°. This demonstrates the conservation of energy—you may gain a stronger signal at certain elevation angles, but you will lose gain at others. In this case, the N6BV/1 station always felt "weak" towards Japan on 20 meters, because the dominant angles are low.

Examination of the detailed data output from *HFTA* shows that at an elevation angle of 5°, there are 6159

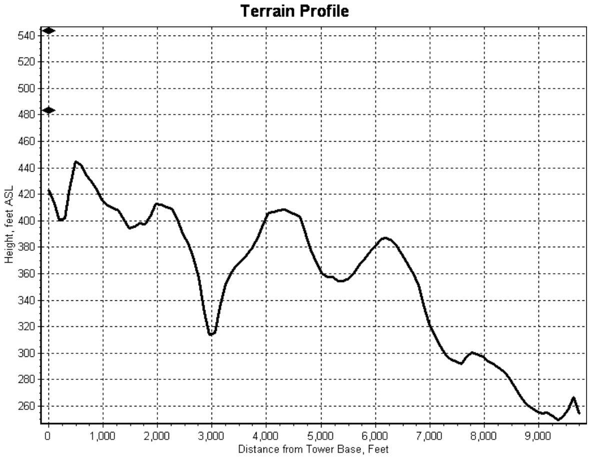

Fig 44—Terrain of N6BV in Windham, NH, towards Japan. *HFTA* identifies 52 different points where diffraction can occur.

diffraction components. There are many, many signals bouncing around off the terrain on their trip to Japan! Note that because of blockage of some parts of the terrain, the 60-foot high Yagi cannot illuminate all the diffraction points, while the higher 120-foot Yagi is able to see these diffraction points.

It is fascinating to reflect on the thought that received signals coming down from the ionosphere to the receiver are having encounters with the terrain, but from the opposite direction. It's not surprising, given these kinds of interactions, that transmitting and receiving might not be totally reciprocal.

The 120/60-foot stack in Fig 45 achieves its peak gain of 17.3 dBi at 11° elevation, where it is about 3 dB stronger than the single Yagi at 120 feet. It maintains this 3-dB advantage over most of the range of incoming signals from Japan. This difference in performance between the stack and each antenna by itself was observed many times on the air. Much of the time when comparisons are being made, however, the small differences in signal are difficult to measure meaningfully, especially when the QSB varies signals by 20 dB or so during a typical QSO. It should be noted that the stack usually exhibited less fading compared to each antenna by itself.

USING *HFTA*

Manually Generating a Terrain Profile

The *HFTA* program uses two distinct algorithms to generate the far-field elevation pattern. The first is a simple reflection-only Geometric Optics (GO) algorithm. The second is the diffraction algorithm using the Uniform Theory of Diffraction (UTD). These algorithms work with a digitized representation of the terrain profile for a single azimuthal direction—for example, towards Japan or towards Europe.

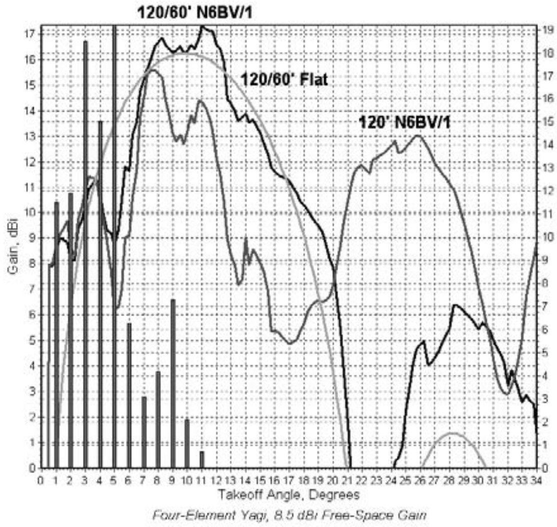

Fig 45—Elevation responses computed by *HFTA* for N6BV/1 terrain shown in Fig 44, for a stack of two 4-element 20-meter Yagis at 120 and 60 feet, together with the response for a single Yagi at 120 feet and a 120/60-foot stack over flat ground for reference. The response due to many diffraction and reflection components is quite complicated!

You can generate a terrain file manually using a topographic map and a ruler or a pair of dividers. The HFTA.PDF file (accessed by clicking on the **Help** button) and on the accompanying CD-ROM gives complete instructions on how to create a terrain file manually (or automatically). The manual process is simple enough in concept. Mark on your US Geological Survey 7.5-minute map the exact location of your tower. You will find 7.5-minute maps available from some local sources, such as large hardware stores, but the main contact point is the U.S. Geological Survey, Denver, CO 80225 or Reston, VA 22092. Call 1-800-MAPS-USA. Ask for the folder describing the topographic maps available for your geographic area. Many countries outside the USA have topographic charts also. Most are calibrated in meters. To use these with *HFTA*, you will have to convert meters to feet by multiplying meters by 3.28 or else inserting a single line at the very beginning of the disk file, saying "meters" for *HFTA* to recognize meters automatically.

Mark off a pencil line from the tower base, in the azimuthal direction of interest, perhaps 45° from New England to Europe, or 335° to Japan. Then measure the distance from the tower base to each height contour crossed by the pencil line. Enter the data at each distance/height into an ASCII computer file, whose filename extension is "PRO," standing for *profile*.

Fig 46 shows a portion of the USGS paper map for the N6BV QTH in Windham, NH, along with lines scribed in several directions towards various parts of Europe and the Far East. Note that the elevation heights of the intermediate contour lines are labeled manually in pencil in order to make sense of things. It is very easy to get confused unless you do this!

The terrain model used by *HFTA* assumes that the terrain is represented by flat *plates* connecting the elevation points in the *.PRO file with straight lines. The model is two dimensional, meaning that range and elevation are the only data for a particular azimuth. In effect, *HFTA* assumes that the width of a terrain plate is wide relative to its length. Obviously, the world is three-dimensional. If your shot in a particular direction involves aiming your Yagi down a canyon with steep walls, then it's pretty likely that your actual elevation pattern will be different than what *HFTA* tells you. The signals must careen horizontally from wall to wall, in addition to being affected by the height changes of the terrain. *HFTA* isn't designed to do canyons.

To get a true 3-D picture of the full effects of terrain, a terrain model would have to show azimuth, along with range and elevation, point-by-point for about two miles in every direction around the base of the tower. After you go through the pain of creating a profile for a single azimuth, you'll appreciate the immensity of the process if you were you try to create a full 360° 3D profile manually.

Terrain Data from the Internet

At one time digitized terrain data commonly available from the Internet didn't have sufficient resolution to be accurate enough for *HFTA*. Nowadays, the complete, accurate set of USGS topographic 7.5-minute maps are available at no cost on the Internet. You can use a program called *MicroDEM*, written by Professor Peter Guth at the US Naval Academy, to quickly and easily produce terrain data files suitable for *HFTA* from topographic data files. Dr Guth and the US Naval Academy have graciously allowed ARRL to include the *MicroDEM* program on the CD-ROM accompanying this book. It should be noted that besides automatically creating terrain profiles for *HFTA*, *MicroDEM* is a full-featured mapping program on its own.

Instructions for using *MicroDEM* are in the Help file for *HFTA* (HFTA.PDF), which you can access from the *HFTA* main window by clicking on the **Help** button. **Fig 47** shows a screen capture of the *MicroDEM* program for the N6BV/1 location in New Hampshire for an azimuth of 45° into Europe. The black/white rendering of the screen capture doesn't do justice to the same information in color. The computed terrain profile is plotted in the window at the right of Fig 47 and the data file is shown in the inset window at the top right.

Using *MicroDEM* and on-line USGS topographic map data, you can also automatically create up to 360 terrain profiles with 1° spacing of azimuths in a few seconds. (Specifying a 1° spacing is really overkill; most operator choose to create 72 profiles with 5° spacing.) On a topographic DEM (digital elevation model) map that covers the geographic

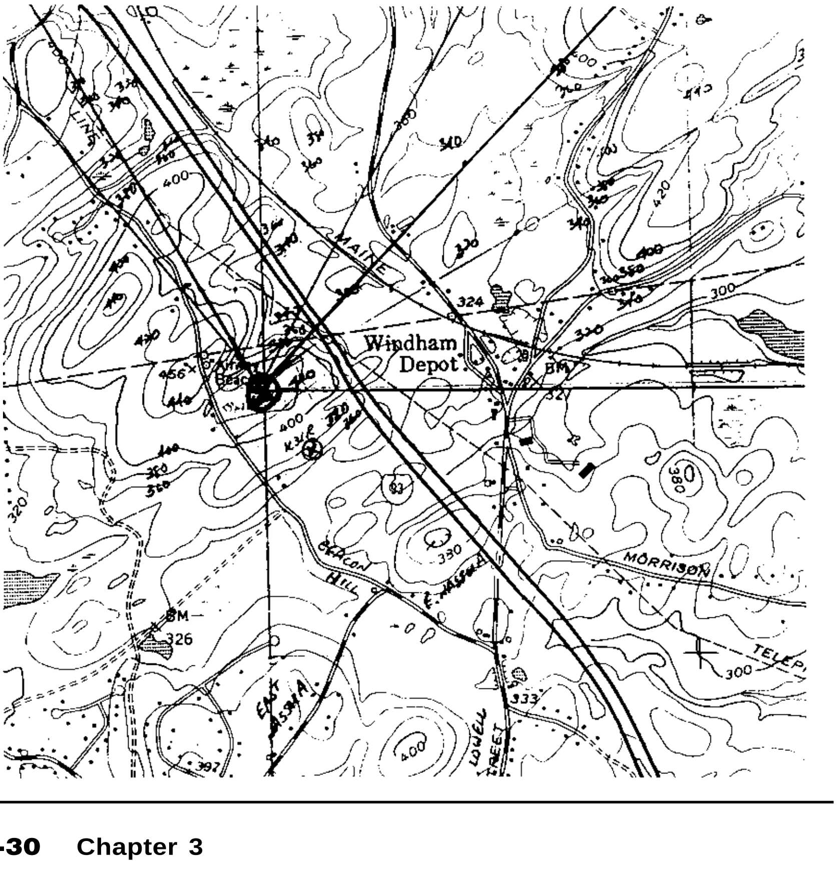

Fig 46—A portion of USGS 7.5 minute topographic map, showing N6BV/1 QTH, together with marks in direction of Europe and Japan from tower base. Note that the elevations contours were marked by hand to help eliminate confusion. This required a magnifying glass and a steady hand!

area of interest, you simply specify the latitude and longitude of a tower's location—found using a GPS receiver—and then ask *MicroDEM* for a **Viewshed**. See the *HFTA* **Help** file for the details.

Compare this automated several-second *MicroDEM* process to creating manual profiles on a paper topographic map—It can take up to an hour of meticulous measurements to manually create a single terrain profile!

Algorithm for Ray-Tracing the Terrain

Once a terrain profile is created, there are a number of mechanisms that *HFTA* takes into account as a ray travels over that terrain:

1. Classical ray reflection, with Fresnel ground coefficients.
2. Direct diffraction, where a diffraction point is illuminated directly by an antenna, with no intervening terrain features blocking the direct illumination.
3. When a diffracted ray is subsequently reflected off the terrain.
4. When a reflected ray encounters a diffraction point and causes another series of diffracted rays to be generated.
5. When a diffracted ray hits another diffraction point, generating another whole series of diffractions.

Certain unusual, bowl-shaped terrain profiles, with sheer vertical faces, can conceivably cause signals to reflect or diffract in a backwards direction, only to be reflected back again in the forward direction by the sheer-walled terrain to the rear. *HFTA* does not accommodate these interactions, mainly because to do so would increase the computation time too much. It only evaluates terrain in the forward direction along one azimuth of interest.

Fig 48 shows a portion of an *HFTA* screen capture in the direction towards Europe from the N6BV/1 location in New Hampshire on 21.2 MHz. It compares the results for a 90/60/30-foot stack of TH7DX tribanders to the same stack over flat land, and to a single antenna at 70 feet over flat ground. The 70-foot single antenna represents a pretty typical station on 15 meters. The terrain produces excellent gain at lower elevation angles compared to the same stack over flat ground. The stack is very close to or superior to the single 70-foot high Yagi at all useful elevation angles. Terrain can indeed exhibit a profound effect on the launch of signals into the ionosphere—for good or for bad.

HFTA's Internal Antenna Model

The operator selects the antenna used inside *HFTA* to be anything from a dipole to an 8-element Yagi. The default assumes a simple cosine-squared mathematic response, equivalent to a 4-element Yagi in free space. *HFTA* traces rays only in the forward direction from the tower along the azimuth of interest. This keeps the algorithms reasonably simple and saves computing time.

HFTA considers each antenna in a stack as a separate *point source*. The simulation begins to fall apart if a traveling wave type of antenna like a rhombic is used, particularly if the terrain changes under the antenna—that is, the ground is not flat under the entire antenna. For a typical Yagi, even a long-boom one, the point-source assumption is reasonable. The internal antenna model also assumes that the Yagi is horizontally polarized. *HFTA* does not do vertically polarized antennas, as discussed previously. The documentation for *HFTA* also cautions the user to work with practical spacings between stacked Yagis—0.5 λ or more

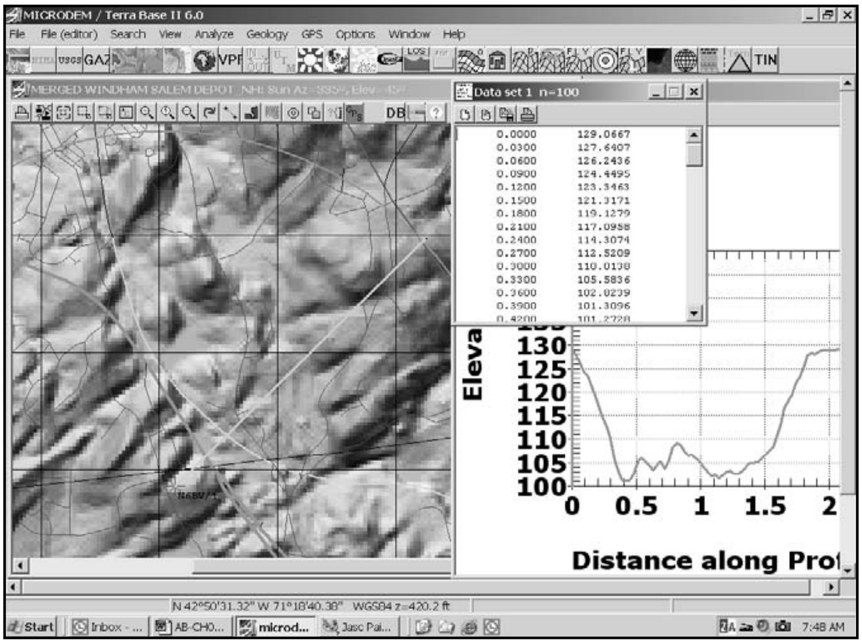

```
0.0000      129.0667
0.0300      127.6407
0.0600      126.2436
0.0900      124.4495
0.1200      123.3463
0.1500      121.3171
0.1800      119.1279
0.2100      117.0958
0.2400      114.3074
0.2700      112.5209
0.3000      110.0138
0.3300      105.5836
0.3600      102.0239
0.3900      101.3096
0.4200      101.2728
```

```
N 42°50'31.32" W 71°18'40.38"   WGS84 z=420.2 ft
```

Fig 47—A screen-capture of the *MicroDEM* program, showing the topographic map for the same terrain shown in Fig 46, together with the computed terrain profile along an azimuth of 45° on the path towards Europe from the N6BV/1 location in Windham, NH.

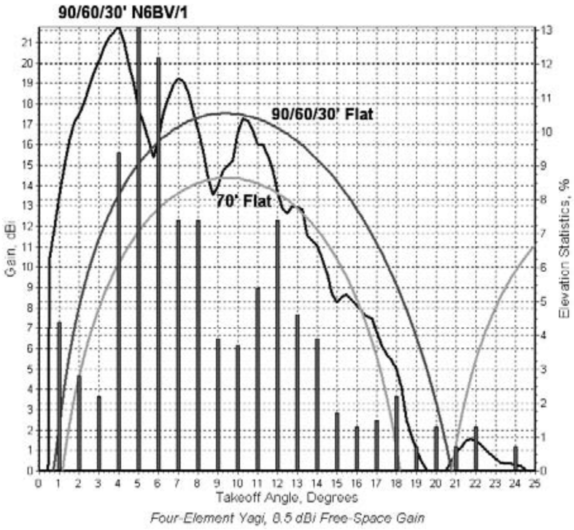

Fig 48—The 21-MHz elevation response for a stack of three TH7DX Yagis mounted on a single tower at 90/60/30 feet, at the N6BV/1 QTH for a 45° azimuth towards Europe. The terrain focuses the energy at low elevation angles compared to the same stack over flat ground. This illustrates once again the conservation of energy—Energy squeezed down into low elevation angles is stolen from other, higher, angles.

because *HFTA* doesn't explicitly model mutual coupling between Yagis in a stack.

HFTA compares well with the measurements for the horizontal antennas described earlier by Jim Breakall, WA3FET, using a helicopter in Utah. Breakall's measurements were done with a 15-foot high horizontal dipole.

More Details About *HFTA*

Frequency Coverage

HFTA can be used on frequencies higher than the HF bands, although the graphical resolution is only 0.25°. The patterns above about 100 MHz thus look rather grainy. The UTD is a *high-frequency-asymptotic* solution, so in theory the results become more realistic as the frequency is raised. Keep in mind too that *HFTA* is designed to model launch angles for skywave propagation modes, including E- and F-layer, and even Sporadic-E. Since by definition the ionospheric launch angles include only those above the horizon, direct line-of-sight UHF modes involving negative launch angles are not considered in *HFTA*.

See the HFTA.PDF documentation file for further details on the operation of *HFTA*. This file, as well as sample terrain profiles for some *big-gun* stations, is located on the CD-ROM accompanying this book.

BIBLIOGRAPHY

Source material and more extended discussion of the topics covered in this chapter can be found in the references listed below and in the textbooks listed at the end of Chapter 2, Antenna Fundamentals.

B. Boothe, "The Minooka Special," *QST*, Dec 1974.

Brown, Lewis and Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proc. IRE* , Jun 1937.

G. Brown, "The Phase And Magnitude Of Earth Currents Near Radio Transmitting Antennas," *Proc. IRE*, Feb 1935, pp 168-182.

R. Collin and F. Zucker, *Antenna Theory*, Chap 23 by J. Wait, Inter-University Electronics Series (New York: McGraw-Hill, 1969), Vol 7, pp 414-424.

T. Hullick, W9QQ, "A Two-Element Vertical Parasitic Array For 75 Meters," *QST*, Dec 1995, pp 38-41.

R. Jones, "A 7-MHz Vertical Parasitic Array," *QST*, Nov 1973.

T. Larsen, "The E-Field and H-Field Losses Around Antennas With a Radial Ground Wire System," *Journal of Research of the National Bureau of Standards*, D. Radio Propagation, Vol 66D, No. 2, Mar-Apr 1962, pp 189-204.

D. A. McNamara, C. W. I. Pistorius, J. A. G. Malherbe, *Introduction to the Geometrical Theory of Diffraction* (Norwood, MA: Artech House, 1994).

C. J. Michaels, "Some Reflections on Vertical Antennas," *QST*, Jul 1987.

C. J. Michaels, "Horizontal Antennas and the Compound Reflection Coefficient," *The ARRL Antenna Compendium, Vol 3* (Newington: ARRL, 1992).

R. Severns, N6LF, "Verticals, Ground Systems and Some History," *QST*, Jul 2000, pp 38-44.

J. Sevick, "The Ground-Image Vertical Antenna," *QST*, Jul 1971.

J. Sevick, "The W2FMI 20-Meter Vertical Beam," *QST*, Jun 1972.

J. Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, Mar 1973.

J. Sevick, "A High Performance 20-, 40- and 80-Meter Vertical System," *QST*, Dec 1973.

J. Sevick, "The Constant-Impedance Trap Vertical," *QST*, Mar 1974.

J. Stanley, "Optimum Ground Systems for Vertical Antennas," *QST*, Dec 1976.

F. E. Terman, *Radio Engineers' Handbook*, 1st ed. (New York, London: McGraw-Hill Book Co, 1943).

Radio Broadcast Ground Systems, available from Smith Electronics, Inc, 8200 Snowville Rd, Cleveland, OH 44141.

The last major study that appeared in the Amateur literature on the subject of local terrain as it affects DX appeared in four *QST* "How's DX" columns, by Clarke Greene, K1JX, from October 1980 to January 1981. Greene's work was an update of a landmark Sep 1966 *QST* article entitled "Station Design for DX," by Paul Rockwell, W3AFM. The long-range profiles of several prominent, indeed legendary, stations in Rockwell's article are fascinating: W3CRA, W4KFC and W6AM.

Chapter 4

Antenna Modeling & System Planning

OVERVIEW: ANTENNA ANALYSIS BY COMPUTER

As pointed out in Chapter 3, The Effects of Ground, irregular local terrain can have a profound effect on the launch of HF signals into the ionosphere. A *system approach* is needed to create a scientifically planned station. We pointed out in Chapter 3 that antenna modeling programs do not generally take into account the effects of irregular terrain, and by "irregular" we mean any sort of ground that is not flat. Most modeling programs based on *NEC-2* or *MININEC* do model reflections, but they do not model diffractions.

On the other hand, while a ray-tracing program like *HFTA* (HF Terrain Assessment) does take into account diffraction, it doesn't explicitly factor in the mutual impedance between an antenna and the ground. Instead, *HFTA* makes the basic assumption that the antenna is mounted sufficiently high above ground so that the mutual impedance between an antenna and the ground is minimal.

In this chapter we'll look at modeling the antennas themselves on the PC. We'll evaluate some typical antennas over flat ground and also in free space. Once characterized—or even optimized for certain characteristics—these antennas can then be analyzed over real terrain using *HFTA* and the other tools discussed in Chapter 3.

A Short History of Antenna Modeling

With the proliferation of personal computers since the early 1980s, amateurs and professionals alike have made significant strides in computerized antenna system analysis. It is now possible for the amateur with a relatively inexpensive computer to evaluate even complicated antenna systems. Amateurs can obtain a keener grasp of the operation of antenna systems—a subject that has been a great mystery to many in the past. We might add that modern computing tools allow hams to debunk overblown claims made about certain antennas.

The most commonly encountered programs for antenna analysis are those derived from a program developed at US government laboratories called *NEC*, short for "Numerical Electromagnetics Code." *NEC* uses a so-called *Method of Moments* algorithm. This intriguing name derives from a mathematical convention dealing with how "momentous" the accumulated error becomes when certain simplifying assumptions are made about the current distribution along an antenna wire. If you want to delve into details about the method of moments, John Kraus, W8JK, has an excellent chapter in his book *Antennas*, 2nd edition. See also the article "Programs for Antenna Analysis by the Method of Moments," by Bob Haviland, W4MB, in *The ARRL Antenna Compendium, Vol 4*.

The mathematics behind the method-of-moments algorithm are pretty formidable, but the basic principle is simple. An antenna is broken down into a number of straight-line wire *segments*, and the field resulting from the RF current in each segment is evaluated by itself and also with respect to other mutually coupled segments. Finally, the field from each contributing segment is vector-summed to yield the total field, which can be computed for any elevation or azimuth angle desired. The effects of flat-earth ground reflections, including the effect of ground conductivity and dielectric constant, may be evaluated as well.

In the early 1980s, *MININEC* was written in BASIC for use on personal computers. Because of limitations in memory and speed typical of personal computers of the time, several simplifying assumptions were necessary in *MININEC* and these limited potential accuracy. Perhaps the most significant limitation was that perfect ground was assumed to be directly under the antenna, even though the radiation pattern in the far field did take into account real ground parameters. This meant that antennas modeled closer than approximately 0.2 λ over ground sometimes gave erroneous impedances and inflated gains, especially for horizontal polarization. Despite some limitations, *MININEC* represented a remarkable leap forward in analytical capability. See Roy Lewallen's (W7EL) "*MININEC*—the Other Edge of the Sword" in Feb 1991 *QST* for an excellent treatment on pitfalls when using *MININEC*.

Because source code was made available when *MININEC* was released to the public, a number of programmers produced some very capable commercial versions for the amateur market, many incorporating exciting graphics showing antenna patterns in 2D or 3D. These programs also simplify the creation of models for popular antenna types, and several come with libraries of sample antennas.

By the end of the 1980s, the speed and capabilities

Commercial Implementations of *MININEC* and *NEC-2* Programs

Ever since the source code for *NEC-2* and *MININEC* came into in the public domain, enterprising programmers have been upgrading, extending and improving these programs. There are a number of "freeware" versions available nowadays, and there are also a variety of commercial implementations.

This sidebar deals only with the most popular commercial versions, programs that many hams use. You should keep in mind that whatever program you choose will require an investment in learning time, if not in dollars. Your time is valuable, of course, and so is the ability to swap modeling files you create with other modelers. Other peoples' modeling files, particularly when you are just starting out, are a great way to learn how the "experts" do their modeling. For example, there are archives of *EZNEC/ELNEC* files available on the Internet, since this popular modeling program has been around for a number of years. (*ELNEC* is the DOS-only, *MININEC*-core predecessor of EZNEC.)

The following table summarizes the main features and the pricing as of 2006 for some popular commercial antenna modeling programs. The programs that use the *NEC-4* core require separate licenses from Lawrence Livermore National Laboratories.

Commercial Implementations of *MININEC* and *NEC 2* programs

Name	***EZNEC 4.0* *(+ ver.)***	***EZNEC-M Pro***	***EZNEC/4***	***NEC-Win Plus +***	***NEC-Win Pro***	***GNEC***	***Antenna Model***
Manufacturer	Roy Lewallen	Roy Lewallen	Roy Lewallen	Nittany Scientific	Nittany Scientific	Nittany Scientific	Teri Software
Core	*NEC-2*	*NEC-2*	*NEC-4*	*NEC-2*	*NEC-2*	*NEC-2/NEC-4*	*MININEC*
Operating System	Windows 32-bit	Windows 32-bit	Windows 32-bit	Windows 32-bit	Windows 32-bit	Windows 32-bit	Windows 32-bit
Number Segments	500 (1500, + ver.)	20,000	20,000	10,000	10,000	80,000	Limited by memory
NEC-Card Inputs	No	Yes	Yes	Yes	Yes	Yes	No
Other Input	ASCII (NEC, + ver.)	ASCII, NEC	ASCII, NEC	CAD *.DXF	CAD *.DXF	CAD *.DXF	No
Wires by Equation	No	No	No	Yes	Yes	Yes	Yes
Source Setting	By %	By %	By %	By %	By %	By %	By %
Source Type	Current/ Voltage/Split	Current/ Voltage/Split	Current /Voltage/Split	Current/ Voltage/Split	All types	All types	Current/Voltage
R + *j* X Loads	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RLC Loads	Series, Parallel, Trap	Series, Parallel, Trap	Series, Parallel, Trap	Series, Parallel	Series, Parallel	Series, Parallel	Series, Parallel
True Trap Loads	Yes	Yes	Yes	No	No	No	No
Laplace Loads	Yes	Yes	Yes	Yes	Yes	Yes	No
Conductivity Table	Yes *	Yes *	Yes *	Yes	Yes	Yes	Yes
Average Gain Test	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Transmission Lines	Yes	Yes	Yes	Yes	Yes	Yes	No
View Geometry	Excellent	Excellent	Excellent	Good	Good	Good	Very Good
Geometry Checking	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Easy Height Change	Yes	Yes	Yes	No	No	No	No
Polar Plots	ARRL, linear-dB Az/El, Circ. (+ ver.)	ARRL, linear-dB Az/El, Circ.	ARRL, linear-dB Az/El, Circ.	ARRL, linear-dB Az/El Patterns	ARRL, linear-dB Az/El Patterns	ARRL, linear-dB Az/El Patterns	ARRL, Linear -dB Az/El Patterns
Rectangular Plots	SWR	SWR	SWR	SWR, Zin	SWR, Zin, Az/El, Near/Far Plots, Currents	SWR, Zin, Az/El, Near/Far Plots, Currents	Gain, SWR, F/B, F/R, Rin, Xin
Operating Speed	Fast	Fast	Fast	Very Fast	Very Fast	Very Fast	Slow
Smith Chart	Data for Ext. Smith program	Data for Ext. Smith program	Data for Ext. Smithprogram	No	Yes	Yes	Yes
Near/Far Field Tables	Both	Both	Both	Far	Both	Both	Both
Ground Wave Analysis	No	Yes	Yes	No	Yes	Yes	No
Pricing	$89 Web; $99 CD-ROM, $139 (+ ver.)	$450	$600; must have *NEC-4* license	$150	$425	$795	$85

* Wire conductivity is the same for all wires.

Excellent, Very Good, Good ratings done by *Antenna Book* editor.

of personal computers had advanced to the point where PC versions of *NEC* became practical, and several versions are now available to amateurs. The most recent public-domain version is *NEC-2* and this is the computational core that we'll use as an example throughout this chapter.

Like *MININEC*, *NEC-2* is a general-purpose modeling package and it can be difficult to use and relatively slow in operation for certain specialized antenna forms. Thus, custom commercial software has been created for more user-friendly and speedier analysis of specific antenna varieties, mainly Yagi arrays. See Chapter 11, HF Yagi Arrays. Also see the sidebar, "Commercial Implementations of *MININEC* and *NEC-2* Programs."

For this edition of *The ARRL Antenna Book*, Roy Lewallen, W7EL, has graciously provided a special version of his *EZNEC 3.0* program, called *EZNEC-ARRL*. This version works with the specific antenna models also bundled on the CD-ROM. Please note that this ARRL-specific version of *EZNEC* is limited to a maximum of 20 segments (we'll explain segments later) for all models except for the special ones included on this CD-ROM. You can find information on how to purchase the full-fledged version of *EZNEC* in the **Help** section of the *EZNEC-ARRL* program.

The following material on antenna modeling is by necessity a summary, since entire books have been written on this subject. Serious modelers would be well-advised to enroll in the online Antenna Modeling course, part of the ARRL Certification and Continuing Education series. L. B. Cebik, W4RNL created the ARRL Antenna Modeling course and it contains a great deal of information, tips and techniques concerning modeling by computer. See: **http://www.arrl.org/news/stories/2002/02/06/2/** for more information. We also strongly recommend that you read the Help files in *EZNEC-ARRL*. There is a wealth of practical information on the finer points of antenna modeling there.

THE BASICS OF ANTENNA MODELING

This chapter will discuss the following antenna-modeling topics for an *NEC-2*-based modeling software, using *EZNEC-ARRL* as an example:

- Program outputs
- Wire geometry
- Segmentation, warnings and limitations
- Source (feed point) placement
- Environment, including ground types and frequency
- Loads and transmission lines
- Testing the adequacy of a model

PROGRAM OUTPUTS

Instruction manuals for software programs traditionally start out describing in detail the input data needed by the program. They then demonstrate the output data the program can generate. We feel it is instructive, however, to turn things around and start out with a brief overview of the output from a typical antenna-modeling program.

We'll look at the output from public-domain *NEC-2*. Next, we'll look at the output information available from commercial adaptations of *NEC-2*, using *EZNEC-ARRL* provided by W7EL. After this brief overview of the output data, we'll look in detail at the input data needed to make a modeling program work. In the following discussions it will be very instructive if you to bring up *EZNEC-ARRL* on your computer and open the specific modeling files used in each example. [From now on in this chapter we'll refer merely to *EZNEC* rather than *EZNEC 3.0*, the official name or *EZNEC-ARRL*, a specialized subset of *EZNEC 3.0*. Where there are specific differences between *EZNEC 3.0* and the limited-edition *EZNEC-ARRL* we'll identify them.]

Native *NEC-2*

The native *NEC-2* program produces pages and pages of output formatted for a mainframe "line printer." You may be old enough to remember the stacks of green-and-white, tractor-feed, 132-column computer paper that such a line printer produced. Corporate MIS departments stored untold number of boxes of that paper.

Native *NEC-2* was written in the Fortran language, which stands for *Formula Translation*. Programmers used punched cards to enter the program itself and its accompanying input data into huge mainframe computers. To say that the paper output from *NEC-2* is massive, even intimidating, is putting it mildly. There is a strong distinction between "useful information" and "raw data" and the raw output from native *NEC-2* bombards the user with raw data.

Commercial versions that use the *NEC-2* computational core shield the user from the ugliness of raw line-printer output, as well as punched-card input (or disk surrogates for punched cards). Commercial versions like *EZNEC* do produce output numerical tables where this is useful. These tables show parameters such as the source impedance and SWR at a single frequency, or the characteristics of a load or a transmission line. But as the old saying goes, "One picture is worth a thousand words." This is as true for modeling programs as it is for other endeavors dealing with reams of numbers. Thus, most commercial modeling software packages create *graphs* for the user. *EZNEC* produces the following types of graphs:

- Polar (linear-dB or ARRL-style) graphs of the far-field elevation and azimuth responses.
- 3-D wire-frame graph of the total far-field response.
- Graph of the SWR across a frequency band.
- Graphical display of the RF currents on various conductors in a model.
- Rotatable, zoomable 3-D views of the wires used to make a model.
- Output to programs capable of generating Smith charts.

Fig 1A and 1B shows the computed far-field 2-D elevation and azimuth patterns for a 135-foot long horizontal dipole, mounted in a flattop configuration 50 feet above flat ground. These figures were generated using *EZNEC* at 3.75 MHz. Fig 1C shows a 3-D wire-frame picture of the far-field response, but this time at 14.2 MHz. For comparison, **Table 1** shows a short portion of the line-printer output for the azimuth pattern at 3.75 MHz. The actual printout is many pages long. One picture can indeed replace thousands of numbers!

Fig 2 shows the computed SWR curve over the frequency range 3.0 to 4.0 MHz for this dipole, fed with lossless 50-Ω transmission line. *EZNEC* generated this plot using the "SWR" button. Figs 1 and 2 are typical of the kind of graphical outputs that commercial implementations of the *NEC-2* computing core can produce.

Now, let's get into the details of what kind of input data is required to run a typical method-of-moments antenna-modeling program.

PROGRAM INPUTS: WIRE GEOMETRY

Coordinates in an X, Y and Z World

The most difficult part of using a *NEC*-type of modeling program is setting up the antenna's geometry—you must condition yourself to think in three-dimensional, Cartesian coordinates. Each end point of a wire is represented by three numbers: an x, y and z coordinate. These coordinates represent the distance from the origin (x-axis), the width of an antenna (y-axis), and the height (z-axis).

An example should help sort things out. **Fig 3** shows a simple model of a 135-foot center-fed dipole, made of #14 copper wire placed 50 feet above flat ground. The common term for this antenna is *flattop dipole*. For convenience, the ground is located at the *origin* of the coordinate system, at (0, 0, 0) feet, directly under the center of the dipole. **Fig 4** shows the *EZNEC* spreadsheet-like input data for this antenna. (Use model file: **Ch4-Flattop Dipole.EZ**.) *EZNEC* allows you to specify the type of conductor material from its main window, using the **Wire Loss** button to open a new window. We will click on the **Copper** button for this dipole.

Above the origin, at a height of 50 feet on the z-axis, is the dipole's *feed point*, called a *source* in *NEC* terminology. The width of the dipole goes toward the left (that is, in the "negative-y" direction) one-half the overall length of 135 feet, or –67.5 feet. Toward the right, our dipole's other end is at +67.5 feet. The x-axis dimension of our dipole is zero, meaning that the dipole wire is parallel to and directly above the x-axis. The dipole's ends are thus represented by two points, whose coordinates are (0, –67.5, 50) and (0, 67.5, 50) feet. The use of parentheses with a sequential listing of (x, y, z) coordinates is a common practice among antenna modelers to describe a wire end point.

Fig 3B includes some other useful information about this antenna beyond the wire geometry. Fig 3B overlays the wire geometry, the current distribution along the wire and the far-field azimuth response, in this case at an elevation angle of 30°.

Although not shown specifically in Fig 3, the thickness of the antenna is the diameter of the wire, #14 gauge. Note that native *NEC* programs specify the *radius* of the

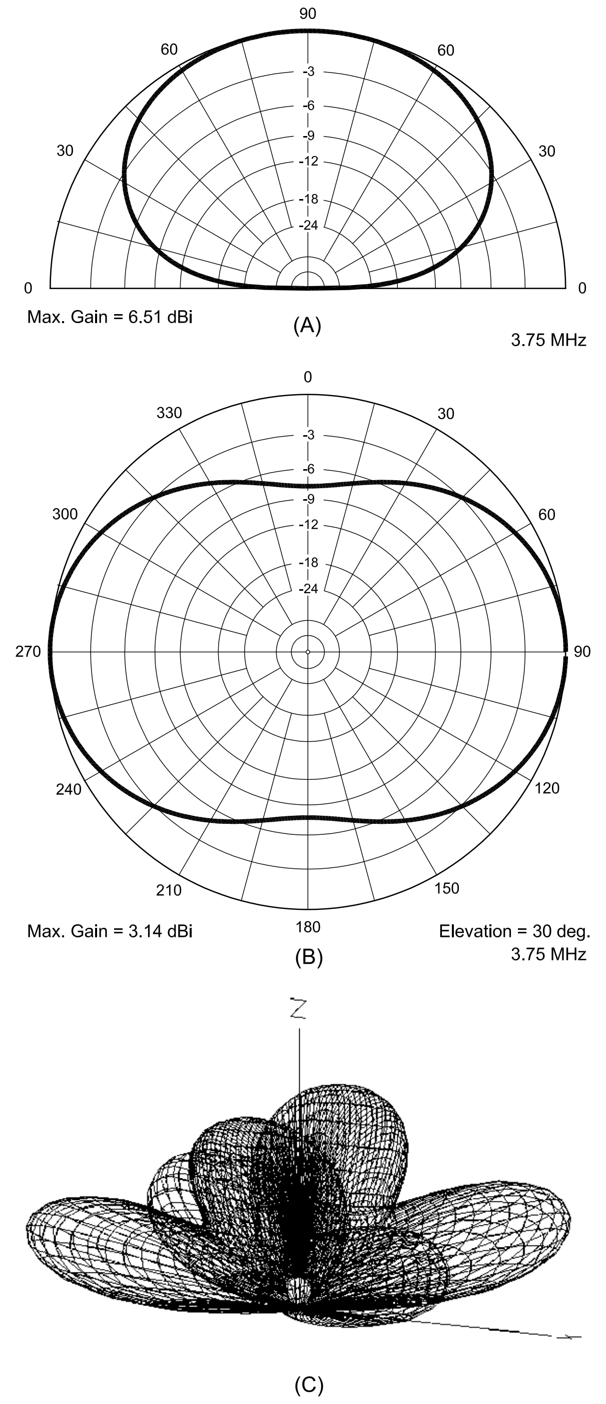

Fig 1—At A, far-field elevation-plane pattern for a 135-foot-long horizontal dipole, 50 feet above flat ground, at 3.5 MHz. At B, the far-field azimuth-plane pattern at an elevation angle of 30°.

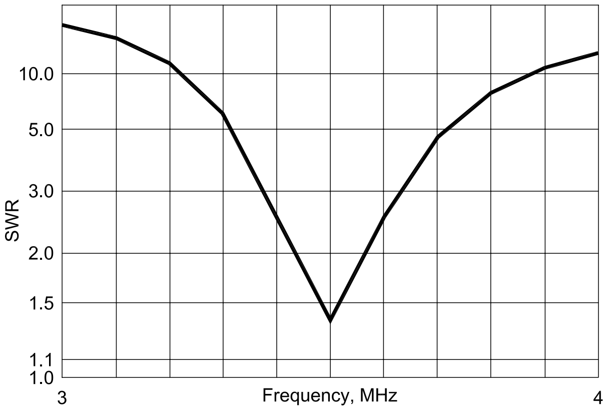

Fig 2—SWR curve for 135-foot flattop dipole over the frequency range 3.0 to 4.0 MHz for a 50-Ω feed line. This antenna is an example and is not optimized for the amateur band.

wire, rather than the diameter, but programs like *EZNEC* use the more intuitive diameter of a wire rather than the radius. *EZNEC* (and other commercial programs) also allows the user to specify the wire as an AWG gauge, such as #14 or #22, for example.

We've represented our simple dipole in Fig 3 using a single, straight wire. In fact, all antenna models created for method-of-moment programs are made of combinations of straight wires. This includes even complex antennas, such as helical antennas or round loops. (The mathematical basis for modeling complex antennas is that they can be simulated using straight-wire *polygons*. A circular loop, for example, can be modeled using an octagon.)

Segmentation and Specifying a Source Segment

We've specified the physical geometry of this simple one-wire dipole. Now several more modeling details surface—you must specify the number of *segments* into which the dipole is divided for the method-of-moment analysis and you must somehow feed the antenna. The *NEC-2* guideline for setting the number of segments is to use at least 10 segments per half-wavelength. This is a general rule of thumb, however, and in many models more dense segmentation is mandatory for good accuracy.

In Fig 3, we've specified that the dipole be divided into 11 segments for operation on the 80-meter band. This follows the rule of thumb above, since the 135-foot dipole is about a half-wavelength long at 3.5 MHz.

Setting the Source Segment

The use of 11 segments, an odd rather than an even number such as 10, places the dipole's feed point (the

Table 1
Portion of line-printer output from *NEC-2* for 135-foot dipole.

- - - RADIATION PATTERNS - - -

- - ANGLES - -		*- POWER GAINS -*			*- - - POLARIZATION - - -*			*- - - E(THETA) - - -*		*- - - E(PHI) - - -*	
Theta Degrees	*Phi Degrees*	*Vert dB*	*Hor dB*	*Total dB*	*Axial Ratio*	*Tilt Degrees*	*Sense*	*Magnitude Volts*	*Phase Degrees*	*Magnitude Volts*	*Phase Degrees*
60.00	0.00	–999.99	3.14	3.14	0.00000	90.00	LINEAR	0.00000E+00	0.00	6.62073E-01	–66.87
60.00	1.00	–37.87	3.13	3.14	0.00301	89.52	LEFT	5.89772E-03	–86.64	6.61933E-01	–66.87
60.00	2.00	–31.85	3.13	3.13	0.00603	89.04	LEFT	1.17915E-02	–86.64	6.61512E-01	–66.87
60.00	3.00	–28.33	3.12	3.12	0.00904	88.56	LEFT	1.76776E-02	–86.64	6.60812E-01	–66.87
60.00	4.00	–25.84	3.11	3.11	0.01206	88.08	LEFT	2.35520E-02	–86.64	6.59834E-01	–66.87
60.00	5.00	–23.91	3.09	3.10	0.01508	87.59	LEFT	2.94109E-02	–86.64	6.58577E-01	–66.87
60.00	6.00	–22.34	3.07	3.08	0.01810	87.11	LEFT	3.52504E-02	–86.64	6.57045E-01	–66.87
60.00	7.00	–21.01	3.05	3.06	0.02112	86.62	LEFT	4.10669E-02	–86.63	6.55237E-01	–66.87
60.00	8.00	–19.87	3.02	3.04	0.02415	86.14	LEFT	4.68565E-02	–86.63	6.53158E-01	–66.87
60.00	9.00	–18.86	2.99	3.02	0.02718	85.65	LEFT	5.26156E-02	–86.63	6.50808E-01	–66.87
60.00	10.00	–17.96	2.95	2.99	0.03022	85.15	LEFT	5.83405E-02	–86.63	6.48190E-01	–66.87
60.00	11.00	–17.15	2.91	2.96	0.03327	84.66	LEFT	6.40278E-02	–86.63	6.45308E-01	–66.86
60.00	12.00	–16.42	2.87	2.92	0.03631	84.16	LEFT	6.96739E-02	–86.63	6.42165E-01	–66.86
60.00	13.00	–15.75	2.83	2.89	0.03937	83.66	LEFT	7.52755E-02	–86.63	6.38764E-01	–66.86
60.00	14.00	–15.13	2.78	2.85	0.04243	83.16	LEFT	8.08291E-02	–86.63	6.35108E-01	–66.86
60.00	15.00	–14.56	2.72	2.80	0.04550	82.65	LEFT	8.63317E-02	–86.62	6.31203E-01	–66.86
60.00	16.00	–14.03	2.66	2.76	0.04858	82.14	LEFT	9.17800E-02	–86.62	6.27051E-01	–66.86
60.00	17.00	–13.53	2.60	2.71	0.05166	81.62	LEFT	9.71711E-02	–86.62	6.22657E-01	–66.85
60.00	18.00	–13.07	2.54	2.66	0.05475	81.10	LEFT	1.02502E-01	–86.62	6.18027E-01	–66.85

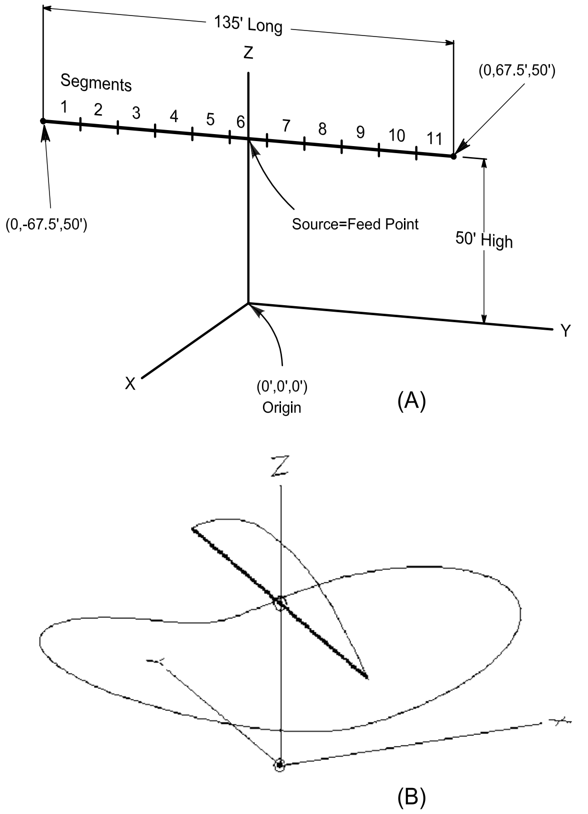

Fig 3—At A, simple model for a 135-foot long horizontal dipole, 50 feet above the ground. The dipole is over the y-axis. The wire has been segmented into 11 segments, with the center of segment number 6 as the feed point. The left-hand end of the antenna is -67.5 feet from the center feed point and that the right-hand end is at 67.5 feet from the center. At B, *EZNEC* "View Antenna" drawing, showing geometry of wire and the x, y and z axes. Overlaid on the wire geometry drawing are the current distribution along the wire and the far-field azimuthal response at an elevation angle of 30°.

source in *NEC*-speak, a word choice that can befuddle beginners) right at the antenna's center, at the center of segment number six. In concert with the "EZ" in its name, *EZNEC* makes choosing the source segment easy by allowing the user to specify a percentage along the wire, in this case 50% for center feeding.

At this point you may very well be wondering why no center insulator is shown in the middle of our center-fed dipole. After all, a real dipole would have a center insulator. However, method-of-moment programs assume that a source generator is placed across an infinitely small gap in the antenna wire. While this is convenient from a mathematical point of view, the unstated use of such an infinitely small gap often confuses newcomers to the world of antenna modeling. We'll get into more details, caveats and limitations in source placement later in this

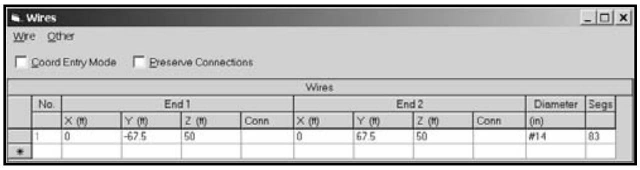

☐ Coord Entry Mode ☐ Preserve Connections

Wires

	No.	End 1				End 2				Diameter	Segs
		X (ft)	Y (ft)	Z (ft)	Conn	X (ft)	Y (ft)	Z (ft)	Conn	(in)	
	1	0	-67.5	50		0	67.5	50		#14	83
*											

Fig 4—*EZNEC* "Wires" spreadsheet for simple flattop dipole in Fig 3. The numbers shown are in feet, except for the wire diameter, which *EZNEC* allows you to specify as an AWG gauge, in this case #14. Note that 83 segments have been specified for this antenna for analysis over the range from 3.5 to 29.7 MHz.

chapter. For now, just trust us that the model we've just described with 11 segments, fed at segment 6, will work well over the full amateur band from 3.5 to 4.0 MHz.

Now, let's consider what would happen if we want to use our 135-foot long dipole on all HF amateur bands from 3.5 to 29.7 MHz, rather than just from 3.5 to 4.0 MHz. Instead of feeding such an antenna with coax cable, we would feed it with open-wire line and use an antenna tuner in the shack to create a 50-Ω load for the transmitter. To comply with the segmentation rule above, the number of segments used in the model should vary with frequency—or at least be segmented at or above the minimum recommended level at the highest frequency used. This is because a half wavelength at 29.7 MHz is 16.6 feet, while a half wavelength at 3.5 MHz is 140.6 feet. So the number of segments for proper operation on 29.7 MHz should be 10 × 135/16.6 = 81. We'll be a little more conservative than the minimum requirement and specify 83 segments. **Fig 4** shows the *EZNEC* input spreadsheet for this model. (Use model file: **Ch4-Multiband Dipole.EZ**.)

The penalty for using more segments in a program like *NEC* is that the program slows down roughly as the square of the segments—double the number of segments and the speed drops to a fourth. If we try to use too few segments, we'll introduce inaccuracies, particularly in computing the feed-point impedance. We'll delve into this area of segmentation density in more detail later when we discuss testing the adequacy of a model.

Segment Length-to-Wire-Diameter Ratio

Even if you're willing to live with the slowdown in computing speed for situations involving a large number of wire segments, you should make sure the ratio between the segment length and the diameter of any wire is greater than 1:1. This is to say that the length of each segment is longer than the diameter of the wire. Doing so stays away from internal limitations in the *NEC* program.

For the #14 wire specified in this simple 135-foot long dipole, it's pretty unlikely that you'll bump up against this limitation for any reasonable level of segmentation. After all, #14 wire has a diameter of 0.064 inches and 135 feet is 1620 inches. To keep above a segment length of 0.064 inches, the maximum number of segments is 1620/

0.064 = 25,312. This is a *very* large number of segments and it would take a very long time to compute, assuming that your program can handle that many segments.

Keeping above a 1:1 ratio in segment length to wire diameter can be more challenging at VHF/UHF frequencies, however. This is particularly true for fairly large "wires" made of aluminum tubing. Incidentally, this is another point where newcomers to antenna modeling can be led astray by the terminology. In a *NEC*-type program, all conductors in a model are considered to be *wires*, even if they consist of hollow aluminum or copper tubes. Surface effect keeps the RF current in any conductor confined to the outer surface of that conductor, and thus it doesn't matter whether the conductor is hollow or solid, or even made using a number of stranded wires twisted together.

Let's look at a half-wave dipole at 420 MHz. This would be about 14.1 inches long. If you use $^1/_4$-inch diameter tubing for this dipole, the maximum segment length meeting the 1:1 diameter-to-length ratio requirement is also $^1/_4$ inches long. The maximum number of segments then would be 14.1/0.25 = 56.4, rounded down to 56. From this discussion you should now understand why method-of-moment programs are known for using a "thin-wire approximation." Really fat conductors can get you into trouble, particularly at VHF/UHF.

Some Caveats and Limitations Concerning Geometry

Example: Inverted-V Dipole

Now, let's get a little more complicated and specify another 135-foot-long dipole, but this time configured as an *inverted V*. As shown in **Fig 5**, you must now specify two wires. The two wires join at the top, at (0, 0, 50) feet. (Again, the program doesn't use a center insulator in the model.)

If you are using a native version of *NEC*, you may have to go back to your high-school trigonometry book to figure out how to specify the end points of our "droopy" dipole, with its 120° included angle. Fig 5 shows the details, along with the trigonometric equations needed. *EZNEC* is indeed more "easy" here, since it allows you to tilt the ends of each wire downwards an appropriate number of degrees (in this case –30° at each end of the dipole) to automatically create an inverted-V configuration. **Fig 6A** shows the *EZNEC* spreadsheet describing this inverted-V dipole with a 120° included angle between the two wires.

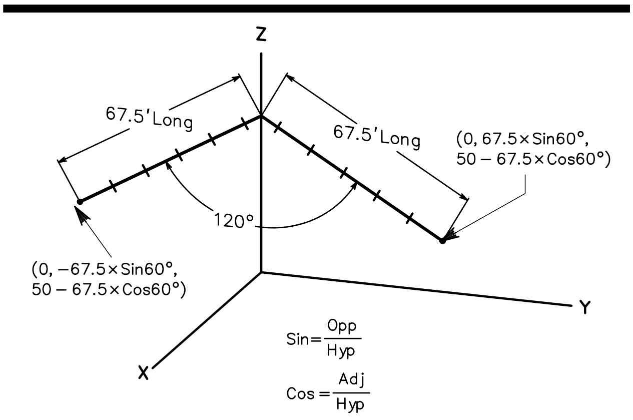

Fig 5—Model for an inverted-V dipole, with an included angle between the two legs of 120°. Sine and cosine functions are used to describe the heights of the end points for the sloping arms of the antenna.

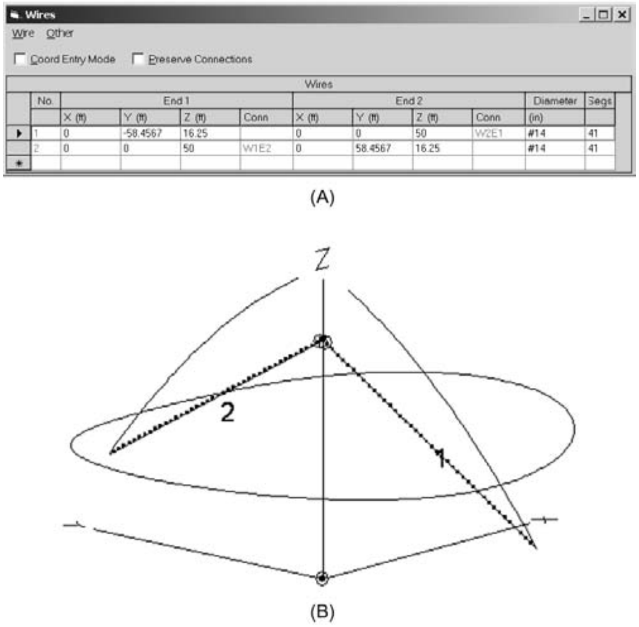

	Wires										
	No.	End 1				End 2				Diameter	Segs
		X (ft)	Y (ft)	Z (ft)	Conn	X (ft)	Y (ft)	Z (ft)	Conn	(in)	
▶	1	0	-58.4567	16.25		0	0	50	W2E1	#14	41
	2	0	0	50	W1E2	0	58.4567	16.25		#14	41
*											

(A)

(B)

Fig 6—At A, *EZNEC* spreadsheet for inverted-V dipole in Fig 5. Now the ends of the inverted-V dipole are 16.25 feet above ground, instead of 50 feet for the flattop dipole. At B, *EZNEC* "View Antenna" drawing, with overlay of geometry, current distribution and azimuth plot.

See the *EZNEC* **Help** section under "Wire Coordinate Shortcuts" for specific instructions on how to use the "elevation rotate end" shortcut "RE–30" to create the sloping wires easily by rotating the end of the wire down 30°. Now the specification of the source becomes a bit more complicated. The easiest way is to specify two sources, one on each end segment at the junction of the two wires. *EZNEC* does this automatically if you specify a so-called *split-source* feed. Fig 6B shows the two sources as two open circles at the top ends of the two wires making up the inverted-V dipole. What *EZNEC* is doing is creating two sources, each on the closest segments on either side of the junction of the two wires. *EZNEC* sums up the two source impedances to provide a single readout.

Navigating in the View Antenna Window

At this point it's worthwhile to explore some of the ways you can see what the wire geometry looks like using the *EZNEC* **View Ant** button on the main window. Bring up the file **Ch4-Inverted V Dipole.EZ** in *EZNEC*, and click on the **View Ant** button. You will see a small inverted-V dipole raised over the (0, 0, 0) origin on the ground directly under the feed point of the inverted-V dipole. First, "rotate" the dipole by holding down the left-mouse button and moving the mouse. You can orient the picture any way you wish.

Let's take a closer look at the junction of the two wires at the feed point. Click the **Center Ant Image** checkbox toward the bottom of the window to anchor the center of the image at the center of the window, and then move the **Zoom** slider upwards to zoom in on the image. At some point the junction of the two slanted wires will move up off the edge of the window, so you will need to click on the left-hand side of the **Z Move Image** slider to bring the junction back into view. Now you should be able to see a zoomed view of the junction, along with the two open circles that represent the location of the split sources in the middle of the segments adjacent to the wire junction.

Now put the mouse cursor over one of the slanted wires and double click the left-mouse button. *EZNEC* will now identify that wire and show its length, as well as the length of each segment on that wire. Pretty slick, isn't it?

Short, Fat Wires and the Acute-Angle Junction

Another possible complication can arise for wires with short, fat segments, particularly ones that have only a small included angle between them. These wire segments can end up inter-penetrating within each other's volumes, leading to problems in a model. Once you think of each wire segment as a thick cylinder, you can appreciate the difficulty in connecting two wires together at their ends. The two wires always inter-penetrate each other's volume to some extent. **Fig 7** depicts this problem graphically for two short, fat wires joined at their ends at an acute angle. A rule of thumb is to avoid creating junctions where more than $^1/_3$ of the wire volumes inter-penetrate. You can achieve this by using longer segment lengths or thinner wire diameters.

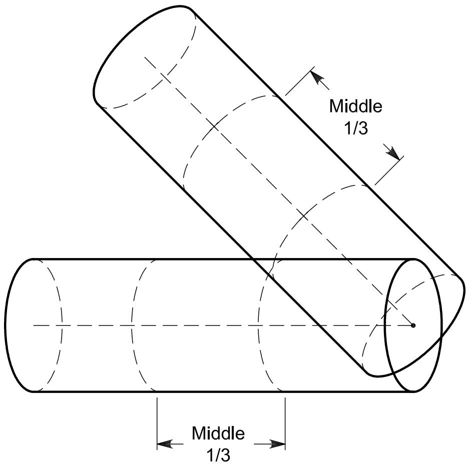

Fig 7—A junction of two short, fat wire segments at an acute angle. This results in inter-penetration of the two wire volumes beyond the middle-1/3 recommended limit.

Some Other Practical Antenna Geometries

A Vertical Half-Wave Dipole

If you turn the 135-foot-long horizontal dipole in Fig 1 on its end you will create a vertical half-wave dipole that is above the origin of the x, y and z axes. See **Fig 8**, where the bottom end of the dipole is placed 8 feet off the ground to keep it away from humans and animals for safety, at (0, 0, 8) feet. The top end is thus at 8 + 135 = 143 feet off the ground at (0, 0, 143). Fig 8 also shows the current distribution and the elevation pattern for this antenna. (Use *EZNEC* model file: **Ch4-Vertical Dipole.EZ**.)

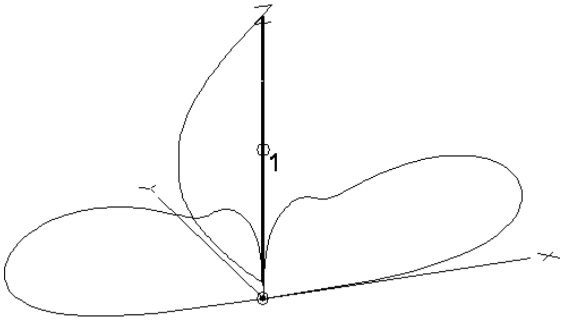

Fig 8—A vertical half-wave dipole, created by turning the dipole in Fig 3 on its end, with a minimum height at the lower end of 8 feet to keep the antenna away from people and animals. The current distribution and the elevation pattern for this antenna are also shown overlaid on the wire geometry.

A Ground-Plane Antenna

The ground-plane model is more complicated than previous ones because a total of five wires are now needed: one for the vertical radiator and four for the radials. **Fig 9** shows the *EZNEC* view for a 20-meter ground plane mounted 15 feet off the ground (perhaps on a garage roof), with the overlay of both the current distribution and the elevation-plane plot. (Use *EZNEC* model file: **Ch4-GP.EZ**.) Note that the source has been placed at the

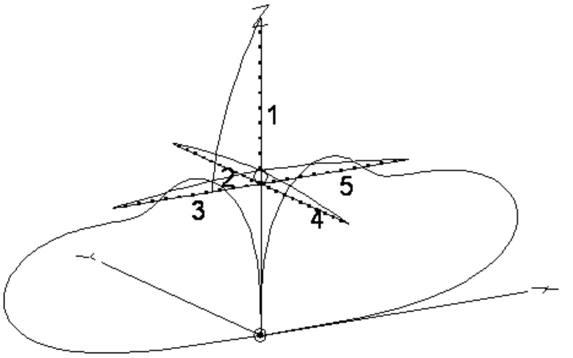

Fig 9—A vertical ground-plane antenna. The radials and the bottom of the vertical radiator are located 15 feet off the ground in this model. The current distribution along each wire and the far-field elevation-plane pattern are overlaid on the antenna geometry.

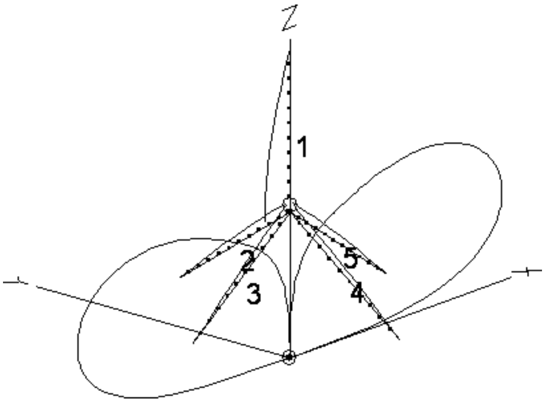

Fig 10—*EZNEC* View Antenna for the ground-plane antenna with its four radials tilted downwards by 40° to improve the SWR at the feed point.

bottom segment of the vertical radiator. Once again, the program needs no bottom insulator, since all five wires are connected together at a common point. *EZNEC* reports that this antenna has a resonant feed-point impedance of about 22 Ω, which would show an SWR of 2.3:1 for a 50-Ω coax feed line if no matching system is used, such as a gamma or hairpin match.

Fig 10 shows the same antenna, except that the radials have now been tilted downwards by 35° to facilitate an almost perfect 50-Ω match (SWR = 1.08:1). In addition, the length of the radiator in this model was shortened by 6 inches to re-resonate the antenna. (Use *EZNEC* model file: **Ch4-Modified GP.EZ**.) The trick of tilting the radials downwards for a ground-plane antenna is an old one, and the modeling programs validates what hams have been doing for years.

A 5-Element Horizontal Yagi

This is a little more challenging modeling exercise. Let's use a 5-element design on a 40-foot boom, but rather than using telescoping aluminum tubing for the elements, we'll use #14 wire. The *SCALE* program included with this book on the CD-ROM converted the aluminum-tubing 520-40.YW to a design using #14 copper wire. **Table 2** shows the element lineup for this antenna. (Later in this chapter we'll see what happens when telescoping aluminum tubing is used in a real-world Yagi design.)

Some explanations of what Table 3 means are in order. First, only one half of each element is shown. The *YW* program (Yagi for Windows), also included on the CD-ROM, computes the other half of the Yagi automatically, essentially mirroring the other half on the opposite side of the boom. Having to enter the dimensions for only half of a real-world Yagi element that uses telescoping aluminum tubing is much easier this way.

Second, the placement of the elements along the boom starts at 0.0 inches for the reflector. The distance between adjacent elements defined in this particular file is the spacing between the element itself and the element just before it. For example, the spacing between the driven element and the reflector is 72 inches, and the spacing between the first director and the driven element is also 72 inches. The spacing between the second director and the first director is 139 inches.

Fig 11A shows the wire geometry for this Yagi array when it is mounted 720 inches (60 feet) above flat ground and Fig 11B shows the *EZNEC* Wires spreadsheet that describes the coordinates. (Use *EZNEC* model file: **Ch4-520-40W.EZ**.) You can see that the x-axis coordinates for the elements have been automatically moved by the *SCALE* program so that the center of the boom is located directly above the origin. This makes it easier to evaluate the effects of stacking different monoband Yagis on a rotating mast in a "Christmas Tree" arrangement. A typical Christmas Tree stack might include 20, 15 and 10-meter monobanders on a single rotating mast sticking out of the top of the tower.

Fig 12 shows the computed azimuth pattern for this Yagi at 14.175 MHz, at an elevation angle of 15°, the angle where the peak of the forward lobe occurs at this

Table 2

520-40W.YW, using #14 wire from 520-40H.YW
14.000 14.174 14.350 MHz

5 elements,	inches
Spacing	.064
0.000	210.923
72.000	200.941
72.000	199.600
139.000	197.502
191.000	190.536

height above flat ground. The antenna exhibits excellent gain at 13.1 dBi, as well as a clean pattern behind the main lobe. The worst-case front-to-rear ratio at any point from 90° to 270° in azimuth is better than 23 dB. *EZNEC* says the feed-point impedance is 25 – *j* 23 Ω, just the right impedance suited for a simple hairpin or gamma match.

A Monoband 2-Element Cubical Quad

Unlike a Yagi, with its elements existing only in the x-y plane, a quad type of beam is a three-dimensional sort of antenna. A quad loop has height in the z-axis, as well as width and length in the x-y plane. Each individual loop for a monoband quad consists of four wires, joined together at the corners. **Fig 13** shows the coordinates for a 2-element 15-meter quad, consisting of a reflector and a driven element on a 10-foot boom.

You can see that the axis of symmetry, the x-axis, runs down the center of this model, meaning that the origin of this particular x, y and z-coordinate scheme is in the center of the reflector. The (0, 0, 0) origin is placed this way for convenience in assigning corner coordinates for each element. For actual placement of the antenna at a particular height above real ground, the heights of all z-axis coordinates are changed accordingly. *EZNEC* has a convenient built-in function to change the height of all wires at a single stroke.

Fig 14 shows the input *EZNEC* spreadsheet for this quad in free space, clearly showing the symmetrical nature of the corner coordinates. (Use *EZNEC* model file: **Ch4-Quad.EZ**.) This is a good place to emphasize that you should enter the wire coordinates in a logical

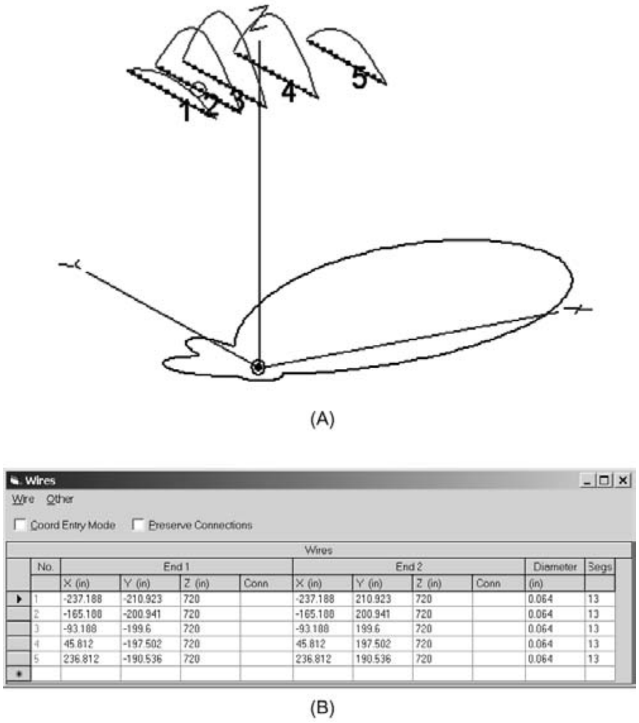

	No.	End 1				End 2				Diameter	Segs
		X (in)	Y (in)	Z (in)	Conn	X (in)	Y (in)	Z (in)	Conn	(in)	
▶	1	-237.188	-210.923	720		-237.188	210.923	720		0.064	13
	2	-165.188	-200.941	720		-165.188	200.941	720		0.064	13
	3	-93.188	-199.6	720		-93.188	199.6	720		0.064	13
	4	45.812	-197.502	720		45.812	197.502	720		0.064	13
	5	236.812	-190.536	720		236.812	190.536	720		0.064	13
*											

Fig 11—At A, geometry for 5-element Yagi on a 40-foot boom, mounted 720 inches (60 feet) above flat ground, with an overlay of current and the azimuth pattern. At B, *EZNEC* Wires spreadsheet for this antenna. This design uses #14 wire for simplicity.

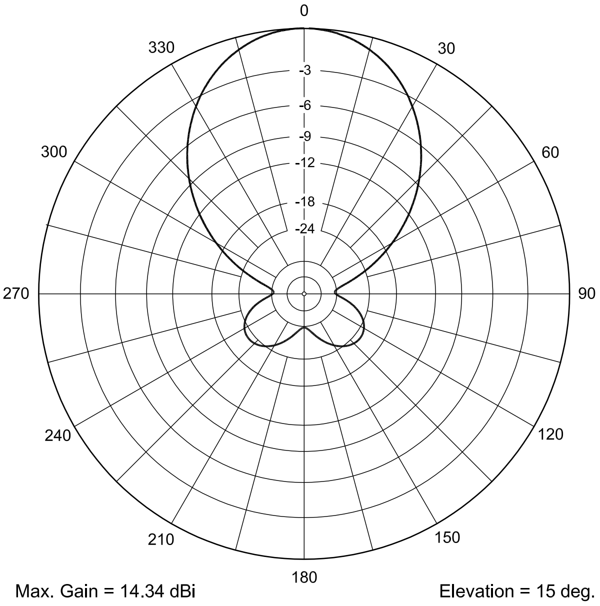

Fig 12—*EZNEC* azimuth-plane pattern at an elevation angle of 15° for #14 wire Yagi described in Fig 11.

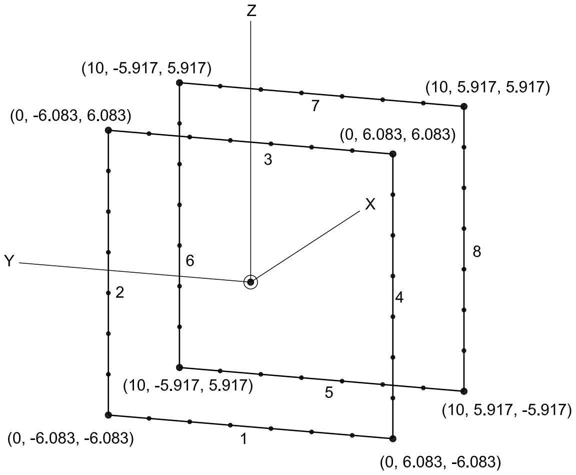

Fig 13—Wire geometry for a 2-element cubical quad, with a reflector and driven element. The x-axis is the axis of symmetry for this free-space model.

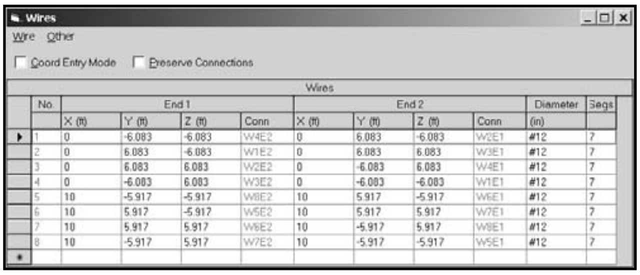

	Wires									
No.	End 1				End 2				Diameter	Segs
	X (ft)	Y (ft)	Z (ft)	Conn	X (ft)	Y (ft)	Z (ft)	Conn	(in)	
1	0	-6.083	-6.083	W4E2	0	6.083	-6.083	W2E1	#12	7
2	0	6.083	-6.083	W1E2	0	6.083	6.083	W3E1	#12	7
3	0	6.083	6.083	W2E2	0	-6.083	6.083	W4E1	#12	7
4	0	-6.083	6.083	W3E2	0	-6.083	-6.083	W1E1	#12	7
5	10	-5.917	-5.917	W8E2	10	5.917	-5.917	W6E1	#12	7
6	10	5.917	-5.917	W5E2	10	5.917	5.917	W7E1	#12	7
7	10	5.917	5.917	W6E2	10	-5.917	5.917	W8E1	#12	7
8	10	-5.917	5.917	W7E2	10	-5.917	-5.917	W5E1	#12	7

Fig 14—*EZNEC* Wires spreadsheet showing the coordinates used for the quad in Fig 13. Note how the x-axis describes the position of an element on the 10-foot boom and also is the axis of symmetry for each element. The values for the z-axis and y-axis vary above and below the axis of symmetry.

sequence. The most obvious example in this particular model is that you should group all the wires associated with a particular element together—for example, the four wires associated with the reflector should be in one place. In Fig 14 you can see that all four wires with an x-coordinate of zero represent the reflector.

It's best to follow a convention in entering wires in a loop structure in a logical fashion. The idea to connect the end point of one wire to the starting point of the next wire. For example, in Fig 13 you can see that the left-hand end of Wire 1 is connected to the bottom of Wire 2, and that the top of Wire 2 connects to the left-hand end of Wire 3. In turn, Wire 3 connects to the top of Wire 4, whose bottom end connects to the right-hand end of Wire 1. The pattern is known as "going around the horn" meaning that the connections proceed smoothly in one direction, in this case in a clockwise direction.

You can see that the entry for the wires making up the elements in the 5-element Yagi in Fig 11B also proceeded in an orderly fashion by starting with the reflector, then the driven element, then director 1, then director 2 and finally director 3. This doesn't mean that you couldn't mix things up, say by specifying the driven element first, followed by director 3, and then the reflector, or whatever. But it's a pretty good bet that doing so in this quasi-random fashion will result in some confusion later on when you revisit a model, or when you let another person see your model.

THE MODELING ENVIRONMENT

The Ground

Above, when considering the 135-foot dipole mounted 50 feet above flat earth, we briefly mentioned the most important environmental item in an antenna model—the ground beneath it. Let's examine some of the options available in the *NEC-2* environment in *EZNEC*:

- Free space
- Perfect ground
- *MININEC* type ground
- "Fast" type ground
- Sommerfeld-Norton ground.

The free space environment option is pretty self-explanatory—the antenna model is placed in free space away from the influence of any type of ground. This option is useful when you wish to optimize certain characteristics of a particular antenna design. For example, you might wish to optimize the front-to-rear ratio of a Yagi over an entire amateur band and this might entail many calculation runs. The free-space ground will run the fastest among all the ground options.

Perfect ground is useful as a reference case, especially for vertically polarized antennas over real ground. Antenna evaluations over perfect ground are shown in most classical antenna textbooks, so it is useful to compare models for simple antennas over perfect ground to those textbook cases.

MININEC type ground is useful when modeling vertical wires, or horizontal wires that are higher than 0.2 λ above ground. A *MININEC* type ground will compute faster than either a "Fast" ground or a Sommerfeld-Norton type of ground because it assumes that the ground under the antenna is perfect, while still taking into account the far-field reflections for ground using user-specified values of ground conductivity and dielectric constant. The fact that the ground under the antenna is perfect allows the *NEC-2* user of a *MININEC* type ground to specify wires that touch (but don't go below) the ground surface, something that only users of the advanced *NEC-4* program can do with the more accurate Sommerfeld-Norton type of ground described below. (*NEC-4* is presently not in the public domain and is strictly restricted and licensed by the US government.) The ability to model grounded wires is useful with vertical antennas. The modeler must be wary of the feed-point source impedances reported for either horizontally or vertically polarized wires because of the perfect-ground assumption inherent in a *MININEC*-type ground.

The "fast" type of ground is a hybrid type of ground that makes certain simplifying assumptions that allow it to be used provided that horizontal wires are higher than about 0.1 λ above ground. With today's fast computers the Sommerfeld-Norton model is preferred.

The Sommerfeld-Norton ground (referred to in *EZNEC* as the "high accuracy" ground) is preferable to the other ground types because it has essentially no practical limitations for wire height. It has the disadvantage that it can run about four times slower than a *MININEC* type of ground, but today's fast computers make that almost a non-issue. Again, *NEC-2*-based programs cannot model wires that penetrate into the ground (although there are work-arounds described below).

As mentioned above, for any type of ground other than perfect ground or free space, the user must specify the conductivity and dielectric constant of the soil. *EZNEC* allows the entry by several user-friendly categories, where σ is conductivity in Siemens/meter and ε is dielectric constant:

- Extremely poor: cities, high buildings (σ = 0.001, ε = 3)
- Very Poor: cities, industrial (σ = 0.001, ε = 5)
- Sandy, dry (σ = 0.002, ε = 10)
- Poor: rocky, mountainous (σ = 0.002, ε = 13)
- Average: pastoral, heavy clay (σ = 0.005, ε = 13)
- Pastoral: medium hills and forestation (σ = 0.006, ε = 13)
- Flat, marshy, densely wooded (σ = 0.0075, ε = 12)
- Pastoral, rich soil, US Midwest (σ = 0.010, ε = 14)
- Very Good: pastoral, rich, central US (σ = 0.0303, ε = 20)
- Fresh water (σ = 0.001, ε = 80)
- Saltwater (σ = 5, ε = 80)

Let's use *EZNEC*'s ability to overlay one or more plots together on one graph to compare the response of the vertical ground plane antenna in Fig 9 for two different types of ground: Saltwater and Poor. Open the **Ch4-GP.EZ** file in *EZNEC*. Click the **Ground Descrip** button and then right-click anywhere in the Media window that opens up. Choose first the "Poor: rocky, mountainous" option button, click **OK** and then **FF Plot**. When the elevation plot appears, click the **File** menu at the top of the main window, and then **Save As**. Choose an appropriate name for the trace, perhaps "Poor Gnd.PF."

Go back and select saltwater as your **Ground Descrip** and follow the same procedure to compute the far-field plot for saltwater ground. Now, add the Poor Gnd.PF trace, by clicking menu selection **File**, **Add Trace**. **Fig 15** shows this comparison, which greatly favors the saltwater environment, particularly at low elevation angles. At 5° the ground plane mounted over saltwater has about a 10 dB advantage compared to its landlocked cousin.

You might be wondering what happens if we move the ground-plane antenna down closer to the ground. The lower limit to how far the radials can approach the lossy earth is 0.001 λ or twice the diameter of the radial wire. A distance of 0.001 λ at 1.8 MHz is about 6 inches, while it is 0.4 inches at 30 MHz. While *NEC-2*-based programs cannot model wires that penetrate the ground, radial systems just above the ground with more than about eight radial wires can provide a work-around to simulate a direct-ground connection.

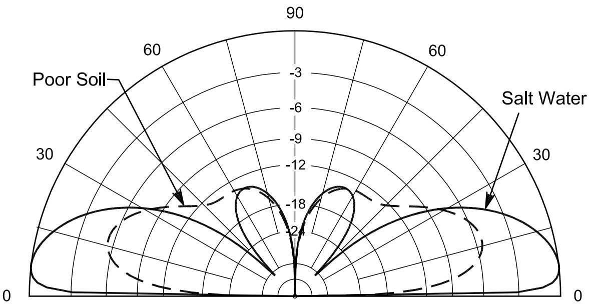

Fig 15—A comparison of the elevation response for the vertical ground plane in Fig 9 over saltwater and over "poor: rocky, mountainous" soil. Saltwater works wonders for verticals, providing excellent low-angle signals.

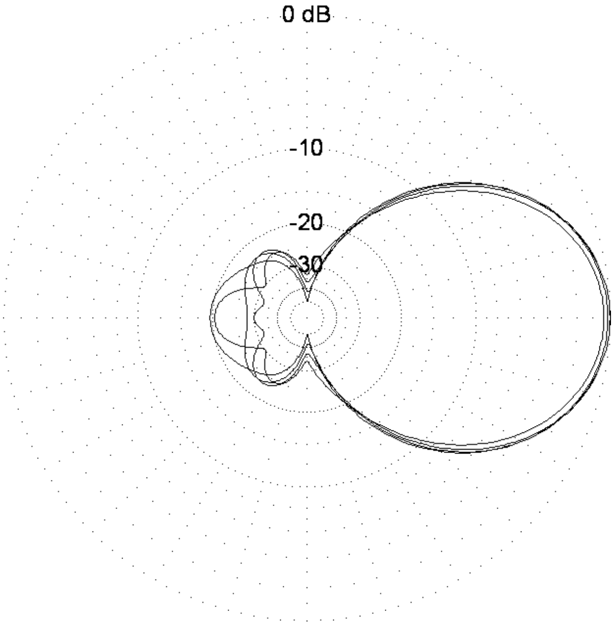

Fig 16—Frequency sweep of 5-element Yagi described in Fig 11, showing how the azimuth pattern changes with frequency.

Modeling Environment: Frequency

It's always a good idea to evaluate an antenna over a *range of frequencies*, rather than simply at a single spot frequency. Trends that become quite apparent on a frequency sweep are frequently lost when looking simply at a single frequency. Native *NEC-2* has built-in frequency sweep capabilities, but once again the commercial programs make the process easier to use and understand. You saw in the SWR curve in Fig 2 the result of one such frequency sweep using *EZNEC*. **Fig 16** shows a frequency sweep of the azimuth response for the 5-element Yagi in Fig 11 across the 20-meter band, using steps of 117 kHz so there are four evaluation frequencies. At 14.0 MHz this Yagi's gain is down a small amount compared to the gain at 14.351 MHz but the rearward pattern is noticeably degraded, dropping to a front-to-back ratio of just under 20 dB.

EZNEC can save to a series of output plot files a frequency sweep of elevation (or azimuth) patterns. In essence, this automates the process described above for saving a plot to disk and then overlaying it on another plot. *EZNEC* can save to a text file for later analysis (or perhaps importation into a spreadsheet) the following parameters, chosen by the user:

- Source data
- Load data
- Pattern data
- Current data
- MicroSmith numeric data
- Pattern analysis summary.

Frequency Scaling

EZNEC has a very useful feature that allows you to create new models *scaled* to a new frequency. You invoke the algorithm used to scale a model from one frequency to another by checking the **Rescale** box after you've clicked the **Frequency** button. *EZNEC* will scale all model dimensions (wire length, height and diameter) except for one specific situation—the wire diameter will stay the same at the new frequency if you originally specified wire size by AWG gauge. For example, #14 copper wire for a half-wave 80-meter dipole will stay #14 copper wire for a 20-meter half-wave dipole. If, however, you specified diameter as a floating point number originally, the diameter will be scaled by the ratio of new to old frequency, along with wire length and height.

Start up *EZNEC* and open up the file **Ch4-520-40W.EZ** for the 5-element 20-meter Yagi on a 40-foot boom. Click the **Frequency** box and then check the **Rescale** check box. Now, type in the frequency of 28.4 MHz and click **OK**. You have quickly and easily created a new 5-element 10-meter Yagi, that is mounted 29.9949 feet high, the exact ratio of 28.4 MHz to 14.1739 MHz, the original design frequency on 20 meters. Click the **FF Plot** button to plot the azimuth pattern for this new Yagi. You will see that it closely duplicates the performance of its 20-meter brother. Click **Src Dat** to see that the source impedance is 25.38 – *j* 22.19 Ω, again very close to the source data for the 20-meter version.

REVISITING SOURCE SPECIFICATION

Sensitivity to Source Placement

Earlier, we briefly described how to specify a source on a particular segment using *EZNEC*. The sources for the relatively simple dipole, Yagi and quad models investigated so far have been in the center of an easy-to-visualize wire. The placement for the source on the vertical ground plane was at the bottom of the vertical radiator, an eminently logical place. In the other cases we specified the position of the source at 50% of the distance along a wire, given that the wire being fed had an odd number of segments. Please note that in each case so far, the feed point (source) has been placed at a relatively low-impedance point, where the current changes relatively slowly from segment to segment.

Now we're going to examine some subtler source-placement problems. *NEC-2* is well-known as being very sensitive to source placement. Significant errors can result from a haphazard choice of the source segment and the segments surrounding it.

Let's return to the inverted-V dipole in Fig 5. The first time we evaluated this antenna (**Ch4-Inverted V Dipole.EZ**) we specified a split source in *EZNEC*. This function uses two sources, one on each of the segments immediately adjacent to the junction of the two downward slanting wires.

Another common method to create a source at the junction of two wires that meet at an angle is to separate these two slanted wires by a short distance and bridge that gap with a short straight wire, which is fed at its center. **Fig 17** shows a close-up of this scheme. In Fig 17 the length of the segments surrounding the short middle wire are purposely made equal to the length of the middle wire. The segmentation for the short middle wire is set to one. **Table 3** lists the source impedance and the maximum gain the *EZNEC* computes for three different models:

1. **Ch4-Inverted V Dipole.EZ** (the original model)
2. **Ch4-Inverted V Dipole Triple Segmentation.EZ**
3. **Ch4-Modified Inverted V Dipole.EZ** (as shown in Fig 17, for the middle wire set to be 2 feet long)
4. **Ch4-Mod Inverted V Poor Segmentation.EZ** (where the number of segments on the two slanted wires have been increased to 200)

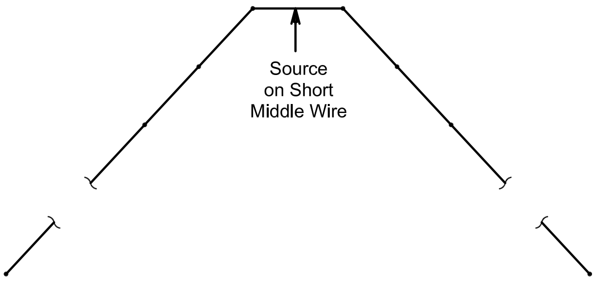

Fig 17—Model of inverted-V dipole using a short center wire on which the source is placed.

Table 3
135-Foot Inverted-V Dipole at 3.75 MHz

Case	*Segments*	*Source Impedance Ω*	*Max. Gain dBi*
1	82	72.64 + *j* 128.2	4.82
2	246	73.19 + *j* 128.9	4.82
3	67	73.06 + *j* 129.1	4.85
4	401	76.21 + *j* 135.2	4.67

Case 2 shows the effect of tripling the number of segments in Case 1. This is a check on the segmentation, to see that the results are stable at a lower level compared to a higher level of segmentation (which theoretically is better, although slower in computation). We purposely set up Case 4 so that the lengths of the segments on either side of the single-segment middle wire are significantly different (0.33 feet) compared to the 2-foot length of the middle wire.

The feed-point and gain figures for the first three models are close to each other. But you can see that the figures for the fourth model are beginning to diverge from the first three, with about a 5% overall change in the reactance and resistance compared to the average values, and about a 3% change in the maximum gain. This illustrates that it is best to keep the segments surrounding the source equal or at least close to equal in length. We'll soon examine a figure of merit called the *Average Gain* test, but it bears mentioning here that the average gain test is very close for the first three models and begins to diverge for the fourth model.

Things get more interesting if the source is placed at a high-impedance point on an antenna—for example, in the center of a full-wave dipole—the value computed for the source impedance will be high, and things will be quite sensitive to the segment lengths. We'll repeat the computations for the same inverted-V models, but this time at twice the operating frequency, at 7.5 MHz.

Table 4 summarizes the results. The impedance is high, as expected. Note that the resistance term varies quite a bit for all four models, a range of about 23% around the average value. Interestingly, the poorly segmented model's resistance falls in between the other three. The reactive terms are closer for all four models but still cover a range of 4% around the average value. The maximum gain shows the same tendancy to be somewhat lower in the fourth model compared to the first three and thus looks as potentially untrustworthy at 7.5 MHz as it does at 3.75 MHz.

This is, of course, but a small sampling of segmentation schemes, and caution dictates that you shouldn't take these results as being representative of all possibilities. Nevertheless, the lesson to be learned here is that the feed-point (source) impedance can vary significantly at a point where the current is changing rapidly, as it does where a high impedance feed is involved. Another general conclusion that can be drawn from Table 5 is that more segments, particularly if they surround the source segment improperly, is not necessarily better.

Table 4
135-Foot Inverted-V Dipole at 7.5 MHz

Case	*Segments*	*Source Impedance Ω*	*Max. Gain dBi*
1	82	2297 – *j* 2668	5.67
2	246	1822 – *j* 2553	5.66
3	67	1960 – *j* 2583	5.66
4	401	2031 – *j* 2688	5.48

Voltage and Current Sources

Before we leave the topic of sources, you should be aware that programs like *EZNEC* and others have the ability to simulate both voltage sources and current sources. Although native *NEC-2* has several source types, voltage sources are the most commonly used by hams. Native *NEC-2* doesn't have a current source, but a current source is nothing more than a voltage source delivering current through a high impedance. Basic network theory says that every Thevenin voltage source has a Norton current source equivalent. Various commercial implementations of *NEC-2* approach the creation of a current source in slightly different fashions. Some use a high value of inductive reactance as a series impedance, while others use a high value of series resistance.

Why would we want to use a current source instead of a voltage source in a model? The general-purpose answer is that models containing a single source at a single feed point can use a voltage source with no problems. Models that employ multiple sources, usually with different amplitudes and different phase shifts, do best with current sources.

For example, *phased arrays* feed RF currents at different amplitudes and phase shifts into two or more elements. The impedances seen at each element may be very different—some impedances might even have negative values of resistance, indicating that power is flowing out of that element into the feed system due to mutual coupling to other elements. Having the ability to specify the amplitude and phase of the current, rather than a feed voltage, at a feed point in a program like *EZNEC* is a valuable tool.

Next, we examine one more important aspect of building a model, setting up loads. After that, we'll look into two tests for the potential accuracy of a model. These tests can help identify source placement, as well as other problems.

LOADS

Many ham antennas, in particular electrically short ones, employ some sort of *loading* to resonate the system. Sometimes loading takes the form of *capacitance hats*, but these can and should be modeled as wires connected to the top of a vertical radiator. A capacitance hat is not the type of loading we'll explore in this section.

Here, the term *loads* refers to discrete inductances, capacitances and resistances that are placed at some point (or points) in an antenna system to achieve certain effects. One fairly common form of a load is a *loading coil* used to resonate an electrically short antenna. Another form of load often seen in ham antennas is a *trap*. *EZNEC* has a special built-in function to evaluate parallel-resonant traps, even at different frequencies beyond their main parallel resonance.

Just for reference, a more subtle type of load is a *distributed* material load. We encountered just such a load in our first model antenna, the 135-foot long flattop dipole—although we didn't identify it specifically as a load at that time. Instead, it was identified as a "wire loss" associated with copper.

The *NEC-2* core program has the capability of simulating a number of built-in loads, including distributed material and discrete loads. *EZNEC* implements the following discrete loads:

- Series R ± *j* X loads.
- Series R-L-C loads, specified in Ω of resistance, µH of inductance and pF of capacitance.
- Parallel R-L-C loads, specified in Ω of resistance, µH of inductance and pF of capacitance.
- Trap loads, specified in Ω of resistance in series with µH of inductance, shunted by pF of capacitance, at a specific frequency.
- Laplace loads, specified as mathematical Laplace coefficients (sometimes used in older modeling programs and left in *EZNEC* for backwards compatibility).

It is important to recognize that the discrete loads in an antenna modeling program *do not radiate* and they have zero size. The *NEC-2* discrete loads are described by L. B. Cebik in his antenna modeling course as being *mathematical loads*. The fact that *NEC-2* loads do not radiate means that the popular mobile antennas that use helical loading coils wound over a length of fiberglass whip cannot be modeled with *NEC-2*, because such coils do radiate.

Let's say that we want to put a air-wound loading coil with an unloaded Q of 400 at the center of a 40-foot long, 50-foot high, flattop dipole so that it is resonant at 7.1 MHz. The schematic of this antenna is shown in **Fig 18**. Examine the modeling file **Ch4-Loaded Dipole.EZ** to see how a discrete series RL load is used to resonate this short dipole at 7.1 MHz, with a feed-point (source) impedance of 25.3 Ω. This requires a series resistance of 1.854 Ω and an inductive reactance of +741.5 Ω. Note that we again used a single wire to model this antenna, and that we placed the load at a point 50% along the length of the wire.

This load represents a 16.62 µH coil with an unloaded Q of 741.5/1.854 = 400, just what we wanted. Let's assume for now that we use a perfect transformer to transform the 25.3-Ω source impedance to 50 Ω. If we now attempt to run a frequency sweep over the whole 40-meter band from 7.0 to 7.3 MHz, the load reactance and resistance will not change, since we specified fixed values for reactance and resistance. Hence, the source impedance will be correct only at the frequency where the reactance and resistance are specified, since the reactance changes with frequency.

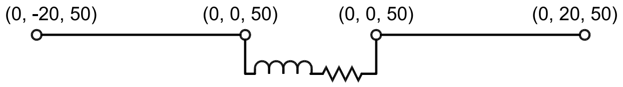

Fig 18—Schematic diagram of a 40-foot long flattop dipole with a loading coil placed at the center. This coil has an unloaded Q of 400 at 7.1 MHz.

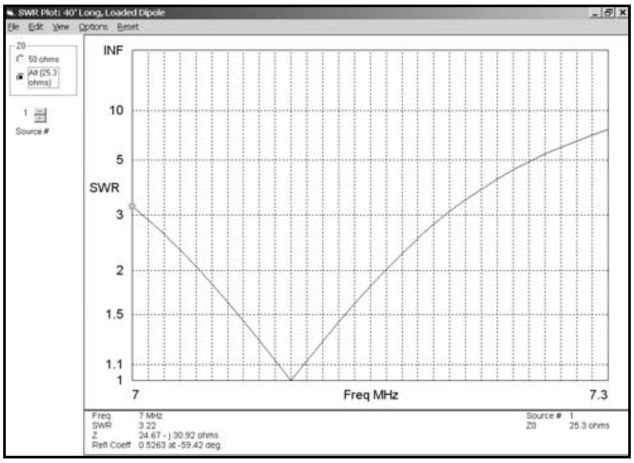

```
Freq        7 MHz                                   Source #   1
SWR         3.22                                    Z0         25.3 ohms
Z           24.67 - j 30.92 ohms
Refl Coeff  0.5263 at -59.42 deg.
```

Fig 19—SWR graph of the loaded 40-foot long flattop dipole shown in Fig 18.

So let's use another load capability and substitute a 16.62 µH coil with a series 1.854-Ω resistance at 7.1 MHz. We'll let *EZNEC* take care of the details of computing both the reactance and the changing series resistance at various frequencies. The degree that both reactance and series loss resistance of the coil change with frequency may be viewed using the **Load Dat** button from the main *EZNEC* window.

Fig 19 shows the computed SWR curve for a 25.3-Ω **Alt SWR Z0** reference resistance. The 2:1 SWR bandwidth is about 120 kHz. As could be expected, the antenna has a rather narrow bandwidth because it is electrically short.

ACCURACY TESTS

There are two tests that can help identify accuracy problems in a model:

- The Convergence test.
- The Average Gain test.

Convergence Test

The idea behind the *Convergence* test is simple: If you increase the segmentation in a particular model and the results changes more than you'd like, then you increase the segmentation until the computations converge to a level that is suitable to you. This process has the potential for being subjective, but simple antenna models do converge quickly. In this section, we'll review several more of the antennas discussed previously to see how they converge.

Let's go back to the simple dipole in Fig 3. The original segmentation was 11 segments, but we'll start with a very low value of segmentation of three, well below the minimum recommended level. **Table 5** shows how the source impedance and gain change with increase in segmentation at 3.75 MHz. For this simple antenna, the gain levels off at 6.50 dBi by the time the segmentation has reached 11 segments. Going to ten times the minimum-recommended level (to 111 segments) results in an increase of only 0.01 dBi in the gain.

Arguably, the impedance has also stabilized by the time we reach a segmentation level of 11 segments, although purists may opt for 23 segments. The tradeoff is a slowdown in computational speed.

Let's see how the 5-element Yagi model converges with changes in segmentation level. **Table 6** shows how the source impedance, gain, 180° front-to-back ratio and worst-case front-to-rear ratio change with segmentation density. By the time the segmentation has reached 11 segments per wire, the impedance and gain have stabilized quite nicely, as has the F/R. The 180° F/B is still increasing with segmentation level until about 25 segments, but a relatively small shift in frequency will change the maximum F/B level greatly. For example, with 11 segments per wire, shifting the frequency to 14.1 MHz—a shift of only 0.5%— will change the maximum 180° F/B from almost 50 dB down to 27 dB. For this reason the F/R is considered a more reliable indicator of the adequacy of the segmentation level than is F/B.

Average Gain Test

The theory behind the *Average Gain* test is a little more involved. Basically, if you remove all intentional losses in a model, and if you place the antenna either in free space or over perfect ground, then all the power fed to the antenna should be radiated by it. Internally, the program runs a full 3-D analysis, adding up the power in all directions and then dividing that sum by the total power fed to the antenna. Since *NEC-2* is very sensitive about source placement, as mentioned before, the Average Gain test is a good indicator that something is wrong with the specification of the source.

Various commercial versions of *NEC-2* handle the Average Gain test in different ways. *EZNEC* requires the operator to turn off all distributed losses in wires or set to zero any discrete resistive losses in loads. Next you set the ground environment to free space (or perfect ground) and request a 3-D pattern plot. *EZNEC* will then report the average gain, which will be 1.000 if the model has no problems. The average gain can be lower or higher than 1.000, but if it falls within the range 0.95 to 1.05 it is usually considered adequate.

Table 5
135-Foot Flattop Dipole at 3.75 MHz

Segments	*Source Impedance Ω*	*Max. Gain dBi*
3	85.9 + *j* 128.0	6.34
5	86.3 + *j* 128.3	6.45
7	86.8 + *j* 128.8	6.48
11	87.9 + *j* 129.5	6.50
23	88.5 + *j* 130.3	6.51
45	89.0 + *j* 130.8	6.51
101	89.4 + *j* 131.1	6.51

Table 6
5-element Wire Yagi at 14.1739 MHz

Segments	*Source Impedance Ω*	*Max. Gain dBi*	*180° F/B dB*	*F/R dB*
3	28.5 – *j* 30.6	12.79	23.2	22.4
5	26.3 – *j* 25.6	13.02	30.5	23.1
7	25.6 – *j* 24.0	13.07	34.8	23.1
11	25.1 – *j* 22.9	13.09	39.9	23.1
25	24.9 – *j* 22.0	13.10	43.7	23.1
99	24.7 – *j* 21.5	13.10	44.2	23.1

As L. B. Cebik, W4RNL, stated in his ARRL Certification and Continuing Education Course on antenna modeling: "Like the convergence test, the average gain test is a necessary but not a sufficient condition of model reliability." Pass both tests, however, and you can be pretty well sure that your model represents reality. Pass only one test, and you have reason to worry about how well your model represents reality.

Once again, open the model file **Ch4-Mod Inverted V Poor Segmentation.EZ** and set **Wire Loss** to zero, **Ground Type** to Free Space and **Plot Type** to 3-Dimensional. Click on the **FF Plot** button. *EZNEC* will report that the Average Gain is 0.955 = –0.2 dB. This is very close to the lower limit of 0.95 considered valid for excellent accuracy. This is a direct result of forcing the segment lengths adjacent to the source segment to be considerably shorter than the source segment's length. The gain reported using this test would be approximately –0.2 dB from what it should be—just what Table 4 alludes to also.

Now, let's revisit the basic model **Ch4-Inverted V Dipole.EZ** and look at Case 2 in Table 4. Case 2 amounts to a Convergence test for the basic inverted-V model. Since the impedance and gain changes were small comparing the basic model to the one using three times the number of segments, the model passed the Convergence test. The Average Gain test for the basic model yields a value of 0.991, well within the limits for good accuracy. This model has thus passed both tests and can be considered accurate.

Running the Average Gain test for the 5-element Yagi (using 11 segments per wire and whose convergence we examined in Table 6) yields a value of 0.996, again well within the bounds indicating a good model. And the simple flattop dipole with 11 segments at 3.75 MHz yields an Average Gain result of 0.997, again indicating a very accurate model.

OTHER POSSIBLE MODEL LIMITATIONS

Programs based on the *NEC-2* core computational code have several well-documented limitations that you should know about. Some limitations have been removed in the restricted-access *NEC-4* core (which is not generally available to users), but other limitations still exist, even in *NEC-4*.

Closely Spaced Wires

If wires are spaced too close to each other, the *NEC-2* core can run into problems. If the segments are not carefully aligned, there also can be problems with accuracy. The worst-case situation is where two wires are so close together that their volumes actually merge into each other. This can happen where wires are thick, parallel to each other and close together. You should keep parallel wires separated by at least several diameters.

For example, #14 wire is 0.064 inches in diameter. The rule then is to keep parallel #14 wires separated by more than 2 × 0.064 = 0.128 inches. And you should run the Convergence test to assure yourself that the solution is indeed converging when you have closely spaced wires, especially if the two wires have different diameters. To model antennas containing closely spaced wires, very often you will need many more segments than usual and you must also carefully ensure that the segments line up with each other.

Things can get a little more tricky when wires cross over or under each other, simply because such crossings are sometimes difficult to visualize. Again, the rule is to keep crossing wires separated by more than two diameters from each other. And if you intend to join two wires together, make sure you do so at the ends of the two wires, using identical end coordinates. When any or all of these rules are violated, the Convergence and Average Gain tests will usually warn you of potential inaccuracies.

Parallel-Wire Transmission Lines and LPDAs

A common example of problems with closely spaced wires is when someone attempts to model a parallel-wire transmission line. *NEC-2*-based programs usually do not work as well in such situations as do *MININEC*-based programs. The problems are compounded if the diameters are different for the two wires simulating a parallel-wire transmission line. In *NEC-2* programs, it is usually better to use the built-in "perfect transmission line" function than to try to model closely spaced parallel wires as a transmission line.

For example, a Log Periodic Dipole Array (LPDA) is composed of a series of elements fed using a transmission line that reverses the phase 180° at each element. In other words, the elements are connected to a transmission line that reverses connections left-to-right at each element. It is cumbersome to do so, but you could model such a transmission line using separate wires in *EZNEC*, but it is a potentially confusing and a definitely painstaking process. Further, the accuracy of the resulting model is usually suspect, as shown by the Average Gain test.

It is far easier to use the **Trans Lines** function from the *EZNEC* main window to accurately model an LPDA. See **Fig 20**, which shows the **Trans Lines** window for the **9302A.EZ** 16-element LPDA. There are 15 transmission lines connecting the 16 elements, placed at the 50% point on each element, with a 200-Ω characteristic impedance and with Reversed connections.

Fat Wires Connected to Skinny Wires

Another inherent limitation in the *NEC-2* computational core shows up when modeling several popular ham-radio antennas: many Yagis and some quads.

Tapered Elements

As mentioned before, many Yagis are built using telescoping aluminum tubing. This technique saves weight and makes for a more flexible and usually stronger element, one that can survive wind and ice loading better than a "monotaper" element design. Many vertical antennas are also constructed using telescoping aluminum tubing.

Unfortunately, native *NEC-2* doesn't model accurately such *tapered elements*, as they are commonly called. There is, however, a sophisticated and accurate work-around for such elements, called the *Leeson corrections*. The Leeson corrections, derived by Dave Leeson, W6NL, from pioneering work by Schelkunoff at Bell Labs, compute the diameter and length of an element that is electrically equivalent to a tapered element. This monotaper element is much easier to use in a pro-

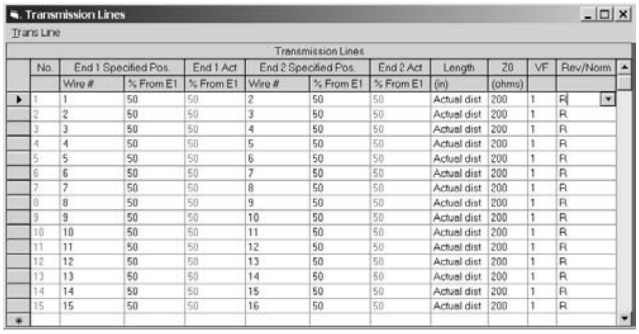

Transmission Lines

No.	End 1 Specified Pos.		End 1 Act	End 2 Specified Pos.		End 2 Act	Length	Z0	VF	Rev/Norm
	Wire #	% From E1	% From E1	Wire #	% From E1	% From E1	(in)	(ohms)		
1	1	50	50	2	50	50	Actual dist	200	1	R
2	2	50	50	3	50	50	Actual dist	200	1	R
3	3	50	50	4	50	50	Actual dist	200	1	R
4	4	50	50	5	50	50	Actual dist	200	1	R
5	5	50	50	6	50	50	Actual dist	200	1	R
6	6	50	50	7	50	50	Actual dist	200	1	R
7	7	50	50	8	50	50	Actual dist	200	1	R
8	8	50	50	9	50	50	Actual dist	200	1	R
9	9	50	50	10	50	50	Actual dist	200	1	R
10	10	50	50	11	50	50	Actual dist	200	1	R
11	11	50	50	12	50	50	Actual dist	200	1	R
12	12	50	50	13	50	50	Actual dist	200	1	R
13	13	50	50	14	50	50	Actual dist	200	1	R
14	14	50	50	15	50	50	Actual dist	200	1	R
15	15	50	50	16	50	50	Actual dist	200	1	R

Fig 20—Transmission-line window for the 9302A.EZ 16-element LPDA. Note that the transmission lines going between elements are "reversed," meaning that they are 180° out-of-phase at each element, a requirement for properly feeding an LPDA.

Table 7

5-element Yagi at 14.1739 MHz with Telescoping Aluminum Elements

	With Leeson Corrections			Without Leeson Corrections		
Freq. MHz	*Source Impedance Ω*	*Gain dBi*	*F/R dB*	*Source Impedance Ω*	*Gain dBi*	*F/R dB*
14.0	23.2 – *j* 26.5	14.82	23.3	22.4 – *j* 12.7	14.92	23.1
14.1	22.7 – *j* 20.5	14.87	22.8	18.6 – *j* 12.5	14.70	21.6
14.2	22.8 – *j* 14.8	14.87	22.7	6.6 – *j* 4.6	14.01	16.2
14.3	22.5 – *j* 11.9	14.76	21.5	1.9 + *j* 10.6	10.61	3.1
14.4	14.5 – *j* 10.5	14.45	19.9	1.6 + *j* 23.7	11.15	–11.4

gram like *NEC-2*. See Chapter 2, Antenna Fundamentals, for more information on the Leeson corrections.

EZNEC and other *NEC-2* programs can automatically invoke the Leeson corrections, providing that some basic conditions are met—and happily, these conditions are true for the telescoping aluminum-tubing elements commonly used as Yagi elements. *EZNEC* gives you the ability to disable or enable Leeson corrections, under the **Option** menu, under **Stepped Diameter Correction**, *EZNEC*'s name for the Leeson corrections. Open the modeling file **520-40H.EZ**, which contains tapered aluminum tubing elements and compare the results using and without using the Leeson corrections.

Table 7 lists the differences over the 20-meter band, with the 5-element Yagi at a height of 70 feet above flat ground. You can see that the non-Leeson corrected figures are very different from the corrected ones. At 14.3 MHz, the pattern for the non-corrected Yagi has degenerated to a F/R of 3.1 db, while at 14.4 MHz, just outside the top of the Amateur band, the pattern for the non-corrected antenna actually has reversed. Even at 14.2 MHz, the non-corrected antenna shows a low source impedance, while the corrected version exhibits smooth variations in gain, F/R and impedance across the whole band, just as the actual antenna exhibits.

Some Quads

Some types of cubical quads are made using a combination of aluminum tubing and wire elements, particularly in Europe where the "Swiss" quad has a wide following. Again, *NEC-2*-based programs don't handle such tubing/wire elements well. It is best to avoid modeling this type of antenna, although there are some ways to attempt to get around the limitations, ways that are beyond the scope of this chapter.

NEAR-FIELD OUTPUTS

FCC regulations set limits on the maximum permissible exposure (MPE) allowed from the operation of radio transmitters. These limits are expressed in terms of the electric (V/m) and magnetic fields (A/m) close to an antenna. *NEC-2*-based programs can compute the electric and magnetic *near fields* and the FCC accepts such computations to demonstrate that an installation meets their regulatory requirements. See Chapter 1, Safety, in this book.

Table 8

E- and H-Field Intensities for 1500 W into 5-Element Yagi at 70 Feet on 14.2 MHz

Height Feet	*H-Field (A/m)*	*E-Field (V/m)*
0	0.04	4.1
10	0.03	13.8
20	0.04	20.6
30	0.06	22.6
40	0.08	25.8
50	0.10	33.8
60	0.12	41.5
70	0.12	44.3

We'll continue to use the 5-element Yagi at 70 feet to demonstrate a near-field computation. Open **Ch4-520-40H.EZ** in *EZNEC* and choose **Setups** and then **Near Field** from the menu at the top of the main window. Let's calculate the E-field and H-field intensity for a power level of 1500 W (chosen using the **Options**, **Power Level** choices from the main menu) in the main beam at a fixed distance, say 50 feet, from the tower base. We'll do this at various heights, using 10-foot increments of height, in order to see the lobe structure of the Yagi at 70 feet height.

Table 8 summarizes the total H- and E-field intensities as a function of height. As you might expect, the fields are strongest directly in line with the antenna at a height of 70 feet. At ground level, the total fields are well within the FCC limits for rf exposure for both fields. In fact, the fields are within the FCC limits if someone were to stand at the tower base, directly under the antenna.

ANTENNA MODELING SUMMARY

This section on antenna modeling is by necessity only a brief introduction to the science of antenna modeling. The subject is partly art as well as science because there are usually several ways of creating a model for a particular antenna or antennas.

Indeed, the presence of other wires surrounding a particular antenna can affect the performance of that antenna. Finally, there are the practical aspects of putting a actual antenna up in the real world. We'll explore this next.

Practical Aspects, Designing Your Antenna System

The most important time spent in putting together an antenna system is the time spent in planning. In Chapter 3, The Effects of Ground, we outlined the steps needed to evaluate how your local terrain can affect HF communications. There we emphasized that you need to compare the patterns resulting from your own terrain to the statistically relevant elevation angles needed for coverage of various geographic areas. (The elevation-angle statistics were developed in Chapter 23, Radio Wave Propagation and are located on the CD-ROM included with this book, as is the terrain-assessment program *HFTA*.)

The implicit assumptions in Chapter 3 are (1) that you know where you want to talk to, and (2) that you'd like the most effective system possible. At the start of such a theoretical analysis, cost is no object. Practical matters, like cost or the desires of your spouse, can come later! After all, you're just checking out all the possibilities. If nothing else, you will use the methodology in Chapter 3 to evaluate any property you are considering buying so that you can build your "dream station."

Next, in the first part of this chapter we described modeling tools used to evaluate different types of antennas. These modeling tools can help you evaluate what type of antenna might be suitable to your own particular style of operating. Do you want a Yagi with a lot of rejection of received signals from the rear? Let's say that terrain analysis shows that you need an antenna at least 50 feet high. Do you really need a steel tower, or would a simple dipole in the trees serve your communication needs just fine? How about a vertical in your backyard? Would that be inconspicuous enough to suit your neighbors and your own family, yet still get you on the air?

In short, using the techniques and tools we've presented in Chapters 3, 23 and here in Chapter 4, you can scientifically plan an antenna *system* that will be best suited for your own particular conditions. Now, however, you have to get practical. Thinking through and planning the installation can save a lot of time, money and frustration. While no one can tell you the exact steps you should take in developing your own master plan, this section, prepared originally by Chuck Hutchinson, K8CH, should help you with some ideas.

WHAT DO YOU REALLY WANT?

Begin planning by spelling out your communications *desires*. What bands are you interested in? Who (or where) do you want to talk to? When do you operate? How much time and money are you willing to spend on an antenna system? What physical limitations affect your master plan?

From the answers to the above questions, begin to formulate *goals*—short, intermediate and long range. Be realistic about those goals. Remember that there are three station effectiveness factors that are under your control. These are: operator skill, equipment in the shack, and the antenna system. There is no substitute for developing operating skills. Some tradeoffs are possible between shack equipment and antennas. For example, a high-power amplifier can compensate for a less than optimum antenna. By contrast, a better antenna has advantages for receiving as well as for transmitting.

Consider your *limitations*. Are there regulatory restrictions on antennas in your community? Are there any deed restrictions or covenants that apply to your property? Do other factors (finances, family considerations, other interests, and so forth) limit the type or height of antennas that you can erect? All of these factors must be investigated because they play a major role determining the type of antennas you erect.

Chances are that you won't be able to immediately do all you desire. Think about how you can budget your resources over a period of time. Your resources are your money, your time available to work, materials you may have on hand, friends that are willing to help, etc. One way to budget is to concentrate your initial efforts on a given band or two. If your major interest is in chasing DX, you might want to start with a very good antenna for the 14-MHz band. A simple multiband antenna could initially serve for other frequencies. Later you can add better antennas for those other bands.

SITE PLANNING

A map of your property or proposed antenna site can be of great help as you begin to consider alternative antennas. You'll need to know the size and location of buildings, trees and other major objects in the area. Be sure to note compass directions on your map. Graph or quadrille paper (or a simple CAD program) can be very useful for this purpose. See **Fig 21** for an example. It's a good idea to make a few photocopies of your *site map* so you can mark on the copies as you work on your plans.

Use your map to plan antenna layouts and locations of any supporting towers or masts. If your plan calls for more than one tower or mast, think about using them as supports for wire antennas. As you work on a layout, be sure to think in three dimensions even though the map shows only two.

Be sensitive to your neighbors. A 70-foot guyed tower in the front yard of a house in a residential neighborhood is not a good idea (and probably won't comply with local ordinances!). You probably will want to locate that tower in the back yard.

ANALYSIS

Use the information earlier in this chapter and in Chapters 3 and 23 to analyze antenna patterns in both

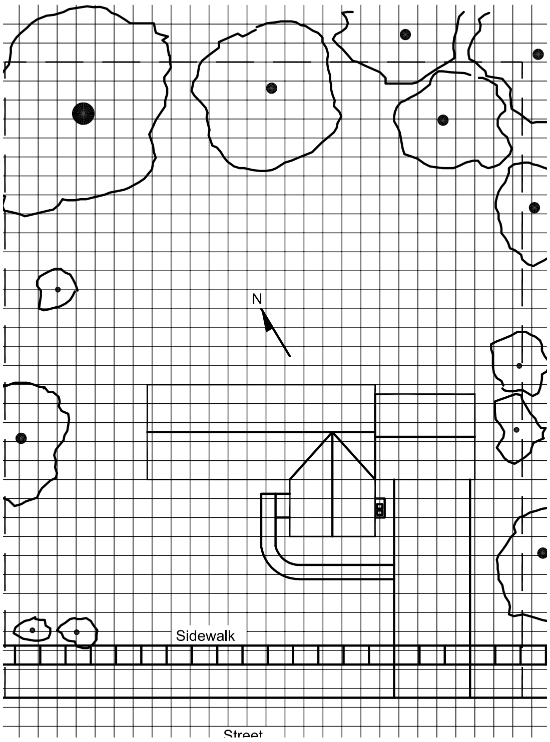

Fig 21—A site map such this one is a useful tool for planning your antenna installation.

horizontal and vertical planes towards geographic areas of interest. If you want to work DX, you'll want antennas that radiate energy at low as well as intermediate angles. An antenna pattern is greatly affected by the presence of ground and by the local topography of the ground. Therefore, be sure to consider what effect ground will have on the antenna pattern at the height you are considering. A 70-foot high antenna is approximately $^1/_2$, 1, $1^1/_2$ and 2 wavelengths (λ) high on 7, 14, 21 and 28 MHz respectively. Those heights are useful for long-distance communications. The same 70-foot height represents only λ/4 at 3.5 MHz, however. Most of the radiated energy from a dipole at that height would be concentrated straight up. This condition is not great for long-distance communication, but can still be useful for some DX work and excellent for short-range communications.

Lower heights can be useful for communications. However, it is generally true that "the higher, the better" as far as communications effectiveness is concerned. This general rule of thumb, of course, should be tempered by an exact analysis of your local terrain. Being located at the top of a steep hill can mean that you can use lower tower heights to achieve good coverage.

There may be cases where it is not possible to install low-frequency dipoles at λ/4 or more above the ground. A vertical antenna with many radials is a good choice for long-distance communications. You may want to install both a dipole and a vertical for the 3.5- or 7-MHz bands. On the 1.8-MHz band, unless extremely tall supports are available, a vertical antenna is likely to be the most useful for DXing. You can then choose the antenna that performs best for a given set of conditions. A low dipole will generally work better for shorter-range communications, while the vertical will generally be the better performer over longer distances.

Consider the azimuthal pattern of fixed antennas. You'll want to orient any fixed antennas to favor the directions of greatest interest to you.

BUILDING THE SYSTEM

When the planning is completed, it is time to begin construction of the antenna system. Chances are that you can divide that construction into a series of *phases* or steps. Say, for example, that you have lots of room and that your long-range plan calls for a pair of towers, one 100-feet high, and the other 70-feet high, to support monoband Yagi antennas. The towers will also support a horizontal 3.5-MHz dipole, for DX work. On your map you've located them so the 80-meter dipole will be broadside to Europe. You decide to build the 70-foot tower with a "triband" beam and 80- and 40-meter inverted-V dipoles to begin the project.

In your master plan you design the guys, anchors and all hardware for the 70-foot tower to support the load of stacked 4-element 10- and 15-meter monobanders Yagis. So you make sure you buy a heavy-duty rotator and the stout mast needed for the monoband antennas later. Thus you avoid having to buy, and then sell, a medium-duty rotator and lighter weight tower equipment later on when you upgrade the station. You could have saved money in the long run by putting up a monoband beam for your favorite band, but you decided that for now it is more important to have a beam on 14, 21 and 28 MHz, so you choose a commercial triband Yagi.

The second step of your plan calls for installing the second tower and stacking a 2-element 40-meter and a 4-element 20-meter monoband Yagi on it. You also plan to replace the tribander on the 70-foot tower with stacked 4-element 10- and 15-meter monoband Yagis. Although this is still a "dream system" you can now apply some of the modeling techniques discussed earlier in this chapter to determine the overall system performance.

Modeling Interactions at Your Dream Station

In this analysis we're going to assume that you have sufficient real estate to separate the 70- and 100-foot towers by 150 feet so that you can easily support an 80-meter dipole between them. We'll also assume that you want the 80-meter dipole to have its maximum response at a heading of 45° into Europe from your location in Newington, Connecticut. The dipole will also have a lobe

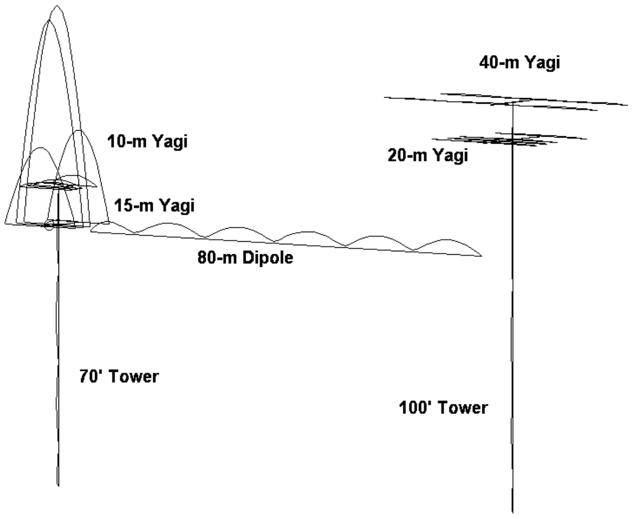

Fig 22—Layout for two-tower antenna system, at 70 and 100 feet high and 150 feet apart. The 70-foot tower has a 4-element 10-meter Yagi at 80 feet on a 10-foot rotating mast and a 4-element 15-meter Yagi at 70 feet. An 80-meter dipole goes from the 70-foot tower to the 100-foot tower, which holds a 2-element 40-meter Yagi at 110 feet and a 4-element 20-meter Yagi at 100 feet. In this figure all the rotatable Yagis are facing the direction of Europe and the currents on the 15-meter Yagi are shown. Note the significant amount of current re-radiated by the nearby 80-meter dipole.

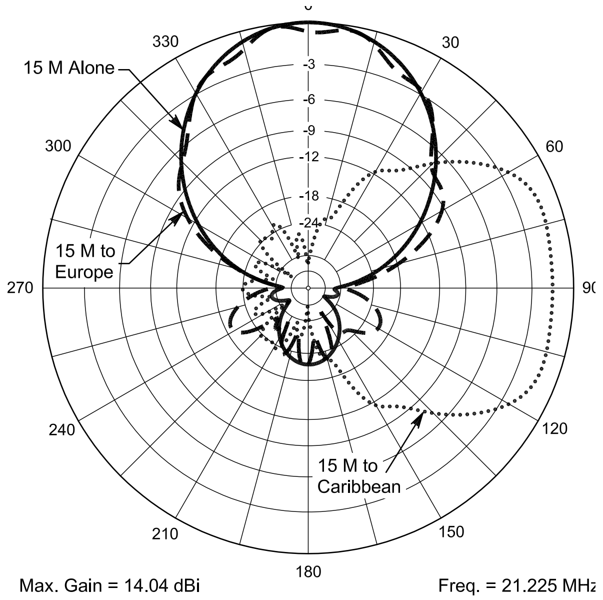

Fig 23—An overlay of azimuth patterns. The solid line is the radiation pattern for the 15-meter Yagi all by itself. The dashed line is the pattern for the 15-meter Yagi, as affected by all the other antennas. The dotted line is the pattern for the 15-meter Yagi when it is pointed toward the Caribbean, with the Yagis on the 100-foot tower pointed toward the 70-foot tower. The peak response of the 15-meter Yagi has dropped by about 1.5 dB.

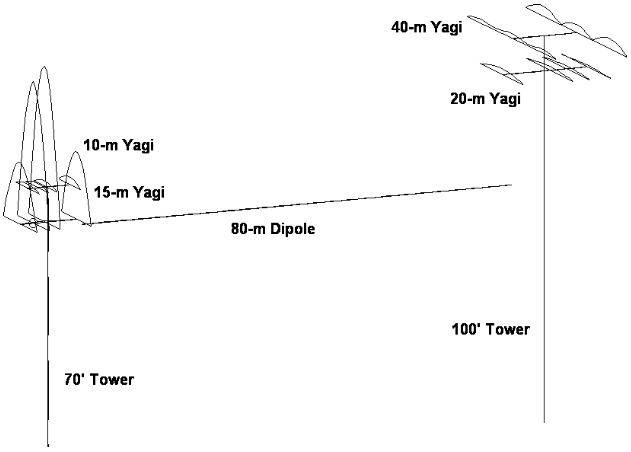

Fig 24—The layout and 15-meter currents when the Yagis on the 100-foot tower are pointed toward the 70-foot tower. The 15-meter Yagi has been rotated to face the direction of the 100-foot tower (toward the Caribbean).

facing 225° towards the USA and New Zealand, making it a good antenna for both domestic contacts and DX work.

Let's examine the interactions that occur between the rotatable Yagis for 10, 15, 20 and 40 meters. See **Fig 22**, which purposely exaggerates the magnitude of the currents on the 4-element 15-meter Yagi mounted at 70 feet. Here, both sets of Yagis have been rotated so that they are pointing into Europe. There is a small amount of current radiated onto the 10-meter antenna but virtually no current is radiated onto the 40- and 20-meter Yagis. This is good.

However, significant current is radiated onto, and then re-radiated, by the 80-meter dipole. This undesired current affects the radiation pattern of the 15-meter antenna, as shown in **Fig 23**, which overlays the pattern of the 4-element 15-meter Yagi by itself with that of the Yagi interacting with the other antennas. You can see "ripples" in the azimuthal response of the 15-meter Yagi due to the effects of the 80-meter dipole's re-radiation. The magnitude of the ripples is about 1 dB at worst, so they don't seriously affect the forward pattern (into Europe), but the rearward lobes are degraded somewhat, to just below 20 dB.

Fig 23 also shows the worst-case situation for the 15-meter Yagi. Here, the 15- and 10-meter stack has been turned clockwise 90°, facing the Caribbean, while the 40- and 20-meter Yagis on the 100-foot tower have been turned counter-clockwise 90° (in the direction of Japan) to face the 70-foot tower holding the 10/15-meter Yagis. You can see the layout and the currents in **Fig 24**. Now the 40- and 20-meter Yagis re-radiate some 15-meter

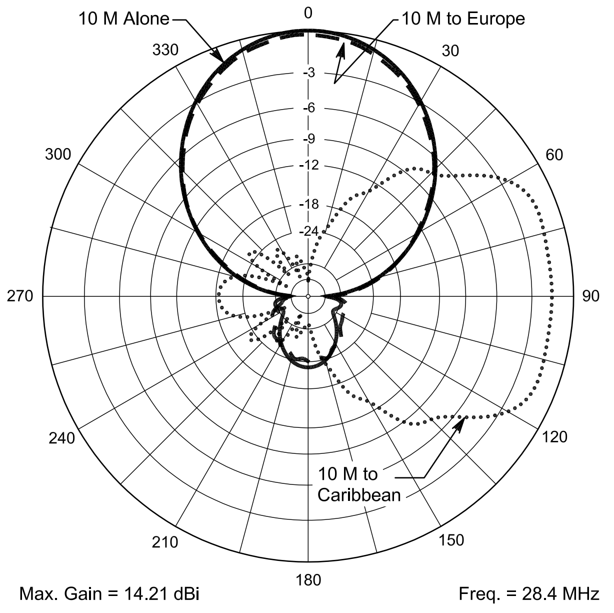

Fig 25—The radiation patterns for the 10-meter Yagi. The solid line is the 10-meter Yagi by itself. The dashed line is for the same Yagi, with all other antenna interactions. The dotted line shows the worst-case pattern, with the stacked Yagis on the 100-foot tower facing the 70-foot tower and the 10-meter Yagi pointed toward the Caribbean. Again, the peak response of the 10-meter Yagi has dropped about 1.5 dB in the worst-case situation.

energy and reduce the maximum gain by about 1.5 dB. Note that in this direction the 80-meter dipole no longer has 15-meter energy radiated onto it by the 15-meter Yagi.

The shape of the patterns will change depending on whether you specify "current" or "voltage" sources in the models for the other antennas, since this effectively opens up or shorts the feed points at the other antennas so far as 15-meter energy is concerned. In practice, this means that the interaction between antennas will vary somewhat depending on the length of the feed lines going to each antenna and whether each feed line is open-circuited or short-circuited when it is not in use.

You can now see that interactions between various antennas pointing in different directions can be significant in a real-world antenna system. In general, higher-frequency antennas are affected by re-radiation from lower-frequency antennas, rather than the other way around. Thus the presence of a 10- or 15-meter stack does not affect the 20-meter Yagi at all.

Modeling can also help determine the minimum stacking distance required between monoband Yagis on the same rotating mast. In this case, stacking the 10- and 15-meter monobanders 10 feet apart holds down interaction between them so that the pattern and gain of the 10-meter Yagi is not impacted adversely. **Fig 25** demonstrates this in the European direction, where the patterns for the 10-meter beam by itself looks very clean compared to the same Yagi separated by 10 feet from the 15-meter Yagi below it. The worst-case situation is pointing towards the Caribbean, when the 40- and 20-meter stack is facing the 70-foot tower. This drops the 10-meter gain down about 1.5 dB from maximum, indicating significant interaction is occurring.

In this situation you might find it best to place the 70-foot tower in the direction closest to the Caribbean if this direction is very important to you. Doing so will, however, cause the pattern in the direction of the Far East to be affected on 10 and 15 meters. You have the modeling tools necessary to evaluate various configurations to achieve whatever is most important to you.

COMPROMISES

Because of limitations, most amateurs are never able to build their dream antenna system. This means that some compromises must be made. Do not, under any circumstances, compromise the *safety* of an antenna installation. Follow the manufacturer's recommendations for tower assembly, installation and accessories. Make sure that all hardware is being used within its ratings.

Guyed towers are frequently used by radio amateurs because they cost less than more complicated unguyed or freestanding towers with similar ratings. Guyed towers are fine for those who can climb, or those with a friend who is willing to climb. But you may want to consider an antenna tower that folds over, or one that cranks up (and down). Some towers crank up (and down) and fold over too. See **Fig 26**. That makes for convenient access to antennas for adjustments and maintenance without climbing. Crank-up towers also offer another advantage. They allow antennas to be lowered during periods of no operation, such as for aesthetic reasons or during periods of high winds.

A well-designed monoband Yagi should outperform a multiband Yagi. In a monoband design the best adjustments can be made for gain, front-to-rear ratio (F/R) and matching, *but only for a single band*. In a multiband design, there are always tradeoffs in these properties for the ability to operate on more than one band. Nevertheless, a multiband antenna has many advantages over two or more single band antennas. A multiband antenna requires less heavy-duty hardware, requires only one feed line, takes up less space and it costs less.

Apartment dwellers face much greater limitations in their choice of antennas. For most, the possibility of a tower is only a dream. (One enterprising ham made arrangements to purchase a top-floor condominium from a developer. The arrangements were made before construction began, and the plans were altered to include a roof-top tower installation.) For apartment and condominium dwellers, the situation is still far from hopeless. A later section presents ideas for consideration.

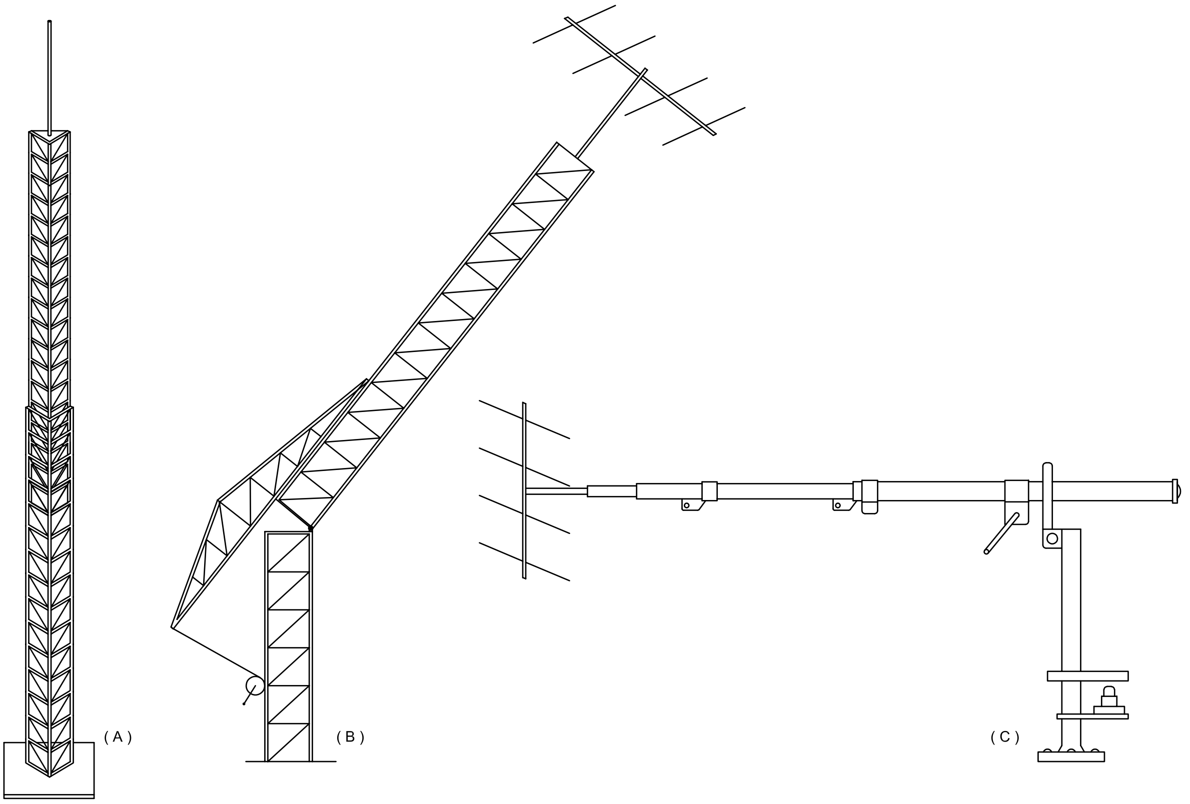

Fig 26—Alternatives to a guyed tower are shown here. At A, the crank-up tower permits working on antennas at reduced height. It also allows antennas to be lowered during periods of no operation. Motor-driven versions are available. The fold-over tower at B and the combination at C permit working on antennas at ground level.

EXAMPLES

You can follow the procedure previously outlined to put together modest or very large antenna systems. What might a ham put together for antennas when he or she wants to try a little of everything, and has a modest budget? Let's suppose that the goals are (1) low cost, (2) no tower, (3) coverage of all HF bands and the repeater portion of one VHF band, and (4) the possibility of working some DX.

After studying the pages of this book, the station owner decides to first put up a 135-foot center-fed antenna. High trees in the back yard will serve as supports to about 50 feet. This antenna will cover all the HF bands by using a balanced feeder and an antenna tuner. It should be good for DX contacts on 10 MHz and above, and will probably work okay for DX contacts on the lower bands. However, her plan calls for a vertical for 3.5 and 7 MHz to enhance the DX possibilities on those bands. For VHF, a chimney-mounted vertical is included.

ANOTHER EXAMPLE

A licensed couple has bigger ambitions. Goals for their station are (1) a good setup for DX on 14, 21 and 28 MHz, (2) moderate cost, (3) one tower, (4) ability to work some DX on 1.8, 3.5 and 7 MHz, and (5) no need to cover the CW portion of the bands.

After considering the options, the couple decides to install a 65-foot guyed tower. A large commercial triband Yagi will be mounted on top of the tower. The center of a trap dipole tuned for the phone portion of the 3.5- and 7-MHz bands will be supported by a wooden yardarm installed at the 60-foot level of the tower, with ends drooping down to form an inverted V. An inverted L for 1.8 MHz starts near ground level and goes up to a similar yardarm on the opposite side of the tower. The horizontal portion of the inverted L runs away from the tower at right angles to the trap dipole. Later, the husband will experiment with sloping antennas for 3.5 MHz. If those experiments are not successful, a $\lambda/4$ vertical will be used on that band.

Apartment Possibilities

A complete and accurate assessment of antenna types, antenna placement and feed-line placement is very important for the apartment dweller. Among the many possibilities for types are balcony antennas, *invisible* ones (made of fine wire), vertical antennas disguised as flag poles or as masts with a TV antenna on top, and indoor antennas.

A number of amateurs have been successful negotiating with the apartment owner or manager for permission to install a short mast on the roof of the building. Coaxial lines and rotator control cables might be routed through conduit troughs or through ductwork. If you live in one of the upper stories of the building, routing the cables over the edge of the roof and in through a window might be the way to go. There is a story about one amateur who owns a triband beam mounted on a 10-foot mast. But even with such a short mast, he is the envy of all his amateur friends because of his superb antenna height. His mast stands on top of a 22-story apartment building.

Usually the challenge is to find ways to install antennas that are unobtrusive. That means searching out antenna locations such as balconies, eaves, nearby trees, etc. For example, a simple but effective balcony antenna is a dangling vertical. Attach a thin wire to the tip of a mobile whip or a length of metal rod or tubing. Then mount the rigid part of the antenna horizontally on the balcony rail, dangling the wire over the edge. The antenna is operated against the balcony railing or other metallic framework. A matching network is usually required at the antenna feed point. Metal in the building will likely give a directivity effect, but this may be of little consequence and perhaps even an advantage. The antenna may be removed and stored when not in use.

Frequently, the task of finding an inconspicuous route for a feed line is more difficult than the antenna installation itself. When Al Francisco, K7NHV, lived in an apartment, he used a tree-mounted vertical antenna. The coax feeder exited his apartment through a window and ran down the wall to the ground. Al buried the section of line that went from under the window to a nearby tree. At the tree, a section of enameled wire was connected to the coax center conductor. He ran the wire up the side of the tree away from foot traffic. A few short radials completed the installation. The antenna worked fine, and was never noticed by the neighbors.

See Chapters 6, Low-Frequency Antennas, and Chapter 15, Portable Antennas, for ideas about low-frequency and portable antennas that might fit into your available space. Your options are limited as much by your imagination and ingenuity as by your pocketbook. Another option for apartment dwellers is to operate away from home. Some hams concentrate on mobile operation as an alternative to a fixed station. It is possible to make a lot of contacts on HF mobile. Some have worked DXCC that way.

Suppose that you like VHF contests. Because of other activities, you are not particularly interested in operating VHF outside the contests. Why not take your equipment and antennas to a hilltop for the contests? Many hams combine a love for camping or hiking with their interest in radio.

Antennas for Limited Space

It is not always practical to erect full-size antennas for the HF bands. Those who live in apartment buildings may be restricted to the use of minuscule radiators because of house rules, or simply because the required space for full-size antennas is unavailable. Other amateurs may desire small antennas for aesthetic reasons, perhaps to keep peace with neighbors who do not share their enthusiasm about high towers and big antennas. There are many reasons why some amateurs prefer to use physically-shortened antennas. This section discusses proven designs and various ways of building and using them effectively. You will find that modeling antennas by computer, even compromised “stealth antennas,” can help you determine the most practical system possible for your particular circumstances—before you go through the effort of stringing up wires.

Few compromise antennas are capable of delivering the performance you can expect from the full-size variety. But the patient and skillful operator can often do as well as some who are equipped with high power and full-size antennas. Someone with a reduced-size antenna may not be able to “bore a hole” in the bands as often and with the commanding dispatch enjoyed by those who are better equipped, but DX can be worked successfully when band conditions are suitable.

INVISIBLE ANTENNAS

We amateurs don’t regard our antennas as eyesores; in fact, we almost always regard them as works of art! But there are occasions when having an outdoor or visible antenna can present problems.

When we are confronted with restrictions—self-imposed or otherwise—we can take advantage of a number of options toward getting on the air and radiating at least a moderately effective signal. In this context, a poor antenna is certainly better than no antenna at all! This section describes a number of techniques that enable us to use indoor antennas or “invisible” antennas outdoors.

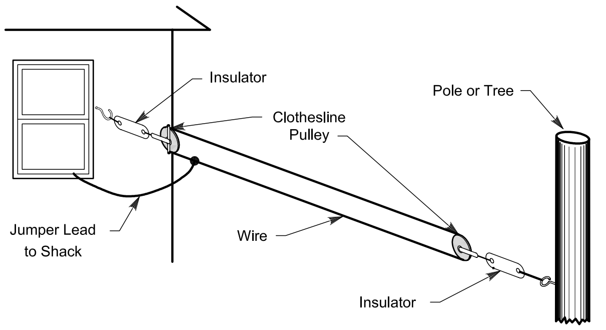

Fig 27—The clothesline antenna is more than it appears to be.

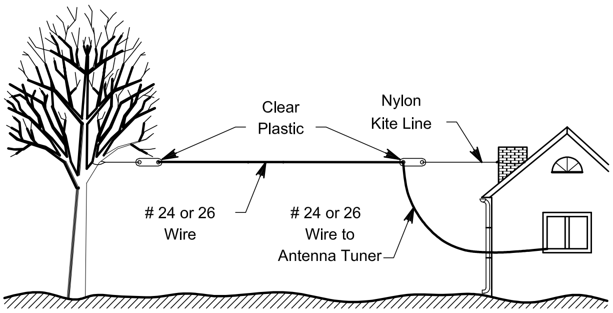

Fig 28—The "invisible" end-fed antenna.

Many of these systems will yield good-to-excellent results for local and DX contacts, depending on band conditions at any given time. *The most important consideration is that of not erecting any antenna that can present a hazard (physical or electrical) to humans, animals and buildings. Safety first!*

Clothesline Antenna

Clotheslines are sometimes attached to pulleys (**Fig 27**) so that the user can load the line and retrieve the laundry from a back porch. Laundry lines of this variety are accepted parts of the neighborhood "scenery," and can be used handily as amateur antennas by simply insulating the pulleys from their support points. This calls for the use of a conducting type of clothesline, such as heavy gauge stranded electrical wire with Teflon or vinyl insulation. A high quality, flexible steel cable (stranded) is suitable as a substitute if you don't mind cleaning it before clothing is hung on it.

A jumper wire can be brought from one end of the line to the ham shack when the station is being operated. If a good electrical connection exists between the wire clothesline and the pulley, a permanent connection can be made by connecting the lead-in wire between the pulley and its insulator. An antenna tuner can be used to match the "invisible" random-length wire to the transmitter and receiver.

Invisible Long Wire

A wire antenna is not actually a "long wire" unless it is one wavelength or greater in length. Yet many amateurs refer to (relatively) long physical spans of conductor as *long wires*. For the purpose of this discussion we will assume we have a fairly long span of wire, and refer to it as an *end-fed* wire antenna.

If we use small-diameter enameled wire for our end-fed antenna, chances are that it will be very difficult to see against the sky and neighborhood scenery. The smaller the wire, the more invisible the antenna will be. The limiting factor with small wire is fragility. A good compromise is #24 or #26 magnet wire for spans up to 130 feet; lighter-gauge wire can be used for shorter spans, such as 30 or 60 feet. The major threat to the longevity of fine wire is icing. Also, birds may fly into the wire and break it. Therefore, this style of antenna may require frequent service or replacement.

Fig 28 illustrates how you might install an invisible end-fed wire. It is important that the insulators also be lacking in prominence. Tiny Plexiglas blocks perform this function well. Small-diameter clear plastic medical vials are suitable also. Some amateurs simply use rubber bands for end insulators, but they will deteriorate rapidly from sun and air pollutants. They are entirely adequate for short-term operation with an invisible antenna, however.

Rain Gutter and TV Antennas

A great number of amateurs have taken advantage of standard house fixtures when contriving inconspicuous antennas. A very old technique is the use of the gutter and downspout system on the building. This is shown in **Fig 29**, where a lead wire is routed to the operating room

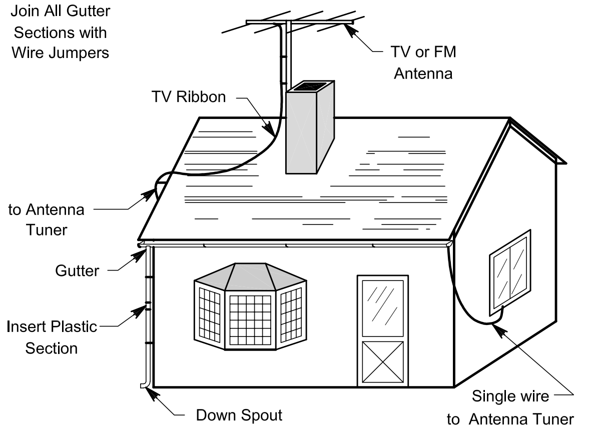

Fig 29—Rain gutters and TV antenna installations can be used as inconspicuous Amateur Radio antennas.

from one end of the gutter trough. We must assume that the wood to which the gutter is affixed is dry and of good quality to provide reasonable electrical insulation. The rain gutter antenna may perform quite poorly during wet weather or when there is ice and snow on it and the house roof.

All joints between gutter and downspout sections must be bonded electrically with straps of braid or flashing copper to provide good continuity in the system. Poor joints can permit rectification of RF and subsequently cause TVI and other harmonic interference. Also, it is prudent to insert a section of plastic downspout about 8 feet above ground to prevent RF shocks or burns to passersby while the antenna is being used. Improved performance may result if you join the front and back gutters of the house with a jumper wire to increase the area of the antenna.

Fig 29 also shows a TV or FM antenna that can be employed as an invisible amateur antenna. Many of these antennas can be modified easily to accommodate the 144- or 222-MHz bands, thereby permitting the use of the 300-Ω line as a feeder system. Some FM antennas can be used on 6 meters by adding #10 bus wire extensions to the ends of the elements, and adjusting the match for an SWR of 1:1. If 300-Ω line is used it will require a balun or antenna tuner to interface the line with the station equipment.

For operation in the HF bands, the TV or FM antenna feeders can be tied together at the transmitter end of the span and the system treated as a random length wire. If this is done, the 300-Ω line will have to be on TV standoff insulators and spaced well away from phone and power company service entrance lines. Naturally, the TV or FM radio must be disconnected from the system when it is used for amateur work! Similarly, masthead amplifiers and splitters must be removed from the line if the system is to be used for amateur operation. If the system is mostly vertical, a good RF ground system with many radials around the base of the house should be used to improve performance.

A very nice top-loaded vertical can be made from a length of TV mast with a large TV antenna on the top. Radials can be placed on the roof or at ground level with the TV “feed line” acting as part of the vertical. There is an extensive discussion of loaded verticals and radial systems in Chapter 6, Low-Frequency Antennas.

Flagpole Antennas

We can exhibit our patriotism and have an invisible amateur antenna at the same time by disguising our antenna as shown in **Fig 30**. The vertical antenna is a wire that has been placed inside a plastic or fiberglass pole.

The flagpole antenna shown is structured for a single amateur band, and it is assumed that the height of the pole corresponds to a quarter wavelength for the chosen band. The radials and feed line can be buried in the ground as shown. In a practical installation, the sealed end of the coax cable would protrude slightly into the lower end of the plastic pole.

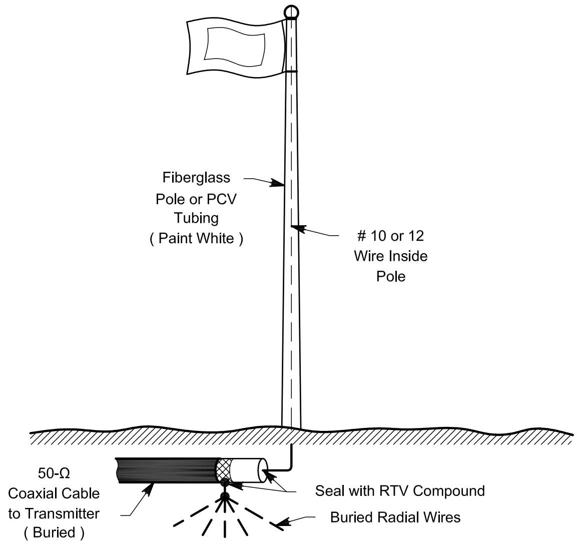

Fig 30—A flagpole antenna.

If a large-diameter fiberglass pole were available, a multiband trap vertical may be concealed inside it. Or you might use a metal pole and bury a water-tight box at its base, containing fixed-tuned matching networks for the bands of interest. The networks could then be selected remotely by means of relays inside the box. A 30-foot flagpole would provide good results in this kind of system, provided it was used in conjunction with a buried radial system.

Still another technique is one that employs a wooden flagpole. A small diameter wire can be stapled to the pole and routed to the coax feeder or matching network. The halyard could by itself constitute the antenna wire if it were made from heavy-duty insulated hookup wire. There are countless variations for this type of antenna, and they are limited only by the imagination of the amateur.

Other Invisible Antennas

Some amateurs have used the metal fence on apartment verandas as antennas, and have had good results on the upper HF bands (14, 21 and 28 MHz). We must presume that the fences were not connected to the steel framework of the building, but rather were insulated by the concrete floor to which they were affixed. These verandah fences have also been used effectively as ground systems (counterpoises) for HF-band vertical antennas put in place temporarily after dark.

One amateur in New York City uses the fire escape on his apartment building as a 7-MHz antenna, and he reports good success working DX stations with it. Another apartment dweller makes use of the aluminum frame on his living room picture window as an antenna for 21 and 28 MHz. He works it against the metal conductors of the baseboard heater in the same room.

Many jokes have been told over the years about *bedspring antennas*. The idea is by no means absurd. Bedsprings and metal end boards have been used to advantage as antennas by many apartment dwellers as 14, 21 and 28 MHz radiators. A counterpoise ground can be routed along the baseboard of the room and used in combination with the bedspring. It is important to remember that any independent (insulated) metal object of reasonable size can serve as an antenna if the transmitter can be matched to it. An amateur in Detroit once used his Shopsmith craft machine (about 5 feet tall) as a 28 MHz antenna. He worked a number of DX stations with it when band conditions were good.

A number of operators have used metal curtain rods and window screens for VHF work, and found them to be acceptable for local communications. Best results with any of these makeshift antennas will be had when the "antennas" are kept well away from house wiring and other conductive objects.

INDOOR ANTENNAS

Without question, the best place for your antenna is outdoors, and as high and in the clear as possible. Some of us, however, for legal, social, neighborhood, family or landlord reasons, are restricted to indoor antennas. Having to settle for an indoor antenna is certainly a handicap for the amateur seeking effective radio communication, but that is not enough reason to abandon all operation in despair.

First, we should be aware of the reasons why indoor antennas do not work well. Principal faults are:

- Low height above ground—the antenna cannot be placed higher than the highest peak of the roof, a point usually low in terms of wavelength at HF
- The antenna must function in a lossy RF environment involving close coupling to electrical wiring, guttering, plumbing and other parasitic conductors, besides dielectric losses in such nonconductors as wood, plaster and masonry
- Sometimes the antenna must be made small in terms of a wavelength
- Usually it cannot be rotated.

These are appreciable handicaps. Nevertheless, global communication with an indoor antenna is still possible, although you must be sure that you are not exposing anyone in your family or nearby neighbors to excessive radiation. See Chapter 1, Safety, in this book.

Some practical points in favor of the indoor antenna include:

- Freedom from weathering effects and damage caused by wind, ice, rain and sunlight (the SWR of an attic antenna, however, can be affected somewhat by a wet or snow-covered roof).
- Indoor antennas can be made from materials that would be altogether impractical outdoors, such as aluminum foil and thread (the antenna need support only its own weight).
- The supporting structure is already in place, eliminating the need for antenna masts.
- The antenna is readily accessible in all weather conditions, simplifying pruning or tuning, which can be accomplished without climbing or tilting over a tower.

Empiricism

A typical house or apartment presents such a complex electromagnetic environment that it is impossible to predict theoretically which location or orientation of the indoor antenna will work best. This is where good old fashioned cut-and-try, use-what-works-best *empiricism* pays off. But to properly determine what really is most suitable requires an understanding of some antenna measuring fundamentals.

Unfortunately, many amateurs do not know how to evaluate performance scientifically or compare one antenna with another. Typically, they will put up one antenna and try it out on the air to see how it "gets out" in comparison with a previous antenna. This is obviously a very poor evaluation method because there is no way to know if the better or worse reports are caused by changing band conditions, different S-meter characteristics or any of several other factors that could influence the reports received.

Many times the difference between two antennas or between two different locations for identical antennas amounts to only a few decibels, a difference that is hard to discern unless instantaneous switching between the two is possible. Those few decibels are not important under strong signal conditions, of course, but when the going gets rough, as is often the case with an indoor antenna, a few dB can make the difference between solid copy and no possibility of real communication.

Very little in the way of test equipment is needed for casual antenna evaluation, other than a communications receiver. You can even do a qualitative comparison by ear, if you can switch antennas instantaneously. Differences of less than 2 dB, however, are still hard to discern. The same is true of S-meters. Signal strength differences of less than a decibel are usually difficult to see. If you want to measure that last fraction of a decibel, you should use a good ac voltmeter at the receiver audio output (with the AGC turned off).

In order to compare two antennas, switching the coaxial transmission line from one to the other is necessary. No elaborate coaxial switch is needed; even a simple double-throw toggle or slide switch will provide more than 40 dB of isolation at HF. See **Fig 31**. Switching by means of manually connecting and disconnecting coaxial lines is not recommended because that takes too long.

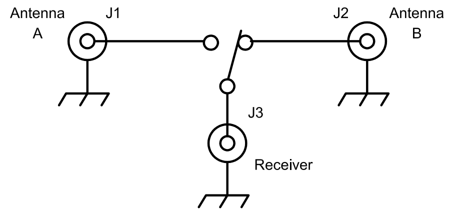

Fig 31—When antennas are compared on fading signals, the time delay involved in disconnecting and reconnecting coaxial cables is too long for accurate measurements. A simple slide switch will do well for switching coaxial lines at HF. The four components can be mounted in a tin can or any small metal box. Leads should be short and direct. J1 through J3 are coaxial connectors.

Fading can cause signal-strength changes during the changeover interval.

Whatever difference shows up in the strength of the received signal will be the difference in performance between the two antennas in the direction of that signal. For this test to be valid, both antennas must have nearly the same feed-point impedance, a condition that is reasonably well met if the SWR is below 2:1 on both antennas.

On ionospheric propagated signals (sky wave) there will be constant fading, and for a valid comparison it will be necessary to take an average of the difference between the two antennas. Occasionally, the inferior antenna will deliver a stronger signal to the receiver, but in the long run the law of averages will put the better antenna ahead.

Of course with a ground-wave signal, such as that from a station across town, there will be no fading problems. A ground-wave signal will enable the operator to properly evaluate the antenna under test in the direction of the source. The results will be valid for ionospheric-propagated signals at low elevation angles in that direction. On 28 MHz, all sky-wave signals arrive and leave at low angles. But on the lower bands, particularly 3.5 and 7 MHz, we often use signals propagated at high elevation angles, almost up to the zenith. For these angles a ground-wave test between local stations may not provide a proper evaluation of the antenna, and use of sky wave signals becomes necessary.

Dipoles

At HF the most practical indoor antenna is usually the dipole. Attempts to get more gain with parasitic elements will usually fail because of close proximity to the ground or coupling to house wiring. Beam antenna dimensions determined outdoors will not usually be valid for an attic antenna because the roof structure will cause dielectric loading of the parasitic elements. It is usually more worthwhile to spend time optimizing the location and performance of a dipole than to try to improve results with parasitic elements.

Most attics are not long enough to accommodate half-wave dipoles for 7 MHz and below. If this is the case, some folding of the dipole will be necessary. The final shape of the antenna will depend on the dimensions and configuration of the attic. Remember that the center of the dipole carries the most current and therefore does most of the radiating. This part should be as high and unfolded as possible. Because the dipole ends radiate less energy than the center, their orientation is not as important. They do carry the maximum voltage, nevertheless, so care should be taken to position the ends far enough from other conductors to avoid arcing.

The dipole may end up being L-shaped, Z-shaped, U-shaped or some indescribable corkscrew shape, depending on what space is available, but reasonable performance can often be had even with such a non-straight arrangement. **Fig 32** shows some possible configurations. Multiband operation is possible with the use of open-wire feeders and an antenna tuner.

One alternative not shown here is the aluminum-foil dipole, which was conceived by Rudy Stork, KA5FSB. He suggests mounting the dipole behind wallpaper or in the attic, with portability, ease of construction and adjustment, and economy in design among its desirable features. This antenna should also display reasonably good bandwidth resulting from the large area of its conductor material. If coaxial feed is used, some pruning of an attic antenna to establish minimum SWR at the band center will be required. Tuning the antenna outdoors and then installing it inside is usually not feasible since the behavior of the antenna will not be the same when placed in the attic. Resonance will be affected somewhat if the antenna is bent.

Even if the antenna is placed in a straight line, parasitic conductors and dielectric loading by nearby wood structures can affect the impedance. Trap and loaded dipoles are shorter than the full-sized versions, but are comparable performers. Trap dipoles are discussed in Chapter 7, Multiband Antennas, and loaded dipoles in Chapter 6, Low-Frequency Antennas.

Dipole Orientation

Theoretically a vertical dipole is most effective at low radiation angles, but practical experience shows that the horizontal dipole is usually a better indoor antenna. A high horizontal dipole does exhibit directional effects at low radiation angles, but you will not be likely to see much, if any, directivity with an attic-mounted dipole. Some operators place two dipoles at right angles to each other with provisions at the operating position for switching between the two. Their reasoning is the radiation patterns will inevitably be distorted in an unpredictable manner by nearby parasitic conductors. There will be little coupling between the dipoles if they are oriented a right

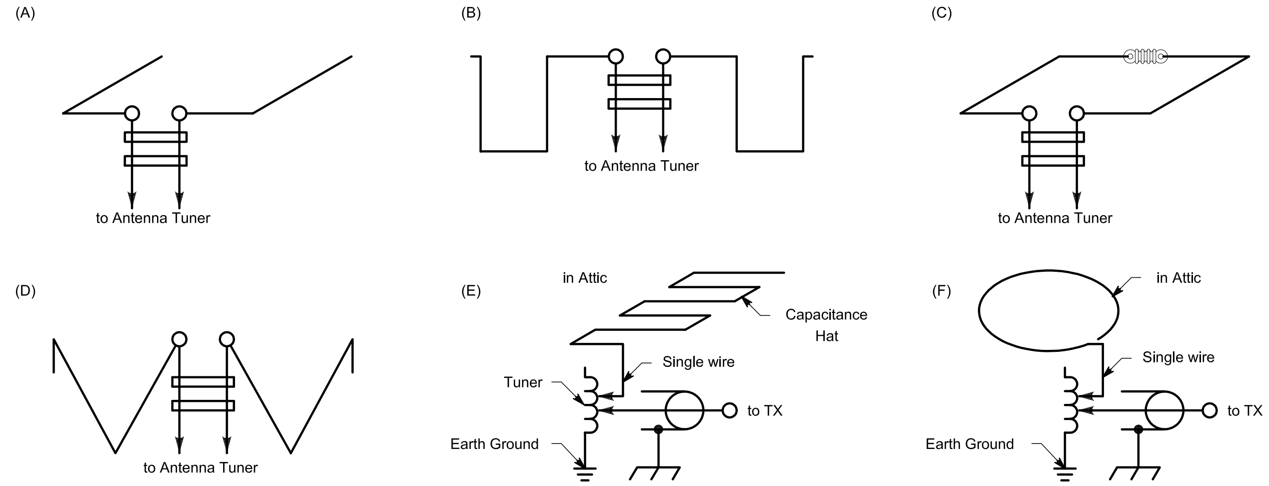
(A)
to Antenna Tuner
(B)
to Antenna Tuner
(C)
to Antenna Tuner
(D)
to Antenna Tuner
(E)
in Attic
Capacitance Hat
Single wire
Tuner
to TX
Earth Ground
(F)
in Attic
Single wire
to TX
Earth Ground

Fig 32—Various configurations for small indoor antennas. See text for discussion.

angles to each other as shown in **Figs 33A** and 33B. There will be some coupling with the arrangement shown in Fig 33C, but even this orientation is preferable to a single dipole.

With two antennas mounted 90° apart, you may find that one dipole is consistently better in nearly all directions, in which case you will want to remove the inferior dipole, perhaps placing it someplace else. In this manner the best spots in the house or attic can be determined experimentally.

Parasitic Conductors

Inevitably, any conductor in your house near a quarter wave in length or longer at the operating frequency will be parasitically coupled to your antenna. The word *parasitic* is particularly appropriate in this case because these conductors usually introduce losses and leave less energy for radiation into space. Unlike the parasitic elements in a beam antenna, conductors such as house wiring and plumbing are usually connected to lossy objects such as earth, electrical appliances, masonry or other objects that dissipate energy. Even where this energy is reradiated, it is not likely to be in the right phase in the desired direction; it is, in fact, likely to be a source of RFI.

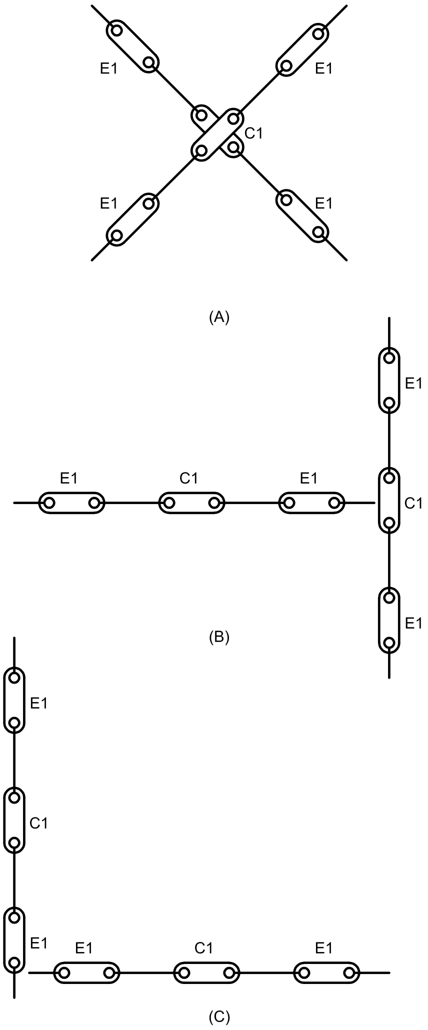
E1
E1
C1
E1
E1
(A)
E1
E1
C1
E1
C1
E1
(B)
E1
C1
E1
E1
C1
E1
(C)

Fig 33—Ways to orient a pair of perpendicular dipoles. The orientation at A and B will result in no mutual coupling between the two dipoles, but there will be some coupling in the configuration shown at C. End (EI) and center (CI) insulators are shown.

There are, however, some things that can be done about parasitic conductors. The most obvious is to reroute them at right angles to the antenna or close to the ground, or even underground—procedures that are usually not feasible in a finished home. Where these conductors cannot be rerouted, other measures can be taken. Electrical wiring can be broken up with RF chokes to prevent the flow of radio-frequency currents while permitting 60-Hz current (or audio, in the case of telephone wires) to flow unimpeded. A typical RF choke for a power line can be 100 turns of #10 insulated wire close wound on a length of 1-inch diameter plastic pipe. Of course one choke will be needed for each conductor. A three-wire line calls for three chokes. The chokes can be simplified by winding them bifilar or trifilar on a single coil form.

THE RESONANT BREAKER

Obviously, RF chokes cannot be used on conductors such as metal conduit or water pipes. But it is still possible, surprising as it may seem, to obstruct RF currents on such conductors without breaking the metal. The resonant breaker was first described by Fred Brown, W6HPH, in Oct 1979 *QST*.

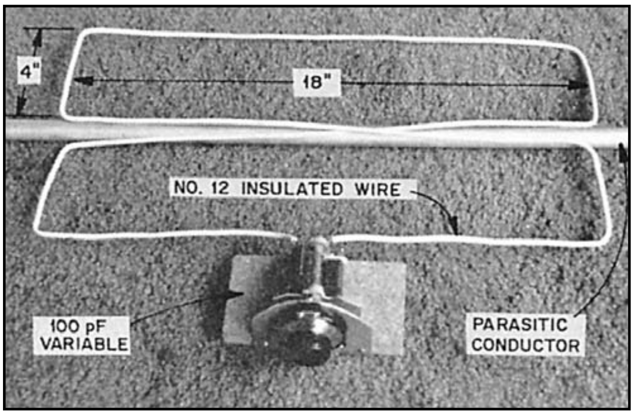

Fig 34—A "resonant breaker" such as shown here can be used to obstruct radio-frequency currents in a conductor without the need to break the conductor physically. A vernier dial is recommended for use with the variable capacitor because tuning is quite sharp. The 100-pF capacitor is in series with the loop. This resonant breaker tunes from 14 through 29.7 MHz. Larger models may be constructed for the lower frequency bands.

Fig 34 shows a method of accomplishing this. A figure-eight loop is inductively coupled to the parasitic conductor and is resonated to the desired frequency with a variable capacitor. The result is a very high impedance induced in series with the pipe, conduit or wire. This impedance will block the flow of radio-frequency currents. The figure-eight coil can be thought of as two turns of an air-core toroid and since the parasitic conductor threads through the hole of this core, there will be tight coupling between the two. Inasmuch as the figure-eight coil is parallel resonated, transformer action will reflect a high impedance in series with the linear conductor.

Before you bother with a *resonant breaker* of this type, be sure that there is a significant amount of RF current flowing in the parasitic conductor, and that you will therefore benefit from installing one. The relative magnitude of this current can be determined with an RF current probe of the type described in Chapter 27, Antenna and Transmission-Line Measurements. According to the rule of thumb regarding parasitic conductor current, if it measures less than $^1/_{10}$ of that measured near the center of the dipole, the parasitic current is generally not large enough to be of concern.

The current probe is also needed for resonating the breaker after it is installed. Normally, the resonant breaker will be placed on the parasitic conductor near the point of maximum current. When it is tuned through resonance, there will be a sharp dip in RF current, as indicated by the current probe. Of course, the resonant breaker will be effective only on one band. You will need one for each band where there is significant current indicated by the probe.

Power-Handling Capability

So far, our discussion have not considered the full power-handling capability of an indoor antenna. Any tendency to flash over must be determined by running full power or, preferably, somewhat more than the peak power you intend to use in regular operation. The antenna should be carefully checked for arcing or RF heating before you do any operating. Bear in mind that attics are indeed vulnerable to fire hazards. A potential of several hundred volts exists at the ends of a dipole fed by the typical Amateur Radio transmitter. If a power amplifier is used, there could be a few thousand volts at the ends of the dipole. Keep your antenna elements well away from other objects. *Safety first*!

Construction Details and Practical Considerations

Ultimately the success of an antenna project depends on the details of how the antenna is fabricated. A great deal of construction information is given in other chapters of this book. For example the construction of HF Yagis is discussed in Chapter 11, Quad arrays in Chapter 12, VHF antennas in Chapter 18, and in Chapter 20 there is an excellent discussion of antenna materials, particularly wire and tubing for elements. Here is still more helpful antenna construction information.

END EFFECT

If the standard expression $\lambda/2 \approx 491.8/f$ (MHz) is used for the length of a $\lambda/2$ wire antenna, the antenna will resonate at a somewhat lower frequency than is desired. The reason is that in addition to the effect of the conductor diameter and ground effects (Chapter 3, The Effects of Ground) an additional "loading" effect is caused by the insulators used at the ends of the wires to support the antenna. The insulators and the wire loops that tie the insulators to the antenna add a small amount of capacitance to the system. This capacitance helps to tune the antenna to a slightly lower frequency, in much the same way that additional capacitance in any tuned circuit lowers the resonant frequency. In an antenna this is called *end effect*. The current at the ends of the antenna does not quite reach zero because of the end effect, as there is some current flowing into the end capacitance. Note that the computations used to create Figs 2 through 7 in Chapter 2, Antenna Fundamentals, did not take into account any end effect.

End effect increases with frequency and varies slightly with different installations. However, at frequencies up to 30 MHz (the frequency range over which wire antennas are most commonly used), experience shows that the length of a practical $\lambda/2$ antenna, including the effect of diameter and end effect, is on the order of 5% less than the length of a half wave in space. As an average, then, the physical length of a resonant $\lambda/2$ wire antenna can be found from:

$$\lambda = \frac{491.8 \times 0.95}{f\ (\text{MHz})} \approx \frac{468}{f\ (\text{MHz})} \qquad \text{(Eq 1)}$$

Eq 1 is reasonably accurate for finding the physical length of a $\lambda/2$ antenna for a given frequency, but does not apply to antennas longer than a half wave in length. In the practical case, if the antenna length must be adjusted to exact frequency (not all antenna systems require it) the length should be "pruned" to resonance. Note that the use of plastic-insulated wire will typically lower the resonant frequency of a half-wave dipole about 3%.

INSULATORS

Wire antennas must be insulated at the ends. Commercially available insulators are made from ceramic, glass or plastic. Insulators are available from many Amateur Radio dealers. RadioShack and local hardware stores are other possible sources. Acceptable homemade insulators may be fashioned from a variety of material including (but not limited to) acrylic sheet or rod, PVC tubing, wood, fiberglass rod or even stiff plastic from a discarded container. **Fig 35** shows some homemade insulators. Ceramic or glass insulators will usually outlast the wire, so they are highly recommended for a safe, reliable, permanent installation. Other materials may tear under stress or break down in the presence of sunlight. Many types of plastic do not weather well.

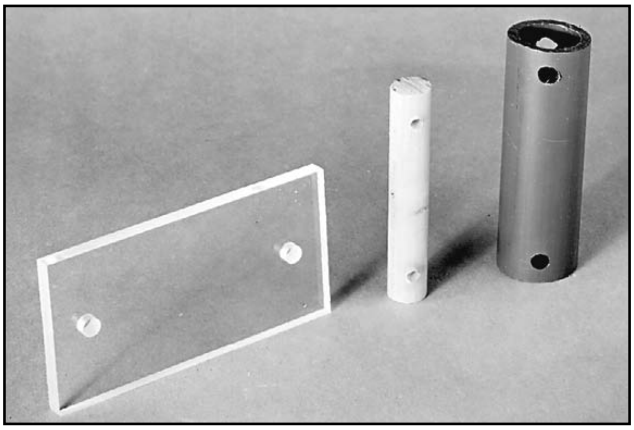

Fig 35—Some ideas for homemade antenna insulators.

INSTALLING TRANSMISSION LINES

Many wire antennas require an insulator at the feed point. Although there are many ways to connect the feed line, there are a few things to keep in mind. If you feed your antenna with coaxial cable, you have two choices. You can install an SO-239 connector on the center insu-

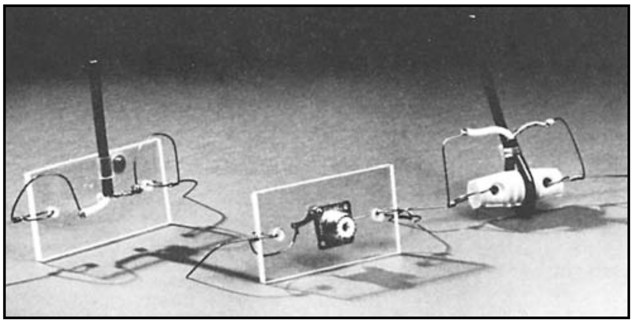

Fig 36—Some homemade dipole center insulators. The one in the center includes a built-in SO-239 connector. Others are designed for direct connection to the feed line.

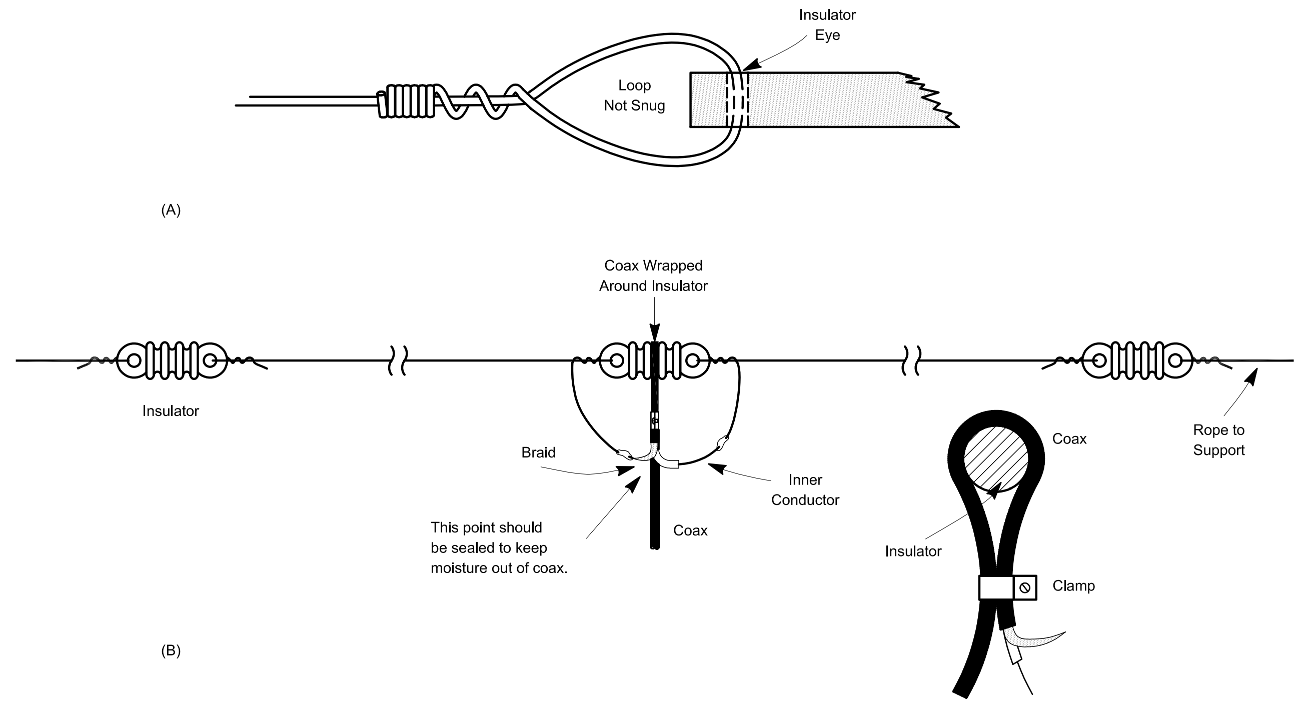
Insulator
Eye
Loop
Not Snug
(A)
Coax Wrapped
Around Insulator
Insulator
Braid
Inner
Conductor
This point should
be sealed to keep
moisture out of coax.
Coax
Coax
Rope to
Support
Insulator
Clamp
(B)

Fig 37—Details of dipole antenna construction. At A, the end insulator connection is shown. At B, the completed antenna is shown. A balun (not shown) is often used at the feed point, since this is a balanced antenna.

lator, as shown by the center example in **Fig 36,** and use a PL-259 on the end of your coax, or you can separate the center conductor from the braid and connect the feed line directly to the antenna wire as shown in the other two examples in Fig 36 and the example in **Fig 37**. Although it costs less to connect direct, the use of connectors offers several advantages. Coaxial cable braid soaks up water like a sponge unless it is very well waterproofed. If you do not adequately seal the antenna end of the feed line, water will find its way into the braid. Water in the feed line will lead to contamination, rendering the coax useless long before its normal lifetime is up. Many hams waterproof the coax, first with vinyl electrical tape, and then using a paint-on material called “PlastiDip,” which is sold by RadioShack (part number 910-5166 for the white variety).

It is not uncommon for water to drip from the end of the coax inside the shack after a year or so of service if the antenna connection is not properly waterproofed. Use of a PL-259/SO-239 combination (or connector of your choice) makes the task of waterproofing connections much easier. Another advantage to using the PL-259/SO-239 combination is that feed-line replacement is much easier, should that become necessary.

Whether you use coaxial cable, ladder line, or twin lead to feed your antenna, an often overlooked consideration is the mechanical strength of the connection. Wire antennas and feed lines tend to move a lot in the breeze, and

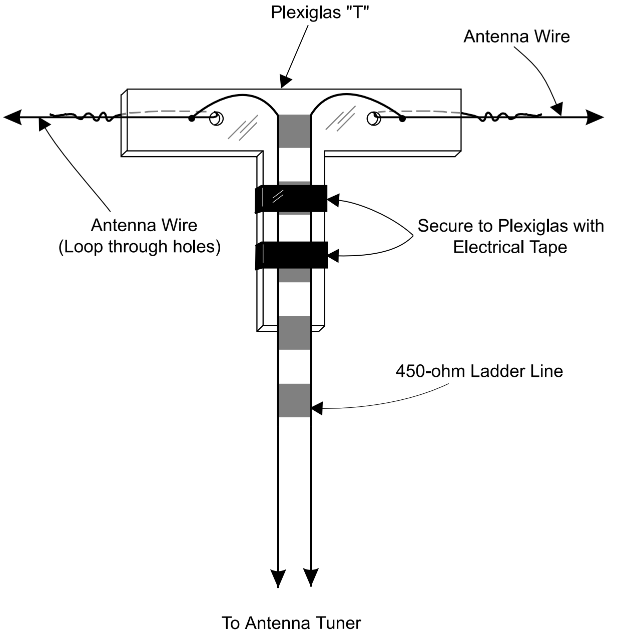
Plexiglas "T"
Antenna Wire
Antenna Wire
(Loop through holes)
Secure to Plexiglas with
Electrical Tape
450-ohm Ladder Line
To Antenna Tuner

Fig 38—A piece of cut Plexiglas can be used as a center insulator and to support a ladder-line feeder. The Plexiglas acts to reduce the flexing of the wires where they connect to the antenna. Use thick Plexiglas in areas subject to high winds.

unless the feed line is attached securely, the connection will weaken with time. The resulting failure can range from a frustrating intermittent electrical connection to a complete separation of feed line and antenna. Fig 37 and **Fig 38** illustrate different ways of attaching either coax or ladder line to the antenna securely.

When open-wire feed line is used, the conductors of the line should be anchored to the insulator by threading them through the eyes of the insulator two or three times, and twisting the wire back on itself before soldering. A slack tie wire should then be used between the feeder conductor and the antenna, as shown in Fig 38. (The tie wires may be extensions of the line conductors themselves.) When window-type line is suspended from an antenna in a manner such as that shown in Fig 38, the line should be twisted—at several twists per foot—to prevent stress hardening of the wire because of constant flexing in the wind.

When using plastic-insulated open-wire line, the tendency of the line to twist and short out close to the antenna can be counteracted by making the center insulator of the antenna longer than the spacing of the line, as shown in Fig 38. In severe wind areas, it may be necessary to use ¼-inch thick Plexiglas for the center insulator rather than thinner material.

RUNNING THE FEED LINE FROM THE ANTENNA TO THE STATION

Chapter 24, Transmission Lines, contains some general guidelines for installing feed lines. More detailed information is contained in this section. Whenever possible, the transmission line should be lead away from the antenna at a 90° angle to minimize coupling from the antenna to the transmission line. This coupling can cause unequal currents on the transmission line, which will then radiate and it can detune the antenna.

Except for the portion of the line in close proximity to the antenna, coaxial cable requires no particular care in running from the antenna to the station entrance, other than protection from mechanical damage. If the antenna is not supported at the center, the line should be fastened to a post more than head high located under the center of the antenna, allowing enough slack between the post and the antenna to take care of any movement of the antenna in the wind. If the antenna feed point is supported by a tower or mast, the cable can be taped to the mast at intervals or to one leg of the tower.

Coaxial cable rated for direct burial can be buried a few inches in the ground to make the run from the antenna to the station. A deep slit can be cut by pushing a square-end spade full depth into the ground and moving the handle back and forth to widen the slit before removing the spade. After the cable has been pushed into the slit with a piece of 1-inch board 3 or 4 inches wide, the slit can be tamped closed. Many hams run coax cables through PVC pipe buried in the ground deeper than the frost line and slanted downwards slightly so that water will drain, rather than pooling inside the length of the pipe.

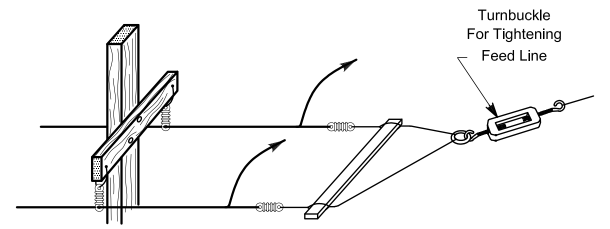

Fig 39—A support for open-wire line. The support at the antenna end of the line must be sufficiently rigid to stand the tension of the line.

Solid ribbon or the newer *window* types of line should be kept reasonably well spaced from other conductors running parallel to it for more than a few feet. TV-type standoff insulators with strap clamp mountings can be used for running this type of line down a mast or tower leg. Similar insulators of the screw-in type can be used in supporting the line on wooden poles for a long run.

Open-wire lines with bare conductors require frequent supports to keep the lines from twisting and shorting out, as well as to relieve the mechanical strain. One method of supporting a long horizontal run of heavy open-wire line is shown in **Fig 39**. The line must be anchored securely at a point under the feed point of the antenna. Window-type line can be supported similarly with wire links fastened to the insulators or with black cable ties (ones not affected by UV radiation from the sun).

To keep the line clear of pedestrians and vehicles, it is usually desirable to anchor the feed line at the eaves or rafter line of the station building (see **Fig 40**), and then drop it vertically to the point of entrance. The points of anchorage and entrance should be chosen to permit the vertical drop without crossing windows for aesthetic reasons.

If the station is located in a room on the ground floor, one way of bringing coax transmission line into the house is to go through the outside wall below floor level, feed it through the basement or crawl space and then up to the station through a hole in the floor. When making the entrance hole in the side of the building, suitable measurements should be made in advance to be sure the hole will go through the sill 2 or 3 inches above the foundation line (and between joists if the bore is parallel to the joists). The line should be allowed to sag below the entrance hole level outside the building to allow rain water to drip off.

Open-wire line can be fed in a similar manner, although it will require a separate hole for each conductor. Each hole should be insulated with a length of polystyrene or Lucite tubing. If available, ceramic tubes salvaged from old-fashioned *knob and tube* electrical

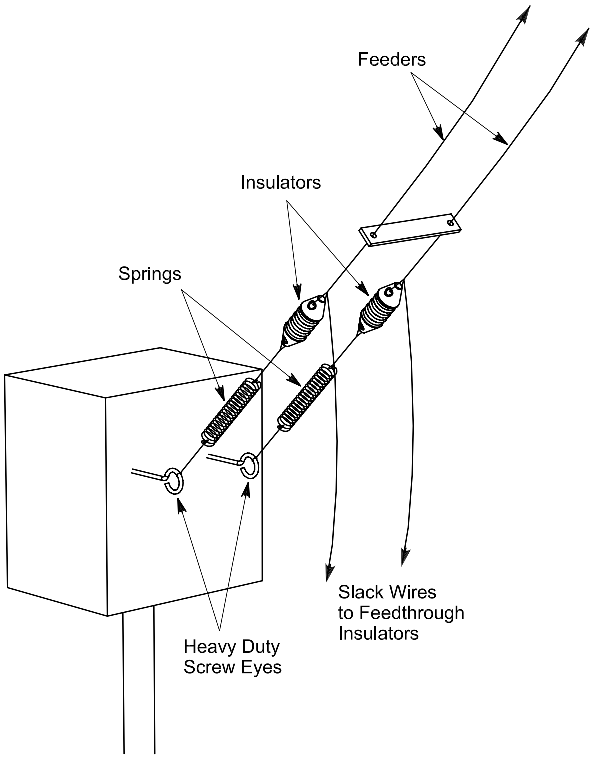

Fig 40—Anchoring open-wire line at the station end. The springs are especially desirable if the line is not supported between the antenna and the anchoring point.

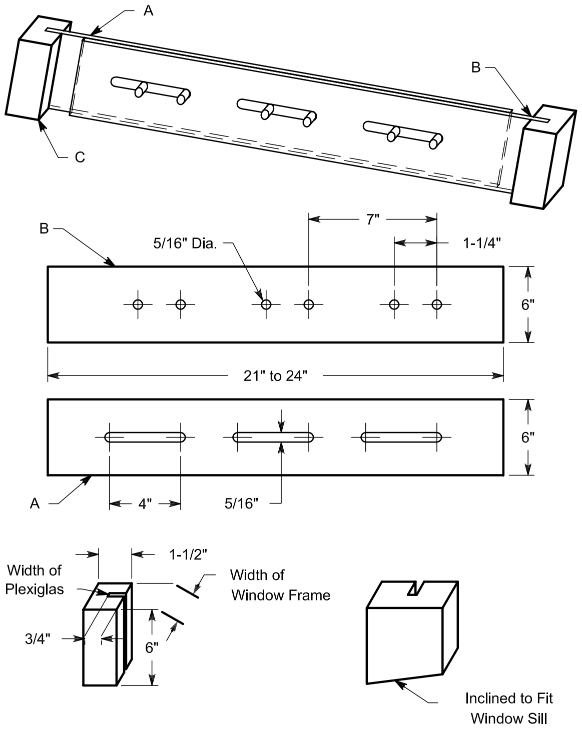

Fig 41—An adjustable window lead-in panel made up of two sheets of Lucite or Plexiglas. A feedthrough connector for coax line can be made as shown in Fig 28. Ceramic feedthrough insulators are suitable for open-wire line. (W1RVE)

installations, work very well for this purpose. Drill the holes with a slight downward slant toward the outside of the building to prevent rain seepage. With window ladder line, it will be necessary to remove a few of the spreader insulators, cut the line before passing through the holes (allowing enough length to reach the inside) and splice the remainder on the inside.

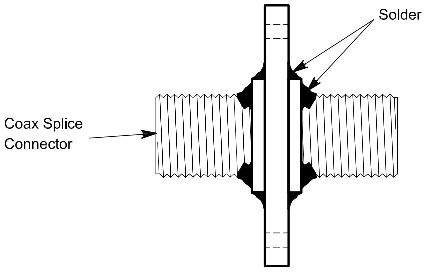

Fig 42—Feedthrough connector for coax line. An Amphenol 83-1J (PL-258) connector, the type used to splice sections of coax line together, is soldered into a hole cut in a brass mounting flange. An Amphenol bulkhead adapter 83-1F may be used instead.

If the station is located above ground level, or if there is other objection to the procedure described above, entrance can be made at a window, using the arrangement shown in **Fig 41**. An Amphenol type 83-1F (UG-363) connector can be used as shown in **Fig 42**; ceramic feedthrough insulators can be used for open-wire line. Ribbon line can be run through clearance holes in the panel, and secured by a winding of tape on either side of the panel, or by cutting the retaining rings and insulators from a pair of TV standoff insulators and clamping one on each side of the panel.

LIGHTNING PROTECTION

Two or three types of lightning arresters for coaxial cable are available on the market. If the antenna feed point is at the top of a well-grounded tower, the arrester can be fastened securely to the top of the tower for grounding purposes. A short length of cable, terminated in a coaxial plug, is then run from the antenna feed point to one receptacle of the arrester, while the transmission line is run from the other arrester receptacle to the station. Such arresters may also be placed at the entrance point to the station, if a suitable ground connection is available at that point (or arresters may be placed at both points for added insurance).

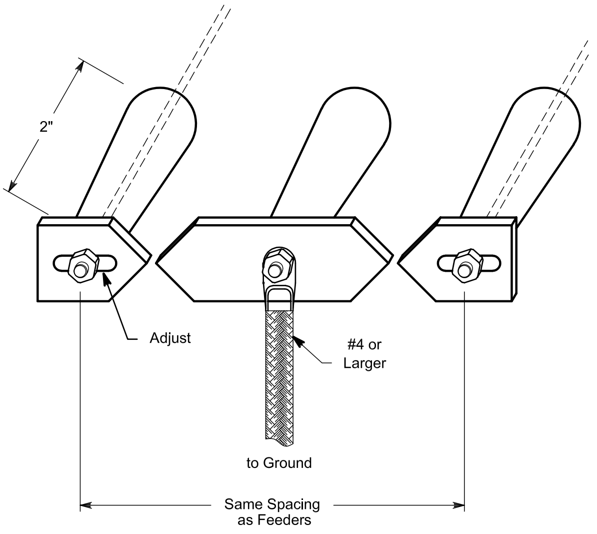

Fig 43—A simple lightning arrester for open-wire line made from three standoff or feedthrough insulators and sections of $^1/_8 \times ^1/_2$-inch brass or copper strap. It should be installed in the line at the point where the line enters the station. The heavy ground lead should be as short and as direct as possible. The gap setting should be adjusted to the minimum width that will prohibit arcing when the transmitter is operated.

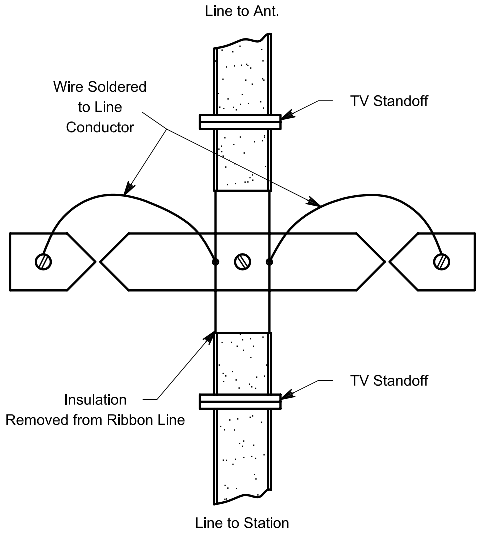

Fig 44—The lightning arrester of Fig 39 may be used with 300-Ω ribbon line in the manner shown here. The TV standoffs support the line an inch or so away from the grounded center member of the arrester

The construction of a homemade arrester for open-wire line is shown in **Fig 43**. This type of arrester can be adapted to ribbon line an inch or so away from the center member of the arrester, as shown in **Fig 44**. Sufficient insulation should be removed from the line where it crosses the arrester to permit soldering the arrester connecting leads.

Lightning Grounds

Lightning-ground connecting leads should be of conductor size equivalent to at least #10 wire. The #8 aluminum wire used for TV-antenna grounds is satisfactory. Copper braid $^3/_4$-inch wide (Belden 8662-10) is also suitable. The conductor should run in a straight line to the grounding point. The ground connection may be made to a water pipe system (if the pipe is not plastic), the grounded metal frame of a building, or to one or more $^5/_8$-inch ground rods driven to a depth of at least 8 feet. More detailed information on lightning protection is contained in Chapter 1, Safety.

A central grounding panel for coax cables coming into the house is highly recommended. See **Fig 45** for a photo of the homemade grounding panel installed by Chuck Hutchinson, K8CH, at his Michigan home. The coax cables screwed into dual-female feed-through UHF connectors. K8CH installed this aluminum panel under the outside grill for a duct that provided combustion air to an unused fireplace. He used ground strap to connect to ground rods located under the panel. See the ARRL book *Simple and Fun Antennas* for more information about ground panels.

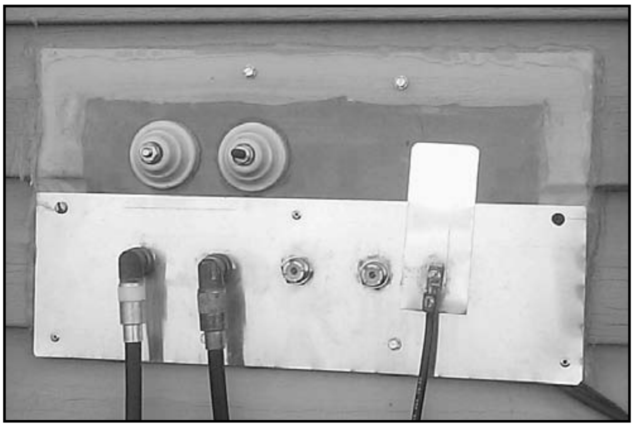

Fig 45—K8CH's coax entry panel mounted on exterior wall (later covered by grill that provides combustion to an unused fireplace). The ground braid goes to a ground rod located beneath the panel. (Photo courtesy: *Simple and Fun Antennas for Hams*)

Before a lightning storm approaches, a prudent ham will disconnect all feed lines, rotor lines and control lines inside the shack to prevent damage to sensitive electronics. When lightning is crashing about outside, you certainly don't want that lightning inside your shack!

Chapter 5

Loop Antennas

A loop antenna is a closed-circuit antenna—that is, one in which a conductor is formed into one or more turns so its two ends are close together. Loops can be divided into two general classes, those in which both the total conductor length and the maximum linear dimension of a turn are very small compared with the wavelength, and those in which both the conductor length and the loop dimensions begin to be comparable with the wavelength.

A "small" loop can be considered to be simply a rather large coil, and the current distribution in such a loop is the same as in a coil. That is, the current has the same phase and the same amplitude in every part of the loop. To meet this condition, the total length of conductor in the loop must not exceed about 0.1 λ. Small loops are discussed later in this chapter, and further in Chapter 14, Direction Finding Antennas.

A "large" loop is one in which the current is not the same either in amplitude or phase in every part of the loop. This change in current distribution gives rise to entirely different properties compared with a small loop.

Half-Wave Loops

The smallest size of "large" loop generally used is one having a conductor length of $^1/_2$ λ. The conductor is usually formed into a square, as shown in **Fig 1**, making each side $^1/_8$ λ long. When fed at the center of one side, the current flows in a closed loop as shown in Fig 1A. The current distribution is approximately the same as on a $^1/_2$ λ wire, and so is maximum at the center of the side opposite the terminals X-Y, and minimum at the terminals themselves. This current distribution causes the field strength to be maximum in the plane of the loop and in the direction looking from the low-current side to the high-current side. If the side opposite the terminals is opened at the center as shown in Fig 1B (strictly speaking, it is then no longer a loop because it is no longer a closed circuit), the direction of current flow remains unchanged but the maximum current flow occurs at the terminals. This reverses the direction of maximum radiation.

The radiation resistance at a current antinode (which is also the resistance at X-Y in Fig 1B) is on the order of 50 Ω. The impedance at the terminals in Fig 1A is a few thousand ohms. This can be reduced by using two identical loops side by side with a few inches spacing between them and applying power between terminal X on one loop and terminal Y on the other.

Unlike a $^1/_2$ λ dipole or a small loop, there is no direction in which the radiation from a loop of the type shown in Fig 1 is zero. There is appreciable radiation in the direction perpendicular to the plane of the loop, as well as to the "rear"—the opposite direction to the arrows shown. The front-to-back (F/B) ratio is approximately 4 to 6 dB. The small size and the shape of the directive pattern result in a loss of about 1 dB when the field strength in the optimum direction from such a

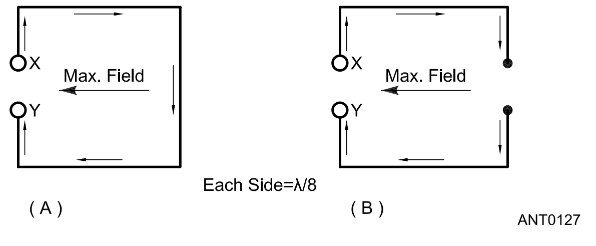

Fig 1—Half-wave loops, consisting of a single turn having a total length of $^1/_2$ λ.

loop is compared with the field from a $^1/_2$ λ dipole in its optimum direction.

The ratio of the forward radiation to the backward radiation can be increased, and the field strength likewise increased at the same time to give a gain of about 1 dB over a dipole, by using inductive reactances to "load" the sides joining the front and back of the loop. This is shown in **Fig 2**. The reactances, which should have a value of approximately 360 Ω, decrease the current in the sides in which they are inserted and increase it in the side having terminals. This increases the directivity and thus increases the efficiency of the loop as a radiator. Lossy coils can reduce this advantage greatly.

One-Wavelength Loops

Loops in which the conductor length is 1 λ have different characteristics than $^1/_2$-λ loops. Three forms of 1 λ loops are shown in **Fig 3**. At A and B the sides of the squares are equal to $^1/_4$ λ, the difference being in the point at which the terminals are inserted. At C the sides of the triangle are equal to $^1/_3$ λ. The relative direction of current flow is as shown in the drawings. This direction reverses halfway around the perimeter of the loop, as such reversals always occur at the junction of each $^1/_2$-λ section of wire.

The directional characteristics of loops of this type are opposite in sense to those of a small loop. That is, the radiation is maximum perpendicular to the plane of the loop and is minimum in either direction in the plane containing the loop. If the three loops shown in Fig 3 are mounted in a vertical plane with the terminals at the bottom, the radiation is horizontally polarized. When the terminals are moved to the center of one vertical side in Fig 3A, or to a side corner in B, the radiation is vertically polarized. If the terminals are moved to a side corner in C, the polarization will be diagonal, containing both vertical and horizontal components.

In contrast to straight-wire antennas, the electrical length of the circumference of a 1-λ loop is shorter than the actual length. For a loop made of bare #18 wire and operating at a frequency of 14 MHz, where the ratio of conductor length to wire diameter is large, the loop will be close to resonance when

$$\text{Length}_{\text{feet}} = \frac{1032}{f_{\text{MHz}}}$$

The radiation resistance of a resonant 1 λ loop is approximately 120 Ω, under these conditions. Since the loop dimensions are larger than those of a $^1/_2$-λ dipole, the radiation efficiency is high.

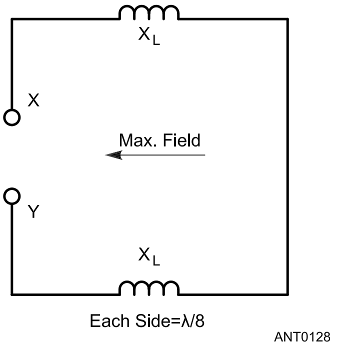

Fig 2—Inductive loading in the sides of a $^1/_2$-λ loop to increase the directivity and gain. Maximum radiation or response is in the plane of the loop, in the direction shown by the arrow.

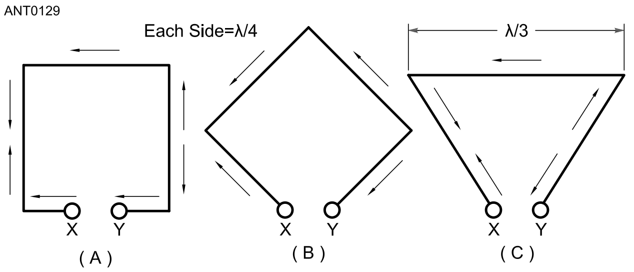

Fig 3—At A and B, loops having sides $^1/_4$ λ long, and at C having sides $^1/_3$ λ long (total conductor length 1 λ). The polarization depends on the orientation of the loop and on the position of the feed point (terminals X-Y) around the perimeter of the loop.

In the direction of maximum radiation (that is, broadside to the plane of the loop, regardless of the point at which it is fed) the 1-λ loop will show a small gain over a $^1/_2$-λ dipole. Theoretically, this gain is about 1 dB, and measurements have confirmed that it is of this order.

The 1-λ loop is more frequently used as an element of a directive antenna array (the quad and delta-loop antennas described in Chapter 12, Quad Arrays) than singly, although there is no reason why it cannot be used alone. In the quad and delta loop, it is nearly always driven so that the polarization is horizontal.

Small Loop Antennas

The electrically small loop antenna has existed in various forms for many years. Probably the most familiar form of this antenna is the ferrite *loopstick* found in portable AM broadcast-band receivers. Amateur applications of the small loop include direction finding, low-noise directional receiving antennas for 1.8 and 3.5 MHz, and small transmitting antennas. Because the design of transmitting and receiving loops requires some different considerations, the two situations are examined separately in this section. This information was written by Domenic M. Mallozzi, N1DM.

The Basic Loop

What is and what is not a small loop antenna? By definition, the loop is considered to be electrically small when its total conductor length is less than 0.1 λ—0.085 is the number used in this section. This size is based on the fact that the current around the perimeter of the loop must be in phase. When the winding conductor is more than about 0.085 λ long, this is no longer true. This constraint results in a very predictable figure-eight radiation pattern, shown in **Fig 4**.

The simplest loop is a 1-turn untuned loop with a load connected to a pair of terminals located in the center of one of the sides, as shown in **Fig 5**. How its pattern is developed is easily pictured if we look at some "snapshots" of the antenna relative to a signal source. **Fig 6** represents a loop from above, and shows the instantaneous radiated voltage wave. Note that points A and B of the loop are receiving the same instantaneous voltage. This means that no current will flow through the loop, because there is no current flow between points of equal potential. A similar analysis of **Fig 7**, with the loop turned 90° from the position represented in Fig 6, shows that this position of the loop provides maximum response. Of course, the voltage derived from the passing wave is small because of the small physical size of the loop. Fig 4 shows the ideal radiation pattern for a small loop.

The voltage across the loop terminals is given by

$$V = \frac{2\pi A N E \cos\theta}{\lambda} \qquad \text{(Eq 1)}$$

where

V = voltage across the loop terminals
A = area of loop in square meters
N = number of turns in the loop
E = RF field strength in volts per meter
θ = angle between the plane of the loop and the signal source (transmitting station)
λ = wavelength of operation in meters

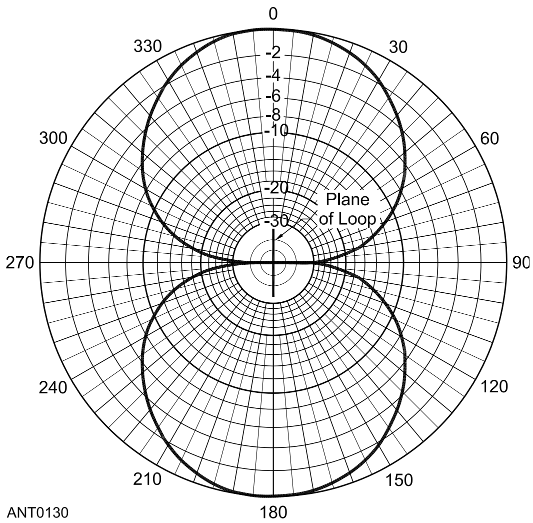

Fig 4—Calculated small loop antenna radiation pattern.

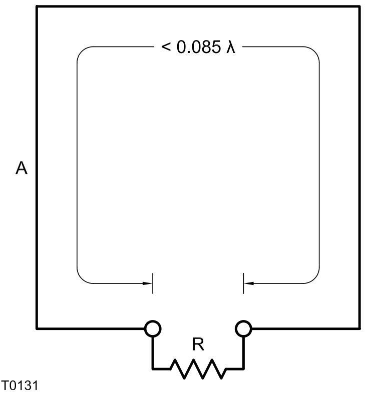

Fig 5—Simple untuned small loop antenna.

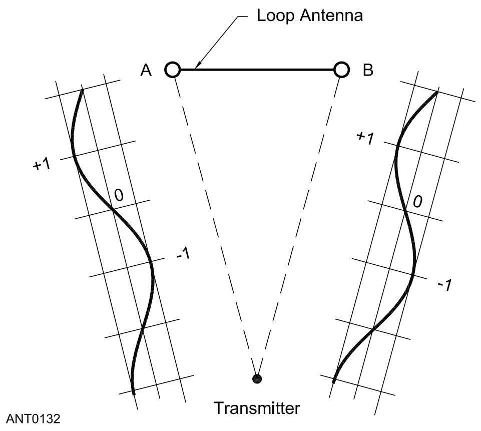

Fig 6—Example of orientation of loop antenna that does not respond to a signal source (null in pattern).

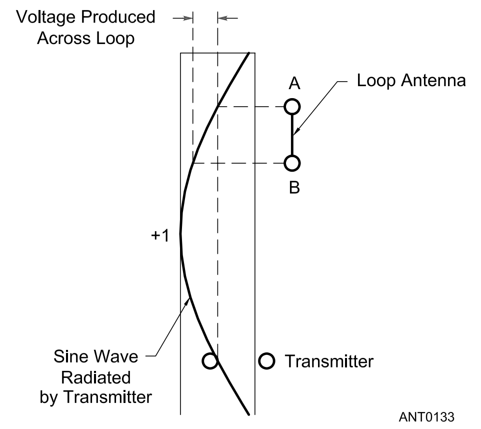

Fig 7—Example of orientation of loop antenna for maximum response.

This equation comes from a term called *effective height*. The effective height refers to the height (length) of a vertical piece of wire above ground that would deliver the same voltage to the receiver. The equation for effective height is

$$h = \frac{2\pi N A}{\lambda} \tag{Eq 2}$$

where h is in meters and the other terms are as for Eq 1.

A few minutes with a calculator will show that, with the constraints previously stated, the loop antenna will have a very small effective height. This means it will deliver a relatively small voltage to the receiver, even with a large transmitted signal.

TUNED LOOPS

We can tune the loop by placing a capacitor across the antenna terminals. This causes a larger voltage to appear across the loop terminals because of the Q of the parallel resonant circuit that is formed.

The voltage across the loop terminals is now given by

$$V = \frac{2\pi A N E Q \cos\theta}{\lambda} \tag{Eq 3}$$

where Q is the loaded Q of the tuned circuit, and the other terms are as defined above.

Most amateur loops are of the tuned variety. For this reason, all comments that follow are based on tuned-loop antennas, consisting of one or more turns. The tuned-loop antenna has some particular advantages. For example, it puts high selectivity up at the “front” of a receiving system, where it can significantly help factors such as dynamic range. Loaded Q values of 100 or greater are easy to obtain with careful loop construction.

Consider a situation where the inherent selectivity of the loop is helpful. Assume we have a loop with a Q of 100 at 1.805 MHz. We are working a DX station on 1.805 MHz and are suffering strong interference from a local station 10 kHz away. Switching from a dipole to a small loop will reduce the strength of the off-frequency signal by 6 dB (approximately one S unit). This, in effect, increases the dynamic range of the receiver. In fact, if the off-frequency station were further off frequency, the attenuation would be greater.

Another way the loop can help is by using the nulls in its pattern to null out on-frequency (or slightly off-frequency) interference. For example, say we are working a DX station to the north, and just 1 kHz away is another local station engaged in a contact. The local station is to our west. We can simply rotate our loop to put its null to the west, and now the DX station should be readable while the local will be knocked down by 60 or more dB. This obviously is quite a noticeable difference. Loop nulls are very sharp and are generally noticeable only on ground-wave signals (more on this later).

Of course, this method of nulling will be effective only if the interfering station and the station being worked are not in the same direction (or in exact opposite directions) from our location. If the two stations were on the same line from our location, both the station being worked and the undesired station would be nulled out. Luckily the nulls are very sharp, so as long as the stations are at least 10° off axis from each other, the loop null will be usable.

A similar use of the nulling capability is to eliminate local noise interference, such as that from a light dimmer in a neighbor’s house. Just put the null on the offending light dimmer, and the noise should disappear.

Now that we have seen some possible uses of the small loop, let us look at a bit of detail about its design. First, the loop forms an inductor having a very small ratio of winding length to diameter. The equations for finding inductance given in most radio handbooks assume that the inductor coil is longer than its diameter. However, F. W. Grover of the US National Bureau of Standards has provided equations for inductors of common cross-sectional shapes and small length-to-diameter ratios. (See the Bibliography at the end of this chapter.) Grover’s equations are shown in **Table 1**. Their use will yield relatively accurate numbers; results are easily worked out with a scientific calculator or home computer.

The value of a tuning capacitor for a loop is easy to calculate from the standard resonance equations. The only matter to consider before calculating this is the value of distributed capacitance of the loop winding. This capacitance shows up between adjacent turns of the coil because of their slight difference in potential. This causes each turn to appear as a charge plate. As with all other capacitances, the value of the distributed capacitance is based on the physical dimensions of the coil. An exact mathematical analysis of its value is a complex problem. A simple approximation is given by Medhurst (see Bibliography) as:

$$C = HD \tag{Eq 4}$$

where

C = distributed capacitance in pF
H = a constant related to the length-to-diameter ratio of the coil (**Table 2** gives H values for length-to-diameter ratios used in loop antenna work.)
D = diameter of the winding in cm

Medhurst's work was with coils of round cross section. For loops of square cross section the distributed capacitance is given by Bramslev (see Bibliography) as

$$C = 60S \qquad \text{(Eq 5)}$$

where

C = the distributed capacitance in pF
S = the length of the side in meters

If you convert the length in this equation to centimeters, you will find Bramslev's equation gives results in the same order of magnitude as Medhurst's equation.

This distributed capacitance appears as if it were a capacitor across the loop terminals. Therefore, when determining the value of the tuning capacitor, the distributed capacitance must be subtracted from the total capacitance required to resonate the loop. The distributed capacitance also determines the highest frequency at which a particular loop can be used, because it is the minimum capacitance obtainable.

Electrostatically Shielded Loops

Over the years, many loop antennas have incorporated an electrostatic shield. This shield generally takes the form of a tube around the winding, made of a conductive but nonmagnetic material (such as copper or aluminum). Its purpose is to maintain loop balance with respect to ground, by forcing the capacitance between all portions of the loop and ground to be identical. This is illustrated in **Fig 8**. It is necessary to maintain electrical loop balance to eliminate what is referred to as the *antenna effect*. When the antenna becomes unbalanced it appears to act partially as a small vertical antenna. This vertical pattern gets superimposed on the ideal figure-eight pattern, distorting the pattern and filling in the nulls. The type of pattern that results is shown in **Fig 9**.

Adding the shield has the effect of somewhat reducing the pickup of the loop, but this loss is generally offset by the increase in null depth of the loops. Proper balance of the loop antenna requires that the load on the loop also be balanced. This is usually accomplished by use of a balun transformer or a balanced input preamplifier. One important point regarding the shield is that it cannot form a continuous electrical path around the loop perimeter, or it will appear as a shorted coil turn. Usually the insulated break is located opposite the feed point to maintain symmetry. Another point to be considered is that the shield should be of a much larger diameter than the loop winding, or it will lower the Q of the loop.

Table 1
Inductance Equations for Short Coils (Loop Antennas)

Triangle:

$$L\,(\mu H) = 0.006N^2\,s \left[\ln\left(\frac{1.1547\,sN}{(N+1)\,\ell}\right) + 0.65533 + \frac{0.1348\,(N+1)\,\ell}{sN} \right]$$

Square:

$$L\,(\mu H) = 0.008N^2\,s \left[\ln\left(\frac{1.4142sN}{(N+1)\ell}\right) + 0.37942 + \frac{0.3333(N+1)\ell}{sN} \right]$$

Hexagon:

$$L\,(\mu H) = 0.012N^2\,s \left[\ln\left(\frac{2sN}{(N+1)\ell}\right) + 0.65533 + \frac{0.1348(N+1)\ell}{sN} \right]$$

Octagon:

$$L\,(\mu H) = 0.016N^2 \left[\ln\left(\frac{2.613sN}{(N+1)\ell}\right) + 0.75143 + \frac{0.07153(N+1)\ell}{sN} \right]$$

where

N = number of turns
s = side length in cm
ℓ = coil length in cm

Note: In the case of single-turn coils, the diameter of the conductor should be used for ℓ.

Table 2
Values of the Constant H for Distributed Capacitance

Length to Diameter Ratio	*H*
0.10	0.96
0.15	0.79
0.20	0.78
0.25	0.64
0.30	0.60
0.35	0.57
0.40	0.54
0.50	0.50
1.00	0.46

Various construction techniques have been used in

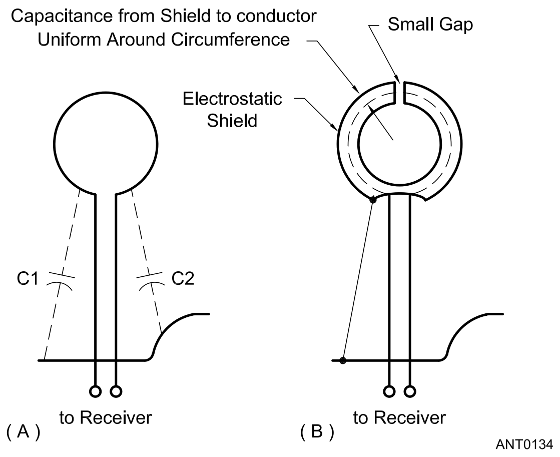

Fig 8—At A, the loop is unbalanced by capacitance to its surroundings. At B, the use of an electrostatic shield overcomes this effect.

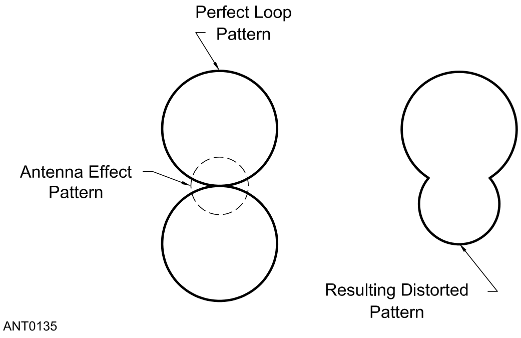

Fig 9—Distortion in loop pattern resulting from antenna effect.

making shielded loops. Genaille located his loop winding inside aluminum conduit, while True constructed an aluminum shield can around his winding. Others have used pieces of Hardline to form a loop, using the outer conductor as a shield. DeMaw used flexible coax with the shield broken at the center of the loop conductor in a multiturn loop for 1.8 MHz. Goldman uses another shielding method for broadcast receiver loops. His shield is in the form of a barrel made of hardware cloth, with the loop in its center. (See Bibliography for above references.) All these methods provide sufficient shielding to maintain the balance. It is possible, as Nelson shows, to construct an unshielded loop with good nulls (60 dB or better) by paying great care to symmetry.

LOOP Q

As previously mentioned, Q is an important consideration in loop performance because it determines both the loop bandwidth and its terminal voltage for a given field strength. The loaded Q of a loop is based on four major factors. These are (1) the intrinsic Q of the loop winding, (2) the effect of the load, (3) the effect of the electrostatic shield, and (4) the Q of the tuning capacitor.

The major factor is the Q of the winding of the loop itself. The ac resistance of the conductor caused by skin effect is the major consideration. The ac resistance for copper conductors may be determined from

$$R = \frac{0.996 \times 10^{-6}\sqrt{f}}{d} \qquad \text{(Eq 6)}$$

where

R = resistance in ohms per foot
f = frequency, Hz
d = conductor diameter, inches

The Q of the inductor is then easily determined by taking the reactance of the inductor and dividing it by the ac resistance. If you are using a multiturn loop and are a perfectionist, you might also want to include the loss from conductor proximity effect. This effect is described in detail later in this chapter, in the section on transmitting loops.

Improvement in Q can be obtained in some cases by the use of Litz wire (short for *Litzendraht*). Litz wire consists of strands of individual insulated wires that are woven into bundles in such a manner that each conductor occupies each location in the bundle with equal frequency. Litz wire results in improved Q over solid or stranded wire of equivalent size, up to about 3 MHz.

Also, the Q of the tuned circuit of the loop antenna is determined by the Q of the capacitors used to resonate it. In the case of air variables or dipped micas this is not usually a problem. But if variable-capacitance diodes are used to remotely tune the loop, pay particular attention to the manufacturer's specification for Q of the diode at the frequency of operation. The tuning diodes can have a significant effect on circuit Q.

Now we consider the effect of load impedance on loop Q. In the case of a directly coupled loop (as in Fig 5), the load is connected directly across the loop terminals, causing it to be treated as a parallel resistance in a parallel-tuned RLC circuit. Obviously, if the load is of a low value, the Q of the loop will be low. A simple way to correct this is to use a transformer to step up the load impedance that appears across the loop terminals. In fact, if we make this transformer a balun, it also allows us to use our unbalanced receivers with the loop and maintain loop symmetry. Another solution is to use what is referred to as an inductively coupled loop, such as DeMaw's four turn electrostatically shielded loop. A one-turn link is connected to the receiver. This turn is wound with the four-turn loop. In effect, this builds the transformer into the antenna.

Another solution to the problem of load impedance on loop Q is to use an active preamplifier with a high impedance balanced input and unbalanced output. This method also has the advantage of amplifying the low-level output voltage of the loop to where it can be used with a receiver of even mediocre sensitivity. In fact, the Q of the loop when used with a balanced preamplifier having high input impedance may be so high as to be unusable in certain applications. An example of this situation would occur where a loop is being used to receive a 5 kHz wide AM signal at a frequency where the bandwidth of the loop is only 1.5 kHz. In this case the detected audio might be very distorted. The solution to this is to put a Q-degrading resistor across the loop terminals.

FERRITE-CORE LOOP ANTENNAS

The ferrite-core loop antenna is a special case of the air-core receiving loops considered up to now. Because of its use in every AM broadcast-band portable radio, the ferrite-core loop is, by quantity, the most popular form of the loop antenna. But broadcast-band reception is far from its only use; it is commonly found in radio-direction-finding equipment and low-frequency-receiving systems (below 500 kHz) for time and frequency standard systems. In recent years, design information on these types of antennas has been a bit sparse in the amateur literature, so the next few paragraphs are devoted to providing some details.

Ferrite-loop antennas are characteristically very small compared to the frequency of use. For example, a 3.5-MHz version may be in the range of 15 to 30 cm long and about 1.25 cm in diameter. Earlier in this chapter, effective height was introduced as a measure of loop sensitivity. The effective height of an air-core loop antenna is given by Eq 2.

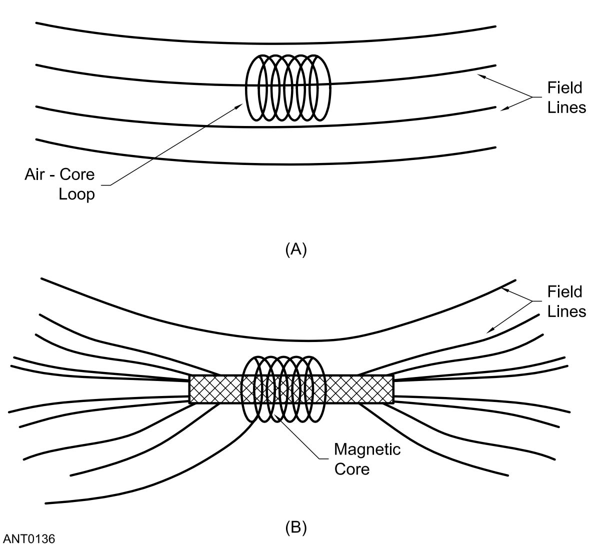

Fig 10—At A, an air-core loop has no effect on nearby field lines. B illustrates the effect of a ferrite core on nearby field lines. The field is altered by the reluctance of the ferrite material.

If an air-core loop is placed in a field, in essence it cuts the lines of flux without disturbing them (**Fig 10A**). On the other hand, when a ferrite (magnetic) core is placed in the field, the nearby field lines are redirected into the loop (Fig 10B). This is because the reluctance of the ferrite material is less than that of the surrounding air, so the nearby flux lines tend to flow through the loop rather than passing it by. (Reluctance is the magnetic analogy of resistance, while flux is analogous to current.) The reluctance is inversely proportional to the permeability of the rod core, μ_{rod}. (In some texts the rod permeability is referred to as effective permeability, μ_{eff}). This effect modifies the equation for effective height of a ferrite-core loop to

$$h = \frac{2\pi N A \mu_{rod}}{\lambda} \qquad \text{(Eq 7)}$$

where

h = effective height (length) in meters
N = number of turns in the loop
A = area of loop in square meters
μ_{rod} = permeability of the ferrite rod
λ = wavelength of operation in meters

This obviously is a large increase in “collected” signal. If the rod permeability were 90, this would be the same as making the loop area 90 times larger with the same number of turns. For example, a 1.25-cm diameter ferrite-core loop would have an effective height equal to an air-core loop 22.5 cm in diameter (with the same number of turns).

By now you might have noticed we have been very careful to refer to rod permeability. There is a very important reason for this. The permeability that a rod of ferrite exhibits is a combination of the material permeability or μ, the shape of the rod, and the dimensions of the rod. In ferrite rods, μ is sometimes referred to as initial permeability, μ_i, or toroidal permeability, μ_{tor}. Because most amateur ferrite loops are in the form of rods, we will discuss only this shape.

The reason that μ_{rod} is different from μ is a very complex physics problem that is well beyond the scope of this book. For those interested in the details, books by Polydoroff and by Snelling cover this subject in considerable detail. (See Bibliography.) For our purposes a simple explanation will suffice. The rod is in fact not a perfect director of flux, as is illustrated in **Fig 11**. Note that some lines impinge on the sides of the core and also exit from the sides. These lines therefore would not pass through all the turns of the coil if it were wound from one end of the core to the other. These flux lines are referred to as *leakage flux*, or sometimes as flux leakage.

Leakage flux causes the flux density in the core to be nonuniform along its length. From Fig 11 it can be seen that the flux has a maximum at the geometric center of the length of the core, and decreases as the ends of the core are approached. This causes some noticeable effects. As a short coil is placed at different locations along a long core, its inductance will change. The maximum inductance exists when the coil is centered on the rod. The Q of a short coil on a long rod is greatest at the center. On the other hand, if you require a higher Q than this, it is recommended that you spread the coil turns along the whole length of the core, even though this will result in a lower value of inductance. (The inductance can be increased to the original value by adding turns.) **Fig 12** gives the relationship of rod permeability to material permeability for a variety of values.

The change in μ over the length of the rod results in an adjustment in the term μ_{rod} for its so called "free ends" (those not covered by the winding). This adjustment factor is given by

$$\mu' = \mu_{rod} \sqrt[3]{\frac{a}{b}} \qquad \text{(Eq 8)}$$

where

μ' = the corrected permeability
a = the length of the core
b = the length of the coil

This value of μ' should be used in place of μ_{rod} in Eq 7 to obtain the most accurate value of effective height.

All these variables make the calculation of ferrite loop antenna inductance somewhat less accurate than for the air-core version. The inductance of a ferrite loop is given by

$$L = \frac{4\pi N^2 A \mu_{rod} \times 10^{-4}}{\ell} \qquad \text{(Eq 9)}$$

where

L = inductance in μH
N = number of turns
A = cross-sectional area of the core in square mm
ℓ = magnetic length of core in mm

Experiments indicate that the winding diameter should be as close to that of the rod diameter as practical in order to maximize both inductance value and Q. By using all this information, we may determine the voltage at the loop terminals and its signal-to-noise ratio (SNR). The voltage may be determined from

$$V = \frac{2\pi A N \mu' Q E}{\lambda} \qquad \text{(Eq 10)}$$

where

V = output voltage across the loop terminals
A = loop area in square meters
N = number of turns in the loop winding
μ' = corrected rod permeability
Q = loaded Q of the loop
E = RF field strength in volts per meter
λ = wavelength of operation in meters

Lankford's equation for the sensitivity of the loop for a 10 dB SNR is

$$E = \frac{1.09 \times 10^{-10} \lambda \sqrt{f\ L\ b}}{A N \mu' \sqrt{Q}} \qquad \text{(Eq 11)}$$

where

f = operating frequency in Hz
L = loop inductance in henrys
b = receiver bandwidth in Hz

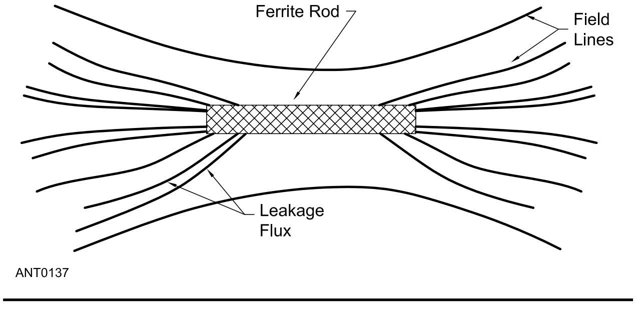

Fig 11—Example of magnetic field lines near a practical ferrite rod, showing leakage flux.

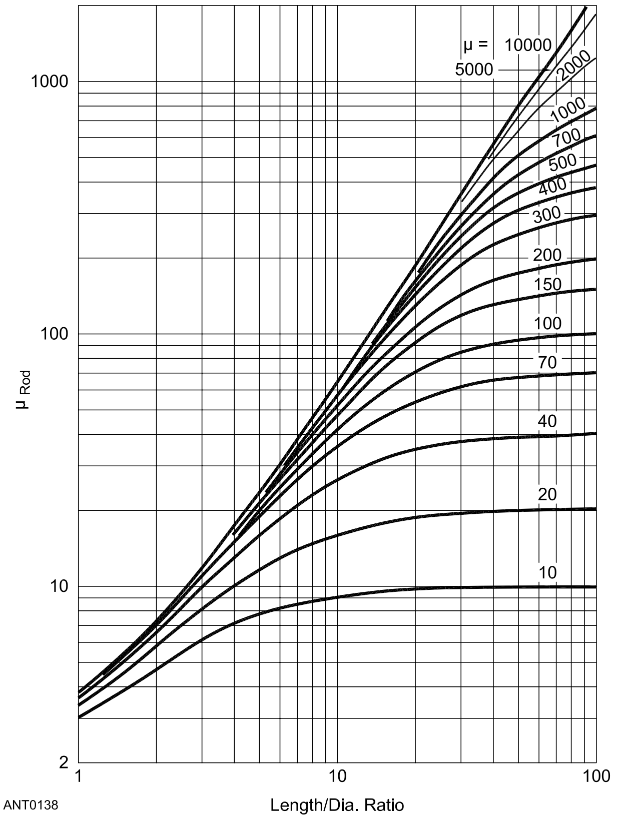

Fig 12—Rod permeability, μ_{rod}, versus material permeability, μ, for different rod length-to-diameter ratios.

Similarly, Belrose gives the SNR of a tuned loop antenna as

$$SNR = \frac{66.3 N A \mu_{rod} E}{\sqrt{b}} \sqrt{\frac{Qf}{L}} \quad \text{(Eq 12)}$$

From this, if the field strength E, μ_{rod}, b, and A are fixed, then Q or N must increase (or L decrease) to yield a better SNR. Higher sensitivity can also be obtained (especially at frequencies below 500 kHz) by bunching ferrite cores together to increase the loop area over that which would be possible with a single rod. High sensitivity is important because loop antennas are not the most efficient collectors of signals, but they do offer improvement over other receiving antennas in terms of SNR. For this reason, you should attempt to maximize the SNR when using a small loop receiving antenna. In some cases there may be physical constraints that limit how large you can make a ferrite-core loop.

After working through Eq 11 or 12, you might find you still require some increase in antenna system gain to effectively use your loop. In these cases the addition of a low noise preamplifier may be quite valuable even on the lower frequency bands where they are not commonly used. Chapter 14 contains information on such preamplifiers.

The electrostatic shield discussed earlier with reference to air-core loops can be used effectively with ferrite-core loops. (Construction examples are presented in Chapter 14.) As in the air-core loop, a shield will reduce electrical noise and improve loop balance.

PROPAGATION EFFECTS ON NULL DEPTH

After building a balanced loop you may find it does not approach the theoretical performance in the null depth. This problem may result from propagation effects. Tilting the loop away from a vertical plane may improve performance under some propagation conditions, to account for the vertical angle of arrival. Basically, the loop performs as described above only when the signal is arriving perpendicular to the axis of rotation of the loop. At incidence angles other than perpendicular, the position and depth of the nulls deteriorate.

The problem can be even further influenced by the fact that if the loop is situated over less than perfectly conductive ground, the wave front will appear to tilt or bend. (This bending is not always detrimental; in the case of Beverage antennas, sites are chosen to take advantage of this effect.)

Another cause of apparent poor performance in the null depth can be from polarization error. If the polarization of the signal is not completely linear, the nulls will not be sharp. In fact, for circularly polarized signals, the loop might appear to have almost no nulls. Propagation effects are discussed further in Chapter 14.

SITING EFFECTS ON THE LOOP

The location of the loop has an influence on its performance that at times may become quite noticeable. For ideal performance the loop should be located outdoors and clear of any large conductors, such as metallic downspouts and towers. A VLF loop, when mounted this way, will show good sharp nulls spaced 180° apart if the loop is well balanced. This is because the major propagation mode at VLF is by ground wave. At frequencies in the HF region, a significant portion of the signals is propagated by sky wave, and nulls are often only partial.

Most hams locate their loop antennas near their operating position. If you choose to locate a small loop indoors, its performance may show nulls of less than the expected depth, and some skewing of the pattern. For precision direction finding there may be some errors associated with wiring, plumbing, and other metallic construction members in the building. Also, a strong local signal may be reradiated from the surrounding conductors so that it cannot be nulled with any positioning of the loop. There appears to be no known method of curing this type of problem. All this should not discourage you from locating a loop indoors; this information is presented here only to give you an idea of some pitfalls. Many hams have reported excellent results with indoor mounted loops, in spite of some of the problems.

Locating a receiving loop in the field of a transmitting antenna may cause a large voltage to appear at the receiver antenna terminals. This may be sufficient to destroy sensitive RF amplifier transistors or front-end protection diodes. This can be solved by disconnecting your loop from the receiver during transmit periods. This can obviously be done automatically with a relay that opens when the transmitter is activated.

LOOP ANTENNA ARRAYS

Arrays of loop antennas, both in combination with each other and with other antenna types, have been used for many years. The arrays are generally used to cure some “deficiency” in the basic loop for a particular application, such as a 180° ambiguity in the null direction, low sensitivity, and so forth.

A Sensing Element

For direction-finding applications the single loop suffers the problem of having two nulls that are 180° apart. This leads to an ambiguity of 180° when trying to find the direction to a transmitting station from a given location. A sensing element (often called a *sense antenna*) may be added to the loop, causing the overall antenna to have a cardioid pattern and only one null. The sensing element is a small vertical antenna whose height is equal to or greater than the loop effective height. This vertical is physically close to the loop, and when its omnidirectional pattern is adjusted so that its amplitude and phase are equal to one of the loop lobes, the patterns combine to form a cardioid. This antenna can be made quite compact by use of a ferrite loop to form a portable DF antenna for HF direction finding. Chapter 14 contains additional information and construction projects using sensing elements.

Arrays of Loops

A more advanced array that can develop more diverse patterns consists of two or more loops. Their outputs are combined through appropriate phasing lines and combiners to form a phased array. Two loops can also be formed into an array that can be rotated without physically turning the loops themselves. This method was developed by Bellini and Tosi in 1907 and performs this apparently contradictory feat by use of a special transformer called a *goniometer*. The goniometer is described in Chapter 14.

Aperiodic Arrays

The aperiodic loop array is a wide-band antenna. This type of array is useful over at least a decade of frequency, such as 2 to 20 MHz. Unlike most of the loops discussed up to now, the loop elements in an aperiodic array are untuned. Such arrays have been used commercially for many years. One loop used in such an array is shown in **Fig 13**. This loop is quite different from all the loops discussed so far in this chapter because its pattern is not the familiar figure eight. Rather, it is omnidirectional.

The antenna is omnidirectional because it is purposely unbalanced, and also because the isolating resistor causes the antenna to appear as two closely spaced short monopoles. The loop maintains the omnidirectional characteristics over a frequency range of at least four or five to one. These loops, when combined into end-fire or broadside phased arrays, can provide quite impressive performance. A commercially made end-fire array of this type consisting of four loops equally spaced along a 25-meter baseline can provide gains in excess of 5 dBi over a range of 2 to 30 MHz. Over a considerable portion of this frequency range, the array can maintain F/B ratios of 10 dB. Even though the commercial version is very expensive, an amateur version can be constructed using the information provided by Lambert. One interesting feature of this type of array is that, with the proper combination of hybrids and combiners, the antenna can simultaneously feed two receivers with signals from different directions, as shown in **Fig 14**. This antenna may be especially interesting to one wanting a directional receiving array for two or more adjacent amateur bands.

SMALL TRANSMITTING LOOP ANTENNAS

The electrically small transmitting-loop antenna involves some different design considerations compared to receiving loops. Unlike receiving loops, the size limitations of the antenna are not as clearly defined. For most purposes, any transmitting loop whose physical circumference is less than ¼ λ can be considered "small." In most cases, as a consequence of their relatively large size (when compared to a receiving loop), transmitting loops have a nonuniform current distribution along their circumference. This leads to some performance changes from a receiving loop.

The transmitting loop is a parallel-tuned circuit with a large inductor acting as the radiator. As with the receiving

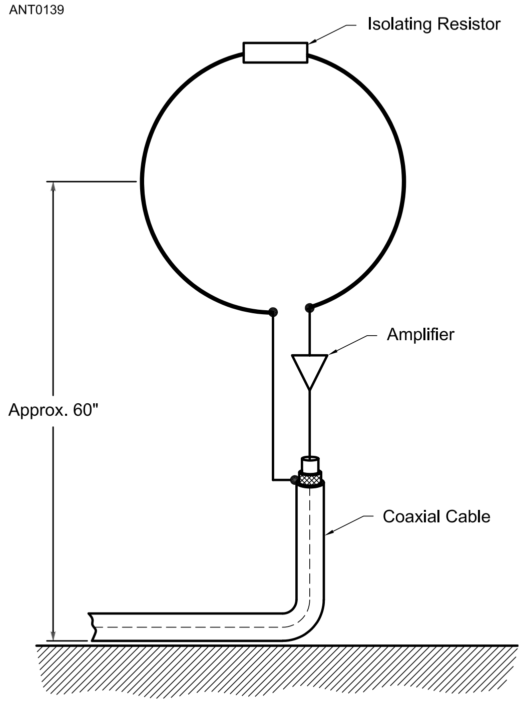

Fig 13—A single wide-band loop antenna used in an aperiodic array.

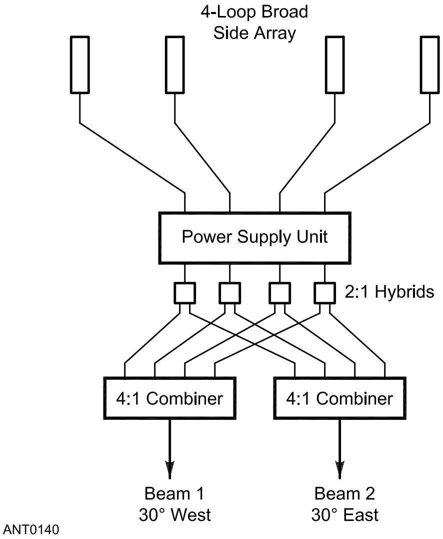

Fig 14—Block diagram of a four-loop broadside array with dual beams separated by 60° in azimuth.

Table 3
Transmitting Loop Equations

$$X_L = 2\pi fL \text{ ohms}$$

$$Q = \frac{f}{\Delta f} = \frac{X_L}{2\left(R_R + R_L\right)}$$

$$R_R = 3.12 \times 10^4 \left[\frac{NA}{\lambda^2}\right]^2 \text{ ohms}$$

$$V_C = \sqrt{PX_L\,Q}$$

$$I_L = \sqrt{\frac{PQ}{X_L}}$$

where

X_L = inductive reactance, ohms
f = frequency, Hz
Δf = bandwidth, Hz
R_R = radiation resistance, ohms
R_L = loss resistance, ohms (see text)
N = number of turns
A = area enclosed by loop, square meters
λ = wavelength at operating frequency, meters
V_C = voltage across capacitor
P = power, watts
I_L = resonant circulating current in loop

loop, the calculation of the transmitting-loop inductance may be carried out with the equations in Table 1. Avoid equations for long solenoids found in most texts. Other fundamental equations for transmitting loops are given in **Table 3**.

In the March 1968 *QST*, Lew McCoy, W1ICP, introduced the so-called “Army Loop” to radio amateurs. This was an amateur version of a loop designed for portable use in Southeast Asia by Patterson of the US Army and described in 1967. The Army Loop is diagrammed in **Fig 15A**, showing that this is a parallel tuned circuit fed by a tapped-capacitance impedance-matching network.

The Hart “high-efficiency” loop was introduced in the June 1986 *QST* by Ted Hart, W5QJR. It is shown schematically in Fig 15B and has the series-tuning capacitor separate from the matching network. The Hart matching network is basically a form of gamma match. Other designs have used a smaller loop connected to the transmission line to couple into the larger transmitting loop.

The approximate radiation resistance of a loop in ohms is given by

$$R_R = 3.12 \times 10^4 \left(\frac{NA}{\lambda^2}\right)^2 \qquad \text{(Eq 13)}$$

where

N = number of turns
A = area of loop in square meters
λ = wavelength of operation in meters

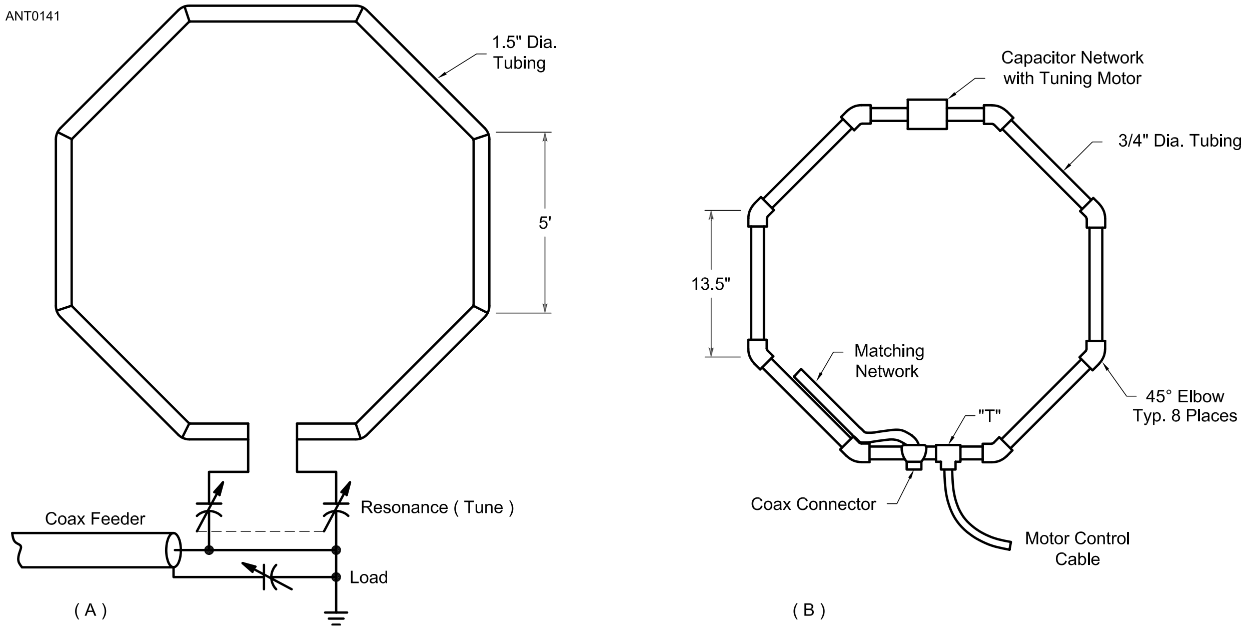

Fig 15—At A, a simplified diagram of the Army Loop. At B, the W5QJR loop, which is described in more detail later in this chapter.

The radiation resistance of a small transmitting loop is usually very small. For example, a 1-meter diameter, single-turn circular loop has a radius of 0.5 meters and an enclosed area of $\pi \times 0.5^2 = 0.785$ m^2. Operated at 14.0 MHz, the free-space wavelength is 21.4 meters and this leads to a computed radiation resistance of only $3.12 \times 10^{-4} (0.785/21.4^2)^2 = 0.092\ \Omega$.

Unfortunately the loop also has losses, both ohmic and from skin effect. By using this information, the radiation efficiency of a loop can be calculated from

$$\eta = \frac{R_R}{R_R + R_L} \qquad \text{(Eq 14)}$$

where

η = antenna efficiency, %

R_R = radiation resistance, Ω

R_L = loss resistance, Ω, which includes the loop's conductor loss plus the loss in the series-tuning capacitor.

A simple ratio of R_R versus R_L shows the effects on the efficiency, as can be seen from **Fig 16**. The loss resistance is primarily the ac resistance of the conductor. This can be calculated from Eq 6. A transmitting loop generally requires the use of copper conductors of at least ¾ inch in diameter in order to obtain reasonable efficiency. Tubing is as useful as a solid conductor because high-frequency currents flow only along a very small depth of the surface of the conductor; the center of the conductor has almost no effect on current flow.

Note that the R_L term above also includes the effect of the tuning capacitor's loss. Normally, the unloaded Q of a capacitor can be considered to be so high that any loss in the tuning capacitor can be neglected. For example, a very high-quality tuning capacitor with no mechanical wiping contacts, such as a vacuum-variable or a transmitting butterfly capacitor, might have an unloaded Q of about 5000. This implies a series loss resistance of less than about 0.02 Ω for a capacitive reactance of 100 Ω. This relatively tiny loss resistance can become significant, however, when the radiation resistance of the loop is only on the order of 0.1 Ω! Practical details for curbing capacitor losses are covered later in this chapter.

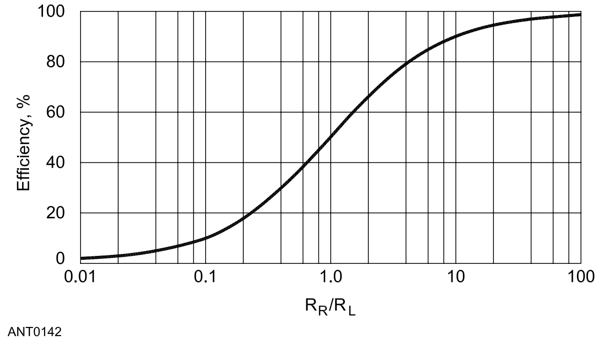

Fig 16—Effect of ratio of R_R/R_L on loop efficiency.

In the case of multiturn loops there is an additional loss related to a term called *proximity effect*. The proximity effect occurs in cases where the turns are closely spaced (such as being spaced one wire diameter apart). As these current-carrying conductors are brought close to each other, the current density around the circumference of each conductor gets redistributed. The result is that more current per square meter is flowing at the surfaces adjacent to other conductors. This means that the loss is higher than a simple skin-effect analysis would indicate, because the current is bunched so it flows through a smaller cross section of the conductor than if the other turns were not present.

As the efficiency of a loop approaches 90%, the proximity effect is less serious. But unfortunately, the less efficient the loop, the worse the effect. For example, an 8-turn transmitting loop with an efficiency of 10% (calculated by the skin-effect method) actually only has an efficiency of 3% because of the additional losses introduced by the proximity effect. If you are contemplating construction of a multiturn transmitting loop, you might want to consider spreading the conductors apart to reduce this effect. G. S. Smith includes graphs that detail this effect in his 1972 IEEE paper.

The components in a resonated transmitting loop are subject to both high currents and voltages as a result of the large circulating currents found in the high-Q tuned circuit formed by the antenna. This makes it important that any fixed capacitors have a high RF current rating, such as transmitting micas or the Centralab 850 series. Be aware that even a 100-W transmitter can develop currents in the tens of amperes, and voltages across the tuning capacitor in excess of 10,000 V. This consideration also applies to any conductors used to connect the loop to the capacitors. A piece of #14 wire may have more resistance than the rest of the loop conductor!

It is therefore best to use copper strips or the braid from a piece of large coax cable to make any connections. Make the best electrical connection possible, using soldered or welded joints. Using nuts and bolts should be avoided, because at RF these joints generally have high resistance, especially after being subjected to weathering.

An unfortunate consequence of having a small but high-efficiency transmitting loop is high loaded Q, and therefore limited bandwidth. This type of antenna may require retuning for frequency changes as little as 5 kHz. If you are using any wide-band mode such as AM or FM, this might cause fidelity problems and you might wish to sacrifice a little efficiency to obtain the required bandwidth.

A special case of the transmitting loop is that of the ferrite-loaded loop. This is a logical extension of the transmitting loop if we consider the improvement that a ferrite core makes in receiving loops. The use of ferrites in a transmitting loop is still under development. (See the Bibliography reference for DeVore and Bohley.)

PRACTICAL COMPACT TRANSMITTING LOOPS

The ideal small transmitting antenna would have performance equal to a large antenna. A small loop antenna can approach that performance except for a reduction in bandwidth, but that effect can be overcome by retuning. This section is adapted and updated from material written by Robert T. (Ted) Hart, W5QJR.

As pointed out above, small antennas are characterized by low radiation resistance. For a typical small antenna, such as a short dipole, loading coils are often added to achieve resonance. However, the loss inherent in the coils can result in an antenna with low efficiency. If instead of coils a large, low-loss capacitor is added to a low-loss conductor to achieve resonance, and if the antenna conductor is bent to connect the ends to the capacitor, a loop is formed.

Based on this concept, the small loop is capable of relatively high efficiency, compared to its coil-loaded cousin. In addition, the small loop, when mounted vertically, can radiate efficiently over the wide range of elevation angles required on the lower frequency bands. This is because it has both high-angle and low-angle response. See **Fig 17**, which shows the elevation response for a compact transmitting loop only 16.2 inches wide at 14.2 MHz. This loop is vertically polarized and its bottom is 8 feet above average ground, which has a conductivity of 5 mS/m and a dielectric constant of 13. For comparison, Fig 17 also shows the responses of three other reference antennas—the same small loop flipped sideways at a height of 30 feet to produce horizontal radiation, a full-sized ¼-λ ground plane antenna mounted 8 feet above average ground using two tuned radials, and finally a simple ½ λ flattop dipole mounted 30 feet above flat ground. The considerably smaller transmitting loop comes to within 3 dB of the larger ¼-λ vertical at a 10° elevation angle, and it is far stronger for high elevation angles because it does not have the null at high elevation angles that the ground plane has. Of course, this characteristic does make it more susceptible to strong signals received at high elevation angles. Incidentally, just in case you were wondering, adding more radials to the λ/4 ground plane doesn't materially improve its performance when mounted at an 8-foot height on 20 meters.

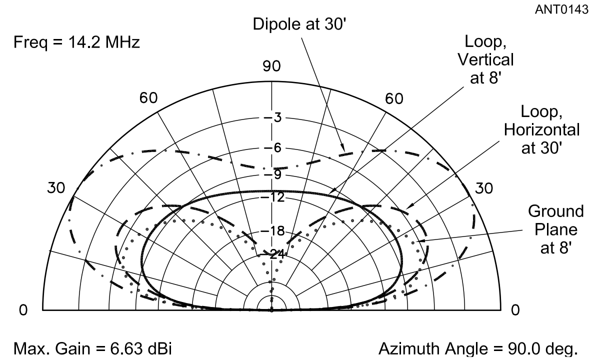

Fig 17—Elevation-plane plot at 14.2 MHz, showing response of an 8.5-foot circumference octagonal copper loop (width of 16.2 inches), compared to a full-sized λ/4 ground-plane vertical with two elevated λ/4 radials, the same small loop flipped horizontally at a height of 30 feet, and lastly, a λ/2 flattop dipole also at a height of 30 feet. Both the λ/4 ground-plane vertical and the vertically polarized loop are elevated 8 feet above typical ground, with σ = 5 mS/m and ε = 13. The low vertically polarized loop is surprisingly competitive, only down about 2.5 dB compared to the far larger ground plane at low elevation angles. Note that the vertical loop has both high-angle as well as low-angle radiation, and hence would be better at working close-in local stations than the ground-plane vertical, with its deep nulls at higher angles. The simple flattop dipole, however, is better than either vertical because of the poor ground reflection for a vertically polarized compared to a horizontally polarized signal.

The simple horizontal dipole in Fig 17 would be the clear winner in any shootout because its horizontally polarized radiation does not suffer as much attenuation at reflection from ground as does a vertically polarized wave. The case is not quite so clear-cut, however, for the small loop mounted horizontally at 30 feet. While it does have increased gain at medium elevation angles, it may not be worth the effort needed to mount it on a mast, considering the slight loss at low angles compared to its twin mounted vertically only 8 feet above ground.

A physically small antenna like the 16.2-inch-wide vertically polarized loop does put out an impressive signal compared to far larger competing antennas. Though somewhat ungainly, it is a substantially better performer than most mobile whips, for example. The main deficiency in a compact transmitting loop is its narrow bandwidth—it must be accurately tuned to the operating frequency. The use of a remote motor drive allows the loop to be tuned over a wide frequency range.

For example, for fixed-station use, two loops could be constructed to provide continuous frequency coverage from 3.5 to 30 MHz. A loop with an 8.5 foot circumference, 16 inches wide, could cover 10 through 30 MHz and a loop with a 20-foot circumference, 72 inches wide, could cover 3.5 to 10.1 MHz.

Table 4 presents summary data for various size loop antennas for the HF amateur bands. Through computer analysis, the optimum size conductor was determined to be 3/4-inch rigid copper water pipe, considering both performance and cost. Performance will be compromised, but only slightly, if 5/8-inch flexible copper tubing is used. This tubing can easily be bent to any desired shape, even a circle. The rigid 3/4-inch copper pipe is best used with 45° elbows to make an octagon.

The loop circumference should be between 1/4 and 1/8 λ at the operating frequency. It will become self-resonant above 1/4 λ, and efficiency drops rapidly below 1/8 λ. In the frequency ranges shown in Table 4, the high fre-

Table 4

Design Data for Loops

Loop Circumference = 8.5' (Width = 32.4"), Vertically Polarized

Frequency, MHz	10.1	14.2	21.2	29.0
Max Gain, dBi	–4.47	–1.42	+1.34	+2.97
Max Elevation Angle	40°	30°	22°	90°
Gain, dBi @10°	–8.40	–4.61	–0.87	+0.40
Total Capacitance, pF	145	70	29	13
Peak Capacitor kV	23	27	30	30

Loop Circumference = 8.5' (Width = 32.4"), Horizontally Polarized, @30'

Frequency, MHz	10.1	14.2	21.2	29.0
Max Gain, dBi	–3.06	+1.71	+5.43	+6.60
Max Elevation Angle	34°	28°	20°	16°
Gain, dBi @10°	–9.25	–3.11	+2.61	+5.34
Total Capacitance, pF	145	70	29	13
Peak Capacitor kV	23	27	30	30

Loop Circumference = 20' (Width = 6'), Vertically Polarized

Frequency, MHz	3.5	4.0	7.2	10.1
Max Gain, dBi	–7.40	–6.07	–1.69	–0.34
Max Elevation Angle	68°	60°	38°	30°
Gain, dBi @10°	–11.46	–10.12	–5.27	–3.33
Capacitance, pF	379	286	85	38
Peak Capacitor kV	22	24	26	30

Loop Circumference = 20' (Width = 6'), Horizontally Polarized, @30'

Frequency, MHz	3.5	4.0	7.2	10.1
Max Gain, dBi	–13.32	–10.60	–0.20	+3.20
Max Elevation Angle	42°	42°	38°	34°
Gain, dBi @10°	–21.62	–18.79	–7.51	–3.22
Capacitance, pF	379	286	85	38
Peak Capacitor kV	22	24	26	30

Loop Circumference = 38' (Width = 11.5'), Vertically Polarized

Frequency, MHz	3.5	4.0	7.2
Max Gain, dBi	–2.93	–2.20	–0.05
Max Elevation Angle	46°	42°	28°
Gain, dBi @10°	–6.48	–5.69	–2.80
Capacitance, pF	165	123	29
Peak Capacitor kV	26	27	33

Notes: These loops are octagonal in shape, constructed with ¾-inch copper water pipe and soldered 45° copper elbows. The gain figures assume a capacitor unloaded $Q_C = 5000$, typical for vacuum-variable type of tuning capacitor. The bottom of the loop is assumed to be 8 feet high for safety and the ground constants are "typical" at conductivity = 5 mS/m and dielectric constant = 13. Transmitter power is 1500 W. The voltage across the tuning capacitor for lower powers goes down with a multiplier of $\sqrt{\frac{P}{1500}}$. For example, at 100 W using the 38-foot-circumference loop at 7.2 MHz, the peak voltage would be $33\text{ kV} \times \sqrt{\frac{100}{1500}} = 8.5\text{ kV}$.

quency is tuned with a minimum capacitance of about 29 pF—including stray capacitance.

The low frequency listed in Table 4 is that where the loop response is down about 10 dB from that of a full-sized elevated ground plane at low elevation angles suitable for DX work. **Fig 18** shows an overlay at 3.5 MHz of the elevation responses for two loops: one with an 8.5-foot circumference and one with a 20-foot circumference, together with the response for a full-sized 80-meter ground plane elevated 8 feet off average ground with 2 tuned radials. The 20-foot circumference loop holds its own well compared to the full-sized ground plane.

Controlling Losses

Contrary to earlier reports, adding quarter-wave ground radials underneath a vertically polarized transmitting loop doesn't materially increase loop efficiency. The size of the conductor used for a transmitting loop, however, does directly affect several interrelated aspects of loop performance.

Data for Table 4 was computed for ¾-inch copper water pipe (nominal OD of 0.9 inch). Note that the efficiency is higher and the Q is lower for loops having a circumference near ¼ λ. Larger pipe size will reduce the loss resistance, but the Q increases. Therefore the bandwidth decreases, and the voltage across the tuning capacitor increases. The voltage across the tuning capacitor for high-power operation can become very impressive, as shown in Table 4. Rigid ¾-inch copper water pipe is a good electrical compromise and can also help make a small-diameter loop mechanically sturdy.

The equivalent electrical circuit for the loop is a parallel resonant circuit with a very high Q, and therefore a narrow bandwidth. The efficiency is a function of radiation resistance divided by the sum of the radiation plus loss resistances. The radiation resistance is much less than 1 Ω, so it is necessary to minimize the loss resistance, which is largely the skin-effect loss of the conductor, assuming that the tuning capacitor has very low loss. Poor construction techniques must be avoided. All joints in the loop must be brazed or soldered.

However, if the system loss is too low, for example by using even larger diameter tubing, the Q may become excessive and the bandwidth may become too narrow for practical use. These reasons dictate the need for a complete analysis to be performed before proceeding with the construction of a loop.

There is another source of additional loss in a completed loop antenna besides the conductor and capacitor losses. If the loop is mounted near lossy metallic conductors, the large magnetic field produced will induce currents into those conductors and be reflected as losses in the loop. Therefore the loop should be as far from other conductors as possible. If you use the loop inside a building constructed with large amounts of iron or near ferrous materials, you will simply have to live with the loss if the loop cannot otherwise be relocated.

The Tuning Capacitor

Fig 19 demonstrates the selection of loop size versus tuning capacitance for any desired operating frequency range for the HF amateur bands. This is for octagonal-shaped loops using ¾-inch copper water pipe with 45° copper elbows. For example, a capacitor that varies from 5 to 50 pF, used with a loop 10 feet in circumference, tunes from 13 to 27 MHz (represented by the left dark vertical bar). A 25 to 150-pF capacitor with a 13.5-foot loop circumference covers the 7 to 14.4-MHz range, represented by the right vertical bar.

Fig 20 illustrates how the 29-MHz elevation pattern becomes distorted and rather bulbous-looking for the 10-foot circumference loop, although the response at low elevation angles is still better than that of a full-sized ground-plane antenna.

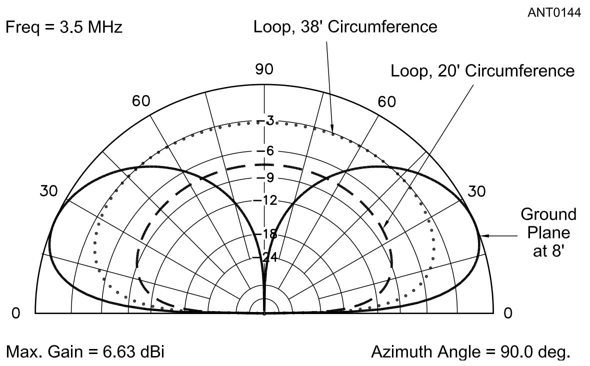

Fig 18—Elevation-plane response of three antennas at 3.5 MHz—a 20-foot circumference octagonal copper loop, a 38-foot circumference copper loop and a full-sized λ/4 ground plane with two elevated radials. The bottom of each antenna is mounted 8 feet above ground for safety. The 38-foot circumference loop (which has a "wingspan" of 11.5 feet) is fairly competitive with the much large ground-plane, being down only about 4 dB at low elevation angles. The 20-foot circumference loop is much more lossy, but with its top only about 14 feet off the ground is very much of a "stealth" antenna.

Air Variable Capacitors

Special care must be taken with the tuning capacitor if an air-variable type is used. The use of a split-stator capacitor eliminates the resistance of wiper contacts, resistance that is inherent in a single-section capacitor. The ends of the loop are connected to the stators, and the rotor forms the variable coupling path between the stators. With this arrangement the value of capacitance is divided by two, but the voltage rating is doubled.

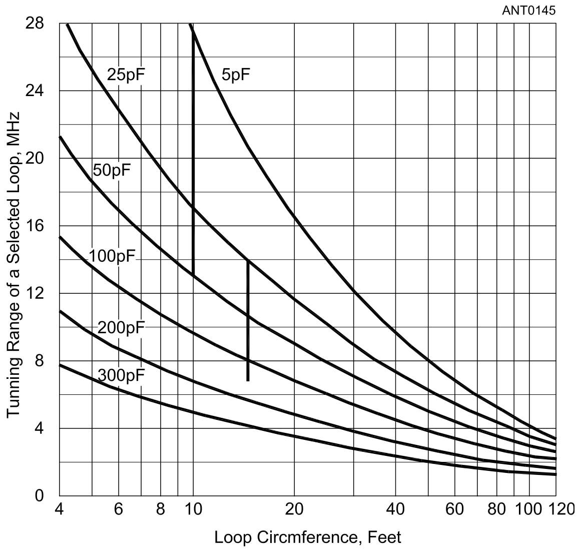

Fig 19—Frequency tuning range of an octagon-shaped loop using ¾-inch copper water pipe, for various values of tuning capacitance and loop circumference.

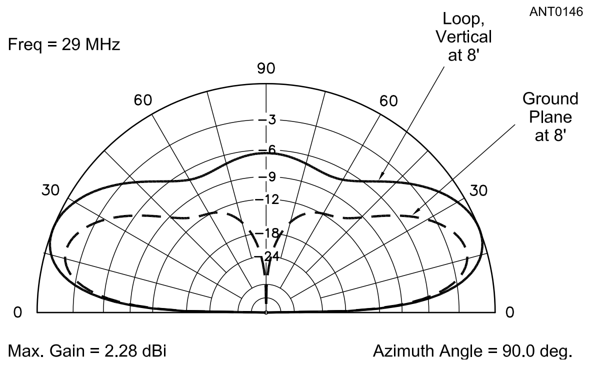

Fig 20—Elevation-plane plot for a 16.2-inch wingspan octagonal copper loop at 29 MHz, compared to a λ/4 ground-plane antenna with two resonant elevated radials. The gains at low angles are almost identical, but the loop exhibits more gain at medium and high elevation angles. Again, the bottom of each antenna is located 8 feet above ground for safety.

You must carefully select a variable capacitor for transmitting-loop application—that is, all contacts must be welded, and no mechanical wiping contacts are allowed. For example, if the spacers between plates are not welded to the plates, there will be loss at each joint, and thus degraded loop efficiency. (Earlier transmitting loops exhibited poor efficiency because capacitors with wiping contacts were used.)

There are several suitable types of capacitors for this application. A vacuum variable is an excellent choice, provided one is selected with an adequate voltage rating. Unfortunately, those capacitors are very expensive.

W5QJR used a specially modified air-variable capacitor in his designs. This had up to 340 pF maximum per section, with ¼-inch spacing, resulting in 170 pF when both sections were in series as a butterfly capacitor. Another alternative is to obtain a large air variable, remove the aluminum plates, and replace them with copper or double-sided PC board material to reduce losses. Connect all plates together on the rotor and on the stators. Solder copper straps to the capacitor for soldering to the loop itself.

The spacing between plates in an air-variable capacitor determines the voltage-handling capability, rated at 75,000 V per inch. For other power ratings, multiply the spacing (and voltage) by the square root of the ratio of your power to 1000 W. For example, for 100 W, the ratio would be = 0.316.

A Teflon-Insulated Trombone Variable Capacitor

Another type of variable capacitor discussed in the amateur literature for use with a compact transmitting loop is the so-called "trombone" type of capacitor. **Fig 21** shows a practical trombone capacitor created by Bill Jones, KD7S, for Nov 1994 *QST*. This capacitor uses downward pointing extensions of the two ¾-inch OD main conductor copper pipes, with a Teflon-insulated trombone section made of ½-inch ID copper pipe. The trombone telescopes into the main pipes, driven by a lead screw and a 180-rpm gear-head motor. Like the butterfly air variable capacitor, the trombone works without lossy wiper contacts. Jones' capacitor varied from 12 pF (including strays) to almost 60 pF, making it suitable to tune his 3-foot circumference loop from 14 to 30 MHz at the 100-W level.

KD7S used 5-mil (0.005 inch) thick Teflon sheet as an insulator. Since Teflon is conservatively rated at more than 1 kV per mil of thickness, the voltage breakdown capability of this capacitor is well in excess of 5 kV. The parts list is given in **Table 5**.

Table 5

KD7S Loop-Tuning Capacitor Parts List for Nominal 50-pF Capacitor

Qty	*Description*
2	10-inch length of ¾-inch-ID type M copper water pipe
2	10-inch length of ½-inch-ID type M copper water pipe
1	3-inch length of ½-inch-ID type M copper water pipe
2	½-inch, 90° copper elbows
2	¾-inch, 90° copper elbows
2	10 × 22-inch piece of 0.005-inch-thick Teflon sheet plastic
1	12-inch length of #8-32 threaded brass rod
1	#8-32 brass shoulder nut
2	22 × 5½ × ¼-inch ABS plastic sheet (top and bottom covers)
3	1 × 5½ × ¼-inch ABS plastic sheet (end pieces and center) brace/guide
2	1 × 22 × ¼-inch ABS plastic sheet (side rails)
1	50 to 200-rpm gear-head dc motor
1	DPDT center-off toggle switch (up/down control)
2	SPDT microswitches (limit switches)
50 feet	3-conductor control cable
1	Enclosure for control switch

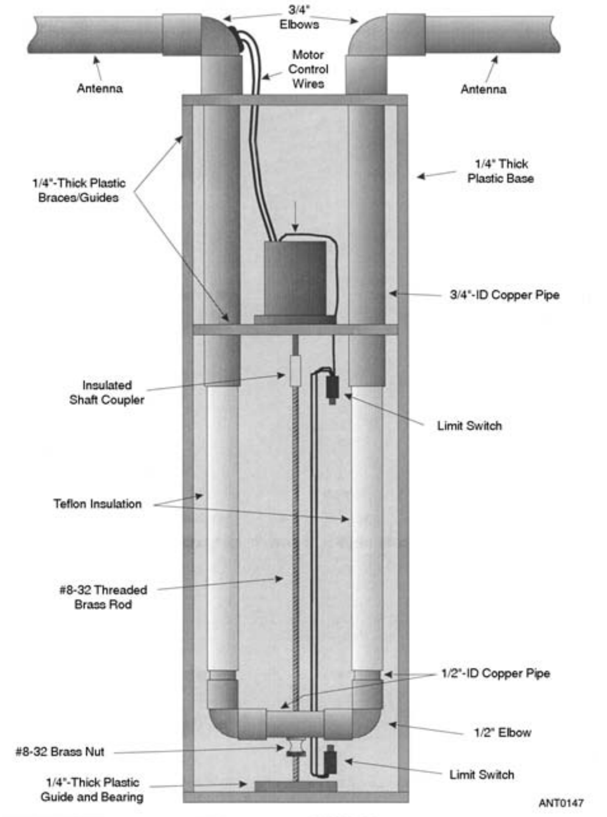

Fig 21—A practical trombone capacitor designed by Bill Jones, KD7S, for his compact transmitting loop. This capacitor has a tuning range from 12 to almost 60 pF, and can withstand at least 5 kV peak. The 10-inch ½-inch ID tubes are covered with Teflon-sheet insulation and slide into the ¾-inch ID copper pipes.

A short length of plastic tubing connects the threaded brass rod to the motor. The tubing acts as an insulator and a flexible coupling to smooth out minor shaft-alignment errors. The other end of the rod is threaded into a brass nut soldered to the crossbar holding the $^1/_2$-inch pipes together. Jones used a 12-V motor rated at 180 rpm, but it has sufficient torque to work with as little as 4 V applied. Instead of a sophisticated variable duty-cycle speed control circuit, he used an LM327 adjustable voltage regulator to vary the motor-control voltage from 4 to 12 V. Tuning speeds ranged from 11 seconds per inch at 12 V to 40 seconds per inch at 4 V. The higher speed is necessary to jump from band to band in a reasonable length of time. The lower speed makes it easy to fine-tune the capacitor to any desired frequency within a band.

When building the capacitor, keep in mind that the smaller tubes must telescope in and out of the larger tubes with silky smoothness. Any binding will cause erratic tuning. For the same reason, the #8-32 brass threaded rod must be straight and properly aligned with the brass nut. *Take your time with this part of the project*.

Perhaps the easiest way to form the insulator is to precut a length of Teflon sheet to the proper size. Place a lengthwise strip of double-sided tape on the tube to secure one end of the Teflon sheet. Begin wrapping the Teflon around the tube while keeping it as tight as possible. *Don't allow wrinkles or ridges to form*. Secure the other end with another piece of tape. Once both tubes are covered, ensure they are just short of being a snug fit inside the larger tubes. Confirm that the insulation completely overlaps the open end of the small tubes. If not, the capacitor is certain to arc internally with more than a few watts of power applied to it.

Route the motor wiring inside the antenna pipes to minimize the amount of metal within the field of the antenna. Bring the wires out next to the coaxial connector. A three-wire system allows the use of limit switches to restrict the movement of the trombone section. Be sure to solder together all metal parts of the capacitor. Use a small propane torch, a good quality flux and 50/50 solid solder. Do not use acid-core solder! Clean all parts to be joined with steel wool prior to coating them with flux.

A Cookie-Sheet and Picture-Frame-Glass Variable Capacitor

In Vol 2 of *The ARRL Antenna Compendium* series, Richard Plasencia, WØRPV, described a clever high-voltage variable capacitor he constructed using readily available materials. See **Fig 22**, which shows Plasencia's homebrew high-voltage variable capacitor, along with the coil and other parts used in his homemade antenna coupler. This capacitor could be varied from 16 to 542 pF and tested at a breakdown of 12,000 V.

The capacitor sits on four PVC pillars and consists of two $4^1/_2 \times 4^1/_2$-inch aluminum plates separated by a piece of window glass that is $8^1/_2 \times 5^1/_2$ inches in size. The lower plate is epoxied to the glass. The upper plate is free to move in a wooden track epoxied to the upper surface of the glass. The motor is reversible and moves the upper capacitor plate by rotating a threaded rod in a wing nut pinned to a tab on the capacitor plate. The four pillars are cut from PVC pipe to insulate the capacitor from the chassis and to elevate it into alignment with the motor shaft.

WØRPV used a piece of 0.063-inch thick single-weight glass that exhibited a dielectric constant of 8. He removed the glass from a dime-store picture frame. In time-honored ham fashion, he improvised his wooden tracks for the upper capacitor plate from a single wooden paint stirrer, and for the capacitor plates, he used aluminum cookie sheets.

The wooden track for the upper plate is made by splitting the wooden paint stirrer with a knife into one narrow and one wide strip. The narrow strip is cemented on top and overhangs the movable plate, creating a slotted track. Since the wood is supported by the glass plate, its insulating qualities are of no importance.

The principle of operation is simple. The reversible motor turns a threaded $^1/_4$-inch rod with a pitch of 20 threads to the inch. This rod engages a wing nut attached to the movable capacitor plate. Although WØRPV grounded his capacitor's movable plate with a braid, an insulator similar to that used in the trombone capacitor above should be used to isolate the lead-screw mechanism. Several pieces of braid made from RG-8 coax shield should be used to connect to the ends of the compact transmitting loop conductors to form low-loss connections.

WØRPV used a 90-rpm motor from a surplus vending machine. It moved his variable capacitor plate $4^1/_2$ inches, taking about a minute to travel from one end to the other. Since he wished to eliminate the complexity and dubious reliability of limit switches when used outdoors, he monitored the motor's dc current through two 3 Ω, 2W resistors

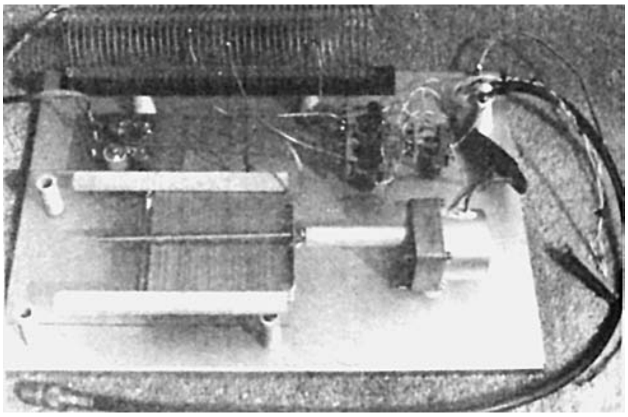

Fig 22—The picture-frame-glass variable capacitor design of Richard Plasencia, WØRPV. Two aluminum plates separated by a piece of glass scavenged from a picture frame create a variable capacitor that can withstand 12,000 V, with a variable range from 16 to 542 pF.

placed in series with each lead of the motor and shunted by red LEDs at the control box. When the motor stalled by jamming up against the PVC limit stop or against the inside of the plastic mounting box, the increased motor current caused one or the other of the LEDs to light up.

TYPICAL LOOP CONSTRUCTION

After you select the electrical design for your loop application, you must consider how to mount it and how to feed it. If you wish to cover only the upper HF bands of 20 through 10 meters, you will probably choose a loop that has a circumference of about 8.5 feet. You can make a reasonably sturdy loop using 1-inch diameter PVC pipe and $^5/_8$-inch flexible copper tubing bent into the shape of a circle. Robert Capon, WA3ULH, did this for a QRP-level transmitting loop described in May 1994 *QST*. **Fig 23** shows a picture of his loop, with PVC H-frame stand.

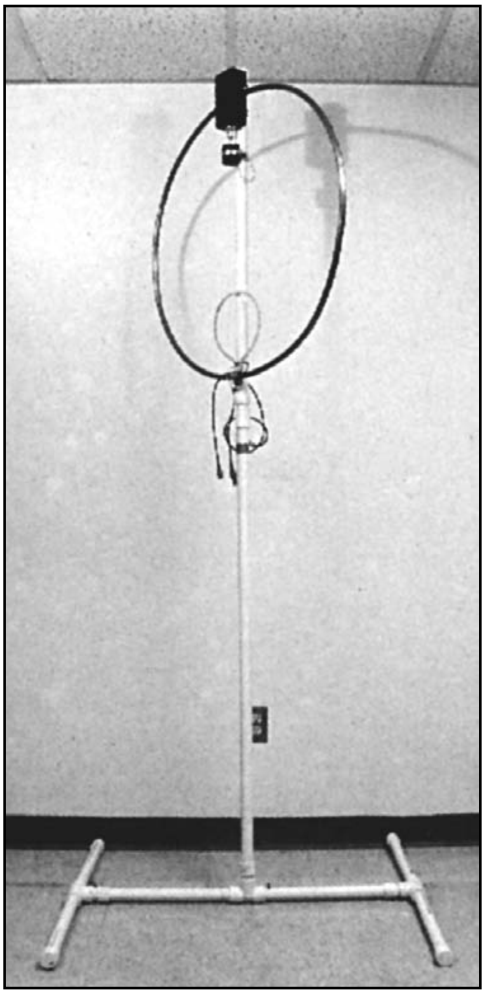

Fig 23—Photo of compact transmitting loop designed by Robert Capon, WA3ULH. This uses a 1-inch PVC H-frame to support the loop made of flexible $^5/_8$-inch copper tubing. The small coupling loop made of RG-8 coax braid couples the loop to the coax feed line. The tuning capacitor and drive motor are at the top of the loop, shown here in the ARRL Laboratory during testing.

This loop design used a 20-inch long coupling loop made of RG-8 coax to magnetically couple into the transmitting loop rather than the gamma-match arrangement used by W5QJR in his loop designs. The coupling loop was fastened to the PVC pipe frame using 2-inch long #8 bolts that also held the main loop to the mast.

A more rugged loop can be constructed using rigid $^3/_4$-inch copper water pipe, as shown in the W5QJR design in **Fig 24**. While a round loop is theoretically a bit more efficient, an octagonal shape is much easier to construct. The values presented in Table 4 are for octagons.

For a given loop circumference, divide the circumference by 8 and cut eight equal-length pieces of $^3/_4$ inch copper water pipe. Join the pieces with 45° elbows to form the octagon. With the loop lying on the ground on scraps of 2 × 4 lumber, braze or solder all joints.

W5QJR made a box from clear plastic to house his air-variable capacitor and drive motor at the top of the loop. The side of the box that mounts to the loop and the capacitor should be at least ¼-inch thick, preferably $^3/_8$-inch. The remainder of the box can be $^1/_8$-inch plastic sheet. He mounted the loop to the plastic using $^1/_4$-inch bolts (two on either side of center) after cutting out a section of pipe 2 inches wide in the center. On the motor side of the capacitor, he cut the pipe and installed a copper T for the motor wiring.

W5QJR's next step was to solder copper straps to the loop ends and to the capacitor stators, then he remounted the loop to the plastic. If you insert wood dowels, the pipe will remain round when you tighten the bolts. Next he installed the motor drive cable through the loop and connected it to the motor. Antenna rotator cable is a good choice for this cable. He completed the plastic box using short pieces of aluminum angle and small sheet-metal screws to join the pieces.

The loop was then ready to raise to the vertical position. Remember, no metal is allowed near the loop. W5QJR made a pole of 2 × 4-inch lumber with 1 × 4-inch boards on either side to form an I section. He held the boards together with $^1/_4$-inch bolts, 2 feet apart and tied rope guys to the top. This made an excellent mast up to 50 feet high. The pole height should be one foot greater than the loop diameter, to allow room for cutting grass or weeds at the bottom of the loop. W5QJR installed a pulley at the top so that his loop could be raised, supported by rope. He supported the bottom of the loop by tying it to the pole and tied guy ropes to the sides of the loop to keep it from rotating in the wind. By moving the anchor points, he could rotate his loop in the azimuth plane.

W5QJR used a gamma-matching arrangement made of flexible $^1/_4$-inch copper tubing to couple the loop to the transmission line. In the center of one leg, he cut the pipe and installed a copper T. Adjacent to the T, he installed a mount for the coax connector. He made the mount from copper strap, which can be obtained by splitting a short piece of pipe and hammering it flat.

While the loop was in the vertical position he cut a piece of $^1/_4$-inch flexible copper tubing the length of one

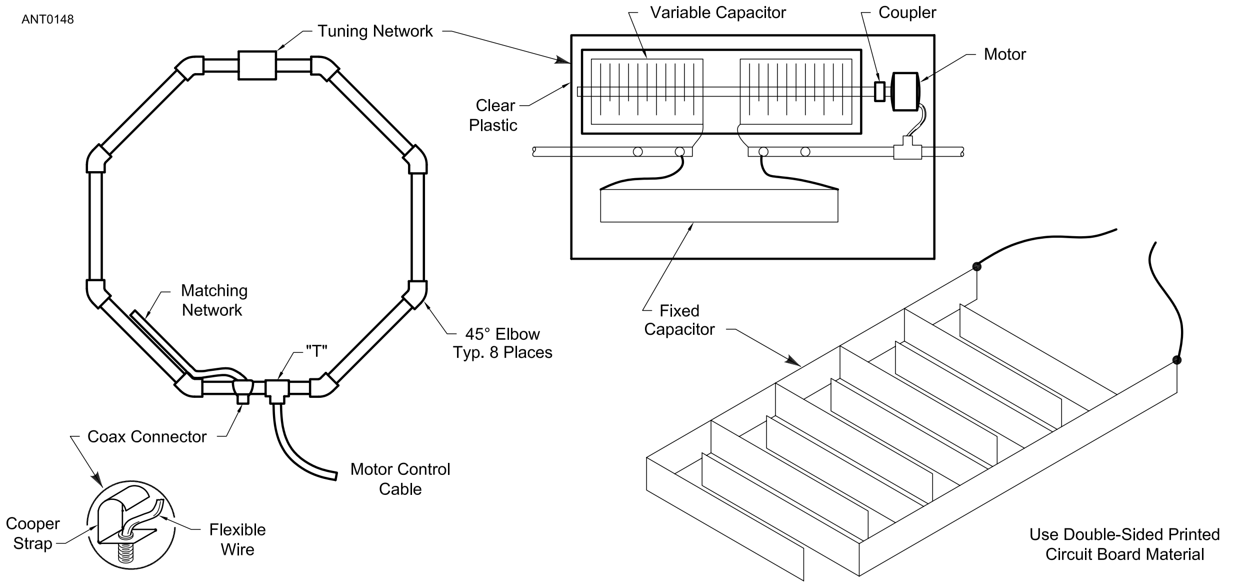

Fig 24—Octagonal loop construction details. Table 4 gives loop design data for various frequency ranges.

of the straight sides of the loop. He then flattened one end and soldered a piece of flexible wire to the other. He wrapped the tubing with electrical tape for insulation and connected the flexible wire to the coax connector. He then installed the tubing against the inside of the loop, held temporarily in place with tape. He soldered the flat part to the loop, ending up with a form of gamma match, but without reactive components. This simple feed provided better than 1.7:1 SWR over a 2:1 frequency range. For safety, he installed a good ground rod under the loop and connected it to the strap for the coax connector, using large flexible wire.

TUNE-UP PROCEDURE

The resonant frequency of the loop can be readily found by setting the receiver to a desired frequency and rotating the capacitor (by remote control) until signals peak. The peak will be very sharp because of the high Q of the loop.

Turn on the transmitter in the tune mode and adjust either the transmitter frequency or the loop capacitor for maximum signal on a field-strength meter, or for maximum forward signal on an SWR bridge. Adjust the matching network for minimum SWR by bending the matching line. Normally a small hump in the $^1/_4$-inch tubing line, as shown in Fig 24, will give the desired results. For a loop that covers two or more bands, adjust the feed to give equally low SWR at each end of the tubing range. The SWR will be very low in the center of the tuning range but will rise at each end.

If there is metal near the loop, the additional loss will reduce the Q and therefore the impedance of the loop. In those cases it will be necessary to increase the length of the matching line and tap higher up on the loop to obtain a 50-Ω match.

PERFORMANCE COMPARISON

As previously indicated, a compact transmitting loop can provide performance approaching full-size dipoles and verticals. To illustrate one case, a loop 100 feet in circumference would be 30 feet high for 1.8 MHz. However, a good dipole would be 240 feet ($^1/_2$ λ) in length and at least 120 feet high ($^1/_4$ λ). A $^1/_4$-λ vertical would be 120 feet tall with a large number of radials on the ground, each 120 feet in length. The smaller loop could replace both of those antennas with only a moderate degradation in performance and a requirement for a high-voltage variable capacitor.

On the higher frequencies, the same ratios apply, but full-size antennas are less dramatic. However, very few city dwellers can erect good verticals even on 7 MHz with a full-size counterpoise. Even on 14 MHz a loop about 3 feet high can work the world.

Other than trading small size for narrow bandwidth and a high-voltage capacitor, the compact transmitting loop is an excellent antenna and should find use where large antennas are not practical.

The Loop Skywire

Are you looking for a multiband HF antenna that is easy to construct, costs nearly nothing and yet works well? You might want to try this one. The *Loop Skywire* antenna is a full-sized horizontal loop. Early proponents suggested that the antenna could be fed with coaxial cable with little concern for losses, but later analysis proved that this was a bit of wishful thinking—the relatively low values for SWR across multiple bands indicate that cable losses were part and parcel of performance. The best way to feed this versatile antenna is with open-wire ladder line, with an antenna tuner in the shack to present the transmitter with a low value of SWR.

THE DESIGN

The Loop Skywire is shown in **Fig 25**. The antenna has one wavelength of wire in its perimeter at the design or fundamental frequency. If you choose to calculate L_{total} in feet, the following equation should be used:

$$L_{total} = \frac{1005}{f}$$

where f equals the frequency in MHz.

Given any length of wire, the maximum possible area the antenna can enclose is with the wire in the shape of a circle. Since it takes an infinite number of supports to hang a circular loop, the square loop (four supports) is the most practical. Further reducing the area enclosed by the wire loop (fewer supports) brings the antenna closer to the properties of the folded dipole, and both harmonic-impedance and feed-line voltage problems can result. Loop geometries other than a square are thus possible, but remember the two fundamental requirements for the Loop Skywire—its horizontal position and maximum enclosed area.

There is another great advantage to this antenna system. It can be operated as a vertical antenna with top-hat loading on other bands as well. This is accomplished by simply keeping the feed line run from the antenna to the shack as vertical as possible and clear of objects. Both feed-line conductors are then tied together, and the antenna is fed against a good ground.

CONSTRUCTION

Antenna construction is simple. Although the loop can be made for any band or frequency of operation, the following two Loop Skywires are good performers. The 10-MHz band can also be operated on both.

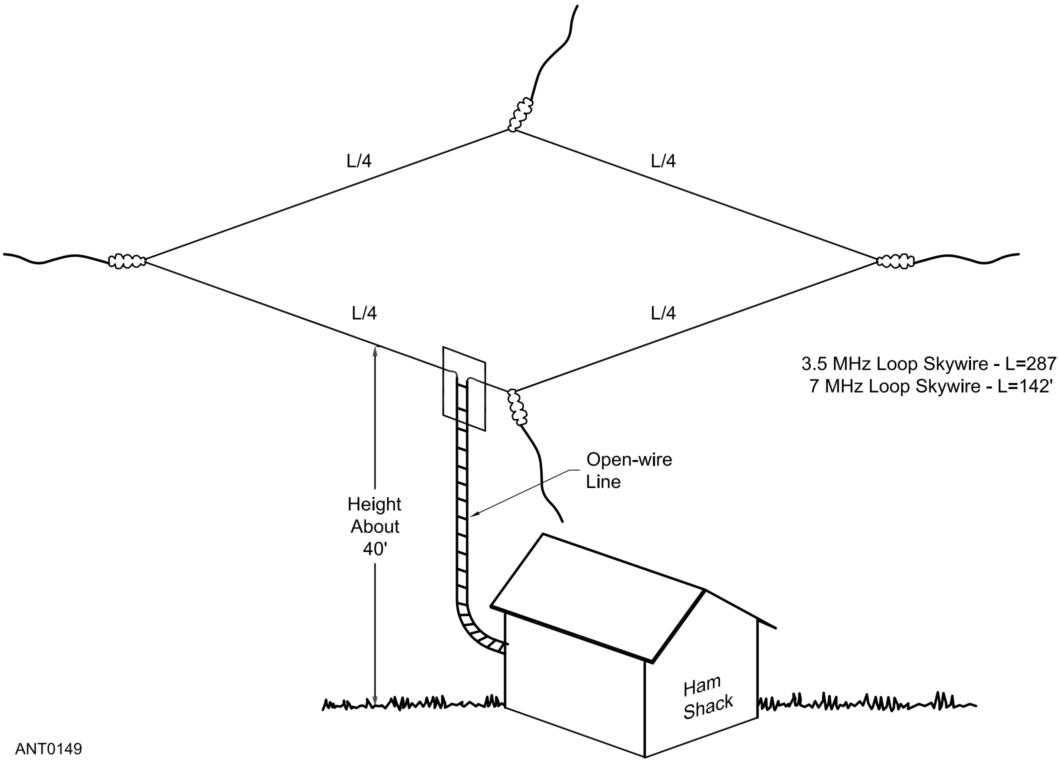

Fig 25—A complete view of the Loop Skywire. The square loop is erected horizontal to the earth.

3.5-MHz Loop Skywire
(3.5-28 MHz loop and 1.8-MHz vertical)
Total loop perimeter: 272 feet
Square side length: 68 feet

7-MHz Loop Skywire
(7-28 MHz loop and 3.5-MHz vertical)
Total loop perimeter: 142 feet
Square side length: 35.5 feet

The actual total length can vary from the above by a few feet, as the length is not at all critical. Do not worry about tuning and pruning the loop to resonance. No signal difference will be detected on the other end when that method is used.

Bare #14 copper wire is used in the loop. **Fig 26** shows the placement of the insulators at the loop corners. Two common methods are used to attach the insulators. Either lock or tie the insulator in place with a loop wire tie, as shown in Fig 26A, or leave the insulator free to "float" or slide along the wire, Fig 26B. Most loop users float at least two insulators. This allows pulling the slack out of the loop once it is in the air, and eliminates the need to have all the supports exactly placed for proper tension in each leg. Floating two opposite corners is recommended.

Fig 27A shows the azimuth-plane performance on 7.2 MHz of a 142-foot long, 7-MHz Loop Skywire, 40 feet high at an elevation angle of 10º, compared to a regular flattop $^{1}/_{2}$-λ dipole at a height of 30 feet. The loop comes into its own at higher frequencies. Fig 27B shows the response at 14.2 MHz, compared again to a $^{1}/_{2}$–λ 14.2-MHz

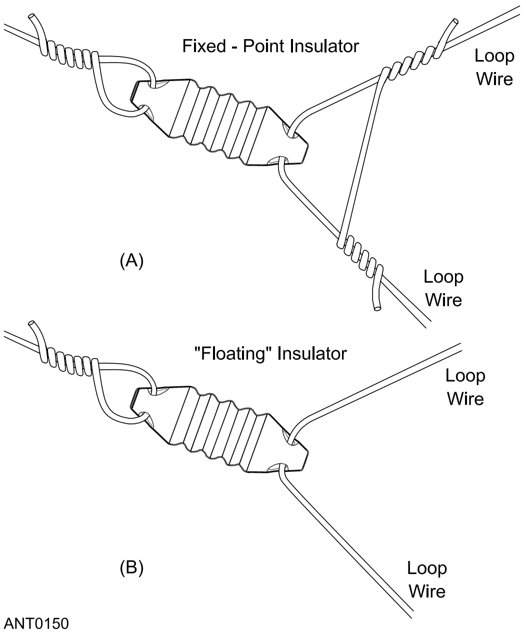

Fig 26—Two methods of installing the insulators at the loop corners.

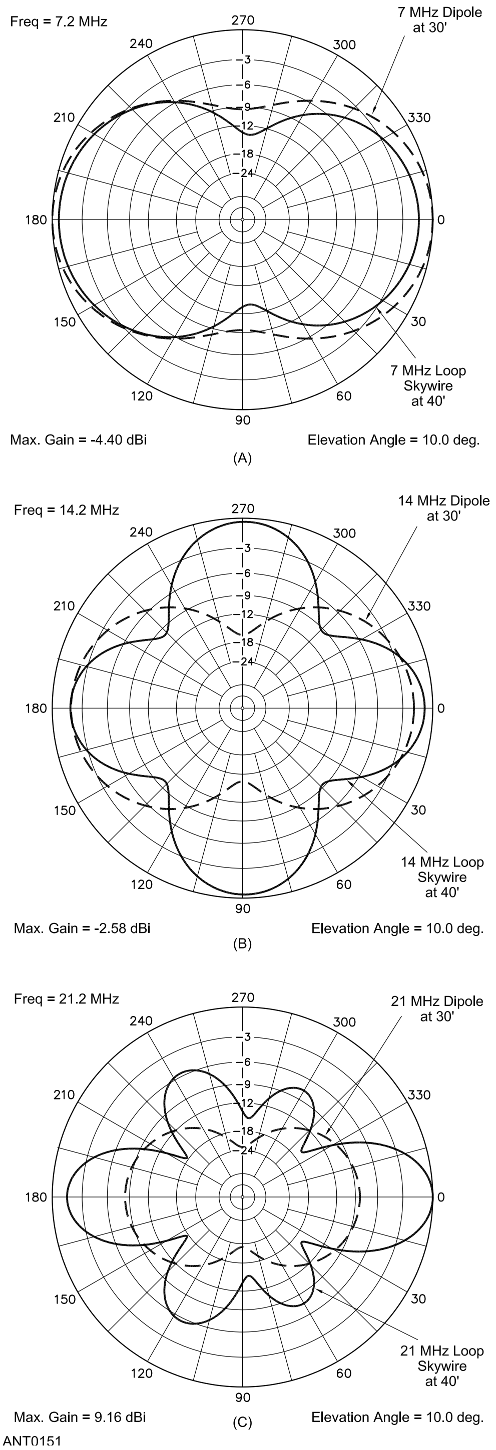

Fig 27—At A, azimuth-plane response of 142-foot long, 7-MHz Loop Skywire, 40 feet in the air at 7.2 MHz, compared with ½-λ dipole 30 feet in the air. At B, response of same Loop Skywire at 14.2 MHz, compared with ½-λ 14.2-MHz dipole 30 feet in the air. Now the loop has some advantage in certain directions. At C, response of the same Loop Skywire at 21.2 MHz compared to a 21.2-MHz dipole at 30 feet. Here, the Loop Skywire has more gain in almost all directions than the simple dipole. All azimuth-plane patterns were made at 10° elevation.

dipole at a height of 30 feet. Now the loop has several lobes that are stronger than the dipole. Fig 27C shows the response at 21.2 MHz, compared to a dipole. Now the loop has superior gain compared to the $^1/_2$-λ dipole at almost any azimuth. In its favored direction on 21.2 MHz, the loop is 8 dB stronger than the dipole.

The feed point can be positioned anywhere along the loop that you wish. However, most users feed the Skywire at a corner. **Fig 28** depicts a method of doing this, using a piece of plexiglass to provide insulation as well as strain relief for the open-wire ladder line. It is advantageous to keep the feed-point mechanicals away from the corner support. Feeding a foot or so from one corner allows the feed line to exit more freely. This method keeps the feed line free from the loop support.

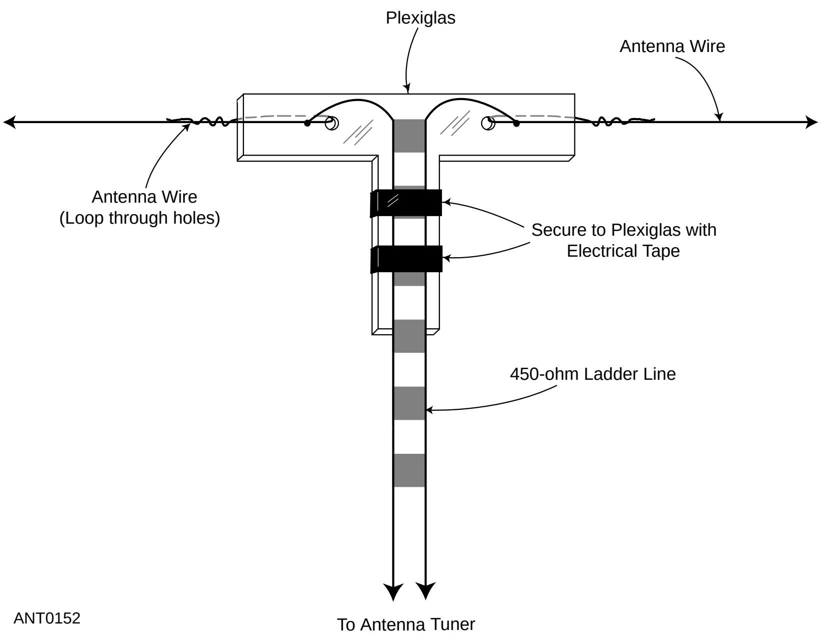

Fig 28—Most users feed the Skywire at a corner. A high-impedance weather-resistant insulator should be used for the feed-point insulator.

Generally a minimum of four supports is required. If trees are used for supports, then at least two of the ropes or guys used to support the insulators should be counterweighted and allowed to move freely. The feedline corner is almost always tied down, however. Very little tension is needed to support the loop (far less than that for a dipole). Thus, counterweights are light. Several such loops have been constructed with bungie cords tied to three of the four insulators. This eliminates the need for counterweighting.

Recommended height for the antenna is 40 feet or more. The higher the better, especially if you wish to use the loop in the vertical mode. However, successful local and DX operation has been reported in several cases with the antenna at 20 feet. **Fig 29** shows the feed arrangement for using the Loop Skywire as a top-loaded vertical fed against ground on the lower bands.

Because the loop is high in the air and has considerable electrical exposure to the elements, proper methods should be employed to eliminate the chance of induced or direct lightning hazard to the shack and operator. Some users simply completely disconnect the antenna from the antenna tuner and rig and shack during periods of possible lightning activity.

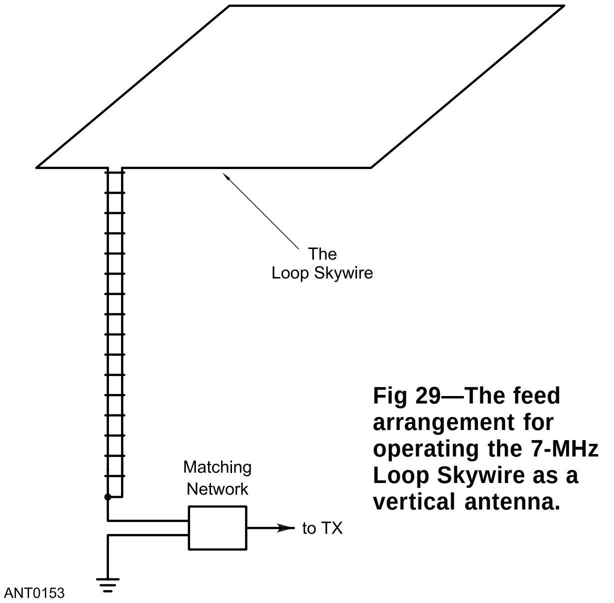

Fig 29—The feed arrangement for operating the 7-MHz Loop Skywire as a vertical antenna.

7-MHz Loop

An effective but simple 7MHz antenna that has a theoretical gain of approximately 1 dB over a dipole is a full-wave, closed vertical loop. Such a loop need not be square, as illustrated in **Fig 30A**. It can be trapezoidal, rectangular, circular, or some distorted configuration in between those shapes. For best results, however, you should attempt to make the loop as square as possible. The more rectangular the shape, the greater the cancellation of energy in the system, and the less effective it will be. In the limiting case, the antenna loses its identity as a loop and becomes a folded dipole.

You can feed the loop in the center of one of the vertical sides if you want vertical polarization. For horizontal polarization, you feed either of the horizontal sides at the center. Since optimum directivity occurs at right angles to the plane of the loop (or in more simple terms, broadside to the loop), you should hang the loop to radiate the maximum amount in some favored direction.

Fig 31A shows the azimuthal response at a takeoff angle of 15°, a typical angle for 40-meter DX, for vertical and horizontal feed systems over ground with "average" conductivity and dielectric constant. Fig 31A includes, for reference, the response of a flattop dipole 50 feet high. For DX work on 40 meters, the vertically polarized loop can perform as well as or substantially better than either a horizontally polarized loop or a flattop dipole, particularly in the azimuthal nulls of the dipole.

For the low elevation angles that favor DX work, the optimal feed point is at the center of one of the vertical wires. Feeding the loop at one of the corners at the bottom gives a

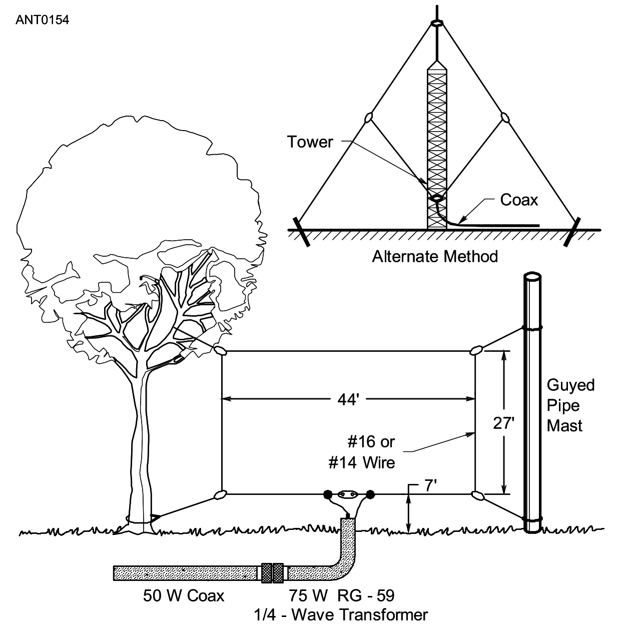

Fig 30—At A, details of the rectangular full-wave loop. The dimensions given are for operation at 7.05 MHz. The height above ground was 7 feet in this instance, although improved performance should result if the builder can install the loop higher above ground without sacrificing length on the vertical sides. At B, illustration how a single supporting structure can be used to hold the loop in a diamond-shaped configuration. Feeding the diamond at the lower tip provides radiation in the horizontal plane. Feeding the system at either side will result in vertical polarization of the radiated signal.

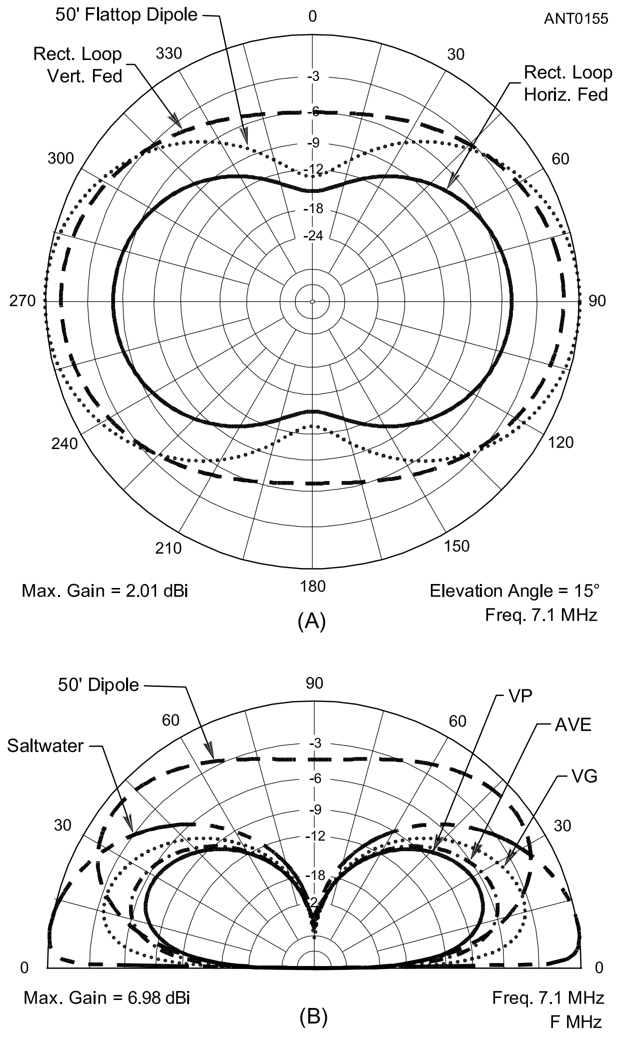

Fig 31—At A, azimuthal plane responses for the vertically and horizontally polarized 7-MHz loop, compared to a flattop 50-foot high dipole, all at a takeoff angle of 15° for DX work. The solid line is for feeding the loop horizontally at the bottom; the dashed line is for feeding the loop vertically at a side, and the dotted line is for a simple flattop horizontal dipole at 50 feet in height. For DX work, the vertically polarized loop is an excellent performer.

compromise result for both local and DX work. The actual impedance is roughly the same at each point: bottom horizontal center, corner or vertical side center.

Fig 31B demonstrates how the gain for vertical polarization changes over different type of grounds: saltwater, very poor ground (conductivity = 1 mS/m, dielectric constant = 5) very good (conductivity = 30 mS/m, dielectric constant = 20) and average ground (conductivity = 5 mS/m, dielectric constant = 13). Again, for reference a 50-foot high flattop dipole's elevation response is included. As has been mentioned previously in other chapters, a seaside location is a wonderful environment for verticals!

Just how you erect such a loop will depend on what is available in your backyard. Trees are always handy for supporting loop antennas. A disadvantage to the rectangular loop shown in Fig 30A is that two 34-foot high supports are needed, although in many instances your house may be high enough to serve as one of these supports. If you have a tower higher than about 50 feet, Fig 30B demonstrates how you can use it to support a diamond-shaped loop for 40 meters. The elevation and azimuthal responses are almost the same for either loop configuration, rectangular- or diamond-shaped.

The overall length of the wire used in a loop is determined in feet from the formula 1005/f (MHz). Hence, for operation at 7.125 MHz the overall wire length will be 141 feet. The matching transformer, an electrical $^{1}/_{4}$ λ of 75Ω coax cable, can be computed by dividing 246 by the operating frequency in MHz, then multiplying that number by the velocity factor of the cable being used. Thus, for operation at 7.125 MHz, 246/7.125 MHz = 34.53 feet. If coax with solid polyethylene insulation is used, a velocity factor of 0.66 must be employed. Foam-polyethylene coax has a velocity factor of 0.80. Assuming RG-59 is used, the length of the matching transformer becomes 34.53 (feet) × 0.66 = 22.79 feet, or 22 feet, 9$^{1}/_{2}$ inches.

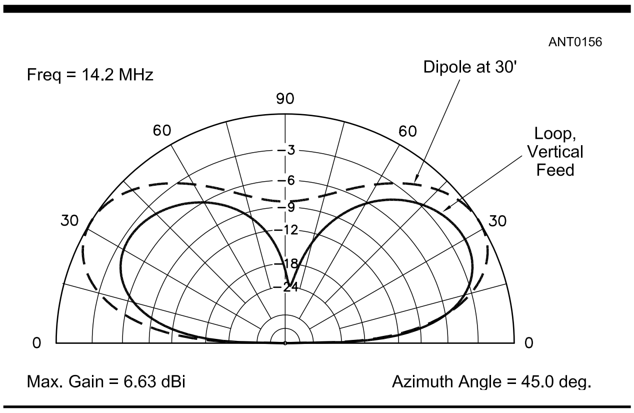

Fig 32—Elevation-plane response of 7-MHz loop used on 14.2 MHz. This is for a feed point at the center of one of the two vertical wires. The dashed line is the response of a flattop 20-meter dipole at 30 feet in height for comparison.

This same loop antenna in Fig 30A fed vertically may be used on the 14 and 21MHz bands, although its pattern will not be as good as that on its fundamental frequency and you will have to use an open-wire transmission line to feed the loop for multiband use. **Fig 32** shows the response at the peak lobe of the loop, at a 45° angle to the plane of the loop, compared to the peak response for a simple halfwave 20-meter dipole, 30 feet high. The gain from a simple flattop dipole, mounted at 30 feet, will be superior to the loop operated on a harmonic frequency.

A Receiving Loop for 1.8 MHz

You can use a small shielded-loop antenna to improve reception under certain conditions, especially at the lower amateur frequencies. This is particularly true when high levels of man-made noise are prevalent, when the second-harmonic energy from a nearby broadcast station falls in the 1.8MHz band, or when interference exists from some other amateur station in the immediate area. A properly constructed and tuned small loop will exhibit approximately 30 dB of front-to-side response, the minimum response being at right angles to the plane of the loop. Therefore, noise and interference can be reduced significantly or completely nulled out, by rotating the loop so that it is sideways to the interference-causing source.

Generally speaking, small shielded loops are far less responsive to man-made noise than are the larger antennas used for transmitting and receiving. But a trade-off in performance must be accepted when using the loop, for the strength of received signals will be 10 or 15 dB less than when using a full-size resonant antenna. This condition is not a handicap on 1.8 or 3.5 MHz, provided the station receiver has normal sensitivity and overall gain. Because a front-to-side ratio of 30 dB may be expected, a shielded loop can be used to eliminate a variety of receiving problems if made rotatable, as shown in **Fig 33**.

To obtain the sharp bidirectional pattern of a small loop, the overall length of the conductor must not exceed 0.1 λ. The loop of **Fig 34** has a conductor length of 20 feet. At 1.81 MHz, 20 feet is 0.037 λ. With this style of loop, 0.037 λ is about the maximum practical dimension if you want to tune the element to resonance. This limitation results from the distributed capacitance between the shield and inner conductor of the loop. RG-59 was used for the loop element in this example. The capacitance per foot for this cable is 21 pF, resulting in a total distributed

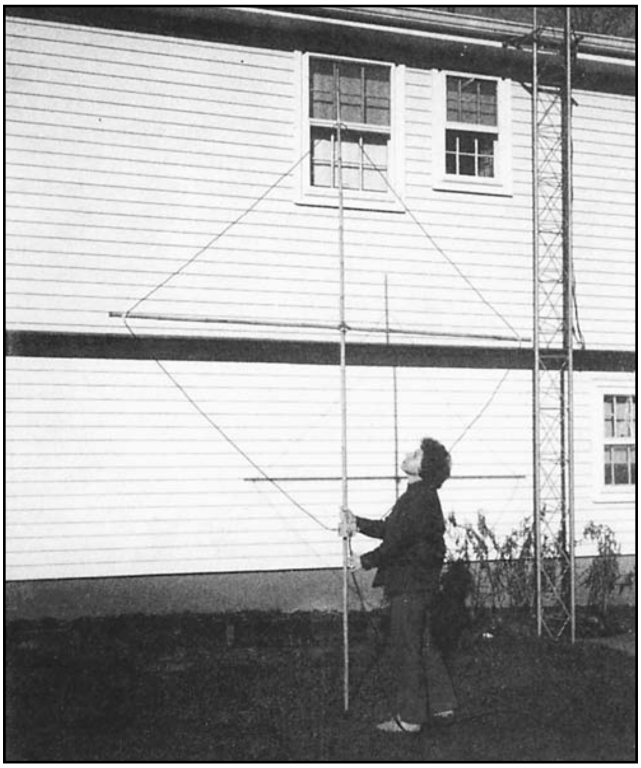

Fig 33—Jean DeMaw, W1CKK, tests the 1.8-MHz shielded loop. Bamboo cross arms are used to support the antenna.

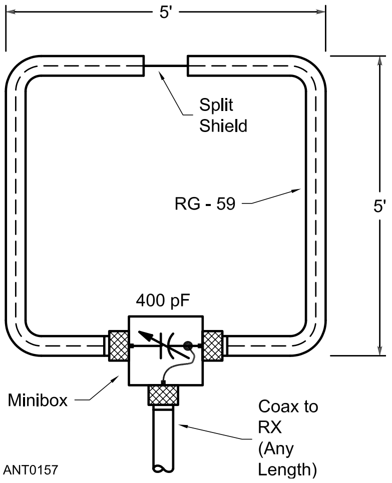

Fig 34—Schematic diagram of the loop antenna. The dimensions are not critical provided overall length of the loop element does not exceed approximately 0.1 λ. Small loops which are one half or less the size of this one will prove useful where limited space is a consideration.

capacitance of 420 pF. An additional 100 pF was needed to resonate the loop at 1.810 MHz.

Therefore, the approximate inductance of the loop is 15 µH. The effect of the capacitance becomes less pronounced at the higher end of the HF spectrum, provided the same percentage of a wavelength is used in computing the conductor length. The ratio between the distributed capacitance and the lumped capacitance used at the feed point becomes greater at resonance. These facts should be contemplated when scaling the loop to those bands above 1.8 MHz.

There will not be a major difference in the construction requirements of the loop if coaxial cables other than RG-59 are used. The line impedance is not significant with respect to the loop element. Various types of coaxial line exhibit different amounts of capacitance per foot, however, thereby requiring more or less capacitance across the feed point to establish resonance.

Shielded loops are not affected noticeably by nearby objects, and therefore they can be installed indoors or out after being tuned to resonance. Moving them from one place to another does not significantly affect the tuning.

You can see in the model shown in Fig 33 that a supporting structure was fashioned from bamboo poles. The X frame is held together at the center with two U bolts. The loop element is taped to the cross-arms to form a square. You could likely use metal cross arms without seriously degrading the antenna performance. Alternatively, wood can be used for the supporting frame.

A Minibox was used at the feed point of the loop to hold the resonating variable capacitor. In this model a 50 to 400-pF compression trimmer was used to establish resonance. You must weatherproof the box for outdoor installations.

Remove the shield braid of the loop coax for one inch directly opposite the feed point. You should treat the exposed areas with a sealing compound once this is done.

In operation this receiving loop has proven very effective for nulling out second-harmonic energy from local broadcast stations. During DX and contest operations on 1.8 MHz it helped prevent receiver overloading from nearby 1.8-MHz stations that share the band. The marked reduction in response to noise has made the loop a valuable station accessory when receiving weak signals. It is not used all of the time, but is available when needed by connecting it to the receiver through an antenna selector switch. Reception of European stations with the loop has been possible from New England at times when other antennas were totally ineffective because of noise.

It was also discovered that the effects of approaching storms (with attendant atmospheric noise) could be nullified considerably by rotating the loop away from the storm front. It should be said that the loop does not exhibit meaningful directivity when receiving sky-wave signals. The directivity characteristics relate primarily to ground-wave signals. This is a bonus feature in disguise, for when nulling out local noise or interference, one is still able to copy sky-wave signals from all compass points!

For receiving applications it is not necessary to match the feed line to the loop, though doing so may enhance the performance somewhat. If no attempt is made to obtain an SWR of 1, the builder can use 50 or 75-Ω coax for a feeder, and no difference in performance will be observed. The Q of this loop is sufficiently low to allow the operator to peak it for resonance at 1.9 MHz and use it across the entire 1.8MHz band. The degradation in performance at 1.8 and 2 MHz will be so slight that it will be difficult to discern.

An Indoor Stealth Loop

Ted Phelps, W8TP, wrote an article in *The ARRL Antenna Compendium, Vol 7* describing his attic-mounted wire loop antenna, fed with an automatic antenna tuner. Here is a shortened version of that article.

If you drive down my street in Whitechapel Village in Newark, DE, trying to find my ham location by looking for my antenna, you wouldn't find it. Even if you pulled up in front of my condo, you wouldn't notice any telltale signs, because my multiband antenna is completely hidden. It's in the attic of my two-bedroom condominium in a small retirement community completed in 1999.

Before the move, I had given considerable thought to what type of antenna I might use, if any. I already knew that permanent outdoor types were out of the question, due to restrictive real estate and condo association rules. So I planned a clause for any sales contract I might sign, specifically mentioning amateur radio and my desire to set up a station in my new living quarters. I decided I would not move where my lifelong hobby would be severely restricted or prohibited.

That meant that to be reasonably sure I could continue enjoying Amateur Radio as before, I would have to install an indoor antenna that could perform as well as a typical outdoor system. What kind? In Ohio, I had tried a horizontally polarized attic dipole made with #14 wire. It didn't work very well—it was just too low to the ground.

I learned about a high-tech method of remote antenna tuning using an antenna coupler which contains a microprocessor. I found this kind of automatic tuner available from two American manufacturers and within a reasonable price range. Although our move was still a few months ahead, I purchased a model SG-230 antenna coupler made by SGC Inc, Bellevue, WA, for use in Delaware.

Fig 35 shows the final dimensions of my hidden loop, which is a single-turn rectangular loop, erected in a north-south vertical plane and made from nearly 78 feet of #6 stranded, aircraft primary wire in a PVC jacket, held taut at the lower corners and supported by a pulley and guy rope at each upper corner. Because it's vertically polarized, it supports low-angle radiation reasonably well. By the way, if you're wondering why I used such a relatively large-gauge wire as #6 for the loop antenna, it was readily available from my son-in-law!

Fig 36 shows my completed condo unit. Note the doghouse dormer on the roof about 12 feet above ground at the attic floor. This is the level of my hidden loop's base leg.

In constructing my system I had to overcome RFI problems on my own premises. Each condo unit has its own electronic security panel on an upper shelf in a closet. As soon as I applied moderate power to my radio and loop, the Fire Alarm sounded and firefighters came to my door! The burglar/intrusion signal was triggered a couple of times, too. Working with a security installation technician, I found that *there was no ground wire connected to my security panel.* "We don't bother with that," said the tech, and then, reacting to my surprise, connected a #14 ground between the security panel and the house water-pipe ground. I then installed a ferrite bead on each lead entering the security panel. I also placed ferrite beads on keyer-paddle leads, GFCI electrical outlets, etc. Those

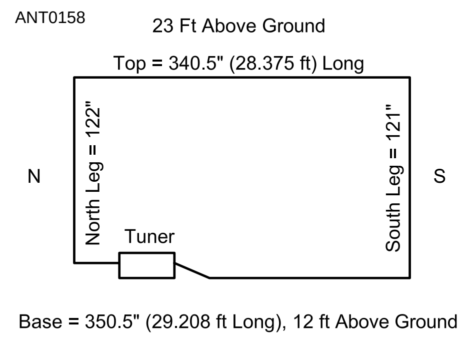
ANT0158
23 Ft Above Ground
Top = 340.5" (28.375 ft) Long
North Leg = 122"
South Leg = 121"
N
S
Tuner
Base = 350.5" (29.208 ft Long), 12 ft Above Ground

Fig 35—Diagram showing layout of W8TP's indoor hidden loop antenna.

Fig 36—Can you see W8TP's antenna in this photograph? Of course you can't—it's hidden from view inside his attic!

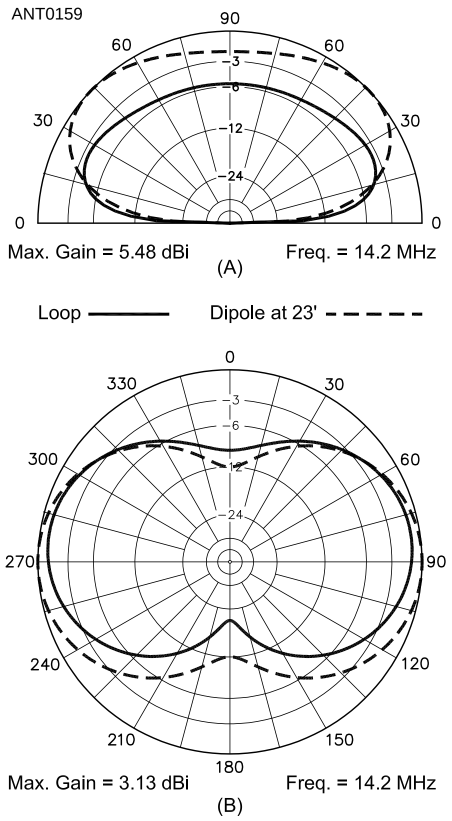

Fig 37—At A, computed elevation pattern at 14.2 MHz for W8TP's hidden loop (solid line), compared to a 20-meter dipole (dashed line) at a height of 23 feet. At B, a comparison of the azimuth patterns at a 20°-elevation angle for W8TP's loop (solid line) and the same 20-meter dipole (dashed line). The loop has a slightly asymmetrical response because it is fed at a corner, but its performance is competitive to an outdoor dipole. In fact, it has superior low-angle performance, typical of a vertically polarized antenna compared to a low horizontal antenna.

measures seem to have eliminated my RFI problems.

When we moved into our condo, I took the obvious precaution of not using a linear amplifier. I took extra care to establish a single-point ground for my station equipment by connecting all equipment grounds to the cover plate of the dedicated metallic outlet box behind the operating position, and thence to a separate ground rod in our front yard. I use a 1-kW RL Drake low-pass filter in the transceiver-antenna feed line.

Is this indoor antenna system safe? I believe so. In the attic it is not at all close to our living space. It is fixed firmly in place and unlike most amateur antennas it is *out of the weather*! I therefore do not use a quick-grounding system for times when a thunderstorm approaches.

Fig 37 shows the computed elevation and azimuth patterns on 20 meters. The tuner is able to hold the SWR down low enough so that my JRC-245 transceiver can operate through its internal antenna tuner.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

C. F. W. Anderson, "A Crossed-Loop/Goniometer DF Antenna for 160 Meters," *The ARRL Antenna Compendium, Vol 1* (Newington: ARRL, 1985), pp 127-132.

J. S. Belrose, "Ferromagnetic Loop Aerials," *Wireless Engineer*, Feb 1955, pp 41-46.

J. S. Belrose, "An Update on Compact Transmitting Loops," *QST*, Nov 1993, pp 37-40.

D. S. Bond, *Radio Direction Finders*, 1st ed. (New York: McGraw-Hill Book Co, 1944).

G. Bramslev, "Loop Aerial Reception," *Wireless World*, Nov 1952, pp 469-472.

R. W. Burhans, "Experimental Loop Antennas for 60 kHz to 200 kHz," *Technical Memorandum (NASA) 71*, (Athens, OH: Ohio Univ, Dept of Electrical Engr), Dec 1979.

R. W. Burhans, "Loop Antennas for VLF-LF," *Radio-Electronics*, Jun 1983, pp 83-87.

D. DeMaw, "Beat the Noise with a Scoop Loop," *QST*, Jul 1977, pp 30-34.

M. F. DeMaw, *Ferromagnetic-Core Design and Application Handbook* (Englewood Cliffs, NJ: Prentice-Hall Inc, 1981).

D. DeMaw and L. Aurick, "The Full-Wave Delta Loop at Low Height," *QST*, Oct 1984, pp 24-26.

R. Devore and P. Bohley, "The Electrically Small Magnetically Loaded Multiturn Loop Antenna," *IEEE Trans on Ant and Prop*, Jul 1977, pp 496-505.

T. Dorbuck, "Radio Direction-Finding Techniques," *QST*, Aug 1975, pp 30-36.

R. G. Fenwick, "A Loop Array for 160 Meters," *CQ*, Apr 1986, pp 25-29.

D. Fischer, "The Loop Skywire," *QST*, Nov 1985, pp 20-22. Also "Feedback," *QST*, Dec 1985, p 53.

R. A. Genaille, "V.L.F. Loop Antenna," *Electronics World*, Jan 1963, pp 49-52.

G. Gercke, "Radio Direction/Range Finder," *73*, Dec 1971, pp 29-30.

R. S. Glasgow, *Principles of Radio Engineering* (New York: McGraw-Hill Book Co, Inc, 1936).

S. Goldman, "A Shielded Loop for Low Noise Broadcast Reception," *Electronics*, Oct 1938, pp 20-22.

F. W. Grover, *Inductance Calculation-Working Formulas and Tables* (New York: D. VanNostrand Co, Inc, 1946).

J. V. Hagan, "A Large Aperture Ferrite Core Loop Antenna for Long and Medium Wave Reception," *Loop Antennas Design and Theory*, M.G. Knitter, Ed. (Cambridge, WI: National Radio Club, 1983), pp 37-49.

T. Hart, "Small, High-Efficiency Loop Antennas," *QST*, Jun 1986, pp 33-36.

T. Hart, *Small High Efficiency Antennas Alias The Loop* (Melbourne, FL: W5QJR Antenna Products, 1985)

H. Hawkins, "A Low Budget, Rotatable 17 Meter Loop," *QST*, Nov 1997, p 35.

F. M. Howes and F. M. Wood, "Note on the Bearing Error and Sensitivity of a Loop Antenna in an Abnormally Polarized Field," *Proc IRE*, Apr 1944, pp 231-233.

B. Jones, KD7S, "You Can Build: A Compact Loop Antenna for 30 Through 12 Meters," Nov 1994 *QST*, pp 33-36.

J. A. Lambert, "A Directional Active Loop Receiving Antenna System," *Radio Communication*, Nov 1982, pp 944-949.

D. Lankford, "Loop Antennas, Theory and Practice," *Loop Antennas Design and Theory*, M. G. Knitter, Ed. (Cambridge, WI: National Radio Club, 1983), pp 10-22.

D. Lankford, "Multi-Rod Ferrite Loop Antennas," *Loop Antennas Design and Theory*, M. G. Knitter, Ed. (Cambridge, WI: National Radio Club, 1983), pp 53-56.

G. Levy, "Loop Antennas for Aircraft," *Proc IRE*, Feb 1943, pp 56-66. Also see correction, *Proc IRE*, Jul 1943, p 384.

J. Malone, "Can a 7 foot 40m Antenna Work?" *73*, Mar 1975, pp 33-38.

R. Q. Marris, "An In-Room 80-Meter Transmitting Multiturn Loop Antenna," *QST*, Feb 1996, pp 43-45. Feedback. *QST*, May 1996, p 48.

L. G. McCoy, "The Army Loop in Ham Communications," *QST*, Mar 1968, pp 17, 18, 150, 152. (See also Technical Correspondence, *QST*, May 1968, pp 49-51 and Nov 1968, pp 46-47.)

R. G. Medhurst, "HF Resistance and Self Capacitance of Single Layer Solenoids," *Wireless Engineer*, Feb 1947, pp 35-43, and Mar 1947, pp 80-92.

G. P. Nelson, "The NRC FET Altazimuth Loop Antenna," *N.R.C. Antenna Reference Manual Vol. 1*, 4th ed., R. J. Edmunds, Ed. (Cambridge, WI: National Radio Club, 1982), pp 2-18.

K. H. Patterson, "Down-To-Earth Army Antenna," *Electronics*, Aug 21, 1967, pp 111-114.

R. C. Pettengill, H. T. Garland and J. D. Meindl, "Receiving Antenna Design for Miniature Receivers," *IEEE Trans on Ant and Prop*, Jul 1977, pp 528-530.

T. Phelps, W8TP, "A Hidden Loop Antenna," *The ARRL Antenna Compendium, Vol 7* (Newington: ARRL, 2002), pp 160-162.

R. Z. Plasencia, W0RPV, "Remotely Controlled Antenna Coupler," *The ARRL Antenna Compendium, Vol 2* (Newington: ARRL, 1989), pp 182-186.

W. J. Polydoroff, *High Frequency Magnetic Materials—Their Characteristics and Principal Applications* (New York: John Wiley and Sons, Inc, 1960).

E. Robberson, "QRM? Get Looped," *Radio and Television News*, Aug 1955, pp 52-54, 126.

G. S. Smith, "Radiation Efficiency of Electrically Small Multiturn Loop Antennas," *IEEE Trans on Ant and Prop*, Sep 1972, pp 656-657.

E. C. Snelling, *Soft Ferrites—Properties and Applications* (Cleveland, OH: CRC Press, 1969).

G. Thomas, "The Hot Rod—An Inexpensive Ferrite Booster Antenna," *Loop Antennas Theory and Design*, M. G. Knitter, Ed. (Cambridge, WI: National Radio Club, 1983), pp 57-62.

J. R. True, "Low-Frequency Loop Antennas," *Ham Radio*, Dec 1976, pp 18-24.

E. G. VonWald, "Small-Loop Antennas," *Ham Radio*, May 1972, pp 36-41.

"An FET Loop Amplifier with Coaxial Output," *N.R.C. Antenna Reference Manual, Vol 2*, 1st ed., R. J. Edmunds, Ed. (Cambridge, WI: National Radio Club, Oct 1982), pp 17-20.

Aperiodic Loop Antenna Arrays (Hermes Electronics Ltd, Nov 1973).

Reference Data for Radio Engineers, 6th ed. (Indianapolis: Howard W. Sams & Co, subsidiary of ITT, 1977).

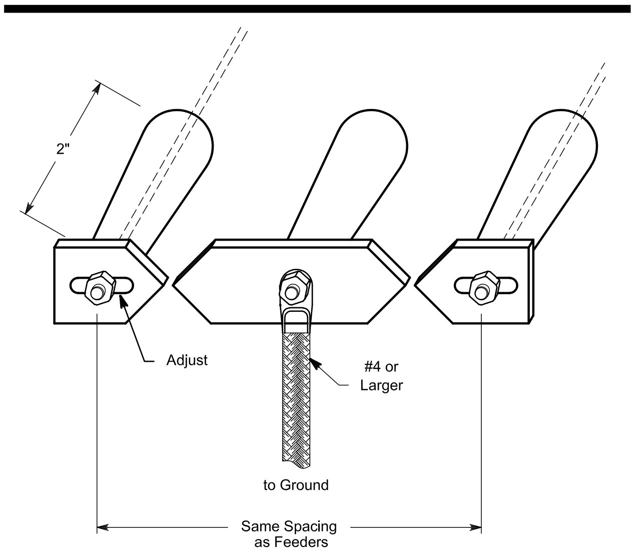

Fig 43—A simple lightning arrester for open-wire line made from three standoff or feedthrough insulators and sections of $^{1}/_{8}$ × $^{1}/_{2}$-inch brass or copper strap. It should be installed in the line at the point where the line enters the station. The heavy ground lead should be as short and as direct as possible. The gap setting should be adjusted to the minimum width that will prohibit arcing when the transmitter is operated.

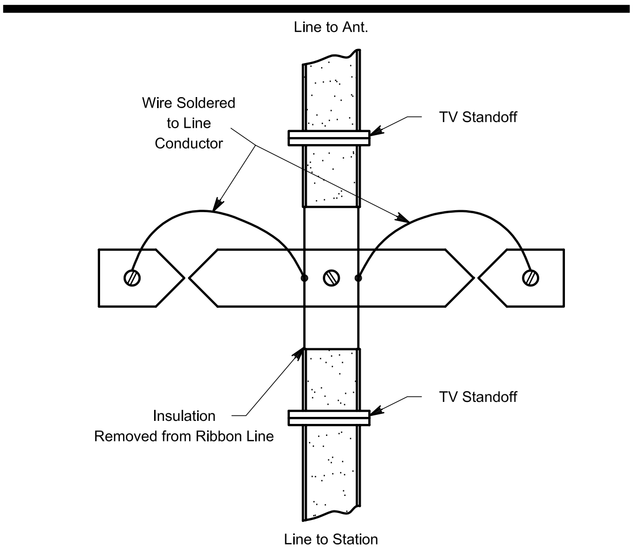
Line to Ant.
Wire Soldered to Line Conductor
TV Standoff
TV Standoff
Insulation Removed from Ribbon Line
Line to Station

Fig 44—The lightning arrester of Fig 39 may be used with 300-Ω ribbon line in the manner shown here. The TV standoffs support the line an inch or so away from the grounded center member of the arrester

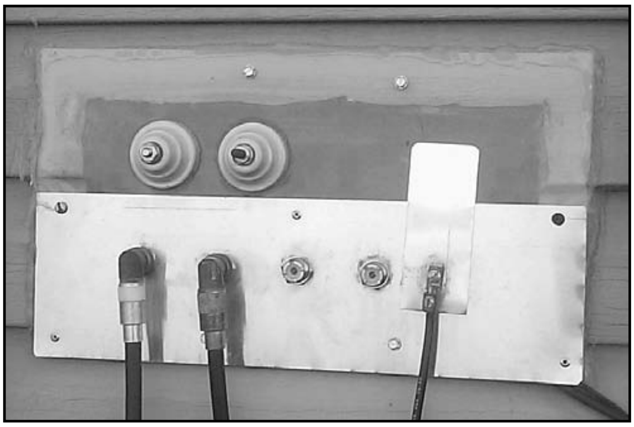

Fig 45—K8CH's coax entry panel mounted on exterior wall (later covered by grill that provides combustion to an unused fireplace). The ground braid goes to a ground rod located beneath the panel. (Photo courtesy: *Simple and Fun Antennas for Hams*)

The construction of a homemade arrester for open-wire line is shown in **Fig 43**. This type of arrester can be adapted to ribbon line an inch or so away from the center member of the arrester, as shown in **Fig 44**. Sufficient insulation should be removed from the line where it crosses the arrester to permit soldering the arrester connecting leads.

Lightning Grounds

Lightning-ground connecting leads should be of conductor size equivalent to at least #10 wire. The #8 aluminum wire used for TV-antenna grounds is satisfactory. Copper braid $^{3}/_{4}$-inch wide (Belden 8662-10) is also suitable. The conductor should run in a straight line to the grounding point. The ground connection may be made to a water pipe system (if the pipe is not plastic), the grounded metal frame of a building, or to one or more $^{5}/_{8}$-inch ground rods driven to a depth of at least 8 feet. More detailed information on lightning protection is contained in Chapter 1, Safety.

A central grounding panel for coax cables coming into the house is highly recommended. See **Fig 45** for a photo of the homemade grounding panel installed by Chuck Hutchinson, K8CH, at his Michigan home. The coax cables screwed into dual-female feed-through UHF connectors. K8CH installed this aluminum panel under the outside grill for a duct that provided combustion air to an unused fireplace. He used ground strap to connect to ground rods located under the panel. See the ARRL book *Simple and Fun Antennas* for more information about ground panels.

Before a lightning storm approaches, a prudent ham will disconnect all feed lines, rotor lines and control lines inside the shack to prevent damage to sensitive electronics. When lightning is crashing about outside, you certainly don't want that lightning inside your shack!

Chapter 6

Low-Frequency Antennas

In theory there is no difference between antennas at 10 MHz and up and those for lower frequencies. In reality however, there are often important differences. It is the size of the antennas, which increases as frequency is decreased, that creates practical limits on what can be realized physically at reasonable cost.

At 7.3 MHz, 1λ = 133 feet and by the time we get to 1.8 MHz, 1 λ = 547 feet. Even a λ/2 dipole is very long on 160 meters. The result is that the average antenna for these bands is quite different from the higher bands, where Yagis and other relatively complex antennas dominate. In addition, vertical antennas can be more useful at low frequencies than they are on 20 meters and above because of the low heights (in wavelengths) usually available for horizontal antennas on the low bands. Much of the effort on the low bands is focused on how to build simple but effective antennas with limited resources. This section is devoted to antennas for use on amateur bands between 1.8 to 7 MHz.

The Importance of Low Angles for Low-Band DXing

In Chapter 3, The Effects of Ground, we emphasized the importance of matching the elevation response of your antennas as closely as possible to the range of elevation angles needed for communication with desired geographic areas. **Fig 1** shows the statistical 40-meter elevation angles needed over the entire 11-year solar cycle to cover the path from Boston, Massachusetts, to all of Europe. These angles range from 1° (at 9.6% of the time when the 40-meter band is open to Europe) to 28° (at 0.3% of the time).

Fig 1 also overlays the elevation pattern response of a 100-foot high flattop dipole on the elevation-angle statistics, illustrating that even at this height the coverage is hardly optimum to cover all the necessary elevation angles. While Fig 1 is dramatic in its own right, the data can be viewed in another way that emphasizes even more the importance of low elevation angles. **Fig 2** plots the the *cumulative distribution function*, the total percentage of time 40 meters is open from Boston to Europe, at or below each elevation angle. For example, Fig 2 says that 40 meters is open to Europe from Boston 50% of the time at an elevation angle of 9° or less. The band is open 90%

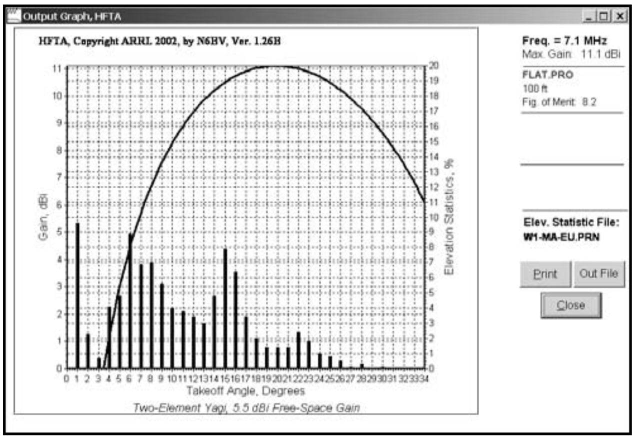

HFTA, Copyright ARRL 2002, by N6BV, Ver. 1.26B

Two-Element Yagi, 5.5 dBi Free-Space Gain

Freq. = 7.1 MHz
Max. Gain: 11.1 dBi

FLAT.PRO
100 ft
Fig. of Merit: 8.2

Elev. Statistic File:
W1-MA-EU.PRN

Fig 1—Screen capture from *HFTA* (HF Terrain Assessment) program showing elevation response for 100-foot high dipole over flat ground on 7.1 MHz, with bar-graph overlay of the statistical elevation angles needed over the whole 11-year solar cycle from New England (Boston) to all of Europe. Even a 100-foot high antenna cannot cover all the necessary angles.

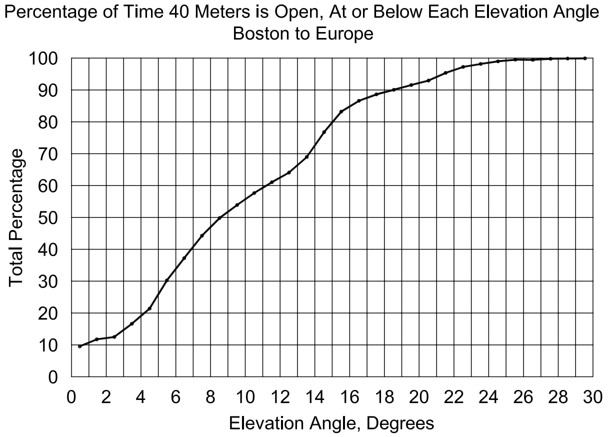

Fig 2—Another way of looking at the elevation statistics from Fig 1. This shows the percentage of time the 40-meter band is open, at or below each elevation angle, on the path from Boston to Europe. For example, the band is open 50% of the time at an angle of 9° or lower. It is open 90% of the time at an angle of 19° or lower.

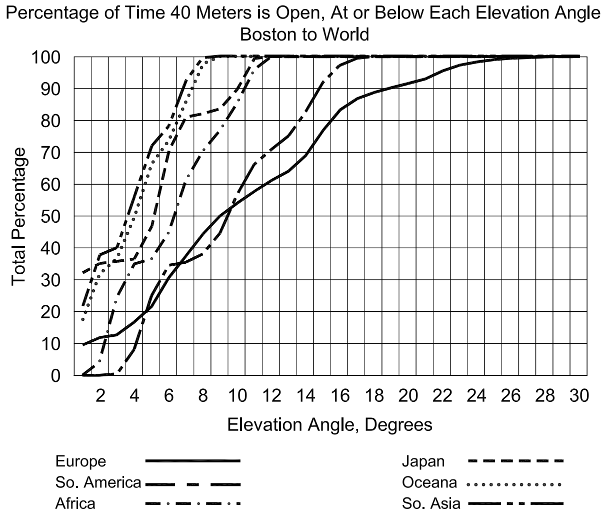

Fig 3—The percentage of time the 40-meter band is open, at or below each elevation angle, for various DX paths from Boston: to Europe, South America, southern Africa, Japan, Oceania and south Asias. The angles are predominantly quite low. For example, on the path from Boston to Japan, 90% of the time when the 40-meter band is open, it is open at elevation angles less than or equal to 10°. Achieving good performance at these low takeoff angles requires very high horizontally polarized antennas, or efficient vertically polarized antennas.

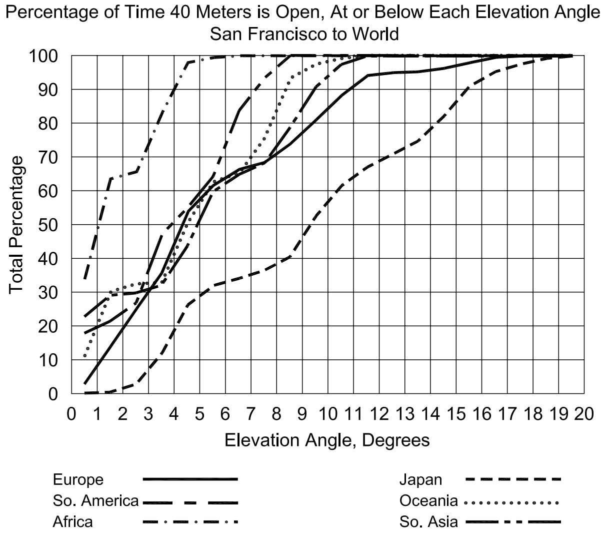

Fig 4—The 40-meter statistics from the West Coast: from San Francisco to the rest of the DX world. Here, 90% of the time the path to Europe is open, it is at takeoff angles less than or equal to 11°. No wonder the hams living on mountain tops do best into Europe from the West Coast.

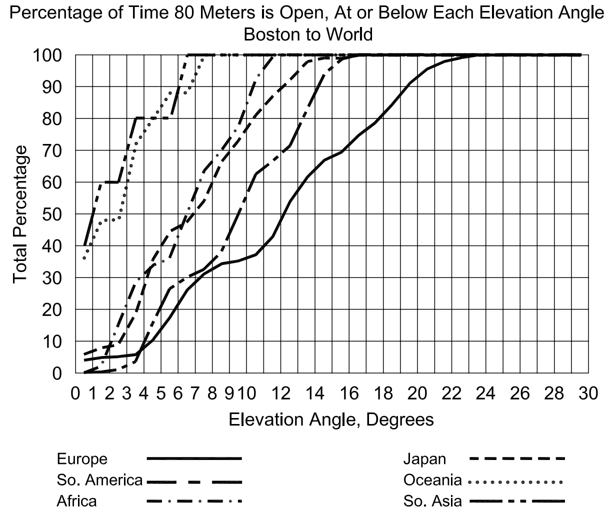

Fig 5—The situation on 80 meters from Boston to the rest of the DX world. Into Europe, 90% of the time the elevation angle is less than or equal to 20°. Into Japan from Boston, 90% of the time the angle is less than or equal to 12°.

of the time at an elevation angle of 19° or less.

Fig 3 plots the 40-meter elevation-angle data for six major geographic areas around the world from Boston. In general, the overall range of elevation angles for far-distant locations is smaller, and the angles are lower than for closer-in areas. For example, from Boston to southern Asia (India), 50% of the time the takeoff angles are 4° or less. On the path to Japan from Boston, the takeoff angles is less than or equal to 6° about 70% of the time. These are low angles indeed.

Fig 4 shows similar data for the 40-meter band from San Francisco, California, to the rest of the world. The path to southern Africa from the US West Coast is a very long-distance path, open some 65% of the time it is open at angles of 2° or less! The 40-meter path to Japan involves takeoff angles of 10° or less more than 50% of the time. If you are

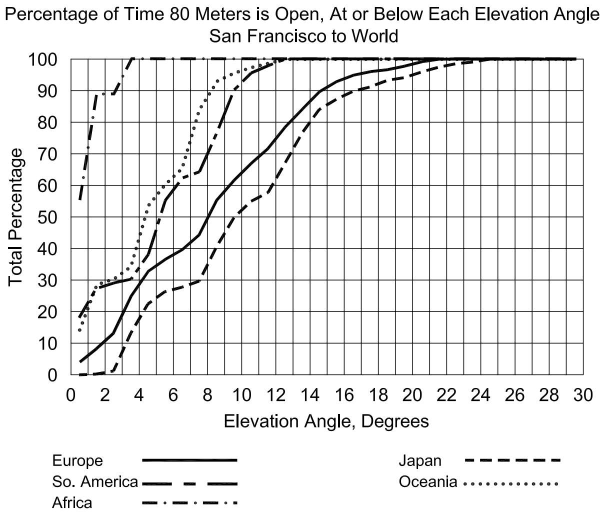

Fig 6—From San Francisco to the rest of the world on 80 meters: 90% of the time on the path to Japan, the takeoff angle is less than or equal to 17°; 50% of the time the angle is less than or equal to 10°; 25% of the time the angle is less than or equal to 6°. A horizontally polarized antenna would have to be 600 feet above flat ground to be optimum at 6°!

fortunate enough to have a 100-foot high flattop dipole for 40 meters, at a takeoff angle of 10° the response would be down about 3 dB from its peak level at 20°. At an elevation angle of 5° the response would be about 8 dB down from peak. You can see why the California stations located on mountain tops do best on 40 meters for DXing.

Fig 5 shows the same percentage-of-time data for the 80-meter band from Boston to the world. Into Europe from Boston, the 80-meter elevation angle is 13° or less more than 50% of the time. Into Japan from Boston, 90% of the time the band is open is at a takeoff angle of 13° or less. (Note that these elevation statistics are computed for "undisturbed" ionospheric conditions. There are times when the incoming angles are affected by geomagnetic storms, and generally speaking the elevation angles rise under these conditions.)

Fig 6 shows the 80-meter data from San Francisco to the world. Low elevation angles dominate in this graph and high horizontal antennas would be necessary to optimal coverage. In fact, 50% of the time for all paths, the elevation angle is less than 10°.

In the rest of this chapter, we'll often compare horizontally polarized antennas at practical heights with vertically polarized antennas, usually at takeoff angles of 5° or 10°, angles useful for DX work. But first, let us look at situations where *high* takeoff angles are most useful.

Short/Medium-Range Communications

Not all hams are interested in working stations thousands of miles from them. Traffic handlers and rag chewers may, in fact, only be interested in *nearby* communications—perhaps out to 600 miles from their location.

For example, a ham in Boston may want to talk with his brother-in-law in Cleveland, OH, a path that is just over 550 miles away. Or an operator in Buffalo, NY, may be the net control station (NCS) for a regional net involving the states of New York and New Jersey. She needs to cover distances up to about 300 miles away.

Depending on the time of day, the most appropriate ham frequencies needed for nearby communications are the 40 and 80/75-meter bands, with 160 meters also a possibility during the night hours, particularly during low portions of the sunspot cycle. The elevation angles involved in such nearby distances are usually high, even almost directly overhead for distances beyond ground-wave coverage (which may be as short as a few miles on 40 meters). For example, the distance between the Massachusetts cities of Boston and Worchester is about 40 miles. On 40 meters, 40 miles is beyond ground-wave coverage. So you will need sky-wave signals that use the ionosphere to communicate between these two cities, where the elevation angle is 83°—very nearly straight up.

Hams using vertical antennas for communications with nearby stations may well find that their signals will be below the noise level typical on the lower bands, especially if they aren't running maximum legal power. Such relatively short-range paths involve so-called *NVIS*, "Near Vertical Incidence Skywave," a fancy name for HF communication systems covering nearby geographic areas. The US military discusses NVIS out to about 500 miles, encompassing the territory a brigade might cover. Elevation angles needed to cover distances from 0 to 500 miles range from about 40º to 90º. This also covers the circumstances involved in amateur communications, particularly in emergency situations.

The following section is adopted from the article "What's the Deal About NVIS?" that appeared in December 2005 *QST*. This article used an example of a hypothetical earthquake in San Francisco to analyze HF emergency communication requirements.

HAM RADIO RESPONSE IN NATURAL DISASTERS

One of San Francisco's somewhat less endearing nicknames is "the city that waits to die." When the *Big Earthquake* does come, you can be assured that all the cell phones and the land-line telephones will be totally jammed, making calling in or out of the San Francisco Bay Area virtually impossible. The same thing occurred in Manhattan on September 11, 2001. The Internet will also be severely affected throughout northern California because of its trunking via the facilities of the telephone network. Commercial electricity will be out in wide areas because power lines will be down. It's virtually certain that water mains will be out of commission too.

If the repeaters on the hills around the San Francisco Bay Area haven't been damaged by the shaking itself, there will be some ham VHF/UHF voice coverage in the intermediate area, at least until the backup batteries run down. But connecting to the dysfunctional telephone system will be difficult at best through amateur repeaters.

With little or no telephone coverage, an obvious need for ham radio communications to aid disaster relief would be from San Francisco to Sacramento, the state capital. Sacramento is 75 miles northeast of the Bay Area, well outside VHF/UHF coverage, so amateur HF will be required on this radio circuit. On-the-ground communications directly between emergency personnel (including the armed-forces personnel who will be brought into the rescue and rebuilding effort) will often be difficult on VHF/UHF since San Francisco is a hilly place. So HF will probably be needed even for short distance, operator-to-operator or operator-to-communications center work. Throughout the city, portable HF stations will have to be quickly set up and staffed to provide such communications.

Hams used to half jokingly call short range HF communications on 40 and 80 meters "cloud warming." This is an apt description, because the takeoff angles needed to launch HF signals up into the ionosphere and then down again to a nearby station are almost directly upwards. **Table 1** lists the distance and takeoff angles from San Francisco to various cities around the western part of the USA. The distance between San Francisco and Sacramento is about 75 miles, and the optimum takeoff angle is about 78°. Launching such a high-angle signal is best done using horizontally polarized antennas mounted relatively close to the ground.

GEOGRAPHIC COVERAGE FOR NVIS

Figure 7A shows the geographic area coverage around San Francisco for a 100-W, 7.2-MHz station using an inverted V dipole. The center of this antenna is 20 feet above flat

Table 1
Average Elevation Angles for Target Destinations from San Francisco

Location	*Distance Miles*	*Average Elevation Angle, Degrees*
San Jose, CA	43	80
Sacramento, CA	75	78
Fresno, CA	160	63
Reno, NV	185	60
Los Angeles	350	44
San Diego	450	42
Portland, OR	530	30
Denver, CO	950	18
Dallas, TX	1500	8

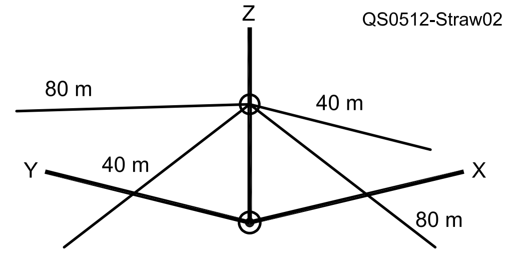

Fig 8—Layout for two band inverted V dipoles for 40 and 80 meters. The two dipoles are fed together at the center and are laid out at right angles to each other to minimize interaction between them. Each end of both dipoles is kept 8 feet above ground for personnel safety.

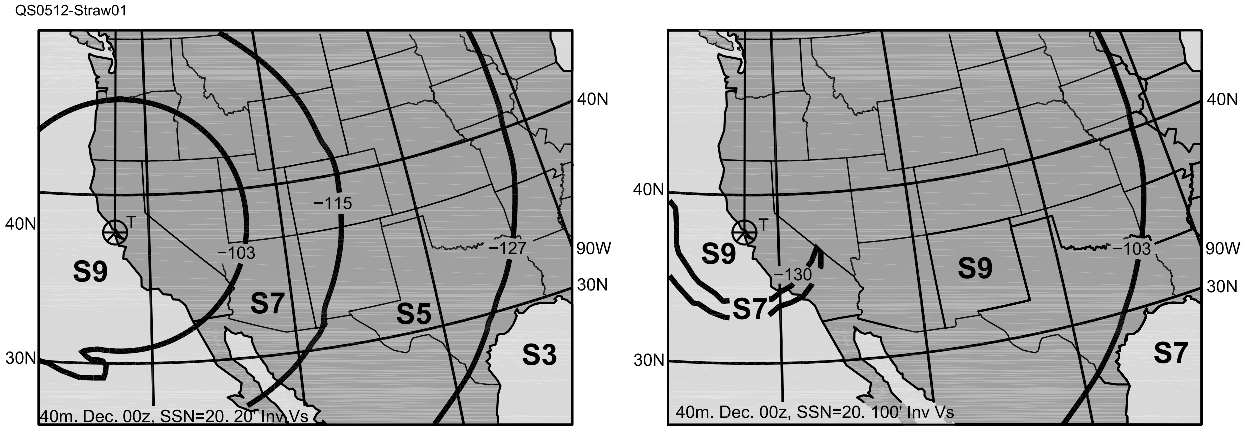

(A) (B)

Fig 7—At A, Predicted 40 meter geographic coverage plot for a 100 W transmitter in December at 0000 UTC (near sunset), for a SSN (Smoothed Sunspot Number) of 20. The antennas used are 20 foot-high inverted V dipoles. At B, 40 meter coverage for same date and time, but for 100 foot-high flattop dipoles. Most of California is well covered with S9 signals in both cases, but there is more susceptibility in the higher dipole case to thunderstorm crashes coming from outside California, for example from Arizona or even Texas. Such noise can interfere with communications inside California.

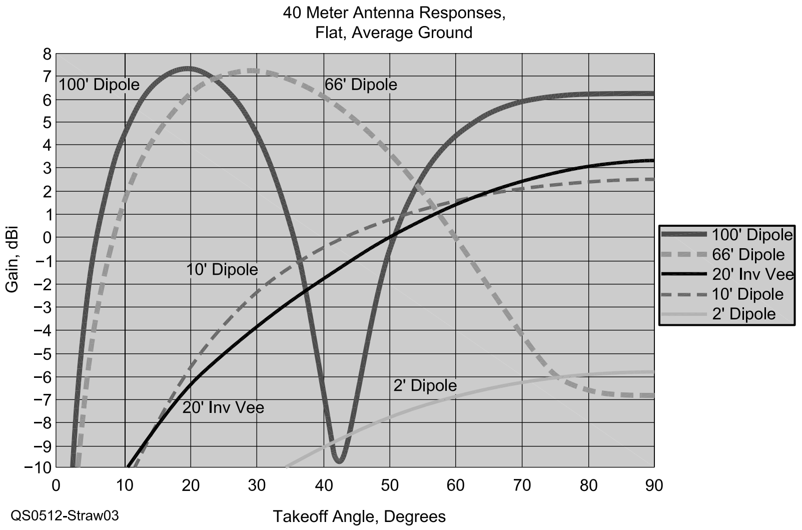

Fig 9—Elevation plots for different 40 meter antennas above flat ground with average ground characteristics (5 mS/m conductivity and dielectric constant of 13). The 10 foot-high flattop dipole and the 20 foot-high inverted V dipoles both have close to the same characteristics. Note that there is a null in the response of the 100 foot-high flattop dipole at a 42° elevation angle. The gain there is roughly that of a 2 foot-high dipole!

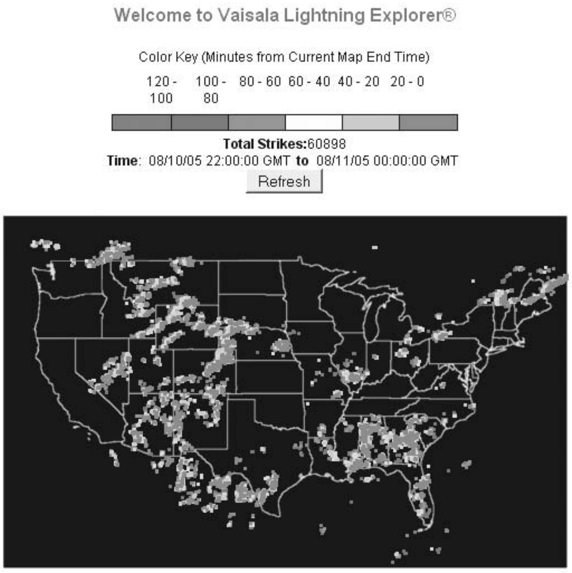
Welcome to Vaisala Lightning Explorer®

Color Key (Minutes from Current Map End Time)

120 - 100 | 100 - 80 | 80 - 60 | 60 - 40 | 40 - 20 | 20 - 0

Total Strikes: 60898

Time: 08/10/05 22:00:00 GMT **to** 08/11/05 00:00:00 GMT

Fig 10—The distribution of lightning strikes across the USA for August 10, 2005 from 2200 to 0000 UTC, in the afternoon California time. There are lots of lightning strikes in the US during the summer—60,898 of them in this two-hour period! (*Courtesy Vaisala Lightning Explorer.*)

ground and the ends are 8 feet high. An actual implementation of such an antenna could be as an 80-meter inverted V, fed in parallel with a 40-meter inverted V dipole at a 90° angle. See **Fig 8**. The 8-foot height puts the ends high enough to prevent RF burns to humans (or most animals). The low height of the antenna above ground means that the azimuthal pattern is omnidirectional for high elevation angles.

Fig 7 was generated using the *VOAAREA* program, part of the *VOACAP* propagation-prediction suite, for the month of December. This was for 0000 UTC, close to sundown, for a low period of solar activity (Smoothed Sunspot Number, SSN of 20). The receiving stations were also assumed to be using identical inverted-V dipoles.

You can see that almost the whole state of California is covered with S9 signals, minus only a thin slice of land near the Mexican border in the southeast portion of the state, where the signal drops to S7. Signals from Texas are predicted to be only S5 or less in strength. Signals (or thunderstorm static) coming from, say, Louisiana would be several S units weaker than signals from central Texas.

Now take a look at Fig 7B. Here, the date, time and solar conditions remain the same, but now the antennas are 100-foot high flattop dipoles. California is still blanketed with S9 signals, save for an interesting crescent-shaped slice near Los Angeles, where the signal drops down to S7. Close investigation of this intriguing drop in signal strength reveals that the necessary elevation angle, 44°, from San Francisco to this part of southern California falls in the first null of the 100-foot high antenna's elevation pattern. See **Fig 9**, which shows the elevation patterns for five 40-meter antennas at different heights. In the null at a 44° takeoff angle, the 100-foot high dipole is just about equal to a 2-foot high dipole. We'll discuss 2-foot high dipoles in more detail later.

For most of California, the problem with 100-foot high 40-meter antennas is that interfering signals from Texas, Colorado or Washington State will *also* be S9 in San Francisco. So will static crashes coming from thunderstorms all over the West and much of the Gulf Coast. (Ed Farmer, AA6ZM, joked once that the Army doesn't have any problem with interfering signals—they just call in an airstrike. We hams don't generally have this ability, although we occasionally call in the FCC.) See **Fig 10**, which shows a typical distribution of thunderstorms across the US in the late afternoon, California time, in mid-August. There certainly are a lot of thunderstorms raging around the country in the summer.

The signal-to-noise and signal-to-interference ratios for a 20-foot high inverted V dipole will be superior for medium-

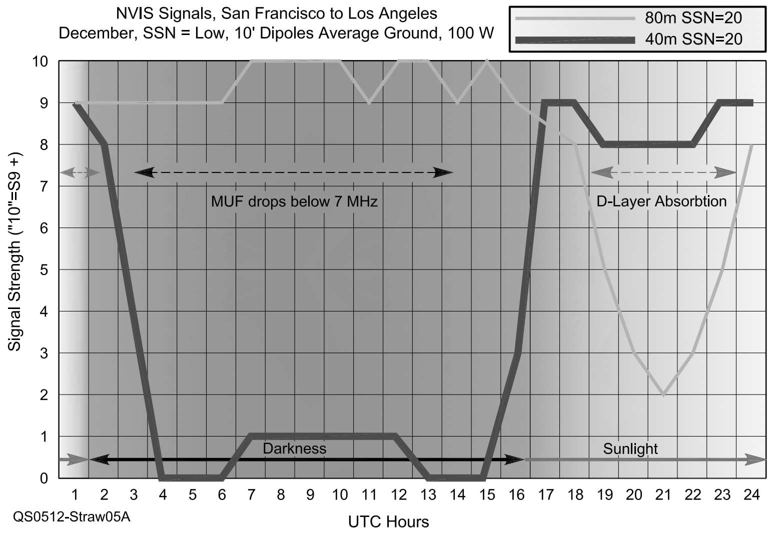

Fig 11—VOACAP calculations for a 350 mile path from San Francisco to Los Angeles, using 10 foot-high flattop dipoles. This plot shows the signal strength in S Units ("S10" = S9+10) for a worst-case month/SSN combination—winter solstice, in December, for a low level of solar activity (SSN = 20). The 40-meter signal drops to a very low level during the night because the MUF drops well below 7.2 MHz. The 80-meter signal drops in the afternoon because of D-layer absorption. For 24-hour communications on this path, the rule of thumb is to select 40 meters during the day and 80 meters during the night.

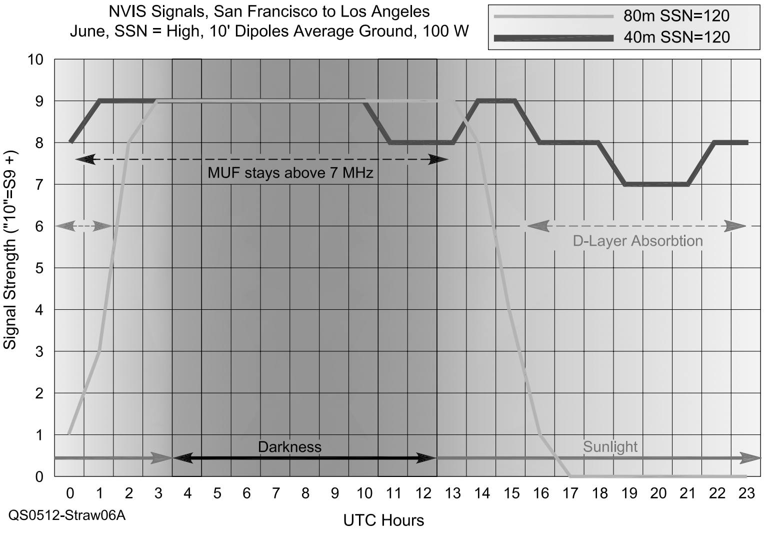

Fig 12—Signal strengths for the San Francisco to Los Angeles path for a worst-case month/SSN combination—summer solstice, in June, for a high level of solar activity (SSN = 120). Now 80 meters drops out more dramatically during the daylight hours, due to increased D-layer absorption. At this high level of solar activity, 40 meters remains open 24 hours with reasonable signal levels. However, the NVIS rule-of-thumb still holds: Use 40 meters during the day; 80 meters at night.

range distances, say out to 500 miles from the center, compared to a 100-foot high antenna. The 20-foot high antenna can discriminate against medium-angle thunderstorm noise in the late afternoon coming from the Arizona desert, although it wouldn't help much for thunderstorms in the Sierra Nevada in central Nevada, which are arriving in San Francisco at high angles, along with the desired NVIS signals.

This is the essence of what NVIS means. NVIS exploits the difference in elevation pattern responses of low horizontally polarized antennas compared to higher horizontal antennas, or even verticals. Over the years, many hams have been lead to believe that higher is always better. This is not quite so true for consistent coverage of medium or short distance signals!

If NVIS only involved putting up a low horizontally polarized antenna on 40 meters the story would end here. However, real cloud warming is more complicated. It also involves the intelligent choice of more than just one operating frequency to achieve reliable all day, all-night communications coverage.

Fig 11 shows the signal strength predicted using *VOACAP* for the 350-mile path from San Francisco to Los Angeles for the month of December for a period of low solar activity (SSN of 20). The antennas used in this case are 10-foot high dipoles, just for some variety. These act almost like 20-foot high Inverted V dipoles. December at a low SSN was chosen as a worst-case scenario because the *winter solstice* occurs on December 21. This is the day that has the fewest hours of daylight in the year. (Contrast this with the *summer solstice*, on June 21, which has the most hours of daylight in the year.) Note that the upper signal limit in Fig 11 is "S10"—a fictitious quantity that allows easier graphing.

S10 is equivalent to S9+, or at least S9+10 dB.

The 40-meter curve in Fig 11 shows that the MUF (maximum usable frequency) actually drops below the 7.2 MHz amateur band after sunset. The signal becomes quite weak for about 14 hours during the night, from about 0300 to 1700 UTC. In a period of low solar activity the 40-meter band thus becomes strictly a *daytime band* on this medium-distance path.

The 80-meter curve in Fig 11 shows strong signals after dusk, through the night and up until about an hour after sunrise. After sunrise, 80 meters starts to suffer absorption in the D layer of the ionosphere and hence the signal strength drops. Here, 80 meters is a true *nighttime band*.

Let's see what happens from San Francisco to Los Angeles during a period of high solar activity (SSN of 120) during the summer solstice in June. **Fig 12** shows that 40 meters now stays open all hours of the day due to the greater number of hours of sunlight in June and because the ionosphere becomes more highly ionized by higher solar activity. Meanwhile, 80 meters still remains a nighttime band during these conditions on this path.

Now, let's look at a shorter-distance path—our 75-mile emergency communications path from San Francisco to Sacramento. We'll again use June during the summer solstice, at a high level of solar activity (SSN of 120) because this represents another worst-case scenario. **Fig 13** shows that 40 meters remains open on this path all day, dropping to a lower signal level just before sunrise. At sunrise, the MUF drops close to 7.2 MHz. 80 meters is still mainly a nighttime band to Sacramento, even though it does yield workable signal levels even during the daylight hours. However, 40 meters is better from 1200 to 0400 UTC, so 40 would be still the right daytime band for this path during the day.

CHOOSING THE RIGHT NVIS FREQUENCY

You can see that a pattern is developing here for efficient NVIS short/medium-distance communications out to 500 miles:

- You should pick a frequency on 40 meters during the day.
- You should pick a frequency on 80 meters during the night.
- You should choose an antenna that emphasizes moderate to high elevation angles, from 40° to almost directly overhead at 90°.

"What about 60 meters?" you might ask. The characteristics on 60 meters fall in-between 40 and 80 meters, although it resembles 40 meters more closely. With characteristics close to that of 40, but with only five channels available and a 50-W power limit, the 60-meter band is of low utility for serious NVIS use.

What about 160 meters? For 100-W level radios, even at the worst-case month or during low solar activity, the critical frequency doesn't fall below 3.8 MHz often enough to destroy the ability to communicate, even for short distances. That is a relief, considering that installing a 160-meter half-wave dipole involves a 255-foot wingspan, and it would need to be elevated at least 30 feet in the center. A short loaded vertical such as a 160-meter mobile whip would have poor response at the high elevation angles needed for NVIS. You could probably put a monster 160-meter horizontal dipole up at a permanent location, but hauling such a thing around in the field would not be an easy task.

SOME OTHER OBSERVATIONS ABOUT NVIS—STRATEGY

You could pose the question about whether NVIS is an operating *mode* or whether it is actually an operating *strategy*. We maintain that NVIS is a strategy. It involves choosing both appropriate frequencies and then appropriate antennas for those frequencies. Fig 13 does show that on short-distance

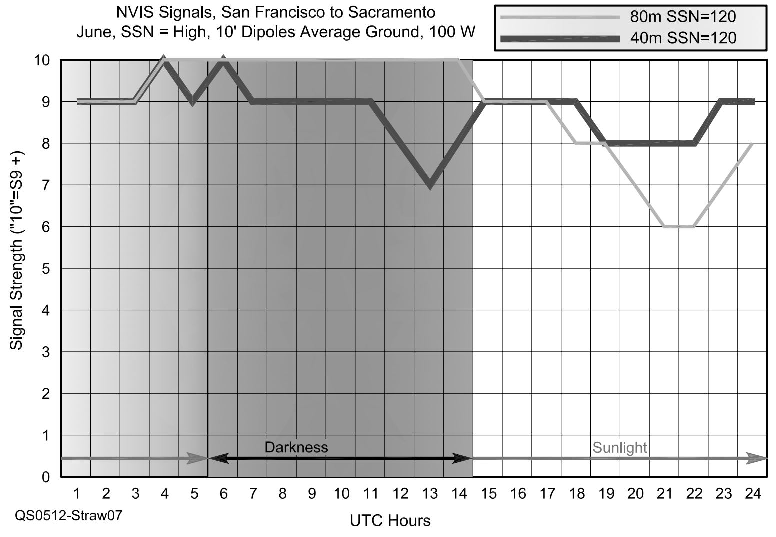

Fig 13—Signal strengths for a 75 mile path—San Francisco to Sacramento. This is for June and SSN = 120. Either band could be used successfully over the full 24-hour period because signal levels are always higher than S6. But the simple NVIS rule-of-thumb still holds: Use 40 meters during the day; 80 meters at night. This simplifies giving instructions to operators unaccustomed to the use of HF.

paths, such as between San Francisco and Sacramento, you could stay on 80 meters all day and night. But if you have to give a single rule-of-thumb to operators who are not very experienced at operating HF, we would tell them to operate on the higher frequency band during the day and on the lower frequency band at night.

SOME OTHER OBSERVATIONS ABOUT NVIS—ANTENNA HEIGHT

Some NVIS aficionados have advocated placing dipoles only a few feet over ground, something akin to saying, "If low is good for NVIS, then lower must be even better." Now we are not claiming that a very low antenna *won't* work in specific instances—for example, covering a small state such as Rhode Island or even just the San Francisco Bay Area.

It certainly is convenient to mount a 40-meter dipole on some 2-foot high red traffic cones! You should be very skeptical, however, about the ability of such antennas to cover all of a large state, such as California or Texas, especially on 80 meters. **Fig 14** shows the computed elevation responses for a number of 80-meter antennas, including a 2-foot-high dipole.

Fig 15B shows the 80 meter geographic coverage plot for 2-foot-high flattop dipoles, compared with the plot in Fig 15A for 20-foot-high inverted V dipoles on both ends of the path. The 2-foot-high dipoles produce about two S-units less signal across all of California than the 20-foot-high inverted V dipoles, at 0300 UTC in December, with an SSN of 20. The reason is that a low dipole will suffer more losses in the ground under it.

The differential between California signals and possible interfering signals from, say, New Mexico, is predicted to be four S-units, the same as it is for the higher inverted V dipole

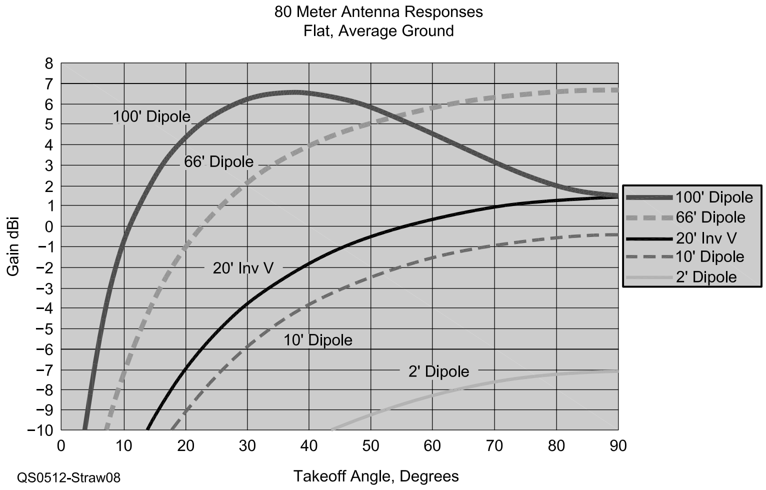

Fig 14—Elevation response patterns for 80 meter antennas over average soil. The shapes track each other rather well, remaining parallel for heights from 2 to 66 feet over flat ground. The 2 foot dipole is substantially down, about 9 dB, from the 20 foot inverted V dipole at all angles.

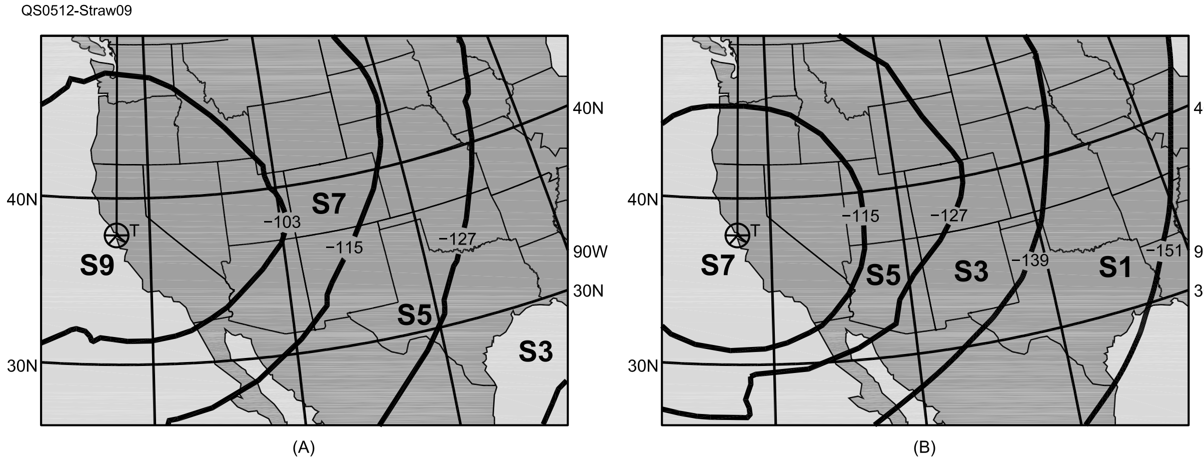

(A)

(B)

Fig 15—Geographic coverage plots for December, SSN = 20, 0300 UTC. At A, antennas are 20 foot-high inverted V dipoles over Average soil. At B, antennas are 2-foot-high flattop dipoles over Average soil. The response for the 2-foot-high antennas is down about 2 S Units, 8 to 12 dB for a typical communications receiver.

at 20 feet. Thus there is no real advantage in terms of signal-to-interference ratio or signal-to-noise ratio (for thunderstorm static crashes) for either height. This is because the shape of all the response curves in Fig 14 below 20 feet essentially track each other in parallel.

However, the lower the antenna, the lower the transmitted signal strength. Physics remain physics. And if you are in an emergency situation operating on batteries, you could reduce power from 100 W to 10 W with a 20-foot high inverted-V antenna and still maintain the same signal strength as a 2-foot high dipole at 100W.

LOW NVIS ANTENNAS AND LOCAL POWERLINE NOISE

Some advocates of really low antennas have stated that the received noise is much lower than that received from higher antennas, and this therefore leads to better signal-to-noise ratios (SNR). How much this is true depends on the source of the noise. If the noise comes from distant thunderstorms, then the SNR advantage going to a 2-foot antenna from a 20-foot-high one is insignificant, as Fig 15 indicates.

If noise is from an arcing insulator on a HV power line half a mile away, that noise will arrive at the antenna as a ground-wave signal. We calculate that the 2-foot antenna receives 4.4 dB less noise by groundwave than a 20-foot-high inverted V dipole. However, at an incoming elevation angle of 45º—suitable for a signal going from Los Angeles to San Francisco—the signal would be down 7.1 dB on the low dipole compared to the higher antenna. The net loss in SNR for the 2-foot-high dipole is thus 7.1– 4.4 or 2.7 dB. Close, but no cigar. Summarizing about really low NVIS antennas:

- A 2-foot-high dipole yields weaker signals, but without an SNR advantage compared to its more elevated brethren.
- A 2-foot-high dipole is a lot easier to trip over at night. We would call this a "knee biter" (or maybe an "ankle biter" if you're really tall).
- You (and your dog) can easily get RF burns from an antenna that is only 2 feet off the ground.

This is not a winning strategy to make friends or QSOs, it seems. But still, a really low dipole may serve your short-range communication needs just fine. But remember, that just as "higher is better" isn't universally true for NVIS (or even longer range) applications, "lower is better" isn't a panacea either.

ELEVATION ANGLES FOR MODERATE DISTANCES ON 75/80 METERS

Fig 16 shows the elevation angles statistics for a 75-meter, 550-mile path from Boston to Cleveland, together with overlays of the elevation patterns for several different types of antennas. These elevation statistics cover all parts of the 11-year solar cycle for this path. The responses for the popular G5RV antenna (described later in this chapter) are shown for two different heights above flat ground: 50 and 100 feet. An 80-meter half-wave sloper ("full sloper") and an 80-meter ground-plane antenna are also shown. All antenna patterns are for "average ground" constants of 5 mS/m conductivity and a dielectric constant of 13.

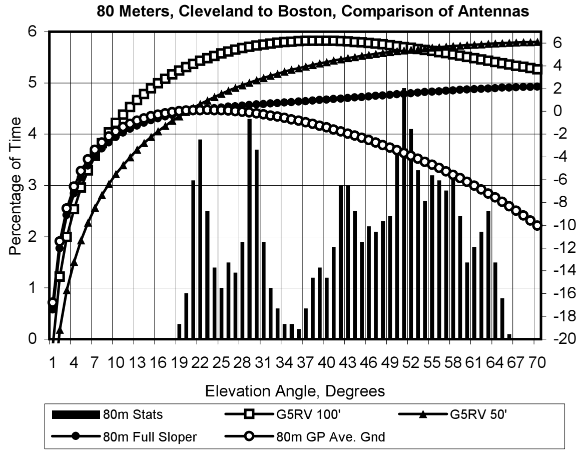

Fig 16—80/75-meter elevation statistics for all portions of the 11-year solar cycle for the path from Cleveland, Ohio, to Boston, Massachusetts, together with the elevation responses for four different multiband antennas. The 100-foot high horizontally polarized G5RV performs well over the entire range of necessary takeoff elevation angles.

At the statistically most significant takeoff angles around 50°, the two horizontally polarized G5RV antennas are about equal. At the second-highest elevation peak near 30°, the 100-foot G5RV has about a 4-dB advantage over its lower counterpart. The full sloper has comparable performance to the 100-foot high G5RV from 1° to about 20° and then gradually rises to its peak at angles higher than 70°. The full sloper is superior to the 50-foot horizontal G5RV at low takeoff elevation angles. The 80-meter ground plane has a deep null directly overhead. At an elevation angle of 70° it is down some 16 dB compared to the 50-foot high horizontal G5RV.

The advantage of antennas suitable for high-angle radiation was vividly demonstrated during a 75-meter QSO one fall evening between N6BV/1 in southern New Hampshire and W1WEF in central Connecticut. This involved a distance of about 100 miles and W1WEF was using his Four Square vertical array. Although W1WEF's signal was S9 on the Four Square, N6BV/1 suggested an experiment. Instead of connecting the so-called "dump power" connector on his Comtek ACB-4 hybrid phasing coupler to a 50-Ω dummy load (the normal configuration), W1WEF switched the dump power to his 100-foot high 80-meter horizontal dipole. W1WEF's signal came up more than 20 dB! The approximately 100-W of power that would otherwise be "wasted" in the dummy load was converted to useful signal.

ELEVATION ANGLES FOR MODERATE DISTANCES ON 40 METERS

Fig 17 shows the situation for the 40-meter band, from Boston to Cleveland, together with the same antennas used for 80 meters in Fig 16. Note that the 100-foot high horizontally

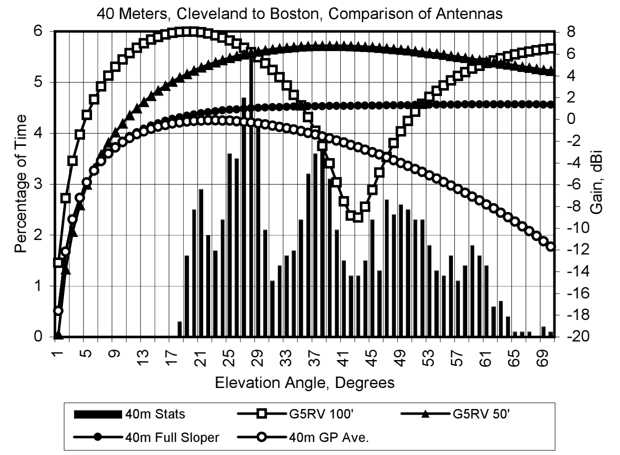

Fig 17—40-meter elevation statistics for the Cleveland to Boston path, together with elevation patterns for four antennas. Here, the 100-foot high horizontally polarized G5RV would have a null in the middle of the range of elevation angles needed for consistent performance on this path. For multiband use on this path to relatively nearby stations, the 50-foot high horizontal antenna would be a better choice than the 100-foot high antenna.

polarized G5RV has about a 16-dB null at an elevation angle of 43°. This doesn't affect things for low elevation angles, but it certainly has a profound effect on signals arriving between about 30° to 60°, especially when compared to the 50-foot high horizontal G5RV. The 40-meter full sloper beats out the high horizontal antenna from about 35° to 50°. And the ground plane is obviously not the antenna of choice for this moderate-range path from Boston to Cleveland, although it is still a good performer on longer-distance paths, with their low takeoff angles.

A 100-foot high multiband dipole is about ⅜-λ high on 75/80 meters. It is an excellent antenna for general-purpose local and DXing operation. But the same dipole used on 40 meters becomes ¾-λ high. At that height, the nulls in its elevation pattern give large holes in coverage for nearby 40-meter contacts. Many operators have found that a 40- to 50-foot high dipole on 40 meters gives them far superior performance for close-in QSOs, when compared to a high dipole, or even a high 2-element 40-meter Yagi.

NVIS SUMMARY

The use of NVIS strategies to cover close-in and intermediate distance communications within about 600 miles involves the intelligent choice of low HF frequencies. As a rule-of-thumb for ham band NVIS, 40 meters is recommended for use during the day; 80 meters during the night.

NVIS involves the choice of antennas suitable for this strategy. Horizontally polarized dual-band 80 and 40-meter flattop dipoles that are mounted higher than about 10 feet high will work adequately for portable operations. Dual-band 80 and 40-meter inverted V dipoles supported 20 feet above the ground at the center can also work well in portable operations.

Single-band 40-meter flattop antennas about 30 feet high and 80-meter flattop antennas about 60 feet high can do a good job for fixed locations.

Horizontal Antennas for the Low Bands

As shown in Chapter 3, The Effects of Ground, and here, radiation angles from horizontal antennas are a very strong function of the height above ground in wavelengths. Typically for DX work heights of λ/2 to 1 λ are considered to be a minimum. As we go down in frequency these heights become harder to realize. For example, a 160-meter dipole at 70 feet is only 0.14 λ high. This antenna will be very effective for local and short distance QSOs but not very good for DX work. Despite this limitation, horizontal antennas are very popular on the lower bands because the low frequencies are often used for short range communications, local nets and rag chewing. Also horizontal antennas do not require extensive ground systems to be efficient.

DIPOLE ANTENNAS

Half-wave dipoles and variations of these can be a very good choice for a low band antenna. A variety of possibilities are shown in **Fig 18**. An untuned or "flat" feed line is a logical choice on any band because the losses are low, but this generally limits the use of the antenna to one band. Where only single-band operation is wanted, the λ/2 antenna fed with open-wire line is one of the most popular systems on the 3.5 and 7-MHz bands.

If the antenna is a single-wire affair, its impedance is in the vicinity of 60 Ω, depending on the height and the ground characteristics. The most common way to feed the antenna is with 50- or 75-Ω coaxial line. Heavy coaxial lines present support problems because they are a concentrated weight at the center of the antenna, tending to pull the center of the antenna down. This can be overcome by using an auxiliary pole to take at least some of the weight of the line. The line should come away from the antenna at right angles, and it can be of any length.

Folded Dipoles

A folded dipole (Fig 18B and C) has an impedance of about 300 Ω, and can be fed directly with any length of 300-Ω line. The folded dipole can be made of ordinary wire spaced by lightweight wooden or plastic spacers, 4 or 6 inches long, or a piece of 300 or 450-Ω twin-lead or ladder line.

A folded dipole can be fed with a 600-Ω open wire line with only a 2:1 SWR, but a nearly perfect match can be ob-

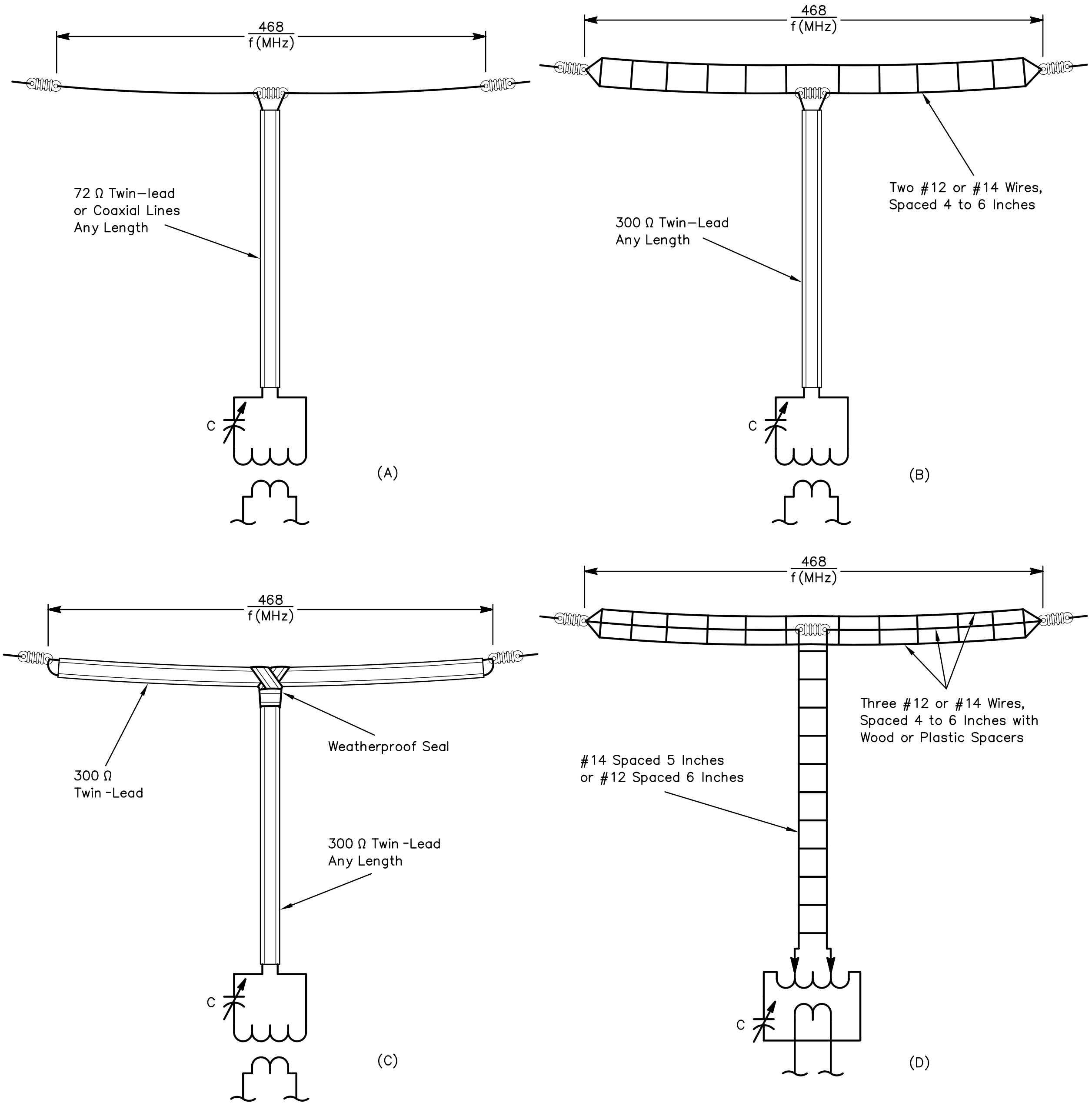
468 / f (MHz)
72 Ω Twin—lead or Coaxial Lines Any Length
C
(A)
468 / f (MHz)
Two #12 or #14 Wires, Spaced 4 to 6 Inches
300 Ω Twin—Lead Any Length
C
(B)
468 / f (MHz)
300 Ω Twin -Lead
Weatherproof Seal
300 Ω Twin -Lead Any Length
C
(C)
468 / f (MHz)
Three #12 or #14 Wires, Spaced 4 to 6 Inches with Wood or Plastic Spacers
#14 Spaced 5 Inches or #12 Spaced 6 Inches
C
(D)

Fig 18—Half-wavelength antennas for single band operation. The multiwire types shown in B, C and D offer a better match to the feeder over a somewhat wider range of frequencies but otherwise the performances are identical. The feeder should run away from the antenna at a right angle for as great a distance as possible. In the coupling circuits shown, tuned circuits should resonate to the operating frequency. In the series-tuned circuits of A, B, and C, high L and low C are recommended, and in D the inductance and capacitance should be similar to the output-amplifier tank, with the feeders tapped across at least ½ the coil. The tapped-coil matching circuit shown in Chapter 25 can be substituted in each case.

tained with a three-wire dipole fed with either 450-Ω ladder line or 600-Ω open wire line. One advantage of the two- and three-wire antennas over the single wire is that they offer a better match over a wider band. This is particularly important if full coverage of the 3.5-MHz band is contemplated.

Inverted-V Dipole

The halves of a dipole may be sloped to form an inverted V, as shown in **Fig 19**. This has the advantages of requiring only a single high support and less horizontal space. There will be some difference in performance between a normal horizontal dipole and the inverted V as shown by the radiation patterns in **Fig 20**. There is small loss in peak gain and the pattern is less directional.

Sloping of the wires results in a raising of the resonant frequency and a decrease in feed-point impedance and bandwidth. Thus, for the same frequency, the length of the dipole must be increased somewhat. The angle at the apex is not critical, although it should probably be made no smaller than 90°. Because of the lower impedance, a 50-Ω line should be used. For those who are dissatisfied with anything but a perfect match, the usual procedure is to adjust the angle for lowest SWR while keeping the dipole resonant by adjustment of length. Bandwidth may be increased by using multiconductor elements, such as a cage configuration.

PHASED HORIZONTAL ARRAYS

Phased arrays with horizontal elements, which provide some directional gain, can be used to advantage at 7 MHz, if they can be placed at least 40 feet above ground. At 3.5 MHz heights of 70 feet or more are needed for any real advantage.

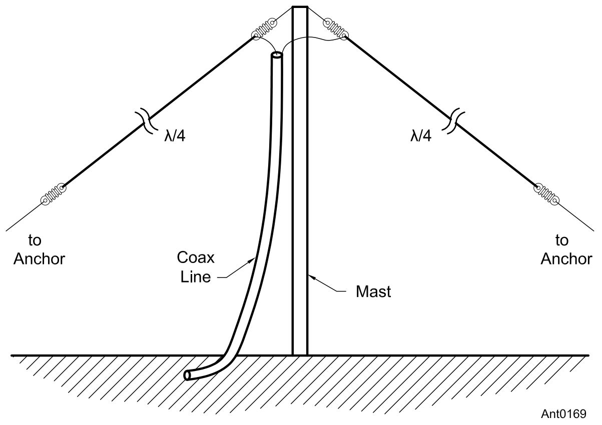

Fig 19—The inverted-V dipole. The length and apex angle should be adjusted as described in the text.

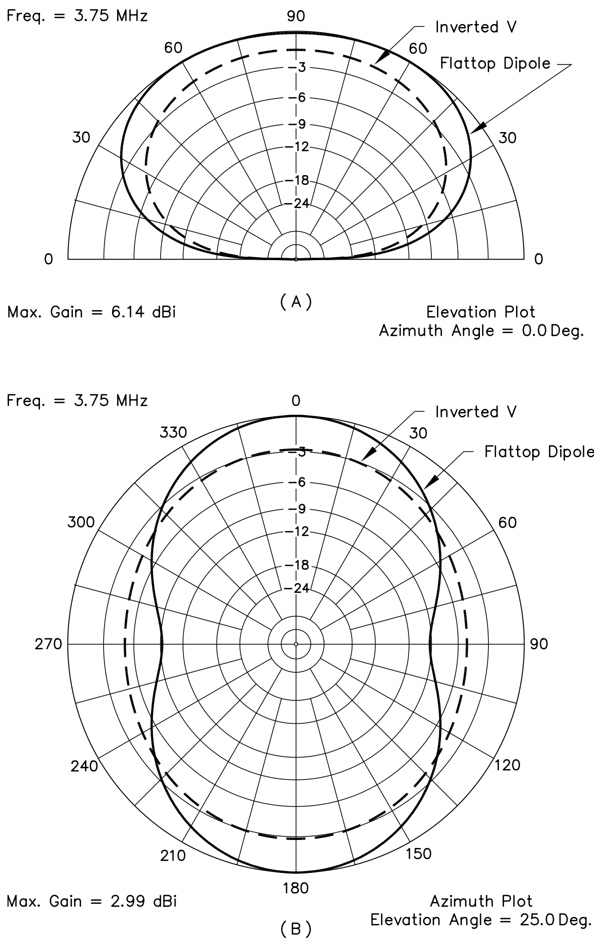

Fig 20—At A, elevation and at B, azimuthal radiation patterns comparing a normal 80-meter dipole and an inverted-V dipole. The center of both dipoles is at 65 feet and the ends of the inverted V are at 20 feet. The frequency is 3.750 MHz.

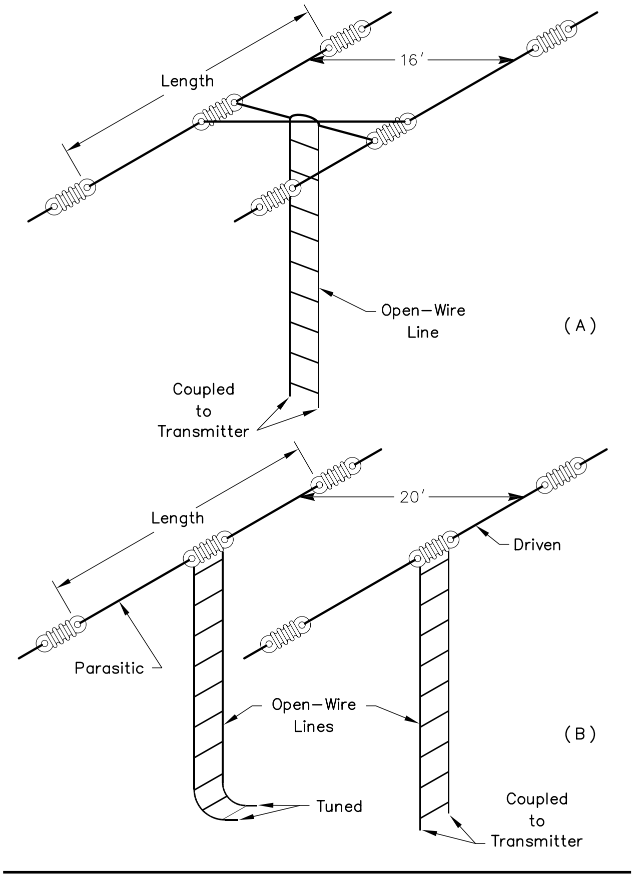

Fig 21—Directional antennas for 7 MHz. To realize any advantage from these antennas, they should be at least 40 feet high. At A, system is bidirectional. At B, system is unidirectional in a direction depending upon the tuning conditions of the parasitic element. The length of the elements in either antenna should be exactly the same, but any length from 60 to 150 feet can be used.
If the length of the antenna at A is between 60 and 80 feet, the antenna will be bidirectional along the same line on both 7 and 14 MHz. The system at B can be made to work on 7 and 14 MHz in the same way, by keeping the length between 60 and 80 feet.

Many of the driven arrays discussed in Chapter 8 and even some of the Yagis discussed in Chapter 11 can be used as fixed directional antennas. If a bidirectional characteristic is desired, the W8JK array, shown in **Fig 21A**, is a good one. If a unidirectional characteristic is required, two elements can be mounted about 20 feet apart and provision included for tuning one of the elements as either a director or reflector, as shown in Fig 21B.

The parasitic element is tuned at the end of its feed line with a series or parallel-tuned circuit (whichever would normally be required to couple power into the line), and the proper tuning condition can be found by using the system for receiving and listening to distant stations along the line to the rear of the antenna. Tuning the feeder to the parasitic element can minimize the received signals from the back of the antenna. This is in effect adjusting the antenna for maximum front-to-back ratio. Maximum front-to-back does not occur at the same point as maximum forward gain but the loss in forward gain is very small. Adjusting the antenna for maximum forward gain (peaking received signals in the forward direction) may increase the forward gain slightly but will almost certainly result in relatively poor front-to-back ratio.

A MODIFIED EXTENDED DOUBLE ZEPP

If the distance between the available supports is greater than λ/2 then a very simple form of a single wire collinear array can be used to achieve significant gain. The *extended double Zepp* antenna has long been used by amateurs and is discussed in Chapter 8, Multielement Arrays. A simple variation of this antenna with substantially improved bandwidth can be very useful on 3.5 and 7.0 MHz. The following material has been taken from an article by Rudy Severns, N6LF, in *The ARRL Antenna Compendium Vol 4*.

The key to improving the characteristics of a standard double-extended Zepp is to modify the current distribution. One of the simplest ways to do this is to insert a reactance(s) in series with the wire. This could either be an inductor(s) or a capacitor(s). In general, a series capacitor will have a higher Q and therefore less loss. With either choice it is desirable to use as few components as possible.

As an initial trial at 7 MHz, only two capacitors, one on each side of the antenna, were used. The value and position of the capacitors was varied to see what would happen. It quickly became clear that the reactance at the feed point could be tuned out by adjusting the capacitor value, making the antenna look essentially like a resistor over the entire band. The value of the feed-point resistance could be varied from less than 150 Ω to over 1500 Ω by changing the location of the capacitors and adjusting their values to resonate the antenna.

A number of interesting combinations were created. The one ultimately selected is shown in **Fig 22**. The antenna is 170 feet in length. Two 9.1 pF capacitors are located 25 feet out each side of the center. The antenna is fed with 450-Ω transmission line and a 9:1 three-core Guanella balun used at the transmitter to convert to 50 Ω. The transmission line can be any convenient length and it operates with a very low SWR.

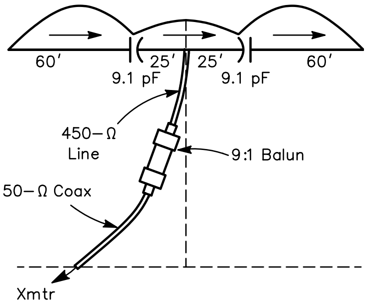

Fig 22—Schematic for modified N6LF Double Extended Zepp. Overall length is 170 feet, with 9.1 pF capacitors placed 25 feet each side of center.

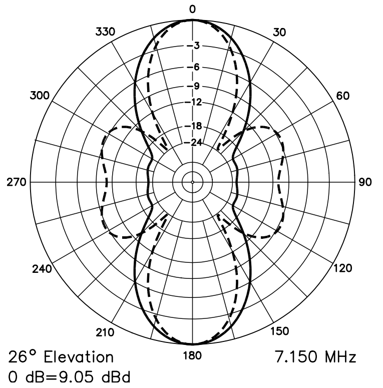

Fig 23—Azimuth pattern for N6LF Double Extended Zepp (solid line), compared to classic Double Extended Zepp (dashed line). The main lobe for the modified antenna is slightly broader than that of the classic model, and the sidelobes are suppressed better.

That's all there is to it. The radiation pattern, overlaid with that for a standard DEZepp for comparison, is shown in **Fig 23**. The sidelobes are now reduced to below 20 dB. The main lobe is now 43° wide at the 3-dB points, as opposed to 35° for the original DEZepp. The antenna has gain over a dipole for > 50° now and the gain of the main lobe has dropped only 0.2 dB below the original DEZepp.

Experimental Results

The antenna was made from #14 wire and the capacitors were made from 3.5-inch sections of RG-213, shown in **Fig 24A**. Note that great care should be taken to seal out moisture in these capacitors. The voltage across the capacitor for 1.5 kW will be about 2000 V so any corona will quickly destroy the capacitor.

A silicon sealant was used and then both ends covered

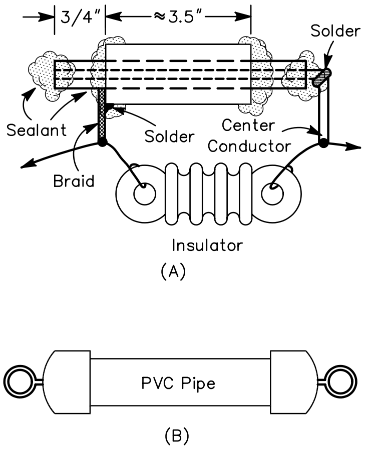

Fig 24—Construction details for series capacitor made from RG-213 coaxial cable. At A, the method used by N6LF is illustrated. At B, a suggested method to seal capacitor better against weather is shown, using a section of PVC pipe with end caps.

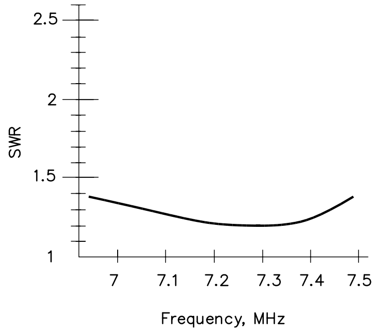

Fig 25—Measured SWR curve across 40-meter band for N6LF DEZepp.

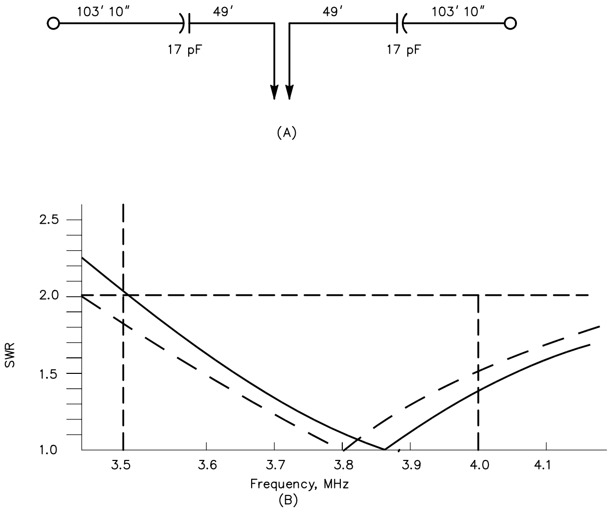

Fig 26—75/80-meter modified Double Extended Zepp, designed using *NEC Wires*. At A, a schematic is shown for antenna. At B, SWR curve is shown across 75/80-meter band. Solid line shows measured curve for W7ISV antenna, which was pruned to place SWR minimum higher in the band. The dashed curve shows the computed response when SWR minimum is set to 3.8 MHz.

with coax seal, finally wrapping it with plastic tape. The solder balls indicated on the drawing are to prevent wicking of moisture through the braid and the stranded center conductor. This is a small but important point if long service out in the weather is expected. An even better way to protect the capacitor would be to enclose it in a short piece of PVC pipe with end caps, as shown in Fig 24B.

Note that all RG-8 type cables do not have exactly the same capacitance per foot and there will also be some end effect adding to the capacitance. If possible the capacitor should be trimmed with a capacitance meter. It isn't necessary to be too exact—the effect of varying the capacitance ±10% was checked and the antenna still worked fine.

The results proved to be close to those predicted by the computer model. **Fig 25** shows the measured value for SWR across the band. These measurements were made with a Bird directional wattmeter. The worst SWR is 1.35:1 at the low end of the band.

Dick Ives, W7ISV, erected an 80-meter version of the antenna, shown in **Fig 26A**. The series capacitors are 17 pF. Since he isn't interested in CW, Dick adjusted the length for the lowest SWR at the high end of the band, as shown in the SWR curve (Fig 26B). The antenna could have been tuned somewhat lower in frequency and would then provide an SWR < 2:1 over the entire band, as indicated by the dashed line.

This antenna provides wide bandwidth and moderate gain over the entire 75/80-meter band. Not many antennas will give you that with a simple wire structure.

Vertical Antennas

On the low bands quarter-wave high vertical antennas become increasingly attractive, especially for DX work, because they provide a means for lowering the radiation angle. This is especially true where practical heights for horizontally polarized antennas are too low. In addition, verticals can be very simple and unobtrusive structures. For example, it is very easy to disguise a vertical as a flagpole. In fact an actual flagpole may be used as a vertical. Performance of a vertical is determined by several factors:

- Height of the vertical portion of the radiator
- The ground or counterpoise system efficiency, if one is used
- Ground characteristics in the near- and far-field regions
- The efficiency of loading elements and matching networks

THE HALF-WAVE VERTICAL DIPOLE (HVD)

The simplest form of vertical is that of a half-wave vertical dipole, an HVD. This is a horizontal dipole turned 90º so that it is perpendicular to the ground under it. Of course, the top end of such an antenna must be at least a half wave above the ground or else it would be touching the ground. This poses quite a construction challenge if the builder wants a free-standing low-frequency antenna. Hams fortunate enough to have tall trees on their property can suspend wire HVDs from these trees. Similarly, hams with two tall towers can run rope catenaries between them to hold up an HVD.

A vertical half-wave dipole has some operational advantages compared to a more-commonly used vertical configuration—the quarter-wave vertical used with some sort of above-ground counterpoise or an on-ground radial system. See **Fig 27** and 27B, which shows the two configurations discussed here. In each case, the lowest part of each antenna is 8 feet above ground, to prevent passersby from being able to touch any live wire. Each antenna is assumed to be made of #14 wire resonant on 80 meters.

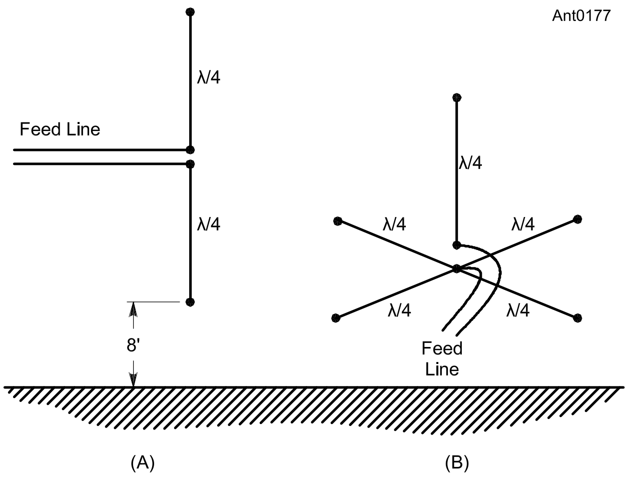

Fig 27—At A, a 80-meter half-wave vertical dipole elevated 8 feet above the ground. The feed line is run perpendicularly away from the dipole. At B, a "ground plane" type of quarter-wave vertical, with four elevated resonant radials. Both antennas are mounted 8 feet above the ground to keep them away from passersby.

Feeding a Half-Wave Vertical Dipole

Fig 28 compares elevation patterns for the two antennas for "average ground." You can see that the half-wave vertical dipole has about 1.5 dB higher peak gain, since it compresses the vertical elevation pattern down somewhat closer to the horizon than does the quarter-wave ground plane. Another advantage to using a half-wave radiator besides higher gain is that less horizontal "real estate" is needed compared to a quarter-wave vertical with its horizontal radials.

The obvious disadvantage to an HVD is that it is taller than a quarter-wave ground plane. This requires a higher support (such as a taller tree) if you make it from wire, or a longer element if you make it from telescoping aluminum tubing.

Another problem is that theory says you must dress the feed line so that it is perpendicular to the half-wave radiator. This means you must support the coax feed line above ground for some distance before bringing the coax down to ground level. A question immediately arises: How far must you go out horizontally with the feed line before going to ground level to eliminate common-mode currents that are radiated onto the coax shield? Such common-mode currents will affect the feed-point impedance as well as the radiation pattern for the antenna system. Quite a bit of distortion in the azimuthal pattern can be created if common-mode currents aren't suppressed, usually by using a common-mode choke, also known as a current balun.

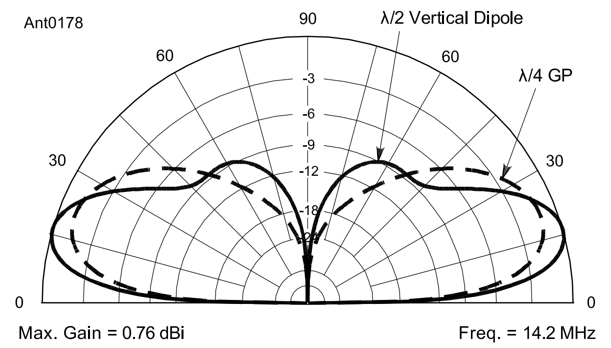

Fig 28—A comparison of the elevation patterns for the two antennas in Fig 27. The peak gain of the HVD is about 1.5 dB higher than that for the quarter-wave ground-plane radiator with radials.

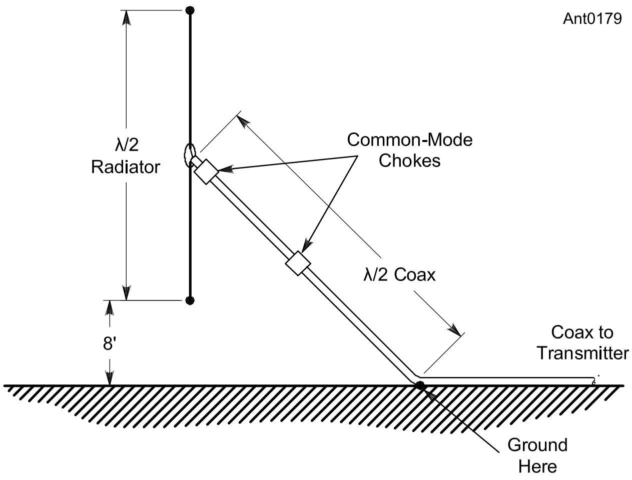

Fig 29—A 20-meter HVD whose bottom is 8 feet above ground. This is fed with a λ/2 of RG-213 coax. This system uses a common-mode choke at the feed point and another λ/4 down the line. The resulting azimuthal radiation pattern is within 0.4 dB of being perfectly circular. The "wingspan" of this antenna system is 27 feet from the radiator to the point where the coax comes to ground level.

Constructing such a common-mode choke is very simple: Three large ferrite beads are slipped over the coax (before the connectors are soldered on or else they won't fit!) and taped in place. The only problem with this scheme is that an additional support (some sort of "skyhook") is required to support the coax horizontally. Let's try to simplify the installation, by slanting the feed-line coax down to ground from the feed point at a fairly steep angle of about 30º from vertical. See **Fig 29**.

Note that the bottom end of the coax in Fig 29 is grounded to a ground rod. This serves several purposes—this serves as a mechanical connection to hold the coax in place and it provides some protection against lightning strikes. Now, as a purely practical matter, just how picky are we being here? What if we skip the second common-mode choke and just use one at the feed point? The computer models predicts that there will be some distortion in the azimuthal pattern—about 1.1 dB worth. Whether this is serious is up to you. However, you may find other problems with common-mode currents on the coax shield—problems such as RF in the shack or variable SWR readings depending on the way coax is routed in the shack. The addition of three extra ferrite beads to suppress the common-mode currents is cheap insurance.

Later in this chapter we'll discuss shortened vertical antennas, ones arranged both as vertical dipoles and as vertical monopoles with radial systems.

MONOPOLE VERTICALS WITH GROUND PLANE RADIALS

For best performance the vertical portion of a ground-plane type of antenna should be λ/4 or more, but this is not an absolute requirement. With proper design, antennas as short as 0.1 λ or even less can be efficient and effective. Antennas shorter than λ/4 will be reactive and some form of loading and perhaps a matching network will be required.

If the radiator is made of wire supported by nonconducting material, the approximate length for λ/4 resonance can be found from:

$$\ell_{\text{feet}} = \frac{234}{f_{\text{MHz}}} \qquad \text{(Eq 1)}$$

For tubing, the length for resonance must be shorter than given by the above equation, as the length-to-diameter ratio is lower than for wire (see Chapter 2, Antenna Fundamentals). For a tower, the resonant length will be shorter still. In any case, after installation the antenna length (height) can be adjusted for resonance at the desired frequency.

The effect of ground characteristics on losses and elevation pattern is discussed in detail in Chapter 3, The Effects of Ground. The most important points made in that discussion are the effect of ground characteristics on the radiation pattern and the means for achieving low ground-loss resistance in a buried ground system. As ground conductivity increases, low-angle radiation improves. This makes a vertical very attractive to those who live in areas with good ground conductivity. If your QTH is on a saltwater beach, then a vertical would be very effective, even when compared to horizontal antennas at great height.

When a buried-radial ground system is used, the efficiency of the antenna will be limited by the loss resistance of the ground system. The ground can be a number of radial wires extending out from the base of the antenna for about λ/4. Driven ground rods, while satisfactory for electrical safety and for lightning protection, are of little value as an RF ground for a vertical antenna, except perhaps in marshy or beach areas. As pointed out in Chapter 3, many long radials are desirable. In general, however, a large number of short radials are preferable to only a few long radials, although the best system would have 60 or more radials longer than λ/4. An elevated system of radials or a ground screen (*counterpoise*) may be used instead of buried radials, and can result in an efficient antenna.

ELEVATED RADIALS AND COUNTERPOISES

Elevated radials, isolated from ground, can be used in place of an extensive buried radial system. Work by Al Christman, K3LC (ex-KB8I), has shown that 4 to 8 elevated radials can provide performance comparable to a 120 λ/4-long buried wires. This is especially important for the low bands, where such a buried ground system is very large and impractical for most amateurs. An elevated ground system is sometimes referred to as a *ground plane* or *counterpoise*. **Fig 30** compares buried and elevated ground systems, showing the difference in current flow in the two systems.

An elevated ground can take several forms. A number of wires arranged with radial symmetry around the base of the antenna is shown in Fig 30B. Four radials are normally

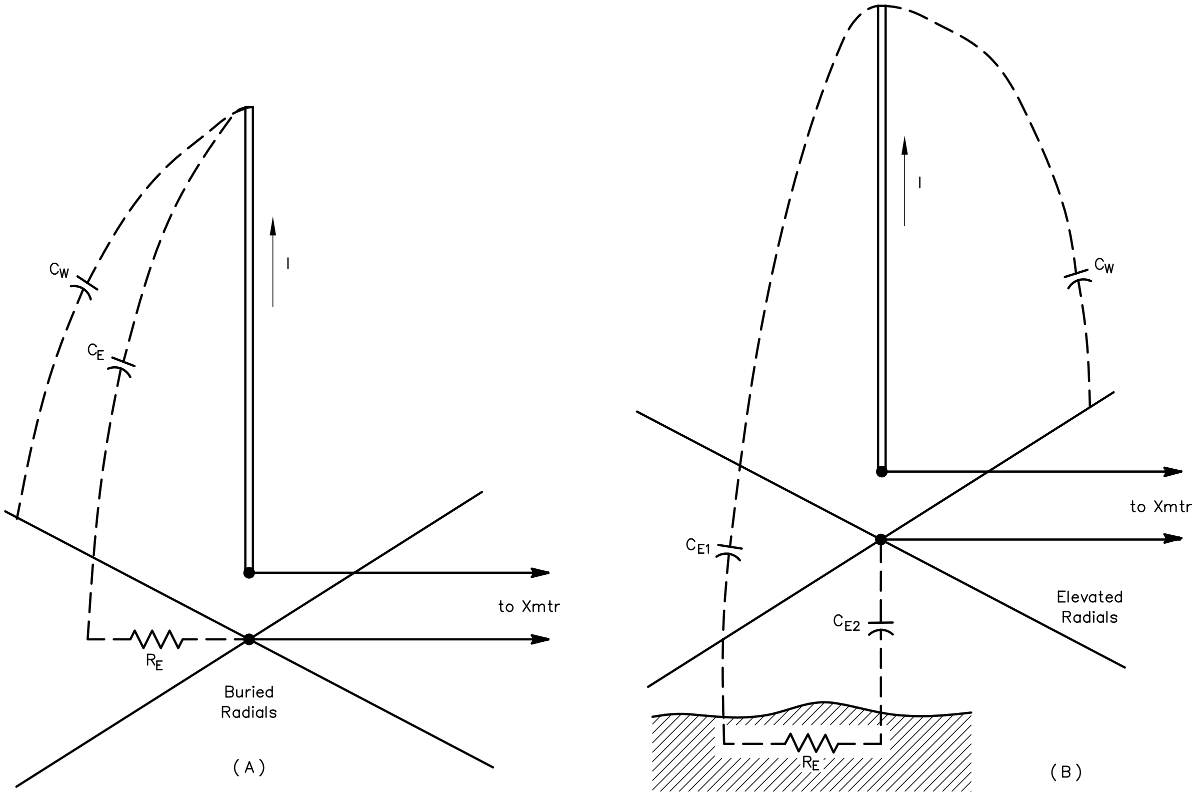

Fig 30—How earth currents affect the losses in a short vertical antenna system. At A, the current through the combination of C_E and R_E may be appreciable if C_E is much greater than C_W, the capacitance of the vertical to the ground wires. This ratio can be improved (up to a point) by using more radials. By raising the entire antenna system off the ground, C_E (which consists of the series combination of C_{E1} and C_{E2}) is decreased while C_W stays the same. The radial system shown at B is sometimes called a *counterpoise*.

used, but as few as two, or as many as eight, can be used. For a given height of vertical, the length of the radials can be adjusted to resonate the antenna. For a λ/4 vertical, the radials are normally λ/4 long.

In the case of a multiband vertical, two or more sets of radials, with different lengths, may be interleaved. The radials associated with each band are adjusted for resonance on their associated band.

A counterpoise is most commonly a system of elevated radials, where the radial wires are interconnected with jumpers, as shown in **Fig 31**. As illustrated in Fig 30, the purpose of the elevated-ground system is to provide a return path for the displacement currents flowing in the vicinity of the antenna. The idea is to minimize the current flowing through the ground itself, which is usually very lossy. By raising the radials above ground most of the current will flow in the radials, which are good conductors. This allows a simple radial system to provide a very efficient ground. However, there is a price to be paid for this.

The ground system now has a direct effect on the feedpoint impedance, introducing reactance as well as resistance, and is relatively narrow band. For a given vertical height, the

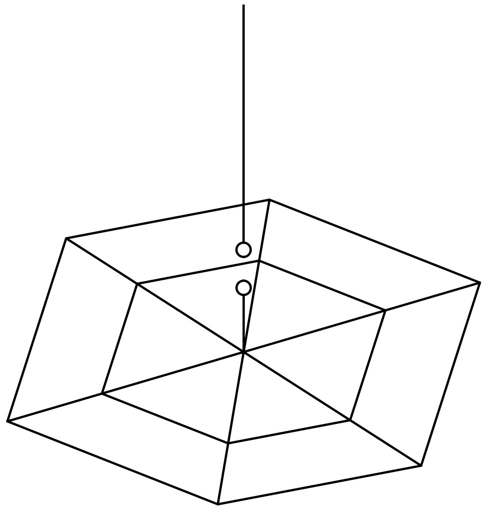

Fig 31—Counterpoise, showing the radial wires connected together by cross wires. The length of the perimeter of the individual meshes should be < λ/4 to prevent undesired resonances. Sometimes the center portion of the counterpoise is made from wire mesh.

Table 2

Illustration of the effect of variable vertical height (L_1) on elevated radial length (L2) and R_R.
#12 wire, elevated 5 feet over average ground at 3.525 MHz.

L_1 (λ)	L_1 (feet)	L_2 (feet)	R_R (Ω)
0.225	62.8	94	28.8
0.25	69.8	67	38.4
0.27	75.0	45	51.0
0.3	83.7	24	75.9

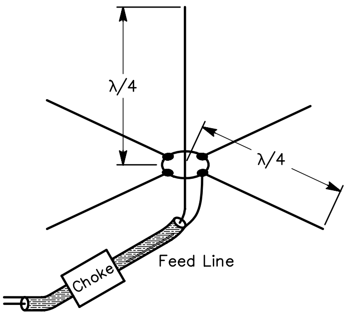

Fig 32—The ground-plane antenna. Power is applied between the base of the vertical radiator and the center of the ground plane, as indicated in the drawing. Decoupling from the transmission line and any conductive support structure is highly desirable.

radial length must be adjusted to resonate the antenna. The length of the radials must be readjusted for each band if a multiband vertical is used. As pointed out above, this usually means the installation of a set of radials for each band. To minimize current flowing in the ground, the antenna, ground plane and feed line must be isolated from ground for RF. More on this later.

The height of the vertical does not have to be exactly λ/4. Other lengths may be used and the antenna may be resonated by adjusting the length of the radials. **Table 2** gives a comparison between three different vertical lengths in an antenna using four elevated radials at 3.525 MHz.

An important feature of Table 2 is the dramatic reduction in radial length (L_2) with even a small increase in vertical height (L_1). For example, increasing the height by 5 feet reduces the radial length by 22 feet on 80 meters. On the other hand even a small decrease in L_1 can cause a substantial increase in L_2. This would be very undesirable, since the area required by the radials is already considerable. Notice also that the small increase in height raises R_R to 51 Ω. This trick of increasing the height slightly to reduce the size of the elevated ground system and to increase the input resistance can be very useful. In a following section the use of *top loading* for short antennas will be discussed. Top loading can also be used on a λ/4 vertical to achieve the same effect as increasing the height—the ability to use shorter radials and a better match.

GROUND-PLANE ANTENNAS

The ground-plane antenna is a λ/4 vertical with four radials, as shown in **Fig 32**. The entire antenna is elevated above ground. A practical example of a 7-MHz ground-plane antenna is given in **Fig 33**. As explained earlier, elevating the antenna reduces the ground loss and lowers the radiation angle somewhat. The radials are sloped downward to make the feed-point impedance closer to 50 Ω.

The feed-point impedance of the antenna varies with the height above ground, and to a lesser extent varies with the ground characteristics. **Fig 34** is a graph of feed-point resistance (R_R) for a ground-plane antenna with the radials parallel to the ground. R_R is plotted as a function of height above ground. Notice that the difference between perfect ground and average ground (ε=13 and σ = 0.005 S/m) is small, except when quite close to ground. Near ground R_R is between 36 and 40 Ω. This is a reasonable match for 50-Ω feed line but as the antenna is raised above ground R_R drops to approximately 22 Ω, which is not a very good match. The feed-point resistance can be increased by sloping the radials downward, away from the vertical section.

The effect of sloping the radials is shown in **Fig 35**. The

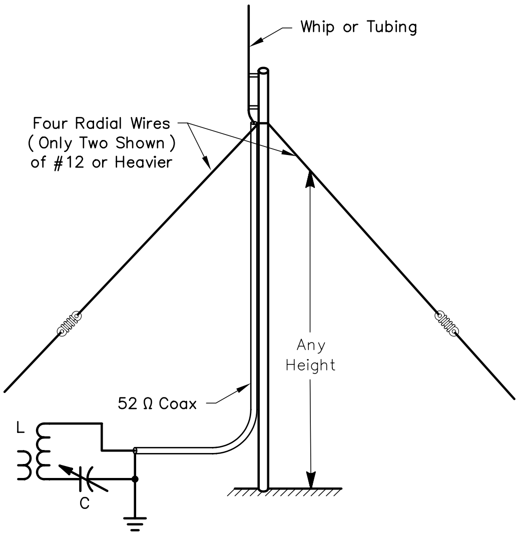

Fig 33—A ground-plane antenna is effective for DX work on 7 MHz. Although its base can be any height above ground, losses in the ground underneath will be reduced by keeping the bottom of the antenna and the ground plane as high above ground as possible. Feeding the antenna directly with 50-Ω coaxial cable will result in a low SWR. The vertical radiator and the radials are all λ/4 long electrically. Contrary to popular myth, the radials need not necessarily be 5% longer than the radiator. Their physical length will depend on their length-to-diameter ratios, the height over ground and the length of the vertical radiator, as discussed in text.

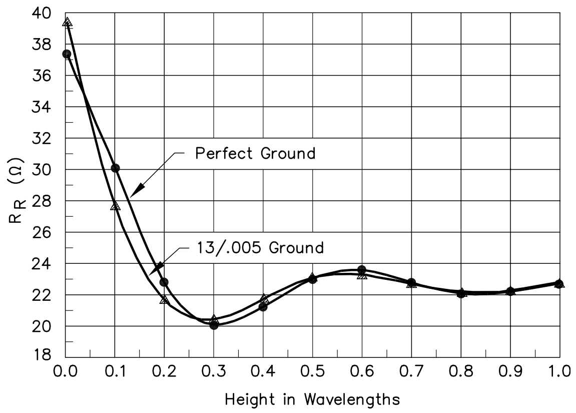

Fig 34—Radiation resistance of a 4-radial ground-plane antenna as a function of height over ground. Perfect and average ground are shown. Frequency is 3.525 MHz. Radial angle (θ) is 0°.

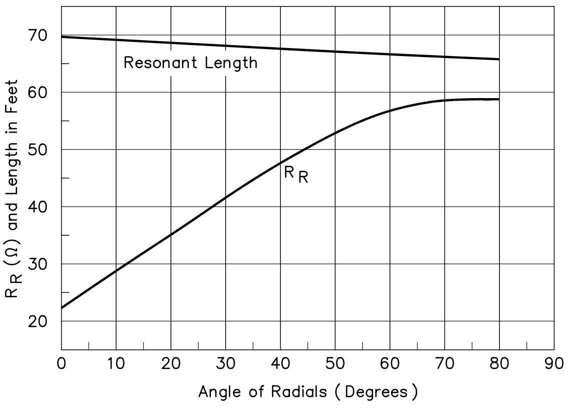

Fig 35—Radiation resistance and resonant length for a 4-radial ground-plane antenna > 0.3 λ above ground as a function of radial droop angle (θ).

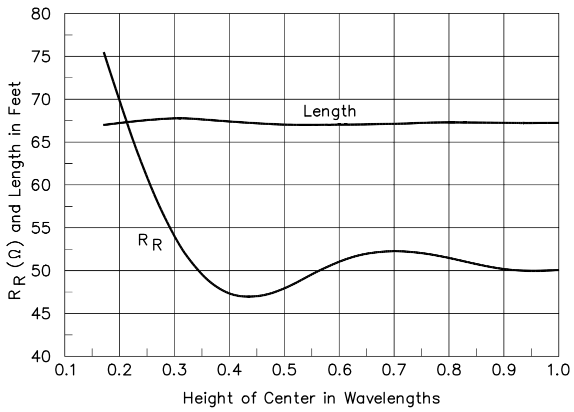

Fig 36—Radiation resistance and resonant length for a 4-radial ground-plane antenna for various heights above average ground for radial droop angle θ = 45°.

graph is for an antenna well above ground (> 0.3 λ). Notice that R_R = 50 Ω when the radials are sloped downward at an angle of 45°, a convenient value. The resonant length of the antenna will vary slightly with the angle. In addition, the resonant length will vary a small amount with height above the ground. It is for these reasons, as well as the effect of conductor diameter, that some adjustment of the radial lengths is usually required. When the ground-plane antenna is used on the higher HF bands and at VHF, the height above ground is usually such that a radial sloping angle of 45° will give a good match to 50-Ω feed line.

The effect of height on R_R with a radial angle of 45° is shown in **Fig 36**. At 7 MHz and lower, it is seldom possible to elevate the antenna a significant portion of a wavelength and the radial angle required to match to 50-Ω line is usually of the order of 10° to 20°. To make the vertical portion of the antenna

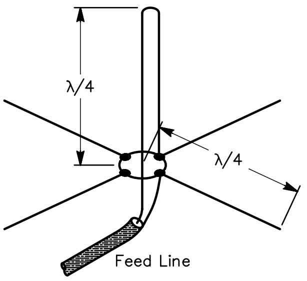

Fig 37—The folded monopole antenna. Shown here is a ground plane of four λ/4 radials. The folded element may be operated over an extensive counterpoise system or mounted on the ground and worked against buried radials and the earth. As with the folded dipole antenna, the feed-point impedance depends on the ratios of the radiator conductor sizes and their spacing.

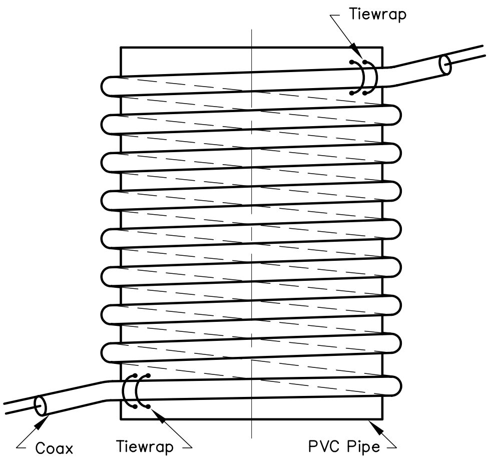

Fig 38—A choke balun with sufficient impedance to isolate the antenna properly can be made by winding coaxial cable around a section of plastic pipe. Suitable dimensions are given in the text.

as long as possible, it may be better to accept a slightly poorer match and keep the radials parallel to ground.

The principles of the folded dipole (Fig 18) can also be applied to the ground-plane antenna, as shown in **Fig 37**. This is the *folded monopole* antenna. The feed-point resistance can be controlled by the number of parallel vertical conductors and the ratios of their diameters.

As mentioned earlier, it is important in most installations to isolate the antenna from the feed line and any conductive supporting structure. This is done to minimize the return current conducted through the ground. A return current on the feed line itself or the support structure can drastically alter the radiation pattern, usually for the worse. For these reasons, a balun (see Chapter 26, Coupling the Line to the Antenna) or other isolation scheme must be used. 1:1 baluns are effective for the higher bands but at 3.5 and 1.8 MHz commercial baluns often have too low a shunt inductance to provide adequate isolation. It is very easy to recognize when the isolation is inadequate. When the antenna is being adjusted while watching an isolated impedance or SWR meter, adjustments may be sensitive to your touching the instrument. After adjustment and after the feed line is attached, the SWR may be drastically different. When the feed line is inadequately isolated, the apparent resonant frequency or the length of the radials required for resonance may also be significantly different from what you expect.

In general, an isolation choke inductance of 50 to 100 μH will be needed for 3.5 and 1.8-MHz ground-plane antennas. One of the easiest ways to make the required isolation choke is to wind a length of coaxial cable into a coil as shown in **Fig 38**. For 1.8 MHz, 30 turns of RG-213 wound on a 14-inch length of 8-inch diameter PVC pipe, will make a very good isolation choke that can handle full legal power continuously. A smaller choke could be wound on 4-inch diameter plastic drain pipe using RG-8X or a Teflon insulated cable. The important point here is to isolate or decouple the antenna from the feed line and support structure.

A full-size ground-plane antenna is often a little impractical for 3.5-MHz and quite impractical for 1.8 MHz, but it can be used at 7 MHz to good advantage, particularly for DX work. Smaller versions can be very useful on 3.5 and 1.8 MHz.

EXAMPLES OF VERTICALS

There are many possible ways to build a vertical antenna—the limits are set by your ingenuity. The primary problem is creating the vertical portion of the antenna with sufficient height. Some of the more common means are:

- A dedicated tower
- Using an existing tower with an HF Yagi on top
- A wire suspended from a tree limb or the side of a building
- A vertical wire supported by a line between two trees or other supports
- A tall pole supporting a conductor
- Flagpoles
- Light standards
- Irrigation pipe
- TV masts

If you have the space and the resources, the most straightforward means is to erect a dedicated tower for a vertical. While this is certainly an effective approach, many amateurs do not have the space or the funds to do this, especially if they already have a tower with an HF antenna on the top. The existing tower can be used as a top-loaded vertical, using shunt feed and a ground radial system. A system like this is shown in **Fig 39B**.

For those who live in an area with tall trees, it may be possible to install a support rope between two trees, or between a tree and an existing tower. (Under no circumstances should you use an active utility pole!) The vertical portion of the antenna can be a wire suspended from the support line to ground, as shown in Fig 39C. If top loading is needed, some or all of the support line can be made part of the antenna.

Your local utility company will periodically have older power poles that they no longer wish to keep in service. These are sometimes available at little or no expense. If you see a power line under reconstruction or repair in your area you might stop and speak with the crew foreman. Sometimes they will have removed older poles they will not use again and will have to haul them back to their shop for disposal. Your offer for local "disposal" may well be accepted. Such a pole can be used in conjunction with a tubing or whip extension such as that shown in Fig 39A. Power poles are not your only option. In some areas of the US, such as the southeast or northwest, tall poles made directly from small conifers are available.

Freestanding (unguyed) flagpoles and roadway illumination standards are available in heights exceeding 100 feet. These are made of fiberglass, aluminum or galvanized steel. All of these are candidates for verticals. Flagpole suppliers are listed under "Flags and Banners" in your Yellow Pages. For lighting standards (lamp posts), you can contact a local electrical hardware distributor. Like a wooden pole, a fiberglass flagpole does not require a base insulator, but metal poles do. Guy wires will be needed.

One option to avoid the use of guys and a base insulator is to mount the pole directly into the ground as originally intended and then use shunt feed. If you want to keep the pole grounded but would like to use elevated radials, you can attach a cage of wires (four to six) at the top as shown in Fig 39D. The cage surrounds the pole and allows the pole (or tower for that matter) to be grounded while allowing elevated radials to be used. The use of a cage of wires surrounding the pole or tower is a very good way to increase the effective diameter. This reduces the Q of the antenna, thereby increasing the bandwidth. It can also reduce the conductor loss, especially if the pole is galvanized steel, which is not a very good RF conductor.

Aluminum irrigation tubing, which comes in diameters of 3 and 4 inches and in lengths of 20 to 40 feet, is widely available in rural areas. One or two lengths of tubing con-

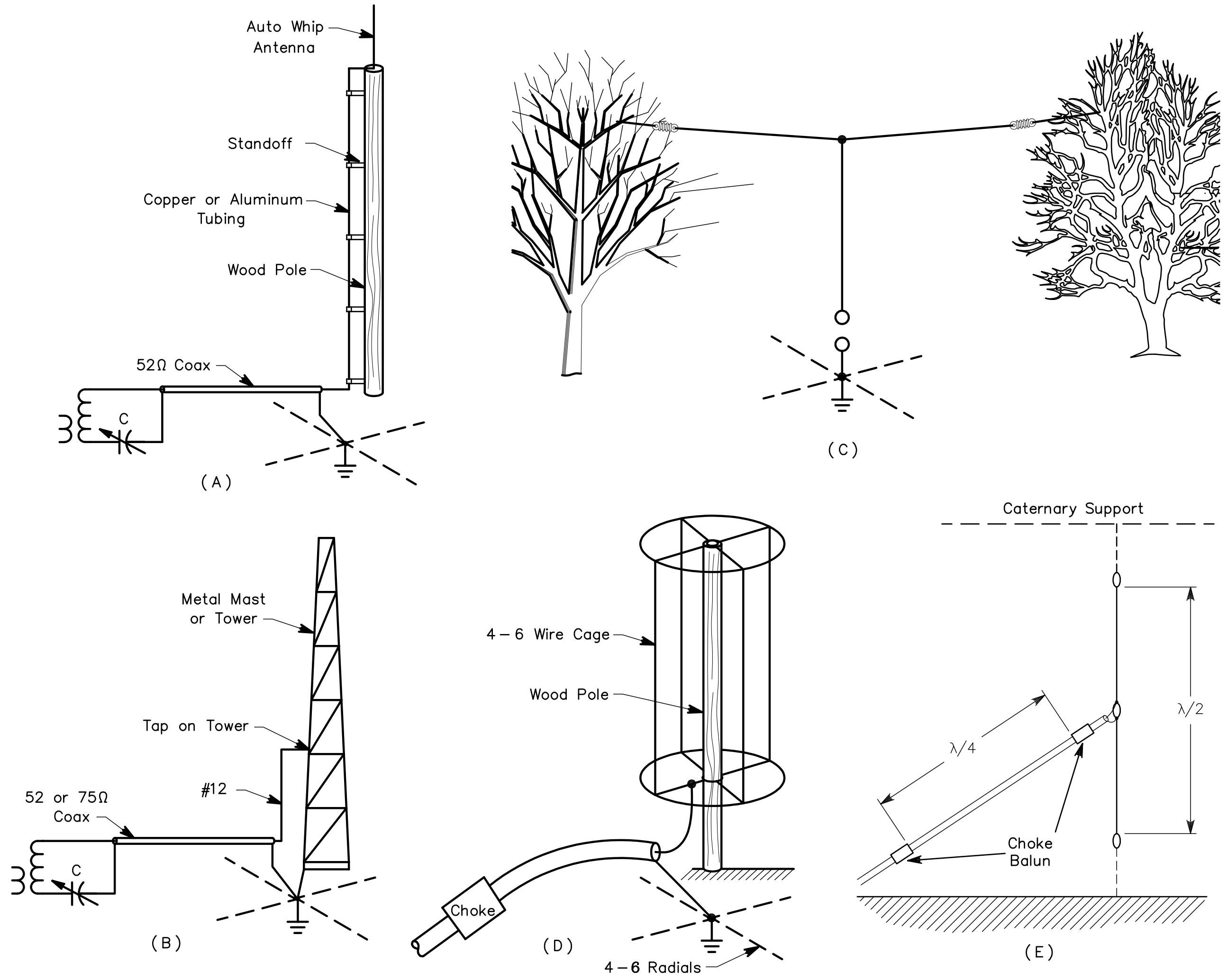

Fig 39—Vertical antennas are effective for 3.5- or 7-MHz work. The λ/4 antenna shown at A is fed directly with 50-Ω coaxial line, and the resulting SWR is usually less than 1.5 to 1, depending on the ground resistance. If a grounded antenna is used as at B, the antenna can be shunt fed with either 50- or 75-Ω coaxial line. The tap for best match and the value of C will have to be found by experiment. The line running up the side of the antenna should be spaced 6 to 12 inches from the antenna. If tall trees are available the antenna can be supported from a line suspended between the trees, as shown in C. If the vertical section is not long enough then the horizontal support section can be made of wire and act as top loading. A pole or even a grounded tower can be used with elevated radials if a cage of four to six wires is provided as shown in D. The cage surrounds the pole which may be wood or a grounded conductor.

nected together can make a very good vertical when guyed with non-conducting line. It is also very lightweight and relatively easy to erect. A variety of TV masts are available which can also be used for verticals.

1.8 TO 3.5-MHz VERTICAL USING AN EXISTING TOWER

A tower can be used as a vertical antenna, provided that a good ground system is available. The shunt-fed tower is at its best on 1.8 MHz, where a full λ/4 vertical antenna is rarely possible. Almost any tower height can be used. If the beam structure provides some top loading, so much the better, but anything can be made to radiate—if it is fed properly. W5RTQ (now K6SE) uses a self-supporting, aluminum, crank-up, tilt-over tower, with a TH6DXX tribander mounted at 70 feet. Measurements showed that the entire structure has about the same properties as a 125-foot vertical. It thus works quite well as an antenna on 1.8 and 3.5 MHz for DX work requiring low-angle radiation.

Preparing the Structure

Usually some work on the tower system must be done before shunt-feeding is tried. If present, metallic guys should be broken up with insulators. They can be made to simulate top loading, if needed, by judicious placement of the first insulators. Don't overdo it; there is no need to "tune the radiator to resonance" in this way since a shunt feed is employed. If the tower is fastened to a house at a point more than about one-fourth of the height of the tower, it may be desirable to insulate the tower from the building. Plexiglas sheet, ¼-inch or more thick, can be bent to any desired shape for this purpose, if it is heated in an oven and bent while hot.

All cables should be taped tightly to the tower, on the inside, and run down to the ground level. It is not necessary to bond shielded cables to the tower electrically, but there should be no exceptions to the down-to-the-ground rule.

A good system of buried radials is very desirable. The ideal would be 120 radials, each 250 feet long, but fewer and shorter ones must often suffice. You can lay them around corners of houses, along fences or sidewalks, wherever they can be put a few inches under the surface, or even on the earth's surface. Aluminum clothesline wire may be used extensively in areas where it will not be subject to corrosion. Neoprene-covered aluminum wire will be better in highly acid soils. Contact with the soil is not important. Deep-driven ground rods and connection to underground copper water pipes may be helpful, if available, especially to provide some protection from lightning.

Installing the Shunt Feed

Principal details of the shunt-fed tower for 1.8 and 3.5 MHz are shown in **Fig 40**. Rigid rod or tubing can be used for the feed portion, but heavy gauge aluminum or copper wire is easier to work with. Flexible stranded #8 copper wire is used at W5RTQ (now K6SE) for the 1.8-MHz feed, because when the tower is cranked down, the feed wire must come down with it. Connection is made at the top, 68 feet, through a 4-foot length of aluminum tubing clamped to the top of the tower, horizontally. The wire is clamped to the tubing at the outer end, and runs down vertically through standoff insulators. These are made by fitting 12-inch lengths of PVC plastic water pipe over 3-foot lengths of aluminum tubing. These are clamped to the tower at 15- to 20-foot intervals, with the bottom clamp about 3 feet above ground. These lengths allow for adjustment of the tower-to-wire spacing over a range of about 12 to 36 inches, for impedance matching.

The gamma-match capacitor for 1.8 MHz is a 250-pF variable with about ⅙-inch plate spacing. This is adequate for power levels up to about 200 W. A large transmitting or a vacuum-variable capacitor should be used for high-power applications.

Tuning Procedure

The 1.8-MHz feed wire should be connected to the top of the structure if it is 75 feet tall or less. Mount the standoff insulators so as to have a spacing of about 24 inches between wire and tower. Pull the wire taut and clamp it in place at the bottom insulator. Leave a little slack below to permit adjustment of the wire spacing, if necessary.

Adjust the series capacitor in the 1.8-MHz line for minimum reflected power, as indicated on an SWR meter connected between the coax and the connector on the capacitor housing. Make this adjustment at a frequency near the middle of your expected operating range. If a high SWR is indicated, try moving the wire closer to the tower. Just the

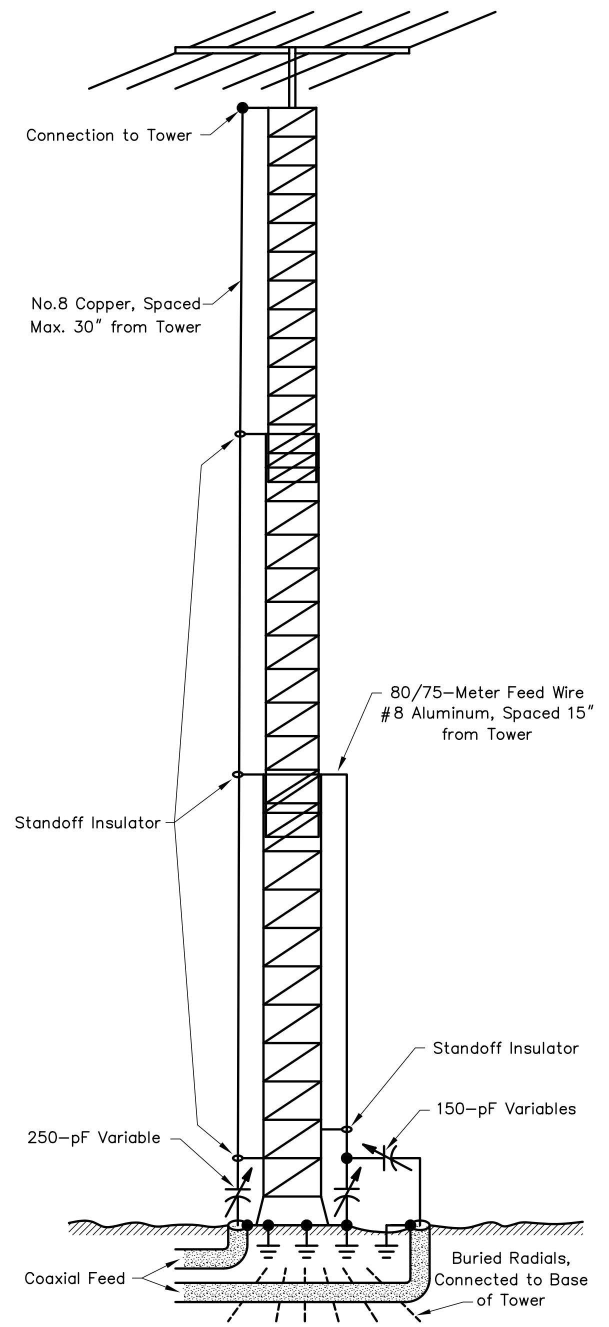

Fig 40—Principal details of the shunt-fed tower at W5RTQ (now K6SE). The 1.8-MHz feed, left side, connects to the top of the tower through a horizontal arm of 1-inch diameter aluminum tubing. The other arms have standoff insulators at their outer ends, made of 1-foot lengths of plastic water pipe. The connection for 3.5-4 MHz, right, is made similarly, at 28 feet, but two variable capacitors are used to permit adjustment of matching with large changes in frequency.

lower part of the wire need be moved for an indication as to whether reduced spacing is needed. If the SWR drops, move all insulators closer to the tower, and try again.

If the SWR goes up, increase the spacing. There will be a practical range of about 12 to 36 inches. If going down to 12 inches does not give a low SWR, try connecting the top a bit farther down the tower. If wide spacing does not make it, the omega match shown for 3.5-MHz work should be tried. No adjustment of spacing is needed with the latter arrangement, which may be necessary with short towers or installations having little or no top loading.

The two-capacitor arrangement in the omega match is also useful for working in more than one 25-kHz segment of the 1.8-MHz band. Tune up on the highest frequency, say 1990 kHz, using the single capacitor, making the settings of wire spacing and connection point permanent for this frequency. To move to the lower frequency, say 1810 kHz, connect the second capacitor into the circuit and adjust it for the new frequency. Switching the second capacitor in and out then allows changing from one segment to the other, with no more than a slight retuning of the first capacitor.

SIMPLE, EFFECTIVE, ELEVATED GROUND-PLANE ANTENNAS

This section describes a simple and effective means of using a grounded tower, with or without top-mounted antennas, as an elevated ground-plane antenna for 80 and 160 meters. It first appeared in a June 1994 *QST* article by Thomas Russell, N4KG.

From Sloper to Vertical

Recall the quarter-wavelength sloper, also known as the *half sloper*. [The half sloper is covered later in this chapter in more detail.—*Ed.*] It consists of an isolated quarter wavelength of wire, sloping from an elevated feed point on a grounded tower. Best results are usually obtained when the feed point is somewhere below a top-mounted Yagi antenna. You feed a sloper by attaching the center conductor of a coaxial cable to the wire and the braid of the cable to the tower leg. Now, imagine four (or more) slopers, but instead of feeding each individually, connect them together to the center conductor of a single feed line. Voilà! Instant elevated ground plane.

Now, all you need to do is determine how to tune the antenna to resonance. With no antennas on the top of the tower, the tower can be thought of as a fat conductor and should be approximately 4% shorter than a quarter wavelength in free space. Calculate this length and attach four insulated quarter-wavelength radials at this distance from the top of the tower. For 80 meters, a feed point 65 feet below the top of an unloaded tower is called for. The tower guys must be broken up with insulators for all such installations. For 160 meters, 130 feet of tower above the feed point is needed.

What can be done with a typical grounded-tower-and-Yagi installation? A top-mounted Yagi acts as a large capacitance hat, top loading the tower. Fortunately, top loading is the most efficient means of loading a vertical antenna.

The examples in **Table 3** should give us an idea of how much top loading might be expected from typical amateur antennas. The values listed in the *Equivalent Loading* column tell us the approximate vertical height replaced by the antennas listed in a top-loaded vertical antenna. To arrive at the remaining amount of tower needed for resonance, subtract these numbers from the non-loaded tower height needed for resonance. Note that for all but the 10-meter antennas, the equivalent loading equals or exceeds a quarter wavelength on 40 meters. For typical HF Yagis, this method is best used only on 80 and 160 meters.

Construction Examples

Consider this example: A TH7 triband Yagi mounted on a 40-foot tower. The TH7 has approximately the same overall dimensions as a full-sized 3-element 20-meter beam, but has more interlaced elements. Its equivalent loading is estimated to be 40 feet. At 3.6 MHz, 65 feet of tower is needed without loading. Subtracting 40 feet of equivalent loading, the feed point should be 25 feet below the TH7 antenna.

Ten quarter-wavelength (65-foot) radials were run from a nylon rope tied between tower legs at the 15-foot level, to various supports 10 feet high. Nylon cord was tied to the insulated, stranded, #18 wire, without using insulators. The radials are all connected together and to the center of an exact half wavelength (at 3.6 MHz) of RG-213 coax, which will repeat the antenna feed impedance at the other end. **Fig 41** is a drawing of the installation. The author used a Hewlett-Packard low-frequency impedance analyzer to measure the input impedance across the 80-meter band. An exact resonance (zero reactance) was seen at 3.6 MHz, just as predicted. The radiation resistance was found to be 17 Ω. The next question is, how to feed and match the antenna.

One good approach to 80-meter antennas is to tune them to the low end of the band, use a low-loss transmission line, and switch an antenna tuner in line for operation in the higher portions of the band. With a 50-Ω line, the 17-Ω radiation resistance represents a 3:1 SWR, meaning that an antenna tuner should be in-line for all frequencies. For short runs, it

Table 3
Effective Loading of Common Yagi Antennas

Antenna	*Boom Length (feet)*	*S (area, ft^2)*	*Equivalent Loading (feet)*
3L 20	24	768	39
5L 15	26	624	35
4L 15	20	480	31
3L 15	16	384	28
5L 10	24	384	28
4L 10	18	288	24
3L 10	12	192	20
TH7	24	—	40 (estimated)
TH3	14	—	27 (estimated)

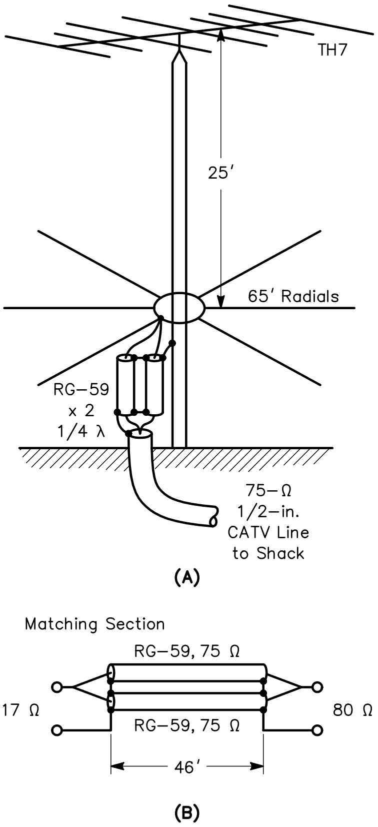

Fig 41—At A, an 80-meter top-loaded, reverse-fed elevated ground plane, using a 40-foot tower carrying a TH7 triband Yagi antenna. At B, dimensions of the 3.6-MHz matching network, made from RG-59.

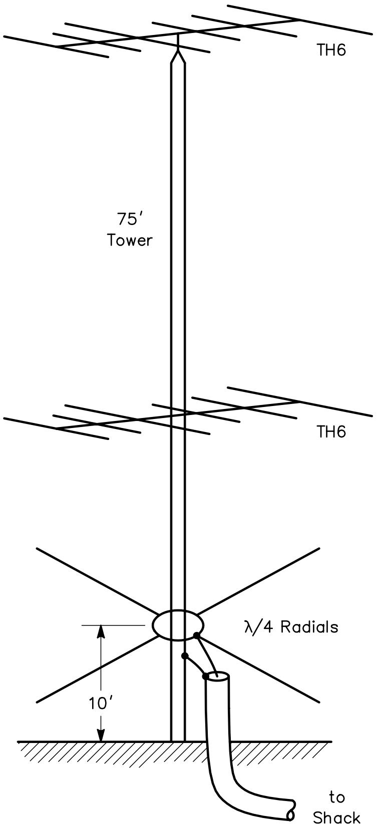

Fig 42—A 160-meter antenna using a 75-foot tower carrying stacked triband Yagis.

would be permissible to use RG-8 or RG-213 directly to the tuner. If you have a plentiful supply of low-loss 75-Ω CATV rigid coax, you can take another approach.

Make a quarter-wave (70 feet × 0.66 velocity factor = 46 feet) 37-Ω matching line by paralleling two pieces of RG-59 and connecting them between the feed point and a run of the rigid coax to the transmitter. The magic of quarter-wave matching transformers is that the input impedance (R_i) and output impedance (R_o) are related by:

$$Z_0^2 = R_i \times R_o \qquad \text{(Eq 2)}$$

For $R_i = 17\ \Omega$ and $Z_0 = 37\ \Omega$, $R_o = 80\ \Omega$, an almost perfect match for the 75-Ω CATV coax. The resulting 1.6:1 SWR at the transmitter is good enough for CW operation without a tuner.

160-Meter Operation

On the 160-meter band, a resonant quarter-wavelength requires 130 feet of tower above the radials. That's a pretty tall order. Subtracting 40 feet of top loading for a 3-element 20-meter or TH7 antenna brings us to a more reasonable 90 feet above the radials. Additional top loading in the form of more antennas will reduce that even more.

Another installation, using stacked TH6s on a 75-foot tower, is shown in **Fig 42**. The radials are 10 feet off the ground.

PHASED VERTICALS

Two or more vertical antennas spaced apart can be operated as a single antenna system to obtain additional gain and a directional pattern. There is an extensive discussion of phased arrays in Chapter 8, Multielement Arrays. Much of the material in Chapter 8 is useful for low-band antennas.

The Half-Square Antenna

The *half-square* antenna is a very simple form of vertical two-element phased array that can be very effective on the low bands. The following section was originally presented in *The ARRL Antenna Compendium Vol 5*, by Rudy Severns, N6LF.

A simple modification to a standard dipole is to add two λ/4 vertical wires, one at each end, as shown in **Fig 43**. This makes a *half-square antenna*. The antenna can be fed at one corner (low-impedance, current fed) or at the lower end of one of the vertical wires (high-impedance, voltage fed).

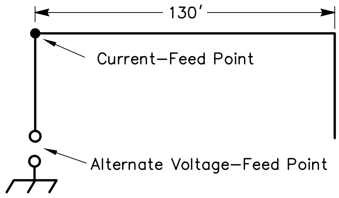

Fig 43—Typical 80-meter half-square, with λ/4-high vertical legs and a λ/2-long horizontal leg. The antenna may be fed at the bottom or at a corner. When fed at a corner, the feed point is a low-impedance, current-feed. When fed at the bottom of one of the wires against a small ground counterpoise, the feed point is a high-impedance, voltage-feed.

Other feed arrangements are also possible.

The "classical" dimensions for this antenna are λ/2 (131 feet at 3.75 MHz) for the top wire and λ/4 (65.5 feet) for the vertical wires. However, there is nothing sacred about these dimensions! They can vary over a wide range and still obtain nearly the same performance.

This antenna is two λ/4 verticals, spaced λ/2, fed in-phase by the top wire. The current maximums are at the top corners. The theoretical gain over a single vertical is 3.8 dB. An important advantage of this antenna is that it does not require the extensive ground system and feed arrangements that a conventional pair of phased λ/4 verticals would.

Comparison to a Dipole

In the past, one of the things that has turned off potential users of the half-square on 80 and 160 meters is the perceived need for λ/4 vertical sections. This forces the height to be > 65 feet on 80 meters and > 130 feet on 160 meters. That's not really a problem. If you don't have the height there are several things you can do. For example, just fold the ends in, as shown in **Fig 44**. This compromises the performance surprisingly little.

It is helpful to compare the examples given in Figs 43 and 44 to dipoles at the same height. Two heights, 40 and 80 feet, and average, very good and sea water grounds, were used for this comparison. It is also assumed that the lower end of the vertical wires had to be a minimum of 5 feet above ground.

At 40 feet the half-square is really mangled, with only 35-foot long (≈ λ/8) vertical sections. The elevation-plane comparison between this antenna and a dipole of the same height is shown in **Fig 45**. Over average ground the half-square is superior below 32° and at 15° is almost 5 dB better. That is a worthwhile improvement. If you have very good soil conductivity, like parts of the lower Midwest and South, then the half-square will be superior below 38° and at 15° will be nearly 8 dB better. For those fortunate few with saltwater frontal property the advantage at 15° is 11 dB! Notice also that above 35°, the response drops off rapidly. This is great for DX but is not good for local work.

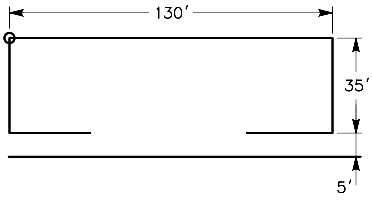

Fig 44—An 80-meter half-square configured for 40-foot high supports. The ends have been bent inward to reresonate the antenna. The performance is compromised surprisingly little.

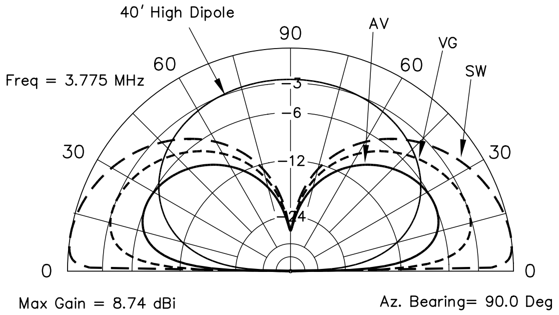

Fig 45—Comparison of 80-meter elevation response of 40-foot high, horizontally polarized dipole over average ground and a 40-foot high, vertically polarized half-square, over three types of ground: average (conductivity σ = 5 mS/m, dielectric constant ε = 13), very good (σ = 30 mS/m, ε = 20) and salt water (σ = 5000 mS/m, ε = 80). The quality of the ground clearly has a profound effect on the low-angle performance of the half-square. Even over average ground, the half-square outperforms the low dipole below about 32°.

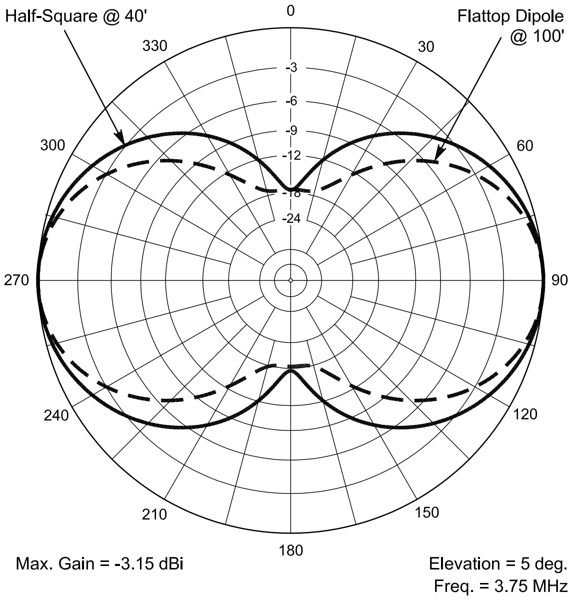

Fig 46—80-meter azimuth patterns for shortened half-square antenna (solid line) shown in Fig 44, compared with flattop dipole (dashed line) at 100 feet height. Average ground is assumed for these cases.

Fig 46 shows the azimuthal-plane pattern for the 80-meter half-square antenna in Fig 44, but this time compared with the response of a flattop horizontal dipole that is 100 feet high. These comparisons are for average ground and are for an elevation angle of 5°. The message here is that the lower your dipole and the better your ground, the more you have to gain by switching from a dipole to a half-square. The half-square antenna looks like a good bet for DXing.

Changing the Shape of the Half Square

Just how flexible is the shape? There are several common distortions of practical importance. Some have very little effect but a few are fatal to the gain. Suppose you have either more height and less width than called for in the standard version or more width and less height, as shown in **Fig 47A**.

The effect on gain from this type of dimensional variation is given in **Table 4**. For a top length (L_T) varying between 110 and 150 feet, where the vertical wire lengths (L_V) readjusted to resonate the antenna, the gain changes only by 0.6 dB. For a 1-dB change the range of L_T is 100 to 155 feet, a pretty wide range.

Another variation results if we vary the length of the horizontal top wire and readjust the vertical wires for resonance, while keeping the top at a constant height. See Fig 47B. **Table 5** shows the effect of this variation on the peak gain. For a range of L_T = 110 to 145 feet, the gain changes only 0.65 dB.

The effect of bending the ends into a V shape, as shown in Fig 47C, is given in **Table 6**. The bottom of the antenna is kept at a height of 5 feet and the top height (H) is either 40 or 60 feet. Even this gross deformation has only a relatively small effect on the gain. Sloping the ends outward as shown in Fig 47D and varying the top length also has only a small effect on the gain. While this is good news because it allows you dimension the antenna to fit different QTHs, not all distortions are so benign.

Table 4

Variation in Gain with Change in Horizontal Length, with Vertical Height Readjusted for Resonance (see Fig 47A)

L_T (feet)	*L_V (feet)*	*Gain (dBi)*
100	85.4	2.65
110	79.5	3.15
120	73.7	3.55
130	67.8	3.75
140	61.8	3.65
150	56	3.05
155	53	2.65

Table 5

Variation in Gain with Change in Horizontal Length, with Vertical Length Readjusted for Resonance, but Horizontal Wire Kept at Constant Height (see Fig 47B)

L_T (feet)	*L_V (feet)*	*Gain (dBi)*
110	78.7	3.15
120	73.9	3.55
130	68	3.75
140	63	3.35
145	60.7	3.05

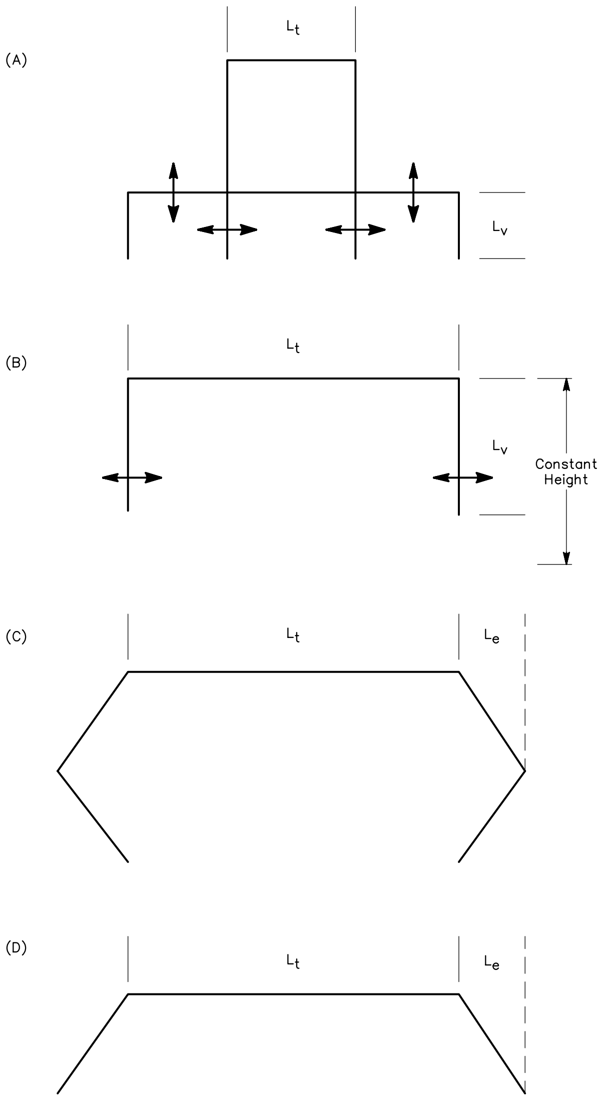

Fig 47—Varying the horizontal and vertical lengths of a half-square. At A, both the horizontal and vertical legs are varied, while keeping the antenna resonant. At B, the height of the horizontal wire is kept constant, while its length and that of the vertical legs is varied to keep the antenna resonant. At C, the length of the horizontal wire is varied and the legs are bent inwards in the shape of "vees." At D, the ends are sloped outward and the length of the flattop portion is varied. All these symmetrical forms of distortion of the basic half-square shape result in small performance losses. At E, a "halfwave vertical dipole" (HVD) with the feed coax isolated with common-mode choke baluns to keep RF current off the coax shield.

Table 6

Gain for Half-Square Antenna, Where Ends Are Bent Into V-Shape (see Fig 47 C)

Height ⇒	*H=40 feet*	*H=40 feet*	*H=60 feet*	*H=60 feet*
L_T (feet)	*L_e (feet)*	*Gain (dBi)*	*L_e (feet)*	*Gain (dBi)*
40	57.6	3.25	52.0	2.75
60	51.4	3.75	45.4	3.35
80	45.2	3.95	76.4	3.65
100	38.6	3.75	61.4	3.85
120	31.7	3.05	44.4	3.65
140	—	—	23	3.05

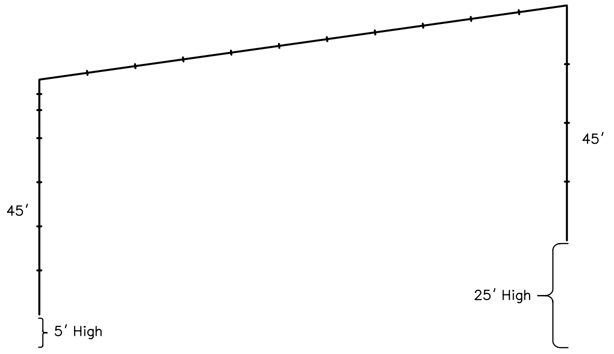

Fig 48—An asymmetrical distortion of the half-square antenna, where the bottom of one leg is purposely made 20 feet higher than the other. This type of distortion does affect the pattern!

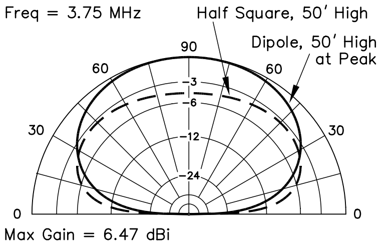

Fig 49—Elevation pattern for the asymmetrical half-square shown in Fig 48, compared with pattern for a 50-foot high dipole. This is over average ground, with a conductivity of 5 mS/m and a dielectric constant of 13. Note that the zenith-angle null has filled in and the peak gain is lower compared to conventional half-square shown in Fig 43 over the same kind of ground.

Suppose the two ends are not of the same height, as illustrated in **Fig 48**, where one end of the half-square is 20 feet higher than the other. The elevation-plane radiation pattern for this antenna is shown in **Fig 49** compared to a dipole at 50 feet. This type of distortion does affect the pattern. The gain drops somewhat and the zenith null goes away. The nulls off the end of the antenna also go away, so that there is some end-fire radiation. In this example the difference in height is fairly extreme at 20 feet. Small differences of 1 to 5 feet do not affect the pattern seriously.

If the top height is the same at both ends but the length of the vertical wires is not the same, then a similar pattern distortion can occur. The antenna is very tolerant of symmetrical distortions but it is much less accepting of asymmetrical distortion.

What if the length of the wires is such that the antenna is not resonant? Depending on the feed arrangement, that may or may not matter. We will look at that issue later on, in the section on patterns versus frequency. The half-square antenna, like the dipole, is very flexible in its proportions.

Half-Square Feed-Point Impedance

There are many different ways to feed the half-square. Traditionally the antenna has been fed either at the end of one of the vertical sections, against ground, or at one of the upper corners as shown in Fig 43.

For voltage feed at the bottom against ground, the impedance is very high, on the order of several thousand ohms. For current feed at a corner, the impedance is much lower and is usually close to 50 Ω. This is very convenient for direct feed with coax.

The half-square is a relatively high-Q antenna (Q ≈ 17). **Fig 50** shows the SWR variation with frequency for this feed arrangement. An 80-meter dipole is not particularly wideband either, but a dipole will have less extreme variation in SWR than the half-square.

Patterns Versus Frequency

Impedance is not the only issue when defining the bandwidth of an antenna. The effect on the radiation pattern of changing frequency is also a concern. For a voltage-fed half-square, the current distribution changes with frequency. For an antenna resonant near 3.75 MHz, the current distribution is nearly symmetrical. However, above and below resonance the current distribution increasingly becomes asymmetrical. In effect, the open end of the antenna is constrained to be a voltage maximum but the feed point can behave less as a voltage point and more like a current maxima. This allows the current distribution to become asymmetrical.

The effect is to reduce the gain by –0.4 dB at 3.5 MHz and by –0.6 dB at 4 MHz. The depth of the zenith null is reduced from –20 dB to –10 dB. The side nulls are also reduced. Note that this is exactly what happened when the antenna was made physically asymmetrical. Whether the asymmetry is due to current distribution or mechanical arrangements, the antenna pattern will suffer.

When current-feed at a corner is used, the asymmetry introduced by off-resonance operation is much less, since both ends of the antenna are open circuits and constrained to be voltage maximums. The resulting gain reduction is only –0.1 dB. It is interesting that the sensitivity of the pattern to changing frequency depends on the feed scheme used.

Of more concern for corner feed is the effect of the transmission line. The usual instruction is to simply feed

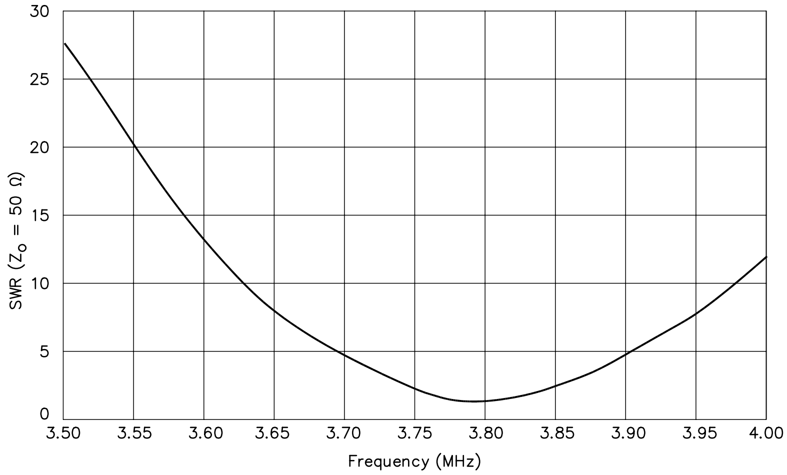

Fig 50—Variation of SWR with frequency for current-fed half-square antenna. The SWR bandwidth is quite narrow.

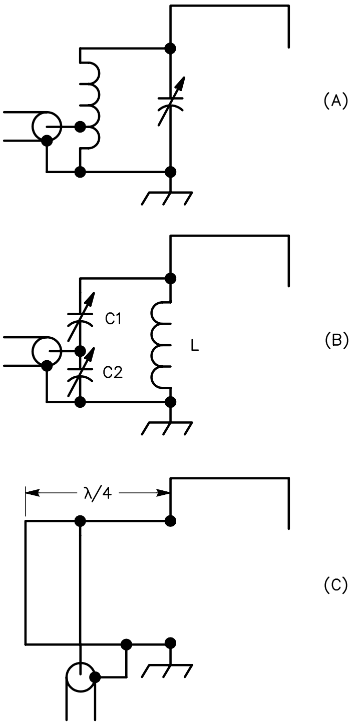

Fig 51—Typical matching networks used for voltage-feeding a half-square antenna.

the antenna using coax, with the shield connected to vertical wire and the center conductor to the top wire. Since the shield of the coax is a conductor, more or less parallel with the radiator, and is in the immediate field of the antenna, you might expect the pattern to be seriously distorted by this practice. This arrangement seems to have very little effect on the pattern. The greatest effect is when the feed line length was near a multiple of λ/2. Such lengths should be avoided.

Of course, you may use a choke balun at the feed point if you desire. This might reduce the coupling to the feed line even further but it doesn't appear to be worth the trouble. In fact, if you use an antenna tuner in the shack to operate away from resonance with a very high SWR on the transmission line, a balun at the feed point would take a beating.

Voltage-Feed at One End of Antenna: Matching Schemes

Several straightforward means are available for narrow-band matching. However, broadband matching over the full 80-meter band is much more challenging. Voltage feed with a parallel-resonant circuit and a modest local ground, as shown in **Fig 51**, is the traditional matching scheme for this antenna. Matching is achieved by resonating the circuit at the desired frequency and tapping down on the inductor in Fig 51A or using a capacitive divider (Fig 51B). It is also possible to use a λ/4 transmission-line matching scheme, as shown in Fig 51C.

If the matching network shown in Fig 51B is used, typical values for the components would be: L = 15 µH, C1 = 125 pF and C2 = 855 pF. At any single point the SWR can be made very close to 1:1 but the bandwidth for SWR < 2:1 will be very narrow at <100 kHz. Altering the L-C ratio doesn't make very much difference. The half-square antenna has a well-earned reputation for being narrow-band.

Short Vertical Antennas

On the lower frequencies it becomes increasingly difficult to accommodate a full λ/4 vertical height and full-sized λ/4 radials, or even worse, a full-sized half-wave vertical dipole (HVD). In fact, it is not absolutely necessary to make the antenna full size, whether it is an HVD, a grounded monopole antenna or a ground-plane type of monopole antenna. The size of the antenna can be reduced by half or even more and still retain high efficiency and the desired radiation pattern. This requires careful design, however. And if high efficiency is maintained, the operating bandwidth of the shortened antenna will be reduced because the shortened antenna will have a higher Q.

This translates into a more rapid increase of reactance away from resonance. The effect can be mitigated to some extent by using larger-diameter conductors. Even doing this however, bandwidth will be a problem, particularly on the 3.5 to 4-MHz band, which is very wide in proportion to the center frequency.

If we take a vertical monopole with a diameter of 2 inches and a frequency of 3.525 MHz and progressively shorten it from λ/4 in length, the feed-point impedance and efficiency (using an inductor at the base to tune out the capacitive reactance) will vary as shown in **Table 7**. In this example perfect ground and conductor are assumed. Real ground will not make a great difference in the impedance but will introduce ground loss, which will reduce the efficiency further. Conductor loss will also reduce efficiency. In general, higher R_R will result in better efficiency.

The important point of Table 7 is the drastic reduction in radiation resistance R_R as the antenna gets shorter. This combined with the increasing loss resistance of the inductor (R_L) used to tune out the increasing base reactance (X_C) reduces the efficiency.

BASE LOADING A SHORT VERTICAL ANTENNA

The base of the antenna is a convenient point at which to add a loading inductor, but it is usually not the lowest loss point at which an inductor, of a given Q, could be placed. There is an extensive discussion of the optimum location of the loading in a short vertical as a function of ground loss and inductor Q in Chapter 16 for mobile antennas, which by necessity are electrically and physically short. This information should be reviewed before using inductive loading.

On the accompanying CD-ROM is a copy of the program *MOBILE.EXE*. This is an excellent tool for designing short, inductively loaded antennas. In most cases, where top loading (discussed below) is not used, the optimum point is near or a little above the middle of the vertical section. Moving the loading coil from the base to the middle of the vertical antenna can make an important difference, increasing R_R and reducing the inductor loss. For example, in an antenna operating at 3.525 MHz, if we make L_1 = 34.9 feet (0.125 λ) the amount of loading inductor placed at the center is 25.2 µH. This resonates the antenna. In this configuration R_R will increase from 6.8 Ω (base loading) to 13.5 Ω (center loading). This substantially increases the efficiency of the antenna, depending on the ground loss and conductor resistances.

Instead of a lumped inductance being inserted at some point in the antenna, it is also possible to use "continuous loading," where the entire radiator is wound as a small diameter coil. The effect is to distribute the inductive loading all along the radiator. In this version of inductive loading the coil is the radiator. An example of a short vertical using this principle is given later in this chapter.

OTHER WAYS OF LOADING A SHORT ANTENNA FOR RESONANCE

Inductive loading is not the only, or even the best, way to compensate for reduced antenna height. *Capacitive top loading* can also be used as indicated in **Fig 52** to bring a vertical monopole to resonance. **Table 8** gives information on a shortened 3.525-MHz vertical using top loading. The vertical portion (L_1) is made from 2-inch tubing. The top loading is also 2-inch tubing extending across the top like a T. The length of the top loading T (±L_2) is adjusted to resonate the

Table 7

Effect of Shortening a Vertical Radiator Below λ/4 Using Inductive Base Loading.

Frequency is 3.525 MHz and for the Inductor Q_L = 200. Ground and conductor losses are omitted.

Length (feet)	*Length (λ)*	*R_R (Ω)*	*X_C (Ω)*	*R_L (Ω)*	*Efficiency (%)*	*Loss (dB)*
14	0.050	0.96	–761	3.8	20	–7.0
20.9	0.075	2.2	–533	2.7	45	–3.5
27.9	0.100	4.2	–395	2.0	68	–1.7
34.9	0.125	6.8	–298	1.5	82	–0.86
41.9	0.150	10.4	–220	1.1	90	–0.44
48.9	0.175	15.1	–153	0.77	95	–0.22
55.8	0.200	21.4	–92	0.46	98	–0.09
62.8	0.225	29.7	–34	0.17	99	–0.02

Table 8

Effect of Shortening a Vertical Using Top Loading

L_1 (feet)	*L_2 (feet)*	*Length (λ)*	*R_R (Ω)*
14.0	48.8	0.050	4.0
20.9	38.6	0.075	8.5
27.9	30.1	0.100	14.0
34.9	22.8	0.125	19.9
41.9	17.3	0.150	25.5
48.9	11.9	0.175	30.4
55.8	7.0	0.200	33.9
62.8	2.4	0.225	35.7

antenna. Again, the ground and the conductors are assumed to be perfect in Table 8.

For a given vertical height, resonating the antenna with top loading results in much higher radiation resistance R_R—2 to 4 times. In addition, the loss associated with the loading element will be smaller. The result is a more efficient antenna for low heights. A comparison of R_R for both capacitive top loading and inductive base loading is given in **Fig 53**. For heights below 0.15 λ the length of the top-loading elements becomes impractical but there are other, potentially more useful, top-loading schemes.

A multiwire system such as the one shown in **Fig 54** has more capacitance than the single-conductor arrangement, and thus does not need to be as long to resonate at a given frequency. This design does, however, require extra supports for the additional wires. Ideally, an arrangement of this sort should be in the form of a cross, but parallel wires separated by several feet give a considerable increase in capacitance

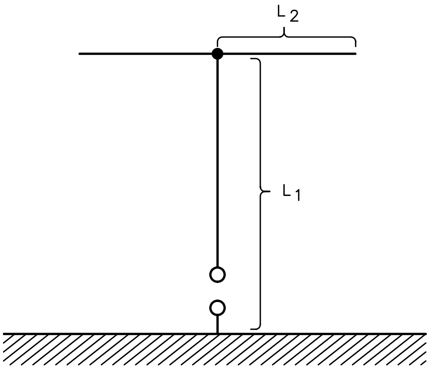

Fig 52—Horizontal wire used to top load a short vertical.

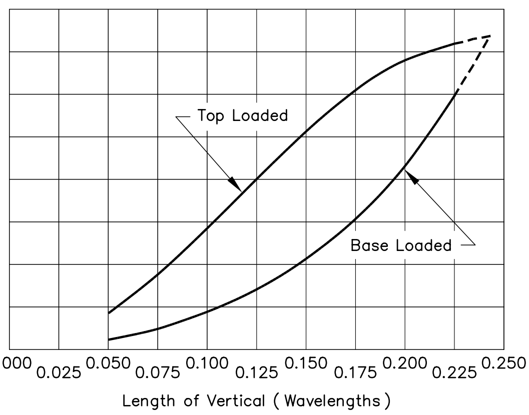

Fig 53—Comparison of top (capacitive) and base (inductive) loading for short verticals. Sufficient loading is used to resonate the antenna.

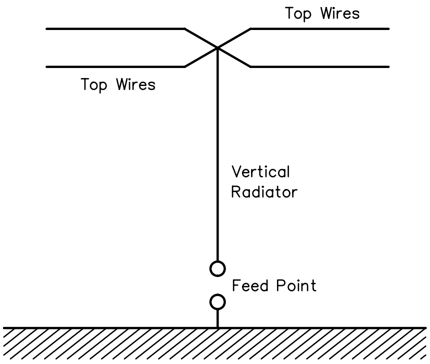

Fig 54—Multiple top wires can increase the effective capacitance substantially. This allows the use of shorter top wires to achieve resonance.

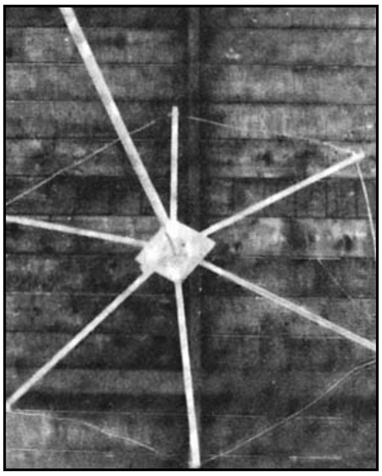

Fig 55—A close-up view of the capacitance hat for a 7-MHz vertical antenna. The ½-in. diameter radial arms terminate in a loop of copper wire.

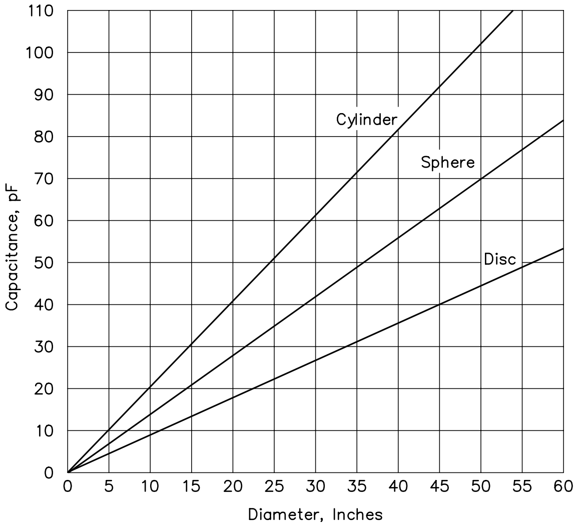

Fig 56—Capacitance of sphere, disc and cylinder as a function of their diameters. The cylinder length is assumed equal to its diameter.

over a single wire.

The top loading can be supplied by a variety of metallic structures large enough to have the necessary self-capacitance. For example, as shown in **Fig 55**, a multi-spoked structure with the ends connected together can be used. One simple way to make a capacitance hat is to take four to six 8-foot fiberglass CB mobile whips, arrange them like spokes in a wagon wheel and connect the ends with a peripheral wire. This arrangement will produce a 16-foot diameter hat that is economical and very durable, even when loaded with ice. Practically any sufficiently large metallic structure can be used for this purpose, but simple geometric forms such as the sphere, cylinder and disc are preferred because of the relative ease with which their capacitance can be calculated.

The capacitance of three geometric forms can be estimated from the curves of **Fig 56** as a function of their size. For the cylinder, the length is specified equal to the diameter. The sphere, disc and cylinder can be constructed from sheet metal, if such construction is feasible, but the capacitance will be practically the same in each if a "skeleton" type of construction with screening or networks of wire or tubing are used.

Finding Capacitance Hat Size

The required size of a capacitance hat may be determined from the following procedure. The information in this section is based on a September 1978 *QST* article by Walter Schulz, K3OQF. The physical length of a shortened antenna can be found from:

$$h_{\text{inches}} = \frac{11808}{F_{\text{MHz}}} \qquad \text{(Eq 3)}$$

where h = length in inches.

Thus, using an example of 7 MHz and a shortened length of 0.167 λ, h = 11808/7 × 0.167 = 282 inches, equivalent to 23.48 feet.

Consider the vertical radiator as an open-ended transmission line, so the impedance and top loading may be determined. The characteristic impedance of a vertical antenna can be found from

$$Z_0 = 60 \left(\ln\left(\frac{4h}{d}\right) - 1 \right) \qquad \text{(Eq 4)}$$

where

ln = natural logarithm

h = length (height) of vertical radiator in inches (as above)

d = diameter of radiator in inches

The vertical radiator for this example has a diameter of 1 inch. Thus, for this example,

$$Z_0 = 60 \left(\ln\left(\frac{4 \times 281}{1}\right) - 1 \right) = 361\ \Omega$$

The capacitive reactance required for the amount of top loading can be found from

$$Z_0 = 60\ \ln\left(\frac{4 \times 281}{1}\right) - 1 = 361\ \Omega \qquad \text{(Eq 5)}$$

where

X = capacitive reactance, ohms

Z_0 = characteristic impedance of antenna (from Eq 4)

θ = amount of electrical loading, degrees.

This value for a 30° hat is 361/tan 30° = 625 Ω. This capacitive reactance may be converted to capacitance with the following equation,

$$C = \frac{10^6}{2\pi f X_C} \qquad \text{(Eq 6)}$$

where

C = capacitance in pF

f = frequency, MHz

X_C = capacitive reactance, ohms (from above).

For this example, the required C = $10^6/(2\ \pi \times 7 \times 625)$ = 36.4 pF, which may be rounded to 36 pF. A disc capacitor is used in this example. The appropriate diameter for 36 pF of hat capacitance can be found from Fig 56. The disc diameter that yields 36 pF of capacitance is 40 inches.

The skeleton disc shown in Fig 55 is fashioned into a wagonwheel configuration. Six 20-inch lengths of ½-inch OD aluminum tubing are used as spokes. Each is connected to the hub at equidistant intervals. The outer ends of the spokes terminate in a loop made of #14 copper wire. Note that the loop increases the hat capacitance slightly, making a better approximation of a solid disc. The addition of this hat at the top of a 23.4-foot radiator makes it quarter-wave resonant at 7 MHz.

After construction, some slight adjustment in the radiator length or the hat size may be required if resonance at a specific frequency is desired. From Fig 53, the radiation resistance of a 0.167-λ high radiator is seen to be about 13 Ω without top loading. With top loading Rr ≈ 25 Ω or almost double.

THE COMPACT VERTICAL DIPOLE

A variation on the HVD (half-wave vertical dipole) theme is the *compact vertical dipole*, or CVD. The CVD uses capacitance-hat loading on each end of a shortened vertical radiator, as shown in **Fig 57C**. Some call this "top hat" and "bottom hat" loading. Les Moxon, G6XN, called this method of loading a shortened antenna simply "end loading." The vertical wire for his 40-meter CVD is 25 feet high, with 15-foot long horizontal loading wires on each side, top and bottom.

K8CH Compact Vertical Dipole

The top loading wires needn't be perpendicular to the vertical radiator, although that is convenient if you construct the antenna from aluminum tubing. K8CH described a wire 30-meter CVD in the 2006 Edition of *The ARRL Handbook.* This used #14 wire throughout, with sloping top loading wires. It was designed to be suspended from a tree at least

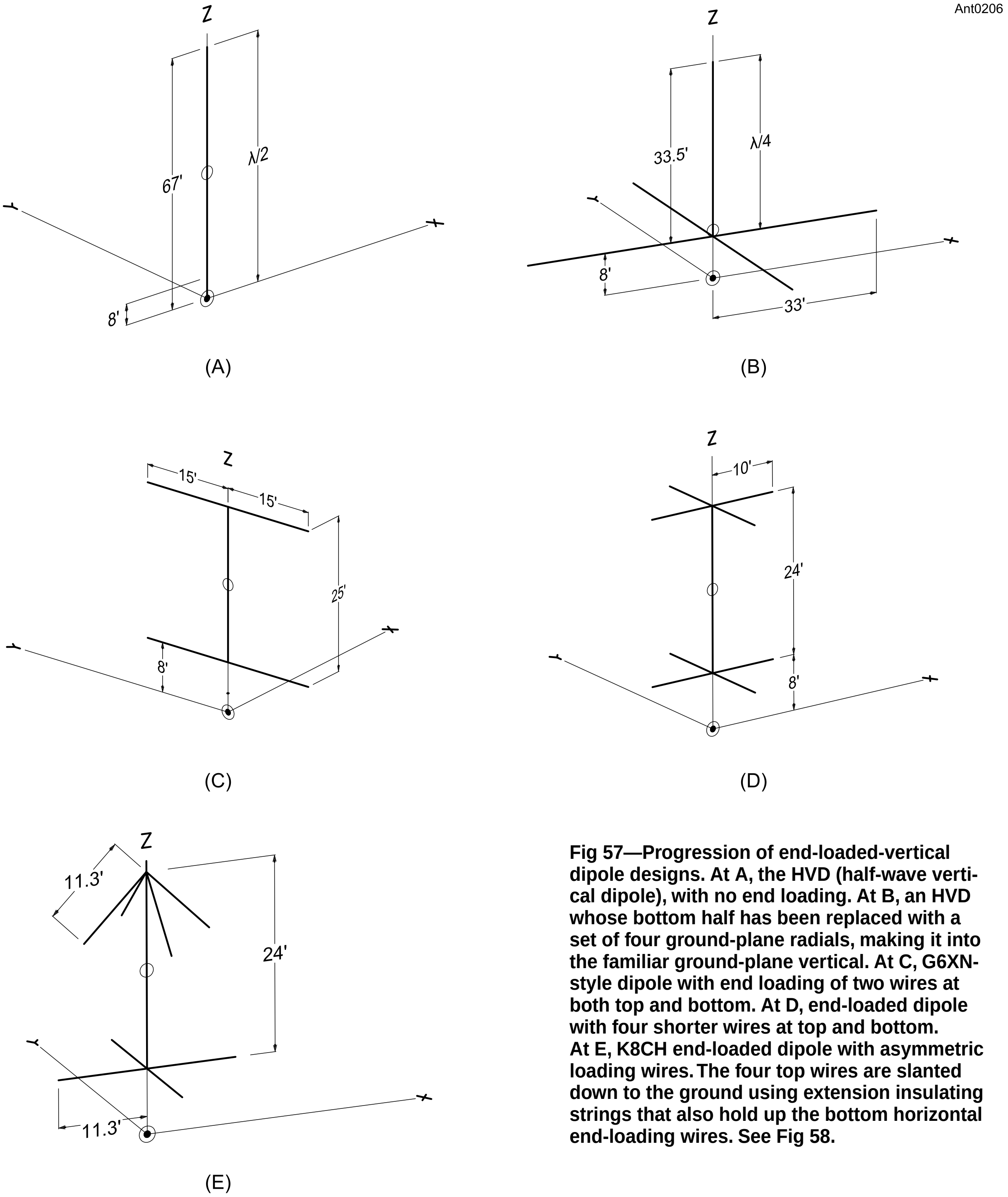

Fig 57—Progression of end-loaded-vertical dipole designs. At A, the HVD (half-wave vertical dipole), with no end loading. At B, an HVD whose bottom half has been replaced with a set of four ground-plane radials, making it into the familiar ground-plane vertical. At C, G6XN-style dipole with end loading of two wires at both top and bottom. At D, end-loaded dipole with four shorter wires at top and bottom.
At E, K8CH end-loaded dipole with asymmetric loading wires. The four top wires are slanted down to the ground using extension insulating strings that also hold up the bottom horizontal end-loading wires. See Fig 58.

32 feet high to keep all wires 8 feet or more above humans and animals. The vertical radiating wire is 24 feet long, and the eight top and bottom end-loading wires are all 5 feet 9 inches long. See Fig 57E and **Table 9**, CVD 1.

The top loading wires slant at an angle of 45º down to the ground, using insulating strings that also support the ends of the bottom loading wires, holding them out so that they are horizontal. See **Fig 58**. There is a small loss of gain because of the “umbrella” shape of the top loading wires, and the 2:1 SWR bandwidth is diminished slightly from the horizontal case. This isn’t a problem on a narrow band like 30 meters.

A 40-meter version of this antenna, CVD 6, also uses a 24-foot long vertical radiator, 8 feet high at its bottom. It uses 11.3-foot long radial wires made of #14 wire, again with the top four slanted down at 45º. This CVD has a 2:1

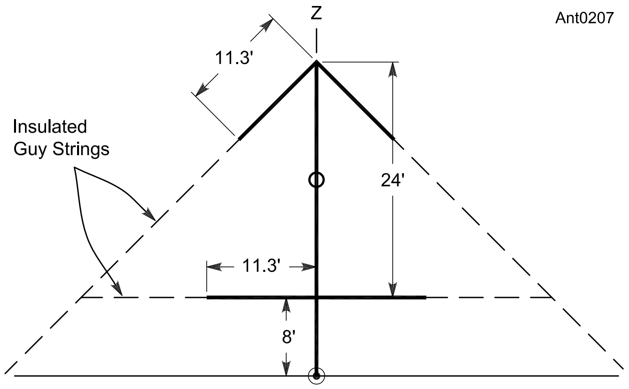

Fig 58—Layout of CVD made using #14 wire suspended from a tree branch.

SWR bandwidth of 250 kHz. It is directly fed with 50-Ω coax with ferrite-core common-mode choke baluns in the middle of the vertical radiator. An additional choke balun is used where the coax reaches ground level in order to knock down common-mode currents that might otherwise radiate onto the coax shield.

You should note the comparison of 40-meter ground-plane vertical antennas in Table 9. GP 1 is for horizontal radials that are 8 feet off the ground, while GP 2 has its radials only 2 feet high. There is some loss in gain because of the proximity of the lossy ground. The 40-meter CVD 1 and CVD 2 cases illustrate the same effect of being close to the lossy ground.

Some of the cases in Table 9 require Center Load coils to bring the antenna to resonance. Where the loading coil inductance is equal to the "Hairpin Coil" inductance, the loading coil also serves as a hairpin matching coil. Where the amount of Hairpin Coil inductance is less than the Load Coil inductance, a match is achieved by tapping the Center Load Coil symmetrically out from the center.

80-Meter CVD

The size of a CVD becomes a real challenge on 80 meters, requiring either very tall support structures or multiple loading methods to keep the vertical radiator to a reasonable length. The CVD 2 design in Table 9 shows a K8CH-style CVD wire antenna whose vertical radiator is 46.5 feet long. It requires a 54.5-foot high tree to keep the bottom end of the vertical radiator 8 feet above ground for safety. Compare this to an HVD that requires a 143-foot support of some sort to keep it 8 feet off the ground at the bottom. The CVD 2 sacrifices some 0.7 dB in gain for this difference in size, and about 75 kHz in 2:1 SWR bandwidth.

The CVD 2 would require retuning when going from

Table 9
Variations on a Vertical Center-Fed Dipole

Name	*Style Fig 57*	*Vertical Length feet*	*Spoke Length feet*	*Min. Ht feet*	*Max. Gain dBi*	*2:1 SWR kHz*	*Hairpin Coil µH*	*Center Load µH*
20 Meters								
GP	B	17.53	16.53	8	0.29	400	—	—
CVD 1	C	13	7.57	8	0.12	625	—	—
CVD 2	D	12	5.1	8	0.00	550	0.68	0.68
CVD 3	E	12.15	5.6	8	–0.01	450	0.5	0.5
30 Meters								
GP	B	24.54	23.14	8	0.04	400	—	—
CVD 1	E	24	5.33	8	–0.2	500	—	—
CVD 2	E	17	7.60	8	–0.36	400	0.82	0.82
40 Meters								
HVD	A	66	—	8	0.13	450	—	—
GP 1	B	35	33	8	–0.12	325	—	—
GP 2	B	34.5	33	2	–0.37	400	—	—
CVD 1	C	25	15	8	–0.42	450	—	—
CVD 2	C	24	15	2	–1.09	400	—	—
CVD 3	D	24	10	8	–0.55	425	—	0.25
CVD 4	D	24	5	8	–0.85	225	—	8.7
CVD 5	D	16	8	8	–1.18	175	0.94	6.8
CVD 6	E	24	11.3	8	–0.59	250	—	—
80 Meters								
HVD	A	135	—	8	0.19	225	—	—
GP	B	65.5	61	8	0.11	200	—	—
CVD 1	C	48.3	30	8	–0.27	200	—	—
CVD 2	E	46.5	21.9	8	–0.50	150	—	—

CW to phone, probably by changing the length of the four bottom horizontal wires equally.

LINEAR LOADING

Another alternative to inductive loading is *linear loading*. This little-understood method of shortening radiators can be applied to almost any antenna configuration—including parasitic arrays. Although commercial antenna manufacturers make use of linear loading in their HF antennas, relatively few hams have used it in their own designs. Linear loading can be used to advantage in many antennas because it introduces relatively little loss, does not degrade directivity patterns, and has low enough Q to allow reasonably good bandwidth. Some examples of linear-loaded antennas are shown in **Fig 59**.

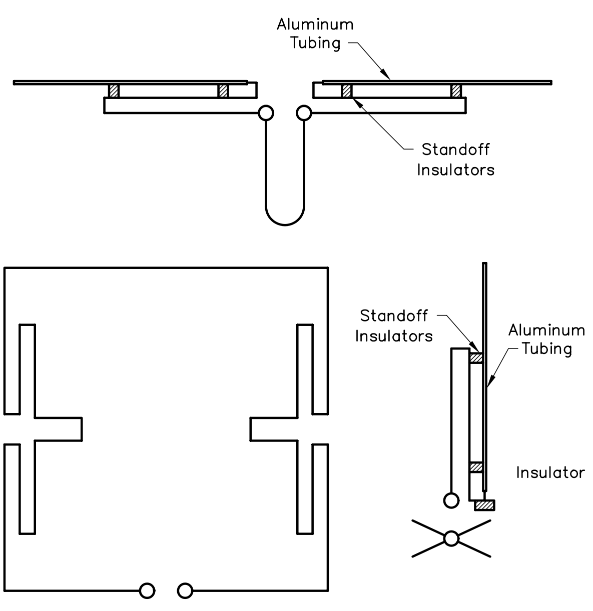

Fig 59—Some examples of linear loading. The small circles indicate the feed points of the antennas.

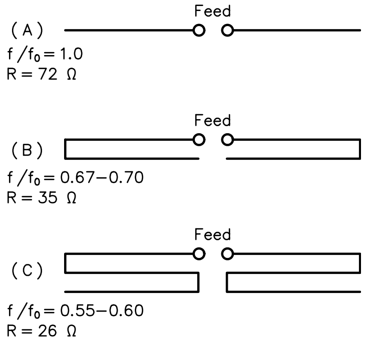

Fig 60—Wire dipole antennas. The ratio f/f_0 is the measured resonant frequency divided by frequency f_0 of a standard dipole of same length. R is radiation resistance in ohms. At A, standard single-wire dipole. At B, two-wire linear-loaded dipole, similar to folded dipole except that side opposite feed line is open. At C, three-wire linear-loaded dipole.

Since the dimensions and spacing of linear-loading devices vary greatly from one antenna installation to another, the best way to employ this technique is to try a length of conductor 10% to 20% longer than the difference between the shortened antenna and the full-size dimension for the linear-loading device. Then use the "cut-and-try" method, varying both the spacing and length of the loading device to optimize the match. A hairpin at the feed point can be useful in achieving a 1:1 SWR at resonance.

Linear-Loaded Short Wire Antennas

More detail on linear loading is provided in this section, which was originally presented in *The ARRL Antenna Compendium Vol 5* by John Stanford, NN0F. Linear loading can significantly reduce the required length for resonant antennas. For example, it is easy to make a resonant antenna that is as much as 30 to 40% shorter than an ordinary dipole for a given band. The shorter overall lengths come from bending back some of the wire. The increased self-coupling lowers the resonant frequency. These ideas are applicable to short antennas for restricted space or portable use.

Experiments

The results of the measurements are shown in **Fig 60** and are also consistent with values given by Rashed and Tai from an earlier paper. This shows several simple wire antenna configurations, with resonant frequencies and impedance (radiation resistance). The reference dipole has a resonant frequency f_0 and resistance $R = 72\ \Omega$. The f/f_0 values give the effective reduced frequency obtained with the linear loading in each case. For example, the two-wire linear-loaded dipole has its resonant frequency lowered to about 0.67 to 0.70 that of the simple reference dipole of the same length.

The three-wire linear-loaded dipole has its frequency reduced to 0.55 to 0.60 of the simple dipole of the same length. As you will see later, these values will vary with conductor diameter and spacing.

The two-wire linear-loaded dipole (Fig 60B) looks almost like a folded dipole but, unlike a folded dipole, it is open in the middle of the side opposite where the feed line is attached. Measurements show that this antenna structure has a resonant frequency lowered to about two-thirds that of the reference dipole, and R equal to about 35 Ω. A three-wire linear-loaded dipole (Fig 60C) has even lower resonant frequency and R about 25 to 30 Ω.

Linear-loaded monopoles (one half of the dipoles in Fig 60) working against a radial ground plane have similar resonant frequencies, but with only half the radiation resistance shown for the dipoles.

A Ladder-Line Linear-Loaded Dipole

Based on these results, NN0F next constructed a linear loaded dipole as in Fig 60B, using 24 feet of 1-inch ladder line (the black, 450-Ω plastic kind widely available) for the dipole length. He hung the system from a tree using nylon fishing line, about 4 feet from the tree at the top, and about 8 feet from the ground on the bottom end. It was slanted at about a 60° angle to the ground. This antenna resonated at 12.8 MHz and had a measured resistance of about 35 Ω. After the resonance measurements, he fed it with 1-inch ladder open-wire line (a total of about 100 feet to the shack).

For brevity, this is called a vertical *LLSD* (linear-loaded short dipole). A tuner resonated the system nicely on 20 and 30 meters. On these bands the performance of the vertical LLSD seemed comparable to his 120-foot long, horizontal center-fed Zepp, 30 feet above ground. In some directions where the horizontal, all-band Zepp has nulls, such as toward Siberia, the vertical LLSD was definitely superior. This system also resonates on 17 and 40 meters. However, from listening to various signals, NNØF had the impression that this length LLSD is not as good on 17 and 40 meters as the horizontal 120-foot antenna.

Using Capacitance End Hats

He also experimented with an even shorter resonant length by trying an LLSD with capacitance end-hats. The hats, as expected, increased the radiation resistance and lowered the resonant frequency. Six-foot long, single-wire hats were used on each end of the previous 24-foot LLSD, as shown in **Fig 61**. The antenna was supported in the same way as the previous vertical dipole, but the bottom-end hat wire was only inches from the grass. This system resonated at 10.6 MHz with a measured resistance of 50 Ω.

If the dipole section were lengthened slightly, by a foot or so, to about 25 feet, it should hit the 10.1-MHz band and be a good match for 50-Ω coax. It would be suitable for a restricted space, shortened 30-meter antenna. Note that this antenna is only about half the length of a conventional 30-meter dipole, needs no tuner, and has no losses due to traps. It does have the loss of the extra wire, but this is essentially negligible.

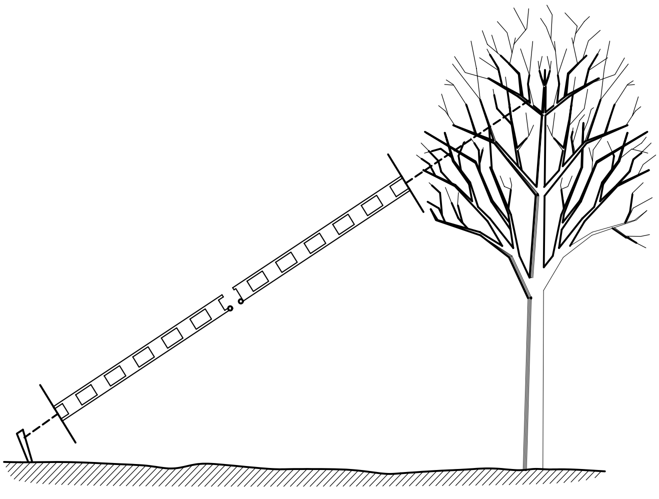

Fig 61—Two-wire linear-loaded dipole with capacitance end hats. Main dipole length was constructed from 24 feet of "windowed" ladder line. The end-hat elements were stiff wires 6 feet long. The antenna was strung at about a 60° angle from a tree limb using monofilament fishing line. Measured resonant frequency and radiation resistance were 10.6 MHz and 50 Ω.

Any of the linear-loaded dipole antennas can be mounted either horizontally or vertically. The vertical version can be used for longer skip contacts—beyond 600 miles or so—unless you have rather tall supports for horizontal antennas to give a low elevation angle. Using different diameter conductors in linear-loaded antenna configurations yields different results, depending on whether the larger or small diameter conductor is fed. NNØF experimented with a vertical ground-plane antenna using a 10-foot piece of electrical conduit pipe (⅝ inch OD) and #12 copper house wire.

Fig 62 shows the configuration. The radial ground system was buried a couple of inches under the soil and is not shown. Note that this is not a folded monopole, which would have either A or B grounded.

The two conductors were separated by 2 inches, using plastic spreaders held onto the pipe by stainless-steel hose clamps obtained from the local hardware store. Hose clamps intertwined at right angles were also used to clamp the pipe on electric fence stand-off insulators on a short 2×4 post set vertically in the ground.

The two different diameter conductors make the antenna characteristics change, depending on how they are configured. With the antenna bridge connected to the larger diameter conductor (point A in Fig 62), and point B unconnected, the system resonated at 16.8 MHz and had R = 35 Ω. With the bridge at B (the smaller conductor), and point A left unconnected, the resonance lowered to 12.4 MHz and R was found to be about 24 Ω.

The resonant frequency of the system in Fig 62 can be adjusted by changing the overall height, or for increasing the frequency, by reducing the length of the wire. Note that a 3.8-MHz resonant ground plane can be made with height only about half that of the usual 67 feet required, if the smaller conductor is fed (point B in Fig 62). In this case, the pipe would be left unconnected electrically. The lengths given above can be scaled to determine a first-try attempt for your favorite band. Resonant lengths will, however, depend on the conductor diameters and spacing.

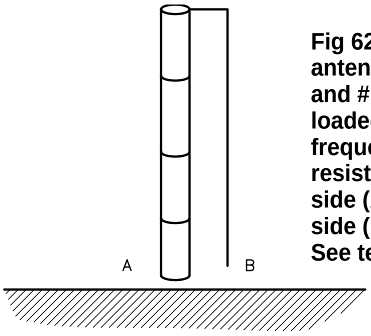

Fig 62—Vertical ground-plane antenna with a 10-foot pipe and #12 wire as the linear-loaded element. Resonant frequency and radiation resistance depend on which side (A or B) is fed. The other side (B or A) is *not* grounded. See text for details.

The same ideas hold for a dipole, except that the lengths should be doubled from those of the ground plane in Fig 62. The resistance will be twice that of the ground plane. Say, how about a shortened 40-meter horizontal beam to enhance your signal?!

COMBINED LOADING

As an antenna is shortened further the size of the top loading device will become larger and at some point will be impractical. In this situation inductive loading, usually placed directly between the capacitance "hat" and the top of the antenna, can be added to resonate the antenna. An alternative would be to use linear loading in place of inductive loading. The previous section contained an example of end loading combined with linear loading.

SHORTENING THE RADIALS

Very often the space required by full-length radials is simply not available. Like the vertical portion of the antenna, the radials can also be shortened and loaded in very much the same way. An example of end loaded radials is given in **Fig 63A**. Radials half the usual length can be used with little reduction in efficiency but, as in the case of top loading, the antenna Q will be higher and the bandwidth reduced. As shown in Fig 63B**,** inductive loading can also be used. As long as they are not made too short (down to 0.1λ) loaded radials can be efficient—with careful design.

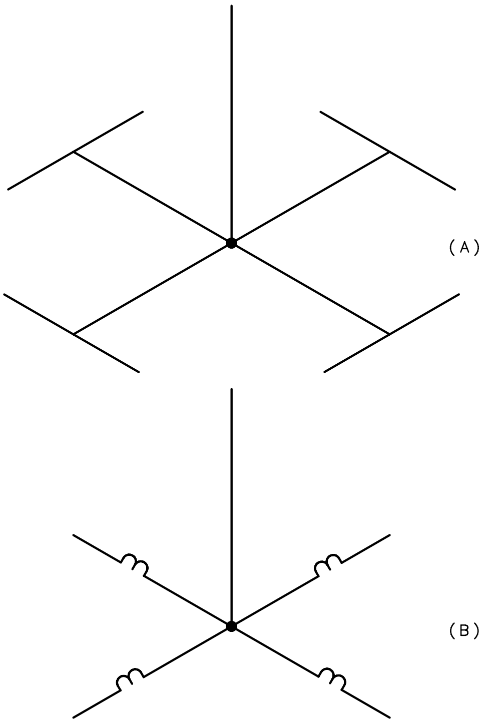

Fig 63—Radials may be shortened by using either capacitive (A) or inductive (B) loading. In extreme cases both may be used but the operating bandwidth will be limited.

GENERAL RULES

The steps in designing an efficient short vertical antenna system are:

- Make the vertical section as long as possible.
- Make the diameter of the vertical section as large as possible. Tubing or a cage of smaller wires will work well.
- Provide as much top and/or bottom loading as possible.
- If the top/bottom loading is insufficient, resonate the antenna with a high-Q inductor placed between the hat and the top of the antenna.
- For buried-ground systems, use as many radials (> 0.2 λ) as possible. 40 or more is best.
- If an elevated ground plane is used, use 4 to 8 radials, 5 or more feet above ground.
- If shortened radials must be used then capacitive loading is preferable to inductive loading.

EXAMPLES OF SHORT VERTICALS

A 6-Foot-High 7-MHz Vertical Antenna

Figs 64 through 67 give details for building short, effective vertical quarter-wavelength radiators. This information was originally presented by Jerry Sevick, W2FMI.

A short vertical antenna, properly designed and installed, approaches the efficiency of a full-size resonant quarter-wave antenna. Even a 6-foot vertical on 7 MHz can produce an exceptional signal. Theory tells us that this should be possible, but the practical achievement of such a result requires an understanding of the problems of ground losses, loading, and impedance matching.

The key to success with shortened vertical antennas lies in the efficiency of the ground system with which the antenna is used. A system of at least 60 radial wires is rec-

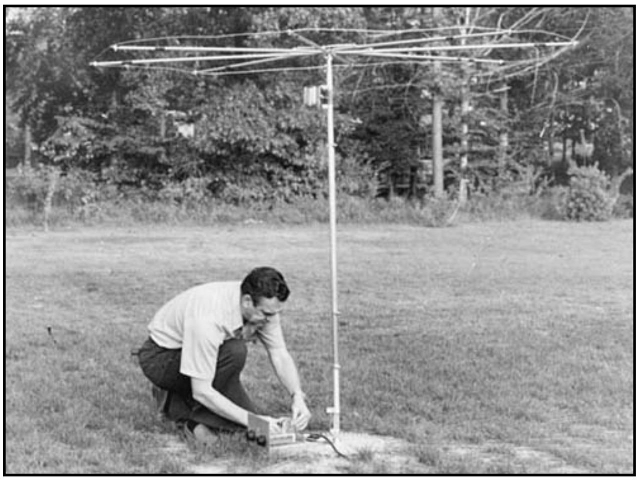

Fig 64—Jerry Sevick, W2FMI, adjusts the 6-foot high, 40-meter vertical.

ommended for best results, although the builder may elect to reduce the number at the expense of some performance. The radials can be tensioned and pinned at the far ends to permit on-the-ground installation, which will enable the amateur to mow the lawn without the wires becoming entangled in the mower blades. Alternatively, the wires can be buried in the ground, where they will not be visible. There is nothing critical about the wire size for the radials. Radials made of 28, 22, or even 16-gauge wire, will provide the same results. The radials should be at least 0.2 λ long (27 feet or greater on 7 MHz).

A top hat is formed as illustrated in **Fig 65**. The diameter is 7 feet, and a continuous length of wire is connected to the spokes around the outer circumference of the wheel. A loading coil consisting of 14 turns of B&W 3029 Miniductor stock (2½-inch dia, 6 TPI, #12 wire) is installed 6 inches below the top hat (see Fig 65). This antenna exhibits a feed-point impedance of 3.5 Ω at 7.21 MHz. For operation above or below this frequency, the number of coil turns must be decreased or increased, respectively. Matching is accomplished by increasing the feed-point impedance to 14 Ω through addition of a 4:1-transformer, then matching 14 Ω to 50 Ω (feeder impedance) by means of a pi network. The 2:1 SWR bandwidth for this antenna is approximately 100 kHz.

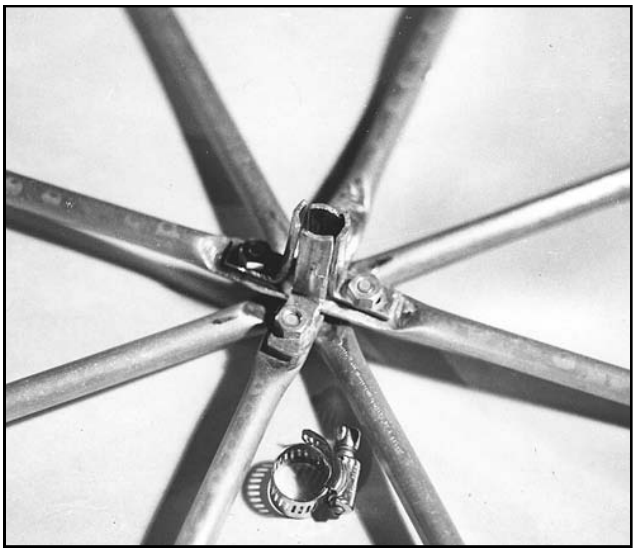

Fig 65—Construction details for the top hat. For a diameter of 7 feet, ½-in. aluminum tubing is used. The hose clamp is made of stainless steel and is available at Sears. The rest of the hardware is aluminum.

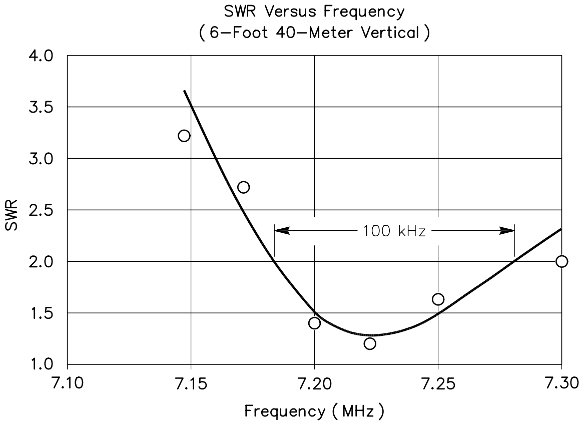

Fig 66—Standing-wave ratio of the 6-foot vertical using a 7-foot top hat and 14 turns of loading 6 inches below the top hat.

More than 200 contacts with the 6-foot antenna have indicated the efficiency and capability of a short vertical. Invariably at distances greater than 500 or 600 miles, the short vertical yields excellent signals. Similar antennas can be scaled and constructed for bands other than 7 MHz. The 7-foot-diameter-top hat was tried on a 3.5-MHz vertical, with an antenna height of 22 feet. The loading coil had 24 turns and was placed 2 feet below the top hat. On-the-air results duplicated those on 40-meters. The bandwidth was 65 kHz.

Short verticals such as these have the ability to radiate and receive almost as well as a full-size quarter-wave. Trade-offs are in lowered input impedances and bandwidths. However, with a good radial system and a proper design, these trade-offs can be made entirely acceptable.

Short Continuously Loaded Verticals

While there is the option of using lumped inductance to achieve resonance in a short antenna, the antenna can also be helically wound to provide the required inductance. This is shown in **Fig 68**. Shortened quarter-wavelength vertical antennas can be made by forming a helix on a long cylindrical insulator. The diameter of the helix must be small in terms of λ to prevent the antenna from radiating in the axial mode.

Acceptable form diameters for HF-band operation are from 1 inch to 10 inches when the practical aspects of antenna construction are considered. Insulating poles of fiberglass, PVC tubing, treated bamboo or wood, or phenolic are suitable for use in building helically wound radiators. If wood or bamboo is used the builder should treat the material with at least two coats of exterior spar varnish prior to winding the antenna element. The completed structure should be given two more coats of varnish, regardless of the material used for the coil form. Application of the varnish will help weatherproof the antenna and prevent the coil turns from changing position.

No strict rule has been established concerning how short a helically wound vertical can be before a significant drop in performance is experienced. Generally, one should use the greatest amount of length consistent with available space. A guideline might be to maintain an element length of 0.05 wavelength or more for antennas which are electrically a quarter wavelength long. Thus, use 13 feet or more of stock for an 80-meter antenna, 7 feet for 40 meters, and so on.

A quarter-wavelength helically wound vertical can be used in the same manner as a full-size vertical. That is, it can be worked against an above-ground wire radial system (four or more radials), or it can be ground-mounted with

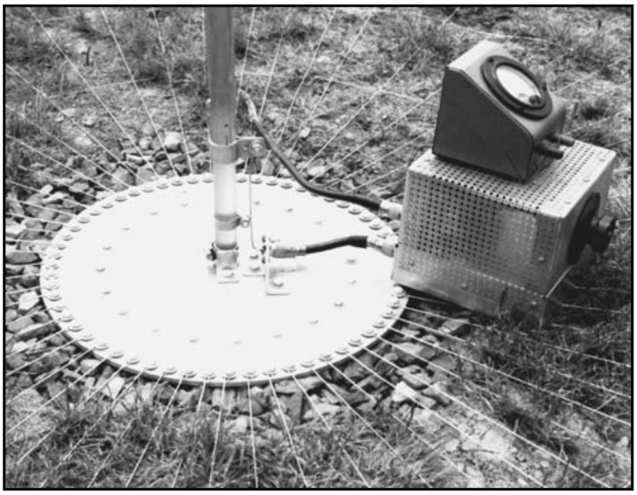

Fig 67—Base of the vertical antenna showing the 60 radial wires. The aluminum disc is 15 inches in diameter and ¼ inches thick. Sixty tapped holes for ¼-20 aluminum hex-head bolts form the outer ring and 20 form the inner ring. The inner bolts were used for performance comparisons with more than 60 radials. The insulator is polystyrene material (phenolic or Plexiglas suitable) with a 1-inch diameter. Also shown is the impedance bridge used for measuring input resistance.

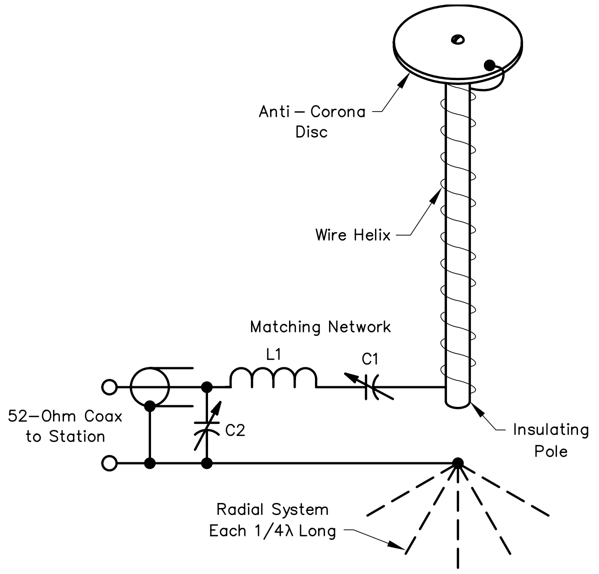

Fig 68—Helically wound ground-plane vertical. Performance from this type of antenna is comparable to that of many full-size λ/4 vertical antennas. The major design trade-off is usable bandwidth. All shortened antennas of this variety are narrow-band devices. At 7 MHz, in the example illustrated here, the bandwidth between the 2:1 SWR points will be on the order of 50 kHz, half that amount on 80 meters, and twice that amount on 20 meters. Therefore, the antenna should be adjusted for operation in the center of the frequency band of interest.

radials buried or lying on the ground. Some operators have reported good results when using antennas of this kind with four helically wound radials cut for resonance at the operating frequency. The latter technique should capture the attention of those persons who must use indoor antennas.

Winding Information

There is no hard-and-fast formula for determining the amount of wire needed to establish resonance in a helical antenna. The relationship between the length of wire needed for resonance and a full quarter wave at the desired frequency depends on several factors. Some of these are wire size, diameter of the turns, and the dielectric properties of the form material, to name a few. Experience has indicated that a section of wire approximately one half wavelength long, wound on an insulating form with a linear pitch (equal spacing between turns) will come close to yielding a resonant quarter wavelength. Therefore, an antenna for use on 160 meters would require approximately 260 feet of wire, spirally wound on the support.

No specific rule exists concerning the size or type of wire one should use in making a helix. Larger wire sizes are, of course, preferable in the interest of minimizing I^2R losses in the system. For power levels up to 1000 W it is wise to use a wire size of #16 or larger. Aluminum clothesline wire is suitable for use in systems where the spacing between turns is greater than the wire diameter. Antennas requiring close-spaced turns can be made from enameled magnet wire or #14 vinyl jacketed, single-conductor house wiring stock. Every effort should be made to keep the turn spacing as large as is practical to maximize efficiency.

A short rod or metal disc should be made for the top or high-impedance end of the vertical. This is a necessary part of the installation to assure reduction in antenna Q. This broadens the bandwidth of the system and helps prevent extremely high amounts of RF voltage from being developed at the top of the radiator. (Some helical antennas act like Tesla coils when used with high-power transmitters, and can actually catch fire at the high-impedance end when a stub or disc is not used.) Since the Q-lowering device exhibits some additional capacitance in the system, it must be in place before the antenna is tuned.

Tuning and Matching

Once the element is wound it should be mounted where it will be used, with the ground system installed. The feed end of the radiator can be connected temporarily to the ground system. Use a dip meter to check the antenna for resonance by coupling the dipper to the last few turns near the ground end of the radiator. Add or remove turns until the vertical is resonant at the desired operating frequency.

It is impossible to predict the absolute value of feed impedance for a helically wound vertical. The value will depend upon the length and diameter of the element, the ground system used with the antenna, and the size of the disc or stub atop the radiator. Generally speaking, the radiation resistance will be very low—approximately 3 to 10 Ω. An L network of the kind shown in Fig 68 can be used to increase the impedance to 50 Ω. The Q_L (loaded Q) of the network inductors is low to provide reasonable bandwidth, consistent with the bandwidth of the antenna. Network values for other operating bands and frequencies can be determined by using the reactance values listed below. The design center for the network is based on a radiation resistance of 5 Ω. If the exact feed impedance is known, the following equations can be used to determine precise component values for the matching network. (See Chapter 25, Coupling the Transmitter to the Line, for additional information on L-network matching.)

XA QRL (Eq 7)

$$X_{C2} = 50\sqrt{\frac{R_L}{50 - R_L}} \qquad \text{(Eq 8)}$$

$$X_{L1} = X_{C1} + \frac{R_L\,50}{X_{C2}} \qquad \text{(Eq 9)}$$

where

X_{C1} = capacitive reactance of C1
X_{C2} = capacitive reactance of C2
X_{L1} = inductive reactance of L1
Q = loaded Q of network
R_L = radiation resistance of antenna

Example: Find the network constants for a helical antenna with a feed impedance of 5 Ω at 7 MHz, Q = 3:

$$X_{C1} = 3 \times 5 = 15$$

$$X_{C2} = \sqrt{\frac{5}{50 - 5}} = 16.666$$

$$X_{L1} = 15 + \frac{250}{16.666} = 30$$

Therefore, C1 = 1500 pF, C2 = 1350 pF, and L1 = 0.7 µH. The capacitors can be made from parallel or series combinations of transmitting micas. L_1 can be a few turns of large Miniductor stock. At RF power levels of 100 W or less, large compression trimmers can be used at C1 and C2 because the maximum RMS voltage at 100 W (across 50 Ω) will be 50 V. At, say, 800 W there will be approximately 220 V RMS developed across 50 Ω. This suggests the use of small transmitting variables at C1 and C2, possibly connected in parallel with fixed values of capacitance to constitute the required amount of capacitance for the network.

By making some part of the network variable, it will be possible to adjust the circuit for an SWR of 1:1 without knowing precisely what the antenna feed impedance is. Actually, C1 is not required as part of the matching network. It is included here to bring the necessary value for L1 into a practical range.

Fig 68 illustrates the practical form a typical helically wound ground-plane vertical might take. Performance from this type antenna is comparable to that of many full-size quarter-wavelength vertical antennas. The major design trade-off is in usable bandwidth. All shortened antennas of this variety are narrow-band devices. At 7 MHz, in the example illustrated here, the bandwidth between the 2:1 SWR points will be on the order of 50 kHz, half that amount on 80 meters, and twice that amount on 20 meters. Therefore, the antenna should be adjusted for operation in the center of the frequency spread of interest.

SHORTENED DIPOLES

As shown in preceding sections, there are a number of ways to load antennas so they may be reduced in size without severe reductions in effectiveness. Loading is always a compromise; the best method is determined by the amount of space available and the band(s) to be worked.

The simplest way to shorten a dipole is shown in **Fig 69**. If you do not have sufficient length between the supports, simply hang as much of the center of the antenna as possible between the supports and let the ends hang down. The ends can be straight down or may be at an angle as indicated but in either case should be secured so that they do not move in the wind. As long as the center portion between the supports is at least ¼, the radiation pattern will be very nearly the same

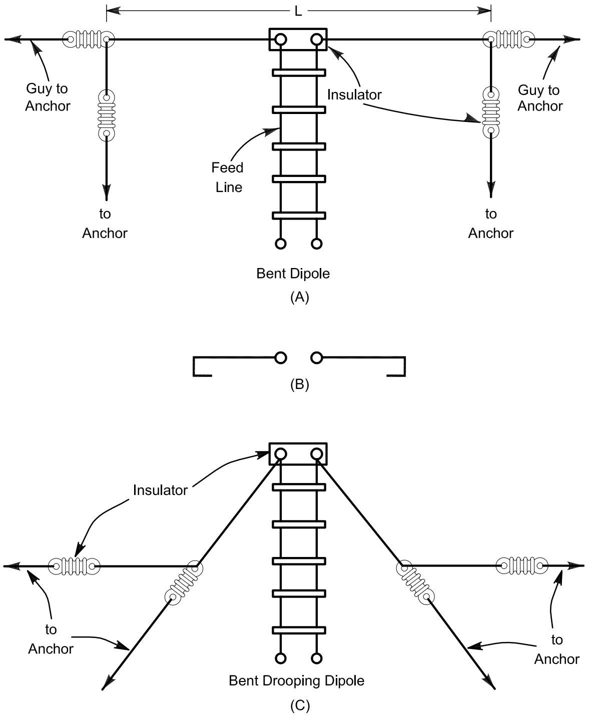

Fig 69—When space is limited, the ends may be bent downward as shown at A, or back on the radiator as shown at B. The bent dipole ends may come straight down or be led off at an angle away from the center of the antenna. An inverted V at C can be erected with the ends bent parallel to the ground when the support structure is not high enough.

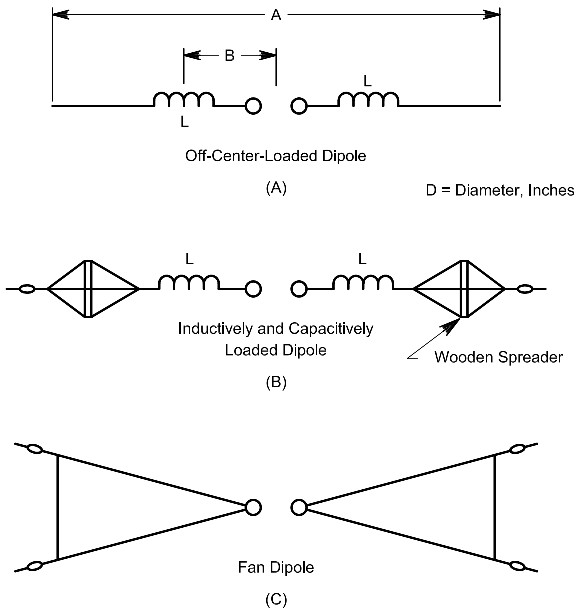

Fig 70—At A is a dipole antenna lengthened electrically with off-center loading coils. For a fixed dimension A, greater efficiency will be realized with greater distance B, but as B is increased, L must be larger in value to maintain resonance. If the two coils are placed at the ends of the antenna, in theory they must be infinite in size to maintain resonance. At B, capacitive loading of the ends, either through proximity of the antenna to other objects or through the addition of capacitance hats, will reduce the required value of the coils. At C, a fan dipole provides some electrical lengthening as well as broadbanding.

as a full-length dipole.

The resonant length of the wire will be somewhat shorter than a full-length dipole and can best be determined by experimentally adjusting the length of ends, which may be conveniently near ground. Keep in mind that there can be very high potentials at the ends of the wires and for safety the ends should be kept out of reach. Letting the ends hang down as shown is a form of capacitive end loading. While it is efficient, it will also reduce the matching bandwidth—as does any form of loading.

The most serious drawback associated with inductive loading is high loss in the coils themselves. It is important that you use inductors made from reasonably large wire or tubing to minimize this problem. Close winding of turns should also be avoided if possible. A good compromise is to use some off-center inductive loading in combination with capacitive end loading, keeping the inductor losses small and the efficiency as high as possible.

Some examples of off-center coil loading and capacitive-end loading are shown in **Fig 70**. This technique was described by Jerry Hall, K1TD in Sep 1974 *QST*. In the equation below, the diameter D is that of the wire used for the antenna, in inches. The frequency f is expressed in MHz.

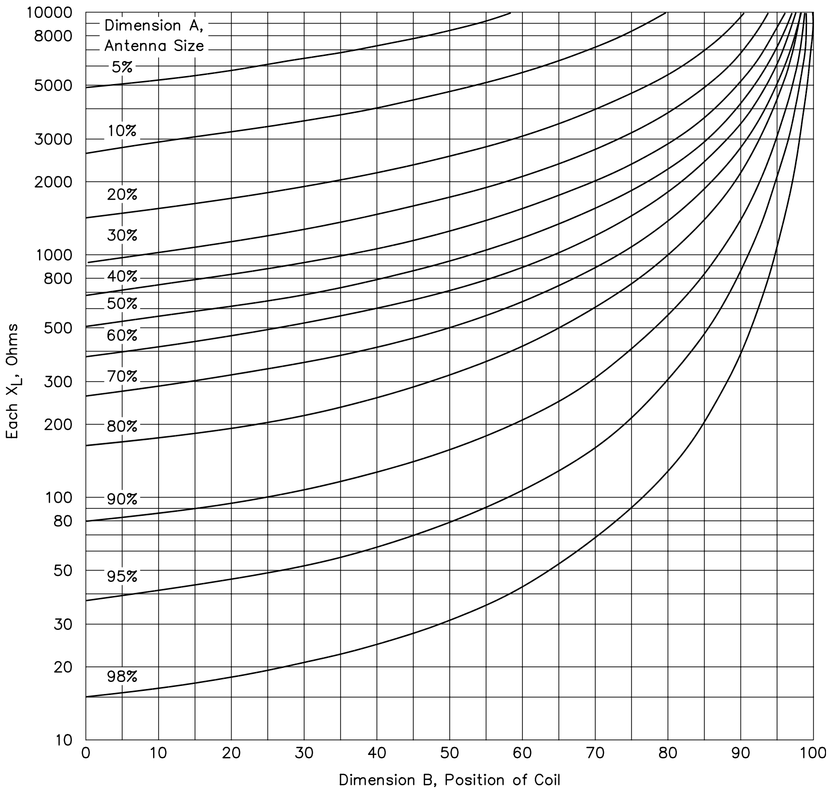

Fig 71—Chart for determining approximate inductance values for off-center-loaded dipoles. See Fig 70A. At the intersection of the appropriate curve from the body of the chart for dimension A and proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at the left. Dimension A is expressed as percent length of the shortened antenna with respect to the length of a half-wave dipole of the same conductor material. Dimension B is expressed as the percentage of coil distance from the feed point to the end of the antenna. For example, a shortened antenna, which is 50% or half the size of a half-wave dipole (one-quarter wavelength overall) with loading coils positioned midway between the feed point and each end (50% out), would require coils having an inductive reactance of approximately 950 Ω at the operating frequency for antenna resonance.

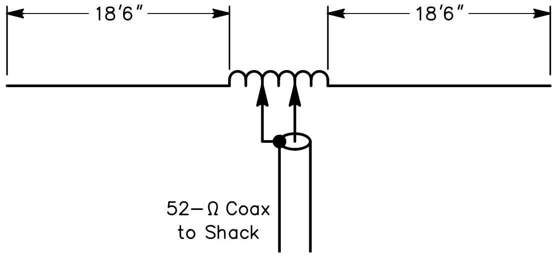

Fig 72—The W0SVM "Shorty Forty" center-loaded antenna. Dimensions given are for 7.0 MHz. The loading coil is 5 inches long and 2½ inches diameter. It has a total of 30 turns of #12 wire wound at 6 turns per inch (Miniductor 3029 stock).

For the antennas shown, the longer the overall length (dimension A, Fig 70A, in feet) and the farther the loading coils are from the center of the antenna (dimension B, also in feet), the greater the efficiency of the antenna. As dimension B is increased, however, the inductance required to resonate the antenna at the desired frequency increases.

Approximate inductive reactances for single-band resonance (for the antenna in Fig 70A only) may be determined with the aid of **Fig 71** or from Eq 10 below. The final values will depend on the proximity of surrounding objects in individual installations and must be determined experimentally. The use of high-Q low-loss coils is important for maximum efficiency.

A dip meter or SWR indicator is recommended for use during adjustment of the system. Note that the minimum inductance required is for a center-loaded dipole. If the inductive reactance is read from Fig 66 for a dimension B of zero, one coil having approximately twice this reactance can be used near the center of the dipole. **Fig 72** illustrates this idea. This antenna was conceived by Jack Sobel, W0SVM, who dubbed the 7-MHz version the "Shorty Forty."

$$X_L = \frac{10^6}{34\pi f}\left(\frac{\left(\ln \frac{24\left(\frac{234}{f}\right) - B}{D} - 1 \right)\left(\left(1 - \frac{fB}{234}\right)^2 - 1 \right)}{\frac{234}{f} - B} - \frac{\left(\ln \frac{24\left(\frac{A}{2} - B\right)}{D} - 1 \right)\left(\left(\frac{\frac{fA}{2} - fB}{234} \right)^2 - 1 \right)}{\frac{A}{2} - B} \right) \quad \text{(Eq 10)}$$

Inverted-L Antennas

The antenna shown in **Fig 73** is called an *inverted-L* antenna. It is simple and easy to construct and is a good antenna for the beginner or the experienced 1.8-MHz DXer. Because the overall electrical length is made somewhat greater than λ/4, the feed-point resistance is on the order of 50 Ω, with an inductive reactance. That reactance is canceled by a series capacitor as indicated in the figure. For a vertical section length of 60 feet and a horizontal section length of 115 feet, the input impedance is ≈ 40 + *j* 300 Ω. Longer vertical or horizontal sections would increase the input impedance. The azimuthal radiation pattern is slightly asymmetrical with ≈1 to 2-dB increase in the direction opposite to the horizontal wire. This antenna requires a good buried ground system or elevated radials and will have a 2:1 SWR bandwidth of about 50 kHz.

This antenna is a form of top-loaded vertical, where the top loading is asymmetrical. This results in both vertical and horizontal polarization because the currents in the top wire do not cancel like they would in a symmetrical-T vertical. This is not necessarily a bad thing because it eliminates the zenith null present in a true vertical. This allows for good communication at short ranges as well as for DX.

A yardarm attached to a tower or a tree limb can be used to support the vertical section. As with any vertical, for best results the vertical section should be as long as possible. A good ground system is necessary for good results—the better the ground, the better the results.

If you don't have the space for the inverted L shown in Fig 73 (with its 115-foot horizontal section) and if you don't have a second tall supporting structure to make the top

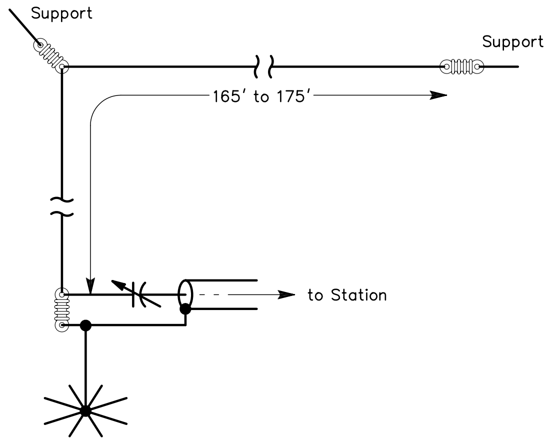

Fig 73—The 1.8-MHz inverted L. Overall wire length is 165 to 175 feet. The variable capacitor has a capacitance range from 100 to 800 pF, at 3 kV or more. Adjust antenna length and variable capacitor for lowest SWR.

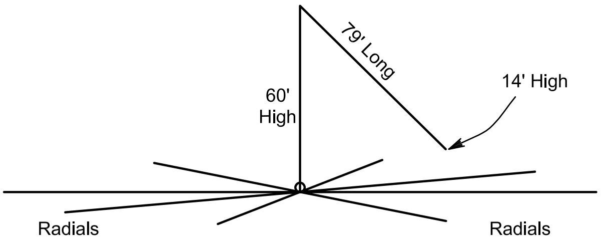

Fig 74—Sketch showing a modified 160-meter inverted L, with a single supporting 60-foot high tower and a 79-foot long slanted top-loading wire. The feed-point impedance is about 12 Ω in this system, requiring a quarter-wave matching transformer made of paralleled 50-Ω coaxes.

wire horizontal, consider sloping the top wire down towards ground. **Fig 74** illustrates such a setup, with a 60-foot high vertical section and a 79-foot sloping wire. As always, you will have to adjust the length of the sloping wire to fine-tune the resonant frequency. For a good ground radial system, the feed-point impedance is about 12 Ω, which may be transformed to 50 Ω with a 25-Ω quarter-wave transformer consisting of two paralleled 50-Ω quarter-wave coaxes. The peak gain will decrease about 1 dB compared to the inverted L shown in Fig 73. **Fig 75** overlays the elevation responses for average ground conditions. The 2:1 SWR bandwidth will be about 30 kHz, narrower than the larger system in Fig 73.

If the ground system suggested for Figs 73 and 74 is not practical, you can use a single elevated radial as shown in **Fig 76**. For the dimensions shown in the figure $Z_i = 50 + j\,498\ \Omega$, requiring a 175-pF series resonating capacitor. The azimuthal radiation pattern is shown in **Fig 77** compared to

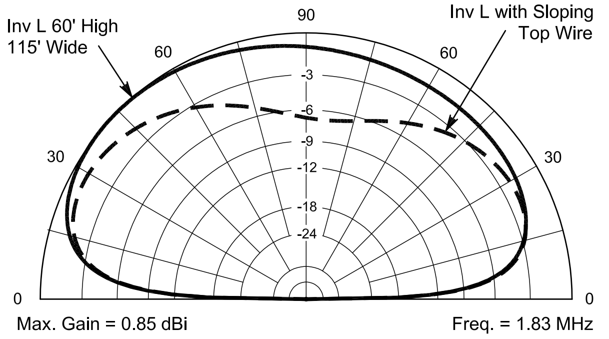

Fig 75—Overlay of the elevation responses for the inverted-L antennas in Fig 73 (solid line) and Fig 74 (dashed line). The gains are very close for these two setups, provided that the ground radial system for the antenna in Fig 74 is extensive enough to keep ground losses low.

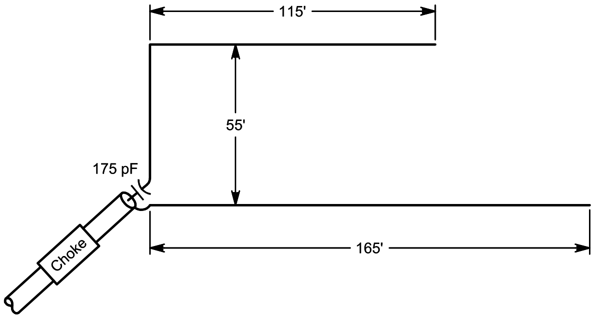

Fig 76—A single elevated radial can be used for the inverted-L. This changes the directivity slightly. The series tuning capacitor is approximately 175 pF for this system.

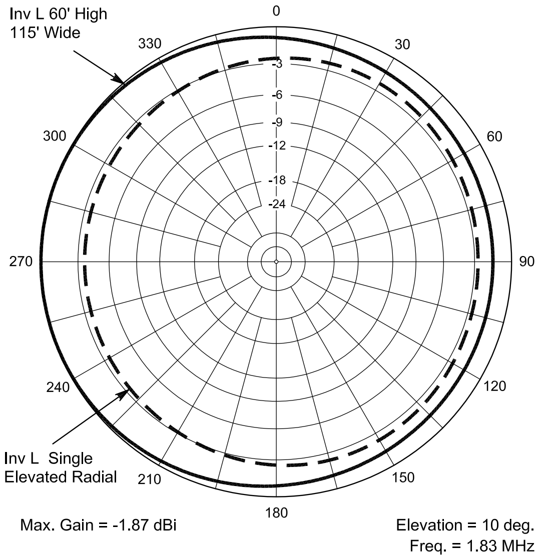

Fig 77—Azimuthal pattern comparison for inverted-L antennas shown in Fig 73 (solid line) and the compromise, single-radial system in Fig 76 (dashed line). This is for a takeoff angle of 10°.

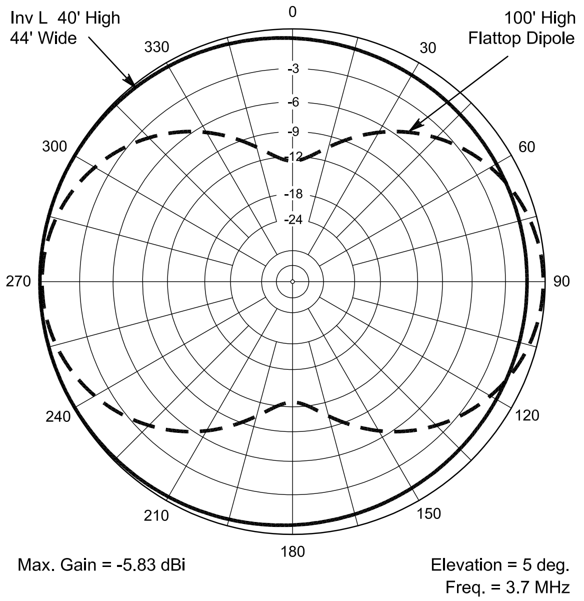

Fig 78—Azimuthal pattern at a takeoff angle of 5° for an 80-meter version of the inverted L (solid line) in Fig 73, compared to the response for a 100-foot high flattop dipole (dashed line).

the inverted L in Fig 68. Note that the 1 to 2-dB asymmetry is now in the direction of the horizontal wires, just the opposite of that for a symmetrical ground system. The 2:1 SWR bandwidth is about 40 kHz, assuming that the series capacitor is adjusted at 1.83 MHz for minimum SWR.

Fig 78 shows the azimuthal response at a 5° elevation angle for an 80-meter version of the inverted L in Fig 73. The peak response occurs at an azimuth directly behind the direction in which the horizontal portion of the inverted L points. For comparison, the response for a 100-foot high flattop dipole is also shown. The top wire of this antenna is only 40 feet high and the 2:1 SWR bandwidth is about 150 kHz wide with a good, low-loss ground-radial system.

Fig 78 illustrates that the azimuth response of an inverted L is nearly omnidirectional. This gives such an antenna an advantage in certain directions compared to a flattop dipole, which is constrained by its supporting mounts (such as trees or towers) to favor fixed directions. For example, the flattop dipole in Fig 78 is at its weakest at azimuths of 90° and 270°, where it is down about 12 dB compared to the inverted L. Hams who are fortunate enough to have high rotary dipoles or rotatable low-band Yagis have found them to be very effective antennas indeed.

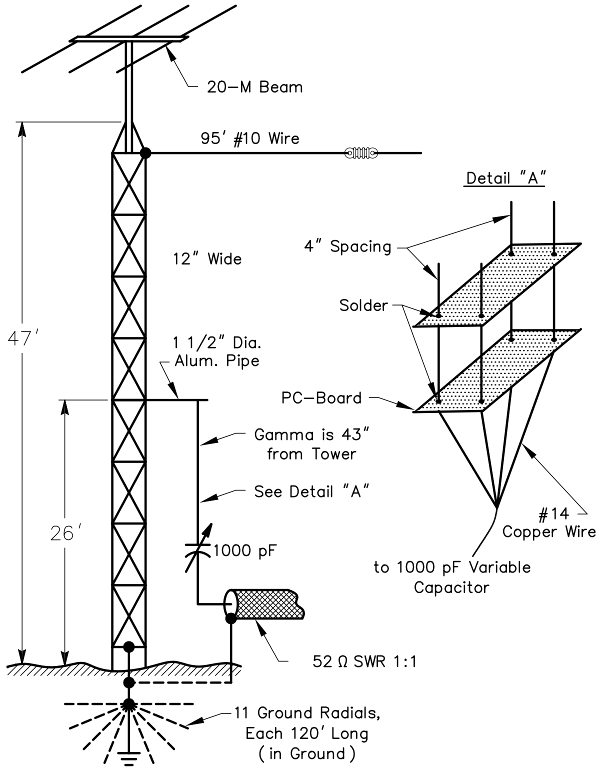

Fig 79—Details and dimensions for gamma-match feeding a 50-foot tower as a 1.8-MHz vertical antenna. The rotator cable and coaxial feed line for the 14-MHz beam is taped to the tower legs and run into the shack from ground level. No decoupling networks are necessary.

A DIFFERENT APPROACH

Fig 79 shows the method used by Doug DeMaw, W1FB, to gamma match his self-supporting 50-foot tower operating as an inverted L. A wire cage simulates a gamma rod of the proper diameter. The tuning capacitor is fashioned from telescoping sections of 1½ and 1½-inch aluminum tubing with polyethylene tubing serving as the dielectric. This capacitor is more than adequate for power levels of 100 W. The horizontal wire connected to the top of the tower provides the additional top loading.

Sloper Antennas

Sloping dipoles and λ/2 dipoles can be very useful antennas on the low bands. These antennas can have one end attached to a tower, tree or other structure and the other end near ground level, elevated high enough so that passersby can't contact them, of course. The following section gives a number of examples of these types of antennas.

THE HALF-WAVE SLOPING DIPOLE

If you have a sufficiently high support, you can install a halfwave dipole sloping downwards toward ground to provide vertical as well as horizontal polarization. This antenna is popularly known as a *sloper* or a *halfwave sloper*. The amount of slope from horizontal can vary from 0°, where the dipole is in a flattop configuration, all the way to 90°, where the dipole becomes fully vertical. The latter configuration is sometimes called a *Halfwave Vertical Dipole* (HVD).

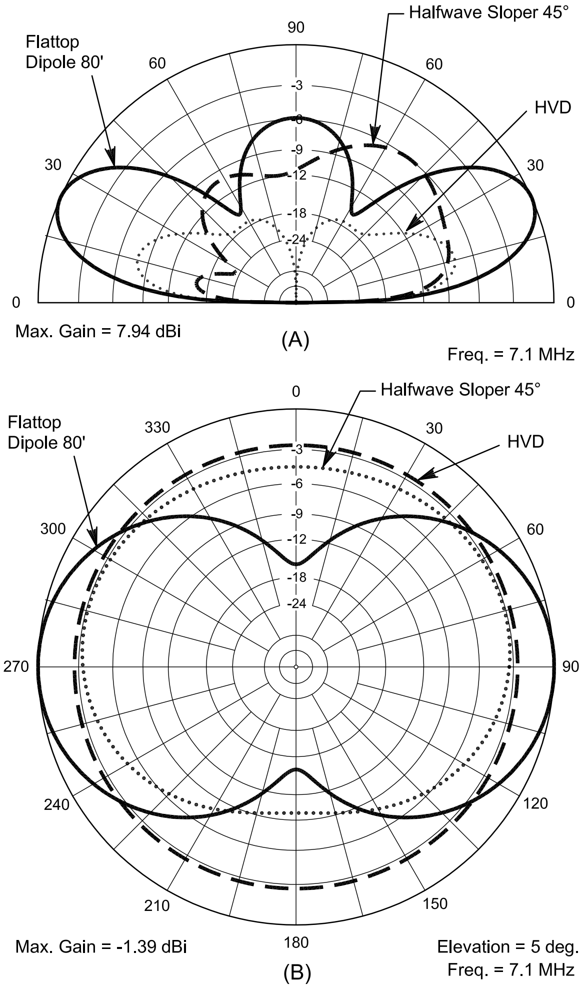

Fig 80—At A, the azimuthal responses for a flattop dipole (solid line), a dipole whose end has been tilted down 45° (dashed line), and a HVD (halfwave vertical dipole, dotted line). All these were modeled over average ground, with a conductivity of 5 mS/m and a dielectric constant of 13. Note that the tilted dipole exhibits about 5 dB front-to-back ratio, although its maximum gain is less than either the HVD or flattop dipole. At B, the elevation-plane patterns for the same antennas. Note that the tilted halfwave dipole (dashed line) has more energy at higher elevation angles than either the flattop dipole or HVD.

The question arises when contemplating a vertical halfwave dipole or a halfwave sloping dipole about how to treat the feed line to make sure it doesn't accidentally become part of the radiating system. The ideal situation would be to bring the feed line out perpendicular to the vertical or sloping wire for an infinite distance. Obviously, that isn't very practical because the feed line eventually has to be connected to a transmitter located near the ground. An intensive modeling study on feeding an HVD was done for the book *Simple and Fun Antennas for Hams*. This study indicated that a slant angle down to the ground of as little as 30° from a vertical radiator can work with only minor interaction, provided that common-mode decoupling chokes were employed at the feed point and a quarter-wavelength down the line from the feed point. These common-mode chokes can consist of either discrete ferrite beads placed over the outer jacket of the coaxial line or multiple turns of the coax itself to form a choke.

Fig 80A compares the 40-meter azimuthal patterns at a DX takeoff angle of 5° for three configurations: a flattop dipole, a dipole tilted down 45° and an HVD (halfwave vertical dipole). These are computed for ground with average conductivity and dielectric constant, and for a maximum height of 80 feet in each configuration. The sloping halfwave dipole exhibits about 5 dB of front-to-back ratio, although even at its most favored direction it doesn't quite have the same maximum gain as the HVD or the flattop dipole.

The reason why the maximum gain for the sloper is less than the other two configurations, even while still exhibiting some front-to-back pattern, is shown in Fig 80A, which shows the elevation-plane patterns for the same antennas, each at the azimuth of maximum gain. The halfwave sloper distributes much of its energy higher in elevation than the HVD, lowering the peak-gain potential of the sloper.

You can also see from Fig 80B that the 80-foot high horizontal dipole would perform much better than either the HVD or halfwave sloper for close-in local contacts, which occur at high elevation angles. On the other hand, except for the greater gain exhibited in the flattop dipole's most favored directions, the HVD has more gain than the other antennas at low elevation angles. While the HVD's omnidirectional pattern is a plus for transmitting, it may be a problem for receiving, where local noise may be coming from specific directions (such as power lines) and may also be predominantly vertically polarized. In such cases, a horizontally polarized flattop dipole may be a considerably better receiving antenna than a vertically polarized antenna of any sort. We've already mentioned the fact that a rotary flattop dipole high in the air can be a very effective antenna on the low bands.

THE QUARTER-WAVELENGTH “HALF SLOPER”

Perhaps one of the easiest antennas to install is the λ/4 sloper shown in **Fig 81**. As pointed out above, a sloping λ/2 dipole is known among radio amateurs as a *sloper* or sometimes as a *full sloper*. If only one half of it is used, it becomes a *half sloper*. The performance of the two types of sloping antennas is similar—They exhibit some directivity in the direction of the slope and radiate vertically polarized energy at low angles respective to the horizon. The amount of directivity will range from 3 to 6 dB, depending upon the individual installation, and will be observed in the slope direction.

The main advantage of the half sloper over the full half-wave-long sloping dipole is that its supporting tower needn’t be as high. Both the half sloper and the full sloper place the feed point (the point of maximum current) high above lossy ground. But the half-sloper only needs half as much wire to build the antenna for a given amateur band. The disadvantage of the half sloper is that it is sometimes difficult or even impossible to obtain a low SWR when using coaxial-cable feed, especially without a good isolating choke balun. (See the section above on isolating ground-plane antennas.)

Other factors that affect the feed impedance are tower height, height of the attachment point, enclosed angle between the sloper and the tower, and what is mounted atop the tower (HF or VHF beams). Further, the quality of the ground under the tower (ground conductivity, radials, etc) has a marked effect on the antenna performance. The final SWR can vary (after optimization) from 1:1 to as high as 6:1. Generally speaking, the closer the low end of the slope wire is to ground, the more difficult it will be to obtain a good match.

Basic Recommendations for a Half Sloper

The half sloper can be an excellent DX type of antenna. Hams usually install theirs on a metal supporting structure such as a mast or tower. The support needs to be grounded at the lower end, preferably to a buried or on-ground radial system. If a nonconductive support is used, the outside of the coax braid becomes the return circuit and should be grounded at the base of the support. As a starting point you can attach the sloper so the feed point is approximately λ/4 above ground. If the tower is not high enough to permit this, the antenna should be fastened as high on the supporting structure as possible. Start with an enclosed angle of approximately 45°, as indicated in Fig 81. Cut the wire to the length determined from

$$\ell = \frac{260}{f_{MHz}} \qquad \text{(Eq 11)}$$

This will allow sufficient extra length for pruning the wire for the lowest SWR. A metal tower or mast becomes an operating part of the half sloper system. In effect, it and the slope wire function somewhat like an inverted-V dipole antenna. In other words, the tower operates as the missing half of the dipole. Hence its height and the top loading (beams) play a significant role.

Detailed modeling indicates that a sufficiently large

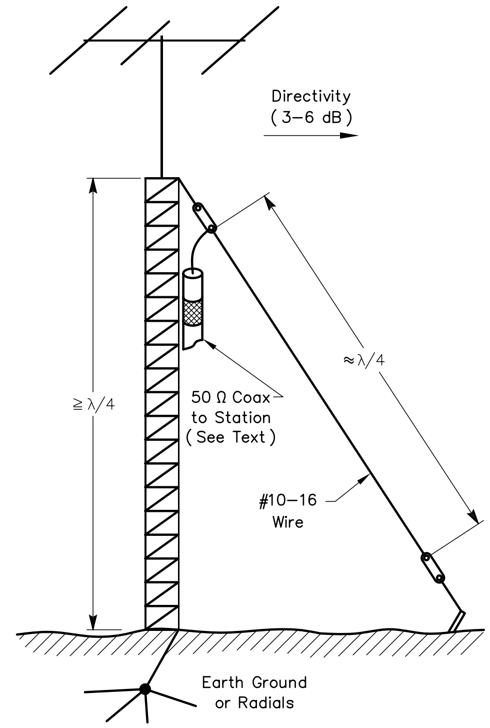

Fig 81—The λ/4 “half sloper” antenna.

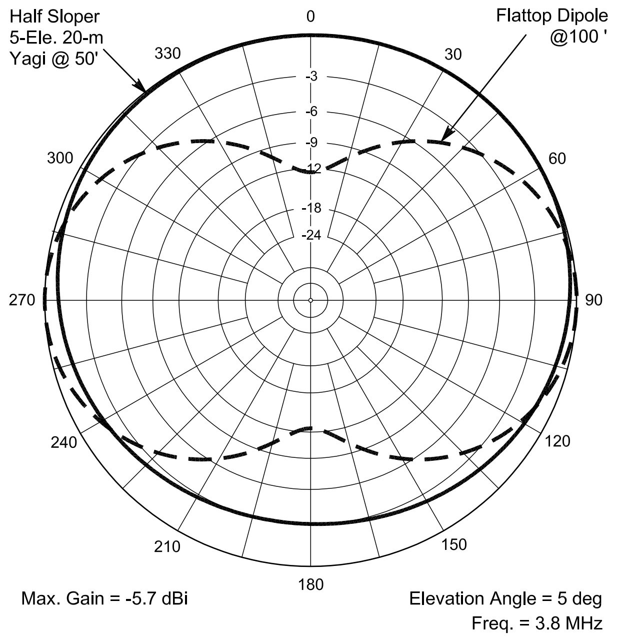

Fig 82—Radiation pattern for a typical half sloper (solid line) mounted on a 50-foot high tower with a large 5-element 20-meter beam on the top compared to that for a flattop dipole (dashed line) at 100 feet. At a 5° takeoff angle typical for DX work on 80 meters, the two antennas are pretty comparable in the directions favored by the high dipole. In other directions, the half sloper has an advantage of more than 10 dB.

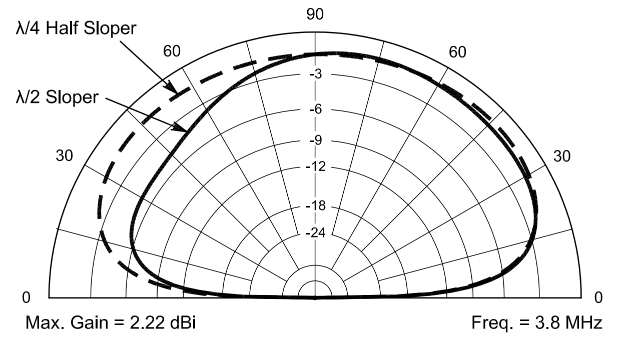

Fig 83—Comparison of elevation patterns for a full-sized halfwave sloper (solid line) on a 100-foot tower and a half sloper (dashed line) on a 50-foot tower with a 5-element 20-meter Yagi acting as a top counterpoise. The performance is quite comparable for these two systems.

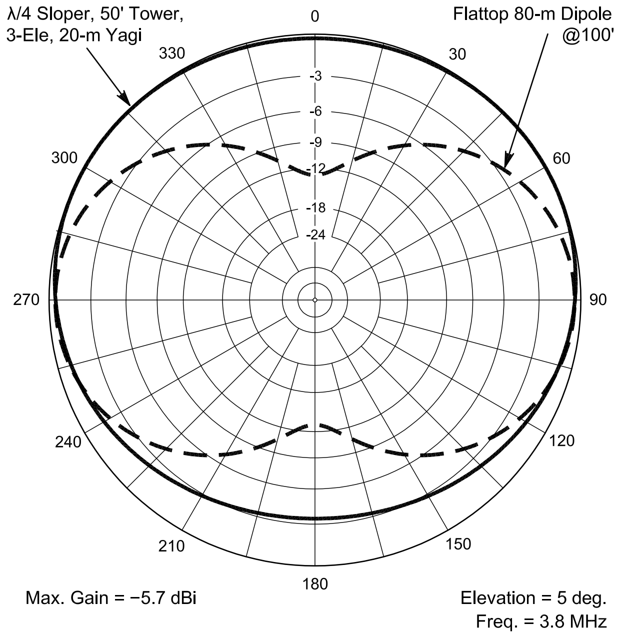

Fig 84—Comparing the azimuthal response of a half sloper (solid line) on a 50-foot tower with a 3-element 20-meter Yagi on top to that of a flattop dipole (dashed line) at 100 feet. The two are again quite comparable at a 5° takeoff angle.

mass of metal (that is, a large, "Plumber's Delight" Yagi) connected to the top of the tower acts like enough of a "top counterpoise" that the tower may be removed from the model with little change in the essential characteristics of the half-sloper system. Consider an installation using a freestanding 50-foot tower with a large 5-element 20-meter Yagi on top. This Yagi is assumed to have a 40-foot boom oriented 90° to the direction of the slanted 80-meter half-sloper wire. The best SWR that could be reached by changing the length and slant angle for this sloper is 1.67:1, representing a feed-point impedance of 30.1 – *j* 2.7 Ω. The peak gain at 3.8 MHz is 0.97 dBi at an elevation angle of 70°. **Fig 82** shows the azimuth-plane pattern for this half sloper, compared to a 100-foot high flattop dipole for reference, at an elevation angle of 5°.

Removing the tower from the model resulted in a feed-point impedance of 30.1 – *j* 1.5 Ω and a peak gain of 1.17 dBi. The tower is obviously not contributing much in this setup, since the mass of the large 20-meter Yagi is acting like an elevated counterpoise all by itself. It's interesting to rotate the boom of the model Yagi and observe the change in SWR that occurs on the half-sloper antenna. With the boom turned 90°, the SWR falls to 1.38:1. This level of SWR change could be measured with amateur-type instrumentation.

On the other hand, substituting a smaller 3-element 20-meter Yagi with an 18-foot boom in the model does result in significant change in feed-point impedance and gain when the tower is removed from the model, indicating that the "counterpoise effect" of the smaller beam is insufficient by itself. Interestingly enough, the best SWR for the half sloper/tower and the 3-element Yagi (with its boom inline with the half sloper is 1.33:1), changing to 1.27:1 with the boom turned 90°. Such a small change in SWR would be difficult to measure using typical amateur instrumentation.

In any case, the 50-Ω transmission line feeding a half sloper should be taped to the tower leg at frequent intervals to make it secure. The best method is to bring it to earth level, then route it to the operating position along the surface of the ground if it can't be buried. This will ensure adequate RF decoupling, which will help prevent RF energy from affecting the equipment in the station. Rotator cable and other feed lines on the tower or mast should be treated in a similar manner.

Adjustment of the half sloper is done with an SWR indicator in the 50-Ω transmission line. A compromise can usually be found between the enclosed angle and wire length, providing the lowest SWR attainable in the center of the chosen part of an amateur band. If the SWR "bottoms out" at 2:1 or lower, the system will work fine without using an antenna tuner, provided the transmitter can work into the load. Typical optimum values of SWR for 3.5 or 7-MHz half slopers are between 1.3:1 and 2:1. A 100-kHz bandwidth is normal on 3.5 MHz, with 200 kHz being typical at 7 MHz.

If the lowest SWR possible is greater than 2:1, the attachment point can be raised or lowered to improve the match. Readjustment of the wire length and enclosed angle may be necessary when the feed-point height is changed. If the tower is guyed, the guy wires will need to be insulated from the tower and broken up with additional insulators to prevent resonance.

At this point you may be curious about which antenna is better—a full sloper or a half sloper. The peak gain for each antenna is very nearly identical. **Fig 83** overlays the elevation-plane pattern for the full-sized halfwave sloper on a 100-foot tower and for the half sloper shown in Fig 81 on a 50-foot tower with a 5-element 20-meter Yagi on top. The

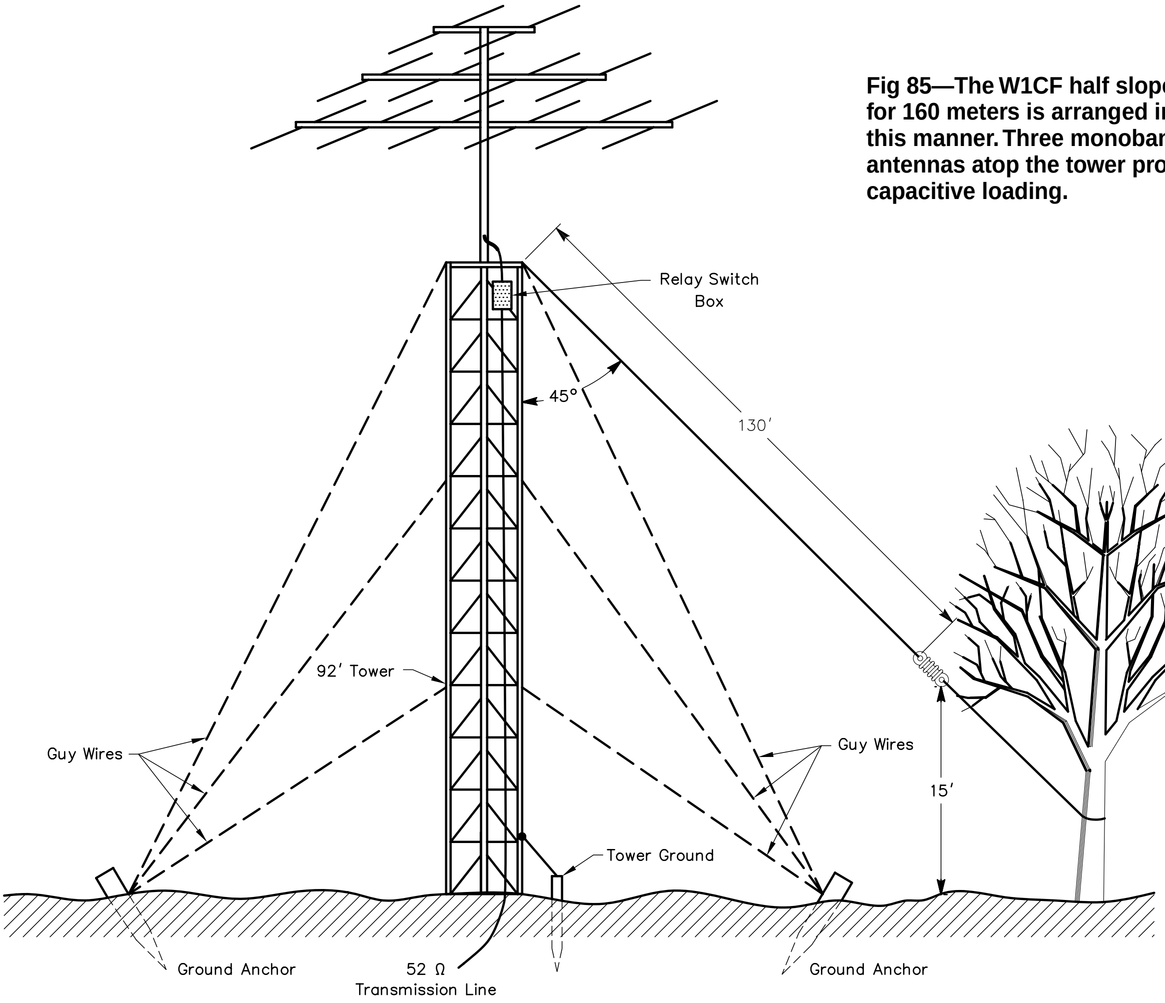

Relay Switch Box
45°
130'
92' Tower
Guy Wires
Guy Wires
15'
Tower Ground
Ground Anchor
52 Ω Transmission Line
Ground Anchor

Fig 85—The W1CF half sloper for 160 meters is arranged in this manner. Three monoband antennas atop the tower provide capacitive loading.

full-sized halfwave sloper has more front-to-back ratio, but it is only a few dB more than the half sloper. **Fig 84** compares the azimuthal patterns at a 5° takeoff angle for a 100-foot high flattop dipole and a half-sloper system on a 50-foot tower with a 3-element 20-meter Yagi on top.

Despite the frustration some have experienced trying to achieve a low SWR with some half-sloper installations, many operators have found the half sloper to be an effective and low-cost antenna for DX work.

1.8-MHz ANTENNA SYSTEMS USING TOWERS

The half sloper discussed above for 80 or 40-meter operation will also perform well on 1.8 MHz where vertically polarized radiators can achieve the low takeoff angles needed on Topband. Prominent 1.8-MHz operators who have had success with the half sloper antenna suggest a minimum tower height of 50 feet. Dana Atchley, W1CF (SK), used the configuration sketched in **Fig 85**. He reported that the uninsulated guy wires act as an effective counterpoise for the sloping wire. In **Fig 86** is the feed system used by Doug

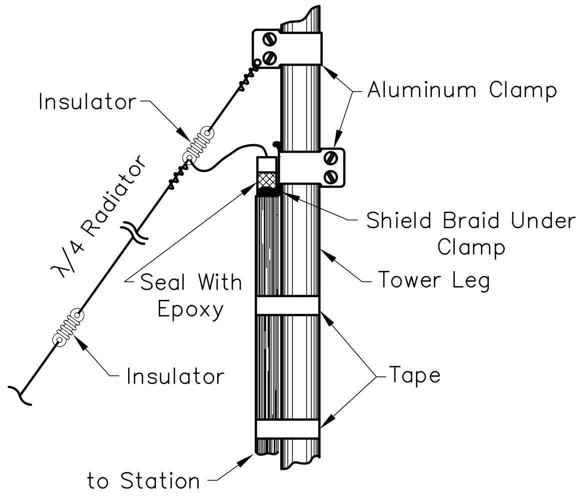

Insulator
Aluminum Clamp
λ/4 Radiator
Shield Braid Under Clamp
Seal With Epoxy
Tower Leg
Insulator
Tape
to Station

Fig 86—Feed system used by W1FB for 1.8 MHz half sloper on a 50-foot self-supporting tower.

DeMaw, W1FB (SK), on a 50-foot self-supporting tower. The ground for the W1FB system is provided by buried radials connected to the tower base.

As described previously, a tower can also be used as a true vertical antenna, provided a good ground system is used. The shunt-fed tower is at its best on 1.8 MHz, where a full λ/4 vertical antenna is rarely possible. Almost any tower height can be used. An HF beam at the top provides some top loading.

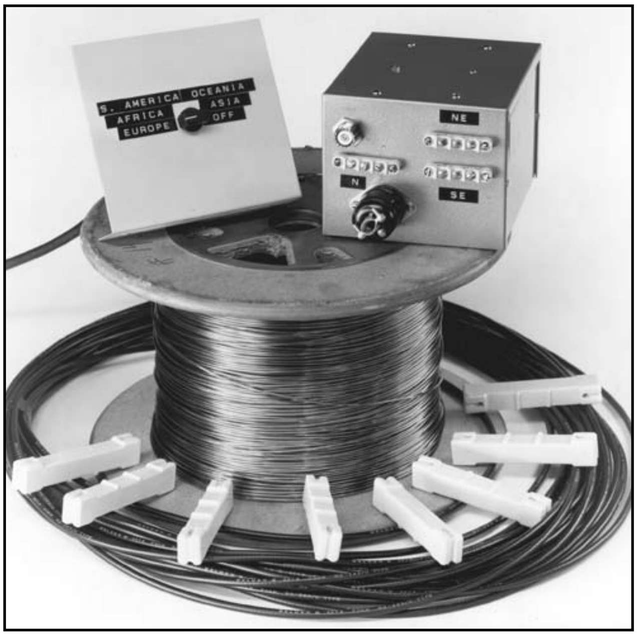

Fig 88—The basic materials required for the sloper system. The control box appears at the left, and the relay box at the right.

THE K1WA 7-MHz "SLOPER SYSTEM"

One of the more popular antennas for 3.5 and 7 MHz is the half-wave long sloping dipole described previously. David Pietraszewski, K1WA, made an extensive study of sloping dipoles at different heights with reflectors at the 3-GHz frequency range. From his experiments, he developed the novel 7-MHz antenna system described here. With several sloping dipoles supported by a single mast and a switching network, an antenna with directional characteristics and forward gain can be simply constructed. This 7-MHz system uses several "slopers" equally spaced around a common center support. Each dipole is cut to λ/2 and fed at the center with 50-Ω coax. The length of each feed line is 36 feet.

All of the feed lines go to a common point on the support (tower) where the switching takes place. The line length of 36 feet is just over 3λ/8, which provides a useful quality. At 7 MHz, the coax looks inductive to the antenna when the end at the switching box is open circuited. This has the effect of adding inductance at the center of the sloping dipole element, which electrically lengthens the element. The 36-foot length of feed line serves to increase the length of the element about 5%. This makes any unused element appear to be a reflector.

The array is simple and effective. By selecting one of the slopers through a relay box located at the tower, the system becomes a parasitic array that can be electrically rotated. All but the driven element of the array become reflectors.

The physical layout is shown in **Fig 87**, and the basic materials required for the sloper system are shown in **Fig 88**. The height of the support point should be about 70 feet, but can be less and still give reasonable results. The upper portion of the sloper is 5 feet from the tower, suspended by rope. The wire makes an angle of 60° with the ground.

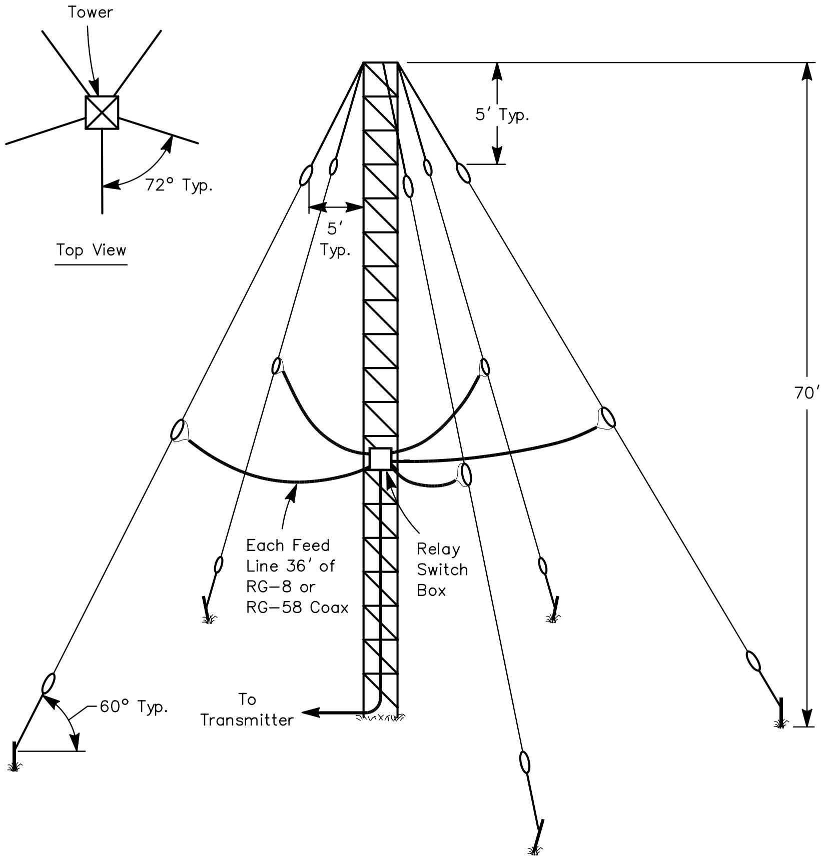

Fig 87—Five sloping dipoles suspended from one support. Directivity and forward gain can be obtained from this simple array. The top view shows how the elements should be spaced around the support.

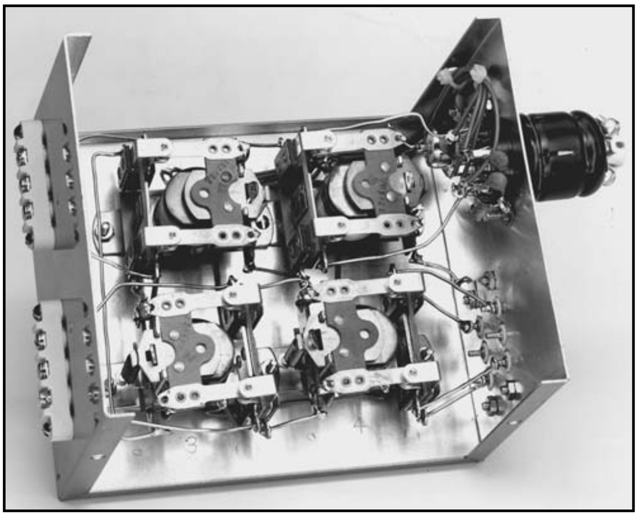

Fig 89—Inside view of relay box. Four relays provide control over five antennas. See text. The relays pictured here are Potter and Brumfeld type MR11D.

In **Fig 89**, the switch box is shown containing all the necessary relays to select the proper feed line for the desired direction. One feed line is selected at a time and the feed lines of those remaining are opened, **Fig 90**. In this way the array is electrically rotated. These relays are controlled from inside the shack with an appropriate power supply and rotary switch. For safety reasons and simplicity, 12-volt dc relays are used. The control line consists of a five conductor cable, one wire used as a common connection; the others go to the four relays. By using diodes in series with the relays and a dual-polarity power supply, the number of control wires can be reduced, as shown in Fig 90B.

Measurements indicate that this sloper array provides up to 20 dB front-to-back ratio and forward gain of about 4 dB over a single half-wave sloper. **Fig 91** shows the azimuthal pattern (at a 5° takeoff angle) for the K1WA array, compared to a 100-foot high flattop dipole and a full sloper suspended from a 50-foot tower. These patterns were calculated for average ground conditions. Just for fun, look at

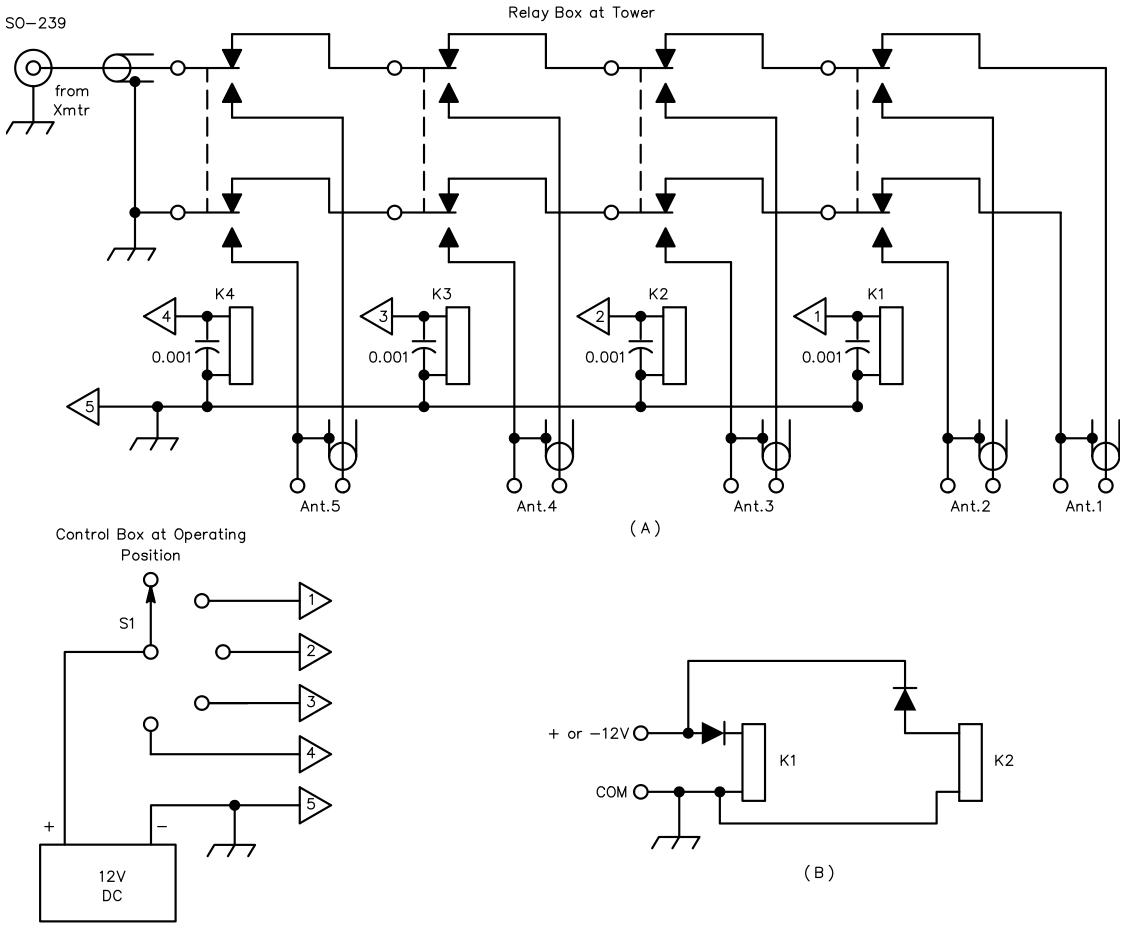

Fig 90—Schematic diagram for sloper control system. All relays are 12-volt dc, DPDT, with 8-A contact ratings. At A, the basic layout, excluding control cable and antennas. Note that the braid of the coax is also open-circuited when not in use. Each relay is bypassed with 0.001-µF capacitors. The power supply is a low current type. At B, diodes are used to reduce the number of control wires when using dc relays. See text.

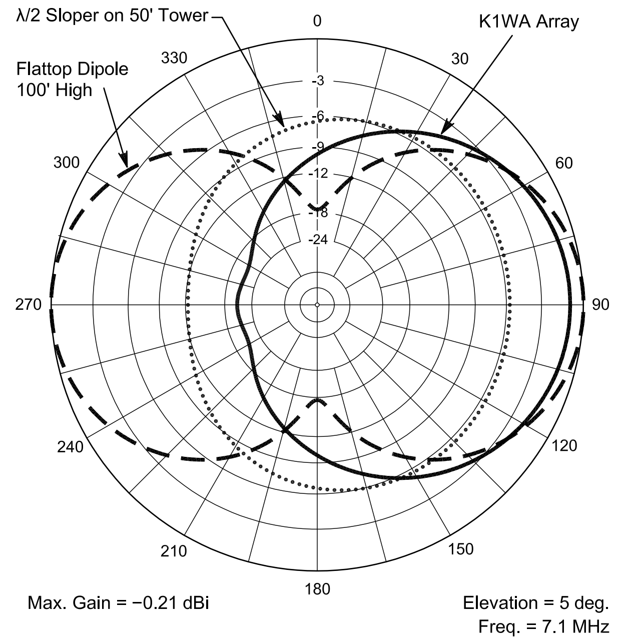

Fig 91—Azimuth pattern for K1WA 40-meter sloper array (solid line), compared to a flattop dipole (dashed line) at 100 feet and a halfwave full sloper on a 50-foot tower (dotted line). The K1WA array has an excellent front-to-back ratio and almost as much gain as the high flattop dipole. These patterns are for average ground.

Fig 92, which shows a comparison between a 100-foot high flattop dipole and a K1WA array placed over saltwater. Now that's a real barnburner at low takeoff angles! Such a seaside system would be very competitive with a rotatable 2-element "shorty-40" type of Yagi.

If one direction is the only concern, the switching system can be eliminated and the reflectors should be cut 5% longer than the resonant frequency. The feature worth noting is the good F/B ratio. By arranging the system properly, a null can be placed in an unwanted direction, thus making it an effective receiving antenna. In the tests conducted with this antenna, the number of reflectors used were as few as one and as many as five. The optimum combination appeared to occur with four reflectors and one driven element. No tests were conducted with more than five reflectors. This same array can be scaled to 3.5 MHz for similar results.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

D. Atchley, Jr., "Putting the Quarter-Wave Sloper to Work on 160," *QST*, Jul 1979, pp 19-20.

J. Belrose, "Transmission Line Low Profile Antennas," *QST*, Dec 1975, pp 19-25.

L. Braskamp, AA6GL, "*MOBILE,* a Computer Program for Short HF Verticals," *The ARRL Antenna Compendium Vol 4*, pp 92-96.

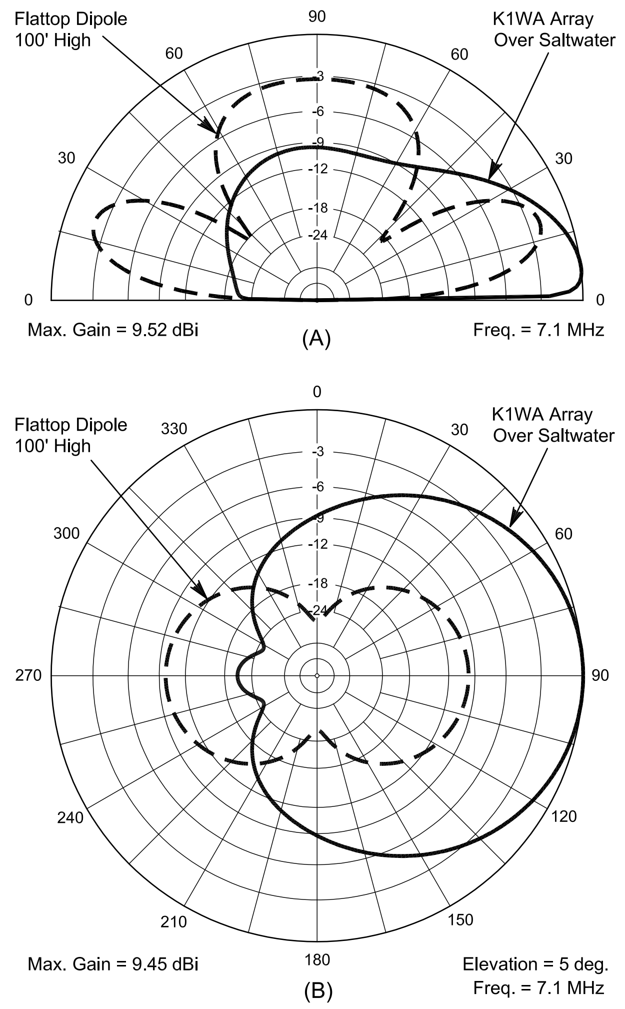

Fig 92—The K1WA 40-meter sloper array over saltwater, compared to a 100-foot high flattop dipole. At A, azimuth patterns. At B, elevation patterns. Now the sloper array really comes into its own!

G. H. Brown, "The Phase and Magnitude of Earth Currents Near Radio Transmitting Antennas," *Proc. IRE*.

G. H. Brown, R. F. Lewis and J. Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proc. IRE* , Jun 1937, pp 753-787.

P. Carr, N4PC, "A Two Band Half-Square Antenna With Coaxial Feed," *CQ*, Sep 1992, pp 40-45.

P. Carr, N4PC, "A DX Antenna For 40 Meters," *CQ*, Sep 1994, pp 40-43.

A. Christman, KB8I, "Elevated Vertical Antennas for the Low Bands," *The ARRL Antenna Compendium Vol 5*, pp 11-18.

Curry Communications, PO Box 1884, Burbank, CA 91507, a kit manufacturer of lowfer CW transceivers.

D. DeMaw, "Additional Notes on the Half Sloper," *QST*, Jul 1979, p 20.

P. Dodd, *The LF Experimenter's Source Book* (RSGB, 1996).

A. C. Doty, Jr., J. A. Frey and H. J. Mills, "Efficient Ground Systems for Vertical Antennas," *QST*, Feb 1983, pp 9-12.

S. Ford, "*Low*fing on 1750 Meters," *QST*, Oct 1993, pp 67-68.

R. Fosberg, "Some Notes on Ground Systems for 160 Meters," *QST*, Apr 1965, pp 65-67.

J. Hall, "Off-Center-Loaded Dipole Antennas," *QST*, Sep 1974, pp 28-34.

G. Hubbell, "Feeding Grounded Towers as Radiators," *QST*, pp 32-33, 140, 142.

C. Hutchinson and R. D. Straw, *Simple and Fun Antennas for Hams* (Newington, Ct: ARRL, 2002).

P. H. Lee, *The Amateur Radio Vertical Antenna Handbook*, 1st edition (Port Washington, NY: Cowan Publishing Corp., 1974).

C. J. Michaels, "Some Reflections on Vertical Antennas," *QST*, July 1987, pp 15-19; feedback *QST*, Aug 1987, p 39.

Rashed and C. Tai, "A New Class of Wire Antennas," 1982 International Symposium Digest, *Antennas and Propagation*, Vol 2, IEEE.

T. Russell, N4KG, "Simple, Effective, Elevated Ground-Plane Antennas," *QST*, June 1994, pp 45-46

F. J. Schnell, "The Flagpole Deluxe," *QST*, Mar 1978, pp 29-32.

R. Severns, N6LF, "Using the Half-Square Antenna for Low-Band Dxing," *The ARRL Antenna Compendium Vol 5*, pp 35-42.

R. Severns, N6LF, "Broadbanding the Half-Square Antenna For 80-meter Dxing," *The ARRL Antenna Compendium Vol 5*, pp 43-44.

R. Severns, N6LF, "An Improved Double Extended Zepp," *The ARRL Antenna Compendium Vol 4*, pp 78-80.

J. Sevick, "The Ground-Image Vertical Antenna," *QST*, Jul 1971, pp 16-19, 22.

J. Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, Mar 1973, pp 13-18, 41.

J. Sevick, "The Constant-Impedance Trap Vertical," *QST* , Mar 1974, pp 29-34.

J. Sevick, "Short Ground-Radial Systems for Short Verticals," *QST*, Apr 1978, pp 30-33.

J. Sevick, W2FMI, *Transmission Line Transformers* (Noble Publishing, Atlanta, 1996).

W. Smith, W6BCX, "Bet My Money On A Bobtail Beam," *CQ*, Mar 1948, pp 21-23, 92.

J. Stanford, NNØF, "Linear-Loaded Short Wire Antennas," *The ARRL Antenna Compendium Vol 5*, pp 105-107.

J. Swank, W8HXR, "The S-Meter Bender," *73*, Jun 1978, pp 170-173.

J. Tyskewicz, "The Heli-Rope Antenna," *QST*, Jun 1971.

B. Vester, K3BC, "The Half-Square Antenna," *QST*, Mar 1974, pp 11-14.

A. D. Watt, *VLF Radio Engineering* (Pergamon Press, 1967) (out of print).

Prediction of Sky-wave Field Strength at Frequencies Between About 150 and 1700 kHz, ITU Doc 3/14, Radiocommunication Study Groups, Feb 1995.

Chapter 7

Multiband Antennas

For operation in a number of bands, such as those between 3.5 and 30 MHz, it would be impractical for most amateurs to put up a separate antenna for each band. But this is not necessary—a dipole, cut for the lowest frequency band to be used, can be operated readily on higher frequencies. To do so, one must be willing to accept the fact that such harmonic-type operation leads to a change in the directional pattern of the antenna, both in the azimuth and the elevation planes (see Chapter 2, Antenna Fundamentals, and Chapter 3, The Effects of Ground).

You can see from discussions in Chapter 6, Low-Frequency Antennas, that you should carefully plan the height at which you install a multiband horizontally polarized antenna. This is one aspect of multiband antennas. Another important thing to consider is that you should be willing to use so-called *tuned feeders*. A center-fed single-wire antenna can be made to accept power and radiate it with high efficiency on any frequency higher than its fundamental resonant frequency and, with a reduction in efficiency and bandwidth, on frequencies as low as one half the fundamental.

In fact, it is not necessary for an antenna to be a full half-wavelength long at the lowest frequency. An antenna can be considerably shorter than ½ λ, even as short as ¼ λ, and still be a very efficient radiator. The use of such short antennas results in stresses, however, on other parts of the system, for example the antenna tuner and the transmission line. This will be discussed in some detail in this chapter.

Methods have been devised for making a single antenna structure operate on a number of bands while still offering a good match to a transmission line, usually of the coaxial type. It should be understood, however, that a multiband antenna is not *necessarily* one that will match a given line on all bands on which you intend to use it. Even a relatively short whip type of antenna can be operated as a multiband antenna with suitable loading for each band. Such loading may be in the form of a coil at the base of the antenna on those frequencies where loading is needed, or it may be incorporated in the tuned feeders running from the transmitter to the base of the antenna.

This chapter describes a number of systems that can be used on two or more bands. Beam antennas, such as Yagis or quads, are treated separately in later chapters.

Simple Wire Antennas

The simplest multiband antenna is a random length of #12 or #14 wire. Power can be fed to the wire on practically any frequency using one or the other of the methods shown in **Fig 1**. If the wire is made either 67 or 135 feet long, it can also be fed through a tuned circuit, as in **Fig 2**. It is advantageous to use an SWR bridge or other indicator in the coax line at the point marked "X."

If you have installed a 28- or 50-MHz rotary beam, in many cases it may be possible to use the beam's feed line as an antenna on the lower frequencies. Connecting the two wires of the feeder together at the station end will give a random-length wire that can be conveniently coupled to the transmitter as in Fig 1. The rotary system at the far end will serve only to *end-load* the wire and will not have much other effect.

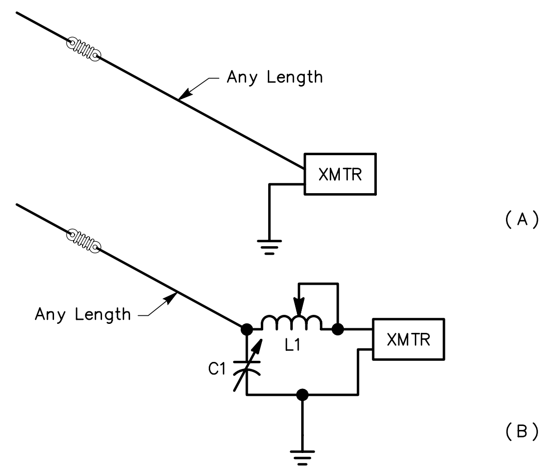

Fig 1—At A, a random-length wire driven directly from the pi-network output of a transmitter. At B, an L network for use in cases where sufficient loading cannot be obtained with the arrangement at A. C1 should have about the same plate spacing as the final tank capacitor in a vacuum-tube type of transmitter; a maximum capacitance of 100 pF is sufficient if L1 is 20 to 25 µH. A suitable coil would consist of 30 turns of #12 wire, 2½ inches diameter, 6 turns per inch. Bare wire should be used so the tap can be placed as required for loading the transmitter.

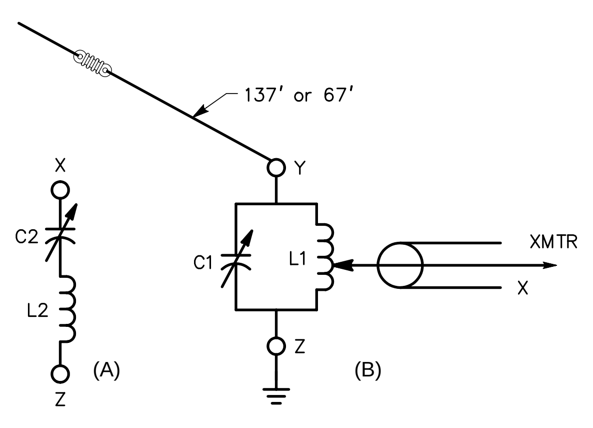

Fig 2—If the antenna length is 137 feet, a parallel-tuned coupling circuit can be used on each amateur band from 3.5 through 30 MHz, with the possible exception of the WARC 10-, 18- and 24-MHz bands. C1 should duplicate the final tank tuning capacitor and L1 should have the same dimensions as the final tank inductor on the band being used. If the wire is 67 feet long, series tuning can be used on 3.5 MHz as shown at the left; parallel tuning will be required on 7 MHz and higher frequency bands. C2 and L2 will in general duplicate the final tank tuning capacitor and inductor, the same as with parallel tuning. The L network shown in Fig 1B is also suitable for these antenna lengths.

One disadvantage of all such directly fed systems is that part of the antenna is practically within the station, and there is a good chance that you will have some trouble with RF feedback. RF within the station can often be minimized by choosing a length of wire so that the low feed-point impedance at a current loop occurs at or near the transmitter. This means using a wire length of $\lambda/4$ (65 feet at 3.6 MHz, 33 feet at 7.1 MHz), or an odd multiple of $\lambda/4$ (¾-λ is 195 feet at 3.6 MHz, 100 feet at 7.1 MHz). Obviously, this can be done for only one band in the case of even harmonically related bands, since the wire length that presents a current loop at the transmitter will present a voltage loop at two (or four) times that frequency.

When you operate with a random-length wire antenna, as in Figs 1 and 2, you should try different types of grounds on the various bands, to see what gives you the best results. In many cases it will be satisfactory to return to the transmitter chassis for the ground, or directly to a convenient metallic water pipe. If neither of these works well (or the metallic water pipe is not available), a length of #12 or #14 wire (approximately $\lambda/4$ long) can often be used to good advantage. Connect the wire at the point in the circuit that is shown grounded, and run it out and down the side of the house, or support it a few feet above the ground if the station is on the first floor or in the basement. It should not be connected to actual ground at any point.

END-FED ANTENNAS

When a straight-wire antenna is fed at one end with a two-wire transmission line, the length of the antenna portion becomes critical if radiation from the line is to be held to a minimum. Such an antenna system for multiband operation is the *end-fed Zepp* or *Zepp-fed* antenna shown in **Fig 3**. The antenna length is made $\lambda/2$ long at the lowest operating frequency. (This name came about because the first documented use of this sort of antennas was on the *Zeppelin* airships.) The feeder length can be anything that is convenient, but feeder lengths that are multiples of $\lambda/4$ generally give trouble with parallel currents and radiation from the feeder portion of the system. The feeder can be an open-wire line of #14 solid

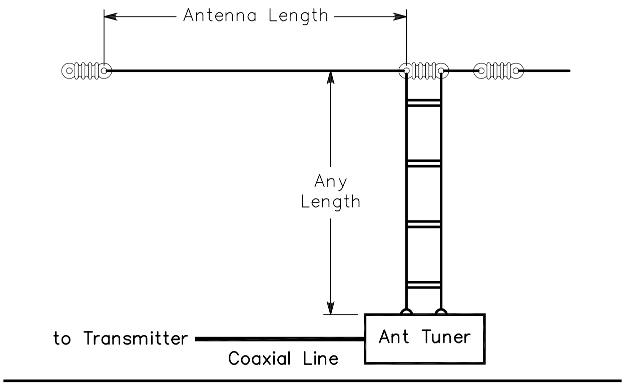

Fig 3—An end-fed Zepp antenna for multiband use.

copper wire spaced 4 or 6 inches with ceramic or plastic spacers. Open-wire TV line (not the type with a solid web of dielectric) is a convenient type to use. This type of line is available in approximately 300- and 450-Ω characteristic impedances.

If you have room for only a 67-foot flat top and yet want to operate in the 3.5-MHz band, the two feeder wires can be tied together at the transmitter end and the entire system treated as a random-length wire fed directly, as in Fig 1. The simplest precaution against parallel currents that could cause feed-line radiation is to use a feeder length that is not a multiple of λ/4. An antenna tuner can be used to provide multiband coverage with an end-fed antenna with any length of open-wire feed line, as shown in Fig 3.

CENTER-FED ANTENNAS

The simplest and most flexible (and also least expensive) all-band antennas are those using open-wire parallel-conductor feeders to the center of the antenna, as in **Fig 4**. Because each half of the flat top is the same length, the feeder currents will be balanced at all frequencies unless, of course, unbalance is introduced by one half of the antenna being closer to ground (or a grounded object) than the other. For best results and to maintain feed-current balance, the feeder should run away at right angles to the antenna, preferably for at least λ/4.

Center feed is not only more desirable than end feed because of inherently better balance, but generally also results in a lower standing wave ratio on the transmission line, provided a parallel-conductor line having a characteristic impedance of 450 to 600 Ω is used. TV-type open-wire line is satisfactory for all but possibly high power installations (over 500 W), where heavier wire and wider spacing is desirable to handle the larger currents and voltages.

The length of the antenna is not critical, nor is the length of the line. As mentioned earlier, the length of the antenna can be considerably less than λ/2 and still be very effective. If the overall length is at least λ/4 at the lowest frequency, a quite usable system will result. The only difficulty that may exist with this type of system is the matter of coupling the antenna-system load to the transmitter. Most modern transmitters are designed to work into a 50-Ω coaxial load. With this type of antenna system a coupling network (an antenna tuner) is required.

Feed-Line Radiation

The preceding sections have pointed out means of reducing or eliminating feed-line radiation. However, it should be emphasized that any radiation from a transmission line is not "lost" energy and is not necessarily harmful. Whether or not feed-line radiation is important depends entirely on the antenna system being used. For example, feed-line radiation is not desirable when a directive array is being used. Such feed-line radiation can distort the desired pattern of such an array, producing responses in unwanted directions. In other words, you want radiation only from the directive array, rather than from the directive array and the feed line. See Chapter 26, Coupling the Line to the Antenna, for a detailed discussion of this topic.

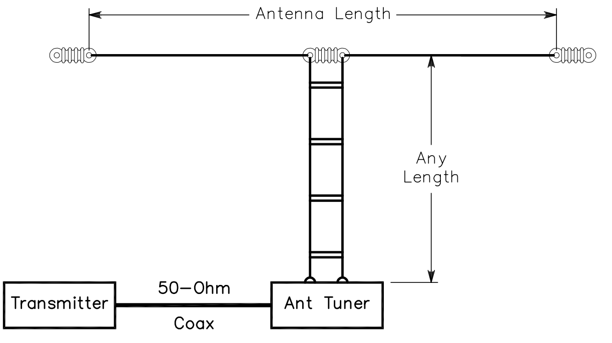

Fig 4—A center-fed antenna system for multiband use.

On the other hand, in the case of a multiband dipole where general coverage is desired, if the feed line happens to radiate, such energy could actually have a desirable effect. Antenna purists may dispute such a premise, but from a practical standpoint where you are not concerned with a directive pattern, much time and labor can be saved by ignoring possible transmission-line radiation.

THE 135-FOOT, 80 TO 10-METER DIPOLE

As mentioned previously, one of the most versatile antennas around is a simple dipole, center-fed with open-wire transmission line and used with an antenna tuner in the shack. A 135-foot long dipole hung horizontally between two trees or towers at a height of 50 feet or higher works very well on 80 through 10 meters. Such an antenna system has significant gain at the higher frequencies.

Flattop or Inverted-V Configuration?

There is no denying that the inverted-V mounting configuration (sometimes called a *drooping dipole*) is very convenient, since it requires only a single support. The flattop configuration, however, where the dipole is mounted horizontally, gives more gain at the higher frequencies. **Fig 5** shows the 80-meter azimuth and elevation patterns for two 135-foot long dipoles. The first is mounted as a flattop at a height of 50 feet over flat ground with a conductivity of 5 mS/m and a dielectric constant of 13, typical for average soil. The second dipole uses the same length of wire, with the center apex at 50 feet and the ends drooped down to be suspended 10 feet off the ground. This height is sufficient so that there is no danger to passersby from RF burns.

At 3.8 MHz, the flattop dipole about 4 dB more peak gain than its drooping cousin. On the other hand, the inverted-V configuration gives a pattern that is more omnidirectional than the flattop dipole, which has nulls off the ends of the wire. Omnidirectional coverage may be more important to net operators, for example, than maximum gain.

Fig 6 shows the azimuth and elevation patterns for the same two antenna configurations, but this time at 14.2 MHz. The flattop dipole has developed four distinct lobes at a 10° elevation angle, an angle typical for 20-meter skywave communication. The peak elevation angle gain of 9.4 dBi occurs at about 17° for a height of 50 feet above flat ground for the flattop dipole. The inverted-V configuration is again nominally more omnidirectional, but the peak gain is down some 6 dB from the flattop.

The situation gets even worse in terms of peak gain at 28.4 MHz for the inverted-V configuration. Here the peak gain is down about 8 dB from that produced by the flattop dipole, which exhibits eight lobes at this frequency with a maximum gain of 10.5 dBi at about 7° elevation. See the comparisons in **Fig 7**.

Whatever configuration you choose to mount the 135-foot dipole, you will want to feed it with some sort of low-loss open-wire transmission line. So-called *window* 450-Ω ladder line is popular for this application. Be sure to twist the line about three or four turns per foot to keep it from twisting excessively in the wind. Make sure also that you provide some mechanical support for the line at the junction with the dipole

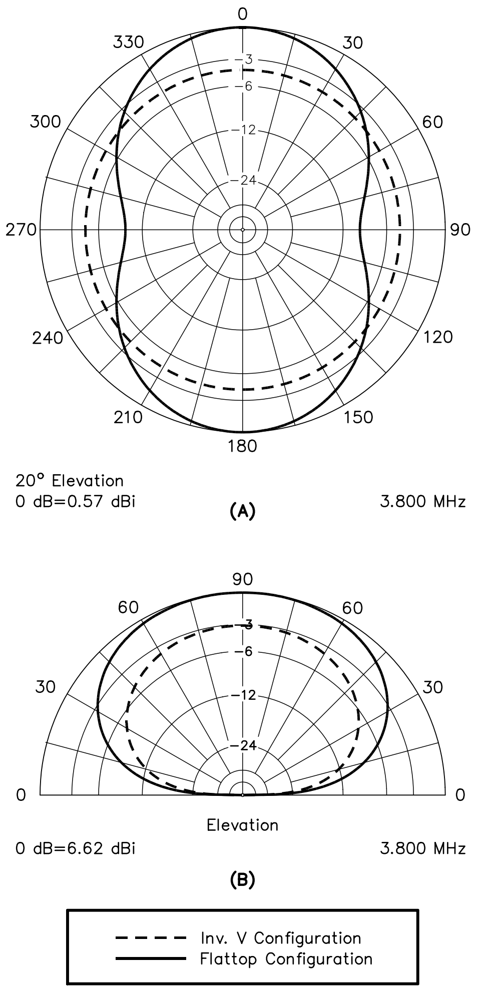

Fig 5—Patterns on 80 meters for 135-foot, center-fed dipole erected as a horizontal flattop dipole at 50 feet, compared with the same dipole installed as an inverted V with the apex at 50 feet and the ends at 10 feet. The azimuth pattern is shown at A, where the dipole wire lies in the 90° to 270° plane. At B, the elevation pattern, the dipole wire comes out of the paper at a right angle. On 80 meters, the patterns are not markedly different for either flattop or inverted-V configuration.

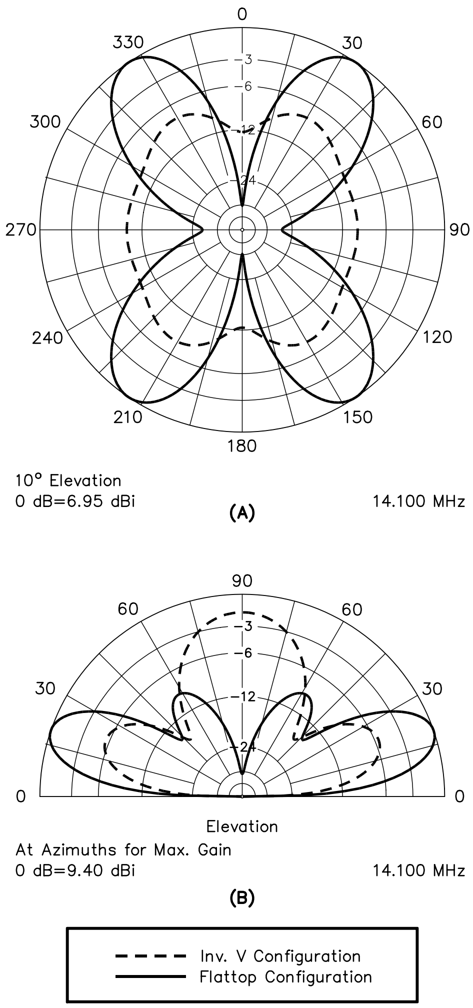

Fig 6—Patterns on 20 meters for two 135-foot dipoles. One is mounted horizontally as a flattop and the other as an inverted V with 120° included angle between the legs. The azimuth pattern is shown in A and the elevation pattern is shown in B. The inverted V has about 6 dB less gain at the peak azimuths, but has a more uniform, almost omnidirectional, azimuthal pattern. In the elevation plane, the inverted V has a fat lobe overhead, making it a somewhat better antenna for local communication, but not quite so good for DX contacts at low elevation angles.

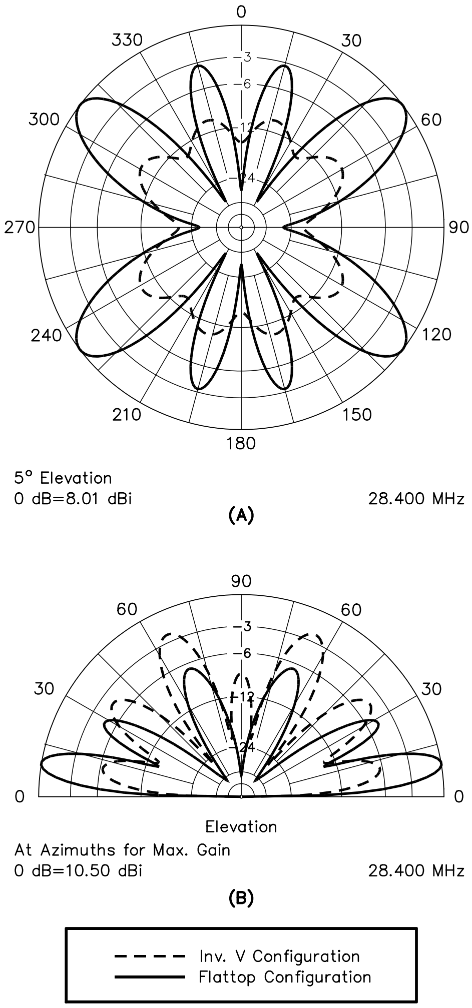

Fig 7—Patterns on 10 meters for same antenna configurations as in Figs 7 and 8. Once again, the inverted-V configuration yields a more omnidirectional pattern, but at the expense of almost 8 dB less gain than the flattop configuration at its strongest lobes.

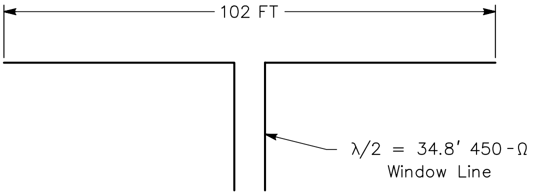

Fig 8—The G5RV multiband antenna covers 3.5 through 30 MHz. Although many amateurs claim it may be fed directly with 50-Ω coax on several amateur bands, Louis Varney, its originator, recommends the use of a matching network on bands other than 14 MHz.

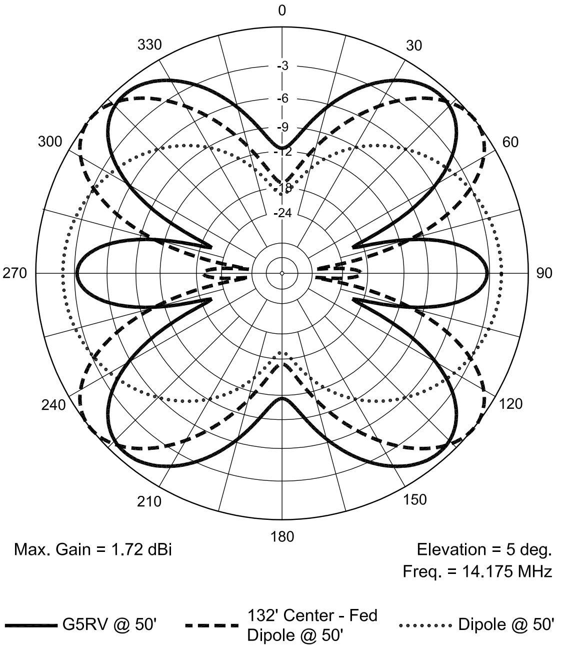

Fig 9—Azimuth pattern at a 5° takeoff angle for a 102-foot long, 50-foot high G5RV dipole (solid line). For comparison, the response for a 132-foot long, center-fed dipole at 50 feet height (dashed line) and a 33-foot long half wave 20-meter dipole at 50 feet (dotted line) are also shown. The longest antenna exhibits about 0.5 dB more gain than the G5RV, although the response is more omnidirectional for the G5RV—an advantage for a wire antenna that is not usually rotatable.

wires. This will prevent flexing of the transmission-line wire, since excessive flexing will result in breakage.

THE G5RV MULTIBAND ANTENNA

A multiband antenna that does not require a lot of space, is simple to construct, and is low in cost is the G5RV. Designed in England by Louis Varney (G5RV) some years ago, it has become quite popular in the US. The G5RV design is shown in **Fig 8**. The antenna may be used from 3.5 through 30 MHz. Although some amateurs claim it may be fed directly with 50-Ω coax on several amateur bands with a low SWR, Varney himself recommended the use of an antenna tuner on bands other than 14 MHz (see Bibliography). In fact, an analysis of the G5RV feed-point impedance shows there is *no* length of balanced line of *any* characteristic impedance that will transform the terminal impedance to the 50 to 75-Ω range on all bands. (Low SWR indication with coax feed and no matching network on bands other than 14 MHz may indicate excessive losses in the coaxial line.)

Fig 9 shows the 20-meter azimuthal pattern for a G5RV at a height of 50 feet over flat ground, at an elevation angle of 5° that is suitable for DX work. For comparison, the response for two other antennas is also shown in Fig 9—a standard half wave 20-meter dipole at 50 feet and a 132-foot long center-fed dipole at 50 feet. The G5RV on 20 meters is, of course, longer than a standard half wave dipole and it exhibits about 2 dB more gain compared to that dipole. With four lobes making it look rather like a four-leaf clover, the azimuth pattern is more omnidirectional than the two-lobed dipole. The 132-foot center-fed dipole is longer than the G5RV and it has about 0.5 dB more gain than the G5RV, also exhibiting four major lobes, along with two strong minor lobes in the plane of the wire. Overall, the azimuthal response for the G5RV is more omnidirectional than the comparison antennas.

The G5RV patterns for other frequencies are similar to those shown for the 135-foot dipole previously for other frequencies. Incidentally, you may be wondering why a 132-foot dipole is shown in Fig 9, rather than the 135-foot dipole described earlier. The 132-foot overall length describes another antenna that we'll discus in the next section on Windom antennas.

The portion of the G5RV antenna shown as horizontal in Fig 8 may also be installed in an inverted-V dipole arrangement, subject to the same loss of peak gain mentioned above for the 135-foot dipole. Or instead, up to ⅙ of the total length of the antenna at each end may be dropped vertically, semi-vertically, or bent at a convenient angle to the main axis of the antenna, to cut down on the requirements for real estate.

THE WINDOM ANTENNA

An antenna that enjoyed popularity in the 1930s and into the 1940s was what we now call the *Windom*. It was known at the time as a "single-feeder Hertz" antenna, after being described in Sep 1929 *QST* by Loren G. Windom, W8GZ (see Bibliography).

The Windom antenna, shown in **Fig 10**, is fed with a single wire, attached approximately 14% off center. In theory, this location provides a match for the single-wire transmission line, which is worked against an earth ground. Because the single-wire feed line is not inherently well balanced and because it is brought to the operating position, "RF in the shack" and a potential radiation hazard may be experienced with this antenna.

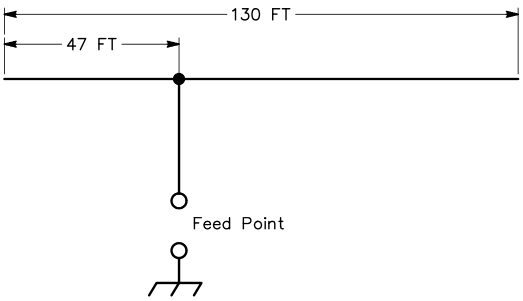

Fig 10—The Windom antenna, cut for a fundamental frequency of 3.75 MHz. The single-wire feeder, connected 14% off center, is brought into the station and the system is fed against ground. The antenna is also effective on its harmonics.

Later variations of the off-center fed Windom moved the attachment point slightly to accommodate balanced 300-Ω ribbon line. One relatively recent variation is called the "Carolina Windom," apparently because two of the designers, Edgar Lambert, WA4LVB, and Joe Wright, W4UEB, lived in coastal North Carolina (the third, Jim Wilkie, WY4R, lived

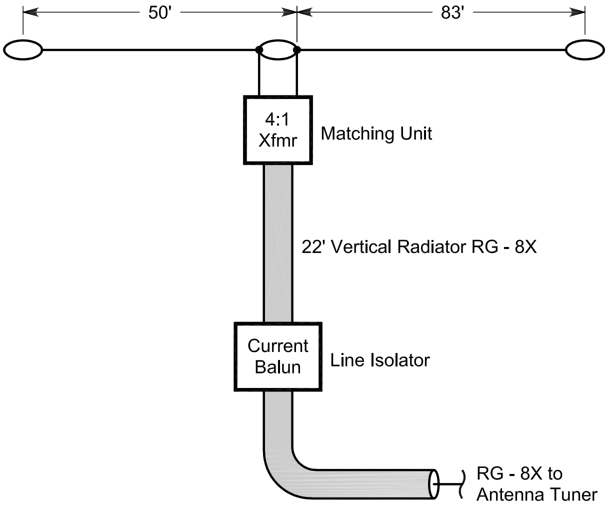

Fig 11—Layout for flattop "Carolina Windom" antenna.

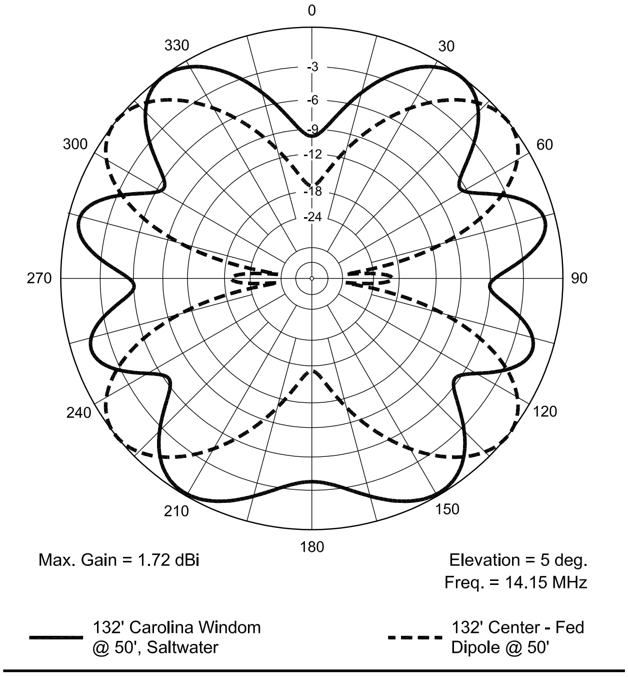

Fig 12—20-meter azimuth patterns for a 132-foot long off-center fed Carolina Windom and a 132-foot long center-fed flattop dipole on 20 meters, both at a height of 50 feet above saltwater. The response for the Carolina Windom is more omnidirectional because the vertically polarized radiation from the 22-foot long vertical RG-8X coax fills in the deep nulls.

in nearby Norfolk, Virginia). One of the interesting parts about the Carolina Windom is that it turns a potential disadvantage—feed line radiation—into a potential advantage.

Fig 11 is a diagram of a flattop Carolina Windom, which uses a 50-foot wire joined with an 83-foot wire at the feed-point insulator. This resembles the layout shown in Fig 10 for the original W8GZ Windom. The "Vertical Radiator" for the Carolina Windom is a 22-foot piece of RG-8X coax, with a "Line Isolator" (current-type choke balun) at the bottom end and a 4:1 "Matching Unit" at the top. The system takes advantage of the asymmetry of the horizontal wires to induce current onto the braid of the vertical coax section. Note that the matching unit is a voltage-type balun transformer, which purposely does not act like a common-mode current choking balun. You must use an antenna tuner with this system to present a 1:1 SWR to the transmitter on the amateur bands from 80 through 10 meters.

The radiation resulting from current induced onto the 22-foot vertical coax section tends to fill in the deep nulls that would be present if the 132-feet of horizontal wire were symmetrically center fed. Over saltwater, the vertical radiator can give significant gain at the low elevation angles needed for DX work. Indeed, field reports for the Carolina Windom are most impressive for stations located near or on saltwater. Over average soil the advantage of the additional vertically polarized component is not quite so evident. **Fig 12** compares a 50-foot high Carolina Windom on 14 MHz over saltwater to a 50-foot high, 132-foot long, flattop center-fed dipole. The Carolina Windom has a more omnidirectional azimuthal pattern, a desirable characteristic in a 132-foot long wire antenna that is not normally rotated to favor different directions.

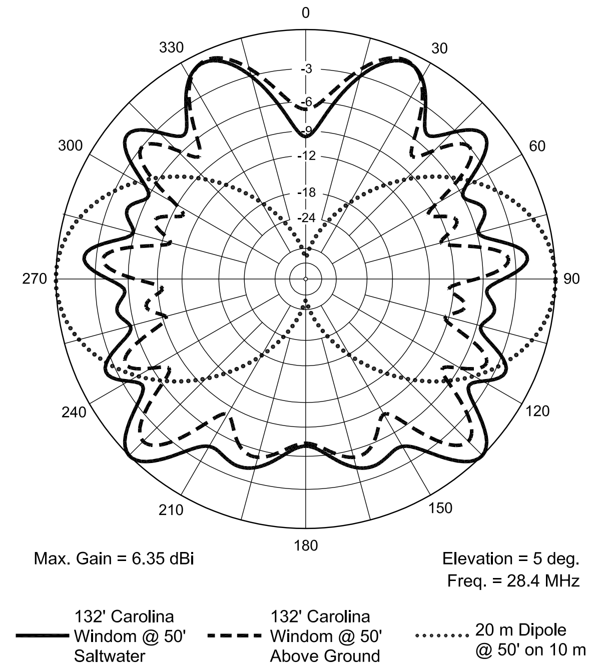

Fig 13—10-meter azimuthal responses for a 132-foot long, 50-foot high Carolina Windom over saltwater (solid line) and over average ground (dashed line), compared to that for a 20-meter half-wave dipole at 50 feet (dotted line).

Another advantage of the Carolina Windom over a traditional Windom is that the coax feed line hanging below the common-mode current choke does not radiate, meaning that there will be less "RF in the shack." Since the feed line is not always operating at a low SWR on various ham bands, use the minimum length of feed coax possible to hold down losses in the coax.

Fig 13 shows the azimuth responses for a 50-foot long flattop Carolina Windom on 28.4 MHz over saltwater and over average soil. The pattern for a 50-foot high, flattop 20-meter dipole operated on 28.4 MHz is also shown, since this 20-meter dipole can also be used as a multiband antenna, when fed with open-wire transmission line rather than with coax. Again, the Carolina Windom exhibits a more omnidirectional pattern, even if the pattern is somewhat lopsided at the bottom.

MULTIPLE-DIPOLE ANTENNAS

The antenna system shown in **Fig 14** consists of a group of center-fed dipoles, all connected in parallel at the point where the transmission line joins them. The dipole elements are *stagger-tuned*. That is, they are individually cut to be λ/2 at different frequencies. Chapter 9, Broadband Antenna Matching, discusses stagger tuning of dipole antennas to attain a low SWR across a broad range of frequencies. An extension of the stagger tuning idea is to construct multiwire dipoles cut for different bands.

In theory, the 4-wire antenna of Fig 14 can be used with a coaxial feeder on five bands. The four wires are prepared as parallel-fed dipoles for 3.5, 7, 14, and 28 MHz. The 7-MHz dipole can be operated on its 3rd harmonic for 21-MHz operation to cover a fifth band. However, in practice it has been found difficult to get a good match to coaxial line

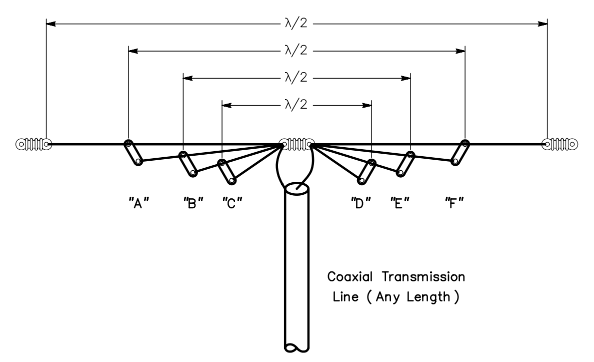

Fig 14—Multiband antenna using paralleled dipoles all connected to a common low-impedance transmission line. The half-wave dimensions may be either for the centers of the various bands or selected to fit favorite frequencies in each band. The length of a half wave in feet is 468/ frequency in MHz, but because of interaction among the various elements, some pruning for resonance may be needed on each band.

on all bands. The λ/2 resonant length of any one dipole in the presence of the others is not the same as for a dipole by itself due to interaction, and attempts to optimize all four lengths can become a frustrating procedure. The problem is compounded because the optimum tuning changes in a different antenna environment, so what works for one amateur may not work for another. Even so, many amateurs with limited antenna space are willing to accept the mismatch on some bands just so they can operate on those frequencies using a single coax feed line.

Since this antenna system is balanced, it is desirable to use a balanced transmission line to feed it. The most desirable type of line is 75-Ω transmitting twin-lead. However, either 52-Ω or 75-Ω coaxial line can be used. Coax line introduces some unbalance, but this is tolerable on the lower frequencies. An alternative is to use a balun at the feed point, fed with coaxial cable.

The separation between the dipoles for the various frequencies does not seem to be especially critical. One set of wires can be suspended from the next larger set, using insulating spreaders (of the type used for feeder spreaders) to give a separation of a few inches. Users of this antenna often run some of the dipoles at right angles to each other to help reduce interaction. Some operators use inverted-V-mounted dipoles as guy wires for the mast that supports the antenna system.

An interesting method of construction used successfully by Louis Richard, ON4UF, is shown in **Fig 15**. The antenna has four dipoles (for 7, 14, 21 and 28 MHz) constructed from 300-Ω ribbon transmission line. A single length of ribbon makes two dipoles. Thus, two lengths, as shown in the sketch, serve to make dipoles for four bands. Ribbon with copper-clad steel conductors (Amphenol type 14-022) should be used because all of the weight, including that of the feed line, must be supported by the uppermost wire.

Two pieces of ribbon are first cut to a length suitable for the two halves of the longest dipole. Then one of the conductors in each piece is cut to proper length for the next band higher in frequency. The excess wire and insulation is stripped away. A second pair of lengths is prepared in the same manner, except that the lengths are appropriate for the next two higher frequency bands.

A piece of thick polystyrene sheet drilled with holes for anchoring each wire serves as the central insulator. The shorter pair of dipoles is suspended the width of the ribbon below the longer pair by clamps also made of poly sheet. Intermediate spacers are made by sawing slots in pieces of poly sheet so they will fit the ribbon snugly.

The multiple-dipole principle can also be applied to vertical antennas. Parallel or fanned λ/4 elements of wire or tubing can be worked against ground or tuned radials from a common feed point.

OFF-CENTER-FED DIPOLES

Fig 16 shows an off-center-fed or *OCF* dipole. Because it is similar in appearance to the Windom of Fig 12, this antenna is often mistakenly called a “Windom,” or sometimes a “coax-fed Windom.” The two antennas are not the same, since the Windom is worked against its image in the ground, while one leg is worked against the other in the OCF dipole.

It is not necessary to feed a dipole antenna at its center, although doing so will allow it to be operated with a relatively low feed-point impedance on its fundamental and *odd* harmonics. (For example, a 7-MHz center-fed half-wave dipole can also be used for 21-MHz operation.) By contrast, the OCF dipole of Fig 16, fed ⅓ of its length from one end, may be used on its fundamental and *even* harmonics. Its free-space antenna-terminal impedance at 3.5, 7 and 14 MHz is on the order of 150 to 200 Ω. A 1:4 step-up transformer at the feed point should offer a reasonably good match to 50- or 75-Ω line, although some commercially made OCF dipoles use a 1:6 transformer.

At the 6th harmonic, 21 MHz, the antenna is three wavelengths long and fed at a voltage loop (maximum), instead of a current loop. The feed-point impedance at this frequency

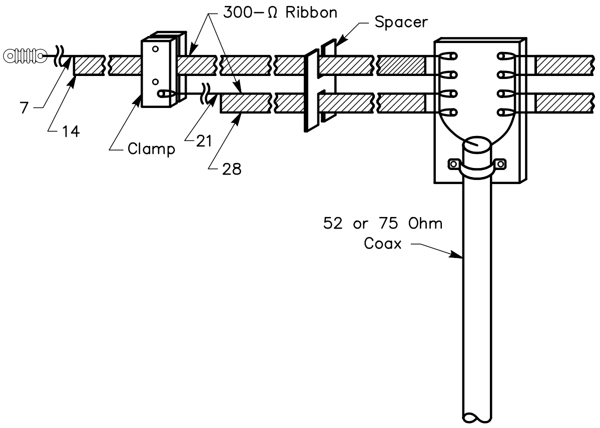

Fig 15—Sketch showing how the twin-lead multiple-dipole antenna system is assembled. The excess wire and insulation are stripped away.

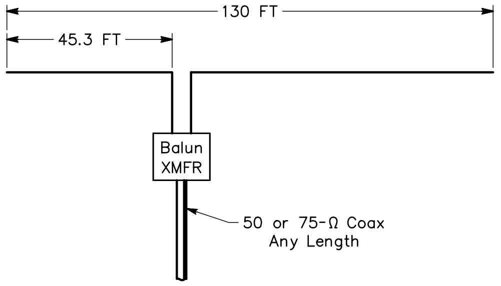

Fig 16—The off-center-fed (OCF) dipole for 3.5, 7 and 14 MHz. A 1:4 or 1:6 step-up current balun is used at the feed point.

is high, a few thousand ohms, so the antenna is unsuitable for use on this band.

Balun Requirements

Because the OCF dipole is not fed at the center of the radiator, the RF impedance paths of the two wires at the feed point are unequal. If the antenna is fed directly with coax (or a balanced line), or if a voltage step-up transformer is used, then voltages of equal magnitude (but opposite polarity) are applied to the wires at the feed point. Because of unequal impedances, the resulting antenna currents flowing in the two wires will not be equal. This also means that antenna current can flow on the feeder—on the *outside* of a coaxial line. (You may recall that this is how the Carolina Windom works, actually inducing current onto a carefully chosen length of coax, choked at its bottom end, so that it acts as a vertical radiator.)

How much current flows on the coax shield depends on the impedance of the RF current path down the outside of the feed line. In general, this is not a desirable situation. To prevent radiation, equal *currents* are required at the feed point, with the same current flowing in and out of the short leg as in and out of the long leg of the radiator. A *current* or *choke* type of balun provides just such operation. (Current baluns are discussed in detail in Chapter 26, Coupling the Line to the Antenna.)

Trap Antennas

By using tuned circuits of appropriate design strategically placed in a dipole, the antenna can be made to show what is essentially fundamental resonance at a number of different frequencies. The general principle is illustrated by **Fig 17**.

Even though a trap-antenna arrangement is a simple one, an explanation of how a trap antenna works can be elusive. For some designs, traps are resonated in our amateur bands, and for others (especially commercially made antennas) the traps are resonant far outside any amateur band.

A trap in an antenna system can perform either of two functions, depending on whether or not it is resonant at the operating frequency. A familiar case is where the trap is parallel-resonant in an amateur band. For the moment, let us assume that dimension A in Fig 17 is 32 feet and that each L/C combination is resonant in the 7-MHz band. Because of its parallel resonance, the trap presents a high impedance at that point in the antenna system. The electrical effect at 7 MHz is that the trap behaves as an insulator. It serves to divorce the outside ends, the B sections, from the antenna. The result is easy to visualize—we have an antenna system that is resonant in the 7-MHz band. Each 33-foot section (labeled A in the drawing) represents λ/4, and the trap behaves as an insulator. We therefore have a full-size 7-MHz antenna.

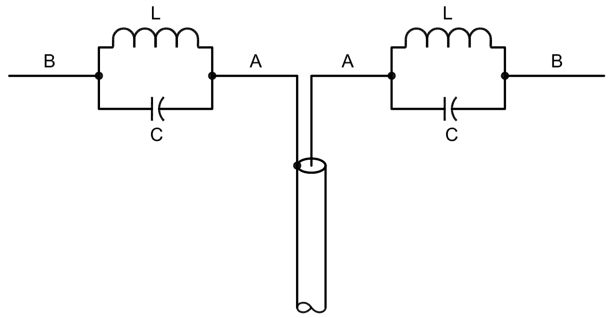

Fig 17—A trap dipole antenna. This antenna may be fed with 50-Ω coaxial line. Depending on the L/C ratio of the trap elements and the lengths chosen for dimensions A and B, the traps may be resonant either in an amateur band or at a frequency far removed from an amateur band for proper two-band antenna operation.

The second function of a trap, obtained when the frequency of operation is *not* the resonant frequency of the trap, is one of electrical loading. If the operating frequency is below that of trap resonance, the trap behaves as an inductor; if above, as a capacitor. Inductive loading will electrically lengthen the antenna, and capacitive loading will electrically shorten the antenna.

Let's carry our assumption a bit further and try using the antenna we just considered at 3.5 MHz. With the traps resonant in the 7-MHz band, they will behave as inductors when operation takes place at 3.5 MHz, electrically lengthening the antenna. This means that the total length of sections A and B (plus the length of the inductor) may be something less than a physical λ/4 for resonance at 3.5 MHz. Thus, we have a two-band antenna that is shorter than full size on the lower frequency band. But with the electrical loading provided by the traps, the overall electrical length is λ/2. The total antenna length needed for resonance in the 3.5-MHz band will depend on the L/C ratio of the trap elements.

The key to trap operation off resonance is its L/C ratio, the ratio of the value of L to the value of C. At resonance, however, within practical limitations the L/C ratio is immaterial as far as electrical operation goes. For example, in the antenna we've been discussing, it would make no difference for 7-MHz operation whether the inductor were 1 µH and the capacitor were 500 pF (the reactances would be just below 45 Ω at 7.1 MHz), or whether the inductor were 5 µH and the capacitor 100 pF (reactances of approximately 224 Ω at 7.1 MHz). But the choice of these values will make a significant difference in the antenna size for resonance at 3.5 MHz. In the first case, where the L/C ratio is 2000, the necessary length of section B of the antenna for resonance at 3.75 MHz would be approximately 28.25 feet. In the second case, where the L/C ratio is 50,000, this length need be only 24.0 feet, a difference of more than 15%.

The above example concerns a two-band antenna with trap resonance at one of the two frequencies of operation. On each of the two bands, each half of the dipole operates as an electrical λ/4. However, the same band coverage can be obtained with a trap resonant at, say, 5 MHz, a frequency quite removed from either amateur band. With proper selection of the L/C ratio and the dimensions for A and B, the trap will act to shorten the antenna electrically at 7 MHz and lengthen it electrically at 3.5 MHz. Thus, an antenna that is intermediate in physical length between being full size on 3.5 MHz and full size on 7 MHz can cover both bands, even though the trap is not resonant at either frequency. Again, the antenna operates with electrical λ/4 sections. Note that such non-resonant traps have less RF current flowing in the trap components, and hence trap losses are less than for resonant traps.

Additional traps may be added in an antenna section to cover three or more bands. Or a judicious choice of dimensions and the L/C ratio may permit operation on three or more bands with just a pair of identical traps in the dipole.

An important point to remember about traps is this. If the operating frequency is below that of trap resonance, the trap behaves as an inductor; if above, as a capacitor. The above discussion is based on dipoles that operate electrically as λ/2 antennas. This is not a requirement, however. Elements may be operated as electrical 3/2 λ, or even 5/2 λ, and still present a reasonable impedance to a coaxial feeder. In trap antennas covering several HF bands, using electrical lengths that are odd multiples of λ/2 is often done at the higher frequencies.

To further aid in understanding trap operation, let’s now choose trap L and C components that each have a reactance of 20 Ω at 7 MHz. Inductive reactance is directly proportional to frequency, and capacitive reactance is inversely proportional. When we shift operation to the 3.5-MHz band, the inductive reactance becomes 10 Ω, and the capacitive reactance becomes 40 Ω. At first thought, it may seem that the trap would become capacitive at 3.5 MHz with a higher capacitive reactance, and that the extra capacitive reactance would make the antenna electrically shorter yet. Fortunately, this is not the case. The inductor and the capacitor are connected in parallel with each other.

$$Z = \frac{-jX_L X_C}{X_L + X_C} \qquad \text{(Eq 1)}$$

where *j* indicates a reactive impedance component, rather than resistive. A positive result indicates inductive reactance, and a negative result indicates capacitive. In this 3.5-MHz case, with 40 Ω of capacitive reactance and 10 Ω of inductive, the equivalent series reactance is 13.3 Ω inductive. This inductive loading lengthens the antenna to an electrical λ/2 overall at 3.5 MHz, assuming the B end sections in Fig 17 are of the proper length.

With the above reactance values providing resonance at 7-MHz, X_L equals X_C, and the theoretical series equivalent is infinity. This provides the insulator effect, divorcing the ends.

At 14 MHz, where $X_L = 40\ \Omega$ and $X_C = 10\ \Omega$, the resultant series equivalent trap reactance is 13.3 Ω capacitive. If the total physical antenna length is slightly longer than 3/2 λ at 14 MHz, this trap reactance at 14 MHz can be used to shorten the antenna to an electrical 3/2 λ. In this way, 3-band operation is obtained for 3.5, 7 and 14 MHz with just one pair of identical traps. The design of such a system is not straightforward, however, for any chosen L/C ratio for a given total length affects the resonant frequency of the antenna on both the 3.5 and 14-MHz bands.

Trap Losses

Since the tuned circuits have some inherent losses, the efficiency of a trap system depends on the unloaded Q values of the tuned circuits. Low-loss (high-Q) coils should be used, and the capacitor losses likewise should be kept as low as possible. With tuned circuits that are good in this respect—comparable with the low-loss components used in transmitter tank circuits, for example—the reduction in efficiency compared with the efficiency of a simple dipole is small, but tuned circuits of low unloaded Q can lose an appreciable portion of the power supplied to the antenna.

The commentary above applies to traps assembled from conventional components. The important function of a trap that is resonant in an amateur band is to provide a high isolating impedance, and this impedance is directly proportional to Q. Unfortunately, high Q restricts the antenna bandwidth, because the traps provide maximum isolation only at trap resonance.

FIVE-BAND W3DZZ TRAP ANTENNA

C. L. Buchanan, W3DZZ, created one of the first trap antennas for the five pre-1979 WARC amateur bands from 3.5 to 30 MHz. Dimensions are given in **Fig 18**. Only one set of traps is used, resonant at 7 MHz to isolate the inner (7-MHz) dipole from the outer sections. This causes the overall system to be resonant in the 3.5-MHz band. On 14, 21 and 28 MHz the antenna works on the capacitive-reactance principle just outlined. With a 75-Ω twin-lead feeder, the SWR with this antenna is under 2:1 throughout the three highest frequency

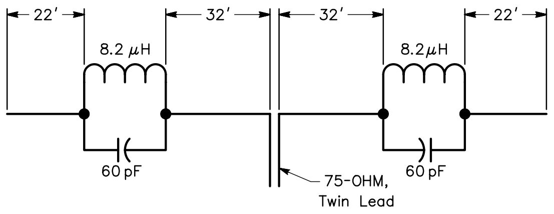

Fig 18—Five-band (3.5, 7, 14, 21 and 28 MHz) trap dipole for operation with 75-Ω feeder at low SWR (C. L. Buchanan, W3DZZ). The balanced (parallel-conductor) line indicated is desirable, but 75-Ω coax can be substituted with some sacrifice of symmetry in the system. Dimensions given are for resonance (lowest SWR) at 3.75, 7.2, 14.15 and 29.5 MHz. Resonance is very broad on the 21-MHz band, with SWR less than 2:1 throughout the band.

bands, and the SWR is comparable with that obtained with similarly fed simple dipoles on 3.5 and 7 MHz.

Trap Construction

Traps frequently are built with coaxial aluminum tubes (usually with polystyrene tubing in-between them for insulation) for the capacitor, with the coil either self-supporting or wound on a form of larger diameter than the tubular capacitor. The coil is then mounted coaxially with the capacitor to form a unit assembly that can be supported at each end by the antenna wires. In another type of trap devised by William J. Lattin, W4JRW (see Bibliography at the end of this chapter), the coil is supported inside an aluminum tube and the trap capacitor is obtained in the form of capacitance between the coil and the outer tube. This type of trap is inherently weatherproof.

A simpler type of trap, easily assembled from readily available components, is shown in **Fig 19**. A small transmitting-type ceramic "doorknob" capacitor is used, together with a length of commercially available coil material, these being supported by an ordinary antenna strain insulator. The circuit constants and antenna dimensions differ slightly from those of Fig 18, in order to bring the antenna resonance points closer to the centers of the various phone bands. Construction data are given in **Fig 20**. If a 10-turn length of inductor is used, a half turn from each end may be used to slip through the anchor holes in the insulator to act as leads.

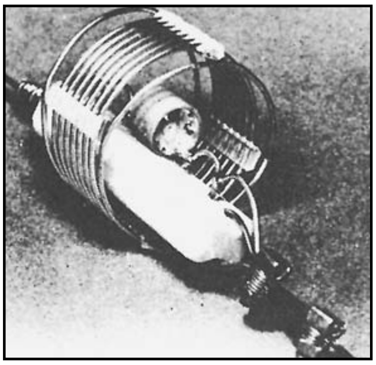

Fig 19—Easily constructed trap for wire antennas (A. Greenburg, W2LH). The ceramic insulator is 4¼ inches long (Birnback 688). The clamps are small service connectors available from electrical supply and hardware stores (Burndy KS90 servits).

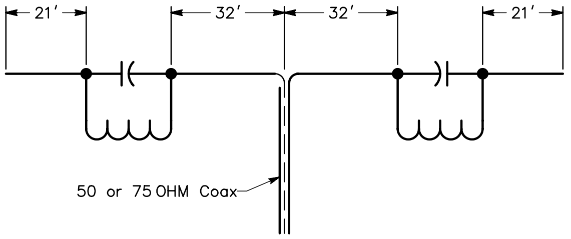

Fig 20—Layout of multiband antenna using traps constructed as shown in Fig 21. The capacitors are 100 pF each, transmitting type, 5000-volt dc rating (Centralab 850SL-100N). Coils are 9 turns of #12 wire, 2½ inches diameter, 6 turns per inch (B&W 3029) with end turns spread as necessary to resonate the traps to 7.2 MHz. These traps, with the wire dimensions shown, resonate the antenna at approximately the following frequencies on each band: 3.9, 7.25, 14.1, 21.5 and 29.9 MHz (based on measurements by W9YJH).

The components used in these traps are sufficiently weatherproof in themselves so that no additional weatherproofing has been found necessary. However, if it is desired to protect them from the accumulation of snow or ice, a plastic cover can be made by cutting two discs of polystyrene slightly larger in diameter than the coil, drilling at the center to pass the antenna wires, and cementing a plastic cylinder on the edges of the discs. The cylinder can be made by wrapping two turns or so of 0.02-inch poly or Lucite sheet around the discs, if no suitable ready-made tubing is available. Plastic drinking glasses and 2-liter soft-drink plastic bottles are easily adaptable for use as impromptu trap covers.

TWO W8NX MULTIBAND, COAX-TRAP DIPOLES

Over the last 60 or 70 years, amateurs have used many kinds of multiband antennas to cover the traditional HF bands. The availability of the 30, 17 and 12-meter bands has expanded our need for multiband antenna coverage.

Two different antennas are described here. The first covers the traditional 80, 40, 20, 15 and 10-meter bands, and the second covers 80, 40, 17 and 12 meters. Each uses the same type of W8NX trap—connected for different modes of operation—and a pair of short capacitive stubs to enhance coverage. The W8NX coaxial-cable traps have two different modes: a high- and a low-impedance mode. The inner-conductor windings and shield windings of the traps are connected in series for both modes. However, either the low- or high-impedance point can be used as the trap's output terminal. For low-impedance trap operation, only the center conductor turns of the trap windings are used. For high-impedance operation, all turns are used, in the conventional manner for a trap. The short stubs on each antenna are strategically sized and located to permit more flexibility in adjusting thc resonant frequencies of the antenna.

80, 40, 20, 15 and 10-Meter Dipole

Fig 21 shows the configuration of the 80, 40, 20, 15 and 10-meter antenna. The radiating elements are made of #14

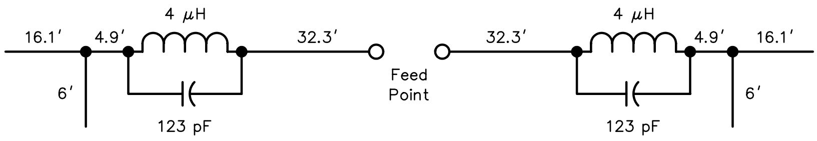

Fig 21—A W8NX multiband dipole for 80, 40, 20, 15 and 10 meters. The values shown (123 pF and 4 μH) for the coaxial-cable traps are for parallel resonance at 7.15 MHz. The low-impedance output of each trap is used for this antenna.

stranded copper wire. The element lengths are the wire span lengths in feet. These lengths do not include the lengths of the pigtails at the balun, traps and insulators. The 32.3-foot-long inner 40-meter segments are measured from the eyelet of the input balun to the tension-relief hole in the trap coil form. The 4.9-foot segment length is measured from the tension-relief hole in the trap to the 6-foot stub. The 16.l-foot outer-segment span is measured from the stub to the eyelet of the end insulator.

The coaxial-cable traps are wound on PVC pipe coil forms and use the low-impedance output connection. The stubs are 6-foot lengths of ⅛-inch stiffened aluminum or copper rod hanging perpendicular to the radiating elements. The first inch of their length is bent 90° to permit attachment to the radiating elements by large-diameter copper crimp connectors. Ordinary #14 wire may be used for the stubs, but it has a tendency to curl up and may tangle unless weighed down at the end. You should feed the antenna with 75-Ω coax cable using a good 1:1 balun.

This antenna may be thought of as a modified W3DZZ antenna due to the addition of the capacitive stubs. The length and location of the stub give the antenna designer two extra degrees of freedom to place the resonant frequencies within the amateur bands. This additional flexibility is particularly helpful to bring the 15 and 10-meter resonant frequencies to more desirable locations in these bands. The actual 10-meter resonant frequency of the original W3DZZ antenna is somewhat above 30 MHz, pretty remote from the more desirable low frequency end of 10 meters.

80, 40, 17 and 12-Meter Dipole

Fig 22 shows the configuration of the 80, 40, 17 and 12-meter antenna. Notice that the capacitive stubs are attached immediately outboard after the traps and are 6.5 feet long, ½ foot longer than those used in the other antenna. The traps are the same as those of the other antenna, but are connected for the high-impedance parallel-resonant output mode. Since only four bands are covered by this antenna, it is easier to fine tune it to precisely the desired frequency on all bands. The 12.4-foot tips can he pruned to a particular 17-meter frequency with little effect on the 12-meter frequency. The stub lengths can be pruned to a particular 12-meter frequency with little effect on the 17-meter frequency. Both such pruning adjustments slightly alter the 80-meter resonant frequency. However, the bandwidths of the antennas are so broad on 17 and 12 meters that little need for such pruning exists. The 40-meter frequency is nearly independent of adjustments to the capacitive stubs and outer radiating tip elements. Like the first antennas, this dipole is fed with a 75-Ω balun and feed line.

Fig 23 shows the schematic diagram of the traps. It explains the difference between the low and high-impedance modes of the traps. Notice that the high-impedance terminal is the output configuration used in most conventional trap applications. The low-impedance connection is made across only the inner conductor turns, corresponding to one-half of the total turns of the trap. This mode steps the trap's impedance down to approximately one-fourth of that of the high-impedance level. This is what allows a single trap design to be used for two different multiband antennas.

Fig 24 is a drawing of a cross-section of the coax trap shown through the long axis of the trap. Notice that the traps are conventional coaxial-cable traps, except for the added low-impedance output terminal. The traps are 8¾ close-spaced turns of RG-59 (Belden 8241) on a 2⅜-inch-OD PVC pipe (schedule 40 pipe with a 2-inch ID) coil form. The forms are 4⅛ inches long. Trap resonant frequency is very sensitive to the outer diameter of the coil form, so check it carefully. Unfortunately, not all PVC pipe is made with the same wall thickness. The trap frequencies should be checked with a dip meter and general-coverage receiver and adjusted to within 50 kHz of the 7150 kHz resonant frequency before installation. One inch is left over at each end of the coil forms to allow for the coax feed-through holes and holes for tension-relief attachment of the antenna radiating elements to the traps. Be sure to seal the ends of the trap coax cable with RTV sealant to prevent moisture from entering the coaxial cable.

Also, be sure that you connect the 32.3-foot wire element at the start of the inner conductor winding of the trap. This avoids detuning the antenna by the stray capacitance of the coaxial-cable shield. The trap output terminal (which has the shield stray capacitance) should be at the outboard side of

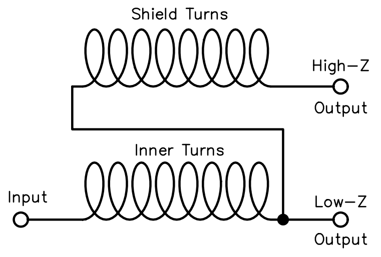

Fig 23—Schematic for the W8NX coaxial-cable trap. RG-59 is wound on a 2⅜-inch OD PVC pipe.

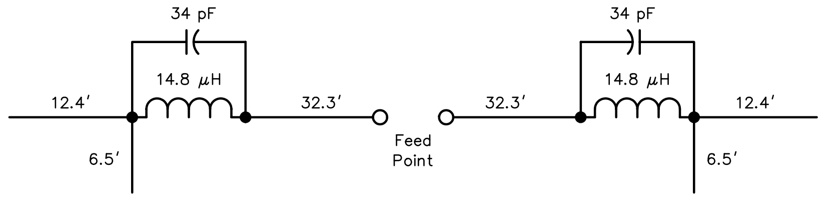

Fig 22—A W8NX multiband dipole for 80, 40, 17 and 12 meters. For this antenna, the high-impedance output is used on each trap. The resonant frequency of the traps is 7.15 MHz.

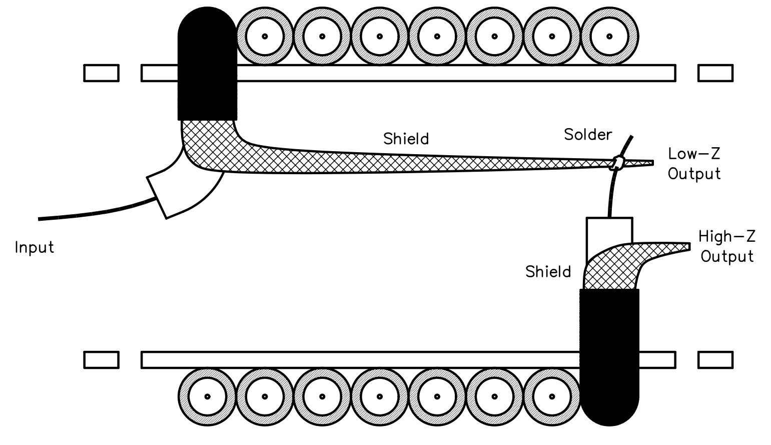

Fig 24—Construction details of the W8NX coaxial-cable trap.

the trap. Reversing the input and output terminals of the trap will lower the 40-meter frequency by approximately 50 kHz, but there will be negligible effect on the other bands.

Fig 25 shows a coaxial-cable trap. Further details of the trap installation are shown in **Fig 26**. This drawing applies specifically to the 80, 40, 20, 15 and l0-meter antenna, which uses the low-impedance trap connections. Notice the lengths of the trap pigtails: 3 to 4 inches at each terminal of the trap. If you use a different arrangement, you must modify the span lengths accordingly. All connections can be made using crimp connectors rather than by soldering. Access to the trap's interior is attained more easily with a crimping tool than with a soldering iron.

Performance

The performance of both antennas has been very satisfactory. W8NX uses the 80, 40, 17 and 12-meter version because it covers 17 and 12 meters. (He has a tribander for 20, 15 and 10 meters.) The radiation pattern on 17 meters is that of a $\frac{3}{2}$-wave dipole. On 12 meters, the pattern is that of a $\frac{5}{2}$-wave dipole. At his location in Akron, Ohio, the antenna runs essentially east and west. It is installed as an inverted V, 40 feet high at the center, with a 120° included angle between the legs. Since the stubs are very short, they radiate little power and make only minor contributions to the radiation patterns. In theory, the pattern has four major lobes on 17 meters, with maxima to the northeast, southeast, southwest and northwest. These provide low-angle radiation into Europe, Africa, South Pacific, Japan and Alaska. A narrow pair of minor broadside lobes provides north and south coverage into Central America, South America and the polar regions.

There are four major lobes on 12 meters, giving nearly end-fire radiation and good low-angle east and west coverage. There are also three pairs of very narrow, nearly broadside, minor lobes on 12 meters, down about 6 dB from the major end-fire lobes. On 80 and 40 meters, the antenna has the usual figure-8 patterns of a half-wave-length dipole.

Both antennas function as electrical half-wave dipoles on 80 and 40 meters with a low SWR. They both function as odd-harmonic current-fed dipoles on their other operating frequencies, with higher, but still acceptable, SWR. The presence of the stubs can either raise or lower the input impedance of the antenna from those of the usual third and fifth harmonic dipoles. Again W8NX recommends that 75-Ω,

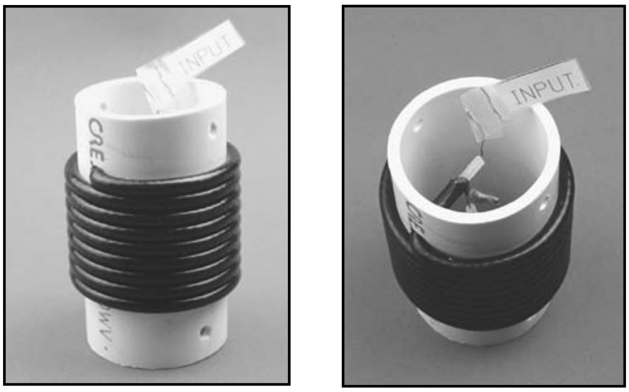

Fig 25—Other views of a W8NX coax-cable trap.

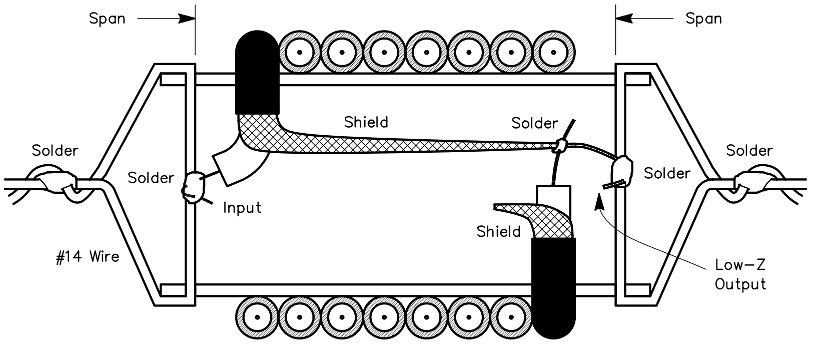

Fig 26—Additional construction details for the W8NX coaxial-cable trap.

rather than 50-Ω, feed line be used because of the generally higher input impedances at the harmonic operating frequencies of the antennas.

The SWR curves of both antennas were carefully measured using a 75 to 50-Ω transformer from Palomar Engineers inserted at the junction of the 75-Ω coax feed line and a 50-Ω SWR bridge. The transformer is required for accurate SWR measurement if a 50-Ω SWR bridge is used with a 75-Ω line. Most 50-Ω rigs operate satisfactorily with a 75-Ω line, although this requires different tuning and load settings in the final output stage of the rig or antenna tuner. The author uses the 75 to 50-Ω transformer only when making SWR measurements and at low power levels. The transformer is rated for 100 W, and when he runs his 1-kW PEP linear amplifier the transformer is taken out of the line.

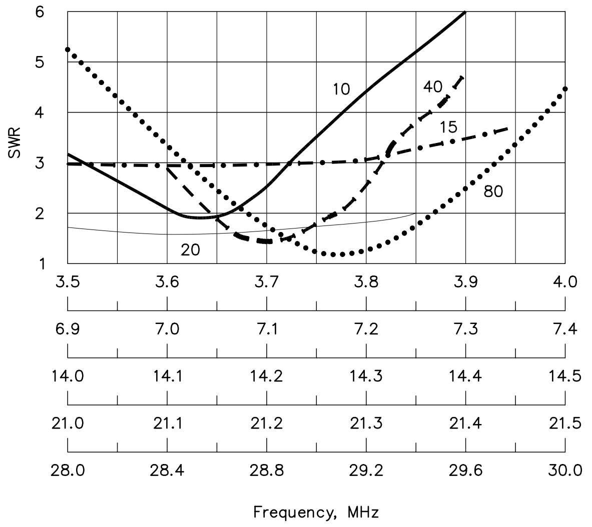

Fig 27—Measured SWR curves for an 80, 40, 20, 15 and 10-meter antenna, installed as an inverted-V with 40-ft apex and 120° included angle between legs.

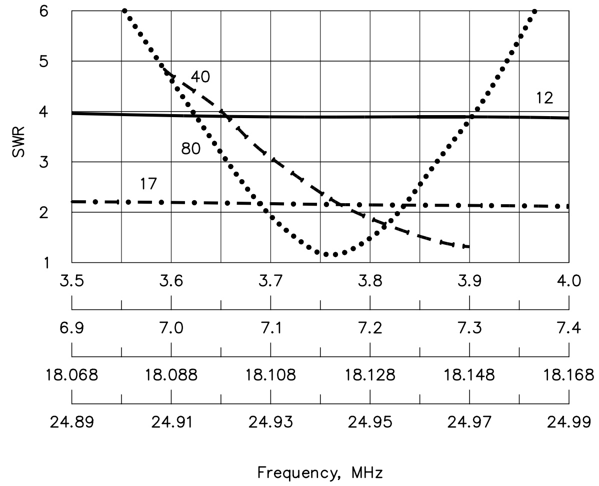

Fig 28—Measured SWR curves for an 80, 40, 17 and 12-meter antenna, installed as an inverted-V with 40-ft apex and 120° included angle between legs.

Fig 27 gives the SWR curves of the 80, 40, 20, 15 and 10-meter antenna. Minimum SWR is nearly 1:1 on 80 meters, 1.5:1 on 40 meters, 1.6:1 on 20 meters, and 1.5:1 on 10 meters. The minimum SWR is slightly below 3:1 on 15 meters. On 15 meters, the stub capacitive reactance combines with the inductive reactance of the outer segment of the antenna to produce a resonant rise that raises the antenna input resistance to about 220 Ω, higher than that of the usual 3/2-wavelength dipole. An antenna tuner may be required on this band to keep a solid-state final output stage happy under these load conditions.

Fig 28 shows the SWR curves of the 80, 40, 17 and 12-meter antenna. Notice the excellent 80-meter performance with a nearly unity minimum SWR in the middle of the band. The performance approaches that of a full-size 80-meter wire dipole. The short stubs and the low-inductance traps shorten the antenna somewhat on 80 meters. Also observe the good 17-meter performance, with the SWR being only a little above 2:1 across the band.

But notice the 12-meter SWR curve of this antenna, which shows 4:1 SWR across the band. The antenna input resistance approaches 300 Ω on this band because the capacitive reactance of the stubs combines with the inductive reactance of the outer antenna segments to give resonant rises in impedance. These are reflected back to the input terminals. These stub-induced resonant impedance rises are similar to those on the other antenna on 15 meters, but are even more pronounced.

Too much concern must not be given to SWR on the feed line. Even if the SWR is as high as 9:1 *no destructively high voltages will exist on the transmission line.* Recall that transmission-line voltages increase as the square root of the SWR in the line. Thus, 1 kW of RF power in 75-Ω line corresponds to 274 V line voltage for a 1:1 SWR. Raising the SWR to 9:1 merely triples the maximum voltage that the line must withstand to 822 V. This voltage is well below the 3700-V rating of RG-11, or the 1700-V rating of RG-59, the two most popular 75-Ω coax lines. Voltage breakdown in the traps is also very unlikely. As will be pointed out later, the operating power levels of these antennas are limited by RF power dissipation in the traps, not trap voltage breakdown or feed-line SWR.

Trap Losses and Power Rating

Table 1 presents the results of trap Q measurements and extrapolation by a two-frequency method to higher frequencies above resonance. W8NX employed an old, but recently calibrated, Boonton Q meter for the measurements. Extrapolation to higher-frequency bands assumes that trap resistance losses rise with skin effect according to the square root of frequency, and that trap dielectric loses rise directly with frequency. Systematic measurement errors are not increased by frequency extrapolation. However, random measurement

Table 1
Trap Q

Frequency (MHz)	*3.8*	*7.15*	*14.18*	*18.1*	*21.3*	*24.9*	*28.6*
High Z out (Ω)	101	124	139	165	73	179	186
Low Z out (Ω)	83	103	125	137	44	149	155

Table 2
Trap Loss Analysis: 80, 40, 20, 15, 10-Meter Antenna

Frequency (MHz)	*3.8*	*7.15*	*14.18*	*21.3*	*28.6*
Radiation Efficiency (%)	96.4	70.8	99.4	99.9	100.0
Trap Losses (dB)	0.16	1.5	0.02	0.01	0.003

Table 3
Trap Loss Analysis: 80, 40, 17, 12-Meter Antenna

Frequency (MHz)	*3.8*	*7.15*	*18.1*	*24.9*
Radiation Efficiency (%)	89.5	90.5	99.3	99.8
Trap Losses (dB)	0.5	0.4	0.03	0.006

errors increase in magnitude with upward frequency extrapolation. Results are believed to be accurate within 4% on 80 and 40 meters, but only within l0 to 15% at 10 meters. Trap Q is shown at both the high- and low-impedance trap terminals. The Q at the low-impedance output terminals is 15 to 20% lower than the Q at the high-impedance output terminals.

W8NX computer-analyzed trap losses for both antennas in free space. Antenna-input resistances at resonance were first calculated, assuming lossless, infinite-Q traps. They were again calculated using the Q values in Table 1. The radiation efficiencies were also converted into equivalent trap losses in decibels. **Table 2** summarizes the trap-loss analysis for the 80, 40, 20, 15 and 10-meter antenna and **Table 3** for the 80, 40,17 and 12-meter antenna.

The loss analysis shows radiation efficiencies of 90% or more for both antennas on all bands except for the 80, 40, 20, 15 and 10-meter antenna when used on 40 meters. Here, the radiation efficiency falls to 70.8%. A 1-kW power level at 90% radiation efficiency corresponds to 50-W dissipation per trap. In W8NX's experience, this is the trap's survival limit for extended key-down operation. SSB power levels of 1 kW PEP would dissipate 25 W or less in each trap. This is well within the dissipation capability of the traps.

When the 80, 40, 20, 15 and 10-meter antenna is operated on 40 meters, the radiation efficiency of 70.8% corresponds to a dissipation of 146 W in each trap when 1 kW is delivered to the antenna. This is sure to burn out the traps—even if sustained for only a short time. Thus, the power should be limited to leas than 300 W when this antenna is operated on 40 meters under prolonged key-down conditions. A 50% CW duty cycle would correspond to a 600-W power limit for normal 40-meter CW operation. Likewise, a 50% duty cycle for 40-meter SSB corresponds to a 600-W PEP power limit for the antenna.

The author knows of no analysis where the burnout wattage rating of traps has been rigorously determined. Operating experience seems to be the best way to determine trap burnout ratings. In his own experience with these antennas, he's had no traps burn out, even though he operated the 80, 40, 20, 15 and 10-meter antenna on the critical 40-meter band using his AL-80A linear amplifier at the 600-W PEP output level. He did not make a continuous, key-down, CW operating test at full power purposely trying to destroy the traps!

Some hams may suggest using a different type of coaxial cable for the traps. The dc resistance of 40.7 Ω per 1000 feet of RG-59 coax seems rather high. However, W8NX has found no coax other than RG-59 that has the necessary inductance-to-capacitance ratio to create the trap characteristic reactance required for the 80, 40, 20, 15 and 10-meter antenna. Conventional traps with wide-spaced, open-air inductors and appropriate fixed-value capacitors could be substituted for the coax traps, but the convenience, weatherproof configuration and ease of fabrication of coaxial-cable traps is hard to beat.

Multiband Vertical Antennas

There are two basic types of vertical antennas; either type can be used in multiband configurations. The first is the ground-mounted vertical and the second, the ground plane. These antennas are described in detail in Chapter 6, Low-Frequency Antennas.

The efficiency of any ground-mounted vertical depends a great deal on near-field earth losses. As pointed out in Chapter 3, The Effects of Ground, these near-field losses can be reduced or eliminated with an adequate radial system. Considerable experimentation has been conducted on this subject by Jerry Sevick, W2FMI, and several important results were obtained. It was determined that a radial system consisting of 40 to 50 radials, 0.2 λ long, would reduce the earth losses to about 2 Ω when a λ/4 radiator was being used. These radials should be on the earth's surface, or if buried, placed not more than an inch or so below ground. Otherwise, the RF current would have to travel through the lossy earth before reaching the radials. In a multiband vertical system, the radials should be 0.2 λ long for the lowest band, that is, 55 feet long for 3.5-MHz operation. Any wire size may be used for the radials. The radials should fan out in a circle, radiating from the base of the antenna. A metal plate, such as a piece of sheet copper, can be used at the center connection.

The other common type of vertical is the ground-plane antenna. Normally, this antenna is mounted above ground with the radials fanning out from the base of the antenna. The vertical portion of the antenna is usually an electrical λ/4, as is each of the radials. In this type of antenna, the system of radials acts somewhat like an RF choke, to prevent RF currents from flowing in the supporting structure, so the number of radials is not as important a factor as it is with a ground- mounted vertical system. From a practical standpoint, the customary number of radials is four or five. In a multiband configuration, λ/4 radials are required for each band of operation with the ground-plane antenna.

This is not so with the ground-mounted vertical antenna, where the ground plane is relied upon to provide an image of the radiating section. Note that even quarter-wave-long radials are greatly detuned by their proximity to ground—radial resonance is not necessary or even possible. In the ground-mounted case, so long as the ground-screen radials are approximately 0.2 λ long at the lowest frequency, the length will be more than adequate for the higher frequency bands.

Short Vertical Antennas

A short vertical antenna can be operated on several bands by loading it at the base, the general arrangement being similar to Figs 1 and 2. That is, for multiband work the vertical can be handled by the same methods that are used for random-length wires.

A vertical antenna should not be longer than about ¾ λ at the highest frequency to be used, however, if low-angle radiation is wanted. If the antenna is to be used on 28 MHz and lower frequencies, therefore, it should not be more than approximately 25 feet high, and the shortest possible ground lead should be used.

Another method of feeding is shown in **Fig 29**. L1 is a loading coil, tapped to resonate the antenna on the desired band. A second tap permits using the coil as a transformer for matching a coax line to the transmitter. C1 is not strictly necessary, but may be helpful on the lower frequencies, 3.5 and 7 MHz, if the antenna is quite short. In that case C1 makes it possible to tune the system to resonance with a coil of reasonable dimensions at L1. C1 may also be useful on other bands as well, if the system cannot be matched to the feed line with a coil alone.

The coil and capacitor should preferably be installed at the base of the antenna, but if this cannot be done a wire can be run from the antenna base to the nearest convenient location for mounting L1 and C1. The extra wire will of course be a part of the antenna, and since it may have to run through unfavorable surroundings it is best to avoid using it if at all possible.

This system is best adjusted with the help of an SWR indicator. Connect the coax line across a few turns of L1 and take trial positions of the shorting tap until the SWR reaches its lowest value. Then vary the line tap similarly; this should bring the SWR down to a low value. Small adjustments of both taps then should reduce the SWR to close to 1:1. If not, try adding C1 and go through the same procedure, varying C1 each time a tap position is changed.

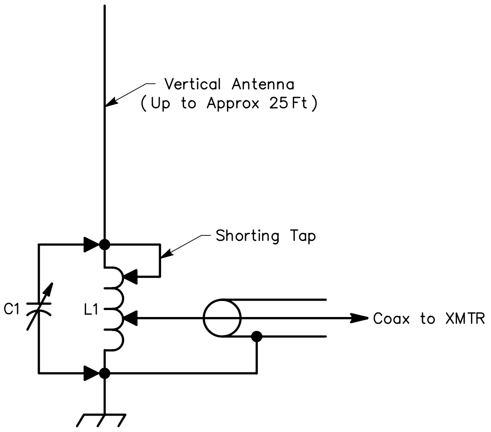

Fig 29—Multiband vertical antenna system using base loading for resonating on 3.5 to 28 MHz. L1 should be wound with bare wire so it can be tapped at every turn, using #12 wire. A convenient size is 2½ inches diameter, 6 turns per inch (such as B&W 3029). Number of turns required depends on antenna and ground lead length, more turns being required as the antenna and ground lead are made shorter. For a 25-foot antenna and a ground lead of the order of 5 feet, L1 should have about 30 turns. The use of C1 is explained in the text. The smallest capacitance that will permit matching the coax cable should be used; a maximum capacitance of 100 to 150 pF will be sufficient in any case.

Trap Verticals

The trap principle described in Fig 17 for center-fed dipoles also can be used for vertical antennas. There are two principal differences. Only one half of the dipole is used, the ground connection taking the place of the missing half, and the feed-point impedance is one half the feed-point impedance of a dipole. Thus it is in the vicinity of 30 Ω (plus the ground-connection resistance), so 52-Ω cable should be used since it is the commonly available type that comes closest to matching.

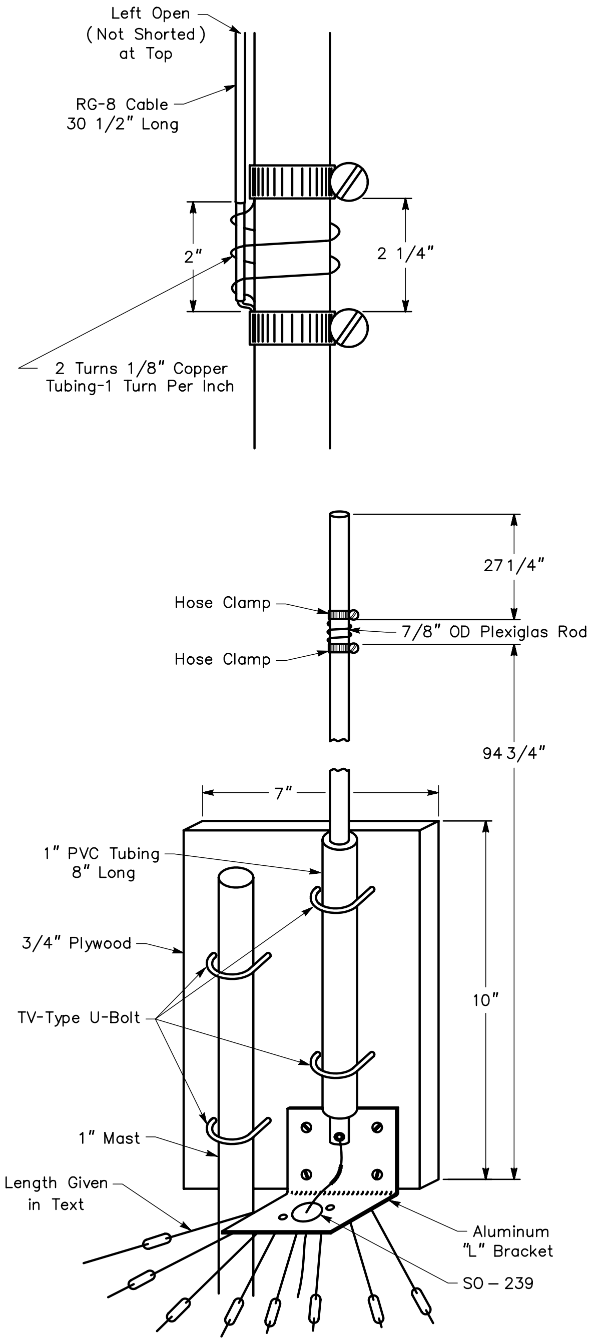
Left Open
(Not Shorted)
at Top
RG-8 Cable
30 1/2" Long
2"
2 1/4"
2 Turns 1/8" Copper
Tubing-1 Turn Per Inch
27 1/4"
Hose Clamp
7/8" OD Plexiglas Rod
Hose Clamp
94 3/4"
7"
1" PVC Tubing
8" Long
3/4" Plywood
TV-Type U-Bolt
10"
1" Mast
Length Given
in Text
Aluminum
"L" Bracket
SO – 239

Fig 30—Constructional details of the 21- and 28-MHz dual-band antenna system.

A TRAP VERTICAL FOR 21 AND 28 MHZ

Simple antennas covering the upper HF bands can be quite compact and inexpensive. The two-band vertical ground plane described here is highly effective for long-distance communication when installed in the clear.

Figs 30, **31** and **32** show the important assembly details. The vertical section of the antenna is mounted on a ¾-inch thick piece of plywood board that measures 7 × 10 inches. Several coats of exterior varnish or similar material will help protect the wood from inclement weather. Both the mast and the radiator are mounted on the piece of wood by means of TV U-bolt hardware. The vertical is electrically isolated from the wood with a piece of 1-inch diameter PVC tubing. A piece approximately 8 inches long is required, and it is of the schedule-80 variety. To prepare the tubing, you must slit it along the entire length on one side. A hacksaw will work quite well. The PVC fits rather snugly on the aluminum tubing and

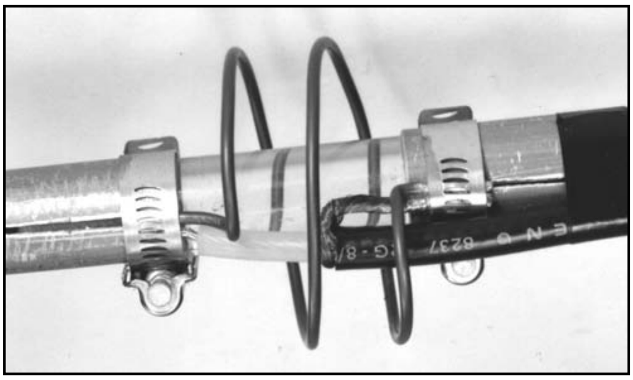

Fig 31—A close-up view of a trap. The coil is 3 inches in diameter. The leads from the coaxial-cable capacitor should be soldered directly to the pigtails of the coil. These connections should be coated with varnish after they have been secured under the hose clamps.

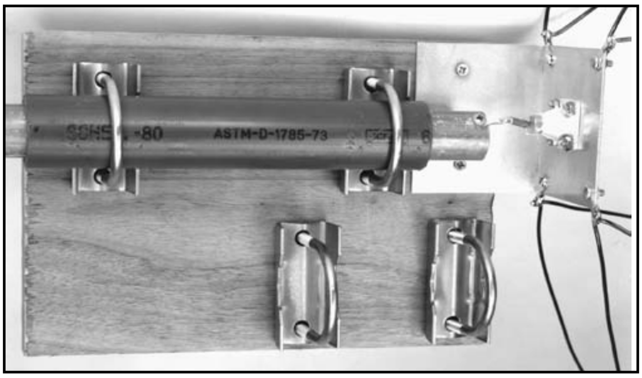

Fig 32—The base assembly of the 21- and 28-MHz vertical. The SO-239 coaxial connector and hood can be seen in the center of the aluminum L bracket. The U bolts are TV-type antenna hardware. The plywood should be coated with varnish or similar material.

will have to be "persuaded" with the aid of a hammer. Mount the mast directly on the wood with no insulation.

Use an SO-239 coaxial connector and four solder lugs on an L-shaped bracket made from a piece of aluminum sheet. Solder a short length of test probe wire, or inner conductor of RG-58 cable, to the inner terminal of the connector. A UG-106 connector hood is then slid over the wire and onto the coaxial connector. Then bolt the hood and connector to the aluminum bracket. Two wood screws are used to secure the aluminum bracket to the plywood, as shown in the drawing and photograph. Solder the free end of the wire coming from the connector to a lug mounted on the bottom of the vertical radiator. Fill any space between the wire and where it passes through the hood with GE silicone sealant or similar material to keep moisture out. The eight radials, four for each band, are soldered to the four lugs on the aluminum bracket. Separate the two sections of the vertical member with a piece of clear acrylic rod. Approximately 8 inches of ⅞-inch OD material is required. You must slit the aluminum tubing lengthwise for several inches so the acrylic rod may be inserted. The two pieces of aluminum tubing are separated by 2¼ inches.

The trap capacitor is made from RG-8 coaxial cable and is 30.5 inches long. RG-8 cable has 29.5 pF of capacitance per foot and RG-58 has 28.5 pF per foot. RG-8 cable is recommended over RG-58 because of its higher breakdown-voltage capability. The braid should be pulled back 2 inches on one end of the cable, and the center conductor soldered to one end of the coil. Solder the braid to the other end of the coil. Compression type hose clamps are placed over the capacitor/coil leads and put in position at the edges of the aluminum tubing. When tightened securely, the clamps serve a two-fold purpose—they keep the trap in contact with the vertical members and prevent the aluminum tubing from slipping off the acrylic rod. The coaxial-cable capacitor runs upward along the top section of the antenna. This is the side of the antenna to which the braid of the capacitor is connected. Place a cork or plastic cap in the very top of the antenna to keep moisture out.

Installation and Operation

The antenna may be mounted in position using a TV-type tripod, chimney, wall or vent mount. Alternatively, a telescoping mast or ordinary steel TV mast may be used, in which case the radials may be used as guys for the structure. The 28-MHz radials are 8 feet 5 inches long, and the 21-MHz radials are 11 feet 7 inches.

Any length of 50-Ω cable may be used to feed the antenna. The SWR at resonance should be on the order of 1.2:1 to 1.5:1 on both bands. The reason the SWR is not 1:1 is that the feed-point resistance is something other than 50 Ω—closer to 35 or 40 Ω.

The Open-Sleeve Antenna

Although only recently adapted for the HF and VHF amateur bands, the open-sleeve antenna has been around since 1946. The antenna was invented by Dr J. T. Bolljahn, of Stanford Research Institute. This section on sleeve antennas was written by Roger A. Cox, WB0DGF.

The basic form of the open-sleeve monopole is shown in **Fig 33**. The open-sleeve monopole consists of a base-fed central monopole with two parallel closely spaced parasitics, one on each side of the central element, and grounded at each base. The lengths of the parasitics are roughly one half that of the central monopole.

Impedance

The operation of the open sleeve can be divided into two modes, an antenna-mode and a transmission-line mode. This is shown in **Fig 34**.

The antenna-mode impedance, Z_A, is determined by the length and diameter of the central monopole. For sleeve lengths less than that of the monopole, this impedance is essentially independent of the sleeve dimensions.

The transmission-line mode impedance, Z_T, is determined by the characteristic impedance, end impedance, and length of the 3- wire transmission line formed by the central monopole and the two sleeve elements. The characteristic impedance, Z_c, can be determined by the element diameters and spacing if all element diameters are equal, and is found from

$$Z_c = 207 \log 1.59 (D/d) \qquad \text{(Eq 2)}$$

where

D = spacing between the center of each sleeve element and the center of the driven element

d = diameter of each element

This is shown graphically in **Fig 35**. However, since the end impedance is usually unknown, there is little need to know the characteristic impedance. The transmission-line mode impedance, Z_T, is usually determined by an educated guess and experimentation.

As an example, let us consider the case where the central monopole is λ/4 at 14 MHz. It would have an antenna mode impedance, Z_A, of approximately 52 Ω, depending upon the ground conductivity and number of radials. If two sleeve elements were added on either side of the central monopole, with each approximately half the height of the monopole and at a distance equal to their height, there would be very little

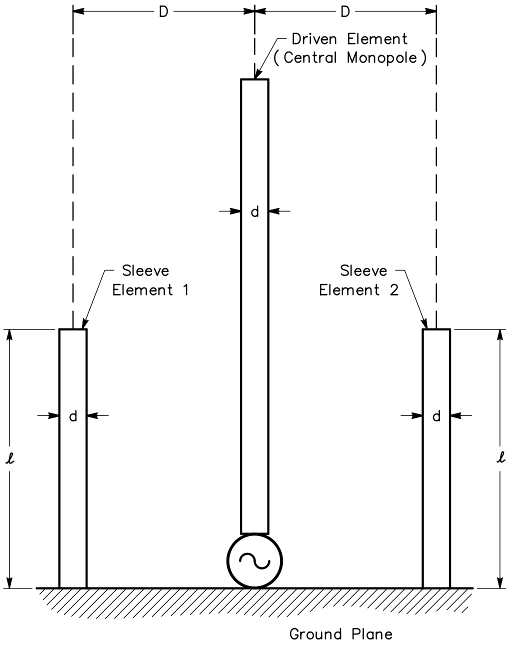

Fig 33—Diagram of an open-sleeve monopole.

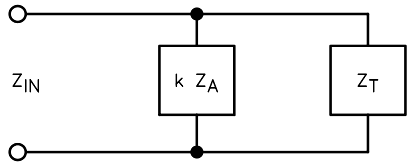

Fig 34—Equivalent circuit of an open-sleeve antenna.

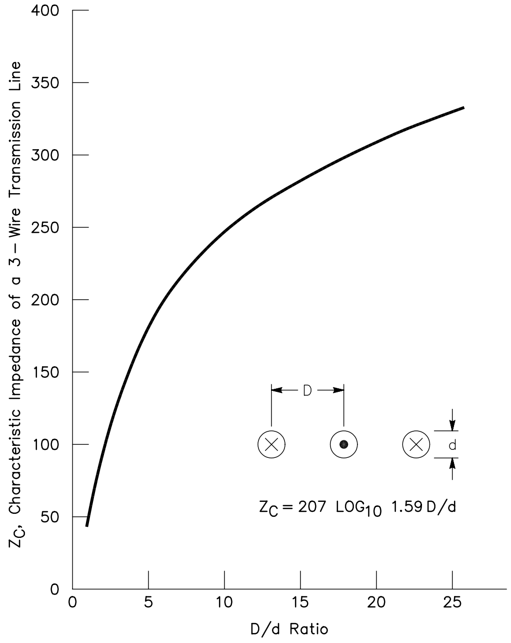

Fig 35—Characteristic impedance of transmission-line mode in an open-sleeve antenna.

effect on the antenna mode impedance, Z_A, at 14 MHz.

Also, Z_T at 14 MHz would be the end impedance transformed through a $\lambda/8$ section of a very high characteristic impedance transmission line. Therefore, Z_T would be on the order of 500-2000 Ω resistive plus a large capacitive reactance component. This high impedance in parallel with 52 Ω would still give a resultant impedance close to 52 Ω.

At a frequency of 28 MHz, however, Z_A is that of an end-fed half-wave antenna, and is on the order of 1000-5000 Ω resistive. Also, Z_T at 28 MHz would be on the order of 1000 to 5000 Ω resistive, since it is the end impedance of the sleeve elements transformed through a quarter-wave section of a very high characteristic impedance 3-wire transmission line. Therefore, the parallel combination of Z_A and Z_T would still be on the order of 500 to 2500 Ω resistive.

If the sleeve elements were brought closer to the central monopole such that the ratio of the spacing to element diameter was less than 10:1, then the characteristic impedance of the 3-wire transmission line would drop to less than 250 Ω. At 28 MHz, Z_A remains essentially unchanged, while Z_T begins to edge closer to 52 Ω as the spacing is reduced. At some particular spacing the characteristic impedance, as determined by the D/d ratio, is just right to transform the end impedance to exactly 52 Ω at some frequency. Also, as the spacing is decreased, the frequency where the impedance is purely resistive gradually increases.

The actual impedance plots of a 14/28-MHz open-sleeve monopole appear in **Figs 36** and **37**. The length of the central monopole is 195.5 inches, and of the sleeve elements 89.5 inches. The element diameters range from 1.25 inches at the bases to 0.875 inch at each tip. The measured impedance of the 14-MHz monopole alone, curve A of Fig 36, is quite high. This is probably because of a very poor ground plane under the antenna. The addition of the sleeve elements raises this impedance slightly, curves B, C and D.

As curves A and B in Fig 37 show, an 8-inch sleeve spacing gives a resonance near 27.8 MHz at 70 Ω, while a 6-inch spacing gives a resonance near 28.5 MHz at 42 Ω. Closer spacings give lower impedances and higher resonances. The optimum spacing for this particular antenna would be somewhere between 6 and 8 inches. Once the spacing is found, the lengths of the sleeve elements can be tweaked slightly for a choice of resonant frequency.

In other frequency combinations such as 10/21, 10/24, 14/21 and 14/24-MHz, spacings in the 6 to 10-inch range work very well with element diameters in the 0.5 to 1.25-inch range.

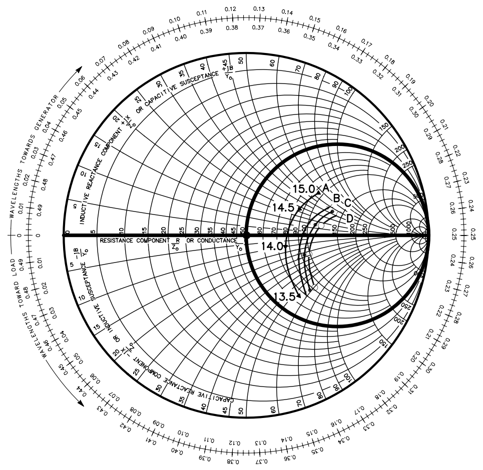

Fig 36—Impedance of an open-sleeve monopole for the frequency range 13.5-15 MHz. Curve A is for a 14 MHz monopole alone. For curves B, C and D, the respective spacings from the central monopole to the sleeve elements are 8, 6 and 4 inches. See text for other dimensions.

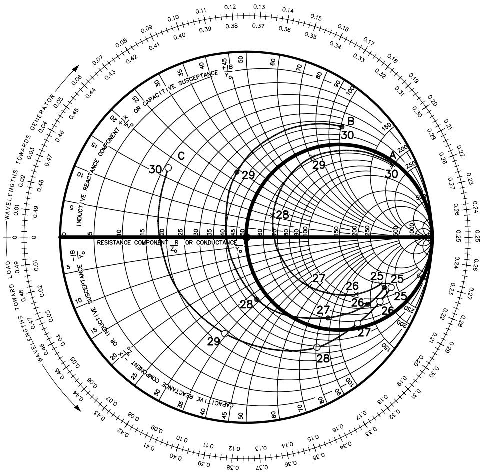

Fig 37—Impedance of the open-sleeve monopole for the range 25-30 MHz. For curves A, B and C the spacings from the central monopole to the sleeve elements are 8, 6 and 4 inches, respectively.

Bandwidth

The open-sleeve antenna, when used as a multiband antenna, does not exhibit broad SWR bandwidths unless, of course, the two bands are very close together. For example, **Fig 38** shows the return loss and SWR of a single 10-MHz vertical antenna. Its 2:1 SWR bandwidth is 1.5 MHz, from 9.8 to 11.3 MHz. Return loss and SWR are related as given by the following equation.

$$\text{SWR} = \frac{1+k}{1-k} \tag{Eq 3}$$

where

$$k = 10^{\frac{R_L}{20}}$$

RL = return loss, dB

When sleeve elements are added for a resonance near 22 MHz, the 2:1 SWR bandwidth at 10 MHz is still nearly 1.5 MHz, as shown in **Fig 39**. The total amount of spectrum under 2:1 SWR increases, of course, because of the additional band, but the individual bandwidths of each resonance are virtually unaffected.

The open-sleeve antenna, however, can be used as a broadband structure, if the resonances are close enough to overlap. With the proper choices of resonant frequencies, sleeve and driven element diameters and sleeve spacing, the SWR "hump" between resonances can be reduced to a value less than 3:1. This is shown in **Fig 40**.

Current Distribution

According to H. B. Barkley (see Bibliography at the end of this chapter), the total current flowing into the base of the open-sleeve antenna may be broken down into two components, that contributed by the antenna mode, I_A, and that contributed by the transmission-line mode, I_T. Assuming that the sleeves are approximately half the height of the central monopole, the impedance of the antenna mode, Z_A, is very low at the resonant frequency of the central monopole, and

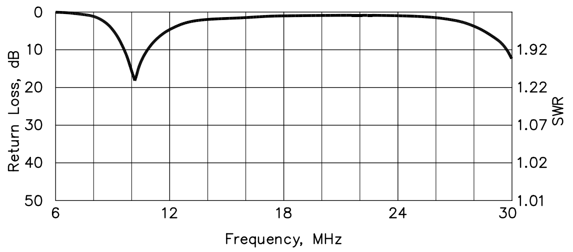

Fig 38—Return loss and SWR of a 10 MHz vertical antenna. A return loss of 0 dB represents an SWR of infinity. The text contains an equation for converting return loss to an SWR value.

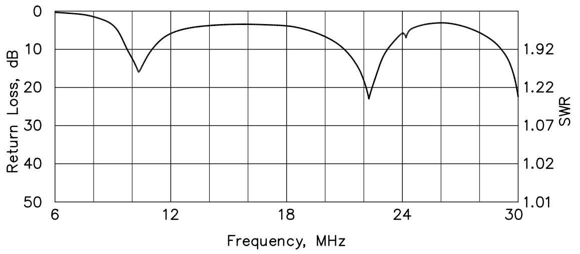

Fig 39—Return loss and SWR of a 10/22 MHz open-sleeve vertical antenna.

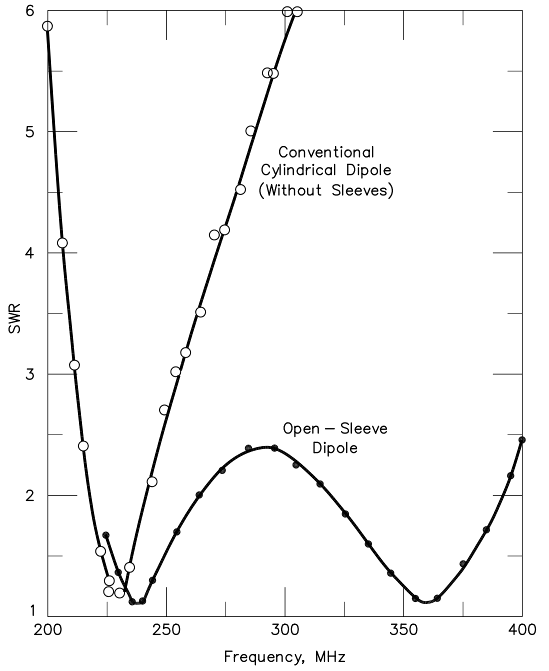

Fig 40—SWR response of an open-sleeve dipole and a conventional dipole.

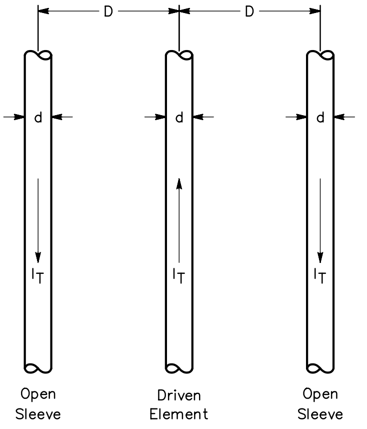

Fig 41—Current distribution in the transmission-line mode. The amplitude of the current induced in each sleeve element equals that of the current in the central element but the phases are opposite, as shown.

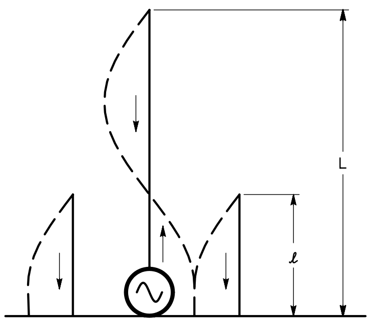

Fig 42—Total current distribution with λ= L/2.

the impedance of the transmission-line mode, Z_T, is very high. This allows almost all of the current to flow in the antenna mode, and I_A is very much greater than I_T. Therefore, the current on the central λ/4 monopole assumes the standard sinusoidal variation, and the radiation and gain characteristics are much like those of a normal λ/4 vertical antenna.

However, at the resonant frequency of the sleeves, the impedance of the central monopole is that of an end fed half-wave monopole and is very high. Therefore I_A is small. If proper element diameters and spacings have been used to match the transmission line mode impedance, Z_T, to 52 Ω, then I_T, the transmission line mode current, is high compared to I_A.

This means that very little current flows in the central monopole above the tops of the sleeve elements, and the radiation is mostly from the transmission-line mode current, I_T, in all three elements below the tops of the sleeve elements. The resulting current distribution is shown in **Figs 41** and **42** for this case.

Radiation Pattern and Gain

The current distribution of the open-sleeve antenna where all three elements are nearly equal in length is nearly that of a single monopole antenna. If, at a particular frequency, the elements are approximately λ/4 long, the current distribution is sinusoidal.

If, for this and other length ratios, the chosen diameters and spacings are such that the two sleeve elements approach an interelement spacing of λ/8, the azimuthal pattern will show directivity typical of two in-phase vertical radiators, approximately λ/8 apart. If a bi-directional pattern is needed, then this is one way to achieve it.

Spacings closer than this will produce nearly circular azimuthal radiation patterns. Practical designs in the 10 to 30 MHz range using 0.5 to 1.5-inch diameter elements will produce azimuthal patterns that vary less than ±1 dB.

If the ratio of the length of the central monopole to the length of the sleeves approaches 2:1, then the elevation pattern of the open-sleeve vertical antenna at the resonant frequency of the sleeves becomes slightly compressed. This is because of the in-phase contribution of radiation from the λ/2 central monopole.

As shown in **Fig 43**, the 10/21-Mhz open-sleeve vertical antenna produces a lower angle of radiation at 21.2 MHz with a corresponding increase in gain of 0.66 dB over that of the 10-MHz vertical alone. At length ratios approaching 3:1, the

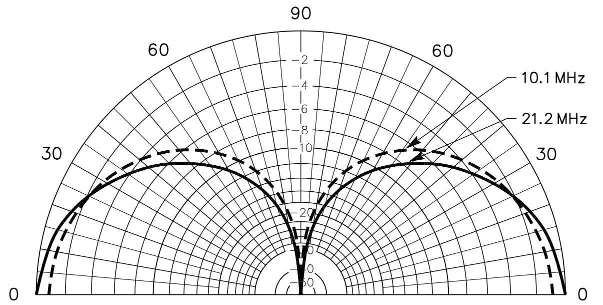

Fig 43—Vertical-plane radiation patterns of a 10/21-MHz open-sleeve vertical antenna on a perfect ground plane. At 10.1 MHz the maximum gain is 5.09 dBi, and 5.75 dBi at 21.2 MHz.

antenna mode and transmission-line mode impedance become nearly equal again, and the central monopole again carries a significant portion of the antenna current. The radiation from the top λ/2 combines constructively with the radiation from the λ/4 sleeve elements to produce gains of up to 3 dB more than just a quarter-wave vertical element alone.

Length ratios in excess of 3.2:1 produce higher level sidelobes and less gain on the horizon, except for narrow spots near the even ratios of 4:1, 6:1, 8:1, etc. These are where the central monopole is an even multiple of a half-wave, and the antenna-mode impedance is too high to allow much antenna-mode current.

Up to this point, it has been assumed that only λ/4 resonance could be used on the sleeve elements. The third, fifth, and seventh-order resonances of the sleeve elements and the central monopole element can be used, but their radiation patterns normally consist of high-elevation lobes, and the gain on the horizon is less than that of a λ/4 vertical.

Practical Construction and Evaluation

The open-sleeve antenna lends itself very easily to home construction. For the open-sleeve vertical antenna, only a feed-point insulator and a good supply of aluminum tubing are needed. No special traps or matching networks are required. The open-sleeve vertical can produce up to 3 dB more gain than a conventional λ/4 vertical. Further, there is no reduction in bandwidth, because there are no loading coils.

The open-sleeve design can also be adapted to horizontal dipole and beam antennas for HF, VHF and UHF. A good example of this is Telex/Hy-Gain's Explorer 14 triband beam which utilizes an open sleeve for the 10/15-meter driven element. The open-sleeve antenna is also very easy to model in computer programs such as *NEC* and *MININEC*, because of the open tubular construction and lack of traps or other intricate structures.

In conclusion, the open-sleeve antenna is an antenna experimenters delight. It is not difficult to match or construct, and it makes an ideal broadband or multiband antenna.

The Coupled-Resonator Dipole

A variation of the open-sleeve system above is the coupled-resonator system described by Gary Breed, K9AY, in an article in *The ARRL Antenna Compendium, Vol 5*, entitled “The Coupled-Resonator Principle: A Flexible Method for Multiband Antennas.” The following is condensed from that article.

In 1995, *QST* published two antenna designs that use an interesting technique to get multiband coverage in one antenna. Rudy Severns, N6LF, described a wideband 80 and 75-meter dipole using this technique, and Robert Wilson, AL7KK, showed us how to make a three-band vertical. Both of these antennas achieve multi-frequency operation by placing resonant conductors very close to a driven dipole or vertical—with no physical connection.

The Coupled-Resonator Principle

As we all know, nearby conductors can interact with an antenna. Our dipoles, verticals and beams can be affected by nearby power lines, rain gutters, guy wires and other metallic materials. The antennas designed by Severns and Wilson use this interaction intentionally, to combine the resonances of several conductors at a single feed point. While other names have been used, I call the behavior that makes these antennas work the *coupled-resonator* (C-R) principle.

Take a look at **Fig 44**, which illustrates the general idea. Each figure shows the SWR at the feed point of a dipole, over a range of frequencies. When this dipole is all alone, it will have a very low SWR at its half-wave resonant frequency (Fig 44A). Next, if we take another wire or tubing conductor and start bringing it close to the dipole, we will see a “bump” in the dipole’s SWR at the resonant frequency of this new wire. See Fig 44B. We are beginning to the see the effects of interaction between the two conductors. As we bring this new conductor closer, we reach a point where the SWR “bump” has grown to a very deep dip—a low SWR. We now have a good match at both the original dipole’s resonant frequency and the frequency of the new conductor, as illustrated in Fig 44C.

We can repeat this process for several more conductors at other frequencies to get a dipole with three, four, five, six, or more resonant frequencies. The principle also applies to verticals, so any reference to a dipole can be considered to be valid for a vertical, as well.

We can write a definition of the C-R principle this

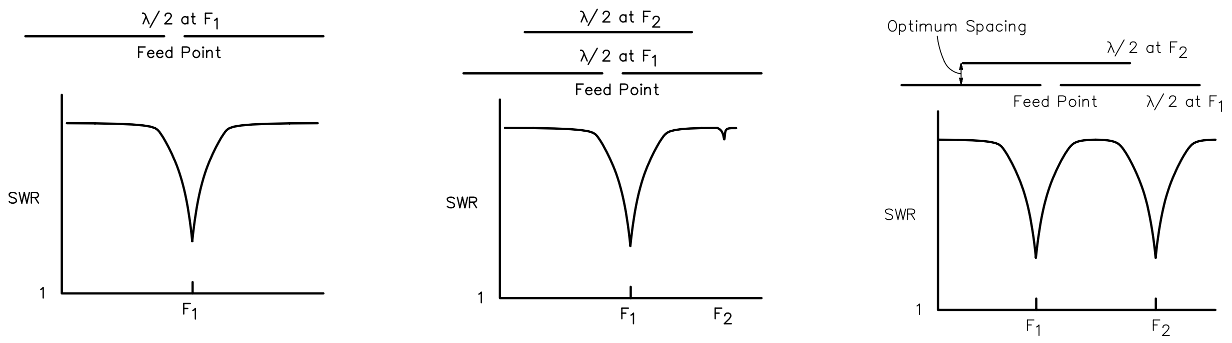

Fig 44—At A, the SWR of a dipole over a wide frequency range. At B, a nearby conductor is just close enough to interact with the dipole. At C, when the second conductor is at the optimum spacing, the combination is matched at both frequencies.

way: *Given a dipole (or vertical) at one frequency and an additional conductor resonant at another frequency, there is an optimum distance between them that results in the resonance of the additional conductor being imposed upon the original dipole, resulting in a low SWR at both resonant frequencies.*

Some History

In the late 1940s, the *coaxial sleeve* antenna was developed (**Fig 45A**), covering two frequencies by surrounding a dipole or monopole with a cylindrical tube resonant at the higher of the desired frequencies. In the 1950s, Gonset briefly marketed a two-band antenna based on this design. Other experimenters soon determined that two conductors at the second frequency, placed on either side of the main dipole or monopole, would make a skeleton representation of a cylinder (Fig 45B). This is called the *open-sleeve* antenna. The Hy-Gain Explorer tribander uses this method in its driven element to obtain resonance in the 10-meter band. Later on, a few antenna developers finally figured out that these extra conductors did not need to be added in pairs, and that a single conductor at each frequency could add the extra resonances (Fig 45C). This is the method used by Force 12 in some of their multiband antennas.

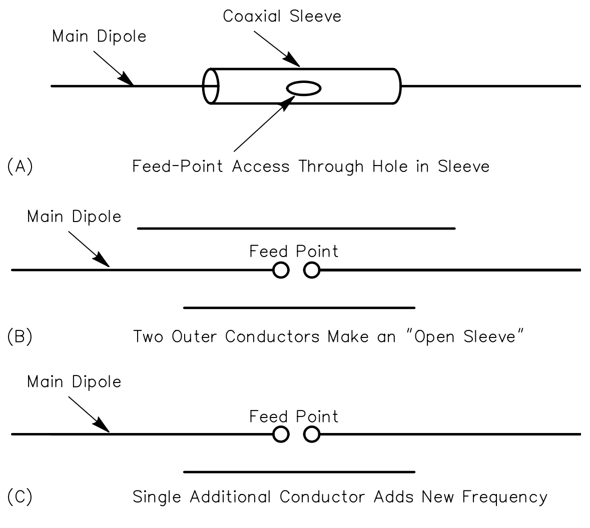
Coaxial Sleeve
Main Dipole
(A)
Feed-Point Access Through Hole in Sleeve
Main Dipole
Feed Point
(B)
Two Outer Conductors Make an "Open Sleeve"
Main Dipole
Feed Point
(C)
Single Additional Conductor Adds New Frequency

Fig 45—Evolution of coupled-resonator antennas: At A, the *coaxial-sleeve* dipole; at B, the *open-sleeve* dipole; and at C, a *coupled-resonator* dipole, the most universal configuration.

This is a perfect example of how science works. A specific idea is discovered, with later developments leading to an underlying general principle. The original coaxial-sleeve configuration is the most specific, being limited to two frequencies and requiring a particular construction method. The open-sleeve antenna is an intermediate step, showing that the sleeve idea is not limited to one configuration.

Finally, we have the coupled-resonator concept, which is the general principle, applicable in many different antenna configurations, for many different frequency combinations. Severns' antenna uses it with a folded dipole, and Wilson uses it with a main vertical that is off-center fed. The author K9AY used it with conventional dipoles and quarter-wave verticals. Other designers have used the principle more subtly, like putting the first director in a Yagi very close to the driven element, broadening the SWR bandwidth the same way Severns' design does with a dipole.

In the past, most antennas built with this single-conductor technique have also been called *open-sleeve* (or *multiple-open-sleeve*) antennas, a term taken from the history of their development. However, the term *sleeve* implies that one conductor must surround another. This is not really a physical or electrical description of the antenna's operation, therefore, K9AY suggests using the term *coupled-resonator,* which is the most accurate description of the general principle.

A Little Math

The interaction that makes the C-R principle work is not random. It behaves in a predictable, regular manner. K9AY derived an equation that shows the relationship between the driven element and the additional resonators for ordinary dipoles and verticals:

$$\frac{\log_{10} d}{\log_{10}(D/4)} = 0.54$$ (Eq 4)

where

d = distance between conductors, measured in wavelengths at the frequency of the chosen additional resonator

D = the diameter of the conductors, also in wavelengths at the frequency of the additional resonator.

Eq 4 assumes they are both the same diameter and that the feed point impedance at both frequencies is the same as a dipole in free space (72 Ω) or a quarter-wave monopole over perfect ground (36 Ω).

The equation only describes the impedance due to the additional resonator. The main dipole element is always part of the antenna, and it may have a fairly low impedance at the additional frequency. This is the case when the frequencies are close together, or when the main element is operating at its third harmonic. At these frequencies, the spacing distance must be adjusted so that the parallel combination of dipole and resonator results in the desired feed-point impedance.

K9AY worked out two correction factors, one to cover a range of impedances and another for frequencies close together. These can be included in the basic equation, which is rearranged below to solve for the distance between the conductors:

$$d = 10^{0.54 \log_{10}(D/4)} \times \frac{Z_0 + 35.5}{109} \times [1 + e^{-[((F_2/F_1) - 1.1) \times 11.3) + 0.1)]}]$$ (Eq 5)

where

d and D are the same as above.

Z_0 = the desired feed point impedance at the frequency of the additional resonator (between 20 and 120 Ω). For a vertical, multiply the desired impedance by two to get Z_0. If you want a 50-Ω feed, use 100 Ω for Z_0.

F_1 = the resonant frequency of the main dipole or vertical.

F_2 = the resonant frequency of the additional conductor. The ratio F_2/F_1 is more than 1.1.

e = 2.7183, the base of natural logarithms.

Eq 5 does not directly allow for conductors of unequal diameters, but it can be used as a starting point if you use the diameter of the driven dipole or vertical element for D in the equation.

Characteristics of C-R Antennas

Here's the important stuff—what's different about C-R antennas, what are they good for and what are their drawbacks? The key points are:

- Multiband operation without traps, stubs or tuners

K9AY's Eq 5 above does indeed yield a good "first-cut" value for the spacing between coupled-resonator elements. **Fig A** shows the spacing, in inches, plotted against the ratio of frequencies, for two coupled resonator elements with different diameters, again expressed in inches. This is for an upper frequency of 28.4 MHz. Beyond a frequency ratio of about 1.5:1 (28.4:18.1 MHz), the spacing flattens out to a fixed distance between elements for each element diameter. For example, if $^1/_2$-inch elements are used at 28.4 and 18.1 MHz, the spacing between the elements is about 3.75 inches.

EZNEC verifies Eq 5's computations. Note that a large number of segments are necessary for each element when they are closely spaced from each other, and the segments on the elements must be closely aligned with each other. Be sure to run the Average Gain test, as well as Segmentation tests. The modeler should also be aware that if mutually coupled resonators are placed along a horizontal boom (as they would be on multiband Yagis using coupled resonators), the higher-frequency elements will act like retrograde directors, producing some gain (or lack of gain, depending on the azimuth being investigated).

For example, in the *EZNEC* file **K9AY C-R 28-21-14 MHz 1 In.EZ**, using 1-inch diameter elements spaced 6 inches apart, if the 28-MHz element is placed 6 inches behind the 14-MHz driven element (with the 21-MHz element placed 6 inches ahead), on 28 MHz the system will have a F/B of 2.6 dB, favoring the rearward direction. On 21 MHz, the system will exhibit a F/B of 1.6 dB, favoring the forward direction. Of course, there are systems where gain and F/B due to the C-R configuration may be put to good use, such as the multiband Yagis mentioned above. However, if the elements are spaced above/below the 14-MHz driven element there is no distortion of the dipole patterns.

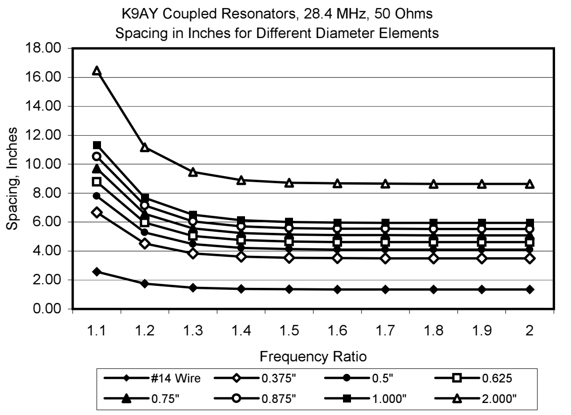

Fig A—Graph of the spacing versus frequency ratio for two Coupled-Resonator elements at 28.4 MHz, for 50-Ω feed-point impedance.

- Flexible impedance matching at each frequency
- Independent fine-tuning at each frequency (little interaction)
- Easily modeled using *MININEC* or *NEC*-based programs
- Pruning process same as a simple dipole
- Can accommodate many frequencies (seven or more)
- Virtually lossless coupling (high efficiency)
- Requires a separate wire or tubing conductor at each frequency
- Mechanical assembly requires a number of insulated supports
- Narrower bandwidth than equivalent dipole
- Capacitance requires slight lengthening of conductors

To begin with, the most obvious characteristic is that this principle can be used to add multiple resonant frequencies to an ordinary dipole or vertical, using additional conductors that are not physically connected. This gives us three variable factors: (1) the diameter of the conductor, (2) its length, and (3) its position relative to the main element.

Having the freedom to control these factors gives us the advantage of *flexibility*; we have a wide range of control over the impedance at each added frequency. Another advantage is that the behavior at each frequency is quite *independent*, once the basic design is in place. In other words, making fine-tuning adjustments at one frequency doesn't change the resonance or impedance at the other frequencies. A final advantage is *efficiency.* With conductors close together, and with a resonant target conductor, coupling is very efficient. Traps, stubs, and compensating networks found on other multiband antennas all introduce lossy reactive components.

There are two main disadvantages of C-R antennas. The first is the relative *complexity* of construction. Several conductors are needed, installed with some type of insulating spacers. Other multiband antennas have their complexities as well (such as traps that need to be mounted and tuned), but C-R antennas will usually be bulkier. The larger size generally means greater windload, which is a disadvantage to some hams.

The other significant disadvantage is *narrower bandwidth,* particularly at the highest of the operating frequencies. We can partially overcome this problem with large conductors that are naturally broad in bandwidth, and in some cases we might even use an extra conductor to put two resonances in one band. It is interesting to note that the pattern is opposite that of trapped antennas. The C-R antenna gets narrower at the highest frequencies of operation, while trap antennas generally have narrowest bandwidth at their lowest frequencies.

There are two special situations that should be noted. First, when the antenna has a resonance near the frequency where the driven dipole is 3/2 λ long (3/4 λ for a vertical), the dipole has a fairly low impedance. The spacing of the C-R element needs to be increased to raise its impedance so that the parallel combination of the main element and C-R element equals the desired impedance (usually 50 Ω). There is also significant antenna current in the part of the main dipole extending beyond the C-R section, contributing to the total radiation pattern. As a result, this particular arrangement radiates as three λ/2 sections in phase, and has about 3 dB gain and a narrower directional pattern compared to a dipole (**Fig 46**). This might be an advantage for antennas covering bands with a frequency ratio of about three, such as 3.5 and 10.1 MHz, 7 and 21 MHz, or 144 and 430 MHz.

The other special situation is when we want to add a new frequency very close to the resonant frequency of the main dipole. An antenna for 80 and 75 meters would be an example of this. Again, the driven dipole has a fairly low impedance at the new frequency. Add the fact that coupling is very strong between these similar conductors and we find that a wide spacing is required to make the antenna work. A dipole resonant at 3.5 MHz and another wire resonant at 3.8 MHz will need to be 3 or 4 feet apart, while a 3.5 MHz and 7 MHz combination might only need to be spaced 4 or 5 inches.

Another useful characteristic of C-R antennas is that they are easily and accurately modeled by computer programs based on either *MININEC* or *NEC*, as long as you stay within each program's limitations. For example, Severns points out that *MININEC* does not handle folded dipoles very well, and *NEC* modeling is required. With ease of computer modeling, a precise answer isn't needed for the design equation given above. An approximate solution will provide a starting point that can quickly be adjusted for optimum dimensions.

The added resonators have an effect on the lengths of all conductors, due to the capacitance between the conductors. Capacitance causes antennas to look electrically shorter, so

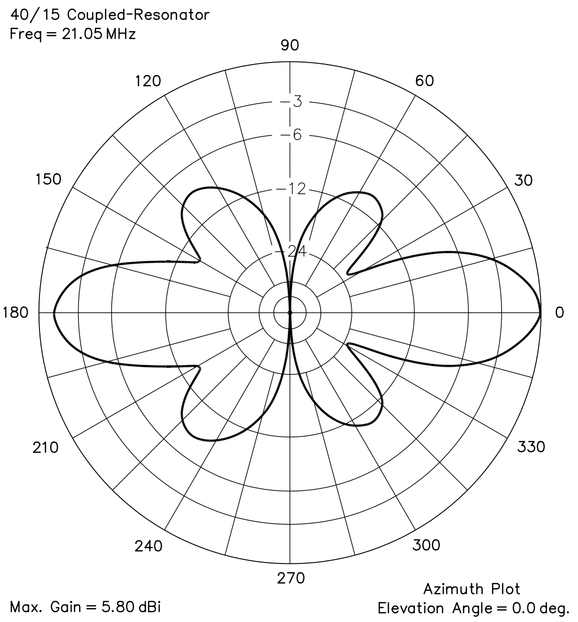

Fig 46—Radiation pattern for the special case of a C-R antenna with the additional frequency at the third harmonic of the main dipole resonant frequency.

each element needs to be about 1% or 2% longer than a simple dipole at the same frequency. As a rule of thumb, use 477/f (in feet) instead of the usual 468/f when calculating dipole length, and 239/f instead of 234/f for a λ/4 vertical.

A 30/17/12-Meter Dipole

To show how a C-R antenna is designed, let's build a dipole to cover all three WARC bands. We'll use #12 wire, which has a diameter of 0.08 inches, and the main dipole will be cut for the 10.1 MHz band. From the equation above, the spacing between the main dipole and the 18-MHz resonator should be 2.4 inches for 72 Ω, or 1.875 inches for 50 Ω. At 24.9 MHz, the spacing to the resonator for that band should be 2.0 inches for 72 Ω, or 1.62 inches for 50 Ω. Of course, this antenna will be installed over real ground, not in free space, so these spacing distances may not be exact. Plugging these numbers into your favorite antenna-modeling program will let you optimize the dimensions for installation at the height you choose.

For those of you who like to work with real antennas, not computer-generated ones, the predicted spacing is accurate enough to build an antenna with minimum trial-and-error. You should use a nice round number just larger than the calculated spacing for 50 Ω. For this antenna, K9AY decided that the right spacing for the desired height would be 2 inches for the 18 MHz resonator and 1.8 inches for the 24.9 MHz resonator. For simplicity of construction, he just used 2 inches for both, figuring that the worst he would get is a 1.2:1 SWR if the numbers were a little bit off. Like all dipoles, the impedance varies with height above ground, but the 2-inch spacing results in an excellent match on the two additional bands, at heights of more than 25 feet.

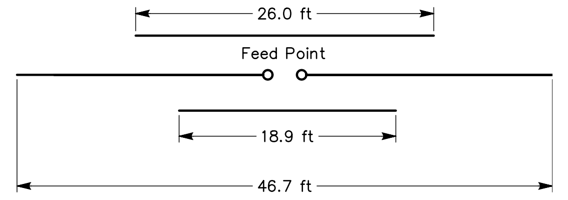

Fig 47—Dimensions of a C-R dipole for the 30, 17 and 12-meter bands.

The final dimensions of the dipole for 10.1, 18.068 and 24.89 MHz are shown in **Fig 47**. These are the final pruned lengths for a straight dipole installed at a height of about 40 feet. If you put up the antenna as an inverted V, you will need each wire to be a bit longer. Pruning this type of antenna is just like a dipole—if it's resonant too low in frequency, it's too long and the appropriate wire needs to be shortened. So, you can cut the wires just a little long to start with and easily prune them to resonance.

A final note: if you want to duplicate this antenna design, remember that the 2-inch spacing is just for #12 wire! The required spacing for a C-R antenna is related to the conductor diameter. This same antenna built with #14 wire needs under 1½-inch spacing, while a 1-inch aluminum-tubing version requires about 7-inch spacing.

Summary

The coupled-resonator principle is one more weapon in the antenna designer's arsenal. It's not the perfect method for all multiband antennas, but what the C-R principle offers is an alternative to traps and tuners, in exchange for using more wire or aluminum. Although a C-R antenna requires more complicated construction, its main attraction is in making a multiband antenna that can be built with no compromise in matching or efficiency.

HF Discone Antennas

The material in this section is adapted from an article by Daniel A. Krupp, W8NWF, in *The ARRL Antenna Compendium, Vol 5*. The name *discone* is a contraction of the words disc and cone. Although people often describe a discone by its design-center frequency (for example, a "20-meter discone"), it works very well over a wide frequency range, as much as several octaves. **Fig 48** shows a typical discone, constructed of sheet metal for UHF use. On lower frequencies, the sheet metal may be replaced with closely spaced wires and/or aluminum tubing.

The dimensions of a discone are determined by the lowest frequency of use. The antenna produces a vertically polarized signal at a low-elevation angle and it presents a good match for 50-Ω coax over its operating range. One advantage of the discone is that its maximum current area is near the top of the antenna, where it can radiate away from ground clutter. The cone-like skirt of the discone radiates the signal—radiation from the disc on top is minimal. This is because the currents flowing in the skirt wires essentially all go in the same direction, while the currents in the disc elements oppose each other and cancel out. The discone's omnidirectional characteristics make it ideal for roundtable QSOs or for a Net Control station.

Electrical operation of this antenna is very stable, with no changes due to rain or accumulated ice. It is a self-contained antenna—unlike a traditional ground-mounted vertical radiator, the discone does not rely on a ground-radial system for efficient operation. However, just like any other vertical

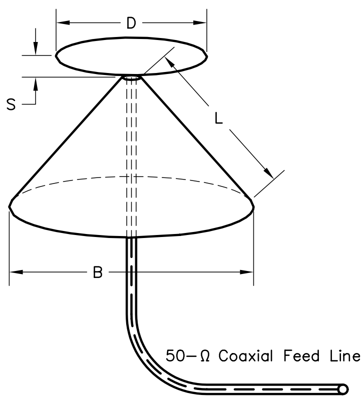

Fig 48—Diagram of VHF/UHF discone, using a sheet-metal disc and cone. It is fed directly with 50-Ω coax line. The dimensions L and D, together with the spacing S between the disc and cone, determine the frequency characteristics of the antenna. $L = 246 / f_{MHz}$ for the lowest frequency to be used. Diameter D should be from 0.67 to 0.70 of dimension L. The diameter at the bottom of the cone B is equal to L. The space S between disc and cone can be 2 to 12 inches, with the wider spacing appropriate for larger antennas.

antenna, the quality of the ground in the Fresnel area will affect the discone's far-field pattern.

Both the disc and cone are inherently balanced for wind loading, so torque caused by the wind is minimal. The entire cone and metal mast or tower can be connected directly to ground for lightning protection.

Unlike a trap vertical or a triband beam, discone antennas are not adjusted to resonate at a particular frequency in a ham band or a group of ham bands. Instead, a discone functions as a sort of high-pass filter, efficiently radiating RF all the way from the low-frequency design cutoff to the high-frequency limits imposed by the physical design.

While VHF discones have been available out-of-the-box for many years, HF discones are rare indeed. Some articles have dealt with HF discones, where the number of disc elements and cone wires was minimized to cut costs or to simplify construction. While the minimalist approach is fine if the sought-after results really are obtained, W8NWF believes in building his discones without compromise.

History of the Discone

The July 1949 and July 1950 issues of *CQ* magazine both contained excellent articles on discones. The first article, by Joseph M. Boyer, W6UYH, said that the discone was developed and used by the military during World War II. (See Bibliography.) The exact configuration of the top disc and cone was the brainchild of Armig G. Kandonian. Boyer described three VHF models, plus information on how to build them, radiation patterns, and most importantly, a detailed description of how they work. He referred to the discone as a type of "coaxial taper transformer."

The July 1950 article was by Mack Seybold, W2RYI. He described an 11-MHz version he built on his garage roof. The mast actually fit through the roof to allow lowering the antenna for service. Seybold stated that his 11-MHz discone would load up on 2 meters but that performance was down 10 dB compared to his 100-MHz Birdcage discone. He commented that this was caused by the relatively large spacing between the disc and cone. Actually, the performance degradation he found was caused by the wave angle lifting upward at high frequencies. The cone wires were electrically long, causing them to act like long wire antennas. See **Fig 49**.

W8NWF's First Discone: the A-Frame Discone

The first discone was one designed to cover 20 through 10 meters without requiring an antenna tuner. The cone assembly uses 18-foot long wires, with a 60° included apex angle and a 12-foot diameter disc assembly. See **Fig 50** and **Fig 51**. The whole thing was assembled on the ground, with the feed coax and all guys attached. Then with the aid of some friends, it was pulled up into position.

The author used a 40-foot tall wooden "A-frame" mast, made of three 22-foot-long 2 × 4s. He primed the mast with sealer and then gave it two coats of red barn paint to make it look nice and last a long time. The disc hub was a 12-inch length of 3-inch schedule-40 PVC plumbing pipe. The PVC is very tough, slightly ductile, and easy to drill and cut. PVC is well suited for RF power at the feed point of the antenna.

Three 12-foot by 0.375-inch OD pieces of 6061 aluminum, with 0.058-inch wall thickness, were used for the 12-foot diameter top disc. These were cut in half to make the center portions of the six telescoping spreaders. Four twelve foot by 0.250-inch OD (0.035-inch wall thickness) tubes were cut into 12 pieces, each 40 inches long. This gave extension tips for each end of the six spreaders.

See **Fig 52** for details on the disc hub assembly. W8NWF started by drilling six holes straight through the PVC for the six spreaders, accurately and squarely, starting

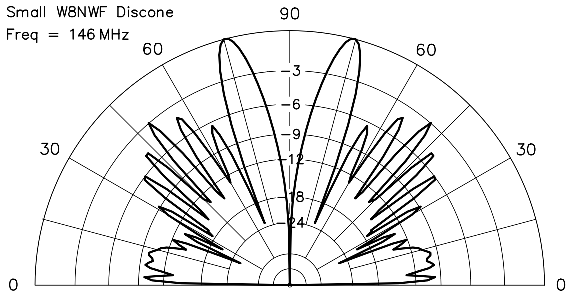

Fig49—Computed elevation plot over average ground for W8NWF's small discone at 146 MHz, ten times its design frequency range. The cone wires are acting as long-wire antennas, distorting severely the low-elevation angle response, even though the feed-point impedance is close to 50 Ω.

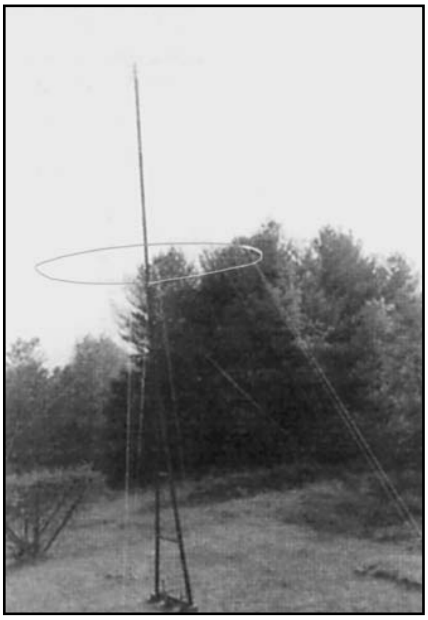

Fig 50—Photo of W8NWF's original A-frame mounted HF discone.

about two inches down from the top and spaced radially every 30°. Each hole is 0.375 inches below the plane of the previous one. Take great care in drilling—a poor job now will look bad from the ground up for a long time! It's a good idea to make up a paper template beforehand. Tape this to the PVC hub and then drill the holes, which should make for a close fit with the elements. If you goof, start over with a new piece of PVC—it's cheap.

Each six-foot spreader tube was secured exactly in the center to clear a 6-32 threaded brass rod that secured the elements mechanically and electrically. A two-foot long by ¼-inch OD wooden dowel was inserted into the middle of each six-foot length of tubing. The dowel added strength and also prevented crushing the element when the nuts on the threaded rod were tightened.

Insert the 40-inch long extensions four inches into each end of the six-foot spreaders. Mark and drill holes to pin the telescoping tips, plus holes big enough to clear #18 soft-drawn copper wire. This was for the inner circumferential wire for the disc. Drill a single hole for #18 wire about ¼ inch from each extension element tip, through which passes the outer circumferential wire. Finally, insert all 6-foot elements into the PVC hub and line up the holes in the center so the brass rod could be inserted through the middle to secure the elements.

The next step is to "chisel to fit" the top of my wooden mast to allow the PVC to slide down on it about six or seven inches. For convenience, place the whole mast assembly in a horizontal position on top of two clothesline poles and one stepladder.

Place the disc head assembly over the top of the mast, but don't secure it yet. This allows for rotation while adding the disc spreader extensions. A tip for safety: tie white pieces of cloth to the ends of elements near eye level. Just remember to remove them before raising the antenna.

For a long-lasting installation, use an anti-corrosion compound, such as Penetrox, when assembling the aluminum antenna elements. As the extensions are added, secure them in the innermost of the two holes with a short piece of #18 wire. Then run a wire through the remaining holes looping each element as you go. This gives added support laterally to the elements. Next add a #18 wire to the tips of the extensions in the same fashion. This provides even more

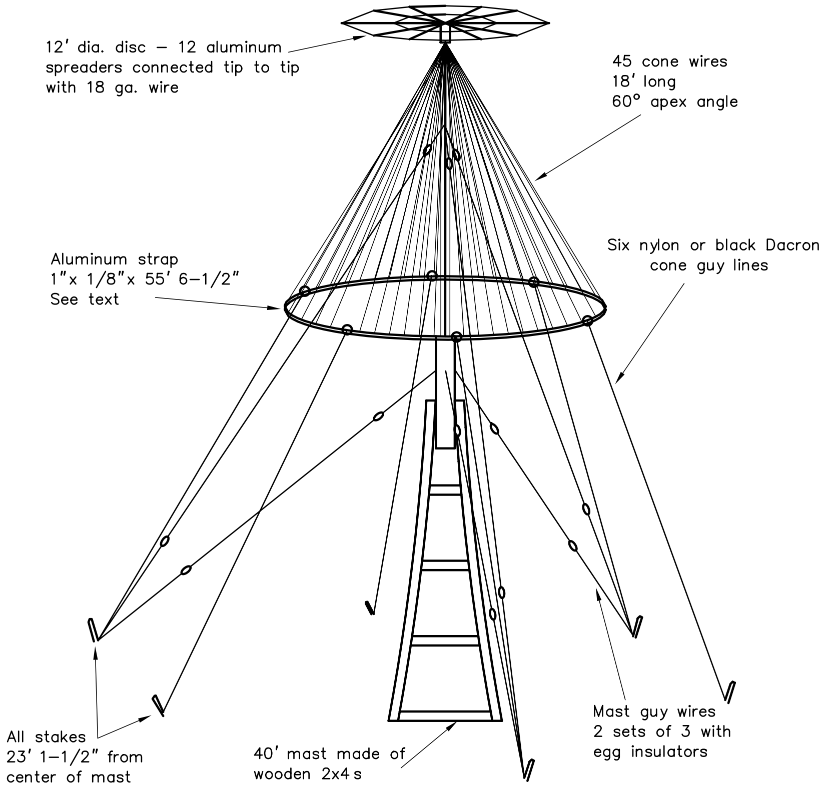
12' dia. disc — 12 aluminum spreaders connected tip to tip with 18 ga. wire
45 cone wires 18' long 60° apex angle
Aluminum strap 1"x 1/8"x 55' 6-1/2" See text
Six nylon or black Dacron cone guy lines
All stakes 23' 1-1/2" from center of mast
40' mast made of wooden 2x4 s
Mast guy wires 2 sets of 3 with egg insulators

Fig 51—Detailed drawing of the A-frame discone for 14 to 30 MHz. The disc assembly at the top of the A-frame is 12 feet in diameter. There are 45 cone wires, each 18 feet long, making a 60° included angle of the cone. This antenna works very well over the design frequency range.

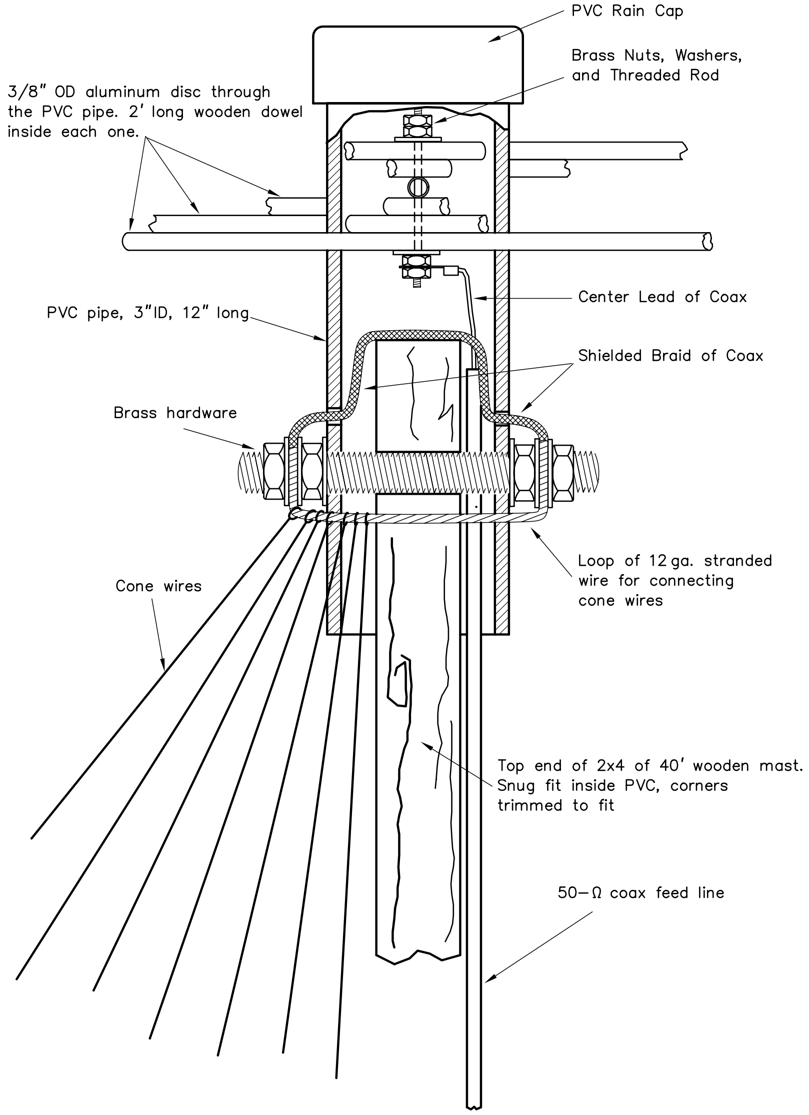
PVC Rain Cap
Brass Nuts, Washers, and Threaded Rod
3/8" OD aluminum disc through the PVC pipe. 2' long wooden dowel inside each one.
Center Lead of Coax
PVC pipe, 3"ID, 12" long
Shielded Braid of Coax
Brass hardware
Loop of 12 ga. stranded wire for connecting cone wires
Cone wires
Top end of 2x4 of 40' wooden mast. Snug fit inside PVC, corners trimmed to fit
50-Ω coax feed line

Fig 52—Details of the top hub for the A-frame discone. The three-inch PVC pipe was drilled to hold the six spreaders making up the top disc. Connections for the shield of the feed coax were made to the disc. The coax center conductor was connected to the cone-wire assembly by means of a loop of #12 stranded wire encircling the outside of the PVC hub.

physical stability as well as making electrical connections.

Next, pin the PVC disk hub to the wooden mast with a ⅜-inch threaded rod. This is also the point where the cone wires are attached, using a loop of #12 stranded copper wire around the PVC. Solder each cone wire to this loop, together with the coax shield braid. Make sure the loop of #12 wire is large enough to make soldering possible without burning the PVC with the soldering iron.

Connect the coax center conductor to the disc assembly by securing it with the same 6-32 threaded rod that ties all the disc elements together. Make sure to use coax-seal compound to keep moisture out of the coax. The coax is then fed down the mast and secured in a few places to provide strain relief and to keep it out of the way of the cone wires.

Use two sets of three guy wires. Break these up with egg insulators, just to be sure there won't be any interaction with the antenna. Use 45 wires of #18 soft-drawn copper wire for the cone, 18 feet long each. Cut them a little long so they can be soldered to the connecting loop.

A difficult task is now at hand—keeping all the cone wires from getting tangled! Solder each of the 45 cone wires to the loop of #12 wire, spacing each wire about ¼ inch from the last one for an even distribution all the way around.

The cone base is 18 feet in diameter to provide a 60° included angle. At the base of the cone, use five 12-foot long aluminum straps, 1 inche wide by ⅛ inch thick, overlapping 8¼ inches and fastened together with aluminum rivets. Drill holes along the strap every 15 inches to secure the cone wires.

Make sure to handle the aluminum strap carefully while fastening the cone wire ends; too sharp of a bend could possibly break it. Fasten six small-diameter nylon lines to the cone-base aluminum strap to stabilize the cone. These cone-guys share the same guy stakes as the mast guy lines.

After cutting the nylon lines, heat the frayed ends of each with a small flame to prevent unraveling. Apply several coats of clear protective spray to the disk head assembly, after checking that all hardware is tight. A rain cap at the top of the PVC disc hub completes construction.

Putting It All Up

You are going to need a lot of help now to raise this antenna. Have the whole process fully thought out before trying to raise it. You should have the spot selected for the base of the mast and some pipes driven into the ground to prevent the mast from slipping sideways as it is being pulled up. The three guy stakes should be in place, 23 feet, 1½ inches from mast center. Of course, the guys should have been cut to the correct length, with some extra. Be sure the coax transmission line will come off the mast where it should. A long length of rope to an upper and lower guy line is used to pull up the whole works.

The author used an old trick of standing an extension ladder vertically near the antenna base with the pull lines looped over the top rung to get a good lift angle. The weight added to the mast from the antenna disc assembly and cone wires is about 26 pounds, most of it from the cone assembly. Use two strong people to pull up the antenna slowly so that the other helpers on the guy wires and cone guy lines have time to move about as required. As the antenna rises to the vertical position, if there are no snafus, the guy lines can be secured. Then tie the six cone lines to stakes.

A Really Big Discone

When an opportunity arose to buy a 64-foot self-supporting TV tower, the author jumped at the chance to implement a full 7 to 30-MHz discone. His new tower had eight sections, each eight feet long. Counting the overlap between sections, the cone wires would come off the tower at about the 61.5-foot mark.

W8NWF took some liberties with the design of this larger discone compared to the first one, which he had done strictly "by the book." The first change was to make the cone wires 70 feet long, even though the formula said they should be 38 feet long. Further, the cone wires would not be connected together at the bottom. With the longer cone wires, he felt that 75 and 80-meter operation might be a possibility.

The second major change was to widen the apex angle out from 60° to about 78°. Modeling said this should produce a flatter SWR over the frequency spectrum and would also give a better guy system for the tower.

The topside disc assembly would be 27 feet in diameter and have 16 radial spreaders, using telescoping aluminum tubing tapering from ⅝ to ½ to ⅜ inches OD. All spreaders were made from 0.058-inch wall thickness 6063-T832 aluminum tubing, available from Texas Towers. A section of 10-inch PVC plumbing pipe would be used as the hub for construction of the disc assembly.

Construction Details for the Large Discone.

While installing the tower, the author had left the top section on the ground. This allowed him to fit the disc head assembly precisely to it. **Fig 53** shows the overall plan for the large discone. The 10-inch diameter PVC hub was designed to slip over the tower top section, but was a little too large. So a set of shims was installed on the three legs at the top of the tower for a just-right fit. Drilling the PVC pipe for the eight ⅝-inch OD elements was started about an inch down from the top. W8NWF purposely staggered the drilled holes in the same fashion as the hub for the smaller antenna. See **Fig 54**.

Again, three-foot sections of ½-inch wooden dowel were used to strengthen the ⅝-inch center portion of each spreader. Instead of using a loop of #12 wire for connecting the cone wires, as had been done on the smaller discone, he drilled 36 holes in the PVC hub. These holes are small enough so

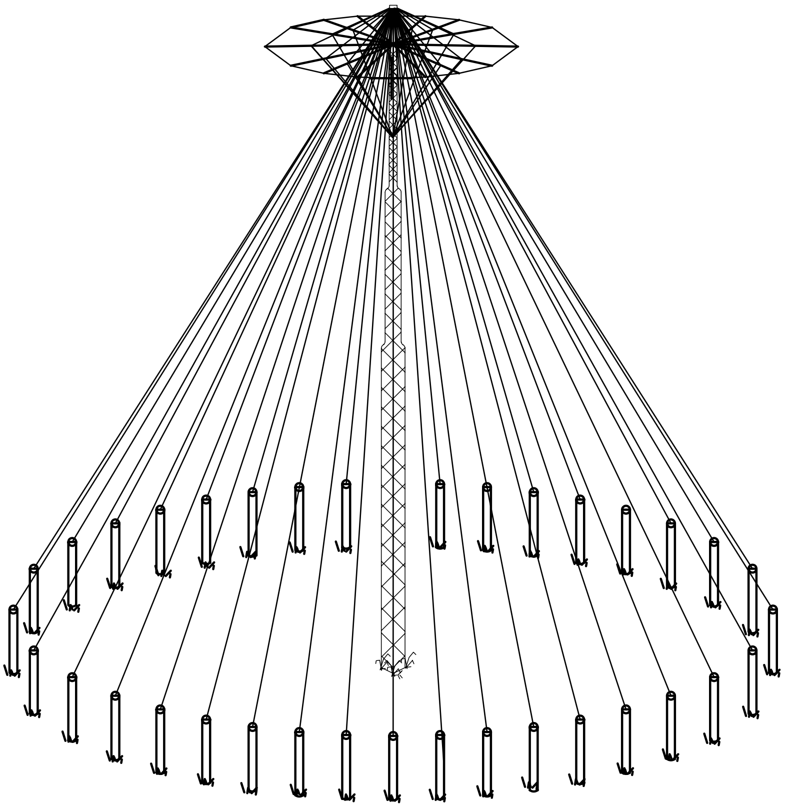

Fig 53—The large W8NWF discone, designed for operation from 7 to 14 MHz, but useable with a tuning network in the shack for 3.8 MHz.

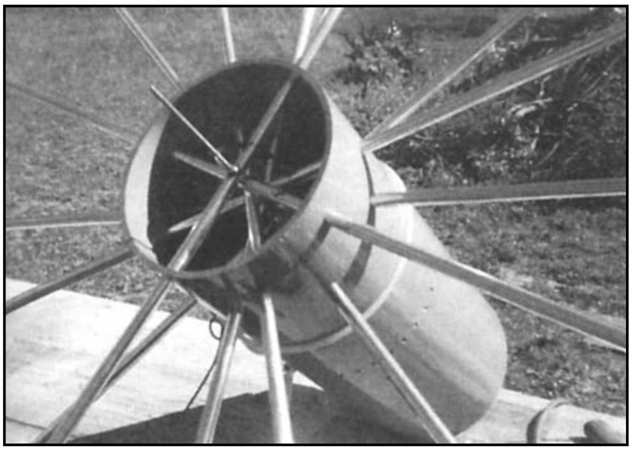

Fig 54—Photo showing details of the hub assembly for the large discone, including the threaded brass rod that connects the radial spreaders together. The 10-inch PVC pipe is drilled to accommodate the radial spreaders. Each spreader is reinforced with a three-foot long wooden dowel inside for crush resistance. Note the row of holes drilled below the lowest spreader. Each of the 36 cone wire passes through one of these holes.

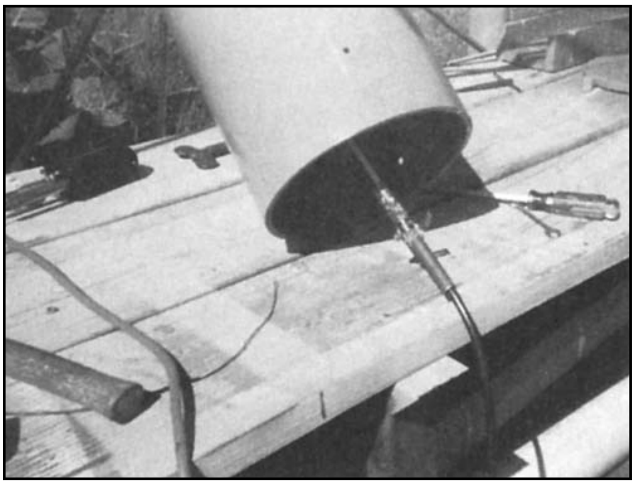

Fig 55—Details of the copper pipe slipped over the feed coax. The coax shield has been folded back over the copper pipe and secured with two stainless-steel hose clamps. The cone wires are also laid against the copper pipe and secured with additional hose clamps.

that the PVC hub would not be weakened appreciably. He drilled the circles of holes for the cone wires about 6 inches below the disc spreaders.

He prepared a three-foot long piece of RG-213 coax, permanently fastened on one end to the antenna, with a female type-N connector at the other end. Type-N fittings were used because of their superior waterproofing abilities. The coax center lead was connected with a terminal lug under a nut on the brass threaded rod securing the disc spreaders. The coax shield braid was folded back over a six-inch long copper pipe and clamped to it with a stainless-steel hose clamp. See **Fig 55** for details.

The plan was that after the top disc assembly had been hoisted up and attached at the top of the tower, individual cone wires would be fed, one at a time, through the small holes drilled in the PVC. They were to be laid against the copper pipe and secured with stainless-steel hose clamps.

The ½ and ⅜-inch OD spreader extension tips were secured in place with two aluminum pop-rivets at each joint. Again, the author used anti-oxidant compound on all spreader junctions. He drilled a hole horizontally near the tip of each ⅜-inch tip all around the perimeter to allow a #8 aluminum wire to circle the entire disc. A small stainless-steel sheet-metal screw was threaded into the end of each element to secure the wire.

In parallel with the aluminum wire, a length of small-diameter black Dacron line was run, securing it in a couple of places between each set of spreaders with UV-resistant plastic tie-wraps. The reason for doing this was to hold the aluminum wire in position and to prevent it from dangling, in case it should break some years in the future. Two coats of clear protective spray were applied for protection.

A truss system helps prevent the disc from sagging due to its own weight. See **Fig 56** for details. This shows the completed disk assembly mounted on the

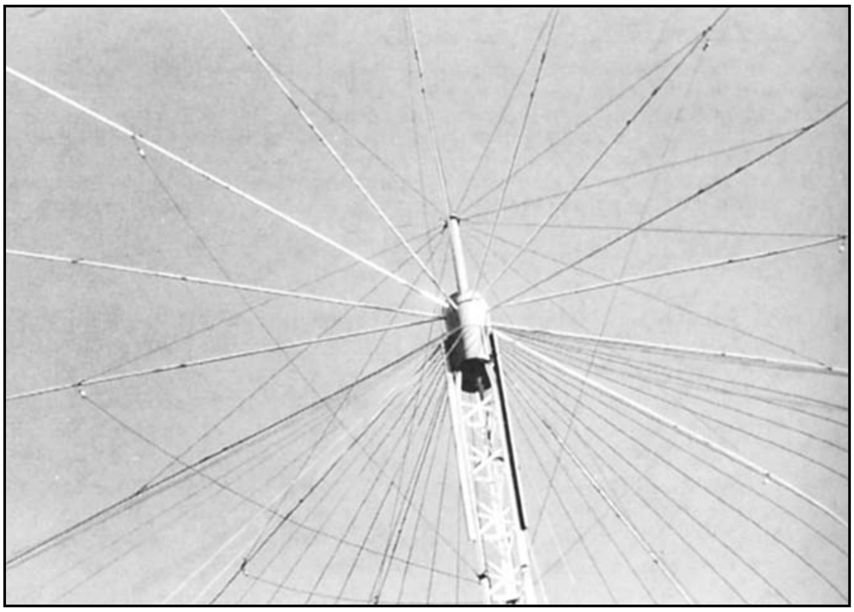

Fig 56—Photo of the spreader hub assembly, showing the truss ropes above and below the radial spreaders. This is a very rugged assembly!

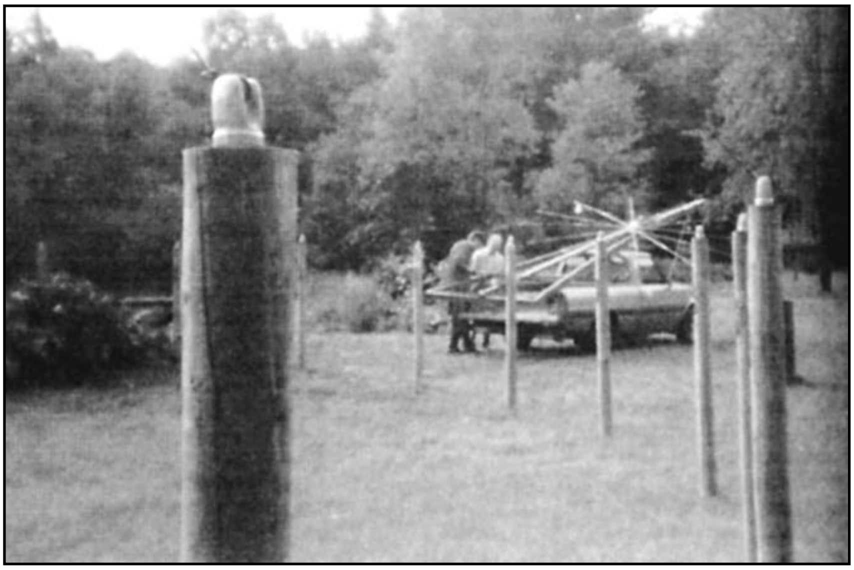

Fig 57—Photo showing some of the fence posts used to hold individual cone wires to keep them off the ground and out of harm's way. The truck in the background is carting away the A-frame discone for installation at KA8UNO's QTH.

top of the tower. A 3-foot length of 2-inch PVC pipe was used for a truss mast above the disc assembly, notching the bottom of the pipe so that it would form a saddle over the top couple of spreaders. This gave a good foothold. He cut a circle of thin sheet aluminum to fit over the 10-inch PVC to serve as a rain cap. The cap has a hole in the center for the two-inch PVC truss mast to pass through, thereby holding it down tight. The author sprayed a few light coats of paint over the PVC for protection from ultraviolet radiation from the sun.

Sixteen small-diameter black Dacron ropes were connected at the top of the truss support mast, with the other ends fastened to the disc spreaders, halfway out. Another rain cap was added to the top of the two-inch PVC truss mast. Eight lengths of the same small diameter Dacron rope were added halfway out the length of every other spreader. These ropes are meant to be tied back to the tower, to prevent updrafts from blowing the disc assembly upwards. Small egg insulators were used near the spot where the eight bottom trusses were tied to the disc spreaders, just to be sure there would be no RF leakage in rainy weather.

Hoisting the completed disc assembly to the top of the tower can be done easily, with the assistance of at least two others. The trickiest part is to get the disc assembly from its position sitting flat on the ground to the vertical position needed for hoisting it up the tower without damaging it. The disc assembly weighs about 35 pounds. Someone at the top of the tower will receive the disc as it is hoisted up by gin pole, and can mount it on the tower top.

You should prepare three 6-foot long metal braces going over the outside of the PVC to fasten to the tower legs. They really beef things up.

In plastic irrigation pipe buried between the house and tower base, the author ran 100 feet of 9086 low-loss coax to the shack. For cone wires, he was able to obtain some #18 copperclad steel wire, with heavy black insulation that looked a lot like neoprene. The cone system takes a lot of wire: 36 × 70 feet = 2520 feet, plus some extra at each end for termination. You'd be well advised to look around at hamfests to save money.

As each cone wire was connected at the top of the tower, a helper should place the other end at its proper spot below. The lower end of each cone wire is secured to an insulator screwed into a fencepost. See **Fig 57**. There are 36 treated-pine fenceposts, each standing about 5½ feet tall, 45 feet from the tower base to hold the lower end of the cone wires. This makes mowing the grass easier and the cone wires are less likely to be tripped over too.

On the final trip down the tower, the eight Dacron downward-truss lines were tied back to the tower about 6 feet below the disc assembly. The author's tower has three ground rods driven near the base, connected with heavy copper wire to the three tower legs.

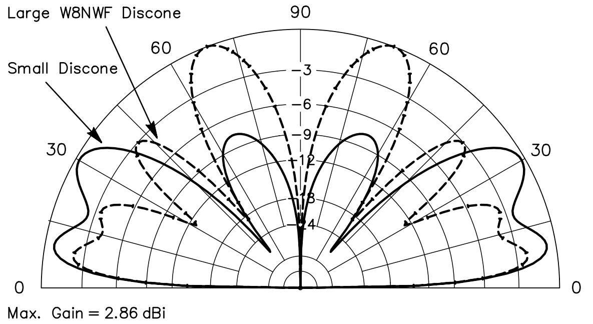

Fig 58—Computed patterns showing elevation response of small discone at 28.5 MHz compared to that of the larger discone at 28.5 MHz. The cone wires are clearly too long for efficient operation on 10 meters, producing unwanted high-angle lobes that rob power from the desirable low-elevation angles.

Performance Tests

On the air tests proved to be very satisfying. Loading up on 40 meters was easy—the SWR was 1:1 across the entire band. W8NWF can work all directions very well and receives excellent signal reports from DX stations. When he switches to his long (333 foot) center-fed dipole for comparison, he finds the dipole is much noisier and that received signals are weaker. During the daytime, nearby stations (less than about 300 to 500 miles) can be louder with the dipole, but the discone can work them just fine also.

The author happily reports that this antenna even works well on 75 meters. As you might expect, it doesn't present a 1:1 match. However, the SWR is between 3.5:1 and 5.5:1 across the band. W8NWF uses an antenna tuner to operate the discone on 75. It seems to get out as well on 75 as it does on 40 meters.

The SWR on 30 meters is about 1.1:1. On 20 meters the SWR runs from 1.05:1 at 14.0 MHz to 1.4:1 at 14.3 MHz. The SWR on the 17, 15, 12 and 10-meter bands varies, going up to a high of 3.5:1 on 12 meters.

Radiation Patterns for the Discones

From modeling using *NEC/Wires* by K6STI, W8NWF verified that the low-angle performance for the bigger antenna is worse than that for the smaller discone on the upper frequencies. See **Fig 58** for an elevation-pattern comparison on 10 meters for both antennas, with average ground constants. The azimuth patterns are simply circles. Radiation patterns produced by antenna modeling programs are very helpful to determine what to expect from an antenna.

The smaller discone, which was built by the book, displays good, low-angle lobes on 20 through 10 meters. The frequency range of 14 through 28 MHz is an octave's worth of coverage. It met his expectations in every way by covering this frequency span with low SWR and a low angle of radiation.

The bigger discone, with a modified cone suitable for use on 75 meters, presents a little different story. The low-angle lobe on 40 meters works well, and 75-meter performance also is good, although an antenna tuner is necessary on this band. The 30-meter band has a good low-angle lobe but secondary high-angle lobes are starting to hurt performance. Note that 30 meters is roughly three times the design frequency of the cone. On 20 and 17 meters there still are good low-angle lobes but more and more power is wasted in high-angle lobes.

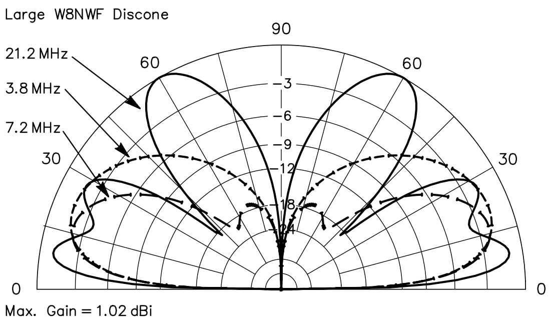

Fig 59—Computed elevation-response patterns for the larger W8NWF discone for 3.8, 7.2 and 21.2 MHz operation. Again, as in Fig 58, the pattern degrades at 21.2 MHz, although it is still reasonably efficient, if not optimal.

The operation on 15, 12, and 10 meters continues to worsen for the larger discone. The message here is that although a discone may have a decent SWR as high as 10 times the design frequency, its radiation pattern is not necessarily good for low-angle communications. See **Fig 59** for a comparison of elevation patterns for 3.8, 7.2 and 21.2 MHz on the larger discone.

A discone antenna built according to formula will work predictably and without any adjustments. One can modify the antenna's cone length and apex angle without fear of rendering it useless. The broadband feature of the discone makes it attractive to use on the HF bands. The low angle of radiation makes DX a real possibility, and the discone is also much less noisy on receive than a dipole.

Probably the biggest drawback to an HF discone is its bulky size. There is no disguising this antenna! However, if you live in the countryside you should be able to put up a nice one.

Harmonic Radiation from Multiband Antennas

Since a multiband antenna is intentionally designed for operation on a number of different frequencies, any harmonics or spurious frequencies that happen to coincide with one of the antenna resonant frequencies will be radiated with very little, if any, attenuation. Particular care should be exercised, therefore, to prevent such harmonics from reaching the antenna.

Multiband antennas using tuned feeders have a certain inherent amount of built-in protection against such radiation, since it is nearly always necessary to use a tuned coupling circuit (antenna tuner) between the transmitter and the feeder. This adds considerable selectivity to the system and helps to discriminate against frequencies other than the desired one.

Multiple dipoles and trap antennas do not have this feature, since the objective in design is to make the antenna show as nearly as possible the same resistive impedance in all the amateur bands the antenna is intended to cover. It is advisable to conduct tests with other amateur stations to determine whether harmonics of the transmitting frequency can be heard at a distance of, say, a mile or so. If they can, more selectivity should be added to the system since a harmonic that is heard locally, even if weak, may be quite strong at a distance because of propagation conditions.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2, Antenna Fundamentals.

H. B. Barkley, *The Open-Sleeve As A Broadband Antenna*, Technical Report No. 14, U.S. Naval Postgraduate School, Monterey, CA, Jun 1955.

W. M. Bell, “A Trap Collinear Antenna,” *QST*, Aug 1963, pp 30-31.

J. S. Belrose, “The HF Discone Antenna,” *QST*, Jul 1975, pp 11-14, 56.

J. Belrose and P. Bouliane, “The Off-Center-Fed Dipole Revisited: A Broadband, Multiband Antenna,” *QST*, Oct 1990, p 40.

H. J. Berg, “Multiband Operation with Paralleled Dipoles,” *QST*, Jul 1956.

E. L. Bock, J. A. Nelson and A. Dorne, “Sleeve Antennas,” *Very High Frequency Techniques*, H. J. Reich, ed. (New York: McGraw-Hill, 1947), Chap 5.

J. T. Bolljahn and J. V. N. Granger, “Omnidirectional VHF and UHF Antennas,” *Antenna Engineering Handbook* , H. Jasik, ed. (New York: McGraw-Hill, 1961) pp 27-32 through 27-34.

J. M. Boyer, W6UYH, “Discone--40 to 500 Mc Skywire,” *CQ*, July 1949, p 11.

G. A. Breed, K9AY, “Multi-Frequency Antenna Technique Uses Closely-Coupled Resonators,” *RF Design*, November 1994. US Patent 5,489,914, “Method of Constructing Multiple-Frequency Dipole or Monopole Antenna Elements Using Closely-Coupled Resonators,” Gary A. Breed, Feb 6, 1996.

G. H. Brown, “The Phase and Magnitude of Earth Currents Near Radio Transmitting Antennas,” *Proc. IRE*, Feb 1935.

G. H. Brown, R. F. Lewis and J. Epstein, “Ground Systems as a Factor in Antenna Efficiency,” *Proc. IRE* , Jun 1937, pp 753-787.

C. L. Buchanan, “The Multimatch Antenna System,” *QST*, Mar 1955.

R. A. Cox, “The Open-Sleeve Antenna,” *CQ*, Aug 1983, pp 13-19.

D. DeMaw, “Lightweight Trap Antennas—Some Thoughts,” *QST*, Jun 1983, p 15.

W. C. Gann, “A Center-Fed ‘Zepp’ for 80 and 40,” *QST*, May 1966.

D. Geiser, “An Inexpensive Multiband VHF Antenna,” *QST*, Dec 1978, pp 28-29.

A. Greenberg, “Simple Trap Construction for the Multiband Antenna,” *QST*, Oct 1956.

G. L. Hall, “Trap Antennas,” Technical Correspondence, *QST*, Nov 1981, pp 49-50.

W. Hayward, “Designing Trap Antennas,” Technical Correspondence, *QST*, Aug 1976, p 38.

R. H. Johns, “Dual-Frequency Antenna Traps,” *QST*, Nov 1983, p 27.

A. G. Kandoian, “Three New Antenna Types and Their Applications,” *Proc IEE*, Vol 34, Feb 1946, pp 70W-75W.

R. W. P. King, *Theory of Linear Antennas* (Cambridge, MA: Harvard Univ Press, 1956), pp 407-427.

W. J. Lattin, “Multiband Antennas Using Decoupling Stubs,” *QST*, Dec 1960.

W. J. Lattin, “Antenna Traps of Spiral Delay Line,” *QST*, Nov 1972, pp 13-15.

M. A. Logan, “Coaxial-Cable Traps,” Technical Correspondence, *QST*, Aug 1985, p 43.

J. R. Mathison, “Inexpensive Traps for Wire Antennas,” *QST*, Aug 1977, p 18.

L. McCoy, “An Easy-to-Make Coax-Fed Multiband Trap Dipole,” *QST*, Dec 1964.

M. Mims, “The Mims Signal Squirter,” *QST*, Dec 1939, p 12.

G. E. O’Neil, “Trapping the Mysteries of Trapped Antennas,” *Ham Radio*, Oct 1981, pp 10-16.

B. Orr, W6SAI, “Radio FUNdamentals,” The Open-Sleeve Dipole, *CQ*, February 1995.

E. W. Pappenfus, “The Conical Monopole Antenna,” *QST*, Nov 1966, pp 21-24.

L. Richard, “Parallel Dipoles of 300-Ohm Ribbon,” *QST*, Mar 1957.

R. Severns, N6LF, “A Wideband 80-Meter Dipole,” *QST*, July 1995.

R. R. Shellenback, “Try the ‘TJ’,” *QST*, Jun 1982, p 18.

R. R. Shellenback, “The JF Array,” *QST*, Nov 1982, p 26.

H. Scholle and R. Steins, “Eine Doppel-Windom Antenna fur Acht Bander,” *cq-DL*, Sep 1983, p 427. (In English: *QST*, Aug 1990, pp 33-34.)

T. H. Schiller, N6BT, President, Force 12, US Patent 5,995,061, “No loss, multi-band, adaptable antenna,” Nov 30, 1999.

M. Seybold, W2RYI, “The Low-Frequency Discone,” *CQ*, July 1950, p 13.

D. P. Shafer, “Four-Band Dipole with Traps,” *QST*, Oct 1958.

R. C. Sommer, “Optimizing Coaxial-Cable Traps,” *QST*, Dec 1984, p 37.

L. Varney, “The G5RV Multiband Antenna . . . Up-to-Date,” *The ARRL Antenna Compendium*, *Vol 1* (Newington: ARRL, 1985), pp 86.

R. Wilson, AL7KK, “The Offset Multiband Trapless Antenna (OMTA),” *QST*, October 1995.

L. G. Windom, “Notes on Ethereal Antennas,” *QST*, Sep 1929, pp 19-22, 84.

W. I. Orr, W6SAI, editor, “The Low-Frequency Discone,” *RADIO Handbook*, 14th Edition, (Editors and Engineers, 1956), p 369.

Chapter 8

Multielement Arrays

The gain and directivity offered by an array of elements represents a worthwhile improvement both in transmitting and receiving. *Power gain* in an antenna is the same as an equivalent increase in the transmitter power. But unlike increasing the power of your own transmitter, antenna gain works equally well on signals received from the favored direction. In addition, the directivity reduces the strength of signals coming from the directions not favored, and so helps discriminate against interference.

One common method of obtaining gain and directivity is to combine the radiation from a group of $\lambda/2$ dipoles to concentrate it in a desired direction. A few words of explanation may help make it clear how power gain is obtained.

In **Fig 1**, imagine that the four circles, A, B, C and D, represent four dipoles so far separated from each other that the coupling between them is negligible. Also imagine that point P is so far away from the dipoles that the distance from P to each one is exactly the same (obviously P would have to be much farther away than it is shown in this drawing). Under these conditions the fields from all the dipoles will add up at P if all four are fed RF currents in the same phase.

Let us say that a certain current, I, in dipole A will produce a certain value of field strength, E, at the distant point P. The same current in any of the other dipoles will produce the same field at P. Thus, if only dipoles A and B are operating, each with a current I, the field at P will be 2E. With A, B and C operating, the field will be 3E, and with all four operating with the same I, the field will be 4E. Since the power received at P is proportional to the square of the field strength, the relative power received at P is 1, 4, 9 or 16, depending on whether one, two, three or four dipoles are operating.

Now, since all four dipoles are alike and there is no coupling between them, the same power must be put into each in order to cause the current I to flow. For two dipoles the relative power input is 2, for three dipoles it is 3, for four dipoles 4, and so on. The actual gain in each case is the relative received (or output) power divided by the relative input power. Thus we have the results shown in **Table 1**. The power ratio is directly proportional to the number of elements used.

It is well to have clearly in mind the conditions under which this relationship is true:

1) The fields from the separate antenna elements must be in-phase at the receiving point.

2) The elements are identical, with equal currents in all elements.

3) The elements must be separated in such a way that the current induced in one by another is negligible; that is, the radiation resistance of each element must be the same as it would be if the other elements were not there.

Very few antenna arrays meet all these conditions exactly. However, the power gain of a directive array using dipole elements with optimum values of element spacing is approximately proportional to the number of elements.

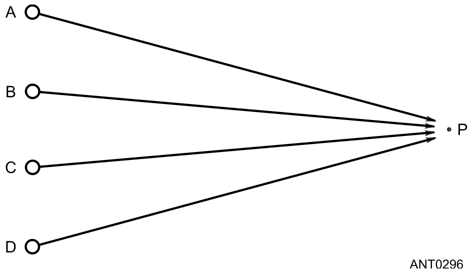

Fig 1—Fields from separate antennas combine at a distant point, P, to produce a field strength that exceeds the field produced by the same power in a single antenna.

Table 1
Comparison of Dipoles with Negligible Coupling (See Fig 1)

Dipoles	*Relative Output Power*	*Relative Input Power*	*Power Gain*	*Gain in dB*
A only	1	1	1	0
A and B	4	2	2	3
A, B and C	9	3	3	4.8
A, B, C and D	16	4	4	6

Another way to say this is that a gain of approximately 3 dB will be obtained each time the number of elements is doubled, assuming the proper element spacing is maintained. It is possible, though, for an estimate based on this rule to be in error by a ratio factor of two or more (gain error of 3 dB or more), especially if mutual coupling is *not* negligible.

DEFINITIONS

An *element* in a multi-element directive array is usually a λ/2 radiator or a λ/4 vertical element above ground. The length is not always an exact electrical half or quarter wavelength, because in some types of arrays it is desirable that the element show either inductive or capacitive reactance. However, the departure in length from resonance is ordinarily small (not more than 5% in the usual case) and so has no appreciable effect on the radiating properties of the element.

Antenna elements in multi-element arrays of the type considered in this chapter are always either *parallel*, as in **Fig 2A** , or *collinear* (end-to-end), as in Fig 2B. Fig 2C shows an array combining both parallel and collinear elements. The elements can be either horizontal or vertical, depending on whether horizontal or vertical polarization is desired. Except for space communications, there is seldom any reason for mixing polarization, so arrays are customarily constructed with all elements similarly polarized.

A *driven element* is one supplied power from the transmitter, usually through a transmission line. A *parasitic element* is one that obtains power solely through coupling to another element in the array because of its proximity to such an element.

A *driven array* is one in which all the elements are driven elements. A *parasitic array* is one in which one or more of the elements are parasitic elements. At least one element must be a driven element, since you must somehow introduce power into the array.

A *broadside array* is one in which the principal direction of radiation is perpendicular to the axis of the array and to the plane containing the elements, as shown in **Fig 3**. The elements of a broadside array may be collinear, as in Fig 3A, or parallel (two views in Fig 3B).

An *end-fire array* is one in which the principal direction of radiation coincides with the direction of the array axis. This definition is illustrated in **Fig 4**. An end-fire array must consist of parallel elements. They cannot be collinear, as λ/2 elements do not radiate straight off their ends. A Yagi is a familiar form of an end-fire array.

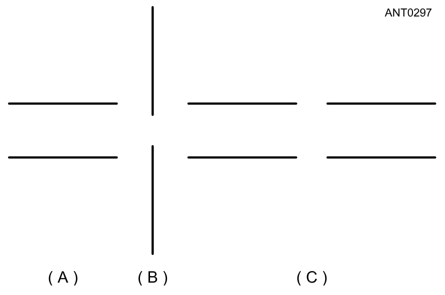

Fig 2—At A, parallel and at B, collinear antenna elements. The array shown at C combines both parallel and collinear elements.

A *bidirectional array* is one that radiates equally well in either direction along the line of maximum radiation. A bidirectional pattern is shown in **Fig 5A**. A *unidirectional array* is one that has only one principal direction of radiation, as the pattern in Fig 5B shows.

The *major lobes* of the directive pattern are those in which the radiation is maximum. Lobes of lesser radiation intensity are called *minor lobes*. The *beamwidth* of a directive antenna is the width, in degrees, of the major lobe between the two directions at which the relative radiated power is equal to one half its value at the peak of the lobe. At these *half-power points* the field intensity is equal to 0.707 times its maximum value, or in other words, is down 3 dB from the maximum. **Fig 6** shows a lobe having a beamwidth of 30°.

Unless specified otherwise, the term *gain* as used in this

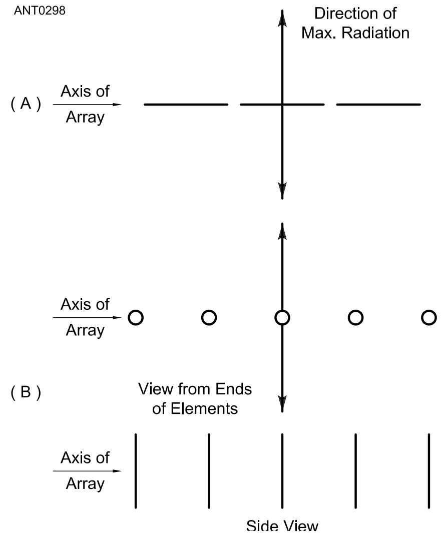

Fig 3—Representative broadside arrays. At A, collinear elements, with parallel elements at B.

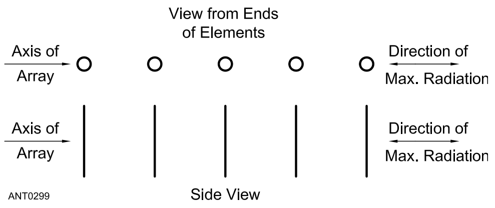
View from Ends of Elements
Axis of Array
Direction of Max. Radiation
Axis of Array
Direction of Max. Radiation
ANT0299
Side View

Fig 4—An end-fire array. Practical arrays may combine both broadside directivity (Fig 3) and end-fire directivity, including both parallel and collinear elements.

section is the power gain over an isotropic radiator in free space. The gain can also be compared with a λ/2 dipole of the same orientation and height as the array under discussion and having the same power input. Gain may either be measured experimentally or determined by calculation. Experimental measurement is difficult and often subject to considerable error, for two reasons. First, errors normally occur in measurement because the accuracy of simple RF measuring equipment is relatively poor—even high-quality instruments suffer in accuracy compared with their low-frequency and dc counterparts. And second, the accuracy depends considerably on conditions—the antenna site, including height, terrain characteristics, and surroundings—under which the measurements are made.

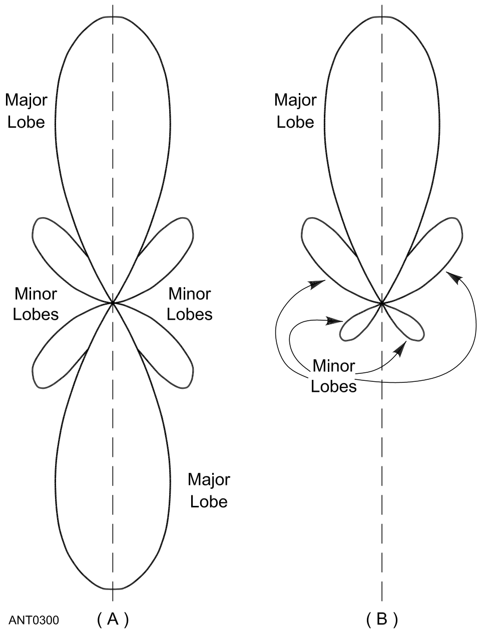

Fig 5—At A, typical bidirectional pattern and at B, unidirectional directive pattern. These drawings also illustrate the application of the terms *major* and *minor* to the pattern lobes.

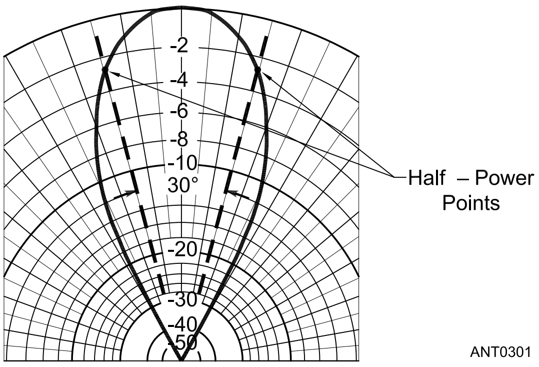

Fig 6—The width of a beam is the angular distance between the directions at which the received or transmitted power is half the maximum power (–3 dB). Each angular division of the pattern grid is 5°.

Calculations are frequently based on the measured or theoretical directive patterns of the antenna (see Chapter 2). The theoretical gain of an array may be determined approximately from:

$$G = 10 \log \frac{41{,}253}{\theta_H \ \theta_V} \qquad \text{(Eq 1)}$$

where

G = decibel gain over a dipole in its favored direction
θ_H = horizontal half-power beamwidth in degrees
θ_V = vertical half-power beamwidth in degrees.

This equation, strictly speaking, applies only to lossless antennas having approximately equal and narrow E- and H-plane beam widths—up to about 20°—and no large minor lobes. The E and H planes are discussed in Chapter 2. The error may be considerable when the formula is applied to simple directive antennas having relatively large beam widths. The error is in the direction of making the calculated gain larger than the actual gain.

Front-to-back ratio (F/B) is the ratio of the power radiated in the favored direction to the power radiated in the opposite direction. See Chapter 11 for a discussion of front-to-back ratio, and its close cousin, *worst-case front-to-rear ratio*.

Phase

The term *phase* has the same meaning when used in connection with the currents flowing in antenna elements as it does in ordinary circuit work. For example, two currents are in-phase when they reach their maximum values, flowing in the same direction, at the same instant. The direction of current flow depends on the way in which power is applied to the element.

This is illustrated in **Fig 7**. Assume that by some means an identical voltage is applied to each of the elements at the ends marked A. Assume also that the coupling between the elements is negligible, and that the instantaneous polarity of the voltage is such that the current is flowing away from the point at which the voltage is applied. The arrows show the assumed current directions. Then the currents in elements 1 and 2 are in-phase, since they are flowing in the same direction in space and are caused by the same voltage. However, the current in element 3 is flowing in the *opposite* direction in space because the voltage is applied to the opposite end of the element. The current in element 3 is therefore 180° out-of-phase with the currents in elements 1 and 2.

The phasing of driven elements depends on the direction of the element, the phase of the applied voltage, and the point at which the voltage is applied. In many systems used by amateurs, the voltages applied to the elements are exactly

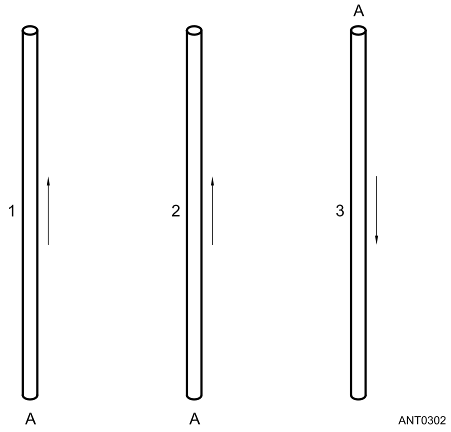

Fig 7—This drawing illustrates the phase of currents in antenna elements, represented by the arrows. The currents in elements 1 and 2 are in phase, while that in element 3 is 180° out of phase with 1 and 2.

in or exactly out-of-phase with each other. Also, the axes of the elements are nearly always in the same direction, since parallel or collinear elements are invariably used. The currents in driven elements in such systems therefore are usually either exactly in or exactly out-of-phase with the currents in other elements.

It is possible to use phase differences of less than 180° in driven arrays. One important case is where the current in one set of elements differs by 90° from the current in another set. However, making provision for proper phasing in such systems is considerably more complex than in the case of simple 0° or 180° phasing, as described in a later section of this chapter.

In parasitic arrays the phase of the currents in the parasitic elements depends on the spacing and tuning, as described later.

Ground Effects

The effect of the ground is the same with a directive antenna as it is with a simple dipole antenna. The reflection factors discussed in Chapter 3 may therefore be applied to the vertical pattern of an array, subject to the same modifications mentioned in that chapter. In cases where the array elements are not all at the same height, the reflection factor for the mean height of the array may be used for a close approximation. The mean height is the average of the heights measured from the ground to the centers of the lowest and highest elements.

MUTUAL IMPEDANCE

Consider two $\lambda/2$ elements that are fairly close to each other. Assume that power is applied to only one element, causing current to flow. This creates an electromagnetic field, which induces a voltage in the second element and causes current to flow in it as well. The current flowing in element 2 will in turn induce a voltage in element 1, causing additional current to flow there. The total current in 1 is then the sum (taking phase into account) of the original current and the induced current.

With element 2 present, the amplitude and phase of the resulting current in element 1 will be different than if element 2 were not there. This indicates that the presence of the second element has changed the impedance of the first. This effect is called *mutual coupling*. Mutual coupling results in a *mutual impedance* between the two elements. The mutual impedance has both resistive and reactive components. The actual impedance of an antenna element is the sum of its self-impedance (the impedance with no other antennas present) and its mutual impedances with all other antennas in the vicinity.

The magnitude and nature of the feed-point impedance of the first antenna depends on the amplitude of the current induced in it by the second, and on the phase relationship between the original and induced currents. The amplitude and phase of the induced current depend on the spacing between the antennas and whether or not the second antenna is tuned to resonance.

In the discussion of the several preceding paragraphs, power is applied to only one of the two elements. Do not interpret this to mean that mutual coupling exists only in parasitic arrays! It is important to remember that mutual coupling exists between any two conductors that are located near one another.

Amplitude of Induced Current

The induced current will be largest when the two antennas are close together and are parallel. Under these conditions the voltage induced in the second antenna by the first, and in the first by the second, has its greatest value and causes the largest current flow. The coupling decreases as the parallel antennas are moved farther apart.

The coupling between collinear antennas is comparatively small, and so the mutual impedance between such antennas is likewise small. It is not negligible, however.

Phase Relationships

When the separation between two antennas is an appreciable fraction of a wavelength, a measurable period of time elapses before the field from antenna 1 reaches antenna 2. There is a similar time lapse before the field set up by the current in number 2 gets back to induce a current in number 1. Hence the current induced in antenna 1 by antenna 2 will have a phase relationship with the original current in antenna 1 that depends on the spacing between the two antennas.

The induced current can range all the way from being completely in-phase with the original current to being completely out-of-phase with it. If the currents are in-phase, the total current is larger than the original current and the antenna feed-point impedance is reduced. If the currents are out-of-phase, the total current is smaller and the impedance is increased. At intermediate phase relationships the impedance will be lowered or raised depending on whether the induced current is mostly in or mostly out-of-phase with the original current.

Except in the special cases when the induced current is exactly in or out-of-phase with the original current, the induced current causes the phase of the total current to shift with respect to the applied voltage. Consequently, the presence of a second antenna nearby may cause the impedance of an antenna to be reactive—that is, the antenna will be detuned from resonance—even though its self-impedance is entirely resistive. The amount of detuning depends on the magnitude and phase of the induced current.

Tuning Conditions

A third factor that affects the impedance of antenna 1 when antenna 2 is present is the tuning of number 2. If antenna 2 is not exactly resonant, the current that flows in it as a result of the induced voltage will either lead or lag the phase it would have if the antenna were resonant. This causes an additional phase advance or delay that affects the phase of the current induced back in antenna 1. Such a phase lag has an effect similar to a change in the spacing between self-resonant antennas. However, a change in tuning is not exactly equivalent to a change in spacing because the two methods do not have the same effect on the amplitude of the induced current.

MUTUAL IMPEDANCE AND GAIN

The mutual coupling between antennas is important because it can have a significant effect on the amount of current that will flow for a given amount of power supplied. And it is the amount of *current* flowing that determines the field strength from the antenna. Other things being equal, if the mutual coupling between two antennas is such that the currents are greater for the same total power than would be the case if the two antennas were not coupled, the power gain will be greater than that shown in Table 1.

On the other hand, if the mutual coupling is such as to reduce the current, the gain will be less than if the antennas were not coupled. The term *mutual coupling*, as used in this paragraph, assumes that the mutual impedance between elements is taken into account, along with the added effects of propagation delay because of element spacing and element tuning or phasing.

The calculation of mutual impedance between antennas is a complex problem. Data for two simple but important cases are graphed in Figs 8 and 9. These graphs do not show the mutual impedance, but instead show a more useful quantity—the feed-point resistance measured at the center of an antenna as it is affected by the spacing between two antennas.

As shown by the solid curve in **Fig 8**, the feed-point resistance at the center of either antenna, when the two are self-resonant, parallel, and operated in-phase, decreases as the spacing between them is increased until the spacing is about 0.7 λ. This is a broadside array. The maximum gain is achieved from a pair of such elements when the spacing is in this region, because the current is larger for the same power and the fields from the two arrive in-phase at a distant point placed on a line perpendicular to the line joining the two antennas.

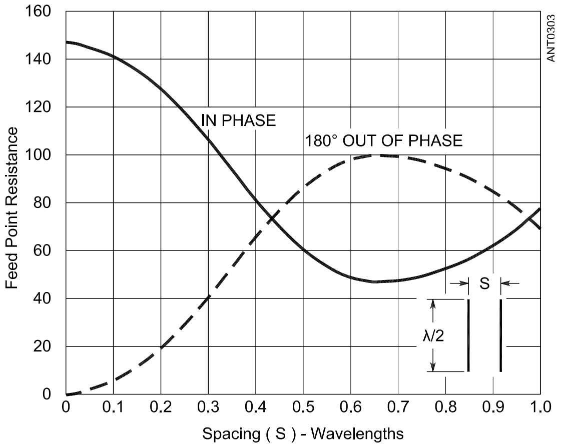

Fig 8—Feed-point resistance measured at the center of one element as a function of the spacing between two parallel ½-λ self-resonant antenna elements. For ground-mounted ¼-λ vertical elements, divide these resistances by two.

The dashed line in Fig 8, representing two antennas operated 180° out-of-phase (end-fire), cannot be interpreted quite so simply. The feed-point resistance decreases with spacing decreasing less than about 0.6 λ in this case. However, for the range of spacings considered, only when the spacing is 0.5 λ do the fields from the two antennas add up exactly in phase at a distant point in the favored direction. At smaller spacings the fields become increasingly out-of-phase, so the total field is less than the simple sum of the two. Smaller spacings thus decrease the gain at the same time that

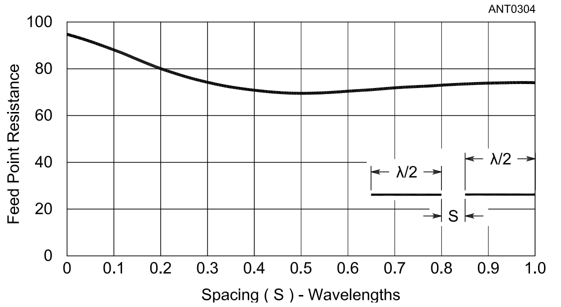

Fig 9—Feed-point resistance measured at the center of one element as a function of the spacing between the ends of two collinear self-resonant ½-λ antenna elements operated in phase.

the reduction in feed-point resistance is increasing it. For a lossless antenna, the gain goes through a maximum when the spacing is in the region of ⅛ λ.

The feed-point resistance curve for two collinear elements in-phase, **Fig 9**, shows that the feed-point resistance decreases and goes through a broad minimum in the region of 0.4 to 0.6-λ spacing between the adjacent ends of the antennas. As the minimum is not significantly less than the feed-point resistance of an isolated antenna, the gain will not exceed the gain calculated on the basis of uncoupled antennas. That is, the best that two collinear elements will give, even with optimum spacing, is a power gain of about 2 (3 dB). When the separation between the ends is very small—the usual method of operation—the gain is reduced.

GAIN AND ARRAY DIMENSIONS

The gain of an array is principally determined by the dimensions of the array, so long as there are a minimum number of elements. A good example of this is the relationship between boom length, gain and number of elements for an array such as a Yagi. **Fig 10** compares the gain versus boom length for Yagis with different numbers of elements. For given number of elements, notice that the gain increases as the boom length increases, up to a maximum. Beyond this point, longer boom lengths result in less gain for a given number of elements. This observation does not mean that it is always desirable to use only the minimum number of elements. Other considerations of array performance, such as front-to-back ratio, minor lobe amplitudes or operating bandwidth, may make it advantageous to use more than the minimum number of elements for a given array length. A specific example of this is presented in a later section in a comparison between a half-square, a bobtail curtain and a Bruce array.

In a broadside array the gain is a function of both the length and width of the array. The gain can be increased by adding more elements (with additional spacing) or by using longer elements (>λ/2), although the use of longer elements requires proper attention to current phase in the elements. In general, in a broadside array the element spacing that gives maximum gain for a minimum number of elements, is in the range of 0.5 to 0.7 λ. Broadside arrays with elements spaced for maximum gain will frequently have significant side lobes and associated narrowing of the main lobe beamwidth. Side lobes can be reduced by using more than the minimum number of elements, spaced closer than the maximum gain distance.

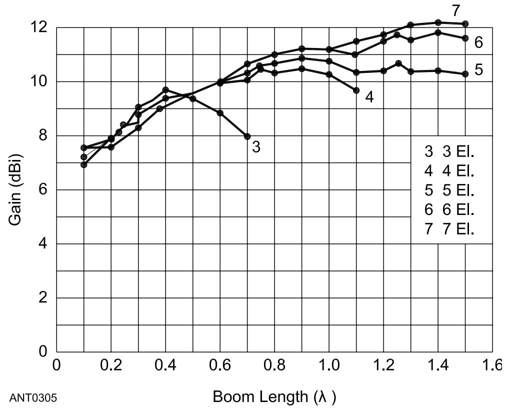

Fig 10—Yagi gain for 3, 4, 5, 6 and 7-element beams as a function of boom length. (From *Yagi Antenna Design*, J. Lawson, W2PV.)

Additional gain can be obtained by expanding the array into a third dimension. An example of this is the stacking of endfire arrays in a broadside configuration. In the case of stacked short endfire arrays, maximum gain occurs with spacings in the region of 0.5 to 0.7 λ. However, for longer higher-gain end-fire arrays, larger spacing is required to achieve maximum gain. This is important in VHF and UHF arrays, which often use long-boom Yagis.

PARASITIC ARRAYS

The foregoing applies to multi-element arrays of both types, driven and parasitic. However, there are special considerations for driven arrays that do not necessarily apply to parasitic arrays, and vice versa. Such considerations for Yagi and quad parasitic arrays are presented in Chapters 11 and 12. The remainder of this chapter is devoted to driven arrays.

Driven Arrays

Driven arrays in general are either broadside or end-fire, and may consist of collinear elements, parallel elements, or a combination of both. From a practical standpoint, the maximum number of usable elements depends on the frequency and the space available for the antenna. Fairly elaborate arrays, using as many as 16 or even 32 elements, can be installed in a rather small space when the operating frequency is in the VHF range, and more at UHF. At lower frequencies the construction of antennas with a large number of elements is impractical for most amateurs.

Of course the simplest of driven arrays is one with just two elements. If the elements are collinear, they are always fed in-phase. The effects of mutual coupling are not great, as illustrated in Fig 9. Therefore, feeding power to each element in the presence of the other presents no significant problems. This may not be the case when the elements are parallel to each other. However, because the combination of spacing and phasing arrangements for parallel elements is infinite, the number of possible radiation patterns is endless.

This is illustrated in **Fig 11**. When the elements are fed in-phase, a broadside pattern always results. At spacings of less than ⅝ λ with the elements fed 180° out-of-phase, an end-fire pattern always results. With intermediate amounts of phase difference, the results cannot be so simply stated. Patterns evolve that are not symmetrical in all four quadrants.

Because of the effects of mutual coupling between the two driven elements, for a given power input greater or lesser currents will flow in each element with changes in spacing and phasing, as described earlier. This, in turn, affects the gain of the array in a way that cannot be shown merely by plotting the *shapes* of the patterns, as has been done in Fig 11. Therefore, supplemental gain information is also shown in Fig 11, adjacent to the pattern plot for each combination of spacing and phasing. The gain figures shown are referenced to a single element. For example, a pair of elements fed 90° apart at a spacing of λ/4 will have a gain in the direction of maximum radiation of 3.1 dB over a single element.

Current Distribution in Phased Arrays

In the plots of Fig 11, the two elements are assumed to be identical and self-resonant. In addition, currents of equal amplitude are assumed to be flowing at the feed point of each element, a condition that most often will not exist in practice without devoting special consideration to the feeder system. Such considerations are discussed in the next section of this chapter.

Most literature for radio amateurs concerning phased arrays is based on the assumption that if all elements in the array are identical, the *current distribution* in all the elements will be identical. This distribution is presumed to be that of a single, isolated element, or nearly sinusoidal. However, information published in the professional literature as early as the 1940s indicates the existence of dissimilar current distributions among the elements of phased arrays. (See Harrison and King references in the Bibliography.) Lewallen, in July 1990 *QST*, pointed out the causes and effects of dissimilar current distributions.

In essence, even though the two elements in a phased array may be identical and have exactly equal currents of the desired phase flowing *at the feed point*, the amplitude and phase relationships degenerate with departure from the feed point. This happens any time the phase relationship is not 0° or 180°. Thus, the field strengths produced at a distant point by the individual elements may differ. This is because the field from each element is determined by the *distribution* of the current, as well as its magnitude and phase.

The effects are minimal with shortened elements—verticals less than λ/4 or dipoles less than λ/2 long. The effects on radiation patterns begin to show at the above resonant lengths, and become profound with longer elements—λ/2 or longer verticals and 1 λ or longer center-fed elements. These effects are less pronounced with thin elements. The amplitude and phase degeneration takes place because the currents in the array elements are not sinusoidal. Even in two-element arrays with phasing of 0° or 180°, the currents are not sinusoidal, but in these two special cases they do remain identical.

The pattern plots of Fig 11 take element current distributions into account. The visible results of dissimilar distributions are incomplete nulls in some patterns and the development of very small minor lobes in others. For example, the pattern for a phased array with 90° spacing and 90° phasing has traditionally been published in amateur literature as a cardioid with a perfect null in the rear direction. Fig 11, calculated for 7.15-MHz self-resonant dipoles of #12 wire in free space, shows a minor lobe at the rear and only a 33-dB front-to-back ratio.

It is characteristic of broadside arrays that the power gain is proportional to the length of the array but is substantially independent of the number of elements used, provided the optimum element spacing is not exceeded. This means, for example, that a five-element array and a six-element array will have the same gain, provided the elements in both are spaced so the overall array length is the same. Although this principle is seldom used for the purpose of reducing the number of elements because of complications introduced in feeding power to each element in the proper phase, it does illustrate the fact that there is nothing to be gained, in terms of more gain, by increasing the number of elements if the space occupied by the antenna is not increased proportionally.

Generally speaking, the maximum gain in the smallest linear dimensions will result when the antenna combines both broadside and end-fire directivity and uses both parallel and collinear elements. In this way the antenna is spread over a greater volume of space, which has the same effect as extending its length to a much greater extent in one linear direction.

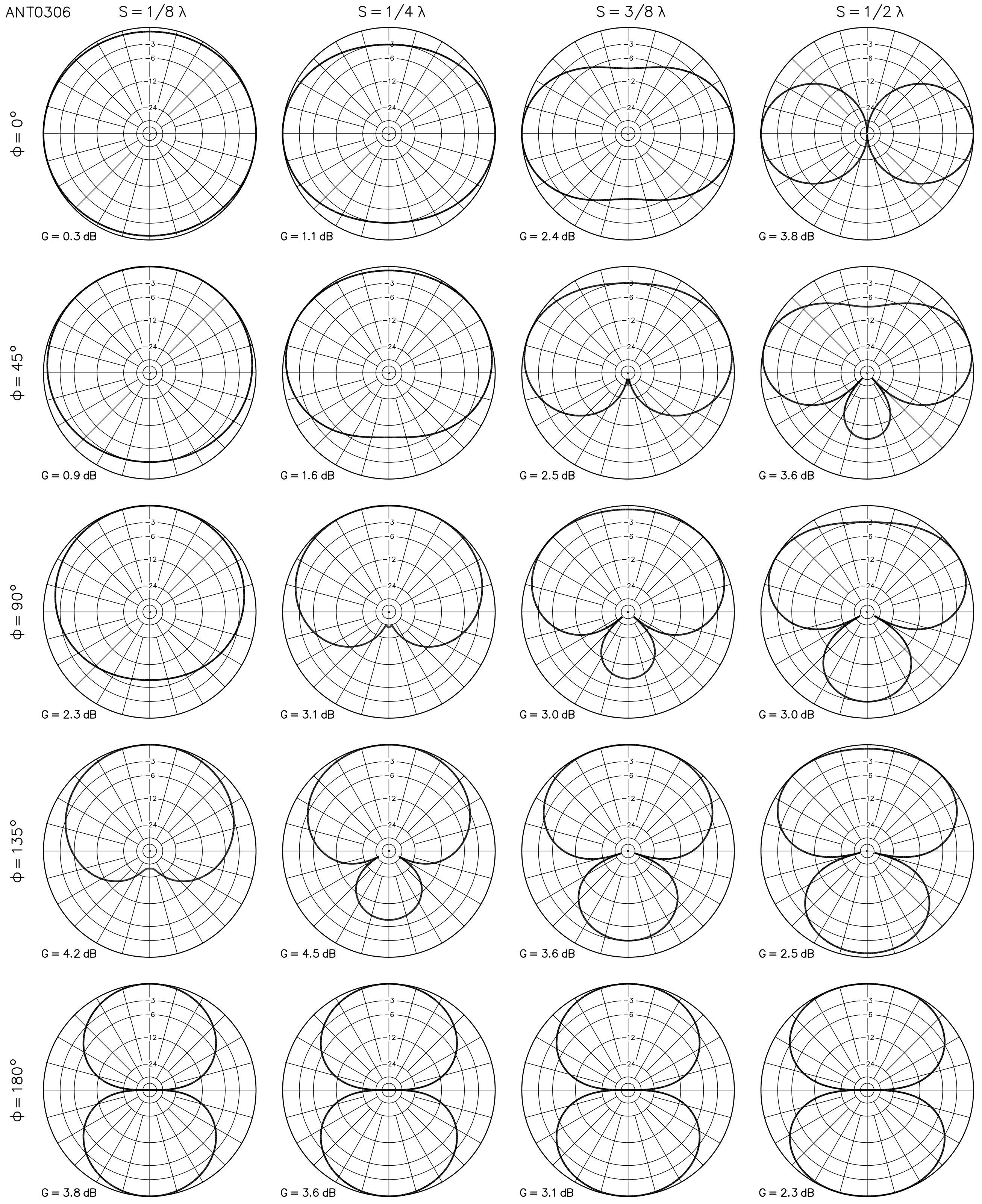

Fig 11—H-plane patterns of two identical parallel driven elements, spaced and phased as indicated (S = spacing, φ = phasing). The elements are aligned with the vertical (0°-180°) axis, and the element nearer the 0° direction (top of page) is of lagging phase at angles other than 0°. The two elements are assumed to be thin and self-resonant, with equal-amplitude currents flowing at the feed point. See text regarding current distributions. The gain figure associated with

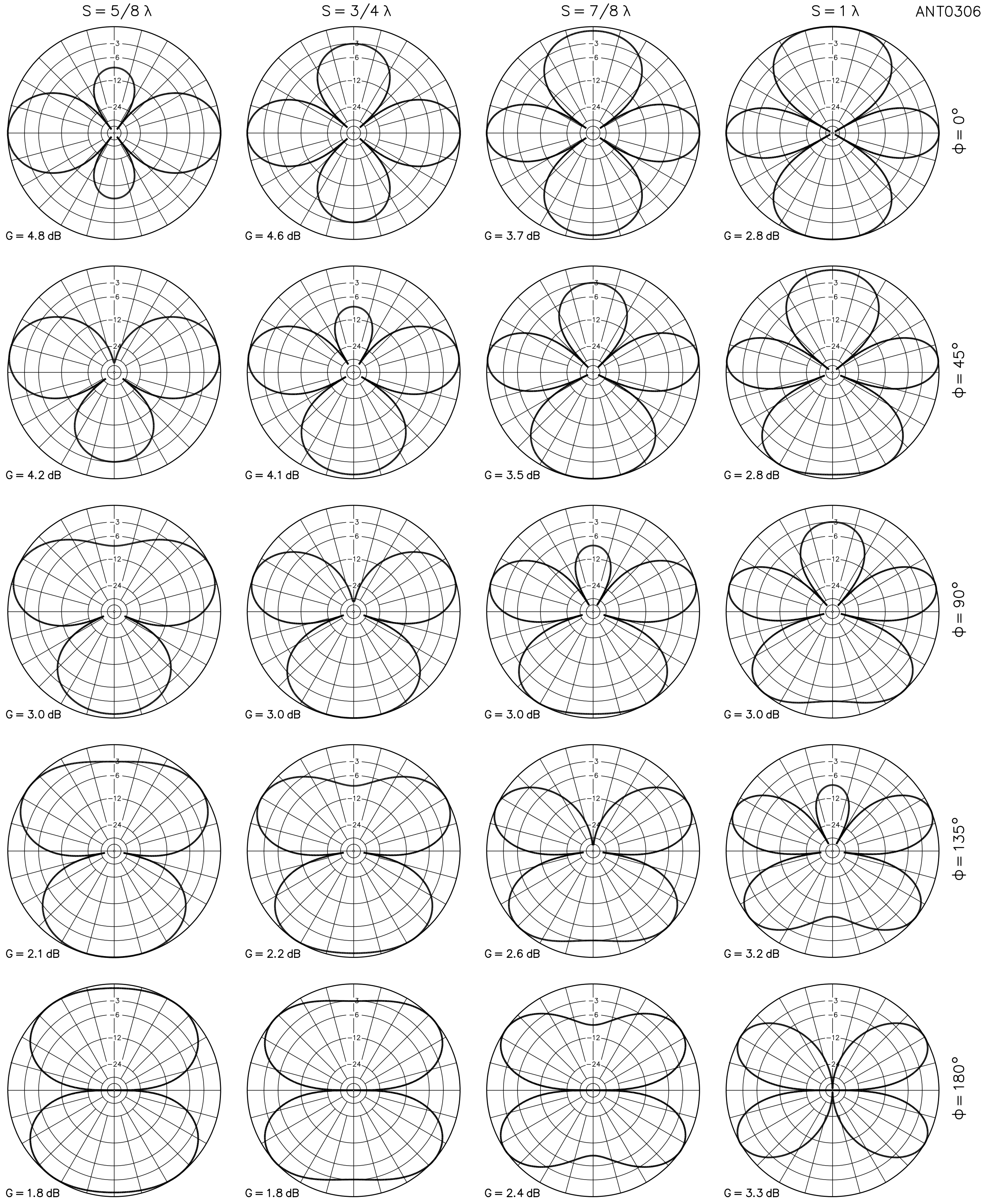

each pattern indicates that of the array over a single element. The plots represent the horizontal or azimuth pattern at a 0° elevation angle of two ¼-λ vertical elements over a perfect conductor, or the free-space vertical or elevation pattern of two horizontal ½-λ elements when viewed on end, with one element above the other. (Patterns computed with *ELNEC*—see Bibliography.)

Phased Array Techniques

Phased antenna arrays have become increasingly popular for amateur use, particularly on the lower frequency bands, where they provide one of the few practical methods to obtain substantial gain and directivity. This section on phased-array techniques was written by Roy Lewallen, W7EL.

The operation and limitations of phased arrays, how to design feed systems to make them work properly and how to make necessary tests and adjustments are discussed in the pages that follow. The examples deal primarily with vertical HF arrays, but the principles apply to VHF/UHF arrays and arrays made from other element types as well.

OVERVIEW

Much of this chapter is devoted to techniques for feeding phased arrays. Many people who have a limited acquaintance with phased array techniques believe this is a simple problem, consisting only of connecting array elements through "phasing lines" consisting of transmission lines of the desired electrical lengths. Unfortunately, except for a very few special cases, this approach won't achieve the desired array pattern.

Other proposed universal solutions, such as hybrid couplers or Wilkinson or other power dividers, also usually fail to achieve the necessary phasing. These approaches sometimes produce—often more by accident than design—results good enough to mislead the user into believing that the simple approach is working as planned. Confusion can result when an approach fails to work in different circumstances. This section will explain why the simple solutions don't work as often thought, and how to design feed systems that do consistently produce the desired results.

Very briefly, the reason why the simple phasing-line approach fails is that the delay of current or voltage in a transmission line equals the line's electrical length only if the line is terminated in its characteristic impedance. And in phased arrays, element feed-point impedances are profoundly affected by mutual coupling.

Consequently, even if each element has the correct impedance when isolated, it won't when all elements are excited. Furthermore, transmission lines that are not terminated in their characteristic impedance will transform both the voltage and current magnitude. The net result is that the array elements will have neither the correct magnitudes nor phases of current necessary for proper operation except in a few special cases. This isn't a minor effect of concern only to perfectionists, but often a major one that causes significant pattern distortion and poor or mislocated nulls. The problem is examined in greater depth later.

Power dividers and hybrid couplers also fail to achieve the desired result for different reasons, which will be discussed below, although in one common application hybrid couplers fortuitously provide results that are acceptable to many users. This chapter will show how to design array feed systems that *will* produce predicted element currents and array patterns.

Various *EZNEC* models are provided to illustrate concepts presented in this chapter. They can all be viewed with the *EZNEC-ARRL* software furnished on the CD included with this book. Step-by-step instructions for the examples are given in **Appendix A**.

FUNDAMENTALS OF PHASED ARRAYS

The performance of a phased array is determined by several factors. Most significant among these are the characteristics of a single element, reinforcement or cancellation of the fields from the elements and the effects of mutual coupling. To understand the operation of phased arrays, it is first necessary to understand the operation of a single antenna element.

Of primary importance is the strength of the field produced by the element. The field radiated from a linear (straight) element, such as a dipole or vertical monopole, is proportional to the sum of the elementary currents flowing in each part of the antenna element. For this discussion it is important to understand what determines the current in a single element.

The amount of current flowing at the base of a ground mounted vertical or ground-plane antenna is given by the familiar formula

$$I = \sqrt{\frac{P}{R}} \qquad \text{(Eq 2)}$$

where

P is the power supplied to the antenna
R is the feed-point resistance.

R consists of two parts, the loss resistance and the radiation resistance. The loss resistance, R_L, includes losses in the conductor, in the matching and loading components and dominantly (in the case of ground-mounted verticals) in ground losses. The power "dissipated" in the radiation resistance, R_R, is actually the power that is radiated, so maximizing the power dissipated by the radiation resistance is desirable. However, the power dissipated in the loss resistance truly is lost as heat, so resistive losses should be made as small as possible.

The radiation resistance of an element can be derived from electromagnetic field theory, being a function of antenna length, diameter and geometry. Graphs of radiation resistance versus antenna length are given in Chapter 2. The radiation resistance of a thin resonant $\lambda/4$ ground-mounted vertical is about 36 Ω. A resonant $\lambda/2$ dipole in free space has a radiation resistance of twice this amount, about 73 Ω. Reducing the antenna lengths by one half drops the radiation resistances to approximately 7 and 14 Ω, respectively.

The radiation resistance of a large variety of antennas can easily be determined by using *EZNEC-ARRL*, which is included on the CD in the back of this book. The radiation

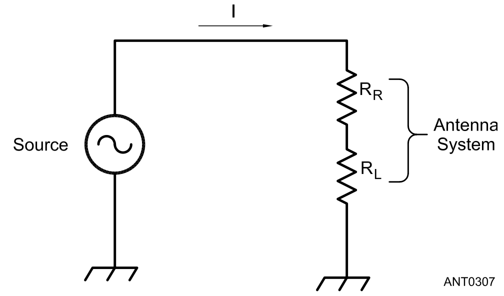

Fig 12—Simplified equivalent circuit for a single-element resonant antenna. R_R represents the radiation resistance, and R_L the ohmic losses in the total antenna system.

resistance is simply the feed-point resistance (the resistive part of the feed-point impedance) when all losses have been set to zero.

Radiation Efficiency

To generate a stronger field from a given radiator, it is necessary to increase the power P (the brute-force solution), decrease the loss resistance R_L (by putting in a more elaborate ground system for a vertical, for instance) or to somehow decrease the radiation resistance R_R so more current will flow with a given power input. This can be seen by expanding the formula for base current as

$$I = \sqrt{\frac{P}{R_R + R_L}} \qquad \text{(Eq 3)}$$

Splitting the feed-point resistance into components R_R and R_L easily leads to an understanding of element efficiency. The efficiency of an element is the proportion of the total power that is actually radiated. The roles of R_R and R_L in determining efficiency can be seen by analyzing a simple equivalent circuit, shown in **Fig 12**.

The power dissipated in R_R (the radiated power) equals I^2R_R. The total power supplied to the antenna system is

$$P = I^2 (R_R + R_L) \qquad \text{(Eq 4)}$$

so the efficiency (the fraction of supplied power that is actually radiated) is

$$\text{Eff} = \frac{I^2 R_R}{I^2 (R_R + R_L)} = \frac{R_R}{R_R + R_L} \qquad \text{(Eq 5)}$$

Efficiency is frequently expressed in percent, but expressing it in decibels relative to a 100%-efficient radiator gives a better idea of what to expect in the way of signal strength. The field strength of an element relative to a lossless but otherwise identical element, in dB, is

$$\text{FSG} = 10 \log \frac{R_R}{R_R + R_L} \qquad \text{(Eq 6)}$$

where FSG = field strength gain, dB.

For example, information presented by Sevick in March 1973 *QST* shows that a λ/4 ground-mounted vertical antenna with four 0.2-λ radials has a feed-point resistance of about 65 Ω (see the Bibliography at the end of this chapter). The efficiency of such a system is 36/65 = 55.4%. It is rather disheartening to think that, of 100 W fed to the antenna, only 55 W is being radiated, with the remainder literally warming up the ground. Yet the signal will be only 10 log (36/65) = –2.57 dB relative to the same vertical with a perfect ground system. In view of this information, trading a small reduction in signal strength for lower cost and greater simplicity may become an attractive consideration.

So far, only the current at the base of a resonant antenna has been discussed, but the field is proportional to the sum of currents in each tiny part of the antenna. The field is a function of not only the magnitude of current flowing at the feed point, but also the distribution of current along the radiator and the length of the radiator. Nothing can be done at the feed point to change the current distribution, so for a given element the field strength is proportional to the feed point current. However, changing the radiator length or loading it at some point other than the feed point will change the current distribution.

More information on shortened or loaded radiators may be found in Chapters 2 and 6 and in the Bibliography references of this chapter. The current distribution is also changed by mutual coupling to other array elements, although for most arrays this has only a minor effect on the pattern. This is discussed later in more detail. A few other important facts follow.

1) If there is no loss, the field from even an infinitesimally short radiator is less than ½ dB weaker than the field from a half-wave dipole or quarter-wave vertical. Without loss, all the supplied power is radiated regardless of the antenna length, so the only factor influencing gain is the slight difference in the patterns of very short and λ/2 antennas. The small pattern difference arises from different current distributions. A short antenna has a very low radiation resistance, resulting in a heavy current flow over its short length. In the absence of loss, this generates a field strength comparable to that of a longer antenna. Where loss is present—that is, in practical antennas—shorter radiators usually don't do so well, since the low radiation resistance leads to lower efficiency for a given loss resistance. If care is taken, reasonably short antennas can achieve good efficiency.

2) Care has to be taken in calculating the efficiency of folded antennas. (See Chapter 6.) Folding transforms both the radiation resistance and loss resistance by the same factor, so their ratio and therefore the efficiency remains the same. It's easy to show that in a ground-mounted vertical array, folding reduces the current flowing from the feed line to the ground system by a factor of two due to the impedance transformation. However, the folded antenna has an additional connection to ground, which also carries half the original ground current. The result is that the same amount of current flows

into the ground system, whether unfolded or folded, resulting in the same ground system loss. Analyses purporting to show otherwise invariably transform the radiation resistance but neglect to also transform the loss resistance and reach an incorrect conclusion.

3) The current flowing in an element with a given power input can be increased or decreased by mutual coupling to other elements. The effect is equivalent to changing the element radiation resistance. Mutual coupling is sometimes thought of as a minor effect, but often it is not minor!

Field Reinforcement and Cancellation

The mechanism by which phased arrays produce gain, and the role of mutual coupling in determining gain, were covered earlier in this chapter. One important point that can't be emphasized enough is that all antennas must abide by the law of *conservation of energy*. No antenna can radiate more power than supplied to it. The total amount of power it radiates is the amount it's supplied, less the amount lost as heat. This is true of all antennas, from the smallest "rubber ducky" to the most gigantic array.

Gain

Gain is strictly a relative measure, so the term is completely meaningless unless accompanied by a statement of just what it is relative to. One useful measure for phased array gain is *gain relative to a single similar element*. This is the increase in signal strength that would be obtained by replacing a single element by an array made from elements just like it. In some instances, such as investigating what happens to array performance when *all* elements become more lossy, it's useful to state gain relative to a more absolute, although unattainable standard: a lossless element.

And the most universal reference for gain is another unattainable standard, the *isotropic radiator*. This fictional antenna radiates absolutely equally in all directions. It's very useful because the field strength resulting from any power input is readily calculated, so if the gain relative to this standard is known, the field strength is also known for any radiated power. Gain relative to this reference is referred to as dBi, and it's the standard used by most modeling programs including *EZNEC-ARRL*. To find the gain of an array relative to a single element or other reference antenna such as a dipole, model both the array and the single element or other reference antenna in the same environment and subtract their dBi gains. Don't rely on some assumption about the gain of a single element—many people assume values that can be very wrong.

Nulls

Pattern nulls are very often more important to users of phased arrays than gain because of their importance in reducing both man-made and natural interference when receiving. Consequently, a good deal of emphasis is, and should be, placed on achieving good pattern nulls. Unfortunately, good nulls are much more difficult to achieve than gain and they are much more sensitive to array and feed-system imperfections.

As an illustration, consider two elements that each produce a field strength of, say, exactly 1 millivolt per meter (mV/m) at some distance many wavelengths from the array. In the direction in which the fields from the elements are in-phase, a total field of 2 mV/m results. In the direction in which they're out-of-phase, zero field results. The ratio of maximum to minimum field strength of this array is 2/0, or infinity.

Now suppose, instead, that one field is 10% high and the other 10% low—1.1 and 0.9 mV/m, respectively. In the forward direction, the field strength is still 2 mV/m, but in the canceling direction, the field will be 0.2 mV/m. The front-to-back ratio has dropped from infinity to 2/0.2, or 20 dB. (Actually, slightly more power is required to redistribute the field strengths this way, so the forward gain is reduced—but by only a small amount, less than 0.1 dB.) For most arrays, unequal fields from the elements have a minor effect on forward gain, but a major effect on pattern nulls. This is illustrated by ***EZNEC* Example: Nulls** in Appendix A.

Even with perfect current balance, deep nulls aren't assured. **Fig 13** shows the minimum spacing required for total field reinforcement or cancellation. If the element spacing isn't adequate, there may be no direction in which the fields are completely out-of-phase (see curve B of Fig 13). Slight physical and environmental differences between elements will invariably affect null depths, and null depths will also vary with elevation angle.

However, a properly designed and fed array can produce very impressive nulls. The key to producing good nulls, like producing gain, is controlling the strengths and phases of the fields from the elements. Just how to accomplish that is the subject of most of the remainder of this section. But be sure to keep in mind that producing good nulls is generally a much more difficult task than producing approximately the predicted gain.

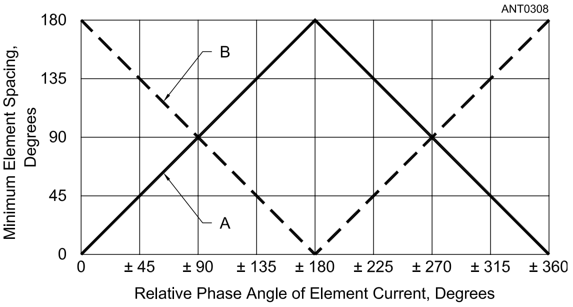

Fig 13—Minimum element spacing required for total field reinforcement, curve A, or total field cancellation, curve B. Total cancellation results in pattern nulls in one or more directions. Total reinforcement does not necessarily mean there is gain over a single element, as the effects of loss and mutual coupling must also be considered.

Mutual Coupling

Mutual coupling was discussed briefly earlier in this chapter. Because it has an important and profound effect on both the performance and feed system design of phased arrays, it will be covered in greater depth here.

Mutual coupling refers to the effects which the elements in an array have on each other. Mutual coupling can occur intentionally or entirely unintentionally. People with multiple antennas on a small lot (or car top) often discover that a better description of their system is a single antenna with multiple feed points. Current is induced in conductors in various antennas by mutual coupling, causing them to act like parasitic elements, which re-radiate and distort the antenna's pattern. The effects of mutual coupling are present whether or not the elements are driven.

Suppose that two driven elements are many wavelengths from each other. Each has some voltage and current at its feed point. For each element, the ratio of this voltage to current is the element self-impedance. If the elements are brought close to each other, the current in each element will change in magnitude and phase because of coupling with the field from the other element. The field from the first element changes the current in the second. This changes the field from the second, which alters the current in the first, and so forth until an equilibrium condition is reached in which the currents in all elements (hence, their fields) are totally interdependent.

The feed-point impedances of all elements also are changed from their values when far apart, and all are dependent on each other. In a driven array, the changes in feed-point impedances can cause additional changes in element currents, because the operation of many feed systems depends on the element feed-point impedances. Significant mutual coupling occurs at spacings as great as a wavelength or more.

Connecting the elements to a feed system to form a driven array does not eliminate the effects of mutual coupling. In fact, in many driven arrays the mutual coupling has a greater effect on antenna operation than the feed system does. All feed-system designs must account for the impedance changes caused by mutual coupling if the desired current balance and phasing are to be achieved.

Several general statements can be made regarding the effects of mutual coupling on phased-array systems.

1) The resistances and reactances of all elements of an array generally will be substantially different from the values when the elements are isolated (that is, very distant from other elements).

2) If the elements of a two-element array are identical and have equal currents that are in-phase or 180° out-of-phase, the feed-point impedances of the two elements will be equal. But they will be different than for an isolated element. If the two elements are part of a larger array, their impedances can be very different from each other.

3) If the elements of a two-element array have currents that are neither in-phase (0°) nor out-of-phase (180°), their feed-point impedances will not be equal. The difference will be considerable in typical amateur arrays.

4) The feed-point resistances of the elements in a closely spaced, 180° out-of-phase array will be very low, resulting in poor efficiency due to ohmic losses unless care is taken to minimize loss. This is also true for any other closely spaced array with significant predicted gain.

It's essential to realize that this is *not* a minor effect and one that can be overlooked or ignored. See ***EZNEC* Example—Mutual Coupling** in Appendix A for an illustration of these phenomena.

Loss Resistance, Mutual Coupling and Antenna Gain

Loss reduces the effects of mutual coupling because the feed-point impedance change resulting from mutual coupling is effectively in series with loss resistance. If the loss is great enough, two important results occur. First, the feed-point impedance becomes independent of the presence of nearby current-carrying elements. This greatly simplifies feed system design—the simple "phasing-line" or hybrid-coupler feed system described below is adequate *provided that all elements are physically identical and the feed point of each element is matched to the* Z_0 *of the feed line and, if used, the hybrid coupler.*

The impedance matching restrictions are necessary to insure that the phasing line or hybrid coupler performs as expected. Identical elements are needed so that equal element currents will result in equal fields from the elements.

In the absence of mutual coupling effects, the maximum gain of an array of identical elements relative to a single (similarly lossy) element is simply 10 log(N), where N is the number of elements—providing that spacing is adequate for the fields to fully reinforce in some direction. If spacing is less, maximum gain will also be less. Of course, the array gain relative to a single *lossless* element will be very low, most likely a sizeable negative number when expressed in dB. So intentionally introducing loss isn't a wise idea for a transmitting array. It is sometimes an advantageous thing to do for a receiving array, however, as explained in the following section.

High-gain close-spaced arrays, such as the W8JK phased array (see *EZNEC-ARRL* example file **ARRL_W8JK.EZ** and accompanying Antenna Notes file), and most parasitic arrays depend heavily on mutual coupling to achieve their gain. Introduction of any loss to these arrays, which reduces the mutual coupling effects, has a profound effect on the gain. Consequently, parasitic or close-spaced driven arrays often produce disappointing results when made from grounded vertical elements unless each has a fairly elaborate (and therefore very low-loss) ground system.

If you place two low-loss elements very close together and feed them in-phase, mutual coupling reduces the array gain to essentially that of a single element, so there's no advantage to this configuration over a single element. However, if you have a single lossy element, for example a short vertical having a relatively poor ground system, you can improve the gain by up to 3 dB by adding a second, close spaced, element and ground system and feeding the two in-phase. Another way to look at this technique is that you're putting two equal ground system resistances in parallel, which effectively cuts the loss in half. The gain you can realize in practice depends on such things as the ground system overlap, but it might be a practical way to improve transmitting array performance in some situations.

FEEDING PHASED ARRAYS

The previous section explains why the fields from the elements must be very close to the ratios required by the array design. Since the field strengths are proportional to the currents in the elements, controlling the fields requires controlling the element currents. Since the desired current ratio is 1:1 for virtually all two-element and for most (but not all) larger amateur arrays, special attention is paid to methods of assuring equal element currents. But we will examine other current ratios also.

The Role of Element Currents

The field from a conductor is proportional to the current flowing on it. So if we're to control the relative strengths and phases of the fields from the elements, we have to control their currents. We usually do this by controlling the currents at the element feed points. But because the field from an element depends on the current everywhere along the element, elements having identical feed-point currents will produce different fields if they have different current distributions—that is, if the way the current varies along the lengths of the elements is different.

A previous section explained that mutual coupling alters the current distribution, so in many arrays the current distributions will be different on the elements and consequently the relationship between the overall fields won't be the same as that between the feed-point currents. Fortunately, this effect is relatively minor in thin, λ/4 monopole or λ/2 dipole elements. The most common arrays are made from elements in this category, so we can generally get very nearly the desired ratio of fields by effecting the same ratio of feed point currents. Exceptions are detailed immediately below.

Feed-point vs Element Current

For most antennas, environmental factors are likely to cause greater performance anomalies than current distribution differences, and both can be corrected with minor feed system adjustments. The difference between field and feed-point current ratios can become very significant, however, if the elements are very fat and/or close to λ/2 (monopole) or 1 λ (dipole) long. In those cases, most of the feed systems described here won't produce the desired field ratios without major adjustment or modification, except in the special cases of 2-element arrays with identical elements having feed point currents in-phase or 180° out-of-phase. In those special cases, the element current distributions are the same for the same reason the feed-point impedances are equal. This is explained later in the feed system sections.

To get an idea of just how large an element must be to disturb the pattern of an array with correct feed-point currents, a two-element cardioid array of quarter wave vertical elements was modeled at 10 MHz. With thin, 0.1-inch diameter elements, the front/back ratio was 35 dB, the very small reverse lobe caused by slightly unequal element current distributions. Increasing the element diameter to 20 inches decreased the front/back ratio to 20 dB. Returning the front/back ratio of the array of 20-inch elements to >35 dB required changing the feed-point current ratio from the nominal value of 1.0 at an angle of 90° to 0.88 at 83°.

The same array was first modeled with 0.1 inch diameter elements, where it has a front/back ratio of 35 dB, then the elements were lengthened. The front/back ratio dropped to 20 dB at an element length of 36 feet, or about 0.37 λ. In that case, adjustment of the feed-point current ratio to about 0.9 at about 83° restored a good front/back ratio.

In the discussion and development which follow, the assumption is made that the fields will be very nearly proportional to the feed-point currents. If the elements are fat or long enough to make this assumption untrue, some adjustment of feed-point current ratio will be necessary to achieve the desired pattern, particularly nulls. Most feed systems can be designed for any current ratio. Modeling will reveal the ratio required for the desired pattern, and then the feed system can be designed accordingly.

COMMON PHASED-ARRAY FEED SYSTEMS

This section will first describe several popular approaches to feeding phased arrays that often don't produce the desired results. It will describe why they don't work as well as hoped. It also briefly discusses systems that could be used, but that often aren't appropriate or optimum for amateur arrays.

This will be followed in the next section by detailed descriptions of array feed systems that *do* produce the predicted element current ratios and array patterns.

The "Phasing-Line" Approach

For an array to produce the desired pattern, the element currents must have the required magnitude and the required phase relationship. As explained above, this can generally be achieved well enough by causing the feed point currents to have that same relationship.

On the surface, this sounds easy—just make sure that the difference in electrical lengths of the feed lines to the elements equals the desired phase angle. Unfortunately, this approach doesn't necessarily achieve the desired result. The first problem is that the phase shift through the line is not equal to its electrical length. The current (or, for that matter, voltage) delay in a transmission line is equal to its electrical length in only a few special cases—cases which don't exist in most amateur arrays! The impedance of an element in an array is frequently very different from the impedance of an isolated element and the impedances of all the elements in an array can be different from each other.

See the ***EZNEC* Example—Mutual Coupling** in Appendix A for a graphic illustration of the effect of mutual coupling on feed-point impedance. Also look at the Four-Square array example in the **Phased Array Design Examples** section. The array in that example has one element with a *negative* feed-point resistance, if ground loss is low. Without mutual coupling, the resistance of that same element would be about 36 Ω plus ground loss.

Because of mutual coupling, the elements seldom provide a matched load for the element feed lines. The effect of mismatch on phase shift can be seen in **Fig 14**. Observe what happens to the phase of the current and voltage on a line terminated by a purely resistive impedance that is lower than the characteristic impedance of the line (Fig 14A). At a point 45° from the load the current has advanced less than 45°, and the voltage more than 45°. At 90° from the load both are advanced 90°. At 135° the current has advanced more and the voltage less than 135°. This apparent slowing down and speeding up of the current and voltage waves is caused by interference between the forward and reflected waves. It occurs on any line that is not terminated with a pure resistance equal to its characteristic impedance. If the load resistance is greater than the characteristic impedance of the line, as shown in Fig 14B, the voltage and current exchange angles. Adding reactance to the load causes additional phase shift. The *only cases in which the current (or voltage) delay is equal to the electrical length of the line are*

1) When the line is flat; that is, terminated in a purely resistive load equal to its characteristic impedance;

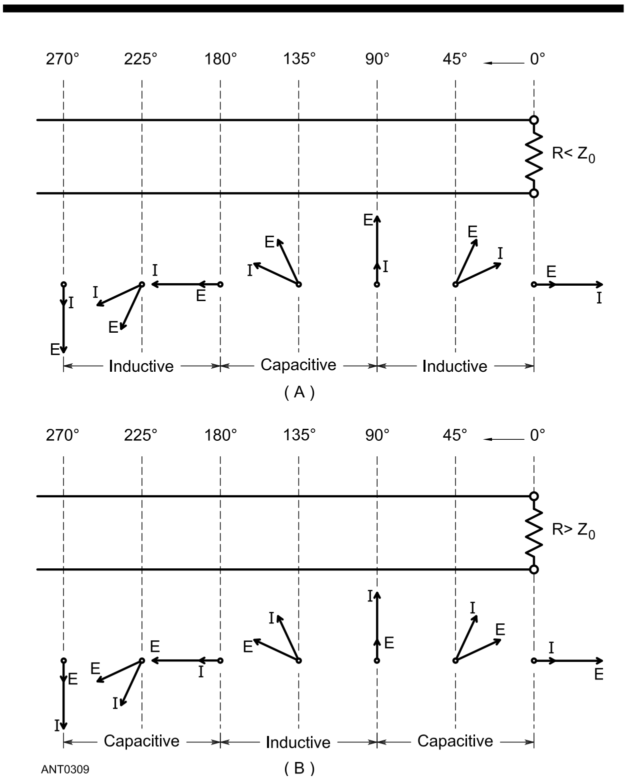

Fig 14—Resultant voltages and currents along a mismatched line. At A, R less than Z_0, and at B, R greater than Z_0.

2) When the line length is an integral number of half wavelengths;

3) When the line length is an odd number of quarter wavelengths and the load is purely resistive; and

4) When other specific lengths are used for specific load impedances.

Just how much phase error can be expected if two feed lines are simply hooked up to form an array? There is no simple answer. Some casually designed feed systems might deliver satisfactory results, but most will not. See the ***EZNEC* Example—"Phasing-Line" Feed** in Appendix A for the typical consequences of using this sort of feed system.

A second problem with simply connecting feed lines of different lengths to the elements is that the lines will change the *magnitudes* of the currents. The magnitude of the current (or voltage) out of a line does not equal the magnitude into that line, except in cases 1, 2 and 4 above. The feed systems presented later in this chapter assure currents that are correct in both magnitude and phase.

The elementary phasing-line approach *will* work in three very special but common situations. If the array consists of only two identical elements and those elements are fed in-phase, mutual coupling will modify the element impedances, but both will be modified exactly the same amount. Consequently, if the two elements are fed through equal-length transmission lines, the lines will transform and delay the currents by the same amount and result in equal, in-phase currents at the element feed points.

Similarly, an array of two identical elements fed 180º out-of-phase will have the same feed-point impedances and can be fed with two lines of any length so long as one line is an electrical half wavelength longer than the other. But this can't be extended to any two elements in a larger array, since mutual coupling to the other elements can result in different feed-point impedances. Methods will be described later which do assure a correct current ratio in this situation.

The third application in which the phasing-line approach works is in receiving arrays where the elements are very short in terms of wavelength and/or very lossy. In either of these cases, mutual coupling between elements is much less than an element's self-impedance. This allows the elements to be individually matched to the feed lines, with no significant change taking place when the elements are formed into an array. Under those conditions, the transmission lines can be matched and the lines used as simple delay lines with easily predictable phase shift and with no transformation of current or voltage magnitude other than cable loss. This is discussed in the later section on receiving antennas.

Many arrays *can* be correctly fed with a feed system consisting of only transmission lines, but the technique requires knowledge of the element feed-point impedances in a correctly fed array. Line lengths can then be computed that provide the correct ratio of currents into those particular load impedances. The line lengths generally differ by amount that's considerably different from the element phase angle difference, and appropriate line lengths can't always be found for all arrays. This technique is described more fully in the **"The Simplest Phased Array Feed System—That Works"** section later in the chapter and illustrated in the examples in **Phased Array Design Examples**.

The Wilkinson Divider

The *Wilkinson divider*, sometimes called the *Wilkinson power divider*, was once heavily promoted as a means to distribute power among the elements of a phased array. While it's a very useful device for other purposes, it won't produce the desired current ratios in antenna elements. In most phased arrays, element feed-point resistances are different and therefore require different amounts of power to achieve the desired equal magnitude currents. (See the section on mutual coupling above.) A Wilkinson divider is intended to deliver equal powers, not currents, to multiple loads. And it won't even do that when the load impedances are different.

The Hybrid Coupler

Hybrid couplers are promoted as solving the problem of achieving equal magnitude currents with a 90º phase difference between elements. Unfortunately, they provide equal magnitude, quadrature (90° phased) currents only when the load impedances are equal and correct. And this simply isn't true of arrays with quadrature-fed elements, except for arrays consisting of short and/or lossy elements, usually suitable only for receiving. In those arrays, the hybrid coupler can be useful for the same reasons as the phasing-line approach, discussed in an earlier section.

At the time of this writing, hybrid couplers are being used in a popular commercial product for phasing at least one type of array. Reports are that it works satisfactorily. However, this shouldn't be taken as proof that the hybrid coupler forces equal magnitude 90º phased currents in loads of arbitrary impedances. No passive network, including the hybrid coupler, is capable of doing that. See **The "Magic Bullet"** below for more information.

Large Array Feed Systems

The author once worked on a radar system where the transmit array consisted of over 5,000 separate dipole elements and the receive array over 4,000 pairs of crossed dipoles, all over a metal reflecting plane, which was the sloping side of a 140 foot high building. In such large arrays, each element is in essentially the same environment as every other element except near the array edges, so almost all elements have very nearly the same feed-point impedance. While producing the phase shifts and magnitude tapers is a considerable mathematical challenge, the problem of unequal element feed-point impedances can largely be ignored. Consequently, feed methods for these large arrays are generally not suitable for typical amateur arrays of a few elements.

The Broadcast Approach

Networks can be designed to transform the element base impedances from their values in an excited array to, say, 50 Ω resistive. Then another network can be inserted at the junction of the feed lines to properly divide the power among the elements (not necessarily equally!). And finally, additional networks must be added to correct for the phase shifts and magnitude transformations of the other networks. This general approach is used by the broadcast industry, in installations that are typically adjusted only once for a particular frequency and pattern.

Although this technique can be used to correctly feed any type of array, design is difficult and adjustment is tedious, since all adjustments interact. When the relative currents and phasings are adjusted, the feed-point impedances change, which in turn affect the element currents and phasings, and so on. A further disadvantage of using this method is that switching the array direction is generally impossible. Information on applying this technique to amateur arrays can be found in Paul Lee's book, listed in the Bibliography.

The "Magic Bullet"

For about 15 years prior to this writing, this *Antenna Book* section has contained a specification for a hypothetical passive circuit that would provide equal-magnitude, 90° phased currents into two loads without respect to the load impedances. This would be a circuit to which we could connect any two elements and guarantee that they'd have exactly the correct currents.

Along with the specifcation was a request that any person knowing of such a circuit would contact the author (Roy Lewallen, W7EL) or the book editors. During this time, only a single response was received, in 1996. It was from Kevin Schmidt, W9CF, who has formulated a mathematical proof that such a circuit—in fact, one resulting in any relative phase other than 0º or 180º—*cannot exist* if restricted to reciprocal elements. (That is, it can't exist unless directional components such as ferrite circulators are used.) This means that, in order to design a network to feed elements at any other phase angle other than 0 or 180º, we *must* know the impedance of at least one element, and correct feed system operation depends on that impedance. There's no way around this requirement. At the time of this writing, Schmidt's proof can be found at **fermi.la.asu.edu/w9cf/articles/magic/index.html**.

RECOMMENDED FEED METHODS FOR AMATEUR ARRAYS

The following feed methods are able to produce element feed-point currents having a desired magnitude and phase relationship, resulting in desirable and predictable patterns. Most methods require knowing the feed-point impedance of one or more array elements *when the array element currents are the correct values*. This isn't possible to measure directly, because if the element currents were correct, the feed system would already be working properly and no further design would be necessary.

By far the easiest way to get this information, if possible, is by computer modeling. Modeling programs such as *EZNEC-ARRL* (included on the CD in the back of this book) allow you to construct an ideal array with perfect element currents, then look at the resulting feed-point impedances. Because of its simplicity and versatility, this approach is highly recommended and it's the one used for the array design examples in this chapter.

Some feed systems allow adjustment, so even an approximate result provides an adequate starting point on which to base the feed-system design. There are several other alternatives to computer modeling. One is to first eliminate the effects of coupling of the element to be measured from all other elements, usually by open circuiting the feed points of the other elements. Then the feed-point impedance of the element is measured. Next, the impedance change due to mutual coupling from all other elements has to be calculated, based on the intended currents in the other elements, their lengths and their distances from the element being measured. Mutual impedance (which is not the same as the impedance *change* due to mutual coupling) between each pair of elements must be known for this calculation and it can be determined by measurement, calculation or from a graph.

The latter two methods are possible only for the simplest element types and measurement is very difficult to do accurately because it involves resolving very small differences between two relatively large values. Accuracy of a calculated result will be reduced if any elements are relatively fat (that is, they have a large diameter, because this impacts the current distribution) or they aren't perfectly straight and parallel.

So the only situations where you're likely to get good results from approaches other than modeling are the very easiest ones to model! And modeling allows determination of the feed-point impedances of many antennas that areimpossible to calculate by manual or graphical methods. Therefore, the manual approach isn't discussed or used here. Appendix B, on the CD, contains equations and manual techniques from previous editions of *The ARRL Antenna Book*, for those who are interested. You can also find a great deal of additional information in many of the texts listed in the Bibliography, particularly Jasik and Johnson.

Current Forcing with λ/4 Lines—Elements In-Phase or 180° Out-of-Phase

The feed method introduced here has been used in its simplest form to feed television receiving antennas and other arrays, as presented by Jasik, pages 2-12 and 24-10 or Johnson, on his page 2-14. However, until first presented in *The ARRL Antenna Book*, this feed method was not widely applied to amateur arrays.

The method takes advantage of an interesting property of λ/4 transmission lines. (All references to lengths of lines are electrical length and lines are assumed to have negligible loss.) See **Fig 15**. The magnitude of the *current out* of a λ/4 transmission line is equal to the *input voltage* divided by the characteristic impedance of the line. This is independent of the load impedance. In addition, the phase of the output current lags the phase of the input voltage by 90°, also independent of the load impedance. These properties can be used to advantage in feeding arrays with certain phase angles between elements.

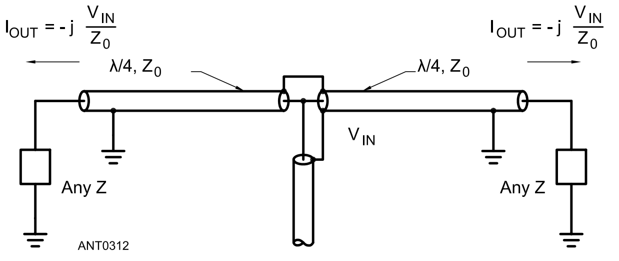
$I_{OUT} = -j \frac{V_{IN}}{Z_0}$
$I_{OUT} = -j \frac{V_{IN}}{Z_0}$
λ/4, Z_0
λ/4, Z_0
V_{IN}
Any Z
Any Z
ANT0312

Fig 15—A useful property of λ/4 transmission lines; see text. This property is utilized in the "current-forcing" method of feeding an array of coupled elements.

If any number of loads are connected to a common driving point through λ/4 lines of equal impedance, the currents in the loads will be *forced* to be equal and in-phase, regardless of the load impedances. So any number of in-phase elements can be correctly fed using this method, regardless of how their impedances might have been changed by mutual coupling. Arrays that require unequal currents can be fed through λ/4 lines of unequal impedances to achieve other current ratios.

The properties of λ/2 lines also are useful. Since the current out of a λ/2 line equals the input current shifted 180°, regardless of the load impedance, any number of half wavelengths of line may be added to the basic λ/4, and the current and phase forcing property will be preserved. For example, if one element is fed through a λ/4 line and another element is fed from the same point through a 3λ/4 line of the same characteristic impedance, the currents in the two elements will be forced to be equal in magnitude and 180° out-of-phase, regardless of the feed-point impedances of the elements.

If an array of two, and only two, identical elements is fed in-phase or 180° out-of-phase with equal magnitude currents, both elements have the same feed-point impedance. The reason is that each element sees exactly the same thing when looking at the other. In an in-phase array, each sees another element with an identical current; in an out-of-phase array, each sees another element with an equal magnitude current that's 180° out-of-phase, the same distance away in both cases. This isn't true in something like a 90° fed array, where one element sees another with a current leading its current by 90°, while the other sees another element with a lagging current.

With arrays fed in-phase or 180° out-of-phase, feeding the elements through equal lengths of feed line (in-phase) or lengths differing by 180° (out-of-phase) will lead to the correct current magnitude ratio and phase difference, regardless of the line length and regardless of how much the element feed-point impedances depart from the lines' Z_0.

Unless the feed-point impedances equal the line Z_0 or the lines are an integral number of half wavelengths long, the magnitudes of the currents out of the lines will not be equal to the input magnitudes, and the phase will not be shifted an amount equal to the electrical lengths of the lines. But both lines will produce the same transformation and phase shift because their load impedances are equal, resulting in a properly fed array. In practice, however, feed-point impedances of elements frequently are different even in these arrays, because of such things as different ground systems (for ground mounted vertical elements), proximity to buildings or other antennas, or different heights above ground (for horizontal or elevated vertical elements).

In many larger arrays, two or more elements must be fed either in-phase or out-of-phase with equal currents, but coupling to other elements can cause their impedances to change unequally—sometimes extremely so. Using the current-forcing method allows the feed system designer to ignore all these effects, while guaranteeing equal and correctly phased currents in any combination and number of 0° and 180° fed elements.

This method is used to develop feed systems for the Four Square and 4-element rectangular arrays in the **Practical Array Design** section. The front and rear elements of a Four-Square antenna provide a good example of elements having very different feed-point impedances that are forced to have equal out-of-phase currents.

"The Simplest Phased Array Feed System—That Works"

This is the title of an article in *The ARRL Antenna Compendium*, *Vol 2*, which describes how arrays can be fed with a feed system consisting of only transmission lines. (The article is available for viewing at **eznec.com/Amateur/Articles/Simpfeed.pdf** and is also on the CD included in the back of this book, along with the program *Arrayfeed1*, which solves the equations presented in the article.)

As explained earlier in the Phasing Line section, this method requires knowing what the element feed-point impedances will be in a correctly fed array. Feed-line lengths can then be computed, for most but not all arrays. These lengths will produce the desired current ratio in array elements that do present those feed-point impedances. If you know the load impedances connected to transmission lines whose inputs are connected to a common source, it's simple to calculate the resulting load currents for any transmission line lengths. However, the reverse problem is much more difficult; that is, given the load impedances and desired currents to calculate the required cable lengths.

One way to solve the problem is to choose some feed-line lengths, solve for the currents, examine the answer, adjust the feed-line lengths, and try again until the desired currents are obtained. The author used this iterative approach, using first a programmable calculator and later a computer, for some time before developing a direct way of solving for the transmission-line lengths. The direct solution method is described briefly in the *Compendium* article.

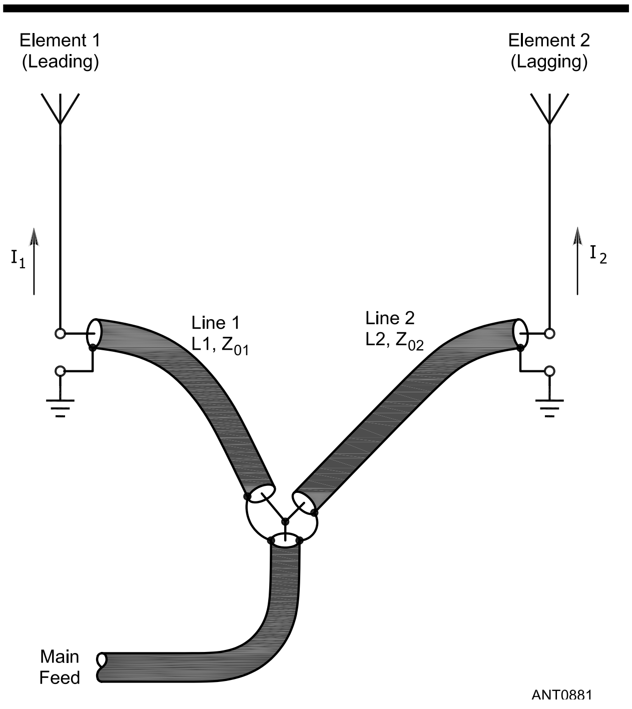

Fig 16—"Simplest" feed system for 2-element array. No matching or phasing network is used here, only transmission lines.

Fig 16 shows the basic so-called "simplest" system applied to a two element array. Although it resembles an elementary phasing-line system as described earlier, the critical difference is that the lengths of Lines 1 and 2 are calculated to provide the correct current relative magnitude transformation and phase shift when terminated with the actual feed-point impedances.

The advantage to using this "simplest" feed system is indeed its simplicity. It's no more complicated than the elementary phasing-line approach but actually works as planned. The disadvantage over some other methods is that there's no convenient adjustment to compensate for environment factors, array imperfections or inaccurately known feed-point impedances.

Also, while unusual, it's possible that no suitable feedline lengths can be found for some arrays, or at least none with practical feed-line characteristic impedances. The difference in electrical feed-line lengths almost never equals the difference in phase angles between element currents. This is because of the different line delays resulting from different feed-point impedances.

Program *Arrayfeed1*, included on the CD included with this book, can do the calculations for any two elements (alone or in a larger array), a Four-Square array or a rectangular array in which two in-phase elements are driven at any current magnitude and phase relative to the other two in-phase elements. These possibilities cover a large number of common arrays.

Arrayfeed1 can also be applied to other types of arrays using the method described in the **Feeding Larger Arrays** section in Appendix B (on the CD). The required knowledge of element feed-point impedances in a correctly fed array can be obtained using *EZNEC-ARRL*, also included on the CD. Examples of the design of a "simplest" feed system for several different arrays using *EZNEC-ARRL* and *Arrayfeed1* can be found in the **Phased Array Design Examples** section.

When a solution is possible for a given choice of line characteristic impedances, a second solution with different lengths is always available. See the comments in the introductory part of the **Phased Arrrray Design Examples** section about choosing the solution to use.

An Adjustable L-Network Feed System

Adjustment of the current ratio of any two elements requires varying two independent quantities; for example the magnitude and phase of the current ratio. Two degrees of freedom—adjustments that are at least partially independent—are required. The "simplest" all-transmission line feed system described earlier adjusts the lengths of the two transmission lines to achieve the correct ratio.

But if the antenna characteristics aren't well known—for example, if the ground resistance isn't known even approximately—then the initial "simplest" design won't be optimum and adjustment can be difficult and tedious. The current-forcing method produces correct currents independently of the element characteristics, so it doesn't require adjustment as long as the elements are identical. But it's suitable only for feeding elements in-phase or 180° out-of-phase and a few fixed current-magnitude ratios.

The addition of a simple network as shown in **Fig 17** allows you to easily adjust feeding of element pairs at other relative phase angles and/or magnitude ratios. Any desired current ratio (magnitude and phase) can be obtained with two elements fed with any lengths of wire, equal or unequal, by adding a network.

However, calculations for the general case are complex. The problem becomes much simpler if the transmission lines are restricted to lengths of odd multiples of λ/4, forming a modified "forcing" system that includes an added network. There are at least three additional advantages of this scheme. One is that a λ/4 line is easy to measure, even if the velocity factor isn't known. This is described in the **Practical Aspects of Phased Array Design** section.

Second is that the feed system becomes completely insensitive to the feed-point impedance of one of the two elements. And the third is that the transmission lines of "forcing" systems feeding groups of elements in larger arrays can be used in place of the normal λ/4 lines. This greatly simplifies both the design of feed systems for larger arrays and the feed systems themselves. Note that both lines can be changed to 3λ/4 if necessary to span the physical distance between elements, but *both* lines must be the same 3λ/4 length.

This basic feed method can be used for any pair of elements, or for two groups of elements having forced equal currents. (See **Feeding Four Element and Larger Arrays** below.) Many networks can accomplish the desired function, but a simple L network is adequate for most feed systems. The network can be designed to produce a phase lead or phase lag. The basic two-element L-network feed system is shown in Fig 17. Many variations of this general method can be used, but the equations, program, and method to be discussed here apply only to the feed system shown.

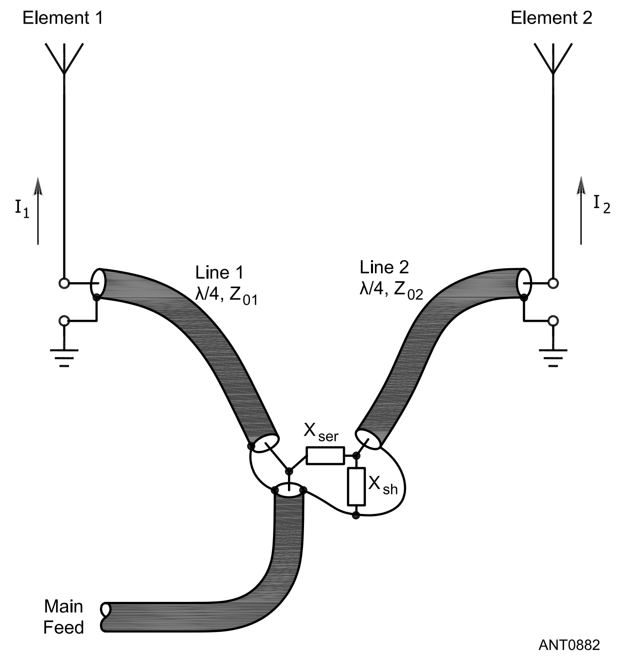

Fig 17—The addition of a simple L-network to Fig 16 allows you to easily adjust feeding of element pairs at other relative phase angles and/or magnitude ratios.

If the phase angle of I_2/I_1 is negative (element 2 is lagging element 1), the L network will usually resemble a low-pass network (X_{ser} is an inductor and X_{sh} is a capacitor). But if the phase angle is positive (element 2 lagging element 1), the L network will resemble a high-pass network (X_{ser} is a capacitor and X_{sh} is an inductor). However, some current ratios and feed-point impedances could result in both components being inductors or both being capacitors.

If it's desired to maintain symmetry in the feed system, X_{ser} can be divided into two components, each being inserted in series with a transmission line conductor. If X_{ser} is an inductor, the new components will each have half the value of the original X_{ser}, as shown in **Fig 18**. If X_{ser} is a capacitor, each of the new components will be twice the original value of X_{ser}.

Because of the current-forcing properties of λ/4 lines, we need to make the ratio of *voltages* at the inputs of the lines equal to the desired ratio of *currents* at the output ends of the lines; that is, at the element feed points. The job of the L network is to provide the desired voltage transformation. If the output-to-input voltage ratio of the network is, say, 2.0 at an angle of –60°, then the ratio of element currents (I_2/I_1)

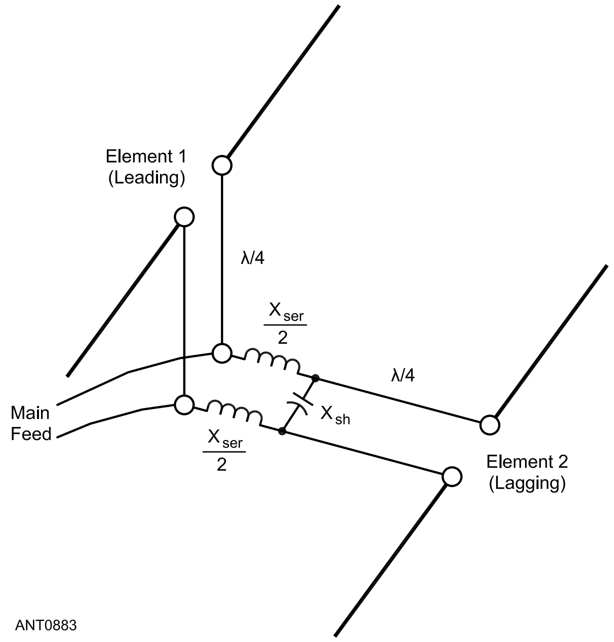

Fig 18—Symmetrical feed system similar to Fig 17, where feed network is split into two symmetrical parts.

will be 2.0 at an angle of –60°. The voltage transformation of the network is affected by the impedance of element 2, but not by the impedance of element 1. So only the impedance of element 2 must be known to design the feed system.

Equations for designing the L network are given in Appendix B, but the program *Arrayfeed1* is included on the CD to make it unnecessary to solve them. The feed-point impedance of the lagging element or group of elements must be known in order to design the network. This can best be determined by modeling the array with *EZNEC-ARRL*. The impedance can be manually calculated for some simple element and array types by using the equations in Appendix B, but those same types of element and array are simple to model.

Examples of the design of L-network feed systems for several different arrays using *EZNEC-ARRL* and *Arrayfeed1* can be found in the **Phased Array Design Examples** section. A similar application of this feed system and a spreadsheet program for calculation was developed by Robye Lahlum, W1MK, and described in *Low-Band DXing* (see the Bibliography). *Arrayfeed1* can be used for the applications of the feed system described in that book if desired.

Additional Considerations

Feeding 4-Element and Larger Arrays

Both the simplest and L-network feed systems described above can be extended to feeding larger arrays having two groups of elements in which all the elements in a group are in-phase or 180° out-of-phase with each other—basically, any group that can be fed with the current-forcing method. The elements in each group are connected to a common point with λ/4 or 3λ/4 lines to force the currents to be in the correct ratio within the group. Then the "simplest" or L-network feed system can be used to produce the correct phasing between the two groups, just as it does between two individual elements.

Two common arrays that fit this description are the Four Square and the 4-element rectangular array. But more elaborate arrays could be constructed and fed using this method, such as a pair of binomial arrays. (A single binomial array is described in the **Phased Array Design Example** section below.) The *Arrayfeed1* program incorporates additional calculations necessary for designing Four Square and 4-element rectangular arrays. The general procedure for adapting the feed methods to other larger arrays can be found in Appendix B.

What If the Elements Aren't Identical?

Getting the desired pattern requires getting the correct relative magnitude and phase of the *fields* from the elements. If the elements are identical, which we've generally assumed up to this point, then producing currents of the desired magnitude and phase will create the desired fields (neglecting mutual coupling current distribution effects, discussed elsewhere).

But what if the elements aren't identical? Fortunately, the feed systems described here can still be used for any 2-element array and some more complex arrays, provided that the system can be accurately modeled. But a slightly different approach is required than for identical elements.

The first step is to model the array with a current source at the feed point of each element. Next, the magnitudes and phases of the model source currents are varied until the desired pattern is achieved. Then the ratio of feed-point (source) currents is calculated and this value, along with the feed-point impedances reported by the model, are used for the feed system design. The feed system will produce the same ratio of currents as the model, resulting in the same pattern.

In general, this approach won't work with shunt fed towers or gamma-fed elements because of the difficulty of accurately modeling those structures. See **Shunt and Gamma Fed Towers and Elements** for more information.

Shunt and Gamma Fed Towers and Elements

In a shunt, gamma, or similarly fed tower or element, the feed-point current isn't the same as the main current flowing in the element. The ratio between the feed-point current and element current isn't a constant, but depends on a number of factors. The ratio of currents in shunt or gamma fed elements are typically different—often vastly different—from the currents at the feed points. This complicates the design of feed systems for arrays of these elements.

An even more limiting problem is that the feed-point impedances are difficult to determine. The feed-point impedances of one or more elements in a properly fed array must be known in order to design a feed system for anything but

2-element in-phase or 180° out-of-phase arrays.

The only practical way to get this information for a shunt or gamma fed array is by modeling an array having the desired element currents. But Cebik has pointed out ("Two Limitations of NEC-4"—see the Bibliography) that many common antenna analysis programs, including *EZNEC-ARRL*, have difficulty accurately modeling folded dipoles with unequal diameter wires. The same problem applies to shunt and gamma-fed elements when the element diameter is significantly different from the diameter of the shunt or gamma feed wire. Without accurate feed-point impedances, feed systems can't be designed to work without adjustment. It might be possible to get reasonably accurate results from a *MININEC*-based modeling program, but there are a number of issues which must be given great care when using one. (See Lewallen, "*MININEC*—The Other Edge of the Sword," listed in the Bibliography.)

If such a *MININEC* program is available, you would have to model the complete array including feed system, with sources at the normal feed points in the shunt or gamma wires. Next, you would have to adjust the magnitudes and phases of the sources to produce the desired pattern. The reported source impedances and currents would be the ones you would use to design the feed system. It's likely that some adjustment would be necessary, so an adjustable system such as the L-network feed system described later would be best.

Loading, Matching and Other Networks

Adding a component such as a loading inductor in series with an element or element feed point won't change the ratio of element current to feed-point current. As a result, a feed system designed to produce a particular ratio of element currents will still function properly if the elements contain series components. The extra feed-point impedance introduced by the loading component(s) must be considered when designing the feed system, however. Similarly, end or top loading won't alter the relationship between feed-point and element current, provided that the current distribution in the elements is essentially the same. (See **Feed-point vs Element Current** in a previous section.)

However, insertion of any *shunt* component, or a network containing a shunt component, *will* alter the relationship between feed-point and element current because it will divert part of the feed-line current that would otherwise flow into the antenna. As a result, a feed system designed to deliver correct currents at the feed points will produce incorrect element currents and therefore an incorrect pattern. Therefore, any components or networks other than a series loading component should be avoided at any place in the feed system on the antenna side of the point at which the feed system splits to go to the various elements.

There are a few exceptions to this rule. If the feed-point impedances of the elements when in the excited array are equal, then identical networks with or without shunt components can be put at the feed points of the elements and the proper element current ratio maintained—so long as the feed system is designed to deliver the proper feed-point current ratio with the networks in place. Equal element impedances occur in arrays having only two identical elements fed in-phase or 180º out-of-phase, or arrays of any number of elements where the elements are electrically short and/or very lossy.

Baluns in Phased Arrays

For purposes of achieving the correct array pattern, baluns aren't usually required when feeding grounded vertical elements with coaxial cable feed lines. However, a balun might be desirable if current induced onto the outside of the feed line by mutual coupling to the elements is causing RF in the shack. And with arrays of dipole or other elevated elements, baluns can be important to achieve the proper element current ratio, as explained below.

First, however, the general rules for using baluns in phased arrays will be stated. Here, "main feed line" means the feedline going from the transmitter or receiver to the common point where the system splits to feed the various elements. "Phasing-system lines" means any transmission lines between that common point and any element. The rules:

- **Rule 1:** A balun or baluns (more specifically, a current, sometimes called a choke balun) should be used as necessary to suppress unbalanced current on the main feed line. This usually isn't required when feeding grounded elements with coaxial feed line from an unbalanced rig or tuner. Unbalanced current can occur on either coax or parallel-conductor line.

"Baluns: What They Do and How They Do It", listed in the Bibliography, describes conducted-imbalance (common-mode) currents. Imbalance can also be caused by mutual coupling to the array elements. Common-mode currents have at least two undesirable effects on array performance. First, the imbalance current can flow from the main feed line to the phasing system lines, not necessarily splitting in the right proportion to maintain the correct element current ratio. This can affect the array pattern. In practice, however, this effect is likely to be small unless the common-mode current is unusually large. Even a small common-mode current, however, results in main feed-line radiation, and even a small amount of radiation can significantly degrade array pattern nulls. Any type of current balun can be used on the main feed line, at any place along the line, without any effect on the array pattern except to the extent that it reduces common mode current.

- **Rule 2:** No balun or any other component or network should be inserted in any phasing system line that will alter the line length or characteristic impedance. This means that baluns in phasing system lines must be of a type made from the phasing line itself. Options are the W2DU type balun, consisting of ferrite cores placed along the outside of the feed line; an air-core balun made by winding part of the line into an approximately self-resonant or otherwise high-impedance coil; or winding part of the line onto a ferrite core or rod to make a several-turn winding. When coaxial cable is used, the feed system characteristics are dictated by the inside of the cable. Any cores or winding of the outside prevents common-mode current on the outside, but otherwise have no effect on the phasing performance. This rule applies equally to parallel wire line, where the balun affects only common-mode current (equivalent to current on the outside of coax) while the phasing performance depends on differential mode current (equivalent to the current on the inside of coax).

Baluns are important when feeding dipole or other elevated arrays, unless a fully balanced tuner is used. This is because common-mode current represents a diversion of some of the current that should be going to the array elements. The presence of common-mode current means that the element currents are being altered from the desired ratio and therefore the pattern won't as intended. A balun should be placed wherever a path for current exists other than along a parallel-line conductor or on the inside of a coaxial line. Such a path exists, for example, where a coaxial cable connects to a dipole, as shown in Fig 1 of the balun article referenced above. Or a path can exist where a parallel-conductor transmission line connects to an unbalanced tuner or to a coaxial line, as shown in Fig 2 of that article. In both those cases, a path exists for a common-mode current to flow on the outside of the coax cable. A balun creates a high impedance to this current, thereby reducing its magnitude. But remember that all baluns must conform to the rules above.

Fig19 shows recommended balun locations for a coax-fed dipole array using an L-network feed system.

Receiving Arrays

While it might not be entirely intuitive, an array designed for a particular gain and pattern for transmitting that considers mutual coupling, element currents, field reinforcement and cancellation, and so forth, will perform exactly the same when receiving. So a receiving array can be designed by approaching the problem as though the array were to be used for transmitting.

However, at HF and below, the system requirements for transmitting and receiving antennas are different, so receiving-only arrays can be designed that aren't suitable for transmitting but are perfectly adequate for receiving in that frequency range. The reason, described in more detail in Chapter 13 of this book, is that at HF and below atmospheric noise is typically much greater than a receiver's internally generated noise. Lowering a receiving antenna's gain and efficiency reduces the signal and atmospheric noise both by the same factor. Because the overall noise is for practical purposes all atmospheric noise, the signal/noise ratio isn't affected by antenna efficiency.

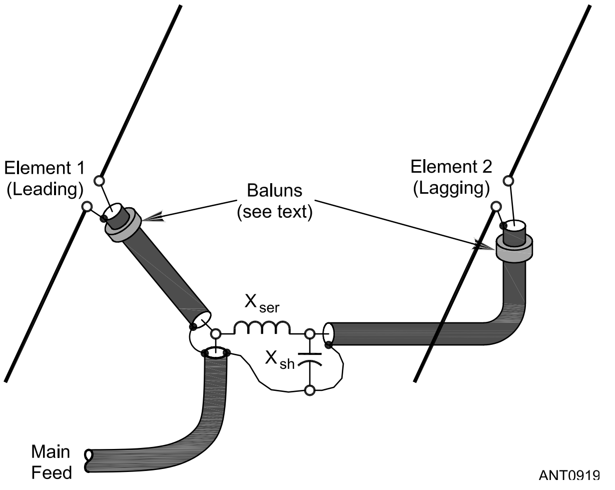

Fig 19—Adding choke baluns to a two-dipole feed system to get rid of common-mode currents radiated onto the coax shields.

Of course, a point can be reached where the atmospheric noise is so reduced by inefficiency that the receiver itself becomes the dominant source of noise, but this typically doesn't happen until the antenna is extremely inefficient. When transmitting, reduced efficiency lowers the transmitted signal, but it has no effect on the receiving station's noise. So reduced efficiency of a transmitting antenna results in a reduced signal/noise ratio at the receiving end, and consequently should be avoided.

Mutual coupling effects can be minimized by increasing the loss (and therefore reducing the efficiency) of the elements, or by reducing the element sizes to a small fraction of a wavelength. Doing the second without the first isn't usually a good idea because the feed-point impedance tends to change rapidly with frequency for very small elements, making an antenna that works well over only a narrow bandwidth. But increasing loss broadens the bandwidth, even for small elements, as well as reducing mutual coupling effects. So this approach is often taken for designing a receiving-only array. With mutual coupling effects minimized because of loss, feed-system design becomes relatively simple, provided a few simple rules are followed. See **Loss Resistance, Mutual Coupling, and Antenna Gain**, above.

Phased Array Design Examples

This section, also written by Roy Lewallen, W7EL, presents examples of feed-system design for several kinds of array using the design principles given in previous sections. All but the last example array are assumed to be made of λ/4 vertical elements. The last example is for a halfwave-dipole array, which illustrates that exactly the same method can be used for arrays of any shape of elements, including dipole, square (quad) and triangular. Likewise, the methods shown here apply equally well to VHF and UHF arrays. The first example includes more detail than the remaining ones, so you should read it before the others.

General Array Design Considerations

If either the "simplest" feed system (Fig 16) or L-network feed system (Fig 17) is used, the feed-point impedance of one or more elements—*when the array elements all have the correct currents*—must be known. By far the best way to determine this is by modeling. If accurate modeling isn't practical for some reason, an estimation should be made from an approximate model, and you should expect to have to adjust the feed system after building and installing it.

Manual calculation methods for some simple configurations are given in Appendix B (on the CD), but calculation is tedious and, as stated earlier, the configurations for which this method works are the very ones which are easiest to model. *EZNEC-ARRL* (also included on the CD) is used in the following examples to determine feed-point impedance. Space doesn't permit detailed instructions here on creating the models, so they are included in complete form. They should provide a convenient starting point for any variations you might like to try. See the *EZNEC-ARRL* manual (accessed by clicking **Help/Contents** in the main *EZNEC-ARRL* window) for help in using this program.

In the following examples, vertical elements are close to λ/4 high and dipole elements close to λ/2, and their lengths have been adjusted for resonance when all other elements are absent or open circuited. There's actually no need in practice to make the elements self-resonant—it's simply used as a convenient reference point for these examples. You'll also find it interesting to see how much reactance is present at the feed points of the elements when in the arrays, knowing that it's very nearly zero when only one element is present.

In any real grounded vertical array, there is ground loss associated with each element. The amount of loss depends on the length and number of ground radials, and on the type and wetness of the ground under and around the antenna. This resistance becomes part of the feed-point resistance, so it must be included in the model used to determine feed-point impedance. The 90° Fed, 90° Spaced Array example below discusses how this is done. **Fig 20** gives resistance values for typical ground systems, based on measurements by Sevick (July 1971 and March 1973 *QST*). The values of feed system components based on Fig 20 will be reasonably close to correct, even if the ground characteristics are somewhat different than Sevick's.

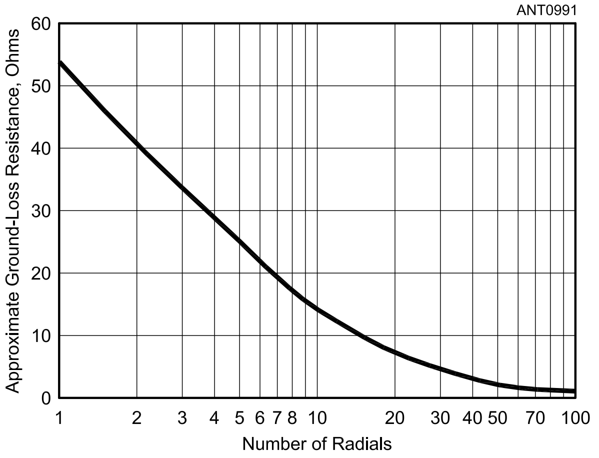

Fig 20—Approximate ground system loss resistance of a resonant λ/4 ground-mounted vertical element versus the number of radials, based on measurements by Jerry Sevick, W2FMI. Moderate length radials (0.2 to 0.4 λ) were used for the measurements. The exact resistance, especially for only a few radials, will depend on the nature of the soil under the antenna. Add 36 Ω for the approximate feed-point resistance of a thin resonant λ/4 vertical.

Feed systems for the design example arrays to follow are based on the resistance values given below.

Number of Radials	*Loss Resistance, Ω*
4	29
8	18
16	9
Infinite	0

Elevated radial systems also have some ground loss, although it can be considerably less than a system with the same number of buried radials. This loss will be automatically included in the feed-point impedance of a model which includes the elevated radials, so no further estimation is required. Be sure to use Perfect, High-Accuracy ground type when modeling an elevated radial system with *EZNEC-ARRL*. In other *NEC-2* based programs, this might be referred to as Sommerfeld type ground. More information can be found in the *EZNEC-ARRL* manual.

The matter of matching the array for the best SWR on the feed line to the station is not dealt with here, since it's a separate problem from that of the main topic, which is designing feed systems to produce a desired pattern. Some of the simpler arrays provide a match that is close to 50 or 75 Ω, so no further matching is required. However, as shown by program *Arrayfeed1*, many larger arrays present a less favorable impedance for direct connection and will require matching if a low SWR on the main feedline is required. If matching is necessary, the appropriate network should be placed in the single feed line running to the station. Attempts to improve the match by adjustment of the phasing L network, individual element lengths, matching at the element feed points or individual element feeder lengths will usually ruin the current balance of the array. Program *TLW*, included on the CD, can be used for designing an appropriate matching network. Additional information on impedance matching may be found in Chapters 25 and 26 of this book.

Choosing Arrayfeed1 Solutions

When designing a feed system for a two element array, *Arrayfeed1* program allows you to choose the characteristic impedances of the two transmission lines going to the elements, which don't have to be equal, so you have your choice of more than one solution. However, directional array switching is much more difficult if the lines have different impedances, so in general you should use the same characteristic impedances.

For larger arrays, *Arrayfeed1* requires the feed lines to all elements to have the same impedance. In choosing the transmission line impedance values, usually you can simply use convenient impedances. But in general, you should avoid solutions where component reactance (X) values are vastly different (say, more than three times or less than one third as large) as the line characteristic impedances. Such networks will become more critical to adjust, and both the impedance and pattern will change more rapidly with changes in frequency. You can usually avoid this situation by choosing feed-line impedances that are in the same ballpark as the element feed-point impedances. The last example in the Practical Array Design section illustrates this problem and its solution.

When designing a "simplest" feed system, the most broadbanded and least critical system is usually one where the difference in electrical feed-line lengths is closest to the relative element phase angle. Here, "broadbanded" means that the pattern changes less with frequency, not necessarily that the SWR changes less. However, an array that's broadbanded in the pattern sense is usually also relatively broadbanded with respect to SWR.

Arrayfeed1 reports the impedance seen at the main array feed point. While it might be tempting to choose the solution producing the lowest SWR on the main feedline, you'll end up with a less critical and more broadbanded system if you base your choice on the criteria given above, and provide separate impedance matching at the array's main feed point when necessary.

90° Fed, 90° Spaced Vertical Array

This example illustrates the design of both "simplest" and L-network feed systems for a 2-element, 90° spaced and fed vertical array. The first task when using either feed system is to determine the feed-point impedances of the elements when placed in an array having the desired element currents. The "simplest" feed system method requires knowledge of both element impedances, while the L-network system requires you to know only one. Actually, it's equally easy to determine both as it is to find just one, using *EZNEC-ARRL*. (Appendix B contains equations for those interested in manual methods or for more insight as to how the impedances come about.) The first step is to specify the antenna we want. For this example, we'll specify:

- Frequency: 7.15 MHz.
- Two identical, one inch (2.54 cm) diameter, 33 feet (10.06 meter) long elements spaced 90 electrical degrees, with element currents equal in magnitude and 90° apart in-phase.
- 8 buried radial wires, 0.3 λ long, under each element.

A model of this antenna has been created and furnished with *EZNEC-ARRL*. So the next step is to start *EZNEC-ARRL*, click the **Open** button, enter **ARRL_Cardioid_Example** in the text box (or double-click it in the file list) to open example file **ARRL_Cardioid_Example.EZ**.

This *EZNEC* example model uses a *MININEC*-type ground, which is the same as perfect ground when calculating antenna currents and impedances. A real antenna would have some additional resistive loss due to the finite conductivity of the ground system. The only way to model a buried radial ground system with an *NEC-2* based program like *EZNEC* is to create radial wires just above the ground (using the Real, High-Accuracy ground type), because *NEC-2* can't handle buried conductors.

This provides only a moderate approximation of a buried system. Another way to estimate ground-system resistance is to measure the feed-point impedance of a single element, then subtract from that the resistance reported for a model of that element over perfect (or *MININEC*-type) ground. For most uses, however, an adequate approximation can be made by simply referring to the graph of Fig 20. As stated previously, the feed system design depends on the feed-point impedances of the elements, which in turn depend on the ground system resistance. So the ground system resistance must be known, approximately anyway, before designing the feed system. At the end of this example we'll investigate the effect of changes in the ground system or errors in estimating the resistance on the pattern.

For 8 radials, Fig 20 shows the ground system resistance to be about 18 Ω. This is included in the example model as a simple resistive load at the feed point of each element. Click the **Src Dat** button to see the feed-point impedances of the two elements. In this model, Source 1 is at the base of Wire 1 (element 1), and Source 2 is at the base of Wire 2 (element 2). Notice in the **Source Data** display that the Source 1 current has been specified at 1 amp at 0°, and Source 2 is 1 amp at –90°. So the Source 2 element is the lagging element. You should see impedances of 37.53 – *j*19.1 Ω for element 1 and 68.97 + *j*18.5 Ω for element 2. These are the feed-point impedances resulting when the array is ideally fed, with equal magnitude and 90° phased currents. Record these values for use in *Arrayfeed1*.

Click the **FF Plot** button to generate a plot of the azimuth pattern at an elevation angle of 10°. In the 2D Plot Window, open the **File** menu and select **Save Trace As**. Enter **Cardioid_Ideal Feed** in the **File Name** box, then click **Save**. This saves the cardioid pattern plot so you can compare it later to the pattern you get with the transmission line feed system.

Now it's time to design the feed system. Refer to the appropriate subheading below for the design of each of the two kinds of feed systems. Both systems use program *Arrayfeed1* program.

"Simplest" (Transmission Line Only) Feed System

Start *Arrayfeed1*. In the **Array Type** frame, select **Two Element**. In Feed System Type, select **"Simplest."** In the **Inputs** frame, enter the following values:

Frequency MHz = 7.15; Feed-point impedances – Leading Element: R ohms = 37.53, X ohms = –19.1; Lagging Element: R ohms = 68.97, X ohms = 18.5 (these are the element R and X values from *EZNEC-ARRL*). We'll be discussing the array input impedance, so check the **Calc Zin** box near the lower left corner of the main window if it's not already checked.

We're free to choose any transmission-line characteristic impedances we want, so long as we can get cables with those impedances. And the two cables don't have to have the same characteristic impedances. Each choice will lead to a different set of solutions. But sometimes a solution isn't possible, which then requires choosing different line impedances. Let's try 50 Ω for both lines. Enter **50** in both **Z0** boxes.

Finally, enter **1** for the lagging:leading **I Mag**, and **–90** for the **Phase**. Click **Find Solutions**. The result is no solution! So enter **75** into both the line impedance boxes and click **Find Solutions** again. You should now see two sets of results in the **Solutions** frame, electrical lengths of 68.80° and 156.03° for the first solution and 131.69° and 185.00° for the second. (Notice that the difference in length between the two lines isn't 90° for either solution, although the first solution is quite close. It's normal for the feed-line length difference to be different than the phase difference, due to the unequal element feed-point impedances caused by mutual coupling.)

The solution with a line length difference closest to the element phase difference is usually preferable. Also, all else being equal, the solution with shortest lines is better providing that the lines will physically reach the elements. This is because the current magnitude and phase will change less with frequency than for a longer-length solution. However, there might be some cases where the change with frequency luckily compensates for the changing electrical distance between elements, so it's not a bad idea to model both solutions unless you plan on using the antenna over only a narrow frequency range.

In this case, the first solution looks best in all respects. The sum of the two lines in the first solution is about 225 electrical degrees. Assuming the lines have a velocity factor of 0.66, the total length of the lines will be more than 148 physical degrees. Since our two elements are spaced 90 physical degrees apart, the lines will comfortably reach. If they didn't, we could either use the second solution's lengths, use cable with a higher velocity factor or add a half wavelength to both the line lengths in the first solution.

The impedance **Zin** shown by *Arrayfeed1* is the impedance at the input to the feed system, so it's the impedance that will be seen by the main feed line. The second solution provides nearly a perfect match for a 50-Ω transmission line. But the first solution is good for nearly all applications. Also a 50-Ω line connected to the first solution's feed system would have an SWR of only 1.65:1, which wouldn't require any matching under most circumstances. Normal line loss would reduce the SWR even more at the transmitter end of the feed coax.

To find the required physical line lengths, enter the cable velocity factor and make your choice of units in the **Physical Lengths** frame. The design is now complete; all you have to do is cut two lines to the specified lengths and connect one from a common feed point to each element as shown in Fig 16, or the screen capture from *Arrayfeed1* shown in **Fig 21**.

Next, we'll design an L-network feed system for the same array.

L Network Feed System

In *Arrayfeed1*, select **L Network** in the **Feed System Type** frame. The program doesn't need to know the leading element impedance to calculate the L-network values, but it does need it to calculate the array input impedance. If you want to know the impedance, check the **Zin** box at the lower left corner of the main window, otherwise you can uncheck it and the input box for the leading element Z will disappear. The values from the "Simplest" analysis should still be present in the appropriate boxes; if not, refer to the "Simplest" feed system design above and re-enter the values. Again, we'll use 75 Ω for the line impedances, since that gave us a

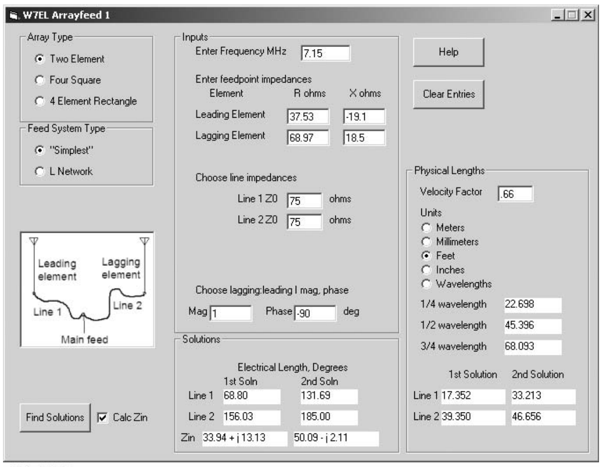

W7EL Arrayfeed 1

Array Type
● Two Element
○ Four Square
○ 4 Element Rectangle

Feed System Type
● "Simplest"
○ L Network

Inputs
Enter Frequency MHz: 7.15
Enter feedpoint impedances

Element	R ohms	X ohms
Leading Element	37.53	-19.1
Lagging Element	68.97	18.5

Choose line impedances
Line 1 Z0: 75 ohms
Line 2 Z0: 75 ohms

Choose lagging:leading I mag, phase
Mag: 1
Phase: -90 deg

Solutions
Electrical Length, Degrees

	1st Soln	2nd Soln
Line 1	68.80	131.69
Line 2	156.03	185.00
Zin	33.94 + j 13.13	50.09 - j 2.11

☑ Calc Zin

Physical Lengths
Velocity Factor: .66
Units
○ Meters
○ Millimeters
● Feet
○ Inches
○ Wavelengths
1/4 wavelength: 22.698
1/2 wavelength: 45.396
3/4 wavelength: 68.093

	1st Solution	2nd Solution
Line 1	17.352	33.213
Line 2	39.350	46.656

ANT0992

Fig 21—Screen capture from *Arrayfeed1* program for "Simplest" 2-element phased array shown in Fig 16 and whose feed-point impedances are modeled by *EZNEC-ARRL*.

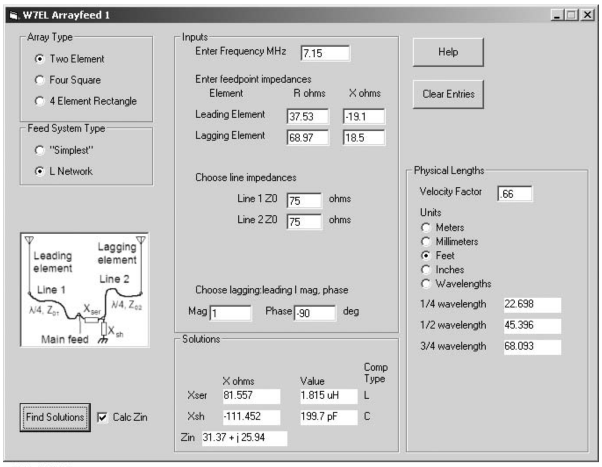

Array Type: ● Two Element ○ Four Square ○ 4 Element Rectangle

Feed System Type: ○ "Simplest" ● L Network

Inputs

Enter Frequency MHz 7.15

Enter feedpoint impedances

Element	R ohms	X ohms
Leading Element	37.53	-19.1
Lagging Element	68.97	18.5

Choose line impedances

Line 1 Z0 75 ohms

Line 2 Z0 75 ohms

Choose lagging:leading I mag, phase

Mag 1 Phase -90 deg

Solutions

	X ohms	Value	Comp Type
Xser	81.557	1.815 uH	L
Xsh	-111.452	199.7 pF	C

Zin 31.37 + j 25.94

☑ Calc Zin

Physical Lengths

Velocity Factor .66

Units: ○ Meters ○ Millimeters ● Feet ○ Inches ○ Wavelengths

1/4 wavelength 22.698

1/2 wavelength 45.396

3/4 wavelength 68.093

ANT0993

Fig 22—Screen capture from *Arrayfeed1* program for L-network feed system using "current-forcing" properties of λ/4 feed lines.

solution for the "Simplest" feed system. This feed system is more versatile, though, so we could use 50-Ω lines with this feed system if desired.

Click **Find Solution** and see the results in the **Solution** frame. See screen capture in **Fig 22**. With 75-Ω lines, the L network consists of a series inductor of 1.815 µH and a shunt capacitor of 199.7 pF, connected as shown in the diagram in the left part of the program window. To find the physical length of the λ/4 lines, enter the velocity factor and choice of units in the **Physical Lengths** frame.

The main feed-point impedance of 31.37 + *j*25.94 Ω would result in about a 2.2:1 SWR on a 50-Ω feed line, which would be acceptable for many applications. It could easily be reduced to 1.6:1 by the simple addition of a series capacitor of 25.94 Ω reactance (858 pF) at the main feed point or, of course, reduced to 1:1 with a simple L network or other matching system designed with the *TLW* program.

Pattern Verification and Effect of Loss Resistance—"Simplest" System

EZNEC-ARRL doesn't have the capability to model an L-network, so *EZNEC-ARRL* verification of the pattern and the effect of various modifications can be done only for the "simplest" feed system.

EZNEC model **ARRL_Cardioid_TL_Example.EZ** has been created to model the "simplest" feed system just designed. Open it with *EZNEC-ARRL*. In the View Antenna Display, you can see the transmission lines connecting to the source midway between the antennas. In *EZNEC*, the physical locations of the ends of transmission line models don't have to be the same as the physical locations, so the view isn't a precise representation of what the actual setup would look like. (You can find more about this in **ARRL_Cardioid_TL_Example.txt**, the Antenna Notes file that accompanies example file **ARRL_Cardioid_TL_Example.EZ.**)

Click **FF Plot** to generate a 2D pattern of the antenna. In the 2D Plot Window, open the **File** menu and select **Add Trace**. Select **Cardioid—Ideal Feed** (which you saved earlier) and click **Open**. The added plot overlays perfectly, indicating that the pattern using this feed system is identical to the pattern we got with perfect current sources at each feed point.

To check the feed-point currents, click the **Currents** button. In the resulting table, you can see that **Wire 1 Segment 1** current is 0.56467 A at a phase of –56.73° and **Wire 2 Segment 1** current is 0.56467 A at –146.7°. (If you get the correct phase angles but wrong magnitudes, open the main window **Options** menu, select **Power level**, and make sure the **Absolute V, I sources** box is checked.) The ratio is 1.0000 at an angle of –89.97°, which is within normal error bounds for the desired 1 at –90°.

As a check on *Arrayfeed1*, click the **Src Dat** button to find the impedance seen by the source. This would be the impedance at the main feedline connection in the real array. *EZNEC-ARRL* reports 33.96 + *j*13.11 Ω, very close to the 33.94 + *j*13.13 Ω given by *Arrayfeed1* in Fig 21. Small differences of this order are normal and to be expected. This provides a further check that the *EZNEC-ARRL* model is correctly analyzing the *Arrayfeed1* feed system.

This *EZNEC-ARRL* model uses lossless transmission lines of a fixed physical length rather than a fixed electrical length (number of degrees), so they'll behave like real lines as the frequency is changed. By changing the *EZNEC* frequency and re-running the 2D plot, you can see that the front-to-back ratio degrades at 7.0 and 7.3 MHz. A slight adjustment of one or more line lengths, or a new *Arrayfeed1* solution at a slightly different frequency might produce a better compromise for some uses.

Other things you can try are to evaluate the second *Arrayfeed1* solution, or to try using different line impedances. (Keep the two line impedances equal if you anticipate doing array direction switching.) The effect of varying ground system resistance can also be evaluated by clicking the **Loads** line in the main window and changing the load resistance values. For example, if the ground system resistance were 9 Ω instead of the 18 Ω we have assumed, the front/back ratio would drop from about 32 to about 20 dB. Note that changing the *EZNEC* ground conductivity in this model has no effect on the feed-point current ratio. With a *MININEC* type ground, it's used only for pattern calculation—the ground is assumed perfect during impedance and current calculations, and the only ground system loss resistance in the model is what we've specifically put in as loads.

Not surprisingly, the forward gain is affected very little by changes in frequency or ground system loss. To find the gain relative to a single element, compare the reported dBi gain of **ARRL_Cardioid_Example** with the same model with one of the elements deleted. You'll find it's very close to 3.0 dB. The 90° fed, 90° spaced array is a special case of array where the effects of mutual coupling on the two elements are opposite and cancel, resulting in the same gain as if mutual coupling didn't exist. But mutual coupling most certainly does exist!

The second solution presented a more favorable main feed-point impedance, so it would be tempting to use that one instead of the first solution. Replacing the feed-line lengths with the second solution lengths to model the second solution shows that the front/back ratio deteriorates more at the band edges when the second solution is used. This might be tolerable if restricted frequency use is anticipated. But it does illustrate that the solution with shorter lines is generally more broadbanded and that the choice of solution shouldn't in general be based on the one giving the most favorable impedance.

A Three-Element Binomial Broadside Array

An array of three in-line elements spaced λ/2 apart and fed in-phase gives a pattern that is generally bidirectional. If the element currents are equal, the resulting pattern has a forward gain of 5.7 dB (for lossless elements) compared to a single element, but it has substantial side lobes. If the currents are tapered in a binomial coefficient 1:2:1 ratio (twice the current in the center element as in the two end elements), the gain drops slightly to just under 5.3 dB, the main lobes widen and the side lobes disappear.

The array is shown in **Fig 23**, and an *EZNEC-ARRL* model of the antenna over perfect ground to show the ideal pattern is provided as **ARRL_Binomial_Example.EZ**. To obtain a 1:2:1 current ratio in the elements, each end element is fed through a 3λ/4 line of impedance Z_0. Line lengths of 3λ/4 are chosen because λ/4 lines will not physically reach. The center element is fed from the same point through two parallel 3λ/4 lines of the same characteristic impedance. This is equivalent to feeding it through a line of impedance $Z_0/2$. The currents are thus forced to be in-phase and to have the correct ratio. **ARRL_Binomial_TL_Example.EZ** is an *EZNEC-ARRL* model that shows this feed system with lossless transmission lines. The reader is encouraged to experiment with this model to see the effect of changes in frequency, the addition of loss resistance (as resistive loads at the element feed points) and other alterations on the array pattern and gain. You should also replace the perfect ground with *MININEC* type of ground to show how radiation patterns over real ground differ from the theoretical perfect-ground pattern.

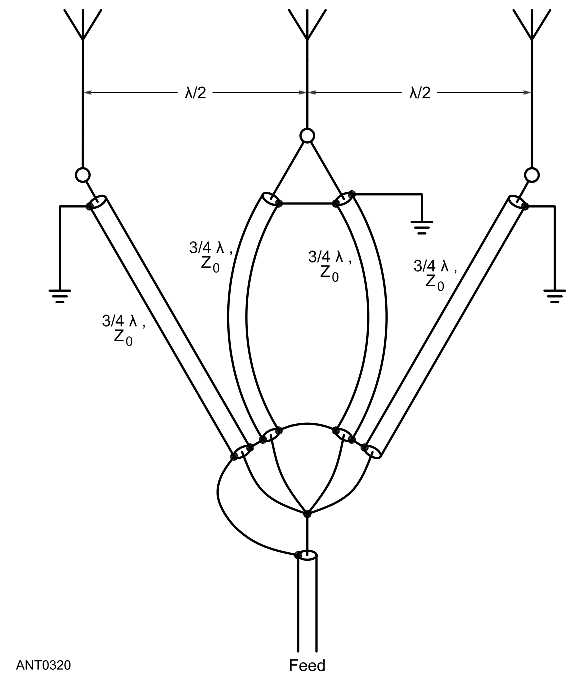

Fig 23—Feed system for the three element 1:2:1 binomial array. All feed lines are ¾ electrical wavelength long and have the same characteristic impedance.

A "Four Square" Array

Several types of feed system are used for feeding this popular array, and most share a common problem—they don't provide the correct element current ratio—although a number of them produce a workable approximation. The feed systems described here are capable of producing exactly the correct current ratio. The only significant variable is the element feed-point impedances, so the quality of the result depends on your ability to model the feed-point impedances of a correctly fed array. As in the examples above, *EZNEC-ARRL* will be used for that purpose and *Arrayfeed1* for the design of the feed system itself.

In this array (see **Fig 24**), four elements are placed in a square with λ/4 sides. (A variation of the Four Square uses wider spacing.) The rear and front elements (1 and 4) are 180° out-of-phase with each other. The side elements (2 and 3) are in phase with each other and 90° delayed from the front ele-

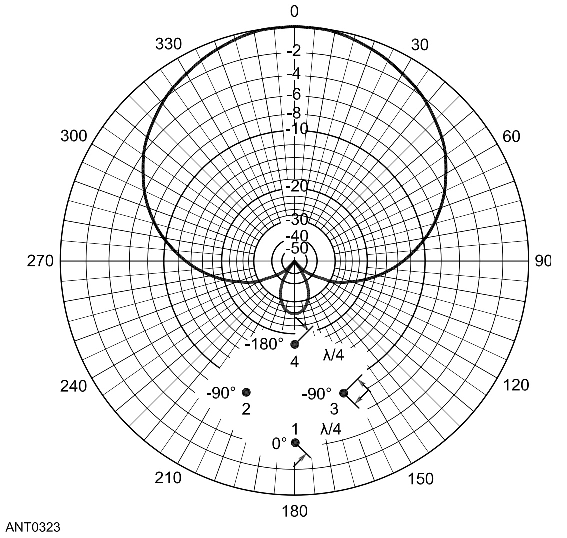

Fig 24—Pattern and layout of the four-element Four-Square array. Gain is referenced to a single similar element; add 5.5 dB to the scale values shown.

ment. The magnitudes of the currents in all four elements are equal. The front and rear elements can be forced to be 180° out-of-phase and to have equal currents by using the current-forcing method described earlier. One element is connected to a line that is either λ/4 or 3λ/4 long, the other to a line that is λ/2 longer, and the two lines to a common point.

Likewise, the two side element currents are forced to be equal by connecting them to a common point via λ/4 or 3λ/4 lines. **Fig 25** shows the basic current-forcing system.

If the pattern is to be electrically rotated, it is necessary to bring lines from all four elements to a common location. If solid-polyethylene dielectric coaxial cable, which has a dielectric constant of 0.66, is used, λ/4 lines won't reach the center of the array. So 3λ/4 lines must be used. Alternatively, you can use λ/4 lines with foam or other dielectric having a velocity factor of more than about 0.71 (plus a little extra margin). These will reach to the center. Whatever your choice, three of the lines must be the same length and the fourth must be λ/2 longer.

In this array, the side elements (2 and 3) have equal impedances, but the rear and front (1 and 4) are different from each other, and both are different from the side elements. We have to know the feed-point impedances of the front, rear and side elements in order to design the "simplest" feed system, but only the side element impedances are needed to design the L-network system. Knowledge of all feed-point impedances is necessary if the array main feedpoint impedance **Zin** is to be calculated. *EZNEC-ARRL* model **4Square_Example.EZ** shows a 40-meter Four Square array with 18 Ω of loss resistance at each element, to approximate an 8-radial per element ground system. (See the cardioid array example above for more information about modeling ground system loss.) Opening the file in *EZNEC-ARRL* and clicking the **Src Dat** button gives the following impedances:

Source 1: 16.4 – *j*15.85 Ω
Sources 2 and 3: 57.47 – *j*19.44 Ω
Source 4: 77.81 + *j*54.8 Ω

It's interesting to note that the resistive part of source 1 is less than the 18 Ω of loss resistance we intentionally added to simulate ground system loss. That means that the element 1 feed-point resistance would be negative if the ground resistance were less than about an ohm and a half. This isn't uncommon in phased arrays and simply means that the element is feeding power *into* the feed system. This power is coming via mutual coupling from the other elements.

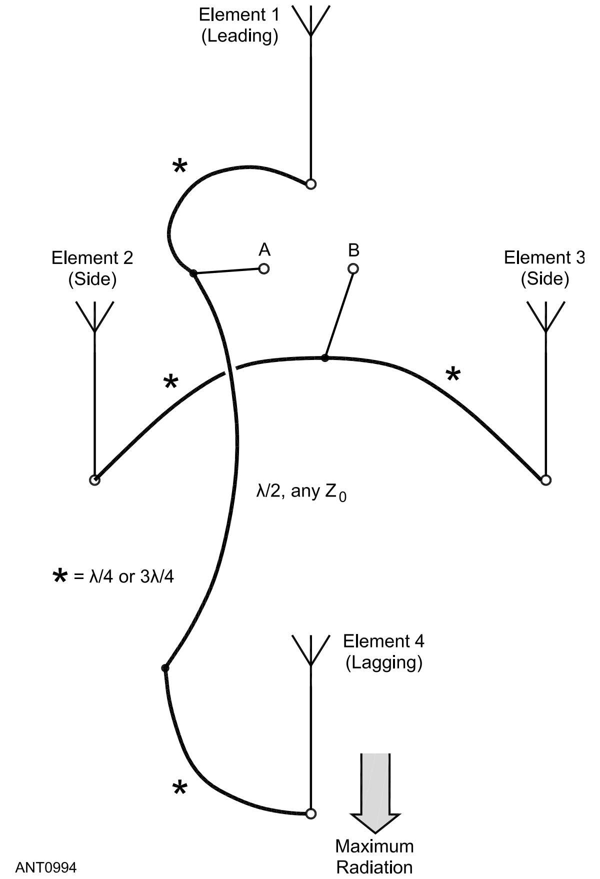

Fig 25—"Simplest" feed system for the Four-Square array in Fig 24. Grounds and cable shields have been omitted for clarity.

"Simplest" (Transmission Line Only) Feed System

To design a "simplest" feed system, start *Arrayfeed1*. In the **Array Type** frame, select **4 Square**, and select **"Simplest"** in the **Feed System Type** frame. In the **Inputs** frame, enter the frequency and the impedances from *EZNEC-ARRL*:

Frequency = 7.15 MHz
Leading Element: R = 16.4, X = –15.85
Side elements: R = 57.47, X = –19.44
Lagging Element: R = 77.81, X = 54.8

We'll try using 50 Ω for all lines, so enter **50** into the next three boxes.

Enter **1** for the lagging:leading I magnitude and **–90** for the phase.

Click **Find Solutions**.

The result is shown in the **Solutions** frame, shown in **Fig 26**. As always when any solution exists, there are two to choose from. The one with the shortest lines is generally preferable, so we'll choose it. For this example, we'll use λ/4 lines with velocity factor of 0.82. So enter **0.82** in the **Velocity Factor** box in the **Physical Lengths** frame, and read the physical lengths from the bottom of that frame. The λ/4 lines (marked in the *Arrayfeed1* diagram with an asterisk) are 28.2 feet, line 1 is 7.483 feet and line 2 is 51.668 feet. The "simplest" feed system is shown in Fig 26, and the complete feed system consists of this connected to the array of Fig 25.

EZNEC-ARRL model **ARRL_4Square_TL_Example.EZ** is simulates the array fed with this system. Comparison of the pattern plot to one from ideal-current model **ARRL_4Square_Example.EZ** and examination of

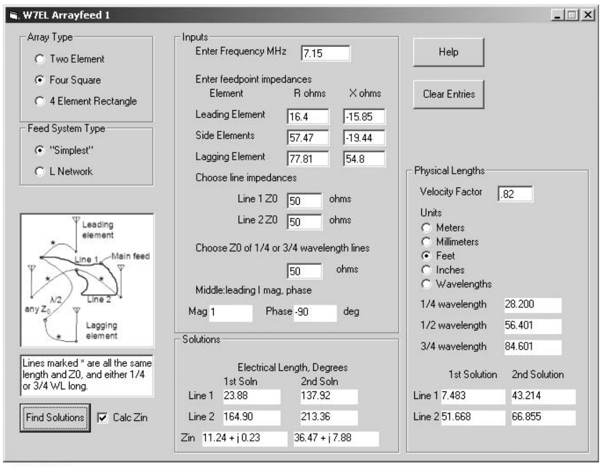

Array Type: ○ Two Element ● Four Square ○ 4 Element Rectangle

Feed System Type: ● "Simplest" ○ L Network

Inputs

Enter Frequency MHz 7.15

Enter feedpoint impedances

Element	R ohms	X ohms
Leading Element	16.4	-15.85
Side Elements	57.47	-19.44
Lagging Element	77.81	54.8

Choose line impedances

Line 1 Z0 50 ohms

Line 2 Z0 50 ohms

Choose Z0 of 1/4 or 3/4 wavelength lines 50 ohms

Middle:leading I mag, phase

Mag 1 Phase -90 deg

Solutions

Electrical Length, Degrees

	1st Soln	2nd Soln
Line 1	23.88	137.92
Line 2	164.90	213.36
Zin	11.24 + j 0.23	36.47 + j 7.88

Physical Lengths

Velocity Factor .82

Units: ○ Meters ○ Millimeters ● Feet ○ Inches ○ Wavelengths

1/4 wavelength 28.200

1/2 wavelength 56.401

3/4 wavelength 84.601

	1st Solution	2nd Solution
Line 1	7.483	43.214
Line 2	51.668	66.855

Lines marked * are all the same length and Z0, and either 1/4 or 3/4 WL long.

☑ Calc Zin

ANT0995

Fig 26—Screen capture from *Arrayfeed1* for "Simplest" feed system for Four Square feed system shown in Fig 25.

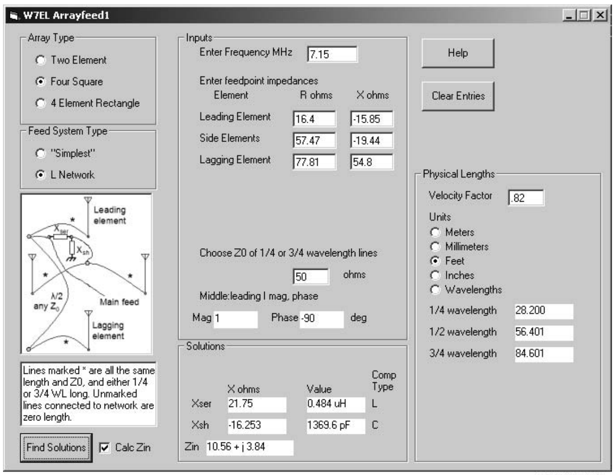

Array Type: ○ Two Element ● Four Square ○ 4 Element Rectangle

Feed System Type: ○ "Simplest" ● L Network

Inputs

Enter Frequency MHz 7.15

Enter feedpoint impedances

Element	R ohms	X ohms
Leading Element	16.4	-15.85
Side Elements	57.47	-19.44
Lagging Element	77.81	54.8

Choose Z0 of 1/4 or 3/4 wavelength lines 50 ohms

Middle:leading I mag, phase

Mag 1 Phase -90 deg

Solutions

	X ohms	Value	Comp Type
Xser	21.75	0.484 uH	L
Xsh	-16.253	1369.6 pF	C

Zin 10.56 + j 3.84

Physical Lengths

Velocity Factor .82

Units: ○ Meters ○ Millimeters ● Feet ○ Inches ○ Wavelengths

1/4 wavelength 28.200

1/2 wavelength 56.401

3/4 wavelength 84.601

Lines marked * are all the same length and Z0, and either 1/4 or 3/4 WL long. Unmarked lines connected to network are zero length.

☑ Calc Zin

ANT0134

Fig 27—L-network setup for Four-Square array in Fig 25, fed with λ/4 (or 3λ/4) current-forcing feed system.

the element currents verify that the feed system is producing the desired pattern and element currents. You can use **ARRL_4Square_Example.EZ** to investigate the effect of frequency change, ground loss and other changes on the array gain and pattern.

L-Network Feed System

To design the L-network feed system, simply change the **Feed System Type** to **L Network** and click **Find Solutions**. The results you should see are a 0.484 µH inductor for the series component X_{ser}, and a 1369.6 pF capacitor for the shunt component X_{sh}. The L-network feed system is shown in **Fig 27**, and the complete feed system consists of this L network connected to the array of Fig 25.

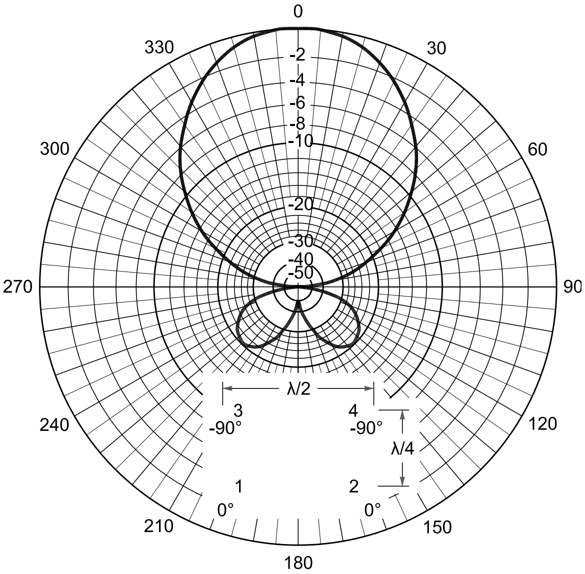

ANT0321

Fig 28—Pattern and layout of the four-element rectangular array. Gain is referenced to a single similar element; add 6.8 dB to the scale values shown.

EZNEC-ARRL doesn't have the ability to directly model the L network, so it's unable to model the complete system. However, the system has been modeled using the network capability of *EZNEC v.5* and found to work as designed. Arrays have also been built using this feed system and the element currents measured, with exactly the expected results.

This array is more sensitive to adjustment than the 2-element 90° fed, 90° spaced array. Adjustment procedures and a method of remotely switching the direction of this array are described in the **Practical Aspects of Phased Array Design** section that follows.

A 4-Element Rectangular Array

The 4-element rectangule array shown with its pattern in **Fig 28** has appeared numerous times in amateur publications. However, many of the accompanying feed systems fail to deliver currents in the proper amounts and phases to the various elements. The array can be correctly fed using the principles discussed in this chapter and the design methods that follow.

Elements 1 and 2 can be forced to be in-phase and to have equal currents by feeding them through 3λ/4 lines. (As in the binomial and Four Square array examples, 3λ/4 lines are chosen because λ/4 lines won't physically reach.) The currents in elements 3 and 4 can similarly be forced to be equal and in-phase. **Fig 29** shows the "current-forcing" feed system. Elements 3 and 4 are made to have currents of equal magnitude but of 90° phase difference from elements 1 and 2 by use of either a "simplest" all-transmission line feed system or an L-network feed system. Both will be designed in this example.

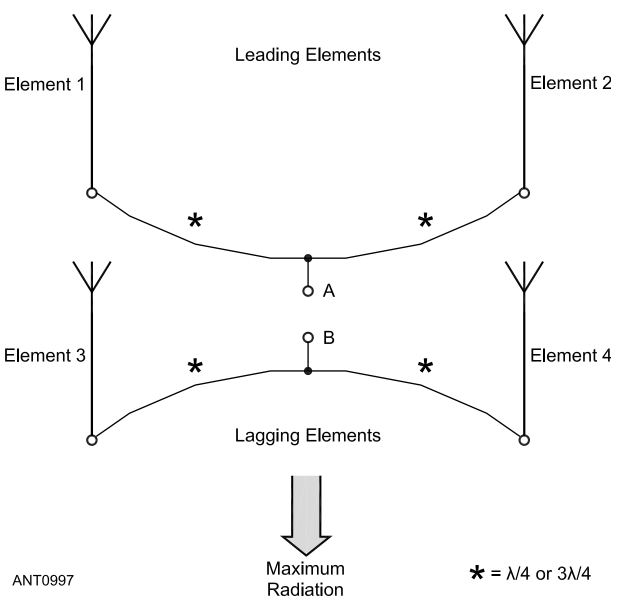

Fig 29—"Simplest" feed system for four-element rectangular array, using four equal-length λ/4 (or 3λ/4) cables.

For this array, we have to know the feed-point impedances of two elements (one of each pair) in order to design either type of feed system. *EZNEC-ARRL* model **Rectangular_Example.EZ** shows a 20-meter rectangular array with 18 Ω of loss resistance at each element, again to approximate an 8-radial per element ground system. (See the cardioid array example above for more information about modeling ground system loss.) Open the file in *EZNEC-ARRL* and click the **Src Dat** button to find the following feed-point impedances:

Sources 1 and 2: 21.44 – *j*21.29 Ω
Sources 3 and 4: 70.81 – *j*5.232 Ω

"Simplest" (Transmission- Line Only) Feed System

To design a "simplest" feed system, start program *Arrayfeed1*. In the **Array Type** frame, select **4 Element Rectangle**, and select **"Simplest"** in the **Feed System Type** frame. In the **Inputs** frame, enter the frequency and the impedances from *EZNEC-ARRL*:

Frequency = 14.15 MHz
Leading Elements R = 21.44, X = –21.29
Lagging Elements R = 70.81, X = –5.232

We'll use 50 Ω for all lines, so enter **50** into the next three boxes.

Enter **1** for the lagging:leading I magnitude and **–90** for the phase.

Click Find Solutions.

The result is "No Solution"—This combination of line impedances can't be used. Several other combinations also produce this result, but making lines 1 and 2 each 75 Ω and the 3λ/4 lines 50 Ω does produce a solution. Enter **75** into the **Line 1 Z0** and **Line 2 Z0** boxes, and leave **50** in the **Choose Z0 of ¼ or ¾ wavelength lines** box, then click the Find Solutions button. There won't be any problem making lines 1 and 2 reach, so we'll choose the first solution because the lines are shorter. The physical lengths of all the lines are shown in the **Physical Lengths** frame when the velocity factor is entered in the appropriate box. Assuming that we use coax with a velocity factor of 0.66 (and the example frequency of 14.15 MHz), the lengths are:

Line 1: 4.982 feet
Line 2: 20.153 feet
3λ/4 lines (marked with an asterisk in the *Arrayfeed1* diagram): 34.408 feet

The lines are connected following the diagram in the upper left part of the *Arrayfeed1* window. This completes the "simplest" feed system design. *EZNEC-ARRL* model **Rectangular_TL_Example.EZ** simulates an array fed with this system.

Comparison of the pattern plot to one from ideal-current **Rectangular_Example.EZ**, and examination of the element currents verify that the feed system is producing the desired pattern and element currents.

L-Network Feed System

To design the L-network feed system using *Arrayfeed1*, change the **Feed System Type** to **L Network** and click **Find Solutions**. The resulting L-network values are a 0.484 µH inductor for the series component X_{ser} and a 1369.6 pF capacitor for the shunt component X_{sh}.

120° Fed, 60° Spaced Dipole Array

This example shows the design of "simplest" and L network feed systems for a 2-element 20- meter dipole array, rather than a vertical array. No special accommodation is required for the array made from dipoles rather than vertical elements—the same methods can be used regardless of element shape. This example also shows that both the "simplest" and L-network feed systems can readily be applied to elements that use phase angles other than 90°.

Any 2-element array made with identical elements spaced λ/2 or closer and having equal magnitude currents with a relative phase angle of 180° minus the spacing will produce a unidirectional pattern with a good null to the rear. In practice, very close spacings lead to very low feed-point resistances, with consequent losses and very narrowband characteristics. But this 60° spaced array is well within the range of practical realization. File **ARRL_Dipole_Array_Example.EZ** is a model created for this array, with ideal element currents. Open this file in *EZNEC-ARRL* and click **FF Plot** to show the pattern at an elevation angle of 10°. You can save this pattern for later comparison to the pattern with a "simplest" feed system by opening the **File** menu in the **2D Plot** window, selecting **Save Trace As**, entering a name for the trace file and clicking **Save**.

Following the same procedure as in the previous examples, we begin the array design by finding the element feed-point impedances in the ideally fed array using *EZNEC-ARRL* numbers. Having already opened **ARRL_Dipole_Array_Example.EZ**, all that's needed is to click **Src Dat.** The results are:

Leading element (source 1) : 36.16 – *j*46.05 Ω
Lagging element (source 2) : 49.56 + *j*51.47 Ω

"Simplest" (Transmission Line Only) Feed System

Select **Two Element** for the **Array Type** in *Arrayfeed1* and **"Simplest"** for the **Feed System Type**. Enter the frequency of 14.15 MHz and enter the element feed-point impedances from *EZNEC-ARRL* into the appropriate boxes in the **Inputs** frame. For line impedances, the section describing the "simplest" feed system recommends against choosing one which is very different from the element feed-point impedances, but for fun let's try 300 Ω for the two lines and see what happens. Enter **300** in the **Line 1 Z0** and **Line 2 Z0** boxes. Finally, enter the **lagging:leading I mag, phase** of **1** for **Mag** and **–120** for **Phase**.

Click **Find Solutions**. For this example we'll assume that TV type twinlead with a velocity factor of 0.8 is being used. So enter **0.8** for the **Velocity Factor** and read the physical line lengths in the **Physical Lengths** frame. A model of the array using the first solution has been created as **ARRL_Dipole_Array_TL_Example.EZ**. Open this file in *EZNEC-ARRL* and click **FF Tab**. You should see that the plot is virtually identical to the one saved earlier from the ideal-current model. Note the gain and front/back ratio or 8.79 dBi and 31.01 dB respectively reported in the data box below the 2D plot.

Don't subtract 2.15 dB to find the gain relative to a single element! This isn't a free-space model, and the gain of a single dipole over ground is much greater than 2.15 dBi. Instead, delete one of the elements in **ARRL_Dipole_Array_Example.EZ** to find the gain of a single element and subtract that value from the array gain. You can use the undo feature or re-open the file to restore the array.

Now, go back to the model with the "simplest" feed system in *EZNEC-ARRL* and change the Frequency to **14.0** MHz. Click **FF Tab** again. The gain has decreased a little, to 8.54 dBi and the front/back ratio has also decreased, to 21.8 dB. At 14.3 MHz, the gain is slightly higher, 9.04 dBi, but the front/back is again worse, down to 18.64 dB. But this isn't bad overall.

Let's take a look at the second solution. Click the **Trans Lines** line in the main *EZNEC-ARRL* window to open the Transmission Lines Window. Change the length of the first line to **26.856** feet, the second to **28.356** feet, and press the **Enter** key to complete the change. Change the **Frequency** back to **14.15** MHz and click **FF Tab**. You should see exactly the same pattern as for both the first solution and for the ideal current model. But now change the **Frequency** to **14.0** MHz, click FF Tab, and look at the pattern.

What happened? The gain has dropped to 5.95 dBi and the front/back to only 3.1 dB. The array is now nearly bi-directional! It's almost as bad at 14.3 MHz. So we've created a terribly touchy system. The chance of its working correctly even at the design frequency is slim, because there are inevitably some differences between the model and real antenna.

We did have a clue this might happen. As stated in the section describing the "simplest" feed system, the best choices for line Z_0 and for the resulting solution give a difference in electrical line lengths about equal to the desired phase delay of the current. The difference in electrical line lengths for the first solution was about 152°—not as close to the 120° current phase difference as we'd like, but much better than the mere 9.7° difference of the lines for the second solution. While the 300-Ω line Z_0 is quite different from the element feed-point impedances, the first solution result is quite good. If desired, you can try other line impedance values into *Arrayfeed1* and evaluate the results with *EZNEC-ARRL*.

Please see the information about baluns in the **Baluns in Phased Arrays** section. Baluns are placed the same as in Fig 19, which shows the L-network feed system.

L-Network Feed System

To design an L network feed system, change the *Arrayfeed1* **Feed System Type** to **L Network** and click **Find Solutions**. The results aren't good ones to use. The component reactance magnitudes of about 1573 and 2619 Ω are more than five times the 300-Ω Z_0 of the feed lines. As explained in the section describing the L-network feed system, it's undesirable to have such a large ratio of component reactance to line Z_0. Among other problems, the inductor and capacitor values are quite extreme and capacitor stray inductance and inductor capacitance would have a significant impact on performance.

The problem occurs because the feed-line impedance we chose is much larger than the element feed-point impedances, so the λ/4 lines transform the feed-point impedances to much higher values at the L network and main feed point. This feed system would be extremely critical, narrowbanded and difficult to adjust. We can do better by choosing feed-line impedances that aren't too drastically different than the element feed-point impedances. In this case, 50 or 75 Ω would be a much better choice than 300. Let's try 75.

In *Arrayfeed1*, change the **Line 1 Z0** and **Line 2 Z0** impedances from **300** to **75** and click **Find Solutions**. L-network component reactance magnitudes are now about 98 and 164 Ω, much better than before. This will be a relatively uncritical and broadbanded feed system.

Again, be sure to read the information about baluns in the **Baluns in Phased Arrays** section. Fig 19 shows the completed feed system including baluns.

PRACTICAL ASPECTS OF PHASED ARRAY DESIGN

With almost any type of antenna system, there is much that can be learned from experimenting with, testing and using various array configurations. In this section, Roy Lewallen, W7EL, shares the benefit of years of his experience from actually building, adjusting and using phased arrays. There is much more work to be done in most of the areas covered here, and Roy encourages the reader to build on this work.

Adjusting Phased Array Feed Systems

If a phased array is constructed only to achieve forward gain, adjusting it is seldom worthwhile. This is because the forward gain of most arrays is quite insensitive to either the magnitude or phase of the relative currents flowing in the elements. If, however, good rejection of unwanted signals is desired, adjustment may be required. And achieving very deep nulls will almost surely require some adjustment.

The in-phase and 180° out-of-phase current-forcing method supplies very well-balanced and well-phased currents to elements without adjustment. If the pattern of an array fed using this method is unsatisfactory, it's generally the result of environmental differences—where the elements, even though furnished with correct currents, aren't generating the correct fields. Such an array can be optimized in a single direction, but a more general approach than the current-forcing method must be taken. Some possibilities are described by Paul Lee and Forrest Gehrke (see Bibliography).

Unlike the current-forcing method, the "simplest" and L-network feed systems described earlier in this chapter are dependent on the self and mutual impedance of one or more elements. The required transmission-line lengths or L-network component values can be computed to a high level of precision, but the results are only as good as the knowledge of the relevant feed-point impedances.

While the simplest feed system doesn't readily lend itself to adjustment, the components of an L network can easily be made adjustable or can be experimentally changed in increments. A practical approach is to model the array as accurately as possible, design and build the feed system based on the model results and then adjust the network for the best performance.

Simple arrays such as the two-element 90° fed and spaced array can be adjusted as follows. Place a low-power signal source at a distance from the array (preferably several wavelengths), in the direction a null should be. While listening to the signal on a receiver connected to the array, alternately adjust the two L-network components for the best rejection of the signal.

This has proved to be a very good way to adjust 2-element arrays. However, variable results were obtained when a Four-Square array was adjusted using this technique. The probable reason is that more than one combination of current balance and phasing can produce a null in a given direction but each produces a different overall pattern. So a different method must be used for adjusting more complex arrays. This involves actually measuring the element currents in some way, and adjusting the network until the currents are correct. After adjusting the currents, small adjustments can be made to deepen the null(s) if desired.

Measuring Element Currents

You can measure the element currents two ways. One way is to measure them directly at the element feed points, as shown in **Fig 30**. A dual-channel oscilloscope is required to monitor the currents. This method is the most accurate and it provides a direct indication of the actual relative magnitudes and phases of the element currents. The current probe is shown in **Fig 31**.

Instead of measuring the element currents directly, you could measure them indirectly by measuring the voltages on the feed lines an electrical λ/4 or 3λ/4 distance from the array. The voltages at these points are directly proportional to the element currents. This introduces additional variables that can reduce the accuracy of the result, but the method generally produces adequate performance. The 2-element arrays fed with the L-network system and all the four element arrays presented earlier have λ/4 or 3λ/4 lines from all elements to a common location, making this second measurement method convenient. The voltages can be observed with a dual-channel oscilloscope, or, to adjust for equal-magnitude currents and 90° phasing, you can use the test circuit shown in **Fig 32**.

The test circuit is connected to the feed lines of two elements that are to be adjusted for 90° phasing (such as elements 1 and 2, or 2 and 4 of the Four-Square array of Figs 24 and 25). Adjust the L-network components alternately until both meters read zero. Proper operation of the test circuit can be verified by disconnecting one of the inputs. The *phase* output

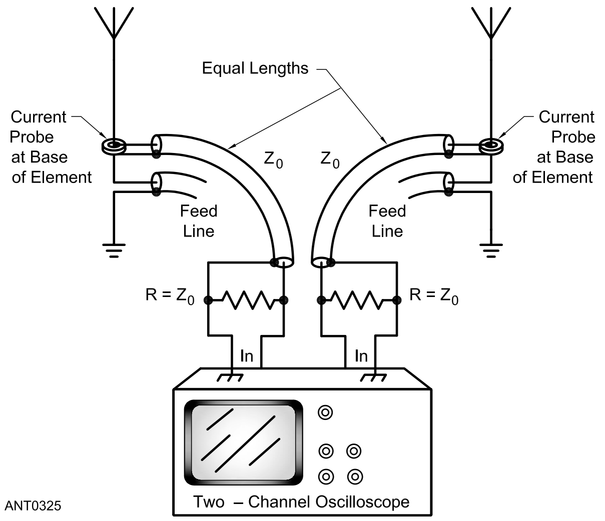
Equal Lengths
Current Probe at Base of Element
Current Probe at Base of Element
Z_0
Z_0
Feed Line
Feed Line
$R = Z_0$
$R = Z_0$
In
In
Two – Channel Oscilloscope
ANT0325

Fig 30—One method of measuring element currents in a phased array. Details of the current probe are given in Fig 31. Caution: Do not run high power to the antenna system for this measurement, or damage to the test equipment may result.

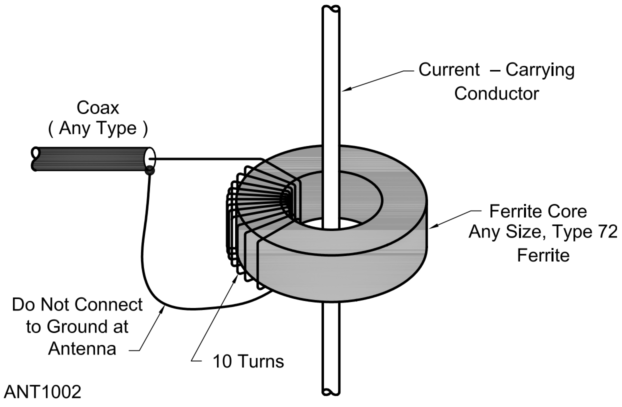
Current – Carrying Conductor
Coax (Any Type)
Ferrite Core Any Size, Type 72 Ferrite
Do Not Connect to Ground at Antenna
10 Turns
ANT1002

Fig 31—The current probe for use in the test setup of Fig 30. The ferrite core is of type 72 material, and may be any size. The coax line must be terminated at the opposite end with a resistor equal to its characteristic impedance. You should build this probe in a plastic or metal box to provide mechanical ruggedness.

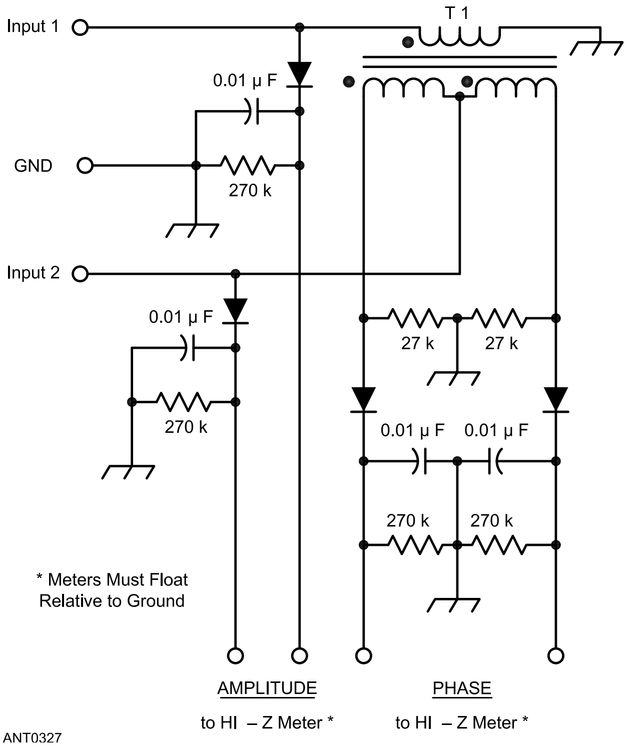

Fig 32—Quadrature test circuit. All diodes are germanium, such as 1N34A, 1N270, or equiv. Hot carrier or silicon diodes can be used at higher power levels. All resistors are ¼ or ½ W, 5% tolerance. Capacitors are ceramic. Alligator clips are convenient for making the input and ground connections to the array.

T1—7 trifilar turns on an Amidon FT-37-43, -75, -77, or equivalent ferrite toroid core.

should remain close to zero. If not, there is an undesirable imbalance in the circuit, which must be corrected. Another means of verification is to first adjust the L network so the tester indicates correct phasing (zero volts at the *phase* output). Then reverse the tester input connections to the elements. The *phase* output should remain close to zero.

Directional Switching of Arrays

One ideal directional-switching method would take the entire feed system, including the lines to the elements and physically rotate it. The smallest possible increment of rotation would depend on the symmetry of the array—the feed system would need to rotate until the array again *looks* the same to it. For example, any 2-element array can be rotated 180° (although that wouldn't accomplish anything if the array is bidirectional to begin with). The 4-element rectangular array of Figs 28 and 29 can also be reversed, and the Four-Square array of Figs 24 and 25 can be switched in 90° increments.

Smaller switching increments can be accomplished only by reconfiguring the feed system, including any network if used, effectively creating a different kind of array. Switching in smaller increments than dictated by symmetry will create a different pattern in some directions than in others, and must be thoughtfully done to maintain equal and properly phased element currents. The methods illustrated here will deal only with switching in increments related to the array symmetry, except for one: a 2-element broadside/end-fire array.

In all arrays, the success of directional switching depends on the elements and ground systems being identical so that equal element currents result in equal fields. It's even more important in arrays fed with any method other than current forcing, because the effectiveness of those methods depends on the element feed-point impedances. Few of us can afford the luxury of having an array many wavelengths away from all other conductors, so an array will nearly always perform somewhat differently in each direction. The array should be adjusted when steered in the direction requiring the most signal rejection in the nulls. Forward gain will, for all practical purposes, be equal in all the switched directions, since gain is much more tolerant of error than are nulls.

Basic Switching Methods

Following is a discussion of basic switching methods, how to power relays through the main feed line and other practical considerations. In diagrams, grounds are frequently omitted to aid clarity, but connections of the ground conductors must be carefully made. In fact, it is recommended that the ground conductors be switched just as the center conductors are, as explained in more detail in **Improving Array Switching Systems** below. In all cases, interconnecting lines must be very short.

A pair of elements spaced λ/2 apart can readily be switched between broadside and end-fire bidirectional pat-

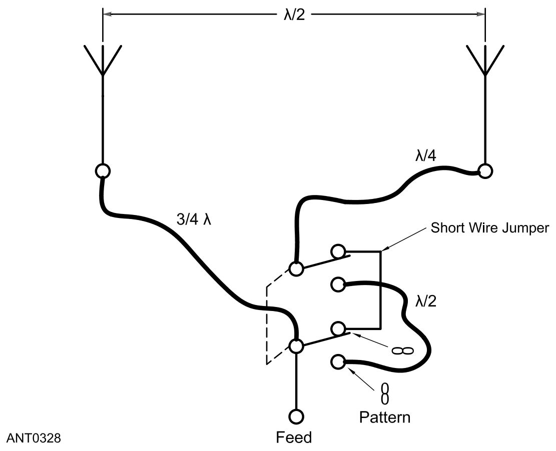

Fig 33—Two-element broadside/end-fire switching. All lines must have the same characteristic impedance. Grounds and cable shields have been omitted for clarity.

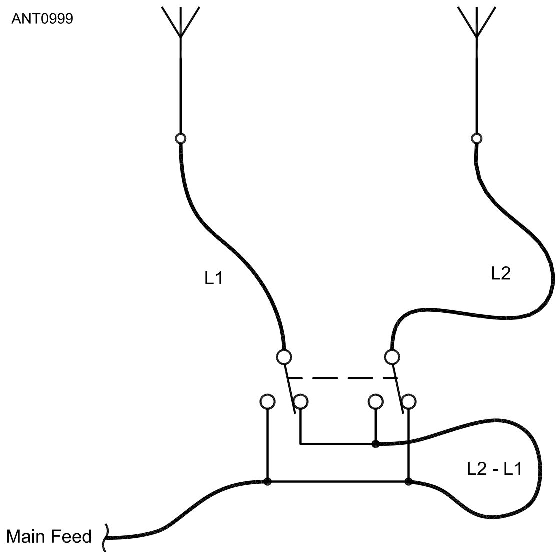

Fig 34—Directional switching for 90°, 90° spaced 2-element array fed with a "simplest" feed system.

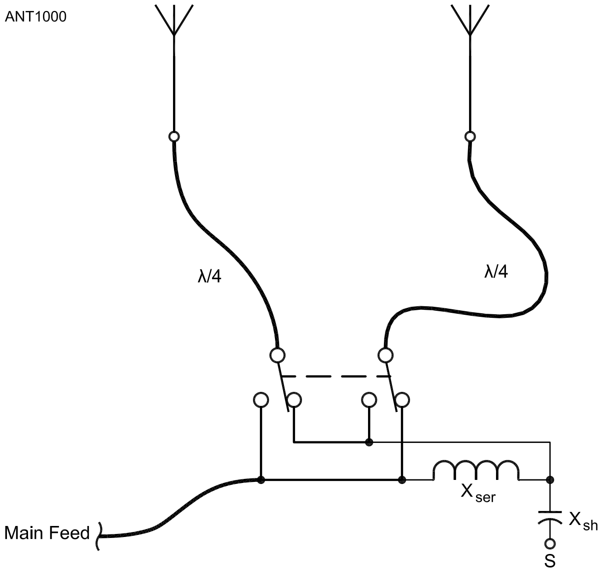

Fig 35—Directional switching for 90°, 90° spaced 2-element array fed with an L-network, current-forcing feed system.

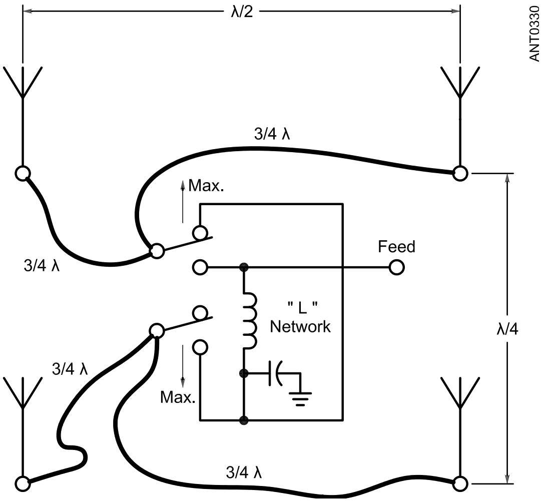

Fig 36—Directional switching of a four-element rectangular array. All interconnections must be very short. As usual, grounds and cable shields have been omitted for clarity.

terns, using the current-forcing properties of λ/4 lines. The method is shown in **Fig 33**. The switching device can be a relay powered via a separate cable or by dc sent along the main feed line.

Fig 34 shows directional switching of a 90° fed, 90° spaced array fed with a "simplest" feed system, where L1 and L2 are the required lengths of the two feed lines. **Fig 35** shows how to switch the same array when fed with an L-network, current-forcing system.

The rectangular array of Fig 28 can be switched in a similar manner, as shown in **Fig 36**. To switch a "simplest" fed rectangular array, use the switching circuit of Fig 34, but connect the two equal length lines to points A and B of Fig 29 in place of the two elements shown in Fig 34.

Switching the direction of an array in increments of 90°, when permitted by symmetry, requires at least two relays. A method of 90° switching of the Four-Square array with L-network feed is shown in **Fig 37**.

Powering Relays Through Feed Lines

All of the above switching methods can be implemented without additional wires to the switch box. A single-relay system is shown in **Fig 38A**, and a two-relay system in

Fig 38B. Small 12 or 24-V dc power relays can be used in either system at power levels up to at least a few hundred watts. Do not attempt to change directions while transmitting, however. Blocking capacitors C1 and C2 should be good quality ceramic or transmitting mica units of 0.01 to 0.1 µF. No problems have been encountered using 0.1 µF, 300-V monolithic ceramic units at RF output levels up to 300 W. C2 may be omitted if the antenna system is an open circuit at dc. C3 and C4 should be ceramic, 0.001 µF or larger.

In Fig 38B, capacitors C5 through C8 should be selected with the ratings of their counterparts in Fig 38A, as given above. Electrolytic capacitors across the relay coils, C9 and C10 in Fig 38B, should be large enough to prevent the relays from buzzing, but not so large as to make relay operation too slow. Final values for most relays will be in the range from 10 to 100 µF. They should have a voltage rating of at least double the relay coil voltage. Some relays do not require this capacitor. All diodes are 1N4001 or similar. A rotary switch may be used in place of the two toggle switches in the two-relay system to switch the relays in the desired sequence.

Improving Array-Switching Systems

The extra circuitry involved in switching arrays can degrade array performance by altering the relative currents fed to each element. One common cause is current sharing in common ground conductors, even when connections are kept very short. The author has seen a 30° phase shift in voltage along a 4-inch piece of #12 wire in a 40-meter array feed system.

When the two conductors of a feedline are physically separated from each other, the impedance increases. This is especially true when the main lines are coaxial cables. If currents from two elements share the ground conductor of a split line, a relatively large voltage drop results. Voltage changes λ/4 from the elements translate to current changes at the elements. Although keeping all leads extremely short is sometimes adequate, the best way to reduce current sharing problems is to keep the two conductors of each transmission line as close together as possible, and switch both conductors of each line rather than just a single or "hot" conductor.

An example of a carefully designed switching system is shown in **Fig 39**. It avoids the problem of shared ground conductor currents, as well as another common problem, namely that effective line lengths are often different along different switching paths. Notice how the path from the main feed point travels through a single line to each element with no common ground connections to other lines except at the

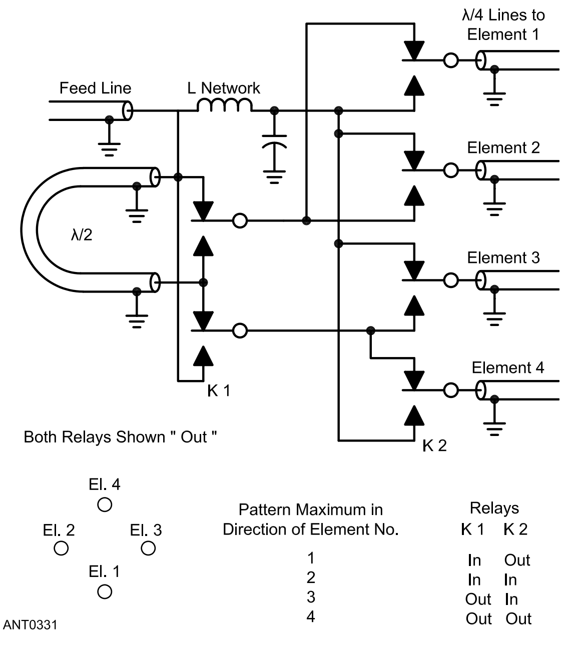
λ/4 Lines to
Element 1
Feed Line
L Network
Element 2
λ/2
Element 3
Element 4
K 1
K 2
Both Relays Shown " Out "
El. 4
El. 2
El. 3
El. 1
Pattern Maximum in Direction of Element No.
Relays
K 1 K 2
1 In Out
2 In In
3 Out In
4 Out Out
ANT0331

Fig 37—Directional switching of the four-square array. All interconnections must be very short.

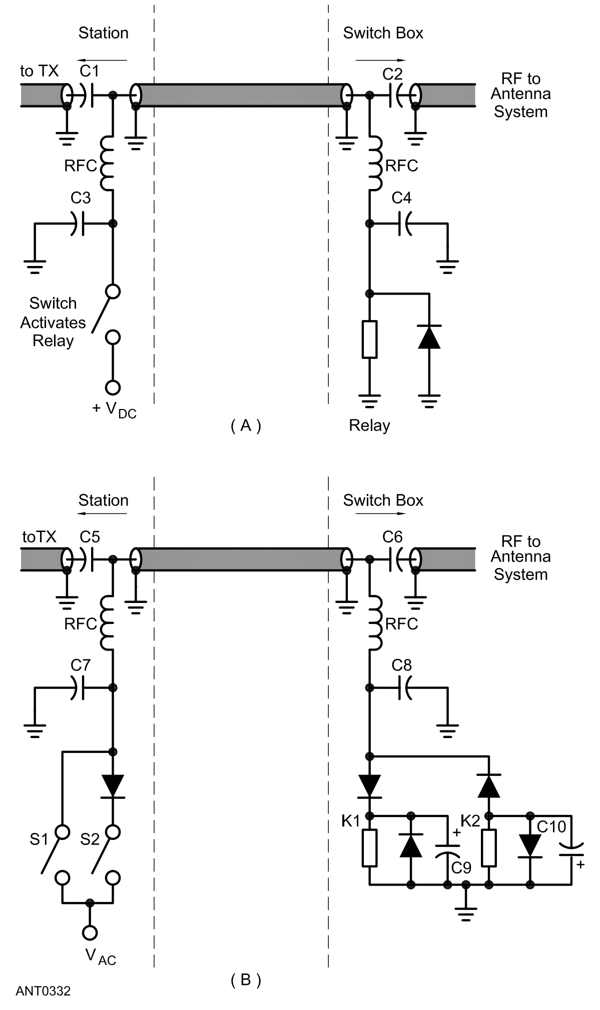
Station
Switch Box
to TX
C1
C2
RF to Antenna System
RFC
RFC
C3
C4
Switch Activates Relay
+ V_{DC}
(A)
Relay
Station
Switch Box
toTX
C5
C6
RF to Antenna System
RFC
RFC
C7
C8
S1
S2
K1
K2
C10
C9
V_{AC}
(B)
ANT0332

Fig 38—Remote switching of relays. See text for component information. A one-relay system is shown at A, and a two-relay system at B. In B, S1 activates K1, and S2 activates K2.

main feed point. Notice also that the distance doesn't change as the direction is switched. The λ/4 lines going to the two elements must be shortened by the length ℓ of the lines on the feed side of the relays, so that the total line length from the main feed point to each element is λ/4 (or 3λ/4).

You can see that in either relay position, there's an open ended stub of length ℓ connected at the main feed point and another at the output end of the L network. These will add capacitance at those points. Extra C at the main feed point will alter the overall impedance seen by a transmitter, but won't otherwise have any effect on the array or its performance. The one at the output of the L network will, however, change the transformation and phase shift properties of the network. But it's easy to compensate—the value of the shunt capacitor element is simply reduced by the amount of the C added by the stub. The amount of C for any kind of transmission line can be calculated from:

$$C(\text{pf / ft}) = \frac{1017}{Z_0 \text{VF}}$$

or

$$C(\text{pf / m}) = \frac{3336}{Z_0 \text{VF}}$$

where Z_0 = the characteristic impedance of the line and VF = the velocity factor. This works out to 31 pF/foot or 101 pF/meter for 50-Ω solid polyethylene insulated coax, which has a velocity factor of 0.66.

The general principles illustrated in Fig 39 can be extended to other switching systems. If switching the ground conductors as described above isn't practical, use of a metal box for the switching circuitry is recommended, so that the relatively large surface area of the box can be used for the common ground conductors, minimizing their inductance. Always keep leads extremely short.

Measuring the Electrical Length of Feed Lines

When using the feed methods described earlier the feed lines must be very close to the correct length. For best results, they should be correct within 1% or so. This means that a line that is intended to be, say, λ/4 at 7 MHz, should actually be λ/4 at some frequency within 70 kHz of 7 MHz. A simple but accurate method to determine at what frequency a line is λ/4 or λ/2 is shown in **Fig 40A**. The far end of the line is short circuited with a very short connection. A signal is applied to the input and the frequency is swept until the impedance at the input is a minimum. This is the frequency at which the

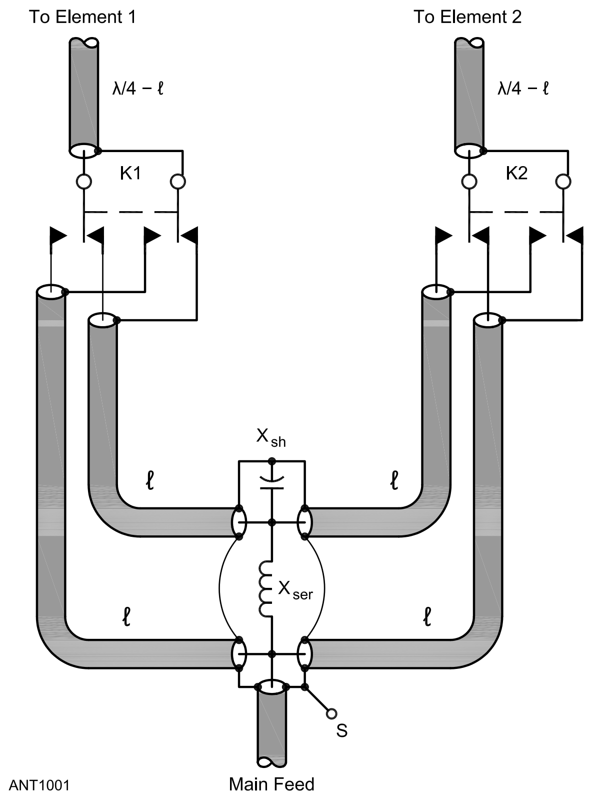

Fig 39—A carefully designed L-network, current-forcing switching system that switches both hot and shield conductors in feed coaxes.

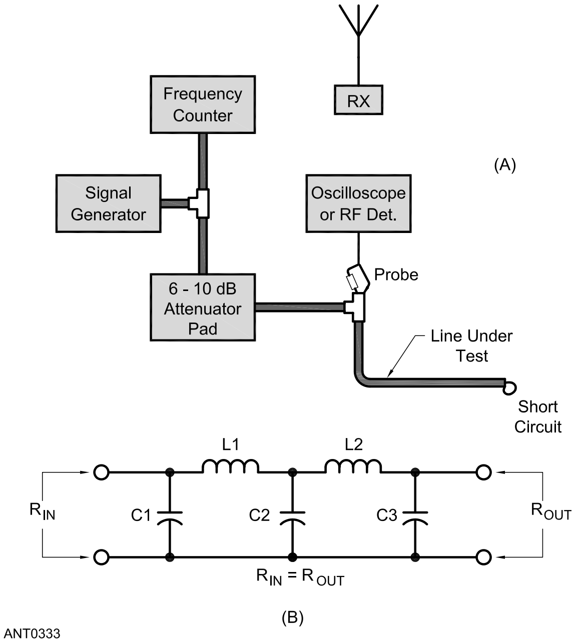

Fig 40—At A, the setup for measurement of the electrical length of a transmission line. The receiver may be used in place of the frequency counter to determine the frequency of the signal generator. The signal generator output must be free of harmonics; the half-wave harmonic filter at B may be used outboard if there is any doubt. It must be constructed for the frequency band of operation. Connect the filter between the signal generator and the attenuator pad.

C1, C3—Value to have a capacitive reactance = R_{IN}.
C2—Value to have a capacitive reactance = ½ R_{IN}.
L1, L2—Value to have an inductive reactance = R_{IN}.

line is λ/2. Either the frequency counter or the receiver may be used to determine this frequency. The line is, of course, λ/4 at one half the measured frequency.

The detector can be a simple diode detector or an oscilloscope may be used if available. A 6 to 10 dB attenuator pad is included to prevent the signal generator from looking into a short circuit at the measurement frequency. The signal generator output must be free of harmonics. If there is any doubt, an outboard low-pass filter, such as a half-wave harmonic filter, should be used. The half-wave filter circuit is shown in Fig 40B, and must be constructed for the frequency band of operation.

Another satisfactory method is to use a noise or resistance bridge or antenna analyzer at the input of the line, again looking for a low impedance at the input while the output is short circuited. Simple resistance bridges are described in Chapter 27.

Dip oscillators have been found to be unsatisfactory. The required coupling loop has too great an effect on measurements.

Measuring Element Self and Mutual Impedances

The need for measuring element self and mutual impedances has been made largely unnecessary with the ready availability of modeling software. Few amateurs appreciate the considerable difficulty of making accurate impedance measurements and accurate mutual impedance measurements are very difficult even with professional test equipment and skills. Despite the limitations of computer modeling, results very often are better than measured values because of the multiple factors affecting measurement accuracy.

Those who are interested in measuring self and mutual impedances can find more detailed information about doing so in Appendix B. The information there is from earlier editions of *The ARRL Antenna Book*.

Broadside Arrays

Broadside arrays can be made up of collinear or parallel elements or combinations of the two. This section was contributed by Rudy Severns, N6LF.

COLLINEAR ARRAYS

Collinear arrays are always operated with the elements in-phase. (If alternate elements in such an array are out-of-phase, the system simply becomes a harmonic type of antenna.) A collinear array is a broadside radiator, the direction of maximum radiation being at right angles to the line of the antenna.

POWER GAIN

Because of the nature of the mutual impedance between collinear elements, the feed-point resistance (compared to a single element, which is ≈73 Ω) is increased as shown earlier in this chapter (Fig 9). For this reason the power gain does not increase in direct proportion to the number of elements. The gain with two elements, as the spacing between them is varied, is shown by **Fig 41**. Although the gain is greatest when the end-to-end spacing is in the region of 0.4 to 0.6 λ, the use of spacings of this order is inconvenient constructionally and introduces problems in feeding the two elements. As a result, collinear elements are almost always operated with their ends

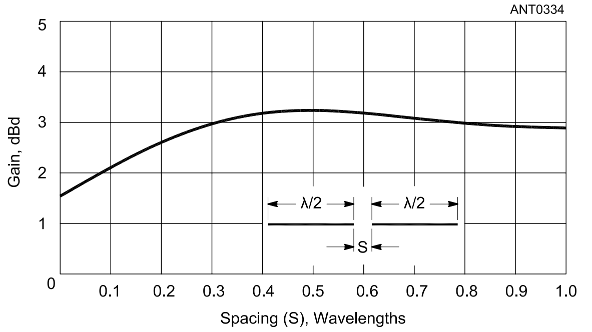

Fig 41—Gain of two collinear λ/2 elements as a function of spacing between the adjacent ends.

quite close together—in wire antennas, usually with just a strain insulator between.

With very small spacing between the ends of adjacent elements the theoretical power gain of collinear arrays, assuming the use of #12 copper wire, is approximately as follows over a dipole in free space:

2 collinear elements—1.6 dB
3 collinear elements—3.1 dB
4 collinear elements—3.9 dB

More than four elements are rarely used.

DIRECTIVITY

The directivity of a collinear array, in a plane containing the axis of the array, increases with its length. Small secondary lobes appear in the pattern when more than two elements are used, but the amplitudes of these lobes are low enough so that they are usually not important. In a plane at right angles to the array the directive diagram is a circle, no matter what the number of elements. Collinear operation, therefore, affects only E-plane directivity, the plane containing the antenna.

When a collinear array is mounted with the elements vertical, the antenna radiates equally well in all geographical directions. An array of such *stacked* collinear elements tends to confine the radiation to low vertical angles.

If a collinear array is mounted horizontally, the directive pattern in the vertical plane at right angles to the array is the same as the vertical pattern of a simple λ/2 antenna at the same height (Chapter 3).

TWO-ELEMENT ARRAYS

The simplest and most popular collinear array is one using two elements, as shown in **Fig 42**. This system is commonly known as *two half-waves in phase*. The directive pattern in a plane containing the wire axis is shown in **Fig 43**, which shows superimposed patterns for a dipole and 2, 3 and 4-element collinear arrays. Depending on the conductor size, height, and similar factors, the impedance at the feed point can be expected to be in the range of 4 to 6 kΩ, for wire antennas. If the elements are made of tubing having a low λ/dia (wavelength to diameter) ratio, values as low as 1 kΩ are representative. The system can be fed through an open-wire tuned line with negligible loss for ordinary line lengths, or a matching section may be used if desired.

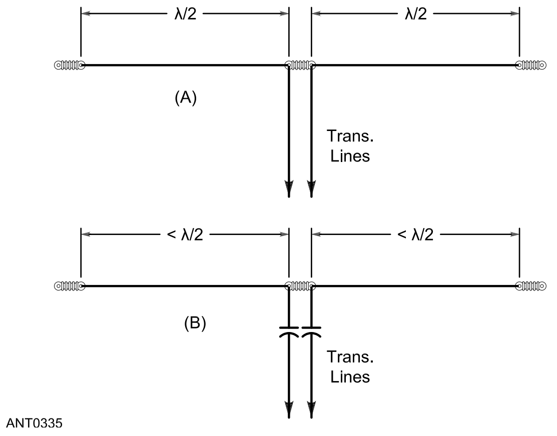

Fig 42—At A, two-element collinear array (two half-waves in phase). The transmission line shown would operate as a tuned line. A matching section can be substituted and a nonresonant line used if desired, as shown at B, where the matching section is two series capacitors.

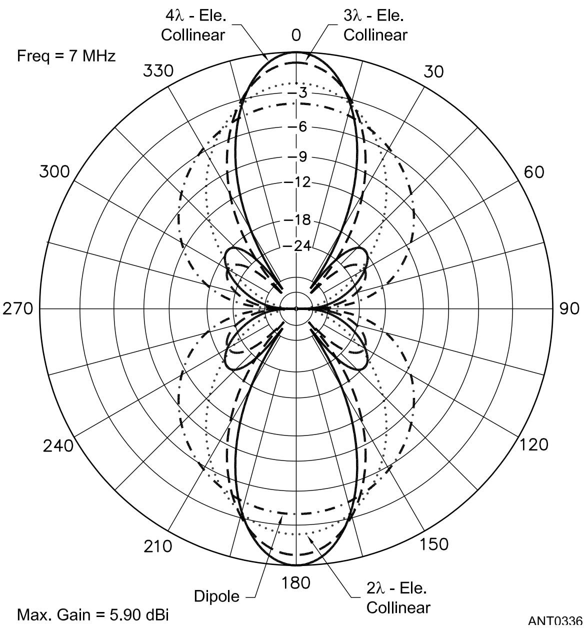

Fig 43—Free-space E-plane directive diagram for dipole, 2, 3 and 4-element collinear arrays. The solid line is a 4-element collinear; the dashed line is for a 3-element collinear; the dotted line is for a 2-element collinear and the dashed-dotted line is for a λ/2 dipole.

A number of arrangements for matching the feed line to this antenna are described in Chapter 26. If elements somewhat shorter than λ/2 are used, then additional matching schemes can be employed at the expense of a slight reduction in gain. When the elements are shortened two things happen—the impedance at the feed-point drops and the impedance has inductive reactance that can be tuned out with simple series capacitors, as shown in Fig 42B.

Note that these capacitors must be suitable for the power level. Small *doorknob* capacitors such as those frequently used in power amplifiers, are suitable. By way of an example, if each side of a 40-meter 2-element array is shortened from 67 to 58 feet, the feed-point impedance drops from nearly 6000 Ω to about 1012 Ω with an inductive reactance of 1800 Ω. The reactance can be tuned out by inserting 25 pF capacitors at the feed-point. The 1012 Ω resistance can be transformed to 200 Ω using a λ/4 matching section made of 450-Ω ladder line and then transformed to 50 Ω with a 4:1 balun. Shortening the array as suggested reduces the gain by about 0.5 dB.

Another scheme that preserves the gain is to use a 450-Ω

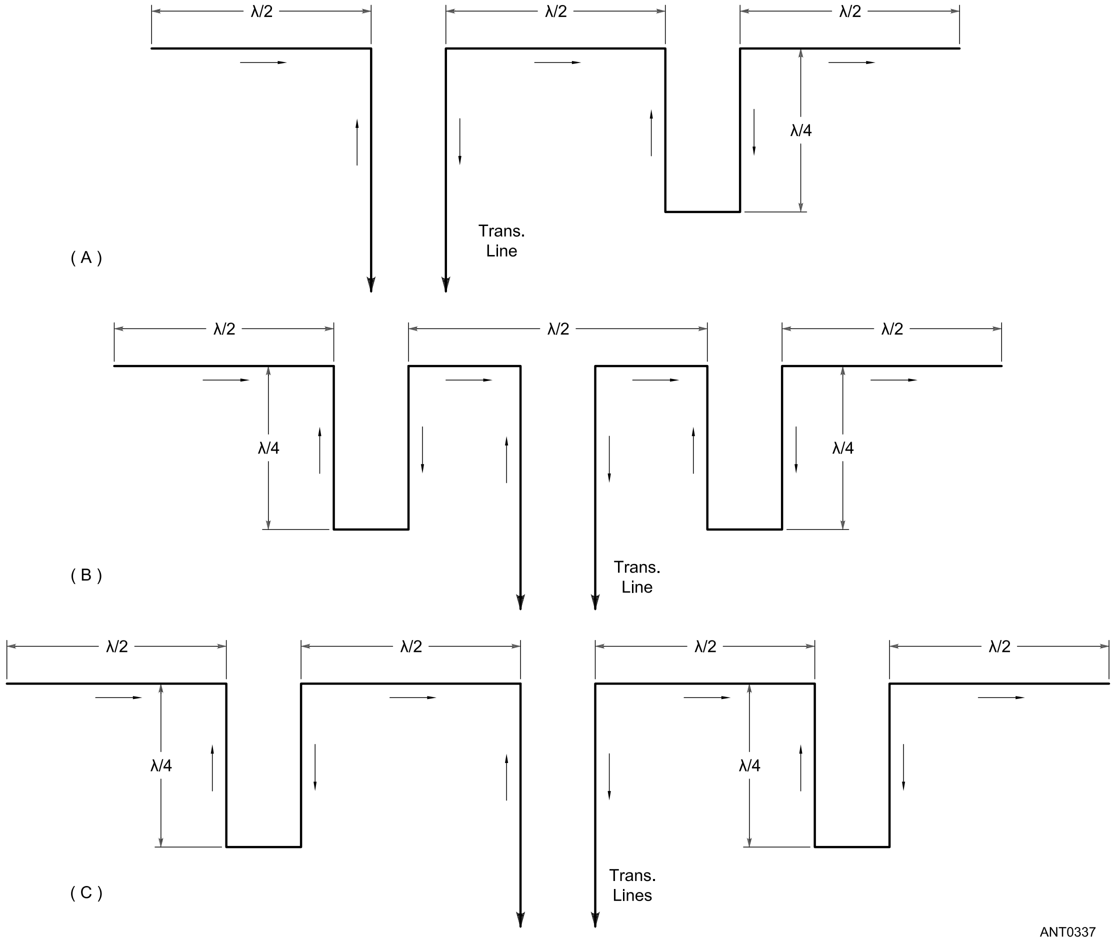

Fig 44—Layouts for 3- and 4-element collinear arrays. Alternative methods of feeding a 3-element array are shown at A and B. These drawings also show the current distribution on the antenna elements and phasing stubs. A matched transmission line can be substituted for the tuned line by using a suitable matching section.

λ/4 matching section and shorten the antenna only slightly to have a resistance of 4 kΩ. The impedance at the input of the matching section is then near 50 Ω and a simple 1:1 balun can be used. Many other schemes are possible. The free-space E-plane response for a 2-element collinear array is shown in Fig 43, compared with the responses for more elaborate collinear arrays described below.

THREE- AND FOUR-ELEMENT ARRAYS

In a long wire the direction of current flow reverses in each λ/2 section. Consequently, collinear elements cannot simply be connected end to end; there must be some means for making the current flow in the same direction in all elements. When more than two collinear elements are used it is necessary to connect *phasing* stubs between adjacent elements in order to bring the currents in all elements in-phase. In **Fig 44A** the direction of current flow is correct in the two left-hand elements because the shorted λ/4 transmission line (*stub*) is connected between them. This stub may be looked upon simply as the alternate λ/2 section of a long-wire antenna folded back on itself to cancel its radiation. In Fig 44A the part to the right of the transmission line has a total length of three half wavelengths, the center half wave being folded back to form a λ/4 phase-reversing stub. No data are available on the impedance at the feed point in this arrangement, but various considerations indicate that it should be over 1 kΩ.

An alternative method of feeding three collinear elements is shown in Fig 44B. In this case power is applied at the center of the middle element and phase-reversing stubs are used between this element and both of the outer elements. The impedance at the feed point in this case is somewhat over 300 Ω and provides a close match to 300 Ω line. The SWR will be less than 2:1 when 600-Ω line is used. Center feed of this type is somewhat preferable to the arrangement in Fig 44A because the system as a whole is balanced. This assures more uniform power distribution among the elements. In Fig 44A, the right-hand element is likely to receive somewhat less power than the other two because a portion of the input power is radiated by the middle element before it can reach the element located at the extreme right.

A four-element array is shown in Fig 44C. The system is symmetrical when fed between the two center elements as shown. As in the three-element case, no data are available on the impedance at the feed point. However, the SWR with a 600 Ω line should not be much over 2:1.

Fig 43 compares the directive patterns of 2, 3 and 4-element arrays. Collinear arrays can be extended to more than four elements. However, the simple 2-element collinear array is the type most frequently used, as it lends itself well to multi-band operation. More than two collinear elements are seldom used because more gain can be obtained from other types of arrays.

ADJUSTMENT

In any of the collinear systems described, the lengths of the radiating elements in feet can be found from the formula $468/f_{MHz}$. The lengths of the phasing stubs can be found from the equations given in Chapter 26 for the type of line used. If the stub is open-wire line (500 to 600 Ω impedance) you may assume a velocity factor of 0.975 in the formula for a λ/4 line. On-site adjustment is, in general, an unnecessary refinement. If desired, however, the following procedure may be used when the system has more than two elements.

Disconnect all stubs and all elements except those directly connected to the transmission line (in the case of feed such as is shown in Fig 44B leave only the center element connected to the line). Adjust the elements to resonance, using the still-connected element. When the proper length is determined, cut all other elements to the same length. Make the phasing stubs slightly long and use a shorting bar to adjust their length. Connect the elements to the stubs and adjust the stubs to resonance, as indicated by maximum current in the shorting bars or by the SWR on the transmission line. If more than three or four elements are used it is best to add elements two at a time (one at each end of the array), resonating the system each time before a new pair is added.

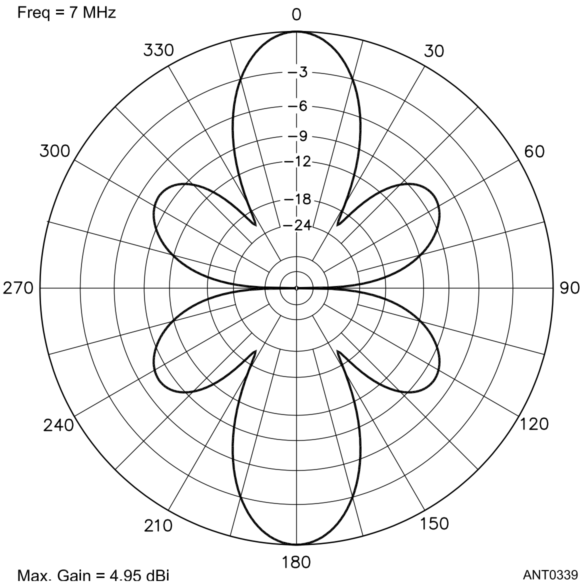

Fig 46—E-plane pattern for the extended double Zepp of Fig 45. This is also the horizontal directional pattern when the elements are horizontal. The axis of the elements lies along the 90°-270° line. The free-space array gain is approximately 4.95 dBi.

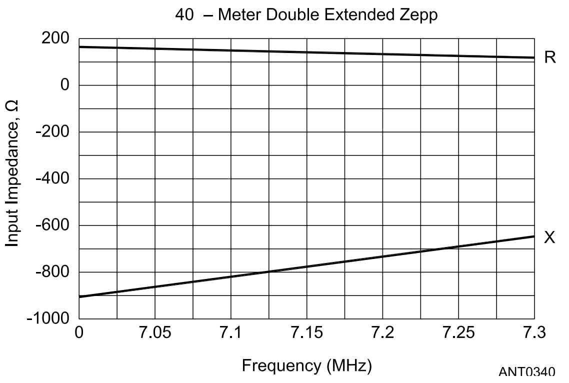

Fig 47—Resistive and reactive feed-point impedance of a 40-meter extended double Zepp in free space.

THE EXTENDED DOUBLE ZEPP

One method to obtain higher gain that goes with wider spacing in a simple system of two collinear elements is to make the elements somewhat longer than λ/2. As shown in **Fig 45**, this increases the spacing between the two in-phase λ/2 sections at the ends of the wires. The section in the center carries a current of opposite phase, but if this section is short the current will be small; it represents only the outer ends of a λ/2 antenna section. Because of the small current and short length, the radiation from the center is small. The optimum length for each element is 0.64 λ. At greater lengths the system tends to act as a long-wire antenna, and the gain decreases.

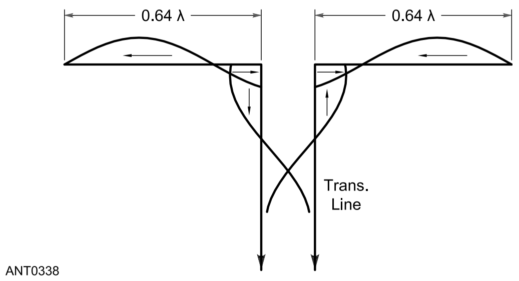

Fig 45—The extended double Zepp. This system gives somewhat more gain than two λ-sized collinear elements.

This system is known as the *extended double Zepp*. The gain over a λ/2 dipole is approximately 3 dB, as compared with about 1.6 dB for two collinear λ/2 dipoles. The directional pattern in the plane containing the axis of the antenna is shown in **Fig 46**. As in the case of all other collinear arrays, the free-space pattern in the plane at right angles to the antenna elements is the same as that of a λ/2 antenna—circular.

This antenna is not resonant at the operating frequency so that the feed-point impedance is complex (R ± *j* X). A typical example of the variation of the feed-point impedance over the band for a 40-meter double-extended Zepp is shown in **Fig 47**. This antenna is normally fed with open-wire transmission line to an antenna tuner. Other matching arrangements are, of course, possible. A method for transforming the feed-point impedance to 450 Ω and eliminating the minor lobes is given in Chapter 6.

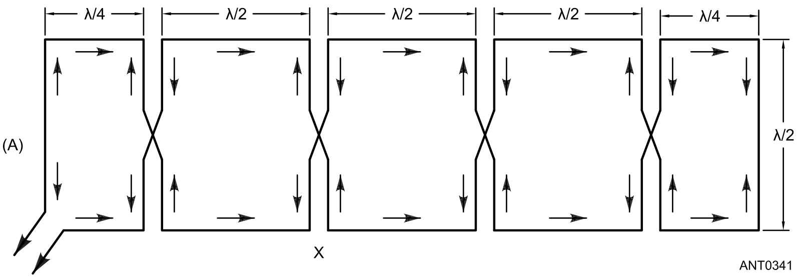

Fig 48—Typical Sterba array, an 8-element version.

THE STERBA ARRAY

Two collinear arrays can be combined to form the *Sterba array*, often called the *Sterba curtain*. An 8-element example of a Sterba array is shown in **Fig 48**. The four λ/4 elements joined on the ends are equivalent to two λ/2 elements. The two collinear arrays are spaced λ/2 and the λ/4 phasing lines connected together to provide λ/2 phasing lines. This arrangement has the advantage of increasing the gain for a given length and also increasing the E-plane directivity, which is no longer circular. An additional advantage of this array is that the wire forms a closed loop. For installations where icing is a problem a low voltage dc or low frequency (50 or 60 Hz) ac current can be passed through the wire to heat it for deicing. The heating current is isolated from RF by decoupling chokes. This is standard practice in commercial installations.

The number of sections in a Sterba array can be extended as far as desired but more than four or five are rarely used because of the slow increase in gain with extra elements, the narrow H-plane directivity and the appearance of multiple sidelobes. When fed at the point indicated the impedance is about 600 Ω. The antenna can also be fed at the point marked X. The impedance at this point will be about 1 kΩ. The gain of the 8-element array in Fig 48 will be between 7 to 8 dB over a single element.

Parallel Broadside Arrays

To obtain broadside directivity with parallel elements the currents in the elements must all be in-phase. At a distant point lying on a line perpendicular to the axis of the array and also perpendicular to the plane containing the elements, the fields from all elements add up in phase. The situation is like that pictured in Fig 1 in this chapter, where four parallel λ/2 dipoles were fed together a broadside array.

Broadside arrays of this type theoretically can have any number of elements. However, practical limitations of construction and available space usually limit the number of broadside parallel elements.

POWER GAIN

The power gain of a parallel-element broadside array depends on the spacing between elements as well as on the number of elements. The way in which the gain of a two-element array varies with spacing is shown in **Fig 49**. The greatest gain is obtained when the spacing is in the vicinity of 0.67 λ.

The theoretical gains of broadside arrays having more than two elements are approximately as follows:

No. of Parallel Elements	*dB Gain with λ/2 Spacing*	*dB Gain with 3λ/4 Spacing*
3	5.7	7.2
4	7.1	8.5
5	8.1	9.4
6	8.9	10.4

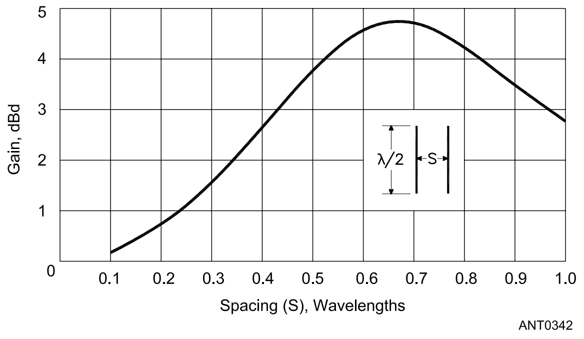

Fig 49—Gain as a function of the spacing between two parallel elements operated in-phase (broadside).

The elements must, of course, all lie in the same plane and all must be fed in-phase.

DIRECTIVITY

The sharpness of the directive pattern depends on spacing between elements and number of elements. Larger element spacing will sharpen the main lobe, for a given number of elements, up to a point as was shown in Fig 41. The two-element array has no minor lobes when the spacing is λ/2, but small minor lobes appear at greater spacings. When three or more elements are used the pattern always has minor lobes.

Other Forms Of Broadside Arrays

For those who have the available room, multi-element arrays based on the broadside concept have something to offer. The antennas are large but of simple design and noncritical dimensions; they are also very economical in terms of gain per unit of cost.

Large arrays can often be fed at several different points. However, the pattern symmetry may be sensitive to the choice of feed point within the array. Non-symmetrical feed points will result in small asymmetries in the pattern but these are not usually of great concern.

Arrays of three and four elements are shown in **Fig 50**. In the 3-element array with λ/2 spacing at A, the array is fed at the center. This is the most desirable point in that it tends to keep the power distribution among the elements uniform. However, the transmission line could alternatively be connected at either point B or C of Fig 50A, with only slight skewing of the radiation pattern.

When the spacing is greater than λ/2, the phasing lines must be 1 λ long and are not transposed between elements. This is shown Fig 50B. With this arrangement, any element spacing up to 1 λ can be used, if the phasing lines can be folded as suggested in the drawing.

The 2-element array at C is fed at the center of the system to make the power distribution among elements as uniform as possible. However, the transmission line could be connected at either point B, C, D or E. In this case the section of phasing line between B and D must be transposed to make the currents flow in the same direction in all elements. The 4-element array at C and the 3-element array at B have approximately the same gain when the element spacing in the array at B is 3λ/4.

An alternative feeding method is shown in Fig 50D. This system can also be applied to the 3-element arrays, and will result in better symmetry in any case. It is necessary only to move the phasing line to the center of each element, making connection to both sides of the line instead of one only.

The free-space pattern for a 4-element array with λ/2 spacing is shown in **Fig 51**. This is also approximately the pattern for a 3-element array with 3λ/4 spacing.

Larger arrays can be designed and constructed by following the phasing principles shown in the drawings. No accurate figures are available for the impedances at the various feed points indicated in Fig 50. You can estimate it to be in the vicinity of 1 kΩ when the feed point is at a junction between the phasing line and a λ/2 element, becoming smaller as the number of elements in the array is increased. When the feed point is midway between end-fed elements as in Fig 50C, the feed-point impedance of a 4-element array is in the vicinity of 200 to 300 Ω, with 600-Ω open-wire phasing lines. The impedance at the feed point with the antenna shown at D should be about 1.5 kΩ.

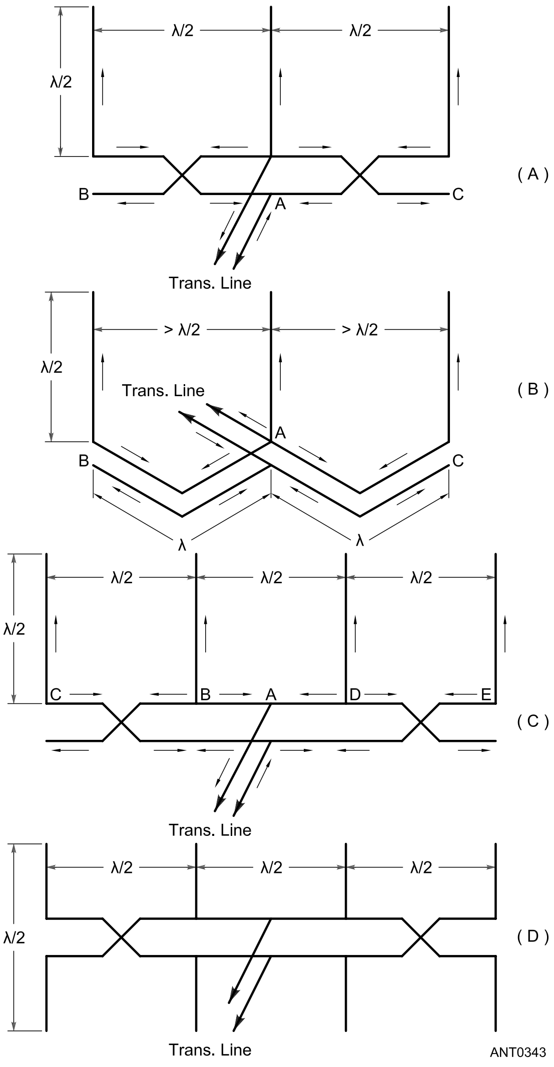

Fig 50—Methods of feeding 3- and 4-element broadside arrays with parallel elements.

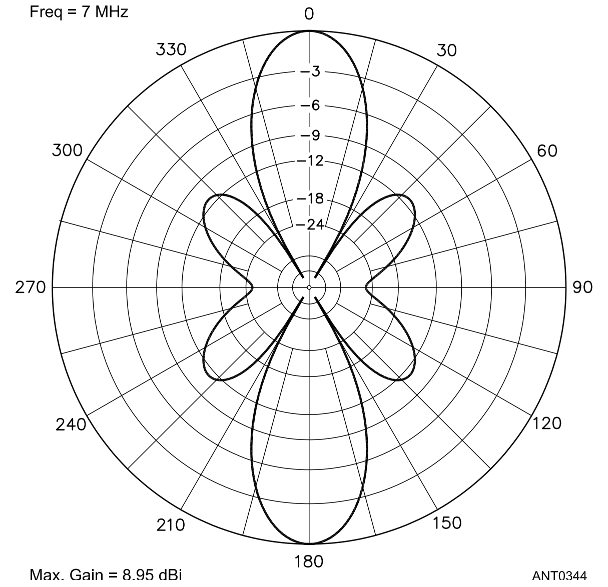

Fig 51—Free-space E-plane pattern of a 4-element broadside array using parallel elements (Fig 50). This corresponds to the horizontal directive pattern at low wave angles for a vertically polarized array over ground. The axis of the elements lies along the 90°-270° line.

NON-UNIFORM ELEMENT CURRENTS

The pattern for a 4-element broadside array shown in Fig 51 has substantial side lobes. This is typical for arrays more than λ/2 wide when equal currents flow in each element. Sidelobe amplitude can be reduced by using non-uniform current distribution among the elements. Many possible current amplitude distributions have been suggested. All of them have reduced current in the outer elements and greater current in the inner elements. This reduces the gain somewhat but can produce a more desirable pattern. One of the common current distributions is called *binomial current grading*. In this scheme the ratio of element currents is set equal to the coefficients of a polynomial. For example:

$$\begin{aligned}
&1\,x+1, \quad \Rightarrow \ 1,1 \\
&(x+1)^2 = 1x^2 + 2x + 1, \quad \Rightarrow \ 1, 2, 1 \\
&(x+1)^3 = 1x^3 + 3x^2 + 3x + 1, \quad \Rightarrow \ 1, 3, 3, 1 \\
&(x+1)^4 = 1x^4 + 4x^3 + 6x^2 + 6x + 1, \quad \Rightarrow \ 1, 4, 6, 4, 1
\end{aligned} \qquad \text{(Eq 7)}$$

In a 2-element array the currents are equal, in a 3-element array the current in the center element is twice that in the outer elements, and so on.

HALF-SQUARE ANTENNA

On the low-frequency bands (40, 80 and 160 meters) it becomes increasingly difficult to use λ/2 elements because of their size. The half-square antenna is a 2-element broadside array with λ/4-high vertical elements and λ/2 horizontal spacing. See **Fig 52**. The free-space H-plane pattern for this array is shown in **Fig 53**. The antenna gives modest (4.2 dBi) but useful gain and has the advantage of only λ/4 height. Like all vertically polarized antennas, real-world performance depends directly on the characteristics of the ground surrounding it.

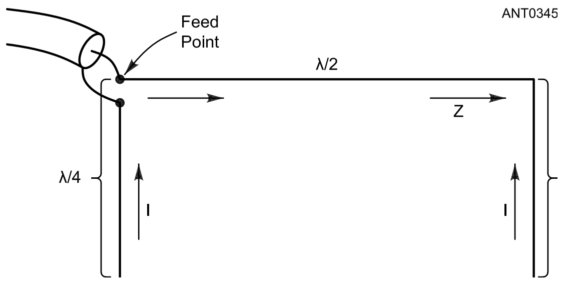

Fig 52—Layout for the half-square antenna.

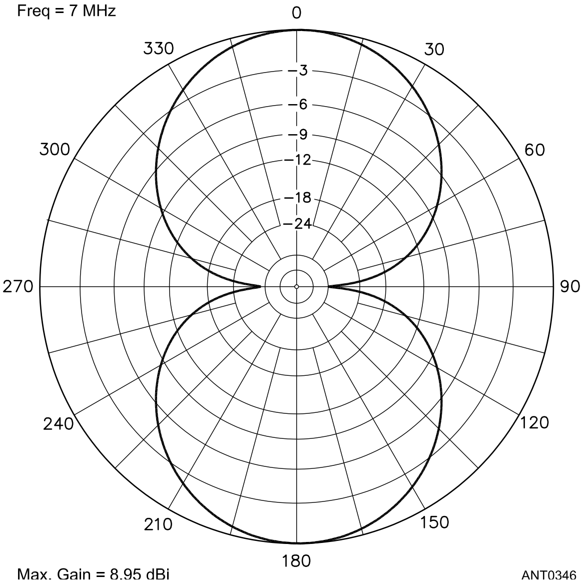

Fig 53—Free-space E-plane directive pattern for the half-square antenna.

The half-square can be fed either at the point indicated or at the bottom end of one of the vertical elements using a voltage-feed scheme, such as that shown in **Fig 54** for the bobtail curtain. The feed-point impedance is in the region of 50 Ω when fed at a corner as shown in Fig 52. A typical SWR plot is shown in **Fig 55**. Chapter 6 has a detailed discussion of the half-square antenna with several variations, together with practical considerations.

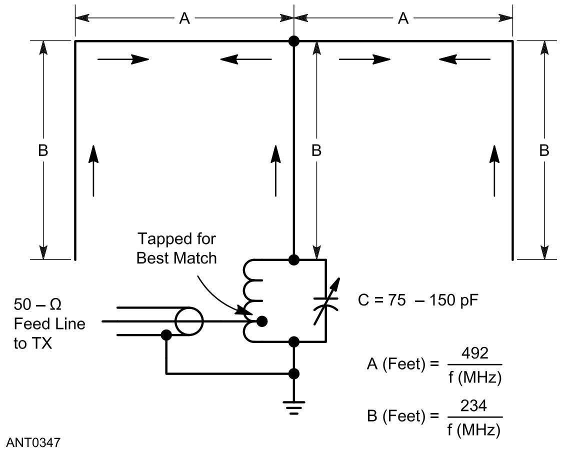

Fig 54—The bobtail curtain is an excellent low-angle radiator having broadside bidirectional characteristics. Current distribution is represented by the arrows. Dimensions A and B (in feet, for wire antennas) can be determined from the equations.

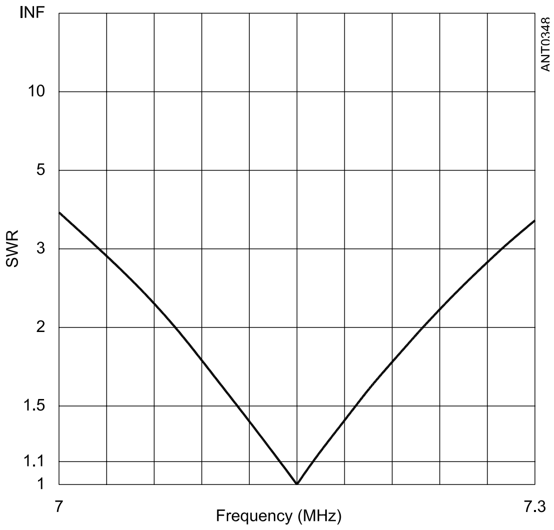

Fig 55—Typical SWR plot for a 40-meter half-square antenna fed at one corner. Antenna in free space.

BOBTAIL CURTAIN

The antenna system in Fig 54 uses the principles of co-phased verticals to produce a broadside, bidirectional pattern providing approximately 5.1 dB of gain over a single λ/4 element. The antenna performs as three in-phase, top-fed vertical radiators approximately λ/4 in height and spaced approximately λ/2. It is most effective for low-angle signals and makes an excellent long-distance antenna for 1.8, 3.5 or 7 MHz.

The three vertical sections are the actual radiating components, but only the center element is fed directly. The two horizontal parts, A, act as phasing lines and contribute very little to the radiation pattern. Because the current in the center element must be divided between the end sections, the current distribution approaches a binomial 1:2:1 ratio. The radiation pattern is shown in **Fig 56**.

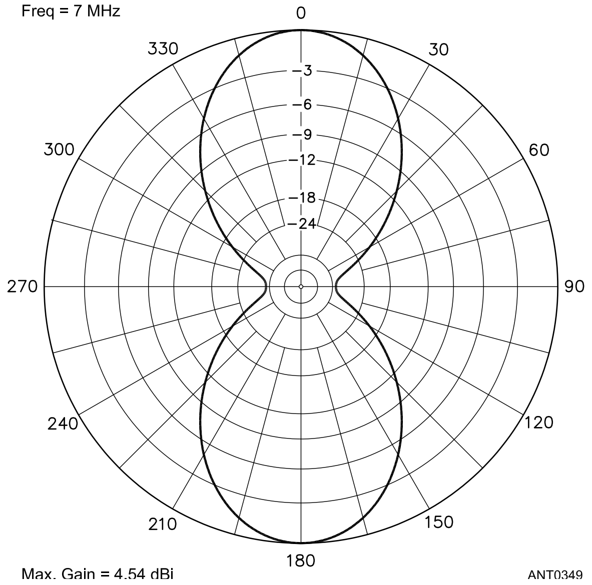

Fig 56—Calculated free-space E-plane directive diagram of the bobtail curtain shown in Fig 54. The array lies along the 90°-270° axis.

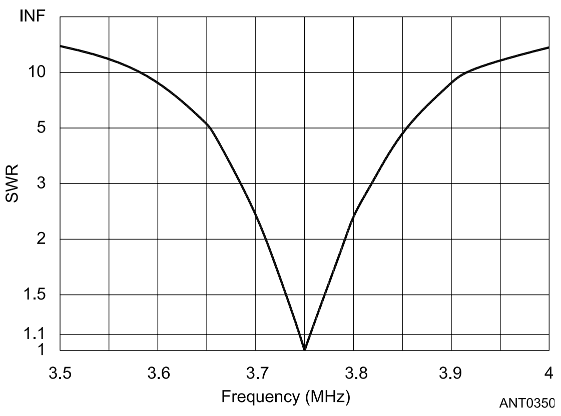

Fig 57—Typical SWR plot for an 80-meter bobtail curtain in free space. This is a narrow-band antenna.

The vertical elements should be as vertical as possible. The height for the horizontal portion should be slightly greater than B, as shown in Fig 54. The tuning network is resonant at the operating frequency. The L/C ratio should be fairly low to provide good loading characteristics. As a starting point, a maximum capacitor value of 75 to 150 pF is recommended, and the inductor value is determined by C and the operating frequency. The network is first tuned to resonance and then the tap point is adjusted for the best match. A slight readjustment of C may be necessary. A link coil consisting of a few turns can also be used to feed the antenna.

A feeling for the matching bandwidth of this antenna can be obtained by looking at a feed point located at the top end of the center element. The impedance at this point will be approximately 32 Ω. An SWR plot (for $Z_0 = 32\ \Omega$) for an 80-meter bobtail curtain at this feed-point is shown in **Fig 57**. However, it is not advisable to actually connect a feed line at this point since it would detune the array and alter the pattern. This antenna is relatively narrow band. When fed at the bottom of the center element as shown in Fig 54, the SWR can be adjusted to be 1:1 at one frequency but the operating bandwidth for SWR < 2:1 may be even narrower than Fig 57 shows. For 80-meters, where operation is often desired in the CW DX window (3.510 MHz) and in the phone DX window (3.790 MHz), it will be necessary to retune the matching network as you change frequency. This can be done by switching a capacitor in or out, manually or remotely with a relay.

While the match bandwidth is quite narrow, the radiation pattern changes more slowly with frequency. **Fig 58** shows the variation in the pattern over the entire band (3.5 to 4.0 MHz). As would be expected, the gain increases with frequency because the antenna is larger in terms of wavelengths. The general shape of the pattern, however, is quite stable.

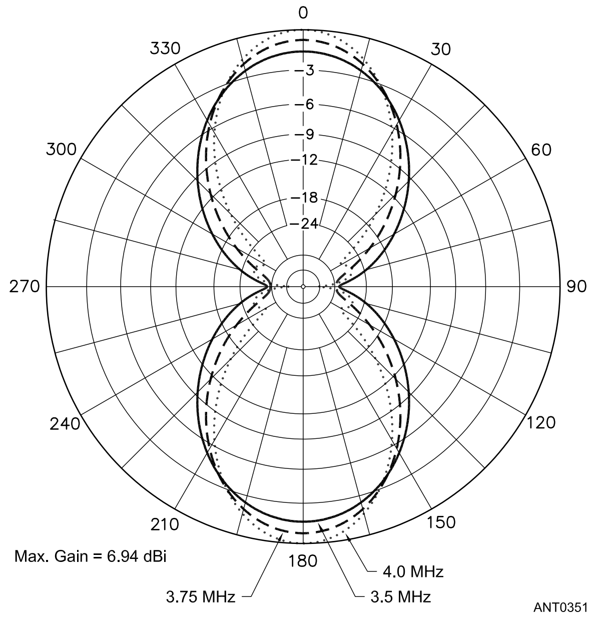

Fig 58—80-meter bobtail curtain's free-space E-plane pattern variation over the 80-meter band.

THE BRUCE ARRAY

Four variations of the Bruce array are shown in **Fig 59**. The Bruce is simply a wire folded so that the vertical sections carry large in-phase currents, while the horizontal sections carry small currents flowing in opposite directions with respect to the center of a section (indicated by dots). The radiation is vertically polarized. The gain is proportional to the length of the array but is somewhat smaller than you can obtain from a broadside array of λ/2 elements of the same length. This is because the radiating portion of the elements is only λ/4.

The Bruce array has a number of advantages:

1) The array is only λ/4 high. This is especially helpful on 80 and 160 meters, where the height of λ/2 supports becomes impractical for most amateurs.

2) The array is very simple. It is just a single piece of wire folded to form the array.

3) The dimensions of the array are very flexible. Depending on the available distance between supports, any number of elements can be used. The longer the array, the greater the gain.

4) The shape of the array does not have to be exactly 1.05 λ/4 squares. If the available height is short but the array can be made longer, then shorter vertical sections and longer horizontal sections can be used to maintain gain and resonance. Conversely, if more height is available but width is restricted then longer vertical sections can be used with shorter horizontal sections.

5) The array can be fed at other points more convenient for a particular installation.

6) The antenna is relatively low Q, so that the feed-point impedance changes slowly with frequency. This is very helpful on 80 meters, for example, where the antenna can be relatively broadband.

7) The radiation pattern and gain is stable over the width of an amateur band.

Note that the nominal dimensions of the array in Fig 59 call for section lengths = 1.05 λ/4. The need to use slightly longer elements to achieve resonance is common in large wire arrays. A quad loop behaves in the same manner. This is quite different from wire dipoles, which are typically *shortened* by 2-5% to achieve resonance.

Fig 60 shows the variations in gain and pattern for 2 to 5-element 80-meter Bruce arrays. **Table 2** lists the gain over a vertical λ/2 dipole, a 4-radial ground-plane vertical and the size of the array. The gain and impedance parameters listed are for free space. Over real ground the patterns and gain will depend on the height above ground and the ground characteristics. Copper loss using #12 conductors in included.

Worthwhile gain can be obtained from these arrays, especially on 80 and 160 meters, where any gain is hard to come by. The feed-point impedance is for the center of a vertical section. From the patterns in Fig 60 you can see that sidelobes start to appear as the length of the array is increased beyond 3λ/4. This is typical for arrays using equal currents in the elements.

It is interesting to compare the bobtail curtain (Fig 54) with a 4-element Bruce array. **Fig 61** compares the radiation patterns for these two antennas. Even though the Bruce is shorter (3λ/4) than the bobtail (1 λ), it has slightly more

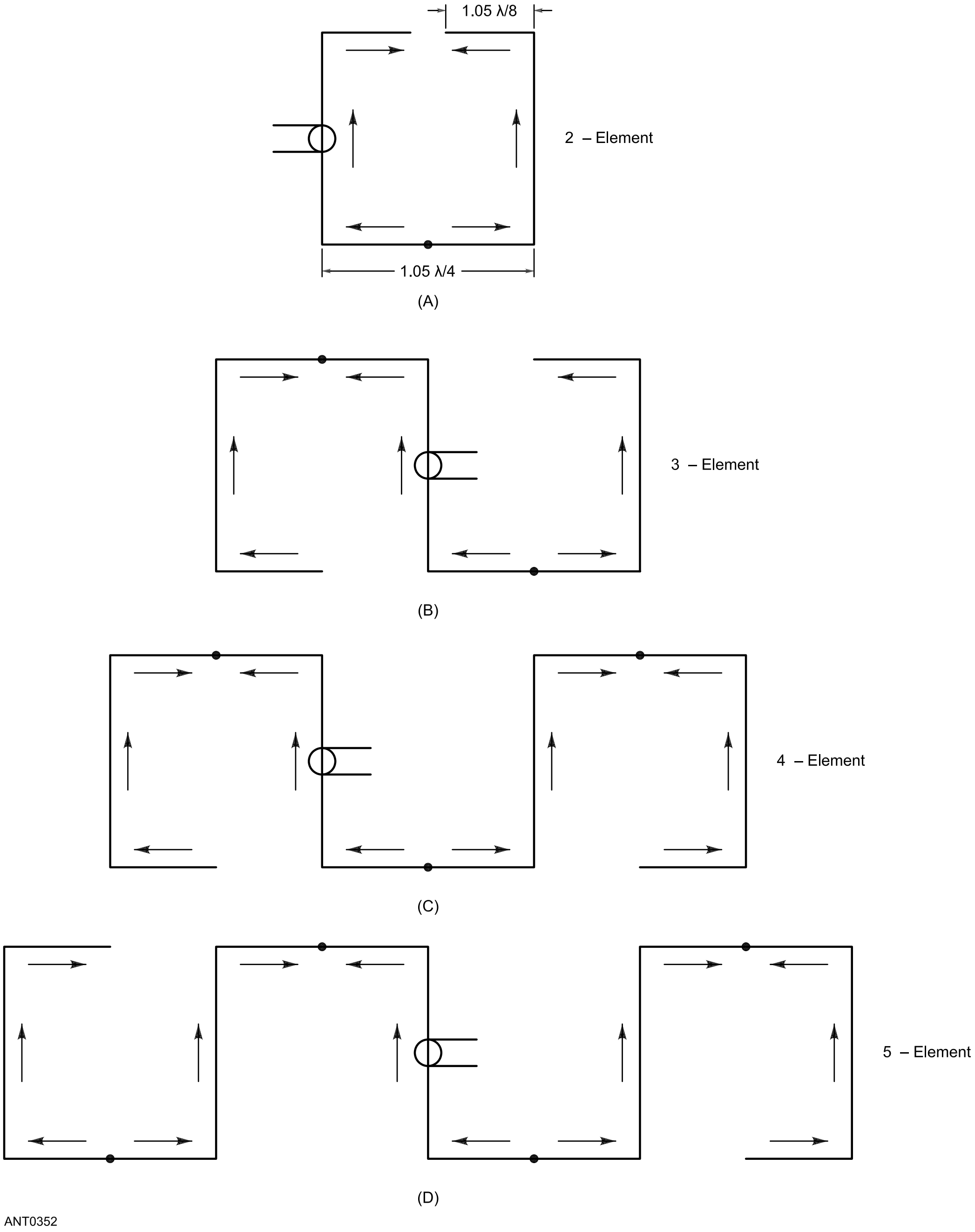
1.05 λ/8
2 – Element
1.05 λ/4
(A)
3 – Element
(B)
4 – Element
(C)
5 – Element
(D)
ANT0352

Fig 59—Various Bruce arrays: 2, 3, 4 and 5-element versions.

Table 2
Bruce array length, impedance and gain as a function of number of elements

Number Elements	*Gain Over λ/2 Vertical Dipole*	*Gain over λ/4 Ground-Plane*	*Array Length Wavelengths*	*Approx. Feed Z, Ω*
2	1.2 dB	1.9 dB	¼	130
3	2.8 dB	3.6 dB	½	200
4	4.3 dB	5.1 dB	¾	250
5	5.3 dB	6.1 dB	1	300

gain. The matching bandwidth is illustrated by the SWR curve in **Fig 62**. The 4-element Bruce has over twice the match bandwidth (200 kHz) than does the bobtail (75 kHz in Fig 57). Part of the gain difference is due to the binomial current distribution—the center element has twice the current as the outer elements in the bobtail. This reduces the gain slightly so that the 4-element Bruce becomes competitive. This is a good example of using more than the minimum number of elements to improve performance or to reduce size. On 160 meters the 4-element Bruce will be 140 feet shorter than the bobtail, a significant reduction. If additional space is available for the bobtail (1 λ) then a 5-element Bruce could

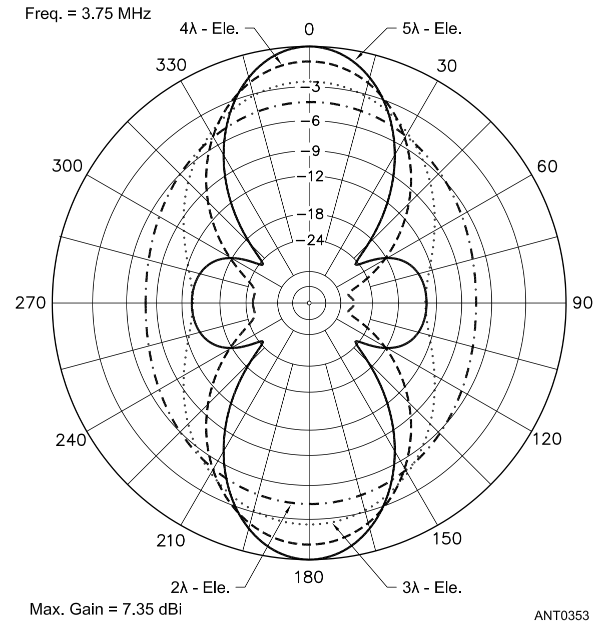

Fig 60—80-meter free-space E-plane directive patterns for the Bruce arrays shown in Fig 59. The 5-element's pattern is a solid line; the 4-element is a dashed line; the 3-element is a dotted line, and the 2-element version is a dashed-dotted line.

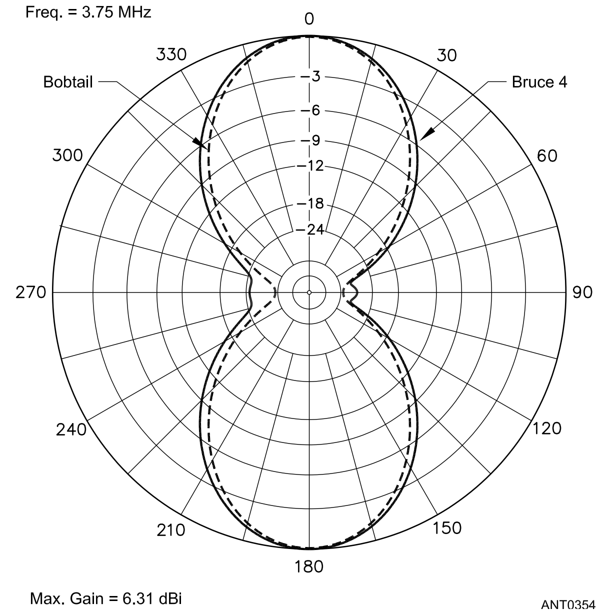

Fig 61—Comparison of free space patterns of a 4-element Bruce array (solid line) and a 3-element bobtail curtain (dashed line).

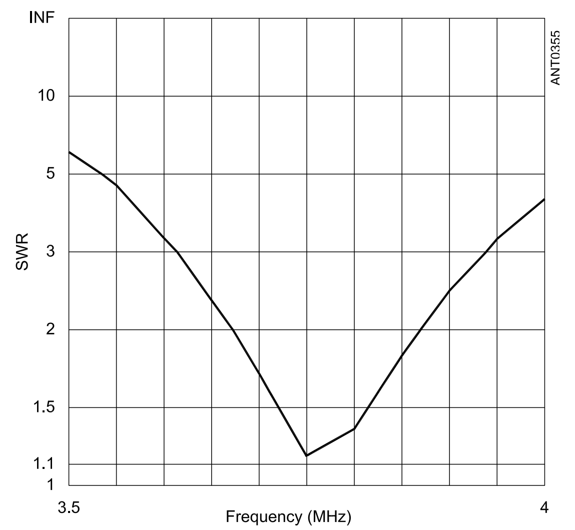

Fig 62—Typical SWR curve for a 4-element 80-meter Bruce array.

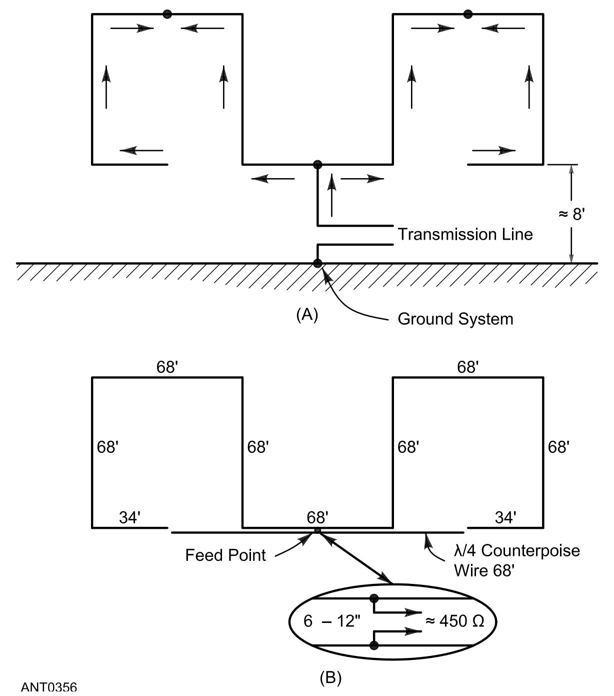

Fig 63—Alternate feed arrangements for the Bruce array. At A, the antenna is driven against a ground system and at B, it uses a two-wire counterpoise.

be used, with a small increase in gain but also introducing some sidelobes.

The 2-element Bruce and the half-square antennas are both 2-element arrays. However, since the spacing between radiators is greater in the half-square (λ/2) the gain of the half-square is about 1 dB greater. If space is available, the half-square would be a better choice. If there is not room for a half-square then the Bruce, which is only half as long (λ/4), may be a good alternative. The 3-element Bruce, which has the same length (λ/2) as the half-square, has about 0.6 dB more gain than the half-square and will have a wider match bandwidth.

The Bruce antenna can be fed at many different points and in different ways. In addition to the feed points indicated in Fig 59, you may connect the feed line at the center of any of the vertical sections. In longer Bruce arrays, feeding at one end will result in some current imbalance among the elements but the resulting pattern distortion is small. Actually, the feed-point can be anywhere along a vertical section. One very convenient point is at an outside corner. The feed-point impedance will be higher (about 600 Ω). A good match for 450-Ω ladder-line can usually be found somewhere on the vertical section. It is important to recognize that feeding the antenna at a voltage node (dots in Fig 59) by breaking the wire and inserting an insulator, completely changes the current distribution. This will be discussed in the section on endfire arrays.

A Bruce can be fed unbalanced against ground or against a counterpoise as shown in **Fig 63**. Because it is a vertically polarized antenna, the better the ground system, the better the performance. As few as two elevated radials can be used as shown in Fig 63B, but more radials can also be used to improve the performance, depending on local ground constants. The original development of the Bruce array in the late 1920s used this feed arrangement.

FOUR-ELEMENT BROADSIDE ARRAY

The 4-element array shown in **Fig 64** is commonly known as the *lazy H*. It consists of a set of two collinear elements and a set of two parallel elements, all operated in-phase to give broadside directivity. The gain and directivity will depend on the spacing, as in the case of a simple parallel-element broadside array. The spacing may be chosen between the limits shown on the drawing, but spacings below 3λ/8 are not worthwhile because the gain is small. Estimated gains compared to a single element are:

3λ/8 spacing—4.2 dB
λ/2 spacing—5.8 dB
5λ/8 spacing—6.7 dB
3λ/4 spacing—6.3 dB

Half-wave spacing is generally used. Directive patterns for this spacing are given in **Figs 65** and **66**. With λ/2 spacing between parallel elements, the impedance at the junction of the phasing line and transmission line is resistive and in the vicinity of 100 Ω. With larger or smaller spacing the impedance at this junction will be reactive as well as resistive. Matching stubs are recommended in cases where a non-resonant line is to be used. They may be calculated and adjusted as described in Chapter 26.

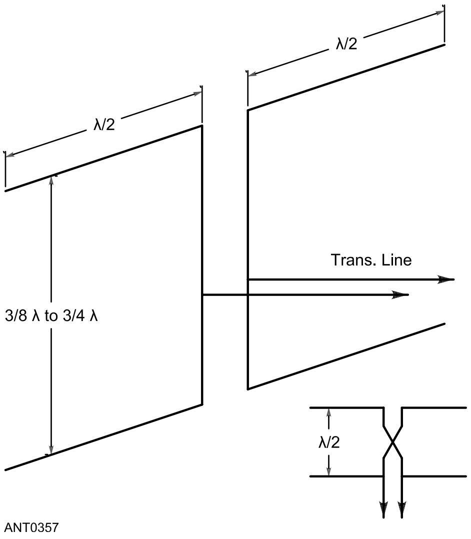

Fig 64—Four-element broadside array ("lazy H") using collinear and parallel elements.

The system shown in Fig 64 may be used on two bands having a 2-to-1 frequency relationship. It should be designed for the higher of the two frequencies, using 3λ/4 spacing between parallel elements. It will then operate on the lower frequency as a simple broadside array with 3λ/8 spacing.

An alternative method of feeding is shown in the small diagram in Fig 64. In this case the elements and the phasing line must be adjusted exactly to an electrical half wavelength. The impedance at the feed point will be resistive and on the order of 2 kΩ.

THE BI-SQUARE ANTENNA

A development of the lazy H, known as the *bi-square antenna*, is shown in **Fig 67**. The gain of the bi-square is somewhat less than that of the lazy-H, but this array is attractive because it can be supported from a single pole. It has a circumference of 2 λ at the operating frequency, and is horizontally polarized.

The bi-square antenna consists of two 1 λ radiators, fed 180° out-of-phase at the bottom of the array. The radiation resistance is 300 Ω, so it can be fed with either 300- or 600-Ω line. The free space gain of the antenna is about 5.8 dBi, which is 3.7 dB more than a single dipole element. Gain may be increased by adding a parasitic reflector or director. Two bi-square arrays can be mounted at right angles and switched to provide omnidirectional coverage. In this way, the antenna wires may be used as part of the guying system for the pole.

Although it resembles a loop antenna, the bi-square is not a true loop because the ends opposite the feed point

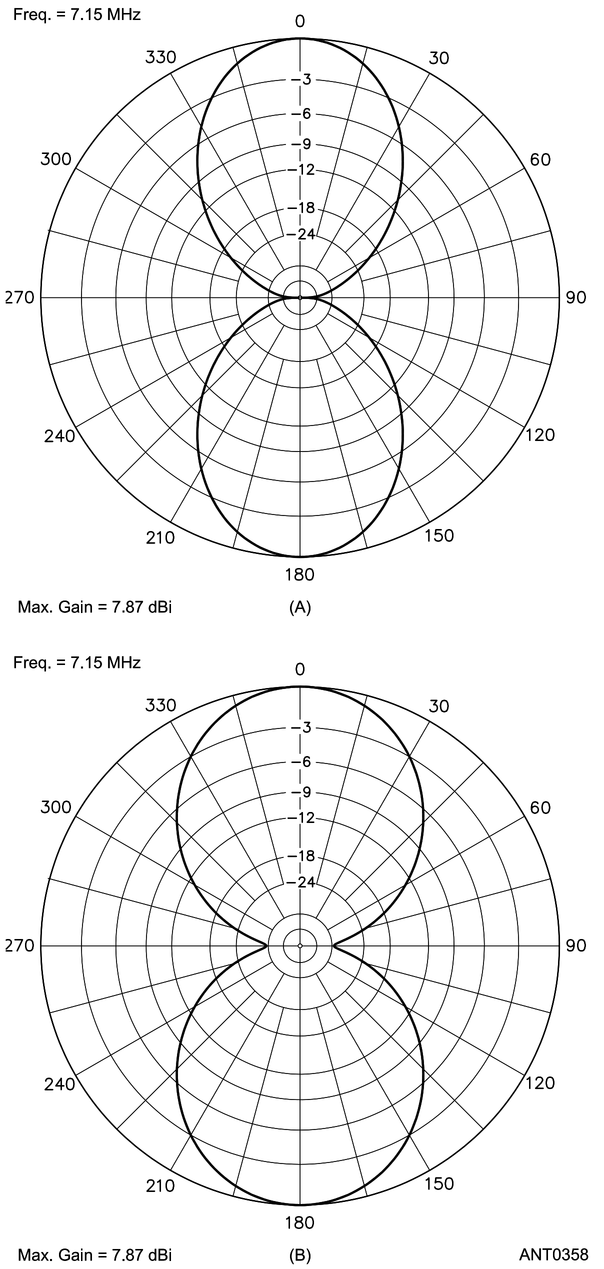

Fig 65—Free-space directive diagrams of the 4-element antenna shown in Fig 64. At A is the E-plane pattern. The axis of the elements lies along the 90°-270° line. At B is the free-space H-plane pattern, viewed as if one set of elements is above the other from the ends of the elements.

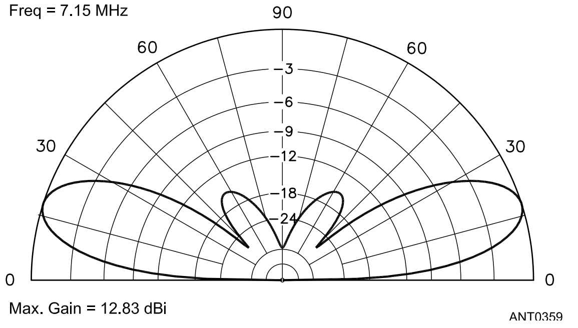

Fig 66—Vertical pattern of the 4-element broadside antenna of Fig 64, when mounted with the elements horizontal and the lower set λ/4 above flat ground. Stacked arrays of this type give best results when the lowest elements are at least λ/2 high. The gain is reduced and the wave angle raised if the lowest elements are too close to ground.

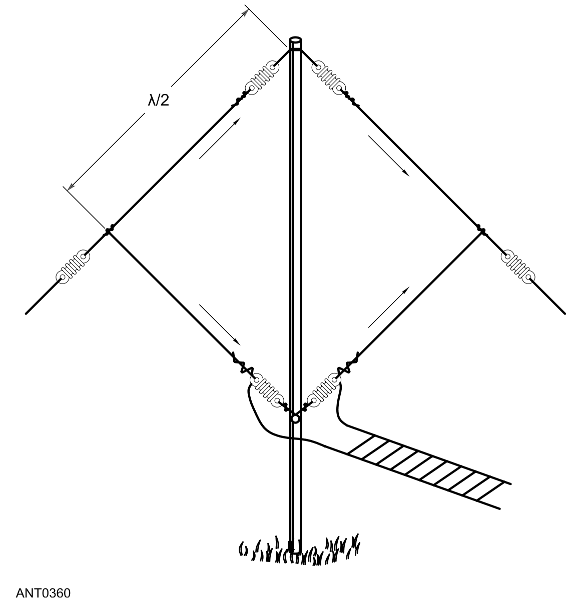

Fig 67—The bi-square array. It has the appearance of a loop, but is not a true loop because the conductor is open at the top. The length of each side, in feet, is 480/f (MHz).

are open. However, identical construction techniques can be used for the two antenna types. Indeed, with a means of remotely closing the connection at the top for lower frequency operation, the antenna can be operated on two harmonically related bands. As an example, an array with 17 feet per side can be operated as a bi-square at 28 MHz and as a full-wave loop at 14 MHz. For two-band operation in this manner, the side length should favor the higher frequency. The length of a closed loop is not as critical.

End-Fire Arrays

The term *end-fire* covers a number of different methods of operation, all having in common the fact that the maximum radiation takes place along the array axis, and that the array consists of a number of parallel elements in one plane. End-fire arrays can be either bidirectional or unidirectional. In the bidirectional type commonly used by amateurs there are only two elements, and these are operated with currents 180° out-of-phase. Even though adjustment tends to be complicated, unidirectional end-fire driven arrays have also seen amateur use, primarily as a pair of phased, ground-mounted λ/4 vertical elements. Extensive discussion of this array is contained in earlier sections of this chapter.

Horizontally polarized unidirectional end-fire arrays see little amateur use except in log-periodic arrays (described in Chapter 10). Instead, horizontally polarized unidirectional arrays usually have parasitic elements (described in Chapter 11) and are called Yagis.

TWO-ELEMENT END-FIRE ARRAY

In a 2-element array with equal currents out-of-phase, the gain varies with the spacing between elements as shown in **Fig 68**. The maximum gain occurs in the neighborhood of 0.1 λ spacing. Below that the gain drops rapidly due to conductor loss resistance.

The feed-point resistance for either element is very low at the spacings giving greatest gain, as shown in Fig 8 earlier in this chapter. The spacings most frequently used are λ/8 and λ/4, at which the resistances of center-fed λ/2 elements are about 9 and 32 Ω, respectively.

The effect of conductor resistance on gain for various spacings is shown in **Fig 69**. Because current along the element is not constant (it is approximately sinusoidal), the resistance shown is the equivalent resistance (R_{eq}) inserted at the center of the element to account for the loss distributed along the element.

The equivalent resistance of a λ/2 element is one half the ac resistance (R_{ac}) of the complete element. R_{ac} is usually >> R_{dc} due to skin effect. For example, a 1.84 MHz dipole using #12 copper wire will have the following R_{eq}:

Wire length = 267 feet
Rdc = 0.00159 [/foot] × 267 [feet] = 0.42
Fr = Rac/Rdc = 10.8
Req =(Rdc/2) × Fr = 2.29

For a 3.75 MHz dipole made with #12 wire, Req = 1.59 Ω. In Fig 69, it is clear that end-fire antennas made with #12 or smaller wire will limit the attainable gain because of losses. There is no point in using spacings much less than λ/4 if you use wire elements. If instead you use elements made of aluminum tubing then smaller spacings can be used to increase gain. However, as the spacing is reduced below λ/4 the increase in gain is quite small even with good conductors. Closer spacings give little gain increase but can drastically reduce the operating bandwidth due to the rapidly increasing Q of the array.

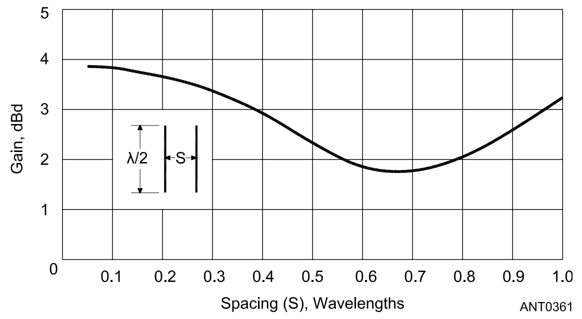

Fig 68—Gain of an end-fire array consisting of two elements fed 180° out-of-phase, as a function of the spacing between elements. Maximum radiation is in the plane of the elements and at right angles to them at spacings up to λ/2, but the direction changes at greater spacings.

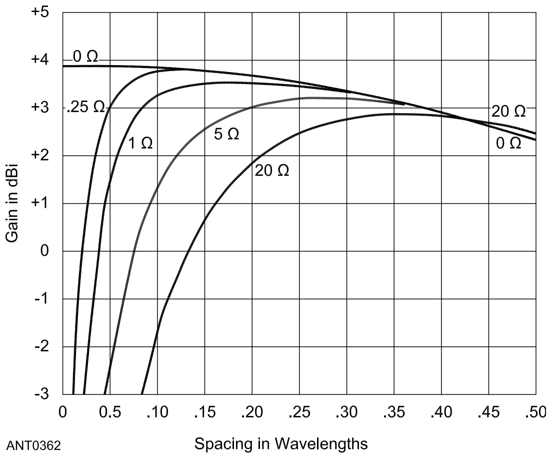

Fig 69—Gain over a single element of two out-of-phase elements in free space as a function of spacing for various loss resistances.

Unidirectional End-Fire Arrays

Two parallel elements spaced λ/4 apart and fed equal currents 90° out-of-phase will have a directional pattern in the plane at right angles to the plane of the array. See **Fig 70**. The maximum radiation is in the direction of the element in which the current lags. In the opposite direction the fields from the two elements cancel.

When the currents in the elements are neither in-phase nor 180° out-of-phase, the feed-point resistances of the elements are not equal. This complicates the problem of feeding equal currents to the elements, as discussed in earlier sections.

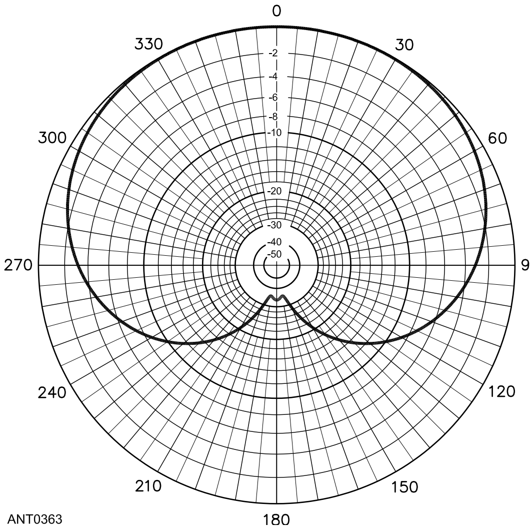

Fig 70—Representative H-plane pattern for a 2-element end-fire array with 90° spacing and phasing. The elements lie along the vertical axis, with the uppermost element the one of lagging phase. Dissimilar current distributions are taken into account. (Pattern computed with *ELNEC*.)

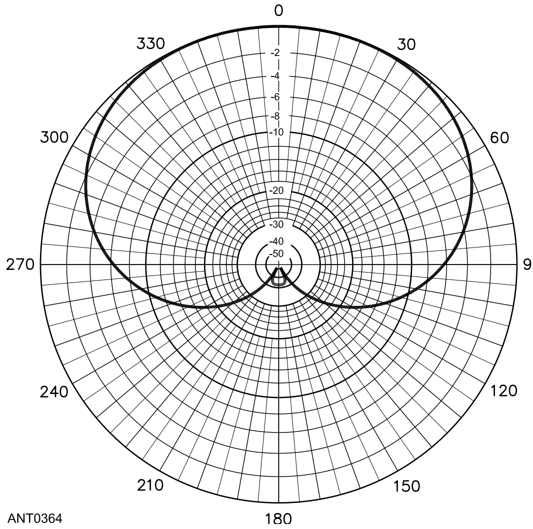

Fig 71—H-plane pattern for a 3-element end-fire array with binomial current distribution (the current in the center element is twice that in each end element). The elements are spaced λ/4 apart along the 0°-180° axis. The center element lags the lower element by 90°, while the upper element lags the lower element by 180° in phase. Dissimilar current distributions are taken into account. (Pattern computed with *ELNEC*.)

More than two elements can be used in a unidirectional end-fire array. The requirement for unidirectivity is that there must be a progressive phase shift in the element currents equal to the spacing, in electrical degrees, between the elements. The amplitudes of the currents in the various elements also must be properly related. This requires binomial current distribution. In the case of three elements, this requires that the current in the center element be twice that in the two outside elements, for 90° (λ/4) spacing and element current phasing. This antenna has an overall length of λ/2. The directive diagram is shown in **Fig 71**. The pattern is similar to that of Fig 70, but the 3-element binomial array has greater directivity, evidenced by the narrower half-power beamwidth (146° versus 176°). Its gain is 1.0 dB greater.

THE W8JK ARRAY

As pointed out earlier, John Kraus, W8JK, described his bidirectional flat-top W8JK beam antenna in 1940. See **Fig 72**. Two λ/2 elements are spaced λ/8 to λ/4 and driven 180° out-of-phase. The free-space radiation pattern for this antenna, using #12 copper wire, is given in **Fig 73**. The pattern is representative of spacings between λ/8 and λ/4 where the gain varies less than 0.5 dB. The gain over a dipole is about 3.3 dB (5.4 dBi referenced to an isotropic radiator), a worthwhile improvement. The feed-point impedance (including wire resistance) of *each* element is about 11 Ω for λ/8 spacing and 33 Ω for λ/4 spacing. The feed-point impedance at the center connection will depend on the length and Z_0 of the connecting transmission line.

Kraus gave a number of other variations for end-fire arrays, some of which are shown in **Fig 74**. The ones fed at the center (A, C and E) are usually horizontally polarized flat-top beams. The end-fed versions (B, D & F) are usually vertically polarized, where the feed point can be conveniently near ground.

A practical variation of Fig 74B is given in **Fig 75**. In this example, the height is limited to λ/4 so the ends can be bent over as shown, producing a 2-element Bruce array. This reduces the gain somewhat but allows much shorter supports, an important consideration on the low bands. If additional height is available, then you can achieve some additional gain. The upper ends can be bent over to fit the available height. The feed-point impedance will greater than 1 kΩ.

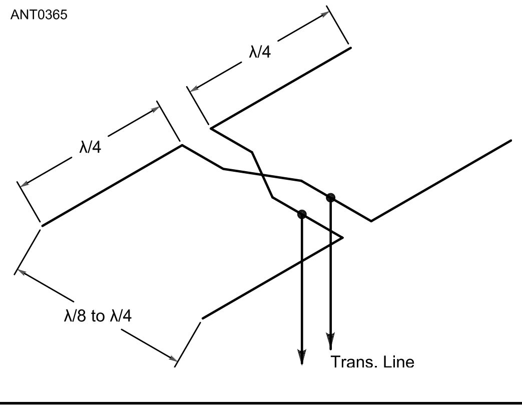

Fig 72—A 2-element W8JK array.

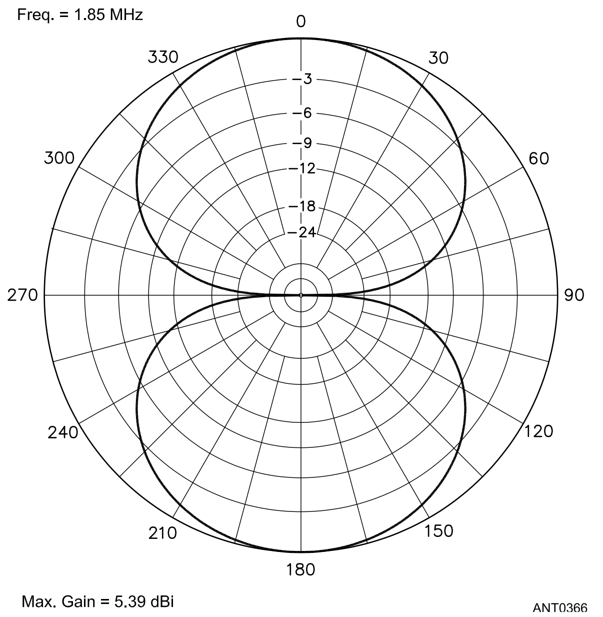

Fig 73—Free-space E-plane pattern for the 2-element W8JK array

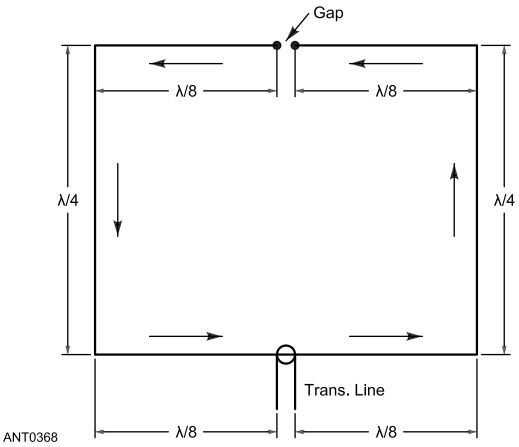
Gap
λ/8
λ/8
λ/4
λ/4
Trans. Line
λ/8
λ/8
ANT0368

Fig 75—A 2-element end-fire array with reduced height.

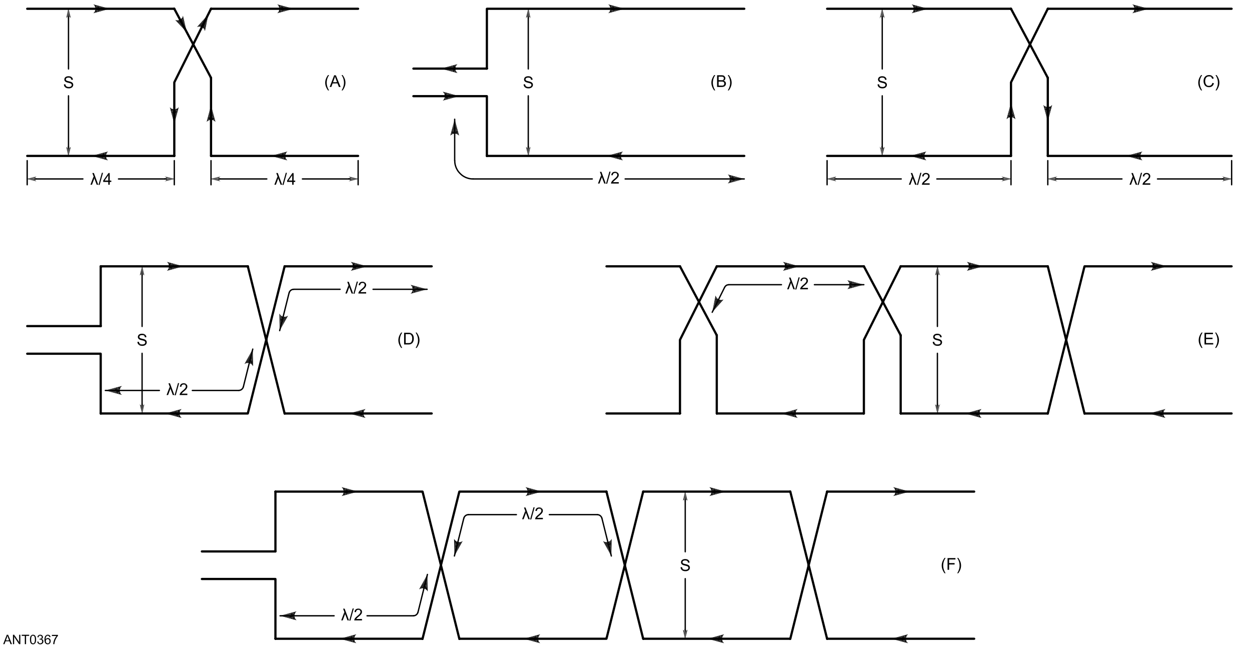
S
(A)
λ/4
λ/4
S
(B)
λ/2
S
(C)
λ/2
λ/2
λ/2
S
(D)
λ/2
λ/2
S
(E)
λ/2
λ/2
S
(F)
λ/2
ANT0367

Fig 74—Six other variations of W8JK “flat-top beam” antennas.

FOUR-ELEMENT END-FIRE AND COLLINEAR ARRAYS

The array shown in **Fig 76** combines collinear in-phase elements with parallel out-of-phase elements to give both broadside and end-fire directivity. It is a *two-section W8JK*. The approximate free-space gain using #12 copper wire is 4.9 dBi with λ/8 spacing and 5.4 dBi with λ/4 spacing. Directive patterns are given in **Figs 77** for free space, and in **Fig 78** for heights of 1 λ and λ/2 above flat ground.

The impedance between elements at the point where the phasing line is connected is of the order of several thousand ohms. The SWR with an unmatched line consequently is quite high, and this system should be constructed with open-wire line (500 or 600 Ω) if the line is to be resonant. With λ/4 element spacing the SWR on a 600 Ω line is estimated to be in the vicinity of 3 or 4:1.

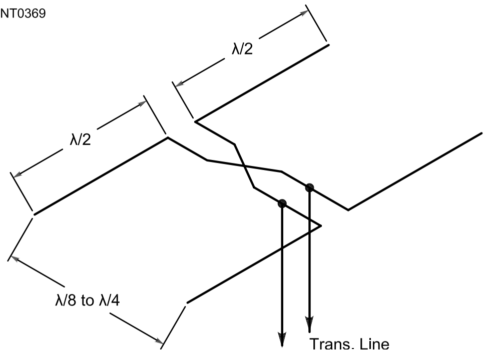

Fig 76—A 4-element array combining collinear broadside elements and parallel end-fire elements, popularly known as a two-section W8JK array.

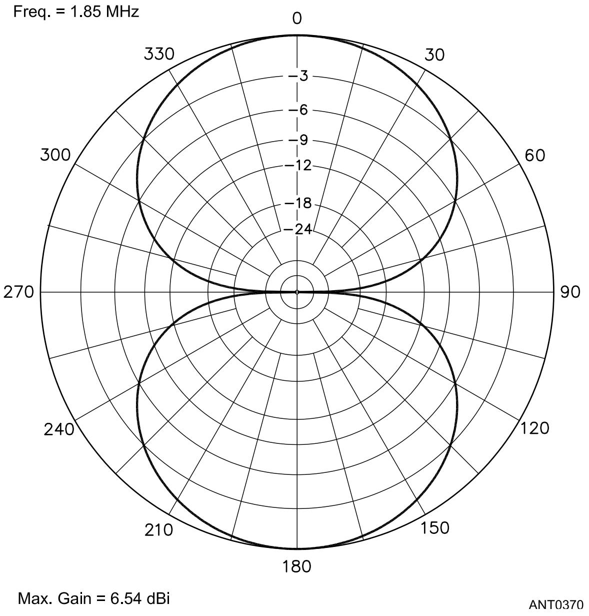

Fig 77—Free-space E-plane pattern for the antenna shown in Fig 76, with λ/8 spacing. The elements are parallel to the 90°-270° line in this diagram. Less than a 1° change in half-power beamwidth results when the spacing is changed from λ/8 to λ/4.

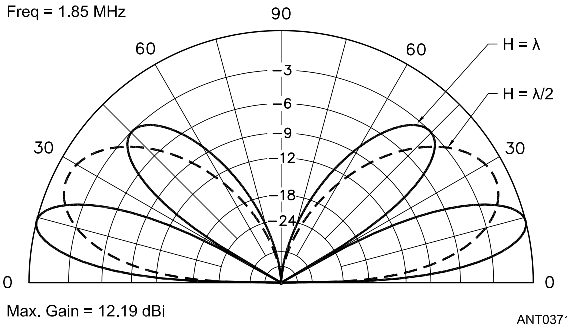

Fig 78—Elevation-plane pattern for the 4-element antenna of Fig 76 when mounted horizontally at two heights over flat ground. Solid line = 1 λ high; dashed line = λ/2 high.

To use a matched line, you could connect a closed stub 3λ/16 long at the transmission-line junction shown in Fig 76. The transmission line itself can then be tapped on this matching section at the point resulting in the lowest line SWR. This point can be determined by trial.

This type of antenna can be operated on two bands having a frequency ratio of 2 to 1, if a resonant feed line is used. For example, if you design for 28 MHz with λ/4 spacing between elements, you can also operate on 14 MHz as a simple 2-element end-fire array having λ/8 spacing.

Combination Driven Arrays

You can readily combine broadside, end-fire and collinear elements to increase gain and directivity, and this is in fact usually done when more than two elements are used in an array. Combinations of this type give more gain, in a given amount of space, than plain arrays of the types just described. Since the combinations that can be worked out are almost endless, this section describes only a few of the simpler types.

The accurate calculation of the power gain of a multi-element array requires a knowledge of the mutual impedances between all elements, as discussed in earlier sections. For approximate purposes it is sufficient to assume that each set (collinear, broadside, end-fire) will have the gains as given earlier, and then simply add up the gains for the combination. This neglects the effects of cross-coupling between sets of elements. However, the array configurations are such that the mutual impedances from cross-coupling should be relatively small, particularly when the spacings are λ/4 or more, so the estimated gain should be reasonably close to the actual gain. Alternatively, an antenna modeling program, such as *EZNEC*, can give good estimates of all parameters for a real-world antenna, providing that you take care to model all applicable parameters.

FOUR-ELEMENT DRIVEN ARRAYS

The array shown in **Fig 79** combines parallel elements with broadside and end-fire directivity. The smallest array (physically)—3λ/8 spacing between broadside and λ/8 spacing between end-fire elements—has an estimated gain of 6.5 dBi and the largest—3λ/4 and λ/4 spacing, respectively—about 8.4 dBi. Typical directive patterns for a λ/4 × λ/2 array are given in **Figs 80** and **81**.

The impedance at the feed point will not be purely resistive unless the element lengths are correct and the phasing lines are exactly λ/2 long. (This requires somewhat less than λ/2 spacing between broadside elements.) In this case the impedance at the junction is estimated to be over 10 kΩ. With other element spacings the impedance at the junction will be reactive as well as resistive, but in any event the SWR will be quite large. An open-wire line can be used as a resonant line, or a matching section may be used for non-resonant operation.

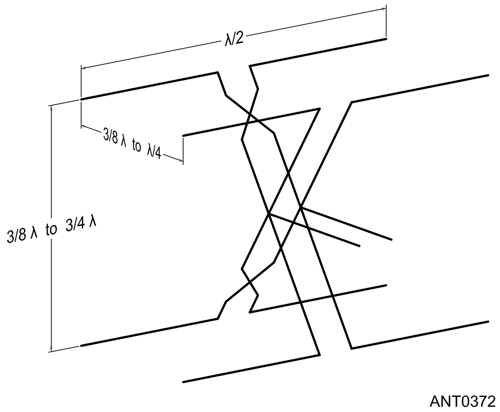

Fig 79—Four-element array combining both broadside and end-fire elements.

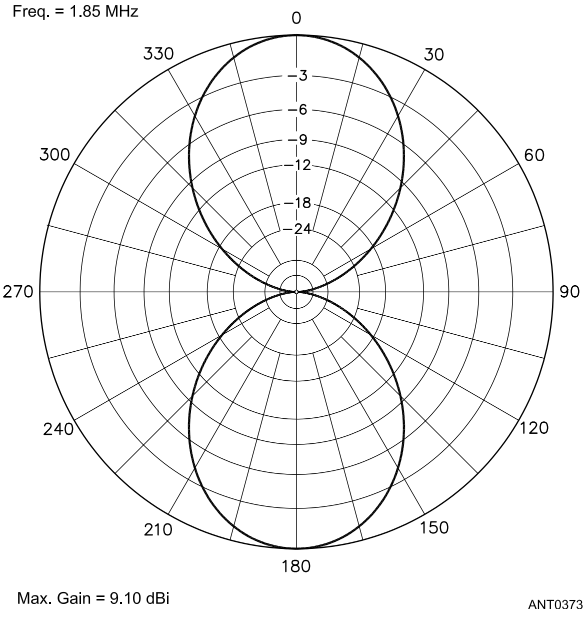

Fig 80—Free-space H-plane pattern of the 4-element antenna shown in Fig 79.

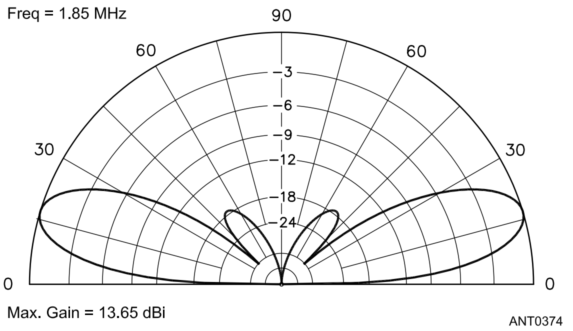

Fig 81—Vertical pattern of the antenna shown in Fig 79 at a mean height of 3λ/4 (lowest elements λ/2 above flat ground) when the antenna is horizontally polarized. For optimum gain and low wave angle the mean height should

EIGHT-ELEMENT DRIVEN ARRAYS

The array shown in **Fig 82** is a combination of collinear and parallel elements in broadside and end-fire directivity. Common practice in a wire antenna is to use λ/2 spacing for the parallel broadside elements and λ/4 spacing for the end-fire elements. This gives a free-space gain of about 9.1 dBi. Directive patterns for an array using these spacings are similar to those of Figs 80 and 81, but are somewhat sharper.

The SWR with this arrangement will be high. Matching stubs are recommended for making the lines non-resonant. Their position and length can be determined as described in Chapter 26.

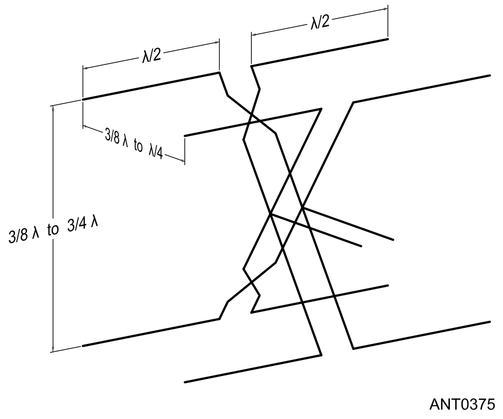

Fig 82—Eight-element driven array combining collinear and parallel elements for broadside and end-fire directivity.

This system can be used on two bands related in frequency by a 2-to-1 ratio, providing it is designed for the higher of the two, with 3λ/4 spacing between the parallel broadside elements and λ/4 spacing between the end-fire elements. On the lower frequency it will then operate as a 4-element antenna of the type shown in Fig 79, with 3λ/8 broadside spacing and λ/8 end-fire spacing. For two-band operation a resonant transmission line must be used.

PHASING ARROWS IN ARRAY ELEMENTS

In the antenna diagrams of preceding sections, the relative direction of current flow in the various antenna elements and connecting lines was shown by arrows. In laying out any antenna system it is necessary to know that the phasing lines are properly connected; otherwise the antenna may have entirely different characteristics than anticipated. The phasing may be checked either on the basis of current direction or polarity of voltages. There are two rules to remember:

1) In every λ/2 section of wire, starting from an open end, the current directions reverse. In terms of voltage, the polarity reverses at each λ/2 point, starting from an open end.
2) Currents in transmission lines always must flow in opposite directions in adjacent wires. In terms of voltage, polarities always must be opposite.

Examples of the use of current direction and voltage polarity are given at A and B, respectively, in **Fig 83**. The λ/2 points in the system are marked by small circles. When current in one section flows toward a circle, the current in the next section must also flow toward it, and vice versa. In the 4-element antenna shown at A, the current in the upper right-hand element cannot flow toward the transmission line because then the current in the right-hand section of the phasing line would have to flow upward and thus would be flowing in the same direction as the current in the left-hand wire. The phasing line would simply act like two wires in parallel in such a case. Of course, all arrows in the drawing could be reversed, and the net effect would be unchanged.

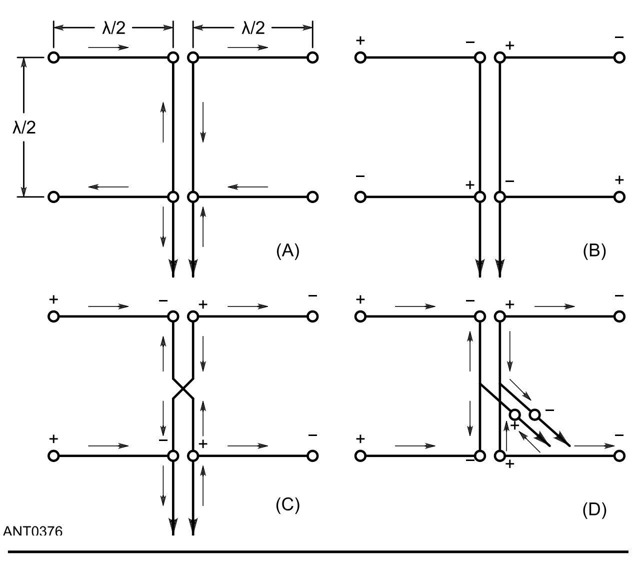

Fig 83—Methods of checking the phase of currents in elements and phasing lines.

C shows the effect of transposing the phasing line. This transposition reverses the direction of current flow in the lower pair of elements, as compared with A, and thus changes the array from a combination collinear and end-fire arrangement into a collinear-broadside array.

The drawing at D shows what happens when the transmission line is connected at the center of a section of phasing line. Viewed from the main transmission line, the two parts of the phasing line are simply in parallel, so the half wavelength is measured from the antenna element along the upper section of phasing line and thence along the transmission line. The distance from the lower elements is measured in the same way. Obviously, the two sections of phasing line should be the same length. If they are not, the current distribution becomes quite complicated; the element currents are neither in-phase nor 180° out-of-phase, and the elements at opposite ends of the lines do not receive the same current. To change the element current phasing at D into the phasing at A, simply transpose the wires in one section of the phasing line. This reverses the direction of current flow in the antenna elements connected to that section of phasing line.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

D. W. Atchley, H. E. Stinehelfer, and J. F. White, "360°-Steerable Vertical Phased Arrays," *QST*, Apr 1976, pp 27-30.

G. H. Brown, "Directional Antennas," *Proc. IRE*, Vol 25, No. 1, Jan 1937, pp 78-145.

G. H. Brown, R. F. Lewis and J. Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proc. IRE*, Jun 1937, pp 753-787.

G. H. Brown and O. M. Woodward, Jr., "Experimentally Determined Impedance Characteristics of Cylindrical Antennas," *Proc. IRE*, Apr 1945.

L. B. Cebik, "Two Limitations of NEC-4", **www.cebik.com/model/fd.html**.

A. Christman, "Feeding Phased Arrays: An Alternate Method," *Ham Radio*, May 1985, pp 58-59, 61-64.

J. Devoldere, *ON4UN's Low-Band DXing*, 4th ed. (Newington, CT: ARRL, 2005).

EZNEC is an antenna-modeling computer program for Microsoft Windows. See **eznec.com** for full information.

F. Gehrke, "Vertical Phased Arrays," in six parts, *Ham Radio*, May-Jul, Oct and Dec 1983, and May 1984.

C. Harrison, Jr, and R. King, "Theory of Coupled Folded Antennas," *IRE Trans on Antennas and Propagation*, Mar 1960, pp131-135.

W. Hayward and D. DeMaw, *Solid State Design for the Radio Amateur* (Newington, CT: ARRL, 1977).

W. Hayward, *Radio Frequency Design* (Newington, CT: ARRL, 1994).

H. Jasik, *Antenna Engineering Handbook*, 1st ed. (New York: McGraw-Hill, 1961). Later editions are edited by Richard C. Johnson.

R. King and C. Harrison, Jr, "Mutual and Self-Impedance for Coupled Antennas," *Journal of Applied Physics* , Vol 15, Jun 1944, pp 481-495.

R. King, "Self- and Mutual Impedances of Parallel Identical Antennas," *Proc. IRE*, Aug 1952, pp 981-988.

R. W. P. King, *Theory of Linear Antennas* (Cambridge, MA: Harvard Univ Press, 1956), p 275ff.

H. W. Kohler, "Antenna Design for Field-Strength Gain," *Proc. IRE*, Oct 1944, pp 611-616.

J. D. Kraus, "Antenna Arrays with Closely Spaced Elements," *Proc. IRE*, Feb, 1940, pp 76-84.

J. D. Kraus, *Antennas*, 2nd ed. (New York: McGraw-Hill Book Co., 1988).

Johnson, Richard C., *Antenna Engineering Handbook*, 3rd ed. (New York: McGraw-Hill Inc., 1993). This is a later edition of the volume by the same name edited by H. Jasik.

E. A. Laport, *Radio Antenna Engineering* (New York: McGraw-Hill Book Co, 1952).

J. L. Lawson, "Simple Arrays of Vertical Antenna Elements," *QST*, May 1971, pp 22-27.

P. H. Lee, N6PL *The Amateur Radio Vertical Antenna Handbook*, 2nd ed. (Hicksville, NY: CQ Publishing, Inc., 1984).

R. W. Lewallen, "Baluns: What They Do and How They Do It," *The ARRL Antenna Compendium, Vol 1* (Newington: ARRL, 1985). Also available for viewing at **eznec.com/Amateur/Articles/Baluns.pdf**.

R. Lewallen, "The Impact of Current Distribution on Array Patterns," Technical Correspondence, *QST*, Jul 1990, pp 39-40. Also available for viewing at **eznec.com/Amateur/Articles/Current_Dist.pdf**.

R. Lewallen, "*MININEC*—The Other Edge of the Sword", *QST*, Feb 1991, pp 18-22. *ELNEC*, referenced in the article, is no longer available.

M. W. Maxwell, "Some Aspects of the Balun Problem," *QST*, Mar 1983, pp 38-40.

J. Sevick, "The Ground-Image Vertical Antenna," *QST*, Jul 1971, pp 16-19, 22.

J. Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, Mar 1973, pp 13-28,41.

E. J. Wilkinson, "An N-Way Hybrid Power Divider," *IRE Transactions on Microwave Theory and Techniques*, Jan, 1960.

Radio Broadcast Ground Systems, available from Smith Electronics, Inc, 8200 Snowville Rd, Cleveland, OH 44141.

Appendix A—*EZNEC-ARRL* Examples

This appendix contains step-by-step procedures using *EZNEC-ARRL* (included on the *Antenna Book* CD) to illustrate various topics discussed in the main chapter. A standard *EZNEC* program type of v. 4.0 or later may also be used. Different versions, program types and calculating engines may give results that are slightly different from those shown in the examples. However, any differences should be insignificantly small.

EZNEC Example—Mutual Coupling

Illustrates the effect of mutual coupling on feed-point impedance. Open the **ARRL_Cardioid.EZ** file, which is mounted over "perfect" ground. Click the **View Ant** button to see a diagram of the antenna, a 2-element array of vertical elements. Click on the **Wires** line in the main window to open the Wires Window. Click the button at the left of the Wire 2 line, and then press the **Delete** key on your keyboard to delete wire #2. After clicking **Ok**, note that one of the verticals has disappeared from the View Antenna display, leaving a single element. Click **Src Dat** and note that the feed-point impedance of this single vertical is about $37 + j\,1\ \Omega$--it's very nearly resonant.

Next, in the Wires Window, open the **Edit** menu at the top and click **Undo delete wire(s)** to restore the second element. Click **Src Dat** again and notice that the feed-point impedance of wire #1 is now about $21 - j\,19\ \Omega$. The feed-point impedance of the second element, which is identical to the first, is about $52 + j\,21\ \Omega$. This difference, and the change from the self-impedance of $37 + j\,1\ \Omega$, is due to mutual coupling. As you see, it's not at all a minor effect.

As an additional exercise, change the magnitude or phase angle of the source at the base of wire #2 (click **Sources** in the main window), and see how this changes the feed-point impedances of both elements. You should be able to confirm each of the four points enumerated in the **MUTUAL COUPLING** section.

EZNEC Example—Nulls

Illustrates the effect of current magnitude on nulls and gain. Again, open the **ARRL_Cardioid.EZ** file. Click the **FF Plot** button to generate the azimuth pattern of an ideal array. Save the plot for future reference as follows: In the plot window, open the **File** menu and select **Save Trace As**. Enter the name **Cardioid** and click **Save**. Now, in the main window click on the **Sources** line to open the Sources Window. Change the magnitude of source 1 from 1 to 1.1, and of source 2 from 1 to 0.9 and press **Enter** on your keyboard so that *EZNEC-ARRL* will accept the last change.

Click **FF Plot** to generate a pattern with the new currents. In the plot window, open the File menu and select **Add**

Trace. Enter the name **Cardioid** and click Open. You should now see the original plot and new plot overlaid. Notice that the null is much less deep with the altered currents, but the forward patterns are nearly identical. By clicking on the names of the traces, **Primary** and **Cardioid**, you can see in turn the gain and front-to-back ratio of each of the traces. The original, **Cardioid**, has a front-to-back ratio of about 32 dB, while Primary, the new plot, has a ratio of about 22.5 dB. The forward gain, however, differs by only 0.02 dB, a completely insignificant amount.

EZNEC Example—"Phasing-Line" Feed

Illustrates the effect of using a "phasing-line" feed. Open the **ARRL_CardTL.EZ** file. This is a model of an array fed with transmission lines whose lengths were designed using the *Arrayfeed1* program to take into account the actual load impedances of elements in a phased array. This model is mounted over "perfect" ground.

Click the **View Ant** button to show the array. Note that the lengths of the lines from the source (circle) to the elements don't represent the actual physical lengths of the lines. In the main window, click on the **Trans Lines** line to open the Transmission Lines window. In it you can see that the lengths of the feed lines, both of which are connected to the same source, are about 81° and 155°, a difference of 74° rather than 90°.

In the main window, click the **Currents** button and take a look at the current shown for segment 1 of wires 1 and 2. These are the currents at the element feed points. The ratio of the magnitude of currents is 4.577/4.561 = 1.003, and the phase difference is –56.3°–(–147.5°) = 91.2°. (A more accurate determination of feed-line lengths with program *Arrayfeed1* gives lengths of 80.61° and 153.70°, resulting in a current ratio of 1.000 at a phase of 90.02°. But the resulting pattern is very nearly the same.) But let's see what happens when we make the lines exactly 90° different in length.

First, click the **FF Plot** button to generate the azimuth pattern of the original model. Save the plot for future reference as follows: In the plot window, open the **File** menu and select **Save Trace As**. Enter the name **CardTL** and click **Save**. Now in the Transmission Lines Window, change the length of line number 1 from 80.56° to 90°. *Important:* In the line 1 Length box, enter **90d** to make the line 90° long. If you omit the **d**, it will become 90 meters long! Similarly, change the length of line 2 to 180° by entering **180d** in the line 2 Length box, then press **Enter** on your keyboard so that *EZNEC* will accept the last change.

Click **FF Plot** to generate a pattern with the new line lengths. In the plot window, open the **File** menu and select **Add Trace**. Enter the name **CardTL** and click **Open**. You should now see the original plot and new plot overlaid together. Notice that the gain of the modified model is about 1 dB greater than the original, but the front-to-back ratio has deteriorated to about 10 dB.

Experiment with different combinations of line lengths that differ by 90°--for example, 45° and 135° (don't forget the '**d**'!), or change the impedance of one or both lines and you'll see that you can get a wide variety of patterns. None, however, are likely to be as close to the ideal cardioid pattern as the original.

Chapter 9

Broadband Antenna Matching

Antenna systems that provide a good impedance match to the transmitter over a wide frequency range have been a topic of interest to hams for many years. Most emphasis has been focused on the 80-meter band since a conventional half-wave dipole will provide better than 2:1 SWR over only about one-third of the 3.5 to 4.0 MHz band. The advantage of a broadband match is obvious—fewer adjustments during tune-up and an antenna tuner may not be required.

This chapter was written by Frank Witt, AI1H, who has written numerous articles in *QST* and in *The ARRL Antenna Compendium* series on this subject. See the Bibliography for details.

The term *broadband antennas* has frequently been used to describe antenna systems that provide a wideband impedance match to the transmitter. This is something of a misnomer, since most antennas are good radiators over a wide range of frequencies and are therefore "broadband antennas" by definition. The problem is getting the energy to the antenna so it can be radiated. Antenna tuners solve this problem in some cases, although losses in transmission lines, baluns and the antenna tuner itself can be excessive. Also worthy of mention is that antenna directional properties are usually frequency dependent. In this chapter, we discuss broadbanding the impedance match to the transmitter only.

GENERAL CONCEPTS

The Objective

Fig 1 shows a simplified system: the transmitter, an SWR meter and the transmitter load impedance. Modern transceivers are designed to operate properly into a 50-Ω load and they will deliver full power into the load impedance, at their rated level of distortion, when the SWR is less than about 1.5:1. Loads beyond this limit may cause the transceiver to protect itself by lowering the output power.

Many transceivers have built-in automatic antenna tuners that permit operation for loads outside the 1.5:1 SWR range, but the matching range is often limited, particularly on the lower-frequency bands. In practice, barefoot operation is simplified if the load SWR is held to less than 2:1.

For high-power amplifiers, it is also best to keep the load SWR to less than about 2:1, since output tuning components are commonly rated to handle such loads. You can see that the primary function of the SWR meter in Fig 1 is to measure the suitability of the load impedance so far as the transmitter is concerned. Henceforth, we will use the term *load SWR* as a description of the transmitter's load.

The SWR meter is actually reading a circuit condition at a single point in the system. In fact, the meter really measures magnitude of the reflection coefficient, but is calibrated in SWR. The relationships between the complex load impedance, Z_L, the magnitude of the reflection coefficient, $|\rho|$, and the load SWR are as follows:

$$|\rho| = \left| \frac{Z_L - 50}{Z_L + 50} \right| \quad \text{(Eq 1)}$$

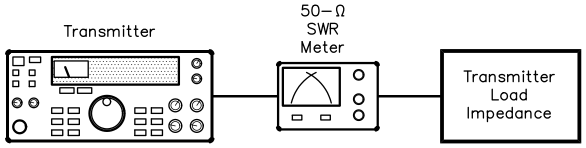

Fig 1—Basic antenna system elements at the output of a transmitter.

$$\text{SWR} = \frac{1+|\rho|}{1-|\rho|} \qquad \text{(Eq 2)}$$

where 50 Ω is used as the reference impedance, since the design load impedance for the transmitter is assumed to be 50 Ω.

An important point is that you need not know the output impedance of the transmitter to design a broadband matching network. The issue of the actual value of the output impedance of typical RF power amplifiers is a matter of continuing controversy in amateur circles. Fortunately, this issue is not important for the design of broadband matching networks, since the load SWR is independent of the output impedance of the transmitter. Our objective is to design the matching network so that the load SWR (with a 50-Ω reference resistance) is less than some value, say 2:1, over as wide a band as possible.

Resonant Antennas

The broadband matching techniques described here apply to antennas operating near resonance. Typical resonant antennas include half-wave dipoles, quarter-wave verticals over a ground plane and full-wave loops. To design a broadband matching network, you must know the antenna feed-point impedance near resonance. **Fig 2** shows the antenna-impedance equivalent circuit near resonance. Although the series RLC circuit is an approximation, it is good enough to allow us to design matching networks that significantly increase the band over which a good match to the transmitter is achieved.

Note that the impedance is defined by F_0, the resonant frequency, R_A, the antenna resistance at resonance, and Q_A, the antenna Q. R_A is actually the sum of the radiation resistance and any loss resistance, including conductor losses and losses induced by surrounding objects, such as the ground below the antenna. R_A is frequency dependent, but it is sufficient to assume it is fixed during the matching network design process. Minor adjustments to the matching network will correct for the frequency dependence of R_A.

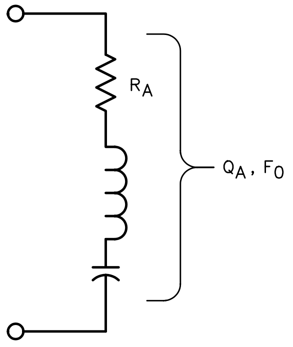

Fig 2—The equivalent circuit of a resonant antenna. The simple series RLC approximation applies to many resonant antenna types, such as dipoles, monopoles, and loops. R_A, Q_A and F_0 are properties of the entire antenna, as discussed in the text.

The antenna resistance and Q depend on the physical properties of the antenna itself, the properties of ground and the height above ground. Consider as an example an 80-meter horizontal half-wave dipole made from #12 wire located over average ground (dielectric constant = 13 and conductivity = 5 mS/m). **Figs 3** and **4** show how feed-point resistance and Q vary with height. It is clear from these figures that there are wide gyrations in antenna parameters. The better these parameters are known, the more successful we will be at designing the broadband matching network. For horizontal dipoles, Figs 3 and 4 may be used to get a good idea of antenna resistance and Q, so long as the height is scaled in proportion to wavelength. For example, a 160-meter dipole at a height of 100 feet would have about the same resistance and Q as an 80-meter dipole at 50 feet.

For optimum results, the resistance and Q can be

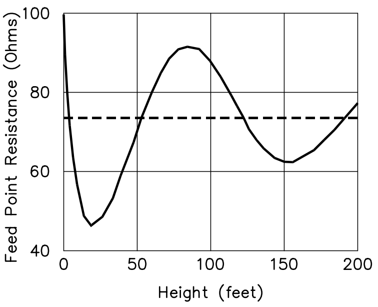

Fig 3—Feed-point resistance (solid line) for 80-meter horizontal dipole at resonance versus height over ground. The free-space value is shown as a dashed line.

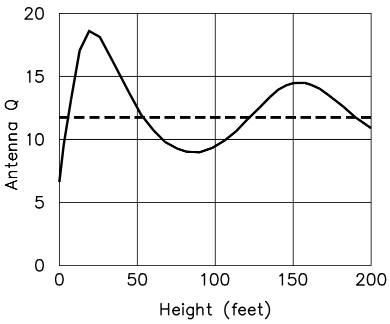

Fig 4—Antenna Q for 80-meter horizontal dipole (solid line) versus height. The free-space value is shown as a dashed line.

determined from computer simulation (using a program like *EZNEC,* for example) or measurement (using a low power SWR/Z meter such as the MFJ-259B or the Autek Research VA1 or RF-1). R_A can be computed using SWR measurements at resonance. **Fig 5** shows the typical bowl-shaped SWR curve as a function of frequency for a resonant antenna. S_0 is the SWR at resonance. You must take into account the loss of the feed line if the measurement must be made at the end of a long 50-Ω transmission line. If the line loss is low or if the measurement is made at the antenna terminals, the formulas given below apply.

For $R_A > 50\ \Omega$

$$R_A = S_0 \times 50 \qquad \text{(Eq 3)}$$

For $R_A < 50\ \Omega$

$$R_A = \frac{50}{S_0} \qquad \text{(Eq 4)}$$

There is an ambiguity, since we do not know if R_A is greater than or less than 50 Ω. A simple way of resolving this ambiguity is to take the same SWR measurement at resonance with a 10-Ω non-inductive resistor added in series with the antenna impedance. If the SWR goes up, $R_A > 45\ \Omega$. If the SWR goes down, $R_A < 45\ \Omega$. This will resolve the ambiguity.

Q_A may be determined by measuring the SWR bandwidth. Define BW_2 as the bandwidth over which the resonant antenna SWR is less than 2:1 (in the same units as F_0). See Fig 5.

For $R_A > 50\ \Omega$

$$Q_A = \frac{F_0}{BW_2}\sqrt{2.5 S_0 - S_0^{\,2} - 1} \qquad \text{(Eq 5)}$$

For $R_A < 50\ \Omega$

$$Q_A = \frac{F_0}{BW_2 \times S_0}\sqrt{2.5 S_0 - S_0^{\,2} - 1} \qquad \text{(Eq 6)}$$

Again, you must take into account the loss of the feed line if you make the measurement at the end of a long 50-Ω transmission line.

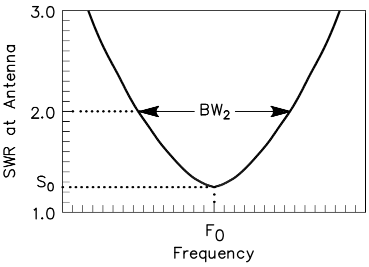

Fig 5—SWR at the antenna versus frequency in the vicinity of resonance.

For horizontal and inverted-V half-wave dipoles, a very good approximation for the antenna Q when R_A is known is given by:

$$Q_A = \frac{93.9\left[\ln\frac{8110}{D F_0} - 1\right]}{R_A} \qquad \text{(Eq 7)}$$

where D = the diameter of wire, in inches.

Loss

When you build a resonant antenna with some matching scheme designed to increase the bandwidth, you might overlook the loss introduced by the broadband matching components. Loss in dB is calculated from:

$$\text{Loss} = -10\log\frac{\text{Power radiated by antenna}}{\text{Total power from transmitter}} \qquad \text{(Eq 8)}$$

or, alternatively, define efficiency in % as

$$\text{Efficiency} = \frac{100 \times \text{Power radiated by antenna}}{\text{Total power from transmitter}} \qquad \text{(Eq 9)}$$

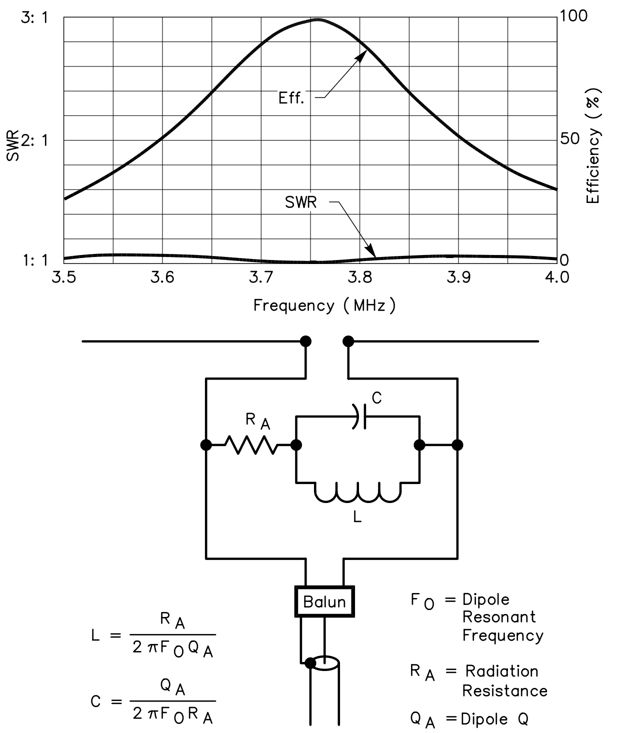
3: 1
2: 1
1: 1
SWR
100
50
0
Efficiency (%)
Eff.
SWR
3.5
3.6
3.7
3.8
3.9
4.0
Frequency (MHz)
C
R_A
L
Balun
$L = \frac{R_A}{2\pi F_0 Q_A}$
$C = \frac{Q_A}{2\pi F_0 R_A}$
F_O = Dipole Resonant Frequency
R_A = Radiation Resistance
Q_A = Dipole Q

Fig 6—Matching the dipole with a complementary RLC network greatly improves the SWR characteristics, nearly 1:1 across the 3.5-MHz band. However, the relative loss at the band edges is greater than 5 dB.

An extreme degree of bandwidth broadening of an 80-meter dipole is illustrated in **Fig 6**. This approach is not recommended, but it is offered here to make a point. The broadening is accomplished by adding resistive losses. From network theory we obtain the RLC (resistor, inductor, capacitor) matching network shown. The network provides the complement of the antenna impedance. Note that the SWR is virtually 1:1 over the entire band (and beyond), but the efficiency falls off dramatically away from resonance.

The band-edge efficiency of 25 to 30% shown in Fig 6 means that the antenna has about 5 to 6 dB of loss relative to an ideal dipole. At the band edges, 70 to 75% of the power delivered down the transmission line from the transmitter is heating up the matching-network resistor. For a 1-kW output level, the resistor must have a power rating of at least 750 W! Use of an RLC complementary network for broadbanding is not recommended, but it does illustrate how resistance (or losses) in the matching network can significantly increase the apparent antenna SWR bandwidth.

The loss introduced by any broadband matching approach must be taken into consideration. Lossy matching networks will usually provide more match bandwidth improvement than less lossy ones.

SIMPLE BROADBAND MATCHING TECHNIQUES

The Cage Dipole

You can increase the match bandwidth of a single-wire dipole by using a thick radiator, one with a large diameter. The gain and radiation pattern are essentially the same as that of a thin-wire dipole. The radiator does not necessarily have to be solid; open construction such as shown in **Fig 7** may be used.

The theoretical SWR response of an 80-meter cage dipole having a 6-inch diameter is shown in **Fig 8**. BW_2 for this antenna fed with 50-Ω line is 287 kHz, and the Q is approximately 8. Its 2:1-SWR frequency range is 1.79 times broader than a dipole with a Q of 13, typical for thin-wire dipoles.

There are other means of creating a thick radiator, thereby gaining greater match bandwidth. The bow-tie and the fan dipole make use of the same Q-lowering principle as the cage for increased match bandwidth. The broadbanding techniques described below are usually more practical than the unwieldy cage dipole.

Stagger-Tuned Dipoles

A single-wire dipole exhibits a relatively narrow bandwidth in terms of coverage for the 3.5 to 4.0-MHz band. A technique that has been used for years to cover the entire band is to use two dipoles, one cut for the CW portion and one for the phone portion. The dipoles are connected in parallel at the feed point and use a single feeder. This technique is known as *stagger tuning*.

Fig 9 shows the theoretical SWR response of a pair of stagger-tuned dipoles fed with 50-Ω line. No mutual coupling between the wires is assumed, a condition

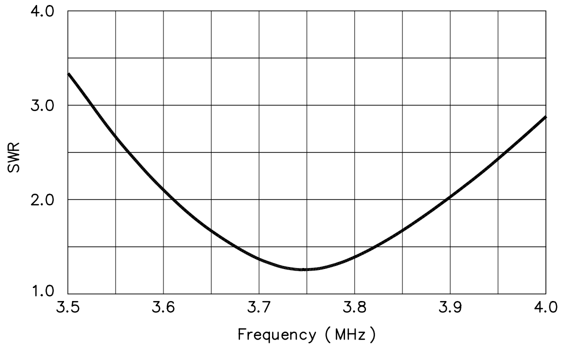

Fig 8—Theoretical SWR versus frequency response for a cage dipole of length 122 feet 6 inches and a spreader diameter of 6 inches, fed with 50-Ω line. The 2:1-SWR bandwidth frequencies are 3.610 and 3.897 MHz, with a resulting BW_2 of 287 kHz.

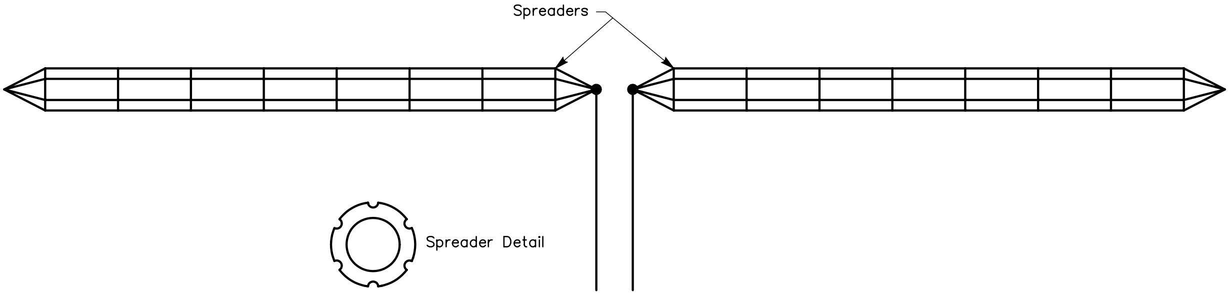

Fig 7—Construction of a cage dipole, which has some resemblance to a round birdcage. The spreaders need not be of conductive material, and should be lightweight. Between adjacent conductors, the spacing should be 0.02 λ or less. The number of spreaders and their spacing should be sufficient to maintain a relatively constant separation of the radiator wires.

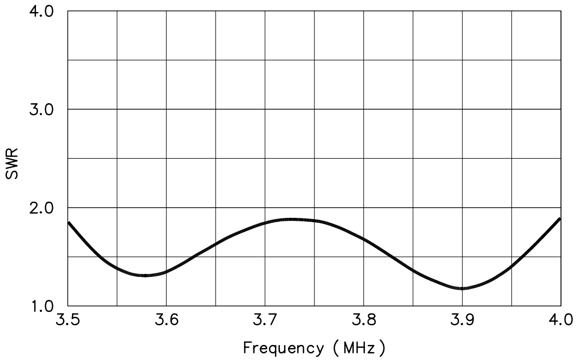

Fig 9—Theoretical SWR response of two stagger-tuned dipoles. They are connected in parallel at the feed point and fed with 50-Ω line. The dipoles are of wire such as #12 or #14, with total lengths of 119 and 132 feet.

that would exist if the two antennas were mounted at right angles to one another. As Fig 9 shows, the SWR response is less than 1.9:1 across the entire band.

A difficulty with such crossed dipoles is that four supports are required for horizontal antennas. A more common arrangement is to use inverted V dipoles with a single support, at the apex of each element. The radiator wires can also act as guy lines for the supporting mast.

When the dipoles are mounted at something other than a right angle, mutual coupling between them comes into play. This causes interaction between the two elements—tuning of one by length adjustment will affect the tuning of the other. The interaction becomes most critical when the two dipoles are run parallel to each other, suspended by the same supports, and the wires are close together. Finding the optimum length for each dipole for total band coverage can become a tedious and frustrating process.

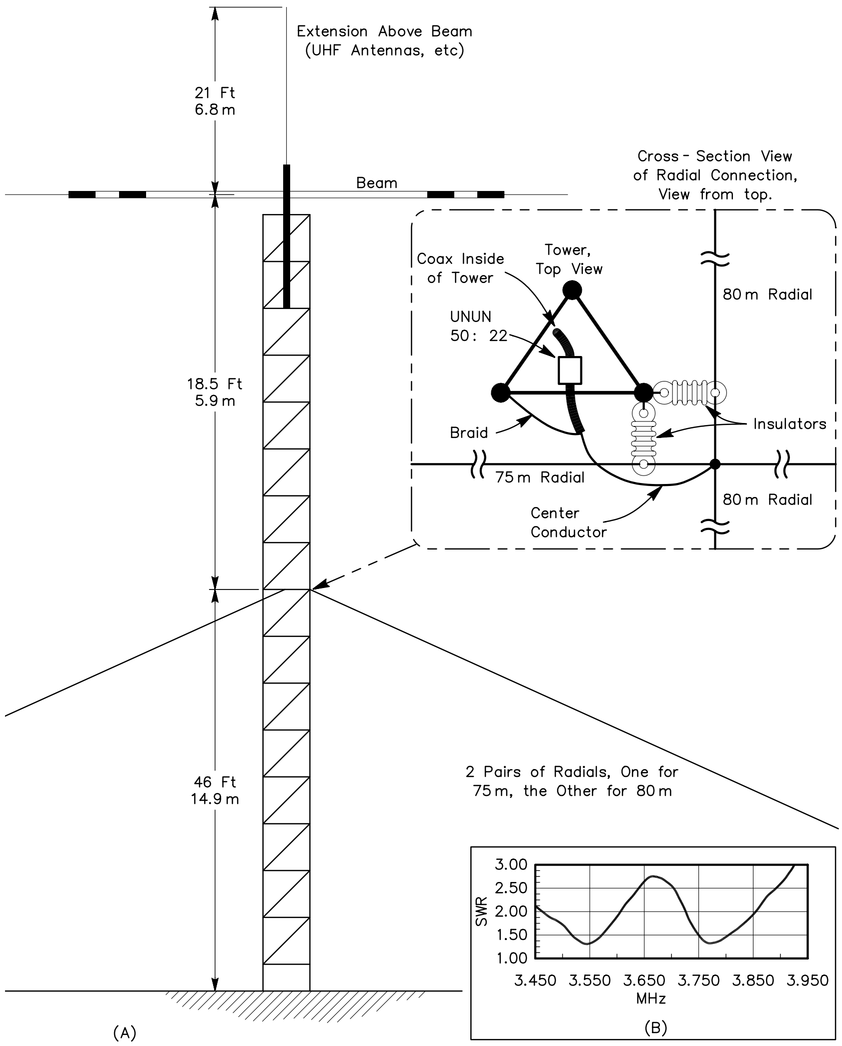
Extension Above Beam
(UHF Antennas, etc)
21 Ft
6.8 m
Beam
18.5 Ft
5.9 m
46 Ft
14.9 m
Cross - Section View
of Radial Connection,
View from top.
Coax Inside
of Tower
Tower,
Top View
UNUN
50 : 22
80 m Radial
Braid
Insulators
75 m Radial
80 m Radial
Center
Conductor
2 Pairs of Radials, One for
75 m, the Other for 80 m
3.00
2.50
2.00
1.50
1.00
SWR
3.450
3.550
3.650
3.750
3.850
3.950
MHz
(A)
(B)

Fig 10—At A, details of the inverse-fed ground-plane at W4PK. The inset shows the feed-point details. Keep the two sets of radials as close to 90· as possible. Don't try to resonate the antenna by adjusting radial lengths. At B, SWR curve of the inverse-fed ground plane antenna with stagger-tuned radials. The objective is to have a low SWR for both the CW and SSB 80/75-meter DX windows.

Stagger-Tuned Radials

A variation of the use of stagger tuning has been applied by Sam Leslie, W4PK. See his article entitled “Broadbanding the Elevated, Inverse-Fed Ground Plane Antenna” in *The ARRL Antenna Compendium, Vol 6*. An existing tower with a large beam and VHF/UHF antennas on top is used as an 80-meter monopole. See **Fig 10A**. The feed point is part way up the tower, at a point where the metal above it makes up an electrical quarter wavelength. Four elevated quarter-wave radials are used as a ground plane. The radials droop away from the tower. They are joined together at the tower but not connected to the tower. The antenna is fed with 50-Ω coax and an impedance step-down autotransformer (50:22 Ω unun) located at the tower. The shield of the coax is carried through the transformer and connected to the tower. The coax “hot side” is connected to the junction of the radials. A Q-section made from two paralleled RG-59 coax cables (Z_0 = 75/2 = 37.5 Ω) may be used instead of the autotransformer.

The broadband match is achieved by cutting two opposing radials for one end of the desired match range (75 meters) and cutting the other two for the other end of the range (80 meters). See Fig 10B for the resultant SWR versus frequency characteristic.

FEED-LINE IMPEDANCE MISMATCHING

The simplest broadband matching network is a transformer at the junction of the transmission line and the antenna. See Frank Witt’s article entitled “Match Bandwidth of Resonant Antenna Systems” in October 1991 *QST*. Observe that the impedance of the series RLC resonant antenna model of Fig 2 increases at frequencies away from resonance. The result is that more bandwidth is achieved when SWR at resonance is not exactly 1:1. For example, if the antenna is fed with a low-loss 50-Ω transmission line, the maximum SWR bandwidth is obtained if the effective R_A = 50/1.25 = 40 Ω. The improvement in BW_2, the 2:1 SWR bandwidth, over a perfect match at resonance is only 6%, however. This condition of R_A = 40 Ω can be achieved by installing a transformer at the junction of the feed line and the antenna.

If the line loss is larger, you can gain more benefit in terms of bandwidth by deliberately mismatching at the antenna. If the matched-line loss is 2 dB, the improvement rises to 18%. To achieve this requires an R_A = 28.2 Ω. Again, the improvement over a perfect match at resonance is modest.

It is interesting to compare the 2:1 SWR bandwidth (at the transmitter end of the line) for the case of a lossless line versus the case of 2-dB matched-line loss. BW_2 for the lossy line case is 1.95 times that of the lossless line case, so substantial match bandwidth improvement occurs, but at the cost of considerable loss. The band edge loss for the lossy-line case is 3 dB!

This example is provided mostly as a reference for comparison with the more desirable broadbanding techniques below. It also explains why some installations with long feed lines show large match bandwidths.

PARALLEL-TUNED CIRCUITS AT THE ANTENNA TERMINALS

As mentioned previously, a resonant antenna has an equivalent circuit that may be represented as a series RLC circuit. We can cancel out some of the antenna reactance away from resonance with a parallel-tuned circuit connected across the antenna terminals. The impedance level of the parallel-tuned circuit must be low enough to be effective and must have the same resonant frequency as the antenna. You can use a parallel-tuned circuit made with lumped LC components or with transmission-line segments. A quarter-wave transmission line with a short at the far end, or a half-wave line with an open at the far end, each behave like a parallel-tuned LC circuit.

The Double Bazooka

The response of the controversial *double bazooka* antenna is shown in **Fig 11**. This antenna actually consists of a dipole with two quarter-wave coaxial resonator stubs connected in series. The stubs are connected across the antenna terminals.

Not much bandwidth enhancement is provided by this resonator connection because the impedance of the matching network is too high. The antenna offers a 2:1-SWR bandwidth frequency range that is only 1.14 times that of a simple dipole with the same feeder. And the bandwidth enhancement is partially due to the “fat” antenna wires composed mostly of the coax shield. No

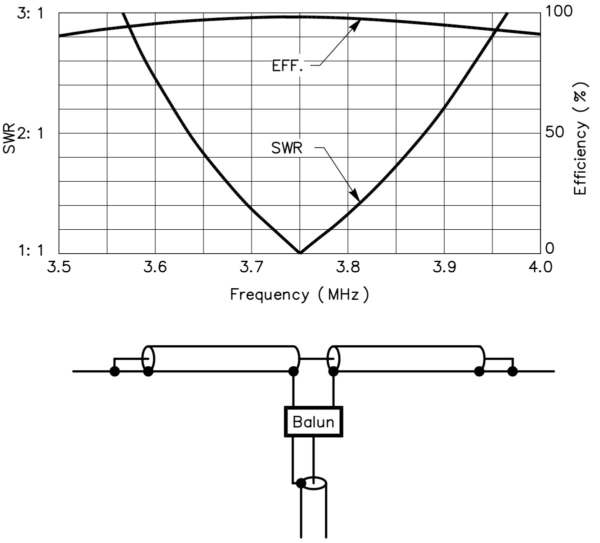

Fig 11—The double bazooka, sometimes called a coaxial dipole. The antenna is self-resonant at 3.75 MHz. The resonator stubs are 43.23-foot lengths of RG-58A coax.

improvement in antenna gain or pattern over a thin-wire dipole can be expected from this antenna.

The Crossed Double Bazooka

A modified version of the double-bazooka antenna is shown in **Fig 12**. In this case, the impedance of the matching network is reduced to one-fourth of the impedance of the standard double-bazooka network. The lower impedance provides more reactance correction, and hence increases the bandwidth frequency range noticeably, to 1.55 times that of a simple dipole. Notice, however, that the efficiency of the antenna drops to about 80% at the 2:1-SWR points. This amounts to a loss of approximately 1 dB. The broadbanding, in part, is caused by the losses in the coaxial resonator stubs (made of RG-58A coax), which have a remarkably low Q (only 20).

MATCHING NETWORK DESIGN

Optimum Matching

You can improve the match bandwidth by using a transformer or a parallel-tuned circuit at the antenna terminals. A combination of these two techniques can yield a very effective match bandwidth enhancement. Again, you can make the impedance transformation and implement the tuned circuit using lumped-circuit elements (transformers, inductors and capacitors) or with distributed-circuit elements (transmission lines).

In 1950, an article by R. M. Fano presented the theory of broadband matching of arbitrary impedances. This classic work was limited to lossless matching networks. Unfortunately, matching networks realized with transmission lines have enough loss to render the Fano results useful as only a starting point in the design process. A theory of optimum matching with lossy matching networks was presented in an article by Frank Witt entitled “Optimum Lossy Broadband Matching Networks for Resonant Antennas” that appeared in April 1990 *RF Design* and was summarized and extended in “Broadband Matching with the Transmission Line Resonator,” in *The ARRL Antenna Compendium Vol 4*.

Using a matching network, the 2:1 SWR bandwidth can be increased by a factor of about 2.5. Instead of the familiar bowl-shaped SWR versus frequency characteristic seen in Fig 5, the addition of the matching network yields the W-shaped characteristic of **Fig 13**. The maximum SWR over the band is S_M. The band edges are F_L and F_H. Thus the bandwidth, BW, and the center frequency, F_0, are

$$BW = F_H - F_L \qquad \text{(Eq 10)}$$

$$F_0 = \sqrt{F_H \times F_L} \qquad \text{(Eq 11)}$$

Fig 14 shows the pertinent equivalent circuit for a typical broadband antenna system. The parallel LC matching network is defined by its resonant frequency, F_0 (which is the same as the resonant frequency of the antenna), its Q and its impedance level. The impedance level is the reactance of either the matching network inductance or capacitance at F_0. The optimum transmitter load resistance,

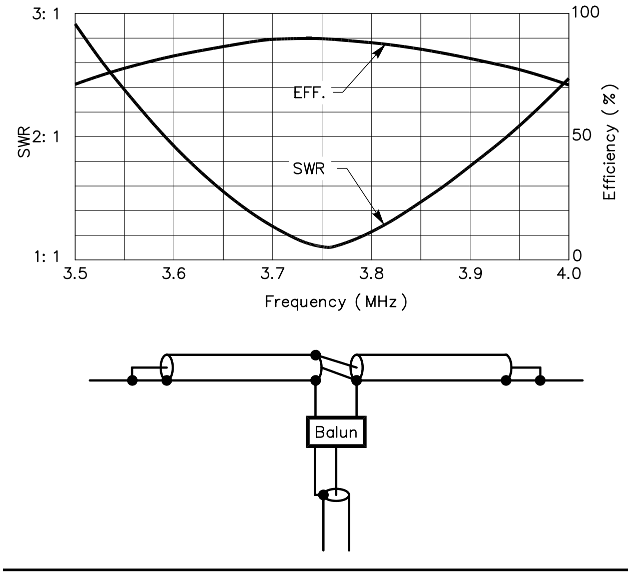

Fig 12—The crossed double bazooka yields bandwidth improvement by using two quarter-wave resonators, parallel connected, as a matching network. As with the double bazooka, the resonator stubs are 43.23-foot lengths of RG-58A coax.

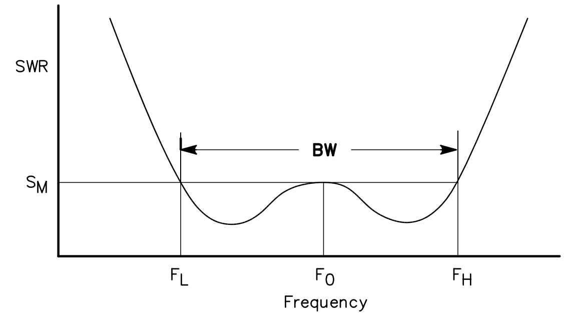

Fig 13—W-shaped SWR versus frequency characteristic that results when a transformer and parallel LC matching network are used.

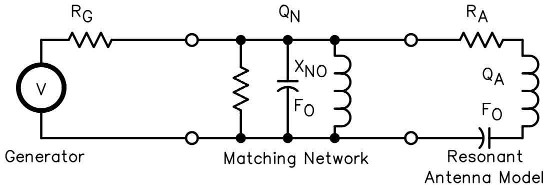

Fig 14—Assumed equivalent circuit for the broadband matching system. The antenna and the matching network have the same resonant frequency.

R_G, is achieved using an impedance transformer between the transmitter and the input to the parallel-tuned circuit. Thus for a transmitter whose optimum load resistance is 50 Ω, the required impedance transformer would have an impedance transformation ratio of 50:R_G.

The goal is to keep the SWR over the band as low as possible. There is a best or minimum value of S_M that may be achieved over the entire band. It depends on the desired match bandwidth and the Q's of the antenna and matching network and is given by the following formula:

$$S_{M\,min} = \frac{\sqrt{B_N^{\,2}+1} + \sqrt{B_N^{\,2} + 1 + \frac{2Q_A}{Q_N}\left(1+\frac{Q_A}{2\,Q_N}\right)}}{2\left(1+\frac{Q_A}{2\,Q_N}\right)} \qquad \text{(Eq 12)}$$

where $B_N = \frac{BW}{F_0}$ and

Q_A = antenna Q

Q_N = matching network Q.

The reactance at resonance of the parallel tuned LC matching network is given by:

$$X_{N0} = \frac{R_A}{Q_A}\left[\left(1+\frac{Q_A}{2Q_N}\right)S_{M\,min}^{\;\;2} - \frac{Q_A}{2Q_N}\right] \qquad \text{(Eq 13)}$$

where R_A = antenna resistance in Ω.

The transmitter load resistance is calculated from:

$$R_G = \frac{S_{M\,min}\,R_A\,Q_N\,X_{N0}}{R_A + Q_N\,X_{N0}} \qquad \text{(Eq 14)}$$

The loss of the matching network is always greatest at the edges of the band. This loss, L_{MNE}, in dB, is given by:

$$L_{MNE} = 10\log\left[1+\frac{R_A}{Q_N\,X_{N0}}\left(B_N^{\,2}+1\right)\right] \qquad \text{(Eq 15)}$$

Now you have all the information needed to design an optimal broadband matching network. An 80-meter dipole will serve as an example to illustrate the procedure. The following parameters are assumed:

F_L = 3.5 MHz
F_H = 4.0 MHz
R_A = 57.2 Ω
Q_A = 13
Q_N = 40.65

The values of R_A and Q_A are reasonable values for an 80-meter inverted V with an apex 60 feet above ground. Q_N = 40.65 is the calculated value of the Q of a resonator made from RG-213 or similar cable in the middle of the 80-meter band. The results are:

BW = 0.5 MHz (from Eq 10)
F_0 = 3.742 MHz (from Eq 11)
S_{Mmin} = 1.8 (from Eq 12)
X_{N0} = 15.9 Ω (from Eq 13)
R_G = 94.8 Ω (from Eq 14)
L_{MNE} = 1.32 dB (from Eq 15)

Shown in **Fig 15** is the plot of SWR and matching-network loss versus frequency for this example. The reference resistance for the SWR calculation is R_G or 94.8 Ω. Since the desired load resistance for the transmitter is 50 Ω, we need some means to transform from 94.8 Ω to 50 Ω. The examples given in the rest of this chapter will show a variety of ways of achieving the necessary impedance transformation.

The meaning of X_{N0} = 15.9 Ω is that the inductor and the capacitor in the matching network would each have a reactance of 15.9 Ω at 3.742 MHz. Methods for obtaining a particular value of X_{N0} are shown in the fol-

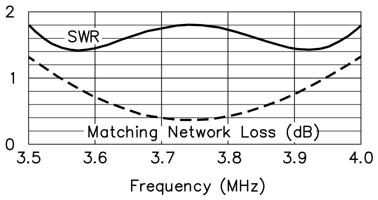

Fig 15—SWR and matching-network loss for example. Note that S_M is exactly 1.8 and the band-edge loss is exactly 1.32 dB, as calculated.

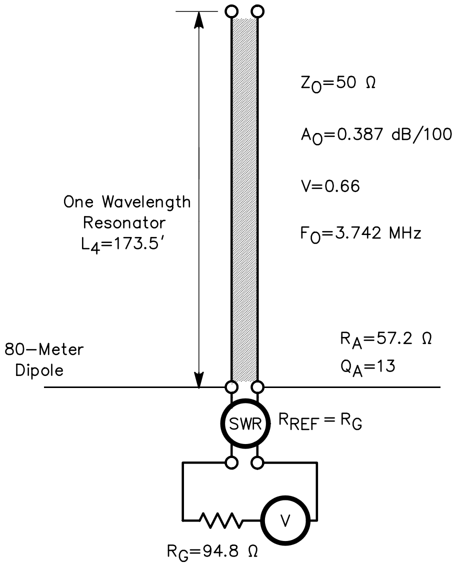

Fig 16—80-meter dipole example with a 1-λ resonator as the matching network.

lowing examples as well. The particular value of X_{N0} of this example may be obtained with a one-wavelength resonator (open-circuited at the far end) made from RG-213. The system described in this example is shown in **Fig 16**.

The broadband antenna system shown in Fig 16 has a desirable SWR characteristic, but the feed line to the transmitter is not yet present. Fortunately, the same cable segment that makes up the resonator may be used as the feed line. A property of a feed line whose length is a multiple of λ/2 (an even multiple of λ/4) is that its input impedance is a near replica of the impedance at the far end. This is exactly true for lossless lines at the frequencies where the λ/2 condition is true. Fortunately, practical lines have low-enough loss that the property mentioned above is true at the resonant frequency. Off resonance, the desired reactance cancellation we are looking for from the resonant line takes place. The very same design equations apply.

In **Fig 17**, the antenna shown in Fig 16 is moved to the far end of the 1λ resonator. The SWR and loss for this case are shown in **Fig 18**, which should be compared with Fig 15. The SWR is only slightly degraded, but the loss is about the same at the band edges. We have picked up a feed line that is 1λ long (173.5 feet) *and* broadband matching. You can make the SWR curve virtually the same as that of Fig 15 if the transmitter load resistance is increased from 94.8 Ω to 102 Ω. From a practical point of view, this degree of design refinement is unnecessary, but is instructive to know how the SWR characteristic may be controlled. You may apply the same approach when the resonator is an odd multiple of a wavelength, but in that case the transmitter or antenna must be connected λ/4 from the shorted end.

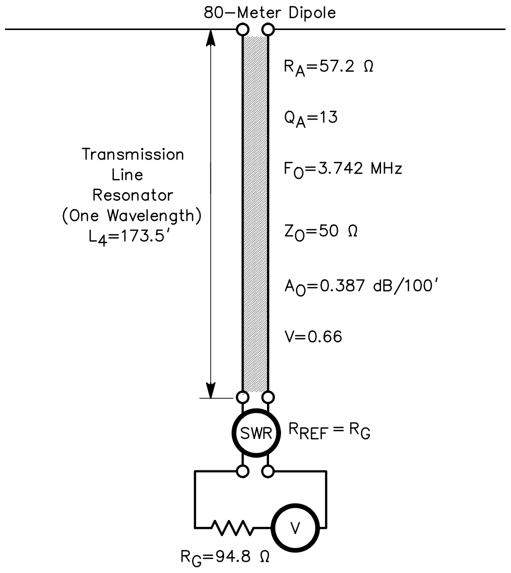

Fig 17—The antenna is at the far end of the 1-λ line. The line does double duty—broadband matching and signal transport.

Chebyshev Matching

Although it is not as efficient from the standpoint of bandwidth improvement and loss, Chebyshev matching is useful for some purposes. This arrangement yields a transmitter load impedance of 50 + *j* 0 Ω at two frequencies in the band. The matching network circuit is the same as the optimum matching case, but the parameters are different. In this case, we assume that the antenna parameters, F_0, R_A and Q_A, the network Q and the maximum SWR over the band are specified. The bandwidth, impedance level and generator resistance are given by:

$$BW = \frac{F_0}{Q_A Q_N} \left\{ \frac{(Q_A + Q_N)[(Q_A + 2Q_N)S_M - Q_A](S_M - 1)}{S_M} \right\}^{\frac{1}{2}} \qquad \text{(Eq 16)}$$

$$X_{N0} = \frac{R_A}{Q_A}\left[S_M + \frac{Q_A}{Q_N}(S_M - 1)\right] \qquad \text{(Eq 17)}$$

$$R_G = R_A\left(S_M - \frac{Q_A}{Q_A + Q_N}\right) \qquad \text{(Eq 18)}$$

The loss at the band edges is given by Eq 15. The frequencies where the transmitter load is 50 + *j* 0 Ω, F_L and F_H, are given by:

$$F_L = \sqrt{F_0^{\,2} + F_M^{\,2}} - F_M \qquad \text{(Eq 19)}$$

where

$$F_M = \frac{F_0}{2Q_A}\left[(S_M - 1)\left(1 + \frac{Q_A}{Q_N}\right)\right]^{\frac{1}{2}} \qquad \text{(Eq 20)}$$

$$F_H = \frac{F_0^{\,2}}{F_L} \qquad \text{(Eq 21)}$$

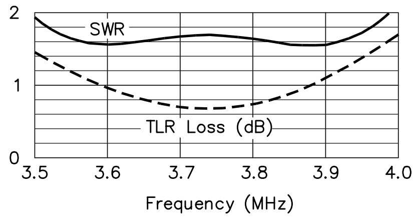

Fig 18—SWR and loss for a 1-λ 50-Ω feed line. The antenna system model is the one shown in Fig 17.

LC-MATCHING NETWORKS

Before the theory of optimal matching above was developed, Frank Witt, AI1H, described in his article “Broadband Dipoles—Some New Insights” in October 1986 *QST* the basic principle and some applications of LC matching networks. **Fig 19** shows an LC matching network that is a modified version of one originally proposed by Alan Bloom, N1AL. Note that the basic ingredients are present: the parallel-tuned circuit for reactance cancellation and a transformer.

The network also functions as a voltage balun by connecting the shield of the coaxial feed line to the center tap of the inductor. The capacitor is connected across the entire coil to obtain practical element values. The SWR response and efficiency offered by a network of lumped components is shown in **Fig 20**. The 2:1-SWR bandwidth with 50-Ω line is 460 kHz. The design is a Chebyshev match, where SWR = 1:1 is achieved at two frequencies. For the same configuration, more match bandwidth and less loss could have been realized if the optimal-match theory had been applied.

The efficiency at the band edges for the antenna system shown in Fig 20 is 90% (Loss = 0.5 dB). This low loss is due to the high Q of the LC matching network ($Q_N \approx 200$). The very low-impedance level required ($X_{N0} = 12.4\ \Omega$) cannot be easily realized with practical inductor-capacitor values. It is for this reason that the coil is tapped and serves as an autotransformer with multiple taps. The required impedance transformation ratio of 1:2.8 is easily achieved with this arrangement.

80-Meter DXer’s Delight

A version of an antenna system with an LC broadbanding network is dubbed the *80-meter DXer’s Delight*. This antenna has SWR minima near 3.5 and 3.8 MHz. A single antenna permits operation with a transmitter load impedance of about 50 Ω in the DX portions of the band, both CW and phone. The efficiency and SWR characteristic are shown in **Fig 21**.

Fig 22 shows the method of construction. The selection of a capacitor for this application must be made carefully, especially if high power is to be used. For the capacitor described in the caption of Fig 22, the allowable peak power (limited by the breakdown voltage) is 2450 W. However, the allowable average power (limited by the RF current rating of 4 A) is only 88 W! These limits apply at the 1.8:1 SWR points.

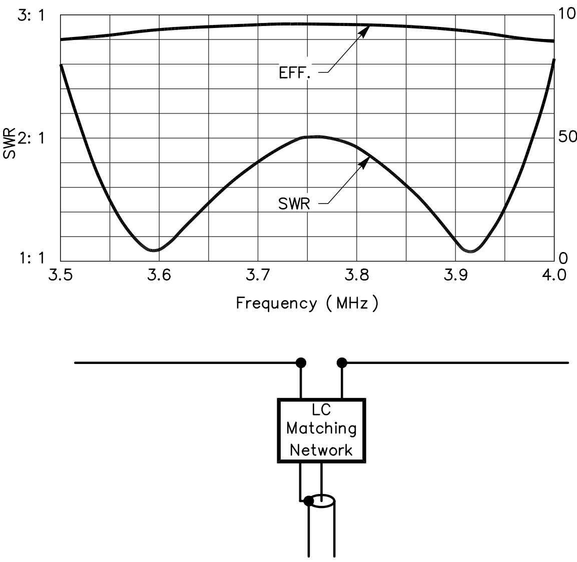

Fig 20—Efficient broadband matching with a lumped element LC network. The antenna in this example has a resistance of 72 Ω and a Q of 12. The matching network has an impedance level of 12.4 Ω, and a Q of 200. The feed line is 50-Ω coax, and the matching network provides an impedance step-up ratio of 2.8:1.

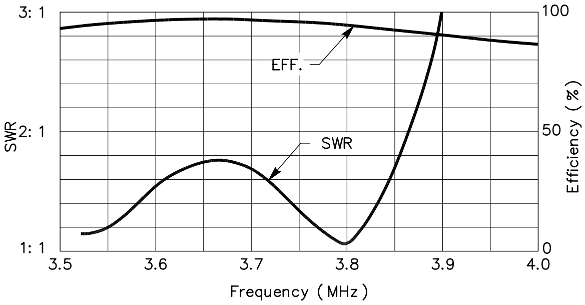

Fig 21—The 80-meter DXer’s Delight permits operation with a 50-Ω transmitter load in the DX portions of the band, both CW and phone. This network was designed to provide a broadband match for an inverted V with F_0 = 3.67 MHz, an antenna resistance of 60 Ω and a Q of 13. The matching network has an impedance level of 9 Ω, and a Q of 220. The feed line is 50-Ω coax, and the matching network provides an impedance step-up ratio of 2.0:1.

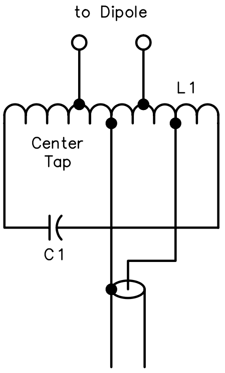

Fig 19—A practical LC matching network that provides reactance compensation, impedance transformation and balun action.

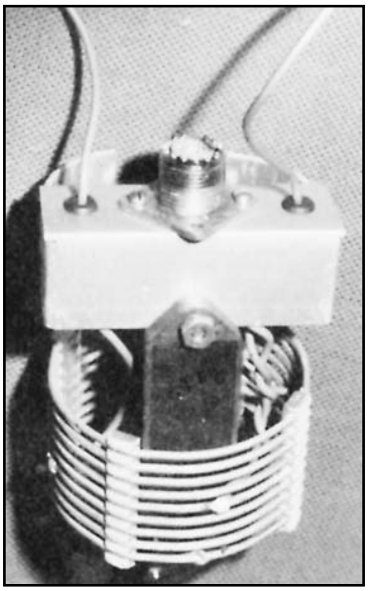

Fig 22—A method of constructing the LC broadband matching network. Components must be chosen for high Q and must have adequate voltage and current ratings. The network is designed for use at the antenna feed point, and should be housed in a weather-resistant package. The component values are as follows:

C1—400 pF transmitting mica rated at 3000 V, 4 A (RF)

L1—4.7 µH, 8½ turns of B&W coil stock, type 3029 (6 turns per in., 2½-in. dia, #12 wire). The LC circuit is brought to midband resonance by adjusting an end tap on the inductor. The primary and secondary portions of the coil have 1¾ and 2½ turns, respectively.

PROPERTIES OF TRANSMISSION-LINE RESONATORS

One way around the power limitations of LC broadband matching networks is to use a transmission line as the resonator. Transmission line resonators, TLRs, typically have higher power-handling capability because the losses, though higher, are distributed over the length of the line instead of being concentrated in the lumped-LC components. Transmission-line resonators must be multiples of a quarter wavelength. For a parallel-tuned circuit, the odd-multiple λ/4 lines must be shorted at the far end and the even-multiple λ/4 lines must be open circuited.

The length of the matching network resonator is n electrical quarter wavelengths. The impedance level, Q, and line length are given by:

$$X_{N0} = \frac{4 Z_0}{n \pi}$$

or, solving for Z_0,

$$Z_0 = \frac{n \pi X_{N0}}{4} \qquad \text{(Eq 22)}$$

$$Q_N = \frac{2.774 F_0}{A_0 V} \qquad \text{(Eq 23)}$$

$$L_N = \frac{245.9 n V}{F_0} \qquad \text{(Eq 24)}$$

where

Z_0 = line characteristic impedance in Ω

A_0 = matched line loss in dB per 100 feet

V = velocity factor, and

L_N = line length in feet.

These equations may be used to find the properties of the transmission line resonator in the systems shown in Figs 16 and 17. The line is RG-213 with the following properties:

$Z_0 = 50\ \Omega$

$V = 0.66$

A_0 = 0.4 dB/100 feet at 4 MHz

We need the matched loss at 3.742 MHz. Since the line loss is approximately proportional to $\sqrt{\text{frequency}}$,

$$A_{0F2} = A_{0F1} \sqrt{\frac{F_2}{F_1}} \qquad \text{(Eq 25)}$$

where A_{0F1} and A_{0F2} are the matched line losses at frequencies F_1 and F_2, respectively. Hence,

A_0 = 0.387 dB/100 feet at 3.742 MHz

$Q_N = 40.65$ (from Eq 23)

This is the Q of any resonator made from RG-213 for a frequency of 3.742 MHz. It does not matter how many quarter wavelengths (n) make up the cable segment. Earlier we calculated that the required impedance level, X_{N0}, is 15.9 Ω. Therefore, the required impedance level is:

$$Z_0 = \frac{n \pi \times 15.9}{4} = 12.5\, n\, \Omega \qquad \text{(from Eq 22)}$$

This means that you could use a quarter-wavelength segment with a short at the far end if Z_0 = 12.5 Ω, or a half-wavelength segment with a open at the far end if Z_0 = 25 Ω. Our example used an open-circuited full-wavelength resonator (n = 4) with a Z_0 of 50 Ω. This example was crafted to make this happen (so that it fits RG-213 properties), but it illustrates that some juggling is required to obtain the right resonator properties when it is made from a transmission line, since we are stuck with the cable's characteristic impedance.

The TLR Transformer

A way around the limitation created by available characteristic impedance values is to use the *transmission-line resonator transformer*. The basic idea is to make the connection to the TLR at an intermediate point along the line instead of at the end of the line. It is analogous to using taps on a parallel-tuned LC resonant circuit to achieve transformer action (See Fig 19). The nature of the transformer action is seen in **Fig 23**, where the ends of the transformer are designated 1-1 and 2-2. The impedance ratio of the transformer, N_Z, is approximated by

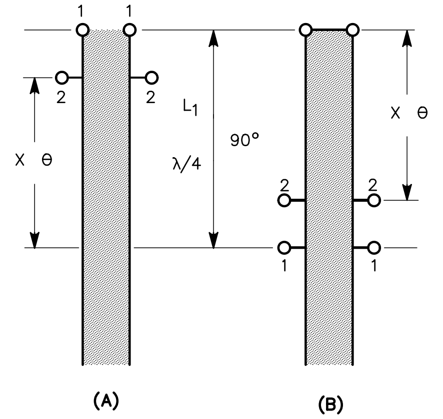

Fig 23—Transformer action in the TLR. The definition of transformer terminals depends on whether the TLR end is open- or short-circuited. θ is the distance between the minimum of the voltage standing wave (at resonance) and the connection point, expressed as an electrical angle. The distance x is that same distance expressed in feet.

$$N_Z = \sin^2 \theta \qquad \text{(Eq 26)}$$

The distance in feet, x, is given by

$$x = \frac{\theta}{90} \times \frac{\lambda}{4} = \frac{\theta}{90} \times \frac{245.9\ V}{F_0} \qquad \text{(Eq 27)}$$

This approach may be used for the connection of the TLR to the transmitter or to the antenna or both. An example will help make the power of the TLR transformer clear. We want to broadband match an 80-meter dipole (where F_0 is 3.742 MHz, R_A is 65 Ω and Q_A is 13). The feed line is RG-213 50-Ω cable, and the total distance between the antenna and transmitter is less than 100 feet.

Since we know that Z_0 = 50 Ω, we will design a prototype system with a 3/4 λ TLR (n=3). Since n is odd, the TLR must be shorted at one end, and we will place the short at the transmitter end of the line. The intermediate step of using a prototype system will enable us to design a system with the final TLR, but without transformer action. Then we'll calculate the proper connection points along the TLR. **Fig 24** shows the prototype system and the system with built-in transformers.

Note that the prototype antenna is connected to the open end of the TLR. The prototype generator (transmitter) is connected at a multiple of λ/2 from the antenna. Here, at resonance, an approximate replica of the antenna will appear. We will use primed variables to indicate prototype antenna system elements. The prototype antenna differs from the actual antenna in impedance level, so only R_A will change. F_0 and Q_A remain the same.

Starting with the 3/4 λ TLR, we have:

X_{N0} = 21.2 Ω (from Eq 22)

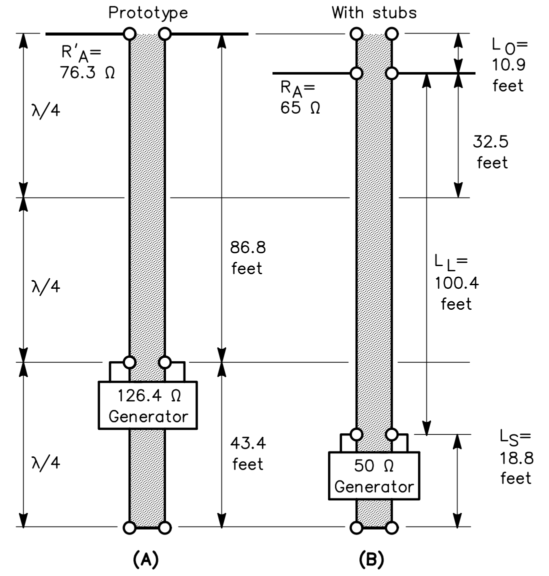

Fig 24—An optimized antenna system with a 3/4 λ TLR made from RG-213 coaxial cable. The prototype system at A is a convenient intermediate step in the design process. At B is the configuration with TLR transformers. It has an open stub, L_O, a link, L_L (which serves as the feed line), and a shorted stub, L_S.

The prototype antenna resistance at resonance, R'_A, is calculated by rearranging Eq 13.

$$R'_G = \frac{X_{N0}\, Q_A}{\left(1 + \frac{Q_A}{2 Q_N}\right) S_{M\,min}^{\;2} - \frac{Q_A}{2\, Q_N}} \qquad \text{(Eq 28)}$$

Earlier we calculated that Q_N = 40.65 and S_{Mmin} = 1.8. Now we find

R'_A = 76.3 Ω (from Eq 28)
R'_G = 126.4 Ω (from Eq 14)

The electrical angles (in degrees) and connection locations (in feet) are given by

$$\theta_A = \sin^{-1} \sqrt{\frac{R_A}{R'_A}} \qquad \text{(Eq 29)}$$

$$\theta_G = \sin^{-1} \sqrt{\frac{R_G}{R'_G}} \qquad \text{(Eq 30)}$$

$$X_A = \frac{\theta_A}{90} L_1 \qquad \text{(Eq 31)}$$

$$X_G = \frac{\theta_G}{90} L_1 \qquad \text{(Eq 32)}$$

where

$$L_1 = \frac{\lambda}{4} = \frac{245.9\ V}{F_0}$$

The lengths of the stubs and link of Fig 24 are given by

$L_O = L_1 = x_A$ (Eq 33)

$L_S = x_G$ (Eq 34)

$L_L = 3L_1 - L_O - L_S$ (Eq 35)

Since L_1 = 43.4 feet,

$\theta_A = 67.4°$ (from Eq 29)

$\theta_G = 39.0°$ (from Eq 30)

L_O = 10.9 feet (from Eq 33)

L_S = 18.8 feet (from Eq 34)

L_L = 100.4 feet (from Eq 35)

These dimensions are summarized in Fig 24. If required, additional 50-Ω cable may be added between the transmitter and the junction of the shorted stub and the link. A remarkable aspect of the TLR transformer, as demonstrated in this example, is that it is a transformer with multiple taps distributed over a great distance, in this case over 100 feet.

The calculated SWR and loss of the antenna system with the TLR transformer are shown in **Fig 25A**. We mentioned earlier that the stub calculations are an approximation. Fig 25 bears this out, but the changes necessary to achieve the optimized results are relatively small. By changing the lengths and dipole resonant frequency slightly, the target optimized SWR characteristic of Fig 25B results. See **Table 1** for a summary of the changes required.

To demonstrate the significant improvement in match bandwidth that the TLR transformer provides, the SWR and loss of the same dipole fed with 100.4 feet of RG-213 cable is shown in Fig 25C. The loss added by the two stubs used to obtain match bandwidth enhancement is negligible (0.2 to 0.5 dB).

A TLR transformer may be used by itself to achieve a low-loss narrow-band impedance transformation, functioning like a λ/4 Q-section. It is different in nature from a Q-section, in that it is a true impedance transformer, while a Q section is an *impedance inverter*. The TLR transformer may be designed into the antenna system and is a useful tool when a narrow-band impedance transformation is needed.

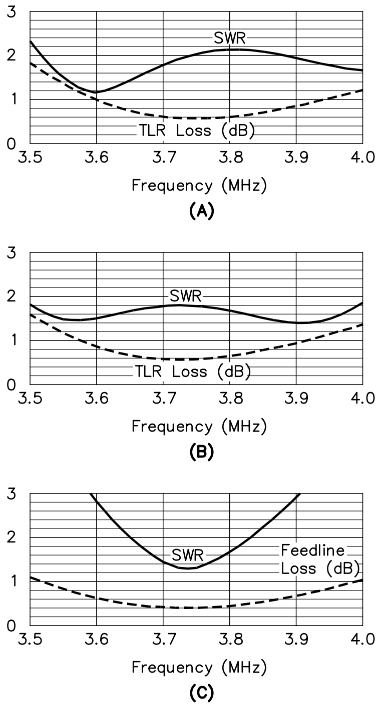

Fig 25—The SWR and loss of the TLR transformer antenna system. At A, the results are shown with the calculated line lengths. At B, the improvement after the lengths were optimized is shown in Table 1. At C, for comparison, SWR and loss for the same setup as at A, but with the stubs removed.

Table 1

Calculated and Optimized Parameters Using TLR Transformers

	Calculated	*Optimized*
Dipole resonant frequency (F_0)	3.742 MHz	3.710 MHz
Open stub (L_O)	10.9 feet	13.1 feet
Shorted stub (L_S)	18.8 feet	20.1 feet
Link (L_L)	100.4 feet	101.0 feet

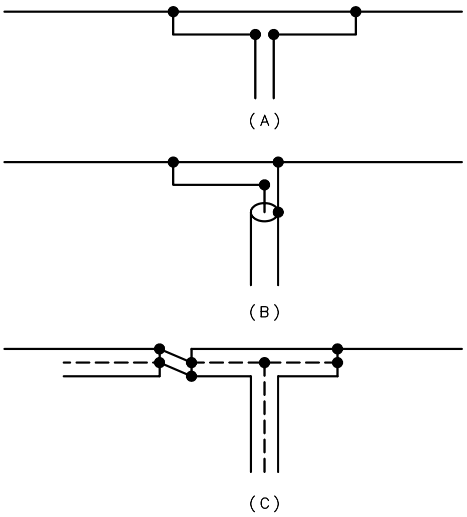

Fig 26—Dipole matching methods. At A, the T match; at B, the gamma match; at C, the coaxial resonator match.

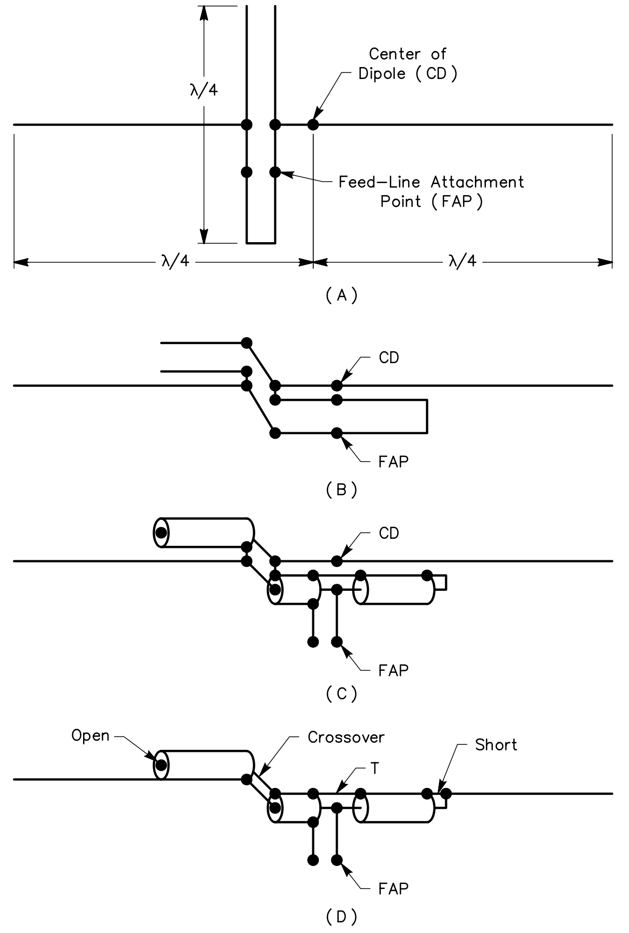

Fig 27—Evolution of the coaxial-resonator-match broadband dipole. At A, a TLR transformer is used to match the feed line to the off-center-fed dipole. The match and dipole are made collinear at B. At C, the balanced transmission-line TLR transformer of A and B is replaced by a coaxial version. Because the shield of the coax can serve as a part of the dipole radiator, the wire adjacent to the coax match may be eliminated, D.

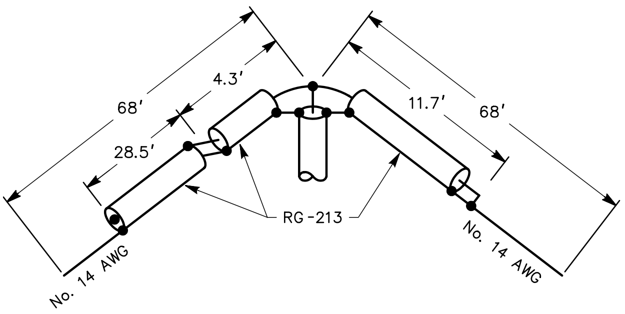

Fig 28—Dimensions for the 80-Meter MHz DX Special, an antenna optimized for the phone and CW DX portions of the 3.5-MHz band. The coax segment lengths total to one quarter-wavelength. The overall length is the same as that of a conventional inverted V dipole.

TRANSMISSION-LINE RESONATORS AS PART OF THE ANTENNA

The Coaxial-Resonator Match

This material is condensed from articles that appeared in April 1989 *QST* and *The ARRL Antenna Compendium, Vol 2*. The coaxial-resonator match, a concept based on the TLR transformer, performs the same function as the T match and the gamma match, that is, matching a transmission line to a resonant dipole. These familiar matching devices as well as the coaxial resonator match are shown in **Fig 26**. The coaxial resonator match has some similarity to the gamma match in that it allows connection of the shield of the coaxial feed line to the center of the dipole, and it feeds the dipole off center. The coaxial resonator match has a further advantage: It can be used to broadband the antenna system while providing an impedance match.

The coaxial resonator match is a resonant transformer made from a quarter-wave long piece of coaxial cable. **Fig 27** shows the evolution of the coaxial-resonator match. Now it becomes clear why coaxial cable is used for the quarter-wave TLR transformer—interaction between the dipole and the matching network is minimized. The effective dipole feed point is located at the crossover, which is an off-center feed point. In effect, the match is physically located “inside” the dipole. Currents flowing on the inside of the shield of the coax are associated with the resonator; currents flowing on the outside of the shield of the coax are the usual dipole currents. Skin effect provides a degree of isolation and allows the coax to perform its dual function. The wire extensions at each end make up the remainder of the dipole, making the overall length equal to one half-wavelength.

A useful feature of an antenna using the coaxial-resonator match is that the entire antenna is at the same dc potential as the feed-line potential, thereby avoiding charge buildup on the antenna. Hence, noise and the potential of lightning damage are reduced.

Fig 28 shows the detailed dimensions of an 80-meter inverted-V dipole using a coaxial-resonator match. This design provides a load SWR better than 1.6:1 from 3.5 to 3.825 MHz and has been named the *80-Meter DX Special*. The coax is an electrical quarter wavelength, has a short at one end, an open at the other end, a strategically placed crossover, and is fed at a tee junction. (The crossover is made by connecting the shield of one coax segment to the center conductor of the adjacent segment and by connecting the remaining center conductor and shield in a similar way.) The antenna is constructed as an inverted-V dipole with a 110° included angle and an apex at 60 feet. The measured SWR versus frequency is shown in **Fig 29**. Also in Fig 29

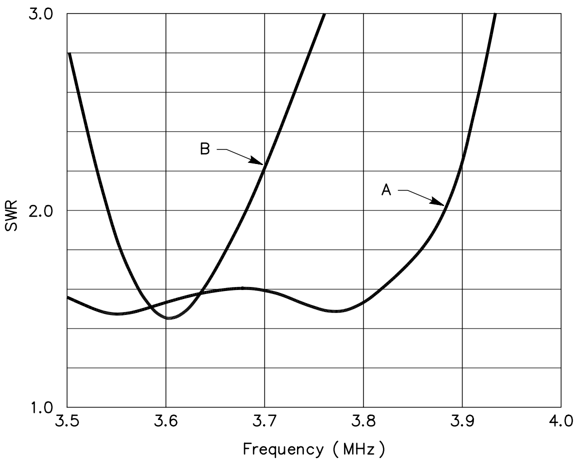

Fig 29—Measured SWR performance of the 80-Meter DX Special, curve A. Note the substantial broadbanding relative to a conventional uncompensated dipole, curve B.

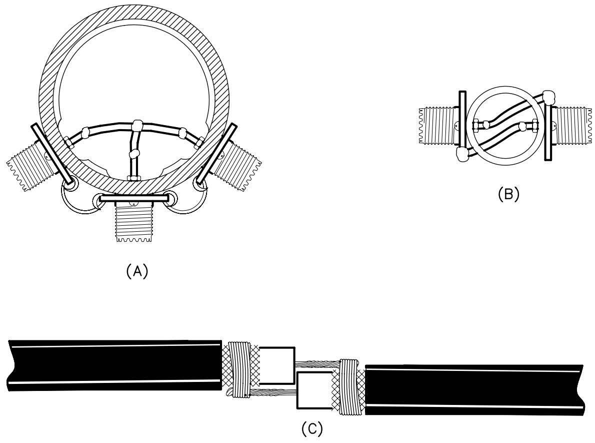

Fig 30—T and crossover construction. At A, a 2-inch PVC pipe coupling can be used for the T, and at B, a 1-inch coupling for the crossover. These sizes are the nominal inside diameters of the PVC pipe that is normally used with the couplings. The T could be made from standard UHF hardware (an M-358 T and a PL-258 coupler). An alternative construction for the crossover is shown at C, where a direct solder connection is made.

is the SWR characteristic for an uncompensated inverted-V dipole made from the same materials and positioned exactly like the broadband version.

The 80-Meter DX Special is made from RG-213 coaxial cable and #14 AWG wire, and is fed with 50-Ω coax. The coax forming the antenna should be cut so that the stub lengths of Fig 28 are within $^{1}/_{2}$ inch of the specified values. PVC plastic-pipe couplings and SO-239 UHF chassis connectors can be used to make the T and crossover connections, as shown in **Fig 30A** and 30B. Alternatively, a standard UHF T connector and coupler can be used for the T, and the crossover may be a soldered connection (Fig 30C). RG-213 is used because of its ready availability, physical strength, power handling capability, and moderate loss.

Cut the wire ends of the dipole about three feet longer than the lengths given in Fig 28. If there is a tilt in the SWR-frequency curve when the antenna is first built, it may be flattened to look like the shape given in Fig 29 by increasing or decreasing the wire length. Each end should be lengthened or shortened by the same amount.

A word of caution: If the coaxial cable chosen is not RG-213 or equivalent, the dimensions will have to be modified. The following cable types have about the same characteristic impedance, loss and velocity factor as RG-213 and could be substituted: RG-8, RG-8A, RG-10, RG-10A and RG-215. If the Q of the dipole is particularly high or the radiation resistance is unusually low because of different ground characteristics, antenna height, surrounding objects and so on, then different segment lengths will be required.

What is the performance of this broadband antenna relative to that of a conventional inverted-V dipole? Aside from the slight loss (about 0.75 dB at band edges, less elsewhere) because of the non-ideal matching network, the broadband version will behave essentially the same as a dipole cut for the frequency of interest. That is, the radiation patterns for the two cases will be virtually the same.

The Snyder Dipole

A commercially manufactured antenna utilizing the principles described in the Matching Network Design section is the Snyder dipole. Patented by Richard D. Snyder in late 1984 (see Bibliography), it immediately received much public attention through an article that Snyder published. Snyder's claimed performance for the antenna is a 2:1 SWR bandwidth of 20% with high efficiency.

The configuration of the Snyder antenna is like that of the crossed-double bazooka of Fig 12, with 25-Ω line used for the cross-connected resonators. In addition, the antenna is fed through a 2:1 balun, and exhibits a W-shaped SWR characteristic like that of Fig 13. The SWR at band center, based on information in the patent document, is 1.7 to 1. There is some controversy in professional circles regarding the claims for the Snyder antenna.

The Improved Crossed-Double Bazooka

The following has been condensed from an article by Reed Fisher, W2CQH, that appeared in *The ARRL Antenna Compendium Volume 2*. The antenna is shown in **Fig 31**. Note that the half-wave flat-top is constructed of sections of RG-58 coaxial cable. These sections of coaxial cable serve as quarter-wave shunt stubs that are essentially connected *in parallel* at the feed point in the crossed double-bazooka fashion. At an electrical quarter

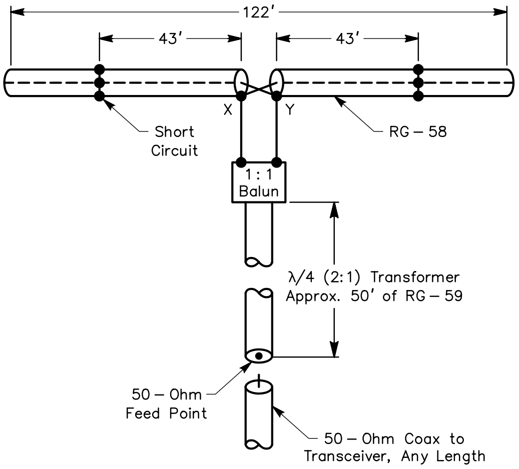
122'
43'
43'
X
Y
Short
Circuit
RG – 58
1 : 1
Balun
λ/4 (2:1) Transformer
Approx. 50' of RG – 59
50 – Ohm
Feed Point
50 – Ohm Coax to
Transceiver, Any Length

Fig 31—Details of the W2CQH broadband-matched 80-meter dipole.

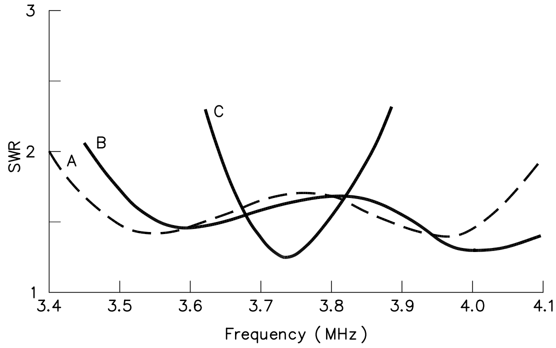

Fig 32—SWR curves for the improved double bazooka. Curve A, the theoretical curve with 50-Ω stubs and a λ/4, 75-Ω matching transformer. Curve B, measured response of the same antenna, built with RG-58 stubs and an RG- 59 transformer. Curve C, measurements from a dipole without broadbanding. Measurements were made at W2CQH with the dipole horizontal at 30 feet.

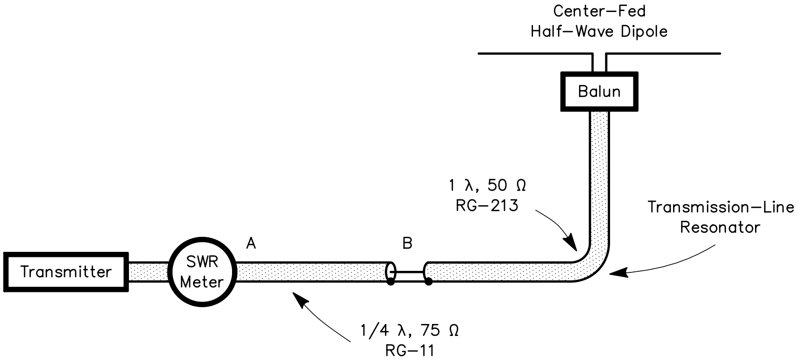
Center–Fed
Half–Wave Dipole
Balun
1 λ, 50 Ω
RG–213
Transmission–Line
Resonator
A
B
Transmitter
SWR
Meter
1/4 λ, 75 Ω
RG–11

Fig 33—This simple broadband antenna system resembles a conventional 80-meter dipole except for the λ/4-wavelength 75-Ω segment. The lengths of the Q-section, TLR (including balun) and dipole are 43.3 feet, 170.5 feet, and 122.7 feet, respectively.

wavelength (43 feet) from each side of the feed point X-Y, the center conductor is shorted to the braid of the coaxial cable.

The parallel stubs provide reactance compensation. The necessary impedance transformation at the antenna feed point is provided by the quarter-wave Q-section constructed of a 50-foot section of 75-Ω coaxial cable (RG-59). A W2DU 1:1 current balun is used at the feed point. See **Fig 32** for the SWR versus frequency for this antenna with and without the broadband matching network.

This antenna is essentially that of the crossed double bazooka antenna shown in Fig 12, with the addition of the 75-Ω Q-section. The improvement in match bandwidth is substantial. The antenna at W2CQH is straight and nearly horizontal with an average height of about 30 feet.

TRANSMISSION-LINE RESONATORS AS PART OF THE FEED LINE

A Simple Broadband Dipole for 80 Meters

This system was described in an article in September 1993 *QST*. It has the advantage that the radiator is the same as that of a conventional half-wave wire dipole. Thus the antenna is light weight, easy to support and has small wind and ice loading. The broadband matching network is integrated into the feed line.

The system is shown in **Fig 33**. It is a variation of the example shown in the Matching Network Design section of this chapter as shown in Fig 17. The TLR is a 1λ length of RG-213 50-Ω coax and the impedance transformation is accomplished with a 75-Ω Q-section

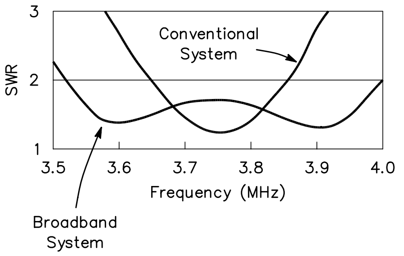

Fig 34—Measured SWR versus frequency for the broadband and conventional antenna systems.

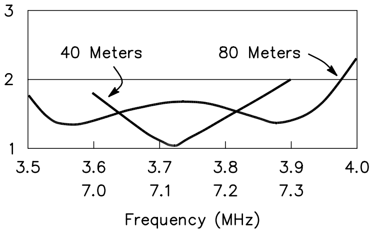

Fig 35—Measured SWR for the 80- and 40-meter multiband antenna system.

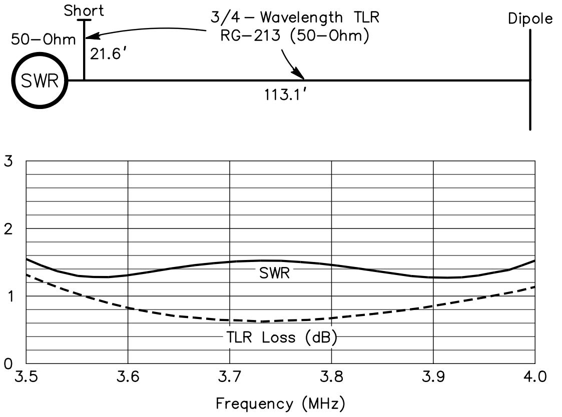

Fig 36—Broadbanding with the TLR transformer. The dimensions shown apply for an 80-meter horizontal dipole 80 feet above average ground.

made of RG-11. The antenna is an inverted-V with a 140° included angle and an apex height of 60 feet. The wire size is #14, but is not critical. The balun is a W2DU 1:1 current type. It easily handles the legal power limit over the entire 80-meter band. The measured SWR is shown in **Fig 34**.

A unique advantage of this antenna is that by paralleling a 40-meter dipole at the feed point and sharing the feed line, operation on both 80 and 40 meters is possible. The reason for this is that the lengths of the Q-section and the TLR are multiples of a half-wavelength on 40 meters. To minimize the interaction, the two dipoles should be spaced from each other away from the feed point. First tune the 80-meter antenna and then the 40-meter one. **Fig 35** shows the result of adding a 40-meter dipole to the system shown in Fig 33. Each 40-meter dipole leg is 34.4 feet long. Note that the SWR on 80 meters changes very little compared to Fig 34. No change was made to the 80-meter dipole or to the transmission line.

80-Meter Dipole with the TLR Transformer

The total length of the feed line in the system just described above in Fig 33 is quite long, about 214 feet. For installations with shorter runs, the TLR transformer concept may be employed to advantage. An example provided in "Optimizing the 80-Meter Dipole" in *The ARRL Antenna Compendium, Vol 4* is shown in **Fig 36**. Here the TLR is $^{3}/_{4}$ λ long overall and is made from RG-213 50-Ω cable. The design applies for an 80-meter horizontal dipole, 80 feet above average ground, with an assumed antenna F_0, R_A, and Q_A of 3.72 MHz, 92 Ω and 9, respectively. (See Figs 3 and 4). The lengths must be revised for other dipole parameters.

For this design, the feed line is 113.1 feet long and the shorted stub at the transmitter end is 21.6 feet long. You can add any length of 50-Ω cable between the transmitter and the junction of the shorted stub and the feed line. The shorted stub gives a dc path to ground for both sides of the dipole, preventing charge buildup and some lightning protection. For multiband operation with an antenna tuner, the stub should be removed.

The general application of the TLR transformer concept has a tap at both the transmitter end and a tap at the antenna end. (See Fig 24.) In this case, the concept was applied only at the transmitter end, since very little impedance transformation was required at the antenna end. The location of the TLR transformer is not obvious. Consider a $^{3}/_{4}$ λ TLR with a short at one end, and an open at the other end where the antenna is connected. The generator for no impedance transformation would be located λ/4 (43.6 feet @ 3.72 MHz) from the short. By connecting the transmitter 21.6 feet from the short, the required impedance transformation for a transmitter optimized for 50-Ω loads is achieved.

BIBLIOGRAPHY

Source material and more extended discussion of the topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

A. Bloom, "Once More with the 80-Meter Dipole," Technical Correspondence, *QST*, Jun 1985, p 42.

R. E. Fisher, "A Simple, Broadband 80-Meter Dipole Antenna," *The ARRL Antenna Compendium, Vol 2* (Newington, CT: ARRL, 1989), pp 119-123.

R. M. Fano, "Theoretical Limitations on the Broadband Matching of Arbitrary Impedances," *Journal of the Franklin Inst*, Jan 1950, pp 57-83, and Feb 1950, pp 139-155.

J. Hall, "The Search for a Simple, Broadband 80-Meter Dipole," *QST*, Apr 1983, pp 22-27.

J. Hall, "Maxcom Antenna Matcher and Dipole Cable Kit," Product Review, *QST*, Nov 1984, pp 53-54.

R. C. Hansen, "Evaluation of the Snyder Dipole," *IEEE Trans on Antennas and Prop*, Vol AP-35, No. 2, Feb 1987, pp 207-210.

A. Harbach, "Broadband 80-Meter Antenna," *QST*, Dec 1980, pp 36-37 (presentation of a cage antenna).

S. Leslie, "Broadbanding the Elevated, Inverse-Fed Ground Plane Antenna," *The ARRL Antenna Compendium, Vol 6* (Newington, CT: 1999), pp 209-211.

R. D. Snyder, "The Snyder Antenna," *RF Design*, Sep/Oct 1984, pp 49-51.

R. D. Snyder, "Broadband Antennae Employing Coaxial Transmission Line Sections," United States Patent no. 4,479,130, issued Oct 23, 1984.

C. Whysall, "The 'Double Bazooka' Antenna," *QST*, Jul 1968, pp 38-39.

F. Witt, "A Simple Broadband Dipole for 80 Meters," *QST*, Sept 1993, pp 27-30, 76.

F. J. Witt, "Broadband Dipoles—Some New Insights," *QST*, Oct 1986, pp 27-37.

F. Witt, "Broadband Matching with the Transmission Line Resonator," *The ARRL Antenna Compendium, Vol 4*, (Newington, CT: 1995), pp 30-37.

F. Witt, "How to Design Off-Center-Fed Multiband Wire Antennas Using that Invisible Transformer in the Sky," *The ARRL Antenna Compendium, Vol 3,* (Newington, CT: ARRL, 1992), p 74.

F. Witt, "Match Bandwidth of Resonant Antenna Systems," *QST*, Oct 1991, pp 21-25

F. Witt, "Optimizing the 80-Meter Dipole," *The ARRL Antenna Compendium, Vol 4* (Newington, CT: 1995), pp 38-48.

F. Witt, "Optimum Lossy Broadband Matching Networks for Resonant Antennas," *RF Design*, Apr 1990, pp 44-51 and Jul 1990, p 10.

F. Witt, "The Coaxial Resonator Match and the Broadband Dipole," *QST*, Apr 1989, pp 22-27.

F. Witt, "The Coaxial Resonator Match, *The ARRL Antenna Compendium, Vol 2* (Newington, CT: ARRL, 1989), pp 110-118.

Chapter 10

Log Periodic Arrays

This chapter was contributed by L. B. Cebik, W4RNL. The *Log Periodic Dipole Array* (*LPDA*) is one of a family of frequency-independent antennas. Alone, an LPDA forms a directional antenna with relatively constant characteristics across a wide frequency range. It may also be used with parasitic elements to achieve specific characteristic within a narrow frequency range. Common names for such hybrid arrays are the *log-cell Yagi* or the *Log-Yag*. We shall look at the essential characteristics of both types of arrays in this chapter.

The LPDA is the most popular form of log-periodic systems, which also include zig-zag, planar, trapezoidal, slot, and V forms. The appeal of the LPDA version of the log periodic antenna owes much to its structural similarity to the Yagi-Uda parasitic array. This permits the construction of directional LPDAs that can be rotated—at least within the upper HF and higher frequency ranges. Nevertheless, the LPDA has special structural as well as design considerations that distinguish it from the Yagi. A number of different construction techniques for both wire and tubular elements are illustrated later in this chapter.

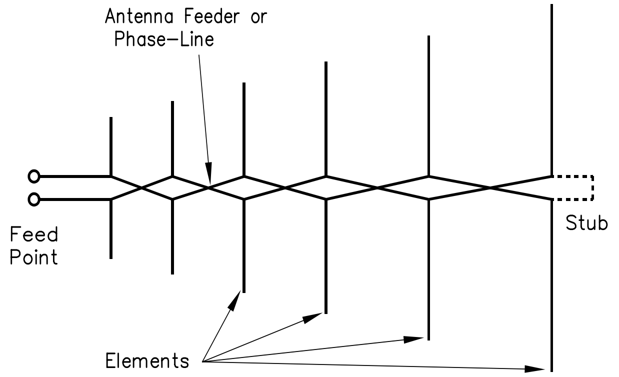

Fig 1—The basic components of a log periodic dipole array (LPDA). The forward direction is to the left in this sketch. Many variations of the basic design are possible.

The LPDA in its present form derives from the pioneering work of D. E. Isbell at the University of Illinois in the late 1950s. Although you may design LPDAs for large frequency ranges—for example, from 3 to 30 MHz or a little over 3 octaves—the most common LPDA designs that radio amateurs use are limited to a one-octave range, usually from 14 to 30 MHz. Amateur designs for this range tend to consist of linear elements. However, experimental designs for lower frequencies have used elements shaped like inverted Vs, and some versions use vertically oriented $^1/_4$ λ elements over a ground system.

Fig 1 shows the parts of a typical LPDA. The structure consists of a number of linear elements, the longest of which is approximately $^1/_2$ λ long at the lowest design frequency. The shortest element is usually about $^1/_2$ λ long at a frequency well above the highest operating frequency. The antenna feeder, also informally called the *phase-line*, connects the center points of each element in the series, with a phase reversal or *cross-over* between each element. A stub consisting of a shorted length of parallel feed line is often added at the back of an LPDA.

The arrangement of elements and the method of feed yield an array with relatively constant gain and front-to-back ratio across the designed operating range. In addition, the LPDA exhibits a relatively constant feed-point impedance, simplifying matching to a transmission line.

BASIC DESIGN CONSIDERATIONS

For the amateur designer, the most fundamental facets of the LPDA revolve around three interrelated design variables: α (alpha), τ (tau), and σ (sigma). Any one of the three variables may be defined by reference to the other two.

Fig 2 shows the basic components of an LPDA. The angle α defines the outline of an LPDA and permits every dimension to be treated as a radius or the consequence of a radius (R) of a circle. The most basic structural dimensions are the element lengths (L), the distance (R) of each element from the apex of angle α, and the distance between elements (D). A single design constant, τ, defines all of these relationships in the following manner:

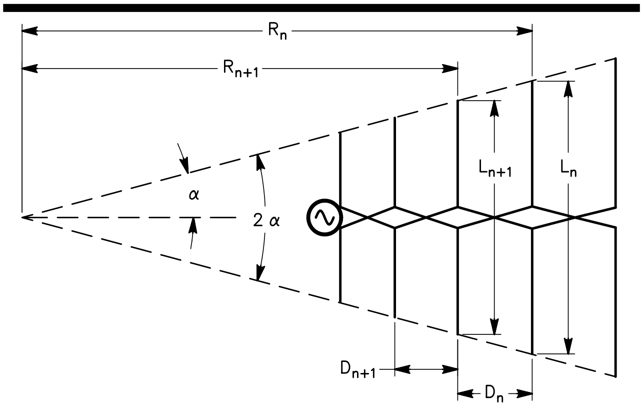

Fig 2—Some fundamental relationships that define an array as an LPDA. See the text for the defining equations.

$$\tau = \frac{R_{n+1}}{R_n} = \frac{D_{n+1}}{D_n} = \frac{L_{n+1}}{L_n} \qquad \text{(Eq 1)}$$

where element n and n+1 are successive elements in the array working toward the apex of angle α. The value of τ is always less than 1.0, although effective LPDA design requires values as close to 1.0 as may be feasible.

The variable τ defines the relationship between successive element spacings but it does not itself determine the initial spacing between the longest and next longest elements upon which to apply τ successively. The initial spacing also defines the angle α for the array. Hence, we have two ways to determine the value of σ, the relative spacing constant:

$$\sigma = \frac{1 - \tau}{4 \tan \alpha} = \frac{D_n}{2L_n} \qquad \text{(Eq 2)}$$

where D_n is the distance between any two elements of the array and L_n is the length of the longer of the two elements. From the first of the two methods of determining the value of σ, we may also find a means of determining α when we know both τ and σ.

For any value of τ, we may determine the optimal value of σ:

$$\sigma_{opt} = 0.243\tau - 0.051 \qquad \text{(Eq 3)}$$

The combination of a value for τ and its corresponding optimal value of σ yields the highest performance of which an LPDA is capable. For values of τ from 0.80 through 0.98, the value of optimal σ varies from 0.143 to 0.187, in increments of 0.00243 for each 0.01 change in τ. However, using the optimal value of σ usually yields a total array length that is beyond ham construction or tower/mast support capabilities. Consequently, amateur LPDAs usually employ compromise values of τ and σ that yield lesser but acceptable performance.

For a given frequency range, increasing the value of τ increases both the gain and the number of required elements. Increasing the value of σ increases both the gain and the overall boom length. A τ of 0.96—which approaches the upper maximum recommended value for τ—yields an optimal σ of about 0.18, and the resulting array grows to over 100 feet long for the 14 to 30 MHz range. The maximum free space gain is about 11 dBi, with a front-to-back ratio that approaches 40 dB. Normal amateur practice, however, uses values of τ from about 0.88 to 0.95 and values of σ from about 0.03 to 0.06.

Standard design procedures usually set the length of the rear element for a frequency about 7% lower than the lowest design frequency and use the common dipole formula ($L_{feet} = 468/f_{MHz}$) to determine its length (5% lower than a free-space half wavelength, where $L_{feet} = 493.56/f_{MHz}$). The upper frequency limit of the design is ordinarily set at about 1.3 times the highest design frequency. Since τ and σ set the increment between successive element lengths, the number of elements becomes a function of when the shortest element reaches the dipole length for the adjusted upper frequency.

The adjusted upper frequency limit results from the behavior of LPDAs with respect to the number of active elements. See **Fig 3**, which shows an edge view of a 10-element LPDA for 20 through 10 meters. The vertical lines represent the peak relative current magnitude for each element at the specified frequency. At 14 MHz, virtually every element of the array shows a significant cur-

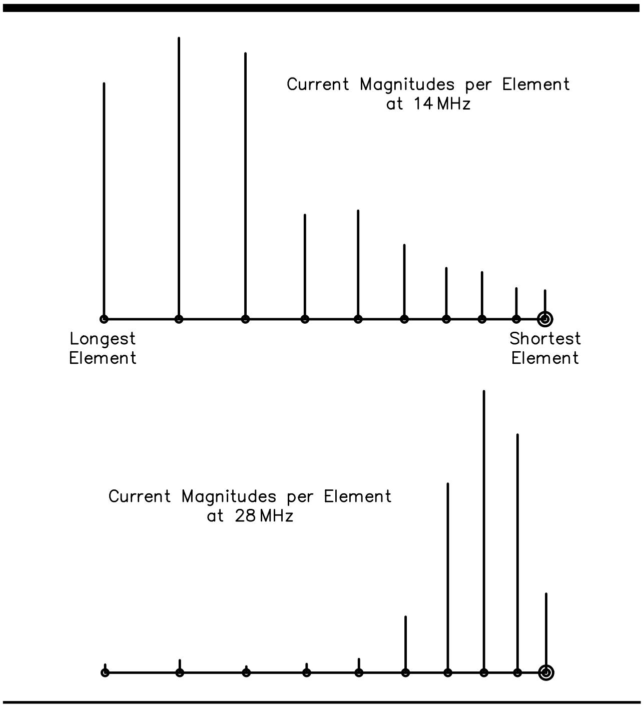

Fig 3—The relative current magnitude on the elements of an LPDA at the lowest and highest operating frequencies for a given design. Compare the number of "active" elements, that is, those with current levels at least $^1/_{10}$ of the highest level.

rent magnitude. However, at 28 MHz, only the forward 5 elements carry significant current. Without the extended design range to nearly 40 MHz, the number of elements with significant current levels would be severely reduced, along with upper frequency performance.

The need to extend the design equations below the lowest proposed operating frequency varies with the value of τ. In **Fig 4**, we can compare the current on the rear elements of two LPDAs, both with a σ value of 0.04. The upper design uses a τ of 0.89, while the lower design uses a value of 0.93. The most significant current-bearing element moves forward with increases in τ, reducing (but not wholly eliminating) the need for elements whose lengths are longer than a dipole for the lowest operating frequency.

LPDA Design and Computers

Originally, LPDA design proceeded through a series of design equations intended to yield the complete specifications for an array. More recent techniques available to radio amateurs include basic LPDA design software and antenna modeling software. One good example of LPDA design software is *LPCAD28* by Roger Cox, WB0DGF. A copy of this freeware program is on the CD-ROM that accompanies this volume. The user begins by specifying the lowest and highest frequencies in the design. He then enters either his selected values for τ and σ or his choices for the number of elements and the total length of the array. With this and other input data, the program provides a table of element lengths and spacings, using the adjusted upper and lower frequency limits described earlier.

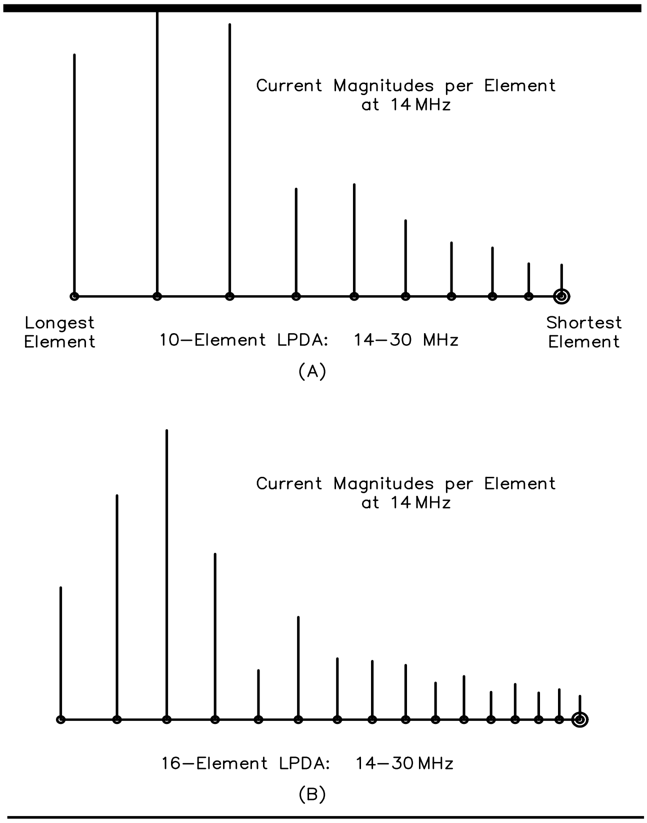

Fig 4—Patterns of current magnitude at the lowest operating frequency of two different LPDA designs: a 10-element low-τ design and a 16-element higher-τ design.

The program also requests the diameters of the longest and shortest elements in the array, as well as the diameter of average element. From this data, the program calculates a recommended value for the phase-line connecting the elements and the approximate resistive value of the input impedance. Among the additional data that *LPCAD28* makes available is the spacing of conductors to achieve the desired characteristic impedance of the phase-line. These conductors may be round—as we would use for a wire phase-line—or square—as we might use for double-boom construction.

An additional vital output from *LPCAD28* is the conversion of the design into antenna modeling input files of several formats, including versions for *AO* and *NEC4WIN* (both *MININEC*-based programs), and a version in the standard *.NEC format usable by many implementations of *NEC-2* and *NEC-4*, including *NECWin Plus*, *GNEC*, and *EZNEC Pro*. Every proposed LPDA design should be verified and optimized by means of antenna modeling, since basic design calculations rarely provide arrays that require no further work before construction. A one-octave LPDA represents a segment of an arc defined by α that is cut off at both the upper and the lower frequency limits. Moreover, some of the design equations are based upon approximations and do not completely predict LPDA behavior. Despite these limitations, most of the sample LPDA designs shown later in this chapter are based directly upon the fundamental calculations. Therefore, the procedure will be outlined in detail before turning to hybrid log-cell Yagi concepts.

Modeling LPDA designs is most easily done on a version of *NEC*. The transmission line (TL) facility built into *NEC-2* and *NEC-4* alleviates the problem of modeling the phase-line as a set of physical wires, each section of which has a set of constraints in *MININEC* at the right-angle junctions with the elements. Although the *NEC* TL facility does not account for losses in the lines, the losses are ordinarily low enough to neglect.

NEC models do require some careful construction to obtain the most accurate results. Foremost among the cautions is the need for careful segmentation, since each element has a different length. The shortest element should have about 9 or 11 segments, so that it has sufficient segments at the highest modeling frequency for the design. Each element behind the shortest one should have a greater number of segments than the preceding element by the inverse of the value of τ. However, there is a further limitation. Since the transmission line is at the center of each element, *NEC* elements should have an odd number of segments to hold the phase-line centered. Hence, each segmentation value calculated from the inverse of τ must be rounded up to the nearest odd integer.

Initial modeling of LPDAs in *NEC-2* should be done with uniform-diameter elements, with any provision for stepped-diameter element correction turned off. Since these correction factors apply only to elements within about 15% of dipole resonance at the test frequency, models with stepped-diameter elements will correct for only a few elements at any test frequency. The resulting combination of corrected and uncorrected elements will not yield a model with assured reliability.

Once one has achieved a satisfactory model with uniform-diameter elements, the modeling program can be used to calculate stepped-diameter substitutes. Each uniform-diameter element, when extracted from the larger array, will have a resonant frequency. Once this frequency is determined, the stepped-diameter element to be used in final construction can be resonated to the same frequency. Although *NEC-4* handles stepped diameter elements with much greater accuracy than *NEC-2*, the process just described is also applicable to *NEC-4* models for the greatest precision.

LPDA Behavior

Although LPDA behavior is remarkably uniform over a wide frequency range compared to narrow-band designs, such as the Yagi-Uda array, it nevertheless exhibits very significant variations within the design range. **Fig 5** shows several facets of these behaviors. Fig 5 shows the free-space gain for three LPDA designs using 0.5-inch diameter aluminum elements. The designations for each model list the values of τ (0.93, 0.89, and 0.85) and of σ (0.02, 0.04, and 0.06) used to design each array. The resultant array lengths are listed with each designator. The total number of elements varies from 16 for "9302" to 10 for "8904" to 7 for "8504."

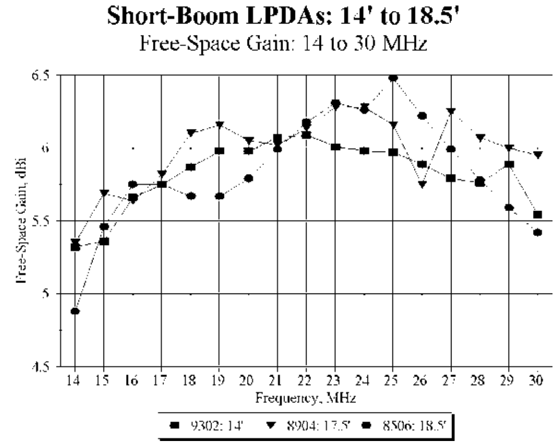

Fig 5—The modeled free-space gain of 3 relatively small LPDAs of different design. Note the relationship of the values of τ and of σ for these arrays with quite similar performance across the 14-30 MHz span.

First, the gain is never uniform across the entire frequency span. The gain tapers off at both the low and high ends of the design spectrum. Moreover, the amount of gain undulates across the spectrum, with the number of peaks dependent upon the selected value of τ and the resultant number of elements. The front-to-back ratio tends to follow the gain level. In general, it ranges from less than 10 dB when the free-space gain is below 5 dBi to over 20 dB as the gain approaches 7 dBi. The front-to-back ratio may reach the high 30s when the free-space array gain exceeds 8.5 dBi. Well-designed arrays, especially those with high values of τ and σ, tended to have well-controlled rear patterns that result in only small differences between the 180° front-to-back ratio and the averaged front-to-rear ratio.

Since array gain is a mutual function of both τ and σ, average gain becomes a function of array length for any given frequency range. Although the gain curves in Fig 5 interweave, there is little to choose among them in terms of average gain for the 14 to 18-foot range of array lengths. Well-designed 20- to 10-meter arrays in the 30-foot array length region are capable of about 7 dBi free-space gain, while 40-foot arrays for the same frequency range can achieve about 8 dBi free-space gain.

Exceeding an average gain of 8.5 dBi requires at least a 50-foot array length for this frequency range. Long arrays with high values of τ and σ also tend to show smaller excursions of gain and of front-to-back ratio in the overall curves. In addition, high-τ designs tend to show higher gain at the low frequency end of the design spectrum.

The frequency sweeps shown in Fig 5 are widely spaced at 1 MHz intervals. The evaluation of a specific design for the 14 to 30-MHz range should decrease the interval between check points to no greater than 0.25 MHz in order to detect frequencies at which the array may show a *performance weakness*. Weaknesses are frequency regions in the overall design spectrum at which the array shows unexpectedly lower values of gain and front-to-back ratio. In Fig 5 note the unexpected decrease in gain of model "8904" at 26 MHz. The other designs also have weak points, but they fall between the frequencies sampled.

In large arrays, these regions may be quite small and may occur in more than one frequency region. The weakness results from the harmonic operation of longer elements to the rear of those expected to have high current levels. Consider a 7-element LPDA about 12.25-foot long for 14 to 30 MHz using 0.5-inch aluminum elements. At 28 MHz, the rear elements operate in a harmonic mode, as shown by the high relative current magnitude curves in **Fig 6**. The result is a radical decrease in gain, as shown in the "No Stub" curve of **Fig 7**. The front-to-back ratio also drops as a result of strong radiation from the long

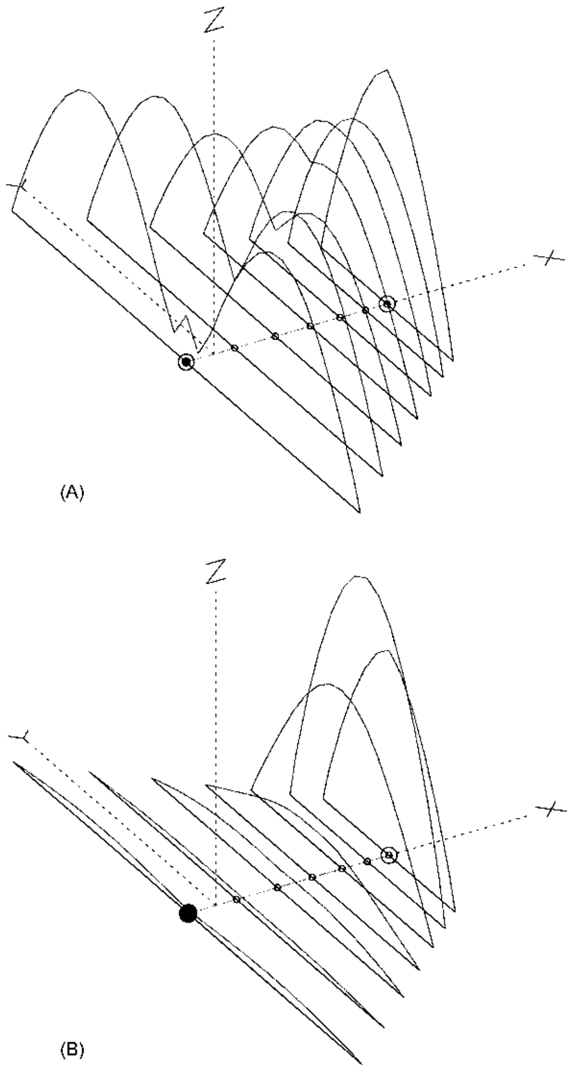

Fig 6—The relative current magnitude on the elements of model "8504" at 28 MHz without and with a stub. Note the harmonic operation of the rear elements before a stub is added to suppress such operation.

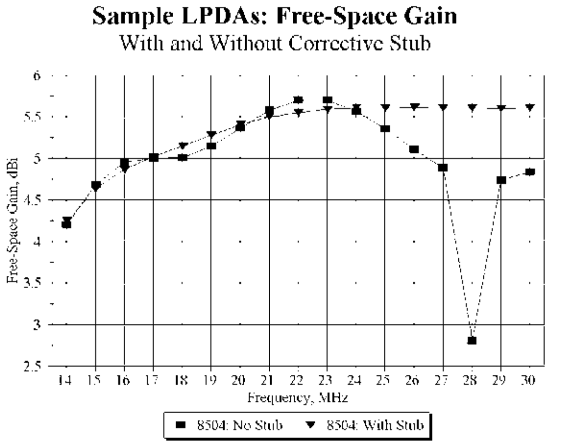

Fig 7—A graph of the gain of model "8504" showing the frequency region in which a "weakness" occurs and its absence once a suitable stub is added to the array.

elements to the rear of the array.

Early designs of LPDAs called for terminating transmission-line stubs as standard practice to help eliminate such weak spots in frequency coverage. In contemporary designs, their use tends to be more specific for eliminating or moving frequencies that show gain and front-to-back weakness. (Stubs have the added function of keeping both sides of each element at the same level of static charge or discharge.) The model dubbed "8504" was fitted (by trial and error) with an 18-inch shorted stub of 600-Ω transmission line. As Fig 6B shows, the harmonic operation of the rear elements is attenuated. The "stub" curve of Fig 7 shows the smoothing of the gain curve for the array throughout the upper half of its design spectrum. In some arrays showing multiple weaknesses, a single stub may not eliminate all of them. However, it may move the weaknesses to unused frequency regions. Where full-spectrum operation of an LPDA is necessary, additional stubs located at specific elements may be needed.

Most LPDA designs benefit (with respect to gain and front-to-back ratio) from the use of larger-diameter elements. Elements with an average diameter of at least 0.5-inch are desirable in the 14 to 30 MHz range. However, standard designs usually presume a constant element length-to-diameter ratio. In the case of *LPCAD28*, this ratio is about 125:1, which assumes an even larger diameter. To achieve a relatively constant length-to-diameter ratio in the computer models, you can set the diameter of the shortest element in a given array design and then increase the element diameter by the inverse of τ for each succeeding longer element. This procedure is often likely to result in unreasonably large element diameters for the longest elements, relative to standard amateur construction practices.

Since most amateur designs using aluminum tubing for elements employ stepped-diameter (tapered) elements, roughly uniform element diameters will result unless the LPDA mechanical design tries to lighten the elements at the forward end of the array. This practice may not be advisable, however. Larger elements at the high end of the design spectrum often counteract (at least partially) the natural decrease in high-frequency gain and show improved performance compared to smaller diameter elements.

An alternative construction method for LPDAs uses wire throughout. At every frequency, single-wire elements reduce gain relative to larger-diameter tubular elements. An alternative to tubular elements appears in **Fig 8**. For each element of a tubular design, there is a roughly equivalent 2-wire element that may be substituted. The

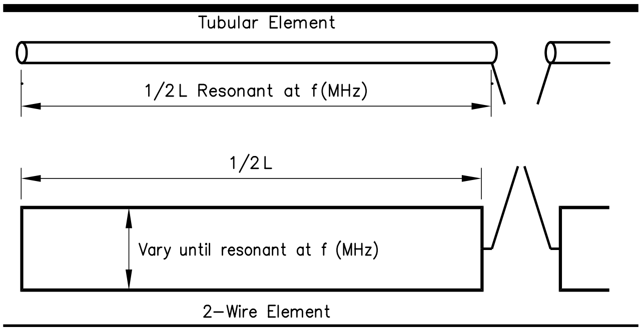
Tubular Element
1/2 L Resonant at f (MHz)
1/2 L
Vary until resonant at f (MHz)
2–Wire Element

Fig 8—A substitute for a large-diameter tubular element composed of two wires shorted at both the outer ends and at the center junction with the phase-line.

spacing between the wire is determined by taking one of the modeled tubular elements and finding its resonant frequency. A two-wire element of the same length is then constructed with shorts at the far ends and at the junctions with the phase-line. The separation of the two wires is adjusted until the wire element is resonant at the same frequency as the original tubular element. The required separation will vary with the wire chosen for the element. Models used to develop these substitutes must pay close attention to segmentation rules for *NEC* due to the short length of segments in the end and center shorts, and to the need to keep segment junctions as exactly parallel as possible with close-spaced wires.

Feeding and Constructing the LPDA

Original design procedures for LPDAs used a single, ordinarily fairly high, characteristic impedance for the phase-line (antenna feeder). Over time, designers realized that other values of impedance for the phase line offered both mechanical and performance advantages for LPDA performance. Consequently, for the contemporary designer, phase-line choice and construction techniques are almost inseparable considerations.

High-impedance phase lines (roughly 200 Ω and higher) are amenable to wire construction similar to that used with ordinary parallel-wire transmission lines. They require careful placement relative to a metal boom used to support individual elements (which themselves must be insulated from the support boom). Connections also require care. If the phase-line is given a half-twist between each element, the construction of the line must ensure constant spacing and relative isolation from metal supports to maintain a constant impedance and to prevent shorts.

Along with the standard parallel-wire line, shown in **Fig 9A**, there are a number of possible LPDA structures using booms. The booms serve both to support the elements and to create relatively low-impedance (under 200 Ω) phase-lines. Fig 9B shows the basics of a twin circular tubing boom with the elements cross-supported by insulated rods. Fig 9C shows the use of square tubing with the elements attached directly to each tube by through-bolts. Fig 9D illustrates the use of L-stock, which may be practical at VHF frequencies. Each of these sketches is incomplete, however, since it omits the necessary stress analyses that determine the mechanical feasibility of a structure for a given LPDA project.

The use of square boom material requires some adjustment when calculating the characteristic impedance of the phase-line. For conductors with a circular cross-section,

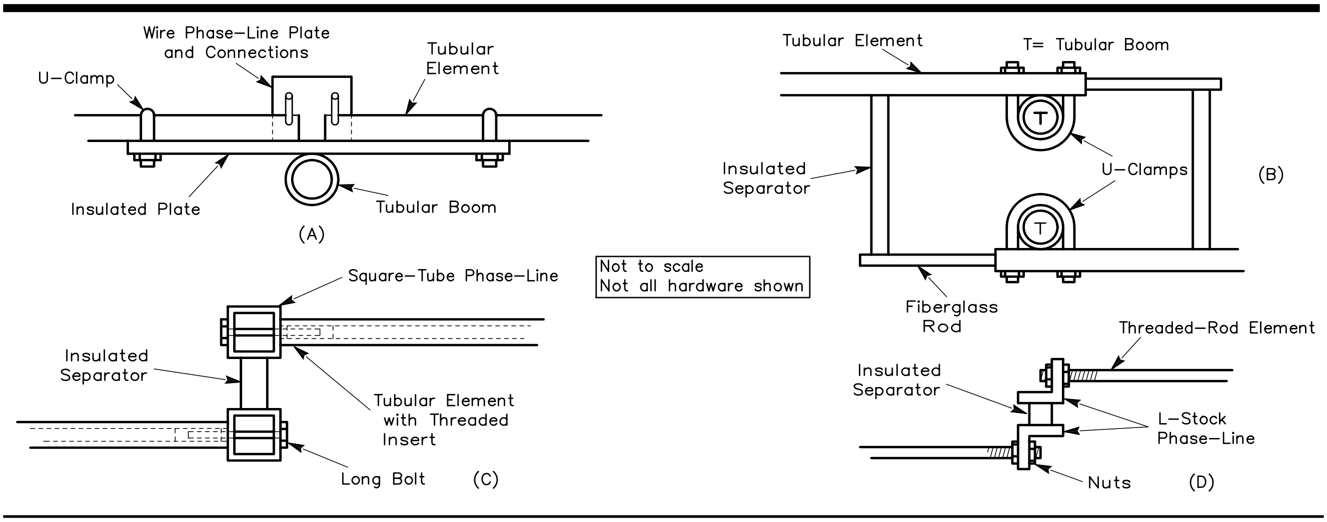
Wire Phase–Line Plate and Connections
U–Clamp
Tubular Element
Insulated Plate
Tubular Boom
(A)
Square–Tube Phase–Line
Insulated Separator
Tubular Element with Threaded Insert
Long Bolt
(C)
Not to scale
Not all hardware shown
Tubular Element
T= Tubular Boom
T
Insulated Separator
U–Clamps
(B)
Fiberglass Rod
Threaded–Rod Element
Insulated Separator
L–Stock Phase–Line
Nuts
(D)

Fig 9—Four (of many) possible construction techniques, shown from the array end. In A, an insulated plate supports and separates the wires of the phase-line, suitable with wire or tubular elements. A dual circular boom phase-line also supports the elements, which are cross-supported for boom stability. Square tubing is used in C, with the elements joined to the boom/phase-line with through-bolts and an insert in each half element. The L-stock shown in D is useful for lighter VHF and UHF arrays.

$$Z_0 = 120 \cosh^{-1} \frac{D}{d} \tag{Eq 4}$$

where D is the center-to-center spacing of the conductors and d is the outside diameter of each conductor, both expressed in the same units of measurement. Since we are dealing with closely spaced conductors, relative to their diameters, the use of this version of the equation for calculating the characteristic impedance (Z_0) is recommended. For a square conductor,

$$d \approx 1.18\ w \tag{Eq 5}$$

where d is the approximate equivalent diameter of the square tubing and w is the width of the tubing across one side. Thus, for a given spacing, a square tube permits you to achieve a lower characteristic impedance than round conductors. However, square tubing requires special attention to matters of strength, relative to comparable round tubing.

Electrically, the characteristic impedance of the LPDA phase-line tends to influence other performance parameters of the array. Decreasing the phase-line Z_0 also decreases the feed-point impedance of the array. For small designs with few elements, the decrease is not fully matched by a decrease in the excursions of reactance. Consequently, using a low impedance phase-line may make it more difficult to achieve a 2:1 or less SWR for the entire frequency range. However, higher-impedance phase-lines may result in a feed-point impedance that requires the use of an impedance-matching balun.

Decreasing the phase-line Z_0 also tends to increase LPDA gain and front-to-back ratio. There is a price to be paid for this performance improvement—weaknesses at specific frequency regions become much more pronounced with reductions in the phase-line Z_0. For a specific array you must weigh carefully the gains and losses, while employing one or more transmission line stubs to get around performance weaknesses at specific frequencies.

Depending upon the specific values of τ and σ selected for a design, you can sometimes select a phase-line Z_0 that provides either a 50-Ω or a 75-Ω feed-point impedance, holding the SWR under 2:1 for the entire design range of the LPDA. The higher the values of τ and σ for the design, the lower the reactance and resistance excursions around a central value. Designs using optimal values of σ with high values of τ show a very slight capacitive reactance throughout the frequency range. Lower design values obscure this phenomenon due to the wide range of values taken by both resistance and reactance as the frequency is changed.

At the upper end of the frequency range, the source resistance value decreases more rapidly than elsewhere in the design spectrum. In larger arrays, this can be overcome by using a variable Z_0 phase-line for approximately the first 20% of the array length. This technique is, however, difficult to implement with anything other than wire phase-lines. Begin with a line impedance about half of the final value and increase the wire spacing evenly until it reaches its final and fixed spacing. This technique can sometimes produce smoother impedance performance across the entire frequency span and improved high-frequency SWR performance.

Designing an LPDA requires as much attention to designing the phase-line as to element design. It is always useful to run models of the proposed design through several iterations of possible phase-line Z_0 values before freezing the structure for construction.

Special Design Corrections

The curve for the sample 8504 LPDA in Fig 7 revealed several deficiencies in standard LPDA designs. The weakness in the overall curve was corrected by the use of a stub to eliminate or move the frequency at which rearward elements operated in a harmonic mode. In the course of describing the characteristics of the array, we have noted several other means to improve performance. Fattening elements (either uniformly or by increasing their diameter in step with τ) and reducing the characteristic impedance of the phase-line are capable of small improvements in performance. However, they cannot wholly correct the tendency of the array gain and front-to-back ratio to fall off at the upper and lower limits of the LPDA frequency range.

One technique sometimes used to improve performances near the frequency limits is to design the LPDA for upper and lower frequency limits much higher and lower than the frequencies of use. This technique unnecessarily increases the overall size of the array and does not eliminate the downward performance curves. Increasing the values of τ and σ will usually improve perfor-

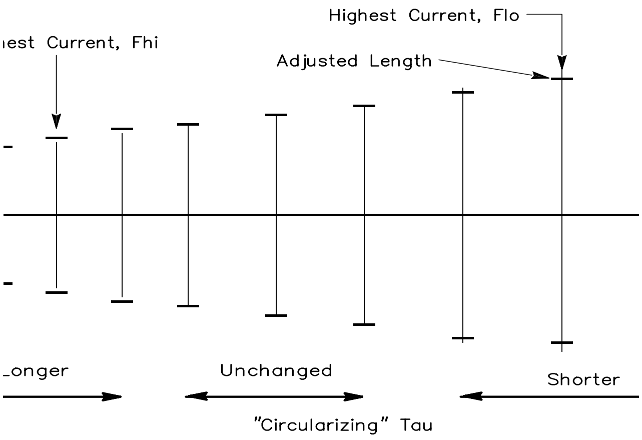

Fig 10—A before and after sketch of an LPDA, showing the original lengths of the elements and their adjustments from diminishing the value of τ at both ends of the array. See the text for the amount of change applicable to each element.

mance at no greater cost in size than extending the frequency range. Increasing the value of τ is especially effective in improving the low frequency performance of an LPDA.

Working within the overall size limits of a standard design, one may employ a technique of *circularizing* the value of τ for the rear-most and forward-most elements. See **Fig 10**, which is not to-scale relative to overall array length and width. Locate (using an antenna-modeling program) the element with the highest current at the lowest operating frequency, and the element with the highest current at the highest operating frequency. The adjustments to element lengths may begin with these elements or—at most—one element further toward the array center. For the first element (counting from the center) to be modified, reduce the value of τ by about 0.5%. For a rearward element, use the inverse of the adjusted value of τ to calculate the new length of the element relative to the unchanged element just forward of the change. For a forward element, use the new value of τ to calculate the new length of the element relative to the unchanged element immediately to the rear of it.

For succeeding elements outward, calculate new values of τ from the adjusted values, increasing the increment of decrease with each step. Second adjusted elements may use values of τ about 0.75% to 1.0% lower than the values just calculated. Third adjusted elements may use an increment of 1.0% to 1.5% relative to the preceding value.

Not all designs require extensive treatment. As the values of τ and σ increase, fewer elements may require adjustment to obtain the highest possible gain at the frequency limits, and these will always be the most outward elements in the array. A second caution is to check the feed-point impedance of the array after each change to ensure that it remains within design limits.

Fig 11 shows the free-space gain curves from 14 to 30 MHz for a 10-element LPDA with an initial τ of 0.89 and a σ of 0.04. The design uses a 200-Ω phase-line, 0.5-inch aluminum elements, and a 3-inch 600-Ω stub. The lowest curve shows the modeled performance across the design frequency range with only the stub. Performance at the frequency limits is visibly lower than within the peak performance region. The middle curve shows the effects of circularizing τ. Average performance levels have improved noticeably at both ends of the spectrum.

In lieu of, or in addition to, the adjustment of element lengths, you may also add a parasitic director to an LPDA, as shown in **Fig 12**. The director is cut roughly for the highest operating frequency. It may be spaced between 0.1 λ and 0.15 λ from the forward-most element of the LPDA. The exact length and spacing should be determined experimentally (or from models) with two factors in mind. First, the element should not adversely effect the feed-point impedance at the highest operating frequencies. Close spacing of the director has the greatest effect on this impedance. Second, the exact spacing and element length should be set to have the most desired effect on the overall performance curve of the array. The mechanical impact of adding a director is to increase overall array length by the spacing selected for the element.

The upper curve in Fig 11 shows the effect of adding a director to the circularized array already equipped with a stub. The effect of the director is cumulative, increasing the upper range gain still further. Note that the added parasitic director is not just effective at the highest

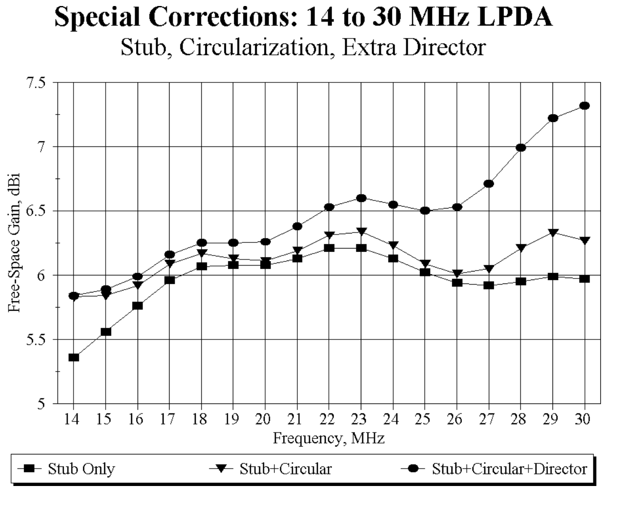

Fig 11—The modeled free space gain from 14 to 30 MHz of an LPDA with τ of 0.89 and σ of 0.04. Squares: just a stub to eliminate a weakness; Triangles: with a stub and circularized elements, and Circles: with a stub, circularized elements and a parasitic director.

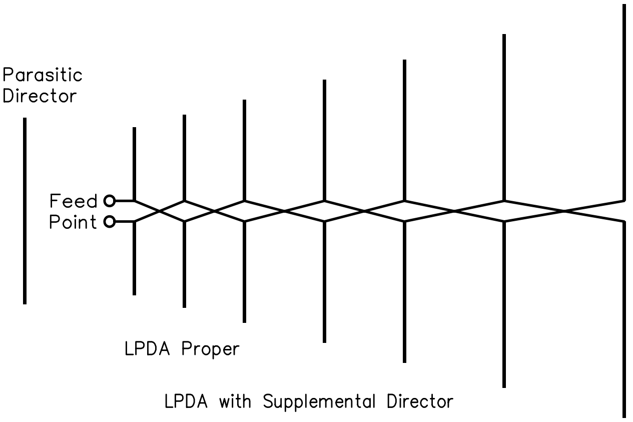

Fig 12—A generalized sketch of an LPDA with the addition of a parasitic director to improve performance at the higher frequencies within the design range.

frequencies within the LPDA design range. It has a perceptible effect almost all the way across the frequency span of the array, although the effect is smallest at the low-frequency end of the range.

The addition of a director can be used to enhance upper frequency performance of an LPDA, as in the illustration, or simply to equalize upper frequency performance with mid-range performance. High-τ designs, with good low-frequency performance, may need only a director to compensate for high-frequency gain decrease. One potential challenge to adding a director to an LPDA is sustaining a high front-to-back ratio at the upper frequency range.

Throughout the discussion of LPDAs, the performance curves of sample designs have been treated at all frequencies alike, seeking maximum performance across the entire design frequency span. Special compensations are also possible for ham-band-only LPDA designs. They include the insertion of parasitic elements within the array as well as outside the initial design boundaries. In addition, stubs may be employed not so much to eliminate weaknesses, but only to move them to frequencies outside the range of amateur interests.

A DESIGN PROCEDURE FOR AN LPDA

The following presents a systematic step-by-step design procedure for an LPDA with any desired bandwidth. The procedure requires some mathematical calculations, but a common calculator with square-root, logarithmic, and trigonometric functions is completely adequate. The notation used in this section may vary slightly from that used earlier in this chapter.

1) Decide on an operating bandwidth B between f_1, lowest frequency and f_n, highest frequency:

$$B = \frac{f_n}{f_1} \qquad \text{(Eq 6)}$$

2) Choose τ and σ to give the desired estimated average gain.

$$0.8 \le \tau \le 0.98 \text{ and } 0.03 \le \sigma \le \sigma_{opt} \qquad \text{(Eq 7)}$$

where σ_{opt} is calculated as noted earlier in this chapter.

3)Determine the value for the cotangent of the apex half-angle α from

$$\cot \alpha = \frac{4\sigma}{1-\tau} \qquad \text{(Eq 8)}$$

Although α is not directly used in the calculations, cot α is used extensively.

4) Determine the bandwidth of the active region B_{ar} from

$$B_{ar} = 1.1 + 7.7\left(1-\tau\right)^2 \cot \alpha \qquad \text{(Eq 9)}$$

5) Determine the structure (array) bandwidth B_s from

$$B_S = B \times B_{ar} \qquad \text{(Eq 10)}$$

6) Determine the boom length L, number of elements N, and longest element length ℓ_1.

$$L_n = \left(1 - \frac{1}{B_S}\right) \cot \alpha \times \frac{\lambda_{max}}{4} \qquad \text{(Eq 11)}$$

$$\lambda_{max} = \frac{984}{f_1} \qquad \text{(Eq 12)}$$

$$N = 1 + \frac{\log B_s}{\log \frac{1}{\tau}} = 1 + \frac{\ln B_S}{\ln \frac{1}{\tau}} \qquad \text{(Eq 13)}$$

$$\lambda_{1ft} = \frac{492}{f_1} \qquad \text{(Eq 14)}$$

Usually the calculated value for N will not be an integral number of elements. If the fractional value is more than about 0.3, increase the value of N to the next higher integer. Increasing the value of N will also increase the actual value of L over that obtained from the sequence of calculations just performed.

Examine L, N and l_1 to determine whether or not the array size is acceptable for your needs. If the array is too large, increase f_1 or decrease σ or τ and repeat steps 2 through 6. Increasing f_1 will decrease all dimensions. Decreasing σ will decrease the boom length. Decreasing τ will decrease both the boom length and the number of elements.

7) Determine the terminating stub Z_t. (Note: For many HF arrays, you may omit the stub, short out the longest element with a 6-inch jumper, or design a stub to overcome a specific performance weakness.) For VHF and UHF arrays calculate the stub length from

$$Z_t = \frac{\lambda_{max}}{8} \qquad \text{(Eq 15)}$$

8) Solve for the remaining element lengths from

$$\ell_n = \tau \ell_{n-1} \qquad \text{(Eq 16)}$$

9) Determine the element spacing $d_{1\text{-}2}$ from

$$d_{1-2} = \frac{\left(\ell_1 - \ell_2\right) \cot \alpha}{2} \qquad \text{(Eq 17)}$$

where ℓ_1 and ℓ_2 are the lengths of the rearmost elements, and d_{1-2} is the distance between the elements with the lengths ℓ_1 and ℓ_2. Determine the remaining element-to-element spacings from

$$d_{(n-1)-n} = \tau \, d_{(n-2)-(n-1)} \qquad \text{(Eq 18)}$$

10) Choose R_0, the desired feed-point resistance, to give the lowest SWR for the intended balun ratio and feedline impedance. R_0, the mean radiation resistance level of the LPDA input impedance, is approximated by:

$$R_0 = \frac{Z_0}{\sqrt{1+\frac{Z_0}{4\sigma' Z_{AV}}}} \qquad \text{(Eq 19)}$$

where the component terms are defined and/or calculated in the following way.

From the following equations, determine the necessary antenna feeder (phase-line) impedance, Z_0:

$$Z_0 = \frac{{R_0}^2}{8\sigma' Z_{AV}} + R_0\sqrt{\left(\frac{R_0}{8\sigma' Z_{AV}}\right)^2+1} \qquad \text{(Eq 20)}$$

σ' is the mean spacing factor and is given by

$$\sigma' = \frac{\sigma}{\sqrt{\tau}} \qquad \text{(Eq 21)}$$

Z_{AV} is the average characteristic impedance of a dipole and is given by

$$Z_{AV} = 120\left[\mathrm{1n}\left(\frac{\ell_n}{\mathrm{diam}_n}\right) - 2.25\right] \qquad \text{(Eq 22)}$$

The ratio, l_n/diam_n is the length-to-diameter ratio of the element n.

11) Once Z_0 has been determined, select a combination of conductor size and spacing to achieve that impedance, using the appropriate equation for the shape of the conductors. If an impractical spacing results for the antenna feeder, select a different conductor diameter and repeat step 11. In severe cases it may be necessary to select a different R_0 and repeat steps 10 and 11. Once a satisfactory feeder arrangement is found, the LPDA design is complete.

A number of the LPDA design examples at the end of this chapter make use of this calculation method. However, the resultant design should be subjected to extensive modeling tests to determine whether there are performance deficiencies or weaknesses that require modification of the design before actual construction.

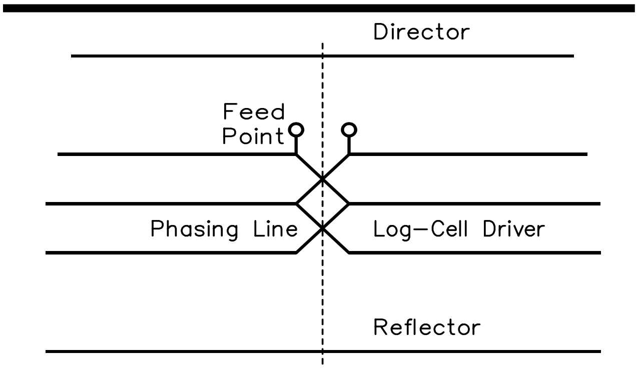

Fig 13—A sketch of a typical monoband log-cell Yagi. The reflector is, in principle, optional. The log-cell may have from 2 to 5 (or more elements). There may be one or more directors.

Log-Cell Yagis

Fig 12 showed an LPDA with an added parasitic director. Technically, this converts the original design into a hybrid Log-Yag. However, the term *Log-Yag* (or more generally the *log-cell Yagi*) is normally reserved for monoband designs that employ two or more elements in a single-band LPDA arrangement, together with (usually) a reflector and one or more directors. The aim is to produce a monoband directive array with superior directional qualities over a wider bandwidth than can be obtained from many Yagi-Uda designs. Log-cells have also been successfully used as wide-band driver sections for multiband Yagi beams.

Fig 13 illustrates the general outline of a typical log-cell Yagi. The driver section consists of a log periodic array designed for the confines of a single amateur band or other narrow range of frequencies. The parasitic reflector is usually spaced about 0.085 λ behind the rear element of the log cell, while the parasitic director is normally placed between 0.13 and 0.15 λ ahead of the log cell.

Early log-cell Yagis tended to be casually designed. Most of these designs have inferior performance compared with present-day computer-optimized Yagis of the same boom length. Some were designed by adding one or more parasitic directors to simple phased pairs of elements. Although good performance is possible, the operating bandwidth of these designs is small, suitable only for the so-called WARC bands. However, when the log-cell is designed as a narrowly spaced monoband log periodic array, the operating bandwidth increases dramatically. Operating bandwidth here refers not just to the SWR bandwidth, but also to the gain and front-to-back bandwidth.

The widest operating bandwidths require log cells of 3 to 4 elements for HF bands like 20 meters, and 4 to 5 elements for bands as wide as 10 meters. (The bandwidth of the 20-meter band is approximately 2.4% of the center frequency, while the bandwidth of the 10-meter band approaches 9.4% of the center frequency.) A practical limit to σ for log cells used within parasitic arrays is about 0.05. Slightly higher gain may be obtained from higher values of σ, but at the cost of a much longer log cell. The limiting figure for σ results in a practical value for τ between 0.94 and 0.95 to achieve a cell with the desired bandwidth characteristics.

An array designed according to these principles has an overall length that varies with the size of the log cell. A typical array with a 4-element log-cell and single parasitic elements fore and aft is a bit over 0.35 λ long, while a 5-element log-cell Yagi will be between 0.4 and 0.45 λ long. Spacing the reflector more widely (for example, up

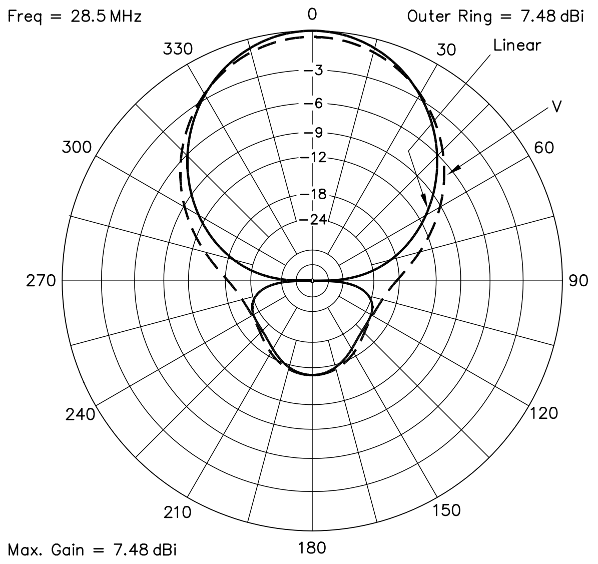

Fig 14—Overlaid free-space azimuth patterns for virtually identical 5-element log-cell Yagis (with 3-element log-cells), one having the elements V'd forward at 40° from linear (dashed), the other with linear elements (solid).

to 0.25 λ) has little effect on either gain or front-to-back ratio. Wider spacing of the director will also have only a small effect on gain, since the arrangement is already close to the boom-length limit recommended for director-driver-reflector arrays. Further lengthening of the boom should be accompanied by the addition of one or more directors to the array, if additional gain is desired from the design.

Compared to a modern-day Yagi of the same boom length, the log-cell Yagi is considerably heavier and exhibits a higher wind load due to the requirements of the log-cell driver. Yagis with 3- and 4-elements within the boom lengths just given are capable of peak free-space gain values of 8.2 to 8.5 dBi, while sustaining a high front-to-back ratio. Peak front-to-back ratios are typically in the vicinity of 25 dB. However, Yagi gain tends to decrease below the design frequency (and increase above it), while the front-to-back ratio tends to taper off as one moves away from the design frequency. For the largest log-cell sizes, log-cell Yagis of the indicated boom lengths are capable of sustaining at least 8.2 dBi free-space gain over the entire band, with front-to-back ratios of over 30 dB across the operating bandwidth.

The feed-point impedance of a log-cell Yagi is a function of both the cell design and the influence of the parasitic elements. However, for most cell designs and common phase-line designs, you can achieve a very low variation of resistance and reactance across a desired band. In many cases, the feed impedance will form a direct match for the standard 50-Ω coaxial cable used by most amateur installations. (In contrast, the high-gain, high-front-to-back Yagis used for comparison here have feed-point impedances ranging from 20 to 25 Ω.)

A common design technique used in some LPDA and log-cell Yagi designs is to bend the elements forward to form a series of Vs. A forward angle on each side of the array centerline of about 40° relative to a linear element has been popular. In some instances, the mechanical design of the array may dictate this element formation. However, this arrangement has no special benefits and possibly may degrade performance.

Fig 14 shows the free-space azimuth patterns of a single 5-element log-cell Yagi in two versions: with the elements linear and with the elements bent forward 40°. The V-array loses about $^{1}/_{2}$ dB gain, but more significantly, it loses considerable signal rejection from the sides. Similar comparisons can be obtained from pure LPDA designs and from Yagi-Uda designs when using elements in the vicinity of $^{1}/_{2}$ λ. Unless mechanical considerations call for arranging the elements in a V, the technique is not recommended.

Ultimately, the decision to build and use a log-cell Yagi involves balancing the additional weight and wind-load requirements of this design against the improvements in operating bandwidth for all of the major operating parameters, especially with respect to the front-to-back ratio and the feed-point impedance.

Wire Log-Periodic Dipole Arrays for 3.5 or 7 MHz

These wire log-periodic dipole arrays for the lower HF bands are simple in design and easy to build. They are designed to have reasonable gain, be inexpensive and lightweight, and may be assembled with stock items found in large hardware stores. They are also strong—they can withstand a hurricane! These antennas were first described by John J. Uhl, KV5E, in *QST* for August, 1986. **Fig 15** shows one method of installation. You can use the information here as a guide and point of reference for building similar LPDAs.

If space is available, the antennas can be rotated or repositioned in azimuth after they are completed. A 75-foot tower and a clear turning radius of 120 feet around the base of the tower are needed. The task is simplified

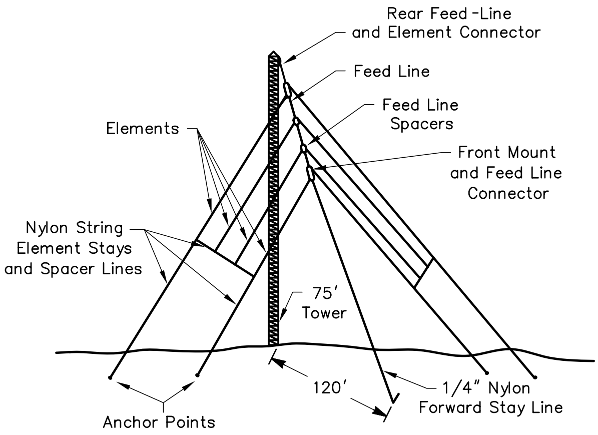

Fig 15—Typical lower-HF wire 4-element log periodic dipole array erected on a tower.

if you use only three anchor points, instead of the five shown in Fig 15. Omit the two anchor points on the forward element, and extend the two nylon strings used for element stays all the way to the forward stay line.

DESIGN OF THE LOG-PERIODIC DIPOLE ARRAYS

Design constants for the two arrays are listed in **Tables 1** and **2**. The preceding sections of this chapter contain the design procedure for arriving at the dimensions and other parameters of these arrays. The primary differences between these designs and one-octave upper HF arrays are the narrower frequency ranges and the use of wire, rather than tubing, for the elements. As design examples for the LPDA, you may wish to work through the step-by-step procedure and check your results against the values in Tables 1 and 2. You may also wish to compare these results with the output of an LPDA design software package such as *LPCAD28*.

From the design procedure, the feeder wire spacings for the two arrays are slightly different, 0.58 inch for the 3.5-MHz array and 0.66 inch for the 7-MHz version. As a compromise toward the use of common spacers for both bands, a spacing of 5/8 inch is quite satisfactory. Surprisingly, the feeder spacing is not at all critical here from a matching standpoint, as may be verified from $Z_0 = 276 \log (2S/\text{diam})$ and from Eq 4. Increasing the spacing to as much as 3/4 inch results in an R_0 SWR of less than 1.1:1 on both bands.

Constructing the Arrays

Construction techniques are the same for both the 3.5 and the 7-MHz versions of the array. Once the designs are completed, the next step is to fabricate the fittings; see **Fig 16** for details. Cut the wire elements and feed lines to the proper sizes and mark them for identification. After the wires are cut and placed aside, it will be difficult to remember which is which unless they are marked. When you have finished fabricating the connectors and cutting all of the wires, the antenna can be assembled. Use your ingenuity when building one of these antennas; it isn't necessary to duplicate these LPDAs precisely.

The elements are made of standard #14 stranded cop-

Table 1

Design Parameters for the 3.5-MHz Single-Band LPDA

f1 = 3.3 MHz
f_n = 4.1 MHz
B = 1.2424
τ = 0.845
σ = 0.06
Gain = 5.9 dBi = 3.8 dBd
cot α = 1.5484
B_{ar} = 1.3864
B_s = 1.7225
L = 48.42 feet
N = 4.23 elements (decrease to 4)
Zt = 6-inch short jumper
R_0 = 208 Ω
Z_{AV} = 897.8 Ω
σ' = 0.06527
Z_0 = 319.8 Ω
Antenna feeder:
#12 wire spaced 0.58 inches
Balun: 4:1
Feed line: 52-Ω coax

Element lengths:
$\ell 1$ = 149.091 feet
$\ell 2$ = 125.982 feet
$\ell 3$ = 106.455 feet
$\ell 4$ = 89.954 feet
Element spacings:
d_{12}=17.891 feet
d_{23} = 15.118 feet
d_{34} = 12.775 feet
Element diameters
All = 0.0641 inches
ℓ/diameter ratios:
ℓ/diam_4 = 16840
ℓ/diam_3 = 19929
ℓ/diam_2 = 23585
ℓ/diam_1 = 27911

Table 2

Design Parameters for the 7-MHz Single-Band LPDA

f1 = 6.9 MHz
f_n = 7.5 MHz
B = 1.0870
τ = 0.845
σ = 0.06
Gain = 5.9 dBi = 3.8 dBd
cot α = 1.5484
B_{ar} = 1.3864
B_s = 1.5070
L = 18.57 feet
N = 3.44 elements (increase to 4)
Z_t = 6-inch short jumper
R_0 = 208 Ω
Z_{AV} = 809.3 Ω
σ' = 0.06527
Z_0 = 334.2 Ω
Antenna feeder:
#12 wire spaced 0.66 inches
Balun: 4:1
Feed line: 52-Ω coax

Element lengths:
$\ell 1$ = 71.304 feet
$\ell 2$ = 60.252 feet
$\ell 3$ = 50.913 feet
$\ell 4$ = 43.022 feet
Element spacings:
d_{12} = 8.557 feet
d_{23} = 7.230 feet
d_{34} = 6.110 feet
Element diameters:
All = 0.0641 inches
ℓ/diameter ratios:
$\ell 4/\text{diam}_4$ = 8054
$\ell 3/\text{diam}_3$ = 9531
$\ell 2/\text{diam}_2$ = 11280
$\ell 1/\text{diam}_1$ = 13349

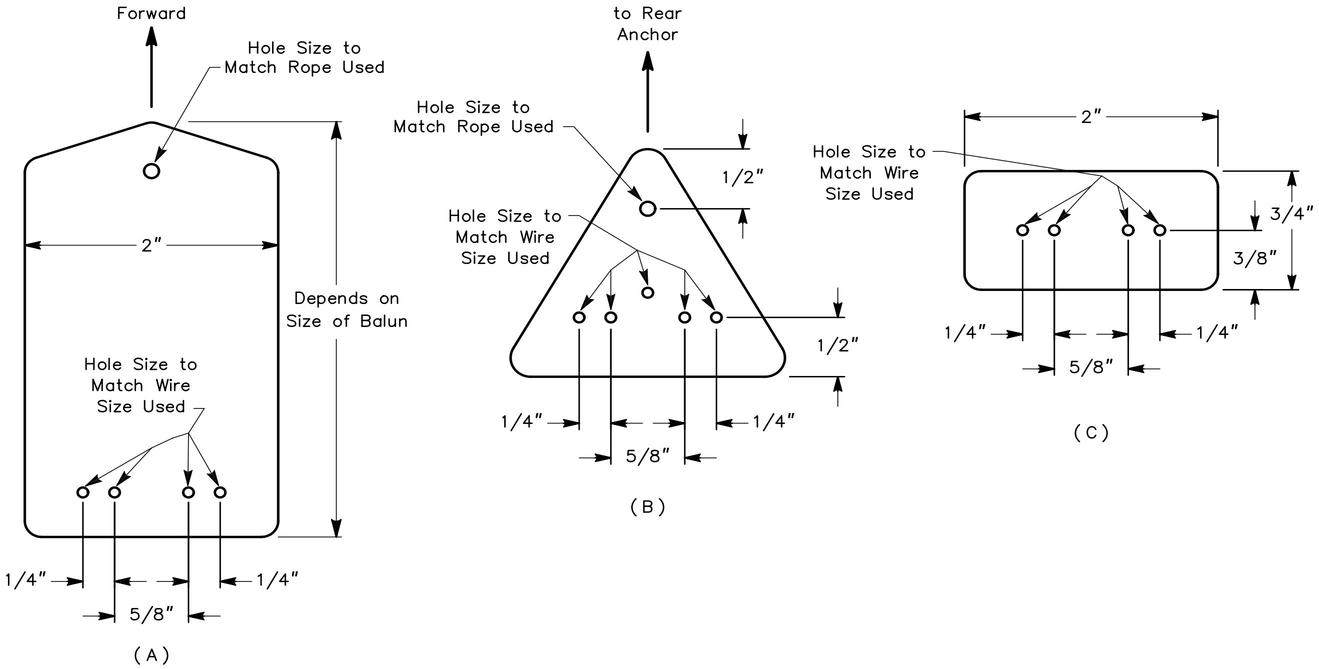
Forward
Hole Size to Match Rope Used
2"
Depends on Size of Balun
Hole Size to Match Wire Size Used
1/4"
1/4"
5/8"
(A)
to Rear Anchor
Hole Size to Match Rope Used
1/2"
Hole Size to Match Wire Size Used
1/2"
1/4"
1/4"
5/8"
(B)
2"
Hole Size to Match Wire Size Used
3/4"
3/8"
1/4"
1/4"
5/8"
(C)

Fig 16—Pieces for the LPDA that require fabrication. At A is the forward connector, made from $^1/_2$-inch Lexan (polycarbonate). At B is the rear connector, also made from $^1/_2$-inch Lexan. At C is the pattern for the phase-line spacers, made from $^1/_4$-inch Plexiglas. Two spacers are required for the array.

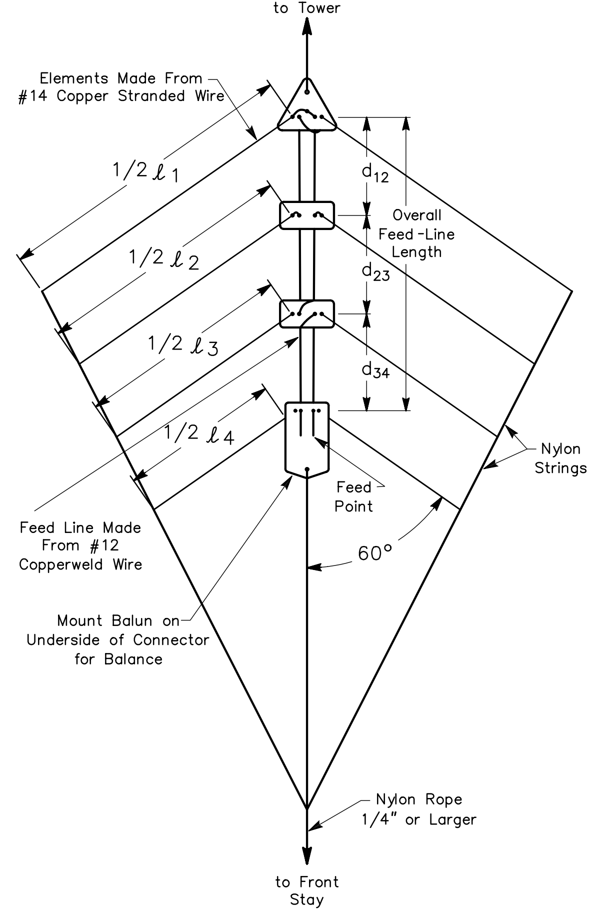
to Tower
Elements Made From #14 Copper Stranded Wire
1/2 ℓ 1
1/2 ℓ 2
1/2 ℓ 3
1/2 ℓ 4
d12
d23
d34
Overall Feed-Line Length
Nylon Strings
Feed Point
60°
Feed Line Made From #12 Copperweld Wire
Mount Balun on Underside of Connector for Balance
Nylon Rope 1/4" or Larger
to Front Stay

Fig 17—The generic layout for the lower HF wire LPDA. Use a 4:1 balun on the forward connector. See Tables 1 and 2 for dimensions.

per wire. The two parallel feed lines are made of #12 solid copper-coated steel wire, such as Copperweld. Copperweld will not stretch when placed under tension. The front and rear connectors are cut from $^1/_2$-inch thick Lexan sheeting, and the feed-line spacers from $^1/_4$-inch Plexiglas sheeting.

Study the drawings carefully and be familiar with the way the wire elements are connected to the two feed lines, through the front, rear and spacer connectors. Details are sketched in **Figs 17** and **18**. Connections made in the way shown in the drawings prevent the wire from breaking. All of the rope, string, and connectors must be made of materials that can withstand the effects of tension and weathering. Use nylon rope and strings, the type that yachtsmen use. Fig 15 shows the front stay rope coming down to ground level at a point 120 feet from the base of a 75-foot tower. Space may not be available for this arrangement in all cases. An alternative installation technique is to put a pulley 40 feet up in a tree and run the front stay rope through the pulley and down to ground level at the base of the tree. The front stay rope will have to be tightened with a block and tackle at ground level.

Putting an LPDA together is not difficult if it is assembled in an orderly manner. It is easier to connect the elements to the feeder lines when the feed-line

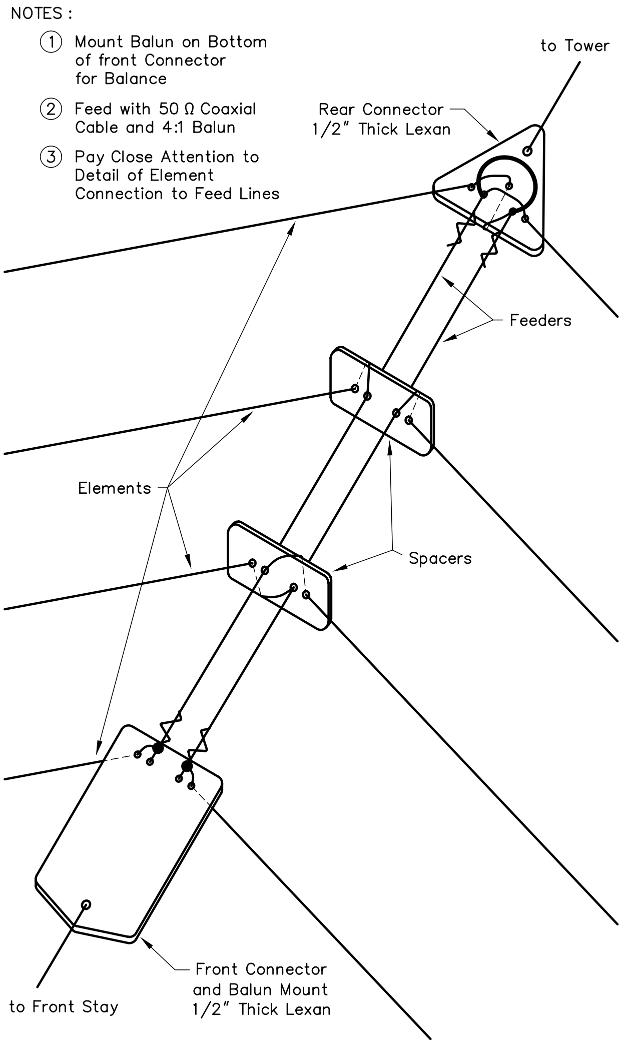
NOTES :
① Mount Balun on Bottom of front Connector for Balance
② Feed with 50 Ω Coaxial Cable and 4:1 Balun
③ Pay Close Attention to Detail of Element Connection to Feed Lines
to Tower
Rear Connector 1/2" Thick Lexan
Feeders
Elements
Spacers
Front Connector and Balun Mount 1/2" Thick Lexan
to Front Stay

Fig 18—Details of the electrical and mechanical connections of the elements to the phase-line. Knots in the nylon rope stay line are not shown.

assembly is stretched between two points. Use the tower and a block and tackle. Attaching the rear connector to the tower and assembling the LPDA at the base of the tower makes raising the antenna into place a much simpler task. Tie the rear connector securely to the base of the tower and attach the two feeder lines to it. Then thread the two feed-line spacers onto the feed line. The spacers will be loose at this time, but will be positioned properly when the elements are connected. Now connect the front connector to the feed lines. A word of caution: Measure accurately and carefully! Double-check all measurements before you make permanent connections.

Connect the elements to the feeder lines through their respective plastic connectors, beginning with element 1, then element 2, and so on. Keep all of the element wires securely coiled. If they unravel, you will have a tangled mess of kinked wire. Recheck the element-to-feeder connections to ensure proper and secure junctions. (See Figs 17 and 18.) Once you have completed all of the element connections, attach the 4:1 balun to the underside of the front connector. Connect the feeder lines and the coaxial cable to the balun.

You will need a separate piece of rope and a pulley to raise the completed LPDA into position. First secure the eight element ends with nylon string, referring to Figs 15 and 17. The string must be long enough to reach the tie-down points. Connect the front stay rope to the front connector, and the completed LPDA is now ready to be raised into position. While raising the antenna, uncoil the element wires to prevent their getting away and tangling up into a mess. Use care! Raise the rear connector to the proper height and attach it securely to the tower, then pull the front stay rope tight and secure it. Move the elements so they form a 60° angle with the feed lines, in the direction of the front, and space them properly relative to one another. By adjusting the end positions of the elements as you walk back and forth, you will be able to align all the elements properly. Now it is time to hook your rig to the system and make some contacts.

Performance

The reports received from these LPDAs were compared with an inverted-V dipole. All of the antennas are fixed; the LPDAs radiate to the northeast, and the dipole to the northeast and southwest. The apex of the dipole is at 70 feet, and the 40- and 80-meter LPDAs are at 60 and 50 feet, respectively. Basic array gain was apparent from many of the reports received. During pileups, it was possible to break in with a few tries on the LPDAs, yet it was impossible to break in the same pileups using the dipole. The gain of the LPDAs is several dB over the dipole. For additional gain, experimenters may wish to try a parasitic director about $^{1}/_{8}\,\lambda$ ahead of the array. Director length and spacing from the forward LPDA element should be field-adjusted for maximum performance while maintaining the impedance match across each of the bands.

Wire LPDA systems offer many possibilities. They are easy to design and to construct: real advantages in countries where commercially built antennas and parts are not available at reasonable cost. The wire needed can be obtained in all parts of the world, and cost of construction is low. If damaged, the LPDAs can be repaired easily with pliers and solder. For those who travel on DXpeditions where space and weight are large considerations, LPDAs are lightweight but sturdy, and they perform well.

5-Band Log Periodic Dipole Array

A rotatable log periodic array designed to cover the frequency range from 13 to 30 MHz is pictured in **Fig 19**. This is a large array having a free-space gain that varies from 6.6 to over 6.9 dBi, depending upon the operating portion of the design spectrum. This antenna system was originally described by Peter D. Rhodes, WA4JVE, in Nov 1973 *QST*. A measured radiation pattern for the array appears in **Fig 20**.

The characteristics of this array are:

1) Half-power beamwidth, 43° (14 MHz)
2) Design parameter $\tau = 0.9$
3) Relative element spacing constant $\sigma = 0.05$
4) Boom length, L = 26 feet
5) Longest element $\lambda 1$ = 37 feet 10 inches. (A tabulation

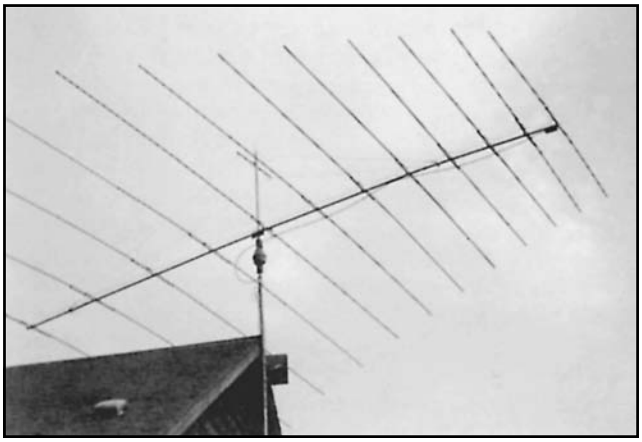

Fig 19—The 13-30 MHz log periodic dipole array.

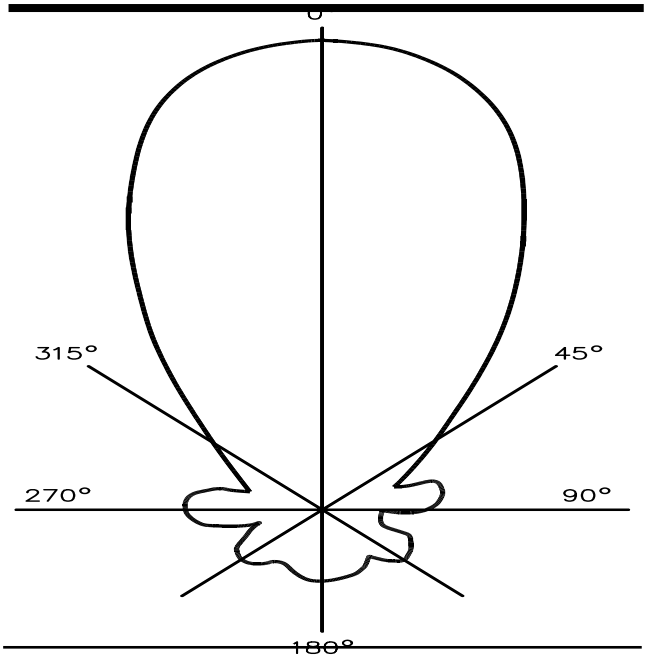

Fig 20—Measured radiation pattern of the 13-30 MHz LPDA. The front-to-back ratio is about 14 dB at 14 MHz and increases to 21 dB at 28 MHz.

Table 3
13-30 MHz LPDA Dimensions, feet

Ele. No.	*Length*	$d_{n-1,n}$ *(spacing)*	*Nearest Resonant*
1	37′ 10.2″	—	
2	34′ 0.7″	3′ 9.4″ = d_{12}	14 MHz
3	30′ 7.9″	3′ 4.9″ = d_{23}	
4	27′ 7.1″	3′ 0.8″ = d_{34}	
5	24′ 10.0″	2′ 9.1″ = d_{45}	18 MHz
6	22′ 4.2″	2′ 5.8″ = d_{56}	21 MHz
7	20′ 1.4″	2′ 2.8″ = d_{67}	
8	18′ 1.2″	2′ 0.1″ = d_{78}	24.9 MHz
9	16′ 3.5″	1′ 9.7″ = d_{89}	28 MHz
10	14′ 7.9″	1′ 7.5″ = $d_{9,10}$	
11	13′ 2.4″	1′ 5.6″ = $d_{10,11}$	
12	11′ 10.5″	1′ 3.8″ = $d_{11,12}$	

Table 4
Materials list: 13-30 MHz LPDA

Material Description	*Quantity*
1) Aluminum tubing—0.047″ wall thickness	
1″—12′ or 6′ lengths	126 lineal feet
⅞″—12′ lengths	96 lineal feet
⅞″—6′ or 12′ lengths	66 lineal feet
¾″—8′ lengths	16 lineal feet
2) Stainless-steel hose clamps—2″ max	48 ea
3) Stainless-steel hose clamps—1¼″ max	26 ea
4) TV type U bolts	14 ea
5) U bolts, galv. type	
5/16″ × 1½″	4 ea
¼″ × 1″	2 ea
6) 1″ ID polyethylene water-service pipe 160 lb/in.^{2} test, approx. 1¼″ OD	20 lineal feet
A) 1¼″ × 1¼ × 1/8″ aluminum Angle—6' lengths	30 lineal feet
B) 1″ × ¼″ aluminum bar—6′ lengths	12 lineal feet
7) 1¼″ top rail of chain-link fence	26 lineal feet
8) 1:1 toroid balun	1 ea
9) 6-32 × 1″ stainless steel screws	24 ea
6-32 stainless steel nuts	48 ea
#6 solder lugs	24 ea
10) #12 copper feeder wire	60 lineal feet
11A) 12″ × 8″ × ¼″ aluminum plate	1 ea
B) 6″ × 4″ × ¼″ aluminum plate	1 ea
12A) ¾″ galv. Pipe	3 lineal feet
B) 1″ galv. pipe-mast	5 lineal feet
13) Galv. guy wire	50 lineal feet
14) ¼″ × 2″ turnbuckles	4 ea
15) ¼″ × 1½″ eye bolts	2 ea
16) TV guy clamps and eye bolts	2 ea

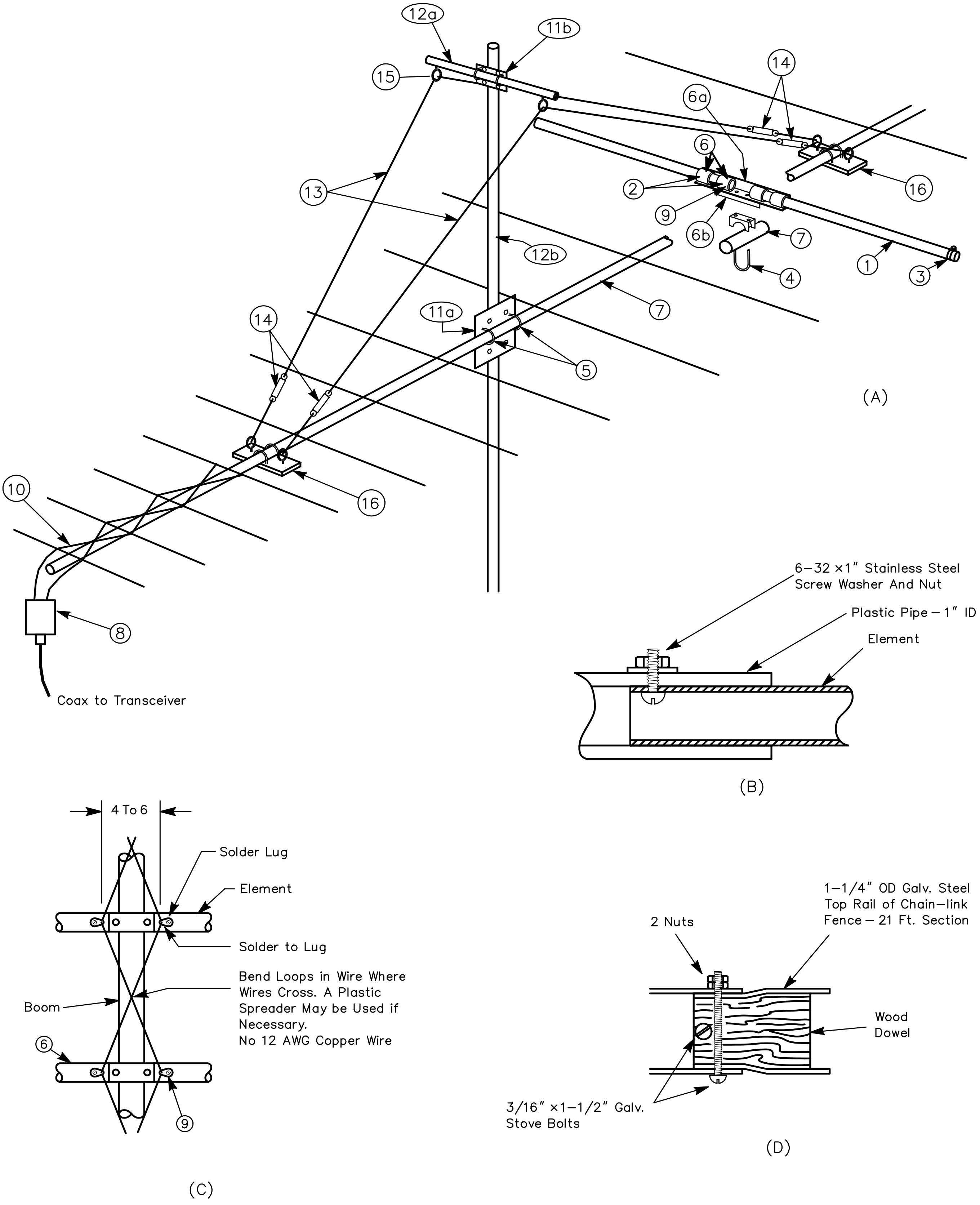
12a
11b
15
14
6a
6
2
9
6b
7
4
1
3
16
13
12b
11a
7
5
14
16
10
8
Coax to Transceiver
(A)
6–32 ×1" Stainless Steel
Screw Washer And Nut
Plastic Pipe – 1" ID
Element
(B)
4 To 6
Solder Lug
Element
Solder to Lug
Bend Loops in Wire Where
Wires Cross. A Plastic
Spreader May be Used if
Necessary.
No 12 AWG Copper Wire
Boom
6
9
(C)
2 Nuts
1–1/4" OD Galv. Steel
Top Rail of Chain–link
Fence – 21 Ft. Section
Wood
Dowel
3/16" ×1–1/2" Galv.
Stove Bolts
(D)

Fig 21—Construction diagrams of the 13-30 MHz LPDA. B and C show the method of making electrical connection between the phase-line and each half-element. D shows how the boom sections are joined.

Table 5
Element Material Requirements: 13-30 MHz LPDA

Ele.	***1″***		***$^{7}/_{8}$″***		***$^{3}/_{4}$″***		***$1^{1}/_{4}$″***	***1″***
No.	***Tubing***		***Tubing***		***Tubing***		***Angle***	***Bar***
	Len.	*Qty*	*Len.*	*Qty*	*Len.*	*Qty*	*Len.*	*Len.*
1	6′	2	6′	2	8′	2	3′	1′
2	6′	2	12′	2	—	—	3′	1′
3	6′	2	12′	2	—	—	3′	1′
4	6′	2	8.5′	2	—	—	3′	1′
5	6′	2	7′	2	—	—	3′	1′
6	6′	2	6′	2	—	—	3′	1′
7	6′	2	5′	2	—	—	2′	1′
8	6′	2	3.5′	2	—	—	2′	1′
9	6′	2	2.5′	2	—	—	2′	1′
10	3′	2	5′	2	—	—	2′	1′
11	3′	2	4′	2	—	—	2′	1′
12	3′	2	4′	2	—	—	2′	1′

of element lengths and spacings appears in **Table 3**.)

6) Total weight, 116 pounds
7) Wind-load area, 10.7 square feet
8) Required input impedance (mean resistance), R_0 = 72 Ω, Z_t = 6-inch jumper #18 wire
9) Average characteristic dipole impedance, Z_{AV}: 337.8 Ω
10) Impedance of the feeder, Z_0: 117.1 Ω
11) Feeder: #12 wire, close spaced
12) With a 1:1 toroid balun at the input terminals and a 72-Ω coax feed line, the maximum SWR is 1.4:1.

The mechanical assembly uses materials readily available from most local hardware stores or aluminum supply houses. The materials needed are given in **Table 4**. In the construction diagram, **Fig 21**, the materials are referenced by their respective material list numbers. The photograph shows the overall construction, and the drawings show the details. **Table 5** gives the required tubing lengths to construct the elements.

Experimenters may wish to improve the performance of the array at both the upper and lower frequency ends of the design spectrum so that it more closely approaches the performance in the middle of the design frequency range. The most apt general technique for raising both the gain and the front-to-back ratio at the frequency extremes would be to circularize τ as described earlier in this chapter. However, other techniques may also be applied.

The Telerana

The Telerana (Spanish for *spider web*) is a rotatable log periodic antenna that is lightweight, easy to construct and relatively inexpensive to build. Designed to cover 12.1 to 30 MHz, it was co-designed by George Smith, W4AEO, and Ansyl Eckols, YV5DLT, and first described by Eckols in *QST* for Jul 1981. Some of the design parameters are as follow.

1) τ = 0.9
2) σ = 0.05
3) Gain = 4.5 to 5.5 dBi (free-space) depending upon frequency
4) Feed arrangement: 400-Ω feeder line with 4:1 balun, fed with 52-Ω coax. The SWR is 1.5:1 or less in all amateur bands.

The array consists of 13 dipole elements, properly spaced and transposed, along an open wire feeder having an impedance of approximately 400 Ω. See **Figs 22** and **23**. The array is fed at the forward (smallest) end with a 4:1 balun and RG-8 cable placed inside the front arm and leading to the transmitter. An alternative feed method is to use open wire or ordinary TV ribbon and a tuner, eliminating the balun.

The frame that supports the array (**Fig 24**) consists of four 15-foot fiberglass vaulting poles slipped over short nipples at the hub, appearing like wheel spokes (**Fig 25**). Instead of being mounted directly into the fiberglass, the hub mounts into short metal tubing sleeves that are inserted into the ends of each arm to prevent crushing and splitting the fiberglass. The necessary holes are drilled to receive the wires and nylon.

A shopping list is provided in **Table 6**. The center hub is made from a $1^{1}/_{4}$-inch galvanized four-outlet cross or X and four 8-inch nipples (Fig 25). A 1-inch diameter X may be used alternatively, depending on the diameter of the fiberglass. A hole is drilled in the bottom of the hub to allow the cable to be passed through after welding the hub to the rotator mounting stub.

All four arms of the array must be 15 feet long. They should be strong and springy to maintain the tautness of the array. If vaulting poles are used, try to obtain all of them with identical strength ratings.

The forward spreader should be approximately 14.8 feet long. It can be much lighter than the four main arms, but must be strong enough to keep the lines rigid.

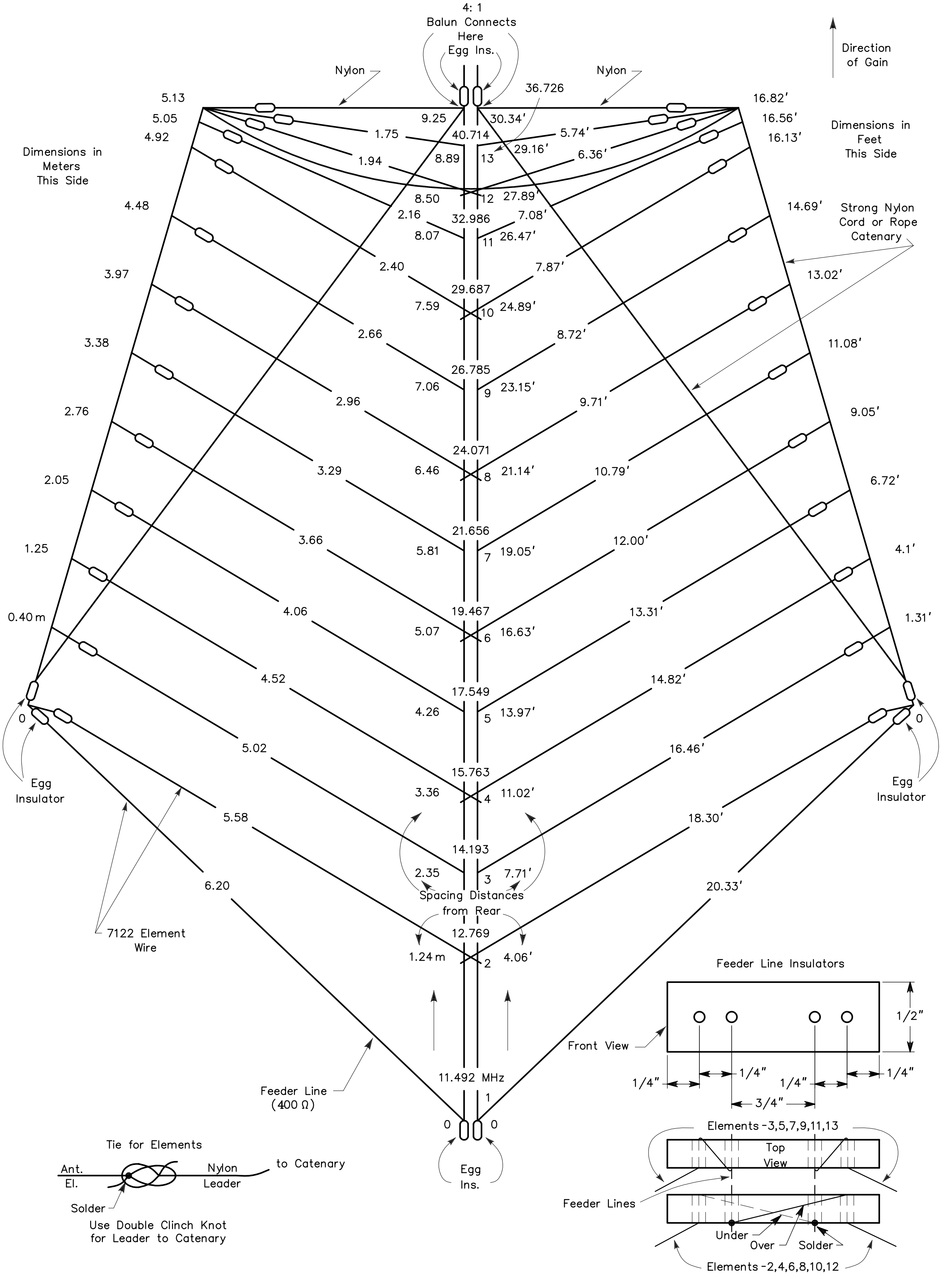
4: 1
Balun Connects
Here
Egg Ins.
Direction
of Gain
Nylon
Nylon
36.726
5.13
16.82′
5.05
9.25
30.34′
16.56′
4.92
1.75
40.714
5.74′
16.13′
Dimensions in
Feet
This Side
Dimensions in
Meters
This Side
1.94
8.89
13
29.16′
6.36′
8.50
12
27.89′
4.48
14.69′
Strong Nylon
Cord or Rope
Catenary
2.16
32.986
7.08′
8.07
11
26.47′
3.97
2.40
7.87′
13.02′
29.687
7.59
10
24.89′
3.38
2.66
8.72′
11.08′
26.785
7.06
9
23.15′
2.76
2.96
9.71′
9.05′
24.071
6.46
8
21.14′
2.05
3.29
10.79′
6.72′
21.656
1.25
3.66
5.81
7
19.05′
12.00′
4.1′
19.467
0.40 m
4.06
13.31′
1.31′
5.07
6
16.63′
4.52
17.549
14.82′
4.26
5
13.97′
0
0
5.02
16.46′
15.763
Egg
Insulator
Egg
Insulator
3.36
4
11.02′
5.58
18.30′
14.193
2.35
3
7.71′
6.20
20.33′
Spacing Distances
from Rear
12.769
7122 Element
Wire
1.24 m
2
4.06′
Feeder Line Insulators
1/2″
Front View
1/4″
1/4″
1/4″
1/4″
3/4″
11.492 MHz
1
Feeder Line
(400 Ω)
0
0
Elements -3,5,7,9,11,13
Top
View
Tie for Elements
Ant.
El.
Nylon
Leader
to Catenary
Egg
Ins.
Feeder Lines
Solder
Use Double Clinch Knot
for Leader to Catenary
Under
Over
Solder
Elements -2,4,6,8,10,12

Fig 22—The overall configuration of the spider web antenna. Nylon monofilament line is used from the ends of the elements to the nylon cords. Use nylon line to tie every point where lines cross. The forward fiberglass feeder lies on the feeder line and is tied to it. Both metric and English measurements are shown, except for the illustration of the feed-line insulator. Use soft-drawn copper or stranded wire for elements 2 through 12. Element 1 should use 7/22 flexible wire or #14 AWG Copperweld.

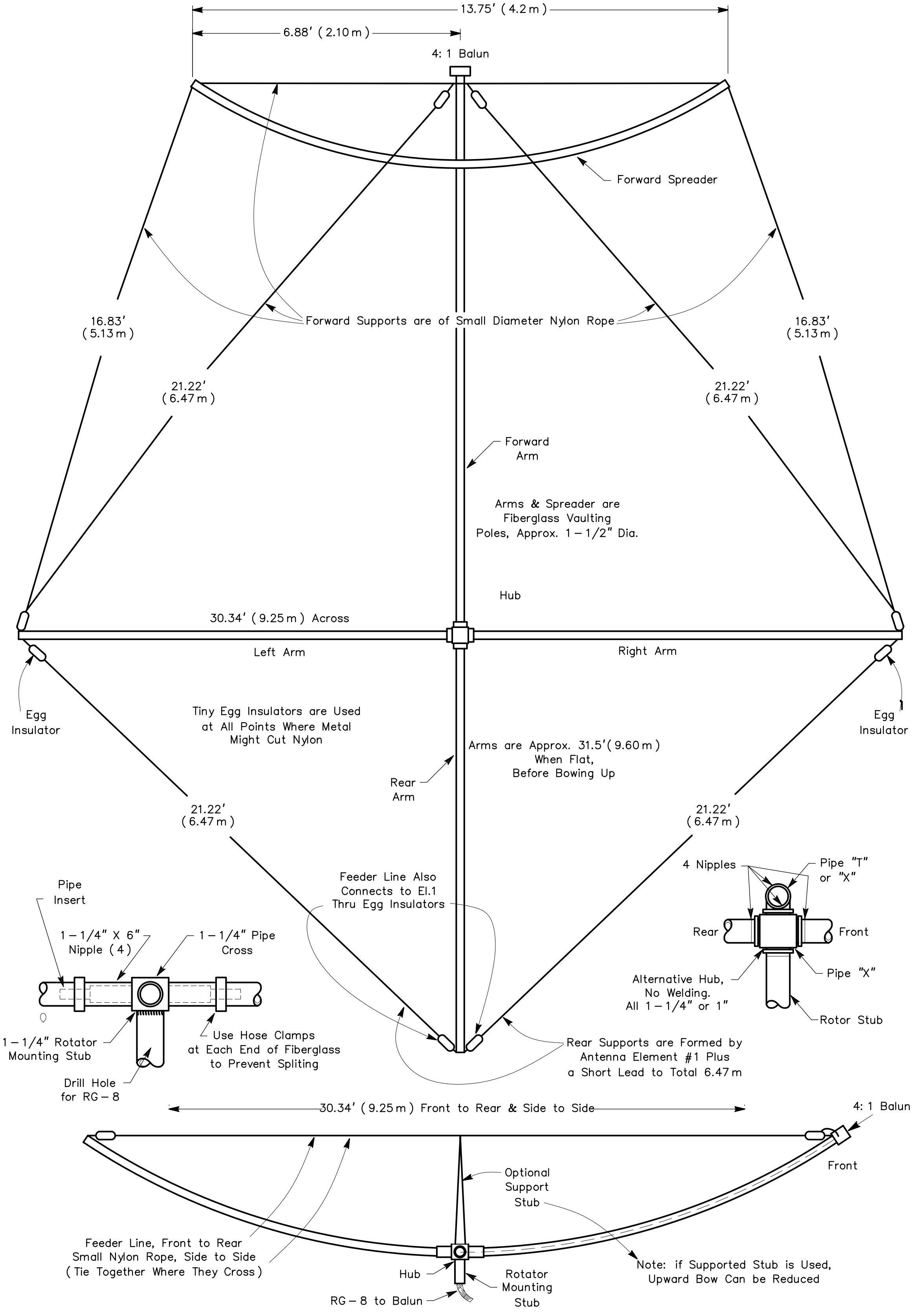
13.75′ (4.2 m)
6.88′ (2.10 m)
4: 1 Balun
Forward Spreader
Forward Supports are of Small Diameter Nylon Rope
16.83′ (5.13 m)
16.83′ (5.13 m)
21.22′ (6.47 m)
21.22′ (6.47 m)
Forward Arm
Arms & Spreader are Fiberglass Vaulting Poles, Approx. 1 − 1/2″ Dia.
Hub
30.34′ (9.25 m) Across
Left Arm
Right Arm
Egg Insulator
Egg Insulator
Tiny Egg Insulators are Used at All Points Where Metal Might Cut Nylon
Arms are Approx. 31.5′(9.60 m) When Flat, Before Bowing Up
Rear Arm
21.22′ (6.47 m)
21.22′ (6.47 m)
Feeder Line Also Connects to El.1 Thru Egg Insulators
4 Nipples
Pipe "T" or "X"
Rear
Front
Alternative Hub, No Welding. All 1 − 1/4″ or 1″
Pipe "X"
Rotor Stub
Pipe Insert
1 − 1/4″ X 6″ Nipple (4)
1 − 1/4″ Pipe Cross
1 − 1/4″ Rotator Mounting Stub
Use Hose Clamps at Each End of Fiberglass to Prevent Spliting
Drill Hole for RG − 8
Rear Supports are Formed by Antenna Element #1 Plus a Short Lead to Total 6.47 m
30.34′ (9.25 m) Front to Rear & Side to Side
4: 1 Balun
Front
Optional Support Stub
Feeder Line, Front to Rear Small Nylon Rope, Side to Side (Tie Together Where They Cross)
Hub
Rotator Mounting Stub
RG − 8 to Balun
Note: if Supported Stub is Used, Upward Bow Can be Reduced

Fig 23—The frame construction of the spider web antenna. Two different hub arrangements are illustrated.

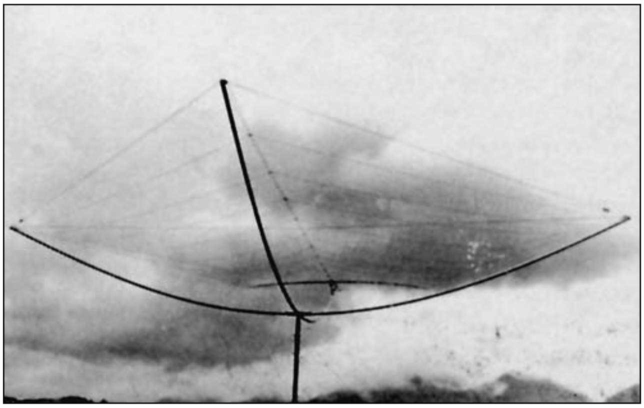

Fig 24—Although the spider web antenna resembles a rotatable clothes line, it is much larger, as indicated by Figs 22 and 23. However, the antenna can be lifted by hand.

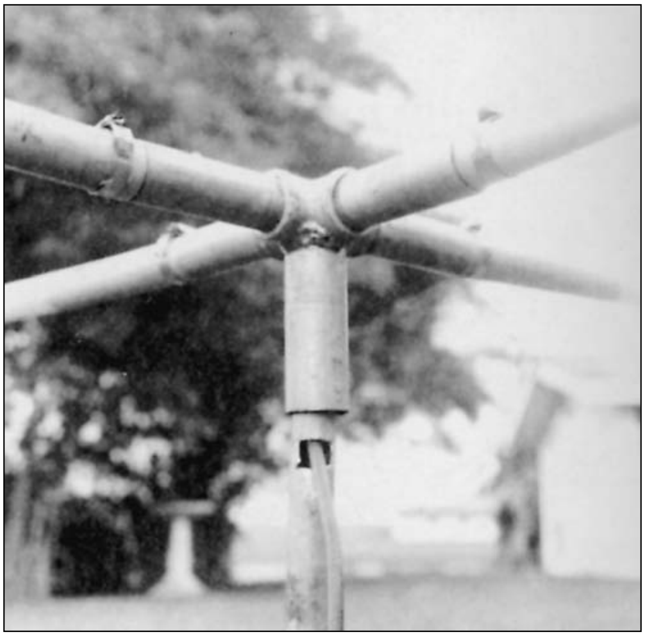

Fig 25—The simple arrangement of the spider web antenna hub. See Fig 23 and the text for details.

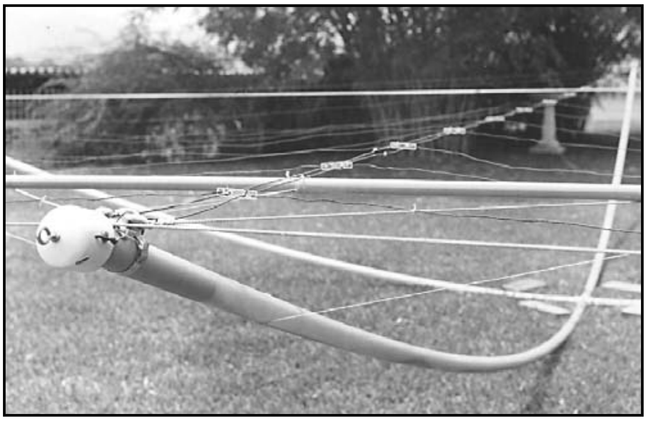

Fig 26—The elements, balun, transmission line and main bow of the spider web antenna.

Table 6
Shopping List for the Telerana

1—$1\frac{1}{4}$-inch galvanized, 4-outlet cross or X.
4—8-inch nipples.
4—15-foot long arms. Vaulting poles suggested. These must be strong and all of the same strength (150 lb) or better.
1—Spreader, 14.8 foot long (must not be metal).
1—4:1 balun unless open-wire or TV cable is used.
12—Feed-line insulators made from Plexiglas or fiberglass.
36—Small egg insulators.
328 feet copper wire for elements; flexible 7/22 is suggested.
65.6 feet (20 m) #14 Copperweld wire for interelement feed line.
164 feet (50 m) strong $\frac{1}{8}$-inch dia cord.
1—Roll of nylon monofilament fishing line, 50 lb test or better.
4—Metal tubing inserts go into the ends of the fiberglass arms.
2—Fiberglass fishing-rod blanks.
4—Hose clamps.

If tapered, the spreader should have the same measurements from the center to each end. *Do not use metal for this spreader*.

Building the frame for the array is the first construction step. Once the frame is prepared, then everything else can be built onto it. Begin by assembling the hub and the four arms, letting them lie flat on the ground with the rotator stub inserted in a hole in the ground. The tip-to-tip length should be about 31.5 feet each way. A hose clamp is used at each end of the arms to prevent splitting. Place the metal inserts in the outer ends of the arms, with 1 inch protruding. The mounting holes should have been drilled at this point. If the egg insulators and nylon cords are mounted to these tube inserts, the whole antenna can be disassembled simply by bending up the arms and pulling out the inserts with everything still attached.

Choose the arm to be at the front end. Mount two egg insulators at the front and rear to accommodate the inter-element feeder. These insulators should be as close as possible to the ends.

At each end of the cross-arm on top, install a small pulley and string nylon cord across and back. Tighten the cord until the upward bow reaches 3 feet above the hub. All cords will require retightening after the first few days because of stretching. The cross-arm can be laid on its side while preparing the feeder line. For the front-to-rear bowstring it is important to use a wire that will not stretch, such as #14 Copperweld. This bowstring is actually the inter-element transmission line. See **Fig 26**.

Secure the rear ends of the feeder to the two rear insulators, soldering the wrap. Before securing the fronts, slip the 12 insulators onto the two feed lines. A rope can be used temporarily to form the bow and to aid in mounting the feeder line. The end-to-end length of the feeder

Improving the Telerana

In *The ARRL Antenna Compendium, Vol 4*, Markus Hansen, VE7CA, described how he modified the Telerana to improve the front-to-back ratio on 20 and 15 meters. In addition, he added a trapped 30/40-meter dipole that functions as a top truss system to stabilize the modified Telerana in strong uprising winds that otherwise could turn the antenna into an “inside-out umbrella.”

Fig A shows the layout for the modified Telerana, and Table A lists the lengths and spacings for the #14 wire elements. Note that VE7CA used tuning stubs to tweak the 15 and 20-meter reflector wires for best rearward pattern. The construction techniques used by VE7CA are the same as for the original Telerana. Fig B shows a side view of the additional 40/30-meter-dipole truss sytem.

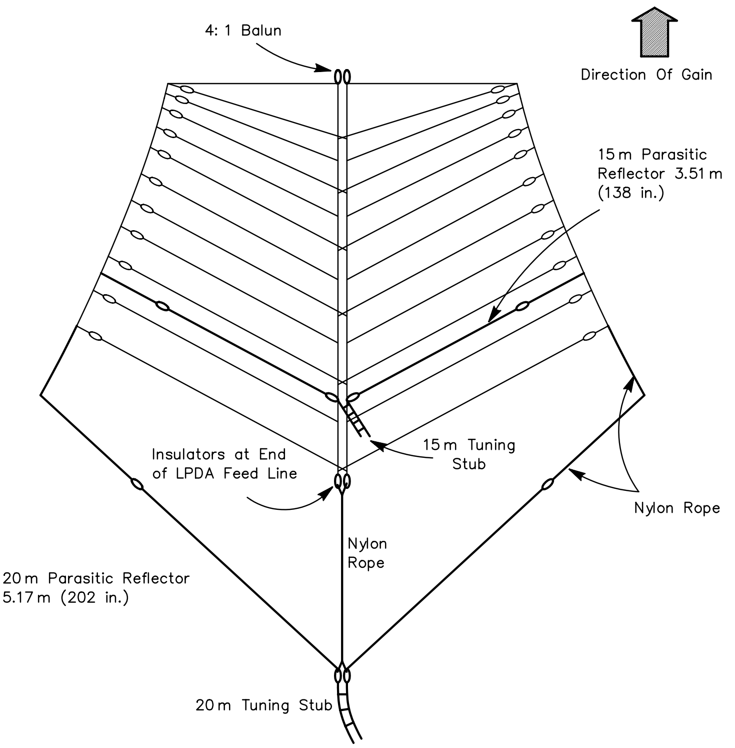
4:1 Balun
Direction Of Gain
15 m Parasitic Reflector 3.51 m (138 in.)
Insulators at End of LPDA Feed Line
15 m Tuning Stub
Nylon Rope
Nylon Rope
20 m Parasitic Reflector 5.17 m (202 in.)
20 m Tuning Stub

Table A

Element Lengths and Spacings, in Inches

Element Number	*1/2 Element Length (inches)*	*Total Distance (inches)*
R1	202.0	0.0
L1	210.1	102.0
L2	191.2	140.7
R2	138.0	158.3
L3	174.0	175.9
L4	158.3	207.9
L5	144.1	237.0
L6	131.1	261.5
L7	119.3	285.6
L8	108.6	307.6
L9	98.8	327.6
L10	89.9	345.8
L11	82.4	364.5

Note: the reflector lengths do not include the length of the tuning stubs.

Fig A—Physical layout of modified Telerana with 20 and 15-meter reflectors added (in place of first two elements in original Telerana). Note the tuning stubs for the added reflectors.

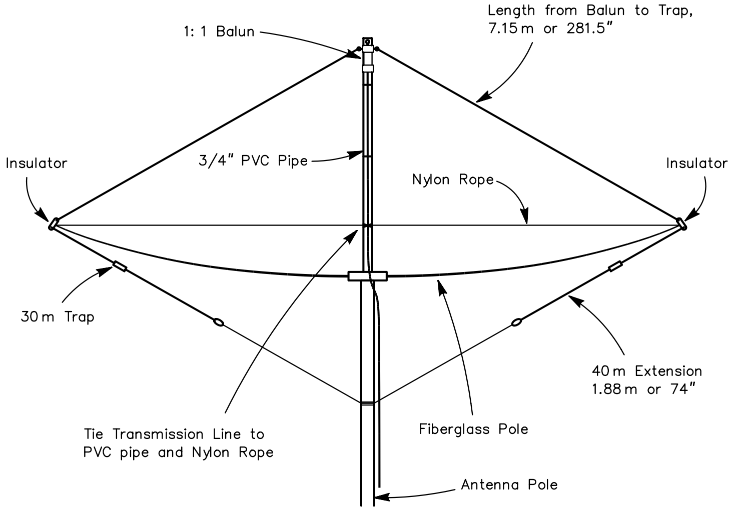
1:1 Balun
Length from Balun to Trap, 7.15 m or 281.5″
Insulator
3/4″ PVC Pipe
Nylon Rope
Insulator
30 m Trap
40 m Extension 1.88 m or 74″
Tie Transmission Line to PVC pipe and Nylon Rope
Fiberglass Pole
Antenna Pole

Fig B—Side view of 30/40-meter addition to the modified Telerana, using 3/4-inch PVC pipe as a vertical stabilizer and support for the 30/40-meter trapped dipole.

should be 30.24 feet.

Now lift both bows to their upright position and tie the feeder line and the cross-arm bowstring together where they cross, directly over and approximately 3 feet above the hub.

The next step is to install the number 1 rear element from the rear egg insulators to the right and left cross-arms using other egg insulators to provide the proper element length. Be sure to solder the element halves to the transmission line. Complete this portion of the construction by installing the nylon cord catenaries from the front arm to the cross-arm tips. Use egg insulators where needed to prevent cutting the nylon cords.

When preparing the fiberglass forward spreader, keep in mind that it should be 14.75 feet long before bowing and is approximately 13.75 feet across when bowed. Secure the center of the bowstring to the end of the front arm. Lay the spreader on top of the feed line, then tie the feeder to the spreader with nylon fish line. String the catenary from the spreader tips to the cross-arm tips.

At this point of assembly, prepare antenna elements 2 through 13. There will be two segments for each element. At the outer tip make a small loop and solder the wrap. The loop will be for the nylon leader. Measure the length plus 0.4 inch for wrapping and soldering the element segment to the feeder. Seven-strand #22 antenna wire is suggested for the element wires. Slide the feedline insulators to their proper position and secure them temporarily.

The drawings show the necessary transposition scheme. Each element half of elements 1, 3, 5, 7, 9, 11 and 13 is connected to its own side of the feeder, while elements 2, 4, 6, 8, 10 and 12 cross over to the opposite side of the transmission line.

There are four holes in each of the transmission-line insulators (see Fig 22). The inner holes are for the transmission line, and the outer ones are for the elements. Since the array elements are slanted forward, they should pass through the insulator from front to back, then back over the insulator to the front side and be soldered to the transmission line. The small drawings of Fig 22 show the details of the element transpositions.

Everywhere that lines cross, tie them together with nylon line, including all copper-nylon and nylon-nylon junctions. Careful tying makes the array much more rigid. However, all elements should be mounted loosely before you try to align the whole thing. Tightening any line or element affects all the others. There will be plenty of walking back and forth before the array is aligned properly. Expect the array to be firm but not extremely taut.

The Pounder: A Single-Band 144-MHz LPDA

The 4-element Pounder LPDA pictured in **Fig 27** was developed by Jerry Hall, K1TD, for the 144-148 MHz band. Because it started as an experimental antenna, it utilizes some unusual construction techniques. However, it gives a very good account of itself, exhibiting a theoretical free-space gain of about 7.2 dBi and a front-to-back ratio of 20 dB or better. The Pounder is small and light. It weighs just 1 pound, and hence its name. In addition, as may be seen in **Fig 28**, it can be disassembled and reassembled quickly, making it an excellent antenna for portable use. This array also serves well as a fixed-station antenna, and may be changed easily to either vertical or horizontal polarization.

The antenna feeder consists of two lengths of $^1/_2$ × $^1/_2$ × $^1/_{16}$-inch angle aluminum. The use of two facing flat surfaces permits the builder to obtain a lower characteristic impedance than can be obtained from round conductors with the same spacing. The feeder also serves as the boom for the Pounder. In the first experimental model, the array contained only two elements with a spacing of 1 foot, so a boom length of 1 foot was the primary design requirement for the 4-element version. **Table 7** gives the calculated design data for the 4-element array.

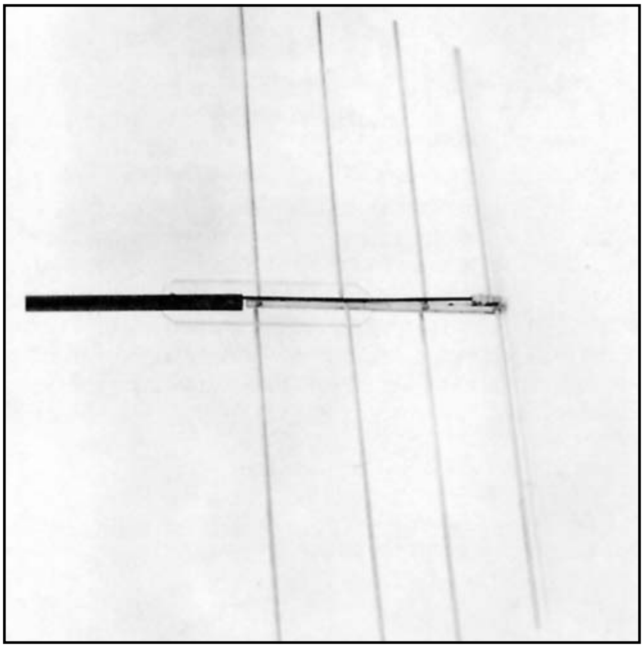

Fig 27—The 144 MHz Pounder. The boom extension at the left of the photo is a 40-inch length of slotted PVC tubing, $^7/_8$-inch outer diameter. The tubing may be clamped to the side of a tower or attached to a mast with a small boom-to-mast plate. Rotating the tubing at the clamp will provide for either vertical or horizontal polarization.

Construction

You can see the general construction approach for

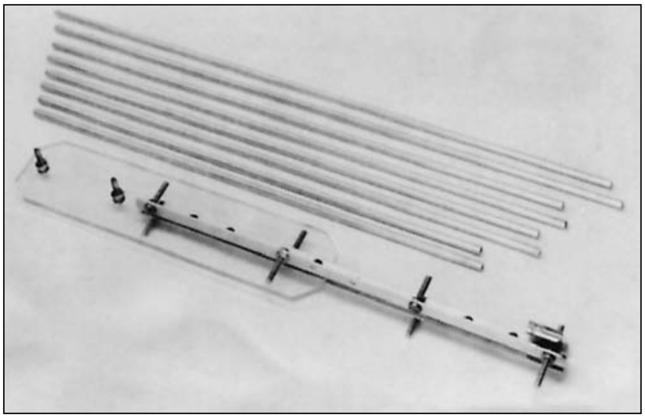

Fig 28—One end of each half element is tapped to fasten onto boom-mounted screws. Disassembly of the array consists of merely unscrewing 8 half elements from the boom. The entire disassembled array creates a small bundle only 21 inches long.

Table 7

Design Parameters for the 144-MHz Pounder

$f1 = 143$ MHz
$f_n = 148$ MHz
$B = 1.0350$
$\tau = 0.92$
$\sigma = 0.053$
Gain = 7.2 dBi = 5.1 dBd
$\cot \alpha = 2.6500$
$B_{ar} = 1.2306$
$B_s = 1.2736$
L = 0.98 feet
N = 3.90 elements (increase to 4)
Zt = none
$R_0 = 52\ \Omega$
$Z_{AV} = 312.8\ \Omega$
$\sigma' = 0.05526$
$Z_0 = 75.1\ \Omega$

Element lengths:
$\ell 1 = 3.441$ feet
$\ell 2 = 3.165$ feet
$\ell 3 = 2.912$ feet
$\ell 4 = 2.679$ feet
Element spacings:
$d_{12} = 0.365$ feet
$d_{23} = 0.336$ feet
$d_{34} = 0.309$ feet
Element diameters:
All = 0.25 inches
ℓ/diameter ratios:
$\ell 4/\text{diam}_4 = 128.6$
$\ell 3/\text{diam}_3 = 139.8$
$\ell 2/\text{diam}_2 = 151.9$
$\ell 1/\text{diam}_1 = 165.1$

Antenna feeder: $^1/_2 \times {}^1/_2 \times {}^1/_{16}''$ angle aluminum spaced $^1/_4''$
Balun: 1:1 (see text)
Feed line: 52-Ω coax (see text)

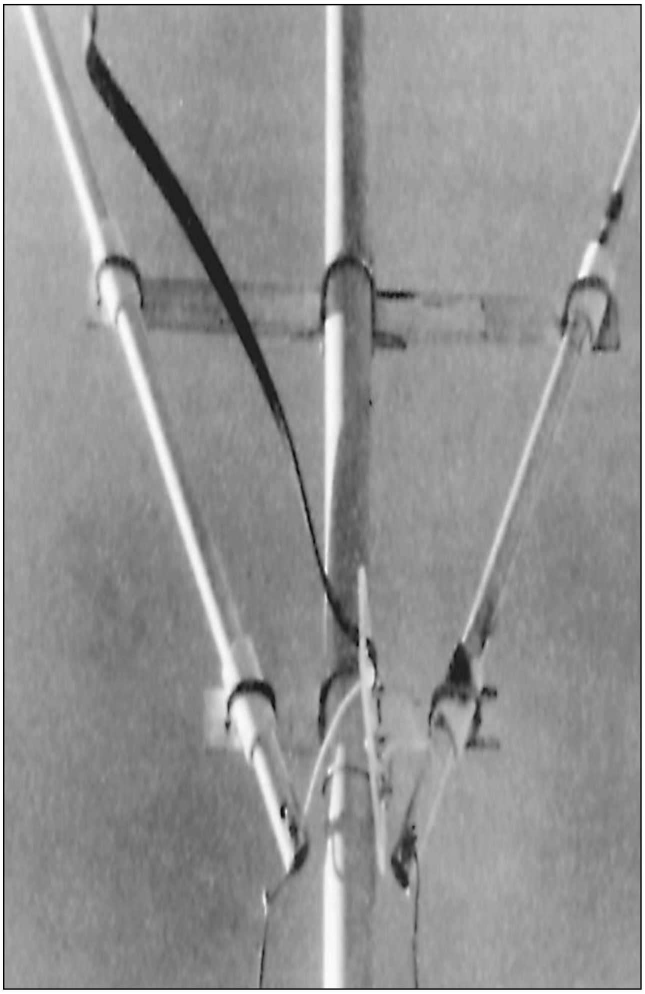

Fig 29—A close-up view of the boom, showing an alternative mounting scheme. This photo shows an earlier 2-element array, but the boom construction is the same for 2 or 4 elements. See the text for details.

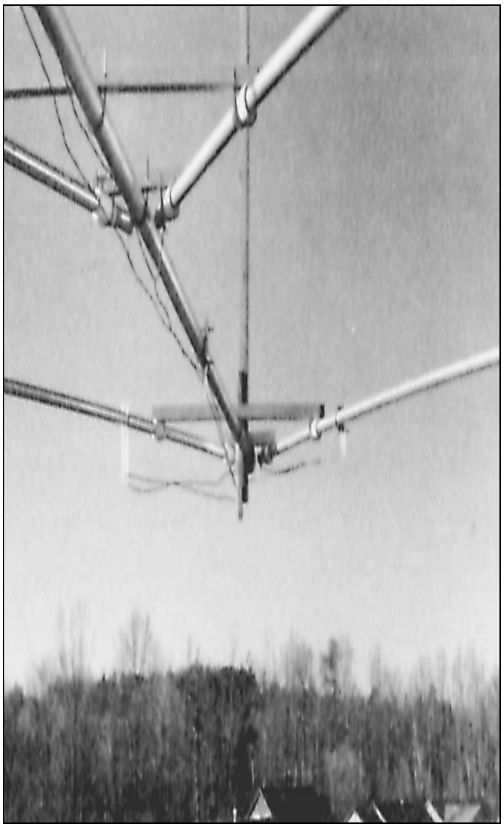

Fig 30—The feed arrangement, using a right-angle chassis-mounted BNC connector, modified by removing a portion of the flange. A short length of bus wire connects the center pin to the opposite feeder conductor.

the Pounder in the photographs. Drilled and tapped pieces of Plexiglas sheet, 1/4-inch thick, serve as insulating spacers for the angle aluminum feeder. Two spacers are used, one near the front and one near the rear of the array. Four #6-32 × 1/4-inch pan head screws secure each aluminum angle section to the Plexiglas spacers, as shown in **Figs 29** and **30**. Use flat washers with each screw to prevent it from touching the angle stock on the opposite side of the spacer. Be sure the screws are not so long as to short out the feeder! A clearance of about 1/16 inch is sufficient. If you have doubts about the screw lengths, check the assembled boom for a short with your ohmmeter on a Megohm range.

Either of two mounting techniques may be used for the Pounder. As shown in Figs 27 and 28, the rear spacer measures 10 × 2 1/2 inches, with 45º corners to avoid sharp points. This spacer also accommodates a boom extension of PVC tubing, which is attached with two #10-32 × 1-inch screws. This tubing provides for side mounting the Pounder away from a mast or tower.

An alternative support arrangement is shown in Fig 29. Two 1/2 × 3-inch Plexiglas spacers are used at the front and rear of the array. Each spacer has four holes drilled 5/8 inch apart and tapped with #6-32 threads. Two screws enter each spacer from either side to make a tight aluminum-Plexiglas-aluminum sandwich. At the center of the boom, secured with only two screws, is a 2 × 18-inch strip of 1/4-inch Plexiglas. This strip is slotted about 2 inches from each end to accept hose clamps for mounting the Pounder atop a mast. As shown, the strip is attached for vertical polarization. Alternate mounting holes, visible on the now-horizontal lip of the angle stock, provide for horizontal polarization. Although sufficient, this mounting arrangement is not as sturdy as that shown in Fig 27.

The elements are lengths of thick-wall aluminum tubing, 1/4-inch OD. The inside wall conveniently accepts a #10-32 tap. The threads should penetrate the tubing to a depth of at least 1 inch. Eight #10-32 × 1-inch screws are attached to the boom at the proper element spacings and held in place with #10-32 nuts, as shown in Fig 28. For assembly, the elements are then simply screwed into place.

Note that with this construction arrangement, the two halves of any individual element are not precisely collinear; their axes are offset by about 3/4 inch. This offset does not seem to affect performance.

The Feed Arrangement

Use care initially in mounting and cutting the elements to length. To obtain the 180° crossover feed arrangement, the element halves from a single section of the feeder/boom must alternate directions. That is, for half-elements attached to one of the two pieces of angle stock, elements 1 and 3 will point to one side, and elements 2 and 4 to the other. This arrangement may be seen by observing the element-mounting screws in Fig 28. Because of this mounting scheme, the length of tubing for an element "half" is not simply half of the length given in Table 7. After final assembly, halves for elements 2 and 4 will have a slight overlap, while elements 1 and 3 are extended somewhat by the boom thickness. The best procedure is to cut each assembled element to its final length by measuring from tip to tip.

The Pounder may be fed with RG-58 or RG-59 coax and a BNC connector. A modified right-angle chassis-mount BNC connector is attached to one side of the feeder/boom assembly for cable connection, Fig 30. The modification consists of cutting away part of the mounting flange that would otherwise protrude from the boom assembly. This leaves only two mounting-flange holes, but these are sufficient for a secure mount. A short length of small bus wire connects the center pin to the opposite side of the feeder, where it is secured under the mounting-screw nut for the shortest element.

For operation, you may secure the coax to the PVC boom extension or to the mast with electrical tape. You should use a balun, especially if the Pounder is operated with vertical elements. A choke type of balun is satisfactory, formed by taping 6 turns of the coax into a coil of 3 inches diameter, but a bead balun is preferred (see Chapter 26). The balun should be placed at the point where the coax is brought away from the boom. If the mounting arrangement of Fig 29 is used with vertical polarization, a second balun should be located approximately 1/4 wavelength down the coax line from the first. This will place it at about the level of the lower tips of the elements. For long runs of coax to the transmitter, a transition from RG-58 to RG-8 or from RG-59 to RG-9 is suggested, to reduce line losses. Make this transition at some convenient point near the array.

No shorting feeder termination is used with the array described here. The antenna feeder (phase-line) Z_0 of this array is in the neighborhood of 120 Ω, and with a resulting feed-point impedance of about 72 Ω. The theoretical mean SWR with 52 Ω line is 72/52 or 1.4 to 1. Upon array completion, the measured SWR (52-Ω line) was found to be relatively constant across the band, with a value of about 1.7 to 1. The Pounder offers a better match to 72-Ω coax.

Being an all-driven array, the Pounder is more immune to changes in feed-point impedance caused by nearby objects than is a parasitic array. This became obvious during portable use when the array was operated near trees and other objects… the SWR did not change noticeably with antenna rotation toward and away from those objects. Consequently, the Pounder should behave well in a restricted environment, such as an attic. Weighing just one pound, this array indeed does give a good account of itself.

Log Periodic-Yagi Arrays

Several possibilities exist for constructing high-gain arrays that use the log periodic dipole array concept. One technique is to add parasitic elements to the LPDA to increase both the gain and the front-to-back ratio for a specific frequency within the passband of the LPDA. The LPDA-Yagi combination is simple in concept. It utilizes an LPDA group of driven elements, along with parasitic elements at normal Yagi spacings from the active elements of the LPDA.

The LPDA-Yagi combinations are endless. An example of a single-band high-gain design is a 2- or 3-element LPDA for 21.0 to 21.45 MHz with the addition of two or three parasitic directors and one parasitic reflector. The name Log-Yag (log-cell Yagi) array has been coined for these hybrid antennas. The LPDA portion of the array is of the usual design to cover the desired bandwidth, and standard Yagi design procedures are used for the parasitic elements. Information in this section is based on a Dec 1976, *QST* article by P. D. Rhodes, K4EWG, and J. R. Painter, W4BBP, "The Log-Yag Array."

THE LOG-YAG ARRAY

The Log-Yag array, with its added parasitic elements, provides higher gain and greater directivity than would be realized with the LPDA alone. Yagi arrays require a long boom and wide element spacing for wide bandwidth and high gain, because the Q of the Yagi system increases as the number of elements is increased or as the spacing between adjacent elements is decreased. An increase in the Q of the Yagi array means that the total operating bandwidth of the array is decreased, and the gain and front-to-back ratio specified in the design are obtainable only over small portions of the band. [Older Yagi designs did indeed exhibit the limitations mentioned here. But modern, computer-aided design has resulted in wideband Yagis, provided that sufficient elements are used on the boom to allow stagger tuning for wide-band coverage. See Chapter 11.—*Ed.*]

The Log-Yag system overcomes this difficulty by using a multiple driven element cell designed in accordance with the principles of the log periodic dipole array. Since this log cell exhibits both gain and directivity by itself, it is a more effective wide-band radiator than a simple dipole driven element. The front-to-back ratio and gain of the log cell can then be improved with the addition of a parasitic reflector and director.

It is not necessary for the parasitic element spacings to be large with respect to wavelength, since the log cell is the determining factor in the array bandwidth. As well, the element spacings within the log cell may be small with respect to a wavelength without appreciable deterioration of the cell gain. For example, decreasing the relative spacing constant (σ) from 0.1 to 0.05 will decrease the array gain by less than 1 dB.

A Practical Example

The photographs and figures show a Log-Yag array for the 14-MHz amateur band. The array design takes the form of a 4-element log cell, a parasitic reflector spaced at 0.085 λ_{max}, and a parasitic director spaced at 0.15 λ_{max} (where λ_{max} is the longest free-space wavelength within the array passband). Array gain is almost unaffected with reflector spacings from 0.08 λ to 0.25 λ, and the increase in boom length is not justified. The function of the reflector is to improve the front-to-back ratio of the log cell, while the director sharpens the forward lobe and decreases the half-power beamwidth. As the spacing between the parasitic elements and the log cell decreases, the parasitic elements must increase in length.

The log cell is designed to meet upper and lower band limits with $\sigma = 0.05$. The design parameter τ is dependent on the structure bandwidth, B_s. When the log periodic design parameters have been found, the element length and spacings can be determined.

Array layout and construction details can be seen in Figs 31 through 34. Characteristics of the array are given in **Table 8**.

The method of feeding the antenna is identical to that of feeding the log periodic dipole array without the parasitic elements. As shown in **Fig 31**, a balanced feeder is required for each log-cell element, and all adjacent elements are fed with a 180° phase shift by alternating connections. Since the Log-Yag array will be covering a relatively small bandwidth, the radiation resistance of the narrow-band log cell will vary from 80 to 90 Ω (tubing

Table 8
Log-Yag Array Characteristics

1) Frequency range	14 to 14.35 MHz
2) Operating bandwidth	B = 1.025
3) Design parameter	$\tau = 0.946657$
4) Apex half angle	$\alpha = 14.921°$; cot $\alpha = 3.753$
5) Half-power beamwidth	42° (14 to 14.35 MHz)
6) Bandwidth of structure	Bs = 1.17875
7) Free-space wavelength	λ_{max} = 70.28 feet
8) Log cell boom length	L = 10.0 feet
9) Longest log element	$\ell 1$ = 35.14 feet (a tabulation of element lengths and spacings is given in Table 9)
10) Forward gain (free space)	8.2 dBi
11) Front-to-back ratio	32 dB (theoretical)
12) Front-to-side ratio	45 dB (theoretical)
13) Input impedance	Z_0 = 37 Ω
14) SWR	1.3 to 1 (14 to 14.35 MHz)
15) Total weight	96 pounds
16) Wind-load area	8.5 sq feet
17) Reflector length	36.4 feet at 6.0 foot spacing
18) Director length	32.2 feet at 10.5 foot spacing
19) Total boom length	26.5 feet

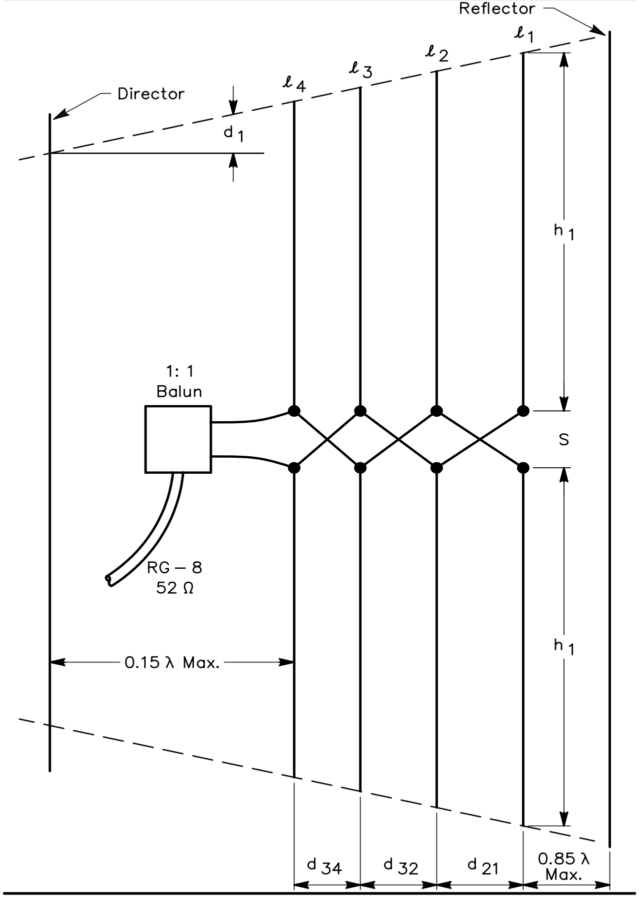
Reflector
ℓ1
ℓ2
ℓ3
ℓ4
Director
d1
h1
1: 1
Balun
S
RG – 8
52 Ω
h1
0.15 λ Max.
d34
d32
d21
0.85 λ
Max.

Fig 31—Layout of the Log-Yag array.

Table 9

Log-Yag Array Dimensions

Element	*Length Feet*	*Spacing Feet*
Reflector	36.40	6.00 (Ref. to ℓ1)
ℓ1	35.14	3.51 (d_{12})
ℓ2	33.27	3.32 (d_{23})
ℓ3	31.49	3.14 (d_{34})
ℓ4	29.81	10.57 (ℓ4 to dir.)
Director	32.20	

elements) depending on the operating bandwidth. The addition of parasitic elements lowers the log-cell radiation resistance. Hence, it is recommended that a 1:1 balun be connected at the log-cell input terminals and 50-Ω coaxial cable be used for the feed line.

The measured radiation resistance of the 14-MHz Log-Yag is 37 Ω over the frequency range from 14.0 to 14.35 MHz. It is assumed that tubing elements will be used. However, if a wire array is used, then the radiation resistance R_0 and antenna-feeder input impedance Z_0 must be calculated so that the proper balun and coax may be used. The procedure is outlined in detail in an earlier part of this chapter. However, programs such as *LPCAD28* are also suitable to automate the calculations.

Table 9 has array dimensions. **Tables 10** and **11** contain lists of the materials necessary to build the Log-Yag array.

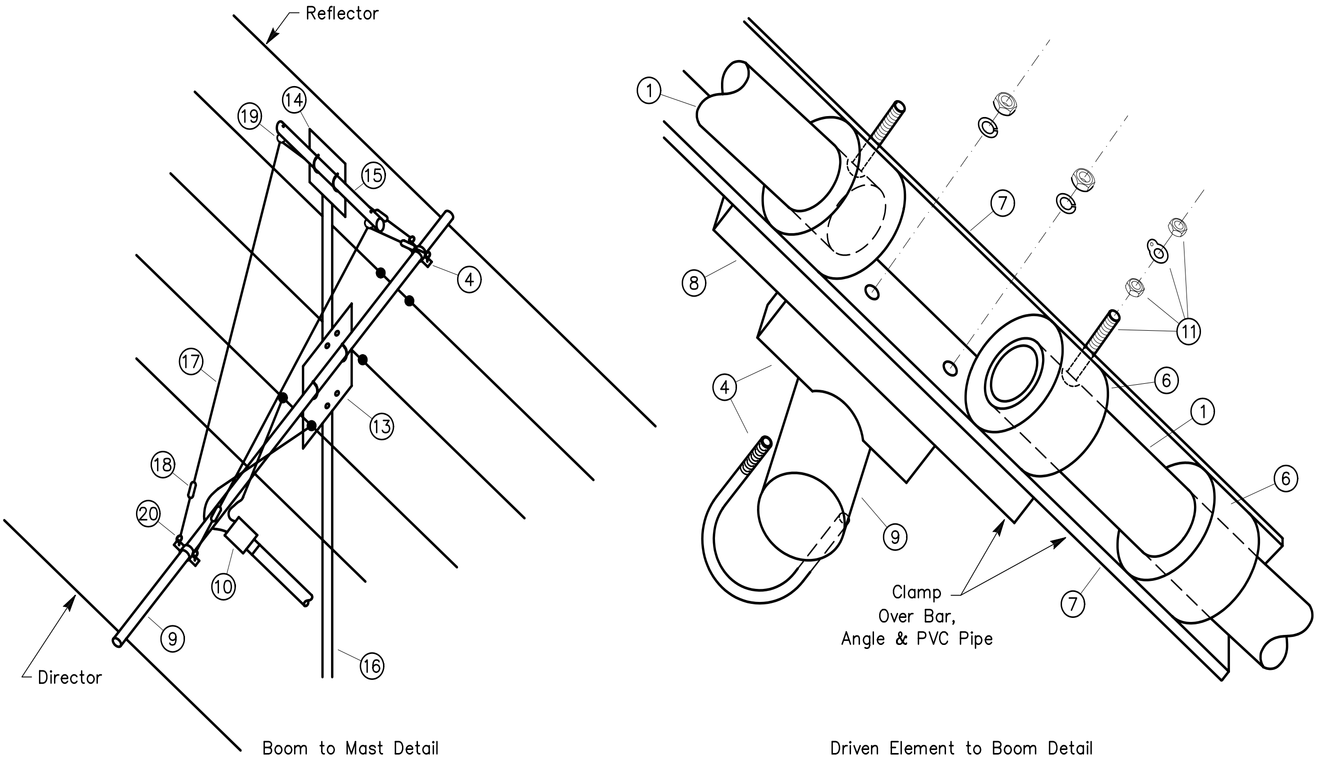
Reflector
⑭
⑲
⑮
④
⑰
⑬
⑱
⑳
⑩
⑨
⑯
Director
Boom to Mast Detail
①
⑦
⑧
⑪
④
⑥
①
⑥
⑨
Clamp
Over Bar,
Angle & PVC Pipe
⑦
Driven Element to Boom Detail

Fig 32—Assembly details. The numbered components refer to Table 11.

Table 10
Element Material Requirements: Log-Yag Array

	1-in. Tubing		***⅞-in. Tubing***		***¾-in. Tubing***		***1¼-in. Angle***	***1¼-in. Bar***
	Len. Feet	*Qty*	*Len. Feet*	*Qty*	*Len. Feet*	*Qty*	*Len. Feet*	*Len. Feet*
Reflector	12	1	6	2	8	2	None	None
ℓ1	6	2	6	2	8	2	3	1
ℓ2	6	2	6	2	8	2	3	1
ℓ3	6	2	6	2	6	2	3	1
ℓ4	6	2	6	2	6	2	3	1
Director	12	1	6	2	6	2	None	None

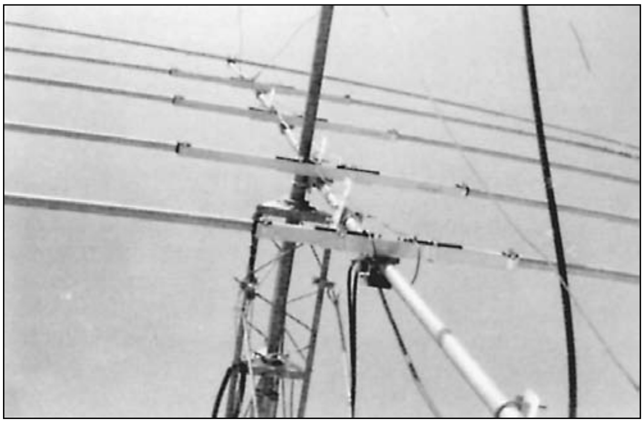

Fig 33—The attachment of the elements to the boom.

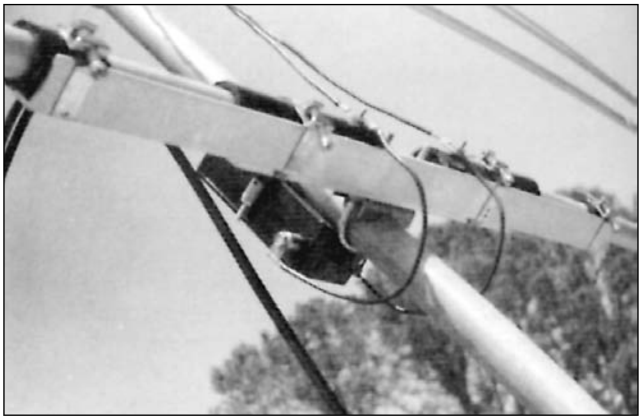

Fig 34—Looking from the front to the back of the Log-Yag array. A truss provides lateral and vertical support.

Table 11
Materials List, Log-Yag Array

1) Aluminum tubing—0.047 in. wall thickness
   1 in.—12 ft lenths, 24 lin. ft
   1 in.—12 ft or 6 ft lengths, 48 lin. ft
   ⅞ in.—12 ft or 6 ft lengths, 72 lin. ft
   ¾ in.—8 ft lengths, 48 lin. ft
   ¾ in.—6 ft lengths, 36 lin. ft
2) Stainless steel hose clamps—2 in. max, 8 ea
3) Stainless steel hose clamps—1¼ in. max, 24 ea
4) TV-type U bolts—1½ in., 6 ea
5) U bolts, galv. type: 5/16 in. × 1½ in., 6 ea
5A) U bolts, galv. type: ¼ in. × 1 in., 2 ea
6) 1 in. ID water-service polyethylene pipe 160 lb/in.^{2} test, approx. 1⅜ in. OD, 7 lin. ft
7) 1¼ in. × 1¼ in. × ⅛ in. aluminum angle—6 ft lengths, 12 lin. ft
8) 1 in. × ¼ in. × ¼ in. aluminum angle—6 ft lengths, 6 lin. ft
9) 1¼ in. top rail of chain-link fence, 26.5 lin. ft
10) 1:1 toroid balun, 1 ea
11) No. 6-32 × 1 in. stainless steel screws, 8 ea
    No. 6-32 stainless steel nuts, 16 ea
    No. 6 solder lugs, 8 ea
12) #12 copper feed wire, 22 lin. ft
13) 12 in. × 6 in. × ¼ in. aluminum plate, 1 ea
14) 6 in. × 4 in. × ¼ in. aluminum plate, 1 ea
15) ¾ in. galv. pipe, 3 lin. ft
16) 1 in. galv. pipe—mast, 5 lin. ft
17) Galv. guy wire, 50 lin. ft
18) ¼ in. × 2 in. turnbuckles, 4 ea
19) ¼ in. × 1½ in. eye bolts, 2 ea
20) TV guy clamps and eyebolts, 2 ea

BIBLIOGRAPHY

Source material and more extended discussion of the topics covered in this chapter can be found in the references listed below and in the textbooks listed at the end of Chapter 2.

D. Allen, “The Log Periodic Loop Array (LPLA) Antenna,” *The ARRL Antenna Compendium, Vol 3,* pp 115-117.

C. A. Balanis, *Antenna Theory, Analysis and Design,* 2nd Ed. (New York: John Wiley & Sons, 1997) Chapter 9.

P. C. Butson and G. T. Thompson, “A Note on the Calculation of the Gain of Log-Periodic Dipole Antennas,” *IEEE Trans on Antennas and Propagation*, Vol AP-24, No. 1, Jan 1976, pp 105-106.

R. L. Carrel, “The Design of Log-Periodic Dipole Antennas,” *1961 IRE International Convention Record*.

L. B. Cebik, "The Monoband Log-Cell Yagi Revisited," (4-part series) *National Contest Journal*, Jan-Jul 2000.

L. B. Cebik, "Some Preliminary Notes on Standard Design LPDAs for 3-30 MHz," (2-part series), *QEX*, Mar-May 2000.

R. H. DuHamel and D. E. Isbell, "Broadband Logarithmically Periodic Antenna Structures," *1957 IRE National Convention Record*, Part 1.

A. Eckols, "The Telerana: A Broadband 13- to 30-MHz Directional Antenna," *QST*, Jul 1981, pp 24-27.

J. Fisher, "Development of the W8JF Waveram: A Planar Log-Periodic Quad Array," *The ARRL Antenna Compendium, Vol 1*, pp 50-54.

M. Hansen, "The Improved Telerana, with Bonus 30/40 meter Coverage," *The ARRL Antenna Compendium, Vol 4*, pp 112-117.

K. Heitner, "A Wide-Band, Low-Z Antenna—New Thoughts on Small Antennas," *The ARRL Antenna Compendium, Vol 1*, pp 48-49.

D. E. Isbell, "Log-Periodic Dipole Arrays," *IRE Transactions on Antennas and Propagation*, Vol. AP-8, No. 3, May 1960.

J. D. Kraus, *Antennas*, 2nd Ed. (New York: McGraw-Hill, 1988), Chapter 15.

R. A. Johnson, ed., *Antenna Engineering Handbook*, 3rd Ed. (New York: McGraw-Hill, 1993), Chapters 14 and 26.

K. Luetzelschwab, "Log Periodic Dipole Array Improvements," *The ARRL Antenna Compendium, Vol 6*, pp 74-76.

D. A. Mack, "A Second-Generation Spiderweb Antenna," *The ARRL Antenna Compendium, Vol 1*, pp 55-59.

P. E. Mayes and R. L. Carrel, "Log Periodic Resonant-V Arrays," *IRE Wescon Convention Record*, Part 1, 1961.

P. E. Mayes, G. A. Deschamps, and W. T. Patton, "Backward Wave Radiation from Periodic Structures and Application to the Design of Frequency Independent Antennas," *Proc. IRE*.

J. J. Meyer, "A Simple Log-Yag Array for 50 MHz," *The ARRL Antenna Compendium, Vol 1*, pp 62-63.

C. T. Milner, "Log Periodic Antennas," *QST*, Nov 1959, pp 11-14.

W. I. Orr and S. D. Cowan, *Beam Antenna Handbook*, pp 251-253.

P. D. Rhodes, "The Log-Periodic Dipole Array," *QST*, Nov 1973, pp 16-22.

P. D. Rhodes and J. R. Painter, "The Log-Yag Array," *QST*, Dec 1976, pp 18-21.

P. D. Rhodes, "The Log-Periodic V Array," *QST*, Oct 1979, pp 40-43.

P. D. Rhodes, "The K4EWG Log Periodic Array," *The ARRL Antenna Compendium, Vol 3*, pp 118-123

V. H. Rumsey, *Frequency Independent Antennas* (New York: Academic Press, 1966).

F. Scholz, "A 14-30 MHz LPDA for Limited Space," *The ARRL Antenna Compendium, Vol 2*, pp 96-99

W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*, 2nd Ed. (New York: John Wiley & Sons, 1998), Chapter 6.

J. J. Uhl, "Construct a Wire Log-Periodic Dipole Array for 80 or 40 Meters," *QST*, Aug 1976, pp 21-24.

R. F. Zimmer, "Three Experimental Antennas for 15 Meters," *CQ*, Jan 1983, pp 44-45.

R. F. Zimmer, "Development and Construction of 'V' Beam Antennas," *CQ*, Aug 1983, pp 28-32.

Chapter 11

HF Yagi Arrays

Along with the dipole and the quarter-wave vertical, radio amateurs throughout the world make extensive use of the Yagi array. Hidetsugu Yagi and Shintaro Uda, two Japanese university professors, invented the Yagi in the 1920s. Uda did much of the developmental work, while Yagi introduced the array to the world outside Japan through his writings in English. Although the antenna should properly be called a *Yagi-Uda* array, it is commonly referred to simply as a *Yagi*.

The Yagi is a type of endfire multielement array. At the minimum, it consists of a single driven element and a single parasitic element. These elements are placed parallel to each other, on a supporting boom spacing them apart. This arrangement is known as a 2-element Yagi. The parasitic element is termed a *reflector* when it is placed behind the driven element, opposite to the direction of maximum radiation, and is called a *director* when it is placed ahead of the driven element. See **Fig 1**. In the VHF and UHF spectrum, Yagis employing 30 or more elements are not uncommon, with a single reflector and multiple directors. See Chapter 18, VHF and UHF Antenna Systems, for details on VHF and UHF Yagis. Large HF arrays may employ 10 or more elements, and will be covered in this chapter.

The gain and directional pattern of a Yagi array is determined by the relative amplitudes and phases of the currents induced into all the parasitic elements. Unlike the directly driven multielement arrays considered in Chapter 8, Multielement Arrays, where the designer must compensate for mutual coupling between elements, proper Yagi operation *relies on* mutual coupling. The current in each parasitic element is determined by its spacing from both the driven element and other parasitic elements, and by the tuning of the element itself. Both length and diameter affect element tuning.

For about 50 years amateurs and professionals created Yagi array designs largely by "cut and try" experimental techniques. In the early 1980s, Jim Lawson, W2PV, described in detail for the amateur audience the fundamental mathematics involved in modeling Yagis. His book *Yagi Antenna Design* is highly recommended for serious antenna designers. The advent of powerful microcomputers and sophisticated computer antenna modeling software in the mid 1980s revolutionized the field of Yagi design for the radio amateur. In a matter of minutes, a computer can

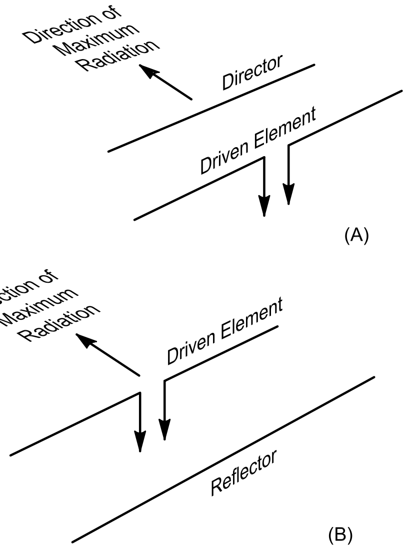
Direction of Maximum Radiation
Director
Driven Element
(A)
Direction of Maximum Radiation
Driven Element
Reflector
(B)

Fig 1—Two-element Yagi systems using a single parasitic element. At A the parasitic element acts as a director, and at B as a reflector. The arrows show the direction in which maximum radiation takes place.

try 100,000 or more different combinations of element lengths and spacings to create a Yagi design tailored to meet a particular set of high-performance parameters. To explore this number of combinations experimentally, a human experimenter would take an unimaginable amount of time and dedication, and the process would no doubt suffer from considerable measurement errors. With the computer tools available today, an antenna can be designed, constructed and then put up in the air, with little or no tuning or pruning required.

Yagi Performance Parameters

There are three main parameters used to characterize the performance of a particular Yagi—*forward gain*, *pattern* and *drive impedance/SWR*. Another important consideration is *mechanical strength*. It is very important to recognize that each of the three electrical parameters should be characterized over the frequency band of interest in order to be meaningful. Neither the gain, SWR nor the pattern measured at a single frequency gives very much insight into the overall performance of a particular Yagi.

Poor designs have been known to reverse their directionality over a frequency band, while other designs have excessively narrow SWR bandwidths, or overly "peaky" gain response. Finally, an antenna's ability to survive the wind and ice conditions expected in one's geographical location is an important consideration in any design. Much of this chapter will be devoted to describing detailed Yagi designs that are optimized for a good balance between gain, pattern and SWR over various amateur bands, and that are designed to survive strong winds and icing.

YAGI GAIN

Like any other antenna, the gain of a Yagi must be stated in comparison to some standard of reference. Designers of phased vertical arrays often state gain referenced to a single, isolated vertical element. See the section on "Phased Array Techniques" in Chapter 8, Multielement Arrays.

Many antenna designers prefer to compare gain to that of an *isotropic radiator in free space*. This is a theoretical antenna that radiates equally well in all directions, and by definition, it has a gain of 0 *dBi* (dB isotropic). Many radio amateurs, however, are comfortable using a dipole as a standard reference antenna, mainly because it is *not* a theoretical antenna.

In free space, a dipole does not radiate equally well in all directions—it has a figure-eight azimuth pattern, with deep nulls off the ends of the wire. In its favored directions, a free-space dipole has 2.15 dB gain compared to the isotropic radiator. You may see the term *dBd* in amateur literature, meaning gain referenced to a dipole in free space. Subtract 2.15 dB from gain in dBi to convert to gain in dBd.

Assume for a moment that we take a dipole out of "free space," and place it one wavelength over the ocean, whose saltwater makes an almost perfect ground. At an elevation angle of 15°, where sea water-reflected radiation adds in phase with direct radiation, the dipole has a gain of about 6 dB, compared to its gain when it was in free space, isolated from any reflections. See Chapter 3, The Effects of the Earth.

It is perfectly legitimate to say that this dipole has a gain of 6 dBd, although the term "dBd" (meaning "dB dipole") makes it sound as though the dipole somehow has gain over itself! Always remember that gain expressed in dBd (or dBi) refers to the *counterpart antenna in free space*. The gain of the dipole over saltwater in this example can be rated at either 6 dBd (over a dipole in free space), or as 8.15 dBi (over an isotropic radiator in free space). Each frame of reference is valid, as long as it is used consistently and clearly. In this chapter we will often switch between Yagis in free space and Yagis over ground. To prevent any confusion, gains will be stated in dBi.

Yagi free-space gain ranges from about 5 dBi for a small 2-element design to about 20 dBi for a 31-element long-boom UHF design. The length of the boom is the main factor determining the gain a Yagi can deliver. Gain as a function of boom length will be discussed in detail after the sections below defining antenna response patterns and SWR characteristics.

RESPONSE PATTERNS—FRONT-TO-REAR RATIO

As discussed in Chapter 2, Antenna Fundamentals, for an antenna to have gain, it must concentrate energy radiated in a particular direction, at the expense of energy radiated in other directions. Gain is thus closely related to an antenna's directivity pattern, and also to the losses in the antenna. **Fig 2** shows the *E-plane* (also called *E-field*, for electric field) and *H-plane* (also called *H-field*, for magnetic field) pattern of a 3-element Yagi in free space, compared to a dipole, and an isotropic radiator. These patterns were generated using the computer program *NEC-2*, which is highly regarded by antenna professionals for its accuracy and flexibility.

In free space there is no Earth reference to determine whether the antenna polarization is horizontal or vertical, and so its response patterns are labeled as E-field (electric) or H-field (magnetic). For a Yagi mounted over ground rather than in free space, if the

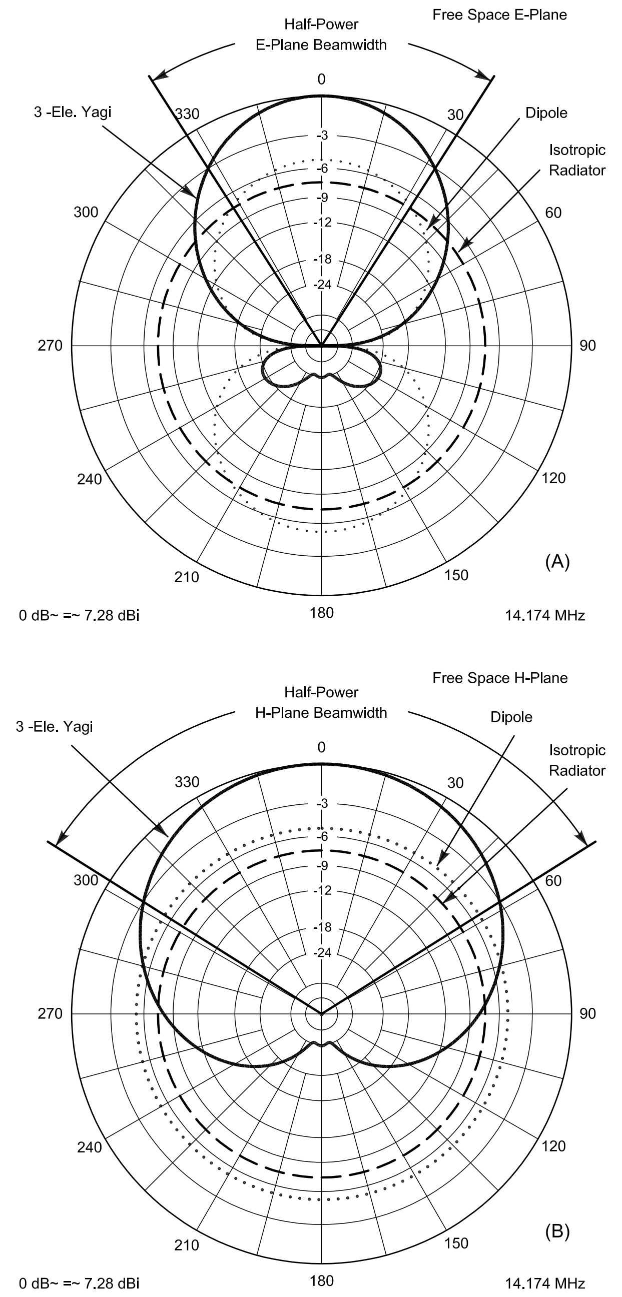

Fig 2—E-Plane (electric field) and H-Plane (magnetic field) response patterns for 3-element 20-meter Yagi in free space. At A the E-Plane pattern for a typical 3-element Yagi is compared with a dipole and an isotropic radiator. At B the H-Plane patterns are compared for the same antennas. The Yagi has an E-Plane half-power beamwidth of 66°, and an H-Plane half-power beamwidth of about 120°. The Yagi has 7.28 dBi (5.13 dBd) of gain. The front-to-back ratio, which compares the response at 0° and at 180°, is about 35 dB for this Yagi. The front-to-rear ratio, which compares the response at 0° to the largest lobe in the rearward 180° arc behind the antenna, is 24 dB, due to the lobes at 120° and 240°.

E-field is parallel to the earth (that is, the elements are parallel to the earth) then the antenna polarization is horizontal, and its E-field response is then usually referred to as its *azimuth* pattern. Its H-field response is then referred to as its *elevation* pattern.

Fig 2A demonstrates how this 3-element Yagi in free space exhibits 7.28 dBi of gain (referenced to isotropic), and has 5.13 dB gain over a free-space dipole. The gain is in the forward direction on the graph at 0° azimuth, and the forward part of the lobe is called the *main lobe*. For this particular antenna, the angular width of the E-plane main lobe at the half power, or 3 dB points compared to the peak, is about 66°. This performance characteristic is called the antenna's azimuthal *half-power beamwidth*.

Again as seen in Fig 2A, this antenna's response in the reverse direction at 180° azimuth is 34 dB less than in the forward direction. This characteristic is called the antenna's *front-to-back ratio*, and it describes the ability of an antenna to discriminate, for example, against interfering signals coming directly from the rear, when the antenna is being used for reception. In Fig 2A there are two sidelobes, at 120° and at 240° azimuth, which are about 24 dB down from the peak response at 0°. Since interference can come from any direction, not only directly off the back of an antenna, these kinds of sidelobes limit the ability to discriminate against rearward signals. The term *worst-case front-to-rear ratio* is used to describe the worst-case rearward lobe in the 180°-wide sector behind the antenna's main lobe. In this case, the worst-case front-to-rear ratio is 24 dB.

In the rest of this chapter the worst-case front-to-rear ratio will be used as a performance parameter, and will be abbreviated as "F/R." For a dipole or an isotropic radiator, Fig 2A demonstrates that F/R is 0 dB. Fig 2B depicts the H-field response for the same 3-element Yagi in free space, again compared to a dipole and an isotropic radiator in free space. Unlike the E-field pattern, the H-field pattern for a Yagi does not have a null at 90°, directly over the top of the Yagi. For this 3-element design, the H-field half-power beamwidth is approximately 120°.

Fig 3 compares the azimuth and elevation patterns for a horizontally polarized 6-element 14-MHz Yagi, with a 60-foot boom mounted one wavelength over ground, to a dipole at the same height. As with any horizontally polarized antenna, the height above ground is the main factor determining the peaks and nulls in the elevation pattern of each antenna. Fig 3A shows the E-field pattern, which has now been labeled as the Azimuth pattern. This antenna has a half-power azimuthal beamwidth of about 50°, and at an elevation angle of 12° it exhibits a forward gain of 16.02 dBi, including about 5 dB of ground reflection gain over relatively poor ground, with a dielectric constant of 13 and conductivity of 5 mS/m. In free space this Yagi has a gain of 10.97 dBi.

The H-field elevation response of the 6-element Yagi has a half-power beamwidth of about 60° in free space,

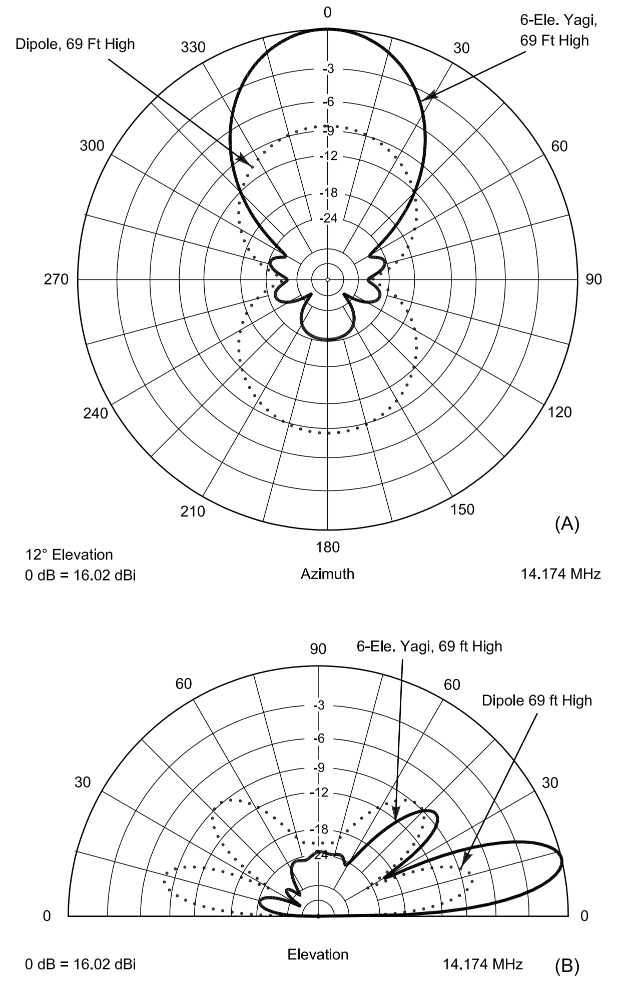

Fig 3—Azimuth pattern for 6-element 20-meter Yagi on 60-foot long boom, mounted 60 feet over ground. At A, the azimuth pattern at 12° elevation angle is shown, compared to a dipole at the same height. Peak gain of the Yagi is 16.04 dBi, or just over 8 dB compared to the dipole. At B, the elevation pattern for the same two antennas is shown. Note that the peak elevation pattern of the Yagi is compressed slightly lower compared to the dipole, even though they are both at the same height over ground. This is most noticeable for the Yagi's second lobe, which peaks at about 40°, while the dipole's second lobe peaks at about 48°. This is due to the greater free-space directionality of the Yagi at higher angles.

but as shown in Fig 3B, the first lobe (centered at 12° in elevation) has a half-power beamwidth of only 13° when the antenna is mounted one wavelength over ground. The dipole at the same height has a very slightly larger first-lobe half-power elevation beamwidth of 14°, since its free-space H-field response is omnidirectional.

Note that the free-space H-field directivity of the Yagi suppresses its second lobe over ground (at an elevation angle of about 40°) to 8 dBi, while the dipole's response at its second lobe peak (at about 48°) is at a level of 9 dBi.

The shape of the azimuthal pattern for a Yagi operated over real ground will change slightly as the Yagi is placed closer and closer to earth. Generally, however, the azimuth pattern doesn't depart significantly from the free-space pattern until the antenna is less than 0.5 λ high. This is just over 17 feet high at 28.4 MHz, and just under 35 feet at 14.2 MHz, heights that are not difficult to achieve for most amateurs. Some advanced computer programs can optimize Yagis at the exact installation height.

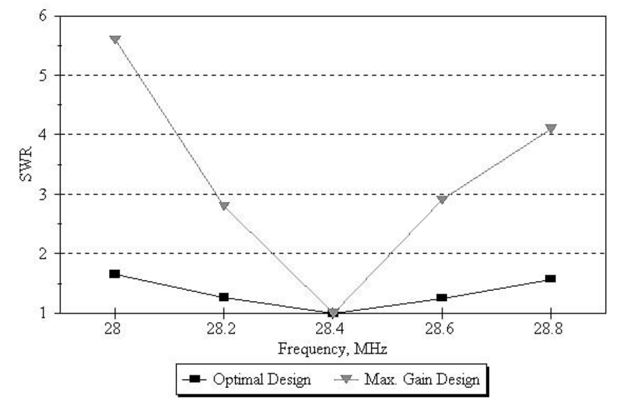

Fig 4—SWR over the 28.0 to 28.8 MHz portion of the 10-meter band for two different 3-element Yagi designs. One is designed strictly for maximum gain, while the second is optimized for F/R pattern and SWR over the frequency band. A Yagi designed only for maximum gain usually suffers from a very narrow SWR bandwidth.

DRIVE IMPEDANCE AND SWR

The impedance at the driven element in a Yagi is affected not only by the tuning of the driven element itself, but also by the spacing and tuning of nearby parasitic elements, and to a lesser extent by the presence of ground. In some designs that have been tuned solely for maximum gain, the driven-element impedance can fall to very low levels, sometimes less than 5 Ω. This can lead to excessive losses due to conductor resistance, especially at VHF and UHF. In a Yagi that has been optimized solely for gain, conductor losses are usually compounded by large excursions in impedance levels with relatively small changes in frequency. The SWR can thus change dramatically over a band and can create additional losses in the feed cable. **Fig 4** illustrates the SWR over the 28 to 28.8 MHz portion of the 10-meter amateur band for a 5-element Yagi on a 24-foot boom, which has been tuned for maximum forward gain at a spot frequency of 28.4 MHz. Its SWR curve is contrasted to that of a Yagi designed for a good compromise of gain, SWR and F/R.

Even professional antenna designers have difficulty accurately measuring forward gain. On the other hand, SWR can easily be measured by professional and amateur alike. Few manufacturers would probably want to advertise an antenna with the narrow-band SWR curve shown in Fig 4!

Monoband Yagi Performance Optimization

DESIGN GOALS

The previous section discussing driven-element impedance and SWR hinted at possible design trade-offs among gain, pattern and SWR, especially when each parameter is considered over a frequency band rather than at a spot frequency. Trade-offs in Yagi design parameters can be a matter of personal taste and operating style. For example, one operator might exclusively operate the CW portions of the HF bands, while another might only be interested in the Phone portions. Another operator may want a good pattern in order to discriminate against signals coming from a particular direction; someone else may want the most forward gain possible, and may not care about responses in other directions.

Extensive computer modeling of Yagis indicates that the parameter that must be compromised most to achieve wide bandwidths for front-to-rear ratio and SWR is forward gain. However, not much gain must be sacrificed for good F/R and SWR coverage, especially on long-boom Yagis. Although 10 and 7-MHz Yagis are not rare, the HF bands from 14 to 30 MHz are where Yagis are most often found, mainly due to the mechanical difficulties involved with making sturdy antennas for lower frequencies. The highest HF band, 28.0 to 29.7 MHz, represents the largest percentage bandwidth of the upper HF bands, at almost 6%. It is difficult to try to optimize in one design the main performance parameters of gain, worst-case F/R ratio and SWR over this large a band. Many commercial designs thus split up their 10-meter designs into antennas covering one of two bands: 28.0 to 28.8 MHz, and 28.8 to 29.7 MHz. For the amateur bands below 10 meters, optimal designs that cover the entire band are more easily achieved.

DESIGN VARIABLES

There are only a few variables available when one is designing a Yagi to meet certain design goals. The variables are:

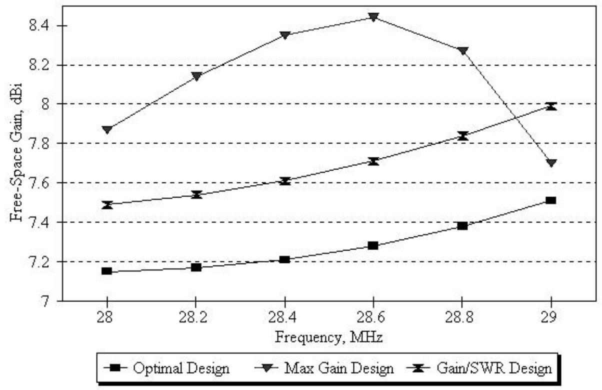

(A)

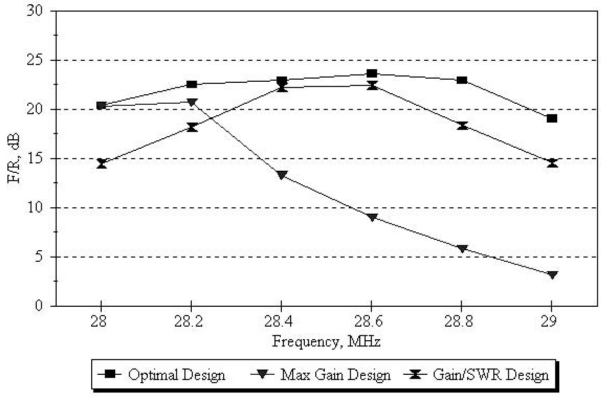

(B)

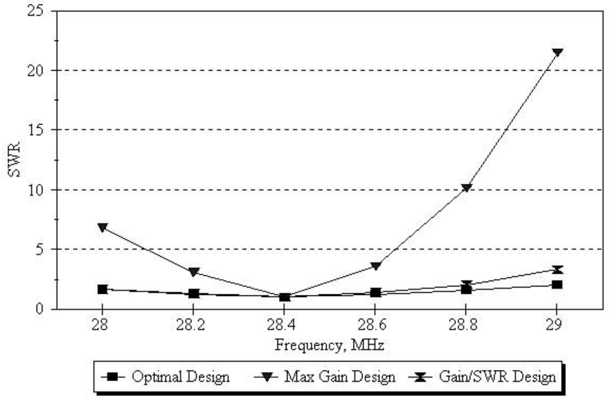

(C)

Fig 5—Comparisons of three different 3-element 10-meter Yagi designs using 8-foot booms. At A, gain comparisons are shown. The Yagi designed for the best compromise of gain and SWR sacrifices an average of about 0.5 dB compared to the antenna designed for maximum gain. The Yagi designed for optimal F/R, gain and SWR sacrifices an average of 1.0 dB compared to the maximum-gain case, and about 0.4 dB compared to the compromise gain and SWR case. At B, the front-to-rear ratio is shown for the three different designs. The antenna designed for optimal combination of gain, F/R and SWR maintains a F/R higher than 20 dB across the entire frequency range, while the antenna designed strictly for gain has a F/R of 3 dB at the high end of the band. At C, the three antenna designs are compared for SWR bandwidth. At the high end of the band, the antenna designed strictly for gain has a very high SWR.

1. The physical length of the boom
2. The number of elements on the boom
3. The spacing of each element along the boom
4. The tuning of each element
5. The type of matching network used to feed the array.

GAIN AND BOOM LENGTH

As pointed out earlier, the gain of a Yagi is largely a function of the length of the boom. As the boom is made longer, the maximum gain potential rises. For a given boom length, the number of elements populating that boom can be varied, while still maintaining the antenna's gain, provided of course that the elements are tuned properly. In general, putting more elements on a boom gives the designer added flexibility to achieve desired design goals, especially to spread the response out over a frequency band.

Fig 5A is an example illustrating gain versus frequency for three different types of 3-element Yagis on 8-foot booms. The three antennas were designed for the lower end of the 10-meter band, 28.0 to 28.8 MHz, based on the following different design goals:

Antenna 1: Maximum mid-band gain, regardless of F/R or SWR across the band
Antenna 2: SWR less than 2:1 over the frequency band; best compromise gain, with no special consideration for F/R over the band.
Antenna 3: "Optimal" case: F/R greater than 20 dB, SWR less than 2:1 over the frequency band; best compromise gain.

Fig 5B shows the F/R over the frequency band for these three designs, and Fig 5C shows the SWR curves over the frequency band. Antenna 1, the design that strives strictly for maximum gain, has a poor SWR response over the band, as might be expected after the previous section discussing SWR. The SWR is 10:1 at 28.8 MHz and rises to 22:1 at 29 MHz. At 28 MHz, at the low end of the band, the SWR of the maximum-gain design is more than 6:1. Clearly, designing for maximum gain alone produces an unacceptable design in terms of SWR bandwidth. The F/R for Antenna 1 reaches a high point of about 20 dB at the low-frequency end of the band, but falls to only 3 dB at the high-frequency end.

Antenna 2, designed for the best compromise of gain while the SWR across the band is held to less than 2:1, achieves this goal, but at an average gain sacrifice of 0.7 dB compared to the maximum gain case. The F/R for this design is just under 15 dB over the band. This design is fairly typical of many amateur Yagi designs before the advent of computer modeling and optimization programs. SWR can easily be measured, and experimental optimization for forward gain is a fairly straightforward procedure. By contrast, overall pattern optimization is not a trivial thing to achieve experimentally, particularly for antennas with more than four or five elements.

Antenna 3, designed for an optimum combination of F/R, SWR and gain, compromises forward gain an average of 1.0 dB compared to the maximum gain case, and about 0.4 dB compared to the compromise gain/SWR case. It achieves its design objectives of more than 20 dB F/R over the 28.0 to 28.8 MHz portion of the band, with an SWR less than 2:1 over that range.

Fig 6A shows the free-space gain versus frequency for the same three types of designs, but for a bigger 5-element 10-meter Yagi on a 20-foot boom. Fig 6B shows the variation in F/R, and Fig 6C shows the SWR curves versus frequency. Once again, the design that concentrates solely on maximum gain has a poor SWR curve over the band, reaching just over 6:1 toward the high end of the band. The difference in gain between the maximum gain case and the optimum design case has narrowed for this size of boom to an average of under 0.5 dB. This comes about because the designer has access to more variables in a 5-element design than he does in a 3-element design, and he can stagger-tune the various elements to spread the response out over the whole band.

Fig 7A, **B** and **C** show the same three types of designs, but for a 6-element Yagi on a 36-foot boom. The SWR bandwidth of the antenna designed for maximum gain has improved compared to the previous two shorter-boom examples, but the SWR still rises to more than 4:1 at 28.8 MHz, while the F/R ratio is pretty constant over the band, at a mediocre 11 dB average level. While the antenna designed for gain and SWR does hold the SWR below 2:1 over the band, it also has the same mediocre level of F/R performance as does the maximum-gain design.

The optimized 36-foot boom antenna achieves an excellent F/R of more than 22 dB over the whole 28.0 to 28.8 MHz band. Again, the availability of more elements and more space on the 36-foot long boom gives the designer more flexibility in broadbanding the response over the whole band, while sacrificing only 0.3 dB of gain compared to the maximum-gain design.

Fig 8A, **B**, and **C** show the same three types of 10-meter designs, but now for a 60-foot boom, populated with eight elements. With eight elements and a very long boom on which to space them out, the antenna designed solely for maximum gain can achieve a much better SWR response across the band, although the SWR does rise to more than 7:1 at the very high end of the band. The SWR remains less than 2:1 from 28.0 to 28.7 MHz, much better than for shorter-boom, maximum-gain designs. The worst-case F/R ratio is never better than 19 dB, however, and remains around 10 dB over much of the band. The antenna designed for the best compromise gain and SWR loses only about 0.1 dB of gain compared to the maximum-gain design, but does little better in terms of F/R across the band.

Contrasted to these two designs, the antenna optimized for F/R, SWR and gain has an outstanding pattern, exhibiting an F/R of more than 24 dB across the entire

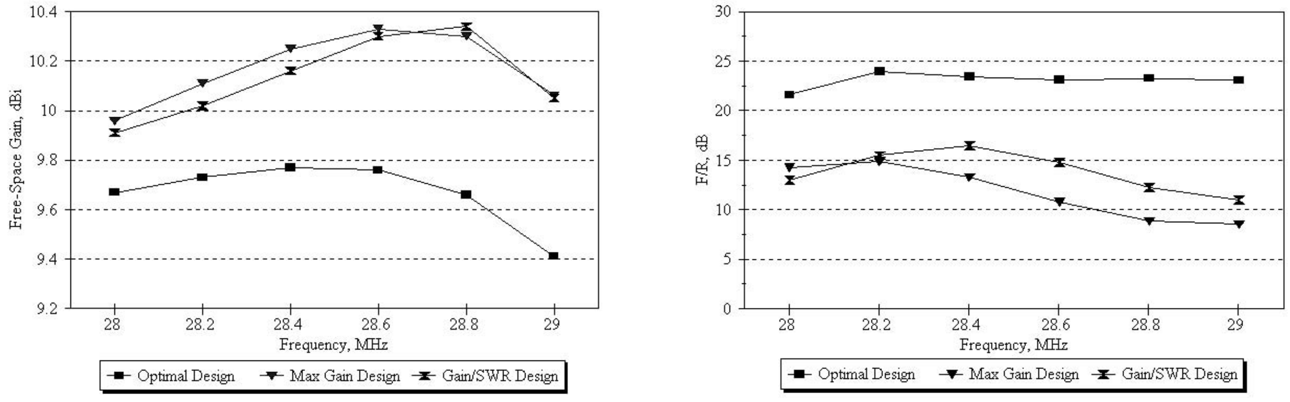

(A)

(B)

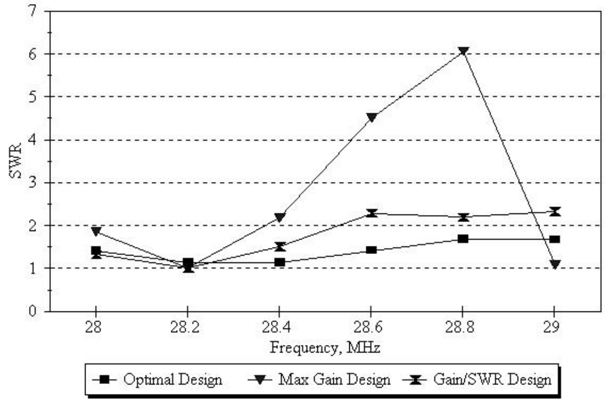

(C)

Fig 6—Comparisons of three different designs for 5-element 10-meter Yagis on 20-foot booms. At A, the gain of three different 5-element 10-meter Yagi designs are graphed. The difference in gain between the three antennas narrows because the elements can be stagger-tuned to spread the response out better over the desired frequency band. The average gain reduction for the fully optimized antenna design is about 0.5 dB. At B, the optimal antenna displays better than 22 dB F/R over the band, while the Yagi designed for gain and SWR displays on average 10 dB less F/R throughout the band. At C, the SWR bandwidth is compared for the three Yagis. The antenna designed strictly for forward gain has a poor SWR bandwidth and a high peak SWR of 6:1 at 28.8 MHz.

band, while keeping the SWR below 2:1 from 28.0 to 28.9 MHz. It must sacrifice an average of only 0.4 dB compared to the maximum gain design at the low end of the band, and actually has more gain than the maximum gain and gain/SWR designs at the high-frequency end of the band.

The conclusion drawn from these and many other detailed comparisons is that designing strictly for maximum mid-band gain yields an inferior design when the antenna is examined over an entire frequency band, especially in terms of SWR. Designing a Yagi for both gain and SWR will yield antennas that have mediocre rearward patterns, but that lose relatively little gain compared to the maximum gain case, at least for designs with more than three elements.

However, designing a Yagi for an optimal combination of F/R, SWR and gain results in a loss of gain less than 0.5 dB compared to designs designed only for gain and SWR. **Fig 9** summarizes the forward gain achieved for the three different design types versus boom length, as expressed in wavelength.

Except for the 2-element designs, the Yagis described in the rest of this chapter have the following design goals over a desired frequency band:

1. Front-to-rear ratio over the frequency band of more than 20 dB
2. SWR over the frequency band less than 2:1
3. Maximum gain consistent with points 1 and 2 above

Just for fun, **Fig 10** shows the gain versus boom length for theoretical 20-meter Yagis that have been designed to meet the three design goals above. The 31-element design for 14 MHz would be wondrous to behold. Sadly, it is unlikely that anyone will build one, considering that the boom would be 724 feet long! However, such a design *does* become practical when scaled to 432 MHz. In fact, a K1FO 22-element and a K1FO 31-element Yagi are the prototypes for the theoretical 14-MHz long-boom designs. See Chapter 18, VHF and UHF Antenna Systems.

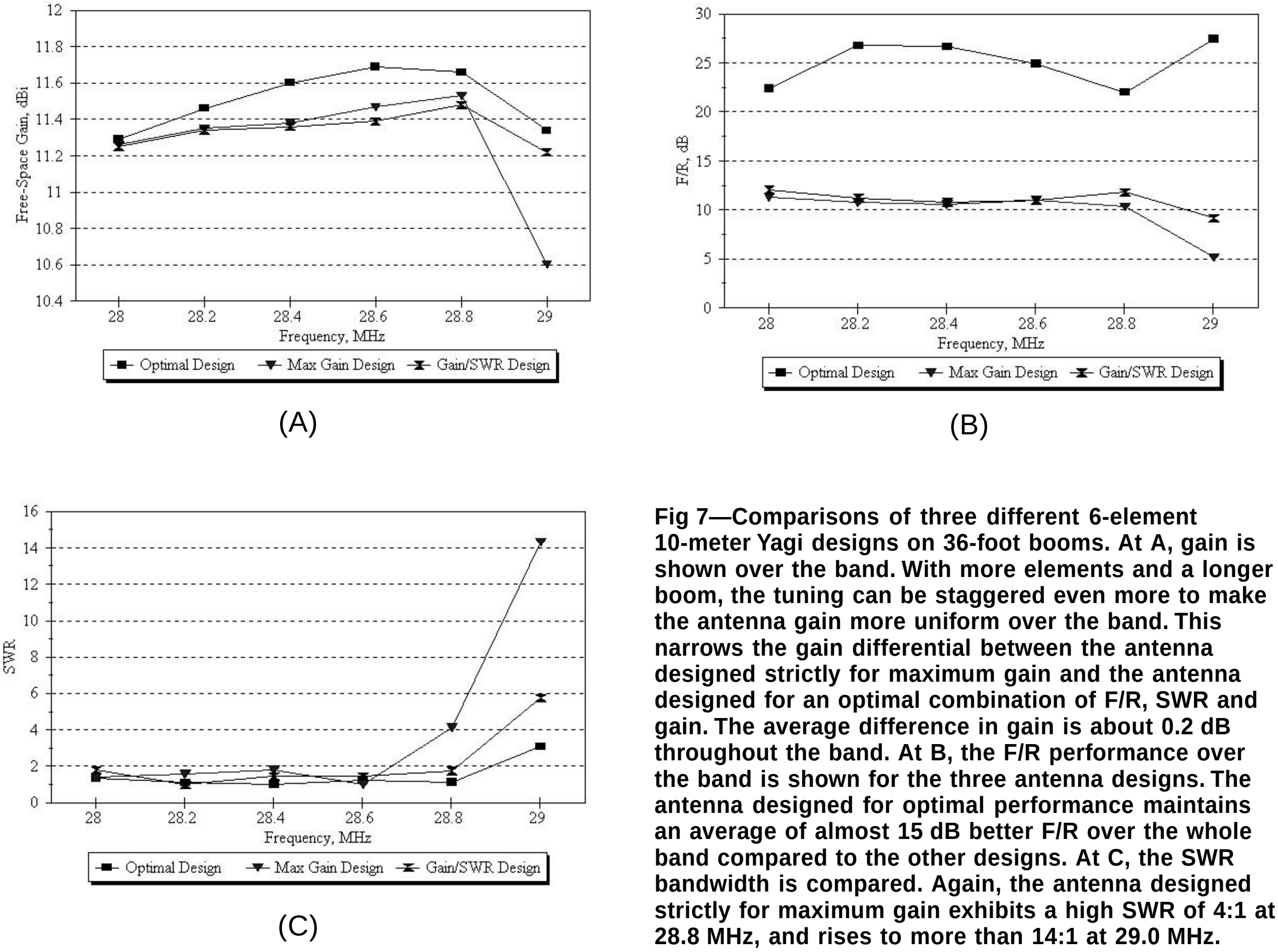

(A)

(B)

(C)

Fig 7—Comparisons of three different 6-element 10-meter Yagi designs on 36-foot booms. At A, gain is shown over the band. With more elements and a longer boom, the tuning can be staggered even more to make the antenna gain more uniform over the band. This narrows the gain differential between the antenna designed strictly for maximum gain and the antenna designed for an optimal combination of F/R, SWR and gain. The average difference in gain is about 0.2 dB throughout the band. At B, the F/R performance over the band is shown for the three antenna designs. The antenna designed for optimal performance maintains an average of almost 15 dB better F/R over the whole band compared to the other designs. At C, the SWR bandwidth is compared. Again, the antenna designed strictly for maximum gain exhibits a high SWR of 4:1 at 28.8 MHz, and rises to more than 14:1 at 29.0 MHz.

OPTIMUM DESIGNS AND ELEMENT SPACING

Two-Element Yagis

Many hams consider a 2-element Yagi to give “the most bang for the buck” among various Yagi designs, particularly for portable operations such as Field Day. A 2-element Yagi has about 4 dB of gain over a simple dipole (sometimes jokingly called a “one-element Yagi”) and gives a modest F/R of about 10 dB to help with rejection of interference on receive. By comparison, going from a 2-element to a 3-element Yagi increases the boom length by about 50% and adds another element, a 50% increase in the number of elements—for a gain increase of about 1 dB and another 10 dB in F/R.

Element Spacing in Larger Yagis

One of the more interesting results of computer modeling and optimization of high-performance Yagis with four or more elements is that a distinct pattern in the element spacings along the boom shows up consistently. This pattern is relatively independent of boom length, once the boom is longer than about 0.3 λ.

The reflector, driven element and first director of these optimal designs are typically bunched rather closely together, occupying together only about 0.15 to 0.20 λ of the boom. This pattern contrasts sharply with older designs, where the amount of boom taken up by the reflector, driven element and first director was typically more than 0.3 λ. **Fig 11** shows the element spacings for an optimized 6-element, 36-foot boom, 10-meter design, compared to a W2PV 6-element design with constant spacing of 0.15 λ between all elements.

A problem arises with such a bunching of elements toward the reflector end of the boom—the wind loading of the antenna is not equal along the boom. Unless properly compensated, such new-generation Yagis will act like windvanes, punishing, and often breaking, the rotators trying to turn, or hold, them in the wind. One successful solution to windvaning has been to employ “dummy elements” made of PVC piping. These nonconducting elements are placed on the boom close to the last director so

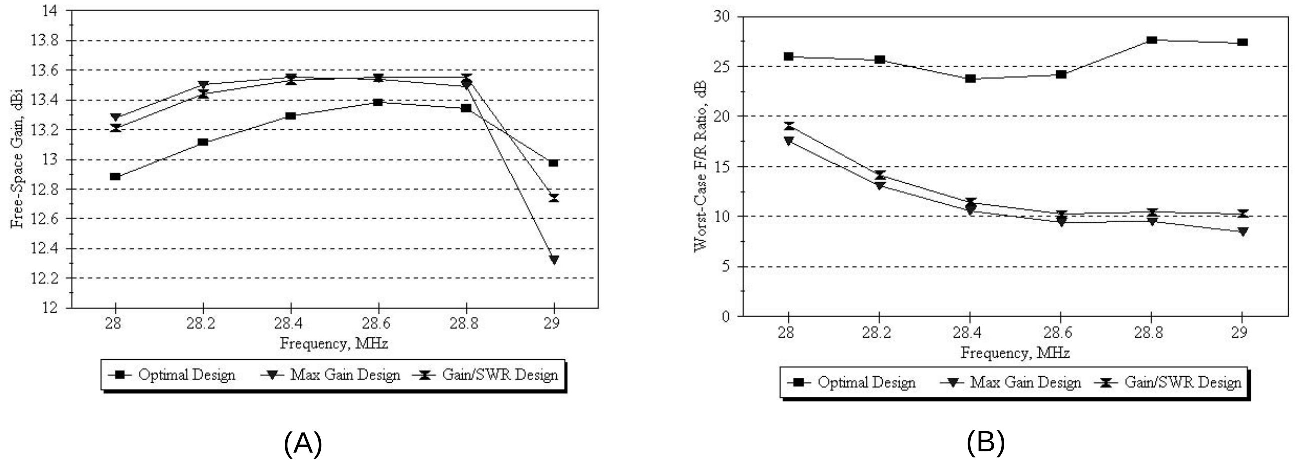

(A)

(B)

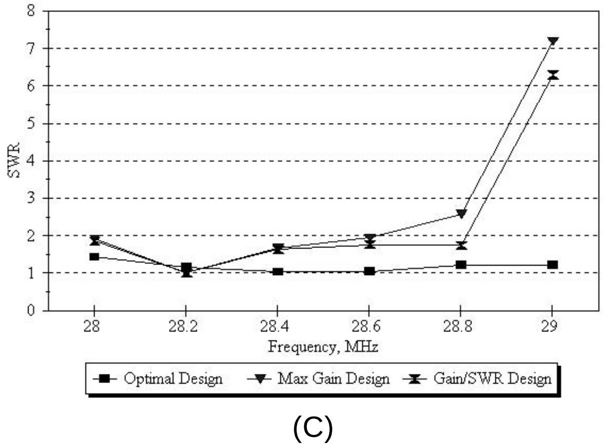

(C)

Fig 8—Comparisons of three different 8-element 10-meter Yagi designs using 60-foot booms. At A, gain is shown over the frequency band. With even more freedom to stagger-tune elements and a very long boom on which to place them, the average antenna gain differential over the band is now less than 0.2 dB between the three design cases. At B, an excellent 24 dB F/R for the optimal design is maintained over the whole band, compared to the average of about 12 dB for the other two designs. At C, the SWR differential over the band is narrowed between the three designs, again because there are more variables available to broaden the bandwidth.

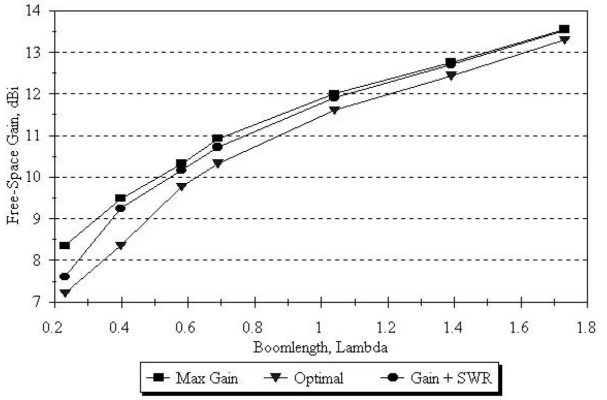

Fig 9—Gain versus boom length for three different 10-meter design goals. The goals are: (1) designed for maximum gain across band, (2) designed for a compromise of gain and SWR, and (3) designed for optimal F/R, SWR and gain across 28.0 to 28.8 MHz portion of 10-meter band. The gain difference is less than 0.5 dB for booms longer than approximately 0.5 λ.

the windload is equalized at the mast-to-boom bracket. In addition, it may be necessary to insert a small amount of lead weight at one end of the boom in order to balance the antenna weight.

Despite the relatively close spacing of the reflector, driven element and first director, modern optimal Yagi designs are not overly sensitive to small changes in either element length or spacing. In fact, these antennas can be constructed from design tables without excessive concern about close dimensional tolerances. In the HF range up to 30 MHz, building the antennas to the nearest $^1/_8$-inch results in performance remarkably consistent with the computations, without any "tweaking" or fine-tuning when the Yagi is on the tower.

ELEMENT TUNING

Element tuning (or *self-impedance*) is a complex function of the effective electrical length of each element and the effective diameter of the element. In turn, the effective length and diameter of each element is related to

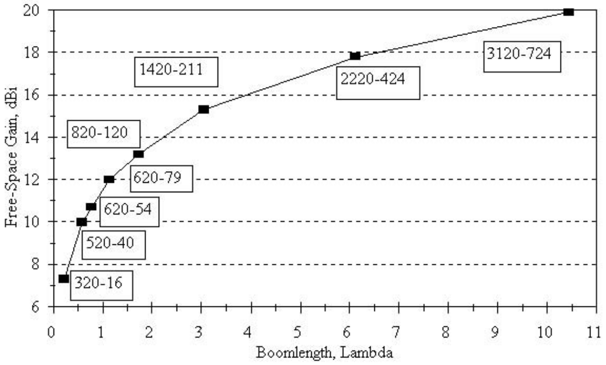

Fig 10—Theoretical gain versus boom length for 20-meter Yagis designed for optimal combination of F/R, SWR and gain across the entire 14.0 to 14.35 MHz band. The theoretical gain approaches 20 dBi for a gigantic 724-foot boom, populated with 31 elements. Such a design on 20 meters is not too practical, of course, but can readily be achieved on a 24-foot boom on 432 MHz.

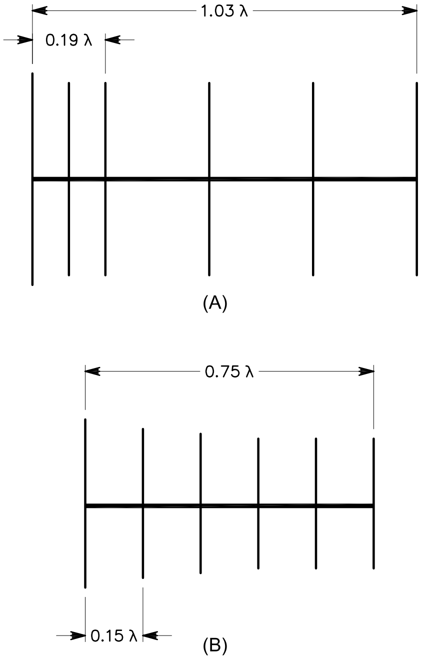

Fig 11—Tapering spacing versus constant element spacing. At A, illustration of how the spacing of the reflector, driven element and first director (over the first 0.19 λ of the boom) of an optimally designed Yagi is bunched together compared to the Yagi at B, which uses constant 0.15 λ spacing between all elements. The optimally designed antenna has more than 22 dB F/R and an SWR less than 1.5:1 over the frequency band 28.0 to 28.8 MHz.

the taper schedule (if telescoping aluminum tubing is used, the most common method of construction), the length of each telescoping section, the type and size of mounting bracket used to secure the element to or through the boom, and the size of the Yagi boom itself. See the section entitled “Antenna Frequency Scaling,” and “Tapered Elements” in Chapter 2, Antenna Fundamentals, of this book for details about element tuning as a function of tapering and element diameter. Note especially that Yagis constructed using wire elements will perform very differently compared to the same antenna constructed with elements made of telescoping aluminum tubing.

The process by which a modern Yagi is designed usually starts out with the selection of the longest boom possible for a given installation. A suitable number of elements of a given taper schedule are then placed on this boom, and the gain, pattern and SWR are calculated over the entire frequency band of interest to the operator. Once an electrical design is chosen, the designer must then ensure the mechanical integrity of the antenna design. This involves verifying the integrity of the boom and each element in the face of the wind and ice loading expected for a particular location. The section entitled “Construction with Aluminum Tubing” in Chapter 20, Antenna Materials and Accessories, of this book shows details of tapered telescoping aluminum elements for the upper HF bands. In addition, the ARRL book *Physical Design of Yagi Antennas*, by Dave Leeson, W6NL (ex-W6QHS), describes the mechanical design process for all portions of a Yagi antenna very thoroughly, and is highly recommended for serious Yagi builders.

Specific Monoband Yagi Designs

The detailed Yagi design tables that follow are for two taper schedules for HF Yagis covering the 14 through 30-MHz amateur bands. The heavy-duty elements are designed to survive at least 120-mph winds without icing, or 85-mph winds with 1/4-inch radial ice. The medium-duty elements are designed to survive winds greater than 80 mph, or 60-mph winds with 1/4-inch radial ice.

For 10.1 MHz, the elements shown are capable of surviving 105-mph winds, or 93-mph winds with 1/4-inch radial ice. For 7.1 MHz the elements shown can survive 93-mph winds, or 69-mph winds with 1/4-inch radial ice. For these two lower frequency bands, the elements and the booms needed are very large and heavy. Mounting, turning and keeping such antennas in the air is not a trivial task.

Each element is mounted above the boom with a heavy rectangular aluminum plate, by means of U-bolts with saddles, as shown in Fig 35 in Chapter 18, VHF and UHF Antenna Systems for a 6-meter Yagi. This method of element mounting is rugged and stable, and because the element is mounted away from the boom, the amount of element detuning due to the presence of the boom is minimal. The element dimensions given in each table already take into account any element detuning due to the boom-to-element mounting plate. For each element, the length of the tip determines the tuning, since the inner tubes are fixed in diameter and length.

Half Elements

Each design shows the dimensions for *one-half* of each element, mounted on *one side* of the boom. The other half of each element is symmetrical, mounted on the other side of the boom. The use of a tubing sleeve inside the center portion of the element is recommended, so that the element is not crushed by the mounting U-bolts. Unless otherwise noted, each section of tubing is made of 6061-T6 aluminum tubing, with a 0.058-inch wall thickness. This wall thickness ensures that the next standard size of tubing can telescope with it. Each telescoping section is inserted 3 inches into the larger tubing, and is secured by one of the methods shown in Fig 11 in Chapter 20, Antenna Materials and Accessories.

Matching System

Each antenna is designed with a driven-element length appropriate for a hairpin type of matching network. The driven-element's length may require slight readjustment for best match, particularly if a different matching network is used. *Do not change* either the lengths or the telescoping tubing schedule of the parasitic elements—they have been optimized for best performance and will not be affected by tuning of the driven element!

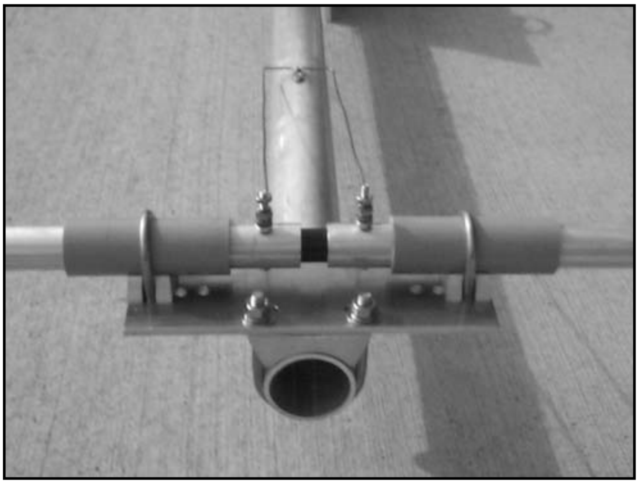

Fig 12—Typical construction techniques for an HF Yagi. This photo shows a hairpin match on a driven element that uses a fiberglass insulator (wrapped in black vinyl tape for protection against UV). Muffler clamps and saddles mount the element to the boom, while U-bolts and saddles mount the element to the boom-to-element plate. The gray PVC sleeves insulate the element from the plate. The feed coax is connected to the two bolts that also connect to the hairpin wire. Note that the hairpin is grounded at its opposite end to dissipate static charges that might otherwise build up.

Fig 12 is a photograph of the driven element for a 2-element 17-meter Yagi built by Chuck Hutchinson, K8CH, for the ARRL book *Simple and Fun Antennas for Hams.* The aluminum tubing on each side of the boom was 1-inch OD, and the two pieces were mechanically joined together with a 3/4-inch OD fiberglass insulator. Chuck wound electrical tape over the insulator to protect the fiberglass from the sun's UV.

Chuck used 3-inch lengths of 1-inch sunlight-resistant PVC conduit, split lengthwise, to make the grey outer insulators for the driven element. The aluminum plates came from DX Engineering, as did the stainless-steel U-bolts and saddle clamps. These saddles ensured that the elements don't rotate on the 2-inch OD boom in the heavy winds in his part of rural Michigan.

You can see the bolts used to pin the center fiberglass insulator to the aluminum tubing, while also providing an electrical connection for the #12 hairpin wire and for the feed-line coax, which uses ferrite beads over the coax's outer vinyl jacket to make a common-mode current-type of balun (not shown in Fig 12). Note that the center of the hairpin is connected to the boom using a grounding lug for some measure of protection from static buildup.

10-METER YAGIS

Fig 13 describes the electrical performance of eight optimized 10-meter Yagis with boom lengths between 6 to 60 feet. The end of each boom includes 3 inches of space for the reflector and last-director (or driven element for the 2-element designs) mounting plates. Fig 13A shows the free-space gain versus frequency for each antenna; 13B shows the front-to-rear ratio, and 13C shows the SWR versus frequency. Each antenna with three or more elements was designed to cover the lower half of the 10-meter band from 28.0 to 28.8 MHz, with SWR less than 2:1 and F/R better than 20 dB over that range.

Fig 13D shows the taper schedule for two types of 10-meter elements. The heavy-duty design can survive 125-mph winds with no icing, and 88-mph winds with $^1/_4$-inch of radial ice. The medium-duty design can handle 96-mph winds with no icing, and 68-mph winds with $^1/_4$-inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.250-inch thick flat aluminum plate, 4 inches wide by 4 inches long. Each element except for the insulated driven element, is centered on the plate, held by two stainless-steel U-bolts with saddles. Another set

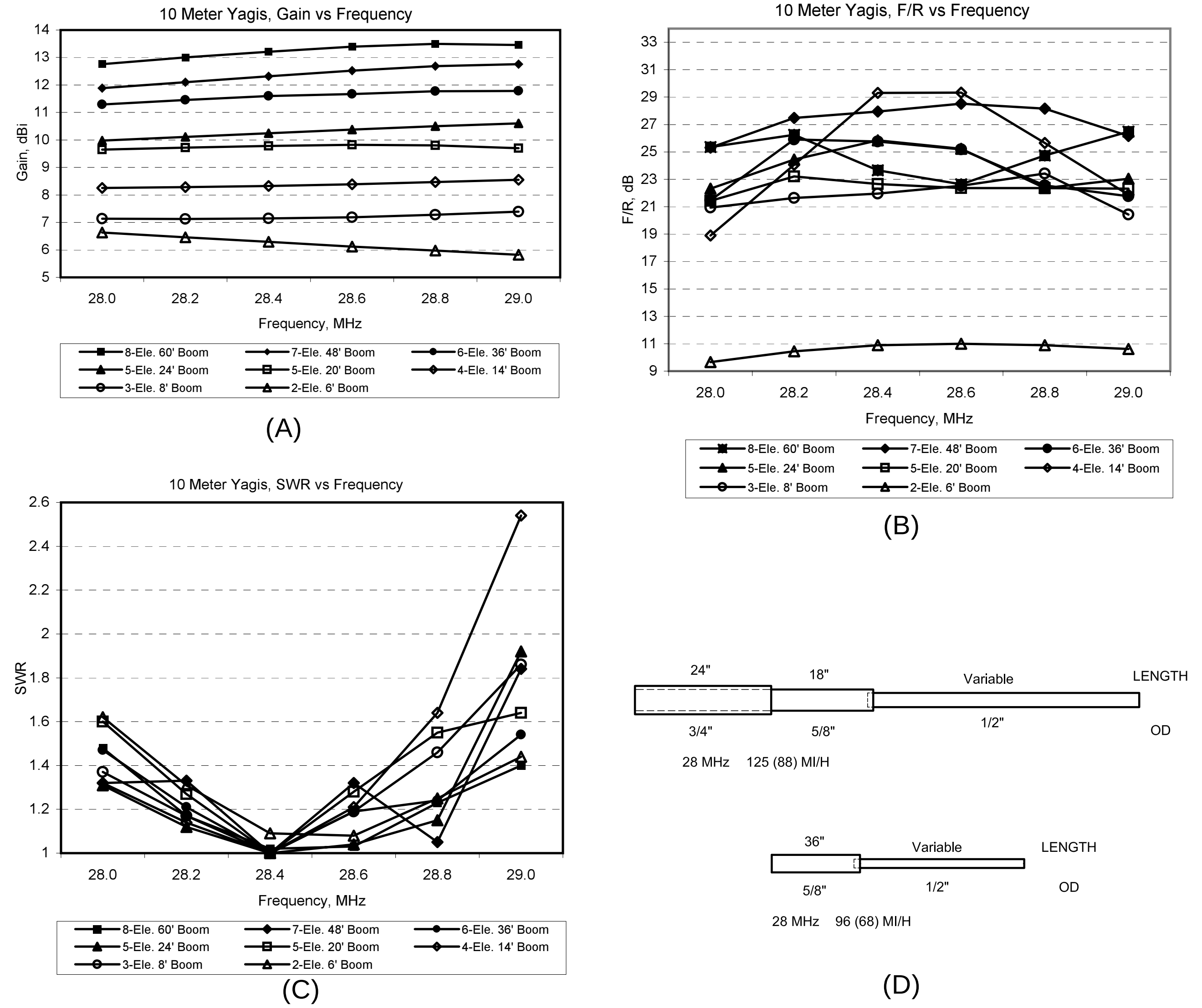

Fig 13—Gain, F/R and SWR performance versus frequency for optimized 10-meter Yagis. At A, gain is shown versus frequency for eight 10-meter Yagis whose booms range from 6 feet to 60 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the frequency range 28.0 to 28.8 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR is shown over the frequency range. At D, the taper schedule is shown for heavy-duty and for medium-duty 10-meter elements. The heavy-duty elements can withstand 125-mph winds without icing, and 88-mph winds with $^1/_4$-inch radial ice. The medium-duty elements can survive 96-mph winds without icing, and 68-mph winds with $^1/_4$-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

Table 1

Optimized 10-Meter Yagi Designs

Two-element 10-meter Yagi, 6 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		210-06H.YW	210-06M.YW
Reflector	0.000"	66.000"	71.500"
Driven Element	66.000"	57.625"	63.000"

Three-element 10-meter Yagi, 8 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		310-08H.YW	310-08M.YW
Reflector	0.000"	66.750"	71.875"
Driven Element	36.000"	57.625"	62.875"
Director 1	54.000"	53.125"	58.500"
Compensator	12" behind Dir. 1	19.000"	18.125"

Four-element 10-meter Yagi, 14 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		410-14H.YW	410-14M.YW
Reflector	0.000"	66.000"	72.000"
Driven Element	36.000"	58.625"	63.875"
Director 1	36.000"	57.000"	62.250"
Director 2	90.000"	47.750"	53.125"
Compensator	12" behind Dir. 2	22.000"	20.500"

Five-element 10-meter Yagi, 24 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		510-24H.YW	510-24M.YW
Reflector	0.000"	65.625"	70.750"
Driven Element	36.000"	58.000"	63.250"
Director 1	36.000"	57.125"	62.375"
Director 2	99.000"	55.000"	60.250"
Director 3	111.000"	50.750"	56.125"
Compensator	12" behind Dir. 3	28.750"	26.750"

Six-element 10-meter Yagi, 36 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		610-36H.YW	610-36M.YW
Reflector	0.000"	66.500"	71.500"
Driven Element	37.000"	58.500"	64.000"
Director 1	43.000"	57.125"	62.375"
Director 2	98.000"	54.875"	60.125"
Director 3	127.000"	53.875"	59.250"
Director 4	121.000"	49.875"	55.250"
Compensator	12" behind Dir. 4	32.000"	29.750"

Seven-element 10-meter Yagi, 48 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		710-48H.YW	710-48M.YW
Reflector	0.000"	65.375"	70.500"
Driven Element	37.000"	59.000"	64.250"
Director 1	37.000"	57.500"	62.750"
Director 2	96.000"	54.875"	60.125"
Director 3	130.000"	52.250"	57.625"
Director 4	154.000"	52.625"	58.000"
Director 5	116.000"	49.875"	55.250"
Compensator	12" behind Dir. 5	35.750"	33.750"

Eight-element 10-meter Yagi, 60 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		810-60H.YW	810-60M.YW
Reflector	0.000"	65.000"	70.125"
Driven Element	42.000"	58.000"	63.500"
Director 1	37.000"	57.125"	62.375"
Director 2	87.000"	55.375"	60.625"
Director 3	126.000"	53.250"	58.625"
Director 4	141.000"	51.875"	57.250"
Director 5	157.000"	52.500"	57.875"
Director 6	121.000"	50.125"	55.500"
Compensator	12" behind Dir. 6	59.375"	55.125"

These 10-meter Yagi designs are optimized for > 20 dB F/R, and SWR < 2:1 over frequency range from 28.000 to 28.800 MHz, for heavy-duty elements (125 mph wind survival) and for medium-duty (96 mph wind survival). For coverage from 28.8 to 29.7 MHz, subtract 2.000 inches from end of each element, but leave element spacings the same as shown here. Only element tip dimensions are shown, and all dimensions are inches. See Fig 13D for element telescoping tubing schedule. Torque compensator element is made of 2.5" OD PVC water pipe placed 12 inches behind last director. Dimensions shown for compensators is one-half of total length, centered on boom.

of U-bolts with saddles is used to secure the mounting plate to the boom.

Electrically each mounting plate is equivalent to a cylinder, with an effective diameter of 2.405 inches for the heavy-duty element, and 2.310 inches for the medium-duty element. The equivalent length on each side of the boom is 2 inches. These dimensions are incorporated in the files for the *YW* (Yagi for Windows) computer modeling program on the CD-ROM accompanying this book to simulate the effect of the mounting plate.

The second column in **Table 1** shows the spacing of each element relative to the next element in line on the boom, starting at the reflector, which itself is defined as being at the 0.000-inch reference point on the boom. The boom for antennas less than 30 feet long can be constructed of 2-inch OD tubing with 0.065-inch wall thickness. Designs larger than 30 feet long should use 3-inch OD heavy-wall tubing for the boom. Because each boom has extra space at each end, the reflector is actually placed 3 inches from the end of the boom. For example, in the 310-08H.YW design (3 elements on an 8-foot boom), the driven element is placed 36 inches ahead of the reflector, and the director is placed 54 inches ahead of the driven element.

The next columns give the lengths for the variable tips for the heavy-duty and then the medium-duty elements. In the example above for the 310-08H.YW Yagi, the heavy-duty reflector tip, made out of 1/2-inch OD tubing, sticks out 66.750 inches from the 5/8-inch OD tubing. Note that each telescoping piece of tubing overlaps 3 inches inside the piece into which it fits, so the overall length of 1/8-inch OD tubing is 69.750 inches long for the reflector. The medium-duty reflector tip has 71.875 inches protruding from the 5/8-inch OD tube, and is 74.875 inches long overall. As previously stated, the dimensions are not extremely critical, although measurement accu-

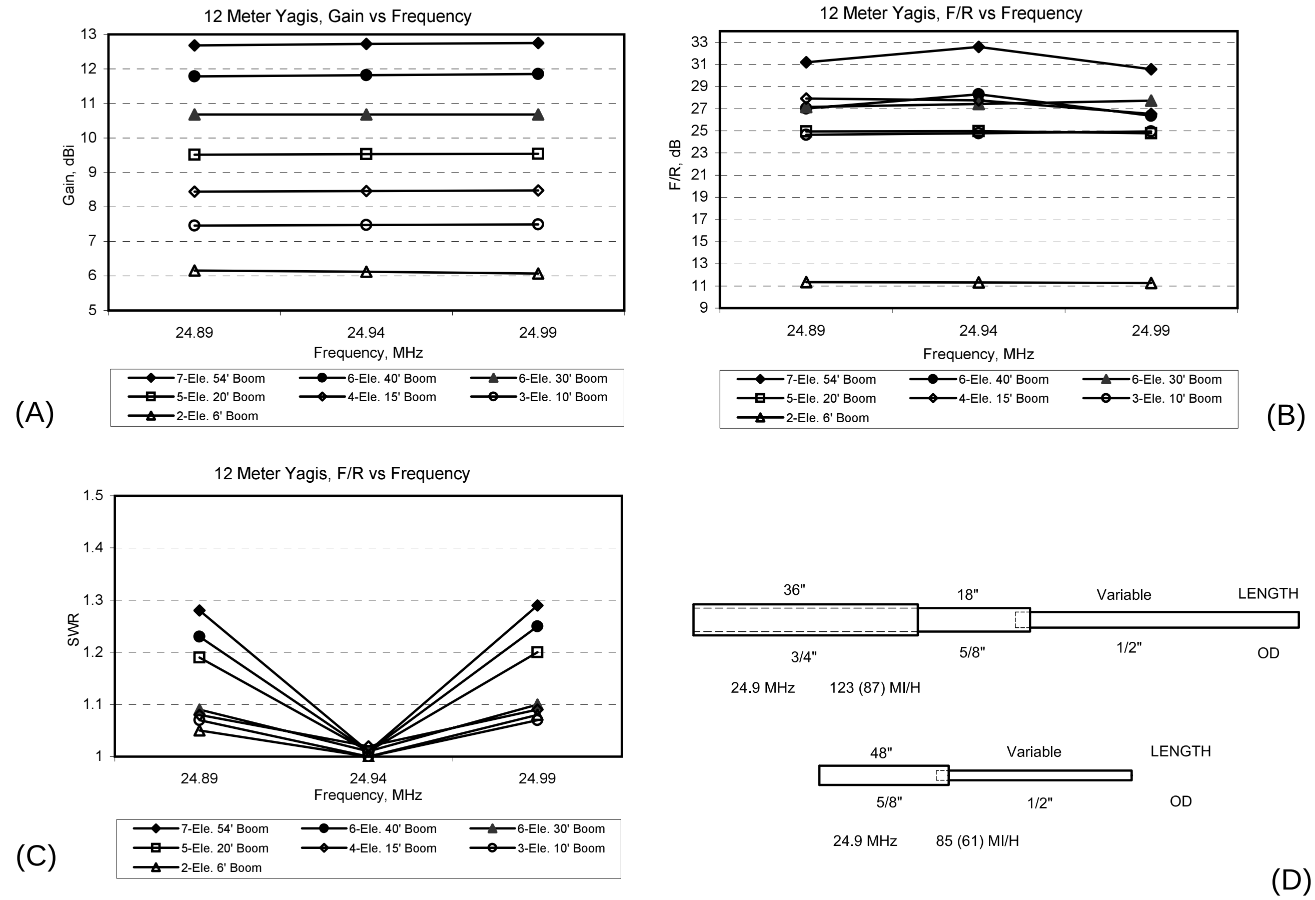

Fig 14—Gain, F/R and SWR performance versus frequency for optimized 12-meter Yagis. At A, gain is shown versus frequency for seven 12-meter Yagis whose booms range from 6 feet to 54 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the narrow 12-meter band 24.89 to 24.99 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule for heavy-duty and for medium-duty 12-meter elements is shown. The heavy-duty elements can withstand 123-mph winds without icing, and 87-mph winds with 1/4-inch radial ice. The medium-duty elements can survive 85-mph winds without icing, and 61-mph winds with 1/4-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

Table 2
Optimized 12-Meter Yagi Designs

Two-element 12-meter Yagi, 6 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		212-06H.YW	212-06M.YW
Reflector	0.000"	67.500"	72.500"
Driven Element	66.000"	59.500"	65.000"

Three-element 12-meter Yagi, 10 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		312-10H.YW	312-10M.YW
Reflector	0.000"	69.000"	73.875"
Driven Element	40.000"	60.250"	65.250"
Director 1	74.000"	54.000"	59.125"
Compensator	12" behind Dir. 1	13.625"	12.000"

Four-element 12-meter Yagi, 15 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		412-15H.YW	412-15M.YW
Reflector	0.000"	66.875"	71.875"
Driven Element	46.000"	61.000"	66.000"
Director 1	46.000"	58.625"	63.750"
Director 2	82.000"	50.875"	56.125"
Compensator	12" behind Dir. 2	16.375"	14.500"

Five-element 12-meter Yagi, 20 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		512-20H.YW	512-20M.YW
Reflector	0.000"	69.750"	74.625"
Driven Element	46.000"	62.250"	67.000"
Director 1	46.000"	60.500"	65.500"
Director 2	48.000"	55.500"	60.625"
Director 3	94.000"	54.625"	59.750"
Compensator	12" behind Dir. 3	22.125"	19.625"

Six-element 12-meter Yagi, 30 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		612-30H.YW	612-30M.YW
Reflector	0.000"	68.125"	73.000"
Driven Element	46.000"	61.750"	66.750"
Director 1	46.000"	60.250"	65.250"
Director 2	73.000"	52.375"	57.625"
Director 3	75.000"	57.625"	62.750"
Director 4	114.000"	53.625"	58.750"
Compensator	12" behind Dir. 4	30.000"	26.250"

Six-element 12-meter Yagi, 40 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		612-40H.YW	612-40M.YW
Reflector	0.000"	67.000"	71.875"
Driven Element	46.000"	60.125"	65.500"
Director 1	46.000"	57.375"	62.500"
Director 2	91.000"	57.375"	62.500"
Director 3	157.000"	57.000"	62.125"
Director 4	134.000"	54.375"	59.500"
Compensator	12" behind Dir. 4	36.500"	31.625"

Seven-element 12-meter Yagi, 54 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		712-54H.YW	712-54M.YW
Reflector	0.000"	68.000"	73.000"
Driven Element	46.000"	60.500"	65.500"
Director 1	46.000"	56.750"	61.875"
Director 2	75.000"	58.000"	63.125"
Director 3	161.000"	55.625"	60.750"
Director 4	174.000"	56.000"	61.125"
Director 5	140.000"	53.125"	58.375"
Compensator	12" behind Dir. 5	43.125"	37.500"

These 12-meter Yagi designs were optimized for > 20 dB F/R, and SWR < 2:1 over frequency range from 24.890 to 24.990 MHz, for heavy-duty elements (123 mph wind survival) and for medium-duty (85 mph wind survival). Only element tip dimensions are shown, and all dimensions are inches. See Fig 14D for element telescoping tubing schedule. Torque compensator element is made of 2.5" OD PVC water pipe placed 12" behind last director. Dimensions shown for compensators is one-half of total length, centered on boom.

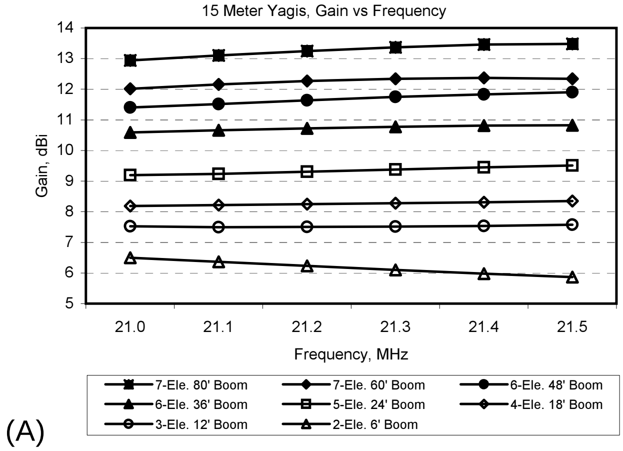

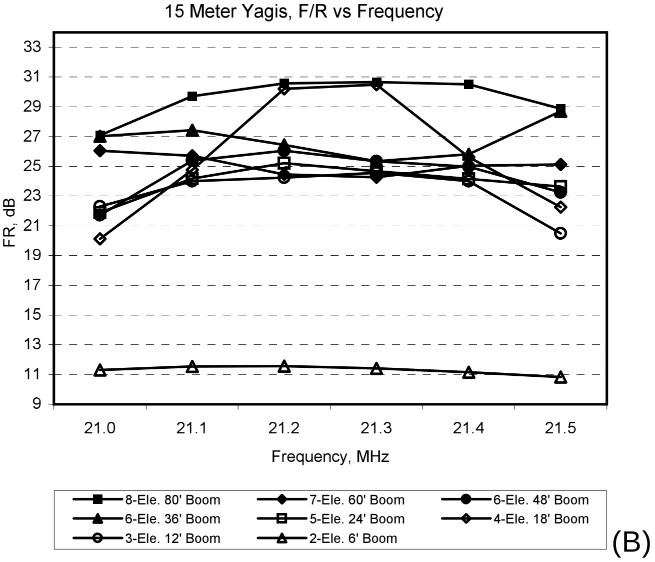

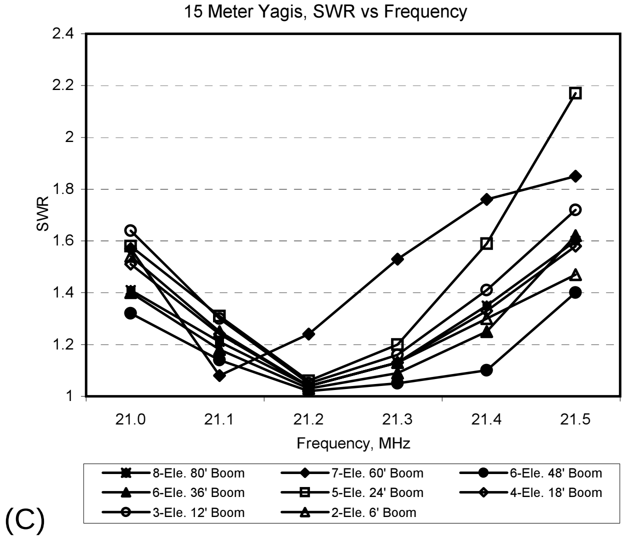

Fig 15—Gain, F/R and SWR performance versus frequency for optimized 15-meter Yagis. At A, gain versus frequency is shown for eight 15-meter Yagis whose booms range from 6 feet to 80 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the frequency range 21.0 to 21.45 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule for heavy-duty and for medium-duty 15-meter elements is shown. The heavy-duty elements can withstand 124-mph winds without icing, and 90-mph winds with $^1/_4$-inch radial ice. The medium-duty elements can survive 86-mph winds without icing, and 61-mph winds with $^1/_4$-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

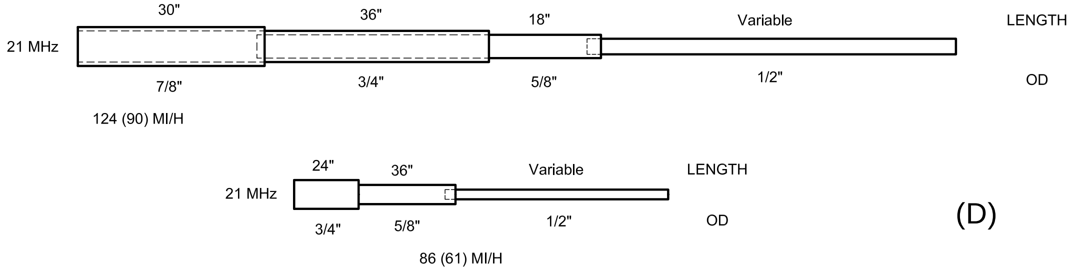

racy to $^1/_8$ inch is desirable.

The last row in each variable tip column shows the length of one-half of the "dummy element" torque compensator used to correct for uneven wind loading along the boom. This compensator is made from 2.5 inches OD PVC water pipe mounted to an element-to-boom plate like those used for each element. The compensator is mounted 12 inches behind the last director, the first director in the case of the 3-element 310-08H.YW antenna. Note that the heavy-duty elements require a correspondingly longer torque compensator than do the medium-duty elements.

12-METER YAGIS

Fig 14 describes the electrical performance of seven optimized 12-meter Yagis with boom lengths between 6 to 54 feet. The end of each boom includes 3 inches of space for the reflector and last director (or driven element) mounting plates. The narrow frequency width of the 12-meter band allows the performance to be optimized easily. Fig 14A shows the free-space gain versus frequency for each antenna; 14B shows the front-to-rear ratio, and 14C shows the SWR versus frequency. Each antenna with three or more elements was designed to

Table 3
Optimized 15-Meter Yagi Designs

Two-element 15-meter Yagi, 6 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		215-06H.YW	215-06M.YW
Reflector	0.000"	62.000"	85.000"
Driven Element	66.000"	51.000"	74.000"

Three-element 15-meter Yagi, 12 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		315-12H.YW	315-12M.YW
Reflector	0.000"	62.000"	84.250"
Driven Element	48.000"	51.000"	73.750"
Director 1	92.000"	43.500"	66.750"
Compensator	12" behind Dir. 1	34.750"	37.625"

Four-element 15-meter Yagi, 18 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		415-18H.YW	415-18M.YW
Reflector	0.000"	61.000"	83.500"
Driven Element	56.000"	51.500"	74.500"
Director 1	56.000"	48.000"	71.125"
Director 2	98.000"	36.625"	60.250"
Compensator	12" behind Dir. 2	20.875"	18.625"

Five-element 15-meter Yagi, 24 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		515-24H.YW	515-24M.YW
Reflector	0.000"	62.000"	84.375"
Driven Element	48.000"	52.375"	75.250"
Director 1	48.000"	47.875"	71.000"
Director 2	52.000"	47.000"	70.125"
Director 3	134.000"	41.000"	64.375"
Compensator	12" behind Dir. 3	40.250"	35.125"

Six-element 15-meter Yagi, 36 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		615-36H.YW	615-36M.YW
Reflector	0.000"	61.000"	83.375"
Driven Element	53.000"	52.000"	75.000"
Director 1	56.000"	49.125"	72.125"
Director 2	59.000"	45.125"	68.375"
Director 3	116.000"	47.875"	71.000"
Director 4	142.000"	42.000"	65.375"
Compensator	12" behind Dir. 4	45.500"	39.750"

Seven-element 15-meter Yagi, 48 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		615-48H.YW	615-48M.YW
Reflector	0.000"	62.000"	84.000"
Driven Element	48.000"	52.000"	75.000"
Director 1	48.000"	51.250"	74.125"
Director 2	125.000"	48.000"	71.125"
Director 3	190.000"	45.500"	68.750"
Director 4	161.000"	42.000"	65.375"
Compensator	12" behind Dir. 4	51.500"	45.375"

Seven-element 15-meter Yagi, 60 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		715-60H.YW	715-60M.YW
Reflector	0.000"	59.750"	82.250"
Driven Element	48.000"	52.000"	75.000"
Director 1	48.000"	52.000"	74.875"
Director 2	93.000"	49.500"	72.500"
Director 3	173.000"	44.125"	67.375"
Director 4	197.000"	45.500"	68.750"
Director 5	155.000"	41.750"	65.125"
Compensator	12" behind Dir. 5	58.500"	51.000"

Eight-element 15-meter Yagi, 80 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		815-80H.YW	815-80M.YW
Reflector	0.000"	62.000"	84.000"
Driven Element	56.000"	52.500"	75.500"
Director 1	48.000"	51.500"	74.375"
Director 2	115.000"	48.375"	71.500"
Director 3	164.000"	45.750"	69.000"
Director 4	202.000"	43.125"	66.500"
Director 5	206.000"	44.750"	68.000"
Director 6	163.000"	40.875"	64.250"
Compensator	12" behind Dir. 6	95.000"	83.375"

These 15-meter Yagi designs are optimized for > 20 dB F/R, and SWR < 2:1 over entire frequency range from 21.000 to 21.450 MHz, for heavy-duty elements (124 mph wind survival) and for medium-duty (86 mph wind survival). Only element tip dimensions are shown. See Fig 15D for element telescoping tubing schedule. All dimensions are in inches. Torque compensator element is made of 2.5" OD PVC water pipe placed 12" behind last director, and dimensions shown for compensators is one-half of total length, centered on boom.

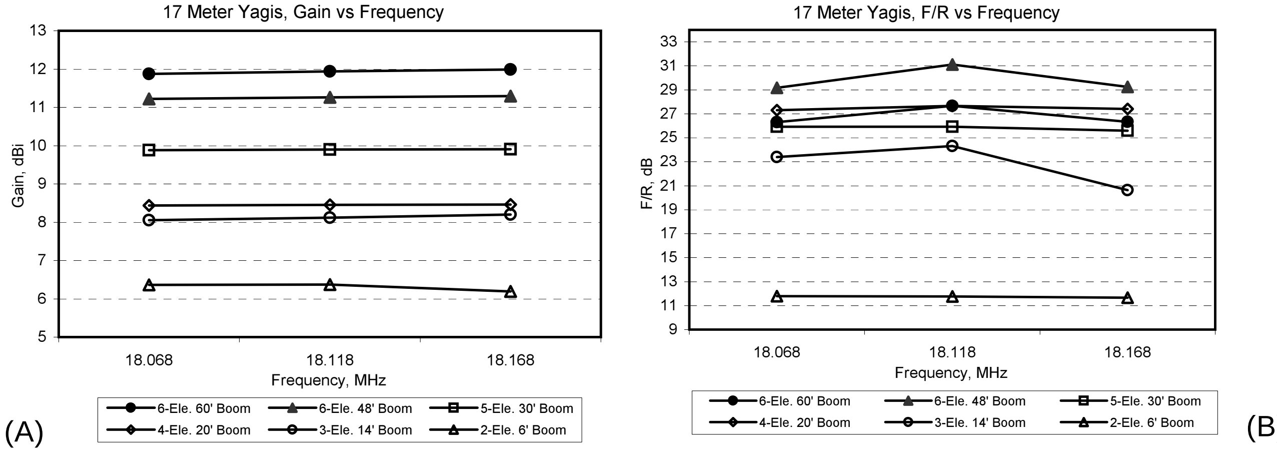

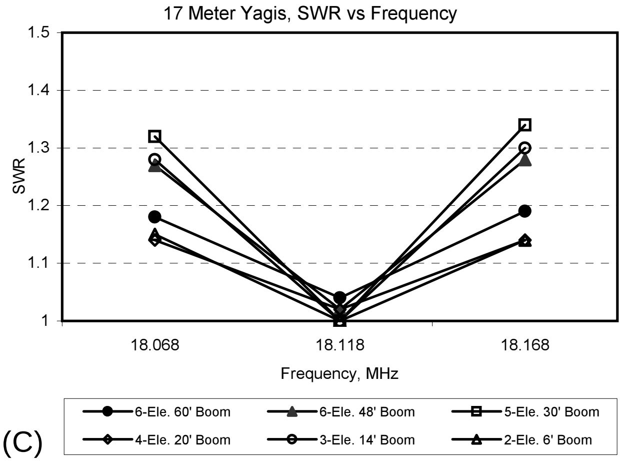

Fig 16—Gain, F/R and SWR performance versus frequency for optimized 17-meter Yagis. At A, gain versus frequency is shown for six 17-meter Yagis whose booms range from 6 feet to 60 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the narrow 17-meter band 18.068 to 18.168 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule for heavy-duty and for medium-duty 10-meter elements is shown. The heavy-duty elements can withstand 123-mph winds without icing, and 89-mph winds with 1/4-inch radial ice. The medium-duty elements can survive 83-mph winds without icing, and 59-mph winds with 1/4-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

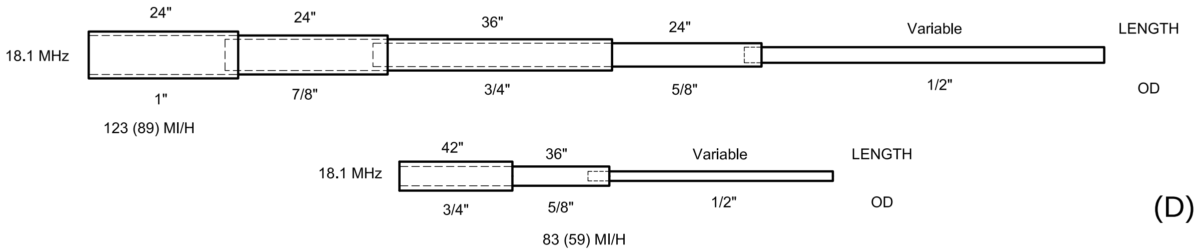

cover the narrow 12-meter band from 24.89 to 24.99 MHz, with SWR less than 2:1 and F/R better than 20 dB over that range.

Fig 14D shows the taper schedule for two types of 12-meter elements. The heavy-duty design can survive 123-mph winds with no icing, and 87-mph winds with 1/4 inch of radial ice. The medium-duty design can handle 85-mph winds with no icing, and 61-mph winds with 1/4 inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.375 inch thick flat aluminum plate, 5 inches wide by 6 inches long.

Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 2.945 inches for the heavy-duty element, and 2.857 inches for the medium-duty element. The equivalent length on each side of the boom is 3 inches. As usual, the torque compensator is mounted 12 inches behind the last director.

15-METER YAGIS

Fig 15 describes the electrical performance of eight optimized 15-meter Yagis with boom lengths between 6 feet to a spectacular 80 feet. The end of each boom includes 3 inches of space for the reflector and last-director (or driven element) mounting plates. Fig 15A shows the free-space gain versus frequency for each antenna; 15B shows the worst-case front-to-rear ratio, and 15C shows the SWR versus frequency. Each antenna with three or more elements was designed to cover the full 15-meter band from

Table 4

Optimized 17-meter Yagi Designs

Two-element 17-meter Yagi, 6 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		217-06H.YW	217-06M.YW
Reflector	0.000"	61.000"	89.000"
Driven Element	66.000"	48.000"	76.250"

Three-element 17-meter Yagi, 14 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		317-14H.YW	317-14M.YW
Reflector	0.000"	61.500"	91.500"
Driven Element	65.000"	52.000"	79.500"
Director 1	97.000"	46.000"	73.000"
	12" behind Dir. 1	12.625"	10.750"

Four-element 17-meter Yagi, 20 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		417-20H.YW	417-20M.YW
Reflector	0.000"	61.500"	89.500"
Driven Element	48.000"	54.250"	82.625"
Director 1	48.000"	52.625"	81.125"
Director 2	138.000"	40.500"	69.625"
Compensator	12" behind Dir. 2	42.500"	36.250"

Five-element 17-meter Yagi, 30 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		517-30H.YW	517-30M.YW
Reflector	0.000"	61.875"	89.875"
Driven Element	48.000"	52.250"	80.500"
Director 1	52.000"	49.625"	78.250"
Director 2	93.000"	49.875"	78.500"
Director 3	161.000"	43.500"	72.500"
Compensator	12" behind Dir. 3	54.375"	45.875"

Six-element 17-meter Yagi, 48 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		617-48H.YW	617-48M.YW
Reflector	0.000"	63.000"	90.250"
Driven Element	52.000"	52.500"	80.500"
Director 1	51.000"	45.500"	74.375"
Director 2	87.000"	47.875"	76.625"
Director 3	204.000"	47.000"	75.875"
Director 4	176.000"	42.000"	71.125"
Compensator	12" behind Dir. 4	68.250"	57.500"

Six-element 17-meter Yagi, 60 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		617-60H.YW	617-60M.YW
Reflector	0.000"	61.250"	89.250"
Driven Element	54.000"	54.750"	83.125"
Director 1	54.000"	52.250"	80.750"
Director 2	180.000"	46.000"	74.875"
Director 3	235.000"	44.625"	73.625"
Director 4	191.000"	41.500"	70.625"
Compensator	12" behind Dir. 4	62.875"	53.000"

These 17-m Yagi designs are optimized for > 20 dB F/R, and SWR < 2:1 over entire frequency range from18.068 to 18.168 MHz, for heavy-duty elements (123 mph wind survival) and for medium-duty (83 mph wind survival). Only element tip dimensions are shown. All dimensions are in inches. Torque compensator element is made of 2.5" OD PVC water pipe placed 12" behind last director, and dimensions shown for compensators is one-half of total length, centered on boom.

21.000 to 21.450 MHz, with SWR less than 2:1 and F/R ratio better than 20 dB over that range.

Fig 15D shows the taper schedule for two types of 15-meter elements. The heavy-duty design can survive 124-mph winds with no icing, and 90-mph winds with $^{1}/_{4}$ inch of radial ice. The medium-duty design can handle 86-mph winds with no icing, and 61-mph winds with $^{1}/_{4}$ inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.375-inch thick flat aluminum plate, 5 inches wide by 6 inches long.

Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 3.0362 inches for

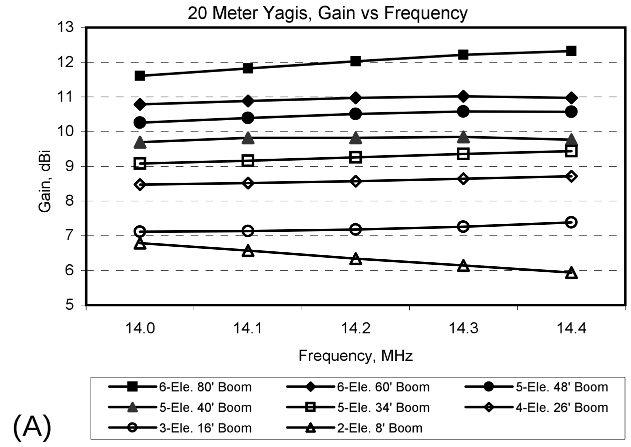

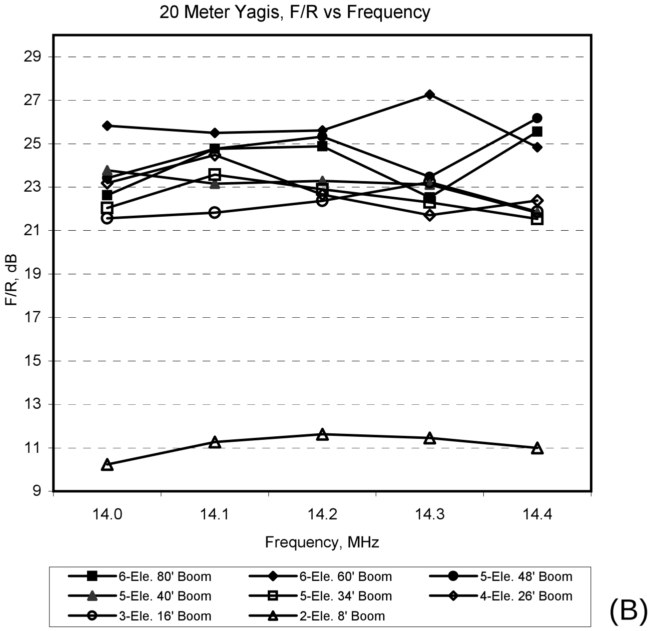

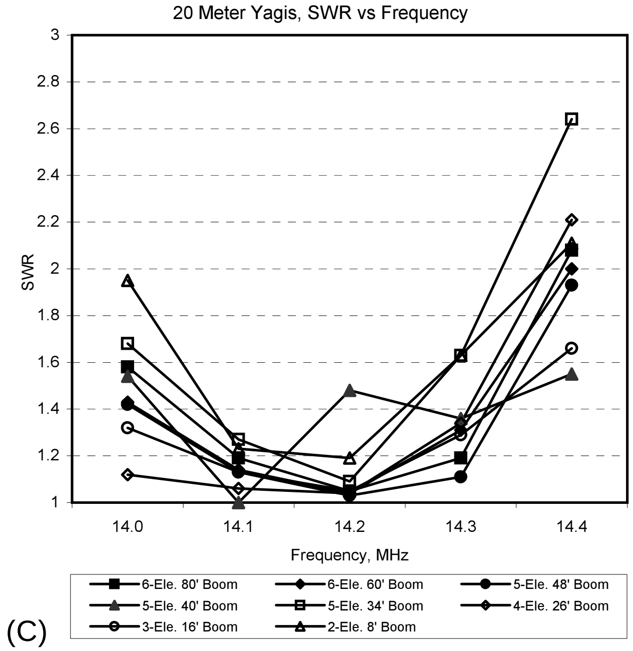

Fig 17—Gain, F/R and SWR performance versus frequency for optimized 20-meter Yagis. At A, gain versus frequency is shown for eight 20-meter Yagis whose booms range from 8 feet to 80 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the frequency range 14.0 to 14.35 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule for heavy-duty and for medium-duty 20-meter elements is shown. The heavy-duty elements can withstand 122-mph winds without icing, and 89-mph winds with $^1/_4$-inch radial ice. The medium-duty elements can survive 82-mph winds without icing, and 60-mph winds with $^1/_4$-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

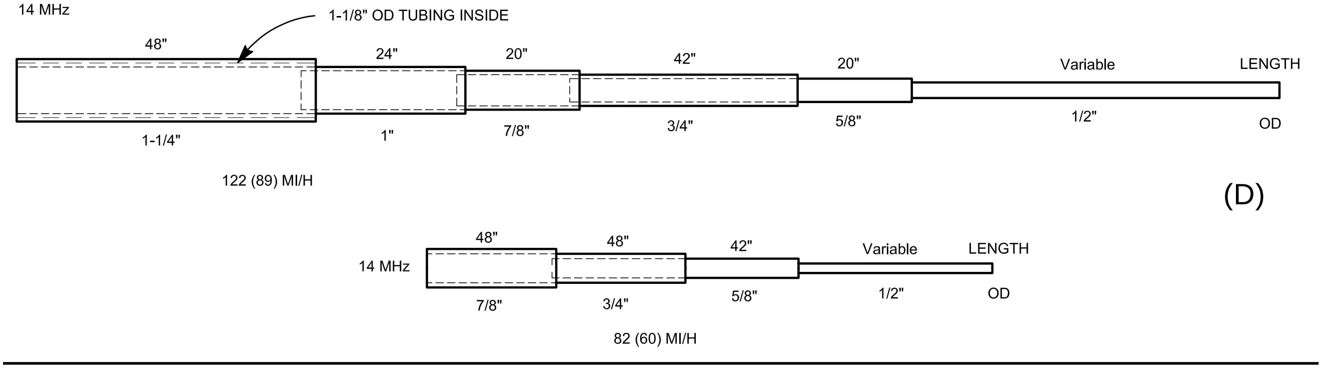

the heavy-duty element, and 2.9447 inches for the medium-duty element. The equivalent length on each side of the boom is 3 inches. As usual, the torque compensator is mounted 12 inches behind the last director.

17-METER YAGIS

Fig 16 describes the electrical performance of six optimized 17-meter Yagis with boom lengths between 6 to a heroic 60 feet. As usual, the end of each boom includes 3 inches of space for the reflector and last director (or driven element) mounting plates. Fig 16A shows the free-space gain versus frequency for each an-

Table 5

Optimized 20-Meter Yagi Designs

Two-element 20-meter Yagi, 8 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		220-08H.YW	220-08M.YW
Reflector	0.000"	66.000"	80.000"
Driven Element	90.000"	46.000"	59.000"

Three-element 20-meter Yagi, 16 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		320-16H.YW	320-16M.YW
Reflector	0.000"	69.625"	81.625"
Driven Element	80.000"	51.250"	64.500"
Director 1	106.000"	42.625"	56.375"
Compensator	12" behind Dir. 1	33.375"	38.250"

Four-element 20-meter Yagi, 26 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		420-26H.YW	420-26M.YW
Reflector	0.000"	65.625"	78.000"
Driven Element	72.000"	53.375"	65.375"
Director 1	60.000"	51.750"	63.875"
Director 2	174.000"	38.625"	51.500"
Compensator	12" behind Dir. 2	54.250"	44.250"

Five-element 20-meter Yagi, 34 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		520-34H.YW	520-34M.YW
Reflector	0.000"	68.625"	80.750"
Driven Element	72.000"	52.250"	65.500"
Director 1	71.000"	45.875"	59.375"
Director 2	68.000"	45.875"	59.375"
Director 3	191.000"	37.000"	51.000"
Compensator	12" behind Dir. 3	69.250"	56.250"

Five-element 20-meter Yagi, 40 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		520-40H.YW	520-40M.YW
Reflector	0.000"	68.375"	80.500"
Driven Element	72.000"	53.500"	66.625"
Director 1	72.000"	51.500"	64.625"
Director 2	139.000"	48.375"	61.750"
Director 3	191.000"	38.000"	52.000"
Compensator	12" behind Dir. 3	69.750"	56.750"

Five-element 20-meter Yagi, 48 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		520-48H.YW	520-48M.YW
Reflector	0.000"	66.250"	78.500"
Driven Element	72.000"	53.000"	66.000"
Director 1	88.000"	50.500"	63.750"
Director 2	199.000"	47.375"	60.875"
Director 3	211.000"	39.750"	53.625"
Compensator	12" behind Dir. 3	70.325"	57.325"

Six-element 20-meter Yagi, 60 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		620-60H.YW	620-60M.YW
Reflector	0.000"	67.000"	79.250"
Driven Element	84.000"	51.500"	65.000"
Director 1	91.000"	45.125"	58.750"
Director 2	130.000"	41.375"	55.125"
Director 3	210.000"	46.875"	60.375"
Director 4	199.000"	39.125"	53.000"
Compensator	12" behind Dir. 4	72.875"	59.250"

Six-element 20-meter Yagi, 80 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		620-80H.YW	620-80M.YW
Reflector	0.000"	66.125"	78.375"
Driven Element	72.000"	52.375"	65.500"
Director 1	122.000"	49.125"	62.500"
Director 2	229.000"	44.500"	58.125"
Director 3	291.000"	42.625"	56.375"
Director 4	240.000"	38.750"	52.625"
Compensator	12" behind Dir. 4	78.750"	64.125"

These 20-meter Yagi designs are optimized for > 20 dB F/R, and SWR < 2:1 over entire frequency range from 14.000 to 14.350 MHz, for heavy-duty elements (122 mph wind survival) and for medium-duty (82 mph wind survival). Only element tip dimensions are shown. See Fig 17 for element telescoping tubing schedule. All dimensions are in inches. Torque compensator element is made of 2.5" OD PVC water pipe placed 12" behind last director, and dimensions shown for compensators is one-half of total length, centered on boom.

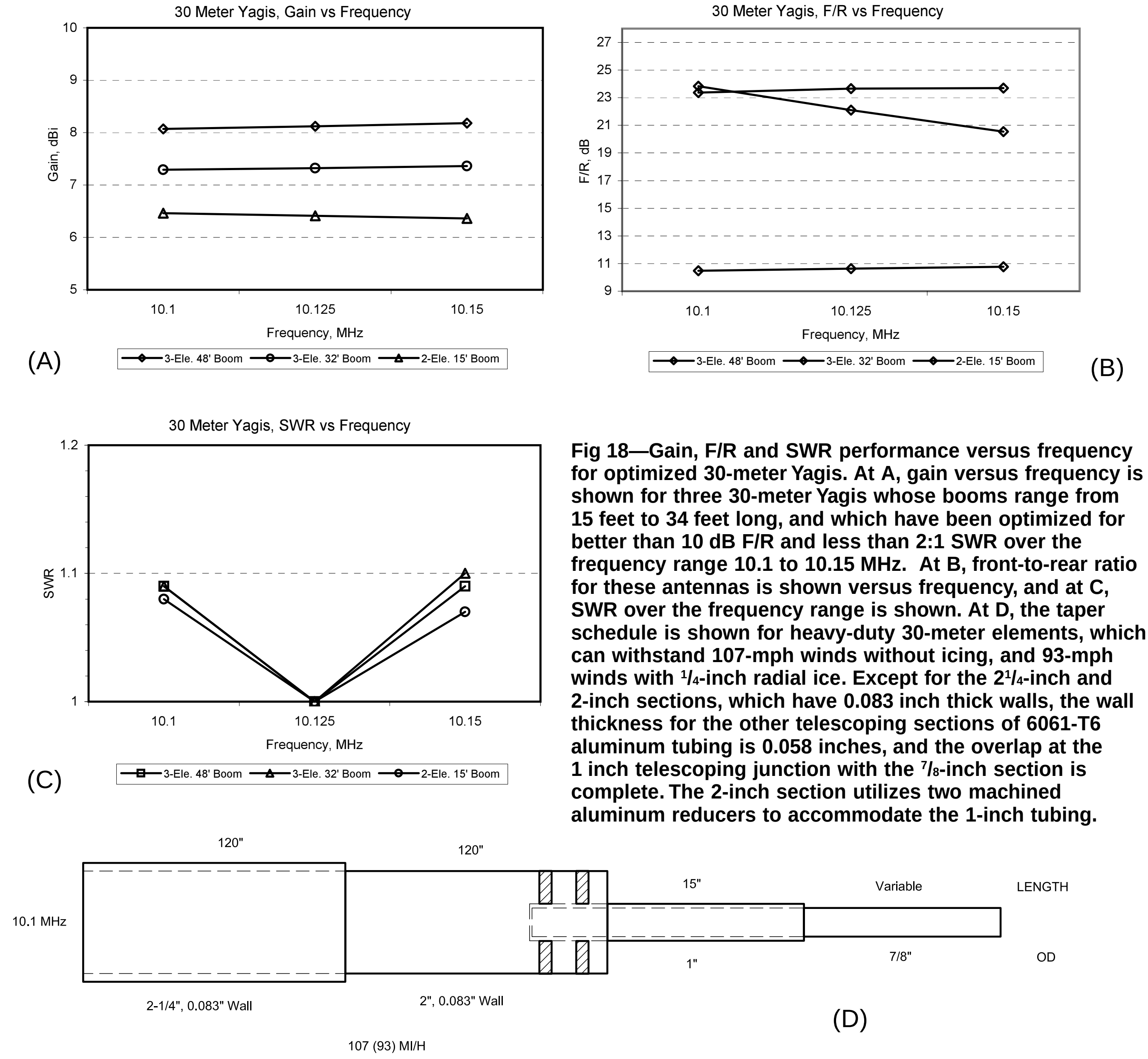

Fig 18—Gain, F/R and SWR performance versus frequency for optimized 30-meter Yagis. At A, gain versus frequency is shown for three 30-meter Yagis whose booms range from 15 feet to 34 feet long, and which have been optimized for better than 10 dB F/R and less than 2:1 SWR over the frequency range 10.1 to 10.15 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule is shown for heavy-duty 30-meter elements, which can withstand 107-mph winds without icing, and 93-mph winds with $^{1}/_{4}$-inch radial ice. Except for the 2$^{1}/_{4}$-inch and 2-inch sections, which have 0.083 inch thick walls, the wall thickness for the other telescoping sections of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at the 1 inch telescoping junction with the $^{7}/_{8}$-inch section is complete. The 2-inch section utilizes two machined aluminum reducers to accommodate the 1-inch tubing.

tenna; 16B shows the worst-case front-to-rear ratio, and 16C shows the SWR versus frequency. Each antenna with three or more elements was designed to cover the narrow 17-meter band from 18.068 to 18.168 MHz, with SWR less than 2:1 and F/R ratio better than 20 dB over that range.

Fig 16D shows the taper schedule for two types of 17-meter elements. The heavy-duty design can survive 123-mph winds with no icing, and 83-mph winds with $^{1}/_{4}$-inch of radial ice. The medium-duty design can handle 83-mph winds with no icing, and 59-mph winds with ¼ inch of radial ice.

The element-to-boom mounting plate for these Yagis is a 0.375-inch thick flat aluminum plate, 6 inches wide by 8 inches long. Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 3.5122 inches for the heavy-duty element, and 3.3299 inches for the medium-duty element. The equivalent length on each side of the boom is 4 inches. As usual, the torque compensator is mounted 12 inches behind the last director.

20-METER YAGIS

Fig 17 describes the electrical performance of eight optimized 20-meter Yagis with boom lengths between 8 to a giant 80 feet. As usual, the end of each boom includes 3 inches of space for the reflector and last director (driven element) mounting plates. Fig 17A shows the free-space gain versus frequency for each antenna; 17B shows the front-to-rear ratio, and 17C shows the SWR versus frequency. Each antenna with three or more ele-

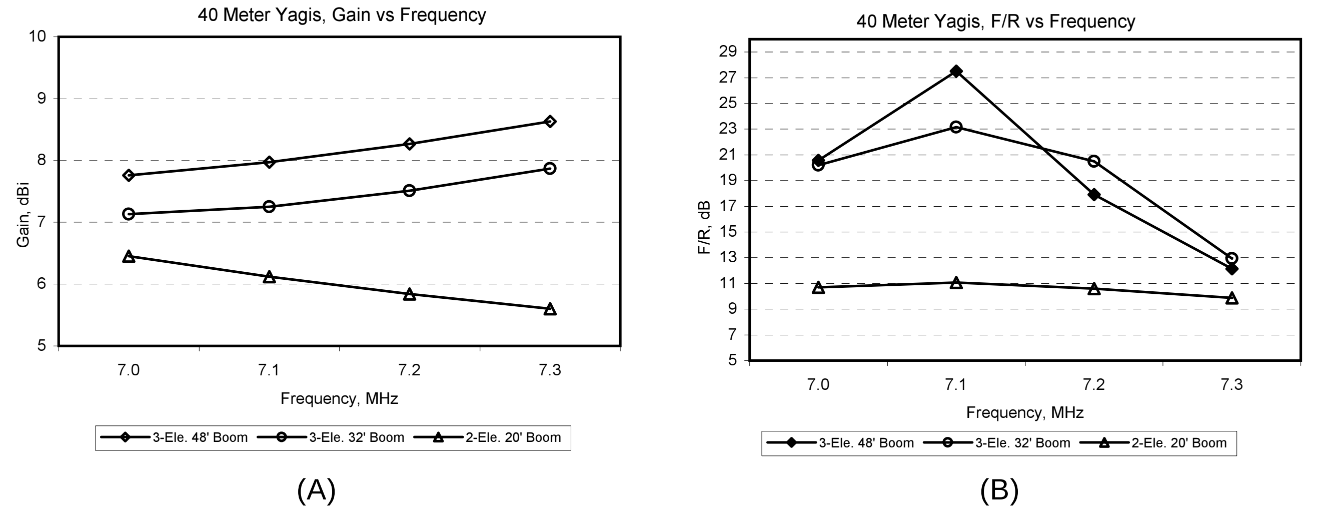

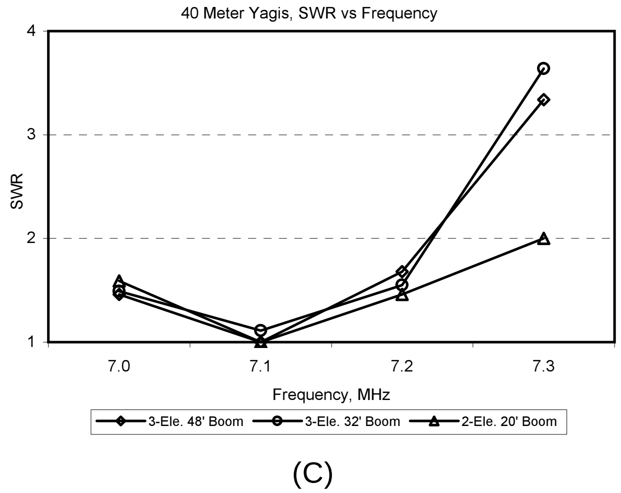

Fig 19—Gain, F/R and SWR performance versus frequency for optimized 40-meter Yagis. At A, gain versus frequency is shown for three 40-meter Yagis whose booms range from 20 feet to 48 feet long, and which have been optimized for better than 10 dB F/R and less than 2:1 SWR over the frequency range 7.0 to 7.2 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule is shown for heavy-duty 40-meter elements, which can withstand 107-mph winds without icing, and 93-mph winds with $^1/_4$-inch radial ice. Except for the $2^1/_4$-inch and 2-inch sections, which have 0.083 inch thick walls, the wall thickness for the other telescoping sections of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at the end telescoping junction is 3 inches. The 2-inch section utilizes two machined aluminum reducers to accommodate the 1-inch tubing.

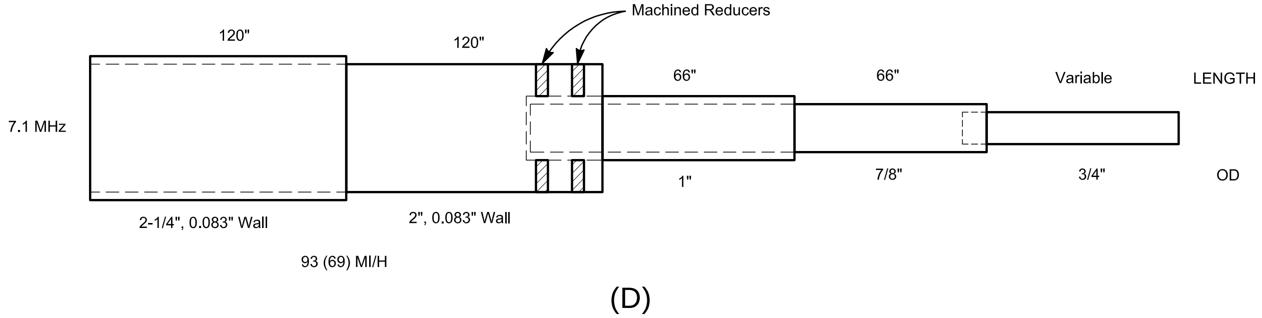

ments was designed to cover the complete 20-meter band from 14.000 to 14.350 MHz, with SWR less than 2:1 and F/R ratio better than 20 dB over that range.

Fig 17D shows the taper schedule for two types of 20-meter elements. The heavy-duty design can survive 122-mph winds with no icing, and 89-mph winds with $^1/_4$ inch of radial ice. The medium-duty design can handle 82-mph winds with no icing, and 60-mph winds with $^1/_4$ inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.375-inch thick flat aluminum plate, 6 inches wide by 8 inches long. Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 3.7063 inches for the heavy-duty element, and 3.4194 inches for the medium-duty element. The equivalent length on each side of the boom is 4 inches. As usual, the torque compensator is mounted 12 inches behind the last director.

30-METER YAGIS

Fig 18 describes the electrical performance of three

Table 6

Optimized 30-Meter Yagi Designs

Two-element 30-meter Yagi, 15 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		230-15H.YW
Reflector	0.000"	50.250"
Driven Element	174.000"	14.875"

3-element 30-meter Yagi, 22 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		330-22H.YW
Reflector	0.000	59.375
Driven Element	135.000	35.000
Director 1	123.000	19.625

Three-element 30-meter Yagi, 34 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		330-34H.YW
Reflector	0.000"	53.750"
Driven Element	212"	29.000"
Director 1	190"	14.500"

These 30-m Yagi designs are optimized for > 10 dB F/R, and SWR < 2:1 over entire frequency range from 10.100 to 10.150 MHz for heavy-duty elements (105 mph wind survival). Only element tip dimensions are shown. See Fig 18D for element telescoping tubing schedule. All dimensions are in inches. No torque compensator element is required.

Table 7

Optimized 40-Meter Yagi Designs

Two-element 40-meter Yagi, 20 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		240-20H.YW
Reflector	0.000"	85.000"
Driven Element	234.000"	35.000"

Three-element 40-meter Yagi, 32 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		340-32H.YW
Reflector	0.000"	90.750"
Driven Element	196.000"	55.875"
Director 1	182.000"	33.875"

Three-element 40-meter Yagi, 48 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		340-48H.YW
Reflector	0.000"	81.000"
Driven Element	300.000"	45.000"
Director 1	270.000"	21.000"

These 40-m Yagi designs are optimized for > 10 dB F/R, and SWR < 2:1 over low-end of frequency range from 7.000 to 7.200 MHz, for heavy-duty elements (95 mph wind survival). Only element tip dimensions are shown. See Fig 19D for element telescoping tubing schedule. All dimensions are in inches. No wind torque compensator is required.

optimized 30-meter Yagis with boom lengths between 15 to 34 feet. Because of the size and weight of the elements alone for Yagis on this band, only 2-element and 3-element designs are described. The front-to-rear ratio requirement for the 2-element antenna is relaxed to be greater than 10 dB over the band from 10.100 to 10.150 MHz, while that for the 3-element designs is kept at greater than 20 dB over that frequency range.

As usual, the end of each boom includes 3 inches of space for the reflector and last director mounting plates. Fig 18A shows the free-space gain versus frequency for each antenna; 18B shows the worst-case front-to-rear ratio, and 18C shows the SWR versus frequency.

Fig 18D shows the taper schedule for the 30-meter elements. Note that the wall thickness of the first two sections of tubing is 0.083 inches, rather than 0.058 inches. This heavy-duty element design can survive 107-mph winds with no icing, and 93-mph winds with $^{1}/_{4}$ inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.500-inch thick flat aluminum plate, 6 inches wide by 24 inches long. Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 4.684 inches. The equivalent length on each side of the boom is 12 inches. These designs require no torque compensator.

40-METER YAGIS

Fig 19 describes the electrical performance of three optimized 40-meter Yagis with boom lengths between 20 to 48 feet. Like the 30-meter antennas, because of the size and weight of the elements for a 40-meter Yagi, only 2-element and 3-element designs are described. The front-to-rear ratio requirement for the 2-element antenna is relaxed to be greater than 10 dB over the band from 7.000 to 7.300 MHz, while the goal for the 3-element designs is 20 dB over the frequency range of 7.000 to 7.200 MHz. It is exceedingly difficult to hold the F/R greater than 20 dB over the entire 40-meter band without sacrificing excessive gain with a 3-element design.

As usual, the end of each boom includes 3 inches of space for the reflector and last director mounting plates. Fig 19A shows the free-space gain versus frequency for each antenna; 19B shows the front-to- rear ratio, and 19C shows the SWR versus frequency.

Fig 19D shows the taper schedule for the 40-meter elements. Note that the wall thickness of the first two sections of tubing is 0.083 inches, rather than 0.058 inches. This element design can survive 93-mph winds with no icing, and 69-mph winds with $^{1}/_{4}$ inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.500-inch thick flat aluminum plate, 6 inches wide by 24 inches long. Electrically each mounting plate is equivalent to a cylinder, with an effective diameter of 4.684 inches. The equivalent length on each side of the boom is 12 inches. These designs require no torque compensator.

Modifying Monoband Hy-Gain Yagis

Enterprising amateurs have long used the Telex Communications Hy-Gain "Long John" series of HF monobanders as a source of top-quality aluminum and hardware for customized Yagis. Often-modified older models include the 105BA for 10 meters, the 155BA for 15 meters, and the 204BA and 205BA for 20 meters. Newer Hy-Gain designs, the 105CA, 155CA and 205CA, have been redesigned by computer for better performance.

Hy-Gain antennas have historically had an excellent reputation for superior mechanical design, and Hy-Gain proudly points out that many of their monobanders are still working after more than 30 years. In the older designs the elements were purposely spaced along the boom to achieve good weight balance at the mast-to-boom bracket, with electrical performance as a secondary goal. Thus, the electrical performance was not necessarily optimum, particularly over an entire amateur band. Newer Hy-Gain designs are electrically superior to the older ones, but because of their strong concern for weight-balance are still not optimal by the definitions used in this chapter.

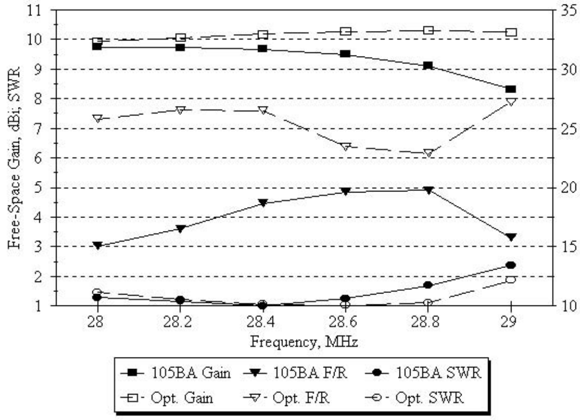

Fig 20—Gain, F/R and SWR over the 28.0 to 28.8 MHz range for original and optimized Yagis using Hy-Gain hardware. Original 105BA design provided excellent weight balance at boom-to-mast bracket, but compromised the electrical performance somewhat because of non-optimum spacing of elements. Optimized design requires wind torque-balancing compensator element, and compensating weight at director end of boom to rebalance weight. The F/R ratio over the frequency range for the optimized design is more than 23 dB. Each element uses the original Hy-Gain taper schedule and element-to-boom clamp, but the length of the tip is changed per Table 8.

Table 8
Optimized Hy-Gain 20-Meter Yagi Designs

Optimized 204BA, Four-element 20-meter Yagi, 26 foot boom

Element	*Spacing*	*Element Tip*
File Name		BV204CA.YW
Reflector	0.000"	56.000"
Driven Element	85.000"	52.000"
Director 1	72.000"	61.500"
Director 2	149.000"	50.125"

Optimized 205CA, Five-element 20-meter Yagi, 34 foot boom

Element	*Spacing*	*Element Tip*
File Name		BV205CA.YW
Reflector	0.000"	62.625"
Driven Element	72.000"	53.500"
Director 1	72.000"	63.875"
Director 2	74.000"	61.625"
Director 3	190.000"	55.000"

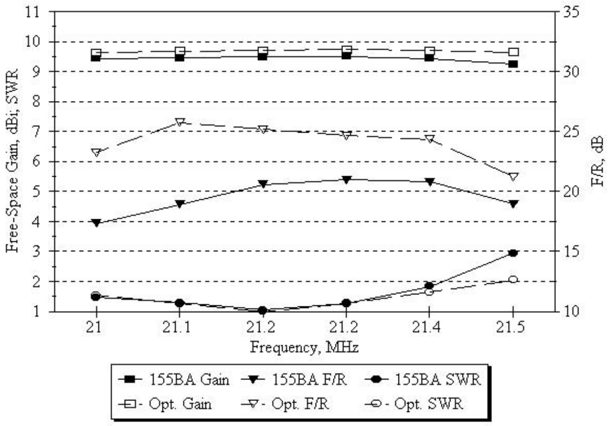

Fig 21—Gain, F/R and SWR over the 21.0 to 21.45 MHz band for original and optimized Yagis using Hy-Gain hardware. Original 155BA design provided excellent weight balance at boom-to-mast bracket, but compromised the electrical performance somewhat because of non-optimum spacing of elements. Optimized design requires wind torque-balancing compensator element, and compensating weight at director end of boom to rebalance weight. The F/R ratio over the frequency range for the optimized design is more than 22 dB. Each element uses the original Hy-Gain taper schedule and element-to-boom clamp, but the length of the tip is changed per Table 9.

Table 9
Optimized Hy-Gain 15-Meter Yagi Designs

Optimized 155BA, Five-element 15-meter Yagi, 24 foot boom

Element	*Spacing*	*Element Tip*
File Name		BV155CA.YW
Reflector	0.000"	64.000"
Driven Element	48.000"	65.500"
Director 1	48.000"	63.875"
Director 2	82.750"	61.625"
Director 3	127.250"	55.000"

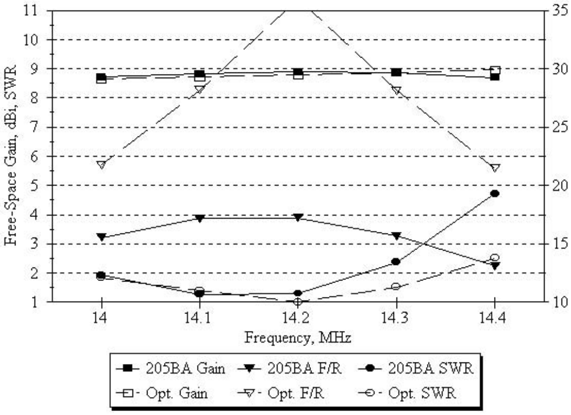

Fig 22—Gain, F/R and SWR over the 14.0 to 14.35 MHz band for original and optimized Yagis using Hy-Gain hardware. Original 205BA design provided good weight balance at boom-to-mast bracket, but compromised the electrical performance because of non-optimum spacing of elements. Optimized design requires wind torque-balancing compensator element, and compensating weight at director end of boom to rebalance weight. The F/R ratio over the frequency range for the optimized design is more than 23 dB, while the original design never went beyond 17 dB of F/R. Each element uses the original Hy-Gain taper schedule and element-to-boom clamp, but the length of the tip is changed per Table 10.

Table 10

Optimized Hy-Gain 10-Meter Yagi Designs

Optimized 105BA, Five-element 10-meter Yagi, 24 foot boom

Element	*Spacing, inches*	*Element Tip*
File Name		BV105CA.YW
Reflector	0.000"	44.250"
Driven Element	40.000"	53.625"
Director 1	40.000"	52.500"
Director 2	89.500"	50.500"
Director 3	112.250"	44.750"

With the addition of wind torque-compensation dummy elements, and with extra lead weights, where necessary, at the director end of the boom for weight-balance, the electrical performance can be enhanced, using the same proven mechanical parts.

Fig 20 shows the computed gain, F/R ratio and SWR for a 24-foot boom, 10-meter optimized Yagi (modified 105BA) using Hy-Gain hardware. **Fig 21** shows the same for a 26-foot boom 15-meter Yagi (modified 155BA), and **Fig 22** shows the same for a 34-foot boom (modified 205BA) 20-meter Yagi. **Tables 8** through **10** show dimensions for these designs. The original Hy-Gain taper schedule is used for each element. Only the length of the end tip (and the spacing along the boom) is changed for each element.

Multiband Yagis

So far, this chapter has discussed monoband Yagis—that is, Yagis designed for a single Amateur-Radio frequency band. Because hams have operating privileges on more than one band, multiband coverage has always been very desirable.

INTERLACING ELEMENTS

In the late 1940s, some experimenters tried interlacing Yagi elements for different frequencies on a single boom, mainly to cover the 10 and 20-meter bands (at that time the 15-meter band wasn't yet available to hams). The experimenters discovered, to their considerable chagrin, that the mutual interactions between different elements tuned to different frequencies are very difficult to handle.

Adjusting a lower-frequency element usually results in interaction with higher-frequency elements near it. In effect, the lower-frequency element acts like a retrograde reflector, throwing off the effectiveness of the higher-frequency directors nearby. Element lengths and the spacing between elements can be changed to improve performance of the higher-frequency Yagi, but the resulting compromise is rarely equal to that of an optimized monoband Yagi. A reasonable compromise for portable operation may be found in Chapter 15, Portable Antennas, by VE7CA.

TRAPPED MULTIBANDERS

Multiband Yagis using a single boom can also be made using traps. Traps allow an element to have multiple resonances. See Chapter 7, Multiband Antennas, for details on trap designs. Commercial vendors have sold trapped antennas to hams since the 1950s and surveys show that after simple wire dipoles and multiband verticals, trapped triband Yagis are the most popular antennas in the Amateur Radio service.

The originator of the trapped tribander was Chester Buchanan, W3DZZ, in his Mar 1955 *QST* article, "The Multimatch Antenna System." On 10 meters this rather unusual tribander used two reflectors (one dedicated and one with traps) and two directors (one dedicated and one with traps). On 20 and 15 meters three of the five elements were active using traps. The W3DZZ tribander employed 12 traps overall, made with heavy wire and concentric tubular capacitors to hold down losses in the traps. Each trap was individually fine tuned after construction before mounting it on an element.

Another example of a homemade tribander was the 26-foot boom 7-element 20/15/10-meter design described by Bob Myers, W1XT (ex-W1FBY) in Dec 1970 *QST*. The W1FBY tribander used only two sets of traps in the driven element, with dedicated reflectors and directors for each frequency band. Again, the traps were quite robust in this design to minimize trap losses, using $^7/_{16}$-inch aluminum tubing for the coils and short pieces of RG-8 coax as high-voltage tuning capacitors.

Only a relatively few hams actually built tribanders for themselves, mainly because of the mechanical complexity and the close tolerances required for such antennas. The traps themselves must be constructed quite accurately for reproducible results, and they must be carefully weatherproofed for long life in rain, snow, and often polluted or corrosive atmospheres.

Christmas Tree Stacks

Another possible method for achieving multiband coverage using monoband Yagis is to stack them in a "Christmas tree" arrangement. See **Fig 23**. For an installation covering 20, 15 and 10 meters, you could mount on the rotating mast just at the top of the tower the 20-meter monobander. Then perhaps 9 feet above that you would mount the 15-meter monobander, followed by the 10-meter monoband Yagi 7 feet further up on the mast. Another configuration would be to place the 10-meter Yagi in between the lower 20-meter and upper 15-meter Yagis. Whatever the arrangement, the antenna in the middle of such a Christmas-tree always suffers the most interaction from the lowest-frequency Yagi.

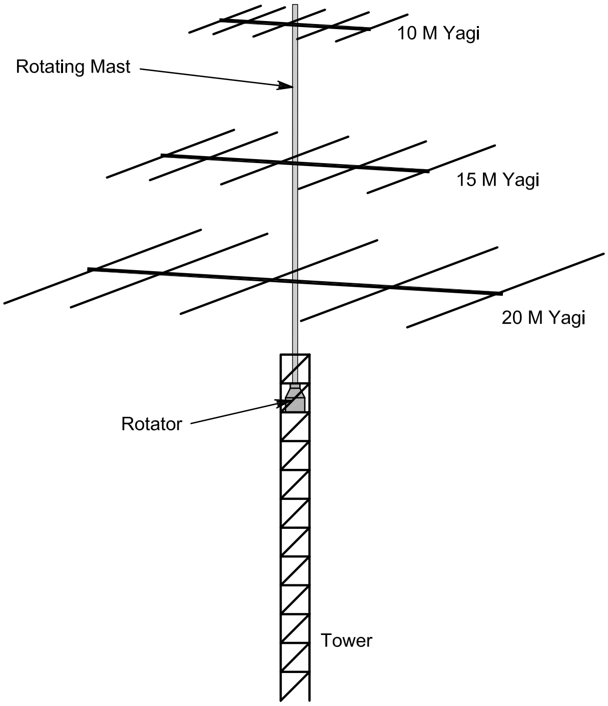

Fig 23—"Christmas Tree" stack of 20/15/10-meter Yagis spaced vertically on a single rotating mast.

Dave Leeson, W6NL (ex-W6QHS), mentions that the 10-meter Yagi in his closely stacked Christmas Tree (15 meters at the top, 10 meters in the middle, and 20 meters at the bottom of the rotating mast) loses "substantial gain" because of serious interaction with the 20-meter antenna. (N6BV and K1VR calculated that the free-space gain in the W6NL stack drops to 5 dBi, compared to about 9 dBi with no surrounding antennas.) Monobanders are *definitely not* universally superior to tribanders in multiband installations. In private conversations, W6NL has indicated that he would not repeat this kind of short Christmas Tree installation again.

Forward Staggering

Some hams have built multiband Yagis on a common boom, using a technique called *forward staggering*. This means that that most (or all) of the higher-frequency elements are placed in front of any lower-frequency elements—in other words, most of the elements are not interlaced. Richard Fenwick, K5RR, described his triband Yagi design in Sep 1996 *QEX* magazine. This uses forward-stagger and open-sleeve design techniques and was optimized using several sophisticated modeling programs.

Fenwick's tribander used a 57-foot, 3-inch OD boom to hold 4 elements on 20 meters, 4 elements on 15 meters and 5 elements on 10 meters. **Fig 24** shows the element placement for the K5RR tribander. Most hams, of course, don't have the real-estate or the large rotator needed to turn such a large, but elegant solution to the interaction problem!

Force 12 C3 "Multi-Monoband" Triband Yagi

Antenna manufacturer Force 12 also uses forward-stagger layouts and patented combinations of open- and closed-sleeve drive techniques extensively in their product line of multiband antennas, which they call "multi-monoband Yagis." **Fig 25** shows the layout for the popular Force 12 C3 triband Yagi. The C3 uses no traps, thereby avoiding any losses due to traps. The C3 consists of three 2-element Yagis on an 18-foot boom, using full-sized elements designed to withstand high winds.

The C3 feed system employs open-sleeves, where the 20-meter driver element is fed with coax through a common-mode current balun and parasitically couples to the closely spaced 15-meter driver and the two 10-meter drivers to yield a feed-point impedances close to 50 Ω on all three bands. See the section on open-sleeve dipoles in Chapter 7, Multiband Antennas.

Note the use of the forward-stagger technique in the C3, especially on 10 meters. To reduce interaction with

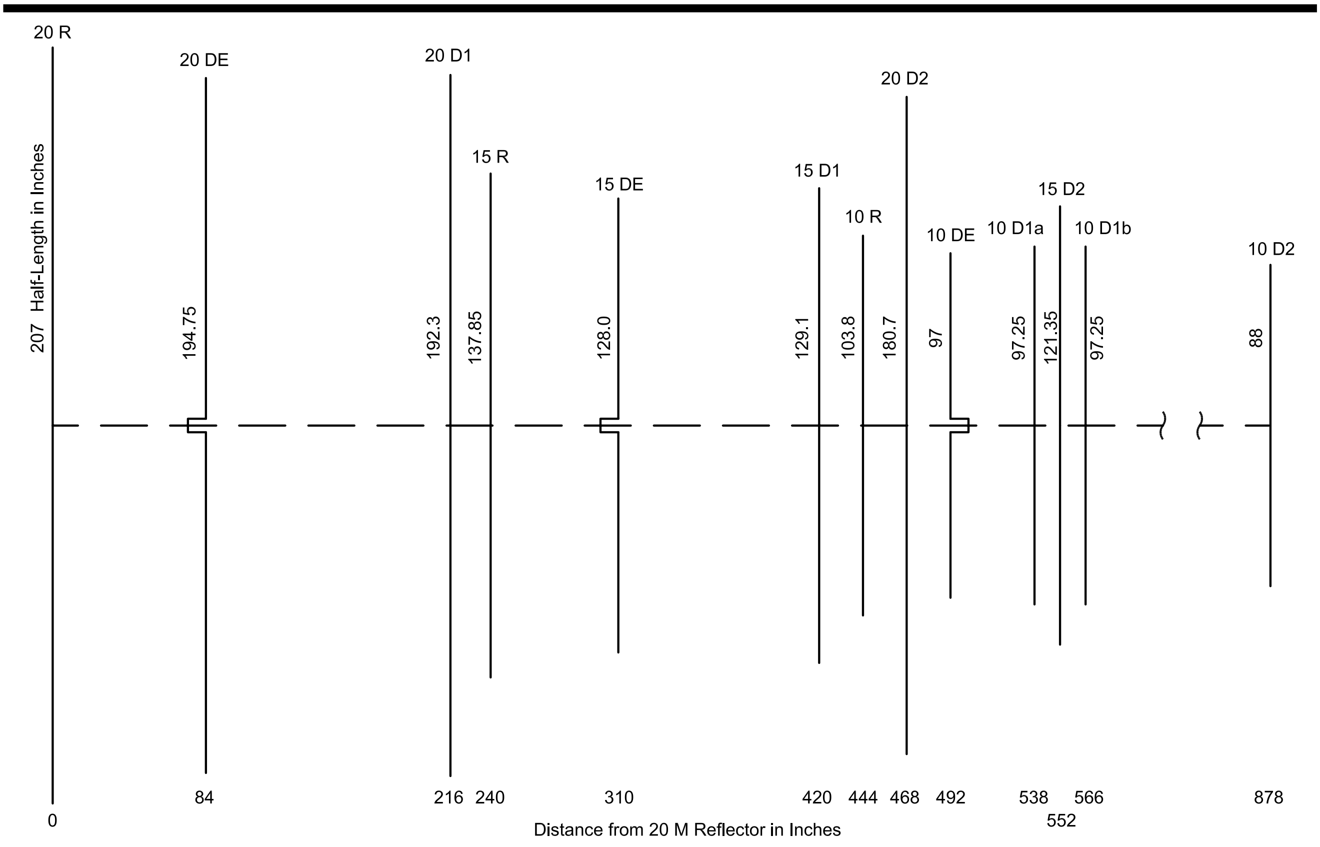
20 R
20 DE
20 D1
20 D2
15 R
15 DE
15 D1
10 R
15 D2
10 DE
10 D1a
10 D1b
10 D2
207 Half-Length in Inches
194.75
192.3
137.85
128.0
129.1
103.8
180.7
97
97.25
121.35
97.25
88
0
84
216
240
310
420
444
468
492
538
566
552
878
Distance from 20 M Reflector in Inches

Fig 24—Dimensions of K5RR's trapless tribander using "forward stagger" and open-sleeve techniques to manage interaction between elements for different frequencies.

the lower-frequency elements behind it, the 10-meter portion of the C3 is mounted on the boom ahead of all the lower-frequency elements, with the main 10-meter parasitic element (#7) acting as a director. The lower-frequency elements behind the 10-meter section act as retrograde reflectors, gaining some improvement of the gain and pattern compared to a monoband 2-element Yagi. A simplified *EZNEC* model of the C3 is included on the CD-ROM accompanying this book.

On 15 meters, the main parasitic element (#2) is a dedicated reflector, but the other elements ahead on the boom act like retrograde directors to improve the gain and pattern somewhat over a typical 2-element Yagi with a reflector. On 20 meters, the C3 is a 2-element Yagi with a dedicated reflector (#1) at the back end of the boom.

The exact implementation of any Yagi, of course, depends on the way the elements are constructed using telescoping aluminum tubing. The C3 type of design is no exception.

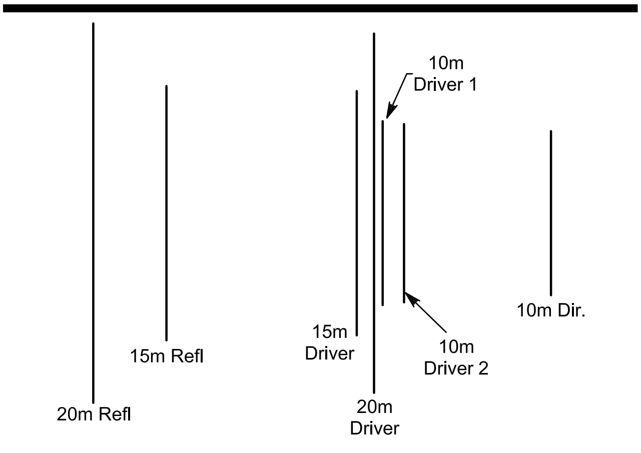
10m
Driver 1
10m Dir.
15m Refl
15m
Driver
10m
Driver 2
20m Refl
20m
Driver

Fig 25—Layout of Force 12 C3 multiband Yagi. Note that the 10-meter (driver/director) portion of the antenna is "forward staggered" ahead of the 15-meter (reflector/driver) portion, which in turn is placed ahead of the 20-meter (reflector/driver) portion. The antenna is fed at the 20-meter driver, which couples parasitically to the 15-meter driver and the two 10-meter drivers.

Stacked Yagis

Monoband parasitic arrays are commonly stacked either in broadside or collinear fashion to produce additional directivity and gain. In HF amateur work, the most common broadside stack is a vertical stack of identical Yagis on a single tower. This arrangement is commonly called a *vertical stack*. At VHF and UHF, amateurs often employ collinear stacks, where identical Yagis are stacked side-by-side at the same height. This arrangement is called a *horizontal stack*, and is not usually found at HF, because of the severe mechanical difficulties involved with large, rotatable side-by-side arrays.

Fig 26 illustrates the two different stacking arrangements. In either case, the individual Yagis making up the stack are generally fed in phase. There are times, however, when individual antennas in a stacked array are purposely fed out of phase in order to emphasize a particular elevation pattern. See Chapter 17, Repeater Antenna Systems, for such a case where elevation pattern steering is implemented for a repeater station.

Let's look at the reasons hams stack Yagis:

- For more gain
- For a wider elevation footprint in a target geographical area
- For azimuthal diversity—two or more directions at once
- For less fading
- For less precipitation static

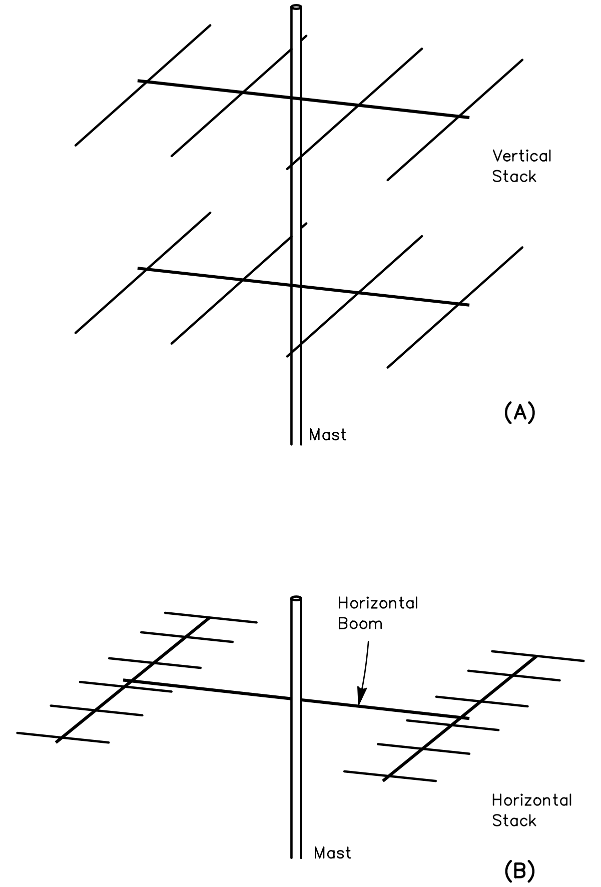

Fig 26—Stacking arrangements. At A, two Yagis are stacked vertically (broadside) on the same mast. At B, two Yagis are stacked horizontally (collinear) side-by-side. At HF the vertical stack is more common because of mechanical difficulties involved with large HF antennas stacked side-by-side, whereas at VHF and UHF the horizontal stack is common.

STACKS AND GAIN

Fig 27 compares the elevation responses for three antenna systems of 4-element 15-meter Yagis. The response for the single Yagi at a height of 120 feet peaks at an elevation of about 5°, with a second peak at 17° and a third at 29°. When operated by itself, the 60-foot high Yagi has its first peak at about 11° and its second peak beyond 34°.

The basic principle of a vertically stacked HF array is that it takes energy from higher-angle lobes and concentrates that energy into the main elevation lobe. The main lobe of the 120/60-foot stack peaks about 7° and is about 2 dB stronger than either the 60- or 120-foot antenna by itself. The shape of the left-hand side of the stack's main lobe is determined mainly by the 120-foot antenna's response. The right-hand side of the stack's main lobe is "stretched" rightwards (toward higher angles) mainly by the 60-foot Yagi, while the shape follows the curve of the 120-foot Yagi.

Look at the second and third lobes of the stack, which appear about 18° and 27°. These are about 14 dB down from the stack's peak gain, showing that energy has indeed been extracted from them. By contrast, look at the levels of the second and third lobes for the individual Yagis at 60 and 120 feet. These higher-angle lobes are almost as strong as the first lobes.

The stack squeezes higher-angle energy into its main elevation lobe, while maintaining the frontal lobe azimuth pattern of a single Yagi. This is the reason why many state-of-the-art contest stations are stacking arrays of relatively short-boom antennas, rather than stacking long-boom, higher-gain Yagis. A long-boom HF Yagi narrows the azimuthal pattern (and the elevation pattern too), making pointing the antenna more critical and making it more difficult to spread a signal over a wide azimuthal area, such as all of Europe and Asiatic Russia at one time.

STACKS AND WIDE ELEVATION FOOTPRINTS

Detailed studies using sophisticated computer models of the ionosphere have revealed that coverage

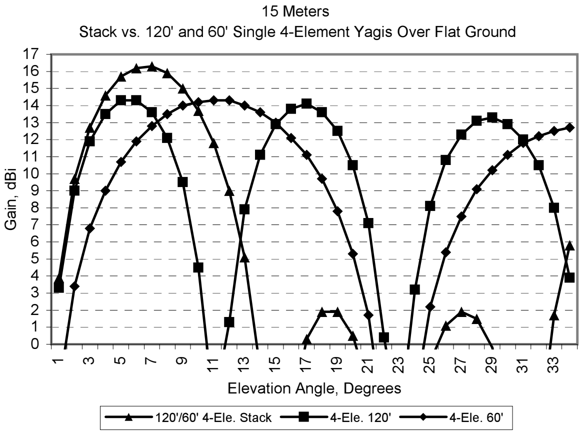

Fig 27—Comparison of elevation patterns on 15 meters for a stack of 4-element Yagis at 120 and 60 feet and individual Yagis at those two heights. The shape of the stack's response is determined mainly by that of the top antenna.

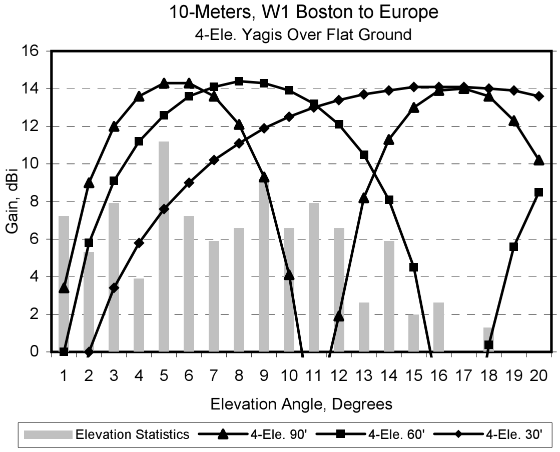

Fig 28—Comparison of elevation patterns and elevation-angle statistics for individual 10-meter TH7DX tribanders mounted over flat ground aiming from New England to Europe. No single antenna can cover the wide range of angles needed—from 1° to 18°.

of a wide range of elevation angles is necessary to ensure consistent DX or contest coverage on the HF bands. These studies have been conducted over all phases of the 11-year solar cycle, and for numerous transmitting and receiving QTHs throughout the world.

Chapter 23, Radio Wave Propagation, covers these studies in more detail, and the CD-ROM accompanying this book contains a huge number of elevation-angle statistical tables for locations all around the world. The *HFTA* (HF Terrain Assessment) program on the CD-ROM can not only compute antenna elevation patterns over irregular local terrain, but it can compare them directly to the elevation-angle statistics for a particular target geographic area.

A 10-Meter Example

Fig 28 shows the 10-meter elevation-angle statistics for the New England path from Boston, Massachusetts, to all of the continent of Europe. The statistics are overlaid with the computed elevation response for three individual 4-element Yagis, at three heights: 90, 60 and 30 feet above flat ground. In terms of wavelength, these heights are 2.60 λ, 1.73 λ and 0.86 λ high.

You can see that the 90-foot high Yagi covers the lower elevation angles best, but it has a large null in its response centered at about 11°. This null puts a big hole in the coverage for some 22% of all the times the 10-meter band is open to Europe. At those angles where the 90-foot Yagi exhibits a null, the 60-foot Yagi would be effective, and so would the 30-foot Yagi. If that is the only antenna you have, the 90-foot high Yagi would be too high for good coverage of Europe from New England.

The peak statistical elevation angle into Europe is 5°, and this occurs about 11% of all the times the 10-meter band is open to Europe from Boston. At an elevation of 5° the 30-foot high Yagi would be down almost 7 dB compared to the 90-foot high Yagi, but at 11° the 90-foot Yagi would be more than 22 dB down from the 30-foot Yagi. There is no single height at which one Yagi can optimally cover all the necessary elevation angles, especially to a large geographic area such as Europe—although the 60-foot high antenna is arguably the best compromise for a single height. To cover all the possibilities to Europe, however, you need a 10-meter antenna system that can cover equally well the entire range of elevation angles from 1° to 18°.

Fig 29 compares elevation-angle statistics for two 10-meter paths from New England to Europe and to Japan. The elevation angles needed for communications with the Far East are very low. Overlaid on Fig 29 for comparison are the elevation responses over flat ground for three different antenna systems, using identical 4-element Yagis:

- Three Yagis, stacked at 90, 60 and 30 feet
- Two Yagis, stacked at 70 and 40 feet
- One Yagi at 90 feet.

The best coverage of all the necessary angles on 10 meters to Europe is with the stack of three Yagis at 90/60/30 feet. The two-Yagi stack at 70 and 40 feet comes in a close second to Europe, and for elevation angles higher than about 9° the 70/40-foot stack is actually superior to the 90/60/30-foot stack.

Both of the stacks illustrated here give a wider *elevation footprint* than any single antenna, so that all the angles can be covered automatically without having to switch from higher to lower antennas manually. This is perhaps the major benefit of using stacks, but not the only

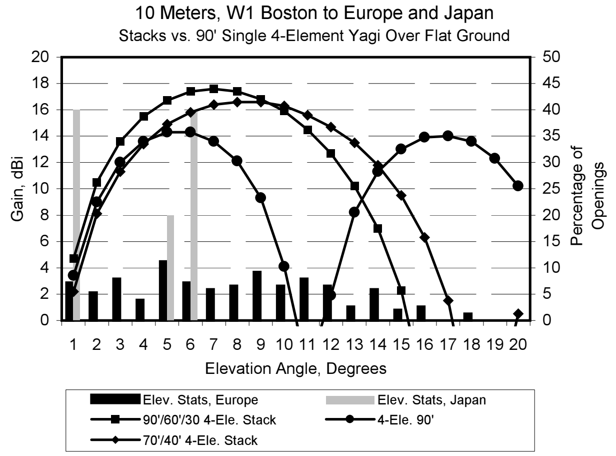

Fig 29—Combinations of 4-element Yagis over flat ground. The elevation-angle statistics into Japan from New England (Boston) are represented by the black vertical bars, while the grey vertical bars represent the elevation-angle statistics to Europe. The 90/60/30-foot stack has the best *elevation footprint* into Japan, although the 70/40-foot stack performs well also.

one, as we'll see.

To Japan, the necessary range of elevation angles is considerably smaller than that needed to a larger geographic target area like Europe. The 90/60/30-foot stack is still best on the basis of having higher gain at low angles, although the two-Yagi stack at 70 and 40 feet is a good choice too. Note that the single 90-foot high Yagi's performance is very close to the 70/40-foot stack of two Yagis at low angles, but the two-Yagi stack is superior to the single 90-foot antenna for angles higher than about 5° on 10 meters.

A 15-Meter Example

The situation is similar on 15 meters from New England to Europe. On 15 meters, the range of angles needed to fully cover Europe is 1° to 28°. This large range of angles makes covering all the angles even more challenging. Ken Wolff, K1EA, a devoted contest operator and the author of the famous *CT* contest logging program, put it very clearly when he wrote in the bulletin for the Yankee Clipper Contest Club:

"Suppose you have 15-meter Yagis at 120 feet and 60 feet, but can feed only one at a time. A 15-meter beam at 120 feet has its first maximum at roughly 5° and the first minimum at 10°. The Yagi at 60 feet has a maximum at 10° and a minimum at 2°. At daybreak, the band is just opening, signals are arriving at 3° or less and the high Yagi outperforms the low one by 5-10 dB. Late in the morning, western Europeans are arriving at angles of 10° or more, while UA6 is still arriving at 4-5°. Western Europe can be 20-30 dB louder on the low antenna than the high! What to do? Stack em!"

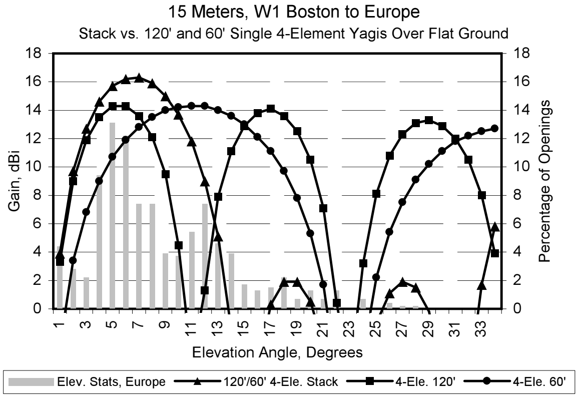

Fig 30—Comparison of elevation patterns for K1EA's illustration about 15-meter Yagis mounted over flat ground, with elevation-angle statistics to Europe added. The stack at 120 and 60 feet yields a better footprint over the range of 3° to 11° at its half-power points, better than either antenna by itself.

Fig 30 illustrates K1EA's scenario, showing the elevation statistics to Europe from Massachusetts and the elevation responses for a 120- and a 60-foot high, 4-element Yagi, both over flat ground, together with the response for both antennas operated as vertical stack. The half-power beamwidth of the stack's main lobe is 6.9°, while that for the 120-foot antenna by itself is 5.5° and that for the 60-foot antenna by itself is 11.1°. The half-power beamwidth numbers by themselves can be deceiving, mainly because the stack starts out with a higher gain. A more meaningful observation is that the stack has equal to or more gain than either of the two individual antennas from 1° to about 10°.

Is such a stack of 15-meter Yagis at 120 and 60 feet optimal for the New England to Europe path? No, it isn't, as we'll explore later, but the stack is clearly better than either antenna by itself for the scenario K1EA outlined above.

A 20-Meter Example

Take a look now at **Fig 31**, which overlays elevation-angle statistics for Europe (gray vertical bars) and Japan (black vertical bars) from Boston on 20 meters, plus the elevation responses for four different sets of antennas mounted over flat ground. Just for emphasis, the highest antenna is a 200-foot high 4-Element Yagi. It is clearly too high for complete coverage of all the needed angles into Europe. A number of New England operators have verified that this is true—a really high Yagi will open the 20-meter band to Europe in the morning and may shut it down in the afternoon, but during the middle of the day the high antenna gets soundly beaten by lower antennas.

To Japan, however, from New England the range of angles needed narrows considerably on 20 meters,

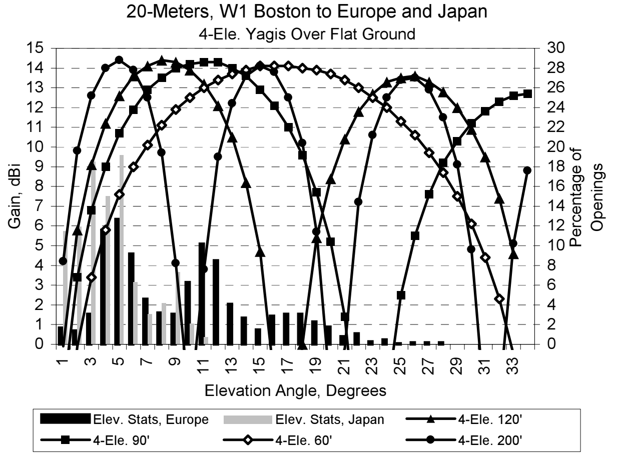

Fig 31—Comparison of elevation patterns for individual 20-meter Yagis over flat ground, compared with the range of elevation angles needed on this band from New England to Europe (gray bars) and to Japan (black bars). For fun, the response of a 200-foot high Yagi is included—this antenna is far too high to cover the needed range of angles to Europe because of its deep nulls at critical angles, like 10°. But the 200 footer is great into Japan!

from 1° to only 11°. For these angles, the 200-foot Yagi is the best antenna to work Japan from New England on 20 meters.

This is true provided that the antenna is aiming out over flat ground. The actual, generally irregular, terrain in various directions can profoundly modify the takeoff angles favored by an antenna system, particularly on steep hills. There will be more discussion on this important topic later on.

SPARE ME THE NULLS!

Now, let's look closely at some other 20-meter antennas in Fig 28, the ones at 120 and 60 feet. At an elevation angle of 8° the difference in elevation response between the 60- and 120-foot high Yagis is just over 3 dB. Can you really notice a change of 3 dB on the air? Signals on the HF bands often rise and fall quickly due to fading, so differences of 2 or 3 dB are difficult to discern. Consequently, the difference between a Yagi at 120 feet and one at 60 feet may be difficult to detect at elevation angles covered well by both antennas. But a *deep null* in the elevation response is very noticeable.

Back in 1990, when editor Dean Straw, N6BV, put up his 120-foot tower in Windham, New Hampshire, his first operational antenna was a 5-element triband Yagi, with 3 elements on 40 and 4 elements on both 20 and 15 meters. Just as the sun was going down on a late August day Straw finished connecting the feed line in the shack. The antenna seemed to be playing like it should, with a good SWR curve and a good pattern when it was rotated. So N6BV/1 called a nearby friend, John Dorr, K1AR, on the telephone and asked him to get on the air to make some signal comparisons on 20 meters into Europe.

Straw was shocked that every European they worked that evening said his signal was several S units weaker than K1AR's. Dorr was using a 4-element 20-meter monobander at 90 feet, which at first glance should have been comparable to Straw's 4-element antenna at 120 feet. But N6BV really shouldn't have been so shocked—in New England, the elevation angles from Europe late in the day on 20 meters are almost always higher than 11°, and that is true for the entire solar cycle.

The N6BV/1 station was located on a small hill, while K1AR was located on flat terrain towards Europe. The elevation response for N6BV/1's 120-foot high Yagi fell right into a deep null at 11°. This was later confirmed many times in the following eight years that the N6BV/1 station was operational. During the early morning opening on 20 meters into Europe, the top antenna was always very close to or equal to the stack of three TH7DX tribanders at 90/60/30 feet on the same tower. But in the afternoon the top antenna was *always* decidedly worse than the stack, so much so that Straw often wondered whether something had gone wrong with the top antenna!

So what's the moral to this short tale? It's simple: *The gain you can achieve, while useful, is not so important as the deep nulls you can avoid by using a stack.*

STACKING DISTANCES BETWEEN YAGIS

So far, we've examined stacks as a means of achieving more gain over an individual Yagi, while also matching the antenna system's response to the range of elevation angles needed for particular propagation paths. Most importantly, we seek to avoid nulls in the elevation response. Earlier we asked whether a 120/60-foot stack was optimal for the path from New England to Europe on 15 meters. Let's examine how the stacking distance between individual antennas affects the performance of a stack.

Fig 32 shows overlays of various combinations of 15-meter Yagis. Just for reference, a plot for a single 60-foot high Yagi is also included. Let's start by looking at the most widely spaced stack in the group: the 120/30-foot stack. Here, the spacing is obviously too large, since the second lobe is actually stronger than the first lobe. In terms of wavelength, the 90-foot spacing between antennas in this stack is 1.94 λ, a large spacing indeed.

There is a great deal of folklore and superstition among amateurs about stacking distances for HF arrays. For years, high-performance stacked Yagi arrays have been used for weak-signal DXing on the VHF and UHF bands. The most extreme example of weak-signal work is EME work (Earth-Moon-Earth, also called *moonbounce*) because of the huge path losses incurred on the way to and from the Moon. The most successful arrays used for

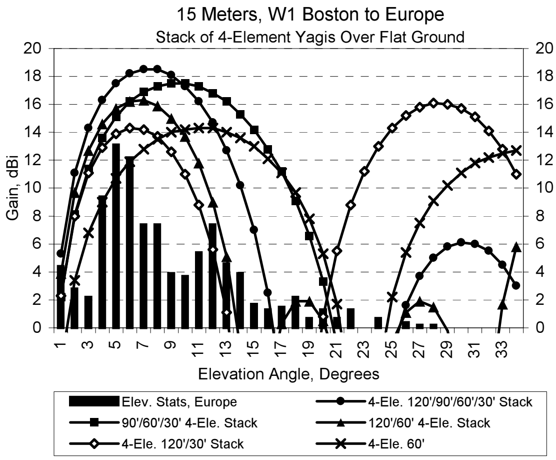

Fig 32—Various stacks towards Europe from New England for 15-meters. The stack at 120 and 30 feet is clearly suboptimal, since the second lobe is higher than the first lobe. The 120/60-foot stack is better in this regard, but is still not as good a performer as the 90/60/30-foot stack. It's debatable whether going to four Yagis in the 120/90/60/30-foot stack is a good idea because it drops below the performance of the 90/60/30-foot stack at about 10° in elevation. The exact distance between practical HF Yagis is not critical to obtain the benefits of stacking. For a stack of tribanders at 90, 60 and 30 feet, the distance in wavelengths between individual antennas is 0.87 λ at 28.5 MHz, 0.65 λ at 21.2 MHz, and 0.43 λ at 14.2 MHz.

moonbounce have low sidelobe levels and very narrow frontal lobes that give huge amounts of gain. The low sidelobes help minimize received noise, since the receive levels for signals that do manage to bounce off the Moon and return to Earth are exceedingly weak.

But HF work is different from moonbounce in that rigorously trying to minimize high-angle lobes is far less crucial at HF, where we've already shown that the main goal is to achieve gain over a wide elevation-plane footprint without any disastrous nulls in the pattern. The gain gradually increases as spacing in terms of wavelength is increased between individual Yagis in a stack, and then decreases slowly once the spacing is greater than about 1.0 λ. The difference in gain between spacings of 0.5 λ to 1.0 λ for a stack of typical HF Yagis amounts to only a fraction of a decibel. Stacking distances on the order of 0.6 λ to 0.75 λ give best gain commensurate with good patterns.

While the stack at 120/60 feet in Fig 32 doesn't have the second-lobe-stronger problem the 120/30-foot stack has, 60 feet between antennas is 1.29 λ, again outside the normal range of HF stack spacings. As a consequence, the 120/60-foot stack doesn't cover the range of elevation angles as well as it could, and is inferior to both the 90/60/30-foot stack and the 120/90/60/30-foot stack. The 120/60-foot two-Yagi stack needs at least one more antenna placed in-between to spread out the elevation-range coverage and to provide more gain.

It could be debated, but the 90/60/30-foot stack seems optimal for coverage of all the angles into Europe from New England on 15 meters. Note that the 30-foot spacing between Yagis is 0.65 λ on 21.2 MHz, right in the middle of the range of typical stack spacings.

Switching Out Yagis in the Stack

Still, the extra gain that is available at low elevation angles from a 120/90/60/30-foot high, four-Yagi stack in Fig 32 is alluring. For those statistically possible, but less likely, occasions when the elevation angle is higher than about 12°, it would be advantageous to switch out the top 120-foot Yagi and operate with only the lower three Yagis in a stack. (This also allows the top antenna to be rotated in another direction, an aspect we'll explore later.) There are even times when the incoming angles are really high and when the top two antennas might be switched out to create a 60/30-foot stack. Later in this chapter we'll explore flexible circuitry for such stack switching.

Stacking Distance and Lobes at HF

Let's look a little more closely at how a stack achieves gain and a wide elevation footprint. **Fig 33** shows a rectangular X-Y graph of the elevation response from 0° to 180° for two 3-element 15-meter Yagis (with 12-foot booms) spaced 30 feet apart (0.65 λ at 21.2 MHz), but mounted at two different heights: 95/65 and 85/55 feet. The rectangular plot gives more resolution than is possible on a polar plot. Note that the heights shown represent typical stacking heights on 15 meters—there's nothing magic about these choices. The free-space H-Plane pattern for the 30-foot spaced stack is also shown for reference.

The worst-case overhead elevation lobe, which ranges from about 60° to 120° in elevation (±30° from straight overhead at 90°), is about 14.7 dB down for the 95/65-foot stack. The overhead lobe peaks broadly at an elevation angle of about 82°. The overhead lobe for the lower 85/55-foot stack occurs at an elevation of about 64°, where it is 19 dB down.

The F/B for both 3-element sets of heights is about 15 dB, well down from the excellent 32 dB F/B for each Yagi by itself. The degradation of F/B is mainly due to mutual coupling to its neighbor in the stack.

The ground-reflection pattern in effect "modulates" the free-space pattern of the individual Yagi, but in a complex and not always intuitive manner. This is quite evident for the 85/55-foot stack at near-overhead angles. In this region things become complicated indeed, because the fourth and fifth lobes due to ground reflections are interacting with the free-space pattern of the stack.

Because the spacing remains constant at 30 feet for

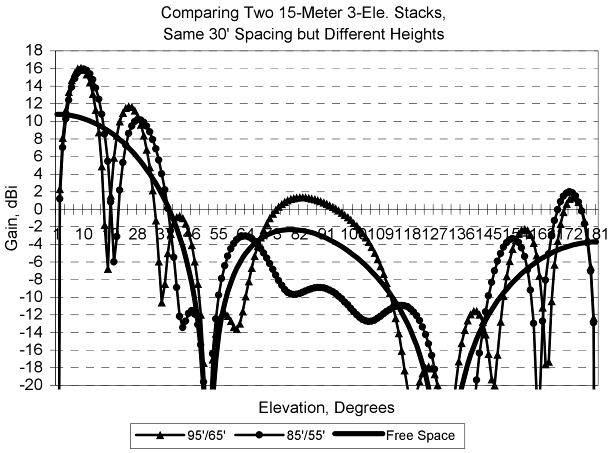

Fig 33—Rectangular plot comparing two 15-meter stacks of 3-element Yagis—each antenna is spaced 30 feet from its partner, but at different heights. The lobes are a complicated function of the antenna height, not the spacing, since that remains constant.

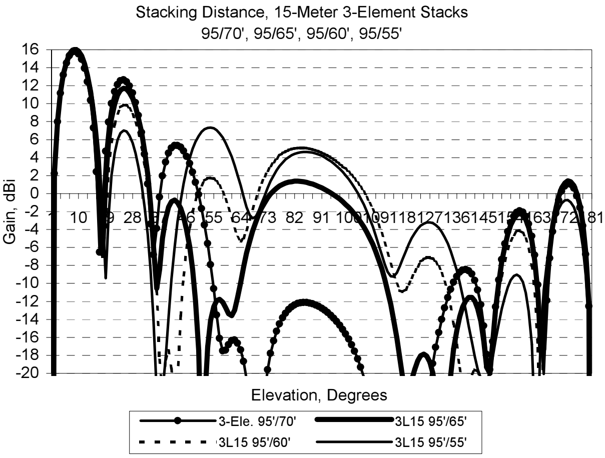

Fig 34—Four spacing scenarios for two 3-element 15-meter Yagis. Things get very complicated. The optimal spacing in terms of stacking gain is 30 feet, which is 0.65 λ. The near-overhead lobes turn out to be ugly looking, but unimportant for skywave propagation.

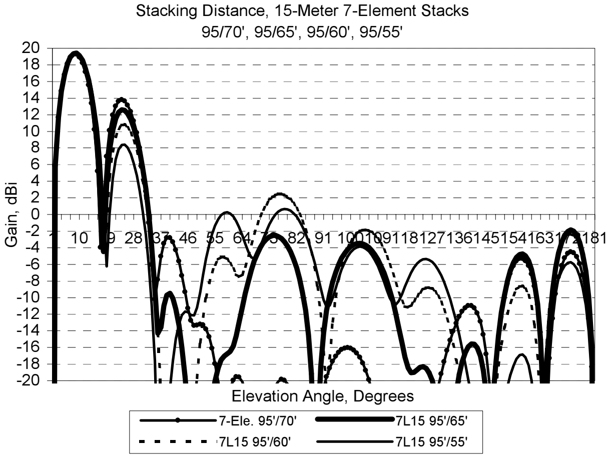

Fig 35—Four spacing scenarios for two large 7-element 15-meter Yagis (on 64-foot booms). Again, a 0.65 λ spacing (30 feet) provides the most stacking gain.

these pairs of antennas, however, the main determinant for the upper-elevation angle lobes is the distance of the horizontally polarized antennas above the ground, not the spacing between them.

Changing the Stack Spacing

Fig 34 demonstrates just how complicated things get for four different spacing scenarios. Here, the lower Yagi in the stack is moved down in 5-foot increments from the 95/70 feet level, to 95/65, 95/60 and 95/55 feet. The closest spacing, 25 feet in the 95/70-foot stack, yields nominally the "cleanest" pattern in the overhead region from 60° to 120°. The worst-case overhead lobe for the 95/70-foot stack is down 28 dB from peak. The F/B is again about 15 dB.

The worst case overhead lobe for the widest spacing, 40 feet in the 95/55-foot stack, is about 11 dB down from peak. The F/B has increased marginally, but is still only about 16 dB. It is difficult to pinpoint directly whether the spacing or the height above ground is the major determinant for the various lobe amplitudes for the 3-element stack. We'll soon look closely at whether the overhead lobe is important or not for HF work.

Longer Boom Length and Stack Spacing

Fig 35 shows the same type overlay of elevation plots, but this time for two 7-element 15-meter Yagis on gigantic 64-foot booms. These Yagis are also spaced 30 feet apart (0.65 λ at 21.2 MHz), mounted at the same four sets of heights in Fig 34. As you'd expect, the free-space elevation pattern for a stacked pair of 7-element Yagis on 64-foot booms is narrower than that for a stacked pair of 3-element Yagis on 12-foot booms. The intrinsic F/B of the longer Yagi is also better than the F/B of the shorter antenna. As a result, all lobes beyond the main lobe of the stacked 7-element pair are lower for both sets of heights than their 3-element counterparts. The worst-case overhead lobe for the 7-element 95/65-foot pair is about 22 dB down at 76° and the F/B at 172° is greater than 21 dB for all four sets of heights.

Table 11 summarizes the main performance characteristics for four sets of stacked Yagis. The first entry for each boom length is for the Yagi by itself at a height of 95 feet. Stacked configurations are next listed in order of gain. The column labeled "Worst lobe, dB re Peak" is the

Table 11
Example, Spacing Between 15–Meter Yagis

Antenna	*Peak Gain dBi*	*Worst Lobe dB re Peak*	*Worst Lobe Angle,°*	*F/B dB*	*Overhead Lobe dB re Peak*
3–Ele., 12' boom					
By itself 95'	13.2	–0.9	21	28.8	–17.5
95'/65' (Δ 30')	16.08	–4.5	25	14.9	–14.7
95'/60' (Δ 35')	16.01	–6.2	24	15.1	–10.9
95'/70' (Δ 25')	15.81	–3.2	24	14.8	–28
95'/55' (Δ 40')	15.71	–8.7	24	16.4	–11
95'/75' (Δ 20')	15.34	–2.3	23	16.3	–17.2
4–Ele., 18' boom					
By itself 95'	13.92	–1	21	28.3	–20.4
95'/65' (Δ 30')	16.63	–4.5	23	18.5	–17.3
95'/60' (Δ 35')	16.6	–6.2	24	18.2	–13.1
95'/55' (Δ 40')	16.36	–8.7	24	19.8	–13.2
95'/70' (Δ 25')	16.36	–3.3	24	20.4	–31.8
95'/75' (Δ 20')	15.92	–2.5	23	25.9	–19
5–Ele., 23' boom					
By itself 95'	14.26	–1.1	21	27.9	–22.3
95'/65' (Δ 30')	16.86	–4.6	24	20.8	–19
95'/60' (Δ 35')	16.86	–6.3	24	20.7	–14.4
95'/55' (Δ 40')	16.67	–8.8	24	23.5	–14.4
95'/70' (Δ 25')	16.59	–3.4	24	24.9	–34.4
95'/75' (Δ 20')	16.18	–2.6	23	34.3	–20.2
7–Ele., 64' boom					
By itself 95'	17.93	–2.2	21	28.9	–17.1
95'/65' (Δ 30')	19.39	–6.9	24.3	21.4	–21.9
95'/60' (Δ 35')	19.38	–8.6	24	21.4	–16.9
95'/55' (Δ 40')	19.29	–10.9	24	25.0	–18.6
95'/70' (Δ 25')	19.26	–5.5	23	24	–35.3
95'/75' (Δ 20')	19.08	–4.6	23	27	–23.4

amplitude of the second lobe due to ground reflections, and the elevation angle of that second lobe is listed as well.

Besides the 3- and 7-element designs discussed above, we've also added 4- and 5-element designs in Table 11. Over the range of stacking distances between 20 and 40 feet on 15 meters (0.43 λ to 0.86 λ), the peak gain for the 3-element stacks changes less than 0.75 dB, with the 30-foot spacing exhibiting the highest gain. The differences between peak gains versus stacking distance become smaller as the boom length increases. For example, for the 64-foot boom Yagi, the gain varies 19.39 – 19.08 = 0.31 dB for stack spacings from 20 to 40 feet.

In other words, changing the spacing from 20 to 40 feet (0.43 λ to 0.86 λ) doesn't change the gain significantly for boom lengths from 12 to 64 feet (0.26 λ to 1.38 λ). From the point of view of gain, the vertical spacing between individual antennas in an HF stack is not critical.

The worst-case lobes (generally speaking, the second lobe due to ground reflections) are highest for a Yagi operated by itself. After all, a single Yagi doesn't benefit from the redistribution of energy from higher-angle lobes into the main lobe that a stack gives. Thus, the 3-element, 12-foot boom Yagi by itself at 95 feet would have a second lobe at 21° that is only 0.9 dB down from the main lobe, while the stack of two such antennas at a 30-foot (0.65 λ) spacing at 95/65 feet would have a second lobe down 4.5 dB. As the spacing between antennas in a vertical stack increases, the second lobe is suppressed more, up to 8.7 dB at a 40-foot (0.86 λ) spacing.

Since the free-space elevation pattern for a 3-element Yagi is wider than that for a 7-element Yagi, the second lobe due to ground reflection will be somewhat reduced. This is true for all longer-boom antennas operating by themselves over ground. Used in stacks, the second lobe's amplitude will vary depending on spacing between antennas, but they range only about 6 dB.

The front-to-back ratio will also tend to increase with longer boom lengths on a properly designed Yagi. Table 11 shows that the F/B is somewhat better for closer spacings between antennas in a stack, a rather non-intuitive result, considering that the mutual coupling should be greater for closer antennas. For example, the 5-element Yagi stack with a 20-foot spacing has a exceptional F/B of 34.3 dB, compared to a F/B of 21.4 dB with the 30-foot spacing distance that gives nominally the most gain. High values of F/B, however, rarely hold over a wide frequency range because of the very critical phasing relationships necessary to get a deep null, so the difference between 34.3 and 21.4 dB would rarely be noticeable in practice.

The near-overhead lobe structure (between 60º to 120º in elevation) tends also to be lower for smaller stack spacings—for all boom lengths—peaking in this example at a spacing of 25 feet for the boom lengths considered here. Since the peak gain actually occurs with smaller spacing between Yagis in this 7-element stack, even relatively large and messy looking overhead lobes are not subtracting from the stacking gain. In the next section we'll now examine whether this overhead lobe is important or not.

Are Higher-Angle Lobes Important?

We've already shown that the exact spacing between HF Yagis is not critical for stacking gain. Further, the heights (and hence spacing) of the individual Yagis in a stack interact in a complicated fashion to determine higher-angle lobes.

Let's examine the relevance of such higher-angle lobes for stacked HF Yagis, this time in terms of interference reduction on receive. As Chapter 23, Radio Wave Propagation, points out, few DX signals arrive at elevation angles greater than about 30°. In fact, DX signals only propagate at elevation angles in the range from 1° to 30° on all the bands where operators might reasonably expect to stack Yagis—nominally from 7 to 29.7 MHz.

You should remember that the definition of the *critical frequency* for HF propagation is the highest frequency for which a wave launched directly overhead at 90° elevation is reflected back down to Earth, rather than being lost into outer space. The maximum critical frequency for extremely high levels of solar flux is about 15 MHz. In other words, high overhead angles do not propagate signals on the upper HF bands.

However, some domestic signals do arrive at relatively high elevation angles. Let's look at some scenarios where higher angles might be encountered and how the elevation patterns of typical HF stacks affect these signals. Let's examine a situation where a medium-range interfering station is on the same heading as a more distant target station.

We'll examine a typical scenario involving stations in Atlanta, Boston and Paris. The heading from Atlanta to Paris is 49º, the same heading as Atlanta to Boston. In other words, the Atlanta station would have to transmit over (and listen through) a Boston station for communication with Paris. The distance between Atlanta and Boston is about 940 miles, while the distance from Atlanta to Paris is about 4350 miles. Ground wave signals obviously cannot travel either of these distances at 21 MHz (ground wave coverage is less than about 10 miles at this frequency), and so the propagation between Atlanta to Boston and Atlanta to Paris will be entirely by means of the ionosphere.

Let's evaluate the situation on 15 meters in the month of October. We'll assume a smoothed sunspot number (SSN) of 100 and that each station puts 1500 W of power into theoretical isotropic antennas that have +10 dBi of gain at all elevation and azimuth angles. [We use such theoretical isotropic antennas because they make it easier to work in *VOACAP*. We will factor in real-world stacks later.] *VOACAP* predicts that the signal from Boston will be S9 + 8 dB in Atlanta at 1400 UTC, arriving at an elevation angle of 21.3º on a single F_2 hop. This elevation angle is higher than commonly encountered angles for DX signals, but it is still far away from near-overhead angles.

The signal from Paris into Atlanta is predicted to be about S6 for the same theoretical isotropic antennas, at an incoming elevation angle of 6.4º on three F_2 hops. The S6 level validates the rule-of-thumb that each extra hop loses approximately 10 dB of signal strength, assuming that each S unit is about 4 dB, typical for modern receivers.

Now look at **Fig 36**, which shows the response for a stack of 3-element Yagis at 90/60/30 feet over flat ground, along with the response for a similar stack of 7-element Yagis. Again, we'll assume that all three stations are using such 3-element 90/60/30-foot stacks. The stations in Atlanta and Boston point their stacks into Europe and the Parisian station points his stack towards the USA. The gain of the Atlanta array at 6.4º into Paris will be about 16 dBi, or 6 dB more than the isotropic array with its +10 dBi of gain selected for use in *VOACAP*. Similarly, the French station's transmitted signal will enjoy a 6 dB gain advantage over the isotropic array used in the *VOACAP* calculation, and thus the French signal into Atlanta will now be S6 + 12 dB, or about S9.

By comparison, the interfering signal from Boston into Atlanta will be reduced by the rearward pattern of his array, which will launch a signal at 180º – 21.3º = 158.7º in elevation at the single F_2 mode from Boston to Atlanta. From Fig 33, the Boston station's gain at this rearward elevation is going to drop from the isotropic's +10 dBi of gain down to –11 dBi, a drop of 21 dB. The signal into the Atlanta receiver will also be reduced by the pattern of the Atlanta array on receive, which has a gain of about 0 dBi at 21.3º, compared to the isotropic's +10 dBi gain at 6.4º, a net drop of 10 dB.

Thus, the Boston station's signal will drop by about 21 + 10 = 31 dB, bringing the interfering signal from Boston, which would be S9 + 8 dB for isotropic antennas, down to about S3 due to the combined effects of the arrays. This is a very significant reduction in interference. But you will note that the reduction has nothing to

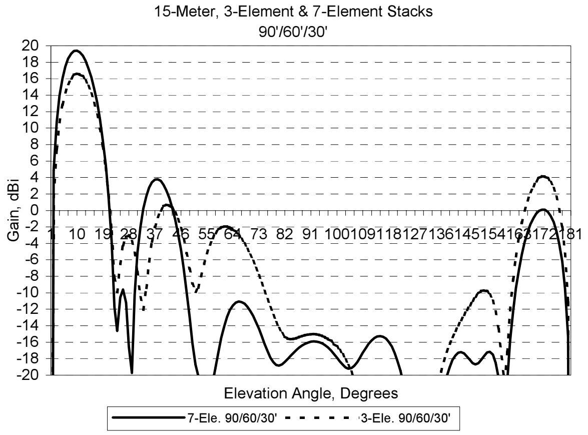

Fig 36—Stacks of three 3-element and 7-element Yagis on 15 meters at 90/60/30 feet heights. The F/B for the 7-element stack is superior to the 3-element stack mainly because the F/B is intrinsically better for the long-boom design.

do with the near-overhead lobes, dealing as it does with the trailing edge of the main lobe and the F/B lobe.

Even Higher Elevation Angles

Now let's evaluate a station that is even closer to Boston, say a station in Philadelphia. The heading from Philadelphia to Paris is 53º and the distance is 3220 miles. On the same day in October as above, *VOACAP* predicts a signal strength of S8 from Paris to Philadelphia, at a 2.7º elevation angle on two F_2 hops. Again, the *VOACAP* computations assume isotropic antennas with +10 dBi gain at all three stations. The gain of the 3-element stacks at both ends of the circuit at 2.7º is also about +10 dBi, so the signal level from Paris to Philadelphia would be S8 with the 3-element stacks.

Now *VOACAP* computes the elevation angle from Philadelphia to Boston as 56.3º, on one F_2 hop launched at an azimuth of 53º, well within the azimuthal beamwidth of the stack. *VOACAP* says the predicted signal strength for isotropic antennas with +10 dBi of gain is less than S1!

What's happening here? Boston and Philadelphia are within the "skip" region on 21 MHz and signals are skipping right over Boston from Philadelphia (and vice versa). Actual signals would be much weaker than they would be with theoretical isotropic antennas because of the actual patterns of the transmitting and receiving stacks. At an elevation angle of 56.3º the receiving stack would have a gain of –10 dBi, while at an elevation of 180º – 56.3º = 123.7º the transmitting stack would be down to –10 dBi as well. The net reduction for the stacks compared to isotropics with +10 dBi gain each would be 40 dB, putting the interfering signal well into the receiver noise.

You can safely say that near-overhead angles don't enter into the picture, simply because signals at intermediate distances are in the ionospheric skip zone and interfering signals are very weak in that zone already.

Even in situations where having a poor front-to-back ratio might be beneficial—because it alerts stations tuning across your signal that you are occupying that frequency—the ionosphere doesn't cooperate for intermediate-distance signals that are in the skip zone. Often two stations may be on the same frequency without either knowing that the other is there.

Ground Wave?

What happens, you might wonder, for ground-wave signals? Let's look at a situation where the interfering station is in the same direction as the desired target, but is only 5 miles away. Unfortunately, his signal is S9 + 50 dB. Even reducing the level by 30 dB, a huge number, is still going to make his signal 20 dB stronger than signals from your desired target location! There is not much you can do about ground-wave signals and fretting about optimizing stack heights to discriminate against local signals is generally futile.

Stacking Distances for Multiband Yagis

By definition, a stack of multiband Yagis (such as a "tribander" covering 20/15/10 meters) has a constant vertical spacing between antennas in terms of feet or meters, but not in terms of wavelength. Tribanders are no different than monobanders in terms of optimal spacing between individual antennas. Again, the difference in gain between spacings of 0.5 λ and 1.0 λ for a stack of triband Yagis amounts to only a fraction of a decibel. Furthermore, the main practical constraint that limits choice of stacking distances between any kind of Yagis, multiband or monoband, is the spacing between guy wire sets on the tower itself.

Summary, Stacking Distances

In short, let us summarize that there is nothing magical about stacking distances for practical HF Yagis—a good rule-of-thumb is a stacking distance of 0.65 λ. This is 23 feet on 10 meters, 30 feet on 15 meters and 45 feet on 20 meters for monoband stacks. Practically speaking, however, you've only got limited places where you can mount antennas on the tower—mainly where guy wires allow you to place them. This is especially applicable if you wish to rotate lower antennas on the tower, where you must clear the guys from up above.

STACKS AND FADING

The following is derived from an article by Fred Hopengarten, K1VR, and Dean Straw, N6BV, in a Feb 1994 *QST* article. Using stacked Hy-Gain TH7DXs or TH6DXXs at their respective stations, they have solicited a number of reports from stations, mainly in Europe, to compare various combinations of antennas in stacks and as single antennas. The peak gain of the stack is usually just a little bit higher than that for the best of the single antennas, which is not surprising. Even a large stack has no more than about 6 dB of gain over a single Yagi at a height favoring the prevailing elevation angle. Fading on the European path can easily be 20 dB or more, so it is very confusing to try to make definitive comparisons. They have noticed over many tests that the stacks are much less susceptible to fading compared to single Yagis. Even within the confines of a typical SSB bandwidth, frequency-selective fading occasionally causes the tonal quality of a voice to change on both receive and transmit, often dramatically becoming fuller on the stacks, and tinnier on the single antennas. This doesn't happen all the time, but is often seen. They have also observed often that the depth of a fade is less, and the period of fading is longer, on the stacks compared to single antennas.

Exactly *why* stacks exhibit less fading is a fascinating subject, for which there exist a number of speculative ideas, but little hard evidence. Some maintain that stacks outperform single antennas because they can afford *space diversity* effects, where by virtue of the difference in physical placement one antenna will randomly pick up signals that another one in another physical location might not hear.

This is difficult to argue with, and equally difficult to prove scientifically. A more plausible explanation about why stacked Yagis exhibit superior fading performance is that their narrower frontal elevation lobes can discriminate against undesired propagation modes. Even when band conditions favor, for example, a very low 3° elevation angle on 10 or 15 meters from New England to Western Europe, there are signals, albeit weaker ones, that arrive at higher elevation angles. These higher-angle signals have traveled longer distances on their journey through the ionosphere, and thus their signal levels and their phase angles are different from the signals traversing the primary propagation mode. When combined with the dominant mode, the net effect is that there is both destructive and constructive fading. If the elevation response of a stacked antenna can discriminate against signals arriving at higher elevation angles, then in theory the fading will be reduced. Suffice it to say: In practice, stacks do reduce fading.

STACKS AND PRECIPITATION STATIC

The top antenna in a stack is often much more affected by rain or snow precipitation static than is the lower antenna. N6BV and K1VR have observed this phenomenon, where signals on the lower antenna by itself are perfectly readable, while S9+ rain static is rendering reception impossible on the higher antenna or on the stack. This means that the ability to select individual antennas in a stack can sometimes be extremely important.

STACKS AND AZIMUTHAL DIVERSITY

Azimuthal diversity is a term coined to describe the situation where one of the antennas in a stack is purposely pointed in a direction different from the main direction of the stack. During most of the time in a DX contest from the East Coast, the lower antennas in a stack are pointed into Europe, while the top antenna is often rotated toward the Caribbean or Japan. In a stack of three identical Yagis, the first-order effect of pointing one antenna in a different direction is that one-third of the transmitter power is diverted from the main target area. This means that the peak gain is reduced by 1.8 dB, not a very large amount considering that signals are often 10 to 20 dB over S9 anyway when the band is open from New England to Europe.

Fig 37 shows the 3D pattern of a pair of 4-element Yagis fed in-phase at 95 and 65 feet, but where the lower antenna has been rotated 180° to fire in the –X direction. The backwards lobe peaks at a higher elevation angle because the antenna doing the radiating in this direction is lower on the tower. The forward lobe peaks at a lower angle because its main radiator is higher.

THE N6BV/1 ANTENNA SYSTEM—BRUTE FORCE FEEDING

The N6BV/1 system in Windham, New Hampshire, was located on the crest of a small hill about 40 miles from Boston, and could be characterized as a good, but not dominant, contesting station. A number of top-10 contest results were achieved from that station in the 1990s before N6BV returned to California.

There was a single 120-foot high Rohn 45 tower, guyed at 30-foot intervals, with a 100-foot horizontal spread from tower base to each guy point so there was sufficient room for rotation of individual Yagis on the tower. Each set of guy wires employed heavy-duty insulators at 57-foot intervals, to avoid resonances in the 80 through 10-meter amateur bands. There were five Yagis on the tower. A heavy-duty 12-foot long steel mast with 0.25-inch walls was at the top of the tower, turned by an Orion 2800 rotator. Two thrust bearings were used above the rotator, one at the top plate of the tower itself, and the other about 2 feet down in the tower on a modified rotator shelf plate. The two thrust bearings allowed the rotator to be removed for service.

At the top of the mast, 130 feet high, was a 5-element, computer-optimized 10-meter Yagi, which was a modified Create design on a 24-foot boom. The element tuning was modified from the stock antenna in order to achieve higher gain and a better pattern over the band. At the top of the tower (120-foot level) was mounted a Create 714X-3 triband Yagi. This was a large tribander, with a 32-foot boom and five elements. Three elements were active on 40 meters, four were active on 20 meters and four were active on 15 meters. The 40-meter elements were loaded with coils, traps and

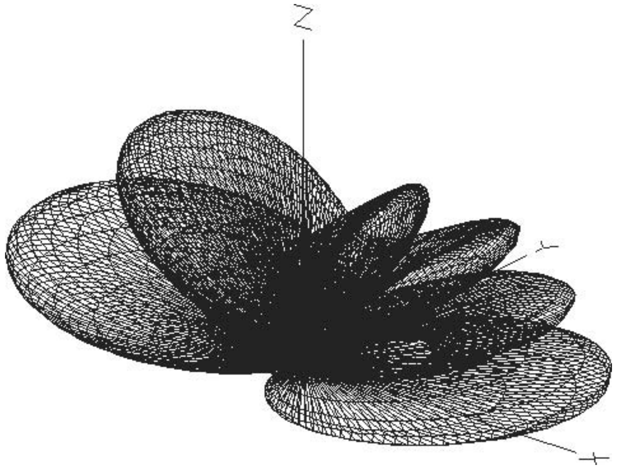

Fig 37—3D representation of the pattern for two 4-element 15-meter Yagis, with the top antenna at 95 and the bottom at 65 feet, but pointed in the opposite direction.

capacitance hats, and were approximately 46 feet long. A triband 20/15/10-meter Hy-Gain TH7DX tribander was fixed into Europe at the 90-foot level on the tower, just above the third set of guys.

At the 60-foot level on the tower, just above the second set of guys, there was a "swinging-gate" side-mount bracket, made by DX Engineering of Oregon. A Hy-Gain *Tailtwister* rotator turned a TH7DX on this side mount. (Note that both the side mount and the element spacings of the TH7DX itself prevented full rotation around the tower—about 280° of rotation was achieved with this system.) At the 30-foot level, just above the first set of guys, was located the third TH7DX, also fixed on Europe.

All five Yagis were fed with equal lengths of Belden 9913 low-loss coaxial cable, each measured with a noise bridge to ensure equal electrical characteristics. At each feed point a ferrite-bead choke balun (using seven large beads) was placed on the coax. All five coaxial cables went to a relay switch box mounted at the 85-foot level on the tower. **Fig 38** shows the schematic for the switch box, which was fed with 250 feet of 75-Ω, 0.75-inch OD Hardline coaxial cable.

The stock DX Engineering remote switch box was modified by adding relay K6, so that either the 130-foot or the 120-foot rotating antenna could be selected through a second length of 0.75-inch Hardline going to the shack. This created a *Multiplier* antenna, independent of the *Main* antennas. A second band could be monitored in this fashion while calling CQ using the main antennas on another band. Band-pass filters were required at the multiplier receiver to prevent overload from the main transmitter.

The 0.75-inch Hardline had very low losses, even when presented with a significant amount of SWR at the switch-box end. This was important, because unlike K1VR's system, no attempt was made at N6BV to maintain a constant SWR when relays K1 through K5 were switched in or out. This seemingly cavalier attitude came about because of several factors. First, there were many

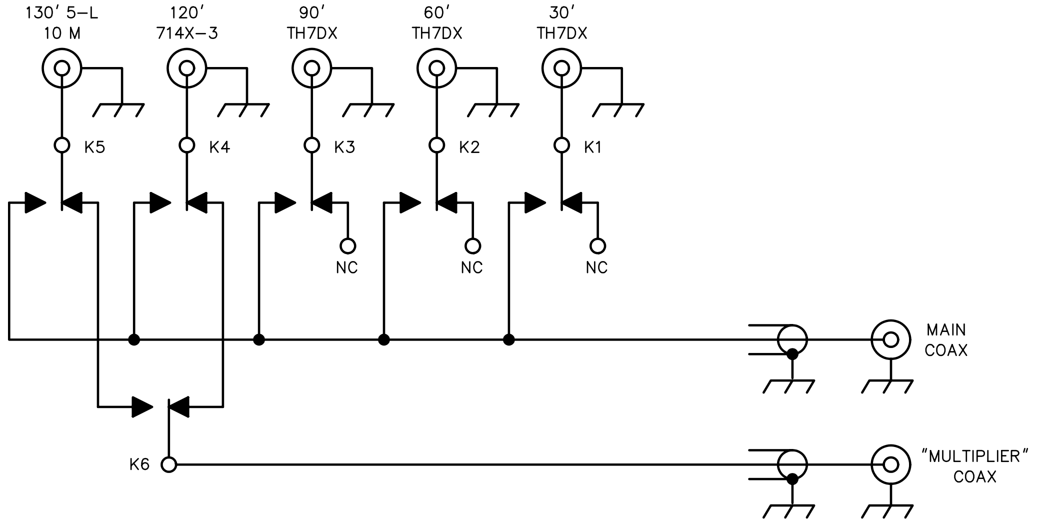

Fig 38—N6BV/1 switch box system. This uses a modified DX Engineering remote switch box, with relay K6 added to allow selection of either of the two top antennas (5-element 10-meter Yagi or 40/20/15-meter triband 714X-3) as a "multiplier" antenna. There is no special provision for SWR equalization when any or all of the Yagis are connected in parallel as a stack fed by the Main coaxial cable. Each of the five Yagis is fed with equal lengths of flexible Belden 9913 coax, so phasing can be maintained on any band. The Main and "Multiplier" coaxes going to the shack are 0.75" OD 75-Ω Hardline cables.

different combinations of antennas that could be used together in this system. Each relay coil was independently controlled by a toggle switch in the shack. N6BV was unable to devise a matching system that did not become incredibly complex because of the numerous impedance combinations used over all the five bands.

Second, the worst-case additional transmission line loss due to a 4:1 SWR mismatch when four antennas were connected in parallel on 10 meters was only 0.5 dB. It was true that the linear amplifier had to be retuned slightly when combinations of antennas were switched in and out, but this was a small penalty to pay for the reduced complexity of the switching and matching networks. The 90/60/30-foot stack into Europe was used for about 95% of the time during DX contests, so the small amount of amplifier retuning for other antenna combinations was considered only a minor irritation.

WHY TRIBANDERS?

Without a doubt, the most common question K1VR and N6BV have been asked is: "Why did you pick *tribanders* for your stacks?" Triband antennas were chosen with full recognition that they are compromise antennas. Other enterprising amateurs have built stacked tribander arrays. Bob Mitchell, N5RM, is a prominent example, with his so-called *TH28DX* array of four TH7DX tribanders on a 145-foot-high rotating tower. Mitchell employed a rather complex system of relay-selected tuned networks to choose either the upper stacked pair, the lower stacked pair or all four antennas in stack. Others in Texas have also had good results with their tribander stacks. Contester Danny Eskenazi, K7SS, has very successfully used a pair of stacked KT-34XA tribanders for years.

A major reason why tribanders were used is that over the years both authors have had good results using TH6DXX or TH7DX antennas. They are ruggedly built, mechanically and electrically. They are able to withstand New England winters without a whimper, and their 24-foot long booms are long enough to produce significant gain, despite trap-loss compromises. Amateurs speculating about trap losses in tribanders freely bandy about numbers between 0.5 and 2 dB. Both N6BV and K1VR are comfortable with the lower figure, as are the Hy-Gain engineers.

Consider this: If 1500 W of transmitter power is going into an antenna, a loss of 0.5 dB amounts to 163 W. This would create a significant amount of heat in the six traps that are on average in use on a TH6DXX, amounting to 27 W per trap. If the loss were as high as 1 dB, this would be 300 W total, or 50 W per trap. Common sense says that if the overall loss were greater than about 0.5 dB, the traps would act more like big *firecrackers* than resonant circuits! A long-boom tribander like the TH6DXX or TH7DX also has enough space to employ elements dedicated to different bands, so the compromises in element spacing usually found on short-boom 3 or 4-element tribanders can be avoided.

Another factor in the conscious choice of tribanders was first-hand frustration with the serious interaction that can result from stacking monoband antennas closely together on one mast in a Christmas Tree configuration. N6BV's worst experience was with the ambitious 10 through 40-meter Christmas Tree at W6OWQ in the early 1980s. This installation used a Tri-Ex SkyNeedle tubular crankup tower with a rotating 10-foot-long heavy-wall mast. The antenna suffering the greatest degradation was the 5-element 15-meter Yagi, sandwiched 5 feet below the 5-element 10-meter Yagi at the top of the mast, and 5 feet above the full-sized 3-element 40-meter Yagi, which also had five 20-meter elements interlaced on its 50-foot boom.

The front-to-back ratio on 15 meters was at best about 12 dB, down from the 25+ dB measured with the bottom 40/20-meter Yagi removed. No amount of fiddling with element spacing, element tuning or even orientation of the 15-meter boom with respect to the other booms (at 90° or 180°, for example) improved its performance. Further, the 20-meter elements had to be lengthened by almost a foot *on each end of each element* in order to compensate for the effect of the interlaced 40-meter elements. It was a lucky thing that the tower was a motorized crankup, because it went up and down hundreds of times as various experiments were attempted!

Interaction due to close proximity to other antennas in a short Christmas Tree can definitely destroy carefully optimized patterns of individual Yagis. Nowadays, such interaction can be modeled using a computer program such as *EZNEC* or *NEC*. A gain reduction of as much as 2 to 3 dB can easily result due to close vertical spacing of monobanders, compared to the gain of a single monoband antenna mounted in the clear. Curiously enough, at times such a reduction in gain can be found even when the front-to-back ratio is not drastically degraded, or when the front-to-back occasionally is actually *improved*.

If you plan on stacking monoband Yagis—for example, putting only 15-meters Yagis on a single tower, with your other monoband stacks on other towers—do make sure you model the system to see if any interactions occur. You may be quite surprised.

Finally, in the N6BV/1 installation, triband antennas were chosen because the system was meant to be as simple as possible, given a certain desired level of performance, of course. Triband antennas make for less mechanical complexity than do an equivalent number of monobanders. There were five Yagis on the N6BV/1 tower, yielding gain from 40 to 10 meters, as opposed to using 12 or 13 monobanders on the tower.

THE K1VR ARRAY: A MORE ELEGANT APPROACH TO MATCHING

The K1VR stacked array is on a 100-foot high Rohn

25 tower, with sets of guy wires at 30, 60 and 90 feet, made of nonconducting Phillystran. Phillystran is a nonmetallic Kevlar rope covered by black polyethylene to protect against the harmful effects of the sun's ultraviolet rays. A caution about Phillystran: Don't allow tree branches to rub against it. It is designed to work in tension, but unlike steel guy wire, it does not tolerate abrasion well.

Both antennas are Hy-Gain TH6DXX tribanders, with the top one at 97 feet and the bottom one at 61 feet. The lower antenna is rotated by a Telex Ham-M rotator on a homemade swinging-gate side mount, which allows it to be rotated 300° around the tower without hitting any guy wires or having an element swing into the tower. At the 90-foot point on the tower, a 2-element 40-meter Cushcraft Yagi has been mounted on a RingRotor so it can be rotated 360° around the tower.

After several fruitless attempts trying to match the TH6DXX antennas so that either could be used by itself or together in a stack, K1VR settled on using a relay-selected broadband toroidal matching transformer. When both triband antennas are fed together in parallel as a stack, it transforms the resulting 25-Ω impedance to 50 Ω. The transformer is wound on a T-200A powdered-iron core, available from Amidon, Palomar Engineering or Ocean State Electronics. Two lengths of twin RG-59 coax (sometimes called Siamese or WangNet), four turns each, are wound on the core. Two separate RG-59 cables could be used, but the Siamese-twin cable makes the assembly look much more tidy. The shields of the RG-59 cables are connected in series, and the center conductors are connected in parallel. See **Fig 39** for details.

Fig 40 shows the schematic of the K1VR switch box, which is located in the shack. Equal electrical lengths of 50-Ω Hardline are brought from the antennas into the shack and then to the switch box. Inside the box, the relay contacts were soldered directly to the SO-239 chassis connectors to keep the wire lengths down to the absolute minimum. K1VR used a metal box that was larger than

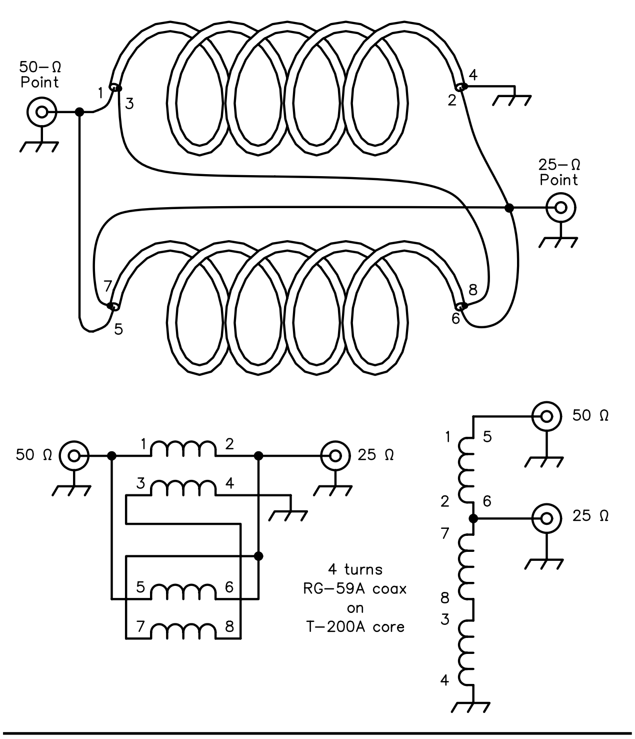
50-Ω Point
1
3
4
2
25-Ω Point
7
8
5
6
50 Ω
1
2
3
4
5
6
7
8
25 Ω
4 turns RG-59A coax on T-200A core
50 Ω
1
5
2
6
7
25 Ω
8
3
4

Fig 39—Diagram for matching transformer for K1VR stacked tribander system. The core is powdered iron-core T-200A, with four turns of two RG-59A or "Siamese" coax cables. Center conductors are connected in parallel and shields are connected in series to yield 0.667:1 turns ratio, close to desired 25-Ω to 50-Ω transformation.

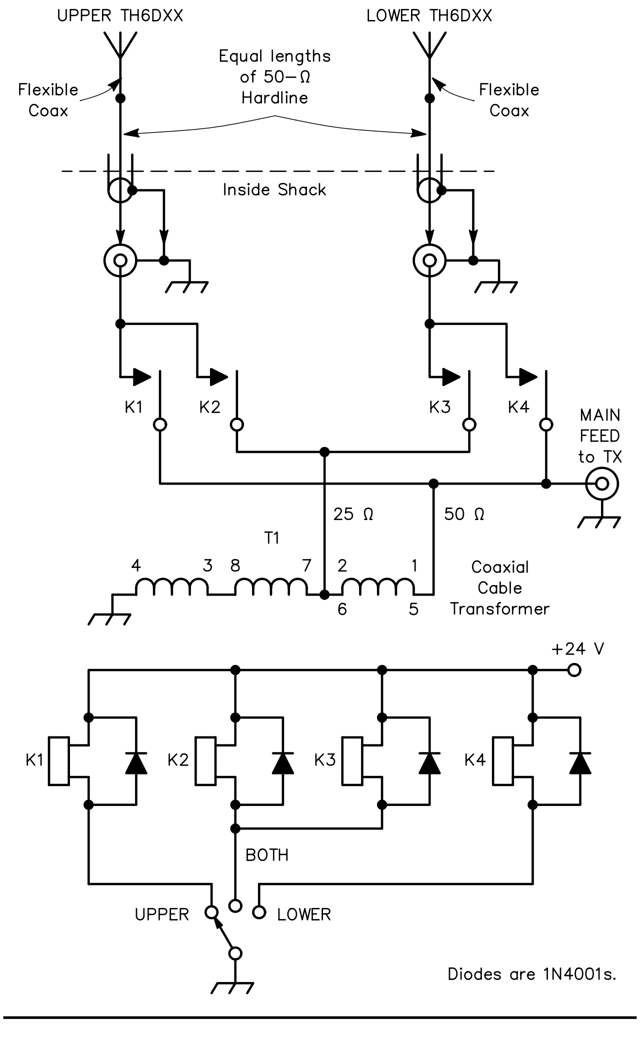
UPPER TH6DXX
LOWER TH6DXX
Equal lengths of 50-Ω Hardline
Flexible Coax
Flexible Coax
Inside Shack
K1
K2
K3
K4
MAIN FEED to TX
25 Ω
50 Ω
T1
4
3
8
7
2
1
6
5
Coaxial Cable Transformer
+24 V
K1
K2
K3
K4
BOTH
UPPER
LOWER
Diodes are 1N4001s.

Fig 40—Relay switch box for K1VR stacked tribander system. Equal lengths of 50-Ω Hardline (with equal lengths of flexible 50-Ω cable at each antenna to allow rotation) go to the switch box in the shack. The SWR on all three bands for Upper, Lower or Both switch positions is very close to constant.

might appear necessary because he wanted to mount the toroidal transformer with plenty of clearance between it and the box walls. The toroid is held in place with a piece of insulation foam board. Before placing the switch box in service, the system was tested using two 50-Ω dummy loads, with equal lengths of cable connected in parallel to yield 25 Ω. The maximum SWR measured was 1.25:1 at 14 MHz, 1.3:1 at 21 MHz and 1.15:1 at 28 MHz, and the core remained cold with 80 W of continuous output power.

One key to the system performance is that K1VR made the electrical lengths of the two hardlines the same (within 1 inch) by using a borrowed TDR (time domain reflectometer). Almost as good as Hardline, K1VR points out, would be to cut exactly the same length of cable from the same 500-foot roll of RG-213. This eliminates manufacturing tolerances between different rolls of cable.

K1VR's experience over the last 10 years has been that at the beginning of the 10 or 15-meter morning opening to Europe the upper antenna is better. Once the band is wide open, both antennas are fed in phase to cast a bigger shadow, or footprint, on Europe. By mid-morning, the lower antenna is better for most Europeans, although he continues to use the stack in case someone is hearing him over a really long distance path throughout Europe. He reports that it is always very pleasant to be called by a 4S7 or HS0 or VU2 when he is working Europeans at a fast clip!

SOME SUGGESTIONS FOR STACKING TRIBANDERS

It is unlikely that many amateurs will try to duplicate exactly K1VR's or N6BV's contest setups. However, many hams already have a tribander on top of a moderately tall tower, typically at a height of about 70 feet. It is not terribly difficult to add another, identical tribander at about the 40-foot level on such a tower. The second tribander can be pointed in a fixed direction of particular interest (such as Europe or Japan), or it can be rotated around the tower on a side mount or a Ring Rotor. If guy wires get in the way of rotation, the antenna can usually be arranged so that it is fixed in a single direction.

Insulate the guy wires at intervals to ensure that they don't shroud the lower antenna electrically. A simple feed system consists of equal-length runs of surplus 0.5-inch 75-Ω Hardline (or more expensive 50-Ω Hardline, if you are really obsessed by SWR) from the shack up the tower to each antenna. Each tribander is connected to its respective Hardline feeder by means of an equal length of flexible coaxial cable, with a ferrite choke balun, so that the antenna can be rotated.

Down in the shack, the two hardlines can simply be switched in and out of parallel to select the upper antenna only, the lower antenna only, or the two antennas as a stack. See **Fig 41**. Any impedance differences can be handled as stated previously, simply by retuning the linear amplifier, or by means of the internal antenna tuner (included in most modern transceivers) when the transceiver is run barefoot. The extra performance experienced in such a system will be far greater than the extra decibel or two that modeling calculates.

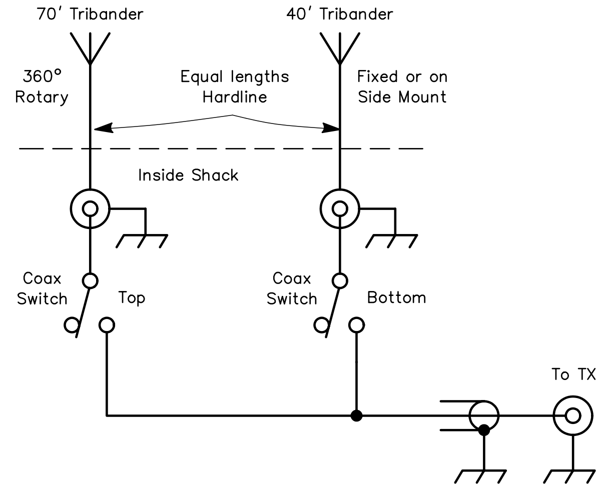

Fig 41—Simple feed system for 70/40-foot stack of tribanders. Each tribander is fed with equal lengths of 0.5-inch 75-Ω Hardline cables (with equal lengths of flexible coax at the antenna to allow rotation), and can be selected singly or in parallel at the operator's position in the shack. Again, no special provision is made in this system to equal SWR for any of the combinations.

THE WX0B APPROACH TO STACK MATCHING AND FEEDING

Earlier we mentioned how useful it would be to switch various antennas in or out of a stack, depending on the elevation angles that need to be emphasized at that moment. Jay Terleski, WX0B, of Array Solutions has designed switchable matching systems, called *StackMatches*, for stacks of monoband or multiband Yagis.

The StackMatch uses a 50-Ω to 22.25-Ω broadband transmission-line transformer to match combinations of up to three Yagis in a stack. See **Fig 42** for a schematic of the StackMatch. For selection of any 50-Ω Yagi by itself, no matching transformer is needed and Relay IN routes RF directly to the common bus going to Relay 1, 2 and 3. For selection of two Yagis together the parallel impedance is 50/2 = 25 Ω and Relay IN routes RF to the matching transformer. The SWR is 25/22.25 = 1.1:1. For three Yagis used together, the parallel impedance is 50/3 = 16.67 Ω, and the SWR is 22.25/16.67 = 1.3:1.

The broadband transformer consists of four trifilar turns of #12 enamel-insulated wire wound on a Ferrite Corporation FT-240 2.4-inch OD core made of #61

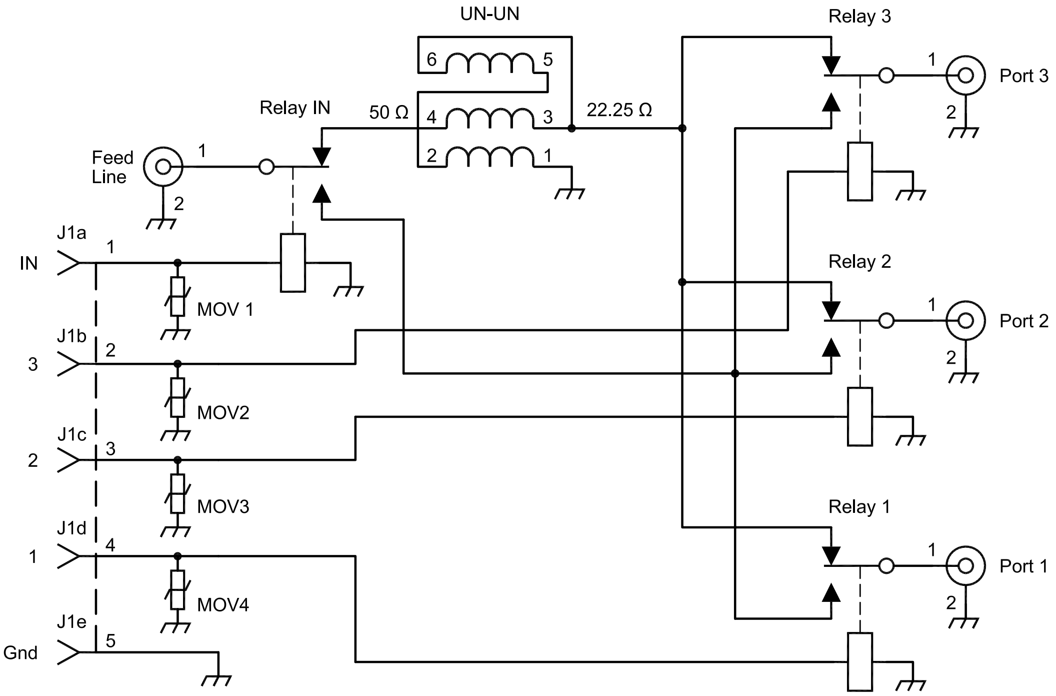

Fig 42—Schematic of WX0B's StackMatch 2000 switchbox, which uses a broadband transmission line transformer using trifilar #12 enamel-insulated wires. (*Courtesy Array Solutions.*)

material ($\mu = 125$). WX0B uses 10-A relays enclosed in plastic cases to do the RF switching, selected by a control box at the operating position. (10-A relays can theoretically handle $10\ A^2 \times 50\ \Omega = 5000$ W.) **Fig 43** shows a photo of the transmission-line transformer and StackMaster PCB.

The control/indicator box uses a diode matrix to switch various combinations of antennas in/out of the stack. Three LEDs lined up vertically on the front panel indicate which antennas in a stack are selected.

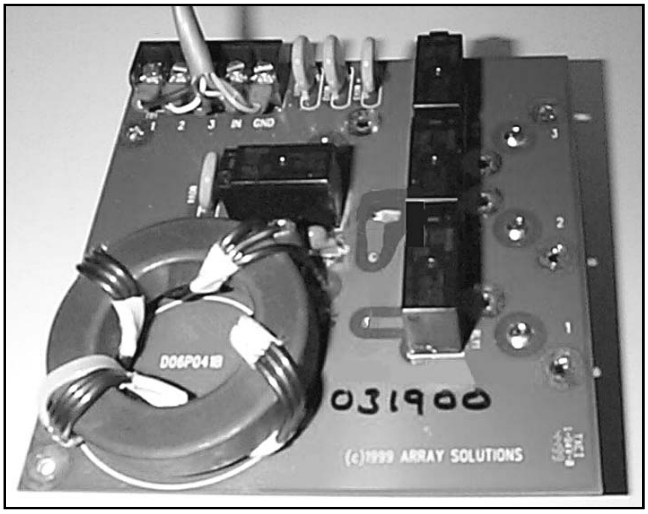

Fig 43—Inside view of StackMatch. (*Photo courtesy Array Solutions.*)

"BIP/BOP" OPERATION

The contraction "BIP" means "both in-phase," while "BOP" means "both out-of-phase." BIP/BOP refer to stacks containing two Yagis, although the term is commonly used for stacks containing more than two Yagis. In theory, feeding a stack with the antennas out-of-phase will shift the elevation response higher than in-phase feeding.

Fig 44 shows a rectangular plot comparing BIP/BOP operation of two 3-element 15-meter Yagis at heights of $2\ \lambda$ and $1\ \lambda$ (93 and 46 feet) over flat ground. The BOP pattern is the higher-angle lobe and the two lobes cross over about 14°. The maximum amplitude of the BOP stack's gain is about ½ dB less than the BIP pair. For reference, the pattern of a single 46-foot high Yagi is overlaid on the pattern for the stacks.

The most common method for feeding one Yagi 180° out-of-phase is to include an extra electrical half wavelength of feed-line coax going to one of the antennas. This method obviously works on a single frequency band and thus is not applicable to stacks of multiband Yagis, such as tribanders. For such multiband stacks, feeding only the lower antenna(s)—by switching out higher antenna(s) in the stack—is a practical method for achieving better coverage at medium or high elevation angles.

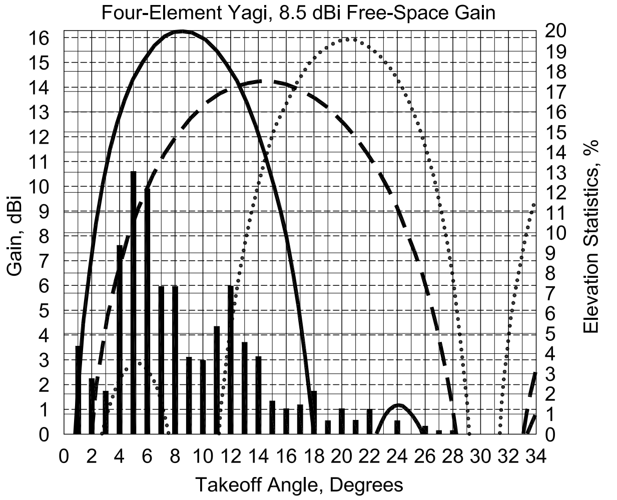

Fig 44—*HFTA* screen shot of "BIP/BOP" operation of two 4-element 15-meter Yagis at 93 and 46 feet above flat ground. The elevation response in BOP (both out-of-phase) operation is shifted higher, peaking at about 21°, compared to the BIP (both in-phase) operation where the peak is at 8°. The dashed line is response of single Yagi at 46 feet.

STACKING DISIMILAR YAGIS

So far we have been discussing vertical stacks of identical Yagis. Less commonly, hams have successfully stacked dissimilar Yagis. For example, consider a case where two 5-element 10-meter Yagis are placed 46 and 25 feet above flat ground, with a 7-element 10-meter Yagi at 68 feet on the same tower. See **Fig 45**, which is a schematic of the layout for this stack. Note that the driven element for the top 7-element Yagi is well behind the vertical plane of the driven elements for the two 5-element Yagis. This offset distance must be compensated for with a phase shift in the drive system for the top Yagi.

Fig 46 shows the elevation-pattern responses for uncompensated (equal-length feed lines) and the compensated (additional 150° of phase shift to top Yagi) stacks. These patterns were computed using *EZNEC ARRL*, which is included with this book. Not only is about 1.7 dB of maximum gain lost, but the peak elevation angle is shifted upwards by 11° from the optimal takeoff angle of 8°—where some 10 dB of gain is also lost. Without compensation, this is a severe distortion of the stack's elevation pattern.

For RG-213 coax, the extra length needed to provide an additional 150° of phase shift = 150°/360° λ = 0.417 λ = 9.53 feet at 28.4 MHz. This was computed using the program *TLW* (Transmission Line for Windows) included on the CD-ROM accompanying this book.

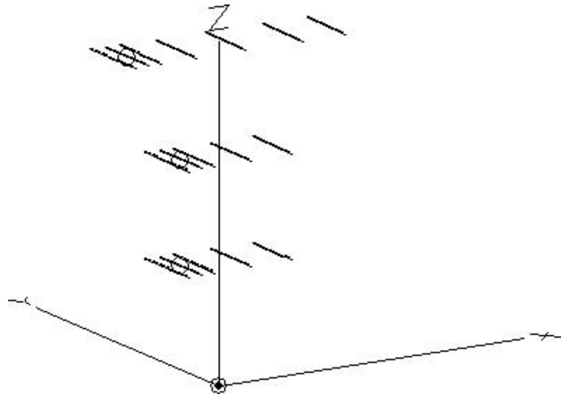

Fig 45—Stacking dissimilar Yagis. In this case a 7-element 10-meter Yagi is stacked over two 5-element Yagis. Note the displacement of the 7-element Yagi's driven element compared to the position of the two 5-element Yagis. This leads to an undesired phase shift for the higher antenna.

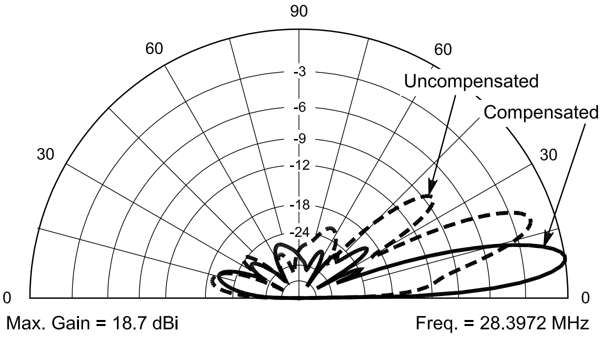

Fig 46—Comparison of elevation responses for 7/5/5-element 10-meter stacks, with and without compensation for driven-element offset.

It is not always possible to compensate for dissimilar Yagis in a stack with a simple length of extra coax, so you should be sure to model such combinations to make sure that they work properly. A safe alternative, of course, is to stack only identical Yagis, feeding all of them with equal lengths of coax to ensure in-phase operation.

Real-World Terrain and Stacks

So far, the stacking examples shown have been for flat ground. Things can become a lot more complicated when you deal with real-world irregular terrains! See Chapter 3, The Effects of Ground, for a description of the *HFTA* (High Frequency Terrain Assessment) program that is included with this book.

Fig 47 shows the *HFTA*-computed 20-meter elevation responses towards Europe (at an azimuth of 45°) for three antennas at the N6BV/1 location in Windham, New Hampshire. Overlaid as a bar graph are the elevation-angle statistics for the path to all of Europe from New England (Massachusetts). The stack at 90/60/30 feet clearly covers all the angles needed best at 14 MHz. The N6BV 120-foot Yagi has a severe null in the region from about 7° to about 20°, with the deepest part of that null occurring at about 13° and is roughly comparable to the 90/60/30-foot stack between 2° to 7°.

In practice, the 120-foot Yagi was indeed comparable to the stack during morning openings to Europe on 20 meters, when the elevation angles are typically about 5°. In the New England afternoon, when the elevation angles typically rise to about 11°, the 120-foot Yagi was always distinctly inferior to the stack.

For reference, the response of a single 120-foot high Yagi over flat ground is also shown. Note that the N6BV 120-foot high Yagi has about 3 dB more gain at a 5° takeoff angle than does its flatland counterpart. This additional gain is due to the focusing effects of the local terrain, which had about a 3° downwards slope towards Europe.

Fig 48 shows the *HFTA*-computed 15-meter elevation responses towards Europe for the 90/60/30-foot stack at 90/60/30 feet at N6BV/1, compared to the same 120-foot high Yagi and a 90/60/30-foot stack, but this time over flat ground. Again, the N6BV/1 terrain towards Europe has a significant effect on the gain of the stack compared to that of an identical stack over flat ground. In fact, the peak gain of 20.1 dBi at a 4° elevation angle is close to moon-bounce levels.

OPTIMIZING OVER LOCAL TERRAIN

There are only a small number of possibilities to optimize an installation over local terrain:

- Change the antenna height(s) above ground.
- Stack two (or more) Yagis.
- Change the spacing between stacked Yagis.
- Move the tower back from a cliff (or a hill).
- BIP/BOP (Both In Phase/Both Out of Phase).

The *HFTA* program on the CD-ROM accompanying this book can be used, together with Digital Elevation Model (DEM) topographic data available on the Internet, to evaluate all these options.

SO NEAR, YET SO FAR

It is sometimes very surprising to compare elevation responses for different towers located at various points on the same property, particularly when that prop-

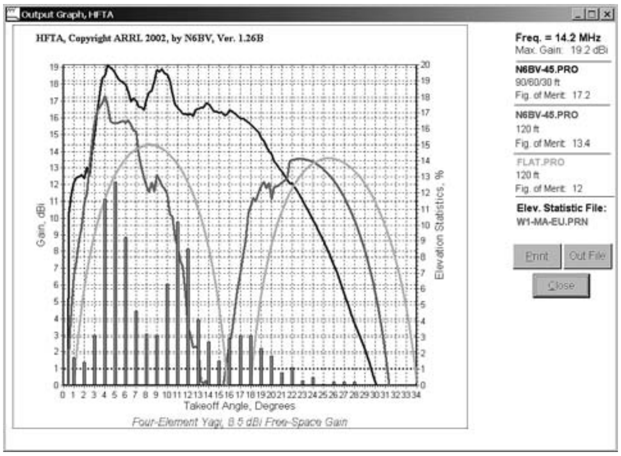

HFTA, Copyright ARRL 2002, by N6BV, Ver. 1.26B

Four-Element Yagi, 8.5 dBi Free-Space Gain

Freq. = 14.2 MHz
Max. Gain: 19.2 dBi

N6BV-45.PRO
90/60/30 ft
Fig. of Merit: 17.2

N6BV-45.PRO
120 ft
Fig. of Merit: 13.4

FLAT.PRO
120 ft
Fig. of Merit: 12

Elev. Statistic File:
W1-MA-EU.PRN

Fig 47—*HFTA* screen shot showing how complicated things become when real-world irregular terrain is analyzed. This is the 20-meter elevation pattern for the N6BV/1 station location in Windham, NH, for the 90/60/30-foot stack of triband TH7DX Yagis and a 4-element Yagi at 120 feet on the same tower. For comparison, the response of a 120-foot Yagi over flat ground is also included.

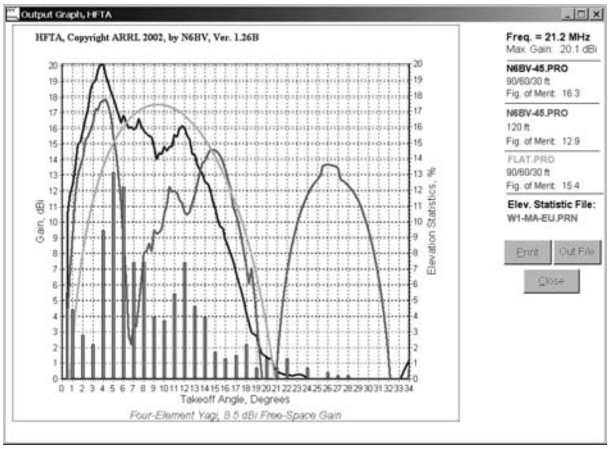

HFTA, Copyright ARRL 2002, by N6BV, Ver. 1.26B

Four-Element Yagi, 8.5 dBi Free-Space Gain

Freq. = 21.2 MHz
Max. Gain: 20.1 dBi

N6BV-45.PRO
90/60/30 ft
Fig. of Merit: 16.3

N6BV-45.PRO
120 ft
Fig. of Merit: 12.9

FLAT.PRO
90/60/30 ft
Fig. of Merit: 15.4

Elev. Statistic File:
W1-MA-EU.PRN

Fig 48—*HFTA* screen shot showing the 15-meter elevation pattern for the N6BV/1 station location in Windham, NH, for the 90/60/30-foot stack of triband TH7DX Yagis and a 4-element Yagi at 120 feet on the same tower. For comparison, the response of a 120-foot Yagi over flat ground is also included.

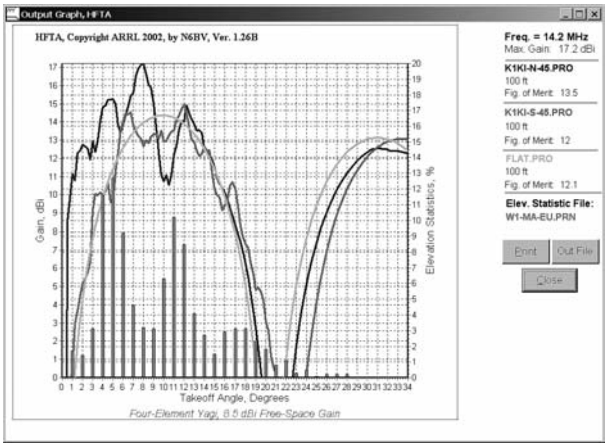

Freq. = 14.2 MHz
Max. Gain: 17.2 dBi

K1KI-N-45.PRO
100 ft
Fig. of Merit: 13.5

K1KI-S-45.PRO
100 ft
Fig. of Merit: 12

FLAT.PRO
100 ft
Fig. of Merit: 12.1

Elev. Statistic File:
W1-MA-EU.PRN

Fig 49—*HFTA* screen shot showing the 20-meter elevation pattern for K1KI's North and South towers, with 100-foot high 4-element Yagis pointing into Europe at an azimuth of 45°. The responses are surprisingly different for two towers separated by only 600 feet.

erty is located in the mountains. **Fig 49** shows the computed elevation responses for three 100-foot high 14-MHz Yagis over three terrains towards Europe: from the North tower at K1KI's location in West Suffield, Connecticut, from the South tower at K1KI, and over flat ground. The elevation response from the South tower follows that over flat ground well, while the response from the North tower is quite a bit stronger at low elevation angles—about 1.5 dB on average, as the Figure of Merit shows from *HFTA*.

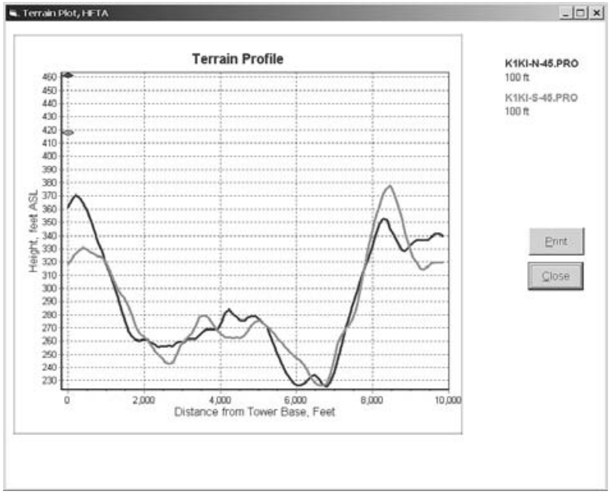

K1KI-N-45.PRO
100 ft

K1KI-S-45.PRO
100 ft

Fig 50—K1KI's terrain profiles for the North and South towers at an azimuth of 45° into Europe.

Fig 50 shows the reason why this happens—the terrain from the North tower slopes down quickly towards Europe, while the terrain from the South tower goes out almost 900 feet before starting to fall off. These two towers are about 600 feet apart.

Moxon Rectangle Beams

LB Cebik, W4RNL, has written extensively about the *Moxon rectangle*, an antenna invented by Les Moxon, G6XN, derived from a design by VK2ABQ. The Moxon rectangle beam takes less space horizontally than a conventional 2-element Yagi design, yet it offers nearly the same amount of gain and a superior front-to-back ratio. And as an additional benefit, the drive-point impedance is close to 50 Ω, so that it doesn't need a matching section.

For example, rather than a "wingspan" of 17 feet for the reflector in a conventional 2-element 10-meter Yagi, the Moxon rectangle is 13 feet wide, a saving of almost 25%. The Moxon rectangle W4RNL created for *The ARRL Antenna Compendium, Vol 6*, had an SWR less than 2:1 from 28.0 to 29.7 MHz, with a gain over ground of 11 dBi. It had a F/B of 15 dB at 28.0 MHz, more than 20 dB at 28.4 MHz, and 12 dB at 29.7 MHz.

The Moxon rectangle relies on controlling the spacing (hence controlling the coupling) between the ends of the driven element tips and the ends of the reflector tips, which are both bent toward each other. See **Fig 51**, which shows the general outline for W4RNL's 10-meter aluminum Moxon rectangle. The tips of the elements are kept a fixed distance from each other by PVC spacers. The closed rectangular mechanical assembly gives some rigidity to the design, keeping it stable in the wind. W4RNL described other Moxon rectangle designs using wire elements in June 2000 *QST*.

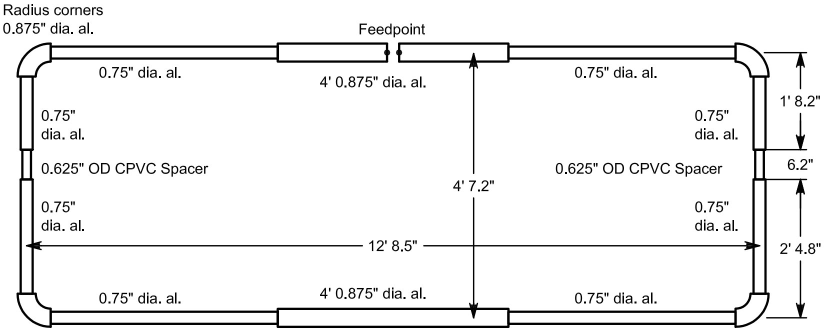

Fig 51—General outline of the 10-meter aluminum Moxon rectangle, showing tubing dimensions.

Chapter 12

Quad Arrays

Chapter 11, HF Yagi Arrays, discussed Yagi arrays as systems of approximately half-wave dipole elements that are coupled together mutually. You can also employ other kinds of elements using the same basic principles of analysis. For example, loops of various types may be combined into directive arrays. A popular type of parasitic array using loops is the *quad antenna*, in which loops having a perimeter of about one wavelength are used in much the same way as half-wave dipole elements in the Yagi antenna.

Clarence Moore, W9LZX, created the quad antenna in the early 1940s while he was at the Missionary Radio Station HCJB in Quito, Ecuador. He developed the quad to combat the effects of corona discharge at high altitudes. The problem at HCJB was that their large Yagi was literally destroying itself by melting its own element tips. This occurred due to the huge balls of corona it generated in the thin atmosphere of the high Andes mountains. Moore reasoned correctly that closed loop elements would generate less high voltage—and hence less corona—than would the high impedances at the ends of a half-wave dipole element.

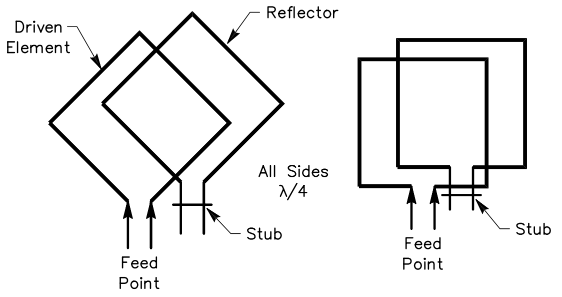

Fig 1—The basic two-element quad antenna, with driven-element loop and reflector loop. The driven loops are electrically one wavelength in circumference ($^1/_4$ wavelength on a side); the reflectors are slightly longer. Both configurations shown give horizontal polarization. For vertical polarization, the driven element should be fed at one of the side corners in the arrangement at the left, or at the center of a vertical side in the "square" quad at the right.

Fig 1 shows the original version of the two-element quad, with a driven element and a parasitic reflector. The square loops may be mounted either with the corners lying on horizontal and vertical lines, as shown at the left, or with two sides horizontal and two vertical (right). The feed points shown for these two cases will result in horizontal polarization, which is commonly used.

Since its inception, there has been controversy whether the quad is a better performer than a Yagi. Chapter 11 showed that the three main electrical performance parameters of a Yagi are gain, response patterns (front-to-rear ratio, F/R) and drive impedance/SWR. Proper analysis of a quad also involves checking all these parameters across the entire frequency range over which you intend to use it. Both a quad and a Yagi are classified as "parasitic, end-fire arrays." Modern antenna modeling by computer shows that monoband Yagis and quads with the same boom lengths and optimized for the same performance parameters have gains within about 1 dB of each other, with the quad slightly ahead of the Yagi.

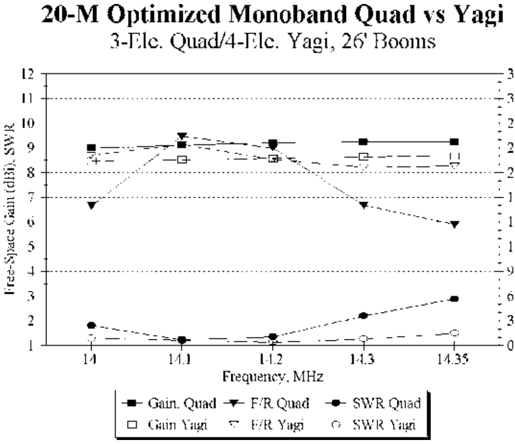

Fig 2—Comparison of gain, F/R and SWR over the 14.0 to 14.35-MHz range for an optimized three-element quad and an optimized three-element Yagi, both on 26-foot booms. The quad exhibits almost 0.5 dB more gain for the same boom length, but doesn't have as good a rearward pattern over the whole frequency range compared to the Yagi. This is evidenced by the F/R curve. The quad's SWR curve is also not quite as flat as the Yagi. The quad's design emphasizes gain more than the other two parameters.

Fig 2 plots the three parameters of gain, front-to-rear ratio (F/R) and SWR over the 14.0 to 14.35-MHz band for two representative antennas—a monoband three-element quad and a monoband four-element Yagi. Both of these have 26-foot booms and both are optimized for the best compromise of gain, F/R and SWR across the whole band.

While the quad in Fig 2 consistently exhibits about 0.5 dB more gain over the whole band, its F/R pattern toward the rear isn't quite as good as the Yagi's over that span of frequencies. This quad attains a maximum F/R of 25 dB at 14.1 MHz, but it falls to 17 dB at the bottom end of the band and 15 dB at the top. On the other hand, the Yagi's F/R stays consistently above 21 dB across the whole 20-meter band. The quad's SWR rises to just under 3:1 at the top end of the band, but stays below 2:1 from 14.0 to almost 14.3 MHz. The Yagi's SWR remains lower than 1.5:1 over the whole band.

The reason the Yagi in Fig 2 has more consistent responses for gain, F/R and SWR across the whole 20-meter band is that it has an additional parasitic element, giving two additional variables to play with—that is, the length of that additional element and the spacing of that element from the others on the boom.

Yagi advocates point out that it is easier to add extra elements to a Yagi, given the mechanical complexities of adding another element to a quad. Extra parasitic elements give a designer more flexibility to tailor all performance parameters over a wide frequency range. Quad designers have historically opted to optimize strictly for gain and, as stated before, they can achieve as much as 1 dB more gain than a Yagi with the same length boom. But in so doing, a quad designer typically has to settle for front-to-rear patterns that are peaked over more narrow frequency ranges. The 20-meter quad plots in Fig 2 actually represent an even-handed approach, where the gain is compromised slightly to obtain a more consistent pattern and SWR across the whole band.

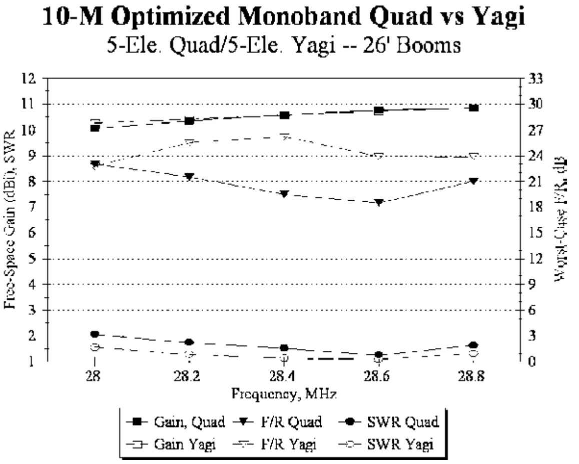

Fig 3—Comparison of gain, F/R and SWR over the 28.0 to 28.8-MHz range for an optimized five-element quad and an optimized five-element Yagi, both on 26-foot booms. The gain advantage of the quad is about 0.25 dB at the low end of the band. The F/R is more peaked in frequency for the quad, however, than the Yagi.

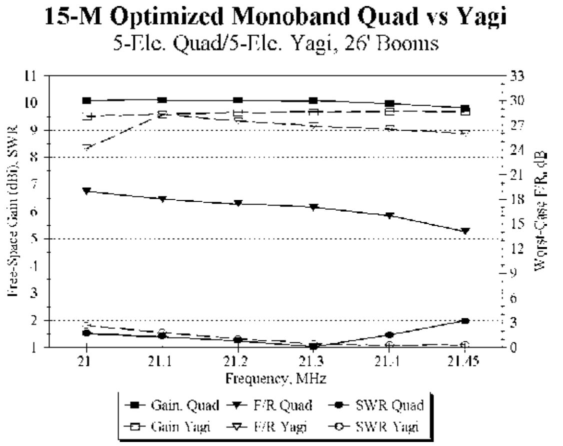

Fig 4—Comparison of gain, F/R and SWR over the 21.0 to 21.45-MHz range for an optimized 5-element quad and optimized 5-element Yagi, both on 26-foot booms. The quad enjoys a gain advantage of about 0.5 dB over most of the band. Its rearward pattern is not as good as the Yagi, which remains higher than 24 dB across the whole range, compared to the quad, which remains in the 16-dB average range.

Fig 3 plots gain, F/R and SWR for two 10-meter monoband designs: a five-element quad and a five-element Yagi, both placed on 26-foot booms. The quad now has the same degrees of freedom as the Yagi, and as a consequence the pattern and SWR are more consistent across the range from 28.0 to 28.8 MHz. The quad's F/R remains above about 18.5 dB from 28.0 to 28.8 MHz. Meanwhile, the Yagi maintains an F/R of greater than 22 dB over the same range, but has almost 0.8 dB less gain compared to the quad at the low end of the band, eventually catching up at the high end of the band. The SWR for the quad is just over 2:1 at the bottom of the band, but remains less than 2:1 up to 28.8 MHz. The SWR on the Yagi remains less than 1.6:1 over the whole band.

Fig 4 shows the performance parameters for two 15-meter monoband designs: a five-element quad and a five-element Yagi, both on 26-foot booms. The quad is still the leader in gain, but has a less optimal rearward pattern and a somewhat less flat SWR curve than the Yagi. One thing should be noted in Figs 2 through 4. The F/R pattern on the Yagi is largely determined by the response at the 180° point, directly in back of the frontal lobe. This point is usually referred to when discussing the "front-to-back ratio."

The quad on the other hand has what a sailor might term "quartering lobes" (referring to the direction back towards the "quarterdeck" at the stern of a sailing vessel) in the rearward pattern. These quartering lobes are often

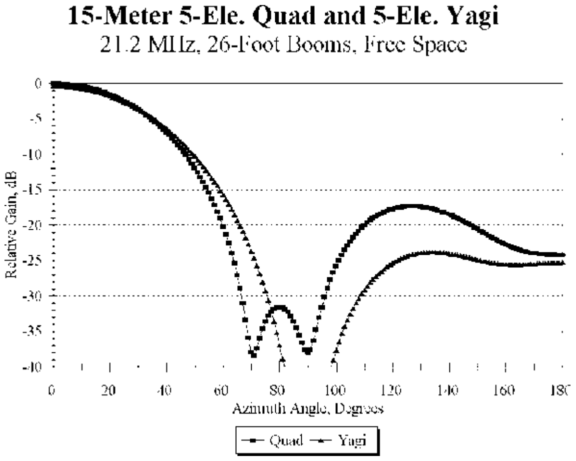

Fig 5—Comparing the pattern of the 15-meter quad and Yagi shown in Fig 4. The quad has a slightly narrower frontal beamwidth (it has 0.5 dB more gain than the Yagi), but has higher "rear quartering" sidelobes at about 125° (with a twin sidelobe, not shown, at 235°). These sidelobes limit the worst-case front-to-rear (F/R) to about 17 dB, while the F/B (at 180°, directly at the back of the quad) is more than 24 dB for each antenna.

Table 1
Dimensions for Optimized Monoband Quads in Figs 2, 3 and 4, on 26-Foot Booms

	14.2 MHz	*21.2 MHz*	*28.4 MHz*
Reflector	73′ 9"	49′ 6"	37′ 3"
R-DE Spacing	17′ 8"	7′	6′ 4"
Driven Element	71′ 8"	47′ 6"	35′ 9"
DE-D1 Spacing	8′ 3"	5′	5′ 6"
Director 1	68′ 7"	46′ 8"	34′ 8"
D1-D2 Spacing	—	6′ 8"	6′ 9"
Director 2	—	46′ 10"	35′ 2"
D2-D3 Spacing	—	7′ 4"	7′ 5"
Director 3	—	45′ 8"	34′ 2"
Feed method	Direct 50 Ω	Direct 50 Ω	Direct 50 Ω

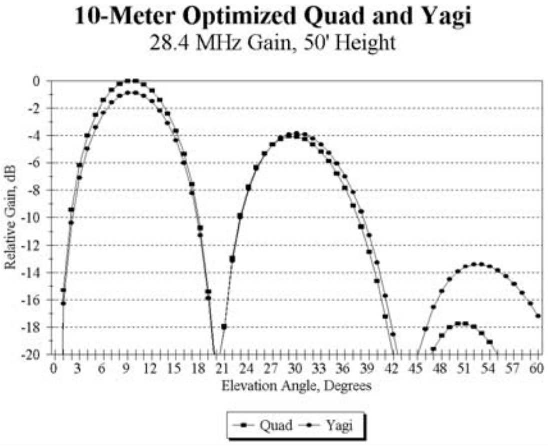

Fig 6—A comparison on 10 meters between an optimized five-element quad and an optimized five-element Yagi, both mounted 50 feet high over flat ground and both employing 26-foot booms. There is no appreciable difference in the peak elevation angle for either antenna. In other words, a quad does not have an appreciable elevation-angle advantage over a Yagi mounted at the same boom height. Note that the quad achieves its slightly higher gain by taking energy from higher-angle lobes and concentrating that energy in the main elevation lobe. This is a process that is similar to what happens with stacked Yagis.

worse than the response at 180°, directly in back of the main beam. **Fig 5** overlays the free-space E-Field responses of the 15-meter quad and Yagi together. At 21.2 MHz, the quad actually has a front-to-back ratio (F/B) of about 24 dB, excellent in anyone's book. The Yagi at 180° has a F/B of about 25 dB, again excellent.

However, at an azimuth angle of about 125° (and at 235° azimuth on the other side of the main lobe) the quad's "quartering lobe" is down only some 17 dB, setting the worst-case F/R at 17 dB also. As explained in Chapter 11, the reason F/R is more important than just the F/B is that on receive signals can come from any direction, not just from directly behind the main beam.

Table 1 lists the dimensions for the three computer-optimized monoband quads shown in Figs 2, 3 and 4.

Is a Quad Better at Low Heights than a Yagi?

Another belief held by some quads enthusiasts is that they need not be mounted very high off the ground to give excellent DX performance. Quads are somehow supposed to be greatly superior to a Yagi at the same height above ground. Unfortunately, this is mainly wishful thinking.

Fig 6 compares the same two 10-meter antennas as in Fig 3, but this time with each one mounted on a 50-foot tower over flat ground, rather than in theoretical free space. The quad does indeed have slightly more gain than a Yagi with the same boom length, as it has in free space. This is evidenced by the very slight compression of the quad's main lobe, but is more obvious when you look at the third lobe, which peaks at about 53° elevation. In effect, the quad squeezes some energy out of its second and third lobes and adds that to the first lobe. However, the difference in gain compared to the Yagi is only 0.8 dB for this particular quad design at a 9° elevation angle. And while it's true that every dB counts, you can also be certain that on the air you wouldn't be able to tell the difference between the two antennas. After all, a 10- to 20-dB variation in the level of signals is pretty common because of fading at HF.

Multiband Quads

On the other hand, one of the valid reasons quads have remained popular over the years is that antenna homebrewers can build multiband quads far more easily than they can construct multiband Yagis. In effect, all you have to do with a quad is add more wire to the existing support arms. It's not quite as simple as that, of course, but the idea of ready expandability for other bands is very appealing to experimenters.

Like the Yagi, the quad does suffer from interactions between wires of different frequencies, but the degree of interaction between bands is usually less for a quad. The higher-frequency bands are the ones that often suffer most from any interaction, for both Yagis and quads. For example, the 10- and 15-meter bands are usually the ones affected most by nearby 20-meter wires in a triband quad, while the 20-meter elements are not affected by the 10- or 15-meter elements.

Modern computer modeling software can help you counteract at least some of the interaction by allowing you to do virtual "retuning" of the quad on the computer screen — rather than clinging precariously to your tower fiddling with wires. However, the programs (such as *NEC-2* or *EZNEC*) that can model three-dimensional wire antennas such as quads typically run far more slowly than those designed for monoband Yagis (such as *YW* included with this book). This makes optimizing rather tedious, but you use the same considerations for tradeoffs between gain, pattern (F/R) and SWR over the operating bandwidth as you do with monoband Yagis.

CONSTRUCTING A QUAD

The parasitic element shown in Fig 1 is tuned in much the same way as the parasitic element in a Yagi antenna. That is, the parasitic loop is tuned to a lower frequency than the driven element when the parasitic is to act as a reflector, and to a higher frequency when it is to act as a director. Fig 1 shows the parasitic element with an adjustable tuning stub, a convenient method of tuning since the resonant frequency can be changed simply by changing the position of the shorting bar on the stub. In practice, it has been found that the length around the loop should be approximately 3.5% greater than the self-resonant length if the element is a reflector, and about 3.0% shorter than the self-resonant length if the parasitic element is a director. Approximate formulas for the loop lengths in feet are:

$$\text{Driven Element} = \frac{1008}{f_{\text{MHz}}}$$

$$\text{Reflector} = \frac{1045}{f_{\text{MHz}}}$$

$$\text{Director} = \frac{977}{f_{\text{MHz}}}$$

These are valid for quad antennas intended for operation below 30 MHz and using uninsulated #14 stranded copper wire. At VHF, where the ratio of loop circumference to conductor diameter is usually relatively small, the circumference must be increased in comparison to the wavelength. For example, a one-wavelength loop constructed of $^1/_4$-inch tubing for 144 MHz should have a circumference about 2% greater than in the above equation for the driven element.

Element spacings on the order of 0.14 to 0.2 free-space wavelengths are generally used. You would employ the smaller spacings for antennas with more than two elements, where the structural support for elements with larger spacings tends to become challenging. The feed-point impedances of antennas having element spacings on this order have been found to be in the 40- to 60-Ω range, so the driven element can be fed directly with coaxial cable with only a small mismatch.

For spacings on the order of 0.25 wavelength (physically feasible for two elements, or for several elements at 28 MHz) the impedance more closely approximates the impedance of a driven loop alone—that is, 80 to 100 Ω. The feed methods described in Chapter 26 can be used, just as in the case of the Yagi.

Making It Sturdy

The physical sturdiness of a quad is directly proportional to the quality of the material used and the care with which it is constructed. The size and type of wire selected for use with a quad antenna is important because it will determine the capability of the spreaders to withstand high winds and ice. One of the more common problems confronting the quad owner is that of broken wires. A solid conductor is more apt to break than stranded wire under constant flexing conditions. For this reason, stranded copper wire is recommended. For 14-, 21- or 28-MHz operation, #14 or #12 stranded wire is a good choice. Soldering of the stranded wire at points where flexing is likely to occur should be avoided.

You may connect the wires to the spreader arms in many ways. The simplest method is to drill holes through the fiberglass at the appropriate points on the arms and route the wires through the holes. Soldering a wire loop across the spreader, as shown later, is recommended. However, you should take care to prevent solder from flowing to the corner point where flexing could break it.

While a boom diameter of 2 inches is sufficient for smaller quads using two or even three elements for 14, 21 and 28 MHz, when the boom length reaches 20 feet or longer a 3-inch diameter boom is highly recommended. Wind creates two forces on the boom, vertical and horizontal. The vertical load on the boom can be reduced with a guy-wire truss cable. The horizontal forces on the boom are more difficult to relieve, so 3-inch diameter tubing is desirable.

Generally speaking, three grades of material can be used for quad spreaders. The least expensive material is bamboo. Bamboo, however, is also the weakest material normally used for quad construction. It has a short life, typically only three or four years, and will not withstand a harsh climate very well. Also, bamboo is heavy in contrast to fiberglass, which weighs only about a pound per 13-foot length. Fiberglass is the most popular type of spreader material, and will withstand normal winter climates. One step beyond the conventional fiberglass arm is the pole-vaulting arm. For quads designed to be used on 7 MHz, surplus "rejected" pole-vaulting poles are highly recommended. Their ability to withstand large amounts of bending is very desirable. The cost of these poles is high, and they are difficult to obtain. See Chapter 21 for dealers and manufacturers of spreaders.

Diamond or Square?

The question of how to orient the spreader arms has been raised many times over the years. Should you mount the loops in a diamond or a square configuration? Should one set of spreaders be horizontal to the earth as shown in Fig 1 (right), or should the wire itself be horizontal to the ground (spreaders mounted in the fashion of an X) as shown in Fig 1 (left)? From the electrical point of view, there is not enough difference in performance to worry about.

From the mechanical point of view there is no question which version is better. The diamond quad, with the associated horizontal and vertical spreader arms, is capable of holding an ice load much better than a system where no vertical support exists to hold the wire loops upright. Put another way, the vertical poles of a diamond array, if sufficiently strong, will hold the rest of the system erect. When water droplets are accumulating and forming into ice, it is very reassuring to see water running down the wires to a corner and dripping off, rather than just sitting there on the wires and freezing. The wires of a loop (or several loops, in the case of a multiband antenna) help support the horizontal spreaders under a load of ice. A square quad will droop severely under heavy ice conditions because there is nothing to hold it up straight.

Of course, in climates where icing is not a problem, many amateurs point out that they like the aesthetics of the square configuration. There are thousands of square-configuration quads in temperate areas around the world.

Another consideration will enter into your choice of orientation for a quad. You must mount a diamond quad somewhat higher on the mast or tower than for an equivalent square array, just to keep the bottom spreader away from the tower guys when you rotate the antenna.

Two Multiband Quads

This section describes two multiband quad designs. The first is a large triband 20/15/10-meter quad built on a 26-foot boom made of 3-inch irrigation tubing. This antenna has three elements on 20 meters, four elements on 15 meters, and five elements on 10 meters. **Fig 7** shows a photograph of the five-element triband quad.

The second project is a compact two-element triband quad on an 8-foot boom that covers 20, 17, 15, 12 and 10 meters. We call this a "pentaband" quad since it covers five bands. This antenna uses five concentric wire loops mounted on each of the two sets of spreaders. Either antenna may be constructed in a diamond or square configuration.

While the same basic construction techniques are employed for both multiband quads, the scale of the larger triband antenna makes it a far more ambitious undertaking! The large quad requires a strong tower and a rugged rotator. It also requires a fair amount of real estate in order to raise the quad to the top of the tower without entangling trees or other antennas.

A FIVE-ELEMENT, 26-FOOT BOOM TRIBAND QUAD

Five sets of element spreaders are used to support the three elements used on 14 MHz, four elements on 21 MHz and five elements on 28-MHz. We chose to use four elements on 15 meters in this design (rather than the five we could have been employed on this length of boom) because the difference in optimized performance wasn't great enough to warrant the extra complexity of using five elements. The dimensions are listed in **Table 2**, and are designed for center frequencies of 14.175, 21.2 and 28.4 MHz.

The spacing between elements has been chosen to provide good compromises in performance consistent with boom length and mechanical construction. You can see that the element spacings for 20 meters are quite different from those for the optimized monoband design. This is because the same set of spreaders is used for all

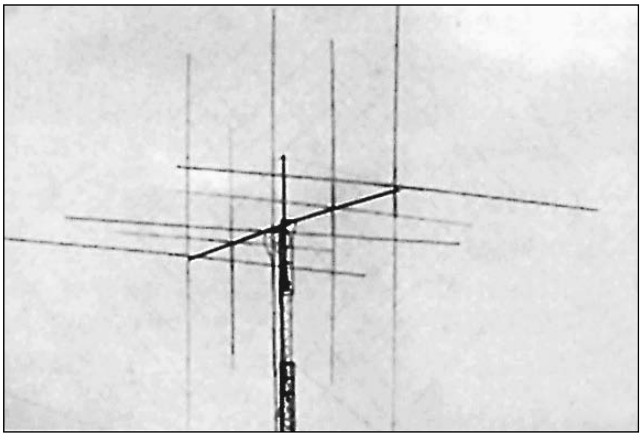

Fig 7—Photo of the three-band, five-element quad antenna.

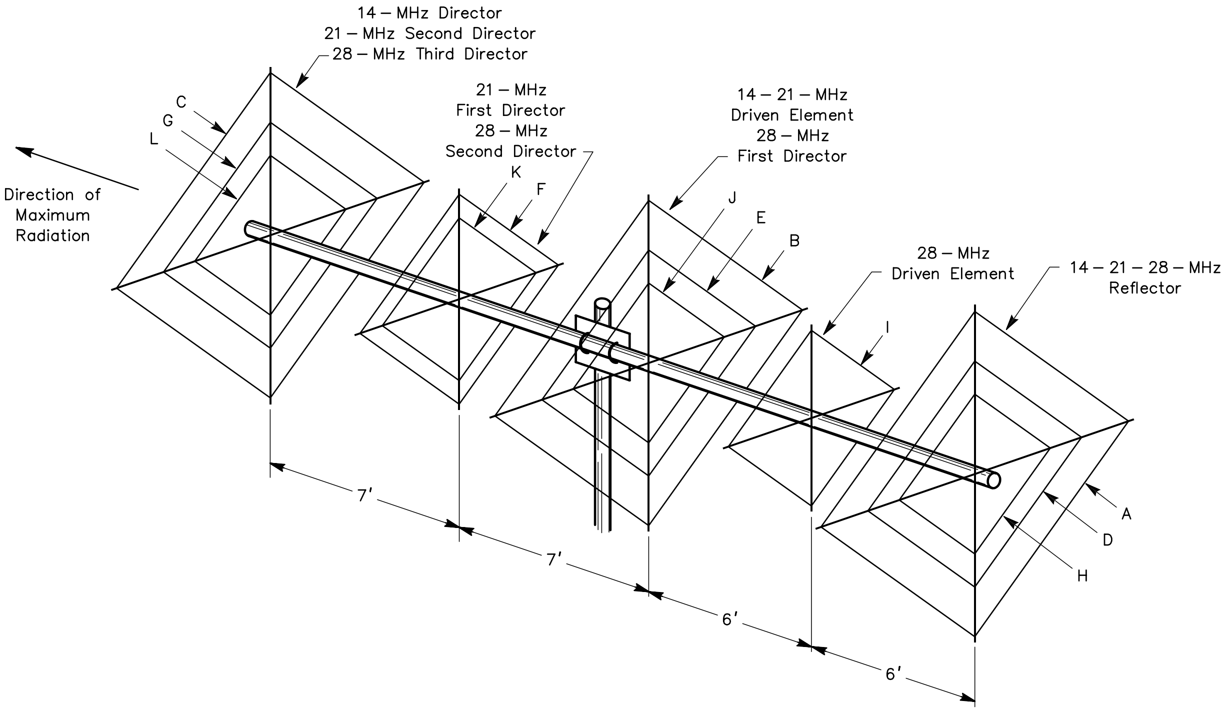
14 – MHz Director
21 – MHz Second Director
28 – MHz Third Director
21 – MHz
First Director
28 – MHz
Second Director
14 – 21 – MHz
Driven Element
28 – MHz
First Director
C
G
L
Direction of
Maximum
Radiation
K
F
J
E
B
28 – MHz
Driven Element
14 – 21 – 28 – MHz
Reflector
I
A
D
H
7′
7′
6′
6′

Fig 8—Layout for the three-band, five-element quad, not drawn to scale. See Table 2 for dimensions.

three bands on three out of the five elements, and the higher-frequency bands dictate the spacing because they are more critical.

Each of the parasitic loops is closed (ends soldered together) and requires no tuning. **Fig 8** shows the physical layout of the triband quad. **Fig 9** plots the computed free-space gain, front-to-rear ratio and SWR response across the 20-meter band. With only a few degrees of freedom in tuning and spacing of the three elements, it is impossible to spread the response out to cover the entire 20-meter band. The compromise design results in a rearward pattern that varies from a worst-case of just under 10 dB at the high end of the band, to a peak F/R of just under 19 dB at 14.2 MHz, in the phone portion of the band. The F/R is about 11 dB at the low end of the band.

Table 2
Three-Band Five-Element Quad on 26-Foot Boom

	14.15 MHz	*21.2 MHz*	*28.4 MHz*
Reflector	72′ 6"	49′ 4"	36′ 8"
R-DE Spacing	12′	12′	6′
Driven Element	71′	47′ 6"	35′ 4"
DE-D1 Spacing	14′	7′	6′
Director 1	68′ 6"	46′ 8"	34′ 8"
D1-D2 Spacing	—	14′	7′
Director 2	—	46′ 5"	34′ 8"
D2-D3 Spacing	—	—	7′
Director 3	—	—	34′
Feed method	Direct 50 Ω	Direct 50 Ω	Direct 50 Ω

The SWR remains under 3:1 for the entire 20-meter band, rising to 2.8:1 at the high end. The feed system for this triband quad consists of three separate 50-Ω coax

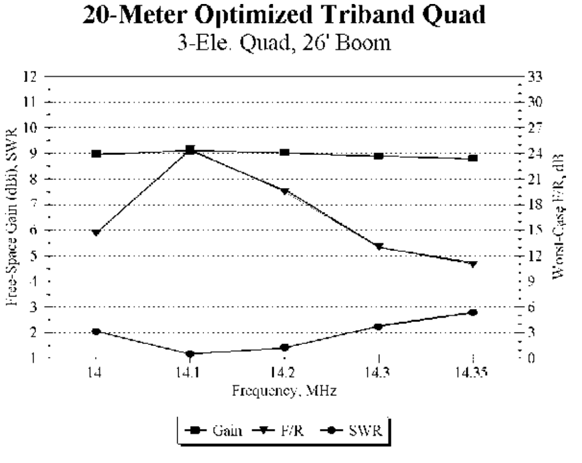

Fig 9—Computed performance of the triband, five-element quad over the 20-meter band. The direct 50-Ω feed system holds the SWR below 2.8: 1 across the whole band. This could be improved with a gamma-match system tuned to 14.1 MHz if the builder really desires a low SWR. The F/R peaks at 14.1 MHz and remains above 10 dB across the whole band.

lines, one per driven element, together with a relay switchbox mounted to the boom so that a single coax can be used back to the operating position. Each feed line uses a ferrite-bead balun to control common-mode currents and preserve the radiation pattern and each coax going to the switchbox is cut to be an electrical three-quarter wavelength on 15 meters. This presents a short at the unused driven elements since modeling indicated that the 15-meter band is adversely affected by the presence of the 20-meter driven element if it is left open-circuited. If you use RG-213 coax, the $^{3}/_{4}$-λ electrical length of each feed line is 23 feet long at 21.2 MHz. This is sufficient physical length to reach each driven element from the switchbox.

Fig 10 shows the free-space response for the 15-meter band. The rearward response is roughly 15 dB across the band. This is a result of the residual interaction between the 20-meter elements on 15 meters, and no further tuning could improve the F/R. Note how flat the SWR curve is. This SWR characteristic is what gives the quad the reputation of being "wideband." A flat SWR curve, however, is not necessarily a good indicator of optimal performance for directional antennas like quads or Yagis, particularly multiband designs where compromises must be made by physical necessity.

Fig 11 shows the characteristics of the 10-meter portion of the two-element triband quad. The response favors the low-phone band, with the F/R falling to about 12 dB at the low end of the frequency range and rising to just about 23 dB at 28.4 MHz. The SWR curve is once again relatively flat across the major portion of the band up to 28.8 MHz.

Construction

The most obvious problem related to quad antennas is the ability to build a structurally sound system. If high winds or heavy ice are a normal part of the environment, special precautions are necessary if the antenna is to survive a winter season. Another stumbling block for would-be quad builders is the installation of a three-dimensional system (assuming a Yagi has only two important dimensions) on top of a tower—especially if the tower needs guy wires for support. With proper planning, however, many of these obstacles can be overcome. For example, a tram system may be used.

Both multiband quad arrays use fiberglass spreaders (see Chapter 21 for suppliers). Bamboo is a suitable substitute (if economy is of great importance). However, the additional weight of the bamboo spreaders over fiberglass is an important consideration. A typical 12-foot bamboo pole weighs about 2 pounds; the fiberglass type weighs less than a pound. By multiplying the difference times 8 for a two-element array, times 12 for a three-element antenna, and so on, it quickly becomes apparent that fiberglass is worth the investment if weight is an important factor. Properly treated, bamboo has a useful life of three or four years, while fiberglass life is probably 10 times longer.

Spreader supports (sometimes called *spiders*) are available from many different manufacturers. If the builder is keeping the cost at a minimum, he should consider building his own. The expense is about half that of a commercially manufactured equivalent and, according to some authorities, the homemade arm supports described below are less likely to rotate on the boom as a result of wind pressure.

A 3-foot length of steel angle stock, 1 inch per side, is used to interconnect the pairs of spreader arms. The steel is drilled at the center to accept a muffler clamp of sufficient size to clamp the assembly to the boom. The fiberglass is clamped to the steel angle stock with auto-

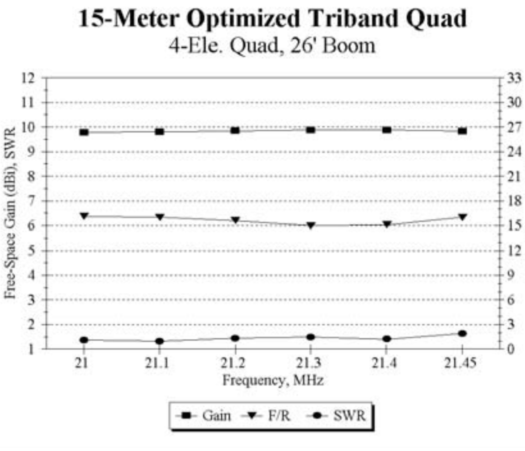

Fig 10—Computed performance of the triband, five-element quad over the 15-meter band. There is some degree of interaction with the 20-meter elements, limiting the worst-case F/R to about 15 dB. The gain and SWR curves are relatively flat across the band.

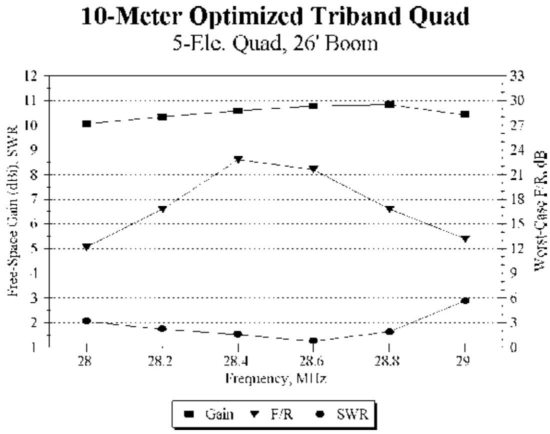

Fig 11—Computed performance of the triband, five-element quad over the 10-meter band. The F/R is higher than 12 dB across the band from 28.0 to 29.0 MHz, but the SWR rises at the top end of the band beyond 2:1. The free-space gain is higher than 10 dBi across the band.

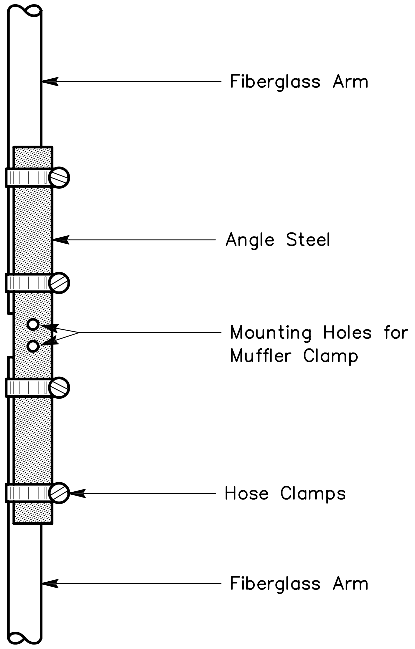
Fiberglass Arm
Angle Steel
Mounting Holes for
Muffler Clamp
Hose Clamps
Fiberglass Arm

Fig 12—Details of one of two assemblies for a spreader frame. The two assemblies are joined back-to-back to form an X with a muffler clamp mounted at the position shown.

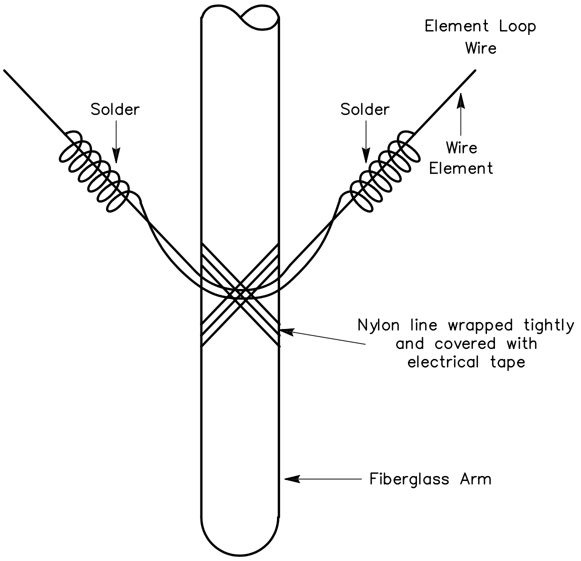
Element Loop
Wire
Solder
Solder
Wire
Element
Nylon line wrapped tightly
and covered with
electrical tape
Fiberglass Arm

Fig 14—An alternative method of assembling the wire of a quad loop to the spreader arm.

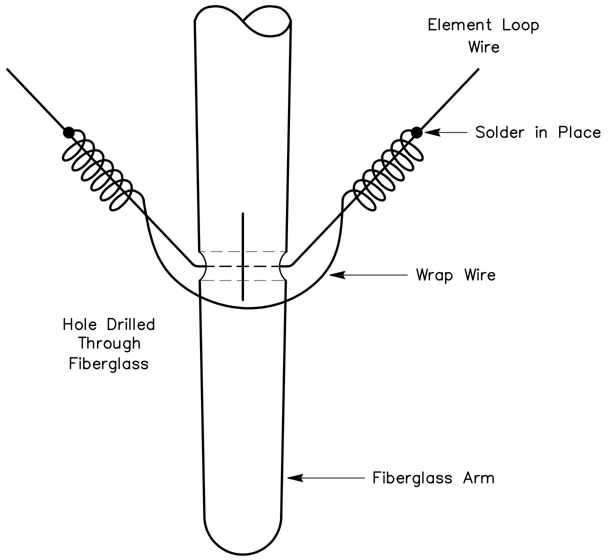
Element Loop
Wire
Solder in Place
Wrap Wire
Hole Drilled
Through
Fiberglass
Fiberglass Arm

Fig 13—A method of assembling a corner of the wire loop of a quad element to the spreader arm.

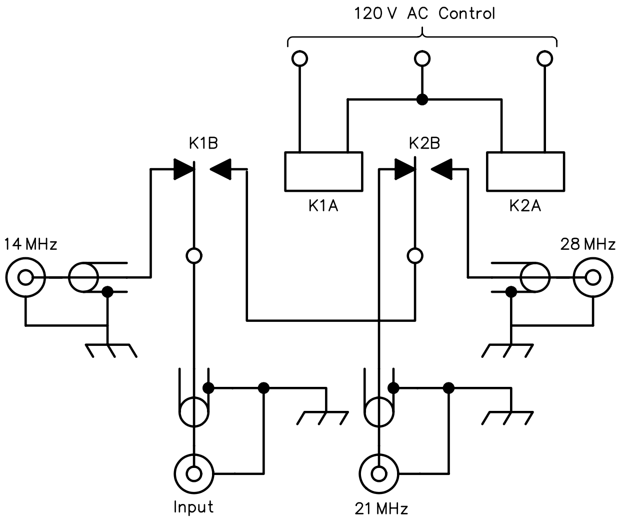
120 V AC Control
K1B
K2B
K1A
K2A
14 MHz
28 MHz
Input
21 MHz

Fig 15—Suitable circuit for relay switching of bands for the three-band quad. A three-wire control cable is required. K1, K2—any type of relay suitable for RF switching, coaxial type not required (Potter and Brumfeld MR11A acceptable; although this type has double-pole contacts, mechanical arrangements of most single-pole relays make them unacceptable for switching of RF).

motive hose clamps, two per pole. Each quad-loop spreader frame consists of two assemblies of the type shown in **Fig 12**.

Connecting the wires to the fiberglass can be done in a number of different ways. Holes can be drilled at the proper places on the spreader arms and the wires run through them. A separate wrap wire should be included at the entry/exit point to prevent the loop from slipping. Details are presented in **Fig 13**. Some amateurs have experienced cracking of the fiberglass, which might be a result of drilling holes through the material. However, this seems to be the exception rather than the rule. The model described here has no holes in the spreader arms; the wires are attached to each arm with a few layers of

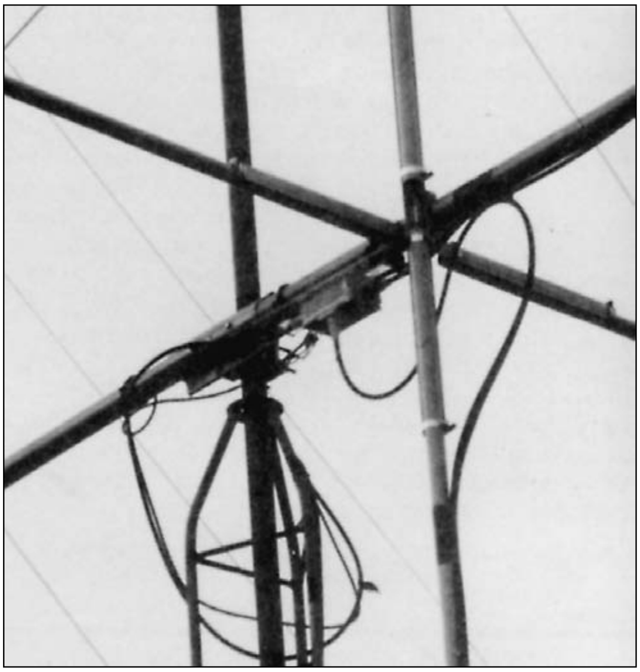

Fig 16—The relay box is mounted on the boom near the center. Each of the spreader-arm fiberglass poles is attached to steel angle stock with hose clamps.

plastic electrical tape and then wrapped approximately 20 times in a crisscross fashion with $^{1}/_{8}$-inch diameter nylon string, followed by more electrical tape for UV protection, as shown in **Fig 14**.

The wire loops are left open at the bottom of each driven element where the feed-line coaxes are attached. All of the parasitic elements are continuous loops of wire; the solder joint is at the base of the diamond.

Although you could run three separate coax cables down to the shack, we suggest that you install a relay box at the center of the boom. A three-wire control system may be used to apply power to the proper relay for changing bands. The circuit diagram of a typical configuration is presented in **Fig 15** and its installation is shown in **Fig 16**.

Every effort must be placed upon proper construction if you want to have freedom from mechanical problems. Hardware must be secure or vibration created by the wind may cause separation of assemblies. Solder joints should be clamped in place to keep them from flexing, which might fracture a connection point.

A TWO-ELEMENT, 8-FOOT BOOM PENTABAND QUAD

This two-element pentaband (20/17/15/12/10-meter) quad uses the same construction techniques as its big brother above. Since only two elements are used, the boom can be less robust for this antenna, at 2 inches diameter rather than 3 inches. Those who like really rugged antennas can still use the 3-inch diameter boom, of course.

Table 3 lists the element dimensions for the pentaband quad. The following plots show the performance for each of the five bands covered. The feed system for the pentaband quad uses five, direct 50-Ω coaxes, one to each driven element. These five coaxes are cut to be $^{3}/_{4}$-λ electrically on 10 meters (17 feet, 2 inches for RG-213 at 28.4 MHz). In this design the 10-meter band is the one most affected by the presence of the other driven elements if they are left unshorted. The $^{3}/_{4}$-λ lines open-circuited at the switchbox are long enough physically to reach all elements from a centrally mounted switchbox. This length assumes that the switchbox open-circuits the unused coaxes. If the switchbox short-circuits unused coaxes (as several commercial switchboxes do), then use $^{1}/_{2}$-λ long lines to feed all five driven elements (11 feet,

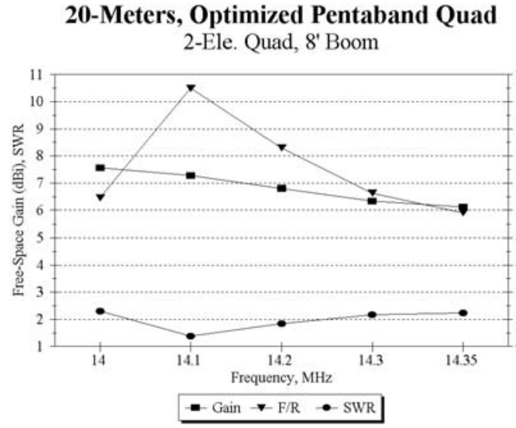

Fig 17—Computed performance of the pentaband two-element quad on 20 meters. With the simple direct-feed system, the SWR rises to about 2.3:1 at the low end of the band. A gamma match can bring the SWR down to 1:1 at 14.1 MHz, if desired.

Table 3
Five-Band Two-Element Quad on 8-Foot Boom

	14.2 MHz	*18.1 MHz*	*21 MHz*	*24.9 MHz*	*28.4 MHz*
Reflector	72′ 4"	56′ 4"	48′ 6"	40′ 11$^{1}/_{4}$"	37′ 5$^{1}/_{2}$"
R-DE Spacing	8′	8′	8′	8′	8′
Driven Element	69′ 10$^{1}/_{2}$"	54′ 10$^{1}/_{2}$"	46′ 7"	39′ 10$^{1}/_{2}$"	34′ 6"

5 inches for RG-213 at 28.4 MHz).

The SWR curves do not necessarily go down to 1:1 because of this simple, direct feed system. If anyone is bothered by this, of course they can always implement individual matching systems, such as gamma matches. Most amateurs would agree that such a degree of complexity is not warranted. The worst-case SWR is less than 2.3:1 on each band, even with direct feed on 20 meters. With typical lengths of coaxial feed line from the shack to the switchbox at the antenna, say 100 feet of RG-213, the SWR at the transmitter would be less than 2.0:1 on all bands due to losses in the feed line.

Fig 17 shows the computed responses for the pentaband quad over the 20-meter band. With only two degrees of freedom (spacing and element tuning) there is not much that can be done to spread the response out over the entire 20-meter band. Nonetheless, the performance over the band is still pretty reasonable for an antenna this small. The F/R pattern peaks at 19 dB at 14.1 MHz and falls to about 10 dB at either end of the band. The free-space gain varies from about 7.5 dBi to just above 6 dBi, comparable to a short-boom three-element Yagi. The SWR curve remains below 2.3:1 across the band. If you were to employ a gamma match tuned at 14.1 MHz, you could limit the peak SWR to less than 2.0:1, and this would still occur at 14.0 MHz.

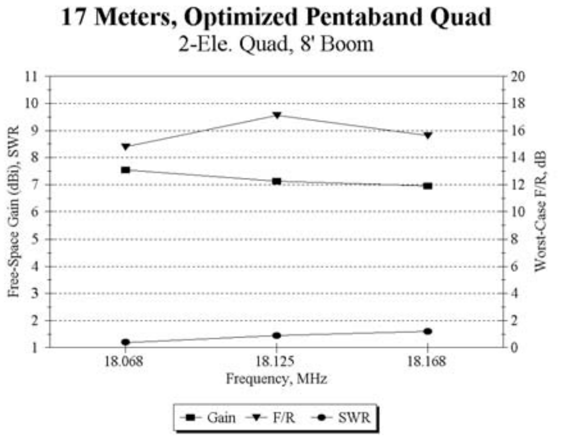

Fig 18—Computed performance of the pentaband two-element quad on 17 meters. There is some interaction with the other elements, but overall the performance is satisfactory on this band.

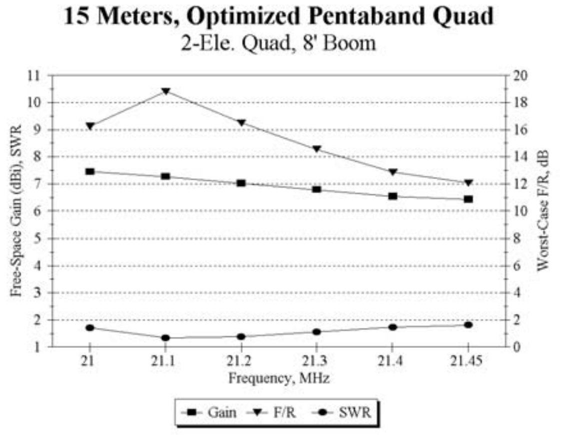

Fig 19—Computed performance of the pentaband two-element quad on 15 meters. The performance is acceptable across the whole band.

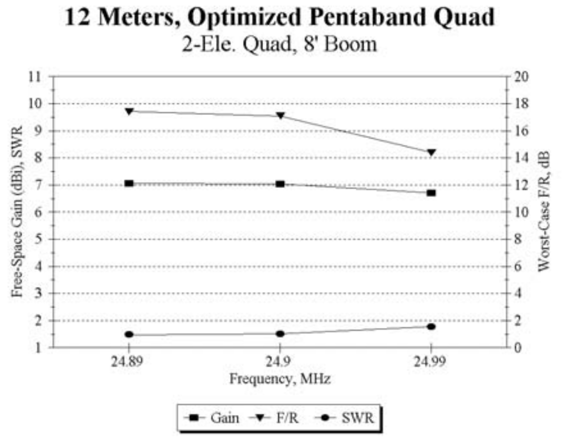

Fig 20—Computed performance of the pentaband two-element quad on 12 meters.

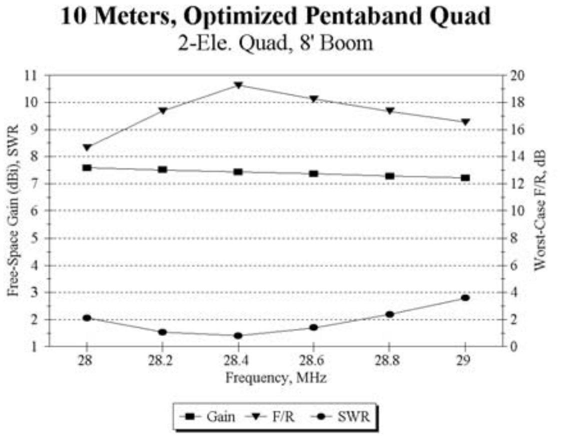

Fig 21—Computed performance of the pentaband two-element quad on 10 meters. The SWR curve is slightly above the target 2:1 at the low end of the band and rises to about 2.2:1 at 28.8 MHz. This unlikely to be a problem, even with rigs with automatic power-reduction due to SWR, since the SWR at the input of a typical coax feed line will be lower than that at the antenna due to losses in the line.

On 17 meters, **Fig 18** shows that the other elements are affecting 18 MHz, even with element-length optimization. Careful examination of the current induced on the other elements shows that the 20-meter driven element is interacting on 18 MHz, deteriorating the pattern and gain slightly. Even still, the performance on 17 meters is reasonable, especially for a five-band quad on an 8-foot boom.

On 15 meters, the interactions seems to have been contained, as **Fig 19** demonstrates. The F/R peaks at 21.1 MHz, at 19 dB and remains better than 12 dB past the top of the band. The SWR curve is low across the whole band.

On 12 meters, the interaction between bands is minor, leading to the good results shown in **Fig 20**. The SWR change across this band is quite flat, which isn't surprising given the narrow bandwidth of the 12-meter band.

On 10 meters, the interaction seems to have been tamed well by computer-tuning of the elements. The F/R remains higher than about 14 dB from 28 to 29 MHz. The SWR remains below 2.2:1 up to about 28.8 MHz, while the gain is relatively flat across the band at more than 7.2 dBi in free space. See **Fig 21**.

Overall, this pentaband quad is physically compact and yet it provides good performance across all five bands. It is competitive with commercial Log Periodic Dipole Array (LPDA) designs and triband Yagi designs that employ longer booms.

BIBLIOGRAPHY

Source material and more extended discussions of the topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

P. S. Carter, C. W. Hansell and N. E. Lindenblad, "Development of Directive Transmitting Antennas by R.C.A. Communications," *Proc. IRE*, Oct 1931.

C. Cleveland, "More'n One Way to Switch an Antenna," Technical Correspondence, *QST*, Nov 1986.

D. Cutter, "Simple Switcher," *73*, May 1980.

D. DeMaw, "A Remote Antenna Switcher for HF," *QST*, Jun 1986.

M. G. Knitter, Ed., *Loop Antennas—Design and Theory* (Cambridge, WI: National Radio Club, 1983).

H. Landskov, W7KAR, "Evolution of a Quad Array," *QST*, Mar 1977.

J. Lindsay, "Quads and Yagis," *QST*, May 1968.

F. E. Terman, *Radio Engineering*, 3rd ed. (New York: McGraw-Hill Book Co, 1947).

E. M. Williams, "Radiating Characteristics of Short-Wave Loop Aerials," *Proc. IRE*, Oct 1940.

Chapter 13

Long-Wire and Traveling-Wave Antennas

The power gain and directive characteristics of electrically long wires (that is, wires that are long in terms of wavelength), as described in Chapter 2, make them useful for long-distance transmission and reception on the higher frequencies. Long wires can be combined to form antennas of various shapes that increase the gain and directivity over a single wire. The term *long wire*, as used in this chapter, means any such configuration, not just a straight-wire antenna.

Long Wires Versus Multielement Arrays

In general, the gain obtained with long-wire antennas is not as great, when the space available for the antenna is limited, as you can obtain from the multielement phased arrays in Chapter 8 or from a parasitic array such as a Yagi or quad (Chapters 11 or 12). However, the long-wire antenna has advantages of its own that tend to compensate for this deficiency. The construction of long-wire antennas is simple both electrically and mechanically, and there are no especially critical dimensions or adjustments. The long-wire antenna will work well and give satisfactory gain and directivity over a 2-to-1 frequency range. In addition, it will accept power and radiate well on any frequency for which its overall length is not less than about a half wavelength. Since a wire is not electrically long, even at 28 MHz, unless its physical length is equal to at least a half wavelength on 3.5 MHz, any long-wire can be used on all amateur bands that are useful for long-distance communication.

Between two directive antennas having the same theoretical gain, one a multielement array and the other a long-wire antenna, many amateurs have found that the long-wire antenna seems more effective in reception. One possible explanation is that there is a *diversity effect* with a long-wire antenna because it is spread out over a large distance, rather than being concentrated in a small space, as would be the case with a Yagi, for example. This may raise the average level of received energy for ionospheric-propagated signals. Another factor is that long-wire antennas have directive patterns that can be extremely sharp in the horizontal (azimuthal) plane. This is an advantage that other types of multielement arrays do not have, but it can be a double-edged sword too. We'll discuss this aspect in some detail in this chapter.

General Characteristics of Long-Wire Antennas

Whether the long-wire antenna is a single wire running in one direction or is formed into a V-beam, rhombic, or some other configuration, there are certain general principles that apply and some performance features that are common to all types. The first of these is that the power gain of a long-wire antenna as compared with a half-wave dipole is not considerable until the antenna is really long (its length measured in wavelengths rather than in a specific number of feet). The reason for this is that the fields radiated by elementary lengths of wire along the antenna do not combine, at a distance, in as simple a fashion as the fields from half-wave dipoles used in other types of directive arrays.

There is no point in space, for example, where the distant fields from all points along the wire are exactly in phase (as they are, in the optimum direction, in the case of two or more collinear or broadside dipoles when fed with in-phase currents). Consequently, the field strength at a distance is always less than would be obtained if the same length of wire were cut up into properly phased and separately driven dipoles. As the wire is made longer, the fields combine to form increasingly intense main lobes, but these lobes do not develop appreciably until the wire is several wavelengths long. See **Fig 1**.

The longer the antenna, the sharper the lobes become, and since it is really a hollow cone of radiation about the wire in free space, it becomes sharper in both planes. Also, the greater the length, the smaller the angle with the wire at which the maximum radiation lobes occur. There are four main lobes to the directive patterns of long-wire antennas; each makes the same angle with respect to the wire.

Fig 2A shows the azimuthal radiation pattern of a 1-λ long-wire antenna, compared with a $^{1}/_{2}$-λ dipole. Both antennas are mounted at the same height of 1 λ above flat ground (70 feet high at 14 MHz, with a wire length of 70 feet) and both patterns are for an elevation angle of 10°, an angle suitable for long-distance communication on 20 meters. The long-wire in Fig 2A is oriented in the 270° to 90° direction, while the dipole is aligned at right angles so that its characteristic figure-8 pattern goes left-to-right. The 1-λ long-wire has about 0.6 dB more gain than the dipole, with four main lobes as compared to the two lobes from the dipole.

You can see that the two lobes on the left side of Fig 2A are about 1 dB down compared to the two lobes on the right side. This is because the long-wire here is fed at the left-hand end in the computer model. Energy is radiated as a wave travels down the wire and some energy is also lost to ohmic resistance in the wire and the ground. The forward-going wave then reflects from the open-circuit at the right-hand end of the wire and reverses direction, traveling toward the left end, still radiating as it

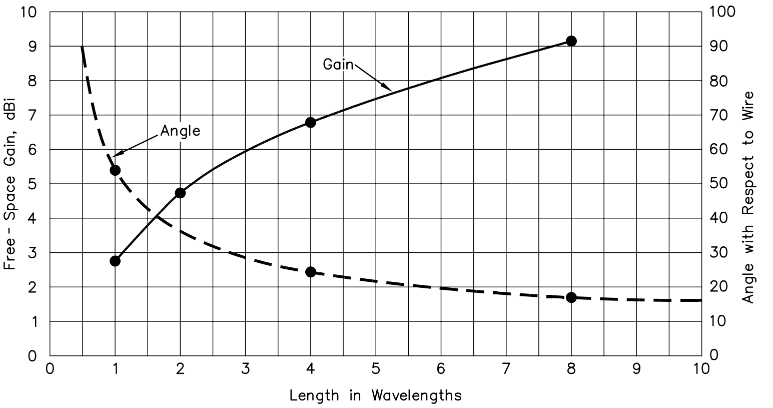

Fig 1—Theoretical gain of a long-wire antenna, in dBi, as a function of wire length. The angle, with respect to the wire, at which the radiation intensity is maximum also is shown.

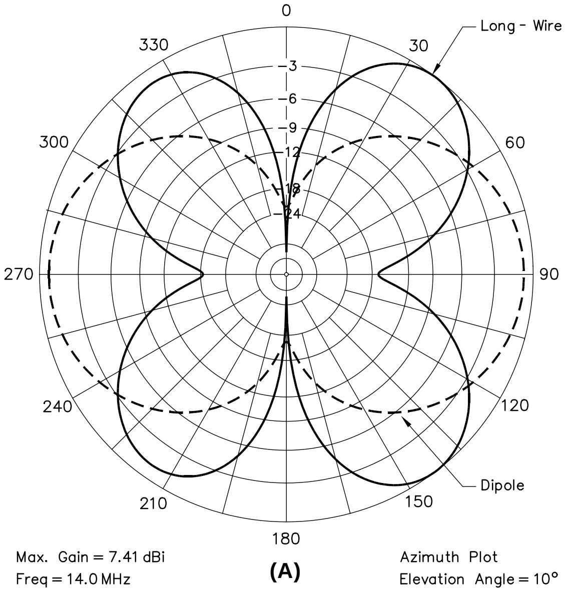

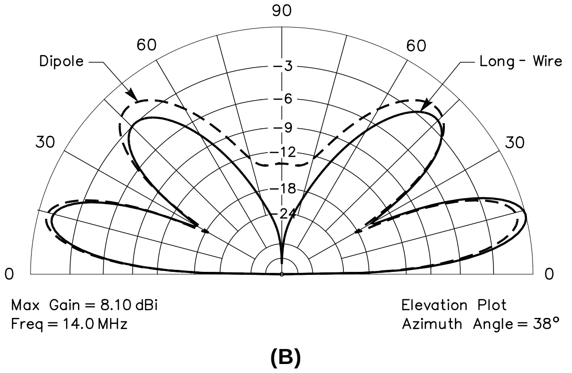

Fig 2—At A, comparison of azimuthal patterns for a 1-λ long-wire antenna (solid line) and a ½-λ dipole (dashed line) at an elevation angle of 10°. Each antenna is located 1 λ (70 feet) over flat ground at 14 MHz. At B, the elevation-plane patterns at peak azimuth angles for each antenna. The long-wire has about 0.6 dB more gain than the dipole.

travels. An antenna operating in this way has much the same characteristics as a transmission line that is terminated in an open circuit—that is, it has standing waves on it. Unterminated long-wire antennas are often referred to as *standing wave antennas*. As the length of a long-wire antenna is increased, a moderate front-to-back ratio results, about 3 dB for very long antennas.

Fig 2B shows the elevation-plane pattern for the long-wire and for the dipole. In each case the elevation pattern is at the azimuth of maximum gain—at an angle of 38° with respect to the wire-axis for the long-wire and at 90° for the dipole. The peak elevation for the long-wire is very slightly lower than that for the dipole at the same height above ground, but not by much. In other words, the height above ground is the main determining factor for the shape of the main lobe of a long-wire's elevation pattern, as it is for most horizontally polarized antennas.

The shape of the azimuth and elevation patterns in Fig 2 might lead you to believe that the radiation pattern is simple. **Fig 3** is a 3-D representation of the pattern from a 1-λ long-wire that is 1 λ high over flat ground. Besides the main low-angle lobes, there are strong lobes at higher angles. Things get even more complicated when the length of the long-wire increases.

Directivity

Because many points along a long wire are carrying currents in different phases (with different current amplitudes as well), the field pattern at a distance becomes more complex as the wire is made longer. This complexity is manifested in a series of minor lobes, the number of which increases with the wire length. The intensity of radiation from the minor lobes is frequently as great as, and sometimes greater than, the radiation from a half-wave dipole. The energy radiated in the minor lobes is not available to improve the gain in the major lobes, which is another reason why a long-wire antenna must be long to give appreciable gain in the desired directions.

Fig 4 shows an azimuthal-plane comparison between a 3-λ (209 feet long) long-wire and the comparison ½-λ dipole. The long-wire now has 8 minor lobes besides the four main lobes. Note that the angle the main lobes make with respect to the axis of the long-wire (also left-to-right in Fig 4) becomes smaller as the length of the long-wire increases. For the 3-λ long-wire, the main lobes occur 28° off the axis of the wire itself.

Other types of simple driven and parasitic arrays do not have minor lobes of any great consequence. For that reason they frequently seem to have much better directivity than long-wire antennas, because their responses in undesired directions are well down from their response in the desired direction. This is the case even if a multielement array and a long-wire antenna have the same peak gain in the favored direction. **Fig 5** compares the same 3-λ long-wire with a 4-element Yagi and a ½-λ dipole, again both at the same height as the long-wire. Note that the Yagi has only a single backlobe, down about 21 dB from its broad main lobe, which has a 3-dB beamwidth of 63°. The 3-dB beamwidth of the long-wire's main lobes (at a 28° angle from the wire axis) is far more narrow, at only 23°.

For amateur work, particularly with directive antennas that cannot be rotated, the minor lobes of a long-wire antenna have some advantages. Although the nulls in the

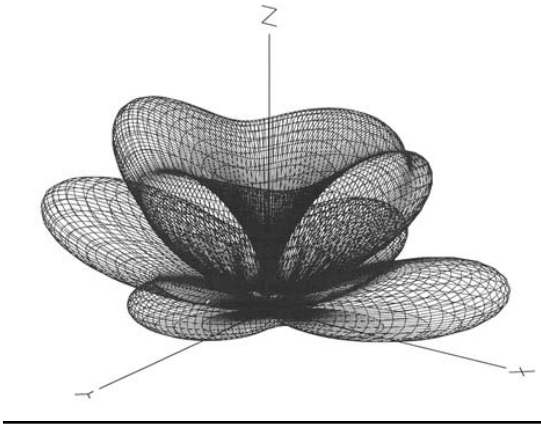

Fig 3—A 3-D representation of the radiation pattern for the 1-λ long-wire shown in Fig 2. The pattern is obviously rather complex. It gets even more complicated for wires longer than 1 λ.

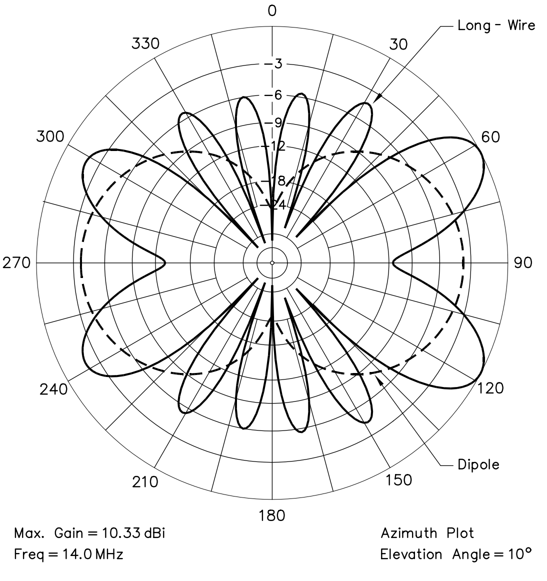

Fig 4—An azimuthal-plane comparison between a 3-λ (209 feet long) long-wire (solid line) and the comparison ½-λ dipole (dashed line) at 70 feet high (1 λ) at 14 MHz.

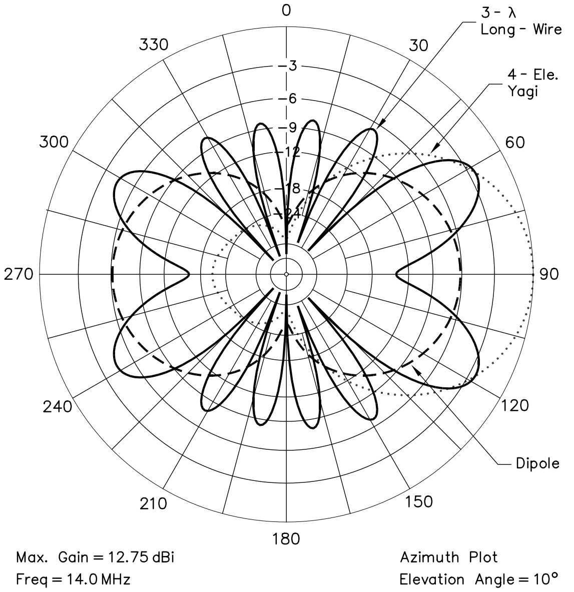

Fig 5—A comparison between the 3-λ long-wire (solid line) in Fig 4, a 4-element 20-meter Yagi on a 26-foot boom (dotted line), and a ½-λ dipole (dashed line), again at a height of 70 feet. The main lobes of the long-wire are very narrow compared to the wide frontal lobe of the Yagi. The long-wire exhibits an azimuthal pattern that is more omnidirectional in nature than a Yagi, particularly when the narrow, deep nulls in the long-wire's pattern are filled-in due to irregularities in the terrain under its long span of wire.

computer model in Fig 5 are deeper than 30 dB, they are not so dramatic in actual practice. This is due to irregularities in the terrain that inevitably occur under the span of a long wire. In most directions the long-wire antenna will be as good as a half-wave dipole, and in addition will give high gain in the most favored directions, even though that is over narrow azimuths.

Fig 6A compares the azimuth responses for a 5-λ long-wire (350 feet long at 14 MHz) to the same 4-element Yagi and dipole. The long-wire now exhibits 16 minor lobes in addition to its four main lobes. The peaks of these sidelobes are down about 8 dB from the main lobes and they are stronger than the dipole, making this long-wire antenna effectively omnidirectional. Fig 6B shows the elevation pattern of the 5-λ long-wire at its most effective azimuth compared to a dipole. Again, the shape of the main lobe is mainly determined by the long-wire's height above ground, since the peak angle is only just a bit lower than the peak angle for the dipole. The long-wire's elevation response breaks up into numerous lobes above the main lobes, just as it does in the azimuth plane.

For the really ambitious, **Fig 7** compares the performance for an 8-λ (571 feet) long-wire antenna with a 4-element Yagi and the $^1/_2$-λ dipole. Again, in actual practice, the nulls would tend to be filled in by terrain irregularities, so a very long antenna like this would be a pretty potent performer.

Calculating Length

In this chapter, lengths are discussed in terms of wave-

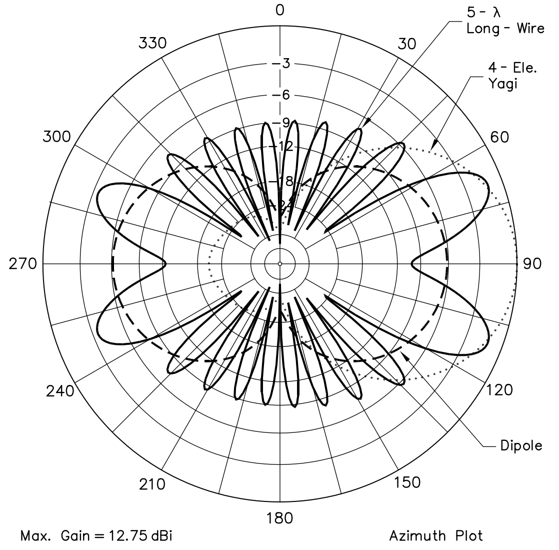

(A)

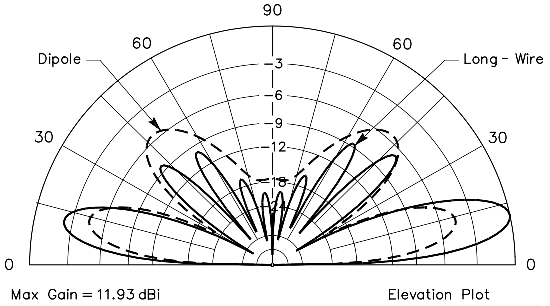

(B)

Fig 6—At A, the azimuth responses for a 5-λ long-wire (350 feet long at 14 MHz—solid line) to the same 4-element Yagi (dotted line) and dipole (dashed line) as in Fig 5. At B, the elevation-plane responses for the long-wire (solid line) and the dipole (dashed line) by themselves. Note that the elevation angle giving peak gain for each antenna is just about the same. The long-wire achieves gain by compressing mainly the azimuthal response, squeezing the gain into narrow lobes; not so much by squeezing the elevation pattern for gain.

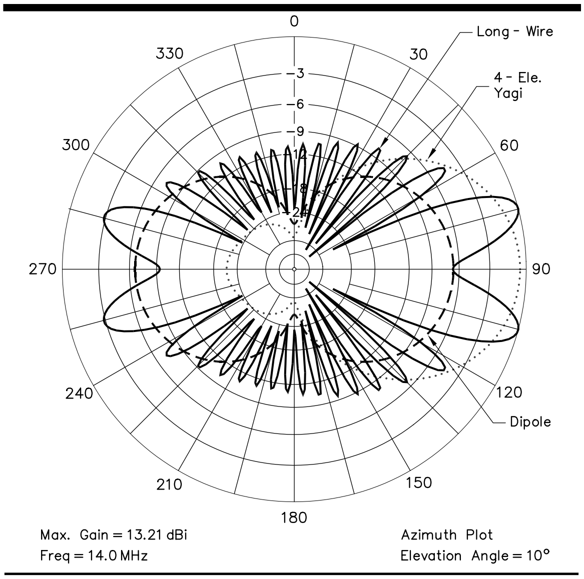

Fig 7—The azimuthal-plane performance for an 8-λ (571 feet) long-wire antenna (solid line), compared with a 4-element Yagi (dotted line) and a ½-λ dipole (dashed line).

lengths. Throughout the preceding discussion the frequency in the models was held at 14 MHz. Remember that a long-wire that is 4 λ long at 14 MHz is 8 λ long at 28 MHz.

There is nothing very critical about wire lengths in an antenna system that will work over a frequency range including several amateur bands. The antenna characteristics change very slowly with length, except when the wires are short (around one wavelength, for instance). There is no need to try to establish exact resonance at a particular frequency for proper antenna operation.

The formula for determining the lengths for harmonic wires is:

$$\text{Length (feet)} = \frac{984\,(N - 0.025)}{f\,(\text{MHz})} \qquad \text{(Eq 1)}$$

where N is the antenna length in wavelengths. In cases where precise resonance is desired for some reason (for obtaining a resistive load for a transmission line at a particular frequency, for example) it is best established by trimming the wire length until the standing-wave ratio on the line is minimum.

Tilted Wires

In theory, it is possible to maximize gain from a long-wire antenna by tilting it to favor a desired elevation take-off angle. Unfortunately, the effect of real ground under the antenna negates the possible advantages of tilting, just as it does when a Yagi or other type of parasitic array is tilted from horizontal. You would do better keeping a long-wire antenna horizontal, but raising it higher above ground, to achieve more gain at low takeoff angles.

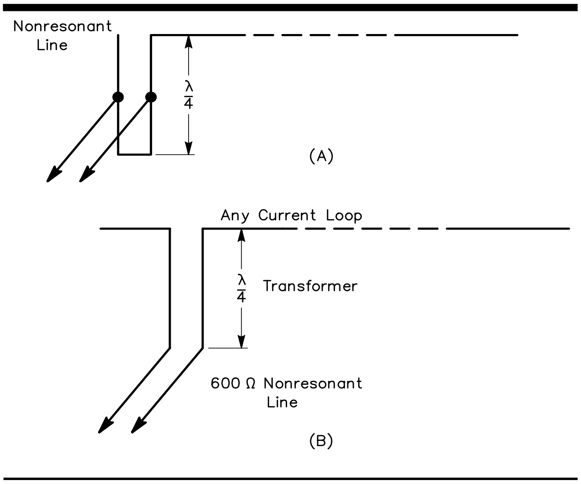

Fig 8—Methods for feeding long single-wire antennas.

Feeding Long Wires

A long-wire antenna is normally fed at the end or at a current loop. Since a current loop changes to a node when the antenna is operated at any even multiple of the frequency for which it is designed, a long-wire antenna will operate as a true long wire on all bands only when it is fed at the end.

A common method of feeding a long-wire is to use a resonant open-wire line. This system will work on all bands down to the one, if any, at which the antenna is only a half wave long. Any convenient line length can be used if you match the transmitter to the line's input impedance using an antenna tuner, as described in Chapter 25.

Two arrangements for using nonresonant lines are given in **Fig 8**. The one at A is useful for one band only since the matching section must be a quarter wave long, approximately, unless a different matching section is used for each band. In B, the λ/4 transformer (Q-section) impedance can be designed to match the antenna to the line, as described in Chapter 26. You can determine the value of radiation resistance using a modern modeling program or you can actually measure the feed-point impedance. Although it will work as designed on only one band, the antenna can be used on other bands by treating the line and matching transformer as a resonant line. In this case, as mentioned earlier, the antenna will not radiate as a true long wire on even multiples of the frequency for which the matching system is designed.

The end-fed arrangement, although the most convenient when tuned feeders are used, suffers the disadvantage that there is likely to be a considerable antenna current on the line. In addition, the antenna reactance changes rapidly with frequency. Consequently, when the wire is several wavelengths long, a relatively small change in frequency—a fraction of the width of a band—may require major changes in the adjustment of the antenna tuner. Also, the line becomes unbalanced at all frequencies between those at which the antenna is resonant. This leads to a considerable amount of radiation from the line. The unbalance can be overcome by using multiple long wires in a V or rhombic shape, as described below.

COMBINATIONS OF LONG WIRES

The directivity and gain of long wires may be increased by using two wires placed in relation to each other such that the fields from both combine to produce the greatest possible field strength at a distant point. The principle is similar to that used in designing the multi-element arrays described in Chapter 8.

Parallel Wires

One possible method of using two (or more) long wires is to place them in parallel, with a spacing of $^1/_2$ λ or so, and feed the two in phase. In the direction of the wires the fields will add in phase. However, the takeoff angle is high directly in the orientation of the wire, and this method will result in rather high-angle radiation even if the wires are several wavelengths long. With a parallel arrangement of this sort the gain should be about 3 dB over a single wire of the same length, at spacings in the vicinity of $^1/_2$ wavelength.

The V-Beam Antenna

Instead of using two long wires parallel to each other, they may be placed in the form of a horizontal V, with the included angle between the wires equal to twice the angle made by the main lobes referenced to the wire axis for a single wire of the same physical length. For example, for a leg length of 5 λ, the angle between the legs of a V should be about 42°, twice the angle of 21° of the main lobe referenced to the long-wire's axis. See Fig 6A.

The plane directive patterns of the individual wires combine along a line in the plane of the antenna and bisecting the V, where the fields from the individual wires reinforce each other. The sidelobes in the azimuthal pattern are suppressed by about 10 dB, so the pattern becomes essentially bidirectional. See **Fig 9**.

The included angle between the legs is not particularly critical. This is fortunate, especially if the same antenna is used on multiple bands, where the electrical length varies directly with frequency. This would normally require different included angles for each band. For multiband V-antennas, a compromise angle is usually chosen to equalize performance. **Fig 10** shows the azimuthal pattern for a V-beam with 1-λ legs, with an included angle of 75° between the legs, mounted 1 λ above flat ground. This is for a 10° elevation angle. At 14 MHz the antenna has two 70-foot high, 68.5-foot long legs, separated at their far ends by 83.4 feet. For comparison the azimuthal patterns for the same 4-element Yagi and ½-λ dipole used previously for the long-wires are overlaid on the same plot. The V has about 2 dB more gain than the dipole but is down some 4 dB compared to the Yagi, as expected for relatively short legs.

Fig 11 shows the azimuthal pattern for the same antenna in Fig 10, but at 28 MHz and at an elevation angle of 6°. Because the legs are twice as long electrically at 28 MHz, the V-beam has compressed the main lobe into a narrow beam that now has a peak gain equal to the Yagi, but with a 3-dB beamwidth of only 18.8°. Note that you could obtain about 0.7 dB more gain at 14 MHz, with a 1.7-dB degradation of gain at 28 MHz, if you increase the included angle to 90° rather than 75°.

Fig 12 shows the azimuthal pattern for a V-beam with 2-λ legs (137 feet at 14 MHz), with an included angle of 60° between them. As usual, the assumed height is 70 feet, or 1 λ at 14 MHz. The peak gain for the V-beam is just about equal to that of the 4-element Yagi, although the 3-dB

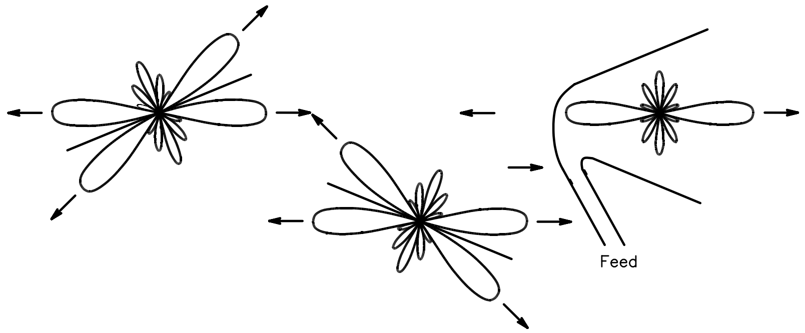

Fig 9—Two long wires and their respective patterns are shown at the left. If these two wires are combined to form a V with an angle that is twice that of the major lobes of the wires and with the wires excited out of phase, the radiation along the bisector of the V adds and the radiation in the other directions tends to cancel.

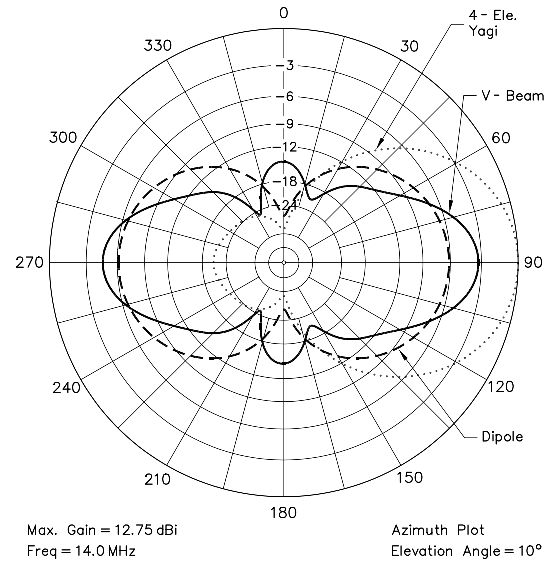

Fig 10— Azimuthal-plane pattern at 10° elevation angle for a 14-MHz V-beam (solid line) with 1-λ legs (68.5 feet long), using an included angle of 75° between the legs. The V-beam is mounted 1 λ above flat ground, and is compared with a ½-λ dipole (dashed line) and a 4-element 20-meter Yagi on a 26-foot boom (dotted line).

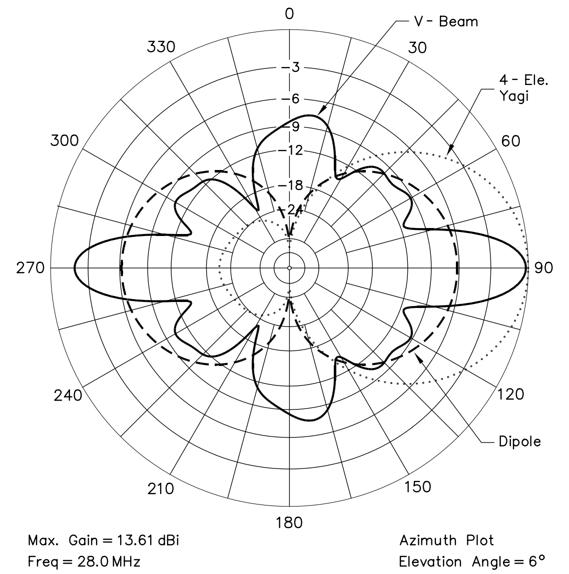

Fig 11—The same V-beam as in Fig 10 at 28 MHz (solid line), at an elevation angle of 6°, compared to a 4-element Yagi (dotted line) and a dipole (dashed line). The V-beam's pattern is very narrow, at 18.8° at the 3-dB points, requiring accurate placement of the supports poles to aim the antenna at the desired geographic target.

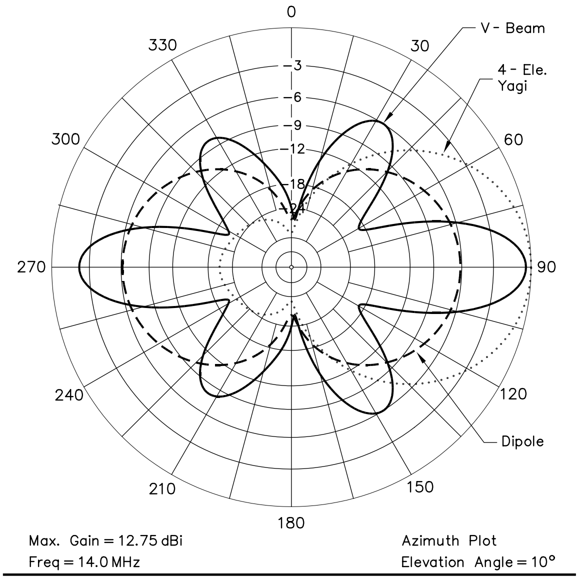

Fig 12—Azimuthal pattern for a V-beam (solid line) with 2-λ legs (137 feet at 14 MHz), with an included angle of 60° between them. The height is 70 feet, or 1 λ, over flat ground. For comparison, the response for a 4-element Yagi (dotted line) and a dipole (dashed line) are shown. The 3-dB beamwidth has decreased to 23.0°.

nose beamwidth is narrow, at 23°. This makes setting up the geometry critical if you want to maximize gain into a particular geographic area. While you might be able to get away with using convenient trees to support such an antenna, it's far more likely that you'll have to use carefully located towers to make sure the beam is aimed where you expect it to be pointed.

For example, in order to cover all of Europe from San Francisco, an antenna must cover from about 11° (to Moscow) to about 46° (to Portugal). This is a range of 35° and signals from the V-beam in Fig 12 would be down some 7 dB over this range of angles, assuming the center of the beam is pointed exactly at a heading of 28.5°. The 4-element Yagi on the other hand would cover this range of azimuths more consistently, since its 3-dB beamwidth is 63°.

Fig 13 shows the same V-beam as in Fig 12, but this time at 28 MHz. The peak gain of the main lobe is now about 1 dB stronger than the 4-element Yagi used as a reference, and the main lobe has two nearby sidelobes that tend to broaden out the azimuthal response. At this frequency the V-beam would cover all of Europe better from San Francisco.

Fig 14 shows a V-beam with 3-λ (209 feet at 14 MHz) legs with an included angle of 50° between them. The peak gain is now greater than that of a 4-element Yagi, but the 3-dB beamwidth has been reduced to 17.8°, making aim-

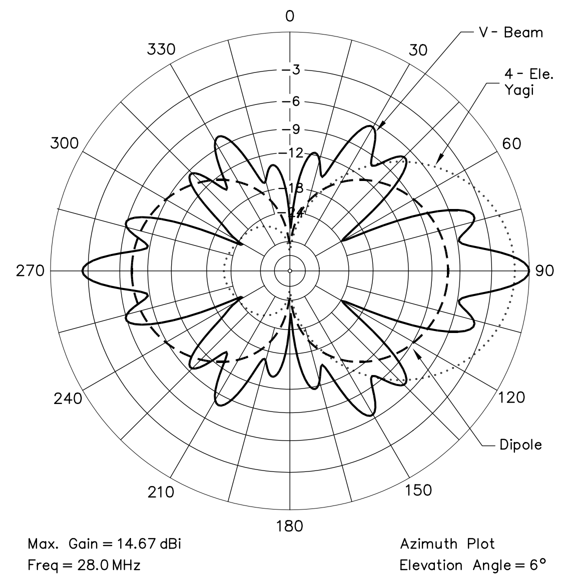

Fig 13—The same 2-λ per leg V-beam (solid line) as in Fig 12, but at 28 MHz and at a 6° takeoff elevation angle. Two sidelobes have appeared flanking the main lobe, making the effective azimuthal pattern wider at this frequency.

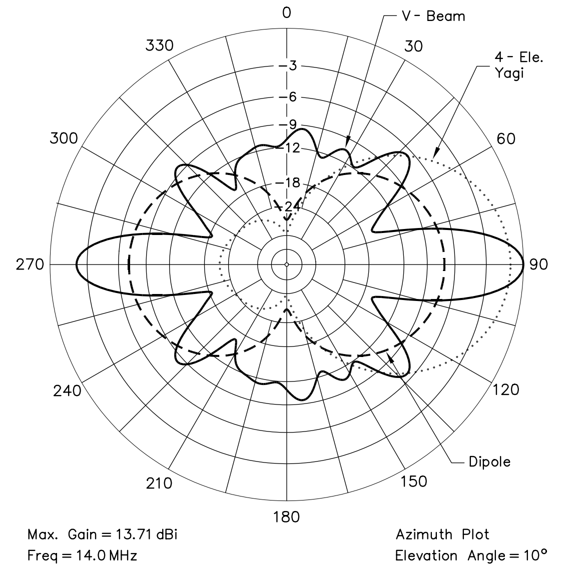

Fig 14—A V-beam (solid line) with 3-λ (209 feet at 14 MHz) legs using an included angle of 50° between them, compared to a 4-element Yagi (dotted line) and a dipole (dashed line). The 3-dB beamwidth has now decreased to 17.8°.

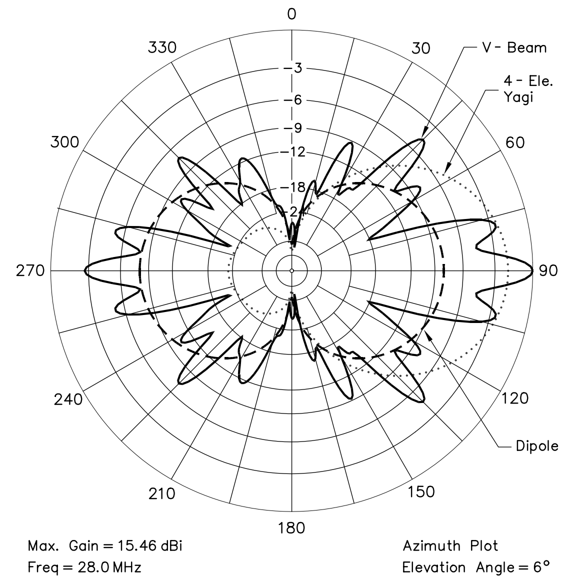

Fig 15—The same 209-foot/leg V-beam as Fig 14, but at 28 MHz. Again, the two close-in sidelobes tend to spread out the azimuthal response some at 28 MHz.

ing the antenna even more critical. **Fig 15** shows the same V-beam at 28 MHz. Here again, the main lobe has nearby sidelobes that broaden the effective azimuth to cover a wider area.

Fig 16 shows the elevation-plane response for the same 209-foot leg V-beam at 28 MHz (3-λ at 14 MHz), compared to a dipole at the same height of 70 feet. The higher-gain V-beam suppresses higher-angle lobes, essentially stealing energy from them and concentrating it in the main beam at 6° elevation.

The same antenna can be used at 3.5 and 7 MHz. The gain will not be large, however, because the legs are not very long at these frequencies. **Fig 17** compares the V-beam versus a horizontal ½-λ 40-meter dipole at 70 feet. At low elevation angles there is about 2 dB of advantage on 40 meters. **Fig 18** shows the same type of comparison for 80 meters, where the 80-meter dipole is superior at all angles.

Other V Combinations

A gain increase of about 3 dB can be had by stacking two V-beams one above the other, a half wavelength apart, and feeding them with in-phase currents. This will result in a lowered angle of radiation. The bottom V should be at least a quarter wavelength above the ground, and preferably a half wavelength. This arrangement will narrow the

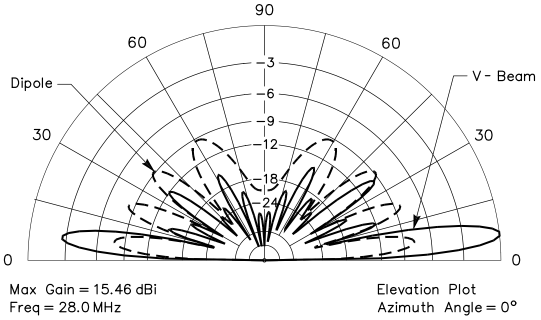

Fig 16—The elevation-plane of the 209-foot/leg V-beam (solid line) compared to the dipole (dashed line). Again, the elevation angle for peak gain corresponds well to that of the simple dipole at the same height.

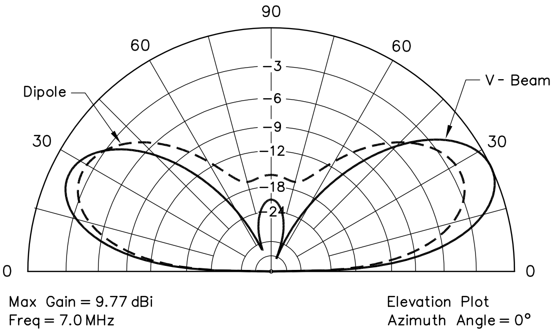

Fig 17—Elevation pattern for the same 209-foot-per-leg V-beam (solid line), at 7 MHz, compared to a 40-meter dipole (dashed line) at the same height of 70 feet.

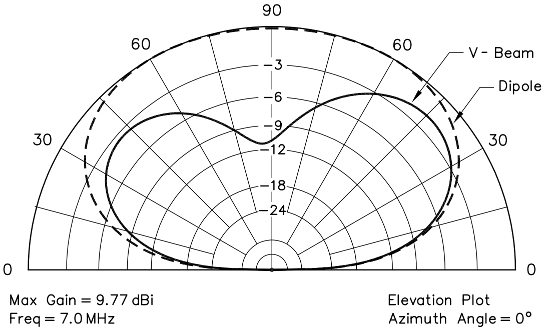

Fig 18—Elevation pattern for the same 209-foot-per-leg V-beam (solid line), at 3.5 MHz, compared to an 80-meter dipole at 70 feet (dashed line).

elevation pattern and it will also have a narrow azimuthal pattern.

The V antenna can be made unidirectional by using a second V placed an odd multiple of a quarter wavelength in back of the first and exciting the two with a phase difference of 90º. The system will be unidirectional in the direction of the antenna with the lagging current. However, the V reflector is not normally employed by amateurs at low frequencies because it restricts the use to one band and requires a fairly elaborate supporting structure. Stacked Vs with driven reflectors could, however, be built for the 200- to 500-MHz region without much difficulty.

Feeding the V Beam

The V-beam antenna is most conveniently fed with tuned open-wire feeders with an antenna tuner, since this permits multiband operation. Although the length of the wires in a V-beam is not at all critical, it is important that both wires be the same electrical length. If a single band matching solution is desired, probably the most appropriate matching system is that using a stub or quarter-wave matching section. The adjustment of such a system is described in Chapter 26.

THE RESONANT RHOMBIC ANTENNA

The diamond-shaped or rhombic antenna shown in **Fig 19** can be looked upon as two acute-angle V-beams placed end-to-end. This arrangement is called a *resonant rhombic*. The leg lengths of the resonant rhombic must be an integral number of half wavelengths to avoid reactance at its feed point.

The resonant rhombic has two advantages over the simple V-beam. For the same total wire length it gives somewhat greater gain than the V-beam. A rhombic with 3 λ on a leg, for example, has about 1 dB gain over a V antenna with 6 wavelengths on a leg. **Fig 20** compares the azimuthal pattern at a 10° elevation for a resonant rhombic with 3 λ legs on 14 MHz, compared to a V-beam with 6 λ legs at the same height of 70 feet. The 3-dB nose beamwidth of the resonant rhombic is only 12.4° wide, but the gain is very high at 16.26 dBi.

The directional pattern of the rhombic is less frequency sensitive than the V when the antenna is used over a wide frequency range. This is because a change in frequency causes the major lobe from one leg to shift in one direction while the lobe from the opposite leg shifts the other way. This automatic compensation keeps the direction the same over a considerable frequency range. The disadvantage of the rhombic as compared with the V-beam is that an additional support is required.

The same factors that govern the design of the V-beam apply in the case of the resonant rhombic. The optimal apex angle A in Fig 19 is the same as that for a V having an equal leg length. The diamond-shaped antenna also can be operated as a terminated antenna, as described later in this chapter, and much of the discus-

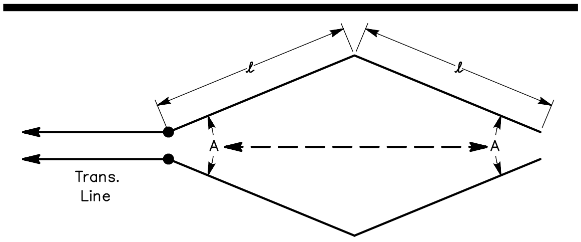

Fig 19—The resonant rhombic or diamond-shaped antenna. All legs are the same length, and opposite angles of the diamond are equal. Length ℓ is an integral number of half wavelengths for resonance.

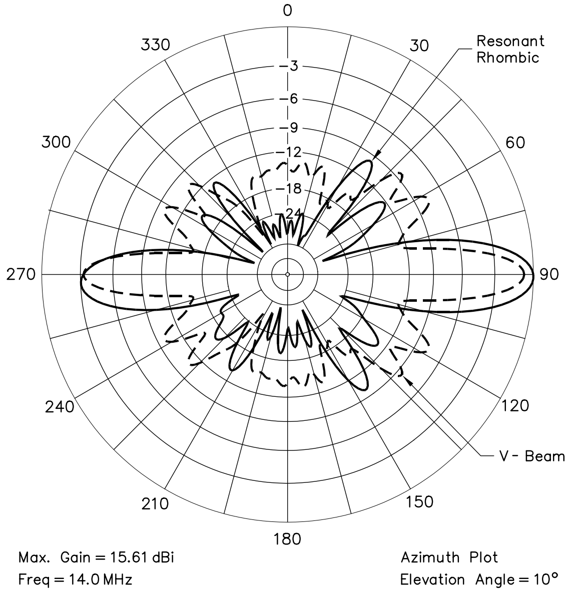

Fig 20—Azimuthal-plane pattern of resonant (unterminated) rhombic (solid line) with 3-λ legs on 14 MHz, at a height of 70 feet above flat ground, compared with a 6-λ per leg V-beam (dashed line) at the same height. Both azimuthal patterns are at a takeoff angle of 10°. The sidelobes for the resonant rhombic are suppressed to a greater degree than those for the V-beam.

sion in that section applies to the resonant rhombic as well.

The resonant rhombic has a bidirectional pattern, with minor lobes in other directions, their number and intensity depending on the leg length. In general, these sidelobes are suppressed better with a resonant rhombic than with a V-beam. When used at frequencies below the VHF region, the rhombic antenna is always mounted with the plane containing the wires horizontal. The polarization in this plane, and also in the perpendicular plane that bisects the rhombic, is horizontal. At 144 MHz and above, the dimensions are such that the antenna can be mounted with the plane containing the wires vertical if vertical polarization is desired.

When the rhombic antenna is to be used on several HF amateur bands, it is advisable to choose the apex angle, A, on the basis of the leg length in wavelengths at 14 MHz. Although the gain on higher frequency bands will not be quite as favorable as if the antenna had been designed for the higher frequencies, the system will still work well at the low angles that are necessary at such frequencies.

The resonant rhombic has lots of gain, but you must not forget that this gain comes from a radiation pattern that is very narrow. This requires careful placement of the supports for the resonant rhombic to cover desired geographic areas. This is definitely not an antenna that allows you to use just any convenient trees as supports!

The resonant rhombic antenna can be fed in the same way as the V-beam. Resonant feeders are necessary if the antenna is to be used in several amateur bands.

TERMINATED LONG-WIRE ANTENNAS

All the antenna systems considered so far in this chapter have been based on operation with standing waves of current and voltage along the wire. Although most hams use antenna designs based on using resonant wires, resonance is by no means a necessary condition for the wire to radiate and intercept electromagnetic waves efficiently, as discussed in Chapter 2. The result of using nonresonant wires is reactance at the feed point, unless the antenna is terminated with a resistive load.

In **Fig 21**, suppose that the wire is parallel with the ground (horizontal) and is terminated by a load Z equal to its characteristic impedance, Z_{ANT}. The wire and its image in the ground create a transmission line. The load Z can represent a receiver matched to the line. The *terminating resistor* R is also equal to the Z_{ANT} of the wire. A wave coming from direction X will strike the wire first at its far end and sweep across the wire at some angle until it reaches the end at which Z is connected. In so doing, it will induce voltages in the antenna, and currents will flow as a result. The current flowing toward Z is the useful output of the antenna, while the current flowing backwards toward R will be absorbed in R. The same thing is true of a wave coming from the direction X'. In such an antenna there are no standing waves, because all received power is absorbed at either end.

The greatest possible power will be delivered to the load Z when the individual currents induced as the wave sweeps across the wire all combine properly on reaching the load. The currents will reach Z in optimum phase when the time required for a current to flow from the far end of the antenna to Z is exactly one-half cycle longer than the time taken by the wave to sweep over the antenna. A half cycle is equivalent to a half wavelength greater than the distance traversed by the wave from the instant it strikes

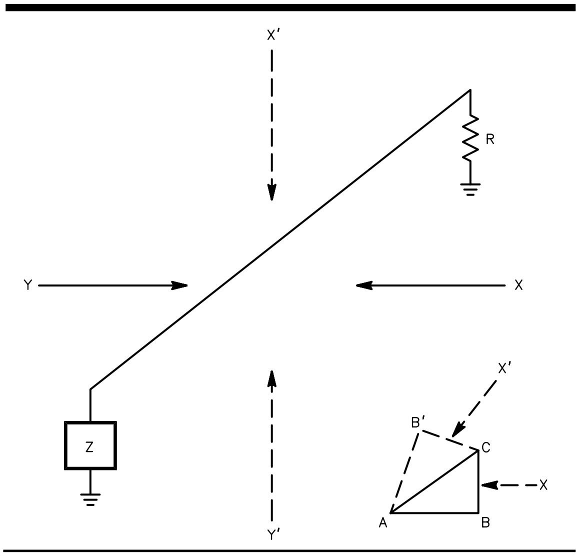

Fig 21—Layout for a terminated long-wire antenna.

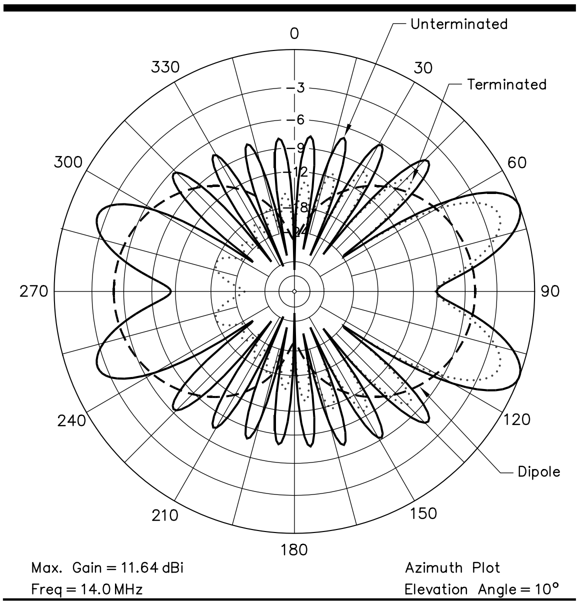

Fig 22—Azimuthal-plane pattern for 5-λ long-wire antenna at 14 MHz and 70 feet above flat ground. The solid line shows the long-wire terminated with 600-Ω to ground, while the dashed line is for the same antenna unterminated. For comparison, the response for a ½-λ dipole is overlaid with the two other patterns. You can see that the terminated long-wire has a good front-to-back pattern, but it loses about 2 dB in forward gain compared to the unterminated long-wire.

the far end of the antenna to the instant that it reaches the near end. This is shown by the small drawing, where AC represents the antenna, BC is a line perpendicular to the wave direction, and AB is the distance traveled by the wave in sweeping past AC. AB must be one-half wavelength shorter than AC. Similarly, AB' must be the same length as AB for a wave arriving from X'.

A wave arriving at the antenna from the opposite direction Y (or Y'), will similarly result in the largest possible current at the far end. However, since the far end is terminated in R, which is equal to Z, all the power delivered to R by the wave arriving from Y will be absorbed in R. The current traveling to Z will produce a signal in Z in proportion to its amplitude. If the antenna length is such that all the individual currents arrive at Z in such phase as to add up to zero, there will be no current through Z. At other lengths the resultant current may reach appreciable values. The lengths that give zero amplitude are those which are odd multiples of $^{1}/_{4}\ \lambda$, beginning at $^{3}/_{4}\ \lambda$. The response from the Y direction is greatest when the antenna is any even multiple of $^{1}/_{2}\ \lambda$ long; the higher the multiple, the smaller the response.

Directional Characteristics

Fig 22 compares the azimuthal pattern for a 5-λ long 14-MHz long-wire antenna, 70 feet high over flat ground, when it is terminated and when it is unterminated. The rearward pattern when the wire is terminated with a 600 Ω resistor is reduced about 15 dB, with a reduction in gain in the forward direction of about 2 dB.

For a shorter leg length in a terminated long-wire antenna, the reduction in forward gain is larger—more energy is radiated by a longer wire before the forward wave is absorbed in the terminating resistor. The azimuthal patterns for terminated and unterminated V-beams with 2-λ legs are overlaid for comparison in **Fig 23**. With these relatively short legs the reduction in forward gain is about 3.5 dB due to the terminations, although the front-to-rear ratio approaches 20 dB for the terminated V-beam. Each leg of this terminated V-beam use a 600-Ω non-inductive resistor to ground. Each resistor would have to dissipate about one-quarter of the transmitter power. For average conductor diameters and heights above ground, the Z_{ANT} of the antenna is of the order of 500 to 600 Ω.

THE TERMINATED RHOMBIC ANTENNA

The highest development of the long-wire antenna is the *terminated rhombic*, shown schematically in **Fig 24**. It consists of four conductors joined to form a diamond, or *rhombus*. All sides of the antenna have the same length and the opposite corner angles are equal. The antenna can be considered as being made up of two V antennas placed end to end and terminated by a noninductive resistor to produce a unidirectional pattern. The terminating resistor is connected between the far ends of the two sides, and is made approximately equal to the characteristic impedance of the antenna as a unit. The rhombic may be constructed either horizontally or vertically, but is practically always constructed horizontally at frequencies below 54 MHz,

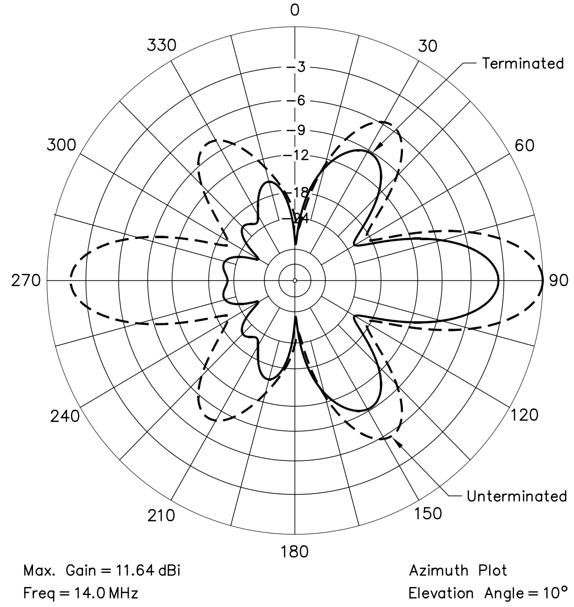

Fig 23—The azimuthal patterns for a shorter-leg V-beam (2-λ legs) when it is terminated (solid line) and unterminated (dashed line). With shorter legs, the terminated V-beam loses about 3.5 dB in forward gain compared to the unterminated version, while suppressing the rearward lobes as much as 20 dB.

since the pole height required is considerably less. Also, horizontal polarization is equally, if not more, satisfactory at these frequencies over most types of soil.

The basic principle of combining lobes of maximum radiation from the four individual wires constituting the rhombus or diamond is the same in either the terminated type or the resonant type described earlier in this chapter.

Tilt Angle

In dealing with the terminated rhombic, it is a matter of custom to talk about the *tilt angle* (ϕ in Fig 24), rather than the angle of maximum radiation with respect to an individual wire. **Fig 25** shows the tilt angle as a function of the antenna leg length. The curve marked “0°” is used for a takeoff elevation angle of 0°; that is, maximum radiation in the plane of the antenna. The other curves show the proper tilt angles to use when aligning the major lobe with a desired takeoff angle. For a 5° takeoff angle, the difference in tilt angle is less than 1° for the range of lengths shown.

The broken curve marked “optimum length” shows the leg length at which maximum gain is obtained at any given takeoff angle. Increasing the leg length beyond the optimum will result in less gain, and for that reason the curves do not extend beyond the optimum length. Note that the

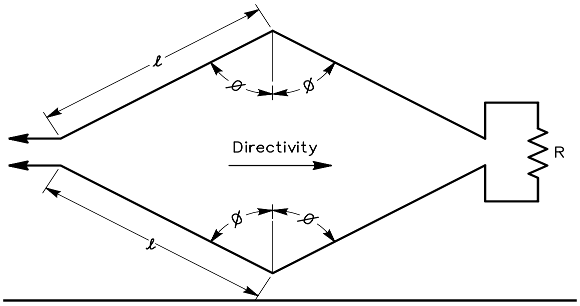

Fig 24—The layout for a terminated rhombic antenna.

optimum length becomes greater as the desired takeoff angle decreases. Leg lengths over 6 λ are not recommended because the directive pattern becomes so sharp that the antenna performance is highly variable with small changes in the angle, both horizontal and vertical, at which an incoming wave reaches the antenna. Since these angles vary to some extent in ionospheric propagation, it does not pay to attempt to try for too great a degree of directivity.

Multiband Design

When a rhombic antenna is to be used over a considerable frequency range, a compromise must be made in the tilt angle. **Fig 26** gives the design dimensions of a suitable compromise for a rhombic that covers the 14 to 30 MHz range well. **Fig 27** shows the azimuth and elevation patterns for this antenna at 14 MHz, at a height of 70 feet over flat ground. The comparison antenna in this case is a 4-element Yagi on a 26-foot boom, also 70 feet above flat ground. The rhombic has about 2.2 dB more gain, but its azimuthal pattern is 17.2° wide at the 3 dB points, and only 26° at the –20 dB points! On the other hand, the Yagi has a 3-dB beamwidth of 63°, making it far easier to aim at a distant geographic location. Fig 27B shows the elevation-plane patterns for the same antennas above. As usual, the peak angle for either horizontally polarized antenna is determined mainly by the height above ground.

The peak gain of a terminated rhombic is less than that of an unterminated resonant rhombic. For the rhombic of Fig 26, the reduction in peak gain is about 1.5 dB. **Fig 28** compares the azimuthal patterns for this rhombic with and without an 800-Ω termination.

Fig 29 shows the azimuth and elevation patterns for the terminated rhombic of Fig 26 when it is operated at 28 MHz. The main lobe becomes very narrow, at 6.9° at the 3-dB points. However, this is partially compensated for by the appearance of two sidelobes each side of the main beam. These tend to spread out the main pattern some. Again, a 4-element Yagi at the same height is used for comparison.

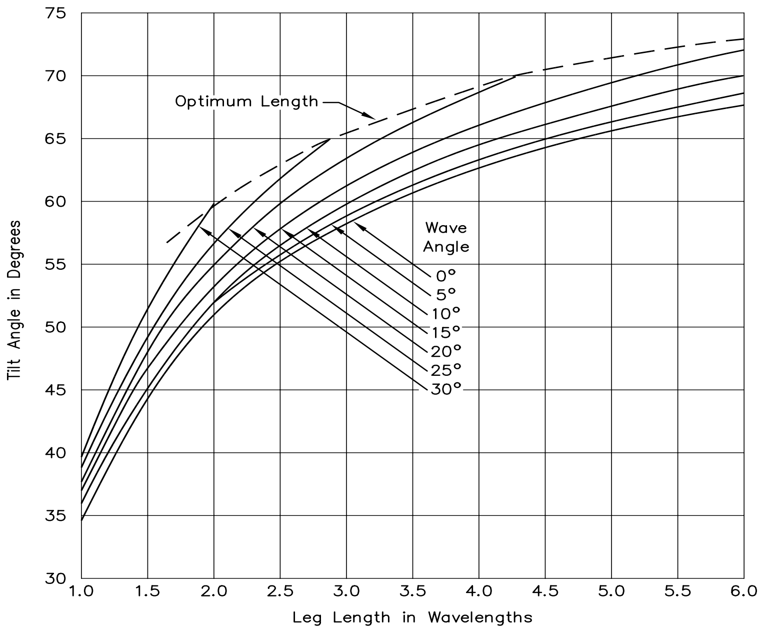

Fig 25—Rhombic-antenna design chart. For any given leg length, the curves show the proper tilt angle to give maximum radiation at the selected takeoff angle. The broken curve marked "optimum length" shows the leg length that gives the maximum possible output at the selected takeoff angle. The optimum length as given by the curves should be multiplied by 0.74 to obtain the leg length for which the takeoff angle and main lobe are aligned.

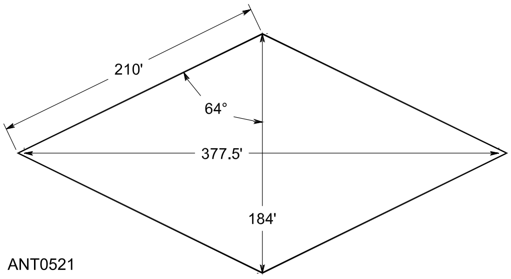

Fig 26—Rhombic antenna dimensions for a compromise design between 14- and 28-MHz requirements, as discussed in the text. The leg length is 6 λ at 28 MHz, 3 λ at 14 MHz.

Termination

Although the difference in the gain is relatively small with terminated or unterminated rhombics of comparable design, the terminated antenna has the advantage that over a wide frequency range it presents an essentially resistive and constant load to the transmitter. In a sense, the power dissipated in the terminating resistor can be considered power that would have been radiated in the other direction had the resistor not been there. Therefore, the fact that some of the power (about one-third) is used up in heating the resistor does not mean that much actual loss in the desired direction.

The characteristic impedance of an ordinary rhombic antenna, looking into the input end, is in the order of 700 to 800 Ω when properly terminated in a resistance at the far end. The terminating resistance required to bring about the matching condition usually is slightly higher than the input impedance because of the loss of energy through radiation by the time the far end is reached. The correct value usually will be found to be of the order of 800 Ω, and should be determined experimentally if the flattest possible antenna is desired. However, for average work a noninductive resistance of 800 Ω can be used with the assurance that the operation will not be far from optimum.

The terminating resistor must be practically a pure resistance at the operating frequencies; that is, its inductance and capacitance should be negligible. Ordinary wire-wound resistors are not suitable because they have far too much inductance and distributed capacitance. Small carbon resistors have satisfactory electrical characteristics but will not dissipate more than a few watts and so cannot be used, except when the transmitter power does not exceed 10 or 20 watts or when the antenna is to be used for reception only. The special resistors designed either for use as dummy antennas or for terminating rhombic antennas should be used in other cases. To allow a factor of safety, the total rated power dissipation of the resistor or resistors should be equal to half the power output of the transmitter.

To reduce the effects of stray capacitance it is desirable to use several units, say three, in series even when one alone will safely dissipate the power. The two end units should be identical and each should have one fourth to one third the total resistance, with the center unit making up the difference. The units should be installed in a weather-

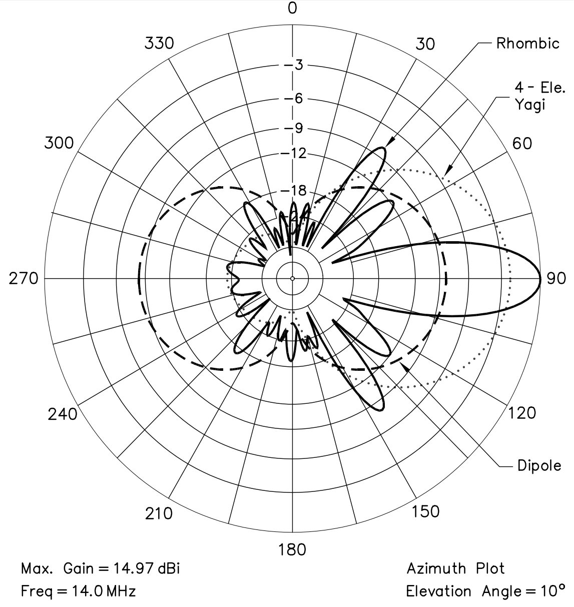

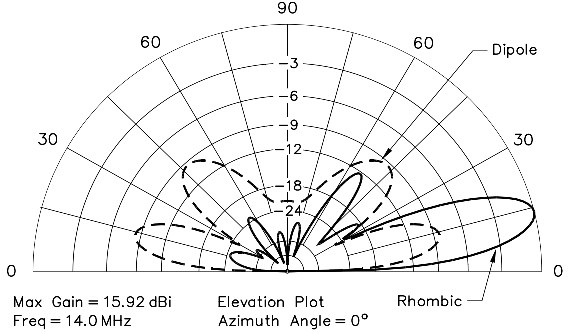

Fig 27—At left, azimuthal pattern for 3-λ (at 14 MHz) terminated rhombic (solid line) shown in Fig 26, compared with 4-element 20-meter Yagi (dotted line) on a 26-foot boom and a 20-meter dipole (dashed line). All antennas are mounted 70 feet (1 λ) above flat ground. The rearward pattern of the terminated rhombic is good and the forward gain exceeds that of the Yagi, but the frontal lobe is very narrow. Above, elevation-plane pattern of terminated rhombic compared to that of a simple dipole at the same height.

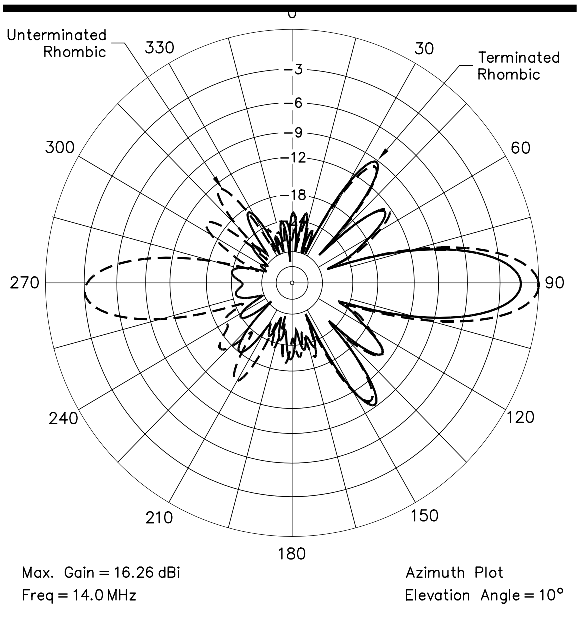

Fig 28—Comparison of azimuthal patterns for terminated (solid line) and unterminated (dashed line) rhombic antennas, using same dimensions as Fig 26 at a frequency of 14 MHz. The gain tradeoff is about 1.5 dB in return for the superior rearward pattern of the terminated antenna.

proof housing at the end of the antenna to protect them and to permit mounting without mechanical strain. The connecting leads should be short so that little extraneous inductance is introduced.

Alternatively, the terminating resistance may be placed at the end of an 800-Ω line connected to the end of the antenna. This will permit placing the resistors and their housing at a point convenient for adjustment rather than at the top of the pole. Resistance wire may be used for this line, so that a portion of the power will be dissipated before it reaches the resistive termination, thus permitting the use of lower wattage lumped resistors.

Multiwire Rhombics

The input impedance of a rhombic antenna constructed as in Fig 26 is not quite constant as the frequency is varied. This is because the varying separation between the wires causes the characteristic impedance of the antenna to vary along its length. The variation in Z_{ANT} can be minimized by a conductor arrangement that increases the capacitance per unit length in proportion to the separation between the wires.

The method of accomplishing this is shown in **Fig 30**. Three conductors are used, joined together at the ends but with increasing separation as the junction between legs is approached. For HF work the spacing between the wires at the center is 3 to 4 feet, which is similar to that used in commercial installations using legs several wavelengths long. Since all three wires should have the same length, the top and bottom wires should be slightly farther from the support than the middle wire. Using three wires in this way reduces the Z_{ANT} of the antenna to approximately 600 Ω, thus providing a better match for practical open-wire line, in addition to smoothing out the impedance variation over the frequency range.

A similar effect (although not quite as favorable) is obtained by using two wires instead of three. The 3-wire system has been found to increase the gain of the

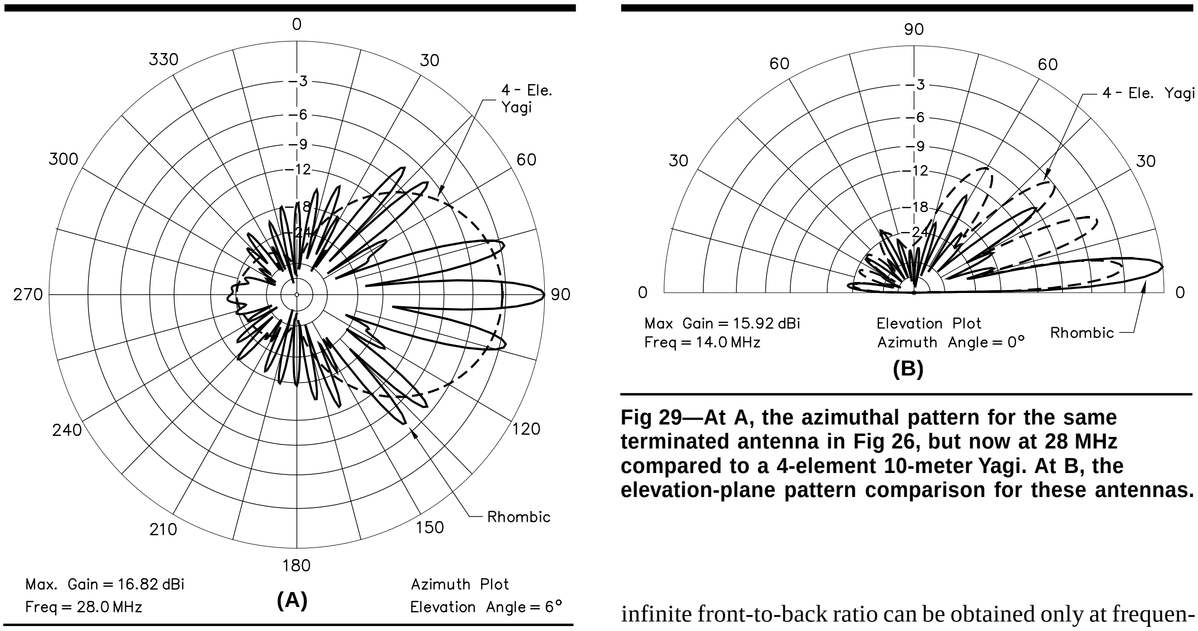

Fig 29—At A, the azimuthal pattern for the same terminated antenna in Fig 26, but now at 28 MHz compared to a 4-element 10-meter Yagi. At B, the elevation-plane pattern comparison for these antennas.

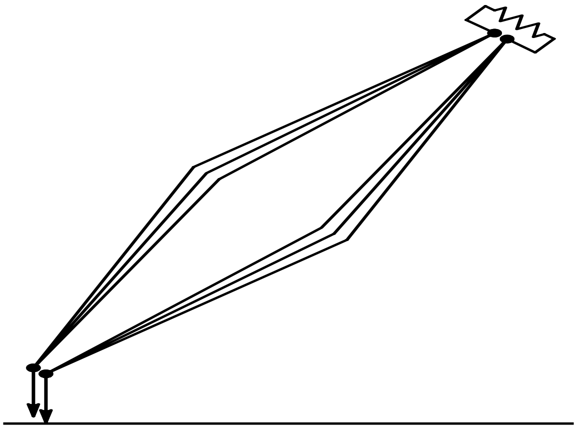

Fig 30—Three-wire rhombic antenna. Use of multiple wires improves the impedance characteristic of a terminated rhombic and increases the gain somewhat.

antenna by about 1 dB over that of a single-conductor version.

Front-to-Back Ratio

It is theoretically possible to obtain an infinite front-to-back ratio with a terminated rhombic antenna, and in practice very large values can be had. However, when the antenna is terminated in its characteristic impedance, the infinite front-to-back ratio can be obtained only at frequencies for which the leg length is an odd multiple of a quarter wavelength. The front-to-back ratio is smallest at frequencies for which the leg length is a multiple of a half wavelength.

When the leg length is not an odd multiple of a quarter wave at the frequency under consideration, the front-to-back ratio can be made very high by decreasing the value of terminating resistance slightly. This permits a small reflection from the far end of the antenna, which cancels out the residual response at the input end. With large antennas, the front-to-back ratio may be made very large over the whole frequency range by experimental adjustment of the terminating resistance. Modification of the terminating resistance can result in a splitting of the back null into two nulls, one on either side of a small lobe in the back direction. Changes in the value of terminating resistance thus permit steering the back null over a small horizontal range so that signals coming from a particular spot not exactly to the rear of the antenna may be minimized.

Methods of Feed

If the broad frequency characteristic of the terminated rhombic antenna is to be utilized fully, the feeder system must be similarly broadbanded. Open-wire transmission line of the same characteristic impedance as that shown at the antenna input terminals (approximately 700 to 800 Ω) may be used. Data for the construction of such lines is given in Chapter 24. While the usual matching stub can be used to provide an impedance transformation to more satisfactory line impedances, this limits the operation of the antenna to a comparatively narrow range of frequencies centering about that for which the stub is adjusted. Probably a more satisfactory arrangement would be to use a coaxial transmission line and a broadband transformer balun at the antenna feed point.

Receiving Wave Antennas

Perhaps the best known type of wave antenna is the *Beverage*. Many 160-meter enthusiasts have used Beverage antennas to enhance the signal-to-noise ratio while attempting to extract weak signals from the often high levels of atmospheric noise and interference on the low bands. Alternative antenna systems have been developed and used over the years, such as loops and long spans of unterminated wire on or slightly above the ground, but the Beverage antenna seems to be the best for 160-meter weak-signal reception. The information in this section was prepared originally by Rus Healy, K2UA (ex-NJ2L).

THE BEVERAGE ANTENNA

A Beverage is simply a directional wire antenna, at least one wavelength long, supported along its length at a fairly low height and terminated at the far end in its characteristic impedance. This antenna is shown in **Fig 31A**. It takes its name from its inventor, Harold Beverage, W2BML.

Many amateurs choose to use a single-wire Beverage because they are easy to install and they work well. The drawback is that Beverages are physically long and they do require that you have the necessary amount of real estate to install them. Sometimes, a neighbor will allow you to put up a temporary Beverage for a particular contest or DXpedition on his land, particularly during the winter months.

Beverage antennas can be useful into the HF range, but they are most effective at lower frequencies, mainly on 160 through 40 meters. The antenna is responsive mostly to low-angle incoming waves that maintain a constant (vertical) polarization. These conditions are nearly always satisfied on 160 meters, and most of the time on 80 meters. As the frequency is increased, however, the polarization and arrival angles are less and less constant and favorable, making Beverages less effective at these frequencies. Many amateurs have, however, reported excellent performance from Beverage antennas at frequencies as high as 14 MHz, especially when rain or snow (precipitation) static prevents good reception on the Yagi or dipole transmitting antennas used on the higher frequencies.

Beverage Theory

The Beverage antenna acts like a long transmission line with one lossy conductor (the earth), and one good conductor (the wire). Beverages have excellent directivity if erected properly, but they are quite inefficient because they are mounted close to the ground. This is in contrast with the terminated long-wire antennas described earlier, which are typically mounted high off the ground. Beverage antennas are not suitable for use as transmitting antennas.

Because the Beverage is a traveling wave, terminated antenna, it has no standing waves resulting from radio signals. As a wave strikes the end of the Beverage from the desired direction, the wave induces voltages along the antenna and continues traveling in space as well. Fig 31B shows part of a wave on the antenna resulting from a desired signal. This diagram also shows the tilt of the wave. The signal induces equal voltages in both directions. The resulting currents are equal and travel in both directions; the component traveling toward the termination end moves *against* the wave and thus builds down to a very low level at the termination end. Any residual signal resulting from this direction of current flow will be absorbed in the termination (if the termination is equal to the antenna impedance). The component of the signal flowing in the other direction, as we will see, becomes a key part of the received signal.

As the wave travels along the wire, the wave in space travels at approximately the same velocity. (There is some phase delay in the wire, as we shall see.) At any given point in time, the wave traveling along in space induces a voltage in the wire in addition to the wave already travel-

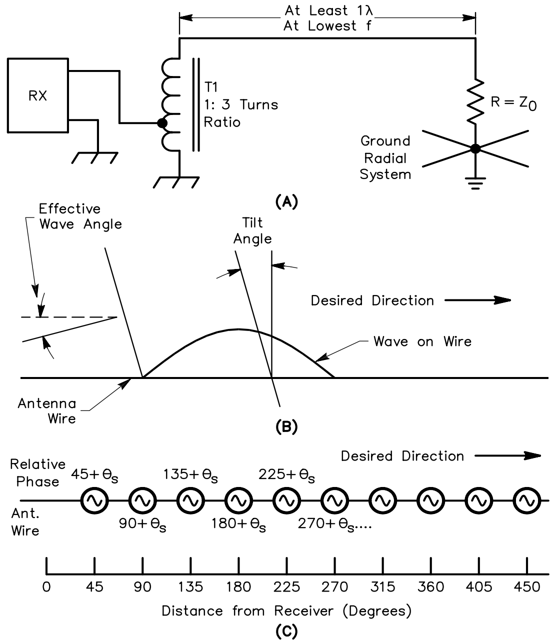

Fig 31—At A, a simple one-wire Beverage antenna with a variable termination impedance and a matching 9:1 autotransformer for the receiver impedance. At B, a portion of a wave from the desired direction is shown traveling down the antenna wire. Its tilt angle and effective takeoff angle are also shown. At C, a situation analogous to the action of a Beverage on an incoming wave is shown. See text for discussion.

ing on the wire (voltages already induced by the wave). Because these two waves are nearly in phase, the voltages add and build toward a maximum at the receiver end of the antenna.

This process can be likened to a series of signal generators lined up on the wire, with phase differences corresponding to their respective spacings on the wire (Fig 31C). At the receiver end, a maximum voltage is produced by these voltages adding in phase. For example, the wave component induced at the receiver end of the antenna will be in phase (at the receiver end) with a component of the same wave induced, say, 270° (or any other distance) down the antenna, after it travels to the receiver end.

In practice, there is some phase shift of the wave on the wire with respect to the wave in space. This phase shift results from the velocity factor of the antenna. (As with any transmission line, the signal velocity on the Beverage is somewhat less than in free space.) Velocity of propagation on a Beverage is typically between 85 and 98% of that in free space. As antenna height is increased to a certain optimum height (which is about 10 feet for 160 meters), the velocity factor increases. Beyond this height, only minimal improvement is afforded, as shown in **Fig 32**. These curves are the result of experimental work done in 1922 by RCA, and reported in a *QST* article (November 1922) entitled “The Wave Antenna for 200-Meter Reception,” by H. H. Beverage. The curve for 160 meters was extrapolated from the other curves.

Phase shift (per wavelength) is shown as a function of velocity factor in **Fig 33**, and is given by:

$$\theta = 360 \left(\frac{100}{k} - 1 \right) \qquad \text{(Eq 2)}$$

where k = velocity factor of the antenna in percent.

The signals present on and around a Beverage antenna are shown graphically in A through D of **Fig 34**. These curves show relative voltage levels over a number of periods of the wave in space and their relative effects in terms of the total signal at the receiver end of the antenna.

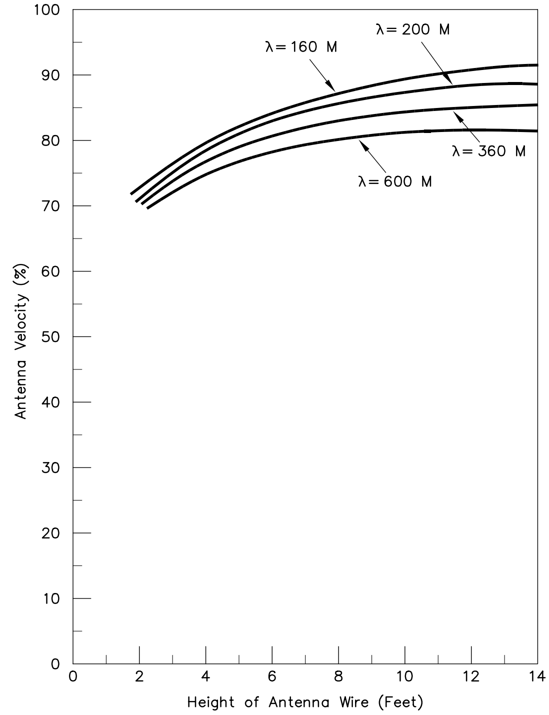

Fig 32—Signal velocity on a Beverage increases with height above ground, and reaches a practical maximum at about 10 feet. Improvement is minimal above this height. (The velocity of light is 100%.)

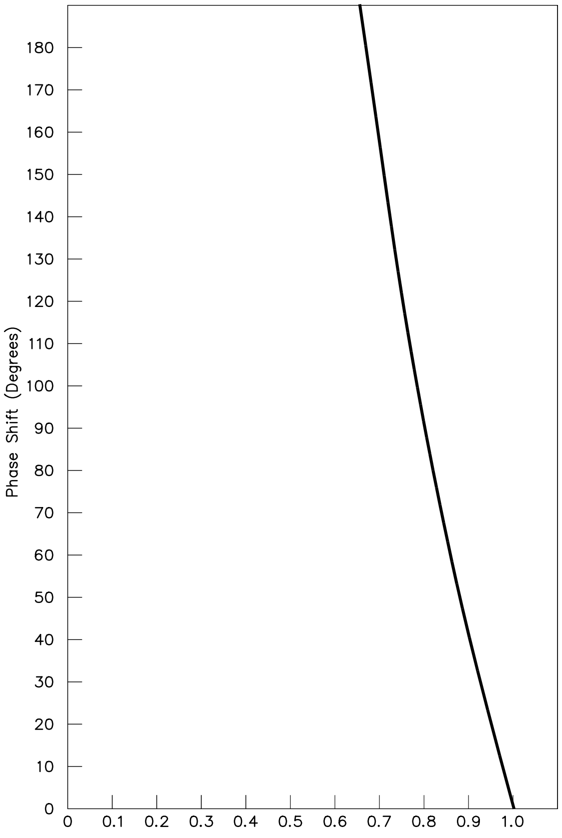

Fig 33—This curve shows phase shift (per wavelength) as a function of velocity factor on a Beverage antenna. Once the phase shift for the antenna goes beyond 90°, the gain drops off from its peak value, and any increase in antenna length will decrease gain.

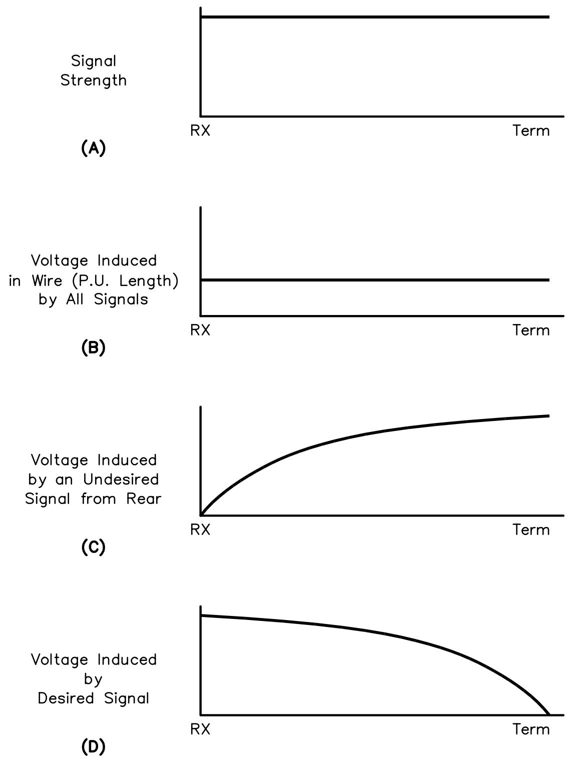

Fig 34—These curves show the voltages that appear in a Beverage antenna over a period of several cycles of the wave. Signal strength (at A) is constant over the length of the antenna during this period, as is voltage induced per unit length in the wire (at B). (The voltage induced in any section of the antenna is the same as the voltage induced in any other section of the same size, over the same period of time.) At C, the voltages induced by an undesired signal from the rearward direction add in phase and build to a maximum at the termination end, where they are dissipated in the termination (if $Z_{term} = Z_0$). The voltages resulting from a desired signal are shown at D. The wave on the wire travels closely with the wave in space, and the voltages resulting add in phase to a maximum at the receiver end of the antenna.

Performance in Other Directions

The performance of a Beverage antenna in directions other than the favored one is quite different than previously discussed. Take, for instance, the case of a signal arriving perpendicular to the wire (90° either side of the favored direction). In this case, the wave induces voltages along the wire that are essentially *in phase*, so that they arrive at the receiver end more or less out of phase, and thus cancel. (This can be likened to a series of signal generators lined up along the antenna as before, but having no progressive phase differences.)

As a result of this cancellation, Beverages exhibit deep nulls off the sides. Some minor sidelobes will exist, as with other long-wire antennas, and will increase in number with the length of the antenna.

In the case of a signal arriving from the rear of the antenna, the behavior of the antenna is very similar to its performance in the favored direction. The major difference is that the signal from the rear adds in phase at the termination end and is absorbed by the termination impedance. **Fig 35** compares the azimuth and elevation patterns for a 2-λ (1062 foot) and a 1-λ (531 foot) Beverage at 1.83 MHz. The wire is mounted 8 feet above flat ground (to keep it above deer antlers and away from humans too) and is terminated with a 500-Ω resistor in each case, although the exact value of the terminating resistance is not very critical. The ground constants assumed in this computer model are conductivity of 5 mS/m and a dielectric constant of 13. Beverage dielectric performance tends to decrease as the ground becomes better. Beverages operated over saltwater do not work as well as they do over poor ground.

For most effective operation, the Beverage should be terminated in an impedance equal to the characteristic impedance Z_{ANT} of the antenna. For maximum signal transfer to the receiver you should also match the receiver's input impedance to the antenna. If the termination impedance is not equal to the characteristic impedance of the antenna, some part of the signal from the rear will be reflected back toward the receiver end of the antenna.

If the termination impedance is merely an open circuit (no terminating resistor), total reflection will result and the antenna will exhibit a bidirectional pattern (still with very deep nulls off the sides). An unterminated Beverage will not have the same response to signals in the rearward direction as it exhibits to signals in the forward direction because of attenuation and reradiation of part of the reflected wave as it travels back toward the receiver end. **Fig 36** compares the response from two 2-λ Beverages, one terminated and the other unterminated. Just like a terminated long-wire transmitting antenna (which is mounted higher off the ground than a Beverage, which is meant only for receiving), the terminated Beverage has a reduced forward lobe compared to its unterminated sibling. The unterminated Beverage exhibits about a 5 dB front-to-back ratio for this length because of the radiation and wire and ground losses that occur before the forward wave gets to the end of the wire.

If the termination is between the extremes (open circuit and perfect termination in Z_{ANT}), the peak direction and intensity of signals off the rear of the Beverage will change. As a result, an adjustable reactive termination can be employed to *steer* the nulls to the rear of the antenna (see **Fig 37**). This can be of great help in eliminating a local interfering signal from a rearward direction (typically 30° to 40° either side of the back direction). Such a scheme doesn't help much for interfering skywave signals because of variations encountered in the ionosphere that constantly shift polarity, amplitude, phase and

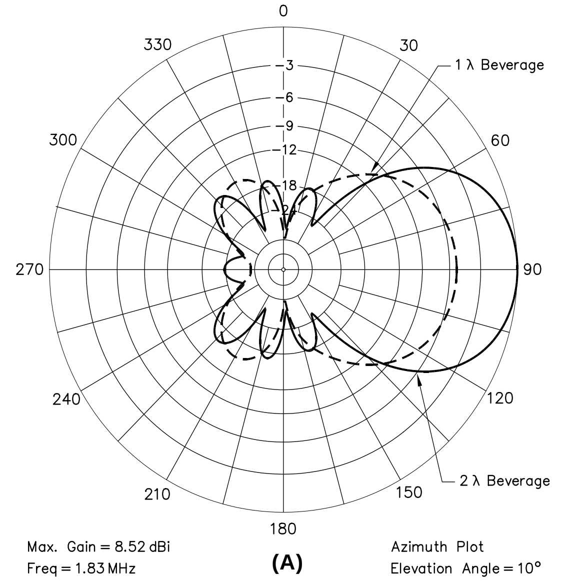

(A)

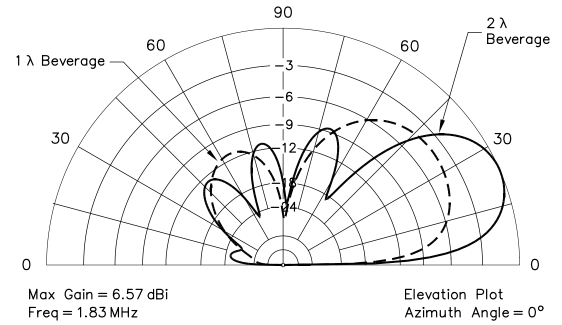

(B)

Fig 35—At A, azimuthal patterns of a 2-λ (solid line) and a 1-λ (dashed line) Beverage antenna, terminated with 550-Ω resistor at 1.83 MHz, at an elevation angle of 10°. The rearward pattern around 180° is more than 20 dB down from the front lobe for each antenna. At B, the elevation-plane patterns. Note the rejection of very high-angle signals near 90°.

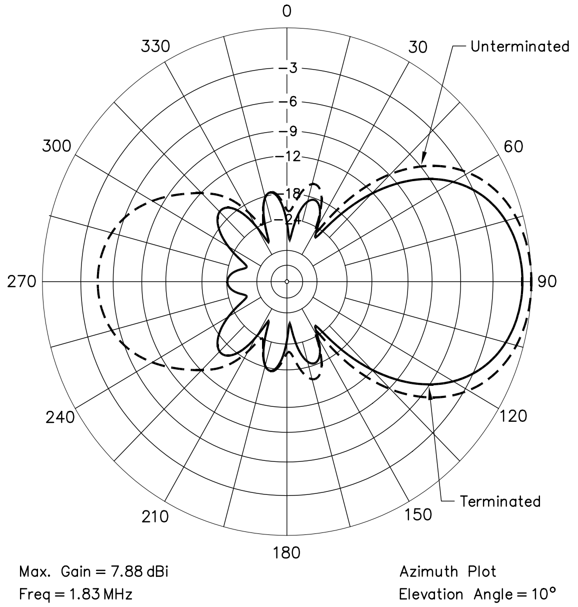

Fig 36—Comparing the azimuthal patterns for a 2-λ Beverage, terminated (solid line) and unterminated (dashed line).

incoming elevation angles.

To determine the appropriate value for a terminating resistor, you need to know the characteristic impedance (surge impedance), Z_{ANT}, of the Beverage. It is interesting to note that Z_{ANT} is not a function of the length, just like a transmission line.

$$Z_{ANT} = 138 \times \log\left(\frac{4h}{d}\right) \qquad \text{(Eq 3)}$$

where

Z_{ANT} = characteristic impedance of the Beverage = terminating resistance needed
h = wire height above ground
d = wire diameter (in the same units as h)

Another aspect of terminating the Beverage is the quality of the RF ground used for the termination. For most types of soil a ground rod is sufficient, since the optimum value for the termination resistance is in the range of 400 to 600 Ω for typical Beverages and the ground-loss resistance is in series with this. Even if the ground-loss resistance at the termination point is as high as 40 or 50 Ω, it still is not an appreciable fraction of the overall terminating resistance. For soil with very poor conductivity, however, (such as sand or rock) you can achieve a better ground termination by laying radial wires on the ground at both the receiver and termination ends. These wires need not be resonant quarter-wave in length, since the ground detunes them anyway. Like the ground counterpoise for a vertical antenna, a number of short radials is better than a few long ones. Some amateurs use chicken-wire ground screens for their ground terminations.

As with many other antennas, improved directivity and gain can be achieved by lengthening the antenna and by arranging several antennas into an array. One item that must be kept in mind is that by virtue of the velocity factor of the antenna, there is some phase shift of the wave on the antenna with respect to the wave in space. Because of this phase shift, although the directivity will continue to sharpen with increased length, there will be some optimum length at which the gain of the antenna will peak. Beyond this

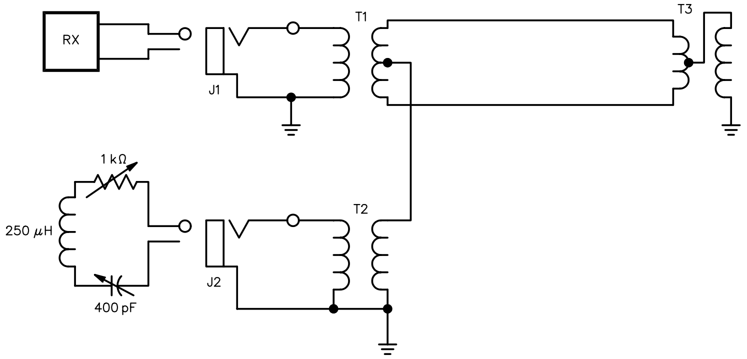

Fig 37—A two-wire Beverage antenna that has provisions for direction switching and null steering in the rear quadrant. Performance improves with height to a point, and is optimum for 1.8-MHz operation at about 10 to 12 feet. Parts identifications are for text reference.

length, the current increments arriving at the receiver end of the antenna will no longer be in phase, and will not add to produce a maximum signal at the receiver end. This optimum length is a function of velocity factor and frequency, and is given by:

$$L = \frac{\lambda}{4\left(\frac{100}{k} - 1\right)} \qquad \text{(Eq 4)}$$

where

L = maximum effective length

λ = signal wavelength in free space (same units as L)

k = velocity factor of the antenna in percent

Because velocity factor increases with height (to a point, as mentioned earlier), optimum length is somewhat longer if the antenna height is increased. The maximum effective length also increases with the number of wires in the antenna system. For example, for a two-wire Beverage like the bidirectional version shown in Fig 37, the maximum effective length is about 20% longer than the single-wire version. A typical length for a single-wire 1.8-MHz Beverage (made of #16 wire and erected 10 feet above ground) is about 1200 feet.

Feed-Point Transformers for Single-Wire Beverages

Matching transformer T1 in Fig 31 is easily constructed. Small toroidal ferrite cores are best for this application, with those of high permeability (μ_i = 125 to 5000) being the easiest to wind (requiring fewest turns) and having the best high-frequency response (because few turns are used). Trifilar-wound autotransformers are most convenient.

Most users are not concerned with a small amount of SWR on the transmission line feeding their Beverages. For example, let us assume that the Z_{ANT} of a particular Beverage is 525 Ω and the terminating resistance is made equal to that value. If a standard 3:1 turns-ratio autotransformer is used at the input end of the antenna, the nominal impedance transformation $50\ \Omega \times 3^2 = 450\ \Omega$. This leads to the terminology often used for this transformer as a *9:1 transformer*, referring to its impedance transformation. The resulting SWR on the feed line going back to the receiver would be 525/450 = 1.27:1, not enough to be concerned about. For a Z_{ANT} of 600 Ω, the SWR is 600/450 = 1.33:1, again not a matter of concern.

Hence, most Beverage users use standard 9:1 (450:50 Ω) autotransformers. You can make a matching transformer suitable for use from 160 to 40 meters using eight trifilar turns of #24 enameled wire wound over a stack of two Amidon FT-50-75 or two MN8-CX cores. See **Fig 38**.

Make your own trifilar cable bundle by placing three 3-foot lengths of the #24 wire side-by-side and twisting them in a hand drill so that there is a uniform twist about one twist-per-inch. This holds the three wires together in a bundle that can be passed through the two stacked cores, rather like threading a needle. Remember that each time you put the bundle through the center of the cores counts as one turn.

After you finish winding, cut the individual wires to leave about $^3/_4$-inch leads, sand off the enamel insulation and tin the wires with a soldering iron. Identify the individual wires with an ohmmeter and then connect them together following Fig 38. Coat the transformer with Q-dope (liquid polystyrene) to finalize the transformer. White glue will work also. See Chapters 25 and 26 and *The ARRL Handbook* for more information on winding toroidal transformers or see Chapter 7 (Special Receiving Antennas) of *ON4UN's Low-Band DXing* book.

The Two-Wire Beverage

The two-wire antenna shown in Fig 37 has the major advantage of having signals from both directions available at the receiver at the flip of a switch between J1 and J2. Also, because there are two wires in the system (equal amounts of signal voltage are induced in both wires), greater signal voltages will be produced.

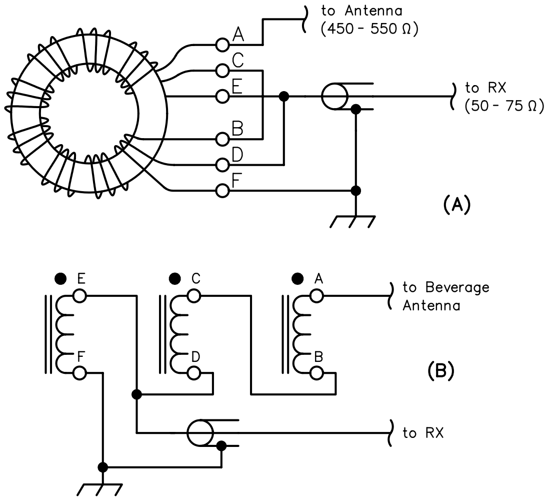

Fig 38—Constructing the feed-point transformer for a single-wire Beverage. See text for details.

A signal from the left direction in Fig 37 induces equal voltages in both wires, and equal in-phase currents flow as a result. The *reflection transformer* (T3 at the right-hand end of the antenna) then inverts the phase of these signals and reflects them back down the antenna toward the receiver, using the antenna wires as a balanced open-wire transmission line. This signal is then transformed by T1 down to the input impedance of the receiver (50 Ω) at J1.

Signals traveling from right to left also induce equal voltages in each wire, and they travel in phase toward the receiver end, through T1, and into T2. Signals from this direction are available at J2.

T1 and T2 are standard 9:1 wideband transformers capable of operating from 1.8 to at least 10 MHz. Like any two parallel wires making up a transmission line, the two-wire Beverage has a certain characteristic impedance—we'll call it Z_1 here—depending on the spacing between the two wires and the insulation between them. T3 transforms the terminating resistance needed at the end of the line to Z_1. Keep in mind that this terminating resistance is equal to the characteristic impedance Z_{ANT} of the Beverage—that is, the impedance of the parallel wires over their images in the ground below. For example, if Z_1 of the Beverage wire is 300 Ω (that is, you used TV twin-lead for the two Beverage wires), T3 must transform the balanced 300 Ω to the unbalanced 500 Ω Z_{ANT} impedance used to terminate the Beverage.

The design and construction of the reflection transformer used in a two-wire Beverage is more demanding than that for the straightforward matching transformer T1 because the exact value of terminating impedance is more critical for good F/B. See Chapter 7 (Special Receiving Antennas) in *ON4UN's Low-Band DXing* for details on winding the reflection transformers for a two-wire Beverage.

Another convenient feature of the two-wire Beverage is the ability to steer the nulls off either end of the antenna while receiving in the opposite direction. For instance, if the series RLC network shown at J2 is adjusted while the receiver is connected to J1, signals can be received from the left direction while interference coming from the right can be partially or completely nulled. The nulls can be steered over a 60° (or more) area off the right-hand end of the antenna. The same null-steering capability exists in the opposite direction with the receiver connected at J2 and the termination connected at J1.

The two-wire Beverage is typically erected at the same height as a single-wire version. The two wires are at the same height and are spaced uniformly—typically 12 to 18 inches apart for discrete wires. Some amateurs construct two-wire Beverages using "window" ladder-line, twisting the line about three twists per foot for mechanical and electrical stability in the wind.

The characteristic impedance Z_{ANT} of a Beverage made using two discrete wires with air insulation between them depends on the wire size, spacing and height and is given by:

$$Z_{ANT} = \frac{69}{\sqrt{\varepsilon}} \times \log\left[\frac{4h}{d}\sqrt{1+\left(\frac{2h}{S}\right)^2}\right] \qquad \text{(Eq 5)}$$

where

Z_{ANT} = Beverage impedance = desired terminating resistance
S = wire spacing
h = height above ground
d = wire diameter (in same units as S and h)
ε = 2.71828

Beverages in Echelon

The pattern of a Beverage receiving antenna is dependent on the terminating resistance used for a particular antenna, as was demonstrated at the extremes by Fig 36. This compared the patterns for a terminated and an unterminated Beverage. The pattern of even a poorly terminated Beverage can be significantly improved by the addition of a second Beverage. The additional Beverage is installed so that it is operated *in echelon,* a word deriving from the fact that the two wires look like the parallel rungs on a ladder. For a practial 160- and 80-meter setup the second Beverage wire is parallel to the first Beverage, spaced from it by about 5 meters, and also staggered 30 meters ahead. See **Fig 38**.

The forward Beverage is fed with a phase difference of +125° such that the total phase, including that due to the forward staggering, is 180°. This forms the equivalent of an end-fire array fed out-of-phase, but it takes advantage of the natural directivity of each Beverage. **Fig 39** compares the pattern of a single 1-λ 160-meter Beverage that is sloppily terminated with two Beverages fed in echelon. The

Beverages in echelon gives a modest additional gain of almost 2 dB. But where the two Beverages in echelon really shine is how they cleans up the rearward pattern—from an average about 15 dB for the single Beverage to more than 25 dB for the two Beverages.

Even at a spacing of 5 meters, there is very little mutual coupling between the two Beverage wires because of their inherently small radiation resistance when they are mounted low above lossy ground. If you adjust for a low SWR (using proper transformers to match the feed-line coaxes), the phase difference will depend solely on the difference in length of the two coaxes feeding the Beverage wires. **Fig 40** shows a wideband feed system designed by Tom Rauch, W8JI, as a "cross-fire" feed system. The 180° wideband phase-inverting transformer allows the system to work on two bands, say 160 and 80 meters. See Chapter 7, Receiving Antennas, in *ON4UN's Low-Band DXing* book, 4th Edition for transformer details.

Practical Considerations

Even though Beverage antennas have excellent directive patterns if terminated properly, gain never exceeds about –3 dBi in most practical installations. However, the directivity that the Beverage provides results in a much higher signal-to-noise ratio for signals in the desired direction than almost any other real-world antenna used at low frequencies.

A typical situation might be a station located in the US Northeast (W1), trying to receive Topband signals from Europe to the northeast, while thunderstorms behind him in the US Southeast (W4) are creating huge static crashes. Instead of listening to an S7 signal with 10-dB over S9 noise and interference on a vertical, the directivity of a Beverage will typically allow you to copy the same signal at perhaps S5 with only S3 (or lower) noise and interference. This is certainly a worthwhile improvement. However, if you are in the middle of a thunderstorm, or if there is a thunderstorm in the direction from which you are trying to receive a signal, no Beverage is going to help you!

There are a few basic principles that must be kept in mind when erecting Beverage antennas if optimum performance is to be realized.

1) Plan the installation thoroughly, including choosing an antenna length consistent with the optimum length values discussed earlier.
2) Keep the antenna as straight and as nearly level as possible over its entire run. Avoid following the terrain under the antenna too closely—keep the antenna level with the average terrain.
3) Minimize the lengths of vertical downleads at the ends of the antenna. Their effect is detrimental to the directive pattern of the antenna. It is best to slope the antenna wire from ground level to its final height (over a distance of 50 feet or so) at the feed-point end. Similar action should be taken at the termination end. Be sure to seal the transformers against weather.

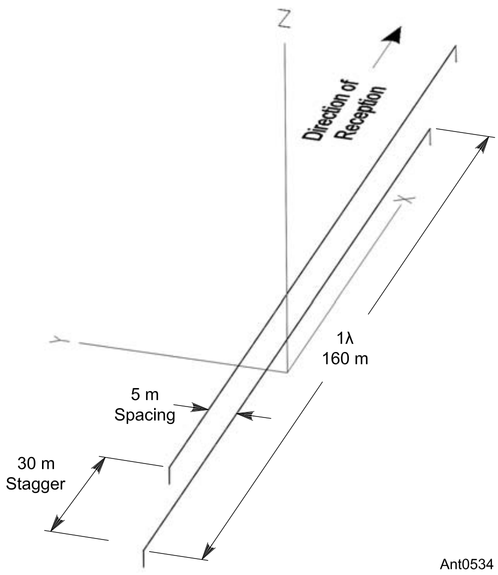

Fig 39—Layout of two 160-meter 1-λ long Beverages in echelon, spaced 5 meters apart, with 30 meter forward stagger. The upper antenna has a 125° phase shift in its feed system.

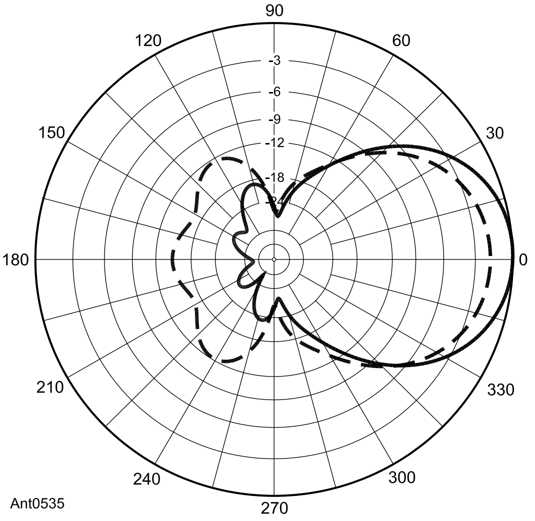

Fig 40—Azimuth pattern at 10° takeoff angle for single Beverage (dashed line) and two Beverages in an echelon end-fire array. The rearward pattern is considerably cleaner on the echelon. Thus, two closely spaced, short Beverages can give considerable improvement over a single short Beverage.

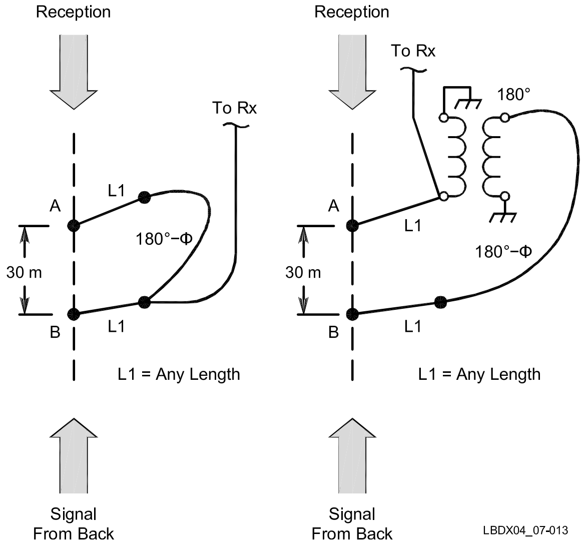

Fig 41—Two ways of feeding the two-Beverage echelon array in Fig 39. On the left, a feed system good for one frequency; on the right, a "cross-fire" feed system good for 1.8 and 3.6 MHz. For this system we want a phase shift due to the coax length of +116° at the back Beverage A. The angle ϕ is thus 180°–116° = 64° long on 160 meters. In the system on the right, a 64° length on 160 meters becomes 128° long on 80 meters. So with the phase-inverting transformer the net phase shift becomes 53° on 80 meters, a reasonable compromise. (*Courtesy W8JI and ON4UN.*)

4) Use a noninductive resistor for terminating a single-wire Beverage. If you live in an area where lightning storms are common, use 2-W terminating resistors, which can survive surges due to nearby lightning strikes.

5) Use high-quality insulators for the Beverage wire where it comes into contact with the supports. Plastic insulators designed for electric fences are inexpensive and effective.

6) Keep the Beverage away from parallel conductors such as electric power and telephone lines for a distance of at least 200 feet. Perpendicular conductors, even other Beverages, may be crossed with relatively little interaction, but do not cross any conductors that may pose a safety hazard.

7) Run the coaxial feed line to the Beverage so that it is not directly under the span of the wire. This prevents common-mode currents from appearing on the shield of the coax. It may be necessary to use a ferrite-bead choke on the feed line if you find that the feed line itself picks up signals when it is temporarily disconnected from the Beverage. See Chapter 26 for details on common-mode chokes.

8) If you use elevated radials in your transmitting antenna system, keep your Beverage feed lines well away from them to avoid stray pickup that will ruin the Beverage's directivity.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

A. Bailey, S. W. Dean and W. T. Wintringham, "The Receiving System for Long-Wave Transatlantic Radio Telephony," *The Bell System Technical Journal*, Apr 1929.

J. S. Belrose, "Beverage Antennas for Amateur Communications," Technical Correspondence, *QST*, Sep 1981.

H. H. Beverage, "Antennas," *RCA Review*, Jul 1939.

H. H. Beverage and D. DeMaw, "The Classic Beverage Antenna Revisited," *QST* , Jan 1982.

B. Boothe, "Weak-Signal Reception on 160—Some Antenna Notes," *QST*, Jun 1977.

E. Bruce, "Developments in Short-Wave Directive Antennas," *Proc IRE*, Aug 1931.

E. Bruce, A. C. Beck and L. R. Lowry, "Horizontal Rhombic Antennas," *Proc IRE*, Jan 1935.

P. S. Carter, C. W. Hansel and N. E. Lindenblad, "Development of Directive Transmitting Antennas by R.C.A. Communications," *Proc IRE*, Oct 1931.

J. Devoldere, *Low-Band DXing* (Newington: ARRL, 1987).

J. Devoldere, *ON4UN's Low-Band DXing* (Newington: ARRL, 1999). See in particular Chapter 7, "Special Receiving Antennas," for many practical details on Beverage antennas.

A. E. Harper, *Rhombic Antenna Design* (New York: D. Van Nostrand Co, Inc).

E. A. Laport, "Design Data for Horizontal Rhombic Antennas," *RCA Review*, Mar 1952.

G. M. Miller, *Modern Electronic Communication* (Englewood Cliffs, NJ: Prentice Hall, 1983).

V. A. Misek, *The Beverage Antenna Handbook* (Wason Rd., Hudson, NH: W1WCR, 1977).

F. E. Terman, *Radio Engineering*, Second Edition (New York: McGraw-Hill, 1937).

Chapter 14

Direction Finding Antennas

The use of radio for direction-finding purposes (RDF) is almost as old as its application for communications. Radio amateurs have learned RDF techniques and found much satisfaction by participating in hidden-transmitter hunts. Other hams have discovered RDF through an interest in boating or aviation, where radio direction finding is used for navigation and emergency location systems.

In many countries of the world, the hunting of hidden amateur transmitters takes on the atmosphere of a sport, as participants wearing jogging togs or track suits dash toward the area where they believe the transmitter is located. The sport is variously known as *fox hunting*, *bunny hunting*, ARDF (Amateur Radio direction finding) or simply transmitter hunting. In North America, most hunting of hidden transmitters is conducted from automobiles, although hunts on foot are gaining popularity.

There are less pleasant RDF applications as well, such as tracking down noise sources or illegal operators from unidentified stations. Jammers of repeaters, traffic nets and other amateur operations can be located with RDF equipment. Or sometimes a stolen amateur rig will be operated by a person who is not familiar with Amateur Radio, and by being lured into making repeated transmissions, the operator unsuspectingly permits himself to be located with RDF equipment. The ability of certain RDF antennas to reject signals from selected directions has also been used to advantage in reducing noise and interference. Through APRS, radio navigation is becoming a popular application of RDF. The locating of downed aircraft is another, and one in which amateurs often lend their skills. Indeed, there are many useful applications for RDF.

Although sophisticated and complex equipment pushing the state of the art has been developed for use by governments and commercial enterprises, relatively simple equipment can be built at home to offer the Radio Amateur an opportunity to RDF. This chapter deals with antennas suitable for that purpose.

RDF by Triangulation

It is impossible, using amateur techniques, to pinpoint the whereabouts of a transmitter from a single receiving location. With a directional antenna you can determine the direction of a signal source, but not how far away it is. To find the distance, you can then travel in the determined direction until you discover the transmitter location. However, that technique can be time consuming and often does not work very well.

A preferred technique is to take at least one additional direction measurement from a second receiving location. Then use a map of the area and plot the bearing or direction measurements as straight lines from points on the map representing the two locations. The approximate location of the transmitter will be indicated by the point where the two bearing lines cross. Even better results can be obtained by taking direction measurements from three locations and using the mapping technique just described. Because absolutely precise bearing measurements are difficult to obtain in practice, the three lines will almost always cross to form a triangle on the map, rather than at a single point. The transmitter will usually be located inside the area represented by the triangle. Additional information on the technique of triangulation and much more on RDF techniques may be found in recent editions of *The ARRL Handbook*.

DIRECTION FINDING SY\STEMS

Required for any RDF system are a directive antenna and a device for detecting the radio signal. In amateur applications the signal detector is usually a transceiver and for convenience it will usually have a meter to indicate signal strength. Unmodified, commercially available portable or mobile receivers are generally quite satisfactory for signal detectors. At very close ranges a simple diode detector and dc microammeter may suffice for the detector.

On the other hand, antennas used for RDF techniques are not generally the types used for normal two-way communications. Directivity is a prime requirement, and here the word *directivity* takes on a somewhat different meaning than is commonly applied to other amateur antennas. Normally we associate directivity with gain, and we think of the ideal antenna pattern as one having a long, thin main lobe. Such a pattern may be of value for coarse measurements in RDF work, but precise bearing measurements are not possible. There is always a spread of a few (or perhaps many) degrees on the *nose* of the lobe, where a shift of antenna bearing produces no detectable change in signal strength. In RDF measure-ments, it is desirable to correlate an exact bearing or compass direction with the position of the antenna. In order to do this as accurately as possible, an antenna exhibiting a *null* in its pattern is used. A null can be very sharp in directivity, to within a half degree or less.

Loop Antennas

A simple antenna for HF RDF work is a small loop tuned to resonance with a capacitor. Several factors must be considered in the design of an RDF loop. The loop must be small in circumference compared with the wavelength. In a single-turn loop, the conductor should be less than 0.08 λ long. For 28 MHz, this represents a length of less than 34 inches (a diameter of approximately 10 inches). Maximum response from the loop antenna is in the plane of the loop, with nulls exhibited at right angles to that plane.

To obtain the most accurate bearings, the loop must be balanced electrostatically with respect to ground. Otherwise, the loop will exhibit two modes of operation. One is the mode of a true loop, while the other is that of an essentially nondirectional vertical antenna of small dimensions. This second mode is called the *antenna effect*. The voltages introduced by the two modes are seldom in phase and may add or subtract, depending upon the direction from which the wave is coming.

The theoretical true loop pattern is illustrated in **Fig 1A**. When properly balanced, the loop exhibits two nulls that are 180° apart. Thus, a single null reading with a small loop antenna will not indicate the exact direction toward the transmitter—only the line along which the transmitter lies. Ways to overcome this ambiguity are discussed later.

When the antenna effect is appreciable and the loop is tuned to resonance, the loop may exhibit little directivity, as shown in Fig 1B. However, by detuning the loop to shift the phasing, a pattern similar to 1C may be obtained. Although this pattern is not symmetrical, it does exhibit a null. Even so, the null may not be as sharp as that obtained with a loop that is well balanced, and it may not be at exact right angles to the plane of the loop.

By suitable detuning, the unidirectional cardioid pattern of Fig 1D may be approached. This adjustment is sometimes used in RDF work to obtain a unidirectional bearing, although there is no complete null in the pattern. A cardioid pattern can also be obtained with a small loop antenna by adding a *sensing element*. Sensing elements are discussed in a later section of this chapter.

An electrostatic balance can be obtained by shielding the loop, as **Fig 2** shows. The shield is represented by the broken lines in the drawing, and eliminates the antenna effect. The response of a well-constructed shielded loop is quite close to the ideal pattern of Fig 1A.

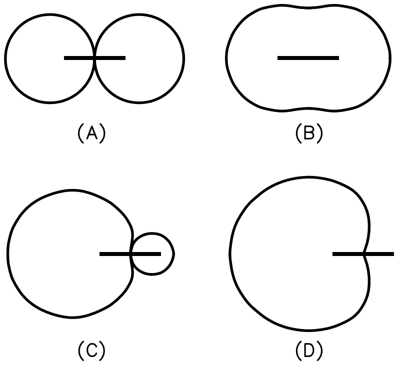

Fig 1—Small-loop field patterns with varying amounts of antenna effect—the undesired response of the loop acting merely as a mass of metal connected to the receiver antenna terminals. The straight lines show the plane of the loop.

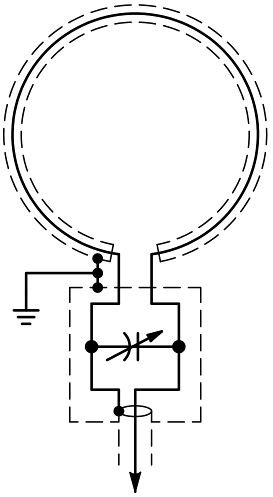

Fig 2—Shielded loop for direction finding. The ends of the shielding turn are not connected, to prevent shielding the loop from magnetic fields. The shield is effective against electric fields.

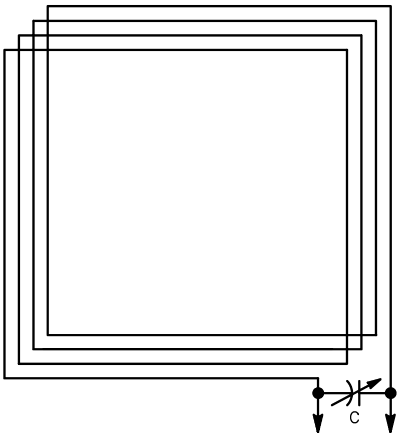

Fig 3—Small loop consisting of several turns of wire. The total conductor length is very much less than a wavelength. Maximum response is in the plane of the loop.

For the low-frequency amateur bands, single-turn loops of convenient physical size for portability are generally found to be unsatisfactory for RDF work. Therefore, multiturn loops are generally used instead. Such a loop is shown in **Fig 3**. This loop may also be shielded, and if the total conductor length remains below 0.08 λ, the directional pattern is that of Fig 1A. A sensing element may also be used with a multiturn loop.

Loop Circuits and Criteria

No single word describes a direction-finding loop of high performance better than *symmetry*. To obtain an undistorted response pattern from this type of antenna, you must build it in the most symmetrical manner possible. The next key word is *balance*. The better the electrical balance, the deeper the loop null and the sharper the maxima.

The physical size of the loop for 7 MHz and below is not of major consequence. A 4-foot diameter loop will exhibit the same electrical characteristics as one which is only an inch or two in diameter. The smaller the loop, however, the lower its efficiency. This is because its aperture samples a smaller section of the wave front. Thus, if you use loops that are very small in terms of a wavelength, you will need preamplifiers to compensate for the reduced efficiency.

An important point to keep in mind about a small loop antenna oriented in a vertical plane is that it is vertically polarized. It should be fed at the bottom for the best null response. Feeding it at one side, rather than at the bottom, will not alter the polarization and will only degrade performance. To obtain horizontal polarization from a small loop, it must be oriented in a horizontal plane, parallel to the earth. In this position the loop response is essentially omnidirectional.

The earliest loop antennas were of the *frame antenna* variety. These were unshielded antennas built on a wooden frame in a rectangular format. The loop conductor could be a single turn of wire (on the larger units) or several turns if the frame was small. Later, shielded versions of the frame antenna became popular, providing electrostatic shielding—an aid to noise reduction from such sources as precipitation static.

Ferrite Rod Antennas

With advances in technology, magnetic-core loop antennas came into use. Their advantage was reduced size, and this appealed especially to the designers of aircraft and portable radios. Most of these antennas contain ferrite bars or cylinders, which provide high inductance and Q with a relatively small number of coil turns.

Magnetic-core antennas consist essentially of turns of wire around a ferrite rod. They are also known as *loopstick* antennas. Probably the best-known example of this type of antenna is that used in small portable AM broadcast receivers. Because of their reduced-size advantage, ferrite-rod antennas are used almost exclusively for portable work at frequencies below 150 MHz.

As implied in the earlier discussion of shielded loops in this chapter, the true loop antenna responds to the magnetic field of the radio wave, and not to the electrical field. The voltage delivered by the loop is proportional to the amount of magnetic flux passing through the coil, and to the number of turns in the coil. The action is much the same as in the secondary winding of a transformer. For a given size of loop, the output voltage can be increased by increasing the flux density, and this is done with a ferrite core of high permeability. A $^{1}/_{2}$-inch diameter, 7-inch rod of Q2 ferrite (m_i = 125) is suitable for a loop core from the broadcast band through 10 MHz. For increased output, the turns may be wound on two rods that are taped together, as shown in **Fig 4**. Loopstick antennas for con-

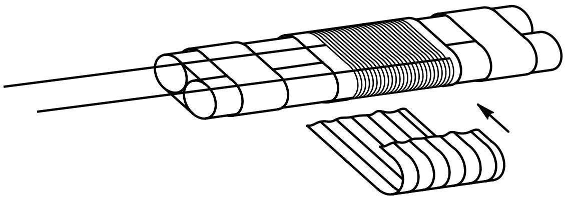

Fig 4—A ferrite-rod or loopstick antenna. Turns of wire may be wound on a single rod, or to increase the output from the loop, the core may be two rods taped together, as shown here. The type of core material must be selected for the intended frequency range of the loop. To avoid bulky windings, fine wire such as #28 or #30 is often used, with larger wire for the leads.

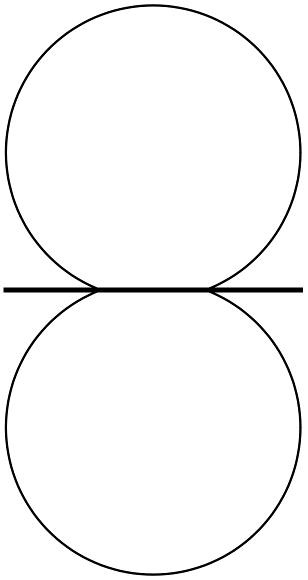

Fig 5—Field pattern for a ferrite rod antenna. The dark bar represents the rod on which the loop turns are wound.

struction are described later in this chapter.

Maximum response of the loopstick antenna is broadside to the axis of the rod as shown in **Fig 5**, whereas maximum response of the ordinary loop is in a direction at right angles to the plane of the loop. Otherwise, the performances of the ferrite-rod antenna and of the ordinary loop are similar. The loopstick may also be shielded to eliminate the antenna effect, such as with a U-shaped or C-shaped channel of aluminum or other form of *trough*. The length of the shield should equal or slightly exceed the length of the rod.

Sensing Antennas

Because there are two nulls that are 180° apart in the directional pattern of a loop or a loopstick, an ambiguity exists as to which one indicates the true direction of the station being tracked. For example, assume you take a bearing measurement and the result indicates the transmitter is somewhere on a line running approximately east and west from your position. With this single reading, you have no way of knowing for sure if the transmitter is east of you or west of you.

If more than one receiving station takes bearings on a single transmitter, or if a single receiving station takes bearings from more than one position on the transmitter, the ambiguity may be worked out by triangulation, as described earlier. However, it is sometimes desirable to have a pattern with only one null, so there is no question about whether the transmitter in the above example would be east or west from your position.

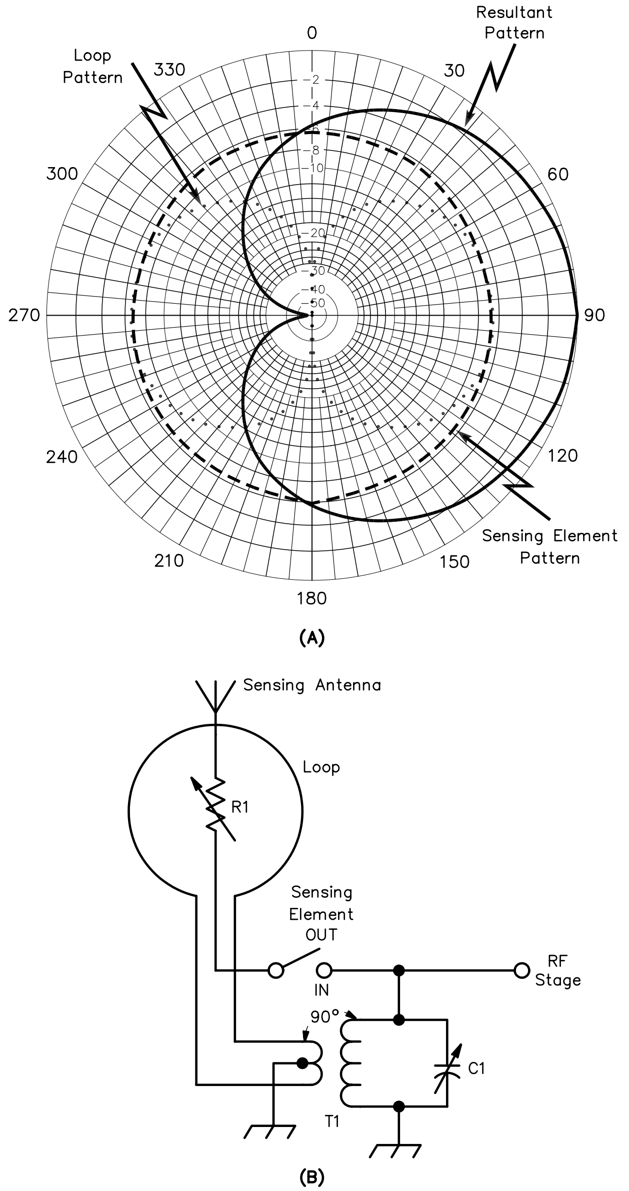

Fig 6—At A, the directivity pattern of a loop antenna with sensing element. At B is a circuit for combining the signals from the two elements. C1 is adjusted for resonance with T1 at the operating frequency.

A loop or loopstick antenna may be made to have a single null if a second antenna element is added. The element is called a *sensing antenna*, because it gives an added sense of direction to the loop pattern. The second element must be omnidirectional, such as a short vertical. When the signals from the loop and the vertical element are combined with a 90° phase shift between the two, a cardioid pattern results. The development of the pattern is shown in **Fig 6A**.

Fig 6B shows a circuit for adding a sensing antenna to a loop or loopstick. R1 is an internal adjustment and is used to set the level of the signal from the sensing antenna. For the best null in the composite pattern, the signals from the loop and the sensing antenna must be of equal amplitude, so R1 is adjusted experimentally during setup. In practice, the null of the cardioid is not as sharp as that of the loop, so the usual measurement procedure is to first use the loop alone to obtain a precise bearing reading, and then to add the sensing antenna and take another reading to resolve the ambiguity. (The null of the cardioid is 90° away from the nulls of the loop.) For this reason, provisions are usually made for switching the sensing element in an out of operation.

PHASED ARRAYS

Phased arrays are also used in amateur RDF work. Two general classifications of phased arrays are end-fire and broadside configurations. Depending on the spacing and phasing of the elements, end-fire patterns may exhibit a null in one direction along the axis of the elements. At the same time, the response is maximum off the other end of the axis, in the opposite direction from the null. A familiar arrangement is two elements spaced $^1/_4$ λ apart and fed 90° out of phase. The resultant pattern is a *cardioid*, with the null in the direction of the leading element. Other arrangements of spacing and phasing for an end-fire array are also suitable for RDF work. One of the best known is the *Adcock array*, discussed in the next section.

Broadside arrays are inherently bidirectional, which means there are always at least two nulls in the pattern. Ambiguity therefore exists in the true direction of the transmitter, but depending on the application, this may be no handicap. Broadside arrays are seldom used for amateur RDF applications however.

The Adcock Antenna

Loops are adequate in RDF applications where only the ground wave is present. The performance of an RDF system for sky-wave reception can be improved by the use of an Adcock antenna, one of the most popular types of end-fire phased arrays. A basic version is shown in **Fig 7**.

This system was invented by F. Adcock and patented in 1919. The array consists of two vertical elements fed 180° apart, and mounted so the system may be rotated. Element spacing is not critical, and may be in the range from 0.1 to 0.75 λ. The two elements must be of identical lengths, but need not be self-resonant. Elements that are shorter than resonant are commonly used. Because neither the element spacing nor the length is critical in terms of wavelengths, an Adcock array may be operated over more than one amateur band.

The response of the Adcock array to vertically polarized waves is similar to a conventional loop, and the directive pattern is essentially the same. Response of the array to a horizontally polarized wave is considerably different from that of a loop, however. The currents induced in the horizontal members tend to balance out regardless of the orientation of the antenna. This effect has been verified in practice, where good nulls were obtained with an experimental Adcock under sky-wave conditions. The same circumstances produced poor nulls with small loops (both conventional and ferrite-loop models).

Generally speaking, the Adcock antenna has attractive properties for amateur RDF applications. Unfortunately, its portability leaves something to be desired, making it more suitable to fixed or semi-portable applications. While a metal support for the mast and boom could be used, wood, PVC or fiberglass are preferable because they are nonconductors and would therefore cause less pattern distortion.

Since the array is balanced, an antenna tuner is required to match the unbalanced input of a typical receiver. **Fig 8** shows a suitable link-coupled network. C2 and C3 are null-balancing capacitors. A low-power signal source is placed some distance from the Adcock

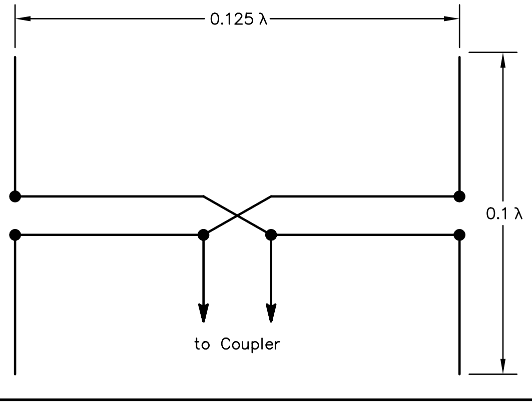

Fig 7—A simple Adcock antenna.

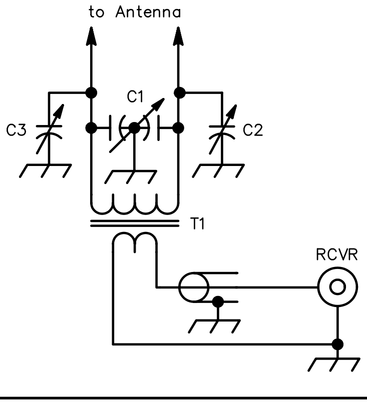

Fig 8—A suitable coupler for use with the Adcock antenna.

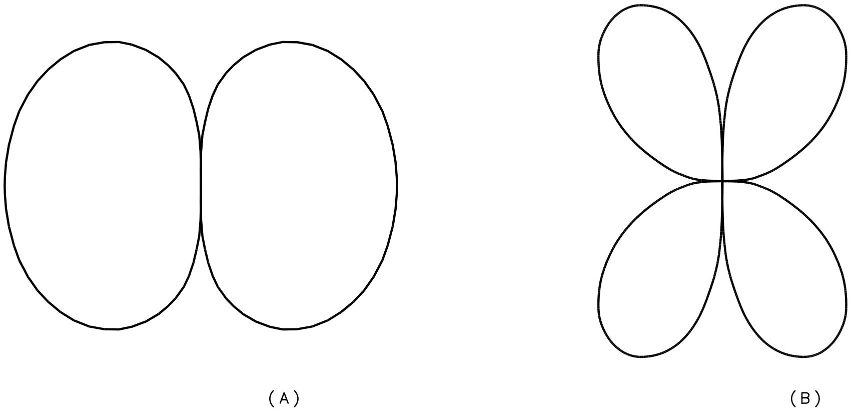

Fig 9—At A, the pattern of the Adcock array with an element spacing of $^{1}/_{2}$ wavelength. In these plots the elements are aligned with the horizontal axis. As the element spacing is increased beyond $^{3}/_{4}$ wavelength, additional nulls develop off the ends of the array, and at a spacing of 1 wavelength the pattern at B exists. This pattern is unsuitable for RDF work.

antenna and broadside to it. C2 and C3 are then adjusted until the deepest null is obtained. The tuner can be placed below the wiring-harness junction on the boom. Connection can be made by means of a short length of 300-Ω twin-lead.

The radiation pattern of the Adcock is shown in **Fig 9A**. The nulls are in directions broadside to the array, and become sharper with greater element spacings. However, with an element spacing greater than 0.75 λ, the pattern begins to take on additional nulls in the directions off the ends of the array axis. At a spacing of 1 λ the pattern is that of Fig 9B, and the array is unsuitable for RDF applications.

Short vertical monopoles are often used in what is sometimes called the *U-Adcock*, so named because the elements with their feeders take on the shape of the letter U. In this arrangement the elements are worked against the earth as a ground or counterpoise. If the array is used only for reception, earth losses are of no great consequence. Short, elevated vertical dipoles are also used in what is sometimes called the *H-Adcock*.

The Adcock array, with two nulls in its pattern, has the same ambiguity as the loop and the loopstick. Adding a sensing element to the Adcock array has not met with great success. Difficulties arise from mutual coupling between the array elements and the sensing element, among other things. Because Adcock arrays are used primarily for fixed-station applications, the ambiguity presents no serious problem. The fixed station is usually one of a group of stations in an RDF network.

LOOPS VERSUS PHASED ARRAYS

Although loops can be made smaller than suitable phased arrays for the same frequency of operation, the phased arrays are preferred by some for a variety of reasons. In general, sharper nulls can be obtained with phased arrays, but this is also a function of the care used in constructing and feeding the individual antennas, as well as of the size of the phased array in terms of wavelengths. The primary constructional consideration is the shielding and balancing of the feed line against unwanted signal pickup, and the balancing of the antenna for a symmetrical pattern.

Loops are not as useful for skywave RDF work because of random polarization of the received signal. Phased arrays are somewhat less sensitive to propagation effects, probably because they are larger for the same frequency of operation and therefore offer some space diversity. In general, loops and loopsticks are used for mobile and portable operation, while phased arrays are used for fixed-station operation. However, phased arrays are used successfully above 144 MHz for portable and mobile RDF work. Practical examples of both types of antennas are presented later in this chapter.

THE GONIOMETER

Most fixed RDF stations for government and commercial work use antenna arrays of stationary elements, rather than mechanically rotatable arrays. This has been true since the earliest days of radio. The early-day device that permits finding directions without moving the elements is called a *radiogoniometer*, or simply a *goniometer*. Various types of goniometers are still used today in many installations, and offer the amateur some possibilities.

The early style of goniometer is a special form of RF transformer, as shown in **Fig 10**. It consists of two fixed coils mounted at right angles to one another. Inside the fixed coils is a movable coil, not shown in Fig 10 to

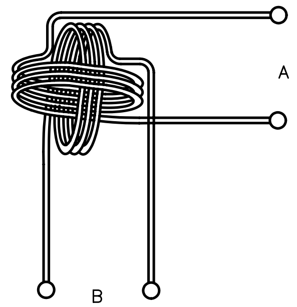

Fig 10—An early type of goniometer that is still used today in some RDF applications. This device is a special type of RF transformer that permits a movable coil in the center (not shown here) to be rotated and determine directions even though the elements are stationary.

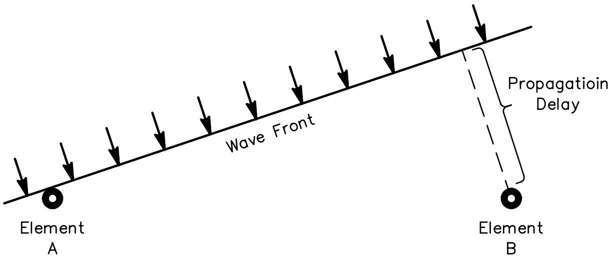

Fig 11—This diagram illustrates one technique used in electronic beam forming. By delaying the signal from element A by an amount equal to the propagation delay, the two signals may be summed precisely in phase, even though the signal is not in the broadside direction. Because this time delay is identical for all frequencies, the system is not frequency sensitive.

avoid cluttering the diagram. The pairs of connections marked A and B are connected respectively to two elements in an array, and the output to the detector or receiver is taken from the movable coil. As the inner coil is rotated, the coupling to one fixed coil increases while that to the other decreases. Both the amplitude and the phase of the signal coupled into the pickup winding are altered with rotation in a way that corresponds to actually rotating the array itself. Therefore, the rotation of the inner coil can be calibrated in degrees to correspond to bearing angles from the station location.

In the early days of radio, the type of goniometer just described saw frequent use with fixed Adcock arrays. A refinement of that system employed four Adcock elements, two arrays at right angles to each other. With a goniometer arrangement, RDF measurements could be taken in all compass directions, as opposed to none off the ends of a two-element fixed array. However, resolution of the 4-element system was not as good as with a single pair of elements, probably because of mutual coupling among the elements. To overcome this difficulty a few systems of eight elements were installed.

Various other types of goniometers have been developed over the years, such as commutator switching to various elements in the array. A later development is the diode switching of capacitors to provide a commutator effect. As mechanical action has gradually been replaced with electronics to “rotate” stationary elements, the word goniometer is used less frequently these days. However, it still appears in many engineering reference texts. The more complex electronic systems of today are called *beam-forming networks*.

Electronic Antenna Rotation

With an array of many fixed elements, beam rotation can be performed electronically by sampling and combining signals from various individual elements in the array. Contingent upon the total number of elements in the system and their physical arrangement, almost any desired antenna pattern can be formed by summing the sampled signals in appropriate amplitude and phase relationships. Delay networks are used for some of the elements before the summation is performed. In addition, attenuators may be used for some elements to develop patterns such as from an array with binomial current distribution.

One system using these techniques is the *Wullenweber* antenna, employed primarily in government and military installations. The Wullenweber consists of a very large number of elements arranged in a circle, usually outside of (or in front of) a circular reflecting screen. Depending on the installation, the circle may be anywhere from a few hundred feet to more than a quarter of a mile in diameter. Although the Wullenweber is not one that would be constructed by an amateur, some of the techniques it uses may certainly be applied to Amateur Radio.

For the moment, consider just two elements of a Wullenweber antenna, shown as A and B in **Fig 11**. Also shown is the wavefront of a radio signal arriving from a distant transmitter. As drawn, the wavefront strikes element A first, and must travel somewhat farther before it strikes element B. There is a finite time delay before the wavefront reaches element B.

The propagation delay may be measured by delaying the signal received at element A before summing it with that from element B. If the two signals are combined directly, the amplitude of the resultant signal will be maximum when the delay for element A exactly equals the propagation delay. This results in an in-phase condition at the summation point. Or if one of the signals is inverted and the two are summed, a null will exist when the element-A delay equals the propagation delay; the signals will combine in a 180° out-of-phase relationship. Either way, once the time delay is known, it may be converted to distance. Then the direction from which the wave is arriving may be determined by trigonometry.

By altering the delay in small increments, the peak of the antenna lobe (or the null) can be steered in azimuth. This is true without regard to the frequency of the incoming wave. Thus, as long as the delay is less than the period of one RF cycle, the system is not frequency sensitive, other than for the frequency range that may be covered satisfactorily by the array elements themselves. Surface acoustic wave (SAW) devices or lumped-constant networks can be used for delay lines in such systems if the system is used only for receiving. Rolls of coaxial cable of various lengths are used in installations for transmitting. In this case, the lines are considered for the time delay they provide, rather than as simple phasing lines. The difference is that a phasing line is ordinarily designed for a single frequency (or for an amateur band), while a delay line offers essentially the same time delay at all frequencies.

By combining signals from other Wullenweber elements appropriately, the broad beamwidth of the pattern from the two elements can be narrowed, and unwanted sidelobes can be suppressed. Then, by electronically switching the delays and attenuations to the various elements, the beam so formed can be rotated around the compass. The package of electronics designed to do this, including delay lines and electronically switched attenuators, is the beam-forming network. However, the Wullenweber system is not restricted to forming a single beam. With an isolation amplifier provided for each element of the array, several beam-forming networks can be operated independently. Imagine having an antenna system that offers a dipole pattern, a rhombic pattern, and a Yagi beam pattern, all simultaneously and without frequency sensitivity. One or more may be rotating while another is held in a particular direction. The Wullenweber was designed to fulfill this type of requirement.

One feature of the Wullenweber antenna is that it can operate 360° around the compass. In many government installations, there is no need for such coverage, as the areas of interest lie in an azimuth sector. In such cases an in-line array of elements with a backscreen or curtain reflector may be installed broadside to the center of the sector. By using the same techniques as the Wullenweber, the beams formed from this array may be slewed left and right across the sector. The maximum sector width available will depend on the installation, but beyond 70° to 80° the patterns begin to deteriorate to the point that they are unsatisfactory for precise RDF work.

RDF SYSTEM CALIBRATION AND USE

Once an RDF system is initially assembled, it should be calibrated or checked out before actually being put into use. Of primary concern is the balance or symmetry of the antenna pattern. A lop-sided figure-8 pattern with a loop, for example, is undesirable; the nulls are not 180° apart, nor are they at exact right angles to the plane of the loop. If you didn't know this fact in actual RDF work, measurement accuracy would suffer.

Initial checkout can be performed with a low-powered transmitter at a distance of a few hundred feet. It should be within visual range and must be operating into a vertical antenna. (A quarter-wave vertical or a loaded whip is quite suitable.) The site must be reasonably clear of obstructions, especially steel and concrete or brick buildings, large metal objects, nearby power lines, and so on. If the system operates above 30 MHz, you should also avoid trees and large bushes. An open field makes an excellent site.

The procedure is to find the transmitter with the RDF equipment as if its position were not known, and compare the RDF null indication with the visual path to the transmitter. For antennas having more than one null, each null should be checked.

If imbalance is found in the antenna system, there are two options available. One is to correct the imbalance. Toward this end, pay particular attention to the feed line. Using a coaxial feeder for a balanced antenna invites an asymmetrical pattern, unless an effective balun is used. A balun is not necessary if the loop is shielded, but an asymmetrical pattern can result with misplacement of the break in the shield itself. The builder may also find that the presence of a sensing antenna upsets the balance slightly, due to mutual coupling. Experiment with its position with respect to the main antenna to correct the error. You will also note that the position of the null shifts by 90° as the sensing element is switched in and out, and the null is not as deep. This is of little concern, however, as the intent of the sensing antenna is only to resolve ambiguities. The sensing element should be switched out when accuracy is desired.

The second option is to accept the imbalance of the antenna and use some kind of indicator to show the true directions of the nulls. Small pointers, painted marks on the mast, or an optical sighting system might be used. Sometimes the end result of the calibration procedure will be a compromise between these two options, as a perfect electrical balance may be difficult or impossible to attain.

The discussion above is oriented toward calibrating portable RDF systems. The same general suggestions apply if the RDF array is fixed, such as an Adcock. However, it won't be possible to move it to an open field. Instead, the array must be calibrated in its intended operating position through the use of a portable or mobile transmitter. Because of nearby obstructions or reflecting objects, the null in the pattern may not appear to indicate the precise direction of the transmitter. Do not confuse this with imbalance in the RDF array. Check for imbalance by rotating the array 180° and comparing readings.

Once the balance is satisfactory, you should make a table of bearing errors noted in different compass directions. These error values should be applied as corrections when actual measurements are made. The mobile or por-

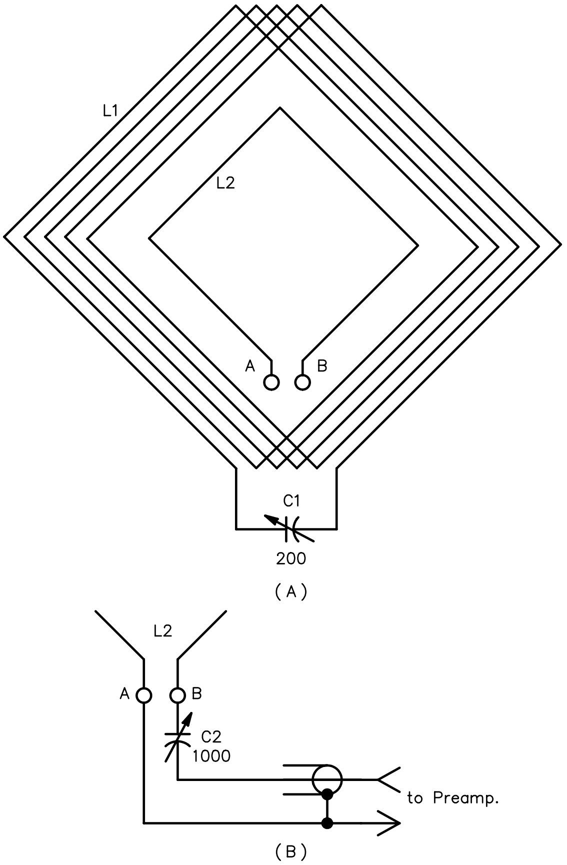

Fig 12—A multiturn frame antenna is shown at A. L2 is the coupling loop. The drawing at B shows how L2 is connected to a preamplifier.

table transmitter should be at a distance of two or three miles for these measurements, and should be in as clear an area as possible during transmissions. The idea is to avoid conduction of the signal along power lines and other overhead wiring from the transmitter to the RDF site. Of course the position of the transmitter must be known accurately for each transmission.

FRAME LOOPS

It was mentioned earlier that the earliest style of receiving loops was the frame antenna. If carefully constructed, such an antenna performs well and can be built at low cost. **Fig 12** illustrates the details of a practical frame type of loop antenna. This antenna was designed by Doug DeMaw, W1FB, and described in *QST* for July 1977. (See the Bibliography at the end of this chapter.) The circuit in Fig 12A is a 5-turn system tuned to resonance by C1. If the layout is symmetrical, good balance should be obtained. L2 helps to achieve this objective by eliminating the need for direct coupling to the feed ter-

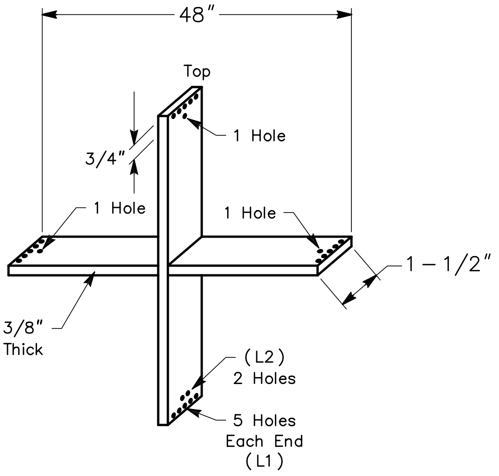

Fig 13—A wooden frame can be used to contain the wire of the loop shown in Fig 12.

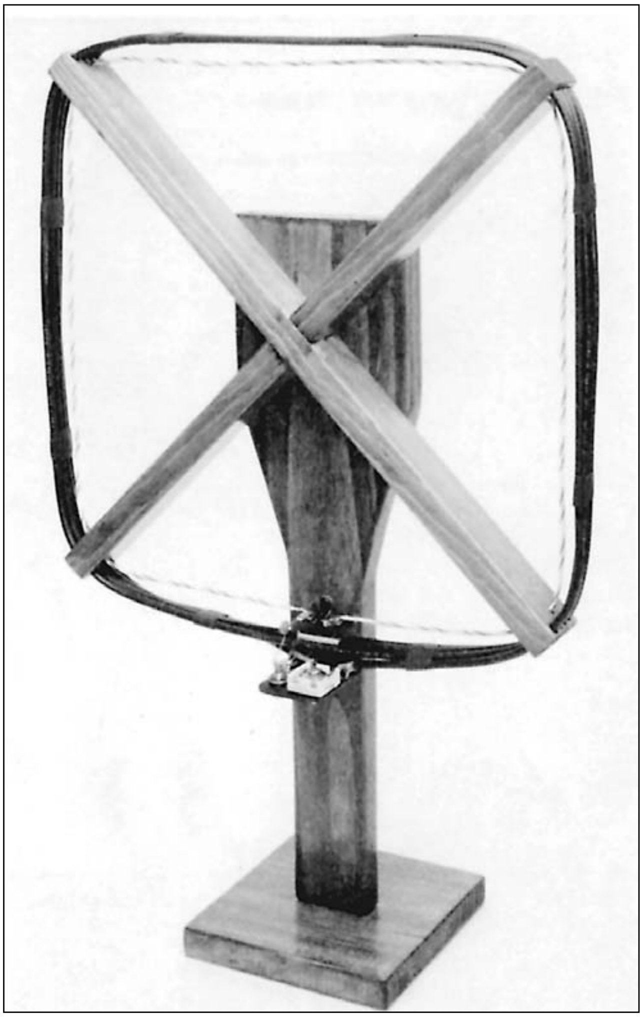

Fig 14—An assembled table-top version of the electrostatically shielded loop. RG-58 cable is used in its construction.

minals of L1. If the loop feed were attached in parallel with C1, a common practice, the chance for imbalance would be considerable.

L2 can be situated just inside or slightly outside of L1; a 1-inch separation works nicely. The receiver or preamplifier can be connected to terminals A and B of L2, as shown in Fig 12B. C2 controls the amount of coupling between the loop and the preamplifier. The lighter the coupling, the higher is the loop Q, the narrower is the frequency response, and the greater is the gain requirement from the preamplifier. It should be noted that no attempt is being made to match the extremely low loop impedance to the preamplifier.

A supporting frame for the loop of Fig 12 can be constructed of wood, as shown in **Fig 13**. The dimensions given are for a 1.8-MHz frame antenna. For use on 75 or 40 meters, L1 of Fig 12A will require fewer turns, or the size of the wooden frame should be made somewhat smaller than that of Fig 13.

SHIELDED FRAME LOOPS

If electrostatic shielding is desired, the format shown in **Fig 14** can be adopted. In this example, the loop conductor and the single-turn coupling loop are made from RG-58 coaxial cable. The number of loop turns should be sufficient to resonate with the tuning capacitor at the operating frequency. Antenna resonance can be checked by first connecting C1 (Fig 12A) and setting it at midrange. Then connect a small 3-turn coil to the loop feed terminals, and couple to it with a dip meter. Just remember that the pickup coil will act to lower the frequency slightly from actual resonance.

In the antenna photographed for Fig 14, the 1-turn coupling loop was made of #22 plastic-insulated wire. However, electrostatic noise pickup occurs on such a coupling loop, noise of the same nature that the shield on the main loop prevents. This can be avoided by using RG-58 for the coupling loop. The shield of the coupling loop should be opened for about 1 inch at the top, and each end of the shield grounded to the shield of the main loop.

Larger single-turn frame loops can be fashioned from aluminum-jacketed Hardline, if that style of coax is available. In either case, the shield conductor must be opened at the electrical center of the loop, as shown in **Fig 15** at A and B. The design example is for 1.8-MHz operation.

To realize the best performance from an electrostatically shielded loop antenna, you must operate it near to and directly above an effective ground plane. An automobile roof (metal) qualifies nicely for small shielded loops. For fixed-station use, a chicken-wire ground screen can be placed below the antenna at a distance of 1 to 6 feet.

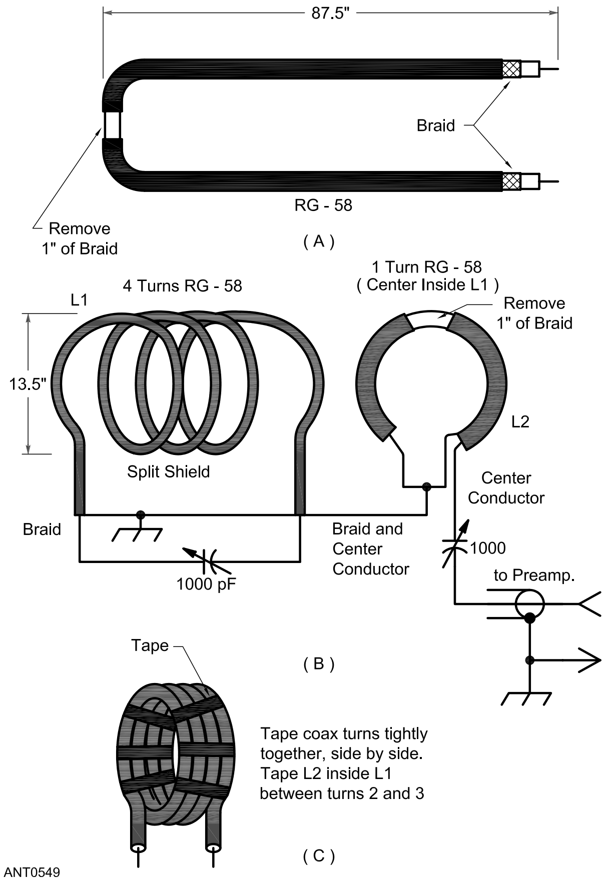
87.5"
Braid
RG - 58
Remove
1" of Braid
(A)
4 Turns RG - 58
L1
1 Turn RG - 58
(Center Inside L1)
Remove
1" of Braid
13.5"
L2
Split Shield
Center
Conductor
Braid
Braid and
Center
Conductor
1000 pF
1000
to Preamp.
Tape
(B)
Tape coax turns tightly
together, side by side.
Tape L2 inside L1
between turns 2 and 3
(C)
ANT0549

Fig 15—Components and assembly details of the shielded loop shown in Fig 14.

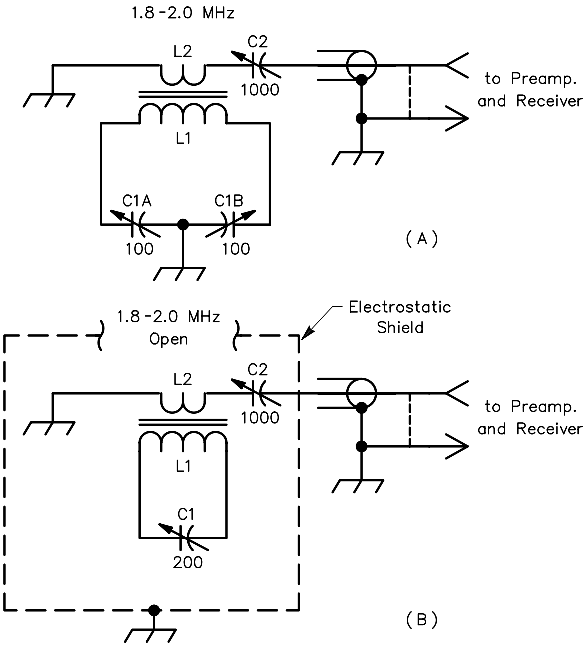
1.8 -2.0 MHz
L2
C2
1000
to Preamp.
and Receiver
L1
C1A
C1B
100
100
(A)
Electrostatic
Shield
1.8 -2.0 MHz
Open
L2
C2
1000
to Preamp.
and Receiver
L1
C1
200
(B)

Fig 16—At A, the diagram of a ferrite loop. C1 is a dual-section air-variable capacitor. The circuit at B shows a rod loop contained in an electrostatic shield channel (see text). A suitable low-noise preamplifier is shown in Fig 19.

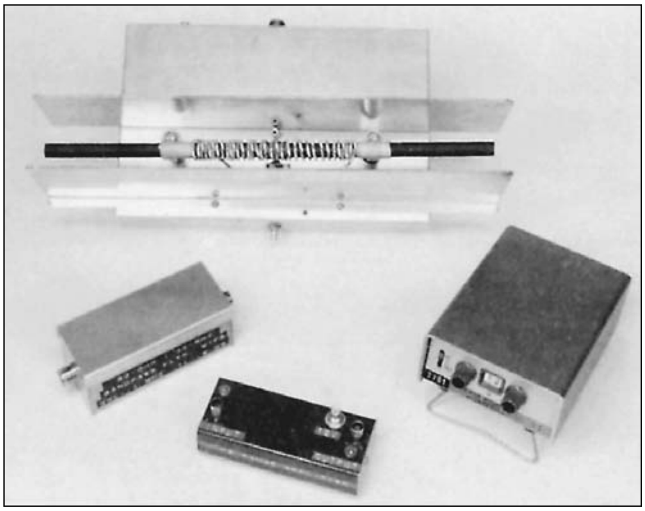

Fig 17—The assembly at the top of the picture is a shielded ferrite-rod loop for 160 meters. Two rods have been glued end to end (see text). The other units in the picture are a low-pass filter (lower left), broadband preamplifier (lower center) and a Tektronix step attenuator (lower right). These were part of the test setup used when the antenna was evaluated.

FERRITE-CORE LOOPS

Fig 16 contains a diagram for a rod loop (loopstick antenna). This antenna was also designed by Doug DeMaw, W1FB, and described in *QST* for July 1977. The winding (L1) has the appropriate number of turns to permit resonance with C1 at the operating frequency. L1 should be spread over approximately 1/3 of the core center. Litz wire will yield the best Q, but Formvar magnet wire can be used if desired. A layer of 3M Company glass tape (or Mylar tape) is recommended as a covering for the core before adding the wire. Masking tape can be used if nothing else is available.

L2 functions as a coupling link over the exact center of L1. C1 is a dual-section variable capacitor, although a differential capacitor might be better toward obtaining optimum balance. The loop Q is controlled by means of C2, which is a mica-compression trimmer.

Electrostatic shielding of rod loops can be effected by centering the rod in a U-shaped aluminum, brass or copper channel, extending slightly beyond the ends of the rod loop (1 inch is suitable). The open side (top) of the channel can't be closed, as that would constitute a shorted turn and render the antenna useless. This can be proved by shorting across the center of the channel with a screwdriver blade when the loop is tuned to an incoming signal. The shield-braid gap in the coaxial loop of Fig 15 is maintained for the same reason.

Fig 17 shows the shielded rod loop assembly. This antenna was developed experimentally for 160 meters and uses two 7-inch ferrite rods, glued together end-to-end with epoxy cement. The longer core resulted in improved sensitivity for weak-signal reception. The other items in the photograph were used during the evaluation tests and are not pertinent to this discussion. This loop and the frame loop discussed in the previous section have bidirectional nulls, as shown in Fig 1A.

Obtaining a Cardioid Pattern

Although the bidirectional pattern of loop antennas

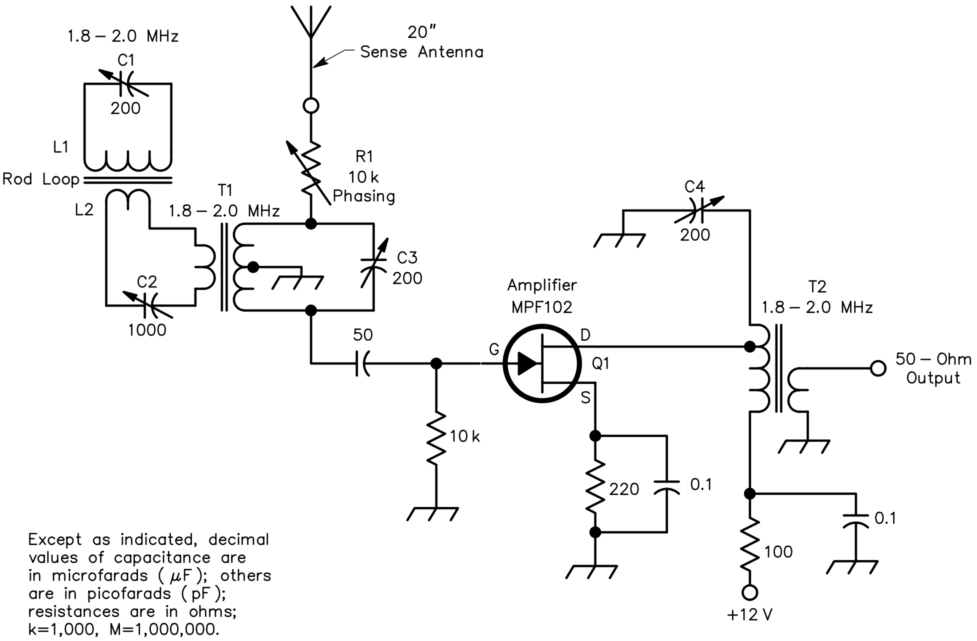
1.8 – 2.0 MHz
C1
200
L1
Rod Loop
L2
T1
1.8 – 2.0 MHz
C2
1000
20"
Sense Antenna
R1
10 k
Phasing
C3
200
50
10 k
Amplifier
MPF102
G
D
Q1
S
220
0.1
C4
200
T2
1.8 – 2.0 MHz
50 – Ohm
Output
0.1
100
+12 V
Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.

Fig 18—Schematic diagram of a rod-loop antenna with a cardioid response. The sensing antenna, phasing network and a preamplifier are shown also. The secondary of T1 and the primary of T2 are tuned to resonance at the operating frequency of the loop. T-68-2 to T-68-6 Amidon toroid cores are suitable for both transformers. Amidon also sells ferrite rods for this type of antenna.

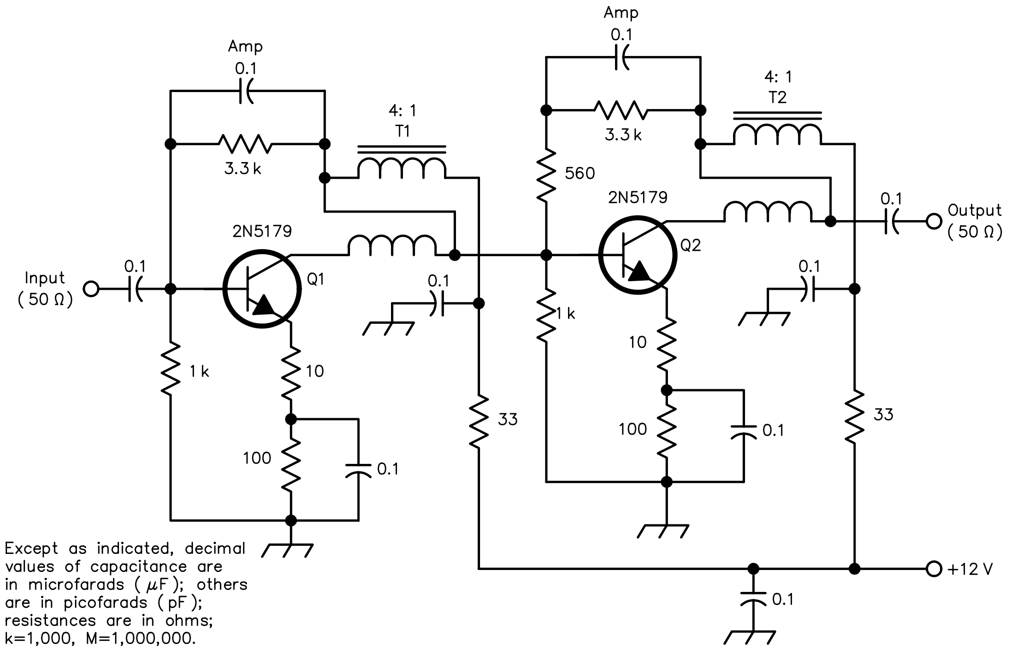

Fig 19—Schematic diagram of a two-stage broadband amplifier patterned after a design by Wes Hayward, W7ZOI. T1 and T2 have a 4:1 impedance ratio and are wound on FT-50-61 toroid cores (Amidon) which have a μ_i of 125. They contain 12 turns of #24 enamel wire, bifilar wound. The capacitors are disc ceramic. This amplifier should be built on double-sided circuit board for best stability.

can be used effectively in tracking down signal sources by means of triangulation, an essentially unidirectional loop response will help to reduce the time spent finding the fox. Adding a sensing antenna to the loop is simple to do, and it will provide the desired cardioid response. The theoretical pattern for this combination is shown in Fig 1D.

Fig 18 shows how a sensing element can be added to a loop or loopstick antenna. The link from the loop is connected by coaxial cable to the primary of T1, which is a tuned toroidal transformer with a split secondary winding. C3 is adjusted for peak signal response at the frequency of interest (as is C4), then R1 is adjusted for minimum back response of the loop. It will be necessary to readjust C3 and R1 several times to compensate for the interaction of these controls. The adjustments are repeated until no further null depth can be obtained. Tests at ARRL Headquarters showed that null depths as great as 40 dB could be obtained with the circuit of Fig 18 on 75 meters. A near-field weak-signal source was used during the tests.

The greater the null depth, the lower the signal output from the system, so plan to include a preamplifier with 25 to 40 dB of gain. Q1 shown in Fig 18 will deliver approximately 15 dB of gain. The circuit of **Fig 19** can be used following T2 to obtain an additional 24 dB of gain. In the interest of maintaining a good noise figure, even at 1.8 MHz, Q1 should be a low-noise device. A 2N4416, an MPF102, or a 40673 MOSFET would be satisfactory. The sensing antenna can be mounted 6 to 15 inches from the loop. The vertical whip need not be more than 12 to 20 inches long. Some experimenting may be necessary in order to obtain the best results. Optimization will also change with the operating frequency of the antenna.

A SHIELDED LOOP WITH SENSING ANTENNA FOR 28 MHz

Fig 20 shows the construction and mounting of a simple shielded 10-meter loop. The loop was designed by Loren Norberg, W9PYG, and described in *QST* for April 1954. (See the Bibliography at the end of this chapter.) It is made from an 18-inch length of RG-11 coax (either solid or foam dielectric) secured to an aluminum box of any convenient size, with two coaxial cable hoods (Amphenol 83-1HP). The outer shield must be broken at the exact center. C1 is a 25-pF variable capacitor, and is connected in parallel with a 33-pF fixed mica padder capacitor, C3. C1 must be tuned to the desired frequency while the loop is connected to the receiver in the same way as it will be used for RDF. C2 is a small differential capacitor used to provide electrical symmetry. The lead-in to the receiver is 67 inches of RG-59 (82 inches if the cable has a foamed dielectric).

The loop can be mounted on the roof of the car with a rubber suction cup. The builder might also fabricate some kind of bracket assembly to mount the loop temporarily in the window opening of the automobile, allowing for loop rotation. Reasonably true bearings may be obtained through the windshield when the car is pointed in the direction of the hidden transmitter. More accurate bearings may be obtained with the loop held out the window and the signal coming toward that side of the car.

Sometimes the car broadcast antenna may interfere with accurate bearings. Disconnecting the antenna from the broadcast receiver may eliminate this trouble.

Sensing Antenna

A sensing antenna can be added to Norberg's loop above to determine which of the two directions indicated by the loop is the correct one. Add a phono jack to the top of the aluminum case shown in Fig 20. The insulated center terminal of the jack should be connected to the side of the tuning capacitors that is common to the center conductor of the RG-59 coax feed line. The jack then takes a short vertical antenna rod of the diameter to fit the jack, or a piece of #12 or #14 solid wire may be soldered to the center pin of a phono plug for insertion in the jack. The sensing antenna can be plugged in as needed. Starting with a length of about four times the loop diameter, the length of the sensing antenna should be pruned until the pattern is similar to that of Fig 1D.

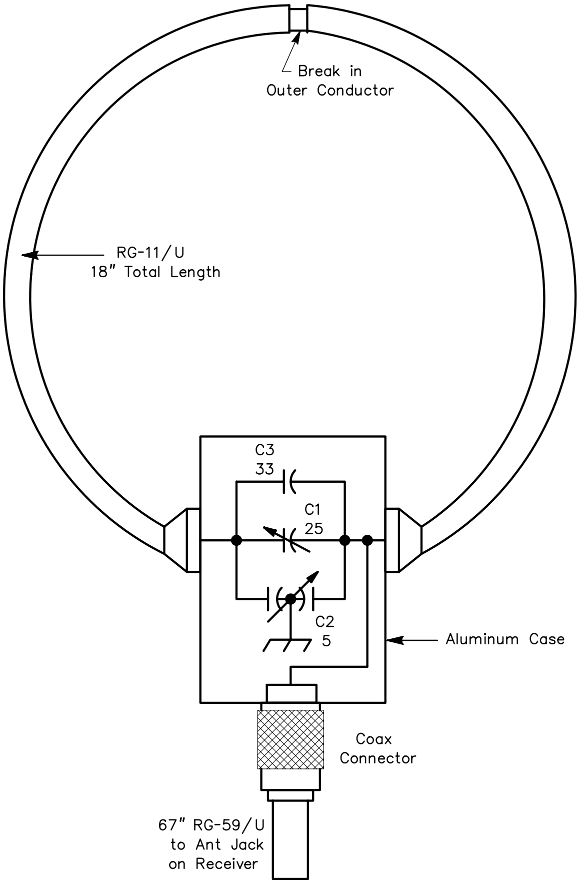

Fig 20—Sketch showing the constructional details of the 28-MHz RDF loop. The outer braid of the coax loop is broken at the center of the loop. The gap is covered with waterproof tape, and the entire assembly is given a coat of acrylic spray.

THE SNOOP LOOP—FOR CLOSE-RANGE HF RDF

Picture yourself on a hunt for a hidden 28-MHz transmitter. The night is dark, very dark. After you take off at the start of the hunt, heading in the right direction, the signal gets stronger and stronger. Your excitement increases with each additional S unit on the meter. You follow your loop closely, and it is working perfectly. You're getting out of town and into the countryside. The roads are unfamiliar. Now the null is beginning to swing rather rapidly, showing that you are getting close.

Suddenly the null shifts to give a direction at right angles to the car. With your flashlight you look carefully across the deep ditch beside the road and into the dark field where you know the transmitter is hidden. There are no roads into the field as far as you can see in either direction. You dare not waste miles driving up and down the road looking for an entrance, for each tenth of a mile counts. But what to do?—Your HF transceiver is mounted in your car and requires power from your car battery.

In a brief moment your decision is made. You park

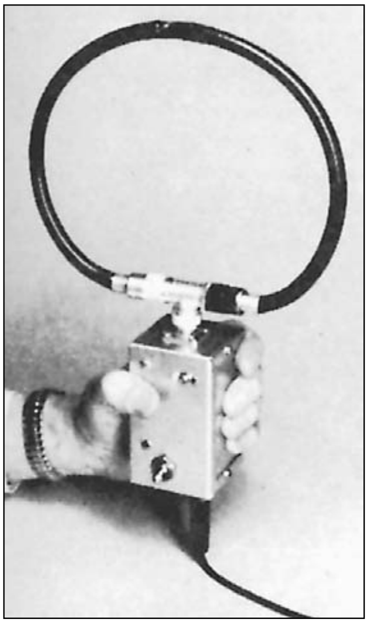

Fig 21—The box containing the detector and amplifier is also the "handle" for the Snoop Loop. The loop is mounted with a coax T as a support, a convenience but not an essential part of the loop assembly. The loop tuning capacitor is screwdriver adjusted. The on-off switch and the meter sensitivity control may be mounted on the bottom.

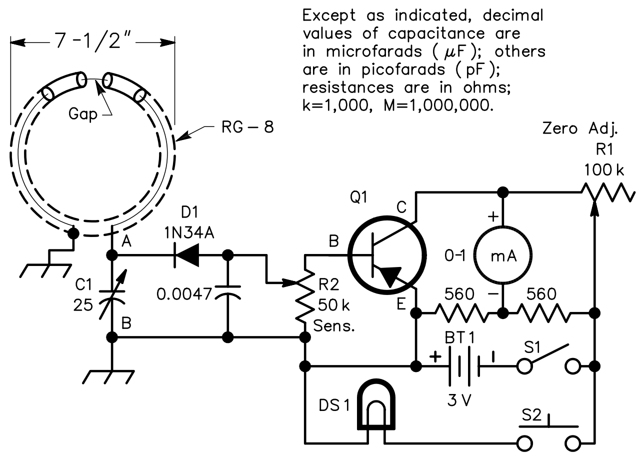
7-1/2″
Gap
RG-8
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.
Zero Adj.
R1
100 k
Q1
C
B
E
D1
1N34A
A
C1
25
B
0.0047
R2
50 k
Sens.
+
0-1
mA
560
−
560
BT1
+
3 V
S1
S2
DS1

Fig 22—The Snoop Loop circuit for 28-MHz operation. The loop is a single turn of RG-8 inner conductor, the outer conductor being used as a shield. Note the gap in the shielding; about a 1-inch section of the outer conductor should be cut out. Refer to Fig 23 for alternative connection at points A and B for other frequencies of operation.

BT1—Two penlight cells.
C1—25-pF midget air padder.
D1—Small-signal germanium diode such as 1N34A or equiv.
DS1—Optional 2-cell penlight lamp for meter illumination, such as no. 222.
Q1—PNP transistor such as ECG102 or equiv.
R1—100-kΩ potentiometer, linear taper. May be PC-mount style.
R2—50-kΩ potentiometer, linear taper.
S1—SPST toggle.
S2—Optional momentary push for illuminating meter.

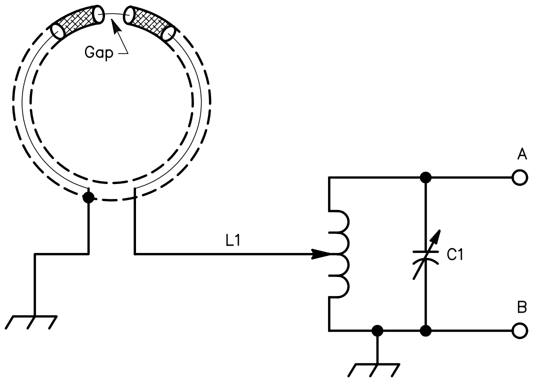

Fig 23—Input circuit for lower frequency bands. Points A and B are connected to corresponding points in the circuit of Fig 22, substituting for the loop and C1 in that circuit. L1-C1 should resonate within the desired amateur band, but the L/C ratio is not critical. After construction is completed, adjust the position of the tap on L1 for maximum signal strength. Instead of connecting the RDF loop directly to the tap on L1, a length of low impedance line may be used between the loop and the tuned circuit, L1-C1.

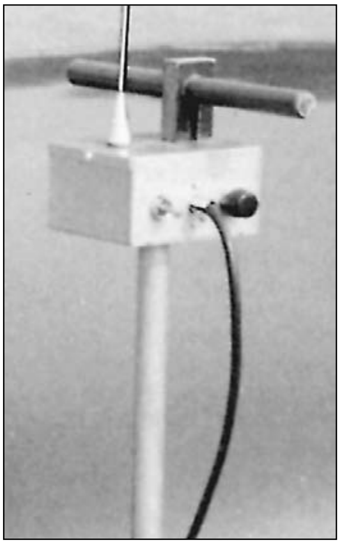

Fig 24—Unidirectional 3.5-MHz RDF using ferrite-core loop with sensing antenna. Adjustable components of the circuit are mounted in the aluminum chassis supported by a short length of tubing.

beside the road, take your flashlight, and plunge into the veldt in the direction your loop null clearly indicated. But after taking a few steps, you're up to your armpits in brush and can't see anything forward or backward. You stumble on in hopes of running into the hidden transmitter—you're probably not more than a few hundred feet from it. But away from your car and radio equipment, it's like the proverbial hunt for the needle in the haystack. What you really need is a portable setup for hunting at close range, and you may prefer something that is inexpensive. The Snoop Loop was designed for just these requirements by Claude Maer, Jr, W0IC, and was described in *QST* for February 1957. (See the Bibliography at the end of this chapter.)

The Snoop Loop is pictured in **Fig 21**. The loop itself is made from a length of RG-8 coax, with the shield broken at the top. A coax T connector is used for convenience and ease of mounting. One end of the coax loop is connected to a male plug in the conventional way, but the center conductor of the other end is shorted to the shield so the male connector at that end has no connection to the center prong. This results in an unbalanced circuit, but seems to give good bidirectional null readings, as well as an easily detectable maximum reading when the grounded end of the loop is pointed in the direction of the transmitter. Careful tuning with C1 will improve this maximum reading. Don't forget to remove one inch of shielding from the top of the loop. You won't get much signal unless you do.

The detector and amplifier circuit for the Snoop Loop

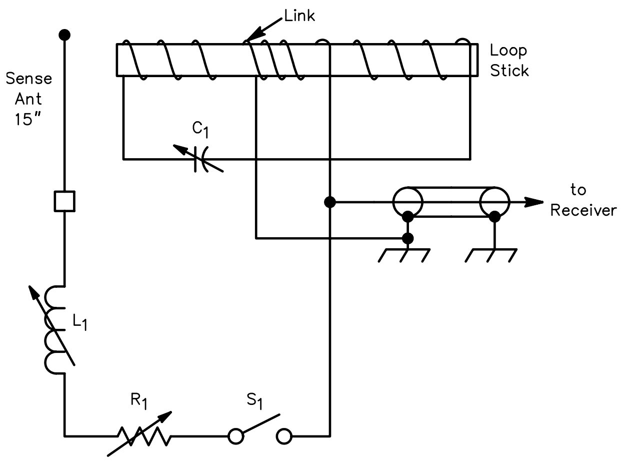

Fig 25—Circuit of the 3.5-MHz direction finder loop.

C1—140 pF variable (125-pF ceramic trimmer in parallel with 15-pF ceramic fixed.
L1—Approximately 140 µH adjustable (Miller No. 4512 or equivalent.
R1—1-kΩ carbon potentiometer.
S1—SPST toggle.
Loopstick—Approximately 15 µH (Miller 705-A, with original winding removed and wound with 20 turns of #22 enamel). Link is two turns at center. Winding ends secured with Scotch electrical tape. This type of ferrite rod may also be found in surplus transistor AM radios.

is shown in **Fig 22**. The model photographed does not include the meter, as it was built for use only with high-impedance headphones. The components are housed in an aluminum box. Almost any size box of sufficient size to contain the meter can be used. At very close ranges, reduction of sensitivity with R2 will prevent pegging the meter.

The Snoop Loop is not limited to the 10-meter band or to a built-in loop. **Fig 23** shows an alternative circuit for other bands and for plugging in a separate loop connected by a low-impedance transmission line. Select coil and capacitor combinations that will tune to the desired frequencies. Plug-in coils could be used. It is a good idea to have the RF end of the unit fairly well shielded, to eliminate signal pickup except through the loop. This little unit should certainly help you on those dark nights in the country. (Tip to the hidden-transmitter operator—if you want to foul up some of your pals using these loops, just hide near the antenna of a 50-kW broadcast transmitter!)

A LOOPSTICK FOR 3.5 MHz

Figs 24 through 26 show an RDF loop suitable for the 3.5-MHz band. It uses a construction technique that has had considerable application in low-frequency marine direction finders. The loop is a coil wound on a ferrite rod from a broadcast-antenna loopstick. The loop was designed by John Isaacs, W6PZV, and described in *QST* for June 1958. Because you can make a coil with high Q using a ferrite core, the sensitivity of such a loop is comparable to a conventional loop that is a foot or so in diameter. The output of the vertical-rod sensing antenna, when properly combined with that of the loop, gives the system the cardioid pattern shown in Fig 1D.

To make the loop, remove the original winding on the ferrite core and wind a new coil, as shown in **Fig 25**. Other types of cores than the one specified may be substituted; use the largest coil available and adjust the winding so that the circuit resonates in the 3.5-MHz band within the range of C1. The tuning range of the loop may be checked with a dip meter.

The sensing system consists of a 15-inch whip and an adjustable inductor that resonates the whip as a quarter-wave antenna. It also contains a potentiometer to control the output of the antenna. S1 is used to switch the sensing antenna in and out of the circuit.

The whip, the loopstick, the inductance L1, the capacitor C1, the potentiometer R1, and the switch S1 are all mounted on a 4 × 5 × 3-inch box chassis, as shown in **Fig 26**. The loopstick may be mounted and protected inside a piece of 1/2-inch PVC pipe. A section of 1/2-inch electrical conduit is attached to the bottom of the chassis box and this supports the instrument.

To produce an output having only one null there must a 90° phase difference between the outputs of the loop and sensing antennas, and the signal strength from each must be the same. The phase shift is obtained by tuning the sensing antenna slightly off frequency, using the slug in L1. Since the sensitivity of the whip antenna is greater than that of the loop, its output is reduced by adjusting R1.

Adjustment

To adjust the system, enlist the aid of a friend with a mobile transmitter and find a clear spot where the transmitter and RDF receiver can be separated by several hundred feet. Use as little power as possible at the transmitter. (Make very sure you don't transmit into the loop if you are using a transceiver as a detector.) With the test transmitter operating on the proper frequency, disconnect the sensing antenna with S1, and peak the loopstick using C1, while watching the S meter on the transceiver. Once the loopstick is peaked, no further adjustment of C1 will be necessary. Next, connect the sensing antenna and turn R1 to minimum resistance. Then vary the adjustable slug of L1 until a maximum reading of the S meter is again obtained. It may be necessary to turn the unit a bit during this adjustment to obtain a higher reading than with the loopstick alone. The last turn of the slug is quite critical, and some hand-capacitance effect may be noted.

Now turn the instrument so that one side (not an end) of the loopstick is pointed toward the test transmitter. Turn R1 a complete revolution and if the proper side was chosen a definite null should be observed on the S meter for one particular position of R1. If not, turn the RDF 180° and try again. This time leave R1 at the setting producing

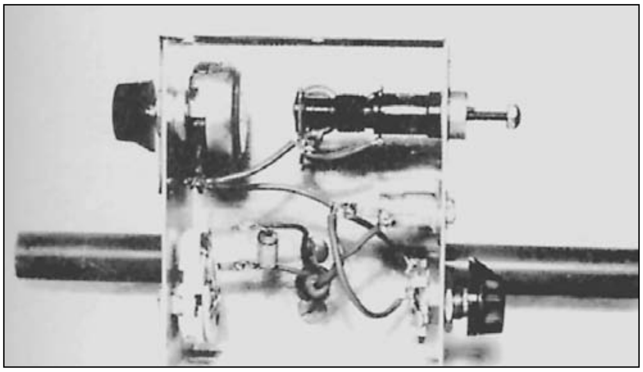

Fig 26—Components of the 3.5-MHz RDF are mounted on the top and sides of a channel-lock type box. In this view R1 is on the left wall at the upper left and C1 is at the lower left. L1, S1 and the output connector are on the right wall. The loopstick and whip mount on the outside.

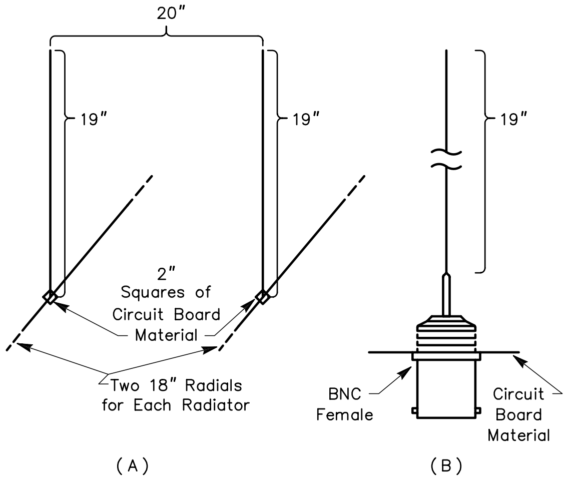

Fig 27—At A is a simple configuration that can produce a cardioid pattern. At B is a convenient way of fabricating a sturdy mount for the radiator using BNC connectors.

the minimum reading. Now adjust L1 very slowly until the S-meter reading is reduced still further. Repeat this several times, first R1, and then L1, until the best minimum is obtained.

Finally, as a check, have the test transmitter move around the RDF and follow it by turning the RDF. If the tuning has been done properly the null will always be broadside to the loopstick. Make a note of the proper side of the RDF for the null, and the job is finished.

A 144-MHZ CARIOID-PATTERN RDF ANTENNA

Although there may be any number of different VHF antennas that can produce a cardioid pattern, a simple design is depicted in **Fig 27**. Two $^{1}/_{4}$-wavelength vertical elements are spaced one $^{1}/_{4}$-λ apart and are fed 90° out of phase. Each radiator is shown with two radials approximately 5% shorter than the radiators. This array was designed by Pete O'Dell, KB1N, and described in *QST* for March 1981.

Computer modeling showed that slight alterations in the size, spacing and phasing of the elements strongly impact the pattern. The results suggest that this system is a little touchy and that the most significant change comes at the null. Very slight alterations in the dimensions caused the notch to become much more shallow and, hence, less usable for RDF. Early experience in building a working model bore this out.

This means that if you build this antenna, you will find it advantageous to spend a few minutes to tune it carefully for the deepest null. If it is built using the techniques presented here, then this should prove to be a small task, well worth the extra effort. Tuning is accomplished by adjusting the length of the vertical radiators, the spacing between them and, if necessary, the lengths of the phasing harness that connects them. Tune for the deepest null on your S meter when using a signal source such as a moderately strong repeater.

This should be done outside, away from buildings and large metal objects. Initial indoor tuning on this project was tried in the kitchen, which revealed that reflections off the appliances were producing spurious readings. Beware too of distant water towers, radio towers, and large office or apartment buildings. They can reflect the signal and give false indications.

Construction is simple and straightforward. Fig 27B shows a female BNC connector (RadioShack 278-105) that has been mounted on a small piece of PC-board material. The BNC connector is held upside down, and the vertical radiator is soldered to the center solder lug. A 12-inch piece of brass tubing provides a snug fit over the solder lug. A second piece of tubing, slightly smaller in diameter, is telescoped inside the first. The outer tubing is crimped slightly at the top after the inner tubing is installed. This provides positive contact between the two tubes. For 146 MHz the length of the radiators is calculated to be about 19 inches. You should be able to find small brass tubing at a hobby store. If none is available in your area, consider brazing rods. These are often available in hardware sections of discount stores. It will probably be necessary to solder a short piece to the top since these come in 18-inch sections. Also, tuning will not be quite as convenient. Two 18-inch radials are added to each element by soldering them to the board. Two 36-inch pieces of heavy brazing rod were used in this project.

The Phasing Harness

As shown in **Fig 28**, a T connector is used with two different lengths of coaxial line to form the phasing harness. This method of feeding the antenna is superior over

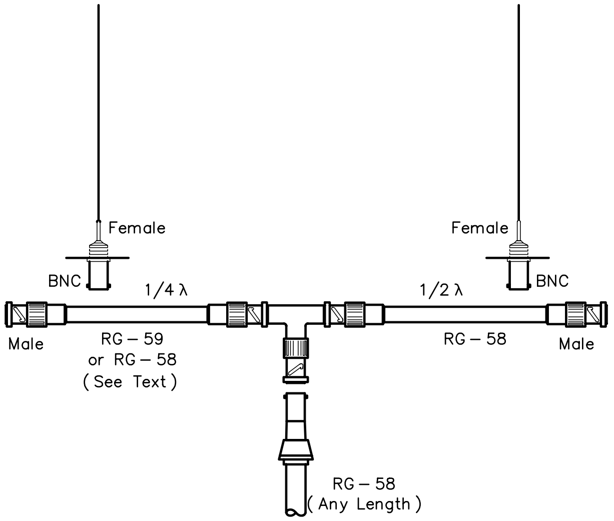

Fig 28—The phasing harness for the phased 144-MHz RDF array. The phasing sections must be measured from the center of the T connector to the point that the vertical radiator emerges from the shielded portion of the upside-down BNC female. Don't forget to take the length of the connectors into account when constructing the harness. If care is taken and coax with solid polyethylene dielectric is used, you should not have to prune the phasing line. With this phasing system, the null will be in a direction that runs along the boom, on the side of the 1/4-wavelength section.

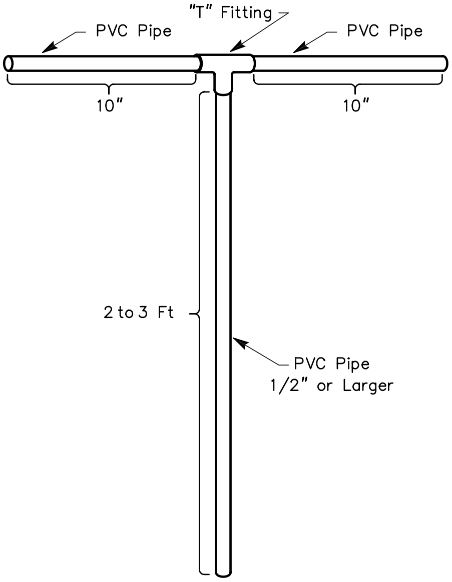

Fig 29—A simple mechanical support for the DF antenna, made of PVC pipe and fittings.

other simple systems to obtain equal currents in the two radiators. Unequal currents tend to reduce the depth of the null in the pattern, all other factors being equal.

The 1/2-wavelength section can be made from either RG-58 or RG-59, because it should act as a 1:1 transformer. With no radials or with two radials perpendicular to the vertical element, it was found that a 1/4-wavelength section made of RG-59 75-Ω coax produced a deeper notch than a 1/4-wavelength section made of RG-58 50-Ω line. However, with the two radials bent downward somewhat, the RG-58 section seemed to outperform the RG-59. Because of minor differences in assembly techniques from one antenna to another, it will probably be worth your time and effort to try both types of coax and determine what works best for your antenna. You may also want to try bending the radials down at slightly different angles for the best null performance.

The most important thing about the coax for the harness is that it be of the highest quality (well-shielded and with a polyethylene dielectric). The reason for avoiding foam dielectric is that the velocity factor can vary from one roll to the next—some say that it varies from one foot to the next. Of course, it can be used if you have test equipment available that will allow you to determine its electrical length. Assuming that you do not want to or cannot go to that trouble, stay with coax having a solid polyethylene dielectric. Avoid coax that is designed for the CB market or do-it-yourself cable-TV market. (A good choice is Belden 8240 for the RG-58 or Belden 8241 for the RG-59.)

Both RG-58 and RG-59 with polyethylene dielectric have a velocity factor of 0.66. Therefore, for 146 MHz a quarter wavelength of transmission line will be 20.2 inches × 0.66 = 13.3 inches. A half-wavelength section will be twice this length or 26.7 inches. One thing you must take into account is that the transmission line is the total length of the cable *and the connectors*. Depending on the type of construction and the type of connectors that you choose, the actual length of the coax by itself will vary somewhat. You will have to determine that for yourself.

Y connectors that mate with RCA phono plugs are widely available and the phono plugs are easy to work with. Avoid the temptation, however, to substitute these for the T and BNC connectors. Phono plugs and a Y connector were tried. The results with that system were not satisfactory. The performance seemed to change from day to day and the notch was never as deep as it should have been. Although they are more difficult to find, BNC T connectors will provide superior performance and are well worth the extra cost. If you must make substitutions, it would be preferable to use UHF connectors (type PL-259).

Fig 29 shows a simple support for the antenna. PVC tubing is used throughout. Additionally, you will need a T fitting, two end caps, and possibly some cement. (By not cementing the PVC fittings together, you will have the option of disassembly for transportation.) Cut the PVC for the dimensions shown, using a saw or a tubing cutter.

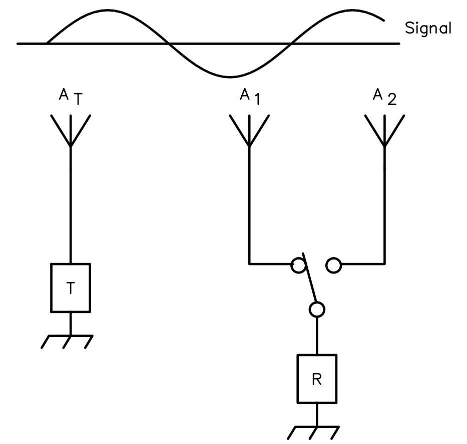

Fig 30—At the left, A_T represents the antenna of the hidden transmitter, T. At the right, rapid switching between antennas A_1 and A_2 at the receiver samples the phase at each antenna, creating a pseudo-Doppler effect. An FM detector detects this as phase modulation.

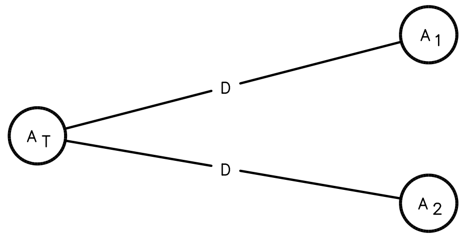

Fig 31—If both receiving antennas are an equal distance (D) from the transmitting antenna, there will be no difference in the phase angles of the signals in the receiving antennas. Therefore, the detector will not detect any phase modulation, and the audio tone will disappear from the output of the detector.

A tubing cutter is preferred because it produces smooth, straight edges without making a mess. Drill a small hole through the PC board near the female BNC of each element assembly. Measure the 20-inch distance horizontally along the boom and mark the two end points. Drill a small hole vertically through the boom at each mark. Use a small nut and bolt to attach each element assembly to the boom.

Tuning

The dimensions given throughout this section are those for approximately 146 MHz. If the signal you will be hunting is above that frequency, then the measurements should be a bit shorter. If you wish to operate below that frequency, then they will need to be somewhat longer. Once you have built the antenna to the rough size, the fun begins. You will need a signal source near the frequency that you will be using for your RDF work. Adjust the length of the radiators and the spacing between them for the deepest null on your S meter. Make changes in increments of 1/4 inch or less. If you must adjust the phasing line, make sure that the 1/4-wavelength section is exactly one-half the length of the half-wavelength section. Keep tuning until you have a satisfactorily deep null on your S meter.

THE DOUBLE-DUCKY DIRECTION FINDER

For direction finding, most amateurs use antennas having pronounced directional effects, either a null or a peak in signal strength. FM receivers are designed to eliminate the effects of amplitude variations, and so they are difficult to use for direction finding without looking at an S meter. Most modern HT transceivers do not have S meters.

This classic "Double-Ducky" direction finder (DDDF) was designed by David Geiser, WA2ANU, and was described in *QST* for July 1981. It works on the principle of switching between two nondirectional antennas, as shown in **Fig 30**. This creates phase modulation on the incoming signal that is heard easily on the FM receiver. When the two antennas are exactly the same distance (phase) from the transmitter, as in **Fig 31**, the tone disappears. (This technique is also known in the RDF literature as *Time-Difference-of-Arrival*, or TDOA, since signals arrive at each antenna at slightly different times, and hence at slightly different phases, from any direction except on a line perpendicular to and halfway in-between the two antennas. Another general term for this kind of two-antenna RDF technique is *interferometer*. *—Ed.*)

In theory the antennas may be very close to each other, but in practice the amount of phase modulation increases directly with the spacing, up to spacings of a half wavelength. While a half-wavelength separation on 2 meters (40 inches) is pretty large for a mobile array, a quarter wavelength gives entirely satisfactory results, and even an eighth wavelength (10 inches) is acceptable.

Think in terms of two antenna elements with fixed spacing. Mount them on a ground plane and rotate that ground plane. The ground plane held above the hiker's head or car roof reduces the needed height of the array and the directional-distorting effects of the searcher's body or other conducting objects.

The DDDF is bidirectional and, as described, its tone null points both toward and away from the signal origin. An L-shaped search path would be needed to resolve the ambiguity. Use the techniques of triangulation described earlier in this chapter.

Specific Design

It is not possible to find a long-life mechanical switch operable at a fairly high audio rate, such as 1000 Hz. Yet we want an audible tone, and the 400- to 1000-Hz range is perhaps most suitable considering audio amplifiers and average hearing. Also, if we wish to use the transmit function of a transceiver, we need a switch that will carry per-

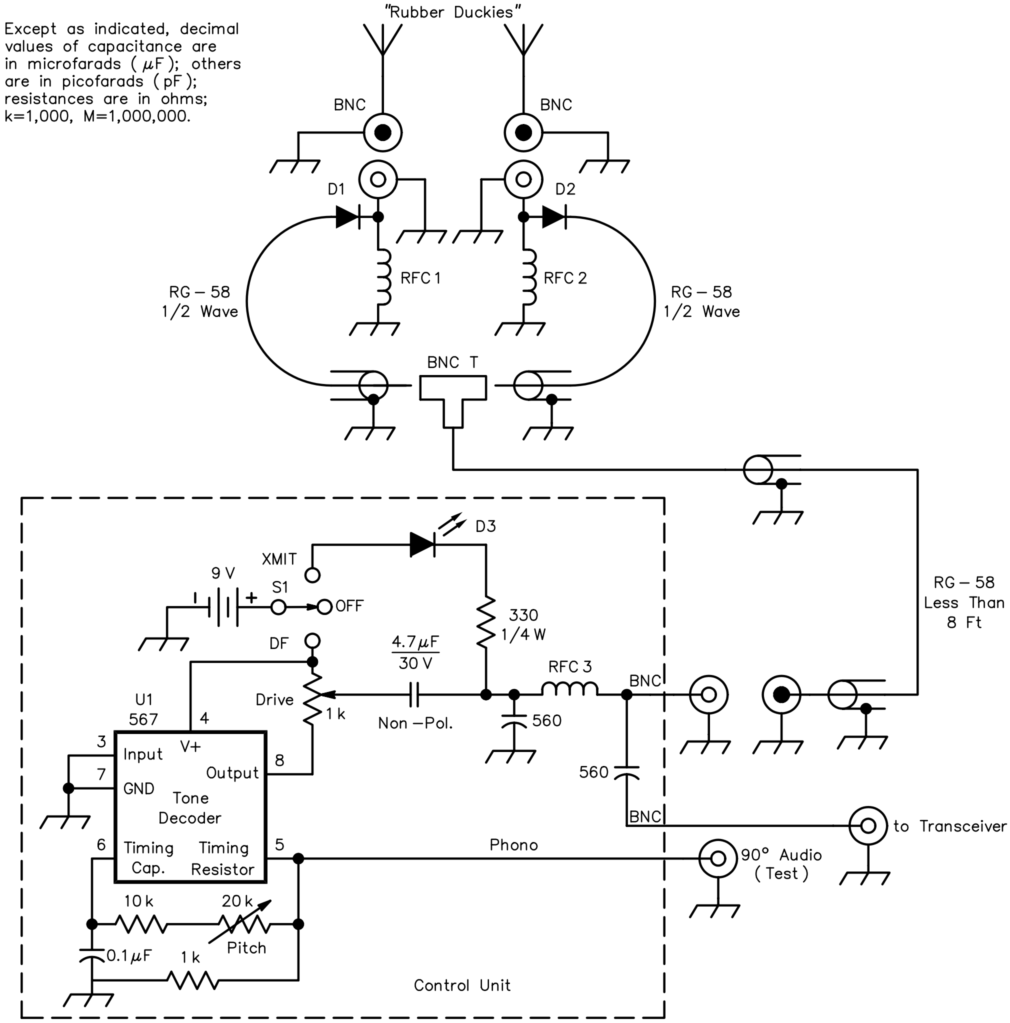
Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.
"Rubber Duckies"
BNC
BNC
D1
D2
RFC 1
RFC 2
RG – 58
1/2 Wave
RG – 58
1/2 Wave
BNC T
D3
XMIT
9 V
S1
OFF
DF
4.7μF
30 V
Non – Pol.
330
1/4 W
RFC 3
BNC
560
560
BNC
RG – 58
Less Than
8 Ft
U1
567
4
Drive
1 k
3
Input
V+
8
Output
7
GND
Tone
Decoder
6
Timing
Cap.
Timing
Resistor
5
Phono
to Transceiver
90° Audio
(Test)
10 k
20 k
Pitch
0.1μF
1 k
Control Unit

Fig 32—Schematic diagram of the DDDF circuit. Construction and layout are not critical. Components inside the broken lines should be housed inside a shielded enclosure. Most of the components are available from RadioShack, except D1, D2, the antennas and RFC1-RFC3. These components are discussed in the text. S1—See text.

haps 10 W without much problem.

A solid-state switch, the PIN diode is used. The intrinsic region of this type of diode is ordinarily bare of current carriers and, with a bit of reverse bias, looks like a low-capacitance open space. A bit of forward bias (20 to 50 mA) will load the intrinsic region with current carriers that are happy to dance back and forth at a 148-MHz rate, looking like a resistance of an ohm or so. In a 10-W circuit, the diodes do not dissipate enough power to damage them.

Because only two antennas are used, the obvious approach is to connect one diode *forward* to one antenna, to connect the other *reverse* to the second antenna and to drive the pair with square-wave audio-frequency ac. **Fig 32** shows the necessary circuitry. RF chokes (Ohmite Z144, J. W. Miller RFC-144 or similar vhf units) are used to let the audio through to bias the diodes while blocking RF. Of course, the reverse bias on one diode is only equal to the forward bias on the other, but in practice this seems sufficient.

A number of PIN diodes were tried in the particular setup built. These were the Hewlett-Packard HP5082-3077, the Alpha LE-5407-4, the KSW KS-3542 and the Microwave Associates M/A-COM 47120. All worked well, but the HP diodes were used because they provided a slightly lower SWR (about 3:1).

A type 567 IC is used as the square-wave generator. The output does have a dc bias that is removed with a nonpolarized coupling capacitor. This minor inconvenience is more than rewarded by the ability of the IC to work well with between 7 and 15 volts (a nominal 9-V minimum is recommended).

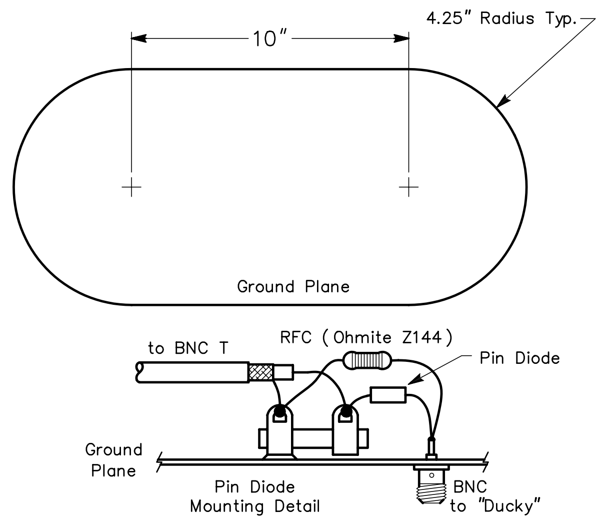

Fig 33—Ground-plane layout and detail of parts at the antenna connectors.

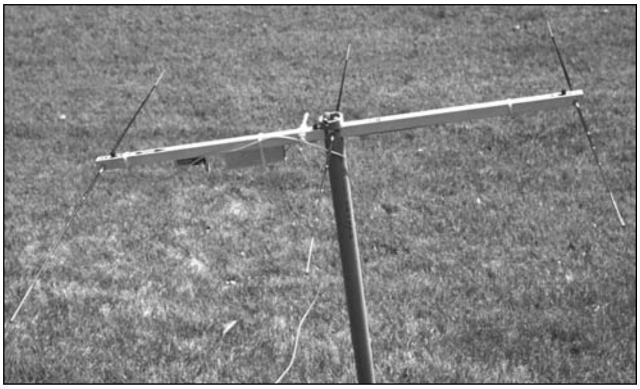

Fig 34—Photo of Fox-Hunting DF Twin 'Tenna set up as a horizontally polarized, 3-element Yagi.

The nonpolarized capacitor is also used for dc blocking when the function switch is set to XMIT. D3, a light-emitting diode (LED), is wired in series with the transmit bias to indicate selection of the XMIT mode. In that mode there is a high battery current drain (20 mA or so). S1 should be a center-off locking type toggle switch. An ordinary center-off switch may be used, but beware. If the switch is left on XMIT you will soon have dead batteries.

Cables going from the antenna to the coaxial T connector were cut to an electrical $^{1}/_{2}$ wavelength to help the open circuit, represented by the reverse-biased diode, look open at the coaxial T. (The length of the line within the T was included in the calculation.)

The length of the line from the T to the control unit is not particularly critical. If possible, keep the total of the cable length from the T to the control unit to the transceiver under 8 feet, because the capacitance of the cable does shunt the square-wave generator output.

Ground-plane dimensions are not critical. See **Fig 33**. Slightly better results may be obtained with a larger ground plane than shown. Increasing the spacing between the pickup antennas will give the greatest improvement. Every doubling (up to a half wavelength maximum) will cut the width of the null in half. A 1° wide null can be obtained with 20-inch spacing.

DDDF Operation

Switch the control unit to DF and advance the drive potentiometer until a tone is heard on the desired signal. Do not advance the drive high enough to distort or "hash up" the voice. Rotate the antenna for a null in the fundamental tone. Note that a tone an octave higher may appear.

If the incoming signal is quite out of the receiver linear region (10 kHz or so off frequency), the off-null antenna aim may present a fairly symmetrical AF output to one side, **Fig 35A**. It may also show instability at a sharp null position, indicated by the broken line on the display in Fig 35B. Aimed to the other side of a null, it will give a greatly increased AF output, Fig 35C. This is caused by the different parts of the receiver FM detector curve used. The sudden tone change is the tip-off that the antenna null position is being passed.

The user should practice with the DDDF to become acquainted with how it behaves under known situations of signal direction, power and frequency. Even in difficult nulling situations where a lot of second-harmonic AF exists, rotating the antenna through the null position causes a very distinctive tone change. With the same frequencies and amplitudes present, the quality of the tone (timbre) changes. It is as if a note were first played by a violin, and then the same note played by a trumpet. (A good part of this is the change of phase of the fundamental and odd harmonics with respect to the even harmonics.) The listener can recognize differences (passing through the null) that would give an electronic analyzer indigestion.

A FOX-HUNTING DF TWIN 'TENNA

Interferometers give sharp bearings, but they lack sensitivity for distant work. Yagis are sensitive, but they provide relatively broad bearings. This project yields an antenna that blends both on a single boom to cover both ends of the hunt. This is a condensation of a October 1998 *QST* article by R. F. Gillette, W9PE.

A good fox-hunting antenna must meet a number of criteria: (1) small size, (2) gain to detect weak signals and (3) high directivity to pinpoint the fox. Small antennas, however, do not normally yield both gain and direc-

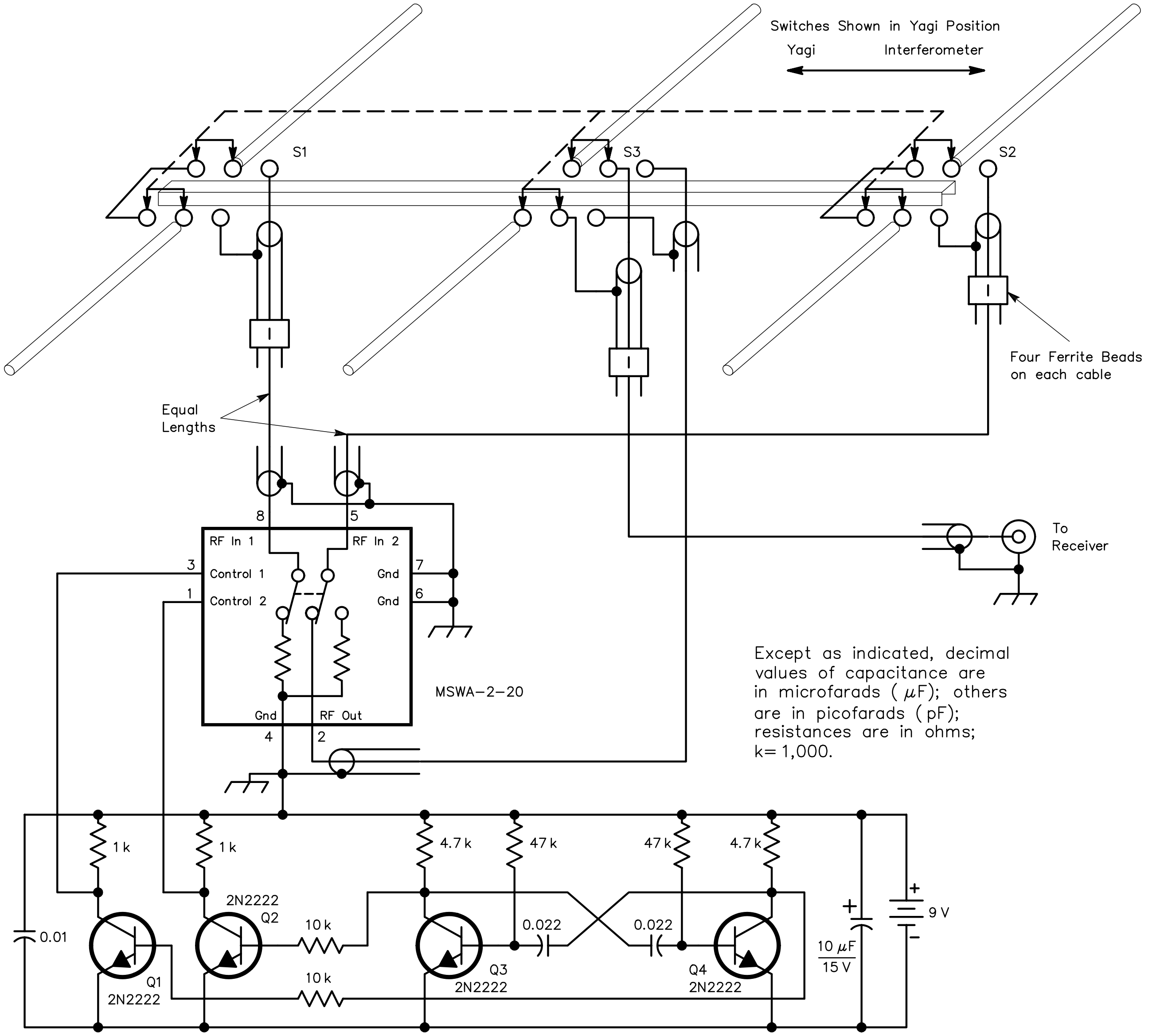
Switches Shown in Yagi Position
Yagi
Interferometer
S1
S3
S2
Four Ferrite Beads on each cable
Equal Lengths
8
5
RF In 1
RF In 2
3
Control 1
1
Control 2
Gnd
7
Gnd
6
Gnd
RF Out
4
2
MSWA–2–20
To Receiver
Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k= 1,000.
1 k
1 k
4.7 k
47 k
47 k
4.7 k
0.01
2N2222
Q2
10 k
10 k
Q1
2N2222
0.022
0.022
Q3
2N2222
Q4
2N2222
10 μF
15 V
9 V
+
+
–

Fig 35—Schematic of the Yagi/interferometer antenna system.

Table 1
Yagi Design

Item	*Overall Length* *(Inches)*	*Boom to Element Tip* *(Inches)*
Director length	34.75	17.00
Director to Driven El. spacing	15.75	16.00
Driven El. length	37.75	18.50
Driven El. to Reflector spacing	15.75	16.00
Reflector length	40.75	20.00

*SWR less than 1.3:1 from 144.5 to 148 MHz

tivity. By combining two antennas, all three requirements are satisfied in a way that makes a nice build-it-yourself project.

This antenna uses slide switches to configure it as either a Yagi or a single-channel interferometer. When used as an interferometer, a GaAs RF microcircuit switches the FM receiver between two matched dipoles at an audio frequency. To make the antenna compact W9PE used hinged, telescopic whips as the elements; they collapse and fold parallel to the boom for storage.

The Yagi

The Yagi is a standard three-element design, based on 0.2-λ spacing between the director, the driven element and the reflector. It yields about 7 dBi gain and a front-to-back ratio of over 15 dB. Because a slide switch is used at the center of each element and the elements have small diameters, their resonant lengths are different from typical ones. **Table 1** shows the sizes used and **Fig 34** shows the Yagi.

To make sure that radiation from the coax does not affect the pattern, the author used some ferrite beads as coaxial choke-baluns. This also prevents objects near the coax from affecting signal-strength readings. The Yagi also has a low SWR; with uncalibrated equipment, he measured less than 1.3:1 over most of the 2-meter band.

This Yagi has a lot more gain than a “rubber ducky,” but we need more directivity for fox hunting. That’s where the interferometer comes in.

An Interferometer

To form the interferometer, the two end elements are converted to dipoles and the center element is disabled. When the three switches in **Fig 35** are thrown to the right, the feed line to the receiver is switched from the center element to the RF switch output, and the end elements are connected via feed lines to the RF switch inputs. With the Yagi’s feed point open and the driven element equidistant from both interferometer antennas, the center element should have no effect on the interferometer. Nonetheless, it’s easier to collapse the driven-element whips and get them out of the way than to worry about spacing.

Now if both interferometer coax cables are of equal length (between the antennas and switch) and the two antennas are the same distance from the transmitter (broadside to it), the signals from both antennas will be in phase. Switching from one antenna to the other will have no effect on the received signal. If one antenna is a little closer to the transmitter than the other, however, there will be a phase shift when we switch antennas.

When the antenna switch is at an audio rate, say 700 Hz, the repeated phase shifts result in a set of 700 Hz sidebands. At this point, all that was needed was a circuit to switch from one antenna to the other at an audio rate. W9PE chose a low-cost Mini-Circuits MSWA-2-20 GaAs RF switch driven by a simple multivibrator and buffer. The GaAs switch is rated to 2.0 GHz, hence this switching concept can easily be scaled to other ham bands. The PC board should work through the 440 MHz ham band. The author suggests adding a ground plane under the RF portion of the PC board and testing it before using it at a higher frequency.

The RF switch is controlled by a set of equal-amplitude, opposite-phase square waves: 0 V at one control port and –8 to –12 V at the other. (Mini Circuits is unclear about maximum voltages for this device. For safety, don’t power it with more than 9 V.—*Ed.*) The opposite controls the other switch position. A 9-V battery was used as the power supply, with the positive terminal grounded. This results in a 0 V control signal to the RF switch when the buffer transistor is off and a Vsat (about 0.2 V less than the –9 V battery: –8.8 V) signal when the buffer transistor is saturated. The multi-vibrator has two outputs, and each drives a buffer resulting in the required equal-and-opposite-phase drive signals.

Table 2
Bill of Materials

Quantity	*Item*
3 ft	$^3/_4$-inch aluminum U channel
6 sets	insulated shoulder washers for elements
1	9 V battery
1	9 V battery connector
1	10 µF, 16 V electrolytic capacitor
1	0.01 µF, 25 V capacitor
2	0.022 µF, 25 V capacitor
2	1 kΩ $^1/_8$ W resistor
2	4.7 kΩ $^1/_8$ W resistor
2	10 kΩ $^1/_8$ W resistor
2	47 kΩ $^1/_8$ W resistor
2	1 kΩ $^1/_8$ W resistor
4	2N2222 transistors
1	Mini-Circuits MSWA-2-20 (Mini-Circuits Labs, 13 Neptune Ave, Brooklyn, NY 11235; tel 718-934-4500, 417-335-5935, fax 718-332-4661; e-mail **sales@minicircuits.com**; URL **www.minicircuits.com**)
3	DPDT slide switch (1$^1/_8$-inch, 29 mm, mounting centers), Stackpole, 3 A, 125 V used
10 ft	50 Ω coax (0.140-inch maximum OD)
1	coaxial connector (receiver dependent)
1	lot, mounting hardware
1	lot, heat-shrink tubing or equal
4	cable ties
1	2 × 3.5-inch single-sided fiberglass PC board
1	1-inch PVC conduit
12	#FB-20 ferrite beads 0.14-inch ID, 0.5-inch long (All Electronics Corp: PO Box 567, Van Nuys, CA 91408-0567; tel 888-826-5432, fax 818-781-2653, e-mail: **allcorp@allcorp.com**; URL **www.allcorp.com/**.)
6	20$^1/_2$-inch telescoping antenna elements (Nebraska Surplus, tel 402-346-4750; e-mail **grinnell@probe.net**)
1	Special resist film (Techniks Inc, PO Box 463, Ringoes, NJ 08551; tel 908-788-8249, fax 908-788-8837; e-mail **techniks@idt.net**; URL **www.techniks.com/**)

Circuit Construction

After he selected the Mini-Circuits RF switch, W9PE realized that its small size would be best handled with a simple PC board. He made the prototype boards with a photocopy transparency technique.

A power on-off switch was not used, as the 9-V bat-

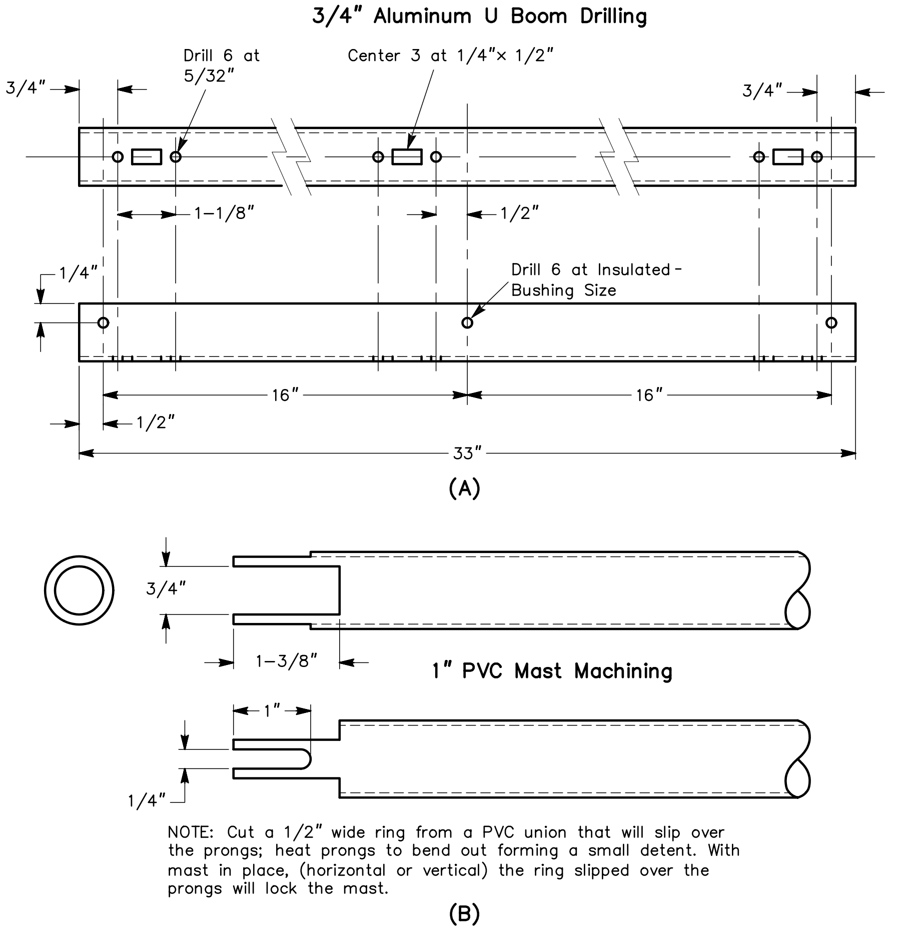
3/4" Aluminum U Boom Drilling
Drill 6 at 5/32"
Center 3 at 1/4"× 1/2"
3/4"
3/4"
1-1/8"
1/2"
1/4"
Drill 6 at Insulated-Bushing Size
16"
16"
1/2"
33"
(A)
3/4"
1-3/8"
1" PVC Mast Machining
1"
1/4"
NOTE: Cut a 1/2" wide ring from a PVC union that will slip over the prongs; heat prongs to bend out forming a small detent. With mast in place, (horizontal or vertical) the ring slipped over the prongs will lock the mast.
(B)

Fig 36—Boom-drilling and mast-machining details.

tery connector serves the same function. The battery fits tightly in the $^3/_4$-inch U channel. W9PE covered the circuit board with a plastic-lined aluminum cover, but plastic film and some aluminum foil, provide the same function. A cable tie will strap either into the U channel.

Table 2 is a complete bill of materials. You can use any telescoping elements, providing that they extend to over 20 inches and have a mounting stud long enough to accommodate the insulated washers. As an alternate to the stud, they can have ends tapped to receive a screw for the insulated mounting. The author picked up his elements at a hamfest from the vendor listed; they are also available from most electronic parts houses. The Mini-Circuit RF switch is a currently available part.

Antenna Construction

Fig 35 shows the antenna schematic. It shows all three switches in the Yagi position; each would slide to the right for interferometer use. Slide switches work pretty well at 2 meters. Each of the elements is mounted to the boom with insulating washers, and a strip of copper stock connects each element to its slide switch. (You can substitute copper braid, solder wick, coax shield or any short, low resistance, low inductance conductor for the copper stock.) This switching arrangement allows you to switch the reflector and director from being parasitic elements (electrically continuous) to being dipoles (center fed).

Because the elements telescope, you can adjust the interferometer dipoles to exactly equal lengths each time you switch the antenna configuration from Yagi to interferometer. Again, choke baluns block RF on the outside of each element's coax.

Caution: Do not transmit when the RF switch is selected. Transmit only when in the Yagi configuration. RF power will destroy the Mini-Circuits RF switch. To be safe, lock out your transmit function. Most HTs have this capability. When using a mobile radio, disconnect the microphone. It is, however, safe to transmit in the Yagi configuration—which is nice for portable operating.

Fig 36 gives dimensions for drilling a standard $^3/_4$-inch aluminum U channel for the boom and shows how the author cut a 1-inch PVC pipe (plastic conduit) for a mast and a mast locking ring. If PVC conduit is not available in your area, PVC water pipe (and a PVC union for the locking ring) will work. This mast allows mounting the antenna for either vertical or horizontal polarization.

Be sure to test the plastic pipe you use for low RF loss. Do this by heating a sample in a microwave oven. Place a pipe sample and a glass of water in the oven. (The sample is not placed in the glass of water; the water keeps the microwave from operating without a load.) Bring the water to a boil, and then carefully check the sample's temperature. If the sample is not hot, its RF loss is low, and the plastic can be used.

Using the Antenna

When starting a hunt, set up the Yagi antenna by placing all switches in the Yagi position. Swing all of the telescoping elements perpendicular to the boom and set the whip lengths to achieve the proper element lengths, while keeping each element symmetrical about the boom. The cables or boom can be marked with the length data.

While the signal is weak, use the Yagi. It has 7 dBi gain, but its bearing resolution is only about 20°. When the signal gets stronger, use the interferometer. It has less gain, but its bearing resolution is better than 1°. If the fox transmitter begins overloading your receiver, collapse the whips (equally) to reduce the gain and continue triangulating. Near the transmitter, you should triangulate both horizontally (azimuth) and vertically (elevation). The antenna works both ways, and the transmitter may be located above or below you.

Antenna Alternative

As an alternative to the telescoping elements, George Holada, K9GLJ, suggested using fixed-length elements with banana plugs matched to banana jacks on the boom. Three pairs would be used for the Yagi, an extra driven-element pair for the interferometer mode and two short-element pairs to reduce the received signal level if an overload condition occurs. He also suggested a PVC boom allowing the elements to be stored inside the boom.

THE FOUR-WAY MOBILE DF SYSTEM

This innovative, yet simple, RDF antenna system was described in an article by Malcolm C. Mallette, WA9BVS, in November 1995 *QST*. It is derived from the TDOA design shown earlier in this chapter by David T. Geiser, WA2ANU, and by a design by Paul Bohrer, W9DUU. (See Bibliography.)

Direction-finding often involves two different activities: DFing on foot and DFing from a vehicle. Often, you must track the signal using a vehicle, then finish the hunt on foot. Whether on foot or in a vehicle, the primary problem you'll encounter when trying to locate the transmitter is multipath reception. Multipath reception involves receiving the same signal by more than one path, one signal from the true direction of the transmitter and others by reflected paths that may come from widely different directions. VHF and UHF signals bounce off almost any object and hide the true source of a signal. For example, if there's a large metal building north of you, a signal from the south may arrive from the north because the signal bounces off the building and back to you.

Multipath reception effects can be defeated by taking a number of readings from different positions and arriving at an average direction. While moving at road speeds in a vehicle, it's possible to take a number of readings from different positions and average them, and it's also possible to average a number of readings over a distance of travel by electronic means. The true bearing to the transmitter can usually be found by either method.

DFing equipment for use on foot is simpler than systems for use on a vehicle. While afoot, you can turn at will or easily rotate an antenna. Turning a vehicle while going down a street may result in a fender-bender if you're not careful!

The simplest DFing system to use while on foot consists of an S-meter-equipped hand-held receiver and a small, hand-held Yagi with an attenuator in line between the antenna and receiver. The attenuator keeps the S meter from pinning. The direction in which the beam points when the strongest signal is received is the direction of the transmitter. Of course, you'll want to take readings at several locations at least a wavelength apart to obtain an average heading, as multipath reception can still cause false readings in some locations.

Another approach that many hams have taken is the simple WA2ANU DDFF—it is now commonly known simply as the "buzzbox." Various commercial versions of the hand-held buzzbox system are available. Some sys-

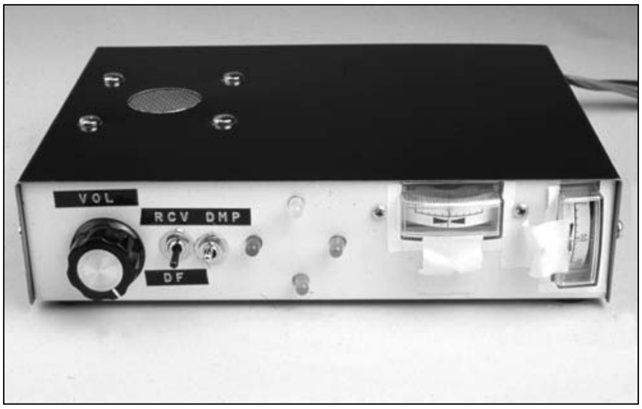

Fig 37—Front panel of the Four-Way DFer. At the extreme left of the front panel is the VOLUME control. Immediately to the right is the RCV/OFF/DF center-off toggle switch, with the damping (DMP) control switch nearby. Four LEDs mounted in a diamond pattern indicate signal direction: front (yellow), right (green), back (orange) and left (red). The horizontally mounted zero-center meter indicates left/right signal reception, the vertically mounted meter displays front/back signal reception. A small speaker is mounted on the top cover.

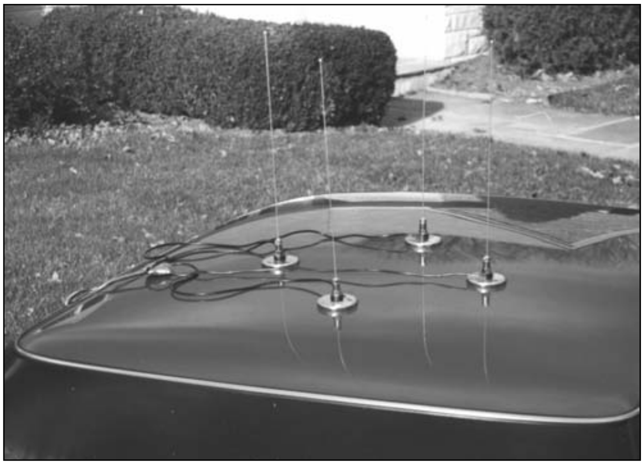

Fig 38—Placement of the four antennas on the author's car roof. The small object to the left of the antennas is the switch board.

tems have been upgraded to indicate whether the signal is arriving from the left or right of your position. The main drawback, however, to the buzzbox is that it's not as sensitive as a simple dipole and not nearly as sensitive as a beam.

In theory, you could take a buzzbox or Yagi/attenuator system in a car, stop periodically, get out and check the direction to the transmitter, then climb back in and drive off. Although this procedure works, it isn't very practical—it takes a long time to find the transmitter.

This design is for a left-right, front-back box (LRFB box) that indicates whether the received signal is to the left or right and whether it is to the front or back of the receiver. The location display consists of four LEDs arranged in a diamond pattern (see the title-page photo). When the top LED is on, the signal is coming from the front. When the top and right LEDs are on, the transmitter is between the front and the right. When only the right LED is on, the signal is directly to the right. When the bottom LED and right LED are on, the transmitter is to the right and to the back. The same pattern occurs around the clock. Therefore, four LEDs indicate eight directions. As most highways and streets have intersections that force a driver to choose moving straight ahead, right or left, the indication is sufficiently precise for practical transmitter hunting.

If the four-LED display is used alone, all parts can be obtained from your local Radio Shack store. Two zero-center 50-µA meters (50-0-50)—M1 and M2—can be used in addition to, or in place of, the LEDs, but RadioShack does not stock such meters. **Fig 37** shows the front panel layout of the LRFB.

The LRFB box uses four mag-mount $^1/_4$-λ antennas placed on the vehicle roof as shown in **Fig 38**. The whips in the mag mounts can be changed to $^1/_4$-λ 440-MHz whips and the antennas placed closer together when switching from 144-MHz to 440-MHz operation.

Circuit Description

See **Fig 39** in the following discussion (pages 26 and 27). U2, a 555 timer, generates a string of square-wave pulses at pin 3. The pulse frequency is determined by the setting of R4. The pulses are fed to the clock input (pin 14) of U3, a 4017 decade counter. On the first pulse, a positive voltage appears at U3, pin 3. On receipt of the second pulse from U2, pin 3 of U3 goes to ground and a positive voltage appears on pin 2. This sequence continues on successive pulses from U2 as pins 3, 2, 4, 7, 10, 1, 5, 6, 9 and 11 go positive in succession.

D1 through D5, and D6 through D10, OR the pulses. The result is that TP3 goes positive on the first pulse from U2, while TP2 is at 0 V. The next pulse of U2 results in TP2 going positive and TP3 going to 0 V. This sequence repeats as the counter goes around to make pin 3 positive again, and recycles.

Antenna Switching

U2 and U3 produce alternating pulses at TP2 and TP3. If we wanted only to alternately turn on and off two antennas, we could use the pulses at TP2 and TP3. The design ensures that the pulses at TP2 are the same length as the pulses at TP3. For the LRFB box, however, we need to switch between the left-right antennas many times, then switch between the front-back antennas many times.

Pin 12 of U3, CARRY OUT, emits a pulse every time U3 counts through its cycle of 10 pulses. The carry pulses from U3 go to U4, pin 14, the clock input of that 4017 counter. As U4 cycles, its output pins pulse; those pulses are, in effect, directed by D11 through D20.

As a result, TP4 is positive during the first 50 pulses from U2 and TP5 is positive during the second 50 pulses of U2. Q1 through Q6 form a quad AND gate. They AND the pulses at TP4 and TP5 with the alternating pulses at TP2 and TP3 so that the result is a pulse at the base of Q9, followed by a pulse at the base of Q10, a pulse at the base of Q9, and so on. The pulses alternate 25 times between Q9 and Q10. Then, as TP4 goes to 0 V, TP5 rises from 0 V to some positive voltage and the alternating pulses appear at the bases of Q7 and Q8. The pulses alternately go to the bases of Q7 and Q8 25 times. Then they alternate between the bases of Q9 and Q10 25 times. This pattern continues as long as the unit is in DF operation.

In **Fig 40**, two leads of a four-conductor-plus-ground cable to the antenna-switch board are connected to points A and B. The same pulses that turn Q7 and Q8 off and on turn on and off the left and right antennas. One of those two antennas is turned on and off in phase with Q7 and the other is turned on and off in phase with Q8. The PHASE switch, S3, determines which antenna is in phase with which transistor. Similarly, the two front/back antennas are turned on and off in phase with Q9 and Q10, and S4 determines which antenna is in phase with which transistor.

The pulses arriving at points A and B turn on and off the diodes connecting the coax of the left and right

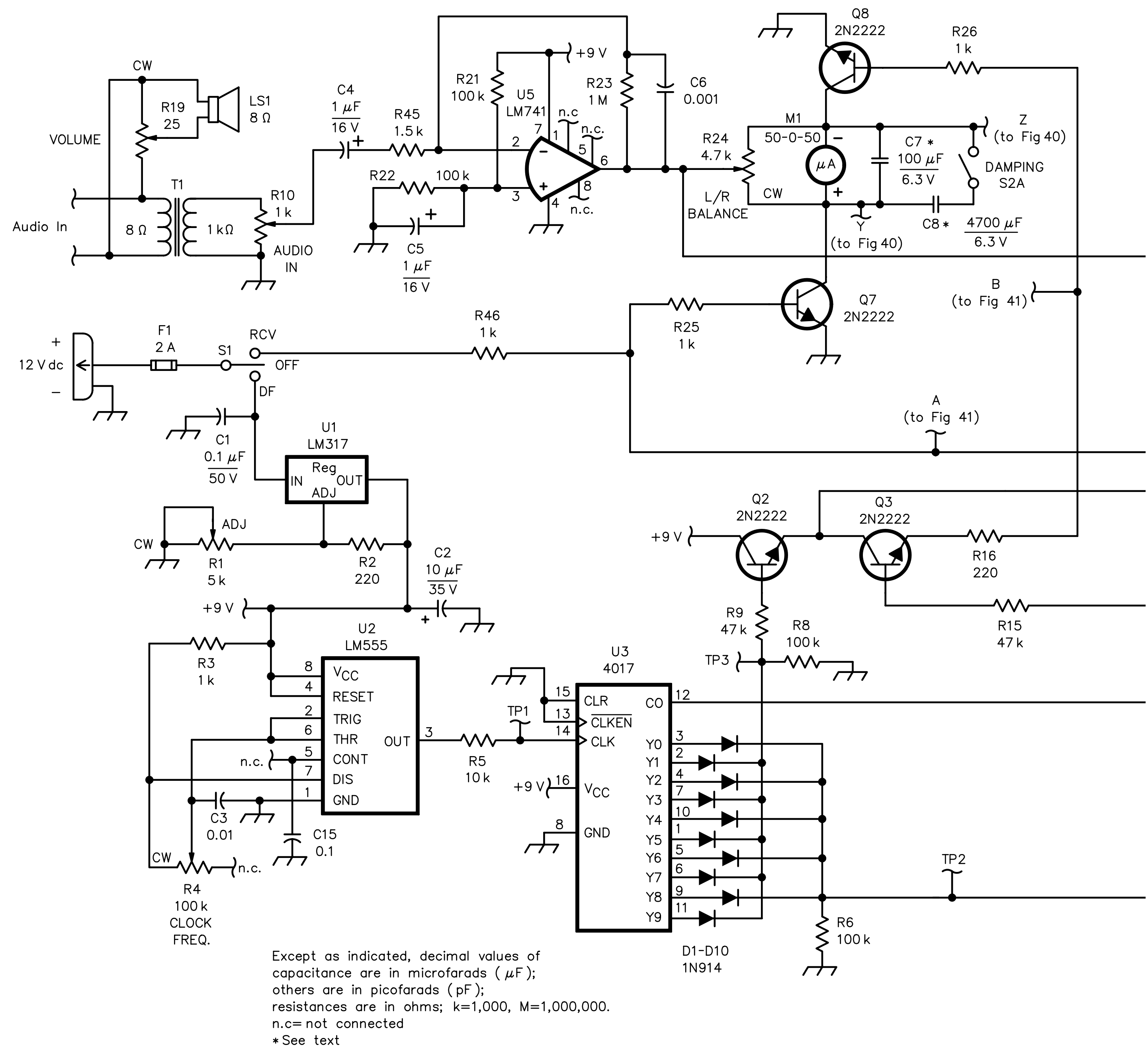
CW
VOLUME
R19
25
LS1
8 Ω
T1
Audio In
8 Ω
1 kΩ
R10
1 k
AUDIO IN
C4
1 µF
16 V
R45
1.5 k
R22
100 k
C5
1 µF
16 V
R21
100 k
U5
LM741
n.c
n.c.
+9 V
R23
1 M
C6
0.001
R24
4.7 k
L/R BALANCE
CW
M1
50-0-50
µA
C7 *
100 µF
6.3 V
C8 *
4700 µF
6.3 V
Z
(to Fig 40)
Y
(to Fig 40)
DAMPING
S2A
Q8
2N2222
R26
1 k
Q7
2N2222
R25
1 k
B
(to Fig 41)
A
(to Fig 41)
+
12 V dc
–
F1
2 A
S1
RCV
OFF
DF
R46
1 k
C1
0.1 µF
50 V
U1
LM317
Reg
IN
OUT
ADJ
R1
5 k
R2
220
C2
10 µF
35 V
+9 V
U2
LM555
VCC
RESET
TRIG
THR
CONT
DIS
GND
OUT
R3
1 k
C3
0.01
C15
0.1
R4
100 k
CLOCK FREQ.
n.c.
R5
10 k
TP1
U3
4017
CLR
CLKEN
CLK
VCC
GND
CO
Y0
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
Q2
2N2222
Q3
2N2222
R16
220
R15
47 k
R9
47 k
R8
100 k
TP3
TP2
R6
100 k
D1-D10
1N914

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.
n.c= not connected
∗See text

Fig 39—Unless otherwise specified, part numbers in parentheses are RadioShack. All fixed-value resistors are $^{1}/_{4}$-W, 5%-tolerance units. Equivalent parts can be substituted.

C1, C15—0.1-µF, 50-V (272-1069)
C2—10-µF, 35-V electrolytic capacitor (272-1025)
C3—0.01-µF, 25-V disc-ceramic capacitor (272-131)
C4, C5—1-µF, 16-V electrolytic capacitor (272-1434)
C6—0.001-µF, 25-V disc-ceramic capacitor (272-126)
C7, C9—100-µF, 6.3-V bipolar (nonpolarized) capacitor; Digi-Key P-1102, available from Digi-Key Corp, 701 Brooks Ave S, PO Box 677, Thief River Falls, MN 56701-0677, tel 800-344-4539, 218-681-6674; fax: 218-681-3880; RadioShack stocks 100-µF, 35-V axial (272-1016) and radial-lead (272-1028) electrolytic capacitors.
C8, C10—4700-µF, 6.3-V bipolar capacitor (made of five 1000-µF, 6.3-V bipolar capacitors), Digi-Key P1106. Standard 1000-µF, 35-V radial and axial-lead electrolytic capacitors are available from RadioShack; a 4700-µF, 35-V axial-lead electrolytic capacitor is also available (272-1022). Note that C7, C8, C9 and C10 are non-polarized capacitors because a small reverse voltage can appear across the meter and capacitors when the system is in use. Standard polarized electrolytics have been used in a number of units using this circuit (the same detector circuit used in W9DUU's unit) without any known ill effects, however.
D1-D20—1N914 silicon switching diode (276-1620 or 276-1122)
F1—2-A fuse (270-1007)
LS1—8-W speaker (40-245)

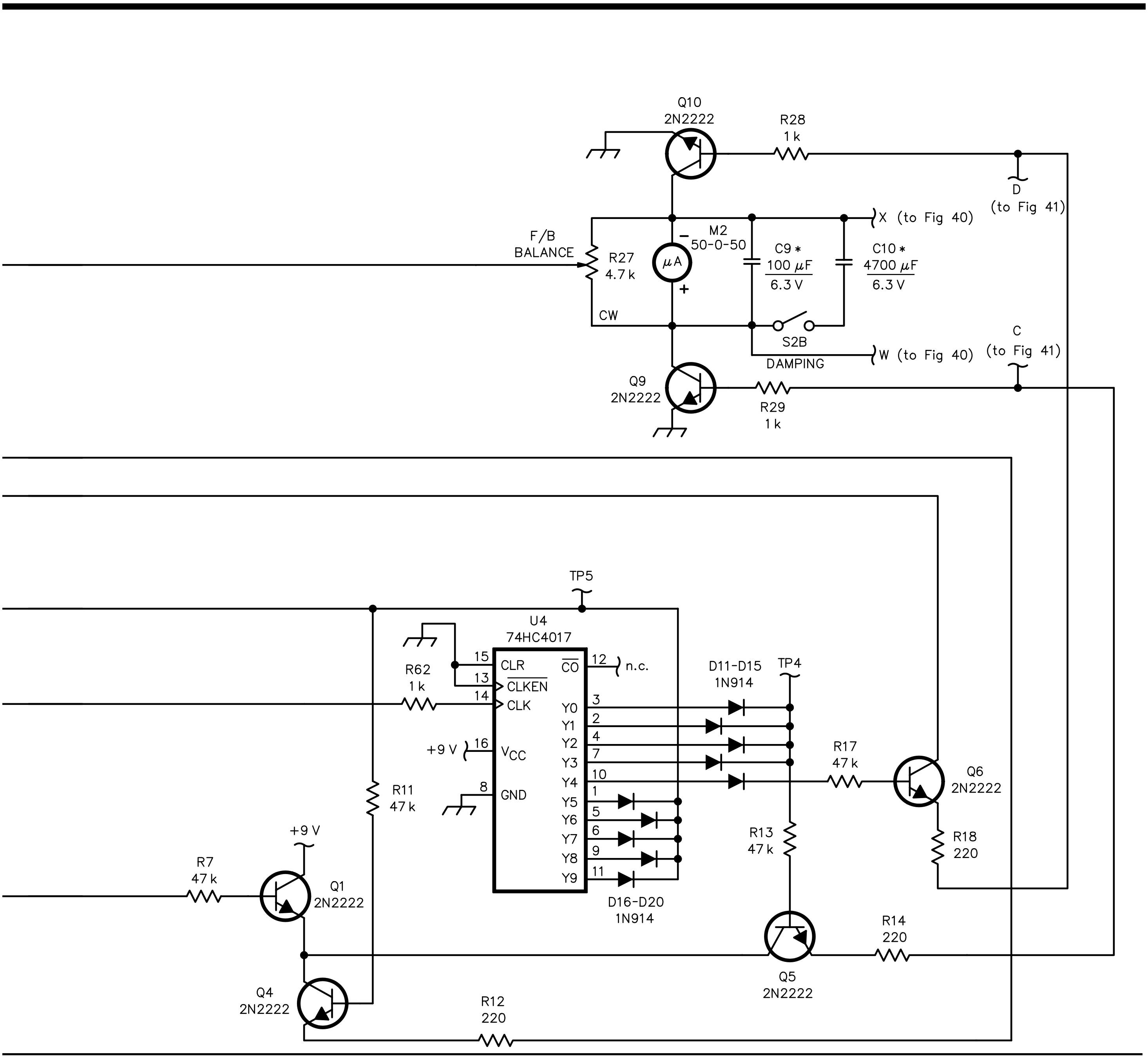
Q10
2N2222
R28
1 k
D
(to Fig 41)
X (to Fig 40)
F/B
BALANCE
R27
4.7 k
M2
50-0-50
μA
C9 *
100 μF
6.3 V
C10 *
4700 μF
6.3 V
CW
S2B
DAMPING
W (to Fig 40)
C
(to Fig 41)
Q9
2N2222
R29
1 k
TP5
U4
74HC4017
15
CLR
13
CLKEN
14
CLK
CO
12
n.c.
R62
1 k
+9 V
16
VCC
8
GND
Y0
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
3
2
4
7
10
1
5
6
9
11
D11-D15
1N914
TP4
D16-D20
1N914
R17
47 k
Q6
2N2222
R18
220
R13
47 k
R11
47 k
+9 V
R7
47 k
Q1
2N2222
Q5
2N2222
R14
220
Q4
2N2222
R12
220

M1, M2—Zero-center, 50-μA meter; optional—see text
Q1-Q10—MPS2222 or 2N2222 NPN silicon general-purpose transistors (276-2009)
R1, R24, R27—4.7 kΩ (271-281); note: many of the fixed-value resistors can be found in RadioShack resistor assortment packages 271-308 and 271-312.
R2, R12, R14, R16, R18—220 Ω (271-1330)
R3, R25, R26, R28, R29—1 kΩ (271-1321)
R4—100-kΩ trimmer potentiometer (271-284)
R5—10 kΩ (271-1335)
R6, R8, R21, R22—100 kΩ (271-1347)
R7, R9, R11, R13, R15, R17—47 kΩ (271-1342)
R10—1-kΩ trimmer potentiometer (271-280)
R19—25-Ω panel-mount potentiometer (271-265A)
R23—1 MΩ (271-1134)
R45—1.5 kΩ (part of 271-312 assortment)
S1—SPDT, center-off switch (275-325)
S2—DPDT switch (275-626)
T1—8-Ω to 1-kΩ audio-output transformer (273-1380)
U1—LM317T, 1.5-A, three-terminal, adjustable voltage regulator (276-1778)
U2—555 timer (276-1723)
U3, U4—4017 decade counter (276-2417)
U5—LM741 op amp (276-007)
Misc: two 8-pin IC sockets (276-1995); two 16-pin IC sockets (276-1992); experimenter's PC board (276-148) or FAR Circuits PC board set; enclosure; four mag-mount antennas, four-conductor shielded cable; in-line fuse holder (270-1281).

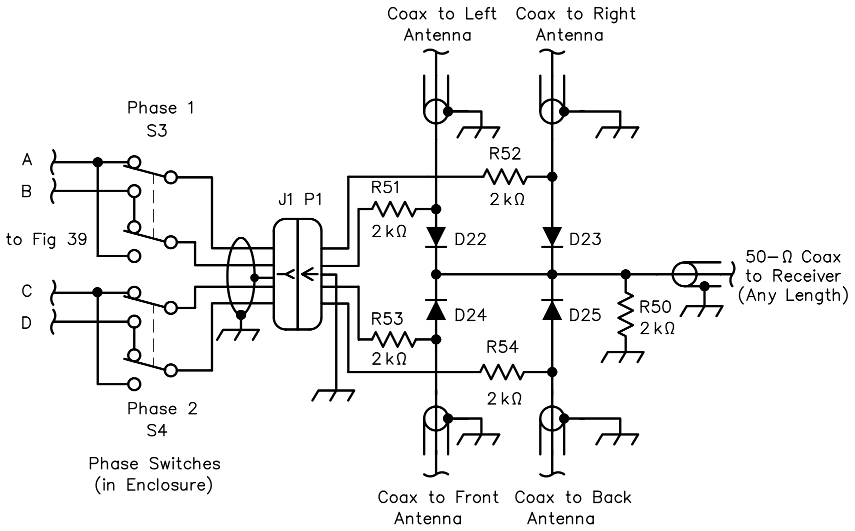

Coax to Left Antenna
Coax to Right Antenna
Phase 1
S3
A
B
to Fig 39
C
D
Phase 2
S4
Phase Switches (in Enclosure)
J1 P1
R51
2 kΩ
R52
2 kΩ
D22
D23
R53
2 kΩ
D24
D25
R54
2 kΩ
R50
2 kΩ
50-Ω Coax to Receiver (Any Length)
Coax to Front Antenna
Coax to Back Antenna

Fig 40—Schematic of the antenna switch board. Part numbers in parentheses are RadioShack. All fixed-value resistors are $^1/_4$-W, 5%-tolerance units. Equivalent parts can be substituted.

D22-D25—1N914 silicon switching diode (276-1620 or 276-1122)
J1—Six-pin female Molex connector (274-236 or 274-155)
P1—Six-pin male Molex connector (274-226 or 274-152)
R50-R54—2.2 kΩ (can be found in RadioShack resistor assortment packages 271-308 and 271-312); also available in pack of five (271-1325)
S3, S4—DPDT switch (275-626)

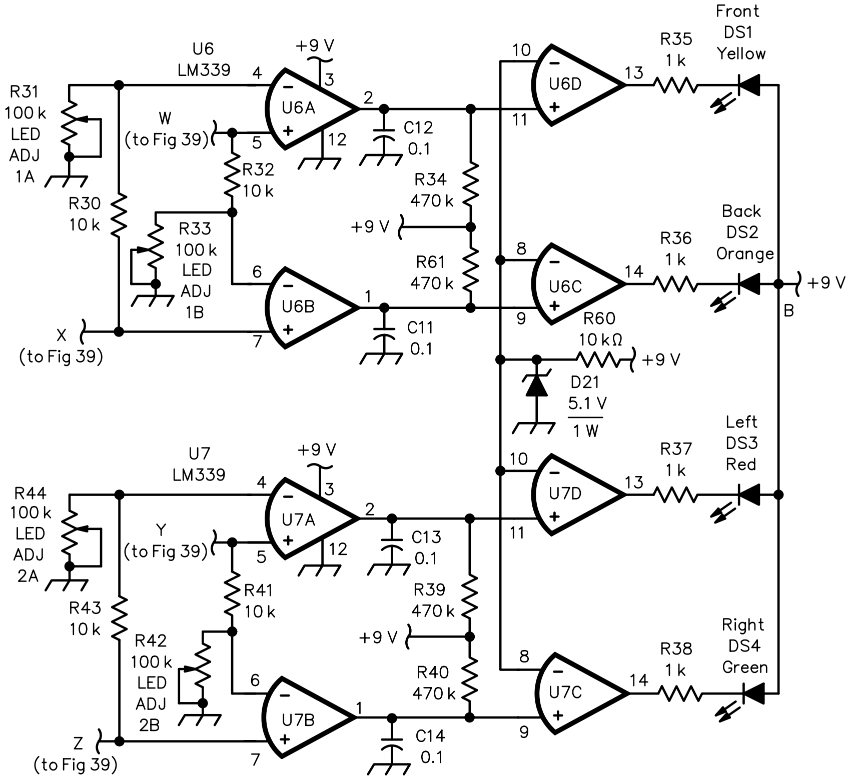

U6
LM339
+9 V
4
3
2
U6A
5
12
R31
100 k
LED
ADJ
1A
W
(to Fig 39)
C12
0.1
R32
10 k
R30
10 k
R33
100 k
LED
ADJ
1B
R34
470 k
+9 V
R61
470 k
6
U6B
1
7
C11
0.1
X
(to Fig 39)
10
U6D
11
13
R35
1 k
Front
DS1
Yellow
8
U6C
9
14
R36
1 k
Back
DS2
Orange
+9 V
B
R60
10 kΩ
+9 V
D21
5.1 V
1 W
U7
LM339
+9 V
4
3
2
U7A
5
12
R44
100 k
LED
ADJ
2A
Y
(to Fig 39)
C13
0.1
R41
10 k
R43
10 k
R42
100 k
LED
ADJ
2B
R39
470 k
+9 V
R40
470 k
6
U7B
1
7
C14
0.1
Z
(to Fig 39)
10
U7D
11
13
R37
1 k
Left
DS3
Red
8
U7C
9
14
R38
1 k
Right
DS4
Green

Fig 41—Schematic of the LED driver circuit. Part numbers in parentheses are RadioShack. All fixed-value resistors are $^1/_4$-W, 5%-tolerance units. Equivalent parts can be substituted.

C11-C14—0.1-µF, 25-V disc-ceramic capacitor (272-135)
D21—1N4733, 5.1-V, 1-W Zener diode (276-565)
DS1-DS4—LEDs; one each red (276-066); green (276-022); yellow (276-021); orange (276-012)
R30, R32, R41, R43—10 kΩ (271-1335)
R34, R39, R40, R61—470 kΩ (271-1354)
R35-R38—1 kΩ (271-1321)
R31, R33, R42, R44—100-kΩ trimmer potentiometer (271-284)
U6, U7—LM339 quad comparator (276-1712)
Misc: two 14-pin IC sockets (276-1999)

antennas to the receiver coax. This occurs 25 times, thereby switching the receiver between the left and right antennas 25 times. Similar switching then occurs between the front and the back antennas from pulses arriving at points C and D.

Detector Circuit

The detector circuit (back again in Fig 39) starts with U5, a 741 op amp that amplifies the receiver's audio output. The audio is fed into R24 and R27, two 4.7-kΩ pots. The zero-center, 50-µA meters across R24 and R27 are optional. The meters, as well as the LEDs, indicate front/back and left/right. Such meters can be expensive unless you find surplus meters, and they're not really necessary.

When the left/right antennas are active, one end of R24 is grounded by Q7 and Q8 on each alternate pulse.

If Q7 is conducting, Q8 is not conducting. On each pulse, one of the left/right antennas is turned on and one end of R24 is grounded. On the next pulse, the other left/right antenna is turned on and the other end of R24 is grounded. If there is a phase difference between the signal received by the left antenna and the signal received by the right antenna, a dc voltage is built up across R24. That voltage causes the quad comparator, U7 in **Fig 41** to turn on DS3 (red) or DS4 (green) L/R LEDs. If the optional left/right meter is installed, it deflects to indicate the direction as do the LEDs.

After 25 cycles, the left and right antennas are both turned off and the front and back antennas are cycled on and off 25 times with the same detection process, producing a voltage across R27 if there is a phase difference between the RF received by the front and the back antennas. That voltage across R27 causes quad comparator U6 to turn on DS1 or DS2.

C9 and C10, for the F/B detector, and C7 and C8, for the L/R detector, damp the voltage swings caused by multipath reception. To control damping, S2A and S2B switch the 4700-µF capacitors in or out of the circuit. You want the greatest amount of damping when you drive through an area with a lot of multipath propagation (as from buildings); a lot of damping helps under those circumstances.

Construction

The prototype was built using a pad-per-hole RadioShack board. However, a PC board makes construction a lot faster. Far Circuits offers a printed-circuit project on their Web site: **www.cl.ais.net/farcir**). Except for the optional meters and the nonpolarized capacitors, most parts are available from RadioShack.

First build the power supply so you can power the unit from your car or another 12-V source. Apply 12 V to the DFer and adjust R1 until U1's output is +9 V. (You can use a 9-V battery and omit the power-supply section, but you'd better take along a spare battery when you go DFing.)

Install U2 and its associated parts. Power up and turn on S1. A string of pulses should appear at TP1. If you have a frequency counter, set R4 for a pulse frequency of 2200 Hz at TP1. If you don't have a counter, connect a 0.1-µF capacitor from TP1 to headphones or a small speaker and set R4 for a tone of about 2 kHz. Later, you'll adjust the clock so the unit works with the passband of your receiver.

Turn off S1 and remove the power source. Install U3 and its diodes. Pin 12 of U3 need not be connected yet. Apply power and turn on S1. At TP2 and TP3, you should find alternating 1100-Hz pulses. If you have a dual-trace scope, you can see that the pulses alternate. If you have a single-trace scope, connect TP3 to TP2 and to the scope input and the trace will appear as a solid line as there is a pulse at either TP2 or TP3 at all times. If you don't have a scope, the tone in a speaker or earphones from TP2 or TP3 will sound half as high (about 1 kHz) as the tone at TP1.

Turn off S1. Install U4 and its diodes. Note that pin 12 of U3 is connected to pin 14 of U4. At TP4 and TP5, there should be long pulses—five times longer than the pulses at TP1, and the pulses should alternate between TP4 and TP5. The pulse frequency should be about 44 Hz. Power down and turn off S1. Install the remaining circuit components. When you power up, 25 alternating pulses should appear at A and B, then 25 alternating pulses should appear at C and D. Use a scope to verify that.

If you're not using the optional panel meters, connect a voltmeter across R24 (L/R BALANCE), using the lowest dc-voltage range. Note that neither end of R24 is grounded. With no audio input, adjust R24 until there is no voltage across it. Do the same with R27 (F/B BALANCE). If you use the optional meters, adjust R24 and R27 so there's no current shown on either meter.

Power down and assemble the rest of the circuit. With power applied, but with no audio input, adjust R31, 33, 42 and 44 so that the four LEDs (DS1 through DS4) are off. The objective of the following adjustments is to get the red and green LEDs to turn on with the same voltage amplitude, but opposite polarity, across R24. Move R24's wiper so a low positive voltage appears across R24, as indicated by the voltmeter connected across R24 or movement of the panel-meter needle. Adjust R44 (LED ADJ 1A) and R42 (LED ADJ 1B) so that the green LED (DS4) comes on when the voltage goes positive at one end of

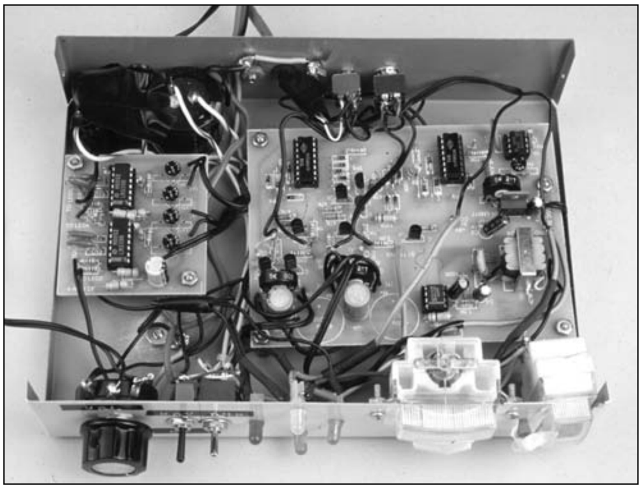

Fig 42—An inside view of one DF unit built into a 2 × 8 × 5¾-inch (HWD) box. Because of the height restriction, the two 4700-µF damping capacitors (C8 and C10) are not mounted on the PC board, but near the rear panel behind the smaller of the two PC boards. One of the 4700-µF damping capacitors is a standard electrolytic, the other is a parallel combination of five 1000-µF, 6.3-V bipolar (nonpolarized) capacitors wrapped in electrical tape.

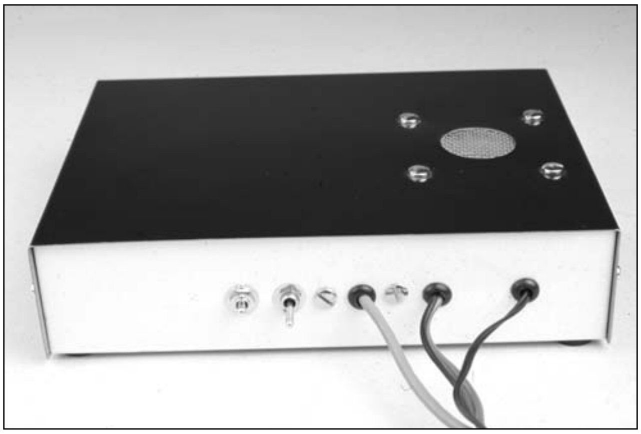

Fig 43—The rear panel of the DF unit supports the two DPDT PHASE toggle switches. Grommets in the panel holes allow abrasion-free passage of the antenna, dc-power and audio cables. The dc power cord is outfitted with an inline fuse holder and a male Jones plug. A six-pin female Molex connector (five pins are used) feeds the four antennas. The audio-input cable is terminated in a $^1/_8$-inch diameter male plug.

R24, but goes off when R24 is adjusted for 0 V across R24.

Next, adjust R24 for a slight negative-voltage indication and adjust R42 and R44 so that the red LED (DS3) comes on, but extinguishes when the voltage across R24 is 0 V. When you're done, adjusting R24's wiper slightly one way should illuminate the red LED. Both LEDs should be off when there's no voltage across R24; rotating R24's wiper slightly in the opposite direction should turn on the green LED.

Connect the voltmeter across R27 or use the panel meter as an indicator. With no audio input, adjust R27 so that there's no voltage across R27. Adjust R31 (F/B LED ADJ 1A) and R33 (F/B LED 1B) so that a F/B LED (DS1) is on when there is a slight positive voltage across R27 and the other F/B LED (DS2) is on when there is a negative voltage across R27.

Switch Board

Assemble the switch board for the four mag-mount antennas (see Fig 40). You can use half of a RadioShack dual pad-per-hole PC board (RS 276-148) as a platform. Lead length is critical only on this board and in the length of the coax from the switching board to the antennas, so avoid wire-wrap construction here.

Feed Lines

The coax lines from the switch board to each of the four antennas must be of equal length. The coax lengths should be long enough to permit each antenna to be placed slightly less than $^1/_4$ λ from its counterpart, at the lowest frequency of operation. (There's a local belief that 27$^1/_2$ inches is the best length. The author used that length and it works, but other lengths might work as well.) An attempt to locate the switch board inside the vehicle and run equal-length 12-foot-long cables to the antennas failed. Keep the switch board on the vehicle's roof.

Use the same type of coax for all lines—that is, don't mix foam and polyethylene dielectric coax on the antenna lines. If the velocity factor of the lines is not equal, a phase shift in the signals will exist even when the transmitter is dead ahead and it will lead you astray.

Solder the coax from the antennas directly to the board—don't use connectors. From the switch board to the receiver, use 50-Ω coax of any type and length. Equip the receiver end of the line with a connector that mates with your receiver's antenna-input jack. Make the four-conductor shielded cable from the main DFer box to the switch board long enough to reach from the LRFB box's operating position to the roof of the car. Construct the antennas so that the whips can be changed easily for use on any frequency in which you're interested.

Mechanical Assembly

Mount the finished PC boards in a metal box of your choice. You can follow the construction method used in the prototype (see **Fig 42** and **Fig 43**), but do ensure that R4 can be adjusted easily with a tuning tool from outside the box. A hole drilled in the box at the right point will suffice. Arrange the LEDs in a diamond pattern on the front panel, with the left LED (red) to the left, the right LED (green) on the right, and the front and back LEDs at the top and the bottom.

Wrap the switch board with tape to waterproof it, or place it in a watertight box. Arrange the mag-mount antennas on top of the vehicle in a diamond shape. The distance between each antenna pair should be less than $^1/_4$ λ at the operating frequency. (In limiting the distance between antennas in a pair—F/B or L/R—to less than $^1/_4$ λ, the author followed W9DUU's example.)

Final Adjustments

It's best to start on 2 meters. Install the 2-meter whips in the antenna bases. Mount the antennas on the top of your vehicle. Identify the left and right antenna bases with L and R, and mark the front and back antenna bases, too.

A good way to find out if the antennas are properly installed is to short either the left or right antenna with a clip lead from the whip to the metal base; the L/R meter will deflect one way and the left or right LED will light. If you short one of the front and back antennas with a clip lead, the front or back LED will come on. If you short one of the L/R antennas and it makes the meter go to the right, it does not necessarily mean you have the PHASE switch in the proper position. That depends on the relative phase determined by the number of audio stages in your receiver, each of which may contribute a shift of 180°.

Connect the coax from the switch board to your FM receiver. It must be an FM receiver; an AM VHF or AM aircraft receiver won't suffice. Connect the audio output of your 2-meter receiver to the LRFB box audio input. If you're using a transceiver, disable its transmit function by removing the mike. You don't want to transmit into the switch board!

Turn on the receiver and place the LRFB box in receive by setting S1 to the RCV position. Center R10. Only one of the antennas will be turned on and the DF operation will be disabled. Back off the squelch and notice that you hear the audio from the speaker of the LRFB box. Switch to an unused simplex channel. Have a friend with an HT stand 20 feet or so in front of the vehicle. With the receiver in the car turned off, turn S1 to DF. The four LEDs should be off. If an LED lights, adjust R24 and R27 for zero voltage. If the LED is still on, readjust R31 and R33, or R42 and R44 as explained earlier. When all four LEDs are off with the antennas connected, no audio from the receiver and no RF input signal, you're ready. Turn on your receiver and have your friend transmit on the 2-meter frequency (simplex) that your receiver is set to. When he transmits, one or more of the LEDs should illuminate. Ignore the front/back LEDs, but check to see if the right or left LED is on. If the right LED is on and your friend is standing to the right of the center of the vehicle front, all is well. Have him walk back and forth in front of the vehicle and notice that when he is to the right, the green LED turns on, and when he's to the left, the red LED glows. If the indications are reversed, that is, if the LRFB box indicates left when your friend is to your right, reverse the position of PHASE switch S3.

When the L/R indicators are working properly, have your friend walk back and forth between a position 20 feet to the front and right of the car and a position 20 feet away to the rear and right of the car. The right LED should stay on, but the front LED should be on when the HT is in front of the antennas, and the back LED should be on when the HT is behind the antennas. If the front/back indications are reversed, reverse the position of PHASE switch S4.

While receiving a signal from your friend's HT, adjust R4 for maximum deflection on any one of the two optional meters, or on a voltmeter (set on the lowest dc voltage range) connected across R24 or R27. Look for the maximum deflection of the meter needle as the meter swings both ways while the signal source moves from back to front or left to right (depending, of course, on which panel meter you're looking at or which resistor—R24 or R27—your voltmeter is across). The audio passband of FM receivers varies, and if you switch from an ICOM IC-27 to an ICOM IC-W2A, for example, you'll have to change U2's master clock frequency. You may encounter receivers that don't require readjusting R4, but such readjustment should be expected. When changing receivers, you may also need to change the position of PHASE switches S3 and S4. This may also be necessary when changing from UHF to VHF, or VHF to UHF on the same receiver, as the number of receiver stages (and, hence, the audio phase) may change from band to band.

Audio Level Adjustments

With meter damping (S2) off, adjust R10 so that you have full deflection of one of the two meters (or your multimeter) with the signal source at a 45° angle from the vehicle (halfway between ahead and right), and a reasonable audio level from the speaker. Turn the rig's volume control all the way up to ensure that the audio circuit doesn't overload. If it overloads, the meters won't deflect. From zero to full blast, the meter should deflect more and more (unless the signal is straight ahead or exactly left or right). If you're using an H-T, maximum deflection of the meter should occur before the volume control is $^3/_4$ of maximum, or before the volume control is at $^1/_2$ of maximum if you're using a mobile rig. If R10 is properly adjusted, turning the volume control to maximum won't cause the meter to fall back toward zero. If increasing the volume causes the meter to deflect less, then R10's setting is too high.

Now have your friend walk around the vehicle with the HT transmitting and notice that the LEDs indicate the signal direction. On 2 meters in a clear field, the indications should be correct 80 or 90% of the time. The erroneous readings that occasionally occur are due to multipath propagation caused by the irregular shape of the vehicle. Slight adjustments in the positions of the L/R and F/B antennas may be necessary to make the zero points fall directly in front of the vehicle (neither left nor right LED on) and at the center of the antennas (neither front nor back LED on).

Try the same procedure on 440 MHz. You may have to flip the PHASE switches when you move to another band, even when using the same receiver. Remember that the total length of the 440-MHz antennas must be $^1/_4\ \lambda$ or less, and the antennas must be placed less than $^1/_4\ \lambda$ apart. The results on 440 MHz probably won't be as consistent as the results on 2 meters, as there is likely to be a lot more multipath propagation caused by the irregular shape of the vehicle.

Fox Hunting

Before heading out to find the fox, check to be certain the LRFB box is working properly. Tune to the fox's frequency and drive off. Turning on the DAMPING switch stabilizes the indication as you drive along. Follow indications generated as you travel over the road or street. If an indication is constant for 15 or 20 seconds while you're moving down the road, it's probably the true direction to the fox. It's possible to have a reflection from a mountain over a long distance down the road, however. When you can hear the fox with no antenna, it's time to get out of the car, switch to the hand-held system and hunt for the fox on foot.

If you switch to a new receiver, you may have to readjust R4, CLOCK FREQ. That's because both receivers may not have the same audio bandwidth. Author WA9BVS forgot this while chasing a balloon once and the results were comical. When chasing balloons with ham radio transmitters, the readings you get are likely to be confusing when the balloon is at a high angle with respect to the plane of the car top. Use a hand-held Yagi to verify the balloon location. Even with a simple buzzbox, you should be able to find a keyed transmitter.

A TAPE-MEASURE BEAM FOR RDFING

Joe Leggio, WB2HOL, designed this antenna while searching for a beam with a really great front-to-back ratio to use in hidden transmitter hunts. It exhibits a very clean pattern and is perfect for RDF use. You can construct this beam using only simple hand tools, and it has been duplicated many times.

WB2HOL's first design requirement was to be able to get in and out of his car easily when hunting for a hidden transmitter. He accomplished this by using steel "tape-measure" elements, which fold easily when putting the antenna into a car and yet are self supporting. They also hold up well while crashing through the underbrush on a fox hunt.

WB2HOL decided to use three elements to keep the boom from getting too long. He used inexpensive

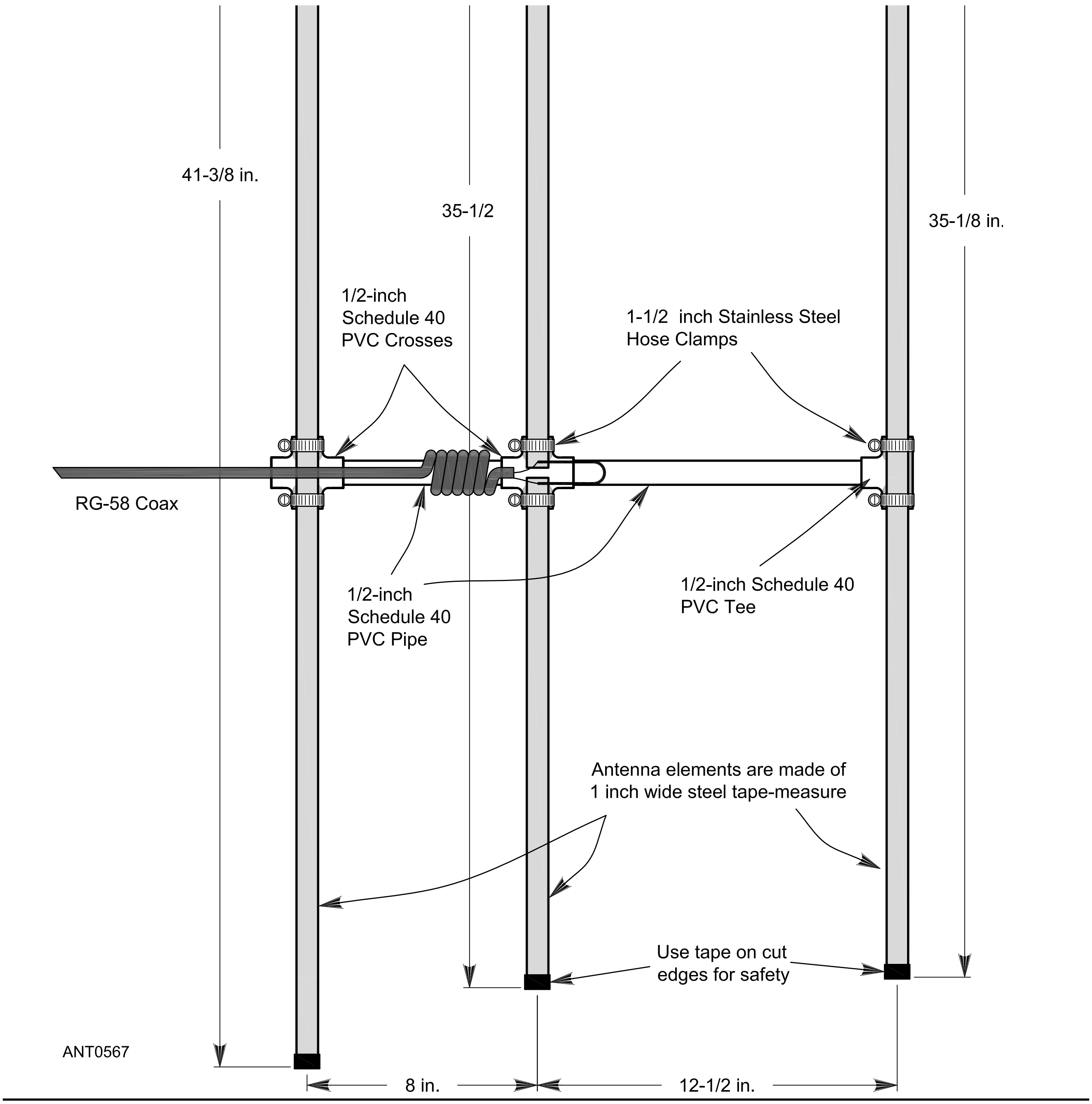
41-3/8 in.
35-1/2
35-1/8 in.
1/2-inch Schedule 40 PVC Crosses
1-1/2 inch Stainless Steel Hose Clamps
RG-58 Coax
1/2-inch Schedule 40 PVC Pipe
1/2-inch Schedule 40 PVC Tee
Antenna elements are made of 1 inch wide steel tape-measure
Use tape on cut edges for safety
ANT0567
8 in.
12-1/2 in.

Fig 44—Tape-measure beam dimensions.

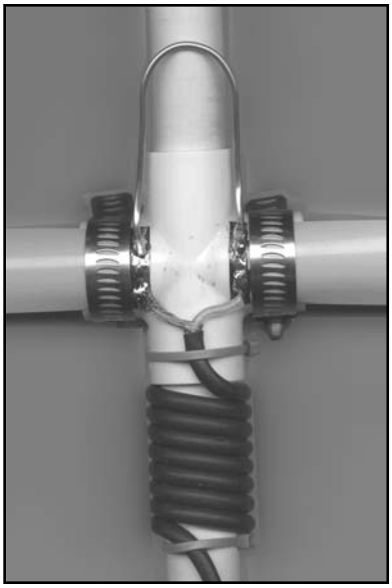

Fig 45—Photo of driven-element mounted to PVC tee using hose clamps. The hairpin match wires is shown here soldered to the tape-measure elements, along with the RG-58 feed line.

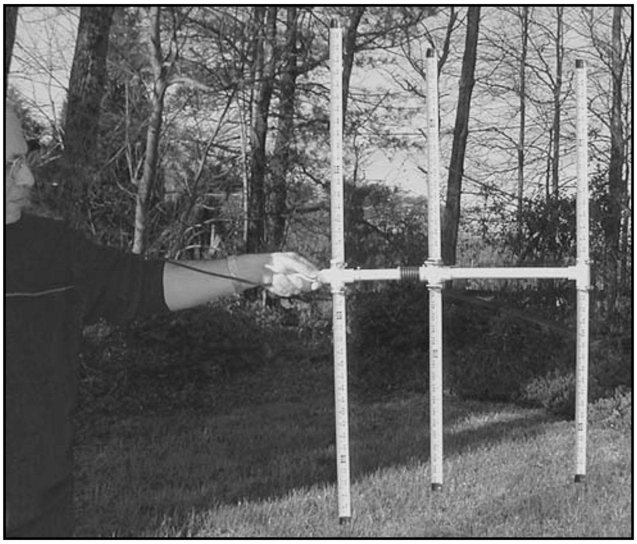

Fig 46—Photo of complete tape-measure beam, ready to hunt foxes!

Schedule-40 PVC pipe, crosses and tees that can be found at any hardware store for the boom and element supports.

He used a simple hairpin match, consisting of a 5-inch length of #14 solid wire bent into the shape of a U, with the two legs about $^{3}/_{4}$ inch apart. This gave in a very good match across the 2-meter band after he tweaked the distance (1-inch on his prototype) between the halves of the driven element for minimum SWR.

You can cut the 1-inch wide tape-measure elements with a pair of shears, chamfering the ends of the elements. Be very careful—the edges are very sharp and will inflict a nasty cut if you are careless. Use some sandpaper to remove the sharp edges and burrs and put some vinyl electrical tape on the ends of the elements to protect yourself from getting cut. Wear safety glasses while cutting the elements. See **Fig 44** for dimensions.

The RG-58 coax feed line is wound into a 8-turn coil along the boom to form a choke balun. The coil is covered with vinyl black tape to secure it to the boom. Make sure you scrape or sand the paint off the tape-measure elements where the feed line is attached. Most tape measures have a very durable paint finish designed to stand up to heavy use. You do not want the paint to insulate your feed-line connection!

If you are careful, you can solder the feed line to the element halves, but take care since the steel tape measure does not solder easily and the PVC supports can be easily melted. Tin the tape-measure elements before mounting them to the PVC cross if you decide to connect the feed line in this fashion.

If you decide not to solder to the tape-measure elements, you can use two other methods to attach the feed line. One method employs ring terminals on the end of the coax. The ring terminals are then secured under self-tapping screws or with 6-32 bolts and nuts into holes drilled in the driven-element halves. However, with this method you cannot fine-tune the antenna by moving the halves of the driven element in and out.

The simplest method is simply to slide the ends of the feed line under the driven element hose clamps and tighten the clamps to hold the ends of the coax. This is low-tech but it works just fine.

WB2HOL used $1^{1}/_{2}$-inch stainless-steel hose clamps to attach each driven-element half to the PVC cross that acts as its support. This allowed him to fine-tune his antenna for lowest SWR simply by loosening the hose clamps and sliding the halves of the driven element in or out to lengthen or shorten the element. He achieved a 1:1 SWR at 146.565 MHz (the local fox-hunt frequency) when the two elements were spaced about 1 inch apart. **Fig 45** shows the hose-clamp method for attaching the driven element to the PVC cross, along with the hairpin wire and feed-line coax.

Some builders have used rubber faucet washers between the tape-measure elements and the PVC-cross fittings on the director and reflector. These allow for the tape to fit the contour of the PVC fitting better and will make the antenna look nicer. It is normal for the reflector and director elements to buckle a bit as they are tightened to the PVC tee and cross if you don't use faucet washers. You can also eliminate the buckling if you use self-tapping screws to attach these elements instead of hose clamps. The beam will not be as rugged, however, as when you use hose clamps.

This beam has been used on fox hunts, on mountain tops, at local public-service events, outdoors, indoors in attics—just about everywhere. The SWR is typically very close to 1:1 once adjusted. Front-to-back performance is exactly as predicted. The null in the rear of the pattern is perfect for transmitter hunts.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

W. U. Amphar, “Unidirectional Loops for Transmitter Hunting,” *QST*, Mar 1955.

G. Bonaguide, “HF DF—A Technique for Volunteer Monitoring,” *QST*, Mar 1984.

D. S. Bond, *Radio Direction Finders*, 1st edition (New York: McGraw-Hill Book Co).

R. E. Cowan and T. A. Beery, “Direction Finding with the Interferometer,” *QST*, Nov 1985.

D. DeMaw, “Beat the Noise with a Scoop Loop,” *QST*, Jul 1977.

D. DeMaw, “Maverick Trackdown,” *QST*, Jul 1980.

T. Dorbuck, “Radio Direction-Finding Techniques,” *QST*, Aug 1975.

D. T. Geiser, “Double-Ducky Direction Finder,” *QST*, Jul 1981.

D. T. Geiser, “The Simple Seeker,” *The ARRL Antenna Compendium*, *Vol 3*, p 126.

N. K. Holter, “Radio Foxhunting in Europe,” Parts 1 and 2, *QST*, Aug and Nov 1976.

J. Isaacs, “Transmitter Hunting on 75 Meters,” *QST*, Jun 1958.

H. Jasik, *Antenna Engineering Handbook*, 1st edition (New York: McGraw-Hill, 1961).

R. Keen, *Wireless Direction Finding*, 3rd edition (London: Wireless World).

J. Kraus, *Antennas*, 2nd edition (New York: McGraw-Hill Book Co, 1988).

J. Kraus, *Electromagnetics*, 4th edition (New York: McGraw-Hill Book Co, 1992).

C. M. Maer, Jr., “The Snoop-Loop,” *QST*, Feb 1957.

M. C. Mallette, “The Four-Way DFer,” *QST*, Nov 1995. A set of two PC boards is available from FAR Circuits for the “Four-Way DFer” project. 18N640 Field Ct, Dundee, IL 60118-9269, tel 708-576-3540 (voice and fax). Note: No component pads for C15 exist. Mount C15 between U2 pin 5 and ground on the bottom (foil) side of the board. A PC-board template package is available from the ARRL free of charge. Send your request for the MALLETTE 4-WAY DFER, along with a business size SASE, to the Technical Department Secretary, 225 Main St, Newington, CT 06111-1494.

L. R. Norberg, “Transmitter Hunting with the DF Loop,” *QST*, Apr 1954.

P. O’Dell, “Simple Antenna and S-Meter Modification for 2-Meter FM Direction Finding,” Basic Amateur Radio, *QST*, Mar 1981.

Ramo and Whinnery, *Fields and Waves in Modern Radio* (New York: John Wiley & Sons, 1944).

F. Terman, *Electronic and Radio Engineering* (New York: McGraw-Hill Book Co, 1955).

Radio Direction Finding , published by the Happy Flyer, 1811 Hillman Ave, Belmont, CA 94002.

For more information on direction finding, see *The ARRL Handbook* and *Transmitter Hunting: Radio Direction Finding Simplified*, by Joe Moell, K0OV, and Thomas Curlee, WB6UZZ. These books are available from your local dealer or can be ordered directly from ARRL.

Chapter 15

Portable Antennas

For many amateurs, the words *portable antennas* may conjure visions of antenna assemblies that can be broken down and carried in a backpack, suitcase, golf bag, or what-have-you, for transportation to some out-of-the way place where they will be used. Or the vision could be of larger arrays that can be disassembled and moved by pickup truck to a Field Day site, and then erected quickly on temporary supports. Portable antennas come in a wide variety of sizes and shapes, and can be used on any amateur frequency.

Strictly speaking, the words "portable antenna" really means *transportable antenna*—one that is moved to some (usually temporary) operating position for use. As such, portable antennas are not placed into service when they are being transported. This puts them in a different class from mobile antennas, which are intended to be used while in motion. Of course this does not mean that mobile antennas cannot be used during portable operation. Rather, true portable antennas are designed to be packed up and moved, usually with quick reassembly being one of the design requisites. This chapter describes antennas that are designed for portability. However, many of these antennas can also be used in more permanent installations.

Any of several schemes can be employed to support an antenna during portable operation. For HF antennas made of wire, probably the most common support is a conveniently located tree at the operating site. Temporary, lightweight masts are also used. An aluminum extension ladder, properly guyed, can serve as a mast for Field Day operation. Such supports are discussed in Chapter 22, Antenna Supports.

A SIMPLE TWIN-LEAD ANTENNA FOR HF PORTABLE OPERATION

The typical portable HF antenna is a random-length wire flung over a tree and end-fed through an antenna tuner. Low-power antenna tuners can be made quite compact, but each additional piece of necessary equipment makes portable operation less attractive. The station can be simplified by using resonant impedance-matched antennas for the bands of interest. Perhaps the simplest antenna of this type is the half-wave dipole, center-fed with 50- or 75-Ω coax. Unfortunately, RG-58, RG-59 or RG-8 cable is quite heavy and bulky for backpacking, and the miniature cables such as RG-174 are too lossy.

A practical solution to the coax problem, developed by Jay Rusgrove, W1VD, and Jerry Hall, K1TD, is to use folded dipoles made from lightweight TV twin-lead. The characteristic impedance of this type of dipole is near 300 Ω, but this can easily be transformed to a 50-Ω impedance. The transformation is obtained by placing a lumped capacitive reactance at a strategic distance from the input end of the line. **Fig 1** illustrates the construction method and gives important dimensions for the twin-lead dipole.

A silver-mica capacitor is shown for the reactive element, but an open-end stub of twin-lead can serve as well, provided it is dressed at right angles to the transmission line for some distance. The stub method has the advantage of easy adjustment of the system resonant frequency.

The dimensions and capacitor values for twin-lead dipoles for the HF bands are given in **Table 1**. To pre-

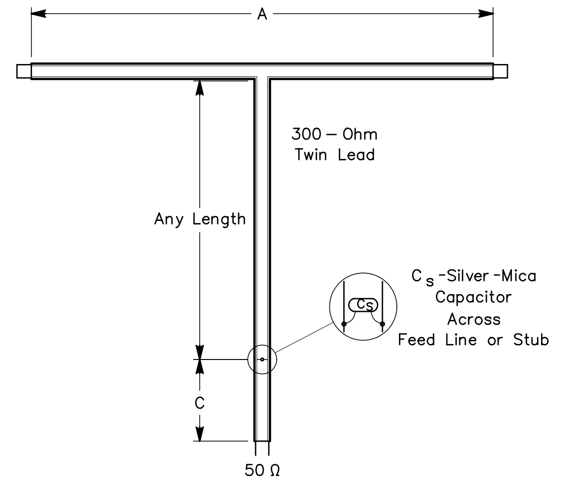

Fig 1—A twin-lead folded dipole makes an excellent portable antenna that is easily matched to 50-Ω equipment. See text and Table 1 for details.

Table 1

Twin-Lead Dipole Dimensions and Capacitor Values

Frequency	*Length A*	*Length C*	C_s	*Stub Length*
3.75 MHz	124' 9½"	13' 0"	289 pF	37' 4"
7.15	65' 5½"	6' 10"	151 pF	19' 7"
10.125	46' 2½"	4' 10"	107 pF	13' 10"
14.175	33' 0"	3' 5½"	76 pF	9' 10½"
18.118	25' 10"	2' 8½"	60 pF	7' 9"
21.225	22' ½"	2' 3½"	51 pF	6' 7"
24.94	18' 9"	1' 11½"	43 pF	5' 7½"
28.5	16' 5"	1' 8½"	38 pF	4' 11"

serve the balance of the feeder, a 1:1 balun must be used at the end of the feed line. In most backpack QRP applications the balance is not critical, and the twin-lead can be connected directly to a coaxial output jack—one lead to the center contact, and one lead to the shell.

Because of the transmission-line effect of the shorted-radiator sections, a folded dipole exhibits a wider bandwidth than a single-conductor type. The antennas described here are not as broad as a standard folded dipole because the impedance-transformation mechanism is frequency selective. However, the bandwidth should be adequate. An antenna cut for 14.175 MHz, for example, will present an SWR of less than 2:1 over the entire 14-MHz band.

ZIP-CORD ANTENNAS

Zip cord is readily available at hardware and department stores, and it's not expensive. The nickname, *zip cord*, refers to that parallel-wire electrical cord with brown or white insulation used for lamps and many small appliances. The conductors are usually #18 stranded copper wire, although larger sizes may also be found. Zip cord is light in weight and easy to work with.

For these reasons, zip cord can be pressed into service as both the transmission line and the radiator section for an emergency dipole antenna system. This information by Jerry Hall, K1TD, appeared in *QST* for March 1979. The radiator section of a zip-cord antenna is obtained simply by "unzipping" or pulling the two conductors apart for the length needed to establish resonance for the operating frequency band. The initial dipole length can be determined from the equation $\ell = 468/f$, where ℓ is the length in feet and f is the frequency in MHz. (It would be necessary to unzip only half the length found from the formula, since each of the two wires becomes half of the dipole.) The insulation left on the wire will have some loading effect, so a bit of length trimming may be needed for exact resonance at the desired frequency.

For installation, you may want to use the electrician's knot shown in **Fig 2** at the dipole feed point. This is a balanced knot that will keep the transmission-line part of the system from unzipping itself under the tension of dipole suspension. This way, if zip cord of sufficient length for both the radiator and the feed line is obtained, a solder-free installation can be made right down to the input end of the line.

(Purists may argue that knots at the feed point will create an impedance mismatch or other complications, but as will become evident in the next section, this is not a major consideration.) Granny knots (or any other variety) can be used at the dipole ends with cotton cord to suspend the system. You end up with a lightweight, low-cost antenna system that can serve for portable or emergency use.

But just how efficient is a zip-cord antenna system? Since it is easy to locate the materials and simple to install, how about using such for a more permanent installation? Upon casual examination, zip cord looks about like 72-Ω balanced feed line. Does it work as well?

Zip Cord as a Transmission Line

To determine the electrical characteristics of zip cord as a radio-frequency transmission line, a 100-foot roll was subjected to tests in the ARRL laboratory with an RF impedance bridge. Zip cord is properly called *parallel power cord*. The variety tested was manufactured for GC Electronics, Rockford, IL, being 18-gauge, brown, plastic-insulated type SPT-1, GC cat. no. 14-118-2G42. Undoubtedly, minor variations in the electrical-characteristics will occur among similar cords from different manufacturers, but the results presented here are probably typical.

The characteristic impedance was determined to be 107 Ω at 10 MHz, dropping in value to 105 Ω at 15 MHz and to a slightly lower value at 29 MHz. The nominal

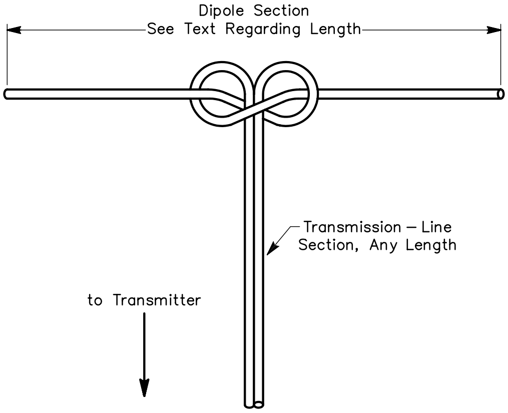

Fig 2—This electrician's knot, often used inside lamp bases and appliances in lieu of a plastic grip, can also serve to prevent the transmission-line section of a zip-cord antenna from unzipping itself under the tension of dipole suspension. To tie the knot, first use the right-hand conductor to form a loop, passing the wire behind the unseparated zip cord and off to the left. Then pass the left-hand wire of the pair behind the wire extending off to the left, in front of the unseparated pair, and thread it through the loop already formed. Adjust the knot for symmetry while pulling on the two dipole wires.

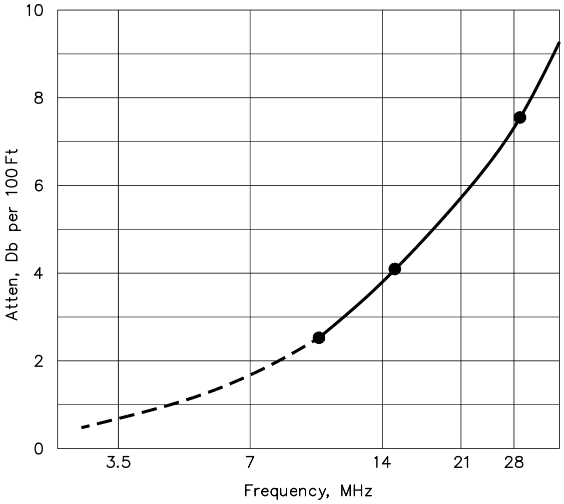

Fig 3—Attenuation of zip cord in decibels per hundred feet when used as a transmission line at radio frequencies. Measurements were made only at the three frequencies where plot points are shown, but the curve has been extrapolated to cover all high-frequency amateur bands.

value is 105 Ω at HF. The velocity factor of the line was determined to be 69.5%.

Who needs a 105-Ω line, especially to feed a dipole? A dipole in free space exhibits a feed-point resistance of 73 Ω, and at heights above ground of less than $^1/_4$ wavelength the resistance can be even lower. An 80-meter dipole at 35 feet over average soil, for example, will exhibit a feed-point resistance of about 35 Ω. Thus, for a resonant antenna, the SWR in the zip-cord transmission line can be 105/35 or 3:1, and maybe even higher in some installations. Depending on the type of transmitter in use, the rig may not like working into the load presented by the zip-cord antenna system.

But the really bad news is still to come—line loss! **Fig 3** is a plot of line attenuation in decibels per hundred feet of line versus frequency. Chart values are based on the assumption that the line is perfectly matched (sees a 105-Ω load as its terminating impedance).

In a feed line, losses up to about 1 decibel or so can be tolerated, because at the receiver a 1-dB difference in signal strength is just barely detectable. But for losses above about 1 dB, beware. Remember that if the total losses are 3 dB, half of your power will be used just to heat the transmission line. Additional losses over those charted in Fig 3 will occur when standing waves are present. (See Chapter 24, Transmission Lines.) The trouble is, you can't use a 50- or 75-Ω SWR instrument to measure the SWR in zip-cord line accurately.

Based on this information, we can see that a hundred feet or so of zip-cord transmission line on 80 meters might be acceptable, as might 50 feet on 40 meters. But for longer lengths and higher frequencies, the losses become appreciable.

Zip Cord Wire as the Radiator

For years, amateurs have been using ordinary copper house wire as the radiator section of an antenna, erecting it without bothering to strip the plastic insulation. Other than the loading effects of the insulation mentioned earlier, no noticeable change in performance has been noted with the insulation present. And the insulation does offer a measure of protection against the weather. These same statements can be applied to single conductors of zip cord.

The situation in a radiating wire covered with insulation is not quite the same as in two parallel conductors, where there may be a leaky dielectric path between the two conductors. In the parallel line, it is the current leakage that contributes to line losses. This leakage current is set up by the voltage potential that exists on the two adjacent wires. The current flowing through the insulation on a single radiating wire is quite small by comparison, and so as a radiator the efficiency is high.

In short, communications can certainly be established with a zip-cord antenna in a pinch on 160, 80, 40, 30 and perhaps 20 meters. For higher frequencies, especially with long line lengths for the feeder, the efficiency of the system is so low that its value becomes questionable.

A TREE-MOUNTED HF GROUND-PLANE ANTENNA

A tree-mounted, vertically polarized antenna may sound silly. But is it really? Perhaps engineering references do not recommend it, but such an antenna does not cost much, is inconspicuous, and it works. This idea was described by Chuck Hutchinson, K8CH, in *QST* for September 1984.

The antenna itself is simple, as shown in **Fig 4**. A piece of RG-58 cable runs to the feed point of the antenna, and is attached to a porcelain insulator. Two radial wires are soldered to the coax-line braid at this point. Another piece of wire forms the radiator. The top of the radiator section is suspended from a tree limb or other convenient support, and in turn supports the rest of the antenna.

The dimensions for the antenna are given in **Fig 5**. All three wires of the antenna are ¼ wavelength long. This generally limits the usefulness of the antenna for portable operation to 7 MHz and higher bands, as temporary supports higher than 35 or 40 feet are difficult to come by. Satisfactory operation might be had on 3.5 MHz with an inverted-L configuration of the radiator, if you can overcome the accompanying difficulty of erecting the antenna at the operating site.

The tree-mounted vertical idea can also be used for fixed station installations to make an invisible antenna. Shallow trenches can be slit for burying the coax feeder and the radial wires. The radiator itself is difficult to see unless you are standing right next to the tree.

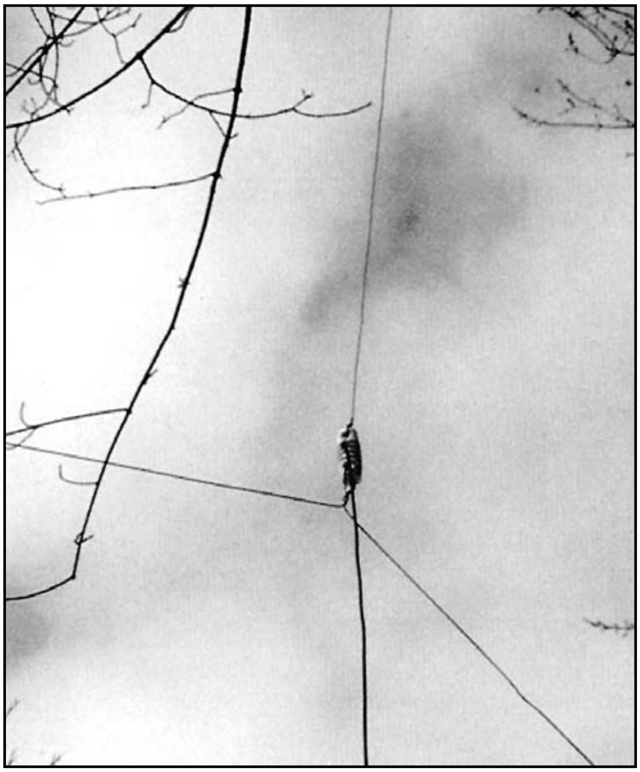

Fig 4—The feed point of the tree-mounted ground-plane antenna. The opposite ends of the two radial wires may be connected to stakes or other convenient points.

A PORTABLE DIPOLE FOR 80 TO 2 METERS

This dipole antenna, described by Robert Johns, W3JIP, in August 1998 *QST*, can be used for any band from 80 through 2 meters. One half of the dipole is an inductively loaded aluminum tube. Its length is adjustable from 4 to 11 feet, depending on how much room is available. The other half is flexible insulated wire that can be spooled out as necessary. The tube is supported by a flagpole bracket attached to a long carpenter's clamp. The clamp mounts the antenna to practically anything: a windowsill, railing, a chair or post. If there is no structure to mount the antenna (a parking lot or the beach), the clamp attaches to two light wooden legs to form a tripod, as shown in **Fig 6**.

The key to mounting flexibility is the large clamp. The key to electrical flexibility is a large adjustable coil that lets you resonate the tube on many bands. The coil is wound with #8 aluminum ground wire from RadioShack. The form is a four-inch (4$^1/_2$-inch OD) styrene drain coupling from the Home Depot or a large plumbing supply store. A movable tap adjusts the inductance to tune the upright tube to the desired band. The wire half of the antenna is always a bit less than λ/4 on each band. Hang it from any convenient support or drape it over bushes to keep it off the ground.

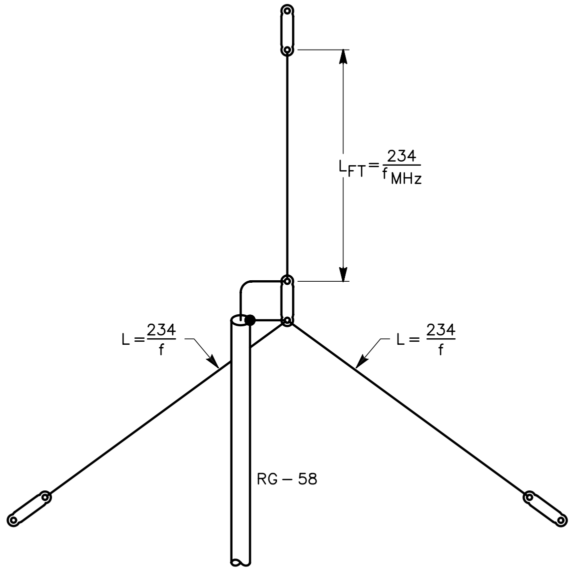

Fig 5—Dimensions and construction of the tree-mounted ground-plane antenna.

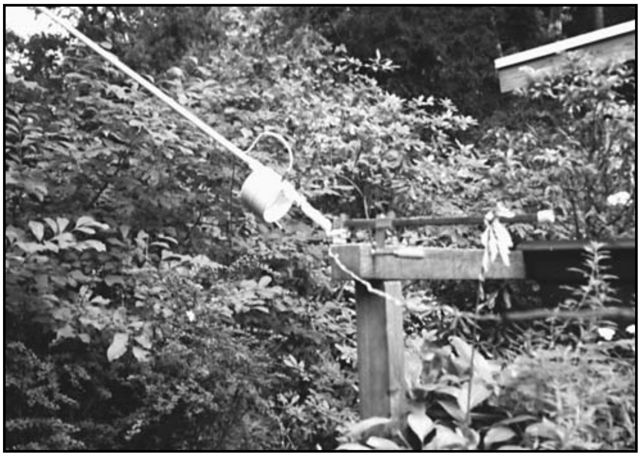

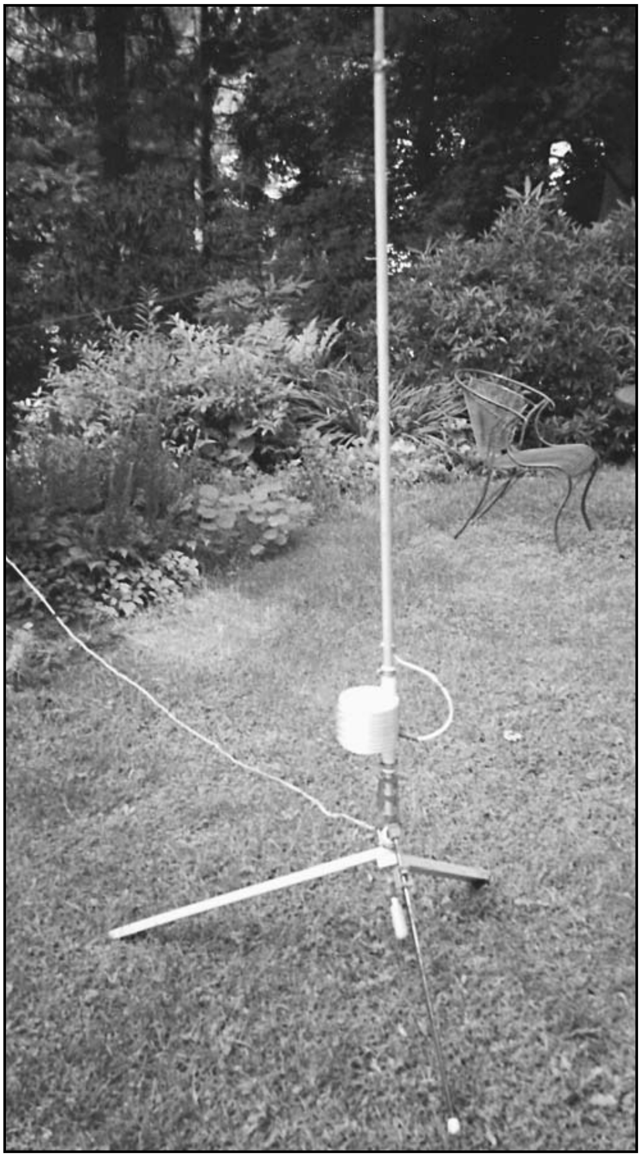

Fig 6—At top, the portable antenna in some of the many places it may be mounted around the house, porch and yard. At bottom, the simple ground-mounted legs that make a tripod.

Construction

The 18-inch carpenter's clamp (sometimes called a *bar clamp*, such as Jorgensen's No. 3718) and flagpole bracket that takes a $^3/_4$-inch pole are common hardware items. Insulate the bracket from the clamp jaw with a 1$^1/_2$-inch length of 1-inch PVC pipe (see **Fig 7**). Hammer the PVC over the end clamp jaw to make it take the shape of the jaw. Secure the flagpole bracket to PVC with a large ground clamp (for $^1/_2$ or 1-inch conduit). The ground clamp includes $^1/_4$-inch bolts; enlarge the flagpole bracket holes to accept them. Some flagpole brackets have an integral cleat; the author hammered the cleat ears flat on his.

Mount an SO-239 chassis connector on the flagpole bracket using RadioShack insulated standoffs (276-1381). The standoffs tightly press the center pin of the SO-239 against the bracket surface; no other connection is needed. Other mounting hardware may require a connection from the coax center conductor to the bracket. The spooled wire's inner end wraps around a standoff and connects to a ring terminal under a screw holding the SO-239 flange to the standoff.

The 1 × 2-inch wooden legs for the tripod are each 30 inches long. Bolt them together at one end with a 1 or 1$^1/_4$-inch-long bolt. Countersink the bolt head and nut below the surface of the wood so they don't interfere with the clamp jaws.

Aluminum Element

You can make this element from three lengths of telescoping aluminum tubing ($^3/_4$, $^5/_8$ and $^1/_2$ inch, 0.058-inch walls). The author used tubing with thinner walls for less weight and easier handling. A 45-inch-long, $^3/_4$-inch tube fits the flagpole bracket. He chose this length because it and the flagpole bracket make $\lambda/4$ on 6 meters. The two outer tubes are both $^5/_8$ inch, made by cutting a seven-foot aluminum clothesline pole in halves. They are

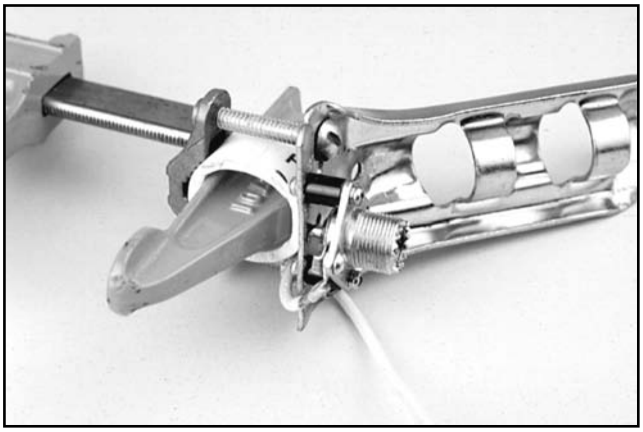

Fig 7—The flagpole bracket that supports the tubing elements is clamped to the long carpenter's clamp, but insulated from it by a small section of 1-inch PVC pipe. A coax connector is mounted to the bracket and the spool of wire is attached to the coax connector.

joined together with a copper coupling ($^5/_8$ inch ID) for $^1/_2$-inch copper pipe. The coupling is bolted to one of the $^5/_8$-inch sections, and a slot is cut in the free half of the coupling. The remaining $^5/_8$-inch tube is inserted there and secured with a hose clamp. See **Fig 8**.

One $^5/_8$-inch tube slides into the $^3/_4$-inch tube to provide a continuously variable element length from about 4 to 7$^1/_2$ feet. This extends from 7$^1/_2$ to 11 feet when the two $^5/_8$-inch sections are joined together.

Loading Coil

The loading coil has 12 turns of bare aluminum wire spaced to fill the 3$^1/_2$-inch length of the drain coupling. Drill $^9/_{64}$-inch holes at the ends of the coil form to accept the ends of the coil wire. (See **Fig 9**.) To wind the coil, feed four inches of wire through the form, make a sharp bend in the wire and start winding away from the nearby mounting hole. Wind 12 turns on the form, spacing them only approximately. Cut the wire for a 4-inch lead and feed it through the other hole in the form. Tighten the wire as best you can and bend it into another acute angle where it passes into the form. Space the turns about equally, but don’t fuss with them. Final spacing will be set after the wire is tightened.

Tighten the coil wire by putting a screwdriver or a needle-nose pliers jaw under one turn, and pry the wire up and away from the surface of the coil form. While this can be done anywhere, it’s best to put these kinks on the backside, away from the mounting shaft. Put kinks in every other turn, removing any slack from the coil and holding the turns in place. Should the coil ever loosen, simply retighten it with a screwdriver. If you prefer, glue the coil turns in place with epoxy or coil dope. Use a thin bead of glue that won’t interfere with the clip that connects to the coil.

The coil form mounts on a nine-inch-long $^3/_4$-inch PVC pipe. (See Fig 9 and **Fig 10**.) The inside diameter is slightly larger than the $^3/_4$-inch aluminum, but slotting the PVC and tightening it with a hose clamp secures the tube. (Use a wide saw to cut these slots, not a hacksaw.) The coil form mounts to the PVC pipe with #6-32 × 1$^1/_2$-inch bolts. A five-inch long, $^3/_4$-inch aluminum tube permanently attaches to one end of the coil assembly and slides into the flagpole bracket. One end of the coil wire connects to this short tube. Flatten the wire end by hammering it on something hard, then drill a $^9/_{64}$-inch hole in the flattened end and attach it to the short tube with a #6-32 × 1-inch bolt. Tighten the bolt until the $^1/_2$-inch tube starts to flatten. This will keep pressure on the aluminum-to-aluminum joint.

The aluminum element slides into the opposite end of the coil assembly. The hose clamp there can be tightened until the element just slides in snugly.

A 12-inch clip lead connects the aluminum element to the coil. Bolt the plain wire end to the $^3/_4$-inch tube three inches from its end. Many alligator clips will fit in the space between the turns of the coil (about $^3/_{16}$ inch), but W3JIP preferred using a solid-copper clip (Mueller TC-1). Cut off most of the jaws, so that only the part close to the hinge grabs the coil. This shorter lever grips very tightly.

Wire and Spool

The lower half of the antenna is insulated wire that is about λ/4 on the band of operation. The wire is pulled from the spool, and the remaining wire forms an inductance that doesn’t add much to the antenna length. The Home Depot sells #12 and #14 insulated stranded copper wire in 50 and 100-foot lengths, on plastic spools. (See **Fig 11**.) A $^1/_2$-inch dowel fits into the spool to make a

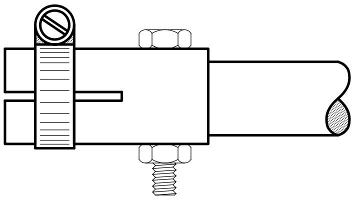

Fig 8—The joint between two sections of $^5/_8$-inch tubing is made from a $^1/_2$-inch copper pipe coupling, bolted to one section and hose clamped to the other.

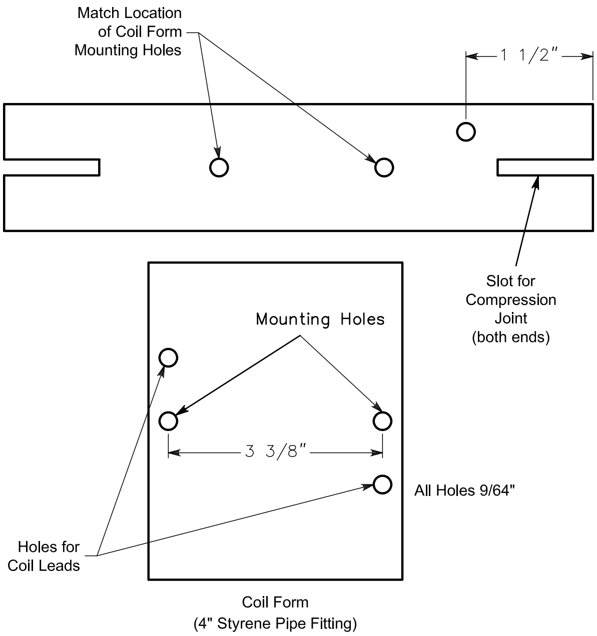

Fig 9—Holes to be drilled in the styrene coupling and the $^3/_4$-inch PVC pipe. All holes are $^9/_{64}$-inch diameter, to provide clearance for #6-32 bolts. The hole 1$^1/_2$ inches from one end holds a bolt that serves as a stop, so that the antenna tube does not slide in too far. Space the holes for coil leads far enough from the mounting holes to clear the $^3/_4$-inch pipe.

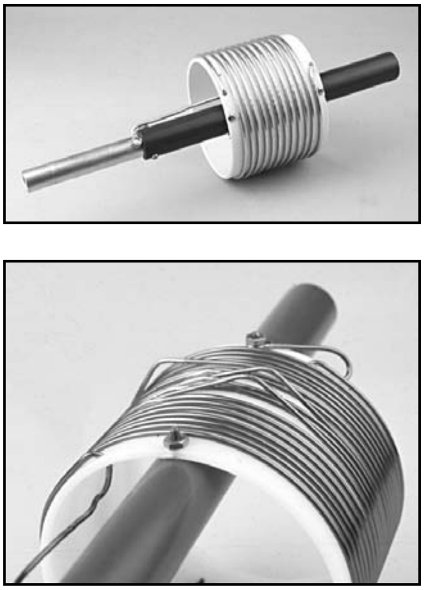

Fig 10—Top photo shows the loading coil for the 20, 30, and 40-meter band coverage. The short aluminum tube on the coil slides into the flagpole bracket, and the tubing element slides into the other end of the PVC pipe. The wire and clip connect the element to the coil. Bottom photo shows the coil for operating the antenna on 80 meters. This is placed in the flagpole bracket and the 40-meter aluminum coil plus the tubing element is inserted into it.

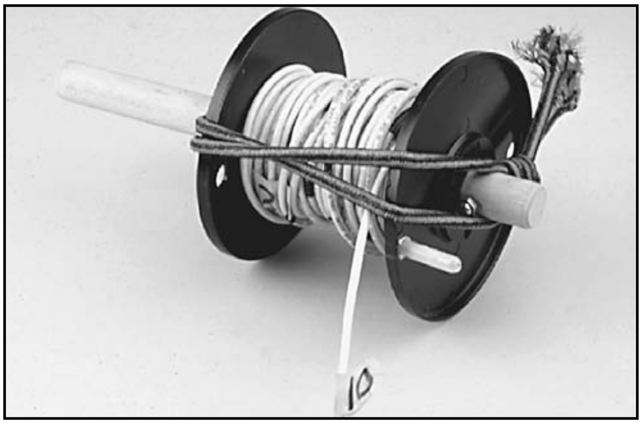

Fig 11—The wire spool has a wooden axle/handle and small handle for winding the wire. A bungee cord stretched over the spool and around the axle prevents the wire from unwinding.

Table 2

	Tubing Length (feet)	*Coil Turns*	*Wire Length (feet)*
6 meters	4	0	4
10 meters	7	0	8
	4	1.7	7
12 meters	8	0	8.3
	4	1.8	9
15 meters	10	0	10.3
	8	1.8	10
	4	3.5	10
17 meters	11	0	13
	8	2.2	13
	4	4.2	13
20 meters	11	1	16
	8	4	16
	4	6.2	16
30 meters	11	4	24
	8	7	24
	4	10.5	24
40 meters	11	9	32
	8	12	32

handle and spool axle. Bolts through the dowel on either side of the spool hold it in place. A crank handle is made by putting a one-inch-long bolt through the spool flange.

The author calibrates the wire on the spool with markers and electrical-tape flags. There's a mark (from a permanent marking pen) at each foot, a black flag every five feet and a length-marked colored flag every 10 feet. Simply mark the length for each band, if you like.

Make sure you prevent the wire from unspooling, especially when it's hanging from a window mounting. A heavy rubber band works, but it doesn't last long. A better solution is a loop of light bungee cord, preferably with a knot for grip. The bungee loop runs from the axle/handle around the spool making a half twist on the way, and then passes over the axle end on the other side of the spool. (See Fig 11.)

Operation

Table 2 lists element length, wire length and coil tap point for various bands. When the number of turns shown is zero, the coil is not needed. On all bands except 6 meters, you can simply bypass the coil with the clip lead—the extra length just lowers the frequency a bit. For 6 meters, the coil *must* be removed. The location of the unspooled wire greatly affects the settings, so these numbers are only starting points. The lengths in the table were taken with the wire one to three feet above ground, draped over and through bushes and flowerbeds. The antenna will still work if the wire is lying on the ground, but it will require less unspooled wire to resonate. A balun is not needed, and an SWR analyzer is very helpful while adjusting this antenna.

The SWR is less than 1.5:1 on all bands, and it's usually below 1.2:1. Occasionally, a band shows a higher SWR (still less than 2:1), but that can always be lowered by adjusting the length or location of the lower wire. Never set up the antenna where it could fall and injure someone, or where the unwary could get an RF burn by touching it.

The author's results with this antenna have been excellent, both from home and on vacation. If you haven't yet operated from a seashore location, be prepared for a pleasant surprise! The good ground afforded by the salt water really makes a difference.

80 Meters

It's easy to add this lower frequency band. Fig 10 shows a 35 µH coil for 80 meters. It's constructed and tightened just like the 40-meter coil, but has 20 turns of #12 magnet wire.

To operate on 75/80 meters, insert the new coil into the flagpole bracket and plug the 40-meter coil into it. Tune across the band with the movable tap on the 40-meter coil. This varies antenna resonance from below 3.5 to above 4.0 MHz, with the full 11 feet of tubing extended. If your version doesn't achieve this tuning range, adjust the spacing of the turns on the 80-meter coil.

The 80-meter coil has a five-inch length of $^3/_4$-inch aluminum tube inserted into one end of the $^1/_2$-inch PVC pipe that supports the coil form. One end of the coil is connected to this aluminum tube. The other end is secured under the bolt that holds the coil form to the PVC pipe. A second clip lead connects the base of the 40-meter coil to the outer end of the larger coil. The length of wire on the spool must also be increased to about 64 feet.

2 Meters

A $\lambda/2$ dipole for 2 meters can be made with about 15 inches of $^1/_2$-inch aluminum tubing in the flagpole bracket, and 18 inches of wire. The tubing element is shorter than normal for 2 meters because the bracket is also part of the antenna. You can also shorten the 6-meter wire a bit and operate the 6-meter antenna as a $3\lambda/2$ on 2 meters, with a somewhat higher SWR.

Continuous Coverage

With easily changed element lengths and a continuously variable loading coil, you may operate the antenna on any frequency from 6.5 to 60 MHz, if coverage for other services is needed. With taps in the 80-meter coil at 8, 11 and 14 turns, the antenna will also tune from 4 to 7 MHz.

THE HF CABOVER ANTENNA

If you have ever had the pleasure of traveling across the country with an HF mobile in a camper, trailer or motor home you may want to duplicate this antenna for use when you park. This antenna was described in *The ARRL Antenna Compendium, Vol* 5, by Jim Ford, N6JF.

The author's camper has limited spots on which to mount an HF mobile antenna. The back ladder is a convenient place to mount a small mobile whip. However, the efficiency of typical mobile center-loaded antennas, depending on coil Q and other assumptions, is often less than 2% for 80 meters and 10% for 40 meters. (These numbers come from the excellent, easy-to-use *MOBILE* antenna design program by Leon Braskamp, AA6GL, which is on the CD-ROM bundled with this book.)

At some locations, N6JF used an 80-meter dipole, which was very efficient and worked great on all bands when fed with open-wire line and an antenna tuner. However, it took over 40 minutes to set up and about 20 minutes to tear down, working with a sling shot and many tree snags. This is too much time to make a schedule or for an early morning departure, although it's OK if you plan to stay for a while. Even more important, there were often too many trees or other barriers (perhaps some even social) to allow putting up the dipole at a campsite.

When this happened he was stuck with the mobile antenna with poor efficiency. There had to be a better antenna for camper operation. The author decided on a large vertical.

A telescoping aluminum extension pole used for roller painting would make a good bottom section for the loaded vertical. These are available at many local home supply centers. The author's was 1 inch in diameter and 6 feet long, telescoping to almost 12 feet. He disassembled both sections and cross-cut a 1-inch slot in the top of the bottom section with a hacksaw to allow compression clamping with a hose clamp. The tip of the top section was fitted with an aluminum plug that had a 3/8-24 hole tapped in it. This procedure was a simple machine-shop operation. The plug was pounded into the top section and is quite snug. He tapped some set screws through the pole into the plug, however, just to be sure. An insulated, lay-down marine antenna mount fit perfectly into the bottom of the aluminum base and was secured with a bolt that also served as the electrical connection from the capacitor matching box. See **Fig 12** showing the aluminum base plate, the laydown mount and the antenna

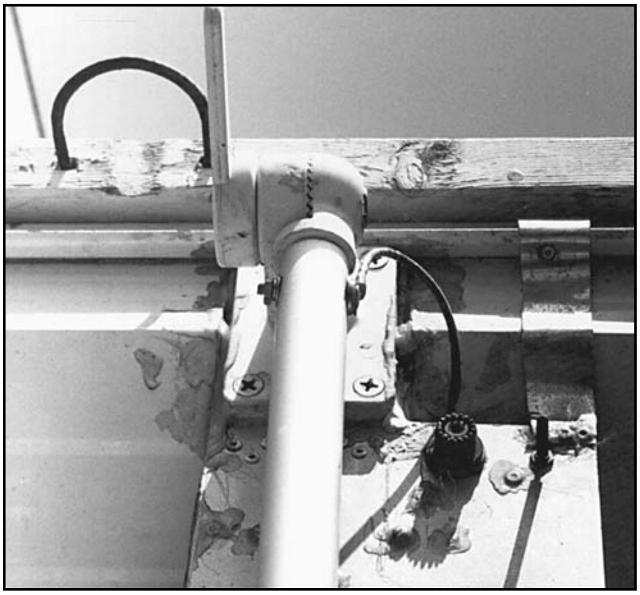

Fig 12—Photo of aluminum base plate, showing details of mounting to the camper. The four banana-plug jacks on the bottom are for extra radials, if desired.

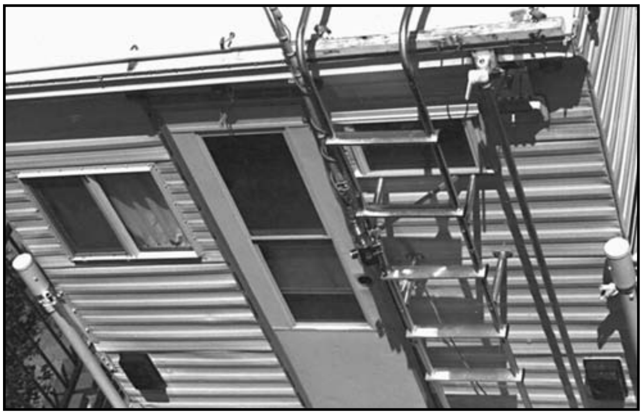

Fig 13—Photo showing the back of the camper, with the antenna in the lowered position, parallel to the ladder. The "Outback" standby mobile antenna is shown clamped to the left side of the ladder.

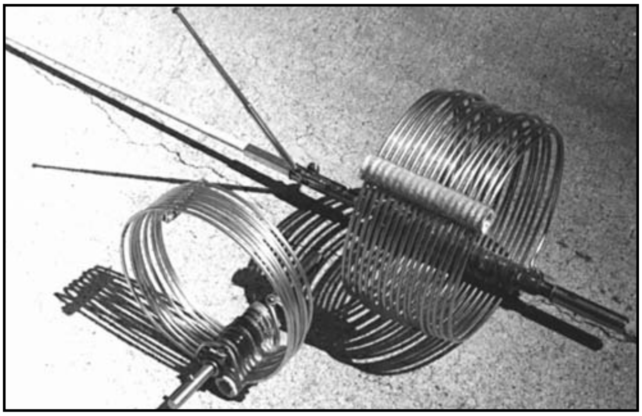

Fig 15—Close-up photo of 80 and 40-meter coils, with top section and telescoping whip antenna with swivel bracket for tuning the top section for the higher bands. Note quick-disconnect connectors at top and bottom of both coils. The top whip is a Fiberglas CB whip, used on 80 and 40 meters.

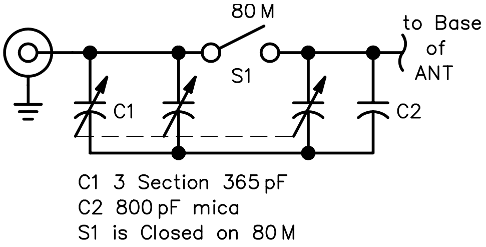

Fig 14—Schematic of tuning capacitor at base plate. C2 is an 800 pF transmitting mica capacitor. C1 is a three-section 365 pF broadcast tuning capacitor. S1 is closed for 80-meter operation.

mast itself lowered down the back of the camper. **Fig 13** shows the layout of the back of the camper, with the antenna on the right-hand side, laid down for travel.

The variable capacitor in the matching box is a surplus three-section 365-pF broadcast tuning capacitor. Two of three sections are connected in parallel and a switch parallels in the third section, along with an extra 800 pF mica capacitor for use on 80 meters. See the schematic in **Fig 14**.

The capacitor assembly was put in a custom-glued Plexiglas box to keep out the weather and mounted to a piece of plate aluminum, along with an SO-239 connector. The aluminum plate was riveted to the camper shell using a lot of aluminum rivets. Do not use steel. N6JF peeled back about a 4-inch wide section of the side of the camper for this direct aluminum-to-aluminum connection. People who are hesitant to modify their campers like this need to find an alternate low-resistance connection method. His camper was old enough not to be an issue.

For an extra low-resistance connection a $1\frac{1}{2}$-inch aluminum strip was added from the top of the base plate to the camper, as shown in Fig 12 near the 80-meter switch. The tuning knob protrudes from the side of the Plexiglas box. The four bottom holes in the plate are for banana plugs to connect ground radials if extra efficiency is desired. However, the roof of a camper is one of the better places for a mobile antenna, so the author seldom hooks up the radials. When he does use them, the tuning changes only a little.

To keep losses down, N6JF used coils wound with aluminum clothes-line wire on old mobile coil center sections with quick disconnect fittings. See **Fig 15**, which shows both the 80 and 40 meter coils, together with the top portion of the antenna. An article by Robert Johns, W3JIP, in October 1992 *QST* described techniques for building your own loading coils. The coils ended with a little more inductance than calculated and the author had to remove some turns. Both coils are spaced at 4 turns per inch. The 8-inch long 80-meter coil has 18 turns. The 7-inch long 40-meter coil has 8 turns.

The matching network is actually an L-section step-up match, using the net inductive reactance of the antenna plus the center loading coil. The PVC coil construction technique was described in W3JIP's *QST* article and a follow-up Technical Correspondence piece in October 1992 *QST*. Basically, it consists of drilling an accurate row of slightly undersized holes along a length of $\frac{1}{2}$-inch PVC pipe and then carefully sawing down the center of the row of holes with a hack saw. Then, you trap the coil wires in the grooves between the two sawed halves and tie the two halves together with string. When you are satisfied that everything is proper, you then tighten the string and apply epoxy glue to make it strong and permanent.

One advantage of aluminum clothes line wire is that it is already coiled at the approximate diameter needed when you buy it and it is easy to position on the coil form.

The clothes line wire had a plastic coating, which wasn't removed except at the contact points. Once the epoxy dries, this method of construction does a good job of holding the finished coil together and it is lightweight.

The computed coil Q from the *MOBILE* antenna program is about 800 for the 80-meter coil and about 400 for the 40-meter coil. The author accidentally made the 40-meter coil 7 inches in diameter instead of a higher-Q 6 inch diameter. Even so, the whole antenna system with a 9-foot top section calculates as being 56% efficient on 80 meters and 85% for 40 meters.

The removable top section for 80 and 40 meters is a full-size fiberglass CB whip from RadioShack. The fiberglass whip has about a #16 hole in the center of it. Be sure to sand and paint the whip for protection against UV and to protect yourself against fiberglass spurs in the hands. N6JF tried a full-size stainless steel CB whip to get a slightly higher capacitance to ground because of larger conductor size but discovered it was far too heavy. That experience reinforced his decision that aluminum was a far more practical coil and base section material for this project.

Quick-disconnect connectors found years ago at a hamfest were used for both coil forms and for the top section. Bands higher than 40 meters don't need any loading coils and the antenna length can be telescoped to get a $^1/_4$ wavelength. Ten meters doesn't require any top section. Be sure to use some NOALOX or similar compound to prevent corrosion and poor connections at all aluminum joints. This is especially true for the telescoping sections and the aluminum rivets. The matching capacitor is in the circuit at all times but when the 80-meter switch is off you can set the capacitor at minimum (about 14 pF) and it is effectively out of the circuit, even at 10 meters. The author has not tried this antenna on power levels greater than about 100 W but the weakest link would probably be the matching capacitor. The voltage at the matching capacitor is low, so 200 W should be no problem.

You can achieve a 1:1 SWR match for 80 or 40 meters and a good SWR is obtained without retuning the base matching capacitor for approximately 100 kHz on 80 meters and most of 40 meters. The top section, however, does not have such low Q and needs to be tuned. The 2:1 SWR bandwidth on 80 meters is about 25 kHz and 150 kHz for 40 meters. Tuning is accomplished by using a telescoping FM portable radio antenna connected to the top-section whip with a stainless steel hose clamp. The maximum length of the telescoping section used was 29 inches, and it collapses down to 7 inches. A whip with an elbow was used to adjust the angle of the whip as well. A telescoping whip half the length would still be long enough for the full adjustment of both bands.

Adjustment of the top section is one of the penalties paid to achieve high efficiency for operation on 80 and 40 meters. Substituting an automatic antenna tuner would likely lose efficiency, particularly on 80 meters since the base resistance calculates to about 10 Ω. N6JF has not tried to make the antenna work on 160 meters. The SWR for the higher bands was good enough without any matching network.

As light as the antenna is, it still won't hold up in a moderate wind without some support guys. N6JF used $^3/_4$-inch and $^1/_2$-inch PVC support pipes in a telescoping arrangement for storage, but expanding to give an approximate 45° support at the top of the bottom section from both directions. One end of the telescoping section was connected to the camper roof with a hinge. The other end formed a snap-fit out of a PVC barrel that was cut lengthwise. See **Fig 16** and **Fig 17** for details. Even though it formed a good snap fit, N6JF didn't trust the joint for strong winds so he glued a piece of Velcro to the joint to close up the open end. Be careful, though, because Velcro deteriorates with exposure.

Another hose clamp near the top of the bottom section holds an $^1/_8$-inch line taken up the ladder to pull up the antenna without the need of an assistant. The disadvantage is you have to climb the roof. Use a non-slip floor mat or something similar to spread the load on the roof and to avoid slipping. Once on the roof, however, the coil is at a height for easy adjustment when the telescoping section is in the down position.

An advantage of being able to assemble the whole antenna on the roof is that you don't need a lot of swing-up room and you can clear trees easily. You can put calibration marks on the upper aluminum section for resonant lengths on the higher bands but just raise the top section up all the way for 80 or 40 meters. Mark also the small

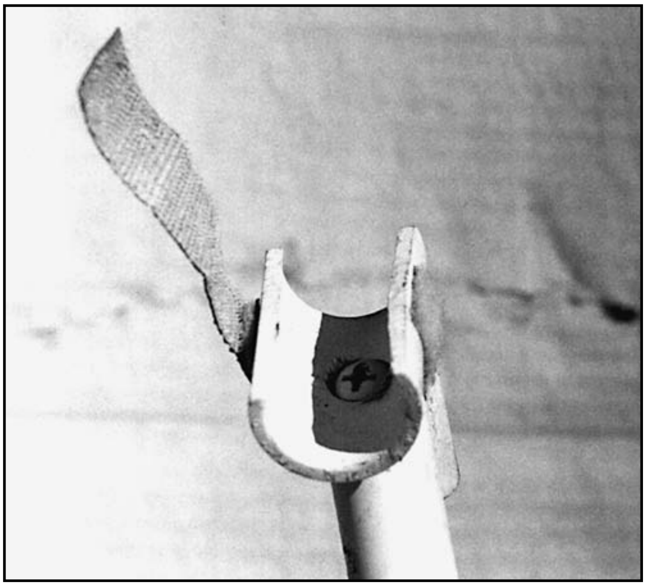

Fig 16—Close-up of one of the snap-on support brackets used to brace the antenna. Note the Velcro pieces used to ensure that the antenna doesn't pop out of the bracket in the wind.

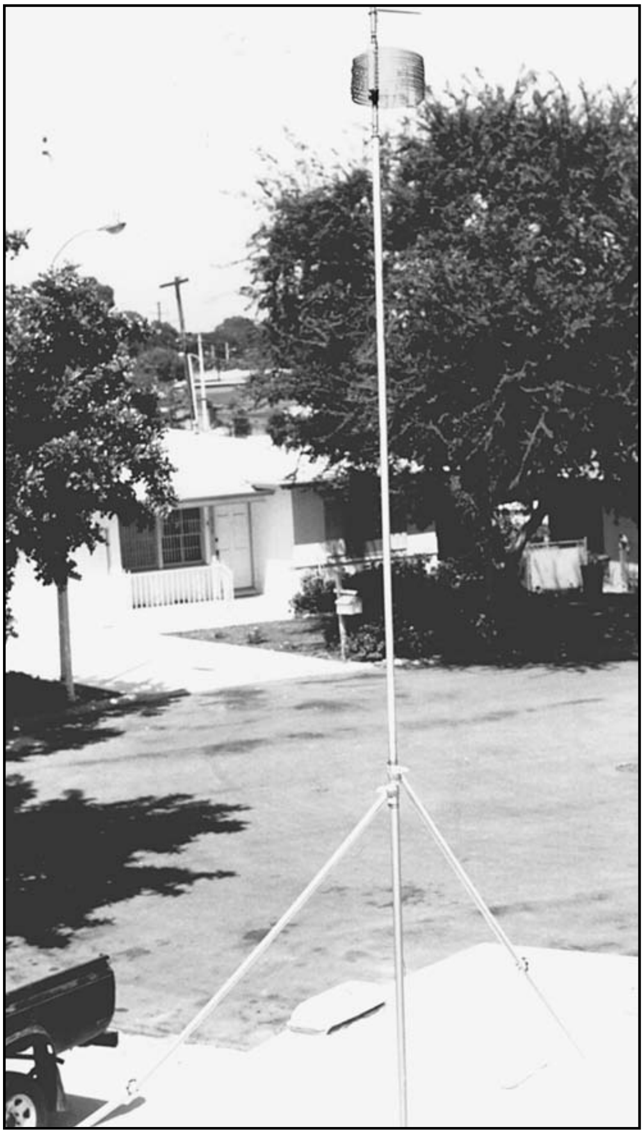

Fig 17—Photo showing the two support bracket poles bracing the bottom section of the antenna. The top tuning whip is evident above the homemade loading coil at the top of the bottom section.

coil tuning whip for 80 or 40 meters, although different locations may require slightly different settings due to detuning from nearby metal objects.

The overall length is about 21 feet. The top of the author's camper is about 10 feet high when on the truck, putting the tip at 31 feet. This antenna is definitely designed for use when you are parked at a fixed location. N6JF can put up this antenna in less than 5 minutes and can take it down in half the time. The success of this project has as much to do with knowing how and where you operate as it does paying attention to mechanical and electrical details. The antenna has been a good compromise between efficiency and convenience.

TWO PORTABLE 6-METER ANTENNAS

These antennas were described by Markus Hansen, VE7CA, in *The ARRL Antenna Compendium, Vol 5*. After years of HF operation, he became enthusiastic about VHF/UHF operation when he found a used Yaesu FT-726R VHF/UHF all-mode transceiver at a reasonable price.

But he became really enthused when he got on 6 meters and discovered the joys of driving to high mountain peaks to operate. Not only does an antenna have to be portable for this kind of operation, it must be easily assembled and disassembled, just in case you have to move quickly to a better location.

A Portable Two-Element Six-Meter Quad

VE7CA's primary objective was to construct a two-element quad using material found in any small town. It should not use a complicated matching network. The Gamma matches commonly used on quads do not hold up well when you are setting up and taking down these antennas in the field. The final design adjusted the distance between the driven and the reflector elements so that the intrinsic feed-point impedance was 50 Ω.

Fig 18 shows the dimensions for the boom and the boom-to-mast bracket. The boom is made from a $27^1/_4$-inch length of 2 × 2. (The actual dimension of 2 × 2 is closer to $1^3/_4$ by $1^3/_4$ inches but it is commonly known in lumber yards as a 2 × 2.) Use whatever material is available in your area, but lightweight wood is preferred, so clear cedar or pine is ideal. Drill the four $^1/_2$-inch holes for the spreaders with a wood bit, two at each end, through one of the faces of the 2 × 2 and the other through the other face. The boom-to-mast bracket is made from $^1/_4$-inch fir plywood.

The spreaders are $^1/_2$-inch dowel. The local lumberyard had a good supply of fir dowels but other species of wood are available. The exact material is not important. Maple is stronger but expensive. Fiberglass would be ideal but it is not always available locally. Cut two of the $^1/_2$-inch dowels to a length of $83^5/_8$ inches for the driven element spreaders and two to 88 inches for the reflector spreaders. **Fig 19** shows one end of the boom with the two spreaders inserted. The mast was made from two six-foot lengths of $1^3/_4$-inch fir dowel. Again, use whatever you may have available. Waterproof all wooden parts with at least two coats of exterior varnish.

While you are at the lumberyard or hardware store look for plastic pipe that fits over the end of the $^1/_2$-inch spreaders. You will need a one-foot length, with some to spare. Cut it into seven equal lengths, approximately one inch long, and one to a length of $1^1/_2$ inches. Drill a $^1/_{16}$-inch hole through the seven equal lengths $^1/_4$ inch from the ends, and two holes one above the other $^1/_4$-inch apart on the $1^1/_2$-inch sleeve. VE7CA used #14 hard-drawn stranded bare copper wire for the elements. Do not use insulated wire unless you are willing to experimentally

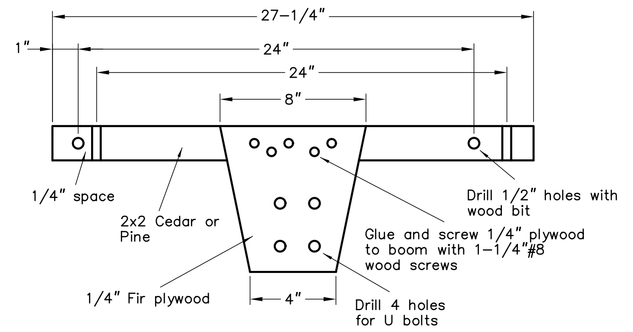
27-1/4"
1"
24"
24"
8"
1/4" space
2x2 Cedar or Pine
Drill 1/2" holes with wood bit
Glue and screw 1/4" plywood to boom with 1-1/4"#8 wood screws
1/4" Fir plywood
4"
Drill 4 holes for U bolts

Fig 18—Dimensions for the boom-to-mast bracket for VE7CA's portable two-element 6-meter quad.

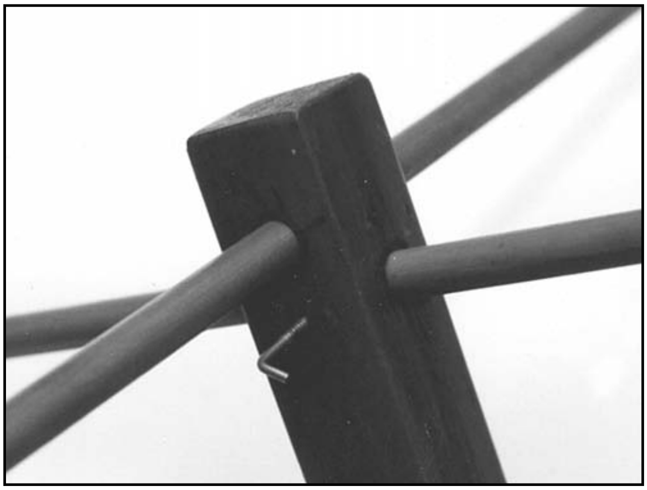

Fig 19—Photo of one end of the VE7CA quad with the two spreaders inserted.

determine the element lengths, since the insulation detunes each element slightly.

Cut the reflector element 251 inches long and slip one end of the wire through the holes you drilled in four of the plastic sleeves. Don't attempt to secure the wire to the plastic sleeves at this point. Cross the end of the reflector elements one inch from their respective ends and twist and solder together. The total circumference of the reflector element should be 249 inches when the ends are connected together.

Cut the driven element wire to 241 inches and slip three of the 1-inch sleeves onto the wire. Again, don't secure the wires to the sleeves yet. Then the ends are passed through the two holes in the $1\frac{1}{2}$-inch pipe. Wrap the ends around the pipe and twist them back onto themselves to secure the wire. The coax feed line is attached directly to the two ends at this point. The circumference of the driven-element loop from the points where the coax is attached should be $236\frac{5}{8}$ inches. Solder the coax feed line to the driven element and waterproof the coax with silicone seal. The author used RG-58, as it is lightweight. The length required for a portable installation is typically not very long, maybe 20 feet, so the loss in the small cable is not excessive. Near the feed point, coil the coax into six turns with an inside diameter of two inches. This is an effective method of choking any RF from flowing on the outside of the coax shield.

Begin assembling the quad by pushing the two reflector spreaders, without wires attached, through one end of the boom and the two shorter driven-element spreaders through the holes in the other end of the boom. Center the spreaders and mark the spreaders with a black felt-tip pen next to the boom. Now insert a $1\frac{1}{2}$ #8 wood screw or a threaded L-hook into the boom so that it just touches one of the spreaders. Take the screw or L-hook out and file the end flat, then reinsert it so that it is just snug against the spreaders. The author only used two L-hooks for the two vertical spreaders; the horizontal spreaders are held in the proper position by the tension of the wire loops. If you use an L-hook, you can unscrew it with your hands—you won't have to worry about leaving the screwdriver at home.

You are now ready to assemble the wire loops. Take the reflector loop and place the four plastic caps over the ends of the reflector spreaders. Equalize the wire lengths between the spreader so that the loop is square. Now, secure the plastic sleeve pipes by tightly wrapping wire around the sleeve and the wire element and soldering the wire in place. See **Fig 20**, a photo showing one of the plastic sleeves slipped over one of the spreader ends, with the wire element through the hole and fastened in place. Follow the same procedure with the driven element.

Fig 21 shows the quad's boom, with the plywood boom-to-mast bracket fastened with wood screws and glue. Two U-bolts are used to attach to the mast. When the quad is raised, the shape of the loop is commonly known as a diamond configuration. The mast consists of two six-foot lengths of doweling joined together with a two-foot length of PVC plastic pipe, held together with wood screws.

Make a slot the width of a #8 wood screw about one inch deep from the top of the plastic PVC pipe and then put the top mast into the plastic pipe. Insert a one inch #8 wood screw into the bottom of the slot you cut into the top of the pipe and tighten only enough so that the top mast can be removed without unscrewing it. VE7CA drove a nail into the end of the lower mast and left it exposed an inch or more. This end is placed in the ground and the nail holds the pole in place. A strip of wood approximately 1 × 3 and long enough to cross over the roof rack of the family van is used to hold the center of the antenna mast to the roof rack of the van with small diameter rope. See **Fig 22** for a photo of the quad in action next to the family van.

To disassemble the quad, lay it on its side, slip the plastic sleeves off the ends of the spreaders and roll up

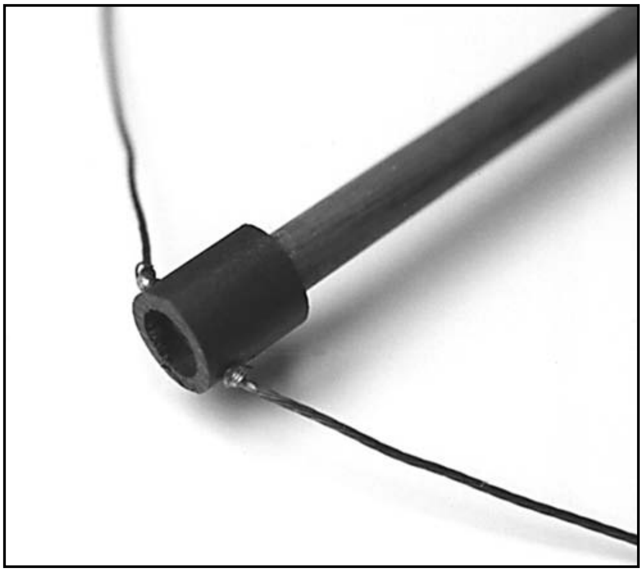

Fig 20—Photo showing one of the plastic sleeves slipped over end of a spreader to provide a mechanical mounting point and support for the wires.

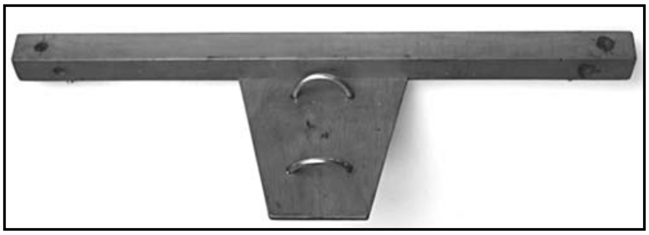

Fig 21—Photo of the two-element quad's boom, with the plywood boom-to-mast bracket secured with wood screws and glue.

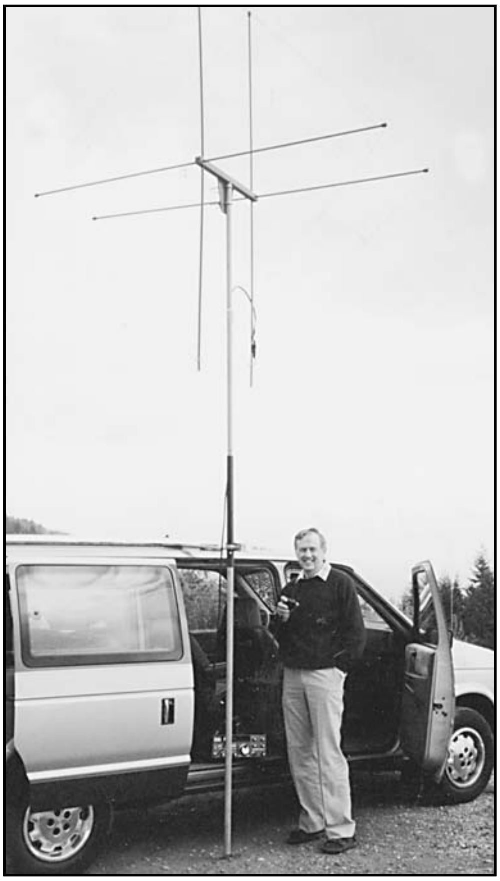

Fig 22—Ready for action! VE7CA has set up his quad next to the family van.

the wire loops. Loosen the L-hooks holding the vertical spreaders in place. Push the spreaders out of the boom, loosen the U-bolts and free the mast from the boom.

That is all there is to it. It takes about two minutes to put it up, or take it down. It is quite sturdy and has survived several high-wind storms.

A Three-Element Portable 6-Meter Yagi

The idea to build a Yagi antenna resulted when the author traded the family van for a compact car. He needed something that would fit into the trunk of the car. At close to 7-feet long, the quad spreaders were too long. Computer modeling showed that a three-element Yagi on a five-foot boom also could pick up about 1.5 dB gain over the short-boom two-element quad. A five-foot boom fits into the trunk or across the back seat of the car, but something had to be done about the nine-foot elements!

One day VE7CA noticed a box of portable-radio telescoping antenna elements at the radio parts store. They were 54 inches long when fully extended. He next found a 60-inch length of aluminum tubing that fit over the end of the telescoping elements. There are many different sizes of telescoping antenna elements, with different diameters. This is where you will have to use your scrounging skills! **Fig 23** shows how the tubing is used as a center section to join two telescoping elements together. It also serves to extend the total length of each element, since two telescoping elements themselves are not long enough to resonate on six meters. See **Table 3** for dimensions and element spacings. Each center section is slotted at both ends with a hacksaw, and stainless-steel hose clamps are used to secure the telescoping elements.

Fig 24 shows the center sections of the three elements with their mounting brackets. A square boom was used to obtain a flat surface to work with. **Fig 25** shows how the reflector is attached to the end of the boom with two $1^1/_2$-inch 10-32 bolts and wingnuts. **Fig 26A** provides the dimensions and details for the reflector and director element-to-boom brackets, which are formed from $^1/_{16}$-inch plate aluminum. The driven element is split in the center and is insulated from the boom. Fig 26B shows details for the driven-element bracket. **Fig 27** is a photo

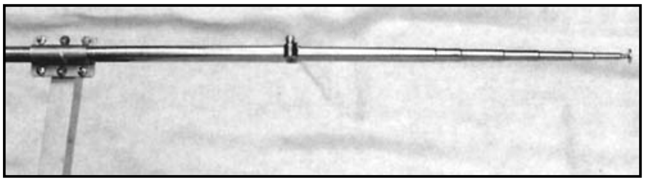

Fig 23—Photo showing a piece of aluminum tubing used as a center section to join the two telescoping tips together.

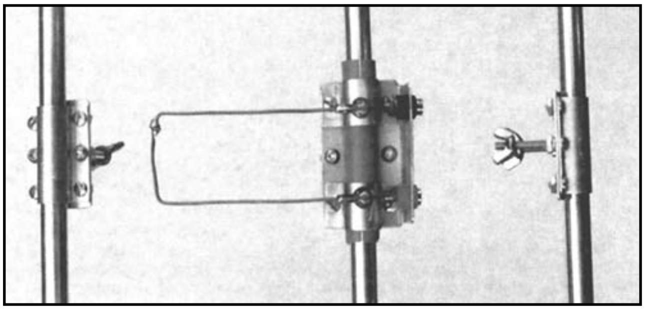

Fig 24—A view of the center sections of the three Yagi elements with their mounting brackets.

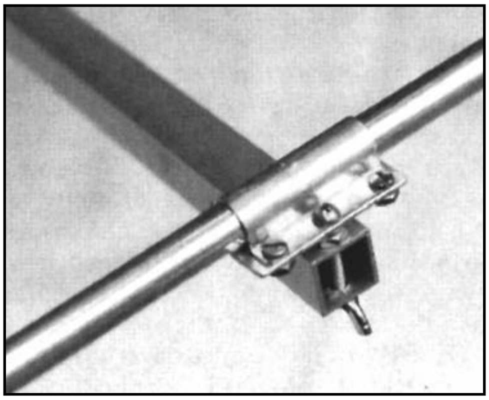

Fig 25—Photo detailing attachment of the reflector to the square-section boom, using two #10 bolts and wingnuts.

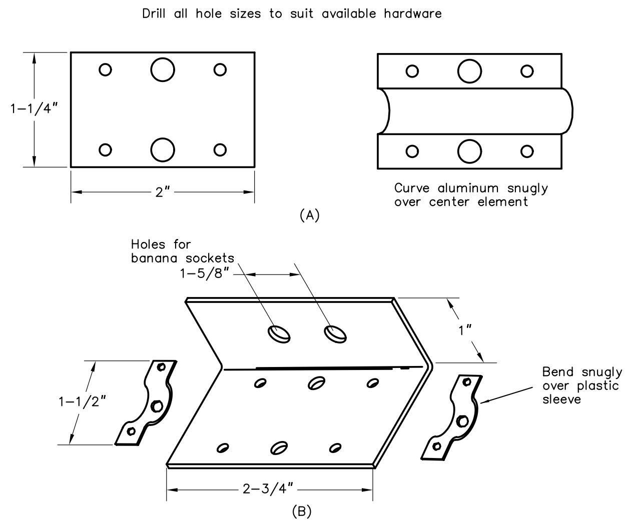
Drill all hole sizes to suit available hardware
1-1/4"
2"
Curve aluminum snugly over center element
(A)
Holes for banana sockets
1-5/8"
1"
1-1/2"
Bend snugly over plastic sleeve
2-3/4"
(B)

Fig 26—At A, details for the reflector and director element-to-boom brackets, made of $^1/_{16}$-inch plate aluminum. At B, details for the driven-element bracket. These are screwed to the square boom.

of the driven element with the hairpin matching wire and the banana plugs used to connect the coax to the driven element. You could use a female PL-259 connector if you wish. VE7CA used #14 solid bare copper wire for the hairpin. It is very durable—even after being severely warped in the car trunk, it can be bent back into shape quickly and easily.

The boom is $^3/_4$-inch square aluminum, 65 inches long. The material was found at a local hardware store. To detach the elements, just loosen the wing nuts and remove the elements from the boom. A similar method was used to attach the support mast to the boom.

As with the quad, a choke balun was used, consisting of a coil of 6 turns of coax with an inside diameter of 2 inches. To tune the hairpin match, assemble the Yagi on its mast and extend the elements. Spray switch contact solution on a cloth and wipe any dirt and grease from the elements. Push the elements together and apart a

Table 3

Three-Element Yagi, Element Lengths and Spacing Along the Boom, and Hairpin Dimensions

Element Along Boom	Spacing Section Ele. (inches)	Center Length (inches)	Telescoping (inches)	Total Length (inches)
Reflector	0	$22^{3}/_{4}$	$51^{1}/_{2}$	$125^{3}/_{4}$
Driven	28	$9^{3}/_{4}$ *	$48^{5}/_{8}$	$58^{3}/_{16}$
Director	$63^{3}/_{8}$	$14^{1}/_{2}$	$51^{1}/_{4}$	117
Hairpin	#14 wire	4 long	$1^{5}/_{8}$ spacing	

* Driven-element uses 2 sections insulated at center

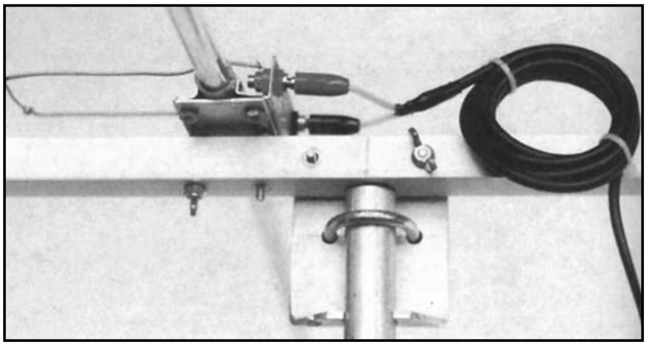

Fig 27—Photo of the driven element, complete with hairpin match and the banana plugs used to connect the coax cable to the driven element.

couple of times so that the contact solution cleans the elements thoroughly. Attach the antenna mast to your vehicle or use whatever method of support you intend to use in the field. Connect an SWR meter and a transmitter to the coax feeding the antenna. VE7CA used two alligator clips soldered together to slide along the two hairpin wires to find the position for the lowest SWR. The dimensions computed by computer were correct! The SWR was below 1.16:1 from 50.05 to 50.2 MHz.

You can take this antenna out of the trunk of the car and assemble it in less than two minutes. One caution: the telescoping elements when fully extended are quite fragile. VE7CA has not broken one as yet, but carrying a spare element just in case would be a good idea.

VE7CA 2-Element Portable HF Triband Yagis

These portable HF wire Yagis were described by Markus Hansen, VE7CA, in November 2001 *QST* and in *The ARRL Antenna Compendium, Vol 7*.

A 20/15/10-Meter Triband 2-Element Yagi

VE7CA wanted a simple 2-element wire Yagi for 20/15/10 meters that could be stored in the ski boot of his car. The basic concept comprises three individual dipole driven elements, one each for 20, 15 and for 10 meters tied to a common feed point, plus three separate reflector elements. The elements are strung between two 2.13-meter (7-foot) long, 2 × 2-inch wood spreaders, each just long enough to fit into the ski boot of his car.

By playing with the reflector-to-driven element spacing and the initial driven-element lengths, VE7CA was able to come up with a feed-point impedance on each band that allowed the use of a single setting for the shorting bar on a hairpin match. The result was a very acceptable match over the lower portions of each band. The layout of the 20/15/10-meter triband wire Yagi is shown in **Fig 28**, with the dimensions provided in **Table 4**.

The dimensions shown in Table 4 are what resulted after tuning for the lowest SWR in the middle of the lower portion of each band. VE7CA set his system up by hanging one end of the antenna from a tree and sloped it downwards at 45º, tying the lower end to a peg in the ground. The height at the feed point was 30 feet.

The feed-point impedance of an antenna is affected by many environmental factors. The presence of a reflector relatively close to the driven element has a major effect, since the impedance at the driven element in a Yagi is affected by the tuning of the driven element itself, by the spacing and length of the reflector element and to a lesser extent the height of the antenna above ground and the character of the soil itself. The real challenge in a multiband Yagi with a single feed line is to obtain a low SWR on all the bands.

The hairpin match is one of the easiest matching systems to make. It is easy to adjust and since wire is the only ingredient, it can be coiled up with the rest of the antenna when the antenna is disassembled. The feed-point impedance of the Yagi with a reflector element spaced 0.1 λ behind the driven element typically produces a resistance around 20 Ω. By shortening the driven element from its resonant length, capacitive reactance is added to the feed-point resistance. This can be cancelled by shunting the feed point with an inductor in the shape of a wire loop resembling a *hairpin*. This causes a step up of the 20-Ω feed-point resistance to 50 Ω.

Table 4
Dimensions for 20/15/10-Meter Tribander

Frequency MHz	*Spacing DE to Refl cm (feet)*	*Driven Element Half Length cm (feet)*	*Reflector Length cm (feet)*	*Hairpin Length cm (inches)*
14.1	213 (6.99)	488 (16.01)	1065 (34.94)	43 (14.9)
21.1	175 (5.74)	335 (10.99)	708 (23.23)	
28.25	125 (4.1)	254 (8.33)	531 (17.42)	

Spacing between hairpin wires is 10 cm (4 inches).

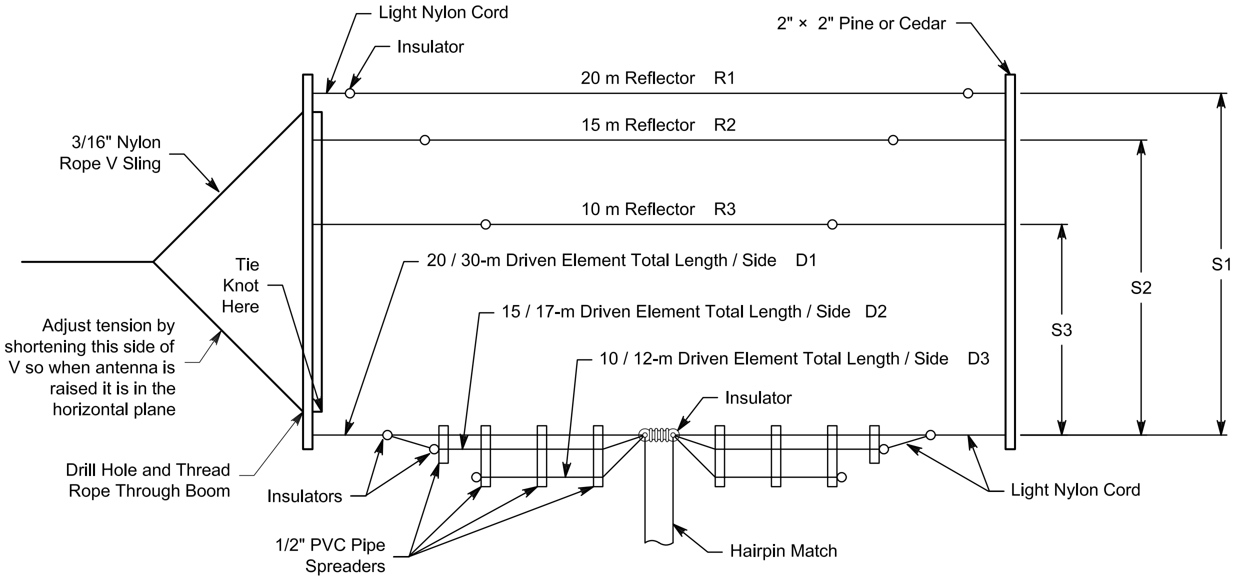
Light Nylon Cord
Insulator
2" × 2" Pine or Cedar
20 m Reflector R1
15 m Reflector R2
10 m Reflector R3
3/16" Nylon Rope V Sling
Tie Knot Here
20 / 30-m Driven Element Total Length / Side D1
15 / 17-m Driven Element Total Length / Side D2
10 / 12-m Driven Element Total Length / Side D3
Adjust tension by shortening this side of V so when antenna is raised it is in the horizontal plane
S1
S2
S3
Insulator
Drill Hole and Thread Rope Through Boom
Insulators
Light Nylon Cord
1/2" PVC Pipe Spreaders
Hairpin Match

Fig 28—Dimensions for VE7CA's 2-element 20/15/10-meter triband Yagi.

The Balun

Fig 29 shows the hairpin match and the common-mode choke balun for the 10/15/20-meter triband wire Yagi. You should let the coax drop straight down from the center insulator and attach it to the center of the hairpin shorting bar. Continue by making a coil using the coax of 8 turns, with a diameter of about 4 inches. This balun will choke off RF flowing along the outside of the coax shield that would otherwise distort the radiation pattern of the antenna. The center of the shorting bar is at a neutral potential, so there is no harm in attaching the coax feed line at that point.

Flattop or Tilted?

If DX is your main interest, then you want to tilt the antenna towards the vertical to emphasize the lower elevation angles. Remember that the radiation pattern is quite dependent on ground conductivity and dielectric constant for a vertically polarized antenna. A location close to saltwater will yield the highest gain and the lowest radiation angle. With very poor soil in the near and far field, the peak radiation angle will be higher and the gain less.

VE7CA has had numerous opportunities to test this out while operating portable at his favorite location. Using two trees as supports he can pull the antenna to horizontal with the feed line about 7 meters above the ground. In this position, with 20 meters open to Europe, he found it difficult to work DX on CW with 3 W of output power. However, when he changed the slope of the antenna so that it is nearly vertical he not only heard more DX stations, but he found it relatively easy to work DX. The sloping antenna always performs much better for working DX than a low horizontal antenna.

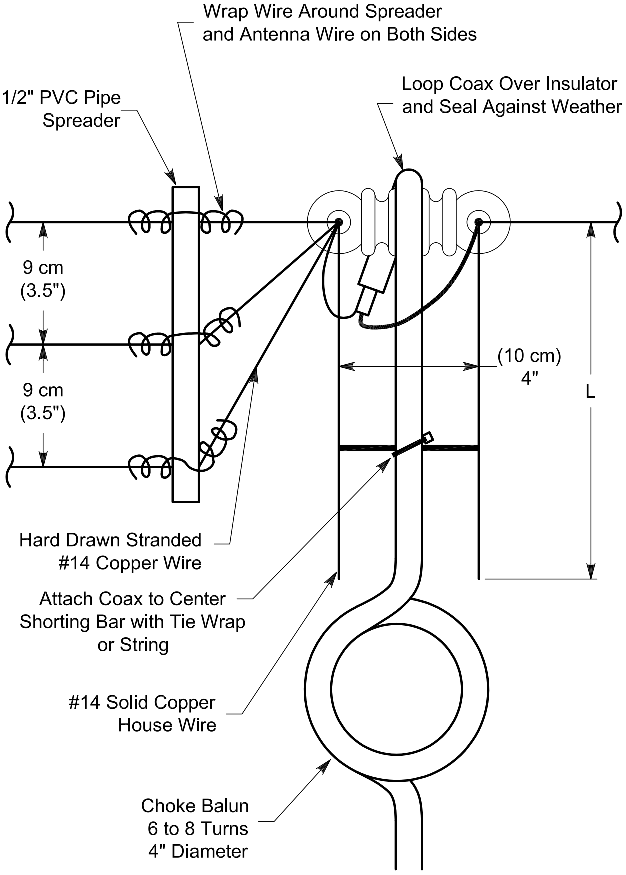
Wrap Wire Around Spreader and Antenna Wire on Both Sides
Loop Coax Over Insulator and Seal Against Weather
1/2" PVC Pipe Spreader
9 cm (3.5")
9 cm (3.5")
(10 cm) 4"
L
Hard Drawn Stranded #14 Copper Wire
Attach Coax to Center Shorting Bar with Tie Wrap or String
#14 Solid Copper House Wire
Choke Balun 6 to 8 Turns 4" Diameter

Fig 29—Details of feed point for 20/15/10-meter triband Yagi. The same mechanical support is provided for the balun and feed coax for the 30/17/12-meter tribander.

A 30/17/12-Meter Triband 2-Element Yagi

When the author attempted to add 17 and 12 meter elements to the existing 20/15/10-meter Yagi model he became exasperated. Adding two more driven elements and reflectors brings many more variables into the equation! It became clear that there was serious interaction between the elements. He could not obtain a workable feed-point impedance on all five bands that could be transformed to 50 Ω using a single hairpin match. There was also serious pattern distortion on 12 meters.

Even building a WARC-only triband Yagi for 30/17/12 meters turned out to be a real challenge. VE7CA had difficulty finding a combination that would allow him to use a single matching system to transform the feed-point impedance of the combined driven elements to 50 Ω. He couldn't create a 30/17/12 triband Yagi using the same design principles as his 20/15/10-meter version. The main problem occurred on 12 meters. Not only was the feed-point impedance unmanageable, but the radiation pattern had four lobes, not the single lobe you'd like from a Yagi.

He decided to try the *Modified Coaxial-Sleeve* method, more aptly termed by K9AY the *Coupled-Resonator (C-R)* in his article in *The ARRL Antenna Compendium, Vol 5*. The K9AY method uses a single driven element, with other elements placed in very close proximity (but not physically connected) to the driven element. By starting with the dimensions suggested by K9AY for a triband 30/17/12-meter dipole, VE7CA was able to develop a 2-element Yagi with acceptable feed-point impedances on all three bands using a single hairpin match. Notice that this WARC design uses a 2 × 2-inch wooden boom length that is 230 cm (7.5 feet) long. Of course, the antenna can't fit into a typical ski boot anymore, so VE7CA had to put it on a roof rack for transportation.

The space between the tightly coupled driven elements is only 3.7 cm (1.5 inches), so you need to use more PVC pipe spreaders than in the 20/15/10 design to make sure the driven-element wires stay as close as possible to the desired spacing without physically touching each other. The driven elements lie in the horizontal plane and the hairpin match and feed line hang down vertically from the center of the 30-meter driven element.

The spacing between the 30-meter driven elements and the other two conductors and the size of the wire all played a part in developing this antenna for a single feed line with the common hairpin match. Do not change the wire size from the recommended #14 for the driven elements unless you are willing to spend a considerable amount of time with a *NEC*-based modeling program retweaking the antenna. This is not the case with the 20/15/10 tri-band Yagi, where any convenient sized wire is acceptable.

Using #14 gauge wire allows all the Yagi antennas in this article to be used at the maximum power levels allowed in North America. The only limiting factor is the power handling capability of the feed line. However, even RG-58 should work for the relatively short length from the feed point down to ground level, where you can change to RG-8 or some other higher-power, lower-loss coaxial cable if you wish. **Fig 30** is a detailed drawing of the 30/17/12-meter driven element. The other dimensions for the 30/17/12-meter triband Yagi are shown in **Table 5**.

Assembly

When you are ready to assemble your wire Yagi, start by attaching the longest reflector element and the driven element assembly to the wood end booms. Do this with the wires and the booms on the ground. Next attach the V slings to both of the booms and with ropes attached to the V slings pull the array up off the ground between two supports (perhaps two trees). A height of 1.5 meters (5 feet) above the ground makes it easy to work on the antenna while you add the other reflector elements and adjust the V slings. Pull them tight so that the array is fairly flat. It won't stay horizontal, because the driven elements are heavier than a single reflector element. So you will need to support the 2 × 2-inch spreaders so they are horizontal. Lean the booms on something at a convenient height, such as the rungs of two step ladders. Now add the two other reflector elements, but don't pull them as tight as the longest reflector. Next attach the feed line.

Table 5
Dimensions for 30/17/12-Meter Tribander

Frequency MHz	*Spacing DE to Refl cm (feet)*	*Driven Element Length cm (feet)*	*Reflector Length cm (feet)*	*Hairpin Length cm (inches)*
10.12	230 (7.5)	713 (23.4) Half	1476 (48.4)	24.5 (9.5)
18.11	165 (5.4)	808 (26.5)	822 (27.0)	
24.91	120 (3.9)	570 (18.7)	606 (19.9)	

Spacing between hairpin wires is 10 cm (4 inches). Note that dimensions for 17 and 12-meter driven elements are full lengths, since they are not broken with insulator in the middle, unlike all driven elements for 20/15/10-meter triband design in Table 4.

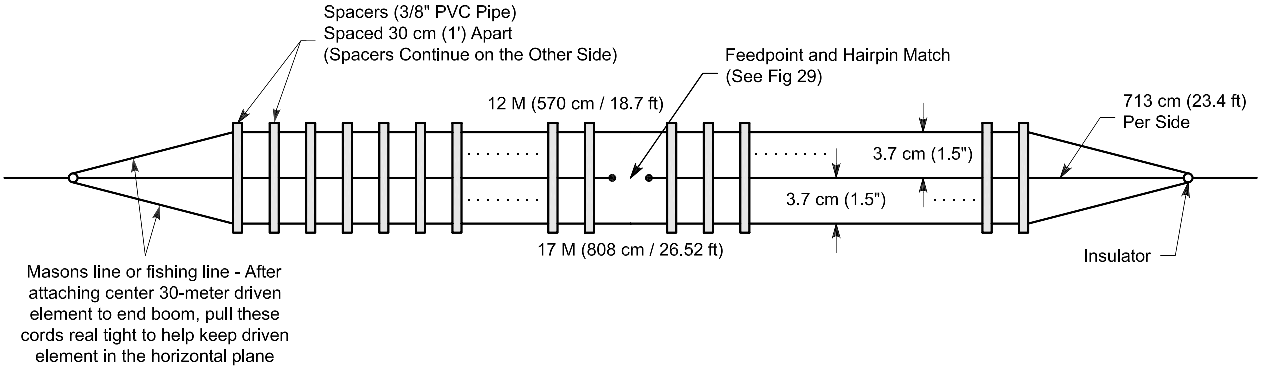
Spacers (3/8" PVC Pipe)
Spaced 30 cm (1') Apart
(Spacers Continue on the Other Side)
Feedpoint and Hairpin Match
(See Fig 29)
12 M (570 cm / 18.7 ft)
713 cm (23.4 ft)
Per Side
3.7 cm (1.5")
3.7 cm (1.5")
17 M (808 cm / 26.52 ft)
Insulator
Masons line or fishing line - After attaching center 30-meter driven element to end boom, pull these cords real tight to help keep driven element in the horizontal plane

Fig 30—Layout of 30-meter driven element with coupled resonators for 17 and 12 meters.

V Slings

Since the author wanted to be able to raise the 30/17/12-meter triband antenna by himself, he again used only one rope on either end of the array. One end goes over a tree limb and the other end is tied to a stake in the ground or some other nearby support, perhaps a tree trunk. Using only one attachment rope on either end makes it very easy to change beam direction by walking the antenna around the antenna support tree or tower. To accomplish this he used two V slings, one on each end attached to the 2 × 2-inch spreaders.

The secret to keeping the antenna level in the horizontal plane is that the V slings are not equilateral in shape. The combined weight of driven elements, balun and feed line is heavier than the reflectors. If the length of the sides of the V are equal, the array will rotate downwards. The driven elements will end up facing the ground, with the reflectors facing up. Adjust the V slings so that the antenna will stay level in the horizontal plane by shortening the length of the side of the V attached to the driven elements. It is quite easy to adjust in the field, and once you have it adjusted it stays balanced.

Once you raise the antenna to its operating position and in the horizontal plane, you can change direction 180° by pulling on the feed line. As you pull, the whole array will slowly turn over. Stop it from turning by holding onto the feed line once the array has swung over to face the opposite direction.

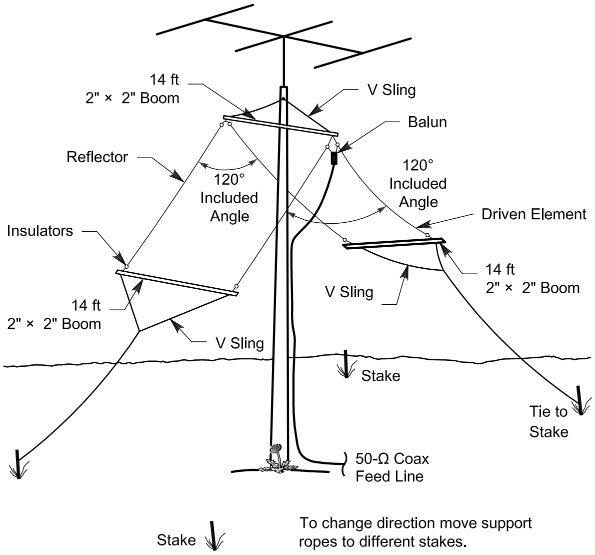
14 ft
2" × 2" Boom
V Sling
Balun
Reflector
120°
Included
Angle
120°
Included
Angle
Driven Element
Insulators
14 ft
2" × 2" Boom
V Sling
14 ft
2" × 2" Boom
V Sling
Stake
Tie to
Stake
50-Ω Coax
Feed Line
Stake
To change direction move support ropes to different stakes.

Fig 31—Layout for inverted-V 40-meter portable wire Yagi suspended from tower.

SWR Adjustment

Since you may situate your antenna in an entirely different position than VE7CA did, you may need to fine tune your antenna. Begin with the dimensions in Table 5 as a starting point. Make a temporary shorting bar using two alligator clips joined by a piece of wire and attach them at the recommended position. Next raise the antenna to the desired final position. Using an antenna analyzer (or transmitter and SWR meter) plot the SWR over all three bands. Start with the lowest band, 30 meters and adjust the shorting bar up or down to find the lowest SWR point in the portion of the band you plan to operate in. (This procedure also works if you wish to adjust the 20/15/10-meter tribander. Start with the lowest frequency.)

You shouldn't have to move the shorting bar very far from the suggested length. Now that you have determined the right shorting-bar position, adjust the other two driven element lengths for the lowest SWR in the portion of those bands in which you plan to operate. You may have to compromise with the position of the shorting bar to find a satisfactory range where the SWR is acceptable on all three bands. After satisfying himself with the position of the shorting bar, VE7CA replaced the alligator clips simply by folding one side of the parallel hair-pin wire lengths over to the other side and soldering it in the position where the alligator clips had been attached. The author does not recommend changing the reflector element lengths unless you are familiar with antenna modeling programs and are willing to model different spacing or element lengths.

40-Meter Wire Yagis

After his November 2001 *QST* article, VE7CA received several requests for a 40-meter wire Yagi. One ham mentioned that he wanted to be able to pull up a 40-meter Yagi between two towers and to be able to flip it over to change direction. Another wanted a 40-meter Yagi he could pull up on a single tower for the winter DX contests and then put it away during the summer. So VE7CA ran four different 40-meter scenarios in his computer models:

1. A horizontal 2-element wire Yagi at 65 feet.
2. A sloping 2-element wire Yagi, with one end at 65 feet and sloping downward 30° from vertical.
3. A sloping 2-element wire Yagi, with one end at 65 feet and sloping downward 65° from vertical.
4. An inverted-V 2-element wire Yagi with the apex at 65 feet and an included angle between the wires of 120°.

Fig 31 shows the layout for an inverted-V system and **Table 6** lists the element and hairpin lengths. Elevation patterns for the 40 meter antennas are compared in **Fig 32**, with a reference antenna of a single flattop dipole at 65 feet. As they say, a picture is worth a thousand words. If your interest is DX, it is very clear that horizontal and high is a very good rule of thumb for most antennas.

Yes, a $^{1}/_{4}$ wave vertical over salt water or 120 $^{1}/_{4}$-wave radials over good ground will produce very low radiation angles, but such systems are not exactly portable and we don't all live near the ocean. Mind you, if you can manage to locate antenna Number 3 (the most vertically oriented wire Yagi) next to the ocean, you would be very happy.

The point here is that if you have two towers and you're not fortunate enough to be located on a saltwater marsh, you should pull the 40-meter array up as high and as horizontal as you can. If you have only one tower and don't need to change direction often, then try the inverted-V configuration. You can still change the direction by walking each end around the tower.

However, even the sloping 40-meter Yagi with one end at 65 feet up a tower (or tree) and the other end attached with a long rope as far as possible from the tower will still put out a very respectable signal. It is directional, and you can walk it around the tower to change direction or you can

Table 6
40-Meter Wire Yagi Configurations

Configuration	*Driven Element cm (feet)*	*Reflector cm (feet)*	*Hairpin Length cm (feet)*
1. Horizontal	1978 (64.90)	2098 (68.83)	Approx 50 cm
2. –30° Sloper	1978 (64.90)	2113 (69.32)	(22 inches)
3. –65° Sloper	1978 (64.90)	2101 (68.93)	
4. Inverted V	2040 (66.93)	2126 (69.75)	

Spacing between driven element and reflector is 427 cm (14 feet). Spacing between parallel hairpin wires is 10 cm (4 inches). The lengths shown above are the total wire length for each element.

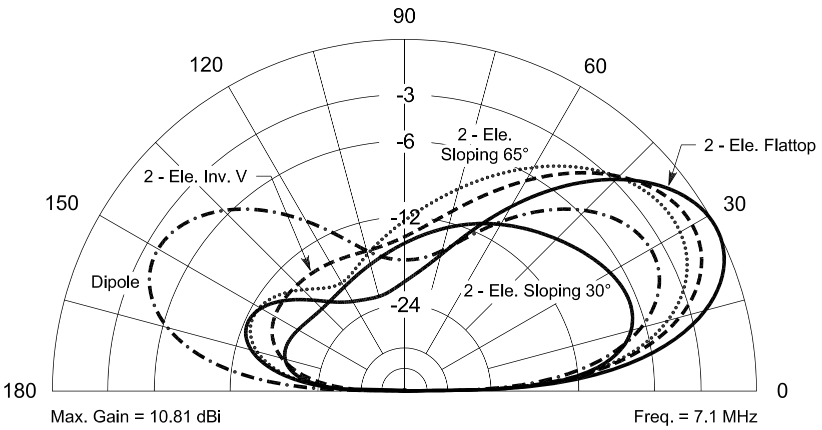

Fig 32—Comparisons of elevation patterns for five 40-meter antennas: a 2-element flattop Yagi at 65 feet; a 2-element inverted-V Yagi at 65 feet; a 2-element Yagi sloped 65° from the vertical plane; a 2-element Yagi sloped 30° from the vertical plane; and a horizontal dipole at 65 feet.

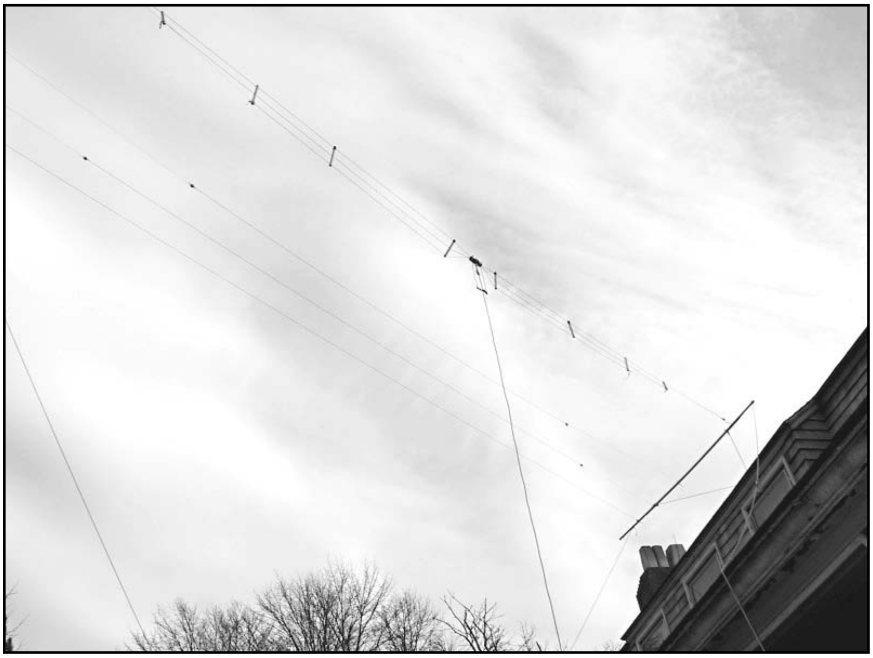

Fig 33 — VE7CA portable 2-element 20/15/10-meter wire Yagi installed at Plimoth Plantation in Plymouth, Massachusetts. (*Photo courtesy K1HTN*).

flip the antenna over and change direction 180° very quickly.

Summary

You don't need a 10-meter (60-foot) high tower, a commercial triband Yagi and a rotator to put out a good signal on the HF bands. Wire Yagis work very well and they can be inexpensive and easy-to-build, using components found at your local hardware store. **Fig 33** is a photograph of a successful portable installation in Massachusetts.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

G. Breed, K9AY, "The Coupled-Resonator Principle: A Flexible Method for Multiband Antennas," *The ARRL Antenna Compendium, Vol 5* (Newington: ARRL, 1996), p 109ff.

R. J. Decesari, "A Portable Quad for 2 Meters," *QST*, Oct 1980, and "Portable Quad for 2 Meters, Part 2," Technical Correspondence, *QST*, Jun 1981, p 39ff.

D. DeMaw, "A Traveling Ham's Trap Vertical," *QST*, Oct 1980, p 28ff.

J. Hall, "Zip-Cord Antennas—Do They Work?" *QST*, Mar 1979, p 31ff.

M. Hansen, "A Portable 2-Element Triband Yagi," *QST*, Nov 2001, p 35ff.

M. Hansen, "Some More VE7CA 2-Element Portable Yagis," *The ARRL Antenna Compendium, Vol7* (Newington: ARRL, 2002, p 183ff.

C. L. Hutchinson, "A Tree-Mounted 30-Meter Ground-Plane Antenna," *QST*, Sep 1974, p 16ff.

C. W. Schecter, "A Deluxe RV 5-Band Antenna," *QST*, Oct 1980, p 38.

Chapter 16

Mobile and Maritime Antennas

Mobile antennas are those designed for use while antennas, most amateurs think of a whip mounted on an automobile or other highway vehicle, perhaps on a recreational vehicle (RV) or maybe on an off-road vehicle. While it is true that most mobile antennas are vertical whips, mobile antennas can also be found in other places. For example, antennas mounted aboard a boat or ship are mobile, and are usually called *maritime antennas*. **Fig 1** shows yet another type of mobile antenna—those for use on handheld transceivers. Because they may be used while in motion, even these antennas are mobile by literal definition.

Pictured in Fig 1 is a telescoping full-size quarter-wave antenna for 144 MHz, and beside it a *stubby* antenna for the same band. The stubby is a helically wound radiator, made of stiff copper wire enclosed in a protective covering of rubber-like material. The inductance of the helical windings provides electrical loading for the antenna. For frequencies above 28 MHz, most mobile installations permit the use of a full-size antenna, but sometimes smaller, loaded antennas are used for convenience. The stubby, for example, is convenient for short-range communications, avoiding the problems of a lengthier, cumbersome antenna attached to a handheld radio.

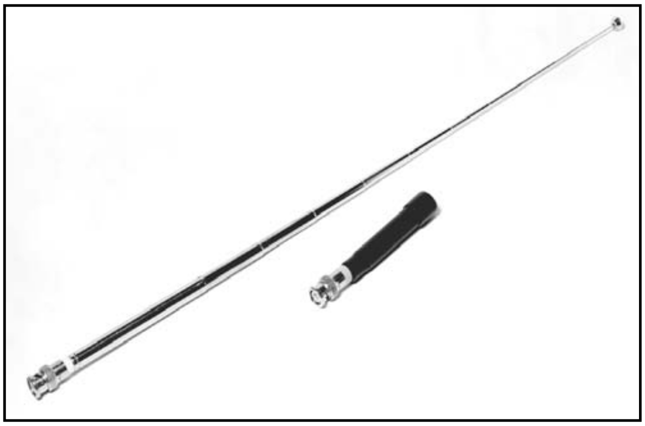

Fig 1—Two mobile antennas—mobile because they may be used while in motion. Shown here are a telescoping $^1/_4$-λ antenna and a "stubby" antenna, both designed for use at 144 MHz. The $^1/_4$-λ antenna is 19 inch long, while the stubby antenna is only $3^1/_2$ inch long. (Both dimensions exclude the length of the BNC connectors. The stubby is a helically wound radiator.

Below 28 MHz, physical size becomes a problem with full-size whips, and some form of electrical loading (as with the stubby) is usually employed. Commonly used loading techniques are to place a coil at the base of the whip (base loading), or at the center of the whip (center loading). These and other techniques are discussed in this chapter.

Few amateurs construct their own antennas for HF mobile and maritime use, since safety reasons dictate very sound mechanical construction. Several construction projects are included, however, in this chapter for those who may wish to build their own mobile antenna. Even if commercially made antennas are installed, most require some adjustment for the particular installation and type of operation desired and the information given here may provide a better understanding of the optimization requirements.

HF-MOBILE FUNDAMENTALS

Fig 2 shows a typical bumper-mounted center-loaded whip suitable for operation in the HF range. Jack Schuster, W1WEF, operates 80 through 2 meters from his car. The antenna could also be mounted on the car body itself (such as a fender), and mounts are available for this purpose. The base spring and tennis ball act as shock absorbers for the bottom of the whip, as the continual flexing while in motion would otherwise weaken the antenna. A short heavy mast section is mounted between the base spring and loading coil. Some models have a mechanism that allows the antenna to be tipped over for adjustment or for fastening to the roof of the car when not in use.

It is also advisable to extend a couple of guy lines from the base of the loading coil to clips or hooks fastened to the roof gutter on the car, or to the trunk and

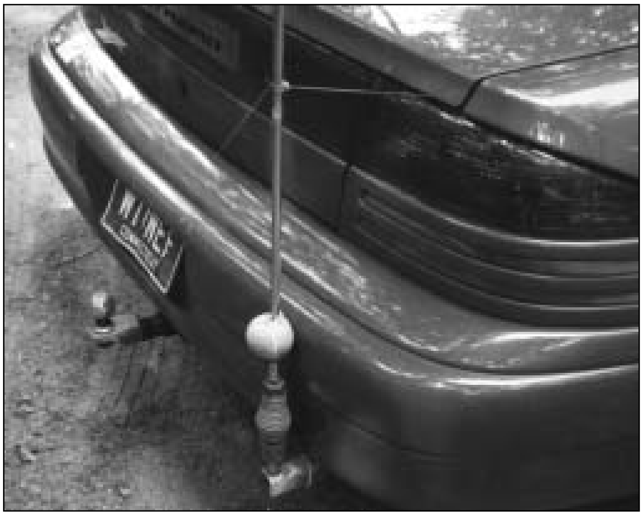

Fig 2—A typical bumper-mounted HF mobile antenna, as used by W1WEF. Note the nylon guy lines and the tennis ball used as a shock absorber. (*Photo courtesy W1WEF.*)

rear bumper, as W1WEF has done. Nylon fishing line (about 40-pound test) is suitable for this purpose. The guy lines act as safety cords and also reduce the swaying motion of the antenna considerably. The feed line to the transmitter is connected to the bumper and base of the antenna. Good low-resistance connections are important here.

Tune-up of the antenna is usually accomplished by changing the height of the adjustable whip section above the precut loading coil. First, tune the receiver and try to determine where the signals seem to peak up. Once this frequency is found, check the SWR with the transmitter on, and find the frequency where the SWR is lowest. Shortening the adjustable section will increase the resonant frequency, and making it longer will lower the frequency. It is important that the antenna be away from surrounding objects such as overhead wires by ten feet or more, as considerable detuning can occur. Once the setting is found where the SWR is lowest at the center of the desired frequency range, the length of the adjustable section should be recorded.

Propagation conditions and ignition noise are usually the limiting factors for mobile operation on 10 through 28 MHz. Antenna size restrictions affect operation somewhat on 7 MHz and much more on 3.5 and 1.8 MHz. From this standpoint, perhaps the optimum band for HF-mobile operation is 7 MHz. The popularity of the regional mobile nets on 7 MHz is perhaps the best indication of its suitability. For local work, 28 MHz is also useful, as antenna efficiency is high and relatively simple antennas without loading coils are easy to build.

As the frequency of operation is lowered, an antenna of fixed length looks (at its feed point) like a decreasing resistance in series with an increasing capacitive reactance. The capacitive reactance must be tuned out, necessitating the use of an equivalent series inductive reactance or loading coil. The amount of inductance required will be determined by the placement of the coil in the antenna system.

Base loading requires the lowest value of inductance for a fixed-length antenna, and as the coil is placed farther up the whip, the necessary value increases. This is because the capacitance of the shorter antenna section (above the coil) to the car body is now lower (higher capacitive reactance), requiring more inductance to tune the antenna to resonance. The advantage is that the current distribution on the whip is improved, increasing the radiation resistance. The disadvantage is that requirement of a larger coil also means the coil size and losses increase. Center loading has been generally accepted as a good compromise with minimal construction problems. Placing the coil $^2/_3$ the distance up the whip seems to be about the optimum position.

For typical antenna lengths used in mobile work, the difficulty in constructing suitable loading coils increases as the frequency of operation is lowered. Since the required resonating inductance gets larger and the radiation resistance decreases at lower frequencies, most of the power is dissipated in the coil's loss resistance and in other ohmic losses. This is one reason why it is advisable to buy a commercially made loading coil with the highest power rating possible, even if only low-power operation is planned.

Coil losses in the higher-power loading coils are usually less (percentage-wise), with subsequent improvement in radiation efficiency, regardless of the power level used. Of course, the above philosophy also applies to homemade loading coils, and design considerations will be considered in a later section.

Once the antenna is tuned to resonance, the input impedance at the antenna terminals will look like a pure resistance. Neglecting losses, this value drops from nearly 15 Ω at 21 MHz to 0.1 Ω at 1.8 MHz for an 8-foot whip. When coil and other losses are included, the input resistance increases to approximately 20 Ω at 1.8 MHz and 16 Ω at 21 MHz. These values are for relatively high-efficiency systems. From this it can be seen that the radiation efficiency is much poorer at 1.8 MHz than at 21 MHz under typical conditions.

Since most modern gear is designed to operate with a 50-Ω transmission line, a matching network is usually necessary, especially with the high-efficiency antennas previously mentioned. This can take the form of either a broadband transformer, a tapped coil, or an LC matching network. With homemade or modified designs, the tapped-coil arrangement is perhaps the easiest one to build, while the broadband transformer requires no adjustment. As the losses go up, so does the input resistance, and in less efficient systems the matching network may not be needed.

The Equivalent Circuit of a Typical Mobile Antenna

In the previous section, some of the general considerations were discussed, and these will now be taken up in more detail. It is customary in solving problems involving electric and magnetic fields (such as antenna systems) to try to find an equivalent network with which to replace the antenna for analysis reasons. In many cases, the network may be an accurate representation over only a limited frequency range. However, this is often a valuable method in matching the antenna to the transmission line.

Antenna resonance is defined as the frequency at which the input impedance at the antenna terminals is purely resistive. The shortest length at which this occurs for a vertical antenna over a ground plane is when the antenna is an electrical quarter wavelength at the operating frequency; the impedance value for this length (neglecting losses) is about 36 Ω. The idea of resonance can be extended to antennas shorter (or longer) than a quarter wave, and means only that the input impedance is purely resistive. As pointed out previously, when the frequency is lowered, the antenna looks like a series RC circuit, as shown in **Fig 3**. For the average 8-foot whip, the capacitive reactance of C_A may range from about –150 Ω at 21 MHz to as high as –8000 Ω at 1.8 MHz, while the radiation resistance R_R varies from about 15 Ω at 21 MHz to as low as 0.1 Ω at 1.8 MHz.

For an antenna less than 0.1 λ long, the approximate radiation resistance may be determined from the following:

$$R_R = 273 \times (\ell f)^2 \times 10^{-8} \qquad \text{(Eq 1)}$$

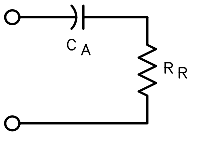

Fig 3—At frequencies below resonance, the whip antenna will show capacitive reactance as well as resistance. R_R is the radiation resistance, and C_A represents the antenna capacitance.

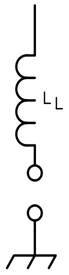

Fig 4—The capacitive reactance at frequencies below the resonant frequency of the whip can be canceled by adding an equivalent inductive reactance in the form of a loading coil in series with the antenna.

where ℓ is the length of the whip in inches, and f is the frequency in megahertz.

Since the resistance is low, considerable current must flow in the circuit if any appreciable power is to be dissipated in the form of radiation in R_R. Yet it is apparent that little current can be made to flow in the circuit as long as the comparatively high series reactance remains.

Antenna Capacitance

Capacitive reactance can be canceled by connecting an equivalent inductive reactance, (coil L_L) in series, as shown in **Fig 4**, thus tuning the system to resonance.

The capacitance of a vertical antenna shorter than a quarter wavelength is given by:

$$C_A = \frac{17\ell}{\left[\left(\ln\frac{24\ell}{D}\right)-1\right]\left[1-\left(\frac{f\ell}{234}\right)^2\right]} \qquad \text{(Eq 2)}$$

where

C_A = capacitance of antenna in pF
ℓ = antenna height in feet
D = diameter of radiator in inches
f = operating frequency in MHz

$$\ln\frac{24\ell}{D} = 2.3\log_{10}\frac{24\ell}{D}$$

Fig 5 shows the approximate capacitance of whip antennas of various average diameters and lengths. For 1.8, 4 and 7 MHz, the loading coil inductance required (when the loading coil is at the base) would be approximately the inductance required to resonate in the desired band (with the whip capacitance taken from the graph). For 10 through 21 MHz, this rough calculation will give more than the required inductance, but it will serve as a

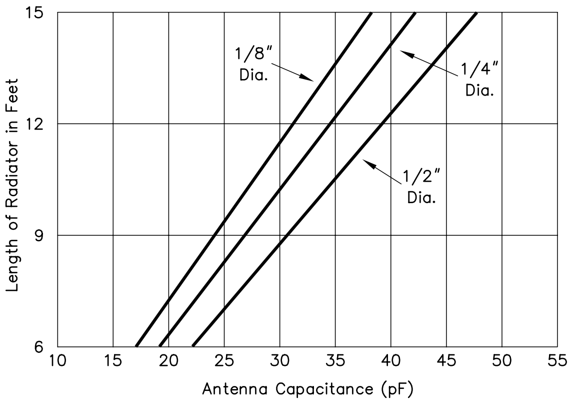

Fig 5—Graph showing the approximate capacitance of short vertical antennas for various diameters and lengths. These values should be approximately halved for a center-loaded antenna.

starting point for the final experimental adjustment that must always be made.

LOADING COIL DESIGN

To minimize loading coil loss, the coil should have a high ratio of reactance-to-resistance (that is, a high unloaded Q). A 4-MHz loading coil wound with small wire on a small-diameter solid form of poor quality, and enclosed in a metal protector, may have a Q as low as 50, with a loss resistance of 50 Ω or more. High-Q coils require a large conductor, air-wound construction, large spacing between turns, and the best insulating material available. A diameter not less than half the length of the coil (not always mechanically feasible) and a minimum of metal in the field of the coil are also necessities for optimum efficiency. Such a coil for 4 MHz may show a Q of 300 or more, with a resistance of 12 Ω or less.

The coil could then be placed in series with the feed line at the base of the antenna to tune out the unwanted capacitive reactance, as shown in Fig 4. Such a method is often referred to as *base-loading*, and many practical mobile antenna systems have been built using this scheme.

Over the years, the question has come up as to whether or not more efficient designs are possible compared with simple base loading. While many ideas have been tried with varying degrees of success, only a few have been generally accepted and incorporated into actual systems. These are center loading, continuous loading, and combinations of the latter with more conventional antennas.

Base Loading and Center Loading

If a whip antenna is short compared to a wavelength *and the current is uniform along the length ℓ*, the electric field strength E, at a distance d, away from the antenna is approximately:

$$E = \frac{120 \pi I \ell}{d \lambda} \qquad \text{(Eq 3)}$$

where

I is the antenna current in amperes

λ is the wavelength in the same units as D and ℓ.

A uniform current flowing along the length of the whip is an idealized situation, however, since the current is greatest at the base of the antenna and goes to a minimum at the top. In practice, the field strength will be less than that given by the above equation, because it is a function of the current distribution on the whip.

The reason that the current is not uniform on a whip antenna can be seen from the circuit approximation shown in **Fig 6**. A whip antenna over a ground plane is similar in many respects to a tapered coaxial cable where the center conductor remains the same diameter along its length, but with an increasing diameter outer conductor. The inductance per unit length of such a cable would increase along the line, while the capacitance per unit length would decrease. In Fig 6 the antenna is represented by a series of LC circuits in which C1 is greater than C2, which is greater than C3, and so on. L1 is less than L2, which is less than succeeding inductances. The net result is that most of the antenna current returns to ground near the base of the antenna, and very little near the top.

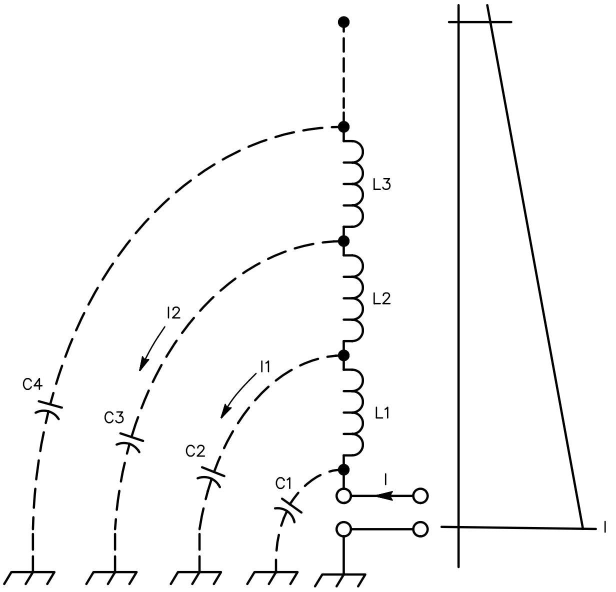

Fig 6—A circuit approximation of a simple whip over a perfectly conducting ground plane. The shunt capacitance per unit length gets smaller as the height increases, and the series inductance per unit length gets larger. Consequently, most of the antenna current returns to the ground plane near the base of the antenna, giving the current distribution shown at the right.

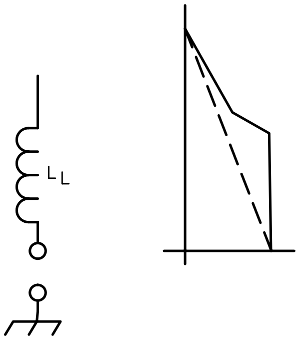

Fig 7—Improved current distribution resulting from center loading.

Two things can be done to improve this distribution and make the current more uniform. One would be to increase the capacitance of the top of the antenna to ground through the use of top loading or a capacitance hat, as discussed in Chapter 6. Unfortunately, the wind resistance of the hat makes it somewhat unwieldy for mobile use. The other method is to place the loading coil farther up the whip, as shown in **Fig 7**, rather than at the base. If the coil is resonant (or nearly so) with the

Table 1
Approximate Values for 8-ft Mobile Whip

f(MHz)	Loading L μH	R_C(Q50) Ω	R_C(Q300) Ω	R_R Ω	Feed R* Ω	Matching L μH
Base Loading						
1.8	345	77	13	0.1	23	3
3.8	77	37	6.1	0.35	16	1.2
7.2	20	18	3	1.35	15	0.6
10.1	9.5	12	2	2.8	12	0.4
14.2	4.5	7.7	1.3	5.7	12	0.28
18.1	3.0	5.0	1.0	10.0	14	0.28
21.25	1.25	3.4	0.5	14.8	16	0.28
24.9	0.9	2.6	—	20.0	22	0.25
29.0	—	—	—	—	36	0.23
Center Loading						
1.8	700	158	23	0.2	34	3.7
3.8	150	72	12	0.8	22	1.4
7.2	40	36	6	3.0	19	0.7
10.1	20	22	4.2	5.8	18	0.5
14.2	8.6	15	2.5	11.0	19	0.35
18.1	4.4	9.2	1.5	19.0	22	0.31
21.25	2.5	6.6	1.1	27.0	29	0.29

R_C = loading coil resistance; R_R = radiation resistance.
*Assuming loading coil Q = 300, and including estimated ground-loss resistance.

Table 2
Suggested Loading Coil Dimensions

Req'd L(μH)	Turns	Wire Size	Dia. Inches	Length Inches
700	190	22	3	10
345	135	18	3	10
150	100	16	2 1/2	10
77	75	14	2 1/2	10
77	29	12	5	4 1/4
40	28	16	2 1/2	2
40	34	12	2 1/2	4 1/4
20	17	16	2 1/2	1 1/4
20	22	12	2 1/2	2 3/4
8.6	16	14	2	2
8.6	15	12	2 1/2	3
4.5	10	14	2	1 1/4
4.5	12	12	2 1/2	4
2.5	8	12	2	2
2.5	8	6	2 3/8	4 1/2
1.25	6	12	1 3/4	2
1.25	6	6	2 3/8	4 1/2

capacitance to ground of the section above the coil, the current distribution is improved as also shown in Fig 7. The result with both top loading and center loading is that the radiation resistance is increased, offsetting the effect of losses and making matching easier.

Table 1 shows the approximate loading coil inductance for the various amateur bands. Also shown in the table are approximate values of radiation resistance to be expected with an 8-foot whip, and the resistances of loading coils—one group having a Q of 50, the other a Q of 300. A comparison of radiation and coil resistances will show the importance of reducing the coil resistance to a minimum, especially on the three lower frequency bands. **Table 2** shows suggested loading-coil dimensions for the inductance values given in Table 1.

OPTIMUM DESIGN OF SHORT COIL-LOADED HF MOBILE ANTENNAS

Optimum design of short HF mobile antennas results from a careful balance of the appropriate loading coil Q-factor, loading coil position in the antenna, ground loss resistance, and the length-to-diameter ratio of the antenna. The optimum balance of these parameters can be realized only through a thorough understanding of how they interact. This section presents a mathematical approach to designing mobile antennas for maximum radiation efficiency. Bruce Brown, W6TWW, in *The ARRL Antenna Compendium Volume 1*, first presented this approach. (See the Bibliography at the end of this chapter.)

The optimum location for a loading coil in an antenna can be found experimentally, but it requires many hours of designing and constructing models and making measurements to ensure the validity of the design. A faster and more reliable way of determining optimum coil location is through the use of a personal computer. This approach allows the variation of any single variable, while observing the cumulative effects on the system. When plotted graphically, the data reveals that the placement of the loading coil is critical if maximum radiation efficiency

Table 3
Variables used in Eqs 4 through 20

A = area in degree-amperes
a = antenna radius in English or metric units
dB = signal loss in decibels
E = efficiency in percent
f (MHz) = frequency in megahertz
H = height in English or metric units
h = height in electrical degrees
h_1 = height of base section in electrical degrees
h_2 = height of top section in electrical degrees
$I = I_{base}$ = 1 ampere base current
k = 0.0128
k_m = mean characteristic impedance
k_{m1} = mean characteristic impedance of base section
k_{m2} =mean characteristic impedance of top section
L = length or height of the antenna in feet
P_I = power fed to the antenna
P_R = power radiated
Q = coil figure of merit
R_C = coil loss resistance in Ω
R_G = ground loss resistance in Ω
R_R = radiation resistance in Ω
X_L = loading-coil inductive reactance

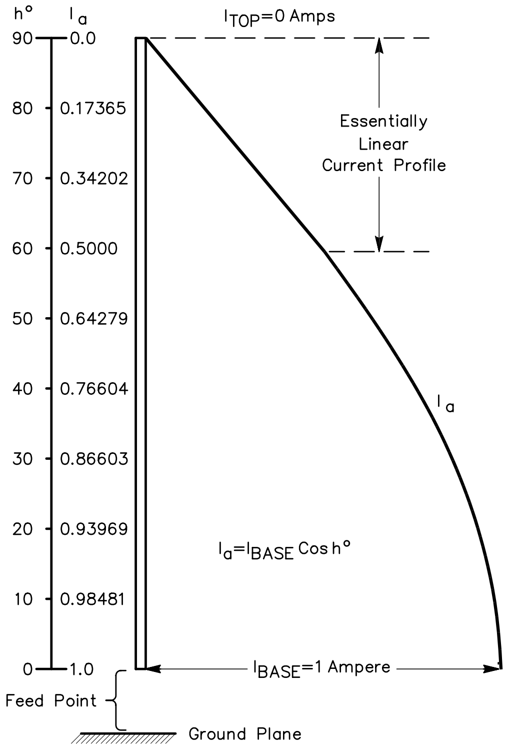

Fig 8—Relative current distribution on a vertical antenna of height h = 90 electrical degrees.

is to be realized. (See the program *MOBILE.EXE*, which is included on the CD-ROM in the back of this book.)

Radiation Resistance

The determination of radiation efficiency requires the knowledge of resistive power losses and radiation losses. Radiation loss is expressed in terms of radiation resistance. Radiation resistance is defined as the resistance that would dissipate the same amount of power that is radiated by the antenna. The variables used in the equations that follow are defined once in the text, and are summarized in **Table 3**. Radiation resistance of vertical antennas shorter than 45 electrical degrees ($^1/_8$ wavelength) is approximately:

$$R_R = \frac{h^2}{312} \qquad \text{(Eq 4)}$$

where

R_R = radiation resistance in Ω
h = antenna length in electrical degrees.

Antenna height in electrical degrees is expressed by:

$$h = \frac{\ell}{984} \times f\left(\text{MHz}\right) \times 360 \qquad \text{(Eq 5)}$$

where

ℓ = antenna length in feet
f (MHz) = operating frequency in megahertz.

End effect is purposely omitted to ensure that an antenna is electrically long. This is so that resonance at the design frequency can be obtained easily by removing a turn or two from the loading coil.

Eq 4 is valid only for antennas having a sinusoidal current distribution and no reactive loading. However, it can be used as a starting point for deriving an equation that is useful for shortened antennas with other than sinusoidal current distributions.

Refer to **Fig 8**. The current distribution on an antenna 90° long electrically ($^1/_4$ wavelength) varies with the cosine of the length in electrical degrees. The current distribution of the top 30° of the antenna is essentially linear. It is this linearity that allows for derivation of a simpler, more useful equation for radiation resistance.

The radiation resistance of an electrically short base-loaded vertical antenna can be conveniently defined in terms of a geometric figure, a triangle, as shown in **Fig 9**. The radiation resistance is given by:

$$R_R = KA^2 \qquad \text{(Eq 6)}$$

where

K is a constant (to be derived shortly)
A = area of the triangular current distribution in degree-amperes.

Degree-ampere area is expressed by

$$A = \frac{1}{2} h \times I_{base} \qquad \text{(Eq 7)}$$

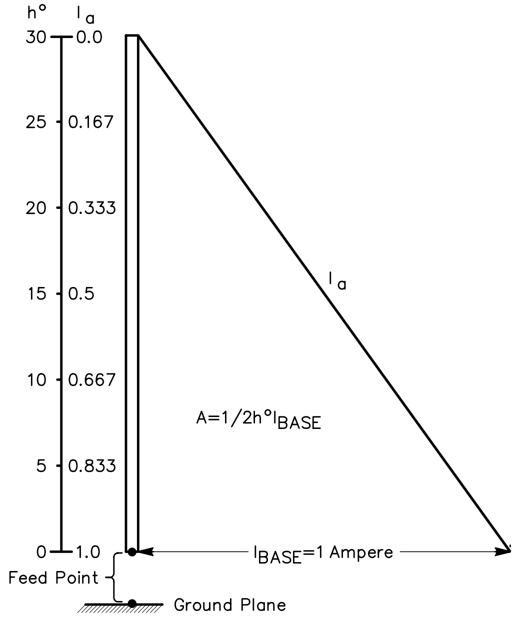

Fig 9—Relative current distribution on a base-loaded vertical antenna of height H = 30 electrical degrees (linearized). The base loading coil is not shown here.

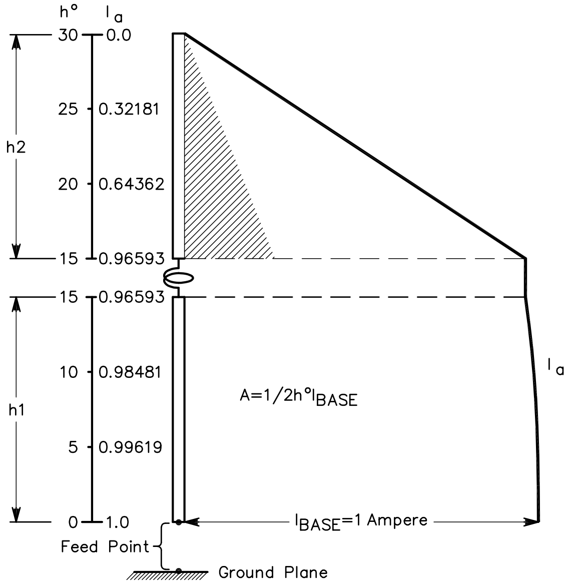

Fig 10—Relative current distribution on a center-loaded antenna with base and top sections each equal to 15 electrical degrees in length. The cross-hatched area shows the current distribution that would exist in the top 15° of a 90°-high vertical fed with 1 ampere at the base.

By combining Eqs 4 and 6 and solving for K, we get

$$K = \frac{h^2}{312 \times A^2} \qquad \text{(Eq 8)}$$

By substituting the values from Fig 9 into Eq 8 we get

$$K = \frac{30^2}{312 \times \left(0.5 \times 30 \times 1\right)^2} = 0.0128$$

and by substituting the derived value of K into Eq 6 we get

$$R_R = 0.0128 \times A^2 \qquad \text{(Eq 9)}$$

Eq 9 is useful for determining the radiation resistance of coil-loaded vertical antennas less than 30° in length. The derived constant differs slightly from that presented by Laport (see Bibliography), as he used a different equation for radiation resistance Eq 4.

When the loading coil is moved up an antenna (away from the feed point), the current distribution is modified as shown in **Fig 10**. The current varies with the cosine of the height in electrical degrees at any point in the base section. Therefore, the current flowing into the bottom of the loading coil is less than the current flowing at the base of the antenna.

But what about the current in the top section of the antenna? The loading coil acts as the lumped constant that it is, and disregarding losses and coil radiation, maintains the same current flow throughout. As a result, the current at the top of a high-Q coil is essentially the same as that at the bottom of the coil. This is easily verified by installing RF ammeters immediately above and below the loading coil in a test antenna. Thus, the coil "forces" much more current into the top section than would flow in the equivalent section of a full 90° long antenna. This occurs as a result of the extremely high voltage that appears at the top of the loading coil. This higher current flow results in more radiation than would occur from the equivalent section of a quarter-wave antenna. (This is true for conventional coils. However, radiation from long thin coils allows coil current to decrease, as in helically wound antennas.)

The cross-hatched area in Fig 10 shows the current that would flow in the equivalent part of a 90° high antenna, and reveals that the degree-ampere area of the whip section of the short antenna is greatly increased as a result of the modified current distribution. The current flow in the top section decreases almost linearly to zero at the top. This can be seen in Fig 10.

The degree-ampere area of Fig 10 is the sum of the triangular area represented by the current distribution in the top section, and the nearly trapezoidal current distribution in the base section. Radiation from the coil is not included in the degree-ampere area because it is small and difficult to define. Any radiation from the coil can be considered a bonus.

The degree-ampere area is expressed by:

$$A = \frac{1}{2}\left[h_1\left(1+\cos h_1\right) + h_2\left(\cos h_1\right)\right] \qquad \text{(Eq 10)}$$

where

h_1 = electrical length in degrees of the base section

h_2 = electrical height in degrees of the top section.

The degree-ampere area (calculated by substituting Eq 10 into Eq 9) can be used to determine the radiation resistance when the loading coil is at any position other than the base of the antenna. Radiation resistance has been calculated with these equations and plotted against loading coil position at three different frequencies for 8- and 11-foot antennas, **Fig 11**. Eight feet is a typical length for commercial antennas, and 11-foot antennas are about the maximum practical length that can be installed on a vehicle.

In Fig 11, the curves reveal that the radiation resistance rises almost linearly as the loading coil is moved up the antenna. They also show that the radiation resistance rises rapidly as the frequency is increased. If the analysis were stopped at this point, one might conclude that the loading coil should be placed at the top of the antenna. This is not so, and will become apparent shortly.

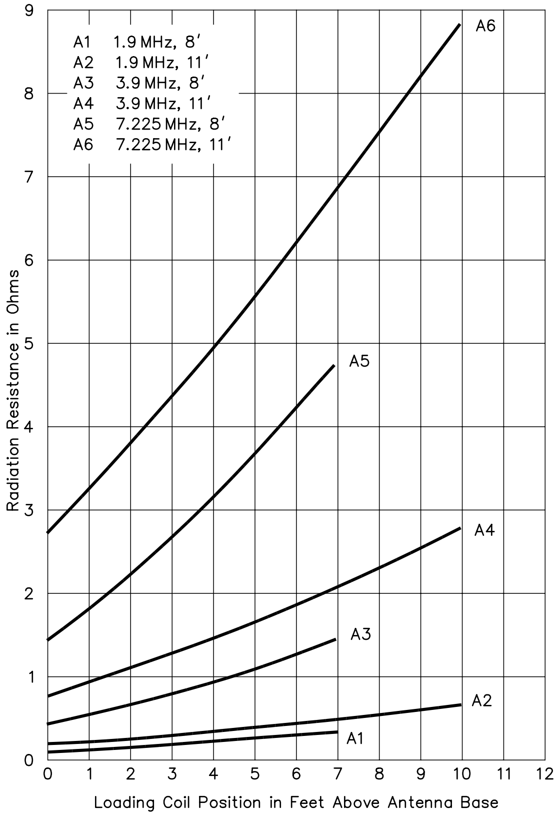

Fig 11—Radiation resistance plotted as a function of loading coil position.

Required Loading Inductance

Calculation of the loading coil inductance needed to resonate a short antenna can be done easily and accurately by using the antenna transmission-line analog described by Boyer in *Ham Radio*. For a base-loaded antenna, Fig 9, the loading coil reactance required to resonate the antenna is given by

$$X_L = -j\,K_m \cot h \qquad \text{(Eq 11)}$$

where

X_L = inductive reactance required

K_m = mean characteristic impedance (defined in Eq 12).

The $-j$ term indicates that the antenna presents capacitive reactance at the feed point. A loading coil must cancel this reactance.

The mean characteristic impedance of an antenna is expressed by

$$K_m = 60\left[\left(\ln\frac{2H}{a}\right) - 1\right] \qquad \text{(Eq 12)}$$

where

H = physical antenna height (excluding the length of the loading coil)

a = radius of the antenna in the same units as H.

From Eq 12 you can see that decreasing the height-to-diameter ratio of an antenna by increasing the radius results in a decrease in K_m. With reference to Eq 11, a decrease in K_m decreases the inductive reactance required to resonate an antenna. As will be shown later, this will increase radiation efficiency. In mobile applications, we quickly run into wind-loading problems if we attempt to use an antenna that is physically large in diameter.

If the loading coil is moved away from the base of the antenna, the antenna is divided into a base and top section, as depicted in Fig 10. The loading coil reactance required to resonate the antenna when the coil is away from the base is given by

$$X_L = j\,K_{m2}\left(\cot h_2\right) - j\,K_{m1}\left(\tan h_1\right) \qquad \text{(Eq 13)}$$

In mobile-antenna design and construction, the top section is usually a whip with a much smaller diameter than the base section. Because of this, it is necessary to compute separate values of K_m for the top and base sections. K_{m1} and K_{m2} are the mean characteristic impedances of the base and top sections, respectively.

Loading coil reactance curves for the 3.8-MHz antennas of Fig 11 have been calculated and plotted in **Fig 12**. These curves show the influence of the loading coil position on the reactance required for resonance. The curves in Fig 12 show that the required reactance decreases with longer antennas. The curves also reveal that the required loading coil reactance grows at an increasingly rapid rate after the coil passes the center of the antenna. Because the highest possible loading coil Q

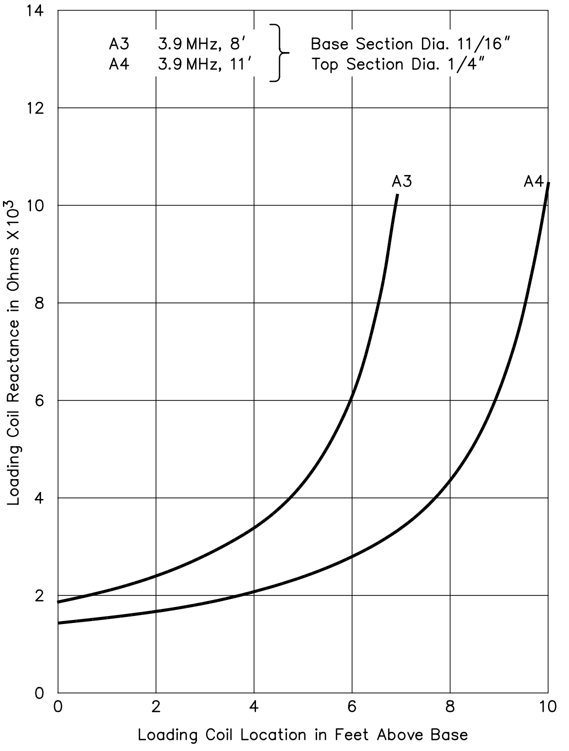

Fig 12—Loading coil reactance required for resonance, plotted as a function of coil height above the antenna base. The resonant frequency is 3.9 MHz.

factor is needed, and because optimum Q is attained when the loading coil diameter is twice the loading coil length, the coil would grow like a smoke ring above the center of the antenna, and would quickly reach an impractical size. It is for this reason that the highest loading coil position is limited to one foot from the top of the antenna in all computations.

Loading Coil Resistance

Loading coil resistance constitutes one of the losses that consumes power that could otherwise be radiated by the antenna. Heat loss in the loading coil is not of any benefit, so it should be minimized by using the highest possible loading coil Q. Loading coil loss resistance is a function of the coil Q and is given by

$$R_C = \frac{X_L}{Q} \qquad \text{(Eq 14)}$$

where

R_C = loading coil loss resistance in Ω
X_L = loading coil reactance
Q = coil figure of merit

Inspection of Eq 14 reveals that, for a given value of inductive reactance, loss resistance will be lower for higher Q coils. Measurements made with a Q meter show that typical, commercially manufactured coil stock produces a Q between 150 and 160 in the 3.8-MHz band.

Higher Q values can be obtained by using larger diameter coils having a diameter-to-length ratio of two, by using larger diameter wire, by using more spacing between turns, and by using low-loss polystyrene supporting and enclosure materials. In theory, loading coil turns should not be shorted for tuning purposes because shorted turns somewhat degrade Q. Pruning to resonance should be done by removing turns from the coil.

In fairness, it should be pointed out that many practical mobile antennas use large-diameter loading coils with shorted turns to achieve resonance. The popular "Texas Bug Catcher" coils come to mind here. Despite general proscriptions against shorting turns, these systems are often more efficient than antennas with small, relatively low-Q, fixed loading coils.

Radiation Efficiency

The ratio of power radiated to power fed to an antenna determines the radiation efficiency. It is given by:

$$E = \frac{P_R}{P_I} \times 100\% \qquad \text{(Eq 15)}$$

where

E = radiation efficiency in percent
P_R = power radiated
P_I = power fed to the antenna at the feed point.

In a short, coil-loaded mobile antenna, a large portion of the power fed to the antenna is dissipated in ground and coil resistances. A relatively insignificant amount of power is also dissipated in the antenna conductor resistance and in the leakage resistance of the base insulator. Because these last two losses are both very small and difficult to estimate, they are here neglected in calculating radiation efficiency.

Another loss worth noting is matching network loss. Because we are concerned only with power fed to the antenna in the determination of radiation efficiency, matching network loss is not considered in any of the equations. Suffice it to say that matching networks should be designed for minimum loss in order to maximize the transmitter power available at the antenna.

The radiation efficiency equation may be rewritten and expanded as follows:

$$E = \frac{I^2 R_R \times 100}{I^2 R_R + I^2 R_G + \left(I \cos h_1\right)^2 R_C} \qquad \text{(Eq 16)}$$

where

I = antenna base current in amperes
R_G = ground loss resistance in Ω
R_C = coil loss resistance in Ω.

Each term of Eq 16 represents the power dissipated in its associated resistance. All the current terms cancel, simplifying this equation to

$$E = \frac{R_R \times 100}{R_R + R_G + R_C \left(\cos^2 h_1\right)} \qquad \text{(Eq 17)}$$

For base-loaded antennas the term $\cos^2 h_1$ drops to unity and may be omitted.

Ground Loss

Eq 14 shows that the total resistive losses in the antenna system are:

$$R_T = R_R + R_G + R_c \left(\cos^2 h_1\right) \qquad \text{(Eq 18)}$$

where R_T is the total resistive loss. Ground loss resistance can be determined by rearranging Eq 18 as follows:

$$R_G = R_T - R_R - R_C \cos^2 h_1 \qquad \text{(Eq 19)}$$

R_T may be measured in a test antenna installation on a vehicle using an R-X noise bridge or an SWR analyzer. You can then calculate R_R and R_C.

Ground loss is a function of vehicle size, placement of the antenna on the vehicle, and conductivity of the ground over which the vehicle is traveling. Only the first two variables can be feasibly controlled. Larger vehicles provide better ground planes than smaller ones. The vehicle ground plane is only partial, so the result is considerable RF current flow (and ground loss) in the ground around and under the vehicle.

By raising the antenna base as high as possible on the vehicle, ground losses are decreased. This results from a decrease in antenna capacitance to ground, which increases the capacitive reactance to ground. This, in turn, reduces ground currents and ground losses.

This effect has been verified by installing the same antenna at three different locations on two different vehicles, and by determining the ground loss from Eq 19. In the first test, the antenna was mounted 6 inches below the top of a large station wagon, just behind the left rear window. This placed the antenna base 4 feet 2 inches above the ground, and resulted in a measured ground loss resistance of 2.5 Ω. The second test used the same antenna mounted on the left rear fender of a mid-sized sedan, just to the left of the trunk lid. In this test, the measured ground loss resistance was 4 Ω. The third test used the same mid-sized car, but the antenna was mounted on the rear bumper. In this last test, the measured ground loss resistance was 6 Ω.

The same antenna therefore sees three different ground loss resistances as a direct result of the antenna mounting location and size of the vehicle. It is important to note that the measured ground loss increases as the antenna base nears the ground. The importance of minimizing ground losses in mobile antenna installations cannot be overemphasized.

Efficiency Curves

With the equations defined previously, a computer

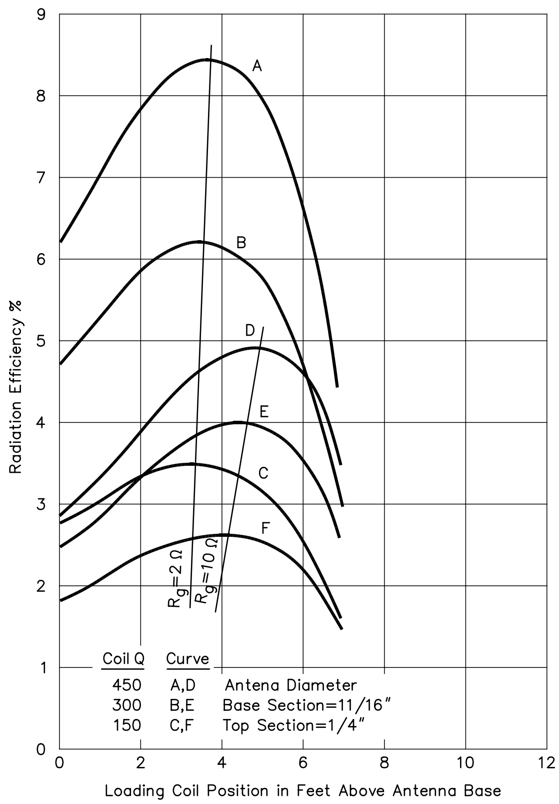

Fig 13—Radiation efficiency of 8-foot antennas at 3.9 MHz.

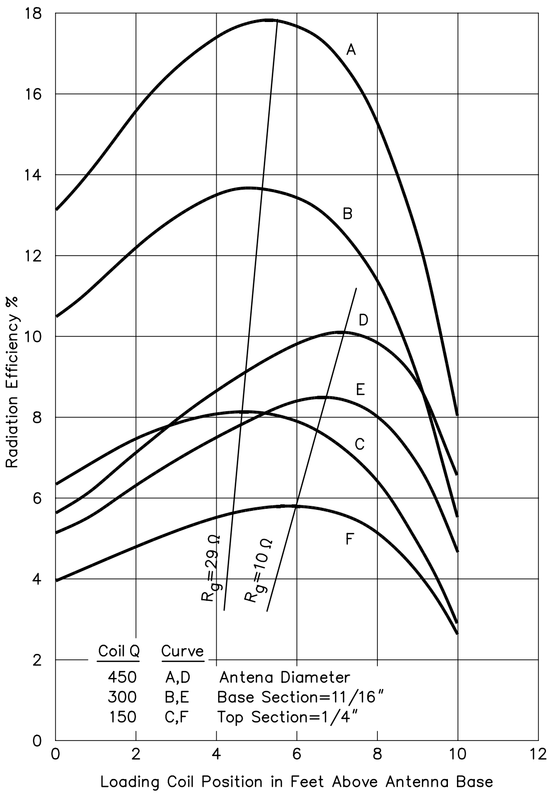

Fig 14—Radiation efficiency of 11-foot antennas at 3.9 MHz.

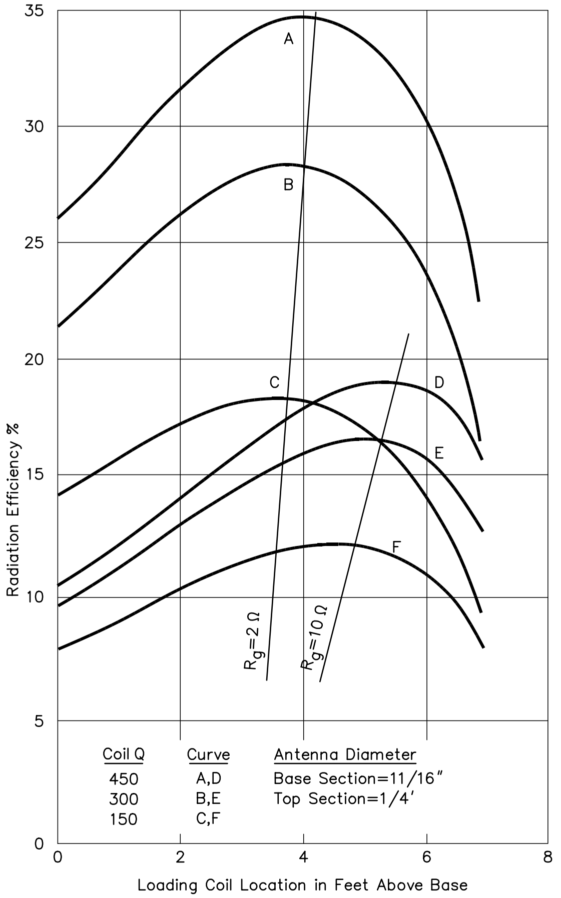

Fig 15—Radiation efficiency of 8-foot antennas at 7.225 MHz.

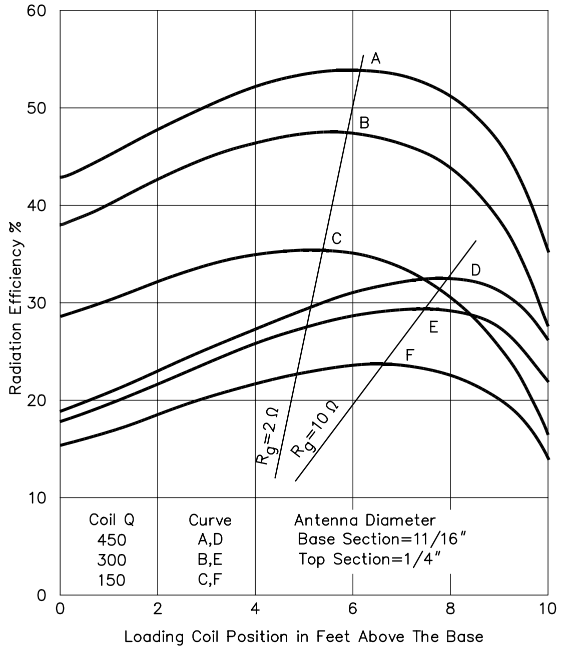

Fig 16—Radiation efficiency of 11-foot antennas at 7.225 MHz.

was used to calculate the radiation efficiency curves depicted in **Figs 13** through **16**. These curves were calculated for 3.8- and 7-MHz antennas of 8- and 11-foot lengths. Several values of loading coil Q were used, for both 2 and 10 Ω of ground loss resistance. For the calculations, the base section is $^1/_2$-inch diameter electrical EMT, which has an outside diameter of $^{11}/_{16}$ inch. The top section is fiberglass bicycle-whip material covered with Belden braid. These are readily available materials, which can be used by the average amateur to construct an inexpensive but rugged antenna.

Upon inspection, these radiation-efficiency curves reveal some significant information:

1) Higher coil Q produces higher radiation efficiencies,
2) longer antennas produce higher radiation efficiencies,
3) higher frequencies produce higher radiation efficiencies,
4) lower ground loss resistances produce higher radiation efficiencies,
5) higher ground loss resistances force the loading coil above the antenna center to reach a crest in the radiation-efficiency curve, and
6) higher coil Q sharpens the radiation-efficiency curves, resulting in the coil position being more critical for optimum radiation efficiency.

Note that the radiation efficiency curves reach a peak and then begin to decline as the loading coil is raised farther up the antenna. This is because of the rapid increase in loading coil reactance required above the antenna center. Refer to Fig 12. The rapid increase in coil size required for resonance results in the coil loss resistance increasing much more rapidly than the radiation resistance. This results in decreased radiation efficiency, as shown in Fig 11.

A slight reverse curvature exists in the curves between the base-loaded position and the one-foot coil-height position. This is caused by a shift in the curve resulting from insertion of a base section of larger diameter than the whip when the coil is above the base.

The curves in Figs 13 through 16 were calculated with constant (but not equal) diameter base and whip sections. Because of wind loading, it is not desirable to increase the diameter of the whip section. However, the base-section diameter can be increased within reason to further improve radiation efficiency. **Fig 17** was calculated for base-section diameters ranging from $^{11}/_{16}$ inch to 3 inches. The curves reveal that a small increase in radiation efficiency results from larger diameter base sections.

The curves in Figs 13 through 16 show that radiation efficiencies can be quite low in the 3.8-MHz band compared to the 7-MHz band. They are lower yet in the 1.8-MHz band. To gain some perspective on what these low efficiencies mean in terms of signal strength, **Fig 18** was calculated using the following equation:

$$\mathrm{dB} = \log \frac{100}{\mathrm{E}} \qquad \text{(Eq 20)}$$

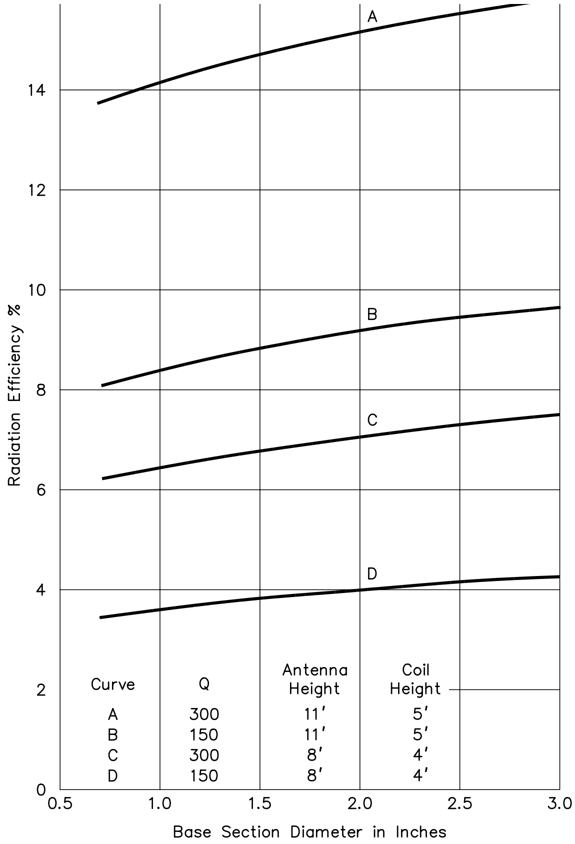

Fig 17—Radiation efficiency plotted as a function of base section diameter. Frequency = 3.9 MHz, ground loss resistance = 2 Ω, and whip section = $^1/_4$-inch diameter.

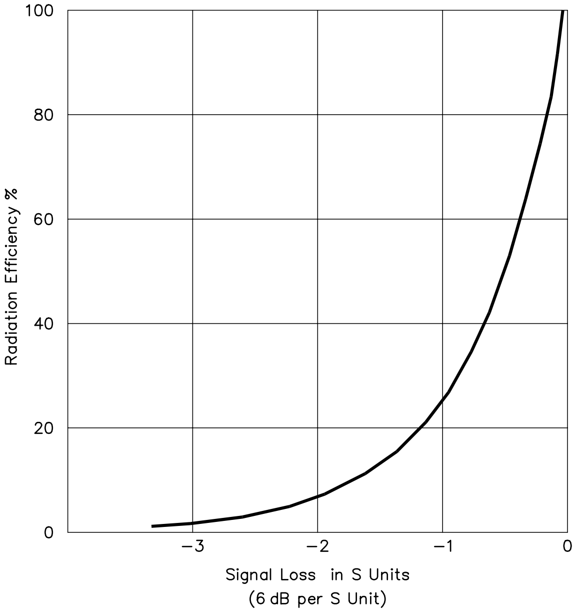

Fig 18—Mobile antenna signal loss as a function of radiation efficiency, compared to a quarter-wave vertical antenna over perfect ground.

where

dB = signal loss in decibels
E = efficiency in percent.

The curve in Fig 18 reveals that an antenna having 25% efficiency has a signal loss of 6 dB (approximately one S unit) below a quarter-wave vertical antenna over perfect ground. An antenna efficiency in the neighborhood of 6% will produce a signal strength on the order of two S units or about 12 dB below the same quarter-wave reference vertical. By careful optimization of mobile-antenna design, signal strengths from mobiles can be made fairly competitive with those from fixed stations using comparable power. And don't forget: Moving your car near saltwater, with its high conductivity, can result in surprisingly strong signals from a mobile station!

Impedance Matching

The input impedance of short, high-Q coil-loaded antennas is quite low. For example, an 8-foot antenna optimized for 3.9 MHz with an unloaded coil Q of 300 and a ground-loss resistance of 2 Ω has a base input impedance of about 13 Ω. This low impedance value causes a standing wave ratio of 4:1 on a 50-Ω coaxial line at resonance. This high SWR is not compatible with the requirements of solid-state transmitters. Also, the bandwidth of shortened vertical antennas is very narrow. This severely limits the capability to maintain transmitter loading over even a small frequency range.

Impedance matching can be accomplished by means of L networks or impedance-matching transformers, but the narrow bandwidth limitation remains. A more elegant solution to the impedance matching and narrow bandwidth problem is to install an automatic tuner at the antenna base. Such a device matches the antenna and coaxial line automatically, and permits operation over a wide frequency range.

The tools are now available to tailor a mobile antenna design to produce maximum radiation efficiency. Mathematical modeling with a personal computer reveals that loading coil Q factor and ground loss resistance greatly influence the optimum loading coil position in a short vertical antenna. It also shows that longer antennas, higher coil Q, and higher operating frequencies produce higher radiation efficiencies.

End effect has not been included in any of the equations to assure that the loading coil will be slightly larger than necessary. Pruning the antenna to resonance should be done by removing coil turns, rather than by shorting turns or shortening the whip section excessively. Shortening the whip reduces radiation efficiency, by both shortening the antenna and moving the optimum coil position. Shorting turns in the loading coil degrades the Q of the coil.

Shortened Dipoles

Mathematical modeling techniques can be applied to shortened dipoles by using zero ground loss resistance and by doubling the computed values of radiation resistance and feed-point impedance. Radiation efficiency, however, does not double. Rather, it remains unchanged, because a second loading coil is required in the other leg of the dipole. The addition of the second coil offsets the gain in efficiency that occurs when the feed-point impedance and radiation resistance are doubled. There is a gain in radiation efficiency over a vertical antenna worked against ground, though, because the dipole configuration allows ground loss resistance to be eliminated from the calculations.

CONTINUOUSLY LOADED ANTENNAS

The design of high-Q air core inductors for RF work is complicated by the number of parameters that must be optimized simultaneously. One of these factors affecting coil Q adversely is radiation from a discrete loading coil. Therefore, the possibility of cutting down other losses while incorporating the coil radiation into that from the rest of the antenna system is an attractive one.

The general approach has been to use a coil made from heavy wire (#14 or larger), with length-to-diameter ratios as high as 21. British experimenters have reported good results with 8-foot overall lengths on the 1.8- and 3.5-MHz bands. The idea of making the entire antenna out of one section of coil has also been tried with some success. This technique is referred to as linear loading. Further information on linear-loaded antennas can be found in Chapter 6.

While going to extremes trying to find a perfect loading arrangement may not improve antenna performance very much, a poor system with lossy coils and high-resistance connections must be avoided if a reasonable signal is to be radiated.

MATCHING TO THE TRANSMITTER

Most modern transmitters require a 50-Ω output load, and because the feed-point impedance of a mobile whip is quite low, a matching network is usually necessary. Although calculations are helpful in the initial design, considerable experimenting is often necessary in final tune-up. This is particularly true for the lower bands, where the antenna is electrically short compared with a quarter-wave whip. The reason is that the loading coil is required to tune out a very large capacitive reactance, and even small changes in component values result in large reactance variations. Since the feed-point resistance is low to begin with, the problem is even more aggravated.

You can transform the low resistance of the whip to a value suitable for a 50-Ω system with an RF transformer or with a shunt-feed arrangement, such as an L network. The latter may only require a shunt coil or shunt capacitor at the base of the whip, since the net series capacitive or inductive reactance of the antenna and its loading coil may be used as part of the network. The following example illustrates the calculations involved.

Assume that a center-loaded whip antenna, 8.5 feet in overall length, is to be used on 7.2 MHz. From Table 1, earlier in this chapter, we see that the feed-point resistance of the antenna will be approximately 19 Ω, and from Fig 5 that the capacitance of the whip, as seen at its base, is approximately 24 pF. Since the antenna is to be center loaded, the capacitance value of the section above the coil will be cut approximately in half, to 12 pF. From this, it may be calculated that a center-loading inductor of 40.7 µH is required to resonate the antenna, that is, to cancel out the capacitive reactance. (This figure agrees with the approximate value of 40 µH shown in Table 1. The resulting feed-point impedance would then be 19 + *j* 0 Ω—a good match, *if* one happens to have a supply of 19-Ω coax.)

Solution: The antenna can be matched to a 52-Ω line such as RG-8 by tuning it either above or below reso-

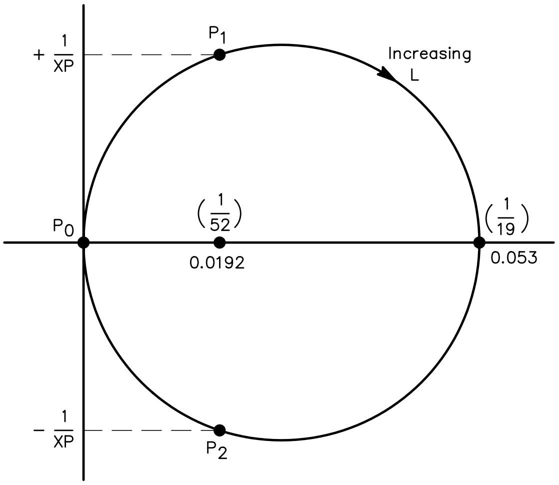

Fig 19—Admittance diagram of the RLC circuit consisting of the whip capacitance, radiation resistance and loading coil discussed in text. The horizontal axis represents conductance, and the vertical axis susceptance. The point P_0 is the input admittance with no whip loading inductance. Points P1 and P2 are described in the text. The conductance equals the reciprocal of the resistance, if no reactive components are present. For a series RX circuit, the conductance is given by

$$G = \frac{R}{R^2 + X^2}$$

and the susceptance is given by

$$B = \frac{-X}{R^2 + X^2}$$

Consequently, a parallel equivalent GB circuit of the series RX one can be found which makes computations easier. This is because conductances and susceptances add in parallel the same way resistances and reactances add in series.

nance and then canceling out the undesired component with an appropriate shunt element, capacitive or inductive. The way in which the impedance is transformed up can be seen by plotting the admittance of the series RLC circuit made up of the loading coil, antenna capacitance, and feed-point resistance. Such a plot is shown in **Fig 19** for a constant feed-point resistance of 19 Ω. There are two points of interest, P1 and P2, where the input conductance is 19.2 millisiemens, corresponding to 52 Ω. The undesired susceptance is shown as $1/X_P$ and $-1/X_P$, which must be canceled with a shunt element of the opposite sign, but with the same magnitude. The value of the canceling shunt reactance, X_P, may be found from the formula:

$$X_p = \frac{R_f\ Z_0}{\sqrt{R_f\left(Z_0 - R_f\right)}} \qquad \text{(Eq 21)}$$

where X_p is the reactance in Ω, R_f is the feed-point resistance, and Z_0 is the feed-line impedance. For Z_0 = 52 Ω and R_f = 19 Ω, $X_p = \pm 39.5$ Ω. A coil or good quality mica capacitor may be used as the shunt element. With the tune-up procedure described later, the value is not critical, and a fixed-value component may be used.

To arrive at point P1, the value of the center loading-coil inductance would be less than that required for resonance. The feed-point impedance would then appear capacitive, and an inductive shunt matching element would then be required. To arrive at point P2, the center loading coil should be more inductive than required for resonance, and the shunt element would need to be capacitive.

The value of the center loading coil required for the shunt-matched and resonated condition may be determined from the equation:

$$L = \frac{10^6}{4\pi^2 f^2 C} \pm \frac{X_S}{2\pi f} \qquad \text{(Eq 22)}$$

where addition is performed if a capacitive shunt is to be used, or subtraction performed if the shunt is inductive, and where L is in µH, f is the frequency in MHz, C is the capacitance of the antenna section being matched in pF, and

$$X_S = \sqrt{R_f\left(Z_0 - R_f\right)} \qquad \text{(Eq 23)}$$

For the example given, where Z_0 = 52 Ω, R_f = 19 Ω, f = 7.2 MHz, and C = 12 pF, X_S is found to be 25.0 Ω. The required antenna loading inductance is either 40.2 µH or 41.3 µH, depending on the type of shunt. Various matching possibilities for this example are shown in **Fig 20**. At A, the antenna is shown as tuned to resonance with L_L, a 40.7 µH coil, but with no provisions included for matching the resulting 19-Ω impedance to the 52-Ω line. At B, L_L has been reduced to 40.2 µH to make the antenna appear net capacitive, and L_M, having a reactance of 39.5 Ω, is added in shunt to cancel the capacitive reactance and transform the feed-point impedance to 50 Ω. The arrangement at C is similar to that at B except that L_L has been increased to 41.3 µH, and C_M (a shunt capacitor having a negative reactance of 39.5 Ω) is added, which also results in a 52-Ω nonreactive termination for the feed line.

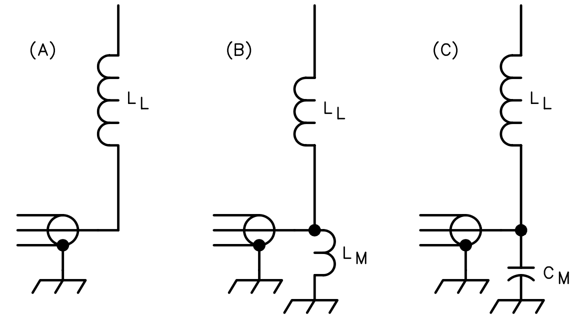

Fig 20—At A, a whip antenna that is resonated with a center loading coil. At B and C, the value of the loading coil has been altered slightly to make the feed-point impedance appear reactive, and a matching component is added in shunt to cancel the reactance. This provides an impedance transformation to match the Z_0 of the feed line. An equally acceptable procedure, rather than altering the loading coil inductance, is to adjust the length of the top section above the loading coil for the best match, as described in the tune-up section of the text.

The values determined for the loading coil in the above example point out an important consideration concerning the matching of short antennas—relatively small changes in values of the loading components will have a greatly magnified effect on the matching requirements. A change of less than 3% in the loading coil inductance value necessitates a completely different matching network! Likewise, calculations show that a 3% change in antenna capacitance will give similar results, and the value of the precautions mentioned earlier becomes clear. The sensitivity of the circuit with regard to frequency variations is also quite critical, and an excursion around practically the entire circle in Fig 19 may represent only 600 kHz, centered around 7.2 MHz, for the above example. This is why tuning up a mobile antenna can be very frustrating unless a systematic procedure is followed.

Tune-Up

Assume that inductive shunt matching is to be used with the antenna in the previous example, Fig 20B, where 39.5 Ω is needed for L_M. This means that at 7.2 MHz, a coil of 0.87 µH will be needed across the whip feed-point terminal to ground. With a 40-µH loading coil in place, the adjustable whip section above the loading coil should be set for minimum height. Signals in the receiver will sound weak and the whip should be lengthened a bit at a time until signals start to peak. Turn the transmitter on and check the SWR at a few frequencies to find where a minimum occurs. If it is below the desired frequency, shorten the whip slightly and check again. It should be moved approximately $^1/_4$ inch at a time until the SWR is minimum at the center of the desired range. If the frequency where the minimum SWR occurs is above the desired frequency, repeat the procedure above, but lengthen the whip only slightly.

If a shunt capacitance is to be used, as in Fig 20C, a value of 560 pF would correspond to the required –39.5 Ω of reactance at 7.2 MHz. With a capacitive shunt, start with the whip in its longest position and shorten it until signals peak up.

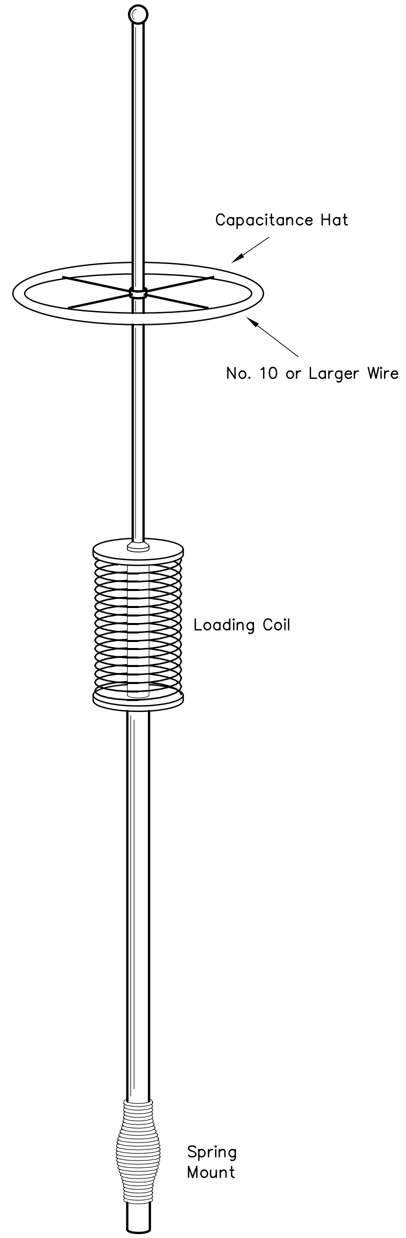
Capacitance Hat
No. 10 or Larger Wire
Loading Coil
Spring
Mount

Fig 21—A capacitance hat can be used to improve the performance of base- or center-loaded whips. A solid metal disc can be used in place of the skeleton disc shown here.

TOP-LOADING CAPACITANCE

Because the loss resistance varies with the inductance of the loading coil, the resistance can be reduced by removing turns from the coil. This must be compensated by adding capacitance to the portion of the mobile antenna that is *above* the loading coil (**Fig 21**). *Capacitance hats*, as they are called, can consist of a single stiff wire, two wires or more, or a disc made up of several wires like the spokes of a wheel. A solid metal disc could also be used, but is less practical for mobile work. The larger the capacitance hat (physically), the greater is the capacitance. The greater the capacitance, the less is the inductance required for resonance at a given frequency.

Capacitance-hat loading is applicable in either base-loaded or center-loaded systems. Since more inductance is required for center-loaded whips to make them resonant at a given frequency, capacitance hats are particularly useful in improving their efficiency.

TAPPED-COIL MATCHING NETWORK

Some of the drawbacks of the L-network can be eliminated by the use of the tapped-coil arrangement shown in **Fig 22**. Tune-up still remains critical, however, although somewhat more straightforward than for an L-network.

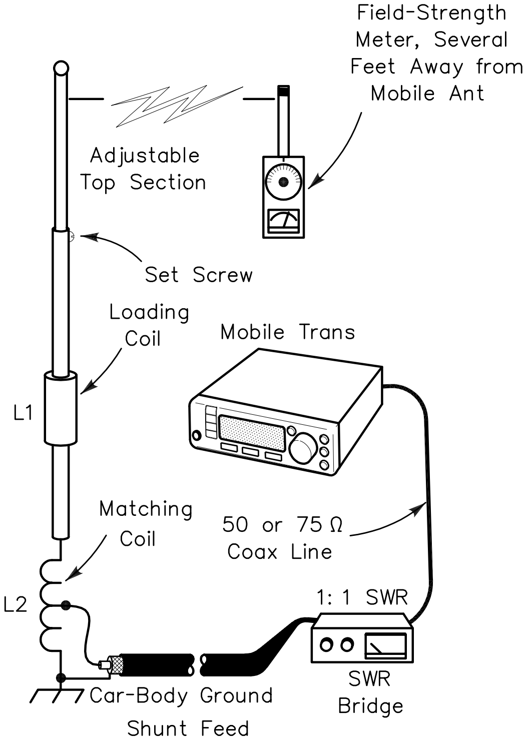
Field-Strength
Meter, Several
Feet Away from
Mobile Ant
Adjustable
Top Section
Set Screw
Loading
Coil
Mobile Trans
L1
Matching
Coil
50 or 75 Ω
Coax Line
L2
1: 1 SWR
SWR
Bridge
Car-Body Ground
Shunt Feed

Fig 22—A mobile antenna using shunt-feed matching. Overall antenna resonance is determined by the combination of L1 and L2. Antenna resonance is set by pruning the turns of L1, or adjusting the top section of the whip, while observing the field-strength meter or SWR indicator. Then adjust the tap on L2 for the lowest SWR.

Coil L2 can be inside the car body, at the base of the antenna, or at the base of the whip. As L2 helps determine the resonance of the antenna, L1 should be tuned to resonance in the desired part of the band with L2 in the circuit. The top section of the whip can be telescoped until a field-strength maximum is found. The tap on L2 is then adjusted for the lowest reflected power. Repeat these two adjustments until no further increase in field strength can be obtained; this point should coincide with the lowest SWR. The number of turns needed for L2 will have to be determined experimentally.

THE "SCREWDRIVER" MOBILE ANTENNA

Imagine QSYing from the bottom of 80 meters to the top of 10 meters, right from the driver's seat. With this antenna you will enjoy a very high Q antenna that has no taps or external adjustments, and that exhibits an SWR under 1.5:1 on all bands. Max Bloodworth, KO4TV, described this antenna in *The ARRL Antenna Compendium, Vol 7*.

The Screwdriver type of antenna was the brainchild of Don Johnson, W6AAQ, who developed it after many years of experimenting with mobile antennas. **Fig 23** is a photo of the completed antenna mounted on KO4TV's truck.

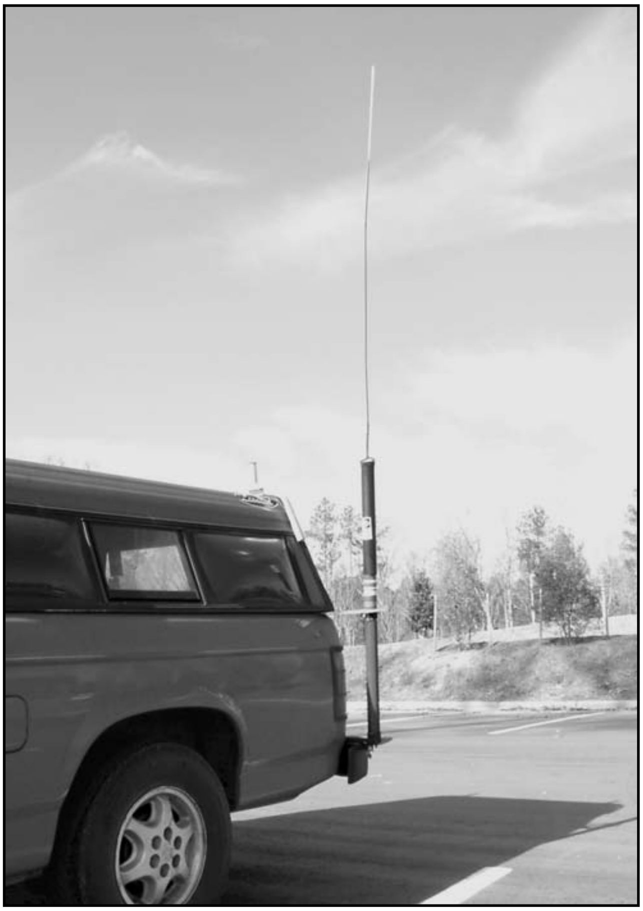

Fig 23—The completed Screwdriver antenna mounted on KO4TV's truck rear bumper. (*Photo courtesy of Gary Pearce, KN4AQ.*)

THE LOADING COIL

The general concept for the Screwdriver is that of a center-loaded antenna with an adjustable loading coil. This is an old idea; however, most previous multi-band antennas required multiple coil taps and an adjustable top whip section, as well as an impedance-matching unit at the base.

The problem of tapping a coil is well known. If you leave one end of the coil open, you have a miniature Tesla Coil, which can cause corona discharge and arcing. If you short the turns, the Q of the loading coil usually drops drastically. The Screwdriver is remarkable in that it does not have any coil taps and yet it can cover a wide range of frequencies with very high Q.

The secret is in the manner of adjusting the coil. It simply slides up or down into a metal tube (the base section), and makes contact with the top of the tube by means of "finger stock," which is made of spring-like Beryllium Copper. The coil is pushed up or down by an ordinary cordless screwdriver; hence the name, Screwdriver. This turns a section of threaded rod and moves the coil form up or down. The section of coil above the finger stock is the active loading coil, and the part of the coil just below the finger stock simply "disappears" into the base tube and is totally out of the circuit. No taps, no loose ends—just a high-Q coil in the middle of two solid metal tubes. Although this antenna is available ready-made from several commercial sources, making one is well within the capabilities of any ham who is relatively handy with ordinary hand tools, and who has access to either a small lathe or a common $1^1/_4$-inch pipe die.

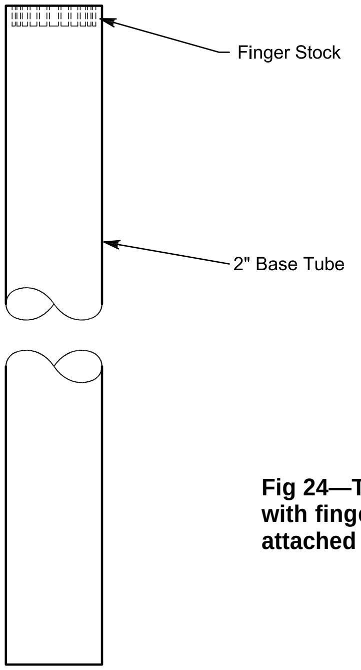

Fig 24—The base tube, with finger stock attached to the top.

CONSTRUCTING YOUR SCREWDRIVER ANTENNA

Construction begins by locating a 2-inch inside-diameter tube, about 3 to 3$^1/_2$ feet long. See **Fig 24**. The author had used such diverse materials as:

- Aluminum irrigation pipe
- Schedule-20 copper tubing
- A stainless-steel hydraulic cylinder
- A section of aluminum from a 6-meter beam that was damaged by a tornado
- And even a brass bedpost salvaged from the local garbage dump.

The brass bedpost looked especially good on his vehicle, which was Burgundy with gold trim. Whatever material you use for the bottom tube, you can either leave it unfinished or painted to your taste.

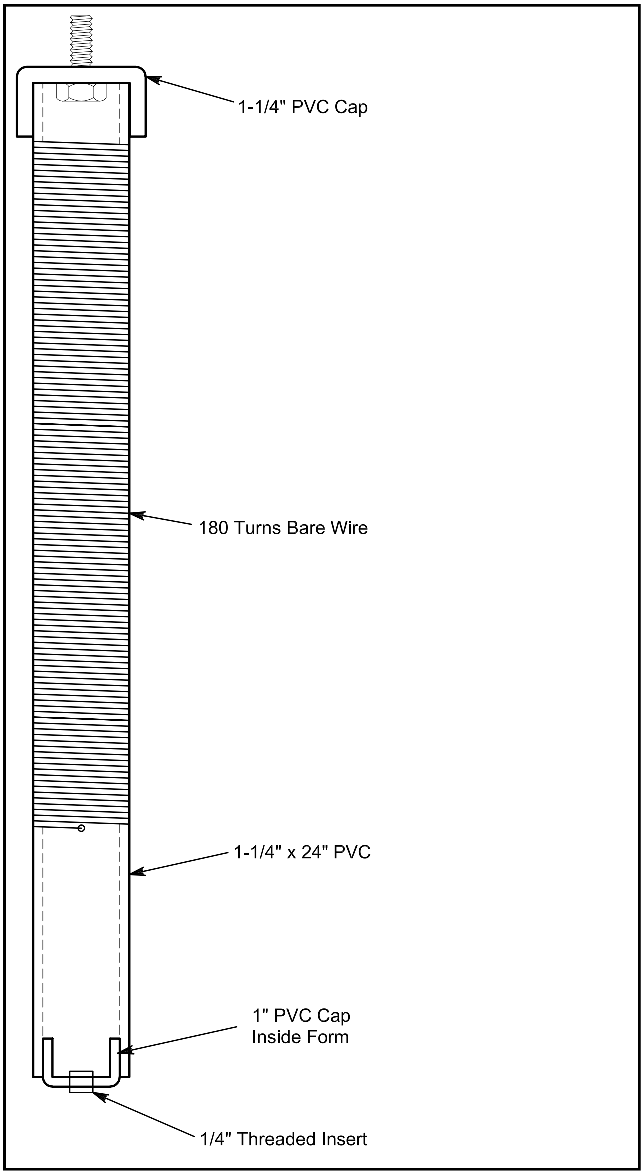

Fig 25—The coil form, with a threaded insert inserted into the bottom PVC cap inside the coil form.

The other essential item is a cordless screwdriver, minus its batteries and switch. You can often find one at yard sales or flea markets for a dollar or so, usually with dead batteries, since it is often just as cheap to buy another one new as it is to buy new batteries. So long as it will fit snugly inside the base tube, the brand name is immaterial. KO4TV has used Skil, Black & Decker, or even a Wal-Mart $8 special, all with equal success. The most difficult part of the job will be making the coil form and winding the loading coil, but with a little patience you should be able to do this satisfactorily.

Bear in mind that it is *not* a "Heathkit" type project, and it will require some innovation and ingenuity on your part. Next, obtain a 2-foot long piece of 1$^1/_4$-inch PVC pipe (*not* CPVC, which is vulnerable to ultraviolet rays from the sun). With either a lathe or a pipe die, thread approximately 20 inches from one end, at 10 to 12 threads per inch (See **Fig 25**). If you use a pipe die, it helps to loosen the cutters slightly, to make the grooves shallower.

When using a lathe, temporarily insert a piece of 1-inch water pipe inside the coil form to hold it steady in the lathe. The grooves should be just deep enough to comfortably hold the wire in place when it is wound. The choice of coil wire is up to you. The author has used #16 or #18 tinned copper, bare copper, or best yet, 17-gauge aluminum electric fence wire. This is available from any

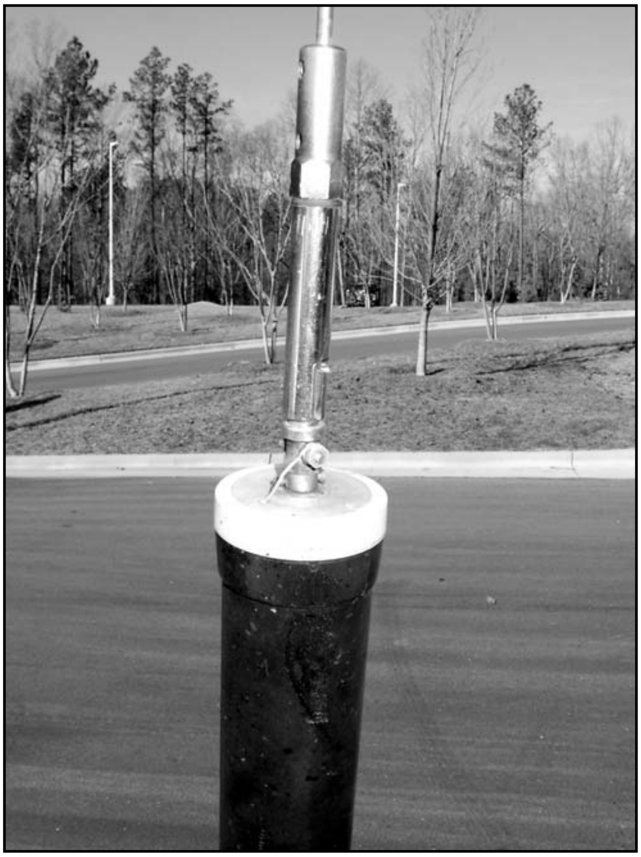

Fig 26—Photo of the top of the loading coil. (*Photo courtesy of Gary Pearce, KN4AQ.*)

farm supply store at a very reasonable cost. It will take about 65 to 70 feet of wire.

Begin by drilling a $^{1}/_{16}$-inch hole through the PVC pipe at the bottom end of the threaded portion. Slip the end of the wire through this hole and tie a knot in it to serve as a stop. You could also insert a nut and bolt through a loop in the wire. Then carefully wind the coil to within an inch of the end of the threaded portion. The preferred method of winding is to put a stick or pipe through the holes of the wire reel and hold it between your feet and the floor, to provide the necessary tension to allow tight winding.

About an inch below the top of the coil, drill a $^{1}/_{16}$-inch hole and thread about 6 or 8 inches of wire through it. This will be the top connection from the coil to the whip. The bottom of the coil has no electrical connection.

The method of attaching the whip to the top of the coil form is up to you. I usually drill a $^{3}/_{8}$-inch hole through a $1^{1}/_{4}$-inch PVC pipe cap and place a $^{3}/_{8}$ × $1^{1}/_{2}$-inch SAE bolt through the hole, with a matching washer and nut on top. The bolt can be drilled and tapped for mounting the whip, which should be about 5 to 6 feet long. (It will be pruned for resonance later.) See **Fig 26**, a photo showing the top of the coil.

A more elegant way of connecting the whip is to obtain a swivel or quick-disconnect antenna fitting from your nearest truck stop. These are widely available for use with CB antennas and are threaded $^{3}/_{8}$-inch SAE. Do not attach the cap to the coil form until after the coil is inserted in the base tube.

Now that you have a coil form and a base tube, the next step will be to plug the bottom of the PVC coil form with a 1-inch PVC pipe cap, which fits snugly inside the coil form. Drill a hole directly in the center of this cap, and install a $^{1}/_{4}$-20 threaded insert, available at most hardware stores for about 25 or 30 cents. These are commonly used in wooden furniture to provide a thread for attaching bolts to wood. See **Fig 27**.

It is imperative that you install this insert squarely. The best method is to screw a short bolt into it and chuck the bolt in a drill press, placing the pipe cap

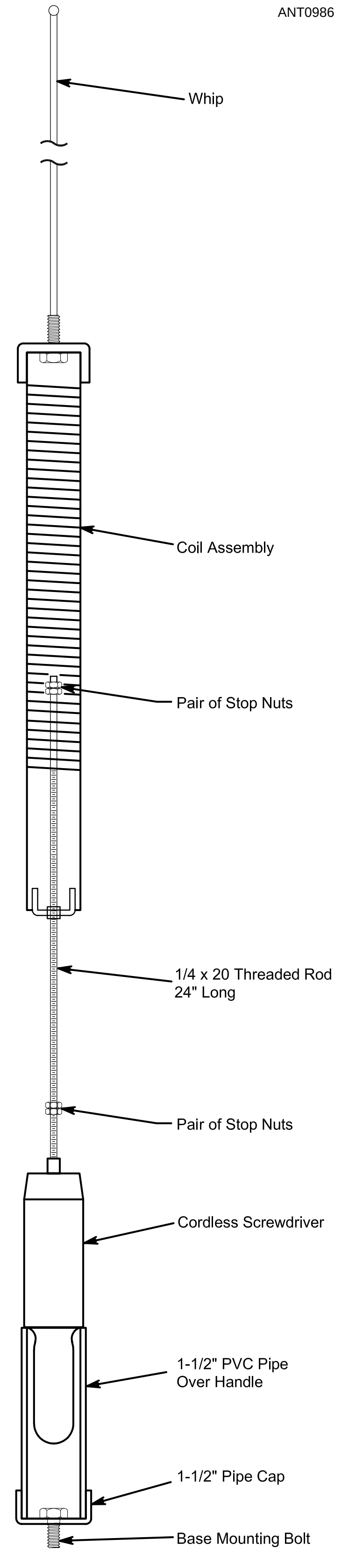
ANT0986
Whip
Coil Assembly
Pair of Stop Nuts
1/4 x 20 Threaded Rod
24" Long
Pair of Stop Nuts
Cordless Screwdriver
1-1/2" PVC Pipe
Over Handle
1-1/2" Pipe Cap
Base Mounting Bolt

Fig 27—The coil assembly and cordless screwdriver drive assembly.

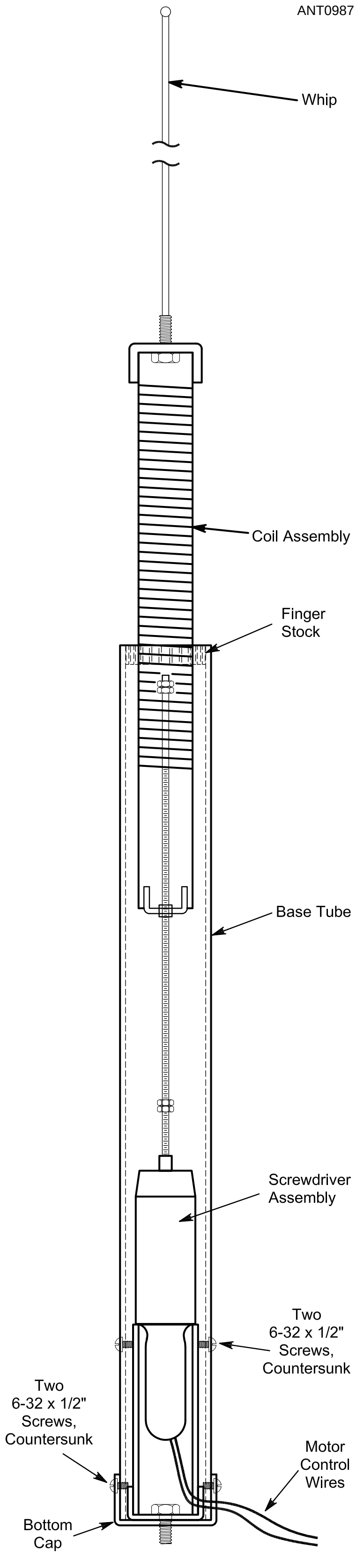
ANT0987
Whip
Coil Assembly
Finger
Stock
Base Tube
Screwdriver
Assembly
Two
6-32 x 1/2"
Screws,
Countersunk
Two
6-32 x 1/2"
Screws,
Countersunk
Motor
Control
Wires
Bottom
Cap

Fig 28—The complete antenna assembly.

squarely on the drill press table. Carefully turn the drill chuck by hand until the insert is properly seated.

The Motor

Now prepare the cordless drill by removing the dead batteries and the switch. Install a pair of 1 or 1.5-Ω, 10-W resistors in series with each motor lead. Connect a 0.1 µF 50-V disc ceramic capacitor across the motor leads to suppress motor noise.

Connect a pair of wires, about a foot long, to the other ends of the resistors. These will power the screwdriver from your car's 12-V system. Next, procure a 20-inch long piece of $^{1}/_{4}$-20 threaded rod (also called "all-thread"). Insert one end of this into the cordless screwdriver chuck, and drill a $^{1}/_{16}$-inch or $^{3}/_{32}$-inch hole through both the chuck and the rod. Secure the rod in place with either a roll or cotter pin, and install a pair of $^{1}/_{4}$-inch nuts on the rod near the screwdriver chuck. This will be the lower stop for the coil travel and will be adjusted later.

Thread the rod into the threaded insert at the bottom of the coil form, and install another pair of stop nuts on the top side of the rod. It will be easier to do this before attaching the bottom plug to the coil form. Again, these nuts will be adjusted later for proper coil travel.

Place the bottom plug into the coil form, drill a couple of holes through the form and the plug and tap them for 6-32 flat head bolts. Countersink the form so that the heads will be flush with the outside. Again, see **Fig 27**.

Finger Stock

Now, let's go back to the base tube. You must install a circle of finger-stock strip on the top of the tube to make contact with the coil, as shown in Fig 24. This can either be soldered or riveted, depending on your choice of tube material. Next, insert the coil and screwdriver assembly through the bottom of the tube, making sure that the coil clears the finger stock without deforming or bending it.

The best method I have found for attaching the screwdriver to the bottom tube is to fit a piece of 1$^{1}/_{2}$-inch PVC pipe, about 8 or 10 inches long, over the handle of the screwdriver and attach it with a couple of flat-head 6-32 bolts. You may have to either grind some material off the handle or wrap it with some duct tape to make a snug fit inside the 1$^{1}/_{2}$-inch PVC pipe, which should slip snugly into the base tube.

Place a matching 1$^{1}/_{2}$-inch PVC pipe cap on the bottom of the pipe, and secure it with a couple of 6-32 screws. Drill a $^{1}/_{4}$-inch hole through one side of the cap, which will serve to pass the wires from the screwdriver motor. Drill a $^{3}/_{8}$-inch hole directly in the center of this cap for the base mount. See **Fig 28**, which is a drawing of the completed antenna assembly.

MOUNTING THE ANTENNA

At this point you must determine exactly how you wish to mount the antenna on your vehicle. You could fasten it to a standard mobile antenna mount, but *do not* use a spring! I prefer to fasten the assembly to a thick metal plate about 6 × 18 inches in size. I mount this to the frame of the car, protruding about 4 or 5 inches from the lower car body, just behind the rear wheel. You could also fasten such a plate to the lower fender sheet metal with large sheet-metal screws. The SO-239 connector is mounted to this plate also. See **Fig 29**.

An upper bracket, made from Plexiglas or similar insulating material, can be mounted from the trunk lip with an L-angle and used to support the top of the lower tube. See **Fig 30**. With a Plexiglas bracket about 4-inches wide, a 2-inch hole can be drilled in it to pass the tube. This will make a snug and rigid fit at the top. Under no circumstances use a metal band around the tube, as this will form a shorted turn and drastically lower the antenna Q.

With a little care, this type of mounting will leave no visible holes in the vehicle, which will likely enhance its trade-in value. If you mount the antenna on a pickup truck, it can be mounted directly on the rear safety bumper with a 1$^{1}/_{2}$ inch pipe floor flange and matching adaptor on the bottom of the tube.

Electrical Connections

After mounting the base, your next job is to connect the wires from the screwdriver to a DPDT, spring-return center-off switch, mounted in a convenient location in your car. See the schematic in **Fig 31**. Connect the coax from the radio to the base tube, with the center conductor going to the base tube and the shield connected to ground at the base of the antenna.

If you like, you could install a simple base-matching network to give a perfect match on the lower bands. Even with no matching unit, SWR is usually under 2:1 on all

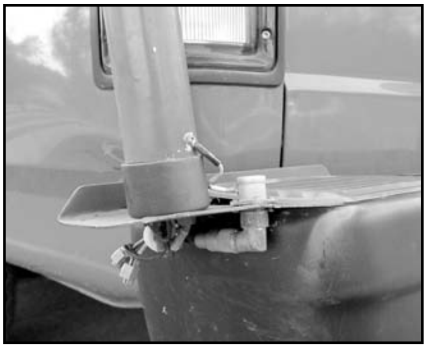

Fig 29—Base-mounting plate for Screw-driver installation on KO4TV's truck. Note the quick-disconnect dc motor con-nections and the SO-239 coax connector.

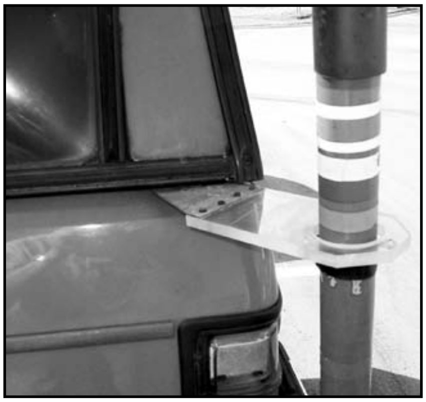

Fig 30—The support Plexiglas plate. (*Photo courtesy of Gary Pearce, KN4AQ.*)

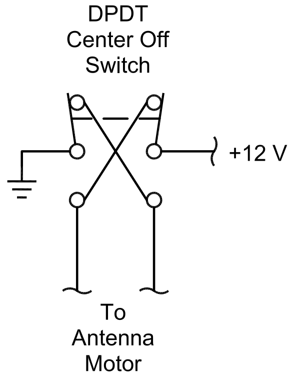

Fig 31—The DPDT center-off spring-loaded switch used to control the screwdriver motor.

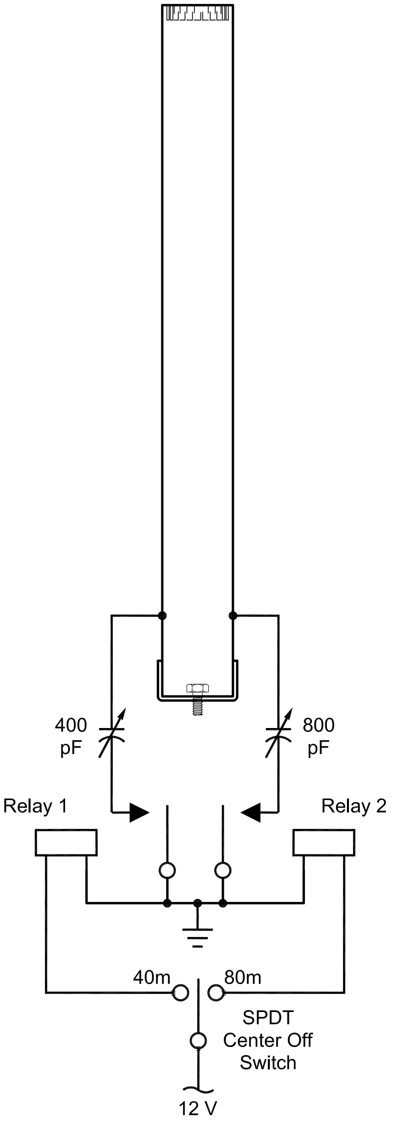

Fig 32—Relay-switched matching capacitors for 40 and 80 meters.

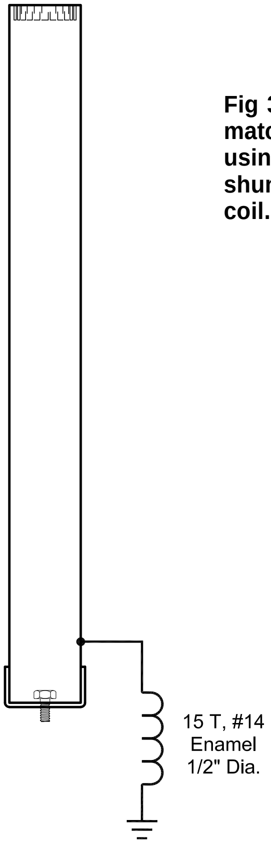

Fig 33—Alternative matching scheme using a single shunt matching coil.

bands. See **Fig 32**. One matching network consists of a couple of trimmer capacitors switched from the bottom of the base tube to ground using a pair of small relays. For 80 meters, approximately 800 pF is required, and for 40 meters about 400 pF. No matching is required for 30 meters or higher.

An alternative matching method is to install a 15-turn coil made from #14 enamel wire, about $^1/_2$ inch diameter, from the base to ground. See **Fig 33**. In either case, adjust the whip length for proper resonance at both ends of the HF spectrum over the range of movement of your coil.

This antenna may have different matching requirements, depending on the vehicle type and mounting location, so you may need to do some experimentation to obtain optimum results. If it doesn't work exactly right at first, be prepared to do a little experimenting.

A COIL COVER

Most builders will want to provide a coil cover, both for protection and appearance. KO4TV's favorite cover is made from a plastic mailing tube, about 24 to 30 inches long and about $^1/_2$ inches in diameter. You can find these at an office supply store. They have a screw-on cap on one end, which can be mounted to the top of the loading coil.

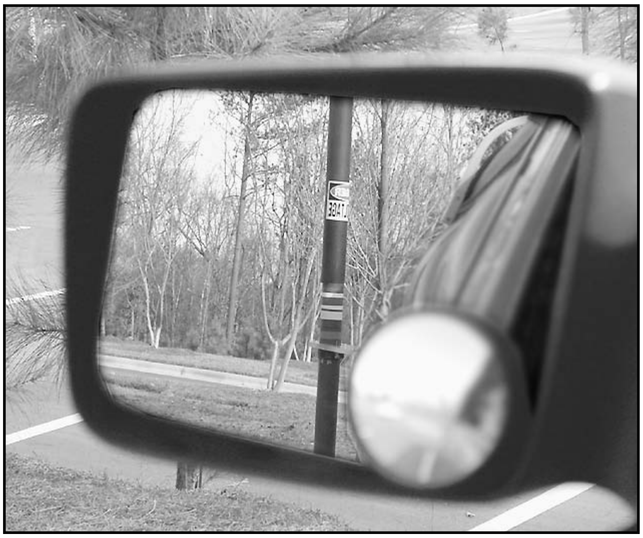

Fig 34—Remote control by the rear-view mirror—tuning by the stripes! (*Photo courtesy of Gary Pearce, KN4AQ.*)

You trim the bottom of the cover to the proper length to allow the coil to be fully retracted into the base tube while still clearing the top bracket. Other builders have used such diverse materials as:

- Several 1-liter soft drink bottles cut off, swaged and glued together
- Sections of Schedule 20 PVC pipe with matching cap
- Empty beef jerky tubes (usually available for the asking at convenience stores)
- Wands from a shop vacuum cleaner, fitted with the cap from an aerosol spray can.

Just let your imagination run wild and you may be surprised at your own ingenuity. Whatever material you do use, paint it to match your vehicle. Just be sure to use non-metallic paint.

TUNING BY THE STRIPES

The author used different colored vinyl tapes to mark the tuning points for the amateur bands along the tuning coil. The lower edge of the supporting Plexiglas middle support acts as a pointer, which he can spot from his rear-view mirror. See **Fig 34**. This may be low-tech tuning method, but it works!

THE PROOF IS IN THE PUDDING

Now, you are ready to enjoy some real mobile communications on all bands, without even stopping the vehicle to QSY. KO4TV's antenna has contacted over 100 DX countries, from as far away as Tasmania and Japan from the East Coast. He enjoys regular contacts with Israel and other stations in the Mid-East, running 100 W or less.

A MOBILE J ANTENNA FOR 144 MHz

The J antenna is a mechanically modified version of the Zepp (Zeppelin) antenna. It consists of a half-wave-length radiator fed by a quarter-wave matching stub. This antenna exhibits an omnidirectional pattern with little high-angle radiation, but does not require the ground plane that $^1/_4$-wave and $^5/_8$-wave antennas do to work properly. The material in this section was prepared by Domenic Mallozzi, N1DM, and Allan White, W1EYI.

Fig 35 shows two common configurations of the J antenna. Fig 35A shows the shorted-stub version that is usually fed with 200- to 600-Ω open-wire line. Some have attempted to feed this antenna directly with coax without a balun, and this usually leads to less than optimum results. Among the problems with such a configuration are a lack of reproducibility and heavy coupling with nearby objects. To eliminate these problems, many

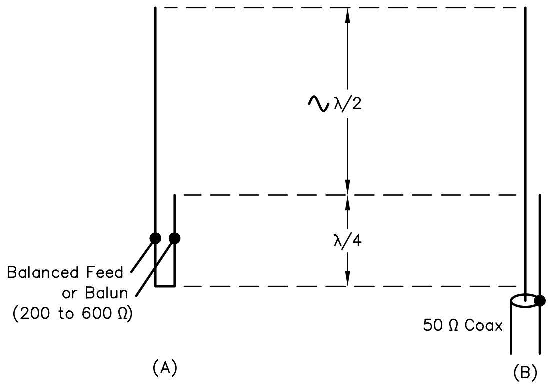

Fig 35—Two configurations of the J antenna.

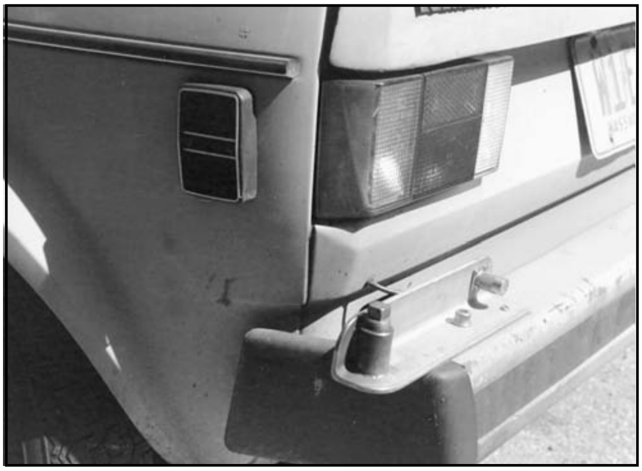

Fig 36—The mount for the mobile J is made from stainless-steel angle stock and secured to the bumper with stainless-steel hardware. Note the $^1/_2$-inch pipe plug and a PL-259 (with a copper disc soldered in its unthreaded end). These protect the mount and connector threads when the antenna is not in use.

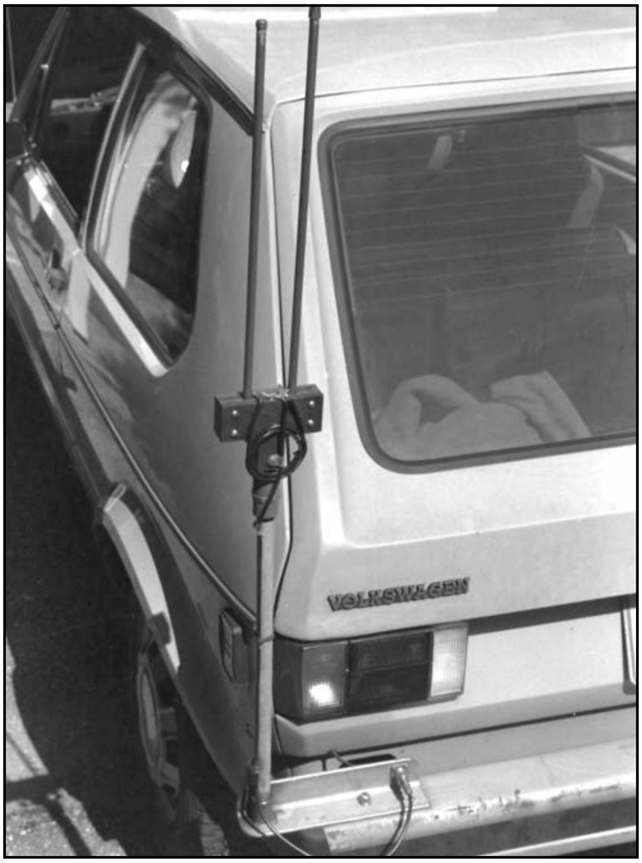

Fig 37—The J antenna, ready for use. Note the bakelite insulator and the method of feed. Tie wraps are used to attach the balun to the mounting block and to hold the coax to the support pipe. Clamps made of flashing copper are used to connect the balun to the J antenna just above the insulating block. The ends of the balun should be weatherproofed.

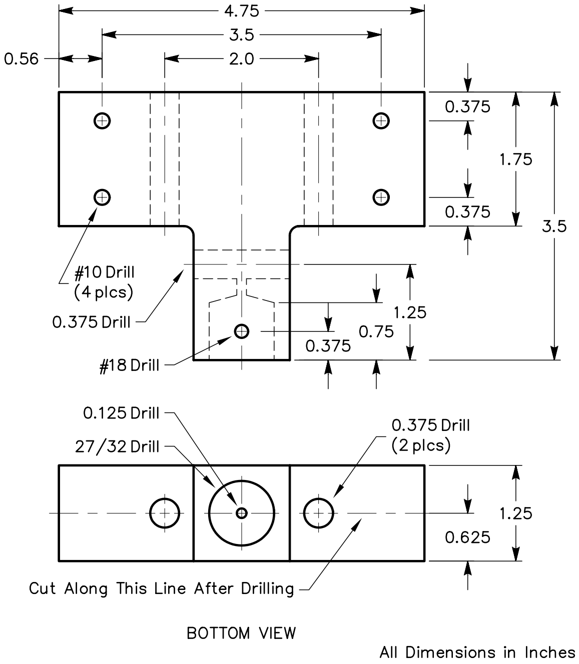

Fig 38—Details of the insulated mounting block. The material is bakelite.

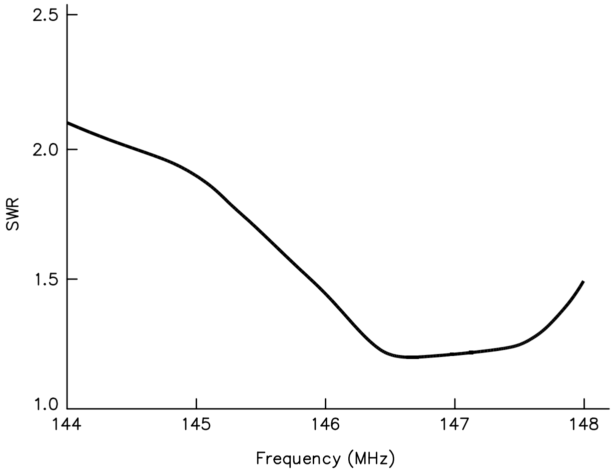

Fig 39—Measured SWR of the mobile J antenna.

amateurs have used a 4:1 half-wave balun between the feed point and a coaxial feed line. This simple addition results in an antenna that can be easily reproduced and that does not interact so heavily with surrounding objects. The bottom of the stub may be grounded (for mechanical or other reasons) without impairing the performance of the antenna.

The open-stub-fed J antenna shown in Fig 35B can be connected directly to low-impedance coax lines with good results. The lack of a movable balun (which allows some impedance adjustment) may make this antenna a bit more difficult to adjust for minimum SWR, however.

The Length Factor

Dr. John S. Belrose, VE2CV, noted in *The Canadian Amateur* that the diameter of the radiating element is important to two characteristics of the antenna—its bandwidth and its physical length. (See Bibliography at the end of this chapter.) As the element diameter is increased, the usable bandwidth increases, while the physical length of the radiating element decreases with respect to the free-space half-wavelength. The increased diameter makes the end effect more pronounced, and also slows the velocity of propagation on the element. These two effects are related to resonant antenna lengths by a factor, “k.” This factor is expressed as a decimal fraction giving the equivalent velocity of propagation on the antenna wire as a function of the ratio of the element diameter to a wavelength. The k factor is discussed at length in Chapter 2.

The length of the radiating element is given by

$$\ell = \frac{5904 \times k}{f} \qquad \text{(Eq 25)}$$

where

ℓ = length in inches
f = frequency in MHz
k = k factor.

The k factor can have a significant effect. For example, if you use a $^{5}/_{8}$-inch diameter piece of tubing for the radiator at 144 MHz, the k value is 0.907 (9.3% shorter than a free-space half wavelength).

The J antenna gives excellent results for both mobile and portable work. The mobile described here is similar to an antenna described by W. B. Freely, K6HMS, in April 1977 *QST*. This design uses mechanical components that are easier to obtain. As necessary with all mobile antennas, significant attention has been paid to a strong, reliable, mechanical design. It has survived not only three New England winters, but also two summers of 370-mile weekend commutes. During this time, it has maintained consistent electrical performance with no noticeable deterioration.

The mechanical mount to the bumper is a 2 × 2-inch stainless steel angle iron, 10 inches long. It is secured to the bumper with stainless steel hardware, as shown in **Fig 36**. A stainless steel $^{1}/_{2}$-inch pipe coupling is welded to the left side of the bracket, and an SO-239 connector is mounted at the right side of the bracket. The bracket is mounted to the bumper so a vertical pipe inserted in the coupling will allow the hatchback of the vehicle to be opened with the antenna installed, **Fig 37**.

A $^{1}/_{2}$-inch galvanized iron pipe supports the antenna so the radiating portion of the J is above the vehicle roof line. This pipe goes into a bakelite insulator block, visible in Fig 37. The insulator block also holds the bottom of the stub. This block was first drilled and then split with a band saw, as shown in **Fig 38**. After splitting, the two portions are weatherproofed with varnish and rejoined with 10-32 stainless hardware. The corners of the insulator are cut to clear the L sections at the shorted end of the stub.

The quarter-wave matching section is made of $^{1}/_{4}$-inch type L copper tubing ($^{5}/_{16}$ inch ID, $^{3}/_{8}$ inch OD). The short at the bottom of the stub is made from two copper L-shaped sections and a short length of $^{1}/_{4}$-inch tubing. Drill a $^{1}/_{8}$-inch hole in the bottom of this piece of tubing to drain any water that may enter or condense in the stub.

A $^{5}/_{16}$-inch diameter brass rod, $1^{1}/_{2}$ to 2 inches long, is partially threaded with a $^{5}/_{16}$ × 24 thread to accept a Larsen whip connector. This rod is then sweated into one of the legs of the quarter-wave matching section. A 40-inch whip is then inserted into the Larsen connector.

The antenna is fed with 50-Ω coaxial line and a coaxial 4:1 half-wave balun. This balun is described at the end of Chapter 26. As with any VHF antenna, use high-quality coax for the balun. Seal all open cable ends and the rear of the SO-239 connector on the mount with RTV sealant.

Adjustment is not complicated. Set the whip so that its tip is 41 inches above the open end of the stub, and adjust the balun position for lowest SWR. Then adjust the height of the whip for the lowest SWR at the center frequency you desire. **Fig 39** shows the measured SWR of the antenna after adjustments are completed.

THE SUPER-J MARITIME ANTENNA

This 144-MHz vertical antenna doesn't have stringent grounding requirements and can be made from easy to find parts. The material in this section was prepared by Steve Cerwin, WA5FRF, who developed the Super-J for use on his boat.

Antennas for maritime use must overcome difficulties that other kinds of mobile antennas normally do not encounter. For instance, the transom of a boat is the logical place to mount an antenna. But the transoms of many boats are composed mostly of fiberglass, and they ride some distance out of the water—from several inches to a few feet, depending on the size of the vessel. Because the next best thing to a ground plane (the water surface) is more than an appreciable fraction of a wavelength away at 144 MHz, none of the popular gain-producing antenna designs requiring a counterpoise are suitable. Also, since a water surface does a good job of assuming the earth's lowest mean elevation (at least on a calm day), anything that can be done to get the radiating part of the antenna up in the air is helpful.

One answer is the venerable J-pole, with an extra in-phase half-wave section added on top—the Super-J antenna. The two vertical half waves fed in phase give outstanding omnidirectional performance for a portable antenna. Also, the J-pole feed arrangement provides the desired insensitivity to height above ground (or water) plus added overall antenna height. Best of all, a $^{1}/_{4}$-wave CB whip provides enough material to build the whole driven element of the antenna, with a few inches to spare. The antenna has enough bandwidth to cover the entire 144-MHz band, and affords a measure of lightning protection by being a grounded design.

Antenna Operation

The antenna is represented schematically in **Fig 40**. The classic J-pole antenna is the lower portion shown between points A and C. The half-wave section between points B and C does most of the radiating. The added half-wave section of the Super-J version is shown between points C and E. The side-by-side quarter-wave elements between points A and B comprise the J feed arrangement.

At first glance, counterproductive currents in the J section between points A and B may seem a waste of element material, but it is through this arrangement that

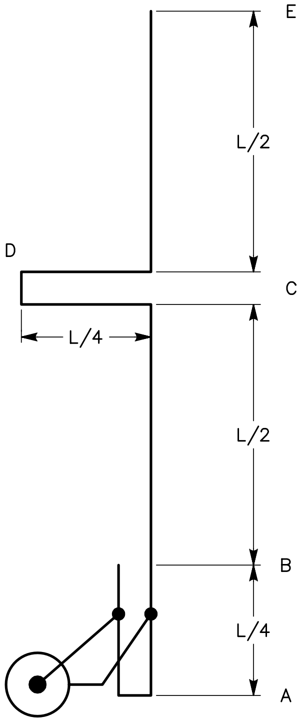

Fig 40—Schematic representation of the Super-J maritime antenna. The radiating section is two half waves in phase.

the antenna is able to perform well in the absence of a good ground. The two halves of the J feed arrangement, side by side, provide a loading mechanism regardless of whether or not a ground plane is present.

The radiation resistance of any antenna fluctuates as a function of height above ground, but the magnitude of this effect is small compared to the wildly changing impedance encountered when the distance from a ground plane element to its counterpoise is varied. Also, the J section adds $^{1}/_{4}$ wavelength of antenna height, reducing the effect of ground height variations even further. Reducing ground-height sensitivity is particularly useful in maritime operation on those days when the water is rough.

The gain afforded by doubling the aperture of a J-pole with the extra half-wave section can be realized only if the added section is excited in phase with the half-wave element B-C. This is accomplished in the Super-J in a conventional manner, through the use of the quarter-wave phasing stub shown between C and D.

Construction and Adjustment

The completed Super-J is shown in **Fig 41**. Details of the individual parts are given in **Fig 42**. The driven element can be liberated from a quarter-wave CB whip antenna and cut to the dimensions shown. All other metal stock can be obtained from metal supply houses or machine shops. Metal may even be scrounged for little or nothing as scraps or remnants, as were the parts for the antenna shown here.

The center insulator and the two J stub spacers are made of $^{1}/_{2}$-inch fiberglass and stainless steel stock, and the end caps are bonded to the insulator sections with epoxy. If you don't have access to a lathe to make the end caps, a simpler one-piece insulator design of wood or fiberglass could be used. However, keep in mind that good electrical connections must be maintained at all joints, and strength is a consideration for the center insulator.

The quarter-wave phasing stub is made of $^{1}/_{8}$-inch stainless steel tubing, **Fig 43**. The line comprising this stub is bent in a semicircular arc to narrow the vertical profile and to keep the weight distribution balanced. This makes for an attractive appearance and keeps the antenna from leaning to one side.

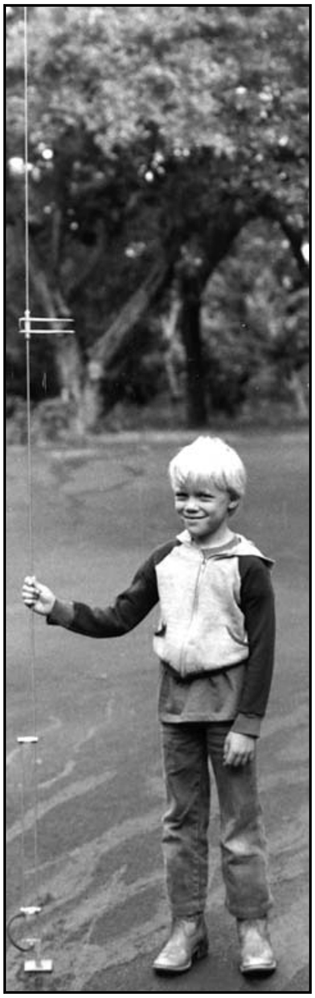

Fig 41—Andy and the assembled Super-J antenna.

The bottom shorting bar and base mounting plate are made of $^{1}/_{4}$-inch stainless steel plate, shown in **Fig 44**. The J stub is made of $^{3}/_{16}$-inch stainless-steel rod stock. The RF connector may be mounted on the shorting bar as shown, and connected to the adjustable slider with a short section of coaxial cable. RTV sealant should be used at the cable ends to keep out moisture. The all-stainless construction looks nice and weathers well in maritime mobile applications.

The antenna should work well over the whole 144-MHz band if cut to the dimensions shown. The only tuning required is adjustment of the sliding feed point for minimum SWR in the center of the band segment you use most. Setting the slider $2^{13}/_{16}$ inch above the top of the shorting bar gave the best match for this antenna and may be used for a starting point. Four turns of coax made into a coil at the feed point or a ferrite-sleeve balun act as a common-mode choke balun to ensure satisfactory performance.

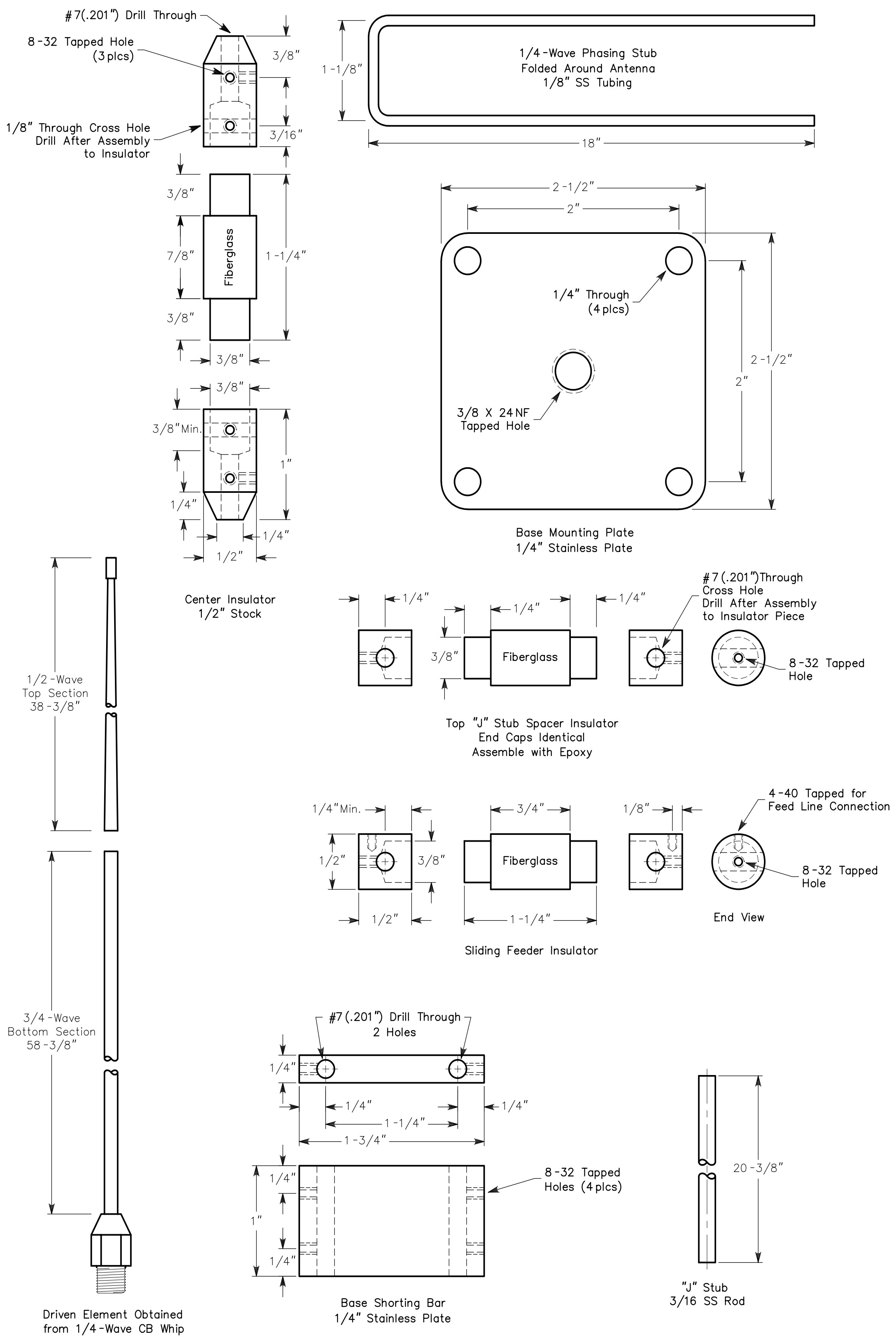
#7(.201") Drill Through
8-32 Tapped Hole (3 plcs)
3/8"
1/8" Through Cross Hole Drill After Assembly to Insulator
3/16"
3/8"
7/8"
Fiberglass
1-1/4"
3/8"
3/8"
3/8"
3/8"Min.
1"
1/4"
1/4"
1/2"
Center Insulator 1/2" Stock
1-1/8"
1/4-Wave Phasing Stub Folded Around Antenna 1/8" SS Tubing
18"
2-1/2"
2"
1/4" Through (4 plcs)
3/8 X 24NF Tapped Hole
2-1/2"
2"
Base Mounting Plate 1/4" Stainless Plate
1/2-Wave Top Section 38-3/8"
3/4-Wave Bottom Section 58-3/8"
Driven Element Obtained from 1/4-Wave CB Whip
1/4"
1/4"
1/4"
3/8"
Fiberglass
#7(.201")Through Cross Hole Drill After Assembly to Insulator Piece
8-32 Tapped Hole
Top "J" Stub Spacer Insulator End Caps Identical Assemble with Epoxy
1/4"Min.
3/4"
1/8"
4-40 Tapped for Feed Line Connection
1/2"
3/8"
Fiberglass
8-32 Tapped Hole
1/2"
1-1/4"
End View
Sliding Feeder Insulator
#7(.201") Drill Through 2 Holes
1/4"
1/4"
1/4"
1-1/4"
1-3/4"
1/4"
1"
1/4"
8-32 Tapped Holes (4 plcs)
Base Shorting Bar 1/4" Stainless Plate
20-3/8"
"J" Stub 3/16 SS Rod

Fig 42—Details of parts used in the construction of the 144-MHz Super-J. Not to scale.

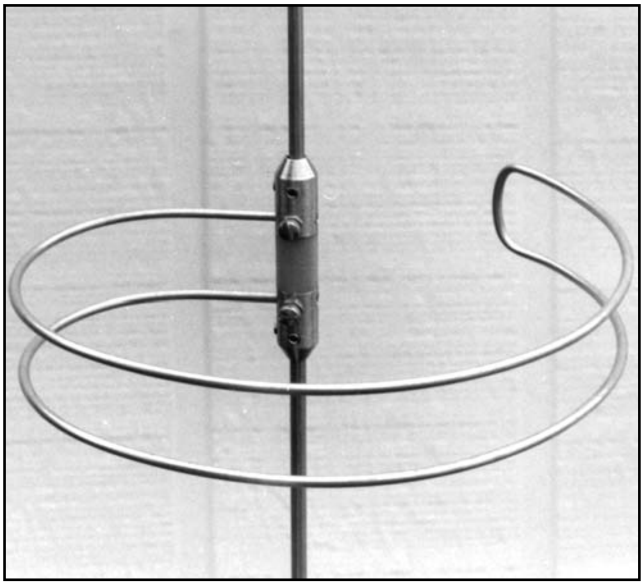

Fig 43—A close-up look at the $^1/_4$-λ phasing section of the Super-J. The insulator fitting is made of stainless-steel end caps and fiberglass rod.

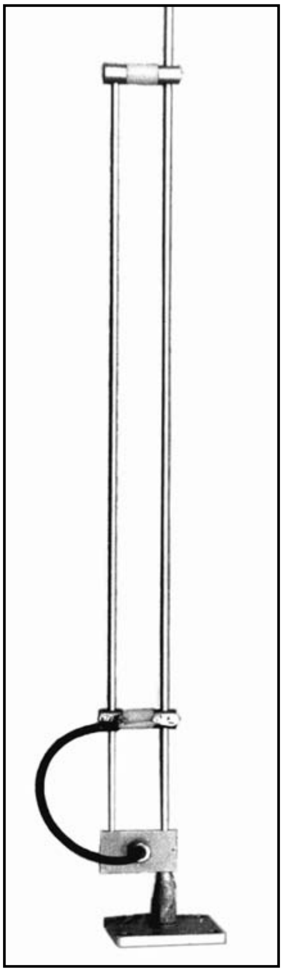

Fig 44—The bottom shorting bar and base mounting plate assembly.

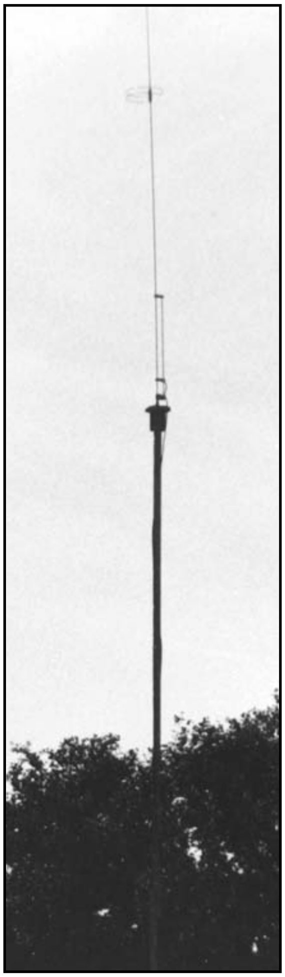

Fig 45—The Super-J in portable use at a field site.

Performance

Initial tests of the Super-J were performed in portable use and were satisfactory, if not exciting. **Fig 45** shows the Super-J mounted on a wooden mast at a portable site. Simplex communication with a station 40 miles away with a 10-W mobile rig was full quieting both ways. Stations were worked through distant repeaters that were thought inaccessible from this location.

Comparative tests between the Super-J and a commercial $^5/_8$-wave antenna mounted on the car showed the Super-J to give superior performance, even when the Super-J was lowered to the same height as the car roof. The mast shown in Fig 45 was made from two 8-foot lengths of 1 × 2-inch pine. (The two mast sections and the Super-J can be easily transported in most vehicles.)

The Super-J offers a gain of about 6 dB over a quarter-wave whip and around 3 dB over a $^5/_8$-wave antenna. Actual performance, especially under less-than-ideal or variable ground conditions, is substantially better than other vertical antennas operated under the same conditions. The freedom from ground-plane radials proves to be a real benefit in maritime mobile operation, especially for those passengers in the back of the boat with sensitive ribs!

A TOP-LOADED 144-MHz MOBILE ANTENNA

Earlier in this chapter, the merits of various loading schemes for shortened whip antennas were discussed. Quite naturally, one might be considering HF mobile operation for the application of those techniques. But the principles may be applied at any frequency. **Fig 46** shows a 144-MHz antenna that is both top and center loaded. This antenna is suitable for both mobile and portable operation, being intended for use on a handheld transceiver. This antenna was devised by Don Johnson, W6AAQ, and Bruce Brown, W6TWW.

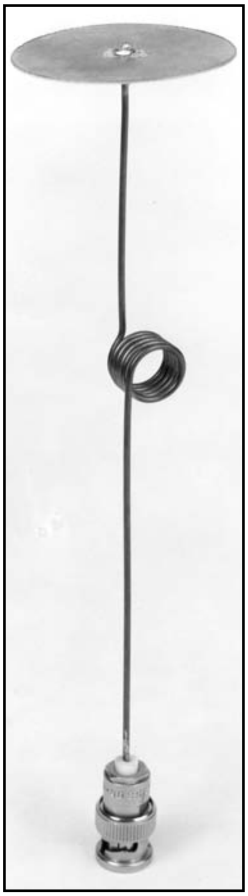

Fig 46—This 144-MHz antenna uses a combination of top and center loading. It offers low construction cost and improved efficiency over continuously loaded rubber-ducky antennas.

A combination of top and center loading offers improved efficiency over continuously loaded antennas such as the “stubby” pictured at the beginning of this chapter. This antenna also offers low construction cost. The only materials needed are a length of stiff wire and a scrap of circuit-board material, in addition to the appropriate connector.

Construction

The entire whip section with above-center loading coil is made of one continuous length of material. An 18-inch length of brazing rod or #14 Copperweld wire is suitable.

In the antenna pictured in Fig 46, the top loading disk was cut from a scrap of circuit-board material, but flashing copper or sheet brass stock could be used instead. Aluminum is not recommended.

The dimensions of the antenna are given in **Fig 47**. First wind the center loading coil. Use a $^{1}/_{2}$-inch bolt, wood dowel, or other cylindrical object for a coil form. Begin winding at a point 3 inches from one end of the wire, and wrap the wire tightly around the coil form. Wind $5^{1}/_{2}$ turns, with just enough space between turns so they don’t touch.

Remove the coil from the form. Next, determine the length necessary to insert the wire into the connector you’ll be using. Cut the long end of the wire to this length plus 4 inches, measured from the center of the coil. Solder the wire to the center pin and assemble the connector. A tight-fitting sleeve made of Teflon or Plexiglas rod may be used to support and insulate the antenna wire inside the shell. An alternative is to fill the shell with epoxy cement, and allow the cement to set while the wire is held centered in the shell.

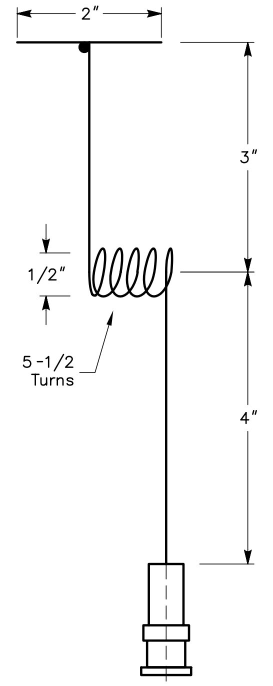

Fig 47—Dimensions for the top-loaded 144-MHz antenna. See text regarding coil length.

The top loading disk may be circular, cut with a hole saw. A circular disk is not required, however—it may be of any shape. Just remember that with a larger disk, less coil inductance will be required, and vice versa. Drill a hole at the center of the disk for mounting it to the wire. For a more rugged antenna, reinforce the hole with a brass eyelet. Solder the disk in place at the top of the antenna, and construction is completed.

Tune-Up

Adjustment consists of spreading the coil turns for the correct amount of inductance. Do this at the center frequency of the range you’ll normally be using. Optimum inductance is determined with the aid of a field strength meter at a distance of 10 or 15 feet.

Attach the antenna to a handheld transceiver operating on low power, and take a field-strength reading. With the transmitter turned off, spread the coil turns slightly, and then take another reading. By experiment, spread or compress the coil turns for the maximum field-strength reading. Very little adjustment should be required. There is one precaution, however. You must keep your body, arms, legs, and head in the same relative position for each field-strength measurement. It is suggested that the transceiver be placed on a nonmetal table and operated at arm’s length for these checks.

Once the maximum field-strength reading is obtained, adjustments are completed. With this antenna in operation, you’ll likely find it possible to access repeaters that are difficult to reach with other shortened antennas. W6AAQ reports that in distant areas his antenna even outperforms a $^{5}/_{8}$-λ vertical.

VHF QUARTER-WAVELENGTH VERTICAL

Ideally, a VHF vertical antenna should be installed over a perfectly flat reflector to assure uniform omnidirectional radiation. This suggests that the center of the automobile roof is the best place to mount it for mobile use. Alternatively, the flat portion of the trunk deck can be used, but will result in a directional pattern because of car-body obstruction.

Fig 48 illustrates how a Millen high-voltage connector can be used as a roof mount for a VHF whip. The hole in the roof can be made over the dome light, thus providing accessibility through the upholstery. RG-59 and the $^{1}/_{4}$-wave matching section, L (Fig 48C), can be routed between the car roof and the ceiling upholstery and brought into the trunk compartment, or down to the dashboard of the car. Instead of a Millen connector, some operators install an SO-239 coax connector on the roof for mounting the whip. The method is similar to that shown in Fig 48.

It has been established that in general, $^{1}/_{4}$-λ vertical antennas for mobile repeater work are not as effective as $^{5}/_{8}$-λ verticals are. With a $^{5}/_{8}$-λ antenna, more of the transmitted signal is directed at a low wave angle, toward the horizon, offering a gain of about 1 dB over the $^{1}/_{4}$-λ vertical. However, in areas where the repeater is located nearby on a very high hill or a mountain top, the $^{1}/_{4}$-λ antenna will usually offer more reliable per-

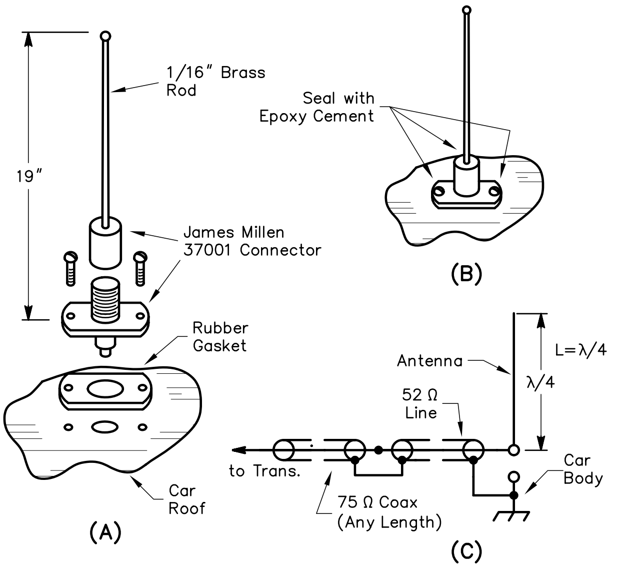

Fig 48—At A and B, an illustration of how a quarter-wavelength vertical antenna can be mounted on a car roof. The whip section should be soldered into the cap portion of the connector and then screwed into the base socket. This arrangement allows for the removal of the antenna when desired. Epoxy cement should be used at the two mounting screws to prevent the entry of moisture through the screw holes. Diagram C is discussed in the text.

formance than a $^{5}/_{8}$-λ antenna. This is because there is more power in the lobe of the $^{1}/_{4}$-λ vertical at higher angles.

144-MHz 5/8-WAVELENGTH VERTICAL

Perhaps the most popular antenna for 144-MHz FM mobile and fixed-station use is the $^{5}/_{8}$-wavelength vertical. As compared to a $^{1}/_{4}$-wavelength vertical, it has 1 dB of gain.

This antenna is suitable for mobile or fixed-station use because it is small, omnidirectional, and can be used with radials or a solid-plane ground (such as a car body). If radials are used, they need be only $^{1}/_{4}$ wavelength long.

Construction

The antenna shown here is made from low-cost materials. **Fig 49** shows the base coil and aluminum mounting plate. The coil form is a piece of low-loss solid rod, such as Plexiglas or phenolic. The dimensions for this and other parts of the antenna are given in **Fig 50**. A length of brazing rod is used as the whip section.

The whip should be 47 inches long. However, brazing rod comes in standard 36-inch lengths, so if used, it is necessary to solder an 11-inch extension to the top of the whip. A piece of #10 copper wire will suffice. Alternatively, a stainless-steel rod can be purchased to make a 47-inch whip. Shops that sell CB antennas should have such rods for replacement purposes on base-loaded antennas. The limitation one can expect with brazing rod is the relative fragility of the material, especially when the threads are cut for screwing the rod into the base coil form. Excessive stress can cause the rod to break where it enters the form. The problem is compli-

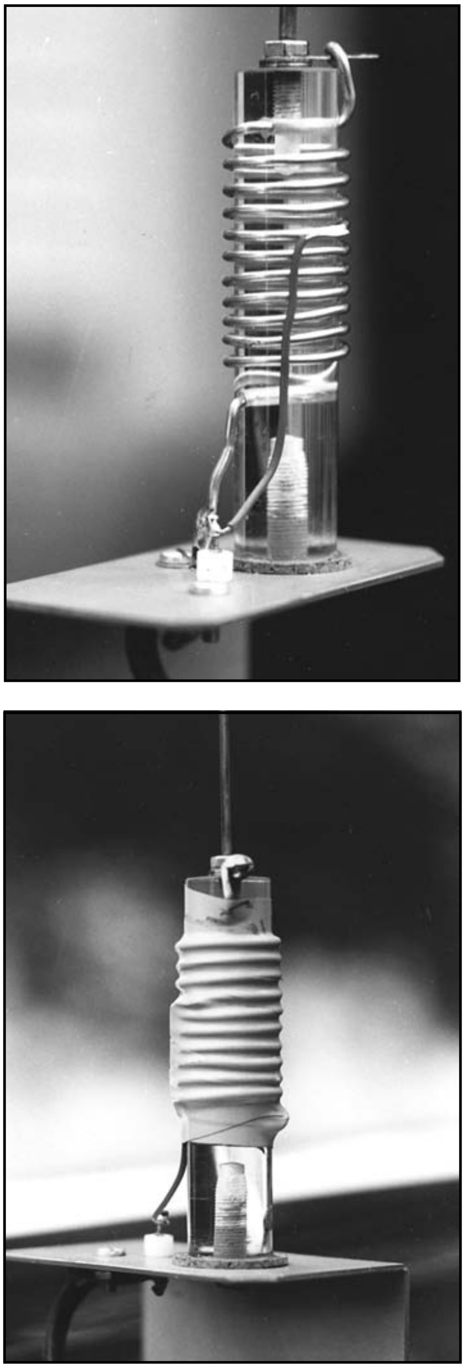

Fig 49—At top, a photograph of the $^{5}/_{8}$-λ vertical base section. The matching coil is affixed to an aluminum bracket that screws onto the inner lip of the car trunk. Above, the completed assembly. The coil has been wrapped with vinyl electrical tape to keep out dirt and moisture.

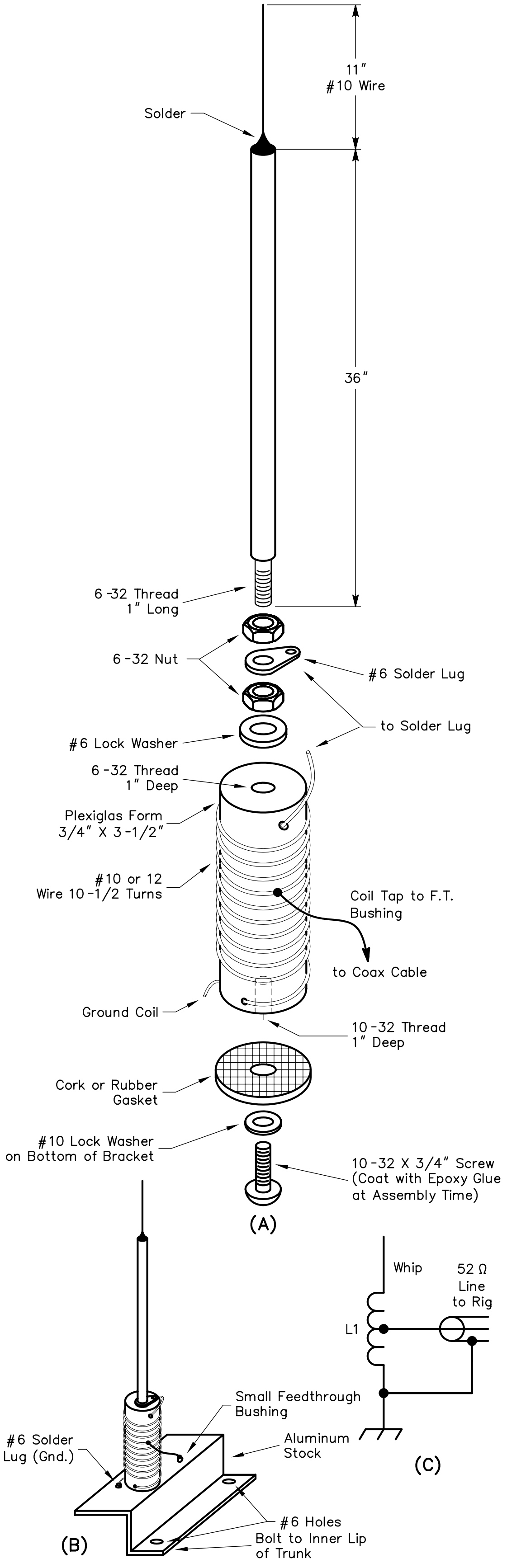
11″
#10 Wire
Solder
36″
6-32 Thread
1″ Long
6-32 Nut
#6 Solder Lug
to Solder Lug
#6 Lock Washer
6-32 Thread
1″ Deep
Plexiglas Form
3/4″ X 3-1/2″
#10 or 12
Wire 10-1/2 Turns
Coil Tap to F.T.
Bushing
to Coax Cable
Ground Coil
10-32 Thread
1″ Deep
Cork or Rubber
Gasket
#10 Lock Washer
on Bottom of Bracket
10-32 X 3/4″ Screw
(Coat with Epoxy Glue
at Assembly Time)
(A)
Whip
52 Ω
Line
to Rig
L1
(C)
Small Feedthrough
Bushing
Aluminum
Stock
#6 Solder
Lug (Gnd.)
#6 Holes
Bolt to Inner Lip
of Trunk
(B)

Fig 50—Structural details for the 2-meter 5/8-λ antenna are provided at A. The mounting bracket is shown at B and the equivalent circuit is given at C.

cated somewhat in this design because a spring is not used at the antenna mounting point. Builders of this antenna can find all kinds of solutions to the problems just outlined by changing the physical design and using different materials when constructing the antenna. The main purpose of this description is to provide dimensions and tune-up information.

The aluminum mounting bracket must be shaped to fit the car with which it will be used. The bracket can be used to effect a no-holes mount with respect to the exterior portion of the car body. The inner lip of the vehicle trunk (or hood) can be the point where the bracket is attached by means of no. 6 or no. 8 sheet-metal screws. The remainder of the bracket is bent so that when the trunk lid or car hood is raised and lowered, there is no contact between the bracket and the moving part. Details of the mounting unit are given in Fig 50B. A 14-gauge metal (or thicker) is recommended for rigidity.

Wind 10$^1/_2$ turns of #10 or #12 copper wire on the $^3/_4$-inch diameter coil form. The tap on L1 is placed approximately four turns below the whip end. A secure solder joint is imperative.

Tune-Up

After the antenna has been mounted on the vehicle, connect an SWR indicator in the 50-Ω transmission line. Key the 144-MHz transmitter and experiment with the coil tap placement. If the whip section is 47 inches long, an SWR of 1:1 can be obtained when the tap is at the right location. As an alternative method of adjustment, place the tap at four turns from the top of L1, make the whip 50 inches long, and trim the whip length until an SWR of 1:1 occurs. Keep the antenna well away from other objects during tune-up, as they may detune the antenna and yield false adjustments for a match.

A 5/8-WAVELENGTH 220-MHz MOBILE ANTENNA

The antenna shown in **Figs 51** and **52** was developed to fill the gap between a homemade $^1/_4$-λ mobile antenna and a commercially made $^5/_8$-λ model. While antennas can be made by modifying CB models, that presents the problem of cost in acquiring the original antenna. The major cost in this setup is the whip portion. This can be any tempered rod that will spring easily.

Construction

The base insulator portion is made of $^1/_2$-inch Plexiglas rod. A few minutes' work on a lathe is sufficient to shape and drill the rod. (The innovative builder can use an electric drill and a file for the lathe work.)

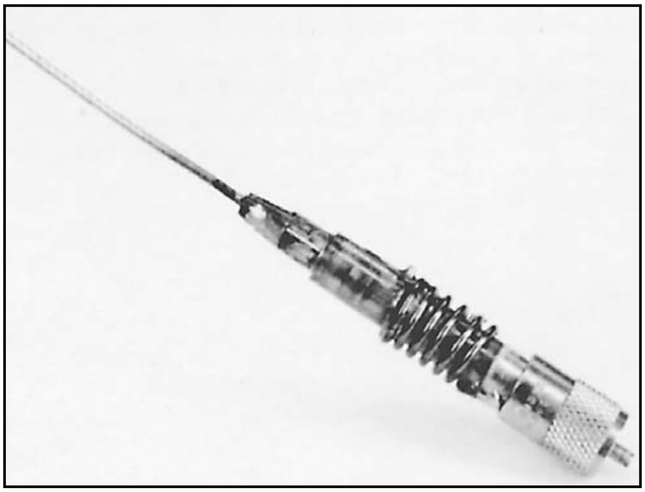

Fig 51—The 220-MHz $^5/_8$-λ mobile antenna. The coil turns are spaced over a distance of 1 inch, and the bottom end of the coil is soldered to the coax connector.

The bottom $^1/_2$ inch of the rod is turned down to a diameter of $^3/_8$ inch. This portion will now fit into a PL-259 UHF connector. A $^1/_8$-inch diameter hole is drilled through the center of the rod. This hole will hold the wires that make the connections between the center conductor of the connector and the coil tap. The connection between the whip and the top of the coil is also run through this opening. A stud is force-fitted into the top of the Plexiglas rod. This allows for removal of the whip from the insulator.

The coil should be initially wound on a form slightly smaller than the base insulator. When the coil is transferred to the Plexiglas rod, it will keep its shape and will not readily move. After the tap point has been determined, a longitudinal hole is drilled into the center of the rod. A #22 wire can then be inserted through the center of the insulator into the connector. This method is also used to attach the whip to the top of the coil. After the whip has been fully assembled, a coating of epoxy cement is applied. This seals the entire assembly and provides some additional strength. During a full winter's use there was no sign of cracking or other mechanical failure. The adjustment procedure is the same as for the 144-MHz version described previously.

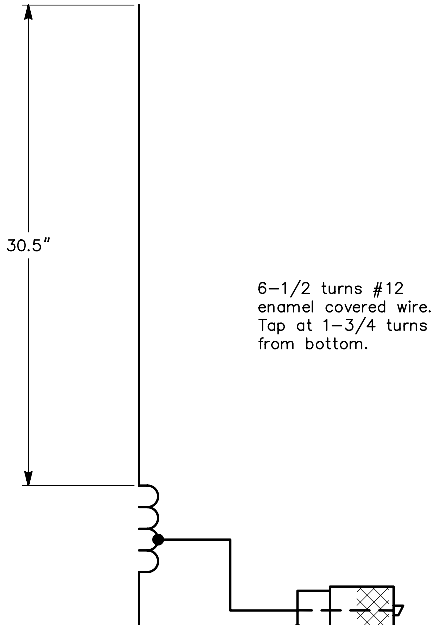

Fig 52—Diagram of the 220-MHz mobile antenna.

HF Antennas For Sailboats

This material was contributed by Rudy Severns, N6LF. Many of the antenna ideas appearing earlier in this chapter can be applied to sailboats. However, the presence of the mast and the rigging, plus the prevalence of non-conducting fiberglass hulls complicates the issue. There are many possibilities for antennas aboard sailboats. This includes both permanently installed antennas and antennas that can be hoisted for temporary use at anchor:

1. ***Permanent***
   Commercial or home-brew automobile-type verticals
   Backstay verticals and slopers
   Shunt feed of uninsulated rigging
2. ***Temporary***
   Sloping dipoles
   Inverted V's
   Yagis

You should remember some basic facts of life on a sailboat:

1) On most boats the spars, standing rigging and some running rigging will be conductors. Stainless steel wire is usually used for the rigging and aluminum for the spars.
2) Topping lifts, running backstays and jackstays all may be made of conducting materials and may often change position while the boat is underway. This changes the configuration of the rigging and may affect radiation patterns and feed-point impedances.
3) Shipboard antennas will always be close to the mast and rigging, in terms of electrical wavelength. Some antennas may in fact be part of the rigging.
4) The feed-point impedance and radiation pattern can be strongly influenced by the presence of the rigging.
5) Because of the close proximity, the rigging is an integral part of the antenna and should be viewed as such.
6) The behavior of a given antenna will depend on the details of the rigging on a particular vessel. The performance of a given antenna can vary widely on different boats, due to differences in dimensions and arrangement of the rigging.
7) Even though you may be floating on a sea of salt water, grounding still requires careful attention!

ANTENNA MODELING

Because of the strong interaction between the rigging and the antenna, accurate prediction of radiation patterns and a reasonable guess at expected feed-point impedance requires that you model both the antenna and the rigging. Unless you do accurately model the system, considerable cut-and-try may be needed. This can be expensive when it has to be done in 1 × 19 stainless steel wire with $300 swaged insulator fittings!

In fact, when your antenna is going to be part of standing rigging, it's a very good idea to try your designs out at the dock. You could temporarily use Copperweld wire and inexpensive insulators in place of the stainless rigging wire and the expensive insulators. This approach can save a good deal of money and aggravation. A wide variety of modeling programs are available and can be very helpful in designing a new antenna but they have to be used with some caution:

1) The rigging will have many small intersection angles and radically different conductor diameters, this can cause problems for *NEC* and *MININEC* programs.
2) You must usually taper the segment lengths near the junctions. This is done automatically in programs like *ELNEC* and *EZNEC*.
3) It is usually necessary to use one wire size for the mast, spars and rigging. Some improvement in accuracy can be obtained by modeling the mast as a cage of 3 or 4 wires.

The predicted radiation patterns will be quite good but the feed-point impedance predictions should be viewed as preliminary. Some final adjustment will usually be required. Because of the wide variation between boats, even those of the same class, each new installation is unique and should be analyzed separately.

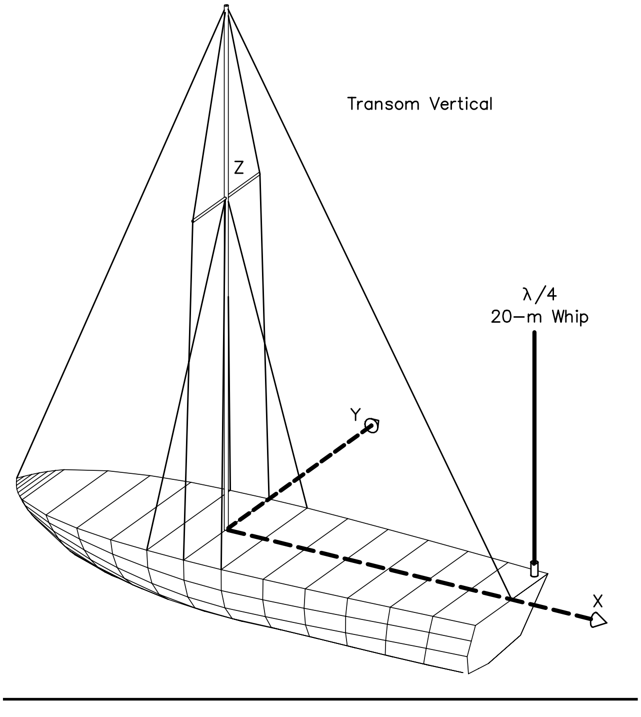

Fig 53—An example of a 20-meter λ/4 whip mounted on the transom. A local ground system must also be provided, as described in the section on grounding.

A SAFETY NOTE

Ungrounded rigging endpoints near deck level can have high RF potentials on them when you transmit. For example, the shrouds on a fiberglass boat connect to chainplates that are bolted to the hull, but are not grounded. These can inflict painful RF burns on the unwary, even while operating at low power! As a general rule all rigging, spars and lifelines near deck level should be grounded. This also makes good sense for lightning protection. For a backstay antenna with its feed point near deck level, a sleeve of heavy wall PVC pipe can be placed over the lower end of the stay as a protective shield.

TRANSOM AND MASTHEAD MOUNTED VERTICALS

A very common antenna for boats is a vertical, either a short mobile antenna or a full λ/4, placed on the transom, as shown in **Fig 53**. Note that in this example the antenna is mounted off to one side—it could also be mounted in the center of the transom. The 20-meter radiation pattern for this antenna as shown in **Fig 54**. Unlike a free standing vertical, this antenna doesn't have an omnidirectional pattern. It is asymmetrical, with a front-to-back ratio of about 13 dB. Further, the angle for maximum gain is offset in the direction the antenna is placed on the transom.

This is a very good example of the profound effect the rigging can have on any antenna used on board a sailboat. Not only is the pattern affected but the feed-point impedance will be reduced from a nominal 36 Ω to 25 to 30 Ω.

The directive gain can be useful—if you point the boat in the right direction! Usually, however, a more uniform omnidirectional pattern is more desirable. It is tempting to suggest putting the vertical at the masthead, perhaps using a 6-foot loaded automobile whip, with the mast and rigging acting as a ground plane. **Fig 55** shows such a system. Unfortunately, this usually doesn't work very well because the overall height of the mast and antenna will very likely be > 5/8 λ. This will result in high-angle lobes, as shown in **Fig 56**. Depending on the mast height, this idea may work reasonably well on 40 or 80 meters, but

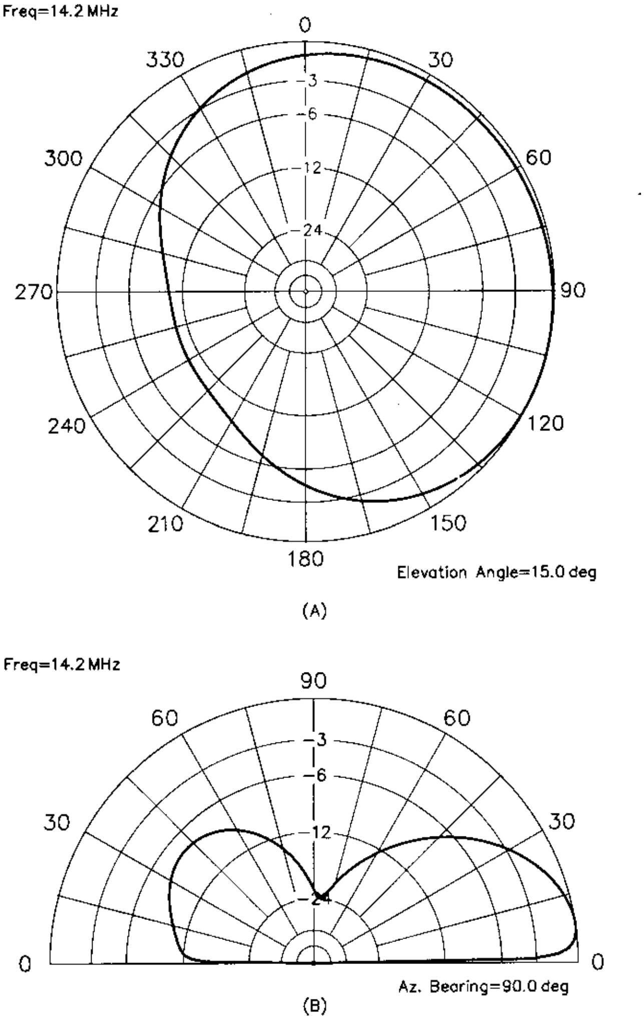

Fig 54—Typical radiation pattern for the λ/4 transom mounted whip in Fig 53.

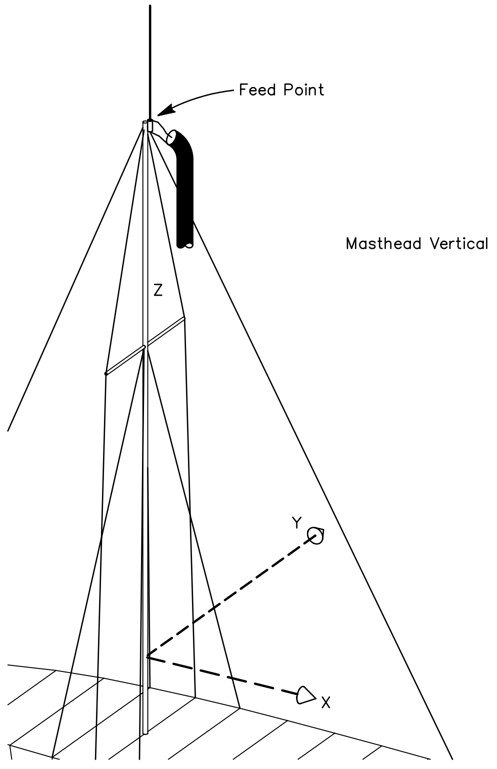

Fig 55—A whip mounted at the masthead. The feed line is fed back down the mast either inside or outside. The base of the mast and the rigging is assumed to be properly grounded.

you will still be faced with severe mechanical stress due to magnified motion at the masthead in rough sea. The masthead is usually reserved to VHF antennas, with their own radial ground plane.

THE BACKSTAY VERTICAL

A portion of the backstay can be insulated and used as a vertical as shown in **Fig 57**. The length of the insulated section will be λ/4 on the lowest band of interest. Typically, due to the loading effect of the rest of the rigging, the resonant length of the insulated section will be shorter than the classic 234/f (MHz) relation, although it can in some case actually be longer. Either modeling or trial adjustment can be used to determine the actual length needed. On a typical 35 to 40-foot sailboat, the lowest band for λ/4 resonance will be 40 meters due to the limited length of the backstay. Examples of the radiation patterns on several bands for such an an-

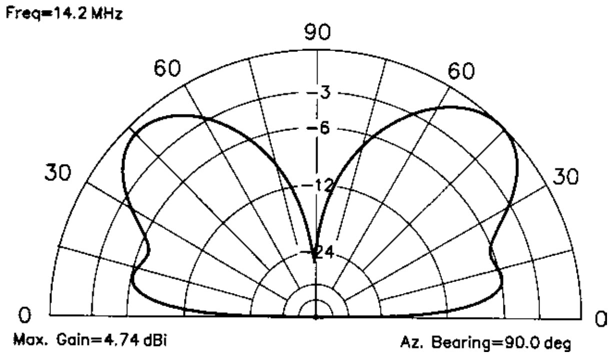

Fig 56—Typical radiation pattern for a masthead-mounted vertical. The multiple vertical lobes are due to the fact the antenna is higher than λ/2.

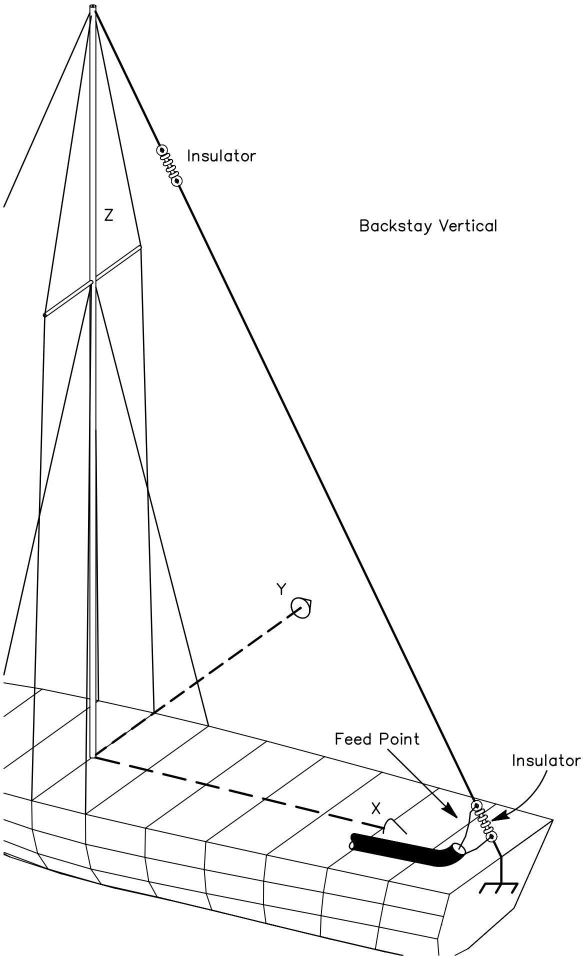

Fig 57—An example of a backstay vertical. A local ground point must be established on the transom next to the base of the backstay.

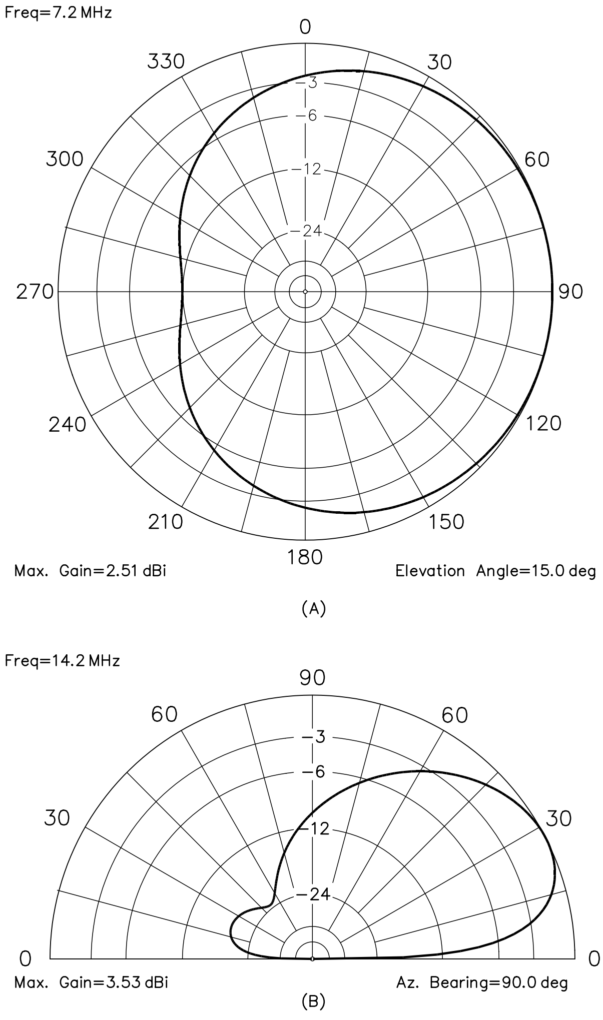

Fig 58—Typical radiation patterns on 40 meters for the backstay vertical in Fig 57.

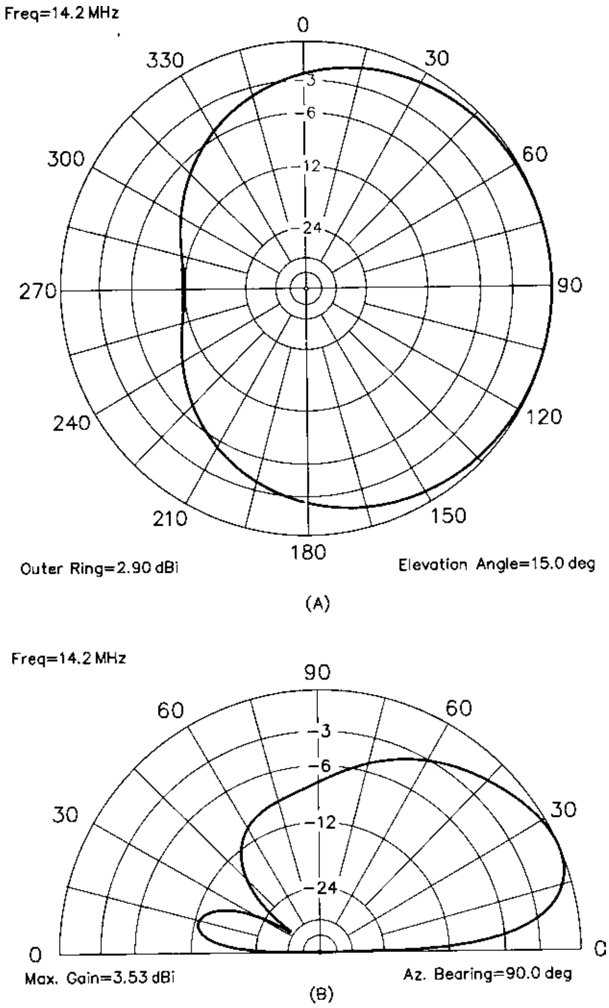

Fig 59—Typical radiation patterns on 20 meters for the backstay vertical in Fig 57.

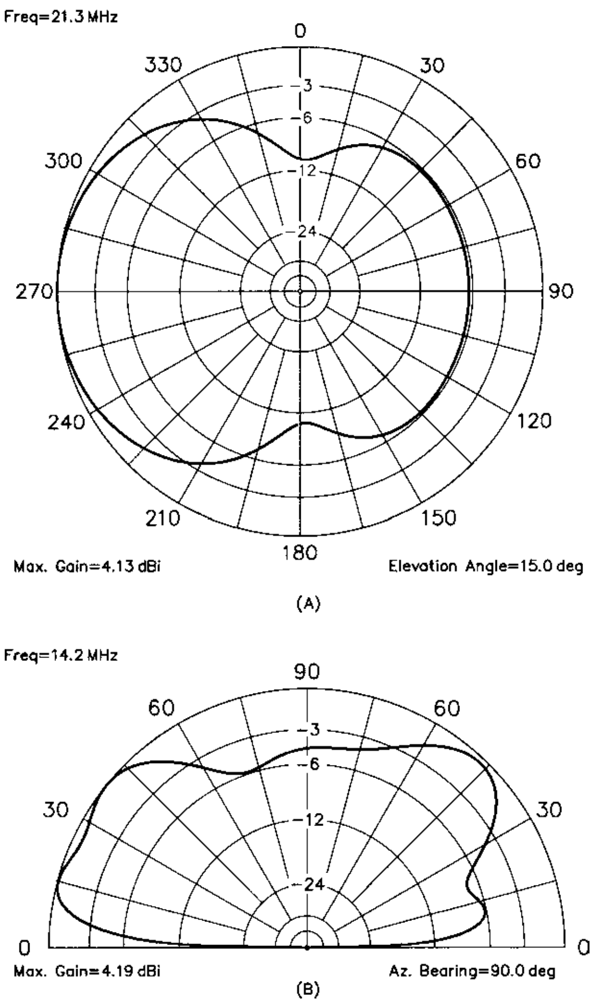

Fig 60—Typical radiation patterns on 15 meters for the backstay vertical in Fig 57.

tenna are given in **Figs 58** through **60**.

The pattern is again quite directional due to the presence of the mast and rigging. On 15 meters, where the antenna is approximately $^{3}/_{4}$ λ, higher angle lobes appear. On 40 and 15 meters, the feed point is near a current maximum and is in the range of 30 to 50 Ω. On 20 meters, however, the feed point is a very high impedance because the antenna is near λ/2 resonance. One way to get around this problem for multiband use is to make the antenna longer than λ/4 on the lowest band. If the lowest band is 40 meters then on 20 meters the feed-point impedance will be much lower. This antenna is non-resonant on any of the bands but can be conveniently fed with a tuner because the feed-point impedances are within the range of commonly available commercial tuners. Tuners specifically intended for marine applications frequently can accommodate very high input impedances, but they tend to be quite expensive.

The sensitivity of the radiation pattern to small details of the mast and rigging is illustrated in **Fig 61**. This is the same antenna as shown in Fig 57 with the exception that the forestay is assumed to be ungrounded. In this particular example, ungrounding the forestay drastically increases the front-to-back ratio. With slightly different dimensions, however, the pattern could have changed in other ways.

High-quality insulators for rigging wire can be quite expensive and represent a potential weak point—if they fail the mast may come down. It is not absolutely necessary to use two insulators in a backstay vertical. As shown in **Fig 62**, the upper insulator may be omitted. The radiation patterns are shown in **Figs 63** and **64**. In this case

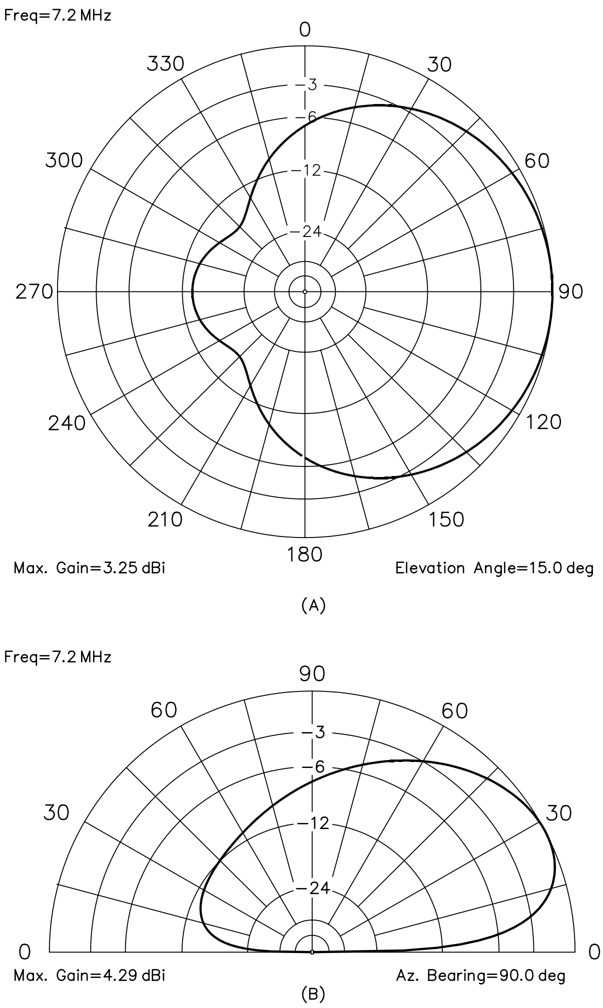

Fig 61—The effect of ungrounding the forestay on the radiation pattern. This is for 40-meter operation.

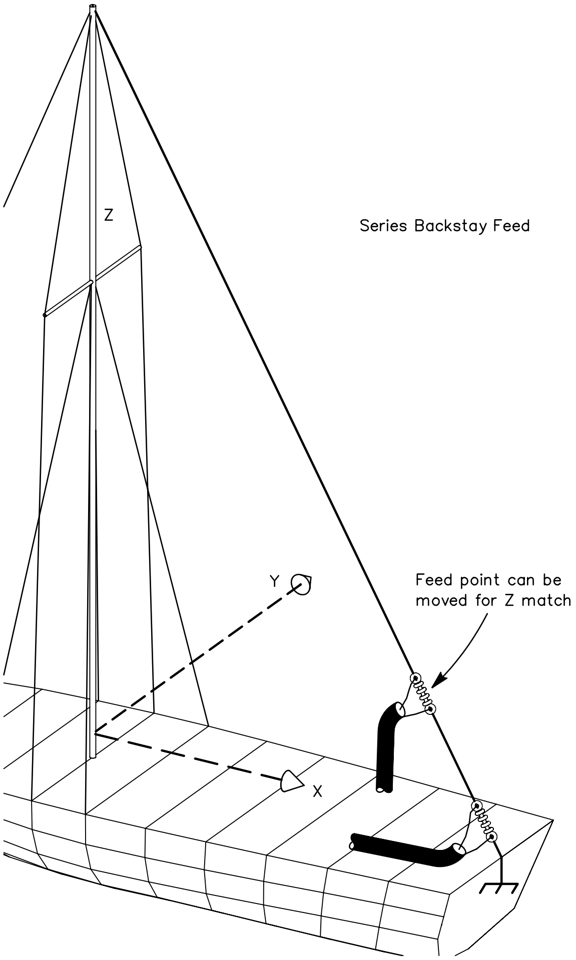

Fig 62—Feeding the backstay without an insulator at the top. The feed point may be moved along the antenna to find a point with a better match on a particular band or to provide a better range of impedances for the tuner to match. The coaxial feed line is taped to the lower portion of the backstay. Again, a good local ground is needed at the base of the backstay.

the pattern is actually more symmetrical than it was with an upper insulator—but this may not hold true for other rigging dimensions. The feed point does not have to be at the bottom of the backstay. As indicated in Fig 62, the feed point can be moved up into the backstay to achieve a better match or a more desirable feed-point impedance variation with frequency. In that case, the center of the coaxial feed line is connected to the upper section and the shield to the lower section. The cable is then taped to the lower portion of the backstay.

If single-band operation is all you want, even the lower insulator can be omitted by using shunt feed. A gamma match would be quite effective for this purpose, as discussed in Chapter 6 when driving a grounded tower.

A 40-METER BACKSTAY HALF SLOPER

A half-sloper antenna can be incorporated into the backstay, as shown in **Fig 65**. This will behave very much the same as the slopers described in Chapter 6. The advantage of this antenna for a sailboat installation is that you don’t need to create a good ground connection at the stern, as you would have to do for a transom-mounted vertical or the backstay vertical just described. This may be more convenient. The mast, shrouds and stays must still be grounded for the half-sloper but the arrangement is somewhat simpler.

TEMPORARY ANTENNAS

Not everyone needs permanent antennas. A variety

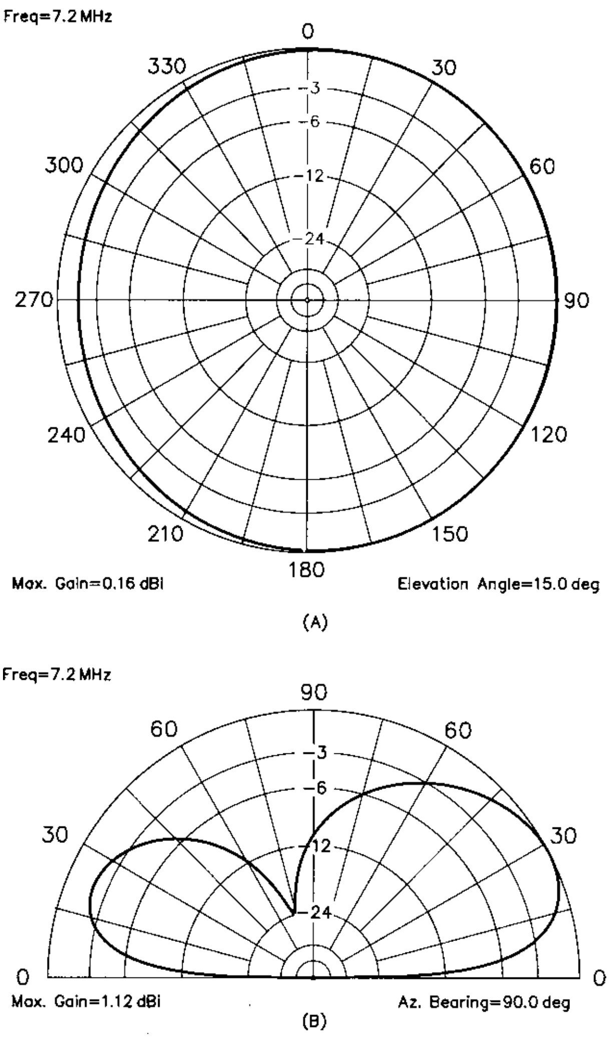

Fig 63—Typical 40-meter radiation pattern in Fig 62.

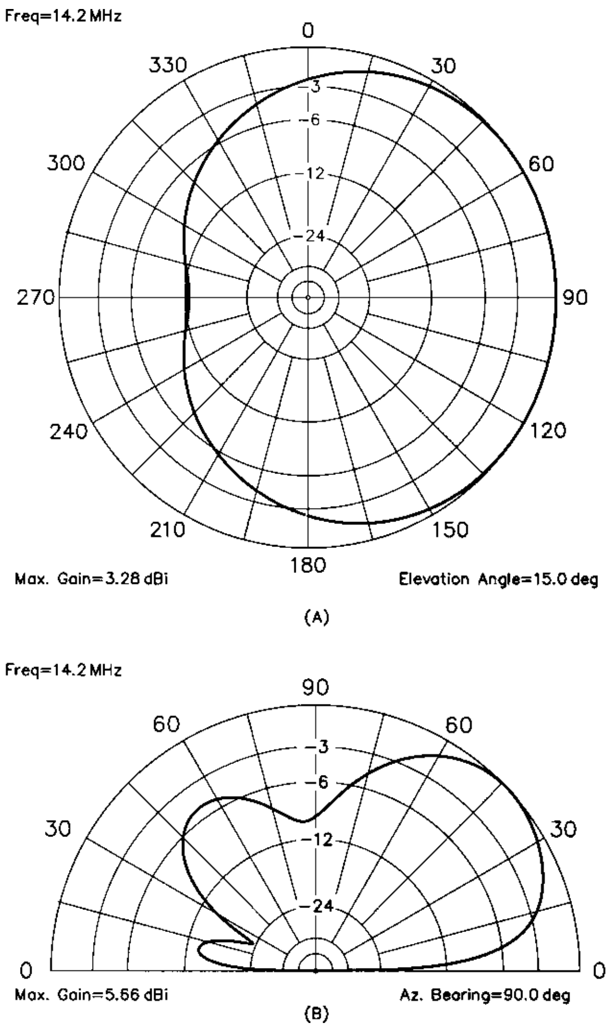

Fig 64—Typical 20-meter radiation pattern in Fig 62.

of temporary antennas can be arranged. A few of these are shown in **Figs 66** through **68**. Of all of these the rigid dipole (Fig 66) will provide the best operation and will have a pattern close to that expected from a freestanding dipole. The other two examples will be strongly affected by their close proximity to the rigging.

GROUNDING SYSTEMS

You may be sitting in the middle of a thousand miles of saltwater. This is great for propagation but you will still have to connect to that ground if you want to use a vertical. There are many possibilities, but the scheme shown in **Fig 69** is representative. First a bonding wire, or better yet a copper strap (it can be very thin!), is connected from bow-to-stern on each side, connecting the forestay, lifeline stanchions, chainplates, bow and stern pulpits and the backstay. Other bonding wires are run from the bow, stern and chainplates on both sides to a common connection at the base of the mast. The fore-and-aft bonding can be attached to the engine and to the keel bolts. The question arises: "What about electrolysis between the keel and propeller if you bond them together?" This has to be dealt with on a case-by-case basis. If your protective zincs are depleting more rapidly after you install a bonding scheme, change it by disconnecting something, the engine-shaft-propeller, for example.

Grounding will vary in every installation and has to be customized to each vessel. However, just as on shore, the better the ground system, the better the performance of the vertical!

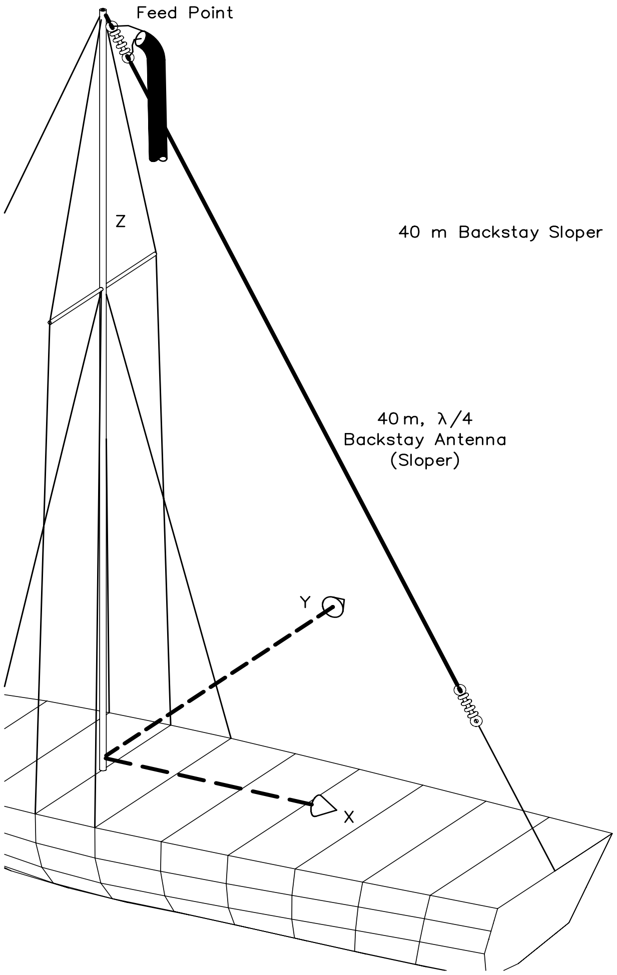
Feed Point
Z
40 m Backstay Sloper
40 m, λ/4
Backstay Antenna
(Sloper)
Y
X

Fig 65—A 40-meter half-sloper fed at the masthead.

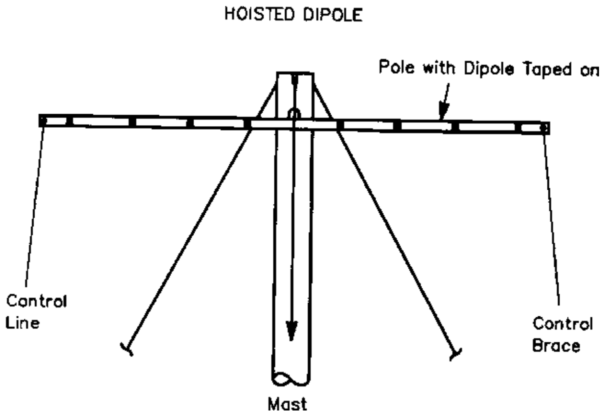
HOISTED DIPOLE
Pole with Dipole Taped on
Control
Line
Control
Brace
Mast

Fig 66—A dipole can be taped to a wood or bamboo pole and hoisted to the masthead with the main halyard while at anchor. It is possible to make this a multiband dipole.

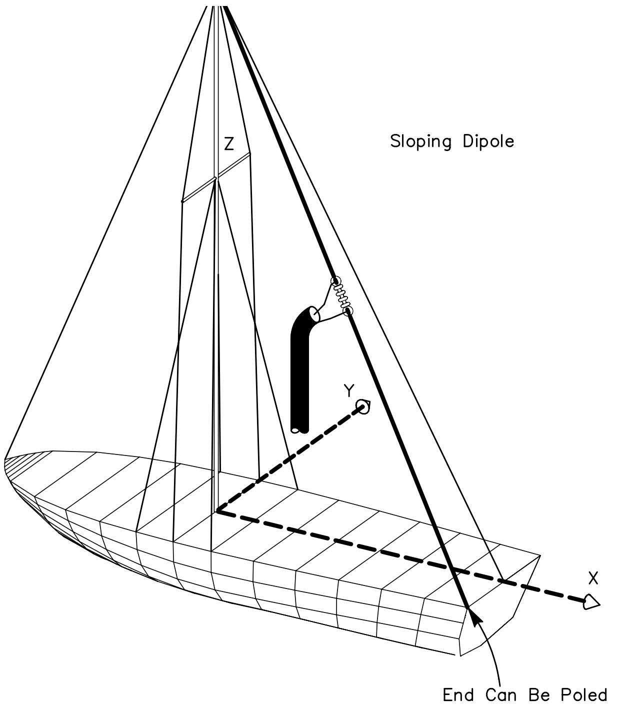
Z
Sloping Dipole
Y
X
End Can Be Poled

Fig 67—One end of dipole can be attached to the main halyard and pulled up to the masthead. The bottom end of the dipole should be pulled out away from the rigging as much as possible to reduce the impact of the rigging on the impedance.

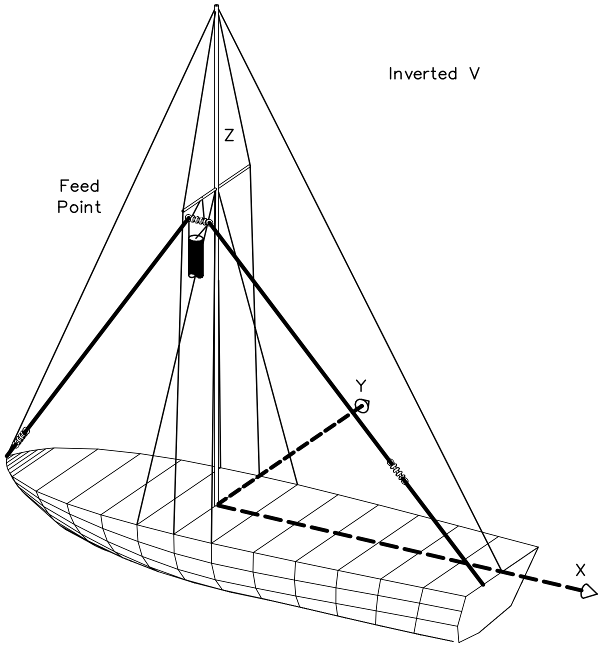
Inverted V
Z
Feed
Point
Y
X

Fig 68—The flag halyard can be used to hoist the center of an inverted V to the spreaders, or alternatively, the main halyard can be used to hoist the center of the antenna to the masthead. Interaction between the rigging and the antenna will be very pronounced and the length of the antenna will have to be adjusted on a cut-and-try basis.

Antennas for Power Boats

Powerboaters are not usually faced with the problems and opportunities created by the mast and rigging on a sailboat. A powerboat may have a small mast, but usually not on the same scale as a sailboat. Antennas for power boats have much more in common with automotive mobile operation, but with some important exceptions:

1) In an automobile, the body is usually metal and it provides a groundplane or counterpoise for a whip antenna. Most modern powerboats, however, have fiberglass hulls. These are basically insulators, and will not work as counterpoises. (On the other hand, metal-hulled power boats can provide nearly ideal grounding!)
2) A height restriction on automotive mobile whips is imposed by clearance limits on highway overpasses and also by the need to sustain wind speeds of up to 80 miles per hour on the highway.
3) In general, powerboats can have much taller antennas that can be lowered for the occasional low bridge.
4) The motion on a powerboat, especially in rough seas, can be quite severe. This places additional mechanical strain on the antennas.
5) On both powerboats and sailboats, operation in a saltwater marine environment is common. This means that a careful choice of materials must be made for the antennas to prevent corrosion and premature failure.

The problem of a ground plane for vertical antennas can be handled in much the same manner as shown in **Fig 69** for sailboats. Since there will most likely not be a large keel structure to connect to and provide a large surface area, additional copper foil can be added inside the hull to increase the counterpoise area. Because of the small area of the propeller, it may be better not to connect to the engine, but to rely instead on increasing the area of the counterpoise and operate it as a true counterpoise—that is, isolated from ground. Sometimes a number of radial wires are used for a vertical, much like that for a ground-plane antenna. This is not a very good idea unless the “wires” are actually wide copper-foil strips that can lower the Q substantially.

The problem is the high voltage present at the ends of normal ground-plane antenna radials. For a boat these radials are likely to be in close proximity to the cabin, which in turn contains both people and electronic equipment. The high potential at the ends of the radials is both a safety hazard and can result in RF coupling back into the equipment, including ham gear, navigational instruments and entertainment devices. The cook is not likely to be happy if he or she gets an RF burn after touching the galley stove! Decoupling the counterpoise from the transmission line, as discussed in Chapter 6, can be very helpful to keep RF out of other equipment.

One way to avoid many of the problems associated with grounding is to use a rigid dipole antenna. On 20 meters and higher, a rigid dipole made up from aluminum tubing, fiberglass poles or some combination of these, can be effective. As shown in **Fig 70A** the halves of the dipole can be slanted upward like rabbit ears to reduce the wingspan and increase the feed-point impedance for a better match to common coax lines. On the lower bands a pair of mobile whips can be used, as shown in Fig 70B. Home-brew coils could also be used.

For short-range communication, a relatively low dipole over saltwater can be effective. However, if long-range communication is needed, then a well-designed vertical, operating over seawater, will work much better. For these to work, of course, you must solve the ground problems associated with a vertical.

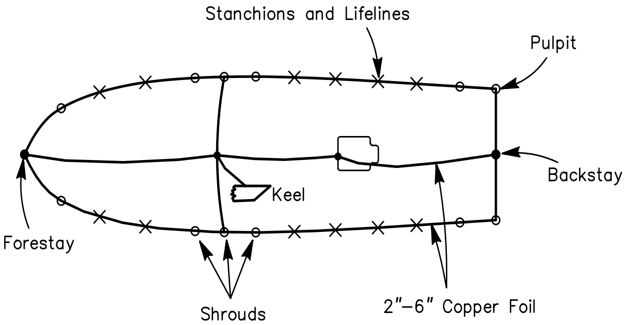

Fig 69—A typical sailboat grounding scheme.

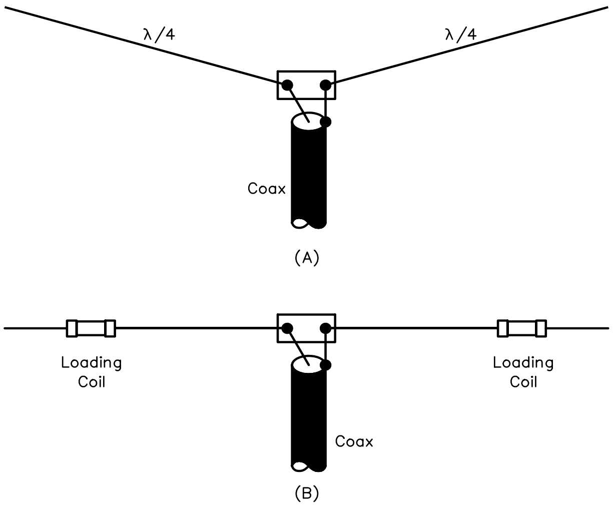

Fig 70—At A, a rigid dipole made from aluminum tubing, fiberglass poles or a combination of these. At B, a pair of mobile whips used as a dipole.

It is not uncommon for large powerboats to have a two or three-element multiband Yagi installed on a short mast. While these can be effective, if they are not mounted high (> $\lambda/2$) they may be disappointing for longer-range communication. Over saltwater, vertical polarization is very effective for longer distances. A simpler, but well-designed, vertical system on a boat may outperform a low Yagi.

The combination of a good ground system and one of the high-quality, motor-tuned multiband mobile whips now available commercially can also be very effective.

BIBLIOGRAPHY

Source material and more extended discussions of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

J. S. Belrose, "Short Antennas for Mobile Operation," *QST*, Sep 1953.

J. S. Belrose, "Vertical J Antenna for 2 meters," *The Canadian Amateur*, Jul/Aug 1979, pp 23-26.

J. M. Boyer, "Antenna-Transmission Line Analog," *Ham Radio*, May 1977.

B. F. Brown, "Tennamatic: An Auto-Tuning Mobile Antenna System," *73*, Jul 1979.

B. F. Brown, "Optimum Design of Short Coil-Loaded High-Frequency Mobile Antennas," *ARRL Antenna Compendium Volume 1* (Newington: ARRL, 1985), p 108.

W. B. Freely, "A Two-meter J Antenna," *QST*, Apr 1977, pp 35-36.

E. A. Laport, *Radio Antenna Engineering* (New York: McGraw-Hill Book Co., 1952), p 23.

C. E. Smith and E. M. Johnson, "Performance of Short Antennas," Proceedings of the IRE, Oct 1947.

F. E. Terman, *Radio Engineering Handbook*, 3rd edition (New York: McGraw-Hill Book Co., 1947), p 74.

Chapter 17

Repeater Antenna Systems

There is an old adage in Amateur Radio that goes, "If your antenna did not fall down last winter, it wasn't big enough." This adage might apply to antennas for MF and HF work, but at VHF things are a bit different, at least as far as antenna size is concerned. VHF antennas are smaller than their HF counterparts, but yet the theory is the same—a dipole is a dipole, and a Yagi is a Yagi, regardless of frequency. A 144-MHz Yagi may pass as a TV antenna, but most neighbors can easily detect a radio hobbyist if a 14-MHz Yagi looms over his property.

Repeater antennas are discussed in this chapter. Because the fundamental operation of these antennas is no different than presented in earlier chapters, there is no need to delve into any exotic theory. Certain considerations must be made and certain precautions must be observed, however, since most repeater operations—amateur and commercial—take place at VHF and UHF.

Basic Concepts

The antenna is a vital part of any repeater installation. Because the function of a repeater is to extend the range of communications between mobile and portable stations, the repeater antenna should be installed in the best possible location to provide the desired coverage. This usually means getting the antenna as high above average terrain as possible. In some instances, a repeater may need to have coverage only in a limited area or direction. When this is the case, antenna installation requirements will be completely different, with certain limits being set on height, gain and power.

Horizontal and Vertical Polarization

Until the upsurge in FM repeater activity in the 1970s, most antennas used in amateur VHF work were horizontally polarized. These days, very few repeater groups use horizontal polarization. (One of the major reasons for using horizontal polarization is to allow separate repeaters to share the same input and/or output frequencies with closer-than-normal geographical spacing.) The vast majority of VHF and UHF repeaters use vertically polarized antennas, and all the antennas discussed in this chapter are of that type.

Transmission Lines

Repeaters provide the first venture into VHF and UHF work for many amateurs. The uninitiated may not be aware that the transmission lines used at VHF become very important because feed-line losses increase with frequency.

The characteristics of feed lines commonly used at VHF are discussed in Chapter 24, Transmission Lines. Although information is provided there for small-diameter RG-58 and RG-59 coaxes, these should not be used except for very short feed lines (25 feet or less). These cables are very lossy at VHF. In addition, the losses can be much higher if fittings and connections are not carefully installed.

The differences in loss between solid-polyethylene dielectric types (RG-8 and RG-11) and those using foam polyethylene are significant at VHF and UHF. If you can afford the line with the least loss, buy it.

If you must bury coaxial cable, check with the cable manufacturer before doing so. Many popular varieties of coaxial cable should not be buried, since the dielectric can become contaminated from moisture and soil chemicals. Some coaxial cables are labeled as non-contaminating. Such a label is the best way to be sure your cable can be buried without damage.

Matching

Losses are lowest in transmission lines that are matched to their characteristic impedances. If there is a mismatch at the end of the line, the losses increase. The *only way* to reduce the SWR on a transmission line is by matching the line *at the antenna*. Changing the length of a transmission line does not reduce the SWR. The SWR is established by the impedance of the line and the impedance of the antenna, so matching must be done at the antenna end of the line.

The importance of matching, so far as feed-line losses are concerned, is sometimes overstressed. But under some conditions, it is necessary to minimize feed-line losses related to SWR if repeater performance is to be consistent. It is important to keep in mind that most VHF/UHF equipment is designed to operate into a 50-Ω load. The output circuitry will not be loaded properly if connected to a mismatched line. This leads to a loss of power, and in some cases, damage to the transmitter.

Repeater Antenna System Design

Choosing a repeater or remote-base antenna system is as close as most amateurs come to designing a commercial-grade antenna system. The term *system* is used because most repeaters utilize not only an antenna and a transmission line, but also include duplexers, cavity filters, circulators or isolators in some configuration. Assembling the proper combination of these items in constructing a reliable system is both an art and a science. In this section prepared by Domenic Mallozzi, N1DM, the functions of each component in a repeater antenna system and their successful integration are discussed. While every possible complication in constructing a repeater cannot be foreseen at the outset, this discussion should serve to steer you along the right lines in solving any problems encountered.

COMPUTING THE COVERAGE AREA FOR A REPEATER

Modern computer programs can show the coverage of a repeater using readily available topographic data from the Internet. In Chapter 3, The Effects of Ground, we described the *MicroDEM* program supplied on the CD-ROM accompanying this book. Dr Peter Guth, the author of *MicroDEM*, built into it the ability to generate terrain profiles that can be used with ARRL's *HFTA* (HF Terrain Assessment) program (also included on the CD-ROM).

MicroDEM has a wide range of capabilities beyond simply making terrain profiles. It can do *LOS* (line of sight) computations, based on visual or radio-horizon considerations. **Fig 1** shows a *MicroDEM* map for the area around Glastonbury, Connecticut. This is somewhat hilly terrain, and as a result the coverage for a repeater placed here on a 30-meter (100-foot) high tower would be somewhat spotty. Fig 1 shows a "Viewshed" on the map, in the form of the white terrain profile strokes in 5° increments around the tower.

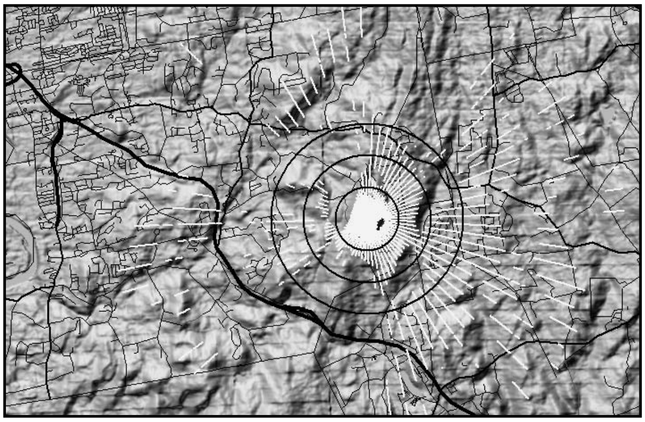

Fig 1—*MicroDEM* topographic map, showing the coverage for a repeater placed on a 30-meter high tower in Glastonbury, CT. The white radial lines indicate the coverage in 5° increments of azimuth around the tower. The range circles are 1000 meters apart.

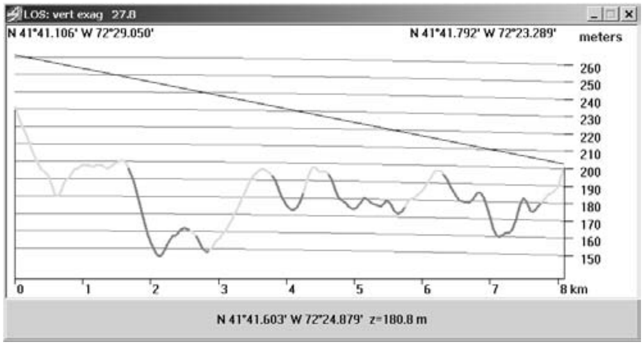

LOS: vert exag 27.8
N 41°41.106' W 72°29.050'
N 41°41.792' W 72°23.289'
meters
260
250
240
230
220
210
200
190
180
170
160
150
0
1
2
3
4
5
6
7
8 km
N 41°41.603' W 72°24.879' z=180.8 m

Fig 2—An "LOS" (line of sight) profile at an azimuth of 80° from the tower in Fig 1. The light-gray portions of the terrain profile are visible from the top of the tower, while the dark portions are blocked by the terrain.

Fig 2 shows the LOS for an azimuth of 80°, from a 30-meter high tower out to a distance of 8000 meters. The light-shaded areas on the profile are those that are illuminated directly by the antenna on the tower, while the dark portions of the profile are those that cannot be seen directly from the tower. This profile assumes that the mobile station is 2 meters high—the height of a 6-foot high person with a handheld radio.

The terrain at an 80° azimuth allows direct radio view from the top of the tower out to about 1.8 km. From here, the downslope prevents direct view until about 2.5 km, where the terrain is briefly visible again from several hundred meters, disappearing from radio view until about 2.8 km, after which it becomes visible until about 3.6 km. Note that other than putting the repeater antenna on a higher tower, there is nothing that can be done to improve repeater coverage over this hilly terrain, although knife-edge diffraction off the hill tops will help fill in coverage gaps.

The Repeater Antenna

The most important part of the system is the antenna itself. As with any antenna, it must radiate and collect RF energy as efficiently as possible. Many repeaters use omnidirectional antennas, but this is not always the best choice. For example, suppose a group wishes to set up a repeater to cover towns A and B and the interconnecting state highway shown in **Fig 3**. The available repeater site is marked on the map. No coverage is required to the west or south, or over the ocean. If an omnidirectional antenna is used in this case, a significant amount of the radiated signal goes in undesired directions. By using an antenna with a cardioid pattern, as shown in Fig 1, the coverage is concentrated in the desired directions. The repeater will be more effective in these locations, and signals from low-power portables and mobiles will be more reliable.

In many cases, antennas with special patterns are more expensive than omnidirectional models. This is an obvious consideration in designing a repeater antenna system. Over terrain where coverage may be difficult in some direction from the repeater site, it may be desirable to skew the antenna pattern in that direction. This can be accomplished by using a phased-vertical array or a com-

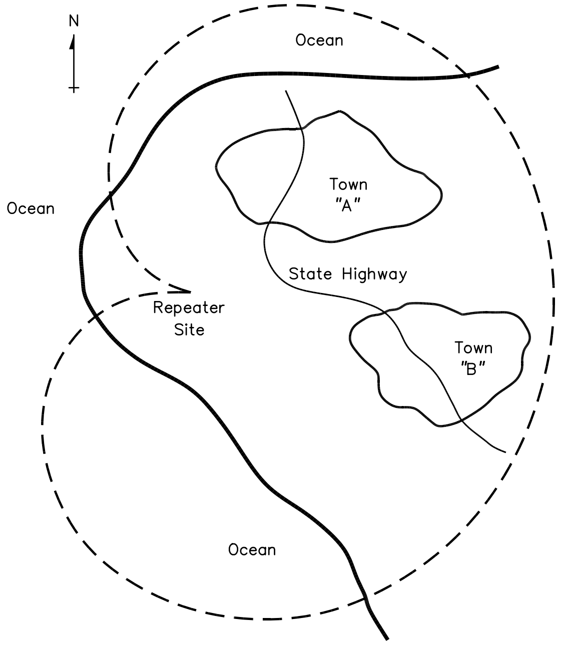

Fig 3—There are many situations where equal repeater coverage is not desired in all directions from the "machine." One such situation is shown here, where the repeater is needed to cover only towns A and B and the interconnecting highway. An omnidirectional antenna would provide coverage in undesired directions, such as over the ocean. The broken line shows the radiation pattern of an antenna that is better suited to this circumstance.

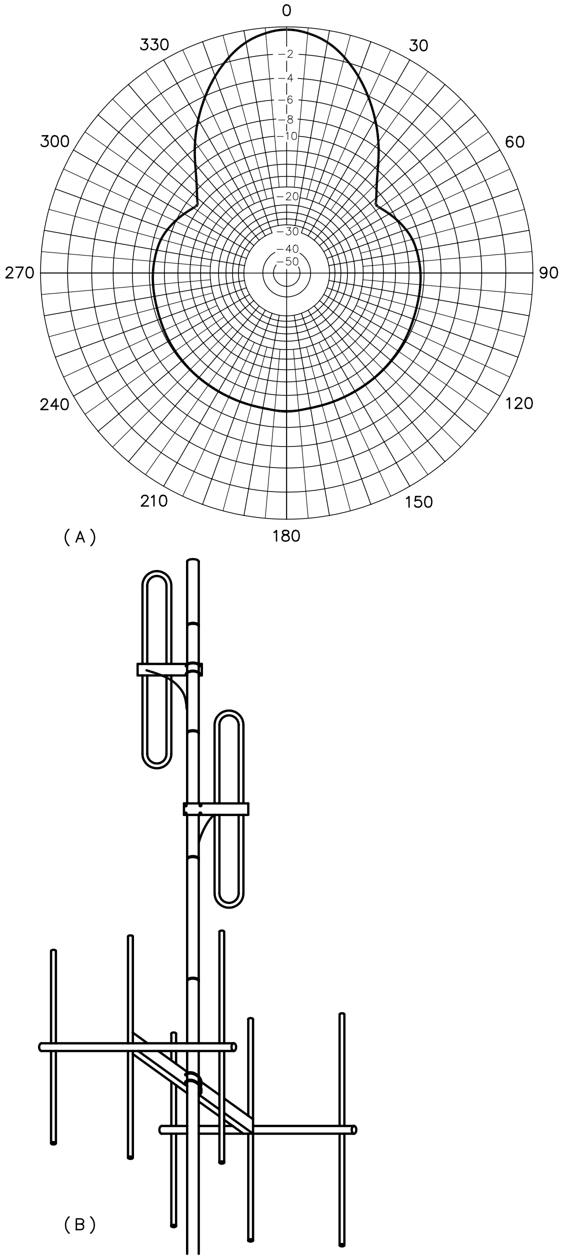

Fig 4—The "keyhole" horizontal radiation pattern at A is generated by the combination of phased Yagis and vertical elements shown at B. Such a pattern is useful in overcoming coverage blockages resulting from local terrain features. (*Based on a design by Decibel Products, Inc.*)

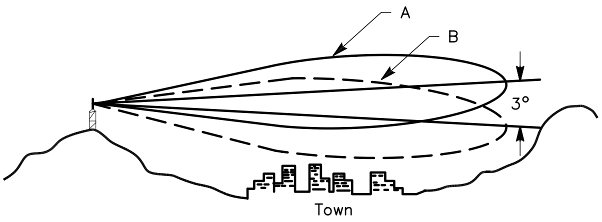

Fig 5—Vertical-beam downtilt is another form of radiation-pattern distortion useful for improving repeater coverage. This technique can be employed in situations where the repeater station is at a greater elevation than the desired coverage area, when a high-gain omnidirectional antenna is used. Pattern A shows the normal vertical-plane radiation pattern of a high-gain omnidirectional antenna with respect to the desired coverage area (the town). Pattern B shows the pattern tilted down, and the coverage improvement is evident.

bination of a Yagi and a phased vertical to produce a “keyhole” pattern. See **Fig 4**.

Repeaters are common on 440 MHz and above, and many groups invest in high-gain omnidirectional antennas. A consequence of getting high gain from an omnidirectional antenna is vertical beamwidth reduction. In most cases, these antennas are designed to radiate their peak gain at the horizon, resulting in optimum coverage when the antenna is located at a moderate height over normal terrain. Unfortunately, in cases where the antenna is located at a very high site (overlooking the coverage area) this may not be the most desirable pattern. The vertical pattern of the antenna can be tilted downward, however, to facilitate coverage of the desired area. This is called *vertical-beam downtilt*.

An example of such a situation is shown in **Fig 5**. The repeater site overlooks a town in a valley. A 450-MHz repeater is needed to serve low-power portable and mobile stations. Constraints on the repeater dictate the use of an antenna with a gain of 11 dBi. (An omnidirec-

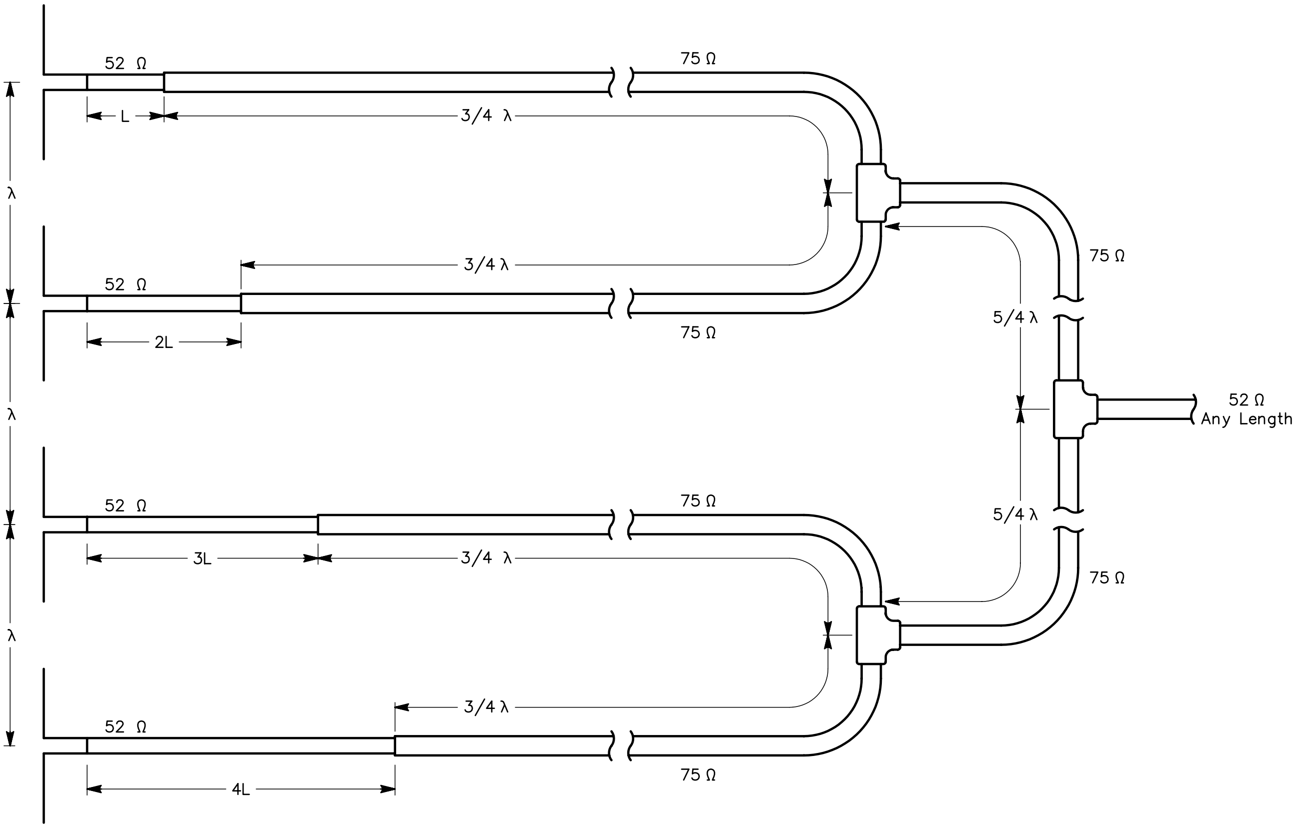

Fig 6—Vertical-beam downtilt can be facilitated by inserting 52-Ω delay lines in series with the 75-Ω feed lines to the collinear elements of an omnidirectional antenna. The delay lines to each element are progressively longer so the phase shift between elements is uniform. Odd $^1/_4$-λ coaxial transformers are used in the main (75-Ω) feed system to match the dipole impedances to the driving point. Tilting the vertical beam in this way often produces minor lobes in the vertical pattern that do not exist when the elements are fed in phase.

tional antenna with this gain has a vertical beamwidth of approximately 6°.) If the repeater antenna has its peak gain at the horizon, a major portion of the transmitted signal and the best area from which to access the repeater exists *above* the town. By tilting the pattern down 3°, the peak radiation will occur in the town.

Vertical-beam downtilt is generally produced by feeding the elements of a collinear vertical array slightly out of phase with each other. Lee Barrett, K7NM, showed such an array in *Ham Radio* magazine. (See the Bibliography at the end of this chapter.) Barrett gives the geometry and design of a four-pole array with progressive phase delay, and a computer program to model it. The technique is shown in **Fig 6**, with a free-space elevation plot showing downtilting in **Fig 7**.

Commercial antennas are sometimes available (at extra cost) with built-in downtilt characteristics. Before ordering such a commercial antenna, make sure that you really require it—they generally are special-order items and are not returnable.

There are disadvantages to improving coverage by means of vertical-beam downtilt. When compared to a standard collinear array, an antenna using vertical-beam downtilt will have somewhat greater extraneous lobes in the vertical pattern, resulting in reduced gain (usually less than 1 dB). Bandwidth is also slightly reduced. The reduction in gain, when combined with the downtilt characteristic, results in a reduction in total coverage area. These trade-offs, as well as the increased cost of a commercial antenna with downtilt, must be compared to the improvement in total performance in a situation where vertical-beam downtilt is contemplated.

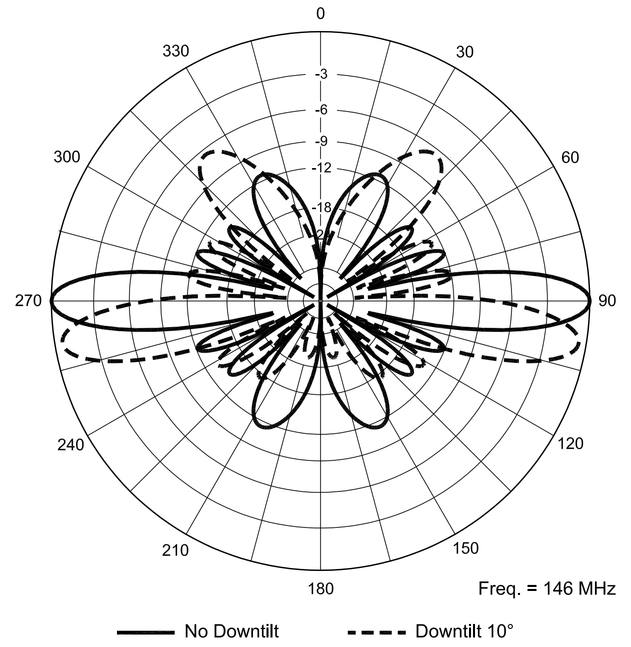

Fig 7—Free-space elevation-plane patterns showing downtilting that results from progressive phase shifts for the feed currents for the dipole in Fig 6.

Top Mounting and Side Mounting

Amateur repeaters often share towers with commercial and public service users. In many of these cases, other antennas are at the top of the tower, so the amateur antenna must be side mounted. A consequence of this arrangement is that the free-space pattern of the repeater antenna is distorted by the tower. This effect is especially noticeable when an omnidirectional antenna is side mounted on a structure.

The effects of supporting structures are most pronounced at close antenna spacings to the tower and with large support dimensions. The result is a measurable increase in gain in one direction and a partial null in the other direction (sometimes 15 dB deep). The shape of the supporting structure also influences pattern distortion. Many antenna manufacturers publish radiation patterns showing the effect of side mounting antennas in their catalogs.

Side mounting is not always a disadvantage. In cases where more (or less) coverage is desired in one direction, the supporting structure can be used to advantage. If pattern distortion is not acceptable, a solution is to mount antennas around the perimeter of the structure and feed them with the proper phasing to synthesize an omnidirectional pattern. Many manufacturers make antennas to accommodate such situations.

The effects of different mounting locations and arrangements can be illustrated with an array of exposed dipoles, **Fig 8**. Such an array is a very versatile antenna because, with simple rearrangement of the elements, it can develop either an omnidirectional pattern or an offset pattern. Drawing A of Fig 8 shows a basic collinear array of four vertical ½-λ elements. The vertical spacing between adjacent elements is 1 λ. All elements are fed in phase. If this array is placed in the clear and supported by a nonconducting mast, the calculated radiation resistance of each dipole element is on the order of 63 Ω. If the feed line is completely decoupled, the resulting azimuth pattern is omnidirectional. The vertical-plane pattern is shown in **Fig 9**.

Fig 8B shows the same array in a side-mounting arrangement, at a spacing of ¼ λ from a conducting mast. In this mounting arrangement, the mast takes on the role of a reflector, producing an F/B on the order of 5.7 dB. The azimuth pattern is shown in **Fig 10**. The vertical pattern is not significantly different from that of Fig 9, except the four small minor lobes (two on either side of the vertical axis) tend to become distorted. They are not as "clean," tending to merge into one minor lobe at some mast heights. This apparently is a function of currents in the supporting mast. The proximity of the mast also alters the feed-point impedance. For elements that are resonant in the configuration of Fig 8A, the calculated impedance in the arrangement of Fig 8B is in the order of 72 + *j* 10 Ω.

If side mounting is the only possibility and an omnidirectional pattern is required, the arrangement of Fig 8C may be used. The calculated azimuth pattern takes on a slight cloverleaf shape, but is within 1.5 dB of being circular. However, gain performance suffers, and the idealized vertical pattern of Fig 9 is not achieved. See **Fig 11**. Spacings other than 1/4-λ from the mast were not investigated.

One very important consideration in side mounting an antenna is mechanical integrity. As with all repeater components, reliability is of great importance. An antenna hanging by the feed line and banging against the tower provides far from optimum performance and reliability. Use a mount that is appropriately secured to the tower and the antenna. Also use good hardware, preferably stainless steel (or bronze). If your local hardware store does not carry stainless steel hardware, try a boating supplier.

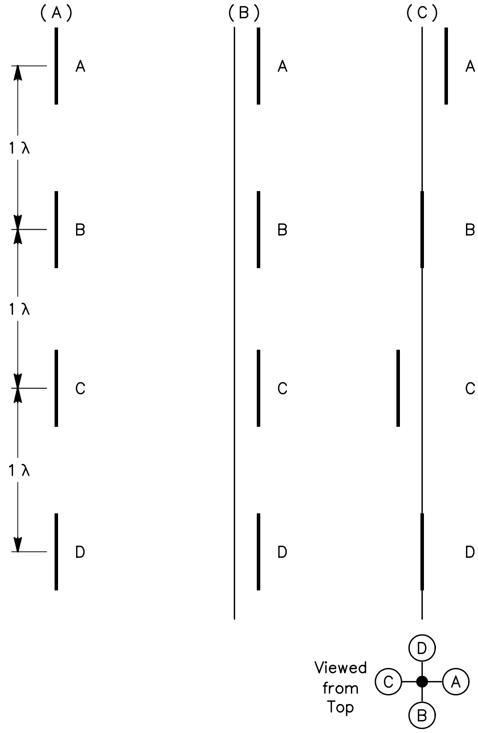

Fig 8—Various arrangements of exposed dipole elements. At A is the basic collinear array of four elements. B shows the same elements mounted on the side of a mast, and C shows the elements in a side-mounted arrangement around the mast for omnidirectional coverage. See text and Figs 9 through 11 for radiation-pattern information.

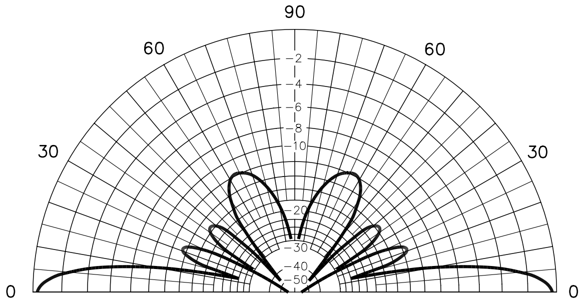

Fig 9—Calculated vertical-plane pattern of the array of Fig 8A, assuming a nonconducting mast support and complete decoupling of the feeder. In azimuth the array is omnidirectional. The calculated gain of the array is 8.6 dBi at 0° elevation; the –3 dB point is at 6.5°.

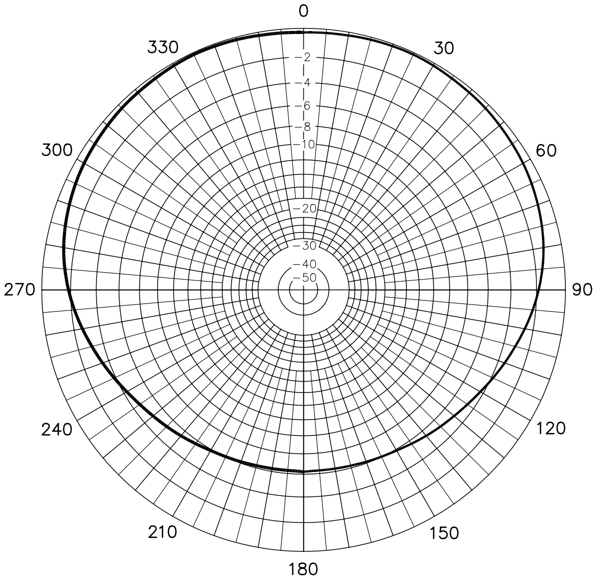

Fig 10—Calculated azimuth pattern of the side-mounted array of Fig 8B, assuming 1/4-λ spacing from a 4-inch mast. The calculated gain in the favored direction, away from the mast and through the elements, is 10.6 dBi.

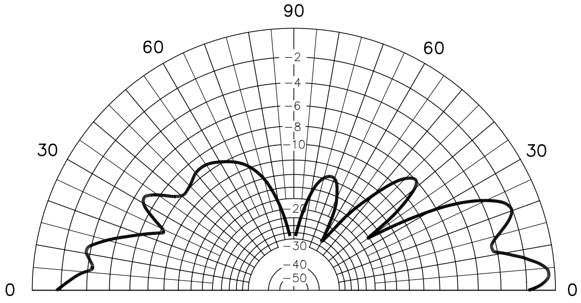

Fig 11—Calculated vertical pattern of the array of Fig 8C, assuming 1/4-λ element spacing from a 4-inch mast. The azimuth pattern is circular within 1.5 dB, and the calculated gain is 4.4 dBi.

Be certain that the feed line is properly supported along its length. Long lengths of cable are subject to contraction and expansion with temperature from season to season, so it is important that the cable not be so tight that contraction causes it to stress the connection at the antenna. This can cause the connection to become intermittent (and noisy) or, at worst, an open circuit. This is far from a pleasant situation if the antenna connection is 300 feet up a tower, and it happens to be the middle of the winter!

Effects of Other Conductors

Feed-line proximity and tower-access ladders or cages also have an effect on the radiation patterns of side-mounted antennas. This subject was studied by Connolly and Blevins, and their findings are given in *IEEE Conference Proceedings* (see the Bibliography at the end of this chapter). Those considering mounting antennas on air-conditioning evaporators or maintenance penthouses on commercial buildings should consult this article. It gives considerable information on the effects of these structures on both unidirectional and omnidirectional antennas.

Metallic guy wires also affect antenna radiation patterns. Yang and Willis studied this and reported the results in *IRE Transactions on Vehicular Communications*. As expected, the closer the antenna is to the guy wires, the greater the effect on the radiation patterns. If the antennas are near the point where the guy wires meet the tower, the effect of the guy wires can be minimized by breaking them up with insulators every 0.75 λ for 2.25 λ to 3.0 λ.

ISOLATION REQUIREMENTS IN REPEATER ANTENNA SYSTEMS

Because repeaters generally operate in full *duplex* (the transmitter and receiver operate simultaneously), the antenna system must act as a filter to keep the transmitter from blocking the receiver. The degree to which the transmitter and receiver must be isolated is a complex problem. It is quite dependent on the equipment used and the difference in transmitter and receiver frequencies (offset). Instead of going into great detail, a simplified example can be used for illustration.

Consider the design of a 144-MHz repeater with a 600-kHz offset. The transmitter has an RF output power of 10 watts, and the receiver has a squelch sensitivity of 0.1 μV. This means there must be at least 1.9×10^{-16} watts at the 52-Ω receiver-antenna terminals to detect a signal. If both the transmitter and receiver were on the same frequency, the isolation (attenuation) required between the transmitter and receiver antenna jacks to keep the transmitter from activating the receiver would be

$$\text{Isolation} = 10 \log \frac{10 \text{ watts}}{1.9 \times 10^{-16} \text{ watts}} = 167 \text{ dB}$$

Obviously there is no need for this much attenuation, because the repeater does not transmit and receive on the same frequency.

If the 10-watt transmitter has noise 600 kHz away from the carrier frequency that is 45 dB below the carrier power, that 45 dB can be subtracted from the isolation requirement. Similarly, if the receiver can detect a 0.1 μV on-frequency signal in the presence of a signal 600 kHz away that is 40 dB greater than 0.1 μV, this 40 dB can also be subtracted from the isolation requirement. Therefore, the isolation requirement is

$$167 \text{ dB} - 45 \text{ dB} - 40 \text{ dB} = 82 \text{ dB}$$

Other factors enter into the isolation requirements as well. For example, if the transmitter power is increased by 10 dB (from 10 to 100 watts), this 10 dB must be added to the isolation requirement. Typical requirements for 144- and 440-MHz repeaters are shown in **Fig 12**.

Obtaining the required isolation is the first problem to be considered in constructing a repeater antenna system. There are three common ways to obtain this isolation:

1) Physically separate the receiving and transmitting antennas so the combination of path loss for the spacing and the antenna radiation patterns results in the required isolation.
2) Use a combination of separate antennas and high-Q filters to develop the required isolation. (The high-Q filters serve to reduce the physical distance required between antennas.)
3) Use a combination filter and combiner system to

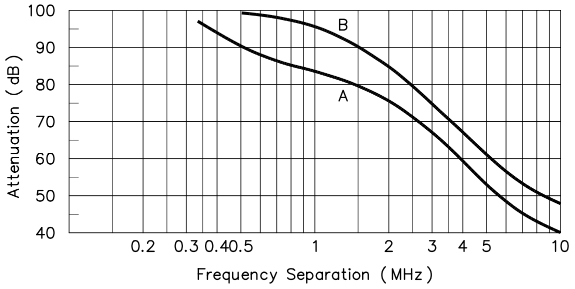

Fig 12—Typical isolation requirements for repeater transmitters and receivers operating in the 132-174 MHz band (Curve A), and the 400-512 MHz band (Curve B). Required isolation in dB is plotted against frequency separation in MHz. These curves were developed for a 100-W transmitter. For other power levels, the isolation requirements will differ by the change in decibels relative to 100 W. Isolation requirements will vary with receiver sensitivity. (The values plotted were calculated for transmitter-carrier and receiver-noise suppression necessary to prevent more than 1 dB degradation in receiver 12-dB SINAD sensitivity.)

allow the transmitter and receiver to share one antenna. Such a filter and combiner is called a *duplexer*.

Repeaters operating on 28 and 50 MHz generally use separate antennas to obtain the required isolation. This is largely because duplexers in this frequency range are both large and very expensive. It is generally less expensive to buy two antennas and link the sites by a committed phone line or an RF link than to purchase a duplexer. At 144 MHz and higher, duplexers are more commonly used. Duplexers are discussed in greater detail in a later section.

Separate Antennas

Receiver desensing (gain limiting caused by the presence of a strong off-frequency signal) can be reduced, and often eliminated, by separation of the transmitting and receiving antennas. Obtaining the 55 to 90 dB of isolation required for a repeater antenna system requires separate antennas to be spaced a considerable distance apart (in wavelengths).

Fig 13 shows the distances required to obtain specific values of isolation for vertical dipoles having horizontal separation (at A) and vertical separation (at B). The isolation gained by using separate antennas is subtracted from the total isolation requirement of the system. For example, if the transmitter and receiver antennas for a 450-MHz repeater are separated horizontally by 400 feet, the total isolation requirement in the system is reduced by about 64 dB.

Note from Fig 13B that a vertical separation of only about 25 feet also provides 64 dB of isolation. Vertical separation yields much more isolation than does horizontal separation. Vertical separation is also more practical than horizontal, since only a single support is required.

An explanation of the significant difference between the two graphs is in order. The vertical spacing requirement for 60 dB attenuation (isolation) at 150 MHz is about 43 feet. The horizontal spacing for the same isolation level is on the order of 700 feet. **Fig 14** shows why this difference exists. The radiation patterns of the antennas at A overlap; each antenna has gain in the direction of the other. The path loss between the antennas is given by

$$\text{Path loss (dB)} = 20 \log \frac{4\pi d}{\lambda} \qquad \text{(Eq 1)}$$

where

d = distance between antennas

λ = wavelength, in the same units as d.

The isolation between the antennas in Fig 14A is the path loss less the antenna gains. Conversely, the antennas at B share pattern nulls, so the isolation is the path loss added to the depth of these nulls. This significantly reduces the spacing requirement for vertical sepa-

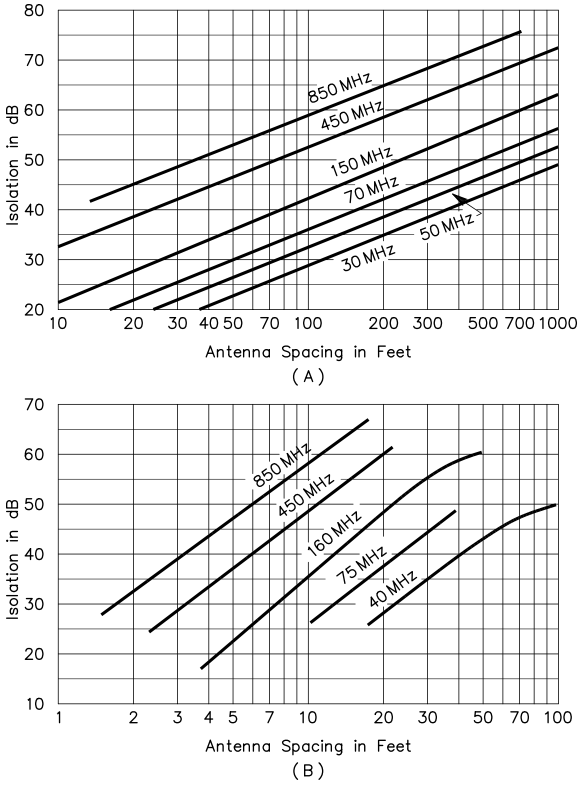

Fig 13—At A, the amount of attenuation (isolation) provided by horizontal separation of vertical dipole antennas. At B, isolation afforded by vertical separation of vertical dipoles.

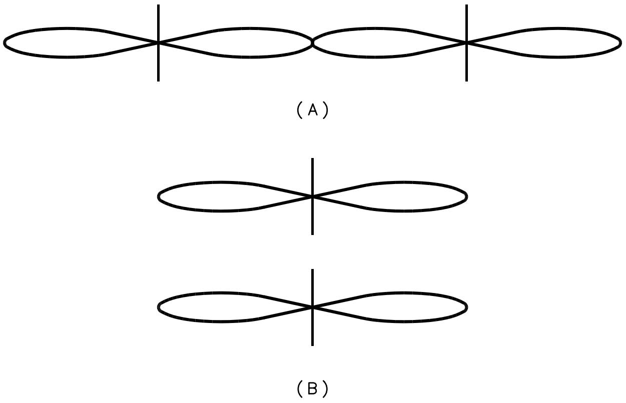

Fig 14—A relative representation of the isolation advantage afforded by separating antennas horizontally (A) and vertically (B) is shown. A great deal of isolation is provided by vertical separation, but horizontal separation requires two supports and much greater distance to be as effective. Separate-site repeaters (those with transmitter and receiver at different locations) benefit much more from horizontal separation than do single-site installations.

ration. Because the depth of the pattern nulls is not infinite, some spacing is required. Combined horizontal and vertical spacing is much more difficult to quantify because the results are dependent on both radiation patterns and the positions of the antennas relative to each other.

Separate antennas have one major disadvantage: They create disparity in transmitter and receiver coverage. For example, say a 50-MHz repeater is installed over average terrain with the transmitter and repeater separated by 2 miles. If both antennas had perfect omnidirectional coverage, the situation depicted in **Fig 15** would exist. In this case, stations able to hear the repeater may not be able to access it, and vice versa. In practice, the situation can be considerably worse. This is especially true if the patterns of both antennas are not omnidirectional. If this disparity in coverage cannot be tolerated, the solution involves skewing the patterns of the antennas until their coverage areas are essentially the same.

Cavity Resonators

As just discussed, receiver desensing can be reduced by separating the transmitter and receiver antennas. But the amount of transmitted energy that reaches the receiver input must often be decreased even farther. Other nearby transmitters can cause desensing as well. A *cavity resonator* (cavity filter) can be helpful in solving these problems. When properly designed and constructed, this type of resonator has very high Q. A commercially made cavity is shown in **Fig 16**.

A cavity resonator placed in series with a transmission line acts as a band-pass filter. For a resonator to operate in series, it must have input and output coupling loops (or probes). A cavity resonator can also be connected across (in parallel with) a transmission line. The cavity then acts as a band-reject (notch) filter, greatly attenuating energy at the frequency to which it is tuned. Only one coupling loop or probe is required for this method of filtering. This type of cavity could be used in the receiver line to "notch" the transmitter signal. Several cavities can be connected in series or parallel to increase the attenuation in a given configuration. The graphs of **Fig 17** show the attenuation of a single cavity (A) and a pair of cavities (B).

The only situation in which cavity filters would not help is the case where the off-frequency noise of the transmitter was right on the receiver frequency. With cavity resonators, an important point to remember is that addition of a cavity across a transmission line may change the impedance of the system. This change can be compensated by adding tuning stubs along the transmission line.

Duplexers

The material in this section was prepared by Domenic Mallozzi, N1DM. Most amateur repeaters in the 144-, 220- and 440-MHz bands use duplexers to obtain the necessary transmitter to receiver isolation. Duplexers have been commonly used in commercial repeaters for many years. The duplexer consists of two high-Q filters. One filter is used in the feed line from the transmitter to the antenna, and another between the antenna and the receiver. These filters must have low loss at the frequency to which they are tuned while having very high attenuation at the surrounding frequencies. To meet the high attenuation requirements at frequencies within as little as 0.4% of the frequency to which they are tuned, the filters usually take the form of cascaded transmission-line cavity filters. These are either band-pass filters, or band-pass filters with a rejection notch. (The rejection notch is tuned to the center frequency of the other filter.) The number of cascaded filter sections is determined by the frequency separation and the ultimate attenuation requirements.

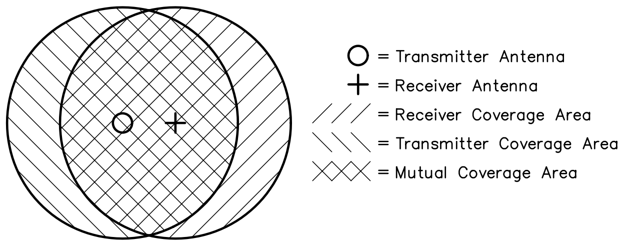
○ = Transmitter Antenna
+ = Receiver Antenna
/// = Receiver Coverage Area
\\\ = Transmitter Coverage Area
XXX = Mutual Coverage Area

Fig 15—Coverage disparity is a major problem for separate-site repeater antennas. The transmitter and receiver coverage areas overlap, but are not entirely mutually inclusive. Solving this problem requires a great deal of experimentation, as many factors are involved. Among these factors are terrain features and distortion of the antenna radiation patterns from supports.

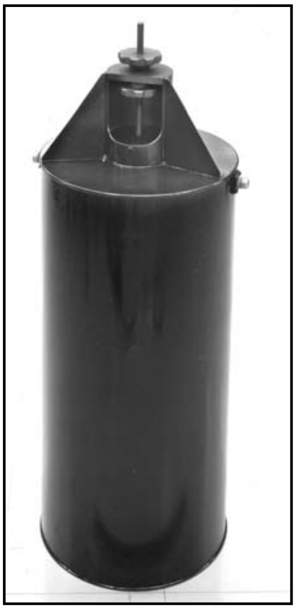

Fig 16—A coaxial cavity filter of the type used in many amateur and commercial repeater installations. Center-conductor length (and thus resonant frequency) is varied by adjustment of the knob (top).

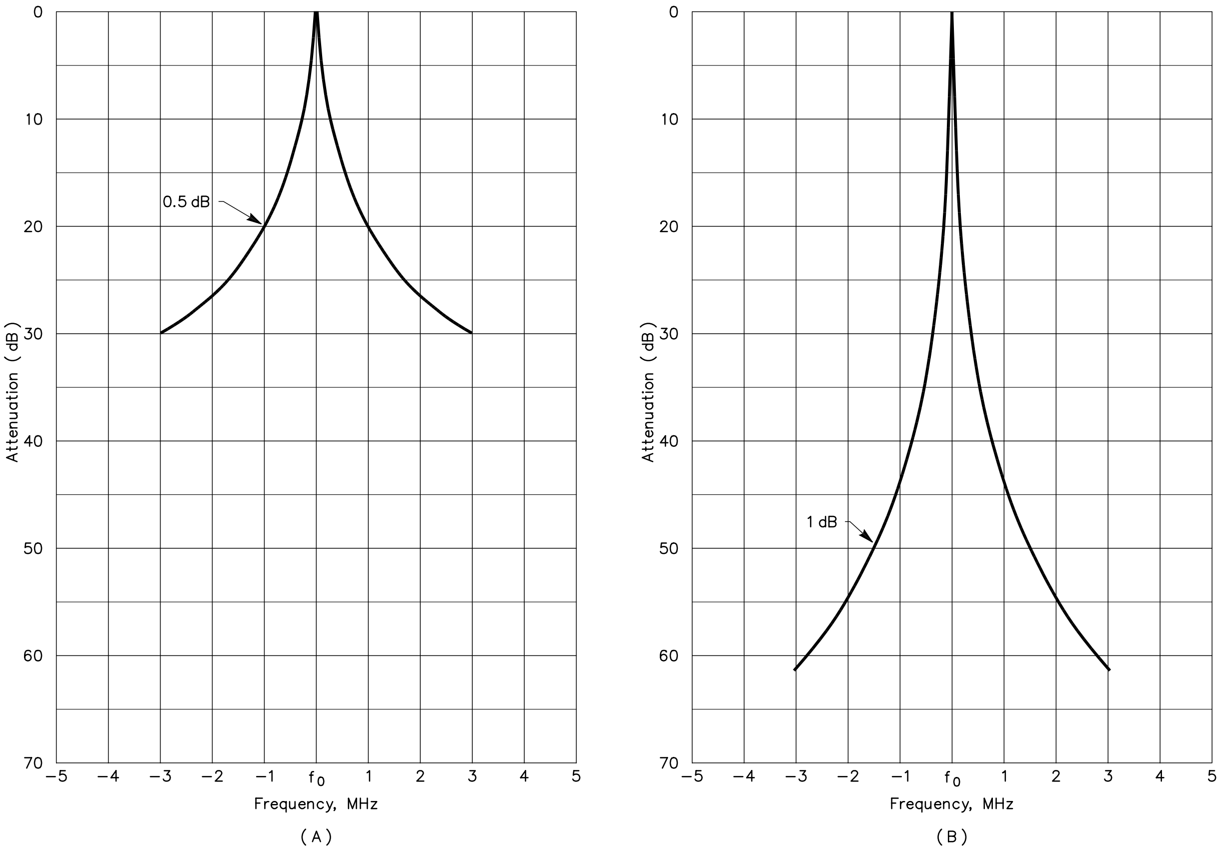

Fig 17—Frequency response curves for a single cavity (A) and two cavities cascaded (B). These curves are for cavities with coupling loops, each having an insertion loss of 0.5 dB. (The total insertion loss is indicated in the body of each graph.) Selectivity will be greater if lighter coupling (greater insertion loss) can be tolerated.

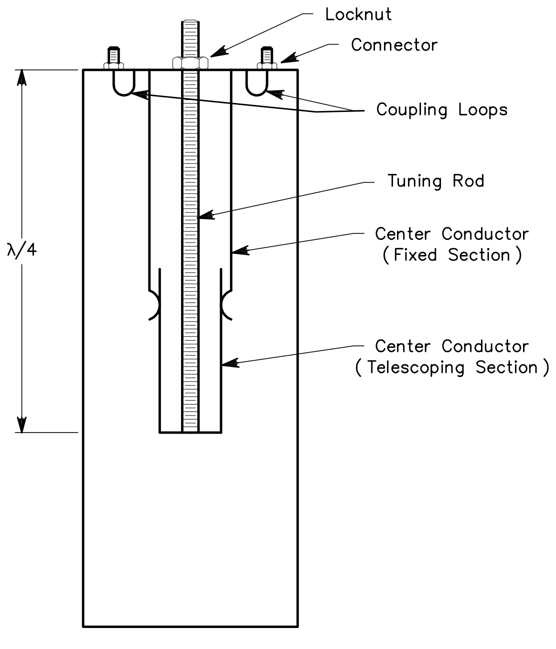

Fig 18—Cutaway view of a typical cavity. Note the relative locations of the coupling loops to each other and to the center conductor of the cavity. A locknut is used to prevent movement of the tuning rod after adjustment.

Duplexers for the amateur bands represent a significant technical challenge, because in most cases amateur repeaters operate with significantly less frequency separation than their commercial counterparts. Information on home construction of duplexers is presented in a later section of this chapter. Many manufacturers market high-quality duplexers for the amateur frequencies.

Duplexers consist of very high-Q cavities whose resonant frequencies are determined by mechanical components, in particular the tuning rod. **Fig 18** shows the cutaway view of a typical duplexer cavity. The rod is usually made of a material that has a limited thermal expansion coefficient (such as Invar). Detuning of the cavity by environmental changes introduces unwanted losses in the antenna system. An article by Arnold in *Mobile Radio Technology* considered the causes of drift in the cavity (see the Bibliography at the end of this chapter). These can be broken into four major categories.

1) Ambient temperature variation (which leads to mechanical variations related to the thermal expansion coefficients of the materials used in the cavity).
2) Humidity (dielectric constant) variation.
3) Localized heating from the power dissipated in the cavity (resulting from its insertion loss).
4) Mechanical variations resulting from other factors (vibration, etc).

In addition, because of the high-Q nature of these cavities, the insertion loss of the duplexer increases when the signal is not at the peak of the filter response. This means, in practical terms, that less power is radiated for a given transmitter output power. Also, the drift in cavities in the receiver line results in increased system noise figure, reducing the sensitivity of the repeater.

As the frequency separation between the receiver and the transmitter decreases, the insertion loss of the duplexer reaches certain practical limits. At 144 MHz, the minimum insertion loss for 600 kHz spacing is 1.5 dB per filter.

Testing and using duplexers requires some special considerations (especially as frequency increases). Because duplexers are very high-Q devices, they are very sensitive to the termination impedances at their ports. A high SWR on any port is a serious problem, because the apparent insertion loss of the duplexer will increase, and the isolation may appear to decrease. Some have found that when duplexers are used at the limits of their isolation capabilities, a small change in antenna SWR is enough to cause receiver desensitization. This occurs most often under ice-loading conditions on antennas with open-wire phasing sections.

The choice of connectors in the duplexer system is important. BNC connectors are good for use below 300 MHz. Above 300 MHz, their use is discouraged because even though many types of BNC connectors work well up to 1 GHz, older style standard BNC connectors are inadequate at UHF and above. Type N connectors should be used above 300 MHz. It is false economy to use marginal quality connectors. Some commercial users have reported deteriorated isolation in commercial UHF repeaters when using such connectors. The location of a bad connector in a system is a complicated and frustrating process. Despite all these considerations, the duplexer is still the best method for obtaining isolation in the 144- to 925 MHz range.

ADVANCED TECHNIQUES

As the number of available antenna sites decreases and the cost of various peripheral items (such as coaxial cable) increases, amateur repeater groups are required to devise advanced techniques if repeaters are to remain effective. Some of the techniques discussed here have been applied in commercial services for many years, but until recently have not been economically justified for amateur use.

One technique worth consideration is the use of *cross-band couplers*. To illustrate a situation where a cross-band coupler would be useful, consider the following example. A repeater group plans to install 144- and 902-MHz repeaters on the same tower. The group intends to erect both antennas on a horizontal cross arm at the 325-foot level. A 325-foot run of $^7/_8$-inch Heliax costs approximately $2000. If both antennas are to be mounted at the top of the tower, the logical approach would require two separate feed lines. A better solution involves the use of a single feed line for both repeaters, along with a cross-band coupler at each end of the line.

The use of the cross-band coupler is shown in **Fig 19**. As the term implies, the coupler allows two signals on different bands to share a common transmission line. Such couplers cost approximately $300 each. In our hypothetical example, this represents a saving of $1400 over the cost of using separate feed lines. But, as with all compromises, there are disadvantages. Cross-band cou-

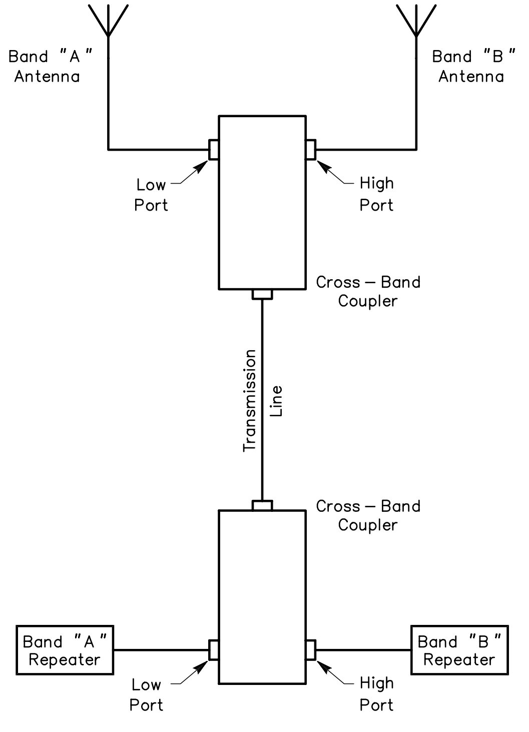

Fig 19—Block diagram of a system using cross-band couplers to allow the use of a single feed line for two repeaters. If the feeder to the antenna location is long (more than 200 feet or so), cross-band couplers may provide a significant saving over separate feed lines, especially at the higher amateur repeater frequencies. Cross-band couplers cannot be used with two repeaters on the same band.

plers have a loss of about 0.5 dB per unit. Therefore, the pair required represents a loss of 1.0 dB in *each* transmission path. If this loss can be tolerated, the cross-band coupler is a good solution.

Cross-band couplers do not allow two repeaters *on the same band* to share a single antenna and feed line. As repeater sites and tower space become more scarce, it may be desirable to have two repeaters on the same band share the same antenna. The solution to this problem is the use of a *transmitter multicoupler*. The multicoupler is related to the duplexers discussed earlier. It is a cavity filter and combiner that allows multiple transmitters and receivers to share the same antenna. This is a common commercial practice. A block diagram of a multicoupler system is shown in **Fig 20**.

The multicoupler, however, is a very expensive device, and has the disadvantage of even greater loss per transmission path than the standard duplexer. For example, a well-designed duplexer for 600 kHz spacing at 146 MHz has a loss per transmission path of approximately 1.5 dB. A four-channel multicoupler (the requirement for two repeaters) has an insertion loss per transmission path on the order of 2.5 dB or more. Another constraint of such a system is that the antenna must present a good match to the transmission line at all frequencies on which it will be used (both transmitting and receiving). This becomes difficult for the system with two repeaters operating at opposite ends of a band.

If you elect to purchase a commercial base-station antenna that requires you to specify a frequency to which the antenna must be tuned, be sure to indicate to the manufacturer the intended use of the antenna and the frequency extremes. In some cases, the only way the manufacturer can accommodate your request is to provide an antenna with some vertical-beam uptilt at one end of the band and some downtilt at the other end of the band. In the case of antennas with very high gain, this in itself may become a serious problem. Careful analysis of the situation is necessary before assembling such a system.

Diversity Techniques for Repeaters

Mobile flutter, "dead spots" and similar problems are a real problem for the mobile operator. The popularity of hand-held transceivers using low power and mediocre antennas causes similar problems. A solution to these difficulties is the use of some form of *diversity reception*. Diversity reception works because signals do not fade at the same rate when received by antennas at different locations (space diversity) or of different polarizations (polarization diversity).

Repeaters with large transmitter coverage areas often have difficulty "hearing" low power stations in peripheral areas or in dead spots. Space diversity is especially useful in such a situation. Space diversity utilizes separate receivers at different locations that are linked to the repeater. The repeater uses a circuit called a *voter* that determines which receiver has the best signal, and then selects the appropriate receiver from which to feed the repeater transmitter. This technique is helpful in urban areas where shadowing from large buildings and bridges causes problems. Space-diversity receiving, when properly executed, can give excellent results. But with the improvement come some disadvantages: added initial cost, maintenance costs, and the possibility of failure created by the extra equipment required. If installed and maintained carefully, problems are generally minimal.

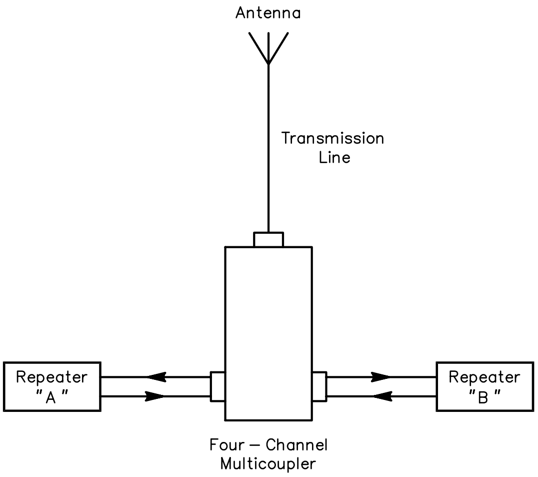

Fig 20—Block diagram of a system using a transmitter multicoupler to allow a single feed line and antenna to be used by two repeaters on one band. The antenna must be designed to operate at all frequencies that the repeaters utilize. More than two repeaters can be operated this way by using a multicoupler with the appropriate number of input ports.

A second improvement technique is the use of *circularly polarized* repeater antennas. This technique has been used in the FM broadcast field for many years, and has been considered for use in the mobile telephone service as well. Some experiments by amateurs have proved very promising, as discussed by Pasternak and Morris (see the Bibliography at the end of this chapter).

The improvement afforded by circular polarization is primarily a reduction in *mobile flutter*. The flutter on a mobile signal is caused by reflections from large buildings (in urban settings) or other terrain features. These reflections cause measurable polarization shifts, sometimes to the point where a vertically polarized signal at the transmitting site may appear to be primarily horizontally polarized after reflection.

A similar situation results from *multipath propagation*, where one or more reflected signals combine with the direct signal at the repeater, having varying effects on the signal. The multipath signal is subjected to large amplitude and phase variations at a relatively rapid rate.

In both of the situations described here, circular polarization can offer considerable improvement. This is because circularly polarized antennas respond equally to all linearly polarized signals, regardless of the plane of polarization. At this writing, there are no known sources of commercial circularly polarized omnidirectional antennas for the amateur bands. Pasternak and Morris describe a circularly polarized antenna made by modifying two commercial four-pole arrays.

EFFECTIVE ISOTROPIC RADIATED POWER (EIRP)

It is useful to know effective isotropic radiated power (EIRP) in calculating the coverage area of a repeater. The FCC formerly required EIRP to be entered in the log of every amateur repeater station. Although logging EIRP is no longer required, it is still useful to have this information on hand for repeater-coordination purposes and so system performance can be monitored periodically.

Calculation of EIRP is straightforward. The PEP output of the transmitter is simply multiplied by the gains and losses in the transmitting antenna system. (These gains and losses are best added or subtracted in decibels and then converted to a multiplying factor.) The following worksheet and example illustrates the calculations.

Feed-line loss ______ dB
Duplexer loss ______ dB
Isolator loss ______ dB
Cross-band coupler loss ______ dB
Cavity filter loss ______ dB

Total losses (L) ______ dB

G (dB) = antenna gain (dBi) –L

where G = antenna system gain. (If antenna gain is specified in dBd, add 2.14 dB to obtain the gain in dBi.)

$M = 10^{G/10}$

where M = multiplying factor

EIRP (watts) = transmitter output (PEP)×M

Example

A repeater transmitter has a power output of 50 W PEP (50-W FM transmitter). The transmission line has 1.8 dB loss. The duplexer used has a loss of 1.5 dB, and a circulator on the transmitter port has a loss of 0.3 dB. There are no cavity filters or cross-band couplers in the system. Antenna gain is 5.6 dBi.

Feed-line loss	1.8 dB
Duplexer loss	1.5 dB
Isolator loss	0.3 dB
Cross-band coupler loss	0 dB
Cavity filter loss	0 dB
Total losses (L)	3.6 dB

Antenna system gain in dB = G = antenna gain (dBi) –L

G = 5.6 dBi – 3.6 dB = 2 dB

Multiplying factor = $M = 10^{G/10}$

$M = 10^{2/10} = 1.585$

EIRP (watts) = transmitter output (PEP) × M

EIRP = 50 W × 1.585 = 79.25 W

If the antenna system is lossier than this example, G may be *negative*, resulting in a multiplying factor less than one. The result is an EIRP that is less than the transmitter output power. This situation can occur in practice, but for obvious reasons is not desirable.

Assembling a Repeater Antenna System

This section will aid you in planning and assembling your repeater antenna system. The material was prepared by Domenic Mallozzi, N1DM. Consult Chapter 23, Radio Wave Propagation, for information on propagation for the band of your interest.

First, a repeater antenna selection checklist such as this will help you in evaluating the antenna system for your needs.

Gain needed ________ dBi
Pattern required ________ Omnidirectional
________ Offset
________ Cardioidal
________ Bidirectional
________ Special pattern (specify)
____
Mounting ________ Top of tower
________ Side of tower

(Determine effects of tower on pattern. Is the result consistent with the pattern required?)

Is downtilt required? ________ Yes
________ No
Type of RF connector ________ UHF
________ N
________ BNC
________ Other (specify)
Size (length) ________
Weight ________
Maximum cost $_______

Table 1 (see next page) has been compiled to provide general information on commercial components available for repeater and remote-base antenna systems. The various components are listed in a matrix format by manufacturer, for equipment designed to operate in the various amateur bands. See Chapter 21, Antenna Products Suppliers, for further information for these components. Although every effort has been made to make this data complete, the ARRL is not responsible for omissions or errors. The listing of a product in Table 1 does not constitute an endorsement by ARRL. Manufacturers are urged to contact the editors with updating information.

Even though almost any antenna can be used for a repeater, the companies indicated in the *Antennas* column in Table 1 are known to have produced heavy-duty antennas to commercial standards for repeater service. Many of these companies offer their antennas with special features for repeater service (such as vertical-beam downtilt). It is best to obtain catalogs of current products from the manufacturers listed, both for general information and to determine which special options are available on their products.

A 144 MHz Duplexer

Obtaining sufficient isolation between the transmitter and receiver of a repeater can be difficult. Many of the solutions to this problem compromise receiver sensitivity or transmitter power output. Other solutions create an imbalance between receiver and transmitter coverage areas. When a duplexer is used, insertion loss is the compromise. But a small amount of insertion loss is more than offset by the use of one antenna for both the transmitter and receiver. Using one antenna assures equal antenna patterns for both transmitting and receiving, and reduces cost, maintenance and mechanical complexity.

As mentioned earlier in this chapter, duplexers may be built in the home workshop. Bob Shriner, WA0UZO, presented a small, mechanically simple duplexer for low-power applications in April 1979 *QST*. Shriner's design is unique, as the duplexer cavities are constructed of circuit-board material. Low cost and simplicity are the result, but with a trade-off in performance. A silver-plated version of Shriner's design has an insertion loss of approximately 5 dB at 146 MHz. The loss is greater if the copper is not plated, and increases as the inner walls of the cavities tarnish.

This duplexer construction project by John Bilodeau, W1GAN, represents an effective duplexer. The information originally appeared in July 1972 *QST*. It is a time-proven project used by many repeater groups, and can be duplicated relatively easily. Its insertion loss is just 1.5 dB.

Fig 21 will help you visualize the requirements for a duplexer, which can be summed up as follows. The duplexer must attenuate the transmitter carrier to avoid overloading the receiver and thereby reducing its sensitivity. It must also attenuate any noise or spurious frequencies from the transmitter on or near the receiver frequency. In addition, a duplexer must provide a proper impedance match between transmitter, antenna, and receiver.

Table 1
Product Matrix Showing Repeater Equipment and Manufacturer by Frequency Band

	Antennas							*Duplexers*				*Cavity Filters*			
Source	*28*	*50*	*144*	*220*	*450*	*902*	*1296*	*144*	*220*	*450*	*902*	*144*	*220*	*450*	*902*
Austin	S	S	S	S	S	S	S								
Celwave	C	C	C	C	C	C		C	C	C	C	C	C	C	C
Comet															
Cushcraft		C	C	C	C										
Dec Prod		C	C	C	C	C		C	C		C	C		C	
MA/COM												C		C	
RF Parts			C		C	C	C								
Sinclair	C	C	C					C		C		C		C	C
TX/RX								C	C	C		C		C	C
Wacom															

	Isolators/Circulators						*Transmitter Combiners*				*Cross-Band Couplers*			
Source	*28*	*50*	*144*	*220*	*450*	*902*	*144*	*220*	*450*	*902*	*0-174 450-512*	*0-512 800-960*	*59-174 806-960*	*406-512 806-960*
Celwave			C	C	C	C	C	C	C	C		S		
Dec Prod			C		C	C		C	C	C				
Sinclair			C		C	C								
TX/RX			C		C	C	C	C	C	C	C	C	C	C
Wacom			C	C	C	C	C	C	C	C		C		C

Abbreviated names above are for the following manufacturers: Austin Antennas, Celwave RF Inc, Cushcraft Corp, Decibel Products Inc, RF Parts, Sinclair Radio Laboratories Inc, TX/RX Systems Inc and Wacom Inc. A manufacturer's contact list appears in Chapter 21.

Key to codes used:
C = catalog (standard) item
S = special-order item

Note: Coaxial cable is not listed, because most manufacturers sell only to dealers.

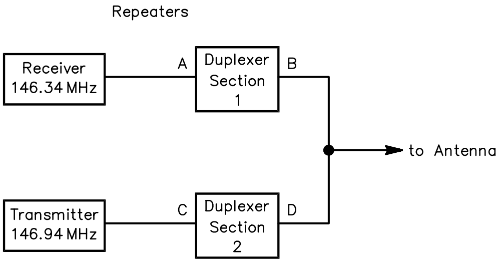

Fig 21—Duplexers permit using one antenna for both transmitting and receiving in a repeater system. Section 1 prevents energy at the transmitter frequency from interfering with the receiver, while section 2 attenuates any off-frequency transmitter energy that is at or near the receiver frequency.

As shown in Fig 21, transmitter output on 146.94 MHz going from point C to D should not be attenuated. However, the transmitter energy should be greatly attenuated between points B and A. Duplexer section 2 should attenuate any noise or signals that are on or near the receiver input frequency of 146.34 MHz. For good reception the noise and spurious signal level must be less than –130 dBm (0 dBm = 1 milliwatt into 50 Ω). Typical transmitter noise 600 kHz away from the carrier frequency is 80 dB below the transmitter power output. For 60 watts of output (+48 dBm), the noise level is –32 dBm. The duplexer must make up the difference between –32 dBm and –130 dBm, or 98 dB.

The received signal must go from point B to A with a minimum of attenuation. Section 1 of the duplexer must also provide enough attenuation of the transmitter energy to prevent receiver overload. For an average receiver, the transmitter signal must be less than –30 dBm to meet this requirement. The difference between the transmitter output of +48 dBm and the receiver overload point of –30 dBm, 78 dB, must be made up by duplexer section 1.

THE CIRCUIT

Fig 22 shows the completed 6-cavity duplexer, and **Fig 23** shows the assembly of an individual cavity. A $^1/_4$-λ resonator was selected for this duplexer design. The length of the center conductor is adjusted by turning a

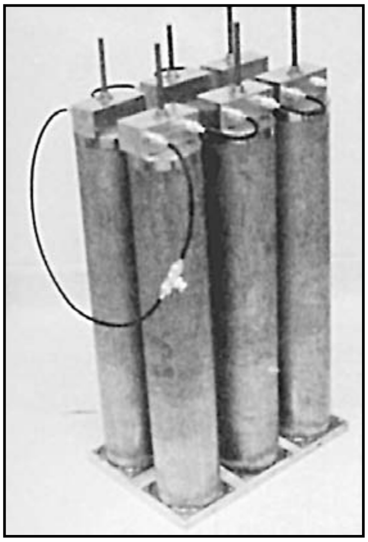

Fig 22—A six-cavity duplexer for use with a 144-MHz repeater. The cavities are fastened to a plywood base for mechanical stability. Short lengths of double-shielded cable are used for connections between individual cavities. An insertion loss of less than 1.5 dB is possible with this design.

threaded rod, which changes the resonant frequency of the cavity. Energy is coupled into and out of the tuned circuit by the coupling loops extending through the top plate.

The cavity functions as a series resonant circuit. When a reactance is connected across a series resonant circuit, an anti-resonant notch is produced, and the resonant frequency is shifted. If a capacitor is added, the notch appears below the resonant frequency. Adding inductance instead of capacitance makes the notch appear above the resonant frequency. The value of the added component determines the spacing between the notch and the resonant frequency of the cavity.

Fig 24 shows the measured band-pass characteristics of the cavity with shunt elements. With the cavity tuned to 146.94 MHz and a shunt capacitor connected from input to output, a 146.34-MHz signal is attenuated

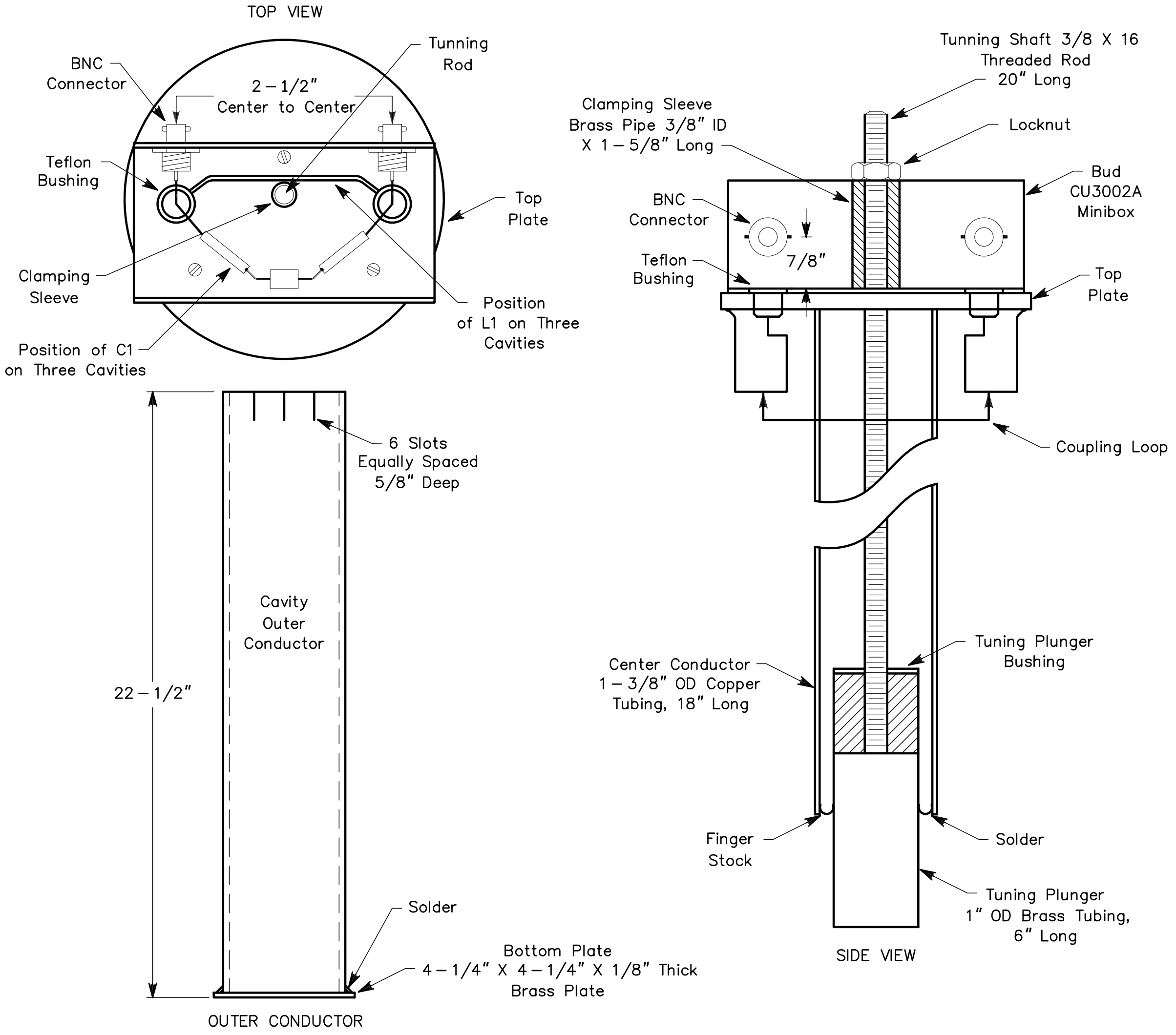
TOP VIEW
BNC Connector
2-1/2" Center to Center
Tunning Rod
Teflon Bushing
Top Plate
Clamping Sleeve
Position of L1 on Three Cavities
Position of C1 on Three Cavities
6 Slots Equally Spaced 5/8" Deep
Cavity Outer Conductor
22-1/2"
Solder
Bottom Plate 4-1/4" X 4-1/4" X 1/8" Thick Brass Plate
OUTER CONDUCTOR
Tunning Shaft 3/8 X 16 Threaded Rod 20" Long
Clamping Sleeve Brass Pipe 3/8" ID X 1-5/8" Long
Locknut
Bud CU3002A Minibox
BNC Connector
7/8"
Teflon Bushing
Top Plate
Coupling Loop
Center Conductor 1-3/8" OD Copper Tubing, 18" Long
Tuning Plunger Bushing
Finger Stock
Solder
Tuning Plunger 1" OD Brass Tubing, 6" Long
SIDE VIEW

Fig 23—The assembly of an individual cavity. A Bud Minibox is mounted on the top plate with three screws. A clamping sleeve made of brass pipe is used to prevent crushing the box when the locknut is tightened on the tuning shaft. Note that the positions of both C1 and L1 are shown, but that three cavities will have C1 installed and three will have L1 in place.

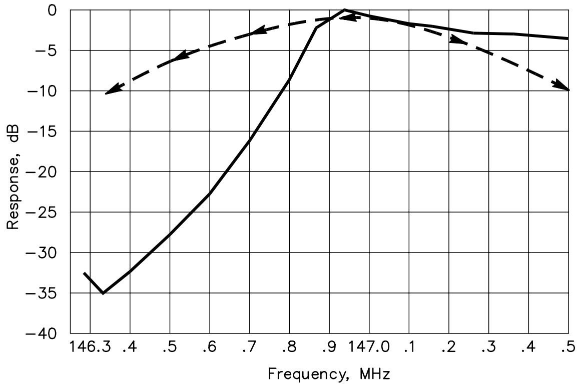

Fig 24—Typical frequency response of a single cavity of the type used in the duplexer. The dotted line represents the passband characteristics of the cavity alone; the solid line for the cavity with a shunt capacitor connected between input and output. An inductance connected in the same manner will cause the rejection notch to be above the frequency to which the cavity is tuned.

by 35 dB. If an inductance is placed across the cavity and the cavity is tuned to 146.34 MHz, the attenuation at 146.94 MHz is 35 dB. Insertion loss in both cases is 0.4 dB. Three cavities with shunt capacitors are tuned to 146.94 MHz and connected together in cascade with short lengths of coaxial cable. The attenuation at 146.34 MHz is more than 100 dB, and insertion loss at 146.94 MHz is 1.5 dB. Response curves for a six-cavity duplexer are given in **Fig 25**.

Construction

The schematic diagram for the duplexer is shown in **Fig 26**. Three parts for the duplexer must be machined; all others can be made with hand tools. A small lathe can be used to machine the brass top plate, the threaded tuning plunger bushing and the Teflon insulator bushing. The dimensions of these parts are given in **Fig 27**.

Type DWV copper tubing is used for the outer conductor of the cavities. The wall thickness is 0.058 inch, with an outside diameter of $4\frac{1}{8}$ inches. You will need a tubing cutter large enough to handle this size (perhaps borrowed or rented). The wheel of the cutter should be tight and sharp. Make slow, careful cuts so the ends will be square. The outer conductor is $22\frac{1}{2}$ inches long.

The inner conductor is made from type M copper tubing having an outside diameter of $1\frac{3}{8}$ inches. A 6-inch length of 1-inch OD brass tubing is used to make the tuning plunger.

The tubing types mentioned above are designations used in the plumbing and steam-fitting industry. Other types may be used in the construction of a duplexer, but you should check the sizes carefully to assure that the parts will fit each other. A greater wall thickness will make the assembly heavier, and the expense will increase accordingly. Soft solder is used throughout the assembly. Unless you have experience with silver solder, do not use it. Eutectic type 157 solder with paste or acid

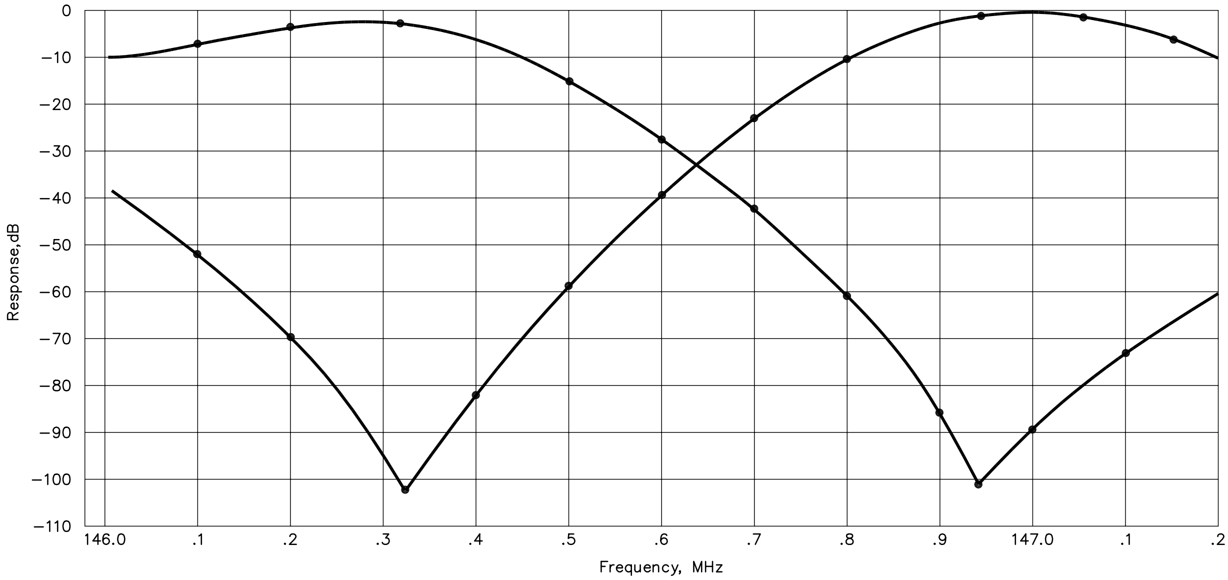

Fig 25—Frequency response of the six-cavity duplexer. One set of three cavities is tuned to pass 146.34 MHz and notch 146.94 MHz (the receiver leg). The remaining set of three cavities is tuned to pass 146.94 MHz and notch 146.34 MHz. This duplexer provides approximately 100 dB of isolation between the transmitter and receiver when properly tuned.

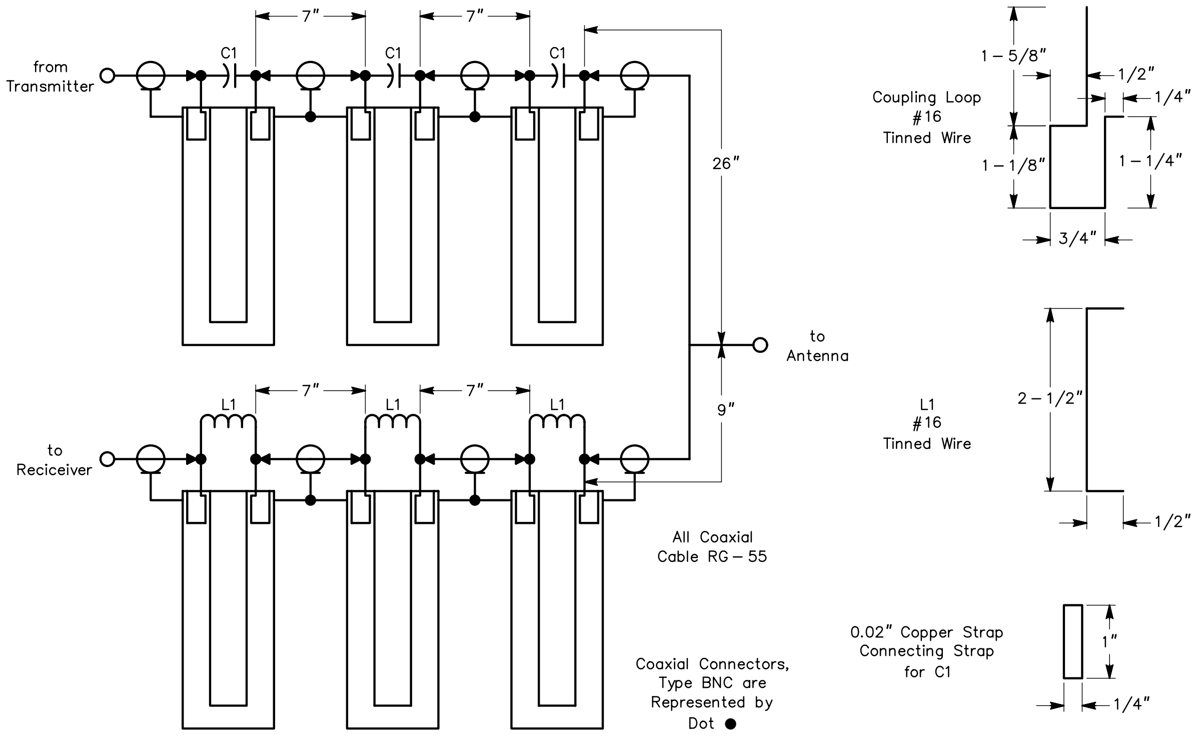

from Transmitter
C1
7"
C1
7"
C1
26"
to Antenna
9"
L1
7"
L1
7"
L1
to Reciceiver
All Coaxial Cable RG – 55
Coaxial Connectors, Type BNC are Represented by Dot ●
Coupling Loop #16 Tinned Wire
1 – 5/8"
1/2"
1/4"
1 – 1/8"
1 – 1/4"
3/4"
L1 #16 Tinned Wire
2 – 1/2"
1/2"
0.02" Copper Strap Connecting Strap for C1
1"
1/4"

Fig 26—Diagram of the six-cavity duplexer. Coaxial cable lengths between cavities are critical and must be followed closely. Double shielded cable and high quality connectors should be used throughout. The sizes and shapes of the coupling loops, L1, and the straps for connecting C1 should be observed.
C1—1.7-11 pF circuit-board mount, E. F. Johnson 189-5-5 or equiv. Set at $^3/_4$ closed for initial alignment.

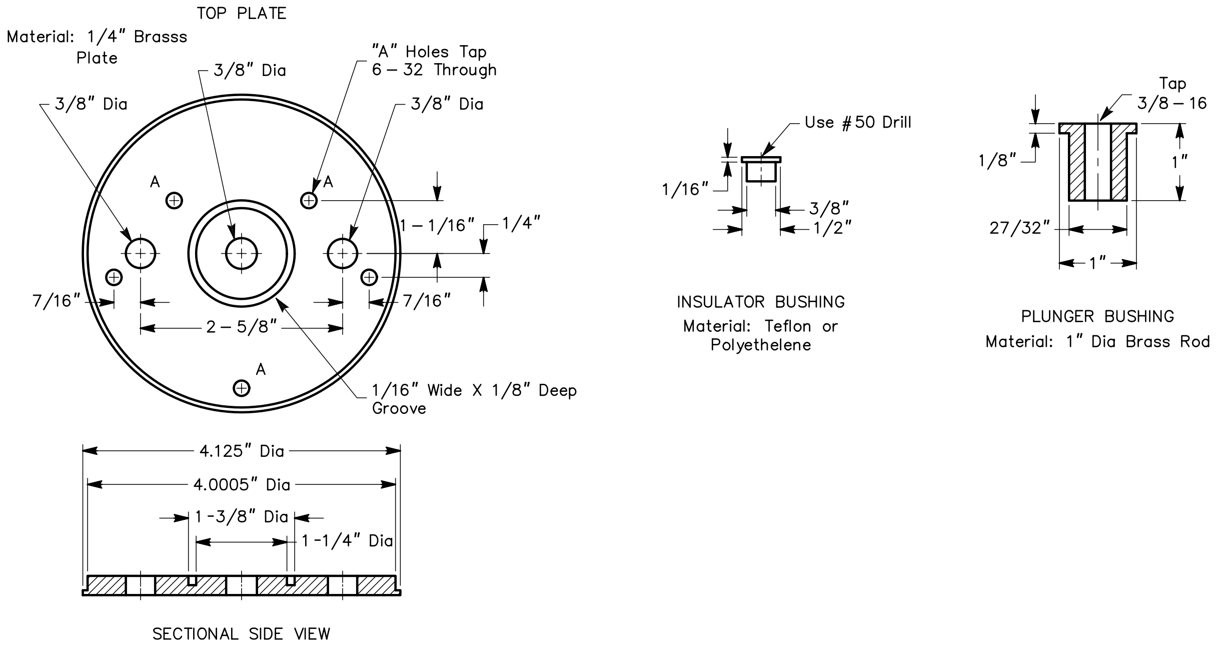

TOP PLATE
Material: 1/4" Brasss Plate
3/8" Dia
"A" Holes Tap 6 – 32 Through
3/8" Dia
3/8" Dia
A
A
1 – 1/16"
1/4"
7/16"
2 – 5/8"
7/16"
A
1/16" Wide X 1/8" Deep Groove
4.125" Dia
4.0005" Dia
1 -3/8" Dia
1 -1/4" Dia
SECTIONAL SIDE VIEW
Use #50 Drill
1/16"
3/8"
1/2"
INSULATOR BUSHING
Material: Teflon or Polyethelene
Tap 3/8 – 16
1/8"
1"
27/32"
1"
PLUNGER BUSHING
Material: 1" Dia Brass Rod

Fig 27—Dimensions for the three parts that require machining. A small metal-working lathe should be used for making these parts.

flux makes very good joints. This type has a slightly higher melting temperature than ordinary tin-lead alloy, but has considerably greater strength.

First solder the inner conductor to the top plate (**Fig 28**). The finger stock can then be soldered inside the lower end of the inner conductor, while temporarily held in place with a plug made of aluminum or stainless steel. While soldering, do not allow the flame from the torch to overheat the finger stock. The plunger bushing is soldered into the tuning plunger and a 20-inch length of threaded rod is soldered into the bushing.

Cut six slots in the top of the outer conductor. They should be $^{5}/_{8}$ inch deep and equally spaced around the tubing. The bottom end of the 4-inch tubing is soldered to the square bottom plate. The bottom plates have holes in the corners so they can be fastened to a plywood base by means of wood screws. Because the center conductor has no support at one end, the cavities must be mounted vertically.

The size and position of the coupling loops are critical. Follow the given dimensions closely. Both loops should be $^{1}/_{8}$ inch away from the center conductor on opposite sides. Connect a solder lug to the ground end of the loop, then fasten the lug to the top plate with a screw. The free end of the loop is insulated by Teflon bushings where it passes through the top plate for connection to the BNC fittings.

Before final assembly of the parts, clean them thoroughly. Soap-filled steel wool pads and hot water work well for this. Be sure the finger stock makes firm contact with the tuning plunger. The top plate should fit snugly in the top of the outer conductor—a large hose clamp tightened around the outer conductor will keep the top plate in place.

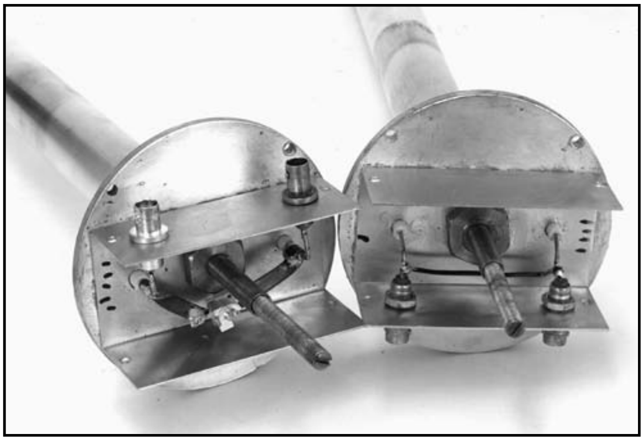

Fig 28—Two of the center conductor and top plate assemblies. In the assembly at the left, C1 is visible just below the tuning shaft, mounted by short straps made from sheet copper. The assembly on the right has L1 in place between the BNC connectors. The Miniboxes are fastened to the top plate by a single large nut in these units. Using screws through the Minibox into the top plate, as described in the text, is preferred.

ADJUSTMENT

After the cavities have been checked for band-pass characteristics and insertion loss, install the anti-resonant elements, C1 and L1. (See Fig 24.) It is preferable to use laboratory test equipment when tuning the duplexer. An option is to use a low-power transmitter with an RF probe and an electronic voltmeter. Both methods are shown in **Fig 29**.

With the test equipment connected as shown in Fig 29A, adjust the signal generator frequency to the desired repeater input frequency. Connect a calibrated step attenuator between points X and Y. With no attenuation, adjust the HP-415 for 0 on the 20-dB scale. You can check the calibration of the 415 by switching in different amounts of attenuation and noting the meter reading. You may note a small error at either high or very low signal levels.

Next, remove the step attenuator and replace it with a cavity that has the shunt inductor, L1, in place. Adjust the tuning screw for maximum reading on the 415 meter. Remove the cavity and connect points X and Y. Set the

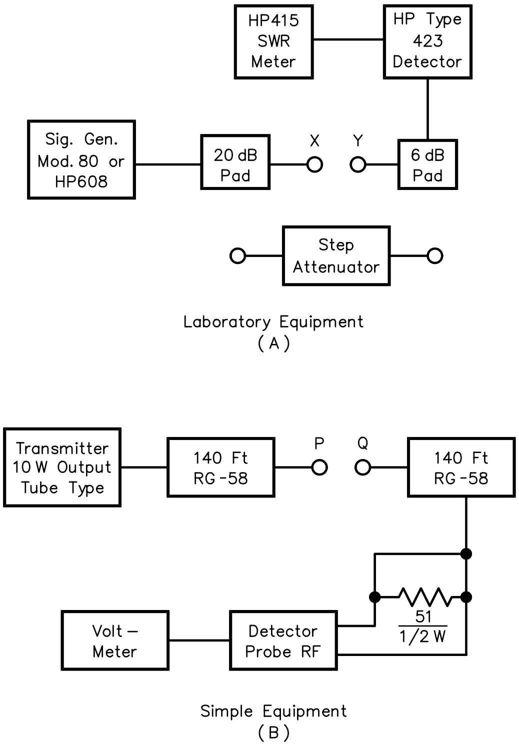

Fig 29—The duplexer can be tuned by either of the two methods shown here, although the method depicted at A is preferred. The signal generator should be modulated by a 1-kHz tone. If the setup shown at B is used, the transmitter should not be modulated, and should have a minimum of noise and spurious signals. The cavities to be aligned are inserted between X and Y in the setup at A, and between P and Q in B.

signal generator to the repeater output frequency and adjust the 415 for a 0 reading on the 20-dB scale.

Reinsert the cavity between X and Y and adjust the cavity tuning for minimum reading on the 415. The notch should be sharp and have a depth of at least 35 dB. It is important to maintain the minimum reading on the meter while tightening the locknut on the tuning shaft.

To check the insertion loss of the cavity, the output from the signal generator should be reduced, and the calibration of the 415 meter checked on the 50-dB expanded scale. Use a fixed 1-dB attenuator to make certain the error is less than 0.1 dB. Replace the attenuator with the cavity and read the loss. The insertion loss should be 0.5 dB or less. The procedure is the same for tuning all six cavities, except that the frequencies are reversed for those having the shunt capacitor installed.

Adjustment with Minimum Equipment

A transmitter with a minimum of spurious output is required. Most modern transmitters meet this requirement. The voltmeter in use should be capable of reading 0.5 volt (or less), full scale. The RF probe used should be rated to 150 MHz or higher. Sections of RG-58 cable are used as attenuators, as shown in Fig 29B. The loss in these 140-foot lengths is nearly 10 dB, and helps to isolate the transmitter in case of mismatch during tuning.

Set the transmitter to the repeater input frequency and connect P and Q. Obtain a reading between 1 and 3 volts on the voltmeter. Insert a cavity with shunt capacitors in place between P and Q and adjust the cavity tuning for a minimum reading on the voltmeter. (This reading should be between 0.01 and 0.05 volt.) The rejection in dB can be calculated by

$$dB = 20 \log (V1/V2)$$

This should be at least 35 dB. Check the insertion loss by putting the receiver on the repeater output frequency and noting the voltmeter reading with the cavity out of the circuit. A 0.5-dB attenuator can be made from a 7-foot length of RG-58. This 7-foot cable can be used to check the calibration of the detector probe and the voltmeter.

Cavities with shunt inductance can be tuned the same way, but with the frequencies reversed. If two or more cavities are tuned while connected together, transmitter noise can cause the rejection readings to be low. In other words, there will be less attenuation.

Results

The duplexer is conservatively rated at 150 watts input, but if constructed carefully should be able to handle as much as 300 watts. Silver plating the interior surfaces of the cavities is recommended if input power is to be greater than 150 watts. A duplexer of this type with silver-plated cavities has an insertion loss of less than 1 dB, and a rejection of more than 100 dB. Unplated cavities should be disassembled at least every two years, cleaned thoroughly, and then retuned.

Miscellaneous Notes

1) Double shielded cable and high quality connectors are *required* throughout the system.
2) The SWR of the antenna should not exceed 1.2:1 for proper duplexer performance.
3) Good shielding of the transmitter and receiver at the repeater is essential.
4) The antenna should have four or more wavelengths of vertical separation from the repeater.
5) Conductors in the near field of the antenna should be well bonded and grounded to eliminate noise.
6) The feed line should be electrically bonded and mechanically secured to the tower or mast.
7) Feed lines and other antennas in the near field of the repeater antenna should be well bonded and as far from the repeater antenna as possible.
8) Individual cavities can be used to improve the performance of separate antenna or separate site repeaters.
9) Individual cavities can be used to help solve intermodulation problems.

BIBLIOGRAPHY

Source material and more extended discussions of the topics covered in this chapter can be found in the references below.

P. Arnold, “Controlling Cavity Drift in Low-Loss Combiners,” *Mobile Radio Technology*, Apr 1986, pp 36-44.

L. Barrett, “Repeater Antenna Beam Tilting,” *Ham Radio*, May 1983, pp 29-35. (See correction, *Ham Radio*, Jul 1983, p 80.)

W. F. Biggerstaff, “Operation of Close Spaced Antennas in Radio Relay Systems,” *IRE Transactions on Vehicular Communications*, Sep 1959, pp 11-15.

J. J. Bilodeau, “A Homemade Duplexer for 2-Meter Repeaters,” *QST*, Jul 1972, pp 22-26, 47.

W. B. Bryson, “Design of High Isolation Duplexers and a New Antenna for Duplex Systems,” *IEEE Transactions on Vehicular Communications*, Mar 1965, pp 134-140.

K. Connolly and P. Blevins, “A Comparison of Horizontal Patterns of Skeletal and Complete Support Structures,” *IEEE 1986 Vehicular Technology Conference Proceedings*, pp 1-7.

S. Kozono, T. Tsuruhara and M. Sakamoto, “Base Station Polarization Diversity Reception for Mobile Radio,” *IEEE Transactions on Vehicular Technology*, Nov 1984, pp 301-306.

J. Kraus, *Antennas*, 2nd ed. (New York: McGraw-Hill Book Co., 1988).

W. Pasternak and M. Morris, *The Practical Handbook of Amateur Radio FM & Repeaters*, (Blue Ridge Summit, PA: Tab Books Inc., 1980), pp 355-363.

M. W. Scheldorf, “Antenna-To-Mast Coupling in Communications," *IRE Transactions on Vehicular Communications*, Apr 1959, pp 5-12.

R. D. Shriner, "A Low Cost PC Board Duplexer," *QST*, Apr 1979, pp 11-14.

W. V. Tilston, "Simultaneous Transmission and Reception with a Common Antenna," *IRE Transactions on Vehicular Communications*, Aug 1962, pp 56-64.

E. P. Tilton, "A Trap-Filter Duplexer for 2-Meter Repeaters," *QST*, Mar 1970, pp 42-46.

R. Wheeler, "Fred's Advice solves Receiver Desense Problem," *Mobile Radio Technology*, Feb 1986, pp 42-44.

R. Yang and F. Willis, "Effects of Tower and Guys on Performance of Side Mounted Vertical Antennas," *IRE Transactions on Vehicular Communications*, Dec 1960, pp 24-31.

Chapter 18

VHF and UHF Antenna Systems

A good antenna system is one of the most valuable assets available to the VHF/UHF enthusiast. Compared to an antenna of lesser quality, an antenna that is well designed, is built of good quality materials, and is well maintained, will increase transmitting range, enhance reception of weak signals and reduce interference problems. The work itself building antennas is by no means the least attractive part of the job. Even with high-gain antennas, experimentation is greatly simplified at VHF and UHF because the antennas are a physically manageable size. Setting up a home antenna range is within the means of most amateurs, and much can be learned about the nature and adjustment of antennas. No large investment in test equipment is necessary.

The Basics

Selecting the best VHF or UHF antenna for a given installation involves much more than scanning gain figures and prices in a manufacturer's catalog. There is no one "best" VHF or UHF antenna design for all purposes. The first step in choosing an antenna is figuring out what you want it to do.

Gain

At VHF and UHF, it is possible to build Yagi antennas with very high gain—15 to 20 dBi—on a physically manageable boom. Such antennas can be combined in arrays of two, four, six, eight or more antennas. These arrays are attractive for EME, tropospheric scatter or other weak-signal communications modes.

Radiation Patterns

Antenna radiation can be made omnidirectional, bidirectional, practically unidirectional, or anything between these conditions. A VHF net operator may find an omnidirectional system almost a necessity, but it may be a poor choice otherwise. Noise pickup and other interference problems are greater with such omnidirectional antennas, and omnidirectional antennas having some gain are especially bad in these respects. Maximum gain and low radiation angle are usually prime interests of the weak-signal DX aspirant. A clean pattern, with lowest possible pickup and radiation off the sides and back, may be important in high-activity areas, or where the noise level is high.

Frequency Response

The ability to work over an entire VHF band may be important in some types of work. Modern Yagis can achieve performance over a remarkably wide frequency range, providing that the boom length is long enough and enough elements are used to populate the boom. Modern Yagi designs in fact are competitive with directly driven collinear arrays of similar size and complexity. The primary performance parameters of gain, front-to-rear ratio and SWR can be optimized over all the VHF or UHF amateur bands readily, with the exception of the full 6meter band from 50.0 to 54.0 MHz, which is an 8% wide bandwidth. A Yagi can be easily designed to cover any 2.0 MHz portion of the 6-meter band with superb performance.

Height Gain

In general, higher is better in VHF and UHF antenna installations. Raising the antenna over nearby obstructions may make dramatic improvements in coverage. Within reason, greater height is almost always worth its cost, but height gain (see Chapter 23, Radio Wave Propagation) must be balanced against increased transmission-line loss. This loss can be considerable, and it increases with frequency. The best available line may not be very good if the run is long in terms of wavelengths. Line loss considerations (see Chapter 24, Transmission Lines) are important in antenna planning.

Physical Size

A given antenna design for 432 MHz has the same gain as the same design for 144 MHz, but being only one-third as large intercepts only one-ninth as much energy in receiving. In other words, the antenna has less pickup efficiency at 432 MHz. To be equal in communication effectiveness, the 432-MHz array should be at least equal in *size* to the 144-MHz antenna, which requires roughly three times as many elements. With all the extra difficulties involved in using the higher frequencies effectively, it is best to keep antennas as large as possible for these bands.

DESIGN FACTORS

With the objectives sorted out in a general way, decisions on specifics, such as polarization, type of transmission line, matching methods and mechanical design must be made.

Polarization

Whether to position antenna elements vertically or horizontally has been widely debated since early VHF pioneering days. Tests have shown little evidence about which polarization sense is most desirable. On long propagation paths there is no consistent advantage either way. Shorter paths tend to yield higher signal levels with horizontally polarized antennas over some kinds of terrain. Man-made noise, especially ignition interference, also tends to be lower with horizontal antennas. These factors make horizontal polarization somewhat more desirable for weak-signal communications. On the other hand, vertically polarized antennas are much simpler to use in omnidirectional systems and in mobile work.

Vertical polarization was widely used in early VHF work, but horizontal polarization gained favor when directional arrays started to become widely used. The major use of FM and repeaters, particularly in the VHF/UHF bands, has tipped the balance in favor of vertical antennas in mobile and repeater use. Horizontal polarization predominates in other communication on 50 MHz and higher frequencies. An additional loss of 20 dB or more can be expected when cross-polarized antennas are used.

TRANSMISSION LINES

Transmission line principles are covered in detail in Chapter 24, Transmission Lines. Techniques that apply to VHF and UHF operation are dealt with in greater detail here. The principles of carrying RF from one location to another via a feed line are the same for all radio frequencies. As at HF, RF is carried principally via open wire lines and coaxial cables at VHF/UHF. Certain aspects of these lines characterize them as good or bad for use above 50 MHz.

Properly built open-wire line can operate with very low loss in VHF and UHF installations. A total line loss under 2 dB per 100 feet at 432 MHz can easily be obtained. A line made of #12 wire, spaced $^3/_4$ inch or more with Teflon spreaders and run essentially straight from antenna to station, can be better than anything but the most expensive coax. Such line can be home made or purchased at a fraction of the cost of coaxial cables, with comparable loss characteristics. Careful attention must be paid to efficient impedance matching if the benefits of this system are to be realized. A similar system for 144 MHz can easily provide a line loss under 1 dB.

Small coax such as RG-58 or RG-59 should never be used in VHF work if the run is more than a few feet. Lines of $^1/_2$-inch diameter (RG-8 or RG-11) work fairly well at 50 MHz, and are acceptable for 144-MHz runs of 50 feet or less. These lines are somewhat better if they employ foam instead of ordinary PE dielectric material. Aluminum-jacket *hardline* coaxial cables with large inner conductors and foam insulation are well worth their cost, and can sometimes be obtained for free from local Cable TV operators as “end runs”—pieces at the end of a roll. The most common CATV cable is $^1/_2$-inch OD 75-Ω hardline. Matched-line loss for this cable is about 1.0 dB/100 feet at 146 MHz and 2.0 dB/100 feet at 432 MHz. Less commonly available from CATV companies is the $^3/_4$-inch 75Ω hardline, sometimes with a black self-healing hard plastic covering. This line has 0.8 dB of loss per 100 feet at 146 MHz, and 1.6 dB loss per 100 feet at 432 MHz. There will be small additional losses for either line if 75-to-50Ω transformers are used at each end.

Commercial connectors for hardline are expensive but provide reliable connections with full waterproofing. Enterprising amateurs have homebrewed low-cost connectors. If they are properly water proofed, connectors and hardline can last almost indefinitely. Hardline must not be bent too sharply, because it will kink. See Chapter 24, Transmission Lines, for details on hardline connectors.

Beware of any “bargains” in coax for VHF or UHF use. Feed-line loss can be compensated to some extent by increasing transmitter power, but once lost, a weak signal can never be recovered in the receiver. Effects of weather on transmission lines should not be ignored. Well-constructed open-wire line works optimally in nearly any weather, and it stands up well. Twin-lead is almost useless in heavy rain, wet snow or icing. The best grades of coax are completely impervious to weather—they can be run underground, fastened to metal towers without insulation and bent into any convenient position with no adverse effects on performance.

WAVEGUIDES

Above 2 GHz, coaxial cable is a losing proposition for communications work. Fortunately, at this frequency the wavelength is short enough to allow practical, efficient energy transfer by an entirely different means. A *waveguide* is a conducting tube through which energy is transmitted in the form of electromagnetic waves. The tube is not considered as carrying a current in the same sense that the wires of a two-conductor line do, but rather as a *boundary* that confines the waves in the enclosed space. Skin effect prevents any electromagnetic effects from being evident outside the guide. The energy is injected at one end, either through capacitive or inductive coupling or by radiation, and is removed from the other end in a like manner. Waveguide merely confines the energy of the fields, which are propagated through it to the receiving end by means of reflections against its inner walls.

Analysis of waveguide operation is based on the assumption that the guide material is a perfect conductor of electricity. Typical distributions of electric and magnetic fields in a rectangular guide are shown in **Fig 1**. The intensity of the electric field is greatest (as indicated by closer spacing of the lines of force) at the center along the X dimension (Fig 1C), diminishing to zero at the end walls. The fields must diminish in this manner, because the existence of any electric field parallel to the walls at the surface would cause an infinite current to flow in a perfect conductor. Waveguides, of course, cannot carry RF in this fashion.

Modes of Propagation

Fig 1 represents the most basic distribution of the electric and magnetic fields in a waveguide. There are an infinite number of ways in which the fields can arrange themselves in a waveguide (for frequencies above the low cutoff frequency of the guide in use). Each of these field configurations is called a *mode*.

The modes may be separated into two general groups. One group, designated *TM* (transverse magnetic), has the magnetic field entirely transverse to the direction of propagation, but has a component of the electric field in that direction. The other type, designated *TE* (transverse electric) has the electric field entirely transverse, but has a component of magnetic field in the direction of propagation. TM waves are sometimes called *E waves*, and TE waves are sometimes called *H waves*, but the TM and TE designations are preferred.

The mode of propagation is identified by the group letters followed by two subscript numerals. For example, TE_{10}, TM_{11}, etc. The number of possible modes increases with frequency for a given size of guide, and there is only one possible mode (called the *dominant mode*) for the lowest frequency that can be transmitted. The dominant mode is the one generally used in amateur work.

Waveguide Dimensions

In rectangular guide the critical dimension is X in Fig 1. This dimension must be more than $^1/_2\ \lambda$ at the lowest frequency to be transmitted. In practice, the Y dimension usually is made about equal to $^1/_2$ X to avoid the possibility of operation in other than the dominant mode.

Cross-sectional shapes other than a rectangle can be used, the most important being the circular pipe. Much the same considerations apply as in the rectangular case.

Wavelength dimensions for rectangular and circular guides are given in **Table 1**, where X is the width of a rectangular guide and r is the radius of a circular guide. All figures apply to the dominant mode.

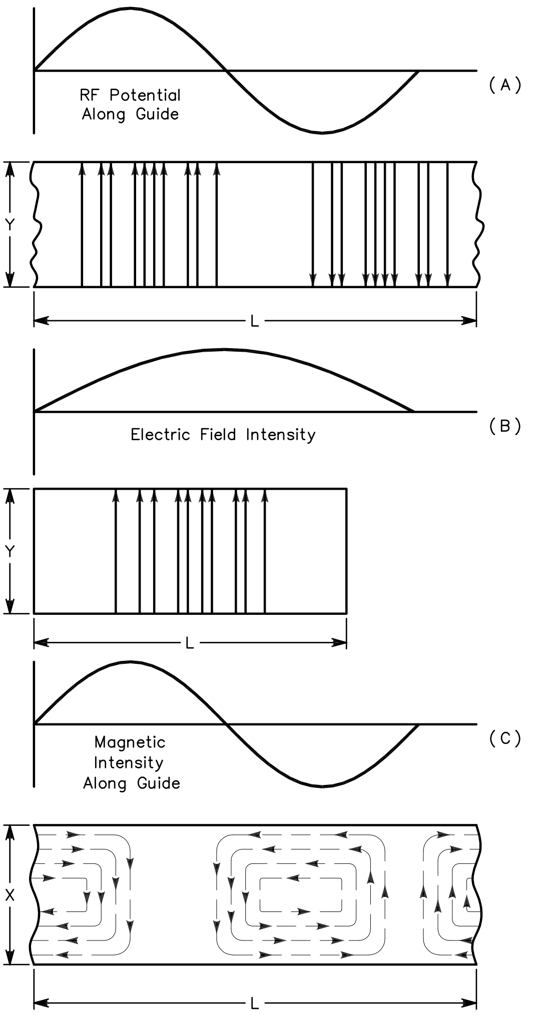

Fig 1—Field distribution in a rectangular waveguide. The TE10 mode of propagation is depicted.

Coupling to Waveguides

Energy may be introduced into or extracted from a waveguide or resonator by means of either the electric or magnetic field. The energy transfer frequently is through a coaxial line. Two methods for coupling to coaxial line are shown in **Fig 2**. The probe shown at A is simply a short extension of the inner conductor of the coaxial line, oriented so that it is parallel to the electric lines of force. The loop shown at B is arranged so that it encloses some of the magnetic lines of force. The point at which maximum coupling is obtained depends upon the mode of propagation in the guide or cavity. Coupling is maximum when the coupling device is in the most intense field.

Coupling can be varied by turning the probe or loop through a 90° angle. When the probe is perpendicular to the electric lines the coupling is minimum. Similarly, when the plane of the loop is parallel to the magnetic lines the coupling is minimum.

If a waveguide is left open at one end it will radiate energy. This radiation can be greatly enhanced by flaring the waveguide to form a pyramidal horn antenna. The horn acts as a transition between the confines of the waveguide and free space. To effect the proper impedance transformation the horn must be at least $^1/_2\ \lambda$ on a side. A horn of this dimension (cutoff) has a unidirectional radiation pattern with a null toward the waveguide transition. The gain at the cutoff frequency is 3 dB, increasing 6 dB with each doubling of frequency. Horns are used extensively in microwave work, both as primary radiators and as feed elements for more elaborate focusing systems. Details for constructing 10-GHz horn antennas are given later in this chapter.

Evolution of a Waveguide

Suppose an open-wire line is used to carry RF energy from a generator to a load. If the line has any appreciable length it must be mechanically supported. The line must be well insulated from the supports if high losses are to be avoided. Because high-quality insulators are difficult to construct at microwave frequencies, the logical alternative is to support the transmission line with $^1/_4\ \lambda$ stubs, shorted at the end opposite the feed line. The open end of such a stub presents an infinite impedance to the transmission line, provided the shorted stub is nonreactive. However, the shorting link has a finite length, and therefore some inductance. The effect of this inductance can be removed by making the RF current flow on the surface of a plate rather than a thin wire. If the plate is large enough, it will prevent the magnetic lines of force from encircling the RF current.

An infinite number of these $^1/_4\ \lambda$ stubs may be connected in parallel without affecting the standing waves of voltage and current. The transmission line may be supported from the top as well as the bottom, and when an infinite number of supports are added, they form the walls of a waveguide at its cutoff frequency. **Fig 3** illustrates how a rectangular waveguide evolves from a two-wire parallel transmission line as described. This simplified analysis also shows why the cutoff dimension is $^1/_2\ \lambda$.

Table 1
Waveguide Dimensions

	Rectangular	*Circular*
Cutoff wavelength	2X	3.41r
Longest wavelength transmitted with little attenuation	1.6X	3.2r
Shortest wavelength before next mode becomes possible	1.1X	2.8r

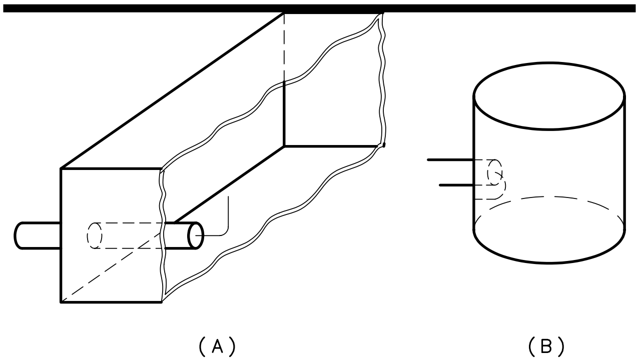

Fig 2—Coupling coaxial line to waveguide and resonators.

While the operation of waveguides is usually described in terms of fields, current does flow on the inside walls, just as on the conductors of a two-wire transmission line. At the waveguide cutoff frequency, the current is concentrated in the center of the walls, and disperses toward the floor and ceiling as the frequency increases.

IMPEDANCE MATCHING

Impedance matching is covered in detail in Chapter 25, Coupling the Transmitter to the Line, and Chapter 26, Coupling the Line to the Antenna. The theory is the same for frequencies above 50 MHz. Practical aspects are similar, but physical size can be a major factor in the choice of methods. Only the matching devices used in practical construction examples later in this chapter are discussed in detail here. This should not rule out consideration of other methods, however, and you should read the relevant portions of both Chapters 25 and 26.

Universal Stub

As its name *universal stub* implies, the double-adjustment stub of **Fig 4A** is useful for many matching purposes. The stub length is varied to resonate the system and the transmission-line attachment point is varied until the transmission line and stub impedances are equal. In practice this involves moving both the sliding short and

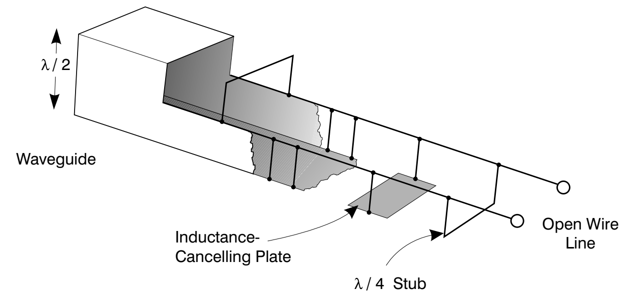

Fig 3—At its cutoff frequency a rectangular waveguide can be thought of as a parallel two-conductor transmission line supported from top and bottom by an infinite number of $^{1}/_{4}$-λ stubs.

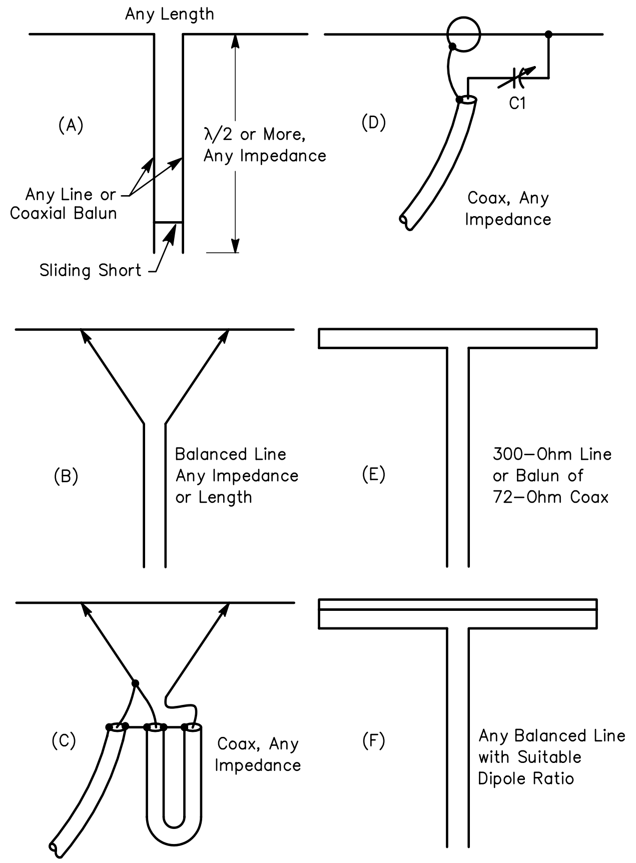

Fig 4—Matching methods commonly used at VHF. The universal stub, A, combines tuning and matching. The adjustable short on the stub and the points of connection of the transmission line are adjusted for minimum reflected power on the line. In the delta match, B and C, the line is fanned out and connected to the dipole at the point of optimum impedance match. Impedances need not be known in A, B or C. The gamma match, D, is for direct connection of coax. C1 tunes out inductance in the arm. A folded dipole of uniform conductor size, E, steps up antenna impedance by a factor of four. Using a larger conductor in the unbroken portion of the folded dipole, F, gives higher orders of impedance transformation.

the point of line connection for zero reflected power, as indicated on an SWR bridge connected in the line.

The universal stub allows for tuning out any small reactance present in the driven part of the system. It permits matching the antenna to the line without knowledge of the actual impedances involved. The position of the short yielding the best match gives some indication of the amount of reactance present. With little or no reactive component to be tuned out, the stub must be approximately $^{1}/_{2}$ λ from load toward the short.

The stub should be made of stiff bare wire or rod, spaced no more than $^{1}/_{20}$ λ apart. Preferably it should be mounted rigidly, on insulators. Once the position of the short is determined, the center of the short can be grounded, if desired, and the portion of the stub no longer needed can be removed.

It is not necessary that the stub be connected directly to the driven element. It can be made part of an open-wire line as a device to match coaxial cable to the line. The stub can be connected to the lower end of a delta match or placed at the feed point of a phased array. Examples of these uses are given later.

Delta Match

Probably the most basic impedance matching device is the *delta match*, fanned ends of an open-wire line tapped onto a $^{1}/_{2}$ λ antenna at the point of the most-efficient power transfer. This is shown in Fig 4B. Both the side length and the points of connection either side of the center of the element must be adjusted for minimum reflected power on the line, but as with the universal stub, you needn't know the impedances. The delta match makes no provision for tuning out reactance, so the universal stub is often used as a termination for it.

At one time, the delta match was thought to be inferior for VHF applications because of its tendency to radiate if improperly adjusted. The delta has come back into favor now that accurate methods are available for measuring the effects of matching. It is very handy for phasing multiple-bay arrays with open-wire lines, and its dimensions in this use are not particularly critical. It should be checked out carefully in applications like that of Fig 4C, where no tuning device is used.

Gamma and T Matches

An application of the same principle allowing direct connection of coax is the *gamma match*, Fig 4D. Because the RF voltage at the center of a $^{1}/_{2}$ λ dipole is zero, the outer conductor of the coax is connected to the element at this point. This may also be the junction with a metallic or wooden boom. The inner conductor, carrying the RF current, is tapped out on the element at the matching point. Inductance of the arm is tuned out by means of C1, resulting in electrical balance. Both the point of contact with the element and the setting of the capacitor are adjusted for zero reflected power, with a bridge connected in the coaxial line.

The capacitance can be varied until the required value is found, and the variable capacitor replaced with a fixed unit of that value. C1 can be mounted in a waterproof box. The maximum required value should be about 100 pF for 50 MHz and 35 to 50 pF for 144 MHz.

The capacitor and arm can be combined in one coaxial assembly, with the arm connected to the driven element by means of a sliding clamp and the inner end of the arm sliding inside a sleeve connected to the center conductor of the coax. An assembly of this type can be constructed from concentric pieces of tubing, insulated by plastic or heat-shrink sleeving. RF voltage across the capacitor is low when the match is adjusted properly, so with a good dielectric, insulation presents no great problem. The initial adjustment should be made with low power. A clean, permanent high-conductivity bond between arm and element is important, since the RF current is high at this point.

Because it is inherently somewhat unbalanced, the gamma match can sometimes introduce pattern distortion, particularly on long-boom, highly directive Yagi arrays. The *T-match*, essentially two gamma matches in series creating a balanced feed system, has become popular for this reason. A coaxial balun like that shown in **Fig 5** is used from the 200 Ω balanced T-match to the unbalanced 50 Ω coaxial line going to the transmitter. See the K1FO Yagi designs later in this chapter for details on practical use of a T-match.

Folded Dipole

The impedance of a $^1/_2$ λ dipole broken at its center is about 70 Ω. If a single conductor of uniform size is folded to make a $^1/_2$ λ dipole as shown in Fig 4E, the impedance is stepped up four times. Such a folded dipole can be fed directly with 300Ω line with no appreciable mismatch. If a 4:1 balun is used, the antenna can be fed with 75Ω coaxial cable. (See balun information presented below.) Higher step-up impedance transformation can be obtained if the unbroken portion is made larger in cross-section than the fed portion, as shown in Fig 4F.

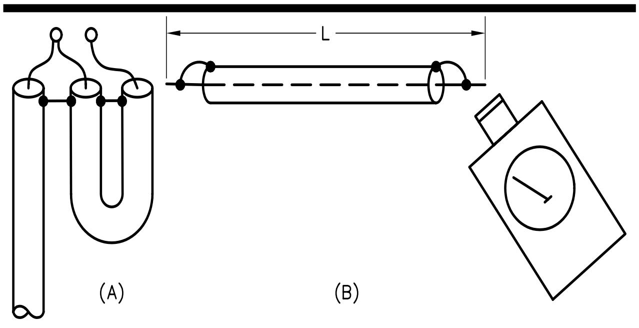

Fig 5—Conversion from unbalanced coax to a balanced load can be done with a ½-λ, coaxial balun at A. Electrical length of the looped section should be checked with a dip meter, with the ends shorted, as at B. The $^1/_2$-λ balun gives a 4:1 impedance step-up.

Hairpin Match

The feed-point resistance of most multielement Yagi arrays is less than 50 Ω. If the driven element is split and fed at the center, it may be shortened from its resonant length to add capacitive reactance at the feed point. Then, shunting the feed point with a wire loop resembling a *hairpin* causes a step-up of the feed-point resistance. The hairpin match is used together with a 4:1 coaxial balun in the 50 MHz arrays described later in this chapter. See Chapter 26, Coupling the Line to the Antenna, for details on the hairpin match.

BALUNS AND ANTENNA TUNERS

Conversion from balanced loads to unbalanced lines (or vice versa) can be performed with electrical circuits, or their equivalents made of coaxial cable. A balun made from flexible coax is shown in Fig 5A. The looped portion is an electrical $^1/_2$ λ. The physical length depends on the velocity factor of the line used, so it is important to check its resonant frequency as shown in Fig 5B. The two ends are shorted, and the loop at one end is coupled to a dip meter coil. This type of balun gives an impedance step-up of 4:1 (typically 50 to 200 Ω, or 75 to 300 Ω).

Coaxial baluns that yield 1:1 impedance transformations are shown in **Fig 6**. The coaxial sleeve, open at the top and connected to the outer conductor of the line at the lower end (A) is the preferred type. At B, a conductor of approximately the same size as the line is used with the outer conductor to form a $^1/_4$ λ stub. Another piece of coax, using only the outer conductor, will serve this purpose. Both baluns are intended to present an infi-

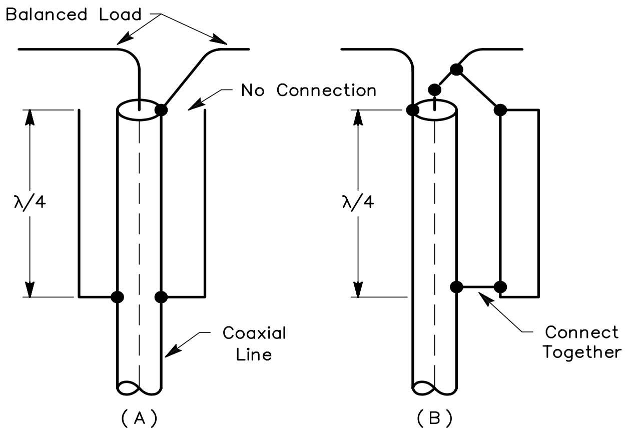

Fig 6—The balun conversion function, with no impedance transformation, can be accomplished with ¼-λ lines, open at the top and connected to the coax outer conductor at the bottom. The coaxial sleeve at A is preferred.

nite impedance to any RF current that might otherwise flow on the outer conductor of the coax.

The functions of the balun and the impedance transformer can be handled by various tuned circuits. Such a device, commonly called an *antenna tuner* or a *Transmatch*, can provide a wide range of impedance transformations. Additional selectivity inherent in the antenna tuner can reduce RFI problems.

THE YAGI AT VHF AND UHF

Without doubt, the Yagi is king of home-station antennas these days. Today's best designs are computer optimized. For years amateurs as well as professionals designed Yagi arrays experimentally. Now we have powerful (and inexpensive) personal computers and sophisticated software for antenna modeling. These have brought us antennas with improved performance, with little or no element pruning required. Chapter 11, HF Yagi Arrays, describes the parameters associated with Yagi-Uda arrays. Except for somewhat tighter dimensional tolerances needed at VHF and UHF, the properties that make a good Yagi at HF also are needed on the higher frequencies. See the end of this chapter for practical Yagi designs.

STACKING YAGIS

Where suitable provision can be made for supporting them, two Yagis mounted one above the other and fed in phase can provide better performance than one long Yagi with the same theoretical or measured gain. The pair occupies a much smaller turning space for the same gain, and their wider elevation coverage can provide excellent results. The wide azimuthal coverage for a vertical stack often results in QSOs that might be missed with a single narrow-beam long-boom Yagi pointed in a different direction. On long ionospheric paths, a stacked pair occasionally may show an *apparent* gain much greater than the measured 2 to 3 dB of stacking gain. (See also the extensive section on stacking Yagis in Chapter 11, HF Yagi Arrays.)

Optimum vertical spacing for Yagis with boom longer than 1 λ or more is about 1 λ (984/50.1 = 19.64 feet), but this may be too much for many builders of 50-MHz antennas to handle. Worthwhile results can be obtained with as little as $^{1}/_{2}$ λ (10 feet), but $^{5}/_{8}$ λ (12 feet) is markedly better. The difference between 12 and 20 feet, however, may not be worth the added structural problems involved in the wider spacing, at least at 50 MHz. The closer spacings give lower measured gain, but the antenna patterns are cleaner in both azimuth and elevation than with 1 λ spacing. Extra gain with wider spacings is usually the objective on 144 MHz and higher-frequency bands, where the structural problems are not as severe.

Yagis can also be stacked in the same plane (collinear elements) for sharper azimuthal directivity. A spacing of $^{5}/_{8}$ λ between the ends of the inner elements yields the maximum gain within the main lobe of the array.

If individual antennas of a stacked array are properly designed, they look like noninductive resistors to the phasing system that connects them. The impedances involved can thus be treated the same as resistances in parallel.

Three sets of stacked dipoles are shown in **Fig 7**. Whether these are merely dipoles or the driven elements of Yagi arrays makes no difference for the purpose of these examples. Two 300 Ω antennas at A are 1 λ apart, resulting in a paralleled feed-point impedance of 150 Ω at the center. (Actually it is slightly less than 150 Ω because of coupling between bays, but this can be neglected for illustrative purposes.) This value remains the same regardless of the impedance of the phasing line. Thus, any convenient line can be used for phasing, as long as the *electrical* length of each line is the same.

The velocity factor of the line must be taken into account as well. As with coax, this is subject to so much variation that it is important to make a resonance check on the actual line used. The method for doing this is shown in Fig 5B. A $^{1}/_{2}$ λ line is resonant both open and shorted, but the shorted condition (both ends) is usually the more convenient test condition.

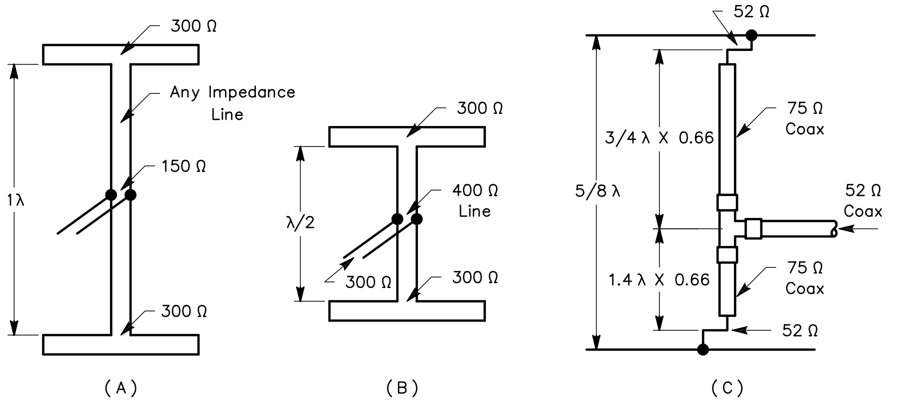

Fig 7—Three methods of feeding stacked VHF arrays. A and B are for bays having balanced driven elements, where a balanced phasing line is desired. Array C has an all-coaxial matching and phasing system. If the lower section is also $^{3}/_{4}$ λ no transposition of line connections is needed.

The impedance transforming property of a $^{1}/_{4}\ \lambda$ line section can be used in combination matching and phasing lines, as shown in Fig 7B and C. At B, two bays spaced $^{1}/_{2}\ \lambda$ apart are phased and matched by a 400-Ω line, acting as a double-Q section, so that a 300-Ω main transmission line is matched to two 300-Ω bays. The two halves of this phasing line could also be $^{3}/_{4}$-λ or $^{5}/_{4}$-λ long, if such lengths serve a useful mechanical purpose. (An example is the stacking of two Yagis where the desirable spacing is more than $^{1}/_{2}\ \lambda$.)

A double-Q section of coaxial line is illustrated in Fig 7C. This is useful for feeding stacked bays that were designed for 50-Ω feed. A spacing of $^{5}/_{8}\ \lambda$ is useful for small Yagis, and this is the equivalent of a full electrical wavelength of solid-dielectric coax such as RG-11.

If one phasing line is electrically $^{1}/_{4}\ \lambda$ and $^{3}/_{4}\ \lambda$ on the other, the connection to one driven element should be reversed with respect to the other to keep the RF currents in the elements in phase—the gamma match is located on opposite sides of the driven elements in Fig 7C. If the number of $^{1}/_{4}\ \lambda$ lengths is the same on either side of the feed point, the two connections should be in the same position, and not reversed. Practically speaking however, you can ensure proper phasing by using exactly equal lengths of line from the same roll of coax. This ensures that the velocity factor for each line is identical.

One marked advantage of coaxial phasing lines is that they can be wrapped around the vertical support, taped or grounded to it, or arranged in any way that is mechanically convenient. The spacing between bays can be set at the most desirable value, and the phasing lines placed anywhere necessary.

Stacking Yagis for Different Frequencies

In stacking horizontal Yagis one above the other on a single rotating support, certain considerations apply when the bays are for different bands. As a very general rule of thumb, the minimum desirable spacing is half the boom length of the higher frequency Yagi.

For example, assume the stacked two-band array of **Fig 8A** is for 50 and 144 MHz. This vertical arrangement is commonly referred to as a *Christmas tree*, because it resembles one. The 50MHz Yagi has 5elements on a 12-foot boom. It tends to look like "ground" to the 8-element 144 MHz Yagi on a 12-foot boom directly above it. [The exact Yagi designs for the examples used in this section are located on the CD-ROM accompanying this book. They may be evaluated as monoband Yagis using the *YW* (Yagi for Windows) program also supplied on the CD-ROM. In each case the bottom Yagi in the stack (at the top of the tower) is assumed to be 20 feet high.]

SWR Change in a Multi-Frequency Stack

Earlier editions of *The ARRL Antenna Book* stated that the feed-point impedance of the higher-frequency antenna would likely be affected the most by the proxim-

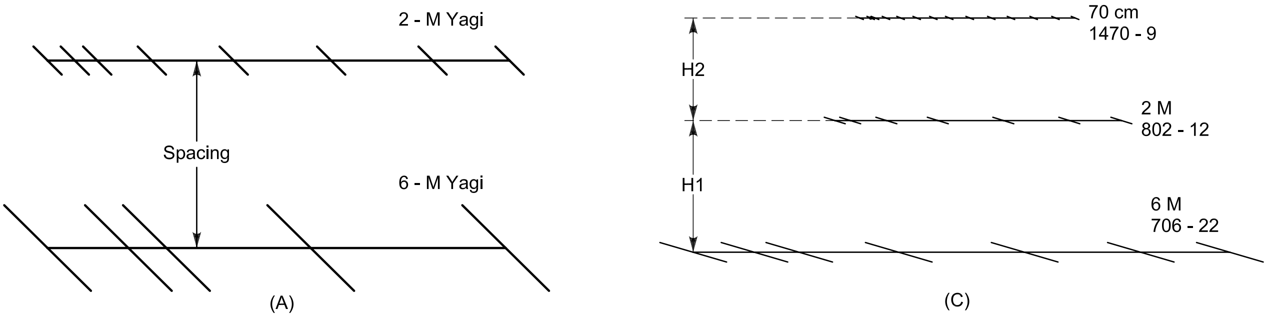

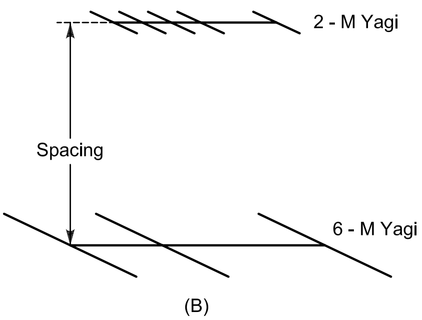

Fig 8—In stacking Yagi arrays one above the other, the minimum spacing between bays (S) should be about half the boom length of the smaller array. Wider spacing is desirable, in which case it should be $\frac{1}{2}\ \lambda$ or some multiple thereof, at the frequency of the smaller array. At A, stack of 8-element 2-meter Yagi on a 12-foot boom over a 5-element 6-meter Yagi, also on a 12-foot boom. At B, 5-element 2-meter beam on a 6-foot boom over a 3-element 6-meter beam on a 4-foot boom. At C, a 14-element 70-cm beam on a 9-foot boom, mounted over a 8-element 2-meter beam on a 12-foot boom and a 7-element 6-meter beam on a 22-foot boom.

ity of the lower-frequency Yagi. Modern computer modeling programs reveal that while the feed-point SWR can indeed be affected, by far the greatest degradation is in the forward gain and rearward pattern of the higher-frequency Yagi when the booms are closely spaced. In fact, the SWR curve is usually not affected enough to make it a good diagnostic indicator of interaction between the two Yagis.

Fig 9 shows an overlay of the SWR curves across the 2-meter band for four configurations: an 8-element 2-meter Yagi by itself, and then over a 5-element 6-meter Yagi with spacings between the booms of 1, 2, 4 and 6 feet. The SWR curves are similar—it would be difficult to see any difference between these configurations using typical amateur SWR indicators for anything but the very closest (1-foot) spacing. For example, the SWR curve for the 2-foot spacing case is virtually indistinguishable from that of the Yagi by itself, while the forward gain has dropped more than 0.6 dB because of interactions with the 6-meter Yagi below it.

Gain and Pattern Degradation Due to Stacking

Fig 10 shows four overlaid rectangular plots of the azimuth response from 0° to 180° for the 8-element 2-meter Yagi described above, spaced 1, 2, 4 and 6 feet over a 5-element 6-meter beam. The rectangular presentation gives more detail than a polar plot. The most closely spaced configuration (with 1-foot spacing between the booms) shows the largest degradation in the forward gain, a drop of 1.7 dB. The worst-case front-to-rear ratio for the 6-foot spacing is 29.0 dB, while it is 36.4 dB for the 1-foot spacing—actually better than the F/R for the 8-element 2-meter Yagi by itself. Performance change due to the nearby presence of other Yagis can be enormously complicated (and sometimes is non-intuitive as well).

What happens when a different kind of 6-meter Yagi is mounted below the 8-element 2-meter Yagi? **Fig 11** compares the change in forward gain and the worst-case F/R performance as a function of spacing between the booms for two varieties of 6-meter Yagis: the 5-element design on a 12-foot boom and a 7-element Yagi on a

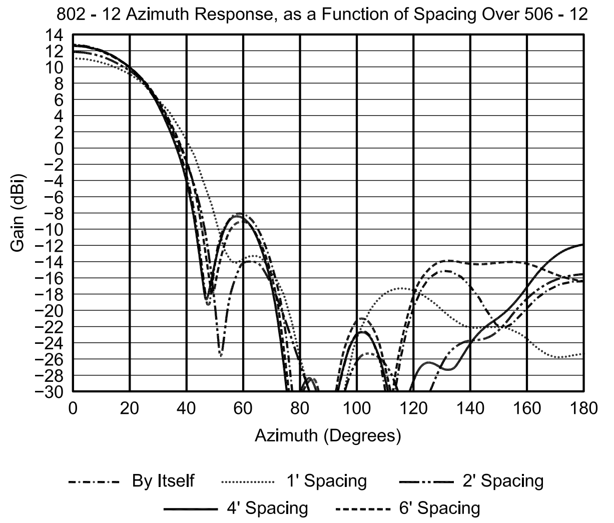

Fig 10—Plots of the 8-element 2-meter Yagi's azimuth response from 0° to 180° for spacing distances from 1 to 6 feet. The sidelobe at about 60° varies about 6 dB over the range of boom spacings, while the shape of worst-case F/R curve varies considerably due to interactions with the lower 6-meter beam. The gain for the 1-foot spacing is degraded by more than 3 dB compared to the 2-meter antenna by itself.

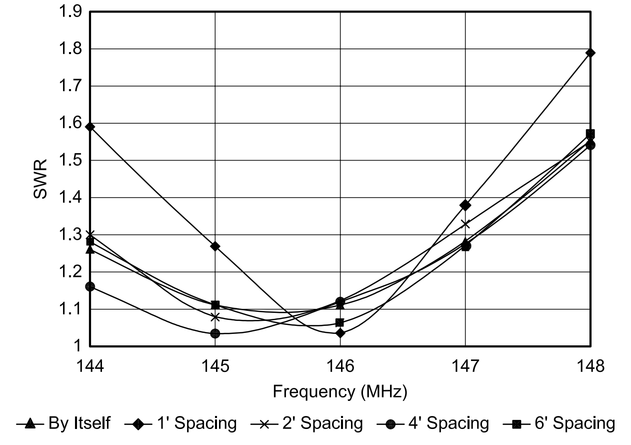

Fig 9—SWR curves for different boom spacing between 8-element 2-meter Yagi on 12-foot boom, over a 5-element 6-meter Yagi on a 12-foot boom. For spacings greater than 1 foot between the booms, differences between the SWR curves are difficult to discern.

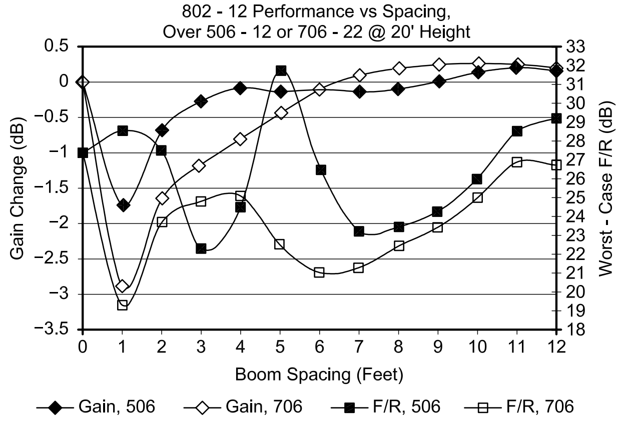

Fig 11—Plot of 8-element 2-meter Yagi's gain and worst-case F/R as a function of distance over two types of 6-meter beams, one on a 12-foot boom and the other on a 22-foot boom. Beyond a spacing of about 5 feet the performance is degraded a minimal amount.

22-foot boom. The spacing of “0 feet” represents the 8-element 2-meter Yagi when it is used alone, with no other antenna nearby. This sets the reference expectations for gain and F/R.

The most severe degradation occurs for the 1-foot spacing, as you might imagine, for both the 12 and 22-foot boom lengths. Over the 5-element 6-meter Yagi, the 2-meter gain doesn’t recover to the reference level of the 8-element 2-meter beam by itself until the spacing is greater than 9 feet. However, the gain is within 0.25 dB of the reference level for spacings of 3 feet or more. Interestingly, the F/R is higher than that of the 2-meter antenna by itself for the 1, 2 and 5-foot spacings and for spacings greater than 11 feet. The 2-meter F/R in the presence of the 12-foot 5-element 6-meter Yagi remains above 20 dB for spacings beyond 1 feet.

Overall, the 2-meter beam performs reasonably well for spacings of 3 feet or more over the 5-element 6-meter Yagi. Put another way, the 2-meter beam’s performance is degraded only slightly for boom spacings greater than 3 feet. A spacing of 3 feet is less than the old rule of thumb that the minimum spacing between booms be greater than one-half the boomlength of the higher-frequency Yagi, which in this case is 6 feet long.

For the 7-element 6-meter Yagi, the 2-meter gain recovers to the reference level for spacings beyond 7 feet, but the F/R is degraded below the reference level for all spacings shown in Fig 11. If we use a gain reduction criterion of less than 0.25 dB and a 20-dB F/R level as the minimum acceptable level, then the spacing must be 5 feet or more over the larger 6-meter Yagi. Again, this is less than the rule of thumb that the minimum spacing between booms be greater than one-half the boomlength of the higher-frequency Yagi.

Now, let’s try a smaller setup of 2- and 6-meter Yagis stacked vertically in a Christmas-tree configuration to see if the rule of thumb for spacing the booms still holds. **Fig 12** shows the performance curves versus boom spacing for a 5-element 2-meter Yagi on a 4-foot boom stacked over a 3-element 6-meter Yagi on a 6-foot boom. Again, the 1-foot spacing produces a substantial gain reduction of about 1.3 dB compared to the reference gain when the 2-meter Yagi is used by itself. Beyond a boom spacing of 3 feet the 2-meter gain drops less than 0.25 dB from the reference level of the 2-meter Yagi by itself and the F/R remains above about 20 dB. In this example, the simple rule of thumb that the minimum spacing between booms be greater than half the boom length (half of 4 feet) of the higher-frequency Yagi does not hold up. However, the same minimum spacing of 3 feet we found for the larger 2-meter Yagi remains true. Three feet spacing is almost 0.5 λ between the booms at the higher frequency.

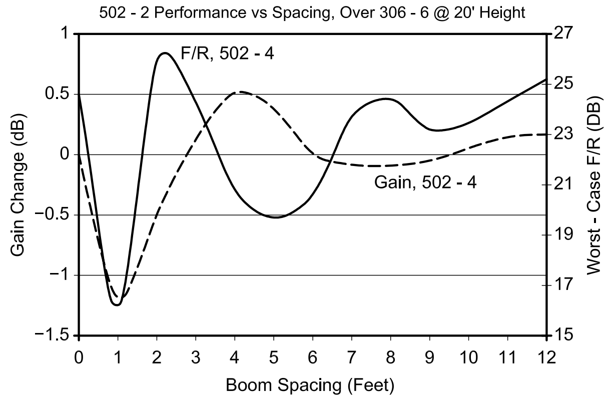

Fig 12— Plot of gain and worst-case F/R of a 5-element 2-meter Yagi on a 4-foot boom as a function of distance over a 3-element 6-meter beam on a 6-foot boom. Beyond a spacing of about 3 feet the performance is degraded a minimal amount.

Adding a 70-cm Yagi to the Christmas Tree

Let’s get more ambitious and set up a larger VHF/UHF Christmas tree, with a 14-element 70-cm Yagi on a 9-foot boom at the top, mounted 5 feet over an 8-element 2-meter Yagi on a 12-foot boom. At the bottom of the stack (at the top of the tower) is either the 5-element 6-meter beam on a 12-foot boom, or a 7-element 6-meter beam on a 22-foot boom. See Fig 8C. As before, we will vary the spacing between the 70-cm Yagi and the 2-meter Yagi below it to assess the interactions that degrade the 70-cm performance.

Fig 13 compares the change in gain and F/R curves as a function of boom spacings between the 70-cm and 2-meter Yagis for the two different 6-meter Yagis (with a fixed distance of 5 feet between the 2-meter and 6-meter Yagis). In this example, the 70-cm Yagi was designed to be an intrinsic 50-Ω feed, where the F/R has been compromised to some extent. Still, the F/R is greater than 20 dB when the 70-cm Yagi is used by itself.

For spacings greater than 4 feet between the 70-cm and 2-meter booms, the 70-cm gain is equal to or even slightly greater than that of the 70-cm antenna by itself. The increase of gain indicates that the elevation pattern of the 70-cm antenna is slightly compressed by the presence of the other Yagis below it. The F/R stays above at 19.5 dB for spacings greater than or equal to 4 feet. This falls just below our desired lower limit of 20 dB, but it is highly doubtful that anyone would notice this 0.5-dB drop in actual operation. A spacing of 4 feet between booms falls under the rule of thumb that the minimum spacing be at least half the boomlength of the higher-frequency Yagi, which in this case is 9 feet.

What should be obvious in this discussion is that you should model the exact configuration you plan to build to avoid unnecessary performance degradation.

Stacking Same-Frequency Yagis

This subject has been examined in some detail in Chapter 11, HF Yagi Arrays. The same basic principles

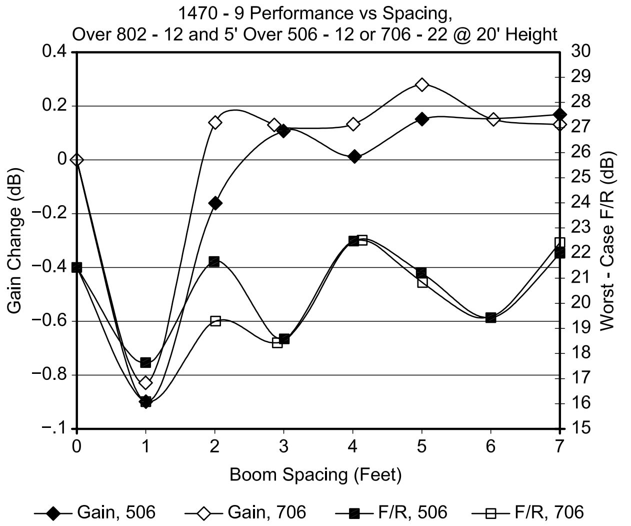

Fig 13—Performance of a 14-element 70-cm Yagi on a 9-foot boom, mounted a variable distance over an 8-element 2-meter Yagi on a 12-foot boom, which is mounted 5 feet above either a 5-element 6-meter Yagi on a 12-foot boom or a 7-element 6-meter Yagi on a 22-foot boom. Beyond a spacing of about 4 feet, the performance of the 70-cm beam is degraded a minimal amount.

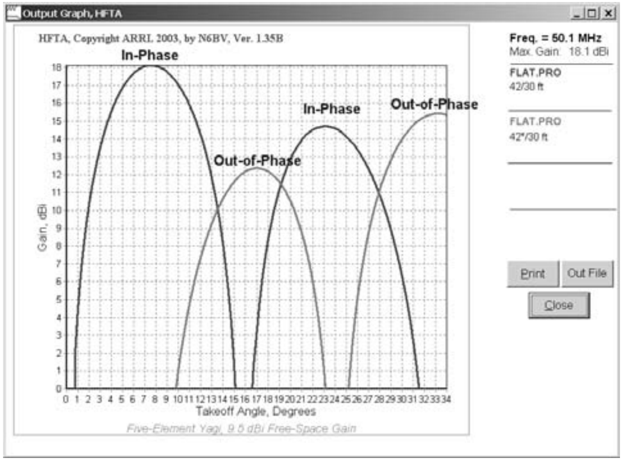

Freq. = 50.1 MHz
Max. Gain: 18.1 dBi

FLAT.PRO
42/30 ft

FLAT.PRO
42"/30 ft

Fig 14—*HFTA* comparison plots of the elevation responses for two 5-element 6-meter Yagis mounted at 42 and 30 feet above flat ground, when they are fed in-phase and out-of-phase. By switching the phasing (adding a half-wavelength of coax to one of the antennas), the elevation angle can be controlled to enhance performance when a Sporadic-E cloud is nearly overhead.

hold at VHF and UHF as they do on HF. That is, the gain increases gradually with increasing spacing between the booms, and then falls off gradually past a certain spacing distance.

At HF, Chapter 11 emphasizes that you should avoid nulls in the antenna's elevation response—so that you can cover all the angles needed for geographic areas of interest. At VHF/UHF, propagation is usually at low elevation angles for most propagation modes, and signals are often extremely weak. Thus, achieving maximum gain is the most common design objective for a VHF/UHF stack. Of secondary importance is the cleanliness of the beam pattern, to discriminate against interference and noise sources.

Six-meter Sporadic-E can sometimes occur at high elevation angles, especially if the E_s cloud is overhead, or nearly overhead. Since Sporadic-E is exactly that, *sporadic*, it's not a good design practice to try to cover a wide range of elevation angles, as you must often do at HF to cover large geographic areas. On 6 meters, you can change to high-angle coverage when necessary. For example, you might switch to a separate Yagi mounted at a low height, or you might provide means to feed stacked antennas out-of-phase. **Fig 14** shows an *HFTA* (HF Terrain Assessment) plot of two 5-element 6-meter Yagis, fed either in-phase or out-of-phase to cover a much wider range of elevation angles than the in-phase stack alone.

Fig 15A shows the change in gain for four 2-meter stacked designs, as a function of the spacing in wavelengths between the booms. The 3-element Yagi is mounted on a 2-foot boom (occupying 0.28 λ of that boom). The 5-element Yagi is on a 4-foot boom (0.51 λ of the boom), while the 8-element Yagi is on a 12-foot boom (1.72 λ of boom). The biggest antenna in the group has 16 elements, on a 27-foot boom (4.0 λ of boom). This range of boom lengths pretty much covers the practical range of antennas used by hams.

The stack of two 3-element Yagis peaks at 3.2 dB of additional gain over a single Yagi for 0.75 λ spacing between the booms. Further increases in spacing see the gain change gradually drop off. Fig 15B shows the worst-case F/R of the four stacks, again as a function of boom length. The F/R of a single 3-element Yagi is just over 24 dB, but in the presence of the second 3-element Yagi in the stack, the F/R of the pair oscillates between 15 to 26 dB, finally remaining consistently over the desired 20-dB level for spacings greater than about 1.7 λ, where the gain has fallen about 0.6 dB from the peak possible gain. A boom spacing of 1.7 λ at 146 MHz is 11.5 feet. Thus you must compromise in choosing the boom spacing between achieving maximum gain and the best pattern.

The increase in gain of the stack of two 5-element Yagis peaks at a spacing of about 1 λ (6.7 feet), where the F/R is an excellent 25 dB. Having more elements on a particular length of boom aids in holding a more consistent F/R in the presence of the second antenna.

The gain increase for the bigger stack of 8-element Yagis peaks at a spacing of about 1.5 λ (10.1 feet), where the F/R is more than 27 dB. The 16-element Yagi's gain

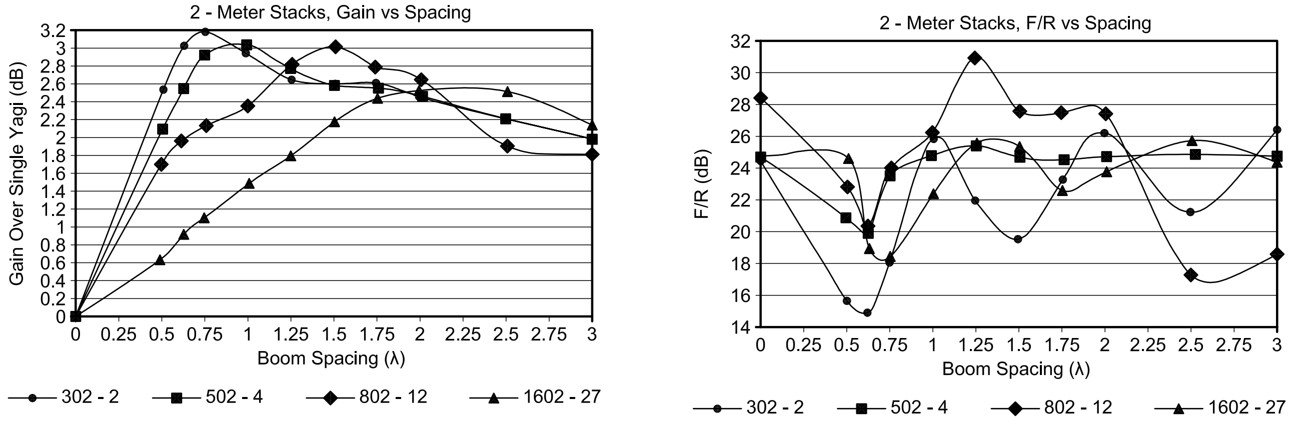

Fig 15—Performance of two different 2-meter Yagis (5-elements on 4-foot boom and 8-elements on 12-foot boom) fed in-phase, as a function of spacing between the booms. Note that the distance is measured in wavelengths.

increase is 2.6 dB for a spacing of about 2.25 λ (15.2 feet), where the F/R remains close to 25 dB. The stacking distance of 15.2 feet for an antenna with a 27-foot long boom may be a real challenge physically, requiring a very sturdy rotating mast to withstand wind pressures without bending.

These examples show that the exact spacing between booms is not overly critical, since the gain varies relatively slowly around the peak. Fig 15A shows that the boom spacing needed to achieve peak gain from a stack increases when higher-gain (longer-boom) individual antennas are used in that stack. It also shows that the increase in maximum gain from stacking decreases for long-boom antennas. Fig 15B shows that beyond boom spacings of about 1 λ, the F/R pattern holds well for Yagi designs with booms longer than about 0.5 λ, which is about 4 feet at 146 MHz.

The plots in Fig 15 are representative of typical modern Yagis. You could simply implement these designs as is, and you'll achieve good results. However, we recommend that you model any specific stack you design, just to make sure. Since the boom spacings are displayed in terms of wavelength, you can extend the results for 2 meters to other bands, provided that you use properly scaled Yagi designs to the other bands too.

You can even tweak the element dimensions and spacings of each Yagi used in a stack to optimize the rearward pattern for a particular stacking distance. This strategy can work out well at VHF/UHF, where stacks are often configured for best gain (and pattern) and are "hard-wired" with fixed lengths of feed lines permanently junctioned together.

This is in contrast to the situation at HF (and even on 6 meters). The HF operator usually wants flexibility to select individual Yagis (or combinations of Yagis) from the stack, to match the array's takeoff angle with ionospheric propagation conditions. See Chapter 11, HF Yagi Arrays. The designer of a flexible HF stack thus usually doesn't try to redo the element lengths and spacings of the Yagis to optimize a particular stack.

Stacking Stacks of Different-Frequency Yagis

The investment in a tower is usually substantial, and most hams want to put as many antennas as possible on a tower, provided that interaction between the antennas can be held to a reasonable level. Really ambitious weak-signal VHF/UHF enthusiasts may want "stacked stacks"—sets of stacked Yagis that cover different bands. For example, a VHF contester might want a stack of two 8-element 2-meter Yagis mounted on the same rotating mast as a stack of two 5-element 6-meter Yagis. Let's assume that the boom length of the 8-element 2-meter Yagis is 12 feet (1.78 λ). We'll assume a boom length of 12 feet (0.61 λ) for the 5-element 6-meter Yagis.

From Fig 15, we find the stacking distance between the 8-element 2-meter beams for peak gain and good pattern is 1.5 λ, or 10 feet, but adequate performance can be had for a boom spacing of 0.75 λ, which is 5 feet on 2 meters.

The boom spacing for two 5-element 6-meter beams is 1 λ for peak stacking gain, but a compromise of 0.625 λ (12 feet) still yields an acceptable gain increase of 2 dB over a single Yagi. The overall height of the rotating mast sticking out of the top of the tower is thus set by the 0.625 λ stacking distance on 6 meters, at 12 feet. In-between the 6-meter Yagis at the bottom and top of the rotating mast we will mount the 2-meter Yagi stack. With only 12 feet available on the mast, the spacing for symmetric placement of the two 2-meter Yagis in-between the 6-meter Yagis dictates a distance of only 4 feet between the 2-meter beams. This is less than optimal.

The performance of the 2-meter stack in this "stack within a stack" is affected by the close spacing, but the interactions are not disastrous. The stacking gain is 1.62 dB more than the gain for a single 8-element 2-meter Yagi and the F/R remains above 20 dB across the 2-meter band.

On 6 meters, the stacking gain for two 5-element 6-meter Yagis spaced 12 feet apart is 2.2 dB more than the gain of a single Yagi, while the F/R pattern remains about 20 dB over the weak-signal portion of the 6-meter band. As described in Chapter 11, HF Yagi Arrays, stacking gives more advantages than merely a gain increase, and 6-meter propagation does require coverage of a range of elevation angles because much of the time ionospheric modes are involved.

Increasing the length of the rotating mast to 18 feet sticking out of the top of the tower will increase performance, particularly on 2 meters. The stacking gain on 6 meters will increase to 2.3 dB while the F/R decreases to 18.5 dB, modest changes both. The 18-foot mast allows the 2-meter Yagis to be spaced 6 feet from each other and 6 feet away from both top and bottom 6-meter antennas. The stacking gain goes to 2.14 dB and the F/R approaches 27 dB in the weak-signal portion of the 2-meter band.

Whether the modest increase in stacking gain is worth the cost and mechanical complexity of stacking two 2-meter Yagis in-between a stack of 6-meter Yagis is a choice left to the operator. Certainly the cost and weight of a rotating mast that is 20 feet long (18 feet out of the top of the tower and 2 feet down inside the tower), a mast that must be sturdy enough to support the antennas in high winds without bending, should give pause to even the most enthusiastic 6-meter weak-signal operator.

QUADS FOR VHF

The quad antenna can be built with inexpensive materials, yet its performance is comparable to other arrays of its size. Adjustment for resonance and impedance matching can be accomplished readily. Quads can be stacked horizontally and vertically to provide high gain, without sharply limiting frequency response. Construction of quad antennas for VHF use is covered later in this chapter.

Stacking Quads

Quads can be mounted side by side or one above the other, or both, in the same general way as other beam antennas. Sets of driven elements can also be mounted in front of a screen reflector. The recommended spacing between adjacent element sides is $^1/_2$ λ. Phasing and feed methods are similar to those employed with other antennas described in this chapter.

Adding Quad Directors

Parasitic elements ahead of the driven element work in a manner similar to those in a Yagi array. Closed loops can be used for directors by making them 5% shorter than the driven element. Spacings are similar to those for conventional Yagis. In an experimental model the reflector was spaced 0.25 λ and the director 0.15 λ. A square array using four 3element bays worked extremely well.

VHF AND UHF QUAGIS

At higher frequencies, especially 420 MHz and above, Yagi arrays using dipole-driven elements can be difficult to feed and match, unless special care is taken to keep the feed-point impedance relatively high by proper element spacing and tuning. The cubical quad described earlier overcomes the feed problems to some extent. When many parasitic elements are used, however, the loops are not nearly as convenient to assemble and tune as are straight cylindrical ones used in conventional Yagis. The *Quagi*, designed and popularized by Wayne Overbeck, N6NB, is an antenna having a full-wave loop driven element and reflector, and Yagi type straight rod directors. Construction details and examples are given in the projects later in this chapter.

COLLINEAR ANTENNAS

The information given earlier in this chapter pertains mainly to parasitic arrays, but the collinear array is worthy of consideration in VHF/UHF operations. This array tends to be tolerant of construction tolerances, making it easy to build and adjust for VHF applications. The use of many collinear driven elements was once popular in very large phased arrays, such as those required in moonbounce (EME) communications, but the advent of computer-optimized Yagis has changed this.

Large Collinear Arrays

Bidirectional curtain arrays of four, six, and eight half waves in phase are shown in **Fig 16**. Usually reflector elements are added, normally at about 0.2 λ behind each driven element, for more gain and a unidirectional pattern. Such parasitic elements are omitted from the sketch in the interest of clarity.

The feed-point impedance of two half waves in phase is high, typically 1000 Ω or more. When they are combined in parallel and parasitic elements are added, the feed impedance is low enough for direct connection to open wire line or twin-lead, connected at the points indicated by black dots. With coaxial line and a balun, it is suggested that the universal stub match, Fig 4A, be used at the feed point. All elements should be mounted at their electrical centers, as indicated by open circles in Fig 16. The framework can be metal or insulating material. The metal supporting structure is entirely behind the plane of the reflector elements. Sheet-metal clamps can be cut from scraps of aluminum for this kind of assembly. Collinear elements of this type should be mounted at their centers (where the RF voltage is zero), rather than at their ends, where the voltage is high and insulation losses and detuning can be harmful.

Collinear arrays of 32, 48, 64 and even 128 elements can give outstanding performance. Any collinear array should be fed at the center of the system, to ensure balanced current distribution. This is very important in large arrays, where sets of six or eight driven elements are treated as "sub arrays," and are fed through a balanced harness. The sections of the harness are resonant lengths, usually of open wire line. The 48-element collinear array for 432 MHz in **Fig 17** illustrates this principle.

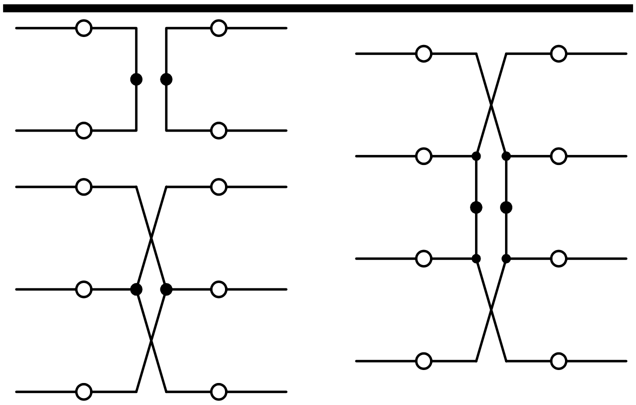

Fig 16—Element arrangements for 8-, 12- and 16-element collinear arrays. Elements are ½ λ long and spaced ½ λ. Parasitic reflectors, omitted here for clarity, are 5% longer and 0.2 λ behind the driven elements. Feed points are indicated by black dots. Open circles show recommended support points. The elements can run through wood or metal booms, without insulation, if supported at their centers in this way. Insulators at the element ends (points of high RF voltage) detune and unbalance the system.

A reflecting plane, which may be sheet metal, wire mesh, or even closely spaced elements of tubing or wire, can be used in place of parasitic reflectors. To be effective, the plane reflector must extend on all sides to at least $^1/_4$ λ beyond the area occupied by the driven elements. The plane reflector provides high F/B ratio, a clean pattern, and somewhat more gain than parasitic elements, but large physical size limits it to use above 420 MHz. An interesting space-saving possibility lies in using a single plane reflector with elements for two different bands mounted on opposite sides. Reflector spacing from the driven element is not critical. About 0.2 λ is common.

THE CORNER REFLECTOR

When a single driven element is used, the reflector screen may be bent to form an angle, giving an improvement in the radiation pattern and gain. At 222 and 420 MHz its size assumes practical proportions, and at 902 MHz and higher, practical reflectors can approach ideal dimensions (very large in terms of wavelengths), resulting in more gain and sharper patterns. The corner reflector can be used at 144 MHz, though usually at much less than optimum size. For a given aperture, the corner reflector does not equal a parabola in gain, but it is simple to construct, broadbanded, and offers gains from about 9 to 14 dBi, depending on the angle and size. This section was written by Paul M. Wilson, W4HHK.

The corner angle can be 90, 60 or 45°, but the side length must be increased as the angle is narrowed. For a 90° corner, the driven element spacing can be anything from 0.25 to 0.7 λ, 0.35 to 0.75 λ for 60°, and 0.5 to

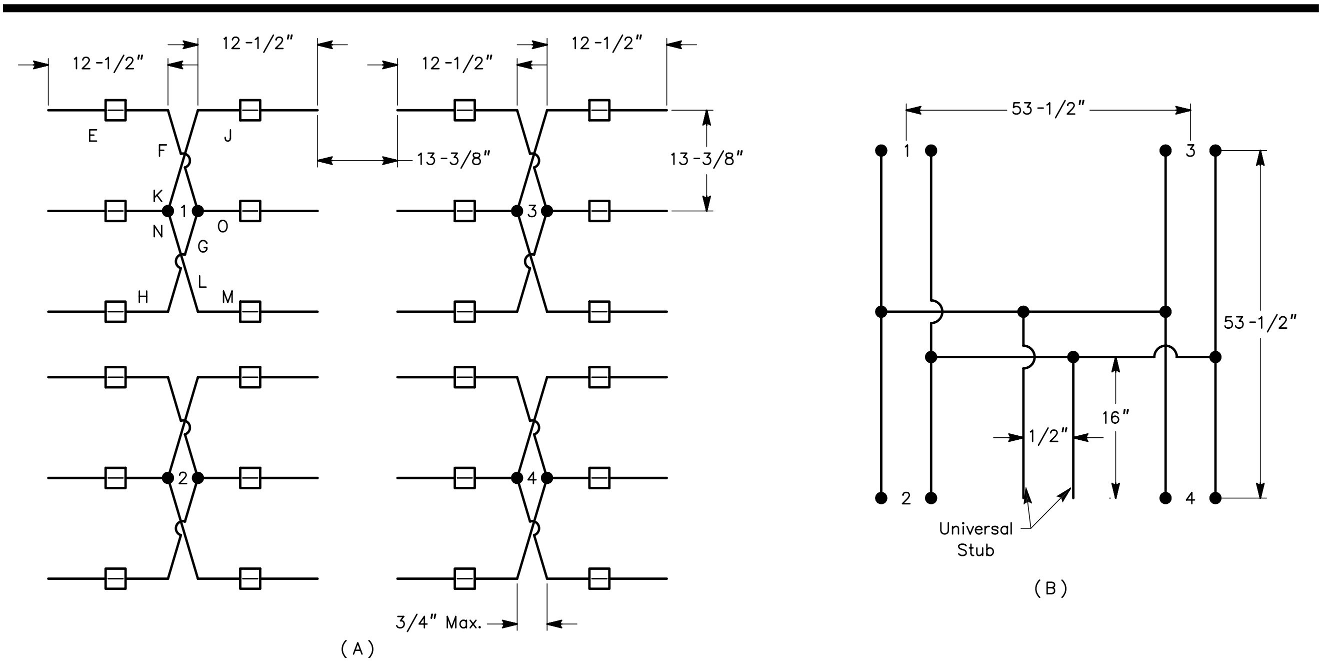

Fig 17—Large collinear arrays should be fed as sets of no more than eight driven elements each, interconnected by phasing lines. This 48-element array for 432 MHz (A) is treated as if it were four 12-element collinear antennas. Reflector elements are omitted for clarity. The phasing harness is shown at B. Squares represent insulators.

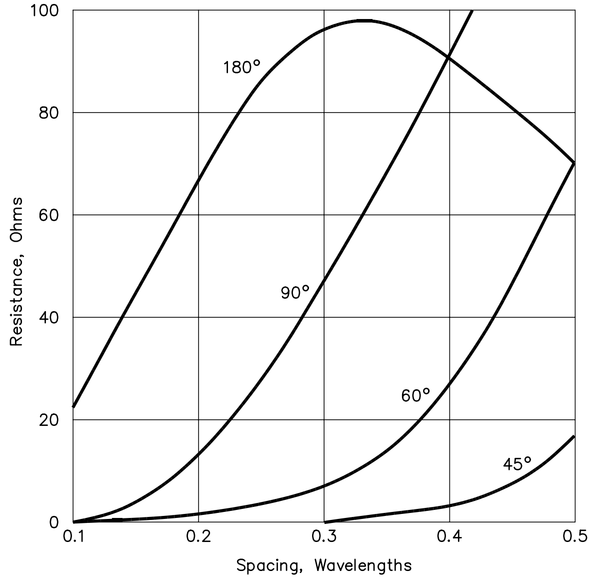

Fig 18—Radiation resistance of the driven element in a corner reflector array for corner angles of 180° (flat sheet), 90°, 60° and 45° as a function of spacing D, as shown in Fig 19.

0.8 λ for 45°. In each case the gain variation over the range of spacings given is about 1.5 dB. Because the spacing is not very critical to gain, it may be varied for impedance-matching purposes. Closer spacings yield lower feed-point impedances, but a folded dipole radiator could be used to raise this to a more convenient level.

Radiation resistance is shown as a function of spacing in **Fig 18**. The maximum gain obtained with minimum spacing is the primary mode (the one generally used at 144, 222 and 432 MHz to maintain reasonable side lengths). A 90° corner, for example, should have a minimum side length (S, **Fig 19**) equal to twice the dipole spacing, or 1 λ long for 0.5-λ spacing. A side length greater than 2 λ is ideal. Gain with a 60° or 90° corner reflector with 1-λ sides is about 10 dB. A 60° corner with 2-λ sides has about 13 dBi gain, and a 45° corner with 3-λ sides has about 14 dBi gain.

Reflector length (L, Fig 19) should be a minimum of 0.6 λ. Less than that spacing causes radiation to increase to the sides and rear, and decreases gain.

Spacing between reflector rods (G, Fig 19) should not exceed 0.06 λ for best results. A spacing of 0.06 λ results in a rear lobe that is about 6% of the forward lobe

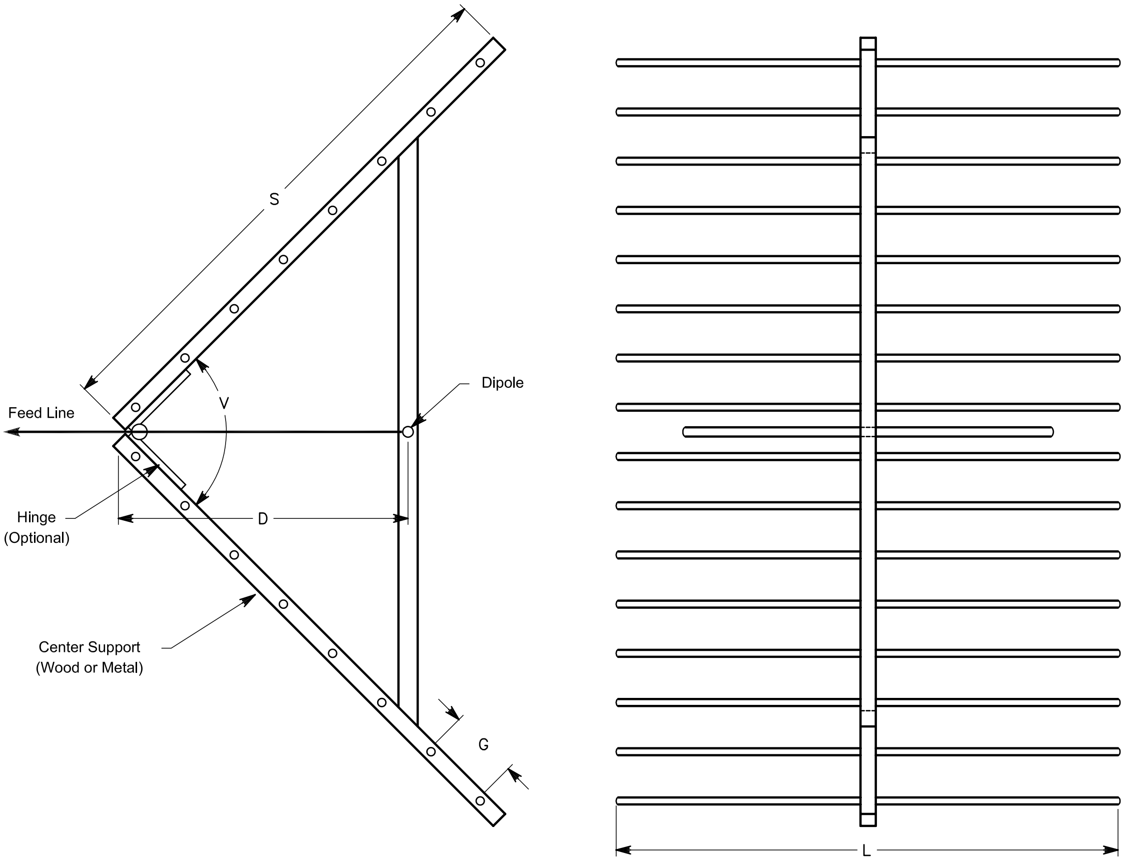

Fig 19—Construction of a corner reflector array. The frame can be wood or metal. Reflector elements are stiff wire or tubing. Dimensions for several bands are given in Table 2. Reflector element spacing, G, is the maximum that should be used for the frequency; closer spacings are optional. The hinge permits folding for portable use.

(down 12 dB). A small mesh screen or solid sheet is preferable at the higher frequencies to obtain maximum efficiency and highest F/B ratio, and to simplify construction. A spacing of 0.06 λ at 1296 MHz, for example, requires mounting reflector rods about every $^{1}/_{2}$ inch along the sides. Rods or spines may be used to reduce wind loading. The support used for mounting the reflector rods may be of insulating or conductive material. Rods or mesh weave should be parallel to the radiator.

A suggested arrangement for a corner reflector is shown in Fig 19. The frame may be made of wood or metal, with a hinge at the corner to facilitate portable work or assembly atop a tower. A hinged corner is also useful in experimenting with different angles. **Table 2** gives the principal dimensions for corner reflector arrays for 144 to 2300 MHz. The arrays for 144, 222 and 420 MHz have side lengths of twice to four times the driven element spacing. The 915 MHz corner reflectors use side lengths of three times the element spacing, 1296 MHz corners use side lengths of four times the spacing, and 2304 MHz corners employ side lengths of six times the spacing. Reflector lengths of 2, 3, and 4 wavelengths are used on the 915, 1296 and 2304 MHz reflectors, respectively. A 4 × 6 λ reflector closely approximates a sheet of infinite dimensions.

A corner reflector may be used for several bands, or for UHF television reception, as well as amateur UHF work. For operation on more than one frequency, side length and reflector length should be selected for the lowest frequency, and reflector spacing for the highest frequency. The type of driven element plays a part in determining bandwidth, as does the spacing to the corner. A fat cylindrical element (small λ/dia ratio) or triangular dipole (bow tie) gives more bandwidth than a thin driven element. Wider spacings between driven element and corner give greater bandwidths. A small increase in gain can be obtained for any corner reflector by mounting collinear elements in a reflector of sufficient size, but

Table 2

Dimensions of Corner Reflector Arrays for VHF and UHF

Freq, MHz	*Side Length S, in.*	*Dipole to Vertex D, in.*	*Reflector Length L, in.*	*Reflector Spacing G, in.*	*Corner Angle, Vo*	*Radiation Resistance, Ω*
144*	65	27½	48	7¾	90	70
144	80	40	48	4	90	150
222*	42	18	30	5	90	70
222	52	25	30	3	90	150
222	100	25	30	Screen	60	70
420	27	8¾	16¼	$2^{5}/_{8}$	90	70
420	54	13½	16¼	Screen	60	70
915	20	6½	25¾	0.65	90	70
915	51	16¾	25¾	Screen	60	65
915	78	25¾	25¾	Screen	45	70
1296	18	4½	27½	½	90	70
1296	48	11¾	27½	Screen	60	65
1296	72	18¼	27½	Screen	45	70
2304	15½	2½	20½	¼	90	70
2304	40	6¾	20½	Screen	60	65
2304	61	10¼	20½	Screen	45	70

*Side length and number of reflector elements somewhat below optimum—slight reduction in gain.

Notes:

915 MHz
Wavelength is 12.9 in.
Side length S is 3 × D, dipole to vertex distance
Reflector length L is 2.0 λ
Reflector spacing G is 0.05 λ

1296 MHz
Wavelength is 9.11 in.
Side length S is 4 × D, dipole to vertex distance
Reflector length L is 3.0 λ
Reflector spacing G is 0.05 λ

2304 MHz
Wavelength is 5.12 in.
Side length S is 6 × D, dipole to vertex distance
Reflector length L is 4.0 λ
Reflector spacing G is 0.05 λ

the simple feed of a dipole is lost if more than two elements are used.

A dipole radiator is usually employed with a corner reflector. This requires a balun between the coaxial line and the balanced feed-point impedance of the antenna. Baluns are easily constructed of coaxial line on the lower VHF bands, but become more difficult at the higher frequencies. This problem may be overcome by using a ground-plane corner reflector, which can be used for vertical polarization. A ground-plane corner with monopole driven element is shown in **Fig 20**. The corner reflector and a $^1/_4\ \lambda$ radiator are mounted on the ground plane, permitting direct connection to a coaxial line if the proper spacing is used. The effective aperture is reduced, but at the higher frequencies, second- or third-mode radiator spacing and larger reflectors can be employed to obtain more gain and offset the loss in effective aperture. A J antenna could be used to maintain the aperture area and provide a match to a coaxial line.

For vertical polarization work, four 90° corner reflectors built back-to-back (with common reflectors) could be used for scanning 360° of horizon with modest gain. Feed-line switching could be used to select the desired sector.

TROUGH REFLECTORS

To reduce the overall dimensions of a large corner reflector the vertex can be cut off and replaced with a plane reflector. Such an arrangement is known as a *trough reflector*. See **Fig 21**. Performance similar to that of the large corner reflector can thereby be had, provided that the dimensions of S and T as shown in Fig 21 do not exceed the limits indicated in the figure. This antenna provides performance very similar to the corner reflector, and presents fewer mechanical problems because the plane center portion is relatively easy to mount on the mast. The sides are considerably shorter, as well.

The gain of both corner reflectors and trough reflectors may be increased by stacking two or more and arranging them to radiate in phase, or alternatively by

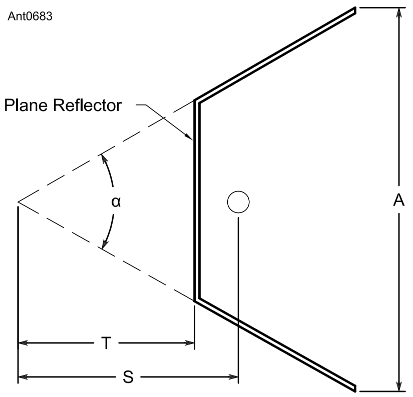

Angle *α*	*Value of S for maximum gain*	*Gain*	*Value of T*
90°	1.5 λ	13.5 dB	1λ - 1.25 λ
60°	1.75 λ	15 dB	1.0 λ
45°	2.0 λ	16 dB	1.9 λ

Fig 21—The trough reflector. This is a useful modification of the corner reflector. The vertex has been cut off and replaced by a simple plane section. The tabulated data shows the gain obtainable for greater values of S than those covered in Table 2, assuming that the reflector is of adequate size.

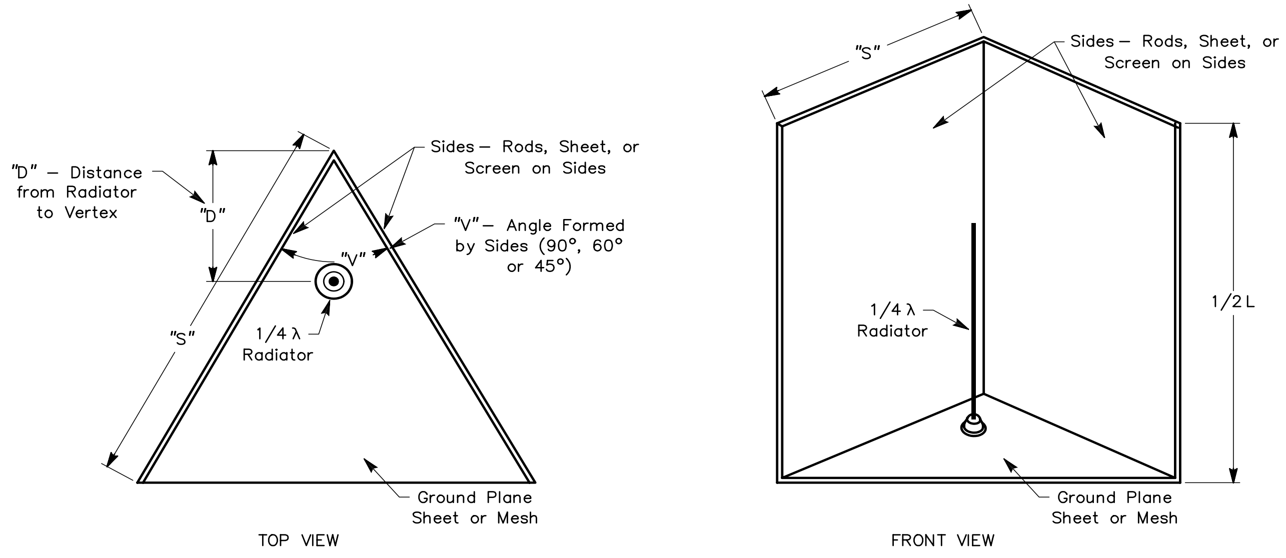
"D" — Distance from Radiator to Vertex
"D"
Sides — Rods, Sheet, or Screen on Sides
"V" — Angle Formed by Sides (90°, 60° or 45°)
"V"
"S"
1/4 λ Radiator
Ground Plane Sheet or Mesh
TOP VIEW
"S"
Sides — Rods, Sheet, or Screen on Sides
1/4 λ Radiator
1/2 L
Ground Plane Sheet or Mesh
FRONT VIEW

Fig 20—A ground-plane corner reflector antenna for vertical polarization, such as FM communications or packet radio. The dimension ½ λ in the front view refers to data in Table 2.

adding further collinear dipoles (fed in phase) within a wider reflector. Not more than two or three radiating units should be used, because the great virtue of the simple feeder arrangement would then be lost.

HORN ANTENNAS FOR THE MICROWAVE BANDS

Horn antennas were briefly introduced in the section on coupling energy into and out of waveguides. For amateur purposes, horns begin to show usable gain with practical dimensions in the 902 MHz band.

It isn't necessary to feed a horn with waveguide. If only two sides of a pyramidal horn are constructed, the antenna may be fed at the apex with a two-conductor transmission line. The impedance of this arrangement is on the order of 300 to 400 Ω. A 60° two-sided pyramidal horn with 18 inch sides is shown in **Fig 22**. This antenna has a theoretical gain of 15 dBi at 1296 MHz, although the feed system detailed in **Fig 23** probably degrades this value somewhat. A $^1/_4$ λ, 150-Ω matching section made from two parallel lengths of twin-lead connects to a bazooka balun made from RG-58 cable and a brass tube. This matching system was assembled strictly for the purpose of demonstrating the two-sided horn in a 50-Ω system. In a practical installation the horn would be fed with open-wire line and matched to 50 Ω at the station equipment.

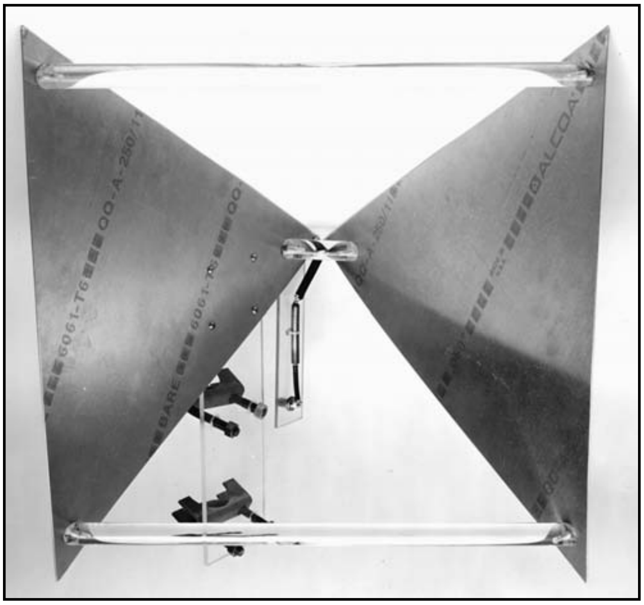

Fig 22—An experimental two-sided pyramidal horn constructed in the ARRL laboratory. A pair of muffler clamps allows mounting the antenna on a mast. This model has sheet-aluminum sides, although window screen would work as well. Temporary elements could be made from cardboard covered with aluminum foil. The horizontal spreaders are Plexiglas rod. Oriented as shown here, the antenna radiates horizontally polarized waves.

PARABOLIC ANTENNAS

When an antenna is located at the focus of a parabolic reflector (dish), it is possible to obtain considerable gain. Furthermore, the beamwidth of the radiated energy will be very narrow, provided all the energy from the driven element is directed toward the reflector. This section was written by Paul M. Wilson, W4HHK.

Gain is a function of parabolic reflector diameter, surface accuracy and proper illumination of the reflector by the feed. Gain may be found from

$$G = 10 \log k \left(\frac{\pi D}{\lambda} \right)^2 \qquad \text{(Eq 1)}$$

where

G = gain over an isotropic antenna, dBi (subtract 2.15 dB for gain over a dipole)
k = efficiency factor, usually about 55%
D = dish diameter in feet
λ = wavelength in feet

See **Table 3** for parabolic antenna gain for the bands 420 MHz through 10 GHz and diameters of 2 to 30 feet.

A close approximation of beamwidth may be found from

$$\psi = \frac{70\lambda}{D} \qquad \text{(Eq 2)}$$

where

ψ = beamwidth in degrees at half-power points (3 dB down)
D = dish diameter in feet
λ = wavelength in feet

At 420 MHz and higher, the parabolic dish becomes a practical antenna. A simple, single feed point eliminates phasing harnesses and balun requirements. Gain is dependent on good surface accuracy, which is more difficult to achieve with increasing frequency. Surface

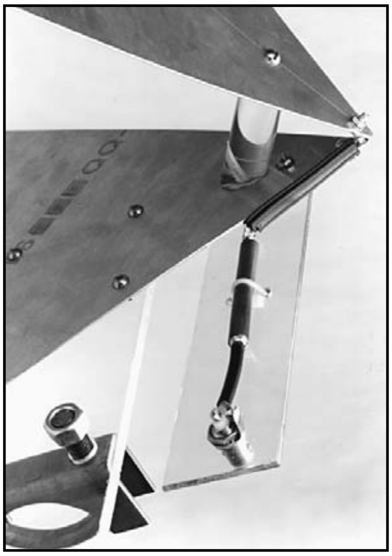

Fig 23—Matching system used to test the horn. Better performance would be realized with open wire line. See text.

Table 3
Gain, Parabolic Antennas*

	Dish Diameter (Feet)						
Frequency	*2*	*4*	*6*	*10*	*15*	*20*	*30*
420 MHz	6.0	12.0	15.5	20.0	23.5	26.0	29.5
902	12.5	18.5	22.0	26.5	30.0	32.5	36.0
1215	15.0	21.0	24.5	29.0	32.5	35.0	38.5
2300	20.5	26.5	30.0	34.5	38.0	40.5	44.0
3300	24.0	30.0	33.5	37.5	41.5	43.5	47.5
5650	28.5	34.5	38.0	42.5	46.0	48.5	52.0
10 GHz	33.5	39.5	43.0	47.5	51.0	53.5	57.0

*Gain over an isotropic antenna (subtract 2.1 dB for gain over a dipole antenna). Reflector efficiency of 55% assumed.

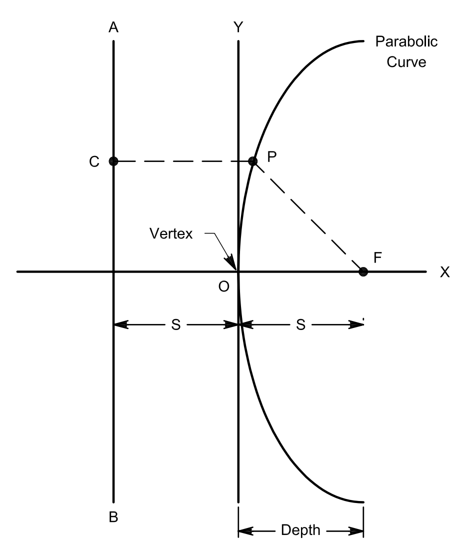

Fig 24—Details of the parabolic curve, $Y^2 = 4SX$. This curve is the locus of points that are equidistant from a fixed point, the focus (F), and a fixed line (AB) that is called the *directrix*. Hence, FP = PC. The focus (F) is located at coordinates S,0.

errors should not exceed $^{1}/_{8}\ \lambda$ in amateur work. At 430 MHz $^{1}/_{8}\ \lambda$ is 3.4 inches, but at 10 GHz it is 0.1476 inch! Mesh can be used for the reflector surface to reduce weight and wind loading, but hole size should be less than $^{1}/_{12}\ \lambda$. At 430 MHz the use of 2-inch hole diameter poultry netting (chicken wire) is acceptable. Fine mesh aluminum screening works well as high as 10 GHz.

A support form may be fashioned to provide the proper parabolic shape by plotting a curve (**Fig 24**) from

$$Y^2 = 4SX$$

as shown in the figure.

Optimum illumination occurs when power at the reflector edge is 10 dB less than that at the center. A circular waveguide feed of correct diameter and length for the frequency and correct beamwidth for the dish focal length to diameter (f/D) ratio provides optimum illumination at 902 MHz and higher. This, however, is impractical at 432 MHz, where a dipole and plane reflector are often used. An f/D ratio between 0.4 and 0.6 is considered ideal for maximum gain and simple feeds.

The focal length of a dish may be found from

$$f = \frac{D^2}{16d} \qquad \text{(Eq 3)}$$

where

f = focal length
D = diameter
d = depth distance from plane at mouth of dish to vertex (see Fig 17)

The units of focal length f are the same as those used to measure the depth and diameter. **Table 4** gives the subtended angle at focus for dish f/D ratios from 0.2 to 1.0. A dish, for example, with a typical f/D of 0.4 requires a 10-dB beamwidth of 130°. A circular waveguide feed with a diameter of approximately 0.7 λ provides nearly optimum illumination, but does not uniformly illuminate the reflector in both the magnetic (TM) and electric (TE) planes. **Fig 25** shows data for plotting radiation patterns from circular guides. The waveguide feed aperture can be modified to change the beamwidth.

One approach used successfully by some experimenters is the use of a disc at a short distance behind the aperture as shown in **Fig 26**. As the distance between the aperture and disc is changed, the TM plane patterns become alternately broader and narrower than with an unmodified aperture. A disc about 2 λ in diameter appears to be as effective as a much larger one. Some experimenters have noted a 1 to 2 dB increase in dish gain with this modified feed. Rectangular waveguide feeds can also be used, but dish illumination is not as uniform as with round guide feeds.

The circular feed can be made of copper, brass, aluminum or even tin in the form of a coffee or juice can, but the latter must be painted on the outside to prevent rust or corrosion. The circular feed must be within a proper size (diameter) range for the frequency being used.

This feed operates in the dominant circular waveguide mode known as the TE_{11} mode. The guide must be large enough to pass the TE_{11} mode with no attenuation, but smaller than the diameter that permits the next higher TM_{01} mode to propagate. To support the desirable TE_{11} mode in circular waveguide, the cutoff frequency, F_C, is given by

$$F_C\,(TE_{11}) = \frac{6917.26}{d\,(\text{inches})} \qquad \text{(Eq 4)}$$

where

F_C = cutoff frequency in MHz for TE_{11} mode
d = waveguide inner diameter

Circular waveguide will support the TM_{01} mode having a cutoff frequency

$$F_C\,(TM_{01}) = \frac{9034.85}{d\,(\text{inches})} \qquad \text{(Eq 5)}$$

The wavelength in a waveguide always exceeds the free-space wavelength and is called guide wavelength, λ_g. It is related to the cutoff frequency and operating frequency by the equation

$$\lambda_g = \frac{11802.85}{\sqrt{f_0^{\,2} - f_C^{\,2}}} \qquad \text{(Eq 6)}$$

where

λ_g = guide wavelength, inches
f_0 = operating frequency, MHz
f_C = TE_{11} waveguide cutoff frequency, MHz

An inside diameter range of about 0.66 to 0.76 l is

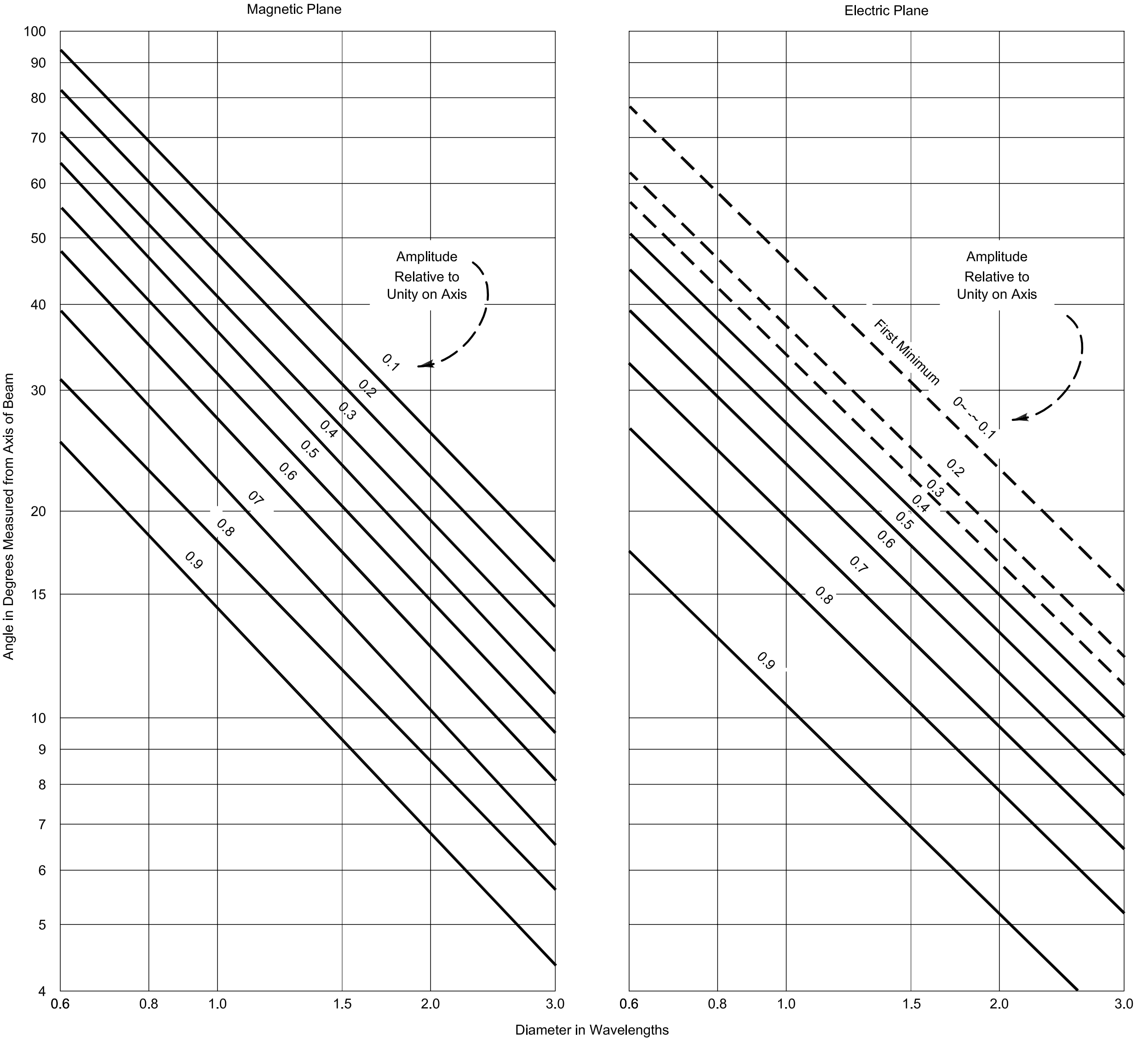

Fig 25—This graph can be used in conjunction with Table 4 for selecting the proper diameter waveguide to illuminate a parabolic reflector.

Table 4
f/D Versus Subtended Angle at Focus of a Parabolic Reflector Antenna

f/D	*Subtended Angle (Deg.)*	*f/D*	*Subtended Angle (Deg.)*
0.20	203	0.65	80
0.25	181	0.70	75
0.30	161	0.75	69
0.35	145	0.80	64
0.40	130	0.85	60
0.45	117	0.90	57
0.50	106	0.95	55
0.55	97	1.00	52
0.60	88		

Taken from graph “f/D vs Subtended Angle at Focus,” page 170 of the 1966 *Microwave Engineers’ Handbook and Buyers Guide*. Graph courtesy of K. S. Kelleher, Aero Geo Astro Corp, Alexandria, Virginia

Table 5
Circular Waveguide Dish Feeds

Freq. (MHz)	*Inside Diameter Circular Waveguide Range (in.)*
915	8.52-9.84
1296	6.02-6.94
2304	3.39-3.91
3400	2.29-2.65
5800	1.34-1.55
10,250	0.76-0.88

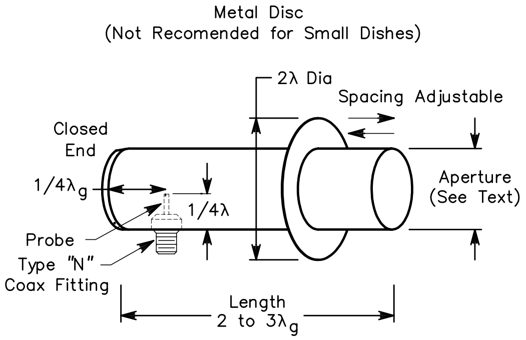

Fig 26—Details of a circular waveguide feed.

suggested. The lower frequency limit (longer dimension) is dictated by proximity to the cutoff frequency. The higher frequency limit (shorter dimension) is dictated by higher order waves. See **Table 5** for recommended inside diameter dimensions for the 902- to 10,000-MHz amateur bands.

The probe that excites the waveguide and makes the transition from coaxial cable to waveguide is $^1/_4$ λ long and spaced from the closed end of the guide by $^1/_4$ guide wavelength. The length of the feed should be two to three guide wavelengths. The latter is preferred if a second probe is to be mounted for polarization change or for polaplexer work where duplex communication (simultaneous transmission and reception) is possible because of the isolation between two properly located and oriented probes. The second probe for polarization switching or polaplexer work should be spaced $^3/_4$ guide wavelength from the closed end and mounted at right angles to the first probe.

The feed aperture is located at the focal point of the dish and aimed at the center of the reflector. The feed mounts should permit adjustment of the aperture either side of the focal point and should present a minimum of blockage to the reflector. Correct distance to the dish center places the focal point about 1 inch inside the feed aperture. The use of a nonmetallic support minimizes blockage. PVC pipe, fiberglass and Plexiglas are commonly used materials. A simple test by placing a material in a microwave oven reveals if it is satisfactory up to 2450 MHz. PVC pipe has tested satisfactorily and appears to work well at 2300 MHz. A simple, clean looking mount for a 4-foot dish with 18 inches focal length, for example, can be made by mounting a length of 4-inch PVC pipe using a PVC flange at the center of the dish. At 2304 MHz the circular feed is approximately 4 inches ID, making a snug fit with the PVC pipe. Precautions should be taken to keep rain and small birds from entering the feed.

Never look into the open end of a waveguide when power is applied, or stand directly in front of a dish while transmitting. Tests and adjustments in these areas should be done while receiving or at extremely low levels of transmitter power (less than 0.1 watt). The US Government has set a limit of 10 mW/cm^2 averaged over a 6-minute period as the safe maximum. Other authorities believe even lower levels should be used. Destructive thermal heating of body tissue results from excessive exposure. This heating effect is especially dangerous to the eyes. The accepted safe level of 10 mW/cm^2 is reached in the near field of a parabolic antenna if the level at $2D^2/\lambda$ is 0.242 mW/cm^2. The equation for power density at the far-field boundary is

$$\text{Power density} = \frac{137.8\ P}{D^2}\ \text{mW/cm}^2 \qquad \text{(Eq 7)}$$

where

P = average power in kilowatts
D = antenna diameter in feet
λ = wavelength in feet

New commercial dishes are expensive, but surplus ones can often be purchased at low cost. Some amateurs build theirs, while others modify UHF TV dishes or circular metal snow sleds for the amateur bands. **Fig 27** shows a dish using the homemade feed just described.

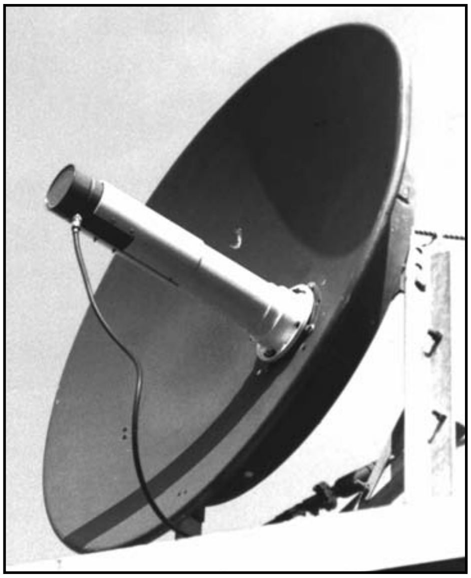

Fig 27—Coffee-can 2304 MHz feed described in text and Fig 26 mounted on a 4-ft dish.

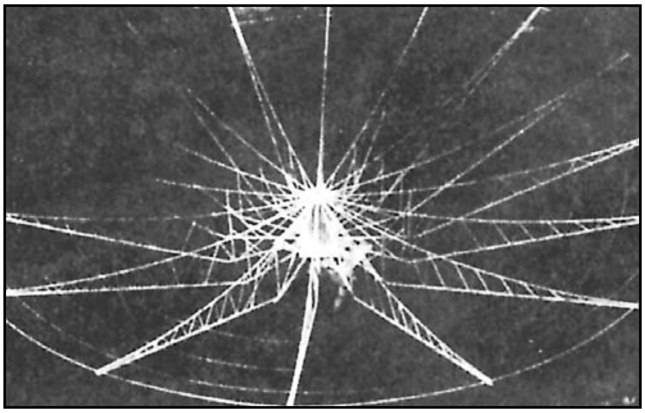

Fig 28—Aluminum framework for a 23-foot dish under construction by ZL1BJQ.

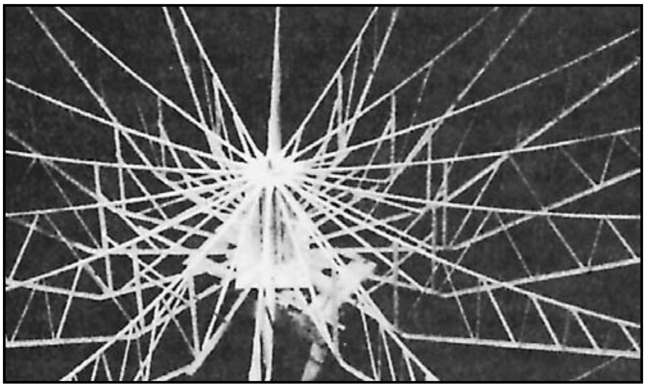

Fig 29—Detailed look at the hub assembly for the ZL1BJQ dish. Most of the structural members are made from ¾-inch T section.

Photos showing a highly ambitious dish project under construction by ZL1BJQ appear in **Figs 28** and **29**. Practical details for constructing this type of antenna are given in Chapter 19. Dick Knadle, K2RIW, described modern UHF antenna test procedures in February 1976 *QST* (see Bibliography). Also see Chapter 19.

OMNIDIRECTIONAL ANTENNAS FOR VHF AND UHF

Local work with mobile stations requires an antenna with wide coverage capabilities. Most mobile work is on FM, and the polarization used with this mode is generally vertical. Some simple vertical systems are described below. Additional material on antennas of this type is presented in Chapter 16, Mobile and Maritime Antennas.

Ground-plane Antennas for 144, 222 and 440 MHz

For the FM operator living in the primary coverage area of a repeater, the ease of construction and low cost of a $^1/_4$ λ ground-plane antenna make it an ideal choice. Three different types of construction are detailed in Figs 30 through 43; the choice of construction method depends upon the materials at hand and the desired style of antenna mounting.

The 144-MHz model shown in **Fig 30** uses a flat piece of sheet aluminum, to which radials are connected with machine screws. A 45° bend is made in each of the radials. This bend can be made with an ordinary bench vise. An SO239 chassis connector is mounted at the center of the aluminum plate with the threaded part of the connector facing down. The vertical portion of the

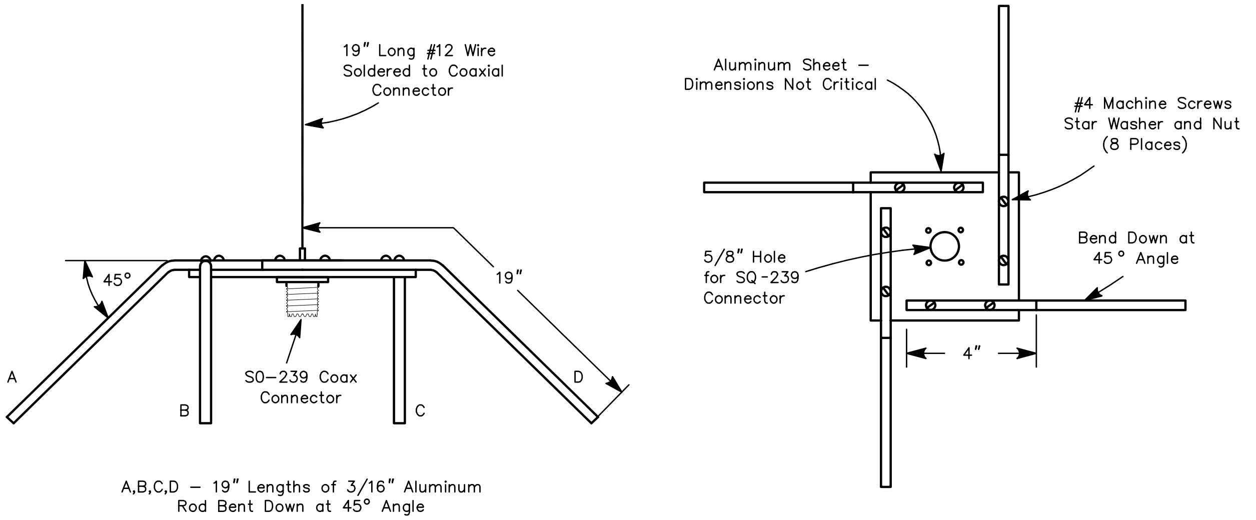
19" Long #12 Wire
Soldered to Coaxial
Connector
45°
19"
A
B
SO–239 Coax
Connector
C
D
A,B,C,D – 19" Lengths of 3/16" Aluminum
Rod Bent Down at 45° Angle
Aluminum Sheet –
Dimensions Not Critical
#4 Machine Screws
Star Washer and Nut
(8 Places)
5/8" Hole
for SQ-239
Connector
Bend Down at
45 ° Angle
4"

Fig 30—These drawings illustrate the dimensions for the 144-MHz ground-plane antenna. The radials are bent down at a 45° angle.

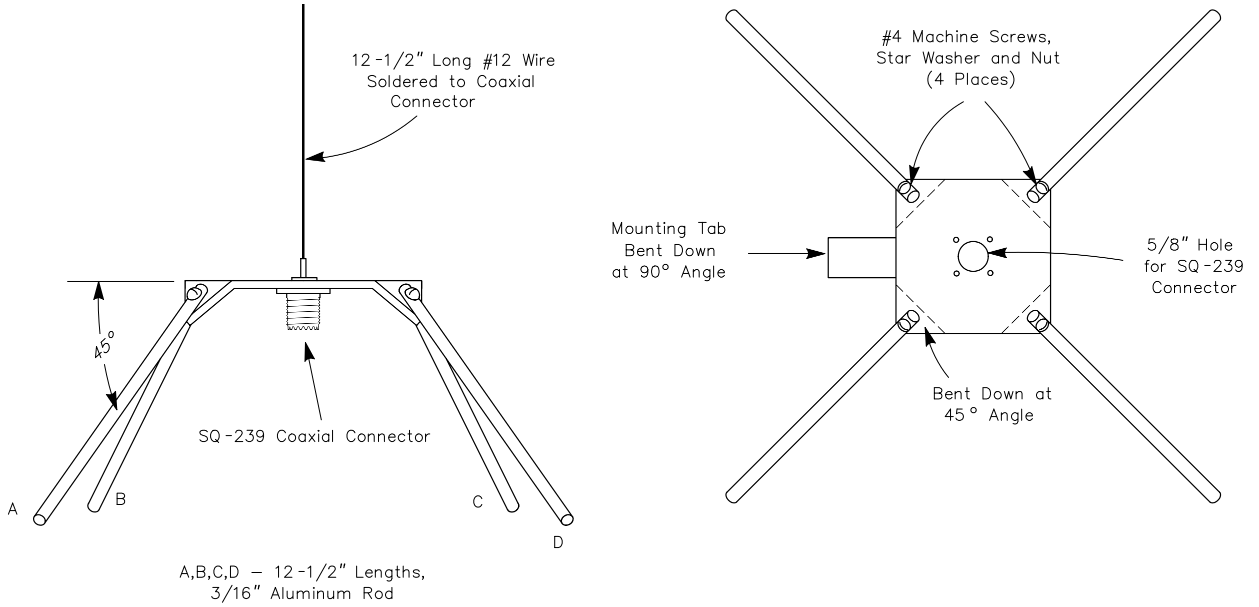
12-1/2" Long #12 Wire
Soldered to Coaxial
Connector
45°
SQ-239 Coaxial Connector
A
B
C
D
A,B,C,D – 12-1/2" Lengths,
3/16" Aluminum Rod
#4 Machine Screws,
Star Washer and Nut
(4 Places)
Mounting Tab
Bent Down
at 90° Angle
5/8" Hole
for SQ-239
Connector
Bent Down at
45 ° Angle

Fig 31—Dimensional information for the 222-MHz ground-plane antenna. Lengths for A, B, C and D are the total distances measured from the center of the SO-239 connector. The corners of the aluminum plate are bent down at a 45° angle rather than bending the aluminum rod as in the 144-MHz model. Either method is suitable for these antennas.

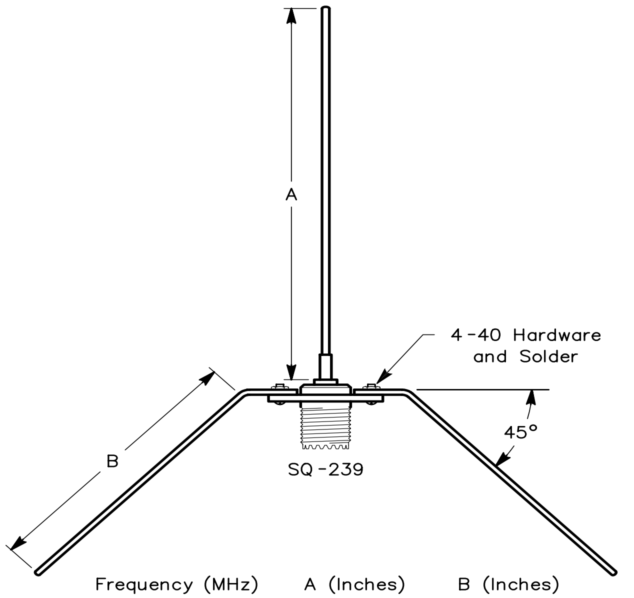

Frequency (MHz)	A (Inches)	B (Inches)
146	19-5/16″	18-11/16″
225	12-5/8″	12″
445	6-3/8″	5-3/4″

Fig 32—Simple ground-plane antenna for the 144-, 222- and 440-MHz bands. The vertical element and radials are $^3/_{32}$- or $^1/_{16}$-in. brass welding rod. Although $^3/_{32}$-in. rod is preferred for the 144-MHz antenna, #10 or #12 copper wire can also be used.

antenna is made of #12 copper wire soldered directly to the center pin of the SO-239 connector.

The 222-MHz version, **Fig 31**, uses a slightly different technique for mounting and sloping the radials. In this case the corners of the aluminum plate are bent down at a 45° angle with respect to the remainder of the plate. The four radials are held to the plate with machine screws, lock washers and nuts. A mounting tab is included in the design of this antenna as part of the aluminum base. A compression type of hose clamp could be used to secure the antenna to a mast. As with the 144-MHz version, the vertical portion of the antenna is soldered directly to the SO-239 connector.

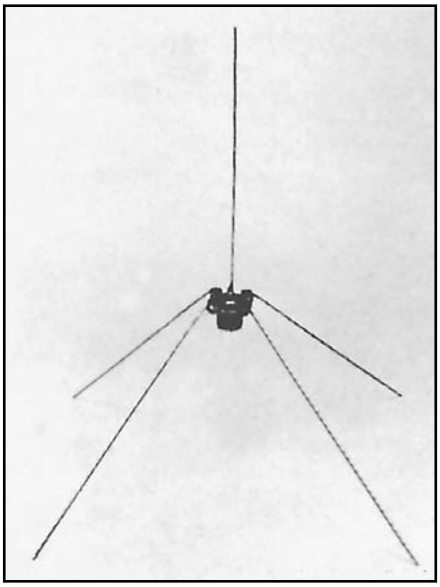

Fig 33—A 440-MHz ground-plane constructed using only an SO-239 connector, no. 4-40 hardware and $^1/_{16}$-in. brass welding rod.

A very simple method of construction, shown in **Figs 32** and **33**, requires nothing more than an SO-239 connector and some no. 4-40 hardware. A small loop formed at the inside end of each radial is used to attach the radial directly to the mounting holes of the coaxial connector. After the radial is fastened to the SO-239 with no. 4-40 hardware, a large soldering iron or propane torch is used to solder the radial and the mounting hardware to the coaxial connector. The radials are bent to a 45° angle and the vertical portion is soldered to the center pin to complete the antenna. The antenna can be mounted by passing the feed line through a mast of ¾-inch ID plastic or aluminum tubing. A compression hose clamp can be used to secure the PL-259 connector, attached to the feed line, in the end of the mast. Dimensions for the 144-, 222- and 440-MHz bands are given in Fig 32.

If these antennas are to be mounted outside it is wise to apply a small amount of RTV sealant or similar material around the areas of the center pin of the connector to prevent the entry of water into the connector and coax line.

The J-Pole Antenna

The J-Pole is a half-wave antenna that is end-fed at its bottom. Since the radiator is longer than that of a 1/4-wave ground-plane antenna, the vertical lobe is compressed down toward the horizon and it has about 1.5 dB of gain compared to the ground-plane configuration. The stub-matching section used to transform the high impedance seen looking into a half-wave to 50 Ω coax is shorted at the bottom, making the antenna look like the letter "J," and giving the antenna its name.

Rigid copper tubing, fittings and assorted hardware can be used to make a really rugged J-pole antenna for 2 meters. When copper tubing is used, the entire assembly can be soldered together, ensuring electrical integrity, and making the whole antenna weatherproof. This

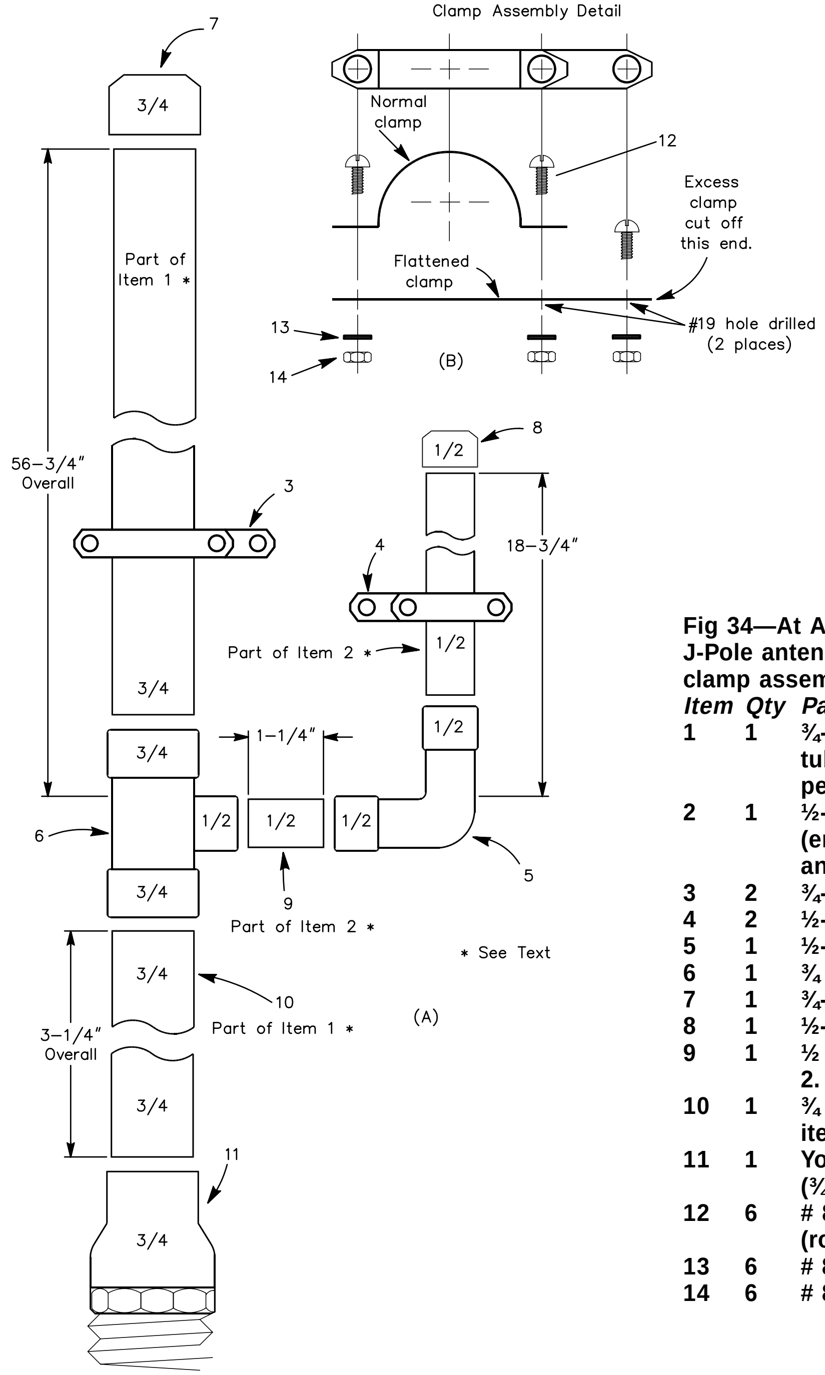
Clamp Assembly Detail
7
3/4
Normal clamp
12
Excess clamp cut off this end.
Part of Item 1 *
Flattened clamp
13
#19 hole drilled (2 places)
14
(B)
8
1/2
56–3/4″ Overall
3
18–3/4″
4
1/2
Part of Item 2 *
3/4
1/2
3/4
1–1/4″
6
1/2
1/2
1/2
5
3/4
9
Part of Item 2 *
* See Text
3/4
10
Part of Item 1 *
(A)
3–1/4″ Overall
3/4
11
3/4

Fig 34—At A, exploded assembly diagram of all-copper J-Pole antenna. At B, detail of clamp assemblies. Both clamp assemblies are the same.

Item	***Qty***	***Part or Material Name***
1	**1**	**¾-inch × 10 foot length of rigid copper tubing (enough for 2 antennas, 60 inches per antenna)**
2	**1**	**½-inch × 10 ft length of rigid copper tubing (enough for 6 antennas, 20 inches per antenna)**
3	**2**	**¾-inch copper pipe clamps**
4	**2**	**½-inch copper pipe clamps**
5	**1**	**½-inch copper elbow**
6	**1**	**¾ × ½-inch copper tee**
7	**1**	**¾-inch copper end cap**
8	**1**	**½-inch copper end cap**
9	**1**	**½ × 1¼-inch copper nipple (Make from item 2. See text)**
10	**1**	**¾ × 3 ¼-inch copper nipple (Make from item 1. See text)**
11	**1**	**Your choice of coupling to mast fitting (¾ × 1 inch NPT used at KD8JB)**
12	**6**	**# 8-32 × ½-inch brass machine screws (round, pan, or binder head)**
13	**6**	**# 8 brass flat washers**
14	**6**	**# 8-32 brass hex nuts**

material came from an article by Michael Hood, KD8JB, in *The ARRL Antenna Compendium,Vol. 4*.

No special hardware or machined parts are used in this antenna, nor are insulating materials needed, since the antenna is always at dc ground. Best of all, even if the parts aren't on sale, the antenna can be built for less than $15. If you only build one antenna, you'll have enough tubing left over to make most of a second antenna.

Construction

Copper and brass is used exclusively in this antenna. These metals get along together, so dissimilar metal corrosion is eliminated. Both metals solder well, too. See **Fig 34**. Cut the copper tubing to the lengths indicated. Item 9 is a $1^{1}/_{4}$-inch nipple cut from the 20-inch length of $^{1}/_{2}$-inch tubing. This leaves $18^{3}/_{4}$ inches for the $^{1}/_{4}$-matching stub. Item 10 is a $3^{1}/_{4}$-inch long nipple cut from the 60-inch length of $^{3}/_{4}$-inch tubing. The $^{3}/_{4}$-wave element should measure $56^{3}/_{4}$-inches long. Remove burrs from the ends of the tubing after cutting, and clean the mating surfaces with sandpaper, steel wool, or emery cloth.

After cleaning, apply a very thin coat of flux to the mating elements and assemble the tubing, elbow, tee, end caps and stubs. Solder the assembled parts with a propane torch and rosin-core solder. Wipe off excess solder with a damp cloth, being careful not to burn yourself. The copper tubing will hold heat for a long time after you've finished soldering. After soldering, set the assembly aside to cool.

Flatten one each of the $^{1}/_{2}$-inch and $^{3}/_{4}$-inch pipe clamps. Drill a hole in the flattened clamp as shown in Fig 34A. Assemble the clamps and cut off the excess metal from the flattened clamp using the unmodified clamp as a template. Disassemble the clamps.

Assemble the $^{1}/_{2}$-inch clamp around the $^{1}/_{4}$-wave element and secure with two of the screws, washers, and nuts as shown in Fig 34B. Do the same with the $^{3}/_{4}$-inch clamp around the $^{3}/_{4}$-wave element. Set the clamps initially to a spot about 4 inches above the bottom of the "J" on their respective elements. Tighten the clamps only finger tight, since you'll need to move them when tuning.

Tuning

The J-Pole can be fed directly from 50-Ω coax through a choke balun (3 turns of the feed coax rolled into a coil about 8 inches in diameter and held together with electrical tape). Before tuning, mount the antenna vertically, about 5 to 10 feet from the ground. A short TV mast on a tripod works well for this purpose. When tuning VHF antennas, keep in mind that they are sensitive to nearby objects—such as your body. Attach the feed line to the clamps on the antenna, and make sure all the nuts and screws are at least finger tight. It really doesn't matter to which element (¾-wave element or stub) you attach the coaxial center lead. The author has done it both ways with no variation in performance. Tune the antenna by moving the two feed-point clamps equal distances a small amount each time until the SWR is minimum at the desired frequency. The SWR will be close to 1:1.

Final Assembly

The final assembly of the antenna will determine its long-term survivability. Perform the following steps with care. After adjusting the clamps for minimum SWR, mark the clamp positions with a pencil and then remove the feed line and clamps. Apply a very thin coating of flux to the inside of the clamp and the corresponding surface of the antenna element where the clamp attaches. Install the clamps and tighten the clamp screws.

Solder the feed line clamps where they are attached to the antenna elements. Now, apply a small amount of solder around the screw heads and nuts where they contact the clamps. Don't get solder on the screw threads! Clean away excess flux with a non-corrosive solvent. After final assembly and erecting/mounting the antenna in the desired location, attach the feed line and secure with the remaining washer and nut. Weather-seal this joint with RTV. Otherwise, you may find yourself repairing the feed line after a couple years.

On-the-Air Performance

Years ago, prior to building the first J-Pole antenna for this station, the author used a standard $^{1}/_{4}$-wave ground plane vertical antenna. While he had no problem working various repeaters around town with a $^{1}/_{4}$-wave antenna, simplex operation left a lot to be desired. The J-Pole performs just as well as a Ringo Ranger, and significantly better than the $^{1}/_{4}$-wave ground-plane vertical.

Practical 6-Meter Yagis

Boom length often proves to be the deciding factor when one selects a Yagi design. **Table 6** shows three 6-meter Yagis designed for convenient boom lengths (6, 12 and 22 feet). The 3-element, 6-foot boom design has 8.0 dBi gain in free space; the 12 foot boom, 5-element version has 10.1 dBi gain, and the 22-foot, 7 element Yagi has a gain of 11.3 dBi. All antennas exhibit better than 22 dB front-to-rear ratio and cover 50 to 51 MHz with better than 1.7:1 SWR.

Half-element lengths and spacings are given in the table. Elements can be mounted to the boom as shown in **Fig 35**. Two muffler clamps hold each aluminum plate to the boom, and two U bolts fasten each element to the plate, which is 0.25 inches thick and 4 × 4 inches square. Stainless steel is the best choice for hardware, however, galvanized hardware can be substituted. Automotive muffler clamps do not work well in this application, because they are not galvanized and quickly rust once exposed to the weather. Please note that the element lengths shown in Table 6 are half the overall element lengths. See page 20-7 to 20-11 in Chapter 20 for practical details of telescoping aluminum elements.

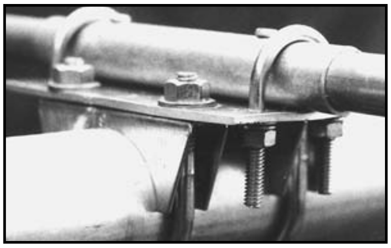

Fig 35—The element to boom clamp. U bolts are used to hold the element to the plate, and 2-in. galvanized muffler clamps hold the plates to the boom.

The driven element is mounted to the boom on a Bakelite or G-10 fiberglass plate of similar dimension to the other mounting plates. A 12-inch piece of Plexiglas rod is inserted into the driven element halves. The Plexiglas allows the use of a single clamp on each side of the element and also seals the center of the elements against moisture. Self-tapping screws are used for electrical connection to the driven element.

Refer to **Fig 36** for driven-element and hairpin match details. A bracket made from a piece of aluminum is used to mount the three SO239 connectors to the driven element plate. A 4:1 transmission-line balun connects the two element halves, transforming the 200 Ω resistance at the hairpin match to 50 Ω at the center connector. Note

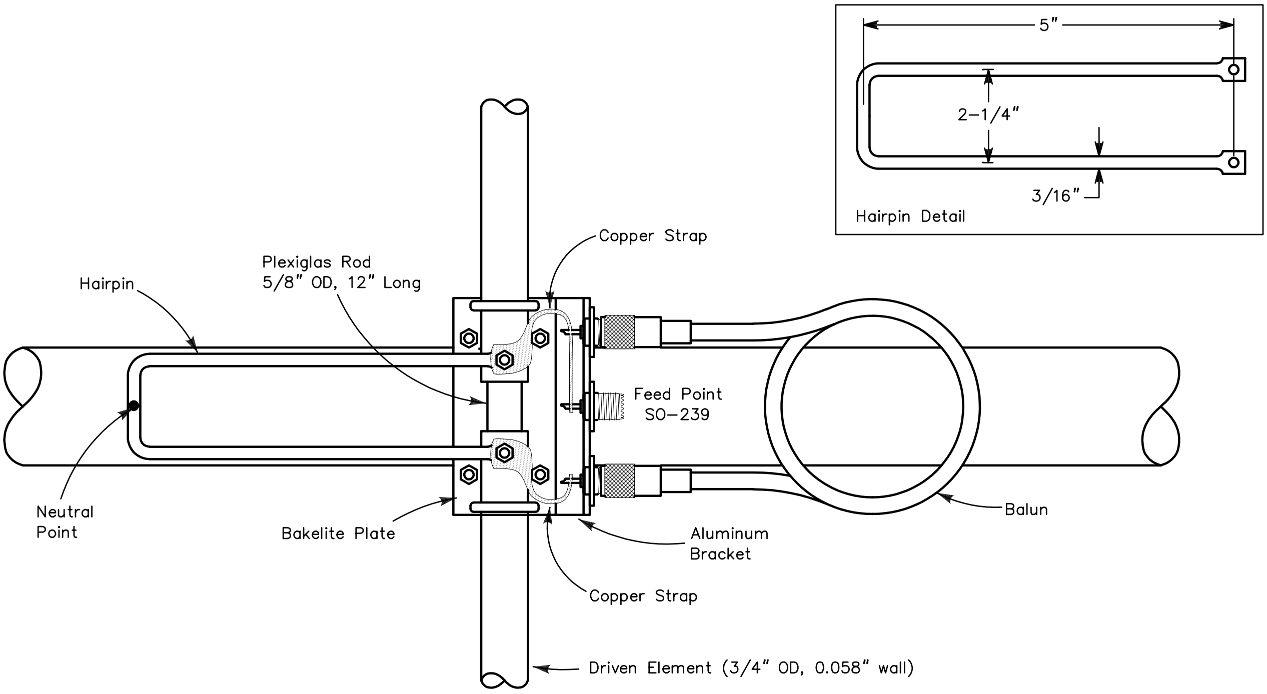

Fig 36—This shows how the driven element and feed system are attached to the boom. The phasing line is coiled and taped to the boom. The center of the hairpin loop may be connected to the boom electrically and mechanically if desired.
Phasing-line lengths:
For cable with 0.80 velocity factor – 7 ft, $10\frac{3}{8}$ in.
For cable with 0.66 velocity factor – 6 ft, $5\frac{3}{4}$ in.

Table 6
Optimized 6-Meter Yagi Designs

	Spacing Between Elements inches	*Seg 1 OD* Length inches*	*Seg2 OD* Length inches*	*Midband Gain F/R*
306-06				
OD		0.750	0.625	
Refl.	0	36	23.500	7.9 dBi
D.E.	24	36	16.000	27.2 dB
Dir. 1	66	36	15.500	
506-12				
OD		0.750	0.625	
Refl.	0	36	24.000	10.1 dBi
D.E.	24	36	17.125	24.7 dB
Dir. 1	12	36	19.375	
Dir. 2	44	36	18.250	
Dir. 3	58	36	15.375	
706-22				
OD		0.750	0.625	
Refl.	0	36	25.000	11.3 dBi
D.E.	27	36	17.250	29.9 dB
Dir. 1	16	36	18.500	
Dir. 2	51	36	15.375	
Dir. 3	54	36	15.875	
Dir. 4	53	36	16.500	
Dir. 5	58	36	12.500	

*See pages 20-7 to 20-11 for telescoping aluminum tubing details.

that the electrical length of the balun is λ/2, but the physical length will be shorter due to the velocity factor of the particular coaxial cable used. The hairpin is connected directly across the element halves. The exact center of the hairpin is electrically neutral and should be fastened to the boom. This has the advantage of placing the driven element at dc ground potential.

The hairpin match requires no adjustment as such. However, you may have to change the length of the driven element slightly to obtain the best match in your preferred portion of the band. Changing the driven-element length will not adversely affect antenna performance. *Do not adjust the lengths or spacings of the other elements—they are optimized already*. If you decide to use a gamma match, add 3 inches to each side of the driven element lengths given in the table for all antennas.

High-Performance Yagis for 144, 222 and 432 MHz

This construction information is presented as an introduction to the three high-performance VHF/UHF Yagis that follow. All were designed and built by Steve Powlishen, K1FO. For years the design of long Yagi antennas seemed to be a mystical black art. The problem of simultaneously optimizing 20 or more element spacings and element lengths presented an almost unsolvable set of simultaneous equations. With the unprecedented increase in computer power and widespread availability of antenna analysis software, we are now able to quickly examine many Yagi designs and determine which approaches work and which designs to avoid.

At 144 MHz and above, most operators desire Yagi antennas two or more wavelengths in length. This length (2λ) is where most classical designs start to fall apart in terms of gain per boom length, bandwidth and pattern quality. Extensive computer and antenna range analysis has proven that the best possible design is a Yagi that has both varying element spacings and varying element lengths.

This design approach starts with closely spaced directors. The director spacings gradually increase until a constant spacing of about 0.4 λ is reached. Conversely, the director lengths start out longest with the first director and decrease in length in a decreasing rate of change until they are virtually constant in length. This method of construction results in a wide gain bandwidth. A bandwidth of 7% of the center frequency at the –1 dB forward-gain points is typical for these Yagis even when they are longer than 10 λ. The log-taper design also reduces the rate of change in driven-element impedance vs frequency. This allows the use of simple dipole driven elements while still obtaining acceptable driven-element SWR over a wide frequency range. Another benefit is that the resonant frequency of the Yagi changes very little as the boom length is increased.

The driven-element impedance also changes moderately with boom length. The tapered approach creates a Yagi with a very clean radiation pattern. Typically, first side lobe levels of ~17 dB in the E plane, ~15 dB in the H plane, and all other lobes at ~20 dB or more are possible on designs from 2 λ to more than 14 λ.

The actual rate of change in element lengths is determined by the diameter of the elements (in wavelengths). The spacings can be optimized for an individual boom length or chosen as a best compromise for most boom lengths.

The gain of long Yagis has been the subject of much debate. Recent measurements and computer analysis by both amateurs and professionals indicates that given an optimum design, doubling a Yagi's boom length will result in a maximum theoretical gain increase of about 2.6 dB. In practice, the real gain increase may be less because of escalating resistive losses and the greater possibility of construction error. **Fig 37** shows the maximum possible gain per boom length expressed in decibels, referenced to an isotropic radiator. The actual number of directors does not play an important part in determining the gain vs boom length as long as a reasonable number of directors are used. The use of more directors per boom length will normally give a wider gain bandwidth, however, a point exists where too many directors will adversely affect all performance aspects.

While short antennas (< 1.5 λ) may show increased gain with the use of quad or loop elements, long Yagis (> 2 λ) will not exhibit measurably greater forward gain or pattern integrity with loop-type elements. Similarly, loops used as driven elements and reflectors will not significantly change the properties of a long log-taper Yagi. Multiple-dipole driven-element assemblies will also not result in any significant gain increase per given boom length when compared to single-dipole feeds.

Once a long-Yagi director string is properly tuned, the reflector becomes relatively non critical. Reflector spacings between 0.15 λ and 0.2 λ are preferred. The spacing can be chosen for best pattern and driven element impedance. Multiple-reflector arrangements will not significantly increase the forward gain of a Yagi which has its directors properly optimized for forward gain. Many multiple-reflector schemes such as tri-reflectors and corner reflectors have the disadvantage of lowering the driven element impedance compared to a single optimum-length reflector. The plane or grid reflector, shown in **Fig 38**, may however reduce the intensity of unwanted rear lobes. This can be used to reduce noise pickup on EME or satellite arrays. This type of reflector will usually increase the driven-element impedance compared to a single reflector. This sometimes makes driven-element matching easier. Keep in mind that even for EME, a plane reflector will add considerable wind load and weight for

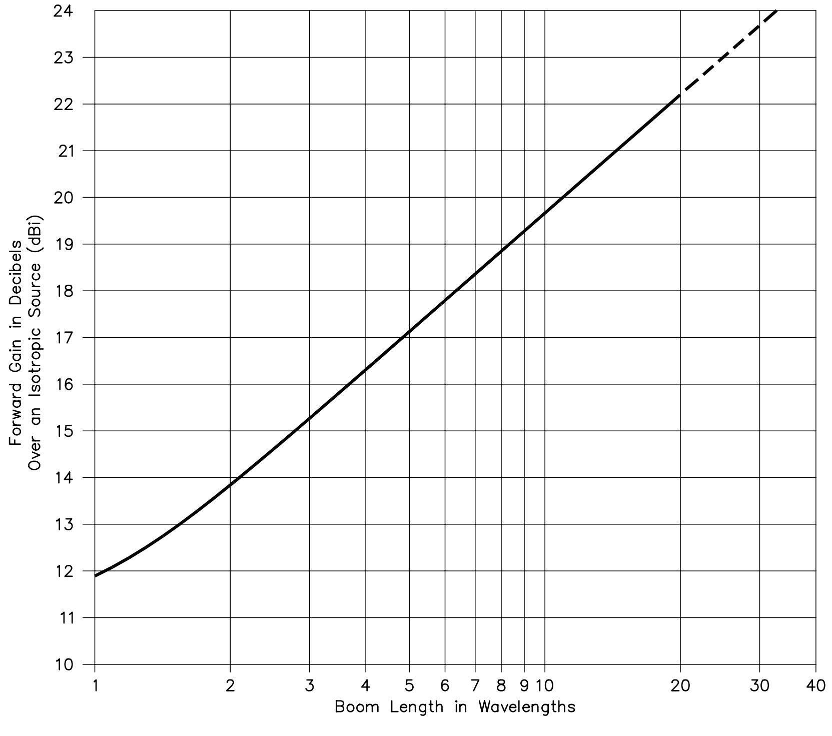

Fig 37—This chart shows maximum gain per boom length for optimally designed long Yagi antennas.

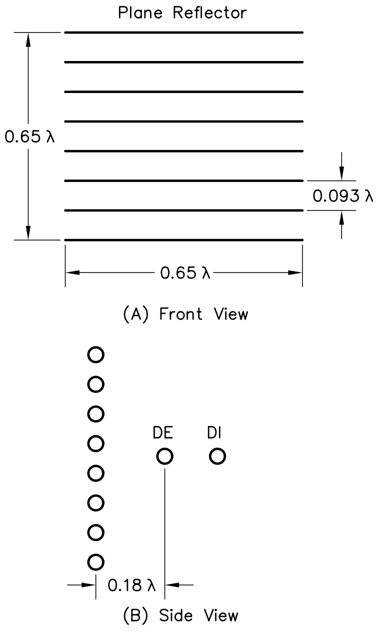

Fig 38—Front and side views of a plane-reflector antenna.

only a few tenths of a decibel of receive signal-to-noise improvement.

Yagi Construction

Normally, aluminum tubing or rod is used for Yagi elements. Hard-drawn enamel-covered copper wire can also be used on Yagis above 420 MHz. Resistive losses are inversely proportional to the square of the element diameter and the square root of its conductivity.

Element diameters of less than $^{3}/_{16}$ inch or 4 mm should not be used on any band. The size should be chosen for reasonable strength. Half-inch diameter is suitable for 50 MHz, $^{3}/_{16}$ to $^{3}/_{8}$ inch for 144 MHz and $^{3}/_{16}$ inch is recommended for the higher bands. Steel, including stainless steel and unprotected brass or copper wire, should not be used for elements.

Boom material may be aluminum tubing, either square or round. High-strength aluminum alloys such as 6061-T6 or 6063-T651 offer the best strength-to-weight advantages. Fiberglass poles have been used (where available as surplus). Wood is a popular low-cost boom material. The wood should be well seasoned and free from knots. Clear pine, spruce and Douglas fir are often used. The wood should be well treated to avoid water absorption and warping.

Elements may be mounted insulated or uninsulated, above or through the boom. Mounting uninsulated elements through a metal boom is the least desirable method unless the elements are welded in place. The Yagi elements will oscillate, even in moderate winds. Over several years this element oscillation will work open the boom holes. This will allow the elements to move in the boom. This will create noise (in your receiver) when the wind blows, as the element contact changes. Eventually the element-to-boom junction will corrode (aluminum oxide is a good insulator). This loss of electrical contact between the boom and element will reduce the boom's effect and change the resonant frequency of the Yagi.

Noninsulated elements mounted above the boom will perform fine as long as a good mechanical connection is made. Insulating blocks mounted above the boom will also work, but they require additional fabrication. One of the most popular construction methods is to mount the elements through the boom using insulating shoulder washers. This method is lightweight and durable. Its main disadvantage is difficult disassembly, making this method of limited use for portable arrays.

If a conductive boom is used, element lengths must be corrected for the mounting method used. The amount of correction is dependent upon the boom diameter in

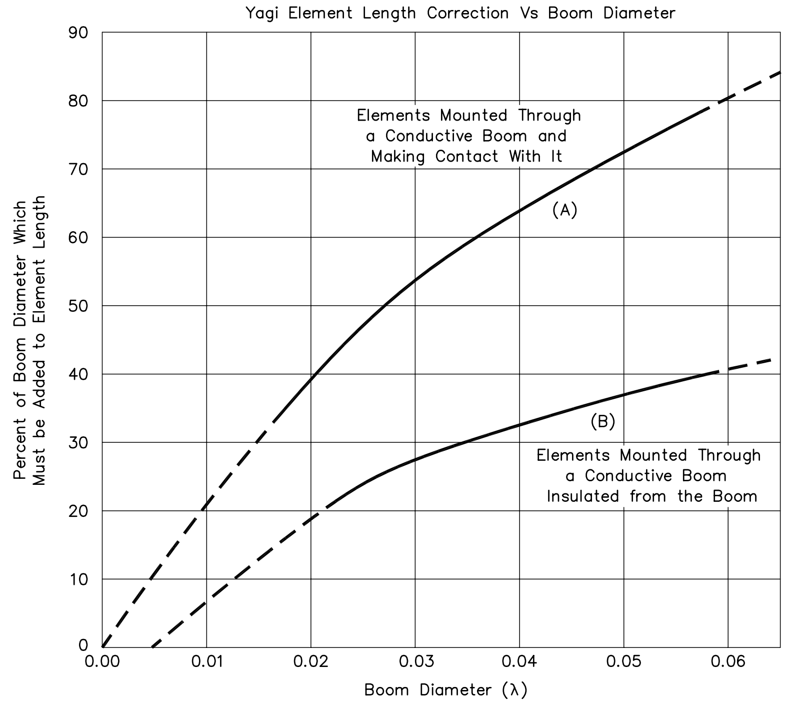

Fig 39—Yagi element correction vs boom diameter. Curve A is for elements mounted through a round or square conductive boom, with the elements in mechanical contact with the boom. Curve B is for insulated elements mounted through a conductive boom, and for elements mounted on top of a conductive boom (elements make electrical contact with the boom). The patterns were corrected to computer simulations to determine Yagi tuning. The amount of element correction is not affected by element diameter.

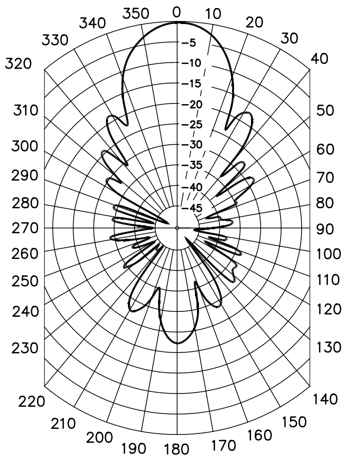

Fig 40—Measured E-plane pattern for the 22-element Yagi. Note: This antenna pattern is drawn on a linear dB grid, rather than on the standard ARRL log-periodic grid, to emphasize low sidelobes.

wavelengths. See **Fig 39**. Elements mounted through the boom and not insulated require the greatest correction. Mounting on top of the boom or through the boom on insulated shoulder washers requires about half of the through-the-boom correction. Insulated elements mounted at least one element diameter above the boom require no correction over the free-space length.

The three following antennas have been optimized for typical boom lengths on each band.

A HIGH-PERFORMANCE 432-MHz YAGI

This 22-element, 6.1-λ, 432-MHz Yagi was originally designed for use in a 12-Yagi EME array built by K1FO. A lengthy evaluation and development process preceded its construction. Many designs were considered and then analyzed on the computer. Next, test models were constructed and evaluated on a home-made antenna range. The resulting design is based on WlEJ's computer-optimized spacings.

The attention paid to the design process has been worth the effort. The 22-element Yagi not only has exceptional forward gain (17.9 dBi), but has an unusually clean radiation pattern. The measured E-plane pattern is shown in **Fig 40**. Note that a 1-dB-per-division axis is used to show pattern detail. A complete description of the design process and construction methods appears in December 1987 and January 1988 *QST*.

Table 7

Specifications for 432-MHz Yagi Family

No. of Ele.	*Boom length(λ)*	*Gain (dBi)**	*F/B ratio (dB)*	*DE imped (Ω)*	*Beamwidth E/H (°)*	*Stacking E/H (inches)*
15	3.4	15.67	21	23	30/32	53/49
16	3.8	16.05	19	23	29/31	55/51
17	4.2	16.45	20	27	28/30	56/53
18	4.6	16.8	25	32	27/29	58/55
19	4.9	17.1	25	30	26/28	61/57
20	5.3	17.4	21	24	25.5/27	62/59
21	5.7	17.65	20	22	25/26.5	63/60
22	6.1	17.9	22	25	24/26	65/62
23	6.5	18.15	27	30	23.5/25	67/64
24	6.9	18.35	29	29	23/24	69/66
25	7.3	18.55	23	25	22.5/23.5	71/68
26	7.7	18.8	22	22	22/23	73/70
27	8.1	19.0	22	21	21.5/22.5	75/72
28	8.5	19.20	25	25	21/22	77/75
29	8.9	19.4	25	25	20.5/21.5	79/77
30	9.3	19.55	26	27	20/21	80/78
31	9.7	19.7	24	25	19.6/20.5	81/79
32	10.2	19.8	23	22	19.3/20	2/80
33	10.6	9.9	23	23	19/19.5	83/81
34	11.0	20.05	25	22	18.8/19.2	84/82
35	11.4	20.2	27	25	18.5/19.0	85/83
36	11.8	20.3	27	26	18.3/18.8	86/84
37	12.2	20.4	26	26	18.1/18.6	87/85
38	12.7	20.5	25	25	18.9/18.4	88/86
39	13.1	20.6	25	23	18.7/18.2	89/87
40	13.5	20.8	26	21	17.5/18	90/88

*Gain is approximate real gain based on gain measurements made on six different-length Yagis.

Like other log-taper Yagi designs, this one can easily be adapted to other boom lengths. Versions of this Yagi have been built by many amateurs. Boom lengths ranged between 5.3 λ (20 elements) and 12.2 λ (37 elements).

The size of the original Yagi (169 inches long, 6.1 λ) was chosen so the antenna could be built from small-diameter boom material ($^{7}/_{8}$-inch and 1 inch round 6061-T6 aluminum) and still survive high winds and ice loading. The 22-element Yagi weighs about 3.5 pounds and has a wind load of approximately 0.8 square feet. This allows a high-gain EME array to be built with manageable wind load and weight. This same low wind load and weight lets the tropo operator add a high-performance 432-MHz array to an existing tower without sacrificing antennas on other bands.

Table 7 lists the gain and stacking specifications for the various length Yagis. The basic Yagi dimensions are shown in **Table 8**. These are free-space element lengths for $^{3}/_{16}$-inch-diameter elements. Boom corrections for the element mounting method must be added in. The element-length correction column gives the length that must be added to keep the Yagi's center frequency optimized for use at 432 MHz. This correction is required to use the same spacing pattern over a wide range of boom lengths. Although any length Yagi will work well, this design is at its best when made with 18 elements or more (4.6 λ). Element material of less than $^{3}/_{16}$-inch diameter is not recommended because resistive losses will reduce the gain by about 0.1 dB, and wet-weather performance will be worse.

Quarter-inch-diameter elements could be used if all elements are shortened by 3 mm. The element lengths are intended for use with a slight chamfer (0.5 mm) cut into the element ends. The gain peak of the array is centered at 437 MHz. This allows acceptable wet-weather performance, while reducing the gain at 432 MHz by only 0.05 dB.

The gain bandwidth of the 22-element Yagi is 31 MHz (at the –1 dB points). The SWR of the Yagi is less than l.4: l between 420 and 440 MHz. **Fig 41** is a network analyzer plot of the driven-element SWR vs frequency. These numbers indicate just how wide the frequency response of a log-taper Yagi can be, even with a simple dipole driven element. In fact, at one antenna gain contest, some ATV operators conducted gain vs frequency measurements from 420 to 440 MHz. The 22-element Yagi beat all entrants including those with so-called broadband feeds.

To peak the Yagi for use on 435 MHz (for satellite use), you may want to shorten all the elements by 2 mm. To peak it for use on 438 MHz (for ATV applications), shorten all elements by 4 mm. If you want to use the Yagi

Table 8

Free-Space Dimensions for 432-MHz Yagi Family

*Element correction is the amount to shorten or lengthen all elements when building a Yagi of that length.
Element lengths are for $^{3}/_{16}$-inch diameter material.

Ele. No.	*Element Position (mm from reflector)*	*Element Length (mm)*	*Element Correction**
Refl	0	340	
DE	104	334	
D1	146	315	
D2	224	306	
D3	332	299	
D4	466	295	
D5	622	291	
D6	798	289	
D7	990	287	
D8	1196	285	
D9	1414	283	
D10	1642	281	–2
D11	1879	279	–2
D12	2122	278	–2
D13	2373	277	–2
D14	2629	276	–2
D15	2890	275	–1
D16	3154	274	–1
D17	3422	273	–1
D18	3693	272	0
D19	3967	271	0
D20	4242	270	0
D21	4520	269	0
D22	4798	269	0
D23	5079	268	0
D24	5360	268	+1
D25	5642	267	+1
D26	5925	267	+1
D27	6209	266	+1
D28	6494	266	+1
D29	6779	265	+2
D30	7064	265	+2
D31	7350	264	+2
D32	7636	264	+2
D33	7922	263	+2
D34	8209	263	+2
D35	8496	262	+2
D36	8783	262	+2
D37	9070	261	+3
D38	9359	261	+3

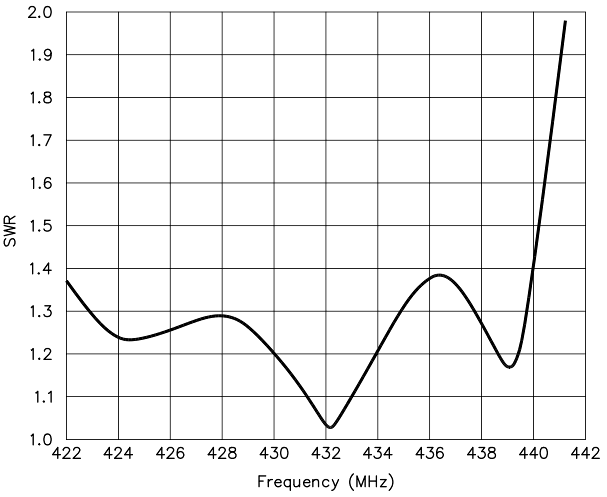

Fig 41—SWR performance of the 22-element Yagi in dry weather.

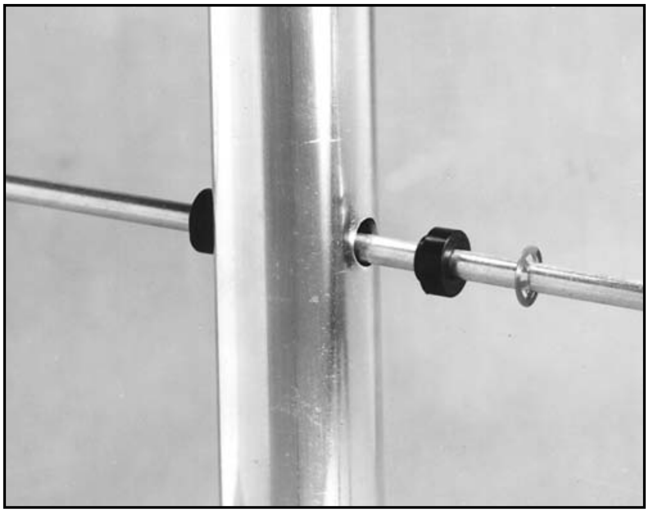

Fig 42—Element-mounting detail. Elements are mounted through the boom using plastic insulators. Stainless steel push-nut retaining rings hold the element in place.

on FM between 440 MHz and 450 MHz, shorten all the elements by 10 mm. This will provide 17.6 dBi gain at 440 MHz, and 18.0 dBi gain at 450 MHz. The driven element may have to be adjusted if the element lengths are shortened.

Although this Yagi design is relatively broadband, it is suggested that close attention be paid to copying the design exactly as built. Metric dimensions are used because they are convenient for a Yagi sized for 432 MHz. Element holes should be drilled within ±2 mm. Element lengths should be kept within ±0.5 mm. Elements can be accurately constructed if they are first rough cut with a hack saw and then held in a vise and filed to the exact length.

The larger the array, the more attention you should pay to making all Yagis identical. Elements are mounted on shoulder insulators and run through the boom (see **Fig 42**). The element retainers are stainless-steel push nuts. These are made by several companies, including Industrial Retaining Ring Co in Irvington, New Jersey, and AuVeco in Ft Mitchell, Kentucky. Local industrial hardware distributors can usually order them for you. The element insulators

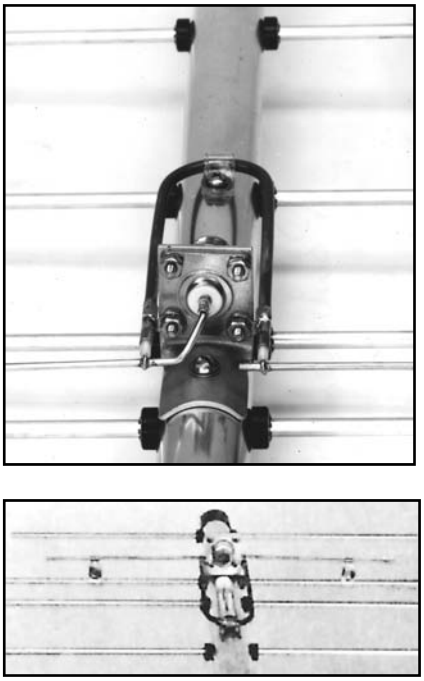

Fig 43—Several views of the driven element and T match.

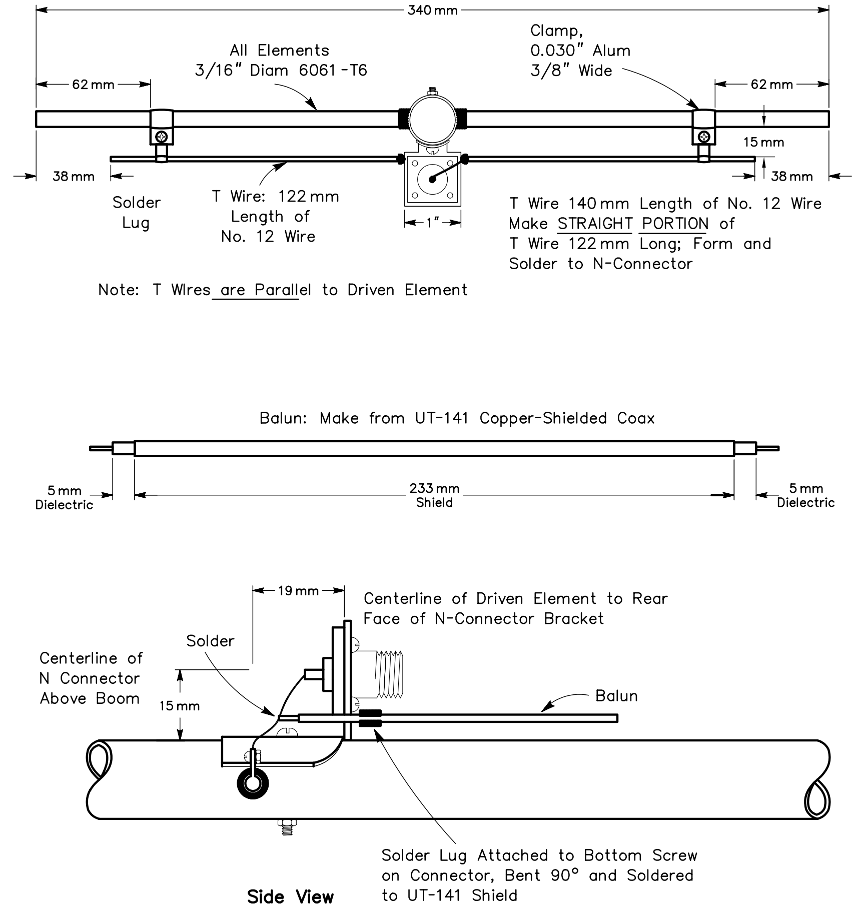
340 mm
All Elements
3/16" Diam 6061-T6
Clamp,
0.030" Alum
3/8" Wide
62 mm
62 mm
15 mm
38 mm
38 mm
Solder
Lug
T Wire: 122 mm
Length of
No. 12 Wire
1"
T Wire 140 mm Length of No. 12 Wire
Make STRAIGHT PORTION of
T Wire 122 mm Long; Form and
Solder to N-Connector
Note: T Wires are Parallel to Driven Element
Balun: Make from UT-141 Copper-Shielded Coax
5 mm
Dielectric
233 mm
Shield
5 mm
Dielectric
19 mm
Centerline of Driven Element to Rear
Face of N-Connector Bracket
Solder
Centerline of
N Connector
Above Boom
15 mm
Balun
Solder Lug Attached to Bottom Screw
on Connector, Bent 90° and Soldered
to UT-141 Shield
Side View

Fig 44—Details of the driven element and T match for the 22-element Yagi. Lengths are given in millimeters to allow precise duplication of the antenna. See text.

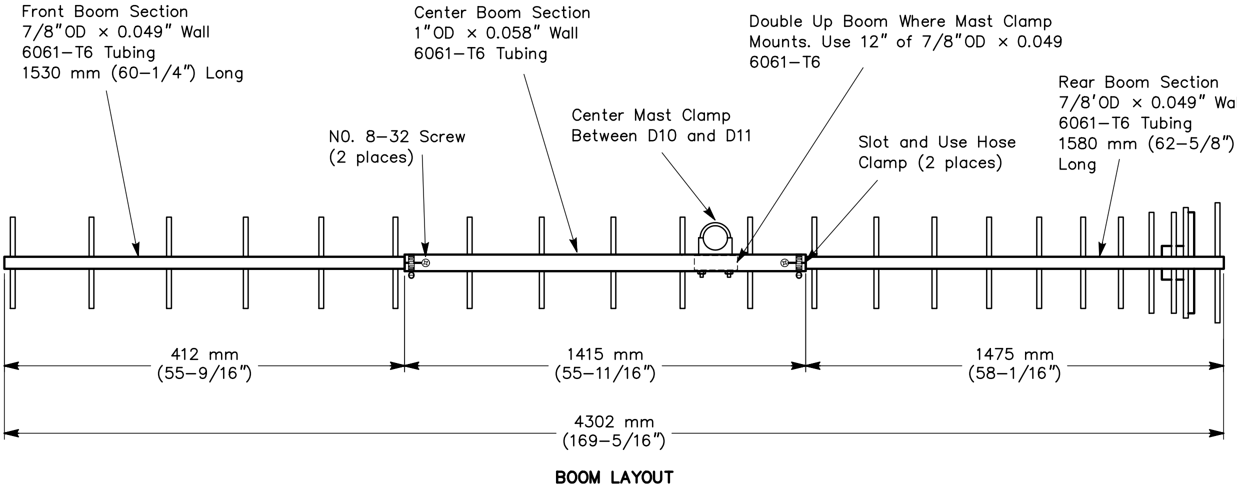
Front Boom Section
7/8"OD × 0.049" Wall
6061–T6 Tubing
1530 mm (60–1/4") Long
Center Boom Section
1"OD × 0.058" Wall
6061–T6 Tubing
Double Up Boom Where Mast Clamp
Mounts. Use 12" of 7/8"OD × 0.049
6061–T6
Rear Boom Section
7/8'OD × 0.049" Wall
6061–T6 Tubing
1580 mm (62–5/8")
Long
Center Mast Clamp
Between D10 and D11
NO. 8–32 Screw
(2 places)
Slot and Use Hose
Clamp (2 places)
412 mm
(55–9/16")
1415 mm
(55–11/16")
1475 mm
(58–1/16")
4302 mm
(169–5/16")
BOOM LAYOUT

Fig 45—Boom-construction information for the 22-element Yagi Lengths are given in millimeters to allow precise duplication of the antenna. See text.

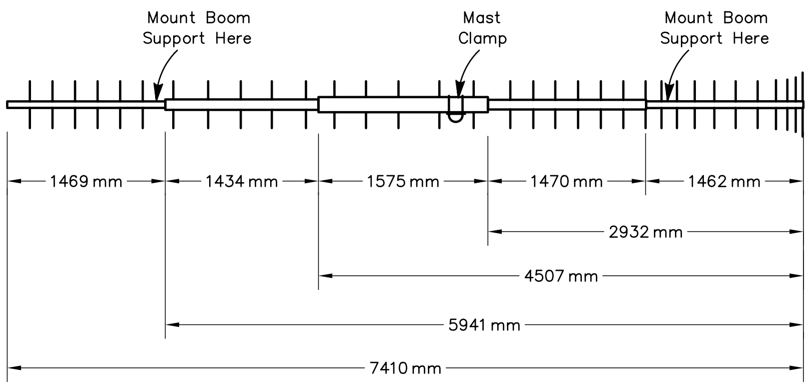

Fig 46—Boom-construction information for the 33-element Yagi. Lengths are given in millimeters to allow precise duplication of the antenna.

are not critical. Teflon or black polyethylene are probably the best materials. The Yagi in the photographs is made with black Delryn insulators, available from C3i in Washington, DC.

The driven element uses a UG-58A/U connector mounted on a small bracket. The UG58A/U should be the type with the press-in center pin. UG-58s with center pins held in by "C" clips will usually leak water. Some connectors use steel retaining clips, which will rust and leave a conductive stripe across the insulator. The T-match wires are supported by the UT-141 balun. RG-303/U or RG-142/U Tefloninsulated cable could be used if UT-141 cannot be obtained. **Fig 43A** and Fig 42B show details of the driven-element construction. Driven element dimensions are given in **Fig 44**.

Dimensions for the 22-element Yagi are listed in **Table 9**. **Fig 45** details the Yagi's boom layout. Element material can be either $^{3}/_{16}$ inch 6061-T6 aluminum rod or hard aluminum welding rod.

A 24-foot-long, 10.6-λ, 33-element Yagi was also built. The construction methods used were the same as the 22-element Yagi. Telescoping round boom sections of 1, $1^{1}/_{8}$, and $1^{1}/_{4}$ inches in diameter were used. A boom support is required to keep boom sag acceptable. At 432 MHz, if boom sag is much more than two or three inches, H-plane pattern distortion will occur. Greater amounts of boom sag will reduce the gain of a Yagi. **Table 10** lists the proper dimensions for the antenna when built with the previously given boom diameters. The boom layout is shown in **Fig 46**, and the driven element is described in **Fig 47**. The 33-element Yagi exhibits the same clean pattern traits as the 22-element Yagi (see **Fig 48**). Measured gain of the 33-element Yagi is 19.9 dBi at 432 MHz. A measured gain sweep of the 33-element Yagi gave a –1 dB gain bandwidth of 14 MHz with the –1 dB points at 424.5 MHz and 438.5 MHz.

A HIGH-PERFORMANCE 144MHZ YAGI

This 144MHz Yagi design uses the latest log-tapered element spacings and lengths. It offers near theoretical gain per boom length, an extremely clean pattern and wide bandwidth. The design is based upon the spacings used in a 4.5-λ 432-MHz computerdeveloped design by W1EJ. It is quite similar to the 432MHz Yagi described elsewhere in this chapter. Refer to that project for additional construction diagrams and photographs.

Mathematical models do not always directly translate into real working examples. Although the computer design provided a good starting point, the author, Steve Powlishen, K1FO, built several test models before the final working Yagi was obtained. This hands-on tuning included changing the element-taper rate in order to obtain the flexibility that allows the Yagi to be built with different boom lengths.

The design is suitable for use from 1.8 λ (10 elements) to 5.1 λ (19 elements). When elements are added to a Yagi, the center frequency, feed impedance and front-to-back ratio will range up and down. A modern tapered design will minimize this effect and allow the builder to select any desired boom length. This Yagi's design capabilities per boom length are listed in **Table 11**.

The gain of any Yagi built around this design will be within 0.1 to 0.2 dB of the maximum theoretical gain at the design frequency of 144.2 MHz. The design is intentionally peaked high in frequency (calculated gain peak is about 144.7 MHz). It has been found that by doing this, the SWR bandwidth and pattern at 144.0 to 144.3 MHz will be better, the Yagi will be less affected by weather and its performance in arrays will be more predictable. This design starts to drop off in performance if built with fewer than 10 elements. At less than 2 λ, more traditional designs perform well.

Table 12 gives free-space element lengths for $^{1}/_{4}$ inch-diameter elements. The use of metric notation allows for much easier dimensional changes during the design stage. Once you become familiar with the metric system, you'll probably find that construction is easier without the burden of cumbersome English fractional units. For $^{3}/_{16}$ inch-diameter elements, lengthen all parasitic elements by 3 mm. If $^{3}/_{8}$ inch diameter elements are used, shorten all of the

Table 9

Dimensions for the 22-Element 432-MHz Yagi

Element Number	*Element Position (mm from reflector)*	*Element Length (mm)*	*Boom Diam (in)*
Refl	30	346	
DE	134	340	
D1	176	321	
D2	254	311	7/8
D3	362	305	
D4	496	301	
D5	652	297	
D6	828	295	
D7	1020	293	
D8	1226	291	
D9	1444	289	
D10	1672	288	
D11	1909	286	
D12	2152	285	1
D13	2403	284	
D14	2659	283	
D15	2920	281	
D16	3184	280	
D17	3452	279	7/8
D18	3723	278	
D19	3997	277	
D20	4272	276	

directors and the reflector by 6 mm. The driven element will have to be adjusted for the individual Yagi if the 12-element design is not adhered to.

For the 12-element Yagi, $^1/_4$-inch diameter elements were selected because smaller-diameter elements become rather flimsy at 2 meters. Other diameter elements can be used as described previously. The 2.5-λ boom was chosen because it has an excellent size and wind load vs gain and pattern trade-off. The size is also convenient; three 6-foot-long pieces of aluminum tubing can be used without any waste. The relatively large-diameter boom sizes ($1^1/_4$ and $1^3/_8$ inches) were chosen, as they provide an extremely rugged Yagi that does not require a boom support. The 12-element 17-foot-long design has a calculated wind survival of close to 120 mph! The absence of a boom support also makes vertical polarization possible.

Table 10

Dimensions for the 33-Element 432-MHz Yagi

Element Number	*Element Position (mm from reflector)*	*Element Length (mm)*	*Boom Diam (in)*
REF	30	348	
DE	134	342	
D1	176	323	
D2	254	313	
D3	362	307	
D4	496	303	1
D5	652	299	
D6	828	297	
D7	1020	295	
D8	1226	293	
D9	1444	291	
D10	1672	290	
D11	1909	288	
D12	2152	287	$1^1/_8$
D13	2403	286	
D14	2659	285	
D15	2920	284	
D16	3184	284	
D17	3452	283	
D18	3723	282	1¼
D19	3997	281	
D20	4272	280	
D21	4550	278	
D22	4828	278	
D23	5109	277	$1^1/_8$
D24	5390	277	
D25	5672	276	
D26	5956	275	
D27	6239	274	
D28	6524	274	1
D29	6809	273	
D30	7094	273	
D31	7380	272	

Longer versions could be made by telescoping smaller-size boom sections into the last section. Some sort of boom support will be required on versions longer than 22 feet. The elements are mounted on shoulder insulators and mounted through the boom. However, elements may

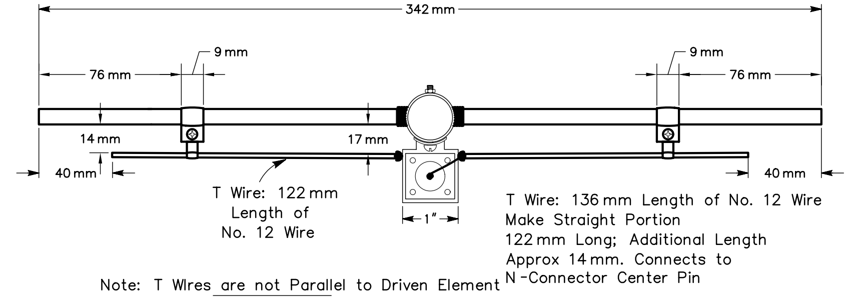
342 mm
9 mm
9 mm
76 mm
76 mm
14 mm
17 mm
40 mm
40 mm
1"
T Wire: 122 mm Length of No. 12 Wire
T Wire: 136 mm Length of No. 12 Wire Make Straight Portion 122 mm Long; Additional Length Approx 14 mm. Connects to N-Connector Center Pin
Note: T Wires are not Parallel to Driven Element

Fig 47—Details of the driven element and T match for the 33-element Yagi. Lengths are given in millimeters to allow precise duplication of the antenna.

Table 11

Specifications for the 144-MHz Yagi Family

No. of Ele.	*Boom Length(λ)*	*Gain (dBd)*	*DE Imped (Ω)*	*FB Ratio (dB)*	*Beamwidth E/H (°)*	*Stacking E/H (°)*
10	1.8	11.4	27	17	39/42	10.2/9.5
11	2.2	12.0	38	19	36/40	11.0/10.0
12	2.5	12.5	28	23	34/37	11.7/10.8
13	2.9	13.0	23	20	32/35	12.5/11.4
14	3.2	13.4	27	18	31/33	12.8/12.0
15	3.6	13.8	35	20	30/32	13.2/12.4
16	4.0	14.2	32	24	29/30	13.7/13.2
17	4.4	14.5	25	23	28/29	14.1/13.6
18	4.8	14.8	25	21	27/28.5	14.6/13.9
19	5.2	15.0	30	22	26/27.5	15.2/14.4

Table 12

Free-Space Dimensions for the 144-MHz Yagi Family

Element diameter is ¼ inch

Element No.	*Element Position (mm from reflector)*	*Element Length*
Refl.	0	1038
DE	312	955
D1	447	956
D2	699	932
D3	1050	916
D4	1482	906
D5	1986	897
D6	2553	891
D7	3168	887
D8	3831	883
D9	4527	879
D10	5259	875
D11	6015	870
D12	6786	865
D13	7566	861
D14	8352	857
D15	9144	853
D16	9942	849
D17	10744	845

be mounted, insulated or uninsulated, above or through the boom, as long as appropriate element length corrections are made. Proper tuning can be verified by checking the depth of the nulls between the main lobe and first side lobes. The nulls should be 5 to 10 dB below the first sidelobe level at the primary operating frequency. The boom layout for the 12-element model is shown in **Fig 49**. The actual corrected element dimensions for the 12-element 2.5-λ Yagi are shown in **Table 13**.

The design may also be cut for use at 147 MHz. There is no need to change element spacings. The element lengths should be shortened by 17 mm for best operation between 146 and 148 MHz. Again, the driven element will have to be adjusted as required.

The driven-element size ($^1/_2$-inch diameter) was chosen to allow easy impedance matching. Any reasonably sized driven element could be used, as long as appropriate length and T-match adjustments are made. Different driven-element dimensions are required if you change the boom length. The calculated natural driven-element impedance is given as a guideline. A balanced T-match was chosen because it's easy to adjust for best SWR and provides a balanced radiation pattern. A 4:1 half-wave coaxial balun is used, although impedance-transforming quarter-wave sleeve baluns could also be used. The cal-

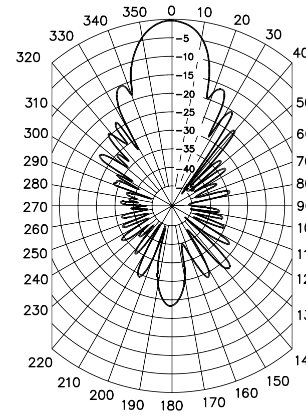

Fig 48—E-plane pattern for the 33-element Yagi. This pattern is drawn on a linear dB grid scale, rather than the standard ARRL log-periodic grid, to emphasize low sidelobes.

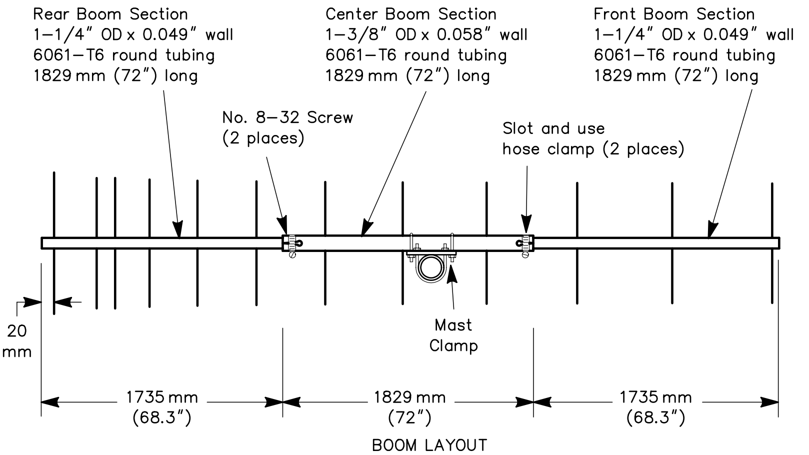
Rear Boom Section
1–1/4″ OD x 0.049″ wall
6061–T6 round tubing
1829 mm (72″) long
Center Boom Section
1–3/8″ OD x 0.058″ wall
6061–T6 round tubing
1829 mm (72″) long
Front Boom Section
1–1/4″ OD x 0.049″ wall
6061–T6 round tubing
1829 mm (72″) long
No. 8–32 Screw
(2 places)
Slot and use
hose clamp (2 places)
20
mm
Mast
Clamp
1735 mm
(68.3″)
1829 mm
(72″)
1735 mm
(68.3″)
BOOM LAYOUT

Fig 49—Boom layout for the 12-element 144-MHz Yagi. Lengths are given in millimeters to allow precise duplication.

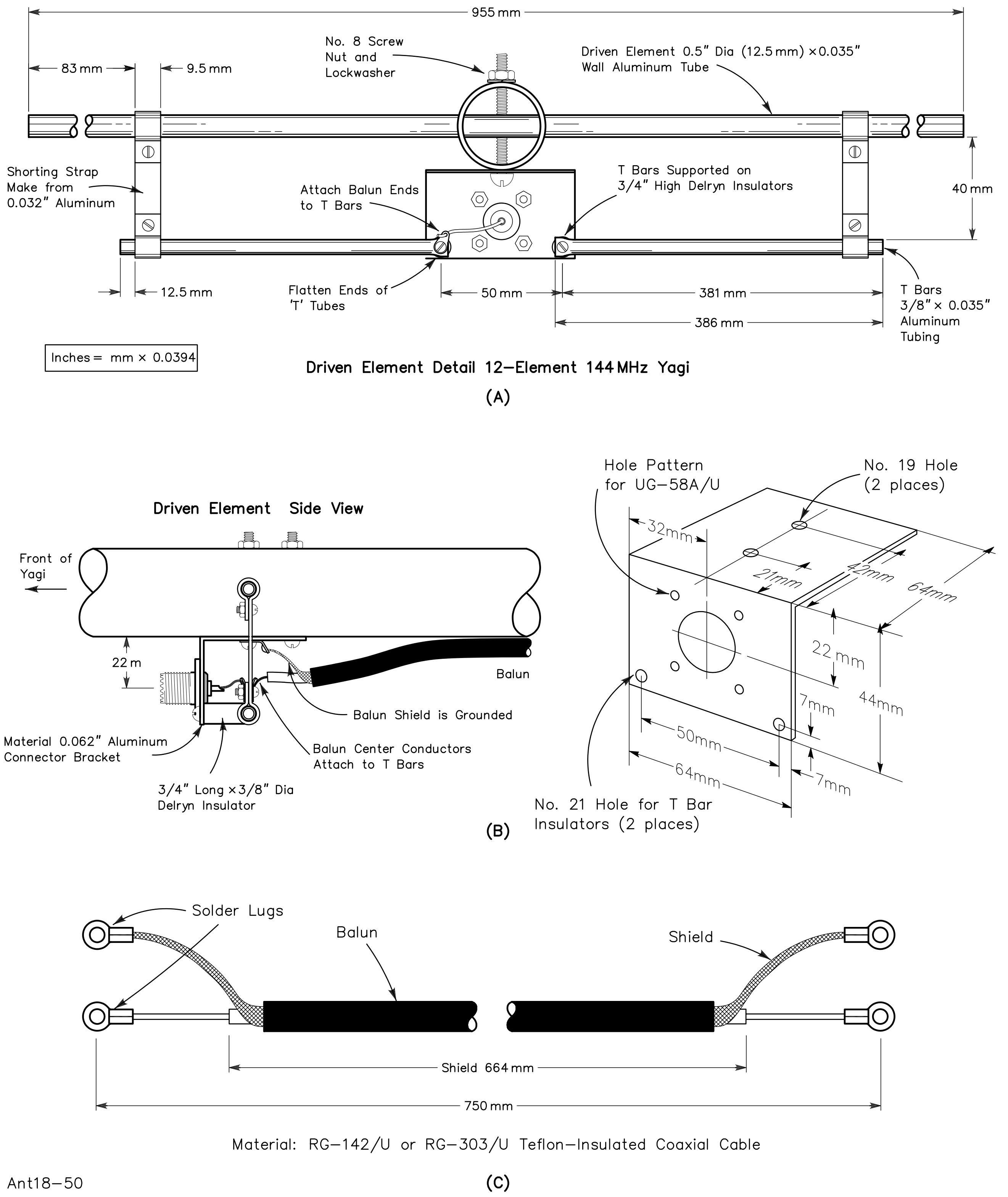
955 mm
83 mm
9.5 mm
No. 8 Screw Nut and Lockwasher
Driven Element 0.5″ Dia (12.5 mm) × 0.035″ Wall Aluminum Tube
Shorting Strap Make from 0.032″ Aluminum
Attach Balun Ends to T Bars
T Bars Supported on 3/4″ High Delryn Insulators
40 mm
12.5 mm
Flatten Ends of 'T' Tubes
50 mm
381 mm
386 mm
T Bars 3/8″ × 0.035″ Aluminum Tubing
Inches = mm × 0.0394
Driven Element Detail 12–Element 144 MHz Yagi
(A)
Driven Element Side View
Front of Yagi
22 m
Balun
Balun Shield is Grounded
Material 0.062″ Aluminum Connector Bracket
Balun Center Conductors Attach to T Bars
3/4″ Long × 3/8″ Dia Delryn Insulator
Hole Pattern for UG–58A/U
No. 19 Hole (2 places)
32mm
21mm
42mm
64mm
22 mm
44mm
7mm
50mm
64mm
7mm
No. 21 Hole for T Bar Insulators (2 places)
(B)
Solder Lugs
Balun
Shield
Shield 664 mm
750 mm
Material: RG–142/U or RG–303/U Teflon–Insulated Coaxial Cable
Ant18–50
(C)

Fig 50—Driven-element detail for the 12-element 144-MHz Yagi. Lengths are given in millimeters to allow precise duplication.

Table 13

Dimensions for the 12-Element 2.5-λ Yagi

Element Number	*Element Position (mm from reflector)*	*Element Length (mm)*	*Boom Diam (in)*
Refl.	0	1044	
DE	312	955	
D1	447	962	$1^{1}/_{4}$
D2	699	938	
D3	1050	922	
D4	1482	912	
D5	1986	904	
D6	2553	898	$1^{3}/_{8}$
D7	3168	894	
D8	3831	889	
D9	4527	885	$1^{1}/_{4}$
D10	5259	882	

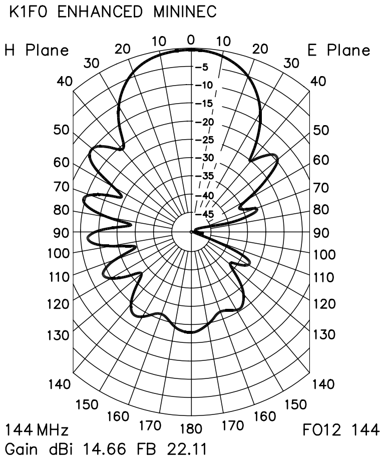

Fig 51—H- and E-plane pattern for the 12-element 144-MHz Yagi.

culated natural impedance will be useful in determining what impedance transformation will be required at the 200-Ω balanced feed point. Chapter 26, Coupling the Line to the Antenna, contains information on calculating folded-dipole and T-match driven-element parameters. A balanced feed is important for best operation on this antenna. Gamma matches can severely distort the pattern balance. Other useful driven-element arrangements are the Delta match and the folded dipole, if you're willing to sacrifice some flexibility. **Fig 50** details the driven-element dimensions.

A noninsulated driven element was chosen for mounting convenience. An insulated driven element may also be used. A grounded driven element may be less affected by static build-up. On the other hand, an insulated driven element allows the operator to easily check his feed lines for water or other contamination by the use of an ohmmeter from the shack.

Table 14

Free-Space Dimensions for the 222-MHz Yagi Family

Element diameter is ${}^{3}/_{16}$-inch.

Element No.	*Element Position (mm from reflector)*	*Element Length (mm)*
Refl.	0	676
DE	204	647
D1	292	623
D2	450	608
D3	668	594
D4	938	597
D5	1251	581
D6	1602	576
D7	1985	573
D8	2395	569
D9	2829	565
D10	3283	562
D11	3755	558
D12	4243	556
D13	4745	554
D14	5259	553
D15	5783	552
D16	6315	551
D17	6853	550
D18	7395	549
D19	7939	548
D20	8483	547

Fig 51 shows computer-predicted E- and H-plane radiation patterns for the 12-element Yagi. The patterns are plotted on a l-dB-per-division linear scale instead of the usual ARRL polar-plot graph. This expanded scale plot is used to show greater pattern detail. The pattern for the 12-element Yagi is so clean that a plot done in the standard ARRL format would be almost featureless, except for the main lobe and first sidelobes.

The excellent performance of the 12-element Yagi is demonstrated by the reception of Moon echoes from several of the larger 144MHz EME stations with only one 12-element Yagi. Four of the 12-element Yagis will make an excellent starter EME array, capable of working many EME QSOs while being relatively small in size. The advanced antenna builder can use the information in Table 11 to design a dream array of virtually any size.

A HIGH-PERFORMANCE 222-MHz YAGI

Modern tapered Yagi designs are easily applied to 222 MHz. This design uses a spacing progression that is in between the 12-element 144-MHz design, and the 22-element 432-MHz design presented elsewhere in this chapter. The result is a design with maximum gain per boom length, a clean, symmetrical radiation pattern, and wide bandwidth. Although it was designed for weak-signal work (tropospheric scatter and EME), the design is

Table 15
Specifications for the 222-MHz Yagi Family

No of Ele.	*Boom Length(λ)*	*Gain (dBd)*	*FB Ratio (dB)*	*DE Imped (Ω)*	*Beamwidth E/H (°)*	*Stacking E/H (feet)*
12	2.4	12.3	22	23	37/39	7.1/6.7
13	2.8	12.8	19	28	33/36	7.8/7.2
14	3.1	13.2	20	34	32/34	8.1/7.6
15	3.5	13.6	24	30	30/33	8.6/7.8
16	3.9	14.0	23	23	29/31	8.9/8.3
17	4.3	14.35	20	24	28/30.5	9.3/8.5
18	4.6	14.7	20	29	27/29	9.6/8.9
19	5.0	15.0	22	33	26/28	9.9/9.3
20	5.4	15.3	24	29	25/27	10.3/9.6
21	5.8	15.55	23	24	24.5/26.5	10.5/9.8
22	6.2	15.8	21	23	24/26	10.7/10.2

suited to all modes of 222-MHz operation, such as packet radio, FM repeater operation and control links.

The spacings were chosen as the best compromise for a 3.9-λ 16-element Yagi. The 3.9-λ design was chosen, like the 12-element 144-MHz design, because it fits perfectly on a boom made from three 6-foot-long aluminum tubing sections. The design is quite extensible, and models from 12 elements (2.4 λ) to 22 elements (6.2 λ) can be built from the dimensions given in **Table 14**. Note that free-space lengths are given. They must be corrected for the element-mounting method. Specifications for various boom lengths are shown in **Table 15**.

Construction

Large-diameter ($1^{1}/_{4}$- and $1^{3}/_{8}$-inch diameter) boom construction is used, eliminating the need for boom supports. The Yagi can also be used vertically polarized. Three-sixteenths-inch-diameter aluminum elements are used. The exact alloy is not critical; 6061-T6 was used, but hard aluminum welding rod is also suitable. Quarter-inch-diameter elements could also be used if all elements are shortened by 3 mm. Three-eighths-inch-diameter elements would require 10-mm shorter lengths. Elements smaller than $^{3}/_{16}$ inch-diameter are not recommended. The elements are insulated and run through the boom. Plastic shoulder washers and stainless steel retainers are used to hold the elements in place. The various pieces needed to build the Yagi may be obtained from C3i in Washington, DC. **Fig 52** details the boom layout for the 16-element Yagi. **Table 16** gives the dimensions for the 16-element Yagi as built. The driven element is fed with a T match and a 4:1 balun. See **Fig 53** for construction details. See the 432-MHz Yagi project elsewhere in this chapter for additional photographs and construction diagrams.

The Yagi has a relatively broad gain and SWR curve, as is typical of a tapered design, making it usable over a wide frequency range. The example dimensions are intended for use at 222.0 to 222.5 MHz. The 16-element Yagi is quite usable to more than 223 MHz. The best compromise for covering the entire band is to shorten all parasitic elements by 4 mm. The driven element will have to be adjusted in length for best match. The position of the T-wire shorting straps may also have to be moved.

The aluminum boom provides superior strength, is lightweight, and has a low wind-load cross section. Aluminum is doubly attractive, as it will long outlast wood and fiberglass. Using state-of-the-art designs, it is unlikely that significant performance increases will be achieved in the next few years. Therefore, it's in your best interest to build an antenna that will last many years. If suitable wood or fiberglass poles are readily available, they may be used without any performance degradation, at least when the wood is new and dry. Use the free-space element lengths given in Table 16 for insulated-boom construction.

The pattern of the 16-element Yagi is shown in **Fig 54**. Like the 144-MHz Yagi, a l-dB-per-division plot is used to detail the pattern accurately. This 16-element design makes a good building block for EME or tropo DX arrays. Old-style narrow-band Yagis often perform

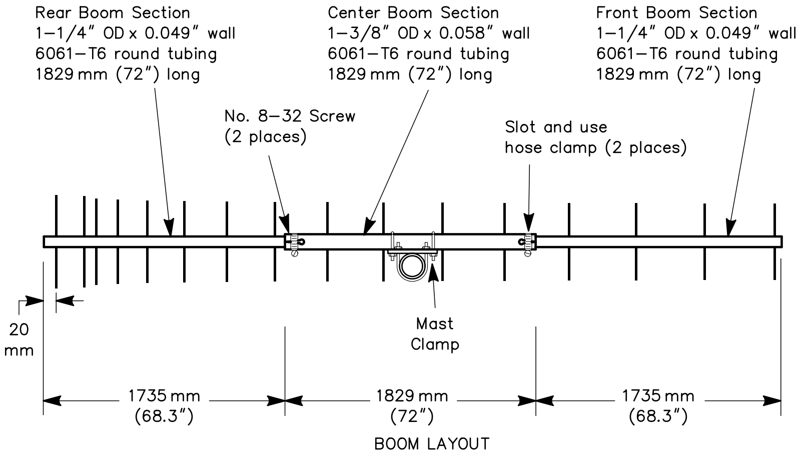

Fig 52—Boom layout for the 16-element 222-MHz Yagi. Lengths are given in millimeters to allow precise duplication.

Table 16
Dimensions for 16-Element 3.9-λ 222-MHz Yagi

Element Number	*Element Position (mm from reflector)*	*Element Length (mm)*	*Boom Diam (in)*
Refl.	0	683	
DE	204	664	
D1	292	630	
D2	450	615	
D3	668	601	1¼
D4	938	594	
D5	1251	588	
D6	1602	583	
D7	1985	580	
D8	2395	576	
D9	2829	572	1⅜
D10	3283	569	
D11	3755	565	
D12	4243	563	
D13	4745	561	1¼
D14	5259	560	

unpredictably when used in arrays. The theoretical 3.0-dB stacking gain is rarely observed. The 16-element Yagi (and other versions of the design) reliably provides stacking gains of nearly 3 dB. (The spacing dimensions listed in Table 15 show just over 2.9 dB stacking gain.) This has been found to be the best compromise between gain, pattern integrity and array size. Any phasing line losses will subtract from the possible stacking gain. Mechanical misalignment will also degrade the performance of an array.

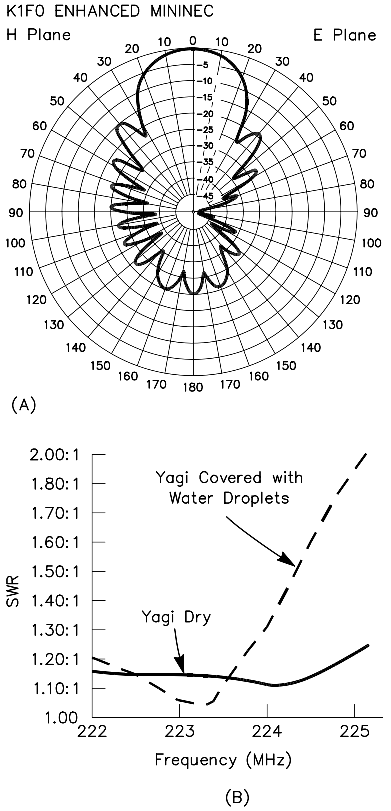

Fig 54—H- and E-plane patterns for the 16-element 222-MHz Yagi at A. The driven-element T-match dimensions were chosen for the best SWR compromise between wet and dry weather conditions. The SWR vs frequency curve shown at B demonstrates the broad frequency response of the Yagi design.

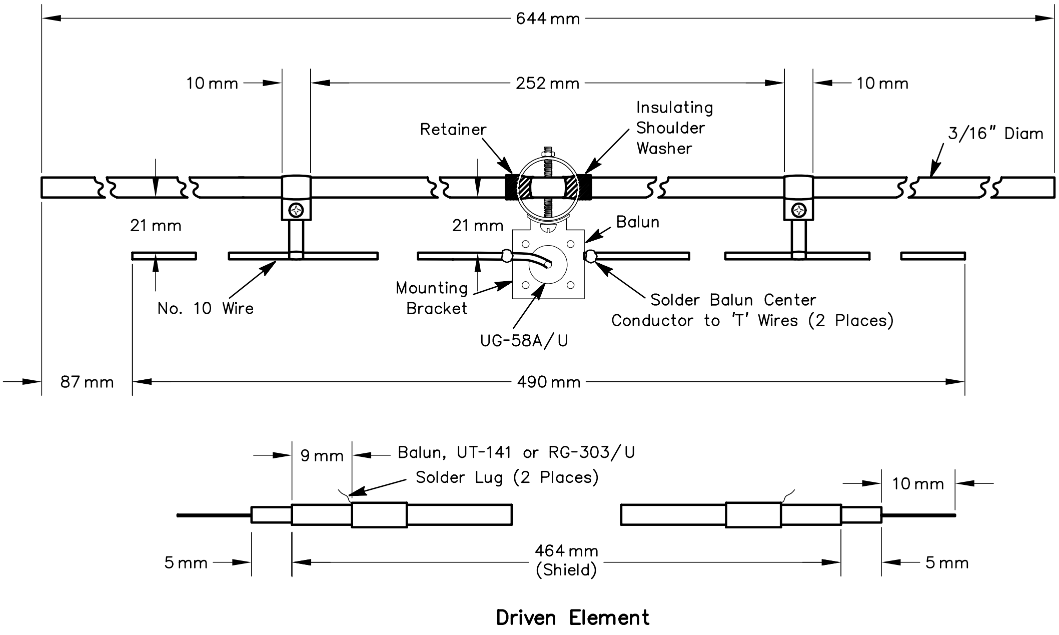
644 mm
10 mm
252 mm
10 mm
Retainer
Insulating Shoulder Washer
3/16" Diam
21 mm
21 mm
Balun
No. 10 Wire
Mounting Bracket
UG-58A/U
Solder Balun Center Conductor to 'T' Wires (2 Places)
87 mm
490 mm
9 mm
Balun, UT-141 or RG-303/U
Solder Lug (2 Places)
10 mm
5 mm
464 mm (Shield)
5 mm
Driven Element

Fig 53—Driven-element detail for the 16-element 222-MHz Yagi. Lengths are given in millimeters to allow precise duplication.

A 144 MHz 2-Element Quad

The basic 2-element quad array for 144 MHz is shown in **Fig 55**. The supporting frame is 1 × 1-inch wood, of any kind suitable for outdoor use. Elements are #8 aluminum wire. The driven element is 1 λ (83 inches) long, and the reflector five percent longer (87 inches). Dimensions are not critical, as the quad is relatively broad in frequency response.

The driven element is open at the bottom, its ends fastened to a plastic block. The block is mounted at the bottom of the forward vertical support. The top portion of the element runs through the support and is held firmly by a screw running into the wood and then bearing on the aluminum wire. Feed is by means of 50-Ω coax, connected to the driven-element loop.

The reflector is a closed loop, its top and bottom portions running through the rear vertical support. It is held in position with screws at the top and bottom. The loop can be closed by fitting a length of tubing over the element ends, or by hammering them flat and bolting them together as shown in the sketch.

The elements in this model are not adjustable, though this can easily be done by the use of stubs. It would then be desirable to make the loops slightly smaller to compensate for the wire in the adjusting stubs. The driven element stub would be trimmed for length and the point of connection for the coax would be adjustable for best match. The reflector stub can be adjusted for maximum gain or maximum F/B ratio, depending on the builder's requirements.

In the model shown only the spacing is adjusted, and this is not particularly critical. If the wooden supports are made as shown, the spacing between the elements can be adjusted for best match, as indicated by an SWR meter connected in the coaxial line. The spacing has little effect on the gain (from 0.15 to 0.25 λ), so the variation in impedance with spacing can be used for matching. This also permits use of either 50- or 75-Ω coax for the transmission line.

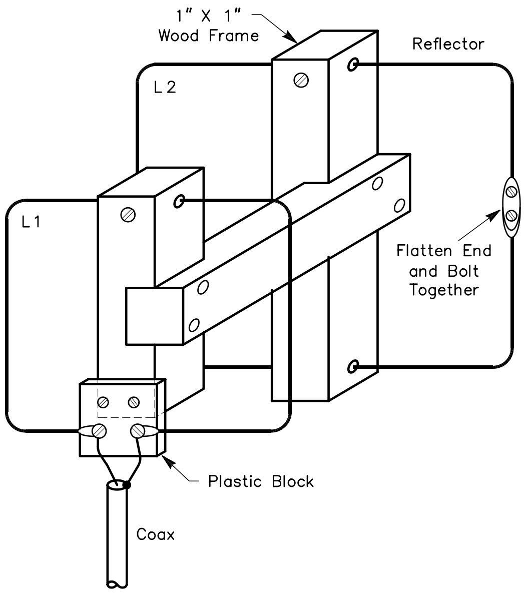

Fig 55—Mechanical details of a 2-element quad for 144 MHz. The driven element, L1, is one wavelength long; reflector L2 is 5% longer. With the transmission line connected as shown here, the resulting radiation is horizontally polarized. Sets of elements of this type can be stacked horizontally and vertically for high gain with broad frequency response. Recommended bay spacing is $^1/_2$ λ between adjacent element sides. The example shown may be fed directly with 50-Ω coax.

A Portable 144 MHz 4-Element Quad

Element spacing for quad antennas found in the literature ranges from 0.14 λ to 0.25 λ. Factors such as the number of elements in the array and the parameters to be optimized (F/B ratio, forward gain, bandwidth, etc), determine the optimum element spacing within this range. The 4-element quad antenna described here was designed for portable use, so a compromise between these factors was chosen. This antenna, pictured in **Fig 56**, was designed and built by Philip D'Agostino, W1KSC.

Based on several experimentally determined correction factors related to the frequency of operation and the wire size, optimum design dimensions were found to be as follows.

$$\text{Reflector length (ft)} = \frac{1046.8}{f_{MHz}} \tag{Eq 8}$$

$$\text{Driven element (ft)} = \frac{985.5}{f_{MHz}} \tag{Eq 9}$$

$$\text{Directors (ft)} = \frac{937.3}{f_{MHz}} \tag{Eq 10}$$

Cutting the loops for 146 MHz provides satisfactory performance across the entire 144MHz band.

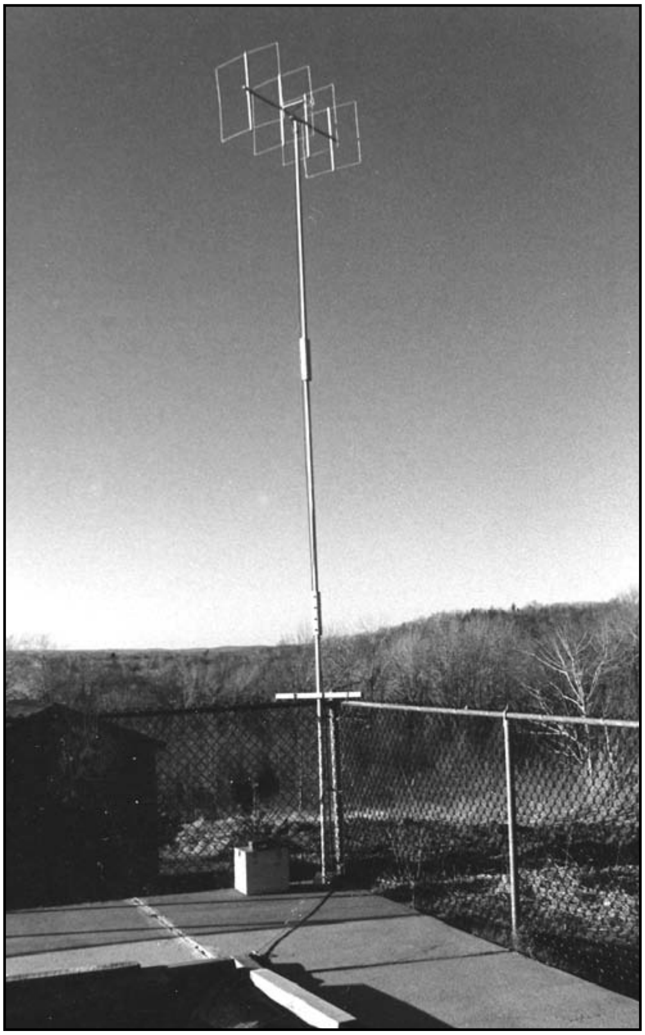

Fig 56—The 4-element 144-MHz portable quad, assembled and ready for operation. Sections of clothes closet poles joined with pine strips make up the mast. (*Photo by Adwin Rusczek, W1MPO)*

Materials

The quad was designed for quick and easy assembly and disassembly, as illustrated in **Fig 57**. Wood (clear trim pine) was chosen as the principal building material because of its light weight, low cost, and ready availability. Pine is used for the boom and element supporting arms. Round wood clothes closet poles comprise the mast material. Strips connecting the mast sections are made of heavier pine trim. Elements are made of no. 8 aluminum wire. Plexiglas is used to support the feed point. **Table 17** lists the hardware and other parts needed to duplicate the quad.

Construction

The elements of the quad are assembled first. The mounting holes in the boom should be drilled to accommodate 1$^{1}/_{2}$ inch no. 8 hardware. Measure and mark the locations where the holes are to be drilled in the element spreaders, **Fig 58**. Drill the holes in the spreaders just

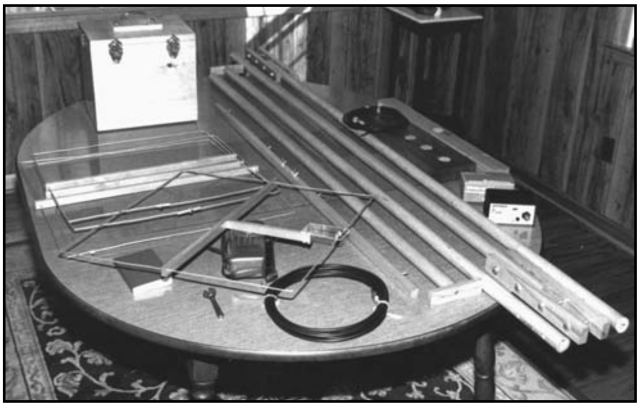

Fig 57—The complete portable quad, broken down for travel. Visible in the foreground is the driven element. The pine box in the background is a carrying case for equipment and accessories. A hole in the lid accepts the mast, so the box doubles as a base for a short mast during portable operation. *(W1MPO photo*)

Table 17

Parts List for the 144 MHz 4-element Quad

Boom: ¾ × ¾ ×48-in. pine

Driven element support (spreader):
½ × ¾ × 21¼ in. pine

Driven element feed point strut: ½ × ¾ × 7½ in. pine

Reflector support (spreader): ½ × ¾ × 22½ in. pine

Director supports (spreaders):
½ × ¾ × 20¼ in. pine, 2 req'd

Mast brackets: ¾ × 1½ × 12 in. heavy pine trim, 4 req'd

Boom to mast bracket: ½ × 1$^{5}/_{8}$ × 5 in. pine

Element wire: Aluminum ground wire
(Radio Shack no. 15-035)

Wire clamps: ¼in. electrician's copper or zinc plated steel clamps, 3 req'd

Boom hardware:
- 6 no. 8-32 × 1½ in. stainless steel machine screws
- 6 no. 8-32 stainless steel wing nuts
- 12 no. 8 stainless steel washers

Mast hardware:
- 8 hex bolts, ¼-20 × 3½ in.
- 8 hex nuts, ¼-20
- 16 flat washers

Mast material: 1$^{5}/_{16}$ in. × 6 ft wood clothes closet poles, 3 req'd

Feed point support plate: 3½ × 2½ in. Plexiglas sheet

Wood preparation materials:
Sandpaper, clear polyurethane, wax

Feed line: 52-W RG-8 or RG-58 cable

Feed line terminals: Solder lugs for no. 8 or larger hardware, 2 req'd

Miscellaneous hardware:
4 small machine screws, nuts, washers; 2 flat-head wood screws

large enough to accept the #8 wire elements. It is important to drill all the holes straight so the elements line up when the antenna is assembled.

Construction of the wire elements is easiest if the directors are made first. A handy jig for bending the elements can be made from a piece of 2 × 3-inch wood cut to the side length of the directors. It is best to start with about 82 inches of wire for each director. The excess can be cut off when the elements are completed. (The total length of each director is 77 inches.) Two bends should initially be made so the directors can be slipped into the spreaders before the remaining corners are bent. See **Fig 59**. Electrician's copper-wire clamps can be used to join the wires after the final bends are made, and they facilitate adjustment of element length. The reflector is made the same way as the directors, but the total length is 86 inches.

The driven element, total length 81 inches, requires special attention, as the feed attachment point needs to be adequately supported. An extra hole is drilled in the driven element spreader to support the feed-point strut, as shown in **Fig 60**. A Plexiglas plate is used at the feed point to support the feed- point hardware and the feed line. The feed-point support strut should be epoxied to the spreader, and a wood screw used for extra mechanical strength.

For vertical polarization, locate the feed point in the center of one side of the driven element, as shown in Fig 60. Although this arrangement places the spreader supports at voltage maxima points on the four loop conductors, D'Agostino reports no adverse effects during operation. However, if the antenna is to be left exposed to the weather, the builder may wish to modify the design to provide support for the loops at current maxima points, such as shown in Fig 60. (The element of Fig 60 should be rotated 90° for horizontal polarization.)

Orient the driven element spreader so that it mounts properly on the boom when the antenna is assembled. Bend the driven element the same way as the reflector and directors, but do not leave any overlap at the feed point. The ends of the wires should be 3/4 inch apart where they mount on the Plexiglas plate. Leave enough excess that small loops can be bent in the wire for attachment to the coaxial feed line with stainless steel hardware.

Drill the boom as shown in **Fig 61**. It is a good idea to use hardware with wing nuts to secure the element spreaders to the boom. After the boom is drilled, clean all the wood parts with denatured alcohol, sand them, and give them two coats of glossy polyurethane. After the polyurethane dries, wax all the wooden parts.

The boom to mast attachment is made next. Square the ends of a 6-foot section of clothes closet pole (a miter box is useful for this). Drill the center holes in both the boom

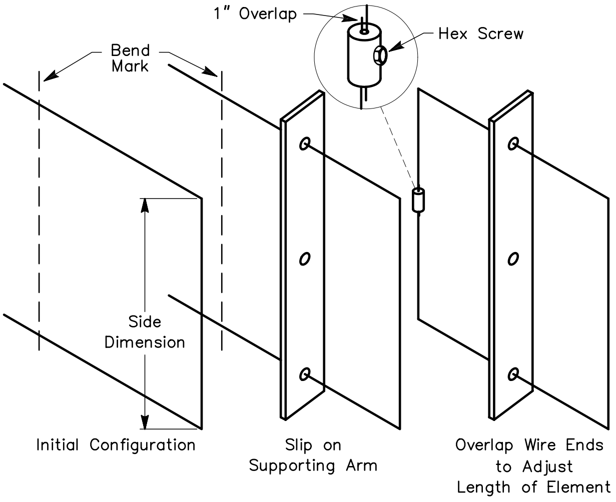

Fig 59—Illustration showing how the aluminum element wires are bent. The adjustment clamp and its location are also shown.

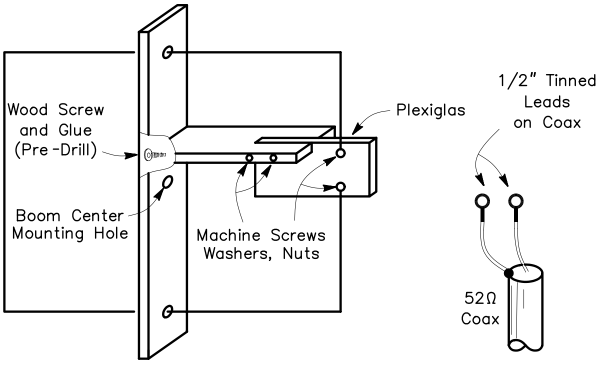
Wood Screw and Glue (Pre-Drill)
Plexiglas
1/2" Tinned Leads on Coax
Boom Center Mounting Hole
Machine Screws Washers, Nuts
52Ω Coax

Fig 60—Layout of the driven element of the 144-MHz quad. The leads of the coaxial cable should be stripped to ½ in. and solder lugs attached for easy connection and disconnection. See text regarding impedance at loop support points.

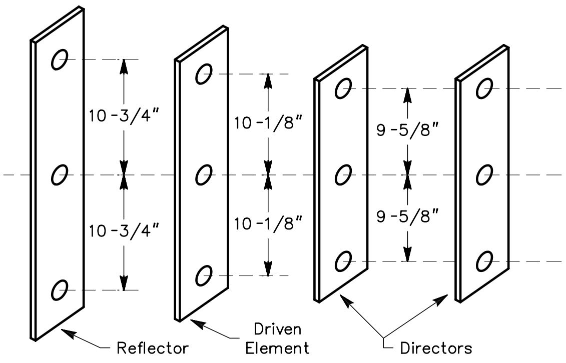

Fig 58—Dimensions for the pine element spreaders for the 144-MHz 4-element quad.

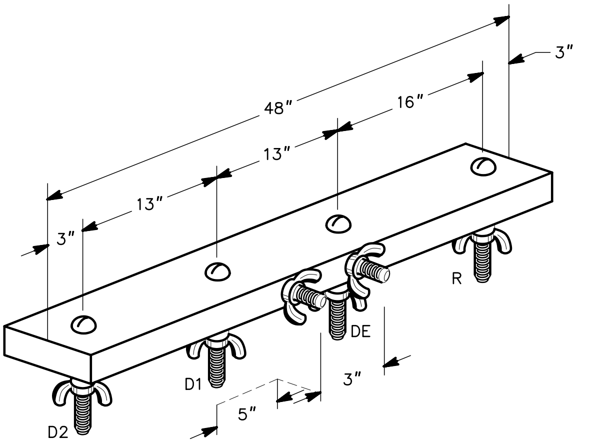

Fig 61—Detail of the boom showing hole center locations and boom to mast connection points.

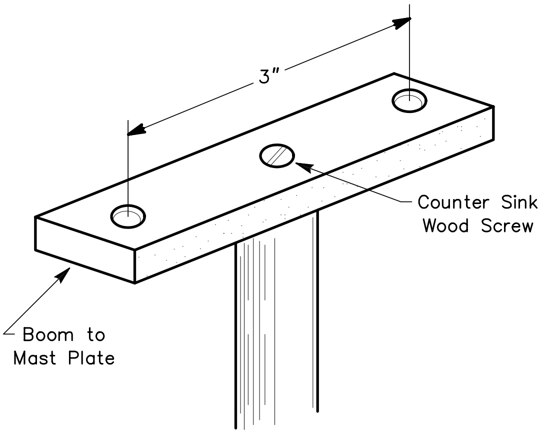

Fig 62—Boom to mast plate for the 144-MHz quad. The screw hole in the center of the plate should be countersunk so the wood screw attaching it to the mast does not interfere with the fit of the boom.

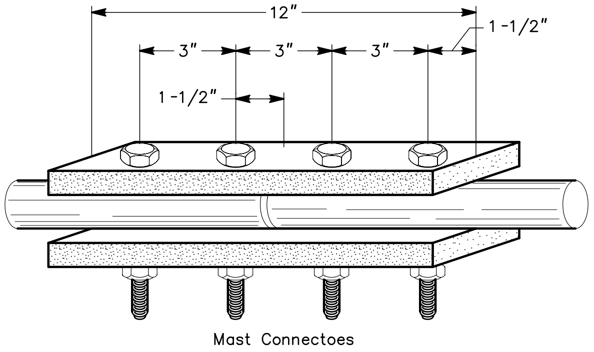

Fig 63—Mast coupling connector details for the portable quad. The plates should be drilled two at a time to ensure the holes line up.

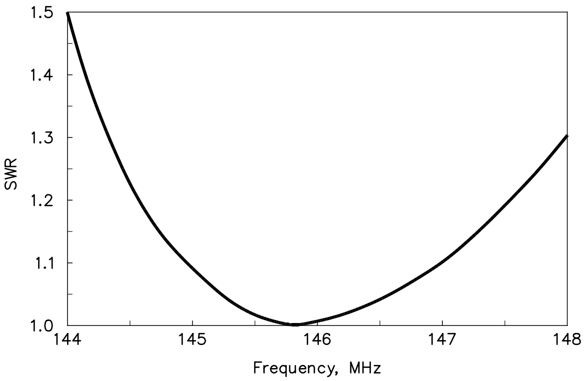

Fig 64—Typical SWR curve for the 144MHz portable quad. The large wire diameter and the quad design provide excellent bandwidth.

attachment piece and one end of the mast section (**Fig 62**). Make certain that the mast hole is smaller than the flat-head screw to be used to ensure a snug fit. Accurately drill the holes for attachment to the boom as shown in Fig 62.

Countersink the hole for the flat-head screw to provide a smooth surface for attachment to the boom. Apply epoxy cement to the surfaces and screw the boom attachment piece securely to the mast section. One 6 foot mast is used for attachment to the other mast sections.

Two additional 6-foot mast sections are prepared next. This brings the total mast height to 18 feet. It is important to square the ends of each pole so the mast stands straight when assembled. Mast-section connectors are made of pine as shown in **Fig 63**. Using $3^{1}/_{2} \times {}^{1}/_{4}$-inch hex bolts, washers, and nuts, sections may be attached as needed, for a total height of 6, 12 or 18 feet. Drill the holes in two connectors at a time. This ensures good alignment of the holes. A drill press is ideal for this job, but with care a hand drill can be used if necessary.

Line up two mast sections end to end, being careful that they are perfectly straight. Use the predrilled connectors to maintain pole straightness, and drill through the poles, one at a time. If good alignment is maintained, a straight 18-foot mast section can be made. Label the connectors and poles immediately so they are always assembled in the same order.

When assembling the antenna, install all the elements on the boom before attaching the feed line. Connect the coax to the screw connections on the driven element support plate and run the cable along the strut to the boom. From there, the cable should be routed directly to the mast and down. Assemble the mast sections to the desired height. The antenna provides good performance, and has a reasonable SWR curve over the entire 144 MHz band (**Fig 64**).

Building Quagi Antennas

The Quagi antenna was designed by Wayne Overbeck, N6NB. He first published information on this antenna in 1977 (see Bibliography). There are a few tricks to Quagi building, but nothing very difficult or complicated is involved. In fact, Overbeck mass produced as many as 16 in one day. **Tables 18** and **19** give the dimensions for Quagis for various frequencies up to 446 MHz.

For the designs of Tables 18 and 19, the boom is *wood* or any other nonconductor (such as, fiberglass or Plexiglas). If a metal boom is used, a new design and new element lengths will be required. Many VHF antenna builders go wrong by failing to follow this rule: If the original uses a metal boom, use the same size and shape metal boom when you duplicate it. If it calls for a wood boom, use a nonconductor. Many amateurs dislike wood booms, but in a salt air environment they outlast aluminum (and surely cost less). Varnish the boom for added protection.

The 144-MHz version is usually built on a 14 foot, 1 × 3 inch boom, with the boom tapered to 1 inch at both ends. Clear pine is best because of its light weight, but construction grade Douglas fir works well. At 222 MHz the boom is under 10 feet long, and most builders use 1 × 2 or (preferably) 3/4 × 1 1/4 inch pine molding stock. At 432 MHz, except for long-boom versions, the boom should be 1/2 inch thick or less. Most builders use strips of 1/2-inch exterior plywood for 432 MHz.

The quad elements are supported at the current maxima (the top and bottom, the latter beside the feed point) with Plexiglas or small strips of wood. See **Fig 65**. The quad elements are made of #12 copper wire, commonly used in house wiring. Some builders may elect to use #10 wire on 144 MHz and #14 on 432 MHz, although this changes the resonant frequency slightly. Solder a type N connector (an SO-239 is often used at 144 MHz) at the midpoint of the driven element bottom side, and close the reflector loop.

Table 19
432MHz, 15-Element, Long Boom Quagi Construction Data

Element Lengths, Inches	*Interelement Spacing, Inches*
R—28	R-DE7
DE—26 5/8	DE-D1—5 1/4
D1—11 3/4	D1-D2—11
D21—11/16	D2-D3—5 7/8
D3—11 5/8	D3-D4—8 3/4
D4—11 9/16	D4-D5—8 3/4
D5—11 1/2	D5-D6—8 3/4
D6—11 7/16	D6-D7—12
D7—11 3/8	D7-D8—12
D8—11 5/16	D8-D9—11 1/4
D9—11 5/16	D9-D10—11 1/2
D10—11 1/4	D10-D11—9 3/16
D11—11 3/16	D11-D12—12 3/8
D12—11 1/8	D12-D13—1-3/4
D13—11 1/16	

Boom: 1 × 2in. × 12-ft Douglas fir, tapered to 5/8 in. at both ends.
Driven element: #12 TW copper wire loop in square configuration, fed at bottom center with type N connector and 52-Ω coax.
Reflector: #12 TW copper wire loop, closed at bottom.
Directors: 1/8 in. rod passing through boom.

Table 18
Dimensions, Eight-Element Quagi

Element			*Frequency*		
Lengths	*144.5 MHz*	*147 MHz*	*222 MHz*	*432 MHz*	*446 MHz*
Reflector[1]	86 5/8"	85"	56 3/8"	28"	27 1/8"
Driven[2]	82"	80"	53½"	26 5/8"	25 7/8"
Directors	35 15/16" to 35" in 3/16" steps	35 5/16" to 34 3/8" in 3/16" steps	23 3/8" to 23¾" in 1/8" steps	11¾" to 11 7/16" in 1/16" steps	11 3/8" to 11 1/16" in 1/16" steps
Spacing					
R-DE	21"	20½"	13 5/8"	7"	6.8"
DE-D1	15¾"	15 3/8"	10¼"	5¼"	5.1"
D1-D2	33"	32½"	21½"	11"	10.7"
D2-D3	17½"	17 1/8"	11 3/8"	5.85"	5.68"
D3-D4	26.1"	25 5/8"	17"	8.73"	8.46"
D4-D5	26.1"	25 5/8"	17"	8.73"	8.46"
D5-D6	26.1"	25 5/8"	17"	8.73"	8.46"
Stacking Distance Between Bays					
	11'	10' 10"	7' 1½"	3'7"	3' 5 5/8"

[1] All #12 TW (electrical) wire, closed loops.
[2] All #12 TW wire loops, fed at bottom.

The directors are mounted through the boom. They can be made of almost any metal rod or wire of about $^1/_8$-inch diameter. Welding rod or aluminum clothesline wire works well if straight. (The designer uses $^1/_8$-inch stainless-steel rod obtained from an aircraft surplus store.)

A TV type U bolt mounts the antenna on a mast. A single machine screw, washers and a nut are used to secure the spreaders to the boom so the antenna can be quickly "flattened" for travel. In permanent installations two screws are recommended.

Construction Reminders

Based on the experiences of Quagi builders, the following hints are offered. First, remember that at 432 MHz even a $^1/_8$-inch measurement error results in performance deterioration. Cut the loops and elements as carefully as possible. No precision tools are needed, but accuracy is necessary. Also make sure to get the elements in the right order. The longest director goes closest to the driven element.

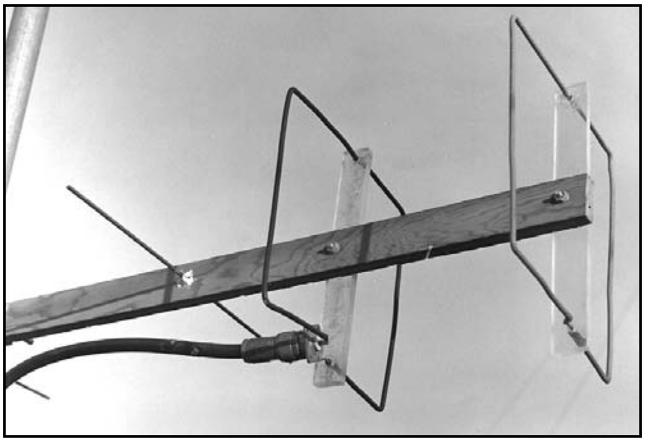

Fig 65—A close-up view of the feed method used on a 432-MHz Quagi. This arrangement produces a low SWR and gain in excess of 13 dBi with a 4-ft 10-in. boom! The same basic arrangement is used on lower frequencies, but wood may be substituted for the Plexiglas spreaders. The boom is ½-in. exterior plywood.

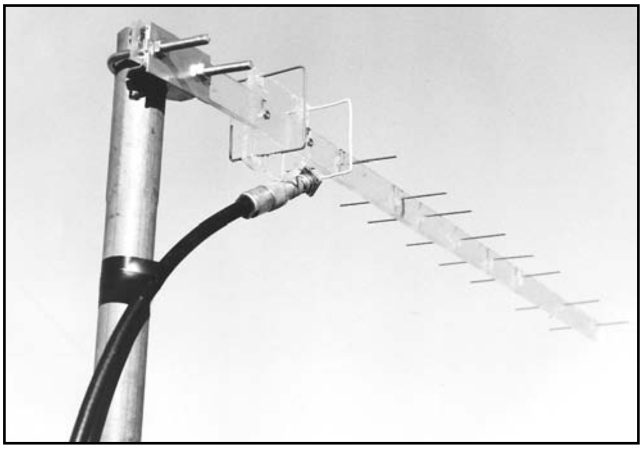

Finally, remember that a balanced antenna is being fed with an unbalanced line. Every balun the designer tried introduced more trouble in terms of losses than the feed imbalance caused. Some builders have tightly coiled several turns of the feed line near the feed point to limit line radiation. In any case, the feed line should be kept at right angles to the antenna. Run it from the driven element directly to the supporting mast and then up or down perpendicularly for best results.

QUAGIS FOR 1296 MHz

The Quagi principle has recently been extended to the 1296-MHz band, where good performance is extremely difficult to obtain from homemade conventional Yagis. **Fig 66** shows the construction and **Table 20** gives the design information for antennas with 10, 15 and 25 elements.

At 1296 MHz, even slight variations in design or building materials can cause substantial changes in performance. The 1296 MHz antennas described here work every time—but only if the same materials are used and the antennas are built *exactly* as described. This is not to discourage experimentation, but if modifications to these 1296-MHz antenna designs are contemplated, consider building one antenna as described here, so a reference is available against which variations can be compared.

The Quagis (and the cubical quad) are built on $^1/_4$-inch thick Plexiglas booms. The driven element and reflector (and also the directors in the case of the cubical quad) are made of insulated #18 AWG solid copper bell wire, available at hardware and electrical supply stores. Other types and sizes of wire work equally well, but the dimensions vary with the wire diameter. Even removing the insulation usually necessitates changing the loop lengths.

Quad loops are approximately square (**Fig 67**), although the shape is relatively uncritical. The element lengths, however, *are* critical. At 1296 MHz, variations

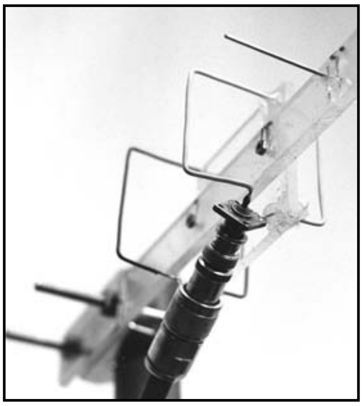

Fig 66—A view of the 10-element version of the 1296-MHz Quagi. It is mounted on a 30-in. Plexiglas boom with a 3 × 3-in. square of Plexiglas to support the driven element and reflector. Note how the driven element is attached to a standard UG-290 BNC connector. The elements are held in place with silicone sealing compound.

of $^1/_{16}$ inch alter the performance measurably, and a $^1/_8$ inch departure can cost several decibels of gain. The loop lengths given are *gross* lengths. Cut the wire to these lengths and then solder the two ends together. There is a $^1/_8$-inch overlap where the two ends of the reflector (and director) loops are joined, as shown in Fig 67.

The driven element is the most important of all. The #18 wire loop is soldered to a standard UG-290 chassis-mount BNC connector as shown in the photographs. This exact type of connector must be used to ensure uniformity in construction. Any substitution may alter the driven element electrical length. One end of the 9$^1/_4$ inch driven loop is pushed as far as it can go into the center pin, and is soldered in that position. The loop is then shaped and threaded through small holes drilled in the Plexiglas support. Finally, the other end is fed into one of the four mounting holes on the BNC connector and soldered. In most cases, the best SWR is obtained if the end of the wire just passes through the hole so it is flush with the opposite side of the connector flange.

Table 20

Dimensions, 1296-MHz Quagi Antennas

Note: All lengths are gross lengths. See text and photos for construction technique and recommended overlap at loop junctions. All loops are made of #18 AWG solid-covered copper bell wire. The Yagi type directors are $^1/_{16}$-in. brass brazing rod. See text for a discussion of director taper.

Feed: Direct with 52-Ω coaxial cable to UG-290 connector at driven element; run coax symmetrically to mast at rear of antenna.

Boom: 1$^1/_4$-in. thick Plexiglas, 30 in. long for 10-element quad or Quagi and 48 in. long for 15-element Quagi; 84 in. for 25-element Quagi.

10-Element Quagi for 1296 MHz

Element	*Length, Inches*	*Construction*	*Element*	*Interelement Spacing, In.*
Reflector	9.5625	Loop	R-DE	2.375
Driven	9.25	Loop	DE-D1	2.0
Director 1	3.91	Brass rod	D1-D2	3.67
Director 2	3.88	Brass rod	D2-D3	1.96
Director 3	3.86	Brass rod	D3-D4	2.92
Director 4	3.83	Brass rod	D4-D5	2.92
Director 5	3.80	Brass rod	D5-D6	2.92
Director 6	3.78	Brass rod	D6-D7	4.75
Director 7	3.75	Brass rod	D7-D8	3.94
Director 8	3.72	Brass rod		

15-Element Quagi for 1296 MHz

The first 10 elements are the same lengths as above, but the spacing from D6 to D7 is 4.0 in.; 07 to D8 is also 4.0 in.

Director 9	3.70	D8-D9	3.75
Director 10	3.67	D9-D10	3.83
Director 11	3.64	D10-D11	3.06
Director 12	3.62	D11-D12	4.125
Director 13	3.59	D12-D13	4.58

25-Element Quagi for 1296 MHz

The first 15 elements use the same element lengths and spacings as the 15-element model. The additional directors are evenly spaced at 3.0-in. intervals and taper in length successively by 0.02 in. per element. Thus, D23 is 3.39 in.

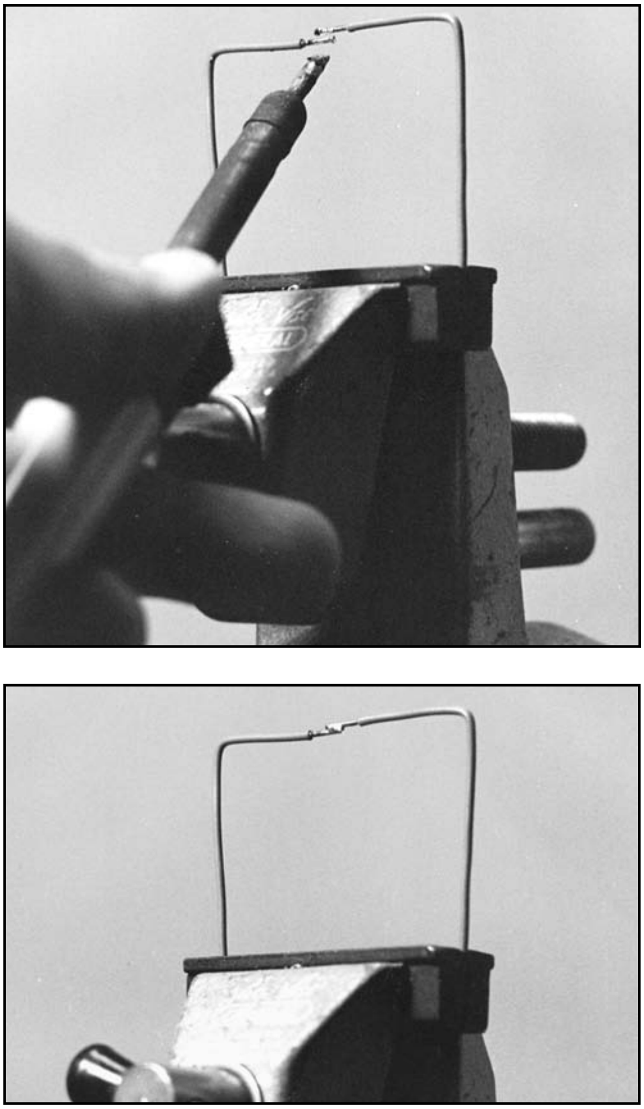

Fig 67—These photos show the construction method used for the 1296-MHz quad type parasitic elements. The two ends of the #18 bell wire are brought together with an overlap of $^1/_8$ in. and soldered.

Loop Yagis for 1296 MHz

Described here are loop Yagis for the 1296-MHz band. The loop Yagi fits into the quad family of antennas, as each element is a closed loop with a length of approximately 1 λ. Several versions are described, so the builder can choose the boom length and frequency coverage desired for the task at hand. Mike Walters, G3JVL, brought the original loop-Yagi design to the amateur community in the 1970s. Since then, many versions have been developed with different loop and boom dimensions. Chip Angle, N6CA, developed the antennas shown here.

Three sets of dimensions are given. Good performance can be expected if the dimensions are carefully followed. Check all dimensions before cutting or drilling anything. The 1296-MHz version is intended for weak-signal operation, while the 1270-MHz version is optimized for FM and mode L satellite work. The 1283-MHz antenna provides acceptable performance from 1280 to 1300 MHz.

These antennas have been built on 6- and 12-foot booms. Results of gain tests at VHF conferences and by individuals around the country show the gain of the 6-foot model to be about 18 dBi, while the 12-foot version provides about 20.5 dBi. Swept measurements indicate that gain is about 2 dB down from maximum gain at ±30 MHz from the design frequency. The SWR, however, deteriorates within a few megahertz on the low side of the design center frequency.

The Boom

The dimensions given here apply only to a ¾-inch OD boom. If a different boom size is used, the dimensions must be scaled accordingly. Many hardware stores

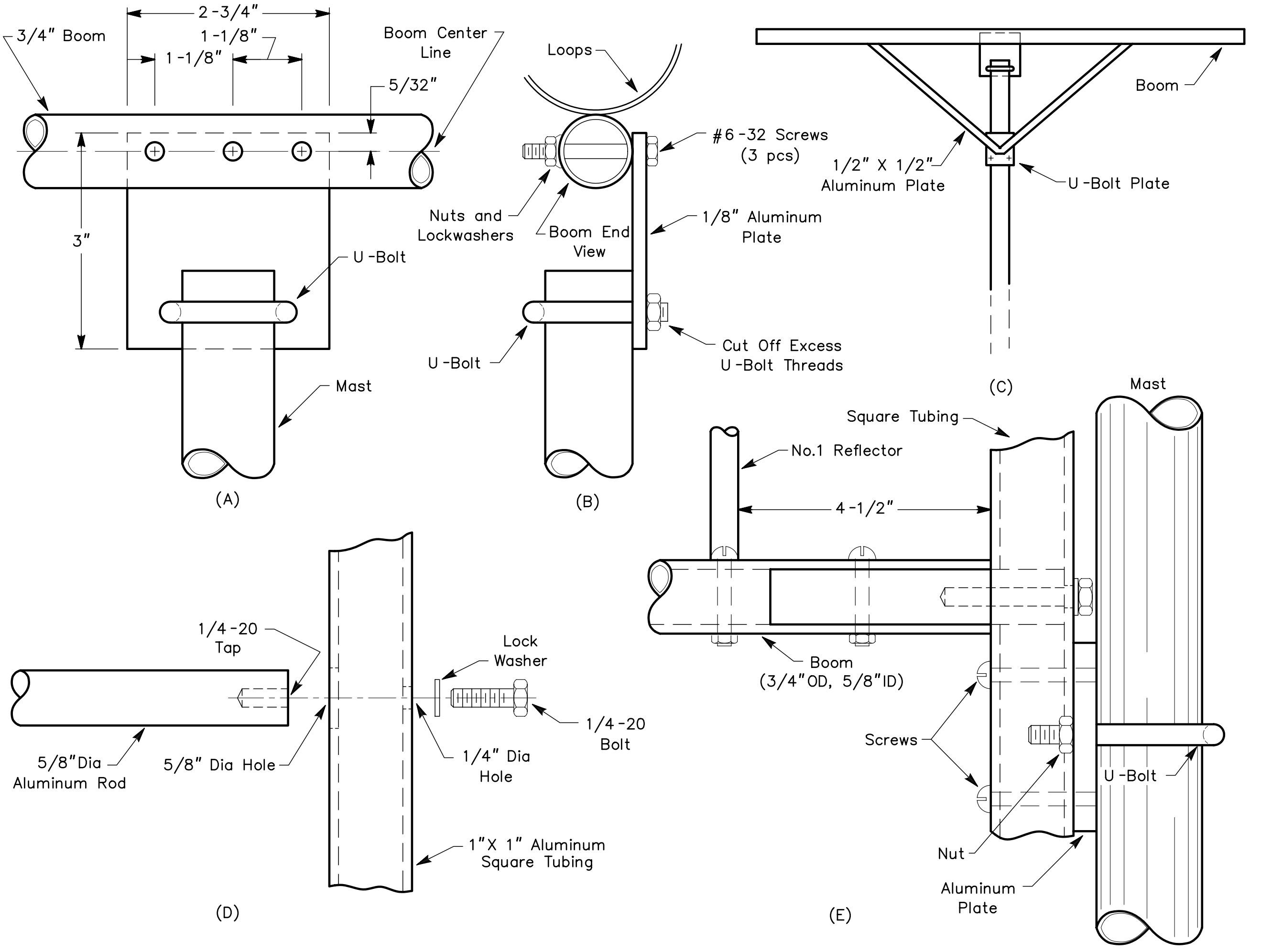
3/4" Boom
2-3/4"
1-1/8"
1-1/8"
5/32"
Boom Center Line
3"
U-Bolt
Mast
(A)
Loops
#6-32 Screws (3 pcs)
Nuts and Lockwashers
Boom End View
1/8" Aluminum Plate
U-Bolt
Cut Off Excess U-Bolt Threads
(B)
Boom
1/2" X 1/2" Aluminum Plate
U-Bolt Plate
(C)
1/4-20 Tap
Lock Washer
1/4-20 Bolt
5/8"Dia Aluminum Rod
5/8" Dia Hole
1/4" Dia Hole
1"X 1" Aluminum Square Tubing
(D)
Mast
Square Tubing
No.1 Reflector
4-1/2"
Boom (3/4"OD, 5/8"ID)
Screws
U-Bolt
Nut
Aluminum Plate
(E)

Fig 68—Loop Yagi boom-to-mast plate details are given at A. At B, the mounting of the antenna to the mast is detailed. A boom support for long antennas is shown at C. The arrangement shown in D and E may be used to rear-mount antennas up to 6 or 7 ft long.

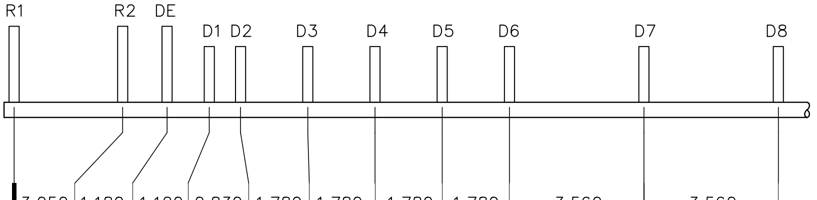

Freq.	Element Spacing (inches)									
1296	3.050	1.180	1.120	0.830	1.780	1.780	1.780	1.780	3.560	3.560
1283	3.090	1.193	1.135	0.841	1.804	1.804	1.804	1.804	3.607	3.607
1270	3.122	1.202	1.146	0.850	1.822	1.822	1.822	1.822	3.644	3.644

Spacing remains constant for all elements from D6 up

Fig 69—Boom drilling dimensions. These dimensions must be carefully followed and the same materials used if performance is to be optimum. Element spacings are the same for all directors after D6—use as many as necessary to fill the boom.

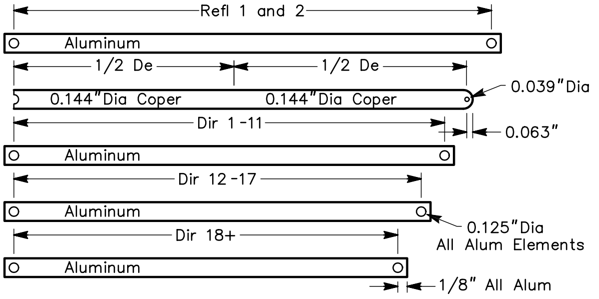

Freq	1270	1283	1296
Refl 1,2	9.929	9.829	9.700
1/2 De	4.752	4.704	4.643
Dir 1-11	8.445	8.359	8.250
Dir 12-17	8.189	8.106	8.000
Dir 18+	7.882	7.802	7.700

Element Lengths (Inches)
(Hole to Hole)

These Dimensions Apply Only to:
0.250" Element Width
0.0325" Element Thickness
0.750" Diaeter Boom

NOTE: All Dimensions are in Inches

Fig 70—Parasitic elements for the loop Yagi are made from aluminum sheet, the driven element from copper sheet. The dimensions given are for ¼-in. wide by 0.0325-in. thick elements only. Lengths specified are hole to hole distances; the holes are located $^1/_8$ in. from each element end.

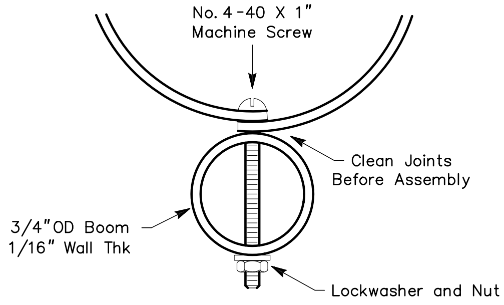

Fig 71—Element-to-boom mounting details.

carry aluminum tubing in 6- and 8-foot lengths, and that tubing is suitable for a short Yagi. If a 12-foot antenna is planned, find a piece of more rugged boom material, such as 6061-T6 grade aluminum. Do not use anodized tubing. The 12foot antenna must have additional boom support to minimize boom sag. The 6 foot version can be rear mounted. For rear mounting, allow 4$^1/_2$ inches of boom behind the last reflector to eliminate SWR effects from the support.

The antenna is attached to the mast with a gusset plate. This plate mounts at the boom center. See **Fig 68**. Drill the plate mounting holes perpendicular to the element mounting holes (assuming the antenna polarization is to be horizontal).

Elements are mounted to the boom with no. 4-40 machine screws, so a series of no. 33 (0.113inch) holes must be drilled along the center of the boom to accommodate this hardware. **Fig 69** shows the element spacings for different parts of the band. Dimensions should be followed as closely as possible.

Parasitic Elements

The reflectors and directors are cut from 0.032-inch thick aluminum sheet and are $^1/_4$ inch wide. **Fig 70** indicates the lengths for the various elements. These lengths apply only to elements cut from the specified material. For best results, the element strips should be cut with a shear. If the edges are left sharp, birds won't sit on the elements.

Drill the mounting holes as shown in Fig 70 after carefully marking their locations. After the holes are drilled, form each strap into a circle. This is easily done by wrapping the element around a round form. (A small juice can works well.)

Mount the loops to the boom with no. 4-40 × 1-inch machine screws, lock washers and nuts. See **Fig 71**. It is best to use only stainless steel or plated-brass hardware. Although the initial cost is higher than for ordinary plated-steel hardware, stainless or brass hardware will not rust and need replacement after a few years. Unless the antenna is painted, the hardware will definitely deteriorate.

Driven Element

The driven element is cut from 0.032-inch copper sheet and is $^1/_4$ inch wide. Drill three holes in the strap as detailed in Fig 69. Trim the ends as shown and form the strap into a loop similar to the other elements. This antenna is like a quad; if the loop is fed at the top or bottom, it is horizontally polarized.

Driven element mounting details are shown in **Fig 72**. A mounting fixture is made from a $^1/_4$-20 × 1$^1/_4$ inch brass bolt. File the bolt head to a thickness of $^1/_8$ inch. Bore a 0.144-inch (no. 27 drill) hole lengthwise through the center of the bolt. A piece of 0.141 inch semi-rigid Hardline (UT-141 or equivalent) mounts through this hole and is soldered to the driven loop feed point. The point at which the UT-141 passes through the copper loop and brass mounting fixture should be left unsoldered at this time to allow for matching adjustments when the antenna is completed, although the range of adjustment is not very large.

The UT-141 can be any convenient length. Attach the connector of your choice (preferably type N). Use a short piece of low-loss RG-8 size cable (or $^1/_2$-inch Hardline) for the run down the boom and mast to the main feed line. For best results, the main feed line should be the lowest loss 50-Ω cable obtainable. Good $^7/_8$-inch Hardline has 1.5 dB of loss per 100 feet and virtually eliminates the need for remote mounting of the transmit converter or amplifier.

Tuning the Driven Element

If the antenna is built carefully to the dimensions given, the SWR should be close to 1:1. Just to be sure, check the SWR if you have access to test equipment. Be sure the signal source is clean, however; wattmeters respond to "dirty" signals and can give erroneous readings. If problems are encountered, recheck all dimensions. If they look good, a minor improvement may be realized by changing the shape of the driven element. Slight bending of reflector 2 may also improve the SWR. When the desired match has been obtained, solder the point where the UT-141 jacket passes through the loop and brass bolt.

Tips for 1296-MHz Antenna Installations

Construction practices that are common on lower frequencies cannot be used on 1296 MHz. This is the most important reason why all who venture to these frequencies are not equally successful. First, when a proven design is used, copy it exactly¾don't change *anything*. This is especially true for antennas.

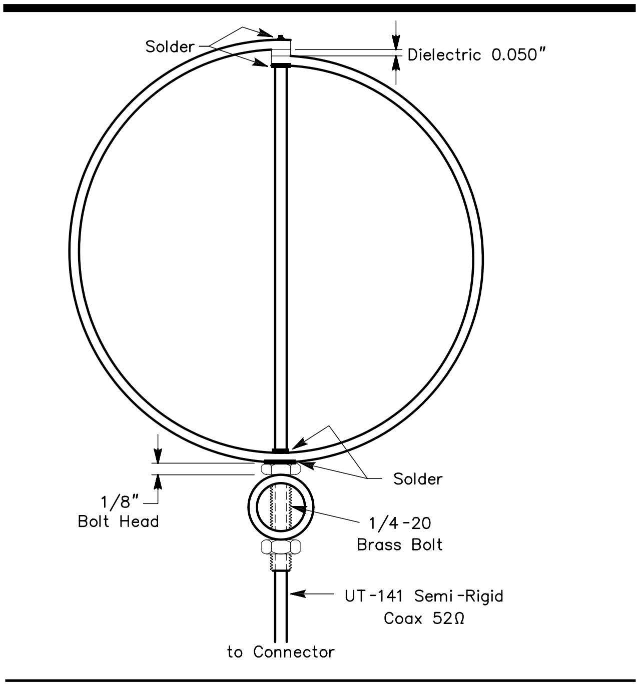

Fig 72—Driven-element details. See Fig 70 and the text for additional information.

Use the best feed line you can get. Here are some realistic measurements of common coaxial cables at 1296 MHz (loss per 100 feet).

RG-8, 213, 214: 11 dB
$^1/_2$ in. foam/copper Hardline: 4 dB
$^7/_8$ in. foam/copper Hardline: 1.5 dB

Mount the antennas to keep feed line losses to an absolute minimum. Antenna height is less important than keeping the line losses low. *Do not* allow the mast to pass through the elements, as is common on antennas for lower frequencies. Cut all U-bolts to the minimum length needed; $^1/_4$ λ at 1296 MHz is only a little over 2 inches. Avoid any unnecessary metal around the antenna.

Trough Reflectors for 432 and 1296 MHz

Dimensions are given in **Fig 73** for 432- and 1296-MHz trough reflectors. The gain to be expected is 16 dBi and 15 dBi, respectively. A very convenient arrangement, especially for portable work, is to use a metal hinge at each angle of the reflector. This permits the reflector to be folded flat for transit. It also permits experiments to be carried out with different apex angles.

A housing is required at the dipole center to prevent the entry of moisture and, in the case of the 432-MHz antenna, to support the dipole elements. The dipole may be moved in and out of the reflector to get either minimum SWR or, if this cannot be measured, maximum gain. If a two-stub tuner or other matching device is used, the dipole may be placed to give optimum gain and the matching device adjusted to give optimum match. In the case of the 1296-MHz antenna, the dipole length can be adjusted by means of the brass screws at the ends of the elements. Locking nuts are essential.

The reflector should be made of sheet aluminum for 1296 MHz, but can be constructed of wire mesh (with twists parallel to the dipole) for 432 MHz. To increase the gain by 3 dB, a pair of these arrays can be stacked so the reflectors are barely separated (to prevent the formation of a slot radiator by the edges). The radiating dipoles must then be fed in phase, and suitable feeding and matching must be arranged. A two-stub tuner can be used for matching either a single- or double-reflector system.

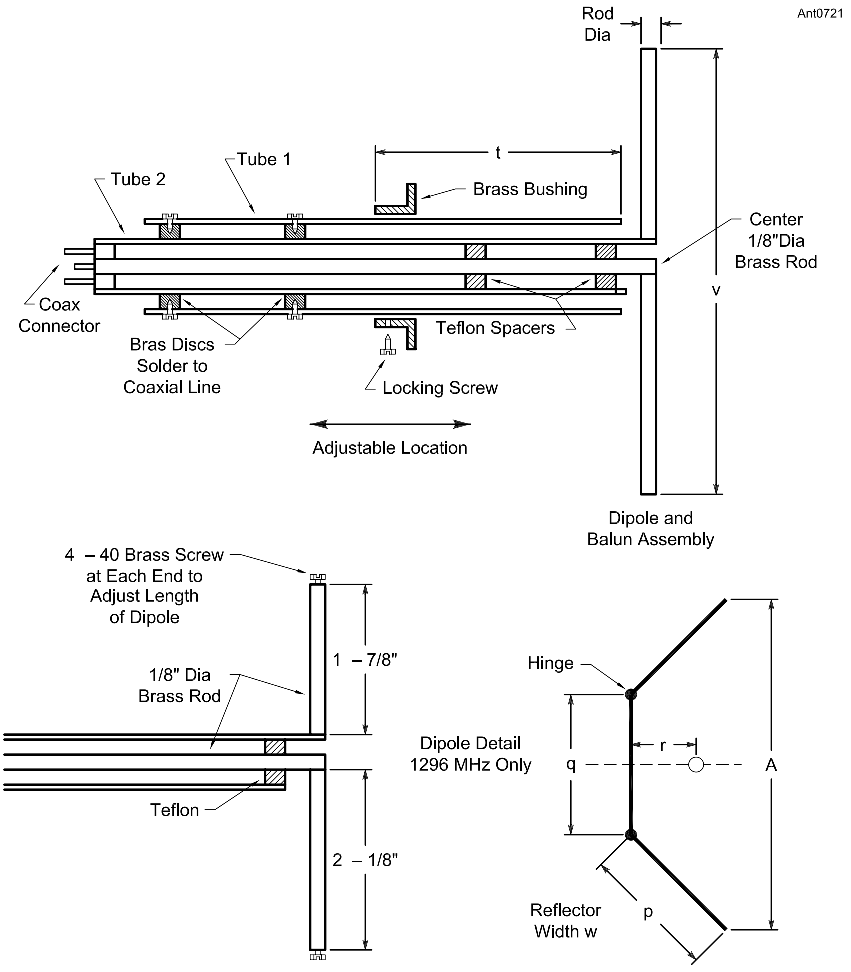
Rod Dia
Ant0721
Tube 1
Tube 2
t
Brass Bushing
Center 1/8"Dia Brass Rod
v
Coax Connector
Bras Discs Solder to Coaxial Line
Teflon Spacers
Locking Screw
Adjustable Location
Dipole and Balun Assembly
4 – 40 Brass Screw at Each End to Adjust Length of Dipole
1/8" Dia Brass Rod
1 – 7/8"
Teflon
2 – 1/8"
Hinge
Dipole Detail 1296 MHz Only
q
r
A
Reflector Width w
p

Dimensions in Inches

Band	A	p	q	r	w	t	u	v	Tube 1 Brass or Copper	Tube 2 Brass or Copper	Gain
432	120	100	50	7 – 1/2	60	6 – 1/2	1/4	12 – 3/4	1 – 1/2" ID 1/16" Wall	5/8" OD 1/16" Wall	16 dBi
1296	45	36	9 – 1/2	2 – 1/2	20	2 – 1/8	1/8	4 – 1/4	3/4" ID	3/8" OD 20 s.w.g.	15 dBi

Fig 73—Practical construction information for trough reflector antennas for 432 and 1296 MHz.

A Horn Antenna for 10 GHz

The horn antenna is the easiest antenna for the beginner on 10 GHz to construct. It can be made out of readily available flat sheet brass. Because it is inherently a broadband structure, minor constructional errors can be tolerated. The one drawback is that horn antennas become physically cumbersome at gains over about 25 dBi, but for most line-of-sight work this much gain is rarely necessary. This antenna was designed by Bob Atkins, KA1GT, and appeared in *QST* for April and May, 1987.

Horn antennas are usually fed by waveguide. When operating in its normal frequency range, waveguide propagation is in the TE_{10} mode. This means that the electric (E) field is across the short dimension of the guide and the magnetic (H) field is across the wide dimension. This is the reason for the E-plane and H-plane terminology shown in **Fig 74**.

There are many varieties of horn antennas. If the waveguide is flared out only in the H-plane, the horn is called an H-plane sectoral horn. Similarly, if the flare is only in the E-plane, an Eplane sectoral horn results. If the flare is in both planes, the antenna is called a pyramidal horn.

For a horn of any given aperture, directivity (gain along the axis) is maximum when the field distribution across the aperture is uniform in magnitude and phase. When the fields are not uniform, side lobes that reduce the directivity of the antenna are formed. To obtain a uniform distribution, the horn should be as long as possible with minimum flare angle. From a practical point of view, however, the horn should be as short as possible, so there is an obvious conflict between performance and convenience.

Fig 75 illustrates this problem. For a given flare angle and a given side length, there is a path-length difference from the apex of the horn to the center of the aperture (L), and from the apex of the horn to the edge of the aperture (L'). This causes a phase difference in the field across the aperture, which in turn causes formation of side lobes, degrading directivity (gain along the axis) of the antenna. If L is large this difference is small, and the field is almost uniform. As L decreases however, the phase difference increases and directivity suffers. An optimum (shortest possible) horn is constructed so that this phase difference is the maximum allowable before side lobes become excessive and axial gain markedly decreases.

The magnitude of this permissible phase difference is different for E-plane and H-plane horns. For the E-plane horn, the field intensity is quite constant across the aperture. For the H-plane horn, the field tapers to zero at the edge. Consequently, the phase difference at the edge of the aperture in the E-plane horn is more critical and should be held to less than 90° ($^1/_4$ λ). In an H-plane horn, the allowable phase difference is 144° (0.4 λ). If the aperture of a pyramidal horn exceeds one wavelength in both planes, the Eplane and Hplane patterns are essentially independent and can be analyzed separately.

The usual direction for orienting the waveguide feed is with the broad face horizontal, giving vertical polarization. If this is the case, the H-plane sectoral horn has a narrow horizontal beamwidth and a very wide vertical beamwidth. This is not a very useful beam pattern for most amateur applications. The E-plane sectoral horn has a narrow vertical beamwidth and a wide horizontal beamwidth. Such a radiation pattern could be useful in a beacon system where wide coverage is desired.

The most useful form of the horn for general applications is the optimum pyramidal horn. In this configuration the two beamwidths are almost the same. The E-plane (vertical) beamwidth is slightly less than the H-plane (horizontal), and also has greater side lobe intensity.

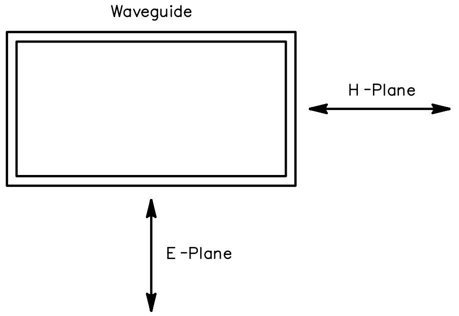

Fig 74—10-GHz antennas are usually fed with waveguide. See text for a discussion of waveguide propagation characteristics.

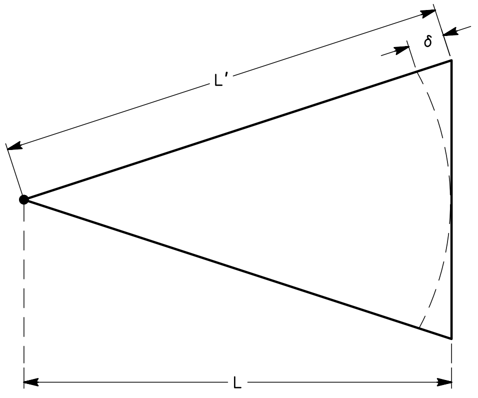

Fig 75—The path-length (phase) difference between the center and edge of a horn antenna is δ.

Building the Antenna

A 10-GHz pyramidal horn with 18.5 dBi gain is shown in **Fig 76**. The first design parameter is usually the required gain, or the maximum antenna size. These are of course related, and the relationships can be approximated by the following:

L = H-plane length (λ) = 0.0654 × gain (Eq 1)
A = H-plane aperture (λ) = 0.0443 × gain (Eq 2)
B = E-plane aperture (λ) = 0.81 A (Eq 3)

where
gain is expressed as a *ratio*; 20 dBi gain = 100
L, A and B are dimensions shown in **Fig 77.**

From these equations, the dimensions for a 20-dBi gain horn for 10.368 GHz can be determined. One wavelength at 10.368 GHz is 1.138 inches. The length (L) of such a horn is 0.0654 × 100 = 6.54 λ. At 10.368 GHz, this is 7.44 inches. The corresponding H-plane aperture (A) is 4.43 λ (5.04 inches), and the E-plane aperture (B), 4.08 inches.

The easiest way to make such a horn is to cut pieces from brass sheet stock and solder them together. Fig 77 shows the dimensions of the triangular pieces for the sides and a square piece for the waveguide flange. (A standard commercial waveguide flange could also be used.) Because the E-plane and H-plane apertures are different, the horn opening is not square. Sheet thickness is unimportant; 0.02 to 0.03 inch works well. Brass sheet is often available from hardware or hobby shops.

Note that the triangular pieces are trimmed at the apex to fit the waveguide aperture (0.9 × 0.4 inch). This necessitates that the length, from base to apex, of the smaller triangle (side B) is shorter than that of the larger (side A). Note that the length, S, of the two different sides of the horn must be the same if the horn is to fit together! For such a simple looking object, getting the parts to fit together properly requires careful fabrication.

The dimensions of the sides can be calculated with simple geometry, but it is easier to draw out templates on a sheet of cardboard first. The templates can be used to build a mock antenna to make sure everything fits together properly before cutting the sheet brass.

First, mark out the larger triangle (side A) on cardboard. Determine at what point its width is 0.9 inch and draw a line parallel to the base as shown in Fig 77. Measure the length of the side S; this is also the length of the sides of the smaller (side B) pieces.

Mark out the shape of the smaller pieces by first drawing a line of length B and then constructing a second line of length S. One end of line S is an end of line B, and the other is 0.2 inch above a line perpendicular to the center of line B as shown in Fig 76. (This procedure is much more easily followed than described.) These smaller pieces are made slightly oversize (shaded area in Fig 77) so you can construct the horn with solder seams on the outside of the horn during assembly.

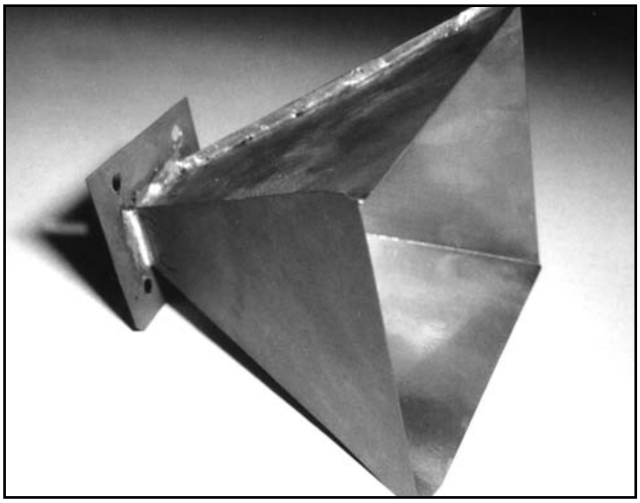

Fig 76—This pyramidal horn has 18.5 dBi gain at 10 GHz. Construction details are given in the text.

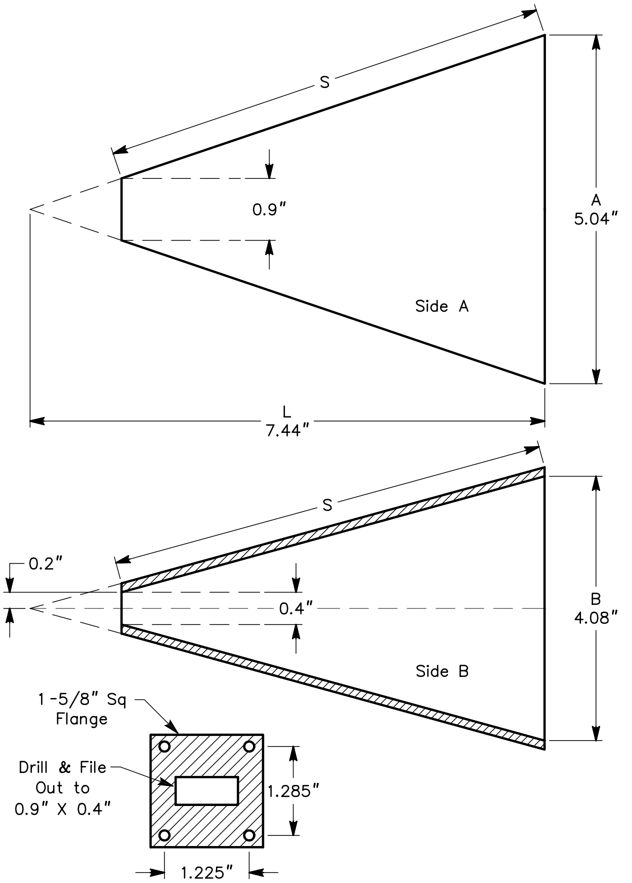

Fig 77—Dimensions of the brass pieces used to make the 10-GHz horn antenna. Construction requires two of each of the triangular pieces (side A and side B).

Cut out two cardboard pieces for side A and two for side B and tape them together in the shape of the horn. The aperture at the waveguide end should measure 0.9 × 0.4 inch and the aperture at the other end should measure 5.04 × 4.08 inches.

If these dimensions are correct, use the cardboard templates to mark out pieces of brass sheet. The brass sheet should be cut with a bench shear if one is available, because scissors type shears tend to bend the metal. Jig the pieces together and solder them on the *outside* of the seams. It is important to keep both solder and rosin from contaminating the inside of the horn; they can absorb RF and reduce gain at these frequencies.

Assembly is shown in **Fig 78**. When the horn is completed, it can be soldered to a standard waveguide flange, or one cut out of sheet metal as shown in Fig 77. The transition between the flange and the horn must be smooth. This antenna provides an excellent performance-to-cost ratio (about 20 dBi gain for about five dollars in parts).

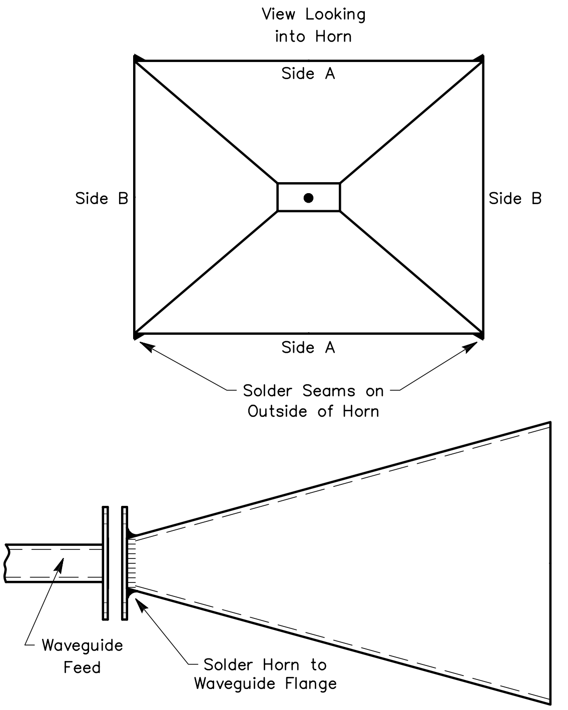
View Looking into Horn
Side A
Side B
Side B
Side A
Solder Seams on Outside of Horn
Waveguide Feed
Solder Horn to Waveguide Flange

Fig 78—Assembly of the 10-GHz horn antenna.

Periscope Antenna Systems

One problem common to all who use microwaves is that of mounting an antenna at the maximum possible height while trying to minimize feed-line losses. The higher the frequency, the more severe this problem becomes, as feeder losses increase with frequency. Because parabolic dish reflectors are most often used on the higher bands, there is also the difficulty of waterproofing feeds (particularly waveguide feeds). Inaccessibility of the dish is also a problem when changing bands. Unless the tower is climbed every time and the feed changed, there must be a feed for each band mounted on the dish. One way around these problems is to use a periscope antenna system (sometimes called a "flyswatter antenna").

The material in this section was prepared by Bob Atkins, KA1GT, and appeared in *QST* for January and February 1984. **Fig 79** shows a schematic representation of a periscope antenna system. A plane reflector is mounted at the top of a rotating tower at an angle of 45°. This reflector can be elliptical with a major to minor axis ratio of 1.41, or rectangular. At the base of the tower is mounted a dish or other type of antenna such as a Yagi, pointing straight up. The advantage of such a system is that the feed antenna can be changed and worked on easily. Additionally, with a correct choice of reflector size, dish size, and dish to reflector spacing, feed losses can be made small, increasing the effective system gain. In fact, for some particular system configurations, the gain of the overall system can be greater than that of the feed antenna alone.

Gain of a Periscope System

Fig 80 shows the relationship between the effective gain of the antenna system and the distance between the reflector and feed antenna for an elliptical reflector. At first sight, it is not at all obvious how the antenna system can have a higher gain than the feed alone. The reason lies in the fact that, depending on the feed to reflector spacing, the reflector may be in the near field (Fresnel) region of the antenna, the far field (Fraunhöffer) region, or the transition region between the two.

In the far field region, the gain is proportional to the reflector area and inversely proportional to the distance between the feed and reflector. In the near field region, seemingly strange things can happen, such as decreasing gain with decreasing feed to reflector separation. The reason for this gain decrease is that, although the reflec-

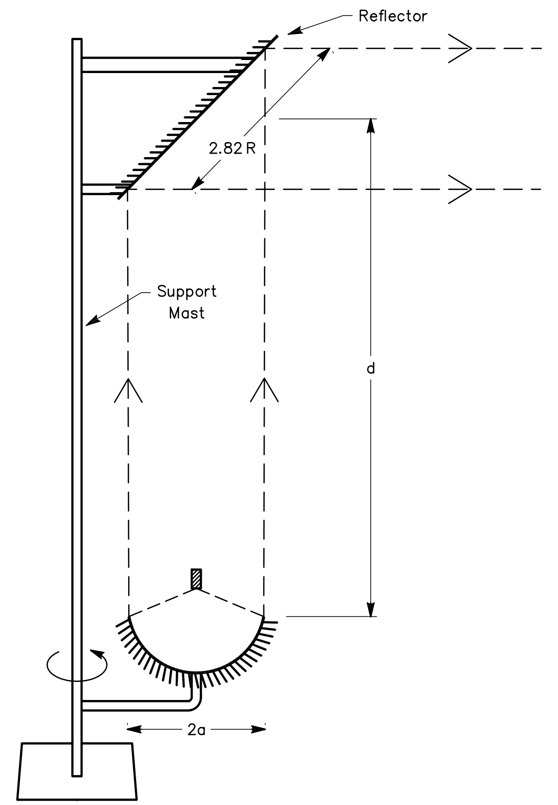

Fig 79—The basic periscope antenna. This design makes it easy to adjust the feed antenna.

tor is intercepting more of the energy radiated by the feed, it does not all contribute in phase at a distant point, and so the gain decreases.

In practice, rectangular reflectors are more common than elliptical. A rectangular reflector with sides equal in length to the major and minor axes of the ellipse will, in fact, normally give a slight gain increase. In the far field region, the gain will be proportional to the area of the reflector. To use Fig 80 with a rectangular reflector, R^2 may be replaced by A / π, where A is the projected area of the reflector. The antenna pattern depends in a complicated way on the system parameters (spacing and size of the elements), but **Table 21** gives an approximation of what to expect. R is the radius of the projected circular area of the elliptical reflector (equal to the minor axis radius), and b is the length of the side of the projected square area of the rectangular reflector (equal to the length of the short side of the rectangle).

For those wishing a rigorous mathematical analysis of this type of antenna system, several references are given in the Bibliography at the end of this chapter.

Mechanical Considerations

There are some problems with the physical construction of a periscope antenna system. Since the antenna gain of a microwave system is high and, hence, its beamwidth narrow, the reflector must be accurately aligned. If the reflector does not produce a beam that is horizontal, the

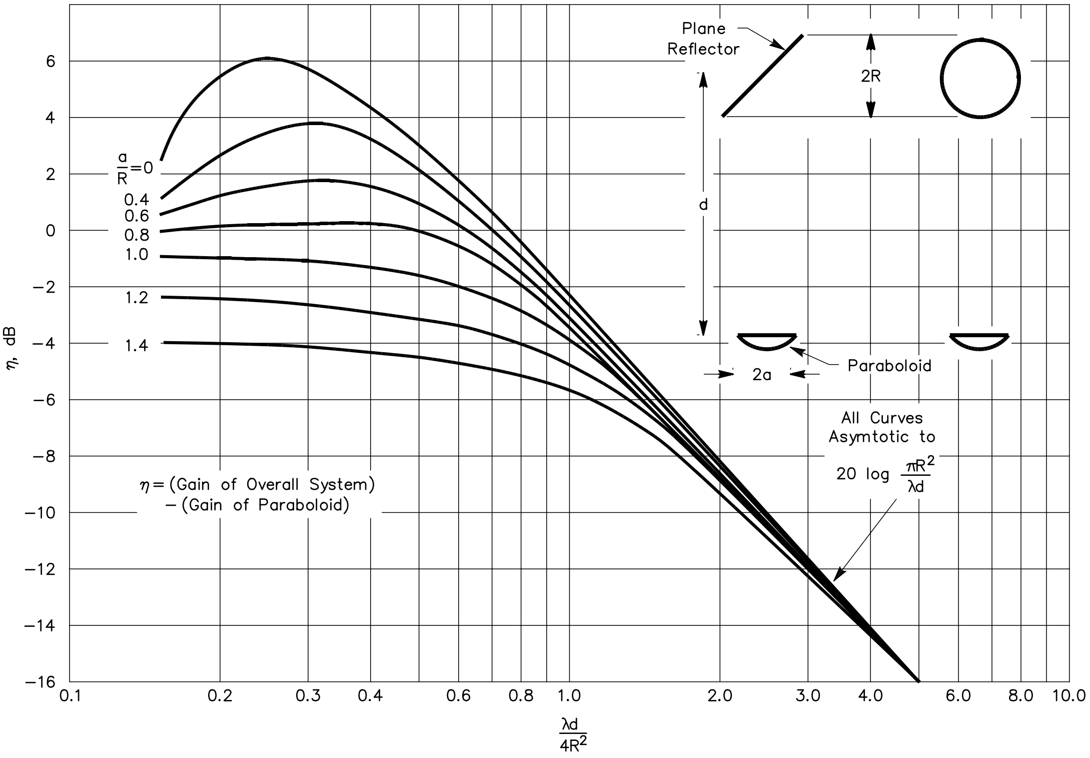

Fig 80—Gain of a periscope antenna using a plane elliptical reflector (after Jasik—see Bibliography).

Table 21
Radiation Patterns of Periscope Antenna Systems

	Elliptical Reflector	*Rectangular Reflector*
3-dB beamwidth, degrees	60 $\lambda/2R$	52 λ/b
6-dB beamwidth, degrees	82 $\lambda/2R$	68 λ/b
First minimum, degrees from axis	73 $\lambda/2R$	58 λ/b
First maximum, degrees from axis	95 $\lambda/2R$	84 λ/b
Second minimum, degrees from axis	130 $\lambda/2R$	116 λ/b
Second maximum, degrees from axis	156 $\lambda/2R$	142 λ/b
Third minimum, degrees from axis	185 $\lambda/2R$	174 λ/b

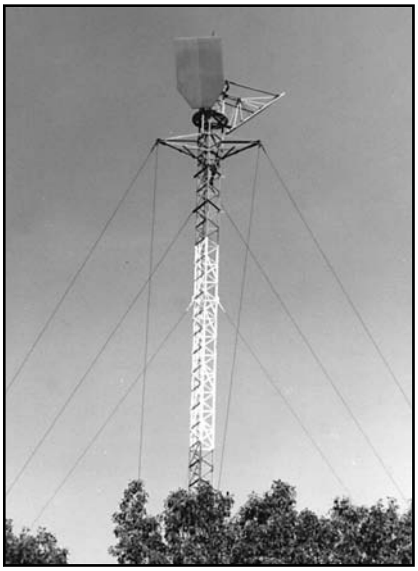

Fig 81—Commercial periscope antennas, such as this one, are often used for point-to-point communication.

useful gain of the system will be reduced. From the geometry of the system, an angular misalignment of the reflector of X degrees in the vertical plane will result in an angular misalignment of 2X degrees in the vertical alignment of the antenna system pattern. Thus, for a dish pointing straight up (the usual case), the reflector must be at an angle of 45° to the vertical and should not fluctuate from factors such as wind loading.

The reflector itself should be flat to better than $^{1}/_{10}$ λ for the frequency in use. It may be made of mesh, provided that the holes in the mesh are also less than $^{1}/_{10}$ λ in diameter. A second problem is getting the support mast to rotate about a truly vertical axis. If the mast is not vertical, the resulting beam will swing up and down from the horizontal as the system is rotated, and the effective gain at the horizon will fluctuate. Despite these problems, amateurs have used periscope antennas successfully on the bands through 10 GHz. Periscope antennas are used frequently in commercial service, though usually for point-to-point transmission. Such a commercial system is shown in **Fig 81**.

Circular polarization is not often used for terrestrial work, but if it is used with a periscope system there is an important point to remember. The circularity sense changes when the signal is reflected. Thus, for right hand circularity with a periscope antenna system, the feed arrangement on the ground should produce left hand circularity. It should also be mentioned that it is possible (though more difficult for amateurs) to construct a periscope antenna system using a parabolically curved reflector. The antenna system can then be regarded as an offset fed parabola. More gain is available from such a system at the added complexity of constructing a parabolically curved reflector, accurate to $^{1}/_{10}$ λ.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2, Antenna Fundamentals.

B. Atkins, "Periscope Antenna Systems," The New Frontier, *QST*, Jan and Feb 1984.

B. Atkins, "Horn Antennas for 10 GHz," The New Frontier, *QST*, Apr and May 1987.

J. Drexler, "An Experimental Study of a Microwave Periscope," *Proc. IRE*, Correspondence, Vol 42, Jun 1954, p 1022.

D. Evans and G. Jessop, *VHF-UHF Manual*, 3rd ed. (London: RSGB), 1976.

N. Foot, "WA9HUV 12foot Dish for 432 and 1296 MHz," The World Above 50 Mc., *QST*, Jun 1971, pp 98-101, 107.

N. Foot, "Cylindrical Feed Horn for Parabolic Reflectors," *Ham Radio*, May 1976.

G. Gobau, "Single-Conductor Surface-Wave Transmission Lines," *Proc. IRE*, Vol 39, Jun 1951, pp 619-624; also see *Journal of Applied Physics*, Vol 21 (1950), pp 1119-1128.

R. E. Greenquist and A. J. Orlando, "An Analysis of Passive Reflector Antenna Systems," *Proc. IRE*, Vol 42, Jul 1954, pp 1173-1178.

G. A. Hatherell, "Putting the G Line to Work," *QST*, Jun 1974, pp 11-15, 152, 154, 156.

D. L. Hilliard, "A 902MHz Loop Yagi Antenna," *QST*, Nov 1985, pp 30-32.

W. C. Jakes, Jr., "A Theoretical Study of an Antenna-Reflector Problem," *Proc. IRE*, Vol 41, Feb 1953, pp 272-274.

H. Jasik, *Antenna Engineering Handbook*, 1st ed. (New York: McGraw-Hill, 1961).

R. T. Knadle, “UHF Antenna Ratiometry,” *QST*, Feb 1976, pp 22-25.

T. Moreno, *Microwave Transmission Design Data* (New York: McGraw- Hill, 1948).

W. Overbeck, “The VHF Quagi,” *QST*, Apr 1977, pp 11-14.

W. Overbeck, “The Long-Boom Quagi,” *QST*, Feb 1978, pp 20-21.

W. Overbeck, “Reproducible Quagi Antennas for 1296 MHz,” *QST*, Aug 1981, pp 11-15.

G. Southworth, *Principles and Applications of Waveguide Transmission* (New York: D. Van Nostrand Co, 1950).

P. P. Viezbicke, “Yagi Antenna Design,” *NBS Technical Note 688* (U. S. Dept. of Commerce/National Bureau of Standards, Boulder, CO), Dec 1976.

D. Vilardi, “Easily Constructed Antennas for 1296 MHz,” *QST*, Jun 1969.

D. Vilardi, “Simple and Efficient Feed for Parabolic Antennas,” *QST*, Mar 1973.

Radio Communication Handbook, 5th ed. (London: RSGB, 1976).

Chapter 19

Antenna Systems for Space Communications

When we consider amateur space communications, we usually think about two basic modes: satellite and earth-moon-earth (EME—also referred to as *moonbounce*). At their essence, both modes communicate using one of the Earth's satellites—our natural satellite (the Moon) or one of a variety of manmade satellites.

There are two main differences between these satellites. The first is one of distance. The Moon is about 250,000 miles from Earth, while man-made satellites can be as far as 36,000 miles away. This 7:1 difference in distance makes a huge difference in the signals that arrive at the satellite, since transmission loss varies as the square of the distance. In other words, the signal arriving at the Moon is 20 dB weaker than that arriving at a geo-synchronous satellite 25,000 miles high, due to distance alone.

The second difference between the Moon and a man-made satellite is that the Moon is a *passive reflector*—and not a very good one at that, since it has a craggy and rather irregular surface, at least when compared to a flat mirror-like surface that would make an ideal reflector. Signals scattered by the Moon's irregular surface are thus weaker than for better reflecting surfaces. By comparison, a man-made satellite is an *active* system, where the satellite receives the signal coming from Earth, amplifies it and then retransmits the signal (usually at a different frequency) using a high-gain antenna. Think of a satellite as an ideal reflector, with gain.

The net result of these differences between a man-made satellite and the Earth's natural satellite is that moonbounce (EME) operation challenges the station builder considerably more than satellite operation, particularly in the area of antennas. Successful EME requires high transmitting power, superb receiver sensitivity and excellent operators capable of pulling weak signals out of the noise. This chapter will first explore antennas suitable for satellite operations and then describe techniques needed for EME work.

Common Ground

There are areas of commonality between satellite and EME antenna requirements, of course. Both require consideration of the effects of polarization and elevation angle, along with the azimuth directions of transmitted and received signals.

On the HF bands, signal polarization is generally of little concern, since the original polarization sense is lost after the signal passes through the ionosphere. At HF, vertical antennas receive sky-wave signals emanating from horizontal antennas, and vice versa. It is not beneficial to provide a means of varying the polarization at HF. With satellite communications, however, because of polarization changes, a signal that would disappear into the noise on one antenna may be S9 on one that is not sensitive to polarization direction. Elevation angle is also important from the standpoint of tracking and avoiding indiscriminate ground reflections that may cause nulls in signal strength.

Antennas for Satellite Work

We have amateur satellites providing links from 2 meters and up, and these provide opportunities to use antennas of many types—from the very simple to some pretty complex ones. This section was written by Dick Jansson, WD4FAB. It covers descriptions of a wide range of satellite antennas and points operators to source material for construction of many of them.

Antennas for LEO Satellites

Antenna design and construction requirements for use with Amateur satellites vary from low-gain antennas for low-earth-orbit (LEO) satellites to higher-gain antennas for the high-altitude elliptical-orbit satellites. You can operate the FM LEO satellites with a basic dual-band VHF/UHF FM transceiver or even a good FM H-T, as some amateurs have managed. Assuming that the transceiver is reasonably sensitive, you can even use a good "rubber duck" antenna. Some amateurs manage to work the FM birds with H-Ts and a multi-element directional antenna such as the popular Arrow Antenna, **Fig 1**. Of course, this means they must aim their antennas at the satellites, even as they cross overhead.

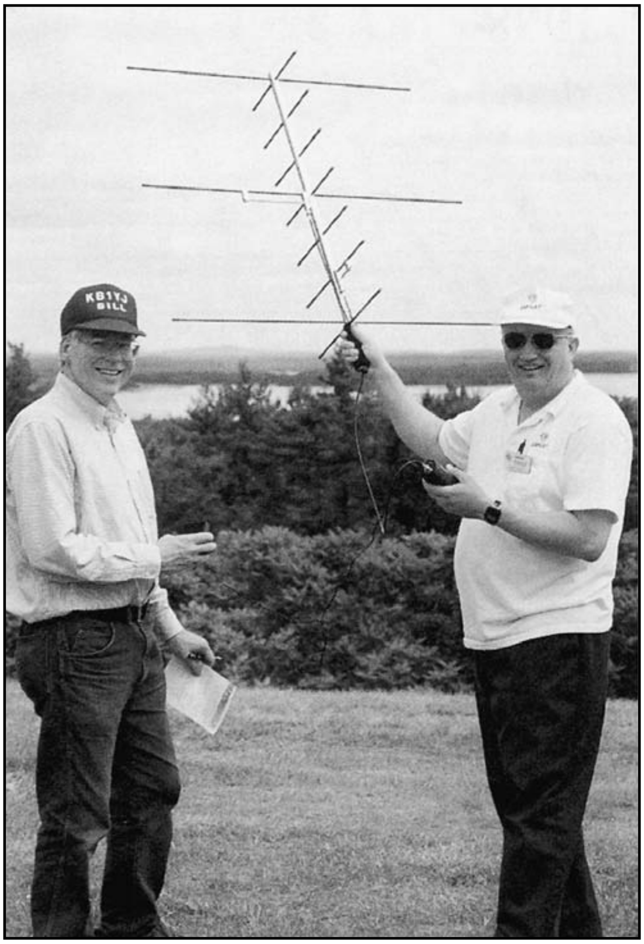

Fig 1—The hand-held "Arrow" gain antenna is popular for LEO FM operations. (*Photo courtesy* The AMSAT Journal, *Sept/Oct 1998.*)

High-quality omnidirectional antennas for LEO service come in quite a number of forms and shapes. M^2 Enterprises has their EB-144 and EB-432 Eggbeater antennas, which have proven to be very useful and do not require any rotators for control. See **Fig 2**. The turnstile-over-reflector antenna has been around for a long time, as shown in **Fig 3**. Other operators have done well using low-gain Yagi antennas, such as those shown in **Fig 4**.

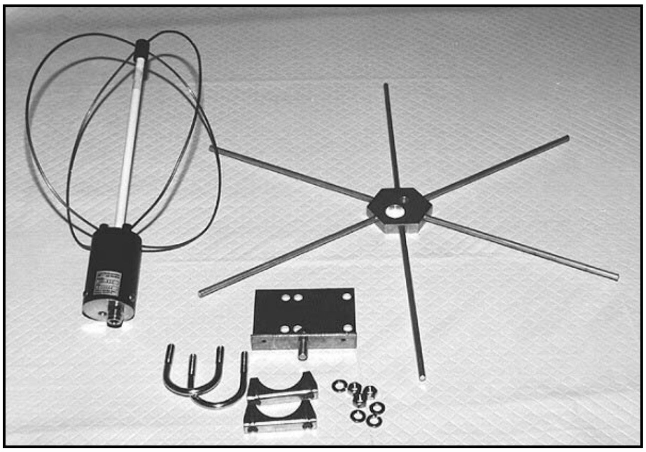

Fig 2—Eggbeater antennas are popular for base station LEO satellite operations. This EB-432 eggbeater antenna for 70 cm is small enough to put in an attic. Antenna gain pattern is helped with the radials placed below the antenna.

For even better performance, at the modest cost of a single, simple TV antenna rotator, check out the fixed-elevation *Texas Potato Masher* antenna by K5OE, **Fig 5.** This antenna provides a dual-band solution for medium-gain directional antennas for LEO satellites. This is a considerable improvement over omnidirectional antennas and does not require an elevation rotator for good performance.

There are still two LEO satellites that work on the 10-meter band, RS-15 and the newly resurrected AO-7. Both have 10-meter downlinks in the range of 29.3 to 29.5 MHz. Low-gain 10-meter antennas, such as dipoles or long-wire antennas, are used to receive these satellites.

Antennas for High-Altitude Satellites

The high-altitude, Phase-3 satellites, such as AO-10 (and the late AO-13), have been around for quite a number of years. The greater distances to the Phase-3 satellites mean that more transmitted power is needed to access them and weaker signals are received on the ground. Successful stations usually require ground-station antennas with significant gain (12 dBi or more), such as a set of high-gain Yagi antennas. See **Fig 6**. Note the use of two Yagi antennas mounted on each boom to provide *circular polarization*, usually referred to as CP.

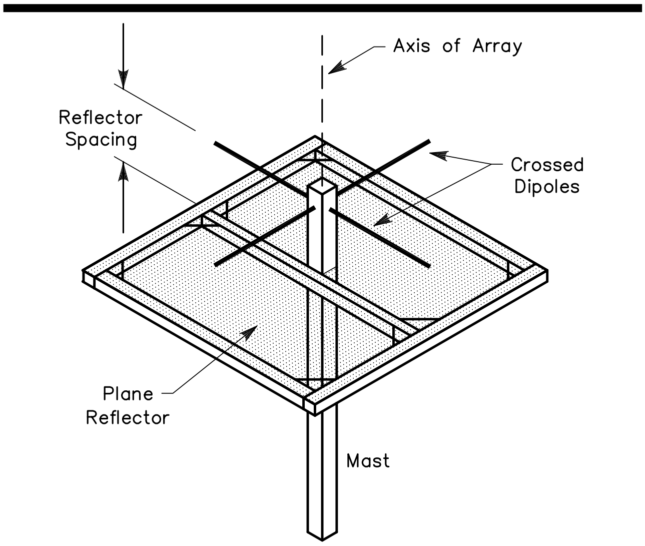

Fig 3—The Turnstile Over Reflector antenna has served well for LEO satellite service for a number of years.

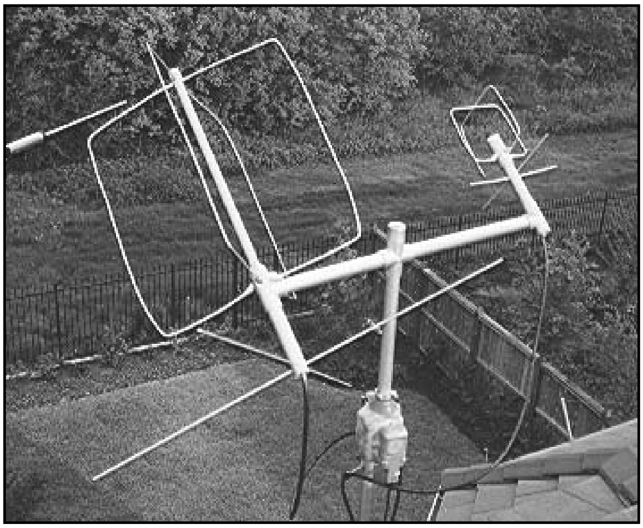

Fig 5—Jerry Brown, K5OE, uses his Texas Potato Masher antennas to work LEO satellites.

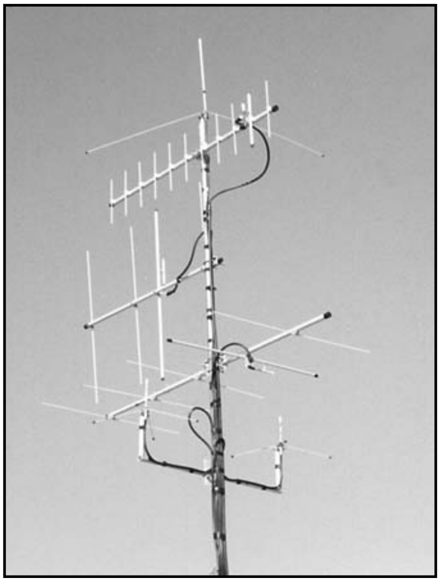

Fig 4—Simple ground plane and Yagi antennas can be used for LEO satellite contacts.

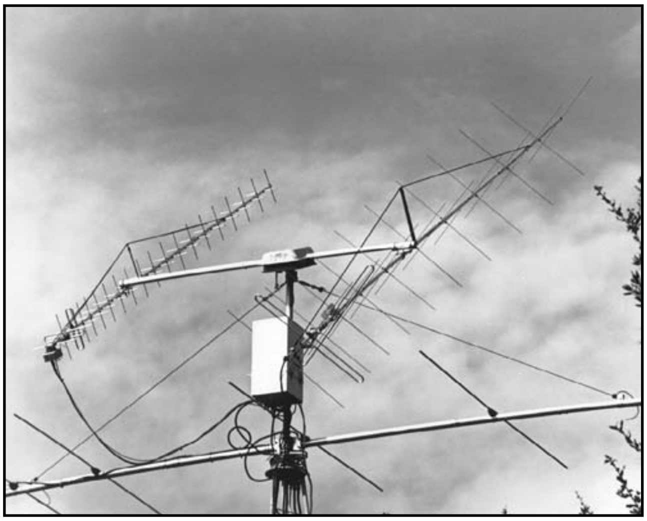

Fig 6—Dick Jansson, WD4FAB, used these 2-meter and 70-cm crossed Yagi's in RHCP for AO-10 and AO-13 operations. The satellite antennas are shown mounted above a 6-meter long-boom Yagi.

CIRCULAR POLARIZATION

Linearly polarized antennas are horizontal or vertical in terms of the antenna's position relative to the surface of the Earth, a reference that loses its meaning in space. The need to use circularly polarized (CP) antennas for space communications is well established. If spacecraft antennas used linear polarization, ground stations would not be able to maintain polarization alignment with the spacecraft because of changing orientations. The ideal antenna for random satellite polarizations is one with a circularly polarized radiation pattern.

There are two commonly used methods for obtaining circular polarization. One is with crossed linear elements such as dipoles or Yagis, as Fig 6 shows. The second popular CP method uses a helical antenna, described below. Other methods also exist, such as with the omnidirectional quadrifilar helix, **Fig 7**.

Polarization *sense* is a critical factor, especially in EME and satellite work. The IEEE standard uses the term "clockwise circular polarization" for a *receding* wave. Amateur technology follows the IEEE standard, calling clockwise polarization for a receding wave as right-hand, or RHCP. Either clockwise or a counter-clockwise (LHCP) sense can be selected by reversing the phasing harness of a crossed-Yagi antenna, see **Fig 8**. The sense of a helical antenna is fixed, determined by its physical construction.

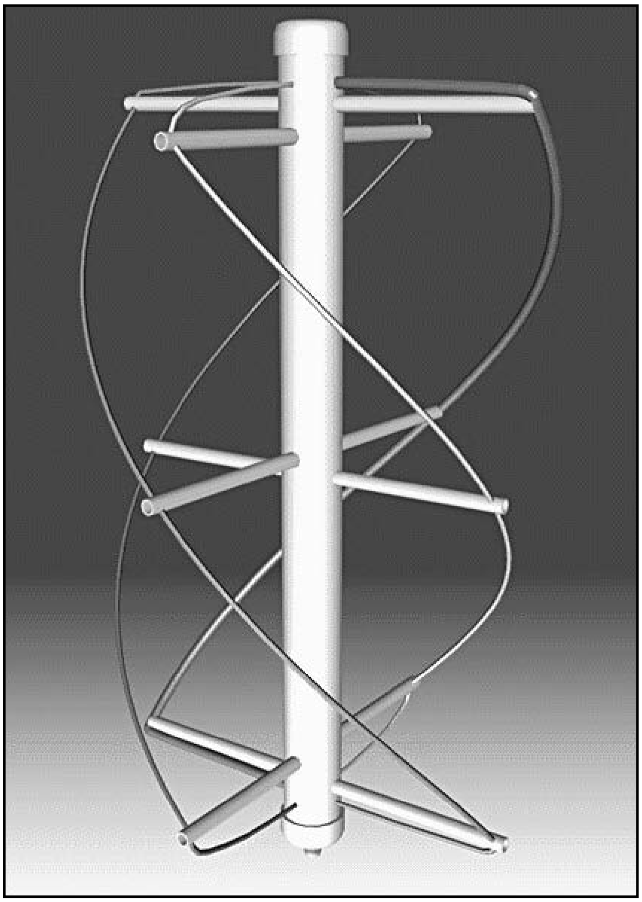

Fig 7—W3KH suggests that quadrifilar antennas can serve well for omnidirectional satellite-station antenna service.

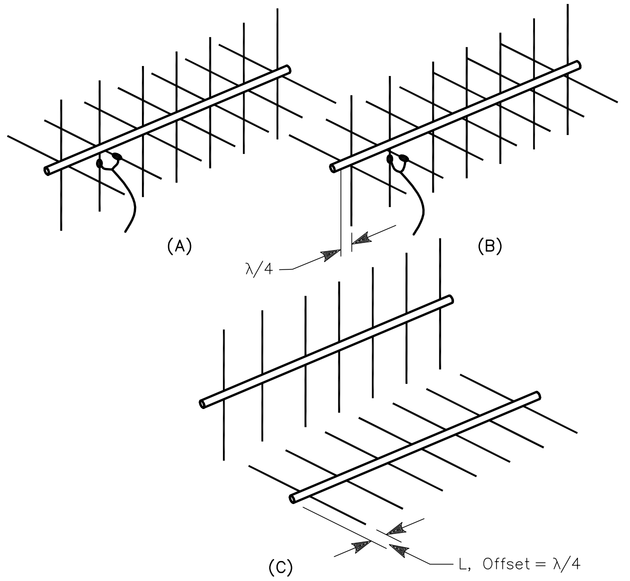

Fig 8—Evolution of the circularly polarized Yagi. The simplest form of crossed Yagi, A, is made to radiate circularly by feeding the two driven elements 90° out of phase. Antenna B has the driven elements fed in phase, but has the elements of one bay mounted ¼ λ forward from those of the other. Antenna C offers elliptical (circular) polarization using separate booms. The elements in one set are perpendicular to those of the other and are ¼ λ forward from those of the other.

See **Fig 9** for construction details.

In working through a satellite with a circularly polarized antenna, it is often convenient to have the capability of switching polarization sense. This is because the sense of the received signal of some of the LEO satellites reverses when the satellite passes its nearest point to you. If the received signal has right-hand circular polarization as the satellite approaches, it may have left-hand circularity as the satellite recedes. There is a sense reversal in EME work, as well, because of a phase reversal of the signal as it is reflected from the surface of the moon. A signal transmitted with right-hand circularity will be returned to the Earth with left-hand circularity. Similarly, the polarization is reversed as it is reflected from a dish antenna, so that for an overall RHCP performance, the feed antenna for the dish needs to be LHCP.

Crossed Linear Antennas

Dipoles radiate linearly polarized signals, and the polarization direction depends on the orientation of the antenna. It two dipoles are arranged for horizontal and vertical dipoles, and the two outputs are combined with the correct phase difference (90°), a circularly polarized wave results. Because the electric fields are identical in magnitude, the power from the transmitter will be equally divided between the two fields. Another way of looking at this is to consider the power as being divided between the two antennas; hence the gain of each is decreased by 3 dB when taken alone in the plane of its orientation.

A 90° phase shift must exist between the two antennas and the simplest way to obtain this shift is to use two feed lines to a coplanar pair of crossed-Yagi antennas. One feed-line section is ¼ λ longer than the other, as shown in Fig 8A. These separate feed lines are then paralleled to a common transmission line to the transmitter or receiver. Therein lies one of the headaches of this system. Assuming negligible coupling between the crossed antennas, the impedance presented to the common transmission line by the parallel combination is one half that of either section alone. (This is not true when there is mutual coupling between the antennas, as in phased arrays.) A practical construction method for implementing a RHCP/LHCP coplanar switched system is shown in **Fig 10**.

Another example of a coplanar crossed-Yagi antenna is shown in **Fig 11**. With this phasing-line method, any mismatch at one antenna will be magnified by the extra ¼ λ of transmission line. This upsets the current balance between the two antennas, resulting in a loss of polarization circularity. Another factor to consider is the attenuation of the cables used in the harness, along with the

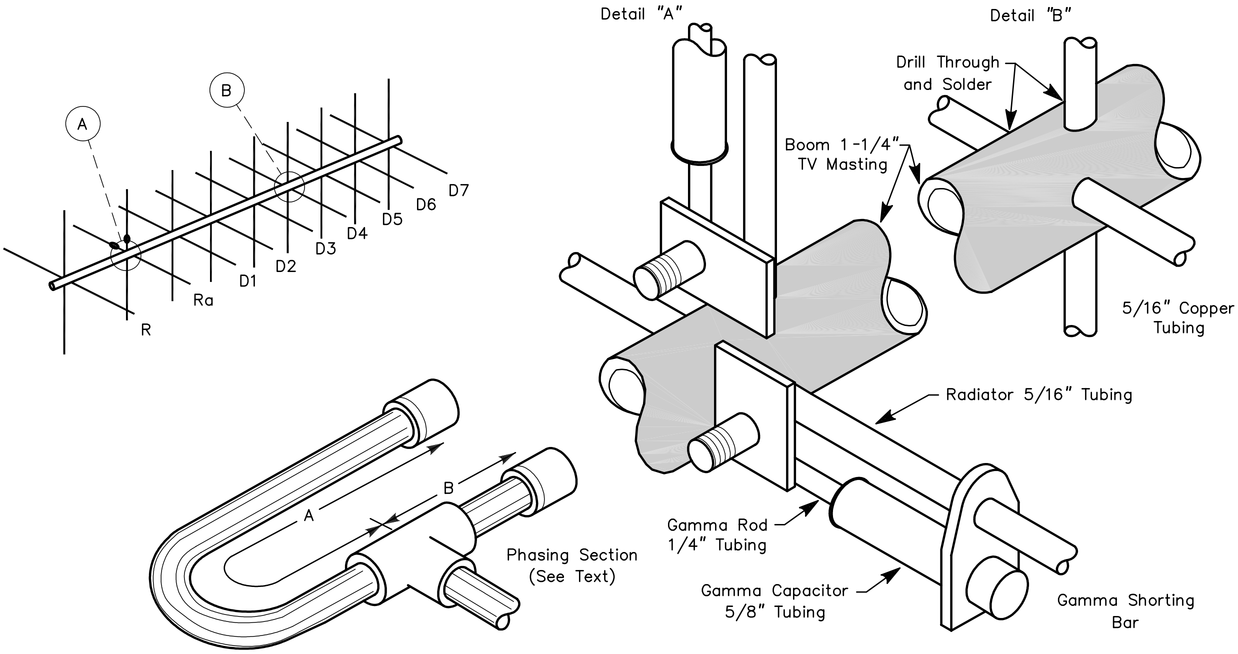
Detail "A"
Detail "B"
Drill Through and Solder
Boom 1-1/4" TV Masting
A
B
D7
D6
D5
D4
D3
D2
D1
Ra
R
5/16" Copper Tubing
Radiator 5/16" Tubing
A
B
Phasing Section (See Text)
Gamma Rod 1/4" Tubing
Gamma Capacitor 5/8" Tubing
Gamma Shorting Bar

Fig 9—Construction details of a co-planar crossed-Yagi antenna.

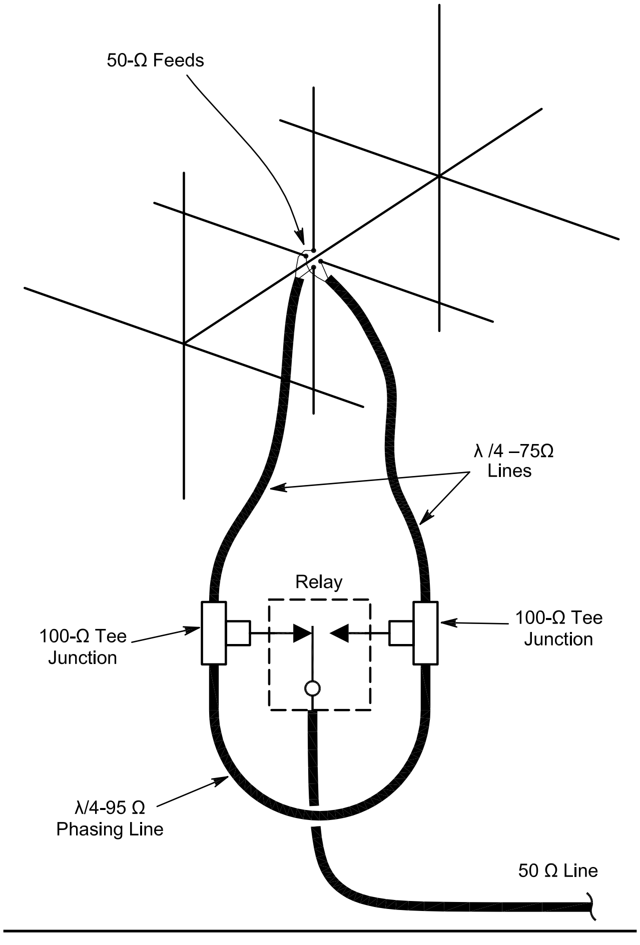
50-Ω Feeds
λ /4 –75Ω Lines
Relay
100-Ω Tee Junction
100-Ω Tee Junction
λ/4-95 Ω Phasing Line
50 Ω Line

Fig 10—Co-planar crossed Yagi, circularly polarized antenna with switchable polarization phasing harness.

connectors. Good low-loss coaxial line should be used. Type-N or BNC connectors are preferable to the UHF variety.

Another method to obtain circular polarization is to use equal-length feed lines and place one antenna $^1/_4$ λ ahead of the other. This offset pair of Yagi-crossed antennas is shown in Fig 8B. The advantage of equal-length feed lines is that identical load impedances will be presented to the common feeder, as shown in **Fig 12**, which shows a fixed circularity sense feed. To obtain a switchable sense feed with the offset Yagi pair, you can use a connection like that of **Fig 13**, although you must compensate for the extra phase added by the relay and connectors.

Fig 8C diagrams a popular method of mounting two separate off-the-shelf Yagis at right angles to each other. The two Yagis may be physically offset by $^1/_4$ λ and fed in parallel, as shown in Fig 8C, or they may be mounted with no offset and fed 90° out of phase. Neither of these arrangements on two separate booms produces true circular polarization. Instead, *elliptical* polarization results from such a system. **Fig 14** is a photo of this type of mounting of Yagis on two booms for elliptical operation.

Helical Antennas

As mentioned, the second method to create a circularly polarized signal is by means of a helical antenna. The axial-mode helical antenna was introduced by Dr John Kraus, W8JK, in the 1940s. **Fig 15** shows examples of S-band (2400-MHz), V-band (145-MHz), and U-band (435-MHz) helical antennas, all constructed by WD4FAB for satellite service.

This antenna has two characteristics that make it

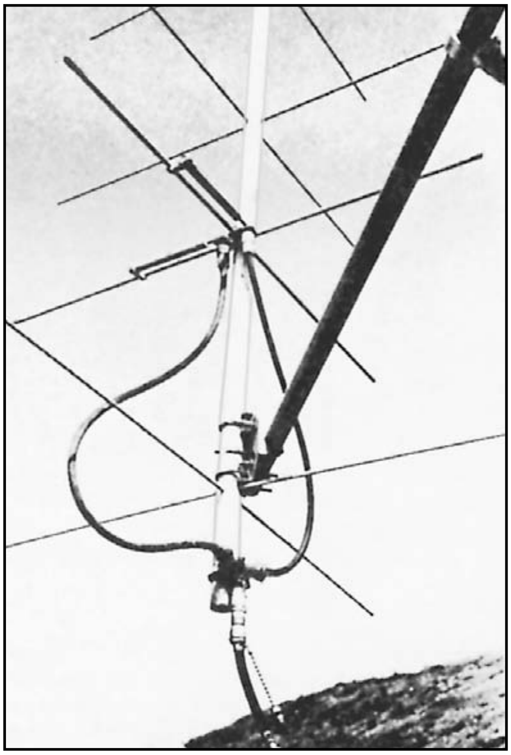

Fig 11—This VHF crossed Yagi design by KH6IJ (Jan 1973 *QST*) illustrates the co-planar, fixed-circularity Yagi.

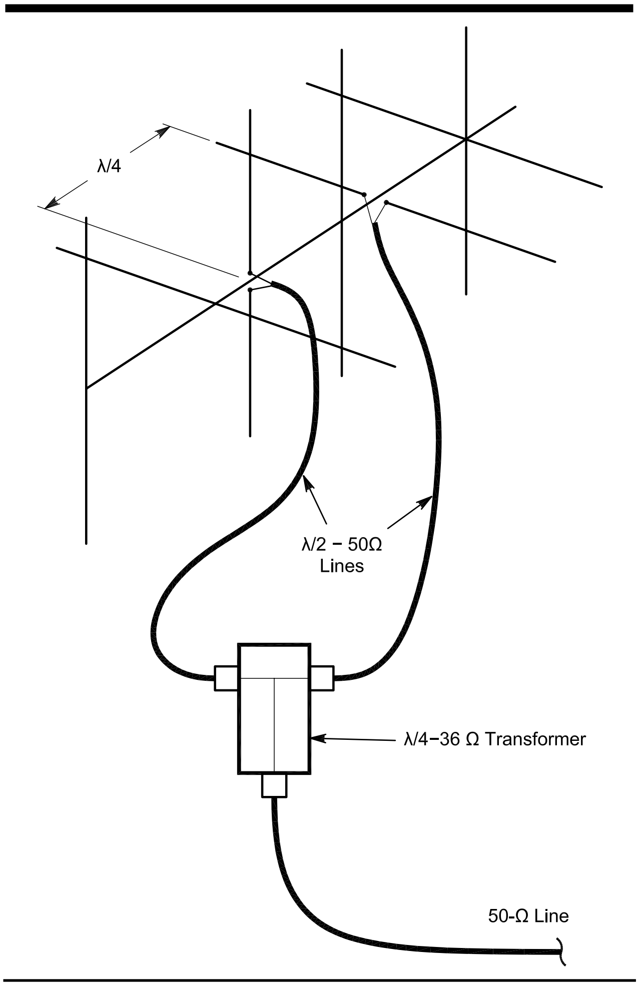

Fig 12—Offset crossed-Yagi circularly polarized antenna-phasing harness with fixed polarization.

especially interesting and useful in many applications. First, the helix is circularly polarized. As discussed earlier, circular polarization is simply linear polarization that continually rotates as it travels through space. In the case of a helical antenna, this rotation is about the axis of the antenna. This can be pictured as the second hand of a watch moving at the same rate as the applied frequency, where the position of the second hand can be thought of as the instantaneous polarization of the signal.

The second interesting property of the helical antenna is its predictable pattern, gain and impedance characteristics over a wide frequency range. This is one of the few antennas that has both broad bandwidth and high gain. The benefit of this property is that, when used for narrow-band applications, the helical antenna is very forgiving of mechanical inaccuracies.

Probably the most common amateur use of the helical antenna is in satellite communications, where the spinning of the satellite antenna system (relative to the earth) and the effects of Faraday rotation cause the polarization of the satellite signal to be unpredictable. Using a linearly polarized antenna in this situation results in deep fading, but with the helical antenna (which responds equally to linearly polarized signals), fading is essentially eliminated.

This same characteristic makes helical antennas useful in polarization-diversity systems. The advantages of circular polarization have been demonstrated on VHF voice schedules over non-optical paths, in cases where linearly polarized beams did not perform satisfactorily.

Another use for the helical antenna is the transmission of color ATV signals. Many beam antennas (when adjusted for maximum gain) have far less bandwidth than the required 6 MHz, or have non-uniform gain over this frequency range. The result is significant distortion of the transmitted and received signals, affecting color reproduction and other features. This problem becomes more aggravated over non-optical paths. The helix exhibits maximum gain (within 1 dB) more than 20 MHz anywhere above 420 MHz.

The helical antenna can be used to advantage with multimode rigs, especially above 420 MHz. Not only does the helix give high gain over an entire amateur band, but it also allows operation on FM, SSB and CW without the need for separate vertically and horizontally polarized antennas.

Helical Antenna Basics

The helical antenna is an unusual specimen in the antenna world, in that its physical configuration gives a

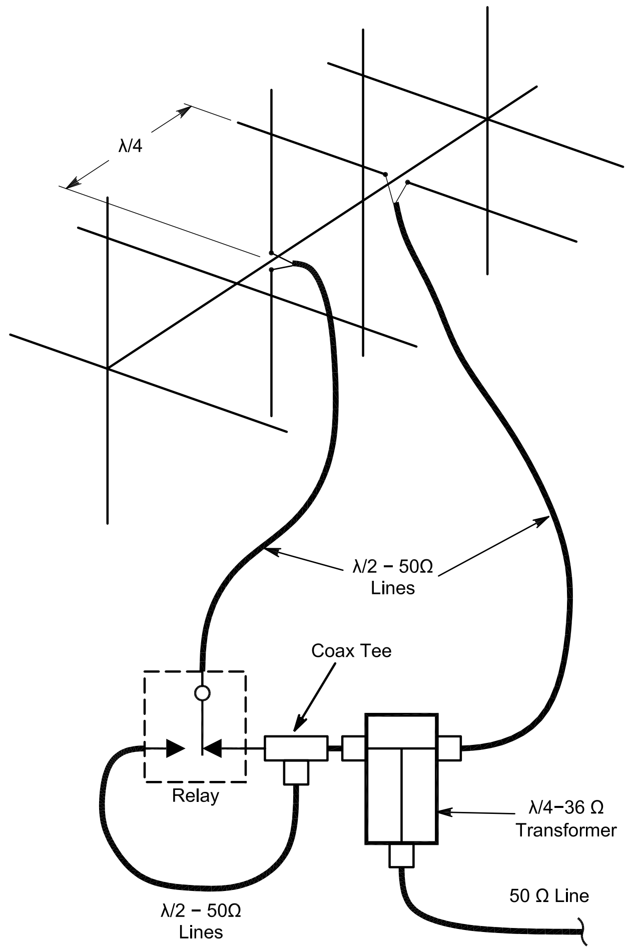

Fig 13—Offset crossed-Yagi circularly polarized antenna-phasing harness with switchable polarization.

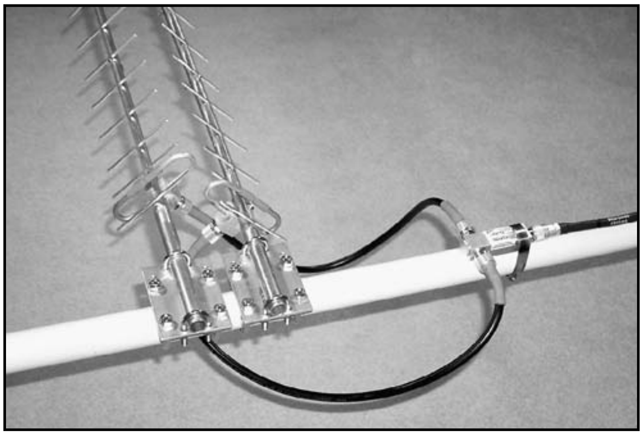

Fig 14—An example of offset crossed-Yagi circularly polarized antennas with fixed polarization. This example is a pair of M^2 23CM22EZA antennas, for L band (1269 MHz), mounted on an elevation boom. (*WD4FAB photo.*)

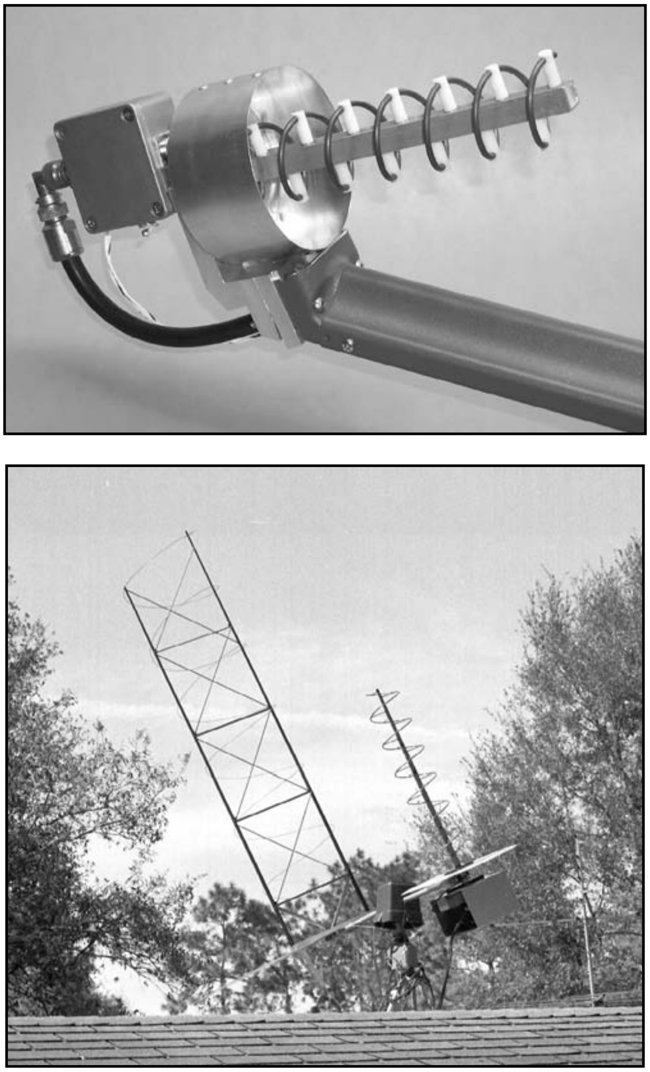

Fig 15—At top, a seven-turn LHCP helical antenna for S-band dish feed for AO-40 service. This helical antenna uses a cupped reflector and has a preamplifier mounted directly to the antenna feed point. At bottom, a pair of helical antennas for AO-10 service on 2 meters and 70 cm. The 2-meter helical antenna is not small! (*WD4FAB photos.*)

hint to its electrical performance. A helix looks like a large air-wound coil with one of its ends fed against a ground plane, as shown in **Fig 16**. The ground plane is a screen of 0.8 λ to 1.1 λ diameter (or on a side for a square ground plane). The circumference (C_λ) of the coil form must be between 0.75 λ and 1.33 λ for the antenna to radiate in the axial mode. The coil should have at least three turns to radiate in this mode. The ratio of the spacing between turns (in wavelengths), S_λ to C_λ, should be in the range of 0.2126 to 0.2867. This ratio range results from the requirement that the pitch angle, α, of the helix be between 12° and 16°, where:

$$\alpha = \arctan \frac{S_\lambda}{C_\lambda} \tag{Eq 1}$$

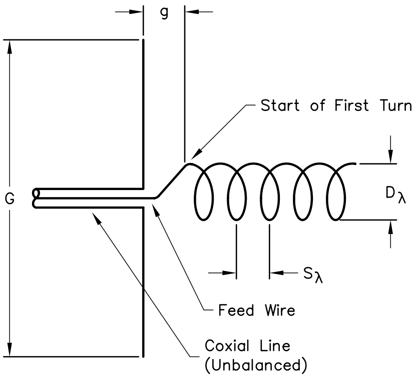

Fig 16—The basic helical antenna and design equations.

These constraints result in a single main lobe along the axis of the coil. This is easily visualized from Fig 15A. The winding of the helix comes away from the cupped reflector with a counterclockwise winding direction for a LHCP. (The winding can also be a clockwise—this results in a RHCP polarization sense.)

A helix with a C_λ of 1λ has a wave propagating from one end of the coil (at the ground plane), corresponding to an instantaneous dipole "across" the helix. The electrical rotation of this dipole produces circularly polarized radiation. Because the wave is moving along the helix conductor at nearly the speed of light, the rotation of the electrical dipole is at a very high rate, and true circular polarization results.

The IEEE definition, in simple terms, is that when viewing the antenna from the feed-point end, a clockwise wind results in right-hand circular polarization (RHCP), and a counterclockwise wind results in left-hand circular polarization (LHCP). This is important, because when two stations use helical antennas over a nonreflective path, both must use antennas with the same polarization sense. If antennas of opposite sense are used, a signal loss of at least 20 dB results from the cross polarization alone.

As mentioned previously, circularly polarized antennas can be used in communications with any linearly polarized antenna (horizontal or vertical), because circularly polarized antennas respond equally to all linearly polarized signals. The gain of a helix is 3 dB less than the theoretical gain in this case, because the linearly polarized antenna does not respond to linear signal components that are orthogonally polarized relative to it.

The response of a helix to all polarizations is indicated by a term called *axial ratio*, also known as *circularity*. Axial ratio is the ratio of amplitude of the polarization that gives maximum response to the amplitude of the polarization that gives minimum response. An ideal circularly polarized antenna has an axial ratio of 1.0. A well-designed practical helix exhibits an axial ratio of 1.0 to 1.1. The axial ratio of a helix is:

$$AR = \frac{2n + 1}{2n} \qquad \text{(Eq 2)}$$

where:

AR = axial ratio
n = the number of turns in the helix

Axial ratio can be measured in two ways. The first is to excite the helix and use a linearly polarized antenna with an amplitude detector to measure the axial ratio directly. This is done by rotating the linearly polarized antenna in a plane perpendicular to the axis of the helix and comparing the maximum and minimum amplitude values. The ratio of maximum to minimum is the axial ratio.

The impedance of the helix is easily predicted. The terminal impedance of a helix is unbalanced, and is defined by:

$$Z = 140 \times C_\lambda \qquad \text{(Eq 3)}$$

where Z is the impedance of the helix in ohms.

The gain of a helical antenna is determined by its physical characteristics. Gain can be calculated from:

$$\text{Gain}\left(\text{dBi}\right) = 11.8 + 10 \log\left(C_\lambda^{\,2}\, nS_\lambda\right) \qquad \text{(Eq 4)}$$

In practice, helical antennas do not deliver the gain in Eq 4 for antennas with turns count greater than about twelve. There will be more discussions in this area when practical antennas are discussed.

The beamwidth of the helical antenna (in degrees) at the half-power points is:

$$BW = \frac{52}{C_\lambda \sqrt{nS_\lambda}} \qquad \text{(Eq 5)}$$

The diameter of the helical antenna conductor should be between 0.006 λ and 0.05 λ, but smaller diameters have been used successfully at 144 MHz. The previously noted diameter of the ground plane (0.8 λ to 1.1 λ) should not be exceeded if you desire a clean radiation pattern. As the ground plane size is increased, the sidelobe levels also increase. Cupped ground planes have been used according to Kraus, as in Fig 15. (The ground plane need not be solid; it can be in the form of a spoked wheel or a frame covered with hardware cloth or screen.)

50-Ω Helix Feed

Joe Cadwallader, K6ZMW, presented this feed method in June 1981 *QST*. Terminate the helix in an N connector mounted on the ground screen at the periphery of the helix. See **Fig 17**. Connect the helix conductor to the N connector as close to the ground screen as possible (**Fig 18**). Then adjust the first quarter turn of the helix to a close spacing from the reflector.

This modification goes a long way toward curing a deficiency of the helix—the 140-Ω nominal feed-point

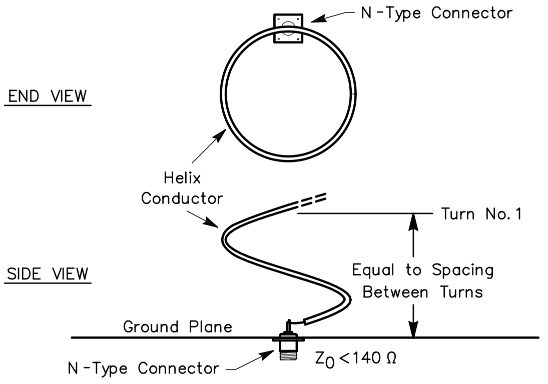

Fig 17—End view and side view of peripherally fed helix.

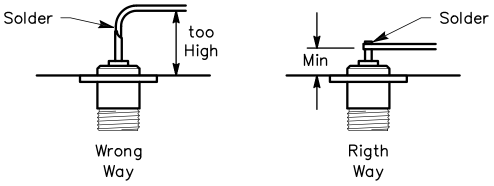

Fig 18—Wrong and right ways to attach helix to a type N connector for 50-Ω feed.

impedance. The traditional λ/4 matching section has proved difficult to fabricate and maintain. But if the helix is fed at the periphery, the first quarter turn of the helix conductor (leaving the N connector) acts much like a transmission line—a single conductor over a perfectly conducting ground plane. The impedance of such a transmission line is:

$$Z_0 = 138 \log \frac{4h}{d} \qquad \text{(Eq 6)}$$

where:

Z_0 = line impedance in ohms
h = height of the center of the conductor above the ground plane
d = conductor diameter (in the same units as h).

The impedance of the helix is 140 Ω a turn or two away from the feed point. But as the helix conductor swoops down toward the feed connector (and the ground plane), h gets smaller, so the impedance decreases. The 140-Ω nominal impedance of the helix is transformed to a lower value. For any particular conductor diameter, an optimum height can be found that will produce a feed-point impedance equal to 50 Ω. The height should be kept very small, and the diameter should be large. Apply power to the helix and measure the SWR at the operating frequency. Adjust the height for an optimum match.

Typically, the conductor diameter may not be large

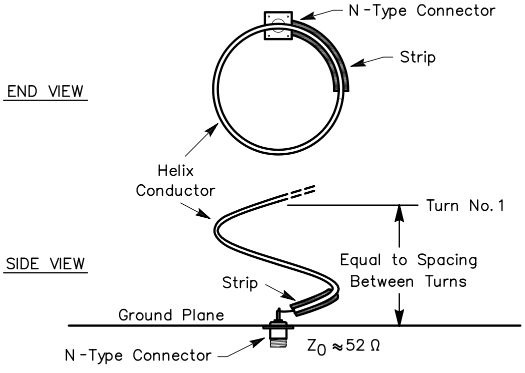

Fig 19—End view and side view of peripherally fed helix with metal strip added to improve transformer action.

enough to yield a 50-Ω match at practical (small) values of h. In this case, a strip of thin brass shim stock or flashing copper can be soldered to the first quarter turn of the helix conductor (**Fig 19**). This effectively increases the conductor diameter, which causes the impedance to decrease further yet. The edges of this strip can be slit every 1/2 inch or so, and the strip bent up or down (toward or away from the ground plane) to tune the line for an optimum match.

This approach yields a perfect match to nearly any coax. The usually wide bandwidth of the helix (70% for less than 2:1 SWR) will be reduced slightly (to about 40%) for the same conditions. This reduction is not enough to be of any consequence for most amateur work. The improvements in performance, ease of assembly and adjustment are well worth the effort in making the helix more practical to build and tune.

ANTENNAS FOR AO-40 OPERATIONS

Antennas for successful operations on AO-40 come in many shapes and sizes. AO-40 has provided amateurs the opportunity to broadly experiment with antennas.

Fig 20 shows the satellite antennas at WD4FAB. The Yagi antennas are used for the U- and L-band AO-40 uplinks and the V-band AO-10 downlink, while the S-band dish antenna is for the AO-40 downlink. These satellite antennas are tower mounted at 63 feet (19 meters) to avoid pointing into the many nearby trees and suffering from the resulting "green attenuation." Of course, satellite antennas do not always need to be mounted high on a tower if dense foliage is not a problem. If satellite antennas are mounted lower down, feed-line length and losses can reduced.

Another benefit, however, to tower mounting of satellite antennas is that they can be used for terrestrial ham communications and contests. The fact that the antennas are set up for CP does not really degrade these other operating activities.

Experience with AO-40 has clearly shown the advan-

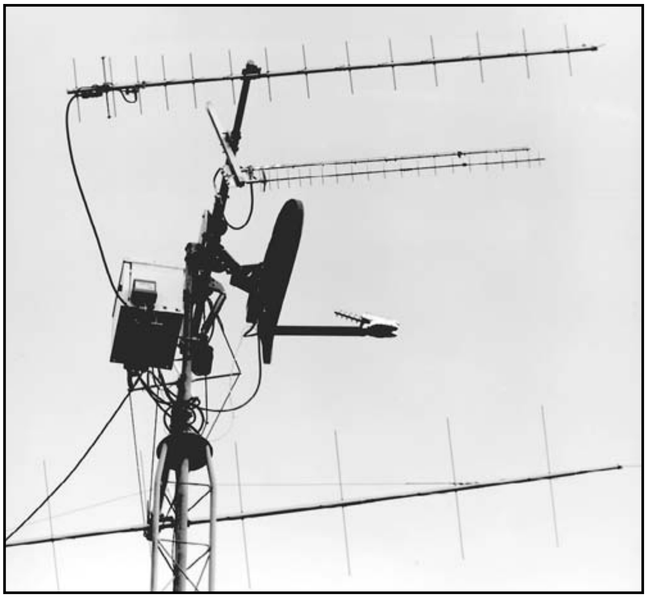

Fig 20—Details of WD4FAB's tower cluster of satellite antennas including a home-brew elevation rotator. Top to bottom: M^2 436-CP30, a CP U-band antenna; two M^2 23CM22EZA antennas in a CP array for L band; "FABStar" dish antenna with helix feed for S band; M^2 2M-CP22, a CP V-band antenna (only partially shown.) To left of dish antenna is a NEMA4 equipment box with an internal 40-W L-band amplifier, and also hosts externally mounted preamplifiers. (*WD4FAB photo*)

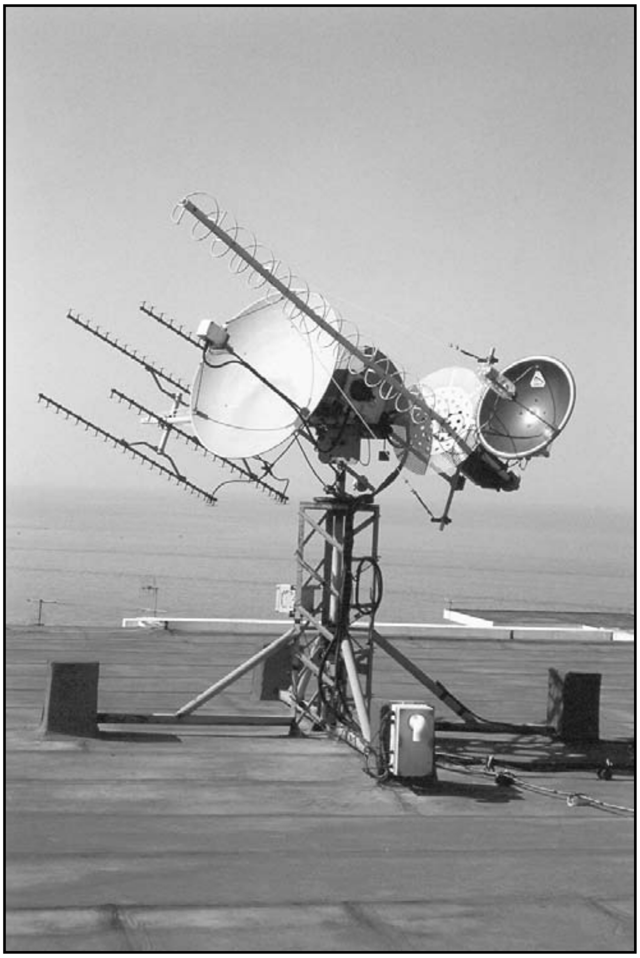

Fig 21—Domenico, I8CVS, has this cluster of satellite antennas for AO-40. Left to right: array of 4 × 23-element Yagi horizontally polarized for L band; 1.2-meter dish with 3-turn helix feed for S band; 15-turn RHCP helical antenna for U band; 60-cm dish for X band. All microwave preamplifiers and power amplifiers are homebrew and are mounted on this antenna cluster. (*I8CVS photo.*)

tages of using RHCP antennas for both the uplink and downlink communications. The antennas shown in Fig 20 are a single-boom RHCP Yagi antenna for U band, a pair of closely spaced Yagi antennas phased for RHCP for L band (see Fig 14), and a helix-fed dish antenna for S band. The antenna gain requirements for U band can easily be met with the gain of a 30-element crossed Yagi. Antennas of this size have boom lengths of 4 to 4$^1/_2$ wavelengths. The enterprising constructor can build a Yagi antenna from one of several references, however most of us prefer to purchase well-tested antennas from commercial sources as M^2 or Hy-Gain. In the past, KLM (now out of business) had offered a 40-element CP Yagi for U-band satellite service, and many of these are still in satisfactory use today.

U-band uplink requirements for AO-40 have clearly demonstrated the need for gain less than 16 to 17 dBic RHCP, with an RF power of less than 50 W PEP at the antenna (≈ 2,500 W-PEP EIRP with a RHCP antenna) depending upon the *squint angle*. (The squint angle is the angle at which the main axis of the satellite is pointed away from your antenna on the ground. If the squint angle is less than half of the half-power beamwidth, the ground station will be within the spacecraft antenna's nominal beam width.)

A gain of 16 to 17 dBic RHCP can be obtained from a 30-element crossed Yagi, AO-13 type antenna, and is good news, considering that the satellite may be over 60,000 km (37,000 miles) from your station. Success on the U-band uplinks to AO-40 is easier than those for L band at wider squint angles more than 20°. At squint angles less than 10°, U-band uplink operation can even be done with 1-5 W power outputs to a RHCP antenna (≈ 200 W-PEP EIRP with RHCP). These lower levels mean that smaller antennas can be used. In practice, these uplinks will produce downlink signals that are 10 to 15 dB above the noise floor, or S7 signals over an S3 noise floor. The beacon will give a downlink S9 signal for these same conditions.

WD4FAB's experience with the AO-40 L-band uplink has demonstrated that 40 W-PEP delivered to an antenna with a gain of ≈ 19dBic (3,000 W-PEP EIRP with RHCP) is needed for operations at the highest altitudes of AO-40 and with squint angles ≤ 15°. This is the pretty compact L-band antenna arrangement with two 22-element antennas in a RHCP array shown in Fig 14 and 20. Other operators have experience that using a 1.2-meter L-band dish antenna and 40 W of RF power (6,100 W-PEP EIRP with RHCP) can also provide a superb uplink for squint angles even up to 25°. A dish antenna can have a practical gain of about 21 to 22 dBic. These uplinks will provide the user a downlink that is 10 to 18 dB above the transponder noise floor. In more practical terms, this is an S7 to 8 signal over a S3 transponder noise floor, a very comfortable armchair copy.

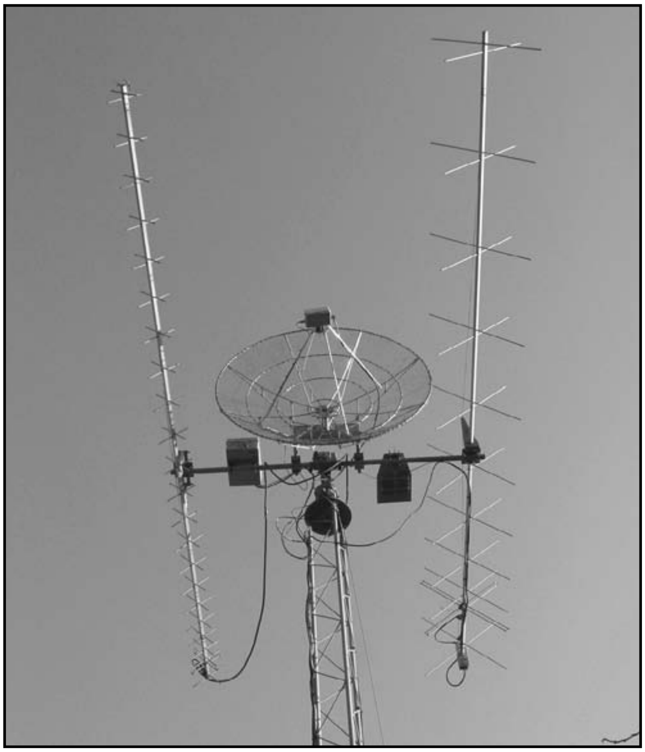

Fig 22—Wilfred Carey, ZS6JT, constructed this cluster of satellite and EME antennas. Left to right: 2 × 23-element offset feed Yagi for U band; 1.64-meter dish with 2¹/₄-turn helix feed for S band; 2 × 11-element coplanar feed Yagi for V band. (*ZS6JT photo.*)

Using the L-band uplink for AO-40, instead of the U-band uplink, allows the use of Yagi antennas that more manageable, since their size for a given gain is only one third of those for U-band. With L band there is a narrower difference between using a dish antenna and a Yagi, since a 21- to 22-dBic dish antenna would be only about 1.2 meters (4 feet) in diameter. However, some of us may not have such “real estate” available on our towers and may seek a lower wind-loading solution offered by Yagis. Long-boom rod-element Yagi, or loop-Yagi antennas are commercially offered by M^2 and DEM, although this band is about the highest for practical Yagis. The example shown in Fig 20 is a pair of rod-element Yagi antennas from M^2 in a CP arrangement with a gain of 18 to 19 dBic.

Other amateurs have successful AO-40 operation with different arrangements. **Fig 21** shows I8CVS’s 4 × 23-element linear array for a 1270 MHz, a 1.2-meter solid dish for 2400 MHz, a 15-turn helical antenna for 435 MHz, and a 60-cm dish for 10,451 MHz. This arrangement clearly shows the advantage and accessibility of having a roof-mounted antenna.

Fig 22 shows ZS6JT’s setup, with a 1.64-meter home-built mesh dish for 2400 MHz and two home-built crossed Yagi antennas, one for 435 MHz and the other for 145 MHz. Note that in these examples, the antennas permit terrestrial communication as well as satellite service. Two of these stations have also maintained the capability to operate the LEO satellites with U- and V-band antennas.

A number of amateurs have taken advantage of the availability of surplus C-band TVRO dishes, since most users of satellite television have moved up to the more

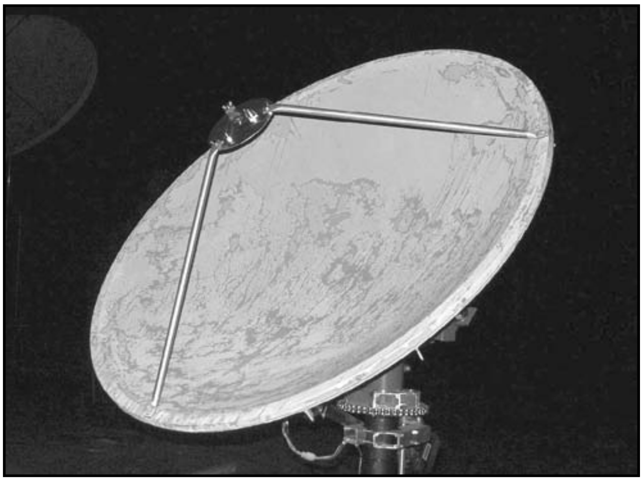

Fig 23—Robert Suding, W0LMD, modified this 4-foot dish antenna with a patch feed for S band and an Az-El mount. (*W0LMD photo.*)

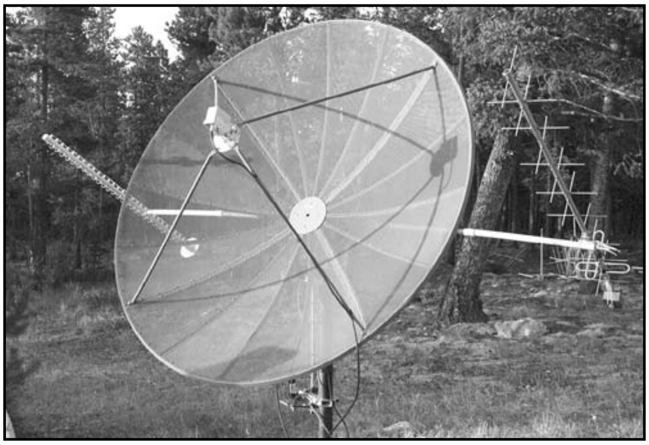

Fig 24—W0LMD graduated to this 8-foot dish with patch feed for S band for AO-40. On the left is a helical antenna for L band and on the right is a 2 × 9-element offset-feed Yagi for U band. A home-brew Az-El mount is provided. (*W0LMD photo.*)

convenient K band using 0.5-meter dishes. Some examples of these dish conversions for satellite communications are shown in **Fig 23**, a W0LMD 4-foot dish with patch feed and Az-El mount. **Fig 24** shows a W0LMD 8-foot dish with patch feed, Az-El mount, a U-band Yagi, and an L-band helical antenna.

Fig 25 is a W0LMD 10-foot dish with tri-band patch feed and Az-El mount; and **Fig 26** is also a W0LMD

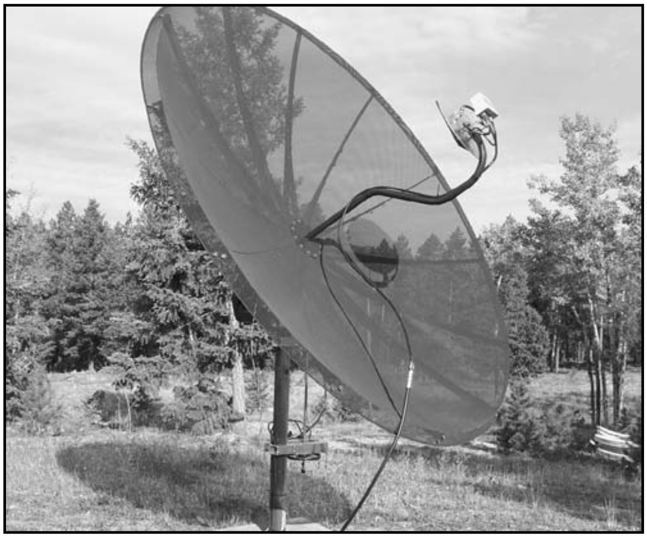

Fig 25—W0LMD increased to this 10-foot dish for AO-40 operations, with a triband patch feed for U, L, and S bands on an Az-El mount. (*W0LMD photo.*)

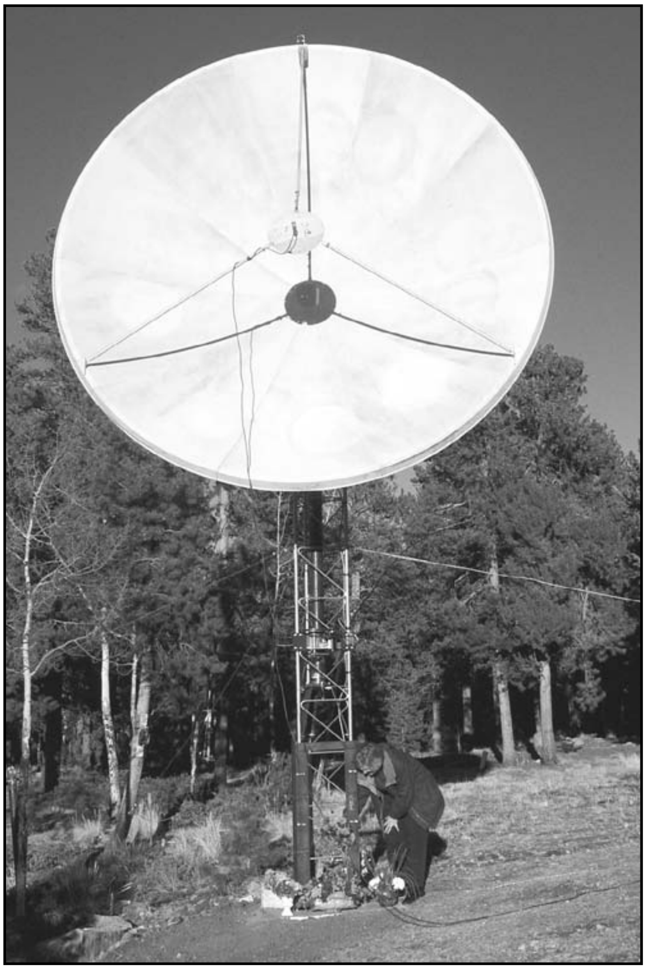

Fig 26—W0LMD found the ultimate in this 14-foot dish for AO-40, with a triband patch feed and Az-El mount. (*W0LMD photo.*)

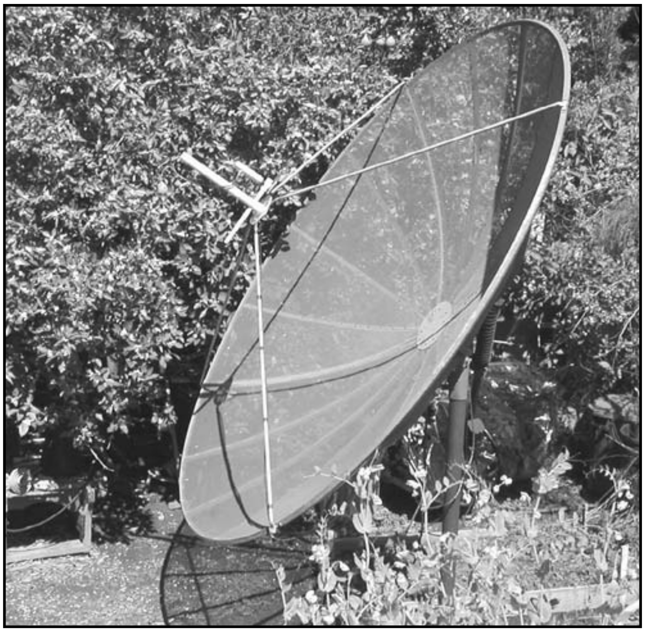

Fig 27—Clair E. Cessna, K6LG, has this 10-foot dish with S-band patch feed. This dish uses the original polar-mounting system and offsets the patch feed to compensate for AO-40's deviation from the Clarke belt. (*K6LG photo.*)

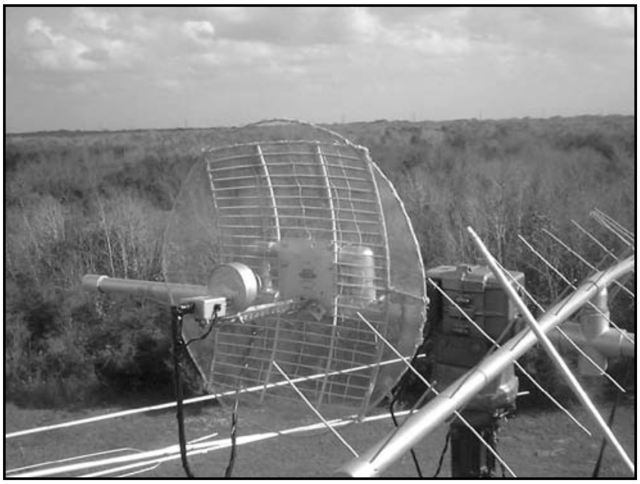

Fig 28—K5GNA's "circularized" mesh modification of an MMDS dish antenna with a helix-CP feed and DEP preamp. The dish modification reduces the spillover loss by making the antenna fully circular. (*K5OE photo.*)

14-foot dish with tri-band patch feed and Az-El mounting. Other operators, like K6LG, have been able to use TVRO dishes, **Fig 27**, with multiband patch feeds and still use, within limits, their polar-mounting system, as will be explained later.

Other hams have taken advantage of other surplus dish situations. **Fig 28** shows modified MMDS dishes, by K5GNA, and **Fig 29**, by K5OE, both using helix feeds. **Fig 30** shows a 75-cm high modified PrimeStar offset feed dish, by WD4FAB, using a longer helical feed antenna because of the higher f/D ratio of this dish configuration. This dish provides 5 dB of Sun noise, which is good performance. These efforts have rewarded their users with superb service on AO-40. Many have experimented with different feed and mounting systems. These experiments will be further illustrated.

One very popular spun-aluminum dish antenna seen in use on AO-40 has been the G3RUH-ON6UG 60-cm unit with its S-band patch feed, **Fig 31**. A kit, complete with a CP-patch feed is available from SSB-USA and has a gain of 21 dBic. It provides a 2.5-dB Sun noise signal. Surplus dishes have not been the only source for antennas for AO-40 operations, since some ingenious operators have even turned to the use of cardboard boxes. See **Figs 32** to **35**.

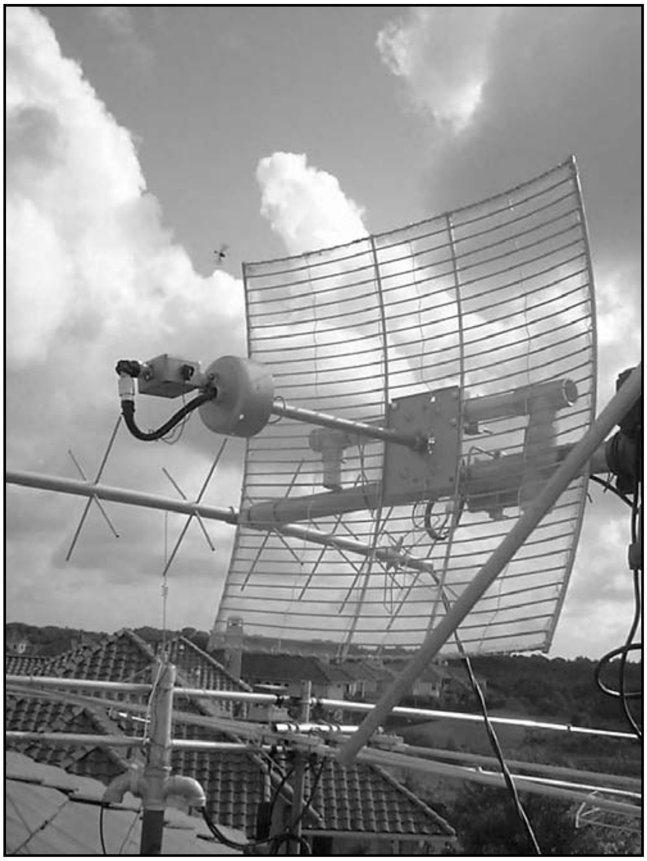

Fig 29—Mesh modification of an MMDS dish antenna by Jerry Brown, K5OE, with a helix-CP feed and DEM preamplifier mounted directly to the helix feed point. (*K5OE photo.*)

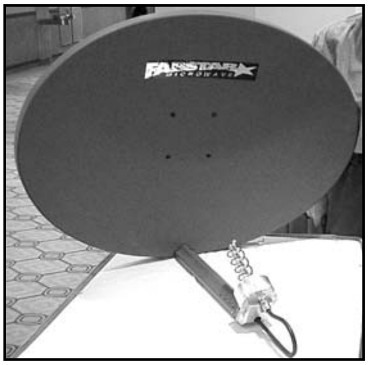

Fig 30—PrimeStar offset-fed dish with WD4FAB's helix-feed antenna. N0NSV was so pleased with the modification that he renamed the dish "FABStar," and made a new label! (*N0NSV photo.*)

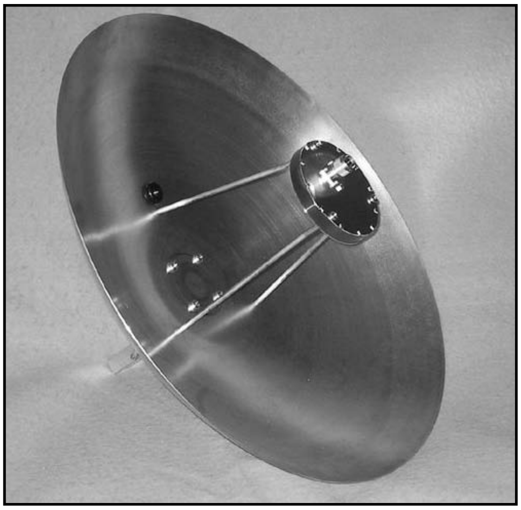

Fig 31—G3RUH's 60-cm spun-aluminum dish with CP-patch feed is available as a kit. This antenna has been popular with many AO-40 operators all over the world.

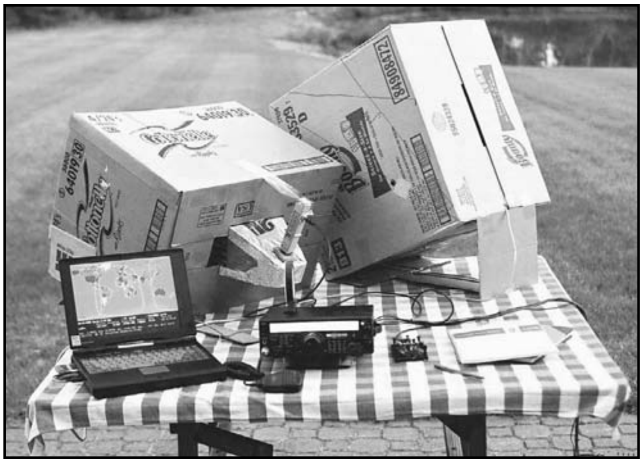

Fig 32—A complete satellite station with the tracking laptop, FT-847 transceiver, and both downlink and uplink cardboard-box antennas.

Parabolic Reflector Antennas

The satellite S-band downlinks have become very popular for a variety of reasons:

- Good performance with physically small downlink antennas
- Availability of good-quality downconverters
- Availability of preamps at reasonable prices.

A number of people advocate S-band operation, including Bill McCaa, K0RZ, who led the team that designed and built the AO-13 S-band transponder and James Miller, G3RUH, who operates one of the AO-40 command stations. Ed Krome, K9EK, and James Miller have published a number of articles detailing construction of preamps, downconverters and antennas for S band.

Some access AO-40's S-band downlink using compact S-band helical antennas. See **Fig 36**. With the demise of AO-40's S1 transmitter and its high-gain downlink antenna, enthusiasts have had to employ high-gain parabolic-dish antennas to use AO-40's S2 downlink, with its lower-gain helical antenna.

W0LMD notes that like a bulb in a flashlight, the

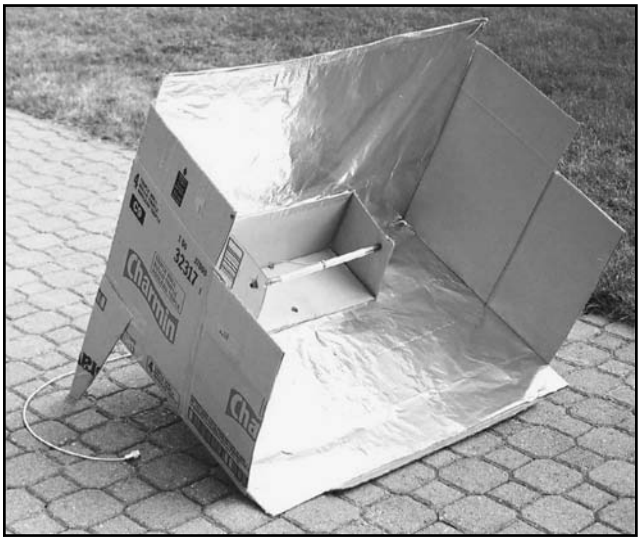

Fig 33—The completed high-performance corner-reflector uplink antenna for U band. Note how the box corners hold the reflectors and dipole feed in place. The rear legs set the antenna elevation to 20°—this gives good coverage at the design latitude but will need modification for other stations.

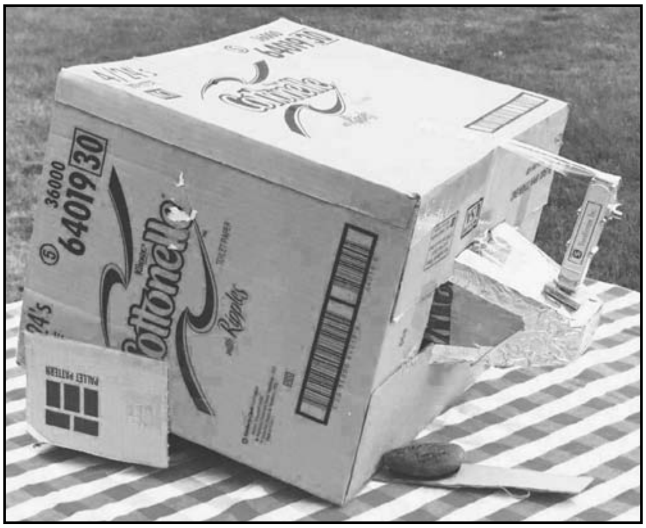

Fig 35—Side view of the downlink pyramidal horn showing the elevation control supports and how the downconverter is attached to the horn. The downconverter is pulled forward by the tape to align the probe wire parallel to the rear surface of the horn.

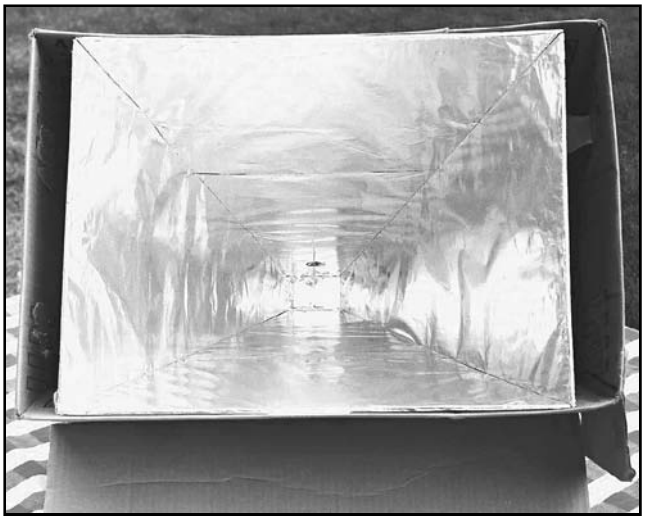

Fig 34—Front view of the downlink pyramidal horn showing how it is mounted in the support carton. Notice the coax probe at the back of the horn.

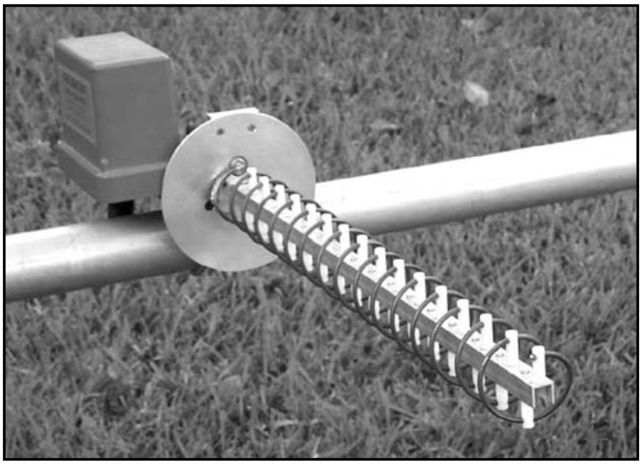

Fig 36—WD4FAB's example of a 16-turn S-band helical antenna for AO-40. This is about the maximum length of any practical helix. Note the SSB UEK2000 downconverter mounted behind the reflector of the antenna. (*WD4FAB photo.*)

parabolic reflector or dish antenna must have a feed source looking into the surface of the dish. Some dishes are designed so that the feed source is mounted directly in front of the dish. This is referred to as a *center-fed dish*. Other dishes are designed so that the feed source is off to one side, referred to as an *off-center-fed dish*, or just offset-fed dish, as shown in Fig 30. The offset-fed dish may be considered a side section of a center-fed dish. The center-fed dish experiences some signal degradation due to blockage of the feed system, but this is usually an insignificantly small amount. The offset-fed dish is initially more difficult to aim, since the direction of reception is not the center axis, as it is for center-fed dishes.

The basic design precepts of parabolic-dish antennas are covered in more detail in the EME Antenna section of this chapter. Dish antenna properties specific to satellite operations are covered here. The dish's parabola can be designed so the focus point is closer to the surface of the dish, referred to a *short-focal-length* dish, or further away from the dish's surface, referred to as a *long-focal-length* dish. To determine the exact focal length, measure the diameter of the dish and the depth of the dish.

$$f = \frac{D^2}{16d} \qquad \text{(Eq 7)}$$

The focal length divided by the diameter of the dish gives the *focal ratio*, commonly shown as f/D. Center-fed dishes usually have short-focal ratios in the range of f/D = 0.3 to 0.45. Offset-fed dishes usually have longer focal lengths, with f/D = 0.45 to 0.80. If you attach two small mirrors to the outer front surface of a dish and then point the dish at the Sun, you can easily find the focus point of the dish. Put the reflector of the patch or helix feed just beyond this point of focus.

An alternate method for finding a dish's focal length is suggested by W1GHZ (ex-N1BWT), who provides a computer program called *HDL_ANT*, available at: **www.w1ghz.org/10g/10g_home.htm**. The method literally measures a solid-surface dish by the dimensions of the bowl of water that it will form when properly positioned. (See: **www.qsl.net/n1bwt/chap5.pdf**.) WD4FAB used this method on the dish of Fig 30, carefully leveling the bowl, plugging bolt holes, and filling it with water to measure the data needed by the W1GHZ Web-site calculation.

While many of us enjoy building our own antennas, surplus-market availability of these small dish antennas makes their construction unproductive. Many AO-40 operators have followed the practices of AO-13 operators using a surplus MMDS linear-screen parabolic reflector antenna, Figs 28 and 29. These grid-dish antennas are often called *barbeque dishes*. K5OE and K5GNA have shown how to greatly improve these linearly polarized reflectors by adapting them for the CP service desired for AO-40. Simple methods can be used to *circularize* a linear dish and to further add to its gain using simple methods to increase the dish area and feed efficiency.

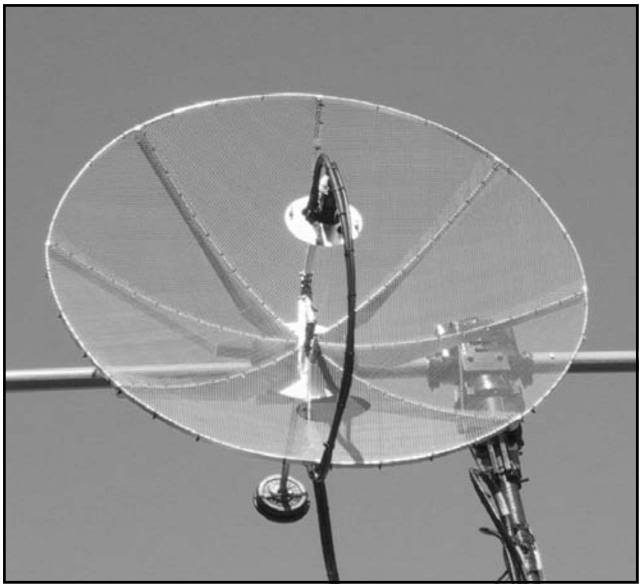

Fig 37—Prototype 1.2-meter dish by Rick Fletcher, KG6IAL, using a dual-band (L and S) patch-feed antenna for AO-40. See text. This kit dish is covered with 1/4-inch mesh. (*KG6IAL photo.*)

Another approach is the construction of a kit-type dish antenna, just becoming available in 1.2-meter and 1.8-meter diameters. This ingenious design by KG6IAL is available from his Web site **www.teksharp.com/**. **Fig 37** shows the prototype of the 1.2-meter dish with an f/D of 0.30. The 1.2-meter dish is fed with a dual-band patch feed for L and S bands. The 1.8-meter dish is designed for up to three bands using a tri-band patch feed for the U, L and S bands. This dish will permit U-band operation. A Central States VHF Society measurement on a similar sized dish (by W0LMD) with a patch feed showed a gain of about 17.1 dBic (actual measurement was 12.0 dBd linearly fed). This performance along with a small V-band (145 MHz) Yagi would permit a very modest satellite antenna assembly for all of the VHF/UHF LEO satellites, as well as AO-40.

The ingenuity of the design of the KG6IAL antenna is that it is constructed of robust, 1/8-inch-thick aluminum sheet that is numerically machined for the parabolic shape of the ribs. The backsides of the ribs are stiffened by a bent flange edge. The panel mesh is attached by using small tie-wraps or small aluminum wire through the mesh and holes provided along the parabolic edge. KG6IAL used 1/4-inch mesh in the prototype antenna to reduce wind loading. A single formed conduit post is provided in the kit for mounting the patch-feed assembly. The post extends rearward to permit the attachment of a counterweight, if needed.

AO-40 has also provided some additional challenges to the ham operator. Besides its well-known S-band downlink, AO-40 also has a K-band downlink in the range of

24.05 GHz. This quite low-powered transmitter has provided a substantial challenge to some operators, such as N1JEZ, K5OE, W5LUA, G3WDG and others. N1JEZ documented his work in *QST* while K5OE shows his K-band work on his Web site and in the *Proceedings of the AMSAT Space Symposium*. See **Fig 38**.

Parabolic Dish Antenna Construction

In the USA large numbers of dishes can be obtained either free or at low cost. But in some parts of the world dishes are not so plentiful, so hams make their own. **Fig 39** shows G3RUH's S-band dish antenna. There are three parts to the dish antenna—the parabolic reflector, the boom and the feed. There are as many ways to construct this as there are constructors. You need not slavishly replicate every nuance of the design. The only critical dimensions occur in the feed system. After construction, you will have a 60-cm diameter S-band RHCP dish antenna with a gain of about 20 dBi and a 3-dB beamwidth of 18°. Coupled with the proper downconverter, performance will be more than adequate for S-band downlink.

The parabolic reflector used for the original antenna was intended to be a lampshade. Several of these aluminum reflectors were located in department-store surplus. The dish is 585 mm in diameter and 110 mm deep, corresponding to an f/D ratio of 585/110/16 = 0.33 and a focal length of 0.33 × 585 = 194 mm. The f/D of 0.33 is a bit too concave for a simple feed to give optimal performance but the price was right, and the under-illumination keeps ground noise pickup to a minimum. The reflector already had a 40-mm hole in the center with three 4-mm holes around it in a 25-mm radius circle.

The boom passes through the center of the reflector and is made from 12.7-mm square aluminum tube. The boom must be long enough to mount to the rotator boom on the backside of the dish. The part of the boom extending through to the front of the dish must be long enough

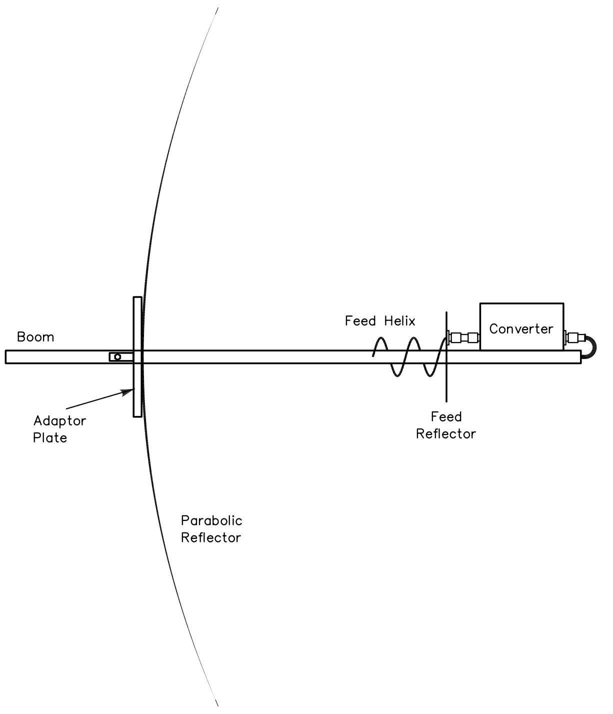

Fig 39—Detail of 60-cm S-band dish antenna with feed.

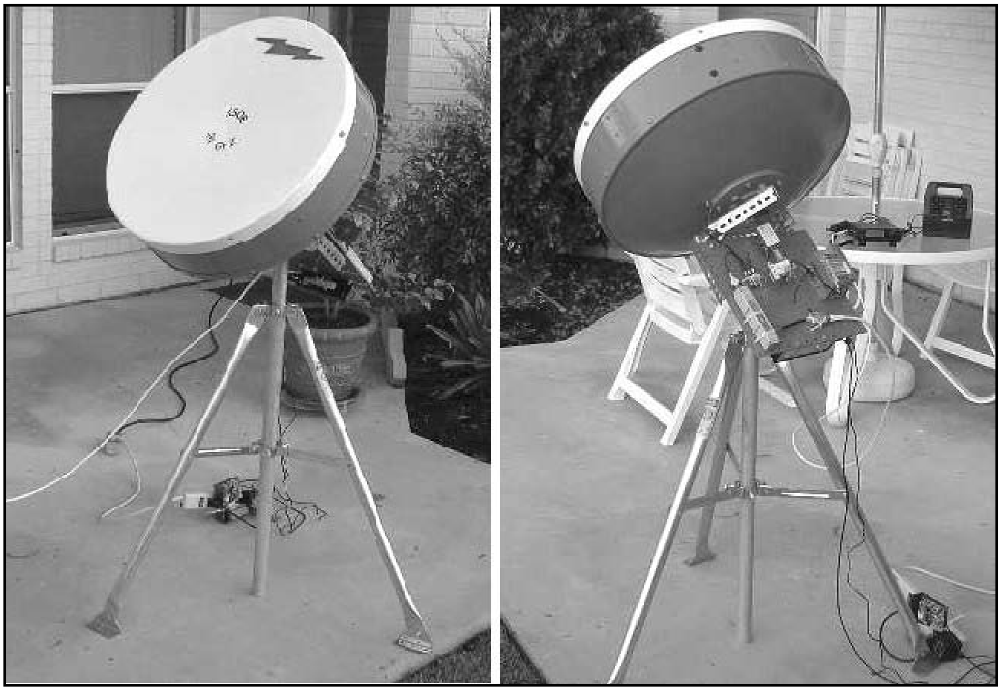

Fig 38—K5OE found this K-band dish on the Web and has set it up for the AO-40 K-band downlink. (*K5OE photo.*)

to mount the feed at the focus. If you choose to mount the downconverter or a preamp near the feed, some additional length will be necessary. Carefully check the requirements for your particular setup.

A 3-mm thick piece of aluminum, 65 mm in diameter, supports the boom at the center of the reflector. Once the center mounting plate is installed, the center boom is attached using four small angle brackets—two on each side of the reflector. See Fig 39 for details of reflector and boom assembly.

A small helix is used for the S-band antenna feed. The reflector for the helix is made from a 125-mm square piece of 1.6-mm thick aluminum. The center of the reflector has a 13-mm hole to accommodate the square center boom described above. The type-N connector is mounted to the reflector about 21.25 mm from the middle. This distance from the middle is, of course, the radius of a helical antenna for S-band. Mount the N connector with spacers so that the back of the connector is flush with the reflector surface. The helix feed assembly is shown in **Fig 40**.

Copper wire, or tubing, about 3.2 mm in diameter is used to form the helix. Wind four turns around a 40-mm diameter form. The turns are wound counterclockwise. This is because the polarization sense is reversed from RHCP when reflected from the dish surface. The wire helix will spring out slightly when winding is complete.

Once the helix is wound, carefully stretch it so that the turns are spaced 28 mm (±1 mm). Make sure the finished spacing of the turns is nice and even. Cut off the first half turn. Carefully bend the first quarter turn about 10° so it will be parallel to the reflector surface once the helix is attached to the N connector. This quarter turn will form part of the matching section.

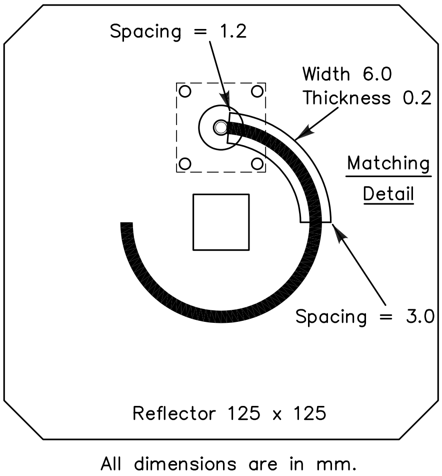

Fig 40—Details of helix feed for S-band dish antennas. The type-N connector is fixed with three screws and is mounted on a 1.6-mm spacer to bring the PTFE molding flush with the reflector. An easier mounting can be using a smaller TNC connector. Reflectors should be 95 to 100 mm in diameter.

Cut a strip of brass, 0.2 mm thick and 6 mm wide, and match the curvature of the first quarter turn of the helix, using a paper pattern. Be careful to get this pattern and subsequent brass cutting done exactly right. Using a large soldering iron and working on a heatproof surface, solder the brass strip to the first $^{1}/_{4}$ turn of the helix. Unless you are experienced at this type of soldering, getting the strip attached just right will require some practice. If it doesn't turn out right, just dismantle, wipe clean and try again.

After tack soldering the end of the helix to the type-N connector, the first $^{1}/_{4}$ turn, with its brass strip in place, should be 1.2 mm above the reflector at its start (at the N connector) and 3.0 mm at its end. Be sure to line up the helix so its axis is perpendicular to the reflector. Cut off any extra turns to make the finished helix have $2^{1}/_{4}$ turns total. Once you are satisfied, apply a generous amount of solder at the point the helix attaches to the N connector. Remember this is all that supports the helix.

Once the feed assembly is completed, pass the boom through the middle hole and complete the mounting by any suitable method. The middle of the helix should be at the geometric focus of the dish. In the figures shown here, the feed is connected directly to the downconverter and then the downconverter is attached to the boom. You may require a slightly different configuration depending on whether you are attaching a downconverter, preamp or just a cable with connector. Angle brackets may be used to secure the feed to the boom in a manner similar to the boom-to-reflector mounting. Be sure to use some method of waterproofing if needed for your preamp and/or downconverter.

Dish Feeds

W0LMD describes in **www.ultimatecharger.com/** that feeding a dish has two major factors that determine the efficiency. Like a flashlight bulb, the feed source should evenly illuminate the entire dish, and none of the feed energy should spillover outside the dish's reflecting surface. No feed system is perfect in illuminating a dish. Losses affect the gain from either under-illuminating or over-illuminating the dish (spillover losses). Typical dish efficiency is 50%. That's 3 dB of lost gain. A great feed system for one dish can be a real lemon on another. A patch feed system is very wide angle, but a helix feed system is narrow angle.

W0LMD has experimented with helical feeds for low f/D antennas ("deep" dishes) shown in **Fig 41**. A short-focal-ratio center-fed dish requires a wide-angle feed system to fully illuminate the dish, making the CP patch the preferred feed system. When used with an offset-fed dish, a patch-type feed system will result in a considerable spillover, or over-illumination loss, with an increased sensitivity to off-axis QRM, due to the higher f/D of this dish. Offset-fed

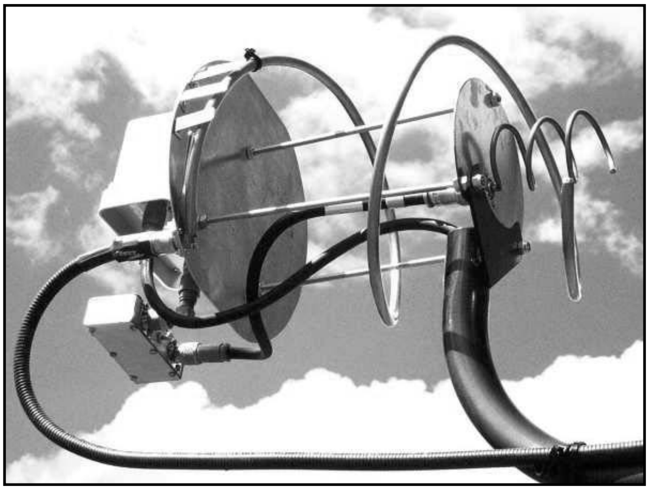

Fig 41—WØLMD's dual helix-dish feed for U and S bands. This early experimental feed was found to be wanting and he then turned to patch feeds for dishes. (*WØLMD photo.*)

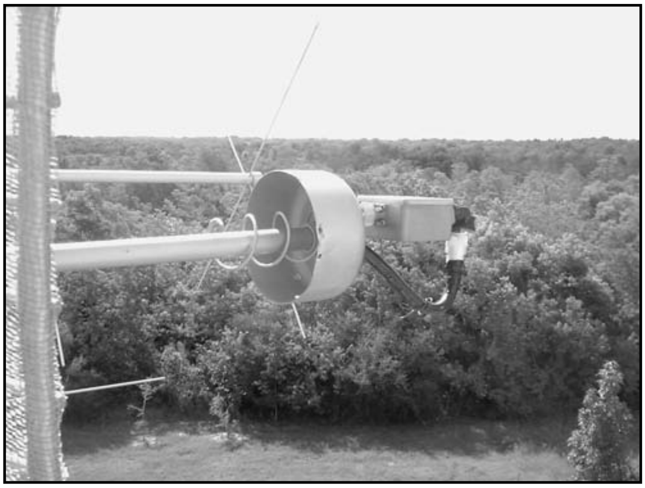

Fig 42—K5OE's helix feed for his MMDS S-band dish antenna. (*K5OE photo.*)

dishes do much better when fed with a helix antenna.

A helix feed is simplicity personified. Mount a type N connector on a flat reflector plate and solder a couple of turns wire to the inner terminal. Designs are anywhere from 2 to 6 turns. The two-turn helices are used for very short-focal-length dishes in the f/D = 0.3 region, and the 6-turn helices are used with longer-focal-length (f/D ~ 0.6) dishes, typically offset-fed dishes. Since AO-40 is right circular and the dish reflection will reverse the polarity, the helix should be wound left circular, looking forward from the connector. Helix feeds work poorly on the short-focal-length dishes but really perform well on the longer-focal-length offset-fed dishes. K5OE shows us the helix feed for his modified MMDS dish in **Fig 42**. This design employs the cupped reflector of W8JK.

A Helix Feed for an Offset-Dish Antenna

This section describes WD4FAB's surplus PrimeStar offset-fed dish antenna with a 7-turn helical feed antenna, shown in Fig 30. This S-band antenna can receive Sun noise 5 dB above sky noise. (Don't try to receive Sun noise with the antenna looking near the horizon, since terrestrial noise will be greater than 5 dB, at least in a big-city environment.) WD4FAB received the dish from NØNSV, who renamed the finished product the "FABStar."

The dish's reflector is a bit out of the ordinary, with the shape of a horizontal ellipse. It is still a single paraboloid, illuminated with an unusual feed horn. At 2401 MHz (S band) we can choose to under-illuminate the sides of the dish while properly feeding the central section, or over-illuminate the center while properly feeding the sides. WD4FAB chose to under-illuminate. The W1GHZ water-bowl measurements showed this to be a dish with a focal point of 500.6 mm and requiring a feed for an f/D = 0.79. The total illumination angle of the feed is 69.8° in the vertical direction and a feed horn with a 3-dB beamwidth of 40.3°. At 50% efficiency this antenna was calculated to provide a gain of 21.9 dBi. A 7-turn helical feed antenna was estimated to provide the needed characteristics for this dish and is shown in **Fig 43**.

The helix is basically constructed as described for the G3RUH parabolic dish above. A matching section for the first $\lambda/4$ turn of the helix is spaced from the reflector at 2 mm at the start and 8 mm at the end of that fractional turn. Modifications of the G3RUH design include the addition of a cup reflector, a design feature used by the originator of the helical antenna, John Kraus, W8JK. For the reflector, a 2-mm thick circular plate is cut for a 94 mm (0.75 λ) diameter with a thin aluminum sheet metal cup, formed with a depth of 47 mm. Employment of the cup enhances the performance of the reflector for a dish feed, as shown by K5OE. (See the K5OE material on the CD-ROM accompanying this book.)

The important information for this 7-turn helical antenna is:

- Boom: 12.7-mm square tube or "C" channel.
- Element: $^1/_8$-inch diameter copper wire or tubing.

Close wind the element on a circular 1.50-inch tube or rod; the finished winding is 40 mm in diameter and spaced to a helical angle of 12.3°, or 28 mm spacing. These dimensions work out for an element circumference of 1.0 λ about the center of the wire.

When WD4FAB tackled this antenna, he felt that the small number of helical element supports used by G3RUH would be inadequate, in view of the real-life bird traffic on the antennas at his QTH. He chose to use PTFE (Teflon) support posts every $^1/_2$ turn. This closer spacing of posts permitted a careful control of the helix-winding diameter and spacing and also made the antenna very

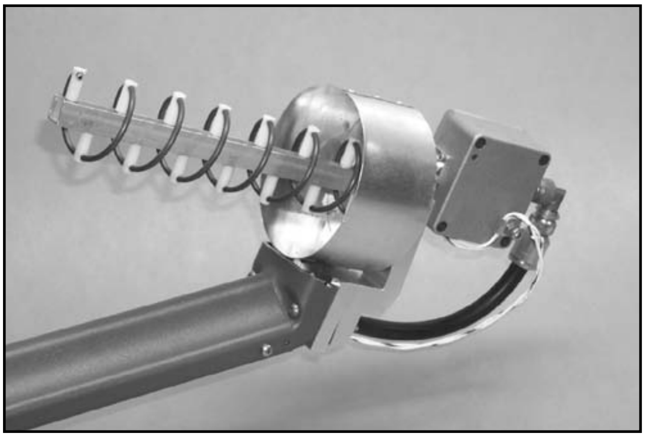

Fig 43—Seven-turn LHCP helix feed for an offset dish, long f/D, antenna, with DEM preamp. (*WD4FAB photo.*)

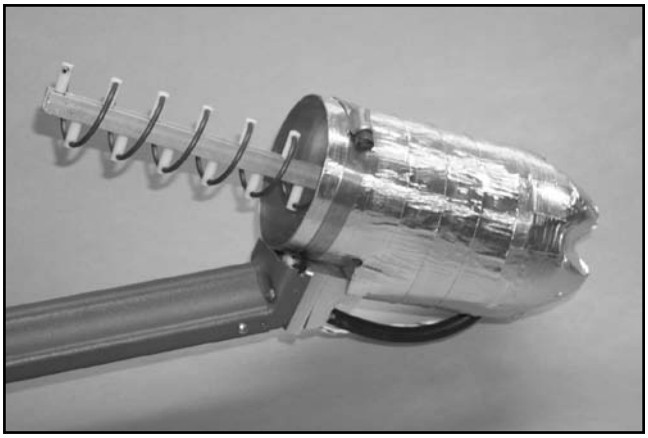

Fig 45—Rain cover for preamp using a two-liter soft-drink bottle with aluminum foil tape for protection from sun damage. (*WD4FAB photo.*)

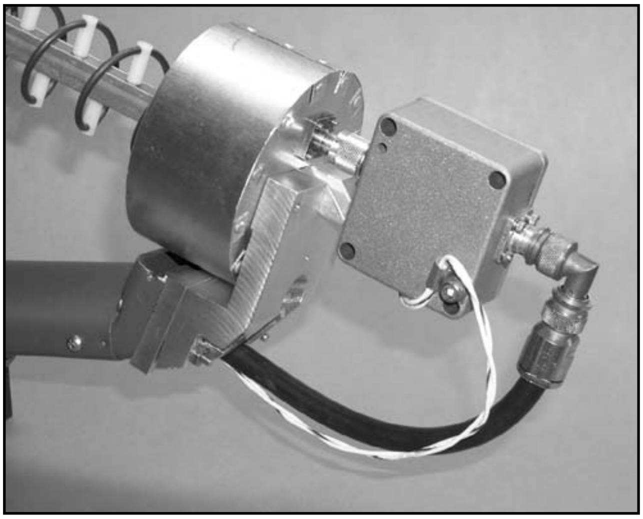

Fig 44—Mounting details of seven-turn helix and preamp. (*WD4FAB photo.*)

robust. He set up a fixture on the drill press to uniformly predrill the holes for the element spacers and boom. Attachment of the reflector is through three very small aluminum angle brackets on the element side of the boom.

Mounting of the helix to the dish requires modification of the dish's receiver-mounting boom. **Fig 44** shows these modifications using a machined mount. NM2A constructed one of these antennas and showed that a machine shop is not needed for this construction. He made a "Z" shaped mount from aluminum-angle plate and then used a spacer from a block of acrylic sheet. The key here is to get the dish focal point at the 1.5-turn point of the feed antenna, which is also at about the lip of the reflector cup.

The W1GHZ data for this focal point is 500.6 mm from the bottom edge of the dish and 744.4 mm from the top edge. A two-string measurement of this point can confirm the focal point, as shown by Wade in his writings. When mounting this feed antenna the constructor must be cautious to aim the feed at the beam-center of the dish, and not the geometric center, as the original microwave horn antenna was constructed. Taking the illumination angle information noted above, the helical feed antenna should be aimed 5.5° down from the geometric center of the dish.

As illustrated in Fig 44, a DEM preamp was directly mounted to the feed helix, using a TNC female connector on the helix, chosen for this case, since N connectors are quite large for this antenna. A male chassis connector should be mounted on the preamp so that the preamp can be directly connected to the antenna without any adaptors. This photo also illustrates how the reflector cup walls were riveted to the reflector plate.

Exposed connectors must be protected from rainwater. Commonly materials such as messy Vinyl Mastic Pads (3M 2200) or Hand Moldable Plastic (Coax Seal) are used. Since this is a tight location for such mastic applications, a rain cover was made instead from a 2-liter soft-drink bottle, **Fig 45**. Properly cutting off the top of the bottle allows it to be slid over the helix reflector cup and secured with a large hose clamp. You must provide UV protection for the plastic bottle and that was done with a wrapping of aluminum foil pressure-sensitive adhesive tape.

There are many methods for mounting this dish antenna to your elevation boom. You must give consideration to the placement of the dish to reduce the wind loading and off-balance to the rotator system. In WD4FAB's FABStar installation, the off-balance issue was not a major factor, as the dish was placed near the center of the elevation boom, between the pillow-block bearing supports. Since there is already a sizeable aluminum plate for these bearings, the dish was located to "cover" part of that plate, so as to not add measurably to the existing wind-loading area of the overall assembly.

A mounting bracket provided with the stock dish clamps to the end of a standard 2-inch pipe stanchion (actual measure: 2.38 inches in diameter). This bracket was turned around on the dish and clamped to the leg of a welded-pipe Tee assembly. See **Fig 46**. Pipe-reducing fittings were machined and fitted in the Tee-top bar, which was sawn in half for clamping over the $1\frac{1}{2}$ inch pipe used for the elevation boom. Bolts were installed through drilled holes and used to clamp this assembly.

Patch Feeds for Dish Antennas

Patch feeds are almost as simple as helix feeds. A patch is typically an N connector on a flat reflector plate with a tuned flat-metal plate soldered to the inner terminal. Sometimes the flat plate is square; sometimes it is rectangular; sometimes it is round. It could have two feed points, 90° out of phase for circular polarization, as used in the construction of the AO-40 U-band antennas. Some patches are rectangular with clipped corners to create a circular radiation pattern.

On 2401 MHz, the plate is 57 mm square and spaced 3 mm away from the reflector. The point of attachment is about halfway between the center and the edge. A round patch for 2401 MHz is about 66 mm in diameter. These patches work well on the shorter focal length center-fed MMDS and TVRO dishes. G3RUH made a CP patch feed for these short f/D dishes, shown in Fig 31 and **Fig 47**.

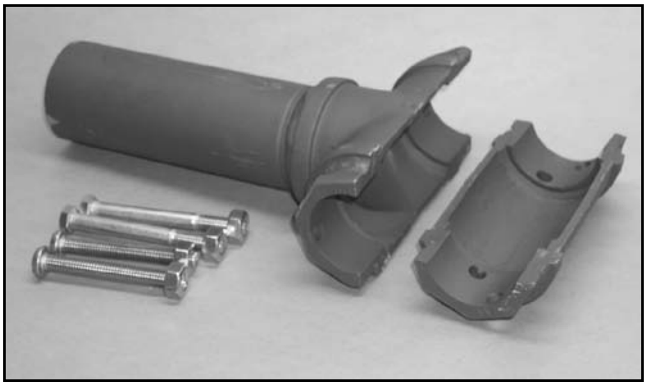

Fig 46—Welded pipefitting mount bracket for FABStar dish antenna. (*WD4FAB photo.*)

Robert, W0LMD, has done a considerable amount of experimenting with patch feeds for his dish antennas. One tri-band feed is shown in **Fig 48**. These are circular patches that have CP properties through the arrangement of the feed point and a small piston-variable capacitor that is offset from the feed point. **Fig 49** shows some of the many patches that Robert has created for his trials.

A No-Tune Dual-Band Feed for Mode L/S

Jerry, K5OE, notes that the AO-40 transponder has two uplink receivers active most of the time for CW/SSB activity. Most operators use U band at 435 MHz (70 cm). Also available, however, are two L-band (23-cm) receivers: L1 at 1269 MHz and L2 at 1268 MHz. The reasons for going to L band can be varied, but there is no arguing the benefits in reduced antenna size and AGC suppression. The types of L-band antennas are varied as well. Many use helices. Others use beams and arrays of beams. Still others use dishes, small and large.

K5OE recently acquired an old UHF TV dish measuring 1.2 meters in diameter. He wanted to use it both to receive on S band at 2401 MHz (13cm) and to transmit on the uplink on L band. He covered it with aluminum mesh and built a dual-helix feed for it, but was unhappy with the L-band performance. It seems the concentric helices interacted with each other substantially. Having had good success with patch feeds on S band, he designed, built and installed a dual-patch feed on a 1.5-meter solid dish for Field Day 2002. This arrangement worked superbly on uplink (with 25 W), but was embarrassingly deaf on receive. This second dual-band feed failure led him to experiment for months with different configurations, leading ultimately to the design presented here. The project goals were:

- Good performance on both S-band receive and L-band uplink.
- An easy-to-produce model using common hardware and simple hand tools.

Patches are better than helices as dish feeds. This revelation came to K5OE while doing investigation and

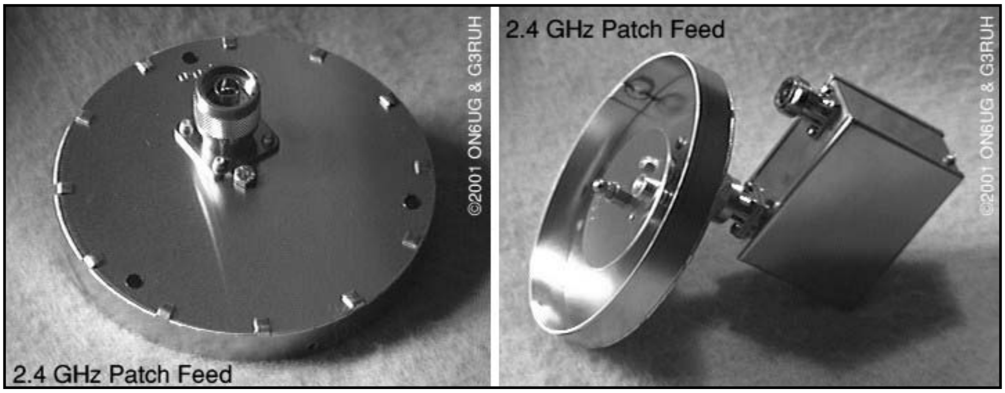

Fig 47—Details of CP-patch feed for short f/D dish antennas by G3RUH and ON6UG.

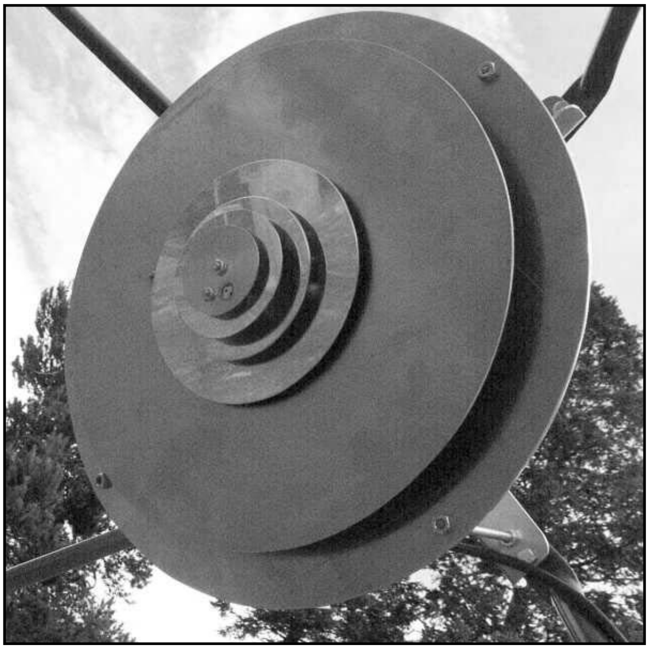

Fig 48—A triband (U, L and S bands) patch-CP feed for large dish antennas for AO-40 service. (*W0LMD photo.*)

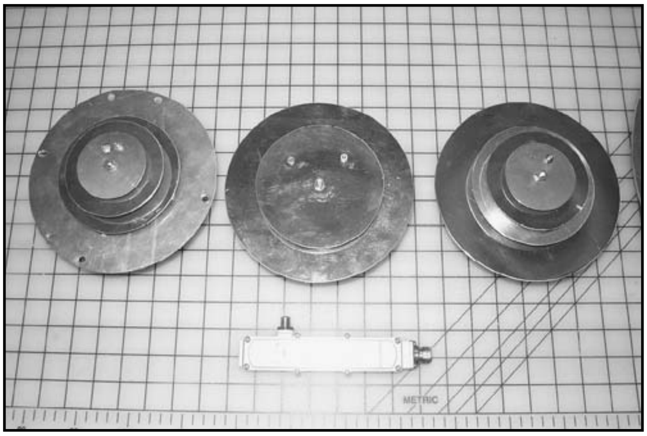

Fig 49—Some of the many experimental CP-patch-feed antennas by W0LMD. (*W0LMD photo.*)

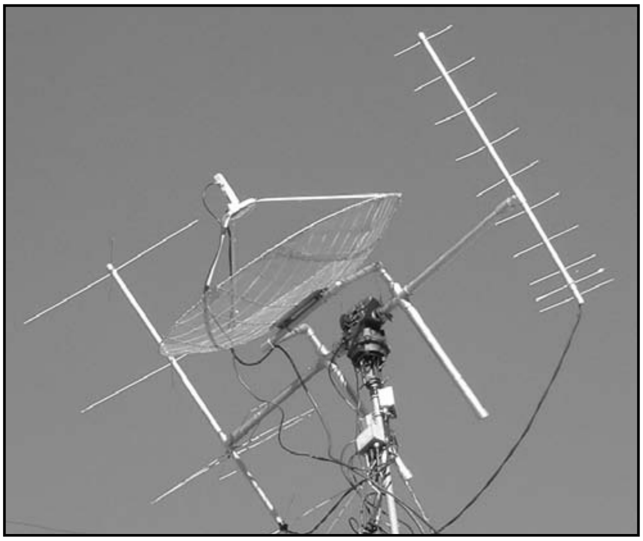

Fig 50—The 1.2-meter dish with dual-band patch feed installed. (*K5OE photo; courtesy of* The AMSAT Journal.)

experimenting with helix antennas. In the middle of this investigative foray, he saw the radiation pattern for the G3RUH patch feed published on James Miller's web site. When he modeled that pattern and input it into the W1GHZ feed pattern program, it produced an amazing 72% efficiency. The best helix he ever modeled has about 60% efficiency. I8CVS recently ran his own antenna range tests of a design similar to the G3RUH patch and produced a similarly impressive pattern.

Then K5OE came across the *truncated corners* square patch design popularized by K3TZ. This AO-40 design here is attributed to 7N1JVW, JF6BCC and JG1IIK. There are references in the literature going back over a decade for this now-common commercial design. The first model K5OE built outperformed his best helix-in-cup design by a full S unit (delta over the noise) on his FT-100 portable setup. Compared to a helix, the patch simply has better illumination efficiency with less spillover from side lobes.

Patch theory is beyond the scope of this article, but can be summarized as building a shape that resonates at the desired frequency, compensated in size by the capacitive inductance between itself and the reflector. A patch can be practically any shape since it basically acts like a parallel-plate transmission line. Current in the patch flows from the feed point to the outer edge(s), where all the radiation occurs. The reputed, but often disputed, circularity of the truncated corner patch is accomplished by effectively designing two antennas into the patch element (of two different diagonal lengths) and feeding them 90° out of phase.

For K5OE's 1.2-meter dish, shown in **Fig 50**, computations predicted 21-dBi gain on L band and almost 27 dBi on S band, with an assumed 50% efficiency:

$$G = 10\log_{10}\left[\eta A\left(\frac{4\pi}{\lambda^2}\right)\right] \qquad \text{(Eq 8)}$$

where

η = efficiency

λ = wavelength in meters.

A = aperture of the dish in meters = $\pi \times r^2$

r = dish radius in meters = diameter/2 in meters = 0.6 meters

At 1269 MHz, λ= 300/1269 = 0.236 meters:

$$G = 10\log_{10}\left[0.50 \times \left(3.14 \times 0.6^2\right) \times \frac{4 \times 3.14}{0.236^2}\right] = 21.1\,\text{dBi}$$

At 2401.5 MHz, λ = 300/2401.5 = 0.125 meters:

$$G = 10\log_{10}\left[0.50 \times \left(3.14 \times 0.6^2\right) \times \frac{4 \times 3.14}{0.125^2}\right] = 26.6\,\text{dB}^2$$

Where does the feed get mounted? The *focal point* is where the parabolic shape of the dish concentrates the reflected signal. In K5OE's case the antenna was placed flat on the garage floor to measure the depth:

$f = D^2 / 16d$ (Eq 9)

where

D = diameter of the dish in inches
d = depth of the dish in inches
$f = 48^2 / (16 \times 7.25) = 19.8$ inches (50.5 cm)

This is just one example of countless combinations of hardware and patch designs. Inherent in this design, however, are five key design and construction features developed from building and empirical testing of a number of patch feeds.

1. The specified dimensions are critical for no-tune operation. **Fig 51** shows the dimensions necessary to build the dual-feed patch. (K5OE recommends you reproduce this sketch accurately on graph paper. When you cut your patches you can lay them on the paper template for checking.) Repeat: These dimensions are critical. Even a 0.5-mm error will throw your resonance off considerably—patches are not broadband.

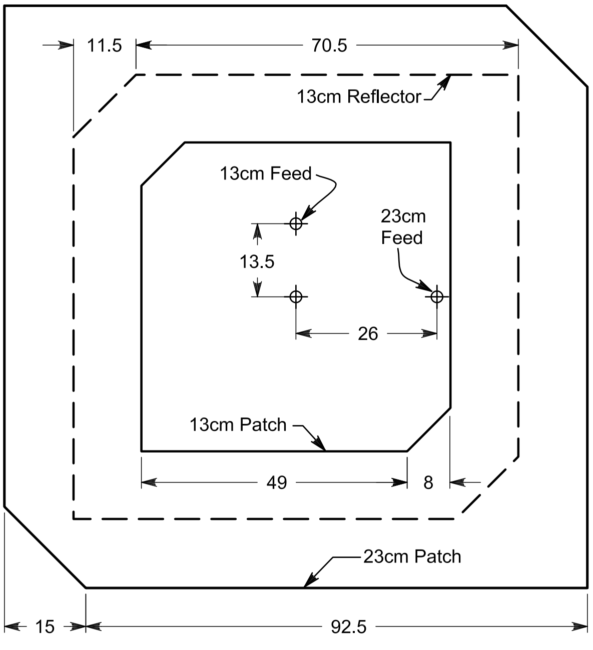

Fig 51—Dual-band patch feed dimensions, in millimeters. (*K5OE diagram; courtesy of* The AMSAT Journal.)

2. The reflector must be rigid. Spacing between the driven element (patch) and the reflector affects the resonant frequency. K5OE found 0.025-inch aluminum sheet and 26-gauge copper sheet acceptable for a single S-band patch feed, but too flimsy for an L-band reflector. Use more rigid material or provide additional stiffening for the L-band reflector, as shown in Figs **52**, **53** and **54**.
3. The patches must be electrically isolated from each other. A metallic center support works for a single patch but creates harmonic-coupling problems when patches are stacked for multiband use. The use of nylon machine screws and nuts helps solve the vexing problem of the S-band patch coupling to the L-band patch.
4. The "straight corners" of the truncated corner patch must be kept clear of any nearby metal. This includes the edges of the feed support or cup, if used. See Fig 54.
5. Feeding the patches at 90° to each other minimizes the electromagnetic interaction between the two antenna fields.

One final design issue deals with the first harmonic of the L-band antenna. You must significantly reduce the potentially destructive effect from the 1269-MHz signal's second harmonic. Severe desense of your receive signal could occur and potentially even overload and damage the first active device in your system. Sensitive preamps and downconverters without a pre-RF-amplifier filter will need an external filter. K5OE has used a G3WDG stub filter rated at 100-dB rejection with good success ahead of his preamp. His current setup, however, uses a AIDC-3731AA downconverter with its internal combline filter providing adequate filtering. Using the downconverter directly at the

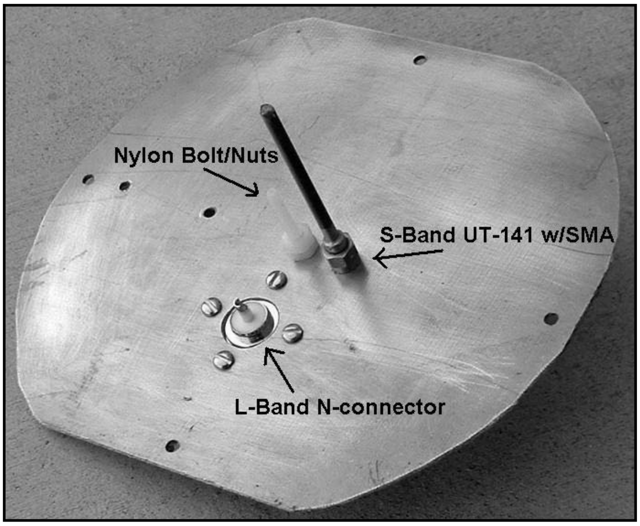

Fig 52—Assembly of the L band reflector. (*K5OE photo; courtesy of* The AMSAT Journal.)

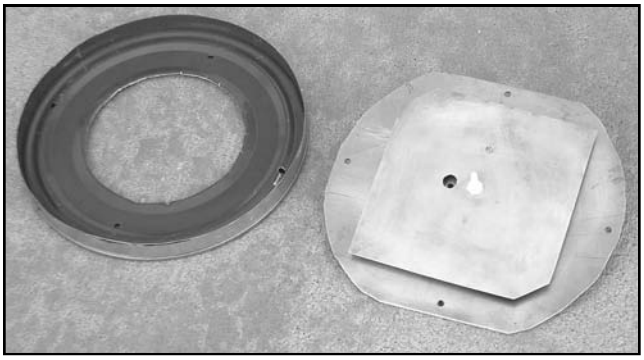

Fig 53—The support, L-band reflector and patch. (*K5OE photo; courtesy of* The AMSAT Journal.)

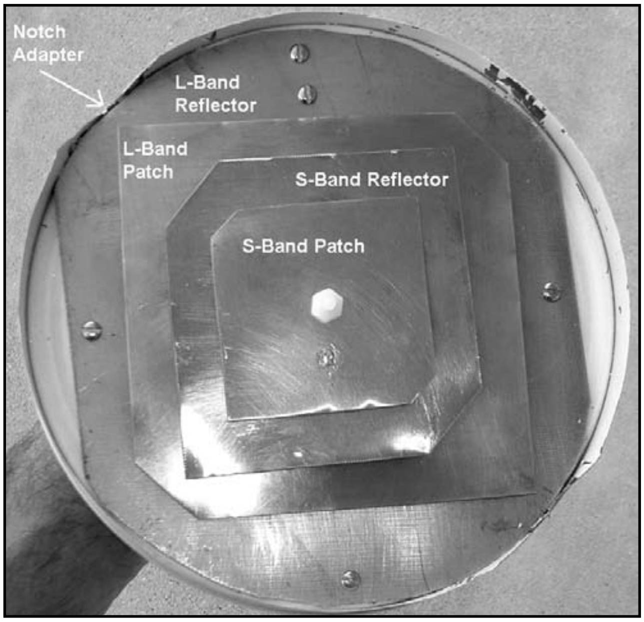

Fig 54—The completed dual-band patch feed. (*K5OE photo; courtesy of* The AMSAT Journal.)

feed point has a noise figure (NF) of 1.0 dB, compared to the cumulative NF of 1.6 dB using a filter and a preamp.

Construction of the feed begins with selection of material for both the electrical parts (the antennas) and the mechanical parts (the support structure). The L-band antenna is constructed using a 6 × 6-inch double-sided circuit board for the reflector and a piece of 26-gauge copper sheet for the driven element (patch). A flanged female type-N connector is used for the feed connection. The S-band antenna is constructed of two pieces of 26-gauge copper sheeting and the feed connection is made with a short piece of UT-141 (0.141-inch copper-clad semirigid coax) terminated in a male SMA fitting. Fig 52 illustrates the assembly of the L-band reflector with the nylon-center support bolt, the L-band N-connector, and the S-band semirigid coax terminated onto an SMA-to-N adapter through the circuit board.

The support structure began life as a paint can, measuring 155 mm in diameter. It was cut down to a 15-mm depth. Cut a hole in the middle of the bottom of the can and trim the PC board to fit inside the can bottom. Use stainless-steel $^3/_8$-inch 4-40 bolts, washers and nuts to secure the PC board to the can bottom. A $1^1/_2$-inch 6-32 nylon bolt is secured through the center of the PC board with two nylon nuts to provide the 6-mm spacing for the L-band patch. Fig 53 shows the L-band patch in position and ready to be soldered to the N-connector. Note the hole through the L-band patch allowing the S-band UT-141 coax to pass (without making contact).

The remainder of the antenna is then assembled in order: First the L-band patch is secured with two nylon nuts and soldered to the N-connector. Then the S-band reflector is secured with one nylon nut to provide 3-mm spacing, and the UT-141 coax shield is soldered to the S-band reflector. Finally, the S-band patch is secured with a single nylon nut (3-mm spacing) and soldered to the center conductor of the UT-141 coax. To summarize the overall order of assembly: L-band reflector, two nylon nuts, L-band patch, two nylon nuts, S-band reflector, one nylon nut, S-band patch, and one nylon nut.

An electrical check with an ohmmeter of the completed feed should show the two reflectors connected, with the patches isolated from the reflectors and from each other. Fig 54 shows the completed feed. Note how the sides of the support are cut out to avoid proximity to the L-band patch and how the L-band and S-band patches are at 90° to each other. **Fig 55** shows the back of the feed, complete with an angle support for the downconverter. The flanged N-connector is for the L-band coax and the male-N adapter is secured from the other side of the feed with the SMA fitting on the UT-141 coax.

For those who are tempted to tune the patch, K5OE recommends doing it with the feed installed on the antenna—since the dish surface affects the feed-point impedance slightly. The feed-point impedance, and thus the resonant frequency, can be changed quite a bit by adjustment of the spacing of just the straight corners. There is no need to change the spacing at the center or the feed—just a slight up or down bending of the straight corners will change the tuning. Do this carefully: a little bit goes a long way. This patch design is very repeatable and will work adequately (an SWR below 1.5:1) with no adjustments.

The antenna performs to the calculated predictions above. On receive, this antenna is 4 S units better than K5OE's 45-cm dish and 3 S units above his 65-cm dish (both other dishes have similar patch feeds and the same downconverter). It also clearly outperforms his previous dual-helix arrangement on the 1.2-meter dish, but he was unable to do a side-by-side comparison.

On transmit, it does equally well, with a decent signal into the satellite with only 10 W measured at the

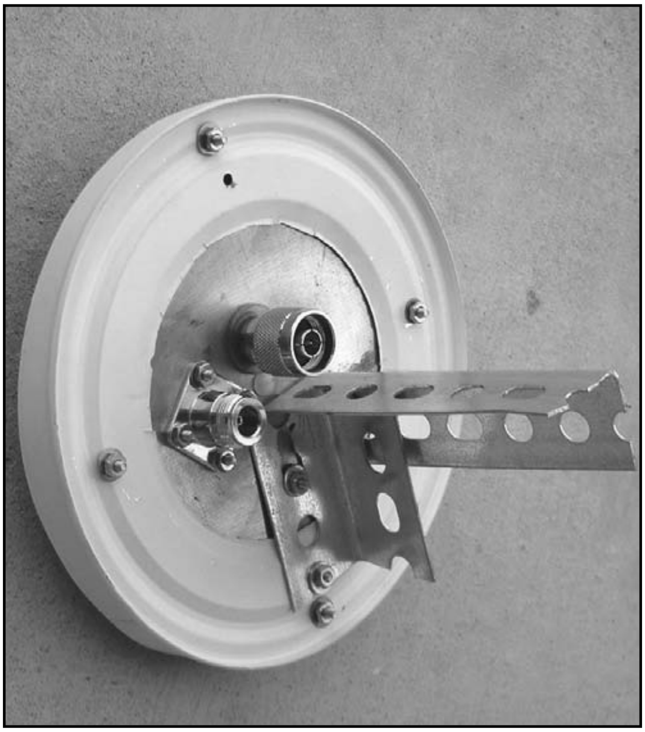

Fig 55—Rear of the completed feed. (K5OE photo; courtesy of *The AMSAT Journal.*)

antenna. The L band is noticeably improved over the helix predecessor. At low squint angles K5OE finds the L-band uplink to be about 1 S unit weaker than his U-band uplink. He later added a small plastic hat to extend over the top of the patches to keep the rain and bird droppings off—both detune the patches when built up between the patch and the reflector.

Though simple and effective, this is merely one way to construct a dual feed. Cookie-tin lids also make excellent supports. Tin snips are a good investment and much easier to use than a hacksaw. Use a flat file to remove burrs from the edges of the patches. Use stainless-steel hardware, most notably $^3/_8$-inch 4-40 machine bolts and nuts for the antenna hardware and $^1/_2$-inch 6-32 for the support-structure connections to the support arms ($^1/_2$-inch aluminum tubing). The copper sheet is much easier to solder to than aluminum. Once completed, the feed received a few coats of white enamel paint to protect the copper and to minimize the visual reflections.

This is not the only dual-band antenna on AO-40. There are many varied, innovative designs available, including G6LVB's simple and effective 1.2-meter homebrew stressed *chicken wire* dish with a dual-G3RUH helix feed. G3WDG has a 3-meter dish with L/S-band helices and a K-band (1.3-cm) feed horn, and W0LMD has developed some popular dual- and tri-band "round" patch feeds. (See the Notes and References, as well as the CD-ROM bundled with this book.)

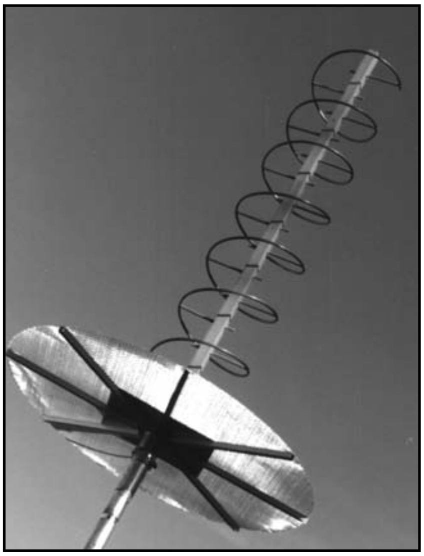

Fig 56—The portable 435-MHz helix assembled and ready for operation. (*W0CY photo.*)

For additional information on constructing antennas, feeds and equipment techniques for use at microwave frequencies, see *The ARRL UHF/Microwave Experimenter's Manual* and *The ARRL UHF/Microwave Projects Manual*. Both of these books have a wealth of information for the experimenter.

PORTABLE HELIX FOR 435 MHZ

Helical antennas for 435 MHz are excellent uplinks for U-band satellite communications. The true circular polarization afforded by the helix minimizes signal *spin fading* that is so predominant in these applications. The antenna shown in **Fig 56** fills the need for an effective portable uplink antenna for OSCAR operation. Speedy assembly and disassembly and light weight are among the benefits of this array. This antenna was designed by Jim McKim, W0CY.

As mentioned previously, the helix is about the most tolerant of any antenna in terms of dimensions. The dimensions given here should be followed as closely as possible, however. Most of the materials specified are available in any well supplied do-it-yourself hardware or building supply store. The materials required to construct the portable helix are listed in **Table 1**.

The portable helix consists of eight turns of $^1/_4$-inch soft-copper tubing spaced around a 1-inch fiberglass tube or maple dowel rod 4 feet, 7 inches long. Surplus aluminum jacket Hardline can be used instead of the copper tubing if necessary. The turns of the helix are supported by 5-inch lengths of $^1/_4$-inch maple dowel mounted through the 1-inch rod in the center of the antenna. **Fig 57A** shows the overall dimensions of the antenna. Each of these support dowels has a V-shaped notch in the end to locate the tubing, as shown in Fig 57B.

The rod in the center of the antenna terminates at the feed-point end in a 4-foot piece of 1-inch ID galva-

Table 1

Parts List for the Portable 435-MHz Helix

Qty	*Item*
1	Type N female chassis mount connector
18 feet	1/4-in. soft copper tubing
4 feet	1-inch ID galvanized steel pipe
1	5 feet × 1-inch fiberglass tube or maple dowel
14	5-inch pieces of 1/4-inch maple dowel (6 feet total)
1	1/8-inch aluminum plate, 10 inches diameter
3	2 × 3/4-inch steel angle brackets
1	30 × 30-inch (round or square) aluminum screen or hardware cloth
8 feet	1/2 × 1/2 × 1/2-inch aluminum channel stock or old TV antenna element stock
3	Small scraps of Teflon or polystyrene rod (spacers for first half turn of helix)
1	1/8 × 5 × 5-inch aluminum plate (boom-to-mast plate)
4	1½-inch U bolts (boom-to-mast mounting)
3 feet	#22 bare copper wire (helix turns to maple spacers)

Assorted hardware for mounting connector, aluminum plate and screen, etc.

nized steel pipe. The pipe serves as a counterweight for the heavier end of the antenna. The 1-inch rod material inside the helix must be nonconductive. Near the point where the nonconductive rod and the steel pipe are joined, a piece of aluminum screen or hardware cloth is used as a reflector screen.

If you have trouble locating the 1/4-inch soft copper tubing, try a refrigeration supply house. The perforated aluminum screening can be cut easily with tin snips. This material is usually supplied in 30 × 30-inch sheets, making this size convenient for a reflector screen. Galvanized 1/4-inch hardware cloth or copper screen could also be used for the screen, but aluminum is easier to work with and is lighter.

A 1/8-inch-thick aluminum sheet is used as the support plate for the helix and the reflector screen. Surplus rack panels provide a good source of this material. **Fig 58** shows the layout of this plate.

Fig 59 shows how aluminum channel stock is used to support the reflector screen. (Aluminum tubing also works well for this. Discarded TV antennas provide plenty of this material if the channel stock is not available.) The screen is mounted on the bottom of the 10-inch aluminum center plate. The center plate, reflector screen and channel stock are connected together with plated hardware or pop rivets. This support structure is very sturdy. Fiberglass tubing is the best choice for the center rod material although maple dowel can be used.

Mount the type-N connector on the bottom of the center plate with appropriate hardware. The center pin should be exposed enough to allow a flattened end of the copper tubing to be soldered to it. Tin the end of the tubing after it is flattened so that no moisture can enter it. If the helix is to be removable from the ground-plane screen, do not solder the copper tubing to the connector. Instead, prepare a small block of brass, drilled and tapped at one side for a 6-32 screw. Drill another hole in the brass block to accept the center pin of the type-N connector, and solder this connection. Now the connection to the copper tubing helix can be made in the field with a 6-32 screw instead of with a soldering iron.

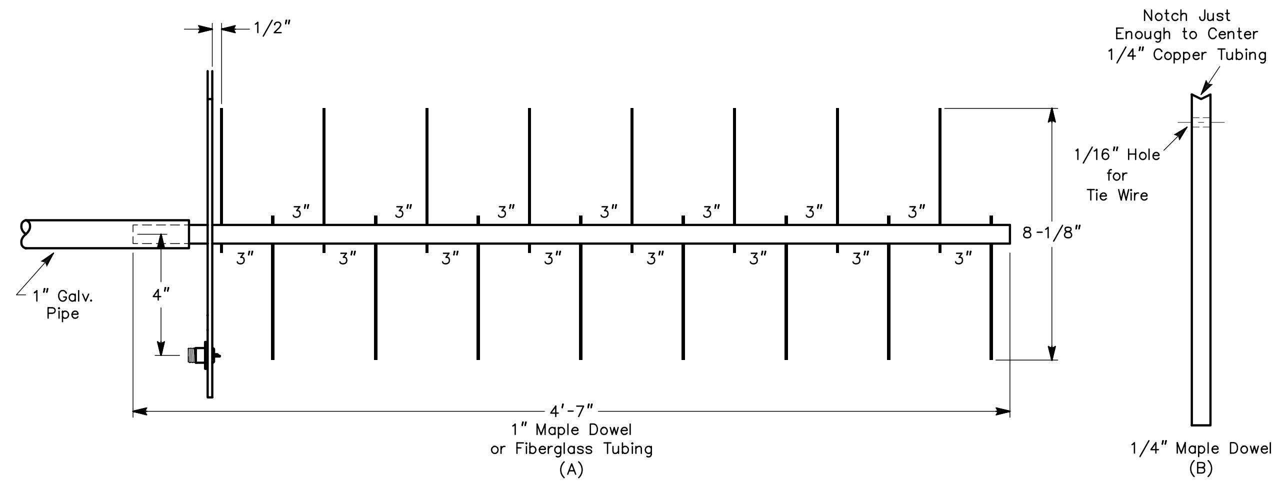

Fig 57—At A, the layout of the portable 435-MHz helix is shown. Spacing between the first 5-inch winding-support dowel and the ground plane is 1/2 inch; all other dowels are spaced 3 inches apart. At B, the detail of notching the winding-support dowels to accept the tubing is shown. As indicated, drill a 1/16-inch hole below the notch for a piece of small wire to hold the tubing in place.

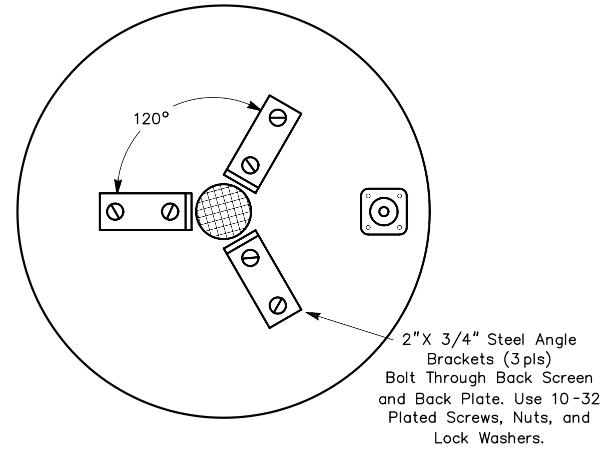

Fig 58—The ground plane and feed-point support assembly are shown. The circular piece is a 10-inch diameter, $^{1}/_{8}$-inch thick piece of aluminum sheet. (A square plate may be used instead.) Three 2 × $^{3}/_{4}$-inch angle brackets are bolted through this plate to the backside of the reflector screen to support the screen on the pipe. The type-N female chassis connector is mounted in the plate 4 inches from the 1-inch diameter center hole.

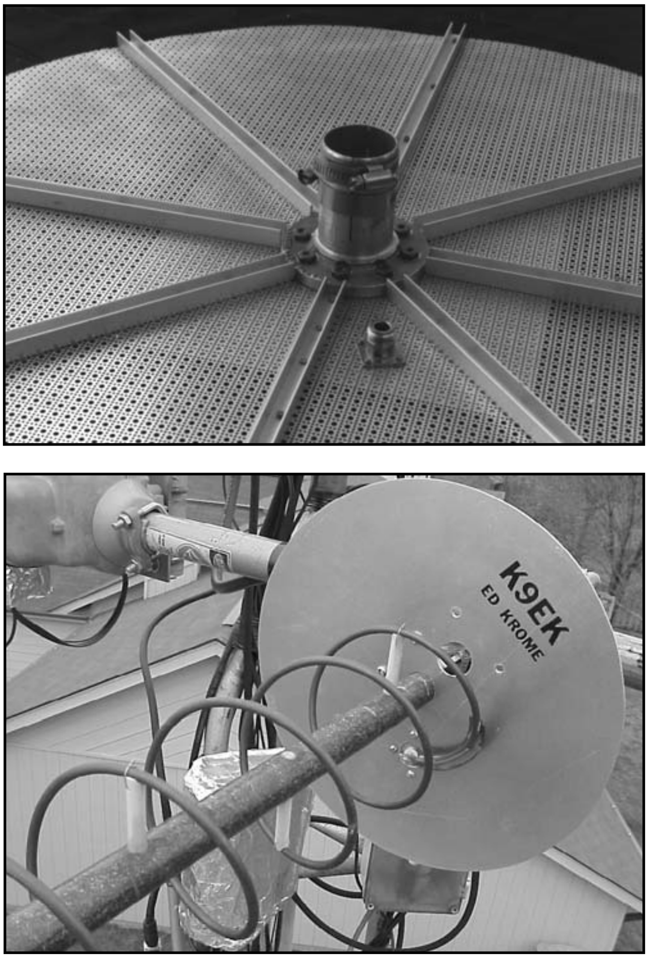

Fig 59—At top, the method of reinforcing the reflector screen with aluminum channel stock is shown. In this version of the antenna, the three angle brackets of Fig 58 have been replaced with a surplus aluminum flange assembly. (*W0CY photo.*) At bottom, this helix view shows the details of a $^{1}/_{4}$-turn matching transformer, as discussed in the text. (*K9EK photo.*)

Refer to Fig 57A. Drill the fiberglass or maple rod at the positions indicated to accept the 5-inch lengths of $^{1}/_{2}$-inch dowel. (If maple doweling is used, the wood must be weatherproofed as described below before drilling.) Drill a $^{1}/_{16}$-inch hole near the notch of each 5-inch dowel to accept a piece of #22 bare copper wire. (The wire is used to keep the copper tubing in place in the notch.) Sand the ends of the 5-inch dowels so the glue will adhere properly, and epoxy them into the main support rod.

Begin winding the tubing in a clockwise direction from the reflector screen end. First drill a hole in the flattened end of the tubing to fit over the center pin of the type-N connector. Solder it to the connector, or put the screw into the brass block described earlier. Carefully proceed to bend the tubing in a circular winding from one support to the next.

See the earlier section entitled "50-Ω Helix Feed" and Figs 19 and 20 to see how the first half-turn of the helix tubing must be positioned close above the reflector assembly. **Fig 59B** shows also an excellent example by K9EK on matching his U-band helical antenna to a 52-Ω feed line. It is important to maintain this spacing, since extra capacitance between the tubing and ground is required for impedance-matching purposes.

Insert a piece of #22 copper wire in the hole in each support as you go. Twist the wire around the tubing and the support dowel. Solder the wire to the tubing and to itself to keep the tubing in the notches. Continue in this way until all eight turns have been wound. After winding the helix, pinch the far end of the tubing together and solder it closed.

Weatherproofing the Wood

A word about preparing the maple doweling is in order. Wood parts must be protected from the weather to ensure long service life. A good way to protect wood is to boil it in paraffin for about half an hour. Any holes to be drilled in the wooden parts should be drilled after the paraffin is applied, since epoxy does not adhere well to wood after it has been coated with paraffin. The small dowels can be boiled in a saucepan. Caution must be exercised here—the wood can be scorched if the paraffin is too hot. Paraffin is sold for canning purposes at most grocery stores. Wood parts can also be protected with three or four coats of spar varnish. Each coat must be allowed to dry fully before another coat is applied.

The fiberglass tube or wood dowel must fit snugly with the steel pipe. The dowel can be sanded or turned down to the appropriate diameter on a lathe. If fiberglass is used, it can be coupled to the pipe with a piece of wood dowel that fits snugly inside the pipe and the tubing. Epoxy the dowel splice into the pipe for a permanent connection.

Drill two holes through the pipe and dowel and bolt them together. The pipe provides a solid mount to the boom of the rotator, as well as most of the weight needed to counterbalance the antenna. More weight can be added to the pipe if the assembly is "front-heavy." (Cut off some of the pipe if the balance is off in the other direction.)

The helix has a nominal impedance of about 105 Ω in this configuration. By varying the spacing of the first half turn of tubing, a good match to 52-Ω coax should be obtainable. When the spacing has been established for the first half turn to provide a good match, add pieces of polystyrene or Teflon rod stock between the tubing and the reflector assembly to maintain the spacing. These can be held in place on the reflector assembly with silicone sealant. Be sure to seal the type-N connector with the same material.

Exposed Antenna Relays and Preamplifiers

For stations using crossed Yagi antennas for CP operation, one feature that has been quite helpful for communicating through most of the LEO satellites, has been the ability to switch polarization from RHCP to LHCP. In some satellite operation this switchable CP ability has been essential. Operation through AO-40 has not shown a great need for such CP agility, since if the satellite is seriously off-pointed the signals are not particularly useable. When AO-40's squint angle is less than 25° the need for LHCP has not been observed. For those using helical antennas or helical-fed dish antennas, we just would not have the choice to switch CP unless an entirely new antenna is added to the cluster for that purpose. Not many of us have the luxury of that kind of space available on our towers.

For stations with switchable-circularity Yagi antennas, experience with exposed circularity switching relays and preamplifiers mounted on antennas have shown that they are prone to failure caused by an elusive mechanism known as *diurnal pumping*. Often these relays are covered with a plastic case, and the seam between the case and PC board is sealed with a silicone sealant. Preamps may also have a gasket seal for the cover, while the connectors can easily leak air. None of these methods create a true hermetic seal and as a result the day/night temperature swings pump air and moisture in and out of the relay or preamp case. Under the right conditions of temperature and moisture content, moisture from the air will condense inside the case when the outside air cools down. Condensed water builds up inside the case, promoting extensive corrosion and unwanted electrical conduction, seriously degrading component performance in a short time.

A solution for those antennas with "sealed" plastic relays, such as the KLM CX series; you can avoid problems by making the modifications shown in **Fig 60**. Relocate the 4:1 balun as shown and place a clear polystyrene plastic refrigerator container over the relay. Notch the container edges for the driven element and the boom so the container will sit down over the relay, sheltering it from the elements. Bond the container in place with a few dabs of RTV adhesive sealant. Position the antenna in an "X" orientation, so neither set of elements is parallel to the ground. The switcher board should now be canted at an angle, and one side of the relay case should be lower than the other. An example for the protective cover for an S-band preamp can be seen in the discussion on feeds for parabolic antennas.

For both the relay and preamp cases, carefully drill a 3/32-inch hole through the low side of the case to provide the needed vent. The added cover keeps rainwater off the

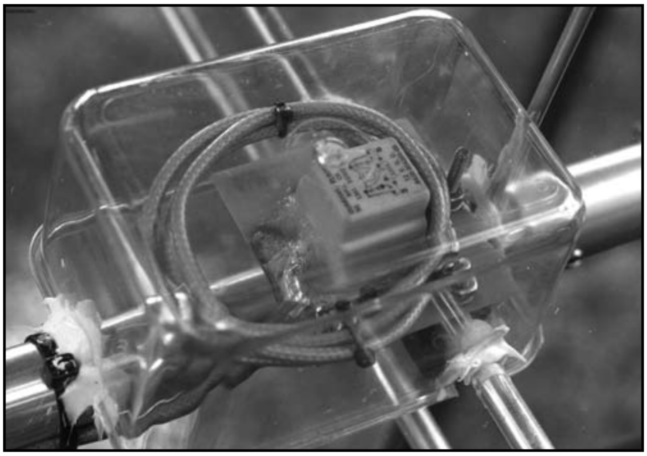

Fig 60—KLM 2M-22C antenna CP switching relay with relocated balun. The protective cover is needed for rain protection, be sure to use a polystyrene kitchen box, see text. (*WD4FAB photo.*)

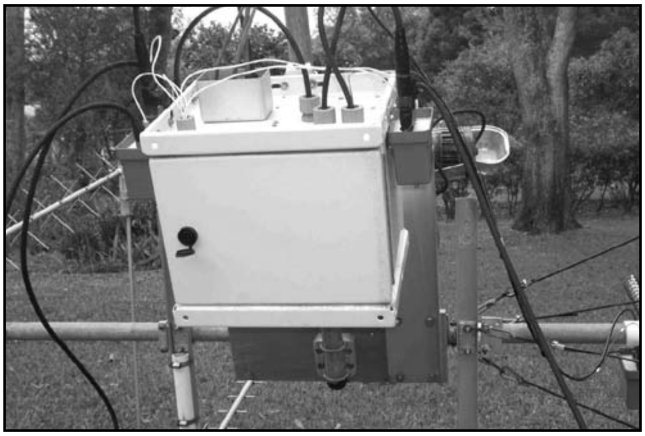

Fig 61—A NEMA4 box is used to shelter the L-band electronics and power supply. The box flanges are convenient for mounting preamplifiers. The box is shown inverted since it is on a tilt-over tower. (*WD4FAB photo.*)

relay and preamp, and the holes will prevent any buildup of condensation inside the relay case. Relays and preamplifiers so treated have remained clean and operational over periods of years without problems.

Another example for the protection of remotely, tower-mounted equipment is shown in Fig 50, illustrating the equipment box and mast-mounted preamplifiers at the top of WD4FAB's tower. The commercial NEMA4-rated equipment box, detailed in **Fig 61** (shown inverted), is used to protect the 23-cm power amplifier and its power supply, as well as a multitude of electrical connections. This steel box is very weather resistant, with an exceptionally good epoxy finish, but it is not sealed and so it will not trap moisture to be condensed with temperature changes. Be sure to use a box with at least a NEMA3 rating for rainwater and dust protection. The NEMA4 rating is just a little better protection than the NEMA3 rating. Using a well-rated equipment box is very well worth the expense of the box. As you can see, the box also provides some pretty good flanges to mount the mast-mounted preamplifiers for three bands. This box is an elegant solution for the simple need of rain shelter for your equipment. See **Fig 62**.

Elevation Control

Satellite antennas need to have elevation control to point up to the sky. This is the "El" part of Az-El control of satellite antennas. Generally, elevation booms for CP satellite antennas need to be non-conducting so that the boom does not affect the radiation pattern of the antenna. In the example shown next, the elevation boom center section is a piece of extra-heavy-wall $1^1/_2$-inch pipe (for greater strength) coupled with a tubular fiberglass-epoxy boom extension on the 70-cm end and a home-brew long extension on the 2-meter end. This uses large PVC pipe reinforced with four braces of Phillystran non-metallic guy cable. (PVC pipe is notoriously flexible, but the Phillystran cables make a quite stiff and strong boom of the PVC pipe.) For smaller installations, a continuous piece of fiberglass-epoxy boom can be placed directly through the elevation rotator.

Elevation boom motion needs to be powered, and one solution by WD4FAB, shown in **Fig 63**, uses a surplus jackscrew drive mechanism. I8CVS has also built his own robust elevation mechanism. See **Fig 64**. Note in each of these applications the methods used to provide bearings for the elevation mechanism. In WD4FAB's case, the elevation axis is a piece of heavy-duty $1^1/_2$-inch pipe,

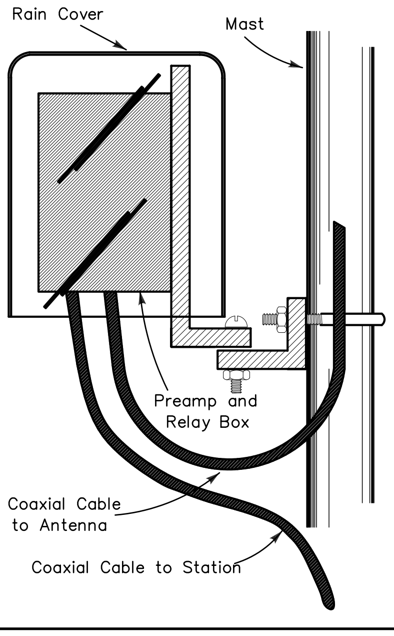

Fig 62—Protection for tower-mounted equipment need not be elaborate. Be sure to dress the cables as shown so that water drips off the cable jacket before it reaches the enclosure. One hazard for such open-bottom enclosures is that of animals liking the cable insulation as a delicacy. Flying insects also like to build their houses in these enclosures.

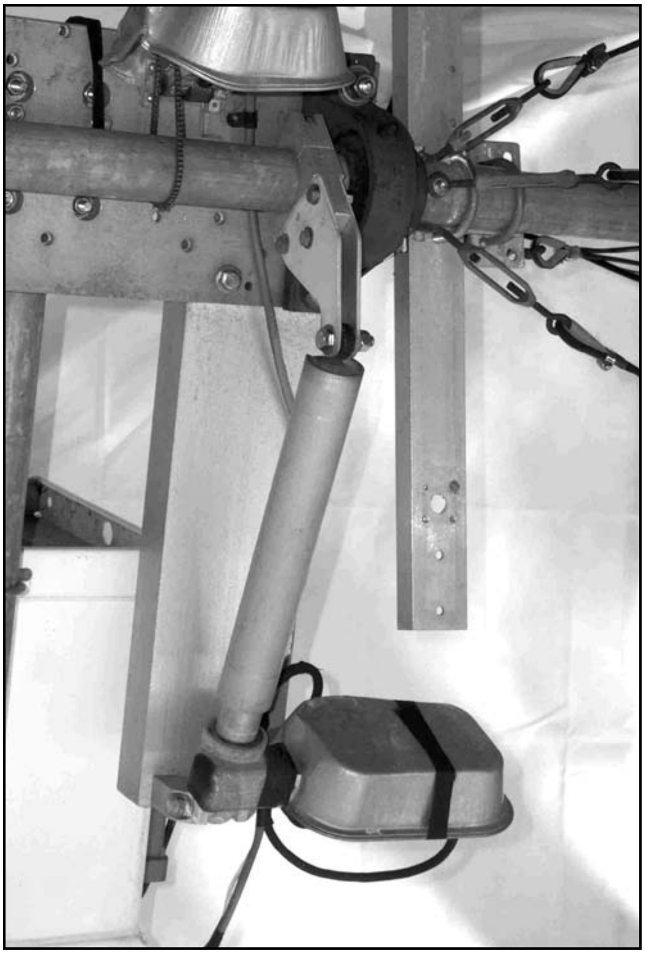

Fig 63—WD4FAB's homebrew elevation rotator drive using a surplus-store drive screw mechanism. Note also the large journal bearing supporting the elevation axis pipe shaft. (*WD4FAB photo.*)

($1^{15}/_{16}$-inch OD) and large 2 inch journal bearings are used for the motion. I8CVS uses a very large hinge to allow his motion.

Robust commercial solutions for Az-El rotators have given operators good service over the years. **See Fig 65**. Manufacturers such as Yaesu and M^2 are among these suppliers. One operator, VE5FP, found a solution for his Az-El needs by using two low-cost, lightweight TV rotators. See Fig 65B.

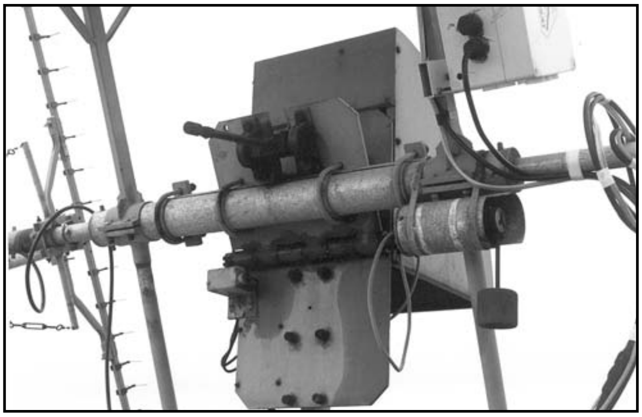

Fig 64—I8CVS's homebrew elevation mechanism using a very large, industrial hinge as the pivot and a jackscrew drive. (*I8CVS photo.*)

CONVERTED C-BAND TVRO DISHES

In working with larger, converted C-band TVRO dishes for AO-40, some operators have used only the polar mount with its jack-screw mechanism. See **Fig 66**. This dish is called *Big Ugly Dish* or just "BUD" by their users. Only using the polar mount mechanism limits the operator in the range of motion, as previously discussed. WØLMD provides for a greater degree of articulation of these dishes through several mechanisms. One of these is a sector-gear elevation drive, shown in **Fig 67**.

For the azimuth motion of our satellite antennas, most use motorized rotator drives, mainly the commercial sources previously mentioned. Most antennas are tower-mounted, allowing the placement of the rotator inside the tower. For the large wind loads of satellite antennas, these commercial rotators become rather expensive.

High loads are also prominent with the use of BUD

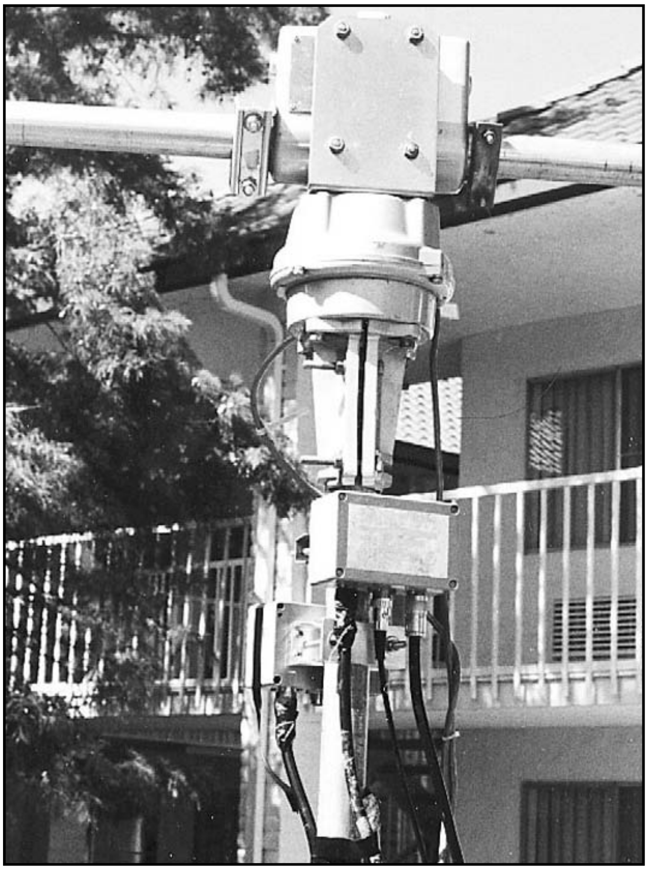

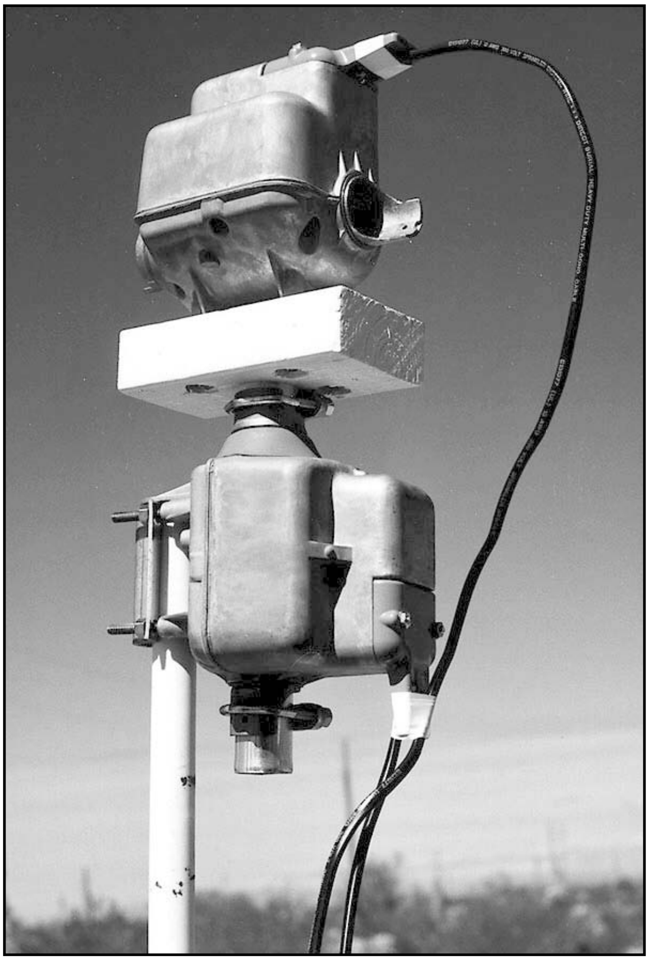

Fig 65—At left, Yaesu Az-El antenna-rotator mounting system is shown. Note that antenna loads must be more carefully balanced on this rotator than in the previously shown systems. At right, VE5FP has a solution for his Az-El rotators by bolting two of them together in his "An Inexpensive Az-El Rotator System", *QST*, December 1998.

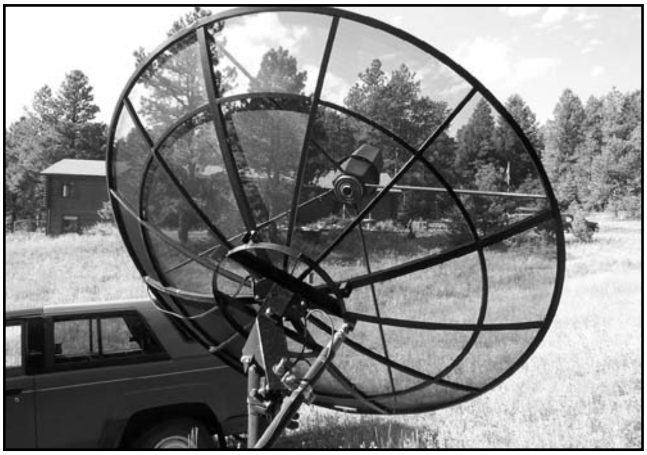

Fig 66—A TVRO dish-drive system is shown on its polar mount, using a protected drive-screw mechanism. (*W0LMD photo.*)

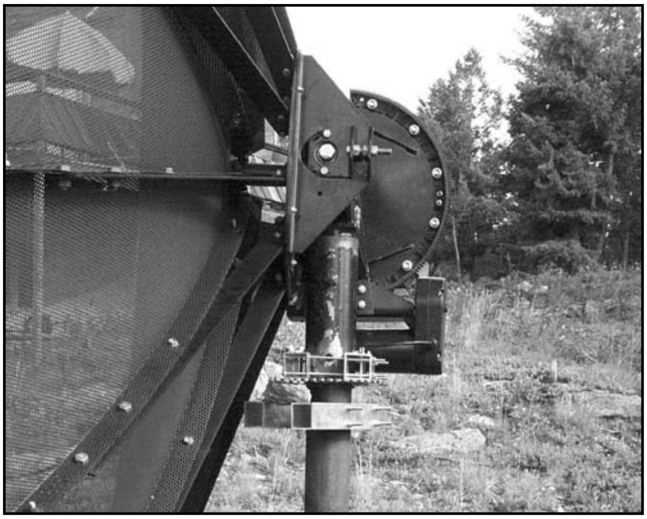

Fig 67—A modified TVRO dish mount is shown using an Az-El mount and a sector-gear drive for the elevation. (*W0LMD photo.*)

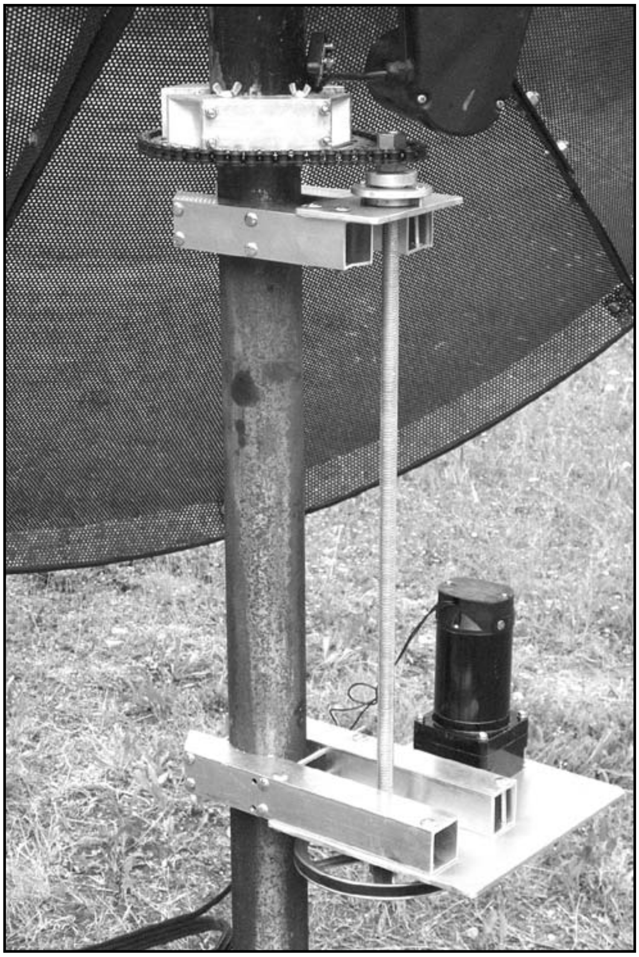

Fig 68—W0LMD constructed a very robust and low-cost Az drive mechanism. (*W0LMD photo.*)

antennas, and W0LMD has again engineered some very robust mechanisms using combinations of motorcycle-chain drives, V-belt drives and gear-head motors, as seen in **Fig 68**. An overall view of one of his BUD antennas is shown in **Fig 69**, showing the Az drive with an El drive that uses a jackscrew mechanism.

Operators through the years have employed many methods for the control of their antenna positions, ranging from true *arm-strong* manual positioning, to manual operation of the powered antenna azimuth and elevation rotators, to fully automated computer control of the rotators. While computer control of the rotators is not essential, life is greatly assisted with their use. For many years, one of the keystone control units for rotators has been the *Kansas City Tracker* (KCT) board installed in your computer. Most satellite-tracking programs can connect to the KCT with ease. One difficulty with the KCT unit is that they are 8-bit digital units, providing positioning precision of 0.35° in elevation and 1.41° in azimuth. For the larger dishes, with their narrow beamwidths, these values of precision are unacceptable. There are other options to replace the KCT unit.

A recent trend for amateur antenna control has been evolving in the form of a standalone controller that translates computer antenna-position information into controller commands with an understanding of antenna-position limits. These boxes, represented by the *EasyTrak* unit, **Fig 70**, from the Tucson Amateur Packet Radio (TAPR) group, have made this capability readily available for many amateurs. This unit is a 10-bit encoder, providing precisions of 0.09° in elevation and 0.35° in azimuth. The computer can also control the operation of your station transceiver through the radio interface provided in *EasyTrak*; you will not need any other radio interface.

Other position readout and control options are available. For many years ham operators have employed synchros, or *selsyns*, for their position readouts. These are

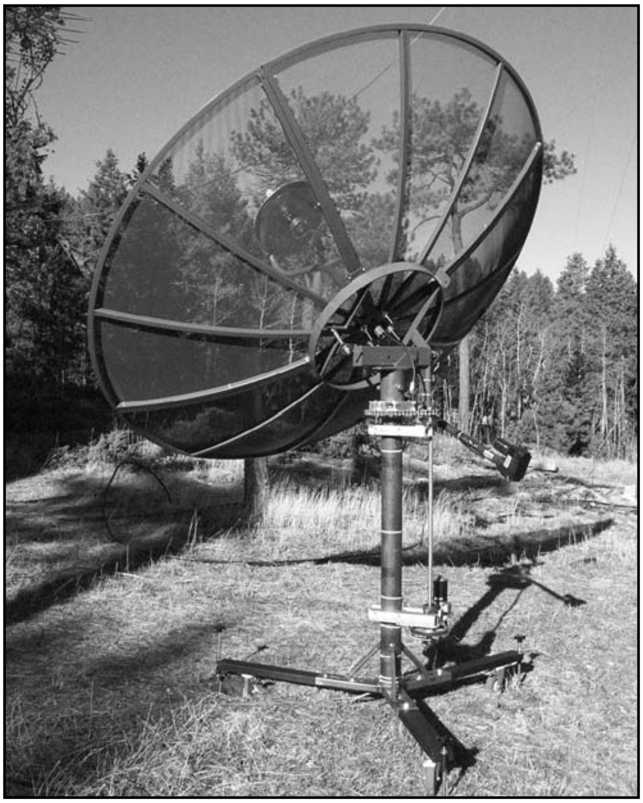

Fig 69—A completed TVRO dish Az-El mounting system is shown, using a jackscrew elevation drive. (*W0LMD photo.*)

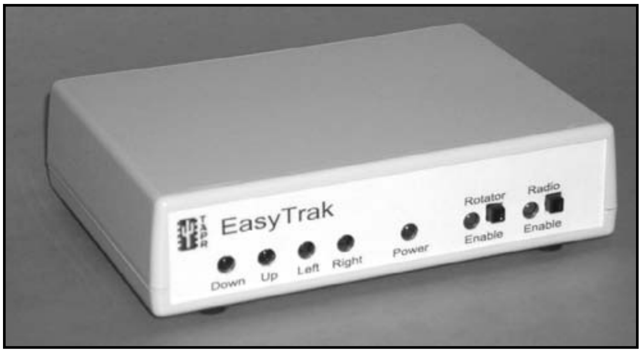

Fig 70—The EasyTrak automated antenna rotator and radio controller by TAPR. (*WD4FAB photo.*)

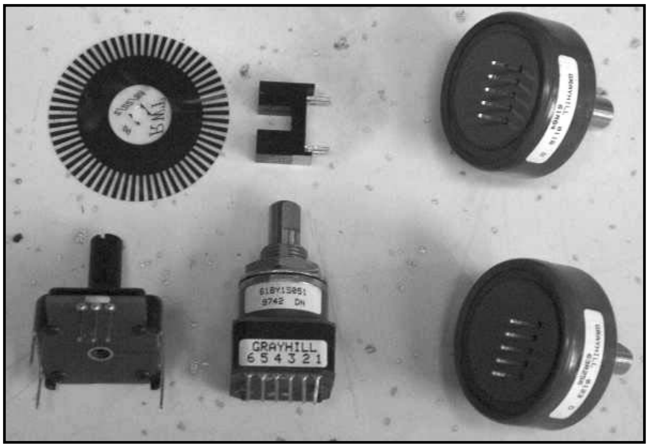

Fig 71—W0LMD has experimented with highly precise optical encoders for his antenna position systems. See text. (*W0LMD photo.*)

specialized transformers, using principles developed over sixty years ago and employed in such devices as surplus "radio compass" steering systems for aircraft. While the position readout of these devices can be quite precise, in general they only provide a visual position indication, one that is not easily adapted to computer control. I8CVS employs such a system at his station and his elevation synchro can be seen in Fig 64, using a weighted arm on the synchro to provide a constant reference to the Earth's gravity vector.

The more up-to-date, computer-friendly position readout methods used these days are usually based on precision potentiometers or digital code wheels. **Fig 71** shows such a digital code-wheel system employed by W0LMD. He notes that such systems, while providing a very high precision of angular position, they are not absolute systems and that once calibrated, they must be continually powered so they do not lose their calibration. Precision potentiometers, on the other hand, provide an absolute position reference, but with a precision that is limited to the quality of the potentiometer, typically 0.5% (0.45° in El and 1.80° in Az) to 1.0%. So the choices have their individual limits, unless a lot of money is spent for very precise commercial systems.

NOTES AND REFERENCES, SATELLITE ANTENNAS

AO-40 frequencies, "Official Transponder Frequency Bandplan for P-3D"; **www.amsat-dl.org/p3dqrg.html**.

Brown, Gerald R., K5OE, "A Helix Feed for Surplus MMDS, Antennas" *Proceedings of the 2001AMSAT-NA Symposium*, October 2001, pp 89-94, also: **members.aol.com/k5oe**.

Brown, Jerry, K5OE, "A K-Band Receiver for AO-40", *Proceedings for the AMSAT-NA Space Symposium, 2002.*

Brown, Gerald R., K5OE, "Build This No-Tune Dual-Band Feed for Mode L/S" *The AMSAT Journal*, Vol 26, No 1, Jan/Feb 2003.

Brown, Gerald R., K5OE, "Dual-Band Dish Feeds for 13/23 cm," *Proceedings of the 2002AMSAT-NA Symposium*, October 2002, pp 123-131.

Brown, Jerry, K5OE, **members.aol.com/k5oe/** MMDS Dishes.

Brown, Gerald R., K5OE, "Patch Feeds," **members.aol.com/k5oe**.

Brown, Jerry, K5OE, "The Texas Potato Masher: A Medium-Gain Directional Satellite Antenna For

LEOs," *The AMSAT Journal*, Vol 22, No. 1, Jan/Feb 1999.

Jansson, Dick, WD4FAB, *QST*, Sep 2002, "Product Review", "M^2 23CM22EZA 1.2 GHz Antenna", p 59.

Kingery, Mike, KE4AZN, "Setting Up for AO-40 L-Band Uplink," *The AMSAT Journal*, May/Jun 2002, pp 14-16, also: **web.infoave.net/~mkmk518/**.

Koehler, Jim, VE5FP, "An Inexpensive Az-El Rotator System", *QST*, Dec 1998.

Kraus, John D., "The Helical Antenna", Chapter 7, *Antennas*, McGraw-Hill Book Company, 1988.

Kraus, John D., W8JK, "16-12, Patch or Microstrip Antennas," *Antennas*, 1988, pp 745-749.

Krome, Ed, K9EK, "Development of a Portable Mode S Ground Station." *The AMSAT Journal*, Vol 16, No. 6, Nov/Dec 1993, pp 25-28.

Krome, Ed, K9EK, "S band Reception: Building the DEM Converter and Preamp Kits," *The AMSAT Journal*, Vol 16, No. 2, Mar/Apr 1993, pp 4-6.

Krome, Ed., K9EK, *Mode S—The Book*, p 96, 109.

Krome, Ed, K9EK, "Mode S: Plug and Play!" *The AMSAT Journal*, Vol 14, No. 1, Jan 1991, pp 21-23, 25.

Long, Howard, G6LVB, "My Shack Configuration—Spring 2002": **www.g6lvb.com/g6lvb_shack_spring_2002.htm**.

McCaa, William D., K0RZ, "Hints on Using the AMSAT-OSCAR 13 Mode S Transponder," *The AMSAT Journal*, Vol 13, No. 1, Mar 1990, pp 21-22.

MacAllister, Andy, W5ACM, "Field Day 2002," *73 Amateur Radio Today*, Sept. 2002, pp 48-52.

Miller, James, G3RUH, "Mode S — Tomorrow's Downlink?" *The AMSAT Journal*, Vol 15, No. 4, Sep/Oct 1992, pp 14-15.

Miller, James S., G3RUH, "'Patch' Feed For S-Band Dish Antennas"; **www.jrmiller.demon.co.uk/products/patch.html**.

Miller, James, G3RUH, "A 60-cm S-Band Dish Antenna," *The AMSAT Journal*, Vol 16 No. 2, Mar/Apr 1993, pp 7-9.

Miller, James, G3RUH, "Small is Best," *The AMSAT Journal*, Vol 16, No. 4, Jul/Aug 1993, p 12.

Monteiro, Anthony, "Work OSCAR 40 with Cardboard-box Antennas!," *QST*, Mar 2003, pp 57-62.

Orban Microwave, "The Basics of Patch Antennas: **www.orbanmicrowave.com/antenna_application_notes.htm**.

Seguin, Mike, N1JEZ, "OSCAR 40 on 24GHz", *QST*, Dec 2002, pp 55-56.

Seydler, Robert, K5GNA,"Modifications of the AIDC 3731 Downconverters", **members.aol.com/k5gna/AIDC3731modifications.doc**.

Suckling, Charles, G3WDG, "K-Band Results From AO-40": **www.g3wdg.free-online.co.uk/kband.htm**.

Suckling, Charles, G3WDG "Notch Filters for AO-40 Mode L/S": **www.g3wdg.free-online.co.uk.notch.htm**.

Suding, Dr. Robert, WØLMD, "Converting TVRO Dish & Dishes For Amateurs" **www.ultimatecharger.com/dish.html**.

Thiel, David & Smith, Staphani, *Switched Parasitic Antennas for Cellular Communications*, Artech House, 2002, "Chapter 3, Patch Antennas," pp 79-96.

Wade, Paul, N1BWT (W1GHZ), *Online Microwave Antenna Handbook*, "Chapter 4, Parabolic Dish Antennas," **www.w1ghz.cx/antbook/chap4.pdf**, Mar. 98. (See CD-ROM bundled with this book.)

Zibrat,Timothy S, K3TZ, "2.4 GHz Patch Design": **www.qsl.net/k3tz/**.

Antenna Systems for EME Communications

This section was updated by David Hallidy, K2DH. As mentioned earlier, the tremendous path loss incurred over an EME circuit places stringent requirements on Earth-station performance. Low-noise receiving equipment, maximum available power and high-gain antenna arrays are required for successful EME operation. Although it is possible to copy some of the better-equipped stations with a low-gain antenna, it is unlikely that such an antenna can provide reliable two-way communications. Antenna gain of at least 20 dBi is required for reasonable EME success. Generally speaking, more antenna gain yields the most noticeable improvement in station performance, since the increased gain improves both the received and transmitted signals.

VHF/UHF EME ANTENNAS

Several types of antennas for 2 meters and 70 cm are popular among EME enthusiasts. Perhaps the most popular antenna for 144-MHz work is an array of either 4 or 8 long-boom (14 to 15 dBi gain) Yagis. The 4-Yagi array provides approximately 20 dB gain, and an 8-Yagi array gives an approximate 3 dB increase over the 4-antenna array. **Fig 72** shows the computed response at a 30° tilt above the horizon for a stack of four 14-element 2-meter Yagis, each with a boomlength of 3.1 λ (22 feet).

At 432 MHz, EME enthusiasts often use 8 or 16 long-boom Yagis in an array. Such Yagis are commercially available or they can be constructed from readily available materials. Chapter 18, VHF and UHF Antenna Systems, has details on some popular Yagi designs.

The main disadvantage of Yagi arrays is that the polarization plane of the individual Yagis cannot be conveniently changed. One way around this is to use cross-polarized Yagis and a relay switching system to select the desired polarization, as described in the previous section. This represents a considerable increase in system complexity to select the desired polarization. Some amateurs have gone so far as to build complicated mechanical systems to allow constant polarization adjustment of all the Yagis in a large array. **Fig 73** shows the K1FO 70-cm EME 16-Yagi array with full polarization control, described in *The ARRL Antenna Compendium*, *Vol 3*. This 432-MHz EME array uses open-wire phasing lines to minimize feed-line losses. **Fig 74** shows the computed response for this array, which employs rugged but lightweight 14-element Yagis on 3.1 λ (7.1 foot) booms. Feed-line losses are not explicitly accounted for in the *EZNEC Professional* computer model, but are estimated to be less than 0.25 dB.

Polarization shift of EME signals at 144 MHz is fairly rapid, and the added complexity of a relay-controlled cross-polarized antenna system or a mechanical polarization adjustment scheme is probably not worth the effort. At 432 MHz, however, where the polarization shifts at a much slower rate, an adjustable polarization system does offer a definite advantage over a fixed one.

The Yagi antenna system used by Ed Stallman, N5BLZ, is shown in **Fig 75**. His system employs twelve 144-MHz long-boom 17-element Yagi antennas. The monster 48-Yagi 2-meter array of Gerald Williamson, K5GW, is shown in **Fig 76**, and the huge 48-Yagi 70-cm EME array of Frank Potts, NC1I, is shown in **Fig 77**.

Although not as popular as Yagis, Quagi antennas (made from both quad and Yagi elements) are sometimes used for EME work. Slightly more gain per unit boom length is possible as compared to the conventional Yagi, at the expense of some robustness. Additional information on the Quagi is presented in Chapter 18, VHF and UHF Antenna Systems.

The collinear array is an older type of antenna for EME work. A 40-element collinear array has approximately

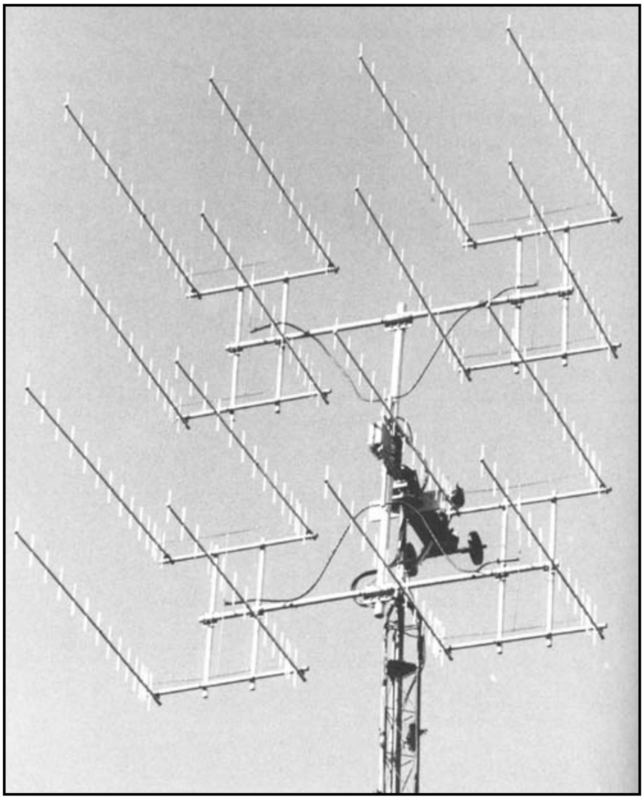

Fig 73—K1FO's variable polarization 16 × 14-element (3.6-λ boom lengths) 432-MHz EME array shown at 2° elevation and vertical polarization. (See *The ARRL Antenna Compendium, Vol 3*.)

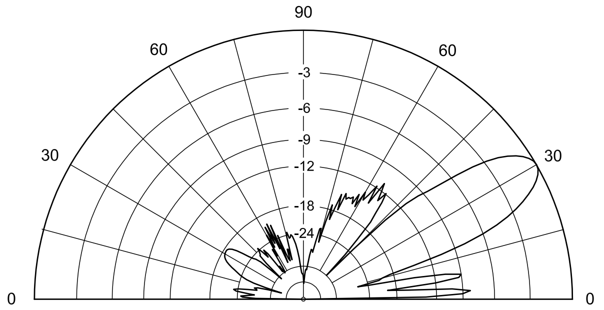

Fig 72—*EZNEC Pro* elevation pattern for four 14-element 2-meter Yagis (3.6-λ boom lengths) at an elevation angle of 30° above the horizon. The computed system gain is 21.5 dBi, suitable for 2-meter EME. This assumes that the phasing system is made of open-wire transmission lines so that feed-line losses can be kept below 0.25 dB.

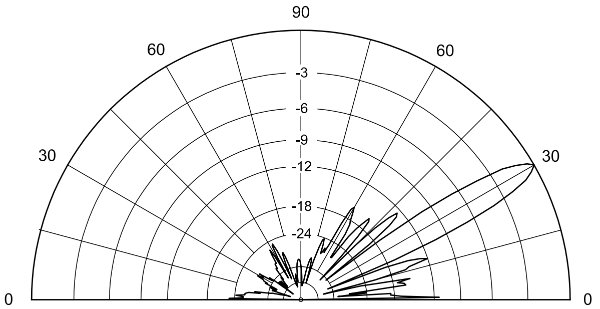

Fig 74—Computed elevation response for K1FO 16-Yagi 432-MHz array shown in Fig 73. (The *EZNEC Pro* model required 2464 segments!) With assumed phasing harness feed-line losses of 0.25 dB, the overall gain exceeds 27.5 dBi.

the same frontal area as an array of four Yagis, but produces approximately 1 to 2 dB less gain. One attraction to a collinear array is that the depth dimension is considerably less than the long-boom Yagis. An 80-element collinear is marginal for EME communications, providing approximately 19 dB gain. As with Yagi and Quagi antennas, the collinear cannot be adjusted easily for polarity changes. From a construction standpoint, there is little difference in complexity and material costs between the collinear and Yagi arrays.

DISH ANTENNAS FOR EME

On 2 meters the minimum antenna gain for reliable EME communications is about 20 dBi. While a few amateurs have had access to parabolic dishes large enough for EME work at 144 and 222 MHz, at those frequencies an array of four long Yagis is equal in gain to a dish 24 feet in diameter! To achieve truly high-gain performance from a dish on 2 meters would require a reflector diameter of nearly 96 feet (providing 32 dBi gain). Such undertakings are generally beyond amateur means, so there has been little work done with dishes at low frequencies, except for the occasional expedition to one of the large radio telescopes that have accommodated amateur EME work.

Microwave Parabolic Dish Antennas

The major problems associated with parabolic dish antennas are mechanical ones. A dish of about 16 feet in diameter is the minimum size required for successful EME operation on 432 MHz. With wind and ice loading, structures of this size place a real strain on the mounting and positioning system. Extremely rugged mounts are required for large dish antennas, especially when used in windy locations. **Fig 78** shows the impressive 7-meter diameter

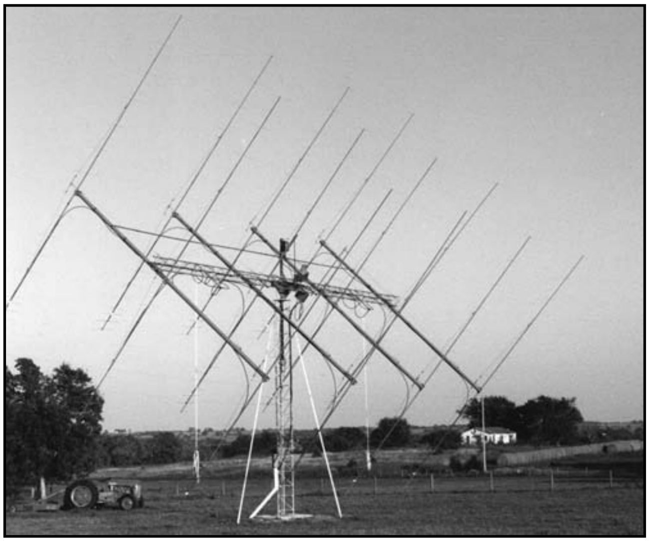

Fig 75—The EME array used at N5BLZ consists of twelve long-boom 144-MHz Yagis. The tractor, lower left, really puts this array into perspective! (*Photo courtesy N5BLZ.*)

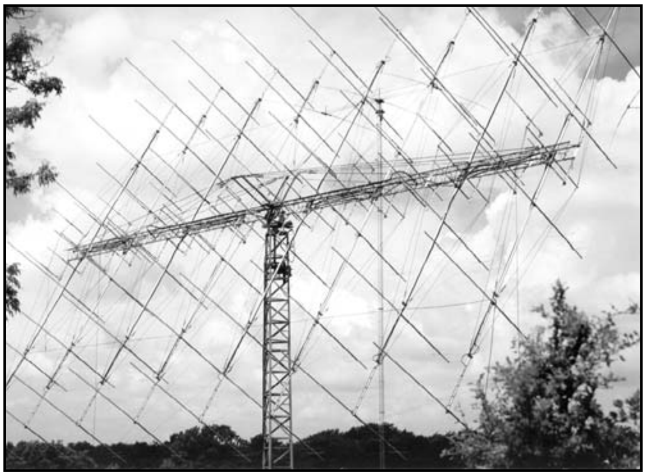

Fig 76—K5GW's huge 48-Yagi 2-meter EME array. (*Photo courtesy K5GW.*)

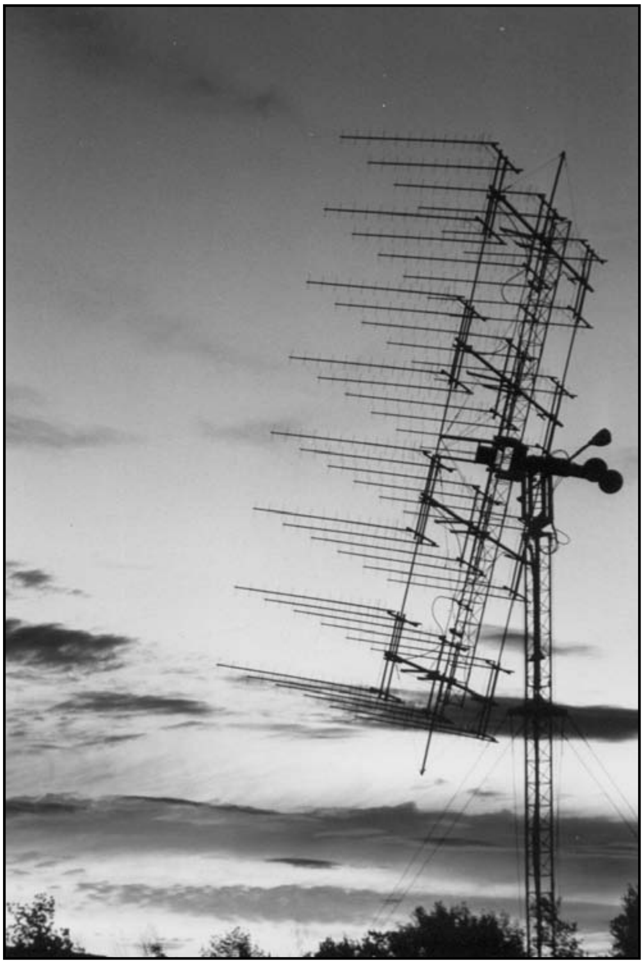

Fig 77—NC1I's magnificent 48-Yagi 70-cm EME array. (*Photo courtesy NC1I.*)

dish built by David Wardley, ZL1BJQ.

Several aspects of parabolic dish antennas make the extra mechanical problems worth the trouble, however. For example, the dish antenna is inherently broadband, and may be used on several different amateur bands by simply changing the feed. An antenna that is suitable for 432 MHz work will most likely be usable on several of the higher amateur bands too. Increased gain is available as the frequency of operation is increased.

Another advantage of a dish is the flexibility of the feed system. The polarization of the feed, and therefore the polarization of the antenna, can be changed with little difficulty. It is a relatively easy matter to devise a system to rotate the feed remotely from the shack to change polarization. Because polarization changes can account for as much as 30 dB of signal attenuation, the rotatable feed can make the difference between consistent communications and no communications at all. Further information on Parabolic Antennas can be found in Chapter 18, VHF and UHF Antenna Systems as well as in the section below.

A 12-FOOT STRESSED HOMEBREW PARABOLIC DISH

Very few antennas evoke as much interest among UHF amateurs as the parabolic dish, and for good reason. First, the parabola and its cousins—Cassegrain, hog horn and Gregorian—are probably the ultimate in high-gain antennas. One of the highest-gain antennas in the world (148 dB) is a parabola. This is the 200-inch Mt. Palomar telescope. (The very short wavelength of light rays causes such a high gain to be realizable.)

Second, the efficiency of the parabola does not change as size increases. With Yagis and collinear arrays, the losses in the phasing harness increase as the array size increases. The corresponding component of the parabola is lossless air between the feed horn and the reflecting surface. If there are a few surface errors, the efficiency of the system stays constant regardless of antenna size. This project was presented by Richard Knadle, K2RIW, in August 1972 *QST*.

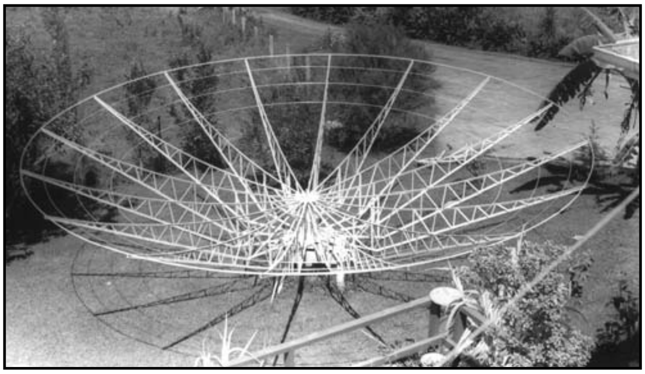

Fig 78—ZL1BJQ's homemade 7-meter (23-foot) parabolic dish, just prior to adding $^1/_2$-inch wire mesh. (*Photo courtesy ZL1BJQ*.)

Some amateurs reject parabolic antennas because of the belief that they are all heavy, hard-to-construct, have large wind-loading surfaces and require precise surface accuracy. However, with modern construction techniques, a prudent choice of materials and an understanding of accuracy requirements, these disadvantages can be largely overcome. A parabola may be constructed with a 0.6 f/D (focal length/diameter) ratio, producing a rather flat dish, which makes it easy to surface and allows the use of recent advances in high-efficiency feed horns. This results in greater gain for a given dish size over conventional designs.

Such an antenna is shown in **Fig 79**. This parabolic dish is lightweight, portable, easy to build, and can be used for 432 and 1296-MHz mountain topping, as well as on 2304, 3456 and 5760 MHz. Disassembled, it fits into the trunk of a car, and can be assembled in 45 minutes.

The usually heavy structure that supports the surface of most parabolic dish antennas has been replaced in this design by aluminum spokes bent into a near parabolic shape by string. These strings serve the triple function of guying the focal point, bending the spokes and reducing the error at the dish perimeter (as well as at the center) to nearly zero. By contrast, in conventional designs, the dish perimeter (which has a greater surface area than the center) is farthest from the supporting center hub. For these reasons, it often has the greatest error. This error becomes more severe when the wind blows.

Here, each of the spokes is basically a cantilevered beam with end loading. The equations of beam bending

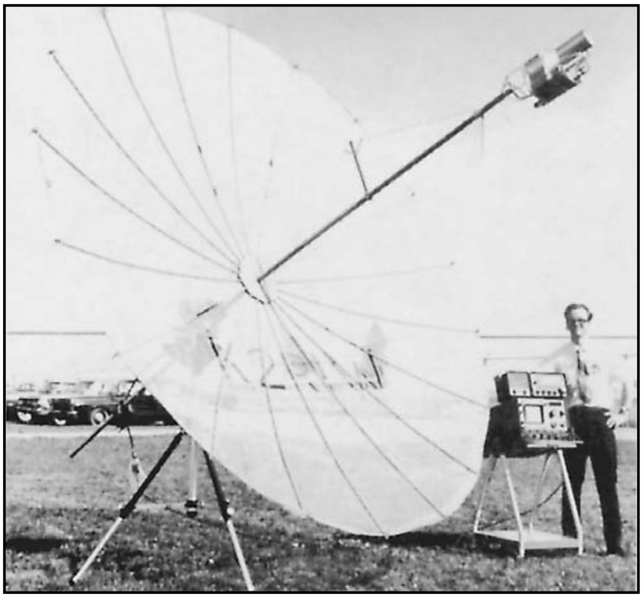

Fig 79—A 12-foot stressed parabolic dish set up for satellite signals near 2280 MHz. A preamplifier is shown taped below the feed horn. The dish was designed by K2RIW, standing at the right. From *QST*, August 1972.

predict a near-perfect parabolic curve for extremely small deflections. Unfortunately the deflections in this dish are not that small and the loading is not perpendicular. For these reasons, mathematical prediction of the resultant curve is quite difficult. A much better solution is to measure the surface error with a template and make the necessary correction by bending each of the spokes to fit. This procedure is discussed later.

The uncorrected surface is accurate enough for 432 and 1296-MHz use. Trophies taken by this parabola in antenna-gain contests were won using a completely natural surface with no error correction. By placing the transmission line inside the central pipe that supports the feed horn, the area of the shadows or blockages on the reflector surface is much smaller than in other feeding and supporting systems, thus increasing gain. For 1296 MHz, a backfire feed horn may be constructed to take full advantage of this feature. At 432 MHz, a dipole and reflector assembly produces 1.5 dB additional gain over a corner-reflector feed system. Because the preamplifier is located right at the horn on 2300 MHz, a conventional feed horn may be used.

Construction

Table 2 is a list of materials required for construction. Care must be exercised when drilling holes in the connecting center plates so assembly problems will not be experienced later. See **Fig 80**. A notch in each plate allows them to be assembled in the same relative positions. The two plates should be clamped together and drilled at the same time. Each of the $18^{1}/_{2}$-inch diameter aluminum spokes has two no. 28 holes drilled at the base to accept no. 6-32 machine screws that go through the center plates. The 6-foot long spokes are cut from standard 12-foot lengths of tubing. A fixture built from a block of aluminum assures that the holes are drilled in exactly the same position in each spoke. The front and back center plates constitute an I-beam type of structure that gives the dish center considerable rigidity.

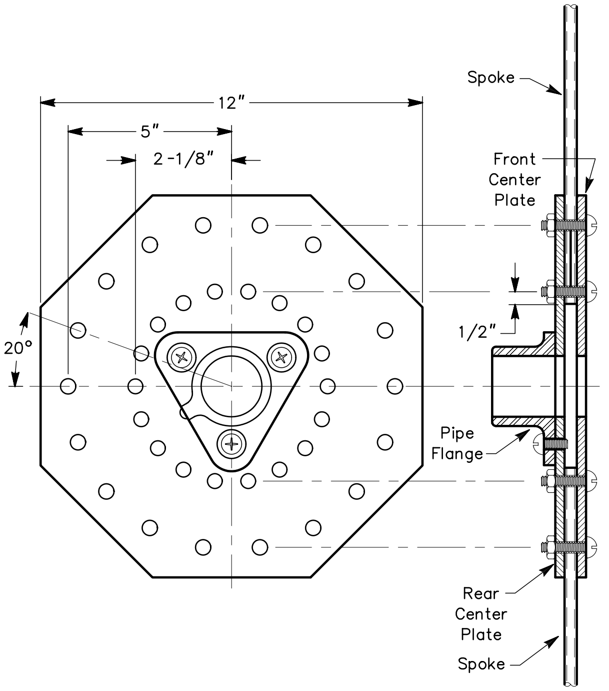

Fig 80—Center plate details. Two center plates are bolted together to hold the spokes in place.

A side view of the complete antenna is shown in **Fig 81**. Aluminum alloy (6061-T6) is used for the spokes, while 2024-T3 aluminum alloy sheet, $^{1}/_{8}$ inch thick, is used for the center plates. (Aluminum has approximately three times the strength-to-weight ratio of wood, and aluminum cannot warp or become water logged.) The end of each of the 18 spokes has an eyebolt facing the dish focal point, which serves a dual purpose:

1) To accept the #9 galvanized fence wire that is routed through the screw eyes to define the dish perimeter, and
2) To facilitate rapid assembly by accepting the S hooks which are tied to the end of each of the lengths of 130-pound test Dacron fishing string.

The string bends the spokes into a parabolic curve; the dish may be adapted for many focal lengths by tightening or slackening the strings. Dacron was chosen because it has the same chemical formula as Mylar. This is a low-stretch material that keeps the dish from changing shape. The galvanized perimeter wire has a 5-inch overlap area that is bound together with baling wire after the spokes have been hooked to the strings.

The aluminum window screening is bent over the perimeter wire to hold it in place on the back of the spokes. Originally, there was concern that the surface perturbations (the spokes) in front of the screening might decrease the gain. The total spoke area is so small, however, that

Table 2
Materials List for the 12-Foot Stressed Parabolic Dish

1) Aluminum tubing, 12 ft × $^{1}/_{2}$ in. OD × 0.049-in. wall, 6061-T6 alloy, 9 required to make 18 spokes.
2) Octagonal mounting plates 12 × 12 × $^{1}/_{8}$ in., 2024-T3 alloy, 2 required.
3) $1^{1}/_{4}$ in. ID pipe flange with setscrews.
4) $1^{1}/_{4}$ in. × 8 ft TV mast tubing, 2 required.
5) Aluminum window screening, 4 × 50 ft.
6) 130-pound test Dacron trolling line.
7) 38 ft #9 galvanized fence wire (perimeter).
8) Two hose clamps, $1^{1}/_{2}$ in.; two U bolts; $^{1}/_{2}$ × 14 in. Bakelite rod or dowel; water-pipe grounding clamp; 18 eye bolts; 18 S hooks.

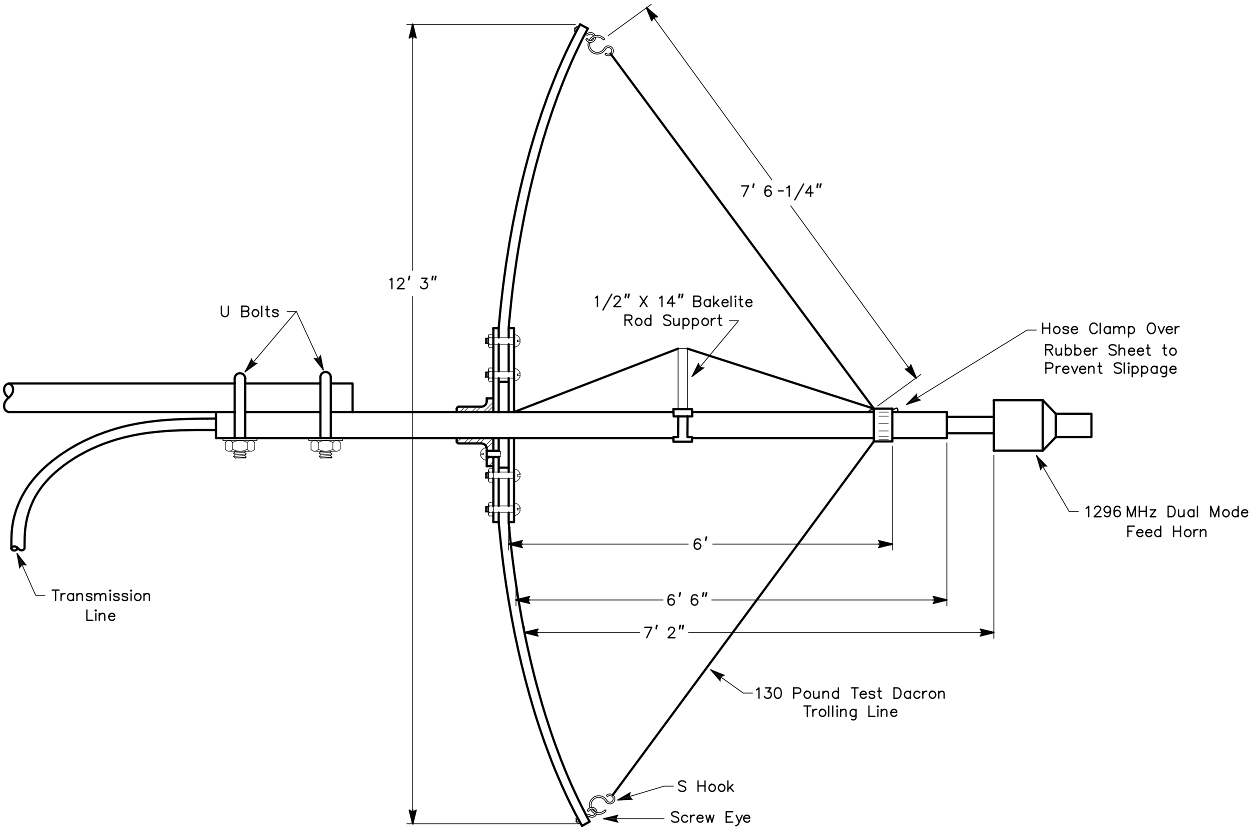
7′ 6-1/4″
12′ 3″
U Bolts
1/2″ X 14″ Bakelite Rod Support
Hose Clamp Over Rubber Sheet to Prevent Slippage
1296 MHz Dual Mode Feed Horn
6′
6′ 6″
7′ 2″
Transmission Line
130 Pound Test Dacron Trolling Line
S Hook
Screw Eye

Fig 81—Side view of the stressed parabolic dish.

this fear proved unfounded.

Placing the aluminum screening in front of the spokes requires the use of 200 pieces of baling wire to hold the screening in place. This would increase the assembly time by at least an hour. For contest and mountaintop operation (when the screening is on the back of the spokes) no fastening technique is required other than bending the screen to overlap the wire perimeter.

The Parabolic Surface

A 4-foot wide roll of aluminum screening 50 feet long is cut into appropriate lengths and laid parallel, with a 3-inch overlap between the top of the unbent spokes and hub assembly. The overlap seams are sewn together on one half of the dish using heavy Dacron thread and a sailmaker's curved needle. Every seam is sewn twice; once on each edge of the overlapped area. The seams on the other half are left open to accommodate the increased overlap that occurs when the spokes are bent into a parabola. The perimeter of the screening is then trimmed. Notches are cut in the 3-inch overlap to accept the screw eyes and S hooks.

The first time the dish is assembled, the screening strips are anchored to the inside surface of the dish and the seams sewn in this position. It is easier to fabricate the surface by placing the screen on the back of the dish frame with the structure inverted. The spokes are sufficiently strong to support the complete weight of the dish when the perimeter is resting on the ground.

The 4-foot wide strips of aluminum screening conform to the compound bend of the parabolic shape very easily. If the seams are placed parallel to the E-field polarization of the feed horn, minimum feedthrough will occur. This feedthrough, even if the seams are placed perpendicular to the E field, is so small that it is negligible. Some constructors may be tempted to cut the screening into pie-shaped sections. This procedure will increase the seam area and construction time considerably. The dish surface appears most pleasing from the front when the screening perimeter is slipped between the spokes and the perimeter wire, and is then folded back over the perimeter wire. In disassembly, the screening is removed in one piece, folded in half, and rolled.

The Horn and Support Structure

The feed horn is supported by 1$^1/_4$-inch aluminum television mast. The Hardline that is inserted into this tubing is connected first to the front of the feed horn, which then slides back into the tubing for support. A setscrew assures that no further movement of the feed horn occurs. During antenna-gain competition the setscrew is omit-

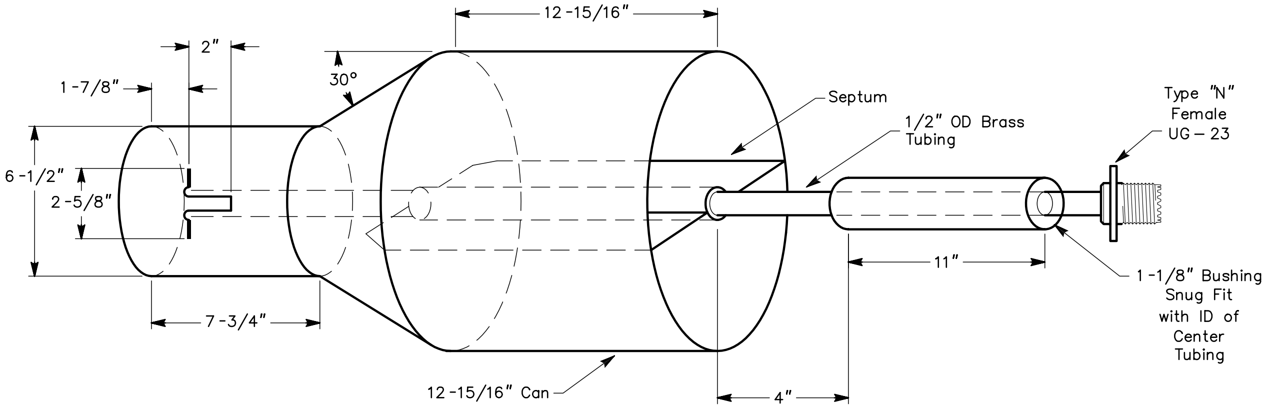

Fig 82—Backfire type 1296-MHz feed horn, linear polarization only. The small can is a Quaker State oil container; the large can is a 50-pound shortening container (obtained from a restaurant, Gold Crisp brand). Brass tubing, $^{1}/_{2}$-inch OD extends from UG-23 connector to dipole. Center conductor and dielectric are obtained from $^{3}/_{8}$-inch Alumafoam coaxial cable. The dipole is made from 3/32-inch copper rod. The septum and 30° section are made from galvanized sheet metal. Styrofoam is used to hold the septum in position. The primary gain is 12.2 dBi.

ted, allowing the $^{1}/_{2}$-inch semirigid CATV transmission line to move in or out while adjusting the focal length for maximum gain. The TV mast is held firmly at the center plates by two setscrews in the pipe flange that is mounted on the rear plate. At 2300 MHz, the dish is focused for best gain by loosening these setscrews on the pipe flange and sliding the dish along the TV mast tubing. (The dish is moved instead of the feed horn.)

The fishing strings are held in place by attaching them to a hose clamp that is permanently connected to the TV tubing. A piece of rubber sheet under the hose clamp prevents slippage and keeps the hose clamp from cutting the fishing string. A second hose clamp is mounted below the first as extra protection against slippage.

The high-efficiency 1296-MHz dual mode feed horn, detailed in **Fig 82**, weighs 5$^{3}/_{4}$ pounds. This weight causes some bending of the mast tubing, but this is corrected by a $^{1}/_{2}$-inch diameter bakelite support, as shown in Fig 81. This support is mounted to a pipe grounding clamp with a no. 8-32 screw inserted in the end of the rod. The bakelite rod and grounding clamp are mounted midway between the hose clamp and the center plates on the mast. A double run of fishing string slipped over the notched upper end of the bakelite rod counteracts bending.

The success of high-efficiency parabolic antennas is primarily determined by feed horn effectiveness. The multiple diameter of this feed horn may seem unusual. This patented dual-mode feed, designed by Dick Turrin, W2IMU, achieves efficiency by launching two different kinds of waveguide modes simultaneously. This causes the dish illumination to be more constant than conventional designs.

Illumination drops off rapidly at the perimeter, reducing spillover. The feed backlobes are reduced by at least 35 dB because the current at the feed perimeter is almost zero; the phase center of the feed system stays constant across the angles of the dish reflector. The larger diameter section is a phase corrector and should not be changed in length. In theory, almost no increase in dish efficiency can be achieved without increasing the feed size in a way that would increase complexity, as well as blockage.

The feed is optimized for a 0.6 f/D dish. The dimensions of the feeds are slightly modified from the original design in order to accommodate the cans. Either feed type can be constructed for other frequencies by changing the scale of all dimensions.

Multiband Use

Many amateurs construct multiband antenna arrays by putting two dishes back to back on the same tower. This is cost inefficient. The parabolic reflector is a completely frequency independent surface, and studies have shown that a 0.6 f/D surface can be steered seven beamwidths by moving the feed horn from side to side before the gain diminishes by 1 dB. Therefore, the best dual-band antenna can be built by mounting separate horns side by side. At worst, the antenna may have to be moved a few degrees (usually less than a beamwidth) when switching between horns, and the unused horn increases the shadow area only slightly. In fact, the same surface can function simultaneously on multiple frequencies, making crossband duplex operation possible with the same dish.

Order of Assembly

1) A single spoke is held upright behind the rear center plate with the screw eye facing forward. Two no. 6-

32 machine screws are pushed through the holes in the rear center plate, through the two holes of the spoke, and into the corresponding holes of the front center plate. Lock washers and nuts are placed on the machine screws and hand tightened.

2) The remaining spokes are placed between the machine screw holes. Make sure that each screw eye faces forward. Machine screws, lock washers and nuts are used to mount all 18 spokes.
3) The no. 6-32 nuts are tightened using a nut driver.
4) The mast tubing is attached to the spoke assembly, positioned properly, and locked down with the setscrews on the pipe flange at the rear center plate. The S hooks of the 18 Dacron strings are attached to the screw eyes of the spokes.
5) The ends of two pieces of fishing string (which go over the bakelite rod support) are tied to a screw eye at the forward center plate.
6) The dish is laid on the ground in an upright position and #9 galvanized wire is threaded through the eyebolts. The overlapping ends are lashed together with baling wire.
7) The dish is placed on the ground in an inverted position with the focus downward. The screening is placed on the back of the dish and the screening perimeter is fastened as previously described.
8) The extension mast tubing (with counterweight) is connected to the center plate with U bolts.
9) The dish is mounted on a support and the transmission line is routed through the tubing and attached to the horn.

Parabola Gain Versus Errors

How accurate must a parabolic surface be? This is a frequently asked question. According to the Rayleigh limit for telescopes, little gain increase is realized by making the mirror accuracy greater than ± $^1/_8$ λ peak error. John Ruze of the MIT Lincoln Laboratory, among others, has derived an equation for parabolic antennas and built models to verify it. The tests show that the tolerance loss can be predicted within a fraction of a decibel, and less than 1 dB of gain is sacrificed with a surface error of ±$^1/_8$ λ. (A $^1/_8$ λ is 3.4 inches at 432 MHz, 1.1 inches at 1296 MHz and 0.64 inch at 2300 MHz.)

Some confusion about requirements of greater than $^1/_8$-λ accuracy may be the result of technical literature describing highly accurate surfaces. Low sidelobe levels are the primary interest in such designs. Forward gain is a much greater concern than low sidelobe levels in amateur work; therefore, these stringent requirements do not apply.

When a template is held up against a surface, positive and negative (±) peak errors can be measured. The graphs of dish accuracy requirements are frequently plotted in terms of RMS error, which is a mathematically derived function much smaller than ± peak error (typically $^1/_3$). These small RMS accuracy requirements have discouraged many constructors who confuse them with ± peak errors.

Fig 83 may be used to predict the resultant gain of various dish sizes with typical errors. There are a couple of surprises, as shown in **Fig 84**. As the frequency is increased for a given dish, the gain increases 6 dB per octave until the tolerance errors become significant. Gain deterioration then increases rapidly. Maximum gain is realized at the frequency where the tolerance loss is 4.3 dB. Notice that at 2304 MHz, a 24-foot dish with ±2-inch peak errors has the same gain as a 6-foot dish with ±1-inch peak errors. This is quite startling, when it is realized that a 24-foot dish has 16 times the area of a 6-foot dish. Each time the diameter or frequency is doubled or halved, the gain changes by 6 dB. Each time all the errors are halved, the frequency of maximum gain is doubled. With this information, the gain of other dish sizes with other tolerances can be predicted.

These curves are adequate for predicting gain, assuming a high-efficiency feed horn is used (as described earlier), which realizes 60% aperture efficiency. At frequencies below 1296 MHz where the horn is large and causes considerable blockage, the curves are somewhat optimistic. A properly built dipole and splasher feed will have about 1.5 dB less gain when used with a 0.6 f/D dish than the dual-mode feed system described.

The worst kind of surface distortion is where the surface curve in the radial direction is not parabolic but gradually departs in a smooth manner from a perfect parabola. The decrease in gain can be severe, because a large area is involved. If the surface is checked with a template, and if reasonable construction techniques are employed, deviations are controlled and the curves represent an upper limit to the gain that can be realized.

If a 24-foot dish with ±2-inch peak errors is being used with 432 and 1296-MHz multiple feed horns, the constructor might be discouraged from trying a 2300-MHz feed because there is 15 dB of gain degradation. The dish will still have 29 dBi of gain on 2300 MHz, however, making it worthy of consideration.

The near-field range of this 12-foot stressed dish (actually 12 feet 3 inches) is 703 feet at 2300 MHz. By using the sun as a noise source and observing receiver noise power, it was found that the antenna had two main lobes about 4° apart. The template showed a surface error (insufficient spoke bending at $^3/_4$ radius), and a correction was made. A recheck showed one main lobe, and the solar noise was almost 3 dB stronger.

Other Surfacing Materials

The choice of surface materials is a compromise between RF reflecting properties and wind loading. Aluminum screening, with its very fine mesh (and weight of 4.3 pounds per 100 square feet) is useful beyond 10 GHz because of its very close spacing. This screening is easy to roll up and is therefore ideal for a portable dish. This close spacing causes the screen to be a 34% filled aperture, bring-

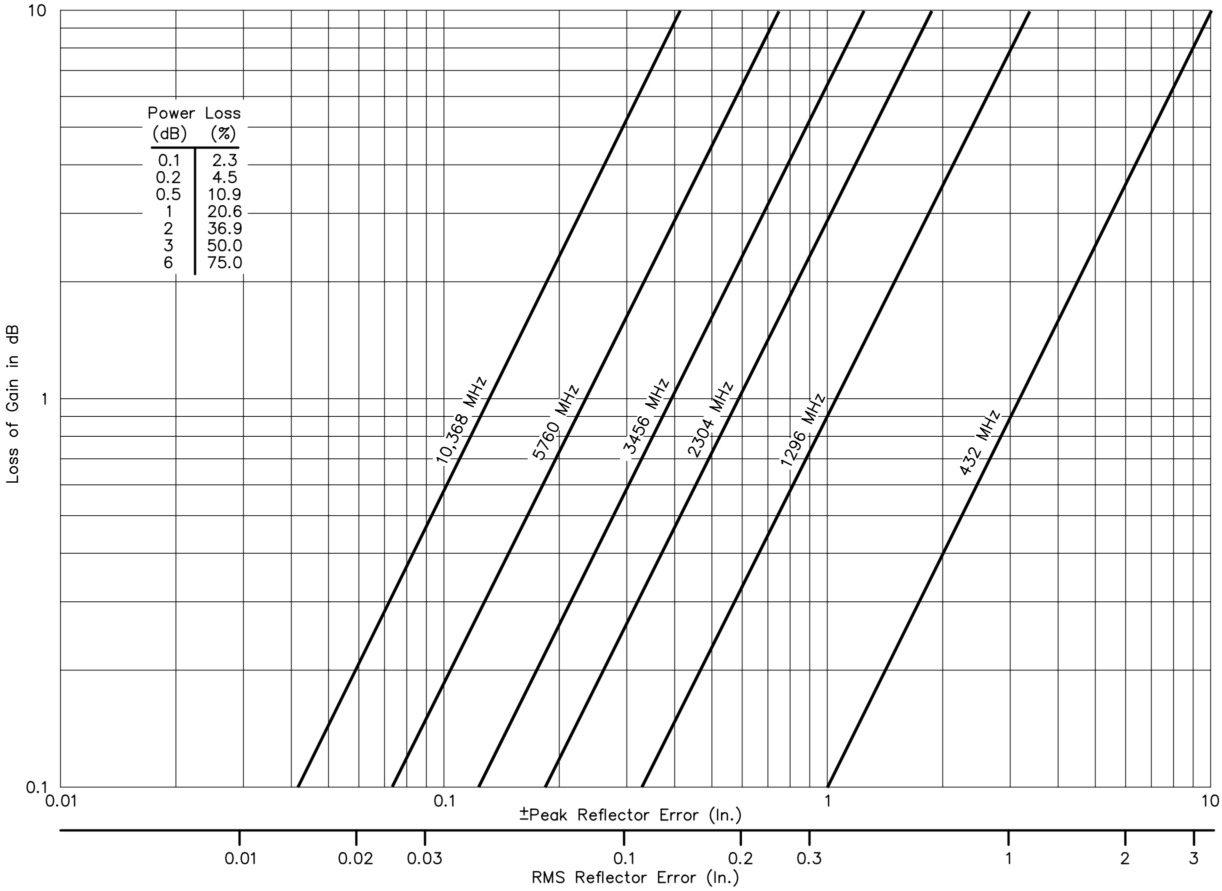

Fig 83—Gain deterioration versus reflector error. By Richard Knadle, K2RIW.

ing the wind force at 60 mph to more than 400 *pounds* on this 12-foot dish. Those considering a permanent installation of this dish should investigate other surfacing materials.

Hexagonal 1-inch poultry netting (chicken wire), which is an 8% filled aperture, is nearly ideal for 432-MHz operation. It weighs 10 pounds per 100 square feet, and exhibits only 81 pounds of force with 60 mph winds. Measurement on a large piece reveals 6 dB of feedthrough at 1296 MHz, however. Therefore, on 1296 MHz, one fourth of the power will feed through the surface material. This will cause a loss of only 1.3 dB of forward gain. Since the low-wind loading material will provide a 30-dBi gain potential, it is still a very good tradeoff.

Poultry netting is very poor material for 2300 MHz and above, because the hole dimensions approach $^1/_2$ λ. As with all surfacing materials, minimum feedthrough occurs when the E-field polarization is parallel to the longest dimension of the surfacing holes.

Hardware cloth with $^1/_2$-inch mesh weighs 20 pounds per 100 square feet and has a wind loading characteristic of 162 pounds with 60 mph winds. The filled aperture is 16%, and this material is useful to 2300 MHz.

A rather interesting material worthy of investigation is $^1/_4$-inch reinforced plastic. It weighs only 4 pounds per 100 square feet. The plastic melts with many universal solvents such as lacquer thinner. If a careful plastic-melting job is done, what remains is the $^1/_4$-inch spaced aluminum wires with a small blob of plastic at each junction to hold the matrix together.

There are some general considerations to be made in selecting surface materials:

1) Joints of screening do not have to make electrical contact. The horizontal wires reflect the horizontal wave. Skew polarizations are merely a combination of horizontal and vertical components which are thus reflected by the corresponding wires of the screening. To a horizontally polarized wave, the spacing and diameter of only the horizontal wires determine the reflection coefficient (see **Fig 85**). Many amateurs have the mistaken impression that screening materials that do not make electrical contact at their junctions are poor reflectors.
2) By measuring wire diameter and spacings between the wires, a calculation of percentage of aperture that is filled can be made. This will be one of the major determining factors of wind pressure when the surfacing material is dry.

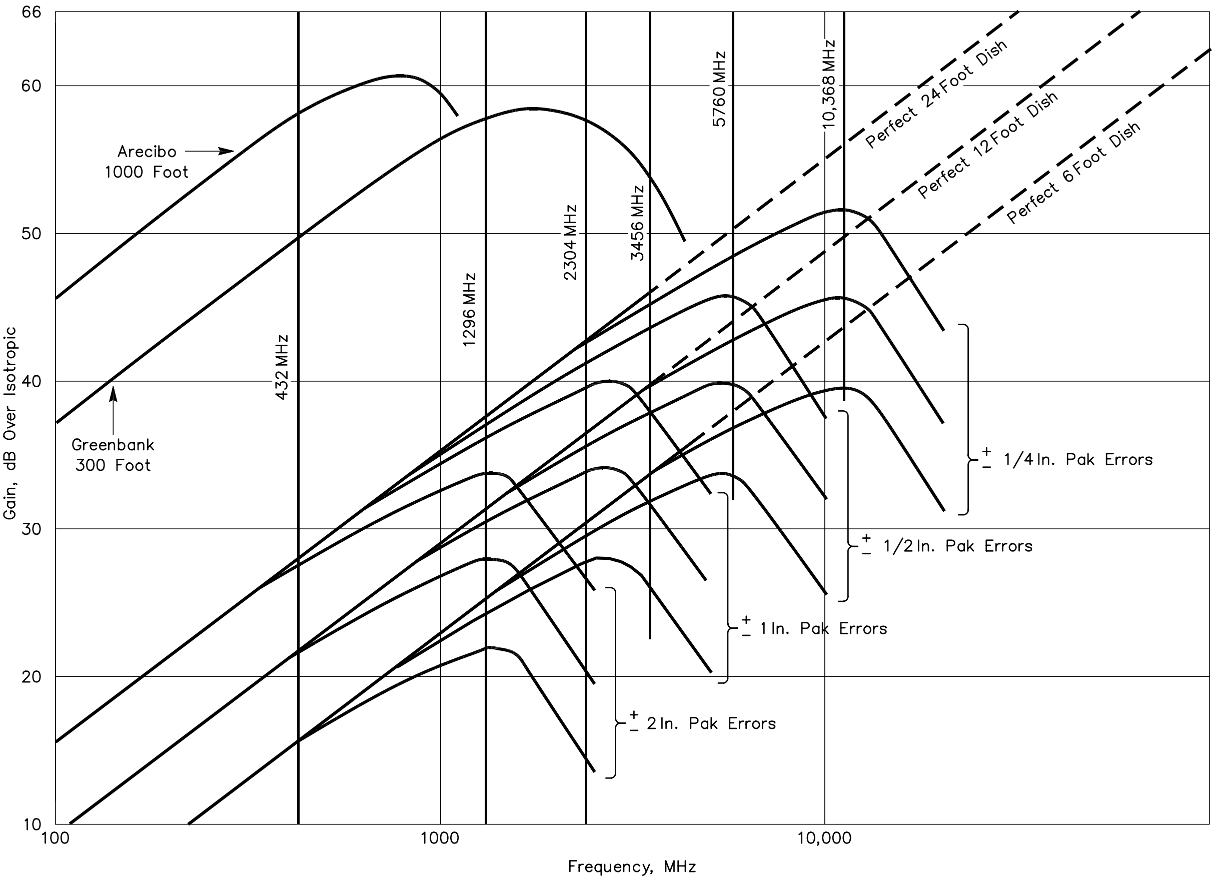

Fig 84—Parabolic-antenna gain versus size, frequency and surface errors. All curves assume 60% aperture efficiency and 10-dB power taper. Graph by K2RIW for ham bands, using display technique of J. Ruze, British IEE.

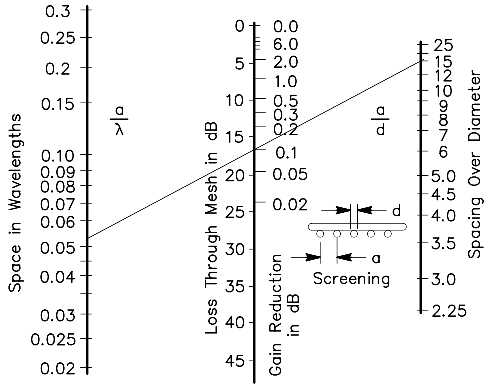

Fig 85—Surfacing material quality.

A Parabolic Template

At and above 2300 MHz (where high surface accuracy is required), a parabolic template should be constructed to measure surface errors. A simple template may be constructed (see **Fig 86**) by taking a 12-foot 3-inch length of 4-foot wide tar paper and drawing a parabolic shape on it with chalk. The points for the parabolic shape are calculated at 6-inch intervals and these points are connected with a smooth curve. For those who wish to use the template with the surface material installed, the template should be cut along the chalk line and stiffened by cardboard or a wood lattice frame. Surface error measurements should take place with all spokes installed and deflected by the fishing lines, as some bending of the center plates does take place. **Fig 87** shows the 12-foot stressed dish built by Franco Marcelo, N2UO.

Variations

All the possibilities of the stressed parabolic antenna have not been explored. For instance, a set of fishing lines or guy wires can be set up behind the dish for error correction, as long as this does not cause permanent bending of the aluminum spokes. This technique also protects the dish against wind loading from the rear. An extended piece of TV mast is an ideal place to hang a counterweight and attach the rear guys. This strengthens the structure considerably.

EME USING SURPLUS TVRO DISH ANTENNAS

Since the 1990s, there has been a significant change

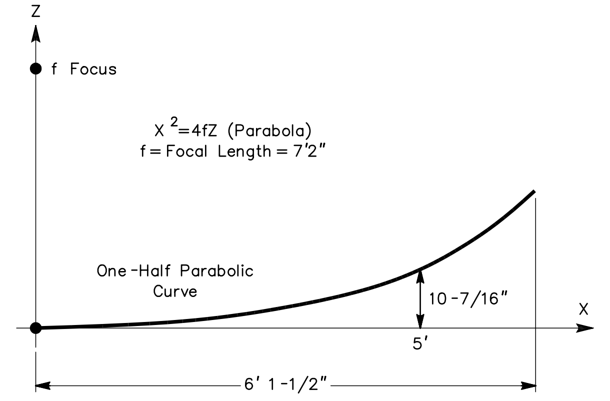

X	Z (to Nearest 1/16″)	X	Z
0.5′	1/8″	3.5′	5-1/8″
1.0′	7/16″	4.0′	6-11/16″
1.5′	15/16″	4.5′	8-1/2″
2.0′	1-11/16″	5.0′	10-7/16″
2.5′	2-5/8″	5.5′	12-11/16″
3.0′	3-3/4″	6.0′	15-1/16″

Fig 86—Parabolic template for 12-foot, 3-inch dish.

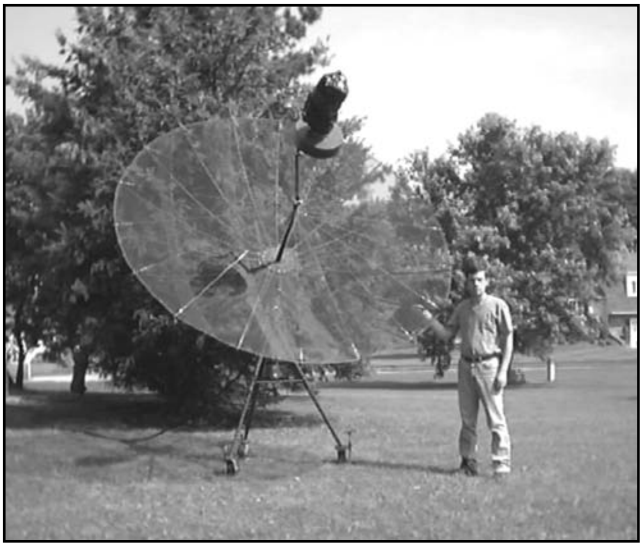

Fig 87—N2UO's homemade 12-foot stressed dish. (*Photo courtesy N2UO.*)

in the systems people use to watch satellite TV broadcasts. Formerly, C-band satellite receivers were used, along with parabolic dish antennas in the 3- to 5-meter diameter range. Now, Ku-band (12-GHz) receivers are the norm, with their associated small (usually 18-inch) dish antennas. This has provided a large body of surplus C-band dishes, which can be used for EME—certainly on the bands at 33 cm and above, and for the larger dishes (5 meters), even at 70 cm. Many times, these dishes and their mounts can be had for the asking, so they truly become an inexpensive way to build a multi-band EME antenna.

This updated article, first presented by David Hallidy, K2DH (ex-KD5RO) in the ARRL *UHF/Microwave Projects Manual*, describes the use of a 3-meter (10-foot) TVRO antenna in such an application. (Also see earlier in this chapter the section describing converted C-Band TVRO Dishes for satellite work.)

Background

Calculations show that a 3-meter dish will have about 30 dBi gain at 1296 MHz. With a state-of-the-art LNA (Low-Noise Amplifier or preamp) at the feed, an efficient feed horn illuminating the dish surface, and 200 W at 1296 MHz, lunar echoes should be easily detected and many stations can be worked. The biggest challenges to such a system are assembling the dish to its mount and steering it to track the Moon. As much as possible, the KISS ("Keep It Simple, Stupid") principle was used to accomplish this task.

In 1987, WA5TNY, KD5RO, KA5JPD, and W7CNK proved that such an EME system could work, even as high as 3.4 and 5.7 GHz, to provide the first EME contacts on those bands. An additional advantage to this (or any) small dish is its ability to be mounted to a trailer and taken out on EME expeditions. It can also be easily disassembled and stored, if necessary.

As can be seen from **Fig 88**, the entire setup is very simple, using a standard amateur tower as the main support for the dish.

Azimuth Drive

In azimuth, direct drive of the main rotating shaft was selected, and a small prop-pitch motor was used. These motors, while not as plentiful as they were some years ago, still turn up with some regularity at flea markets for very little money. The beauty of the prop-pitch motor is that it turns slowly, is reversible, provides very high torque, and requires no braking system (the gear reduction, on the order of 4000:1, provides the necessary braking). Prop-pitch motors are dc motors, and were designed to vary the pitch of propeller blades at start-up, take-off and landing of older large airplanes. Thus, they can be run at different speeds merely by varying the dc voltage to the motor, and can be reversed by reversing the polarity of the dc voltage. By mounting a thrust bearing of the appropriate size at the top of the tower, and mounting the motor directly below it at the end of the rotating shaft that turns the antenna, a simple direct-drive system can be constructed.

The dc power supply and control relays are located in a weatherproof box on the side of the tower, next to the motor. This system requires only 9 V dc at about 5 A to adequately start, turn and stop the prop-pitch motor, and this voltage turns the antenna through 360° of rotation in about $2^1/_2$ minutes. Azimuth position sensing is also a simple task. See **Fig 89**. A linear multi-turn potentiometer is driven by the rotating shaft, using a simple friction drive.

A strip of rubber is attached to the rotating shaft and a wheel is connected to the shaft of the pot. The pot is then mounted so that it presses against the rubber strip, and as the shaft turns so does the pot. If a ten-turn pot is used, and the system is aligned such that the pot is at the center of its rotation when the antenna is pointed approximately south, the pot will not rotate past the end at either extreme of the antenna's rotation (CW/CCW north), and absolute alignment is a simple task of calibrating the change in resistance (change in voltage, when the pot is fed from a constant voltage source) with degrees of rotation (see the discussion on Position Readout for details).

Elevation Drive

The elevation drive is also very simple. Most (nearly all) TVRO setups have a means of moving the dish across the sky to align it with various satellites. To do this, most companies use a device called a *Linear Actuator*. This is a dc motor to which is attached a long lead screw that pulls (or pushes) the outer shell of the actuator in or out to make it longer or shorter. The movable end of the actuator is attached to the dish and the motor end is fixed to the mount. The dish rests on pivots, which allow it to move as the actuator extends/retracts. To convert this type of mount (called a *Polar Mount*) to an Az/El mount is usually very simple.

Fig 90 shows how this can be done. Simply breaking the welds that held the mount in a polar fashion allows the mount to be turned on its side and used to pivot the dish vertically with the linear actuator. Another feature of linear actuators is that they also have some means of feeding their relative position to the satellite receiver. This is usually just a multi-turn potentiometer geared to the lead screw. All we have to do is connect this pot to a readout system, and we can calibrate the lift of the actuator in degrees. We thus have a simple means of rotating the dish and elevating it—but how do we know that it's pointed at the Moon?

Position Readout

Readout of the position of the antenna, in both azimuth and elevation is also a relatively simple task. On the

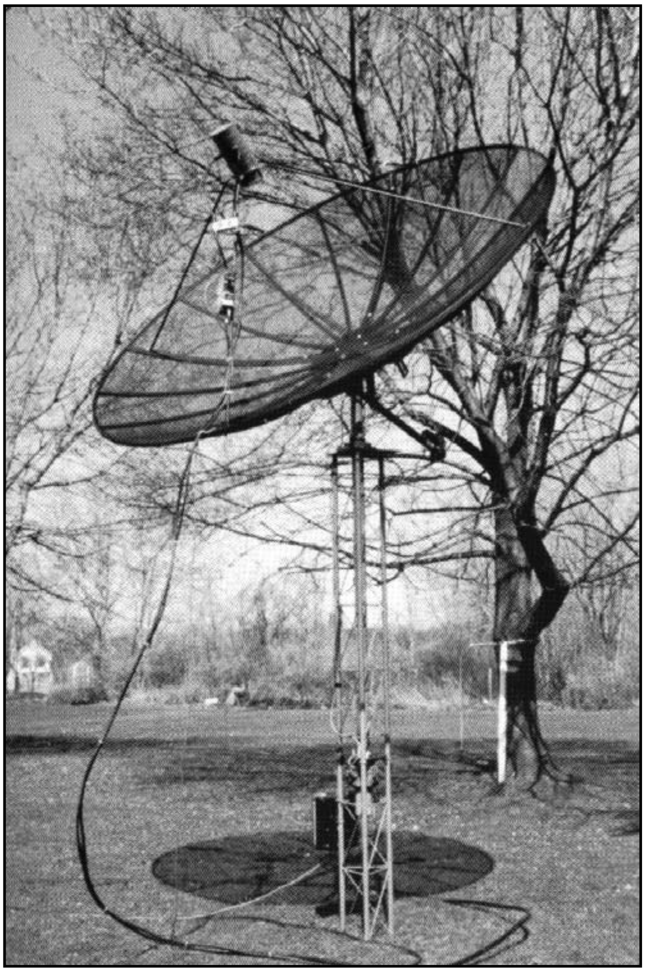

Fig 88—View of K2DH's (ex-KD5RO) complete TVRO antenna installation. (*K2DH photo.*)

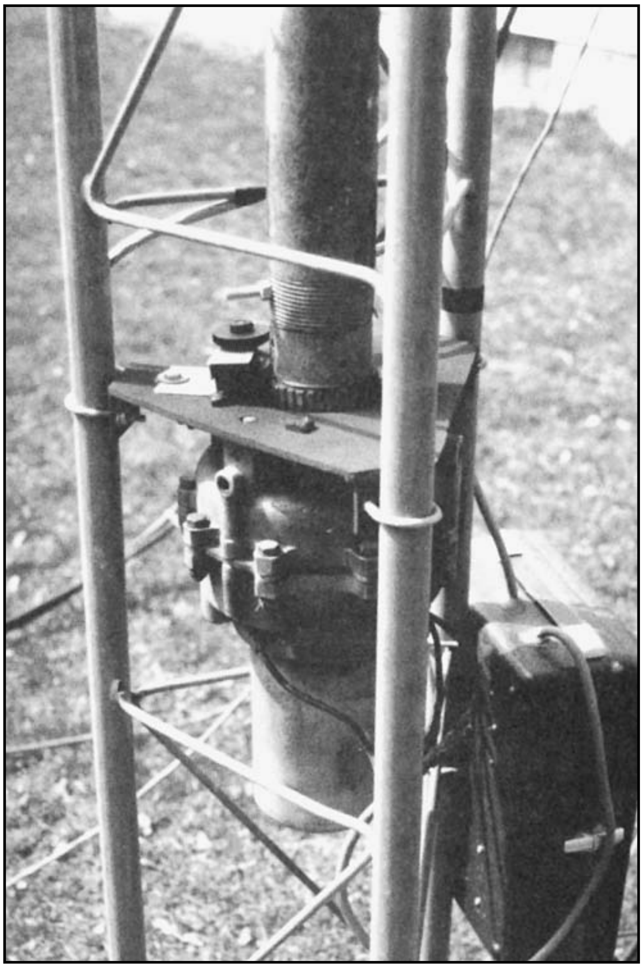

Fig 89—Azimuth rotation systems, showing prop-pitch motor and position sensor.

surplus market there are available Digital Volt Meters (DVMs) using LED or LCD displays that can do this job nicely, and that have more precision than is probably necessary for a dish (or Yagi array) of small size. As mentioned earlier, a multi-turn potentiometer on the elevation-drive mechanism can be used to readout elevation, and the same technique can be used for azimuth readout—a potentiometer coupled to the main rotating shaft that turns the antenna.

When using a pot for readout, the most important thing to know is how many degrees of antenna position change occur (in Az or El) for each turn of the pot. This then can be used to calibrate a voltmeter to read volts directly as degrees—for example, 3.60 V could correspond to 360° azimuth (Clockwise North), and 9.0 V could correspond

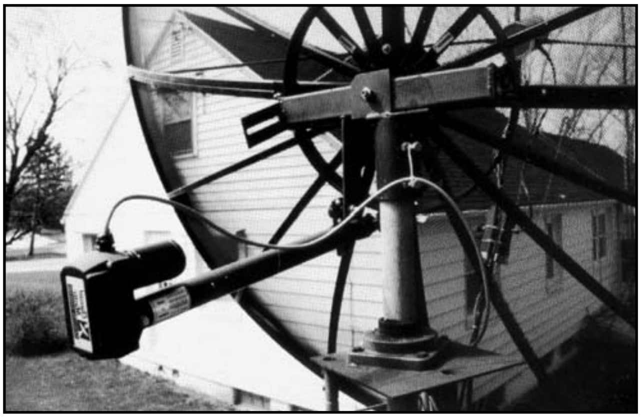

Fig 90—Elevation system, showing modified TVRO mount.

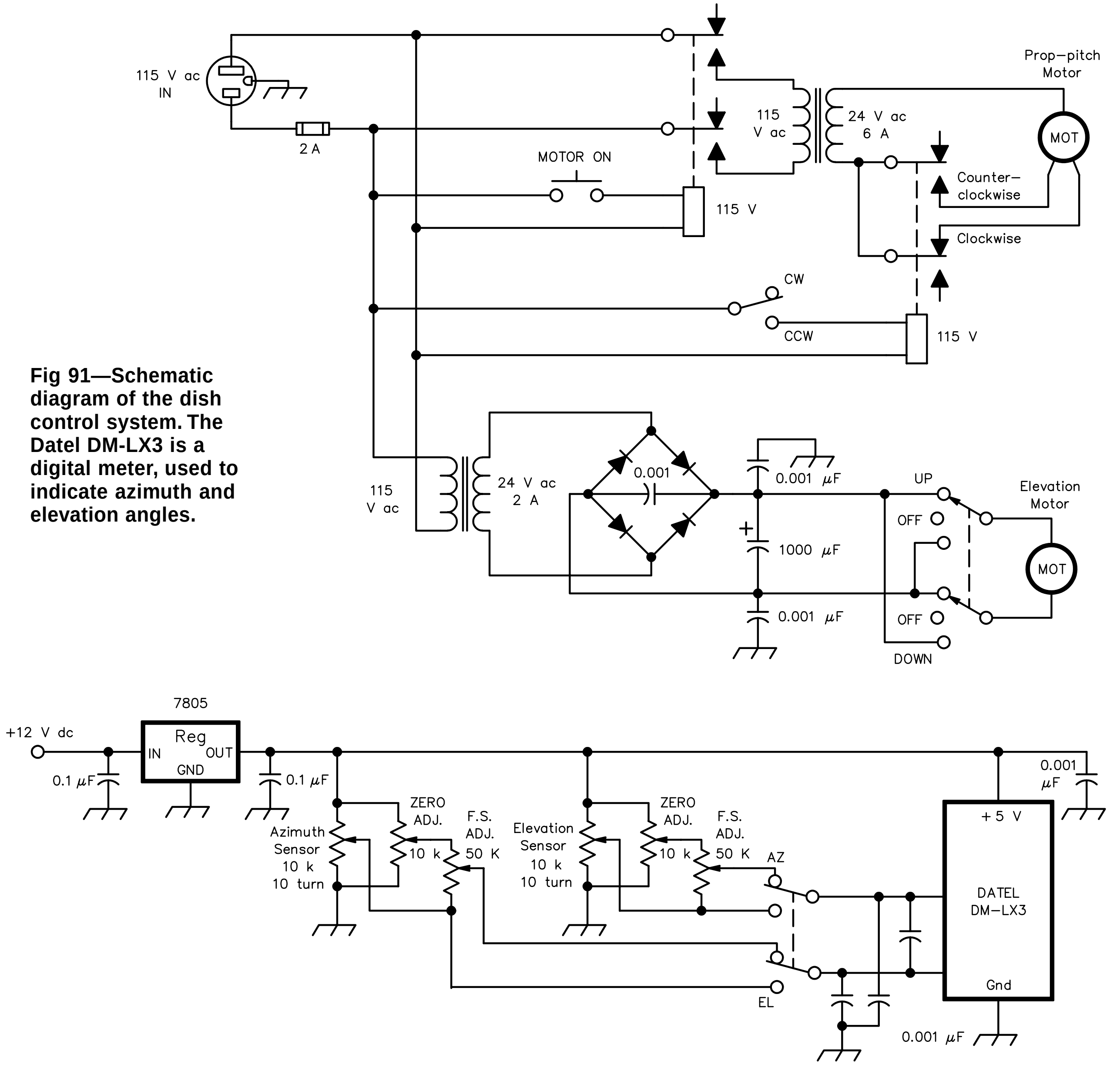

115 V ac
IN
2 A
MOTOR ON
115 V
115
V ac
24 V ac
6 A
Prop–pitch
Motor
MOT
Counter–
clockwise
Clockwise
CW
CCW
115 V
115
V ac
24 V ac
2 A
0.001
0.001 µF
1000 µF
0.001 µF
UP
OFF
OFF
DOWN
Elevation
Motor
MOT
7805
+12 V dc
Reg
IN
OUT
GND
0.1 µF
0.1 µF
ZERO
ADJ.
F.S.
ADJ.
Azimuth
Sensor
10 k
10 turn
10 k
50 K
Elevation
Sensor
10 k
10 turn
ZERO
ADJ.
F.S.
ADJ.
10 k
50 K
AZ
EL
0.001
µF
+5 V
DATEL
DM–LX3
Gnd
0.001 µF

Fig 91—Schematic diagram of the dish control system. The Datel DM-LX3 is a digital meter, used to indicate azimuth and elevation angles.

to 90° elevation (straight up).

A resistance bridge circuit is best used in this application, since it is less sensitive to changes in the supply voltage. The only thing to be careful about is that the DVM must have both the positive (high) and negative (low) inputs isolated from ground (assuming the power supply used to power the DVM is grounded). You could also use a pair of small, cheap Digital Multi-Meters (DMMs), which can sometimes be found for under $10. Because they are battery powered, the isolation issue just discussed is eliminated.

Please see **Fig 91** for a complete schematic of the azimuth, elevation and readout electronics for this antenna-drive system. Also note that while this discussion is geared towards the use of a small dish, the same positioning and readout systems could be used in a Yagi array for 2 meters or 70 cm.

Now that we know where the dish is pointed, how do we know where the Moon is? There are several software programs available to the Amateur for tracking celestial bodies such as the Moon, the Sun, certain stars (usable as noise sources), and even Amateur Satellites. Programs by W9IP, VK3UM, F1EHN and others can be obtained very reasonably and these work well to provide highly accurate position information for tracking.

Feeding the Surplus TVRO Dish

An area that needs particular attention when attempting EME with a small dish is an efficient feed system. An efficient feed system can be a real challenge with TVRO dishes, because many are "deep"—that is, their f/D (focal length to Diameter ratio) is small.

The satellite TV industry used deep dishes because they tend to be quieter, picking up less Earth noise due to spillover effects. A deep dish has a short focal length, and therefore, the feed is relatively close to the surface of the dish. To properly illuminate the reflector out to its edges, a feed horn of relatively wide beamwidth must be used. The feeds designed several years ago by Barry Malowanchuk, VE4MA, are intended for use with just such dishes, and have the advantage of being adjustable to optimize their pattern to the dish in use.

The feed that was used with this dish was modeled after VE4MA's 1296-MHz feed, and a version was even scaled for use at 2304 MHz that worked as well as the original. See **Fig 92** and also see the Notes and References section at the end of this chapter. (Also see the earlier section in the satellite portion of this chapter describing patch feeds for small dishes.)

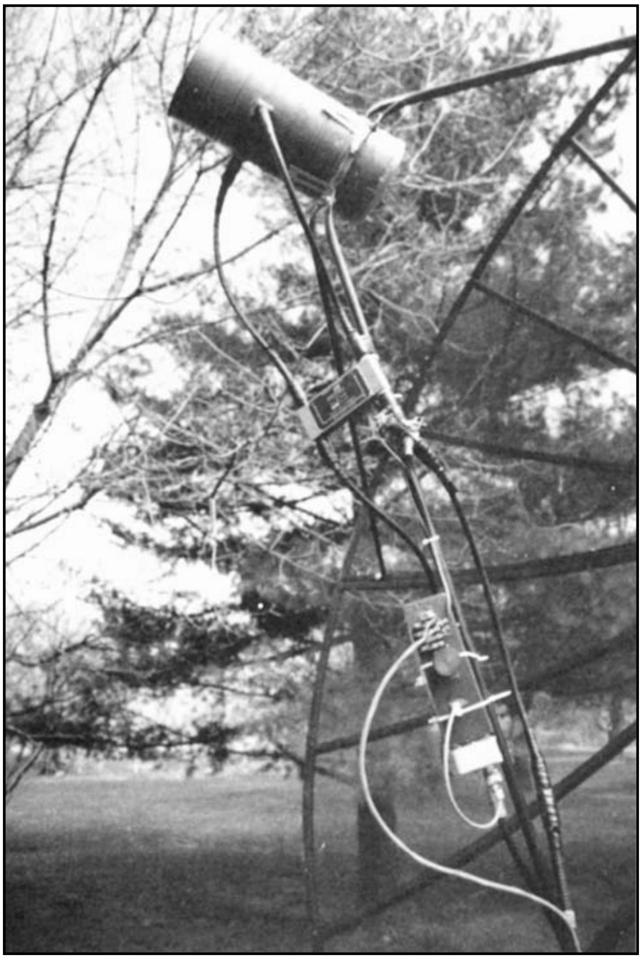

Fig 92—View of feed, showing coffee-can feed horn and hybrid coupler.

SHF EME CHALLENGES

The challenges met when successfully building a station for EME at 900 MHz to 5.7 GHz only become more significant on the SHF bands at 10 GHz and above. Absolute attention to detail is the primary requirement, and this extends to every aspect of the EME antenna system. The dish surface is probably the most difficult problem to solve. As was discussed earlier in this chapter, the shape and accuracy of the reflector contribute directly to the overall gain of the antenna.

But where slight errors in construction can be tolerated at the lower frequencies, the same cannot be said at millimetric wavelengths. Those who have attempted EME on 10 and 24 GHz have discovered that the weight of the dish reflector itself will distort its shape enough to lower the gain to the point where echoes are degraded. Stiffening structures at the back of such dishes are often found necessary. **Fig 93** illustrates the back struts added by Al Ward, W5LUA, to strengthen his dish.

Pointing accuracy is also paramount—a 16-foot dish at 10 GHz has a beamwidth about equal to the diameter of the Moon—0.5°. This means that the echo degradation due to the Moon's movement away from where the dish is pointed is almost immediate, and autotracking systems become more of a necessity than a luxury. At these frequencies, most amateurs actually peak their antennas on

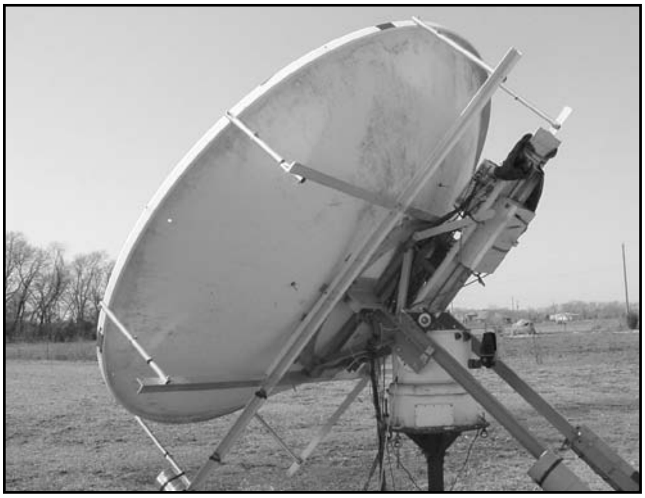

Fig 93—Strengthening struts W5LUA added to the back of his dish to hold down distortion. (*Photo courtesy W5LUA.*)

Moon Noise—the black-body Radiation from the Moon that becomes the dominant source of noise in space.

At these frequencies, the elevation of the Moon above the horizon also plays a role in the ability to communicate, since tropospheric absorption due to water vapor is greatest at low elevation angles (the signal must pass through a greater portion of the troposphere than when the Moon is highly elevated). It is beyond the abilities of most Amateurs to construct their own dishes for these frequencies, so surplus dishes for Ku-band satellite TV (typically 3 meters in diameter) are usually employed, as have high-performance dishes designed for millimetric radar and point-to-point communications at 23 and 38 GHz.

NOTES AND REFERENCES, EME ANTENNAS

W. Allen, "A Mode J Helix," *AMSAT Newsletter*, Jun 1979, pp 30-31.

D. J. Angelakos and D. Kajfez, "Modifications on the Axial-Mode Helical Antenna," *Proc. IEEE*, Apr 1967, pp 558-559.

A. L. Bridges, "Really Zap OSCAR With this Helical Antenna," *73*, in three parts, Jul, Aug and Sep 1975.

K. R. Carver, "The Helicone—A Circularly Polarized Antenna with Low Sidelobe Level," *Proc. IEEE*, Apr 1967, p 559.

M. R. Davidoff, "A Simple 146-MHz Antenna for Oscar Ground Stations," *QST*, Sep 1974, pp 11-13.

M. R. Davidoff, *The Radio Amateur's Satellite Handbook* (Newington: ARRL, 1998-2000).

D. DeMaw, "The Basic Helical Beam," *QST*, Nov 1965, pp 20-25, 170.

D. Evans and G. Jessop, *VHF-UHF Manual, 3rd ed.* (London: RSGB), 1976.

N. Foot, "WA9HUV 12-Foot Dish for 432 and 1296 MHz," The World Above 50 Mc., *QST*, Jun 1971, pp 98-101, 107.

N. Foot, "Cylindrical Feed horn for Parabolic Reflectors," *Ham Radio*, May 1976.

O. J. Glasser and J. D. Kraus, "Measured Impedances of Helical Beam Antennas," *Journal of Applied Physics*, Feb 1948, pp 193-197.

D. Hallidy, "Microwave EME Using A Ten-Foot TVRO Antenna," *The ARRL UHF/Microwave Projects Manual*, 1994 pp 10-9 to 10-13.

G. R. Isely and W. G. Smith, "A Helical Antenna for Space Shuttle Communications," *QST*, Dec 1984, pp 14-18.

H. Jasik, *Antenna Engineering Handbook*, 1st ed. (New York: McGraw-Hill, 1961).

R. T. Knadle, "A Twelve-Foot Stressed Parabolic Dish," *QST*, Aug 1972, pp 16-22.

J. D. Kraus, "Helical Beam Antenna," *Electronics*, Apr 1947, pp 109-111.

J. D. Kraus and J. C. Williamson, "Characteristics of Helical Antennas Radiating in the Axial Mode," *Journal of Applied Physics*, Jan 1948, pp 87-96.

J. D. Kraus, "Helical Beam Antenna for Wide Band Applications," *Proc of the IRE*, Oct 1948, pp 1236-1242.

J. D. Kraus, *Antennas* (New York: McGraw-Hill Book Co., 1950).

J. D. Kraus, "A 50-Ohm Input Impedance for Helical Beam Antenna," *IEEE Transactions on Antennas and Propagation*, Nov 1977, p 913.

J. D. Kraus, *Big Ear* (Powell, OH: Cygnus-Quasar Books, 1976).

T. S. M. MacLean, "Measurements on High-Gain Helical Aerial and on Helicals of Triangular Section," *Proc. of IEE* (London), Jul 1964, pp 1267-1270.

D. M. Mallozzi, "The Tailored Helical," *AMSAT Newsletter*, Mar 1978, pp 8-9.

B. Malowanchuk, "Use of Small TVRO Dishes for EME," *Proceedings of the 21st Conference of The Central States VHF Society*, 1987 pp 68-77.

B. Malowanchuk, "Selection of An Optimum Dish Feed," *Proceedings of the 23rd Conference of The Central States VHF Society*, 1989 pp 35-43.

M. W. Maxwell, *Reflections—Transmission Lines and Antennas* (Newington, CT: ARRL, 1990) [out of print].

T. Moreno, *Microwave Transmission Design Data* (New York: McGraw-Hill, 1948).

K. Nose, "Crossed Yagi Antennas for Circular Polarization," *QST*, Jan 1973, pp 11-12.

K. Nose, "A Simple Az-El Antenna System for Oscar," *QST*, Jun 1973, pp 11-12.

S. Powlishen, K1FO, "432-MHz EME 1990s Style," Parts 1 and 2, *Communications Quarterly*, Oct 1990, pp 29-39; Oct 1991, pp 33-48.

S. Powlishen, K1FO, "Rear-Mount Yagi Arrays for 432-MHz EME: Solving the EME Polarization Problem," *The ARRL Antenna Compendium, Vol 3* (Newington:

ARRL, 1992), pp 79-98.

C. Richards, “The 10 Turn Chopstick Helical (Mk2) for OSCAR 10 432 MHz Uplink,” *Radio Communication*, Oct 1984, pp 844-845.

S. Sander and D. K. Cheng, “Phase Center of Helical Beam Antennas,” *IRE National Convention Record Part 6*, 1958, pp 152-157.

E. A. Scott and H. E. Banta, “Using the Helical Antenna at 1215 Mc.,” *QST*, Jul 1962, pp 14-16.

R. Soifer, W2RS, “More ‘QRP’ EME on 144 MHz,” *QST*, Oct 1990, pp 36-38.

G. Southworth, *Principles and Applications of Waveguide Transmission* (New York: D. Van Nostrand Co., 1950).

B. Sykes, “Circular Polarization and Crossed-Yagi Antennas,” Technical Topics, *Radio Communication*, Feb 1985, p 114.

H. E. Taylor and D. Fowler, “A V-H-F Helical Beam Antenna,” *CQ*, Apr 1949, pp 13-16.

G. Tillitson, “The Polarization Diplexer—A Polaplexer,” *Ham Radio*, Mar 1977.

P. P. Viezbicke, “Yagi Antenna Design,” *NBS Technical Note 688* (U. S. Dept. of Commerce/National Bureau of Standards, Boulder, CO), Dec 1976.

D. Vilardi, “Easily Constructed Antennas for 1296 MHz,” *QST*, Jun 1969.

D. Vilardi, “Simple and Efficient Feed for Parabolic Antennas,” *QST*, Mar 1973.

J. L. Wong and H. E. King, “Broadband Quasi-Tapered Helical Antenna,” *IEEE Transactions on Antennas and Propagation*, Jan 1979, pp 72-78.

Chapter 20

Antenna Materials and Accessories

This chapter contains information on materials amateurs use to construct antennas—what types of material to look for in a particular application, tips on working with and using various materials. Chapter 21 contains information on where to purchase these materials.

Basically, antennas for MF, HF, VHF and the lower UHF range consist simply of one or more conductors that radiate (or receive) electromagnetic waves. However, an antenna system must also include some means to support those conductors and maintain their relative positions—the boom for a Yagi antenna and the halyards for a wire dipole, for example. In this chapter we'll look at materials for those applications, too. Structural supports such as towers, masts, poles, etc. are discussed in Chapter 22.

There are two main types of material used for antenna conductors, wire and tubing. Wire antennas are generally simple and therefore easier to construct, although some arrays of wire elements can become rather complex. When tubing is required, aluminum tubing is used most often because of its light weight. Aluminum tubing is discussed in a subsequent section of this chapter.

Wire Antennas

Although wire antennas are relatively simple, they can constitute a potential hazard unless properly constructed. Antennas should never be run under or over public utility (telephone or power) lines. Several amateurs have lost their lives by failing to observe this precaution.

The National Electric Code® of the National Fire Protection Association contains a section on amateur stations in which a number of recommendations are made concerning minimum size of antenna wire and the manner of bringing the transmission line into the station. Chapter 1 contains more information about this code. The code in itself does not have the force of law, but it is frequently made a part of local building regulations, which are enforceable. The provisions of the code may also be written into, or referred to, in fire and liability insurance documents.

The RF resistance of copper wire increases as the size of the wire decreases. However, in most types of antennas that are commonly constructed of wire (even quite small wire), the radiation resistance will be much higher than the RF resistance, and the efficiency of the antenna will still be adequate. Wire sizes as small as #30, or even smaller, have been used quite successfully in the construction of "invisible" antennas in areas where more conventional antennas cannot be erected. In most cases, the selection of wire for an antenna will be based primarily on the physical properties of the wire, since the suspension of wire from elevated supports places a strain on the wire.

WIRE TYPES

Wire having an enamel coating is preferable to bare wire, since the coating resists oxidation and corrosion. Several types of wire having this type of coating are available, depending on the strength needed. "Soft-drawn" or annealed copper wire is easiest to handle; unfortunately, it stretches considerably under stress. Soft-drawn wire should be avoided, except for applications where the wire will be under little or no tension, or where some change in length can be tolerated. (For example, the length of a horizontal antenna fed at the center with open-wire line is not critical, although a change in length may require some readjustment of coupling to the transmitter.)

"Hard-drawn" copper wire or copper-clad steel wire (also known as Copperweld™) is harder to handle, because it has a tendency to spiral when it is unrolled. These types of wire are ideal for applications where significant stretch cannot be tolerated. Care should be exercised in using this wire to make sure that kinks do not develop—the wire will have a far greater tendency to break at a kink. After the coil has been unwound, suspend the wire a few feet above ground for a day or two before using it. The wire should not be recoiled before it is installed.

Several factors influence the choice of wire type and size. Most important to consider are the length of the unsupported span, the amount of sag that can be tolerated, the stability of the supports under wind pressure, and whether or not an unsupported transmission line is to be suspended from the span. **Table 1** shows the wire diam, current-carrying capacity and resistance of various sizes of copper wire. **Table 2** shows the maximum rated working tensions of hard-drawn and copper-clad steel wire of various sizes. These two tables can be used to select the appropriate wire size for an antenna.

Table 1
Copper-Wire Table

Wire Size AWG (B&S)	*Dia in Mils*[1]	*Dia in mm*	*Turns per Linear Inch Enamel*	*Feet per Pound Bare*	*Ohms per 1000 ft 25°C*	*Cont.-duty current*[2] *Single Wire in Open Air*
1	289.3	7.348	—	3.947	0.1264	—
2	257.6	6.544	—	4.977	0.1593	—
3	229.4	5.827	—	6.276	0.2009	—
4	204.3	5.189	—	7.914	0.2533	—
5	181.9	4.621	—	9.980	0.3195	—
6	162.0	4.115	—	12.58	0.4028	—
7	144.3	3.665	—	15.87	0.5080	—
8	128.5	3.264	7.6	20.01	0.6405	73
9	114.4	2.906	8.6	25.23	0.8077	—
10	101.9	2.588	9.6	31.82	1.018	55
11	90.7	2.305	10.7	40.12	1.284	—
12	80.8	2.053	12.0	50.59	1.619	41
13	72.0	1.828	13.5	63.80	2.042	—
14	64.1	1.628	15.0	80.44	2.575	32
15	57.1	1.450	16.8	101.4	3.247	—
16	50.8	1.291	18.9	127.9	4.094	22
17	45.3	1.150	21.2	161.3	5.163	—
18	40.3	1.024	23.6	203.4	6.510	16
19	35.9	0.912	26.4	256.5	8.210	—
20	32.0	0.812	29.4	323.4	10.35	11
21	28.5	0.723	33.1	407.8	13.05	—
22	25.3	0.644	37.0	514.2	16.46	—
23	22.6	0.573	41.3	648.4	20.76	—
24	20.1	0.511	46.3	817.7	26.17	—
25	17.9	0.455	51.7	1031	33.00	—
26	15.9	0.405	58.0	1300	41.62	—
27	14.2	0.361	64.9	1639	52.48	—
28	12.6	0.321	72.7	2067	66.17	—
29	11.3	0.286	81.6	2607	83.44	—
30	10.0	0.255	90.5	3287	105.2	—
31	8.9	0.227	101	4145	132.7	—
32	8.0	0.202	113	5227	167.3	—
33	7.1	0.180	127	6591	211.0	—
34	6.3	0.160	143	8310	266.0	—
35	5.6	0.143	158	10480	335	—
36	5.0	0.127	175	13210	423	—
37	4.5	0.113	198	16660	533	—
38	4.0	0.101	224	21010	673	—
39	3.5	0.090	248	26500	848	—
40	3.1	0.080	282	33410	1070	—

[1] A mil is 0.001 inch.

[2] Max wire temp of 212° F and max ambient temp of 135° F.

Table 2

Stressed Antenna Wire

American Wire Gauge	***Recommended Tension[1] (pounds)***		***Weight (pounds per 1000 feet)***	
	Copper-clad steel[2]	*Hard-drawn copper*	*Copper-clad steel[2]*	*Hard-drawn copper*
4	495	214	115.8	126.0
6	310	130	72.9	79.5
8	195	84	45.5	50.0
10	120	52	28.8	31.4
12	75	32	18.1	19.8
14	50	20	11.4	12.4
16	31	13	7.1	7.8
18	19	8	4.5	4.9
20	12	5	2.8	3.1

[1] Approximately one-tenth the breaking load. Might be increased 50% if end supports are firm and there is no danger of ice loading.

[2] Copperweld,™ 40% copper.

Wire Tension

If the tension on a wire can be adjusted to a known value, the expected sag of the wire (**Fig 1**) may be determined before installation using Table 2 and the nomograph of **Fig 2**. Even though there may be no convenient method to determine the tension in pounds, calculation of the expected sag for practicable working tensions is often desirable. If the calculated sag is greater than allowable it may be reduced by any one or a combination of the following:

1) Providing additional supports, thereby decreasing the span
2) Increasing the tension in the wire if less than recommended
3) Decreasing the size of the wire

Instructions for Using the Nomograph

1) From Table 2, find the weight (pounds/1000 feet) for the particular wire size and material to be used.
2) Draw a line from the value obtained above, plotted on the weight axis, to the desired span (feet) on the span axis, Fig 2. Note in Fig 1 that the span is one half the distance between the supports.
3) Choose an operating tension level (in pounds) consistent with the values presented in Table 2 (preferably less than the recommended wire tension).
4) Draw a line from the tension value chosen (plotted on the tension axis) through the point where the work axis crosses the original line constructed in step 2, and continue this new line to the sag axis.
5) Read the sag in feet on the sag axis.

Example:

Weight = 11 pounds/1000 feet
Span = 210 feet
Tension = 50 pounds
Answer: Sag = 4.7 feet

These calculations do not take into account the weight of a feed line supported by the antenna wire.

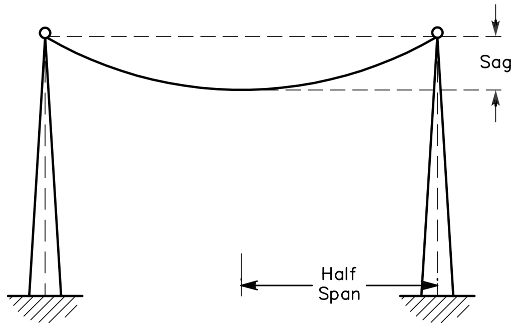

Fig 1—The half span and sag of a long-wire antenna.

Wire Splicing

Wire antennas should preferably be made with unbroken lengths of wire. In instances where this is not feasible, wire sections should be spliced as shown in **Fig 3**. The enamel insulation should be removed for a distance of about 6 inches from the end of each section by scraping with a knife or rubbing with sandpaper until the copper underneath is bright. The turns of wire should be brought up tight around the standing part of the wire by twisting with broad-nose pliers.

The crevices formed by the wire should be completely filled with rosin-core solder. An ordinary soldering iron or gun may not provide sufficient heat to melt solder outdoors; a propane torch is desirable. The joint should be heated sufficiently so the solder flows freely into the joint when the source of heat is removed momentarily. After the joint has cooled completely, it should be wiped clean with a cloth, and then sprayed generously with acrylic to prevent corrosion.

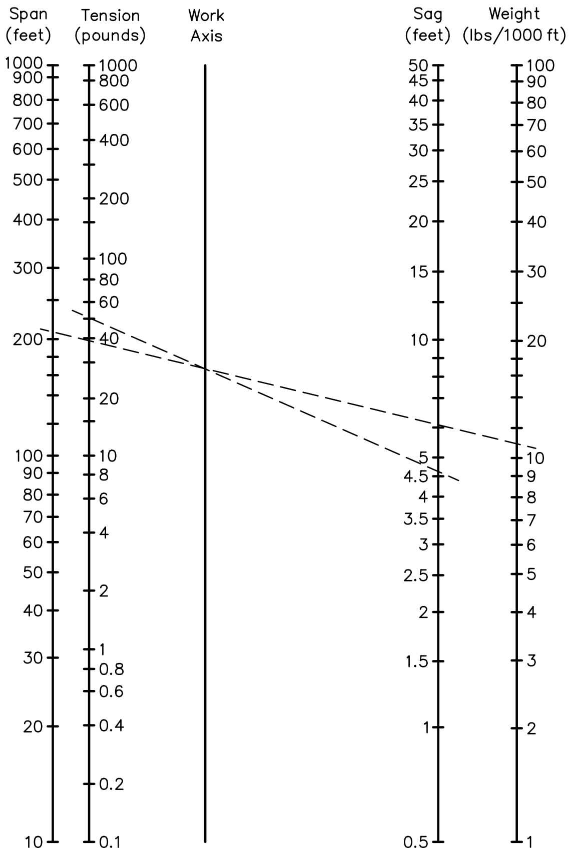

Fig 2—Nomograph for determining wire sag. (John Elengo, Jr, K1AFR)

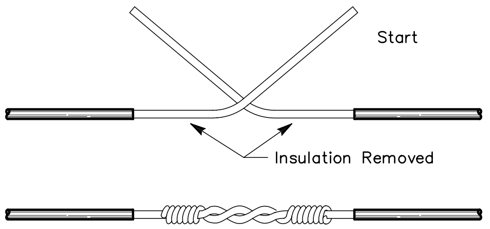

Fig 3—Correct method of splicing antenna wire. Solder should be flowed into the wraps after the connection is completed. After cooling, the joint should be sprayed with acrylic to prevent oxidation and corrosion.

ANTENNA INSULATION

To prevent loss of RF power, the antenna should be well insulated from ground, unless of course it is a shunt-fed system. This is particularly important at the outer end or ends of wire antennas, since these points are always at a comparatively high RF potential. If an antenna is to be installed indoors (in an attic, for instance) the antenna may be suspended directly from the wood rafters without additional insulation, if the wood is permanently dry. Much greater care should be given to the selection of proper insulators when the antenna is located outside where it is exposed to wet weather.

Insulator Leakage

Antenna insulators should be made of material that will not absorb moisture. The best insulators for antenna use are made of glass or glazed porcelain. Depending on the type of material, plastic insulators may be suitable. The length of an insulator relative to its surface area is indicative of its comparative insulating ability. A long thin insulator will have less leakage than a short thick insulator. Some antenna insulators are deeply ribbed to increase the surface leakage path without increasing the physical length of the insulator. Shorter insulators can be used at low-potential points, such as at the center of a dipole. If such an antenna is to be fed with open-wire line and used on several bands, however, the center insulator should be the same as those used at the ends, because high RF potential may exist across the center insulator on some bands.

Insulator Stress

As with the antenna wire, the insulator must have sufficient physical strength to support the stress of the antenna without danger of breakage. Long elastic bands or lengths of nylon fishing line provide long leakage paths and make satisfactory insulators within their limits to resist mechanical strain. They are often used in antennas of the "invisible" type mentioned earlier.

For low-power work with short antennas not subject to appreciable stress, almost any small glass or glazed-porcelain insulator will do. Homemade insulators of Lucite rod or sheet will also be satisfactory. More care is required in the selection of insulators for longer spans and higher transmitter power.

For a given material, the breaking tension of an insulator will be proportional to its cross-sectional area. It should be remembered, however, that the wire hole at the end of the insulator decreases the effective cross-sectional area. For this reason, insulators designed to carry heavy strains are fitted with heavy metal end caps, the eyes being formed in the metal cap, rather than in the insulating material itself. The following stress ratings of antenna insulators are typical:

$^{5}/_{8}$ in. square by 4 in. long—400 lb
1 in. diameter by 7 or 12 in. long—800 lb
1$^{1}/_{2}$ in. diameter by 8, 12 or 20 in. long, with special metal end caps—5000 lb

These are rated breaking tensions. The actual working tensions should be limited to not more than 25% of the breaking rating.

The antenna wire should be attached to the insulators as shown in **Fig 4**. Care should be taken to avoid sharp angular bends in the wire when it is looped through the insulator eye. The loop should be generous enough in size that it will not bind the end of the insulator tightly. If the length of the antenna is critical, the length should be measured to the outward end of the loop, where it passes through the eye of the insulator. The soldering should be done as described earlier for the wire splice.

Strain Insulators

Strain insulators have their holes at right angles, since they are designed to be connected as shown in **Fig 5**. It can be seen that this arrangement places the insulating material under compression, rather than tension. An insulator connected this way can withstand much greater stress. Furthermore, the wire will not collapse if the insulator breaks, since the two wire loops are inter-locked. Because the wire is wrapped around the insulator, however, the leakage path is reduced drastically, and the capacitance between the wire loops provides an additional leakage path. For this reason, the use of the strain insulator is usually confined to such applications as breaking up resonances in guy wires, where high levels of stress prevail, and where the RF insulation is of less importance. Such insulators might be suitable for use at low-potential points on an antenna, such as at the center of a dipole. These insulators may also be fastened in the conventional manner if the wire will not be under sufficient tension to break out the eyes.

Insulators for Ribbon-Line Antennas

Fig 6A shows the sketch of an insulator designed to be used at the ends of a folded dipole or a multiple dipole made of ribbon line. It should be made approximately as shown, out of Lucite or bakelite material about 1/4 inch thick. The advantage of this arrangement is that the strain of the antenna is shared by the conductors and the plastic webbing of the ribbon, which adds considerable strength. After soldering, the screw should be sprayed with acrylic.

Fig 6B shows a similar arrangement for suspending one dipole from another in a stagger-tuned dipole system. If better insulation is desired, these insulators can be wired to a conventional insulator.

PULLEYS AND HALYARDS

Pulleys and halyards commonly used to raise and lower a wire antenna must also be capable of taking the

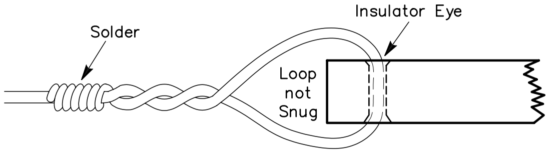

Fig 4—When fastening antenna wire to an insulator, do not make the wire loop too snug. After the connection is complete, flow solder into the turns. Then when the joint has cooled completely, spray it with acrylic.

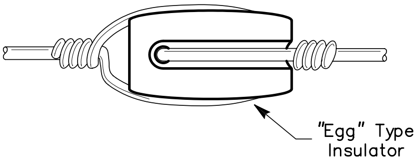

Fig 5—Conventional manner of fastening wire to a strain insulator. This method decreases the leakage path and increases capacitance, as discussed in the text.

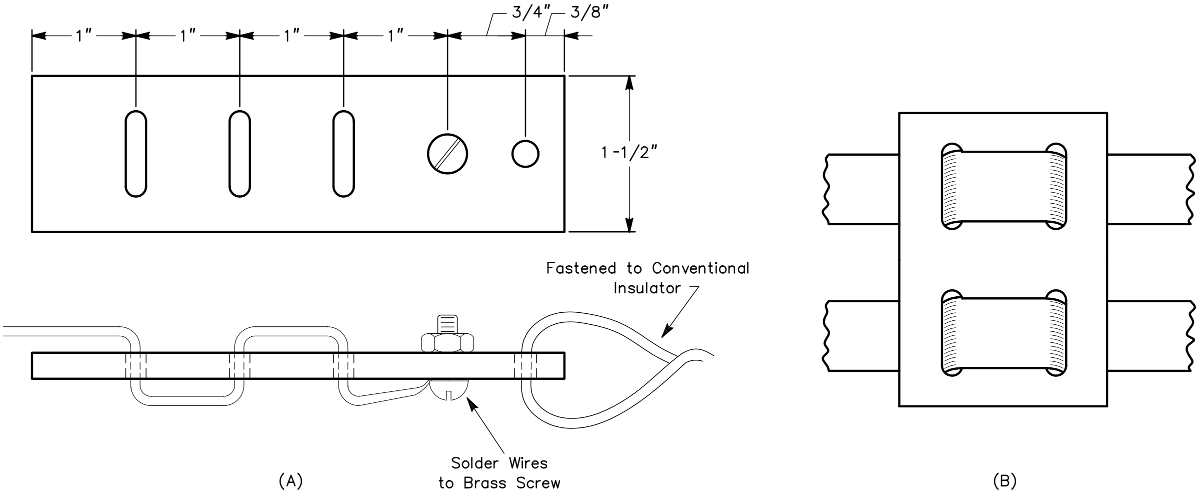

Fig 6—At A, an insulator for the ends of folded dipoles, or multiple dipoles made of 300-ohm ribbon. At B, a method of suspending one ribbon dipole from another in a multiband dipole system.

same strain as the antenna wire and insulators. Unfortunately, little specific information on the stress ratings of most pulleys is available. Several types of pulleys are readily available at almost any hardware store. Among these are small galvanized pulleys designed for awnings and several styles and sizes of clothesline pulleys. Heavier and stronger pulleys are those used in marine work. The factors that determine how much stress a pulley will handle include the diameter of the shaft, how securely the shaft is fitted into the sheath and the size and material of the frame.

Another important factor to be considered in the selection of a pulley is its ability to resist corrosion. Galvanized awning pulleys are probably the most susceptible to corrosion. While the frame or sheath usually stands up well, these pulleys usually fail at the shaft. The shaft rusts out, allowing the grooved wheel to break away under tension.

Most good-quality clothesline pulleys are made of alloys which do not corrode readily. Since they are designed to carry at least 50 feet of line loaded with wet clothing in stiff winds, they should be adequate for normal spans of 100 to 150 feet between stable supports. One type of clothesline pulley has a 4-inch diameter plastic wheel with a ¼-inch shaft running in bronze bearings. The sheath is made of cast or forged corrosion-proof alloy. Some look-alike low-cost pulleys of this type have an aluminum shaft with no bearings. For antenna work, these cheap pulleys are of little long-term value.

Marine pulleys have good weather-resisting qualities, since they are usually made of bronze, but they are comparatively expensive and are not designed to carry heavy loads. For extremely long spans, the wood-sheathed pulleys used in "block and tackle" devices and for sail hoisting should work well.

Table 3
Approximate Safe Working Tension for Various Halyard Materials

Material	*Dia, In.*	*Tension, Lb*
Manila hemp rope	1/4	120
	3/8	270
	1/2	530
	5/8	800
Polypropylene rope	1/4	270
	3/8	530
	1/2	840
Nylon rope	1/4	300
	3/8	660
	1/2	1140
7×11 galvanized sash cord	1/16	30
	1/8	125
	3/16	250
	1/4	450
High-strength stranded galvanized steel guy wire	1/8	400
	3/16	700
	1/4	1200
Rayon-filled plastic clothesline	7/32	60 to 70

Halyards

Table 3 shows the recommended maximum tensions for various sizes and types of line and rope suitable for hoisting halyards. Probably the best type for general amateur use for spans up to 150 or 200 feet is ¼-inch nylon rope. Nylon is somewhat more expensive than ordinary rope of the same size, but it weathers much better. Nylon also has a certain amount of elasticity to accommodate gusts of wind, and is particularly recommended for antennas using trees as supports. A disadvantage of new nylon rope is that it stretches by a significant percentage. After an installation with new rope, it will be necessary to repeatedly take up the slack created by stretching. This process will continue over a period of several weeks, at which time most of the stretching will have taken place. Even a year after installation, however, some slack may still arise from stretching.

Most types of synthetic rope are slippery, and some types of knots ordinarily used for rope will not hold well. **Fig 7** shows a knot that should hold well, even in nylon rope or plastic line.

For exceptionally long spans, stranded galvanized steel sash cord makes a suitable support. Cable advertised as "wire rope" usually does not weather well. A boat winch, sold at marinas and at Sears, is a great convenience in antenna hoisting (and usually a necessity with metal halyards).

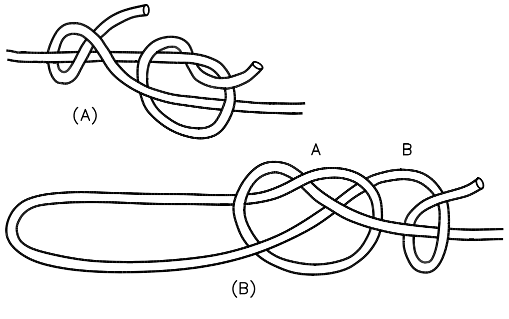

Fig 7—This is one type of knot that will hold with smooth rope, such as nylon. Shown at A, the knot for splicing two ends. B shows the use of a similar knot in forming a loop, as might be needed for attaching an insulator to a halyard. Knot A is first formed loosely 10 or 12 in. from the end of the rope; then the end is passed through the eye of the insulator and knot A. Knot B is then formed and both knots pulled tight. (Richard Carruthers, K7HDB)

Antennas of Aluminum Tubing

Aluminum is a malleable, ductile metal with a mass density of 2.70 grams per cubic centimeter. The density of aluminum is approximately 35% that of iron and 30% that of copper. Aluminum can be polished to a high brightness, and it will retain this polish in dry air. In the presence of moisture, aluminum forms an oxide coating (Al_2O_3) that protects the metal from further corrosion. Direct contact with certain metals, however (especially ferrous metals such as iron or steel), in an outdoor environment can bring about galvanic corrosion of aluminum and its alloys. Some protective coating should be applied to any point of contact between two dissimilar metals. Much of this information about aluminum and aluminum tubing was prepared by Ralph Shaw, K5CAV.

Aluminum is non-toxic; it is used in cooking utensils and to hold and cover "TV dinners" and other frozen foods, so it is certainly safe to work with. The ease with which it can be drilled or sawed makes it a pleasure to work with. Aluminum products lend themselves to many and varied applications.

Aluminum alloys can be used to build amateur antennas, as well as for towers and supports. Light weight and high conductivity make aluminum ideal for these applications. Alloying lowers the conductivity ratings, but the tensile strength can be increased by alloying aluminum with one or more metals such as manganese, silicon, copper, magnesium or zinc. Cold rolling can be employed to further increase the strength.

A four-digit system is used to identify aluminum alloys, such as 6061. Aluminum alloys starting with a 6 contain di-magnesium silicide (Mg_2Si). The second digit indicates modifications of the original alloy or impurity limits. The last two digits designate different aluminum alloys within the category indicated by the first digit.

In the 6000 series, the 6061 and 6063 alloys are a commonly used for antenna applications. Both types have good resistance to corrosion and medium strength. A further designation like T-6 denotes thermal treatment (heat tempering). More information on the available aluminum alloys can be found in **Table 4**.

Table 4

Aluminum Numbers for Amateur Use

	Common Alloy Numbers
Type	*Characteristic*
2024	Good formability, high strength
5052	Excellent surface finish, excellent corrosion resistance, normally not heat treatable for high strength
6061	Good machinability, good weldability
6063	Good machinability, good weldability
7075	Good formability, high strength
	Common Tempers
Type	*Characteristics*
T0	Special soft condition
T3	Hard
T6	Hardest, possibly brittle
TXXX	Three digit tempers—usually specialized high strength heat treatments, similar to T6
	General Uses
Type	*Uses*
2024-T3	Chassis boxes, antennas, anything that will be bent or
7075-T3	Flexed repeatedly
6061-T6	Tubing and pipe; angle channel and bar stock
6063-T832	Tubing and pipe; angle channel and bar stock

SELECTING ALUMINUM TUBING

Table 5 shows the standard sizes of aluminum tubing that are stocked by most aluminum suppliers or distributors in the United States and Canada. Note that all tubing comes in 12-foot lengths (local hardware stores sometimes stock 6- and 8-foot lengths) and larger-diameter sizes may be available in lengths up to 24 feet. Note also that any diameter tubing will fit snugly into the next larger size, if the larger size has a 0.058-inch wall thickness. For example, $^5/_8$-inch tubing has an outside diameter of 0.625 inch. This will fit into $^3/_4$-inch tubing with a 0.058-inch wall, which has an inside diameter of 0.634 inch. A clearance of 0.009 inch is just right for a slip fit or for slotting the tubing and then using hose clamps. Always get the next larger size and specify a 0.058-inch wall to obtain the 0.009-inch clearance.

A little figuring with Table 5 will give you all the information you need to build a beam, including what the antenna will weigh. The 6061-T6 type of aluminum has a relatively high strength and has good workability. It is highly resistant to corrosion and will bend without taking a "set."

SOURCES FOR ALUMINUM

Aluminum can be purchased new, and suppliers are listed in Chapter 21. But don't overlook the local metal scrap yard. The price varies, but between 35 and 60 cents per pound is typical for scrap aluminum. Some aluminum items to look for include aluminum vaulting poles, tent poles, tubing and fittings from scrapped citizen's band antennas, and aluminum angle stock. The scrap yard may even have a section or two of triangular aluminum tower.

Aluminum vaulting poles are 12 or 14 feet long and range in diameter from $1^1/_2$ to $1^3/_4$ inches. These poles are suitable for the center-element sections of large 14-MHz beams or as booms for smaller antennas. Tent poles range in length from 2½ to 4 feet. The tent poles are

Table 5
Aluminum Tubing Sizes

6061-T6 (61S-T6) Round Aluminum Tube In 12-Foot Lengths

	Wall Thickness			Approximate Weight	
Tubing Diameter	*Inches*	*Stubs Ga.*	*ID, Inches*	*Pounds Per Foot*	*Pounds Per Length*
3/16 in.	0.035	(#20)	0.117	0.019	0.228
	0.049	(#18)	0.089	0.025	0.330
1/4 in.	0.035	(#20)	0.180	0.027	0.324
	0.049	(#18)	0.152	0.036	0.432
	0.058	(#17)	0.134	0.041	0.492
5/16 in.	0.035	(#20)	0.242	0.036	0.432
	0.049	(#18)	0.214	0.047	0.564
	0.058	(#17)	0.196	0.055	0.660
3/8 in.	0.035	(#20)	0.305	0.043	0.516
	0.049	(#18)	0.277	0.060	0.720
	0.058	(#17)	0.259	0.068	0.816
	0.065	(#16)	0.245	0.074	0.888
7/16 in.	0.035	(#20)	0.367	0.051	0.612
	0.049	(#18)	0.339	0.070	0.840
	0.065	(#16)	0.307	0.089	1.068
1/2 in.	0.028	(#22)	0.444	0.049	0.588
	0.035	(#20)	0.430	0.059	0.708
	0.049	(#18)	0.402	0.082	0.984
	0.058	(#17)	0.384	0.095	1.040
	0.065	(#16)	0.370	0.107	1.284
5/8 in.	0.028	(#22)	0.569	0.061	0.732
	0.035	(#20)	0.555	0.075	0.900
	0.049	(#18)	0.527	0.106	1.272
	0.058	(#17)	0.509	0.121	1.452
	0.065	(#16)	0.495	0.137	1.644
3/4 in.	0.035	(#20)	0.680	0.091	1.092
	0.049	(#18)	0.652	0.125	1.500
	0.058	(#17)	0.634	0.148	1.776
	0.065	(#16)	0.620	0.160	1.920
	0.083	(#14)	0.584	0.204	2.448
7/8 in.	0.035	(#20)	0.805	0.108	1.308
	0.049	(#18)	0.777	0.151	1.810
	0.058	(#17)	0.759	0.175	2.100
	0.065	(#16)	0.745	0.199	2.399
1 in.	0.035	(#20)	0.930	0.123	1.476
	0.049	(#18)	0.902	0.170	2.040
	0.058	(#17)	0.884	0.202	2.424
	0.065	(#16)	0.870	0.220	2.640
	0.083	(#14)	0.834	0.281	3.372
1 1/8 in.	0.035	(#20)	1.055	0.139	1.668
	0.058	(#17)	1.009	0.228	2.736
1 1/4 in.	0.035	(#20)	1.180	0.155	1.860
	0.049	(#18)	1.152	0.210	2.520
	0.058	(#17)	1.134	0.256	3.072
	0.065	(#16)	1.120	0.284	3.408
	0.083	(#14)	1.084	0.357	4.284
1 3/8 in.	0.035	(#20)	1.305	0.173	2.076
	0.058	(#17)	1.259	0.282	3.384
1 1/2 in.	0.035	(#20)	1.430	0.180	2.160
	0.049	(#18)	1.402	0.260	3.120
	0.058	(#17)	1.384	0.309	3.708
	0.065	(#16)	1.370	0.344	4.128
	0.083	(#14)	1.334	0.434	5.208
	*0.125	1/8 in.	1.250	0.630	7.416
	*0.250	1/4 in.	1.000	1.150	14.832
1 5/8 in.	0.035	(#20)	1.555	0.206	2.472
	0.058	(#17)	1.509	0.336	4.032
1¾ in.	0.058	(#17)	1.634	0.363	4.356
	0.083	(#14)	1.584	0.510	6.120
1 7/8 in.	0.058	(#17)	1.759	0.389	4.668
2 in.	0.049	(#18)	1.902	0.350	4.200
	0.065	(#16)	1.870	0.450	5.400
	0.083	(#14)	1.834	0.590	7.080
	*0.125	1/8 in.	1.750	0.870	9.960
	*0.250	1/4 in.	1.500	1.620	19.920
2 1/4 in.	0.049	(#18)	2.152	0.398	4.776
	0.065	(#16)	2.120	0.520	6.240
	0.083	(#14)	2.084	0.660	7.920
2 1/2 in.	0.065	(#16)	2.370	0.587	7.044
	0.083	(#14)	2.334	0.740	8.880
	*0.125	1/8 in.	2.250	1.100	12.720
	*0.250	1/4 in.	2.000	2.080	25.440
3 in.	0.065	(#16)	2.870	0.710	8.520
	*0.125	1/8 in.	2.700	1.330	15.600
	*0.250	¼ in.	2.500	2.540	31.200

*These sizes are extruded. All other sizes are drawn tubes.

usually tapered; they can be split on the larger end and then mated with the smaller end of another pole of the same diameter. A small stainless-steel hose clamp (sometimes also available at scrap yards!) can be used to fasten the poles at this junction. A 14- or 21-MHz element can be constructed from several tent poles in this fashion. If a longer continuous piece of tubing is available, it can be used for the center section to decrease the number of junctions and clamps.

Other aluminum scrap is sometimes available, such as US Army aluminum mast sections designated AB-85/GRA-4 (J&H Smith Mfg). These are 3 foot sections with a 1 5/8 inch diameter. The ends are swaged so they can be assembled one into another. These are ideal for making a portable mast for a 144-MHz beam or for Field Day applications.

CONSTRUCTION WITH ALUMINUM TUBING

Most antennas built for frequencies of 14 MHz and above are made to be rotated. Constructing a rotatable antenna requires materials that are strong, lightweight and easy to obtain. The materials required to build a suitable antenna will vary, depending on many factors. Perhaps the most important factor that determines the type of hardware needed is the weather conditions normally encountered.

High winds usually don't cause as much damage to an antenna as does ice, especially ice along with high winds. Aluminum element and boom sizes should be selected so the various sections of tubing will telescope to provide the necessary total length.

The boom size for a rotatable Yagi or quad should be selected to provide stability to the entire system. The best diameter for the boom depends on several factors; most important are the element weight, number of elements and overall length. Tubing of 1-¼-inch diameter can easily support three-element 28-MHz arrays and perhaps a two-element 21-MHz system. A 2-inch diameter boom will be adequate for larger 28-MHz antennas or for harsh weather conditions, and for antennas up to three elements on 14 MHz or four elements on 21 MHz. It is not recommended that 2-inch diameter booms be made any longer than 24 feet unless additional support is given to reduce both vertical and horizontal bending forces. Suitable rein-forcement for a long 2-inch boom can consist of a truss or a truss and lateral support, as shown in **Fig 8**.

A boom length of 24 feet is about the point where a 3-inch diameter begins to be very worthwhile. This dimension provides a considerable improvement in overall mechanical stability as well as increased clamping surface area for element hardware. Clamping surface area is extremely important if heavy icing is common and rotation of elements around the boom is to be avoided. Pinning an element to the boom with a large bolt helps in this regard. On smaller diameter booms, however, the elements sometimes work loose and tend to elongate the pinning holes in both the element and the boom. After some time the elements shift their positions slightly (sometimes from day to day!) and give a rather ragged appearance to the system, even though this doesn't generally harm the electrical performance.

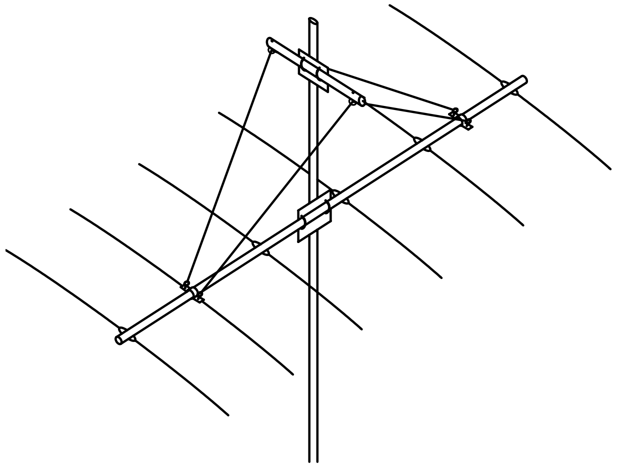

Fig 8—A long boom needs both vertical and horizontal support. The cross bar mounted above the boom can support a double truss to help keep the antenna in position.

A 3-inch diameter boom with a wall thickness of 0.065 inch is satisfactory for antennas up to about a five-element, 14-MHz array that is spaced on a 40-foot long boom. A truss is recommended for any boom longer than 24 feet.

There is no RF voltage at the center of a parasitic element, so no insulation is required in mounting elements that are centered on the boom (driven elements excepted). This is true whether the boom is metal or a nonconducting material. Metal booms have a small "shortening effect" on elements that run through them. With materials sizes commonly employed, this is not more than one percent of the element length, and may not be noticeable in many applications. It is just perceptible with 1/2-inch tubing booms used on 432 MHz, for example. Design-formula lengths can be used as given, if the matching is adjusted in the frequency range one expects to use. The center frequency of an all-metal array will tend to be 0.5 to 1 percent higher than a similar system built of wooden supporting members.

Element Assembly

While the maximum safe length of an antenna element depends to some extent on its diameter, the only laws that specify the minimum diameter of an element are the laws of nature. That is, the element must be rugged enough to survive whatever weather conditions it will encounter.

Fig 9 shows tapered Yagi element designs that will survive winds in excess of 80 mi/h. With a 1/4-inch thickness of radial ice, these designs will withstand winds up to approximately 60 mi/h. (Ice increases the wind area but does not increase the strength of the element.) More rugged designs are shown in **Fig 10**. With no ice loading, these elements will survive in 120-mi/h winds, and in winds exceeding 85 mi/h with 1/4 inch of radial ice. If you lose an antenna made with elements like these, you'll have plenty of company among your neighbors with commercially made antennas!

Figs 9 and 10 show only half elements. When the element is assembled, the largest size tubing for each element should be double the length shown in the drawing, with its center being the point of attachment to the boom. These designs are somewhat conservative, in that they are self-resonant slightly below the frequency indicated for each design. Telescoping the outside end sections to shorter lengths for resonance will increase the survival wind speeds. Conversely, lengthening the outside end sections will reduce the survival wind speeds. [See Bibliography listing for David Leeson (W6NL, ex-W6QHS) at the end of this chapter.]

Fig 11 shows several methods of fastening antenna element sections together. The slot and hose clamp method shown in Fig 11A is probably the best for joints where adjustments are required. Generally, one adjustable joint per element half is sufficient to tune the antenna. Stainless-steel hose clamps (beware—some "stainless

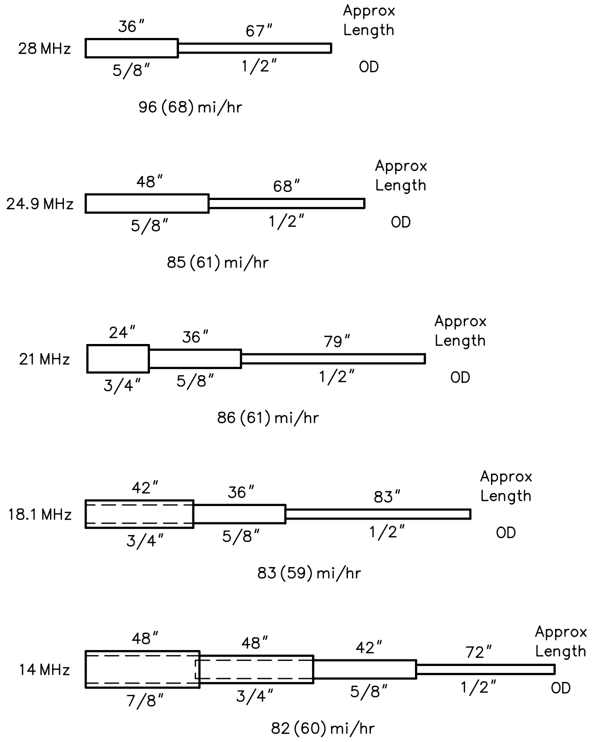
28 MHz
36″
67″
Approx Length
5/8″
1/2″
OD
96 (68) mi/hr
24.9 MHz
48″
68″
Approx Length
5/8″
1/2″
OD
85 (61) mi/hr
21 MHz
24″
36″
79″
Approx Length
3/4″
5/8″
1/2″
OD
86 (61) mi/hr
18.1 MHz
42″
36″
83″
Approx Length
3/4″
5/8″
1/2″
OD
83 (59) mi/hr
14 MHz
48″
48″
42″
72″
Approx Length
7/8″
3/4″
5/8″
1/2″
OD
82 (60) mi/hr

Fig 9—Half-element designs for Yagi antennas. The other side of the element is identical, and the center section should be a single piece twice as long as the length shown here for the largest diameter section. Use 0.058-in.-wall aluminum tubing throughout. Broken lines indicate double tubing thickness, where one tube is inserted into another. The overlap insertion depth into a tube two sizes larger, where shown, should be at least two inches. Maximum survival wind speeds without ice are shown adjacent to each design; values enclosed in parentheses are survival speeds for ¼ inch of radial ice.

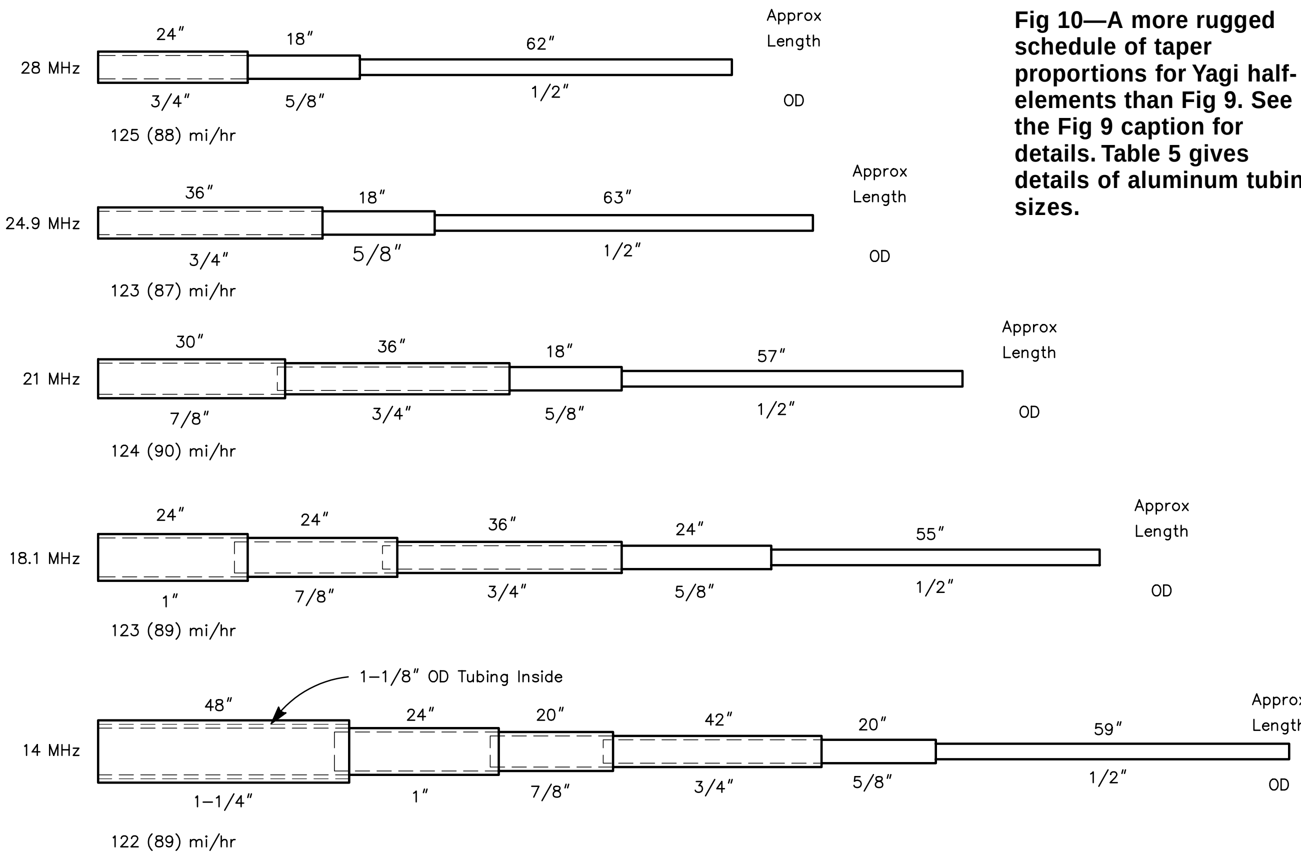
28 MHz
24″
18″
62″
Approx Length
3/4″
5/8″
1/2″
OD
125 (88) mi/hr
24.9 MHz
36″
18″
63″
Approx Length
3/4″
5/8″
1/2″
OD
123 (87) mi/hr
21 MHz
30″
36″
18″
57″
Approx Length
7/8″
3/4″
5/8″
1/2″
OD
124 (90) mi/hr
18.1 MHz
24″
24″
36″
24″
55″
Approx Length
1″
7/8″
3/4″
5/8″
1/2″
OD
123 (89) mi/hr
1–1/8″ OD Tubing Inside
14 MHz
48″
24″
20″
42″
20″
59″
Approx Length
1–1/4″
1″
7/8″
3/4″
5/8″
1/2″
OD
122 (89) mi/hr

Fig 10—A more rugged schedule of taper proportions for Yagi half-elements than Fig 9. See the Fig 9 caption for details. Table 5 gives details of aluminum tubing sizes.

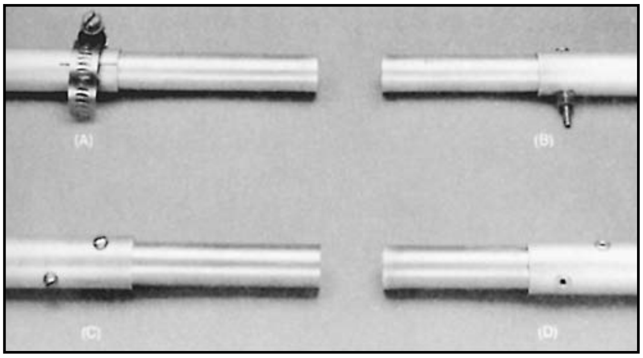

Fig 11—Methods of connecting telescoping tubing sections to build beam elements. See text for a discussion of each method.

Table 6
Hose-Clamp Diameters

	Clamp Diameter (In.)	
Size No.	*Min*	*Max*
06	$^{7}/_{16}$	$^{7}/_{8}$
08	$^{7}/_{16}$	1
10	$^{1}/_{2}$	$1^{1}/_{8}$
12	$^{5}/_{8}$	$1^{1}/_{4}$
16	$^{3}/_{4}$	$1^{1}/_{2}$
20	$^{7}/_{8}$	$1^{3}/_{4}$
24	$1^{1}/_{8}$	2
28	$1^{3}/_{8}$	$2^{1}/_{4}$
32	$1^{5}/_{8}$	$2^{1}/_{2}$
36	$1^{7}/_{8}$	$2^{3}/_{4}$
40	$2^{1}/_{8}$	3
44	$2^{5}/_{16}$	$3^{1}/_{4}$
48	$2^{5}/_{8}$	$3^{1}/_{2}$
52	$2^{7}/_{8}$	$3^{3}/_{4}$
56	$3^{1}/_{8}$	4
64	$3^{1}/_{2}$	$4^{1}/_{2}$
72	4	5
80	$4^{1}/_{2}$	$5^{1}/_{2}$
88	$5^{1}/_{8}$	6
96	$5^{5}/_{8}$	$6^{1}/_{2}$
104	$6^{1}/_{8}$	7

steel” models do not have a stainless screw and will rust) are recommended for longest antenna life. **Table 6** shows available hose-clamp sizes.

Figs 11B, 11C and 11D show possible fastening methods for joints that do not require adjustment. At B, machine screws and nuts hold the elements in place. At C, sheet metal screws are used. At D, rivets secure the tubing. If the antenna is to be assembled permanently, rivets are the best choice. Once in place, they are permanent. They will never work free, regardless of vibration or wind. If aluminum rivets with aluminum mandrels are used, they will never rust. In addition, there is no danger of dissimilar-metal corrosion with aluminum rivets and aluminum antenna elements. If the antenna is to be disassembled and moved periodically, either B or C will work. If machine screws are used, however, take all possible precautions to keep the nuts from vibrating free. Use lock washers, lock nuts and flexible sealant such as silicone bathtub sealant to keep the hardware in place.

Very strong elements can be made by using a double thickness of tubing, made by telescoping one size inside another for the total length. This is usually done at the center of an element where more element strength is desired at at the boom support point, as in the 14-MHz element in Fig 10. Other materials can be used as well, such as wood dowels, fiberglass rods, and so forth.

In each case where a smaller diameter length of tubing is telescoped inside a larger diameter one, it’s a good idea to coat the inside of the joint with Penetrox or a similar substance to ensure a good electrical bond. Antenna elements have a tendency to vibrate when they are mounted on a tower, and one way to dampen the vibrations is by running a piece of clothesline rope through the length of the element. Cap or tape the end of the element to secure the clothesline. If mechanical requirements dictate (a U-bolt going through the center of the element, for instance), the clothesline may be cut into two pieces.

Antennas for 50 MHz need not have elements larger than $^{1}/_{2}$-inch diameter, although up to 1 inch is used occasionally. At 144 and 220 MHz the elements are usually $^{1}/_{8}$ to $^{1}/_{4}$ inch in diameter. For 420 MHz, elements as small as $^{1}/_{16}$ inch diameter work well, if made of stiff rod. Aluminum welding rod of $^{3}/_{32}$ to $^{1}/_{8}$ inch diameter is fine for 420-MHz arrays, and $^{1}/_{8}$ inch or larger is good for the 220-MHz band. Aluminum rod or hard-drawn wire works well at 144 MHz.

Tubing sizes recommended in the paragraph above are usable with most formula dimensions for VHF/UHF antennas. Larger diameters broaden the frequency response; smaller ones sharpen it. Much smaller diameters than those recommended will require longer elements, especially in 50-MHz arrays.

Element Taper and Electrical Length

The builder should be aware of one important aspect of telescoping or tapered elements. When the element diameters are tapered, as shown in Figs 9 and 10, the electrical length is not the same as it would be for a cylindrical element of the same total length. Length corrections for tapered elements are discussed in Chapter 2.

Other Materials for Antenna Construction

Wood is very useful in antenna work. It is available in a great variety of shapes and sizes. Rug poles of wood or bamboo make fine booms. Bamboo is quite satisfactory for spreaders in quad antennas.

Round wood stock (doweling) is found in many hardware stores in sizes suitable for small arrays. Wood is good for the framework of multibay arrays for the higher bands, as it keeps down the amount of metal in the active area of the array. Square or rectangular boom and frame materials can be cut to order in most lumber yards if they are not available from the racks in suitable sizes.

Wood used for antenna construction should be well seasoned and free of knots or damage. Available materials vary, depending on local sources. Your lumber dealer can help you better than anyone else in choosing suitable materials. Joining wood members at right angles can be done with gusset plates, as shown in **Fig 12**. These can be made of thin outdoor-grade plywood or Masonite. Round materials can be handled in ways similar to those used with metal components, with U clamps and with other hardware.

In the early days of Amateur Radio, hardwood was used as insulating material for antennas, such as at the center and ends of dipoles, or for the center insulator of a driven element made of tubing. Wood dowels cut to length were the most common source. To drive out moisture and prevent the subsequent absorption of moisture into the wood, it was treated before use by boiling it in paraffin. Of course today's technology has produced superior materials for insulators in terms of both strength and insulating qualities. However, the technique is worth consideration in an emergency situation or if low cost is a prime requirement. "Baking" the wood in an oven for a short period at 200° F should drive out any moisture. Then treatment as described in the next paragraph should prevent moisture absorption. The use of wood insulators should be avoided at high-voltage points if high power is being used.

All wood used in outdoor installations should be protected from the weather with varnish or paint. A good grade of marine spar varnish or polyurethane varnish will offer protection for years in mild climates, and one or more seasons in harsh climates. Epoxy-based paints also offer good protection.

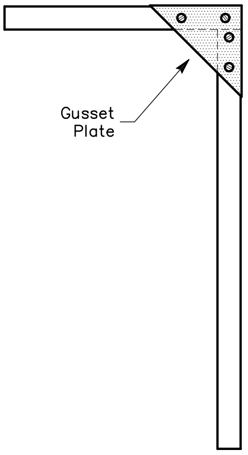

Fig 12—Wood members can be joined at right angles using gusset plates.

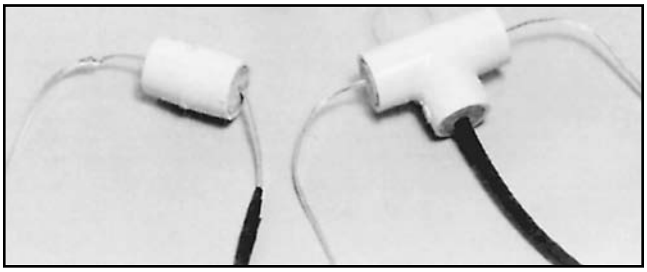

Fig 13—Plastic plumbing parts can be used as antenna center and end insulators.

Plastics

Plastic tubing and rods of various sizes are available from many building-supplies stores. The uses for the available plastic materials are limited only by your imagination. Some amateurs have built beam antennas for VHF using wire elements run inside thin PVC plumbing pipe. The pipe gives the elements a certain amount of physical strength. Other hams have built temporary antennas by wrapping plastic pipe with aluminum foil or other conductive material. Plastic plumbing pipe fittings can also be used to enclose baluns and as the center insulator or end insulators of a dipole, as shown in **Fig 13**. Plastic or Teflon rod can be used as the core of a loading coil for a mobile antenna (**Fig 14**) but the material for this use should be selected carefully. Some plastics become quite warm in the presence of a strong RF field, and the loading-coil core might melt or catch fire!

Fiberglass

Fiberglass poles are the preferred material for spreaders for quad antennas. They are lightweight, they withstand harsh weather well, and their insulating qualities are excellent. One disadvantage of fiberglass poles is that they may be crushed rather easily. Fracturing occurs at the point where the pole is crushed, causing it to lose its strength. A crushed pole is next to worthless. Some amateurs have repaired crushed poles with fiberglass cloth and epoxy, but the original strength is

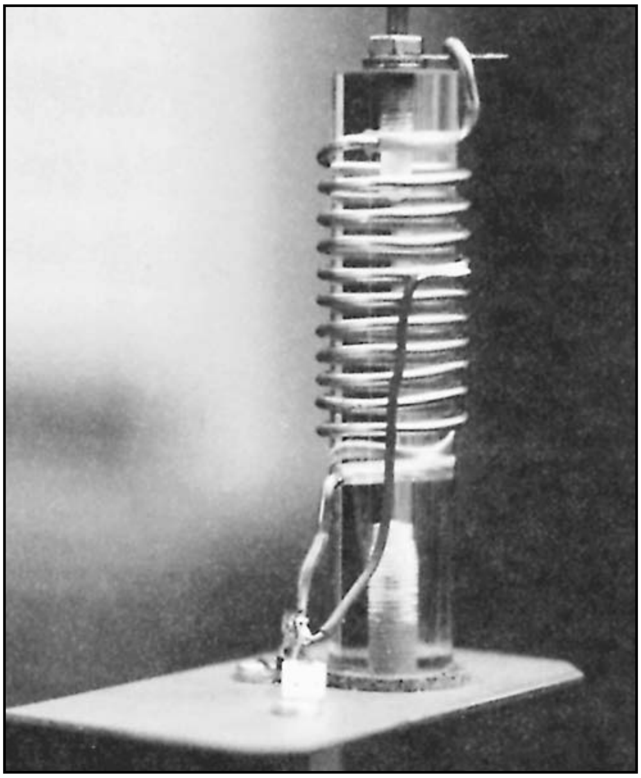

Fig 14—A mobile-antenna loading coil wound on a polystyrene rod.

nearly impossible to regain.

Fiberglass poles can also be used to construct other types of antennas. Examples are helically wound Yagi elements or verticals, where a wire is wound around the pole.

CONCLUSION

The antenna should be put together with good quality hardware. Stainless steel is best for long life. Rust will quickly attack plated steel hardware, making nuts difficult, if not impossible, to remove. If stainless-steel muffler clamps and hose clamps are not available, the next best thing is to have them plated. If you can’t have them plated, at least paint them with a good zinc-chromate primer and a finish coat or two.

Galvanized steel generally has a longer life than plated steel, but this depends on the thickness of the galvanizing coat. Even so, in harsh climates rust will usually develop on galvanized fittings in a few years. For the ultimate in long-term protection, galvanized steel should be further protected with zinc-chromate primer and then paint or enamel before exposing it to the weather.

Good quality hardware is expensive initially, but if you do it right the first time, you won’t have to take the antenna down in a few years and replace the hardware. When the time does come to repair or modify the antenna, nothing is more frustrating than fighting rusty hardware at the top of the tower.

Basically any conductive material can be used as the radiating element of an antenna. Almost any insulating material can be used as an antenna insulator. The materials used for antenna construction are limited mainly by physical considerations (required strength and resistance to outdoor exposure) and by the availability of materials. Don’t be afraid to experiment with radiating materials and insulators.

BIBLIOGRAPHY

Source Material and more extended discussion of topics covered in this chapter can be found in the reference given below.

J. J. Elengo, Jr., “Predicting Sag in Long Wire Antennas,” *QST*, Jan 1966, pp 57-58.

D. B. Leeson, *Physical Design of Yagi Antennas* (Newington, CT: ARRL).

Chapter 21

Antenna Products Suppliers

Antenna Manufacturers Products

Finding parts can be the most difficult aspect of an antenna project. Suppliers of aluminum exist in most major metropolitan areas. They can be found in the Yellow Pages of the phone book. Some careful searching of the Yellow Pages may also reveal sources of other materials and accessories. If you live away from a metropolitan area, try using telephone books for the nearest large metropolitan area; they may be available in the reference section of your local library.

Many dealers and distributors will ship their products by freight or by mail. Tables 1 through 7 list several categories of antenna products and some suppliers of them. Company names have been abbreviated where necessary. Table 8 is an address list arranged alphabetically by company name.

Product lines change often; we recommend that you

Table 1
VHF/UHF/Microwave Antenna Suppliers

Manufacturer	Yagi	Quad	Loop Yagi	Vertical	Mobile	HT	Microwave	Helical	Satellite
Antennaco	✓								
Ant Specialists				✓	✓				
Austin Antenna				✓	✓	✓			
Butternut				✓					
Centurion						✓			
Comet				✓	✓	✓			
Create Design	✓								
Cubex		✓							
Cushcraft	✓		✓	✓	✓				✓
Diamond	✓		✓	✓	✓				
Down East	✓		✓				✓		✓
Eur-AM				✓	✓	✓			
Force 12	✓			✓					
Hustler	✓			✓	✓				
Hy-Gain (MFJ)	✓			✓					✓
Kilo-Tec				✓					
Lakeview				✓	✓	✓			
Larsen				✓	✓				
M^2	✓	✓		✓					
Maldol					✓				
MFJ	✓			✓	✓	✓			
Mosley	✓			✓					
Pro-Am Valor					✓				
RadioShack				✓	✓	✓			
Radio Works					✓				
Sommer	✓								
Spectrum Intl	✓		✓				✓	✓	✓
Tennadyne	LPDA								
Tonna	✓						✓		

request current catalogs from those manufacturers who interest you. In addition, all indications of sales policies and prices for catalogs are given for general information only and are subject to change without notice.

Antenna products for repeaters are listed separately, in Chapter 17.

Table 2
HF Antenna Suppliers

Manufacturer	Yagi	Quad	LPDA	Vertical	Dipole	Mobile	Small Xmtg	Active (RX only)
Alpha Delta				✓	✓	✓		
B & W				✓	✓			
Bilal							✓	
Butternut/Bencher	✓			✓			✓	
Create Design	✓		✓	✓	✓		✓	
Cubex		✓						
Cushcraft	✓		✓	✓	✓			
Flytecraft				✓		✓		
Force 12	✓		✓	✓	✓	✓	✓	
GAP				✓				
Gem Quad		✓						
Grove					✓	✓		✓
High Sierra				✓	✓	✓		
Hustler				✓		✓		
Hy-Gain (MFJ)	✓			✓	✓			
Lakeview/Hamstick						✓		
Lightning Bolt		✓						
M^2	✓				✓			
MFJ	✓	✓	✓	✓	✓	✓	✓	✓
Mosley	✓			✓	✓			
Palomar								✓
Pro-Am Valor						✓		
Roadrunner						✓		
Sommer	✓			✓				
Spi-Ro				✓	✓			
Tennadyne			✓					
Texas Radio					✓	✓		
Radio Works				✓	✓			
Van Gorden				✓	✓			
W1JC					✓		✓	
W9INN					✓			

Table 3
Antenna Parts

Manufacturer	Hardware	Insulators	Traps	Aluminum Tubing	Wire	Transmission Lines
Alexanderson AeroPlane Co				✓		
Barker & Williamson		✓	✓			
Cable X-Perts	✓	✓			✓	✓
Hy-Gain (MFJ)	✓	✓				
Metal & Cable, Inc	✓			✓	✓	
Ocean State Electronics	✓	✓			✓	✓
Spi-Ro		✓	✓		✓	✓
Texas Towers				✓	✓	✓
Radio Works	✓	✓			✓	✓
Van Gorden	✓	✓			✓	✓
The Wireman	✓	✓			✓	✓
W1JC		✓			✓	
W9INN					✓	

Table 4
Suppliers of Quad Antenna Parts

Company	*Material Size*
Cubex Co., Inc.	$1\frac{1}{4}$-$\frac{1}{2}$ in. × 13 ft; $1\frac{1}{2}$ in. × 13 ft cast spiders, boom-to-mast mounts.
Lightning Bolt	Custom-made fiberglass spreaders; any length.
Max-Gain Systems, Inc.	Quad spreaders; fiberglass insulators.

Table 5
Towers, Masts and Accessories

Towers
Aluma
Champion Radio
Create Design
Force 12
Glen Martin Engineering
Heights Tower
Hy-Gain (MFJ)
National Tower
RadioShack (masts only)
Rohn
Rotating Tower Systems
Tashjian Towers Corporation
Texas Tower
Trylon
Universal Manufacturing
US Tower

Climbing and Safety Equipment
Champion Radio Products
ONV
Texas Towers

Rotators
Create Design
Hy-Gain (MFJ)
M^2
RadioShack
The Rotor Doctor
Yaesu

Stacking Frames
(Unless otherwise noted these frames are for use in stacking the manufacturer's own antennas in pairs or in quads. These stacking kits are for VHF or UHF antennas only.)
C3i
Cushcraft
Down East
IIX
Spectrum International

Combiners, Power Dividers and Phasing Harnesses
Byers
C3i
Down East
Spectrum International
Tonna

Table 6
Transmission Lines

Source	Coax	Hardline	Ladder Line
Belden	✓	✓	
Cable X-Perts	✓	✓	✓
International Wire & Cable	✓		
Nemal	✓	✓	
The Radio Works	✓	✓	✓
W1JC			✓
W9INN			✓
The Wireman	✓	✓	✓

Table 7
Transmission Line Instruments and Accessories

Matching Networks
Ameritron
Barker & Williamson
Cubex
ICOM (mobile & fixed)
Kenwood
MFJ
Nye
SGC
Ten-Tec
Texas Radio (mobile only)
Vectronics
XMatch

Ferrite Cores and Rods
Amidon
Palomar
The Wireman

Filters—TVI (Low Pass & High Pass)
K-Com
MFJ

Lightning Arresters
Alpha Delta
Ameritron
Comet
Cushcraft
Hy-Gain (MFJ)
Polyphaser
Radioware
Rohn
The Wireman
Zero Surge Inc

Switches (Manual, Coax)
Alpha Delta
Barker & Williamson
MFJ

Switches (Remote, Coax)
Ameritron
MFJ

SWR and Wattmeters
Autek
Bird
Coaxial Dynamics
MFJ
Nye
Palomar
RF Parts
Texas Radio

Table 8
Suppliers Addresses

We have made every effort to ensure that this list is complete and accurate as of mid 2003. The ARRL takes no responsibility for errors or omissions. Similarly, a listing here does not represent an endorsement of a manufacturer or products by the ARRL. Refer to the product reviews in *QST* for descriptions of particular products that interest you. To the best of our knowledge the suppliers listed are willing to sell products to amateurs by mail unless indicated otherwise. This listing will be updated with each edition of *The Antenna Book* and *The ARRL Handbook* in the TISFIND manufacturer database. Check ads in *QST* and other Amateur Radio publications for any changes to this information. Suppliers who wish to be listed or update their information are urged to contact the editors.

Advanced Composites
1154 S. 300 W.
Salt Lake City, UT 84101
801-467-1204
fax 801-467-4367
e-mail
info@advancedcomposites.com

Advanced Specialties
114 Essex Street
Lodi, NJ 07644
USA
Tel: 800-926-9426
Email: advanspec@aol.com
http://advancedspec.freeyellow.com/

AEA—Div. Tempo Research Corporation
1221 Liberty Way
Vista, CA 92083
760-598-9677
fax 760-598-4898
e-mail tempo@inetworld.com
web www.aea-wireless.com

AFT
132 Boulevard Dauphinot
F-51100 Reims, France
+33 326 070 047
fax +33 326 023 654
e-mail Antennes-ft@wanadoo.fr
web www.f9ft.com

Alexander Aeroplane Co
PO Box 909
Griffin, GA 30224
800-831-2949; 404-229-2329

Alan Broadband Company
(Zap Checker)
93 Arch Street
Redwood City, CA 94062
USA
Tel: 888-369-9627 (orders)
650-369-9627
Fax: 650-369-3788
Email: ABCom@prodigy.net
http://www.alanbroadband.com/

Alpha Delta Communications
PO Box 620
Manchester, KY 40962
606-598-2029
fax 606-598-4413

Aluma Tower Co, Inc
PO Box 2806-AL
Vero Beach, FL 32961-2806
772-567-3423
fax 772-567-3432
e-mail atc-t@alumatower.com
web www.alumatower.com/index.html

Amateur Electronic Supply
5710 W Good Hope Rd
Milwaukee, WI 53223
800-558-0411
web www.aesham.com/

Ameritron Division
116 Willow Road
Starkville, MS 39759
662-323-8211 (Tech)
fax 662-323-6551
e-mail
ameritron@ameritron.com
web www.ameritron.com/ameritron

Amidon Associates, Inc.
1510 E. Edinger Avenue, Unit B
Santa Ana CA 92705
800-898-1883; 714-547-4494
fax 714-547-4433
e-mail sales@amidon-inductive.com
web www.amidon-inductive.com/

Antennaco Inc
102 Armory Road
PO Box 218
Milford, NH 03055-0218
603-673-3153
fax 603-673-4347

Allen Telecom Group
Antenna Specialists Mobile Division
30500 Bruce Industrial Parkway
Cleveland, OH 44139-3996
USA
440-349-8400
fax 440-349-8404
web www.antenna.com/

Array Solutions
350 Gloria Rd
Sunnyvale, Tx 75182
USA
Tel: 972-203-8810
Fax: 972-203-8811
Email: wx0b@arrysolutions.com
http://www.arraysolutions.com

ASA Antenna Sales
PO Box 3461
Myrtle Beach, SC 29578
800-772-2681

Associated Radio Communications
8012 Conser
Overland Park, KS 66204
913-381-5900; 800-497-1457
fax 913-648-3020
e-mail assocrad@tfs.net
web www.associatedradio.com

Austin Amateur Radio Supply
5310 Cammeron Road
Austin, TX 78723
800-423-2604; 512-454-2994
fax 512-454-3069

Austin Antenna, Ltd
10 Main St
Gonic, NH 03839
603-336-6339
fax 603-335-1756

Autek Research
PO Box 7556
Wesley Chapel, FL 33544
813-994-2199
e-mail
Mailbox@autekresearch.com
web www.autekresearch.com/
index.htm

Barker & Williamson Co
603 Cidco Rd
Cocoa, FL 32926
321-639-2545
fax 321-639-2545
e-mail
custsrvc@bwantennas.com
web www.bwantennas.com

Barry Electronics Corp
540 Broadway
New York, NY 10012
212-925-7000
fax 212-925-7001

Belden Wire & Cable
PO Box 1980
Richmond, IN 47374
317-983-5257
fax 317-983-5257
web www.belden.com

Bencher Inc.
831 North Central Avenue
Wood Dale, IL 60191
630-238-1183
fax 630-238-1186
e-mail bencher@bencher.com
web www.bencher.com/

Bilal Company
137 Manchester Dr
Florissant, CO 81816
719-687-0650
web www.isotronantennas.com/

Bird Electronics Corporation
30303 Aurora Rd.
Solon, OH 44139
866-695-4569
e-mail sales@bird-
technologies.com
web www.bird-electronic.com/

Brian Beezley, K6STI
3532 Linda Vista Dr
San Marcos, CA 92069
619-599-4962
e-mail k6sti@n2.net

Burghardt Amateur Center, Inc.
710 10th Street SW
PO Box 73
Watertown, SD 57201
800-927-4261
605-886-7314 (Service)
605-886-6914 (Fax-back
product info)
fax 605-886-3444
e-mail burghart@daknet.com
web www.burghardt-
amateur.com/

Butternut Electronics Co
See Bencher Inc.

Byers Chassis Kits
5120 Harmony Grove Road
Dover, PA 17315
717-292-4901
(6p-9p EST M-F, 8a-4p Sat)
717-292-4901 (24 hrs)

Cable X-Perts Inc.
225 Larkin Drive Suite #6
Wheeling, IL 60090-7209
800-828-3340 (orders only)
847-520-3003 (Tech. Info.)
fax 847-520-3444
e-mail cxp@cablexperts.com
web www.cablexperts.com/

CAL-AV Labs, Inc.
1802 w. Grant road, Ste. 116
Tucson, AZ 85745
USA
Tel: 520-624-1300
888-815-0400 (orders only)
Fax: 520-624-1311
Email: info@cal-av.com
http://www.cal-av.com/

Cardwell Condenser Corp
80 East Montauk Hwy
Lindenhurst, NY 11757
516-957-7200
fax 516-957-7203
web www.cardwellcondenser.com

C.A.T.S.
Formerly known as: Rotor Doctor
7368 SR 105
Pemberville, OH 43450
419-353-2287
fax 419-354-7746
e-mail mailto:craig@rotor-doc.com
web www.rotordoc.com

C-Comm
6115 15th NW
Seattle, WA 98107
206-784-7337; 800-426-6528
fax 206-784-0541

C3i
1401 K Street NW Suite 900
Washington DC 20005
(800) 224-5137
fax (202) 362-8481
e-mail info@C3iusa.com
web www.c3iusa.com/
corpover.html

Centurion International
PO Box 82846
Lincoln, NE 68501-2846
402-467-4491
fax 800-848-3825

Champion Radio Products
Box 572
Woodinville, WA 98072
425-485-7913
206-890-4188 Cell Phone
fax 360-668-1447
e-mail championradio@aol.com
web www.championradio.com/

Coaxial Dynamics
15210 Industrial Parkway
Cleveland, Ohio 44135-3308
800-262-9425
fax 216-267-3142
e-mail cdi-sales@coaxial.com
web www.coaxial.com/
home_frames.htm

Comet North America Inc.
394 Wards Corner Road
Suite 130
Cincinnati, Ohio, 45140
513-831-5000
fax 513-831-7889
e-mail
info@CometNorthAmerica.com
web www.cometnorthamerica.com/

Comet Antenna
See NCG

Comm-Pute
7946 State Street
Midvale, UT 84047
801-567-9944
fax 801-567-9494
e-mail bobwood@xmission.com
web www.comm-pute.com/

Communication Headquarters Inc.
3832 Oleander Drive
Wilmington, NC 28403
910-791-8885
800-688-0073 (orders)
fax 910-452-3891 (orders)
web www.chq-inc.com/

Communications Data Corp
1051 Main St
St Joseph, MI 49085
269-982-0404
fax 269-982-0433
e-mail mailto:did@gtm.net

Comtelco Industries Inc
501 Mitchell Rd
Glendale Heights, IL 60139
800-634-4622; 708-7790-9894
fax 708-798-9799

Create
4-8 Asano-Cho
Kawasaki-Ku
Kawasaki-City, Japan
044 (333) 6681
fax 044 (333) 6598
e-mail email@cd-corp.com

Cubex Co.
228 Hibiscus St, #9
Jupiter, FL 33458
561-748-2830
fax 561-748-2831
e-mail CubexCo@aol.com
web www.cubex.com/

Cushcraft Corp
48 Perimeter Rd.
Manchester, NH 03108
603-627-7877
e-mail sales@cushcraft.com
web www.cushcraft.com/

Dave's Hobby Shop
600 Main St.
Van Buren, AK 72956
USA
Tel: 479-471-0750
Web www.daveswebshop.com

Davis RF Co
See Radioware (distributor)
PO Box 730
Carlisle, MA 01741
978-371-1356
978-369-1738
800-328-4773 (Orders only)
fax 978-369-3484
e-mail davisrfinc@aol.com
web www.davisRF.com

Diamond Antennas
See RF Parts

Down East Microwave
954 Rt 519
Frenchtown, NJ 08825
908-996-3584
fax 908-996-3702
web downeastmicrowave.com/

Dressler Hochfrequenztechnik
GMBH
Werther Strasse 14-16
W-5190 Stolberg
Germany

DX Engineering
POB 440
Peninsula, Ohio 44264
USA
Fax: 330-657-2168
Email:
dxengineering@dxengineering.com
http://www.dxengineering.com/

EDCO - Electronic Distributors
Company
325 Mill Street
Vienna, VA 22180
703-938-8105
fax 703-938-6911
e-mail web www.elecdist.com/
index.html

EUR-AM Antennas
PO Box 225
Moultonboro, NH 03254-0225
603-476-5113
fax 603-476-5113
e-mail info@eur-am.com
web www.eur-am.com/

EZ Hang, Inc.
8645 Tower Dr.
Code C
Laurel, MD 20723
Tel: 540-286-0176
Fax: 202-260-3797
http://www.ezhang.com

Fluidmotion Incorporated
14135 233rd Place SE
Issaquah, WA 98027
425-456-0200; 800-885-8700
fax 425-391-6031
e-mail sales@steppir.com
web www.steppir.com/

Flytecraft
PO Box 3141
Simi Valley, CA 93093
805-583-8173

Force 12
PO Box 1349
Paso Robles, CA 93447
800-248-1985
805-227-1680 (Tech)
fax 805-227-1684
e-mail
mailto:force12e@lightlink.com
web force12inc.com/

GAP Antenna Products
99 North Willow St
Fellsmere, FL 32948
772-571-9922
Email: contact@gapantenna.com
web www.gapantenna.com

Gem Quad Products Ltd
PO Box 291
Boissevain, MB R0K 0E0
Canada
204-534-6184
fax 204-534-6492
e-mail
mailto:gemquad@escape.ca
web www.escape.ca/~gemquad/

GLA Systems
PO Box 425
Caddo Mills, TX 75135
903-527-4163
800-588-2841
fax 214-381-2895
web www.texasbugcatcher.com/

Grove Enterprises Inc.
PO Box 98
Brasstown, NC 28902
800-438-8155 (orders)
704-837-9200 (BBS)
fax 704-837-2216
e-mail nada@grove.net
web www.grove-ent.com/

Ham Radio Outlet
1702 W. Camelback Rd
Phoenix, AZ 85015
web www.hamradio.com/
800-444-4799 Mid Atlantic
800-444-9476 Mountain
800-444-0047 New England
800-644-4476 Northeast
800-444-7927 Southeast
800-854-6046 West

The Ham Station
220 North Fulton Avenue
PO Box 6522
Evansville, IN 47719-0522
800-729-4373 (orders)
812-422-0231 (Tech)
812-422-0252 (Service)
fax 812-422-4253
e-mail sales@hamstation.com
web www.hamstation.com/

Hamtronics, Inc
65-Q Moul Rd
Hilton, NY 14468
716-392-9430
fax 716-392-9420
web www.hamtronics.com

Hamware.de
Int'l: Klaus Bemmerer
Tel: +49 4371 869145
e-mail service@hamware.de
Web www.hamware.de
US: Dillon RF Systems
Email dillonel@mtaonline.net

Heights Tower Systems
1529 Gulf Beach Hwy
Pensacola, FL 32507
850-455-1210
fax 850-455-4355
e-mail info@heightstowers.com
web www.heightstowers.com/

Hi-Q Antennas
21085 Cielo Vista Way
Wildomar, CA 92595
909-674-4862
fax 909-245-2031
e-mail sales@hiqantennas.com
web www.hiqantennas.com/

High Sierra Antennas
Box 2389
Nevada City, CA 95959
888-273-3415; 530-273-3415
fax 530-273-7561
e-mail heath@hsantennas.com
web www.hsantennas.com

Hustler Antennas
see New-Tronics Antenna Corp.

Hy-Gain (also see MFJ)
308 Industrial Park Road
Starkville, MS 39759
800-647-1800 (Sales)
662-323-9358 (Tech)
fax 662-323-6551
web www.hy-gain.com/

ICOM America, Inc.
2380 116th Ave, NE, Suite S
Bellevue, WA 98004
USA
Tel: 800 872-4266
Fax: 425 454-1509
Web www.icomamerica.com

Industrial Communications
Engineers (ICE)
Array Solutions
350 Gloria Rd.
Sunnyvale, TX 75182
972-203-2008; 800-423-2666
fax 972-203-8811
e-mail wx0b@arraysolutions.com
web www.iceradioproducts.com/

IIX Equipment, Ltd
PO Box 9
Oak Lawn, IL 60454
708-423-0605
fax 708-423-1691
web www.w9iix.com/

International Radio
13620 Tyee Road
Umpqua, OR 97486
541-459-5623 9AM-1PM PDT,
Tues-Sat
fax 541-459-5632
e-mail inrad@rosenet.net
web www.qth.com/inrad/

Kenwood USA Corporation
3975 Johns Creek Court,
Suite 300
Suwanee, GA 30024
USA
Tel: 678 474-4700
Fax: 678 474-4730
Web www.kenwoodusa.com

K0XG Systems
1117 Highland Park Drive
Bettendorf, IA 52722
USA
Tel: 563 355-7451
Email K0XG@ K0XG.com
Web www.k0xg.com

Kilo-Tec
PO Box 10
Oakview, CA 93022
805-646-9645

K-Com
PO Box 82
Randolph OH 44265
330-325-2110
fax 330-325-2525
e-mail k-com@worldnet.att.net
web www.k-comfilters.com/
howtoget.asp

Lakeview Co Inc
3620-9A Whitehall Rd
Anderson, SC 29626
(864) 226-6990
fax (864) 225-4565
web www.hamstick.com

Larsen Electronics, Inc
See Radial/Larsen

LDG Electronics
1445 Parran Road
St Leonard, MD 20685
877-890-3003 Toll-Free Orders
410-586-2177 Tech Support
fax 410-586-8475
e-mail ldg@ldgelectronics.com
web www.ldgelectronics.com

Lentini Communications
21 Garfield St
Newington, CT 06111
860-666-6227
800-666-0908 (outside CT)
fax 860-667-3561
e-mail radio@lentinicomm.com
web www.lentinicomm.com

Roy Lewallen, W7EL
PO Box 6658
Beaverton, OR 97007
503-646-2885
fax 503-671-9046
e-mail w7el@eznec.com
web www.eznec.com

Lightning Bolt Antennas
RD #2, RT 19
Volant, PA 16156
724-530-7396
fax 724-530-6796
e-mail lbaquads@ztrain.com
web
www.lightningboltantennas.com/

M^2 Antenna Systems Inc.
4202 N. Selland
Fresno CA, 93722
559-432-8873
fax 559-432-3059
e-mail wyatt@m2inc.com
web www.m2inc.com/

M/A-COM, Inc
(an AMP Company)
1011 Pawtucket Blvd
PO Box 3295
Lowell, MA 01853-3295
508-442-4500
fax 508-442-4436
e-mail sales@macom.com
web www.amp.com

Maldol USA
4711 NE 50th St
Seattle, WA 98105
Answerfax 206-525-1896
fax 206-524-7826
e-mail transtec@cyberquest.com

Glen Martin Engineering
13620 Old Highway 40
Boonville, MO 65233
660-882-2734; 800-486-1223
fax 660-882-7200
web www.glenmartin.com

The Mast Company
Henry Pollock - K4TMC
P.O. Box 1932
Raleigh, NC 27602
Email: k4tmc@aol.com
http://www.geocities.com/tmastco/

MCM Electronics
650 Congress Park Dr
Centerville, OH 45459-4072
800-543-4330
fax 800-765-6960
web www.i-mcm.com

Memphis Amateur Electronics
1465 Wells Station Rd
Memphis, TN 38108
901-683-9125; 800-238-6168
fax 901-682-7165

Metal and Cable Corp, Inc
9241 Ravenna Road, Unit C-10
PO Box 117
Twinsburg, OH 44087
330-425-8455
fax 330-425-3504
web www.metal-cable.com/

MFJ Enterprises
300 Industrial Park Rd
Starkville, MS 39759
662-323-5869
Email
mfjcustserv@mfjenterprises.com
web www.mfjenterprises.com/

MGS (Max-Gain Systems, Inc.)
221 Greencrest Court
Marietta, GA 30068-3825
770-973-6251 (before 9 PM
eastern)
fax 815) 461-7730
e-mail info@mgs4u.com
web www.mgs4u.com/

Michigan Radio
26014 Groesbeck Hwy
Warren, MI 48089
810-771-4711
800-TRU-HAMM (800-878-4266)
Orders only
810-771-4712 (Service)
fax 810-771-6546
web www.michiganradio.com/

Mirage Communications
116 Willow Road
Starkville, MS 39759
662-323-8287
fax 662-323-6551
web www.mirageamp.com/

James Millen Electronics
PO Box 4215BV
Andover, MA 01810-4215
978-975-2711
fax 978-474-8949
e-mail info@jamesmillenco.com
web www.jamesmillenco.com/
millen/millmain.htm

Mosley Electronics
1325 Style Master Drive
Union, MO 63084
800-325-4016 (Orders)
800-9MOSLEY (966-7539)
314-994-7872 (Technical)
fax 314-994-7873
web www.mosley-electronics.com/

Multi-band Antennas (originally
Spider Antennas)
7131 Ownesmouth Avenue -
Suite 563C
Canoga Park, CA 91303
818-341-5460
web www.spiderantenna.com/

Multi-tech Industries, Inc. (MTI)
(formally Radio Switch)
64 South Main Street
PO Box 159
Marlboro, NJ 07746-0159
800-431-3223; 732-431-0550
732-462-6100
fax 732-409-6695
e-mail multitech@sprynet.com
web www.multi-tech-industries.com

National RF, Inc.
7969 Engineer Road, Suite 102
San Diego, CA 92111
858-565-1319
fax 858-571-5909
web www.nationalrf.com/

National Tower Co
PO Box 15417
Shawnee Mission, KS 66285
800-762-5049; 913-888-8864

NCG Companies—Comet Antenna
1275 North Grove St
Anaheim, CA 92806
714-630-4541; 800-962-2611
fax 714-630-7024
web www.cometantenna.com
web www.natcommgroup.com/

Nemal Electronics Inc.
12240 NE 14th Avenue
North Miami, FL 33161
305-893-3924; 305-849-0900
800-522-2253
fax 305-895-8178
e-mail info@nemal.com
web www.nemal.com/

New-Tronics Antenna Corp. - Hustler
1 Newtronics Place
Mineral Wells, TX 76067-9563
877-994-9499 Sales
940-325-1386 Service
fax 940-328-1409
e-mail sales@new-tronics.com
web www.new-tronics.com/main/index.html

Nittany Scientific
1733 West 12600 South
Suite 420
Riverton, UT 84065
Phone/fax (801) 446-1426
e-mail sales@nittany-scientific.com
web www.nittany-scientific.com/

William M. Nye Co
PO Box 1877
Priest River, ID 83856
208-448-1762
fax 208-448-1832

Ocean State Electronics
PO Box 1458
6 Industrial Dr
Westerly, RI 02891
401-596-3080; 800-866-6626
fax 401-596-3590
web www.oselectronics.com

ONV Safety Belt Co.
Box 404
Ramsey, NJ 07446
800-345-5634; 201-327-2462
fax 201-327-2462

OptiBeam
Thomas Schmenger DF2BO
Rastatter Straße 37
D-75179 Pforzheim
Tel./Fax: (0049) 07231 / 45 31 53
Email: Info@optibeam.de

OptiBeam NorthAmerica
Array Solutions
350 Gloria Rd
Sunnyvale, Tx 75182 USA
Tel: 972-203-8810
Fax: 972-203-8811
Email: wx0b@arrysolutions.com
http://www.arraysolutions.com

Orion now sold by M^2

Outbacker Antenna Sales
330 Cedar Glen Circle
Chattanooga, TN 37412
615-899-3390
fax 615-899-6536

Palomar Engineers
PO Box 462222
Escondido, CA 92046
619-747-3343
fax 619-747-3346
e-mail palomar@compuserve.com
web www.palomar-engineers.com/

Palstar, Inc.
9676 N Looney Rd
P O Box 1136-1482
Piqua, Ohio 45356
USA
Tel: 937-773-6255
Fax: 937-773-8003
Email: Palstar@erinet.com
http://www.palstarinc.com/

Phillystran, Inc
151 Commerce Dr
Montgomeryville, PA 18936
215-368-6611
fax 215-362-7956
web www.phillystran.com/

PolyPhaser Corp
2225 Park Place
PO Box 9000
Minden, NV 89423-9000
702-782-2511
fax 702-782-4476
e-mail info@polyphaser.com
web www.polyphaser.com

Quicksilver Radio Products
30 Tremont St.
Meriden, CT 06450
USA
Tel: 203-440-4468
http://www.qsradio.com

Radiall/Larsen Antenna Technologies
3611 NE 112th Avenue
Vancouver, WA 98682
800-ANTENNA (800-268-3662)
360-944-7551
FAX 360-944-7556
fax 800-525-6749
e-mail info@radialllarsen.com
web www.radialllarsen.com/default.htm

Radio Bookstore (Radioware)
PO Box 209
Rindge, NH 03461-0209
603-899-6957
800-457-7373 (10am-6pm EST)
fax 603-899-6826
e-mail radware@radio-ware.com
web www.radiobooks.com/

Radio Center USA
1242 Howell
N. Kansas City, MO 64116
816-459-8832; 800-821-7323

Radio City
2663 County Rd I
Mounds View, MN 55112
612-786-4475; 800-426-2891
fax 612-786-6513
web www.radioinc.com

Radio Engineers/Technitron (see National RF, Inc.)

Radio Shack
(Contact your local store)

Radio Switch Corp
See Multi-tech Industries

Radio Works
PO Box 6159
Portsmouth, VA 23703
757-484-0140
800-280-8327 (Orders)
fax 757-483-1873
e-mail jim@radioworks.com
web radioworks.com/

Raibeam Antennas, Int'l
5638 West Alice Avenue
Glendale, AZ 85302
800-530-1913 (orders)
602-931-9135
e-mail wa7rai@raibeam.com
web www.raibeam.com/

RF Parts Co
435 South Pacific St
San Marcos, CA 92069
760-744-0700; 800-737-2787
760-744-0750 (Tech)
fax 760-744-1943; 888-744-1943
e-mail rfp@rfparts.com
web www.rfparts.com

RF TEC Manufacturing
256 Commerce Rd, Suite 517
Peachtree, GA 30269
770-251-2235
fax 770-502-9827
web www.rftec.com/

Roadrunner Resonator
1850 Swanson, #A20
Lake Havasu, AZ 86403
520-453-7211

ROHN Industries, Inc.
PO Box 2000
Peoria, IL 61656
309-697-4400
fax 309-697-5612
e-mail mail@rohnnet.com
web www.rohnnet.com/Index.htm

Ross Distributing Co
78 South State St
Preston, ID 83263
208-852-0830
fax 208-852-0833
web www.rossdist.com

Rotating Tower Sytems, Inc.
Box 44
Prosper, TX 75058
214-347-2560

Rotor Doctor
See C.A.T.S.

SGC Inc
13737 SE 26th St
Bellevue, WA 98005
425-746-6310
fax 425-746-6384
e-mail sgc@sgcworld.com
web www.sgcworld.com/

Shoestring Antennas
PO Box 425
Keyport, WA 98345
360-697-5399
fax 360-697-8416
e-mail jacksa@juno.com
web www.qth.com/shoestring/

Sky-Pole Manufacturing, Inc
1922 Placentia Ave
Costa Mesa, CA 92627
949-548-5596
fax 949-548-5598

Sommer Antennas
PO Box 710
Geneva, FL 32732
407-349-9114
fax 407-349-2485
e-mail sommer1@ix.netcom.com
web www.sommerantennas.com

Spectrum International, Inc
PO Box 1084
Concord, MA 01742
508-263-2145
fax 508-263-7008

Spi-Ro Manufacturing, Inc.
P.O. Box 189
Jonesborough, TN 37659
800-728-7594
423-913-1615 Tech Support
fax 423-913-2131
e-mail mailto:contact@spiromfg.com
web www.spiromfg.com/

Surplus Sales of Nebraska
1502 Jones St
Omaha, NE 68102-3112
402-346-4750
800-244-4567 (Orders only)
fax 402-346-2939
e-mail grinnell@surplussales.com
web www.surplussales.com/

Tarheel Antennas
913 Old Honeycutt Rd
Fuquay-Varina, NC 27256
919-552-8788
fax 919-552-4970
e-mail tarheelantennas@aol.com
web www.tarheelantennas.com/

Tashjian Towers Corporation
(used to be Tri-Ex Towers)
2183 S Highland Ave.
Sanger, CA 93657
559-495-0307
559-284-9707 Mobile
fax 559-495-0557
e-mail tashjian@MSN.com
web www.karltashjian.com/

Telex Communications, Inc
Hy-Gain (see MFJ)

Ten-Tec, Inc
1185 Dolly Parton Parkway
Sevierville, TN 37862
865-453-7172
865-428-0364 (Repairs)
fax 865-428-4483
web www.tentec.com/

Tennadyne Corp
PO Box 34202
San Antonio, TX 78265-4202
210-599-9064
fax 210-599-9064
(same as voice)
e-mail tennadyn@satx.net
web www.tennadyne.com/

Teri Software
PO Box 277
Lincoln, TX 78948
979-542-7952
e-mail
support@antennamodel.com
web www.antennamodel.com

Texas Radio Products
5 E Upshaw
Temple, TX 76501
817-771-1188

Texas Towers
1108 Summit Ave, Suite 4
Plano, TX 75074
800-272-3467
972-422-7306 (Tech)
fax 972-881-0776
e-mail sales@texastowers.com
web www.texastowers.com/

T.G.M. Communications
121 Devon Street
Stratford, Ontario, N5A 2Z8
Canada
519-271-5928
e-mail tgmc@sympatico.ca
web www3.sympatico.ca/tgmc/
index.html

TIC General
PO Box 1 - 1110 Airport Rd
Thief River Falls, MN 56701
218-681-1119
fax 218-681-8509

Timewave Technology, Inc.
1025 Selby Ave, Suite 101
St Paul, MN 55104
USA
Tel: 651-489-5080
Email: sales@timewave.com
http://www.timewave.com

Traffie Technology
421 Jones Hill Road
Ashby, MA 01431-1801
888-599-BEAM (888-599-2326)
978-386-7900
fax 978-386-7900 (same as voice)

Trylon Manufacturing Co, Ltd
21 South Field Dr
PO Box 186
Elmira, Ontario
N3B 2Z6 Canada
519-669-5421
fax 519-669-8912
e-mail info@trylon.com
web www.trylon.com/home.asp

Tucker Electronics and Computers
1717 Reserve St.
PO Box 551419
Garland, TX 75355-1419
800-527-4642 (Test equipment
and catalog requests)
214-348-8800 (x332)
fax 214-340-5460
web www.tucker.com/

Unadilla Antenna Manufacturing
PO Box 4215
Andover, MA 01810-4215
978-975-2711
978-475-7831 (Office)
fax 978-474-8949
e-mail info@unadilla.com
web www.unadilla.com/

Universal Manufacturing Co
43900 Groesbeck Hwy
Clinton Township, MI 48036
810-463-2560
fax 810-463-2964

Universal Radio
6830 Americana Parkway
Reynoldsburg, OH 43068-4113
USA
Tel: 800-431-3939 (Orders)
614-866-4267 (Technical Info)
Fax: 614-866-2339
Email: dx@universal-radio.com
http://www.universal-radio.com/

US Tower Corp
1220 Marcin St
Visalia, CA 93291
559-733-2438
fax 559-733-7194
web www.ustower.com

Valor Enterprises
1711 Commerce Drive
PO Box 601
Piqua, OH 45356-0601
513-778-0074; 800-543-2197
fax 513-778-0259
web www.valorantenna.com/

Van Gorden Engineering
PO Box 21305
South Euclid, OH 44121

Van Valzah Co
38 W 111 Horseshoe Dr
Batavia, IL 60515-9730
708-406-9210

Vectronics
300 Industrial Park Road
Starkville, MS 39759
800-363-2922; 662-323-5800
fax 662-323-6551
e-mail jshurden@vectronics.com
web www.vectronics.com

XMatch
Paul Schrader, N4XM
7001 Briscoe Lane
Louisville, KY 40228
web http://n4xm.myiglou.com/

XX Towers
814 Hurricane Hill Road
Mason, NH 03048
603-878-1102
fax 603-878-4200
e-mail kc1xx@kc1xx.mv.com
web www.xxtowers.com/

W1JC—TV Evans
113 Stratton Brook Rd
Simsbury, CT 06070
860-658-5579
web pages.prodigy.com/w1jc

W3FF Antennas
2390 Templeton Dr.
Redding, CA 96002
USA
Tel: 530-226-8446
Email: sales@buddipole.com
http://www.buddipole.com

W7FG Vintage Manuals
3300 Wayside Drive
Bartlesville, OK 74006
USA
Tel: 800-807-6146 (orders)
918-333-3754
Email: w7fg@eigen.net
http://www.w7fg.com/

W9INN Antennas
PO Box 393
Mt Prospect, IL 60056
847-394-3414

Wacom Products
PO Box 21145
Waco, TX 76702
716-549-4700
e-mail sales@txrx.com

WB0W, Inc
PO Box 8547
Saint Joseph, MO 64508
USA
Tel: 816-364-2691
Email: WB0W@WB0W.com

The Wireman Inc.
261 Pittman Road
Landrum, SC 29356-9544
800-727-WIRE (800-727-9473)
Orders only
864-895-4195 Technical
fax 803-895-5811
e-mail N8UG@earthlink.net
web thewireman.com/

Yaesu U.S.A. (Vertex Standard)
10900 Walker St.
Cypress, CA 90630
714-827-7600
e-mail
amateurtech@vxstdusa.com
web www.yaesu.com/amateur/
amateur.html

YagiStress
Kurt Andress, K7NV
2835 Wade Street
Minden, NV 89423
702-267-5290
e-mail K7NV@contesting.com
web yagistress.freeyellow.com/

Zero Surge Inc
944 State Rt 12
Frenchtown, NJ 08825
908-996-7700
fax 908-996-7773
e-mail zerosurge1@aol.com
web www.zerosurge.com/

Chapter 22

Antenna Supports

A prime consideration in the selection of a support for an antenna is that of structural safety. Building regulations in many localities require that a permit be obtained in advance of the erection of certain structures, often including antenna poles or towers. In general, localities having such requirements also have building safety codes that must be observed. Such regulations may govern the method and materials used in construction of, for example, a self-supporting tower. Checking with your local government building department before putting up a tower may save a good deal of difficulty later, because a tower would have to be taken down or modified if not approved by the building inspector on safety grounds.

Municipalities have the right and duty to enforce any reasonable regulations having to do with the safety of life or property. The courts generally have recognized, however, that municipal authority does not extend to aesthetic questions. The fact that someone may object to the mere presence of a pole, tower or other antenna structure because in his opinion it detracts from the beauty of the neighborhood is not grounds for refusing to issue a permit for a safe structure to be erected. Since the introduction of PRB-1 (federal preemption of unnecessarily restrictive antenna ordinances), this principle has been borne out in many courts. Permission for erecting amateur towers is more easily obtained than in the recent past because of this legislation.

Even where local regulations do not exist or are not enforced, the amateur should be careful to select a location and a type of support that contribute as much safety as possible to the installation. If collapse occurs, the chances of personal injury or property damage should be minimized by careful choice of design and erection methods. A single injury can be far more costly than the price of a more rugged support, in terms of both monetary loss and damage to the public respect for amateur radio.

This chapter has been reviewed and rewritten by Kurt Andress, K7NV.

TREES AS ANTENNA SUPPORTS

From the beginning of Amateur Radio, trees have been used widely for supporting wire antennas. Trees cost noth-

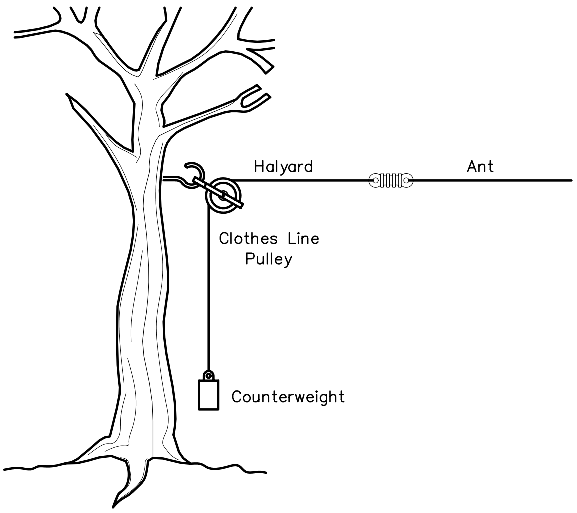

Fig 1—A method of counter weighting to minimize antenna movement and avoid its breaking from tree movement in the wind. The antenna may be lowered without climbing the tree by removing the counterweight and tying additional rope at the bottom end of the halyard. Excess rope may be left at the counterweight for this purpose, as the knot at the lower end of the halyard will not pass through the pulley.

ing to use, and often provide a means of supporting a wire antenna at considerable height. As antenna supports, trees are unstable in the presence of wind, except in the case of very large trees used to support antennas well down from the top branches. As a result, tree-supported antennas must be constructed much more sturdily than is necessary with stable supports. Even with rugged construction, it is unlikely that an antenna suspended from a tree, or between trees, will stand up indefinitely. Occasional repair or replacement usually must be expected.

There are two general methods of securing a pulley to a tree. If the tree can be climbed safely to the desired level, a pulley can be attached to the trunk of the tree, as shown in **Fig 1**. To clear the branches of the tree, the antenna end of the halyard can be tied temporarily to the tree at the pulley level. Then the remainder of the halyard is coiled up, and the coil thrown out horizontally from this level, in the direction in which the antenna runs. It may help to have the antenna end of the halyard weighted.

After attaching the antenna to the halyard, the other end is untied from the tree, passed through the pulley, and brought to ground along the tree trunk in as straight a line as possible. The halyard need only be long enough to reach the ground after the antenna has been hauled up. (Additional rope can be tied to the halyard when it becomes necessary to lower the antenna.)

The other method consists of passing a line over the tree from ground level, and using this line to haul a pulley up into the tree and hold it there. Several ingenious methods have been used to accomplish this. The simplest method employs a weighted pilot line, such as fishing line or mason's chalk line. By grasping the line about two feet from the weight, the weight is swung back and forth, pendulum style, and then heaved with an underhand motion in the direction of the treetop.

Several trials may be necessary to determine the optimum size of the weight for the line selected, the distance between the weight and the hand before throwing, and the point in the arc of the swing where the line released. The weight, however, must be sufficiently large to carry the pilot line back to ground after passing over the tree. Flipping the end of the line up and down so as to put a traveling wave on the line often helps to induce the weight to drop down if the weight is marginal. The higher the tree, the lighter the weight and the pilot line must be. A glove should be worn on the throwing hand, because a line running swiftly through the bare hand can cause a severe burn.

If there is a clear line of sight between ground and a particularly desirable crotch in the tree, it may eventually be possible to hit the crotch after a sufficient number of tries. Otherwise, it is best to try to heave the pilot line completely over the tree, as close to the centerline of the tree as possible. If it is necessary to retrieve the line and start over again, the line should be drawn back very slowly; otherwise the swinging weight may wrap the line around a small limb, making retrieval impossible.

Stretching the line out straight on the ground before throwing may help to keep the line from snarling, but it places extra drag on the line, and the line may snag on obstructions overhanging the line when it is thrown. Another method is to make a stationary reel by driving eight nails, arranged in a circle, through a 1-inch board. After winding the line around the circle formed by the nails, the line should reel off readily when the weighted end of the line is thrown. The board should be tilted at approximately right angles to the path of the throw.

Other devices that have been used successfully to pass a pilot line over a tree are a bow and arrow with heavy thread tied to the arrow, and a short casting rod and spinning reel used by fishermen. The Wrist Rocket slingshot made from surgical rubber tubing and a metal frame has proved highly effective as an antenna-launching device. Still another method that has been used where sufficient space is available is flying a kite to sufficient altitude, walking around the tree until the kite string lines up with the center of the tree, and paying out string until the kite falls to the earth. This method can be used to pass a line over a patch of woods between two higher supports, which may be impossible using any other method.

The pilot line can be used to pull successively heavier lines over the tree until one of adequate size to take the strain of the antenna has been reached. This line is then used to haul a pulley up into the tree after the antenna halyard has been threaded through the pulley. The line that holds the pulley must be capable of withstanding considerable chafing where it passes through the crotch, and at points where lower branches may rub against the standing part. For this reason, it may be advisable to use galvanized sash cord or stranded guy wire for raising the pulley.

Larger lines or cables require special attention when they must be spliced to smaller lines. A splice that minimizes the chances of coming undone when coaxed through the tree crotch must be used. One type of splice is shown in **Fig 2**.

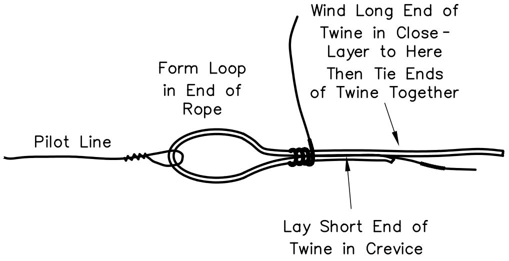

Fig 2—In connecting the halyard to the pilot line, a large knot that might snag in the crotch of a tree should be avoided, as shown.

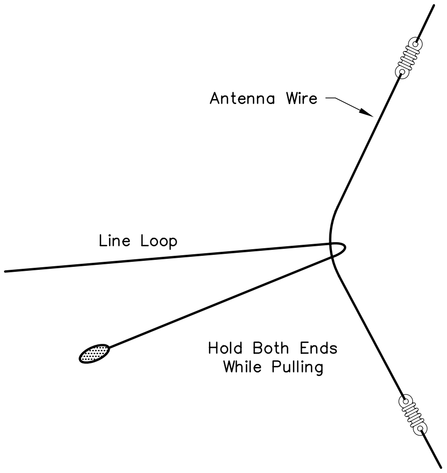

Fig 3—A weighted line thrown over the antenna can be used to pull the antenna to one side of overhanging obstructions, such as tree branches, as the antenna is pulled up. When the obstruction has been cleared, the line can be removed by releasing one end.

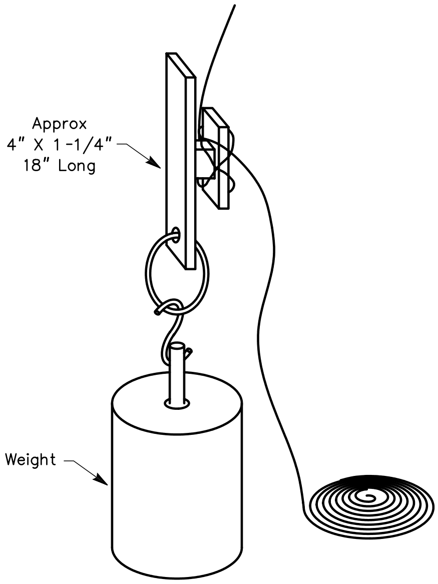

Fig 4—The cleat eliminates the need to untie a knot that may be weather hardened.

The crotch in which the line first comes to rest may not be sufficiently strong to stand up under the tension of the antenna. If, however, the line has been passed over (or close to) the center line of the tree, it will usually break through the lighter crotches and come to rest in a stronger one lower in the tree.

Needless to say, any of the suggested methods should be used with due respect to persons or property in the immediate vicinity. A child's sponge-rubber ball (baseball size) makes a safe weight for heaving a heavy thread line or fishing line.

If the antenna wire snags in the lower branches of the tree when the wire is pulled up, or if other trees interfere with raising the antenna, a weighted line thrown over the antenna and slid to the appropriate point is often helpful in pulling the antenna wire to one side to clear the interference as the antenna is being raised. This is shown in **Fig 3**.

Wind Compensation

The movement of an antenna suspended between supports that are not stable in the wind can be reduced by the use of heavy springs, such as screen-door springs under tension, or by a counterweight at the end of one halyard. This is shown in Fig 1. The weight, which may be made up of junkyard metal, window sash weights, or a galvanized pail filled with sand or stone, should be adjusted experimentally for best results under existing conditions. **Fig 4** shows a convenient way of fastening the counterweight to the halyard. It eliminates the necessity for untying a knot in the halyard, which may have hardened under tension and exposure to the weather.

TREES AS SUPPORTS FOR VERTICAL WIRE ANTENNAS

Trees can often be used to support vertical as well as horizontal antennas. If the tree is tall and has overhanging branches, the scheme of **Fig 5** may be used. The top end of the antenna is secured to a halyard passed over the limb, brought back to ground level, and fastened to the trunk of the tree.

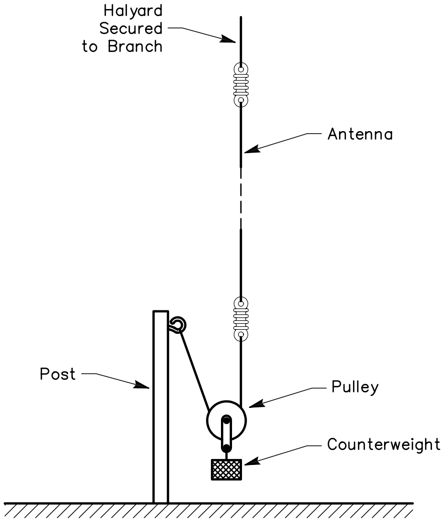

Fig 5—Counterweight for a vertical antenna suspended from an overhanging tree branch.

MAST MATERIALS

Where suitable trees are not available, or a more stable support is desired, light-duty guyed masts are suitable for wire antennas of reasonable span length. At one time, most amateur masts were constructed of lumber, but the TV industry has brought out metal masts that are inexpensive and much more durable than wood. However, there are some applications where wood is necessary or desirable.

A Ladder Mast

A temporary antenna support is sometimes needed for an antenna system for antenna testing, site selection, emergency exercises or Field Day. Ordinary aluminum extension ladders are ideal candidates for this service. They are strong, light, extendable, weatherproof and easily transported. Additionally, they are readily available and can be returned to normal use once the project is concluded. A ladder tower will support a lightweight triband beam and rotator.

With patience and ingenuity one person can erect this assembly. One of the biggest problems is holding the base down while "walking" the ladder to a vertical position. The ladder can be guyed with $^1/_4$-inch polypropylene rope. Rope guys are arranged in the standard fashion with three at each level. If help is available, the ladder can be walked up in its retracted position and extended after the antenna and rotator are attached. The lightweight pulley system on most extension ladders is not strong enough to lift the ladder extension. This mechanism must be replaced (or augmented) with a heavy-duty pulley and rope. Make sure when attaching the guy ropes that they do not foul the operation of the sliding upper section of the ladder.

There is one hazard in this system that must be avoided: Do not climb or stand on the ladder when it is being extended—even as much as one rung. Never stand on the ladder and attempt to raise or lower the upper section. Do all the extending and retracting with the heavy-duty rope and pulley!

If the ladder is to be raised by one person, use the following guidelines. First, make sure the rung-latching mechanism operates properly before beginning. The base must be hinged so that it does not slip along the ground during erection. The guy ropes should be tied and positioned in such a way that they serve as safety constraints in the event that control of the assembly is lost. Have available a device (such as another ladder) for supporting the ladder during rest periods. (See **Fig 6**.)

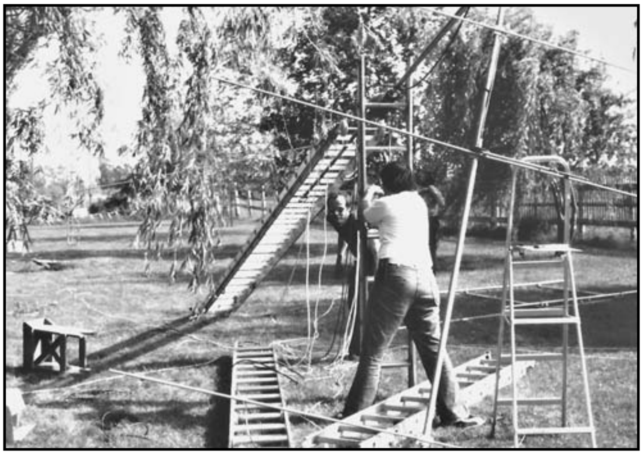

Fig 6—Walking the ladder up to its vertical position. Keith, VE2AQU, supports the mast with a second ladder while Chris, VE2FRJ, checks the ropes. (*Photo by Keith Baker, VE2XL.*)

After the ladder is erect and the lower section guys tied and tightened, raise the upper portion one rung at a time. Do not raise the upper section higher than it is designed to go; safety is far more important than a few extra feet of height.

For a temporary installation, finding suitable guy anchors can be an exercise in creativity. Fence posts, trees, and heavy pipes are all possibilities. If nothing of sufficient strength is available, anchor posts or pipes can be driven into the soil. Sandy soil is the most difficult to work with because it does a very poor job of holding anchors. A discarded car axle can be driven into the ground as an anchor, as its mass and strength are substantial. A chain and car-bumper jack can be used to remove the axle when the operation is done.

Above all else, keep the tower and antenna away from power lines. Make sure that nothing can touch the lines if the assembly falls. Disassemble by reversing the process. Ladder towers are handy for "quickie" antenna supports, but as with any improvisation of support materials, care must be taken to ensure safe construction.

The A-Frame Mast

A light and relatively inexpensive mast is shown in **Fig 7**. In lengths up to 40 feet it is very easy to erect and will stand the pull of ordinary wire antenna systems. The lumber used is 2 × 2-inch straight-grained pine (which many lumber yards know as hemlock) or even fir stock. The uprights can be as long as 22 feet each (for a mast slightly over 40 feet high) and the cross pieces are cut to fit. Four pieces of 2 × 2 lumber, each 22 feet long, provides more than enough. The only other materials required are five $^1/_4$-inch carriage bolts $5^1/_2$ inches long, a few spikes, about 300 feet of stranded or solid galvanized wire for guying, enough glazed porcelain compression ("egg") insulators to break up the guys into sections, and the usual pulley and halyard rope. If the strain insulators are put in every 20 feet, approximately 15 of them will be enough.

After selecting and purchasing the lumber—which should be straight-grained and knot-free—sawhorses or boxes should be set up and the mast assembled as shown in **Fig 8**. At this stage it is wise to give the mast a coat of primer and a coat of outside white latex paint.

After the coat of paint is dry, attach the guys and rig the pulley for the antenna halyard. The pulley anchor should be at the point where the top stays are attached so the backstay will assume the greater part of the load tension. It is better to use wire wrapped around the mast with a small through-bolt to prevent sliding down than to use eyebolts.

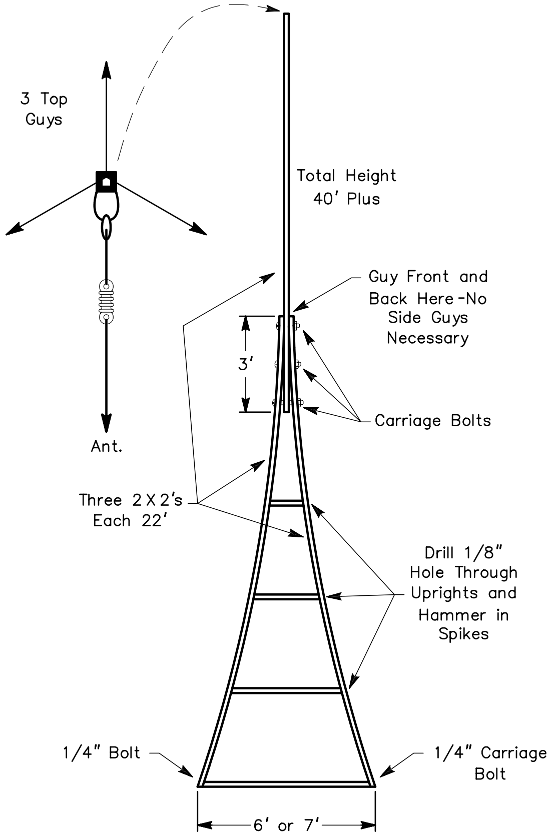
3 Top Guys
Total Height 40′ Plus
Guy Front and Back Here-No Side Guys Necessary
3′
Carriage Bolts
Ant.
Three 2 X 2′s Each 22′
Drill 1/8″ Hole Through Uprights and Hammer in Spikes
1/4″ Bolt
1/4″ Carriage Bolt
6′ or 7′

Fig 7—The A-frame mast is lightweight and easily constructed and erected.

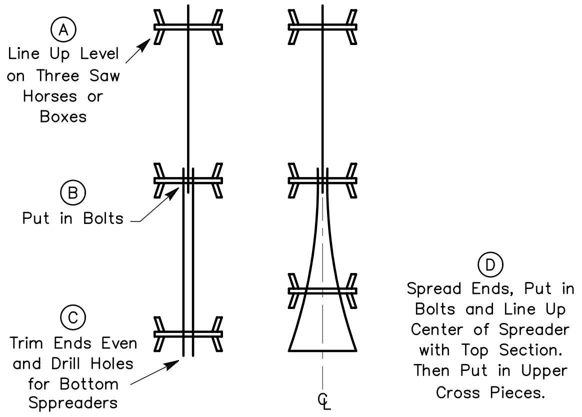
Ⓐ
Line Up Level on Three Saw Horses or Boxes
Ⓑ
Put in Bolts
Ⓒ
Trim Ends Even and Drill Holes for Bottom Sppreaders
Ⓓ
Spread Ends, Put in Bolts and Line Up Center of Spreader with Top Section. Then Put in Upper Cross Pieces.
℄

Fig 8—Method of assembling the A-frame mast on sawhorses.

If the mast is to stand on the ground, a couple of stakes should be driven to keep the bottom from slipping. At this point the mast may be "walked up" by a helper. If it is to go on a roof, first stand it up against the side of the building and then hoist it, from the roof, keeping it vertical. The whole assembly is light enough for two men to perform the complete operation—lifting the mast, carrying it to its permanent berth, and fastening the guys with the mast vertical. It is entirely practical to put up such a mast on a flat area of roof that would be too small to erect a regular tower installation, one that had to be raised vertically on the same spot.

TV Mast Material

TV mast is available in 5- and 10-foot lengths, $1^{1}/_{4}$ inches diameter, in both steel and aluminum. These sections are crimped at one end to permit sections to be joined together. A form that is usually more convenient is the telescoping mast available from many electronic supply houses. The masts may be obtained with three, four or five 10-foot sections, and come complete with guying rings and a means of locking the sections in place after they have been extended. These masts are inherently more suitable for guyed mast installations than the non-telescoping type because the diameters of the sections increase toward the bottom of the mast. For instance, the top section of a 50-foot mast is $1^{1}/_{4}$ inches diameter, and the bottom section is $2^{1}/_{2}$ inches diameter.

Guy rings are provided at 10-foot intervals, but guys may not be required at every point. Guying is essential at the top and at least one other place near the center of the mast. If the mast has any tendency to whip in the wind, or to bow under the load of a horizontal wire antenna, additional guys should be added at the appropriate points.

MAST GUYING

Three guy wires in each set are usually adequate for a mast. These should be spaced equally round the mast. The required number of sets of guys depends on the height of the mast, its natural sturdiness (or stiffness), and the required antenna tension. A 30-foot-high mast usually requires two sets of guys, and a 50-foot mast needs at least three sets. One guy of the top set should be anchored to a point directly opposing the force exerted by the wire antenna. The other two guys of the same set should be spaced 120° with respect to the first, as shown in the inset in Fig 7.

Generally, the top guys should be anchored at distances from the base of the mast at least 60% of the mast height. The distance of the guy anchors from the mast determines the guy loads and the vertical load compressing the mast. At a 60% distance, the load on the guy wire opposite the wire antenna is approximately twice the antenna tension. The compression in the mast will be 1.66 times the antenna tension. With the anchors out 80% of the mast height, the guy tension will be 1.6 times larger than the antenna load and the mast compression will be 1.25 times larger.

Whenever possible, the largest available anchor spacing should be used. The additional compression on the mast, due to closer anchor spacing, increases the tendency of the mast to buckle. Buckling occurs when the compression on the unsupported spans between guys become too great for the unsupported length. The section then bows out laterally and will usually fold over, collapsing the mast. Additional sets of guys reduce the tendency for the mast to buckle under the compression by decreasing the unsupported span lengths and stabilizing the mast, keeping it in a straight line.

A natural phenomenon, called *vortex shedding,* can occur when the wind passes over the sections of a guyed mast. For every section size, shape, and length, there is a wind speed that can cause the sections to oscillate mechanically. When all the sections of an antenna support mast are close to the same size and length, it is possible for all of the mast sections to vibrate simultaneously between the guys. To reduce the potential for this, you can place the guys at locations along the mast that will result in different span lengths. This creates different mechanical resonant frequencies for each span, eliminating the possibility of all sections oscillating at the same time.

When determining the guy locations along the mast to treat this problem, you also need to consider the mast buckling requirements. Since the compression in the mast is greatest in the bottom span, and the least in the top span, the guys should be placed to make the bottom span the shortest and the top span the longest. A general guide for determining the different span lengths is to make the unguyed lengths change by 10 to 20%.

Example: For a 30-foot high mast with three guy sets, the equal-guy locations would be every 10 feet. We can make the center span, 10 feet long, and then make the lower span 15% shorter and the top span 15% longer. While this is not an exact technical method to determine the best solution, the approach will create different mechanical resonant frequencies for the spans, with the span lengths approximately adjusted for the varying buckling requirements.

You can eliminate electrical resonance from conductive guy materials that might cause distortion of the antenna radiation pattern by breaking each guy into nonresonant lengths using strain insulators (see **Figs 9** and **10**). This subject is covered in detail later in this chapter.

Guy Material

When used within their safe load ratings, you may use any of the halyard materials listed in Chapter 20 for the mast guys. Nonmetallic materials have the advantage that there is no need to break them up into sections to avoid unwanted resonant interactions. All of these materials are subject to *stretching,* however, which causes mechanical problems in permanent installations. At rated working loads, dry manila rope stretches about 5%, while nylon rope stretches about 20%. Usually, after a period of wind load and wet/dry cycles, the lines will become fairly stable and require less frequent adjustment.

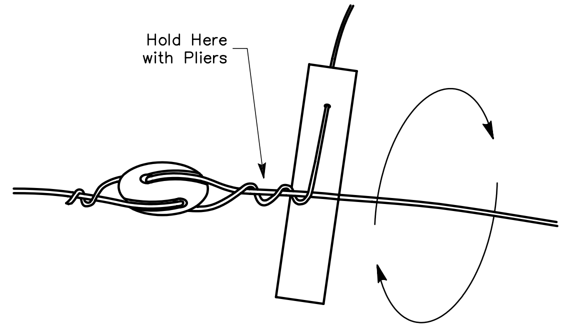

Fig 9—Simple lever for twisting solid guy wires when attaching strain insulators.

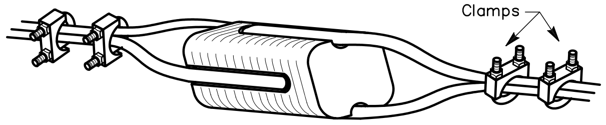

Fig 10—Stranded guy wire should be attached to strain insulators by means of standard cable clamps made to fit the size of wire used.

Solid galvanized steel wire is also widely used for guying. This wire has approximately twice the load ratings of similar sizes of copper-clad wire, but it is more susceptible to corrosion. Stranded galvanized wire sold for guying TV masts is also suitable for light-duty applications, but is also susceptible to corrosion. It is prudent to inspect the guys every six months for signs of deterioration or damage.

Guy Anchors

Figs 11 and **12** show two different kinds of guy anchors. In Fig 11, one or more pipes are driven into the ground at right angles to the guy wire. If a single pipe proves to be inadequate, another pipe can be added in tandem, as shown, and connected with a galvanized steel cable. Heavy-gauge galvanized pipe is preferred for corrosion resistance. Steel fence posts may be used in the same manner. Fig 12 shows a *dead-man* type of

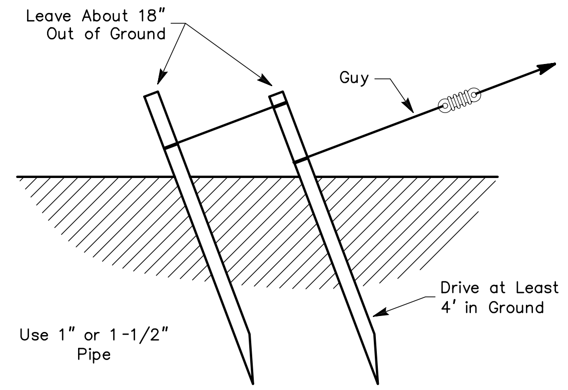

Fig 11—Driven guy anchors. One pipe is usually sufficient for a small mast. For added strength, a second pipe may be added, as shown.

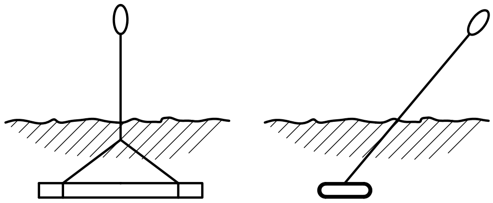

Fig 12—Buried *dead-man* guy anchor (see text).

anchor. The buried anchor may consist of one or more pipes 5 or 6 feet long, or scrap automobile parts, such as bumpers or wheels. The anchors should be buried 3 or 4 feet in the ground. The cable connecting the dead-man to the guys should be galvanized wire rope, like EHS guy cable. You should coat the buried part of the cable with roofing tar, and thoroughly dry it prior to burial to enhance resistance to corrosion.

Also available are some heavy auger-type anchors that screw into the earth. These anchors are usually heavier than required for guying a mast, although they may be more convenient to install. You should conduct annual inspections of the anchors by digging several inches below grade around the anchor to inspect for corrosion.

Trees and buildings may also be used as guy anchors if they are located appropriately. Care should be exercised, however, to make sure that the tree is of adequate size and that any fastening to a building can be made sufficiently secure.

Guy Tension

Many troubles encountered in mast guying are a result of pulling the guy wires too tight. Guy-wire tension should never be more than necessary to correct for obvious bowing or movement under wind pressure. Approximately 10% to 15% of the working load is sufficient. In most cases, the tension needed does not require the use of turnbuckles, with the possible exception of the guy opposite a wire antenna. If any great difficulty is experienced in eliminating bowing from the mast, the guy tension should be reduced or additional sets of guys are required. The mast should be checked periodically, especially after large wind events, to ensure the guys and anchors have not stretched or moved, allowing the mast to get away from the required straight alignment.

ERECTING A MAST OR OTHER SUPPORT

Masts less than 30 feet high usually can be simply walked up after blocking the bottom end securely. Blocking must be done so that the base can neither slip along the ground nor upend when the mast is raised. An assistant should be stationed at each guy wire, and may help by pulling the proper guy wire as the mast nears the vertical position. Halyards can be used in the same manner.

As the mast is raised, it may be helpful to follow the underside of the mast with a scissors rest (**Fig 13**), should a pause in the hoisting become necessary. The rest may also be used to assist in the raising if an assistant mans each leg.

As the mast nears the vertical position, those holding the guy wires should be ready to temporarily fasten the guys to prevent the mast from falling. The guys can then be adjusted until the mast is perfectly straight.

For masts over 30 feet long, a *gin pole* of some form may be required, as shown in Fig 13. Several turns of

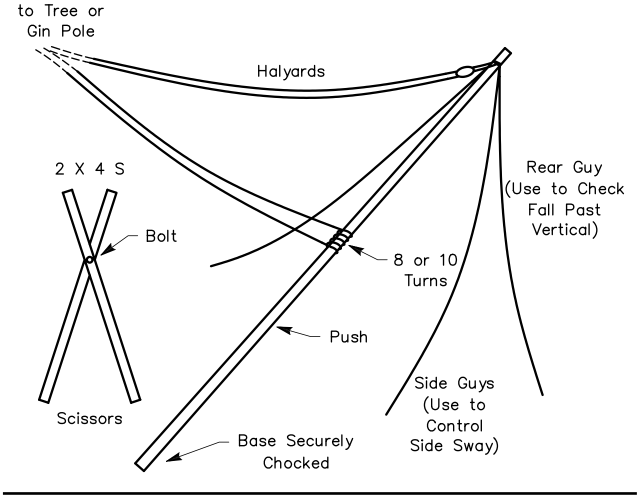

Fig 13—Pulling on a gin line fastened slightly above the center point of the mast and on the halyards can assist in erecting a tall mast. The tensions should be just enough to keep the mast in as straight a line as possible. The "scissors" may be used to push on the under side and to serve as a rest if a pause in raising becomes necessary.

rope are wound around a point on the mast above center. The ends of the rope are then brought together and passed over a tree limb. The rope should be pulled as the mast is walked up to keep the mast from bending at the center. If a tree is not available, a post, such as a 2 × 4, temporarily erected and guyed, can be used. After the mast has been erected, the assisting rope can be removed by walking one end around the mast (inside the guy wires).

Telephone poles and towers are much sturdier supports. Such supports may require no guying, but they are not often used solely for the support of wire antennas because of their relatively high cost. For antenna heights in excess of 50 feet, however, they are usually a most practical form of support.

Tower And Antenna Selection and Installation

The selection of a tower, its height, and the type of antennas and rotator is probably one of the more complex issues faced by station builders. All aspects of the tower, antenna, and rotator system are interrelated, and you should consider the overall system before making any decisions regarding specific system components.

Perhaps the most important consideration for many amateurs is the effect of the antenna system on the surrounding environment. If plenty of space is available for a tower installation and if there is little chance of causing esthetic distress on the part of family members or the neighbors, the amateur is indeed fortunate. Often, the primary considerations are purely financial. For most, however, the size of the property, the effect of the system on others, local ordinances, and the proximity of power lines and poles influence the selection of the tower/antenna system considerably.

The amateur must consider the practical limitations for installation. Some points for consideration are given below:

1) A tower should not be installed in a position where it could fall onto a neighbor's property.
2) The antenna must be located in such a position that it cannot possibly tangle with power lines, either during normal operation or if the structure should fall.
3) Sufficient yard space must be available to position a guyed tower properly. The guy anchors should be between 60% and 80% of the tower height in distance from the base of the tower on level ground—sloping terrain may require larger areas.
4) Provisions must be made to keep children from climbing the support. (Poultry netting around the tower base will serve this need.)
5) Local ordinances should be checked to determine if any legal restrictions affect the proposed installation.

Other important considerations are (1) the total dollar amount to be invested, (2) the size and weight of the antenna desired, (3) the climate, and (4) the ability of the owner to climb a fixed tower.

Most tower manufacturers provide catalogues or data packages that represent engineered tower configurations. These are provided as a convenience for users to help determine the most suitable tower configurations. The most commonly used design specifications for towers are EIA (Electronic Industries Assoc.) RS-222 and UBC (Uniform Building Code). These specifications define how the tower, antenna, and guy loads are determined and applied to the system, and establish general design criteria for the analysis of the tower. Local authorities often require the review and approval of the installation by a state licensed Professional Engineer (P.E.) to obtain building permits. All local authorities in the United States do not subscribe to the same design standards, so often the manufacturers' general-purpose engineering is not applicable.

One of the first things you need to determine in the tower selection process is the type of specification required by the local authorities, if any. Then, you must determine the *Basic Wind Speed* appropriate for the site. The *Basic Wind Speed* used in most specifications is the average wind speed for one mile of wind passing across the structure. It will be a lower value than the peak readings from an anemometer (wind gauge) installed at the site. For example, a Basic Wind Speed of 70 mph could have a maximum value of 80 mph and a minimum of 60 mph, equally distributed during the passage of the mile of wind. Basic wind speeds can be found in tables or maps contained in the appropriate specifications. Often, the basic wind speed used for the location may be obtained from the local permit authority. Check out the Web site at **www.championradio.com**, which contains EIA basic wind speed tables for every county in the USA. UBC speeds are available at almost every local library.

Antenna manufacturers also provide antenna data to assist in the selection process. Unfortunately, antenna mechanical designs do not always follow the same design standards used for towers. Proper antenna selection often means that you must determine the antenna surface areas yourself to avoid overloading the tower. More discussion about this follows later in this chapter.

It is often very helpful to the novice tower installer to visit other local amateurs who have installed towers. Look over their hardware and ask questions. If possible, have a few local experienced amateurs look over your plans—before you commit yourself. They may be able to offer a great deal of help. If someone in your area is planning to install a tower and antenna system, be sure to offer your assistance. There is no substitute for experience when it comes to tower work, and your experience there may prove invaluable to you later.

THE TOWER

Towers for supporting antennas come in a variety of different types. Each type has its own set of benefits and limitations, or conditions and requirements. Often, you can choose a particular tower type by considering issues other than pure mechanical performance. Understanding how each type of tower functions, and what its respective requirements are, are the first steps in making the best tower selection for your own situation.

Guyed Towers

The most common variety of tower is the *guyed tower* made of identical stacked sections, supported by guy cables attached to ground anchors placed symmetrically around the tower. These towers are the most economical, in terms of feet per dollar investment, and are more efficient for carrying antenna loads than non-guyed towers.

The guys resist the lateral loads on the system created by the wind. Since the guys slope down to the ground, horizontal loads due to the wind result in vertical loads applied to the tower at each tower/guy connection. The tower becomes a compression member, trying to resist the column compression generated by the guy reactions. A tower in compression can buckle, so the distance between guy connections along the tower is important.

Tower Bases for Guyed Towers

Another important phenomenon in a guyed tower is stretching of the guy cables. All guys stretch under load and when the wind blows the elongated guys allow the tower to lean over somewhat. If the tower base is buried in the concrete footing—as is commonly done in amateur installations—the bending stress at the tower base can become a significant factor. Towers that have been installed with tapered pier-pin bases much more freely absorb tower leaning, and they are far less sensitive to guy-elongation problems.

The tapered pier-pin tower installation is not without some drawbacks. These installations often require torque-arm guy brackets or six-guy torque-arm assemblies to control tower rotation due to antenna torque. They also require temporary guys when they are being installed to hold the base steady until the permanent guys are mounted. Some climbers also don't like the flexing when they start to climb these types of towers.

On the positive side, pier-pin base towers have all structural members above the concrete footing, eliminating concerns about hidden corrosion that can occur with buried towers. Most decisions regarding the type of base installation are made according to the preference of the tower builder/maintainer. While either type of base configuration can be successfully used, you would be wise to do the stress calculations (or have a professional engineer do them) to ensure safety, particularly when large antenna loads are contemplated and particularly if guys that can easily stretch are used, such as Phillystran guys.

The configuration shown in **Fig 14A** is taken from an older (1983) Unarco-Rohn catalog. This configuration has the top set of guys placed at the top of the tower with the lower set halfway up the tower. This configuration is best for most amateur installations, which usually have the antennas mounted on a rotatable mast extending

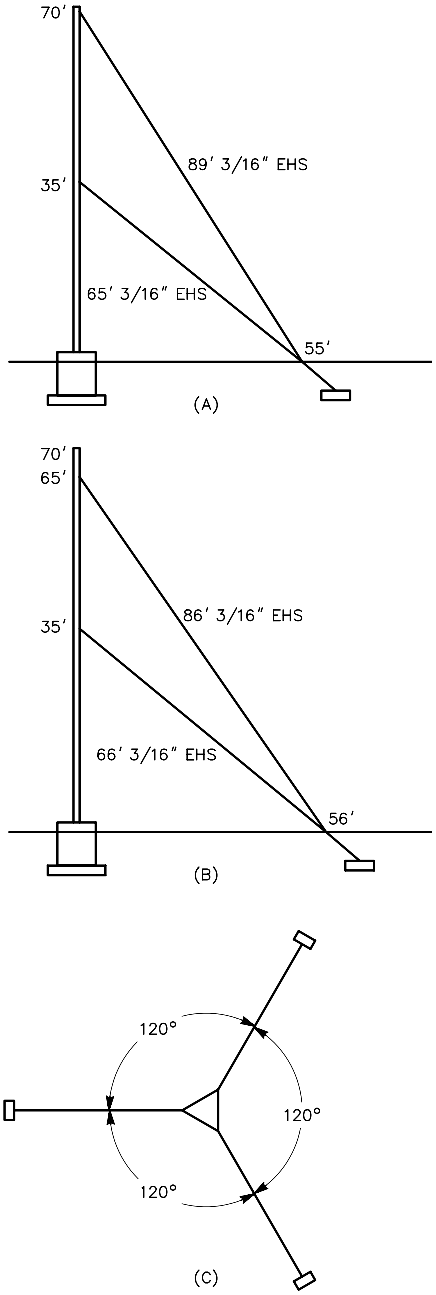

Fig 14—The proper method of installation of a guyed tower. At A, the method recommended for most amateur installations. At B, the method shown in later Rohn catalogs. This places considerable strain on the top section of the tower when large antennas are mounted on the tower.

out the top of the tower—thereby placing the maximum lateral loads when the wind blows at the top of the tower (and the bottom of the rotating mast).

The configuration shown in Fig 14B is from a newer (1998) Rohn catalog. It has 5 feet of unsupported tower extending above the top guy set. The lower guy set is approximately halfway between the top guys and the base. The newer configurations are tailored for commercial users who populate the top region of the tower with fixed arrays and/or dishes. The installation in Fig 14B cannot safely withstand the same amount of horizontal top load as can the configuration shown in Fig 14A, simply because the guys start farther down from the top of the tower.

An overhead view of a guyed tower is given in Fig 14C. Common practice is to use equal angular spacings of 120° between guy wires. If you must deviate from this spacing, the engineering staff of the tower manufacturer or a civil engineer should be contacted for advice.

Amateurs should understand that most catalogs show generic examples of tower configurations that work within the cited design specifications. They are by no means the only solution for any specific tower/antenna configuration. You can usually substantially change the load capability of any given tower by varying the size and number of guys. Station builders are encouraged to utilize the services of professional engineers to get the most out of their guyed towers. Those interested in more generic information about guyed tower behavior can find it at **www.freeyellow.com/members3/yagistress/**.

Unguyed Towers

Another commonly used type of tower is not normally guyed—these are usually referred to as *freestanding* or *self-supporting* towers. Unguyed towers come in three different styles.

One style is comprised of stacked lengths of identical tower sections, just like those used for guyed towers. The only difference is that no guys are used. Manufacturers provide the recommended configurations and allowable loads for this type of installation in their catalogs. Unguyed towers are vastly less capable of supporting antenna loads than their guyed counterparts, but have great utility for light-duty applications—when configured within their capabilities.

The second style utilizes different tower section sizes, varying from large sections at the base and tapering down to smaller sections at the top. This style is more much more efficient for freestanding applications, because the tower is sized for the varying bending loads along the tower length, and is shown in **Fig 15**.

The third style of unguyed towers is commonly called a *crank-up tower*. It is a freestanding tower with telescoping sections that can be extended or retracted with a winch, cable, and pulley mechanism. This allows the tower to be raised and lowered for maintenance and antenna work. It is usually necessary to retract such towers for moderate to heavy winds. Some consider this a disadvantage because they can't operate their antennas at full height when it is windy. Two different forms of the crank-up style, freestanding tower are shown in **Fig 16**. Fig 16A shows the tubular version; Fig 16B shows the triangular space-frame version.

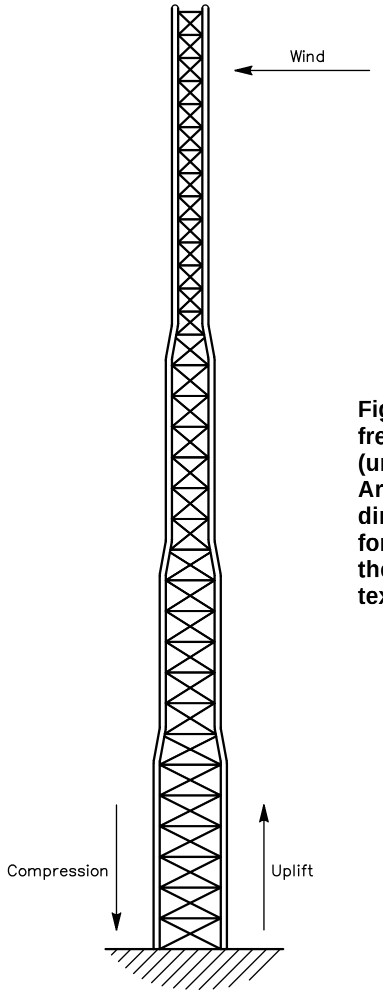

Fig 15—Typical freestanding (unguyed) tower. Arrows indicate the directions of the forces acting upon the structure. See text for discussion.

Some crank-up towers are used with guys and are only retracted for maintenance and antenna work. These towers are specially designed with locking mechanisms between the tower sections to carry the vertical compression created by the guys. *Do not* use guys with normal crank-up towers (those that have no locking devices between sections)! The increased tower compression will be carried by the hoisting cable, which will eventually cause it to fail.

Never climb a crank-up tower unless it is properly nested, with all load removed from the hoisting cable. For general antenna work, this can be accomplished by completely retracting it until the cable becomes loose. When servicing the rotator, the tower must be left partially extended. In this case every tower section must be

Fig 16—Two examples of *crank-up* towers.

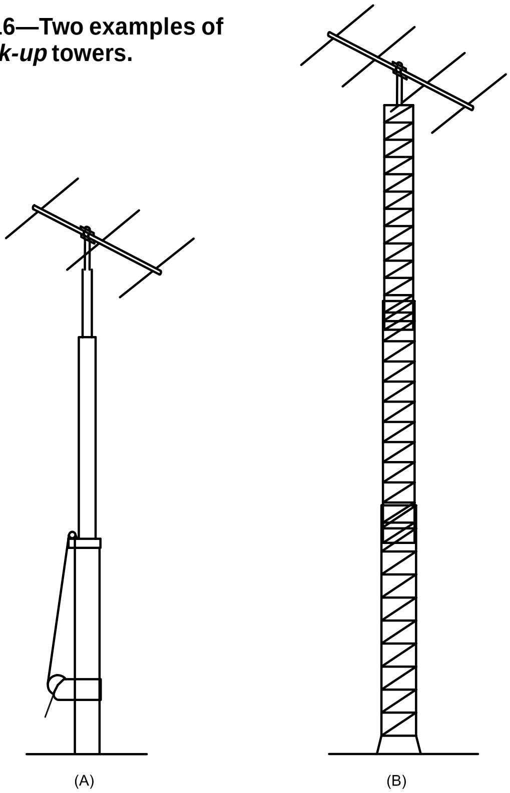

blocked with heavy timber or thick-wall tubes, installed through the tower bracing, until all sections are resting on the blocks and the hoisting cable becomes slack. Safely installing the blocks in an extended crank-up tower can be challenging. The object is to get all the blocks installed without a climber having to scale the unblocked tower, risking loss of limbs should the hoisting cable fail. An extension ladder, capable of reaching the required block elevations is the safest approach. If the necessary equipment or expertise is not available, the tower can be retracted, antennas removed, and leaned over, with the base tilt-over assembly, before extending it to access the rotator. Failure to properly block the tower before climbing can result in serious injury should the cable slip or break!

All freestanding towers share some unique characteristics. Each must support antenna and tower loads only by virtue of the bending strength of the tower sections and the tower footing connection to earth. Because of the large overturning moment at the tower base, freestanding towers require larger concrete footings than guyed towers. They are usually more expensive for the same load capability compared to guyed towers, simply because they require larger heavier tower sections and a larger footing to get the job done. The telescoping mechanisms in crank-up tower require more maintenance too.

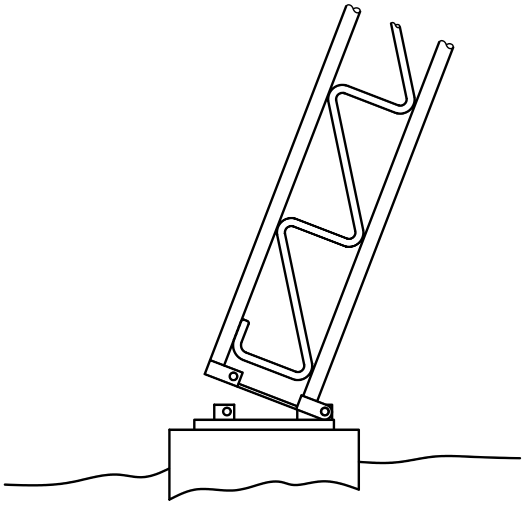

Fig 17—Fold-over or tilting base. There are several different kinds of hinged sections permitting different types of installation. Great care should be exercised when raising or lowering a tilting tower.

Freestanding towers are quite popular, and are often the best solutions for sites with limited space and ascetic concerns. When cranked down, a telescoping tower can maintain a low-profile system, out of sight of the neighbors and family.

Tilt-Over Towers

Some towers have another convenience feature—a hinged section that permits the owner to fold over all or a portion of the tower. The primary benefit is in allowing antenna work to be done close to ground level, without the necessity of removing the antenna and lowering it for service. **Fig 17** shows a hinged base used with stacked, guyed tower sections. The hinged section can be designed for portions of the tower above the base. These are usually referred to as *guyed tilt-over towers*, where a conventional guyed tower can be tilted over for installing and servicing antennas. Many crank-up towers come with optional tilt-over base fixtures that are equipped with a winch and cable system for tilting the fully nested tower from horizontal to vertical positions.

Misuse of hinged sections during tower erection is a dangerously common practice among radio amateurs. Unfortunately, these episodes can end in accidents. If you do not have a good grasp of the fundamentals of physics, it might be wise to avoid hinged towers or to consult an expert if there are any questions about safely installing and using such a tower. It is often far easier (and safer) to erect a regular guyed tower or self-supporting tower with gin pole and climbing belt than it is to try to walk up an unwieldy hinged tower.

The AB-577 Military Surplus Tower

Another light duty tower has found acceptance among many amateurs. Available from assorted military surplus dealers is the AB-577 system. This was designed to be a portable, rapid deployment antenna support for field communications. It is a guyed mast that goes up somewhat like a crank-up. The system consists of several short sections of aluminum tubing, with special end connections for joining them. These can be erected from the base fixture, which has a crank-up type winch-driven elevator platform. The tubing sections are installed in the base fixture and connected to the section above it with an over-center locking *Marmon*-style clamp. Then, the elevator platform is raised with the winch and the new tube is locked in place, high on the base fixture. Then the elevator is lowered to accept the next section. While the tower is extended, the supporting guys are adjusted via the unique *snubber* assemblies at the anchor connection. One person can erect this system, even in windy conditions, when special care is given to keeping the guys properly adjusted during each extension.

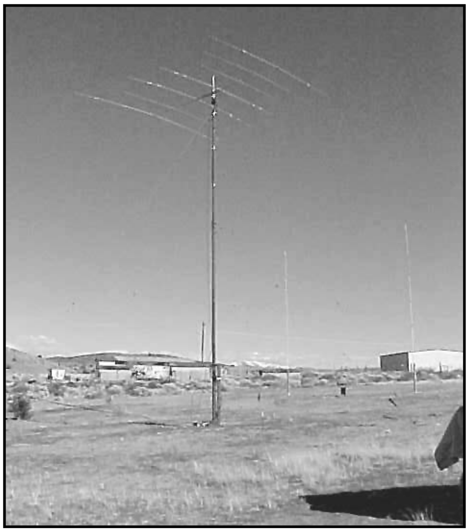

Fig 18—Installation of surplus AB-577 tower with tribander at 45 feet at K7NV. (*Photo by Kurt Andress, K7NV.*)

The standard AB-577 system, with 3 sets of guys, will support a modest triband Yagi at 45 feet. **Fig 18A** shows an installation with a Hy-Gain TH7DX at 45 feet.

TOWER BASES

Tower manufacturers can provide customers with detailed plans for properly constructing tower bases. **Fig 19** is an example of one such plan. This plan calls for a hole that is $3^{1}/_{2} \times 3^{1}/_{2} \times 6$ feet. Steel reinforcement bars are lashed together and placed in the hole. The bars are positioned so that they will be completely embedded in the concrete, yet will not contact any metallic object in the base itself. This is done to minimize the possibility of a direct discharge path for lightning through the base. Should such a lightning discharge occur, the concrete base could be damaged.

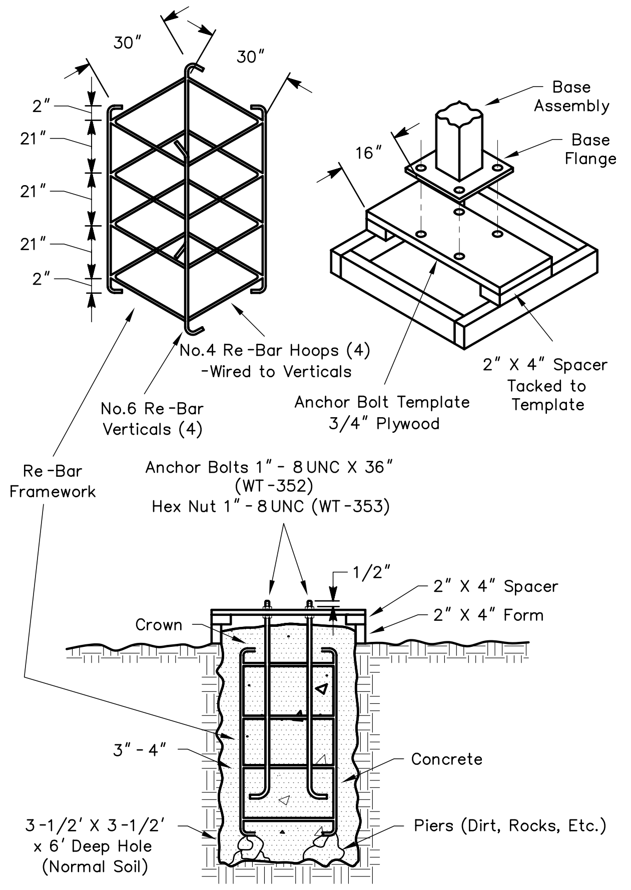
30″
30″
2″
21″
21″
21″
2″
Base Assembly
Base Flange
16″
No.4 Re-Bar Hoops (4) -Wired to Verticals
2″ X 4″ Spacer Tacked to Template
No.6 Re-Bar Verticals (4)
Anchor Bolt Template 3/4″ Plywood
Re-Bar Framework
Anchor Bolts 1″ - 8 UNC X 36″ (WT-352)
Hex Nut 1″ - 8 UNC (WT-353)
1/2″
2″ X 4″ Spacer
2″ X 4″ Form
Crown
3″ - 4″
Concrete
3-1/2′ X 3-1/2′ x 6′ Deep Hole (Normal Soil)
Piers (Dirt, Rocks, Etc.)

Fig 19—Plans for installing concrete base for Wilson ST-77B tower. Although the instructions and dimensions vary from tower to tower, this is representative of the type of concrete base specified by most manufacturers.

Providing suitable paths for the discharge of lightning energy safely for towers is a complex subject. Several companies offer products and guidance. The basic requirements for providing controlled discharge paths for lightning-induced current is to supply a low-impedance grid of conductors from the tower and feed lines to a field of interconnected ground rods around the base of the tower. Generally, the tower, station, and electrical service grounds need to be connected to prevent damaging potential differences from developing between the various components in the system.

A strong wooden form is constructed around the top of the hole. The hole and the wooden form are filled with concrete so that the resultant block will be 4 inches above grade. The anchor bolts are embedded in the concrete, and aligned with the plywood template, before it hardens. The template serves to align the anchor bolts to properly mate with the tower itself. Once the concrete has

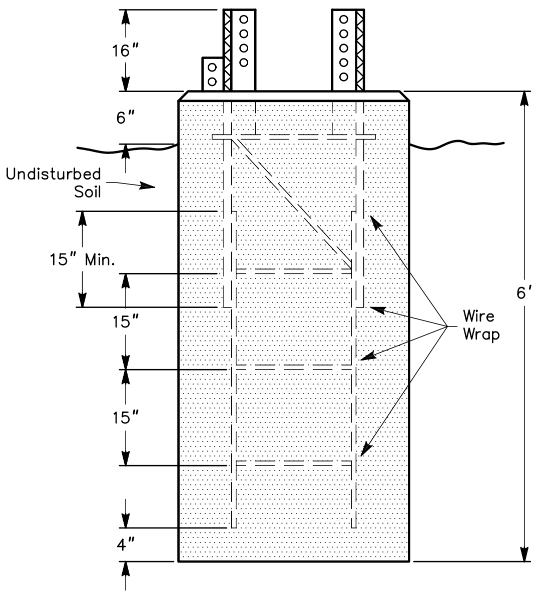

Fig 20—Another example of a concrete base (Tri-Ex LM-470).

cured, the tower base is installed on the anchor bolts and the base connection is adjusted to bring the tower into vertical alignment.

For a tower that bolts to a flat base plate mounted to the footing bolts (as shown in Fig 19), you can bolt the first tower section on the base plate to ensure that the base is level and properly aligned. Use temporary guys to hold things exactly vertical while the concrete cures. (The use of such temporary guys also works well when you place the first tower section in the base hole and plumb it vertically before pouring in the concrete.) Manufacturers can provide specific, detailed instructions for the proper mounting procedure. **Fig 20** shows a slightly different design for a tower base.

The one assumption so far is that *normal* soil is predominant in the area in which the tower is to be installed. Normal soil is a mixture of clay, loam, sand and small rocks. More conservative design parameters for the tower base should be adopted (usually, using more concrete) if the soil is sandy, swampy or extremely rocky. If there are any doubts about the soil, the local agricultural extension office can usually provide specific technical information about the soil in a given area. When this information is in hand, contact the engineering department of the tower manufacturer or a civil engineer for specific recommendations with regard to compensating for any special soil characteristics.

TOWER INSTALLATION

The installation of a tower is not difficult when the proper techniques are used. A guyed tower, in particular, is not hard to erect, because each of the individual sections is relatively lightweight and can be handled with only a few helpers and some good quality rope.

The Gin Pole and Tips on Tower Building

An essential piece of hardware for working on towers is a *gin pole*. This section came from the ARRL book *Simple and Fun Antennas for Hams*. The dictionary describes a gin pole as “a device for moving heavy objects.” See **Fig 21**, which shows a drawing of the Rohn “Erection Fixture” EF2545. This gin pole was designed to work with the nominal 10-foot long sections of Rohn 25 or 45 towers.

We’re going to assume in the following discussion that you are installing Rohn 45, which weighs about 70 pounds. This is a lot of weight, and you must refrain from adding to that during installation. That means, for example, that you *do not* attempt to lift a 10-foot section with the guy wires attached! Neither should you attempt to lift the top section with the rotator and rotor shelf installed. The gin pole (and your ground crew) will not appreciate all that strain.

The main working part of the gin pole is the pulley

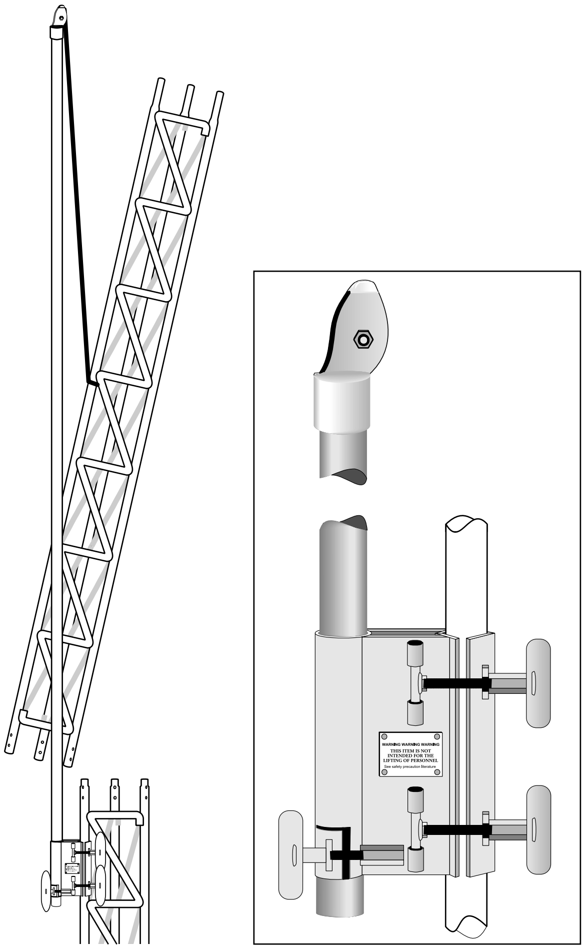

Fig 21—Drawing of Rohn “Erection Fixture” EF2545, also known commonly as a “gin pole.”

mounted at the top of the 12-foot long heavy-wall aluminum tubing. This pulley has a rope going down to the ground crew through the center of the aluminum tube. At the base of the tower, the pull rope should be run through a *snatch block* attached to the tower just above ground level. This block allows the pull rope to be pulled out horizontally away from the tower base. That helps protect ground crew should a tool be dropped by the people on the tower.

An adjustable, sliding clamp towards the bottom of the aluminum tubing is clamped to the tower using a swinging L-bracket-type clamp with two clamping bolts. These have T-bar handles that can be tightened by hand. In fact, this gin pole can be moved and deployed without any tools. The clamp is positioned on the top of the tower section onto which the next tower section is to be installed. Once clamped to the top of the tower, you would loosen the T-bar handle that tightens the clamp against the sliding aluminum tube and slide the tubing up to its maximum extent.

In practice, the following steps are taken as each 10-foot section of tower is installed, one-by-one. We're assuming here that the gin pole starts out on the ground, with at least one person belted in at the top of the tower. We're also assuming that the pull rope has been threaded through the aluminum tube and the top pulley, with a knot tied to prevent it from falling back down the tube.

1. The clamp holding the aluminum tubing is loosened so that the pulley on the tube can be lowered to where it is just above the bottom clamp. Then the T-bar handle for the tube clamp is tightened.
2. The climber lowers a tag rope for the ground crew to tie to the gin-pole pull rope. (This tag rope has been looped through a temporary pulley clipped to the top of the tower. It is also used to pull up tools and other materials.) The ground crew then pulls the gin pole up to the climber, using the tag line rope. Friction of the rope against the top of the pulley-head assembly will prevent the gin-pole assembly from slipping down. [Note that some climbers prefer to "walk" the gin-pole up the tower rather than having it pulled up from the ground below. They free up their hands for climbing and temporarily holding the gin pole by using their belt lanyard looped around the tower as they climb.]
3. Once the gin-pole head reaches the top of the tower, the climber clamps the gin pole clamp securely to the top of the tower.
4. The T-bar handle for the tube clamp is loosened, and the aluminum tube is extended to its maximum height, as shown in Fig 21. Make sure you have tied the free end of the rope coming through the top pulley temporarily to the top of the tower, or else you'll have to lower the gin pole and go through this step again.
5. The free end of the pulley rope is then dropped to the ground, often using a weight such as a medium Crescent wrench or perhaps a hammer to keep the rope from waving about as it dangles down the tower, tangling with every imaginable thing as it proceeds downwards. It's amazing how even a tiny breeze can make an unweighted rope dance like that.
6. The ground crew then ties the free end of the rope *above* the balance point of the tower. For Rohn 25 or 45 there are eight horizontal cross braces per section. You want the crew to tie the rope to the fifth horizontal brace from the bottom. Please remember that you want the bottoms of the tower sections' legs to be pointed downwards, not flipped over, when the bottom of these legs approach you at the top of the tower section the climber is standing on.
7. Once the bottom of the rising tower section is just above the top of the legs of the bottom tower section, the climber guides the tower down onto the top of the three legs, while calling out to the ground crew instructions about *slowly lowering* the new section down onto the legs. See **Fig 22**, which shows the climber guiding the new section of Rohn HBX tower onto the previous section's legs. This process is considerably easier to accomplish if each section of tower has been put together on the ground to make sure that the legs fit together easily. There's nothing more frustrating that trying to manually force-fit tower sections together at the top of a tower.

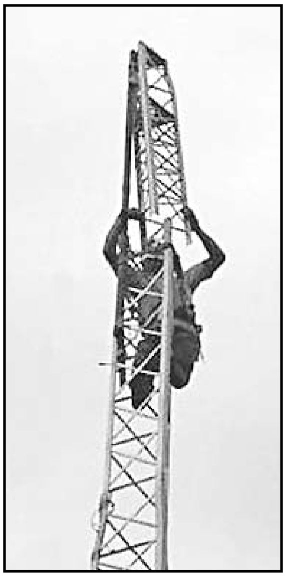

Fig 22—Climber is guiding the new section onto the top of the existing one. The gin pole attached to the left leg is bearing the weight, as the climber gives verbal instructions to the ground crew pulling on the gin-pole rope. (*Photo courtesy Mike Hammer, N2VR.*)

It seems that freight companies don't always handle heavy tower sections very gently and legs easily get bent out of alignment. A careful installer numbers tower sections in the order they've been test-fitted together on the ground, marking them with a laundry marker pen. You should also spray a small amount of WD-40 up inside mating tower legs after test-assembling them to help prevent galling and to ease fitting sections together. [Don't do this to excess—WD-40 is slippery and messy when it runs out of the bottom of tower legs.]

Another caution: Make sure before you start installing any tower that the correct ends of the bottom section's legs, "male" rather than "female," are pointed upwards. A prominent amateur (who will go unnamed) had to have Rohn make and send him a special "gender-bender" flange to turn females into males, since he had installed the base upside-down in the concrete base. You don't want to do that.

8. Once the new tower section has been guided down onto the male ends, the six pinning bolts are inserted and tightened with nuts. Note that Rohn uses two different sized bolts, with the larger diameter one on the bottom.
9. If this section of tower is one where guy wires are to be placed, they can be brought up using the gin pole rope and positioned on the tower. The maximum spacing for Rohn 25 is 30 feet between guy-wire sets, and 40 feet for Rohn 45. Thirty feet of unguyed Rohn 25 tower is wobbly, though safe. Many installers prefer to come down off the tower when setting guy-wire tension, since they do not like to be on a wobbly tower when the ground crew is moving around yanking on guy wires. Many also greatly prefer working on Rohn 45 tower, which is substantially more secure feeling and easier to stand on, with its legs 18 inches apart, while Rohn 25 legs are only 12 inches apart.
10. Finally, you reposition the gin pole for the next section of tower. The T-bar at the clamp is loosened, the tube is dropped down to the level of the clamp, and the climber walks the gin pole up to the top of the section just installed and clamps it there, ready to pull up the next tower section.

Tower Safety

One of the most important aspects of any tower installation project is the safety of all persons involved. See Chapter 1 for details on important safety issues. The use of hard hats is highly recommended for all assistants helping from the ground. Helpers should always stand clear of the tower base to prevent being hit by a dropped tool or hardware. Each person working on the tower must use a good climber's safety belt.

When climbing the tower, if more than one person is involved, one should climb into position before the other begins climbing. The same procedure is required for climbing down a tower after the job is completed. The purpose is to have the non-climbing person stand still so as not to drop any tools or objects on the climbing person, or unintentionally obstruct his movements. When two persons are working on top of a tower, only one should change position (unbelt and move) at a time.

For most installations, a good-quality $^{1}/_{2}$-inch diameter Manila hemp rope can adequately handle the workload for the hoisting tasks. The rope must be periodically inspected to assure that no tearing or chafing has developed, and if the rope should get wet from rain, it should be hung out to dry at the first opportunity. The knots used for connecting hoisting lines and hardware are critical to executing any safe installation, and special attention should be given to this detail for any work party.

Here is an important point regarding safety—the person who climbs the tower should be in charge of what happens with the ground crew. Not only does the person on the tower have a better overall view of the situation below, but also any confusion on the ground can result in serious injury to the climber.

GUY WIRES

In typical guyed tower installations, guy wires may experience loads in excess of 1000 pounds. Since the guys are the primary means of carrying the horizontal wind loads, great care should be taken in their selection and installation.

Guys come in a variety of materials and constructions. Normally, the tower manufacturer or professional engineer will specify the size and type of cable to be used. The most common type of cable used for tower guying is the EHS (Extra High Strength) galvanized steel cable. The EHS cables are very stiff and are the highest strength

Table 1
Guy Cable Comparisons

Cable	*Nominal Dia. Inches*	*Breaking Strength Lbs*	*Weight Lbs/100'*	*Elongation Inches/100'*	*%Elongation*
$^{3}/_{16}$" 1 × 7 EHS	0.188	3990	7.3	6.77	0.56%
$^{1}/_{4}$" 1 × 7 EHS	0.250	6700	12.1	3.81	0.32%
HPTG6700	0.220	6700	3.1	13.20	1.10%
HPTG8000	0.290	8000	3.5	8.90	0.74%
$^{5}/_{16}$" 1 × 7 EHS	0.313	11200	20.5	2.44	0.20%
HPTG11200	0.320	11200	5.5	5.45	0.45%
$^{3}/_{8}$" Fiberglass Rod	0.375	13000	9.7	5.43	0.45%

EHS steel cable information is taken from ASTM A 475-89, the industry standard specification for steel wire rope. The HPTG listings are for *Phillystran* aramid cables, and are based on the manufacturers' data sheets. The elongation (stretch) values are for 100 feet of cable with a 3000-pound load.

cables in the wire rope family. Other steel cables are made to be more flexible for running around pulleys. While these are easier to work with when assembling, they are not as strong as the EHS type, and should be avoided for tower guying. Non-conductive guys, such as *Phillystran* or *pultruded* fiberglass rod have become popular for eliminating resonant interaction with antennas.

Do not attempt to use cheaper cables that don't meet or exceed the criteria for those specified for your installation. Using the wrong cable, or failing to install the cable properly can have disastrous results! **Table 1** shows data for several cables commonly used for tower guying. It is important to note that the minimum breaking strength of the various cables are independent of their elongation (stretch) under load.

Guy Cable Installation

Figs 23 and **24** show methods for tensioning and safety wiring guy-wire turnbuckles. **Fig 25** shows the traditional method for fixing the end of a steel guy wire. A thimble is used to prevent the wire from breaking because of a sharp bend at the point of intersection. Conventional wisdom strongly recommends the use of thimbles that are at least one wire size larger than the cable to provide a more gentle wire bend radius. Three cable clamps follow to hold the wire securely. Be sure to follow the note in Fig 25 for which part of the clip bears against the live (loaded) cable. As a final backup measure, the individual strands of the free end are unraveled and wrapped around the guy wire. It is a lot of work, but it is necessary to ensure a safe and permanent connection.

Fig 26 shows the use of a device that replaces the clamps and twisted strands of wire. These devices are known as *dead ends*, *preformed guy grips,* or *Big Grips* and are commonly used on electrical power poles. They are far more convenient to use than are clamps, and are the recommended method for terminating Phillystran and

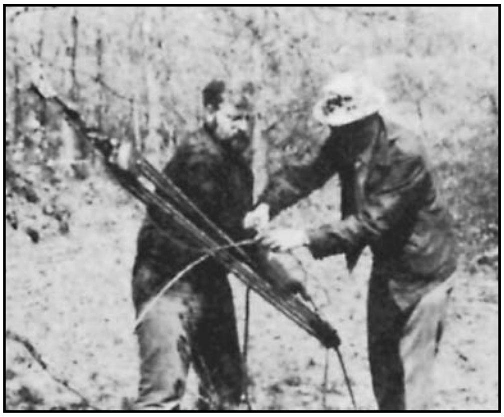

Fig 23—Proper tension can be placed on the guy wires with the aid of a block-and-tackle system. (*Photo by K1WA*)

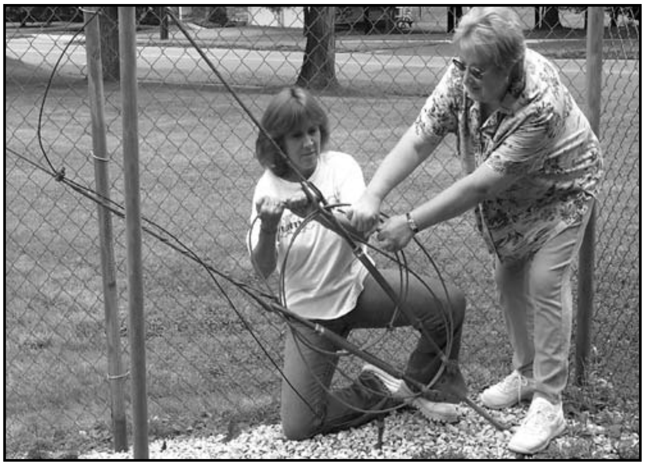

Fig 24—A length of guy cable is used to assure that the turnbuckles remain in place after they are tightened. This procedure is an absolute requirement in guyed tower systems. (*Photo by N4QX*)

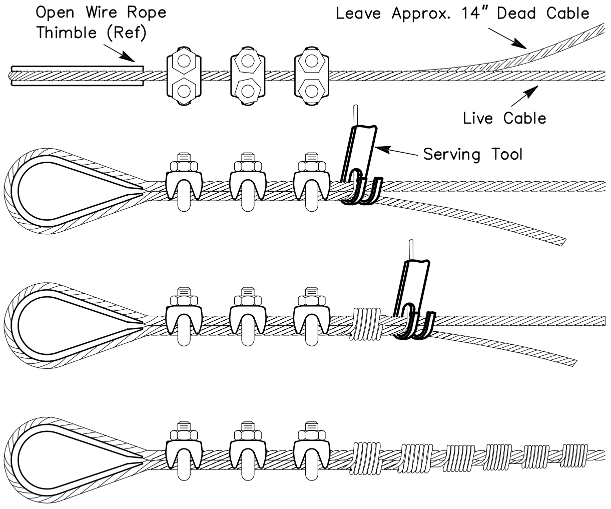

Fig 25—Traditional method for securing the end of a guy wire.

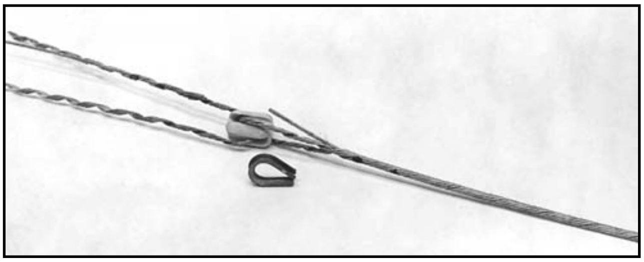

Fig 26—Alternative method for attaching guy wires using *dead ends.* The dead end on the right is completely assembled (the end of the guy wire extends beyond the grip for illustrative purposes). On the left, one side of the dead end is partially attached to the guy wire. In front, a thimble is used where a sharp bend might cause the guy wire or dead end to break.

fiberglass-rod guys. When using the guy grips, it is imperative that the recommended end sleeves are installed over the free end of the grip to prevent ice and falling hardware from sliding down the guy and unraveling the grip connection to the guy. The guy wires must be cut to the proper length. The dead end of each wire is installed into the object to which the guy wire is being attached (use a thimble, if needed to eliminate sharp cable bends). One side of the dead end is then wrapped around the guy wire. The other side of the dead end follows. Using dead ends saves time and trouble, more than making up for their slightly higher cost.

When using the non-conductive guy materials, it is highly recommended that a 25-foot length of EHS steel cable be used at the bottom for connection to the anchor. This serves a valuable purpose. The steel cable is more resistant to damage from ground activity and brush fires, and it is the preferred material for measuring cable pre-tension with commonly available devices.

Fig 27 shows two different methods for attaching guy wires to towers. At Fig 27A, the guy wire is simply looped around the tower leg and terminated in the usual manner. At Fig 27B, a *guy bracket*, with *torque arms* has been added. Even if the torque arms are not required, it is preferred to use the guy bracket to distribute the load from the tower/guy connection to all three tower legs, instead of just one. The torque bracket is more effective resisting torsional loads on the tower than the simpler installation. Rohn offers another guy attachment bracket, called a *Torque Arm Assembly*, that allows six guys to be connected between the bracket and anchors. This is by far the best method of stabilizing a tower against high torque loads, and is recommended for installations with large antennas.

There are two types of commonly used guy anchors. **Fig 28A** depicts an *earth screw*. These are usually 4 to 6 feet long. The screw blade at the bottom typically measures 6 to 8 inches diameter. Fig 28B illustrates two people installing the anchor. The shaft is tilted so that it will be in line with the mean angle of all the guys connecting to the anchor. Earth screws are suitable for use in normal soil where permitted by local building codes. Information about *screw anchors* is available from the manufacturers of these devices. Information from a supplier specializing in this type of anchor can be found at **www.abchance.com**.

The alternative to earth screws is the concrete block anchor. Fig 28C shows the installation of this type of anchor; it is suitable for any soil condition, with the possible exception of a bed of lava rock or coral. Consult the instructions from the manufacturer, or your tower designer, for the precise anchor configuration.

Turnbuckles and associated hardware are used to attach guy wires to anchors and to provide a convenient method for adjusting tension. **Fig 29A** shows a turnbuckle with a single guy wire attached to the eye of the anchor. Turnbuckles are usually fitted with either two eyes, or one eye and one jaw. The eyes are the oval ends, while the jaws are U-shaped with a bolt through each tip.

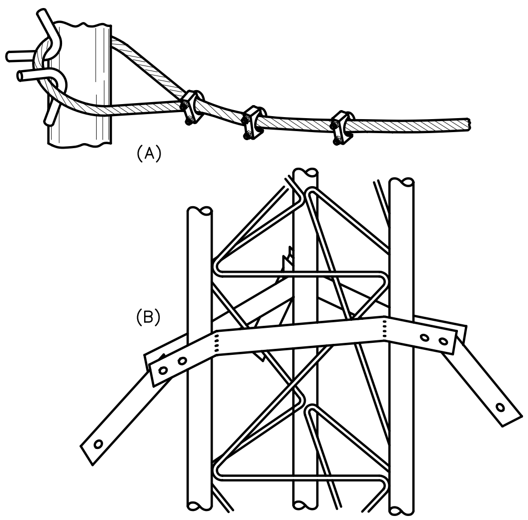

Fig 27—Two methods of attaching guy wires to tower. See text for discussion.

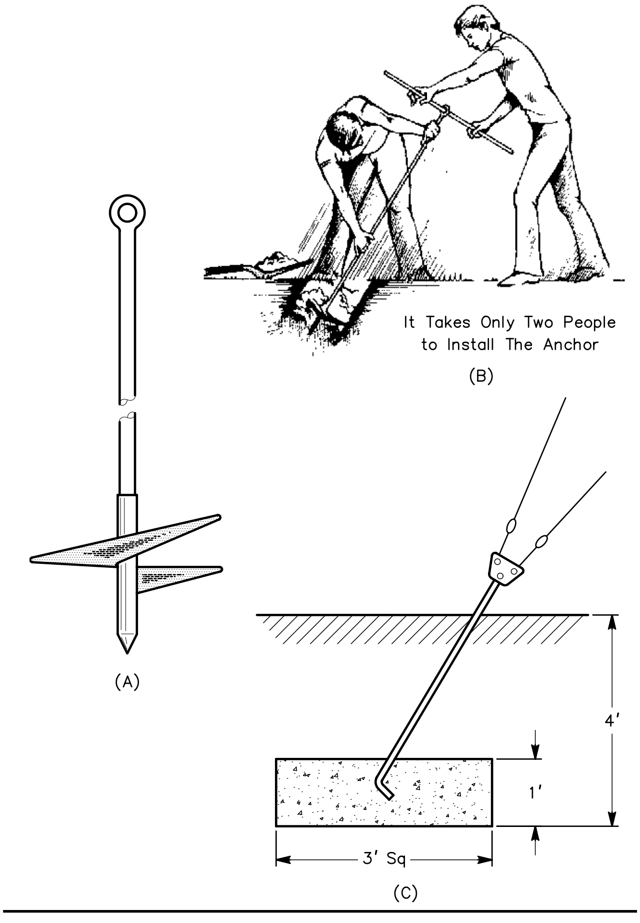

Fig 28—Two standard types of guy anchors. The earth screw shown at A is easy to install and widely available, but may not be suitable for use in certain soils. The concrete anchor is more difficult to install properly, but it is suitable for use with a wide variety of soil conditions and will satisfy most building code requirements.

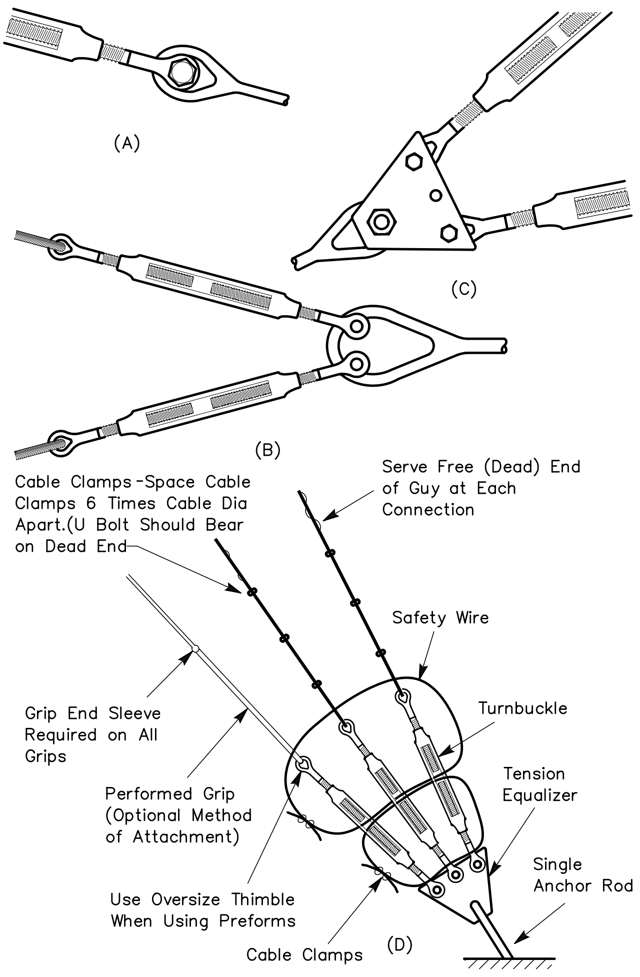
(A)
(C)
(B)
Cable Clamps -Space Cable Clamps 6 Times Cable Dia Apart.(U Bolt Should Bear on Dead End
Serve Free (Dead) End of Guy at Each Connection
Safety Wire
Turnbuckle
Grip End Sleeve Required on All Grips
Performed Grip (Optional Method of Attachment)
Tension Equalizer
Single Anchor Rod
Use Oversize Thimble When Using Preforms
Cable Clamps
(D)

Fig 29—Variety of means available for attaching guy wires and turnbuckles to anchors.

Fig 29B shows two turnbuckles attached to the eye of an anchor. The procedure for installation is to remove the bolt from the jaw, pass the jaw over the eye of the anchor and reinstall the bolt through the jaw, through the eye of the anchor and through the other side of the jaw.

If two or more guy wires are attached to one anchor, *equalizer plates* should be installed (Fig 29C). In addition to providing a convenient point to attach the turnbuckles, the plates pivot slightly to equalize the various guy loads and produce a single load applied to the anchor. Once the installation is complete, a safety wire should be passed through the turnbuckles in a figure-eight fashion to prevent the turnbuckles from turning and getting out of adjustment (Fig 29D).

All guyed towers require the guys to be installed with a certain amount of pre-tension. The tower manufacturer or designer specifies the required pre-tension values, which are usually 10% of the cable breaking strength. Pre-tension is necessary to eliminate looseness in the cable caused by the spiral wire construction and to eliminate excessive dynamic guy and tower motion under wind loading. The recommended method for adjusting the guys is to use a cable tension-measuring device such as the popular *Loos Guy Wire Tensioner*. The guy is gripped with a special clamp, such as the *Klein Cable Grip*, which is connected to the anchor below the eye (or equalizer plate) with a block and tackle arrangement (Fig 23) or a ratcheting come-along. Then the turnbuckle is adjusted to take up the load, the cable grip is released and the final guy tension is adjusted and checked.

When you adjust the guys at each level, you should check the tower for vertical alignment and straightness. This is often done with a transit from two ground points located 90° from each other.

Resonance in Guy Wires

If guy wires are resonant at or near the operating frequency, they can receive and reradiate RF energy. By behaving as parasitic elements, the guy wires may alter and thereby distort the radiation pattern of a nearby antenna. For low frequencies where a dipole or other simple antenna is used, this is generally of little or no consequence. But at the higher frequencies where a unidirectional antenna is installed, it is desirable to avoid pattern distortion if at all possible. The symptoms of re-radiating guy wires are usually a lower front to back ratio and a lower front to side ratio than the antenna is capable of producing. The gain of the antenna and the feed-point impedance will usually not be significantly affected, although sometimes changes in SWR can be noted as the antenna is rotated. (Of course other conductors in the vicinity of the antenna can also produce these same symptoms.)

The amount of re-radiation from a guy wire depends on two factors—its resonant frequency, and the degree of coupling to the antenna. Resonant guy wires near the antenna will have a greater effect on performance than those that are farther away. Therefore, the upper portion of the top level of guy wires should warrant the most attention with horizontally polarized arrays. The lower guy wires are usually closer to horizontal than the top level, but by virtue of their increased distance from the antenna, are not coupled as tightly to the antenna.

To avoid resonance, the guys should be broken up by means of egg or strain insulators. **Fig 30** shows wire lengths that fall within 10% of $^{1}/_{2}$-λ resonance (or a multiple of $^{1}/_{2}$ λ) for all the HF amateur bands. Unfortunately, no single length greater than about 14 feet avoids resonance in all bands. If you operate just a few bands, you can locate greater lengths from Fig 30 that will avoid resonance. For example, if you operate only the 14-, 21- and 24-MHz bands, guy wire lengths of 27 feet or 51 feet would be suitable, along with any length less than 16 feet.

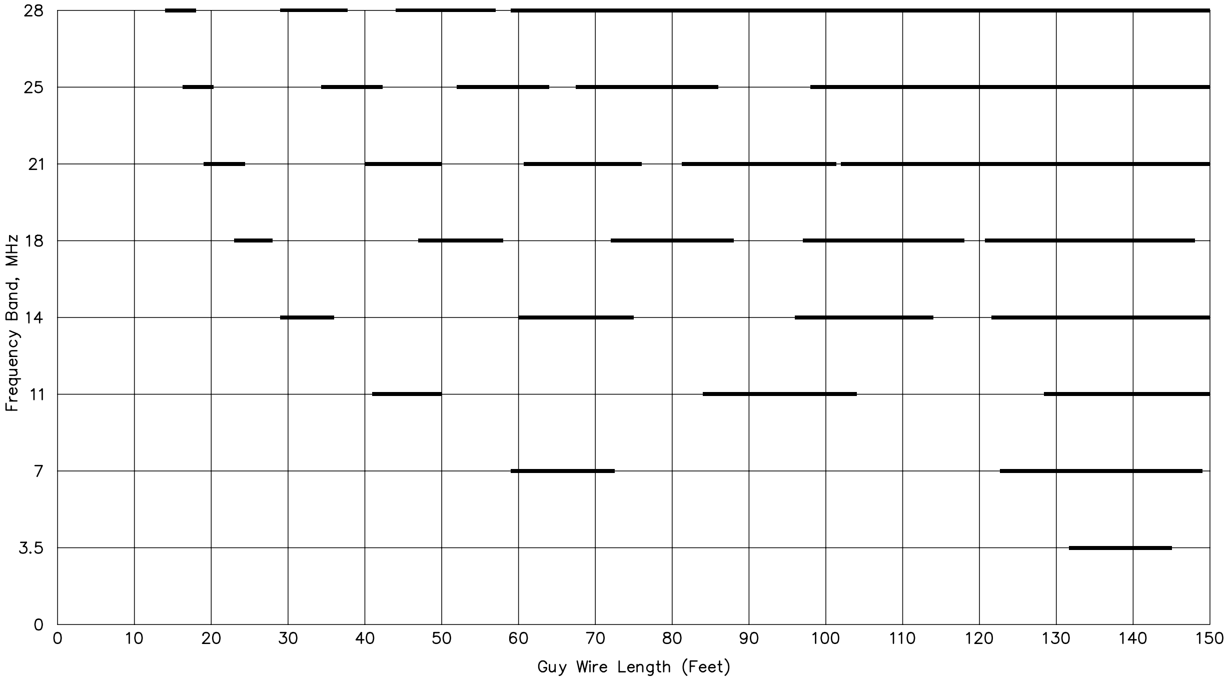

Fig 30—The black bars indicate ungrounded guy wire lengths to avoid for the eight HF amateur bands. This chart is based on resonance within 10% of any frequency in the band. Grounded wires will exhibit resonance at odd multiples of a quarter wavelength. (*By Jerry Hall, K1TD.*)

THE RIGHT TOWER FOR YOUR ANTENNA

Most manufacturers rate their towers in terms of the maximum allowable antenna load that can safely be carried at a specific wind speed. Ensuring that the specific antennas you plan to install meet the tower's design criteria, however, may not always be a straightforward task.

For most towers, the manufacturer assumes that the allowable antenna load is a horizontal force applied at the top of the tower. The allowable load represents a defined amount of exposed antenna area, at a specified wind velocity. Most tower manufacturers rate the load in terms of *Flat Projected Area* (FPA). This is simply the equivalent area of a flat rectangular surface at right angles to the wind. The FPA is not related to the actual shape of the antenna itself, only its rectangular projected area. Some manufacturers provide separate FPAs for antennas made from cylindrical sections and those made from rectangular sections.

In the realm of antenna manufacturers, however, you may encounter another wind load rating called the *Effective Projected Area* (EPA). This attempts to take into account the actual shape of antenna elements. The problem is that there is no agreed-upon standard for the conversion from EPA to load numbers. Different manufacturers may use different conversion factors.

Since most tower manufacturers have provided FPA figures for their towers—allowing us in effect to ignore design-specification details—it would be easiest for us to work only with FPA values for our antennas. This would be fine, if indeed we had good FPA figures for the specific antennas we plan to use! Unfortunately, FPAs are rarely specified for commercially built amateur antennas. Instead, most antenna manufacturers provide effective areas in their specification sheets. You may need to contact the antenna manufacturer directly for the FPA antenna area or for the antenna dimensions so that you can do your own FPA calculations.

Determining Antenna Areas

The method for determining the flat projected area of an antenna is quite simple. We'll use a Yagi antenna as an example. There are two worst-case areas that should be considered here. The first is the FPA of all the elements when the wind blows in the direction along the boom; that is, at right angles to the elements. The second FPA for a Yagi is when the wind is at right angles to the boom. One of these two orientations produces the worst-case exposed antenna area—all other wind angles present lower exposed areas. The idea is to take the highest of the FPAs for these two wind directions and call that the FPA of the antenna structure. See **Fig 31A**.

The element FPA is calculated by multiplying each element's dimension of length by its diameter and then

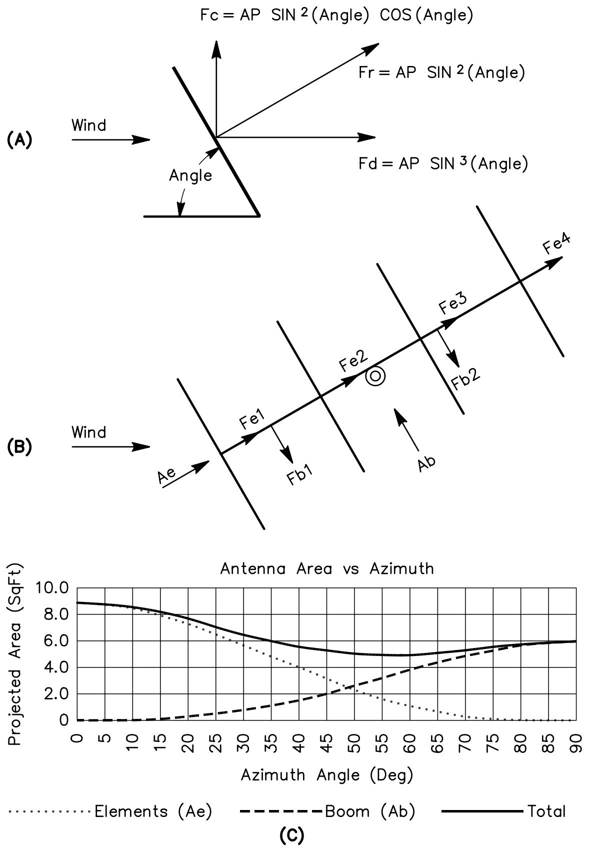

Fig 31—Description of how loads are developed on a Yagi. At A, Fr is the resultant force from the wind load on a generalized member. Fd is the load acting downwind (drag) that creates the load on the tower. Fc is the lateral component of the wind load. The term A is the flat projected area (FPA), which is the broadside area normal to the wind. The term P is the wind pressure. At B, Ae is the total element area, while Ab is the total boom area. All the loads due to the wind act normal to the antenna sections—the force on element #1 (Fe1) acts along the axis of the boom, for example. At C, a plot of the *effective FPA* as a function of the azimuthal wind direction for a Yagi, ignoring drag coefficients. The Yagi in this example has 9.0 square feet of element FPA and 6.0 square feet of boom FPA. The worst-case FPAs occur with the beam pointed in the wind and with the boom broadside to the wind. To determine the actual tower loading, the actual drag coefficients and wind pressures must be used.

summing the FPAs for all elements. The boom's FPA is computed by multiplying the boom's length by its diameter.

The reason for considering two potential peak-load orientations becomes clear when different frequency antennas are stacked on a mast or tower. Some antennas produce peak loads when the elements are broadside to the wind. This is typical of low-frequency Yagis, where the elements are long lengths of aluminum tubing. On the other hand, the boom can dominate the surface area computations in higher-frequency Yagis.

The fundamentals responsible for the need to examine both potential FPAs for Yagis relates to how wind flows over a structure and develops loads. Called *The Cross-Flow Principle*, this was introduced to the communications industry by Dick Weber, K5IU, in 1993. The principle is based on the fact that the loads created by wind flowing across an antenna member only produce forces that are normal to (or perpendicular to) the major axis of the member. The resultant and component load calculations for this method are shown in Fig 31A.

For a Yagi, this means that wind forces on the elements act in-line with the boom, while forces on the boom act in-line with the elements. Fig 31B shows a force diagram for a typical Yagi. Fig 31C shows the FPA for a Yagi rotated through 90° of azimuth.

Antenna Placement on the Mast/Tower

Another important consideration is where the antenna(s) will be placed on the tower. As mentioned before, most generic tower specifications assume that the entire antenna load is applied at the top of the tower. Most amateur installations have a tubular mast extending above the tower top, turned by a rotator mounted down inside the tower. Multiple Yagi antennas are often placed on the mast above the tower top, and you must make sure that both the tower and the mast can withstand the wind forces on the antennas.

For freestanding towers, you can determine how a proposed antenna configuration compares to the tower manufacturer's rating by using an *Equivalent Moment* method. The method computes the bending moment generated at the base of the tower by wind loads on the tower's rated antenna area located right at the top of the tower and compares that to the case when the antenna is mounted on a mast sticking out of the top of the tower.

The exact value of wind pressure is not important, so long as it is the same for both comparisons. The wind load on the tower itself can be ignored because it is the same in both comparisons and the drag coefficients for the antennas can also be ignored if all calculations are performed using flat projected antenna areas, as we've recommended previously.

Keep in mind that this approach does not calculate *actual* loads and moments relevant to any specific tower design standard, but it does allow equivalent comparisons when the wind pressure is constant and all the antenna areas are of the same type. An example is in order.

Fig 32A shows a generic tower configuration, with a concentrated antenna load at the top of the tower. We'll assume that the tower manufacturer rates this tower at 20 square feet of flat projected antenna area. Fig 32B shows a typical amateur installation with a rotating mast and an antenna mounted 7 feet above the top of the tower. To make the calculations easy, we select a wind pressure

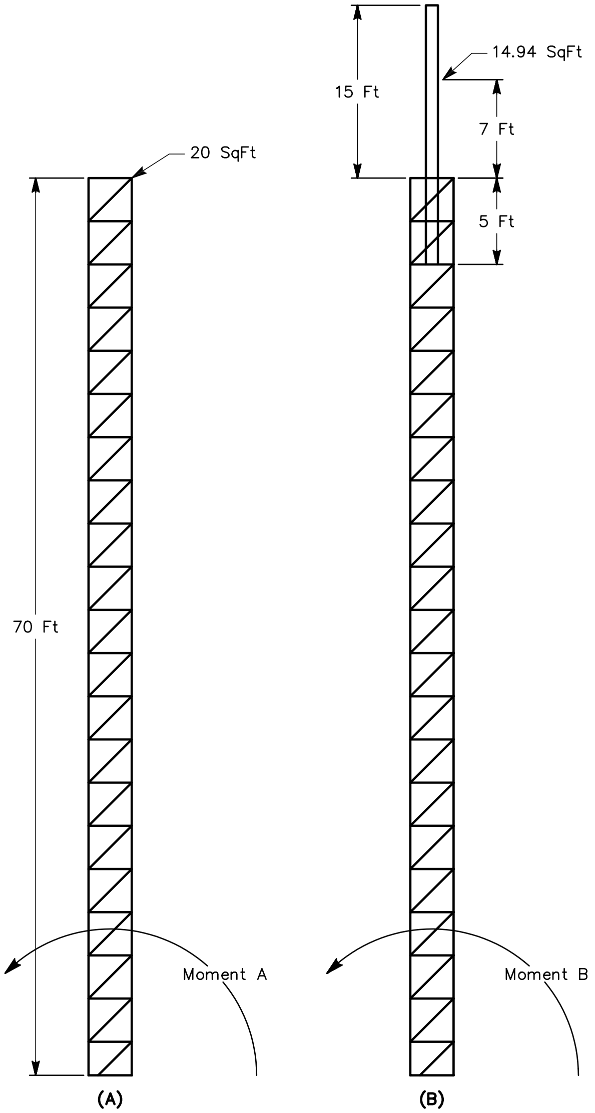

Fig 32—At A, a 70-foot tower rated for 20 square feet of antenna load at the top. At B, the same tower with a 2-inch OD × 20-foot long mast, with an antenna mounted 7 feet above the top of the tower. Both configurations produce the same tower load.

of 1 pound per square foot (1 psf). This makes the tower base moment calculation for Fig 32A:

Antenna load = 20 feet^2 × 1 psf = 20 pounds
Base moment = 70 feet × 20 pounds =
1400 foot-pounds.

This is the target value for the comparison. An equivalent configuration would produce the same base moment. For the configuration in Fig 32B, we assume a tubular 2-inch diameter mast that is 20 feet long, mounted 5 feet down inside the tower. Note that the lattice structure of the tower allows the wind to "see" the whole length of the mast and that we can consider the force distributed along the mast as being a single force concentrated at the mast's center. The flat projected area of the mast by itself, without the antenna, is:

Mast area = 20 feet × 2 inches / 12 inches/foot =
3.33 square feet

The center of the mast is located at a height of 75 feet. Using the same 1-psf-wind load, the base bending moment due to the mast alone is:

Base moment (due to mast) = 3.33 feet^2 × 1 psf ×
75 feet = 249.75 foot-pounds

Including the mast in the configuration reduces the allowable antenna load. The remaining target base moment left for the antenna is found by subtracting the moment due to the mast from the original target value:

New base target moment = 1400 - 249.75 foot-pounds
= 1150.25 foot-pounds.

The antenna in Fig 32B is located at a height of 77 feet. To obtain the allowable antenna area at this elevation we divide the new base target moment by the antenna height, yielding an allowable antenna load of:

1150.25 foot-pounds / 77 feet = 14.94 pounds.

Since we chose a wind load of 1 psf, the allowable antenna FPA has been reduced to 14.94 square feet from 20 square feet. If the projected area of the antenna we are planning to mount in the new configuration is less than or equal to this value, we have satisfied the requirements of the original design. You can use this equivalent-moment method to evaluate different configurations, even ones involving multiple antennas on the mast or situations with additional antennas placed along the tower below the tower top.

For guyed towers, the analyses become much more rigorous to solve. Because the guys and their behaviors are such a significant portion of the tower support mechanism, these designs can become very sensitive to antenna load placements. A general rule of thumb for guyed towers is never to exceed the original tower-top load rating, regardless of distributed loads along its length. Once you redistribute the antenna load placements along a guyed tower, you should do a fresh analysis, just to be sure.

You can run evaluations using the above method for antennas placed on the mast above a guyed tower top. The use of the Equivalent-Moment method for antennas mounted below the top of a guyed tower, however, can become quite suspect, since many generic tower designs have their intermediate guys sized for zero antenna loads lower down the tower. The proper approach in this case is to have a qualified mechanical engineer check the configuration, to see if guy placement and strength is adequate for the additional antennas down the tower.

Mounting the mast and antenna as shown in Fig 32B increases tower loads in the region of the mast. You should investigate these loads to ensure that the tower bracing in that area is sufficient. Now we will consider the problem of bending the rotating mast.

Mast Strength

When you mount antennas on a mast above the tower top, you should examine the bending loads on the mast to ensure that it will be strong enough. This section explains how to perform mast stress calculations for a single sustained wind speed. This procedure does not include height, exposure and gust-response factors found in most tower design standards.

Here are some fundamental formulas and values used to calculate the bending stress in a mast mounted in the top of a tower. The basic formula for wind pressure is:

$$P = .00256\ V^2 \qquad \text{(Eq 1)}$$

where

P is the wind pressure is in pounds per square foot (psf)
V = wind speed in miles per hour (mph)

This assumes an air density for standard temperature and atmospheric pressure at sea level. The wind speed is not the Basic Wind Speed discussed in other sections of this chapter. It is simply a steady state (static) wind velocity.

The formula for calculating the force created by the wind on a structure is:

$$F = P \times A \times C_d \qquad \text{(Eq 2)}$$

where

P = the wind pressure from Eq (1)
A = the flat projected area of the structure (square feet)
C_d = drag coefficient for the shape of the structure's members.

The commonly accepted *drag coefficient* for long cylindrical members like the tubing used for the mast and antenna is 1.20. The coefficient for a flat plate is 2.0.

The formula used to find the *bending stress* in a simple beam like our mast is:

$$\sigma = \frac{M \times c}{I} \qquad \text{(Eq 3)}$$

where

σ = the stress in pounds per square inch (psi)
M = *bending moment* at the base of the mast (inch-pounds)
c = ½ of the mast outside diameter (inches)
I = *moment of inertia* of the mast section (inches^4)

In this equation you must make sure that all values are in the same units. To arrive at the mast stress in pounds per square inch (psi), the other values need to be in inches and pounds also. The equation used to find the moment of inertia for the round tubing mast section is:

$$I = \frac{\pi}{4}(R^4 - r^4) \qquad \text{(Eq 4)}$$

where

I = Moment of Inertia of the section (inches^4)
R = Radius of tube outside diameter (inches)
r = Radius of tube inside diameter (inches)

This value describes the distribution of material about the mast *centroid*, which determines how it behaves under load. The equation used to compute the *bending moment* at the base of the mast (where it is supported by the tower) is:

$$M = (F_M \times L_M) + (F_A \times L_M) \qquad \text{(Eq 5)}$$

where

F_M = wind force from the mast (pounds)
L_M = Distance from tower top to center of mast (inches)
F_A = Wind force from the antenna (pounds)
L_A = Distance from tower top to antenna attachment (inches)

L_M is the distance to the center of the portion of the mast extending above the tower top. Additional antennas can be added to this formula by including their F × L. In the installation shown in Fig 32B, a wind speed of 90 mph, and a mast that is 2 inches OD, with a 0.250-inch wall thickness, the steps for calculating the mast stress are:

1. Calculate the wind pressure for 90 mph, from Eq 1:

$P = .00256\ V^2 = .00256 \times (90)^2 = 20.736$ psf

2. Determine the flat projected area of the mast. The portion of the mast above the tower is 15 feet long and has an outside diameter of 2 inches, which is 2/12 feet.

Mast FPA, A_M = 15 feet × (2 inches / 12 inches/feet) = 2.50 square feet.

3. Calculate the wind load on the mast, from Eq 2:

Mast Force, $F_M = P \times A \times C_d$ = 20.736 psf × 2.50 feet^2 × 1.20 = 62.21 pounds

4. Calculate the wind load on the antenna: From Eq 2

Antenna Force, $F_A = P \times A \times C_d$ = 20.736 psf × 14.94 feet^2 × 1.20 = 371.76 pounds

5. Calculate the mast *Bending Moment*, from Eq 5:

$M = (F_M \times L_M) + (F_A \times L_A)$
= (62.21 pounds × 90 inches) + (371.76 pounds × 84 inches) = 36827 inch-pounds

where L_M = 7.5 feet × 12 inches/foot = 90 inches and L_A = 7.0 feet × 12 inches/foot = 84 inches.

6. Calculate the mast *Moment of Inertia*, from Eq 4:

$$I = \frac{\pi}{4}(R^4 - r^4) = \frac{\pi}{4}(1.0^4 - 0.75^4) = 0.5369 \text{ inches}^4$$

where, for a 2.0-inch OD and 0.250-inch wall thickness tube, R= 1.0 and r = 0.75.

7. Calculate the mast *Bending Stress*, from Eq 3:

$$\sigma = \frac{M \times c}{I} = \frac{36827 \text{ inch-pounds} \times 1.0 \text{ inches}}{0.5369} = 68592 \text{ psi}$$

If the yield strength of the mast material is greater than the calculated bending stress, the mast is considered safe for this configuration and wind speed. If the calculated stress is higher than the mast yield strength, a stronger alloy, or a larger mast, or one with a thicker wall is required.

There are many different materials and manufacturing processes for tubing that may be used for a mast. Yield strengths range from 25,000 psi to nearly 100,000 psi. Knowing the minimum yield strength of the material used for a mast is an important part of determining if it will be safe. Using unknown materials renders efforts from the preceding calculations useless!

When evaluating a mast with multiple antennas attached to it, special care should be given to finding the worst-case condition (wind direction) for the system. What may appear to be the worst load case, by virtue of the combined flat projected antenna areas, may not always be the exposure that creates the largest mast bending moment. Masts with multiple stacked antennas should always be examined to find the exposure that produces the largest mast bending moment. The antenna flat projected areas at 0° and 90° azimuths are particularly useful for this evaluation.

ANTENNA INSTALLATION

All antenna installations are different in some respects. Therefore, thorough planning is the most important first step in installing any antenna. Before anyone climbs the tower, the whole process should be discussed to be sure each crewmember understands what is to be done. And remember that the person on the tower is in charge! Coordinate beforehand what signals and commands are used: "Up" or "Up Slowly" for raising something from the ground; "Down" or "Down Slowly" for the opposite.

"Watch Out!" or 'Watch Out Below!" works for dropped hardware or tools to alert the ground crew below. Remember, once someone is on the tower, no one should be allowed to stand near the base of the tower!

Consider what tools and parts must be assembled and what items must be taken up the tower, and plan alternative actions for possible trouble spots. Extra trips up and down the tower can be avoided by careful planning.

If done properly, the actual work of getting the antenna into position can be executed quite easily with only one person at the top of the tower. The ground crew should do all the heavy work and leave the person on the tower free to guide the antenna into position. Because the ground crew does all the lifting, a large pulley, preferably on a gin pole placed at the top of the tower, is essential. Local radio clubs often have gin poles available for use by their members. Stores that sell tower materials frequently rent gin poles as well.

A gin pole should be placed along the side of the tower so the pulley is no more than 2 feet above the top of the tower (or the point at which the antenna is to be placed). Normally this height is sufficient to allow the antenna to be positioned easily. An important reason that the pulley is placed at this level is that there can be considerable strain on the gin pole when the antenna is pulled away from the tower to maneuver past guy wires.

Sometime the mast to which the antenna will be mounted is used as a place to hang the pulley. You should take care that you don't end up bending the mast by placing the pulley too high on the mast. It may be necessary to back-guy the mast on the opposite side of the tower from which the antenna is raised.

The rope (halyard) through the pulley must be somewhat longer than twice the tower height so that the ground crew can raise the antenna from ground level. The rope should be ½ or $^5/_8$ inch diameter for both strength and ease of handling. Smaller diameter rope is less easily manipulated; it has a tendency to jump out of the pulley track and foul pulley operation.

The first person to climb the tower should carry an end of the halyard so that the gin pole can be lifted and secured to the tower. Those climbing the tower must have safety belts. Belts provide safety and convenience; it is simply impossible to work effectively while hanging onto the tower with one hand.

Once positioned, the gin pole and pulley allow parts and tools to be sent up the tower. A useful trick for sending up small items like bolts and pliers is for a ground crew member to slide them through the rope strands where they are held by the rope for the trip to the top of the tower. Items that might be dislodged by contact with the tower should either be taped or tied to the halyard.
Ever present is the hazard of falling tools or hardware. It is foolish to stand near a tower when someone is working above. Ground crew member should wear hard-hats as extra insurance.

Raising the Antenna Alongside the Tower

A technique that can save much effort in raising the antenna is outlined here. First, the halyard is passed through the gin-pole pulley or the pulley mounted to the mast, and the leading end of the rope is returned to the ground crew, where it is tied to the antenna. The assembled antenna should be placed in a clear area of the yard (or the roof) so the boom points toward the tower. The halyard is then passed *under* the front elements of the beam to a position past the midpoint of the antenna, where it is securely tied to the boom (**Fig 33A**).

Note that once the antenna is installed, the tower worker must be able to reach and untie the halyard from the boom; the rope must be tied less than an arm's length along the boom from the mounting point. If necessary, a large loop may be placed around the first element located beyond the midpoint of the boom, with the knot tied near the center of the antenna. The rope may then be untied easily after completion of the installation. The halyard should be tied temporarily to the boom at the front of the antenna by means of a short piece of light rope or twine.

While the antenna is being raised, the ground crew does all the pulling. As soon as the front of the antenna reaches the top of the mast, the person atop the tower unties the light rope and prevents the front of the antenna from falling, as the ground crew continues to lift the antenna (Fig 33B). When the center of the antenna is even with the top of the tower, the tower worker puts one bolt through the mast and the antenna-mounting bracket on the boom. The single bolt acts as a pivot point and the ground crew continues to lift the back of the antenna with the halyard (Fig 33C). After the antenna is horizontal, the tower worker secures the rest of the mounting bolts and unties the halyard. By using this technique, the tower worker performs no heavy lifting.

Avoiding Guy Wires

Although the same basic methods of installing a Yagi apply to any tower, guyed towers pose a special problem. Steps must be taken to avoid snagging the antenna on the guy wires. With proper precautions, however, even large antennas can be pulled to the top of a tower, even if the

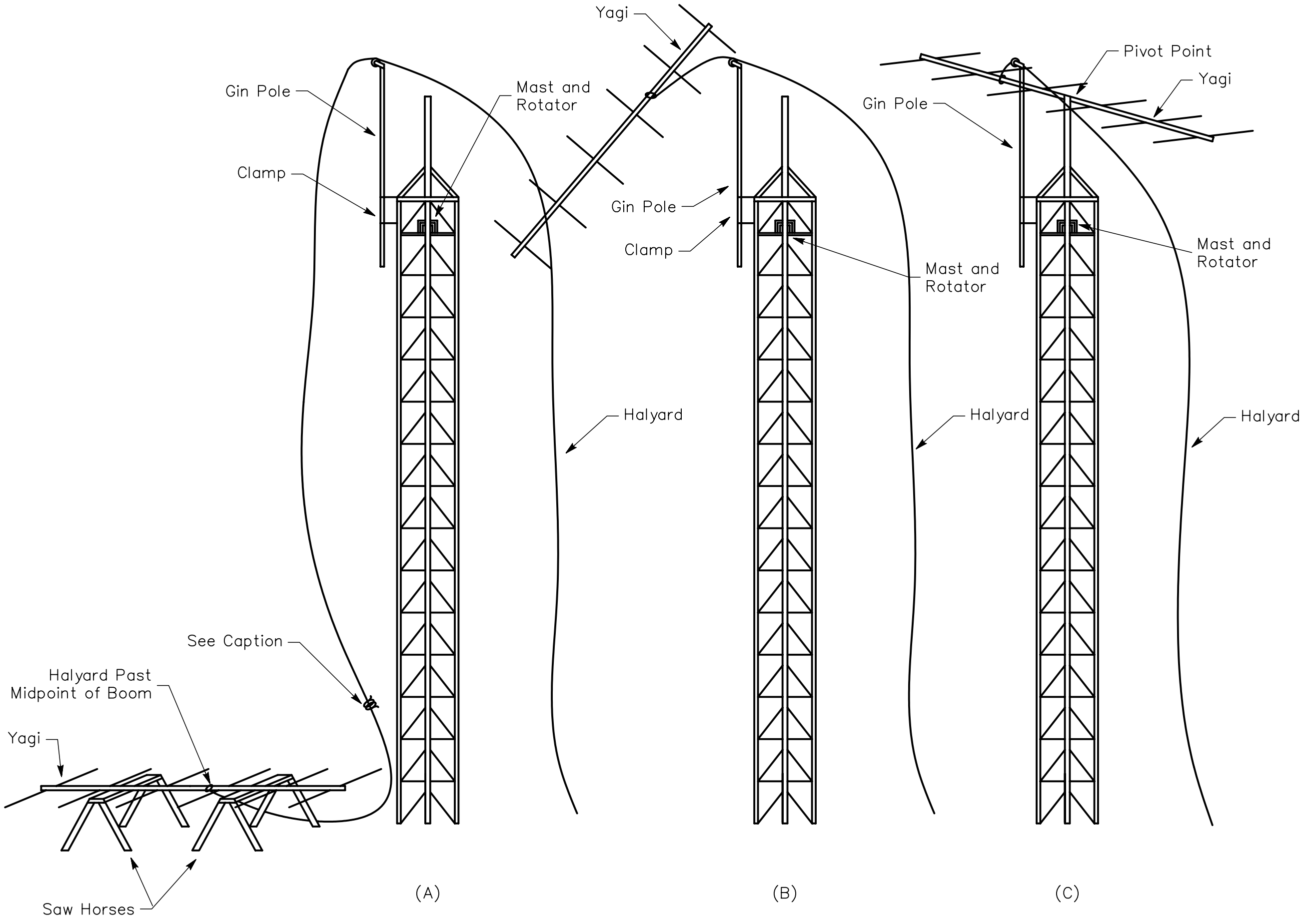

Fig 33—Raising a Yagi antenna alongside the tower. At A the Yagi is placed in a clear area, with the boom pointing toward the tower. The halyard is passed under the elements, then is secured to the boom beyond the midpoint. B shows the antenna approaching the top of the mast. The person on the tower guides it after the lifting rope has been untied from the front of the antenna. At C the antenna is pulled into a horizontal position by the ground crew. The tower worker inserts the pivot bolt and secures it. Note: A short piece of rope is tied around the halyard and the boom at the front of the antenna to stabilize the beam as it is being raised. The tower worker removes it when the boom reaches him at the top of the tower.

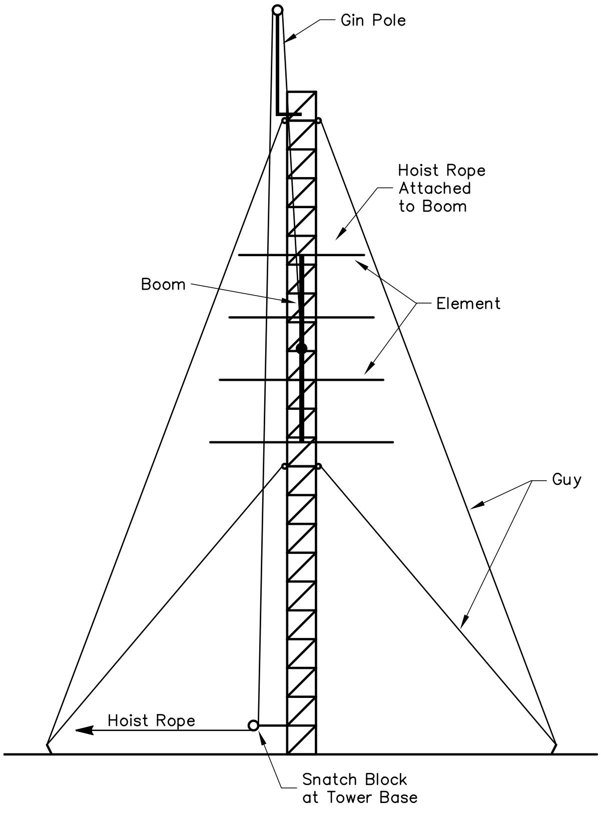
Gin Pole
Hoist Rope
Attached
to Boom
Boom
Element
Guy
Hoist Rope
Snatch Block
at Tower Base

(A)

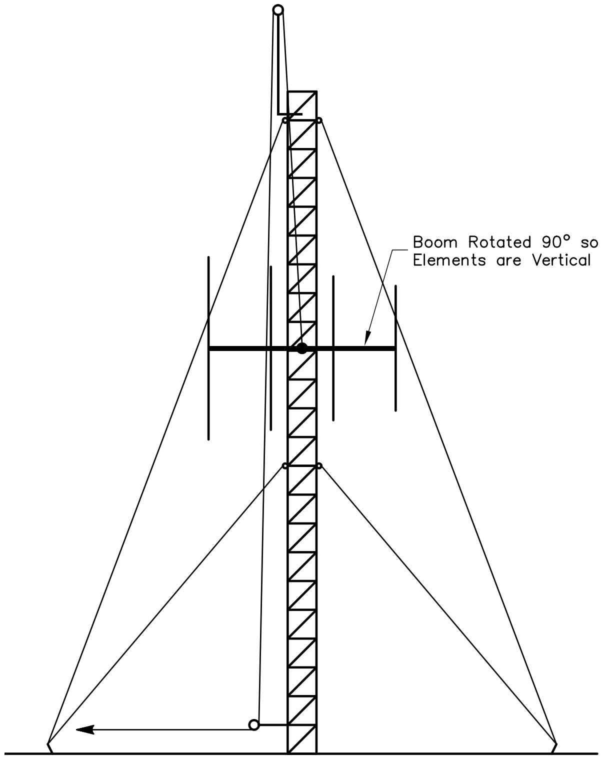
Boom Rotated 90° so
Elements are Vertical

(B)

Fig 34—Building a Yagi partway down the tower. At A, the boom is lashed temporarily to the tower and elements are added, starting at the bottom. At B, the temporary rope securing the boom to the tower is removed and the boom is rotated 90° so that the elements are vertical. At C, the boom is rotated another 90°, "weaving through" guy wires if necessary, until the elements are parallel with the ground, whereupon the boom is secured to the tower.

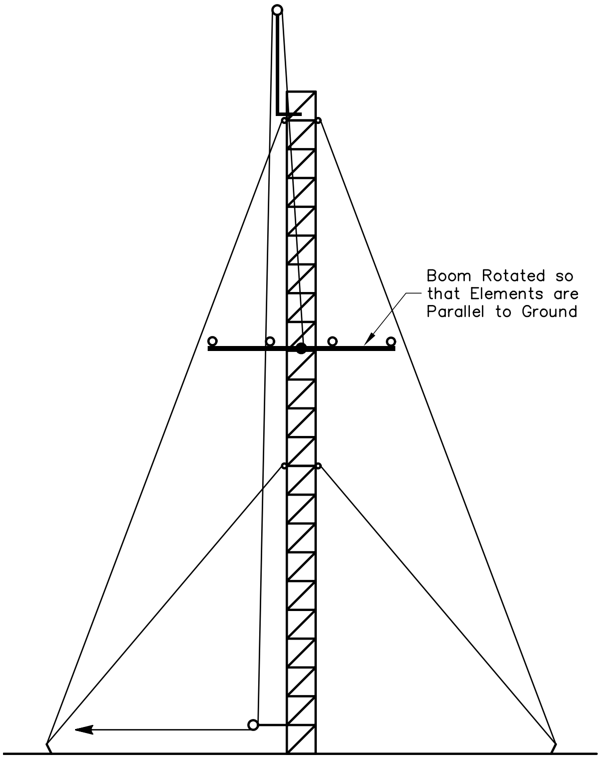
Boom Rotated so
that Elements are
Parallel to Ground

(C)

mast is guyed at several levels.

Sometimes one of the top guys can provide a *track* to support the antenna as it is pulled upward. Insulators in the guys, however, may obstruct the movement of the antenna. A better track made with rope is an alternative. One end of the rope is secured outside the guy anchors. The other end is passed over the top of the tower and back down to an anchor near the first anchor. So arranged, the rope forms a narrow V-track strung outside the guy wires. Once the V-track is secured, the antenna may simply be pulled up, resting on the track.

Another method is to tie a rope to the back of the antenna (but within reach of the center). The ground crews then pull the antenna out away from the guys as the antenna is raised. With this method, some crewmembers are pulling up the antenna to raise it while others are pulling down and out to keep the beam clear of the guys. Obviously, the opposing crews must act in coordination to avoid damaging the antenna. The beam is especially vulnerable when it begins to tip into the horizontal position. If the crew continues to pull out and down against the antenna, the boom can be broken. Another problem with this approach is that the antenna may rotate on the axis of the boom as it is raised. To prevent such rotation, long lengths of twine may be tied to outer elements, one piece on each side of the boom. Ground personnel may then use these *tag lines* to stabilize the antenna. Where this is done, provisions must be made for untying the twine once the antenna is in place.

A third method is to tie the halyard to the center of the antenna. A crewmember, wearing a safety belt, walks the antenna up the tower as the crew on the ground raises it. Because the halyard is tied at the balance point, the tower worker can rotate the elements around the guys. A tag line can be tied to the bottom end of the boom so that a ground worker can help move the antenna around the guys. The tag line must be removed while the antenna is still vertical.

A fourth method is to build the antenna on the tower and then swing it into position. (See also the section below on the PVC Mount.) Building the Yagi on the tower works particularly well for Yagis mounted partway up the tower, as you might do in a stacked array. The technique works best when the vertical spacing between the guys is greater than the length of the Yagi boom.

Fig 34 illustrates the steps involved. A pull rope through a gin-pole or tower-mounted pulley is secured to the boom at the final balance point and the ground crew raises the boom in a vertical position up the tower. A tie rope is used to temporarily secure the upper end of the boom to keep it stable while the boom is being raised. The tower person removes the tie-rope once the boom is raised to the right level and has been temporarily secured to the tower.

The elements are then brought up one at a time and mounted to the boom. It helps if you have a 2- or 3-foot long *spotting mast* temporarily attached to the boom to form a 90° frame of reference. This allows the ground crew to spot from below so that the elements are all lined up in the same plane. After all the elements are mounted and aligned properly, the temporary rope securing the boom to the tower is released, suspending the antenna on the pull rope. The tower person then rotates the boom 90° so that the elements are vertical. Next the elements are rotated 90° into the tower so that they are parallel to the ground. The ground crew then moves the boom up or down using the pull rope to the final point where it is mounted to the tower.

A modification of this technique also works for building a medium-sized Yagi on the top of the tower. This technique will work if the length of the gin pole at maximum safe extension is long enough. See **Fig 35**.

As usual, the gin-pole pull rope is attached to the balance point of the boom and the boom is pulled up the tower in the vertical position, using a rope to temporarily tie the pull rope to the top end of the boom for stability. The boom is temporarily secured to the tower with rope in the vertical position so that the top end is just higher than the top of the tower. In order to clear the gin pole when the elements are mounted and the boom is raised higher to mount the next element, you must tilt the boom slightly so that the element mounted to the top end of the boom will be *behind* the mast. This is very important!

The elements are first mounted to the bottom side of the boom to provide weight down below for stability. Then the top-most element is mounted to the boom. The tower person removes the temporary rope securing the boom to the tower and the ground crew uses the pull rope

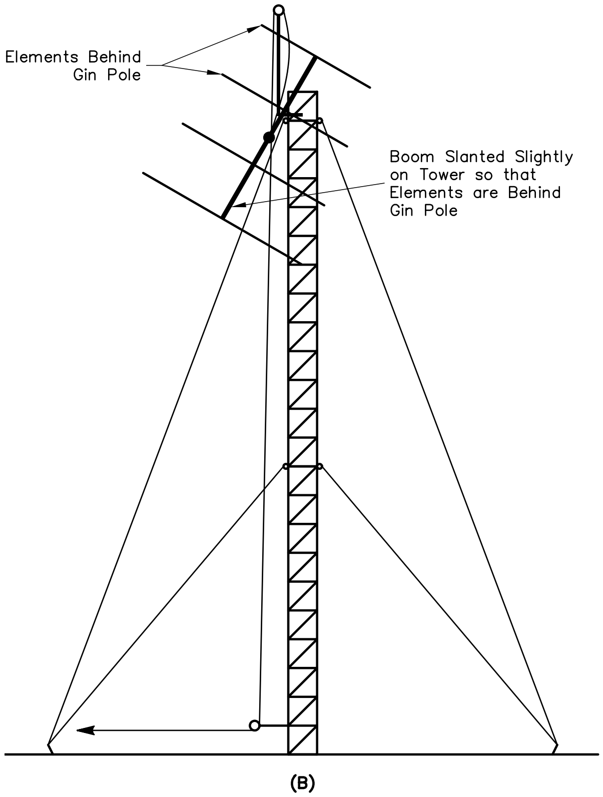
Elements Behind Gin Pole
Boom Slanted Slightly on Tower so that Elements are Behind Gin Pole
(B)

Fig 35—Building a Yagi at the top of the tower. The length of the gin pole must be longer than $^1/_2$ the boom so that the boom can be hoisted upwards to the place where it is mounted to the mast. Usually the boom is initially lashed to the tower slanted slightly from vertical so that the top element ends up *behind* the gin pole. The elements are mounted at the bottom end of the boom first to provide stability. Then the element at the top of the boom is mounted and the boom is moved upwards using the gin-pole hoist rope so that the next-to-top element may be mounted, again behind the gin pole. This process is repeated until all elements are mounted (save possibly the middle element if it can be reached easily from the tower once the beam has been mounted to the mast). Then the boom is tilted to the final position, weaving the elements to clear guy wires if necessary.

to move the mast vertically upwards to the point where the next element from the top can be mounted. Once all the elements are mounted and aligned in the same plane (with perhaps the center element closest to the mast-to-boom bracket left on the ground until later), the temporary securing rope is removed. The boom is now swung so that the elements can be maneuvered to clear the top guy wires. Once the elements are horizontal the boom is secured to the mast and the center element is mounted.

Using a Tram

Another method to get a large Yagi to the top of the tower safely is a *tram*. A tram supports the antenna *under* the tram wire, using a pulley riding on the tram wire. The antenna can thus move more freely—without the friction it would have riding on top of a track rope, as described previously. This puts considerably less strain on the tram wire itself and on the mast to which it is tied on the tower. Some installers prefer to use a wire-rope tramline for its reduced sag.

The tram method uses an easily constructed fixture mounted to the boom of the Yagi to stabilize it from rotating away from the desired attitude as the antenna is raised. A guy-wire cable or heavy rope is fixed to the mast about two feet above the point where the antenna will mount to the mast. A come-along is often used at the ground end to tension the tram wire properly. It is often necessary to back-guy the mast to make sure it doesn't get bent, since the horizontal forces acting on the mast can be considerable in any tram (or track) operation. (Note that the tram technique works well for side-mounted antennas also, where back-guying is not necessary if you are reasonably close to a guy set, as is usually the case.)

Fig 36 is a photograph of a tram fixture built by Kurt Andress, K7NV. This consists of a pulley riding on top of the tramline. This pulley is attached using a caribiner or shackle to two equal-length wires connected to the boom to make an inverted-V shaped *sling*. The two sling wires are secured to the boom using angle irons and muffler clamps so that the antenna is perfectly balanced in the horizontal plane. Balance is very important to make sure the antenna rises properly on the tram without having the boom rotate downward on one end or the other.

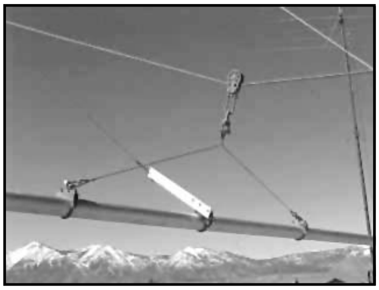

Fig 36—Photo of the tram system used by Kurt Andress, K7NV. (*Photo by K7NV.*)

The hoisting rope running through the tower-mounted gin-pole pulley (and used by the ground crew to pull the antenna up to the tower on the tram line) is attached to a 2-foot piece of angle iron. This is attached to the boom with a muffler clamp. Note that the angle iron is rotated slightly from horizontal so that the plane of the elements is tilted upwards—this allows the elements to clear the guy wires as the antenna is raised. (While the antenna is close to the ground, the angle iron is adjusted so that the elements remain horizontal. Once they are clear of the ground, the angle iron is readjusted to align the elements in the proper direction to clear the guy wires on the tower.) The force of the pull rope along the angle iron also stabilizes the antenna from yawing from side-to-side. Note in Fig 36 that the angle iron is mounted just off-center from where the boom-to-mast plate will be attached so that it clears the mast as the antenna nears the top of the tower. **Fig 37** diagrams how the tram and hoisting line are rigged to the mast.

The tower person directs the activity of the ground crew below and guides the antenna to the mast. Once the end of the pull rope reaches the mast, the tower man ties the boom temporarily to the mast so that he can undo the pull rope from the tram-fixture angle iron and retie it around the boom. Then the antenna can be raised to the point where it can be mounted to the mast.

This technique has been employed to raise Yagis with booms as long as 42 feet at the N6RO contest station on towers as high as 130 feet. As with the track, the tram system requires a good deal of open real estate. While it sounds complicated to set up, you can raise some rather large antennas in less than an hour, once you get the hang of the operation.

THE PVRC MOUNT

The methods described above for hoisting antennas are sometimes not satisfactory for really large, heavy arrays. The best way to handle large Yagis is to assemble them on top of the tower. One way to do this easily is by using the *PVRC Mount*. Many members of the Potomac Valley Radio Club have successfully used this method to install large antennas. Simple and ingenious, the idea involves offsetting the boom from the mast to permit the boom to tilt 360° and rotate axially 360°. This permits the entire length of the boom to be brought alongside the tower, allowing the elements to be attached one by one. (It also allows any part of the antenna to be brought alongside the tower for antenna maintenance.)

See **Figs 38** through **42**. The mount itself consists of a short length of pipe of the same diameter as the rotating mast (or greater), a steel plate, eight U bolts and four pinning bolts. The steel plate is the larger, horizontal one shown in Fig 38. Four U bolts attach the plate to the rotating mast, and four attach the horizontal pipe to the plate. The horizontal pipe provides the offset between the antenna boom and the tower. The antenna boom-to-mast plate is mounted at the outer end of the short pipe. Four bolts are used to ensure that the antenna ends up parallel to the ground, two pinning each plate to the short pipe. When the mast plate pinning bolts are removed and the

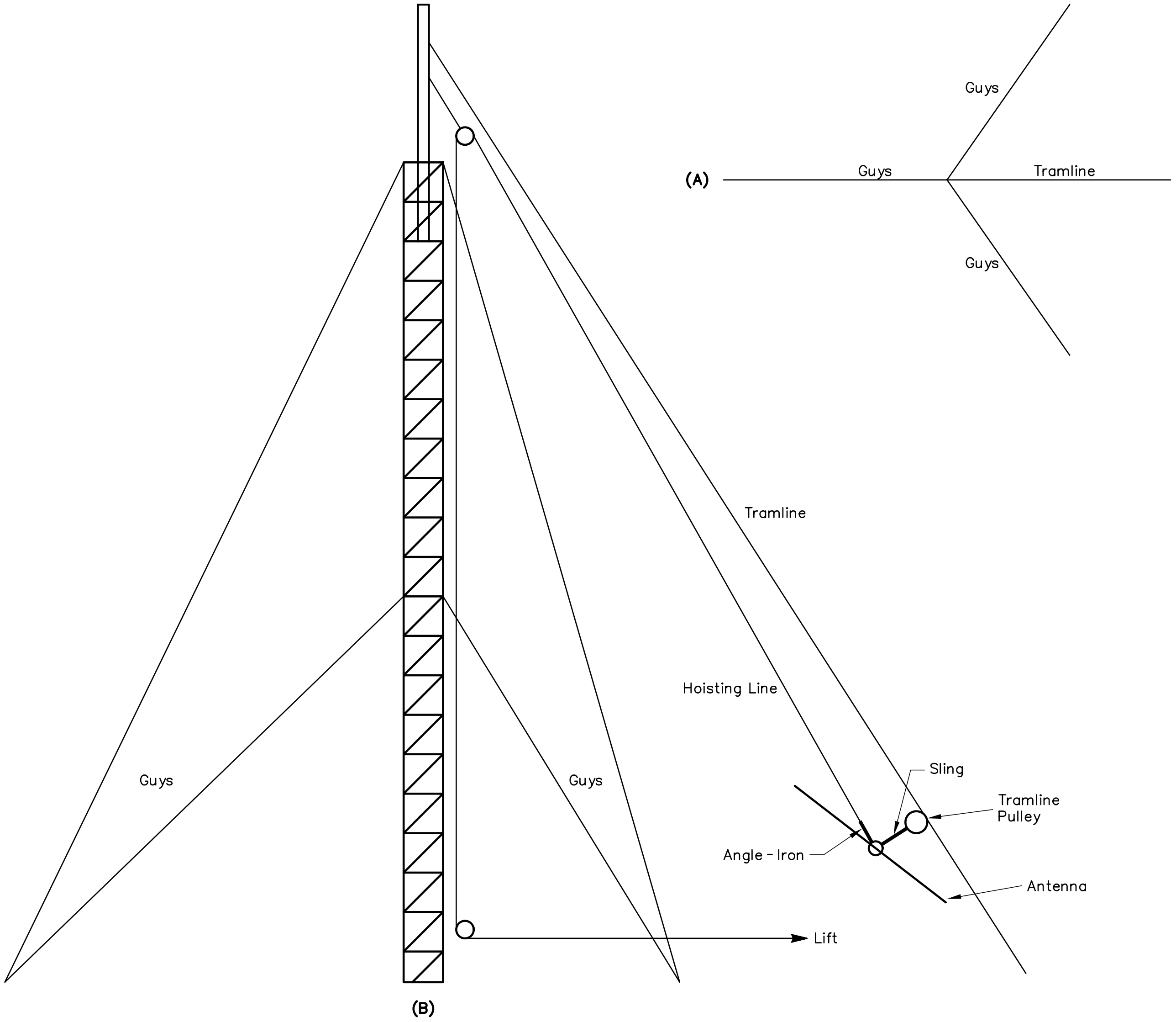

Fig 37—At A, bird's-eye view of tram system used to bring large Yagi antennas from the ground to the top of the tower. At B, side view of rigging used for tramline and hoisting line, along with the sling and tram fixture used to hold the Yagi on the tramline.

four U bolts loosened, the short pipe and boom plate can be rotated through 360°, allowing either half of the boom to come alongside the tower.

First assemble the antenna on the ground. Carefully mark all critical dimensions, and then remove the antenna elements from the boom. Once the rotator and mast have been installed on the tower, a gin pole is used to bring the mast plate and short pipe to the top of the tower. There, the top crew unpins the horizontal pipe and tilts the antenna boom plate to place it in the vertical plane. The boom is attached to the boom plate at the final balance point of the assembled antenna. It is important that the boom be rotated axially so the bottom side of the boom is closest to the tower. This will allow the boom to be tilted without the elements striking the tower.

During installation it may be necessary to loosen one guy wire temporarily to allow for tilting of the boom. As a safety precaution, a temporary guy should be attached to the same leg of the tower just low enough so the assembled antenna will clear it.

The elements are assembled on the boom, starting with those closest to the center of the boom, working out alternately to the farthest director and reflector. This procedure must be followed. If all the elements are put first on one half of the boom, it will be dangerous (if not impossible) to put on the remaining elements. By start-

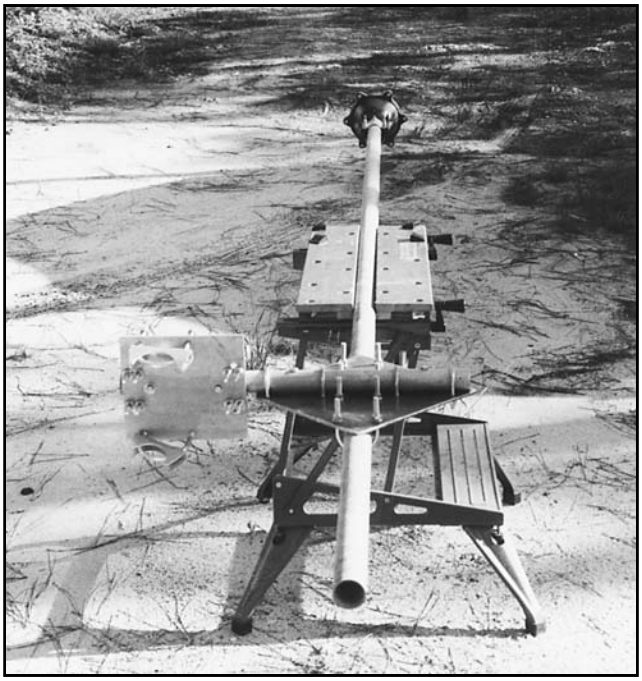

Fig 38—The PVRC mount, boom plate, mast and rotator ready to go. The mast and rotator are installed on the tower first.

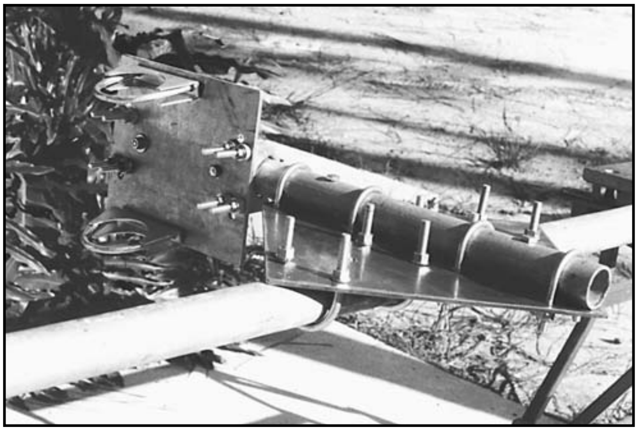

Fig 39—Close-up of the PVRC mount. The long pipe (horizontal in this photo) is the rotating mast. The U bolts in the vertical plate at the left are ready to accept the antenna boom. The heads of two locking pins (bolts) are visible at the midline of the boom plate. The other two pins help secure the horizontal pipe to the large steel mast plate. (The head of the bolt nearest the camera blends in with the right hand leg of the U bolt behind it.)

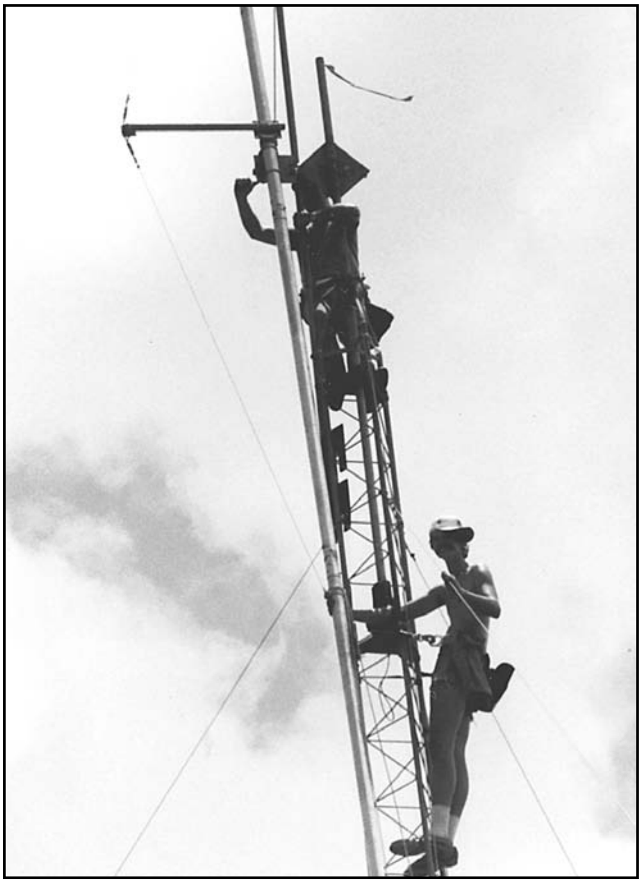

Fig 40—Working at the 70-foot level. A gin pole makes pulling up and mounting the boom to the boom plate a safe and easy procedure.

ing at the middle and working outward, the balance point of the partly assembled antenna will never be so far removed from the tower that tilting of the boom becomes impossible.

When the last element is attached, the boom is brought parallel to the ground, the horizontal pipe is pinned to the mast plate, and the mast plate U bolts tightened. At this point, all the antenna elements will be positioned vertically. Next, loosen the U bolts that hold the boom and rotate the boom axially 90°, bringing the elements parallel to the ground. Tighten the boom bolts and double check all the hardware.

Many long-boom Yagis employ a truss to prevent boom sag. With the PVRC mount, the truss must be attached to a pipe that is independent of the rotating mast. A short length of pipe is attached to the boom as close as possible to the balance point. The truss then moves with the boom whenever the boom is tilted or twisted.

A precaution: Unless you have a really strong rotator, you should consider using this mount mainly for assembling the antenna on the tower. The offset between the boom and the mast with this assembly can generate high torque loads on the rotator. Mounting the boom as close as possible to the mast will minimize the torque when the antenna is pointed into the wind.

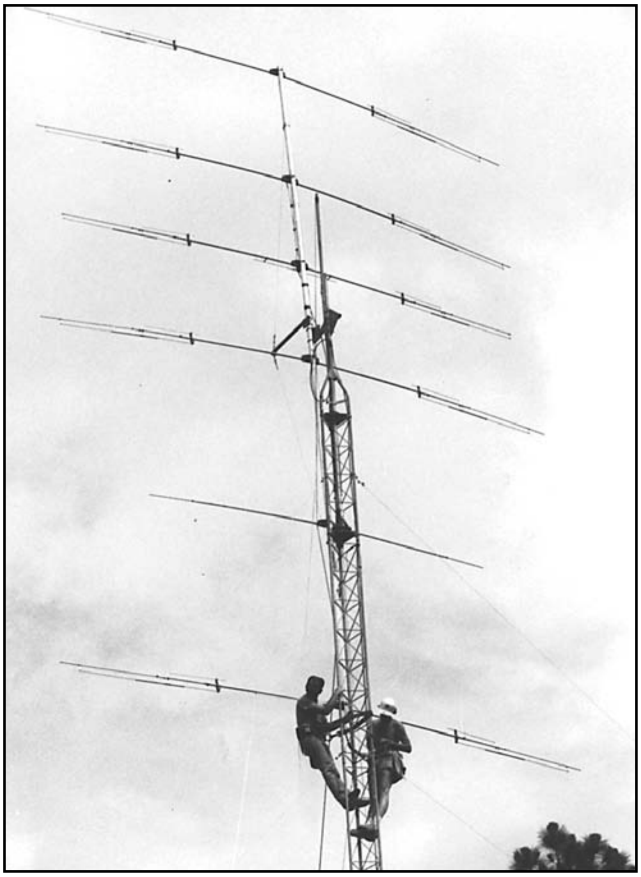

Fig 41—Mounting the last element prior to positioning the boom in a horizontal plane.

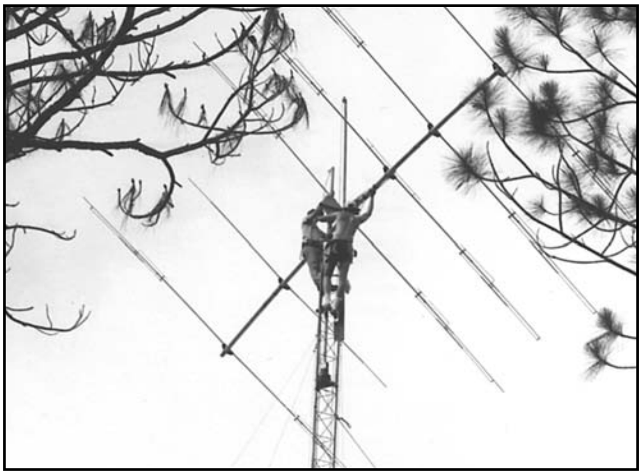

Fig 42—The U bolts securing the short pipe to the mast plate are loosened and the boom is turned to a horizontal position. This puts the elements in a vertical plane. Then the pipe U bolts are tightened and pinning bolts secured. The boom U bolts are then loosened and the boom turned axially 90°.

THE TOWER ALTERNATIVE

A cost saving alternative to the ground-mounted tower is the roof-mounted tripod. Units suitable for small HF or VHF antennas are commercially available. Perhaps the biggest problem with a tripod is determining how to fasten it securely to the roof.

One method of mounting a tripod on a roof is to nail 2 × 6 boards to the undersides of the rafters. Bolts can be extended from the leg mounts through the roof and the 2 × 6s. To avoid exerting too much pressure on the area of the roof between rafters, place another set of 2 × 6s on top of the roof (a mirror image of the ones in the attic). Installation details are shown in **Figs 43** through **46**.

The 2 × 6s are cut 4 inches longer than the outside distance between two rafters. Bolts are cut from a length of ¼-inch-threaded rod. Nails are used to hold the boards in place during installation, and roofing tar is used to seal the area to prevent leaks.

Find a location on the roof that will allow the antenna to turn without obstruction from such things as trees, TV antennas and chimneys. Determine the rafter locations. (Chimneys and vent pipes make good reference points.) Now the tower is set in place atop three 2 × 6s. A plumb line run from the top center of the tower can be used to center it on the peak of the roof. Holes for the mounting bolts can now be drilled through the roof.

Before proceeding, the bottom of the 2 × 6s and the area of the roof under them should be given a coat of roofing tar. Leave about $^1/_8$ inch of clear area around the holes to ensure easy passage of the bolts. Put the tower back in place and insert the bolts and tighten them. Apply tar to the bottom of the legs and the wooden supports, including the bolts. For added security the tripod can be guyed. Guys should be anchored to the frame of the house.

If a rotator is to be mounted above the tripod, pressure will be applied to the bearings. Wind load on the antenna will be translated into a “pinching” of one side of the bearings. Make sure that the rotator is capable of handling this additional stress.

ROTATOR SYSTEMS

There are not that many choices when it comes to antenna rotators for the amateur antenna system. Making the correct decision as to how much capacity the rotator must have is very important to ensure trouble-free operation. Manufacturers generally provide an antenna surface-area rating to help the purchase choose a suitable rotator. The maximum antenna area is linked to the rotator’s torque capability.

Some rotator manufacturers provide additional information to help you select the right size of rotator for the antennas you plan to use. Hy-Gain provides an *Effective Moment* value. Yaesu calls theirs a *K-Factor*. Both of these ratings are torque values in foot-pounds. You can compute the effective moment of your antenna by multiplying the antenna turning radius by its weight. So long

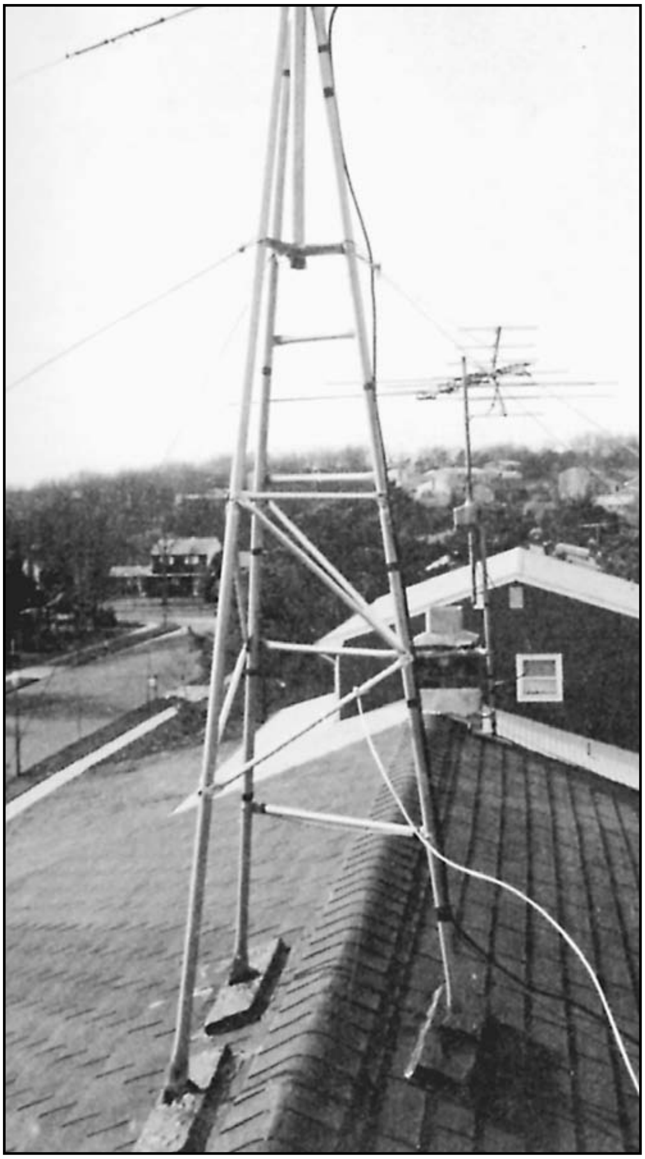

Fig 43—This tripod tower supports a rotary beam antenna. In addition to saving yard space, a roof-mounted tower can be more economical than a ground-mounted tower. A ground lead fastened to the lower part of the frame is for lightning protection. The rotator control cable and the coaxial line are dressed along two of the legs. (*Photo courtesy of Jane Wolfert.*)

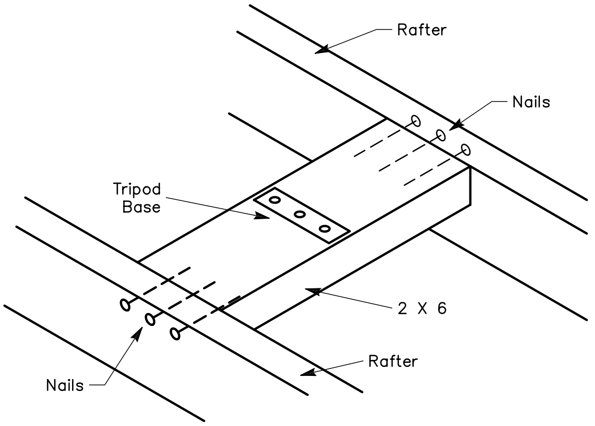

Fig 44—This cutaway view illustrates how the tripod tower is secured to the roof rafters. The leg to be secured to the crosspiece is placed on the outside of the roof. Another cross member is fastened to the underside of the rafters. Bolts, inserted through the roof and the two cross pieces, hold the inner cross member in place because of pressure applied. The inner crosspiece can be nailed to the rafter for added strength.

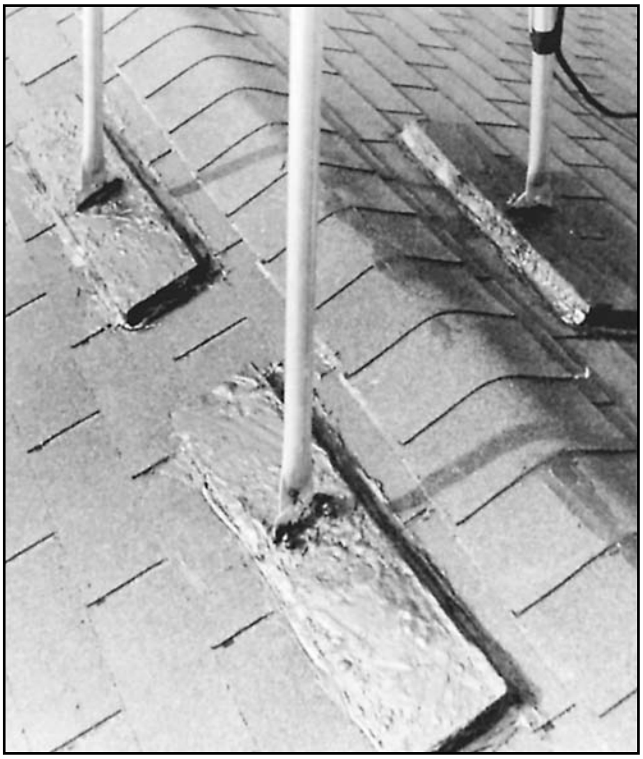

Fig 45—Three lengths of 2 × 6 wood mounted on the outside of the roof and reinforced under the roof by three identical lengths provide a durable means for anchoring the tripod. A thick coat of roofing tar guards against weathering and leaks.

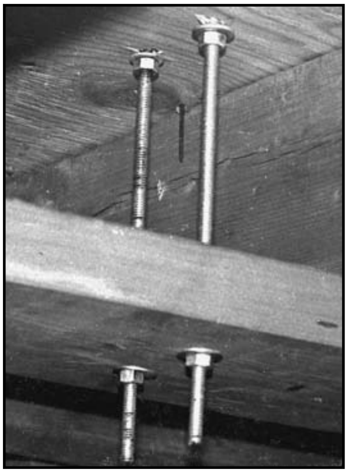

Fig 46—The strengthened anchoring for the tripod. Bolts are placed through two 2 × 6s on the underside of the roof and through the 2 × 6 on the top of the roof, as shown in Fig 43.

as the effective moment rating of the rotator is greater than or equal to the antenna value, the rotator can be expected to provide a useful service life.

There are basically four grades of rotators available to the amateur. The lightest-duty rotator is the type typically used to turn TV antennas. Without much difficulty, these rotators will handle a small 3-element tribander array (14, 21 and 28 MHz) or a single 21- or 28-MHz monoband three-element antenna. The important consideration with a TV rotator is that it lacks braking or holding capability. High winds turn the rotator motor via the gear train in a reverse fashion. Broken gears sometimes result.

The next grade up from the TV class of rotator usually includes a braking arrangement, whereby the antenna is held in place when power is not applied to the rotator. Generally speaking, the brake prevents gear damage on windy days. If adequate precautions are taken, this group of rotators is capable of holding and turning stacked monoband arrays, or up to a five-element 14-MHz system. The next step up in rotator strength is more expensive. This class of rotator will turn just about anything the most demanding amateur might want to install.

A description of antenna rotators would not be complete without the mention of the *prop pitch* class. The prop pitch rotator system consists of a surplus aircraft propeller blade pitch motor coupled to an indicator system and a power supply. There are mechanical problems of installation, however, resulting mostly from the size and weight of these motors. It has been said that a prop pitch rotator system, properly installed, is capable of turning a house. Perhaps in the same class as the prop pitch motor (but with somewhat less capability) is the electric motor of the type used for opening garage doors. These have been used successfully in turning large arrays.

Proper installation of the antenna rotator can provide many years of trouble-free service; sloppy installation can cause problems such as a burned out motor, slippage, binding and casting breakage. Most rotators are capable of accepting mast sizes of different diameters, and suitable precautions must be taken to shim an undersized mast to ensure dead-center rotation. It is very desirable to mount the rotator inside and as far below the top of the tower as possible. The mast absorbs the torsion developed by the antenna during high winds, as well as during starting and stopping.

Some amateurs have used a long mast from the top to the base of the tower. Rotator installation and service can be accomplished at ground level. A mast length of 10 feet or more between the rotator and the antenna will add greatly to the longevity of the entire system by allowing the mast to act as a torsion shock absorber. Another benefit of mounting the rotator 10 feet or more below the antenna is that any misalignment among the rotator, mast and the top of the tower is less significant. A tube at the top of the tower (a sleeve bearing) through which the mast protrudes almost completely eliminates any lateral forces on the rotator casing. All the rotator must do is support the downward weight of the antenna system and turn the array.

While the normal weight of the antenna and the mast is usually not more than a couple of hundred pounds, even with a large system, one can ease this strain on the rotator by installing a thrust bearing at the top of the tower. The bearing is then the component that holds the weight of the antenna system, and the rotator need perform only the rotating task.

Don't forget to provide a loop of coax to allow your beam to rotate properly. Make sure you position the rotator loop so that it doesn't snag on anything. **Fig 47** shows a rotator loop for an elliptically polarized UHF Yagi array. Note that the coax loop is taped to the rotating mast above the top of the tower and to the tower itself.

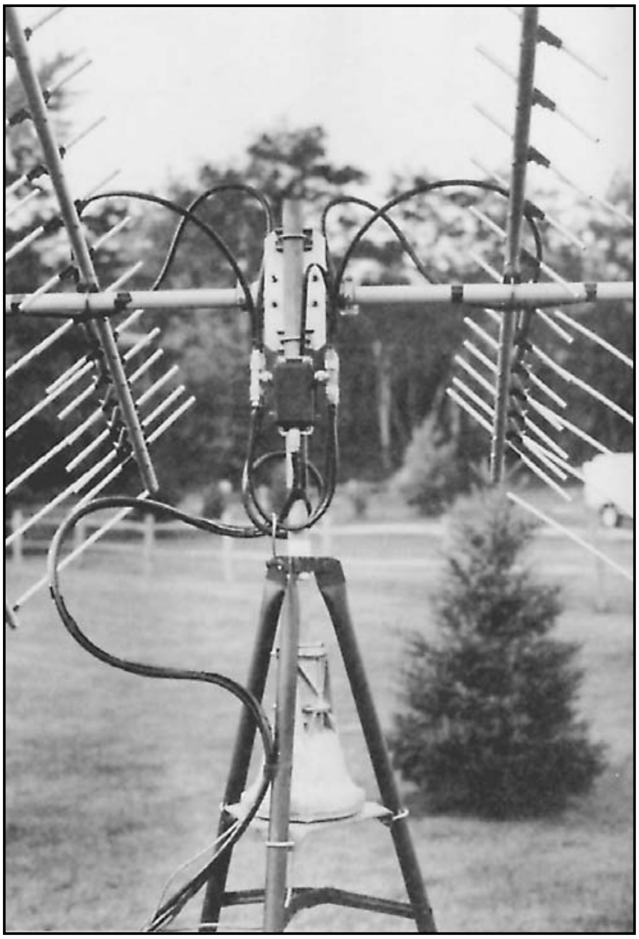

Fig 47—Rotator loop for UHF Yagi array. The feed coax is bundled with a control cable for the polarization relay. The coax/control-cable loop is taped to the rotating mast and to the top of the tower with vinyl electric tape to allow the array to be rotated.

Indicator Alignment

A problem often encountered in amateur installations is that of misalignment between the direction indicator in the rotator control box and the heading of the antenna. With a light duty rotator, this happens frequently when the wind blows the antenna to a different heading. With no brake, the force of the wind can move the gear train and motor of the rotator, while the indicator remains fixed. Such rotator systems have a mechanical stop to prevent continuous rotation during operation, and provision is usually included to realign the indicator against the mechanical stop from inside the shack. During installation, the antenna must be oriented correctly for the mechanical stop position, which is usually north.

In larger rotator systems with an adequate brake, indicator misalignment is caused by mechanical slippage in the antenna boom-to-mast hardware. Many texts suggest that the boom be pinned to the mast with a heavy-duty bolt and the rotator be similarly pinned to the mast. There is a trade-off here. If there is sufficient wind to cause slippage in the couplings without pins, with pins the wind could break a rotator casting. The slippage will act as a clutch release, which may prevent serious damage to the rotator. On the other hand, the amateur might not like to climb the tower and realign the system after each heavy windstorm.

BIBLIOGRAPHY

Source material and more extended discussions of the topics covered in this chapter can be found in the references listed below and in the texts listed at the end of Chapter 2.

L. H. Abraham, “Guys for Guys Who Have To Guy,” *QST*, Jun 1955.

K. Baker, “A Ladder Mast,” *QST*, Jun 1981, p 24.

W. R. Gary, “Toward Safer Antenna Installations,” *QST*, Jan 1980, p 56.

S. F. Hoerner, “Fluid Dynamic Drag,” published by author, Bricktown, New Jersey, 1965, pp 1-10.

C. L. Hutchinson, “A Tree-Mounted 30-Meter Ground-Plane Antenna,” *QST*, Sep 1984, pp 16-18.

C. L. Hutchinson, R. D. Straw, *Simple and Fun Antennas for Hams* (Newington: ARRL, 2002).

M. P. Keown and L. L. Lamb, “A Simple Technique for Tower-Section Separation,” *QST*, Sep 1979, pp 37-38.

P. O’Dell, “The Ups and Downs of Towers,” *QST*, Jul 1981, p 35.

S. Phillabaum, “Installation Techniques for Medium and Large Yagis,” *QST*, Jun 1979, p 23.

C. J. Richards, “Mechanical Engineering in Radar and Communications” (London: Van Nostrand Reinhold Co., 1969) pp 162-165.

D. Weber “Determination of Yagi Wind Loads Using the Cross-Flow Principle,” *Communications Quarterly*, Spring 1993.

A. B. White, “A Delayed Brake Release for the Ham-II,” *QST*, Aug 1977, p 14.

B. White, E. White and J. White, “Assembling Big Antennas on Fixed Towers,” *QST*, Mar 1982, pp 28-29

L. Wolfert, “The Tower Alternative,” *QST*, Nov 1980, p 36.

W. C. Young., “Roark’s Formulas for Stress & Strain” (New York: McGraw-Hill Co., 1989), pp 67, 96

Structural Standards for Steel Antenna Towers and Antenna Supporting Structures, TIA/EIA Standard TIA/EIA-222-F, Electronic Industries Association, Mar 1996. May be purchased from Telecommunications Industry Association (TIA), Standards and Technology Department, 2500 Wilson Boulevard, Arlington VA, 22201 or Global Engineering Documents, 15 Inverness Way East, Englewood, CO, 80112-5704, 1-800-854-7179.

Chapter 23

Radio Wave Propagation

Because radio communication is carried on by means of electromagnetic waves traveling through the Earth's atmosphere, it is important to understand the nature of these waves and their behavior in the propagation medium. Most antennas will radiate the power applied to them efficiently, but no antenna can do all things equally well, under all circumstances. Whether you design and build your own antennas, or buy them and have them put up by a professional, you'll need propagation know-how for best results, both during the planning stages and while operating your station.

For station planning, this chapter contains detailed new information on elevation angles from transmitting locations throughout the world to important areas throughout the world. With this information in hand, you can design your own antenna installation for optimum capabilities possible within your budget. See the CD-ROM in back of this book.

The Nature of Radio Waves

You probably have some familiarity with the concept of electric and magnetic fields. A radio wave is a combination of both, with the energy divided equally between them. If the wave could originate at a point source in free space, it would spread out in an ever-growing sphere, with the source at the center. No antenna can be designed to do this, but the theoretical *isotropic antenna* is useful in explaining and measuring the performance of practical antennas we *can* build. It is, in fact, the basis for any discussion or evaluation of antenna performance.

Our theoretical spheres of radiated energy would expand very rapidly at the same speed as the propagation of light, approximately 186,000 miles or 300,000,000 meters per second. These values are close enough for practical purposes, and are used elsewhere in this book. If one wishes to be more precise, light propagates in a vacuum at the speed of 299.7925 meters per microsecond, and slightly slower in air.

The path of a *ray* traced from its source to any point on a spherical surface is considered to be a straight line—a radius of the sphere. An observer on the surface of the sphere would think of it as being flat, just as the Earth seems flat to us. A radio wave far enough from its source to appear flat is called a *plane wave*. From here on, we will be discussing primarily plane waves.

It helps to understand the radiation of electromagnetic energy if we visualize a plane wave as being made up of electric and magnetic forces, as shown in **Fig 1**. The nature of wave propagation is such that the electric and magnetic lines of force are always perpendicular. The plane containing the sets of crossed lines represents the wave front. The direction of travel is always perpendicular to the wave front; *forward* or *backward* is determined by the relative directions of the electric and magnetic forces.

The speed of travel of a wave through anything but a vacuum is always less than 300,000,000 meters per second. How much less depends on the medium. If it is air, the reduction in propagation speed can be ignored in most discussions of propagation at frequencies below 30 MHz. In the VHF range and higher, temperature and moisture content of the medium have increasing effects on the communication range, as will be discussed later. In solid insulating materials the speed is considerably less. In distilled water (a good insulator) the speed is $^1/_9$ that in free space. In good conductors the speed is so low that the opposing fields set up by the wave front occupy practically the same space as the wave itself, and thus cancel it out. This is the reason for "skin effect" in conductors at high frequencies, making metal enclosures good shields for electrical circuits working at radio frequencies.

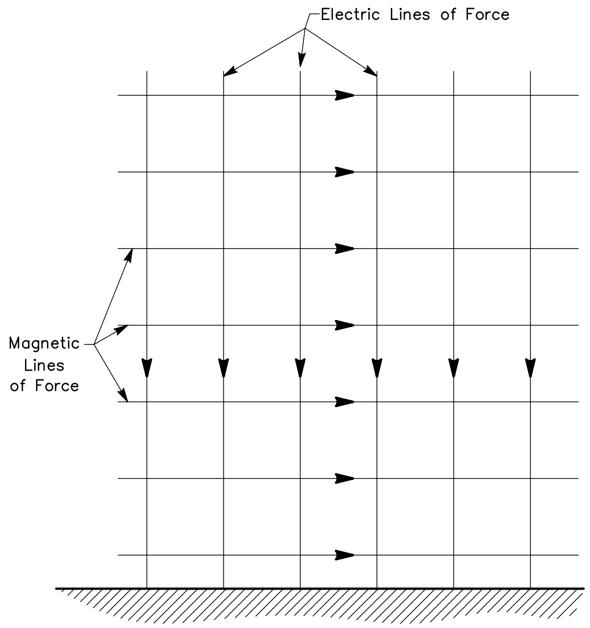

Fig 1—Representation of the magnetic and electric fields of a vertically polarized plane wave traveling along the ground. The arrows indicate instantaneous directions of the fields for a wave traveling perpendicularly out of the page toward the reader. Reversal of the direction of one set of lines reverses the direction of travel. There is no change in direction when both sets are reversed. Such a dual reversal occurs in fact once each half cycle.

Phase and Wavelength

Because the velocity of wave propagation is so great, we tend to ignore it. Only $^1/_7$ of a second is needed for a radio wave to travel around the world—but in working with antennas the *time* factor is extremely important. The wave concept evolved because an alternating current flowing in a wire (antenna) sets up moving electric and magnetic fields. We can hardly discuss antenna theory or performance at all without involving travel time, consciously or otherwise.

Waves used in radio communication may have frequencies from about 10,000 to several billion Hz. Suppose the frequency is 30 MHz. One cycle, or period, is completed in 1/30,000,000 second. The wave is traveling at 300,000,000 meters per second, so it will move only 10 meters during the time that the current is going through one complete period of alternation. The electromagnetic field 10 meters away from the antenna is caused by the current that was flowing one period earlier in time. The field 20 meters away is caused by the current that was flowing two periods earlier, and so on.

If each period of the current is simply a repetition of the one before it, the currents at corresponding instants in each period will be identical. The fields caused by those currents will also be identical. As the fields move outward from the antenna they become more thinly spread over larger and larger surfaces. Their amplitudes decrease with distance from the antenna but they do not lose their identity with respect to the instant of the period at which they were generated. They are, and they remain, *in phase*. In the example above, at intervals of 10 meters measured outward from the antenna, the phase of the waves at any given instant is identical.

From this information we can define both *wave front* and *wavelength*. Consider the wave front as an imaginary surface. On every part of this surface, the wave is in the same phase. The wavelength is the distance between two wave fronts having the same phase at any given instant. This distance must be measured perpendicular to the wave fronts along the line that represents the direction of travel. The abbreviation for wavelength is the Greek letter lambda, λ, which is used throughout this book.

The wavelength will be in the same length units as the velocity when the frequency is expressed in the same time units as the velocity. For waves traveling in free space (and near enough for waves traveling through air) the wavelength is

$$\lambda_{\text{meters}} = \frac{299.7925}{\text{F(MHz)}} \qquad \text{(Eq 1)}$$

There will be few pages in this book where phase, wavelength and frequency do not come into the discussion. It is essential to have a clear understanding of their meaning in order to understand the design, installation, adjustment or use of antennas, matching systems or transmission lines in detail. In essence, *phase* means *time*. When something goes through periodic variations, as an alternating current does, corresponding instants in succeeding periods are *in phase*.

The points A, B and C in **Fig 2** are all in phase. They are corresponding instants in the current flow, at 1-λ intervals. This is a conventional view of a sine-wave alternating current, with time progressing to the right. It also represents a *snapshot* of the intensity of the traveling fields, if distance is substituted for time in the horizontal axis. The distance between A and B or between B and C is one wavelength. The field-intensity distribution follows the sine curve, in both amplitude and polarity, corresponding exactly to the time variations in the current that produced the fields. Remember that this is an *instantaneous* picture—the wave moves outward, much as a wave created by a rock thrown into water does.

Polarization

A wave like that in Fig 1 is said to be *polarized* in the direction of the electric lines of force. The polarization here is vertical, because the electric lines are perpendicular to the surface of the Earth. It is one of the laws of electromagnetics that electric lines touching the surface of a perfect conductor must do so perpendicularly, or else they would have to generate infinite currents in the conductor, an obvious impossibility. Most ground is a rather good conductor at frequencies below about 10 MHz, so waves at these

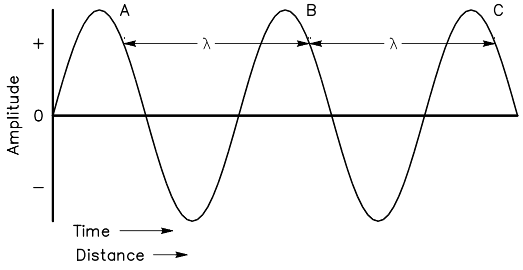

Fig 2—The instantaneous amplitude of both fields (electric and magnetic) varies sinusoidally with time as shown in this graph. Since the fields travel at constant velocity, the graph also represents the instantaneous distribution of field intensity along the wave path. The distance between two points of equal phase such as A-B and B-C is the length of the wave.

frequencies, traveling close to good ground, are mainly vertically polarized. Over partially conducting ground there may be a forward tilt to the wave front; the tilt in the electric lines of force increases as the energy loss in the ground becomes greater.

Waves traveling in contact with the surface of the Earth, called *surface waves*, are of little practical use in amateur communication. This is because as the frequency is raised, the distance over which they will travel without excessive energy loss becomes smaller and smaller. The surface wave is most useful at low frequencies and through the standard AM broadcast band. The surface wave will be covered later. At high frequencies a wave reaching a receiving antenna has had little contact with the ground, and its polarization is not necessarily vertical.

If the electric lines of force are horizontal, the wave is said to be *horizontally polarized*. Horizontally and vertically polarized waves may be classified generally under *linear polarization*. Linear polarization can be anything between horizontal and vertical. In free space, "horizontal" and "vertical" have no meaning, since the reference of the seemingly horizontal surface of the Earth has been lost.

In many cases the polarization of waves is not fixed, but rotates continually, somewhat at random. When this occurs the wave is said to be *elliptically polarized*. A gradual shift in polarization in a medium is known as *Faraday rotation*. For space communication, circular polarization is commonly used to overcome the effects of Faraday rotation. A circularly polarized wave rotates its polarization through 360° as it travels a distance of one wavelength in the propagation medium. The direction of rotation as viewed from the transmitting antenna defines the direction of circularity—right-hand (clockwise) or left-hand (counterclockwise). Linear and circular polarization may be considered as special cases of elliptical polarization.

Field Intensity

The energy from a propagated wave decreases with distance from the source. This decrease in strength is caused by the spreading of the wave energy over ever-larger spheres as the distance from the source increases.

A measurement of the strength of the wave at a distance from the transmitting antenna is its *field intensity*, which is synonymous with *field strength*. The strength of a wave is measured as the voltage between two points lying on an electric line of force in the plane of the wave front. The standard of measure for field intensity is the voltage developed in a wire that is 1 meter long, expressed as volts per meter. (If the wire were 2 meters long, the voltage developed would be divided by two to determine the field strength in volts per meter.)

The voltage in a wave is usually low, so the measurement is made in millivolts or microvolts per meter. The voltage goes through time variations like those of the current that caused the wave. It is measured like any other ac voltage—in terms of the effective value or, sometimes, the peak value. It is fortunate that in amateur work it is not necessary to measure actual field strength, as the equipment required is elaborate. We need to know only if an adjustment has been beneficial, so relative measurements are satisfactory. These can be made easily with home-built equipment.

Wave Attenuation

In free space, the field intensity of the wave varies inversely with the distance from the source, once you are in the radiating far field of the antenna. If the field strength at 1 mile from the source is 100 millivolts per meter, it will be 50 millivolts per meter at 2 miles, and so on. The relationship between field intensity and power density is similar to that for voltage and power in ordinary circuits. They are related by the impedance of free space, which is approximately 377 Ω. A field intensity of 1 volt per meter is therefore equivalent to a power density of

$$P = \frac{E^2}{Z} = \frac{1\,(\text{volt/m})^2}{377\,\Omega} = 2.65\,\text{mW/m}^2 \qquad \text{(Eq 2)}$$

Because of the relationship between voltage and power, the power density therefore varies with the square root of the field intensity, or inversely with the *square* of the distance. If the power density at 1 mile is 4 mW per square meter, then at a distance of 2 miles it will be 1 mW per square meter.

It is important to remember this so-called *spreading loss* when antenna performance is being considered. Gain can come only from narrowing the radiation pattern of an antenna, which concentrates the radiated energy in the desired direction. There is no "antenna magic" by which the total energy radiated can be increased.

In practice, attenuation of the wave energy may be much greater than the inverse-distance law would indicate. The wave does not travel in a vacuum, and the receiving antenna seldom is situated so there is a clear line of sight.

The Earth is spherical and the waves do not penetrate its surface appreciably, so communication beyond visual distances must be by some means that will bend the waves around the curvature of the Earth. These means involve additional energy losses that increase the path attenuation with distance, above that for the theoretical spreading loss in a vacuum.

Bending of Radio Waves

Radio waves and light waves are both propagated as electromagnetic energy. Their major difference is in wavelength, since radio-reflecting surfaces are usually much smaller in terms of wavelength than those for light. In material of a given electrical conductivity, long waves penetrate deeper than short ones, and so require a thicker mass for good reflection. Thin metal however is a good reflector of even long-wavelength radio waves. With poorer conductors, such as the Earth's crust, long waves may penetrate quite a few feet below the surface.

Reflection occurs at any boundary between materials of differing dielectric constant. Familiar examples with light are reflections from water surfaces and window panes. Both water and glass are transparent for light, but their dielectric constants are very different from that of air. Light waves, being very short, seem to bounce off both surfaces. Radio waves, being much longer, are practically unaffected by glass, but their behavior upon encountering water may vary, depending on the purity of that medium. Distilled water is a good insulator; salt water is a relatively good conductor.

Depending on their wavelength (and thus their frequency), radio waves may be reflected by buildings, trees, vehicles, the ground, water, ionized layers in the upper atmosphere, or at boundaries between air masses having different temperatures and moisture content. Ionospheric and atmospheric conditions are important in practically all communication beyond purely local ranges.

Refraction is the bending of a ray as it passes from one medium to another at an angle. The appearance of bending of a straight stick, where it enters water at an angle, is an example of light refraction known to us all. The degree of bending of radio waves at boundaries between air masses increases with the radio frequency. There is slight atmospheric bending in our HF bands. It becomes noticeable at 28 MHz, more so at 50 MHz, and it is much more of a factor in the higher VHF range and in UHF and microwave propagation.

Diffraction of light over a solid wall prevents total darkness on the far side from the light source. This is caused largely by the spreading of waves around the top of the wall, due to the interference of one part of the beam with another. The dielectric constant of the surface of the obstruction may affect what happens to our radio waves when they encounter terrestrial obstructions—but the radio *shadow area* is never totally dark. See Chapter 3, The Effects of Ground, for more information on diffraction.

The three terms, reflection, refraction and diffraction, were in use long before the radio age began. Radio propagation is nearly always a mix of these phenomena, and it may not be easy to identify or separate them while they are happening when we are on the air. This book tends to rely on the words *bending* and *scattering* in its discussions, with appropriate modifiers as needed. The important thing to remember is that any alteration of the path taken by energy as it is radiated from an antenna is almost certain to affect on-the-air results—which is why this chapter on propagation is included in an antenna book.

GROUND WAVES

As we have already seen, radio waves are affected in many ways by the media through which they travel. This has led to some confusion of terms in earlier literature concerning wave propagation. Waves travel close to the ground in several ways, some of which involve relatively little contact with the ground itself. The term *ground wave* has had several meanings in antenna literature, but it has come to be applied to any wave that stays close to the Earth, reaching the receiving point without leaving the Earth's lower atmosphere. This distinguishes the ground wave from a *sky wave*, which utilizes the ionosphere for propagation between the transmitting and receiving antennas.

The ground wave could be traveling in actual contact with the ground, as in Fig 1, where it is called the *surface wave*. Or it could travel directly between the transmitting and receiving antennas, when they are high enough so they can "see" each other—this is commonly called the *direct wave*. The ground wave also travels between the transmitting and receiving antennas by reflections or diffractions off intervening terrain between them. The ground-influenced wave may interact with the direct wave to create a vector-summed resultant at the receiver antenna.

In the generic term ground wave, we also will include ones that are made to follow the Earth's curvature by bending in the Earth's lower atmosphere, or *troposphere*, usually no more than a few miles above the ground. Often called *tropospheric bending*, this propagation mode is a major factor in amateur communications above 50 MHz.

THE SURFACE WAVE

The surface wave travels in contact with the Earth's surface. It can provide coverage up to about 100 miles in the standard AM broadcast band during the daytime, but attenuation is high. As can be seen from **Fig 3**, the attenuation increases with frequency. The surface wave is of little value in amateur communication, except possibly at 1.8 MHz. Vertically polarized antennas must be used, which tends to limit amateur surface-wave communication to where large vertical systems can be erected.

THE SPACE WAVE

Propagation between two antennas situated within line of sight of each other is shown in **Fig 4**. Energy traveling directly between the antennas is attenuated to about the same degree as in free space. Unless the antennas are very high or

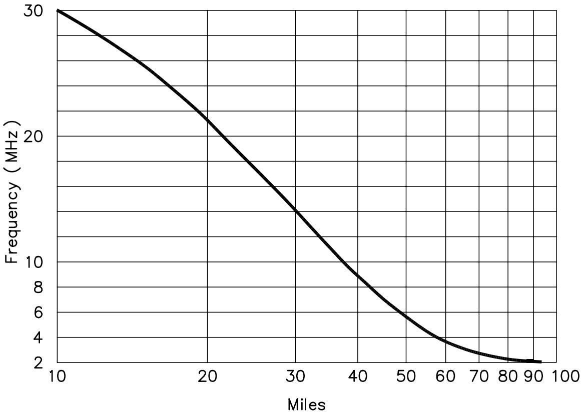

Fig 3—Typical HF ground-wave range as a function of frequency.

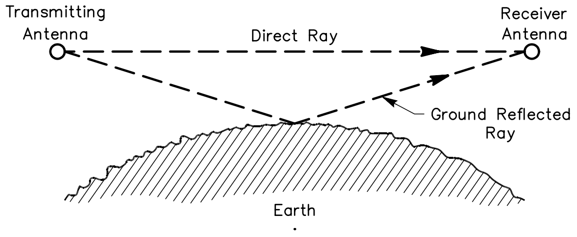

Fig 4—The ray traveling directly from the transmitting antenna to the receiving antenna combines with a ray reflected from the ground to form the space wave. For a horizontally polarized signal a reflection as shown here reverses the phase of the ground-reflected ray.

quite close together, an appreciable portion of the energy is reflected from the ground. This reflected wave combines with direct radiation to affect the actual signal received.

In most communication between two stations on the ground, the angle at which the wave strikes the ground will be small. For a horizontally polarized signal, such a reflection reverses the phase of the wave. If the distances traveled by both parts of the wave were the same, the two parts would arrive out of phase, and would therefore cancel each other. The ground-reflected ray in Fig 4 must travel a little further, so the phase difference between the two depends on the lengths of the paths, measured in wavelengths. The wavelength in use is important in determining the useful signal strength in this type of communication.

If the difference in path length is 3 meters, the phase difference with 160-meter waves would be only 360° × 3/160 = 6.8°. This is a negligible difference from the 180° shift caused by the reflection, so the effective signal strength over the path would still be very small because of cancellation of the two waves. But with 6-meter radio waves the phase length would be 360° × 3/6 = 180°. With the additional 180° shift on reflection, the two rays would add. Thus, the space wave is a negligible factor at low frequencies, but it can be increasingly useful as the frequency is raised. It is a dominant factor in local amateur communication at 50 MHz and higher.

Interaction between the direct and reflected waves is the principle cause of *mobile flutter* observed in local VHF communication between fixed and mobile stations. The flutter effect decreases once the stations are separated enough so that the reflected ray becomes inconsequential. The reflected energy can also confuse the results of field-strength measurements during tests on VHF antennas.

As with most propagation explanations, the space-wave picture presented here is simplified, and practical considerations dictate modifications. There is always some energy loss when the wave is reflected from the ground. Further, the phase of the ground-reflected wave is not shifted exactly 180°, so the waves never cancel completely. At UHF, ground-reflection losses can be greatly reduced or eliminated by using highly directive antennas. By confining the antenna pattern to something approaching a flashlight beam, nearly all the energy is in the direct wave. The resulting energy loss is low enough that microwave relays, for example, can operate with moderate power levels over hundreds or even thousands of miles. Thus we see that, while the space wave is inconsequential below about 20 MHz, it can be a prime asset in the VHF realm and higher.

VHF/UHF PROPAGATION BEYOND LINE OF SIGHT

From Fig 4 it appears that use of the space wave depends on direct line of sight between the antennas of the communicating stations. This is not literally true, although that belief was common in the early days of amateur communication on frequencies above 30 MHz. When equipment became available that operated more efficiently and after antenna techniques were improved, it soon became clear that VHF waves were actually being bent or scattered in several ways, permitting reliable communication beyond visual distances between the two stations. This was found true even with low power and simple antennas. The average communication range can be approximated by assuming the waves travel in straight lines, but with the Earth's radius increased by one-third. The distance to the *radio horizon* is then given as

$$D_{\mathrm{miles}} = 1.415\sqrt{H_{\mathrm{feet}}} \qquad \text{(Eq 3)}$$

or

$$D_{\mathrm{km}} = 4.124\ \sqrt{H_{\mathrm{meters}}} \qquad \text{(Eq 4)}$$

where H is the height of the transmitting antenna, as shown in **Fig 5**. The formula assumes that the Earth is smooth out to the horizon, so any obstructions along the path must be taken into consideration. For an elevated receiving antenna the communication distance is equal to D + D1, that is, the sum of the distances to the horizon of both

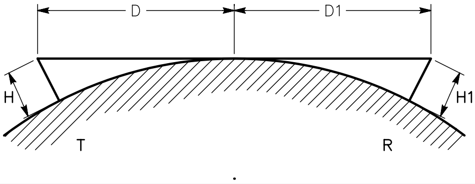

Fig 5—The distance D to the horizon from an antenna of height H is given by equations in the text. The maximum line-of-sight distance between two elevated antennas is equal to the sum of their distances to the horizon as indicated here.

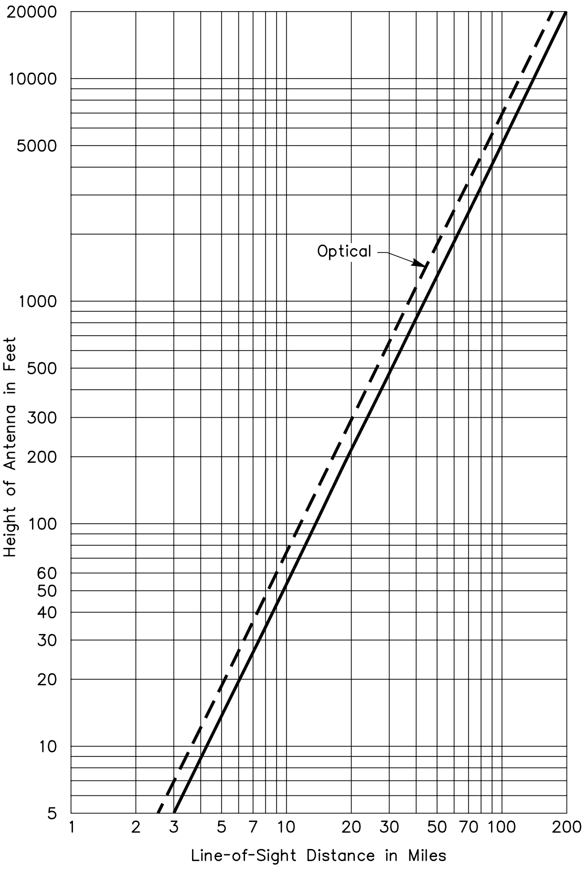

Fig 6—Distance to the horizon from an antenna of given height. The solid curve includes the effect of atmospheric retraction. The optical line-of-sight distance is given by the broken curve.

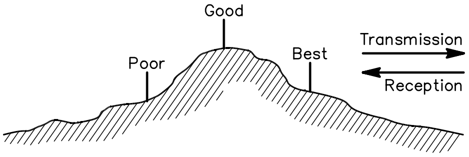

Fig 7—Propagation conditions are generally best when the antenna is located slightly below the top of a hill on the side facing the distant station. Communication is poor when there is a sharp rise immediately in front of the antenna in the direction of communication.

antennas. Radio horizon distances are given in graphic form in **Fig 6**. Two stations on a flat plain, one with its antenna 60 feet above ground and the other 40 feet, could be up to about 20 miles apart for strong-signal line-of-sight communication (11 + 9 mi). The terrain is almost never completely flat, however, and variations along the way may add to or subtract from the distance for reliable communication. Remember that energy is absorbed, reflected or scattered in many ways in nearly all communication situations. The formula or the chart will be a good guide for estimating the potential radius of coverage for a VHF FM repeater, assuming the users are mobile or portable with simple, omnidirectional antennas. Coverage with optimum home-station equipment, high-gain directional arrays, and SSB or CW is quite a different matter. A much more detailed method for estimating coverage on frequencies above 50 MHz is given later in this chapter.

For maximum use of the ordinary space wave it is important to have the antenna as high as possible above nearby buildings, trees, wires and surrounding terrain. A hill that rises above the rest of the countryside is a good location for an amateur station of any kind, and particularly so for extensive coverage on the frequencies above 50 MHz. The highest point on such an eminence is not necessarily the best location for the antenna. In the example shown in **Fig 7**, the hilltop would be a good site in all directions. But if maximum performance to the right is the objective, a point just below the crest might do better. This would involve a trade-off with reduced coverage in the opposite direction. Conversely, an antenna situated on the left side, lower down the hill, might do well to the left, but almost certainly would be inferior in performance to the right.

Selection of a home site for its radio potential is a complex business, at best. A VHF enthusiast dreams of the highest hill. The DX-minded HF ham may be more attracted by a dry spot near a salt marsh. A wide saltwater horizon, especially from a high cliff, just smells of DX. In shopping for ham radio real estate, a mobile or portable rig for the frequencies you're most interested in can provide useful clues.

ANTENNA POLARIZATION

If effective communication over long distances were the only consideration, we might be concerned mainly with radiation of energy at the lowest possible angle above the horizon. However, being engaged in a residential avocation often imposes practical restrictions on our antenna projects. As an example, our 1.8 and 3.5-MHz bands are used primarily for short-distance communication because they serve that purpose with antennas that are not difficult or expensive to put up. Out to a few hundred miles, simple wire antennas for these bands do well, even though their radiation is mostly at high angles above the horizon. Vertical systems might be better for long-distance use, but they require extensive ground systems for good performance.

Horizontal antennas that radiate well at low angles are most easily erected for 7 MHz and higher frequencies—horizontal wires and arrays are almost standard practice for work on 7 through 29.7 MHz. Vertical antennas, such as a single omnidirectional antenna of multiband design, are also used in this frequency range. An antenna of this type may be a good solution to the space problem for a city dweller on a small lot, or even for the resident of an apartment building.

High-gain antennas are almost always used at 50 MHz and higher frequencies, and most of them are horizontal. The principal exception is mobile communication with FM through repeaters, discussed in Chapter 17, Repeater Antenna Systems. The height question is answered easily for VHF enthusiasts—the higher the better.

The theoretical and practical effects of height above ground at HF are treated in detail in Chapter 3, The Effects of Ground. Note that it is the height in *wavelengths* that is important—a good reason to think in the metric system, rather than in feet and inches.

In working locally on any amateur frequency band, best results will be obtained with the same polarization at both stations, except on rare occasions when polarization shift is caused by terrain obstructions or reflections from buildings. Where such a shift is observed, mostly above 100 MHz or so, horizontal polarization tends to work better than vertical. This condition is found primarily on short paths, so it is not too important. Polarization shift may occur on long paths where tropospheric bending is a factor, but here the effect tends to be random. Long-distance communication by way of the ionosphere produces random polarization effects, so polarization matching is of little or no importance. This is fortunate for the HF mobile enthusiast, who will find that even his short, inductively loaded whips work very well at all distances other than local.

Because it responds to all plane polarizations equally, circular polarization may pay off on circuits where the arriving polarization is random, but it exacts a 3-dB penalty when used with a single-plane polarization of any kind. Circular systems find greatest use in work with orbiting satellites. It should be remembered that “horizontal” and “vertical” are meaningless terms in space, where the plane-Earth reference is lost.

Polarization Factors Above 50 MHz

In most VHF communication over short distances, the polarization of the space wave tends to remain constant. Polarization discrimination is high, usually in excess of 20 dB, so the same polarization should be used at both ends of the circuit. Horizontal, vertical and circular polarization all have certain advantages above 50 MHz, so there has never been complete standardization on any one of them.

Horizontal systems are popular, in part because they tend to reject man-made noise, much of which is vertically polarized. There is some evidence that vertical polarization shifts to horizontal in hilly terrain, more readily than horizontal shifts to vertical. With large arrays, horizontal systems may be easier to erect, and they tend to give higher signal strengths over irregular terrain, if any difference is observed.

Practically all work with VHF mobiles is now handled with vertical systems. For use in a VHF repeater system, the vertical antenna can be designed to have gain without losing the desired omnidirectional quality. In the mobile station a small vertical whip has obvious aesthetic advantages. Often a telescoping whip used for broadcast reception can be pressed into service for the 144-MHz FM rig. A car-top mount is preferable, but the broadcast whip is a practical compromise. Tests with at least one experimental repeater have shown that horizontal polarization can give a slightly larger service area, but mechanical advantages of vertical systems have made them the almost unanimous choice in VHF FM communication. Except for the repeater field, horizontal is the standard VHF system almost everywhere.

In communication over the Earth-Moon-Earth (EME) route the polarization picture is blurred, as might be expected with such a diverse medium. If the moon were a flat target, we could expect a 180° phase shift from the moon reflection process. But it is not flat. This plus the moon’s *libration* (its slow oscillation, as viewed from the Earth), and the fact that waves must travel both ways through the Earth’s entire atmosphere and magnetic field, provide other variables that confuse the phase and polarization issue. Building a huge array that will track the moon, and give gains in excess of 20 dB, is enough of a task that most EME enthusiasts tend to take their chances with phase and polarization problems. Where rotation of the element plane has been tried it has helped to stabilize signal levels, but it is not widely employed.

PROPAGATION OF VHF WAVES

The wave energy of VHF stations does not simply disappear once it reaches the radio horizon. It is scattered, but it can be heard to some degree for hundreds of miles, well beyond line-of-sight range. Everything on Earth, and in the regions of space up to at least 100 miles, is a potential forward-scattering agent.

Tropospheric scatter is always with us. Its effects are often hidden, masked by more effective propagation modes on the lower frequencies. But beginning in the VHF range, scatter from the lower atmosphere extends the reliable range markedly if we make use of it. Called *troposcatter*, this is what produces that nearly flat portion of the curves that will be described later (in the section where you can compute reliable VHF coverage range). With a decent station, you can consistently make troposcatter contacts out to 300 miles out on the VHF and even UHF bands, especially if you don’t mind weak signals and something less than 99% reliability. As long ago as the early 1950s, VHF enthusiasts found that VHF contests could be won with high power, big antennas and a good ear for signals deep in the noise. They still can.

Ionospheric scatter works much the same as the tropo version, except that the scattering medium is higher up, mainly the E region of the ionosphere but with some help from the D and F layers too. Ionospheric scatter is useful mainly above the MUF, so its useful frequency range depends on geography, time of day, season, and the state of the Sun. With near maximum legal power, good antennas and quiet locations, ionospheric scatter can fill in the skip zone with marginally readable signals scattered from ionized trails of meteors, small areas of random ionization, cosmic dust, satellites and whatever may come into the antenna patterns at 50 to 150 miles or so above the Earth. It's mostly an E-layer business, so it works all E-layer distances. Good antennas and keen ears help.

Transequatorial propagation (TE) was an amateur 50-MHz discovery in the years 1946-1947. Amateurs of all continents observed it almost simultaneously on three separate north-south paths. These amateurs tried to communicate at 50 MHz, even though the predicted MUF was around 40 MHz for the favorable daylight hours. The first success came at night, when the MUF was thought to be even lower. A remarkable research program inaugurated by amateurs in Europe, Cyprus, Zimbabwe and South Africa eventually provided technically sound theories to explain the then-unknown mode.

It has been known for years that the MUF is higher and less seasonally variable on transequatorial circuits, but the full extent of the difference was not learned until amateur work brought it to light. As will be explained in a later section in more detail, the ionosphere over equatorial regions is higher, thicker and more dense than elsewhere. Because of its more constant exposure to solar radiation, the equatorial belt has high nighttime-MUF possibilities. TE can often work marginally at 144 MHz, and even at 432 MHz on occasion. The potential MUF varies with solar activity, but not to the extent that conventional F-layer propagation does. It is a late-in-the-day mode, taking over about when normal F-layer propagation goes out.

The TE range is usually within about 4000 km (2500 miles) either side of the geomagnetic equator. The Earth's magnetic axis is tilted with respect to the geographical axis, so the TE belt appears as a curving band on conventional flat maps of the world. See **Fig 8**. As a result, TE has a different latitude coverage in the Americas from that from Europe to Africa. The TE belt just reaches into the southern continental US. Stations in Puerto Rico, Mexico and even the northern parts of South America encounter the mode more often than those in favorable US areas. It is no accident that TE was discovered as a result of 50-MHz work in Mexico City and Buenos Aires.

Within its optimum regions of the world, the TE mode extends the usefulness of the 50-MHz band far beyond that of conventional F-layer propagation, since the practical TE MUF runs around 1.5 times that of normal F_2. Both its seasonal and diurnal characteristics are extensions of what is considered normal for 50-MHz propagation. In that part of the Americas south of about 20° North latitude, the existence of TE affects the whole character of band usage, especially in years of high solar activity.

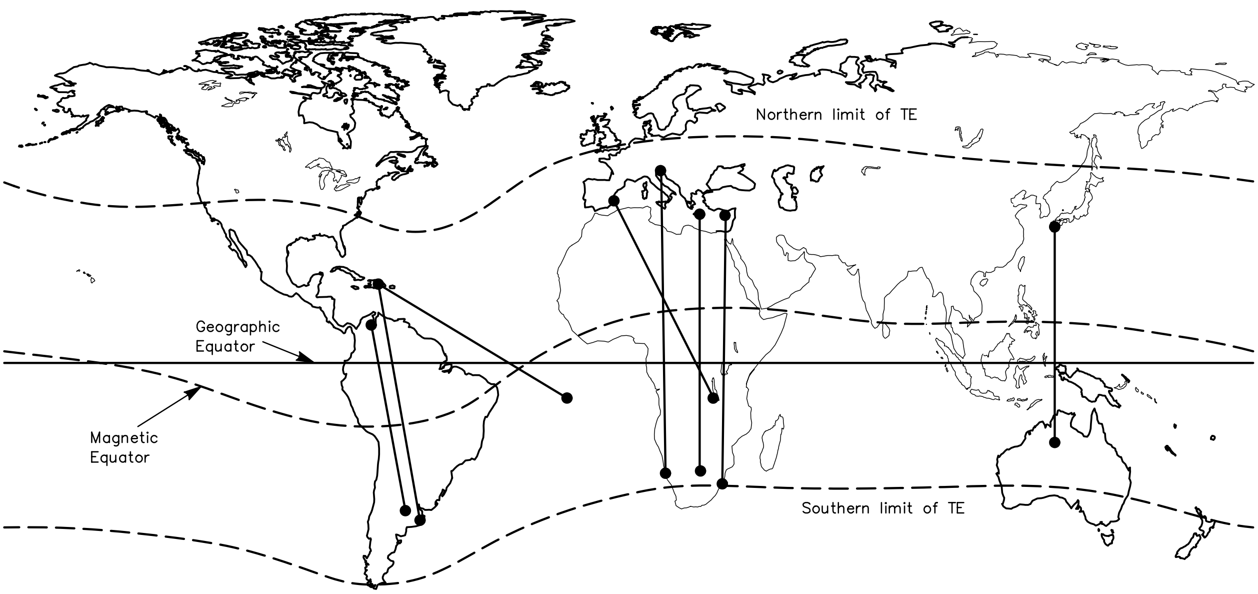

Fig 8—Transequatorial spread-F propagation takes place between stations equidistant across the geomagnetic equator. Distances up to 8000 km (5000 miles) are possible on 28 through 432 MHz. Note that the geomagnetic equator is considerably south of the geographic equator in the Western Hemisphere. (*Figure courtesy of The ARRL Handbook.*)

Weather Effects on VHF/UHF Tropospheric Propagation

Changes in the dielectric constant of the medium can affect propagation. Varied weather patterns over most of the Earth's surface can give rise to boundaries between air masses of very different temperature and humidity characteristics. These boundaries can be anything from local anomalies to air-circulation patterns of continental proportions.

Under stable weather conditions, large air masses can retain their characteristics for hours or even days at a time. See **Fig 9**. Stratified warm dry air over cool moist air, flowing slowly across the Great Lakes region to the Atlantic Seaboard, can provide the medium for east-west communication on 144 MHz and higher amateur frequencies over as much as 1200 miles. More common, however, are communication distances of 400 to 600 miles under such conditions.

A similar inversion along the Atlantic Seaboard as a result of a tropical storm air-circulation pattern may bring VHF and UHF openings extending from the Maritime Provinces of Canada to the Carolinas. Propagation across the Gulf of Mexico, sometimes with very high signal levels, enlivens the VHF scene in coastal areas from Florida to Texas. The California coast, from below the San Francisco Bay Area to Mexico, is blessed with a similar propagation aid during the warmer months. Tropical storms moving west, across the Pacific below the Hawaiian Islands, may provide a transpacific long-distance VHF medium. Amateurs first exploited this on 144, 220 and 432 MHz, in 1957. It has been used fairly often in the summer months since, although not yearly.

The examples of long-haul work cited above may occur infrequently, but lesser extensions of the minimum operating range are available almost daily. Under minimum conditions there may be little more than increased signal strength over paths that are workable at any time.

There is a diurnal effect in temperate climates. At sunrise the air aloft is warmed more rapidly than that near the Earth's surface, and as the Sun goes lower late in the day the upper air is kept warm, while the ground cools. In fair, calm weather such sunrise and sunset temperature inversions can improve signal strength over paths beyond line of sight as much as 20 dB over levels prevailing during the hours of high sun. The diurnal inversion may also extend the operating range for a given strength by some 20 to 50%. If you would be happy with a new VHF antenna, try it first around sunrise!

There are other short-range effects of local atmospheric and topographical conditions. Known as *subsidence*, the flow of cool air down into the bottom of a valley, leaving warm air aloft, is a familiar summer-evening pleasure. The daily inshore-offshore wind shift along a seacoast in summer sets up daily inversions that make coastal areas highly favored as VHF sites. Ask any jealous 144-MHz operator who lives more than a few miles inland.

Tropospheric effects can show up at any time, in any season. Late spring and early fall are the most favored periods, although a winter warming trend can produce strong and stable inversions that work VHF magic almost equal to that of the more familiar spring and fall events.

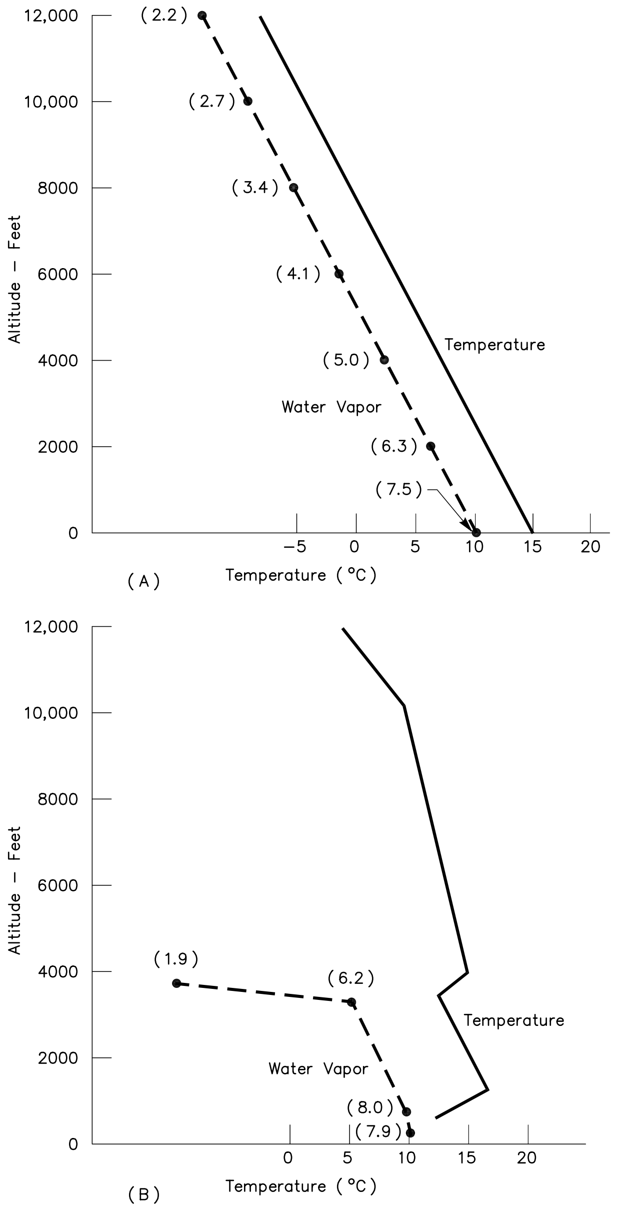

Fig 9—Upper air conditions that produce extended-range communication on the VHF bands. At the top is shown the US Standard Atmosphere temperature curve. The humidity curve (dotted) is what would result if the relative humidity were 70%, from ground level to 12,000 feet elevation. There is only slight refraction under this standard condition. At the bottom is shown a sounding that is typical of marked refraction of VHF waves. Figures in parentheses are the "mixing ratio" —grams of water vapor per kilogram of dry air. Note the sharp break in both curves at about 3500 feet.

Regions where the climate is influenced by large bodies of water enjoy the greatest degree of tropospheric bending. Hot, dry desert areas see little of it, at least in the forms described above.

Tropospheric Ducting

Tropospheric propagation of VHF and UHF waves can influence signal levels at all distances from purely local to something beyond 4000 km (2500 miles). The outer limits are not well known. At the risk of over simplification, we will divide the modes into two classes—extended local and long distance. This concept must be modified depending on the frequency under consideration, but in the VHF range the extended-local effect gives way to a form of propagation much like that of microwaves in a waveguide, called *ducting*. The transition distance is ordinarily somewhere around 200 miles. The difference lies in whether the atmospheric condition producing the bending is localized or continental in scope. Remember, we're concerned here with frequencies in the VHF range, and perhaps up to 500 MHz. At 10 GHz, for example, the scale is much smaller.

In VHF propagation beyond a few hundred miles, more than one weather front is probably involved, but the wave is propagated between the inversion layers and ground, in the main. On long paths over the ocean (two notable examples are California to Hawaii and Ascension Island to Brazil), propagation is likely to be between two atmospheric layers. On such circuits the communicating station antennas must be in the duct, or capable of propagating strongly into it. Here again, we see that the positions and radiation angles of the antennas are important. As with microwaves in a waveguide, the low-frequency limit for the duct is critical. In long-distance ducting it is also very variable. Airborne equipment has shown that duct capability exists well down into the HF region in the stable atmosphere west of Ascension Island. Some contacts between Hawaii and Southern California on 50 MHz are believed to have been by way of tropospheric ducts. Probably all contact over these paths on 144 MHz and higher bands is because of duct propagation.

Amateurs have played a major part in the discovery and eventual explanation of tropospheric propagation. In recent years they have shown that, contrary to beliefs widely held in earlier times, long-distance communication using tropospheric modes is possible to some degree on all amateur frequencies from 50 to at least 10,000 MHz.

RELIABLE VHF COVERAGE

In the preceding sections we discussed means by which amateur bands above 50 MHz may be used intermittently for communication far beyond the visual horizon. In emphasizing distance we should not neglect a prime asset of the VHF band: reliable communication over relatively short distances. The VHF region is far less subject to disruption of local communication than are frequencies below 30 MHz. Since much amateur communication is essentially local in nature, our VHF assignments can carry a great load, and such use of the VHF bands helps solve interference problems on lower frequencies.

Because of age-old ideas, misconceptions about the coverage obtainable in our VHF bands persist. This reflects the thoughts that VHF waves travel only in straight lines, except when the DX modes described above happen to be present. However, let us survey the picture in the light of modern wave-propagation knowledge and see what the bands above 50 MHz are good for on a day-to-day basis, ignoring the anomalies that may result in extensions of normal coverage.

It is possible to predict with fair accuracy how far you should be able to work consistently on any VHF or UHF band, provided a few simple facts are known. The factors affecting operating range can be reduced to graph form, as described in this section. The information was originally published in November 1961 *QST* by D. W. Bray, K2LMG (see the Bibliography at the end of this chapter).

To estimate your station's capabilities, two basic numbers must be determined: station gain and path loss. Station gain is made up of seven factors: receiver sensitivity, transmitted power, receiving antenna gain, receiving antenna height gain, transmitting antenna gain, transmitting antenna height gain and required signal-to-noise ratio. This looks complicated but it really boils down to an easily made evaluation of receiver, transmitter, and antenna performance. The other number, path loss, is readily determined from the nomogram, **Fig 10**. This gives path loss over smooth Earth, for 99% reliability.

For 50 MHz, lay a straightedge from the distance between stations (left side) to the appropriate distance at the right side. For 1296 MHz, use the full scale, right center. For 144, 222 and 432, use the dot in the circle, square or triangle, respectively. Example: At 300 miles the path loss for 144 MHz is 214 dB.

To be meaningful, the losses determined from this nomograph are necessarily greater than simple free-space path losses. As described in an earlier section, communication beyond line-of-sight distances involves propagation modes that increase the path attenuation with distance.

VHF/UHF Station Gain

The largest of the eight factors involved in station design is receiver sensitivity. This is obtainable from **Fig 11**, if you know the approximate receiver noise figure and transmission-line loss. If you can't measure noise figure, assume 3 dB for 50 MHz, 5 for 144 or 222, 8 for 432 and 10 for 1296 MHz, if you know your equipment is working moderately well. These noise figures are well on the conservative side for modern solid-state receivers.

Line loss can be taken from information in Chapter 24 for the line in use, if the antenna system is fed properly. Lay a straightedge between the appropriate points at either side of Fig 11, to find effective receiver sensitivity in decibels below 1 watt (dBW). Use the narrowest bandwidth that is practical for the emission intended, with the receiver you will be using. For CW, an average value for effective work

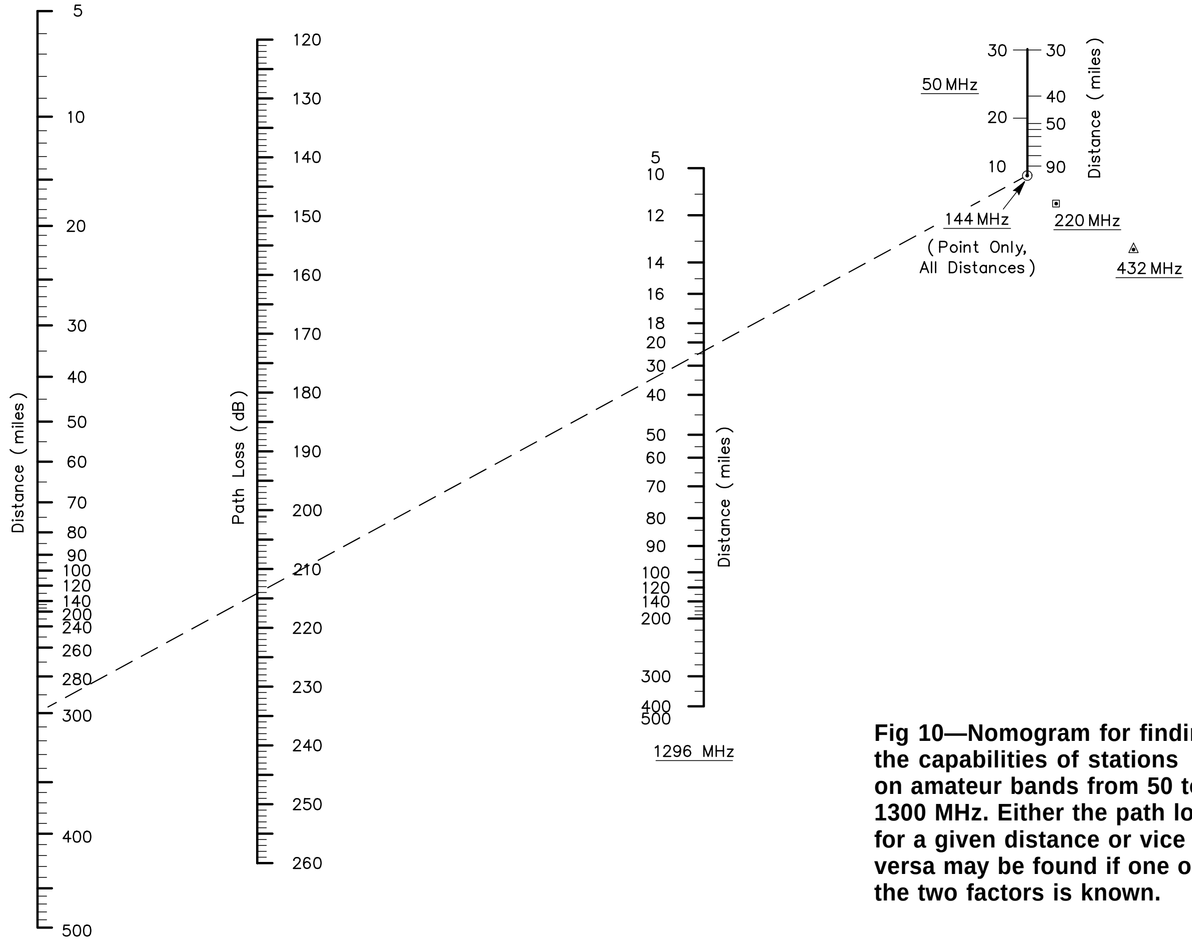

Fig 10—Nomogram for finding the capabilities of stations on amateur bands from 50 to 1300 MHz. Either the path loss for a given distance or vice versa may be found if one of the two factors is known.

is about 500 Hz. Phone bandwidth can be taken from the receiver instruction manual, but it usually falls between 2.1 to 2.7 kHz.

Antenna gain is next in importance. Gains of amateur antennas are often exaggerated. For well-designed Yagis the gain (over isotropic) run close to 10 times the boom length in wavelengths. (Example: A 24-foot Yagi on 144 MHz is 3.6 wavelengths long; 3.6 x 10 = 36, and 10 $\log_{10}$ 36 = 15.5 dBi in free space.) Add 3 dB for stacking, where used properly. Add 4 dB more for ground reflection gain. This varies in amateur work, but averages out near this figure.

We have one more plus factor—antenna height gain, obtained from **Fig 12**. Note that this is greatest for short distances. The left edge of the horizontal center scale is for 0 to 10 miles, the right edge for 100 to 500 miles. Height gain for 10 to 30 feet is assumed to be zero. For 50 feet the height gain is 4 dB at 10 miles, 3 dB at 50 miles, and 2 dB at 100 miles. At 80 feet the height gains are roughly 8, 6 and 4 dB for these distances. Beyond 100 miles the height gain is nearly uniform for a given height, regardless of distance.

Transmitter power output must be stated in decibels above 1 watt. If you have 500 W output, add 10 log (500/1), or 27 dB, to your station gain. The transmission-line loss must be subtracted from the station gain. So must the required signal-to-noise ratio. The information is based on CW work, so the additional signal needed for other modes must be subtracted. Use a figure of 3 dB for SSB. Fading losses must be accounted for also. It has been shown that for distances beyond 100 miles, the signal will vary plus or minus about 7 dB from the average level, so 7 dB must be subtracted from the station gain for high reliability. For distances under 100 miles, fading diminishes almost linearly with distance. For 50 miles, use –3.5 dB for fading.

What It All Means

Add all the plus and minus factors to get the station gain. Use the final value to find the distance over which you can expect to work reliably from the nomogram, Fig 10. Or work it the other way around: Find the path loss for the distance you want to cover from the nomogram and then figure out what station changes will be needed to overcome it.

The significance of all this becomes more obvious when we see path loss plotted against frequency for the various bands, as in **Fig 13**. At the left this is done for 50% reliability. At the right is the same information for 99% reliability.

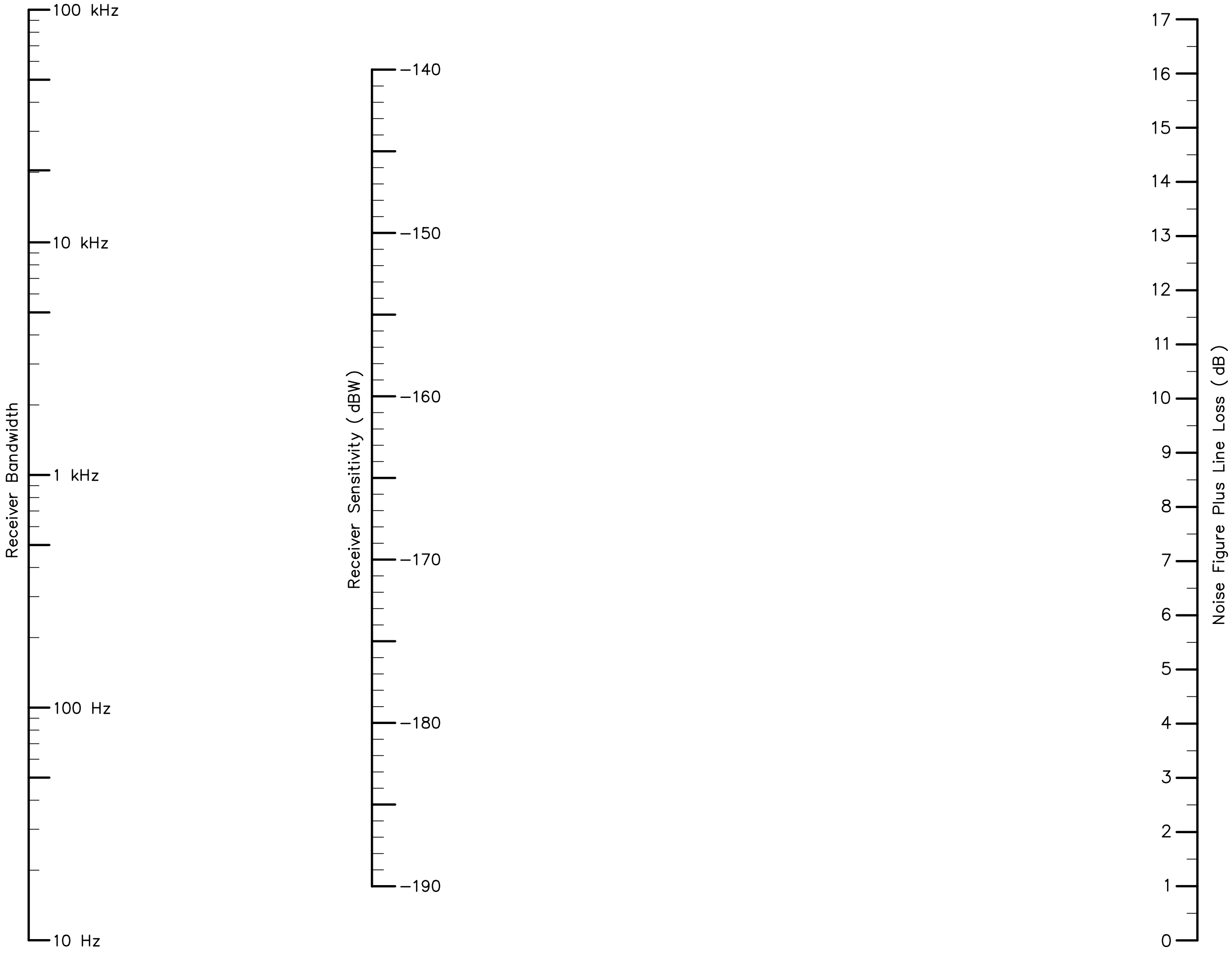

Fig 11—Nomogram for finding effective receiver sensitivity.

For near-perfect reliability, a path loss of 195 dB (easily encountered at 50 or 144 MHz) is involved in 100-mile communication. But look at the 50% reliability curve: The same path loss takes us out to well over 250 miles. Few amateurs demand near-perfect reliability. By choosing our times, and by accepting the necessity for some repeats or occasional loss of signal, we can maintain communication out to distances far beyond those usually covered by VHF stations.

Working out a few typical amateur VHF station setups with these curves will show why an understanding of these factors is important to any user of the VHF spectrum. Note that path loss rises very steeply in the first 100 miles or so. This is no news to VHF operators; locals are very strong, but stations 50 or 75 miles away are much weaker. What happens beyond 100 miles is not so well known to many of us.

From the curves of Fig 13, we see that path loss levels off markedly at what is the approximate limit of working range for average VHF stations using wideband modulation modes. Work out the station gain for a 50-W station with an average receiver and antenna, and you'll find that it comes out around 180 dB. This means you'd have about a 100-mile working radius in average terrain, for good but not perfect reliability. Another 10 dB may extend the range to as much as 250 miles. Just changing from AM phone to SSB and CW makes a major improvement in daily coverage on the VHF bands.

A bigger antenna, a higher one if your present beam is not at least 50 feet up, an increase in power to 500 W from 50 W, an improvement in receiver noise figure if it is presently poor—any of these things can make a big improvement in reliable coverage. Achieve all of them, and you will have very likely tripled your sphere of influence, thanks to that hump in the path-loss curves. This goes a long way toward explaining why using a 10-W packaged station with a small antenna, fun though it may be, does not begin to show what the VHF bands are really good for.

Terrain at VHF/UHF

The coverage figures derived from the above procedure are for average terrain. What of stations in mountainous country? Although an open horizon is generally desirable for the VHF station site, mountain country should not be

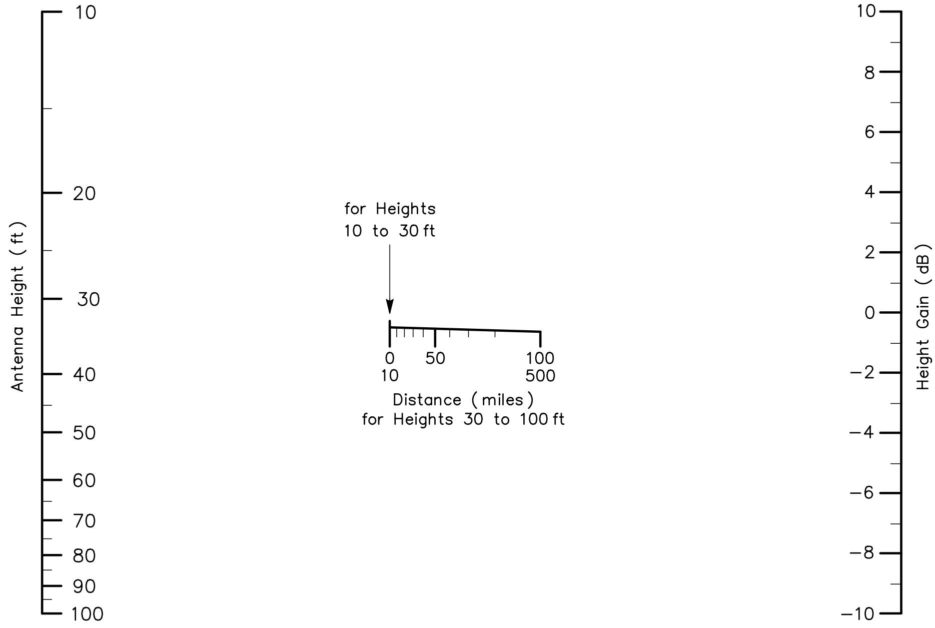

Fig 12—Nomogram for determining antenna-height gain.

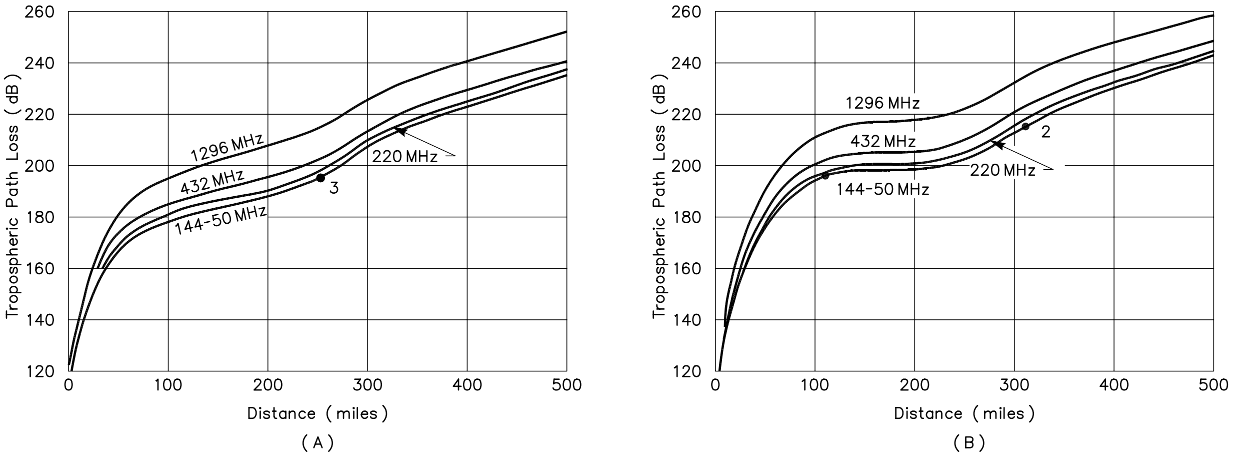

Fig 13—Path loss versus distance for amateur frequencies above 50 MHz. At A are curves for 50% of the time; at B, for 99%. The curves at A are more representative of Amateur Radio requirements.

considered hopeless. Help for the valley dweller often lies in the optical phenomenon known as *knife-edge diffraction*. A flashlight beam pointed at the edge of a partition does not cut off sharply at the partition edge, but is diffracted around it, partially illuminating the shadow area. A similar effect is observed with VHF waves passing over ridges; there is a shadow effect, but not a complete blackout. If the signal is strong where it strikes the mountain range, it will be heard well in the bottom of a valley on the far side. (See Chapter 3, The Effects of Ground, for a more thorough discussion of the theory of diffraction.)

This is familiar to all users of VHF communications equipment who operate in hilly terrain. Where only one ridge lies in the way, signals on the far side may be almost as good as on the near side. Under ideal conditions (a very high and sharp-edged obstruction near the midpoint of a long-enough path so that signals would be weak over average terrain), knife-edge diffraction may yield signals even stronger than would be possible with an open path.

The obstruction must project into the radiation patterns of the antennas used. Often mountains that look formidable to the viewer are not high enough to have an appreciable effect, one way or the other. Since the normal radiation pattern from a VHF array is several degrees above the horizontal, mountains that are less than about three degrees above the horizon, as seen from the antenna, are missed by the radiation from the array. Moving the mountains out of the way would have substantially no effect on VHF signal strength in such cases.

Rolling terrain, where obstructions are not sharp enough to produce knife-edge diffraction, still does not exhibit a complete shadow effect. There is no complete barrier to VHF propagation—only attenuation, which varies widely as the result of many factors. Thus, even valley locations are usable for VHF communication. Good antenna systems, preferably as high as possible, the best available equipment, and above all, the willingness and ability to work with weak signals may make outstanding VHF work possible, even in sites that show little promise by casual inspection.

AURORAL PROPAGATION

The Earth has a *magnetosphere* or magnetic field surrounding it. NASA scientists have described the magnetosphere as a sort of protective "bubble" around the Earth that shields us from the solar wind. Under normal circumstances, there are lots of electrons and protons moving in our magnetosphere, traveling along magnetic lines of force that trap them and keep them in place, neither bombarding the earth nor escaping into outer space.

Sudden bursts of activity on the Sun are sometimes accompanied by the ejection of charged particles, often from so-called Coronal Mass Ejections (CME) because they originate from the Sun's outer coronal region. These charged particles can interact with the magnetosphere, compressing and distorting it. If the orientation of the magnetic field contained in a large blast of solar wind or in a CME is aligned opposite to that of the Earth's magnetic field, the magnetic bubble can partially collapse and the particles normally trapped there can be deposited into the Earth's atmosphere along magnetic lines near the North or South poles. This produces a visible or radio *aurora*. An aurora is visible if the time of entry is after dark.

The visible aurora is, in effect, fluorescence at E-layer height—a curtain of ions capable of refracting radio waves in the frequency range above about 20 MHz. D-region absorption increases on lower frequencies during auroras. The exact frequency ranges depend on many factors: time, season, position with relation to the Earth's auroral regions, and the level of solar activity at the time, to name a few.

The auroral effect on VHF waves is another amateur discovery, this one dating back to the 1930s. The discovery came coincidentally with improved transmitting and receiving techniques then. The returning signal is diffused in frequency by the diversity of the auroral curtain as a refracting (scattering) medium. The result is a modulation of a CW signal, from just a slight burbling sound to what is best described as a "keyed roar." Before SSB took over in VHF work, voice was all but useless for auroral paths. A sideband signal suffers, too, but its narrower bandwidth helps to retain some degree of understandability. Distortion induced by a given set of auroral conditions increases with the frequency in use. 50-MHz signals are much more intelligible than those on 144 MHz on the same path at the same time. On 144 MHz, CW is almost mandatory for effective auroral communication.

The number of auroras that can be expected per year varies with the geomagnetic latitude. Drawn with respect to the Earth's magnetic poles instead of the geographical ones, these latitude lines in the US tilt upward to the northwest. For example, Portland, Oregon, is 2° farther north (geographic latitude) than Portland, Maine. The Maine city's geomagnetic latitude line crosses the Canadian border before it gets as far west as its Oregon namesake. In terms of auroras intense enough to produce VHF propagation results, Portland, Maine, is likely to see about 10 times as many per year. Oregon's auroral prospects are more like those of southern New Jersey or central Pennsylvania.

The antenna requirements for auroral work are mixed. High gain helps, but the area of the aurora yielding the best returns sometimes varies rapidly, so sharp directivity can be a disadvantage. So could a very low radiation angle, or a beam pattern very sharp in the vertical plane. Experience indicates that few amateur antennas are sharp enough in either plane to present a real handicap. The beam heading for maximum signal can change, however, so a bit of scanning in azimuth may turn up some interesting results. A very large array, such as is commonly used for moonbounce (with azimuth-elevation control), should be worthwhile.

The incidence of auroras, their average intensity, and their geographical distribution as to visual sightings and VHF propagation effects all vary to some extent with solar activity. There is some indication that the peak period for auroras lags the sunspot-cycle peak by a year or two. Like sporadic E, an unusual auroral opening can come at any season. There is a marked diurnal swing in the number of auroras. Favored times are late afternoon and early evening, late evening through early morning, and early afternoon, in about that order. Major auroras often start in early afternoon and carry through to early morning the next day.

HF Sky-Wave Propagation

As described earlier, the term *ground wave* is commonly applied to propagation that is confined to the Earth's lower atmosphere. Now we will use the term *sky wave* to describe modes of propagation that use the Earth's ionosphere. First, however, we must examine how the Earth's ionosphere is affected by the Sun.

THE ROLE OF THE SUN

Everything that happens in radio propagation, as with all life on Earth, is the result of radiation from the Sun. The variable nature of radio propagation here on Earth reflects the ever-changing intensity of ultraviolet and X-ray radiation, the primary ionizing agents in solar energy. Every day, solar nuclear reactions are turning hydrogen into helium, releasing an unimaginable blast of energy into space in the process. The total power radiated by the Sun is estimated at 4×10^{23} kW—that is, the number four followed by 23 zeroes. At its surface, the Sun creates about 60 *megawatts* per square meter. That is a very potent transmitter!

The Solar Wind

The Sun is constantly ejecting material from its surface in all directions into space, making up the so-called *solar wind*. Under relatively quiet solar conditions the solar wind blows around 200 miles per second—675,000 miles per hour—taking away about two million tons of solar material each second from the Sun. You needn't worry—the Sun is not going to shrivel up anytime soon. It's big enough that it will take many billions of years before that happens.

A 675,000 mile/hour wind sounds like a pretty stiff breeze, doesn't it? Lucky for us, the density of the material in the solar wind is very small by the time it has been spread out into interplanetary space. Scientists calculate that the density of the particles in the solar wind is less than that of the best vacuum they've ever achieved on Earth. Despite the low density of the material in the solar wind, the effect on the Earth, especially its magnetic field, is very significant.

Before the advent of sophisticated satellite sensors, the Earth's magnetic field was considered to be fairly simple, modeled as if the Earth were a large bar magnet. The axis of this hypothetical bar magnet is oriented about 11° away from the geographic north-south pole. We now know that the solar wind alters the shape of the Earth's magnetic field significantly, compressing it on the side facing the Sun and elongating it on the other side—in the same manner as the tail of a comet is stretched out radially in its orientation from the Sun. In fact, the solar wind is also responsible for the shape of a comet's tail.

Partly because of the very nature of the nuclear reactions going on at the Sun itself, but also because of variations in the speed and direction of the solar wind, the interactions between the Sun and our Earth are incredibly complex. Even scientists who have studied the subject for years do not completely understand everything that happens on the Sun. Later in this chapter, we'll investigate the effects of the solar wind when conditions on the Sun are *not* "quiet." As far as amateur HF skywave propagation is concerned, the results of disturbed conditions on the Sun are not generally beneficial.

Sunspots

The most readily observed characteristic of the Sun, other than its blinding brilliance, is its tendency to have grayish black blemishes, seemingly at random times and at random places, on its fiery surface. (See **Fig 14**.) There are written records of naked-eye sightings of *sunspots* in the Orient back to more than 2000 years ago. As far as is known, the first indication that sunspots were recognized as part of the Sun was the result of observations by Galileo in the early 1600s, not long after he developed one of the first practical telescopes.

Galileo also developed the projection method for observing the Sun safely, but probably not before he had suffered severe eye damage by trying to look at the Sun di-

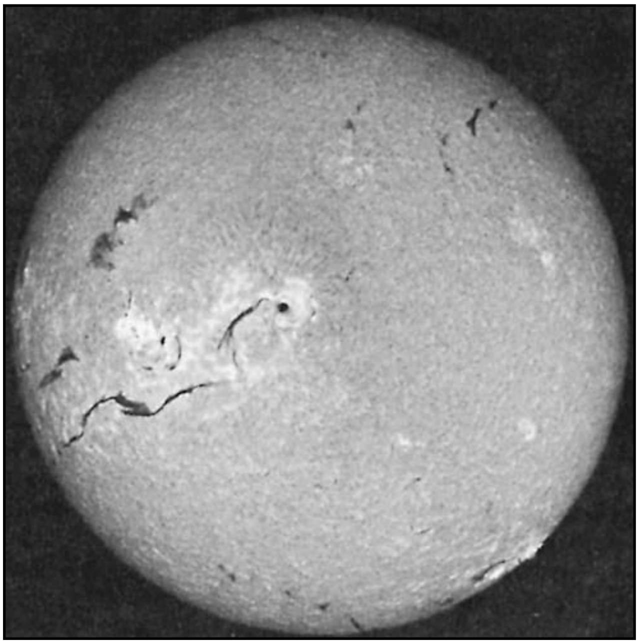

Fig 14—Much more than sunspots can be seen when the sun is viewed through selective optical filters. This photo was taken through a hydrogen-alpha filter that passes a narrow light segment at 6562 angstroms. The bright patches are active areas around and often between sunspots. Dark irregular lines are filaments of activity having no central core. Faint magnetic field lines are visible around a large sunspot group near the disc center. (*Photo courtesy of Sacramento Peak Observatory, Sunspot, New Mexico*).

rectly. (He was blind in his last years.) His drawings of sunspots, indicating their variable nature and position, are the earliest such record known to have been made. His reward for this brilliant work was immediate condemnation by church authorities of the time, which probably set back progress in learning more about the Sun for generations.

The systematic study of solar activity began about 1750, so a fairly reliable record of sunspot numbers goes back that far. (There are some gaps in the early data.) The record shows clearly that the Sun is always in a state of change. It never looks exactly the same from one day to the next. The most obvious daily change is the movement of visible activity centers (sunspots or groups thereof) across the solar disc, from east to west, at a constant rate. This movement was soon found to be the result of the rotation of the Sun, at a rate of approximately four weeks for a complete round. The average is about 27.5 days, the Sun's *synodic* rotation speed, viewed from the perspective of the Earth, which is also moving around the Sun in the same direction as the Sun's rotation.

Sunspot Numbers

Since the earliest days of systematic observation, our traditional measure of solar activity has been based on a count of sunspots. In these hundreds of years we have learned that the average number of spots goes up and down in cycles very roughly approximating a sine wave. In 1848, a method was introduced for the daily measurement of sunspot numbers. That method, which is still used today, was devised by the Swiss astronomer Johann Rudolph Wolf. The observer counts the total number of spots visible on the face of the Sun and the number of groups into which they are clustered, because neither quantity alone provides a satisfactory measure of sunspot activity. The observer's sunspot number for that day is computed by multiplying the number of groups he sees by 10, and then adding to this value the number of individual spots. Where possible, sunspot data collected prior to 1848 have been converted to this system.

As can readily be understood, results from one observer to another can vary greatly, since measurement depends on the capability of the equipment in use and on the stability of the Earth's atmosphere at the time of observation, as well as on the experience of the observer. A number of observatories around the world cooperate in measuring solar activity. A weighted average of the data is used to determine the *International Sunspot Number* or ISN for each day. (Amateur astronomers can approximate the determination of ISN values by multiplying their values by a correction factor determined empirically.)

A major step forward was made with the development of various methods for observing narrow portions of the Sun's spectrum. Narrowband light filters that can be used with any good telescope perform a visual function very similar to the aural function of a sharp filter added to a communications receiver. This enables the observer to see the actual area of the Sun doing the radiating of the ionizing energy, in addition to the sunspots, which are more a by-product than a cause. The photo of Fig 14 was made through such a filter. Studies of the ionosphere with instrumented probes, and later with satellites, manned and unmanned, have added greatly to our knowledge of the effects of the Sun on radio communication.

Daily sunspot counts are recorded, and monthly and yearly averages determined. The averages are used to see trends and observe patterns. Sunspot records were formerly kept in Zurich, Switzerland, and the values were known as *Zurich Sunspot Numbers*. They were also known as Wolf sunspot numbers. The official international sunspot numbers are now compiled at the Sunspot Index Data Center in Bruxelles, Belgium.

The yearly means (averages) of sunspot numbers from 1700 through 2002 are plotted in **Fig 15**. The cyclic nature of solar activity becomes readily apparent from this graph. The duration of the cycles varies from 9.0 to 12.7 years, but averages approximately 11.1 years, usually referred to as the 11-year solar cycle. The first complete cycle to be observed systematically began in 1755, and is numbered Cycle 1. Solar cycle numbers thereafter are consecutive. Cycle 23 began in October, 1996.

The "Quiet" Sun

For more than 60 years it has been well known that radio propagation phenomena vary with the number and size of sunspots, and also with the position of sunspots on the surface of the Sun. There are daily and seasonal variations in the Earth's ionized layers resulting from changes in the amount of ultraviolet light received from the Sun. The 11-year sunspot cycle affects propagation conditions because there is a direct correlation between sunspot activity and ionization.

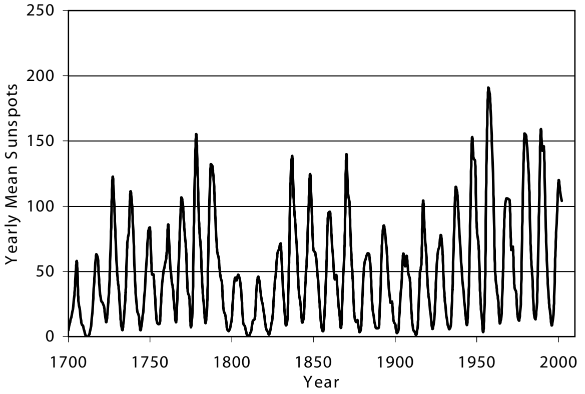

Fig 15—Yearly means of smoothed sunspot numbers from data for 1700 through 2002. This plot clearly shows that sunspot activity takes place in cycles of approximately 11 years duration. There is also a longer-term periodicity in this plot, the Gleissberg 88-year cycle. Cycle 1, the first complete cycle to be examined by systematic observation, began in 1755.

Activity on the surface of the Sun is changing continually. In this section we want to describe the activity of the so-called *quiet Sun*, meaning those times when the Sun is not doing anything more spectacular than acting like a "normal" thermonuclear ball of flaming gases. The Sun and its effects on Earthly propagation can be described in *statistical* terms—that's what the 11-year solar cycle does. You may experience vastly different conditions on any particular day compared to what a long-term average would suggest.

An analogy may be in order here. Have you ever gazed into a relatively calm campfire and been surprised when suddenly a flaming ember or a large spark was ejected in your direction? The Sun can also do unexpected and sometimes very dramatic things. Disturbances of propagation conditions here on Earth are caused by disturbed conditions on the Sun. More on this later.

Individual sunspots may vary in size and appearance, or even disappear totally, within a single day. In general, larger active areas persist through several rotations of the Sun. Some active areas have been identified over periods up to about a year. Because of these continual changes in solar activity, there are continual changes in the state of the Earth's ionosphere and resulting changes in propagation conditions. A short-term burst of solar activity may trigger unusual propagation conditions here on Earth lasting for less than an hour.

Smoothed Sunspot Numbers (SSN)

Sunspot data are averaged or smoothed to remove the effects of short-term changes. The sunspot values used most often for correlating propagation conditions are *Smoothed Sunspot Numbers* (SSN), often called 12-month running average values. Data for 13 consecutive months are required to determine a smoothed sunspot number.

Long-time users have found that the upper HF bands are reliably open for propagation only when the average number of sunspots is above certain minimum levels. For example, between mid 1988 to mid 1992 during Cycle 22, the SSN stayed higher than 100. The 10-meter band was open then almost all day, every day, to some part of the world. However, by mid 1996, few if any sunspots showed up on the Sun and the 10-meter band consequently was rarely open. Even 15 meters, normally a workhorse DX band when solar activity is high, was closed most of the time during the low point in Cycle 22. So far as propagation on the upper HF bands is concerned, the higher the sunspot number, the better the conditions.

Each smoothed number is an average of 13 monthly means, centered on the month of concern. The 1st and 13th months are given a weight of 0.5. A monthly mean is simply the sum of the daily ISN values for a calendar month, divided by the number of days in that month. We would commonly call this value a monthly average.

This may all sound very complicated, but an example should clarify the procedure. Suppose we wished to calculate the smoothed sunspot number for June 1986. We would require monthly mean values for six months prior and six months after this month, or from December 1985 through December 1986. The monthly mean ISN values for these months are

Dec	85	17.3	Jul	86	18.1
Jan	86	2.5	Aug	86	7.4
Feb	86	23.2	Sep	86	3.8
Mar	86	15.1	Oct	86	35.4
Apr	86	18.5	Nov	86	15.2
May	86	13.7	Dec	86	6.8
Jun	86	1.1			

First we find the sum of the values, but using only one-half the amounts indicated for the first and 13th months in the listing. This value is 166.05. Then we determine the smoothed value by dividing the sum by 12: 166.05/12 = 13.8. (Values beyond the first decimal place are not warranted.) Thus, 13.8 is the smoothed sunspot number for June 1986. From this example, you can see that the smoothed sunspot number for a particular month cannot be determined until six months afterwards.

Generally the plots we see of sunspot numbers are averaged data. As already mentioned, smoothed numbers make it easier to observe trends and see patterns, but sometimes this data can be misleading. The plots tend to imply that solar activity varies smoothly, indicating, for example, that at the onset of a new cycle the activity just gradually increases. But this is definitely not so! On any one day, significant changes in solar activity can take place within hours, causing sudden band openings at frequencies well above the MUF values predicted from smoothed sunspot number curves. The durations of such openings may be brief, or they may recur for several days running, depending on the nature of the solar activity.

Solar Flux

Since the late 1940s an additional method of determining solar activity has been put to use—the measurement of *solar radio flux*. The quiet Sun emits radio energy across a broad frequency spectrum, with a slowly varying intensity. Solar flux is a measure of energy received per unit time, per unit area, per unit frequency interval. These radio fluxes, which originate from atmospheric layers high in the Sun's chromosphere and low in its corona, change gradually from day to day, in response to the activity causing sunspots. Thus, there is a degree of correlation between solar flux values and sunspot numbers.

One solar flux unit equals 10^{-22} joules per second per square meter per hertz. Solar flux values are measured daily at 2800 MHz (10.7 cm) at The Dominion Radio Astrophysical Observatory, Penticton, British Columbia, where daily data have been collected since 1991. (Prior to June 1991, the Algonquin Radio Observatory, Ontario, made the measurements.) Measurements are also made at other observatories around the world, at several frequencies. With some variation, the daily measured flux values increase with increasing frequency of measurement, to at least 15.4 GHz. The daily 2800 MHz Penticton value is sent to Boulder, Colorado, where it is incorporated into WWV propagation bulletins (see later section). Daily solar flux information can be of some value in determining current propagation conditions, as sunspot numbers on a given day do not relate directly to maximum usable frequency. Solar flux values are much more reliable for this purpose, when it is averaged over time, as will be discussed later in the section on computer-prediction programs.

Correlating Sunspot Numbers and Solar Flux Values

Based on historical data, an exact mathematical relationship does not exist to correlate sunspot data and solar flux values. Comparing daily values yields almost no correlation. Comparing monthly mean values (often called monthly averages) produces a degree of correlation, but the spread in data is still significant. This is indicated in **Fig 16**, a scatter diagram plot of monthly mean sunspot numbers versus the monthly means of solar flux values adjusted to one astronomical unit. (This adjustment applies a correction for differences in distance between the Sun and the Earth at different times of the year.)

A closer correlation exists when smoothed (12-month running average) sunspot numbers are compared with smoothed (12-month running average) solar flux values adjusted to one astronomical unit. A scatter diagram for smoothed data appears in **Fig 17**. Note how the plot points establish a better defined pattern in Fig 17. The correlation is still no better than a few percent, for records indicate a given smoothed sunspot number does not always correspond with the same smoothed solar flux value, and vice versa.

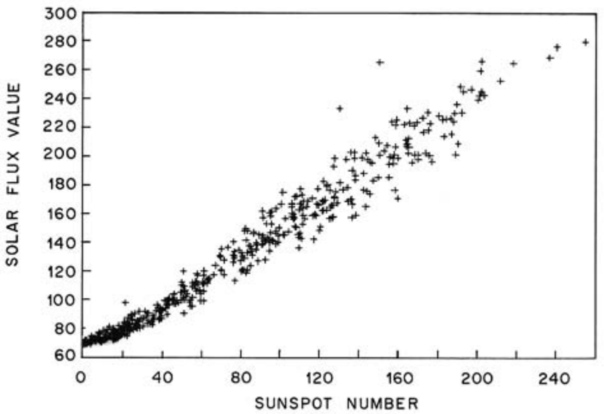

Fig 16—Scatter diagram or X-Y plot of monthly mean sunspot numbers and monthly mean 2800MHz solar flux values. Data values are from February 1947 through February 1987. Each "+" mark represents the intersection of data for a given month. If the correlation between sunspot number and flux values were consistent, all the marks would align to form a smooth curve.

Table 1 illustrates some of the inconsistencies that exist in the historical data. Smoothed or 12-month running average values are shown.

Even though there is no precise mathematical relationship between sunspot numbers and solar flux values, it is helpful to have some way to convert from one to the other. The primary reason is that sunspot numbers are valuable as a long-term link with the past, but the great usefulness of solar flux values are their immediacy, and their direct bearing on our field of interest. (Remember, a smoothed sunspot number will not be calculated until six months after the fact.)

The following mathematical approximation has been derived to convert a smoothed sunspot number to a solar flux value.

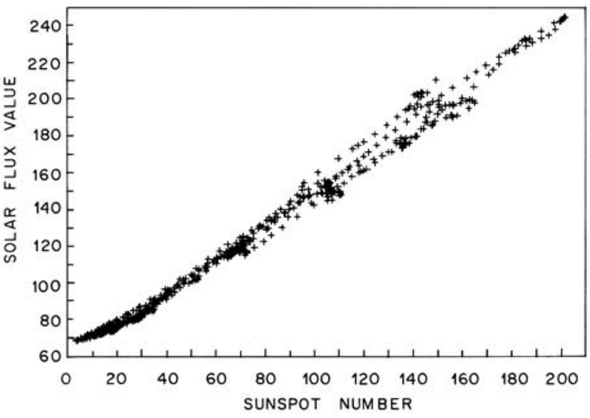

Fig 17—Scatter diagram of smoothed, or 12-month running averages, sunspot numbers versus 2800MHz solar flux values. The correlation of smoothed values is better than for monthly means, shown in Fig 16.

Table 1

Selected Historical Data Showing Inconsistent Correlation Between Sunspot Number and Solar Flux

Month	*Smoothed Sunspot Number*	*Smoothed Solar Flux Value*
May 1953	17.4	75.6
Sept 1965	17.4	78.5
Jul 1985	17.4	74.7
Jun 1969	106.1	151.4
Jul 1969	105.9	151.4
Dec 1982	94.6	151.4
Aug 1948	141.1	180.5
Oct 1959	141.1	192.3
Apr 1979	141.1	180.4
Aug 1981	141.1	203.3

$$F = 63.75 + 0.728\,S + 0.00089\,S^2 \qquad \text{(Eq 5)}$$

where

F = solar flux number

S = smoothed sunspot number

A graphic representation of this equation is given in **Fig 18**. Use this chart to make conversions graphically, rather than by calculations. With the graph, solar flux and sunspot number conversions can be made either way. The equation has been found to yield errors as great as 10% when historical data was examined. (Look at the August 1981 data in Table 1.) Therefore, conversions should be rounded to the nearest whole number, as additional decimal places are unwarranted. To make conversions from flux to sunspot number, the following approximation may be used.

$$S = 33.52\sqrt{85.12 + F} - 408.99 \qquad \text{(Eq 6)}$$

THE IONOSPHERE

There will be inevitable "gray areas" in our discussion of the Earth's atmosphere and the changes wrought in it by the Sun and by associated changes in the Earth's magnetic field. This is not a story that can be told in neat equations, or values carried out to a satisfying number of decimal places. The story must be told, and understood—with its well-known limitations—if we are to put up good antennas and make them serve us well.

Thus far in this chapter we have been concerned with what might be called our "above-ground living space"—that portion of the total atmosphere wherein we can survive without artificial breathing aids, or up to about 6 km (4 miles). The boundary area is a broad one, but life (and radio propagation) undergo basic changes beyond this zone. Somewhat farther out, but still technically within the Earth's atmosphere, the role of the Sun in the wave-propagation picture is a dominant one.

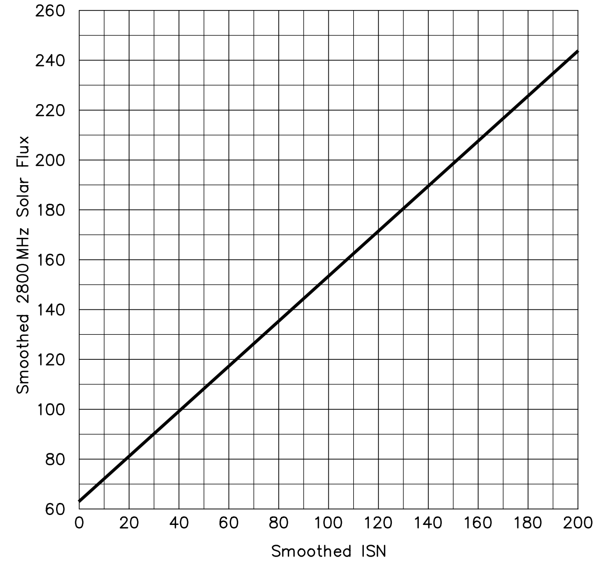

Fig 18—Chart for conversions between smoothed International Sunspot Numbers and smoothed 2800MHz solar flux. This curve is based on the mathematical approximation given in the text.

This is the *ionosphere*—a region where the air pressure is so low that free electrons and ions can move about for some time without getting close enough to recombine into neutral atoms. A radio wave entering this rarefied atmosphere, a region of relatively many free electrons, is affected in the same way as in entering a medium of different dielectric constant—its direction of travel is altered.

Ultraviolet (UV) radiation from the Sun is the primary cause of ionization in the outer regions of the atmosphere, the ones most important for HF propagation. However, there are other forms of solar radiation as well, including both hard and soft x-rays, gamma rays and extreme ultraviolet (EUV). The radiated energy breaks up, or *photoionizes*, atoms and molecules of atmospheric gases into electrons and positively charged ions. The degree of ionization does not increase uniformly with distance from the Earth's surface. Instead there are relatively dense regions (layers) of ionization, each quite thick and more or less parallel to the Earth's surface, at fairly well-defined intervals outward from about 40 to 300 km (25 to 200 miles). These distinct layers are formed due to complex photochemical reactions of the various types of solar radiation with oxygen, ozone, nitrogen and nitrous oxide in the rarefied upper atmosphere.

Ionization is not constant within each layer, but tapers off gradually on either side of the maximum at the center of the layer. The total ionizing energy from the Sun reaching a given point, at a given time, is never constant, so the height and intensity of the ionization in the various regions will also vary. Thus, the practical effect on long-distance communication is an almost continuous variation in signal level, related to the time of day, the season of the year, the distance between the Earth and the Sun, and both short-term and long-term variations in solar activity. It would seem from all this that only the very wise or the very foolish would attempt to predict radio propagation conditions, but it is now possible to do so with a fair chance of success. It is possible to plan antenna designs, particularly the choosing of antenna heights, to exploit known propagation characteristics.

Ionospheric Layer Characteristics

The lowest known ionized region, called the *D layer* (or the *D region*), lies between 60 and 92 km (37 to 57 miles) above the Earth. In this relatively low and dense part of the atmosphere, atoms broken up into ions by sunlight recombine quickly, so the ionization level is directly related to sunlight. It begins at sunrise, peaks at local noon and disappears at sundown. When electrons in this dense medium are set in motion by a passing wave, collisions between particles are so frequent that a major portion of their energy may be used up as heat, as the electrons and disassociated ions recombine.

The probability of collisions depends on the distance an electron travels under the influence of the wave—in other words, on the wavelength. Thus, our 1.8- and 3.5-MHz bands, having the longest wavelengths, suffer the highest daytime absorption loss as they travel through the D layer, particularly for waves that enter the medium at the lowest angles. At times of high solar activity (peak years of the solar cycle) even waves entering the D layer vertically suffer almost total energy absorption around midday, making these bands almost useless for communication over appreciable distances during the hours of high sun. They "go dead" quickly in the morning, but come alive again the same way in late afternoon. The diurnal (daytime) D-layer effect is less at 7 MHz (though still marked), slight at 14 MHz and inconsequential on higher amateur frequencies.

The D region is ineffective in bending HF waves back to Earth, so its role in long-distance communication by amateurs is largely a negative one. It is the principal reason why our frequencies up through the 7-MHz band are useful mainly for short-distance communication during the high-sun hours.

The lowest portion of the ionosphere useful for long-distance communication by amateurs is the *E layer* (also known as the *E region*) about 100 to 115 km (62 to 71 miles) above the Earth. In the E layer, at intermediate atmospheric density, ionization varies with the Sun angle above the horizon, but solar ultraviolet radiation is not the sole ionizing agent. Solar X-rays and meteors entering this portion of the Earth's atmosphere also play a part. Ionization increases rapidly after sunrise, reaches maximum around noon local time, and drops off quickly after sundown. The minimum is after midnight, local time. As with the D layer, the E layer absorbs wave energy in the lower-frequency amateur bands when the Sun angle is high, around mid-day. The other varied effects of E-region ionization will be discussed later.

Most of our long-distance communication capability stems from the tenuous outer reaches of the Earth's atmosphere known as the *F layer*. At heights above 100 miles, ions and electrons recombine more slowly, so the observable effects of the Sun develop more slowly. Also, the region holds its ability to reflect wave energy back to Earth well into the night. The *maximum usable frequency* (MUF) for F-layer propagation on east-west paths thus peaks just after noon at the midpoint, and the minimum occurs after midnight. We'll examine the subject of MUF in more detail later.

Judging what the F layer is doing is by no means that simple, however. The layer height may be from 160 to more than 500 km (100 to over 310 miles), depending on the season of the year, the latitudes, the time of day and, most capricious of all, what the Sun has been doing in the last few minutes and in perhaps the last three days before the attempt is made. The MUF between Eastern US and Europe, for example, has been anything from 7 to 70 MHz, depending on the conditions mentioned above, plus the point in the long-term solar-activity cycle at which the check is made.

During a summer day the F layer may split into two layers. The lower and weaker *F_1 layer*, about 160 km (100 miles) up, has only a minor role, acting more like the E than the *F_2 layer*. At night the F_1 region disappears and the F_2 region height drops somewhat.

Propagation information tailored to amateur needs is transmitted in all information bulletin periods by the ARRL Headquarters station, W1AW. Finally, solar and geomagnetic field data, transmitted hourly and updated eight times daily, are given in brief bulletins carried by the US Time Standard stations, WWV and WWVH, and also on Internet Web sites. But more on these services later.

Bending in the Ionosphere

The degree of bending of a wave path in an ionized layer depends on the density of the ionization and the length of the wave (inversely related to its frequency). The bending at any given frequency or wavelength will increase with increased ionization density and will bend away from the region of most-intense ionization. For a given ionization density, bending increases with wavelength (that is, it decreases with frequency).

Two extremes are thus possible. If the intensity of the ionization is sufficient and the frequency is low enough, even a wave entering the layer perpendicularly will be reflected back to Earth. Conversely, if the frequency is high enough or the ionization decreases to a low-enough density, a condition is reached where the wave angle is not affected enough by the ionosphere to cause a useful portion of the wave energy to return to the Earth. The frequency at which this occurs is called the vertical-incidence *critical frequency.* Each region in the ionosphere has a critical frequency associated with it, and this critical frequency will change depending on the date, time and state of the 11-year solar cycle.

Fig 19 shows a simplified graph of the electron density (in electrons per cubic meter) versus height in the ionosphere (in km) for a particular set of daytime and nighttime conditions. Free electrons are what return the signals you launch into the ionosphere back down to the Earth at some distance from your transmitter—The more free electrons in the ionosphere, the better propagation will be, particularly at higher frequencies.

Electron-density profiles are extremely complicated and vary greatly from one location to the next, depending on a bewildering variety of factors. Of course, this sheer variability makes it all the more interesting and challenging for hams to work each other on ionospheric HF paths!

The following discussion about sounding the ionosphere provides some background information about the scientific instruments used to decipher the highly intricate mechanisms behind ionospheric HF propagation.

SOUNDING THE IONOSPHERE

For many years scientists have *sounded* the ionosphere to determine its communication potential at various elevation angles and frequencies. The word "sound" stems from

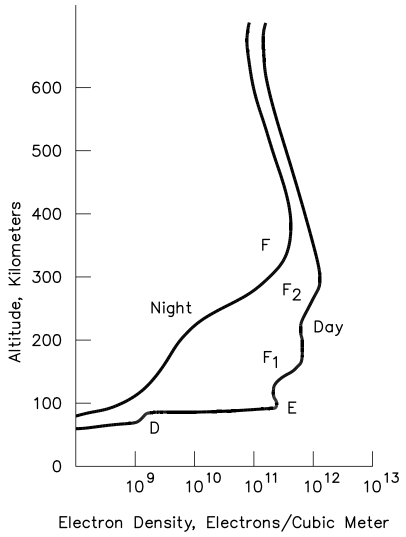

Fig 19—Typical electron densities for nighttime and daytime conditions in the various ionospheric regions.

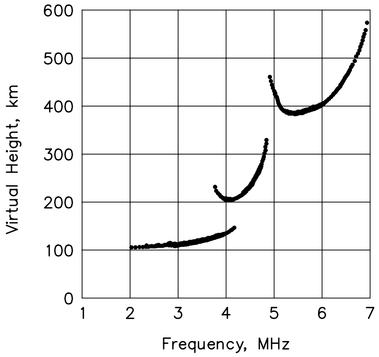

Fig 20—Very simplified ionogram from a vertical-incidence sounder. The lowest trace is for the E region; the middle for the F_1 and the upper trace for the F_2 region.

an old idea—one that has nothing to do with the audio waves that we can hear as "sounds." Long ago, sailors sounded the depths beneath their boats by dropping weighted ropes, calibrated in fathoms, into the water. In a similar fashion, the instrument used to probe the height of the ionosphere is called an *ionosonde*, or ionospheric sounder. It measures distances to various layers by launching a calibrated electronic signal directly up into the ionosphere.

Radar uses the same techniques as ionospheric sounding to detect targets such as airplanes. An ionosonde sends precisely timed pulses into the ionosphere over a range of MF and HF frequencies. The time of reception of an echo reflected from a region in the ionosphere is compared to the time of transmission. The time difference is multiplied by the speed of light to give the apparent distance that the wave has traveled from the transmitter to the ionosphere and back to the receiver. (It is an *apparent* or *virtual distance* because the speed of a wave slows very slightly in the ionosphere, just as the speed of propagation through any medium other than a vacuum slows down because of that medium.)

Another type of ionosonde sweeps the frequency of transmission, from low to high. This is called an "FM-CW," or more colorfully, a "chirp" sounder. Since a received echo takes time to travel from the transmitter up to the reflection point and then back again to the receiver, the echo will be at a lower frequency than the still-moving frequency of the transmitter. The frequency difference is an indication of the height of the echo's reflection off the various ionospheric layers.

Vertical-Incidence Sounders

Most ionosondes are *vertical-incidence sounders*, bouncing their signals perpendicularly off the various ionized regions above it by launching signals straight up into the ionosphere. The ionosonde frequency is swept upwards until echos from the various ionospheric layers disappear, meaning that the critical frequencies for those layers have been exceeded, causing the waves to disappear into space.

Fig 20 shows a highly simplified ionogram for a typical vertical-incidence sounder. The echos at the lowest height at the left-hand side of the plot show that the E region is about 100 km high. The F_1 region shown in the middle of the plot varies from about 200 to 330 km in this example, and the F_2 region ranges from just under 400 km to almost 600 km in height. You can see that the F_1 and F_2 ionospheric regions take a "U" shape, indicating that the electron density varies throughout the layer. In this example, the peak in electron density is at a virtual height of the F_2 region of about 390 km, the lowest point in the F_2 curve.

Scientists can derive a lot of information from a vertical-incidence ionogram, including the critical frequencies for each region, where raising the frequency any higher causes the signals to disappear into space. In Fig 20, the E-region critical frequency (abbreviated f_oE) is about 4.1 MHz. The F_1-region critical frequency (abbreviated f_oF_1) is 4.8 MHz. The F2-region critical frequency (abbreviated f_oF_2) is this simplified diagram is 6.8 MHz.

The observant reader may well be wondering what the subscripted "o" in the abbreviations f_oE, f_oF_1 and f_oF_2 mean. The abbreviation "o" means "ordinary." When an electromagnetic wave is launched into the ionosphere, the Earth's magnetic field splits the wave into two independent waves—the "ordinary" (o) and the "extraordinary" (x) components. The ordinary wave reaches the same height in the ionosphere whether the Earth's magnetic field is present or not, and hence is called "ordinary." The extraordinary wave, however, is greatly affected by the presence of the Earth's magnetic field, in a very complex fashion.

Fig 21 shows an example of an actual ionogram from the vertical Lowell Digisonde at Millstone Hill in Massa-

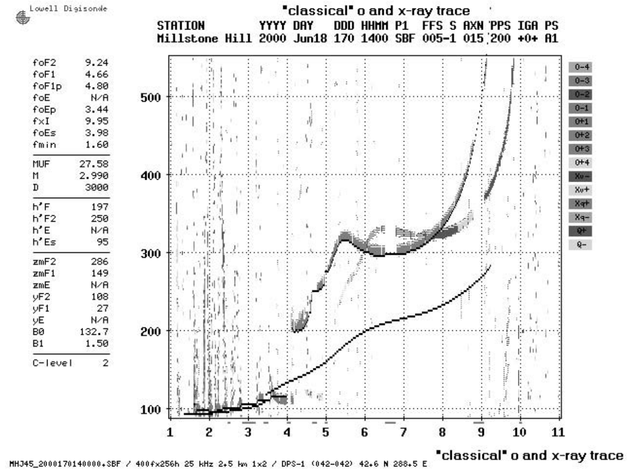

Fig 21—Actual vertical-incidence ionogram from the Lowell Digisonde, owned and operated at Millstone Hill in Massachusetts by MIT. The ordinary (o) and extraordinary (x) traces are shown for heights greater than about 300 km. At the upper left are listed the computer-determined ionospheric parameters, such as f_oF_2 of 9.24 MHz and f_oF_1 at 4.66 MHz.

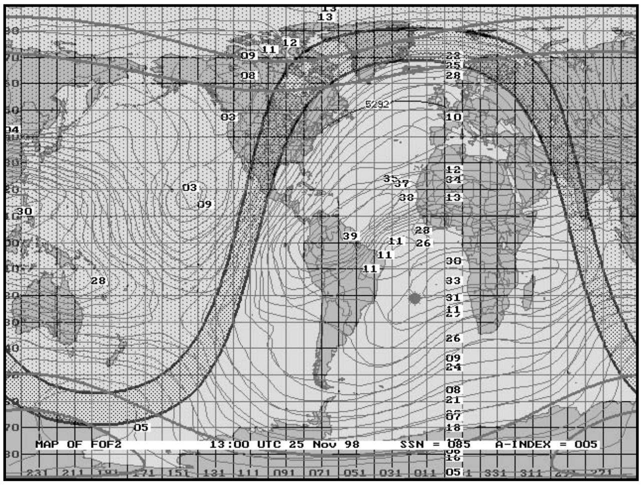

Fig 22—Computer simulation of the f_oF_2 contours for 25 November 1998, for an SSN of 85 and a quiet planetary A_p index of 5. Note the two regions of high f_oF_2 values off the upper and lower west coast of Africa. These are the "equatorial anomalies," regions of high electronic density in the F_2 region that often allow chordal-hop north-south propagation. See also Fig 8. (*PropLab Pro simulation, courtesy of Solar Terrestrial Dispatch.*)

chusetts, owned and operated by the Massachusetts Institute of Technology. This ionogram was made on June 18, 2000, and shows the conditions during a period of very high solar activity. The black-and-white rendition in Fig 21 of the actual color ionogram unfortunately loses some information. However, you can still see that a real ionogram is a lot more complicated looking than the simple simulated one in Fig 20.

The effects of noise and interference from other stations are shown by the many speckled dots appearing in the ionogram. The critical frequencies for various ionospheric layers are listed numerically at the left-hand side of the plot and the signal amplitudes are color-coded by the color bars at the right-hand side of the plot. The x-axis is the frequency, ranging from 1 to 11 MHz.

Compared to the simplified ionogram in Fig 20, Fig 21 shows another trace that appears on the plot from about 5.3 to 9.8 MHz, a trace shifted to the right of the darker ordinary trace. This second trace is the extraordinary (x) wave mentioned above. Since the x and o waves are created by the Earth's magnetic field, the difference in the ordinary and extraordinary traces is about $^1/_2$ the *gyro frequency*, the frequency at which an electron will spiral down a particular magnetic field line. The electron gyro frequency is different at various places around the Earth, being related to the Earth's complicated and changing magnetic field. The extraordinary trace always has a higher critical frequency than the ordinary trace on a vertical-incidence ionogram, and it is considerably weaker than the ordinary trace, especially at frequencies below about 4 MHz because of heavy absorption.

The Big Picture Overhead

There are about 150 vertical-incidence ionosondes around the world. Ionosondes are located on land, even on a number of islands. There are gaps in sounder coverage, however, mainly over large expanses of open ocean. The compilation of all available vertical-incidence data from the worldwide network of ionospheric sounders results in global f_oF_2 maps, such as the map shown in **Fig 22**, a simulation from the highly sophisticated *PropLab Pro* computer program.

This simulation is for 1300 UTC, several hours after East Coast sunrise on Nov 25, 1998, with a high level of solar activity of 85 and a planetary A_p index of 5, indicating calm geomagnetic conditions. The contours of f_oF_2 peak over the ocean off the west coast of Africa at 38 MHz. Over the southern part of Africa, f_oF_2 peaks at 33 MHz.

These two "humps" in f_oF_2 form what is known as the "equatorial anomaly" and are caused by upwelling "fountains" of high electron concentration located in daylight areas about ±20° from the Earth's magnetic dip equator. The equatorial anomaly is important in transequatorial propagation. Those LU stations in Argentina that you can hear on 28 MHz from the US in the late afternoon, even during low portions of the solar cycle when other stations to the south are not coming through, are benefiting from transequatorial propagation, sometimes called "chordal hop" propagation, because signals going through this area remain in the ionosphere without lossy intermediate hops to the ground.

From records of f_oF_2 profiles, the underlying electron densities along a path can be computed. And from the electron density profiles computerized "ray tracing" may be done throughout the ionosphere to determine how a wave propagates from a transmitter to a particular receiver location. *PropLab Pro* can do complex ray tracings that explicitly include the effect of the Earth's magnetic field, even taking into effect ionospheric stormy conditions.

Oblique-Angle Ionospheric Sounding

A more elaborate form of ionospheric sounder is the *oblique ionosonde*. Unlike a vertical-incidence ionospheric sounder, which sends its signals directly overhead, an oblique sounder transmits its pulses obliquely through the ionosphere, recording echos at a receiver located some distance from the transmitter. The transmitter and distant receiver are precisely coordinated in GPS-derived time in modern oblique sounders.

Interpretation of ionograms produced by oblique sounders is considerably more difficult than for vertically incident ones. An oblique ionosonde purposely transmits over a continuous range of elevation angles simultaneously and hence cannot give explicit information about each elevation angle it launches. **Fig 23** shows a typical HF oblique-sounder ionogram for the path from Hawaii to California in March of 1973, during a period of medium-level sunspot activity. The y-axis is calibrated in time delay, in milliseconds. Longer distances involve longer time delays between the start of a transmitted pulse and the reception of the echo. The x-axis in this ionogram is the frequency, just like a vertical-incidence ionogram. Note that the frequency range for this plot extends to 32 MHz, while vertical-incidence ionograms usually don't sweep higher than about 12 MHz.

Six possible modes are shown in this ionogram: $1F_2$, $2F_2$, $3F_2$, $4F_2$ and $5F_2$. These involve multiple modes of propagation (commonly called *hops*) between the ionosphere and reflections from the Earth. For example, at an operating frequency of 14 MHz, there are three modes open during the mid-morning: $2F_2$, $3F_2$ and $4F_2$. We'll discuss multiple hops later in more detail.

The lowest mode, $1F_2$ in Fig 23, employs a single F_2 hop to cover the 3900-km long path from Hawaii to California, but it is only open on 28-MHz. (Note that 3900 km is close to the maximum possible single-hop length for the F_2 region. We'll look at this in more detail later too.) In general, each mode that involves more than a single hop is weaker than a single hop. For example, you can see that the received $5F_2$ echo is weak and broken up because of the accumulation of losses at each ground-level reflection in its five hops, with absorption in the ionosphere all along its complicated path to the receiver.

The trace labeled "FOT" is the *frequency of optimum traffic*, considered the most reliable frequency for communications on this particular circuit and date/time. In this example, the FOT would be near the 21-MHz amateur band.

Another interesting point in Fig 23 is labeled "High Angle Ray." This refers to the Pedersen ray. Before we go into more details about the Pedersen high-angle wave, we need to examine how launch angles affect the way waves are propagated through the ionosphere.

Fig 24 shows a highly simplified situation, with a single ionospheric layer and a smooth Earth. This illustrates several important facts about antenna design for long-distance communication. In Fig 24, Wave #1 is launched at the lowest elevation angle (that is, most nearly horizontal to the horizon). Wave #1 manages to travel from the transmitter to the receiving location at point C in a single hop.

Wave #2 is launched at a higher elevation angle than Wave #1, and penetrates further into the ionospheric layer before it is refracted enough to return to Earth. The ground distance covered from the transmitter to point B is less for

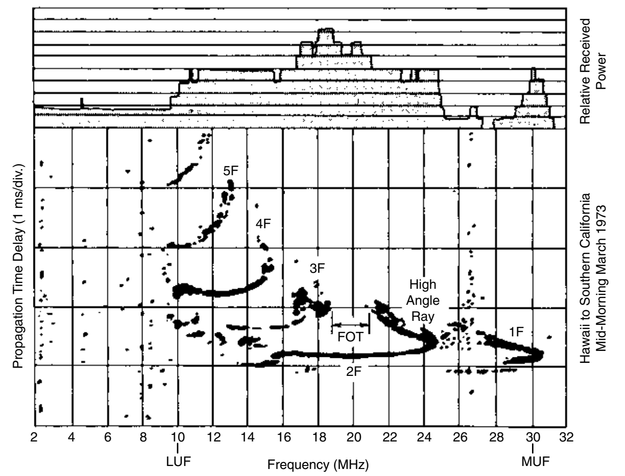

Fig 23—HF oblique-sounder ionogram. This is a typical chirpsounder measurement on a 2500-mile path from Hawaii to southern California during midmorning in March at a medium level of solar activity. Six possible modes (hops) are shown. The "FOT" is the frequency of optimum traffic, considered most reliable for this path/time.

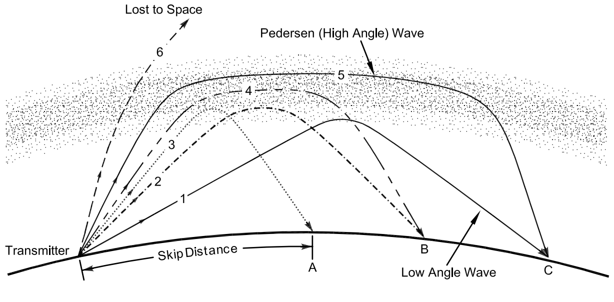

Fig 24—Very simplified smooth-Earth/ionosphere diagram showing how the ground range from transmitter to receiver can vary as the elevation angle is gradually raised. The Pedersen wave, launched at a relatively high angle, has the same ground range as the low-angle wave #1, but is weaker because it travels for a long distance in the ionosphere.

Wave #2 than for lower-angle Wave #1. Wave #3 is launched at a still-higher elevation angle. Like Wave #2 before it, Wave #3 penetrates further into the ionosphere and covers less ground downrange than #2.

Now, we see something very interesting happening for Wave #4, whose launch elevation angle is still higher than #3. Wave #4 penetrates even higher into the ionosphere than #3, reaching the highest level of ionization in our theoretical ionospheric layer, where it is finally refracted sufficiently to bend down to Earth. Wave #4 manages to arrive at the same point B as Wave #2, which was launched at a much lower elevation angle.

In other words, in the sequence from #1 to #3 we have been continually *increasing* the elevation launch angle and the ground range covered from the transmitter to the return of the signal back to Earth has been continually *decreasing*. However, starting with Wave #4, the ground range starts to *increase* with increased elevation angle. A further increase in the elevation angle causes Wave #5 to travel for an even longer distance through the ionosphere, exiting finally at point C, the same ground distance as lowest-angle Wave #1.

Finally, increasing the elevation angle even further results in Wave #6 being lost to outer space because the ionization in the layer is insufficient to bend the wave back to Earth. In other words, Wave #6 has exceeded the *critical angle* for this hypothetical ionospheric layer and this frequency of operation.

Both Waves #4 and #5 in Fig 24 are called "high-angle" or Pedersen waves. Because Wave #5 has traveled a greater distance through the ionosphere, it is always weaker than Wave #1, the one launched at the lowest elevation angle. Pedersen waves are usually not very stable, since small changes in elevation angle can result in large changes in the ground range that these high-angle waves cover.

SKIP DISTANCE

Fig 24 shows that we can communicate with the point on the Earth labeled "A" (where Wave #3 arrives), but not any closer to our transmitter site. When the critical angle is less than 90° (that is, directly overhead) there will always be a region around the transmitting site where an ionospherically propagated signal cannot be heard, or is heard weakly. This area lies between the outer limit of the ground-wave range and the inner edge of energy return from the ionosphere. It is called the *skip zone*, and the distance between the originating site and the beginning of the ionospheric return is called the skip distance. This terminology should not to be confused with ham jargon such as "the skip is in," referring to the fact that a band is open for sky-wave propagation.

The signal may often be heard to some extent within the skip zone, through various forms of scattering (discussed in detail later), but it will ordinarily be marginal in strength. When the skip distance is short, both ground-wave and sky-wave signals may be received near the transmitter. In such instances the sky wave frequently is stronger than the ground wave, even as close as a few miles from the transmitter. The ionosphere is an efficient communication medium under favorable conditions. Comparatively, the ground wave is not.

If the radio wave leaves the Earth at a radiation angle of zero degrees, just at the horizon, the maximum distance that may be reached under usual ionospheric conditions in the F_2 region is about 4000 km (2500 miles).

MULTI-HOP PROPAGATION

As mentioned previously in the discussion about Fig 24, the Earth itself can act as a reflector for radio waves, resulting a multiple hops. Thus, a radio signal can be reflected from the reception point on the Earth back into the ionosphere, reaching the Earth a second time at a still more-distant point. This effect is illustrated in **Fig 25**, where a single ionospheric layer is depicted, although this time we show both the layer and the Earth beneath it as curved rather than flat. The wave identified as "Critical Angle" travels from the transmitter via the ionosphere to point A, in the center of the drawing, where it is reflected upwards and travels through the ionosphere to point B, at the right. This shows a two-hop signal.

As in the simplified case in Fig 24, the distance at which a ray eventually reaches the Earth depends on the launch elevation angle at which it left the transmitting antenna. You have some control of the launch angle by adjusting the height of the antennas you use, as described in Chapter 3, The Effects of Ground.

The information in Fig 25 is greatly simplified. On actual communication paths the picture is complicated by many factors. One is that the transmitted energy spreads over a considerable area after it leaves the antenna. Even with an antenna array having the sharpest practical beam pattern, there is what might be described as a *cone of radiation* centered on the wave lines (rays) shown in the drawing. The reflection/refraction in the ionosphere is also highly variable, and is the cause of considerable spreading and scattering.

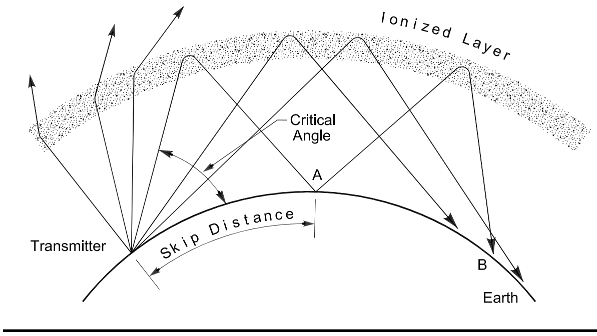

Fig 25—Behavior of waves encountering a simple curved ionospheric layer over a curved Earth. Rays entering the ionized region at angles above the critical angle are not bent enough to be returned to Earth, and are lost to space. Waves entering at angles below the critical angle reach the Earth at increasingly greater distances as the launch angle approaches the horizontal. The maximum distance that may normally be covered in a single hop is 4000 km. Greater distances are covered with multiple hops.

Under some conditions it is possible for as many as four or five signal hops to occur over a radio path, as illustrated by the oblique ionogram in Fig 23. But no more than two or three hops is the norm. In this way, HF communication can be conducted over thousands of miles.

An important point should be recognized with regard to signal hopping. A significant loss of signal occurs with each hop. The D and E layers of the ionosphere absorb energy from signals as they pass through, and the ionosphere tends to scatter the radio energy in various directions, rather than confining it in a tight bundle. The roughness of the Earth's surface also scatters the energy at a reflection point.

Assuming that both waves do reach point B in Fig 25, the low-angle wave will contain more energy at point B. This wave passes through the lower layers just twice, compared to the higher-angle route, which must pass through these layers four times, plus encountering an Earth reflection. Measurements indicate that although there can be great variation in the relative strengths of the two signals—the one-hop signal will generally be from 7 to 10 dB stronger. The nature of the terrain at the mid-path reflection point for the two-hop wave, the angle at which the wave is reflected from the Earth, and the condition of the ionosphere in the vicinity of all the refraction points are the primary factors in determining the signal-strength ratio.

The loss per hop becomes significant at greater distances. It is because of these losses that no more than four or five propagation hops are useful; the received signal becomes too weak to be usable over more hops. Although modes other than signal hopping also account for the propagation of radio waves over thousands of miles, backscatter studies of actual radio propagation have displayed signals with as many as five hops. So the hopping mode is arguably the most prevalent method for long-distance communication.

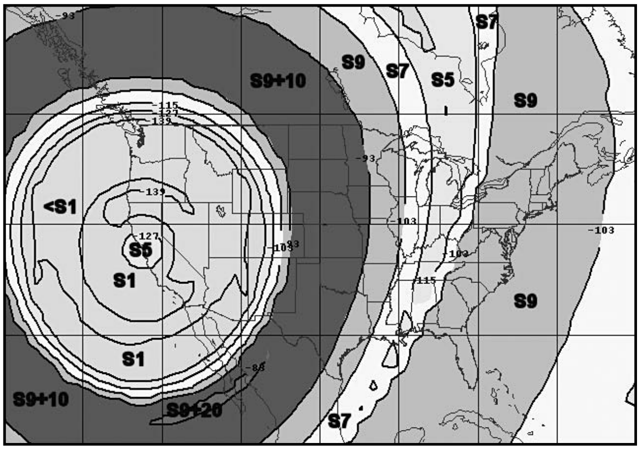

Fig 26—Modified *VOAAREA* plot for 21.2 MHz from San Francisco to the rest of the US, annotated with signal levels in S units, as well as signal contours in dBW (dB below a watt). Antennas are assumed to be 3-element Yagis at 55 feet above flat ground; the transmitter power is 1500 W; the month is November with SSN = 50, a moderate level of solar activity, at 22 UTC. The most obvious feature is the large "skip zone" centered on the transmitter in San Francisco, extending almost a $^{1}/_{3}$ of the distance across the US.

Fig 26 shows another way of looking at propagation—a *geographic area* look. Fig 26 shows 15-meter signal levels across the US as they propagate from a transmitting station in San Francisco. This simulation of propagation conditions is for the month of November, with a medium level of solar activity (SSN = 50) at 22 UTC. Fig 26 was created using the *VOAAREA* software program, part of the *VOACAP* software suite. Transmitter power is assumed to be 1500 W, with 3-element Yagis, 55 feet high, at the transmitter and at each receiving location.

From the transmitter out to about 50 miles, signals are moderate, at about S5 on an S meter. Beyond that coverage area to almost $^{1}/_{3}$ of the way across the country (to Colorado), there is a large and distinctive skip zone, where only very weak signals return to Earth (S1 or less). Beyond Colorado, signals rapidly build up to S9+10 dB across the middle of the US, falling to S9 and then to S7 in the vicinity of Chicago, Illinois. Beyond Chicago, the signals drop to S5 in a swath from Michigan and part of Ohio down to Alabama. All along the US East Coast, signals come back strong at S9.

The reason why the signals in Fig 26 drop down to S5 in the Midwest is that the necessary elevation angles to cover this region in a single F2 hop are extremely low even at a moderate level of solar activity. To achieve launch angles as low as 1° requires either very high antenna heights or a high mountaintop location. Beyond the Midwest, out to the US East Coast, two F2 hops are required, with higher elevation angles and hence greater antenna gain for moderate antenna heights.

Non-Hopping Propagation Modes

Present propagation theory holds that for communication distances of many thousands of kilometers, signals do not always hop in relatively short increments from ionosphere-to-Earth-to-ionosphere and so forth along the entire path. Instead, the wave is thought to propagate *inside* the ionosphere throughout some portion of the path length, tending to be *ducted* in the ionized layer.

As was shown in Fig 24, the high-angle Pedersen ray can also penetrate an ionospheric layer farther than lower-angle rays. In the less-densely ionized upper edge of the layer, the amount of refraction is less, nearly equaling the curvature of the layer itself as it encircles the Earth.

Non-hopping theory of long-distance propagation is further supported by studies of travel times for signals that go completely around the world. The time required is significantly less than would be necessary to hop between the Earth and the ionosphere 10 or more times while circling the Earth.

Propagation between two points thousands of kilometers apart may in fact consist of a combination of ducting and hopping. It may involve combinations of refractions from the E layer and the F layer. Despite all the complex factors involved, most long-distance propagation can be seen to follow certain general rules. Thus, much commercial and military point-to-point communication over long distances employs antennas designed to make maximum use of known radiation angles and layer heights, even on paths where multihop propagation is assumed.

In amateur work, however, we usually try for the lowest practical radiation angle, hoping to keep reflection losses to a minimum. Years of amateur experience have shown this to be a decided advantage under all usual conditions.

The geometry of propagation by means of the F$_2$ layer limits our maximum distance along the Earth's surface to about 4000 km (2500 miles) for a single hop. For higher radiation angles, this same distance may require two or more hops (with higher reflection loss). And fewer hops are better, in most cases. If you have a nearby neighbor who consistently outperforms you on the longer paths, a radiation angle difference in his favor is probably the reason.

MAXIMUM USABLE FREQUENCY

The vertical-incidence critical frequency is the *maximum usable frequency* for local sky-wave high-angle communication. It is also useful in the selection of optimum working frequencies and the determination of the maximum usable frequency for distant points at a given time. The abbreviation "MUF" for maximum usable frequency will be used hereafter.

In geographic middle latitudes, the vertical-incident critical frequency ranges between about 1 and 4 MHz for the E layer, and between 2 and 13 MHz for the F$_2$ layer. The lowest figures are for nighttime conditions in the lowest years of the solar cycle. The highest are for the daytime hours in the years of high solar activity. These are average figures. Critical frequencies have reached as high as 20 MHz briefly during exceptionally high solar activity in the middle latitudes. As was pointed out earlier in Fig 22, f$_o$ F$_2$ levels approaching 40 MHz are possible at low latitudes.

While vertical-incidence critical frequencies are interesting from a scientific point of view, hams are far more concerned about how we can exploit propagation conditions to communicate, preferably at long distances. The MUF for a 4000-km (2500 miles) distance is about 3.5 times the vertical-incidence critical f$_o$ F$_2$ frequency existing at the path midpoint. For one-hop signals, if a uniform ionosphere is assumed, the MUF decreases with shorter distances along the path. This is true because the higher-frequency waves must be launched at higher elevation angles for shorter ranges, and at these launch angles they are not bent sufficiently to reach the Earth. Thus, a lower frequency (where more bending occurs) must be used.

Precisely speaking, a maximum usable frequency or MUF is defined for communication between two specific points on the Earth's surface, for the conditions existing at the time, including the minimum elevation angle that the station can launch at the frequency in use. (This practical form of MUF is sometimes called the *operational MUF*). At the same time and for the same conditions, the MUF from either of these two points to a third point may be different.

Table 2

Time and Frequency Stations Useful for Propagation Monitoring

Call	*Frequency (MHz)*	*Location*
WWV	2.5, 5, 10, 15, 20	Ft Collins, Colorado
WWVH	Same as WWV but no 20	Kekaha, Kauai, Hawaii
CHU	3.330, 7.335, 14.670	Ottawa, Ontario, Canada
RID	5.004, 10.004, 15.004	Irkutsk, USSR*
RWM	4.996, 9.996, 14.996	Novosibirsk, USSR
VNG	2.5, 5, 8.634, 12.984, 16	Lyndhurst, Australia
BPM	5, 5.43, 9.351, 10, 15	Xiang, China
BSF	15	Taoyuan, Taiwan
JJY	2.5, 5, 8, 10, 15	Tokyo, Japan
LOL	5, 10, 15	Buenos Aires, Argentina

*The call, taken from an international table, may not be the one used during actual transmission. Locations and frequencies appear to be accurate as provided.

Therefore, the MUF cannot be expressed broadly as a single frequency, even for any given location at a particular time. The ionosphere is never uniform, and in fact at a given time and for a fixed distance, the MUF changes significantly with changes in compass direction for almost any point on the Earth. Under usual conditions, the MUF will always be highest in the direction toward the Sun—to the east in the morning, to the south at noon (from northern latitudes), and to the west in the afternoon and evening.

For the strongest signals at the greatest distance, especially where the limited power levels of the Amateur Radio Service are concerned, it is important to work fairly near the MUF. It is at these frequencies where signals suffer the least loss. The MUFs can be estimated with sufficient accuracy by using the prediction charts that appear on the ARRL Web site (**www.arrl.org/qst/propcharts/**) or by using a computer prediction program. The CD-ROM bundled in the back of this book contains detailed and summary tables for more than 175 transmitting locations around the world. (See section on "What HF Bands Are Open—Where and When?" later in this chapter.)

MUFs can also be observed, with the use of a continuous coverage communications receiver. Frequencies up to the MUFs are in round-the-clock use today. When you "run out of signals" while tuning upward in frequency from your favorite ham band, you have a pretty good clue as to which band is going to work well, right then. Of course, it helps to know the direction to the transmitters whose signals you are hearing. Shortwave broadcasters know what frequencies to use, and you can hear them anywhere, if conditions are good. Time-and-frequency stations are also excellent indicators, since they operate around the clock. See **Table 2**. WWV is also a reliable source of propagation data, hourly, as discussed in more detail later in this chapter.

The value of working near the MUF is two-fold. Under undisturbed conditions, the absorption loss decreases proportional to the square of a change in frequency. For example, the absorption loss is four times higher at 14 MHz than it is at 28 MHz. Perhaps more important, the hop distance is considerably greater as the MUF is approached. A transcontinental contact is thus much more likely to be made on a single hop on 28 MHz than on 14 MHz, so the higher frequency will give the stronger signal most of the time. The strong-signal reputation of the 28-MHz band is founded on this fact.

LOWEST USABLE FREQUENCY

There is also a lower limit to the range of frequencies that provide useful communication between two given points by way of the ionosphere. *Lowest usable frequency* is abbreviated LUF. If it were possible to start near the MUF and work gradually lower in frequency, the signal would decrease in strength and eventually would disappear into the ever-present "background noise." This happens because signal absorption increases proportional to the square of the lowering of the frequency. The frequency nearest the point where reception became unusable would be the LUF. It is not likely that you would want to work at the LUF, although reception could be improved if the station could increase power by a considerable amount, or if larger antennas could be used at both ends of the path.

For example, when solar activity is very high at the peak of a solar cycle, the LUF often rises higher than 14 MHz on the morning Eastern US-to-Europe path on 20 meters. Just before sunrise in the US, the 20-meter band will be first to open to Europe, followed shortly by 15 meters, and then 10 meters as the Sun rises further. By mid-morning, however, when 10 and 15 meters are both wide open, 20 meters will become very marginal to Europe, even when both sides are running maximum legal power levels. By contrast, stations on 10 meters can be worked readily with a transmitter power of only 1 or 2 watts, indicating the wide range between the LUF and the MUF.

Frequently, the *window* between the LUF and the MUF for two fixed points is very narrow, and there may be no amateur frequencies available inside the window. On occasion the LUF may be higher than the MUF between two points. This means that, for the highest possible frequency that will propagate through the ionosphere for that path, the absorption is so great as to make even that frequency unusable. Under these conditions it is impossible to establish amateur sky-wave communication between those two points, no matter what frequency is used. (It would normally be possible, however, to communicate between either point and other points on some frequency under the existing conditions.) Conditions when amateur sky-wave communication is impossible between two fixed points occur commonly for long distances where the total path is in darkness, and for very great distances in the daytime during periods of low solar activity.

Fig 27 shows a typical propagation prediction from the ARRLWeb members-only site (**www.arrl.org/qst/propcharts/**, previously from the "How's DX" column in *QST*). In this instance, the MUF and the LUF lines are blurred together at about 10 UTC, meaning that the statistical likelihood of any amateur frequency being open for that particular path at that particular time was not very good. Later on, after about 11 UTC, the gap between the MUF and LUF increased, indicating that the higher bands would be open on that path.

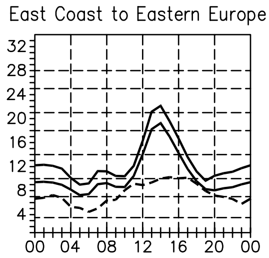

Fig 27—Propagation prediction chart for East Coast of US to Europe. This appeared in December 1994 *QST*, where an average 2800-MHz (10.7-cm) solar flux of 83 was assumed for the mid-December to mid-January period. On 10% of these days, the highest frequency propagated was predicted at least as high as the uppermost curve (the Highest Possible Frequency, or HPF, approximately 21 MHz), and for 50% of the days as high as the middle curve, the MUF. The broken lines show the Lowest Usable Frequency (LUF) for a 1500-W CW transmitter.

DISTURBED IONOSPHERIC CONDITIONS

So far, we have discussed the Earth's ionosphere when conditions at the Sun are undisturbed. There are three general types of major disturbances on the Sun that can affect radio propagation here on the Earth. On the air, you may hear people grousing about *Solar Flares*, *Coronal Holes* or *Sudden Disappearing Filaments*, especially when propagation conditions are not good. Each of these disturbances causes both electromagnetic radiation and ejection of material from the Sun.

Solar Flares

Solar flares are cataclysmic eruptions that suddenly release huge amounts of energy, including sustained, high-energy bursts of radiation from VLF to X-ray frequencies and vast amounts of solar material. Most solar flares occur around the peak of the 11-year solar cycle.

The first Earthly indication of a huge flare is often a visible brightness near a sunspot group, along with increases in UV, X-ray radiation and VHF radio noise. If the geometry between the Sun and Earth is right, intense X-ray radiation takes eight minutes, traveling the 93 million miles to Earth at the speed of light. The sudden increase in X-ray energy can immediately increase RF absorption in the Earth's lowest ionospheric layers, causing a phenomenon known as a *Sudden Ionospheric Disturbance* (SID).

An SID affects all HF communications on the sunlit side of the Earth. Signals in the 2 to 30-MHz range may disappear entirely, and even most background noise may cease in extreme cases. When you experience a big SID, your first inclination may be to look outside to see if your antenna fell down! SIDs may last up to an hour before ionospheric conditions temporarily return to normal.

Between 45 minutes and 2 hours after an SID begins, particles from the mass eruption on the Sun may begin to arrive. These high-energy particles are mainly protons and they can penetrate the ionosphere at the Earth's magnetic poles, where intense ionization can occur, with attendant absorption of HF signals propagating through the polar regions. This is called a *Polar Cap Absorption* (PCA) event and it may last for several days. A PCA results in spectacular auroral displays at high latitudes.

Coronal Holes

As described earlier in the section dealing with auroral propagation at VHF, a second major solar disturbance is a so-called "coronal hole" in the Sun's outer layer (the *corona*). Temperatures in the corona can be more than four million °C over an active sunspot region but more typically are about two million °C. A coronal hole is an area of somewhat lower temperature. Solar-terrestrial scientists have a number of competing theories about how coronal holes are formed.

Matter ejected through this "hole" takes the form of a *plasma*, a highly ionized gas made up of electrons, protons and neutral particles, traveling at speeds up to 1,000 km per second (2 million miles per hour). The plasma becomes part of the solar wind and can affect the Earth's magnetic field, but only if the Sun-Earth geometry is right. A plasma has a very interesting and somewhat bizarre ability. It can lock-in the orientation of the magnetic field where it originates and carry it outwards into space. However, unless the locked-in magnetic field orientation is aligned properly with the Earth's magnetic field, even a large plasma mass may not severely disrupt our magnetosphere, and thence our ionosphere.

Presently, we don't have the ability to predict very long in advance when the Sun might erupt in a disturbance that results in Earthly propagation problems. The SOHO satellite can help determine whether a mass ejection is heading towards Earth, and the ACE satellite about 1 million miles away from Earth can give about an hour's warning whether the imbedded magnetic field in a mass ejection from the Sun might impact the Earth's magnetosphere, causing propagation problems for hams.

Statistically, coronal holes tend to occur most often during the declining phase of the 11-year solar cycle and they can last for a number of solar rotations. This means that a coronal hole can be a "recurring coronal hole," disrupting communications for several days about the same time each month for as long as a year, or even more.

Sudden Disappearing Filaments

A sudden disappearing filament (SDF) is the third major category of solar disturbance that can affect propagation. SDFs take their names from the manner in which they suddenly arch upward from the Sun's surface, spewing huge amounts of matter as plasma out into space in the solar wind. They tend to occur mostly during the rising phase of the 11-year solar cycle.

IONOSPHERIC STORMS

When the conditions are right, a flare, coronal hole or an SDF can launch a plasma cloud into the solar wind, resulting in an *ionospheric storm* here on Earth. Unlike a hurricane or a winter Nor'easter storm in New England, an ionospheric storm is not something we can see with our eyes or feel on our skins. We can't easily measure things occurring in the ionosphere some 200 miles overhead. However, we can see the indirect effects of an ionospheric storm on magnetic instruments located on the Earth's surface, because disturbances in the ionosphere are closely related to the Earth's magnetic field. The term *Geomagnetic Storm* ("Geo" means "Earth" in Greek) is used almost synonymously with ionospheric storm.

During a ionospheric storm, we may experience extraordinary radio noise and interference, especially at HF. You may hear solar radio emissions as increases of noise at VHF. A geomagnetic storm generally adds noise and weakens or disrupts ionospheric propagation for several days. Transpolar signals at 14 MHz or higher may be particularly weak, with a peculiar hollow sound or flutter—even more than normal for transpolar signals.

Depending on the severity of the disturbance to the Earth's geomagnetic field and the consequent disturbance of the ionosphere, propagation may be disrupted completely or it might be at least degraded for a period of time that ranges from a day to three or four days before returning to normal propagation conditions.

What can we do about the solar disturbances and related disturbed ionospheric propagation on Earth? The truth is that we are powerless faced with the truly awesome forces of solar disturbances like flares, coronal holes or sudden disappearing filaments. Perhaps there is some comfort, however, in understanding what has happened to cause our HF bands to be so poor. And as a definite consolation, conditions on the VHF bands are often exceptionally good just when HF propagation is remarkably poor due to solar disturbances. VHF operators enthusiastically look forward to conditions when they can engage in auroral communications—exactly the kind of conditions that have HF operators scratching their heads, wondering where the ionosphere went.

ELEVATION ANGLES FOR HF COMMUNICATION

It was shown in connection with Fig 25 that the distance at which a ray returns to Earth depends on the eleva-

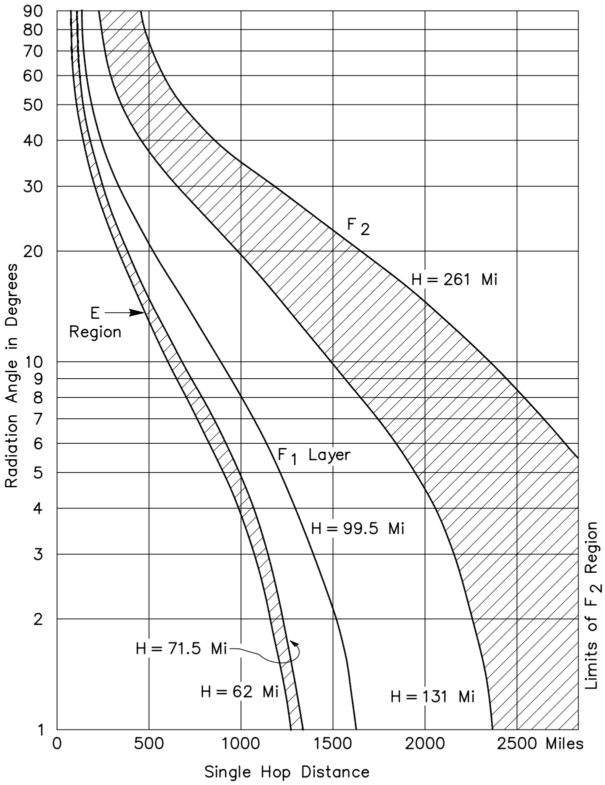

Fig 28—Distance plotted against wave angle (one-hop transmission) for the nominal range of heights for the E and F2 layers, and for the F1 layer.

tion angle at which it left the Earth (also known by other names: takeoff, launch or wave angle). Chapter 3, The Effects of Ground, in this book deals with the effects of local terrain, describing how the elevation angle of a horizontally polarized antenna is determined mainly by its height above the ground.

Although it is not shown specifically in Fig 25, propagation distance also depends on the layer height at the time, as well as the elevation angle. As you can probably imagine, the layer height is a very complex function of the state of the ionosphere and the Earth's geomagnetic field. There is a large difference in the distance covered in a single hop, depending on the height of the E or the F_2 layer. The maximum single-hop distance by the E layer is about 2000 km (1250 miles) or about half the maximum distance via the F_2 layer. Practical communicating distances for single-hop E or F layer work at various wave angles are shown in graphic form in **Fig 28**.

Actual communication experience usually does not fit the simple patterns shown in Fig 25. Propagation by means of the ionosphere is an enormously complicated business (which makes it all the more intriguing and challenging to radio amateurs, of course), even when the Sun is not in a disturbed state. Until the appearance of sophisticated computer models of the ionosphere, there was little definitive information available to guide the radio amateur in the design of his antenna systems for optimal performance over all portions of the 11-year solar cycle. Elevation angle information that had appeared for many years in *The ARRL Antenna Book* was measured for only one transmitting path, during the lowest portion of Solar Cycle 17 in 1934.

The *IONCAP* Computer Propagation Model

Since the 1960s several agencies of the US government have been working on a detailed computer program that models the complex workings of the ionosphere. The program has been dubbed *IONCAP*, short for "Ionospheric Communications Analysis and Prediction Program." *IONCAP* was originally written for a mainframe computer, but later versions have been rewritten to allow them to be run by high-performance personal computers. *IONCAP* incorporates a detailed database covering almost three complete solar cycles. The program allows the operator to specify a wide range of parameters, including detailed antenna models for multiple frequency ranges, noise models tailored to specific local environments (from low-noise rural to noisy residential QTHs), minimum elevation angles suitable for a particular location and antenna system, different months and UTC times, maximum levels of multipath distortion, and finally solar activity levels, to name the most significant of a bewildering array of options.

While *IONCAP* has a well-justified reputation for being very *unfriendly* to use, due to its mainframe, non-interactive background, it is also the one ionospheric model most highly regarded for its accuracy and flexibility, both by amateurs and professionals alike. It is the program used for many years to produce the long-term MUF charts formerly included in the "How's DX" monthly column of *QST* and now available on the Members Only ARRLWeb page.

IONCAP is not well suited for short-term forecasts of propagation conditions based on the latest solar indices received from WWV. It is an excellent tool, however, for long-range, detailed planning of antenna systems and shortwave transmitter installations, such as that for the Voice of America, or for radio amateurs. See the section later in this chapter describing other computer programs that can be used for short-term, interactive propagation predictions.

IONCAP/VOACAP Parameters

The elevation-angle statistical information contained in this section was compiled from thousands of *VOACAP* runs (an improved version of *IONCAP* developed by scientists from VOA, the Voice of America). These runs were done for a number of different transmitting locations throughout the world to important DX locations throughout the world.

Some assumptions were needed for setting *VOACAP* parameters. The transmitting and receiving sites were all assumed to be located on flat ground, with "average" ground conductivity and dielectric constant. Each site was assumed to have a clear shot to the horizon, with a minimum elevation angle less than or equal to 1°. Electrical noise at each

Table 3
Boston, Massachusetts, to All of Europe

Elev	*80 m*	*40 m*	*30 m*	*20 m*	*17 m*	*15 m*	*12 m*	*10 m*
1	4.1	9.6	4.6	1.7	2.1	4.4	5.5	7.2
2	0.8	2.3	7.2	1.4	2.8	2.8	3.7	5.3
3	0.3	0.7	4.3	3.1	2.4	2.2	4.4	7.9
4	0.5	4.1	8.7	11.6	12.2	9.4	8.1	3.9
5	4.6	4.8	7.5	12.7	14.3	13.1	9.2	11.2
6	7.1	8.9	5.5	9.2	9.6	12.2	9.2	7.2
7	8.5	6.9	7.2	4.6	7.9	7.4	10.0	5.9
8	5.1	7.0	5.4	3.2	5.9	7.4	4.8	6.6
9	3.3	5.6	3.2	3.1	2.1	3.9	8.1	9.2
10	1.0	4.0	7.9	6.3	5.1	3.7	11.1	6.6
11	1.9	3.8	9.7	10.2	7.2	5.4	3.7	7.9
12	5.6	3.4	4.8	8.5	6.9	7.4	4.8	6.6
13	11.0	3.0	2.4	4.1	5.9	4.6	3.3	2.6
14	7.6	4.8	2.0	2.7	3.8	3.9	6.3	5.9
15	5.3	7.9	2.0	1.5	2.4	1.7	1.5	2.0
16	2.8	6.4	3.8	2.9	1.5	1.3	2.6	2.6
17	5.0	3.4	4.5	3.1	1.0	1.5	0.0	0.0
18	4.2	2.0	3.1	3.1	2.0	2.2	1.8	1.3
19	5.7	1.4	1.4	2.3	1.3	0.7	0.0	0.0
20	6.6	1.4	1.2	1.8	1.1	1.3	0.7	0.0
21	4.4	1.4	0.5	0.8	0.7	0.7	0.4	0.0
22	2.3	2.4	1.0	1.1	0.6	1.3	0.7	0.0
23	1.3	1.8	0.1	0.3	0.1	0.0	0.0	0.0
24	0.6	1.0	0.5	0.5	0.4	0.7	0.0	0.0
25	0.3	0.8	0.3	0.1	0.4	0.0	0.0	0.0
26	0.0	0.5	0.7	0.2	0.1	0.4	0.0	0.0
27	0.1	0.1	0.1	0.2	0.1	0.2	0.0	0.0
28	0.0	0.3	0.1	0.2	0.0	0.2	0.0	0.0
29	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.0
30	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
31	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
33	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Percentage of time a particular frequency band is open on this specific propagation path.

receiving location was also assumed to be very low.

Transmitting and receiving antennas for the 3.5 to 30-MHz frequency range were specified to be isotropic-type antennas, but with +6 dBi gain, representing a good amateur antenna on each frequency band. These theoretical antennas radiate uniformly from the horizon, up to 90° directly overhead. With response patterns like this, these are obviously not real-world antennas. They do, however, allow the computer program to explore all possible modes and elevation angles.

Looking at the Elevation-Angle Statistical Data

Table 3 shows detailed statistical elevation informa-

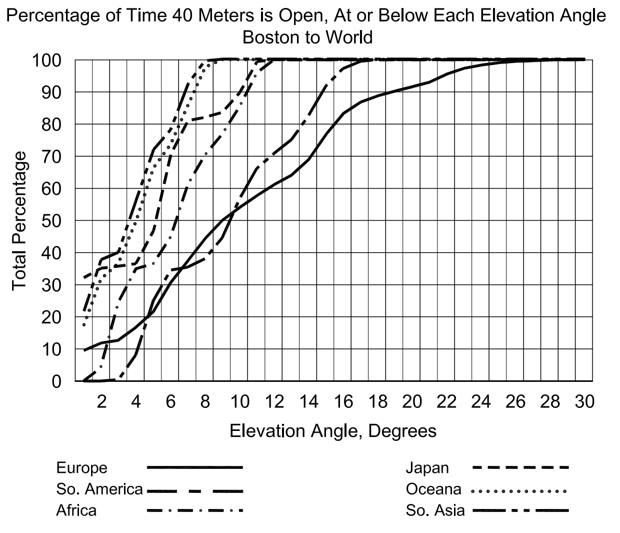

Fig 29—The cumulative distribution function showing the total percentage of time that 40 meters is open, at or below each elevation angle, from Boston to the world. For example, 50% of the time the band is open to Europe from Boston, it is at 10° or less. The angles for DX work are indeed low.

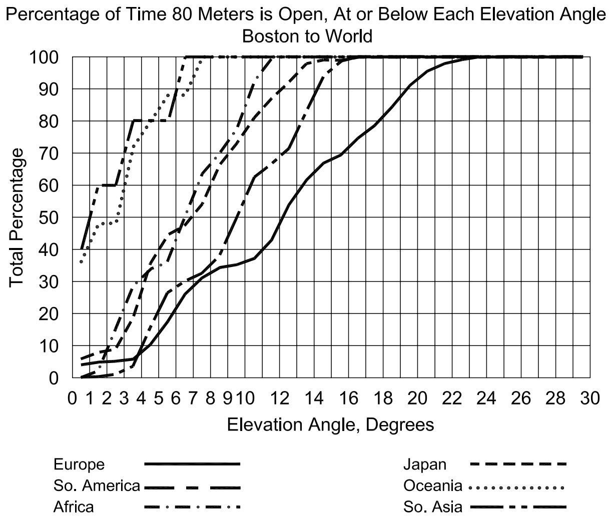

Fig 30—The cumulative distribution function showing the total percentage of time that 80 meters is open, at or below each elevation angle, from Boston to the world. For example, 50% of the time the band is open to Europe from Boston, it is at 13° or less.

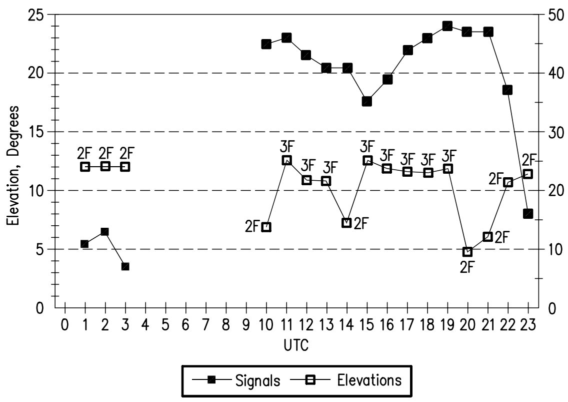

Fig 31—Overlay of signals and elevation angles, together with hop-mode information. This is for one month, October, at one level of solar activity, SSN=70, for the path from Newington, CT, to London, England. The mode of propagation does not closely follow the elevation angle. From 15 to 19 UTC the mode is 3F2 hops, and the elevation angle is approximately 12°. The same elevation angle is required from 23 to 03 UTC, but here the mode is 2F2 hops.

tion for the path from Boston, Massachusetts, near ARRL HQ in Newington, Connecticut, to all of Europe. The data incorporated into Table 3 shows the percentage of time versus elevation angle for all HF bands from 80 meters to 10 meters, over all portions of the 11-year solar cycle. The CD-ROM accompanying this book contains more tables such as this for more than 150 transmitting sites around the world. These tables are used by the *HFTA* program (and earlier *YT* program) and can also be imported into many programs, such as word processors or spreadsheets. Six important areas throughout the world are covered, one per table: all of Europe (from London, England, to Kiev, Ukraine), the Far East (centered on Japan), South America (Paraguay), Oceania (Melbourne, Australia), Southern Africa (Zambia) and South Asia (New Delhi, India).

You may be surprised to see in Table 3 that angles lower than 10° dominate the possible range of incoming angles for this moderate-distance path from New England to Europe. In fact, 1.7% of all the times when the 20-meter band is open to Europe, the takeoff angle is as low as 1°. You should recognize that very few real-world 20-meter antennas achieve much gain at such an extremely low angle—unless they just happen to be mounted about 400 feet high over flat ground or else are located on the top of a tall, steep mountain.

The situation is even more dramatic on 40 and 80 meters. **Fig 29** shows the "cumulative distribution function" of the total percentage of time (derived from Table 3) when 40 meters is open from Boston to the rest of the world, plotted against the elevation angle. For example, into Europe from Boston, 50% of the time when the band is open, it is at 10° or less. Into Japan from Boston, the statistics are even more revealing: 50% of the time when the band is open, the angle is 6° or less, and 90% of the time the angle is 13° or less!

Fig 30 shows the same sort of information for 80 meters from Boston to the world. For 50% of the time from Boston to Europe the elevation angle is 13° or less; at the 90% level the angle is 20° or less. For the path to Japan on 80 meters from Boston, 50% of the time the angle is 8° or less; at the 90% level, the angle is 13° or less. Now, to achieve peak gain on 80 meters at an elevation angle of 8° over flat land, a horizontally polarized antenna must be 500 feet high. You can begin to see why verticals can do very well on long-distance contacts on 80 meters, even when they are mounted over poorly conducting, rocky ground. Clearly, low angles are very important for successful DXing.

The Ionosphere Controls Propagation

You should always remember that it is the *ionosphere* that controls the elevation angles, *not* the transmitting antenna. The elevation response of a particular antenna only determines how strong or weak a signal is, at whatever angle (or angles) the ionosphere is supporting at that particular instant, for that propagation path and for that frequency.

If only one propagation mode is possible at a particular time, and if the elevation angle for that one mode happens to be 5°, then your antenna will have to work satisfactorily at that very low angle or else you won't be able to communicate. For example, if your low dipole has a gain of –10 dBi at 5°, compared to your friend's Yagi on a mountain top with +10 dBi gain at 5°, then you will be down 20 dB compared to his signal. It's not that the elevation angle is somehow *too low*—the real problem here is that you don't have *enough gain* at that particular angle where the ionosphere is supporting propagation. Many "flatlanders" can vividly recall the times when their mountain-top friends could easily work DX stations, while they couldn't even hear a whisper.

Looking at the Data—Further Cautions

A single propagation mode is quite common at the opening and the closing of daytime bands like 20, 15 or 10 meters, when the elevation angle is often lower (but not always) than when the band is wide open. The lower-frequency bands tend to support multiple propagation modes simultaneously. For example, **Fig 31** plots the signal strength (in dBμV) and the elevation angle for the dominant mode (with the strongest signal) over a 24-hour period from Newington to London in October, for a medium-level SSN = 70. The morning opening at 10 UTC starts out with a two-hop $2F_2$ mode (labeled 2F) at an elevation angle of 6°. By 11 UTC the mode has changed to a three-hop $3F_2$ (labeled 3F) at a 12° elevation angle. The band starts to close down with weaker signals after about 23 UTC. Note that this path actually supports both $2F_2$ and $3F_2$ modes most of the time. Either mode may be stronger than the other, depending on the particular time of day.

It is tempting to think that two-hop signals always occur at lower elevation launch angles, while three-hop signals require higher elevation angles. In reality, the detailed workings of the ionosphere are enormously complicated. From 22 UTC to 03 UTC, the elevation angles are higher than 11° for $2F_2$ hops. During much of the morning and early afternoon in Newington (from 11 to 13 UTC, and from 15 to 19 UTC), the angles are also higher than 11°. However, $3F_2$ hops are involved during these periods of time. The number of hops is not directly related to the elevation angles needed—changing layer heights account for this.

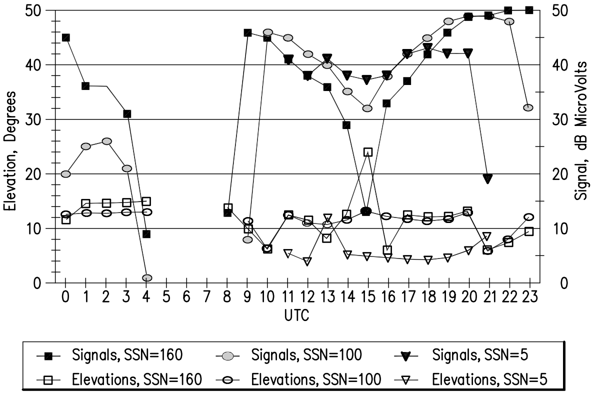

Fig 32—October 20-meter signals and elevation angles for the full range of solar activity, from W1 to England. The elevation angle does not closely follow the level of solar activity. What is important in designing a station capable of covering all levels of solar activity is to have flexibility in antenna elevation pattern response — to cover a wide range of possible angles.

Note that starting around 15 UTC, the mid-morning 20-meter "slump" (down some 10 dB from peak signal level) is caused by higher levels of mainly E-layer absorption when the Sun is high overhead. This condition favors higher elevation angles, since signals launched at lower angles must travel for a longer time through the lossy lower layer.

How does the situation change with different levels of solar activity? **Fig 32** overlays predicted signals and elevation angles for three levels of solar activity in October, again for the Newington-London path. Fig 32 shows the mid-morning slump dramatically when the solar activity is at

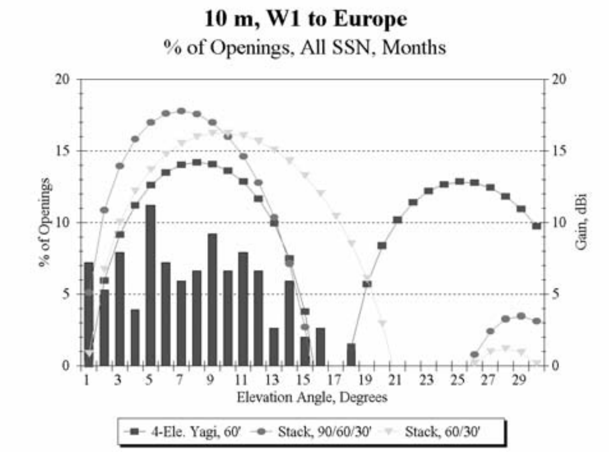

Fig 33—10-meter graph of the percentage of all openings versus elevation angles, together with overlay of elevation patterns over flat ground for three 10-meter antenna systems. Stacked antennas have wider "footprints" in elevation angle coverage for this example from New England to Europe.

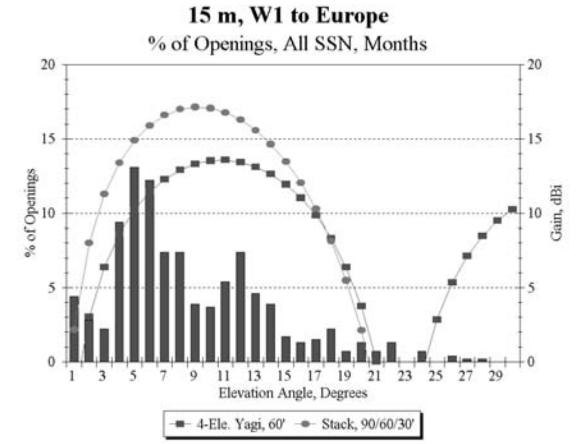

Fig 34—15-meter graph of the percentage of all openings versus elevation angles, together with overlay of elevation patterns over flat ground for two 15-meter antenna systems. Like 10 meters, 15-meter stacked antennas have wider footprints in elevation angle coverage for this example from New England to Europe.

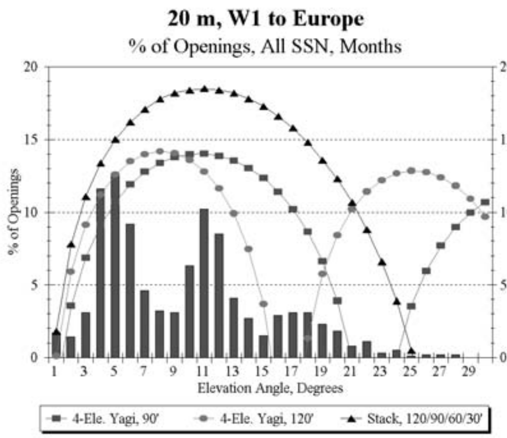

Fig 35—20-meter graph of the percentage of all openings from New England to Europe versus elevation angles, together with overlay of elevation patterns over flat ground for three 20-meter antenna systems.

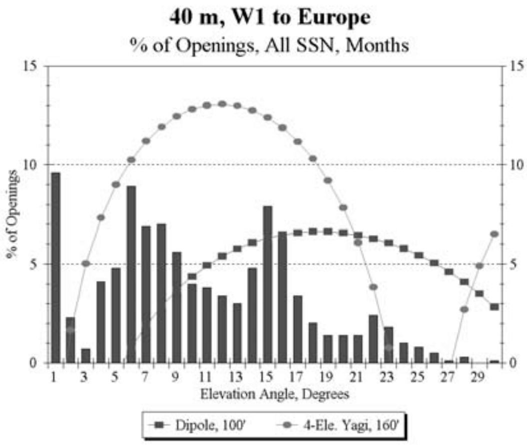

Fig 36—40-meter graph of the percentage of all openings from New England to Europe versus elevation angles, together with overlays of elevation patterns over flat ground for a 100-foot high dipole and a large 4-element Yagi at 160 feet. Achieving gain at very low elevation angles requires very high heights above ground.

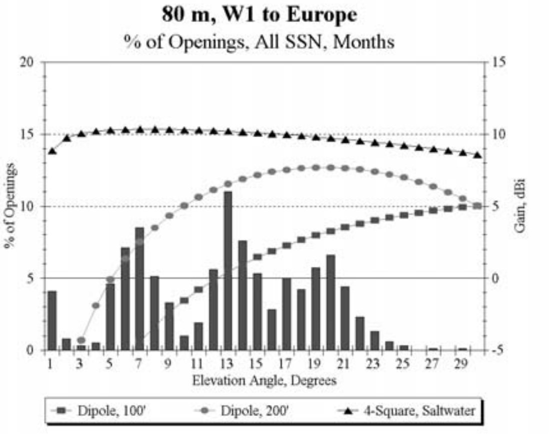

Fig 37—80-meter graph of the percentage of all openings from New England to Europe versus elevation angles, together with overlay of elevation patterns over flat ground for dipoles at two different heights. The 200-foot-high dipole clearly covers the necessary elevation angles better than does the 100-foot-high dipole, although a Four Square vertical array located over saltwater is even better for all angles needed.

a very high level, represented by SSN = 160. At 15 UTC, the signal level drops 35 dB from peak level, and the elevation angle rises all the way to 24°. By the way, as a percentage of all possible openings, the 24° angle occurs only rarely, 0.5% of the time. It barely shows up as a blip in Table 3. Elevation angles are *not* closely related to the level of solar activity.

IONCAP/VOACAP demonstrates that elevation angles do not follow neat, easily identified patterns, even over a 24-hour period—much less over all portions of the solar cycle. Merely looking at the percentage of all openings versus elevation angle, as shown in Table 3, does not tell the whole story, although it is probably the most statistically valid approach to station design, and possibly the most emotionally satisfying approach too. Neither is the whole story revealed by looking only at a snapshot of elevation angles versus time for one particular month, or for one solar activity level.

What is important to recognize is that the most effective antenna system will be one that can cover the *full range* of elevation angles, over the whole spectrum of solar activity, even if the actual angle in use at any one moment in time may not be easy to determine. For this particular path, from New England to all of Europe, an ideal antenna would have equal response over the full range of angles from 1° to 28°. Unfortunately, real-world antennas have a tough time covering such a wide range of elevation angles equally well.

Antenna Elevation Patterns

Figs 33 through **37** show overlays of the same sort of elevation angle information listed in Table 3, together with the elevation response patterns for typical antennas for the HF amateur bands 80, 40, 20, 15 and 10 meters. For example, Fig 34 shows an overlay for 20 meters, with three different types of 20-meter antennas. These are a 4-element Yagi at 90 feet, a 4-element Yagi at 120 feet and a large stack of four 4-element Yagis located at 120, 90, 60 and 30 feet. Each antenna is assumed to be mounted over flat ground. Placement on a hill with a long slope in the direction of interest would lower the required elevation angle by the amount of the hill's slope. For example, if a 10° launch angle is desired, and the antenna is placed on a hill with a slope of –5°, the antenna itself should be designed for a height that would optimize the response at 15° over flat ground—one wavelength high.

In Fig 35, the large stack of four 20-meter Yagis over flat ground comes closest to being ideal, but even this large array will not work well for that very small percentage of time when the angle needed is higher than about 20°. Some hams might conclude that the tiny percentage of time when the angles are very high doesn't justify an antenna tailored for that response. However, when that new DX country pops up on a band, or when a rare multiplier shows up in a contest, doesn't it always seem that the desired signal only comes in at some angle your antenna doesn't cover well? What do you do then, if your only antenna happens to be a large stack?

The answer to this, perhaps unique, high-angle problem lies in switching to using only the top antenna in the stack. In this example, the second elevation lobe of the 120-foot high antenna would cover the angles from 20° to 30° well, much better than the stack does. Note that the top antenna by itself would not be ideal for all conditions. It is simply too high much of the time when the elevation angles are higher than about 12°. The experience of many amateurs on the US East Coast with high 20-meter antennas bears this out—they find that 60 to 90-foot high antennas are far more consistent performers into Europe.

ONE-WAY PROPAGATION

On occasion a signal may be started on the way back toward the Earth by reflection from the F region, only to come down onto the top of the E region and be reflected back up again. This set of conditions is one possible explanation for the often-reported phenomenon called *one-way skip*. The reverse path may not necessarily have the same multilayer characteristic. The effect is more often a difference in the signal strengths, rather than a complete lack of signal in one direction, and many times there may be local noises that mask signals at one end of the path. It is important to remember these sorts of possibilities when a long-distance test with a new antenna system yields apparently conflicting results. Even many tests, on paths of different lengths and headings, may provide data that are difficult to understand. Communication by way of the ionosphere is not always a source of consistent answers to antenna questions.

Fig 37 shows the 80-meter path from New England to Europe with three different antennas. A really high dipole at a height of 200 feet above flat ground would certainly be an impressive antenna. But it would still be overshadowed dramatically by a Four-Square vertical array, at least at the low elevation angles needed often on this path. This is predicated on the Four Square being located over salt water, which provides a virtually perfect RF ground. At an elevation angle of 7°, the Four Square has 7 dB more gain than the 200-foot high dipole.

SHORT OR LONG PATH?

Propagation between any two points on the Earth's surface is usually by the shortest direct route—the *great-circle path* found by stretching a string tightly between the two points on a globe. If an elastic band going completely around the globe in a straight line is substituted for the string, it will show another great-circle path, going "the long way around." The long path may serve for communication over the desired circuit when conditions are favorable along the longer route. There may be times when communication is possible over the long path but not possible at all over the short path. Especially if there is knowledge of this potential at both ends of the circuit, long-path communication may work very well. Cooperation is almost essential, because both the aiming of directional antennas and the timing of the attempts must be right for any worthwhile result. The *IONCAP/VOACAP* computations in the preceding tables were made for short-path azimuths only.

Sunlight is a required element in long-haul communication via the F layer above about 10 MHz. This fact tends to define long-path timing and antenna aiming. Both are essentially the reverse of the "normal" for a given circuit. We know also that salt-water paths work better than overland ones. This can be significant in long-path work.

We can better understand several aspects of long-path propagation if you become accustomed to thinking of the Earth as a ball. This is easy if you use a globe frequently. A

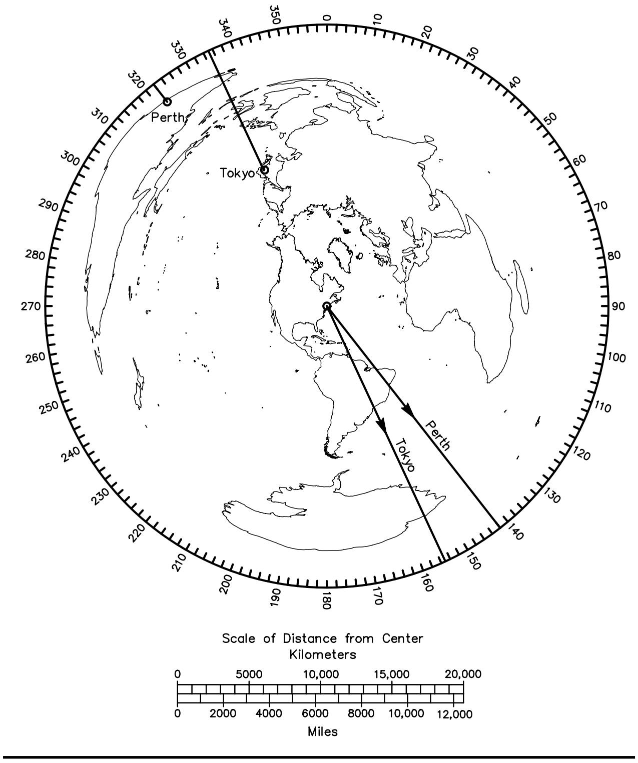

Fig 38—K5ZI's computer-generated azimuthal-equidistant projection centered on Newington, Connecticut. (See Bibliography for ordering information.) Land masses and information showing long paths to Perth and Tokyo have been added. Notice that the paths in both cases lie almost entirely over water, rather than over land masses.

flat map of the world, of the azimuthal-equidistant projection type, is a useful substitute. The ARRL World Map is one, centered on Wichita, Kansas. A similar world map prepared by K5ZI and centered on Newington, Connecticut, is shown in **Fig 38**. These help to clarify paths involving those areas of the world.

Long-Path Examples

There are numerous long-path routes well known to DX-minded amateurs. Two long paths that work frequently and well when 28 MHz is open from the northeastern US are New England to Perth, Western Australia, and New England to Tokyo. Although they represent different beam headings and distances, they share some favorable conditions. By the long path, Perth is close to halfway around the world; Tokyo is about three-quarters of the way. On 28 MHz, both areas come through in the early daylight hours, Eastern Time, but not necessarily on the same days. Both paths are at their best around the equinoxes. (The sunlight is more uniformly distributed over transequatorial paths at these times.) Probably the factor that most favors both is the nature of the first part of the trip at the US end. To work Perth by way of long path, northeastern US antennas are aimed southeast, out over salt water for thousands of miles—the best low-loss start a signal could have. It is salt water essentially all the way, and the distance, about 13,000 miles, is not too much greater than the "short" path.

The long path to Japan is more toward the south, but still with no major land mass at the early reflection points. It is much longer, however, than that to Western Australia. Japanese signals are more limited in number on the long path than on the short, and signals on the average somewhat weaker, probably because of the greater distance.

On the short path, an amateur in the Perth area is looking at the worst conditions—away from the ocean, and out across the huge land mass of North America, unlikely to provide strong ground reflections. The short paths to both Japan and Western Australia, from most of the eastern half of North America, are hardly favorable. The first hop comes down in various western areas likely to be desert or mountains, or both, and not favored as reflection points.

A word of caution: Don't count on the long-path signals always coming in on the same beam heading. There can be notable differences in the line of propagation via the ionosphere on even relatively short distances. There can be more variations on long path, especially on circuits close to halfway around the world. Remember, for a point exactly halfway around, all directions of the compass represent great-circle paths.

FADING

When all the variable factors in long-distance HF communication are taken in account, it is not surprising that signals vary in strength during almost every contact beyond the local range. In VHF communication we can also encounter some fading at distances greater than just to the visible horizon. These are mainly the result of changes in the temperature and moisture content of the air in the first few thousand feet above the ground.

On paths covered by HF ionospheric modes, the causes of fading are very complex—constantly changing layer height and density, random polarization shift, portions of the signal arriving out of phase, and so on. The energy arriving at the receiving antenna has components that have been acted upon differently by the ionosphere. Often the fading is very different for small changes in frequency. With a signal of a wideband nature, such as high-quality FM, or even double-sideband AM, the sidebands may have different fading rates from each other, or from the carrier. This causes severe distortion, resulting in what is termed *selective fading*. The effects are greatly reduced (but still present to some extent) when single-sideband (SSB) is used. Some immunity from fading during reception (but not to the distortion induced by selective fading) can be had by using two or more receivers on separate antennas, preferably with different polarizations, and combining the receiver outputs in what is known as a *diversity* receiving system.

OTHER PROPAGATION MODES

In propagation literature there is a tendency to treat the various propagation modes as if they were separate and distinct phenomena. This they may be at times, but often there is a shifting from one to another, or a mixture of two or more kinds of propagation affecting communication at one time. In the upper part of the usual frequency range for F-region work, for example, there may be enough tropospheric bending at one end (or both ends) to have an appreciable effect on the usable path length. There is the frequent combination of E and F-region propagation in long-distance work. And in the case of the E region, there are various causes of ionization that have very different effects on communication. Finally, there are weak-signal variations of both tropospheric and ionospheric modes, lumped under the term "scatter." We look at these phenomena separately here, but in practice we have to deal with them in combination, more often than not.

Sporadic E (E_s)

First, note that this is *E-subscript-s,* a usefully descriptive term, wrongly written "Es" so often that it is sometimes called "ease," which is certainly not descriptive. *Sporadic E* is ionization at E-layer height, but of different origin and communication potential from the E layer that affects mainly our lower amateur frequencies.

The formative mechanism for sporadic E is believed to be wind shear. This explains ambient ionization being distributed and compressed into a ledge of high density, without the need for production of extra ionization. Neutral winds of high velocity, flowing in opposite directions at slightly different altitudes, produce shears. In the presence of the Earth's magnetic field, the ions are collected at a particular altitude, forming a thin, overdense layer. Data from rockets entering E_s regions confirm the electron density, wind velocities and height parameters.

The ionization is formed in clouds of high density, lasting only a few hours at a time and distributed randomly. They vary in density and, in the middle latitudes in the Northern Hemisphere, move rapidly from southeast to northwest. Although E_s can develop at any time, it is most prevalent in the Northern Hemisphere between May and August, with a minor season about half as long beginning in December (the summer and winter solstices). The seasons and distribution in the Southern Hemisphere are not so well known. Australia and New Zealand seem to have conditions much like those in the US, but with the length of the seasons reversed, of course. Much of what is known about E_s came as the result of amateur pioneering in the VHF range.

Correlation of E_s openings with observed natural phenomena, including sunspot activity, is not readily apparent, although there is a meteorological tie-in with high-altitude winds. There is also a form of E_s, mainly in the northern part of the north temperate zone, that is associated with auroral phenomena.

At the peak of the long E_s season, most commonly in late June and early July, ionization becomes extremely dense and widespread. This extends the usable range from the more common "single-hop" maximum of about 1400 miles to "double-hop" distances, mostly 1400 to 2500 miles. With 50-MHz techniques and interest improving in recent years, it has been shown that distances considerably beyond 2500 miles can be covered. There is also an E_s "link-up" possibility with other modes, believed to be involved in some 50-MHz work between antipodal points, or even long-path communication beyond 12,500 miles.

When E_s is particularly strong and widespread, even the HF bands can suddenly go *short skip* producing exceptionally strong signals from distances that would normally be in the no-signal "skip zone." Editor N6BV distinctly remembers a spectacular 20-meter E_s opening in September 1994, during the "Hiram Percy Maxim/125" anniversary celebration, when he was living in New Hampshire. Signals on 20 meters were 30 to 40 dB over S9 from all along the Eastern Seaboard, from W2 to W4. One exasperated W3 complained that he had been calling in the huge pileup for 20 minutes. N6BV glanced at the S meter and saw that the W3 was 20 dB over S9, normally a very strong 20-meter SSB signal, but not when almost everybody else was 40 dB over S9!

Such short-skip conditions caused by Sporadic E are more common on 10 meters than they are on 15 or 20 meters. They can result in excellent transAtlantic 10-meter openings during the summer months—when 10 meters is not normally open for F_2 ionospheric propagation.

The MUF for E_s is not known precisely. It was long thought to be around 100 MHz, but in the last 25 years or so there have been thousands of 144-MHz contacts during the summer E_s season. Presumably, the possibility also exists at 222 MHz. The skip distance at 144 MHz does average much longer than at 50 MHz, and the openings are usually brief and extremely variable.

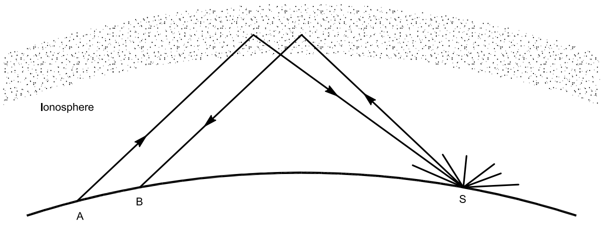

Fig 39—Schematic of a simple backscatter path. Stations A and B are too close to make contact via normal F-layer ionospheric refraction. Signals scattered back from a distant point on the Earth's surface (S), often the ocean, may be accessible to both A and B to create a backscatter circuit. (*Courtesy of The ARRL Handbook.*)

The terms "single" and "double" hop may not be accurate technically, since it is likely that cloud-to-cloud paths are involved. There may also be "no-hop" E_s. At times the very high ionization density produces critical frequencies up to the 50-MHz region, with no skip distance at all. It is often said that the E_s mode is a great equalizer. With the reflecting region practically overhead, even a simple dipole close to the ground may do as well over a few hundred miles as a large stacked antenna array designed for low-angle radiation. It's a great mode for low power and simple antennas on 28 and 50 MHz.

HF Scatter Modes

The term "skip zone" (where no signals are heard) should not be taken too literally. Two stations communicating over a single ionospheric hop can be heard to some degree by other stations at almost any point along the way, unless the two are running low power and using simple antennas. Some of the wave energy is *scattered* in all directions, including back to the starting point and farther.

Backscatter functions like a sort of HF ionospheric radar. **Fig 39** shows a schematic for a simple backscatter path. The signal launched from point A travels through the ionosphere back to earth at Point S, the scattering point. Here, the rough terrain of the land scatters signals in many directions, one of which propagates a weak signal back through the ionosphere to land at point B. Point B would normally be in the no-signal skip zone between A and S. Because backscatter signals arrive from multiple directions, through various paths through the ionosphere, they have a characteristic "hollow" sound, much like you get when you talk into a paper tube with its many internal reflections.

Because backscatter involves mainly scattering from the Earth at the point where the strong ionospherically propagated signal comes down, it is a part of HF over-the-horizon radar techniques. (The infamous 1970s-era "woodpecker" was an over-the-horizon HF radar.) Amateurs using sounding techniques have shown that you can tell to what part of the world a band is usable (single-hop F) by probing the backscatter with a directive antenna and high transmitter power, even when the Earth contact point is open ocean. In fact, that's where the mode is at its best, because ocean waves can be efficient backscatter reflectors.

Backscatter is very useful on 28 MHz, particularly when that band seems dead simply because nobody is active in the right places. The mode keeps the 10-meter band lively in the low years of the solar cycle, thanks to the never-say-die attitude of some users. The mode is also an invaluable tool of 50-MHz DX aspirants, in the high years of the sunspot cycle, for the same reasons. On a high-MUF morning, hundreds of 6-meter beams may zero in on a hot spot somewhere in the Caribbean or South Atlantic, where there is no land, let alone other 6-meter stations—keeping in contact while they wait for the band to open to a place where there is somebody.

Sidescatter is similar to backscatter, except the ground scatter zone is off the direct line between participants. A typical example, often observed during the lowest years of the solar cycle, is communication on 28 MHz between the eastern US (and adjacent areas of Canada) and much of the European continent. Often, this may start as "backscatter chatter" between Europeans whose antennas are turned toward the Azores. Then suddenly the North Americans join the fun, perhaps for only a few minutes, but sometimes much longer, with beams also pointed toward the Azores. Duration of the game can be extended, at times, by careful reorientation of antennas at both ends, as with backscatter. The secret, of course, is to keep hitting the highest-MUF area of the ionosphere and the most favorable ground-reflection points.

The favorable route is usually, but not always, south of the great-circle heading (for stations in the Northern Hemisphere). There can also be sidescatter from the auroral regions. Sidescatter signals are stronger than backscatter signals using the same general area of ground scattering.

Sidescatter signals have been observed frequently on the l4-MHz band, and can take place on any band where there is a large window between the MUF and the LUF. For sidescatter communications to occur, the thing to look for is a common area to which the band is open from both ends of the path (the Azores, in the above example), when there is no direct-path opening. It helps if the common area is in the open ocean, where there is less scattering loss than over land.

GRAY-LINE PROPAGATION

The *gray line*, sometimes called the *twilight zone*, is a band around the Earth between the Sunlit portion and darkness. Astronomers call this the *terminator*. The terminator is a somewhat diffused region because the Earth's atmosphere tends to scatter the light into the darkness. **Fig 40** illustrates the gray line. Notice that on one side of the Earth, the gray line is coming into daylight (sunrise), and on the other side it is coming into darkness (sunset).

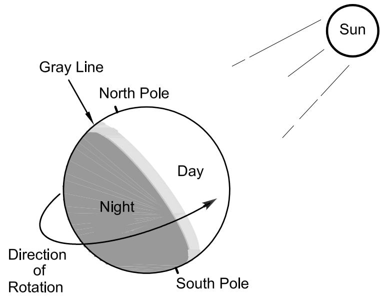

Fig 40—The gray line or terminator is a transition region between daylight and darkness. One side of the Earth is coming into sunrise, and the other is just past sunset.

Propagation along the gray line is very efficient, particularly on the lower bands, especially on 80 or 160 meters, so greater distances can be covered than might be expected for the frequency in use. One major reason for this is that the D layer, which absorbs HF signals, disappears rapidly on the sunset side of the gray line, and has not yet built up on the sunrise side.

The gray line runs generally north and south, but varies as much as 23° either side of the north-south line. This variation is caused by the tilt of the Earth's axis relative to its orbital plane around the Sun. The gray line will be exactly north and south at the equinoxes (March 21 and September 21). On the first day of Northern Hemisphere summer, June 21, it is tilted to the maximum of 23° one way, and on December 21, the first day of winter, it is tilted 23° the other way.

To an observer on the Earth, the direction of the terminator is always at right angles to the direction of the Sun at sunrise or sunset. It is important to note that, except at the equinoxes, the gray-line direction will be different at sunrise from that at sunset. This means you can work different areas of the world in the evening than you worked in the morning.

It isn't necessary to be located inside the twilight zone in order to take advantage of gray-line propagation. The effects can be used to advantage before sunrise and after sunset. This is because the Sun "rises" earlier and "sets" later on the ionospheric layers than it does on the Earth below.

What HF Bands Are Open—Where and When?

The CD-ROM included at the back of this book includes summary and detailed propagation predictions for more than 150 transmitting locations around the world. This propagation data was calculated using *CapMAN*, an upgraded variety of the mainframe propagation program *IONCAP*. The predictions were done for default antennas and powers that are representative of a "big-gun" station. Of course, not everyone has a big-gun station in his/her backyard, but this represents what the ultimate possibilities are, statistically speaking. After all, if the bands aren't open for the big guns, they are unlikely to be open for the "little pistols" too.

Let's see how propagation is affected if the smoothed sunspot number is 0 (corresponding to a smoothed solar flux of about 65), which is classified as a "Very Low" level of solar activity. And we'll examine the situation for a sunspot number of 100 (a smoothed solar flux of 150), which is typical of a "Very High" portion of the solar cycle.

Five-Band Summary Predictions

Tables 4 and **5** are Summary tables showing the predicted signal levels (in S units) from Boston, Massachusetts, to the rest of the world for the month of January. The Boston transmitting site is representative of the entire New England area of the USA. The target geographic receiving regions for the major HF bands from 80 through 10 meters are tabulated versus UTC (Universal Coordinated Time) in hours. Table 4 represents a Very Low level of solar activity, while Table 5 is for a Very High level of solar activity.

Each transmitting location is organized by six levels of solar activity over the whole 11-year solar cycle:

- VL (Very Low: SSN between 0 to 20)
- LO (Low: SSN between 20 to 40)
- ME (Medium: SSN between 40 to 60)
- HI (High: SSN between 60 to 100)
- VH (Very High: SSN between 100 to 150)
- UH (Ultra High: SSN greater than 150)

The receiving geographic regions for each frequency band are abbreviated:

Table 4

Printout of summary propagation table for Boston to the rest of the world, for a Very Low level of solar activity in the month of January. The abbreviations for the target geographic areas are: EU = Europe, FE = Far East, SA = South America, AF = Africa, AS = south Asia, OC = Oceania, and NA = North America.

```
Jan., MA (Boston), for SSN = Very Low, Sigs in S-Units. By N6BV, ARRL.
           80 Meters             40 Meters             20 Meters             15 Meters             10 Meters
UTC  EU FE SA AF AS OC NA  EU FE SA AF AS OC NA  EU FE SA AF AS OC NA  EU FE SA AF AS OC NA  EU FE SA AF AS OC NA  UTC
  0  9  -  9+ 9  9  -  9+  9  8  9+ 9+ 9  2  9+  -  8  9+ 7  4  8  9+  -  -  -  -  -  -  1   -  -  -  -  -  -  2     0
  1  9  -  9+ 9  9  -  9+  9  6  9+ 9+ 9+ 6  9+  -  4  9  4  2  6  9+  -  -  -  -  -  -  1   -  -  -  -  -  -  2     1
  2  9  -  9+ 9+ 8  1  9+  9  6  9+ 9+ 9  8  9+  -  1  8  1  2  3  9+  -  -  -  -  -  -  1   -  -  -  -  -  -  2     2
  3  9  -  9+ 9+ 8  6  9+  9  6  9+ 9+ 9  8  9+  -  -  8  2  2  -  9+  -  -  -  -  -  -  1   -  -  -  -  -  -  2     3
  4  9  -  9+ 9+ 1  8  9+  9  8  9+ 9+ 9  9  9+  -  1  8  7  2  -  9+  -  -  -  -  -  -  1   -  -  -  -  -  -  2     4
  5  9  -  9+ 9+ -  9  9+  9  8  9+ 9+ 8  9  9+  -  1  9  8  2  -  9   -  -  -  -  -  -  1   -  -  -  -  -  -  2     5
  6  9+ -  9+ 9+ -  9  9+  7  8  9+ 9+ 8  9  9+  -  1  9+ 8  -  -  9   -  -  -  -  -  -  1   -  -  -  -  -  -  2     6
  7  9  7  9+ 9  -  9  9+  7  8  9+ 9  8  9  9+  -  1  9+ 1  -  1  8   -  -  -  -  -  -  1   -  -  -  -  -  -  2     7
  8  9  8  9+ 9  -  9  9+  8  9  9+ 9  8  9+ 9+  -  1  9+ -  -  5  9   -  -  -  -  -  -  1   -  -  -  -  -  -  2     8
  9  8  8  9+ 7  6  9  9+  8  9  9+ 9  9  9+ 9+  -  -  9  1  -  7  9   -  -  -  -  -  -  1   -  -  -  -  -  -  2     9
 10  5  8  9+ 4  6  9  9+  9  9  9+ 9  9  9+ 9+  -  3  9  5  -  6  9   -  -  -  -  -  -  1   -  -  -  -  -  -  2    10
 11  3  8  9+ -  5  9  9+  8  9  9+ 7  9  9+ 9+  5  -  9+ 9  5  1* 8   -  -  -  -  -  -  1   -  -  -  -  -  -  2    11
 12  1  8  9  -  4  9  9+  7  9  9+ 4  8  9  9+  9  5  9+ 9+ 9  2* 8   -  -  5  6  -  -  1   -  -  -  -  -  -  2    12
 13  -  6  1  -  -  7  9+  6  8  9+ 1  8  9  9+  9+ 9  9+ 9  9  7  8   4  -  9+ 9  7  -  1   -  -  -  -  -  -  1    13
 14  -  -  -  -  -  1  9+  5  7  8  -  8  8  9+  9+ 9  9+ 9  9  9  9+  7  2* 9+ 9  9  -  8   -  -  5  -  -  -  1    14
 15  -  -  -  -  -  -  9+  4  6  5  -  6  7  9+  9+ 9  9+ 9  9  9  9+  7  5  9+ 9  2  2  5   -  -  5  -  -  -  -    15
 16  -  -  -  -  -  -  9+  5  6  4  2  5  4  9+  9+ 8  9+ 9+ 9  9  9+  5  1  9+ 8  2* 2  9   -  -  5  -  -  -  1    16
 17  -  -  -  -  -  -  9+  6  5  5  5  6  1  9+  9+ 5  9+ 9+ 3  9  9+  -  -  9+ 9  -  3  9+  -  -  5  -  -  -  1    17
 18  1  -  -  -  -  -  9+  8  6  6  7  6  -  9+  9+ 6  9+ 9+ 4  9  9+  -  -  9+ 9  -  7  9+  -  -  5  -  -  -  1    18
 19  3  -  -  2  -  -  9+  9  7  8  8  8  -  9+  6  6  9+ 9+ 6  9  9+  -  -  9+ 9  -  9  9+  -  -  2  -  -  -  1    19
 20  5  -  7  5  -  -  9+  9  8  9+ 9  8  4  9+  1  7  9+ 9+ 8  9  9+  -  -  9+ 4  -  9  9   -  -  -  -  -  -  1    20
 21  8  3  9  8  6  -  9+  9  8  9+ 9+ 9  7  9+  -  8  9+ 9  8  9  9+  -  -  9+ -  -  9  6   -  -  -  -  -  -  1    21
 22  9  3  9+ 9  8  -  9+  9  8  9+ 9+ 9  5  9+  -  9+ 9+ 9  8  9  9+  -  -  9  -  -  7  1   -  -  -  -  -  -  1    22
 23  9  2  9+ 9  9  -  9+  9  8  9+ 9+ 9  4  9+  -  9+ 9+ 9  5  9  9+  -  1  6  -  -  2  3   -  -  -  -  -  -  2    23
     EU FE SA AF AS OC NA  EU FE SA AF AS OC NA  EU FE SA AF AS OC NA  EU FE SA AF AS OC NA  EU FE SA AF AS OC NA
```

Table 5

Printout of summary propagation table for Boston to the rest of the world, for a Very High level of solar activity in the month of January.

```
Jan., MA (Boston), for SSN = Very High, Sigs in S-Units. By N6BV, ARRL.
           80 Meters              40 Meters              20 Meters              15 Meters              10 Meters
UTC  EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   UTC
 0   9+ -  9+ 9+ 8  -  9+   9+ 5  9+ 9+ 9  -  9+   1  9+ 9+ 9+ 9+ 9  9+   -  9  9+ 2  2  9+ 9+   -  1  8  -  -  8  9+     0
 1   9+ -  9+ 9+ 8  -  9+   9+ 4  9+ 9+ 9  2  9+   1  9  9+ 8  9+ 9+ 9+   -  3  9  -  7  9+ 9    -  -  -  -  -  4  2      1
 2   9+ -  9+ 9+ 7  -  9+   9+ 4  9+ 9+ 9  7  9+   1  9  9+ 8  9  9+ 9+   -  -  3  -  -  7  9    -  -  -  -  -  -  2      2
 3   9+ -  9+ 9+ 1  2  9+   9+ 4  9+ 9+ 9  9  9+   -  7  9+ 7  8  9+ 9    -  -  -  -  -  -  -    -  -  -  -  -  -  2      3
 4   9+ -  9+ 9+ -  7  9+   9+ 5  9+ 9+ 8  9  9+   -  5  9+ 9  9  9  9+   -  -  1  -  -  -  -    -  -  -  -  -  -  2      4
 5   9+ -  9+ 9+ -  8  9+   9+ 6  9+ 9+ 7  9  9+   -  5  9+ 9  9  5  9+   -  -  -  -  -  -  -    -  -  -  -  -  -  2      5
 6   9+ -  9+ 9+ -  8  9+   9+ 7  9+ 9+ 7  9  9+   -  8  9+ 8  9  5  9+   -  -  -  -  -  -  -    -  -  -  -  -  -  2      6
 7   9+ -  9+ 9+ -  8  9+   9  8  9+ 9+ 7  9+ 9+   -  9  9+ -  7  9  9+   -  -  1  -  -  -  -    -  -  -  -  -  -  2      7
 8   9  7  9+ 9  -  8  9+   9  8  9+ 9+ 8  9+ 9+   -  9  9+ -  4  9+ 9+   -  -  1  -  -  -  2    -  -  -  -  -  -  2      8
 9   8  7  9+ 7  -  8  9+   9  9  9+ 9  8  9+ 9+   -  6  9+ -  1  9+ 9+   -  -  -  -  -  -  1    -  -  -  -  -  -  2      9
10   5  8  9+ 2  3  8  9+   9  9  9+ 8  8  9  9+   4  -  9+ 9+ 1  5  9    -  -  -  -  -  -  -    -  -  -  -  -  -  2     10
11   1  8  9+ -  4  9  9+   8  9  9+ 5  8  9  9+   9+ 4* 9+ 9+ 7  -  8    -  -  9  9  -  -  -    -  -  -  -  -  -  2     11
12   -  7  8  -  1  9  9+   6  9  9+ 1  8  9  9+   9+ 9  9+ 9  9  1* 9+   9  8* 9+ 9+ 9  5* -    -  2* 9  9  1  1* 2     12
13   -  -  -  -  -  2  9+   4  8  8  -  7  9  9+   9+ 9  9+ 9  9  9+ 9+   9+ 7  9+ 9+ 9+ 3* 9    9  5* 9+ 9+ 9  6* 2     13
14   -  -  -  -  -  -  9+   2  7  4  -  5  8  9+   9+ 9  9+ 8  9  9  9+   9+ 9  9+ 9+ 9+ 9  9+   9  6* 9+ 9+ 9  1* 1     14
15   -  -  -  -  -  -  9    1  5  -  -  4  5  9+   9+ 9  9+ 9  9  9  9+   9+ 9+ 9+ 9+ 9+ 9  9+   9  5  9+ 9+ 6  6  8     15
16   -  -  -  -  -  -  8    3  4  -  -  3  1  9+   9+ 8  9  9  9  9  9+   9+ 9+ 9+ 9+ 9  9+ 9+   9  8  9+ 9+ -  8  9     16
17   -  -  -  -  -  -  8    5  3  -  2  4  -  9+   9+ 8  9+ 9+ 9  9  9+   9+ 9  9+ 9+ 1* 9+ 9+   -  8  9+ 9+ -  8  9+    17
18   -  -  -  -  -  -  9    7  4  2  5  5  -  9+   9+ 9  9+ 9+ 9  9  9+   9+ 9  9+ 9+ 1  9+ 9+   -  7  9+ 9+ -  9+ 9+    18
19   1  -  -  1  -  -  9+   8  5  6  8  7  -  9+   9+ 9  9+ 9+ 9  9  9+   -  9+ 9+ 9+ 2  9  9+   -  6  9+ 9+ -  9+ 9+    19
20   4  -  2  5  -  -  9+   9  6  9  9  8  -  9+   9+ 9  9+ 9+ 9  9  9+   -  8  9+ 9+ 3  9  9+   -  1  9+ 9  -  9  9+    20
21   7  -  8  7  1  -  9+   9+ 7  9+ 9+ 8  1  9+   8  9  9+ 9+ 9  9  9+   -  6  9+ 9+ 3  9  9+   -  -  9+ 5* -  9+ 9+    21
22   9  2  9+ 9  8  -  9+   9+ 7  9+ 9+ 9  4  9+   2  9+ 9+ 9+ 9  9  9+   -  9+ 9+ 9  1  9+ 9+   -  5  9+ 4* -  9  6     22
23   9  -  9+ 9  8  -  9+   9+ 7  9+ 9+ 9  -  9+   1  9+ 9+ 9+ 9  9  9+   -  9+ 9+ 6  -  9  9+   -  7  9+ 2* -  9  2     23
     EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA
```

EU	All of Europe
FE	The Far East, centered on Japan
SA	South America, centered on Paraguay
AF	All of Africa, centered on Zambia
AS	South Asia, centered on India
OC	Oceania, centered on Sydney, Australia
NA	North America, all across the USA

These propagation files show the highest predicted signal strength (in S-units) throughout the generalized receiving area, for a 1500-W transmitter and rather good antennas on both sides of the circuit. The standard antennas are:

- 100-foot high inverted-V dipoles for 80 and 40 meters
- 3-element Yagi at 100 feet for 20 meters
- 4-element Yagi at 60 feet for 15 and 10 meters.

For example, Summary Table 4 shows that in January during a period of Very Low solar activity, 15 meters is open to somewhere in Europe from Boston for only 4 hours, from 13 to 16 UTC, with a peak signal level between S4 and S7. Now look at Table 5, where 15 meters is predicted to be open to Europe during a period of Very High solar activity for 7 hours, from 12 to 18 UTC, with peak signals ranging from S9 to S9+.

Both Tables 4 and 5 represent *snapshots* of predicted signal levels to generalized receiving locations—that is, they are computed for a particular month, from a particular transmitting location, and for a particular level of solar activity. These tables provide summary information that is particularly valuable for someone planning for an operating event such as a DXpedition or a contest.

What happens if you don’t have a big-gun station with high antennas or the 1500-W power assumed in the analyses above? You can discount the S-Meter readings to reflect a smaller station:

- Subtract 2 S units for a dipole instead of a Yagi at same height on 20/15/10 meters.
- Subtract 3 S units for a dipole at 50 feet instead of a Yagi at 100 feet on 20 meters.
- Subtract 1 S unit for a dipole at 50 feet rather than a dipole at 100 feet on 40/80 meters.
- Subtract 3 S units for 100 W rather than 1500 W.
- Subtract 6 S units for 5 W (QRP) rather than 1500 W.

For example, Table 4 predicts an S7 signal into Boston from Europe on 15 meters at 14 UTC. If a European station is using a dipole at 50 feet, with 100 W of power, what

20 Meters: Jan., MA (Boston), for SSN = Very High, Sigs in S-Units. By N6BV, ARRL.

UTC -->

Zone	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
KL7 = 01	9+	9+	9+	7	-	-	-	-	-	-	-	-	-	-	-	3	9+	9+	9+	9+	9+	9+	9+	9+
VO2 = 02	9+	9	9	9	9	9	8	7	5	3	2	1	5	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	9+
W6 = 03	9+	9+	9+	7	7	1	1	5	8	8	3	-	-	1	9	9+	9+	9+	9+	9	9	9+	9+	9+
W0 = 04	9+	9+	9+	8	5	5	5	5	3	2	1	-	-	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+
W3 = 05	4	2	2	2	2	2	2	3	3	3	3	2	1	1	8	9+	9+	9+	9+	9+	9+	9+	9+	9
XE1 = 06	9+	9+	7	9	9+	9+	9+	9+	9+	9+	9	8	9+	9+	9+	9+	9	9	9	9	9+	9+	9+	9+
TI = 07	9+	9+	8	9	9	9	9	9	9	9+	9	9+	9+	9+	9+	9+	9	8	9	9	9+	9+	9+	9+
VP2 = 08	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	9	9+	9+	9+	9+	9+	9	9+	9+	9+	9+	9+	9+	9+
P4 = 09	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9	9	9	9	9+	9+	9+	9+	9+
HC = 10	9+	8	9+	9	9	9	9	9	7	3	1	7	9+	9+	9	5	5	5	7	8	9+	9+	9+	9+
PY1 = 11	9+	9+	9+	9	9	9+	9+	9	8	6	9	9+	8	2	1	-	-	1	4	8	9	9+	9+	9+
CE = 12	9+	9+	9+	9+	9+	9+	9+	9+	9	8	8	9+	9	8	2	1	1	-	1	3	7	9	9+	9+
LU = 13	9+	9+	9+	9+	9+	9+	9+	9+	8	8	8	9+	8	4	2	1	-	-	1	4	8	9	9+	9+
G = 14	-	-	-	-	-	-	-	-	-	-	-	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	2	-
I = 15	-	-	-	-	-	-	-	-	-	-	4	9	9	9	9	9	9	9+	9+	9	8	2	-	-
UA3 = 16	1	1	1	-	-	-	-	-	-	-	-	8	9	9+	9+	9+	9	8	5	-	-	-	-	1
UN = 17	1	-	-	8	7	7	7	1	-	-	-	2	9	9	9	6	-	-	2	4	8	9	5	4
UA9 = 18	6	7	6	6	9	9	9	7	4	1	-	-	8	8	6	6	5	6	7	8	9	9	8	7
UA0 = 19	9+	9	9	6	5	5	8	8	8	4	-	-	2	6	8	8	8	7	4	4	7	9	9+	9+
4X = 20	8	6	3	1	-	3	4	-	-	-	1	8	8	8	8	8	9	9	9+	9	9	8	7	7
HZ = 21	9+	9	4	3	8	8	2	-	-	-	1	7	8	9	8	8	9	9	9	9	9	9	9	9
VU = 22	7	5	8	7	6	7	5	-	-	-	-	6	9	9	9	9	3	2	2	2	8	8	9	8
JT = 23	9	9+	9	5	7	8	8	6	3	-	-	2*	8	8	5	6	8	8	8	8	9	7	5	6
VS6 = 24	9	9	9	5	4	5	7	8	6	1	-	1*	5	7	1	1	1	1	4	2	-	-	-	9
JA1 = 25	9	9	8	7	5	5	8	9	9	6	-	1	1	2	7	7	6	2	-	-	7	9	9+	9
HS = 26	9	9	6	4	2	-	-	2	1	-	-	2*	9	9	9	9	8	7	5	4	5	-	1*	1
DU = 27	9	8	7	-	-	-	5	7	7	1	-	-	1*	9	9	7	6	4	5	3	1*	1*	8	9
YB = 28	9	8	1	-	-	-	-	-	-	-	-	4*	8	9	9	9	8	8	9	9	9	9	9+	9+
VK6 = 29	3*	4*	-	-	-	-	-	-	5	3	-	-	-	5	9	9	9	8	9	9	9	9	9	8
VK3 = 30	1*	-	-	-	-	-	1	3	9	9	4	-	-	9+	9	8	2	1	-	-	1	2*	5*	4*
KH6 = 31	9	9+	9+	9+	8	2	2	6	4	-	-	-	-	-	-	-	9	9	8	7	6	4	6	7
KH8 = 32	-	2	9	9	9	5	5	9	9+	9+	5	-	-	9+	9	9	8	5	3	1	-	-	-	-
CN = 33	-	-	-	-	-	-	-	-	-	-	9	9+	9	9	8	9	9	9+	9+	9+	9+	9+	9+	7
SU = 34	9	8	3	3	-	1	4	-	-	-	2	7	8	8	8	8	9	9	9+	9+	9+	9+	8	8
6W = 35	9+	8	-	-	2	7	5	-	-	-	9+	9+	8	5	4	3	7	9	9+	9+	9+	9+	9+	9+
D2 = 36	9+	9+	5	3	9	9	8	-	-	-	3	-	-	-	-	4	4	7	8	9	9+	9+	9+	9+
5Z = 37	9+	9	2	4	8	8	1	-	-	-	2	-	-	3	5	5	7	8	9	9	9+	9+	9+	9+
ZS6 = 38	9+	9+	8	7	8	9	6	-	-	-	-	-	-	-	1*	1	2	6	8	9	9+	9+	9+	9+
FR = 39	9+	8	2	1	4	1	-	-	-	-	-	-	-	2*	3*	1*	1	3	8	9	9+	9+	9+	9+
FJL = 40	9+	9+	7	4	7	8	7	1	-	-	-	1*	8	9	9	9	9	9	9	9	9+	9+	9+	9+
Zone	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23

UTC -->

Expected signal levels using 1500 W and 3-element Yagis at 100 feet at each station.

Fig 41—The 20-meter page from Detailed propagation-prediction for the month of January, during Very High solar conditions, from Boston to 40 CQ Zones throughout the world. There are similar pages for each month/SSN level for 160, 80, 40, 20, 15 and 10 meters. These Detailed tables are very useful for planning DX work.

would this do to the predicted signal level in Boston? You would compute: S7 – 2 S units (for a dipole instead of Yagi) – 3 S units (100 W rather than 1500 W) = an S2 signal in Boston. A QRP station with a 4-element 15-meter Yagi at 60 feet would yield: S7 – 6 S units = an S1 signal in Boston.

More Detailed Predictions

Let's now look at table in **Fig 41**, which is the Detailed 20-meter page for the same conditions in Table 5: January at a Very High level of solar activity from Boston to the world. There are six such pages per month/SSN level, covering 160, 80, 40, 20, 15 and 10 meters.

In a Detailed prediction table, the world is divided into the 40 CQ Zones, with a particular sample location in each zone. For example, Zone 14 in Western Europe is represented by a location in London, England (call sign G), while Zone 25 is represented by a location in Tokyo, Japan (call sign JA1). Note that Zones with large ham populations are highlighted with dark shadowing for easy identification. For example, Zones 3, 4 and 5 cover the USA, while Zones 14, 15 and 16 cover the majority of Europe. Zone 25 covers the big ham population in Japan.

Let's revisit the example above for computing the signal strength for a station in London, but this time on 20 meters. Again, we'll assume that the G station has a dipole at 50 feet and 100 W of transmitter power. At 14 UTC in Zone 14, the table in Fig 40 predicts a very healthy signal for the reference big-gun station, at S9+. This is a signal at least S9 + 10 dB. Here, we're going to round off the plus 10 dB to 2 S units, giving a fictional 11 S units to start. We discount this for the smaller station: S11 – 3 S units (for a dipole at 50 feet instead of a 3-element Yagi at 100 feet) – 3 S units (100 W rather than 1500 W) = S5 signal in Boston. This is a respectable signal and will probably get through, in the absence of stronger signals calling the Boston station at the same time, of course.

Here's another example of how to use the Detailed propagation-prediction tables. Let's say that at 1230 UTC in January you work a VU2 station in New Delhi on 15 meters from Boston, where the local time is 7:30 AM.

You need a 20-meter contact also for the 5-Band DXCC award, so you quickly check the table in Fig 40 for Zone 22 (VU) and find that the predicted signal strength is S9. Your new VU2 friend is willing to jump to 20 meters and so you QSY to make the contact.

But perhaps you are late leaving for work and so you ask your new VU2 friend to make a schedule with you later that evening. Again, you consult the Detailed prediction table for 20 meters and find that signals are predicted to be S8 or stronger from 20 to 23 UTC, dropping to S7 at 00 UTC. You quickly ask your new friend whether he minds waking up at 4:30 AM his time to make a schedule with you at 2300 UTC, because New Delhi is 5½ hours ahead of UTC. You determined this using the program *GeoClock*, which is included with the software on the CD-ROM in the back of this book and which you run in the background on *Windows*. Luckily, he's a very gracious fellow and agrees to meet you on a specific frequency at that time.

The Detailed propagation-prediction tables give you all the information needed to plan your operations to maximize your enjoyment chasing DX. You can use these tables to plan a 48-hour contest next month, or next year—or you can use them to plan a schedule with your ham cousin on the West Coast on Saturday afternoon.

THE PROPAGATION BIG PICTURE

A newcomer to the HF bands could easily be overwhelmed with the sheer amount of data available in the Summary (and particularly the Detailed) prediction tables on the CD-ROM included with this book. So here's a long-term, "big-picture" view of HF propagation that might help answer some common questions. For example, what month really is the best for working DX around the clock? Or what level of solar activity is necessary to provide an opening between your QTH and somewhere in the South Pacific?

Table 6 is a table showing the number of hours in a day during each month when each major HF band is open to the same receiving areas shown in Tables 4 and 5. The listing is for New England, for three levels of solar activity: Very Low, Medium and Very High. The number of hours are separated in Table 6 by slashes. (Versions of Table 6 for other areas around the US are on the CD-ROM that accompanies this book in **Fig6Tab.PDF**.)

Let's examine the conditions for New England to Europe on 15 meters for October. The entry shows "7/11/17," meaning that for a Very Low level of solar activity, 15 meters is open for 7 hours; for a Medium level, it is open for 11 hours and for a Very High level of solar activity it is open for 17 hours a day.

Even for a Very Low level of solar activity, the month with the most hours available per day from Boston to somewhere in Europe is October, with 7 hours, followed by the next largest month of March, with 6 hours. For a Very High level of solar activity, however, the 15-meter band is open to Europe for 18 hours in April, followed by 17 hours availability in September and October. Arguably, the CQ World Wide Contest Committee picked the very best month for higher-frequency propagation when they chose October for the Phone portion of that contest.

You can easily see that even at a Very High level of solar activity, the summer months are not very good to work DX, particularly on east-west paths. For example, the 10-meter band is very rarely open from New England to Europe after the month of April, even when solar activity is at the highest levels possible. Things pick up after September, even for a Medium level of solar activity. Again, October looks like the most fruitful month in terms of the number of hours 10 meters is open to Europe under all levels of solar conditions.

Ten meters is open more regularly on north-south paths, such as from New England to South America or to southern Africa. It is open as much as 10 hours a day during March and October to deep South America, and 7 hours a day in October to Africa—even during the lowest parts of the solar cycle. (Together with the sporadic-E propagation that 10 meters enjoys during the summer, this band can often be a lot of fun even during the sunspot doldrums. You just have to *be operating* on the band, rather than avoiding it because you know the sunspots are "spotty!")

Now, look at the 20-meter band in Table 6. From New England, twenty is open to somewhere in South America for 24 hours a day, no matter the level of solar activity. Note that Table 6 doesn't predict the level of signals available; it just shows that the band is open with a signal strength greater than 0 on the S meter. Look back at Summary Table 4 for the predicted signal strengths in January at a Very Low level of solar activity. There, you can see that the signal strength from New England into deep South America is always S8 or greater for a big gun station. A lot of the time during the night the band sounds dead, simply because everyone is either asleep or operating on a lower frequency.

For the 40-meter band in Table 6, during the month of January the band is open to Europe for 24 hours a day, whatever the level of solar activity is. Look now at Table 4, and you'll see that the predicted level for Very Low solar activity varies from S4 to S9. Local QRM or QRN would probably disrupt communications on 40 meters in Europe for stateside signals weaker than perhaps S3 or S4. Even though you might well be able to hear Europeans from New England during the day, they probably won't hear you because of local conditions, including local S9+ European stations and atmospheric noise from nearby thunderstorms. New England stations with big antennas can often hear Europeans on 40 meters as early as noontime, but must wait until the late afternoon before the Europeans can hear them above their local noise and QRM.

Let's say that you want to boost your country total on 80 meters by concentrating on stations in the South Pacific. The best months would be from November to

Table 6

The number of hours per day when a particular band is open to the target geographic areas in Table 4, as related to the level of solar activity (Very Low, Medium and Very High). This table is customized for Boston to the rest of the world. Some paths are open 24 hours a day, plus or minus QRM and local QRN, no matter what the level of solar activity is. See CD-ROM for other transmitting locations.

```
MA (Boston)
Hours Open to Each Region for Very-Low/Medium/Very-High SSNs

80 Meters:
Month  Europe   Far East  So. Amer.  Africa   So. Asia  Oceania   No. Amer.
Jan   17/17/16   5/ 4/ 3  17/17/16  16/16/15   8/ 7/ 5  11/10/ 9  24/24/24
Feb   17/16/15   3/ 3/ 2  17/16/16  15/15/14   6/ 4/ 4  10/ 9/ 9  24/24/24
Mar   15/15/14   3/ 2/ 1  16/16/15  15/13/13   4/ 4/ 3   9/ 8/ 7  24/24/24
Apr   13/13/12   1/ 0/ 0  16/16/14  13/13/13   3/ 3/ 1   9/ 8/ 7  24/24/24
May   12/11/10   0/ 0/ 0  16/15/14  12/11/10   2/ 1/ 1   7/ 6/ 6  24/24/24
Jun   10/ 9/ 8   0/ 0/ 0  14/14/14  11/10/10   1/ 1/ 0   6/ 5/ 5  24/24/24
Jul   11/11/ 9   0/ 0/ 0  15/14/14  11/11/11   2/ 1/ 1   7/ 6/ 5  24/24/24
Aug   13/11/11   0/ 0/ 0  16/16/14  13/12/11   3/ 2/ 1   7/ 7/ 6  24/24/24
Sep   14/13/11   2/ 1/ 0  17/16/14  13/13/12   4/ 4/ 2   9/ 8/ 8  24/24/24
Oct   15/15/13   3/ 2/ 1  17/17/16  14/14/13   5/ 4/ 4   9/ 9/ 7  24/24/24
Nov   17/17/15   4/ 4/ 2  17/17/16  16/15/14   8/ 7/ 4  11/10/ 9  24/24/24
Dec   19/18/17   7/ 6/ 4  18/18/17  16/16/16  11/ 9/ 7  12/11/11  24/24/24

40 Meters:
Month  Europe   Far East  So. Amer.  Africa   So. Asia  Oceania   No. Amer.
Jan   24/24/24  15/16/15  24/24/21  21/20/19  21/21/19  19/18/15  24/24/24
Feb   24/24/21  13/11/11  24/23/20  20/19/18  19/19/17  16/15/14  24/24/24
Mar   23/22/19  10/ 9/ 7  24/21/18  19/17/17  17/17/13  13/13/13  24/24/24
Apr   21/19/18   8/ 6/ 4  22/20/18  17/16/15  16/11/ 8  13/13/11  24/24/24
May   19/17/17   5/ 4/ 3  22/18/17  17/16/14   9/ 8/ 5  12/11/10  24/24/24
Jun   17/15/13   4/ 2/ 2  22/18/16  16/15/14   7/ 5/ 5  11/10/ 9  24/24/24
Jul   18/16/15   5/ 4/ 2  24/18/17  17/15/14   8/ 7/ 5  12/11/10  24/24/24
Aug   19/17/16   7/ 5/ 4  24/19/18  18/16/15  11/10/ 6  13/12/11  24/24/24
Sep   22/21/17   9/ 8/ 5  23/20/18  18/17/16  14/11/ 7  13/13/12  24/24/24
Oct   24/23/20  12/11/ 8  24/23/19  20/18/17  17/16/14  16/13/13  24/24/24
Nov   24/24/22  14/13/12  24/24/20  21/19/18  21/20/17  17/17/13  24/24/24
Dec   24/24/24  18/19/22  24/24/21  23/21/19  24/23/22  21/19/18  24/24/24
```

February in terms of the number of hours per day when the 80-meter band is open to Oceania. You can see by reading across the line for each month that the level of solar activity is not hugely important on 80 meters to any location. Common experience (backed by the statistical information in Table 6) is that the 80-meter band is open only marginally longer when sunspots are low.

This is true to a greater extent on 40 meters. Thus you may hear the generalization that the low bands tend to be better during periods of low solar activity, while the upper HF bands (above 10 MHz) tend to be better when the sun is more active.

Table 6 can give you a good handle on what months are the most productive for DXing and contesting. It should be no surprise to most veteran operators that the fall and winter months are the best times to work DX.

20 Meters:

Month	Europe	Far East	So. Amer.	Africa	So. Asia	Oceania	No. Amer.
Jan	13/16/22	15/22/22	24/24/24	20/21/21	18/20/22	18/23/22	24/24/24
Feb	12/18/23	13/21/24	24/24/24	22/22/24	15/21/24	18/23/24	24/24/24
Mar	15/18/24	17/20/24	24/24/24	22/24/24	18/21/24	16/24/24	24/24/24
Apr	15/20/24	19/22/24	24/24/24	21/24/24	19/22/24	18/24/24	24/24/24
May	19/23/24	22/24/24	24/24/24	23/24/24	23/24/24	21/24/24	24/24/24
Jun	22/24/24	24/24/24	24/24/24	24/24/24	24/24/24	24/24/24	24/24/24
Jul	19/24/24	24/24/24	24/24/24	21/24/24	24/24/24	23/24/24	24/24/24
Aug	15/20/24	20/24/24	24/24/24	20/24/24	20/24/24	19/24/24	24/24/24
Sep	16/19/24	17/21/24	24/24/24	21/24/24	18/21/24	17/24/24	24/24/24
Oct	15/21/24	16/20/24	24/24/24	22/24/24	19/22/24	17/24/24	24/24/24
Nov	14/20/23	14/22/24	24/24/24	20/24/24	17/21/24	19/23/24	24/24/24
Dec	11/17/24	13/22/24	24/24/24	17/23/24	12/22/24	16/24/24	24/24/24

15 Meters:

Month	Europe	Far East	So. Amer.	Africa	So. Asia	Oceania	No. Amer.
Jan	4/ 6/ 7	2/ 9/13	12/15/16	9/13/13	3/ 4/ 7	9/12/13	24/15/16
Feb	4/ 7/12	4/10/14	13/18/23	11/13/16	3/ 7/13	8/13/15	22/16/19
Mar	6/ 9/14	2/13/15	14/21/24	13/17/22	5/11/17	10/14/17	15/16/23
Apr	0/10/18	3/13/18	15/23/24	15/18/24	9/15/19	11/15/21	16/16/24
May	1/13/16	6/10/19	17/20/24	14/18/24	13/17/18	10/16/19	20/19/24
Jun	0/ 2/16	0/ 9/15	16/21/24	14/18/24	5/15/18	10/12/20	24/22/22
Jul	0/ 2/16	0/ 5/18	15/19/24	12/18/24	0/12/18	4/12/20	24/22/21
Aug	0/ 2/14	0/ 8/17	14/18/22	13/16/22	0/12/17	6/10/19	22/19/21
Sep	1/10/17	6/13/17	14/16/24	13/17/22	9/14/17	9/14/17	16/16/22
Oct	7/11/17	10/13/17	12/16/22	12/15/22	7/12/17	12/13/15	18/15/22
Nov	5/ 8/14	8/11/14	12/16/22	11/14/17	3/ 7/16	10/13/15	20/16/21
Dec	3/ 6/ 9	2/10/13	12/15/23	8/13/15	2/ 4/12	9/12/14	24/15/18

10 Meters:

Month	Europe	Far East	So. Amer.	Africa	So. Asia	Oceania	No. Amer.
Jan	0/ 1/ 4	0/ 1/ 8	6/11/13	0/ 7/10	0/ 1/ 3	0/ 3/11	23/24/24
Feb	0/ 2/ 7	0/ 2/10	8/12/14	0/ 9/13	0/ 3/ 5	0/ 7/13	24/24/24
Mar	0/ 0/ 8	0/ 1/10	10/14/20	1/11/14	0/ 0/ 8	0/ 7/13	23/24/24
Apr	0/ 0/ 8	0/ 0/ 8	7/14/21	0/12/17	0/ 0/13	0/ 5/11	18/24/24
May	0/ 0/ 0	0/ 0/ 1	7/12/20	1/10/17	0/ 1/12	0/ 2/11	17/20/22
Jun	0/ 0/ 0	0/ 0/ 0	7/11/18	0/ 3/17	0/ 0/ 0	0/ 0/ 2	21/19/23
Jul	0/ 0/ 0	0/ 0/ 0	2/ 9/19	0/ 2/18	0/ 0/ 7	0/ 0/ 6	16/16/24
Aug	0/ 0/ 0	0/ 0/ 0	2/10/17	0/ 1/16	0/ 0/10	0/ 0/ 8	17/17/24
Sep	0/ 0/ 8	0/ 1/10	7/13/18	0/11/16	0/ 0/10	0/ 2/ 9	19/24/24
Oct	0/ 5/ 9	0/ 2/11	10/12/16	7/12/14	0/ 5/ 9	0/ 8/12	24/24/24
Nov	0/ 4/ 8	0/ 3/11	9/12/15	5/10/13	0/ 3/ 6	4/10/12	24/24/24
Dec	0/ 3/ 6	0/ 1/ 8	8/11/13	1/ 8/12	0/ 1/ 4	2/ 7/12	23/23/24

Do-It-Yourself Propagation Prediction

Very reliable methods of determining the MUF for any given radio path have been developed over the last 50 years. As discussed previously, these methods are all based on the smoothed sunspot number (SSN) as the measure of solar activity. It is for this reason that smoothed sunspot numbers hold so much meaning for radio amateurs and others concerned with radio-wave propagation—they are the link to past (and future) propagation conditions.

Early on, the prediction of propagation conditions required tedious work with numerous graphs, along with charts of frequency contours overlaid, or overprinted, on world maps. The basic materials were available from an agency of the US government. Monthly publications provided the frequency-contour data a few months in advance. Only rarely did amateurs try their hand at predicting propagation conditions using these hard-to-use methods.

Today's powerful PCs have given the amateur wonderful tools to make quick-and-easy HF propagation predictions, whether for a contest or a DXpedition. The summary and detailed prediction tables described earlier in this chapter were generated using *CAPMan*, a modernized version of the mainframe *IONCAP* program, on a PC.

While tremendously useful to setting up schedules and for planning strategy for contests, both the Summary and Detailed prediction tables located on the CD-ROM accompanying this book show signal strength. They do not show

Table 7
Features and Attributes of Propagation Prediction Programs

	ASAPS V. 4	*CAPMan*	*VOACAP Windows*	*ACE-HF*	*W6ELProp V. 2.70*	*WinCAP Wizard 2*	*PropLab Pro*
User Friendliness	Good	Good	Good	Excellent	Good	Good	Poor
Operating System	Windows	DOS	Windows	Windows	Windows	Windows	DOS
Uses k or A index	No	Yes	No	No	Yes	Yes	Yes
User library of QTHs	Yes	Yes	Yes	Yes-RX	Yes	Yes	No
Bearings, distances	Yes	Yes	Yes	Yes	Yes	Yes	Yes
MUF calculation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LUF calculation	Yes	Yes	Yes	Yes	No	Yes	Yes
Wave angle calculation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vary minimum wave angle	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Path regions and hops	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Multipath effects	Yes	Yes	Yes	Yes	No	Yes	Yes
Path probability	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Signal strengths	Yes	Yes	Yes	Yes	Yes	Yes	Yes
S/N ratios	Yes	Yes	Yes	Yes	No	Yes	Yes
Long-path calculation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Antenna selection	Yes	Yes	Yes	Yes	Indirectly	Isotropic	Yes
Vary antenna height	Yes	Yes	Yes	Yes	Indirectly	No	Yes
Vary ground characteristics	Yes	Yes	Yes	Yes	No	No	No
Vary transmit power	Yes	Yes	Yes	Yes	Indirectly	Yes	Yes
Graphic displays	Yes	Yes	Yes	Yes	Yes	Yes	2D/3D
UT-day graphs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Area Mapping	No	Yes	Yes	Yes	Yes	No	Yes
Documentation	Yes	Yes	On-line	Yes	Yes	Yes	Yes
Price class	$275+	$89	free†	$99	free§	$29.95+	$150††

Price classes are for late 2003 and subject to change.
†Available on the World Wide Web: **elbert.its.bldrdoc.gov/hf.html**
§Available on the World Wide Web at: **www.qsl.net/w6elprop/**
+ Shipping and handling extra.
††Available on the World Wide Web: **www.spacew.com/www/proplab.html**

other information that is also in the underlying databases used to generate them. They don't, for example, show the dominant elevation angles and neither do they show reliability statistics. You may want to run propagation-prediction software yourself to get into the really "nitty-gritty" details.

Modern programs are designed for quick-and-easy predictions of propagation parameters. See **Table 7** for a listing of a number of popular programs. The basic input information required is the smoothed sunspot number (SSN) or smoothed solar flux, the date (month and day), and the latitudes and longitudes at the two ends of the radio path. The latitude and longitude, of course, are used to determine the great-circle radio path. Most commercial programs tailored for ham use allow you to specify locations by the call sign. The date is used to determine the latitude of the Sun, and this, with the sunspot number, is used to determine the properties of the ionosphere at critical points on the path.

Of course, just because a computer program predicts that a band will be open on a particular path, it doesn't follow that the Sun and the ionosphere will always cooperate! A sudden solar flare can result in a major geomagnetic storm, taking out HF communication anywhere from hours to days. There is still art, as well as a lot of science, in predicting propagation. In times of quiet geomagnetic activity, however, the prediction programs are good at forecasting band openings and closings.

Obtaining Sunspot Number/Solar Flux Data

After you have chosen and then set up a computer program for evaluation of a particular path, you will still need the sunspot number or solar flux level for the period in question. A caution must be stated here—for best accuracy and consistency, use the average of solar flux values taken from actual observations, perhaps from WWV/WWVH, over the previous three or four days. Many amateur packet systems archive WWV flux numbers (plus planetary k_p and A_p indices). Solar flux numbers can vary dramatically from day to day, but the Earth's ionosphere is relatively slow to respond to instantaneous changes in solar radiation. This caveat also

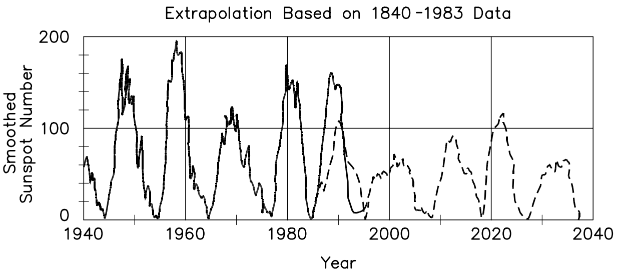

Fig 42—Smoothed sunspot number, with predictions, from 1940 to 2040. This was extrapolated based on data from 1840 to 1983. Cycle 22 actually peaked in Nov 1989, at a monthly smoothed sunspot number of 158. Propagation on the higher frequencies throughout the peak of Cycle 22 was good to excellent, since the monthly smoothed sunspot number stayed at 100 or above from July 1988 through May 1992. (*Courtesy of Naval Ocean Systems Center,* San Diego.*)*

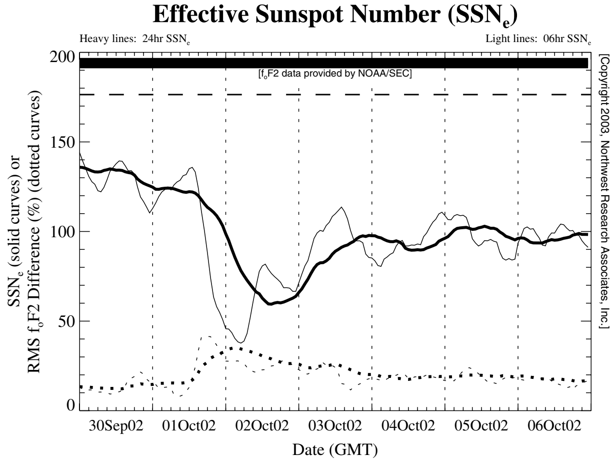

Fig 43—Effective Sunspot Number (SSN$_e$) produced by NWRA. Note large drop in effective SSN due to a geomagnetic storm commencing Oct 1, 2002. (*Courtesy of Northwest Research Associates.*)

holds for sunspot numbers derived, using Fig 18, from WWV/WWVH solar flux numbers.

Fig 42 shows a graph produced in the early 1980s of smoothed sunspot numbers for Solar Cycles 17 through 21, with predictions for Cycles 22 through 26. The graph covers a period of 100 years, from 1940 to 2040, and may be used for making long-term or historical calculations. Just remember that the graph shows smoothed numbers. The solar activity at any given time can be significantly lower or significantly higher than the graph indicates. In fact, Cycle 22 peaked at the end of 1989, as predicted, but with a monthly smoothed sunspot level of 158, quite a bit higher than predicted. Cycle 23 peaked in early 2002 at a level of about 115, a considerably higher level than the predicted value.

WWV PROPAGATION DATA

For the most current data on what the Sun is doing, National Institute of Standards and Technology stations WWV and WWVH broadcast information on solar activity at 18 and 45 minutes past each hour, respectively. These propagation bulletins give the solar flux, geomagnetic A-Index, Boulder K-Index, and a brief statement of solar and geomagnetic activity in the past and coming 24-hour periods, in that order. The solar flux and A-Index are changed daily with the 2118 UT bulletin, the rest every three hours—0018, 0318, 0618 UT and so on. On the Web, up-to-date WWV information can be found at: **ftp://ftp.sel.noaa.gov/pub/latest/wwv.txt** or on the NOAA Web page **www.sec.noaa.gov/**.

Some other useful Web sites are: **dx.qsl.net/propagation/**, **www.dxlc.com/solar**, **hfradio.org/propagation.html**. The Solar Terrestrial Dispatch page contains a wealth of propagation-related information: **www.spacew.com/**. You may also access propagation information on your local PacketCluster. Use the command SH/WWV/*n*, where *n* is the number of spots you wish to see (five is the default).

Another excellent method for obtaining an "equivalent sunspot number" (SSN$_e$) is to go to the Space Weather site of Northwest Research Services: **www.nwra-az.com/spawx/ssne24.html**. NWRA compares real-time ionospheric sounder data around the world with predictions using various levels of SSN looking for the best match. They thus "back into" the actual effective sunspot number. **Fig 43** is a typical NWRA graph, which covers the week ending 6 October 2002. Note the sudden decrease in SSN$_e$ after a geomagnetic storm depressed SSN$_e$ by more than 50%.

The A-Index

The WWV/WWVH A-Index is a daily figure for the state of activity of the Earth's magnetic field. It is updated with the 2118/2145 UT bulletin. The A-Index tells you mainly how yesterday was, but it is very revealing when charted regularly, because geomagnetic disturbances nearly always recur at four-week intervals.

The K-Index

The K-Index (new every three hours) reflects Boulder readings of the Earth's geomagnetic field in the hours just preceding the bulletin data changes. It is the nearest thing to current data on radio propagation available. With new data every three hours, K-Index trend is important. Rising is bad news; falling is good, especially related to propagation on paths involving latitudes above 30° north. Because this is a Boulder, Colorado, reading of geomagnetic activity, it may not correlate closely with conditions in other areas.

The K-Index is also a timely clue to aurora possibilities. Values of 4, and rising, warn that conditions associated with

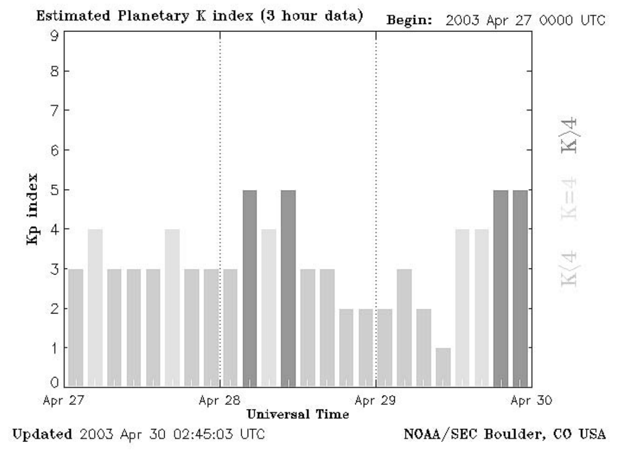

Fig 44—Plot of the estimated planetary k index, K_p, for the last four days. (*Courtesy NOAA/SEC.*)

auroras and degraded HF propagation are present in the Boulder area at the time of the bulletin's preparation. A NOAA Web site that carries up-to-date planetary K_p data is: **www.sel.noaa.gov/ftpmenu/plots/2003_plots/kp.html**. **Fig 44** is a graph from this Web site for four days starting 27 April 2003 to 30 April 2003. This was a period of significant geomagnetic activity, indicated by K_p indices of 5 and 4.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

E. V. Appleton and W.R. Piggott, "Ionospheric Absorption Measurements during a Sunspot Cycle," *J. Atmos. Terr. Phys.*, Vol 3, p 141, 1954.

D. Bilitiza, *International Reference Ionosphere (IRI 90)*, National Space Science Data Center, Greenbelt, MD, 1990.

D. Bray, "Method of Determining VHF/HF Station Capabilities," *QST*, Nov 1961.

A. Brekke, "Physics of the Upper Polar Atmosphere," (John Wiley and Sons: New York, 1997).

R. R. Brown, "Demography, DXpeditions and Magneto-Ionic Theory," *The DX Magazine*, Vol. X, No. 2, p 44, Mar/Apr 1998.

R. R. Brown, *The Little Pistol's Guide to HF Communication* (Sacramento: WorldRadio Books, 1996).

R. R. Brown, "Signal Ducting on the 160 Meter Band," *Communications Quarterly*, p 65, Spring 1998.

R. R. Brown, "Unusual Low-Frequency Signal Propagation at Sunrise," *Communications Quarterly*, p 67, Fall 1998.

R. R. Brown,, "Atmospheric Ozone, a Meteorological Factor in Low-Frequency and 160 Meter Propagation," *Communications Quarterly*, Spring 1999.

K. Davies, *Ionospheric Radio* (London: Peter Peregrinus Ltd, 1990). Excellent technical reference.

R. Garcia, S. Solomon, S. Avery, G. C. Reid, "Transport of Nitric Oxide and the D-Region Winter Anomaly," *J. Geophys. Res*., Vol 92, p 977, 1987.

J. Hall, "Propagation Predictions and Personal Computers," Technical Correspondence, *QST*, Dec 1990, pp 58-59 (description of *IONCAP* as used for ARRL. publications).

E . Harper, *Rhombic Antenna Design* (New York: D. Van Nostrand Co, Inc, 1941).

H. Hertz, *Electric Waves*, translated by D. E. Jones (London: MacMillan, 1893).

R. D. Hunsucker, *Radio Techniques for Probing the Terrestrial Ionosphere* (New York: Springer-Verlag).

R. D. Hunsucker, J. K. Hargreaves, *The High Latitude Ionosphere and Its Effects on Radio Propagation* (Cambridge: Cambridge University Press, 2003).

W. D. Johnston, Computer-calculated and computer-drawn great-circle maps are offered. An 11 × 14-inch map is custom made for your location. Write to K5ZI, PO Box 640, Organ, NM 88052, tele 505-382-7804.

T.L. Killeen, R.M. Johnsson, "Upper Atmospheric Waves, Turbulence, and Winds: Importance for Mesospheric and Thermospheric Studies," **earth.agu.org/revgeophys/killee00/killee00.html**.

J. L. Lynch, "The Maunder Minimum," *QST*, Jul 1976, pp 24-26.

J. C. Maxwell, *A Treatise on Electricity and Magnetism*, Vols I and II (Oxford: Oxford University Press, 1873).

M. W. Maxwell, *Reflections—Transmission Lines and Antennas* (Newington, CT: ARRL, 1990) [out of print].

M. W. Maxwell, *Reflections II—Transmission Lines and Antennas* (Sacramento, CA: 2001).

L. F. McNamara, *The Ionosphere: Communications, Surveillance, and Direction Finding* (Malabar, FL: Krieger Publishing Company, 1991). Another excellent technical reference on propagation.

L. F. McNamara, *Radio Amateur's Guide to the Ionosphere* (Malabar, FL: Krieger Publishing Company, 1994). Excellent, quite-readable text on HF propagation.

A. K. Paul, "Medium Scale Structure of the F Region," *Radio Science*, Volume 24, No. 3, p. 301, 1989.

W. R. Piggott, K. Rawer, "URSI Handbook of Ionogram Interpretation and Reduction," Report UAG-50. World Data Center A for Solar-Terrestrial Physics, Boulder, CO, 1975.

E. Pocock, "Sporadic-E Propagation at VHF: A Review of Progress and Prospects," *QST*, Apr 1988, pp 33-39.

E. Pocock, "Auroral-E Propagation at 144 MHz," *QST*, Dec 1989, pp 28-32.

E. Pocock, "Propagation Forecasting During Solar Cycle 22," *QST*, Jun 1989, pp 18-20.

G. C. Reid, "Ion Chemistry of the D-region," *Advances in Atomic and Molecular Physics*, Vol 12, Academic Press, 1976.

R. B. Rose, "MINIMUF: A Simplified MUF-Prediction Program for Microcomputers," *QST*, Dec 1982, pp 36-38

M. L. Salby, "Fundamentals of Atmospheric Physics," (Academic Press: Boulder, CO, 1996).

S. C. Shallon, W6EL: *MINIPROP PLUS*, a commercially prepared program written for Amateur Radio users; 11058 Queensland St, Los Angeles, CA 90034-3029.

R. D. Straw, *All the Right Angles* (New Bedford, PA: LTA, 1993). Out of print.

R. D. Straw, "*ASAPS* and *CAPMAN*: HF Propagation-Prediction Software for the IBM PC," *QST*, Dec 1994, pp 79-81.

R. D. Straw, "Heavy-Duty HF Propagation-Prediction/Analysis Software," Part 1: *QST*, Sep 1996, pp 28-32; Part 2: *QST*, Oct 1996, pp 28-30.

Chapter 27

Antenna and Transmission-Line Measurements

The principal quantities measured on transmission lines are line current or voltage, and standing-wave ratio (SWR). You make measurements of current or voltage to determine the power input to the line. SWR measurements are useful in connection with the design of coupling circuits and the adjustment of the match between the antenna and transmission line, as well as in the adjustment of these matching circuits.

For most practical purposes a relative measurement is sufficient. An uncalibrated indicator that shows when the largest possible amount of power is being put into the line is just as useful, in most cases, as an instrument that measures the power accurately. It is seldom necessary to know the actual number of watts going into the line unless the overall efficiency of the system is being investigated. An instrument that shows when the SWR is close to 1:1 is all you need for most impedance-matching adjustments. Accurate measurement of SWR is necessary only in studies of antenna characteristics such as bandwidth, or for the design of some types of matching systems, such as a stub match.

Quantitative measurements of reasonable accuracy demand good design and careful construction in the measuring instruments. They also require intelligent use of the equipment, including a knowledge not only of its limitations but also of stray effects that often lead to false results. Until you know the complete conditions of the measurements, a certain amount of skepticism regarding numerical data resulting from amateur measurements with simple equipment is justified. On the other hand, purely qualitative or relative measurements are easy to make and are reliable for the purposes mentioned above.

LINE CURRENT AND VOLTAGE

A current or voltage indicator that can be used with coaxial line is a useful piece of equipment. It need not be elaborate or expensive. Its principal function is to show when the maximum power is being taken from the transmitter; for any given set of line conditions (length, SWR, etc). This will occur when you adjust the transmitter coupling for maximum current or voltage into the transmission line. Although the final-amplifier plate or collector current meter is frequently used for this purpose, it is not always a reliable indicator. In many cases, particularly with a screen-grid tube in the final stage, minimum loaded plate current does not occur simultaneously with maximum power output.

RF VOLTMETER

You can put together a germanium diode in conjunc-

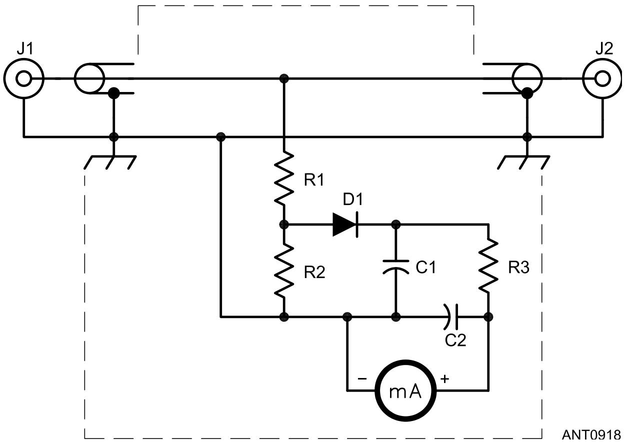

Fig 1—RF voltmeter for coaxial line.

C1, C2—0.005- or 0.01-µF ceramic.
D1—Germanium diode, 1N34A.
J1, J2—Coaxial fittings, chassis-mounting type.
M1—0-1 milliammeter (more sensitive meter may be used if desired; see text).
R1—6.8 kΩ, composition, 1 W for each 100 W of RF power.
R2—680 Ω, $^1/_2$ or 1 W composition.
R3—10 kΩ, $^1/_2$ W (see text).

tion with a low-range milliammeter and a few resistors to form an RF voltmeter suitable for connecting across the two conductors of a coaxial line, as shown in **Fig 1**. It consists of a voltage divider, R1-R2, having a total resistance about 100 times the Z_0 of the line (so the power consumed will be negligible) with a diode rectifier and milliammeter connected across part of the divider to read relative RF voltage. The purpose of R3 is to make the meter readings directly proportional to the applied voltage, as nearly as possible, by swamping the resistance of D1, since the diode resistance will vary with the amplitude of the current through the diode.

You may construct the voltmeter in a small metal box, indicated by the dashed line in the drawing, and fitted with coax receptacles. R1 and R2 should be carbon-composition resistors. The power rating for R1 should be 1 W for each 100 W of carrier power in the matched line; separate 1- or 2-W resistors should be used to make up the total power rating required, to the total resistance as given. Any type of resistor can be used for R3; the total resistance should be such that about 10 V dc will be developed across it at full scale. For example, a 0-1 milliammeter would require 10 kΩ, a 0-500 microammeter would take 20 kΩ, and so on. For comparative measurements only, R3 may be a variable resistor so the sensitivity can be adjusted for various power levels.

In constructing such a voltmeter, you should exercise care to prevent inductive coupling between R1 and the loop formed by R2, D1 and C1, and between the same loop and the line conductors in the assembly. With the lower end of R1 disconnected from R2 and grounded to the enclosure, but without changing its position with respect to the loop, there should be no meter indication when full power is going through the line.

If more than one resistor is used for R1, the units should be arranged end-to-end with very short leads. R1 and R2 should be kept 1/2 inch or more from metal surfaces parallel to the body of the resistor. If you observe these precautions the voltmeter will give consistent readings at frequencies up to 30 MHz. Stray capacitance and stray coupling limit the accuracy at higher frequencies but do not affect the utility of the instrument for comparative measurements.

Calibration

You may calibrate the meter for RF voltage by comparison with a standard such as an RF ammeter. This requires that the line be well matched so the impedance at the point of measurement is equal to the actual Z_0 of the line. Since in that case $P = I^2Z_0$, the power can be calculated from the current. Then $E=\sqrt{PZ_0}$. By making current and voltage measurements at a number of different power levels, you can obtain enough points to draw a calibration curve for your particular setup.

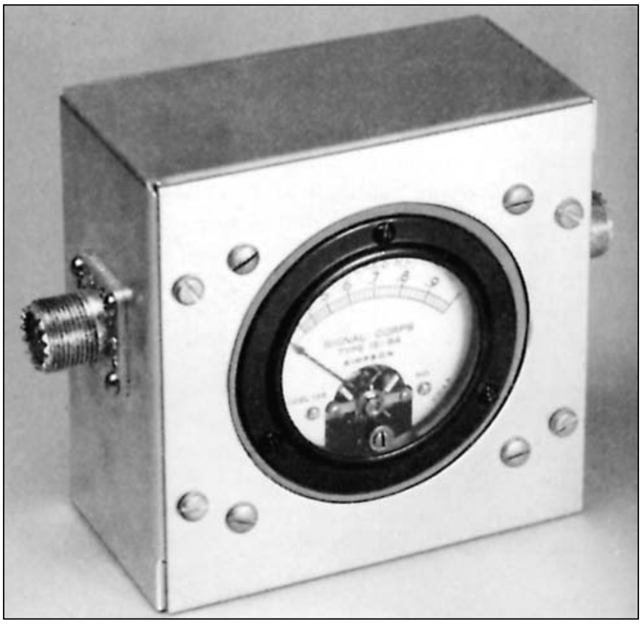

Fig 2—A convenient method of mounting an RF ammeter for use in a coaxial line. This is a metal-case instrument mounted on a thin bakelite panel. The cutout in the metal clears the edge of the meter by about 1/8 inch.

RF AMMETERS

Although they are not as widely available as they used to be, if you can find one on the surplus market or at a hamfest, an RF ammeter is a good way to gauge output power. You can mount an RF ammeter in any convenient location at the input end of the transmission line, the principal precaution being that the capacitance to ground, chassis, and nearby conductors should be low. A bakelite-case instrument can be mounted on a metal panel without introducing enough shunt capacitance to ground to cause serious error up to 30 MHz. When installing a metal-case instrument on a metal panel, you should mount it on a separate sheet of insulating material so that there is 1/8 inch or more separation between the edge of the case and the metal.

A 2-inch instrument can be mounted in a 2 × 4 × 4-inch metal box, as shown in **Fig 2**. This is a convenient arrangement for use with coaxial line. Installed this way, a good quality RF ammeter will measure current with an accuracy that is entirely adequate for calculating power in the line. As discussed above in connection with calibrating RF voltmeters, the line must be closely matched by its load so the actual impedance is resistive and equal to Z_0. The scales of such instruments are cramped at the low end, however, which limits the range of power that can be measured by a single meter. The useful current range is about 3 to 1, corresponding to a power range of about 9 to 1.

SWR Measurements

On parallel-conductor lines it is possible to measure the standing-wave ratio by moving a current (or voltage) indicator along the line, noting the maximum and minimum values of current (or voltage) and then computing the SWR from these measured values. This cannot be done with coaxial line since it is not possible to make measurements of this type inside the cable. The technique is, in fact, seldom used with open lines because it is not only inconvenient but sometimes impossible to reach all parts of the line conductors. Also, the method is subject to considerable error from antenna currents flowing on the line.

Present-day SWR measurements made by amateurs practically always use some form of *directional coupler* or RF-bridge circuit. The indicator circuits themselves are fundamentally simple, but they require considerable care in construction to ensure accurate measurements. The requirements for indicators used only for the adjustment of impedance-matching circuits, rather than actual SWR measurement, are not so stringent, and you can easily make an instrument for this purpose.

BRIDGE CIRCUITS

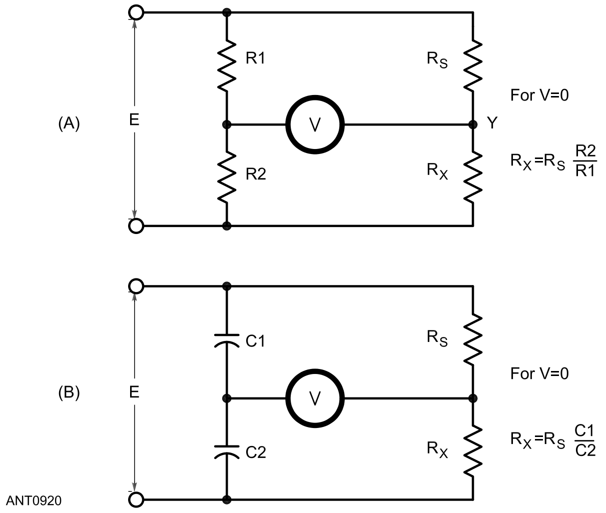

Fig 3—Bridge circuits suitable for SWR measurement. At A, Wheatstone type using resistance arms. At B, capacitance-resistance bridge ("Micromatch"). Conditions for balance are independent of frequency in both types.

Two commonly used bridge circuits are shown in **Fig 3**. The bridges consist essentially of two voltage dividers in parallel, with a voltmeter connected between the junctions of each pair of *arms*, as the individual elements are called. When the equations shown to the right of each circuit are satisfied there is no potential difference between the two junctions, and the voltmeter indicates zero voltage. The bridge is then said to be in *balance*.

Taking Fig 3A as an illustration, if R1 = R2, half the applied voltage, E, will appear across each resistor. Then if $R_S = R_X$, $^1/_2$E will appear across each of these resistors and the voltmeter reading will be zero. Remember that a matched transmission line has essentially a purely resistive input impedance. Suppose that the input terminals of such a line are substituted for R_X. Then if R_S is a resistor equal to the Z_0 of the line, the bridge will be balanced.

If the line is not perfectly matched, its input impedance will not equal Z_0 and hence will not equal R_S, since you chose the latter to be equal to Z_0. There will then be a difference in potential between points X and Y, and the voltmeter will show a reading. Such a bridge therefore can be used to show the presence of standing waves on the line, because the line input impedance will be equal to Z_0 only when there are no standing waves.

Considering the nature of the incident and reflected components of voltage that make up the actual voltage at the input terminals of the line, as discussed in Chapter 24, it should be clear that when $R_S = Z_0$, the bridge is always in balance for the incident component. Thus the voltmeter does not respond to the incident component at any time but reads only the reflected component (assuming that R2 is very small compared with the voltmeter impedance). The incident component can be measured across either R1 or R2, if they are equal resistances. The standing-wave ratio is then

$$\mathrm{SWR} = \frac{E1+E2}{E1-E2} \qquad \text{(Eq 1)}$$

where E1 is the incident voltage and E2 is the reflected voltage. It is often simpler to normalize the voltages by expressing E2 as a fraction of E1, in which case the formula becomes

$$\mathrm{SWR} = \frac{1+k}{1-k} \qquad \text{(Eq 2)}$$

where k = E2/E1.

The operation of the circuit in Fig 3B is essentially the same, although this circuit has arms containing reactance as well as resistance.

It is not necessary that R1 = R2 in Fig 3A; the bridge can be balanced, in theory, with any ratio of these two resistances provided R_S is changed accordingly. In practice, however, the accuracy is highest when the two are equal; this circuit is most commonly used.

A number of types of bridge circuits appear in **Fig 4**, many of which have been used in amateur products or amateur construction projects. All except that at G can have the generator and load at a common potential. At G, the generator and detector are at a common potential. You may interchange the positions of the detector and trans-

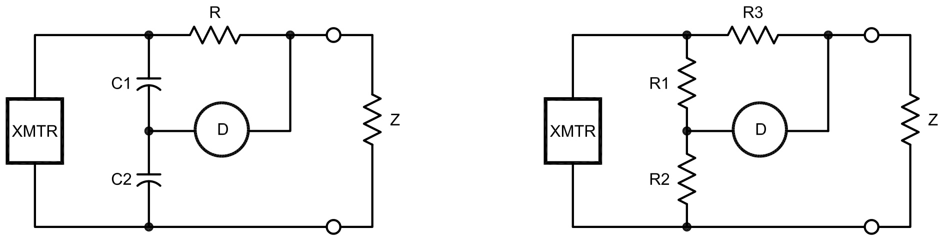
R
C1
XMTR
D
Z
C2
R3
R1
XMTR
D
Z
R2

De Sauty/Wein (Micromatch)
(A)

Balance $Z=\frac{RC1}{C2}$

Christie/Wheatstone (Antenna – Scope)
(B)

Balance $Z=\frac{R2\ R3}{R1}$

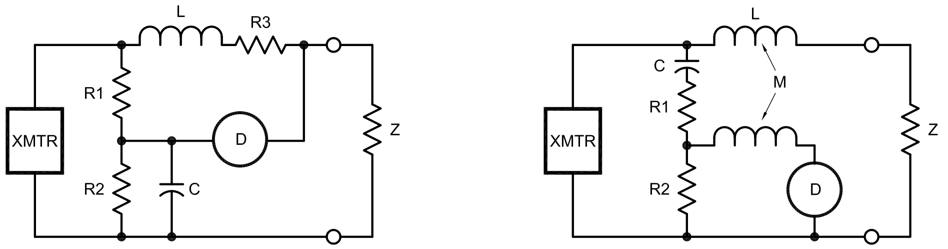
L
R3
R1
XMTR
D
Z
R2
C
L
C
M
R1
XMTR
Z
R2
D

Maxwell (Universal)
(C)

Balance R1 Z=R2 R3=L/C

No Discontinuity: R2→∞,
R3→0, R1=Z

Carey – Foster
(Twin – Lamp, Monomatch Mickey – Match)
(D)

Balance M=C R2 Z
L=M(1+R1/R2)

No Discontinuity: R1+ R2 = Z = $\sqrt{L/C}$

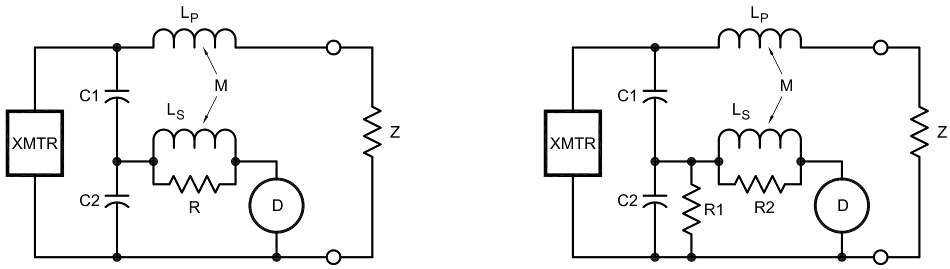
L_P
C1
M
L_S
XMTR
Z
C2
R
D
L_P
C1
M
L_S
XMTR
Z
C2
R1
R2
D

Bruene (Collins Radio)
(E)
Balance (Approx.) Z C1 L_S=M R (C1+C2)
($2\pi f L_S >> R$)
(L_P=M Approx)

Max. 1% Error: $2\pi f L_S \geq 7R$

Phase – Compensated
(F)
Balance: Z R1 C1=M=L_P
L_S=R1 R2 (C1+C2)

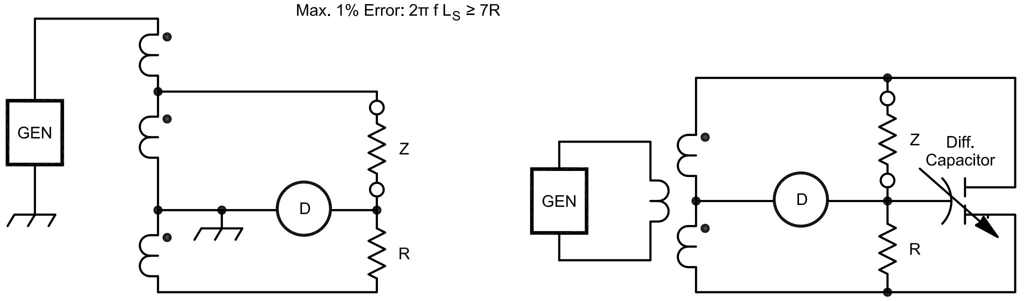
GEN
Z
D
R
GEN
Z
Diff.
Capacitor
D
R

(G)

(H)

Starr's "Hybrid Coil"

Balance: R=Z

ANT0921

(The Differential Capacitor Can Balance Parallel Reactance.)

Fig 4—Various types of SWR indicator circuits and commonly known names of bridge circuits or devices in that they have been used. Detectors (D) are usually semiconductor diodes with meters, isolated with RF chokes and capacitors. However, the detector may be a radio receiver. In each circuit, Z represents the load being measured. (*This information provided by David Geiser, WA2ANU*)

mitter (or generator) in the bridge, and this may be advantageous in some applications.

The bridges shown at D, E, F and H may have one terminal of the generator, detector and load common. Bridges at A, B, E, F, G and H have constant sensitivity over a wide frequency range. Bridges at B, C, D and H may be designed to show no discontinuity (impedance lump) with a matched line, as shown in the drawing. Discontinuities with A, E and F may be small.

Bridges are usually most sensitive when the detector bridges the midpoint of the generator voltage, as in G or H, or in B when each resistor equals the load impedance. Sensitivity also increases when the currents in each leg are equal.

Resistance Bridge

The basic bridge configuration shown in Fig 3B may be home constructed and is reasonably accurate for SWR measurement. A practical circuit for such a bridge is given in **Fig 5** and a representative layout is shown in **Fig 6**. Properly built, a bridge of this design can be used for measurement of SWRs up to about 15:1 with good accuracy.

You should observe these important construction points:

1) Keep leads in the RF circuit short, to reduce stray inductance.
2) Mount resistors two or three times their body diameter away from metal parts, to reduce stray capacitance.
3) Place the RF components so there is as little inductive and capacitive coupling as possible between the bridge arms.

In the instrument shown in Fig 6, the input and line connectors, J1 and J2, are mounted fairly close together so the standard resistor, R_S, can be supported with short leads directly between the center terminals of the connectors. R2 is mounted at right angles to R_S, and a shield partition is used between these two components and the others.

The two 47-kΩ resistors, R5 and R6 in Fig 5, are voltmeter multipliers for the 0-100 microammeter used as an indicator. This is sufficient resistance to make the voltmeter linear (that is, the meter reading is directly proportional to the RF voltage) and no voltage calibration curve is needed. D1 is the rectifier for the reflected volt-

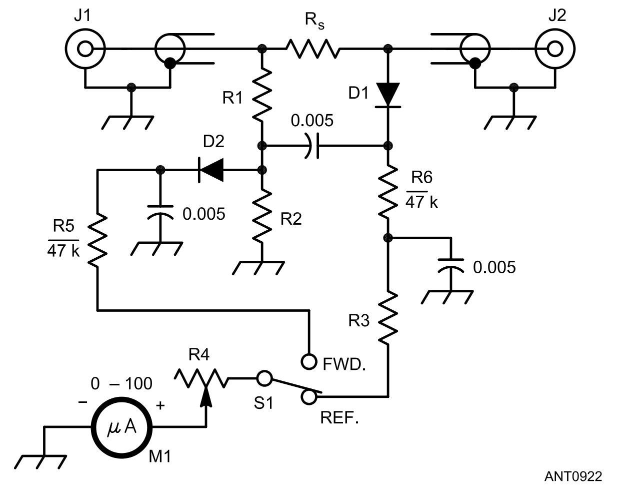

Fig 5—Resistance bridge for SWR measurement. Capacitors are disc ceramic. Resistors are ½-watt composition except as noted below.
D1, D2—Germanium diode, high back resistance type (1N34A, 1N270, etc).
J1, J2—Coaxial connectors, chassis-mounting type.
M1—0-100 dc microammeter.
R1, R2—47 Ω, ½-W composition (see text).
R3—See text.
R4—50-kΩ volume control.
R_S—Resistance equal to line Z_0 (½ or 1 W composition).
S1—SPDT toggle.

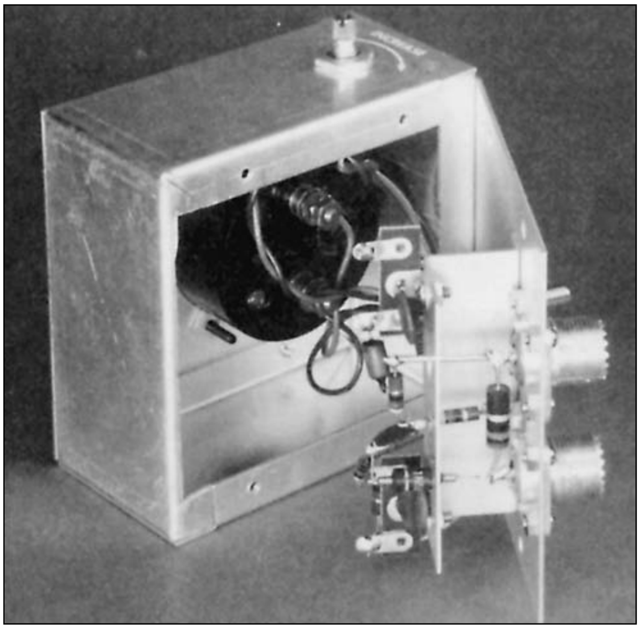

Fig 6—A 2 × 4 × 4-inch aluminum box is used to house this SWR bridge, which uses the circuit of Fig 5. The variable resistor, R4, is mounted on the side. The bridge components are mounted on one side plate of the box and a subchassis formed from a piece of aluminum. The input connector is at the top in this view. R_S is connected directly between the two center posts of the connectors. R2 is visible behind it and perpendicular to it. One terminal of D1 projects through a hole in the chassis so the lead can be connected to J2. R1 is mounted vertically to the left of the chassis in this view, with D2 connected between the junction of R1-R2 and a tie point.

age and D2 is for the incident voltage. Because of manufacturing variations in resistors and diodes, the readings may differ slightly with two multipliers of the same nominal resistance value, so a correction resistor, R3, is included in the circuit. You should select its value so that the meter reading is the same with S1 in either position, when RF is applied to the bridge with the line connection open. In the instrument shown, a value of 1000 Ω was required in series with the multiplier for reflected voltage; in other cases different values probably would be needed and R3 might have to be put in series with the multiplier for the incident voltage. You can determine this by experiment.

The value used for R1 and R2 is not critical, but you should match the two resistors within 1% or 2% if possible. Keep the resistance of R_S as close as possible to the actual Z_0 of the line you use (generally 50 or 75 Ω). Select the resistor by actual measurement with an accurate resistance bridge, if you have one available.

R4 is for adjusting the incident-voltage reading to full scale in the measurement procedure described below. Its use is not essential, but it offers a convenient alternative to exact adjustment of the RF input voltage.

Testing

Measure R1, R2 and R_S with a reliable digital ohmmeter or resistance bridge after completing the wiring. This will ensure that their values have not changed from the heat of soldering. Disconnect one side of the microammeter and leave the input and output terminals of the unit open during such measurements to avoid stray shunt paths through the rectifiers.

Check the two voltmeter circuits as described above, applying enough RF (about 10 V) to the input terminals to give a full-scale reading with the line terminals open. If necessary, try different values for R3 until the reading is the same with S1 in either position.

With J2 open, adjust the RF input voltage and R4 for full-scale reading with S1 in the incident-voltage position. Then switch S1 to the reflected-voltage position. The reading should remain at full scale. Next, short-circuit J2 by touching a screwdriver between the center terminal and the frame of the connector to make a low-inductance short. Switch S1 to the incident-voltage position and readjust R4 for full scale, if necessary. Then throw S1 to the reflected-voltage position, keeping J2 shorted, and the reading should be full scale as before. If the readings differ, R1 and R2 are not the same value, or there is stray coupling between the arms of the bridge. You must read the reflected voltage at full scale with J2 either open or shorted, when the incident voltage is set to full scale in each case, to make accurate SWR measurements.

The circuit should pass these tests at all frequencies at which it is to be used. It is sufficient to test at the lowest and highest frequencies, usually 1.8 or 3.5 and 28 or 50 MHz. If R1 and R2 are poorly matched but the bridge construction is otherwise good, discrepancies in the readings will be substantially the same at all frequencies. A difference in behavior at the low and high ends of the frequency range can be attributed to stray coupling between bridge arms, or stray inductance or capacitance in the arms.

To check the bridge for balance, apply RF and adjust R4 for full scale with J2 open. Then connect a resistor identical with R_S (the resistance should match within 1% or 2%) to the line terminals, using the shortest possible leads. It is convenient to mount the test resistor inside a cable connector (PL-259), a method of mounting that also minimizes lead inductance. When you connect the test resistor the reflected-voltage reading should drop to zero. The incident voltage should be reset to full scale by means of R4, if necessary. The reflected reading should be zero at any frequency in the range to be used. If a good null is obtained at low frequencies but some residual current shows at the high end, the trouble may be the inductance of the test resistor leads, although it may also be caused by stray coupling between the arms of the bridge itself.

If there is a constant low (but not zero) reading at all frequencies the problem is poor matching of the resistance values. Both effects can be present simultaneously. You should make sure you obtain a good null at all frequencies before using your bridge.

Bridge Operation

You must limit the RF power input to a bridge of this type to a few watts at most, because of the power-dissipation ratings of the resistors. If the transmitter has no provision for reducing power output to a very low value—less than 5 W—a simple power-absorber circuit can be made up, as shown in **Fig 7**. Lamp DS1 tends to maintain constant current through the resistor over a fairly

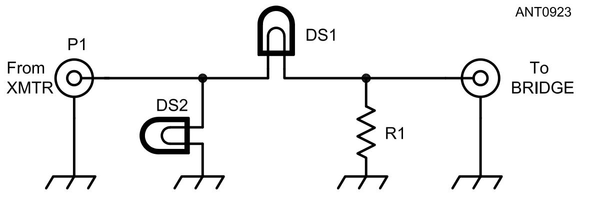

Fig 7—Power-absorber circuit for use with resistance-type SWR bridges when the transmitter has no special provisions for power reduction. For RF powers up to 50 W, DS1 is a 117-V 40-W incandescent lamp and DS2 is not used. For higher powers, use sufficient additional lamp capacity at DS2 to load the transmitter to about normal output; for example, for 250 W output DS2 may consist of two 100-W lamps in parallel. R1 is made from three 1-W 68-Ω resistors connected in parallel. P1 and P2 are cable-mounting coaxial connectors. Leads in the circuit formed by the lamps and R1 should be kept short, but convenient lengths of cable may be used between this assembly and the connectors.

wide power range, so the voltage drop across the resistor also tends to be constant. This voltage is applied to the bridge, and with the constants given is in the right range for resistance-type bridges.

To make a measurement, connect the unknown load to J2 and apply sufficient RF voltage to J1 to give a full-scale incident-voltage reading. Use R4 to set the indicator to exactly full scale. Then throw S1 to the reflected voltage position and note the meter reading. The SWR is then found by using these readings in Eq 1.

For example, if the full-scale calibration of the dc instrument is 100 µA and the reading with S2 in the reflected-voltage position is 40 µA, the SWR is

$$\text{SWR} = \frac{100+40}{100-40} = \frac{140}{60} = 2.33:1$$

Instead of calculating the SWR value, you could use the voltage curve in **Fig 8**. In this example the ratio of reflected to forward voltage is 40/100 = 0.4, and from Fig 8 the SWR value is about 2.3:1.

You may calibrate the meter scale in any arbitrary units, so long as the scale has equal divisions. It is the ratios of the voltages, and not the actual values, that determine the SWR.

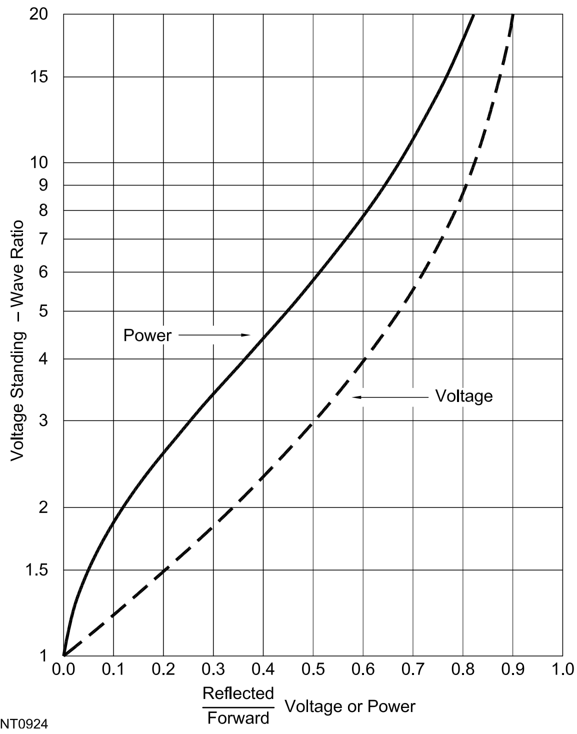

Fig 8—Chart for finding voltage standing-wave ratio when the ratio of reflected-to-forward voltage or reflected-to-forward power is known.

AVOIDING ERRORS IN SWR MEASUREMENTS

The principal causes of inaccuracies within the bridge are differences in the resistances of R1 and R2, stray inductance and capacitance in the bridge arms, and stray coupling between arms. If the checkout procedure described above is followed carefully, the bridge in Fig 5 should be amply accurate for practical use. The accuracy is highest for low standing-wave ratios because of the nature of the SWR calculation; at high ratios the divisor in the equation above represents the difference between two nearly equal quantities, so a small error in voltage measurement may mean a considerable difference in the calculated SWR.

The standard resistor R_S must equal the actual Z_0 of the line. The actual Z_0 of a sample of line may differ by a few percent from the nominal figure because of manufacturing variations, but this has to be tolerated. In the 50- to 75-Ω range, the RF resistance of a composition resistor of $^1/_2$- or 1-W rating is essentially identical with its dc resistance.

Common-Mode Currents

As explained in Chapter 26, there are two ways in which unwanted *common-mode* (sometimes called *antenna*) currents can flow on the outside of a coaxial line—currents radiated onto the line because of its spatial relationship to the antenna and currents that result from the direct connection between the coax outer conductor and (usually) one side of the antenna. The radiated current usually will not be troublesome if the bridge and the transmitter (or other source of RF power for operating the bridge) are shielded so that any RF currents flowing on the outside of the line cannot find their way into the bridge. This point can be checked by inserting an additional section of line ($^1/_8$ to $^1/_4$ electrical wavelength preferably) of the same Z_0. The SWR indicated by the bridge should not change except for a slight decrease because of the additional line loss. If there is a marked change, you may need better shielding.

Parallel-type currents caused by the connection to the antenna without using a common-mode choke balun will change the SWR with variations in line length, even though the bridge and transmitter are well-shielded and the shielding is maintained throughout the system by the use of coaxial fittings. Often, merely moving the transmission line around will cause the indicated SWR to change. This is because the outside of the coax becomes part of the antenna system—being connected to the antenna at the feed point. The outside shield of the line thus constitutes a load, along with the desired load represented by the antenna itself. The SWR on the line then is determined by the composite load of the antenna and the outside of the coax. Since changing the line length (or position) changes one component of this composite load, the SWR changes too.

The remedy for such a situation is to use a good balun or to detune the outside of the line by proper choice of length. Note that this is not a *measurement error*, since what the instrument reads is the actual SWR on the line. However, it is an undesirable condition since the line is usually operating at a higher SWR than it should—and would if the parallel-type current on the outside of the coax were eliminated.

Spurious Frequencies

Off-frequency components in the RF voltage applied to the bridge may cause considerable error. The principal components of this type are harmonics and low-frequency subharmonics that may be fed through the final stage of the transmitter driving the bridge. The antenna is almost always a fairly selective circuit, and even though the system may be operating with a very low SWR at the desired frequency, it is almost always mismatched at harmonic and subharmonic frequencies. If such spurious frequencies are applied to the bridge in appreciable amplitude, the SWR indication will be very much in error. In particular, it may not be possible to obtain a null on the bridge with any set of adjustments of the matching circuit. The only remedy is to filter out the unwanted components by increasing the selectivity of the circuits between the transmitter final amplifier and the bridge.

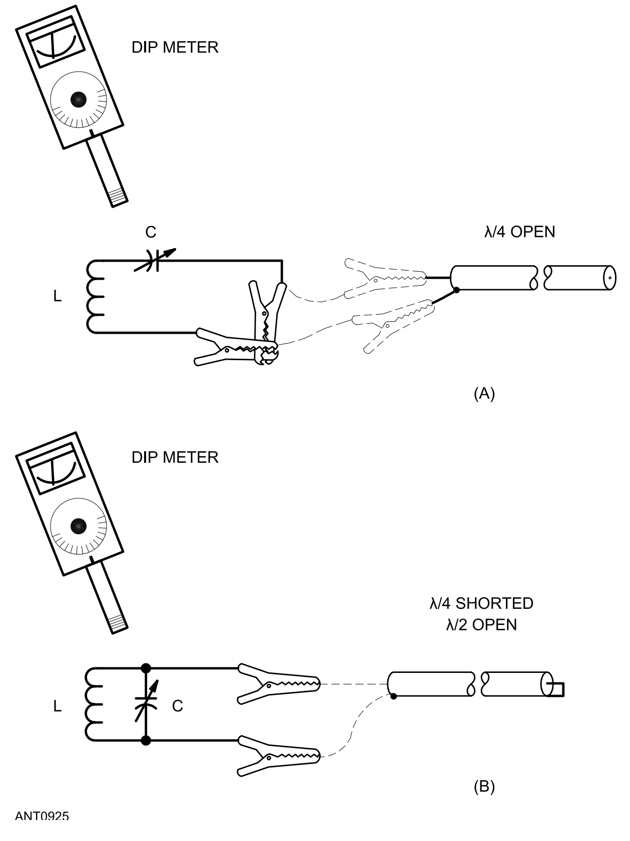

Fig 9—Methods of determining $^1/_4$ and $^1/_2$-λ line lengths. At A, $^1/_4$-λ open-circuited line; at B, $^1/_4$-λ shorted and $^1/_2$-λ open-circuited line.

MEASURING LINE LENGTH

The following material is taken from information in September 1985 *QST* by Charlie Michaels, W7XC (see Bibliography).

There is a popular myth that one may prepare an open quarter-wave line by connecting a loop of wire to one end and trimming the line to resonance (as indicated by a dip meter). This actually yields a line with capacitive reactance equal to the inductive reactance of the loop: a 4-inch wire loop yields a line 82.8° line at 18 MHz; a 2-inch loop yields an 86° line. As the loop size is reduced, line length approaches—but never equals—90°.

To make a quarter-wave open line, parallel connect a coil and capacitor that resonate at the required frequency (see **Fig 9A**). After adjusting the network to resonance, do not make further network adjustments. Open the connection between the coil and capacitor and series connect the line to the pair. Start with a line somewhat longer than required, and trim it until the circuit again resonates at the desired frequency. For a shorted quarter-wave line or an open half-wave line, connect the line in parallel with the coil and capacitor (see Fig 9B).

Another method to accurately measure a coaxial transmission line length uses one of the popular "SWR analyzers," portable hand-held instruments with a tunable low-power signal generator and an SWR bridge. While an SWR analyzer cannot compute the very high values of SWR at the input of a shorted quarter-wave line at the fundamental frequency, most include another readout showing the magnitude of the impedance. This is very handy for finding a low-impedance dip, rather than a high-impedance peak.

At the operating frequency, a shorted quarter-wave line results in a high-impedance open-circuit at the input to that line. At twice the frequency, where the line is now one-half wave long electrically, the instrument shows a low-impedance short-circuit. However, when you are pruning a line to length by cutting off short pieces at the end, it is inconvenient to have to install a short before measuring the response. It is far easier to look for the dip in impedance when a quarter-wave line is terminated in an open circuit.

Again, the strategy is to start with a line physically a little longer than a quarter-wave length. A good rule of thumb is to cut the line 5% longer to take into account the variability in the velocity factor of a typical coax cable. Compute this using:

$$\text{Length (feet)} = 0.25 \times 1.05 \times \text{VF} \times 984/\text{Freq} = \text{VF}/\text{Freq}$$

where

Freq is in MHz

VF is the velocity factor in %.

Plug the coax connector installed at one end of the line into the SWR analyzer and find the frequency for the impedance dip. Prune the line by snipping off short pieces at the end. Once you've pruned the line to the desired frequency, connect the short at the end of the line and recheck for a short circuit at twice the fundamental frequency. Seal the shorted end of the coax and you're done.

REFLECTOMETERS

Low-cost *reflectometers* that do not have a guaranteed wattmeter calibration are not ordinarily reliable for accurate numerical measurement of standing-wave ratio. They are, however, very useful as aids in the adjustment of matching networks, since the objective in such adjustment is to reduce the reflected voltage or power to zero. Relatively inexpensive devices can be used for this, since only good bridge balance is required, not actual calibration of SWR. Bridges of this type are usually frequency-sensitive that is, the meter response increase with increasing frequency for the same applied voltage. When matching and line monitoring, rather than SWR measurement, is the principal use of the device, this is not a serious handicap.

Various simple reflectometers, useful for matching and monitoring, have been described from time to time in *QST* and in *The ARRL Handbook*. Because most of these are frequency sensitive, it is difficult to calibrate them accurately for power measurement, but their low cost and suitability for use at moderate power levels, combined with the ability to show accurately when a matching circuit has been properly adjusted, make them a worthwhile addition to the amateur station.

The Tandem Match—An Accurate Directional Wattmeter

Most SWR meters are not very accurate at low power levels because the detector diodes do not respond to low voltage in a linear fashion. This design uses a compensating circuit to cancel diode nonlinearity. It also provides peak detection for SSB operation and direct SWR readout that does not vary with power level. The following information is condensed from an article by John Grebenkemper, KI6WX, in January 1987 *QST.*

DESIGN PRINCIPLES

Directional wattmeters for Amateur Radio use consist of three basic elements: a directional coupler, a detector and a signal-processing and display circuit. A directional coupler samples forward and reflected-power components on a transmission line. An ideal directional coupler would provide signals proportional to the forward and reflected voltages (independent of frequency), which could then be used to measure forward and reflected power over a wide frequency range. The best contemporary designs work over two decades of frequency.

The detector circuit provides a dc output voltage proportional to the ac input voltage. Most directional wattmeters use a single germanium diode as the detector element. A germanium, rather than silicon, diode is used to minimize diode nonlinearity at low power levels. Diode non-linearity still causes SWR measurement errors unless it is compensated ahead of the display circuit. Most directional wattmeters do not work well at low power levels because of diode nonlinearity.

The signal-processing and display circuits compute and display the SWR. There are a number of ways to perform this function. Meters that display only the forward and reflected power require the operator to compute the SWR manually. Many instruments require that the operator adjust the meter to a reference level while measuring forward power, then switch to measure reflected power on a special scale that indicates SWR. Meters that directly compute the SWR using analog signal-processing circuits have been described by Fayman, Perras,

Table 1
Performance Specifications for the Tandem Match

Power range: 1.5 to 1500 W
Frequency range: 1.8 to 54 MHz
Power accuracy: Better than ± 10% (± 0.4 dB)
SWR accuracy: Better than ± 5%
Minimum SWR: Less than 1.05:1
Power display: Linear, suitable for use with either analog or digital meters
Calibration: Requires only an accurate voltmeter

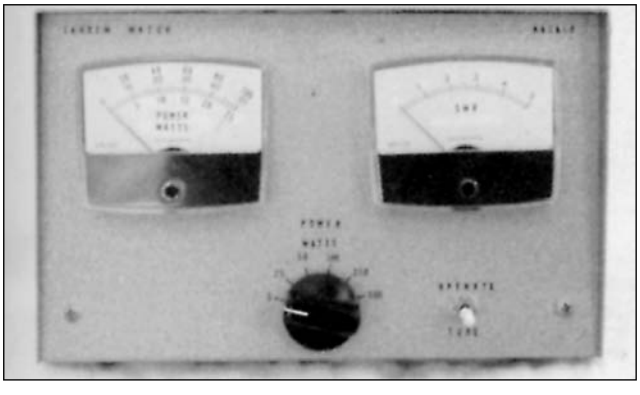

Fig 10—The Tandem Match uses a pair of meters to display net forward power and true SWR simultaneously.

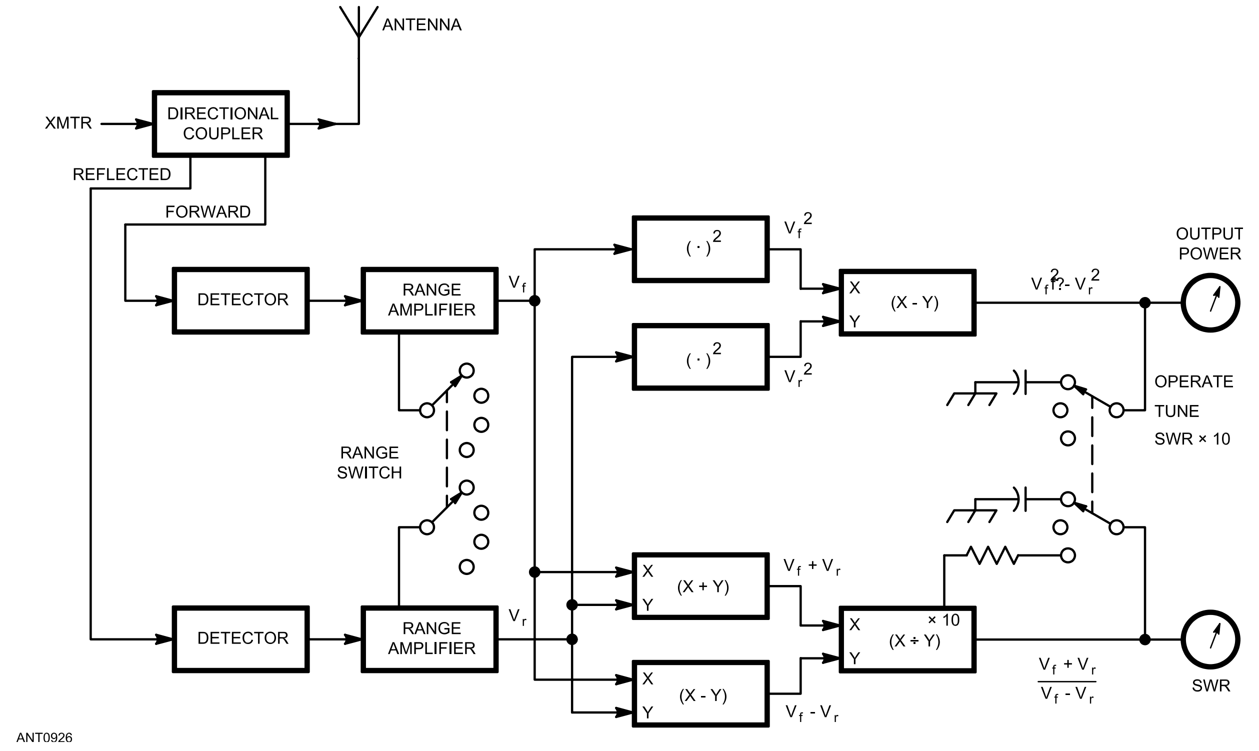

Fig 11—Block diagram of the Tandem Match.

Leenerts and Bailey (see the Bibliography at the end of this chapter).

The next section takes a brief look at several popular circuits that accomplish the functions above and compares them to the circuits used in the Tandem Match. The design specifications of the Tandem Match are shown in **Table l**, and a block diagram is shown in **Fig 11**.

CIRCUIT DESCRIPTION

A directional coupler consists of an input port, an output port and a coupled port. The device takes a portion of the power flowing from the input port to the output port and directs it to the coupled port, but *none* of the power flowing from the output port to the input port is directed to the coupled port.

There are several terms that define the performance of a directional coupler:

1) *Insertion loss* is the amount of power that is lost as the signal flows from the input port to the output port. Insertion loss should be minimized so the coupler doesn't dissipate a significant amount of the transmitted power.

2) *Coupling factor* is the amount of power (or voltage) that appears at the coupled port relative to the amount of power (or voltage) transferred from the input port to the output port. The "flatness" (with frequency) of the coupling factor determines how accurately the directional wattmeter can determine forward and reflected power over a range of frequencies.

3) *Isolation* is the amount of power (or voltage) that appears at the coupled port relative to the amount of power (or voltage) transferred from the output port to the input port.

4) *Directivity* is the isolation less the coupling factor. Directivity dictates the minimum measurable SWR. A directional coupler with 20 dB of directivity measures a 1:1 SWR as l.22:l, but one with 30 dB measures a 1:1 SWR as 1.07:1.

The directional coupler most commonly used in amateur radio was first described in 1959 by Bruene in *QST* (see Bibliography). The coupling factor was fairly flat (± l dB), and the directivity was about 20 dB for a Bruene coupler measured from 3 to 30 MHz. Both factors limit the accuracy of the Bruene coupler for measuring low values of power and SWR. It is a simple directional coupler, however, and it works well over a wide frequency range if great precision is not required.

The coupler used in the Tandem Match (see **Fig 12**) consists of a pair of toroidal transformers connected in tandem. The configuration was patented by Carl G. Sontheimer and Raymond E. Fredrick (US Patent no. 3,426,298, issued February 4, 1969). It has been described by Perras, Spaulding (see Bibliography) and others. With coupling factors of 20 dB or greater, this coupler is suitable for sampling both forward and reflected power.

The configuration used in the Tandem Match works well over the frequency range of 1.8 to 54 MHz, with a

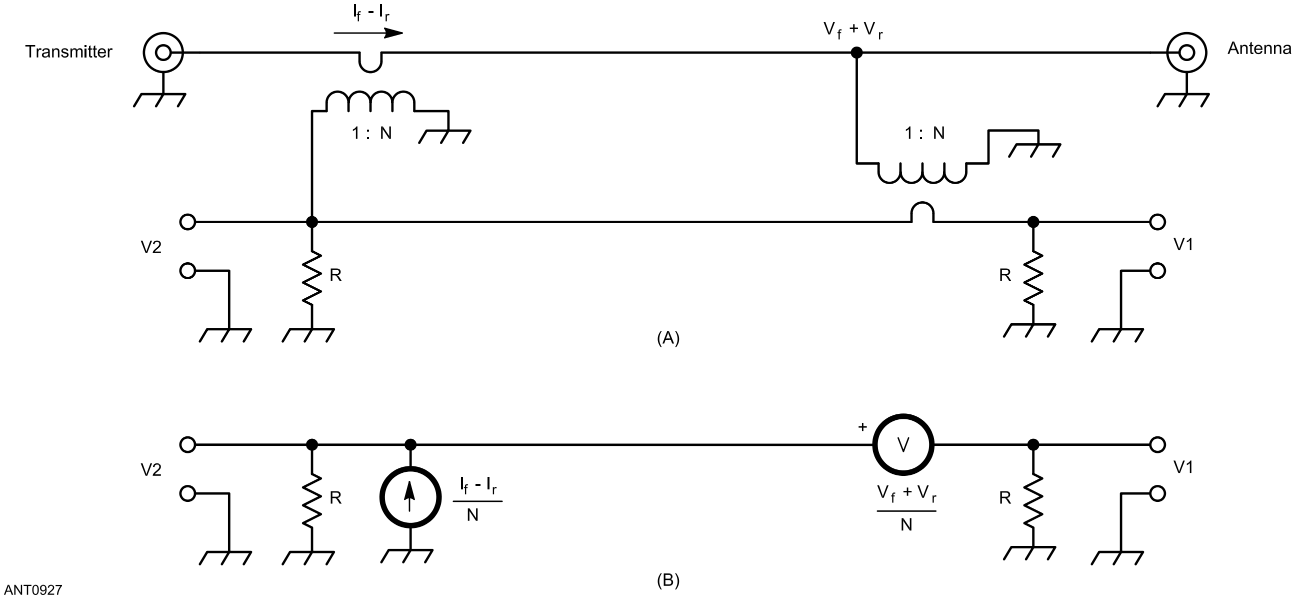
$I_f - I_r$
Transmitter
$V_f + V_r$
Antenna
1 : N
1 : N
V2
R
R
V1
(A)
V2
R
+
V
$\frac{I_f - I_r}{N}$
$\frac{V_f + V_r}{N}$
R
V1
(B)
ANT0927

Fig 12—Simplified diagram of the Tandem Match directional coupler. At A, a schematic of the two transformers. At B, an equivalent circuit.

nominal coupling factor of 30 dB. Over this range, insertion loss is less than 0.1 dB. The coupling factor is flat to within ± 0.1 dB from 1.8 to 30 MHz, and increases to only ±0.3 dB at 50 MHz. Directivity exceeds 35 dB from 1.8 to 30 MHz and exceeds 26 dB at 50 MHz.

The low-frequency limit of this directional coupler is determined by the inductance of the transformer secondary windings. The inductive reactance should be greater than 150 Ω (three times the line characteristic impedance) to reduce insertion loss. The high-frequency limit of this directional coupler is determined by the length of the transformer windings. When the winding length approaches a significant fraction of a wavelength, coupler performance deteriorates.

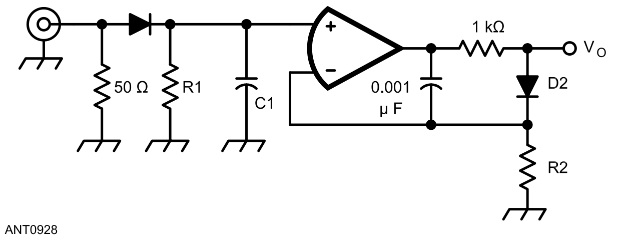
50 Ω
R1
C1
+
–
1 kΩ
V_O
0.001 µ F
D2
R2
ANT0928

Fig 13—Simplified diagram of the detector circuit used in the Tandem Match. The output voltage, V_O is approximately equal to the input voltage. D1 and D2 must be a matched pair (see text). The op amp should have a low offset voltage (less than 1 mV), a low leakage current (less than 1 nA), and be stable over time and temperature. The resistor and capacitor in the feedback path assure that the op amp will be stable.

The coupler described here may overheat at 1500 W on 160 meters (because of the high circulating current in the secondary of T2). The problem could be corrected by using a larger core or one with greater permeability. A larger core would require longer windings; that option would decrease the high-frequency limit.

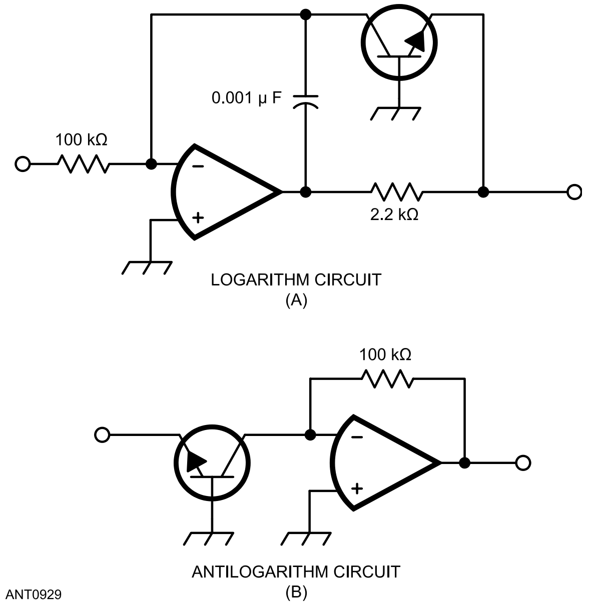
0.001 µ F
100 kΩ
–
+
2.2 kΩ
LOGARITHM CIRCUIT
(A)
100 kΩ
–
+
ANTILOGARITHM CIRCUIT
(B)
ANT0929

Fig 14—Simplified diagrams of the log circuit at A and the antilog circuit at B.

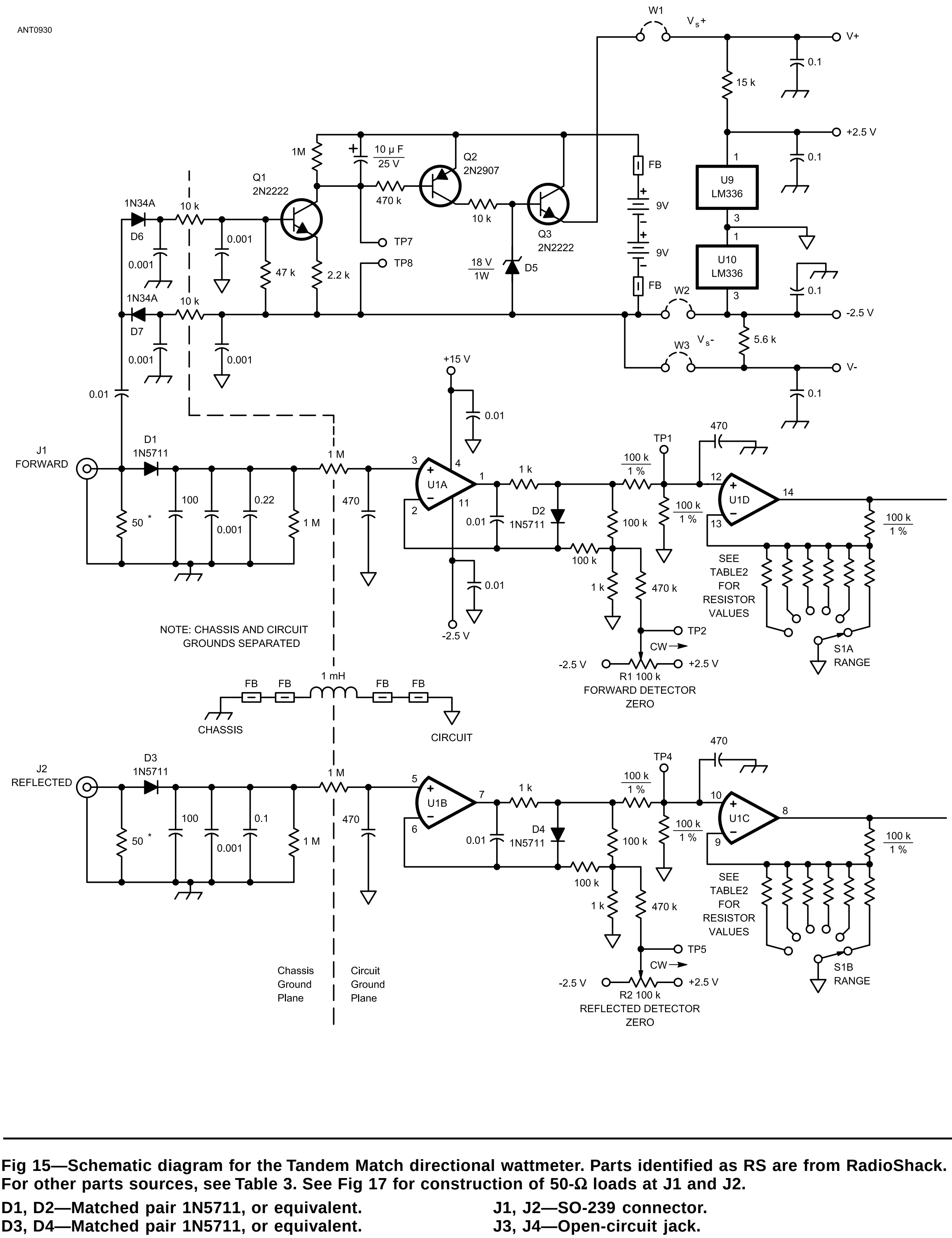
ANT0930
W1
Vs+
V+
0.1
15 k
+2.5 V
0.1
1M
10 µ F
25 V
Q2
2N2907
Q1
2N2222
470 k
FB
U9
LM336
1N34A
10 k
10 k
9V
D6
0.001
0.001
TP7
Q3
2N2222
3
1
TP8
18 V
1W
D5
9V
U10
LM336
0.001
47 k
2.2 k
FB
W2
0.1
1N34A
10 k
3
-2.5 V
D7
W3
Vs-
5.6 k
0.001
0.001
+15 V
V-
0.01
0.1
0.01
470
J1
FORWARD
D1
1N5711
TP1
1 M
100 k
1 %
3
4
1 k
12
14
U1A
1
U1D
100
0.22
470
2
11
100 k
1 %
13
100 k
1 %
50 *
0.001
1 M
0.01
D2
1N5711
100 k
100 k
SEE
TABLE2
FOR
RESISTOR
VALUES
0.01
1 k
470 k
NOTE: CHASSIS AND CIRCUIT
GROUNDS SEPARATED
-2.5 V
TP2
CW
-2.5 V
+2.5 V
S1A
RANGE
R1 100 k
FORWARD DETECTOR
ZERO
1 mH
FB
FB
FB
FB
CHASSIS
CIRCUIT
470
J2
REFLECTED
D3
1N5711
TP4
1 M
100 k
1 %
5
1 k
10
U1B
7
U1C
8
100
0.1
470
6
100 k
1 %
9
100 k
1 %
50 *
0.001
1 M
0.01
D4
1N5711
100 k
100 k
SEE
TABLE2
FOR
RESISTOR
VALUES
1 k
470 k
TP5
CW
S1B
RANGE
-2.5 V
+2.5 V
R2 100 k
REFLECTED DETECTOR
ZERO
Chassis
Ground
Plane
Circuit
Ground
Plane

Fig 15—Schematic diagram for the Tandem Match directional wattmeter. Parts identified as RS are from RadioShack. For other parts sources, see Table 3. See Fig 17 for construction of 50-Ω loads at J1 and J2.

D1, D2—Matched pair 1N5711, or equivalent.
D3, D4—Matched pair 1N5711, or equivalent.
D6, D7—1N34A.
D8-D14—1N914.
FB—Ferrite bead, Amidon FB-73-101 or equiv.
J1, J2—SO-239 connector.
J3, J4—Open-circuit jack.
M1, M2—50 µA panel meter, RS 270-1751.
Q1, Q3, Q4—2N2222 or equiv.
Q2—2N2907 or equiv.

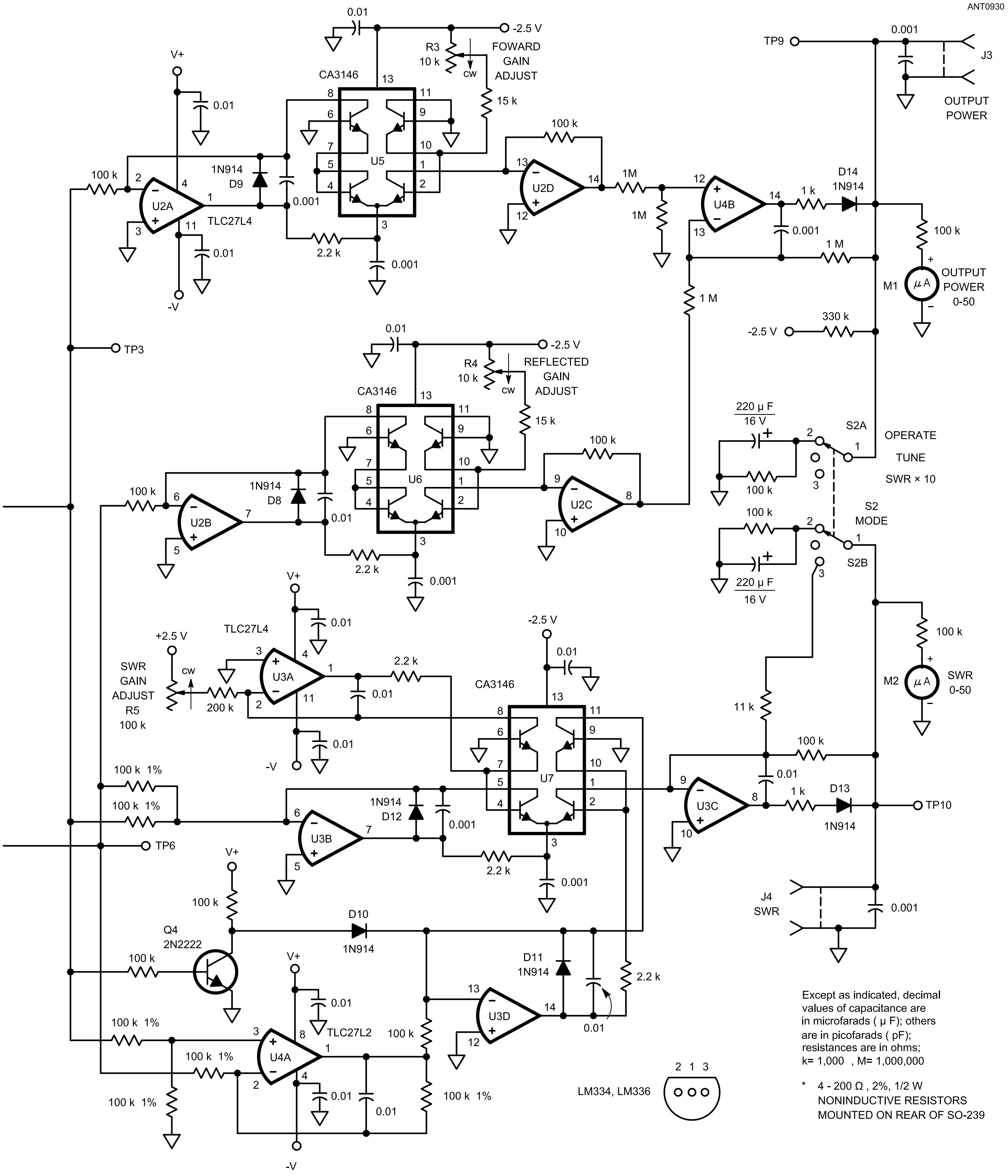
ANT0930
0.01
-2.5 V
R3
10 k
cw
FOWARD
GAIN
ADJUST
15 k
CA3146
13
8
11
6
9
7
10
5
1
4
2
3
U5
V+
0.01
100 k
2
4
U2A
1
3
11
1N914
D9
0.001
TLC27L4
0.01
-V
2.2 k
0.001
100 k
13
U2D
12
14
1M
1M
12
U4B
13
14
1 k
D14
1N914
0.001
1 M
1 M
TP9
0.001
J3
OUTPUT
POWER
100 k
M1
μA
OUTPUT
POWER
0-50
-2.5 V
330 k
TP3
0.01
-2.5 V
R4
10 k
cw
REFLECTED
GAIN
ADJUST
15 k
CA3146
13
8
11
6
9
7
10
5
1
4
2
3
U6
100 k
6
U2B
7
5
1N914
D8
0.01
2.2 k
0.001
100 k
9
U2C
8
10
220 μ F
16 V
2
S2A
1
OPERATE
TUNE
SWR × 10
100 k
3
100 k
2
S2
MODE
1
S2B
3
220 μ F
16 V
V+
0.01
+2.5 V
TLC27L4
3
4
U3A
1
2
11
SWR
GAIN
ADJUST
R5
100 k
cw
200 k
2.2 k
0.01
0.01
-V
-2.5 V
0.01
CA3146
13
8
11
6
9
7
10
5
1
4
2
3
U7
100 k
M2
μA
SWR
0-50
11 k
100 k
0.01
9
U3C
8
10
1 k
D13
1N914
TP10
100 k 1%
100 k 1%
6
U3B
7
5
1N914
D12
0.001
2.2 k
0.001
TP6
V+
100 k
Q4
2N2222
100 k
D10
1N914
D11
1N914
13
U3D
14
12
0.01
2.2 k
J4
SWR
0.001
V+
0.01
100 k 1%
3
8
TLC27L2
U4A
1
2
4
100 k 1%
100 k
100 k 1%
0.01
0.01
100 k 1%
-V
2 1 3
LM334, LM336

Except as indicated, decimal values of capacitance are in microfarads (μ F); others are in picofarads (pF); resistances are in ohms; k= 1,000 , M= 1,000,000

* 4 - 200 Ω , 2%, 1/2 W NONINDUCTIVE RESISTORS MOUNTED ON REAR OF SO-239

R1, R2, R5—100 kΩ, 10-turn, cermet Trimpot.
R3, R4—10 kΩ, 10-turn, cermet Trimpot.
U1-U3—TLC27L4 or TLC27M4 quad op amp (Texas Instruments).
U4—TLC27L2 or TLC27M2 dual op amp.
U5-U7—CA3146 quad transistor array.
U8—LM334 adjustable current source.
U9-U10—LM336 2.5-V reference diode. See text.

Detector Circuits

Most amateur directional wattmeters use a germanium diode detector to minimize the forward voltage drop. Detector voltage drop is still significant, however, and an uncompensated diode detector does not respond to small signals in a linear fashion. Many directional wattmeters compensate for diode nonlinearity by adjusting the meter scale.

The effect of underestimating detected power worsens at low power levels. Under these conditions, the ratio of the forward power to the reflected power is overestimated because the reflected power is always less than the forward power. This results in an instrument that underestimates SWR, particularly as power is reduced. A directional wattmeter can be checked for this effect by measuring SWR at several power levels: the SWR should be independent of power level.

The Tandem Match uses a feedback circuit to compensate for diode nonlinearity. A simplified diagram of the compensated detector is shown in **Fig 13**. When used with the 30-dB directional coupler, the output voltage of this circuit tracks the square root of power over a range from 10 mW to 1.5 kW. The compensated diode detector tracks the peak input voltage down to 30 mV, while an uncompensated germanium-diode detector shows significant errors at peak inputs of 1 V and less. More information about compensated detectors appears in Grebenkemper's *QEX* article, "Calibrating Diode Detectors" (see Bibliography).

The compensation circuit uses the voltage across a feedback diode, D2, to compensate for the voltage drop across the detector diode, D1. (The diodes must be a matched pair.) The average current through D1 is determined by the detector diode load resistor, R1. The peak current through this diode is several times larger than the average current; therefore, the current through D2 must be several times larger than the average current through D1 to compensate adequately for the peal voltage drop across D1. This is accomplished by making the feedback-diode load resistor, R2, several times smaller than R1. The voltage at the output of the compensated detector approximates the peak RF voltage at the input. For Schottky barrier diodes and a 1 MΩ detector-diode load resistor, a 5:1 ratio of R1 to R2 is nearly optimal.

Signal-Processing and Display Circuits

The signal-processing circuitry calculates and displays transmission-line power and SWR. When measuring forward power, most directional wattmeters display the actual forward power present in the transmission line, which is the sum of forward and reflected power if a match exists at the input end of the line. Transmission-line forward power is very close to the net forward power (the actual power delivered to the line) so long as the SWR is low. As the SWR increases, however, forward power becomes an increasingly poor measure of the power delivered to the load. At an SWR of 3:1, a forward power reading of 100 W implies that only 75 W is delivered to the load (the reflected power is 25 W), assuming the transmission-line loss is zero.

The Tandem Match differs from most wattmeters in that *it displays the net forward power, rather than the sum of forward and reflected power.* This is the quantity that must be optimized to result in maximum radiated power (and which concerns the FCC).

The Tandem Match directly computes and displays the transmission-line SWR on a linear scale. As the displayed SWR is not affected by changes in transmitter power, a matching network can be simply adjusted to minimize SWR. Transmatch adjustment requires only a few watts.

The heart of the Tandem Match signal-processing circuit is the analog logarithm and antilogarithm circuitry shown in **Fig 14**. The circuit is based on the fact that collector current in a silicon transistor is proportional to the exponential (antilog) of its base-emitter voltage over a range of collector currents from a few nanoamperes to a few milliamperes when the collector-base voltage is zero (see Gibbons and Horn reference in the Bibliography). Variations of this circuit are used in the squaring circuits to convert voltage to power and in the divider circuit used to compute the SWR. With good op amps, this circuit will work well for input voltages from less than 100 mV to greater than 10 V.

(For the Tandem Match, "good" op amps are quad-packaged, low-power-consumption, unity-gain-stable parts with input bias less than 1 nA and offset voltage less than 5 mV. Op amps that consume more power than those shown may require changes to the power supply.)

CONSTRUCTION

The schematic diagram for the Tandem Match is shown in **Fig 15** (see pages 14 and 15). The circuit is designed to operate from batteries and draw very little power. Much of the circuitry is of high impedance, so take care to isolate it from RF fields. House it in a metal case. Most problems in the prototype were caused by stray RF in the op-amp circuitry.

Directional Coupler

The directional coupler is constructed in its own small ($2\frac{3}{4} \times 2\frac{3}{4} \times 2\frac{1}{4}$-inch) aluminum box (see **Fig 16**). Two pairs of S0-239 connectors are mounted on opposite sides of the box. A piece of PC board is run diagonally across the box to improve coupler directivity. The pieces of RG-8X coaxial cable pass through holes in the PC board.

(Note: Some brands of "mini 8" cable have extremely low breakdown voltage ratings and are unsuitable to carry even 100 W when the SWR exceeds 1:1. See the subsequent section, "High-Power Operation," for details of a coupler made with RG-8 cable.)

Begin by constructing T1 and T2, which are identi-

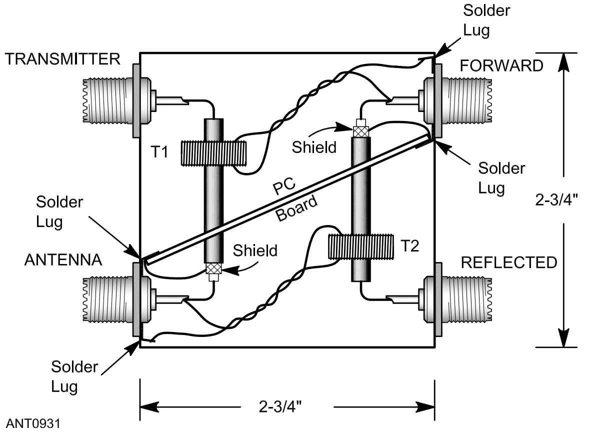

Fig 16—Construction details for the directional coupler.

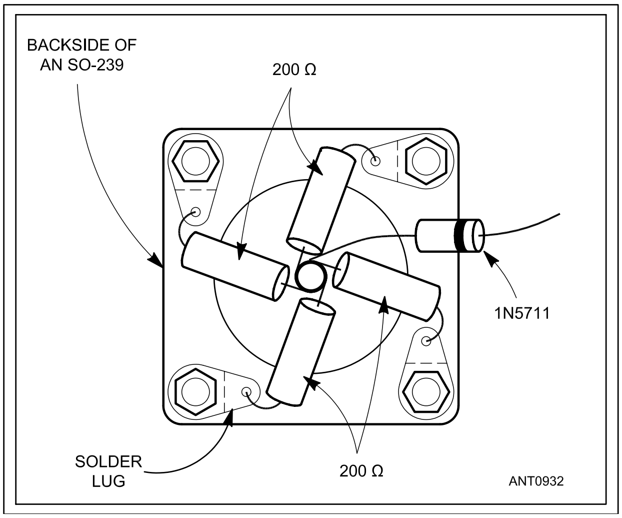

Fig 17—The parallel load resistors mounted on an SO-239 connector. Four 200-Ω, 2%, $^1/_2$-W resistors are mounted in parallel to provide a 50-Ω detector load.

cal except for their end connections. Refer to Fig 16. The primary for each transformer is the center conductor of a length of RG-8X coaxial cable. Cut two cable lengths sufficient for mounting as shown in the figure. Strip the cable jacket, braid and dielectric as shown. The cable braid is used as a Faraday shield between the transformer windings, so it is only grounded at one end. *Important—connect the braid only at one end or the directional-coupler circuit will not work properly!* Wind two transformer secondaries, each 31 turns of #24 enameled wire on an Amidon T50-3 or equivalent powdered-iron core.

Slip each core over one of the prepared cable pieces (including both the shield and the outer insulation). Mount and connect the transformers as shown in Fig 16, with the wire running through separate holes in the copper-clad PC board. The directional coupler can be mounted separately from the rest of the circuitry if desired. If so, use two coaxial cables to carry the forward and reflected-power signals from the directional coupler to the detector inputs. Be aware, however that any losses in the cables will affect power readings.

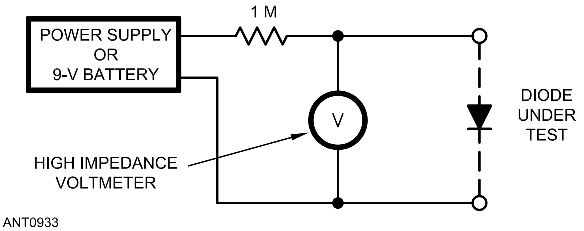

Fig 18—Diode matching test setup.

This directional coupler has not been used at power levels in excess of 100 W. For more information about using the Tandem Match at high power levels, see the section, “High-Power Operation.”

Detector and Signal-Processing Circuits

The detector and signal-processing circuits were constructed on a perforated, copper-clad circuit board. These circuits use two separate grounds—*it is extremely important that the grounds be isolated as shown in the circuit diagram.* Failure to do so may result in faulty circuit operation. Separate grounds prevent RF currents on the cable braid from affecting the op-amp circuitry.

The directional coupler requires good 50-Ω loads. They are constructed on the back of female UHF chassis connectors where the cables from the directional coupler enter the wattmeter housing. Each load consists of four 200-Ω resistors connected from the center conductor of the UHF connector to the four holes on the mounting flange, as shown in **Fig 17**. The detector diode is then run from the center conductor of the connector to the 100-pF and 1000-pF bypass capacitors, which are mounted next to the connector. The response of this load and detector combination measures flat to beyond 500 MHz.

Schottky-barrier diodes (type 1N5711) were used in this design because they were readily available. Any RF-detector diode with a low forward voltage drop (less than 300 mV) and reverse break-down voltage greater than 30 V could be used. (Germanium diodes could be used in this circuit, but performance will suffer. If germanium diodes are used, reduce the resistance values for the detector-diode and feedback-diode load resistors by a factor of 10.)

The detector diodes must be matched. This can be done with dc, using the circuit shown in **Fig 18**. Use a high-impedance voltmeter (10 MΩ or greater). For this project, diodes are matched when their forward voltage drops are equal (within a few millivolts). Diodes from the same batch will probably be sufficiently matched.

Table 2
Range-Switch Resistor Values

Full-Scale Power Level (W)	*Range Resistor (1% Precision) (kΩ)*
1	2.32
2	3.24
3	4.02
5	5.23
10	7.68
15	9.53
20	11.0
25	12.7
30	14.0
50	18.7
100	28.7
150	37.4
200	46.4
250	54 9
300	63.4
500	100.0
1000	237.0
1500	649.0
2000	Open

The rest of the circuit layout is not critical, but keep the lead lengths of the 0.001 and 0.01-pF bypass capacitors short. The capacitors provide additional bypassing for the op-amp circuitry. D6 and D7 form a voltage doubler to detect the presence of a carrier. When the forward power exceeds 1.5 W, Q3 switches on and stays on until about 10 seconds after the carrier drops. (A connection from TP7 to TP9 forces the unit on, even with no carrier present.) The regulated references of +2.5 V and –2.5 V generated by the LM334 and two LM336s are critical. Zener-diode substitutes would significantly degrade performance.

The four op amps in U1 compensate for the nonlinearity of the detector diodes. D1-D2 and D3-D4 are the matched diode pairs discussed above. A RANGE switch selects the meter range. (A six-position switch was used here because it was handy.) The resistor values for the RANGE switch are shown in **Table 2**. Full-scale input power gives an output at U1C or U1D of 7.07 V. The forward and reflected-power detectors are zeroed with R1 and R2.

The forward and reflected-detector voltages are squared by U2, U5 and U6 so that the output voltages are proportional to forward and reflected power. The gain constants are adjusted using R3 and R4 so that an input of 7.07 V to the squaring circuit gives an output of 5 V. The difference between these two voltages is used by U4B to yield an output that is proportional to the power delivered to the transmission line. This voltage is peak detected (by an RC circuit connected to the OPERATE position of the MODE switch) to hold and indicate the maximum power during CW or SSB transmissions. SWR is computed from the forward and reflected voltages by U3, U4 and U7. When no carrier is present, Q4 forces the SWR reading to be zero (that is, when the forward power is less than 2% of the full-scale setting of the RANGE switch). The SWR computation circuit gain is adjusted by R5. The output is peak detected in the OPERATE mode to steady the SWR reading during CW or SSB transmissions.

Transistor arrays (U5, U6 and U7) are used for the log and antilog circuits to guarantee that the transistors will be well matched. Discrete transistors may be used, but accuracy may suffer. A three-position toggle switch selects the three operating modes. In the OPERATE mode, the power and SWR outputs are peak detected and held for a few seconds to allow meter reading during actual transmissions. In the TUNE mode, the meters display instantaneous output power and SWR.

A digital voltmeter is used to obtain more precise readings than are possible with analog meters. The output power range is 0 to 5 V (0 V = 0 W and 5 V = full scale). SWR output varies from 1 V (SWR = 1:1) to 5 V (SWR = 5:1). Voltages above 5 V are unreliable because of voltage limiting in some of the op amp circuits.

Calibration

The directional wattmeter can be calibrated with an accurate voltmeter. All calibration is done with dc voltages. The directional-coupler and detector circuits are inherently accurate if correctly built. To calibrate the wattmeter, use the following procedure:

1) Set the MODE switch to TUNE and the RANGE switch to 100 W or less.
2) Jumper TP7 to TP8. This turns the unit on.
3) Jumper TP1 to TP2. Adjust R1 for 0 V at TP3.
4) Jumper TP4 to TP5. Adjust R2 for 0 V at TP6.
5) Adjust R1 for 7.07 V at TP3.
6) Adjust R3 for 5.00 V at TP9, or a full-scale reading on M1.
7) Adjust R2 for 7.07 V at TP6.
8) Adjust R4 for 0 V at TP9, or a zero reading on M1.
9) Adjust R2 for 4.71 V at TP6.
10) Adjust R5 for 5.00 V at TP10, or a full-scale reading on M2.
11) Set the RANGE switch to its most sensitive scale.
12) Remove the jumpers from TP1 to TP2 and TP4 to TP5.
13) Adjust R1 for 0 V at TP3.
14) Adjust R2 for 0 V at TP6.
15) Remove the jumper from TP7 to TP8.

This completes the calibration procedure. This procedure has been found to equal calibration with expensive laboratory equipment. The directional wattmeter should now be ready for use.

ACCURACY

Performance of the Tandem Match has been compared to other well-known directional couplers and laboratory test equipment, and it equals any amateur directional wattmeter tested. Power measurement accuracy compares well to a Hewlett-Packard HP-436A power meter. The HP meter has a specified measurement error of less than ± 0.05 dB. The Tandem Match tracked the HP436A within +0.5 dB from 10 mW to 100 W, and within ± 0.1 dB from 1 W to 100 W. The unit was not tested above 100 W because a transmitter with a higher power rating was not available.

SWR performance was equally good when compared to the SWR calculated from measurements made with the HP436A and a calibrated directional coupler. The Tandem Match tracked the calculated SWR within ± 5% for SWR values from 1:1 to 5:1. SWR measurements were made at 8 W and 100 W.

OPERATION

Connect the Tandem Match in the 50-Ω line between the transmitter and the antenna matching network (or antenna if no matching network is used). Set the RANGE switch to a range greater than the transmitter output rating and the MODE switch to TUNE. When the transmitter is keyed, the Tandem Match automatically switches on and indicates both power delivered to the antenna and SWR on the transmission line. When no carrier is present, the OUTPUT POWER and SWR meters indicate zero.

The OPERATE mode includes RC circuitry to momentarily hold the peak-power and SWR readings during CW or SSB transmissions. The peak detectors are not ideal, so there could be about 10% variation from the actual power peaks and the SWR reading. The SWR ×10 mode increases the maximum readable SWR to 50:1. This range should be sufficient to cover any SWR value that occurs in amateur use. (A 50-foot open stub of RG-8 yields a measured SWR of only 43:1, or less, at 2.4 MHz because of cable loss. Higher frequencies and longer cables exhibit a lesser maximum SWR.)

It is easy to use the Tandem Match to adjust an antenna matching network: Adjust the transmitter for minimum output power (at least 1.5 W). With the carrier on and the MODE switch set to TUNE or SWR ×10, adjust the matching network for minimum SWR. Once the minimum SWR is obtained, set the transmitter to the proper operating mode and output power. Place the Tandem Match in the OPERATE mode.

DESIGN VARIATIONS

There are several ways in which this design could be enhanced. The most important is to add UHF capability. This would require a new directional-coupler design for the band of interest. (The existing detector circuit should work to at least 500 MHz.)

Those who desire a low-power directional wattmeter can build a directional coupler with a 20-dB coupling factor by decreasing the transformer turns ratio to 10:1. That version should be capable of measuring output power from 1 mW to about 150 W (and it should switch on at about 150 mW).

This change should also increase the maximum operating frequency to about 150 MHz (by virtue of the shorter transformer windings). If you desire 1.8-MHz operation, it may be necessary to change the toroidal core material for sufficient reactance (low insertion loss).

The Tandem Match circuit can accommodate coaxial cable with a characteristic impedance other than 50 Ω. The detector terminating resistors, transformer secondaries and range resistors must change to match the new design impedance.

The detector circuitry can be used (without the directional coupler) to measure low-level RF power in 50-Ω circuits. RF is fed directly to the forward detector (J1, Fig 15), and power is read from the output power meter. The detector is quite linear from 10 µW to 1.5 W.

HIGH-POWER OPERATION

This material was condensed from information by Frank Van Zant, KL7IBA, in July 1989 *QST.* In April 1988, Zack Lau, W1VT, described a directional-coupler circuit (based on the same principle as Grebenkemper's circuit) for a QRP transceiver (see the Bibliography at the end of this chapter). The main advantage of Lau's circuit is a very low parts count.

Grebenkemper used complex log-antilog amplifiers to provide good measurement accuracy. This application gets away from complex circuitry, but retains reasonable measurement accuracy over the 1 to 1500-W range. It also forfeits the SWR-computation feature. Lau's coupler uses ferrite toroids. It works well at low power levels, but the ferrite toroids heat excessively with high power, causing erratic meter readings and the potential for burned parts.

The Revised Design

Powdered-iron toroids are used for the transformers in this version of Lau's basic circuit. The number of turns on the secondaries was increased to compensate for the lower permeability of powdered iron.

Two meters display reflected and forward power (see **Fig 19**). The germanium detector diodes (D1 and D2—1N34) provide fairly accurate meter readings, particularly if the meter is calibrated (using R3, R4 and R5) to place the normal transmitter output at mid scale. If the winding sense of the transformers is reversed, the meters are transposed (the forward-power meter becomes the reflected-power meter, and vice versa).

Construction

Fig 20 shows the physical layout of the coupler. The pickup unit is mounted in a $3^1/_2 \times 3^1/_2 \times$ 4-inch box. The meters, PC-mount potentiometers and HIGH/LOW power

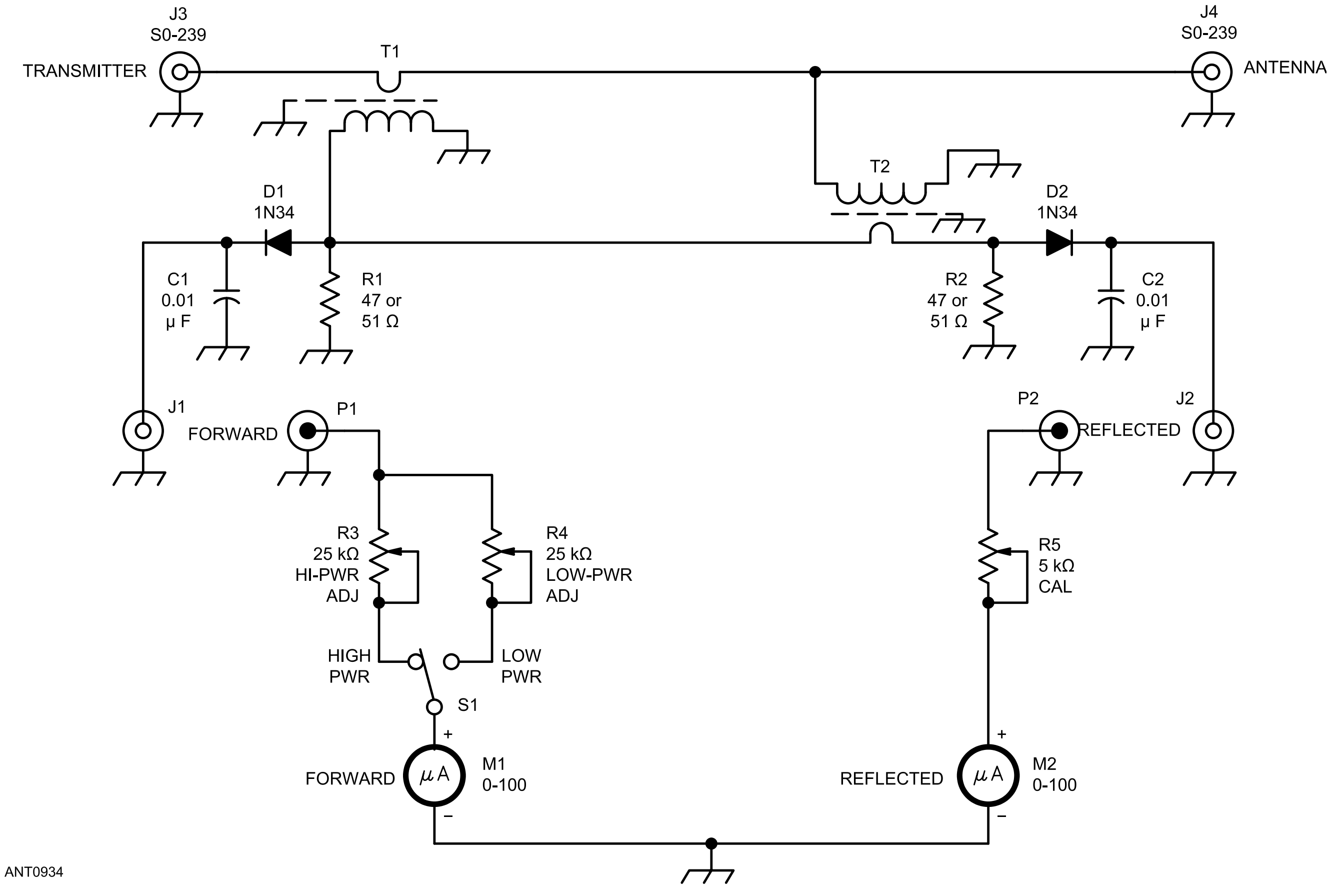
J3
S0-239
TRANSMITTER
T1
J4
S0-239
ANTENNA
T2
D1
1N34
D2
1N34
C1
0.01
μ F
R1
47 or
51 Ω
R2
47 or
51 Ω
C2
0.01
μ F
J1
FORWARD
P1
P2
REFLECTED
J2
R3
25 kΩ
HI-PWR
ADJ
R4
25 kΩ
LOW-PWR
ADJ
R5
5 kΩ
CAL
HIGH
PWR
LOW
PWR
S1
+
FORWARD
μA
M1
0-100
–
+
REFLECTED
μA
M2
0-100
–
ANT0934

Fig 19—Schematic diagram of the high-power directional coupler. D1 and D2 are germanium diodes (1N34 or equiv). R1 and R2 are 47 or 51-Ω, $^1/_2$-W resistors. C1 and C2 have 500-V ratings. The secondary windings of T1 and T2 each consist of 40 turns of #26 to #30 enameled wire on T-68-2 powdered-iron toroid cores. If the coupler is built into an existing antenna tuner, the primary of T1 can be part of the tuner coaxial output line. The remotely located meters (M1 and M2) are connected to the coupler box at J1 and J2 via P1 and P2.

switch are mounted in a separate box or a compartment in an antenna tuner. Parts for this project are available from the suppliers listed in **Table 3**.

The primary windings of Tl and T2 are constructed much as Grebenkemper described, but use RG-8 with its jacket removed so that the core and secondary winding may fit over the cable. The braid is wrapped with fiberglass tape to insulate it from the secondary winding. An excellent alternative to fiberglass tape—with even higher RF voltage-breakdown characteristics—is ordinary plumber's Teflon pipe tape, available at most hardware stores.

The transformer secondaries are wound on T-68-2 powdered-iron toroid cores. They are 40 turns of #26 to #30 enameled wire spread evenly around each core. By using #26 to #30 wire on the cores, the cores slip over the tape-wrapped RG-8 lines. With #26 wire on the toroids, a single layer of tape (slightly more with Teflon tape) over the braid provides an extremely snug fit for the core. Use care when fitting the cores onto the RG-8 assemblies.

After the toroids are mounted on the RG-8 sections,

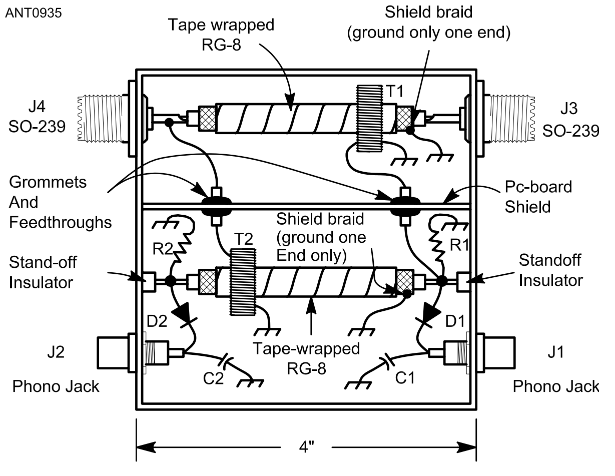
ANT0935
Tape wrapped
RG-8
Shield braid
(ground only one end)
J4
SO-239
T1
J3
SO-239
Grommets
And
Feedthroughs
Pc-board
Shield
Shield braid
(ground one
End only)
R2
T2
R1
Stand-off
Insulator
Standoff
Insulator
D2
D1
J2
Phono Jack
Tape-wrapped
RG-8
C2
C1
J1
Phono Jack
4"

Fig 20—Directional-coupler construction details. Grommets or feedthrough insulators can be used to route the secondary winding of T1 and T2 through the PC board shield. A 3 $^1/_2$ × 3 $^1/_2$ × 4-inch box serves as the enclosure.

Table 3
Parts Sources
(Also see Chapter 21)

Components	*Source*
TLC-series and CA3146 ICs	Newark Electronics 4801 N Ravenswood St Chicago, IL 60640 773-784-5100
LM334, LM336, 1% resistors, trimmer potentiometers	Digi-Key Corporation 701 Brooks Ave S PO Box 677 Thief River Falls, MN 56701 800-344-4539
Toroid cores, Fiberglass tape	Amidon Associates 240 & 250 Briggs Ave Costa Mesa, CA 92626 714-850-4660
Meters	Fair Radio Sales PO Box 1105 Lima, OH 45804 419-227-6573
Toroid cores	Palomar Engineers PO Box 462222 Escondido, CA 92046 760-747-3343
0-150/1500-W-scale meters, A&M model no. 255-138, 1N5711 diodes	Surplus Sales of Nebraska 1502 Jones St Omaha, NE 68102 402-346-4750

coat the assembly with General Cement Corp Polystyrene Q Dope, or use a spot or two of RTV sealant to hold the windings in place and fix the transformers on the RG-8 primary windings.

Mount a PC-board shield in the center of the box, between T1 and T2, to minimize coupling between the transformers. Suspend T1 between the SO-239 connectors and T2 between two standoff insulators. The detector circuits (C1, C2, D1, D2, R1 and R2) are mounted inside the coupler box as shown.

Calibration, Tune Up and Operation

The coupler has excellent directivity. Calibrate the meters for various power levels with an RF ammeter and a 50-Ω dummy load. Calculate I^2R for each power level, and mark the meter faces accordingly. Use R3, R4 and R5 to adjust the meter readings within the ranges. Diode nonlinearities are thus taken into account, and Grebenkemper's signal-processing circuits are not needed for relatively accurate power readings. Start the tune-up process using about 10 W, adjust the antenna tuner for minimum reflected power, and increase power while adjusting the tuner to minimize reflected power.

This circuit has been built into several antenna tuners with good success. The instrument works well at 1.5-kW output on 1.8 MHz. It also works well from 3.5 to 30 MHz with 1.2 and 1.5-kW output.

The antenna is easily tuned for a 1:1 SWR using the null indication provided. Amplifier settings for a matched antenna, as indicated with the wattmeter, closely agreed with those for a 50-Ω dummy load. Checks with a Palomar noise bridge and a Heath Antenna Scope also verified these findings. This circuit should handle more than 1.5 kW, as long as the SWR on the feed line through the wattmeter is kept at or near 1:1. (On one occasion high power was applied while the antenna tuner was not coupled to a load. Naturally the SWR was extremely high, and the output transformer secondary winding opened like a fuse. This resulted from the excessively high voltage across the secondary. The damage was easily and quickly repaired.)

An Inexpensive VHF Directional Coupler

Precision in-line metering devices capable of reading forward and reflected power over a wide range of frequencies are very useful in amateur VHF and UHF work, but their rather high cost puts them out of the reach of many VHF enthusiasts. The device shown in Figs 14 through 16 is an inexpensive adaptation of their basic principles. It can be made for the cost of a meter, a few small parts, and bits of copper pipe and fittings that can be found in the plumbing stocks at many hardware stores.

Construction

The sampler consists of a short section of handmade coaxial line, in this instance, of 50 Ω impedance, with a reversible probe coupled to it. A small pickup loop built into the probe is terminated with a resistor at one end and a diode at the other. The resistor matches the impedance of the loop, not the impedance of the line section. Energy picked up by the loop is rectified by the diode, and the resultant current is fed to a meter equipped with a calibration control.

The principal metal parts of the device are a brass plumbing T, a pipe cap, short pieces of $^3/_4$-inch ID and $^5/_{16}$-inch OD copper pipe, and two coaxial fittings. Other available tubing combinations for 50-Ω line may be usable. The ratio of outer conductor ID to inner conductor OD should be 2.4/1. For a sampler to be used with other

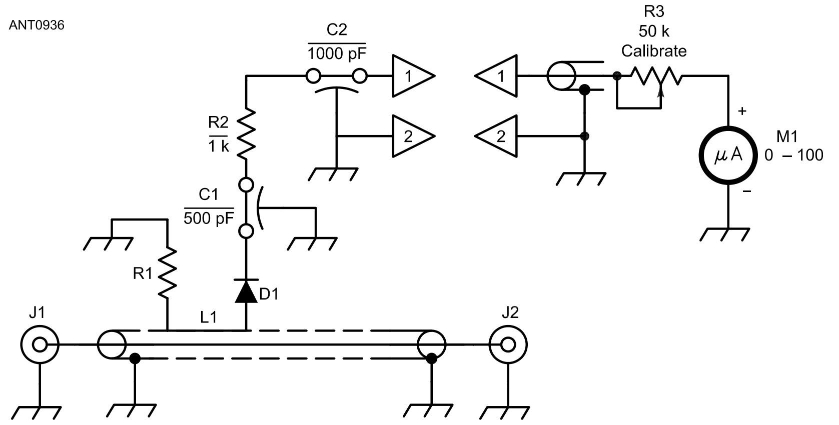

Fig 21—Circuit diagram for the line sampler.
C1—500-pF feedthrough capacitor, solder-in type.
C2—1000-pF feedthrough capacitor, threaded type.
D1—Germanium diode 1N34, 1N60, 1N270, 1N295, or similar.
J1, J2—Coaxial connector, type N (UG-58A).
L1—Pickup loop, copper strap 1-inch long × $^{3}/_{16}$-inch wide. Bend into "C" shape with flat portion $^{5}/_{8}$-inch long.
M1—0-100 µA meter.
R1—Composition resistor, 82 to 100 Ω. See text.
R3—50-kΩ composition control, linear taper.

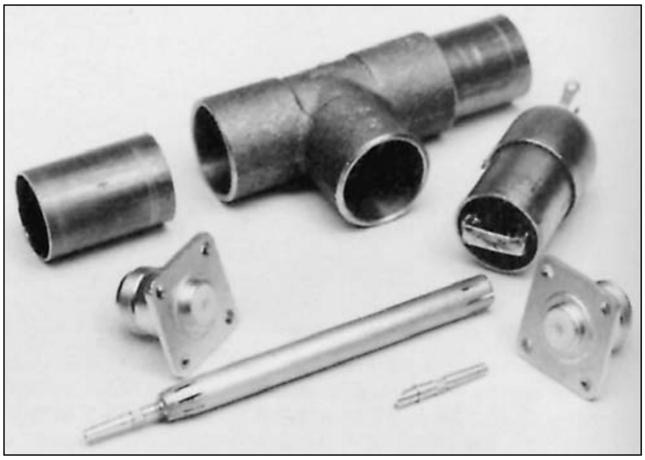

Fig 22—Major components of the line sampler. The brass T and two end sections are at the upper left in this picture. A completed probe assembly is at the right. The N connectors have their center pins removed. The pins are shown with one inserted in the left end of the inner conductor and the other lying in the right foreground.

impedances of transmission line, see Chapter 24 for suitable ratios of conductor sizes. The photographs and **Fig 21** show construction details.

Soldering of the large parts can be done with a 300-watt iron or a small torch. A neat job can be done if the inside of the T and the outside of the pipe are tinned before assembling. When the pieces are reheated and pushed together, a good mechanical and electrical bond will result. If a torch is used, go easy with the heat, as an overheated and discolored fitting will not accept solder well.

Coaxial connectors with Teflon or other heat-resistant insulation are recommended. Type N, with split-ring retainers for the center conductors, are preferred. Pry the split-ring washers out with a knife point or small screwdriver. Don't lose them, as they'll be needed in the final assembly.

The inner conductor is prepared by making eight radial cuts in one end, using a coping saw with a fine-toothed blade, to a depth of $^{1}/_{2}$ inch. The fingers so made are then bent together, forming a tapered end, as shown in **Figs 22** and **23**. Solder the center pin of a coaxial fitting into this, again being careful not to overheat the work.

In preparation for soldering the body of the coax connector to the copper pipe, it is convenient to use a similar fitting clamped into a vise as a holding fixture. Rest the T assembly on top, held in place by its own weight. Use the partially prepared center conductor to assure that the coax connector is concentric with the outer conductor. After being sure that the ends of the pipe are cut exactly perpendicular to the axis, apply heat to the coax fitting, using just enough so a smooth fillet of solder can be formed where the flange and pipe meet.

Before completing the center conductor, check its length. It should clear the inner surface of the connector by the thickness of the split ring on the center pin. File to length; if necessary, slot as with the other end, and solder the center pin in place. The fitting can now be soldered onto the pipe, to complete the 50-Ω line section.

The probe assembly is made from a $1^{1}/_{2}$ inch length of the copper pipe, with a pipe cap on the top to support the upper feedthrough capacitor, C2. The coupling loop is mounted by means of small Teflon standoffs on a copper disc, cut to fit inside the pipe. The disc has four small tabs around the edge for soldering inside the pipe. The diode, D1, is connected between one end of the loop and

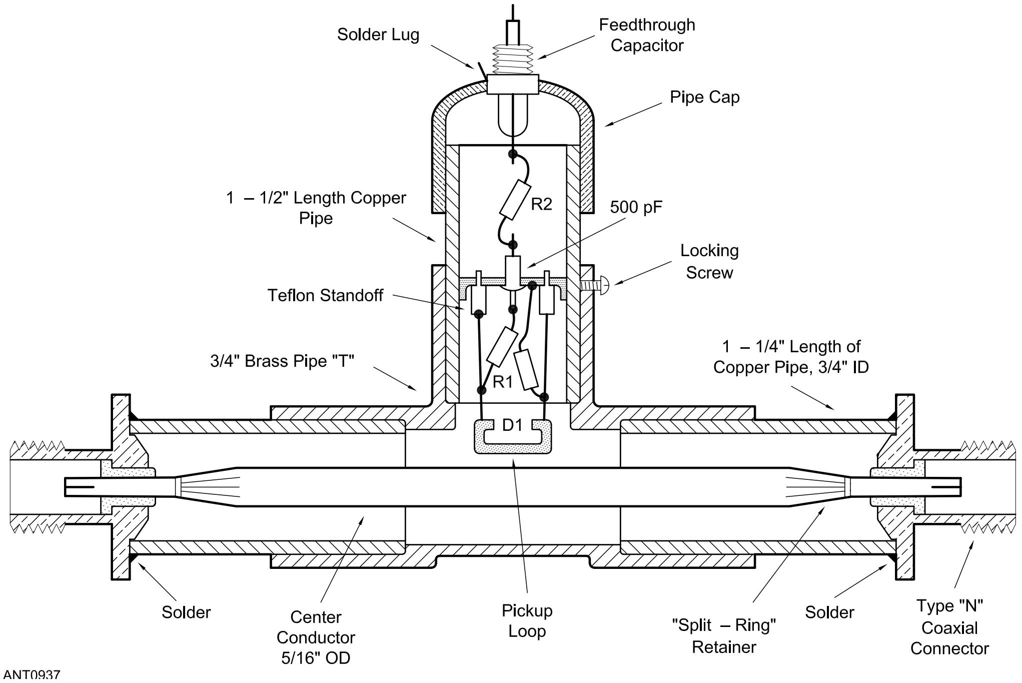
Solder Lug
Feedthrough Capacitor
Pipe Cap
1 – 1/2" Length Copper Pipe
R2
500 pF
Locking Screw
Teflon Standoff
3/4" Brass Pipe "T"
1 – 1/4" Length of Copper Pipe, 3/4" ID
R1
D1
Solder
Center Conductor 5/16" OD
Pickup Loop
"Split – Ring" Retainer
Solder
Type "N" Coaxial Connector
ANT0937

Fig 23—Cross-section view of the line sampler. The pickup loop is supported by two Teflon standoff insulators. The probe body is secured in place with one or more locking screws through holes in the brass T.

a 500-pF feedthrough capacitor, C1, soldered into the disc. The terminating resistor, R1, is connected between the other end of the loop and ground, as directly as possible.

When the disc assembly is completed, insert it into the pipe, apply heat to the outside, and solder the tabs in place by melting solder into the assembly at the tabs. The position of the loop with respect to the end of the pipe will determine the sensitivity of a given probe. For power levels up to 200 watts the loop should extend beyond the face of the pipe about $^5/_{32}$ inch. For use at higher power levels the loop should protrude only $^3/_{32}$ inch. For operation with very low power levels the best probe position can be determined by experiment.

The decoupling resistor, R2, and feedthrough capacitor, C2, can be connected, and the pipe cap put in place. The threaded portion of the capacitor extends through the cap. Put a solder lug over it before tightening its nut in place. Fasten the cap with two small screws that go into threaded holes in the pipe.

Calibration

The sampler is very useful for many jobs even if it is not accurately calibrated, although it is desirable to calibrate it against a wattmeter of known accuracy. A good 50-Ω VHF dummy load is required.

The first step is to adjust the inductance of the loop, or the value of the terminating resistor, for lowest reflected power reading. The loop is the easier to change. Filing it to reduce its width will increase its impedance. Increasing the cross-section of the loop will lower the impedance, and this can be done by coating it with solder. When the reflected power reading is reduced as far as possible, reverse the probe and calibrate for forward power by increasing the transmitter power output in steps and making a graph of the meter readings obtained. Use the calibration control, R3, to set the maximum reading.

Variations

Rather than to use one sampler for monitoring both forward and reflected power by repeatedly reversing the

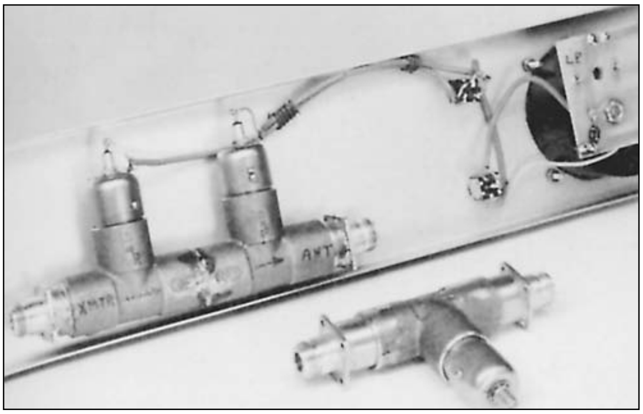

Fig 24—Two versions of the line sampler. The single unit described in detail here is in the foreground. Two sections in a single assembly provide for monitoring forward and reflected power without probe reversal.

probe, it is better to make two assemblies by mounting two T fittings end-to-end, using one for forward and one for reflected power. The meter can be switched between the probes, or two meters can be used.

The sampler described was calibrated at 146 MHz, as it was intended for repeater use. On higher bands the meter reading will be higher for a given power level, and it will be lower for lower frequency bands. Calibration for two or three adjacent bands can be achieved by making the probe depth adjustable, with stops or marks to aid in resetting for a given band. Of course more probes can be made, with each probe calibrated for a given band, as is done in some of the commercially available units.

Other sizes of pipe and fittings can be used by making use of information given in Chapter 24 to select conductor sizes required for the desired impedances. (Since it is occasionally possible to pick up good bargains in 75-Ω line, a sampler for this impedance might be desirable.)

Type-N fittings were used because of their constant impedance and their ease of assembly. Most have the split-ring retainer, which is simple to use in this application. Some have a crimping method, as do apparently all BNC connectors. If a fitting must be used and cannot be taken apart, drill a hole large enough to clear a soldering-iron tip in the copper-pipe outer conductor. A hole of up to $^3/_8$-inch diameter will have very little effect on the operation of the sampler.

A Calorimeter For VHF And UHF Power Measurements

A quart of water in a Styrofoam ice bucket, a roll of small coaxial cable and a thermometer are all the necessary ingredients for an accurate RF wattmeter. Its calibration is independent of frequency. The wattmeter works on the calorimeter principle: A given amount of RF energy is equivalent to an amount of heat, which can be determined by measuring the temperature rise of a known quantity of thermally insulated material. This principle is used in many of the more accurate high-power wattmeters. This procedure was developed by James Bowen, WA4ZRP, and was first described in December 1975 *QST*.

The roll of coaxial cable serves as a dummy load to convert the RF power into heat. RG-174 cable was chosen for use as the dummy load in this calorimeter because of its high loss factor, small size, and low cost. It is a standard 50-Ω cable of approximately 0.11 inch diameter. A prepackaged roll marked as 60 feet long, but measured to be 68 feet, was purchased at a local electronics store. A plot of measured RG-174 loss factor as a function of frequency is shown in **Fig 25**.

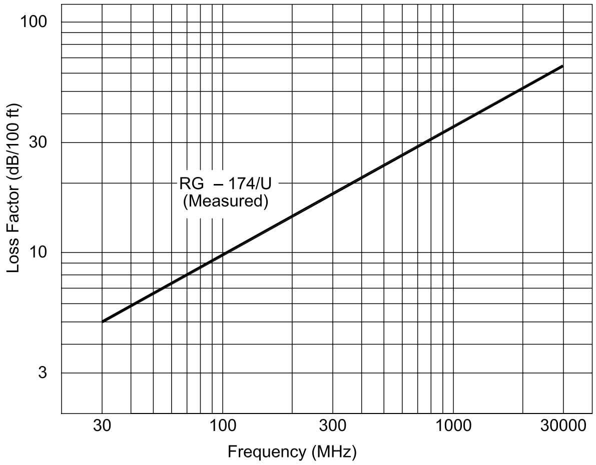

Fig 25—Loss factor of RG-174 coax used in the calorimeter.

In use, the end of the cable not connected to the transmitter is left open-circuited. Thus, at 50 MHz, the reflected wave returning to the transmitter (after making a round trip of 136 feet through the cable) is 6.7 dB × 1.36 = 9.11 dB below the forward wave. A reflected wave 9.11 dB down represents an SWR to the transmitter of 2.08:1. While this value seems larger than would be desired, keep in mind that most 50-MHz transmitters can be tuned to match into an SWR of this magnitude efficiently. To assure accurate results, merely tune the transmitter for maximum power into the load before making

Table 4

Calculated Input SWR for 68 Feet of Unterminated RG-174 Cable

Freq. (MHz)	*SWR*
50	2.08
144	1.35
220	1.20
432	1.06
1296	1.003
2304	1.0003

the measurement. At higher frequencies the cable loss increases so the SWR goes down. **Table 4** presents the calculated input SWR values at several frequencies for 68 feet of RG-174. At 1000 MHz and above, the SWR caused by the cable connector will undoubtedly exceed the very low cable SWR listed for these frequencies.

In operation, the cable is submerged in a quart of water and dissipated heat energy flows from the cable into the water, raising the water temperature. See **Fig 26**. The calibration of the wattmeter is based on the physical fact that one calorie of heat energy will raise one gram of liquid water 1° Celsius. Since one quart of water contains 946.3 grams, the transmitter must deliver 946.3 calories of heat energy to the water to raise its temperature 1° C. One calorie of energy is equivalent to 4.186 joules and a joule is equal to 1 W for 1 second. Thus, the heat capacitance of 1 quart of water expressed in joules is 946.3 × 4.186 = 3961 joules/° C.

The heat capacitance of the cable is small with respect to that of the water, but nevertheless its effect should be included for best accuracy. The heat capacitance of the cable was determined in the manner described below. The 68-foot roll of RG-174 cable was raised to a uniform temperature of 100° C by immersing it in a pan of boiling water for several minutes. A quart of tap water was poured into the Styrofoam ice bucket and its temperature was measured at 28.7° C. the cable was then transferred quickly from the boiling water to the water in the ice bucket. After the water temperature in the ice bucket had ceased to rise, it measured 33.0° C. Since the total heat gained by the quart of water was equal to the total heat lost from the cable, we can write the following equation:

$$(\Delta T_{WATER})(C_{WATER}) = -(\Delta T_{CABLE})(C_{CABLE})$$

where

ΔT_{WATER} = the change in water temperature
C_{WATER} = the water heat capacitance
Δ_{CABLE} = the change in cable temperature
C_{CABLE} = the cable heat capacitance

Substituting and solving:

$$(33.0 - 28.7)(3961) = -(33.0 - 100)(C_{CABLE})$$

Thus, the total heat capacitance of the water and cable in the calorimeter is 3961 + 254 = 4215 joules/º C. Since 1º F = 5/9º C, the total heat capacitance can also be expressed as 4215 × 5/9 = 2342 joules/º F.

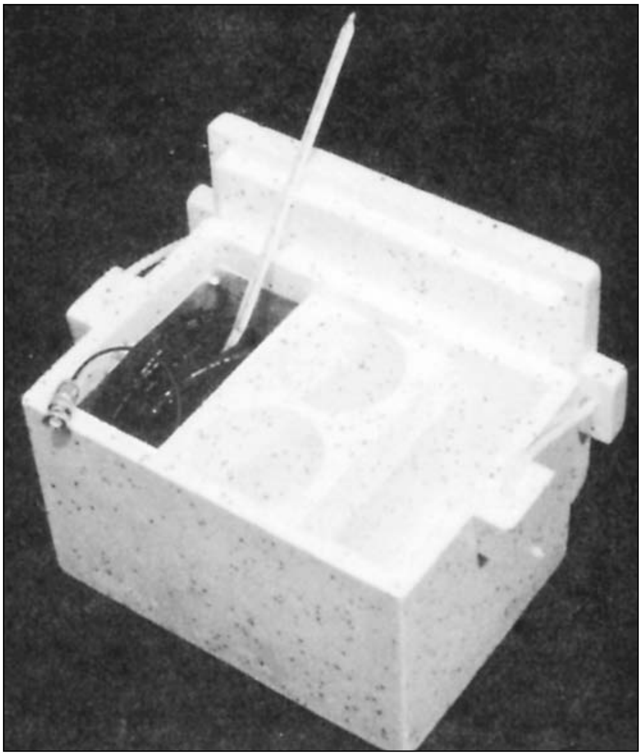

Fig 26—The calorimeter ready for use. The roll of coaxial cable is immersed in one quart of water in the left-hand compartment of the Styrofoam container. Also shown is the thermometer, which doubles as a stirring rod.

Materials and Construction

The quart of water and cable must be thermally insulated to assure that no heat is gained from or lost to the surroundings. A Styrofoam container is ideal for this purpose since Styrofoam has a very low thermal conductivity and a very low thermal capacitance. A local variety store was the source of a small Styrofoam cold chest with compartments for carrying sandwiches and drink cans. The rectangular compartment for sandwiches was found to be just the right size for holding the quart of water and coax.

The thermometer can be either a Celsius or Fahrenheit type, but try to choose one that has divisions for each degree spaced wide enough so that the temperature can be estimated readily to one-tenth degree. Photographic supply stores carry darkroom thermometers, which are ideal for this purpose. In general, glass bulb thermometers are more accurate than mechanical dial-pointer types.

The RF connector on the end of the cable should be a constant-impedance type. A BNC type connector especially designed for use on 0.11-inch diameter cable was located through surplus channels. If you cannot locate one of these, wrap plastic electrical tape around the cable near its end until the diameter of the tape wrap is the same as that of RG-58. Then connect a standard BNC connector for RG-58 in the normal fashion. Carefully seal the opposite open end of the cable with plastic tape or silicone caulking compound so no water can leak into the cable at this point.

Procedure for Use

Pour 1 quart of water (4 measuring cups) into the Styrofoam container. As long as the water temperature is not very hot or very cold, it is unnecessary to cover the top of the Styrofoam container during measurements. Since the transmitter will eventually heat the water several degrees, water initially a few degrees cooler than air temperature is ideal because the average water temperature will very nearly equal the air temperature and heat transfer to the air will be minimized.

Connect the RG-174 dummy load to the transmitter through the shortest possible length of lower loss cable such as RG-8. Tape the connectors and adapter at the RG-8 to RG-174 joint carefully with plastic tape to prevent water from leaking into the connectors and cable at this point. Roll the RG-174 into a loose coil and submerge it in the water. Do not bind the turns of the coil together in any way, as the water must be able to freely circulate among the coaxial cable turns. All the RG-174 cable must be submerged in the water to ensure sufficient cooling. Also submerge part of the taped connector attached to the RG-174 as an added precaution.

Upon completing the above steps, quickly tune up the transmitter for maximum power output into the load. Cease transmitting and stir the water slowly for a minute or so until its temperature has stabilized. Then measure the water temperature as precisely as possible. After the initial temperature has been determined, begin the test transmission, measuring the total number of seconds of key-down time accurately. Stir the water slowly with the thermometer and continue transmitting until there is a significant rise in the water temperature, say 5° to 10°. The test may be broken up into a series of short periods, as long as you keep track of the total key-down time. When the test is completed, continue to stir the water slowly and monitor its temperature. When the temperature ceases to rise, note the final indication as precisely as possible.

To compute the transmitter power output, multiply the calorimeter heat capacitance (4215 for C or 2342 for F) by the difference in initial and final water temperature. Then divide by the total number of seconds of key-down time. The resultant is the transmitter power in watts. A nomogram that can also be used to find transmitter power output is given in **Fig 27**. With a straight line, connect the total number of key-down seconds in the time column to the number of degrees change (F or C) in the temperature rise column, and read off the transmitter power output at the point where the straight line crosses the power-output column.

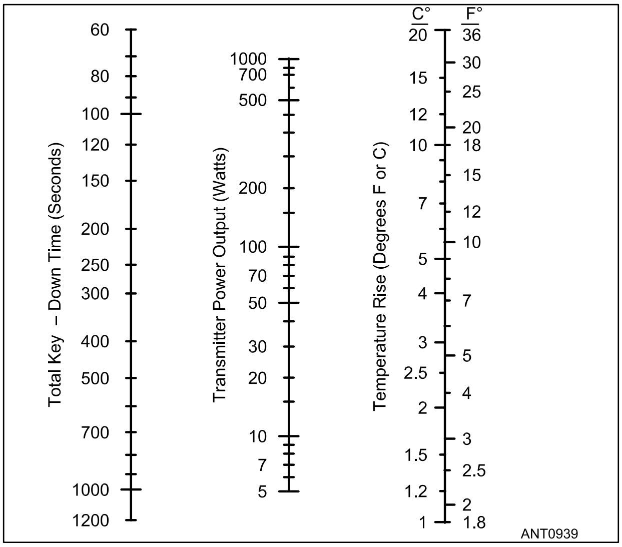

Fig 27—Nomogram for finding transmitter power output for the calorimeter.

Power Limitation

The maximum power handling capability of the calorimeter is limited by the following. At very high powers the dielectric material in the coaxial line will melt because of excessive heating or the cable will arc over from excessive voltage. As the transmitter frequency gets higher, the excessive-heating problem is accentuated, as more of the power is dissipated in the first several feet of cable. For instance, at 1296 MHz, approximately 10% of the transmitting power is dissipated in the first foot of cable. Overheating can be prevented when working with high power by using a low duty cycle to reduce the average dissipated power. Use a series of short transmissions, such as two seconds on, ten seconds off. Keep count of the total key-down time for power calculation purposes. If the cable arcs over, use a larger-diameter cable, such as RG-58, in place of the RG-174. The cable should be long enough to assure that the reflected wave will be down 10 dB or more at the input. It may be necessary to use more than one quart of water in order to submerge all the cable conveniently. If so, be sure to calculate the new value of heat capacitance for the larger quantity of water. Also you should measure the new coaxial cable heat capacitance using the method previously described.

A Noise Bridge For 1.8 Through 30 MHz

The noise bridge, sometimes referred to as an antenna (RX) noise bridge, is an instrument for measuring the impedance of an antenna or other electrical circuits. The unit shown here in **Fig 28**, designed for use in the 1.8 through 30-MHz range, provides adequate accuracy for most measurements. Battery operation and small physical size make this unit ideal for remote-location use. Tone modulation is applied to the wide-band noise generator as an aid for obtaining a null indication. A detector, such as the station receiver, is required for operation.

The noise bridge consists of two parts—the noise generator and the bridge circuitry. See **Fig 29**. A 6.8-V

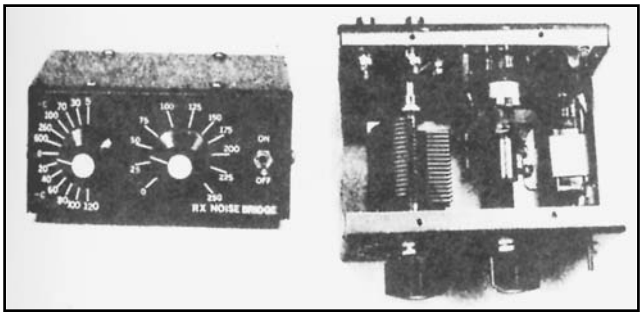
RX NOISE BRIDGE
ON
OFF

Fig 28—Exterior and interior views of the noise bridge. The unit is finished in red enamel. Press-on lettering is used for the calibration marks. Note that the potentiometer must be isolated from ground.

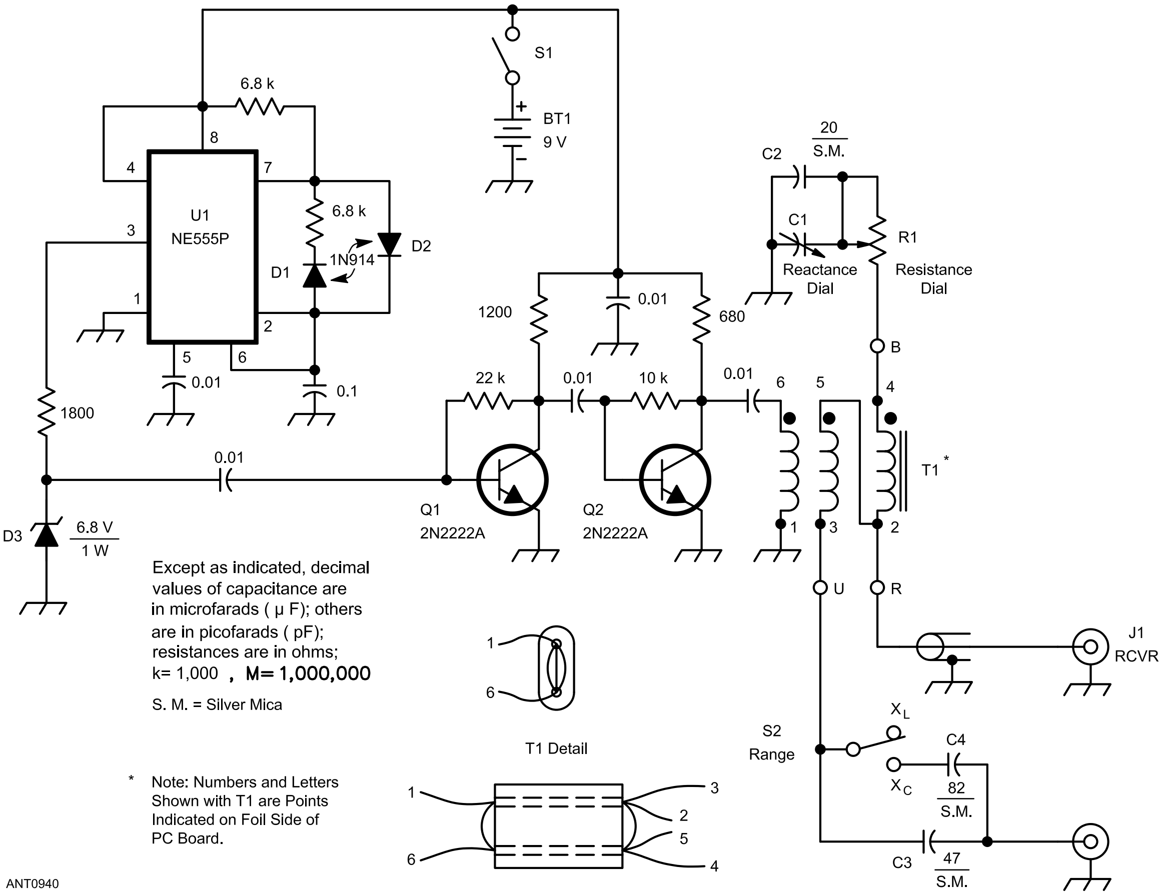
S1
BT1
9 V
6.8 k
8
4
7
3
1
2
5
6
U1
NE555P
6.8 k
D1
1N914
D2
0.01
0.1
1800
0.01
D3
6.8 V
1 W
1200
0.01
680
22 k
0.01
10 k
0.01
Q1
2N2222A
Q2
2N2222A
C2
20
S.M.
C1
R1
Reactance Dial
Resistance Dial
B
6
5
4
T1*
1
3
2
U
R
J1
RCVR
S2
Range
X_L
X_C
C4
82
S.M.
C3
47
S.M.
1
6
T1 Detail
1
6
3
2
5
4
Except as indicated, decimal values of capacitance are in microfarads (μ F); others are in picofarads (pF); resistances are in ohms; k= 1,000 , M= 1,000,000
S. M. = Silver Mica
* Note: Numbers and Letters Shown with T1 are Points Indicated on Foil Side of PC Board.
ANT0940

Fig 29—Schematic diagram of the noise bridge. Use 1/4 W composition resistors. Capacitors are miniature ceramic units unless indicated otherwise. Component designations indicated in the schematic but not called out in the parts list are for text and parts-placement reference only.

BT1—9-V battery, NEDA 1604A or equiv.
C1—15- to 150-pF variable
C2—20-pF mica.
C3—47-pF mica.
C4—82-pF mica.
J1, J2—Coaxial connector.
R1—Linear, 250 Ω, AB type. Use a good grade of resistor.
S1, S2—Toggle, SPST.
T1—Transformer; 3 windings on an Amidon BLN-43-2402 ferrite binocular core. Each winding is three turns of #30 enameled wire. One turn is equal to the wire passing once through both holes in the core. The primary winding starts on one side of the transformer, and the secondary and tertiary windings start on the opposite side.
U1—Timer, NE555 or equiv.

Zener diode serves as the noise source. U1 generates an approximate 50% duty cycle, 1000-Hz square wave signal which is applied to the cathode of the Zener diode. The 1000-Hz modulation appears on the noise signal and provides a useful null detection enhancement effect. The broadband-noise signal is amplified by Q1, Q2 and associated components to a level that produces an approximate S9 signal in the receiver. Slightly more noise is available at the lower end of the frequency range, as no frequency compensation is applied to the amplifier. Roughly 20 mA of current is drawn from the 9-V battery, thus ensuring long battery life—providing the power is switched off after use!

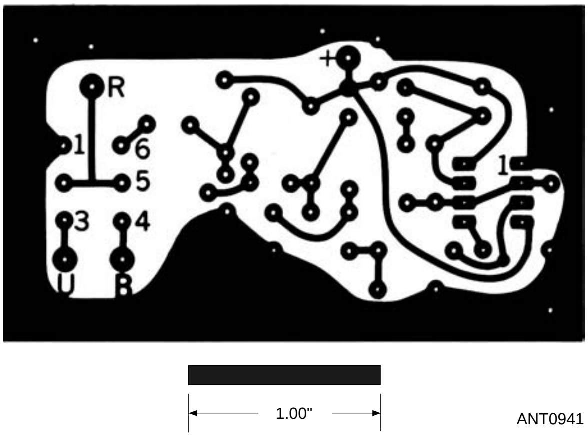

Fig 30—Etching pattern for the noise bridge PC board, at actual size. Black represents copper. This is the pattern for the bottom side of the board. The top side of the board is a complete ground plane with a small amount of copper removed from around the component holes.

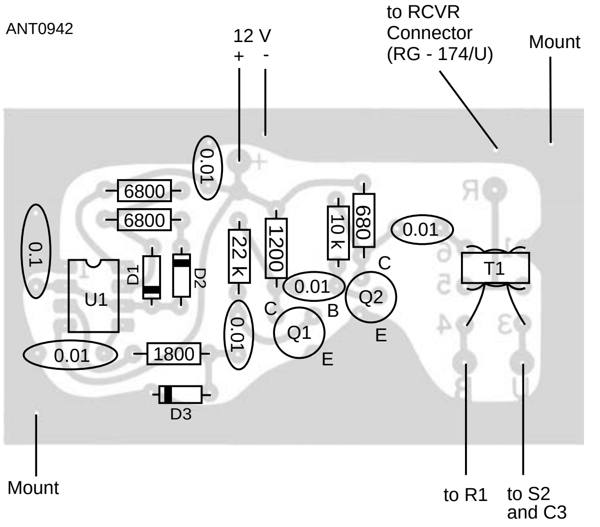

Fig 31—Parts-placement guide for the noise bridge as viewed from the component or top side of the board. Mounting holes are located in two corners of the board, as shown.

The bridge portion of the circuit consists of T1, C1, C2 and R1. T1 is a trifilar wound transformer with one of the windings used to couple noise energy into the bridge circuit. The remaining two windings are arranged so that each one is in an arm of the bridge. C1 and R1 complete one arm and the UNKNOWN circuit, along with C2, comprise the remainder of the bridge. The terminal labeled RCVR is for connection to the detector.

The reactance range of a noise bridge is dependent on several factors, including operating frequency, value of the series capacitor (C3 or C3 plus C4 in Fig 29) and the range of the variable capacitor (C1 in Fig 29). The RANGE switch selects reactance measurements weighted toward either capacitance or inductance by placing C4 in parallel with C3. The zero-reactance point occurs when C1 is either nearly fully meshed or fully unmeshed. The RANGE switch nearly doubles the resolution of the reactance readings.

CONSTRUCTION

The noise bridge is contained in a homemade aluminum enclosure that measures $5 \times 2\frac{3}{8} \times 3\frac{3}{4}$ inches. Many of the circuit components are mounted on a circuit board that is fastened to the rear wall of the cabinet. The circuit-board layout is such that the lead lengths to the board from the bridge and coaxial connectors are at a minimum. An etching pattern and a parts-placement guide for the circuit board are shown in **Figs 30** and **31**.

Care must be taken when mounting the potentiometer, R1. For accurate readings the potentiometer must be well insulated from ground. In the unit shown this was accomplished by mounting the control on a piece of plexiglass, which in turn was fastened to the chassis with a piece of aluminum angle stock.

Additionally, a $\frac{1}{4}$-inch control-shaft coupling and a length of phenolic rod were used to further isolate the control from ground where the shaft passes through the front panel. A high-quality potentiometer is required if good measurement results are to be obtained.

There is no such problem when mounting the variable capacitor because the rotor is grounded. Use a high-quality capacitor; do not try to save money on that component. Two RF connectors on the rear panel are connected to a detector (receiver) and to the UNKNOWN circuit. Do not use plastic-insulated phono connectors (they might influence bridge accuracy at higher frequencies). Use miniature coaxial cable (RG-174) between the RCVR connector and circuit board. Attach one end of C3 to the circuit board and the other directly to the UNKNOWN circuit connector.

Bridge Compensation

Stray capacitance and inductance in the bridge circuit can affect impedance readings. If a very accurate bridge is required, use the following steps to counter the effects of stray reactance. Because the physical location

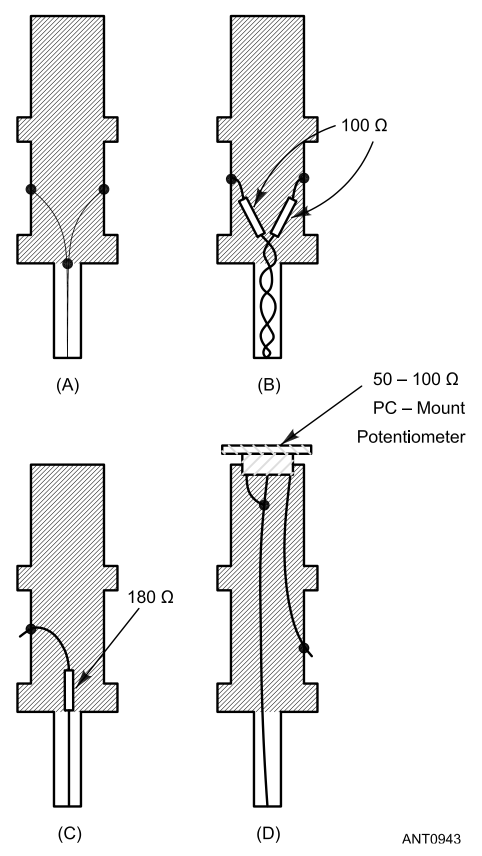

Fig 32—Construction details of the resistive loads used to check and calibrate the noise bridge. Each of the loads is constructed inside a coaxial connector that matches those on the bridge. (Views shown are cross-sections of PL-259 bodies; the sleeves are not shown.) Leads should be kept as short as possible to minimize parasitic inductance. A is a 0-Ω load; B depicts a 50-Ω load; C is a 180-Ω load; D shows a variable-resistance load used to determine the loss in a coaxial cable.

of the board, connectors and controls in the cabinet determine where compensation is needed, there is no provision for the compensation components on the printed circuit board.

Good calibration loads are necessary to check the accuracy of the noise bridge. Four are needed here: a 0-Ω (short-circuit) load, a 50-Ω load, a 180-Ω load, and a variable-resistance load. The short-circuit and fixed-resistance loads are used to check the accuracy of the noise bridge; the variable-resistance load is used when measuring coaxial-cable loss.

Construction details of the loads are shown in **Fig 32**. Each load is constructed inside a connector. When building the loads, keep leads as short as possible to minimize parasitic effects. The resistors must be noninductive (*not* wirewound).

Quarter-watt, carbon-composition resistors should work fine. The potentiometer in the variable-resistance load is a miniature PC-mount unit with a maximum resistance of 100 Ω or less. The potentiometer wiper and one of the end leads are connected to the center pin of the connector; the other lead is connected to ground.

Stray Capacitance

Stray capacitance on the variable-resistor side of the bridge tends to be higher than that on the unknown side. This is so because the parasitic capacitance in the variable resistor, R1, is comparatively high.

The effect of parasitic capacitance is most easily detected using the 180-Ω load. Measure and record the actual resistance of the load, R_L. Connect the load to the UNKNOWN connector, place S2 in the X_L position, tune the receiver to 1.8 MHz, and null the bridge. (See the section, "Finding the Null" for tips.) Use an ohmmeter across R1 to measure its dc resistance. The magnitude of the stray capacitance can be calculated by

$$C_P = C3\left(\sqrt{\frac{R1}{R2}}-1\right) \tag{Eq 4}$$

where

R_L = load resistance (as measured)
R1 = resistance of the variable resistor
C3 = series capacitance.

You can compensate for C_P by placing a variable capacitor, C_C, in the side of the bridge with lesser stray capacitance. If R1 is greater than R_L, stray capacitance is greater on the variable resistor side of the bridge: Place C_C between point U (on the circuit board) and ground. If R1 is less than R_L, stray capacitance is greater on the unknown side: Place C_C between point B and ground. If the required compensating capacitance is only a few picofarads, you can use a gimmick capacitor (made by twisting two short pieces of insulated, solid wire together) for C_C. A gimmick capacitor is adjusted by trimming its length.

Stray Inductance

Parasitic inductance, if present, should be only a few tens of nanohenries. This represents a few ohms of inductive reactance at 30 MHz. The effect is best observed by reading the reactance of the 0-Ω test load at 1.8 and 30 MHz; the indicated reactance should be the same at both frequencies.

If the reactance reading decreases as frequency is increased, parasitic inductance is greater in the known arm, and compensating inductance is needed between point U and C3. If the reactance increases with frequency, the unknown-arm inductance is greater, and compensating inductance should be placed between point B and R1.

Compensate for stray inductance by placing a single-turn coil, made from a 1 to 2-inch length of solid wire, in the appropriate arm of the bridge. Adjust the size of this coil until the reactance reading remains constant from 1.8 to 30 MHz.

Calibration

Good calibration accuracy is necessary for accurate noise-bridge measurements. Calibration of the resistance scale is straightforward. To do this, tune the receiver to a frequency near 10 MHz. Attach the 0-Ω load to the UNKNOWN connector and null the bridge. This is the zero-resistance point; mark it on the front-panel resistance scale. The rest of the resistance range is calibrated by adjusting R1, measuring R1 with an accurate ohmmeter, calculating the increase from the zero point and marking the increase on the front panel.

Most bridges have the reactance scale marked in capacitance because capacitance does not vary with frequency. Unfortunately, that requires calibration curves or non-trivial calculations to arrive at the load reactance. An alternative method is to mark the reactance scale in *ohms* at a reference frequency of 10 MHz. This method calibrates the bridge near the center of its range and displays reactance directly, but it requires a simple calculation to scale the reactance reading for frequencies other than 10 MHz. The scaling equation is:

$$X_{u(f)} = X_{u(10)} \frac{10}{f} \qquad \text{(Eq 5)}$$

where

f = frequency in MHz

$X_{u(10)}$ = reactance of the unknown load at 10 MHz

$X_{u(f)}$ = reactance of the unknown load at f.

A shorted piece of coaxial cable serves as a reactance source. (The reactance of a shorted, low-loss coaxial cable is dependent only on the cable length, the measurement frequency and the cable characteristic impedance.) Radio Shack RG-8M is used here because it is readily available, has relatively low loss and has an almost purely resistive characteristic impedance.

Prepare the calibration cable as follows:

1) Cut a length of coaxial cable that is slightly longer than $^{1}/_{4}$ λ at 10 MHz (about 20 feet for RG-8M). Attach a suitable connector to one end of the cable; leave the other end open-circuited.
2) Connect the 0-Ω load to the noise bridge UNKNOWN connector and set the receiver frequency to 10 MHz. Adjust the noise bridge for a null. Do not adjust the reactance control after the null is found.
3) Connect the calibration cable to the bridge UNKNOWN terminal. Null the bridge by adjusting *only* the variable resistor and the receiver frequency. The receiver frequency should be less than 10 MHz; if it is above 10 MHz, the cable is too short, and you need to prepare a longer one.
4) Gradually cut short lengths from the end of the coaxial cable until you obtain a null at 10 MHz by adjusting only the resistance control. Then connect the cable center and shield conductors at the open end with a short length of braid. Verify that the bridge nulls with zero reactance at 20 MHz.
5) The reactance of the coaxial cable (normalized to 10 MHz) can be calculated from:

$$X_{i(10)} = R_0 \frac{f}{10} \tan\left(2\pi \frac{f}{40}\right) \qquad \text{(Eq 6)}$$

where

$X_{i(10)}$ = cable reactance at 10 MHz

R_0 = characteristic resistance of the coaxial cable (52.5 Ω for Radio Shack RG-8M)

f = frequency in MHz

The results of Eq 6 have less than 5% error for reactances less than 500 Ω, so long as the test-cable loss is less than 0.2 dB. This error becomes significantly less at lower reactances (2% error at 300 Ω for a 0.2-dB-loss cable). The loss in 18 feet of RG-8M is 0.13 dB at 10 MHz. Reactance data for Radio Shack RG-8M is given in **Table 5**.

Table 5

Noise Bridge Calibration Data: Coaxial-Cable Method

This data is for Radio Shack RG-8M cable (R_0 =52.5 Ω) cut to exactly $^{1}/_{4}$ λ at 10 MHz; the reactances and capacitances shown correspond to this frequency.

Reactance				***Capacitance***	
X_i	*f(MHz)*	*X_i*	*f(MHz)*	*C(pF)*	*f(MHz)*
10	3.318	–10	19.376	10	9.798
20	4.484	–20	18.722	20	9.612
30	5.262	–30	18.048	30	9.440
40	5.838	–40	17.368	40	9.280
50	6.286	–50	16.701	50	9.130
60	6.647	–60	16.062	60	8.990
70	6.943	–70	15.471	70	8.859
80	7.191	–80	14.936	80	8.735
90	7.404	–90	14.462	90	8.618
100	7.586	–100	14.044	100	8.508
110	7.747	–110	13.682	110	8.403
120	7.884	–120	13.369	120	8.304
130	8.009	–130	13.097	130	8.209
140	8.119	–140	12.861	140	8.119
150	8.217	–150	12.654		
160	8.306	–160	12.473		
170	8.387	–170	12.313		
180	8.460	–180	12.172		
190	8.527	–190	12.045		
200	8.588	–200	11.932	*C(pF)*	*f(MHz)*
210	8.645	–210	11.831	–10	10.219
220	8.697	–220	11.739	–20	10.459
230	8.746	–230	11.655	–30	10.721
240	8.791	–240	11.579	–40	11.010
250	8.832	–250	11.510	–50	11.328
260	8.872	–260	11.446	–60	11.679
270	8.908	–270	11.387	–70	12.064
280	8.942	–280	11.333	–80	12.484
290	8.975	–290	11.283	–90	12.935
300	9.005	–300	11.236	–100	13.407
350	9.133	–350	11.045	–110	13.887
400	9.232	–400	10.905	–120	14.357
450	9.311	–450	10.798	–130	14.801
500	9.375	–500	10.713	–140	15.211

With the prepared cable and calibration values on hand, proceed to calibrate the reactance scale. Tune the receiver to the appropriate frequency for the desired reactance (given in Table 5, or found using Eq 6). Adjust the resistance and reactance controls to null the bridge. Mark the reactance reading on the front panel. Repeat this process until all desired reactance values have been marked. The resistance values needed to null the bridge during this calibration procedure may be significant (more than 100 Ω) at the higher reactances.

This calibration method is much more accurate than using fixed capacitors across the UNKNOWN connector. Also, you can calibrate a noise bridge in less than an hour using this method.

Finding the Null

In use, a receiver is attached to the RCVR connector and some load of unknown value is connected to the UNKNOWN terminal. The receiver allows us to hear the noise present across the bridge arms at the frequency of the receiver passband. The strength of the noise signal depends on the strength of the noise-bridge battery, the receiver bandwidth/sensitivity and the impedance difference between the known and unknown bridge arms. The noise is stronger and the null more obvious with wide receiver passbands. Set the receiver to the widest bandwidth AM mode available.

The noise-bridge output is heard as a 1000-Hz tone. When the impedances of the known and unknown bridge arms are equal, the voltage across the receiver is minimized; this is a null. In use, the null may be difficult to find because it appears only when both bridge controls approach the values needed to balance the bridge.

To find the null, set C1 to mid-scale, sweep R1 slowly through its range and listen for a reduction in noise (it's also helpful to watch the S meter). If no reduction is heard, set R1 to mid-range and sweep C1. If there is still no reduction, begin at one end of the C1 range and sweep R1. Increment C1 about 10% and sweep R1 with each increment until some noise reduction appears. Once noise reduction begins, adjust C1 and R1 alternately for minimum signal.

MEASURING COAXIAL-CABLE PARAMETERS WITH A NOISE BRIDGE

Coaxial cables have a number of properties that affect the transmission of signals through them. Generally, radio amateurs are concerned with cable attenuation and characteristic impedance. If you plan to use a noise bridge or SWR analyzer to make antenna-impedance measurements, however, you need to accurately determine not just cable impedance and attenuation, but also electrical length. Fortunately, all of these parameters are easy to measure with an accurate noise bridge or SWR analyzer.

Cable Electrical Length

With a noise bridge and a general-coverage receiver, you can easily locate frequencies at which the line in question is a multiple of $^1/_2$ λ, because a shorted $^1/_2$ λ line has a 0-Ω impedance (neglecting line loss). By locating two adjacent null frequencies, you can solve for the length of line in terms of $^1/_2$ λ at one of the frequencies and calculate the line length (overall accuracy is limited by bridge accuracy and line loss, which broadens the nulls). As an interim variable, you can express cable length as the frequency at which a cable is 1 λ long. This length will be represented by f λ. Follow these steps to determine f λ for a coaxial cable.

1) Tune the receiver to the frequency range of interest. Attach the short-circuit load to the noise bridge UNKNOWN connector and null the bridge.
2) Disconnect the far end of the coaxial cable from its load (the antenna) and connect it to the 0-Ω test load. Connect the near end of the cable to the bridge UNKNOWN connector.
3) Adjust the receiver frequency and the noise-bridge resistance control for a null. *Do not change the noise bridge reactance-control setting during this procedure.* Note the frequency at which the null is found; call this frequency f_n. The noise-bridge resistance at the null should be relatively small (less than 20 Ω).
4) Tune the receiver upward in frequency until the next null is found. Adjust the resistance control, if necessary, to improve the null, *but do not adjust the reactance control.* Note the frequency at which this second null is found; this is f n+2 .
5) Solve Eq 7 for n and the electrical length of the cable.

$$n = \frac{2f_n}{f_{n+2} - f_n} \qquad \text{(Eq 7)}$$

$$f_\lambda = \frac{4f_n}{n} \qquad \text{(Eq 8)}$$

where

n = cable electrical length in quarter waves, at f_n
f_l = frequency at which the cable is 1λ
ℓ = cable electrical length, in λ

For example, consider a 74-foot length of Columbia 1188 foam-dielectric cable (velocity factor = 0.78) to be used on the 10-meter band. Based on the manufacturer's specification, the cable is 2.796 λ at 29 MHz. Nulls were found at 24.412 (f_n) and 29.353 (f_{n+2}) MHz. Eq 7 yields n = 9.88, which produces 9.883 MHz from Eq 8 and 2.934 l for Eq 9. If the manufacturer's specification is correct, the measured length is off by less than 5%, which is very reasonable. Ideally, n would yield an integer. The difference between n and the closest integer indicates that there is some error.

This procedure also works for lines with an open circuit as the termination (n will be close to an odd number). End effects from the PL-259 increase the effective length of the coaxial cable; however, this decreases the calculated f_λ.

Cable Characteristic Impedance

The characteristic impedance of the coaxial cable is found by measuring its input impedance at two frequencies separated by $^1/_4\, f_\lambda$. This must be done when the cable is terminated in a resistive load.

Characteristic impedance changes slowly as a function of frequency, so this measurement must be done near the frequency of interest. The measurement procedure is as follows.

1) Place the 50-Ω load on the far end of the coaxial cable and connect the near end to the UNKNOWN connector of the noise bridge. (Measurement error is minimized when the load resistance is close to the characteristic impedance of the cable. This is the reason for using the 50-Ω load.)
2) Tune the receiver approximately $^1/_8$ f below the frequency of interest. Adjust the bridge resistance and reactance controls to obtain a null, and note their readings as R_{f1} and X_{f1}. Remember, the reactance reading must be scaled to the measurement frequency.
3) Increase the receiver frequency by exactly $^1/_4\, f\lambda$. Null the bridge again, and note the readings as R_{f2} and X_{f2}.
4) Calculate the characteristic impedance of the coaxial cable using Eqs 10 through 15. A scientific calculator is helpful for this.

$$\ell = \frac{f_0}{f_\lambda} \qquad \text{(Eq 9)}$$

$$R = R_{f1} \times F_{f2} - X_{f1} \times X_{f2} \qquad \text{(Eq 10)}$$

$$X = R_{f1} \times X_{f2} + X_{f1} \times R_{f2} \qquad \text{(Eq 11)}$$

$$Z = \sqrt{R^2 + X^2} \qquad \text{(Eq 12)}$$

$$R_0 = \sqrt{Z}\cos\left[\frac{1}{2}\tan^{-1}\left(\frac{X}{R}\right)\right] \qquad \text{(Eq 13)}$$

$$X_0 = \sqrt{Z}\sin\left[\frac{1}{2}\tan_{-1}\left(\frac{X}{R}\right)\right] \qquad \text{(Eq 14)}$$

$$Z_0 = R_0 + jX_0 \qquad \text{(Eq 15)}$$

Let's continue with the example used earlier for cable length. The measurements are:

f1 = 29.000 – (9.883 / 8) = 27.765 MHz
R_{f1} = 64 Ω
X_{f1} = –22 Ω × (10 / 27.765) = –7.9 Ω
f2 = 27.765 + (9.883/4) = 30.236 MHz
R_{f2} = 50 Ω
X_{f2} = –24 Ω × (10 / 30.236) = –7.9 Ω

When used in Eqs 10 through 15, these data yield:

R = 3137.59
X = –900.60
Z = 3264.28
R_0 = 56.58 Ω
X_0 = –7.96 Ω

Cable Attenuation

Cable loss can be measured once the cable electrical length and characteristic resistance are known. The measurement must be made at a frequency where the cable presents no reactance. Reactance is zero when the cable electrical length is an integer multiple of λ/4. You can easily meet that condition by making the measurement frequency an integer multiple of $^1/_4$ fλ . Loss at other frequencies can be interpolated with reasonable accuracy. This procedure employs a resistor-substitution method that provides much greater accuracy than is achieved by directly reading the resistance from the noise-bridge scale.

1) Determine the approximate frequency at which you want to make the loss measurement by using

$$n = \frac{4f_0}{f_\lambda} \qquad \text{(Eq 16A)}$$

Round n to the nearest integer, then

$$f1 = \frac{n}{4} f_\lambda \qquad \text{(Eq 16B)}$$

2) If n is odd, leave the far end of the cable open; if n is even, connect the 0-Ω load to the far end of the cable. Attach the near end of the cable to the UNKNOWN connector on the noise bridge.
3) Set the noise bridge to zero reactance and the receiver to f1. Fine tune the receiver frequency and the noise-bridge resistance to find the null.
4) Disconnect the cable from the UNKNOWN terminal, and connect the variable-resistance calibration load in its place. Without changing the resistance setting on the bridge, adjust the load resistor and the bridge reactance to obtain a null.
5) Remove the variable-resistance load from the bridge UNKNOWN terminal and measure the load resistance using an ohmmeter that's accurate at low resistance levels. Refer to this resistance as R_i.
6) Calculate the cable loss in decibels using

$$a\ell = 8.69\,\frac{R_i}{R_0} \qquad \text{(Eq 17)}$$

To continue this example, Eq 16A gives n = 11.74, so measure the attenuation at n = 12. From Eq 16B, f1 = 29.649 MHz. The input resistance of the cable measures 12.1 Ω with 0-Ω load on the far end of the cable; this corresponds to a loss of 1.86 dB.

USING A BRIDGE TO MEASURE THE IMPEDANCE OF AN ANTENNA

The impedance at the end of a transmission line can be easily measured using a noise bridge or SWR analyzer. In many cases, however, you really want to measure the impedance of an antenna—that is, the impedance of the load at the far end of the line. There are several ways to handle this.

1) Measurements can be made with the bridge at the antenna. This is usually not practical because the antenna must be in its final position for the measurement to be accurate. Even if it can be done, making such a measurement is certainly not very convenient.
2) Measurements can be made at the source end of a coaxial cable—if the cable length is an exact integer multiple of $^1/_2$ λ. This effectively restricts measurements to a single frequency.
3) Measurements can be made at the source end of a coaxial cable and corrected using a Smith Chart as shown in Chapter 28. This graphic method can result in reasonable estimates of antenna impedance—as long as the SWR is not too high and the cable is not too lossy. However, it doesn't compensate for the complex impedance characteristics of real-world coaxial cables. Also, compensation for cable loss can be tricky to apply. These problems, too, can lead to significant errors.
4) Last, measurements can be corrected using the transmission-line equation. The *TLW* program included on the CD-ROM in the back of this book, can do these complicated computations for you. This is the best method for calculating antenna impedances from measured parameters, but it requires that you measure the feed-line characteristics beforehand—measurements for which you need access to both ends of the feed line.

The procedure for determining antenna impedance is to first measure the electrical length, characteristic impedance, and attenuation of the coaxial cable connected to the antenna. After making these measurements, connect the antenna to the coaxial cable and measure the input impedance of the cable at a number of frequencies. Then use these measurements in the transmission-line equation to determine the actual antenna impedance at each frequency.

Table 6
Impedance Data for Inverted-V Antenna

Freq (MHz)	R_u *(Ω)*	X_u *@ 10 MHz (Ω)*	X_u *(Ω)*	R_L *(Ω)*	X_L *(Ω)*
27.0	44	85	31.5	24	–65
27.2	60	95	34.9	26	–56
27.4	75	85	31.0	30	–51
27.6	90	40	14.5	32	–42
27.8	90	–20	–7.2	35	–34
28.0	75	–58	–20.7	38	–24
28.2	65	–65	–23.0	40	–19
28.4	56	–52	–18.3	44	–12
28.6	50	–40	–14.0	44	–6
28.8	48	–20	–6.9	47	1
29.0	50	0	0.0	52	8
29.2	55	20	6.8	57	15
29.4	64	30	10.2	63	21
29.6	78	20	6.8	75	26
29.8	85	0	0	78	30
30.0	90	–50	–16.7	89	33

Table 6 and **Fig 33** give an example of such a calculation. The antenna used for this example is a 10-meter inverted V about 30 feet above the ground. The arms of the antenna are separated by a 120° angle. Each arm is exactly 8 feet long, and the antenna is made of #14 wire. The feed line is the 74-foot length of Columbia 1188 characterized earlier.

See Fig 33A. From this plot of impedance measurements, it is very difficult to determine anything about the antenna. Resistance and reactance vary substantially over this frequency range, and the antenna appears to be resonant at 27.7, 29.0 and 29.8 MHz.

The plot in Fig 33B shows the true antenna imped-

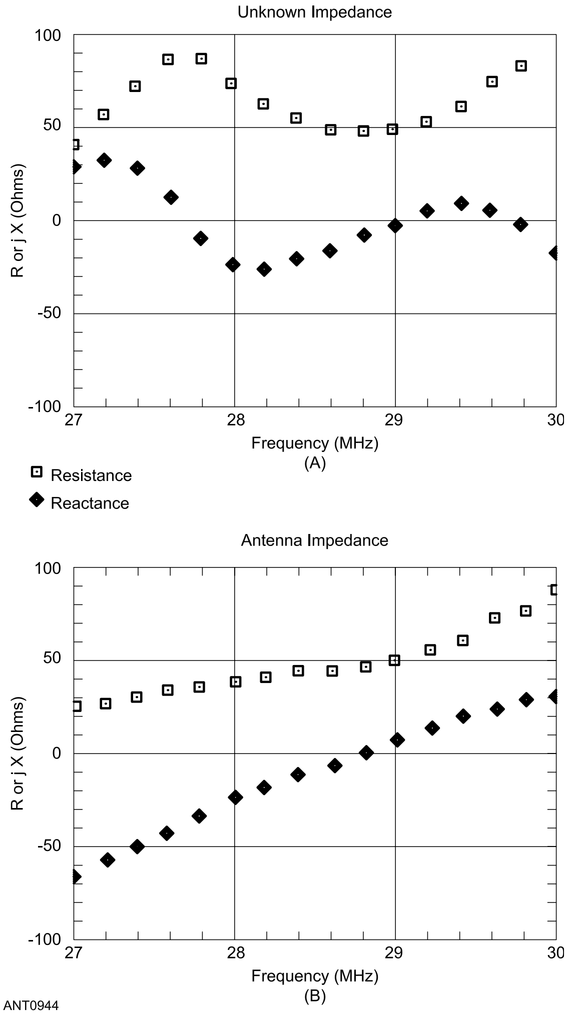

Fig 33–Impedance plot of an inverted-V antenna cut for 29 MHz. At A, a plot of resistances and reactances, measured using the noise bridge, at the end of a 74-foot length of Columbia 1188 coaxial cable. At B, the actual antenna-impedance plot (found using the transmission-line equation to remove the effects of the transmission line).

ance. This plot has been corrected for the effects of the cable using the transmission-line equation. The true antenna resistance and reactance both increase smoothly with frequency. The antenna is resonant at 28.8 MHz, with a radiation resistance at resonance of 47 Ω. This is normal for an inverted V.

When doing the conversions, be careful not to make measurement errors. Such errors introduce more errors into the corrected data. This problem is most significant when the transmission line is near an odd multiple of a $^1/_4$ λ and the line SWR and/or attenuation is high. Measurement errors are probably present if small changes in the input impedance or transmission-line characteristics appear as large changes in antenna impedance. If this effect is present, it can be minimized by making the measurements with a transmission line that is approximately an integer multiple of $^1/_2$ λ.

A Practical Time-Domain Reflectometer

A time-domain reflectometer (TDR) is a simple but powerful tool used to evaluate transmission lines. When used with an oscilloscope, a TDR displays impedance "bumps" (open and short circuits, kinks and so on) in transmission lines. Commercially produced TDRs cost from hundreds to thousands of dollars each, but you can add the TDR described here to your shack for much less. This material is based on a *QST* article by Tom King, KD5HM (see Bibliography), and supplemented with information from the references.

How a TDR Works

A simple TDR consists of a square-wave generator and an oscilloscope. See **Fig 34**. The generator sends a train of dc pulses down a transmission line, and the oscilloscope lets you observe the incident and reflected waves from the pulses (when the scope is synchronized to the pulses).

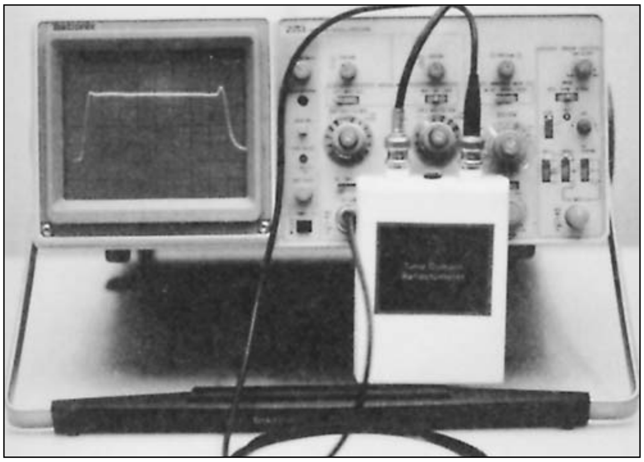

Fig 34—The time-domain reflectometer shown here is attached to a small portable oscilloscope.

A little analysis of the scope display tells the nature and location of any impedance changes along the line. The nature of an impedance disturbance is identified by comparing its pattern to those in **Fig 35**. The patterns are based on the fact that the reflected wave from a disturbance is determined by the incident-wave magnitude and the reflection coefficient of the disturbance. (The patterns shown neglect losses; actual patterns may vary somewhat from those shown.)

The location of a disturbance is calculated with a simple proportional method: The round-trip time (to the disturbance) can be read from the oscilloscope screen (graticule). Thus, you need only read the time, multiply it by the velocity of the radio wave (the speed of light adjusted by the velocity factor of the transmission line) and divide by two. The distance to a disturbance is given by:

$$\ell = \frac{983.6 \times \mathrm{VF} \times \mathrm{t}}{2} \qquad \text{(Eq 18)}$$

where

ℓ = line length in feet

VF = velocity factor of the transmission line (from 0 to 1.0)

t = time delay in microseconds (μs).

The Circuit

The time-domain reflectometer circuit in **Fig 36** consists of a CMOS 555 timer configured as an astable multivibrator, followed by an MPS3646 transistor acting as a 15-ns-risetime buffer. The timer provides a 71-kHz square wave. This is applied to the 50-Ω transmission line under test (connected at J2). The oscilloscope is connected to the circuit at J1.

Construction

An etching pattern for the TDR is shown in **Fig 37**. **Fig 38** is the part-placement diagram. The TDR is

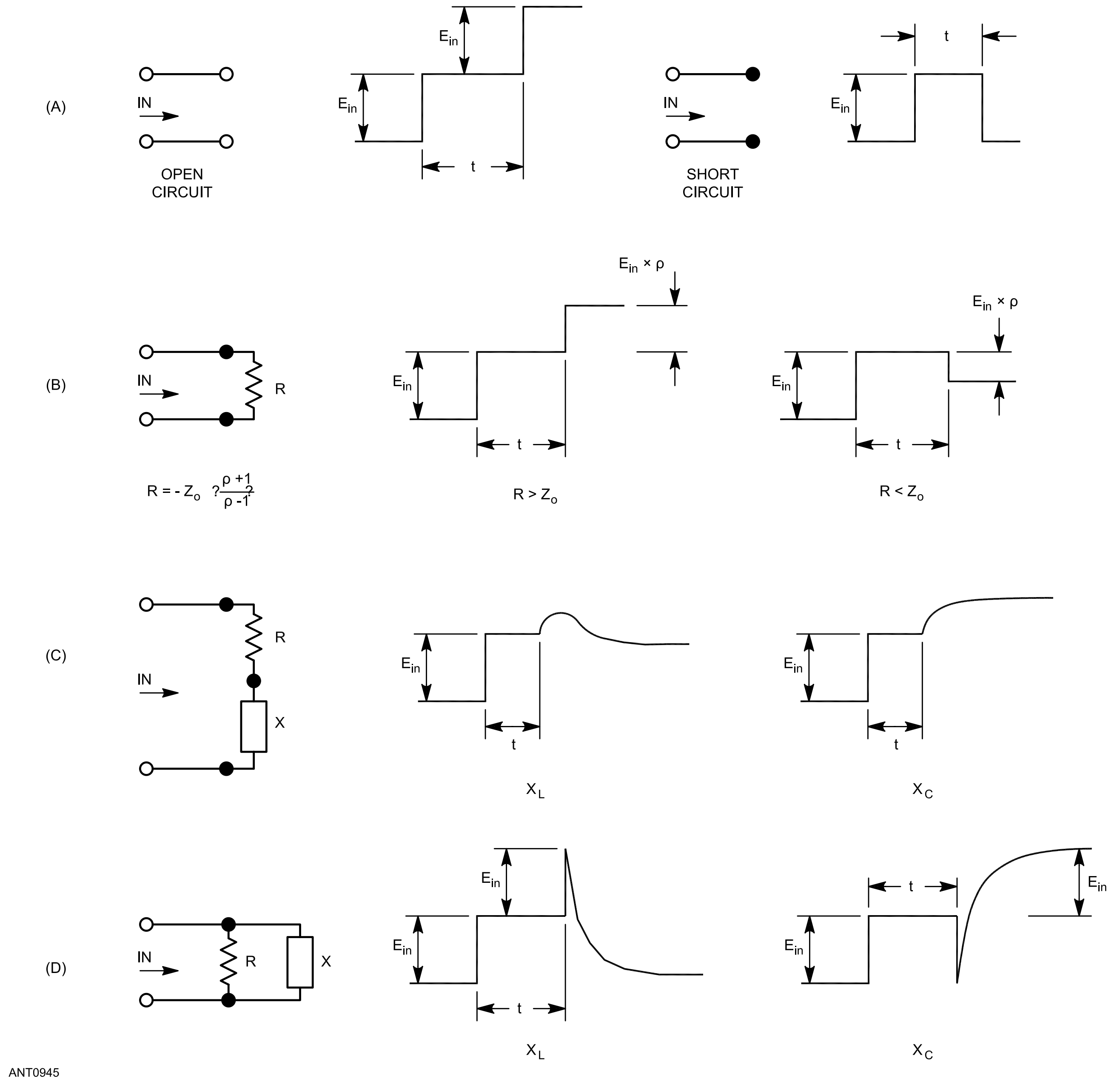

Fig 35—Characteristic TDR patterns for various loads. The location of the load can be calculated from the transit time, t, which is read from the oscilloscope (see text). R values can be calculated as shown (for purely resistive loads only—ρ < 0 when R < Z_0; ρ < 0 when R > Z_0). Values for reactive loads cannot be calculated simply.

designed for a 4 × 3 × 1-inch enclosure (including the batteries). Sl, Jl and J2 are right-angle-mounted components. Two aspects of construction are critical. First use *only* an MPS3646 for Q1. This type was chosen for its good performance in this circuit. If you substitute another transistor, the circuit may not perform properly.

Second, for the TDR to provide accurate measurements, the cable connected to J1 (between the TDR and the oscilloscope) must not introduce impedance mismatches in the circuit. *Do not make this cable from ordinary coaxial cable.* Oscilloscope-probe cable is the best thing to use for this connection.

(It took the author about a week and several phone calls to determine that scope-probe cable isn't "plain old coax." Probe cable has special characteristics that prevent undesired ringing and other problems.)

Mount a binding post at J1 and connect a scope probe to the binding post when testing cables with the TDR. R5 and C2 form a compensation network—much like the networks in oscilloscope probes—to adjust for effects of

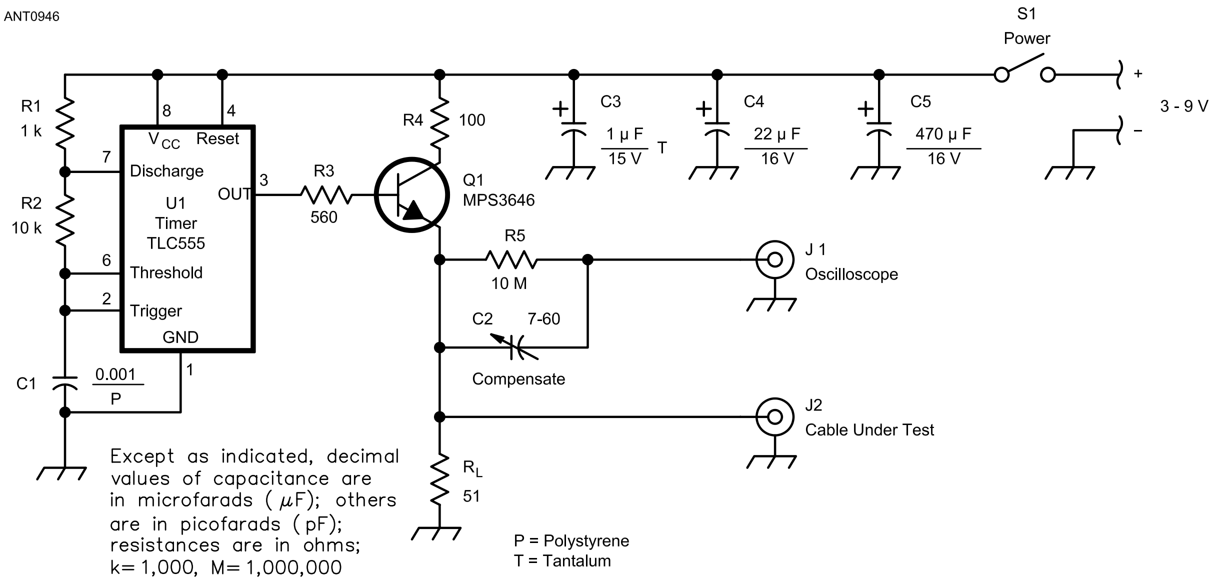
ANT0946
S1
Power
+
3 - 9 V
–
R1
1 k
R2
10 k
C1
0.001
P
8
4
7
6
2
1
3
VCC
Reset
Discharge
U1
Timer
TLC555
Threshold
Trigger
GND
OUT
R3
560
R4
100
Q1
MPS3646
R5
10 M
C2 7-60
Compensate
RL
51
C3
1 μ F
15 V
T
C4
22 μ F
16 V
C5
470 μ F
16 V
J 1
Oscilloscope
J2
Cable Under Test
Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k= 1,000, M= 1,000,000
P = Polystyrene
T = Tantalum

Fig 36—Schematic diagram of the time-domain reflectometer. All resistors are ¼-W, 5% tolerance. U1 is a CMOS 555 timer. Circuit current drain is 10 to 25 mA. When building the TDR, observe the construction cautions discussed in the text. C2 is available from Mouser Electronics, part no. ME242-8050. Right-angle BNC connectors for use at J1 and J2 can be obtained from Newark Electronics, part no. 89N1578. S1 can be obtained from All Electronics, part no. NISW-1. An SPST toggle switch can also be used at S1.

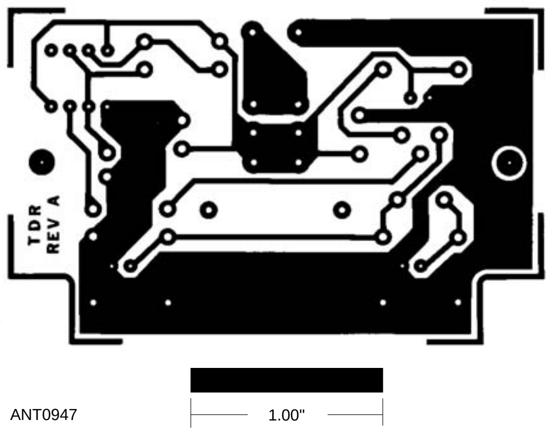
TDR
REV A
ANT0947
1.00"

Fig 37—Full-size PC-board etching pattern for the TDR. Black areas represent unetched copper foil.

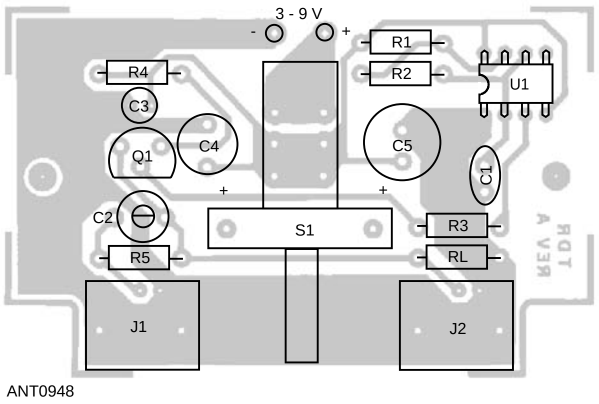
3 - 9 V
-
+
R1
R2
R4
U1
C3
C4
Q1
C5
C1
+
+
C2
S1
R3
R5
RL
J1
J2
TDR
REV A
ANT0948

Fig 38—Part-placement diagram for the TDR. Parts are mounted on the nonfoil side of the board; the shaded area represents an X-ray view of the copper pattern. Be sure to observe the polarity markings of C3, C4 and C5.

the probe wire.

The TDR is designed to operate from dc between 3 and 9 V. Two C cells (in series—3 V) supply operating voltage in this version. The circuit draws only 10 to 25 mA, so the cells should last a long time (about 200 hours of operation). U1 can function with supply voltages as low as 2.25 to 2.5.

If you want to use the TDR in transmission-line systems with characteristic impedances other than 50 Ω, change the value of R_L to match the system impedance as closely as possible.

Calibrating and Using the TDR

Just about any scope with a bandwidth of at least 10 MHz should work fine with the TDR, but for tests in short-length cables, a 50-MHz scope provides for much more accurate measurements. To calibrate the TDR, terminate CABLE UNDER TEST connector, J2, with a 51-Ω

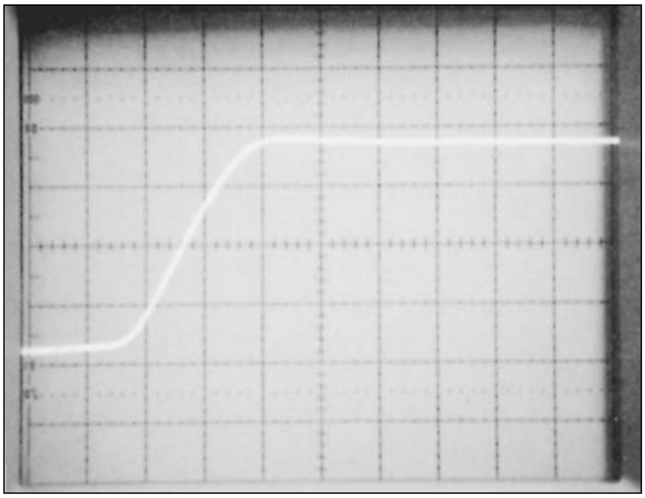

Fig 39—TDR calibration trace as shown on an oscilloscope. Adjust C2 (See Figs 36 and 38) for maximum deflection and sharpest waveform corners during calibration. See text.

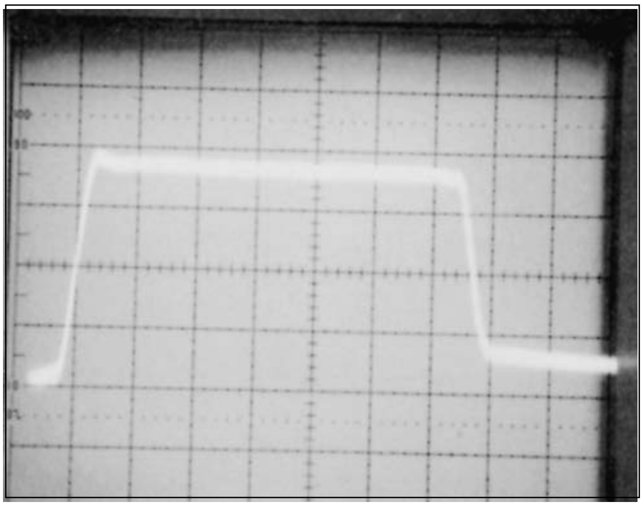

Fig 41—TDR display of the impedance characteristics of the 142-foot Hardline run to the 432-MHz antenna at KD5HM. The scope is set for 0.05 ms per division. See text for discussion.

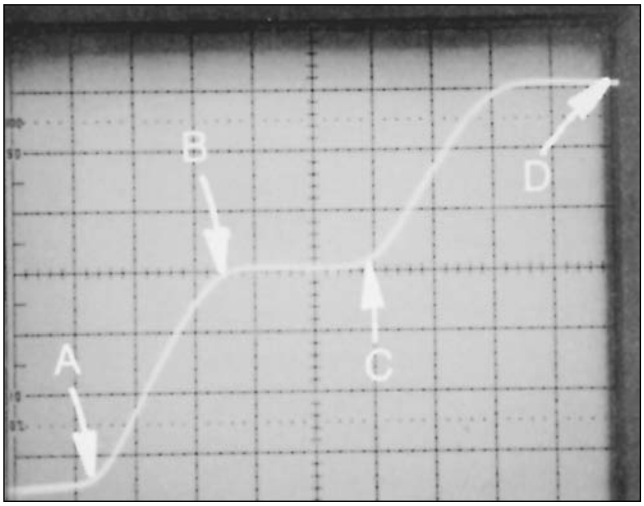

Fig 40—Open-circuited test cable. The scope is set for 0.01 ms per division. See text for interpretation of the waveform.

resistor. Connect the scope vertical input to J1. Turn on the TDR, and adjust the scope timebase so that one square-wave cycle from the TDR fills as much of the scope display as possible (without uncalibrating the timebase). The waveform should resemble **Fig 39**. Adjust C2 to obtain maximum amplitude and sharpest corners on the observed waveform. That's all there is to the calibration process!

To use the TDR, connect the cable under test to J2, and connect the scope vertical input to J1. If the waveform you observe is different from the one you observed during calibration, there are impedance variations in the load you're testing. See **Fig 40**, showing an unterminated test cable connected to the TDR. The beginning of the cable is shown at point A. (AB represents the TDR output-pulse rise time.)

Segment AC shows the portion of the transmission line that has a 50-Ω impedance. Between points C and D, there is a mismatch in the line. Because the scope trace is higher than the 50-Ω trace, the impedance of this part of the line is higher than 50 Ω—in this case, an open circuit.

To determine the length of this cable, read the length of time over which the 50-Ω trace is displayed. The scope is set for 0.01 µs per division, so the time delay for the 50-Ω section is (0.01 µs × 4.6 divisions) = 0.046 µs. The manufacturer's specified velocity factor (VF) of the cable is 0.8. Eq 1 tells us that the 50-Ω section of the cable is

$$\ell = \frac{983.6 \times 0.8 \times 0.046\,\mu\text{s}}{2} = 18.1\ \text{feet}$$

The TDR provides reasonable agreement with the actual cable length—in this case, the cable is really 16.5 feet long. (Variations in TDR-derived calculations and actual cable lengths can occur as a result of cable VFs that can vary considerably from published values. Many cables vary as much as 10% from the specified values.)

A second example is shown in **Fig 41**, where a length of $^3/_4$-inch Hardline is being tested. The line feeds a 432-MHz vertical antenna at the top of a tower. Fig 41 shows that the 50-Ω line section has a delay of (6.6 divisions × 0.05 µs) = 0.33 µs. Because the trace is straight and level at the 50-Ω level, the line is in good shape. The trailing edge at the right-hand end shows where the antenna is connected to the feed line.

To determine the actual length of the line, use the same procedure as before: Using the published VF for the Hardline (0.88) in Eq 1, the line length is

$$\ell = \frac{983.6 \times 0.88 \times 0.33\ \mu\text{s}}{2} = 142.8\ \text{feet}$$

Again, the TDR-derived measurement is in close agreement with the actual cable length (142 feet).

Final Notes

The time-domain reflectometer described here is not frequency specific; its measurements are not made at the frequency at which a system is designed to be used. Because of this, the TDR cannot be used to verify the impedance of an antenna, nor can it be used to measure cable loss at a specific frequency. Just the same, in two years of use, it has never failed to help locate a transmission-line problem. The vast majority of transmission-line problems result from improper cable installation or connector weathering.

Limitations

Certain limitations are characteristic of TDRs because the signal used to test the line differs from the system operating frequency and because an oscilloscope is a broadband device. In the instrument described here, measurements are made with a 71-kHz square wave. That wave contains components at 71 kHz and odd harmonics thereof, with the majority of the energy coming from the lower frequencies. The leading edge of the trace indicates that the response drops quickly above 6 MHz. (The leading edge in Fig 40 is 0.042 µs, corresponding to a period of 0.168 µs and a frequency of 5.95 MHz.) The result is dc pulses of approximately 7 µs duration. The scope display combines the circuit responses to all of those frequencies. Hence, it may be difficult to interpret any disturbance which is narrowband in nature (affecting only a small range of frequencies, and thus a small portion of the total power), or for which the travel time plus pattern duration exceeds 7 µs. The 432-MHz vertical antenna in Fig 41 illustrates a display error resulting from narrow-band response.

The antenna shows as a major impedance disturbance because it is mismatched at the low frequencies that dominate the TDR display, yet it is matched at 432 MHz. For an event that exceeds the observation window, consider a l-µF capacitor across a 50-Ω line. You would see only part of the pattern shown in Fig 35C because the time constant ($1 \times 10^{-6} \times 50 = 50$ ms) is much larger than the 7-µs window.

In addition, TDRs are unsuitable for measurements where there are major impedance changes inside the line section to be tested. Such major changes mask reflections from additional changes farther down the line.

Because of these limitations, TDRs are best suited for spotting faults in dc-continuous systems that maintain a constant impedance from the generator to the load. Happily, most amateur stations would be ideal subjects for TDR analysis, which can conveniently check antenna cables and connectors for short and open-circuit conditions and locate the position of such faults with fair accuracy.

Ground Parameters for Antenna Analysis

This section is taken from an article in *The ARRL Antenna Compendium, Vol 5* by R. P. Haviland, W4MB. In the past, amateurs paid very little attention to the characteristics of the earth (ground) associated with their antennas. There are two reasons for this. First, these characteristics are not easy to measure—even with the best equipment, extreme care is needed. Second, almost all hams have to put up with what they have—there are very few who can afford to move because their location has poor ground conditions! Further, the ground is not a dominant factor in the most popular antennas—a tri-band Yagi at 40 feet or higher, or a 2-meter vertical at roof height, for example.

Even so, there has been a desire and even a need for ground data and for ways to use it. It is very important for vertically polarized antennas. Ground data is useful for antennas mounted at low heights generally, and for such specialized ones as Beverages. The performance of such antennas change a lot as the ground changes.

Importance of Ground Conditions

To see why ground conditions can be important, let us look at some values. For a frequency of 10 MHz, *CCIR Recommendation 368*, gives the distance at which the signal is calculated to drop 10 dB below its free-space level as:

Conductivity (mS/meter)	*Distance for 10 dB Drop (km)*
5000	100
30	15
3	0.3

The high-conductivity condition is for sea-water. Inter-island work in the Caribbean on 40 and 80 meters is easy, whereas 40-meter ground-wave contact is difficult for much of the USA, because of much lower ground conductivity. On the other hand, the Beverage works because of poor ground conductivity.

Fig 42 shows a typical set of expected propagation curves for a range of frequencies. This data is also from *CCIR Recommendation 368* for relatively poor ground, with a dielectric constant of 4 and a conductivity of 3 mS/m (one milliSiemens/meter is 0.001 mho/meter). The same data is available in the *Radio Propagation Handbook*. There are equivalent FCC curves, found in the book *Reference Data for Radio Engineers*, but only the ones near 160 meters are useful. In Florida the author has difficulty hearing stations across town on ground

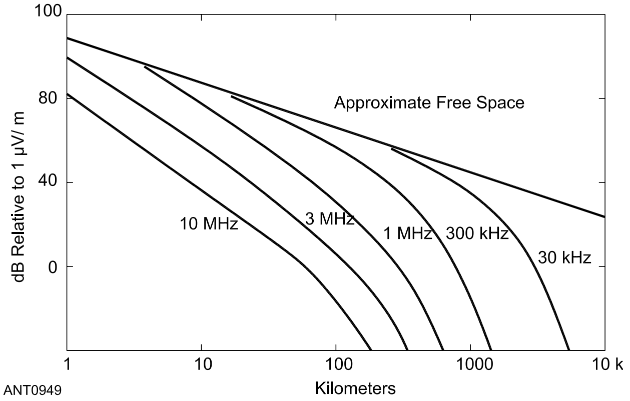

Fig 42—Variation of field strength with distance. Typical field strengths for several frequencies are shown. This is from CCIR data for fairly poor soil, with dielectric constant of 4 and conductivity of 3 mS/m. The curves for good soil are closer to the free-space line, and those for sea water are much closer to the free-space line.

wave, an indication of the poor soil conditions—reflected sky-wave signals are often stronger.

Securing Ground Data

There are only two basic ways to approach this matter of ground data. One is to use generic ground data typical to the area. The second is to make measurements, which haven't really gotten easier. For most amateurs, the best approach seems to be a combination of these—use some simple measurements, and then use the generic data to make a better estimate. Because of equipment costs and measurement difficulties, none of these will be highly accurate for most hams. But they will be much better than simply taking some condition preset into an analysis program. Having a good set of values to plug into an analysis can help you evaluate the true worth of a new antenna project.

Generic Data

In connection with its licensing procedure for broadcast stations, the FCC has published generic data for the

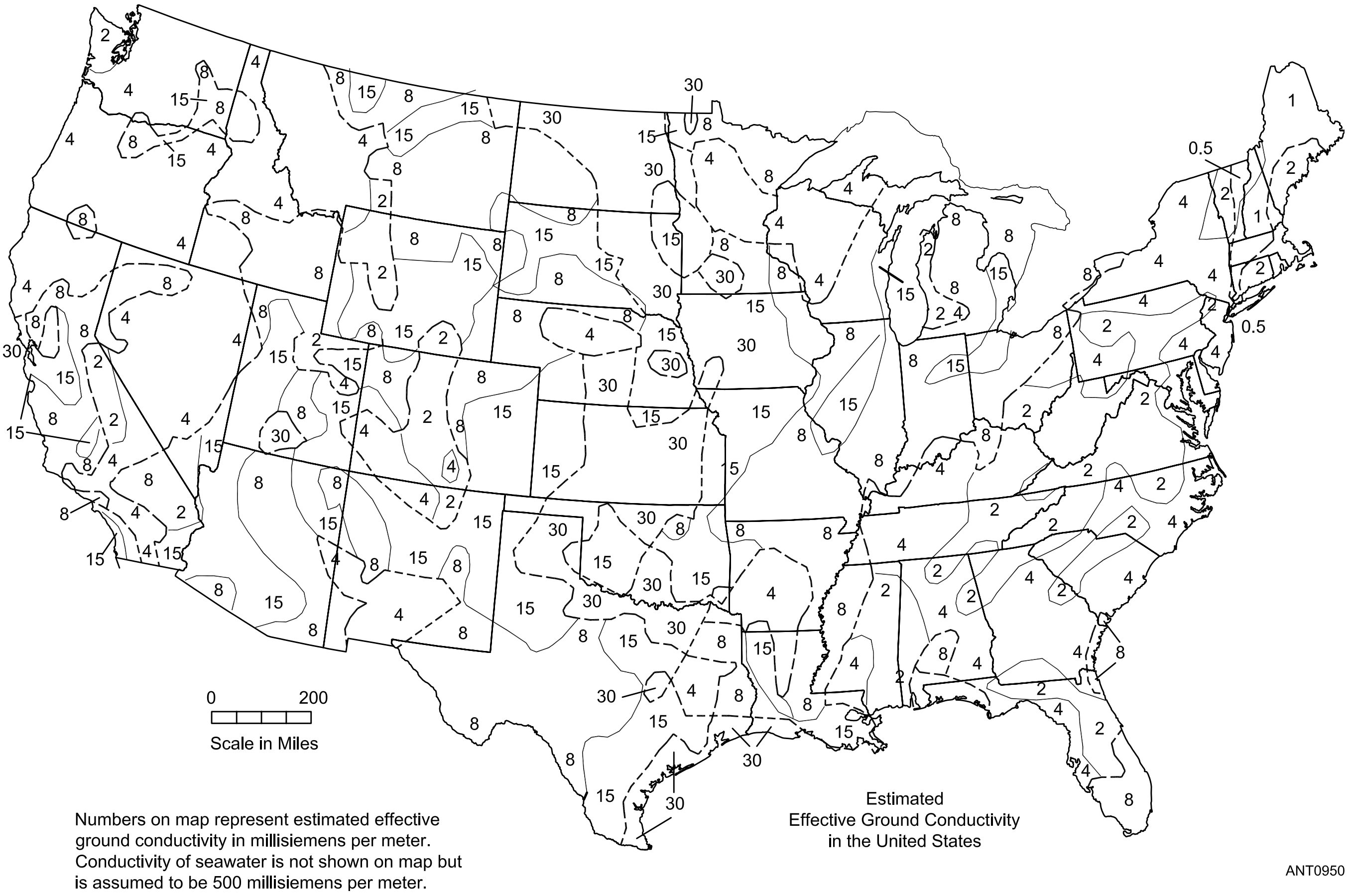

Numbers on map represent estimated effective ground conductivity in millisiemens per meter. Conductivity of seawater is not shown on map but is assumed to be 500 millisiemens per meter.

Fig 43—Estimated effective ground conductivity in the United States. FCC map prepared for the Broadcast Service, showing typical conductivity for continental USA. Values are for the band 500 to 1500 KHz. Values are for flat, open spaces and often will not hold for other types of commonly found terrain, such as seashores, river beds, etc.

entire country. This is reproduced in **Fig 43**, a chart showing the "estimated effective ground conductivity in the United States." A range of 30:1 is shown, from 1 to 30 mS/m. An equivalent chart for Canada has been prepared, originally by DOT, now DOC.

Of course, some judgment is needed when trying to use this data for your location. Broadcast stations are likely to be in open areas, so the data should not be assumed to apply to the center city. And a low site near the sea is likely to have better conductivity than the generic chart for, say, the coast of Oregon. Other than such factors, this chart gives a good first value, and a useful cross-check if some other method is used.

Still another FCC-induced data source is the license application of your local broadcast station. This includes calculated and measured coverage data. This may include specific ground data, or comparison of the coverage curves with the CCIR or FCC data to give the estimated ground conductivity. Another set of curves for ground conditions are those prepared by SRI. These give the conductivity and dielectric constant versus frequency for typical terrain conditions. These are reproduced as **Fig 44** and **Fig 45**. By inspecting your own site, you may select the curve most appropriate to your terrain. The curves are based on measurements at a number of sites across the USA, and are averages of the measured values.

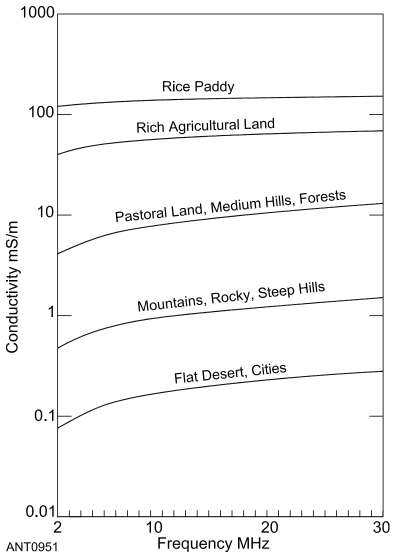

Fig 44—Typical terrain conductivities versus frequency for 5 types of soils. This was measured by SRI. Units are mS/m. Conductivity of seawater is usually taken as 5000 mS/m. Conductivity of fresh water depends on the impurities present, and may be very low. To extrapolate conductivity values (for 500 to 1500 KHz) shown in Fig 43 for a particular geographic area to a different frequency, move from the conductivity at the left edge of Fig 44 to the desired frequency. For example, in rocky New Hampshire, with a conductivity of 1 mS/m at BC frequencies, the effective conductivity at 14 MHz would be approximately 4 mS/m.

Figs 46 through 48 are data derived from these measurements. **Fig 46** gives the ground-dissipation factor. Sea water has low loss (a high dissipation factor), while soil in the desert or in the city is very lossy, with a low dissipation factor. **Fig 47** gives the skin depth, the distance for the signal to decease to 63% of its value at the surface. Penetration is low in high-conductivity areas and deep in low-conductivity soil. Finally, **Fig 48** shows the wavelength in the earth. For example, at 10 meters (30 MHz), the wavelength in sea water is less than 0.3 meters. Even in the desert, the wavelength has been reduced to about 6 meters at this frequency. This is one reason why buried antennas have peculiar properties. Lacking other data, it is suggested that the values of Figs 44 and 45 be used in computer antenna modeling programs.

Measuring Ground Conditions

W2FNQ developed a simple technique to measure low-frequency earth conductivity, which has been used by W2FMI. The test setup is drawn in **Fig 49**, and uses a very old technique of 4-terminal resistivity measurements. For probes of $^{9}/_{16}$-inch diameter, spaced 18 inches and pene-trating 12 inches into the earth, the conductivity is:

$$C = 21\ V_1/V_2 \quad \text{mS/m} \qquad \text{(Eq 19)}$$

The voltages are conveniently measured by a digital voltmeter, to an accuracy of about 2%. In soil suitable for farming, the probes can be copper or aluminum. The strength of iron or copperweld may be needed in hard soils. A piece of 2 × 4 or 4 × 4 with guide holes drilled through it will help maintain proper spacing and vertical alignment of the probes. Use care when measuring—there is a shock hazard. An isolating transformer with a 24-V secondary instead of 115 V will reduce the danger.

Ground conditions vary quite widely over even small

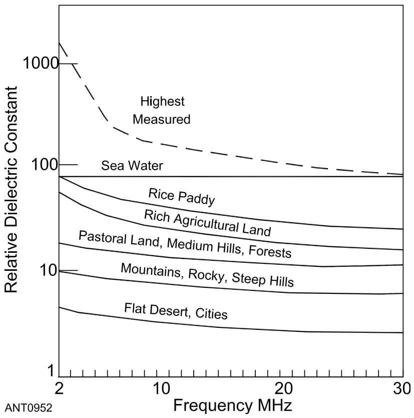

Fig 45—Typical terrain relative dielectric constant for the 5 soil types of Fig 44, plus sea water. The dashed curve shows the highest measured values reported, and usually indicates mineralization.

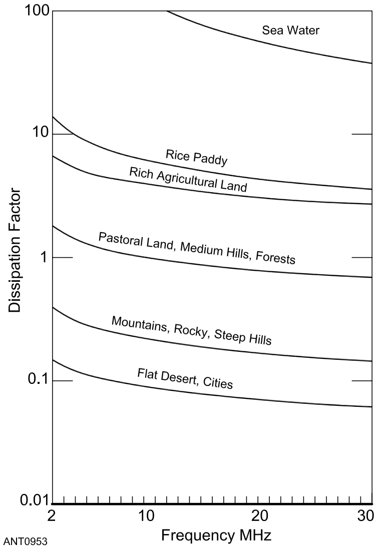

Fig 46—Typical values of dissipation factor. The soil behaves as a leaky dielectric. These curves showing the dimensionless dissipation factor versus frequency for various types of soils and for sea water. The dissipation factor is inversely related to soil conductivity. Among other things, a high dissipation factor indicates that a signal penetrating the soil or water will decrease in strength rapidly with depth.

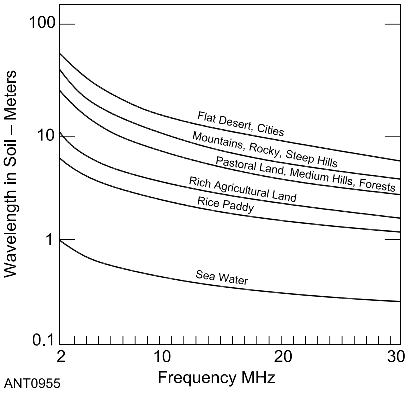

Fig 48—Typical values of wavelength in soil. Because of its dielectric constant, the wavelength in soils and water will be shorter than that for a wave traveling in air. This can be important, since in a Method of Moment the accuracy is affected by the number of analysis segments per wavelength. Depending on the program being used, adjust the number of segments for antennas wholly or partly in the earth, for ground rods, and for antennas very close to earth.

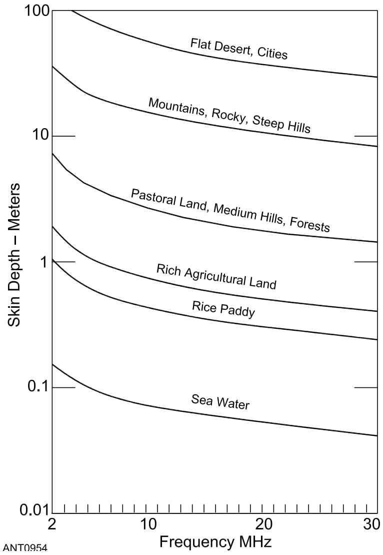

Fig 47—Typical values of skin depth. The skin depth is the depth at which a signal will have decreased to 1/e of its value at the surface (to about 30%). The effective height above ground is essentially the same as the physical height for sea water, but may be much greater for the desert. For practical antennas, this may increase low-angle radiation, but at the same time will increase ground losses.

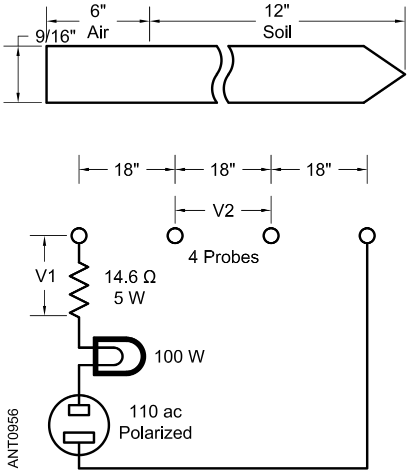

Fig 49—Low-frequency conductivity measurement system. A 60-Hz measuring system devised by W2FNQ and used by W2FMI. The basic system is widely used in geophysics. Use care to be certain that the plug connection is correct. A better system would use a lower voltage and an isolation transformer. Measure the value of V_2 with no power applied—there may be stray ground currents present, especially if there is a power station or an electric railway close.

areas. It is best to make a number of measurements around the area of the antenna, and average the measured values.

While this measurement gives only the low-frequency conductivity, it can be used to select curves in Fig 44 to give an estimate of the conductivity for the common ham bands. Assume that the 60 Hz value is valid at 2 MHz, and find the correct value on the left axis. Move parallel to the curves on the figure to develop the estimated curve for other soil conditions.

A small additional refinement is possible. If the dielectric constant from Fig 45 is plotted against the conductivity from Fig 44 for a given frequency, a scatter plot develops, showing a trend to higher dielectric constant as conductivity increases. At 14 MHz, the relation is:

$$k = \sqrt{1000/C} \qquad \text{(Eq 20)}$$

where k is the dielectric constant and C is the measured conductivity. Using these values in *MININEC* or *NEC* calculations should give better estimates than countrywide average values.

Direct Measurement of Ground Properties

For really good values, both the conductivity and dielectric constant should be measured at the operating frequency. One way of doing this is the two-probe technique described in George Hagn's article (see Bibliography). This was the technique used to secure the data for Figs 44 through 48. The principle is sketched in **Fig 50**. In essence, the two probes form a short, open-circuited, two-wire transmission line. As shown by the equations for such lines, the input impedance is a function of the conductivity and dielectric constant of the medium. A single measurement is difficult to calculate, since the end effect of the two probes must be determined, a complex task if they are pointed for easy driving. The calculation is greatly simplified if a set of measurements is made with several sets of probes that vary in length by a fixed ratio, since the measured difference is largely due to the increased two-wire length, with some change due to the change in soil moisture with depth.

The impedance to be measured is high because of the short line length, so impedance bridges are not really suitable. An RF vector impedance meter, such as the HP-4193A, is probably the best instrument to use, with a RF susceptance bridge, such as the GR-821A, next best. With care, a Q-meter can be substituted. Because of the rarity of these instruments among amateurs, this method of measurement is not explored further here.

Indirect Measurement

Since the terminal impedance and resonant frequency of an antenna change as the antenna approaches earth, measurement of an antenna at one or more heights permits an analysis of the ground characteristics. The technique is to calculate the antenna drive impedance for an assumed ground condition, and compare this with measured values. If not the same, another set of ground conditions is assumed, and the process is repeated. It is best to have a plan to guide the assumptions.

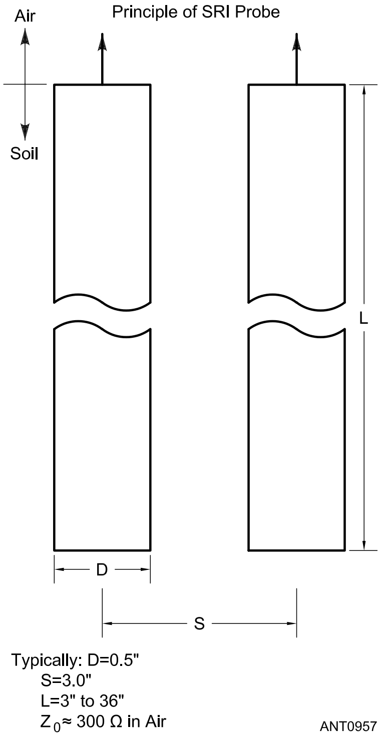

Fig 50—High-frequency conductivity/dielectric constant measurement system. System for measuring ground conditions at frequencies up to about 100 MHz, devised by SRI and used to obtain the data in Figs 44 through 48. Basically, this is a section of transmission line with soil as the dielectric. Requires measurement of high impedances to good accuracy.

In connection with his studies of transmission lines, Walt Maxwell, W2DU, made such measurements on 20, 40 and 80 meters. Some of the data was included in his book *Reflections*. The following example is based on his 80-meter data. Data came from his Table 20-1, for a 66-foot, 2-inch dipole of #14 wire at 40 feet above ground. His table gives an antenna impedance of 72.59 + *j* 1.38 Ω at 7.15 MHz.

Table 7 shows calculated antenna impedances for ground conductivities of three different ground conductivities: 10, 1 and 0.1 mS/m, and for dielectric constants of 3, 15 and 80. The nearest value to the measured drive impedance is for a conductivity of 0.1 mS/m and a dielectric constant of 3. Figs 44 and 45 indicate that these are typical of flat desert and city land. The effect on antenna performance is shown in **Fig 51**. The maximum lobe gain for soil typical of a city is over 2 dB lower than that for the high-conductivity, high-dielectric constant value. Note that the maximum lobe occurs for a radiation angle that is directly overhead.

Table 7

Calculated values of drive resistance, in ohms, for an 40 meter dipole at 40 feet elevation versus conductivity and dielectric constant.

Conductivity	*Dielectric Constant*		
(mS/m)	*3*	*15*	*50*
10	89.78 – *j* 12.12	88.53 – *j* 10.69	88.38 – *j* 7.59
1	80.05 – *j* 17.54	83.72 – *j* 10.23	87.33 – *j* 6.98
0.1	76.44 – *j* 15.69	83.18 – *j* 9.85	97.30 – *j* 6.46

The value measured by W2DU was 72.59 – *j* 1.28 Ω, and compares closest to the poor soil condition of dielectric constant of 3 and conductivity of 0.1 mS/m.

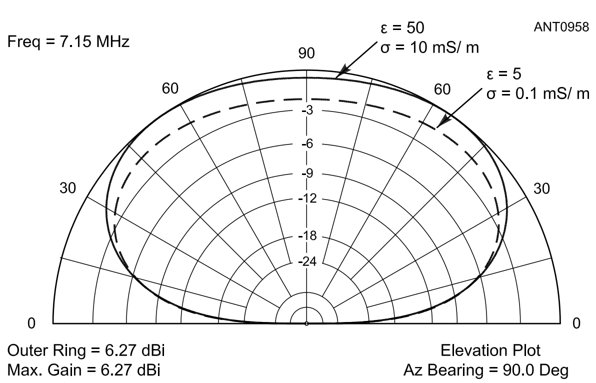

Fig 51—Plot showing computed elevation patterns for 40-foot high, 40-meter dipole for two different ground conditions: poor ground, with dielectric constant of 3 and conductivity of 0.1 mS/m, and good ground, with dielectric constant of 50 and conductivity of 10 mS/m. Note that for a low horizontal antenna, high-angle radiation is most affected by poor ground, with low-angle radiation least affected by ground characteristics.

The ground at the W2DU QTH is a suburban Florida lot, covered with low, native vegetation. The ground is very sandy (a fossil sand dune), and is some 60-70 feet above sea-level. Measurements were made near the end of the Florida dry season. The water table is estimated to be 20 to 30 feet below the surface. Thus the calculated and measured values are reasonably consistent.

In principle, a further analysis, using values around 0.1 mS/m conductivity and 3 for dielectric constant, will give a better ground parameter estimate. However, the results should be taken with a grain of salt, because the opportunities for error in the computer modeling must be considered. The antenna should have no sag, and its length and height should be accurate. The measurement must be with accurate equipment, free from strays, such as current on the outer conductor of the coax. The feed-point gap effect must be estimated. Further, the ground itself under the antenna must be flat and have constant characteristics for modeling to be completely accurate.

Finally, the feed-line length and velocity constant of the transmission line must be accurately measured for transfer of the measured values at the feeding end of the transmission line to the antenna itself. Because of all the possibilities for error, most attempts at precision should be based on measured values at two or three frequencies, and preferably at two or three heights. Orienting the antenna to right angles for another set of measurements may be useful. Obviously, this can involve a lot of detailed work.

The author was not been able to find any guidelines for the best height or frequency. The data in the book *Exact Image Method for Impedance Computation of Antennas Above the Ground* suggests that a height of 0.3 λ will give good sensitivity to ground conditions. Very low heights may give confusing results, since several combinations of ground parameters can give nearly the same drive impedance. Both this data and experience suggest that sensitivity to ground for heights above 0.75 λ is small or negligible.

If an overall conclusion about ground characteristics is needed, we can just restate from the first paragraph—it is not greatly important for the most common horizontally polarized antenna installations. But it's worth taking a look when you need to depart from typical situations, or when the performance of a vertically polarized antenna is contemplated. Then the techniques outlined here can be helpful.

A Switchable RF Attenuator

A switchable RF attenuator is helpful for making antenna-gain comparisons or for plotting antenna radiation patterns. You may switch attenuation in or out of the line leading to the receiver to obtain an initial reference reading on a signal strength meter. Some form of attenuator is also helpful for locating hidden transmitters, where the real trick is pinpointing the signal source from within a few hundred feet. At such a close distance, strong signals may overload the front end of the receiver, making it impossible to obtain any indication of a bearing.

The attenuator of **Figs 52** and **53** is designed for low power levels, not exceeding $^1/_4$ watt. If for some reason the attenuator will be connected to a transceiver, a means of bypassing the unit during transmit periods must be devised. An attenuator of this type is commonly called a *step attenuator*, because any amount of attenuation from 0 dB to the maximum available (81 dB for this particular instrument) may be obtained in steps of 1 dB. As each switch is successively thrown from the OUT to the IN position, the attenuation sections add in cascade to yield the total of the attenuator steps switched in. The maximum attenuation of any single section is limited to 20 dB because leak-through would probably degrade the accuracy of higher values. The tolerance of resistor values also becomes more significant regarding accuracy at higher attenuation values.

A good quality commercially made attenuator will cost upwards from $150, but for less than $25 in parts and a few hours of work, you can build an attenuator at home. It will be suitable for frequencies up to 450 MHz. Double-sided pc board is used for the enclosure. The version of the attenuator shown in Fig 52 has identification lettering etched into the top surface (or front panel) of the unit. This adds a nice touch and is a permanent means of labeling. Of course rub-on transfers or Dymo tape labels could be used as well.

Female BNC single-hole, chassis-mount connectors are used at each end of the enclosure. These connectors

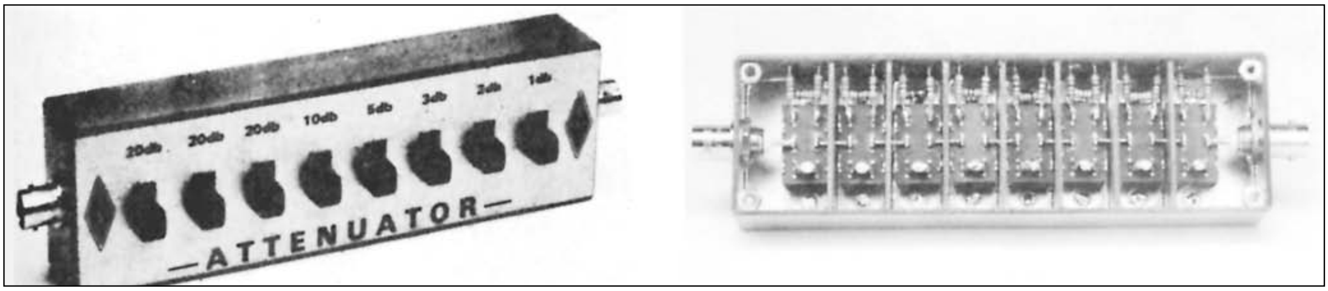

Fig 52—A construction method for a step attenuator. Double-sided circuit-board material, unetched (except for panel identification), is cut to the desired size and soldered in place. Flashing copper may also be used, although it is not as sturdy. Shielding partitions between sections are necessary to reduce signal leakage. Brass nuts soldered at each of the four corners allow machine screws to secure the bottom cover. The practical limit for total attenuation is 80 or 90 dB, as signal leakage around the outside of the attenuator will defeat attempts to obtain much greater amounts.

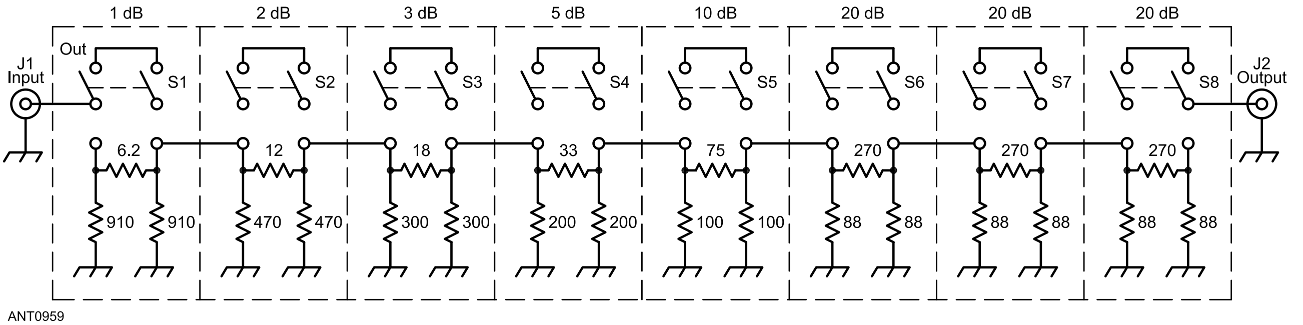

Fig 53—Schematic diagram of the step attenuator, designed for a nominal impedance of 50 Ω. Resistance values are in ohms. Resistors are $^1/_4$-watt, carbon-composition types, 5% tolerance. Broken lines indicate walls of circuit-board material. A small hole is drilled through each partition wall to route bus wire. Keep all leads as short as possible. The attenuator is bilateral; that is, the input and output ends may be reversed.

J1, J2—Female BNC connectors, Radio Shack 278-105 or equiv.

S1-S8, incl.—DPDT slide switches, standard size. (Avoid subminiature or toggle switches.) Stackpole S-5022CD03-0 switches are used here.

provide a means of easily connecting and disconnecting the attenuator.

Construction

After all the box parts are cut to size and the necessary holes made, scribe light lines to locate the inner partitions. Carefully tack-solder all partitions in position. A 25-W pencil type of iron should provide sufficient heat. Dress any pc board parts that do not fit squarely. Once everything is in proper position, run a solder bead all the way around the joints. Caution! Do not use excessive amounts of solder, as the switches must later be fit flat inside the sections. Complete the top, sides, ends and partitions. Dress the outside of the box to suit your taste. For instance, you might wish to bevel the box edges. Buff the copper with steel wool, add lettering, and finish off the work with a coat of clear lacquer or polyurethane varnish.

Using a little lacquer thinner, soak the switches to remove the grease that was added during their manufacture. When they dry, spray the inside of the switches lightly with a TV tuner cleaner/lubricant. Use a sharp drill bit (about 3/16 inch will do), and countersink the mounting holes on the actuator side of the switch mounting plate. This ensures that the switches will fit flush against the top plate. At one end of each switch, bend the two lugs over and solder them together. Cut off the upper halves of the remaining switch lugs. (A close look at Fig 52 will help clarify these steps.)

Solder the series-arm resistors between the appropriate switch lugs. Keep the lead lengths as short as possible and do not overheat the resistors. Now solder the switches in place to the top section of the enclosure by flowing solder through the mounting holes and onto the circuit-board material. Be certain that you place the switches in their proper positions; correlate the resistor values with the degree of attenuation. Otherwise, you may wind up with the 1-dB step at the wrong end of the box how embarrassing!

Once the switches are installed, thread a piece of #18 bare copper wire through the center lugs of all the switches, passing it through the holes in the partitions. Solder the wire at each switch terminal. Cut the wire between the poles of each individual switch, leaving the wire connecting one switch pole to that of the neighboring one on the other side of the partition, as shown in Fig 52. At each of the two end switch terminals, leave a wire length of approximately 1/8 inch. Install the BNC connectors and solder the wire pieces to the connector center conductors.

Now install the shunt-arm resistors of each section. Use short lead lengths. Do not use excessive amounts of heat when soldering. Solder a no. 4-40 brass nut at each inside corner of the enclosure. Recess the nuts approximately 1/16-inch from the bottom edge of the box to allow sufficient room for the bottom panel to fit flush. Secure the bottom panel with four no. 4-40, 1/4-inch machine screws and the project is completed. Remember to use caution, always, when your test setup provides the possibility of transmitting power into the attenuator.

A Portable Field Strength Meter

Few amateur stations, fixed or mobile, are without need of a field-strength meter. An instrument of this type serves many useful purposes during antenna experiments and adjustments. When work is to be done from many wavelengths away, a simple wavemeter lacks the necessary sensitivity. Further, such a device has a serious fault because its linearity leaves much to be desired. The information in this section is based on a January 1973 *QST* article by Lew McCoy, W1ICP.

The field-strength meter described here takes care of these problems. Additionally, it is small, measuring only 4 × 5 × 8 inches. The power supply consists of two 9-volt batteries. Sensitivity can be set for practically any amount desired. However, from a usefulness standpoint, the circuit should not be too sensitive or it will respond to unwanted signals. This unit also has excellent linearity with regard to field strength. (The field strength of a received signal varies inversely with the distance from the source, all other things being equal.) The frequency range includes all amateur bands from 3.5 through

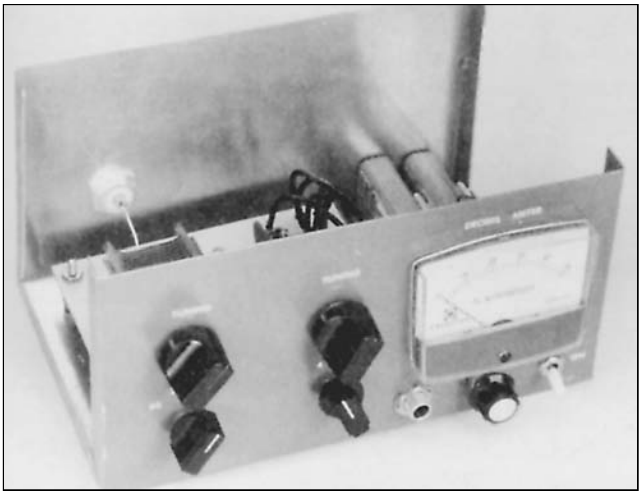

Fig 54—The linear field strength meter. The control at the upper left is for C1 and the one to the right for C2. At the lower left is the band switch, and to its right the sensitivity switch. The zero-set control for M1 is located directly below the meter.

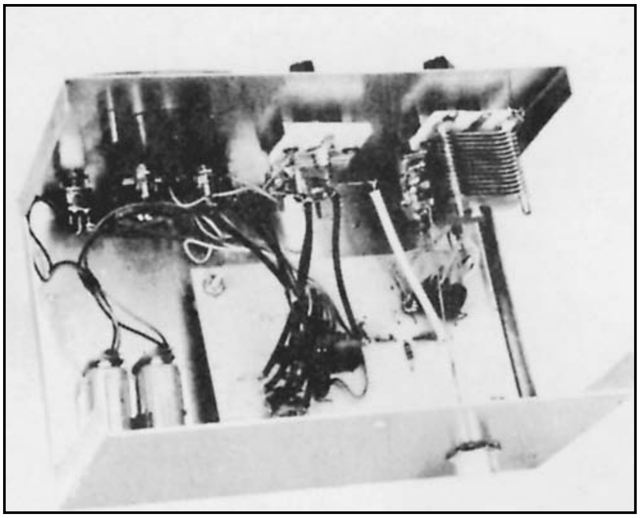

Fig 55—Inside view of the field-strength meter. At the upper right is C1 and to the left, C2. The dark leads from the circuit board to the front panel are the shielded leads described in the text.

148 MHz, with band-switched circuits, thus avoiding the use of plug-in inductors. All in all, it is a quite useful instrument.

The unit is pictured in **Figs 54** and **55**, and the schematic diagram is shown in **Fig 56**. A type 741 op-amp IC is the heart of the unit. The antenna is connected to J1, and a tuned circuit is used ahead of a diode detector. The rectified signal is coupled as dc and amplified in the op amp. Sensitivity of the op amp is controlled by inserting resistors R3 through R6 in the circuit by means of S2.

With the circuit shown, and in its most sensitive setting, M1 will detect a signal from the antenna on the order of 100 µV. Linearity is poor for approximately the first 1/5 of the meter range, but then is almost straight-line from there to full-scale deflection. The reason for the poor linearity at the start of the readings is because of nonlinearity of the diodes at the point of first conduction. However, if gain measurements are being made this is of no real importance, as accurate gain measurements can be made in the linear portion of the readings.

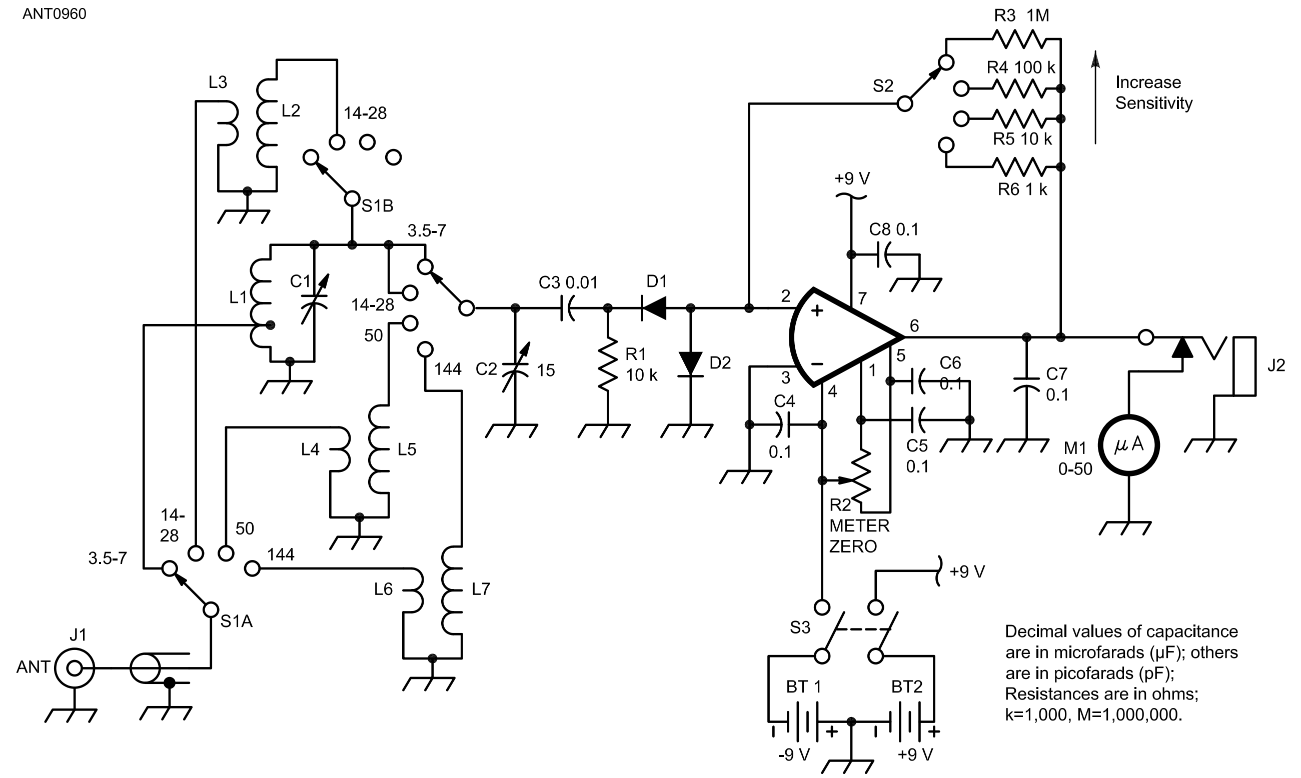
ANT0960
R3 1M
R4 100 k
R5 10 k
R6 1 k
S2
Increase Sensitivity
+9 V
C8 0.1
L3
L2
14-28
S1B
3.5-7
L1
C1
14-28
50
144
C3 0.01
D1
C2 15
R1 10 k
D2
2
7
6
3
5
1
4
C6 0.1
C4 0.1
C5 0.1
C7 0.1
J2
M1 0-50
µA
R2 METER ZERO
L4
L5
14-28
50
144
3.5-7
L6
L7
S1A
J1
ANT
S3
+9 V
BT 1
BT2
-9 V
+9 V
Decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); Resistances are in ohms; k=1,000, M=1,000,000.

Fig 56—Circuit diagram of the linear field strength meter. All resistors are 1/4- or 1/2-W composition types.

C1 — 140 pF variable.
C2 — 15-pF variable
D1, D2 — 1N914 or equiv.
L1 — 34 turns #24 enam. wire wound on an Amidon T-68-2 core, tapped 4 turns from ground end.
L2 — 12 turns #24 enam. wire wound on T-68-2 core.
L3 — 2 turns #24 enam. wire wound at ground end of L2.
L4 — 1 turn #26 enam. wire wound at ground end of L5.
L5 — 12 turns #26 enam. wire wound on T-25-12 core.
L6 — 1 turn #26 enam. wire wound at ground end of L7.
L7 — 1 turn #18 enam. wire wound on T-25-12 core.
M1 — 50 or 100 µA dc.
R2 — 10-kΩ control, linear taper.
S1 — Rotary switch, 3 poles, 5 positions, 3 sections.
S2 — Rotary switch, 1 pole, 4 positions.
S3 — DPST toggle.
U1 — Type 741 op amp. Pin numbers shown are for a 14-pin package.

The 741 op amp requires both a positive and a negative voltage source. This is obtained by connecting two 9-volt batteries in series and grounding the center. One other feature of the instrument is that it can be used remotely by connecting an external meter at J2. This is handy if you want to adjust an antenna and observe the results without having to leave the antenna site.

L1 is the 3.5/7 MHz coil and is tuned by C1. The coil is wound on a toroid form. For 14, 21 or 28 MHz, L2 is switched in parallel with L1 to cover the three bands. L5 and C2 cover approximately 40 to 60 MHz, and L7 and C2 from 130 MHz to approximately 180 MHz. The two VHF coils are also wound on toroid forms.

Construction Notes

The majority of the components may be mounted on an etched circuit board. A shielded lead should be used between pin 4 of the IC and S2. The same is true for the leads from R3 through R6 to the switch. Otherwise, parasitic oscillations may occur in the IC because of its very high gain.

In order for the unit to cover the 144-MHz band, L6 and L7 should be mounted directly across the appropriate terminals of S1, rather than on a circuit board. The extra lead length adds too much stray capacitance to the circuit. It isn't necessary to use toroid forms for the 50- and 144-MHz coils. They were used in the version described here simply because they were available. You may substitute air-wound coils of the appropriate inductance.

Calibration

The field strength meter can be used *as is* for a relative-reading device. A linear indicator scale will serve admirably. However, it will be a much more useful instrument for antenna work if it is calibrated in decibels, enabling the user to check relative gain and front-to-back ratios. If you have access to a calibrated signal generator, connect it to the field-strength meter and use different signal levels fed to the device to make a calibration chart. Convert signal-generator voltage ratios to decibels by using the equation

$$\text{dB} = 20 \log (V1/V2) \quad \text{(Eq 21)}$$

where

V1/V2 is the ratio of the two voltages

log is the common logarithm (base 10)

Let's assume that M1 is calibrated evenly from 0 to 10. Next, assume we set the signal generator to provide a reading of 1 on M1, and that the generator is feeding a 100-µV signal into the instrument. Now we increase the generator output to 200 µV, giving us a voltage ratio of 2:1. Also let's assume M1 reads 5 with the 200-µV input. From the equation above, we find that the voltage ratio of 2 equals 6.02 dB between 1 and 5 on the meter scale. M1 can be calibrated more accurately between 1 and 5 on its scale by adjusting the generator and figuring the ratio. For example, a ratio of 126 µV to 100 µV is 1.26, corresponding to 2.0 dB. By using this method, all of the settings of S2 can be calibrated. In the instrument shown here, the most sensitive setting of S2 with R3, 1 MΩ, provides a range of approximately 6 dB for M1. Keep in mind that the meter scale for each setting of S1 must be calibrated similarly for each band. The degree of coupling of the tuned circuits for the different bands will vary, so each band must be calibrated separately.

Another method for calibrating the instrument is using a transmitter and measuring its output power with an RF wattmeter. In this case we are dealing with power rather than voltage ratios, so this equation applies:

$$\text{dB} = 10 \log (P1/P2) \quad \text{(Eq 22)}$$

where P1/P2 is the power ratio.

With most transmitters the power output can be varied, so calibration of the test instrument is rather easy. Attach a pickup antenna to the field-strength meter (a short wire a foot or so long will do) and position the device in the transmitter antenna field. Let's assume we set the transmitter output for 10 W and get a reading on M1. We note the reading and then increase the output to 20 W, a power ratio of 2. Note the reading on M1 and then use Eq 2. A power ratio of 2 is 3.01 dB. By using this method the instrument can be calibrated on all bands and ranges.

With the tuned circuits and coupling links specified in Fig 56, this instrument has an average range on the various bands of 6 dB for the two most sensitive positions of S2, and 15 dB and 30 dB for the next two successive ranges. The 30-dB scale is handy for making front-to-back antenna measurements without having to switch S2.

An RF Current Probe

The RF current probe of **Figs 57** through **59** operates on the magnetic component of the electromagnetic field, rather than the electric field. Since the two fields are precisely related, as discussed in Chapter 23, the relative field strength measurements are completely equivalent. The use of the magnetic field offers certain advantages, however. The instrument may be made more compact for the same sensitivity, but its principal advantage is that it may be used near a conductor to measure the current flow without cutting the conductor.

In the average amateur location there may be substantial currents flowing in guy wires, masts and towers, coaxial-cable braids, gutters and leaders, water and gas pipes, and perhaps even drainage pipes. Current may be flowing in telephone and power lines as well. All of these RF currents may have an influence on antenna patterns or can be of significance in the case of RFI.

The circuit diagram of the current probe appears in Fig 58, and construction is shown in the photo, Fig 59. The winding data given here apply only to a ferrite rod of the particular dimensions and material specified. Almost any microammeter can be used, but it is usually convenient to use a rather sensitive meter and provide a series resistor to swamp out nonlinearity arising from diode conduction characteristics. A control is also used to adjust instrument sensitivity as required during operation. The tuning capacitor may be almost anything that will cover the desired range.

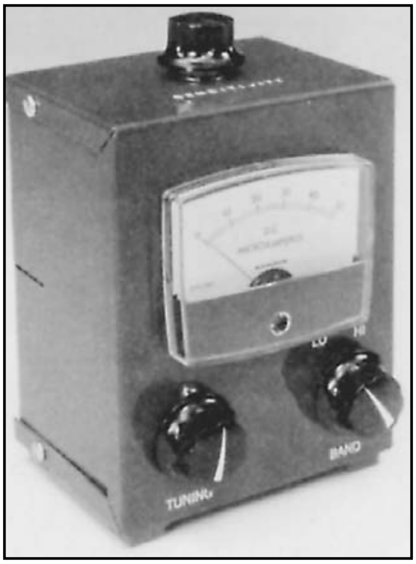

Fig 57—The RF current probe. The sensitivity control is mounted at the top of the instrument, with the tuning and band switches on the lower portion of the front panel. Frequency calibration of the tuning control was not considered necessary for the intended use of this particular instrument, but marks identifying the various amateur bands would be helpful. If the unit is provided with a calibrated dial, it can also be used as an absorption wavemeter.

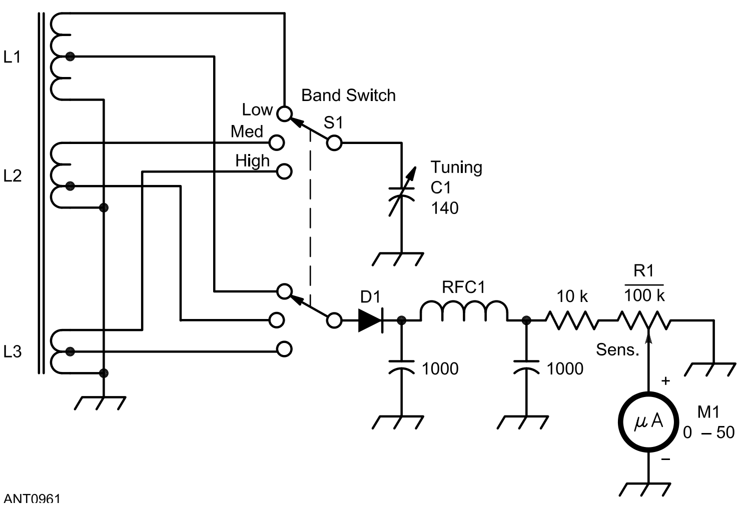

Fig 58—Schematic diagram of the RF current probe. Resistances are in ohms; k = 1000. Capacitances are in picofarads; fixed capacitors are silver mica. Be sure to ground the rotor of C1, rather than the stator, to avoid hand capacitance. L1, L2 and L3 are each close-wound with #22 enameled wire on a single ferrite rod, 4 inch long and $^{1}/_{2}$ inch diameter, with μ = 125 (Amidon R61-50-400). Windings are spaced approximately $^{1}/_{4}$ inch apart.

C1—Air variable, 6-140 pF; Hammarlund HF140 or equiv.
D1—Germanium diode; 1N34A, 1N270 or equiv.
L1—1.6-5 MHz; 30 turns, tapped at 3 turns from grounded end.
L2—5-20 MHz; 8 turns, tapped at 2 turns from grounded end.
L3—17-39 MHz; 2 turns, tapped at 1 turn.
M1—Any microammeter may be used. The one pictured is a Micronta meter, RadioShack no. 270-1751.
R1—Linear taper.
RFC1—1 mH; Miller no. 4642 or equiv. Value is not critical.
S1—Ceramic rotary switch, 1 section, 2 poles, 2 to 6 positions; Centralab PA2002 or PA2003 or equiv.

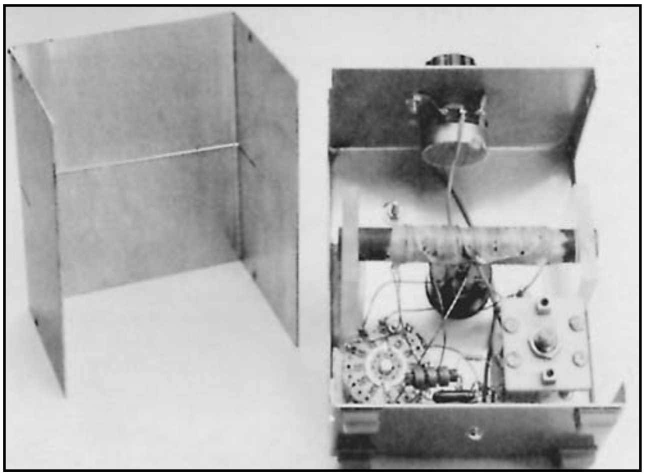

Fig 59—The current probe just before final assembly. Note that all parts except the ferrite rod are mounted on a single half of the 3 × 4 × 5-inch Minibox (Bud CU-2105B or equiv.). Rubber grommets are fitted in holes at the ends of the slot to accept the rod during assembly of the enclosure. Leads in the RF section should be kept as short as possible, although those from the rod windings must necessarily be left somewhat long to facilitate final assembly.

As shown in the photos, the circuit is constructed in a metal box. This enclosure shields the detector circuit from the electric field of the radio wave. A slot must be cut with a hacksaw across the back of the box, and a thin file may be used to smooth the cut. This slot is necessary to prevent the box from acting as a shorted turn.

Using the Probe

In measuring the current in a conductor, the ferrite rod should be kept at right angles to the conductor, and at a constant distance from it. In its upright or vertical position, this instrument is oriented for taking measurements in vertical conductors. It must be laid horizontal to measure current in horizontal conductors.

Numerous uses for the instrument are suggested in an earlier paragraph. In addition, the probe is an ideal instrument for checking the current distribution in antenna elements. It is also useful for measuring RF ground currents in radial systems. A buried radial may be located easily by sweeping the ground. Current division at junctions may be investigated. *Hot spots* usually indicate areas where additional radials would be effective.

Stray currents in conductors not intended to be part of the antenna system may often be eliminated by bonding or by changing the physical lengths involved. Guy wires and other unwanted parasitic elements will often give a tilt to the plane of polarization and can make a marked difference in front-to-back ratios. When the ferrite rod is oriented parallel to the electric field lines, there will be a sharp null reading that may be used to locate the plane of polarization quite accurately. When using the meter, remember that the magnetic field is at right angles to the electric field.

You may also use the current probe as a relative signal strength meter. When making measurements on a vertical antenna, locate the meter at least two wavelengths away, with the rod in a horizontal position. For horizontal antennas, hold the instrument at approximately the same height as the antenna, with the rod vertical.

Antenna Measurements

Of all the measurements made in Amateur Radio systems, perhaps the most difficult and least understood are various measurements of antennas. For example, it is relatively easy to measure the frequency and CW power output of a transmitter, the response of a filter, or the gain of an amplifier. These are all what might be called *bench measurements* because, when performed properly, all the factors that influence the accuracy and success of the measurement are under control. In making antenna measurements, however, the “bench” is probably your backyard. In other words, the environment surrounding the antenna can affect the results of the measurement.

Control of the environment is not at all as simple as it was for the bench measurement, because now the work area may be rather spacious. This section describes antenna measurement techniques that are closely allied to those used in an antenna measuring event or contest. With these procedures you can make measurements successfully and with meaningful results. These techniques should provide a better understanding of the measurement problems, resulting in a more accurate and less difficult task. The information in this section was provided by Dick Turrin, W2IMU, and was originally published in November 1974 *QST*.

SOME BASIC IDEAS

An antenna is simply a transducer or coupler between a suitable feed line and the environment surrounding it. In addition to the efficient transfer of power from feed line to environment, an antenna at VHF or UHF is most frequently required to concentrate the radiated power into a particular region of the environment.

To be consistent while comparing different antennas, you must standardize the environment surrounding the antenna. Ideally, you want to make measurements with the measured antenna so far removed from any objects causing environmental effects that it is literally in outer space—a very impractical situation. The purpose of the measurement techniques is therefore to simulate, under practical conditions, a *controlled environment*. At VHF and UHF, and with practical-size antennas, the environment can be controlled so that successful and accurate measurements can be made in a reasonable amount of space.

The electrical characteristics of an antenna that are most desirable to obtain by direct measurement are: (1) gain (relative to an isotropic source, which by definition has a gain of unity); (2) space-radiation pattern; (3) feed-point impedance (mismatch) and (4) polarization.

Polarization

In general the polarization can be assumed from the geometry of the radiating elements. That is to say, if the antenna consists of a number of linear elements (straight lengths of rod or wire that are resonant and connected to the feed point) the polarization of the electric field will be linear and polarized parallel to the elements. If the elements are not consistently parallel with each other, then the polarization cannot easily be assumed. The following techniques are directed to antennas having polarization that is essentially linear (in one plane), although the method can be extended to include all forms of elliptic (or mixed) polarization.

Feed-Point Mismatch

The feed-point mismatch, although affected to some degree by the immediate environment of the antenna, does not affect the gain or radiation characteristics of an antenna. If the immediate environment of the antenna does not affect the feed-point impedance, then any mismatch intrinsic to the antenna tuning reflects a portion of the incident power back to the source. In a receiving antenna this reflected power is reradiated back into the environment, and can be lost entirely.

In a transmitting antenna, the reflected power travels back down the feed line to the transmitter, where it changes the load impedance presented to that transmitter. The amplifier output controls are customarily altered during the normal tuning procedure to obtain maximum power transfer to the antenna. You can still use a mismatched antenna to its full gain potential, provided the mismatch is not so severe as to cause heating losses in the system, especially the feed line and matching devices. (See also the discussion of additional loss caused by SWR in Chapter 24.)

Similarly, a mismatched receiving antenna may be matched into the receiver front end for maximum power transfer. In any case you should clearly keep in mind that the feed-point mismatch does not affect the radiation characteristics of an antenna. It can only affect the system efficiency when heating losses are considered.

Why then do we include feed-point mismatch as part of the antenna characteristics? The reason is that for efficient system performance, most antennas are resonant transducers and present a reasonable match over a relatively narrow frequency range. It is therefore desirable to design an antenna, whether it be a simple dipole or an array of Yagis, such that the final single feed-point impedance is essentially resistive and matched to the feed line. Furthermore, in order to make accurate, absolute gain measurements, it is vital that the antenna under test accept all the power from a matched-source generator, or that the reflected power caused by the mismatch be measured and a suitable error correction for heating losses be included in the gain calculations. Heating losses may be determined from information contained in Chapter 24.

While on the subject of feed-point impedance, mention should be made of the use of *baluns* in antennas. A balun is simply a device that permits a lossless transition between a balanced system feed line or antenna and an unbalanced feed line or system. If the feed point of an antenna is symmetric, such as with a dipole, and it is desired to feed this antenna with an unbalanced feed line such as coax, you should provide a balun between the line and the feed point. Without the balun, current will be allowed to flow on the outside of the coax. The current on the outside of the feed line will cause radiation, and thus the feed line will become part of the antenna radiation system. In the case of beam antennas, where it is desired to concentrate the radiated energy is a specific direction, this extra radiation from the feed line will be detrimental, causing distortion of the expected antenna pattern. See Chapter 26 for additional details on this problem.

ANTENNA TEST SITE SET-UP AND EVALUATION

Since an antenna is a reciprocal device, measurements of gain and radiation patterns can be made with the test antenna used either as a transmitting or as a receiving antenna. In general and for practical reasons, the test antenna is used in the receiving mode, and the source or transmitting antenna is located at a specified fixed remote site and unattended. In other words the source antenna, energized by a suitable transmitter, is simply required to illuminate or flood the receiving site in a controlled and constant manner.

As mentioned earlier, antenna measurements ideally should be made under free-space conditions. A further restriction is that the illumination from the source antenna be a *plane wave* over the effective aperture (capture area) of the test antenna. A plane wave by definition is one in which the magnitude and phase of the fields are uniform, and in the test-antenna situation, *uniform over the effective area plane of the test antenna*. Since it is the nature of all radiation to expand in a spherical manner at great distance from the source, it would seem to be most desirable to locate the source antenna as far from the test site as possible. However, since for practical reasons the test site and source location will have to be near the earth and not in outer space, the environment must include the effects of the ground surface and other obstacles in the vicinity of both antennas. These effects almost always dictate that the test range (spacing between source and test antennas) be as short as possible consistent with maintaining a nearly error-free plane wave illuminating the test *aperture*.

A nearly error-free plane wave can be specified as one in which the phase and amplitude, from center to edge of the illuminating field over the test aperture, do not deviate by more than about 30° and 1 decibel, respectively. These conditions will result in a gain-measurement error of no more than a few percent less than the true gain. Based on the 30º phase error alone, it can be shown that the minimum range distance is approximately

$$S_{min} = 2\frac{D^2}{\lambda} \qquad \text{(Eq 23)}$$

where D is the largest aperture dimension and λ is the free-space wavelength in the same units as D. The phase error over the aperture D for this condition is $^{1}/_{16}$ wavelength.

Since aperture size and gain are related by

$$\text{Gain} = \frac{4\pi A_e}{\lambda^2} \qquad \text{(Eq 24)}$$

where A_e is the effective aperture area, the dimension D may be obtained for simple aperture configurations. For a square aperture

$$D_2 = G\frac{\lambda^2}{4\pi} \qquad \text{(Eq 25)}$$

that results in a minimum range distance for a square aperture of

$$S_{min} = G\frac{\lambda}{2\pi} \qquad \text{(Eq 26)}$$

and for a circular aperture of

$$S_{min} = G\frac{2\lambda}{\pi^2} \qquad \text{(Eq 27)}$$

For apertures with a physical area that is not well defined or is much larger in one dimension that in other directions, such as a long thin array for maximum directivity in one plane, it is advisable to use the maximum estimate of D from either the expected gain or physical aperture dimensions.

Up to this point in the range development, only the conditions for minimum range length, S_{min}, have been established, as though the ground surface were not present. This minimum S is therefore a necessary condition even under free-space environment. The presence of the ground further complicates the range selection, not in the determination of S but in the exact location of the source and test antennas above the earth.

It is always advisable to select a range whose intervening terrain is essentially flat, clear of obstructions, and of uniform surface conditions, such as all grass or all pavement. The extent of the range is determined by the illumination of the source antenna, usually a Yagi, whose gain is no greater than the highest gain antenna to be measured. For gain measurements the range consists essentially of the region in the beam of the test antenna. For radiation-pattern measurements, the range is considerably larger and consists of all that area illuminated by the source antenna, especially around and behind the test site. Ideally you should choose a site where the test-antenna location is near the center of a large open area and the source antenna is located near the edge where most of the obstacles (trees, poles, fences, etc.) lie.

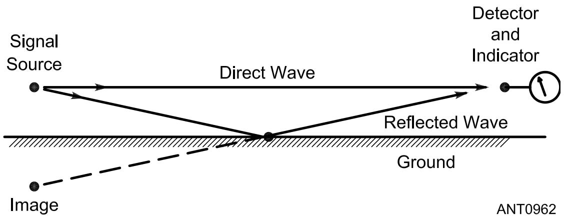

Fig 60—On an antenna test range, energy reaching the receiving equipment may arrive after being reflected from the surface of the ground, as well as by the direct path. The two waves may tend to cancel each other, or may reinforce one another, depending on their phase relationship at the receiving point.

The primary effect of the range surface is that some of the energy from the source antenna will be reflected into the test antenna, while other energy will arrive on a direct line-of-sight path. This is illustrated in **Fig 60**. The use of a flat, uniform ground surface assures that there will be essentially a mirror reflection, even though the reflected energy may be slightly weakened (absorbed) by the surface material (ground). In order to perform an analysis you should realize that horizontally polarized waves undergo a 180º phase reversal upon reflection from the earth. The resulting illumination amplitude at any point in the test aperture is the vector sum of the electric fields arriving from the two directions, the direct path and the reflected path.

If a perfect mirror reflection is assumed from the ground (it is nearly that for practical ground conditions at VHF/UHF) and the source antenna is isotropic, radiating equally in all directions, then a simple geometric analysis of the two path lengths will show that at various point in the vertical plane at the test-antenna site the waves will combine in different phase relationships. At some points the arriving waves will be in phase, and at other points they will be 180º out of phase. Since the field amplitudes are nearly equal, the resulting phase change caused by path length difference will produce an ampli-

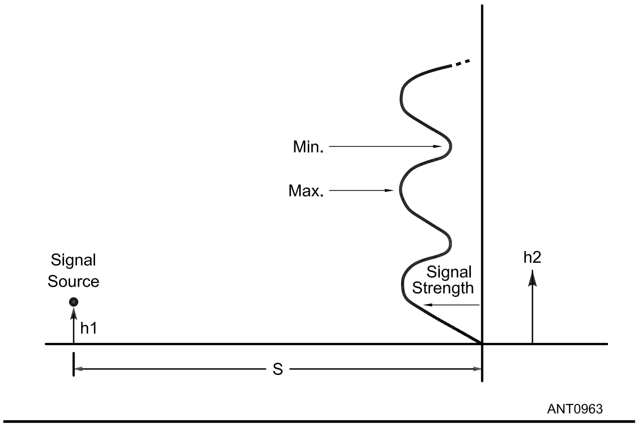

Fig 61—The vertical profile, or plot of signal strength versus test-antenna height, for a fixed height of the signal source above ground and at a fixed distance. See text for definitions of symbols.

tude variation in the vertical test site direction similar to a standing wave, as shown in **Fig 61**.

The simplified formula relating the location of h2 for maximum and minimum values of the two-path summation in terms of h1 and S is

$$h2 = n \frac{\lambda}{4} \times \frac{S}{h1} \qquad \text{(Eq 28)}$$

with n = 0, 2, 4, . . . for minimums and n = 1, 3, 5, . . . for maximums, and S is much larger than either h1 or h2.

The significance of this simple ground reflection formula is that it permits you to determine the approximate location of the source antenna to achieve a nearly plane-wave amplitude distribution *in the vertical direction* over a particular test *aperture size*. It should be clear from examination of the height formula that as h1 is decreased, the vertical distribution pattern of signal at the test site, h2, expands. Also note that the signal level for h2 equal to zero is always zero on the ground regardless of the height of h1.

The objective in using the height formula then is, given an effective antenna aperture to be illuminated from which a minimum S (range length) is determined and a suitable range site chosen, to find a value for h1 (source antenna height). The required value is such that the *first* maximum of vertical distribution at the test site, h2, is at a practical distance above the ground, and at the same time the signal amplitude over the aperture in the vertical direction does not vary more than about 1 dB. This last condition is not sacred but is closely related to the particular antenna under test.

In practice these formulas are useful only to initialize the range setup. A final check of the vertical distribution at the test site must be made by direct measurement. This measurement should be conducted with a small low-gain but unidirectional probe antenna such as a corner reflector or 2-element Yagi that you move along a vertical line over the intended aperture site. Care should be exercised to minimize the effects of local environment around the probe antenna and that the beam of the probe be directed at the source antenna at all times for maximum signal. A simple dipole is undesirable as a probe antenna because it is susceptible to local environmental effects.

The most practical way to instrument the vertical distribution measurement is to construct some kind of vertical track, preferably of wood, with a sliding carriage or platform that may be used to support and move the probe antenna. It is assumed of course that a stable source transmitter and calibrated receiver or detector are available so variations of the order of $^1/_2$ dB can be clearly distinguished.

Once you conduct these initial range measurements successfully, the range is now ready to accommodate any aperture size less in vertical extent than the largest for which S_{min} and the vertical field distribution were selected. Place the test antenna with the center of its aperture at the height h2 where maximum signal was found. Tilt the test antenna tilted so that its main beam is pointed in the direction of the source antenna. The final tilt is found by observing the receiver output for maximum signal. This last process must be done empirically since the apparent location of the source is somewhere between the actual source and its image, below the ground.

An example will illustrate the procedure. Assume that we wish to measure a 7-foot diameter parabolic reflector antenna at 1296 MHz (λ = 0.75 foot). The minimum range distance, S_{min}, can be readily computed from the formula for a circular aperture.

$$S_{min} = 2\frac{D^2}{\lambda} = 2 \times \frac{49}{0.75} = 131 \text{ feet}$$

Now a suitable site is selected based on the qualitative discussion given before.

Next determine the source height, h1. The procedure is to choose a height h1 such that the first minimum above ground (n = 2 in formula) is at least two or three times the aperture size, or about 20 feet.

$$h1 = n\frac{\lambda}{4}\frac{S}{h1} = 2 \times \frac{0.75}{4} \times \frac{131}{20} = 2.5 \text{ feet}$$

Place the source antenna at this height and probe the vertical distribution over the 7-foot aperture location, which will be about 10 feet off the ground.

$$h2 = n\frac{\lambda}{4}\frac{S}{h1} = 1 \times \frac{0.75}{4} \times \frac{131}{2.5} = 9.8 \text{ feet}$$

Plot the measured profile of vertical signal level versus height. From this plot, empirically determine whether the 7-foot aperture can be fitted in this profile such that the 1-dB variation is not exceeded. If the variation exceeds

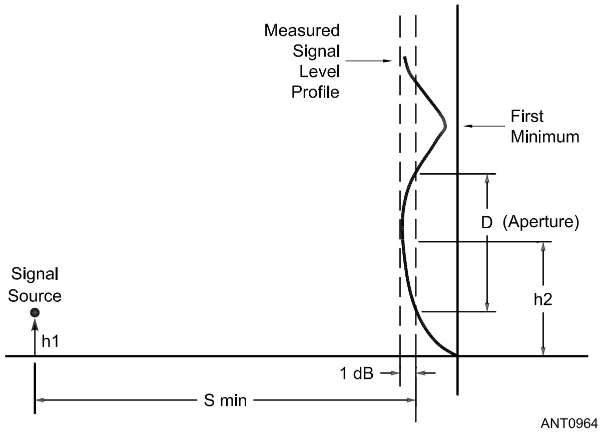

Fig 62—Sample plot of a measured vertical profile.

1 dB over the 7-foot aperture, the source antenna should be lowered and h2 raised. Small changes in h1 can quickly alter the distribution at the test site. **Fig 62** illustrates the points of the previous discussion.

The same set-up procedure applies for either horizontal or vertical linear polarization. However, it is advisable to check by direct measurement at the site for each polarization to be sure that the vertical distribution is satisfactory. Distribution probing in the horizontal plane is unnecessary as little or no variation in amplitude should be found, since the reflection geometry is constant. Because of this, antennas with apertures that are long and thin, such as a stacked collinear vertical, should be measured with the long dimension parallel to the ground.

A particularly difficult range problem occurs in measurements of antennas that have depth as well as cross-sectional aperture area. Long end-fire antennas such as long Yagis, rhombics, V-beams, or arrays of these antennas, radiate as *volumetric arrays* and it is therefore even more essential that the illuminating field from the source antenna be reasonably uniform in depth as well as plane wave in cross section. For measuring these types of antennas it is advisable to make several vertical profile measurements that cover the depth of the array. A qualitative check on the integrity of the illumination for long end-fire antennas can be made by moving the array or antenna axially (forward and backward) and noting the change in received signal level. If the signal level varies less than 1 or 2 dB for an axial movement of several wavelengths then the field can be considered satisfactory for most demands on accuracy. Large variations indicate that the illuminating field is badly distorted over the array depth and subsequent measurements are questionable. It is interesting to note in connection with gain measurements that any illuminating field distortion will always result in measurements that are lower than true values.

ABSOLUTE GAIN MEASUREMENT

Having established a suitable range, the measure-

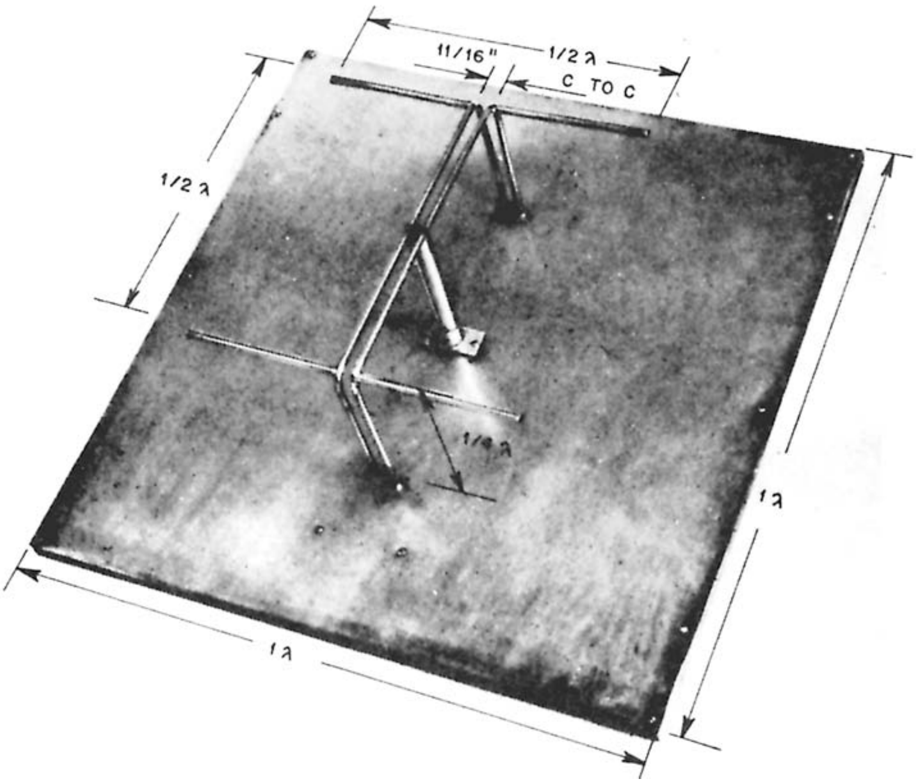

Fig 63—Standard-gain antenna. When accurately constructed for the desired frequency, this antenna will exhibit a gain of 7.7 dB over a dipole radiator, plus or minus 0.25 dB. In this model, constructed for 432 MHz, the elements are $^3/_8$-inch diameter tubing. The phasing and support lines are of $^5/_{16}$-inch diameter tubing or rod.

ment of gain relative to an isotropic (point source) radiator is almost always accomplished by direct comparison with a calibrated standard-gain antenna. That is, the signal level with the test antenna in its optimum location is noted. Then you remove the test antenna and place the standard-gain antenna with its aperture at the center of location where the test antenna was located. Measure the difference in signal level between the standard and the test antennas and add to or subtract from the gain of the standard-gain antenna to obtain the absolute gain of the test antenna. Here, *absolute* means with respect to a point source with a gain of unity, by definition. The reason for using this reference rather than a dipole, for instance, is that it is more useful and convenient for system engineering. We assume that both standard and test antennas have been carefully matched to the appropriate impedance and an accurately calibrated and matched detecting device is being used.

A standard-gain antenna may be any type of unidirectional, preferably planar-aperture, antenna, which has been calibrated either by direct measurement or in special cases by accurate construction according to computed dimensions. A standard-gain antenna has been suggested by Richard F. H. Yang (see Bibliography). Shown in **Fig 63**, it consists of two in-phase dipoles $^1/_2$ λ apart and backed up with a ground plane 1 λ square.

In Yang's original design, the stub at the center is a balun formed by cutting two longitudinal slots of $^1/_8$-inch width, diametrically opposite, on a $^1/_4$-λ section of $^7/_8$-inch rigid 50-Ω coax. An alternative method of feeding is to feed RG-8 or RG-213 coax through slotted $^7/_8$-inch copper tubing. Be sure to leave the outer jacket on the coax to insulate it from the copper-tubing balun section. When constructed accurately to scale for the frequency of interest, this type of standard will have an absolute gain of 9.85 dBi (7.7 dBd gain over a dipole in free space) with an accuracy of ± 0.25 dB.

RADIATION-PATTERN MEASUREMENTS

Of all antenna measurements, the radiation pattern is the most demanding in measurement and the most difficult to interpret. Any antenna radiates to some degree in all directions into the space surrounding it. Therefore, the radiation pattern of an antenna is a three-dimensional representation of the magnitude, phase and polarization. In general, and in practical cases for Amateur Radio communications, the polarization is well defined and only the magnitude of radiation is important.

Furthermore, in many of these cases the radiation in one particular plane is of primary interest, usually the plane corresponding to that of the Earth's surface, regardless of polarization. Because of the nature of the range setup, measurement of radiation pattern can be successfully made only in a plane nearly parallel to the earth's surface. With beam antennas it is advisable and usually sufficient to take two radiation pattern measurements, one in the polarization plane and one at right angles to the plane of polarization. These radiation patterns are referred to in antenna literature as the principal E-plane and H-plane patterns, respectively. *E-plane* means parallel to the electric field that is the polarization plane and *H-plane* means parallel to the magnetic field in free space. The electric field and magnetic field are always perpendicular to each other in a plane wave as it propagates through space.

When the antenna is located over real earth, the terms *Azimuth* and *elevation* planes are commonly used, since the frame of reference is the Earth itself, rather than the electric and magnetic fields in free space. For a horizontally polarized antenna such as a Yagi mounted with its elements parallel to the ground, the azimuth plane is the E-plane and the elevation plane is the H-plane.

The technique to obtain these patterns is simple in procedure but requires more equipment and patience than does making a gain measurement. First, a suitable mount is required that can be rotated in the azimuth plane (horizontal) with some degree of accuracy in terms of azimuth-angle positioning. Second, a signal-level indicator calibrated over at least a 20-dB dynamic range with a readout resolution of at least 2 dB is required. A dynamic range of up to about 40 dB would be desirable but does not add greatly to the measurement significance.

With this much equipment, the procedure is to locate first the area of maximum radiation from the beam antenna by carefully adjusting the azimuth and elevation positioning. These settings are then arbitrarily assigned an azimuth angle of zero degrees and a signal level of zero decibels. Next, without changing the elevation setting (tilt of the rotating axis), the antenna is carefully rotated in azimuth in small steps that permit signal-level readout of 2 or 3 dB per step. These points of signal level corresponding with an azimuth angle are recorded and plotted on polar coordinate paper. A sample of the results is shown on ARRL coordinate paper in **Fig 64**.

On the sample radiation pattern the measured points are marked with an X and a continuous line is drawn in, since the pattern is a continuous curve. Radiation patterns should preferably be plotted on a logarithmic radial scale, rather than a voltage or power scale. The reason is that the log scale approximates the response of the ear to signals in the audio range. Also many receivers have AGC systems that are somewhat logarithmic in response; therefore the log scale is more representative of actual system operation.

Having completed a set of radiation-pattern measurements, one is prompted to ask, "Of what use are they?" The primary answer is as a diagnostic tool to determine if the antenna is functioning as it was intended to. A second answer is to know how the antenna will discriminate against interfering signals from various directions.

Consider now the diagnostic use of the radiation patterns. If the radiation beam is well defined, then there is an approximate formula relating the antenna gain to the

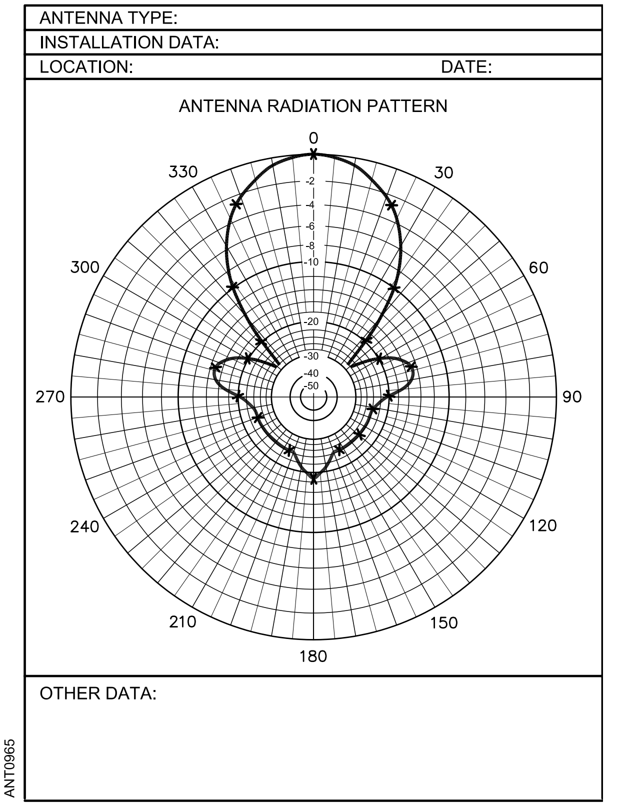

For more information see QST, July 1980, p.26 Copyright 1980, ARRL Inc.

Fig 64—Sample plot of a measured radiation pattern, using techniques described in the text. The plot is on coordinate paper available from ARRL HQ. The form provides space for recording significant data and remarks.

measured half-power beamwidth of the E- and H-plane radiation patterns. The half-power beamwidth is indicated on the polar plot where the radiation level falls to 3 dB below the main beam 0-dB reference on either side. The formula is

$$\text{Gain}\left(\text{dBi}\right) \cong \frac{41{,}253}{\theta_E \phi_H} \qquad \text{(Eq 29)}$$

where θ_E and ϕ_H are the half-power beamwidths in degrees of the E- and H-plane patterns, respectively. This equation assumes a lossless antenna system, where any side-lobes are well suppressed.

To illustrate the use of this equation, assume that we have a Yagi antenna with a boom length of two wavelengths. From known relations (described in Chapter 11) the expected free-space gain of a Yagi with a boom length of 2 λ is about 13 dBi; its gain, G, equals 20. Using the above relationship, the product of $\theta_E \times \phi_H \approx 2062$ square degrees. Since a Yagi produces a nearly symmetric beam shape in cross section, $\theta_E = \phi_H = 45°$. Now if the measured values of θ_E and ϕ_H are much larger than 45°, then the gain will be much lower than the expected 13 dBi.

As another example, suppose that the same antenna (a 2-wavelength-boom Yagi) gives a measured gain of 9 dBi but the radiation pattern half power beamwidths are approximately 45°. This situation indicates that although the radiation patterns seem to be correct, the low gain shows inefficiency somewhere in the antenna, such as lossy materials or poor connections.

Large broadside collinear antennas can be checked for excessive phasing-line losses by comparing the gain computed from the radiation patterns with the direct-measured gain. It seems paradoxical, but it is indeed possible to build a large array with a very narrow beamwidth indicating high gain, but actually having very low gain because of losses in the feed distribution system.

In general, and for most VHF/UHF Amateur Radio communications, gain is the primary attribute of an antenna. However, radiation in other directions than the main beam, called *sidelobe radiation*, should be examined by measurement of radiation patterns for effects such as nonsymmetry on either side of the main beam or excessive magnitude of sidelobes. (Any sidelobe that is less than 10 dB below the main beam reference level of 0 dB should be considered excessive.) These effects are usually attributable to incorrect phasing of the radiating elements or radiation from other parts of the antenna that was not intended, such as the support structure or feed line.

The interpretation of radiation patterns is intimately related to the particular type of antenna under measurement. Reference data should be consulted for the antenna type of interest, to verify that the measured results are in agreement with expected results.

To summarize the use of pattern measurements, if a beam antenna is first checked for gain (the easier measurement to make) and it is as expected, then pattern measurements may be academic. However, if the gain is lower than expected it is advisable to make pattern measurements to help determine the possible causes for low gain.

Regarding radiation pattern measurements, remember that the results measured under proper range facilities will not necessarily be the same as observed for the same antenna at a home-station installation. The reasons may be obvious now in view of the preceding information on the range setup, ground reflections, and the vertical-field distribution profiles. For long paths over rough terrain where many large obstacles may exist, the effects of ground reflection tend to become diffused, although they still can cause unexpected results. For these reasons it is usually unjust to compare VHF/UHF antennas over long paths.

Vector Network Analyzers

The process of building and properly tuning a phased array often involves making a number of different measurements to achieve a desired level of performance, as was pointed out in Chapter 8, Multielement Arrays. This section was written by Rudy Severns, N6LF.

After erecting an array we would like to measure the resonant frequency of each element, the self-impedances of each element and the mutual impedances between the elements. We will also want to know these impedances over the whole operating band to help design a feed network. When building the feed network, we may need to check the values and Qs of the network elements and we will want to determine the electrical lengths of transmission lines.

Final tuning of the array requires that the relative current amplitudes and phases in each element be measured and adjusted, if necessary. We also will want to determine the SWR at the feed point. Doing all of this even moderately well can require quite a bit of equipment, some of which is heavy and requires ac line power. This can be a nuisance in the field, especially if the weather is not cooperating.

Professionals make these measurements by employing a *vector network analyzer* (VNA) or the somewhat simpler, *reflection-transmission test set*. These instruments can make all the necessary measurements quickly and with great accuracy. However, in the past VNAs have been very expensive, out of reach for general amateur use. But thanks to modern digital technology VNAs that work with a laptop computer are now becoming available at prices an amateur might consider. It's even possible to homebrew a VNA[1] with performance that approaches a professional instrument. Considering the cost of even a simple array, investment in a VNA makes sense.

VNAs are based on reflection and transmission measurements. To use a VNA it is very helpful to have a basic understanding of *Scattering Parameters* (S-parameters). Microwave engineers have long used these because they have to work with circuits that are large in terms of wavelength, where measurements of forward and reflected power are easy.

HF arrays are also large in terms of wavelength. The techniques for measuring forward and reverse powers work well even at 160 meters. For example, even though the array elements may be 100 feet apart, you can place your instruments in a central location and run cables out to each element. The effect of the cables from the VNA to the elements can be absorbed in the initial calibration procedure so the measurements read out at the VNA are effectively those at each element. In other words, the measurement reference points can be placed electrically at the base of the element, regardless of the physical location of the instrumentation and the interconnecting cables.

In a completed array with its feed network, the network can be excited by the VNA at the feed point and the relative current amplitudes and phases at each element can be measured over a frequency band. Then, adjustments can be made as needed. When the final values for the current amplitudes and phases are known, these values can be put back into an array model in a program like *EZNEC* to determine the pattern of the array across the whole frequency band.

S-PARAMETERS

In Chapter 24, Transmission Lines, the reflection coefficient rho (ρ) is defined as the ratio of the reflected voltage (V_r) to the incident voltage (V_i):

$$\rho = \frac{V_r}{V_i} \qquad \text{(Eq 30)}$$

If we know the load impedance (Z_L) and the transmission line impedance (Z_0) we can calculate ρ from:

$$\rho = \frac{Z_L - Z_0}{Z_L + Z_0} \qquad \text{(Eq 31)}$$

Keep in mind that ρ is a complex number (a vector), which we represent by either amplitude and phase (|Z|, θ) or by real and imaginary parts (R ± *j* X). The two representations are equivalent. From ρ we can then calculate SWR. That's very handy, but here we want to do something different. If we have an instrument that measures ρ and we know Z_0 then we can determine Z_L from:

$$Z_L = Z_0\left(\frac{1+\rho}{1-\rho}\right) \qquad \text{(Eq 32)}$$

Measuring ρ is one of the things that VNAs do very well. With a VNA, the measurement can be made at one end of a long transmission line with the load at the other end. The effect of the line can be calibrated out, as mentioned above, so that we are in effect measuring right at the load.

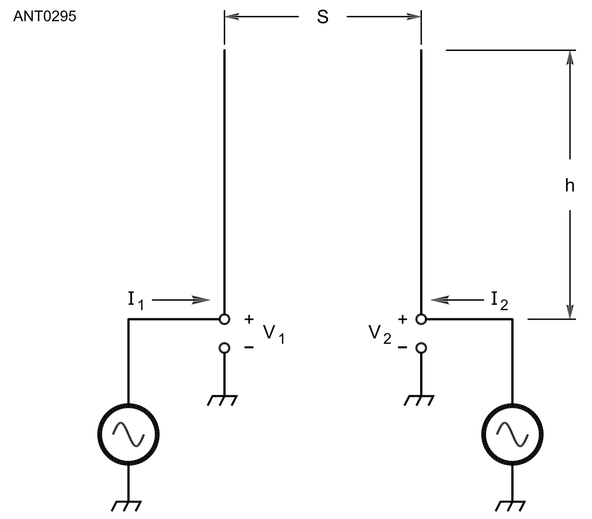

Fig 65—A 2-element array, where h is the element height and S is the spacing between the elements.

This approach can be used directly to measure the impedance and resonant frequency of a single element. By open and short circuiting elements in an array we can determine the mutual as well as self impedances for, and between, all the elements. We can also use this approach to measure component values, inductor Qs, etc.

This is an example of a *one-port* measurement; that is, a load at the end of a transmission line. However, a multi-element array actually behaves as a multi-port network, so to get the most out of a VNA, you need to generalize the above procedure. This is where S-parameters come into play. To illustrate the principles we will use a simple 2-element array like that shown in **Fig 65**.

To design a feed network to drive this array we need to know the input impedance of each element (Z_1 and Z_2) as a function of the drive currents (I_1 and I_2). The input impedances will depend on the self impedance of each element, the coupling between them (the mutual impedance) and the drive currents in each element. To manage this problem we can represent a 2-element array as a two-port network, as shown in **Fig 66**. And we can relate the port voltages, currents and impedances with Eq 33:

$$\begin{aligned} V_1 &= Z_{11}I_1 + Z_{12}I_2 \\ V_2 &= Z_{21}I_1 + Z_{22}I_2 \end{aligned} \qquad \text{(Eq 33)}$$

Normally we know I_1 and I_2 from the design of the array, but we need to determine the resulting element impedances. That's the challenge. Fortunately, an array is a linear network, so $Z_{12} = Z_{21}$, which means we need only determine three variables: the self impedances Z_{11} and Z_{22} and the mutual impedance, Z_{12}.

Once we know Z_{11}, $Z_{12,}$ Z_{22} and are given I_1 and I_2, we can determine the feed-point impedances at each element from:

$$\begin{aligned} Z_1 &= Z_{11} + \left(\frac{I_2}{I_1}\right) Z_{12} \\ Z_2 &= Z_{22} + \left(\frac{I_1}{I_2}\right) Z_{12} \end{aligned} \qquad \text{(Eq 34)}$$

This is the conventional approach. However, there are some problems here. We have to be able to accurately measure either voltages and currents or impedances in multiple elements that may be separated by large fractions of a wavelength. In addition, accurate measurements of current, voltage and impedance become increasingly more difficult as we go up in frequency.

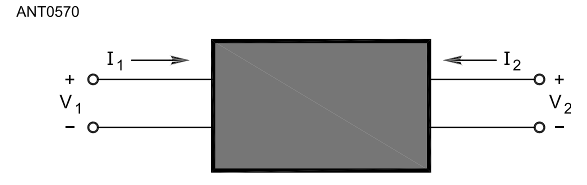

Fig 66—Two-port representation of currents and voltages in the 2-element array in Fig 65.

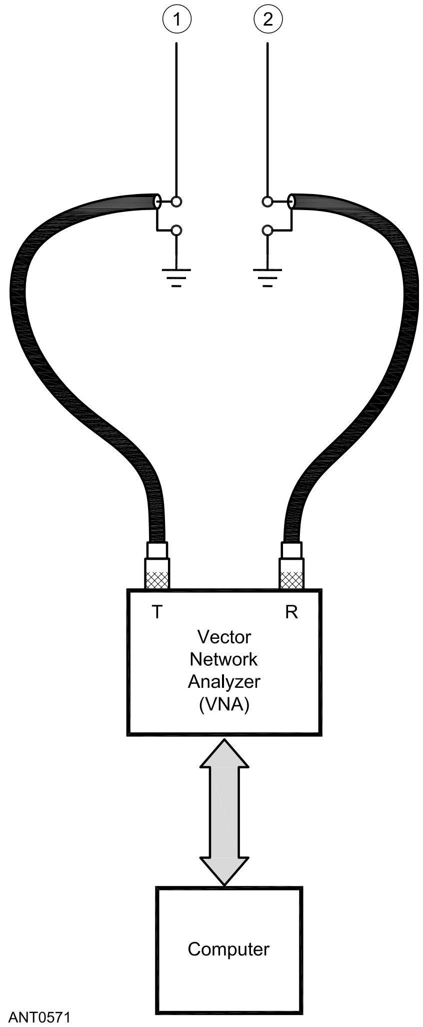

Fig 67—Test setup to measure a 2-element array using a VNA.

It turns out that we can get the information more easily by measuring incident and reflected voltages at the ports and from those measurements determine the feed-point impedances. A VNA is an instrument for measuring these voltages. It turns out to be easier to measure the ratios of two voltages rather than their absolute values.

The measurement setup using a VNA for a 2-element array is shown in **Fig 67**. VNAs usually have at least two RF connections: the transmit port (T) and the receive port (R). Professional units may have more RF connections. The T output provides an signal from a 50-Ω source and the R port is a detector with a 50-Ω input impedance. Basically we have a transmitter and a receiver. The transmit port uses a directional coupler to provide measurements of the forward and reflected signals at that output. The receive port measures the signal transmitted through the network. Transmission lines usually have Z_0 = 50 Ω and may be of any length required by the size of the array.

Using incident and reflected voltages, the two-port network representation is now changed, as shown in **Fig 68**, where:

V_{1i} = incident voltage at port 1
V_{1r} = reflected voltage at port 1
V_{2i} = incident voltage at port 2
V_{2r} = reflected voltage at port 2

In a manner analogous to Eq 33, we can write an expression in terms of the incident and reflected voltages:

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_2 \end{aligned} \qquad \text{(Eq 35)}$$

where:

$$\begin{aligned} &a_1 \frac{V_{1i}}{\sqrt{Z_0}} \qquad b_1 \frac{V_{1r}}{\sqrt{Z_0}} \\ &a_2 = \frac{V_{2i}}{\sqrt{Z_0}} \qquad b_2 = \frac{V_{2r}}{\sqrt{Z_0}} \end{aligned} \qquad \text{(Eq 36)}$$

We see that the a_n and b_n are simply the incident and reflected voltages at the two ports divided by $\sqrt{Z_o}$. Because this is a linear network, $S_{21} = S_{12}$.

What are the S_{ij} quantities? These are called the *S-parameters*, which are defined by:

$$\begin{aligned} S_{11} &\equiv \left.\frac{b_1}{a_1}\right|_{a_2=0} = \left.\frac{V_{1r}}{V_{1i}}\right|_{V_{2i}=0} \\ S_{21} &\equiv \left.\frac{b_2}{a_1}\right|_{a_2=0} = \left.\frac{V_{2r}}{V_{1i}}\right|_{V_{2i}=0} \\ S_{12} &\equiv \left.\frac{b_1}{a_2}\right|_{a_1=0} = \left.\frac{V_{1r}}{V_{2i}}\right|_{V_{1i}=0} \\ S_{22} &\equiv \left.\frac{b_2}{a_2}\right|_{a_1=0} = \left.\frac{V_{2r}}{V_{2i}}\right|_{V_{1i}=0} \end{aligned} \qquad \text{(Eq 37)}$$

Note that the S_{ij} parameters are all ratios of reflected and incident voltages, and they are usually complex numbers. The condition that $a_2 = 0 = V_{2i}$ is the same as saying that port 2 is terminated in a load equal to Z_0 and the network is excited at port 1. This means there is no reflection from the load on port 2, which makes $V_{2i} = 0$. Similarly, if we terminate port 1 with Z_0 and excite port 2, then $V_{1i} = 0 = a_1$.

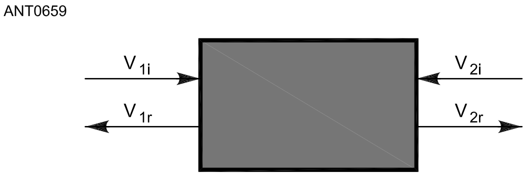

Fig 68—Two-port network with incident and reflected waves.

If we compare Eq 30 to the first line of Eq 37 we see that $S_{11} = \rho_1$, the reflection coefficient at port 1. We can now restate Eq 32 in terms of S_{11}:

$$Z = Z_0\left(\frac{1+S_{11}}{1-S_{11}}\right) \qquad \text{(Eq 38)}$$

where Z is the impedance looking into port 1 with port 2 terminated in Z_0. In the case where port 2 does not exist—that is, you are measuring a single element (for example, measuring element 1 with element 2 open-circuited) or a component, then Z is simply the self impedance (Z_{11} in Eq 33). Since S_{11} is a standard measurement for VNAs you can calculate Z using Eq 38. In many cases the VNA software will do this calculation for you automatically. You can also measure element 2 with element 1 open and determine Z_{22}.

S_{21} represents the ratio of the signal coming out of port 2 (V_{2r}) to the input signal on port 1 (V_{1i}) and is another standard VNA measurement. S_{21} is a measurement of the signal transmission between the ports through the network, or in the case of an array, the signal transmission due to the coupling between the elements. Again, port 2 is terminated in Z_0.

A full-feature VNA will measure all the S_{ij} parameters at once, but most of the lower-cost units of interest to amateurs are what we call *reflection-transmission test sets*. What this means is that they only measure S_{11} and S_{21}. To obtain S_{22} and S_{12} we have to interchange the test cables at the array elements (see Fig 67) and run the measurements again. Normally the software will accommodate this as a second entry and we end up with the full set of S_{ij} parameters.

If we do run a full set of S_{ij} parameters then we can transform these to Z_{ij} (Eq 33) using the following expressions, assuming that $S_{21} = S_{12}$:

$$\begin{aligned} Z_{11} &= \frac{(1+S_{11})(1-S_{22})+S_{12}{}^2}{(1-S_{11})(1-S_{22})-S_{12}{}^2} \\ Z_{22} &= \frac{(1-S_{11})(1+S_{22})+S_{12}{}^2}{(1-S_{11})(1-S_{22})-S_{12}{}^2} \\ Z_{12} &= \frac{(2S_{12})}{(1-S_{11})(1-S_{22})-S_{12}{}^2} \end{aligned} \qquad \text{(Eq 39)}$$

The example to this point has been for a 2-element array. The S-parameters can be determined for an array with any number of elements. In an n-port S-parameter measurement, all ports are terminated in Z_0 at the same time. Measurements are made between one set of ports at a time and repeated until all pairs of ports are measured.

ARRAY MEASUREMENT EXAMPLE

A good way to illustrate the use of a VNA for array measurements is to work through an example with a real array. **Fig 69** is a picture of a 2-element 20-meter phased array built by Mark Perrin, N7MQ.

Each element is λ/4 (self resonant at 14.150 MHz) and spaced λ/4 (17 feet 5 inches). In the ideal case, both elements would have the same current amplitude with a 90º phase difference. This gives the cardioid pattern shown in Chapter 8, Multielement Arrays. There are many schemes for correctly feeding such an array. The one used in this example uses two different 75-Ω transmission lines (one λ/4 and the other λ/2, electrically), as described by Roy Lewallen, W7EL.[2,3]

The first task is to resonate the elements individually. With the VNA set to measure S_{11} phase, we will get a graph like that shown in **Fig 70**.

At the λ/4 resonant frequency (f_r) we will see a sharp phase transition as we go from −180° to +180°. This is typical of any series resonant circuit. The length of each element is adjusted until the desired f_r is achieved. This is a very sensitive measurement. You can see the shift in f_r due to the wind blowing, the length of the element changing as it heats up in the sun or any interactions between the feed line and the antenna as you move the feed line around. In fact this is very good point in the process to make sure everything is mechanically stable and free of unexpected couplings. Usually you will find it necessary place choke baluns on each element to reduce stray coupling.

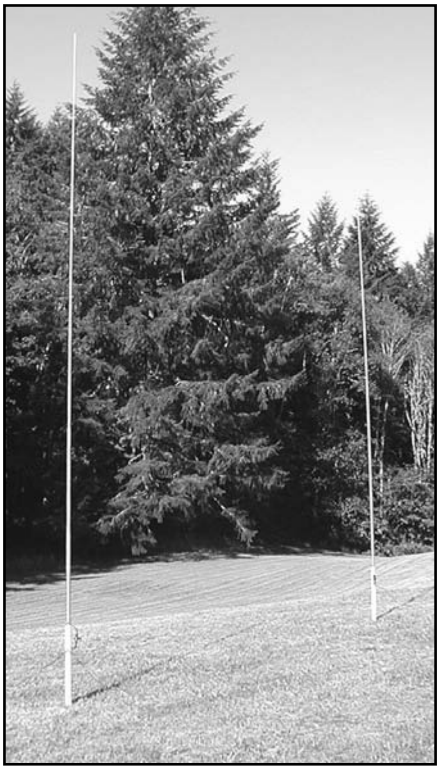

Fig 69—2-element 20-meter phased array (*Photo courtesy N7MQ*).

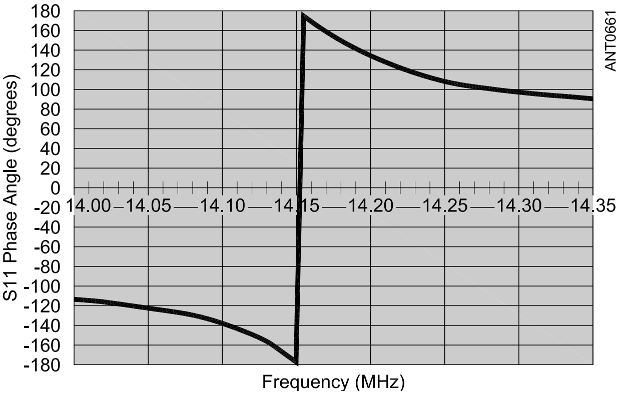

Fig 70—S_{11} phase plot for an individual element.

The next step is to determine the self (Z_{11} and Z_{22}) and mutual (Z_{12}) impedances from which the actual driving point impedances present when the array is excited can be determined. See Chapter 8, Multielement Arrays. There are two ways to go.

First, we can simply use the VNA as an impedance bridge—ie, make two S_{11} measurements at one element, first with the other element open (Z_{11} or Z_{22}) and then with it shorted (Z_1 or Z_2). We can convert the S_{11} measurements to impedances using Eq 38. The value for Z_{12} can be obtained from Eq 40:

$$\begin{aligned} Z_{12} &= \pm\sqrt{Z_{11}\left(Z_{11}-Z_1\right)} \\ Z_{12} &= \pm\sqrt{Z_{22}\left(Z_{22}-Z_2\right)} \end{aligned} \qquad \text{(Eq 40)}$$

The second approach is to do a full two-port S-parameter measurement (S_{11}, S_{21}, S_{12} and S_{22}) and derive the impedances using Eq 38. Both approaches will work but the second approach has the advantage that the ± ambiguity in Eq 40 is eliminated.

For this example, the impedance values from the measurements at 14.150 MHz, turn out to be:

$$\begin{aligned} Z_{11} &= 51.4 + j\,0.35 \\ Z_{22} &= 50.3 + j\,0.299 \\ Z_{12} &= 15.06 - j\,19.26 \end{aligned} \qquad \text{(Eq 41)}$$

With these values we can now determine the feed-point impedances from:

$$\begin{aligned} Z_1{}' &= Z_{11} + \frac{I_2}{I_1} Z_{12} \\ Z_2{}' &= Z_{22} + \frac{I_1}{I_2} Z_{12} \\ \frac{I_1}{I_2} &= -j \end{aligned} \qquad \text{(Eq 42)}$$

Note that –*j* represents the 90° phase shift between the currents. Substituting the values from Eq 41 into Eq 42:

$$\begin{aligned} Z_1{}' &= 32.09 - j\,14.7 \\ Z_2{}' &= 69.6 + j\,15.32 \end{aligned} \qquad \text{(Eq 43)}$$

With these impedances in hand we can now design the feed network. In this particular example however, we have decided to use the λ/4 and λ/2 cables as described by Lewallen[2,3] and accept the results. So we now proceed to cut and trim the two cables to length.

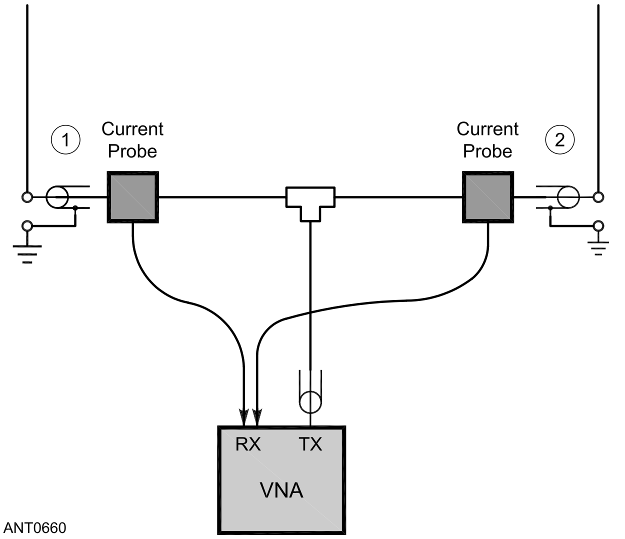

Fig 71—Current phase and amplitude ratio test setup.

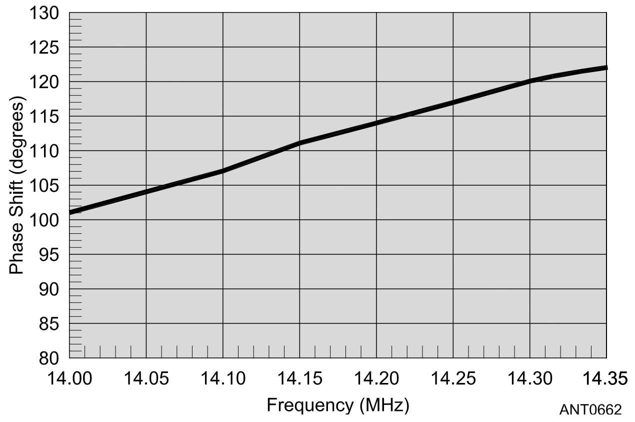

Fig 73—Measured relative current phase shift over the 20-meter band.

Again, there are two ways to go. First we can determine the frequency at which each cable is λ/4 long. At this point the input impedance of the cable will be equivalent to a series-resonant circuit and we can simply measure the phase of S_{11} as we did earlier for f_r and get a plot like that shown in Fig 70. In this example the λ/4 resonant frequencies of the two cables are 7.075 MHz and 14.150 MHz.

The second approach would be to measure S_{21} for each cable at 14.150 MHz. The phase shift in S_{21} tells you how long the cable is, in degrees, at a given frequency. Because there is a small variation in cable characteristics with frequency (dispersion) this approach is slightly more accurate since it is done at the desired operating frequency. But this is not very large effect at HF.

This brings us down to the final measurements, which are to check that the relative current amplitudes and phases between the two elements are correct. We can then determine the feed-point SWR. The phase and amplitude ratios are made using the S_{12} capability of the VNA and the test setup shown in **Fig 71**.

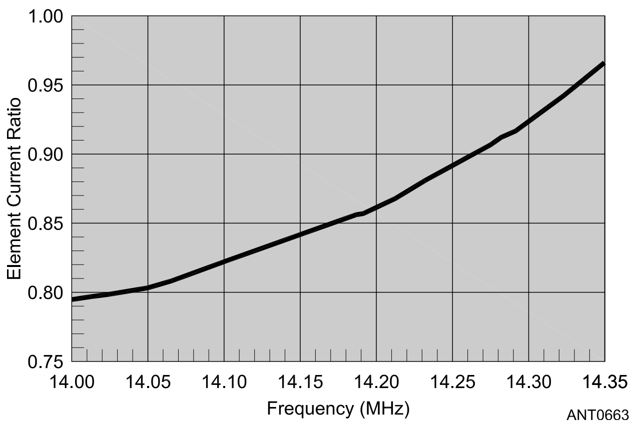

Fig 72—Measured element current ratio over the 20-meter band.

The VNA transmit port is connected to the normal feed point. A current sensor (see Chapter 8, Multielement Arrays, for a discussion of current sensors) is inserted at the base of element 1 and the output of the sensor is returned to the detector or receive port of the VNA. A calibration run is then made to normalize this path. That makes it the reference.

Next, the current sensor is shifted to element 2. The amplitude and phase plots for S_{12} obtained at this point will be the desired relative phase shift and amplitude ratio between the currents in the array when driven at the normal feed point. **Figs 72** and **73** show the behavior of the example array over the 20-meter band. Note that the amplitude ratio has been converted from dB. We can now use these values in a *EZNEC* model of the array to determine the actual radiation pattern.

Obviously the W7EL feed scheme is not perfect, but it has a definite advantage of simplicity. If better performance is desired we can use the values of Z_1' and Z_2' determined earlier to design and fabricate a new feed network and then proceed to evaluate its performance in the same way.

The final measurement is to connect the transmit port of the VNA to the feed point and measure S_{11}. From this we can calculate the SWR:

$$\mathrm{SWR} = \frac{1+|S_{11}|}{1-|S_{11}|} \qquad \text{(Eq 44)}$$

In this example, the return loss, $|S_{11}|$, is about –19 dB over the entire 20-meter band. This corresponds to SWR= 1.25:1.

Notes

[1] Paul Kiciak, N2PK, **http://n2pk.com**.

[2] Roy Lewallen, W7EL, *QST*, Aug 1979, Technical Correspondence, pp 42-43.

[3] Orr and Cowan, *Vertical Antennas*, Radio Amateur Call Book, 1986, pp 148-150.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

A. Bailey, “The Antenna Lab, Parts 1 and 2,” *Radio Communication*, Aug and Sep 1983.

L. Blake, *Transmission Lines and Waveguides* (New York: John Wiley & Sons, 1969), pp 244-251.

J. H. Bowen, “A Calorimeter for VHF and UHF Power Measurements, *QST*, Dec 1975, pp 11-13.

W. Bruene, “An Inside Picture of Directional Wattmeters,” *QST*, Apr 1959.

CCIR Recommendation 368, Documents of the CCIR XII Plenary assembly, ITU, Geneva, 1967.

J. Carr, “Find Fault with Your Coax,” *73*, Oct 1984, pp 10-14.

D. DeMaw, “In-Line RF Power Metering,” *QST*, Dec 1969.

D. Fayman, “A Simple Computing SWR Meter,” *QST*, Jul 1973, pp 23-33.

J. Gibbons and H. Horn, “A Circuit With Logarithmic Response Over Nine Decades,” *IEEE Transactions on Circuit Theory*, Vol CT-11, No. 3, Sep 1964, pp 378-384.

J. Grebenkemper, “Calibrating Diode Detectors,” *QEX*, Aug 1990.

J. Grebenkemper, “The Tandem Match—An Accurate Directional Wattmeter,” *QST*, Jan 1987.

J. Grebenkemper, “Improving and Using R-X Noise Bridges,” *QST*, Aug 1989; Feedback, *QST*, Jan 1990, p 27.

G. H. Hagn, SRI, “HF Ground Measurements at the Lawrence Livermore National Laboratory (LLNL) Field Site,” *Applied Computational Electromagnetics Society Journal and Newsletter*, Vol 3, Number 2, Fall 1988.

T. King, “A Practical Time-Domain Reflectometer,” *QST*, May 1989, pp 22-24.

Z. Lau and C. Hutchinson, “Improving the HW-9 Transceiver,” *QST*, Apr 1988, pp 26-29.

V. G. Leenerts, “Automatic VSWR and Power Meter,” *Ham Radio*, May 1980, pp 34-43.

J. Lenk, *Handbook of Oscilloscopes* (Englewood Cliffs, NJ: Prentice-Hall, 1982), pp 288-292.

I. Lindell, E. Alanen, K. Mannerslo, “Exact Image Method for Impedance Computation of Antennas Above the Ground,” *IEEE Trans. On Antennas and Propagation*, AP-33, Sep 1985.

L. McCoy, “A Linear Field-Strength Meter,” *QST*, Jan 1973.

T. McMullen, “The Line Sampler, an RF Power Monitor for VHF and UHF,” *QST*, Apr 1972.

M. Walter Maxwell, W2DU, *Reflections* (Newington: ARRL, 1990), p 20-3. [Out of print.]

C. Michaels, “Determining Line Lengths,” Technical Correspondence, *QST*, Sep 1985, pp 43-44.

H. Perras, “Broadband Power-Tracking VSWR Bridge,” *Ham Radio*, Aug 1979, pp 72-75.

S. Ramo, J. Whinnery and T. Van Duzer, *Fields and Waves in Communication Electronics* (New York: John Wiley & Sons, 1967), Chap 1.

Reference Data for Radio Engineers, 5th edition (Indianapolis: Howard W. Sams, 1968), Chapter 28.

P. N. Saveskie, *Radio Propagation Handbook* (Blue Ridge Summit, PA: TAB Books, 1960).

J. Sevick, “Short Ground-Radial Systems for Short Verticals,” *QST*, Apr 1978, pp 30-33.

J. Sevick, “Measuring Soil Conductivity,” *QST*, Mar 1981, pp 38- 39.

W. Spaulding, “A Broadband Two-Port S-Parameter Test Set,” *Hewlett-Packard Journal*, Nov 1984.

D. Turrin, “Antenna Performance Measurements,” *QST*, Nov 1974, pp 35-41.

F. Van Zant, “High-Power Operation with the Tandem Match Directional Coupler,” Technical Correspondence, *QST*, Jul 1989, pp 42-43.

R. F. H. Yang, “A Proposed Gain Standard for VHF Antennas,” *IEEE Transactions on Antennas and Propagation*, Nov 1966.

Chapter 28

Smith Chart Calculations

The Smith Chart is a sophisticated graphic tool for solving transmission line problems. One of the simpler applications is to determine the feed-point impedance of an antenna, based on an impedance measurement at the input of a random length of transmission line. By using the Smith Chart, the impedance measurement can be made with the antenna in place atop a tower or mast, and there is no need to cut the line to an exact multiple of half wavelengths. The Smith Chart may be used for other purposes, too, such as the design of impedance-matching networks. These matching networks can take on any of several forms, such as L and pi networks, a stub matching system, a series-section match, and more. With a knowledge of the Smith Chart, the amateur can eliminate much “cut and try” work.

Named after its inventor, Phillip H. Smith, the Smith Chart was originally described in *Electronics* for January 1939. Smith Charts may be obtained at most university book stores. Smith Charts are also available from ARRL HQ. (See the caption for Fig 3.)

It is stated in Chapter 24 that the input impedance, or the impedance seen when “looking into” a length of line, is dependent upon the SWR, the length of the line, and the Z_0 of the line. The SWR, in turn, is dependent upon the load which terminates the line. There are complex mathematical relationships which may be used to calculate the various values of impedances, voltages, currents, and SWR values that exist in the operation of a particular transmission line. These equations can be solved with a personal computer and suitable software, or the parameters may be determined with the Smith Chart. Even if a computer is used, a fundamental knowledge of the Smith Chart will promote a better understanding of the problem being solved. And such an understanding might lead to a quicker or simpler solution than otherwise. If the terminating impedance is known, it is a simple matter to determine the input impedance of the line for any length by means of the chart. Conversely, as indicated above, with a given line length and a known (or measured) input impedance, the load impedance may be determined by means of the chart—a convenient method of remotely determining an antenna impedance, for example.

Although its appearance may at first seem somewhat formidable, the Smith Chart is really nothing more than a specialized type of graph. Consider it as having curved, rather than rectangular, coordinate lines. The coordinate system consists simply of two families of circles—the resistance family, and the reactance family. The resistance circles, **Fig 1**, are centered on the resistance axis (the only straight line on the chart), and are tangent to the outer circle at the right of the chart. Each circle is assigned a value of resistance, which is indicated at the point where the circle crosses the resistance axis. All points along any one circle have the same resistance value.

The values assigned to these circles vary from zero at the left of the chart to infinity at the right, and actually represent a *ratio* with respect to the impedance value assigned to the center point of the chart, indicated 1.0. This center point is called prime center. If prime center is assigned a value of 100 Ω, then 200 Ω resistance is represented by the 2.0 circle, 50 Ω by the 0.5 circle, 20 Ω by the 0.2 circle, and so on. If, instead, a value of 50 is assigned to prime center,

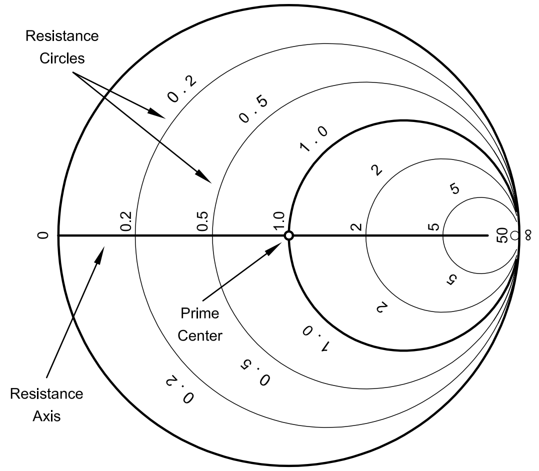

Fig 1—Resistance circles of the Smith Chart coordinate system.

the 2.0 circle now represents 100 Ω, the 0.5 circle 25 Ω, and the 0.2 circle 10 Ω. In each case, it may be seen that the value on the chart is determined by dividing the actual resistance by the number assigned to prime center. This process is called normalizing.

Conversely, values from the chart are converted back to actual resistance values by multiplying the chart value times the value assigned to prime center. This feature permits the use of the Smith Chart for any impedance values, and therefore with any type of uniform transmission line, whatever its impedance may be. As mentioned above, specialized versions of the Smith Chart may be obtained with a value of 50 Ω at prime center. These are intended for use with 50-Ω lines.

Now consider the reactance circles, **Fig 2**, which appear as curved lines on the chart because only segments of the complete circles are drawn. These circles are tangent to the resistance axis, which itself is a member of the reactance family (with a radius of infinity). The centers are displaced to the top or bottom on a line tangent to the right of the chart. The large outer circle bounding the coordinate portion of the chart is the reactance axis.

Each reactance circle segment is assigned a value of reactance, indicated near the point where the circle touches the reactance axis. All points along any one segment have the same reactance value. As with the resistance circles, the values assigned to each reactance circle are normalized with respect to the value assigned to prime center. Values to the top of the resistance axis are positive (inductive), and those to the bottom of the resistance axis are negative (capacitive).

When the resistance family and the reactance family of circles are combined, the coordinate system of the Smith Chart results, as shown in **Fig 3**. Complex impedances (R + *j*X) can be plotted on this coordinate system.

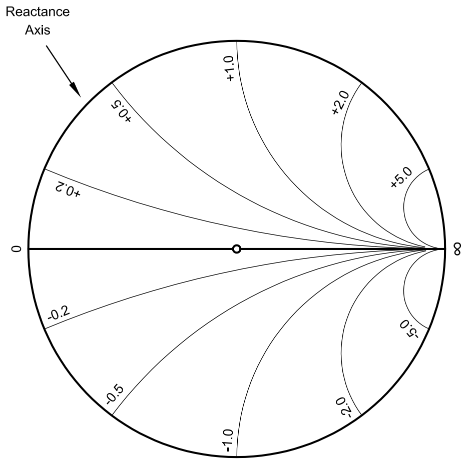

Fig 2—Reactance circles (segments) of the Smith Chart coordinate system.

IMPEDANCE PLOTTING

Suppose we have an impedance consisting of 50 Ω resistance and 100 Ω inductive reactance (Z = 50 + *j* 100). If we assign a value of 100 Ω to prime center, we normalize the above impedance by dividing each component of the impedance by 100. The normalized impedance is then 50/100 + *j* (100/100) = 0.5 + *j* 1.0. This impedance is plotted on the Smith Chart at the intersection of the 0.5 resistance circle and the +1.0 reactance circle, as indicated in Fig 3. Calculations may now be made from this plotted value.

Now say that instead of assigning 100 Ω to prime center, we assign a value of 50 Ω. With this assignment, the 50 + *j* 100 Ω impedance is plotted at the intersection of the 50/50 = 1.0 resistance circle, and the 100/50 = 2.0 positive reactance circle. This value, 1 + *j* 2, is also indicated in Fig 3. But now we have *two* points plotted in Fig 3 to represent the same impedance value, 50 + *j* 100 Ω. How can this be?

These examples show that the same impedance may be plotted at different points on the chart, depending upon the value assigned to prime center. But two plotted points cannot represent the same impedance *at the same time*! It is customary when solving transmission-line problems to assign to prime center a value equal to the characteristic impedance, or Z_0, of the line being used. This value should always be recorded at the start of calculations, to avoid possible confusion later. (In using the specialized charts with the value of 50 at prime center, it is, of course, not necessary to normalize impedances when working with 50-Ω line. The resistance and reactance values may be read directly from the chart coordinate system.)

Prime center is a point of special significance. As

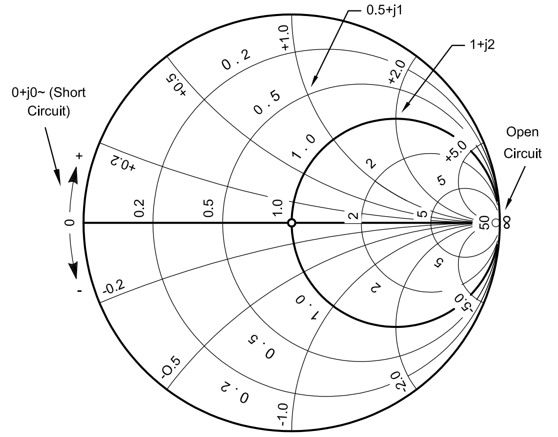

Fig 3—The complete coordinate system of the Smith Chart. For simplicity, only a few divisions are shown for the resistance and reactance values. Various types of Smith Chart forms are available from ARRL HQ. At the time of this writing, five 8½ × 11 inch Smith Chart forms are available for $2.

just mentioned, is is customary when solving problems to assign the Z_0 value of the line to this point on the chart—50 Ω for a 50-Ω line, for example. What this means is that the center point of the chart now represents 50 +*j* 0 ohms–a pure resistance equal to the characteristic impedance of the line. If this were a load on the line, we recognize from transmission-line theory that it represents a perfect match, with no reflected power and with a 1.0 to 1 SWR. Thus, prime center also represents the 1.0 SWR circle (with a radius of zero). SWR circles are also discussed in a later section.

Short and Open Circuits

On the subject of plotting impedances, two special cases deserve consideration. These are short circuits and open circuits. A true short circuit has zero resistance and zero reactance, or 0 + *j* 0). This impedance is plotted at the left of the chart, at the intersection of the resistance and the reactance axes. By contrast, an open circuit has infinite resistance, and therefore is plotted at the right of the chart, at the intersection of the resistance and reactance axes. These two special cases are sometimes used in matching stubs, described later.

Standing-Wave-Ratio Circles

Members of a third family of circles, which are not printed on the chart but which are added during the process of solving problems, are standing-wave-ratio or SWR circles. See **Fig 4**. This family is centered on prime center, and appears as concentric circles inside the reactance axis. During calculations, one or more of these circles may be added with a drawing compass. Each circle represents a value of SWR, with every point on a given circle representing the same SWR. The SWR value for a given circle may be determined directly from the chart coordinate system, by reading the resistance value where the SWR circle crosses the resistance axis to the right of prime center. (The reading where the circle crosses the resistance axis to the left of prime center indicates the inverse ratio.)

Consider the situation where a load mismatch in a length of line causes a 3-to-1 SWR ratio to exist. If we temporarily disregard line losses, we may state that the SWR remains constant throughout the entire length of this line. This is represented on the Smith Chart by drawing a 3:1 constant SWR circle (a circle with a radius of 3 on the resistance axis), as in **Fig 5**. The design of the chart is such that any impedance encountered *anywhere* along the length of this mismatched line will fall on the SWR circle. The impedances may be read from the coordinate system merely by the progressing around the SWR circle by an amount corresponding to the length of the line involved.

This brings into use the wavelength scales, which appear in Fig 5 near the perimeter of the Smith Chart. These scales are calibrated in terms of portions of an electrical wavelength along a transmission line. Both scales start from 0 at the left of the chart. One scale, running

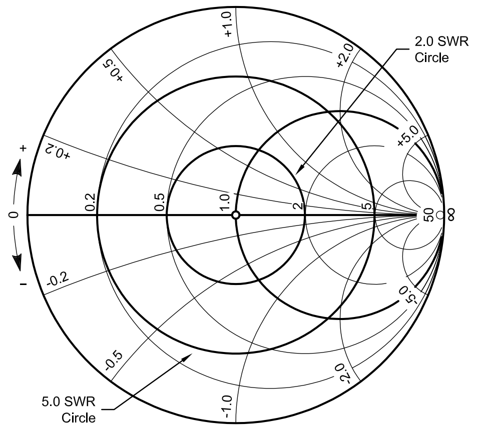

Fig 4—Smith Chart with SWR circles added.

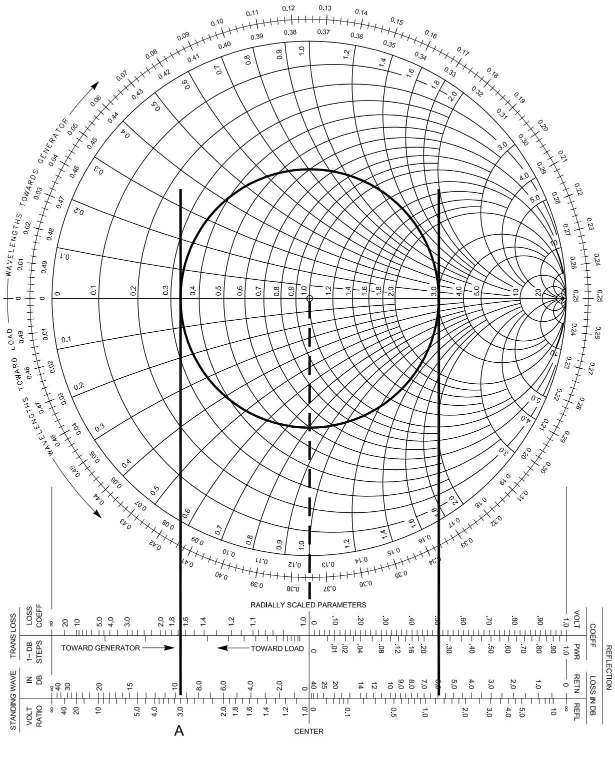

Fig 5—Example discussed in text.

counterclockwise, starts at the generator or input end of the line and progresses toward the load. The other scale starts at the load and proceeds toward the generator in a clockwise direction. The complete circle around the edge of the chart represents $^1/_2$ λ. Progressing once around the perimeter of these scales corresponds to progressing along a transmission line for $^1/_2$ λ. Because impedances repeat themselves every $^1/_2$ λ along a piece of line, the chart may be used for any length of line by disregarding or subtracting from the line's total length an integral, or whole number, of half wavelengths.

Also shown in Fig 5 is a means of transferring the radius of the SWR circle to the external scales of the chart, by drawing lines tangent to the circle. Another simple way to obtain information from these external scales is to transfer the radius of the SWR circle to the external scale with a drawing compass. Place the point of a drawing compass at the center or 0 line, and inscribe a short arc across the appropriate scale. It will be noted that when this is done in Fig 5, the external STANDING-WAVE VOLTAGE-RATIO scale indicates the SWR to be 3.0 (at A)—our condition for initially drawing the circle on the chart (and the same as the SWR reading on the resistance axis).

SOLVING PROBLEMS WITH THE SMITH CHART

Suppose we have a transmission line with a characteristic impedance of 50 Ω and an electrical length of 0.3 λ. Also, suppose we terminate this line with an impedance having a resistive component of 25 Ω and an inductive reactance of 25 Ω (Z = 25 + *j* 25). What is the input impedance to the line?

The characteristic impedance of the line is 50 Ω, so we begin by assigning this value to prime center. Because the line is not terminated in its characteristic impedance, we know that standing waves will exist on the line, and that, therefore, the input impedance to the line will not be exactly 50 Ω. We proceed as follows. First, normalize the load impedance by dividing both the resistive and reactive components by 50 (Z_0 of the line being used). The normalized impedance in this case is 0.5 + *j* 0.5. This is plotted on the chart at the intersection of the 0.5 resistance and the +0.5 reactance circles, as in **Fig 6**. Then draw a constant SWR circle passing through this point. Transfer the radius of this circle to the external scales with the drawing compass. From the external STANDING-WAVE VOLTAGE-RATIO scale, it may be seen (at A) that the voltage ratio of 2.62 exists for this radius, indicating that our line is operating with an SWR of 2.62 to 1. This figure is converted to decibels in the adjacent scale, where 8.4 dB may be read (at B), indicating that the ratio of the voltage maximum to the voltage minimum along the line is 8.4 dB. (This is mathematically equivalent to 20 times the log of the SWR value.)

Next, with a straightedge, draw a radial line from prime center through the plotted point to intersect the wavelengths scale. At this intersection, point C in Fig 6, read a value from the wavelengths scale. Because we are starting from the load, we use the TOWARD GENERATOR or outermost calibration, and read 0.088 λ.

To obtain the line input impedance, we merely find the point on the SWR circle that is 0.3 λ toward the generator from the plotted load impedance. This is accomplished by adding 0.3 (the length of the line in wavelengths) to the reference or starting point, 0.088; 0.3 + 0.088 = 0.388. Locate 0.388 on the TOWARD GENERATOR scale (at D). Draw a second radial line from this point to prime center. The intersection of the new radial line with the SWR circle represents the normalized line input impedance, in this case 0.6 – *j* 0.66.

To find the unnormalized line impedance, multiply by 50, the value assigned to prime center. The resulting value is 30 – *j* 33, or 30 Ω resistance and 33 Ω capacitive reactance. *This* is the impedance that a transmitter must match if such a system were a combination of antenna

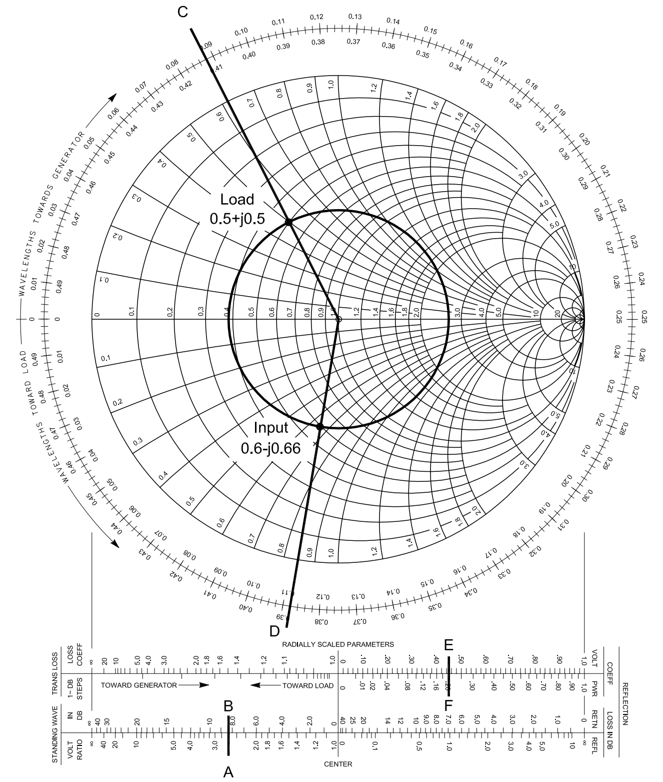

Fig 6—Example discussed in text.

and transmission line. This is also the impedance that would be measured on an impedance bridge if the measurement were taken at the line input.

In addition to the line input impedance and the SWR, the chart reveals several other operating characteristics of the above system of line and load, if a closer look is desired. For example, the voltage reflection coefficient, both magnitude and phase angle, for this particular load is given. The phase angle is read under the radial line drawn through the plot of the load impedance, where the line intersects the ANGLE OF REFLECTION COEFFICIENT scale. This scale is not included in Fig 6, but will be found on the Smith Chart just inside the wavelengths scales. In this example, the reading is 116.6 degrees. This indicates the angle by which the reflected voltage wave leads the incident wave at the load. It will be noted that angles on the bottom half, or capacitive-reactance half, of the chart are negative angles, a "negative" lead indicating that the reflected voltage wave actually lags the incident wave.

The magnitude of the voltage-reflection-coefficient may be read from the external REFLECTION COEFFICIENT VOLTAGE scale, and is seen to be approximately 0.45 (at E) for this example. This means that 45 percent of the incident voltage is reflected. Adjacent to this scale on the POWER calibration, it is noted (at F) that the power reflection coefficient is 0.20, indicating that 20 percent of the incident power is reflected. (The amount of reflected power is proportional to the square of the reflected voltage.)

ADMITTANCE COORDINATES

Quite often it is desirable to convert impedance information to admittance data—conductance and susceptance. Working with admittances greatly simplifies determining the resultant when two complex impedances are connected in parallel, as in stub matching. The conductance values may be added directly, as may be the susceptance values, to arrive at the overall admittance for the parallel combination. This admittance may then be converted back to impedance data, if desired.

On the Smith Chart, the necessary conversion may be made very simply. The equivalent admittance of a plotted impedance value lies diametrically opposite the impedance point on the chart. In other words, an impedance plot and its corresponding admittance plot will lie on a straight line that passes through prime center, and each point will be the same distance from prime center (on the same SWR circle). In the above example, where the normalized line input impedance is 0.6 – *j* 0.66, the equivalent admittance lies at the intersection of the SWR circle and the extension of the straight line passing from point D though prime center. Although not shown in Fig 6, the normalized admittance value may be read as 0.76 + *j* 0.84 if the line starting at D is extended.

In making impedance-admittance conversions, remember that capacitance is considered to be a positive susceptance and inductance a negative susceptance. This corresponds to the scale identification printed on the chart. The admittance in siemens is determined by dividing the normalized values by the Z_0 of the line. For this example the admittance is 0.76/50 + *j* 0.84/50 = 0.0152 + *j* 0.0168 siemen. Of course admittance coordinates may be converted to impedance coordinates just as easily—by locating the point on the Smith Chart that is diametrically opposite that representing the admittance coordinates, on the same SWR circle.

DETERMINING ANTENNA IMPEDANCES

To determine an antenna impedance from the Smith Chart, the procedure is similar to the previous example. The electrical length of the feed line must be known and the impedance value at the input end of the line must be determined through measurement, such as with an impedance-measuring or a good quality noise bridge. In this case, the antenna is connected to the far end of the line and becomes the load for the line. Whether the

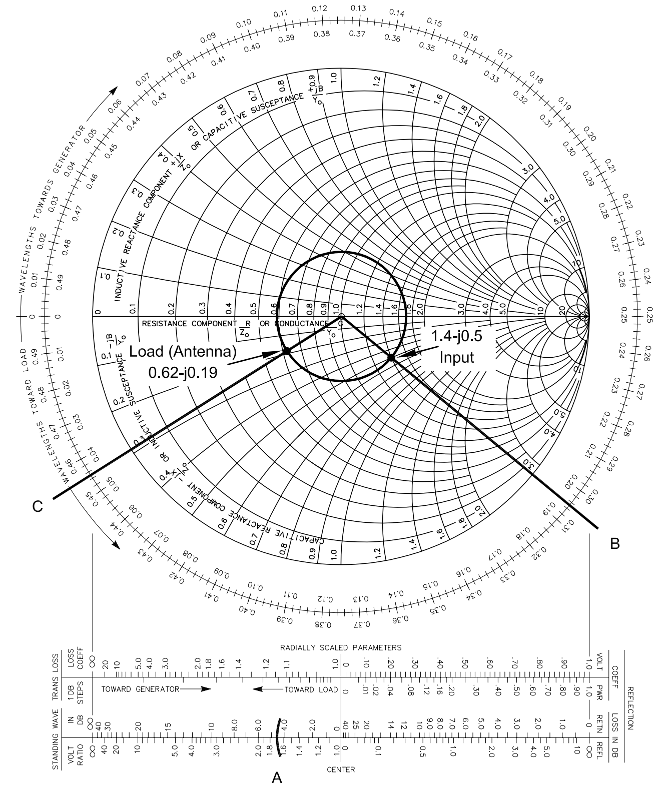

Fig 7—Example discussed in text.

antenna is intended purely for transmission of energy, or purely for reception makes no difference; the antenna is still the terminating or load impedance on the line as far as these measurements are concerned. The input or generator end of the line is that end connected to the device for measurement of the impedance. In this type of problem, the measured impedance is plotted on the chart, and the TOWARD LOAD wavelengths scale is used in conjunction with the electrical line length to determine the actual antenna impedance.

For example, assume we have a measured input impedance to a 50-Ω line of 70 – *j* 25 Ω. The line is 2.35 λ long, and is terminated in an antenna. What is the antenna feed impedance? Normalize the input impedance with respect to 50 Ω, which comes out 1.4 – *j* 0.5, and plot this value on the chart. See **Fig 7**. Draw a constant SWR circle through the point, and transfer the radius to the external scales. The SWR of 1.7 may be read from the VOLTAGE RATIO scale (at A). Now draw a radial line from prime center through this plotted point to the wavelengths scale, and read a reference value (at B). For this case the value is 0.195, on the TOWARD LOAD scale. Remember, we are starting at the generator end of the transmission line.

To locate the load impedance on the SWR circle, add the line length, 2.35 λ, to the reference value from the wavelengths scale; 2.35 + 0.195 = 2.545. Locate the new value on the TOWARD LOAD scale. But because the calibrations extend only from 0 to 0.5, we must first subtract a number of half wavelengths from this value and use only the remaining value. In this situation, the largest integral number of half wavelengths that can be subtracted with a positive result is 5, or 2.5 λ. Thus, 2.545 – 2.5 = 0.045. Locate the 0.045 value on the TOWARD LOAD scale (at C). Draw a radial line from this value to prime center. Now, the coordinates at the intersection of the second radial line and the SWR circle represent the load impedance. To read this value closely, some interpolation between the printed coordinate lines must be made, and the value of 0.62 – *j* 0.19 is read. Multiplying by 50, we get the actual load or antenna impedance as 31 – *j* 9.5 Ω, or 31 Ω resistance with 9.5 Ω capacitive reactance.

Problems may be entered on the chart in yet another manner. Suppose we have a length of 50-Ω line feeding a base-loaded resonant vertical ground-plane antenna which is shorter than $^1/_4$ λ. Further, suppose we have an SWR monitor in the line, and that it indicates an SWR of 1.7 to 1. The line is known to be 0.95 λ long. We want to know both the input and the antenna impedances.

From the information available, we have no impedances to enter into the chart. We may, however, draw a circle representing the 1.7 SWR. We also know, from the definition of resonance, that the antenna presents a purely resistive load to the line, that is, no reactive component. Thus, the antenna impedance must lie on the resistance axis. If we were to draw such an SWR circle and observe the chart with only the circle drawn, we would see two points which satisfy the resonance requirement for the load. These points are 0.59 + *j* 0 and 1.7 + *j* 0. Multiplying by 50, we see that these values represent 29.5 and 85 Ω resistance. This may sound familiar, because, as was discussed in Chapter 24, when a line is terminated in a pure resistance, the SWR in the line equals Z_R/Z_0 or Z_0/Z_R, where Z_R=load resistance and Z_0=line impedance.

If we consider antenna fundamentals described in Chapter 2, we know that the theoretical impedance of a $^1/_4$-λ ground-plane antenna is approximately 36 Ω. We therefore can quite logically discard the 85-Ω impedance figure in favor of the 29.5-Ω value. This is then taken as the load impedance value for the Smith Chart calculations. To find the line input impedance, we subtract 0.5 λ from the line length, 0.95, and find 0.45 λ on the TOWARD GENERATOR scale. (The wavelength-scale starting point in this case is 0.) The line input impedance is found to be 0.63 – *j* 0.20, or 31.5 – *j* 10 Ω.

DETERMINATION OF LINE LENGTH

In the example problems given so far in this chapter, the line length has conveniently been stated in wavelengths. The electrical length of a piece of line depends upon its physical length, the radio frequency under consideration, and the velocity of propagation in the line. If an impedance-measurement bridge is capable of quite reliable readings at high SWR values, the line length may be determined through line input-impedance measurements with short- or open-circuit line terminations. Information on the procedure is given later in this chapter. A more direct method is to measure the physical length of the line and calculate its electrical length from

$$N = \frac{Lf}{984VF} \qquad \text{(Eq 1)}$$

where

N = number of electrical wavelengths in the line
L = line length in feet
f = frequency, MHz
VF = velocity or propagation factor of the line

The velocity factor may be obtained from transmission-line data tables in Chapter 24.

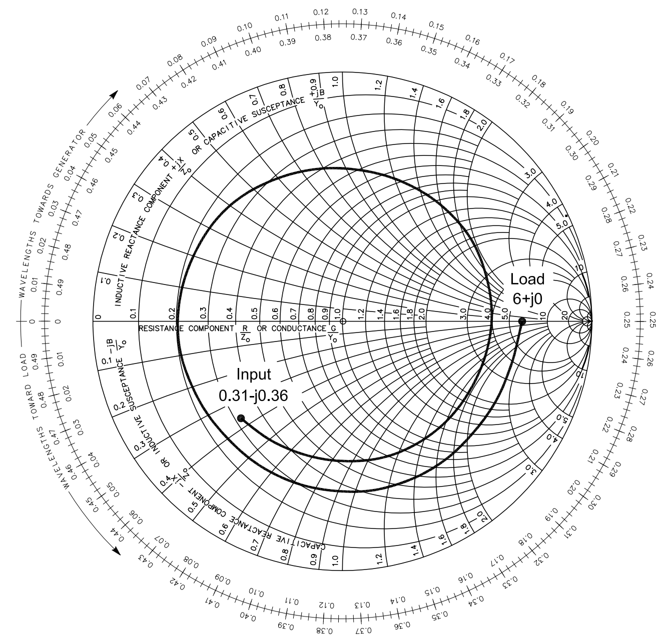
Load
6+j0
Input
0.31-j0.36
RESISTANCE COMPONENT (R/Zo) OR CONDUCTANCE (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo) OR CAPACITIVE SUSCEPTANCE (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (−jX/Zo) OR INDUCTIVE SUSCEPTANCE (−jB/Yo)
WAVELENGTHS TOWARDS GENERATOR
WAVELENGTHS TOWARD LOAD

Fig 8—This spiral is the actual "SWR circle" when line losses are taken into account. It is based on calculations for a 16-ft length of RG-174 coax feeding a resonant 28-MHz 300-Ω antenna (50-Ω coax, velocity factor = 66%, attenuation = 6.2 dB per 100 ft). The SWR at the load is 6:1, while it is 3.6:1 at the line input. When solving problems involving attenuation, two constant SWR circles are drawn instead of a spiral, one for the line input SWR and one for the load SWR.

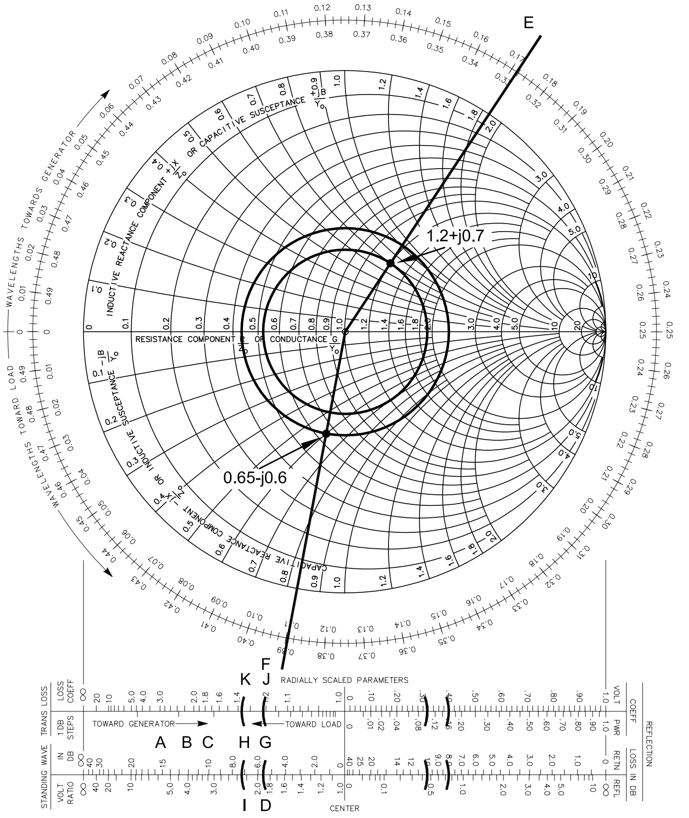
E
1.2+j0.7
0.65-j0.6
F
K
J
RADIALLY SCALED PARAMETERS
TOWARD GENERATOR
TOWARD LOAD
A
B
C
H
G
I
D
CENTER
TRANS LOSS
STANDING WAVE
REFLECTION
LOSS COEFF
1 DB STEPS
VOLT RATIO
IN DB
VOLT COEFF
PWR
RETN LOSS IN DB
REFL

Fig 9—Example of Smith Chart calculations taking line losses into account.

LINE-LOSS CONSIDERATIONS WITH THE SMITH CHART

The example Smith Chart problems presented in the previous section ignored attenuation, or line losses. Quite frequently it is not even necessary to consider losses when making calculations; any difference in readings obtained are often imperceptible on the chart. However, when the line losses become appreciable, such as for high-loss lines, long lines, or at VHF and UHF, loss considerations may become significant in making Smith Chart calculations. This involves only one simple step, in addition to the procedures previously presented.

Because of line losses, as discussed in Chapter 24 the SWR does not remain constant throughout the length of the line. As a result, there is a decrease in SWR as one progresses away from the load. To truly present this situation on the Smith Chart, instead of drawing a constant SWR circle, it would be necessary to draw a spiral inward and clockwise from the load impedance toward the generator, as shown in **Fig 8**. The rate at which the curve spirals toward prime center is related to the attenuation in the line. Rather than drawing spiral curves, a simpler method is used in solving line-loss problems, by means of the external scale TRANSMISSION LOSS 1-DB STEPS. This scale may be seen in **Fig 9**. Because this is only a relative scale, the decibel steps are not numbered.

If we start at the left end of this external scale and proceed in the direction indicated TOWARD GENERATOR, the first dB step is seen to occur at a radius from center corresponding to an SWR of about 9 (at A); the second dB step falls at an SWR of about 4.5 (at B), the third at 3.0 (at C), and so forth, until the

15th dB step falls at an SWR of about 1.05 to 1. This means that a line terminated in a short or open circuit (infinite SWR), and having an attenuation of 15 dB, would exhibit an SWR of only 1.05 at its input. It will be noted that the dB steps near the right end of the scale are very close together, and a line attenuation of 1 or 2 dB in this area will have only slight effect on the SWR. But near the left end of the scale, corresponding to high SWR values, a 1 or 2 dB loss has considerable effect on the SWR.

Using a Second SWR Circle

In solving a problem using line-loss information, it is necessary only to modify the radius of the SWR circle by an amount indicated on the TRANSMISSION-LOSS 1-DB STEPS scale. This is accomplished by drawing a second SWR circle, either smaller or larger than the first, depending on whether you are working toward the load or toward the generator.

For example, assume that we have a 50-Ω line that is 0.282 λ long, with 1-dB inherent attenuation. The line input impedance is measured as 60 + *j* 35 Ω. We desire to know the SWR at the input and at the load, and the load impedance. As before, we normalize the 60 + *j* 35-Ω impedance, plot it on the chart, and draw a constant SWR circle and a radial line through the point. In this case, the normalized impedance is 1.2 + *j* 0.7. From Fig 9, the SWR at the line input is seen to be 1.9 (at D), and the radial line is seen to cross the TOWARD LOAD scale, first subtract 0.500, and locate 0.110 (at F); then draw a radial line from this point to prime center.

To account for line losses, transfer the radius of the SWR circle to the external 1-DB STEPS scale. This radius crosses the external scale at G, the fifth decibel mark from the left. Since the line loss was given as 1 dB, we strike a new radius (at H), one "tick mark" to the left (toward load) on the same scale. (This will be the fourth decibel tick mark from the left of the scale.) Now transfer this new radius back to the main chart, and scribe a new SWR circle of this radius. This new radius represents the SWR at the load, and is read as 2.3 on the external VOLTAGE RATIO scale. At the intersection of the new circle and the load radial line, we read 0.65 – *j* 0.6. This is the normalized load impedance. Multiplying by 50, we obtain the actual load impedance as 32.5 – *j* 30 Ω. The SWR in this problem was seen to increase from 1.9 at the line input to 2.3 (at I) at the load, with the 1-dB line loss taken into consideration.

In the example above, values were chosen to fall conveniently on or very near the "tick marks" on the 1-dB scale. Actually, it is a simple matter to interpolate between these marks when making a radius correction. When this is necessary, the relative distance between marks for each decibel step should be maintained while counting off the proper number of steps.

Adjacent to the 1-DB STEPS scale lies a LOSS COEFFICIENT scale. This scale provides a factor by which the matched-line loss in decibels should be multiplied to account for the increased losses in the line when standing waves are present. These added losses do not affect the SWR or impedance calculations; they are merely the additional dielectric copper losses of the line caused by the fact that the line conducts more average voltage in the presence of standing waves. For the above example, from Fig 9, the loss coefficient at the input end is seen to be 1.21 (at J), and 1.39 (at K) at the load. As a good approximation, the loss coefficient may be averaged over the length of line under consideration; in this case, the average is 1.3. This means that the total losses in the line are 1.3 times the matched loss of the line (1 dB), or 1.3 dB. This is the same result that may be obtained from procedures given in Chapter 24 for this data.

Smith Chart Procedure Summary

To summarize briefly, any calculations made on the Smith Chart are performed in four basic steps, although not necessarily in the order listed.

1) Normalize and plot a line input (or load) impedance, and construct a constant SWR circle.
2) Apply the line length to the wavelengths scales.
3) Determine attenuation or loss, if required, by means of a second SWR circle.
4) Read normalized load (or input) impedance, and convert to impedance in ohms.

The Smith Chart may be used for many types of problems other than those presented as examples here. The transformer action of a length of line—to transform a high impedance (with perhaps high reactance) to a purely resistive impedance of low value—was not mentioned. This is known as "tuning the line," for which the chart is very helpful, eliminating the need for "cut and try" procedures. The chart may also be used to calculate lengths for shorted or open matching stubs in a system, described later in this chapter. In fact, in any application where a transmission line is not perfectly matched, the Smith Chart can be of value.

ATTENUATION AND Z_0 FROM IMPEDANCE MEASUREMENTS

If an impedance bridge is available to make accurate measurements in the presence of very high SWR values, the attenuation, characteristic impedance and velocity factor of any random length of coaxial transmission line can be determined. This section was written by Jerry Hall, K1TD.

Homemade impedance bridges and noise bridges will seldom offer the degree of accuracy required to use this technique, but sometimes laboratory bridges can be found as industrial surplus at a reasonable price. It may also be possible for an amateur to borrow a laboratory type of bridge for the purpose of making some weekend measurements. Making these determinations is not difficult, but the procedure is not commonly known among amateurs. One equation treating complex numbers is used, but the math can be handled with a calculator supporting trig functions. Full details are given in the paragraphs that follow.

For each frequency of interest, two measurements are required to determine the line impedance. Just one measurement is used to determine the line attenuation and velocity factor. As an example, assume we have a 100-foot length of unidentified line with foamed dielectric, and wish to know its characteristics. We make our measurements at 7.15 MHz. The procedure is as follows.

1) Terminate the line in an open circuit. The best "open circuit" is one that minimizes the capacitance between the center conductor and the shield. If the cable has a PL-259 connector, unscrew the shell and slide it back down the coax for a few inches. If the jacket and insulation have been removed from the end, fold the braid back along the outside of the line, away from the center conductor.
2) Measure and record the impedance at the input end of the line. If the bridge measures admittance, convert the measured values to resistance and reactance. Label the values as $R_{oc} + j\,X_{oc}$. For our example, assume we measure $85 + j179\ \Omega$. (If the reactance term is capacitive, record it as negative.)
3) Now terminate the line in a short circuit. If a connector exists at the far end of the line, a simple short is a mating connector with a very short piece of heavy wire soldered between the center pin and the body. If the coax has no connector, removing the jacket and center insulation from a half inch or so at the end will allow you to tightly twist the braid around the center conductor. A small clamp or alligator clip around the outer braid at the twist will keep it tight.
4) Again measure and record the impedance at the input end of the line. This time label the values as $R_{sc} \pm j\,X$. Assume the measured value now is $4.8 - j\,11.2\ \Omega$.

This completes the measurements. Now we reach for the calculator.

As amateurs we normally assume that the characteristic impedance of a line is purely resistive, but it can (and does) have a small capacitive reactance component. Thus, the Z_0 of a line actually consists of $R_0 + j\,X_0$. The basic equation for calculating the characteristic impedance is

$$Z_0 = \sqrt{Z_{oc} \times Z_{sc}} \qquad \text{(Eq 2)}$$

where

$Z_{oc} = R_{oc} + jX_{oc}$

$Z_{sc} = R_{sc} + jX_{sc}$

From Eq 2 the following working equation may be derived.

$$Z_0 = \sqrt{\left(R_{oc}R_{sc} - X_{oc}X_{sc}\right) + j\left(R_{oc}X_{sc} + R_{sc}X_{oc}\right)} \qquad \text{(Eq 3)}$$

The expression under the radical sign in Eq 3 is in the form of R + *j* X. By substituting the values from our example into Eq 3, the R term becomes 85 × 4.8 – 179 × (–11.2) = 2412.8, and the X term becomes 85 × (–11.2) + 4.8 × 179 = –92.8. So far, we have determined that

$$Z_0 = \sqrt{2412.8 - j\,92.8}$$

The quantity under the radical sign is in rectangular form. Extracting the square root of a complex term is handled easily if it is in polar form, a vector value and its angle. The vector value is simply the square root of the sum of the squares, which in this case is

$$\sqrt{2412.8^2 + 92.8^2} = \sqrt{2414.58}$$

The tangent of the vector angle we are seeking is the value of the reactance term divided by the value of the resistance term. For our example this is arctan –92.8/2412.8 = arctan –0.03846. The angle is thus found to be –2.20°. From all of this we have determined that

$$Z_0 = \sqrt{2414.58\,\underline{/-2.20^\circ}}$$

Extracting the square root is now simply a matter of finding

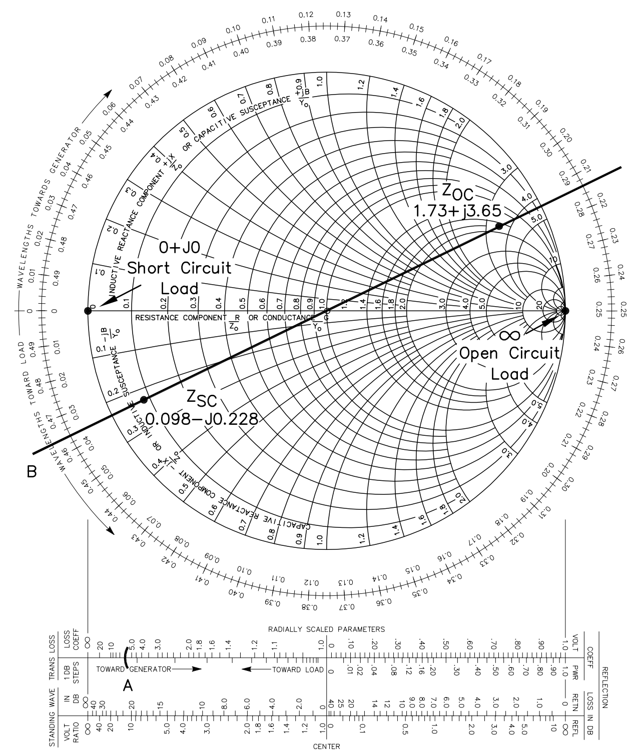

Fig 10—Determining the line loss and velocity factor with the Smith Chart from input measurements taken with open-circuit and short-circuit terminations.

the square root of the vector value, and taking half the angle. (The angle is treated mathematically as an exponent.)

Our result for this example is $Z_0 = 49.1\underline{/-1.1°}$. The small negative angle may be ignored, and we now know that we have coax with a nominal 50-Ω impedance. (Departures of as much as 6 to 8% from the nominal value are not uncommon.) If the negative angle is large, or if the angle is positive, you should recheck your calculations and perhaps even recheck the original measurements. You can get an idea of the validity of the measurements by normalizing the measured values to the calculated impedance and plotting them on a Smith Chart as shown in **Fig 10** for this example. Ideally, the two points should be diametrically opposite, but in practice they will be not quite 180° apart and not quite the same distance from prime center. Careful measurements will yield plotted points that are close to ideal. Significant departures from the ideal indicates sloppy measurements, or perhaps an impedance bridge that is not up to the task.

Determining Line Attenuation

The short circuit measurement may be used to determine the line attenuation. This reading is more reliable than the open circuit measurement because a good short circuit is a short, while a good open circuit is hard to find. (It is impossible to escape some amount of capacitance between conductors with an "open" circuit, and that capacitance presents a path for current to flow at the RF measurement frequency.)

Use the Smith Chart and the 1-DB STEPS external scale to find line attenuation. First normalize the short circuit impedance reading to the calculated Z_0, and plot this point on the chart. See Fig 10. For our example, the normalized impedance is $4.8/49.1 - j\ 11.2\ /\ 49.1$ or $0.098 - j\ 0.228$. After plotting the point, transfer the radius to the 1-DB STEPS scale. This is shown at A of Fig 10.

Remember from discussions earlier in this chapter that the impedance for plotting a short circuit is $0 + j\ 0$, at the left edge of the chart on the resistance axis. On the 1-DB STEPS scale this is also at the left edge. The total attenuation in the line is represented by the number of dB steps from the left edge to the radius mark we have just transferred. For this example it is 0.8 dB. Some estimation may be required in interpolating between the 1-dB step marks.

Determining Velocity Factor

The velocity factor is determined by using the TOWARD GENERATOR wavelength scale of the Smith Chart. With a straightedge, draw a line from prime center through the point representing the short-circuit reading, until it intersects the wavelengths scale. In Fig 10 this point is labeled B. Consider that during our measurement, the short circuit was the load at the end of the line. Imagine a spiral curve progressing from $0 + j\ 0$ clockwise and inward to our plotted measurement point. The wavelength scale, at B, indicates this line length is 0.464 λ. By rearranging the terms of Eq 1 given early in this chapter, we arrive at an equation for calculating the velocity factor.

$$VF = \frac{Lf}{984N} \qquad \text{(Eq 4)}$$

where

VF = velocity factor

L = line length, feet

f = frequency, MHz

N = number of electrical wavelengths in the line

Inserting the example values into Eq 4 yields VF = 100 × 7.15/(984 × 0.464) = 1.566, or 156.6%. Of course, this value is an impossible number—the velocity factor in coax cannot be greater than 100%. But remember, the Smith Chart can be used for lengths greater than $^1/_2$ λ. Therefore, that 0.464 value could rightly be 0.964, 1.464, 1.964, and so on. When using 0.964 λ, Eq 4 yields a velocity factor of 0.753, or 75.3%. Trying successively greater values for the wavelength results in velocity factors of 49.6 and 37.0%. Because the cable we measured had foamed dielectric, 75.3% is the probable velocity factor. This corresponds to an electrical length of 0.964 λ. Therefore, we have determined from the measurements and

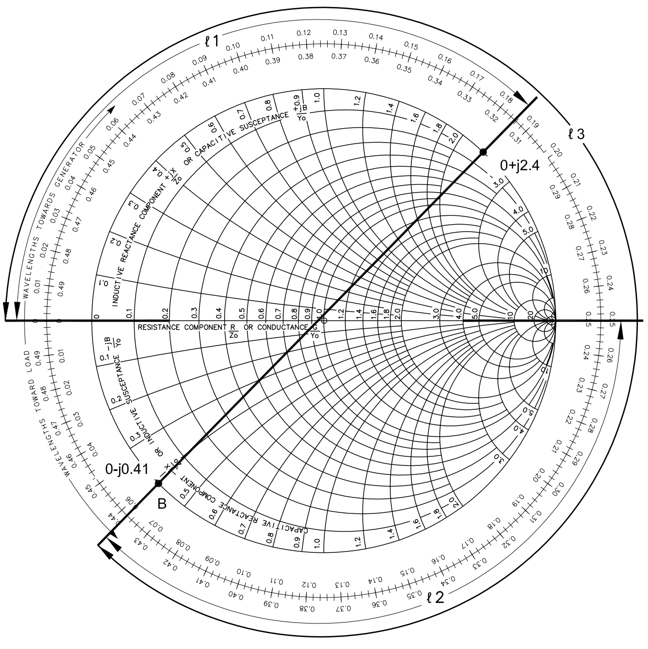
ℓ1
ℓ3
0+j2.4
0-j0.41
B
ℓ2
WAVELENGTHS TOWARDS GENERATOR
WAVELENGTHS TOWARD LOAD
INDUCTIVE REACTANCE COMPONENT
OR CAPACITIVE SUSCEPTANCE
RESISTANCE COMPONENT (R/Z0) OR CONDUCTANCE
INDUCTIVE SUSCEPTANCE
CAPACITIVE REACTANCE COMPONENT

Fig 11—Smith Chart determination of input impedances for short- and open-circuited line sections, disregarding line losses.

calculations that our unmarked coax has a nominal 50-Ω impedance, an attenuation of 0.8 dB per hundred feet at 7.15 MHz, and a velocity factor of 75.3%.

It is difficult to use this procedure with short lengths of coax, just a few feet. The reason is that the SWR at the line input is too high to permit accurate measurements with most impedance bridges. In the example above, the SWR at the line input is approximately 12:1.

The procedure described above may also be used for determining the characteristics of balanced lines. However, impedance bridges are generally unbalanced devices, and the procedure for measuring a balanced impedance accurately with an unbalanced bridge is complicated.

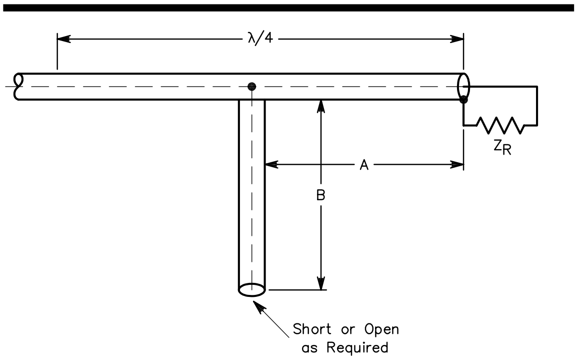

Fig 12—The method of stub matching a mismatched load on coaxial lines.

LINES AS CIRCUIT ELEMENTS

Information is presented in Chapter 24 on the use of transmission-line sections as circuit elements. For example, it is possible to substitute transmission lines of the proper length and termination for coils or capacitors in ordinary circuits. While there is seldom a practical need for that application, lines are frequently used in antenna systems in place of lumped components to tune or resonate elements. Probably the most common use of such a line is in the hairpin match, where a short section of stiff open-wire line acts as a lumped inductor.

The equivalent "lumped" value for any "inductor" or "capacitor" may be determined with the aid of the Smith Chart. Line losses may be taken into account if desired, as explained earlier. See **Fig 11**. Remember that the top half of the Smith Chart coordinate system is used for impedances containing inductive reactances, and the bottom half for capacitive reactances. For example, a section of 600-Ω line 3/16-λ long (0.1875 λ) and short-circuited at the far end is represented by ℓ1, drawn around a portion of the perimeter of the chart. The "load" is a short-circuit, 0 + *j* 0 Ω, and the TOWARD GENERATOR wavelengths scale is used for marking off the line length. At A in Fig 11 may be read the normalized impedance as seen looking into the length of line, 0 + *j* 2.4. The reactance is therefore inductive, equal to 600 × 2.4 = 1440 Ω. The same line when open-circuited (termination impedance = ∞, the point at the right of the chart) is represented by ℓ2 in Fig 11. At B the normalized line-input impedance may be read as 0 – *j* 0.41; the reactance in this case is capacitive, 600 × 0.41 = 246 Ω. (Line losses are disregarded in these examples.) From Fig 11 it is easy to visualize that if ℓ1 were to be extended by 1/4 λ, the total length represented by ℓ3, the line-input impedance would be identical to that obtained in the case represented by ℓ2 alone. In the case of ℓ2, the line is open-circuited at the far end, but in the case of ℓ3 the line is terminated in a short. The added section of line for ℓ3 provides the "transformer action" for which the 1/4-λ line is noted.

The equivalent inductance and capacitance as determined above can be found by substituting these values in the equations relating inductance and capacitance to reactance, or by using the various charts and calculators available. The frequency corresponding to the line length in degrees must be used, of course. In this example, if the frequency is 14 MHz

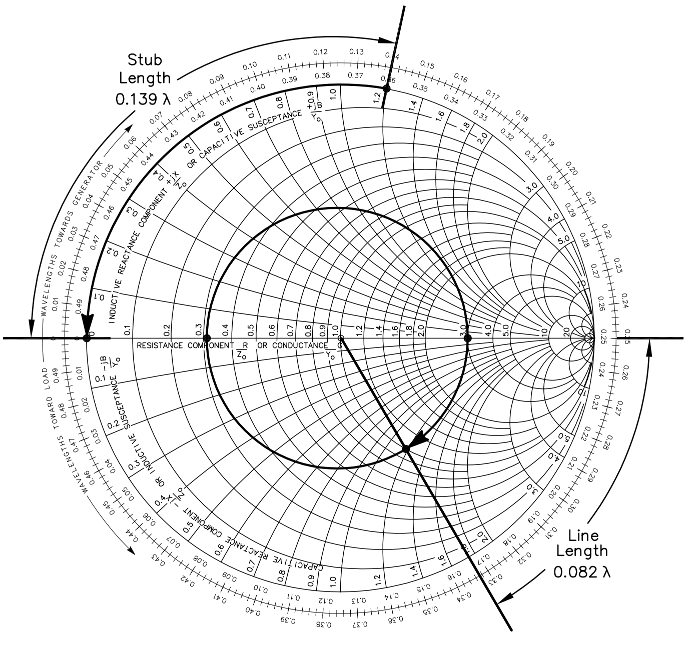

Fig 13—Smith Chart method of determining the dimensions for stub matching.

the equivalent inductance and capacitance in the two cases are 16.4 µH and 46.2 pF, respectively. Note that when the line length is 45° (0.125 λ), the reactance in either case is numerically equal to the characteristic impedance of the line. In using the Smith Chart it should be kept in mind that the electrical length of a line section depends on the frequency and velocity of propagation, as well as on the actual physical length.

At lengths of line that are exact multiples of ¹/₄ λ, such lines have the properties of resonant circuits. At lengths where the input reactance passes through zero at the left of the Smith Chart, the line acts as a series-resonant circuit. At lengths for which the reactances theoretically pass from "positive" to "negative" infinity at the right of the Smith Chart, the line simulates a parallel-resonant circuit.

Designing Stub Matches with the Smith Chart

The design of stub matches is covered in detail in Chapter 26. Equations are presented there to calculate the electrical lengths of the main line and the stub, based on a purely resistive load and on the stub being the same type of line as the main line. The Smith Chart may also be used to determine these lengths, without the requirements that the load be purely resistive and that the line types be identical.

Fig 12 shows the stub matching arrangement in coaxial line. As an example, suppose that the load is an antenna, a close-spaced array fed with a 52-Ω line. Further suppose that the SWR has been measured as 3.1:1. From this information, a constant SWR circle may be drawn on the Smith Chart. Its radius is such that it intersects the right portion of the resistance axis at the SWR value, 3.1, as shown at point B in **Fig 13**.

Since the stub of Fig 12 is connected in parallel with the transmission line, determining the design of the matching arrangement is simplified if Smith Chart values are dealt with as admittances, rather than impedances. (An admittance is simply the reciprocal of the associated impedance. Plotted on the Smith Chart, the two associated points are on the same SWR circle, but diametrically opposite each other.) Using admittances leaves less chance for errors in making calculations, by eliminating the need for making series-equivalent to parallel-equivalent circuit conversions and back, or else for using complicated equations for determining the resultant value of two complex impedances connected in parallel.

A complex impedance, Z, is equal to R + *j* X, as described in Chapter 24. The equivalent admittance, Y, is equal to G – *j* B, where G is the conductive component and B the susceptance. (Inductance is taken as negative susceptance, and capacitance as positive.) Conductance and susceptance values are plotted and handled on the Smith Chart in the same manner as resistance and reactance.

Assuming that the close-spaced array of our example has been resonated at the operating frequency, it will present a purely resistive termination for the load end of the 52-Ω line. From information in Chapter 24, it is known that the impedance of the antenna equals Z_0/SWR = 52/3.1 = 16.8 Ω. (We can logically discard the possibility that the antenna impedance is SWR × Z_0, or 0.06 Ω.) If this 16.8-Ω value were to be plotted as an impedance on the Smith Chart, it would first be normalized (16.8/52 = 0.32) and then plotted as 0.32 + *j* 0. Although not necessary for the solution of this example, this value is plotted at point A in Fig 13. What is necessary is a plot of the admittance for the antenna as a load. This is the reciprocal of the impedance; 1/16.8 Ω equals 0.060 siemen. To plot this point it is first normalized by multiplying the conductance and susceptance values by the Z_0 of the line. Thus, (0.060 + *j* 0) × 52 = 3.1 + *j* 0. This admittance value is shown plotted at point B in Fig 13. It may be seen that points A and B are diametrically opposite each other on the chart. Actually, for the solution of this example, it wasn't necessary to compute the values for either point A or point B as in the above paragraph, for they were both determined from the known SWR value of 3.1. As may be seen in Fig 13, the points are located on the constant SWR circle which was already drawn, at the two places where it intersects the resistance axis. The plotted value for point A, 0.32, is simply the reciprocal of the value for point B, 3.1. However, an understanding of the relationship between impedance and admittance is easier to gain with simple examples such as this.

In stub matching, the stub is to be connected at a point in the line where the conductive component equals the Z_0 of the line. Point B represents the admittance of the load, which is the antenna. Various admittances will be encountered along the line, when moving in a direction indicated by the TOWARD GENERATOR wavelengths scale, but all admittance plots must fall on the constant SWR circle. Moving clockwise around the SWR circle from point B, it is seen that the line input conductance will be 1.0 (normalized Z_0 of the line) at point C, 0.082 λ toward the transmitter from the antenna. Thus, the stub should be connected at this location on the line.

The normalized admittance at point C, the point representing the location of the stub, is 1 – *j* 1.2 siemens, having an inductive susceptance component. A capacitive susceptance having a normalized value of + *j* 1.2 siemens is required across the line at the point of stub connection, to cancel the inductance. This capacitance is to be obtained from the stub section itself; the problem now is to determine its type of termination (open or shorted), and how long the stub should be. This is done by first plotting the susceptance required for cancellation, 0 + *j* 1.2, on the chart (point D in Fig 13). This point represents the input admittance as seen looking into the stub. The "load" or termination for the stub section is found by moving in the TOWARD LOAD direction around the chart, and will appear at the closest point on the resistance/conductance axis, either at the left or the right of the chart. Moving counterclockwise from point D, this is located at E, at the left of the chart, 0.139 λ away. From this we know the required stub length. The "load" at the far end of the stub, as represented on the Smith Chart, has a normalized admittance of 0 + *j* 0 siemen, which is equivalent to an open circuit.

When the stub, having an input admittance of 0 + *j* 1.2 siemens, is connected in parallel with the line at a point 0.082 λ from the load, where the line input admittance is 1.0 – *j* 1.2, the resultant admittance is the sum of the individual admittances. The conductance components are added directly, as are the susceptance components. In this case, 1.0 – *j* 1.2 + *j* 1.2 = 1.0 + *j* 0 siemen. Thus, the line from the point of stub connection to the transmitter will be terminated in a load which offers a perfect match. When determining the physical line lengths for stub matching, it is important to remember that the velocity factor for the type of line in use must be considered.

MATCHING WITH LUMPED CONSTANTS

It was pointed out earlier that the purpose of a matching stub is to cancel the reactive component of line impedance at the point of connection. In other words, the stub is simply a reactance of the proper kind and value shunted across the line. It does not matter what physical shape this reactance takes. It can be a section of transmission line or a "lumped" inductance or capacitance, as desired. In the above example with the Smith Chart solution, a capacitive reactance was required. A capacitor having the same value of reactance can be used just as well. There are cases where, from an installation standpoint, it may be considerably more convenient to connect a capacitor in place of a stub. This is particularly true when open-wire feeders are used. If a variable capacitor is used, it becomes possible to adjust the capacitance to the exact value required.

The proper value of reactance may be determined from Smith Chart information. In the previous example, the required susceptance, normalized, was +*j* 1.2 siemens. This is converted into actual siemens by dividing by the line Z_0; 1.2/52 = 0.023 siemen, capacitance. The required capacitive reactance is the reciprocal of this latter value, 1/0.023 = 43.5 Ω. If the frequency is 14.2 MHz, for instance, 43.5 Ω corresponds to a capacitance of 258 pF. A 325-pF variable capacitor connected across the line 0.082 λ from the antenna terminals would provide ample adjustment range. The RMS voltage across the capacitor is

$$E = \sqrt{P \times Z_0}$$

For 500 W, for example, E = the square root of 500 × 52 = 161 V. The peak voltage is 1.41 times the RMS value, or 227 V.

The Series-Section Transformer

The series-section transformer is described in Chapter 26, and equations are given there for its design. The transformer can be designed graphically with the aid of a Smith Chart. This information is based on a *QST* article by Frank A. Regier, OD5CG. Using the Smith Chart to design a series-section match requires the use of the chart in its less familiar off-center mode. This mode is described in the next two paragraphs.

Fig 14 shows the Smith Chart used in its familiar centered mode, with all impedances normalized to that of the transmission line, in this case 75 Ω, and all constant SWR circles concentric with the normalized value r = 1 at the chart center. An actual impedance is recovered by multiplying a chart reading by the normalizing impedance of 75 Ω. If the actual (unnormalized) impedances represented by a constant SWR circle in Fig 14 are instead divided by a normalizing impedance of 300 Ω, a different picture results. A Smith Chart shows all possible impedances, and so a closed path such as a constant SWR circle in Fig 14 must again be represented by a closed path. In fact, it can be shown that the path remains a circle, but that the constant SWR circles are no longer concentric. **Fig 15** shows the circles that result when the impedances along a mismatched 75-Ω line are normalized by dividing by 300 Ω instead of 75. The constant SWR circles still surround the point corresponding to the characteristic impedance of the line (r = 0.25) but are no longer concentric with it. Note that the normalized impedances read from corresponding points on Figs 14 and 15 are different but that the actual, unnormalized, impedances are exactly the same.

An Example

Now turn to the example shown in **Fig 16**. A complex load of Z_L = 600 + *j* 900 Ω is to be fed with 300-Ω line, and a 75-Ω series section is to be used. These characteristic impedances agree with those used in Fig 15, and thus Fig 15 can be used to find the impedance variation along the 75-Ω series section. In particular, the constant SWR circle which passes through the Fig 15 chart center, SWR = 4 in this case, passes through all the impedances (normalized to 300 Ω) which the 75-Ω series section is able to match to the 300-Ω main line. The length $\ell 1$ of 300-Ω line has the job of transforming the load impedance to some impedance on this matching circle.

Fig 17 shows the whole process more clearly, with all

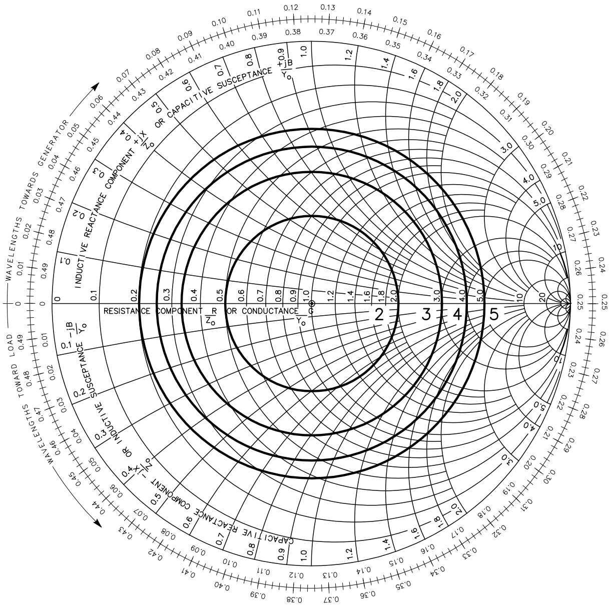

Fig 14—Constant SWR circles for SWR = 2, 3, 4 and 5, showing impedance variation along 75-Ω line, normalized to 75 Ω. The actual impedance is obtained by multiplying the chart reading by 75 Ω.

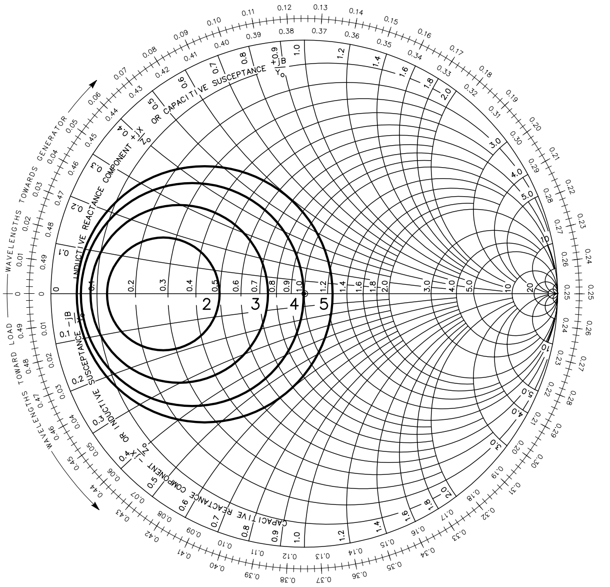

Fig 15—Paths of constant SWR for SWR = 2, 3, 4 and 5, showing impedance variation along 75-Ω line, normalized to 300 Ω. Normalized impedances differ from those in Fig 14, but actual impedances are obtained by multiplying chart readings by 300 Ω and are the same as those corresponding in Fig 14. Paths remain circles but are no longer concentric. One, the matching circle, SWR = 4 in this case, passes through the chart center and is thus the locus of all impedances which can be matched to a 300-Ω line.

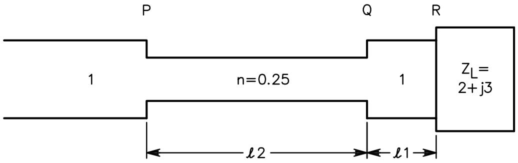

Fig 16—Example for solution by Smith Chart. All impedances are normalized to 300 Ω.

impedances normalized to 300 Ω. Here the normalized load impedance $Z_L = 2 + j\,3$ is shown at R, and the matching circle appears centered on the resistance axis and passing through the points r = 1 and r = n^2 = $(75/300)^2$ = 0.0625. A constant SWR circle is drawn from R to an intersection with the matching circle at Q or Q′ and the corresponding length ℓ1 (or ℓ1′) can be read directly from the Smith Chart. The clockwise distance around the matching circle represents the length of the matching line, from either Q′ to P or from Q to P. Because in this example the distance QP is the shorter of the two for the matching section, we choose the length ℓ1 as shown. By using values from the TOWARD GENERATOR scale, this length is found as 0.045 – 0.213, and adding 0.5 to obtain a positive result yields a value of 0.332 λ.

Although the impedance locus from Q to P is shown in Fig 17, the length ℓ2 cannot be determined directly from this chart. This is because the matching circle is not concentric with the chart center, so the wavelength scales do not apply to this circle. This problem is overcome by forming **Fig 18**, which is the same as Fig 17 except that all normalized impedances have been divided by n = 0.25, resulting in a Smith Chart normalized to 75 Ω instead of 300. The matching circle and the chart center are now concentric, and the series-section length ℓ2, the distance between Q and P, can be taken directly from the chart. By again using the TOWARD GENERATOR scale, this length is found as 0.250 – 0.148 = 0.102 λ.

In fact it is not necessary to construct the entire impedance locus shown in Fig 18. It is sufficient to plot Z_Q/n (Z_Q is read from Fig 17) and $Z_p/n = 1/n$, connect them by a circular arc centered on the chart center, and to determine the arc length ℓ2 from the Smith Chart.

Procedure Summary

The steps necessary to design a series-section transformer by means of the Smith Chart can now be listed:

1) Normalize all impedances by dividing by the characteristic impedance of the main line.
2) On a Smith Chart, plot the normalized load impedance Z_L at R and construct the matching circle

so that its center is on the resistance axis and it passes through the points $r = 1$ and $r = n^2$.

3) Construct a constant SWR circle centered on the chart center through point R. This circle should intersect the matching circle at two points. One of these points, normally the one resulting in the shorter clockwise distance along the matching circle to the chart center, is chosen as point Q, and the clockwise distance from R to Q is read from the chart and taken to be $\ell 1$.
4) Read the impedance Z_Q from the chart, calculate Z_Q/n and plot it as point Q on a second Smith Chart. Also plot $r = 1/n$ as point P.
5) On this second chart construct a circular arc, centered on the chart center, clockwise from Q to P. The length of this arc, read from the chart, represents $\ell 2$. The design of the transformer is now complete, and the necessary physical line lengths may be determined.

The Smith Chart construction shows that two design solutions are usually possible, corresponding to the two intersections of the constant SWR circle (for the load) and the matching circle. These two values correspond to positive and negative values of the square-root radical in the equation for a mathematical solution of the problem. It may happen, however, that the load circle misses the matching circle completely, in which case no solution is possible. The cure is to enlarge the matching circle by choosing a series section whose impedance departs more from that of the main line.

A final possibility is that, rather than intersecting the matching circle, the load circle is tangent to it. There is then but one solution—that of the $^{1}/_{4}$-λ transformer.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

W. N. Caron, *Antenna Impedance Matching* (Newington: ARRL, 1989).

C. MacKeand, “The Smith Chart in BASIC,” *QST*, Nov 1984, pp 28-31.

M. W. Maxwell, *Reflections—Transmission Lines and Antennas* (Newington: ARRL, 1990).

F. A. Regier, Series-Section Transmission-Line Impedance Matching,” *QST*, Jul 1978, pp 14-16.

P. H. Smith, *Electronic Applications of the Smith Chart*, reprint ed. (Malabar, FL: Krieger Pub Co, Inc, 1983).

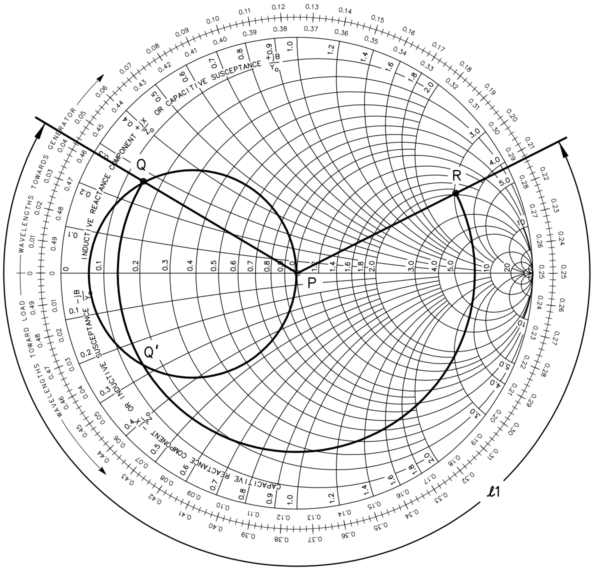

WAVELENGTHS TOWARDS GENERATOR
WAVELENGTHS TOWARD LOAD
INDUCTIVE REACTANCE COMPONENT
OR CAPACITIVE SUSCEPTANCE
CAPACITIVE REACTANCE COMPONENT
OR INDUCTIVE SUSCEPTANCE
Q
R
P
Q′
ℓ1

Fig 17—Smith Chart representation of the example shown in Fig 16. The impedance locus always takes a clockwise direction from the load to the generator. This path is first along the constant SWR circle from the load at R to an intersection with the matching circle at Q or Q′, and then along the matching circle to the chart center at P. Length $\ell 1$ can be determined directly from the chart, and in this example is 0.332 λ.

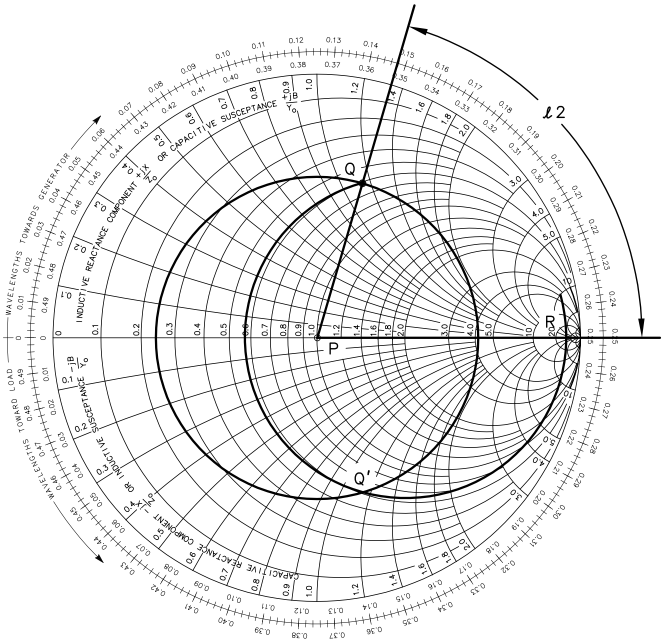

WAVELENGTHS TOWARDS GENERATOR
WAVELENGTHS TOWARD LOAD
INDUCTIVE REACTANCE COMPONENT
OR CAPACITIVE SUSCEPTANCE
CAPACITIVE REACTANCE COMPONENT
OR INDUCTIVE SUSCEPTANCE
ℓ2
Q
R
P
Q′

Fig 18—The same impedance locus as shown in Fig 17 except normalized to 75 Ω instead of 300. The matching circle is now concentric with the chart center, and $\ell 2$ can be determined directly from the chart, 0.102 λ in this case.

Appendix

This appendix contains a glossary of terms, a list of common abbreviations, length conversion information (feet and inches), metric equivalents and antenna-gain-reference data.

Glossary of Terms

This glossary provides a handy list of terms that are used frequently in Amateur Radio conversation and literature about antennas. With each item is a brief definition of the term. Most terms given here are discussed more thoroughly in the text of this book, and may be located by using the index.

Actual ground—The point within the earth's surface where effective ground conductivity exists. The depth for this point varies with frequency and the condition of the soil.

Antenna—An electrical conductor or array of conductors that radiates signal energy (transmitting) or collects signal energy (receiving).

Antenna tuner—A device containing variable reactances (and perhaps a balun). It is connected between the transmitter and the feed point of an antenna system, and adjusted to "tune" or resonate the system to the operating frequency.

Aperture, effective—An area enclosing an antenna, on which it is convenient to make calculations of field strength and antenna gain. Sometimes referred to as the "capture area."

Apex—The feed-point region of a V type of antenna.

Apex angle—The included angle between the wires of a V, an inverted-V dipole, and similar antennas, or the included angle between the two imaginary lines touching the element tips of a log periodic array.

Balanced line—A symmetrical two-conductor feed line that has uniform voltage and current distribution along its length.

Balun—A device for feeding a balanced load with an unbalanced line, or vice versa. May be a form of choke, or a transformer that provides a specific impedance transformation (including 1:1). Often used in antenna systems to interface a coaxial transmission line to the feed point of a balanced antenna, such as a dipole.

Base loading—A lumped reactance that is inserted at the base (ground end) of a vertical antenna to resonate the antenna.

Bazooka—A transmission-line balancer. It is a quarter-wave conductive sleeve (tubing or flexible shielding) placed at the feed point of a center-fed element and grounded to the shield braid of the coaxial feed line at the end of the sleeve farthest from the feed point. It permits the use of unbalanced feed line with balanced feed antennas.

Beamwidth—Related to directive antennas. The width, in degrees, of the major lobe between the two directions at which the relative radiated power is equal to one half its value at the peak of the lobe (half power = –3 dB).

Beta match—A form of hairpin match. The two conductors straddle the boom of the antenna being matched, and the closed end of the matching-section conductors is strapped to the boom.

Bridge—A circuit with two or more ports that is used in measurements of impedance, resistance or standing waves in an antenna system. When the bridge is adjusted for a balanced condition, the unknown factor can be determined by reading its value on a calibrated scale or meter.

Capacitance hat—A conductor of large surface area that is connected at the high-impedance end of an antenna to effectively increase the electrical length. It is sometimes mounted directly above a loading coil to reduce the required inductance for establishing resonance. It usually takes the form of a series of wheel spokes or a solid circular disc. Sometimes referred to as a "top hat."

Capture area—See aperture.

Center fed—Transmission-line connection at the electrical center of an antenna radiator.

Center loading—A scheme for inserting inductive reactance (coil) at or near the center of an antenna element for the purpose of lowering its resonant frequency. Used with elements that are less than $^1/_4$ wavelength at the operating frequency.

Coax—See coaxial cable.

Coaxial cable—Any of the coaxial transmission lines that have the outer shield (solid or braided) on the same axis as the inner or center conductor. The insulating material can be air, helium or solid-dielectric compounds.

Collinear array—A linear array of radiating elements (usually dipoles) with their axes arranged in a straight line. Popular at VHF and above.

Conductor—A metal body such as tubing, rod or wire that permits current to travel continuously along its length.

Counterpoise—A wire or group of wires mounted close to ground, but insulated from ground, to form a low-impedance, high-capacitance path to ground. Used at MF and HF to provide an RF ground for an antenna. Also see ground plane.

Current loop—A point of current maxima (antipode) on an antenna.

Current node—A point of current minima on an antenna.

Decibel—A logarithmic power ratio, abbreviated dB. May also represent a voltage or current ratio if the voltages or currents are measured across (or through) identical impedances. Suffixes to the abbreviation indicate references: dBi, isotropic radiator; dBic, isotropic radiator circular; dBm, milliwatt; dBW, watt.

Delta loop—A full-wave loop shaped like a triangle or delta.

Delta match—Center-feed technique used with radiators that are not split at the center. The feed line is fanned near the radiator center and connected to the radiator symmetrically. The fanned area is delta shaped.

Dielectrics—Various insulating materials used in antenna systems, such as found in insulators and transmission lines.

Dipole—An antenna that is split at the exact center for connection to a feed line, usually a half wavelength long. Also called a "doublet."

Direct ray—Transmitted signal energy that arrives at the receiving antenna directly rather than being reflected by any object or medium.

Directivity—The property of an antenna that concentrates the radiated energy to form one or more major lobes.

Director—A conductor placed in front of a driven element to cause directivity. Frequently used singly or in multiples with Yagi or cubical-quad beam antennas.

Doublet—See dipole.

Driven array—An array of antenna elements which are all driven or excited by means of a transmission line, usually to achieve directivity.

Driven element—A radiator element of an antenna system to which the transmission line is connected.

Dummy load—Synonymous with dummy antenna. A nonradiating substitute for an antenna.

E layer—The ionospheric layer nearest earth from which radio signals can be reflected to a distant point, generally a maximum of 2000 km (1250 ml).

E plane—Related to a linearly polarized antenna, the plane containing the electric field vector of the antenna and its direction of maximum radiation. For terrestrial antenna systems, the direction of the E plane is also taken as the polarization of the antenna. The E plane is at right angles to the H plane.

Efficiency—The ratio of useful output power to input power, determined in antenna systems by losses in the system, including in nearby objects.

EIRP—Effective isotropic radiated power. The power radiated by an antenna in its favored direction, taking the gain of the antenna into account as referenced to isotropic.

Elements—The conductive parts of an antenna system that determine the antenna characteristics. For example, the reflector, driven element and directors of a Yagi antenna.

End effect—A condition caused by capacitance at the ends of an antenna element. Insulators and related support wires contribute to this capacitance and lower the resonant frequency of the antenna. The effect increases with conductor diameter and must be considered when cutting an antenna element to length.

End fed—An end-fed antenna is one to which power is applied at one end, rather than at some point between the ends.

F layer—The ionospheric layer that lies above the E layer. Radio waves can be refracted from it to provide communications distances of several thousand miles by means of single- or double-hop skip.

Feed line—See feeders.

Feeders—Transmission lines of assorted types that are used to route RF power from a transmitter to an antenna, or from an antenna to a receiver.

Field strength—The intensity of a radio wave as measured at a point some distance from the antenna. This measurement is usually made in microvolts per meter.

Front to back—The ratio of the radiated power off the front and back of a directive antenna. For example, a dipole would have a ratio of 1, which is equivalent to 0 dB.

Front to rear—Worst-case rearward lobe in the 180°-wide sector behind an antenna's main lobe, in dB.

Front to side—The ratio of radiated power between the major lobe and that 90° off the front of a directive antenna.

Gain—The increase in effective radiated power in the desired direction of the major lobe.

Gamma match—A matching system used with driven antenna elements to effect a match between the transmission line and the feed point of the antenna. It consists of a series capacitor and an arm that is mounted close to the driven element and in parallel with it near the feed point.

Ground plane—A system of conductors placed beneath an elevated antenna to serve as an earth ground. Also see counterpoise.

Ground screen—A wire mesh counterpoise.

Ground wave—Radio waves that travel along the earth's surface.

H plane—Related to a linearly polarized antenna. The plane containing the magnetic field vector of an antenna and its direction of maximum radiation. The H plane is at right angles to the E plane.

HAAT—Height above average terrain. A term used mainly in connection with repeater antennas in determining coverage area.

Hairpin match—A U-shaped conductor that is connected to the two inner ends of a split dipole for the purpose of creating an impedance match to a balanced feeder.

Harmonic antenna—An antenna that will operate on its fundamental frequency and the harmonics of the fundamental frequency for which it is designed. An end-fed half-wave antenna is one example.

Helical—A helically wound antenna, one that consists of a spiral conductor. If it has a very large winding length to diameter ratio it provides broadside radiation. If the length-to-diameter ratio is small, it will operate in the axial mode and radiate off the end opposite the feed point. The polarization will be circular for the axial mode, with left or right circularity, depending on whether the helix is wound clockwise or counterclockwise.

Helical hairpin—"Hairpin" match with a lumped inductor, rather than parallel-conductor line.

Image antenna—The imaginary counterpart of an actual antenna. It is assumed for mathematical purposes to be located below the earth's surface beneath the antenna, and is considered symmetrical with the antenna above ground.

Impedance—The ohmic value of an antenna feed point, matching section or transmission line. An impedance may contain a reactance as well as a resistance component.

Inverted V—A misnomer, as the antenna being referenced does not have the characteristics of a V antenna. See inverted-V dipole.

Inverted-V dipole—A half-wavelength dipole erected in the form of an upside-down V, with the feed point at the apex. Its radiation pattern is similar to that of a horizontal dipole.

Isotropic—An imaginary or hypothetical point-source antenna that radiates equal power in all directions. It is used as a reference for the directive characteristics of actual antennas.

Lambda—Greek symbol (λ) used to represent a wavelength with reference to electrical dimensions in antenna work.

Line loss—The power lost in a transmission line, usually expressed in decibels.

Line of sight—Transmission path of a wave that travels directly from the transmitting antenna to the receiving antenna.

Litz wire—Stranded wire with individual strands insulated; small wire provides a large surface area for current flow, so losses are reduced for the wire size.

Load—The electrical entity to which power is delivered. The antenna system is a load for the transmitter.

Loading—The process of a transferring power from its source to a load. The effect a load has on a power source.

Lobe—A defined field of energy that radiates from a directive antenna.

Log periodic antenna—A broadband directive antenna that has a structural format causing its impedance and radiation characteristics to repeat periodically as the logarithm of frequency.

Long wire—A wire antenna that is one wavelength or greater in electrical length. When two or more wavelengths long it provides gain and a multilobe radiation pattern. When terminated at one end it becomes essentially unidirectional off that end.

Marconi antenna—A shunt-fed monopole operated against ground or a radial system. In modern jargon, the term refers loosely to any type of vertical antenna.

Matching—The process of effecting an impedance match between two electrical circuits of unlike impedance. One example is matching a transmission line to the feed point of an antenna. Maximum power transfer to the load (antenna system) will occur when a matched condition exists.

Monopole—Literally, one pole, such as a vertical radiator operated against the earth or a counterpoise.

Nichrome wire—An alloy of nickel and chromium; not a good conductor; resistance wire. Used in the heating elements of electrical appliances; also as conductors in transmission lines or circuits where attenuation is desired.

Null—A condition during which an electrical unit is at a minimum. A null in an antenna radiation pattern is a point in the 360-degree pattern where a minima in field intensity is observed. An impedance bridge is said to be "pulled" when it has been brought into balance, with a null in the current flowing through the bridge arm.

Octave—A musical term. As related to RF, frequencies having a 2:1 harmonic relationship.

Open-wire line—A type of transmission line that resembles a ladder, sometimes called "ladder line." Consists of parallel, symmetrical wires with insulating spacers at regular intervals to maintain the line spacing. The dielectric is principally air, making it a low-loss type of line.

Parabolic reflector—An antenna reflector that is a portion of a parabolic revolution or curve. Used mainly at UHF and higher to obtain high gain and a relatively narrow beamwidth when excited by one of a variety of driven elements placed in the plane of and perpendicular to the axis of the parabola.

Parasitic array—A directive antenna that has a driven element and at least one independent director or reflector, or a combination of both. The directors and reflectors are not connected to the feed line. Except for VHF and UHF arrays with long booms (electrically), more than one reflector is seldom used. A Yagi antenna is one example of a parasitic array.

Phasing lines—Sections of transmission line that are used to ensure the correct phase relationship between the elements of a driven array, or between bays of an array of antennas. Also used to effect impedance transformations while maintaining the desired phase.

Polarization—The sense of the wave radiated by an antenna. This can be horizontal, vertical, elliptical or circular (left or right hand circularity), depending on the design and application. (See H plane.)

Q section—Term used in reference to transmission-line matching transformers and phasing lines.

Quad—A parasitic array using rectangular or diamond shaped full-wave wire loop elements. Often called the "cubical quad." Another version uses delta-shaped elements, and is called a delta loop beam.

Radiation pattern—The radiation characteristics of an antenna as a function of space coordinates. Normally, the pattern is measured in the far-field region and is represented graphically.

Radiation resistance—The ratio of the power radiated by an antenna to the square of the RMS antenna current, referred to a specific point and assuming no losses. The effective resistance at the antenna feed point.

Radiator—A discrete conductor that radiates RF energy in an antenna system.

Random wire—A random length of wire used as an antenna and fed at one end by means of an antenna tuner. Seldom operates as a resonant antenna unless the length happens to be correct.

Reflected ray—A radio wave that is reflected from the earth, ionosphere or a man-made medium, such as a passive reflector.

Reflector—A parasitic antenna element or a metal assembly that is located behind the driven element to enhance forward directivity. Hillsides and large man-made structures such as buildings and towers may act as reflectors.

Refraction—Process by which a radio wave is bent and returned to earth from an ionospheric layer or other medium after striking the medium.

Resonator—In antenna terminology, a loading assembly consisting of a coil and a short radiator section. Used to lower the resonant frequency of an antenna, usually a vertical or a mobile whip.

Rhombic—A rhomboid or diamond-shaped antenna consisting of sides (legs) that are each one or more wavelengths long. The antenna is usually erected parallel to the ground. A rhombic antenna is bidirectional unless terminated by a resistance, which makes it unidirectional. The greater the electrical leg length, the greater the gain, assuming the tilt angle is optimized.

Shunt feed—A method of feeding an antenna driven element with a parallel conductor mounted adjacent to a low-impedance point on the radiator. Frequently used with grounded quarter-wave vertical antennas to provide an impedance match to the feeder. Series feed is used when the base of the vertical is insulated from ground.

Stacking—The process of placing similar directive antennas atop or beside one another, forming a "stacked array." Stacking provides more gain or directivity than a single antenna.

Stub—A section of transmission line used to tune an antenna element to resonance or to aid in obtaining an impedance match.

SWR—Standing-wave ratio on a transmission line in an antenna system. More correctly, VSWR, or voltage standing-wave ratio. The ratio of the forward to reflected voltage on the line, and not a power ratio. A VSWR of 1:1 occurs when all parts of the antenna system are matched correctly to one another.

T match—Method for matching a transmission-line to an unbroken driven element. Attached at the electrical center of the driven element in a T-shaped manner. In effect it is a double gamma match.

Tilt angle—Half the angle included between the wires at the sides of a rhombic antenna.

Top hat—See capacitance hat.

Top loading—Addition of a reactance (usually a capacitance hat) at the end of an antenna element opposite the feed point to increase the electrical length of the radiator.

Transmatch—An antenna tuner.

Trap—Parallel L-C network inserted in an antenna element to provide multiband operation with a single conductor.

Unipole—See monopole.

Velocity factor—The ratio of the velocity of radio wave propagation in a dielectric medium to that in free space. When cutting a transmission line to a specific electrical length, the velocity factor of the particular line must be taken into account.

VSWR—Voltage standing-wave ratio. See SWR.

Wave—A disturbance or variation that is a function of time or space, or both, transferring energy progressively from point to point. A radio wave, for example.

Wave angle—The angle above the horizon of a radio wave as it is launched from or received by an antenna. Also called elevation angle.

Wave front—A surface that is a locus of all the points having the same phase at a given instant in time.

Yagi—A directive, gain type of antenna that utilizes a number of parasitic directors and a reflector. Named after one of the two Japanese inventors (Yagi and Uda).

Zepp antenna—A half-wave wire antenna that operates on its fundamental and harmonics. It is fed at one end by means of open-wire feeders. The name evolved from its popularity as an antenna on Zeppelins. In modern jargon the term refers loosely to any horizontal antenna.

Abbreviations

Abbreviations and acronyms that are commonly used throughout this book are defined in the list below. Periods are not part of an abbreviation unless the abbreviation otherwise forms a common English word. When appropriate, abbreviations as shown are used in either singular or plural construction.

-A-
A—ampere
ac—alternating current
AF—audio frequency
AFSK—audio frequency-shift keying
AGC—automatic gain control
AM—amplitude modulation
ANT—antenna
ARRL—American Radio Relay League
ATV—amateur television
AWG—American wire gauge
az-el—azimuth-elevation

-B-
balun—balanced to unbalanced
BC—broadcast
BCI—broadcast interference
BW—bandwidth

-C-
ccw—counterclockwise
cm—centimeter
coax—coaxial cable
CT—center tap
cw—clockwise
CW—continuous wave

-D-
D—diode
dB—decibel
dBd—decibels referenced to a dipole
dBi—decibels referenced to isotropic
dBic—decibels referenced to isotropic, circular
dBm—decibels referenced to one milliwatt
dBW—decibels referenced to one watt
dc—direct current
deg—degree
DF—direction finding
dia—diameter
DPDT—double pole, double throw
DPST—double pole, single throw
DVM—digital voltmeter
DX—long distance communication

-E-
E—ionospheric layer, electric field
ed.—edition
Ed.—editor
EIRP—effective isotropic radiated power
ELF—extremely low frequency
EMC—electromagnetic compatibility
EME—earth-moon-earth
EMF—electromotive force
ERP—effective radiated power
E_S—ionospheric layer (sporadic E)

-F-
f—frequency
F—ionospheric layer, farad
F/B—front to back (ratio)
F/R—worst-case front to rear (ratio)
FM—frequency modulation
FOT—frequency of optimum transmission
ft—foot or feet (unit of length)
F_1—ionospheric layer
F_2—ionospheric layer

-G-
GDO—grid- or gate-dip oscillator
GHz—gigahertz
GND—ground

-H-
H—magnetic field, henry
HAAT—height above average terrain
HF—high frequency (3-30 MHz)
Hz—hertz (unit of frequency)

-I-
I—current
ID—inside diameter
IEEE—Institute of Electrical and Electronic Engineers
in.—inch
IRE—Institute of Radio Engineers (now IEEE)

-J-
j—vector notation

-K-
kHz—kilohertz
km—kilometer
kW—kilowatt
kW—kilohm

-L-
L—inductance
lb—pound (unit of mass)
LF—low frequency (30-300 kHz)
LHCP—left-hand circular polarization
ln—natural logarithm
log—common logarithm
LP—log periodic
LPDA—log periodic dipole array
LPVA—log periodic V array
LUF—lowest usable frequency

-M-
m—meter (unit of length)
m/s—meters per second
mA—milliampere
max—maximum
MF—medium frequency (0.3-3 MHz)
mH—millihenry
MHz—megahertz
mi—mile
min—minute
mm—millimeter
ms—millisecond
mS—millisiemen
MS—meteor scatter
MUF—maximum usable frequency
mW—milliwatt
MW—megohm

-N-
NC—no connection, normally closed
NiCd—nickel cadmium
NIST—National Institute of Standards and Technology
NO—normally open
no.—number

-O-
OD—outside diameter

-P-
p—page (bibliography reference)
P-P—peak to peak
PC—printed circuit
PEP—peak envelope power
pF—picofarad
pot—potentiometer
pp—pages (bibliography reference)
Proc—Proceedings

-Q-
Q—figure of merit

-R-
R—resistance, resistor
RF—radio frequency
RFC—radio frequency choke
RFI—radio frequency interference
RHCP—right-hand circular polarization
RLC—resistance-inductance-capacitance
r/min—revolutions per minute
RMS—root mean square
r/s—revolutions per second
RSGB—Radio Society of Great Britain
RX—receiver

-S-
s—second
S—siemen
S/NR—signal-to-noise ratio
SASE—self-addressed stamped envelope
SINAD—signal-to-noise and distortion
SPDT—single pole, double throw
SPST—single pole, single throw
SWR—standing wave ratio
sync—synchronous

-T-
tpi—turns per inch
TR—transmit-receive
TVI—television interference
TX—transmitter

-U-
UHF—ultra-high frequency (300-3000 MHz)
US—United States
UTC—Universal Time, Coordinated

-V-
V—volt
VF—velocity factor
VHF—very-high frequency (30-300 MHz)
VLF—very-low frequency (3-30 kHz)
Vol—volume (bibliography reference)
VOM—volt-ohm meter
VSWR—voltage standing-wave ratio
VTVM—vacuum-tube voltmeter

-W-
W—watt
WPM—words per minute
WRC—World Radio Conference
WVDC—working voltage, direct current

-X-
X—reactance
XCVR—transceiver
XFMR—transformer
XMTR—transmitter

-Z-
Z—impedance

-Other symbols and Greek letters-
°—degrees
λ—wavelength
λ/dia—wavelength to diameter (ratio)
µ—permeability
µF—microfarad
µH—microhenry
µV—microvolt
Ω—ohm
ϕ—angles
π—3.14159
θ—angles

Length Conversions

Throughout this book, equations may be found for determining the design length and spacing of antenna elements. For convenience, the equations are written to yield a result in feet. (The answer may be converted to meters simply by multiplying the result by 0.3048.) If the result in feet is not an integral number, however, it is necessary to make a conversion from a decimal fraction of a foot to inches and fractions before the physical distance can be determined with a conventional tape measure. Table 1 may be used for this conversion, showing inches and fractions for increments of 0.01 foot. The table deals with only the fractional portion of a foot. The integral number of feet remains the same.

For example, say a calculation yields a result of 11.63 feet, and we wish to convert this to a length we can find on a tape measure. For the moment, consider only the fractional part of the number, 0.63 foot. In Table 1 1ocate the line with "0.6" appearing in the left column. (This is the 7th line down in the body of the table.) Then while staying on that line, move over to the column headed "0.03." Note here that the sum of the column and line heads, 0.6 + 0.03, equals the value of 0.63 that we want to convert. In the body of the table for this column and line we read the equivalent fraction for 0.63 foot, $7^{9}/_{16}$ inches. To that value, add the number of whole feet from the value being converted, 11 in this case. The total length equivalent of 11.63 feet is thus 11 feet $7^{9}/_{16}$ inches.

Similarly, Table 2 may be used to make the conversion from inches and fractions to decimal fractions of a foot. This table is convenient for using measured distances in equations. For example, say we wish to convert a length of 19 feet $7^{3}/_{4}$ inches to a decimal fraction. Considering only the fractional part of this value, $7^{3}/_{4}$ inches, locate the decimal value on the line identified as "7-" and in the column headed "$^{3}/_{4}$," where we read 0.646. This decimal value is equivalent to $7 + {}^{3}/_{4} = 7^{3}/_{4}$ inches. To this value add the whole number of feet from the value being converted for the final result, 19 in this case. In this way, 19 feet $7^{3}/_{4}$ inches converts to 19 + 0.646 = 19.646 feet.

Table 2
Conversion, Inches and Fractions to Decimal Feet

	Fractional Increments							
	0	*$^{1}/_{8}$*	*$^{1}/_{4}$*	*$^{3}/_{8}$*	*$^{1}/_{2}$*	*$^{5}/_{8}$*	*$^{3}/_{4}$*	*$^{7}/_{8}$*
0-	0.000	0.010	0.021	0.031	0.042	0.052	0.063	0.073
1-	0.083	0.094	0.104	0.115	0.125	0.135	0.146	0.156
2-	0.167	0.177	0.188	0.198	0.208	0.219	0.229	0.240
3-	0.250	0.260	0.271	0.281	0.292	0.302	0.313	0.323
4-	0.333	0.344	0.354	0.365	0.375	0.385	0.396	0.406
5-	0.417	0.427	0.438	0.448	0.458	0.469	0.479	0.490
6-	0.500	0.510	0.521	0.531	0.542	0.552	0.563	0.573
7-	0.583	0.594	0.604	0.615	0.625	0.635	0.646	0.656
8-	0.667	0.677	0.688	0.698	0.708	0.719	0.729	0.740
9-	0.750	0.760	0.771	0.781	0.792	0.802	0.813	0.823
10-	0.833	0.844	0.854	0.865	0.875	0.885	0.896	0.906
11-	0.917	0.927	0.938	0.948	0.958	0.969	0.979	0.990

Table 1
Conversion, Decimal Feet to Inches (Nearest 16th)

	Decimal Increments									
	0.00	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
0.0	0-0	$0^{1}/_{8}$	$0^{1}/_{4}$	$0^{3}/_{8}$	$0^{1}/_{2}$	$0^{5}/_{8}$	$0^{3}/_{4}$	$0^{13}/_{16}$	$0^{15}/_{16}$	$1^{1}/_{16}$
0.1	$1^{3}/_{16}$	$1^{5}/_{16}$	$1^{7}/_{16}$	$1^{9}/_{16}$	$1^{11}/_{16}$	$1^{13}/_{16}$	$1^{15}/_{16}$	$2^{1}/_{16}$	$2^{3}/_{16}$	$2^{1}/_{4}$
0.2	$2^{3}/_{8}$	$2^{1}/_{2}$	$2^{5}/_{8}$	$2^{3}/_{4}$	$2^{7}/_{8}$	3-0	$3^{1}/_{8}$	$3^{1}/_{4}$	$3^{3}/_{8}$	$3^{1}/_{2}$
0.3	$3^{5}/_{8}$	$3^{3}/_{4}$	$3^{13}/_{16}$	$3^{15}/_{16}$	$4^{1}/_{16}$	$4^{3}/_{16}$	$4^{5}/_{16}$	$4^{7}/_{16}$	$4^{9}/_{16}$	$4^{11}/_{16}$
0.4	$4^{13}/_{16}$	$4^{15}/_{16}$	$5^{1}/_{16}$	$5^{3}/_{16}$	$5^{1}/_{4}$	$5^{3}/_{8}$	$5^{1}/_{2}$	$5^{5}/_{8}$	$5^{3}/_{4}$	$5^{7}/_{8}$
0.5	6-0	$6^{1}/_{8}$	$6^{1}/_{4}$	$6^{3}/_{8}$	$6^{1}/_{2}$	$6^{5}/_{8}$	$6^{3}/_{4}$	$6^{13}/_{16}$	$6^{15}/_{16}$	$7^{1}/_{16}$
0.6	$7^{3}/_{16}$	$7^{5}/_{16}$	$7^{7}/_{16}$	$7^{9}/_{16}$	$7^{11}/_{16}$	$7^{13}/_{16}$	$7^{15}/_{16}$	$8^{1}/_{16}$	$8^{3}/_{16}$	$8^{1}/_{4}$
0.7	$8^{3}/_{8}$	$8^{1}/_{2}$	$8^{5}/_{8}$	$8^{3}/_{4}$	$8^{7}/_{8}$	9-0	$9^{1}/_{8}$	$9^{1}/_{4}$	$9^{3}/_{8}$	$9^{1}/_{2}$
0.8	$9^{5}/_{8}$	$9^{3}/_{4}$	$9^{13}/_{16}$	$9^{15}/_{16}$	$10^{1}/_{16}$	$10^{3}/_{16}$	$10^{5}/_{16}$	$10^{7}/_{16}$	$10^{9}/_{16}$	$10^{11}/_{16}$
0.9	$10^{13}/_{16}$	$10^{15}/_{16}$	$11^{1}/_{16}$	$11^{3}/_{16}$	$11^{1}/_{4}$	$11^{3}/_{8}$	$11^{1}/_{2}$	$11^{5}/_{8}$	$11^{3}/_{4}$	$11^{7}/_{8}$

Metric Equivalents

Throughout this book, distances and dimensions are usually expressed in English units—the mile, the foot, and the inch. Conversions to metric units may be made by using the following equations:

km = mi × 1.609

m = ft (′) × 0.3048

mm = in. (″) × 25.4

An inch is $^1/_{12}$ of a foot. Tables in the previous section provide information for accurately converting inches and fractions to decimal feet, and vice versa, without the need for a calculator.

Gain Reference

Throughout this book, gain is referenced to an isotropic radiator (dBi) or to an isotropic radiator with circular polarization (dBic).

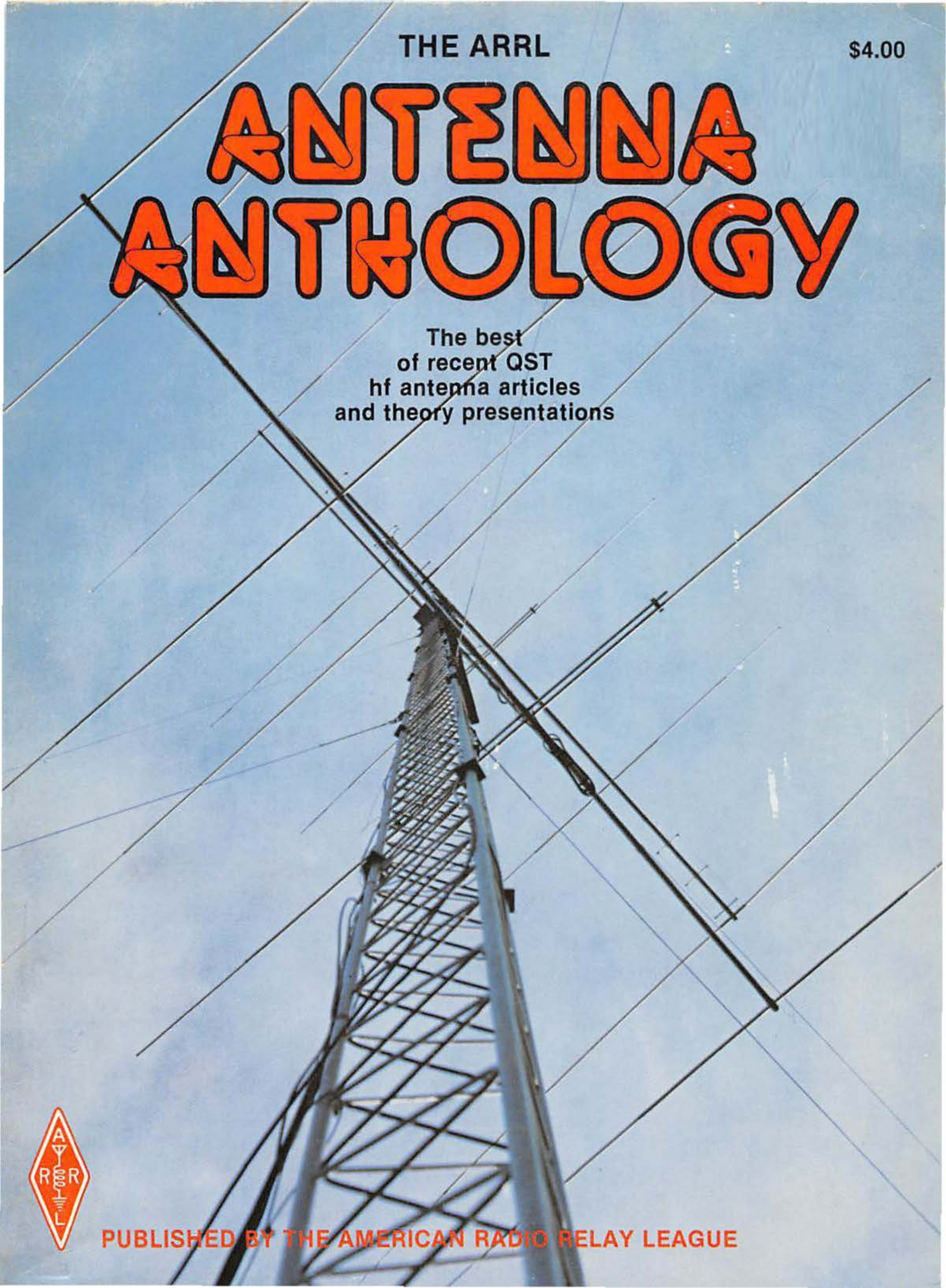

THE ARRL

$4.00

ANTENNA ANTHOLOGY

The best
of recent QST
hf antenna articles
and theory presentations

PUBLISHED BY THE AMERICAN RADIO RELAY LEAGUE

The ARRL
Antenna Anthology

Edited by
Marian S. Anderson, WB1FSB

American Radio Relay League, Inc.
Newington, CT 06111

Copyright © 1978 by

The American Radio Relay League, Inc.

Copyright secured under the Pan-American Convention

International Copyright secured

This work is Publication No. 36 of the Radio Amateur's Library, published by the League. All rights reserved. No part of this work may be reproduced in any form except by written permission of the publisher. All rights of translation are reserved.

Printed in USA

Quedan reservados todos los derechos

Library of Congress Catalog Card Number: 78-71955

$4.00 in USA
$4.50 elsewhere

Foreword

What type of amateur antenna best suits your needs? Does the available space on your property permit the installation of an effective DX antenna? Is there a compromise antenna system that will provide acceptable results over short and long distances while being reasonable in price and modest in complexity? No doubt you have come to grips with these matters repeatedly since obtaining your amateur license. So, where does one turn to find the answers to these sometimes perplexing questions? What one pre-eminent source might one consult among all of the antenna booklets published for amateurs? Why, *QST*, naturally!

This anthology contains what the ARRL hq. staff considers the best of the *QST* antenna articles and theory presentations in recent years. The contributions of numerous amateur authors and designers are contained in this volume for your convenience. This publication was developed not only to serve as a compilation of good articles, but to reach those amateurs who do not have access to past issues of the League's official monthly journal, *QST*.

The editors of this booklet have attempted to select the types of antennas that have greatest appeal to those who like to construct their own radiators, and for the many amateurs who can't justify the expense of purchasing commercial counterparts. A book of this type contains something for everyone who enjoys building and experimenting with antennas. But, if you aren't a builder you will still enjoy the hours of interesting reading provided by these works.

We wish to extend our gratitude to those *QST* authors whose articles have been selected for publication in this volume. Their enthusiasm, ability and generosity typifies clearly what Amateur Radio is all about. Should you have some original antenna ideas of proven performance, send an article to the ARRL, describing your work. Not only may your story be published in *QST* — it might appear in a future edition of this book.

Richard L. Baldwin, W1RU
General Manager

Contents

1 Vertical Antennas
Page 7

2 Yagi Antennas
Page 41

3 Quad Antennas
Page 53

4 Miscellaneous Antenna Types
Page 76

5 Antenna Theory and Test Methods
Page 100

Index
Page 150

Chapter 1

Vertical Antennas

One of the most practical antennas for use where limited space exists is the vertical. This type of antenna is particularly useful to amateurs who live in dense urban areas where city lots are traditionally small. The same general restriction prevails for those who dwell in apartment buildings. In such locations it is often easier to go "up" with an antenna than to erect one horizontally. But the virtues of vertical antennas extend beyond the foregoing considerations: "Verticals," as they often are called, are relatively inexpensive, easy to erect and are capable of performing well for local (stateside) and DX work. A vertical antenna exhibits a low angle of radiation when worked against a good ground system. This makes it an effective DX type of antenna.

This chapter contains reprints of some popular vertical-antenna systems that have appeared in *QST* during recent years. For the most part, the presentations contained herein have been "lifted" in their entirety, thus making it unnecessary for the reader to research the original articles for additional information.

A Two-Band Vertical for the Novice

Build this inexpensive Novice antenna. It is easy to construct and is made from readily available materials.

Knowing of a Novice who had space problems, the author set out to design an antenna for him that could be used on 40 and 15 meters. His previous antenna was a horizontal "V" whose performance wasn't worth the time it took to put it up. Because of his narrow property, the angle between the two wires was only around 40 degrees, no doubt causing considerable signal cancelation. Horizontal antennas were ruled out, so the best antenna for him was a vertical.

Knowing (or caring) little about multiband trap-loaded antennas, it was decided the simplest antenna would be a quarter-wavelength vertical on 40 meters. After a little paper work, and using formulas given in *The ARRL Antenna Book*, the correct length would be 31 feet 10 inches (9.70 m) for 7.175 MHz. This length would also work as a 3/4-wavelength antenna on 15 meters, providing two-band coverage. However, a 32-foot (9.75-m) vertical antenna made roof-mounting too difficult! Mounting the vertical to the side of the house alleviated many guying and support problems.

Materials

The first concern was what to use for antenna materials. Being almost 32 feet high, the vertical would need strength, so the writer thought of steel tubing, but that was much too expensive. With conduit there aren't enough sizes offered, and that left only one more material — aluminum tubing. The latter comes in just about any size desired, and the author had no trouble in finding just what he wanted. Going to the nearest aluminum dealer, the author found his stock to be more than ample.[1]

In order to keep the cost minimal, the author decided to make the base piece from 1-inch (25-mm) diameter thin-wall tubing. Into 12 feet (3.66 m) of this was inserted a 12-foot piece of 7/8-inch (22-mm) diameter thin-wall tubing. This piece was telescoped 2 feet (0.610 m), then bolted and clamped (after the end had been slit twice by a pair of tin snips). It was necessary to use hose clamps on the tubing because the galvanized bolts didn't provide a good bond between the pieces. The top section was 13-1/2-foot (4.11-m) piece of 3/4-inch (19-mm) diameter thin-wall tubing inserted far enough into the lower piece to make the overall antenna length about 31 feet, 10 inches (9.70 m). Because of its large size, all the work must be done while the antenna is

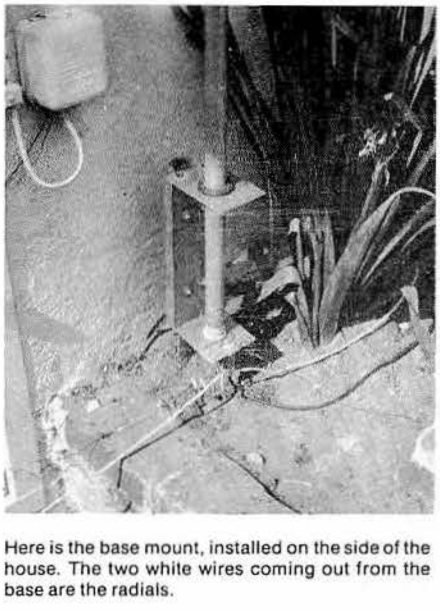

Here is the base mount, installed on the side of the house. The two white wires coming out from the base are the radials.

---

[1] [Editor's Note: TV mast material is also well suited for verticals and comes in 10-foot (3.05-m) lengths, either steel or, preferably, aluminum.]

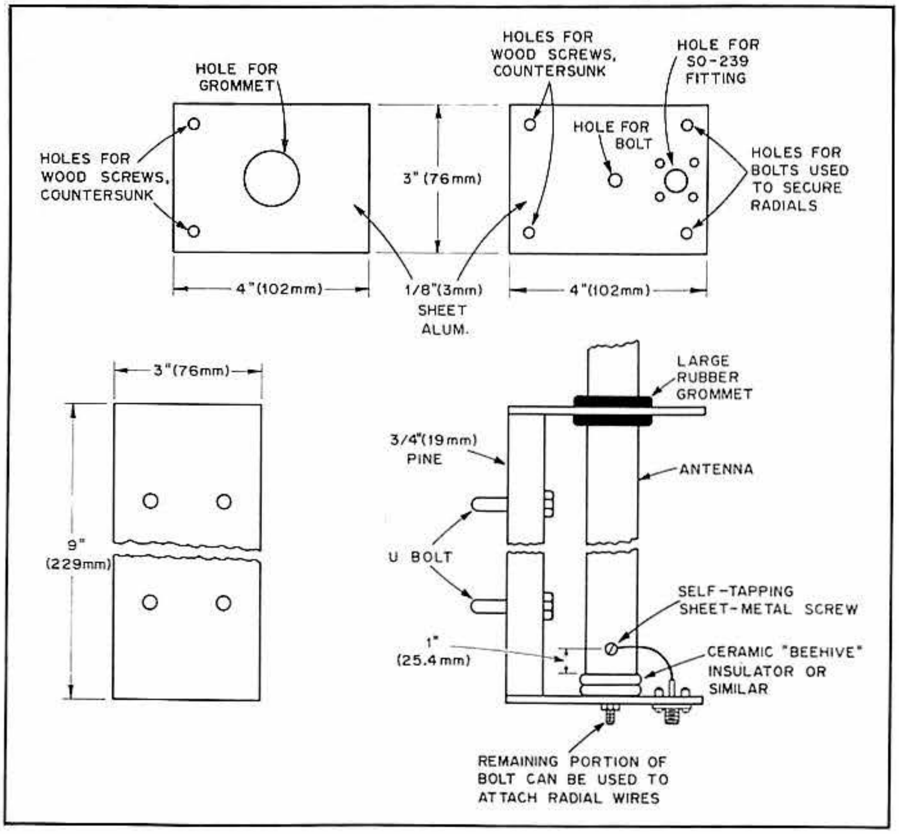
HOLE FOR GROMMET
HOLES FOR WOOD SCREWS, COUNTERSUNK
HOLES FOR WOOD SCREWS, COUNTERSUNK
3" (76mm)
HOLE FOR SO-239 FITTING
HOLE FOR BOLT
HOLES FOR BOLTS USED TO SECURE RADIALS
4"(102mm)
1/8"(3mm) SHEET ALUM.
4"(102mm)
3"(76mm)
9" (229mm)
LARGE RUBBER GROMMET
3/4"(19mm) PINE
ANTENNA
U BOLT
SELF-TAPPING SHEET-METAL SCREW
1" (25.4mm)
CERAMIC "BEEHIVE" INSULATOR OR SIMILAR
REMAINING PORTION OF BOLT CAN BE USED TO ATTACH RADIAL WIRES

Fig. 1 — Details for making the base-mounting section of the vertical.

lying on the ground. Then, the antenna must be disassembled and put up piece by piece. (Do not attempt to put the antenna up in one piece! While the weight isn't excessive, the tubing cannot stand the strain.)

Putting It Up

The original base bracket was made from 3/4-inch (19-mm) pine, but during the first rain the wood got wet and the bracket turned out to be a very poor insulator. So the unit was redesigned to use two pieces of 1/8-inch (3-mm) thick sheet aluminum, a large grommet, and a ceramic standoff insulator (see Fig. 1).

Assembling the two-band vertical was quite simple. First, a 4-foot (1.22-m) stake was driven into the ground about 2-1/2 feet (0.76 m), and about four or five inches (102 or 127 mm) away from the side of the house. Next, the bottom section (with the bracket attached) was slipped over the stake and tightened in place with the U bolts. Then the second section was inserted into the first, clamped and bolted in place. A support bracket was made from scrap aluminum, and a large grommet was installed on it to insulate the antenna from the brace (see Fig. 2). This assembly was then attached to the side of the house (near the gable) with wood screws. Last, the third section was forced through the grommet in the second bracket, and was then inserted into the second piece and clamped. (If adjustment of resonance is desired, it is better not to bolt the final section). A suitable cap was mounted on top of the antenna to keep out moisture.

Tune Up

RG-58/U coax was then connected between the transmitter and antenna with an SWR bridge in the line, and power was applied at 7175 kHz. The SWR was found to be 3:1. This was without any ground system on the antenna. In order to obtain a better match and reduce the SWR, it was apparent that a ground system was required.

According to *The ARRL Antenna Book*, two methods of making a suitable ground system can be used. One way is to connect the coax outer shield directly to ground at the base of the antenna, via a ground rod. However, this doesn't work too well unless the soil is extremely conductive (his wasn't). The other method is to make up some sort of artificial ground in the form of radials.

The author cut a radial from some surplus wire to the same length as the vertical section, attached the radial to the base plate and the SWR dropped to 2:1. Unknown to the author at the time, *The Antenna Book* says that the radials should be 2.5 percent longer than the vertical itself, so actually, they should have been made 32-feet 7-3/4-inches (9.95-m) long. A second radial, similar to the first, was attached and the SWR dropped further (to about 1.4:1). When the soil was wet, such as after a rain or after watering, the SWR takes a nose dive to less than 1.2:1. Since the property space was limited, two radials were all that could be used. Ideally, the most perfect system would consist of a metal disk, one quarter wavelength in radius, buried a few inches below the earth's surface, but this approach is rather impractical! But if the antenna was to be erected in an area *away* from buildings, and guyed, a ground system could be made of chicken wire or other similar material. However, the more radials, the better the artificial ground. Also, they should be laid out in as nearly a spoke-like pattern as possible. Hy-Gain recommended for their AVQ series antennas that a radial system be used consisting of at least two radials for *each* band used. However, for this installation the 40-meter radials operate on their third harmonic at 15 meters, requiring but one set of wires.[2]

The length specified makes the antenna resonant at about 7175 kHz on 40 meters. On 15 meters, at the third harmonic, the antenna is resonant at about 21,425 kHz, but power was applied as low as 21,105 kHz,

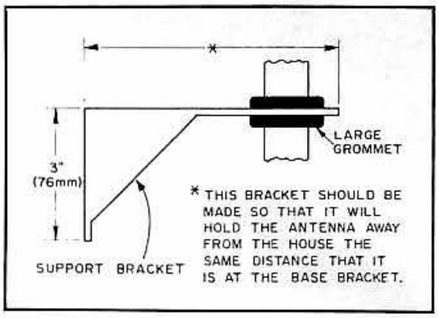
*
LARGE GROMMET
3" (76mm)
* THIS BRACKET SHOULD BE MADE SO THAT IT WILL HOLD THE ANTENNA AWAY FROM THE HOUSE THE SAME DISTANCE THAT IT IS AT THE BASE BRACKET.
SUPPORT BRACKET

Fig. 2 — Details for the support bracket.

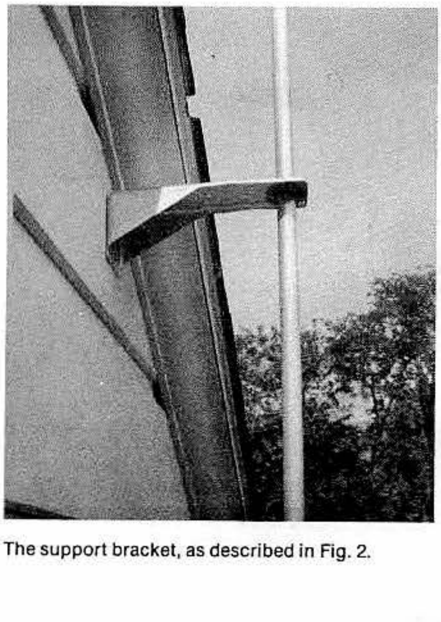

The support bracket, as described in Fig. 2.

---

[2] [Editor's note: Actually, *any* radials, regardless of length, will help the performance of the system. Also when used on 15 meters, the antenna would be three quarter wavelengths long with low-impedance feed, approximately 50 ohms. However, the angle of radiation from the antenna will be higher than that of a quarter-wavelength vertical.]

and the SWR appeared to be only 1.4:1, well within tolerable limits. Bandwidth on both Novice bands was essentially flat.

Additional Information

Total cost will depend greatly on the individual's ability to scrounge parts. If the builder prefers to purchase the principal materials, this list will be useful.

3 — pieces thin-wall aluminum tubing, 1- (25-mm), 7/8- (22-mm), and 3/4-inch (19-mm) diameters. 12 (3.66 m), 12 (3.66 m) and 13-1/2 feet (4.11 m) long, respectively.
2 — U bolts.
1—SO-239 connector.

The other material, including three bolts and nuts, six wood screws, two grommets, two hose clamps, wood and aluminum for brackets, ground stake, and cap will vary from no cost to a few dollars, dependent on available materials. If used or surplus tubing is available one could save a substantial amount of money.

In this instance, the top of the antenna extends about 16 feet (4.88 m) above the peak of the roof, but because of the bracing no guy wires were necessary. If the antenna is to be erected in a high-wind area, guys made from nonconducting monofilament fishing line (or other nonconducting line) could be used.

The owner has received nothing but glowing reports so far. Running about 70 watts input, the first CQ on 40 meters raised W9SVJ in northern Indiana from Glendale, CA. His report, RST 589. On 15 meters the owner has worked ZL3JC, AX7SM and JH1EZZ, and received very good reports.

This vertical has no gain, and probably couldn't come near a Yagi or a quad in performance, but for an antenna costing less than $10 it does one heck of an FB job. This material was originally presented in *QST* by Jerry Arnold, WA6MBP.

Build This Novice Four-Band Vertical

Putting your first amateur station together can be an expensive proposition. One way to cut costs is to keep the antenna simple. Here's how to shave the price and provide four-band operation.

Is operation with one antenna acceptable if it covers the 80-, 40-, 15- and 10-meter bands? For a new Novice that's a reasonable approach. The author's backyard is smaller than that of most urban homes, so full-size dipole or inverted-V antennas were out of the question. Not owning a tower (yet!), it seemed that a ground-mounted vertical antenna would be worth trying.

After reading *The ARRL Antenna Book*, it was decided that a ground-mounted vertical antenna would be easiest to build. Some radial wires could be buried, and the metal fence which encloses the backyard could also be hooked up to enlarge the ground system. The author preferred this type of antenna to one installed above ground, because radials of specific lengths for each of the four bands would have been needed for a roof-mounted, groundplane type of vertical. The buried wires for the ground-mounted antenna could be any convenient length, as long as the available space would permit. From what was read about these antennas, the author believed that reasonable performance could be had even if the ground radials weren't numerous and long, although generally the more, the better.

With the help of W1FB some used aluminum tubing was purchased that would telescope together and give the author a 25-foot (7.62-meter) antenna. The wall thickness of the tubing is 0.058 inch (1.5 mm). Three 10-foot (3.05 m) sections are used. The largest diameter is 1 inch (25 mm). The center telescoping section has a diameter of 7/8 inch (22 mm) and the top piece of tubing has a 3/4-inch (19-mm) diameter. This material, plus hose clamps for holding the sections together, came to $8. An old ceramic rotary switch, a coaxial connector, a feed-through bushing, and a piece of Air Dux coil stock were acquired at a flea market for an additional $3. Two medium-size, ceramic standoff insulators were donated by W1FB. He said they cost him 50 cents each at a swap session. All that remained to collect was a weatherproof box for the loading coil, some 50-ohm coaxial cable and six U bolts. The author's OM, Bob, found some used 1-1/2-inch (38-mm) steel pipe which is 7 feet (2.13 m) long. It is used as a support for the vertical.

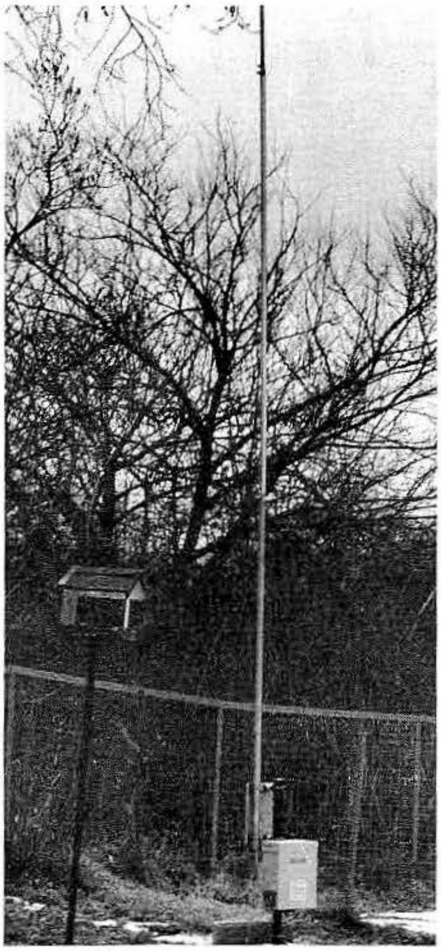
The completed 4-band vertical antenna used at WB1FSB. The box at the base of the antenna contains a loading coil.

Constructing the Antenna

A lawn-edger tool was used to make slits in the lawn, out from the base of the antenna toward the edges of the backyard. The slits were cut to a depth of 2 inches (51 mm). A total of 10 radials were buried in the slits. Some are only 15 feet (4.57 m) in length, while others are 25 feet (7.62 m) long. The metal yard fence was bonded together as needed, using wire jumpers between the fence sections. A single buried wire joined the fence to the common ground point at the base of the antenna.

The OM drove the steel pipe into the ground to a depth of 4 feet (1.22 m), leaving 3 feet (0.91 m) above ground for attaching the vertical antenna and weatherproof box. Construction details are shown in Fig. 3.

Although a wooden box could have been used to house the loading coil, switch and other hardware, an old electrical housing that the author's OM had in his junk box was used (Fig. 4.). It was drilled and

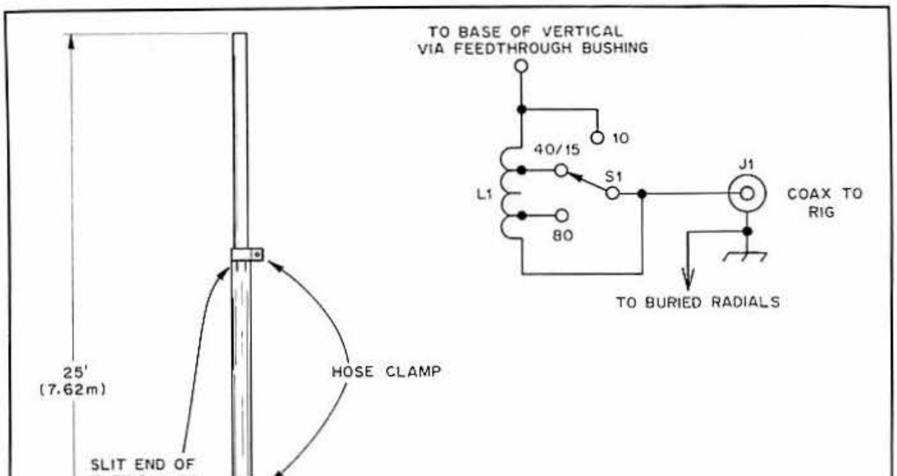
TO BASE OF VERTICAL
VIA FEEDTHROUGH BUSHING
10
40/15
S1
J1
L1
COAX TO RIG
80
TO BURIED RADIALS
25'
(7.62m)
HOSE CLAMP
SLIT END OF TUBING WITH HACK SAW (4 CUTS)
PLATES,1/4 X12X6-INCH(6X304X154mm)
SMALL METAL PLATE AND ONE U BOLT FOR EACH INSULATOR
STEEL PIPE
U BOLT
STANDOFF INSULATOR
METAL OR WOODEN BOX (ATTACHED TO PIPE WITH U BOLTS)
JUMPER WIRE TO VERTICAL
UNDERGROUND COAX CABLE TO RIG
FEEDTHROUGH BUSHING
BAND SWITCH
WIRE TO BURIED RADIALS

Fig. 3 — Dimensional drawing of the four-band Novice antenna. The top ends of the two lower tubing sections are slit four times each by means of a hack saw. This permits a tight joint when the hose clamp is compressed. The vertical is attached to one metal plate with U bolts. The large standoff insulators with metal feet connect this plate to a,pair of small ones. The latter are attached to the steel support pipe by means of two U bolts. The metal box is also affixed to the steel pipe with two U bolts. The band switch is a single-pole three-position ceramic wafer type. L1 can be a 5-inch (127-mm) length of B & W 3029 Miniductor, 2-1/2 inches (64 mm) in diameter, 6 turns per inch of no. 12 wire. See text for alternative mounting methods.

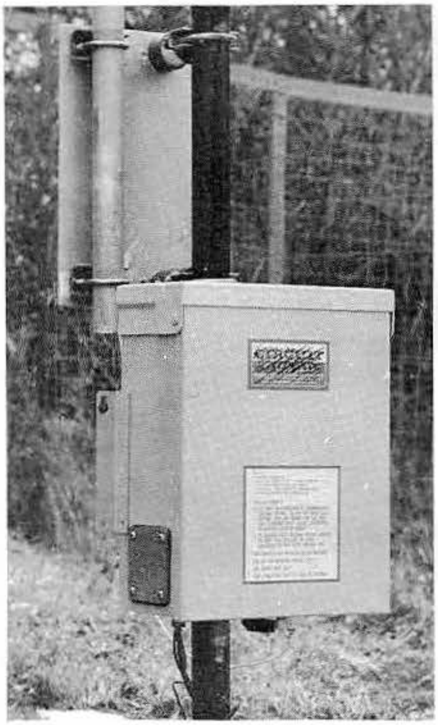

Fig. 4 — Closeup view of the base of the vertical. The aluminum tubing is affixed to a metal plate. The latter is attached to the iron support pipe by means of two surplus standoff insulators. Small aluminum plates are attached to the ends of the insulators to permit them to be fastened to the iron pipe by means of U bolts. The radial wires are connected to the bottom of the coil-housing box.

Fig. 5 — The author checks the SWR of the antenna during final adjustment of the system.

punched on the bottom surface to hold the feedthrough bushing, coaxial connector, switch and ground terminal for the radials.

W1FB designed the antenna, but he wasn't sure that an acceptable impedance match could be had on all four bands without a complex matching network. It was decided to try his idea, so the installation was completed.

Adjusting the Vertical

The author helped her OM install the antenna, then called W1FB for some assistance in tuning the system. He thought we could tune the vertical for 40 meters and make it work okay as a 3/4-wavelength vertical on 15 meters. For use on 80 meters it would be fairly short (63 feet or 19.20 m is the correct length for 3.7 MHz). With base loading it should offer adequate service out to a few hundred miles on 80 meters. Finally, it would operate as a 3/4-wavelength vertical on 10 meters.

A homemade SWR indicator was hooked in the coaxial line at the base of the vertical.[1] A small amount of transmitter power (5 W) was applied at 3725 kHz and the 80-meter switch lead was touched on the turns of the coil until minimum reflected power was indicated (Fig. 5). An SWR of 1:1 was obtained. The wire was then soldered in place on the coil. Next power was fed to the antenna on 7125 kHz, and the 40-meter switch lead was touched to the coil turns until an SWR of 1:1 was read. While using the same coil tap power was supplied to the antenna on 21.1 MHz and the SWR checked. It was approximately 3:1. By moving the coil tap just one turn we were able to get an SWR of 1.5:1 on 15 meters. A recheck on 40 meters followed. The SWR for that band was less than 2:1 — not a bad compromise! The coil was bypassed entirely for operation on 10 meters: an SWR of 2:1

---

[1] DeMaw, "A QRP Man's RF Power Meter," *QST*, June 1973.

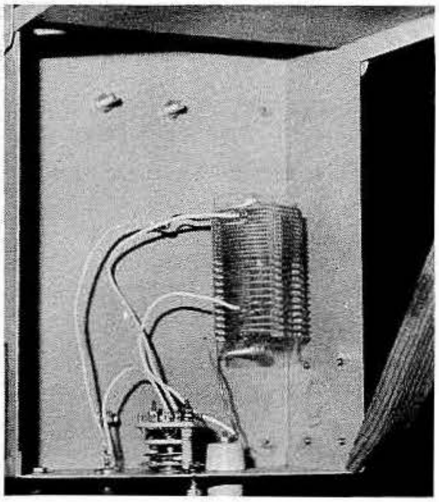

Fig. 6 — Interior view of the coil housing showing the switch, feedthrough bushing, coaxial connector, and ground post for the radials. The coil shown is a piece of Air Dux stock with a tapered pitch. It was obtained at a flea market.

was indicated at 28,000 kHz. The length of the overall antenna for operation on 28,100 kHz should be 25 feet or 7.62 m (3/4-wavelength radiator). However, the switch leads inside the coil housing add to the antenna length. If an SWR of less than 2:1 is desired, break the 10-meter switch lead and insert a 100-pF air variable capacitor. The unwanted reactance can be tuned out by this means and a low SWR will result.

Opening and closing the cover of the metal box had only a minor effect on the SWR. At last an on-the-air test of the system was possible. Fig. 6 shows the interior of the coil and switch housing.

Results

Good signal reports have been received on all bands. The first QSO on 40 meters netted an RST 599 report from North Carolina and many similar reports followed on 80, 15 and 10 meters. The author feels that a WAS award is not too far away now that this antenna is in operation.

An Alternative

There are many ways you can duplicate this design using substitute materials. For example, electrical conduit with couplers between the sections should be satisfactory in place of the aluminum tubing. The entire structure could be made from 2 x 4 inch (51 x 102-mm) lumber. If that is done, the radiator could even be a 25-foot (7.62 m) piece of no. 10 wire, supported on the side of the wood with standoff insulators.

Instead of the mounting method shown in Fig. 3, the vertical pipe could probably be inserted into a 2-foot (.61-m) length of PVC tubing, then clamped to the mounting plate. This would eliminate the need for the two standoff insulators. Better still, four or five wraps of Teflon sheeting (10 mil or 0.25 mm thickness) could be placed over the bottom end of the vertical before clamping it in place on the mounting plate. Teflon can be purchased at most plastic-supply houses.

The author hopes this idea is useful to other Novices who are trying to keep the budget within reasonable limits. This writer likes the way the antenna is working. Others should have good luck with this antenna also. Oh, by the way, the ground radials are made from various scraps of wire. The size isn't important, and they can be insulated or bare. There is quite an assortment of wire types buried in the lawn! This material was originally presented in *QST* by Marian Anderson, WB1FSB.

The Cheapie GP

Keeping an ear on 10 and 15 meters? With the trend of rising solar flux, 15 meters "opens" nearly every day — 10 meters occasionally. Don't miss these openings. Build a Cheapie GP!

Looking for an antenna to cover the 10- and 15-meter bands that won't pinch your pocketbook? Should it contain easy to obtain items and be relatively simple to construct while using ordinary hand tools? Must it be physically small and not a neighborhood eyesore? If your answer to these questions is "yup," you've turned to the right page! Detailed here is a duo-band coaxial-cable-fed, trap groundplane antenna system.

A groundplane antenna consists of a vertical quarter-wavelength radiator and an artificial metallic ground system extending radially from its base. For good results the antenna should be mounted at least one-quarter wavelength above ground. This is approximately 12 feet (3.66 m) on 15 meters. At this height and higher, only four radials for each band are necessary to provide a good ground system. If this antenna was ground mounted, many more radials would be needed to reduce earth losses to an acceptable low level. The antenna has an omnidirectional horizontal or *H*-plane pattern which means radiation from the antenna is essentially the same in all horizontal directions as shown in Fig. 7A. In Fig. 7B the approximate vertical or *E*-plane pattern of the antenna is shown for the antenna mounted one-quarter wavelength above ground. As can be seen, maximum radiation occurs at the low angles above the horizontal plane of the antenna. The antenna has an overhead null so little energy is radiated at high angles. This does not detract from the usefulness of the antenna, since DX signals seldom arrive at high angles on these bands. Instead, maximum radiation is concentrated at the *lower angles* where it will be more effective. Conversely, a horizontal dipole mounted one-quarter wavelength or less above ground will have maximum radiation straight up or nearly so. Lower-angle radiation will not occur unless the dipole is raised to a height of one-half wavelength or more above ground. If it is not possible to mount your horizontal antenna a half wavelength above ground, the vertically polarized groundplane is perhaps the logical choice over a dipole. As a general rule, the higher the groundplane can be located above power lines, gutters, house wiring and the like, the better it will perform.

A Duo-Band System

For the antenna to perform properly on two bands, it must function as a *separate* groundplane vertical on each band. This requirement is met quite simply in the case of the radials since four radials are used for each band. The 10-meter radials are 8 feet, 5 inches (2.56 m) long and the 15-meter radials are 11 feet, 7 inches (3.53 m). When the antenna is operated in the 10-meter mode, the 15-meter radials have little effect on the system. Similarly, when the antenna is used on 15 meters, the 10-meter radials do not appreciably alter the operation of the antenna. The effect of the additional radials is to change slightly the feed point resistance and the resonant frequencies of the antenna.

The vertical member of the antenna must act as a quarter-wavelength radiator on both 10 and 15 meters. To accomplish this a trap is inserted in the 15-meter vertical section. Its approximate location is a

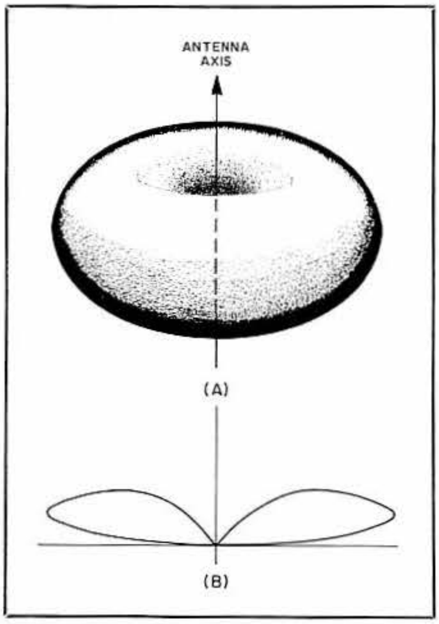

Fig. 7 — (A) Horizontal or H-plane pattern of the groundplane. (B) Approximate vertical or E-plane pattern of the antenna.

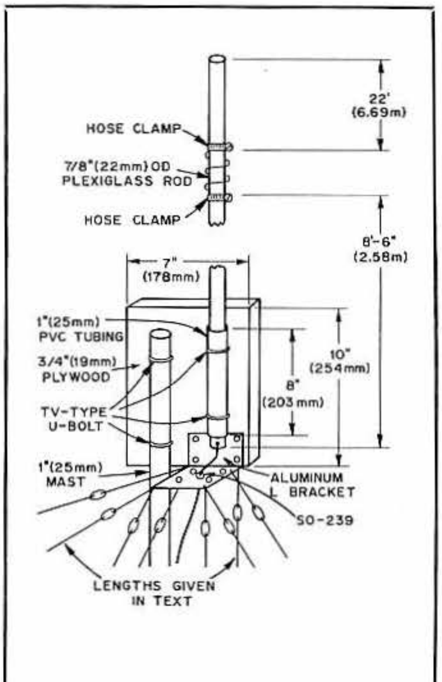

Fig. 8 — Constructional details of the duo-band antenna system.

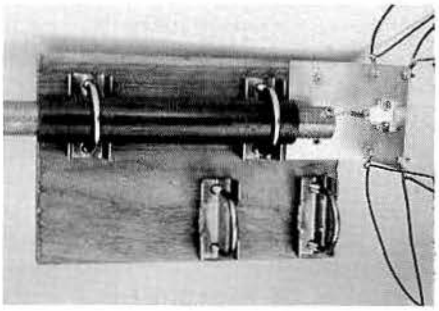

A photograph of the base assembly. The SO-239 coaxial connector and hood can be seen in the center of the aluminum L bracket. The U bolts are TV-type antenna hardware. The plywood should be coated with varnish or similar material.

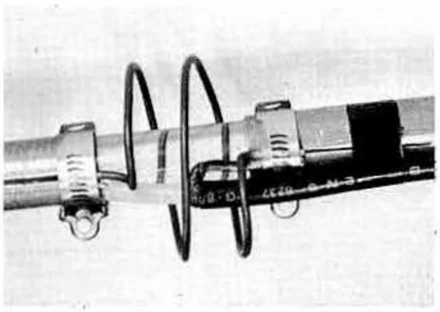

A close-up view of the trap. The leads from the coaxial-cable capacitor should be soldered directly to the pigtails of the coil. These connections should be coated with varnish after they have been secured under the hose clamps.

10-meter quarter wavelength above the base. The trap has a high impedance on 10 meters, thereby electrically divorcing the top section of the antenna when it is operated on that band. The length of the 10-meter section (section below the trap) is somewhat shorter than that of a simple 10-meter-only quarter-wavelength vertical. This is because the trap and top section of the duo-band system add top loading to the 10-meter portion of the antenna, reducing the length required for quarter-wavelength resonance. A coaxial-cable capacitor and a coil made from copper tubing form the trap which is resonant at approximately 28.150 MHz. On 15 meters the trap has a low impedance, effectively connecting together those portions of the antenna above and below the trap. The overall length of the 15-meter radiator is a bit shorter than a simple quarter-wavelength radiator would normally be for that band. This is because at 21 MHz the trap introduces inductance into the radiator — similar to the effects of a loading coil.

Construction

The vertical section of the antenna is mounted to a 3/4-inch (19-mm) thick piece of plywood board that measures 7 X 10 inches (178 mm X 254 mm). Several coats of exterior varnish or similar material will help protect the wood from inclement weather. Both the mast and the radiator are mounted to the piece of wood by means of TV U-bolt hardware. The vertical is electrically isolated from the wood with a piece of 1-inch (24-mm) diameter PVC tubing. A piece approximately 8 inches (203 mm) long is required, and it is of the schedule-80 variety. To prepare the tubing it must be slit along the entire length on one side. A hacksaw will work quite well. The PVC fits rather snugly on the aluminum tubing and will have to be "persuaded" on with the aid of a hammer. The mast is mounted directly to the wood with no insulation. An SO-239 coaxial connector and four solder lugs are mounted to an L-shaped bracket made from a piece of aluminum sheet. A short length of test probe wire, or inner conductor of RG-58/U cable, is soldered to the inner terminal of the connector. A UG-106/U connector hood is then slid over the wire and onto the coaxial connector. The hood and connector are bolted to the aluminum bracket. Two wood screws are used to secure the aluminum bracket to the plywood as shown in the drawing and photograph. The free end of the wire coming from the connector is soldered to a lug which is mounted to the bottom of the vertical radiator. Any space between the wire and where it passes through the hood is filled with GE silicone glue and seal or similar material to keep moisture out. The eight radials are soldered to the four lugs on the aluminum bracket. The two sections of the vertical member are separated by a piece of clear acrylic rod. Approximately 8 inches (203 mm) of 7/8-inch (22-mm) OD material is required. The aluminum tubing must be slit lengthwise for several inches so that the acrylic rod may be inserted. The two pieces of aluminum tubing are separated by 2-1/4 inches (57 mm).

The trap capacitor is made from RG-8/U coaxial cable and is 30.5 inches (775 mm) long. RG-8/U cable has 29.5 pF of capacitance per foot and RG-58/U has 28.5 pF per foot. RG-8/U cable is recommended over RG-58/U because of its higher breakdown-voltage characteristic. The braid should be pulled back 2-inches (51 mm) on one end of the cable, and the center conductor soldered to one end of the coil. Solder the braid to the other end of the coil. Compression-type hose clamps are placed over the capacitor/coil leads and put in position at the edges of the aluminum tubing. When tightened securely, the clamps serve a two-fold purpose — they keep the trap in contact with the vertical members and prevent the aluminum tubing from slipping off the acrylic rod. The coaxial-cable capacitor runs upward along the top section of the antenna. This is the side of the antenna to which the braid of the capacitor is connected. A cork or plastic cap should be placed in the very top of the antenna to keep moisture out.

Installation and Operation

The antenna may be mounted in position using a TV-type tripod, chimney, wall or vent mount. Alternatively, a telescoping mast or ordinary steel TV masting may be used, in which case the radials may be used as guys for the structure.

Any length of 50-ohm cable may be used to feed the antenna. The SWR at resonance should be on the order of 1.2 to 1.5:1 on both bands. The SWR is not 1 because the feedpoint resistance is something other than 50 ohms — closer to 35 or 40 ohms. Nearby metallic objects may also have an effect on the impedance of the antenna. The antenna is resonant at approximately 21.150 MHz and 28.150 MHz.

Some amateurs place too much importance on obtaining a "no reflected power" reading on their SWR indicators. Most ham transmitters will load into lines which exhibit an SWR of up to 2:1, the exception being some of the new broadband, no-tuning transmitters. To demonstrate how little effect an SWR of 1.5 or 2:1 will have on the system, consider the following: If the antenna is fed with 100 feet (30.48 m) of RG-8/U cable and if the antenna is perfectly matched to the line (an SWR of 1), the power loss along the cable will be approximately 0.98 dB on 10 meters. If the SWR on the line is increased to 1.5:1, the additional loss on the line due to the higher SWR is less than 0.1 dB. If the SWR is increased to 2:1, the additional loss over that of the matched condition will be on the order of 0.2 dB. This amount of loss is insignificant, and for this reason a complicated matching network is not necessary. This material was originally presented in *QST* by Jay Rusgrove, W1VD.

A High-Performance, 20-, 40- and 80-Meter Vertical System

Urban dwellers with small lots take heart! Here are the details for building your own DX-effective three-band mini-vertical antenna. Cost-conscious amateurs will be happy with the low cost of this system.

In another article on vertical antennas,[1] we tried to point out some fundamental characteristics of ground-mounted verticals, namely: (1) a good image plane is necessary for efficient operation, (2) a vertical over a good image plane compares favorably with a dipole at an elevation of one-half to one wavelength, (3) a short vertical compromises little in the way of performance.

This paper describes a highly efficient, three-band vertical system for 20, 40 and 80 meters, using elements of the order of an eighth wavelength. The system consists of an 80-meter vertical in parallel with a 20/40-meter trap vertical. Actually, either the 80-meter or the 20/40-meter vertical can be constructed and used alone if one is not interested in triband operation. The input impedances of both antennas are 12-1/2 ohms and they use the same 4:1 matching transformer.[2] The antennas also use the same radial system consisting of 100 radials of no. 15 aluminum wire 50 feet (15.24 m) in length (a lesser number of radials can be used as is discussed later in the article). Because of expected lower sunspot activity and, hence, poorer propagation conditions on the higher bands, the 40- and 80-meter portions of this antenna system, in particular, should prove very effective in DX communication over the next few years.

The first part of this paper deals with the design and tune-up considerations of the 80-meter element; the second part with the 20/40-meter element and the way it is used with the 80-meter vertical forming an efficient triband system. This is followed by reports on performance. Reference is also made to other alternatives for a multiband vertical system.

80-Meter Vertical

The two considerations in designing the 80-meter element of this vertical system were: (1) a good bandwidth for a reasonable height (a height one person can handle); and (2) proper spacing between the 80 and 20/40 portions such that coupling is negligible and both can be used over the same radial system.

Prior to building the shortened vertical described in this article, two others were constructed and tested on the air. One was a 22-foot (6.71 m) vertical[3] which had a 65-kHz bandwidth. The other was a trap vertical[4] with only a 20-kHz bandwidth. Both of these antennas were used separately over the same image plane of 100 radials and were efficient radiators. Therefore, the main reason in going to a slightly longer antenna was simply *to obtain a broader bandwidth.* The eighth-wavelength vertical on 40 meters from the previous work[5] promised considerable improvement by adding only a few feet. The results are shown in Figs. 9 and 10. The total height turned out to be 29 feet (8.83 m). This resulted in a bandwidth of about 140 kHz, a little more than twice the bandwidth of the 22-footer (6.71 m).

The vertical, in Fig. 9, uses a 20-foot (6.10-m) section of thick-wall aluminum tubing. (It was used some years ago as a gin pole for erecting a beam on a 40-foot (12.2 m) tower.) An 8-foot (2.44 m) extension was constructed with 1-inch (25-mm) tubing bolted in place using spacers for centering. The insulator at the bottom is phenolic tubing with a canvas base.[6] It has a 1/2-inch (13-mm) thick wall, is 9 inches

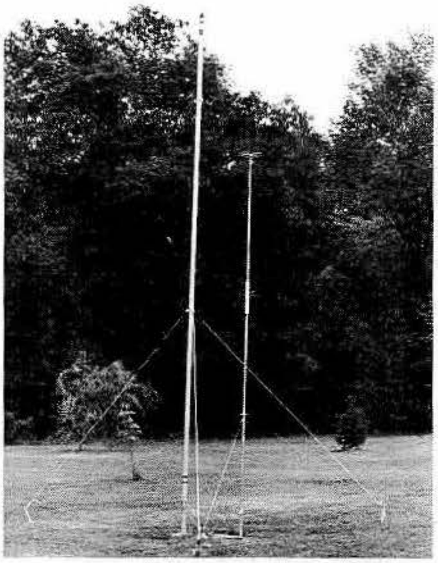

The triband vertical showing the polypropylene guys which provide an extra margin of support.

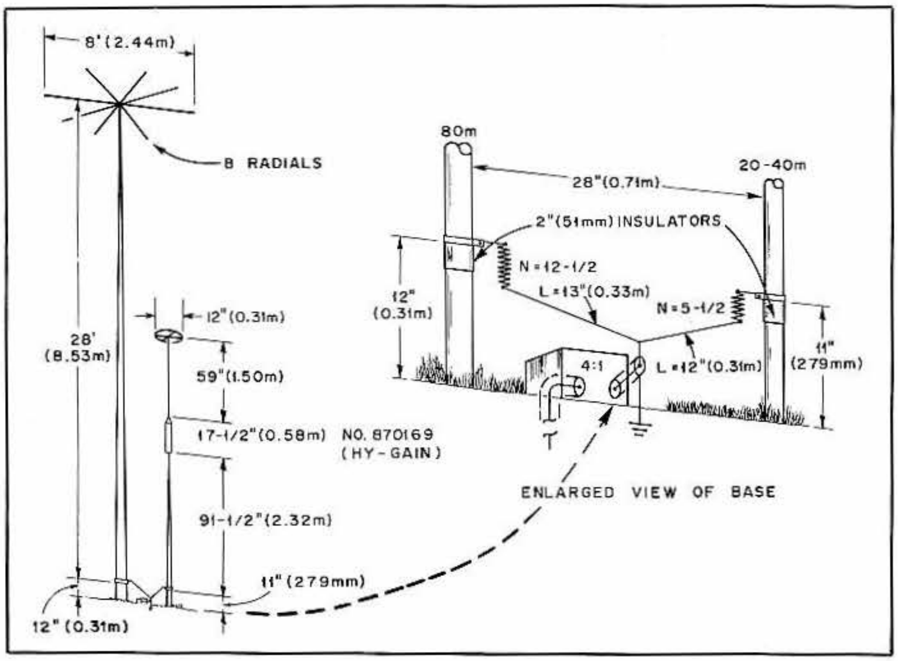

Fig. 9 — The 20-, 40- and 80-meter vertical antenna system. Tuning and construction details are given in the text.

Close-up of the base of the triband vertical antenna.

(229 mm) long and has an ID of 1 inch (25-mm). The bottom aluminum tubing supporting the antenna is 3-1/2 feet (1.06 m) long with 2-1/2 feet (0.76 m) of it placed in cement. The diameter of the hole in the ground is about 1 foot (0.31 m). Even though this construction could probably be self-supporting, three simple polypropylene guys at about the 7-foot (2.13-m) level are used for extra margin of support. The radials at the top use 1/2-inch (13-mm) aluminum tubing.[7] The base loading consists of 12-1/2 turns of a B&W 3029 coil.[8] Actually, 14 turns are on the coil. A shorting stub, as shown in the close-up picture of the base of the two antennas, is used for adjustment. To place the minimum SWR near the low end of the phone band, where much of the DX is worked, 1-1/2 turns were shorted out. The final number of turns employed depends, to some extent, on the number of radials used in the image plane. A simple check is to set the shorting tap at some convenient point, like 12 or 13 turns, and plot the SWR. If the minimum value appears too high in frequency, then add about a half a turn of coil. This half turn chould change the position of the minimum SWR value by about 50 kHz. As can be seen in Fig. 10, the minimum value of SWR is practically 1:1 and occurs at 3.840 MHz. This also verifies the input impedance value of 12-1/2 ohms, as expected from the previous work on short verticals.

20/40-Meter Vertical

In extending the operation of a vertical system over other bands, many alternatives are available. A trap vertical with a 12-1/2-ohm input impedance can be connected in parallel with the 80-meter vertical. Alternatively, the trap vertical can be designed to present an impedance of 50 ohms and thus be connected to the input side of the 4:1 transformer.[9] If broadband operation is desired on only 80 and 40 meters, then one-eighth wavelength verticals on both bands can be used.[10]

For this work, a 20/40-meter trapped vertical having an input impedance of 12-1/2 ohms was used. It offered a rather simple mechanical form of parallel operation as is shown in Fig. 9 and the close-up picture. The bandwidth on 40 meters of 155 kHz appeared acceptable. This bandwidth can be extended by about 50 percent by using the 50-ohm design of a trap vertical.

In the first attempt of parallel operation, the 20/40-meter vertical was placed only 14 inches (355 mm) away from the 80-meter vertical. The coupling appeared excessive. The 80-meter vertical was detuned by approximately 50 kHz. The 20/40-meter vertical also required excessive base loading in order to present an acceptable input impedance. By doubling the spacing between the verticals to 28 inches (711 mm), the interaction between them became negligible. The final values of heights and loadings were practically the same as if the elements were operating alone.

The adjustment of the 20/40-meter vertical is somewhat more complicated than the 80-meter vertical. An impedance bridge, as described in the *ARRL Handbook*, is of considerable help. In this case, there are two degrees of freedom: (1) varying the number of turns at the base, and (2) adjusting the lengths of the vertical sections.

Basically, the tuning is as follows: The tap is set at about 5-1/2 turns and the 20-meter section adjusted to give an acceptable value of SWR, both in position in the band and in magnitude. If the impedance is too high, it can be lowered by increasing the number of turns and lowering the height of the 20-meter section for resonance. After this, the 40-meter portion is then tuned. A plot of the SWR vs. frequency can immediately give an indication of the necessary adjustment of the section above the 20-meter trap. If the minimum value appears too low in frequency, a shortening of the top section is required. In no case should large adjustments be made. A change of a few

Table 1

Some Results With Triband Vertical

Date	*Station Contacted*	*W2FMI Signal Report*	*Freq.(MHz)*	*Input Power SSB (Peak)*	*Comments*
3/13	K5LWL/YV6	59 + 10 dB	7	2 kW	very, very strong signal, only one signal stronger — he used a Yagi at 120 feet (36.58 m)
4/3	WB5HJY	59 + 40 dB	7	2 kW	superior to anything on band
4/4	ZF1SP	59 +	14	2 kW	very, very nice signal
4/4	WA4MUR/4	59 + 40 dB	14	2 kW	best signal on whole band
4/18	KV4HW	59 + 10 dB	7	2 kW	loudest on band
5/9	K6YIY	59	4	2 kW	K1GZL and I are only ones they hear
5/9	W9LZX	59 + 30 dB	4	2 kW	strongest (very consistent signal)
5/12	W2DU	59 +	4	2 kW	tremendous signal
5/17	VK5PB	56-7	4	2 kW	remarkably strong
5/17	W4JNY	59 + 20 dB	4	2 kW	outstanding, certainly one of best
5/22	VK5PB	59 + 10 dB	7	2 kW	really amazing
5/23	ZL3RJ	57-8	4	2 kW	one of the strongest he's heard for some time — pinned the S meter
5/27	WA2BQL	59 + 30 dB	4	200 W	very potent — Stronger than most locals — couldn't believe you were using 200 W

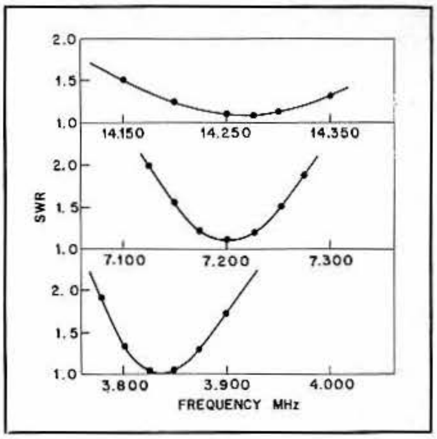

Fig. 10 — Standing wave ratio vs. frequency for the 20-, 40-, and 80-meter vertical antenna system.

inches has considerable effect. If the input impedance on 40 meters appears too low, then the 20-meter section has to be lengthened. This requires that the whole procedure be repeated. In any case, the initial adjustment should be started on the highest band of the trap vertical.

Results

Short verticals have been used by the author during the past year with considerable success. Many DX contacts were made on 40 meters with antennas varying in length from 6 to 33 feet (1.83 to 10.05 m). Since an extensive ground system was used, very little difference in effectiveness was noticed between the antennas.[11] This even includes redesigned trap verticals. As was stated before, the objective of the present investigation was to design a three-band vertical system which not only yielded competitive antennas on the lower bands, but one that was capable of covering a considerable portion of 80 meters. As was seen, a separate one-eighth wavelength antenna connected in parallel with a trap vertical not only gave a bandwidth on 80 meters of 140 kHz where the SWR was less than 2:1, but was short enough to not require considerable help in erection.

During all this time of operation of short verticals on 40 meters, and as of this writing, including several months on 80 meters, very few signal reports were received which did not indicate one of the best signals on the band. Table 1 gives some of the reports and comments received. In only three specific cases on 40 meters have the short verticals been bested by other antenna systems. One is shown in Table 1 where K5LWL/YV6 reported a stronger stateside signal by an amateur using a Yagi at 120 feet (36.58 m). The other cases include a comparison with W2GO of Linden, NJ. On 40 meters, VK5PB reported 6 dB and VK2WC 10 dB in favor of W2GO's signals. He was using a two-element Yagi at 60 feet (18.28 m). The elements were 44-1/2 feet (13.56 m) in length and the boom, 20 feet (6.10 m). On 80 meters, only one other station received a stronger report on direct comparison. This was by W2HCW, when comparing my signals with VK5PB. The difference was 2 to 3 S units. His antenna was an 80-meter Yagi at 120 feet (36.58 m).

Invariably, most amateurs were surprised by the performance of these verticals. In many instances, questions were asked regarding the minimum number and length of radials required for efficient operation of ground-mounted verticals. As was noted in previous articles,[12] the answer depends to some extent on the conductivity of the soil at the respective locations. Poorer soils not only require more radials, but ones that are also considerably longer. Although more experimental work is required in this area, it appears that about 50 radials, 0.2 wavelength long, should generally give good operation. The loss in this case will be approximately 1 to 2 ohms. Doubling the number to 100 radials should reduce the loss to less than 1 ohm. It should be noted that even 1 to 2 ohms of loss are appreciable with these short antennas since their radiation resistances are only 12-1/2 ohms.

Again, the author would like to acknowledge the help, encouragement, and interest shown by the many amateurs during our experimental studies on verticals. Very few antenna laboratories can boast of a greater number of willing and competent field stations. In particular, we would like to thank Al Jones, W2GO, for his considerable help in obtaining comparative reports in Australia and New Zealand. This material was originally presented in *QST* by J. Sevick, W2FMI.

References

[1] Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, March 1973. Also appears elsewhere in this chapter.

[2] Sevick, "The W2FMI 20-Meter Vertical Beam," *QST*, June 1972.

[3] See footnote 1.

[4] Commercial unit.

[5] See footnote 1.

[6] Cadillac Plastic and Chemical Co., P. O. Box 810, Detroit, MI 48232

[7] Construction details on the top hat are also given in the reference in footnote 1.

[8] 2-1/2-inch (64-mm) diameter, 6 tpi, no. 12 wire.

[9] The characterization and design of these trap verticals will be published later.

[10] See footnote 1 for details on the 40-meter vertical.

[11] See footnote 1.

[12] See footnote 1.

A Multiband Vertical Radiator

An efficient multiband radiator with no traps and no loading coils? Impossible; you can't build one without them and have it work! "Oh, yes, sure you can." I doubt it. "You wanna bet? Come on, I'll show you."

What would you say if the author told you this vertical antenna system was 89-percent efficient but had only six radials? (Efficiency is defined as the ratio of radiated power to total power delivered to the antenna.) No, it's no joke. The vertical antenna that is described here is designed to work against ground systems that are less than optimum.

By increasing the radiation resistance by about 200 ohms, we have been able to improve the effectiveness substantially of a vertical radiator with a poor ground system. The value of 200 ohms was selected as the feed-point radiation resistance because a toroid autotransformer with a 1:4 ratio will match 50-ohm coax line very nicely. An added bonus when using such an arrangement is the Faraday screen effect that reduces noise pickup from being transferred to the feed line. The noise reduction over a conventional network was determined to be about 6 dB. See Figs. 11 and 12.

The system is based on a 135-degree antenna height. The radiation resistance at the base of the structure is about 200 ohms,[1] ideal for our matching system. And at the same time the radiation pattern retains a reasonable form. The 10-meter groundplane on top of the structure acts as a top hat. On 7 MHz the phase loops cause the antenna to look like a fat radiator and between the two, top hat and phase loops, the electrical height of the antenna is raised to within the tuning range of the toriodal loading network. On 14 and 21 MHz the phase loops keep the instantaneous current on the structure in phase and eliminate high-angle lobes. When the system is switched to operate on 28 MHz, the groundplane at the top of the structure is fed through the coax line running up the lower portion of the structure.

Construction and Tuning

The photographs and the drawings show how the antenna is constructed. The system was designed so that "storebought" items could be used for the assembly, rather than fabrication of the antenna components. A list of required parts is given in Table 2.

The tuning adjustments of the system are very noncritical. Tuning can be done with an SWR indicator. The best sequence is to start with the highest frequency and work down. The 28-MHz groundplane used as a top hat can be adjusted before it is attached to the rest of the antenna. The Radio Shack 11-meter groundplane that we used comes with a hairpin or "beta" match. With a short length of 50-ohm coax for the feed line, the groundplane should be adjusted for the lowest SWR in the 10-meter band. After the groundplane has been adjusted and attached to the top of the antenna, the entire system can be raised into position for final tune-up on the lower frequencies.

At this point the SWR indicator should be inserted in the 50-ohm feed line from the transmitter, as close to the antenna matching network at the base of the structure as possible. All that remains is the adjustment of the 200-pF capacitor. The

This multiband radiator, just over 36 feet (10.97 m) tall, operates on five amateur bands, 80 through 10 meters.

Table 2

Parts Required for the Multiband Vertical Radiator

3	—10-ft (3.05 m) lengths TV mast; Radio Shack 15-5066.
1	—27-MHz ground plane; Radio Shack 21-901.
3	—100-ft (30.48 m) rolls of 22-gauge hook-up wire; Radio Shack 278-1295.
1	—Amidon Balun kit (see Fig. 12).
1	—Aluminum box for tuner.
1	—28-ft (8.53 m) length of RG-8/U coax line.
4	—8-ft (2.44 m) sections of aluminum tubing (see Fig. 11).
1	—Base insulator; soft drink or (?) bottle.
1	—200-pF capacitor, air variable, surplus.
Misc.	— See text and figures.

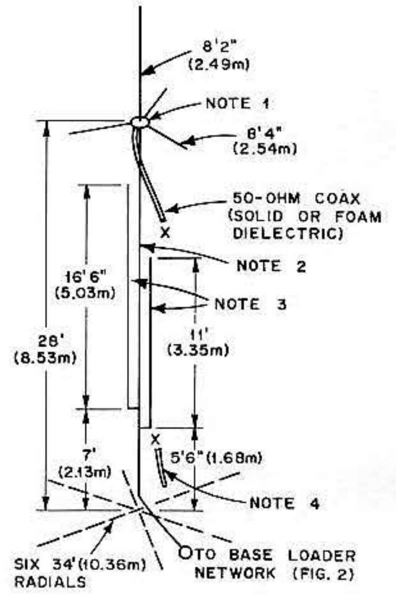

Notes

1) Radio Shack 27-MHz groundplane cut to dimensions shown. Radials are at right angle and connected to mast.
2) Three heavy-duty 10-ft (3.05-m) sections TV mast.
3) Phase loops, sections of telescoping 5/8- (16-mm) and 1/2-inch (13-mm) aluminum tubing spaced 4-1/2 inches (114 mm) from mast. See text.
4) 50-ohm coax such as RG-8/U. At upper end connect center conductor to top element, braid to ground-plane elements. Connect lower end to base loader network (Fig 12).

Fig. 11 — Details of the multiband radiating structure. See Table 2 for parts list. It is very important that a uniform spacing be maintained for the phasing loops, which may be supported at the bottom by shelf brackets. To support the upper portions of these loops, the author used homemade wooden spacers, 1/2 X 1 inch (13 X 25 mm), which were soaked in paraffin and held in place by screws passing through the loop tubing and the main radiator tubing. Homemade Plexiglas spacers may also be used.

---

[1] *The ARRL Antenna Book*, 13th Ed. See Fig. 2-74 and the accompanying text on page 60.

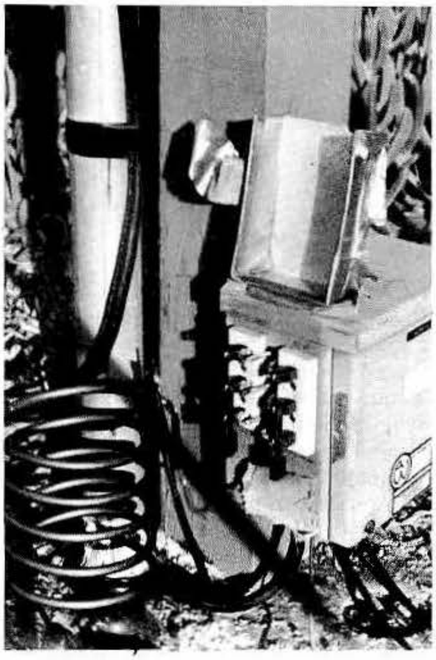

The base of the multiband antenna system. The 200-pF capacitor of the matching network is mounted inside a weatherproofed box, and the 80-meter loading coil is mounted outside. The porcelain knife switches for bandswitching are mounted outside the aluminum box under a homemade hinged cover. The structure itself is supported by a 10-foot (3.05-m) wooden post set 2 feet (0.61 m) in the ground (concrete optional).

Although heavy-duty insulators were used to support the radiating mast from the post as shown here, experience has indicated that insulators are not required. The author suggests that a modified pair of chimney-mount brackets be fastened to the post with lag screws. A short length of PVC pipe or Mylar film may be placed around the mast where it is supported, if you feel it necessary to have the mast insulated from the post.

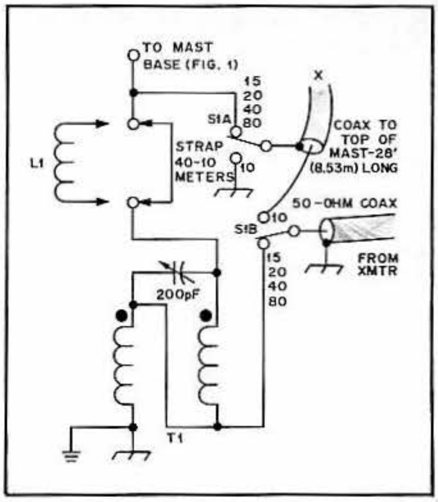

Fig. 12 — Schematic diagram of the base loading network. The connection to the mast base should be made with a short wire, no longer than one foot (0.31-m). The 200-pF capacitor should have a minimum of 1/16-inch (1.6-mm) plate spacing.

L1 — 80-meter loading coil, 0.67 μH; 4 turns of aluminum grounding wire, 3-inch (76-mm) diameter, 1 turn per inch, air core.

S1 — Dpdt, heavy-duty type. Two porcelain knife switches from the local hardware store will suffice, one for each pole.

T1 — Bifilar-wound toroidal autotransformer, 1:4 ratio; 11 turns spaced around 2-inch (51-mm) ID toroidal core: Amidon Balun Kit or equiv. (Amidon Associated, 12033 Otsego St., North Hollywood, CA 91607).

tuning will be different for each band. On 21 MHz the capacitor will be only about five percent meshed, on 14 MHz about half, and on 7 MHz nearly fully meshed. The capacitor on 3.5 MHz, along with the 3.5-MHz coil, can be adjusted to whatever section of the band you wish to work. The bandwidth will be about 400 kHz with less than 2:1 VSWR. To switch bands, all that is required is to set the double-pole, double-throw switch for the proper band and adjust the 200-pF variable capacitor.

All of the antennas constructed so far were cut to the dimensions shown in Fig. 11 and have not required any change. They have been installed in a wide variety of locations and the system design has been broad enough in tolerance to handle all of them. The writer's antenna has been in operation since 1972 and has performed exceptionally well. In the first year on the air, I managed to work 100 countries with less than 250 watts. In several experiments with just the six radials, we made comparative tests between this design and a properly matched quarter-wave vertical over the same radial system. The radiation efficiency was about 60 percent for the quarter wave vs. 89 percent for the "little monster." In the far field the signal improvement was better than 6 dB.

Now there are eight of these antennas on the air, and several more are under construction. All of the hams who have built the antenna have had comparable results to mine.

There are three hams to whom this author would like to give credit. They helped in testing and measuring the system and encouraged me to get it down on paper. They are Bob Crawford, WA6RYZ; Frank Scott, W6WOP and John Campbell, W6NVV. This material was presented originally in *QST* by A. E. Collins, K6VV.

References

The ARRL Antenna Book, ARRL, Newington, CT.

King, *The Theory of Linear Antennas*, Harvard University Press, Cambridge, MA.

Terman, *Radio Engineering*, McGraw-Hill Book Co., New York, NY.

A Triband Vertical Antenna

Keep your vertical antenna simple. Operate three bands without using traps!

The most common vertical is the quarter-wave type, which radiates well and is easily matched to a coax feed line, but has the weakness that it is a single-band antenna. Methods of adapting a vertical for multiband operation include the use of traps, loading coils and elements of different length fed in parallel. The antenna system to be described here was arrived at by treating separately the problems of designing the antenna and of designing the matching system.

The antenna design was easy. A height of 22 feet (6.71 m) is optimum for a vertical antenna for 10, 15 and 20, and several commercial verticals are of this height. It corresponds to about five-eighths wavelength on 10 meters. Greater height results in reduced low-angle radiation on 10, besides being more difficult to erect. A 22-foot (6.71-m) vertical gives better low-angle radiation than a quarter-wave vertical on all three bands and is especially good on 10 meters.

The antenna consists of 22 feet (6.71 m) of 300-ohm twin-lead, with both conductors tied together, held about an inch from a bamboo pole by means of small blocks of insulating material spaced every two feet or so along the pole and held in place with plastic electrical tape. The pole is guyed with nylon fishing line, and the antenna is operated against a groundplane of four quarter-wave radials for each band. The use of twin-lead rather than a single wire simulates a thicker conductor and reduces the impedance, making the antenna easier to match. Even so, the antenna impedances are vastly different on the three bands. The presumed values of impedance are listed in Table 3. These values are taken from published data[1] and assume a cylindrical antenna of specific radius operated against a perfect ground. The antenna is not cylindrical, and it operates against a wire groundplane, so that the impedances listed must be considered approximate only. The results obtained indicate that they are at least fairly close.

Table 3

Approximate impedances at base of 22-foot (6.71-m) ground-plane vertical antenna

Frequency	*Resistance*	*Reactance*
14.250 MHz	100 ohms	+200 ohms
21.375 MHz	1200 ohms	—500 ohms
28.500 MHz	60 ohms	—220 ohms

Table 4

Impedances at input end of 27.8-foot (8.47 -m) matching section of 300-ohm twin-lead terminated by the impedances listed in Table 3.

Frequency	*Resistance*	*Reactance*
14.250 MHz	93 ohms	+177 ohms
21.375 MHz	63 ohms	0 ohms
28.500 MHz	75 ohms	—282 ohms

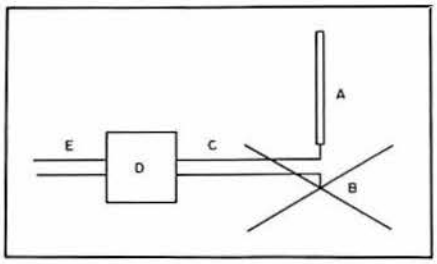

Fig. 13 — The triband vertical antenna system.

A — Vertical radiator, 22 ft (6.71 m) of 300-ohm twin-lead with conductors in parallel.
B — Four radials, each consisting of 1/4λ of wire, for each band (12 wires in all).
C — Matching section, 27.8 ft (8.47 m) of 300-ohm twin-lead.
D — Reactance unit; see Fig. 14.
E — 75-ohm coax to transmitter, any length.

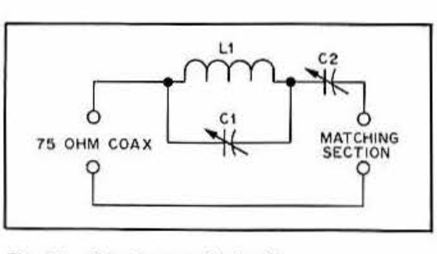

Fig. 14 — Reactance-unit circuit.

L1 — 0.86 µH; 7 turns no. 16 (or heavier). 1-inch (25-mm) dia, 1-inch (25-mm) winding length.
C1 — 30-pF air trimmer set to 23.1 pF.
C2 — 100-pF air trimmer set to 41.7 pF.

If the three antenna matching impedances are normalized to 300 ohms and plotted on a Smith chart, it can be seen that it is possible to obtain a reasonably good match to a coax line on any of the three bands by using an appropriate length of 300-ohm line as a matching section connected between the antenna and the coax line. What is not possible is to find a single length of line that will provide suitable matching on all three bands.

It turns out, however, that a 27.8-foot (8.47 m) section of 300-ohm line having the 0.82 velocity factor which is usual for twin-lead has interesting characteristics. Such a section of line, connected to an antenna having the impedances listed in Table 3, has the input impedances listed in Table 4. This combination of antenna and matching section is shown in Fig. 13.

It will be noted, first of all, that this combination may be used as it stands on 15 meters and fed with any usual coax. What is even more interesting, however, is that the input resistance is near 75 ohms on all three bands, though on 10 and 20 there is also considerable reactance. If a reactance unit could be made which had just the correct value of reactance to cancel out the input reactance of the matching section on all three bands, the unit could be connected in series with the matching-section input and the system fed with 75-ohm coax on all three bands.

Reactance Units

Can such a unit be made? Certainly! It's not hard at all. It consists of a simple seven-turn coil and two trimmer capacitors and is shown in Fig. 14. The trimmers are most easily adjusted with the aid of a grid-dip meter before the reactance unit is installed. C1 should be set so that L1C1 resonates at 35.85 MHz, and C2 can be subsequently adjusted by temporarily connecting it in parallel with L1C1 and varying it until the combination resonates at 21.37 MHz.

With the reactance unit adjusted as described, the antenna when first assembled showed SWRs of 2.1, 1.4, and 1.5 on 10, 15, and 20 meters respectively. Connecting the unit into the feed system apparently increased its shunt capacitance somewhat, for a small reduction in the value of C1 changed the SWR readings to 1.3, 1.5, and 1.4. These values could probably have been improved still more by further adjustment of C1 and C2, but it was not considered worth the trouble. If it is undertaken it should be borne in mind that the adjustments interact, but that C1 has its greatest effect on 10 meters and C2 on 20 meters. In any case it is not possible to obtain a perfect 1:1 match with this antenna system.

An SWR of 2, however, which can easily be bettered on all three bands, is within the loading capability of almost any transmitter and causes an *additional* loss over a perfectly matched line of less than half a dB even when the transmission line is very long. Since such a change in signal strength is undetectable, an SWR of 2 is in practice as good as an SWR of 1, as far as losses are concerned.

---

[1] Jordan, E. C., *Electromagnetic Waves, andRadiating Systems*, Figs. 13-12 and -13, New York: Prentice-Hall, Inc., 1950, pp. 482-3.

A Few Notes

A few observations are in order. The matching section has a fairly high SWR (between 4:1 and 8:1), and it should be spaced well clear of metal objects and should not have sharp bends. TV standoffs are useful here. It is important that the same side of the matching section be connected to the coax shield and to the groundplane. If such a thing were available, 300-ohm coax would probably be preferred for the matching section, but for medium power (180 watts PEP) TV-type twin-lead for the matching section and receiving-type trimmers in the reactance unit have proved satisfactory. The parallel circuit L1C1 does not operate at its resonant frequency, and thus high circulating currents are not encountered. High power might necessitate heavier components. The reactance unit needs protection from the weather and can be built into a small plastic box and sealed with plastic electrical tape.

Although the antenna is matched at three widely separated frequencies, it is not matched at intermediate frequencies and is thus not a broad-band antenna in the usual sense. Nevertheless, it is broad enough for normal ham use. The SWR remains below 2:1 over the entire 15- and 20-meter bands, and also between 28.3 and 29.4 MHz on 10.

All SWRs mentioned in this article have been measured at the transmitter and are thus probably a little optimistic. Transmission-line losses are not known, but the line is not long (about 40 ft. [12.19 m]), and it is doubtful that the SWRs at the reactance unit are much higher than those measured at the transmitter.

This antenna is no match for a quad or a Yagi, but it has provided numerous transatlantic QSOs on all three bands. Because of its effectiveness, convenience, and the cheapness and availability of everything used in its construction, it should appeal to many hams. This material was originally presented in *QST* by Frank A. Regier, OD5CG.

A Multiband Groundplane Vertical Antenna with Tuned Feeders

Who says tuned feeders are used only with horizontal wire antennas? Go multiband with a simple vertical radiator by using a tuned-feeder system — and do it with ease!

For several years, the author has experimented with several versions of this antenna and has found them to be most satisfactory. The antenna shown in Fig. 15 provides not only outstanding performance on 10, 15 and 20 meters, but performance on 40 and 80 meters equivalent to most mobile-antenna installations. Full efficiency on these latter bands can be realized by making the vertical and radial portions proportionately longer.

The antenna system consists of the radiating element, groundplane radials, an open-wire feeder of any convenient length, a Transmatch capable of either series or parallel tuning, and an antenna SWR bridge. This combination is shown in Fig. 16.

Antenna and Radial Lengths

The author preferred to cut the antenna for resonance on one particular ham band (20 meters), but it is not necessary that the antenna be resonant on any band. Efficiency will suffer, of course, on bands where the length is significantly shorter than one-quarter wavelength. Whatever the length chosen for the antenna, the radials should be of about the same length.

Construction

Constructional details are shown in the sketch of Fig. 15. In the event the specified aluminum tubing is not available, thin-walled galvanized electrical conduit, aluminum conduit, or copper pipe may be used.

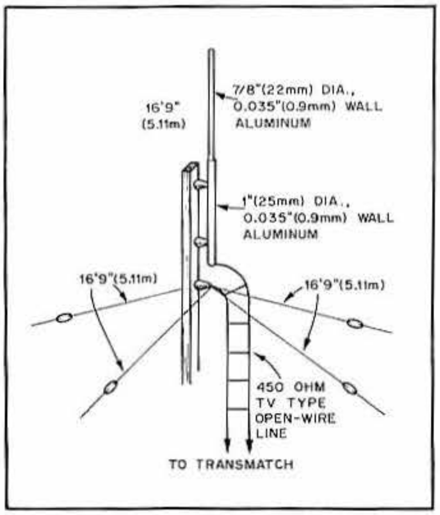

Fig. 15 — Sketch of W4VON's simple multiband antenna.

It is recommended that the vertical element not be supported by drilling holes through the tubing, as this will substantially weaken the structure. An alternative mounting arrangement is shown in Fig. 17. The insulators are porcelain standoff types with a lag-screw insert. They will be found at electrical supply houses (also at Sears), and are sometimes referred to as "saddle" type insulators. The hose clamps are stainless-steel gear type.

Adjustment

Operation of the antenna is simple. The Transmatch (a suitable one, including SWR bridge, is described in the ARRL *Handbooks)* and SWR bridge are connected, and the tuning network adjusted for

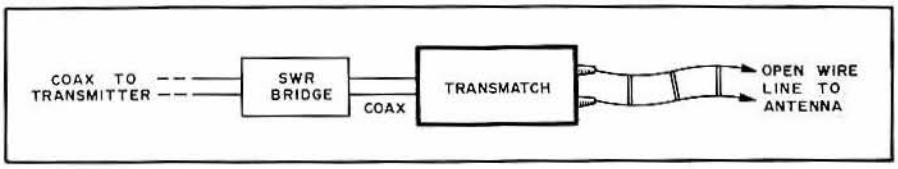

Fig. 16 — Recommended setup for coupling antenna line to low-impedance transmitter output.

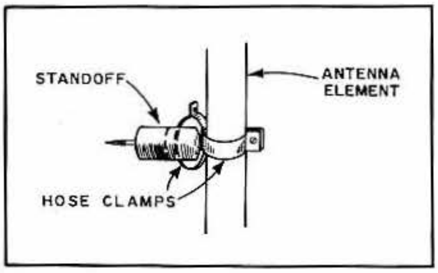

Fig. 17 — Suggested method of mounting radiator element without drilling holes in aluminum tubing.

maximum forward and minimum reflected power. Initially, it will be necessary to determine experimentally whether series or parallel tuning is required for the particular combination of band, feeder length, and antenna length selected.

The tuned-line-fed groundplane vertical antenna gives excellent performance, can be easily constructed in just a few hours even by a beginner, requires a minimum of installation space and costs less than $5, excluding the Transmatch and SWR bridge. The tuned circuit of the Transmatch provides excellent discrimination against harmonic output from the transmitter. Ease of construction and portability make this antenna an ideal one for Field Day use.

In case one wonders about the mismatch between the line and the antenna, the secret is in the use of open-wire line. The loss in such a line with an SWR of 25:1, at 10 meters, is less than the loss in RG-58/U when the latter is matched. This material was originally presented in *QST* by Arthur S. Gillespie, Jr., K4TP.

A Helically Wound Vertical Antenna for the 75-Meter Band

"Around and around she goes, and where she comes out — nobody knows." True, perhaps, with a wheel of fortune, but with a helically wound vertical the rf comes out effectively for DX and local work!

In order to have a full-size vertical antenna for the 75-meter band, the antenna would have to be around 65 feet (19.81 m) tall. That, alone, would require guying, or some other means of support which would result in a structure with no aesthetic appeal to any nonhams (who comprise almost 100 percent of the author's neighborhood). By using a helical configuration, the overall antenna height is reduced, thus eliminating the need for guying. With the helically wound antenna, more equal current and voltage distribution is accomplished (as compared with a lumped impedance from a loading coil). As a result, a better radiation pattern is produced. With this system very little reactance has to be cancelled out, eliminating the need for an overly large "top hat."[1] This antenna will easily take the full legal power limit.

The dark pole to the immediate left of the tower is the WA0WHE helically wound vertical antenna.

Construction details for the antenna are shown in Fig. 19. The following is a list of parts used in constructing the antenna system.

1 — 20-ft (6.10-m) section of 4-inch (102-mm) OD plastic pipe (obtained at a local plumber)

130 ft. (39.62 m) no. 12 plastic-insulated

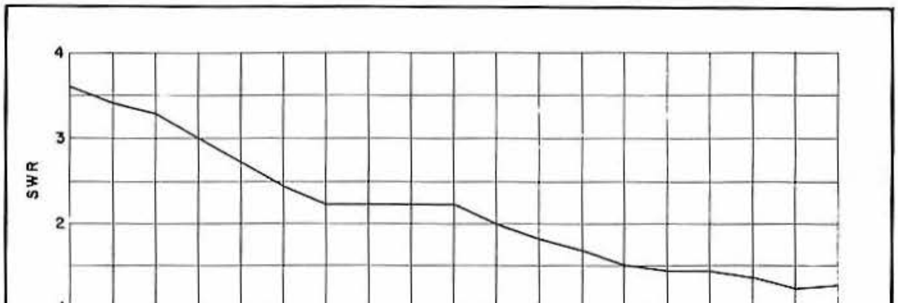

Fig. 18 — Measured SWR of the vertical antenna.

[1] [Editor's Note: Some form of metal top hat should be connected to the last turn of the driven element at the high-impedance end of the radiator. Extremely high levels of rf voltage can develop at the end of the helix, sometimes causing the tip of the antenna to burn. The top hat tends to lower the Q of the antenna, thus reducing the voltage level at the far end. An aluminum pie tin mounted on a ceramic cone insulator works well in this application.]

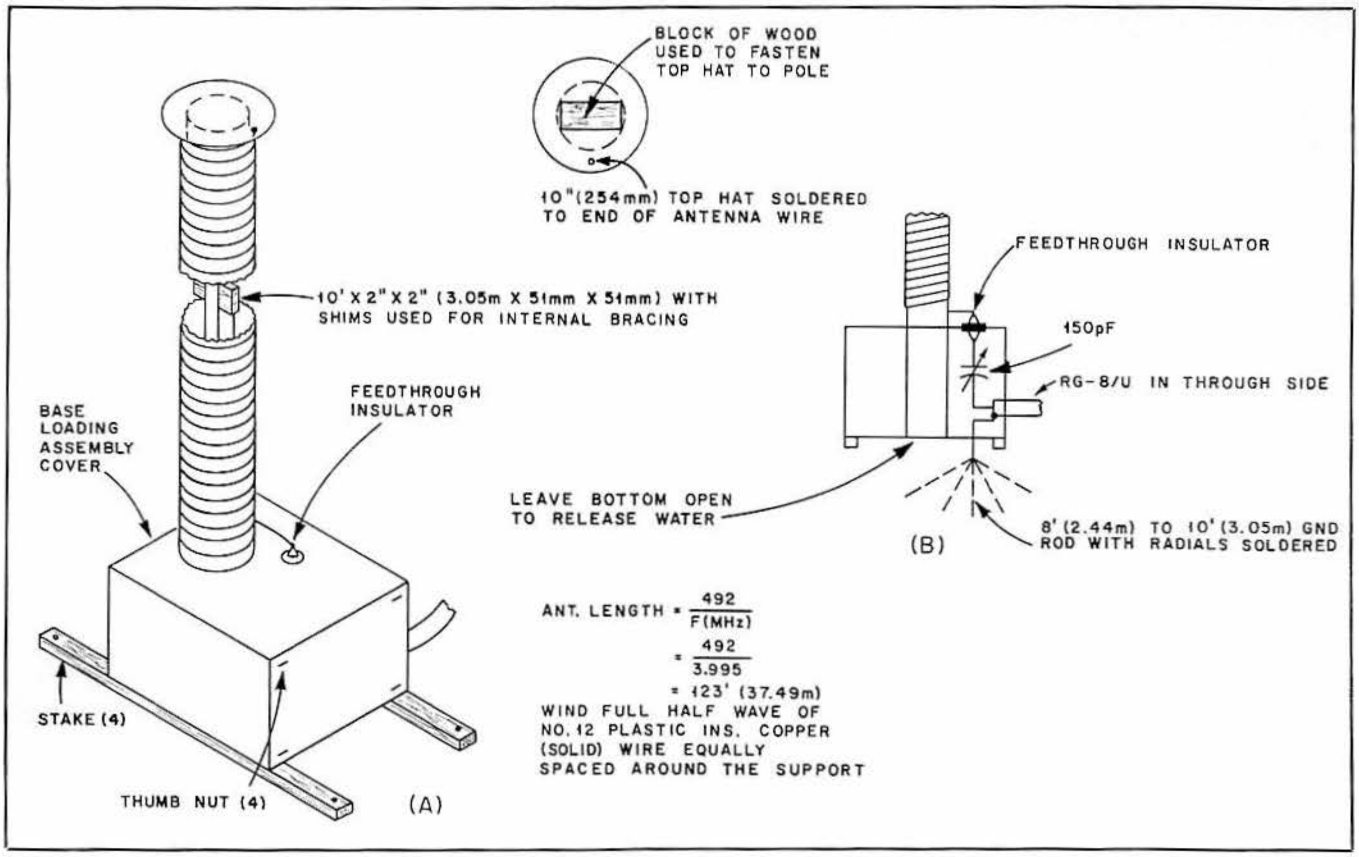
BLOCK OF WOOD USED TO FASTEN TOP HAT TO POLE
10"(254mm) TOP HAT SOLDERED TO END OF ANTENNA WIRE
10' X 2" X 2" (3.05m X 51mm X 51mm) WITH SHIMS USED FOR INTERNAL BRACING
FEEDTHROUGH INSULATOR
BASE LOADING ASSEMBLY COVER
LEAVE BOTTOM OPEN TO RELEASE WATER
ANT. LENGTH = 492 / F(MHz) = 492 / 3.995 = 123' (37.49m)
WIND FULL HALF WAVE OF NO. 12 PLASTIC INS. COPPER (SOLID) WIRE EQUALLY SPACED AROUND THE SUPPORT
STAKE (4)
THUMB NUT (4)
(A)
FEEDTHROUGH INSULATOR
150pF
RG-8/U IN THROUGH SIDE
8' (2.44m) TO 10' (3.05m) GND ROD WITH RADIALS SOLDERED
(B)

Fig. 19 — Construction details of the helically wound antenna.

solid copper wire (allow a little extra, about 2 ft. [0.61 m])

1 — 10-in. (254-mm) disk of sheet metal (top hat; stop at a sheet metal or furnace repairman — I got this free)

260 ft. (79.25 m) no. 14 solid bare copper wire (I used only four radials in this system; use as many as you have room or money for)

20 sq. ft. (1.86 sq. m) of marine plywood for base

2 — 2 × 2 10-ft. (3.05 m) sections, for internal bracing

1 — 8-ft (2.44-m) copper ground rod (for radial junction)

The cost can be reduced considerably if all materials are readily located or already owned.

Fig. 18 shows the results of SWR measurements. The SWR indicator used to make the measurements was a Midland model. The antenna was adjusted for minimum SWR at resonant frequency with the tuning capacitor at the base. The frequency for this antenna is 3995 kHz with 123 feet (37.49 m) of wire wound on the support and one 8-ft. (2.44-m) ground rod and four 65-ft. (19.81-m) radials. There are numerous configurations possible with this system including a broadcasting favorite such as phasing two or more structures. I forgot to mention . . . give the entire structure one or two coats of marine varnish to seal the turns. This material was originally presented in *QST* by Gary L. Ellingson, WA0WHE.

The Ground-Image Vertical Antenna

The name of the game with vertical antennas is the "ground system." Not just any hastily contrived radial ground will do the job right. Read these details for building an effective ground that provides positive results.

Even though a beam antenna supported at the modest height of 40 feet (12.19 m) is a compromise (a 20-meter beam should be higher to be really effective), it still can be an obstacle to good neighborly relations, at least if your neighbors don't appreciate the ecological beauty of such an installation. Moreover, even that modest sort of an installation presents quite a number of engineering problems.

This report presents the results of the first phase of an investigation to find a less conspicuous but equally effective antenna for use at a new location. The author hopes it will provide suggestions for those faced with a similar problem. At the least, it may be of some value to those with a general interest in the subject of antennas.

This first part of this section deals briefly with theoretical considerations, the second with experimental results on quarter-wavelength and five-eighths-wavelength verticals, and the third on the test equipment. It should be pointed out at the outset that the information presented here is the work of a hobby and as such cannot be exhaustive. It is hoped that others will repeat some of the experiments, extend the work, and report the effect in practice of a groundplane system above 3 MHz. To my knowledge such practical data have not been reported.

Theoretical Considerations

A beam antenna possesses the advantage of gain and directivity. Nevertheless, its DX capability is determined primarily by its vertical radiation pattern. A large portion of the radiated energy should be directed between 5 and 25 degrees from the horizon.[1,2] Horizontally polarized antennas yield lower angles of radiation with increase in height above the ground. This is a result of the interference pattern created by reradiation from the earth's surface. Since the earth is a somewhat conducting medium, the electric field tangential to the surface must be approximately zero. This "boundary" condition is brought about by the induced surface currents which create an electric field of opposite phase. This field then combines constructively and destructively with the initial radiation from the antenna. A model for this condition is an image antenna of opposite phase below the earth's surface as a depth equivalent to the height above the surface.[3] In order to get a lobe below 15 degrees, the antenna height must exceed a wavelength. This is greater than 60 feet (18.29 m) on 20 meters.

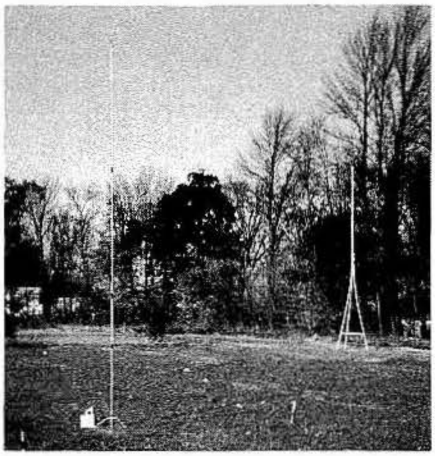

The 20-meter vertical antenna in the foreground, and the 26-foot (7.93-m) test tower.

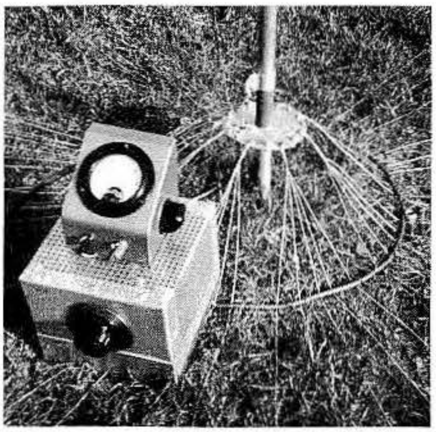

The base of the vertical element of the antenna and the impedance bridge. Forty radials, in bundles of five are fastened to the aluminum base plate and are tied down with a ring made from copper tubing.

On the other hand, a vertical antenna, in the ideal case, possesses an image which is in phase to produce a lobe tangential to the earth's surface. Only when the antenna length is increased to a wavelength, or multiple thereof, does the tangential lobe disappear. This is true whether the antenna is on the ground or suspended in space. Therefore, a vertical beam, i.e., an array of vertical antennas coupled together with an appropriate feed system, on the surface of the earth, seemed to be the logical choice for the new installation since it would appear to meet the following objectives: The system should

1) Exhibit a minimum profile.
2) Be easy to install and tune.
3) Offer a low angle of radiation.
4) Not require a large outlay of money.

The author proceeded, therefore, to construct an array with four vertical elements. When the tests were begun, it immediately became apparent that the simple procedures the writer was using were inadequate to cope with such a complex system. The author had to start anew to develop a test procedure and to build some suitable test equipment. The logical step was to backtrack to a single vertical antenna and use it as a standard on which to develop some basic test standards. As discussed subsequently, the writer found that relatively simple equipment, e.g., a simple impedance bridge, a field-strength meter, and a test oscillator, gave this author all of the data about the system that he needed.

It is important to note that a true ground-image antenna differs substantially from a ground-plane antenna system which relies merely on a few $\lambda/4$ radials above the ground. A true ground-image system results when a sufficient number of radials are used and an image of only the vertical section is sufficient to describe it. Considerable information[4,5] is available on ground-image systems for verticals operating below 3 MHz. The results have shown that some

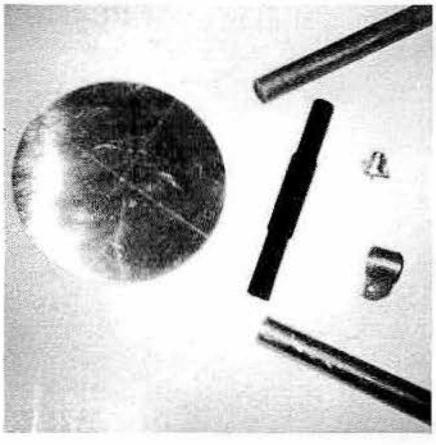

Base hardware for the antennas being tested. The insulator was made from 1-inch (25-mm) maple dowel, turned down to accept the 7/8-inch (22-mm) inside diameter of the ground and antenna tubing.

Table 5

Comparison of responses of 1/4-wavelength and 5/8-wavelength vertical antennas at low radiation angles. Data were taken by field-strength meter mounted on a wooden tower at a distance of 7.5 wavelengths at 14.25 MHz. Field strength (E) is normalized to maximum value obtained with 5/8-wavelength case.

θ	$E(\lambda/4)$	$E(5/8\lambda)$	GAIN OF 5/8-λ ANTENNA
0.1°	0	0.58	∞
0.4°	0	0.62	∞
0.75°	0.62	0.80	3.4 dB
1.1°	0.69	1.0	3.2 dB
1.5°	0.62	0.92	3.4 dB
2.25°	0.48	0.80	4.3 dB
3°	0.41	0.69	4.5 dB

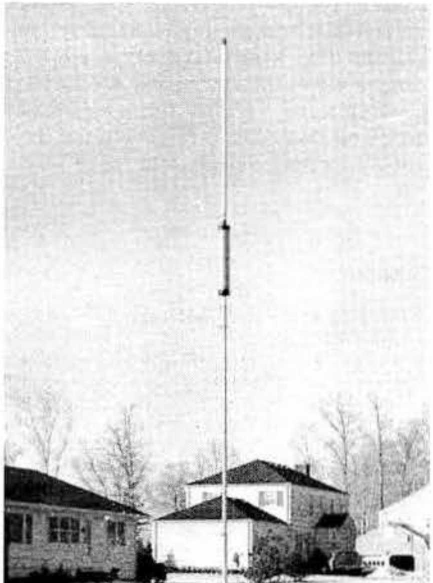

The triband trap vertical antenna, which has an overall height of only 12.5 feet (3.81 m). With the system of 40 radials, performance of this antenna on 20 meters was about the same as that of the 1/4 and 5/8-wavelength vertical elements.

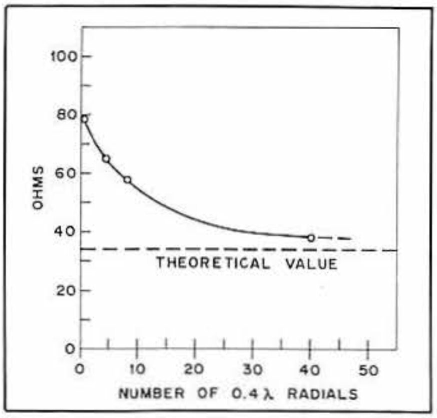

Fig. 20 — The input impedance of a 20-meter quarter-wave vertical antenna as a function of the number of radials 0.4-wavelength long. The four and eight radials consisted of bundles of five wires of no. 18 guage. The 40-radial point was obtained by fanning out the eight bundles.

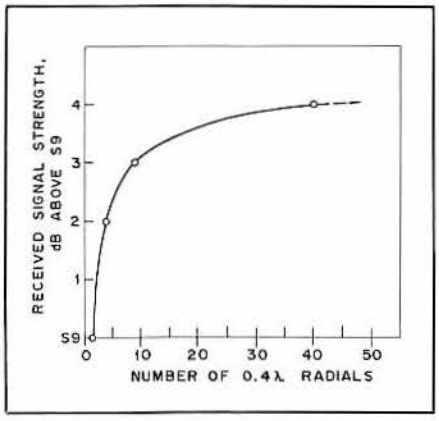

Fig. 21 — The improvement of low-angle radiation of a quarter-wavelength vertical antenna on 20 meters as a function of the number of added radials. A test oscillator was mounted on a wooden tower four wavelengths away at an elevation angle of six degrees from the base of the vertical.

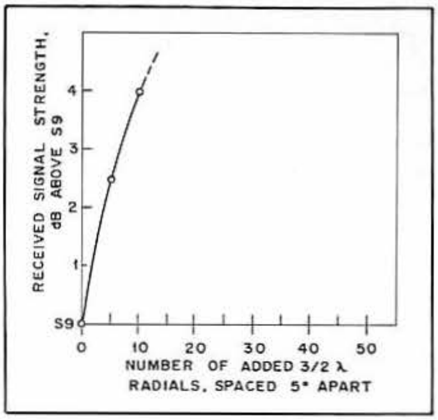

Fig. 22 — The result of interlacing 3/2-wavelength radials in a particular direction. Data were obtained by using a test oscillator on a wooden tower eight wavelengths away, at an elevation angle of 3 degrees from the base of the vertical. The results indicate the advantage of longer wires and the possibility of directional properties of a nonsymmetrical radial system.

100 radials of $\lambda/2$ in length, buried just below the surface, provide an adequate ground system. At higher frequencies, the dielectric effect of the earth becomes important, resulting in severe discrimination of radiation or reception at very low angles.[6,7] At low angles, the waves not only suffer by absorption, but also by a change in phase which results in destructive interference. Since little specific information was available at higher frequencies, the objective of the present work was to measure the input impedance and vertical radiation pattern at low angles of a vertical antenna as a function of the number of radials, in order to determine the feasibility of a vertical-array system as a competitor to beams at high elevation. The results of the tests were most gratifying. The author found that many of the "rules of thumb" which have developed and have been perpetuated to the point where they are practically taken for granted were more myth than truth.

Experimental Results

The classical paper on radial systems, which reports experimental results at 1 and 3 MHz, indicates that a large number of radials 0.4λ long should be used.[8] This appeared to be a good starting point for a 20-meter vertical. Accordingly, to check the number of radials needed, this author used eight bundles of wires, each 25 feet (7.62 m) long, and each made up of five no. 18 copper wires. Each bundle was bolted to a 5 × 1/4-inch (127 × 6-mm) aluminum plate as shown in the photograph. The writer was then able to use each bundle as a radial, measure the input impedance of the system, and then separate the bundles, wire by wire, to increase the number of radials in the system.

Fig. 20 shows the input impedance as a function of the number of radials used. Measurements were made with a simple impedance bridge. (Its construction is discussed later.) The antenna was resonated before each measurement was made, and the difference between the 40-radial system and the 8-radial system resulted only by fanning out the bundles of wires. This technique points out another important feature of ground systems and refutes one of the old myths: since the current carried by the radial system is equally divided among n radial elements, each radial is required to carry only 1/n of the total current. This means that relatively small diameter wire is perfectly adequate.

Moreover, it was found that at the higher frequencies it is best to keep the radials near the surface of the ground. Radials buried more than a few feet become less effective! Thus, you should not rely on the old admonition that radials should be 6 feet (1.83 m) down to be effective. This writer found that burying the wires slightly below the surface is the best way of installing the system. Mechanically the radials can be nailed down, electrically they are most effective, and esthetically they provide little interference to a healthy stand of grass.

It will be seen from Fig. 20 that the input impedance for a vertical antenna is drastically affected by the number of radials. Many radial wires are necessary to prevent an excessive loss of power and to provide a convenient input impedance. Fig. 20 also shows the theoretical impedance limit for an antenna having an effective height-to-radius ratio of 300:1.[9]

Fig. 21 shows the effect on low-angle radiation as a function of the number of radials. These data were obtained by placing a test oscillator on a wooden tower, four wavelengths away. A photograph shows the 20-meter vertical and the 26-foot wooden test tower.

The effect of using longer radials in a particular direction is shown in Fig. 22. Radials of no. 18 wire, 3/2 wavelengths long, were put down between the existing 40 radials. The spacing between the longer wires was five degrees. The considerable improvement indicates the need for longer wires and the directional properties a groundplane could give.

A 5/8-wavelength vertical was also constructed and tested. It consisted of a 40-ft. (12.19-m) telescoping aluminum pole and a loading coil of eight turns of no. 12 wire with a diameter of 2.5 inches (64 mm). The comparison in low-angle radiation at a distance of 7.5 wavelengths is presented in Table 5. A field-strength meter was mounted at different heights on the 26-foot (7.93-m) tower. The groundplane for these data consisted of the 40 0.4-λ radials plus the 11 3/2-λ ones. Measurements were taken in the direction of the added longer radials. The results show the improvement in low angle radiation offered by the 5/8-λ vertical. The input impedance of this longer antenna was found to be 76 ohms.

On-the-Air Checks

From these experiments and measurements the author decided to settle on a 1/4-λ antenna with approximately 40 radials. The installation is shown in the photograph. This writer then proceeded to make on-the-air tests to compare its effectiveness with an inverted-V antenna having its apex at 0.4 wavelength, and with a 5/8-λ vertical using the same ground system. Surprisingly, the 1/4-λ vertical seemed to perform just as well as the much taller 5/8-λ vertical. This could result from the fact that most signals arrive after several hops and the optimum lobe angle is probably as high as 15 to 20 degrees.[10] At that angle, the λ/4 antenna actually enjoys an advantage. With practically all DX contacts, the verticals had a 6- to 8-dB improvement over the inverted V! The only exceptions were at intermediate distances and for local contacts. At about 500 or 600 miles, the inverted V with its higher angle of radiation gave better results. Locally, the verticals gave far superior performance. Improvements of 10 to 15 dB were recorded. A triband trap vertical antenna was also tested on 20 meters and found to be practically the same in impedance and performance. This antenna had an overall height of only 12.5 feet (3.81 m)!

Test Equipment

A most interesting aspect of antenna measurements is that the equipment can be rather simple and in many cases constructed from items most amateurs have in their junk boxes. All that is really needed is some standard to compare with, such as a wattmeter or an S meter that is known to have reasonable accuracy. This writer has used both — the wattmeter in a Drake L-4B linear amplifier and the S meter in a R-4B receiver.

The impedance bridge shown in the photograph was patterned after the one shown in the *ARRL Handbook*. Care was taken in shielding the input and output circuits. The meter was mounted externally in order to minimize stray pickup. A calibration curve was obtained at 14.25 MHz by using many carbon resistors of known values as the load.

The field-strength meter is a simple diode detector and dc amplifier. The instrument was constructed to cover the 10-, 15- and 20-meter bands. Its meter was also mounted externally, The antenna length for the field-strength meter was determined by the strength of the available field. In addition to these pieces of test equipment, a 20-meter transistorized crystal oscillator was used for many of the tests.

Conclusion

The performance of a vertical antenna on the ground is highly dependent upon a good ground system, and, properly installed, the antenna can be a very good performer indeed for DX and local contacts. This is particularly true at the higher frequencies where the dielectric property of the earth plays a major role. Forty radials of no. 18 wire, 0.4λ long, will increase the total radiated power by about 3 dB. Radials can be thin if a sufficient number are used. The thought of using thick wires buried deeply, probably a carryover from lightning grounds, is not valid at higher frequencies. The idea of using only four buried radials, as commonly recommended, is a serious error, and if this article does nothing more than eliminate that misconception, this author will be satisfied. Since the electric field only penetrates the ground for a foot or two at the higher frequencies, the radial wires need be buried only as deep as necessary to escape children's feet and the lawn mower. Finally, these results from the single vertical antenna indicate that a vertical array could be a competitor to the more elaborate horizontal beams and warrants further investigation. This material was originally presented in *QST* by Jerry Sevick, W2FMI.

References

[1] Terman, *Radio Engineering*, McGraw-Hill, New York, 1947, p. 651.

[2] Friis, Feldman, and Sharpless, "The Determination of the Direction of Arrival of Short Radio Waves," *Proceedings of the IRE*, Vol. 22, No. 1, January 1934.

[3] Williams, "Antenna Theory and Design," *The Electrical Design of Antennae*, Vol. 2, Second Edition, Sir Isaac Pitman and Sons Ltd., London, p. 126.

[4] Brown, Lewis, and Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proceedings of the IRE*, Vol. 25, No. 6, June 1937.

[5] Wait, "Input Resistance of L. F. Unipole Aerials," *Wireless Engineer*, May 1955.

[6] Feldman, "The Optical Behavior of the Ground for Short Radio Waves," *Proceedings of the IRE*, Vol. 21, No. 6, June 1933.

[7] Jager, "Effect of the Earth's Surface on Antenna Patterns in the Short Wave Range," *Internat Elekitron Rundschau 1970*, Nr. 4.

[8] See footnote 4.

[9] King and Harrison, "The Impedance of Short, Long, and Capacitively Loaded Antennas with a Critical Discussion of the Antenna Problem," *Journal of Applied Physics*, Vol. 15, February 1944.

[10] See footnote 2.

The W2FMI Ground-Mounted Short Vertical

Just because the vertical is physically short doesn't mean it won't provide excellent results in DX work. Theory says it can be done. Here's how!

A short vertical antenna, properly designed and installed, approaches the efficiency of a full-size resonant quarter-wave antenna. Even a six-foot (1.83 m) vertical on 40 meters can produce an exceptional signal. Theory tells us that this should be possible, but the practical achievement of such a result requires an understanding of the problems of ground losses, loading, and impedance matching. This paper covers design principles which are applicable to all of our amateur bands.

Background

Earlier work by the author[1] on the radiation efficiency of a ground-mounted vertical resulted from the desire to design a simple, inexpensive, low-profile DX beam for 20 meters. This first work reviewed the theoretical considerations involved and pointed out that a good ground system is essential to achieve efficient operation of a resonant quarter-wave antenna. The second work applied these results to a 20-meter beam.[2] In the process of trying to extend the results of these investigations to the 40-, 80- and 160-meter bands, and to possible multiband operation, the need arose to understand the operation of a shortened vertical and the effects of different loading schemes on the input impedance.

The first part of this paper deals with the theoretical considerations of a short vertical antenna and the experimental procedures involved in measuring the various parameters. This is followed by experimental results which show the trade-offs involved in shortening antennas by various loading schemes. Finally, specific designs are given for the 40- and 80-meter bands.

Theoretical and Experimental Considerations

There are several old familiar axioms in Amateur Radio that are not completely understood by the amateur. The author refers specifically to the two sayings: *Make the antenna as long as possible and erect it as high as possible,* and, *A full-size beam is better than a smaller beam, and a tall vertical is always better than a short one.* The first axiom is particularly relevant to horizontal antennas, where height is most important for reliable DX operation.[3] Increasing the length tends to increase the gain in certain directions. But with the vertical antenna, taking it off the ground and feeding it at a voltage point (like the J antenna) in order to eliminate radials can result in poorer performance because of increased earth loss.[4,5] More experimental work remains to be done on this phase of vertical antennas. The second axiom noted above is generally true, but misinterpreted by many. The author refers to those important properties: power gain — important in transmitting and capture cross-section — important in receiving. (Power gain is the gain, in magnitude, compared to an isotropic radiator; cross-section is a fictitious area related to the ability of the antenna to intercept the radiated power and make it available to the receiver for detection.) As will be seen below, very little is compromised in these properties by shortening the antenna.

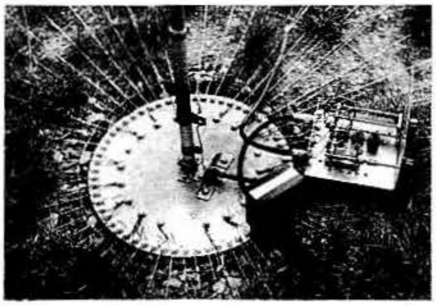

The matching network and base hardware showing 115 radials with the 6-foot (1.83-m) vertical. The pi network is covered with a plastic bag.

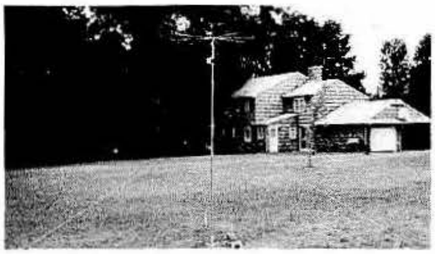

The 10-foot (3.05-m), 40-meter vertical designed for an input impedance of 12.5 ohms in order to use a 4:1 step-up bifilar transformer. The top hat has a diameter of 4 feet (1.22 m) and the coil, placed one foot (0.31 m) below, has 14 turns. Shown at the base is the transformer and impedance bridge.

The author with the 6-foot (1.83-m), 40-meter vertical.

The short antenna has been defined as one that is small compared to a wavelength. In a more exact form, it is defined in such a manner as to simplify the mathematics in the theoretical calculations. King[6] has used the following inequality as the definition where

$$\beta_0 h = \leq 0.5$$

$$\beta_0 = \frac{2\pi}{\lambda}$$

h = half-length of a center-fed antenna or the height of a ground-mounted vertical.

λ = the wavelength

$\beta_0 h$ is actually a quantity which is used to express the height of an antenna in terms of an angle in radians. Thus, this quantity is independent of the frequency. Since 40

Table 6

Inductive Loading

	No. of Turns*	h	$R_{rad.}$
1)	*Base Loading*		
a)	7	24 ft 3 in.	14.5 ohms
b)	10	18 ft 9 in.	7.5 ohms
c)	12	15 ft 7 in.	5.5 ohms
d)	14	13 ft	4 ohms
e)	18	8 ft 10 in.	2 ohms
2)	*Midpoint Loading*		
	*No. of Turns**	*h*	$R_{rad.}$
a)	6	28 ft 5 in.	28.5 ohms
b)	11	24 ft 3 in.	25.2 ohms
c)	15	19 ft 8 in.	16.5 ohms
d)	18	15 ft 8 in.	12.3 ohms
e)	24	14 ft 6 in.	10.5 ohms
3)	*Three-Quarter Point Loading*		
	*No. of Turns**	*h*	$R_{rad.}$
a)	10	29 ft 4 in.	32.5 ohms
b)	18	23 ft	26 ohms
c)	23	21 ft 2 in.	23.5 ohms
d)	24	19 ft 2 in.	22 ohms

*B & W 3029, 2-1/2 in. dia, 6 TPI, no. 12 wire.

1 ft = 0.3048 m

Table 7

Top Hat and Distributed Loading

1) *Top Hat Loading* (4-spoked wheel with 1/8-inch Al wire rim)

	Diameter	*h*	*$R_{rad.}$*
a)	1 ft	30 ft 10 in.	34 ohms
b)	2 ft	28 ft 7 in.	32.5 ohms
c)	4 ft	24 ft	30 ohms
d)	7 ft	19 ft 2 in.	23.5 ohms
e)	4 ft*	23 ft 4 in.	29.4 ohms

2) *Distributed Loading* (Helical Antenna)

	No. of Turns	*h*	*$R_{rad.}$*	*Top Hat at 1.5 ft Above Coil*
a)	111	12 ft	8 ohms	1 ft dia.
b)	105	12 ft	10 ohms	2 ft dia.
c)	113	7 ft	6 ohms	2 ft dia.
d)	75	7 ft	7.5 ohms	4 ft dia.

*8-spoked wheel
1 ft = 0.3048 m

Table 8

Parameters of the 40-Meter Short Vertical Designs

Total Height	*$R_{rad.}$* (ohms)	*No. Turns**	*Dia. Top* (ft)	*Bandwidth*
6 ft	3.5	14	7	100 kHz
8-1/2 ft	7.5	at 6 in. below top hat		
8-1/2 ft	7.5	75 on 7 ft dowel 1-5/8 in. in dia.	4	100 kHz
10 ft	12.5	14 1 ft below top hat	4	125 kHz
15 ft 10 in.	12.5	7 at base	4	540 kHz

*Except for helical antenna, coil wire is same as shown in Table 6.
1 ft = 0.3048 m

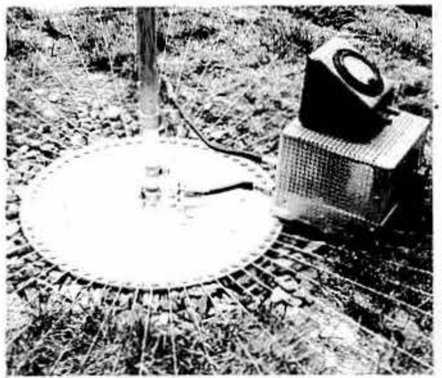

Base of the vertical antenna with 60 radials. The aluminum disk is 15 inches (0.38 m) in diameter and 1/4 inch (6 mm) thick. Sixty tapped holes for 1/4-20 aluminum hex-head bolts form the outer ring and 20 form the inner ring. The insulator is polystyrene material with a 1-inch (25-mm) diameter. Also shown is the impedance bridge for measuring input resistance.
Originally, circular wires were used to connect all radials together. These were positioned every 2 feet (0.61 m), starting from the antenna base. Then, one at a time, they were removed, meanwhile keeping a check on the antenna radiation resistance. There was no apparent change in the radiation resistance, so it was concluded that the interconnecting rings of wire were not needed in the ground system.

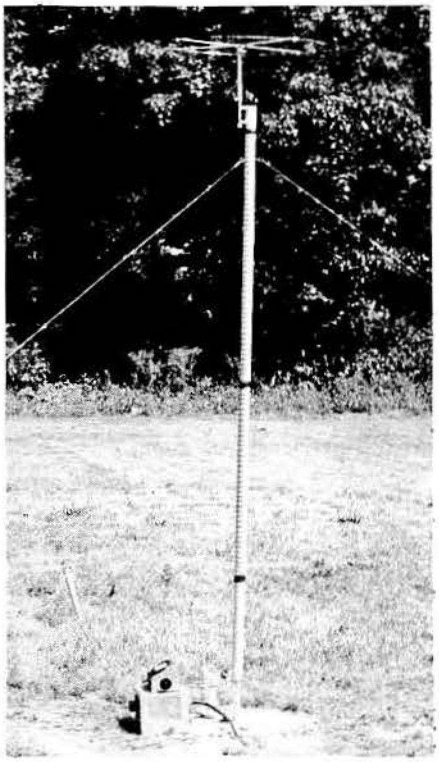

Eight-and-one-half-foot (2.59-m) helical antenna using a 2-foot (0.61-m) top hat 1-1/2 feet (0.46 m) above the 7-foot (2.13-m) helix. The 75 turns have approximately a 3-inch (76-mm) pitch below the midpoint and 1.5-inch (38-mm) pitch above. Very little difference was noted in reversing the pitches. The input impedance was 7.5 ohms.

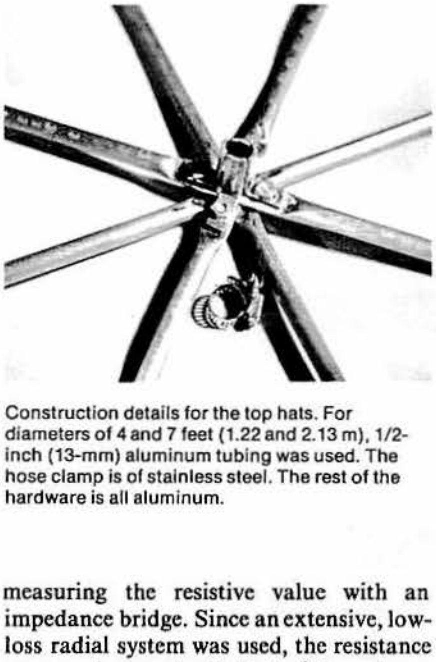

Construction details for the top hats. For diameters of 4 and 7 feet (1.22 and 2.13 m), 1/2-inch (13-mm) aluminum tubing was used. The hose clamp is of stainless steel. The rest of the hardware is all aluminum.

meters was used mostly for experimental results presented in this article, the inequality above assures accurate theoretical calculations for verticals of 11 feet (3.35 m) or less.

The theoretical results show that the power gain for a *very* short antenna, even less than 1-foot (0.31-m) high on 40 meters, is 1.5.[7] This increases slowly to 1.513 for an 11-foot (3.35-m) antenna. These gains are to be compared to about 1.62 for a resonant quarter-wave vertical. It can be seen that this difference amounts to less than 0.4 dB or 0.07 S unit, based on 6 dB per S unit.

In a similar fashion, the capture cross-sections differ by relatively small amounts. For the *very* short antenna, a cross-section value is 0.119 λ^2, while for the quarter-wave vertical it is 0.13 λ^2. This is surprising to many since it is difficult to visualize a vertical of a foot or two (0.31 or 0.61 m) in height on 40 meters, having practically the same receiving ability as a 33-foot (10.06-m) quarter-wave antenna!

But the important property of a short vertical that makes its capture cross-section nearly the equivalent of a full 1/4-λ antenna is its *very small value of input resistance.* Fig 23 shows the theoretical curve and the 40-meter experimental results for the input resistance of a ground-mounted vertical as a function of height.[8] The experimental data were obtained by essentially canceling out the capacitive reactance of the vertical by an inductance at the base of the antenna and measuring the resistive value with an impedance bridge. Since an extensive, low-loss radial system was used, the resistance measured was actually that of the antenna itself, which is called the *radiation resistance.*

Fig. 24 shows the input resistance of a quarter-wave resonant vertical as a function of the number of radials. At the 115-radial point, the input resistance approaches the theoretical value of 35 ohms, which strongly indicates low earth loss and, hence, reliable data in short antenna measurements. The picture also shows the base hardware for 60 radials. The wire is 15-gauge aluminum. There is nothing sacred about 15 gauge. No. 22 wire, or even no. 28, would be just as good.

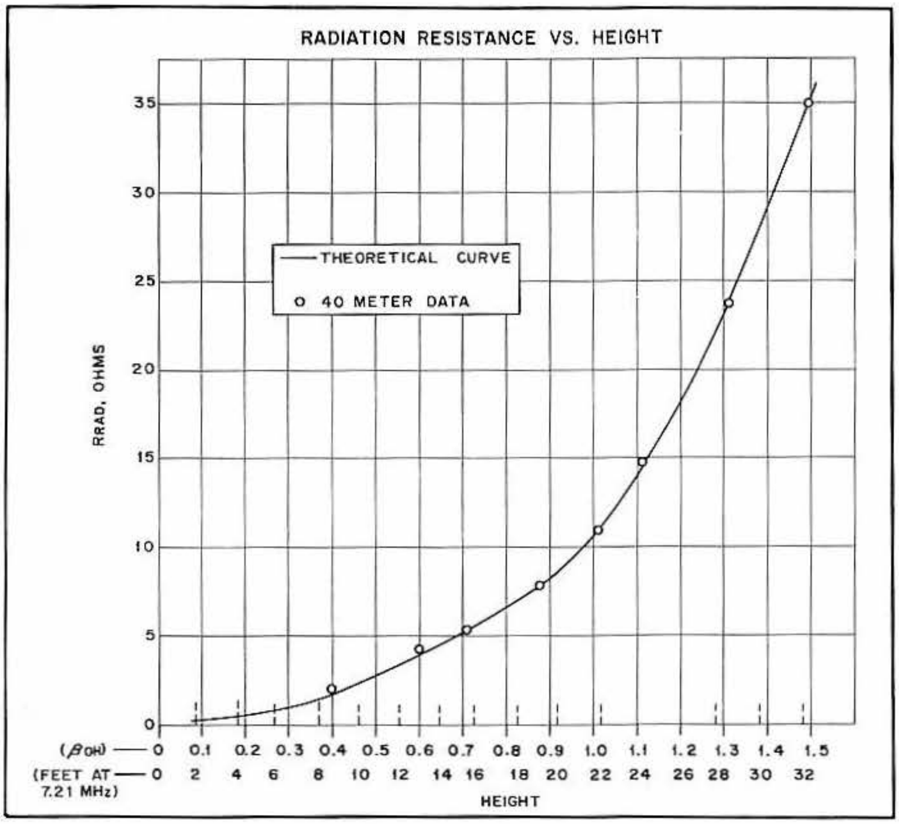

Fig. 23 — Theoretical curve and experimental results for the radiation resistance as function of height.

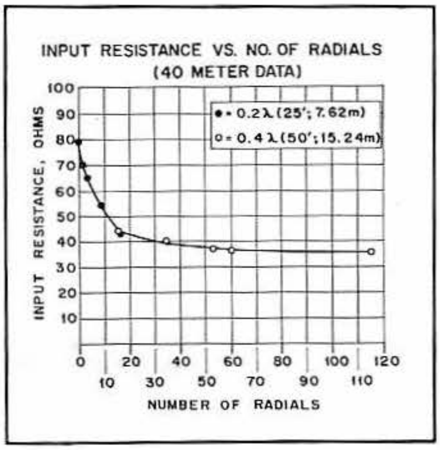

Fig. 24 — Input resistance of a quarter-wave ground-mounted resonant vertical as a function of the number of radials.

It should also be noted that very little difference was noted whether 0.2-λ or 0.4-λ radials were used. At 16 radials, the results are practically identical. In fact, this curve is also similar to the one obtained on 20 meters which was reported earlier.[9] More work remains to be done on the trade-offs in performance versus length of radials for ground-mounted verticals.[10]

Experimental Results

In designing a shortened vertical beam for 40, 80 or 160 meters, a most important consideration is the value of the input impedance of the driven element. As seen in Fig. 23, a shortened vertical using base loading has an extremely low value, particularly at one-eighth wavelength and less. This input impedance is then usually lowered in the presence of other elements in a beam array. The resulting very low value of input impedance makes it difficult to design matching networks. Therefore, an experimental investigation was undertaken to see what increases in radiation resistance could be obtained by other methods of loading, i.e., top hat, three-quarter point, midpoint, and distributed (helical antenna). The results of these experiments are shown in Fig. 25. Table 6 shows the individual points for inductive loading and Table 7 for top hat and distributive loading. Several interesting points were brought out by these experimental results. Fig. 25 shows that top-hat loading yielded the largest value of radiation resistance for a particular height. Surprisingly, the helical antenna[11] yielded a value less than midpoint loading. The three-quarter point and midpoint loading curves were not extended to lower values of height because data were very difficult to obtain below the points shown on the respective curves. The combinations of inductances and lengths below the heights shown on these two curves were probably beyond resonant conditions at the frequency used in the measurements. The other curves were extended by dashed lines indicating no difficulties were encountered in the measurements and other lengths were very possible.

Another interesting aspect of the top-hat loading curve is that a four-spoked wheel approaches to a good degree a solid disk. Doubling to eight spokes only improves the loading by about nine percent as noted in Table 7. Thus, *a few radials on the top of a vertical are very effective.*

Four radials at the base, approximating a ground system, are practically *useless* as noted in Fig. 24. Also, it can be seen from Table 7 that the reduction in length because of top-hat loading is approximately equal to twice the diameter of the disk.

On 160 meters, top hats have been made with sloping wires or struts with success.[12] Because of the sloping nature of this top hat, some cancellation of the antenna current takes place, thus reducing the radiation resistance further.

Although the curves in Fig. 25 were obtained from experiments at 7.21 MHz, these data can be applied to the other bands by proper scaling. For example, by doubling all dimensions, including the number of turns of the loading coils, the radiation resistance values would apply at a frequency of 3.6 MHz. By increasing the dimensions by only 1.85 instead of 2, the results would then apply to a frequency of 3.9 MHz. In like manner, a proper scaling factor could be used to apply these results to a portion of any of the amateur bands.

40- and 80-Meter Short-Vertical Designs

As was stated before, a most important consideration in designing short verticals is a knowledge of the input resistance and how it varies with different types of loading. The objective is to obtain a resistance value large compared to earth losses, so that efficient operation is obtained. In the author's specific case, using 115 radials, practically no earth loss was measured, and, hence, any radiation resistance above a few ohms assured good operation and an opportunity to verify the theoretical predictions for very short vertical antennas.

Since a broad-band, bifilar, 4:1 step-up transformer was available from previous work,[13] the first design was for the shortest vertical having an input resistance of 12.5 ohms. From the curves of Fig. 25, it appeared that a 16- or 17-foot (4.88- or 5.18-m) antenna with a coil of some 13 turns (extrapolated from Table 6) at 8 or 9 feet (2.44 or 2.74 m) from the base would provide the proper impedance at 40 meters. Further, it was decided that some 7 or 8 feet (2.13 or 2.44 m) of length above the coil could be replaced by an eight-spoked top hat having a diameter of 4 feet (1.22 m). The actual design that resulted, after proper tuning, is shown in the picture. The antenna had a total height of only 10 feet (3.05 m) and a 14-turn coil placed 1 foot (0.31 m) below the top hat. This height was about one foot (0.31 m) higher than first expected, but upon careful examination it was noted that the top hat also reduced the radiation resistance, while replacing a section of the vertical portion. Therefore, the height had to be increased somewhat when considering top-loading effects. Also shown in a picture are the construction details for the top hat.

Two other shortened verticals for 40 meters were investigated. One was an 8-1/2-foot (2.59-m) helical using a 4-foot (1.22-m) top hat and 75 turns on a 1-5/8-inch (41-mm), 7-foot (2.13-m) long, wooden dowel. The input impedance was 7.5 ohms and it was matched with a standard pi network. Pictures are shown with the network. Several tests were made by doubling the winding pitch below and above the midpoint, keeping the number of turns constant, with very little difference in results.

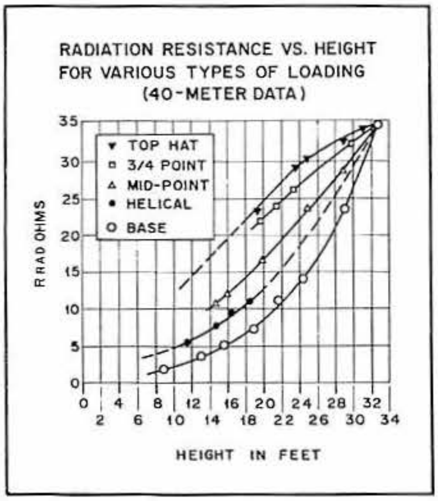

Fig. 25 — Experimental results of radiation resistance as a function of height of antenna for various types of loading (40-meter data).

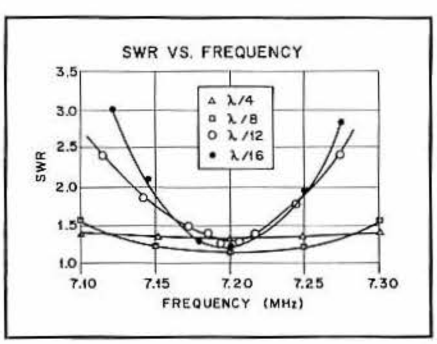

Fig. 26 — Standing wave ratio of various short verticals compared to a resonant quarter-wave antenna . The sixteenth-wavelength antenna is the 8-1/2-foot (2.59-m) helical design, the twelfth-wavelength is the 10-foot (3.05-m) antenna, and the eighth-wavelength is the 15-foot 10-inch (4.83-m) antenna. (All three are described in the text.)

The second antenna, also for 40 meters, using a 7-foot (2.13-m) top-hat, resulted in a matched vertical only 6 feet (1.83 m) high. It had a 14-turn coil, 6 inches (152 mm) below the top hat and an input impedance of only 3.5 ohms. Matching was accomplished with the 4:1 transformer and the pi network. A picture of the antenna and matching network is shown.

A larger 40-meter vertical antenna was also investigated. It had a 4-foot (1.22-m) top hat and 7 turns of base loading which resulted in a height of 15 feet 10 inches (4.83 m) (approximately 1/8 wavelength) and an input impedance of 12.5 ohms. The purpose of this design was to compare its low-angle radiation and bandwidth[14] with the other antennas. Table 8 shows the parameters of these various short vertical designs. Fig. 26 shows the SWR curves of three of the short antennas compared to a quarter-wave resonant vertical and Fig. 27 shows the SWR of the 6-foot (1.83 m) antenna.

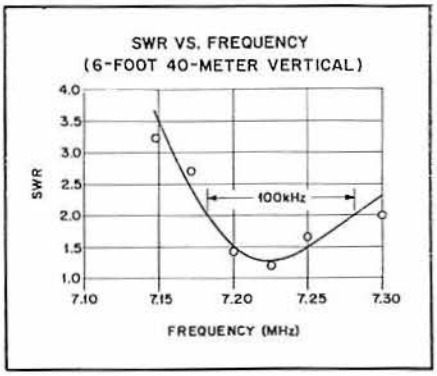

Fig. 27 — Standing wave ratio of the 6-foot (1.83-m) vertical using a 7-foot (2.13-m) top hat and 14 turns of loading 6 inches (152 mm) below the top hat.

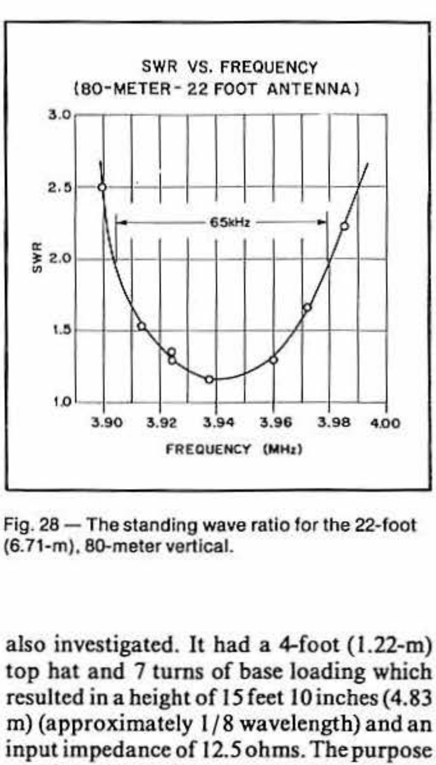

Fig. 28 — The standing wave ratio for the 22-foot (6.71-m), 80-meter vertical.

As can be seen by the SWR curves, shortening an antenna generally decreases its bandwidth. Also by comparing the 8-1/2- and 6-foot (2.59 and 1.83-m) antennas top-hat loading appears to affect the SWR the least. It should also be noted that an eighth-wavelength antenna appears to have a reasonable bandwidth and would probably result in a practical beam design.[15]

And finally, the four short-vertical designs described above were compared with a resonant quarter-wavelength antenna for low-angle radiation and on the air with other amateurs signals. The low-angle radiation measurements were made at three wavelengths in distance and at heights of 1, 3, 6 and 8 feet (0.31, 0.91, 1.83, 2.44 meters). This relates to vertical angles of 0.15, 0.45, 0.90 and 1.2 degrees, respectively. All measurements were made under matched conditions and with a constant 100 watts fed into the antennas under test. In no case were there any appreciable differences noted in the field strength measurements. In fact, the 6-foot (1.83-m) antenna seemed to give slightly higher readings! These measurements certainly tend to verify the theory on the power gain predicted for short verticals. On-the-air checks were again very gratifying and exciting. Over 200 contacts with the 6-foot (1.83-m) antenna strongly indicated the efficiency and capability of a short vertical. Invariably at distances greater than 500 or 600 miles, the short verticals yielded excellent signals.

As was stated previously, all of the results obtained by measurements on 40 meters can be scaled, by the appropriate ratios of frequencies, to other bands. This was tried out on 80 meters. Since a 7-foot (2.13-m) top hat was available (instead of an 8-foot [2.44-m] one), the height turned out to be 22 feet (6.71 m) instead of 20 (6.10 m). The loading coil had 24 turns and was placed 2 feet (0.61 m) below the top hat. On-the-air results duplicated those on 40 meters. The bandwidth was 65 kHz (half of the 40-meter value) as shown in Fig. 28.

Conclusions

Several interesting results came out of this investigation which, even to one schooled in antenna theory, are difficult to believe. The author refers specifically to the ability of very short verticals to radiate and receive as well as a full-size quarter-wave antenna. The differences are practically negligible. But, as was seen, the trade-offs are in lowered input impedances and bandwidths. However, with a good image plane and a proper design, these trade-offs can be entirely acceptable.

As a result of this investigation, other problems were noted which indicate the need for further experimentation. The writer refers specifically to the trade-offs encountered when shortening the radials (to 0.1 wavelength, for example), the efficiency of ground-mounted versus elevated verticals as a function of height above ground, the characteristics of a multiband vertical over a low-loss image plane, and a shortened vertical beam using elements of the order of an eighth wavelength.

Another point should be mentioned in relation to the results reported here. Even though they were obtained on short, ground-mounted verticals, they are valid for center-fed verticals or horizontal antennas as well. The only differences are that the impedance values should be doubled and the effect of the image antenna accounted for.

The author would like to acknowledge the support of many amateurs for their fine words of encouragement and excellent reporting during on-the-air contacts. In this study, some 350 amateurs reported on comparisons between signals from these short verticals and those of other stations. This material was originally presented in *QST* by Jerry Sevick, W2FMI.

References

[1] Sevick "The Ground-Image Vertical Antenna," *QST*, July 1971. Also appears elsewhere in this chapter.
[2] Sevick, "The W2FMI 20-Meter Vertical Beam," *QST*, June 1972.
[3] See footnote 1.
[4] Miller, Poggio, Burke, Seldon, "Analysis of Wire Antennas in the Presence of a Conducting Half-Space. Part II. The Vertical Antenna in Free Space," *Canadian Journal of Physics*, Vol. 50, 1972.
[5] Feldman, "The Optical Behavior of the Ground for Short Radio Waves," *Proc. of the Inst. of Radio Engineers*, Vol. 21, no. 6, June 1933.
[6] King, *The Theory of Linear Antennas*, Harvard Univ. Press, Cambridge, MA, 1956.
[7] See reference in footnote 6, p. 497.
[8] See reference in footnote 6, p. 190.
[9] See footnote 1.
[10] Weeks, *Antenna Engineering*, McGraw-Hill Book Company, 1968, p. 48.
[11] The diameter of the helical antenna was 1-5/8 inches (41 mm). Each helical antenna had a top hat 18 inches (457 mm) above the helix in order to terminate the upper turns with sufficient capacitive reactance.
[12] See reference in footnote 10, p. 44.
[13] See footnote 2.
[14] Bandwidth is defined as the range in frequency where SWR is less than 2:1.
[15] See footnote 2.

Optimizing Vertical Antenna Performance

Do you have tower restrictions, a small yard or patio, almost no room for antennas? If so, here are some suggestions for making your signal bigger in spite of the handicaps.

Antennas never cease to be a hot topic of conversation among amateurs; the interesting articles by John Stanley, K4ERO/HC1[1] and Roger Hostenback, W5EGS,[2] on the optimum grounding of vertical antennas have led the author to communicate these thoughts on the subject. Obstacles at the author's present location (typical of many suburban dwellings) place severe limitations upon where ground wires can be placed. These obstacles are large boulders, areas of concrete, the house, vegetable and flower beds, and the neighbors' property lines. It has not been possible to lay out radials of uniform length or of uniform angular distribution. Nevertheless, the author has had successful results with vertical antennas. Although the writer must make a number of statements which are based upon fragmentary evidence and are lacking in scientific rigor, they should be useful for suggesting how others might proceed to optimize their particular situations.

The writer has three very dissimilar vertical antennas. Much of the discussion will pertain to their individual ground systems, although strictly speaking they cannot be independent; a change of the grounding conductors of one vertical is bound to affect the others. For brevity the writer shall not discuss the methods of impedance matching of the antennas as this is a topic thoroughly covered elsewhere. The author *shall* mention some problems of using dissimilar antennas in a phased array, although the results are preliminary.

The three antennas are (1) a shunt-fed crank-up tower having a three-element triband Yagi on top, the height being variable from 27 to 41 feet (8.23 to 12.50 m); (2) a 40-foot (12.19-m) insulated base-loaded antenna consisting of a 30-foot (9.14-m) pipe mast topped by a whip, located approximately 70 feet (21.34 m) south of the tower and hereafter referred to as the "south antenna"; (3) about 20 feet (6.10 m) east of the tower, a vertical antenna termed the "north antenna," which was described in an earlier *QST* article.[3] Fig. 29 shows the layout. The north antenna has been used with various heights up to 40 feet (12.19 m) but in the present experiments it is used in its "summer" configuration with a height of 25 feet (7.62 m) and with a 7/14-MHz trap located 10 feet (3.05 m) from the base. For 3.5-MHz operation additional loading is supplied at the base.

The Grounding Systems

None of the three verticals have anything approaching an optimum ground system; while the south antenna has the best on paper, it is not necessarily the best in performance. When the south antenna was first placed into operation the grounding system consisted of about 200 feet (60.96 m)

How do you ground a vertical antenna when there is no ground?

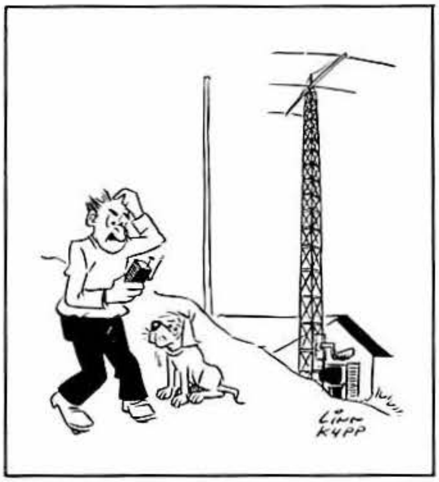

Optimizing ground-wire currents is well worth the time and effort required.

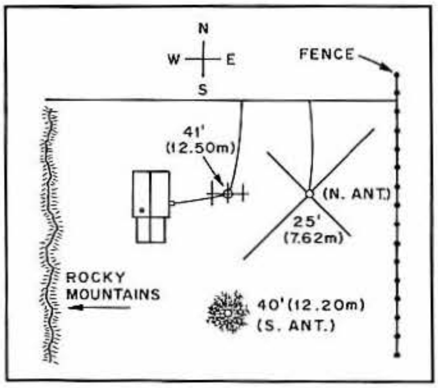

Fig. 29 — The W0JF backyard, with antenna farm.

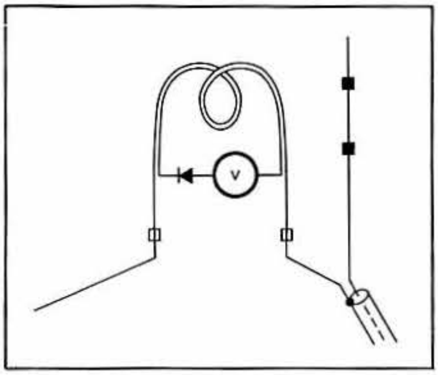

Fig. 30 — Suggested substitute for a thermocouple ammeter.

of wire distributed among a few conductors which went out more or less radially. One was about 60 feet (18.29 m) long, the others 30 feet (9.14 m) or less. On 7 MHz the performance was generally inferior to that of a horizontal half-wave dipole 30 feet (9.14 m) above ground. Later, the ground system was augmented to utilize about 1000 feet (304.80 m) of wire, and the vertical then began outperforming the dipole.

While the author foresaw Stanley's statement that possibly not all of this additional wire would be effective, it was assumed that any wires added at least would do no harm, and the writer tried to put wires in the entire immediate area of the antenna, subject to the constraints mentioned above. Several of the radials have bends to avoid obstacles. In a few cases a single wire goes out from the antenna to a junction some distance away, and then several wires fan outward from the junction. Three wires are connected at the far ends to a few square feet of chicken wire. Where the wires are in shrubbery they are above ground; otherwise they are an inch or two (25 or 51 mm) below. None of the additional wires were as long as 60 feet (18.29 m), and the lengths were generally determined by the obstacles they encountered. Haphazard as this system was, the improvement was spectacular.

While making this change in the south antenna's grounding, the author carried out two simple experiments. First, when operating at 7 MHz, a 33-foot (10.06-m) piece of wire was added, which in free space would be 1/4-λ long, but which when laid on the ground would be longer (in wavelengths). The writer connected it to the rest of the ground system through a thermocouple ammeter, and read the current while applying power. With this wire then cut in half and laid out as separate radials in opposite directions, it was found that the total current increased. Cutting these 1/8-λ pieces in half and distributing the resulting fragments as four radials in directions at 90 degrees to each other caused the total current to decrease. These observations led the author to conclude that for a given total amount of wire the most efficient layout would be radials having lengths of about 1/8 a free-space wavelength. This conclusion is consistent with Stanley's Table 1.[1]

In the other experiment, the author also used a 33-foot (10.06-m) piece of wire, but half of it was no. 34 and half no. 12 wire. With the thick end adjacent to the antenna, the current was higher than with the wire turned around so that the thin end was adjacent. The experiment showed that wire diameter is of some consequence, and it is more important to have heavy conductors close to the antenna.

The Tower and the North Antenna

Grounding of the tower is very heterogeneous; first of all, of course, the bottom of the tower is buried in the ground some 8 feet (2.44 m). There are two radials about 25 feet (7.62 m) long going in a more or less southerly direction, then a short wire going to a water faucet on the wall of the house. Finally, and probably most important, there is a wire that goes northward about 20 feet (6.10 m) and makes a junction with the middle of a 120-foot (36.58-m) east/west wire. The eastern end of this wire is located at the common corner with the property of the neighbor to the northeast and at that point is bonded to his extensive chain link fence. About 25 feet (7.62 m) west of this bond to the fence, there is a junction with a southward wire which is connected to the base of the north antenna. Otherwise, the grounding system of the north antenna, consisting of a few radial wires, is as described in the earlier *QST* article.

Some observations were made by connecting a thermocouple ammeter and a selection of capacitors and inductors in series with various conductors connected to the base of the tower. On 3.5 MHz, with the wire that goes to the fence, the author found a very strong resonance using a 300-pF capacitor. No such resonance was found with this wire of 7 MHz nor on 1.8 MHz, nor did there seem to be any resonance with any of the other wires.

No attempt has been made to bond the upper and lower sections of the tower (where they telescope) together. No such bonding seems necessary, except in the rare case of an extremely strong wind, when there is some receiver noise resulting from poor conductivity between the sections. Cranking the tower up and down seems to have little effect upon the degree of impedance mismatch.

On 3.5 MHz, using a power of about 150 watts, the author had contacts all over the world using this shunt-fed tower, usually not cranked all the way up. The south antenna has also given fairly good account of itself, but with its thousand feet (304 m) of radial wires its performance is still not quite as good as that of the tower with its meager "black magic" grounding.

On 7 MHz, the behavior of the tower is the reverse even though the impedance match is somewhat superior to that on 3.5 MHz. The performance, despite the good match, is far inferior to the south antenna or even the north antenna. Offhand, one would expect the performance, such as it is, to be better with the tower cranked down to minimum height, when it would be close to 1/4λ, but the few reports received on the air indicate better performance is obtained at maximum height. The overall performance of the tower on 7 MHz is a complete mystery.

The author has used the tower on 1.8 MHz only rarely, but has worked as far as 1000 miles, even though there is a very bad impedance mismatch on the feed line. One of the author's future projects is to cure that, since the tower may be magic on the top band, too!

What if You Don't Have a Thermocouple Ammeter?

The preceding discussion has, the author hopes, suggested that individuals may be able to optimize their ground systems by measuring the currents in individual ground wires. In this regard the writer has been fortunate in having available a number of thermocouple ammeters, instruments not commonly available in ham shacks. However, it was found easy to make a substitute which can serve to measure relative current and to detect resonances. Take a short length of twin-lead, speaker cord or similar cable, about a foot (0.305 m) long. Connect one wire of the pair in series with the circuit to be studied (the ground wire). Connect the two ends of the other wire of the pair to a semiconductor diode and a sensitive dc meter in series (Fig. 30). For convenience and for reproducibility in results, coil the cable up and secure it to a terminal board.

Using Dissimilar Antennas in a Phased Array

The author can use any two of the three verticals as a phased array by connecting the feed lines in parallel at the transmitter. In the shack the author has a switchboard that can switch additional lengths of line in series with one antenna or the other. The writer does not attempt to change the impedance match at the bases of the individual antennas, but does generally retune the transmitter as the phasing is changed. Some of the effects observed must be attributed to variations in impedance mismatches rather than to basic changes in directivity. Nevertheless, some gain can usually be obtained by combining the verticals in one way or another.

In the use of two dissimilar antennas, two questions arise which may not apply when identical antennas are used. First, is there really an advantage in using the two of them in a phased array over using the better of the two alone? The relevance of this question may be seen by considering the extreme case where one is very much better than the other. In such a case the field strength of the poorer one is so much less that there is little difference whether it is in or out of phase with the field of the better one, and there can be little or no observable directional effect when the two are used together. Furthermore, the power supplied to the better antenna is cut in half while the other half of the power is largely wasted.

This simple argument suggests that in most cases the two antennas should be nearly equal in the field strengths they produce if they are to be effective in a phased array. However, the actual situation is sufficiently complex that exceptions may be found. For one thing, the better antenna may have "blind" directions in its pattern because of obstructions, and the poorer one may help to fill those in. Also, the impedance effects which have been mentioned before are hard to predict. The author can only suggest that one may find it worthwhile to try using two such antennas together, and may find unexpected advantages in some directions.

On 3.5 MHz the north antenna is generally much poorer than the tower. However, in some directions observers have consistently reported a small gain in using the two together over the tower alone. Incidentally, it should be mentioned that observations made on receiving are not reliable tests, partly because of impedance effects. Also, some of the noise received may be generated locally and its effect may mask the true pattern of the array. The only reliable evaluation is by transmitting to distant observers.

The second question concerning dissimilar antennas involves solving for an unknown quantity: the relative phase of the two antennas. The relative phase of the field of an antenna is *not* determined solely by the length of the transmission line connecting it to a transmitter. The impedance of the antenna usually contains a reactive component, and its presence gives rise to a phase shift. With two identical antennas these phase shifts are the same and do not affect the relative phase between the two antennas, which is determined only by the difference in feed-line lengths. With dissimilar antennas these phase shifts are generally not identical, and they may vary in different ways with shifts in frequency or with variations in the moisture content of the earth. Not unrelated was a discovery made when changing the grounding of the south antenna from 200 feet (60.96 m) of wire to 1000 (304.80 m). There was a substantial change in the phasing when it was used in an array with the tower.

The author has experimentally answered this phasing question by setting up a small probe antenna at a point equally distant from the two antennas and connecting it to a field-strength meter. The line lengths are varied until a minimum (usually quite sharp) is observed. Then it is known that the antennas are 180 degrees out of phase, and the array is a bidirectional end-fire one (W8JK). If the author adds or subtracts 1/2 λ of line from one antenna, the antennas should be in phase and a maximum should be observed. The experimental observations are in agreement with such predictions. If a length of line equal in electrical wavelengths to the spacing between the antennas is then added or subtracted from the feed line of one of them, the combination approximates a unidirectional end-fire array. The favored direction is away from the element which has the longer equivalent feed line (larger phase shift from the transmitter).

To reiterate, the results with phasing dissimilar verticals are very preliminary, and there is little that can be said about them except for the fact that at times very noticeable gains over using the better single antenna are observed.

Vertical Versus Horizontal Antennas

The author used some *horizontal* antennas and observed the differences in performance between them and the vertical antennas at a particular location. Horizontal antennas were not tried on 7 and 3.5 MHz any higher than 30 feet (9.14 m) above ground, however.

The first foothills of the Rocky Mountains rise up immediately to the west of the author's QTH, the nearest ones subtending a vertical angle of about 10 or 15 degrees. As expected, propagation to the west is usually difficult. To the east there are some power wires running north and south. The land slopes off to the north, and has a line-of-sight view for 12 miles from the base of the antenna tower.

The writer finds a tendency for horizontal antennas to be superior to the west, over the mountains, and for verticals to be superior to the east. Standard theory explains these findings as follows: Probably many of the signals coming over the mountains are scattered by the surface of the ground. Those familiar with advanced electromagnetic theory remember that according to the Fresnel equations, the reflection coefficient is greater when the wave is polarized with the electric vector parallel to the surface (i.e., horizontal). Therefore, this scattered radiation is probably horizontally polarized. On the other hand, the electric field from the vertical antenna is perpendicular to the direction of the power wires and is less disturbed by them than the field of a horizontal antenna.

Although very few firm results were presented, the author hopes he has suggested some procedures which will help those facing restrictions on grounding systems they can use. With a little persistence and ingenuity, they should be able to build effective antennas. It *is* possible to evaluate simple arrays yourself and to improve their performance according to your own communication requirements. This material was originally presented in *QST* by Yardley Beers, W0JF.

References

[1] Stanley, "Optimum Ground Systems for Vertical Antennas," *QST*, December 1976.

[2] Hoestenback, "Improving Earth-Ground Characteristics," *QST*, December 1976.

[3] Beers, "Short Antennas for the Lower Frequencies," *QST*, September 1970.

Some Plain Facts about Multiband Vertical Antennas

Does engineering jargon, page-long equations and references to the *IEEE Proceedings* have you gnawing your nails in frustration? Take heart, for these bits of knowledge about multiband vertical antennas are given in plain lingo.

During discussions with newcomers, and old-timers for that matter, it becomes apparent that there is considerable confusion as to what exactly a multiband vertical antenna is. The confusion concerns the method of feed, how much mismatch one can expect, how many radials are required, how the particular antenna is built for multiband use, plus some other points.

This article breaks the subject into simple language and provides the reader with sufficient expertise to assure him that he won't wind up with a system he really doesn't want. Before going into a discussion of the different types of multiband "verticals" we will offer some simple antenna facts.

Some Basic Theory

The term "multiband antenna" has come to mean many things to hams. With trap antennas, tapped coils, random wires, and so forth, there is plenty of reason for the confusion. Simply, a multiband antenna is one that can be used on more than one band. How we make it work on different bands is another story.

Basically, any piece of wire of *any* length can be classed as a multiband antenna. For example, a length of wire 4 feet (1.22 m) long *could* by used on *any* amateur band, from 160 meters on up. But how well the piece of wire would work is a completely different matter.

In the feed point of any antenna there is *radiation resistance*. The energy supplied to an antenna is dissipated in the form of radio waves and in heat losses in the wire and nearby insulating materials. The radiated energy is the useful part, but so far as the transmitter is concerned it represents a power consumption just as much as does the energy lost in heating the wire. In either case the dissipated power is equal to I^2R; in the case of heat losses, R is a real resistance (ohmic losses), but in the case of radiation, R is an *assumed* resistance. This fictitious resistance is the radiation resistance. This brings us to our first important point about multiband antennas.

Whenever one reduces the size (length) of an antenna physically, the radiation resistance is reduced also. As an example, assume we have a 20-meter quarter-wave antenna, which is approximately 16 feet (4.88 m) long. Let's imagine we made it out of no. 40 wire, which has a resistance of about 1 ohm per foot, (0.31 m). The radiation resistance of a resonant quarter-wave vertical operated against a perfect ground is on the order of 35 ohms. In this case, the feed-point impedance of our antenna would be roughly 35 ohms in radiation resistance plus 16 ohms in ohmic resistance. If we were to feed 51 watts into this antenna 16 watts would be dissipated as heat (lost power) and the remainder — 35 watts — would be radiated. Now, suppose we use this same antenna on 80 meters. As mentioned above, when we reduce the size of an antenna physically the radiation resistance is also reduced. On 80 meters our 16-foot (4.88-m) antenna would have a radiation resistance on the order of 1 ohm! However, we would still have the ohmic resistance of 16 ohms. It doesn't take much figuring to realize that just about all of our power would be lost as heat.

Of course we wouldn't use no. 40 wire for such an antenna. More likely the antenna would be made from aluminum tubing and the ohmic losses would be very low, but probably still more than the radiation resistance. There is an old axiom in Amateur Radio that offers some pretty good advice: Always make the antenna as long as possible and erect it as high as possible. Also, there is a joke that goes with that axiom — if such an antenna stays up, it is too small!

At this point we have only mentioned radiation and ohmic resistance in the antenna feed point. These are the two resistances that exist when the antenna is resonant. When the antenna is not resonant, there is reactance present in the feed point. Reactance is also expressed in ohms, but it isn't a real resistance in the sense that power can be dissipated therein. We won't go into a long discussion on reactance because it would take up too much space. An excellent explanation can be found in the League publication, *Understanding Amateur Radio*. Simply, reactance can be likened to a gate or door that stops or hinders the flow of current into a circuit. When an antenna is operated at some frequency other than the resonant frequency there will always be reactance present. Keep in mind that with *any* antenna, multiband or otherwise, we always have a condition on some band or frequency where the antenna is not resonant. Therefore, there will be reactance at the feedpoint.

Types of Vertical Antennas

The basic and most popular type of vertical is one that is a quarter wavelength long and is operated against ground or in a groundplane configuration. The antenna is usually made from tubing and the radials are wire. An ideal groundplane (simulated earth ground) would be a sheet of metal with

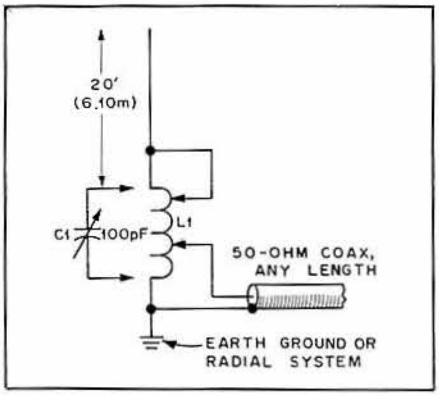

Fig. 31 — A typical multiband vertical antenna. A description of the system is given in the text.

a radius of one-quarter wavelength or more. However, this is only practical at vhf so the customary method is to use wires as the radials. Probably the number one question asked about groundplane antennas is, "how many radials are required?" The answer is simply, the more radials used, the better the antenna will perform, at least up to a certain point. This should not be construed to mean that an antenna with only two or three radials won't work. Such an antenna will work, but for *maximum* performance one should consider 40 or more radials. If the reader is interested in performance data for a few radials versus many, he should read the article in *QST* by Sevick.[1]

The feed-point impedance of quarter-wave groundplane is on the order of 35 ohms. The impedance can be raised by drooping the radials down until a 50-ohm match is obtained. Exactly how much droop is required depends on the number of radials.

The quarter-wave groundplane is essentially a single-band antenna. However, a 40-meter quarter-wave vertical can also be used on 15 meters, a happy circumstance for the Novice. In this case, a 40-meter quarter wave works out to be three quarter waves on 15 and any *odd* multiple of quarter waves will provide a relatively low-impedance feed.

Multiband Verticals

When we get into the field of multiband verticals we find that considerable confusion exists. As pointed out previously, *any* antenna can be *called* a multiband antenna, but how we get power into such an antenna is another matter.

Up until the '50s any amateur multiband antenna was a system that usually consisted of an antenna, tuned feeders, and an antenna coupler. In the early '50s more and more amateurs started to use coaxial cable for feeders, along with band-switching transmitters. The next logical step was the use of a multiband antenna system that required no adjustments and always presented a matched condition to the feed line — in other words, an antenna that had a 50-ohm feed-point impedance on every desired band and frequency within a given band. A logical development was the multiband trap antenna.

By inserting traps in an antenna it was possible to make an antenna "look" like a resonant half-wave dipole in whichever band was used; or, in the case of multiband verticals, making the vertical look like a resonant quarter-wave antenna for the desired band. However, and this is important as far as the newcomer is concerned, to our knowledge there is *no* multiband trap antenna that will provide a perfect match on all bands, *regardless* of what some antenna manufacturers may tell you. Many hams have spent countless hours trying to adjust trap antennas for that "perfect" match when actually, it is just about impossible to obtain such a condition.

Nontrap Multiband Verticals

Several antenna manufacturers sell multiband antennas that consist of a vertical piece of tubing, usually 16 to 20 feet (4.88 to 6.10 m) long. The tubing is used with a loading coil at the ground end. By making appropriate taps and adjustments on the coil the antenna can be matched (or closely matched) on any given band. This type antenna has *no* traps. This in turn means that the coil taps and adjustments *must* be altered when one changes bands. Some misguided amateurs buy these antennas expecting all they need do is put them up and the antenna will work on all bands, automatically. Let's make one point clear: such an antenna *is* a multiband antenna, but requires adjustment *at the antenna* when one changes bands.

Of course, the next question should be, "If the antenna is that simple can't I build my own?" Yes, it is a very simple multiband antenna to make and install. Two or three sections of inexpensive 10-foot (3.05-m) TV mast sections can serve as the vertical radiating element. The mast can be supported on an insulator, such as a beverage bottle, and the mast guyed with nylon line. Fig. 31 shows a diagram of the antenna system. L1 should be a coil made of bare wire, no. 12 or 14, so that it can be tapped at every turn. A convenient coil size is 2-1/2 inches (64 mm) in diameter, six turns per inch (25 mm), such as B & W 3905-1 stock. The number of turns required, assuming 80 meters as the lowest band to be used, should be about 30 turns with an antenna length of 25 feet (7.62 m).

Adjustment of the antenna requires the use of an SWR bridge. Connect the coax line across a few turns of L1 and make a trial position of the shorting tap. Measure the SWR, then try various positions of the shorting tap until the SWR reaches its lowest value. Then vary the line tap similarly. This should bring the SWR down to a low value. Small adjustments of both taps should provide an SWR close to 1. If not, try adding C1 and repeat the adjustment procedure, varying C1 each time until a match is achieved. Radials will enhance the performance of the antenna. The number of radials is up to the individual amateur.

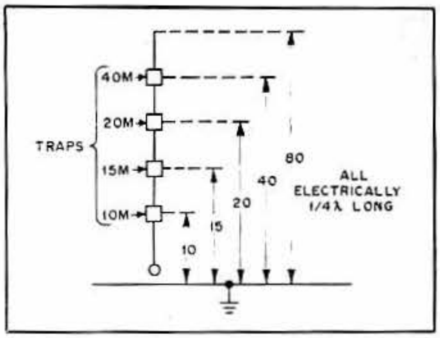

Fig. 32 — Drawing of the theoretical multiband trap vertical. In commercial practice, certain traps may be grouped together, giving the impression that only a single trap is used.

Trap Verticals

As mentioned earlier, traps can be installed in a multiband vertical. These traps are usually parallel-tuned circuits and the objective is to make each section of the antenna work as a quarter-wave vertical or odd multiple thereof on the desired band. Fig. 32 shows an example of this type antenna.

The purpose in using this type antenna is to provide a system that always presents a matched condition for the feed line. Unfortunately, there is so much interaction between various sections of the antenna that it is impossible to come up with a *perfect* match on each band. What is an *acceptable* match is another story.

Amateurs as a whole are inclined to attach too much importance to an SWR of 1. They feel that if their SWR bridge isn't showing an absolute zero reflected power that something is horribly wrong and they won't work out. The plain fact is that using a feed line such as RG-8/U (assume a 100-foot [30.48 m] length) one could have an SWR of as much as 5:1 and have *no* appreciable loss in the system. However, there is one clinker in this thinking!

In many instances commercially made transmitters and transceivers are designed by the manufacturer to work into a 50-ohm load only. They don't allow much leeway from this figure. When there is a mismatch in the antenna system, it can become impossible to load and tune the final amplifier of the transmitter. There just isn't enough tuning range in the tank circuit of the amplifier to handle the reactance that may be present in the load. There is a way around this problem however, and that is using a Transmatch in the feed line to disguise the mismatch.[2] The Transmatch can be adjusted so that the transmitter "sees" a 50-ohm load regardless of the mismatch at the antenna.

The Harmonic Problems

Another consideration should be mentioned. As pointed out earlier, *any* antenna can be a multiband antenna. By the same token any harmonics generated in the transmitter that reach the antenna can be radiated. It is true that a single-band antenna will reject harmonic energy, but *not* completely. In the case of a multiband trap antenna there is no rejection of some harmonics, simply because the antenna is designed to be resonant on all hf amateur bands. The solution to this problem is the use of a selective circuit installed in the feed line. A Transmatch is such a circuit and should provide adequate harmonic rejection.

Some Other Thoughts

The question is frequently asked, "Should I mount my vertical on the ground, or get the base up in the air?" Getting the antenna up in the clear is always better than having it mounted at earth level and surrounded by rain gutters, house wiring, trees, power lines and so forth. However, getting the vertical antenna up in the air also means that radials, as many as possible, should be used. The average installation (if there is such a thing) usually consists of three or four radials (or more) cut for the lowest operating frequency. Such a system should give a good performance.

Another important matter is that of the earth ground. When verticals are mounted at ground level the ground losses can be very important. Too many amateurs buy their verticals, get a 5-foot (1.52-m) long TV-type ground rod and drive it into the earth at the base of the antenna. They think this provides a good ground connection. As a matter of fact, the TV-type ground rods are practically worthless for amateur work. A *good* ground rod is the type used by the power company for home installations. This is a rod that is heavily galvanized, 5/8 inch (16 mm) in diameter, and about 10 feet (3.05 m) long. The amateur should be able to buy these rods from any wholesale electrical supply house. If possible, tie your ground connection to the water-system piping, assuming metal piping is used.

You'll hear the statement from fellow hams that verticals are poor antennas and radiate poorly in all directions. This isn't true, because a vertical *can* by a good antenna, but you have to give it a fighting chance. This material was originally presented in *QST* by Lewis McCoy, W1ICP.

References

[1] Sevick, "The Ground-Image Vertical Antenna" *QST*, July 1971. Also appears elsewhere in this chapter.

[2] A suitable Transmatch is described in the transmission-line chapter of any edition of *The Radio Amateur's Handbook*.

Shunt-Feeding Towers for Operation on the Lower Amateur Frequencies

The 160- and 80-meter DX antenna you've dreamed about may already be standing on your property — just waiting to be used as a DX vertical. Read this presentation and learn how to shunt-feed your tower on 1.8 and 3.5 MHz!

Shunt-feeding a tower used to support beam antennas for 14 MHz or higher frequencies has obvious advantages for the amateur who wants to work all lower bands. If a good ground system is installed, the result may be a very effective antenna for DX work on 80 and 160. The shorter tower installations may work very well on 40 meters, if the effective height above ground is less that 5/8 wavelength. Beyond this height the radiation angle goes higher, and the effectiveness for DX goes down.

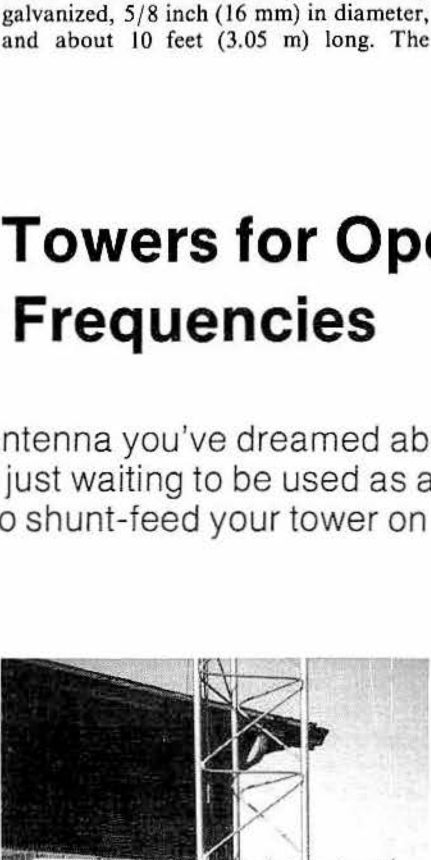

Close-up view of the base of the shunt-fed tower, showing the two bottom arms and their insulators.

The shunt-fed tower is at its best on 160, where a full quarter-wavelength vertical antenna is rarely possible. Almost any tower height can be used. If the beam structure provides some top loading, so much the better — but anything can be made to radiate, if it is fed properly. A self-supporting, aluminum, crank-up, tilt-over tower is used at W5RTQ with a TH6DXX tribander mounted at 70 feet (21.34 m). Measurements showed that the entire structure has about the same properties as a 125-foot (38.10-m) vertical. It thus works quite well on 160 and 80 in DX work requiring low-angle radiation. It can also be used on 40, but results are rather poor because of high radiation angle. WB5FKX, with a 37-foot (11.28-m) tower and a 21-MHz beam, finds that shunt feed on 40 enables him to work DX that he never knew existed before.

Preparing the Structure

Usually some work on the tower system must be done before shunt-feeding is tried. Metallic guys should be broken up with insulators. They can be made to simulate top loading, if needed, by judicious placement of the first insulators. Don't overdo it; there is no need to "tune the radiator to resonance" in this way. If the tower is fastened to a house at a point more than about one-fourth of the height of the tower, it may be desirable to insulate the tower from the building. Plexiglas sheet, 1/4-inch (6-mm) or more thick, can be bent to any desired shape for this purpose, if it is heated in an oven and bent while hot.

All cables should be taped tightly to the tower, preferably on the inside, and run down to the ground level. It is not necessary to bond shielded cables to the tower

Table 9

Shunt-Feed Details for Several Towers Used on 160 Meters

Station	*Height of Tower (ft)*	*"Top-Hat" (Beams)*	*Average Height of Beam (ft)*	*Height of Wire-to-Tower Connection (ft)*	*Wire-to-Tower Spacing (in.)*	*Tuned Value of Gamma Capacitor*
W5RTQ	69	TH6DXX	70	68	30	150 pF
K5PFL	86	Stacked 40, 20, 15, 10 M beams	95	75	24	125 pF
K4PUZ	64	TH6DXX	66	63	8	300 pF
K8KAS	70	TH6DXX	77	68	24	225 pF
W1CER	50	QST 20-M DX Weasel	55	47	44	400 pF

1 ft = 0.3048 m
1 in. = 25.4 mm

electrically, but there should be no exceptions to the down-to-the-ground rule.

No rf problems have developed with rotators, beam traps, or even TV sets where the TV antenna is part of the radiating structure, as is the case at WB5FKX. His only precaution was use of coax on the TV antenna. It would be well to proceed with caution, as every installation could be different in this respect.

Though the effects of ground losses are less severe with the shunt-fed vertical than with the simple quarter-wave antenna, a good system of buried radials is very desirable. The ideal would be 120 radials, each 250 feet (76.20 m) long, but fewer and/or shorter ones must often suffice. You can sneak them around corners of houses, along fences or sidewalks, wherever they can be put a few inches under the surface, or even on the earth surface. Aluminum clothesline wire is used extensively at W5RTQ, and it stands up well. Neoprene-covered aluminum wire may be safer in highly acid soils. Contact with the soil is not important. Deep-driven ground rods, and connection to underground copper water pipes, are good, if usable.

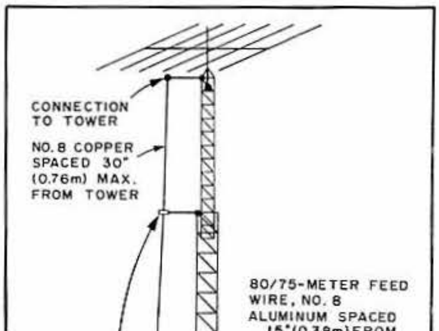
CONNECTION TO TOWER
NO. 8 COPPER SPACED 30" (0.76m) MAX. FROM TOWER
80/75-METER FEED WIRE, NO. 8 ALUMINUM SPACED 15"(0.38m) FROM TOWER
STANDOFF INSULATOR
STANDOFF INSULATOR
150-pF VARIABLES
250-pF VARIABLE
BURIED RADIALS CONNECTED TO BASE OF TOWER
COAXIAL FEED

Fig. 33 — Principal details of the shunt-fed tower at W5RTQ. The 160-meter feed, left side, connects to the top of the tower through a horizontal arm of 1-inch (25-mm) diameter aluminum tubing. The other arms have stand-off insulators at their outer ends, made of 1-foot (0.31-m) lengths of plastic water pipe. The connection for 80/75, right, is made similarly, at 28 feet (8.53 m), but two variable capacitors are used, to permit adjustment of matching with large changes in frequency.

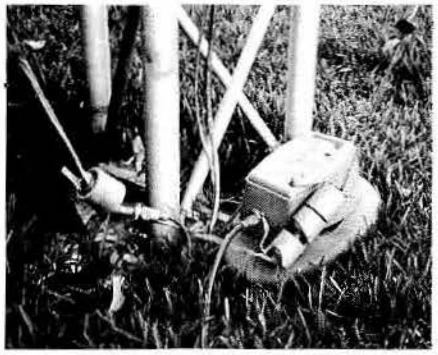

Though variable capacitors are shown in Fig. 33, they were all replaced with fixed-value units except one, which is inside a plastic box at the tower base.

Installing the Shunt Feed

Principal details of the shunt-fed tower for 80 and 160 meters are shown in Fig. 33. Rigid rod or tubing can be used for the feed portion, but heavy gauge aluminum or copper wire is easier to work with. Flexible stranded no. 8 copper wire is used for the 160-meter feed at W5RTQ, because when the tower is cranked down, the feed wire must come down with it. Connection is made at the top, 68 feet (20.73 m), through a 4-foot (1.22-m) length of aluminum tubing clamped to the top of the tower, horizontally. The wire is clamped to the tubing at the outer end, and runs down vertically through standoff insulators. These are made by fitting 12-inch (0.31-m) lengths of PVC plastic water pipe over 3-foot (0.91-m) lengths of aluminum tubing. These are clamped to the tower at 15 to 20-foot (4.57 to 6.10-m) intervals, with the bottom one about 3 feet (0.91 m) above ground. The lengths given allow for adjustment of the tower-to-wire spacing over a range of about 12 to 36 inches (0.31 to 91 m), for impedance matching.

The gamma-match capacitor for 160 is a 250-pF variable with about 1/16-inch (1.6-mm) plate spacing, which is adequate for the power levels presently authorized. The omega match used for 80 and 75 permits retuning for large excursions in frequency encountered in using both cw and phone on this band. Two capacitors are required, each about 150 pF, with plate spacing of about 3/16 inch (4.8 mm), if full power is used. They can be mounted in plastic refrigerator containers for protection against the weather. Use well-insulated knobs to avoid rf burns during the adjustment process.

Separate 50-ohm lines run underground to the station. The shield side of their connectors should be grounded to the base of the tower and to the buried radials as directly as possible.

Tuning Procedure

It is suggested that the 160-meter wire be connected to the top of a structure 75 feet (22.86 m) tall or less. Note, from Table 9, that the monster at K5PFL was fed at 75 feet (22.86 m) above ground. Mount the standoff insulators so as to have a spacing of about 24 inches (0.61 m) between wire and tower. Pull the wire taut and clamp it in place at the bottom insulator. Leave a little slack below to permit adjustment of the wire spacing, if necessary.

Adjust the series capacitor in the 160-meter line for minimum reflected power, as indicated on an SWR meter connected between the coax and the connector on the capacitor housing. Make this adjustment at a frequency near the middle of your expected operating range. If a high SWR is indicated, try moving the wire closer to the tower. Just the lower part of the wire need be moved for an indication as to whether reduced spacing is needed. If the SWR drops, move all insulators closer to the tower, and try again. If the SWR goes up, increase the spacing. There will be a practical range of about 12 to 36 inches (0.31 to .91 m). If going down to 12 inches (0.31 m) does not give a low SWR, try connecting the top a bit farther down the tower. If wide spacing does not make it, the omega match shown for 80-meter work should be tried. No adjustment of spacing is needed with the latter arrangement which may be necessary with short towers or installations having little or no top loading.

The two-capacitor arrangement is also useful for working in more than one 25-kHz segment of the 160-meter band. Tune up on the highest frequency, say 1990 kHz, using the single capacitor, making the settings of wire spacing and connection point permanent for this frequency. To move to the lower frequency, say 1810 kHz, connect the second capacitor into the circuit and adjust it for the new frequency. Switching the second capacitor in and out then allows changing from one segment to the other, with no more than a slight retuning of the first capacitor.

The omega match is recommended for 80-meter operation, because of the wide tuning range required. It was found that the point of connection could be at 28 feet (8.53 m), using a single support at the top of the first tower section. This served as a mount for the second standoff insulator for the 160-meter feed, as well.

A 40-meter feed, not shown, was connected at about 20 feet (6.10 m) up, with about 8 inches (203 mm) spacing between wire and tower.

Substituting Fixed-Value Capacitors

Depending on the frequencies used and the matching problems in a given installation, it may be possible to put fixed-value capacitors in place of some or all of the variables shown in Fig. 33, and then switch them remotely, or by hand, when changing frequency. This was done at W5RTQ. The 160-meter gamma capacitor is now a fixed-value, 150-pF type designed for high rf current service. It does not have to be weatherproofed, and it gives a satisfactory match over the small frequency range used on this band.

Two or three similar capacitors connected in series are used in covering the cw and phone frequencies in the 80-meter band. Operation around 3800 kHz requires about 50 pF in the omega (parallel) capacitor, so the three 150-pF units are used in series. Moving to cw operation near 3500 kHz requires about 75 pF, so only two 150-pF capacitors are used in series. The only variable now in the system is the 150-pF capacitor mounted in the plastic refrigerator box at the base of the tower. It is the series or gamma capacitor in the 80-meter feed. Even this could be replaced with a fixed unit, as it has been found unnecessary to adjust it in the course of normal operation of the station, if only small changes in frequency are made in using the two modes.

Because every tower installation is likely to be different from every other one in some respects, the values and dimensions given here may be subject to change, but the general principles should hold. Proof of this was found in the May 1975 issue of *Ham Radio*,[1] wherein W4OQ describes shunt-feeding in very similar terms. He also presented estimates of the electrical height of various tower-beam combinations that prospective shunt-feed users should find of interest.

It is hoped that this information will inspire others to try shunt-feeding their towers. This method offers an effective way to fire up on the lower frequencies that is useful to many who may have wanted to work these amateur frequencies but have shied away from doing so because of limited antenna space. This material was originally presented in *QST* by Earl W. Cunningham, K6SE.

---

[1] True, "Shunt-Fed Vertical Antennas," *Ham Radio*, May 1975.

The Minooka Special

This 160-meter vertical antenna could be just the one for you, if you reside on property of city-lot size or are a mobile operator. You can make it from readily available, inexpensive material.

Inspiration for the design offered here resulted in part from the author's survey of antennas in use by 160-meter operators, worldwide. An extensive report was compiled, which filled a loose-leaf notebook! A boil-down of the most useful information resulted in an 11-page report which was made available to numerous amateurs. Subsequently, the writer was encouraged to submit some of the data for publication. Approximately 60 operators were polled. The following information from that inquiry should be of interest to those who are curious about "preferred" antennas for top-band use.

Question 1: "If you could put up any antenna for 160, what would it be?" *Result:* 60% favored verticals, 30% said horizontals, and 10% indicated mixed feelings. (The term "vertical" includes various configurations — 1/4, 1/2, and 5/8 wavelength, vertical arrays, and inverted Ls.)

Question 2: "Comparing antennas that an average ham could build, do you prefer verticals or horizontals for 160?" *Result:* 70% said vertical, 17% favored horizontal, 5% inverted L, 2% horizontal and vertical, 2% inverted V, and 3% had no opinion. Reasons given for the responses were, "Because of signal comparisons and past experience. A backyard-compatible vertical is more effective than a backyard-compatible horizontal."

Question 3: "Do you operate mobile on 160? If so, what is the antenna used?" *Result:* Three used base-loaded verticals, eight used center-loaded ones, and one employed a Heliwhip. Some of the operators used capacitance hats. Other questions in the survey dealt with receiving antennas, types of soil and terrain, besides requesting complete details of the present transmitting antenna.

Fig. 34 shows some of the vertical systems used when the poll began in 1969. Additional configurations used by successful 160-meter operators are illustrated in Fig. 35. It can be seen that a great deal of useful "backyard theory" is represented in the systems shown. The experience and trials of others can save countless years of experimenting with antennas (not that some of us can't use the exercise)! It is hoped that these data will be of value to future 160-meter operators.

The Minooka Special

The following is a description of an effective vertical antenna for 160 meters, designed with these objectives in mind:

1) Highly effective for 160-meter DX and local work.
2) Easy to build and adjust.
3) Very economical.
4) Fits neatly into back yard.
5) Reasonable bandwidth.
6) Good for portable and DXpedition work.
7) Can be scaled down for mobile operation.

The resultant antenna (Fig. 36) is top loaded inductively and can be built by anyone from readily available material. Only a dip meter and SWR indicator are needed for tune-up. Many versions were built and tested, ranging from 7-foot (2.13-m) mobile types to 60-foot (18.29-m) backyard

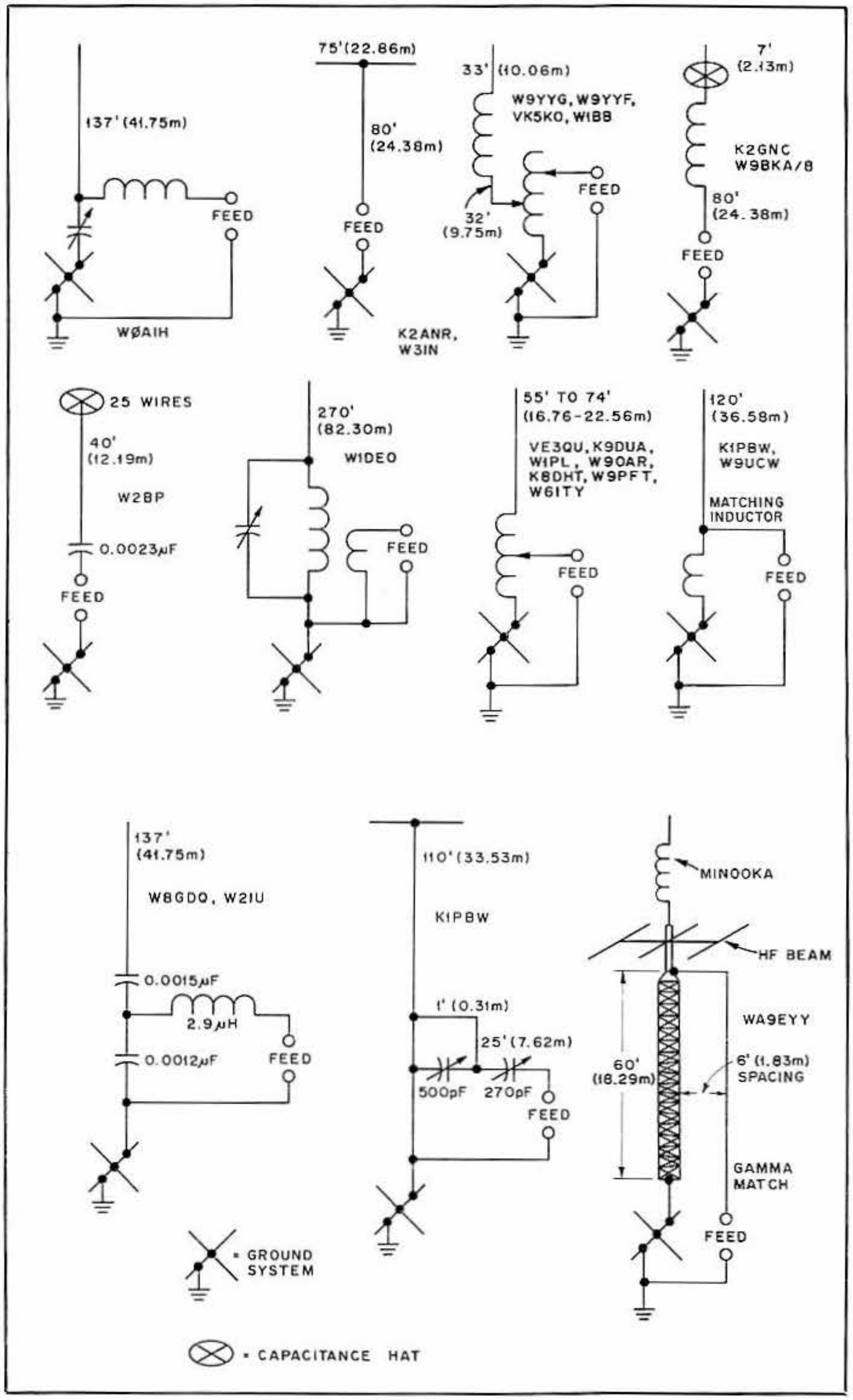
137' (41.75m)
FEED
WØAIH
75'(22.86m)
80'
(24.38m)
FEED
K2ANR,
W3IN
33' (10.06m)
W9YYG, W9YYF,
VK5KO, W1BB
32'
(9.75m)
FEED
7'
(2.13m)
K2GNC
W9BKA/8
80'
(24.38m)
FEED
25 WIRES
40'
(12.19m)
W2BP
0.0023μF
FEED
270'
(82.30m)
W1DEO
FEED
55' TO 74'
(16.76-22.56m)
VE3QU, K9DUA,
W1PL, W9OAR,
K8DHT, W9PFT,
W6ITY
FEED
120'
(36.58m)
K1PBW,
W9UCW
MATCHING
INDUCTOR
FEED
137'
(41.75m)
W8GDQ, W2IU
0.0015μF
2.9μH
0.0012μF
FEED
110'(33.53m)
K1PBW
1' (0.31m)
25'(7.62m)
500pF
270pF
FEED
MINOOKA
HF BEAM
WA9EYY
60'
(18.29m)
6' (1.83m)
SPACING
GAMMA
MATCH
FEED
= GROUND
SYSTEM
= CAPACITANCE HAT

Fig. 34 — Illustrations of vertically polarized 160-meter antennas used by various operators for DX and local work. L and C values are not specified in some of the examples because the data were not furnished by the users.

Table 10

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6
X	5 ft (1.52 m)	8 ft (2.43 m)	4 ft (1.22 m)	4 ft (1.22 m)	19 ft (5.79 m)	3 ft 3 in. (0.99 m)
Y	2 ft (0.61 m)	15 in. (381 mm)	3 ft 6 in. (1.07 m)	4 ft (1.22 m)	11 in. (279 mm)	3 ft (0.91 m)
Z	As long as possible					
Wire Size	No. 20	No. 19	No. 18	No. 16	No. 19	No. 22

Here is a 50-foot (15.24-m) version of the Minooka Special. It is tuned remotely at K9SKX, and is used on 160 through 20 meters. Nylon or polypropylene guy lines should be used.

or DXpedition models. They have been used with good results from 20-foot (6.10-m)-wide backyards in cluttered Chicago, to vast beaches on Carribean islands, and in South America. The fixed-tuned bandwidths vary from 10 kHz for mobile versions, to 50 kHz for the larger fixed-station models (SWR 2:1 or better).

Construction

The physical layout of the antenna is centered around the use of 3/4-inch (19-mm) diameter rigid PVC water pipe. This sturdy tubing has an outside diameter of 1 inch (25 mm). The inner diameter provides an "interference fit" for 1/2-inch (13-mm) EMT conduit (thinwall). A 3/8-24 nut is driven into a 5-inch (127-mm) length of conduit, then aligned and braized into position. Next, the conduit is taper-ground and polished on the opposite end, then driven into the PVC tubing (see table of dimensions).

In the construction of mobile antennas another piece of conduit is driven into the bottom of the PVC coil form and a 3/8-24 bolt is brazed into the bottom end. The bolt will mate with standard mobile antenna mounts. For larger versions of the antenna the PVC material can be mounted in TV masting, or whatever. A standard 8-foot (2.44-m) stainless-steel whip between 3 and 8 feet (0.91 and 2.44 m) in length can be screwed into the top section. For home-installed versions of the antenna a

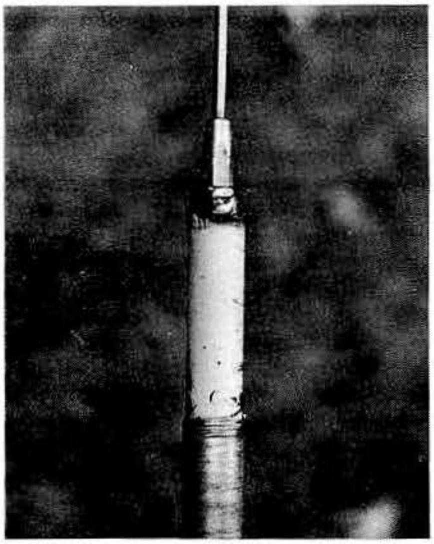

Close-up view of the top section of a coil assembly. This unit was used during mobile operation for over three years (40,000 miles or 64,374 km).

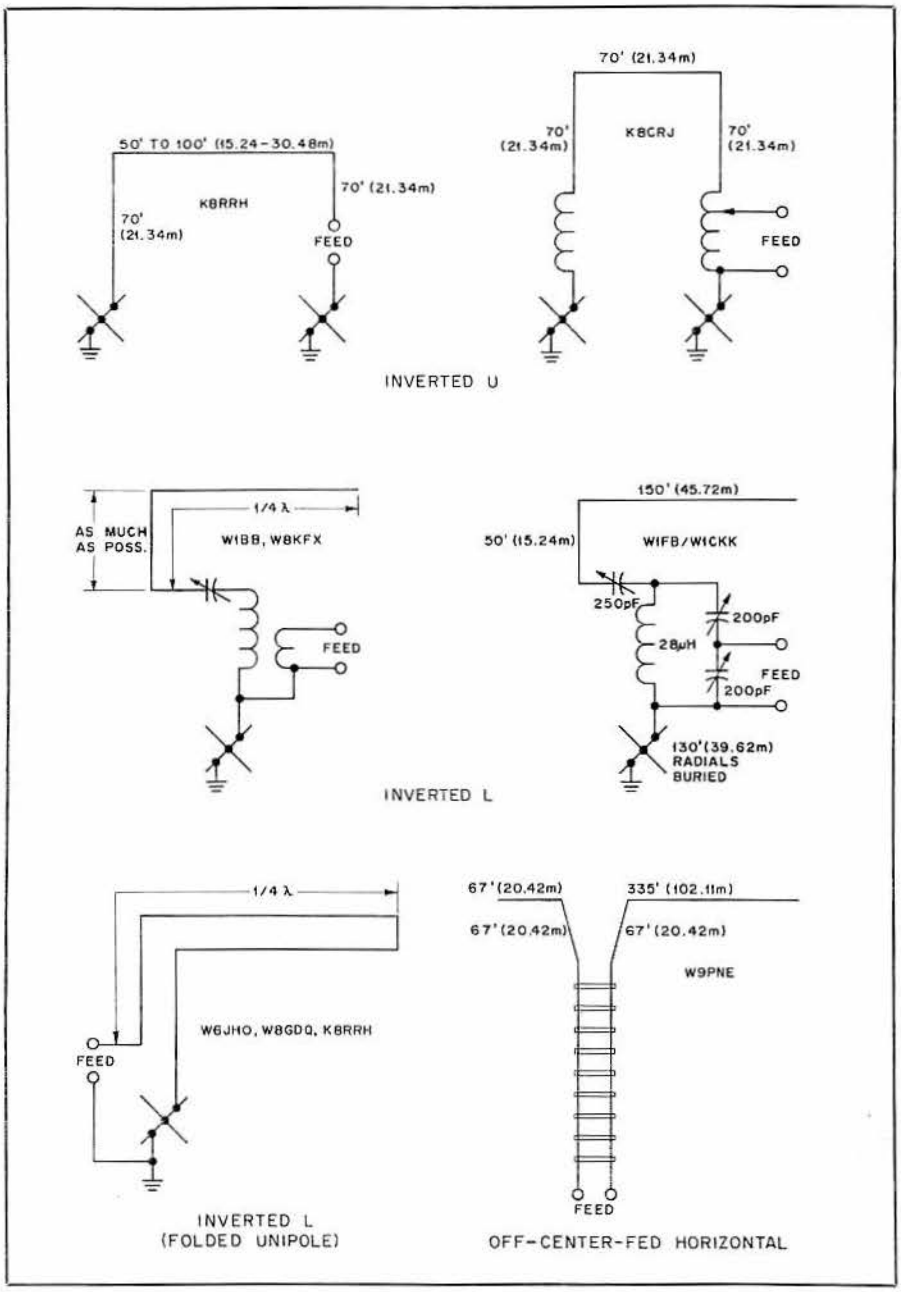
50' TO 100' (15.24-30.48m)
70' (21.34m)
K8RRH
70' (21.34m)
FEED
70' (21.34m)
70' (21.34m)
K8CRJ
70' (21.34m)
FEED
INVERTED U
1/4 λ
AS MUCH AS POSS.
W1BB, W8KFX
FEED
150' (45.72m)
50' (15.24m)
W1FB/W1CKK
250pF
200pF
28μH
FEED
200pF
130'(39.62m) RADIALS BURIED
INVERTED L
1/4 λ
FEED
W6JHO, W8GDQ, K8RRH
INVERTED L (FOLDED UNIPOLE)
67'(20.42m)
335' (102.11m)
67'(20.42m)
67'(20.42m)
W9PNE
FEED
OFF-CENTER-FED HORIZONTAL

Fig. 35 — Examples of additional antennas for 160 meters which proved effective for long-haul work.

A mobile version of the Minooka Special is shown mounted on the author's car.

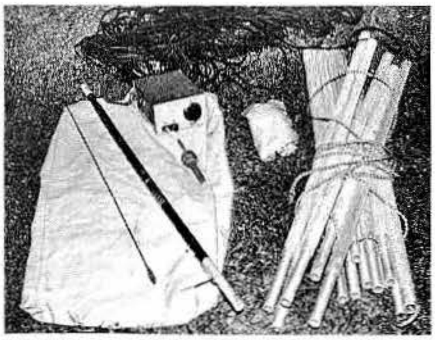

A portable version of the 160-meter vertical antenna. It was used by W9YYG and W9UCW at HK0BKX in January 1974.

3- or 4-foot (0.91- or 22-m) piece of thin-wall tubing can be used in place of the whip. This will save on the cost of materials, and will eliminate the need to have brazing done. The coil wire should be soldered to the conduit to assure a good electrical connection.

Adjustment

Pick a set of dimensions from the table which suits your application, but add a few inches more of coil turns (all turns close wound) than are recommended. This will allow leeway for pruning the system to resonance.

To simplify adjustment it is suggested that the system be assembled first with only the coil and top section (no base section). Place the antenna where it is in the clear (on the car or fixed-station site), and where it can be tuned against the proposed ground system — car body or ground radials.

A three-turn link should be connected temporarily between the lower end of the coil and the ground system. This will permit rough tuning of the system to resonance by inserting a dipper coil into the link and adjusting the coil turns on the antenna until a dip is noted in the desired part of the band. Upon completion of the pruning the constructor can, if he wishes, cover the coil with weatherproof tape or shrink tubing. The antenna should be tuned for roughly 2000 kHz if the entire band is to be used. Adjust the resonance for 1850 kHz if only the low segment will be utilized.

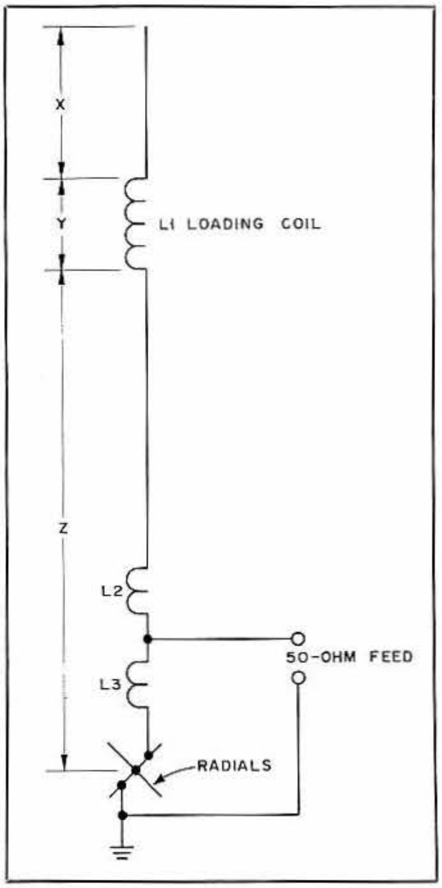

Fig. 36 — Electrical details of the Minooka Special. Table 10 gives specific information concerning dimensions X, Y and Z. L2 may vary in size from 1 to 20 turns, and L3 will contain between 5 and 10 turns, typically. L2 and L3 are made from no. 18 wire, spaced 1/8 inch (3 mm) between turns. The coil diameter is 1-1/2 inches (38 mm). Refer to text for tuning instructions.

Erect the antenna with all of its parts — coil, top and bottom section — and insert inductors L2 and L3 of Fig. 36 as shown. With L3 temporarily out of the system, adjust L2 for the lowerest value of SWR obtainable at the desired operating frequency. Then, place L3 back in the circuit and adjust it for an SWR reading of 1. Addition of the base section later on will not affect the resonant frequency of the overall system materially, provided the base section does not exceed, say, 60 feet (18.29 m).

The foregoing steps are used also in adjusting the mobile version of the antenna. However, because of the small size it is possible to adjust the antenna in completely assembled form. Only L3 of Fig. 36 is needed for mobile antennas. The main coil, L1, is adjusted for resonance, then L3 is set for lowest SWR. L3 can be mounted inside the trunk or under the bumper in a weatherproof enclosure.

For fixed-station operation it is recommended that a good ground system be employed. One should use at least 10 radials of, say, no. 18 or larger wire, 10 to 50 feet (3.05 to 15.24 m) in length. If you can manage 40 radials, 60 feet (18.29 m) in length, all the better.[1]

Concluding Comments

Three fixed-station versions of the Minooka Special have been tested and used at W9UCW. Each was compared against the regular antenna, which is a one-quarter wavelength vertical (130 feet [39.62 m] high), and operates against a radial system that contains 12,000 feet (3657.6 m) of wire. The short verticals were always inferior to the big antenna by approximately 5 dB.

Using version no. 4 from the table (4-foot [1.22-m] base section), the author worked W1s through W6s from the mobile setup. The rig was an HW-18 transceiver. Contacts were also made with VP9 and KV4 stations from the car.

While DXing from San Andres (HK0) reports were received of signal strengths exceeding S9 plus 20 dB from Maine to Washington. Good reports were received from Europe also.

Because the loading coil acts as an rf choke at 3.5 MHz and above, several versions of the antenna have been used successfully from 160 through 20 meters with an appropriate L-network matching section installed at the base of the system.

Acknowledgements

The author wishes to express his gratitude to all who helped make this article possible, . . . particularly K9SKX, WA9EYY, and W9YYF, who assisted with the initial experiments on the Minooka Special. The participants in the antenna survey (a list too large to include here) deserve considerable thanks for their help as well. This material was originally presented in *QST* by Barry A. Boothe, W9UCW.

---

[1] [Editor's Note: The reader should be aware that 1/4-wavelength radials (or longer ones) are recommended for best efficiency in a system of this kind. Detailed information on the subject is given in *The ARRL Antenna Book*, 13th Edition. The recommendations given by the author are no doubt based on the lack of space available in the typical urban lot, around which his design was established. See Hills, "The Grounds Beneath Us." *RSGB Bulletin*, June 1966.]

Inexpensive 5/8-Wave Vertical Antenna with Coaxial Transformer

The author's hastily constructed 5/8-wave vertical antenna (inspired by W0JF's results[1]) consists of a supporting structure made from sections of 2 × 2-inch (51 × 51-mm) lumber on which four lengths of 300-ohm TV ribbon are affixed. Each section of ribbon line is 41 feet 6 inches (12.65 m) long. A length of this line is attached to each side surface of the 2 × 2 support mast, then the ends (top and bottom) of all four line sections are connected in parallel to form a single vertical conductor. The coax-line dimensions are near those of W2JTJ,[2] and are shown in Fig. 37. The SWR varies from a flicker of reflected power at 14,000 kHz to 1.5:1 at 14,350 kHz. (The next task is to get that lowest-SWR point shifted to 14,275 kHz.) The whole thing is leaning against a tree, almost vertical, and seems to work well. This material was originally presented in *QST* by Robert J. Earl, W1DRV.

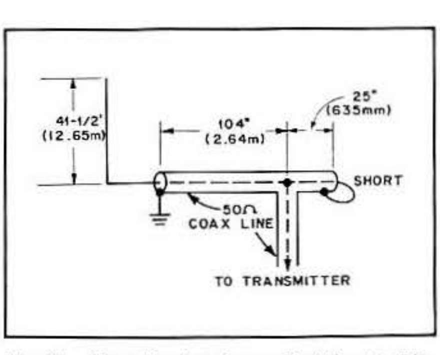

Fig. 37 — Coax-line transformer for 20-meter 5/8-wave vertical antenna.

References

[1] Beers, "Short Antennas for the Lower Frequencies," *QST*, August and September 1970.
[2] Czerwinski, "Coaxial Transformer for Voltage-Fed Antennas," *QST*, June 1961.

A Backyard 160-Meter Vertical

A 160-meter band was the favorite one at the author's QTH this winter. The same could be said for many other amateurs and the simple antenna described here may be of interest.

Some of the old-timers will tell the newcomer to 160 meters that the best antenna for local and DX work is the vertical. Since a full-size vertical would be very large, the question is finding a type that will fit into a closely packed residential area.

As a starter, the writer took an old 40-foot (12.19-m) telescoping mast and insulated it from the ground. (This can be done by setting it on top of a large soft-drink bottle, or clamping it to a wooden post.) The antenna was fed at the base with RG-58A/U 52-ohm coaxial cable. The shield was wired to a radial system and a ground rod. A hint for the radials — buy some aluminum clothesline and cut a number of slits in your yard with a flat spade. Next, push the wire into the slits and tamp the grass back together. If there are any swampy parts of the yard, these seem to make the best ground areas.

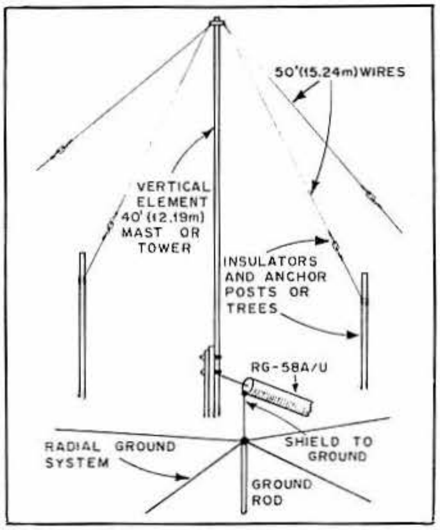
50'(15.24m)WIRES
VERTICAL ELEMENT 40' (12.19m) MAST OR TOWER
INSULATORS AND ANCHOR POSTS OR TREES
RG-58A/U
RADIAL GROUND SYSTEM
SHIELD TO GROUND
GROUND ROD

Fig. 38 — A space-saving vertical for 160 meters. While not always possible, it is desirable to have the wires for the top hat run off at right angles to the mast.

Rather than using a loading network at the base of the antenna, the author constructed a capacitive hat to load it at the top instead. The top hat was made from four 50-foot (15.24-m) pieces of wire strung out from the top of the mast in guy-wire fashion. They are connected electrically to the top, and stretched until they are as close as possible to a 90-degree angle with respect to the mast.

The system described works well with a 40-foot (12.19-m) mast, but other lengths could be used instead. However, the top hat may have to be adjusted in order to bring the system to resonance. While my vertical occupies very little space, it seems to perform as well as do many "full-size" antennas on 160 meters. This material was originally presented in *QST* by Mike Mussler, WB8JJA.

Chapter 2

Yagi Antennas

The antenna systems treated in this chapter are variations of the basic directional gain antenna devised many years ago by Yagi and Uda. Over the years the fundamental design and its variations have been referred to simply as the Yagi. How Mr. Uda was excluded from the credits remains a mystery, as he was a co-developer of the antenna under discussion.

Yagi types of antennas appeal to all manner of hf- vhf- and uhf-band operators. The primary response characteristic of the Yagi is unidirectional, discounting minor response lobes off the front, sides and rear of the antenna. The major lobe concentrates the radiated energy in a desired direction to provide gain in dB and discrimination against interference from unwanted directions.

Yagis can be used singly or in array combinations. In the latter case, each time the size of an array is doubled the system gain can be made to increase by 3 dB — assuming that the phasing, impedance matching, and aperture parameters are correct. Thus, a pair of 3-element Yagis which are "stacked" in an array to meet the foregoing criteria will exhibit 3 dB more gain than a single 3-element Yagi. To effect another 3 dB increase in gain, four 3-element Yagis would need to be placed in a given array of Yagis, and so on.

This chapter deals with the cream of the Yagi crop, so to speak, as taken from past issues of *QST*. When zoning laws and available property permit, the Yagi is preferred over the single-element vertical antennas described in chapter 1.

The 20-Meter DX Weasel

Short antennas have always captured the interest of amateurs.

The antenna presented here is a monoband version of the Bite-Size Beam for 20 and 15 meters published elsewhere in this chapter. It is designed for 20-meter operation only. Most of the parts were "lifted" from the original array.

The reader is cautioned against attempting to figure why some of the dimensions were selected. As a point of interest, they just happened that way. One new twist was added to this system — capacitance hats. The hats provided one noticeable improvement in performance which became obvious during the initial tune-up phase of the project. The hats lowered the Q of the elements and hence broadened the overall response of the system. While only conjecture, one might say that it is this broad tuning which causes the parasitic elements to perform as they do. Element tuning on the Bite-Size Beam[1] was critical. If the dimensions given here are followed carefully, the antenna should provide excellent characteristics for gain and pattern. Front-to-back measurements indicate approximately 20 dB. The front-to-side ratio was measured in excess of 25 dB. There are no objectionable side lobes. The SWR is less than 2:1 at each end of the band when the gamma capacitor and rod are set for proper operation at 14.1 MHz. There is no need to select phone or cw as the primary mode of operation.

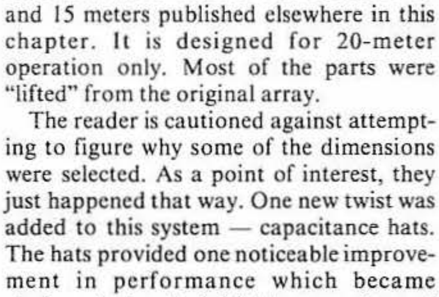
Details of the loading coil and capacitance hat on one of the elements.

Construction

The boom and all the elements are made from 1-1/4-inch (32-mm) diameter aluminum tubing available at most hardware stores. The two boom sections and the two pieces which make up the center portion of the driven element are coupled using 15-inch (0.38-m) sleeves of 1-3/8-inch (35-mm) OD aluminum tubing. Sheet-metal screws should be used to secure the sections within the coupling sleeves.

The loading coils are wound on 1-1/8-inch (29-mm) diameter Plexiglas rod. Details are shown in Fig. 1. Be sure to slit the ends of the aluminum tubing where the compression clamps are placed. The coils are made from no. 14 enameled wire. The specified number of turns are equally spaced to cover the entire 9 inches (229 mm) of Plexiglas.

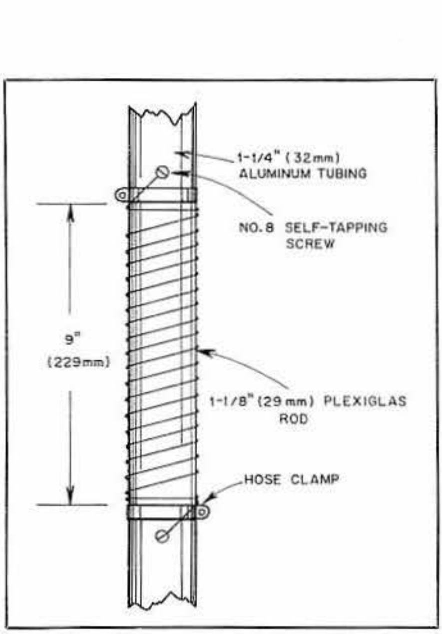

Fig. 1 — Details for joining sections of aluminum tubing and Plexiglas rod.

The capacitance hats are constructed from 3/4-inch (19-mm) angle aluminum. Two pieces 2 feet (0.61 m) in length are required for each hat. The model shown in the diagrams has the angle aluminum

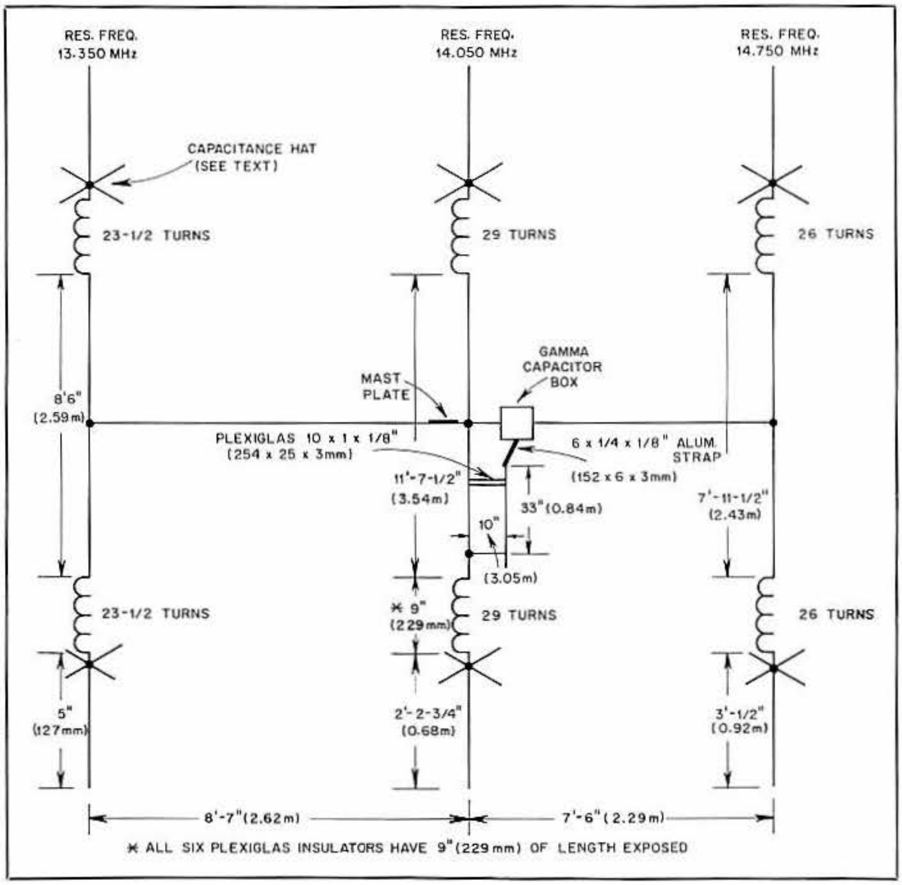

Fig. 2 — Constructional details for the 20-meter beam. The coils on each side of the element are identical. The gamma capacitor is a 140-pF variable unit manufactured by E. F. Johnson Co.

Table 1

Complete Parts List for the Short Beam

Qty	*Material*
2	10-foot (3.05-m) lengths of 1-1/4-inch (32-mm) dia. aluminum tubing (one for the reflector center section, one for the reflector end sections).
3	8-foot (2.44-m) lengths of 1-1/4-inch (32-mm) dia. aluminum tubing (two lengths for the boom, one length for the director element center).
4	6-foot (1.83-m) lengths of 1-1/4-inch (32-mm) dia. aluminum tubing (two lengths for the driven element center, two lengths for the director and driven element ends).
2	15-inch (0.38-m) lengths of 1-5/8-inch (41-mm) dia. aluminum tubing.
1	40-inch (12.19-m) length of 1-3/8-inch (10-mm) dia. aluminum tubing.
4	6-foot (1.83-m) lengths of 3/4-inch (19-mm) angle aluminum.
6	12-inch (0.31-m) lengths of 1-1/8-inch (29-mm) dia. Plexiglas rod.
1	9-inch (229-mm) square, 1/4-inch (6-mm) thick aluminum plate.
8	U bolts.
12	Compression hose clamps.
8	Crutch caps.
38 ft	No. 12 enameled copper wire.
60 ft	No. 14 enameled copper wire.

fastened to the element using aluminum strips and no. 8 sheet-metal screws, but sheet-metal screws alone can be used. Solder lugs are fastened to the ends of the angle aluminum and no. 12 or 14 wire connects the ends of the aluminum, resulting in a square loop. The wires should be soldered at each of the solder lugs.

All of the elements are secured to the boom with TV U-bolt hardware. Plated bolts are desirable to prevent rust from forming. An aluminum plate 9 inches (229 mm) square by 1/4-inch (6-mm) thick was used as the boom-to-mast plate.

A boom strut is recommended because the weight of the elements is sufficient to cause the boom to sag. A 1/8-inch (3-mm) diameter nylon line is plenty strong. A U-bolt clamp is placed on the mast several feet above the antenna and provides the attachment point for the center of the truss line. To reduce the possibility of water accumulating in the element tubing and subsequently freezing, crutch caps are placed over the ends. Rubber feet suitable for keeping furniture from scratching hardwood floors would serve the same purpose.

A piece of Plexiglas was mounted inside an aluminum Minibox to provide support and insulation for the gamma capacitor. A plastic refrigerator box would serve the purpose just as well. The capacitor housing is mounted to the boom by means of a U bolt. The gamma rod is made of 3/8-inch (10-mm) aluminum 40 inches (1.016 m) long and is connected to the gamma capacitor by a 6-inch (152-mm) length of strap aluminum.

Tune-Up and Operation

The builder is encouraged to follow the dimensions given in Fig. 2 as a starting point for the position of the gamma rod shorting strap. Connect the coaxial cable and install the antenna near or at the top of the tower. The gamma capacitor should be adjusted for minimum SWR at 14.1 MHz as indicated by an SWR meter (or power meter) connected in the feed line at the gamma capacitor box. If a perfect match cannot be obtained, a slight repositioning of the gamma short might be required.

Evaluation of performance is always a difficult task. The 20-Meter DX Weasel is mounted on top of a 50-foot (15.24-m) high tower at the QTH of W1FB. The transmitter input power for the testing period was 3 watts. Many European stations as well as a ZL have been contacted with ease. In fact, the Weasel has been responsible for cracking a few pileups on the first call!

Obtaining the Material

Table 1 lists all of the major parts used in the 20-meter Weasel. Most of the U-bolt hardware as well as the aluminum tubing should be available from large hardware stores. The Plexiglas rod, however, presents a bit more of a problem as indicated by mail received by the ARRL Technical Department. Many large plastic supply dealers stock the cast rod. The problem is that there are not very many large dealers! Additionally, Plexiglas is a petroleum-base product which means the cost is no doubt higher than one would expect. Those amateurs who are unable to locate aluminum hardware or Plexiglas should consult the yellow pages to see who their nearest supplier is. This material was originally presented in *QST* by Robert Myers, W1XT and Jay Rusgrove, W1VD.

[1] See page 45.

The HW-40 Micro Beam

A helically wound 2-element 40-meter Yagi.

Disconsolate pleas from apartment dwellers and other amateurs who live on property that is too small in area to permit the use of large 40-meter antennas prompted the authors to develop the antenna described here. The particular dimensions represented in this design were dictated primarily by the kind of fiberglass tubing that was immediately available. Kirk quad spreaders were obtained for use as forms on which to wind the helical elements. Longer (or shorter) elements can be used with good results. This article is intended as a guide in the design of similar antennas.

The HW-40 is scaled down to 28 percent of full size. Two elements are used, each being 18 feet (5.49 m) in length, tip-to-tip. Thus, the elements are just 2 feet (0.61 m) longer than those of a full-size 10-meter Yagi. A 16-foot (4.88-m)-long Reynolds aluminum-tubing boom (two 8-foot [2.44-m] lengths joined) provides 0.12-wavelength spacing between the driven element and reflector. The weight of the array is approximately 25 pounds, making it practical to use a TV-type antenna rotator with the system.

Design Considerations

Some may argue that helically wound antennas will tend to radiate off the ends of the elements (axial mode), and that the condition will prevent the Yagi from exhibiting normal front-to-side and front-to-back characteristics. It was established by Kraus,[1] that axial-mode radiation occurs only when a helix is one wavelength or greater in circumference. Generally, when the circumference is less than two-thirds wavelength, a nearly sinusoidal type of current distribution exists. This is effected by alternate reinforcement and cancellation of the two traveling waves which are directed oppositely. Each is of nearly identical current amplitude (I_O), hence the T_O transmission mode results for both traveling waves on the helix. Because of the foregoing condition the helices of this antenna radiate in the *normal mode* and will exhibit linear polarization characteristics. The diameter of these helices is merely 0.0009469 wavelength, eliminating any opportunity for axial-mode radiation. Front-to-side and front-to-back checks of the beam indicate, indeed, that normal-mode radiation is occurring. A number of relative tests were performed locally at distances up to 20 miles, and the front-to-back characteristics indicate a 10-dB figure. Front-to-side checks confirm a figure of approximately 28 dB. The effect is not so pronounced when receiving high-angle incoming signals. Depending upon the time of day, distance involved, and other propagation factors, various directivity traits will be observed.

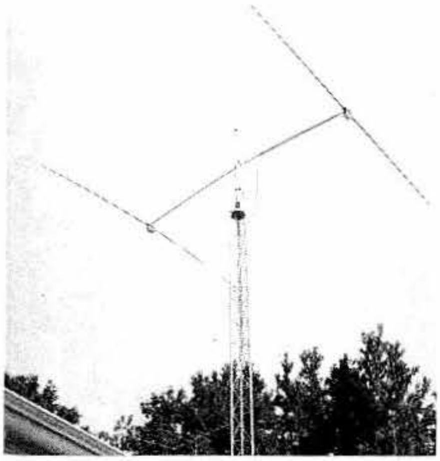

The HW-40 beam, just 28 percent of full-size, is light enough to require only a TV-rotator.

Helically wound elements were chosen in preference to lumped-inductance elements in order to obtain linear voltage and current distribution across the elements, and to make the matching technique less difficult. Even if end-loaded elements with capacitance hats were used, the 18-foot (5.49-m) dimension would make it impossible to use. a gamma match of conventional design. Because of the very low value of radiation resistance of short loaded antennas, other conventional matching schemes would not lend ease to matching a 50-ohm coaxial feeder to the system.

There are certain penalties one must accept when using physically shortened antennas of this kind, and an in-depth treatment of the subject was given by Sevick in *QST*.[2] It was a fortunate circumstance in this case that the feed-point impedance of the beam turned out to be approximately 12 ohms. Therefore, a 4:1 toroidal balun was installed at the feed point to provide a near-perfect match to the 50-ohm transmission line. In addition to encountering low values of feed impedance, the bandwidth of loaded elements is greatly less than with full-size antennas. This beam was tuned for operation at a center frequency of 7050 kHz. It works well from 7025 to 7075 kHz, with an SWR no greater than 2:1. It is flat at 7050 kHz. Those wishing to duplicate this Yagi should be willing to accept this design trade-off in the interest of having an effective directional antenna. It is recommended, therefore, that it be tuned for one's favorite part of the 40-meter band. The use of a Transmatch will, however, permit excursions into portions of the band where the SWR becomes greater than 1:1.

Construction Details

The construction of the 40-meter beam is very simple and requires no special tools or hardware. Two fiberglass 15-meter quad arm spreaders are mounted on an aluminum plate with U bolts, as shown in the photograph. A wooden dowel is inserted approximately 6 inches (152 mm) in the end of each fiberglass arm to prevent the U bolts from crushing the poles. The aluminum mounting plate is equipped with U-bolt hardware for attachment to the 1-1/4-inch (32-mm) diameter boom.

A plastic refrigerator box is mounted on each element support plate and is used to house a Miniductor coil. No. 14 copper wire is used for the elements. The wire is wound directly on the fiberglass poles at a density of *40 turns per foot* (not turns per inch!) for a total of 360 evenly spaced turns. The wire is attached at each end with an automotive hose clamp of the proper size to fit the fiberglass rods. Since the fiberglass is tapered, care must be taken to keep the turns from sliding in the direction of the tips. Several pieces of plastic electrical tape were wrapped around the pole and wire at intervals of 1 foot (0.31 m). All of the element half sections are identical in terms of wire and pitch. Coil dimensions and type are given in Fig. 3.

The driven-element matching system consists of a 4:1 balun transformer and a link tightly coupled to the Miniductor. Complete details are given in Fig. 4.

Mounted at the end of each element (held in place by the hose clamp) is a short section of stiff copper wire used to allow for final tuning of the system as well as to broaden the frequency response of the antenna. Since the overall antenna is very small in relation to a full-sized array, the SWR points of 2:1 are rather close to each other, as mentioned earlier. The antenna shown in the photograph provides an SWR of less than 2:1 within about 25 kHz either side of resonance. Tuning the antenna for phone-band operation should not be difficult and the procedure outlined below should be suitable.

Tuning

The parasitic element is adjusted to be about four percent lower in frequency than

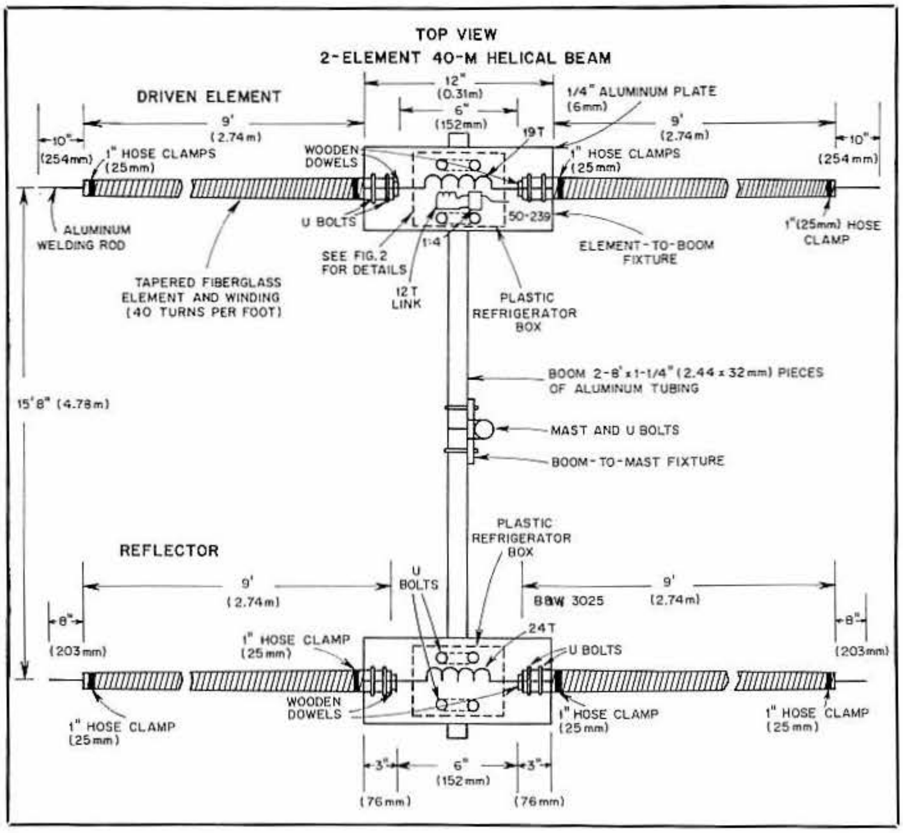
TOP VIEW
2-ELEMENT 40-M HELICAL BEAM
DRIVEN ELEMENT
12" (0.31m)
6" (152mm)
1/4" ALUMINUM PLATE (6mm)
9' (2.74m)
19T
10"
(254mm)
1" HOSE CLAMPS (25mm)
WOODEN DOWELS
U BOLTS
50-239
ALUMINUM WELDING ROD
1:4
SEE FIG. 2 FOR DETAILS
ELEMENT-TO-BOOM FIXTURE
1"(25mm) HOSE CLAMP
TAPERED FIBERGLASS ELEMENT AND WINDING (40 TURNS PER FOOT)
12 T LINK
PLASTIC REFRIGERATOR BOX
BOOM 2-8'x1-1/4" (2.44 x 32mm) PIECES OF ALUMINUM TUBING
15'8" (4.78m)
MAST AND U BOLTS
BOOM-TO-MAST FIXTURE
REFLECTOR
U BOLTS
B&W 3025
24T
8"
(203mm)
1" HOSE CLAMP (25mm)
WOODEN DOWELS
3"
(76mm)

Fig. 3 — Overall dimensions for the 40-meter short beam. The boom consists of two pieces of standard 1-1/4-inch (32-mm) dia do-it-yourself aluminum tubing.

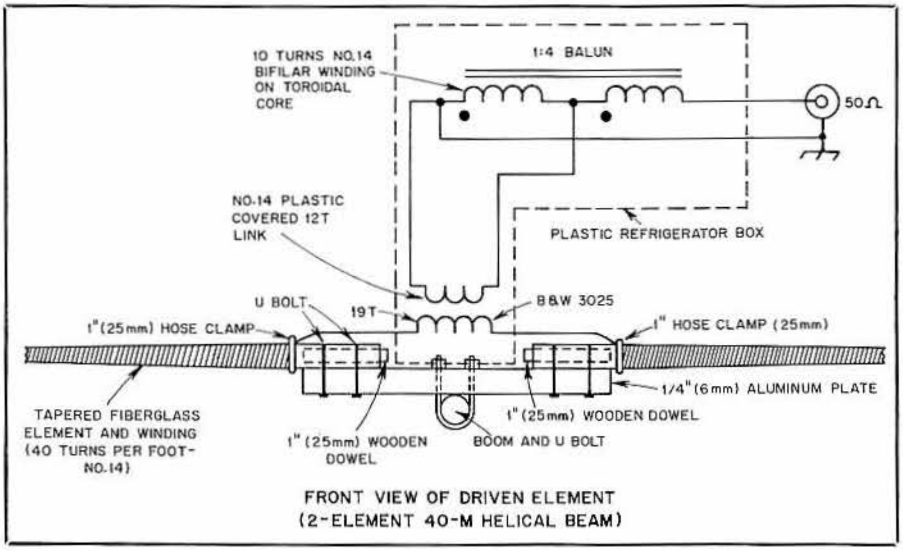
10 TURNS NO.14 BIFILAR WINDING ON TOROIDAL CORE
1:4 BALUN
50Ω
NO.14 PLASTIC COVERED 12T LINK
PLASTIC REFRIGERATOR BOX
U BOLT
19T
B&W 3025
1"(25mm) HOSE CLAMP
1" HOSE CLAMP (25mm)
1/4"(6mm) ALUMINUM PLATE
TAPERED FIBERGLASS ELEMENT AND WINDING (40 TURNS PER FOOT-NO.14)
1" (25mm) WOODEN DOWEL
1" (25mm) WOODEN DOWEL
BOOM AND U BOLT
FRONT VIEW OF DRIVEN ELEMENT
(2-ELEMENT 40-M HELICAL BEAM)

Fig. 4 — Schematic diagram of the balun assembly mounted inside the plastic utility box. The core is a single T-200-2 Amidon. The 12-turn link is wound directly over the 19-turn Miniductor.

the driven element. A grid-dip oscillator was coupled to the center loading coil and the element tips were trimmed (a quarter of an inch [6 mm] at a time!) until the GDO indicated resonance to be at 6.678 MHz. For phone-band use, the ends could be snipped for 6.840 MHz. Adjusting the driven element is simple. Place an SWR meter or power meter at the input connector and cut the end wires (or add some if necessary) to obtain a proper match between line and antenna.

Performance Characteristics

Once the initial near-ground tuning of the elements was completed, the HW-40 was carried aloft and mounted on the 40-foot (12.19-m) tower at W1FB. Final touching up of the matching and element resonances was effected after the beam was in place on the tower. It is important to realize that shortened antennas are sensitive to changes in environment because of their narrow-band characteristics. Thus, height above ground and distance from nearby objects will cause some change in resonance of the elements. The resonant frequency of the HW-40 changed some 150 kHz between tuning at several feet above ground to final tweaking atop the tower.

It should be noted that the 40-meter beam has yet to be tested at an ideal height above ground. A distance of 40 feet (12.19 m) above the earth is not much more than one-quarter wavelength at 7 MHz, though performance thus far has been better than one might expect when using a beam or dipole less than one-half wavelength above ground.

Comparisons were made between the beam and three types of wire antennas . . . an inverted V, a 3/4-wavelength end-fed wire, and a full-wave 40-meter loop (with its theoretical 2-dB gain).[3] The results have been interesting, though at times rather confusing. The confusion results from the time-of-day/propagation syndrome mentioned earlier. At no time, however, did the inverted-V or 3/4-wavelength antennas surpass the performance of the HW-40. The beam was always as good as or better than the other two radiators, regardless of band conditions or time. During contacts with European or South American DX stations, the beam was always the best antenna, usually by two S units or more. During daytime contacts out to, say, 1000 miles, the wire antennas equalled the beam at times.

When comparing the beam to the full-wave loop, results became less easy to evaluate. That is, the antennas took turns being best. Frequently, the loop outperformed the beam (in the favored direction of the loop) both on short-haul and DX contacts. At other times the beam had the edge by a couple of S units. At this time it is impossible to say which antenna would provide the best all-year, all-around service. But, it has been observed that the beam is far less subject to pickup of man-made and atmospheric noise because of its narrow-band characteristics (the loop exhibits the opposite bandwidth trait— broad). Reduction in noise during reception can be a blessing, especially when working with the weaker signals! Another beneficial effect of the narrow-band feature is that strong, near-frequency signals do not so severely affect the receiver with respect to overloading and cross modulation, as is the situation with other types of antennas. The proof of this came when trying to operate while W1AW (two blocks away from W1FB) was transmitting code practice. Until the beam was erected it was impossible to use the 40-meter band, but with the beam connected to the receiver, and with the antenna pointed away from W1AW, it became possible to operate within 10 kHz of W1AW without knowing that station was on the air! So, if some kW operator lives down the street from you, this little beam could put you back in business.

While running approximately 80 watts output, the short beam has been extremely effective in working DX. DX stations now answer our CQs, whereas with the wire antennas (loop excepted) this seldom happened. Signal reports from DX stations average 579, with many 589 and 599 reports. During QRP operation with the beam (2 watts output), a number of DX stations have been raised and worked. The signal reports are lower of course, but RST 559 and 569 indicates that something good must be happening! One unrelated event occurred while testing the beam and loop with the QRP setup: a station in the 4th call area called CQ and was answered. He came back and said, "No QRP lids, CQ kWs only." It's unlikely that the HW-40 beam can be blamed for that, especially since we went on to work five European stations later that day with the 2-watt equipment!

Acknowledgements

The authors wish to express gratitude to those who helped construct, test and erect the 40-meter beam. Special thanks to K1JX, who "clomped" up and down the ARRL test tower and the W1FB station tower numerous times with test gear (and the beam) until the final design goals were met. Thanks also to W1NPG (SK) who labored silently while winding the helices, and subsequently endured the unbearable Connecticut heat-wave temperatures of August, while putting two coats of exterior spar varnish on the completed elements. The element dimensions and turns information were developed by K1TD to test a newly developed calculator program. This material was originally presented in *QST* by Robert Myers, W1XT and Doug DeMaw, W1FB.

References

[1] Kraus, *Antennas*, First Edition, McGraw-Hill Book Co., Chapter 7.

[2] Sevick, "The Ground-Mounted Short Vertical," *QST*, March 1973, p. 13. (See Chapter 1.)

[3] DeMaw, "The Novi-Loop," *QST*, October 1973. (See Chapter 4.)

A Bite-Size Beam for 20 and 15 Meters

Do you want to work DX from your space-restricted QTH with increased effectiveness? If the answer is yes, perhaps this two-band Yagi system is just the one for you.

There is no doubt that a hundred-foot (30.48-m) tower with stacked monobanders for 20 and 15 meters is one way to generate an effective signal in the pileups. But what are the characteristics of this big system which make it perform better than a dipole strung between two trees? First, the tall tower places the antenna up in the clear, away from most of the surrounding ground clutter. Even a relatively short tower (40-foot [12.2-m] variety) provides this advantage in many instances. The ability to point the major radiation lobe of the antenna in the correct direction is another reason the high monobanders outperform the kinky wire. But if the antenna is capable of supporting its own elements, the directional characteristics can be used to good advantage whether the beam is mounted at 100 feet (30.5 m), or at 40 feet (12.2 m). And without any question, the ability to reject interference aids substantially in the reception of weak signals.

A typical one-foot (0.31-m)-per-side triangle tower is self-supporting (when properly installed) to 40 feet in height and requires less than one square foot of property. For many amateurs, the major limiting factor is the inability to turn a monoband beam (or even a large tribander) without hitting some object like a tall tree. Hesitation on behalf of the amateur to place what might be considered by his neighbors an offensive amount of hardware in the sky near his home is another reason beams are ruled out. The 2-element-per-band, 20- and 15-meter, interlaced Yagi presented here, overcomes these problems. It is lightweight and is less than 16-feet (4.88-m) square. And it's probably no more obtrusive than a large TV antenna.

So What if It's Short!

A popular misconception among amateurs is that any element short of full size is no good in an antenna system. Reducing the size of an antenna by 50 percent does lower the efficiency by a decibel or two, but the gain capability of a parasitic array outweighs this small loss in efficiency. Mounting the antenna above the interference-generating neighborhood can greatly reduce susceptibility to man-made noise and certainly aids in the reduction of rf heating to trees, telephone poles, and buildings. It has been said that a brick wall, or dense foliage, can attenuate a signal passed through it by as much as 20 dB. Placing the antenna above these energy-absorbing objects is very desirable.

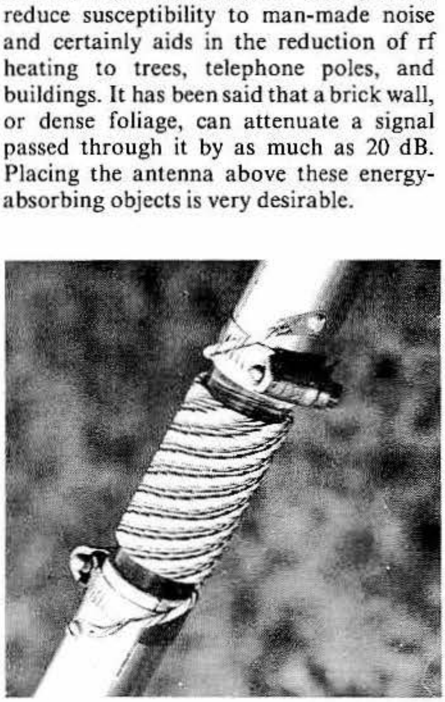

The coils are wound on Plexiglas rod. Compression clamps are used to hold the Plexiglas in position. Sheet-metal screws and solder lugs provide attachment points for the ends of the loading coils.

View of the short beam from the front yard at W1FB.

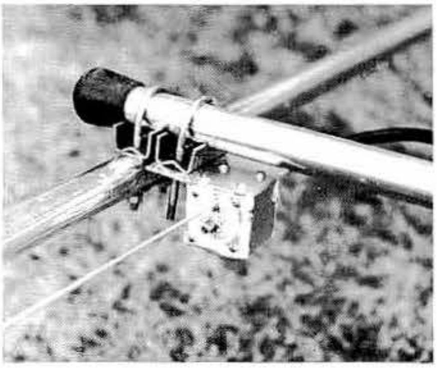

The gamma assembly is held in place by means of a small U bolt. The capacitors are mounted on etched circuit board.

It is a generally accepted fact that traps are difficult to build and adjust, as well as their being lossy. Loading coils, on the other hand, are easy to wind and require essentially no adjustment. If the wire size is large, losses are not a major factor. W2FMI lauded the merits of short antennas in an earlier article explaining that in reality, short (50 percent) elements do not materially depreciate the gain.[1] With these ideas in mind, the authors designed and built a shortened-element, but wide spaced, director array for 20 meters interlaced with a similar 15-meter reflector configuration also having optimum spacing for maximum gain. By placing the 15-meter driven element in front of the 20-meter array, the longer 20-meter elements tend to act as reflectors on 15 meters. There are no unwanted reflectors "in front" nor any directors "behind" the active array — elements which could cause pattern distortion and poor front-to-back ratio.

Construction

The dual-band beam has four elements, the longest of which is 16 feet (4.88 m). All of the elements and the boom are made from 1-1/4-inch (32-mm) diameter aluminum tubing available at most hardware stores. Element sections and boom pieces are joined together by slotting a 10-inch (254-mm) length of 1-1/4-inch (32-mm) tubing with a nibbling tool and compressing it for a snug fit inside the element and boom tubing. Coupling details are shown in Fig. 6.

The loading coils are wound on 1-1/8-inch (29-mm) diameter Plexiglas rod. The rod slips into the element tubing and is held in place with compression clamps. Be sure to slit the end of the aluminum where the compression clamps are placed. The model shown in the photographs has coils made of surplus Teflon-insulated miniature audio coaxial cable with the shield braid and inner conductor shorted together. A suitable substitute would be no. 14 enameled copper wire wound to the same dimensions as those given in Fig. 5.

All of the elements are secured to the boom with common TV U-bolt hardware. Plated bolts are desirable to prevent rust from forming. A 1/4-inch (6-mm) thick boom-to-mast plate is constructed from a few pieces of sheet aluminum cut into 10-inch (254-mm) square sheets and held together with no. 8 hardware. Several cookie tins could be used if sheet aluminum is not available. One local amateur used a plate from a large electrical box as a boom-to-mast bracket. Since it is galvanized, it is quite resistant to the harsh New England winters.

A boom strut (sometimes called a truss) is recommended because the weight of the elements is sufficient to cause the boom to sag a bit. A 1/8-inch (3-mm) diameter nylon line is plenty strong. A U-bolt clamp is

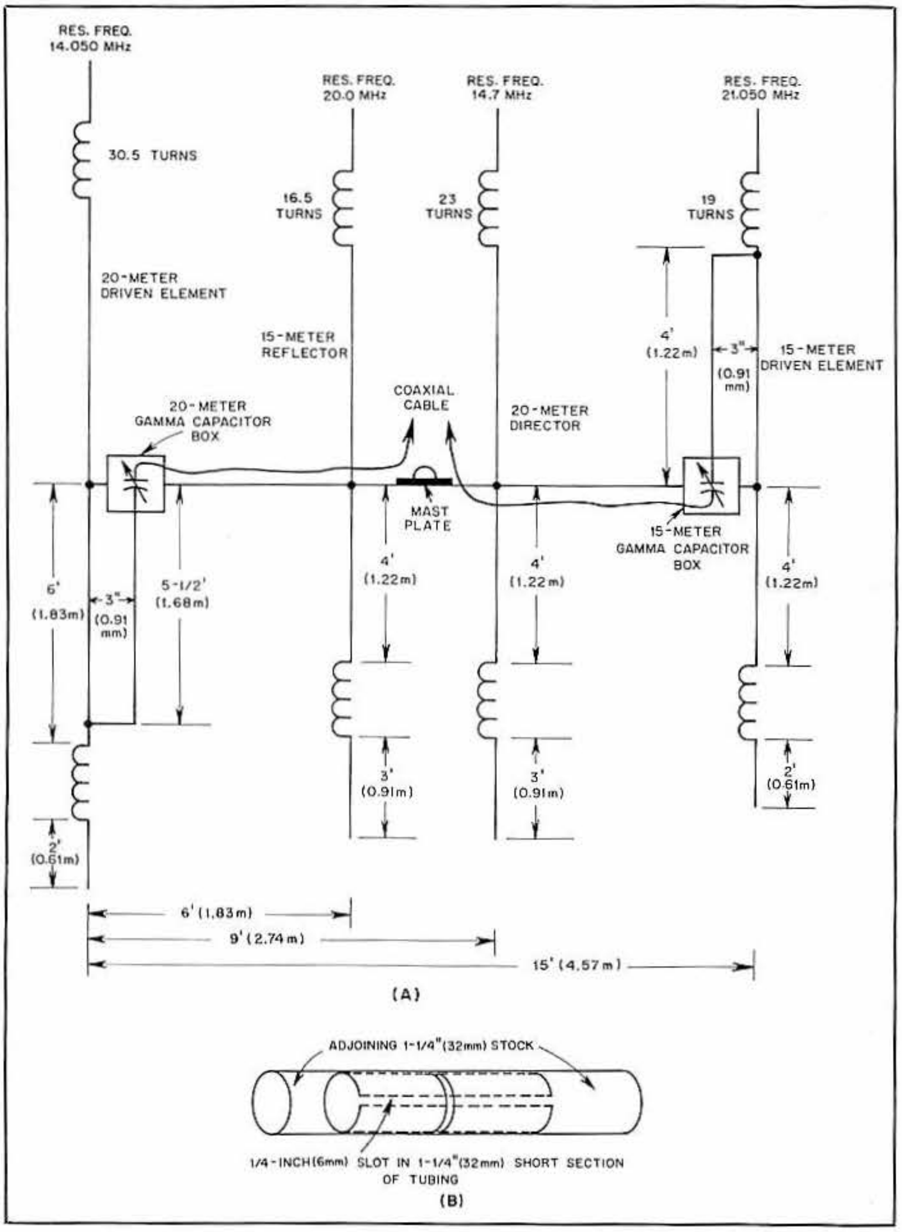
RES. FREQ. 14.050 MHz
RES. FREQ. 20.0 MHz
RES. FREQ. 14.7 MHz
RES. FREQ. 21.050 MHz
30.5 TURNS
16.5 TURNS
23 TURNS
19 TURNS
20-METER DRIVEN ELEMENT
15-METER REFLECTOR
4' (1.22m)
3" (0.91 mm)
15-METER DRIVEN ELEMENT
COAXIAL CABLE
20-METER GAMMA CAPACITOR BOX
20-METER DIRECTOR
MAST PLATE
15-METER GAMMA CAPACITOR BOX
6' (1.83m)
3" (0.91 mm)
5-1/2' (1.68m)
4' (1.22m)
4' (1.22m)
4' (1.22m)
3' (0.91m)
3' (0.91m)
2' (0.61m)
2' (0.61m)
6' (1.83m)
9' (2.74m)
15' (4.57m)
(A)
ADJOINING 1-1/4"(32mm) STOCK
1/4-INCH(6mm) SLOT IN 1-1/4"(32mm) SHORT SECTION OF TUBING
(B)

Fig. 5 — Constructional details for the 20- and 15-meter beam. The coils for each side of the element are identical. The gamma capacitors are each 140-pF variable units manufactured by E. F. Johnson Co. The capacitors are insulated from ground within the container. Since the design is one-half size for each band, the tuning is somewhat critical. The builder is encouraged to follow carefully the dimensions given in the text.

Table 2

Complete Parts List for the Short Beam

Qty	*Material*
9	8-foot (2.44-m) lengths of aluminum tubing, 1-1/4-in. (32 mm) dia
11	U bolts
2	Variable capacitors, 140 pF (E. F. Johnson)
4 ft	Plexiglas cast rod, 1-1/8-in. (29 mm) dia
16	Stainless steel hose clamps, 1-1/2-in. (38 mm) dia
1	Aluminum plate, 8-in. (203 mm) square
10 ft	Aluminum solid rod, 1/4-in. (6 mm) dia
2	Refrigerator boxes, 4 X 4 X 4 in. (102 X 102 X 102 mm)
25 ft	Nylon rope, 1/8-in. (3 mm) dia
16	No. 8 sheet metal screws
16	No. 8 solder lugs
8	Plastic (or rubber) end caps, 1-1/4-in. (32 mm) dia

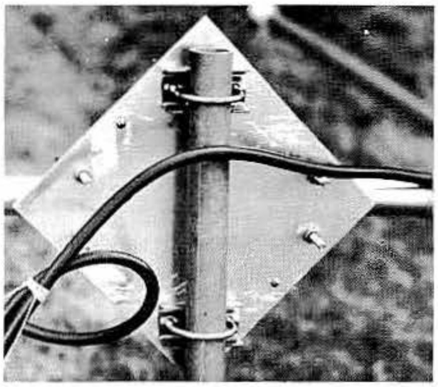

The boom-to-mast plate.

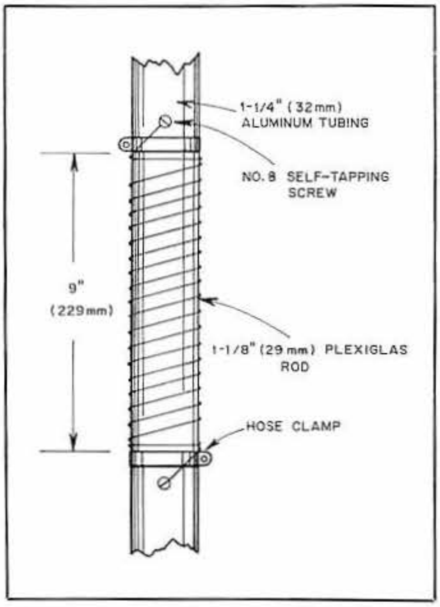

Fig. 6 — Coupling details for joining sections of aluminum tubing.

placed on the mast several feet above the antenna and provides the attachment point for the center of the truss line. To reduce the possibility of water accumulating in the element tubing and subsequently freezing (rupture may be the end result), crutch caps are placed over the element ends. Rubber feet suitable for keeping furniture from scratching hardwood floors would serve the same purpose. In fact, the rubber tips prevent the element ends from damaging surrounding objects during installation.

A heavy-duty steel mast should be used, such as a 1-inch (25-mm)-diameter galvanized water pipe. Steel TV mast is also acceptable. Any conventional TV-type antenna rotator should hold up under load conditions presented by this antenna. Nevertheless, certain precautions should be taken to assure continued trouble-free service. For instance, whenever possible, mount the rotator inside the tower and extend the mast through the tower top sleeve. This procedure relieves the rotator from having to handle lateral pressures during windy weather conditions. A thrust bearing is desirable to reduce downward forces on the rotator bearings.

Hookup and Operation

The monoband nature of the beam requires the use of two coaxial feed lines. The coaxial cable is attached to the 15-meter element (at the front of the beam) at the gamma-capacitor box. The other end of the cable is connected to a surplus 28-V dc single-pole coaxial switch.[2] The cable for the 20-meter element is connected in a similar fashion. The switch allows the use of a single feed line from the shack to a point just below the antenna where the switch is mounted. It is a simple matter to provide voltage to the switch for operation on one of the two bands. At the price of coaxial cable today, a double run of feed line represents a substantial investment and should be avoided if possible.

An etched-circuit board was mounted inside an aluminum Minibox to provide support and insulation for each of the gamma tuning capacitors. Plastic refrigerator boxes available from most department stores would serve just as well. The capacitor housing is mounted to the boom by means of U bolts.

The builder is encouraged to follow the dimensions given in Fig. 5 as a starting point for the position of the gamma rods and shorting bar. Placing the antenna near the top of the tower and then tilting it to allow the capacitors to be reached makes it possible to adjust the capacitors for minimum SWR as indicated by an SWR meter (or power meter) connected in the feed line at the relay. If the SWR cannot be reduced below some nominal figure of approximately 1.4:1, a slight repositioning of the gamma short might be required. The dimensions given are for operation at 14.050 MHz and 21.050 MHz. The SWR climbs above 2:1 about 50 kHz in either direction from the center frequency. Although tests were not conducted at more than 150 watts input to the transmitter there is no reason why the system would not operate correctly with a kilowatt of power supplied to it.

After many months of testing at the QTH of W1FB, several characteristics were noted. First, the antenna withstood several wind and ice storms common to Connecticut. Performance turned out to be what can be expected from a 2-element Yagi. Since the 20-meter portion of the antenna is a director array, the front-to-back ratio is a bit less than 10 dB. On 15 meters, where the system operates with a reflector parasitic element, the front-to-back ratio is considerably better — on the order of 15 dB.

This antenna was mounted atop a 40-foot (12.2-m) tower where previously there was connected a 20-meter sloping dipole pointed at Europe (a multielement array for 2 meters graced the top of the tower!). The number of European stations contacted increased rather dramatically with the installation of the beam. Except in pileups on rare DX stations, DX is now worked as a routine, even though W1FB is a QRP fan and typically runs less than 2 watts of output power.

No doubt the increased effectiveness of the signal was caused by several factors. The antenna certainly has gain (although at ARRL hq. we have no way of making accurate measurements) and is well above the nearby small trees. More importantly, the antenna is now far from the aluminum house siding which once was directly in the pattern of the 20-meter "sloper." Of course, the front-to-back characteristic has the tendency of making DX signals louder in relation to stateside stations and therefore makes receiving much easier.

If previously you've felt that a beam and tower were not possible at your QTH, perhaps re-evaluation of the situation is in order. This antenna, as shown in the photograph, does not appear offensive when viewed from the front yard. Yet the overall amount of property occupied (at ground level) is less than one square foot. And the turning radius is slightly over 11 feet (3.35 m)! This material was originally presented in *QST* by Robert Myers, W1XT and Clarke Greene, K1JX.

References

[1] Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, March 1973. (See Chapter 1.)

[2] Purchased from Fair Radio Sales, Lima, OH.

Linear Loaded 20-Meter Beam

Build this novel minibeam for 20 meters . . . it works!

Full-sized antennas are fine if you have the height and room for them. But much of the advantage is usually lost with average installations. So why bother with a sad-looking monster a few feet atop the house when a trim, smaller model will do the job. This article deals with a method that eliminates the messy and expensive loading coils as well.

Having built an HW-32A transceiver, the author's 20-meter activity was restricted to the phone band. A homemade vertical antenna did a fair job in the states, but not for DX. This writer's thoughts ran to a beam and the aluminum tubing that could be obtained for full-size construction. A 2-element beam was constructed, using a gamma match and information found in the *Handbook*. The beam worked electrically, but the mechanical construction left much to be desired. It flopped in the wind like a wounded butterfly. The author tried to correct this with a wooden framework, but it was just as bad. Something with shorter elements was desired, something that was rigid and would still perform well. But all of the author's research turned up only loading coils or other methods involving a loss compromise.

The author's aluminum supply included four 12-ft (3.66-m) lengths of 7/8-inch (22-mm) diameter tubing. This would give two elements of 24 ft (7.32-m) length and a minimum of sag. The problem was how to mount them and make up the difference in physical and electrical length for a half-wave length. Loading coils were ruled out. The possibility of folding wire under the elements seemed to have merit. (Later, this was found to be called linear loading.) This dictated a split element and standoff supports for the tubing.

Two 2-ft (0.61-m) lengths of aluminum channel were obtained for the element supports (Fig. 7). each 12-ft (3.66-m) section was attached using 1-in. (25-mm) ceramic standoff insulators. A 34-in. (19-mm) birch dowel was used to stiffen and prevent crushing of the tubing where the attachment was made. Element supports were grooved and attached to a 10-ft (3-m) TV mast section using a single 5/16-in. (8-mm) U bolt.

The folded additional loading wire was no. 12 copper house wire with the insulation stripped off, and was supported by two bakelite blocks on each element section. The blocks were anchored to the tubing by small sheet metal screws. Ends of the wire sections were formed into a loop to hold their position.

Tune-Up and Adjustments

My thinking in regard to coaxial-cable-fed dipoles is that they should use a balun or matching device of some kind. This beam, having a split driven element, seemed to call for a 1:1 balun since it was to be fed with RG-58/U (Fig. 8). Formulas found in handbooks for beam-element lengths did not seem to work for the linear-loading design, so cut-and-try was used many times. In final tests, a grid-dip meter was utilized with a calibrated receiver to determine the driven-element total length of 35 ft (10.67 m). A reflector with 10-ft (3.05-m) spacing was used.

It was found that adding a reflector of five-percent greater length lowered the frequency (with the beam near the ground) but on raising the beam to rooftop, the frequency had increased. The 10-ft spacing gave a 3:1 VSWR which was very difficult to reduce. Testing this arrangement on the air seemed to indicate more gain in the back direction than in the front. After many tests and much rebuilding of the wire lengths, the director element approach seemed to give the most encouraging results. The VSWR was down to 1.5:1 and the five-percent shorter director raised the frequency to the phone band. Spacing was reduced to 7 ft 6 in. (2.29 m).

More research brought out the possibility of inductive matching to get the VSWR

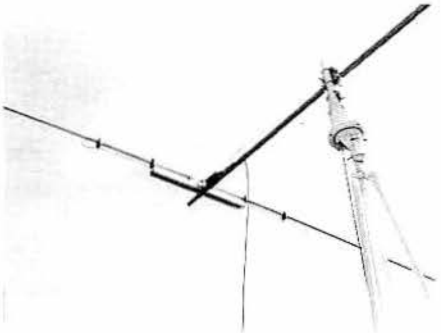

The beam atop the author's tower.

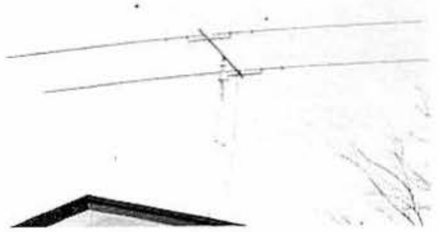

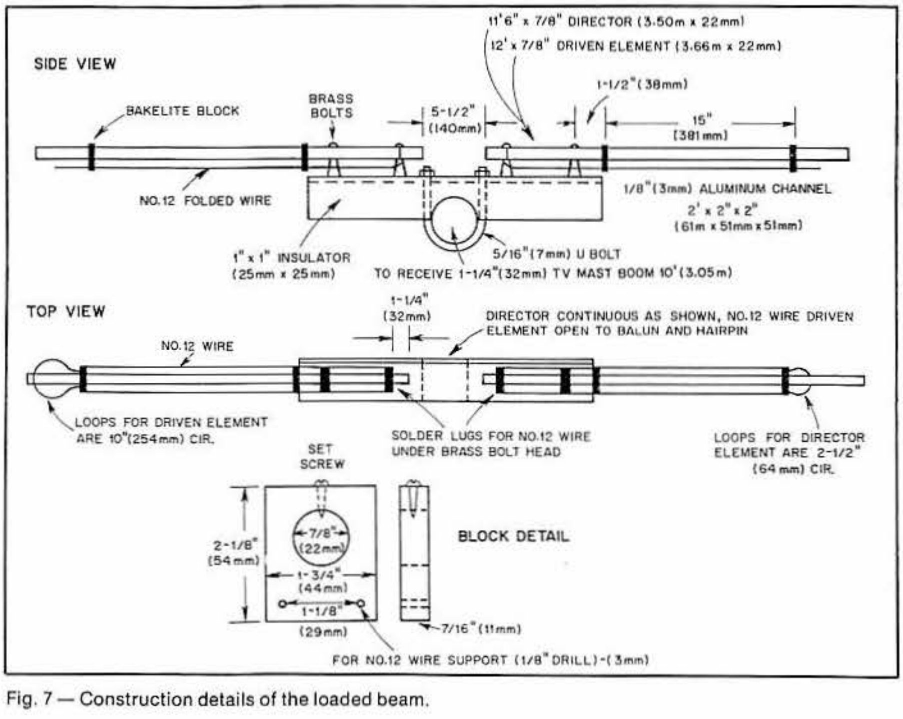
11'6" x 7/8" DIRECTOR (3.50m x 22mm)
12' x 7/8" DRIVEN ELEMENT (3.66m x 22mm)
SIDE VIEW
1-1/2" (38mm)
BAKELITE BLOCK
BRASS BOLTS
5-1/2" (140mm)
15" (381mm)
NO.12 FOLDED WIRE
1/8" (3mm) ALUMINUM CHANNEL 2' x 2" x 2" (61m x 51mm x 51mm)
1" x 1" INSULATOR (25mm x 25mm)
5/16" (7mm) U BOLT
TO RECEIVE 1-1/4" (32mm) TV MAST BOOM 10' (3.05m)
TOP VIEW
1-1/4" (32mm)
DIRECTOR CONTINUOUS AS SHOWN, NO.12 WIRE DRIVEN ELEMENT OPEN TO BALUN AND HAIRPIN
NO.12 WIRE
LOOPS FOR DRIVEN ELEMENT ARE 10" (254mm) CIR.
SOLDER LUGS FOR NO.12 WIRE UNDER BRASS BOLT HEAD
LOOPS FOR DIRECTOR ELEMENT ARE 2-1/2" (64 mm) CIR.
SET SCREW
2-1/8" (54mm)
7/8" (22mm)
1-3/4" (44mm)
1-1/8" (29mm)
BLOCK DETAIL
7/16" (11mm)
FOR NO.12 WIRE SUPPORT (1/8" DRILL)-(3mm)

Fig. 7 — Construction details of the loaded beam.

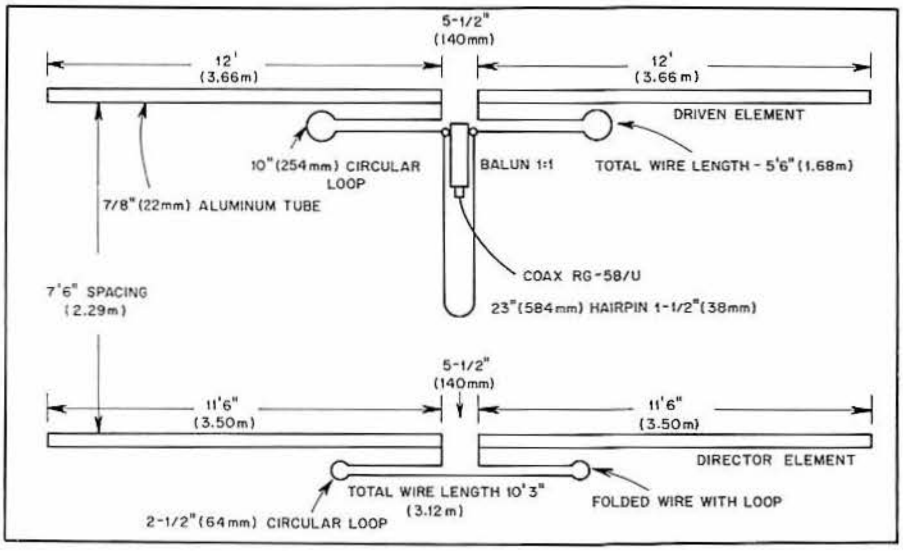

Fig. 8 — Critical electrical dimensions.

down to a more reasonable level. Many hairpin lengths and widths were tried with varying results. The final hairpin was 23-in. long by 1-1/2 in. (580 × 38 mm) wide with a slight fan shape to fit the balun connections. This reduced the VSWR at 14,200 kHz and 14,350 kHz with a flat response in between.

On-the-Air Tests

The help of other amateurs in tests for front-to back ratio was very helpful. Average reports have been 2.5 S units in the forward position over the back position. The author did not have a reference dipole for gain tests, but has heard and worked more DX than in the past 20 years. The beam is mounted on a rooftop 10-ft (3-m) steel tower 38 ft (11.6-m) above the ground. It weighs around 20 pounds and is rotated by an older-type TV rotator. This material was originally presented in *QST* by Cole Collinge, W0YNF.

The Log-Yag Array

The Yagi antenna array has been around for years and years. A relative newcomer to hams is the log-periodic dipole array (LPDA), which offers nearly constant gain over a greater bandwidth than the Yagi. Guess what happens when you cross a Yagi with an LPDA . . . .

With the decline in sunspot activity, a number of amateurs have considered monoband Yagi arrays. The first problem encountered seems to be array length, that is, overall size for a desired gain and bandwidth. The Log-Yag principle, as will be discussed shortly, has produced a system which will provide the amateur with another alternative to the long-boom Yagi, stacked Yagis, or loop-antenna systems. The L-P Yagi (Log-Yag) array is not a new system; many such arrays have been designed and developed by Oliver Swan[1] and others.[2] This article, however, will provide the basic theory of operation, design procedure, and the construction of a practical antenna.

Theory of Operation

The Log-Yag array utilizes an LPDA-driven[3] group of elements, designed to cover a desired bandwidth, in conjunction with parasitic elements to achieve higher gains and greater directivity than would be realized with either the LPDA or Yagi array alone. The Yagi array requires a long boom and wide element spacing for wide bandwidth and high gain. This is because the Q of the Yagi system increases as the number of elements is increased and/or as the spacing between adjacent elements is decreased.[4,5] An increase in the Q of the Yagi array means that the total bandwidth of that array is decreased; optimum gain, front-to-back ratio, and side lobe rejection are obtainable only over small portions of the band. Dr. I. L. Morris, using a high-speed digital computer, has completed extensive research on 4-, 8-, and 10-director Yagi-Uda arrays.[6] His work is comprehensive and is recommended reading for all technically minded amateurs. The parameters varied in his study were element length, spacing, radius and number. As can be seen in Fig. 9, the forward gain and front-to-back ratio

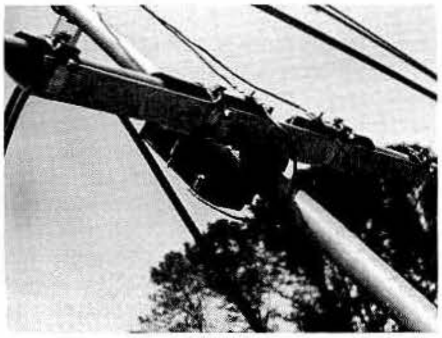

The attachment of the elements to the boom.

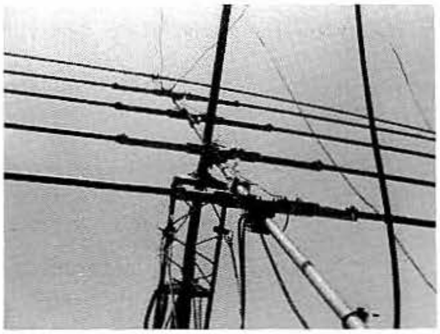

From the front to the back of the Log-Yag array. Note the truss provides lateral and vertical support.

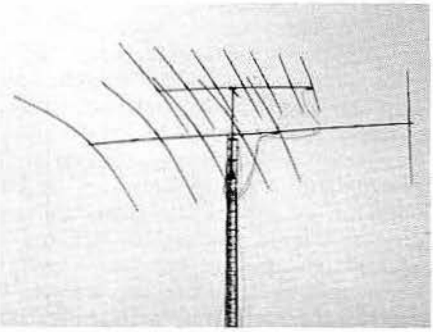

The completed Log-Yag array ready for use. The array mounted above the Log-Yag is a 7-element LPDA for 21 to 30 MHz.

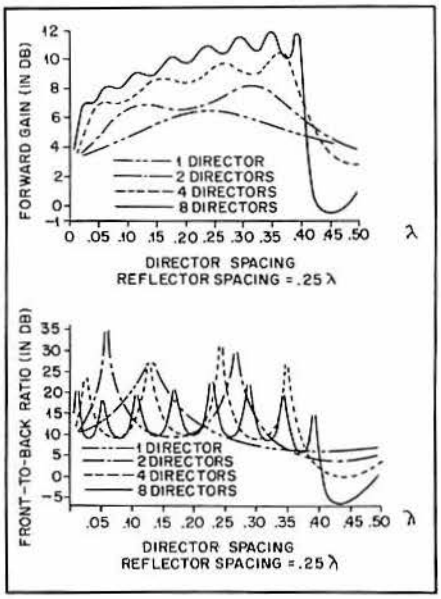

Fig. 9 — The effects of director spacing on various Yagi arrays. The change in gain and front-to-back ratio is plotted for the change in director spacing. The reflector spacing has been held constant.

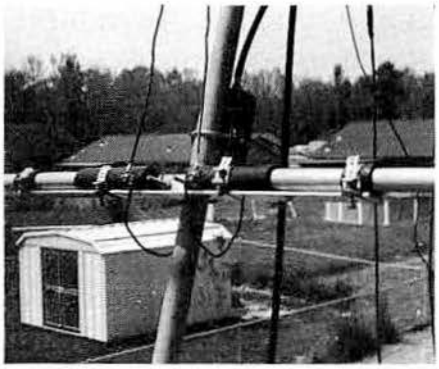

The connections between the balun and the input terminals.

deteriorate sharply as element spacings decrease. If the elements are closely spaced, then as the frequency is shifted either side of the array design frequency the electrical spacing between adjacent elements changes rapidly. This causes a higher SWR and a deterioration of forward gain and front-to-back ratio.

The Log-Yag system overcomes this difficulty by using a multiple driven element "cell" designed in accordance with the principles of the log-periodic dipole array.[7,8] Since this log cell exhibits both gain and directivity by itself, it is a more effective radiator than a simple dipole driven element. The front-to-back ratio and gain of the log cell can be improved with the addition of a parasitic reflector and director. It is not necessary for the parasitic element spacings to be large with respect to wavelength, as in the Yagi array, since the log cell is the determining factor in the array bandwidth. In fact, the element spacings within the log cell may be small with respect to a wavelength without appreciable deterioration of the cell gain. For example, decreasing the relative spacing constant (σ) from 0.1 to 0.5 λ will decrease the gain by less than 1 dB. Hence, a further reduction in boom length. It can be seen that the Log-Yag array will exhibit high theoretical gain (11 dBd), high front-to-back ratio (30 dB), high cross polarization (front-to-side ratio — 45 dB), and a wideband response utilizing boom lengths approximately one half that of a Yagi with similar characteristics.[2]

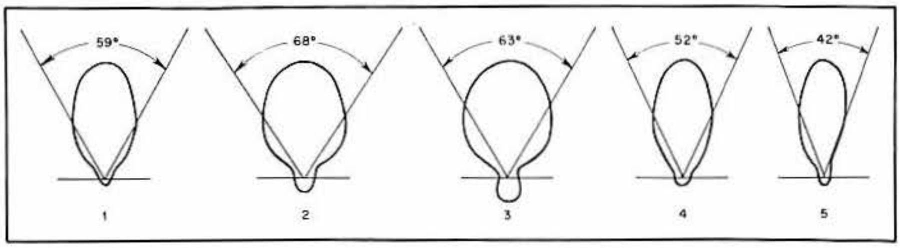

Fig. 10 — Beam patterns of 20-meter arrays.
No. 1 — 3-el. log cell, dir. @ 0.1λ, ref. @ 0.2λ.
No. 2 — 5-el. log periodic, σ = 0.1.
No. 3 — 3-el. log cell, 1st dir. @ 0.1λ, 2nd dir. @ 0.2λ.
No. 4 — 4-el. log cell, ref. @ 0.15λ.
No. 5 — 4-el. log cell, dir. @ 0.15λ, ref. @ 0.085λ (described in this article).

The author has built many monoband 14-MHz Log-Yag arrays in an attempt to find an optimum combination of elements, while holding the boom length to that of a full-sized, three-element monobander Yagi. Relative radiation patterns for various element combinations are found in Fig. 10. The final array design takes the form of a four-element log cell, parasitic reflector spaced at 0.085 λ_{max} and parasitic director spaced at 0.15 λ_{max} where λ_{max} is the longest free-space wavelength within the array passband. It has been found that array gain is almost unaffected with reflector spacings from 0.08 λ to 0.25 λ and the increase in boom length is not justified.[9] The function of the reflector is to improve the front-to-back ratio of the log cell while the director sharpens the forward lobe and decreases the half-power beamwidth. As the spacing between the parasitic elements and the log cell decreases, the parasitic elements must increase in length.[10]

The log cell is designed to meet upper and lower band limits with $\sigma = 0.05\lambda$. The design parameter τ is dependent on the structure bandwidth, B_s. When the log-periodic design parameters have been found, the element length and spacings can be determined. A review of the "Log-Periodic Dipole Array" is recommended though not necessary for the design of the Log-Yag array.[3,7,8]

The method of feeding the antenna is identical to that of feeding the log-periodic dipole array without the parasitic elements. As shown in Fig. 11, a balanced feeder is required for each log-cell element, and all adjacent elements are fed with a 180° phase shift by alternating connections. Since the Log-Yag array will be covering a relatively small bandwidth, the radiation resistance of the narrow-band log cell will vary from 80 to 90 ohms (tubing elements) depending on the operating bandwidth. The addition of parasitic elements lowers the log-cell radiation resistance. Hence, it is recommended that a 1-to-1 balun be connected at the log-cell input terminals and 52-ohm coaxial cable be used for the feed line. The measured radiation resistance of the 14-MHz Log-Yag installed at the author's QTH is 37 ohms, 14.0 to 14.35 MHz. It is assumed that tubing elements will be used. However, if a wire array is used then the radiation resistance R_o and antenna-feeder input impedance Z_o must be calculated so that the proper balun and coax may be used. The procedure is outlined in detail in *The ARRL Antenna Book*.[3,7]

Design Procedure

The following step-by-step design procedure may be used to design any monoband Log-Yag for any desired bandwidth.

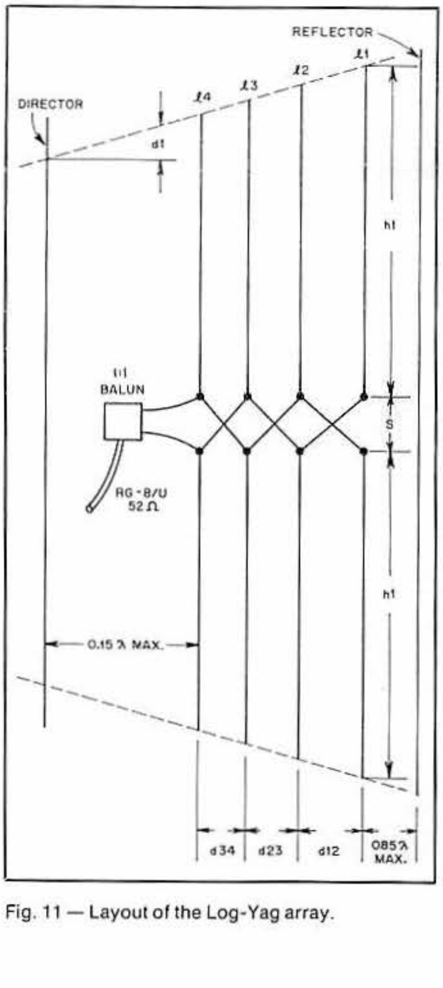

Fig. 11 — Layout of the Log-Yag array.

1) Determine the operating bandwidth, B, between f_1, lowest frequency (band edge), and f_n, highest frequency (band edge).

$$B = \frac{f_n}{f_1}$$

2) Determine the structure bandwidth (log-cell array) B_s.

$$B_s = 1.15B$$

3) Determine the design parameter τ (based on 4-element log cell, note 1).

$$\tau = \frac{1}{\sqrt[3]{B_s}}$$

Note 1. The design parameter τ is chosen for a four-element log cell since it provides the best bandpass for most amateur bands. For log cells with any number of elements

$$\tau = \frac{1}{\sqrt[(n-1)]{B_s}}$$

where n = number of elements within the log cell.

4) Determine the apex half-angle α: Since $\sigma = 0.05$ (relative spacing constant), then

$$\cot \alpha = \frac{0.2}{1 - \tau}$$

5) Determine the longest free-space wavelength λ_{max}, log-cell boom length, L (ft) and longest element length within the log cell l_1 (ft).

$$\lambda_{max} = \frac{984}{f_1 \text{MHz}}$$

$$L = \left[\frac{1}{4}\left(1 - \frac{1}{B_s}\right) \cot \alpha\right] \lambda_{max}$$

and

$$l_1 = \frac{492}{f_1 \text{MHz}}$$

$$l_2 = \tau l_1$$

$$l_3 = \tau l_2$$

$$l_4 = \tau l_3$$

6) Determine the element spacing (d_{12}), distance between elements l_1 and l_2 (ft).

$$d_{12} = \frac{1}{2}\left(l_1 - l_2\right) \quad \cot \alpha$$

and

$$d_{23} = \tau d_{12}$$

$$d_{34} = \tau d_{23}$$

7) Determine the parasitic element lengths (ft) and spacings (ft).

$$l_{REF} = \frac{509.6}{f_1 \text{ MHz}}$$

$$d_{REF} = \frac{84}{f_1 \text{ MHz}}$$

$$l_{DIR} = \frac{450.8}{f_1 \text{ MHz}}$$

$$d_{DIR} = \frac{148}{f_1 \text{ MHz}}$$

This completes the design.

The Finished Log-Yag

The proof is always to be found in the completed and operating product. The

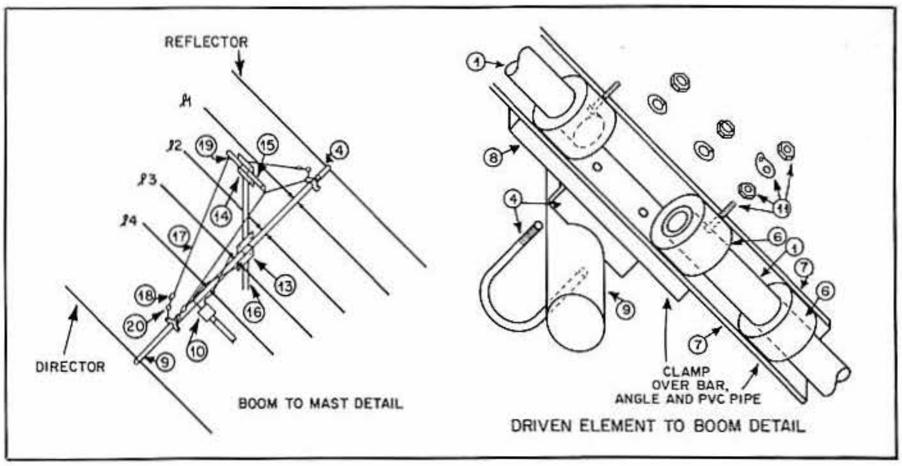

Fig. 12 — Assembly details. The numbered components refer to Table 6.

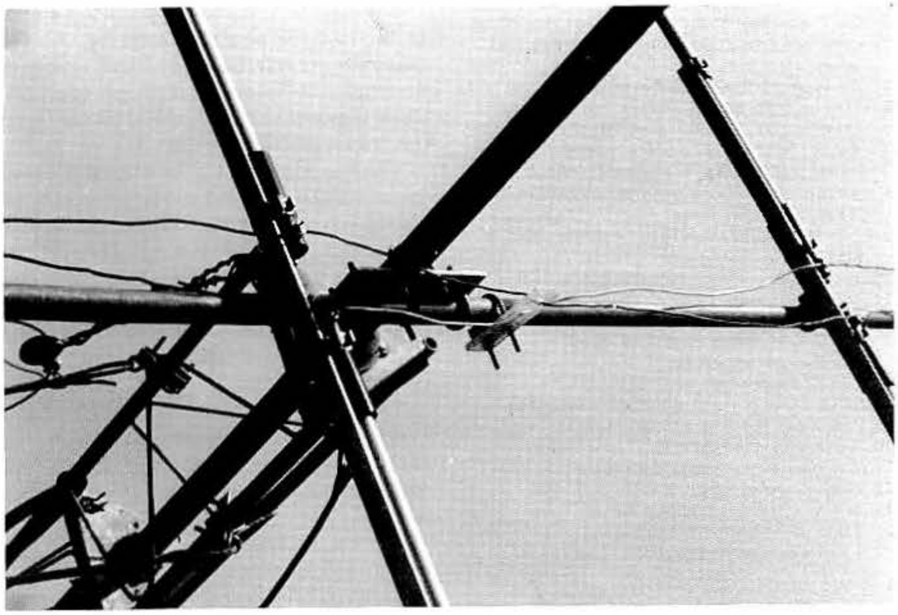

The interconnection between the elements of the log cell changes sides between each element.

Table 3

Array Characteristics

1. Frequency range	14-14.35 MHz
2. Operating bandwidth	B = 1.025
3. Design parameter	τ = 0.946657
4. Apex half angle	a = 14.92°, cot = 3.753
5. Half-power beam width	42° (14-14.35 MHz)
6. Bandwidth of structure	B_s = 1.17875
7. Free-space wavelength	λ_{max} = 70.28 ft
8. Log cell boom length	L = 10.0 ft
9. Longest log element	l_1 = 35.14 ft (a tabulation of element lengths and spacings given in Table 2)
10. Forward gain over dipole	11.5 dB (theoretical)
11. Front-to-back ratio	32 dB (theoretical)
12. Front-to-side ratio	45 dB (theoretical)
13. Input impedance	Z_0 = 37 ohms
14. SWR	1.3 to 1 (14-14.35 MHz)
15. Total weight	96 pounds
16. Wind-load area	8.5 sq. ft
17. Feed-point impedance	Z_0 = 37 ohms
18. Reflector length	36.4 ft @ 6.0 ft spacing
19. Director length	32.2 ft @ 10.5 ft spacing
20. Total boom length	26.5 ft

The mechanical construction of the log cell is identical to that described in *The ARRL Antenna Book*, except for the lengths and spacings.[3,7] Fig. 12 shows how the log cell is constructed as well as the addition of the parasitic elements.

1 ft = 0.30478 m

Table 4

Array Dimensions

Element	*Length Ft*	*Spacing Ft*
Reflector	36.4	6.0 (Ref. to l_1)
l_1	35.14	3.51 (d_{12})
l_2	33.27	3.32 (d_{23})
l_3	31.49	3.14 (d_{34})
l_4	29.81	10.57 (l_4 to dir.)
Director	32.2	

1 ft = 0.30478 m

Table 5

Element Material Requirements

Element	*1-in. Tubing Length (Ft)*	*Qty.*	*7/8-in. Tubing Length (Ft)*	*Qty.*	*3/4-in. Tubing Length (Ft)*	*Qty.*	*1 1/4-in. Angle Length (Ft)*	*1 X 1/4-in. Bar Length (Ft)*
Reflector	12	1	6	2	8	2	None	None
l_1	6	2	6	2	8	2	3	1
l_2	6	2	6	2	8	2	3	1
l_3	6	2	6	2	6	2	3	1
l_4	6	2	6	2	6	2	3	1
Director	12	1	6	2	6	2	None	None

1 ft = 0.30478 m

Table 6

Materials List

1. Aluminum tubing — 0.047 in. wall thickness
    1 in. — 12 ft lengths, 24 lin. ft
    1 in. — 12 ft or 6 ft lengths, 48 lin. ft
    7/8 in. — 12 ft or 6 ft lengths, 72 lin. ft
    3/4 in. — 8 ft lengths, 48 lin. ft
    3/4 in. — 6 ft lengths, 36 lin. ft
2. Stainless steel hose clamps — 2 in. max., 8 ea.
3. Stainless steel hose clamps — 1-1/4 in. max., 24 ea.
4. TV-type U bolts — 1-1/2 in., 6 ea.
5. U bolts, galv. type: 5/16 in. X 1-1/2 in., 4 ea.
6. U bolts, galv. type: 1/4 in. X 1 in., 2 ea.
7. 1 in. ID water-service polyethylene pipe 160 lb/in.^{2} test, approx. 1-3/8 in. OD, 7 lin. ft
8. 1-1/4 in. X 1-1/4 in. X 1/8 in. aluminum angle — 6 ft lengths, 12 lin. ft
9. 1 in. X 1/4 in. aluminum bar — 6 ft lengths, 6 lin. ft
10. 1-1/4 in. top rail of chain-link fence, 26.5 lin. ft
11 1:1 toroid balun, 1 ea.
12. No. 6-32 X 1 in. stainless steel screws, 8 ea.
    No. 6-32 stainless steel nuts, 16 ea.
    No. 6 solder lugs, 8 ea.
13. No. 12 copper feed wire, 22 lin. ft
14 12 in. X 6 in. X 1/4 in. aluminum plate, 1 ea.
15. 6 in. X 4 in. X 1/4 in. aluminum plate, 1 ea.
16. 3/4 in. galv. pipe, 3 lin. ft
17. 1 in. galv. pipe — mast, 5 lin. ft
18. Galv. guy wire, 50 lin. ft
19. 1/4 in. X 2 in. turnbuckles, 4 ea.
20. 1/4 in. X 1-1/2 in. eye bolts, 2 ea.
21. TV guy clamps and eyebolts, 2 ea.

1 in. = 25.4 mm
1 ft = 0.30478 m

author's 14-MHZ Log-Yag on-the-air performance on cw and ssb substantiates the theory. The characteristics of the array are given in Table 3.

The materials needed are given in Table 5. In the construction diagram, Fig. 12, the materials are referenced by their respective material list number. The photographs show the overall construction picture, and the drawings show the details.

The materials should be available from most hardware and electronic stores. However, some have found difficulty in obtaining aluminum tubing. This can be solved by writing to the manufacturer and asking for the name of their distributor nearest your locality. Commercial antenna manufacturers will sell their tubing, but the cost is at a premium.

This array is in operation at K4EWG and W4BBP. The results on the air are nothing short of fantastic! It will give the stacked Yagis and long-boom Yagis a run for their money.

It is the authors' hope that this antenna design will stimulate additional work and research by other amateurs. The field seems wide open, and Yag-Log combinations are endless. The optimum design is by no means achieved in this article. It does seem, however, that a log cell of more than four elements would be necessary only where the array bandwidths, B, exceed 1.03 ($B = f_n/f_1$).

The authors wish to thank George Smith, W4AEO, for his work in substantiating a consideration for log periodic gain.[8] This material was originally presented in *QST* by P. D. Rhodes, K4EWG and J. R. Painter, W4BBP.

References

[1] "The Swan Multidrive 2-Meter Antenna," *QST*, Oct. 1969.
[2] Orr, *Radio Handbook*, 18th edition, pp. 547-548.
[3] Rhodes, "The Log-Periodic Dipole Array," *QST*, November 1973.
[4] *The ARRL Antenna Book*, 13th edition, p. 151.
[5] See ref. 2 above, p. 558.
[6] King, Mack and Sandler, "Arrays of Cylindrical Dipoles," *Yagi-Uda Programme Apdx. V*, 1968, pp. 218-232,468-470.
[7] See ref. 4 above: pp. 160-164, 208-210, and Fig. 9-18.
[8] Smith, "Yes, I've Built Sixteen Log Periodic Antennas!" *73*, March 1975. See pp. 98-99.
[9] Orr, "Antennas," *CQ*, March 1975. See section headed "The KLM Antenna."
[10] See ref. 4 above, p. 204 and Fig. 9-5.

Chapter 3

Quad Antennas

A currently popular form of directional gain antenna is called the "quad." In general performance principles, it compares favorably to the Yagi beams discussed in chapter 2. The proper name for this antenna is the "cubical quad." It was developed by Clarence Moore, W9LZX, in 1942 when he was an engineer for HCJB in Quito, Ecuador. The design came as a solution to large corona discharges at the ends of existing beam-antenna elements. The problem was caused by the highly ionized air at the transmitter location, some two miles above sea level. Operation was in the commercial 25-meter band, which for the intentions of the broadcaster, required a gain type of antenna.

Moore concluded that a gain antenna with no ends was required to handle the 10-kW power without corona or flashover. Thus, the quad was born. Not only did it handle the power without difficulty, it later became a popular amateur antenna after he developed a model for use on 20 meters.

The primary limitation of quad antennas is the unwieldy nature of the "beast," especially at 40 and 20 meters, compared to a Yagi. The supporting framework is subject to extreme stress from ice and wind loading. Nonetheless, the quad has proven to be one of the most effective DX beams that an amateur can build at modest cost. This chapter contains data on various types of quads for a number of amateur bands. It is a compendium of the best on the subject, as published in past issues of *QST*.

Quads and Yagis: Comparisons, Patterns and Working Dimensions

Want a fresh look at the Yagi vs. quad question? This is "must" reading for the antenna-minded.

The material in the following sections will present a discussion of two time-tested antenna arrays.

In this class of arrays the Yagi (more properly the Yagi-Uda) antenna is the grandfather of them all. There are very few antennas which can compete with the electrical operation and simplicity of construction of this array. However, in the early 1940s C. C. Moore, of radio station HCJB, brought into being a parasitic array of loop elements with properties very comparable to those of the Yagi-Uda array. Since its conception this array has become very popular in Amateur Radio circles. It has come to be known as the "cubical quad" or "quad" antenna because of the common use of square loops and box-like construction.[1]

The material to follow will attempt to present an unbiased comparison of these two arrays based upon the experimental and theoretical results of many workers. This material is prefaced by a short section of introductory antenna concepts and definitions so that certain terms and notation will have more meaning when the comparisons between the quad and Yagi antennas are made. Finally, in the last section, some of the experimental patterns of modeled quad and Yagi arrays are presented. A compilation of some of the working dimensions for both types of arrays are also given. This latter material should be of interest to the builder and experimenter.

Introductory Antenna Theory and Definitions: Polarization

Inspection of the radiation field of an antenna will show it to have plane-wave character and to be made up of electric and magnetic field lines which are perpendicular to each other. Polarization of an electromagnetic wave is defined as that direction in which the electric field lines are oriented. If the electric field vector is fixed (in direction) in a plane it is said to be linearly polarized. For example, the polarization of a dipole is linear. If the plane in which this dipole lies is horizontal or parallel to some reference plane (e.g., the surface of the earth), it is called a horizontally polarized antenna. Vertical polarization would result if the dipole were placed perpendicular to the reference plane. It should be noted that if the electric vector does not stay fixed in position but rotates (its tip tracing out an ellipse), then the wave is said to be elliptically polarized. Circular polarization is a special case of elliptical polarization. Elliptical polarization will have a right- or left-hand sense, depending upon the direction of rotation of the electric vector.

Directivity and Gain of an Antenna

The degree to which a particular antenna can concentrate the antenna pattern into a beam is known as *directivity*. Mathematically, it is the ratio of the power density (in watts per square unit of measure) that would be available at an observation point to the power density one would have if the total radiated power were radiated *equally* well in *all* directions.[2] Directivity is a function of the antenna pattern alone and does not take into account antenna efficiency or losses.

The gain of an antenna can be defined in terms of directivity when antenna efficiency is taken into consideration. If K is the antenna efficiency, a number between 0 and 1, then antenna gain is given by

$$G = KD$$

where

G = antenna gain
K = antenna efficiency
D = directivity

The gain in decibels is then given by

$$\text{dB gain} = 10 \log_{10} G.$$

Please note that the above definition has been such that it is a function of the observation point. In amateur circles gain is usually measured at that observation point where the maximum power density is observed. Further, when considering practical Yagi and quad arrays for the hf range, K is normally taken as 1. Gain measured in the above manner is normally called gain with respect to an isotropic antenna. It is interesting to note that since a half-wave dipole concentrates its pattern into certain directions, it will have some gain with respect to an isotropic antenna. This gain with respect to isotropic is found to be

$$G = 1.64$$

or, in dB,

$$G_{dB} = 2.15 \text{ dB}$$

If the vertical and horizontal half-power beam widths are known for a given antenna the following formula will give (somewhat on the generous side) the gain of the antenna with respect to an isotropic antenna

$$G = \frac{40{,}000}{\theta_H \theta_V}$$

where

θ_H = horizontal half-power beam width in degrees.

θ_V = vertical half-power beam width in degrees.

This formula assumes low-side and back-lobe levels for the antenna pattern. If the gain with respect to isotropic is known then the gain with respect to a half-wave dipole is obtained by subtracting 2.15 dB.

Antenna Aperture or Area

This concept is of most use when discussing a receiving antenna. It is a measure of the antenna's ability to gather in available radiated power at some receiving position. Antenna area and gain are *directly* related

$$G = 4\pi \frac{A}{\lambda^2}$$

where

G = antenna gain

λ = operating wavelength

A = area (based upon square units of length measurement, with the length measurement the same as that assumed for λ). It should be noted that one *cannot* have two antennas with the same gain and different areas. One hears this concept misused (in the sense of the above sentence) quite often on the amateur bands. That is to say, if you feel that one antenna has more capture area than another, then it must also have more gain.

Antenna Patterns

The pattern of an antenna may be plotted by recording the received electric field strength or power density as a function of viewing (observation) angle. The properties of most antennas used by radio amateurs (below 30 MHz) can be described by two basic patterns. Consider a pattern due to a radiating half-wave dipole, this pattern to be obtained by moving a sampling or test antenna about the dipole in the horizontal plane which contains the dipole. This pattern is called the horizontal-plane or E-plane pattern (the pattern was taken in a plane containing the electric E-lines, hence the name E-plane pattern). If the sampling antenna is placed directly in front of the horizontal dipole (and in the horizontal plane) and moved up over the top of the dipole (in a plane perpendicular to the horizontal plane), and on around until it returns to the starting position, the result is the vertical-plane or H-plane pattern (since the test antenna moved in a plane parallel to the magnetic H-lines it is called an H-plane pattern).

For Yagi and quad arrays these two patterns will give enough essential information to determine their expected radiating properties.

Comparison of Quad and Yagi Arrays

The cubical quad antenna was first introduced to the radio world in 1942 by Clarence C. Moore at HCJB to alleviate a serious corona problem occurring on a large commercial shortwave Yagi array. It is a parasitic array formed with loop elements which are approximately one wavelength in circumference.

Important to proper operation of a quad antenna are considerations of proper element sizes, spacing between elements, boom length, feed impedance and so forth. The following material presents a discussion of these considerations and a comparison of the quad and Yagi arrays.

Polarization of a Quad Antenna

The polarization of a quad antenna can be determined by considering the current distribution on the driven element. To stress the point that square, diamond, or circular loop all produce essentially the same polarization, consider the current distribution shown in Fig. 1. A little thought reveals that fact that the vertical components of current produce radiated fields that cancel each other, and the horizontal components of current produce a radiated field (broadside to the loop) with all components adding. The polarization of any of these loops is then linear and parallel to an imaginary horizontal line drawn through the current minima. Hence, if a quad loop is fed at the bottom it produces horizontal polarization, and if fed on the side it will have vertical polarization. The geometrical form of the loop, square, diamond or circular, does not mean that the (broadside) radiation has a possible combination of both horizontal and vertical polarization or circular polarization. This seems to be a common misconception in amateur circles. The linear polarization properties introduced above can be verified easily experimentally. It should be further noted that there are no noticeable differences in the radiating properties or impedance of a loop 1 wavelength in circumference whether it be of the circle, diamond or square configuration.

Angle of Radiation for a Horizontally Polarized Quad Antenna

Placement of an antenna above a conducting ground will alter its free-space radiation pattern, positioning the maximum of its main lobe at possibly a new angle with respect to a horizontal reference line. The angular position of this main lobe is called the *angle of radiation.* The angle of radiation of an antenna is a function of its height above the conducting ground and its polarization. Given the height and the polarization, the angle of radiation can be determined by using the concept of the antenna's virtual image or ray theory and reflection of the wave by the ground.

The following conclusions can now be drawn:

The angle of radiation of a *horizontally* polarized quad antenna, at a given boom height, is essentially the *same* as that of a *horizontally* polarized Yagi at the same boom height.

This must be so because the boundary conditions imposed upon the electromag-

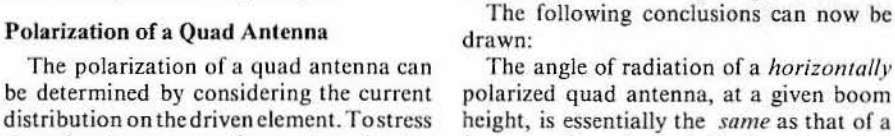

Fig. 1 — Current distribution on circular, diamond and square loops 1 wavelength in circumference.

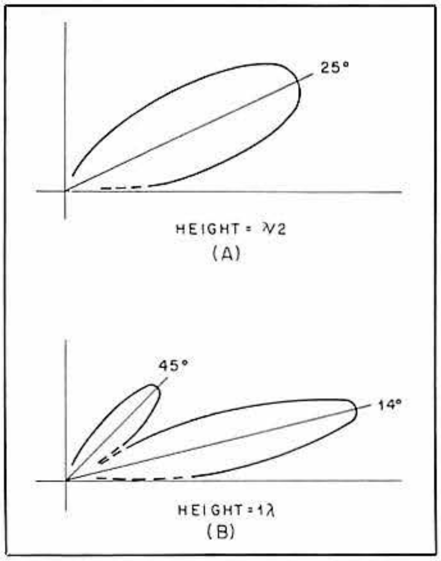

Fig. 2 — Vertical-plane radiation patterns for a horizontally polarized quad or Yagi array above perfectly conducting ground.

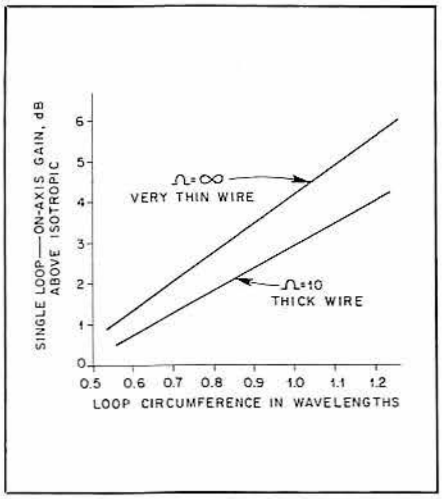

Fig. 3 — On-axis gain of a single loop as a function of loop circumference. (From "Studies of Large Circular Loop Antennas" by S. Adachi and Y. Mushiake *Sci. Rep. Ritu*, B-[Elect. Comm.] vol. 9, no. 2, Tohoku Univ., Sendai, Japan, 1957.)

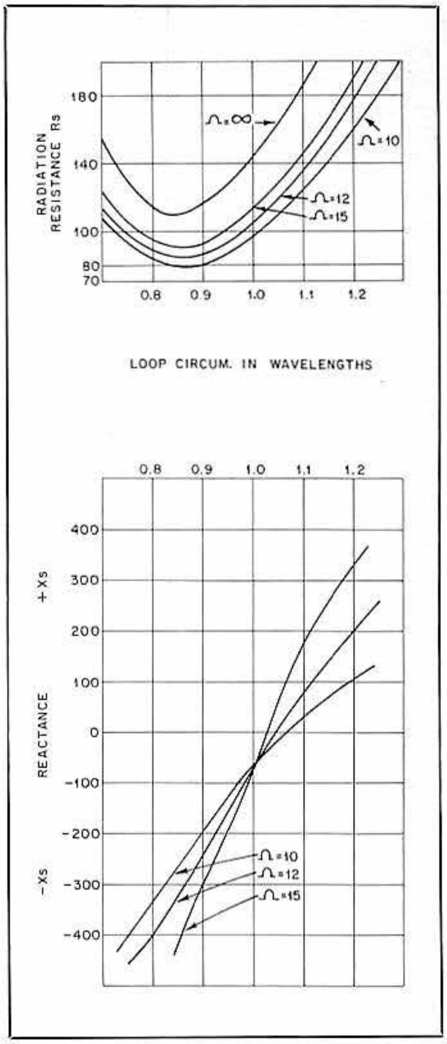

Fig. 4 — Reactance and radiation resistance of a single loop versus circumference. Note: Ω is a measure of the ratio of loop circumference to wire diameter. Ω = 2ln 8a/ρ where a = loop radius and ρ = wire radius.

netic wave at the surface of ground are completely insensitive to the fact that the wave is produced by a quad antenna as opposed to the wave produced by a Yagi antenna.

Typical patterns of a quad *or* Yagi antenna placed at a height of 1/2 λ or 1 λ above a perfect ground are shown in Fig. 2 at A and B.

Gain Comparisons of the Quad Antenna with the Yagi Antenna

The gain of a quad array (of given boom length) seems to be somewhat higher than that of a Yagi array of the same boom length. This gain difference can be explained by considering the theoretical work of Adachi and Mushiake of Tohoku University in Japan. In 1952 and 1957 they published two papers which discussed the current distribution, impedance, and radiating properties of loops of large circumference.[3]

Fig. 3 shows the gain of a single loop as a function of wavelength and wire diameter. Fig. 4 shows the variation of reactance and radiation resistance for a single loop. Please note the dependence upon the circumference-to-wire-diameter ratio in these two figures. They showed that the broadside gain of a single loop 1λ in circumference was approximately 2 dB above that of a half-wave dipole. This result implies that a parasitic array of loop elements should have the same gain advantage when compared to a Yagi array of comparable boom length.

This gain differential for the quad array does appear and can be verified experimentally. Fig. 5 demonstrates this comparison. The points on the figure were obtained as follows: For the antenna under consideration E and H plane patterns were made. The gain was then calculated using the E and H-plane half-power beam widths of these radiation patterns and the formula discussed above. This gain and the boom length of the antenna were then used to determine a point on the graph given in Fig. 5, where gain and boom length represent the parameters under consideration. The data points for many, many antennas, both quad and Yagi, were then placed on the graph and a smooth curve representing an average for each type of antenna was then drawn.

The expected gain difference of 2 dB does indeed appear in favor of the quad antenna. This gain differential is essentially that predicted by the results of the Japanese workers cited above.

The measurements used to obtain the above gain comparisons were performed at 440 MHz. A measuring frequency of 440

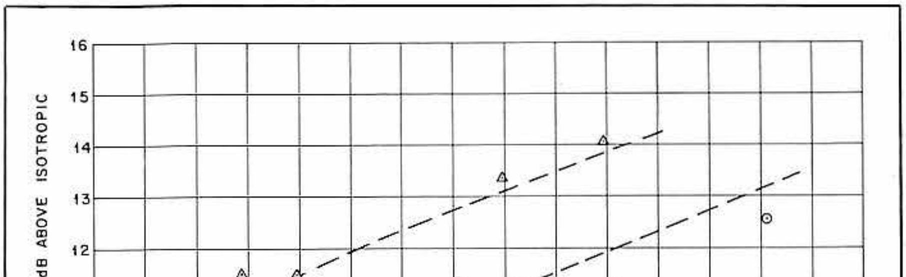

Fig. 5 — Quad and Yagi array gain as a function of boom length.

Table 1

Quad Dimensions

2-element Quad (W0HTH)
Spacing (given below)
Boom length (given below)

Band	*40 M*	*20 M*	*15 M*	*10 M*
Reflector	144 ft 11-1/2 in.	72 ft 4 in.	48 ft 8 in.	35 ft 7 in.
Driven Element	140 ft 11-1/2 in.	70 ft 2 in.	47 ft 4 in.	34 ft 7 in.
Spacing	30 ft	13 ft	10 ft	6 ft 6 in.
Boom length	30 ft	13 ft	10 ft	6 ft 6 in.
Feed method	Directly with 23 ft of RG-11, then any length RG-8 coax.	Directly with 11 ft 7 in. RG-11, then any length RG-8 coax.	Directly with 7 ft 8-1/2 in. RG-11, then any length RG-8 coax.	Directly with 5 ft 8 in. RG-11, then any length RG-8 coax.

(Note that a spider or boomless quad arrangement could be used for the 10/15/20 meter parts of the above dimensions yielding a triband antenna)

4-element Quad* (W0AIW [20 M] /W0HTH** /K0KKU/K0EZH/W6FXB)
Spacing: equal; 10 ft
Boom length: 30 ft

Band	*20 M*		*15 M*	*10 M*
	Phone	CW		
Reflector	72 ft 1-1/2 in.	72 ft 5 in.	48 ft 8 in.	35 ft 8-1/2 in.
Driven Element	70 ft 1-1/2 in.	70 ft 5 in.	47 ft 4 in.	34 ft 8-1/2 in.
Director 1	69 ft 1 in.	69 ft 1 in.	46 ft 4 in.	33 ft 7-1/4 in.
Director 2	69 ft 1 in.	69 ft 1 in.	46 ft 4 in.	33 ft 7-1/4 in.
Feed Method	Directly with 50-ohm coax.		Directly with 50-ohm coax.	Directly with 5 ft 9 in. RG-11, then any length RG-8 coax.

*Common boom used to form a triband array.
**The 2-element 40-meter quad given above is added to form a four-band quad array.

4-element Quad (W0HTH/K8DYZ*/K8YIB*/W7EPA*)
Spacing: equal; 13 ft 4 in.
Boom length: 40 ft

Band	*20 M*	*15 M*	*10 M*
Reflector	72 ft 5 in.	48 ft 4 in.	35 ft 8-1/2 in.
Driven Element	70 ft 5 in.	47 ft 0 in.	34 ft 8-1/2 in.*
Director 1	69 ft 1 in.	46 ft 1 in.	(Directors 1-3 all
Director 2	69 ft 1 in.	46 ft 1 in.	33 ft 7 in.)*
Feed method	Directly with 50-ohm coax.	Directly with 7 ft 9 in. RG-11, then any length 50-ohm coax.	Directly with 50-ohm coax.

*For the 10-meter band the driven element is placed between the 20/15 reflector and 20/15 driven element. The 10-meter reflector is placed on the same frame as the 20/15-meter reflectors and the remaining 10-meter directors are placed on the remaining 20/15-meter frames. The 10-meter portion is then a 5-element quad.

6-element Quad (W0YDM, W7UMJ)
Spacing: equal; 12 ft
Boom length: 60 ft

Band: 20 M	
Reflector	72 ft 1-1/2 in.
Driven Element	70 ft 1-1/2 in.
Directors 1, 2 and 3	69 ft 1 in.
Director 4	69 ft 4 in.
Feed Method	Directly with 50-ohm coax.

1 ft = 0.3048 m

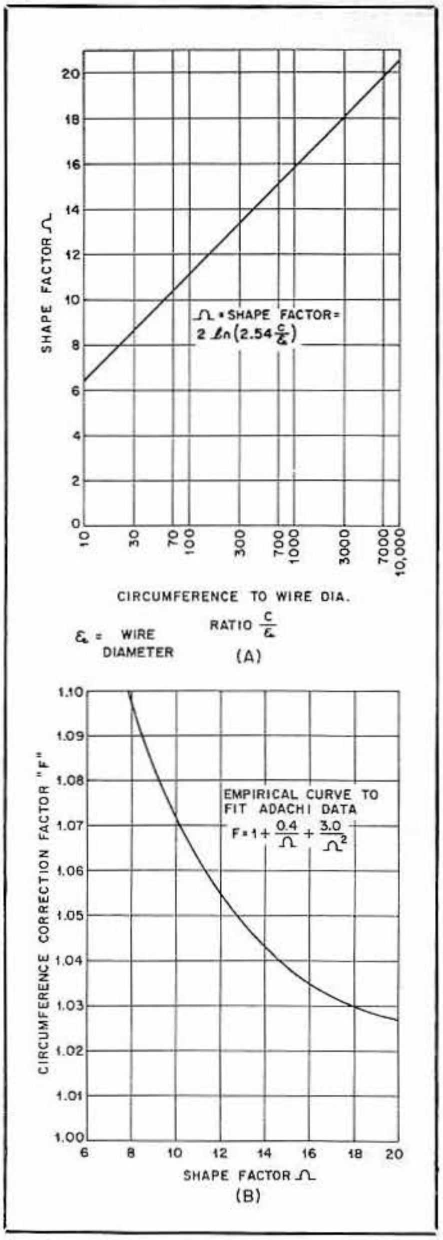

Fig. 6 A: Shape factor as a function of the loop circumference-to-wire diameter ratio. B: Loop "resonant length correction factor" versus shape factor. (Circled values obtained from S. Adachi, and Y. Mushiake, "Studies of Large Circular Loop Antennas" *Sci. Rep. Ritu*, B-[Elect. Comm.] vol. 9, no. 2, Tohoku Univ., Sept. 1957.)

MHz was chosen so that 1 cm (at this frequency) is equivalent to 1 foot of measurement at 14 MHz. Hence, to relate these results directly to the 20-meter amateur band, boom length in cm from the curve can be read as boom length in feet. For example, the gain of a quad on a 40-foot (12.19-m) boom should be slightly more than 1 dB above a quad on a 25-foot (7.62-m) boom. Comparison of the signals from these two antennas at a receiving station will not reveal much difference; however, in the DX pile-ups the longer antenna should have a slight advantage. Another interesting interpretation of these results is to note that a quad on a 30-foot (9.14-m) boom is equivalent to a Yagi on a 55-foot (16.76-m) boom. Hence the boom of a Yagi antenna must be about 1.8 times longer than the boom of a quad in order for the two gains to be comparable.

Impedance Properties and Matching

The driving-point impedance of a quad antenna depends upon the element spacing and number of elements. For 2-element quad antennas the results of Bill Orr in his *Quad Antenna* handbook can be used to determine the driving-point impedance. For multielement quads (3 or more elements)[4] the impedance will range from 40 to 60 ohms. Matching of these multielement quads is most easily done by direct feed with a 50-ohm coaxial line. A 1:1 balun can be used but experimental results show that it does not give performance any better than that obtained with direct feed. A gamma match can be used to obtain an exact match to any chosen coaxial line. For details of using a gamma match on a quad antenna see Bill Orr's *Quad Antenna* handbook.

If a single boom is used to construct a multiband quad antenna, then it is recommended that a separate feed line to each driven element be used.

Approximate Formulas for Quad Antenna Element Lengths

The multielement quad antenna gained its first real popularity following the work of Lee Bergren, W0AIW, in the early 1960s. He published the results of his work in *QST* in 1963.[5] Many experimentalists have used his dimensions for antennas in the 20-meter amateur band and have scaled them to other bands. Based upon Bergren's work and others the following empirical formulas can be used to determine the element lengths of a quad antenna

$$\text{Circumference of driven element} = \frac{1005}{f_{\text{MHz}}} \text{ feet}$$

$$\text{Circumference of reflector element} = \frac{1030}{f_{\text{MHz}}} \text{ feet}$$

$$\text{Circumference of director elements} = \frac{975}{f_{\text{MHz}}} \text{ feet}$$

These results should only be used for quad antennas operating below 30 MHz. It is interesting to note that for quad antennas the reflector and directors are only 2-1/2 to 3 percent different in length from the driven element. This is a smaller change than that normally encountered in Yagi arrays.

The above figures should not be used for vhf or uhf quad antennas. At these higher frequencies the circumference-to-wire-diameter ratio becomes small enough that its effect upon the resonant properties of the loop must be considered. Here again, the results of Adachi and Mushiake are useful and the effects of wire size on the resonant length of the loop can be taken into account. These results are given in graphical form in Fig. 6 where, based upon the circumference-to-wire-diameter ratio, a circumference correction factor can be obtained. Once the correct size of the driven element is determined (using the results of Fig. 6) a quad antenna can be designed by making the reflector approximately 3 percent longer and the directors 3 percent shorter.

The above results can be used to obtain workable dimensions but, just as with approximate formulas for Yagi antenna element lengths, the "best" dimensions would have to be determined experimentally by tuning the array.

A Collection of Quad and Yagi Patterns and Dimensions: E and H Plane Patterns for Quad and Yagi Arrays

The previous section referred to a group of experimental tests, where a series of model antennas were constructed and the E and H plane patterns obtained.

It should be noted that these patterns have direct interpretation in light of the 20-meter band. Recall that the experiments were performed at 440 MHz, where dimensions for boom length, element size, and spacing in cm are directly equivalent to dimensions in feet for the 20-meter band. For example, the patterns for a 440-MHz, 5-element Yagi on a 55-cm boom can also be taken as the pattern for a 14-MHz, 5-element Yagi on a 55-foot boom.

Perusal of these patterns will give a

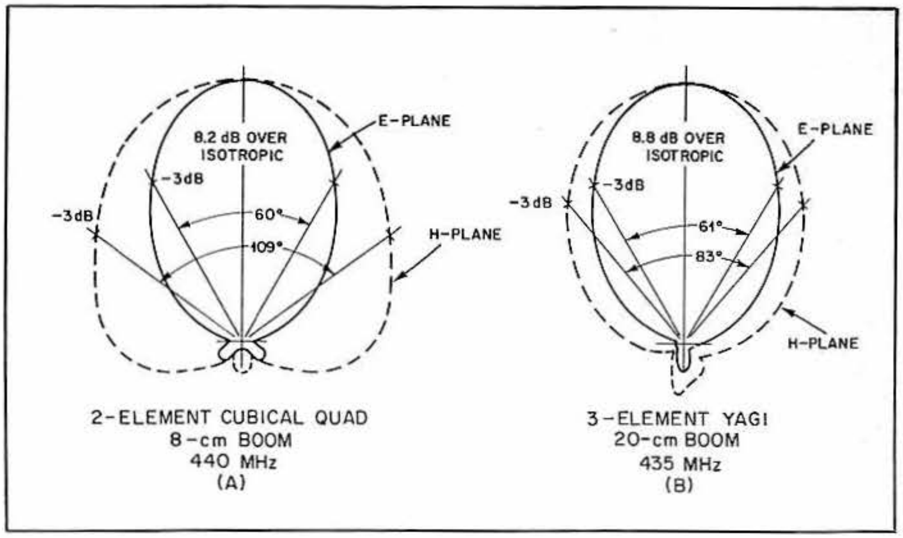

2-ELEMENT CUBICAL QUAD
8-cm BOOM
440 MHz
(A)

3-ELEMENT YAGI
20-cm BOOM
435 MHz
(B)

Fig. 7 — E and H plane radiation patterns for (A) 2-element quad antenna and (B) 3-element Yagi antenna.

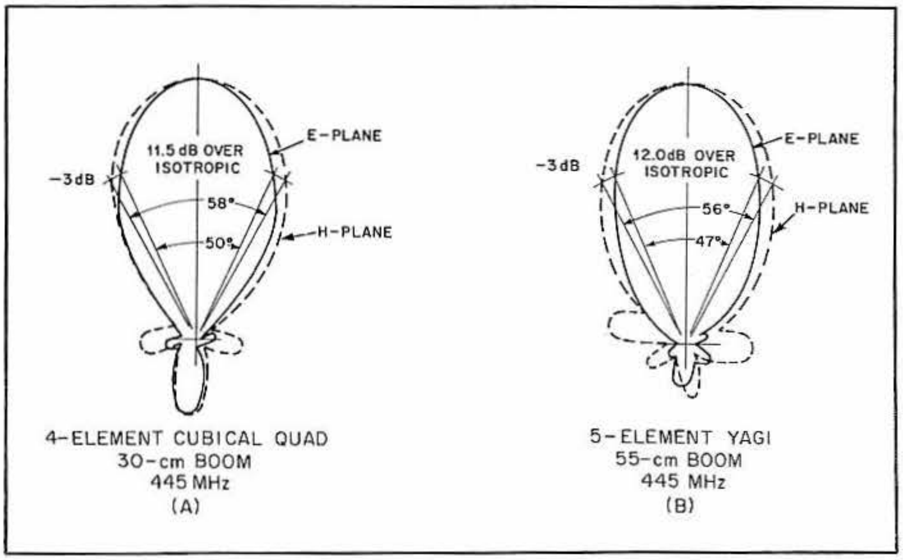

4-ELEMENT CUBICAL QUAD
30-cm BOOM
445 MHz
(A)

5-ELEMENT YAGI
55-cm BOOM
445 MHz
(B)

Fig. 8 — E and H plane radiation patterns for (A) 4-element quad antenna and (B) 5-element Yagi antenna.

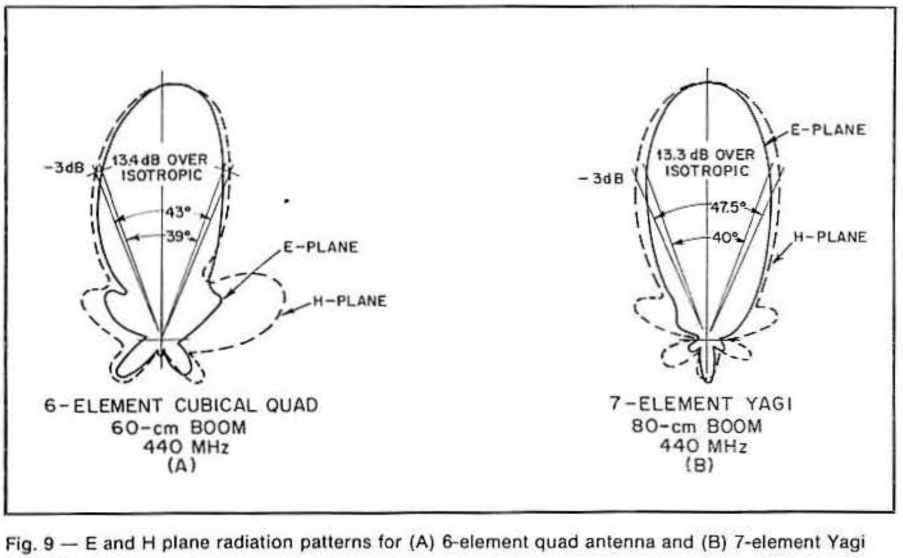

6-ELEMENT CUBICAL QUAD
60-cm BOOM
440 MHz
(A)

7-ELEMENT YAGI
80-cm BOOM
440 MHz
(B)

Fig. 9 — E and H plane radiation patterns for (A) 6-element quad antenna and (B) 7-element Yagi antenna.

visual means of comparing the quad and Yagi arrays. It is interesting to note the dependence of gain and pattern beam width upon array boom length. Without further comment, typical patterns for both types of antennas are given in Figs. 7 through 9.

A Tabulation of Dimensions of Working Arrays

Presented in tabular form are a few representative dimensions for both quad and Yagi antennas. These are dimensions presently in use by a number of amateurs. Most of the Yagi dimensions are due to amateurs across the country who sent their dimensions to Mr. Hal Wolff or the author to be modeled on the antenna range at the University of Denver. The quad dimensions also are those now in use by a number of amateurs in this country and abroad. It should be noted that most of these dimensions, and the resulting model antennas, provided the necessary data used in the previous section to compare the quad and Yagi. Where known, either the user(s) or the originator of a set of dimensions is indicated. The first set of dimensions is for quad antennas, the second set for Yagi antennas.

Summary

A comparison between the quad and Yagi antennas, on the basis of gain, has been given, this comparison being based upon both theoretical and experimental results. The results indicate that if both antennas have the same boom length the gain of the quad array will be about 2 dB higher than that of the Yagi. For either antenna, gain is a function of boom length and will increase approximately 2.5 dB each time the boom length is doubled. Hence, a 2-dB gain advantage is equivalent to a boom-length advantage of 1.8. In other words, for the quad and Yagi arrays to have comparable gain, the boom length of the Yagi antenna must be 1.8 times longer than that of the quad.

The size of the wire used for the loop element (relative to the loop circumference) will have an influence upon the resonance and radiating properties of the loop. The influence of this shape factor — i.e., the ratio of loop circumference to wire diameter — can be accounted for by using the referenced results of Adache and Mushiake.

Table 2

Yagi Dimensions

3-element Yagi (W6SAF)
Boom length: 24 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	34 ft 10-3/4 in.	R—DE: 13 ft 9-1/2 in.
Driven Element	33 ft 2-1/2 in.	DE—D: 10 ft 4-1/2 in.
Director	31 ft. 9-1/2 in.	

3-element Yagi (W0OKC)
Boom length: 26 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	R—DE: 14 ft
Driven Element	33 ft 4 in.	DE—D: 12 ft
Director	31 ft 7 in.	

3-element Yagi (W0CM, formerly W0JYW)
Boom length: 30 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 4-1/2 in.	R—DE: 13 ft
Driven Element	32 ft 2-1/2 in.	DE—D: 17 ft
Director	31 ft 10 in.	
Feed Method	Folded dipole	

4-element Yagi (K0KKU)
Boom length: 30 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	(all elements equally spaced)
Driven Element	33 ft 4 in.	
Director 1	31 ft 9-1/2 in.	
Director 2	31 ft 2 in.	

4-element Yagi (W4PLL)
Boom length: 30 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 4-1/2 in.	(all elements equally spaced)
Driven Element	33 ft 4-1/2 in.	
Director 1	31 ft 9-1/2 in.	
Director 2	31 ft 9-1/2 in.	

4-element Yagi (K6VIZ)
Boom length: 32 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	R—DE: 10 ft
Driven Element	32 ft 8 in.	DE—D1: 10 ft
Director 1	31 ft 8 in.	D1—D2: 12 ft
Director 2	31 ft 1 in.	

4-element Yagi (W0OKC)
Boom length: 36 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	R—DE: 14 ft
Driven Element	33 ft 4 in.	DE—D1: 10 ft 8 in.
Director 1	31 ft 8-1/2 in.	D1—D2: 11 ft 4 in.
Director 2	31 ft 1 in.	

4-element Yagi
Boom length: 40 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	R—DE: 14 ft
Driven Element	33 ft 4 in.	DE—D1: 12 ft
Director 1	31 ft 7 in.	D1—D2: 14 ft
Director 2	31 ft 0 in.	

5-element Yagi (WA6ZZK)
Boom length: 40 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	34 ft 10 in.	R—DE: 10 ft
Driven Element	33 ft 4 in.	DE—D1: 9 ft
Director 1	32 ft 6 in.	D1—D2: 10 ft 6 in.
Director 2	32 ft 2 in.	D2—D3: 10 ft 6 in.
Director 3	31 ft 10 in.	

5-element Yagi
Boom length: 46 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 7-3/4 in.	
Driven Element	34 ft 5-3/8 in.	(All elements equally spaced at 11 ft 6 in.)
Director 1	33 ft 2 in.	
Director 2	31 ft 9-1/2 in.	
Director 3	30 ft 4-1/2 in.	

1 ft = 0.3048 m

In conclusion, it might be stated that one type of antenna (quad or Yagi) is not necessarily better than the other. The choice between a quad or Yagi array will depend critically upon many factors. For example, if single-band operation is of main importance and the geographical area where the antenna is to be used does not have high-wind problems, the best and simplest choice would be a long-boom Yagi antenna. If multiband operation is desired, or if wind problems were to limit the length of the boom to some nominal value, then the quad antenna would be a good choice. Likewise, the choice of available materials in your locale, and the acceptance of the appearance of the antenna by your neighbors or family, constitute other determining factors. The best approach is to survey your requirements and

5-element Yagi (K6EVR)
Boom length: 46 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 4 in.	R—DE: 12 ft
Driven Element	33 ft 5 in.	DE—D1: 10 ft
Director 1	31 ft 8 in.	D1—D1: 12 ft
Director 2	31 ft 8 in.	D2—D3: 12 ft
Director 3	31 ft 8 in.	

5-element Yagi (W8PWH)
Boom length: 52 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 11 in.	R—DE: 13 ft 10 in.
Driven Element	33 ft 4-1/2 in.	DE—D1: 10 ft 5 in.
Director 1	32 ft 10 in.	D1—D2: 13 ft 10 in.
Director 2	31 ft 2 in.	D2—D3: 13 ft 10 in.
Director 3	30 ft 9 in.	

5-element Yagi (W4EX, formerly W4DQH)
Boom length: 54 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 0 in.	R—DE: 13 ft 6 in.
Driven Element	33 ft 4 in.	DE—D1: 11 ft 6 in.
Director 1	32 ft 3 in.	D1—D2: 13 ft
Director 2	32 ft 1 in.	D2—D3: 16 ft
Director 3	31 ft 11 in.	

5-element Yagi (W6RR, formerly W6ITA)
Boom length: 55 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	34 ft 9 in.	R—DE: 16 ft 9 in.
Driven Element	33 ft 7 in.	
Director 1	31 ft 11 in.	(All elements equally spaced)
Director 2	31 ft 9-1/2 in.	
Director 3	31 ft 8 in.	

6-element Yagi
Boom length: 46 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 5 in.	
Driven Element	34 ft 5-1/2 in.	(All elements equally spaced at 9 ft 2-3/8 in.)
Director 1	32 ft 7 in.	
Director 2	32 ft 3 in.	
Director 3	31 ft 4-1/2 in.	
Director 4	31 ft 1/2 in.	

6-element Yagi (W0OKC)
Boom length: 46 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 4 in.	R—DE: 10 ft
Driven Element	33 ft 4 in.	DE—D1: 8 ft
Director 1	31 ft 11 in.	D1—D2: 9 ft
Director 2	31 ft 3-1/2 in.	D2—D3: 9 ft 6 in.
Director 3	31 ft 1/2 in.	D3—D4: 9 ft 6 in.
Director 4	30 ft 11 in.	

6-element Yagi (W0OKC)
Boom length: 74 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 5 in.	R—DE: 16 ft 6 in.
Driven Element	33 ft 4 in.	DE—D1: 9 ft 9 in.
Director 1	31 ft 9-1/2 in.	D1—D2: 14 ft
Director 2	31 ft 2-1/2 in.	D2—D3: 17 ft 6 in.
Director 3	30 ft 11 in.	D3—D4: 21 ft
Director 4	30 ft 8 in.	

7-element Yagi (W6HAW)
Boom length: 80 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 5 in.	R—DE: 14 ft
Driven Element	33 ft 4 in.	DE—D1: 11 ft
Director 1	31 ft 8 in.	D1—D2: 12 ft
Director 2	31 ft 4 in.	D2—D3: 13 ft
Director 3	30 ft 8 in.	D3—D4: 14 ft
Director 4	30 ft 8 in.	D4—D5: 16 ft
Director 5	30 ft 8 in.	

8-element Yagi (WA6EKD)
Boom length: 100 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	34 ft 9 in.	(All elements equally spaced at 14 ft 4 in.)
Driven Element	33 ft 5 in.	
Director 1	32 ft 2 in.	
Director 2	31 ft 11-3/4 in.	
Director 3	31 ft 9-7/8 in.	
Director 4	31 ft 8 in.	
Director 5	31 ft 2 in.	
Director 6	30 ft 11-3/4 in.	

environmental restrictions, make your choice and proceed accordingly.

Acknowledgements

It is very difficult to acknowledge systematically and completely all those who have made the quad antenna a successful communications antenna. However, surely credit must be given to C. C. Moore, the father of the quad antenna, to William Orr, who presented the first written material on the theory and construction of 2-element quads; to Lee Bergren, who showed the feasibility and advantages of using multielement quad arrays; to the Japanese workers, Adache and Mushiake, of Tohoku University, who presented a detailed analysis and experimental verification of the electromagnetic properties of loop antennas of large circumference; and finally, to all the radio amateurs, in all corners of the world, who developed constructional and design techniques and proved or disproved their validity.

The author would like also to acknowledge the help of Mr. Hal Wolff (WA0IOR) for his aid in constructing the model Yagi and quad arrays and running the patterns. These patterns were obtained by using the antenna range of the Denver Research Institute of the University of Denver.[6] Many Denver-area amateurs have entered into these quad experiments, notably Messrs. Jim Snyder (K0ZCM), Butch Ford (WA0IMX), Ed Wood (K0KKU), Keith Farris (W0YDM), Warren Wheeler (W0R-EQ) and many others. The author gratefully acknowledges their assistance, help and encouragement. This material originally presented in *QST* by J. E. Lindsay, Jr., W7ZQ.

References

[1] For a colorful description of C. C. Moore's invention of the cubical quad antenna see the introduction of W. I. Orr's *Quad Antennas*, Radio Publications, Inc.

[2] An antenna which radiates equally well in all directions is called an isotropic radiator or isotropic antenna.

[3] Adachi and Mushiake, "Theoretical formulation for large circular loop antennas by integral equation method," *Repts. of the Research Institute of Elec. Commun.*, Tohoku University, Sendai, Japan, Sci. Rept. Ritu, B(Elec. Commun.), vol. 9, no. 1, 1957.

Idem., "Studies of large circular loop antennas," *Repts. of the Research Institute of Elec. Commun.*, Tohoku University, Sci. Rept. Ritu, B(Elec. Commun.), vol. 9, September 1952.

[4] Element spacing of 0.14λ to 0.2λ being considered.

[5] Bergren, "Multielement Quads," *QST*, May 1963.

[6] At the time these tests were performed (1962-64), the author was head of the Antenna Laboratory, Denver Research Institute, University of Denver. Mr. Hal Wolff was the research technician at this facility.

Additional References

Jasik, *Antenna Engineering Handbook*. New York: McGraw-Hill, 1961, pp. 16-25.

Lindsay, "A circular loop antenna with non-uniform current distribution," *IRE Trans. Antennas and Propagation (Communications)*, vol. AP-8, pp. 439-441, July 1960.

Lindsay, "A parasitic end-fire array of circular loop elements," *IEEE Trans. Antennas and Propagation (Communications)*, vol. AP-15, No. 5, pp. 697-698, Sept. 1967.

A Triband One-Loop Cubical Quad Element

Why not trade gain for economy and structural reliability? This three-band "DX Getter" consists of a single quad loop. It will stay aloft in high winds and won't look like an ugly monster to those who dwell nearby.

The advantages of a rotatable beam antenna are well known to all DXers. A problem arises in many situations regarding just how a reasonable compromise between antenna performance and cost, size and weight can be obtained. Because this author likes to live in peace with the XYL, the neighbors and the local council, the antenna must not be considered dangerous or an eyesore as far as appearance is concerned. In addition, the antenna must not hang over the fence of the 50- × 150-foot (15- × 46-m) block, half of which is already occupied by buildings. With these limitations firmly implanted in mind, the author added such factors as cost and difficulty of erecting and supporting a full-size, 20-meter Yagi or cubical quad. After weighing all the facts, it was felt that the only way out was to construct an economy-style, reduced-size antenna.

A single-loop, cubical-quad triband element is shown in Fig. 10. Such a cubical quad has only one-fourth the weight and wind resistance of a full-sized, three-element cubical quad of the same mechanical strength. The mechanical and installation difficulties are many times smaller; the cost of mast and rotor are also considerably reduced. Yes, the antenna is a compromise and the gain is less than that of a 3-element quad, but this antenna still puts a signal in good DX company. This quad, because of its smaller physical size, will be tolerated in many more locations than its big brother. If desired, this quad could be used as an indoor antenna strung between two opposite room walls.

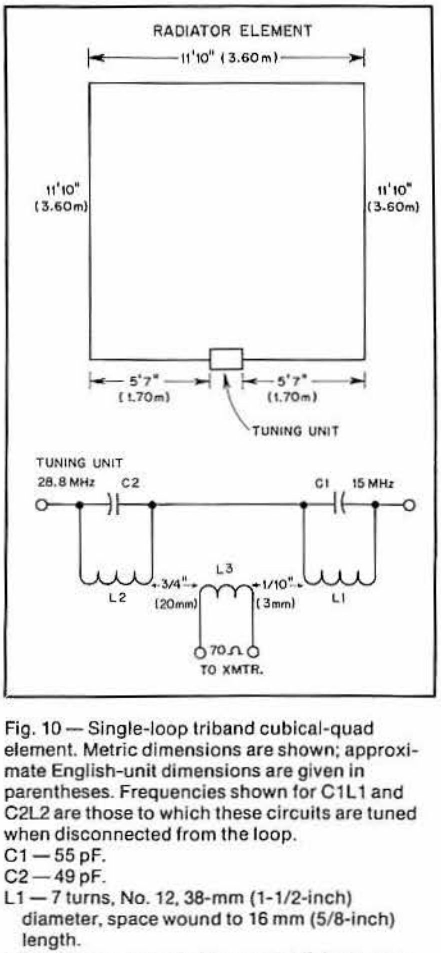

Fig. 10 — Single-loop triband cubical-quad element. Metric dimensions are shown; approximate English-unit dimensions are given in parentheses. Frequencies shown for C1L1 and C2L2 are those to which these circuits are tuned when disconnected from the loop.

C1 — 55 pF.

C2 — 49 pF.

L1 — 7 turns, No. 12, 38-mm (1-1/2-inch) diameter, space wound to 16 mm (5/8-inch) length.

L2 — 4 turns, No. 12, 38-mm (1-1/2-inch) diameter, space wound to 18 mm (11/16-inch) length.

L3 — 4 turns, No. 12, 38-mm (1-1/2-inch) diameter, space wound to 10-mm (3/8-inch) length.

Rather than going to a complicated switching arrangement to change bands, it was decided to apply what the author called the "triband-antenna principle" to this antenna. The theory behind this principle evolves around the same general ideas outlined by Pichitino in *QST*[1] several years ago. Antennas constructed along these guidelines do not operate harmonically and have only the desired resonances. The quad that was built is shortened only at the lowest operating frequency. It is full size at the middle operating frequency and much longer at the highest operating frequency. In addition, no heavy blocking tuned circuits are used near the element ends.

With C1 giving 20-meter resonance depends mainly on the total element length and L1. Fifteen-meter resonance is controlled principally by C1 and L2. Ten-meter resonance is determined by C2 and L2. Once the antenna is constructed, it may be fine-tuned on each band by varying the constants that control the resonant frequencies. After the quad was tuned up, the author found the SWR on each band to be very satisfactory. On 20 meters, the antenna resonated at 14.15 MHz with a resulting SWR of approximately 1:1. With a 15-meter resonance at 21.3 MHz, an SWR of 1.3:1 was measured. On 10 meters, 28.6 MHz was selected as the resonant frequency. The resulting SWR at resonance on this band was 1.5:1. The phase relationship of the fields in L1 and L2 varies from band to band so that too much direct coupling between these coils should be avoided.

Although the author used copper wire to make the loop element, you may desire to use aluminum tubing. The actual constructional details are left to the builder. The materials used and the methods of supporting the antenna are limited only by your imagination. The techniques employed may be as simple or as sophisticated as you may desire. DJ2UT used short pieces of coaxial cable as capacitors and placed them inside the tubing which he used for the element. The coils and inner capacitor ends were sealed and molded in resin. Fifteen- and 10-meter resonances are obtained by shifting copper rings over the ends of L1 and L2. This can be done without affecting the sealed coils in any way.

The author concluded this antenna to be rather efficient based on observations that the tuning network does not become warm and a low SWR has been measured across each of the three bands. This writer wishes to thank OM Sommer, DJ2UT, and his co-workers for the very considerable amount of work carried out, and the many practical ideas which have made successful antennas with the triband principle possible. This material was originally presented in *QST* by Hans F. Ruckert, VK2AOU.

---

[1] Pichitino, "A New Principle in Two-Band Rotary-Beam Design," *QST*, October 1948.

A 2-Element, 15-Meter Quad for the Novice

Why not build this 2-element quad for 15 meters? It is easy to construct, inexpensive and in many ways quite novel.

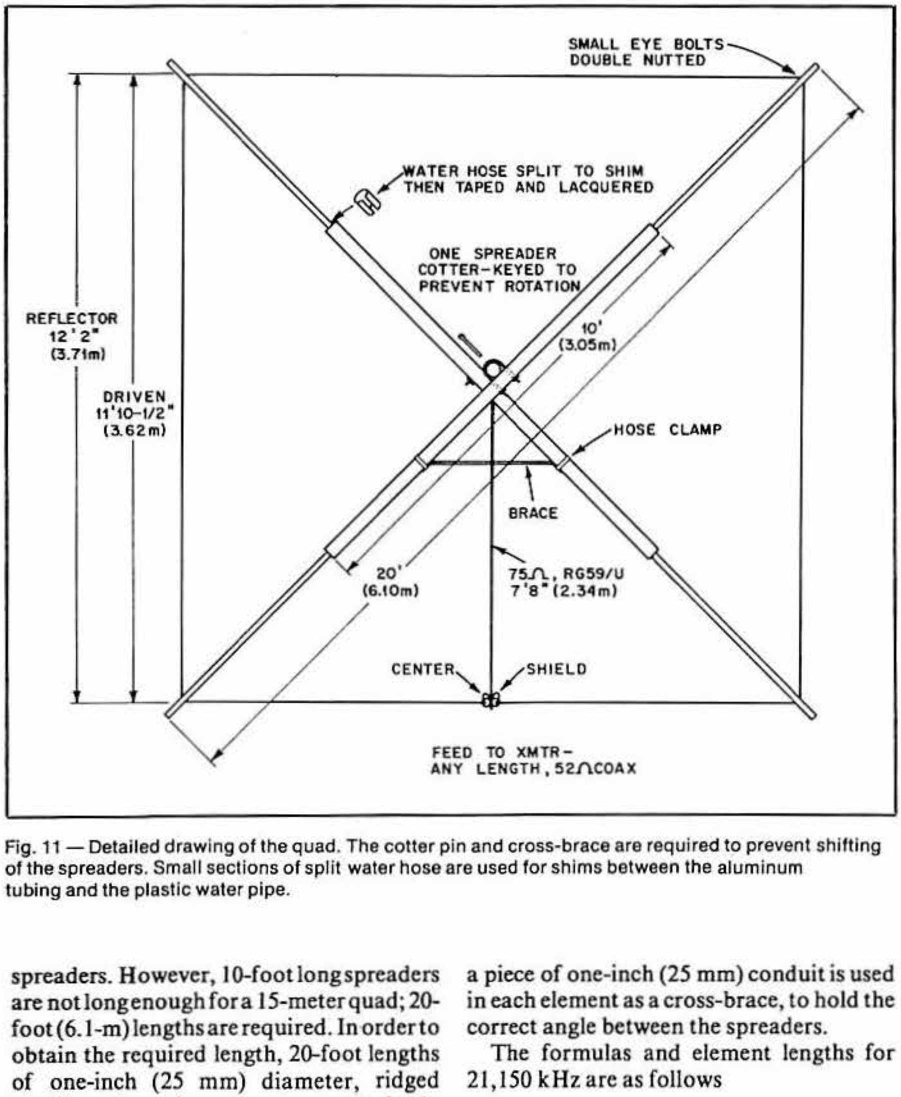
SMALL EYE BOLTS DOUBLE NUTTED
WATER HOSE SPLIT TO SHIM THEN TAPED AND LACQUERED
ONE SPREADER COTTER-KEYED TO PREVENT ROTATION
REFLECTOR 12'2" (3.71m)
DRIVEN 11'10-1/2" (3.62m)
10' (3.05m)
HOSE CLAMP
BRACE
20' (6.10m)
75Ω, RG59/U 7'8" (2.34m)
CENTER
SHIELD
FEED TO XMTR - ANY LENGTH, 52Ω COAX

Fig. 11 — Detailed drawing of the quad. The cotter pin and cross-brace are required to prevent shifting of the spreaders. Small sections of split water hose are used for shims between the aluminum tubing and the plastic water pipe.

Not being blessed with a large area for antennas, the author decided on an optimum-spaced, 2-element monoband quad. The standard formulas were used for sides and spacing, and the antenna is fed with 50-ohm cable through a quarter-wave "Q" section of 75-ohm coax cut to the standard formula. The reflector has a tuning stub.

The driven element grid-dips at 21,150 kHz and the reflector was tuned by using the grid-dip oscillator as a signal generator. The grid-dip meter was placed in a neighbor's home. Then, the back of the quad was aimed at the dip meter, and the reflector was tuned for the lowest indication on the S-meter.

One of the checks made on the antenna was with VE5XH, in Swift Current, SK. Doug gave the following information: The quad has 2 S-units gain over my 1/4-wave groundplane which was at the same height. Strength off the front was S9 and to the rear was S2, indicating a good front-to-back ratio. Because the constructional details are rather novel, the author felt that other hams might be interested in how this antenna was built.

Construction

The author used 1-1/4-inch (32-mm) TV steel mast pipe, 10 feet (3 m) long, for the boom. Also, 10-foot lengths of TV mast sections, but aluminum, are used for the spreaders. However, 10-foot long spreaders are not long enough for a 15-meter quad; 20-foot (6.1-m) lengths are required. In order to obtain the required length, 20-foot lengths of one-inch (25 mm) diameter, ridged plastic water pipe were run completely through the 10-foot aluminum pipe sections. Pieces of rubber hose were used as shims between the aluminum pipe and the plastic pipe and these were taped where the plastic pipe emerges from the aluminum.

The plastic water pipe comes in 20-foot lengths, 80 feet (24.4 m) being required. This pipe was purchased from a local plumbing and lighting outlet; the cost was $10 for the 80 feet. It is ridged plastic, but is not strong enough to support itself; so the aluminum mast pipe is used for the center portion of the spreaders.

The spreaders are clamped to the boom with standard U bolts, and one spreader in each element is pinned to the boom with a 5/16-inch (8 mm) cotter key. Additionally, a piece of one-inch (25 mm) conduit is used in each element as a cross-brace, to hold the correct angle between the spreaders.

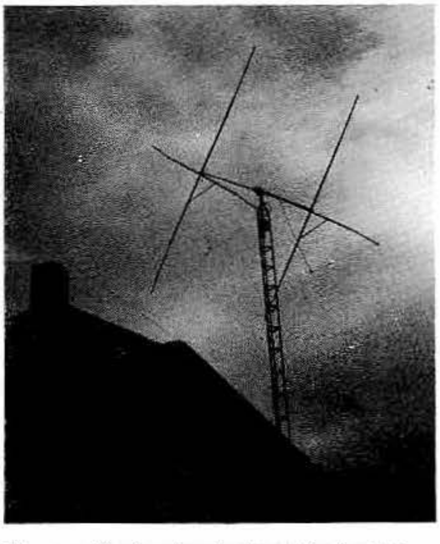

The completed quad on the tower. For tune-up, the quad was temporarily mounted on the garage roof.

The formulas and element lengths for 21,150 kHz are as follows

Driven element

$$\frac{1005}{f(\mathrm{MHz})} = 47 \text{ feet, 6 inches (14.48 m)}$$

Reflector

$$\frac{1030}{f(\mathrm{MHz})} = 48 \text{ feet, 8 inches (14.83 m)}$$

Boom spacing

$$0.2 \text{ wavelength} = 0.2 \times \frac{984}{f(\mathrm{MHz})} = 9 \text{ feet, 3 inches (2.84 m)}$$

75-ohm matching section

$$\frac{246 \times V}{f(\mathrm{MHz})} = \frac{162.4}{f(\mathrm{MHz})} = 7 \text{ feet, 8 inches (2.34 m)}$$

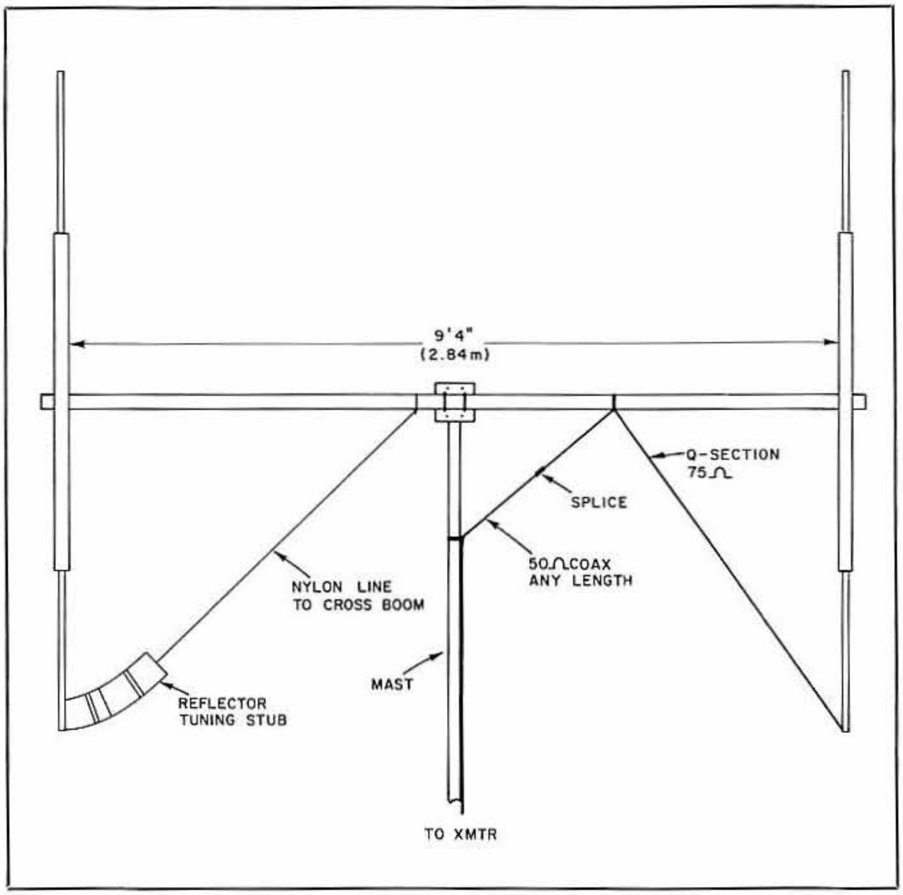

Fig. 12 — Side view, showing the mounting plate, Q matching section, and coax feed.

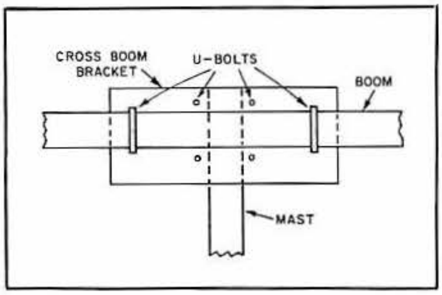

Fig. 13 — The cross-boom bracket, made from 3/8-inch (10 mm) aluminum plate. Ordinary TV U bolts are used to secure it to the boom and mast.

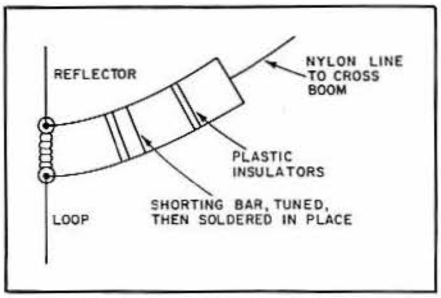

Fig. 14 — The reflector tuning stub is 36 inches (0.91 m) long. The shorting bar, a short piece of wire the same size as used in the stub was approximately 10 inches (254 mm) from the insulator when adjusted for maximum front-to-back ratio. Spacing between wires of the stub may be of the order of a few inches; this spacing is not critical since the stub length is adjusted by the shorting stub in the tune-up procedure.

The author used no. 14 soft-drawn enameled wire for loops and they were secured to the plastic pipe with double-nutted eye bolts. The reflector tuning stub was made from no. 12 copperweld wire. The total length of my stub is 36 inches (0.91 m), and after tuning, the shorting bar ended up approximately 10 inches (254 mm) from the loop.

The SWR showed almost a perfect match across the entire band. Since the quad was put up, the author has worked all continents and 35 countries on the Novice band. Stateside contacts always run S8 to S9. The total cost of the antenna was $28 at the time the system was built. This material was originally presented in *QST* by John Daebelliehn, WA9BJC.

A Practical 40-Meter Quad

One of the most perplexing problems continuously plaguing Field Day committees is how to install a good antenna system. For most clubs, finding a good location is generally the easiest job. The location must then be surveyed by the club antenna specialists. Deciding how many transmitters to use depends mostly on how many indivudual antennas can be set up for simultaneous operation as well as how many operators are available for the 24-hour stretch. Of course, an abundance of equipment is desirable.

Propagation conditions for Field Day and Sweepstakes generally favor the lower bands since they usually remain open all night. Success on the higher frequency bands (10 and 15 meters) depends primarily on catching the short-duration openings. The old trick of tying a rock to the end of a long piece of wire, tossing it into the tallest tree and loading it with a coupler just won't "cut the mustard" anymore. Gain and directivity into the dense population areas is the goal. However, the antenna must be simple to construct, easy to tune, and moderately stable against the wrath of Murphy and Mother Nature.

Our club decided to try a 2-element full-sized cubical quad for 40 meters. The immediate reaction within the more conservative ranks of our club was, "A 40-meter quad? It'll never work — the first breeze will knock it down!" With these objections in mind, we went ahead with our plans. Our location was approximately rectangular in shape, 60 feet wide, 500 feet long (18 × 152 m) and above the rest of the terrain. Of course, it just happened to be one of the highest locations in Winnebago County, IL!

The sketch illustrates the basic construction of the quad. Dimensions were estimated using the formula

$$L(\text{feet}) = \frac{248}{f(\text{MHz})}$$

where L is the length of one side of the diamond.

The parasitic element was constructed as a reflector using an extra long stub. A small relay shorts out the stub, thus converting the parasitic element from a reflector to a director (see Fig. 15).

The resonant frequency of the driven element is 7150 kHz, the middle of the band. The distance L for the driven element is 34 feet 8 inches (10.57 m). The parasitic element dimension is 5 percent smaller than the driven element. The stub is constructed so that the total circumference of the parasitic element with stub in the line is 5 percent greater than the driven element. L for the

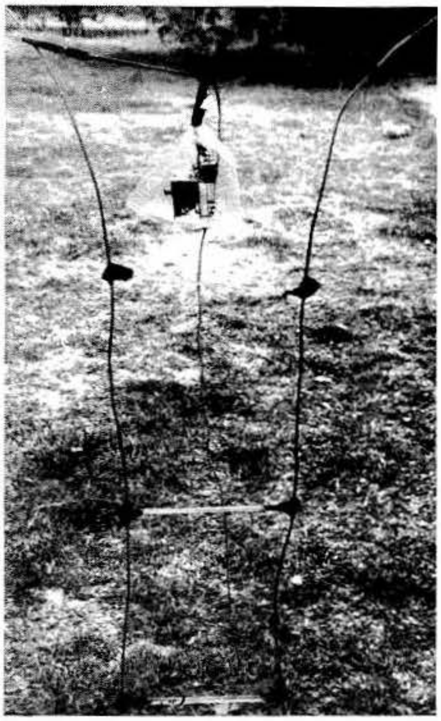

Fig. 15 — A simple relay arrangement is used to short out the stub and change the pattern of the antenna from one direction to another. A small plastic bag provides weatherproofing in the event it rains. (Doesn't it always?)

director is 33 feet (10.06 m). The stub length is 6 feet 9 inches (2.06 m). Eight-inch (203-mm) plastic spacers are placed between the two wires of the stub.

Supporting the quad is very simple. The mast is 50 feet (15.24 m) high, and guyed with nylon ropes. The boom is constructed from two telescoping sections of electrical conduit and is 18 feet (5.49 m) long. The elements are made from no. 14 stranded wire. The driven element is fed directly with RG-58/U coaxial cable. Location of the stub short is determined by the point of maximum front-to-back ratio. The boom-to-mast construction utilizes the support base from a Ham-M rotator. All it takes is a couple of U bolts and the boom-to-mast assembly is complete (see Fig. 16). The wire is strung from the end of the boom and is supported by a small piece of nylon rope and an egg insulator.

Results

Orientation for the quad was east and west. Favoring the eastern direction was the driven element/reflector combination, whereas the driven element/director combination was broadside to the western direction. A local amateur, approximately 5 miles (8 km) west of the Field Day site, assisted in tuning the quad for maximum front-to-back ratio. The end result was approximately 25 dB.

Results were a tremendous surprise! The antenna worked just as planned. The class of entry for simultaneous operation of two transmitters was used. One transmitter operated on 20 and 80 meters; the second transmitter was set up for 15 and 40 meters. Our club, W9AXD/9, scored third place in the two-transmitter class. Our biggest totals came from 40 meters — 693 contacts, over one third of our total number. For the first four hours of the contest, the antenna was aimed east. A rate of 90 contacts per hour prevailed for five solid hours. From that point on it was a steady decline after a sharp drop to about 45 contacts per hour. It was later surmised that the reason for the drop in contact rate was lack of activity and difficulty in finding stations that hadn't been worked before. The East Coast QRM was drastically reduced by changing the switch from "east" to "west." By midmorning we were surprised to find only a dozen W5 stations listed in the log. The antenna may not have been too effective in the southern direction.

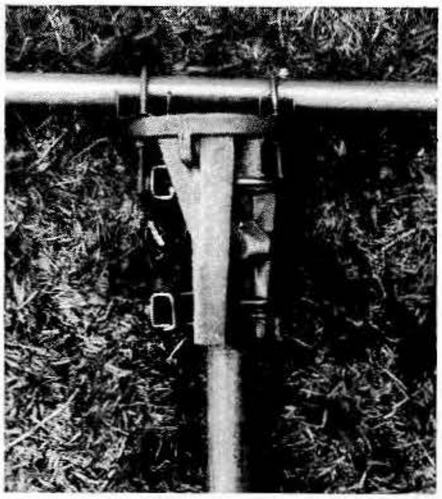

Fig. 16 — The bottom section of a CDE Ham-M rotor can be adapted to serve as a boom-to-mast bracket.

There is no reason why this arrangement could not be used for DX operating. The antenna could be positioned just underneath a beam for the higher frequency bands with the peak of the diamond supported from a boom placed through the top section of the tower.

This material was originally presented in *QST* by Peter H. Grillo, W6RTT.

Build This C-T Quad Beam for Reduced Size

This quad-beam antenna, while smaller in size than a conventional quad, performs like a veteran. The trade secret is to use capacitive tuning.

One of the main drawbacks of the quad beam is its large size compared to the Yagi beam. Previous efforts to reduce the size of the quad have utilized systems of inductive tuning. Loading coils have been used on one or two sides[1] or in the corners. Linear inductors on the vertical sides have also been used.[2] The purpose of this article is to explain a simple method of capacitive tuning to reduce the size of the quad. The C-T (capacitor-tuned) loop is diagrammed in Fig. 17.

It is well-known that the physical size of the dipole antenna can be reduced by replacing a portion of the antenna with a capacitive body and adding an inductor to restore the system to resonance. It seems less well-known that the full-wave loop antenna also may be reduced in physical size while holding the resonant frequency of the system constant by adding capacitance at the voltage loops. No inductors are needed. The capacitance may be added by any of the methods shown in Fig. 18.

The outside dimensions of the C-T loop may be anything chosen by the builder with the larger limit being the full-size, full-wave loop. The limit on the amount of size reduction effectively attainable depends both upon the electrical resistance in the circuit and upon the tendency in loop antennas for the current to become equal at all points in the circuit when the loop becomes very small in proportion to the free-space wavelength at the frequency under consideration. The geometrical shape of the loop is unimportant to basic operation provided that the capacitance is placed at or between the voltage loops. The bottom half of the chart in Fig. 21 shows the effective dimensional range of the C-T loop. The top half shows the tuning range that can

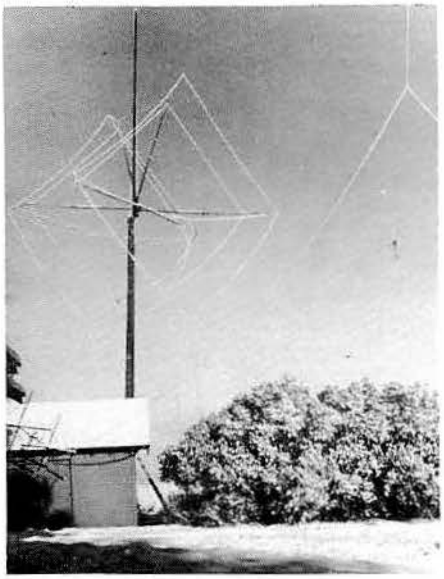

A winter snow sets the scene and accentuates the diamonds of the C-T quad-beam antenna at W7WKB.

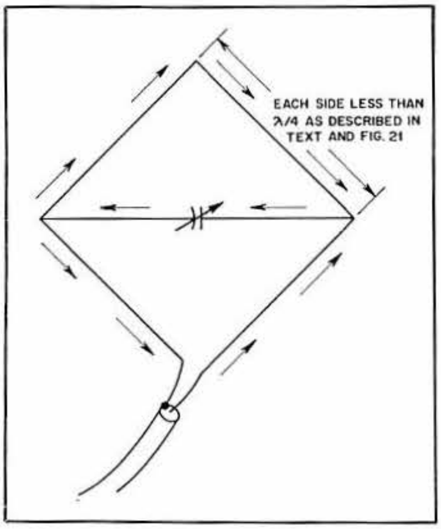

Fig. 17 — The C-T quad design with arrows indicating current flow for 1/2 cycle. The center capacitor is located at the minimum-current (highest-voltage) point in the loop. Each side is less than 1/4-wavelength long. See text and Fig. 21.

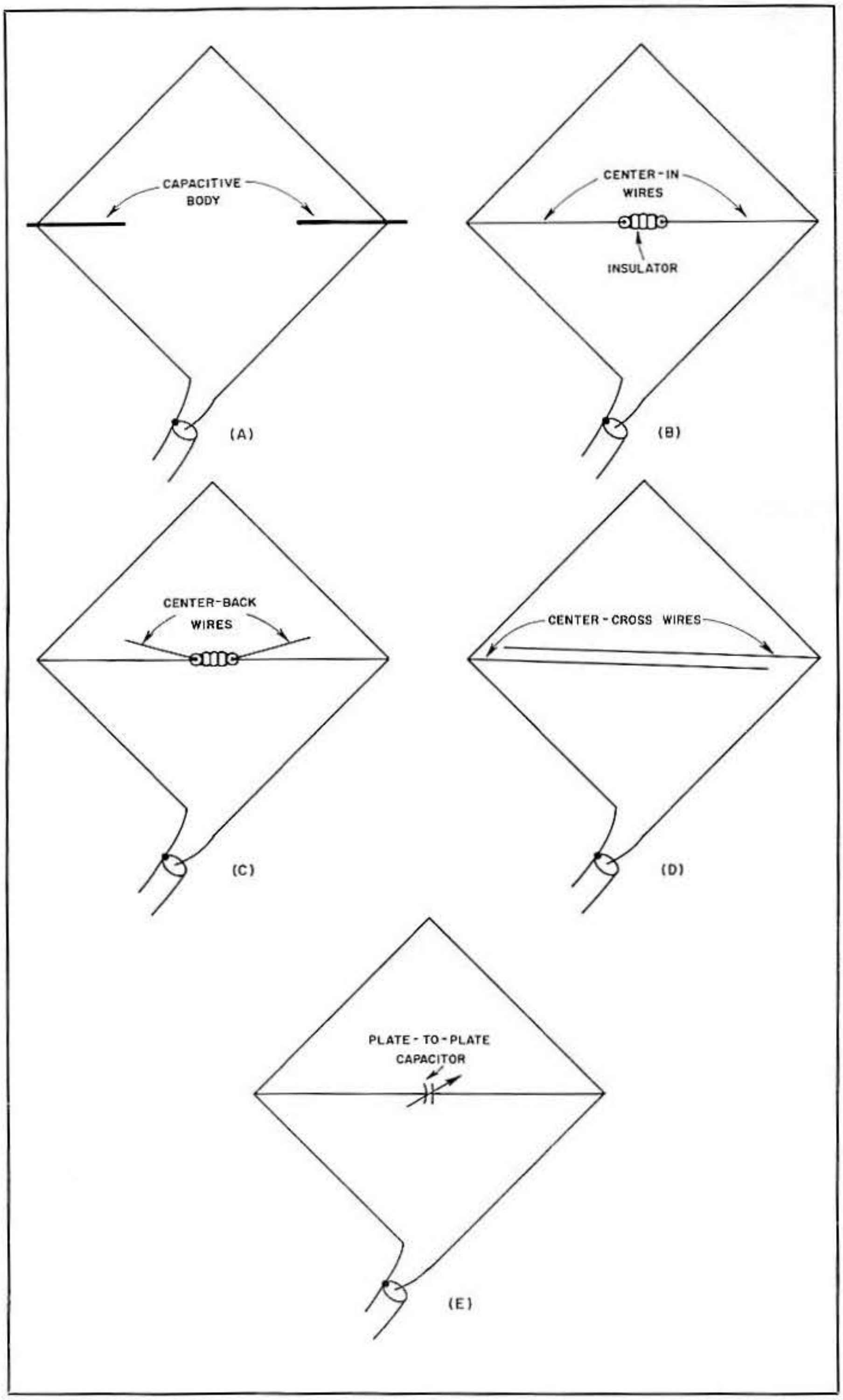

Fig. 18 — Methods of adding capacitance to the C-T loop are shown here. In Fig. 18C, if the center-back wires are placed too closely to the center-in wires, they become ineffective. A 12-inch (0.3048 m) spacing from the end of the center-back wire to the center-in wire is satisfactory. Wires running back toward the loop center may be attached to the outside ends of the center-back wires if more capacitance is needed. In Fig. 18D, the capacitance added by the center cross wires varies with both the length of wire and spacing between wires.

be expected by using various methods of adding capacitance to the diamond quad. These dimensions are intended as guidelines only.

The feedpoint resistance of the full-size, 1-wavelength loop is about 100 ohms. The feedpoint resistance of the C-T loop decreases as it is made smaller but will offer a good match to a 50-ohm coaxial line over the entire usable range as shown in the chart. A gamma match may be used if an exact match is desired.

More than one C-T loop may be used to form the C-T quad beam. A parasitic reflector was used in this installation because of the higher radiation resistance attainable while maintaining good gain. Maintaining the radiation resistance as high as possible is particularly important in small beams, both to keep efficiency high and to prevent the bandwidth from becoming too narrow.

Tuning and Mounting

The use of a grid-dip meter to attain initial band resonance is very helpful. By using a grid-dip oscillator, this beam was tuned initially with the driven element to resonate at the center of the band of interest. The reflector was then tuned to resonate at a frequency three percent lower than the driven element. Final adjustments were made using a test oscillator located about 1000 feet (300 m) away as a reference signal.

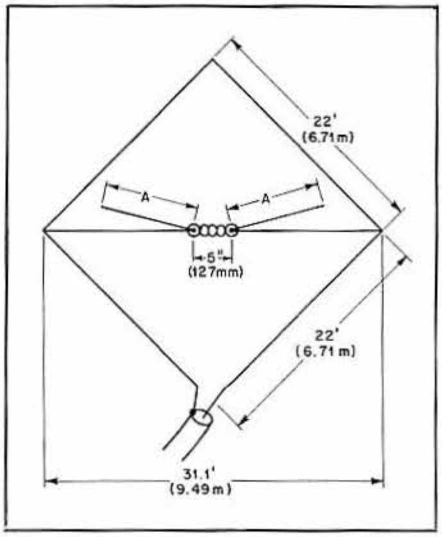

Fig. 19 — Dimensions for the 40-meter C-T quad are shown above. For the 2-element beam the spacing between the driven element and the reflector is 20 feet (6.10 m). The length of the center-back wire (dimension A) is 12 feet, 9 inches (3.89 m) for the reflector, and 9 feet, 11 inches (3.02 m) for the driven element.

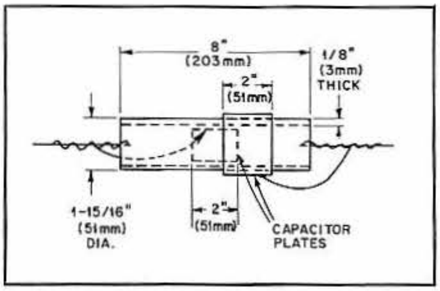

Fig. 20 — A variable capacitor is illustrated here. It can be used to tune the C-T quad. The eight-inch (203-mm) tubular portion is made from plastic water pipe. The cylindrical capacitor plates are formed from a copper sheet. One plate slides inside the eight-inch (203-mm) tube and the other over the outside of the tube. Maximum capacitance is approximately 25 pF.

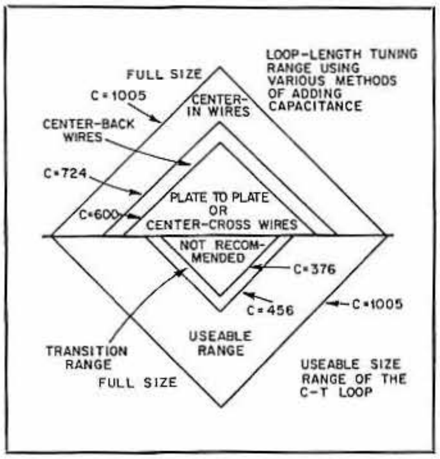

Fig. 21 — The chart shown above, providing dimensional guidelines for a single C-T loop, is based on the equation LF = C, where L is the outside dimension in feet, F is the frequency in MHz, and C is a constant chosen for the size of the C-T loop to be constructed. Use the largest size possible for the best electrical results. This is particularly true when using C-T loop elements to form a parasitic beam.
To use the chart, pick the largest C (up to C = 1005) feasible for your application. Use bottom half of the chart to see if the C-T loop will operate at the dimensional size selected. Then refer to the top half of the chart for the recommended capacitance method.

Table 3

Data for plotting SWR curve. Information in this table is used for plotting the SWR curve for the 40-meter C-T quad beam, as described in the text.

Frequency (MHz)	*SWR*
7.500	4.87
7.450	4.00
7.400	3.08
7.350	2.33
7.300	1.82
7.275	1.78
7.250	1.94
7.200	3.44
7.150	5.25
7.100	8.08

The director was tuned for lowest SWR and the reflector for best gain. If you lack a grid-dipper, initial resonance can be established by listening to your station receiver for a strong signal while making the adjustments. A dramatic increase in signal strength will be noticed at resonance. Tuning is accomplished by shortening or lengthening the capacitance wires.

The single loop or beam may be mounted either horizontally or vertically for horizontal or vertical polarization, but in either case it should be mounted as high above the ground as possible. Loop antennas operating close to ground will show a much higher feedpoint resistance than normal and will waste considerable power in groundheating effects. A rough guideline is to mount the antenna so that the bottom point is at least half the loop diameter above the ground.

The working 40-meter version of the C-T quad beam had a diamond configuration because of the greater mechanical strength and because of the longer center wire for tuning purposes. Final dimensions are shown in Fig. 19. A good quality center insulator should be used because of the high voltages present at the center.

A tubular tuning capacitor was originally installed in the driven element and worked well until the kW power level was reached. It then promptly arced over, leaving a conductive carbon deposit. A version using 2-inch (51-mm) plates as shown in Fig. 20 adequately tuned the reflector even at the kW level for some time. The capacitors, as shown, were not tried under wet conditions, but would probably be satisfactory at power levels up to 400 watts. For long-term use, the method shown in Fig. 18D is preferred over plate-to-plate capacitors because the high voltages are handled more easily.

Because some dielectric heating was noticed in this capacitor, the wire-tuned version was adopted. The wire version is recommended over the methods shown in Fig. 18D and 18E, any time the loop is large enough to be tuned by no more than one doubled-back wire (as shown in Fig. 18C). Tuning wires eliminate the capacitor with its heat loss and allow somewhat better current distribution. Less center radiation occurs because of partial canceling effects in the doubled-back wires. It would be wise initially to cut each center-back wire 12 inches (0.31 m) longer than shown to allow tuning. Pruning each center-back wire by 1 foot (304.8 m) increases the resonant frequency by roughly 100 kHz in the 40-meter version. No. 14 Copperweld wire was used throughout.

Performance

The antenna described performed up to expectations and has prompted many complimentary reports. The SWR curve information shown in Table 3 was attained using a 75-ohm coaxial feed line without any matching network. The SWR would be lower if a 50-ohm line or if a matching network were used. Forward signal strength as compared to the strength to the rear showed a maximum 30-dB difference as measured by means of the receiver S meter. Typical difference, as finally adjusted, was 12 dB.

The gain of the C-T quad over a comparable double Zepp (full wave on 40) was typically at least one S unit better when both antennas were at a height of 45 feet (13.7 m). The Zepp was better on very short skip but the C-T quad consistently was favored on longer skip. When elevated to 65 feet (19.8 m), the C-T quad gave outstanding results. At this height, two S-units improvement over a vertical antenna at distant points was noted many times.

The C-T quad may be built for any band. It should be most useful in meeting requirements for an inexpensive easily tuned wire-beam antenna where the full-size quad is unsuitable for mechanical or economic reasons. This material was orginally presented in *QST* by Roger Sparks, W7WKB.

References

[1] Pinner, "The Short Quad," *QST*, February 1964.
[2] Courtier-Dulton, "Some Notes on a 7-MHz Linear-Loaded Quad." See Chapter 4.

A Nearly Full-Size, Rotatable, 2-Element Quad for 80 Meters

We don't expect readers to rush right out and duplicate this antenna system — but this doesn't mean a lot of hams wouldn't like to!

Many hams have been thinking about new 80-meter antennas to improve their DX capabilities. On the higher amateur frequencies, antenna gain is relatively easy to acquire with a compact Yagi or quad. It is much harder in this respect on 40 meters. On 80 meters, however, the problem of securing any increase in antenna gain over a conventional dipole or groundplane is very difficult indeed, particularly where space is limited. Rhombics, Vs or multi-element collinear arrays become completely impractical on the normal urban or suburban lot. Even a phased vertical array is hard to handle on a city lot, and too often, performance is marginal because it is impossible to install an optimum ground system at such a location.

The author solved this problem at his QTH by constructing an almost full size 2-element quad for 80 meters. A unique tuning arrangement permits this antenna to be operated at any frequency within the 80-meter band with an SWR of close to 1. While certainly not adaptable to every individual's situation or pocketbook, a description of this antenna should be of interest to many amateurs. As far as it is known, this is the first and only set of beams, on one rotatable tower, covering all bands from 2 to 80 meters. By connecting the 2 quad elements in series to form a rotatable, bidirectional loop, the frequency coverage has been extended to 160 meters.

At this point, a few comments might be in order on the circumstances which led to this sizeable 80-meter antenna project. In June 1965, the writer returned to ham radio activity after a QRT of almost 30 years. During the first year of operation, the antennas of K3JH were a conventional commercial tribander for 10, 15 and 20 meters mounted on a 60-foot (18.3-m) tilt-over tower, and a trapped, inverted V for 40 and 80 meters. A bit of DX chasing soon led to the conclusion that better antennas were needed. Since the QTH is on a suburban lot about 175-feet (53.3-m) square, and heavily wooded, a "Christmas tree" array seemed the best alternative. Consequently, in the summer of 1966, with a great deal of help from Bob Scully, W2FXN, a 115-foot (35-m) rotary tower was installed along with full-size monoband beams for all bands from 2 through 40 meters.

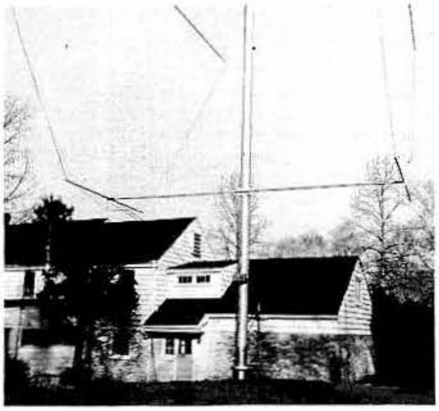

The bottom support of the quad elements is shown in this view, along with the tuning-network box, which is mounted on the mast.

This rotary steel tower is 16 inches (0.41 m) OD at the base with 1-inch (25-mm) thick walls. The tower rests in a 1/2-inch thick steel bearing-tube 15 feet (4.57 m) long and 20-inches (0.51-m) OD. The bearing tube is imbedded in a block of concrete 8 × 8 × 16 feet (2.44 × 2.44 × 4.88 m) which weighs 70 tons. The tower tapers to 5-1/2 inches (140 mm) at the top, and was designed to carry nine full-size monoband beams through 125-mph winds.

Initially, the antenna complement was as follows

2/6meters — Vert. groundplane at 115 feet (35 m)
6 " — 6 el at 113 feet (34.5 m)
2 " — 15 el at 109 feet (33.2 m)
20 " — 5 el at 104 feet (31.7 m)
15 " — 5 el at 96 feet (29.3 m)
10 " — 6 el at 86 feet (26.2 m)
40 " — 3 el at 77 feet (23.5 m)

A rotary ball-bearing ring and clamp at the 70-foot (2.13-m) level on the tower was used to support one end of an inverted L for 80 meters. This long wire, which could be used also on the higher frequencies, was fed through a Matchbox at the base.

Performance on 40 through 2 meters with this antenna system was generally excellent. The L and inverted V on 80, however, left much to be desired, especially when compared to the antenna performance on the other bands.

In attempts to improve this situation, several different 80-meter antennas were installed at various times between 1966 and the spring of 1969. Included in this effort were a top-loaded groundplane and a pair of phased verticals. None of these antennas provided sufficient improvement in DX performance on 80 to be considered satisfactory.

With the advent of the 5-Band DXCC, the question of how to do better on 80 meters again became a matter of concern. It was at this time that the concept of a quad of some sort, to be mounted on the rotary tower, began to emerge. Since 40-meter quads had been constructed previously, and were mechanically feasible, the thought at first was to build a half-size quad for 80 meters, using loading coils. Preliminary calculations indicated the possibility of achieving some gain over a dipole in addition to the obvious advantage of being able to rotate the array.

It was quickly determined that a half-size quad mounted at 57 feet (17.4 m) would fit underneath the 40-meter Yagi. Construction could be quite conventional using fiberglass X frames in each element, and a spacing of 0.15 wavelength.

Design Considerations

The proposed design was discussed with a number of amateurs including Jim Lindsay, W0HJ, Dunc Carter, W5IOU, and Claus Moeller, DL7CM, who were most helpful with advice and suggestions. A search of the literature disclosed that a number of loaded antennas of various types had been built. In most cases, however, performance had been judged empirically, and there was little in the way of specific comparative data on the performance of loaded versus unloaded configurations, or versus a reference dipole. Because of this, the decision was made to build and test a 14-MHz model of a half-size quad. This would enable a direct comparison between the performance of a miniaturized quad and other antennas. Of particular interest was a comparison with a dipole, since this was the more normal antenna used by amateurs on 80 meters.

Henry Pemberton, W3PN, who had become interested in the project, provided the X frames and supports from an old 20-meter quad, for use in the test model. Tom Consalvi, W3EOZ, provided some suitable coil stock for the test design. With this help, the 14-MHz model was quickly constructed. No trouble was experienced in pruning the coils and resonating the loops. Except for one bad piece of insulation on one coil, which promptly burst into flame when rf power was applied to the antenna, the driven element could be fed with a full kilowatt at an SWR of 1. The SWR, however, would rise sharply when the coils, which had no protection from the weather, became wet from rain. All testing, therefore, had to be done on dry days.

Since the 80-meter version was to be mounted on the tower one-quarter wavelength above ground, the 14-MHz model was mounted for the tests at the same relative height. Standard procedures for tuning up the quad were used, and will not be detailed here. Impedance of the loaded loop was measured at 60 ohms, so it could be driven nicely with a 50-ohm line. The bandwidth was 125 kHz, measured between frequencies each side of the resonant frequency where the SWR was 3:1. This was the result expected, due to the heavy loading.

Extensive on-the-air comparisons were made between the model and various other antennas. These included a dipole at the same height above ground as the model, a 3-element tribander at 60 feet (18.3 m), two different inverted Vs at 50 feet (15.2 m) and a 5-element monoband Yagi at 104 feet (31.7 m). The regular antenna switching arrangements in the station were such that almost instantaneous comparisons could be made between the different antennas, minimizing QSB differences in the receiver. In the transmit mode, many amateurs compared signal differences between the test model and the other antennas. It is impossible to list the calls of all those who helped in this way.

After considerable testing, adjustment, readjustment, and minor configuration changes in the model, the results were found to be quite discouraging. The simple conclusion was that a loaded half-size quad had insufficient gain compared to a dipole to warrant proceeding any further with an 80-meter version.

Further Exploration

Before abandoning the 80-meter project, however, we decided to explore some practical ways to increase the size of the elements, preferably without using the normal-type quad X frames, which would be difficult if not impossible to handle when the antenna grew too large. This thinking developed the concept of using the 14-MHz boom at 104 feet (31.7 m) to support diamond-shaped elements at the top, and using a boom and spreaders rather than an X frame to hold the elements at the center. The bottom of the quad elements could be supported easily with a relatively light boom and spreaders near the base of the tower.

In order to reduce the size of the center boom and spreaders, the first model was kite-shaped with an included angle at the top of about 50 degrees. Unfortunately, the results were poor, and compared to a dipole there was no gain. We concluded that this configuration was so squashed together that the antenna was acting like a dipole instead of a quad. A series of configurations was then drawn up on paper. The boom and spreader sizes were varied, but the included angle at the top of each diamond element was kept at 75 degrees or more. This helped to pin down a configuration which would be mechanically feasible, and which would be a reasonable compromise among the various considerations to be taken into account. The one selected is shown in Fig. 22.

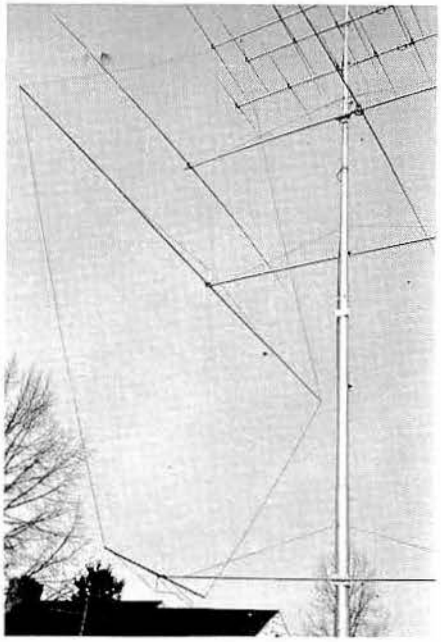

A detailed view of one of the quad elements, showing the method of supporting the loop.

Because these elements are almost full size at 4 MHz, making a 20-meter model quite large, it was decided to build and test a model of this configuration on 15 meters. The test procedures for this model were the same as those previously described for the 20-meter model of the half-size quad. This model showed substantial gain compared to the reference dipole at the same height. Also, the model mounted only 15 feet (4.6 m) above the ground compared favorably in performance with the tribander at 60 feet (18.3 m). The 5-element monobander at 96 feet (29.3 m) consistently provided better gain, as was to be expected. The impedance of the driven element was about 80 ohms. Bandwidth between SWR points of 3:1 was 250 kHz, or about twice what was measured on the half-size model.

Mechanical Considerations

After a lengthy test period, during which many on-the-air comparisons were made, the results were good enough to make the decision to build an 80-meter version of this antenna. Because of the size of the proposed antenna, considerable thought had to be given to the mechanical design to assure longevity comparable to the other beams which are rated for 125-mph winds. The wire in the quad elements is no. 12 stranded copperweld. All other metal in the antenna is aluminum or stainless steel. The two quad elements are suspended from the 20-meter boom at 104 feet (31.7 m). The insulators are five glazed porcelain knobs. Spacing is one-eighth wavelength, 36 feet (10.97 m). The 20-meter boom is 46 feet (14.02 m) long, and is made from 4-inch (102-mm) OD, 1/4-inch (6-mm) wall T-6 aluminum tubing at the center, and similar material tapering from 3-1/2 inches (89 -mm) to 3-inch (76-mm) OD at the ends. A 1/4-inch stainless steel cable supports the boom 18 feet (5.49 m) out from each side of the tower. Originally, the plan was to slide the quad elements down this cable from the tower. Unfortunately, the steel plates holding the Yagi elements to the boom on the other beams were found to be rusting because of poor plating. Rather than dismantle the beams, a crane was brought in so that these plates could be cleaned and painted by a man carried up in a boatswain's chair. At the same time, the elements were hung from the 20-meter beam, and the center boom and spreaders for the quad were installed with relative ease.

Dimensions

The 36-foot (10.97-m) center boom at the 57-foot (17.37-m) level of the tower is made of a single 24-foot (7.32-m) section of 3-1/2-inch (13-mm), 1/4-inch (6-mm) wall, T-6 tubing with 7-foot (2.13-m) sections of 3-inch (76-mm) OD pipe telescoped and bolted at each end. The boom is supported from the tower with a 3/16-inch (4.76-mm) stainless-steel cable. The spreaders at each end of this boom are 70 feet (21.34 m) long. Each consists of a 24-foot (7.32-m) piece of 2-1/2-inch (64-mm) OD, 1/8-inch (3-mm) wall, T-6 tubing at the center, with two short pieces of 2-1/8 × 1/8-inch (54 mm × 3-mm) tubing telescoped and bolted at each end to make up a length of 30 feet (9.14 m). To complete the spreaders, 20-foot (6.10-m) sections of 1-1/2- and 1-1/4-inch (38- and 32-mm) fiberglass poles are attached to each end of the aluminum centerpiece. Use of the fiberglass reduces weight and eliminates a one-quarter wavelength piece of metal from the middle of the quad element. A 1/8-inch (3-mm) cable and strut supports each spreader to minimize sag, which is very slight as may be noted in the photograph. A welded aluminum T structure and stainless steel clamps are used to hold the spreaders on to the ends of the boom. A stainless steel clamp and Teflon grommet is attached to the ends of each spreader to hold the element wires in place. The only function of the spreaders is to hold the two opposite sides of each quad element apart without too much fore and aft flopping around when the antenna is rotated. The length of the spreaders, wind loading, and safety factor dictates the heavy mechanical design.

The lower boom, which is 23 feet (7 m) off the ground, is made of a 24-foot (7.32-m) section of 2-1/4 × 1/4-inch (57- × 6-mm) T-6 tubing with shorter 2- × 1/4-inch (51- × 6-mm) pieces telescoped and bolted at each end to make the total length of 36 feet (10.97 m). To maintain symmetry, 17-foot (5.18-m) spreaders are attached to each end of the lower boom. These spreaders are made of 1-1/2-inch (38-mm) fiberglass. They are attached to the boom with aluminum angles and stainless-steel clamps. A light stainless-steel cable and strut holds the lower spreaders firm against the pull of the quad-element wires, which are attached to each end of the lower spreaders with stainless-steel clamps. The wires are taped along the fiberglass almost to the center of the spreaders, and then go off at right angles to the tuning box, which is mounted on the tower 13 feet (4 m) above ground. The distance from the end of the lower boom to the tower is about 18 feet (5.5 m), and the 36 feet (11 m) of wire which connects the quad proper to the tuning box represents loading.

Tuning

Each element, including the 36 feet (10.97 m) of connecting wire just mentioned, resonates at 4050 kHz. Thus, a small amount of additional inductive loading is required to tune the antenna to resonance within the 80-meter band. This is accomplished by putting two relatively small motor-driven coils in series with each loop of the antenna. The reversible motor for each pair of coils in each loop is controlled by a two-way toggle switch mounted on the antenna control panel in the shack. In the case of the driven element, tuning is accomplished simply by applying power to the antenna and then adjusting the loading inductances by the motor-control switch to the point where the SWR is minimum, usually very close to 1. Tuning the reflector can be accomplished by turning the back of the quad toward a distant signal, and adjusting the reflector loading coils for minimum received signal. This seemed a bit cumbersome to do each time the operating frequency was changed from cw to ssb and vice-versa, so a microswitch was added to the reflector coils and motor assembly. Each revolution of the coils flicks the microswitch, which actuates a light on the control panel in the shack. By counting the blinks of the light, and referring to a chart which shows the resonant frequency of the reflector versus the number of turns of inductance in the loop, the reflector can be tuned to any desired frequency. With this tuning scheme, it is quite easy to tune the quad for best performance at any point in the 80-meter band. The bandwidth between the 3:1 SWR points is close to 200 kHz, so retuning is required only for large frequency changes within the band. Frequency flexibility otherwise is quite good.

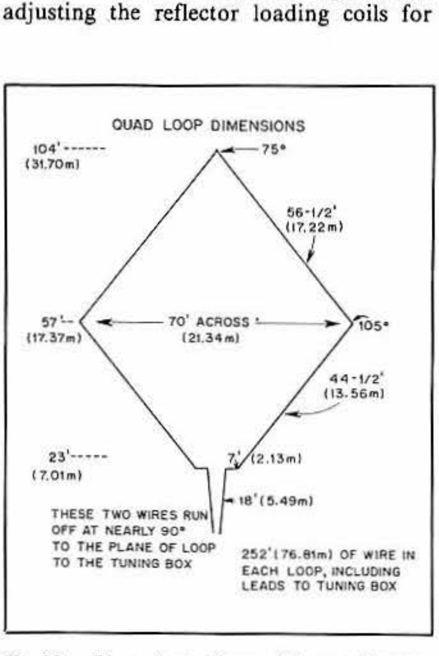
QUAD LOOP DIMENSIONS
104' (31.70m)
75°
56-1/2' (17.22m)
57' (17.37m)
70' ACROSS (21.34m)
105°
44-1/2' (13.56m)
23' (7.01m)
7' (2.13m)
18' (5.49m)
THESE TWO WIRES RUN OFF AT NEARLY 90° TO THE PLANE OF LOOP TO THE TUNING BOX
252' (76.81m) OF WIRE IN EACH LOOP, INCLUDING LEADS TO TUNING BOX

Fig. 22 — Dimensions of one of the quad loops.

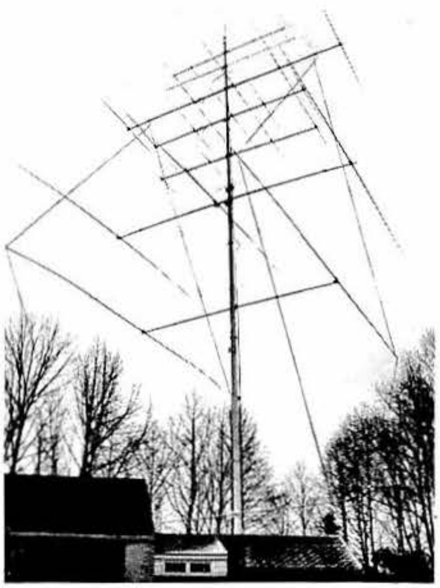

Here is the complete antenna system, 2 through 160 meters, all rotatable, directional arrays.

The impedance of the driven element measures about 75 ohms. The antenna is coupled to the transmitter through a balun and 50-ohm line. The difference between the resonant frequency of the driven element and the reflector, when the latter is tuned for maximum front-to-back ratio, is in the order of only 1-1/2 percent. This probably results from the quad being only a one-quarter wavelength above the ground. The front-to-back ratio is about 20 dB, and the front-to-side, 50 dB. All measurements have an indicated gain over a dipole in the order of 6 to 7 dB. No degradation in performance, or SWR, has been observed during heavy rain. The small loading coils are protected from the weather, of course, in their aluminum tuner box, so the problem, which was encountered with unprotected load coils in the 20-meter test model, has been eliminated.

Tests were conducted with the second element tuned as a director instead of a reflector; there was no noticeable improvement in gain, whereas the front-to-back ratio was diminished. It is felt that the best results are obtained with the reflector.

Some thought was given to driving both elements at a 135-degree phase difference to obtain a cardioid pattern. It is felt, however, that the conventional quad pattern, with deep nulls on each side, and a reasonably good front-to-back ratio, is more desireable for DX. So, nothing further has been done with such a phasing arrangement.

Results

DX results on 80 meters have improved considerably since the new antenna was put on the air. On cw, the gain seems to drop off somewhat as compared to the phone end of the band. Probably this is due to the fact that the relative size of the antenna is smaller at the cw operating frequencies. Nevertheless, reports usually are from one to several S units higher than other U.S. stations with comparable power input, and conventional antennas. In most instances during pile-ups in a recent cw contest, it took only a call or two to get through, whereas, previously, the station was usually last. It was found also that contacts could be made earlier, as the band opened. In several instances like this, it was amazing to get a response from a DX station, and then to hear other U.S. stations calling without success.

Reports from overseas on ssb are outstanding. During a recent phone contest, several dozen DX contacts were made in just a few hours — everyone on the first call.

During the day, on 80 meters, it is quite easy to work into Canada or the Carolinas with good signals at both ends of the circuit. On a dipole, very often the other station is completely unreadable, if not inaudible. This points up the fact that one notices great improvement in reception with the quad compared to a dipole, which on 80 is just as important as being able to transmit a better signal.

Once it was determined that the quad worked well on 80 meters, the possibility of operating the antenna on 160 meters was explored. Since the loops independently resonated at 4050 kHz, and there was considerable inductive loading available in the coils, it was felt that it might be possible to resonate the antenna on 1.8 MHz by putting the two loops in series. Actually, with the load coils tuned to minimum, the series-connected loops resonated at 1775 kHz because of the mutual coupling. By shortening out the coils entirely, the pair of loops were resonated at 1805 kHz. Therefore, the SWR at the low end of the 160-meter band is very low.

Domestic reports on 160 meters have been excellent. Consistent directional effects have been noted by several observers when the two-turn loop is rotated. The pattern is the typical figure eight, with deep nulls off each side of the loop. It remains to evaluate how well the antenna works on 160-meter DX.

Following completion of the antenna in September, there have been several severe storms, with winds gusting as high as 65 mph. These velocities barely moved the wires and spreaders around. It is apparent that the heavy construction is adequate to handle the winds of much higher velocity for which the antenna was designed. Although it is probably one of the largest quads in existence, it certainly appears as though it will stay up a long time, even through rough weather.

In conclusion, this project could never have been completed successfully without the help of many domestic and overseas amateurs. Their reports, advice, and assistance will always be appreciated. Although this has been a sizeable and difficult antenna task, a great deal of satisfaction has been derived from carrying it through from concept to on-the-air operation. The simple lesson has been relearned with regard to antennas — that one cannot get something for nothing. Mini-size antennas are better than no antennas, but there is no substitute for full-size antennas if one wants full-size results. This material was originally presented in *QST* by Joe Hertzberg, N3EA.

The Folded Mini Quad

A full-wavelength loop can take a variety of forms. Despite the shape suggested here, the author claims good results.

An unusual type of full-wavelength wireloop antenna is in use here on 2 and 15 meters. The design, which the author had not seen previously, provides an antenna that occupies minimum space; the antenna is omnidirectional, vertically polarized, broadbanded and requires no groundplane.

It evolves from a full-wavelength loop.[1] The loop, if fed on the side (x-y) as in Fig. 23A, radiates a vertically polarized signal. The radiation pattern is bidirectional, with maximum radiation broadside to the plane of the loop and with the highest field strength in the plane A-A′.

The loop can be deformed into a square, such as in the elements of a quad, and without much loss of efficiency into a rectangle (Fig. 23C). Vertical polarization, bidirectional signal pattern and maximum radiation from the plane A-A′ are retained.

A 90 degree bend at A-A′ and two further right angle bends at B-B′ and C-C′ transforms the flat loop into a cube. The circumference is still a full wavelength and each side of the cube measures 1/10 wavelength. Again, vertical polarization is retained, but the appreciable bidirectional radiation from the sides of the cube adjacent to A-A′ leads to a nearly omnidirectional radiation pattern.

My 2-meter antenna was constructed from aluminum clothesline wire; the length in feet was chosen by 1005/f MHz. It is fed with 50-ohm coaxial cable through a Pomona Electronics no. 1699 adapter and no matching device is used. The SWR is less than 1.2:1 from 144 to 146.5 MHz and climbs to 1.4:1 at 148 MHz. The radiation pattern was measured with a simple field-strength meter, using a half-wave dipole as pickup element and a 20-kΩ resistance in series with the 20-μA meter to improve linearity.[2] The front-to-back ratio is 1.8 dB with no prominent sidelobes apparent, but a rather sharp 3-dB dip is caused by the "shadow" of the feed line and hardware.

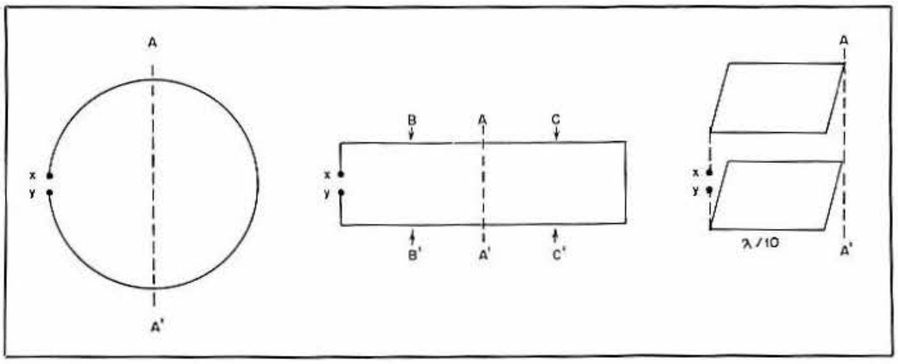

Fig. 23 — Conversion of the full-wavelength loop to the space-saving cubical configuration. The length of the wire and the vertical polarization remain but the pattern changes from bidirectional to omnidirectional. The spacing between the feed point and the adjacent side is not critical. The spacing on 2 meters is about 1-1/2 inches (38 mm), and on 15 meters is about 4 inches (102 mm).

Performance

Fed with 10 watts and used as an indoor antenna, it gives consistent communication with 2-meter fm base and mobile stations within a 10- to 15-mile radius. It performs better on receive and transmit than a 1/4-wave whip at the same location.

A 15-meter version of the antenna was then constructed and suspended in the attic with string and tacks. It is operated with 100 watts and the rig loads almost as well as the 2-meter version, but with a slightly higher SWR. The first contacts were with California and it has since given excellent performance to the West, Midwest and South as well as to South America and Europe.

The close proximity of antenna sections radiating out of phase may lower the efficiency over that of a loop, yet, in terms of space requirement, construction cost and ease of installation (no radials!) and tuning, the antenna is hard to beat. As a full-wave loop it should have a slight edge over a vertical. The author recommends further experimentation and modification, e.g., into the even more space-saving configuration of a cylinder. This material was originally presented in *QST* by Max Blumer, WA1MKP, SK.

References

[1] *The ARRL Antenna Book*, 13th ed., 1974, p. 65.

[2] *The Radio Amateur's Handbook*, 46th ed., 1969, p. 553.

A Convenient Stub-Tuning System for Quad Antennas

The cubical quad has been a popular antenna with amateurs for a number of reasons — relatively light weight, small turning radius, low cost, and good DX performance at rather low heights when compared to other antennas. The authors' experience with quads has demonstrated, however, that best possible performance requires careful parasitic-element tuning. Cutting elements according to various "established" length formulas has often yielded less than satisfactory results. While adjustments for maximum forward gain are generally uncritical, a mediocre front-to-back ratio often results unless time and care are taken to tune the parasitic elements. Factors such as spacing between elements, proximity of the antenna to ground, or influence of other objects (including the other concentric loops in a multiband array) require that the electrical length of the parasitic elements be adjusted using empirical methods if optimum performance is to be realized from the system. It was felt that if one was going to the trouble of erecting a good antenna system, it was worth seeing to it that the antenna was delivering the performance of which it was capable.

These considerations led to the development of the parasitic-element tuning system described here. While the scheme is applicable to quads in general, whether rotary or fixed, the original intended application was in a fixed two-element, full-size 40-meter quad. The descriptions here apply to this 40-meter system, but the general information allows one to follow the same scheme with a quad for any band.

A 40-Meter System

This 40-meter antenna consisted of two non-rotatable full-wave wire loops suspended from a boom which was mounted near the top of a tower. Spacing between elements was 22 feet (6.71 m). The corners of the loops were tied down to convenient anchor points by means of ropes, thereby eliminating the need for spreaders. With the elements thus fixed in place, the antenna was oriented in one general direction and could not be steered elsewhere. A driven-element and reflector combination was used.

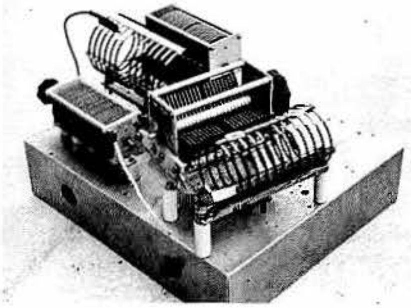

The matching network built by the authors for the modified 40-meter system. Relays, used in the switching circuitry, are not visible. The builder is left to choose a suitable network of his own to meet his operating requirements. Basic design information can be found in *The Radio Amateur's Handbook* or *The ARRL Antenna Book.*

Early experiments with the antenna demonstrated the need for careful reflector tuning. Several trips were made up the tower to alternately prune and add wire to the reflector before the antenna exhibited good performance at the frequencies of interest. But when tuned properly, the quad really performed! In many instances signals which were inaudible while using an 80-foot (24.4-m) high reference dipole were perfectly readable with the quad (at a median height of about 100 feet [30.5 m]). Reception was further enhanced by rejection of unwanted signals off the back of the quad.

Now if only the quad could be made rotatable. . . . Unfortunately, a full-size rotating system (spreaders would be needed) was out of the question. The cost, difficulties of construction and installation, and the severe effects of New England winters discouraged such ideas. The authors, being avid contesters and DXers, were willing to settle for the next best alternative — keeping the fixed quad, but electrically switching its pattern.

The solution involved a tunable stub attached to one of the elements to lengthen or shorten it electrically. With the appropriate stub switched in, the single element could be made to look like either a reflector or director and thus cause the pattern to be reversed. In order that adjustments be made quickly, easily, and from a convenient location (such as the shack or the bottom of the tower), it involved more than the conventional short stub often used with quads. Rather, the situation called for a long stub — long enough to reach the desired remote location — terminated in a variable reactance. Tuning the element, then, involved nothing more than adjustment of the value of reactance which could be provided conveniently by means of a variable capacitor.

The electrical principle is simple. One adds inductive reactance to lengthen an element electrically, or capacitive reactance to shorten it. Typically, the former arrangement is used with quads in a driven-element and reflector combination in which both elements are cut to the same physical length. A stub, shorted at the end, is attached at the center of the bottom section of the reflector. By varying the position of a shorting bar along the stub, the amount of inductive reactance presented to the element can be varied and the electrical length altered correspondingly. Because the stub is short in length, this adjustment must be made at the antenna itself, preferably near or at its final height.

Now, if the stub is cut to some different length which happens to be longer than in the above case, the parasitic element can still be made to see the proper amount of inductive reactance (if it is a reflector) provided the stub is terminated properly with some other reactance (in general, something other than a short). Varying the reactive termination is then equivalent to moving the shorting bar along the conventional stub. This is a consequence of the impedance-transforming property of transmission lines. The long stub can be treated as a transmission line. If we know its characteristic impedance, its length, and the impedance that must be seen at the parasitic-element end of the line, the Smith Chart can tell us what kind of termination is required at the opposite end.

A reflector cut to the same physical dimensions as the driven element — self-resonant at the desired operating frequency — must "see" roughly 150 ohms of inductive reactance at the terminals where the stub is attached. Although little information is available about the use of directors with quads, it is assumed initially that a director, also cut to the same length as the driven element, must see a roughly equal but opposite amount of reactance, or about 150 ohms of capacitive reactance. If one desires to make the termination at the end of the stub a variable capacitor for convenience of tuning, which was the case in the 40-meter system under consideration, the problem can be rephrased slightly: what length of line of a given characteristic impedance is required so that a variable capacitor (a 10- to 350-pF unit was used) causes the parasitic element to see the above reactances? This is easily solved with a Smith Chart. Two somewhat different systems evolved from these ideas. The second, a modification of the first, will be described last.

The design procedure for the 40-meter version went as follows: The driven and parasitic elements were cut to the same length, given by the more or less standard formula, $L = 1005/f$, where L is the length in feet, and f the frequency in MHz. This resulted in a loop circumference of 143 feet, 6 inches (43.74 m) at a frequency of 7.0 MHz. In the practical installation, the loops were triangular in shape rather than

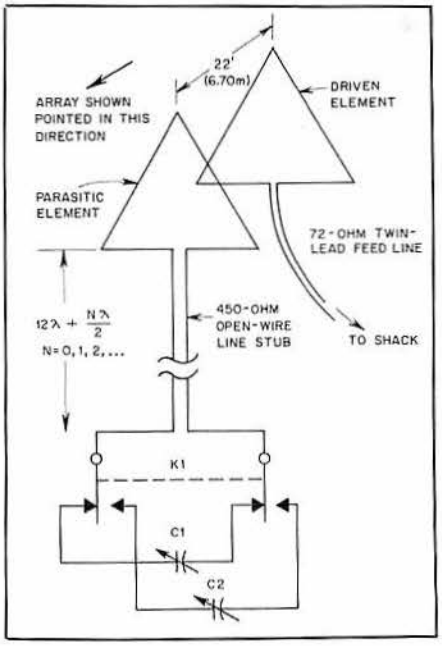
22'
(6.70m)
ARRAY SHOWN POINTED IN THIS DIRECTION
DRIVEN ELEMENT
PARASITIC ELEMENT
72-OHM TWIN-LEAD FEED LINE
12λ + Nλ/2
N=0, 1, 2, ...
450-OHM OPEN-WIRE LINE STUB
TO SHACK
K1
C1
C2

Fig. 24 — Configuration of 2-element switchable 40-meter quad. Triangular loops are shown, although square or diamond loops can be used too. Both elements are cut to the same dimensions: 143 feet, 6 inches (43.74 m) for 7.0-MHz operation. The selection of either C1 or C2 by means of K1 determines which direction the array is pointed. Switching components K1, C1 and C2, can be installed inside the shack or on the ground. The stub line is cut long enough to reach them. Feeding the driven element with 72-ohm Twin Lead results in a low SWR.

C1 — 10- to 350-pF air-variable capacitor, adjusted for director operation as described in text.
C2 — Same as C1, except adjusted for reflector operation.
K1 — Dpdt relay (Potter and Brumfield KA11DG or equivalent).

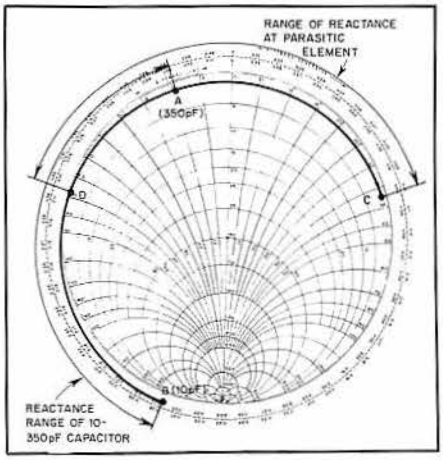

Fig. 25— Example of Smith Chart calculations required to arrive at correct stub dimensions. A 10- to 250-pF capacitor is used as the variable-reactance element which terminates the 450-ohm transmission line used for a parasitic-element stub. The chart is normalized to 450 ohms. The arc between points A and B represents the range of reactance of the capacitor as it is tuned between the extremes of its range. If the capacitor is connected to a 0.12-wavelength long stub (plus any multiple of one-half wavelength) with a 450-ohm characteristic impedance, the range of reactance represented by the arc between points C and D is presented to the parasitic element. This range is sufficient to cause a properly cut parasitic element to look like either a director or reflector. Further discussion of Smith Chart calculations can be found in the latest edition of *The ARRL Antenna Book.*

square or diamond. Shape is relatively uncritical — performance is approximately equivalent for the various geometric configurations, and ease of installation was considered more important. The triangle has the advantage of requiring only three tie-off points — one at the top where it hangs from the boom and two at the bottom — whereas the other two forms require support at four corners. Furthermore, the triangle requires less vertical height — approximately 41 feet (12.5 m) on 40 meters, assuming equal-length sides — as opposed to the diamond which needs about 50 feet (15.25 m). The square requires the least vertical height but is the most difficult to support mechanically. The resultant configuration is shown in Fig. 24.

The stub line was made from 450-ohm open-wire transmission line because of its relatively low cost in commercially made form and its low-loss characteristics. The reactance range of the 10- to 350-pF capacitor by itself at 7 MHz was then plotted on the Smith Chart as shown in Fig. 25. (Recall that $X^L = 2\pi fL$ and $X^C = 1/2\pi fC$ where X^L and X^C represent values of inductive and capacitive reactance respectively, given in ohms, L and C the corresponding values of inductance in henrys and capacitance in farads, and f, the frequency in hertz.) The low-capacitance end of the range is indicated as point B in Fig. 25 with the Smith Chart normalized to 450 ohms in this example, while the other extreme comes out at point A. If the capacitor was then connected at the end of a 0.12-wavelength section of the 450-ohm line (approximately 16 feet [4.88 m] in open-wire line), the impedance seen at the other end of the line would vary between the range of slightly greater than +j300 to -j300, represented by points C and D. This would meet the +j150 and -j150 reactive loading requirements of the parasitic element for reflector and director operation, respectively. Actually a stub length of 0.12 wavelength, plus any multiple of a half wavelength, would work too, because any half-wavelength section of transmission line merely repeats at one end the impedance that appears at the other. Thus, the distance to the desired remote location dictates the minimum required line length. Another advantage of using 450-ohm line, as opposed to a line with a lower characteristic impedance, becomes apparent when working examples on the Smith Chart — for a given capacitor tuning range, the 450-ohm line produces a greater range of reactance change at the antenna end of the line. A 300-ohm line would serve almost as well, however, providing only slightly less tuning flexibility.

The stub was terminated at a dpdt relay which was controlled from the station operating position and which was used to select one of two variable capacitors — one adjusted to provide reactance for reflector operation and the other for director operation, as shown in Fig. 24. In this manner, the pattern could be flipped around 180 degrees instantly from the shack. The driven element was fed directly with 72-ohm transmitting-type Twin-Lead. A low SWR was obtained with no special provisions made for matching.

The director and reflector tuning should be done empirically, as was stated before. Experience has shown that quads tune rather broadly for maximum forward gain. It was found that fairly large changes in the settings of the variable capacitors did not alter the gain significantly. It is best, therefore, to tune for maximum front-to-back ratio, on which the settings had considerably more effect. On-the-air signals arriving from the rear direction can be nulled by tuning the capacitors. The nulls should be fairly pronounced if the

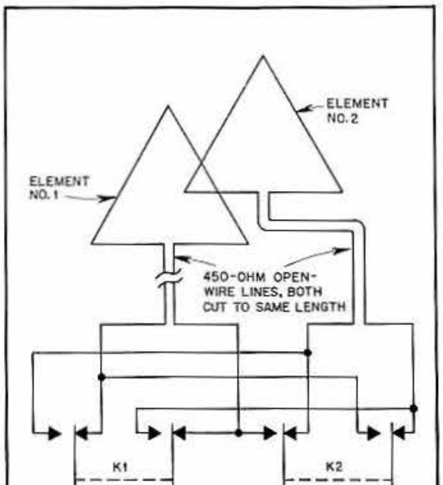
ELEMENT NO. 2
ELEMENT NO. 1
450-OHM OPEN-WIRE LINES, BOTH CUT TO SAME LENGTH
K1
K2
C1
MATCHING NETWORK
COAXIAL FEED LINE (RG-8/U, RG-11/U, ETC.) TO STATION

Fig. 26 — Configuration of modified 2-element switchable 40-meter quad. The lines from each element are both cut to the same length, which can be the same as the parasitic element line in Fig. 24 or other suitable length. Both elements are cut to the same dimensions as in Fig. 24. The positions of K1 and K2 determine whether an element is fed as a driven element or terminated as a reflector. The array pattern can be switched by determining which element operates as a reflector or driven element. Only one tuning capacitor is required, but a matching network is needed to match a coaxial feed line.

C1 — 10- to 250-pF air-variable capacitor, adjusted for reflector operation.
K1, K2 — Dpdt relays operated simultaneously, or both may be replaced by a single four-pole, double-throw relay (Potter and Brumfield KA11DG or equivalent).

system is set up as described here. The exact amount of attenuation of signals off the back of the antenna varies from signal to signal. It was observed that the 40-meter antenna exhibited the greatest front-to-back ratio for signals arriving from distant locations (at low radiation angles). Alternatively, field-strength measurements can be made to determine the proper settings for minimum radiation or reception off the rear of the antenna. In this fashion, one capacitor is adjusted for director operation and the other for reflector operation.

When making large excursions in operating frequency on 40 meters, it is necessary to touch up the tuning. A director tuned for operation in the low end of the cw segment is self-resonant approximately in the middle of the phone portion and provides little usable gain at those higher frequencies. Retuning is done easily, however — something which is not possible using fixed-tuned elements.

A Modified 40-Meter System

The 40-meter system described above provided very good results over a period of months, including stateside and DX contest work and casual DXing. After some thought, however, a modified system was constructed. The new configuration provided some advantages over the orginal scheme.

The same principle of operation was retained. The new system shown in Fig. 26, used only a reflector-type parasitic element, but the switching system now allowed either element to operate as the driven element or reflector. In this manner, the antenna pattern could still be reversed by choosing the appropriate element for the desired function.

Several factors prompted the change. It was found with the original system that a considerably higher front-to-back ratio could be had using a reflector rather than a director. The latter yielded no more than about 10-dB discrimination against signals received off the back of the antenna, even with careful tuning, whereas with the former the figure was approximately as high as 25 dB. Although most available literature made little mention of quad performance using a director, this observation agreed with what little published data could be found on the subject.[1] The reflector-only model also had the advantage that just one tuning adjustment was required, since one stub reflector served both elements in an identical fashion. Furthermore, because the system behaved the same no matter which direction the array was switched, the SWR did not change when switching, and the transmitter was thus always tuned properly. With the original scheme the antenna feed-point impedance changed when switched between director and reflector operation. Although the resultant SWR was not objectionable in either case (it did change, though), it was necessary to retune the transmitter after switching, or tolerate a moderate amount of mistuning.

The design procedure for the new arrangement is basically the same as before. Both elements are cut to the same length. Only the stub requirements for a reflector need be considered now in doing the Smith Chart calculations. Once the stub line length is determined, two such identical-length lines, both made from the same type of transmission line, are cut, one for each element. A four-pole, double-throw relay, or two dpdt relays operated simultaneously, are required in the hookup shown in Fig. 26, to switch the two lines between the parasitic-element termination and the main feed-line matching network.

The new system has one disadvantage, however. A matching network is required between the transmitter and driven element because of the mismatch between the main feedline, which is presumably coax, and the driven-element open-wire feed line. The original system required no special matching. This presented no problem in the 40-meter system considered here since a suitable network was easily constructed. The builder is left to choose a suitable matching network of his own as there are numerous possibilities to take care of the varied matching requirements which may arise under different situations. Depending upon one's preferences and requirements, the modified scheme may or may not prove more attractive.

Additional Remarks

A few general comments can be made at this point. Although a variable capacitor has been designated thus far as the termination for the stub, there is no reason for not using a coil or some other combination of inductance and capacitance if it will work with a stub length which happens to be convenient. The builder can choose for himself. The 10- to 350-pF capacitor described earlier had approximately 0.05-inch (1.27 mm) spacing between plates. While it is difficult to predict the voltages present in the parasitic element (and across the capacitor), no arcing between plates was observed during transmitting periods, even while running one kilowatt transmitter dc input. This suggested the plate spacing was adequate.

The lines from each element should be brought away perpendicular to the elements as much as possible, and open-wire line should be kept reasonably clear of nearby objects, especially metal ones. Also, the velocity factor of the line being used (0.95 for "ladder" line) must be remembered when computing physical lengths.

The authors erected a single diamond-shaped 80-meter quad loop with its bottom close to the ground. The antenna was tried as both a vertically and horizontally polarized radiator. For DX work, the vertically polarized version yielded unquestionably superior results in terms of greater signal strengths at long distances and the beneficial feature of rejection of unwanted high-angle radiation from stateside stations when receiving. A parasitic element to accompany the single loop was not tried. These observations suggest interesting possibilities for vertically polarized ground-mounted quads for low frequencies.

The ideas presented here are applicable to any quad antenna, but should be especially useful in designing large arrays which cannot be turned physically. The design procedure is simple in any case. The feature of continuously variable tuning eliminates the guesswork in cutting fixed-length, fixed-tuned elements. The system will serve well in any situation where, say, climbing the tower to work on the antenna is not convenient.

A fixed, switchable array is an asset to the contestant or DXer who finds it necessary to change antenna headings in a hurry. From the northeastern U.S., a single such array provides good coverage of Europe and Africa in one direction and of the western U.S. and the Pacific in the other. If two suitable supports spaced a reasonable distance apart are available, two arrays of this type, perpendicular to each other, can provide a wide range of coverage. This material was originally presented in *QST* by John E. Kaufmann, WA1CQW and Gary E. Kopec, K8RX.

---

[1] "Technical Topics," *Radio Communication*, February 1973, p. 101.

Some Notes on a 7-MHz, Linear-Loaded Quad

If you like to experiment with antennas, here are enough new ideas about the construction of quad antennas to keep you going.

Most amateurs who have a serious interest in working DX in the 3.5- to 28-MHz bands would like to have antennas providing them with forward gain on all of these bands. Today, a triband trap Yagi or cubical quad may be employed to achieve this objective on 14, 21 and 28 MHz. Rotary beams are, for all practical purposes, out of the question for 3.5 MHz, but it is possible to achieve some gain with a 7-MHz rotary beam without having to pay a fortune for a supporting structure.

The Hy-Gain linear-loaded 2-element Yagi or the Mosley inductively loaded beam for 7 MHz can achieve this objective. The writer, however, prefers quads to Yagis, based on his own experience. Difficulties often arise when one seeks to mount beams for several different bands on the same tower without marked loss of performance on any band. The primary problem is, of course, that a full-sized cubical quad for 7 MHz is a monstrous structure. Such antennas have been erected, but they are not suited for the average amateur with a limited budget.

The first step in building a 7-MHz quad is, therefore, to design a single 7-MHz quad loop of reasonable size which is mechanically strong enough to survive most weather conditions.

Linear Loading

As a general rule, a reduction in antenna size to two-thirds of full size can be effected without a noticeable drop in performance, providing care is taken to limit the losses introduced by the loading devices. This means that one can think in terms of a loop for 7 MHz with 24 feet (7.32 m) on a side instead of 36 feet (10.97 m) on a side. An additional factor in choosing a 24-foot-per-side length was that the longest bamboo spreaders that the writer could obtain were 18 feet (5.5 m). Attempts were made to lengthen these spreaders, but results were not satisfactory mechanically.

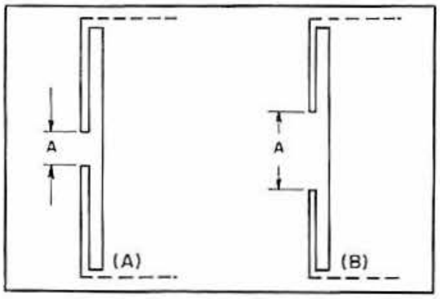

Fig. 27 — A method for linear-loading of one side of a quad element.

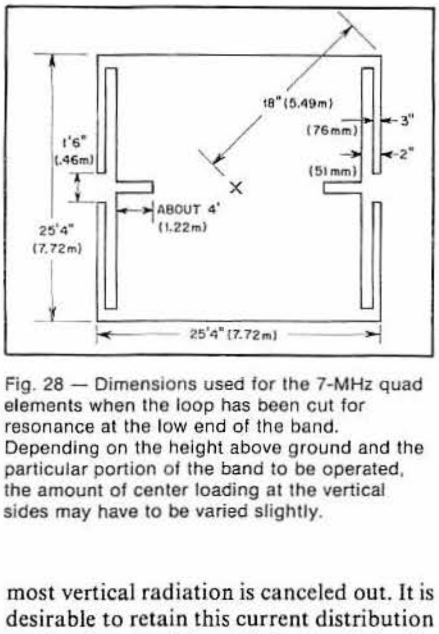

Fig. 28 — Dimensions used for the 7-MHz quad elements when the loop has been cut for resonance at the low end of the band. Depending on the height above ground and the particular portion of the band to be operated, the amount of center loading at the vertical sides may have to be varied slightly.

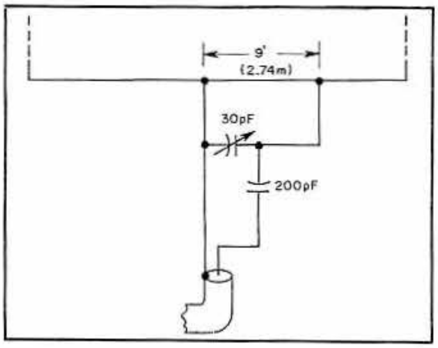

Fig. 29 — Diagram of the matching section used on the 40-meter driven element. The gamma capacitor employs a 100-pF air variable and a 100-pF high-voltage silver mica connected in parallel.

If a 24-foot loop is to be used, the next questions are where to put the loading and what form of loading to use. A full-sized quad loop has symmetrical current distribution. The current maxima occur in the middle of the horizontal sections, and the current minima develop in the middle of the vertical sections. The current distribution in the vertical elements is such that most vertical radiation is canceled out. It is desirable to retain this current distribution in the shortened loop. It is also desirable not to put the loading in the center of the horizontal elements, since maximum radiation occurs at the current maxima, and it does not make sense to cut out the portion of an antenna that does the most effective work. The writer, therefore, chose to put the loading in the vertical sides.

There are two popular methods of loading, stubs or coils. The writer dislikes coils for loading since, if they are wound with a heavy-gauge wire (desirable in order to avoid undue losses), they are bulky and difficult to weatherproof. The choice was therefore stub loading. It would have been possible to insert a stub at each corner of the loop, tied back to the spreaders. But, the current flowing in such stubs would be unbalanced, and radiation in unwanted directions would take place from the stubs themselves. There is also the consideration that if the same spreaders are used for 14-, 21- and 28-MHz loops, stubs at the corners of the 7-MHz loop would have to be mounted well clear of the loops for the other bands.

The original linear-loaded 7-MHz quad element was tried by the writer in a four-band quad some years ago. A single stub in the center of each vertical side was used. The loading stubs were made 14 feet (4.27 m) long and were tied on to a point near the center of the boom. Performance of this quad on 7 MHz was excellent, but the 21-MHz quad mounted on the same boom was almost useless, probably due to pickup at 21 MHz by the 7-MHz loop with subsequent reradiation. An attempt was made to find an efficient method of loading which did not involve long stubs "floating around." The next step was to try folding the wires in the vertical sections back on themselves as shown in Fig. 27A. Current distribution either side of the center part of the vertical elements is symmetrical, as desired. This change resulted in a loop that was resonant near 7 MHz, although its resonant frequency was slightly high. Shortening of the stubs slightly, as shown in Fig. 27B, was tried. This had virtually no effect on resonant frequency. This is no doubt because the increased end-loading effect of the portion A almost equalled the reduction in stub length. Short stubs (abour 4 feet [1.22 m] long) were then

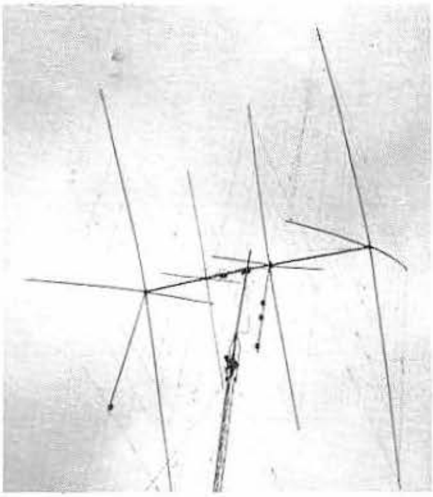

G3FPQ's compact 4-band quad has 2 elements for 7 MHz, 3 elements for 14 MHz, and 4 elements for 21 and 28 MHz. Pairs of Yagi-type directors are used in place of the usual loop director for the three higher frequency bands.

inserted at the center of the vertical sides, at right angles to the vertical sides as shown in Fig. 28, and resonance at 7 MHz was achieved.

A four-band quad for 7, 14, 21 and 28 MHz was then constructed. Much to the writer's delight, there appeared to be no noticeable interference with the pattern of the 21-MHz quad by the 7-MHZ elements. The loading effect of the stubs would be different at 21 MHz since they would include current maxima; the result is to give a loop which, while resonant at 7 MHz, does not appear to have a third-harmonic resonance in or near the 21-MHz band.

Yagi-Type Directors

The reader can probably see from the photograph that the directors for 14, 21 and 28 MHz consist of pairs of stacked Yagi-type elements level with the driven elements. There are two pairs of Yagi directors for the 10- and 15-meter elements at approximately 6-foot, 6-inch (2 m) spacing, and one pair of Yagi directors for 20 meters at approximately 13-foot (4 m) spacing. The overall length of the directors is 31 feet, 3 inches (9.52 m) on 20 meters; 20 feet, 10 inches (6.35 m) on 15 meters; and 15 feet, 2-1/2 inches (4.63 m) on 10 meters. Performance of the author's array on 10, 15, and 20 meters at the moment is little, if any, better than an ordinary 2-element quad. The writer has only recently had an opportunity to start checking out the directors to see what adjustment they need. Current indications are that they require lengthening, but there is a good deal more experimental work to be done. Anyone who wants to try this system must be prepared to do a good deal of cut and try to tune the directors, if any real increase in performance over a standard two-element quad is to be achieved.

One also has to bear in mind that when dealing with multiband quads the resonant frequency of Yagi-type directors is likely to be affected to a far greater extent by the proximity of other elements than would be the case for a quad-loop director. This is because a Yagi director is a comparatively high-Q element. From the theoretical point of view, the drawback to quad-loop directors is that the reactances involved are such that there is a comparatively low degree of mutual coupling between the quad-driven element and the quad-loop director. This, of course, reduces the efficiency of the loop as a director. This problem should not apply to the same extent with Yagi-type directors, and, in theory at least, they should prove to be more efficient than quad-loop directors. The question now is whether practice follows theory or not. This writer won't have an opportunity to conduct any detailed experiments until the summer. When deciding the original length for the Yagi directors, allowance was made for a rise in frequency of approximately 100 kHz. But, the only way to cope with the effects on the director resonant frequency of the proximity of other elements is by field-strength measurements, coupled with a good deal of cut and try. As in any multiband quad, separate feed lines should be used for each band to minimize interaction between bands.

The author's quad has been up for six months now, and final tuning remains to be done to the 7-MHz reflector loop to achieve maximum front-to-back ratio. But the results, even on a cut-and-try basis, are well worth the work that went into the array. This material was originally presented in *QST* by David. L. Courtier-Dutton, G3FPQ.

Spider Quad Mount: Simplified

There seems to be a growing interest in the spider quad. The photograph shows a version of a hub or "spider" mount. Using a 2-foot (0.61-m) length of 2-inch (51-mm) diameter steel pipe, a mount can be constructed easily. Two cuts, 10 inches (254 mm) long, are made through the diameter at each end of the pipe. This leaves 4 inches (102 mm) in the middle for the boom. The cut ends are then bent to an angle to suit the desired spacing. The unit shown in the photograph is bent to 108 degrees. The spreaders may be attached to the boom mount with hose clamps, bolts or wire.

The beauty of this design is that it can be made to any size desired. Also, it can be enlarged to a 4-element version, as shown in Fig. 30.

Fig. 30

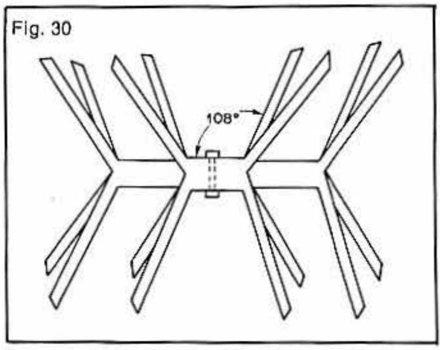

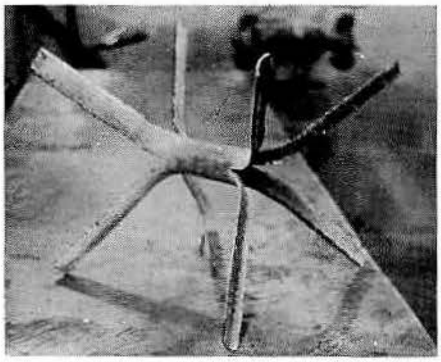

One of these spider mounts is in use at KP4DJI (constructed by KP4DIO and KP4DJI) using bamboo spreaders. The results have been excellent. This material originally presented in *QST* by Lynda B. Crowley, KP4DIP.

Two-Band Quad Matching

While experimenting with a 2-element, 2-band cubical quad antenna the problem of matching the antenna to a 50-ohm transmission line arose. The antenna is a "plain vanilla" quad built around information in W6SAI's book, *Quad Antennas*. The author's method of feeding the antenna *did not* produce fantastic front-to-back ratio or forward gain. It merely made the standing-wave ratio more acceptable to the 32S-3B used at the home QTH.

After building the 2-band quad, the author checked it with a General Radio 1606 impedance bridge and found the 10-meter antenna to be about 118 ohms and the 15-meter antenna to be about 91 ohms at resonance. This is 2.36:1 and 1.82:1, or approximately 2:1 for a 50-ohm transmission line. The feed system described here made the VSWR much nearer the desired 1:1.

The heart of the system is a ferrite-core autotransformer. The measured impedance ratio is very close to 50:107 ohms. The insertion loss is less than 0.2 dB from 15 to 30 MHz when terminated in 107 ohms. When it is terminated in 50 or 200 ohms the insertion loss is less than 0.35 dB from 15 to 30 MHz. Several transformers were built before these figures were obtained, and after a suitable design was achieved it was connected to the quad as shown in Fig. 31. You do not have to sacrifice F/B ratio for a more acceptable VSWR by detuning the reflector stubs; simply adjust the tuning stubs for maximum F/B ratio and then design the transformer for the proper impedance ratio. Also, lowering the VSWR places a larger portion of the desired band within an impedance range the average pi-section tuner will accept.

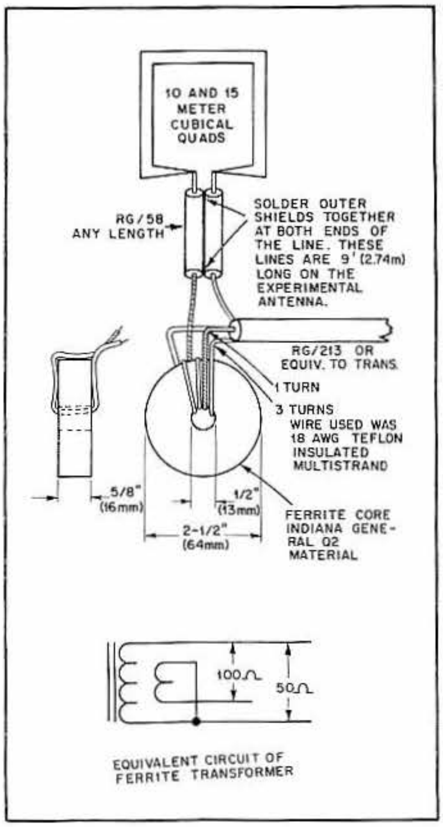
10 AND 15 METER CUBICAL QUADS
RG/58 ANY LENGTH
SOLDER OUTER SHIELDS TOGETHER AT BOTH ENDS OF THE LINE. THESE LINES ARE 9' (2.74m) LONG ON THE EXPERIMENTAL ANTENNA.
RG/213 OR EQUIV. TO TRANS.
1 TURN
3 TURNS WIRE USED WAS 18 AWG TEFLON INSULATED MULTISTRAND
5/8" (16mm)
1/2" (13mm)
2-1/2" (64mm)
FERRITE CORE INDIANA GENERAL Q2 MATERIAL
100Ω
50Ω
EQUIVALENT CIRCUIT OF FERRITE TRANSFORMER

Fig. 31 — Two-band quad matching.

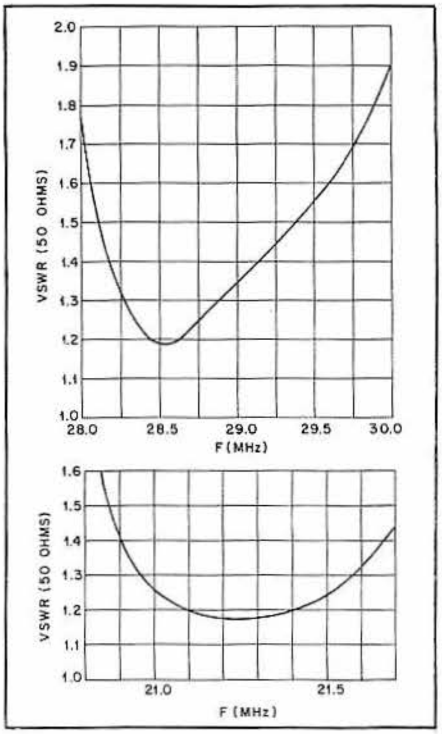

Fig. 32 — Standing-wave ratio curves taken with a laboratory-type VSWR bridge. Measurements were made at the transmitter ends of the transmission line with the antenna array mounted on a steel tower 44 feet (13.4 m) above ground.

In answer to on-the-air inquiries: Yes, performance of the antenna varies considerably from one end of the band to the other.

This antenna feed system is a broadband device, so don't forget to use a low-pass filter. This material was originally presented in *QST* by Toney L. Magnino, W5MVK.

Chapter 4

Miscellaneous Antenna Types

Here is a potpourri of useful antennas for the hf-band enthusiast. This chapter contains information on how to build and operate loop antennas, limited-space antennas and variations of the quad. Once again, the work of numerous *QST* authors has been collected for a timeless volume of useful construction details that can benefit amateurs whose operating objectives are numerous and varied.

For the most part, the antennas described in this section are simple and inexpensive to build. The performance characteristics are consistent with the operating needs of ragchewers, contesters and DX-chasers alike, assuming that the builder does not strive to have the loudest signal on the amateur bands.

Beat the Noise with a "Scoop Loop"

Operators on 160 and 80 meters fight a common foe — noise. Atmospheric QRN is compounded by the presence of man-made noise. Loop antennas for receiving may offer an economical solution to your noise problems. Here are some timely details.

If you've been tempted to massage your receiver with a 10-pound hammer because you can't copy some of the weak signals being heard by your colleagues, you may need a receiving loop! A high ambient noise level on 160 meters has stifled the author's DX work for several years, and more than once was ready to capitulate in favor of the man-made and atmospheric noises that the 75-foot (22.9-m) shunt-fed vertical responded to. "Grrrr," would be the snarl as the author listened to stations with Beverage antennas: They were handing out RST 579 signal reports to DX stations, and the author couldn't find the DX signals in the noise, let alone copy them.

Something had to be done. It was a matter of giving up on 1.8-MHz efforts toward DXCC, or putting forth a zealous attempt toward an improved signal-to-noise ratio. Some effective outdoor receiving antennas were erected (half-wavelength open loop — parallel to the ground, and a 200-foot [61-m] end-fed wire a few feet off the ground), but they did not always respond well to DX signals: It depended on band conditions. Also, there was the danger of receiver front-end damage caused by unwanted excitation of the large receiving antennas (resonant) during the transmit period. They were in close proximity to the vertical antenna, which was used for transmitting purposes. Lethargy prevented the author from installing a protective circuit at the receiver front end, and as a consequence the writer "blew" a JFET preamp and two protective diodes in the front end of the FT-301D transceiver.

Revisiting the Past

While for many years experimenting with small receiving loops for 40, 80 and 160 meters, the efforts were casual. The best of the loops was an outdoor one of the shielded variety, five feet (1.52 m) on a side, and fashioned from RG-59/U coaxial cable.[1] However, it was necessary to go out-of-doors each time to rotate it, and the author wasn't enthusiastic about investing in a rotator to use with the loop.

Recently, it seemed that some work could be done with indoor types of receiving loops. After all, loops of that kind have been around since radio was invented, and plenty of amateurs are using them effectively. After scanning available literature (Terman, Kraus, Jasik and Keen),[2] the author set about the task of building a framework for the chosen design. This time a multiturn loop would be tested (Fig. 1).

Construction

The photograph shows the general structure of the loop. In Fig. 2 are the dimensions of the wooden frame used to contain the antenna wires. There is nothing magic about the size of the antenna: Two pieces of wood that were stored in the workshop were used, each being 48 inches long, 1-1/2 inches wide and 3/8 inch thick (1.22 m × 38 mm × 10 mm).

Arbitrarily, the author decided to make L1 of Fig. 1 a five-turn winding. Holes were drilled (five) in each end of the cross-frame to accommodate the loop turns and hold them in place. Actually, there should be six holes at the bottom point of the frame to permit the final end of L1 to be secured during the threading-through process.

Five additional holes were drilled (Fig. 2) 3/4 inch (19 mm) below the holes for L1. These are used to contain L2, the one-turn output link of the antenna. L2 requires two holes at the bottom member of the frame — again to provide an anchor point for the ends of L2. Alternatively, V-shaped notches can be filed on the ends of the cross-frame. The wire for L1 can be laid in the grooves rather than threaded through the holes shown in Fig. 2. L2 should, however, be wound inside L1, through the five holes shown in the drawing.

A mounting base for the assembled antenna can be made to the builder's specifications. This writer used a piece of 3/4-inch (19 mm) plywood which has dimensions of 30 × 12 inches (0.76 m × 305 mm). A 1-inch (25 mm) diameter hole was drilled in the center of the board to accept an 18-inch (457 mm) length of 1-inch diameter dowel rod. The latter is lashed to the framework of the loop to enable the

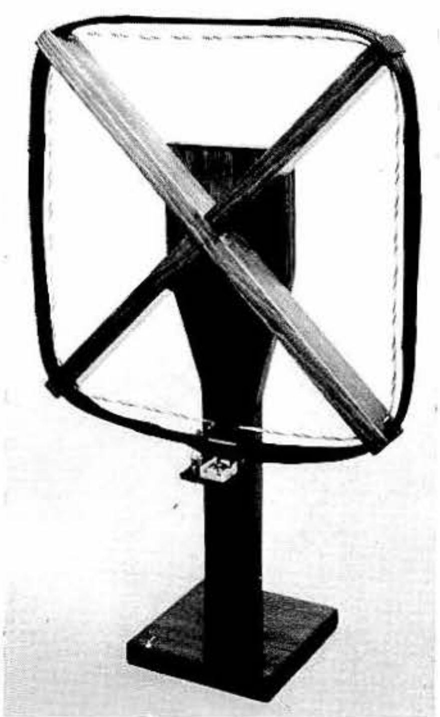

The silhouette is of an indoor version of the four-turn shielded loop described in this article. A wooden framework is used to support the main loop winding and the one-turn pickup link. The resonating capacitor is mounted on an aluminum L bracket along with a phono jack and an 860-pF disk capacitor. The latter is used between the output link and the phono jack instead of a trimmer capacitor. The framework was built by Technical Secretary Marian Anderson's OM, Bob. It has been stained and coated with clear lacquer.

operator to rotate the loop as desired. Although this is a "kluge" type of assembly, it serves its purpose adequately. A *carpenter* should be able to conceive a more practical base and support system for the antenna!

Number 22 insulated hookup wire was used for the loop turns. There is nothing critical about the wire size, but it is best to use the larger gauges. This will make the loop more rugged.

Performance

A variety of feed systems were tried before the author settled for the method illustrated in Fig. 1. Direct connection to the terminals of L1 required careful balancing to assure a uniform loop-response pattern. Unbalance causes pattern distortion, and if it's bad enough the maximum response can occur off the broad side of the loop rather than in the plane of the antenna.

Addition of a one-turn coupling loop, L2, proved to be the easiest method for assuring good balance. Bandwidth of the loop is dependent upon the Q. The latter can be controlled by the amount of capacitance used at C2: the lighter the coupling, the greater the loaded Q. Loop resonance is effected by means of C1. Since there is interaction between L1 and L2, it is necessary to readjust C1 each time C2 is tweaked. In this application it should be stressed that C2 is not used for the purpose of obtaining an SWR of 1. It is purely a *coupling* device. Hence, a 50-ohm condition does not necessarily exist at the loop output terminals.

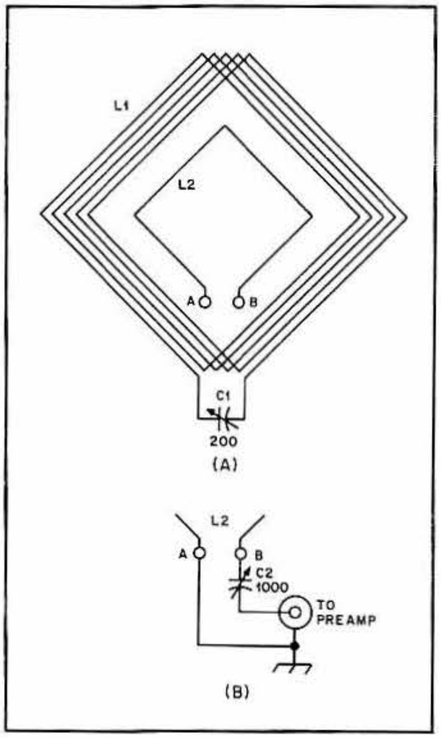

Fig. 1 — Schematic diagram of the receiving loop. C1 and C2 are mica compression trimmers. The illustration at A shows the 1-turn link inside the 5-turn one. At B is an expanded diagram of the feed point for L2.

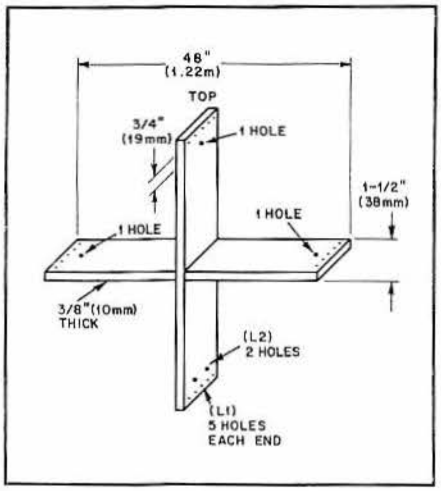

Fig. 2 — Dimensional drawing of the wooden frame which contains the Scoop Loop. Further information concerning the structure is given in the text.

During initial experiments the pickup loop, L2, was connected directly to the 50-ohm input port of a 40-dB broadband preamplifier which was designed by W7ZOI (Fig. 3). Without C2 in the circuit the coupling was excessive. This resulted in low Q, and the loop bandwidth was approximately 150 kHz at the 3-dB points of the response curve. After C2 was added and adjusted to provide a bandwidth of 25 kHz, the loop gain increased 10 dB. Alternatively, one might make L2 smaller and locate it closer to the center of the loop frame in order to lighten the coupling between the two windings. This could be a somewhat tedious mechanical procedure: The writer's inherent lassitude favored the addition of C2 for the purpose!

Performance Characteristics

A test range was set up for checking the null properties and relative gain of the loop. The equipment was arranged in an outdoor location, well away from power wiring and other station antennas. A crystal-controlled weak-signal source (2N2222A oscillator and 9-volt battery) was placed some 100 feet (30.48 m) from the test loop. It contained a 36-inch (0.91-m) whip antenna (vertically mounted) for radiation of the test signal. The 75-foot (22.86-m) transmitting vertical was shorted out at the feed point to prevent it from being resonant and affecting the loop antenna during the test period.

A 50-ohm Tektronix step attenuator was placed in the coaxial line between the loop and the receiver. The attenuator was used to measure accurately the changes in S-meter reading on the FT-301D transceiver which was used as a detector during the measurements.

A null depth of —21 dB was obtained with the loop. An extremely sharp null was noted. As the broad side of the loop was moved toward the signal source, the null began slowly, then fell deeply to —21 dB over the final 10 degrees of rotation. No evidence of pattern distortion was observed. The null was dead off the broad side of the antenna, and maximum response was directly off the thin side of the loop.

A comparison between the 75-foot (22.86-m) transmitting vertical (shunt-fed tower, top loaded and with buried radials) and the wire loop indicated that on groundwave signals the loop (without preamp in line) was —18 dB respective to the vertical. This suggested that in a practical installation a single JFET or dual-gate MOSFET preamp would provide unity gain with the transmitting vertical.

Using the Loop Indoors

The author's home has aluminum siding, and as is the case with all modern homes there is a network of water pipes, phone wires, TV lead-in conductors and house wiring. The author wondered how well the loop would perform within all of these unwanted signal obstacles, some of which form loops by themselves! The ham shack is at ground level inside the family room, and the loop was placed on a stand at floor level. The clutter of unwanted conductors disturbed the directivity of the loop somewhat, but the signal-pickup ability of the antenna was not impaired.

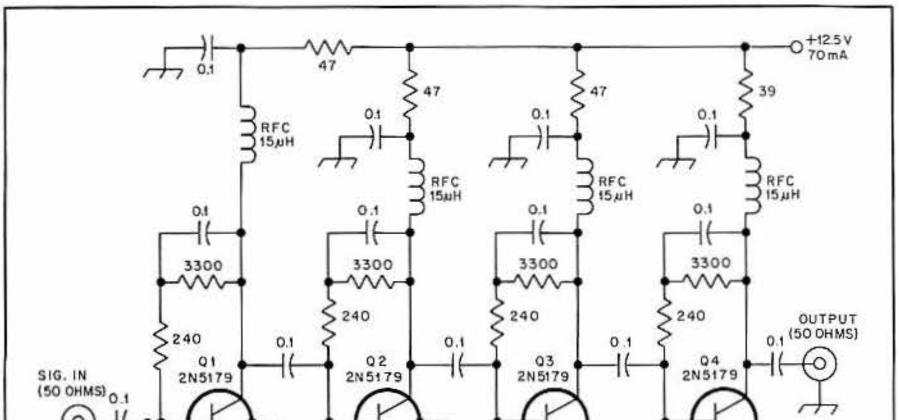

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF) ; OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS ; k = 1 000, M = 1000 000.

Fig. 3 — The W7ZOI broadband 40-dB preamplifier is shown at A. Capacitors are disk ceramic or chip types. Resistors are 1/4-w composition. The upper 3-dB point of the amplifier is at 65 MHz. Heavy feedback is used to stabilize gain and provide a 50-ohm characteristic. At B is a 1-dB attenuator pad which can be used between an antenna of unknown impedance and the filter at C to provide a proper load for the filter. The Butterworth band-pass filter at C is suitable for use with the preamp at A. It will help reject out-of-band signals to improve the IMD characteristics of the preamp. When the preamp is used with nonresonant receiving antennas, it will be necessary to employ the filter to protect the preamp from signals across its response range (bc band through vhf). The preamp and filter would be fine for use with Beverage antennas. C1 and C2 are mica trimmers. L1 and L2 consist of 31 turns of no. 22 enameled wire on Amidon T68-6 toroid cores (5.1 µH).

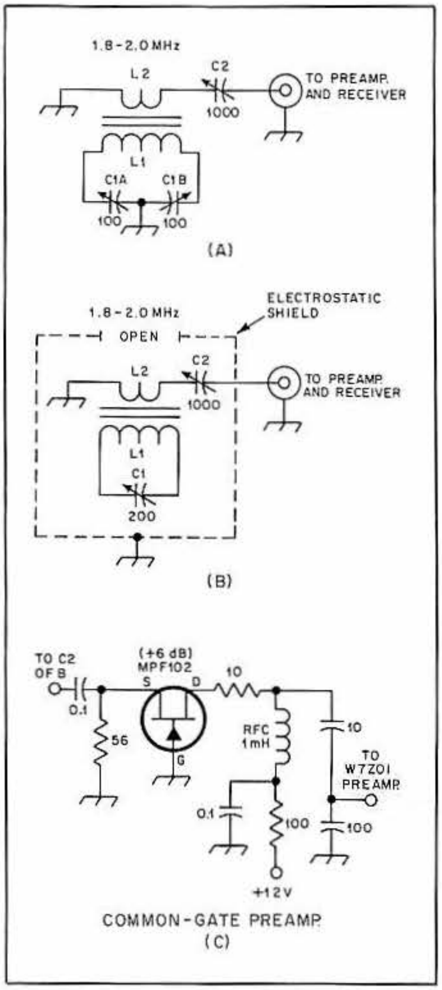

Fig. 4 — The drawing at A shows the circuit of a 7 X 5/8-inch (178 X 16-mm) ferrite-rod loop (see text). At B is the final version which employs an electrostatic shield. C1 and C2 are mica compression trimmers (see text for other details). The circuit at C was used ahead of the W7ZOI preamplifier to improve the system noise figure.

A step attenuator was inserted between the W7ZOI preamplifier and the station receiver so that unity gain could be established between the loop and the 160-meter vertical (22 dB of gain had to be gotten rid of). An antenna switch was added to the system to permit rapid comparisons between the antennas.

Extensive tests proved that the loop outperforms the vertical by a substantial margin when it comes to digging weak signals out of the noise (man-made and atmospheric). The author has copied PA0HIP, PY0ZAE, G3MYI, W6RW, W6BLZ, W5AB and others perfectly Q5 on the loop (indoors) when they couldn't be found, or were unreadable on the vertical. It is only fair to say that on some occasions the vertical was the better antenna for receiving. Propagation conditions and ambient noise levels at a given time will determine which antenna can provide the best reception. Typically, the ambient noise level from the W1FB vertical is S3 on a receiver that is calibrated to indicate S9 on a 50-µV signal.[3] The same receiver shows no noise response when the loop is connected to it. However, static crashes (atmospheric noise peaks) are as loud with the loop as they are with the vertical when a storm is in the immediate area. Noise from distant storms can be discriminated against by turning the loop null toward the storm front. It appears that the major part of the author's ambient noise is man-made in nature. At times it peaks as high as S8.

When the loop is used indoors it responsds to unwanted 15.750-kHz TV spurs, and when appliances are actuated the transients kick the receiver S meter rather hard. This is because the loop is contained within loops formed by house wiring. This type of interference has not been observed while using the vertical antenna, even though it is affixed to the rear of my house. To prove that the house wiring was causing the problem with TV birdies and transients, the loop was located out-of-doors for two nights. The interference vanished!

Some Other Loop Experiments

Ferrite-rod loops have been used to advantage for many years (bc-band radios, direction-finder receivers for boats and aircraft, etc.), so it seemed worthwhile to investigate a model of that type of antenna before concluding the tests. Two versions were tried, each with a preamplifier.[4] The first consisted of a single 7 × 5/8-inch (178 × 16 mm) Amidon Assoc. ferrite rod on which was wound a 40-turn solenoidal coil of no. 22 insulated hookup wire. The loop was tuned to resonance by means of a 100-pF per section split-stator variable. A four-turn output link was wound over the center

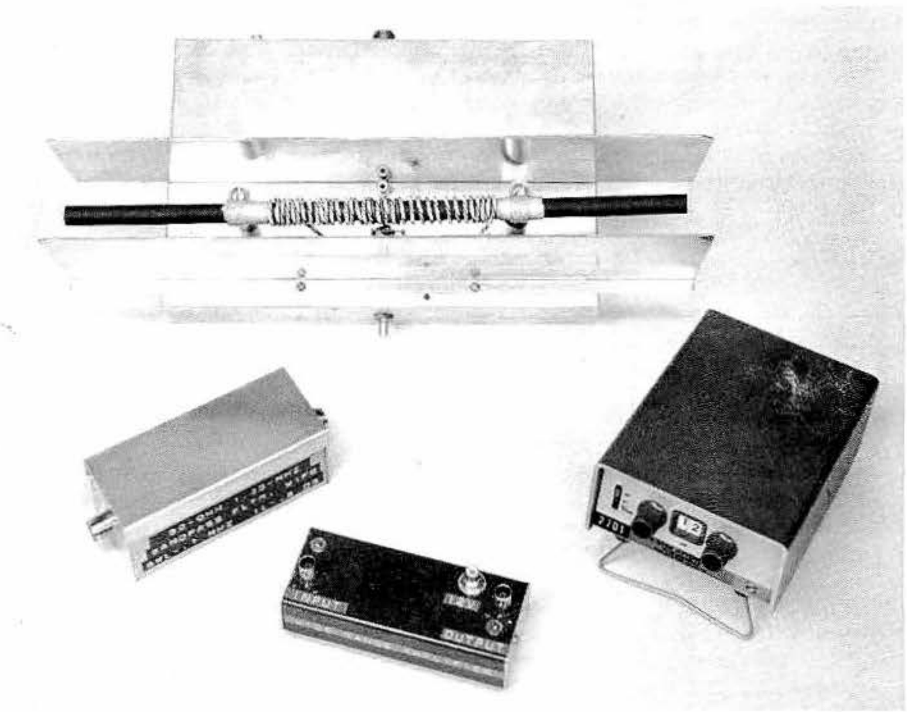

View of the ferrite-rod loop and test setup accessories. In the foreground (left) is the Butterworth band-pass filter. Below it at the center is the broadband preamplifier, and the Tektronix 2701 step attenuator is at the far right.

of the primary winding. The latter occupied approximately 3 inches (76 mm) of the center portion of the rod. A small aluminum chassis served as the foundation for the antenna, and the rod was spaced some 3 inches (76 mm) above the chassis. Performance tests showed the antenna to be less efficient than the large 5-turn wire loop. As referenced to the 75-foot (22.86-m) vertical, the ferrite loop was —38 dB. Furthermore, it was prone to the same induction effects which plagued the wire loop when both were used indoors. The rod-loop null was dismal — some 10 dB!

An improved model of the loop was built by using epoxy cement to join together a pair of 7-inch (178-mm) ferrite rods, end to end. The sensitivity (μV per meter) increases with the length of the rod, but is dependent upon the length of the coil which is wound on the rod. A multilayer coil contained at the center of the rod should be considerably better than a solenoidal coil spread over the major portion of the rod: The shorter the coil, the higher the Q of the antenna.

The author did not have Litz wire available, and was unable to build a multilayer coil, so a short solenoidal winding was placed on the center 5 inches (127 mm) of the 14 × 5/8-inch (356 × 16 mm) rod. It consisted of 32 turns of no. 12 insulated wire, spaced one wire diameter between each turn. The ferrite rods are made of Q2 material (μ = 125).

Fig. 4 shows the schematic diagram of the rod loop. A four-turn output link is wound over the center of L1. C1 is a mica compression trimmer which tunes the loop to resonance. C2 is used as a coupling capacitor to control the loop loading.

It seemed worthwhile to try shielding the loop, so a pair of aluminum L brackets were bolted to the chassis as shown in the photograph. Each bracket is 16 × 3-3/4 inches (406 × 95mm). The shields are spaced 1 inch (25 mm) either side of the loop.

The top of the U-shaped channel must be left open to prevent the shield from acting as a shorted turn. Proof of the effect can be seen by shorting across the open end of the trough (over the center of the rod) with a screwdriver. The received signal will vanish! An electrostatic shield, when used with a loop antenna, should minimize induced fields from nearby conductors and precipitation static, thereby making the system better for indoor use.

Rod-Loop Tests

The same outdoor test setup was employed for checking the performance of the rod loop. The nulls occur off the ends of the rod, and maximum response is found off the broad sides of the rod. A null of —23 dB was obtained with the loop, but efficiency was less than that obtained with the 5-turn wire loop. When compared to the 75-foot (22.86-m) vertical a reading of —31 dB was obtained, 13 dB inferior to the wire loop. The noise figure of the W7ZOI broadband preamplifier was too high for use with the new antenna, so a common-gate JFET pre-preamplifier was added between the loop output and the input to the 'ZOI unit. Ideally, a 30- or 35-dB preamplifier (two JFETs) should be used with the rod loop to assure a low noise figure and ample gain.

With the coupling set for maximum loop gain, the bandwidth (measured with an HP spectrum analyzer) is 10 kHz at the 3-dB points. The loop was measured independent of the shields and preamp for unloaded

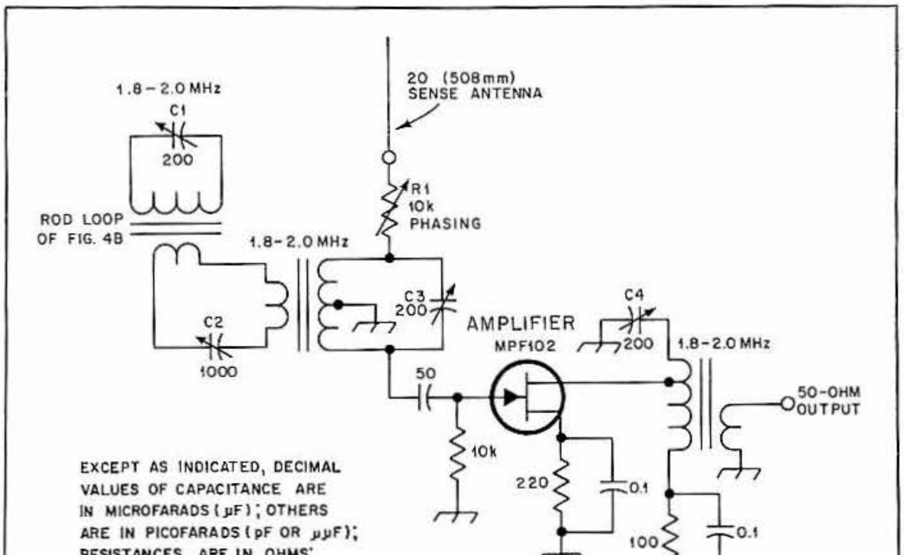

Fig. 5 — Suggested circuit for using a sense antenna with the rod loop. The sense antenna can be mounted 6 to 10 inches (152 to 254 mm) in front of the broadside of the rod loop (centered). C1, C2, C3 and R1 are adjusted alternately for the deepest null possible off the rejection side of the loop (side opposite the sense antenna). It may be necessary to experiment with the length of the sense antenna to obtain the best null in the cardioid response pattern. The tuned circuits for the JFET preamp can be wound on T-68-2 toroid cores.

Q by means of an HP Q meter. A reading of 430 was obtained.

The electrostatic shield cured the unwanted pickup of TV spurs and appliance transients during indoor use. The author would like to mention that the wire and rod loops are affected to some extent by the presence of the 75-foot (22.86-m) vertical. The latter is only 20 feet (6.10 m) away from the operating position. It tends to reradiate noise and signals. Some of that energy is picked up by the loop antennas. If an indoor loop is to be used regularly, or if it is outdoors, but within 100 feet (30.48 m) or so of the vertical, the latter should be shorted out or detuned during the receive period. An spst relay could be slaved to the regular antenna change-over relay to accomplish this.

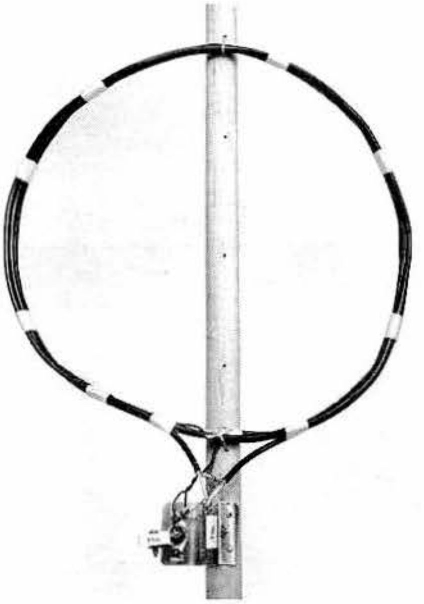

The 4-turn coax loop with electrostatic shielding. The trimmers are visible on aluminum brackets below the loop.

An improvement in performance for either type of loop could be had by adding a sense antenna and appropriate phasing circuit (Fig. 5). This would provide a cardioid pattern and result in a unidirectional response. Greater reduction of QRN and QRM would be effected through that technique. A weatherproof system of that type could be placed out-of-doors for year-round use. A TV type of rotator would be useful for changing the loop position, as desired.

The W1FB 4T-ES Loop

During a final investigation of loop types and performance characteristics, an attempt was made to devise a shielded loop small enough to be practical indoors. The primary objective was to develop a loop from ordinary materials — something that was inexpensive and easy to build. It should have efficiency comparable to that of the 5-turn wire loop described earlier in this report. A reasonable null depth and an undistorted pattern were sought also. From these criteria evolved the 4T-ES (four turn, electrostatic shield) loop shown in Fig. 6.

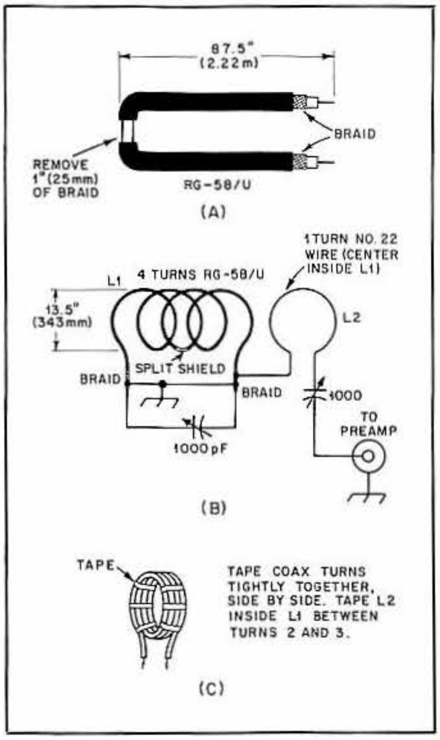

Fig. 6 — Structural details of the W1FB 4T-ES loop antenna. At A is an illustration of the total length of RG-58/U before the loop is formed. An expanded view of the loop turns and pickup link is given at B. Drawing C shows how the loop turns are taped together side by side as the antenna takes its final form.

It consists of four turns of RG-58/U cable which are formed into a circular configuration to provide a diameter of 13-1/2 inches (343 mm). There is nothing critical about the diameter, and RG-59/U coaxial cable could be used instead of the type specified. The turns are formed side by side, then taped at several points to maintain the loop form.

It is necessary to split the shield braid at the electrical midpoint of the loop. This prevents the shield from acting as a shorted turn. The principle is the same as that described for the ferrite-rod loop with its aluminum-channel shield. The braids are joined at the loop feed point and made common to one end of L2. The three grounded elements are then attached to the coaxial connector ground terminal on the loop support frame.

Loop performance was checked by using the test procedure discussed earlier. A null depth of —26 dB resulted. There was no discernible pattern warping, and the null occurred deeply over just a few degrees of rotation. The 4T-ES was 18 dB inferior to the 75-foot (22.86-m) vertical antenna when checking distant signals. The unloaded Q was measured by means of the HP Q meter. A value of 85 was obtained. Because C2 was set for light loading of the loop, a bandwidth (3-dB points) of 20 kHz was obtained when the preamplifier and receiver were connected to the antenna.

Numerous comparisons were made on DX signals from Europe and the Caribbean area during the ARRL DX Contest in March of 1977. In all instances the 4T-ES was as good as or better than the 75-foot (22.86-m) vertical. Ambient noise was considerably lower in amplitude while using the loop, even though all tests were performed with the loop indoors at ground level. There was no reduction in atmospheric noise (which was high that weekend) because the maxima of the loop favored the N.E. and S.W. directions: The storm front was S.W. of Connecticut. However, rotation of the loop caused a drop of 26 dB in the static crashes when the loop nulls were S.W. and N.E. Because of the electrostatic shielding there was minimum noise induction from the house wiring — a contrast to the performance characteristics of the unshielded 5-turn wire loop. The W7ZOI broadband preamplifier and an attenuator were used during the tests. There were some instances when the loop provided a stronger signal response than the vertical, even though the two systems were adjusted for unity gain.

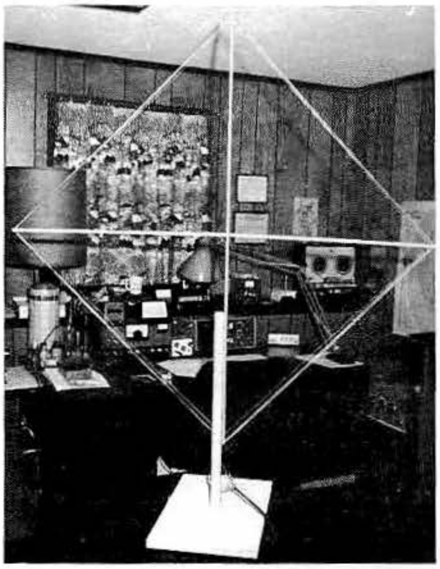

Close-up view of the 160-meter indoor receiving loop at W1FB. The antenna is resting on a piano bench, and is situated near the operating position so that it can be hand-rotated as necessary.

The 4T-ES is suitable for use indoors,

but it could be weatherproofed for outside installations. Epoxy cement could be used to seal the open ends of the coaxial cable of L1, and the trimmer capacitors would survive the natural elements if housed in a metal or plastic container.

In Summary

The loops that have been discussed here should be suitable for use on 40 or 80 meters if scaled accordingly. They may not show much directivity on sky-wave signals because of the "tumbling effect" at the higher frequencies. But QRM from nearby amateurs could be reduced greatly, and man-made noise sources could be rejected significantly by correct orientation of the loop.

The author has attempted to discuss some of the characteristics of small loops which are not treated in most of the references contained in amateur journals. There is plenty of latitude remaining for experimentation, and the reader is encouraged to apply his or her skills in developing better loops for amateur use. Meanwhile, why not build a small loop and scoop those weak signals out of the noise? It could mean some new countries for you on "top band" or 80 meters. This material was originally presented in *QST* by Doug DeMaw, W1FB.

References

[1] DeMaw, "A Receiving Loop for 160 Meters," *QST*, March 1974.

[2] Terman, *Electronic & Radio Engineering*, McGraw-Hill. Kraus, *Antennas*, McGraw-Hill. Jasik, *Antenna Engineering Handbook*, McGraw-Hill. Keen, *Wireless Direction Finding*, 3rd Edition, Iliffe & Sons, Ltd.

[3] DeMaw, "His Eminence — the Receiver," Parts I and 2, *QST*, June and July 1976. Receiver used in combination with an Autek Research QF-1 RC active audio filter.

[4] DeMaw, "Build This Quickie Preamp," *QST*, April 1977.

The Half Square Antenna

Do those 100 countries still elude you for 80-meter DXCC? In terms of dollars versus decibels, a half-square antenna may be your best bet for putting punch into your signal.

With 5-band DXCC as bait, considerable interest has arisen in DX and antennas for the lower frequency ham bands. The author's attempts to get a decent 80-meter DX antenna have gone through an evolutionary cycle with a lot of iterative tries and with steadily improving results. The half-square antenna that is the latest of these has some interesting properties and gives about as many "dBs per buck" as any this author has seen.

Before describing this configuration, it might be interesting to cover briefly the different antennas tried which didn't work as well. The writer's available options were set up by (1) a yard full of tall trees and (2) very little money to spend. The first antenna was an inverted V hung on the 75-foot (22.9-m) tower. This is really a quite good antenna — primarily because it has the high current part of the antenna at the highest support point. It provided a standard against which all succeeding antennas were checked. The next step was to add a parasitic element (director) which didn't seem to help at all. This is perhaps not so surprising with the ground so close. Viewed simply, normal parasitic coupling is dependent upon a fair proportion of the current in the driven element being coupled to the parasitic element. When they are both close to the ground, the reverse polarity of the image of the driven element starts to couple heavily to both the driven element and the parasitic element, making it practically impossible to get the desired parasitic current at any useful spacings.

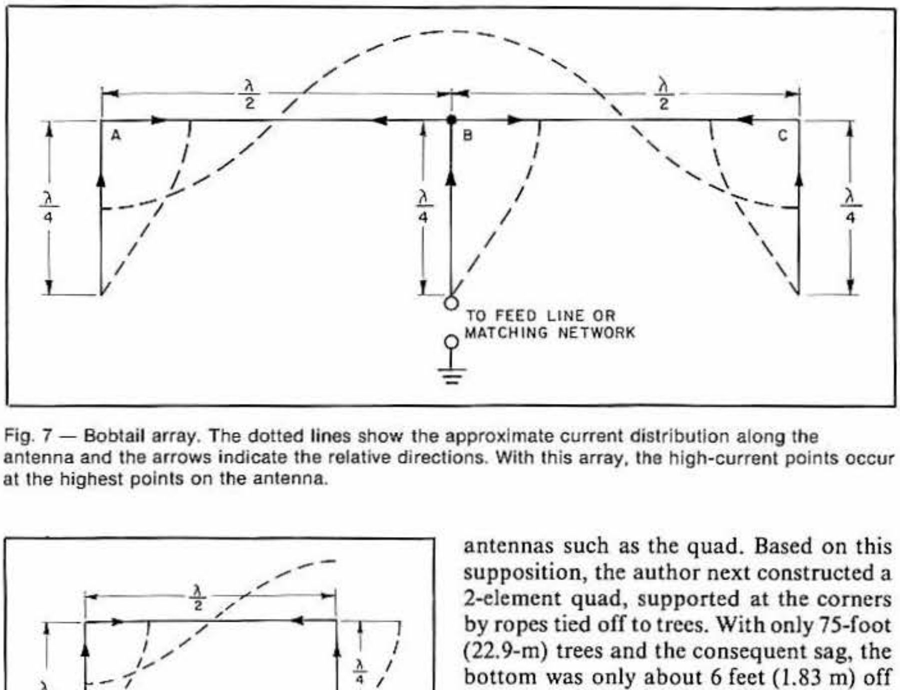

Fig. 7 — Bobtail array. The dotted lines show the approximate current distribution along the antenna and the arrows indicate the relative directions. With this array, the high-current points occur at the highest points on the antenna.

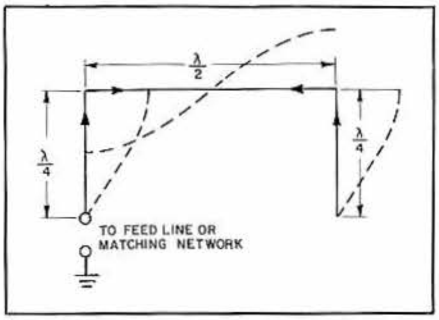

Fig. 8 — The "half-square" evolved when one leg broke off the bobtail. The input impedance of both the bobtail and half-square antenna is very high, which eliminates the need for an elaborate ground system at the feed point. Either a parallel-tuned circuit or a quarter-wave transformer can be used to match to a 50-ohm feed line.

This effect has been shown experimentally to be less with closed-loop type antennas such as the quad. Based on this supposition, the author next constructed a 2-element quad, supported at the corners by ropes tied off to trees. With only 75-foot (22.9-m) trees and the consequent sag, the bottom was only about 6 feet (1.83 m) off the ground. However, the quad did work, had an acceptable feedpoint impedance, and gave a few dB (about 1/2 an S unit) improvement over the inverted V on European contacts. Obviously, the only useful low-angle radiation in this antenna came from the upper half of it, the lower half serving to feed or excite the upper half. A major problem with the quad was keeping it up. It had to be very tightly stretched between trees to keep it from sagging to the ground and as a result the wire often broke.

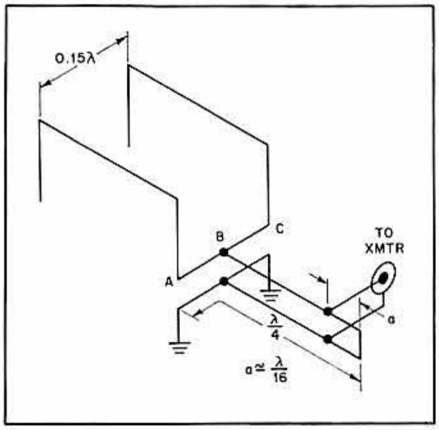

Fig. 9 — Two-element half square array gives greater gain and a very wide bandwidth.

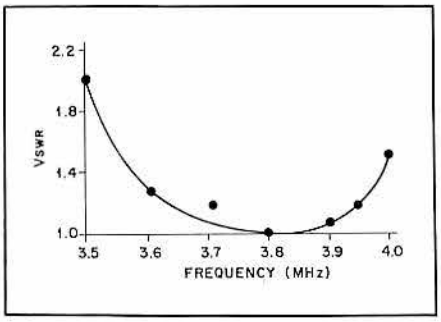

Fig. 10 — Measured SWR of the antenna shown in Fig. 9.

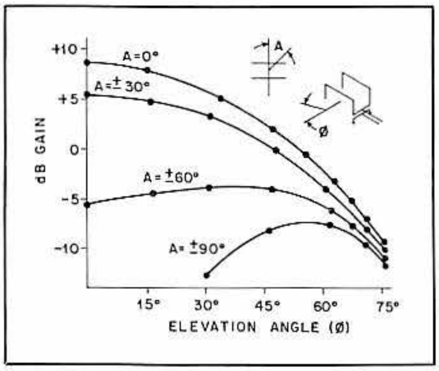

Fig. 11 — Computer-pattern plot for a 2-element half square array with 0.15-wavelength spacing between elements. The angles referred to in the drawing are shown in the inset.

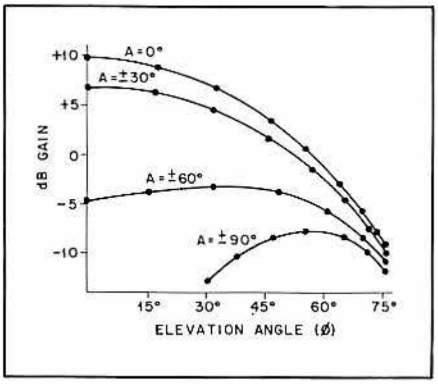

Fig. 12 — Computer-pattern plot for the antenna of Fig. 11 with 0.25-wavelength spacing.

To try another tack which would avoid the parasitic coupling problem an 8JK array with four half-wave elements was tried, all elements being fed. With the span of this being so long and with lots of high-voltage points which had to be kept from touching the trees, the author only managed to get it to a height of about 40 feet (12.19 m). The signal-report results with this were essentially identical to the quad on the path to Europe. The horizontal beamwidth was noticeably less than the quad. It probably would have done better if the writer could have gotten it to 75 or 80 feet (22.86 or 24.38 m) in height.

The next antenna tried was the so-called bobtail which is shown in Fig. 7. This is a broadside vertical array with approximately twice the current in the center leg as the two outer legs and with current distribution in the horizontal wires which tend to cancel horizontally polarized emission.[1] A major reason for trying this was the author's frustration with trying to use the maximum available height of the trees and still be able to feed the antenna and keep the high-voltage points clear of the tree branches. The author had avoided vertical antennas before because of the difficulty in getting a good low-loss ground. Also, the trees looked like great vertically polarized rf energy eaters. The bobtail is particularly nice for tree hanging since its highest support points (A, B, and C in Fig. 7) are all at low impedance and looked as if they could be directly slung over a tree limb with no insulators. It is fed from the bottom at a high-impedance point so the ground system doesn't need to be as good as it does with low-impedance feedpoints. Also, like the inverted V, the bobtail has its maximum current at the highest support point. It can be excited from a parallel-resonant tank circuit or a quarter-wavelength stub.

The bobtail has been an outstanding performer. It's at least an S unit better than the inverted V and works great on DX paths to the Far East (JA, KG6, and VS6). The author put up two of these at right angles and found their interaction to be practically nil even though their center legs are quite close. When switching, the author ground the feed point of the one not in use.

The Half-Square Antenna

With such fine results, it was thought the ultimate "tree-hung" antenna had been discovered. However, after a particularly bad storm, it was noticed that one leg of the bobtail had broken off and yet the antenna feed impedance hadn't changed noticeably. Furthermore, the results compared with the inverted V seemed about the same as the full bobtail. Fig. 8 shows the configuration. It is a 2-element (instead of 3) vertical broadside array with the same current in both legs. It has all the advantages of the bobtail except that the radiation from the horizontal part of the antenna doesn't cancel as well in all directions and some end-fire, horizontally polarized energy does spill off. This might be considered an advantage or a disadvantage, depending on whether you're seeking coverage or discrimination! In discussing this with Ed Watters, WA3LGX, he pointed out the kinship this antenna has with the bisquare array. It's sort of half of a bisquare laid sideways with the ground image antenna making up the other half. That, plus its obvious shape, led the author to dub it the half-square antenna.

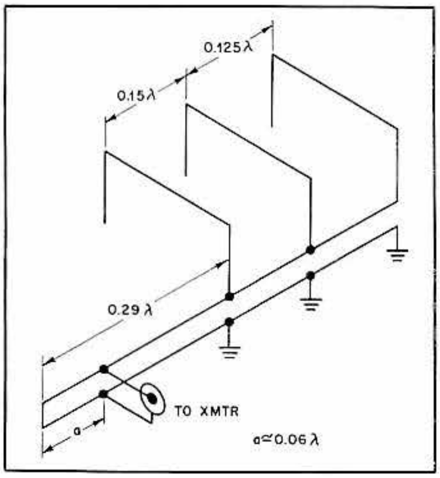

Fig. 13 — A 3-element half-square array.

After using a single half-square for awhile, it was decided to add another element with about 0.15-wavelength spacing. With the new element in place, a check was made to see what parasitic coupling there might be between elements. With several hundred watts fed into the driven element there was no evidence of rf on the parasitic element (using a neon bulb). This was a very gross measurement which tended to confirm what was expected. The very close proximity of the antenna to ground tends to reduce greatly parasitic coupling to a nearby element. Being puzzled why it should appear so small, the only hope was to devise a means of measuring this coupling more accurately.

Having had to tear down the 8JK antenna to get enough wire for this antenna, as a first try for feeding it, the author just used the 8JK open-wire line and stub match as it existed. The one wavelength of wire for each element was taken directly from the 8JK also. Fig. 9 shows the configuration. Unlike the 8JK, which was narrowband (centered around the middle of the band), this configuration gave the very flat SWR shown in Fig. 10 — on the first try! The SWR was not very different whether the feed-line stub was connected to point A, B, or C (Fig. 9). B gives a bidirectional characteristic while connection to A or C gives some unidirectivity. The author is still suspicious of the wide bandwidth and suspects the two elements may be different effective lengths, giving a stagger-tuned effect. If the spacing between these two were increased to a quarter wave and fed at point A or point C, it should theoretically add another couple of dB gain

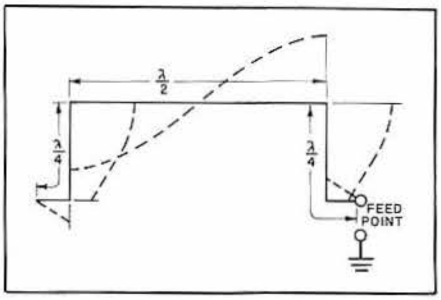

Fig. 14 — Any of the antennas mentioned can be bent as shown in order to keep the current loops at the upper corners.

and give a more significant front-to-back ratio than with the 0.15-wavelength spacing. WA3LGX plugged the dimensions and shape of this antenna into a generalized computer program he has developed for wire antennas and got the curves shown in Figs. 11 and 12. Fig. 11 defines the forward pattern for 0.15-wavelength spacing assuming a perfectly conducting earth. Fig. 12 is for quarter-wave spacing. The actual patterns will be modified at angles below 10° to 15° by the earth's resistivity[2] as is the case with any vertical antenna. The 0.15-wavelength-spaced antenna was used for some time and consistently gave better results than the bobtail on the path to Europe.

The writer took one other step beyond the 2-element array and added another element at 1/8-wavelength spacing (see Fig. 13). The extra element is fed, as are the other two, with an extension of the same open wire line. This antenna is considerably quieter (less QRN) than any of the other 80-meter antennas the writer had and has better than half an S unit advantage over the bobtail on European signals. It doesn't require any retuning to cover the whole band.

With these antennas not insulated at their support points, one might expect some effects when the tree sap rises or in wet conditions. The author hasn't been able to notice any differences from dry weather to monsoon-type rains or wet snow covering everything. The trees are slightly over 1/4-wavelength high on 80 meters. Considered as antennas or absorbers, they would look like grounded 1/4-wavelength vertical radiators. The vertical sections of the half-square antenna that run close to the trees are just the opposite. That is, at the bottom of the antenna the voltage is high and at the top, the voltage is low. This should minimize coupling to the trees, as opposed to, for example, using 1/4-wavelength grounded verticals for the antenna elements.

160-Meter Operation

Now if you look closely, you can see that the 80-meter half-square whether 1, 2 or 3 elements, should make a good antenna for 160 meters. Each element is 1/2-wavelength long on 160 meters, voltage fed on the end, and the high current part of the antenna is at the maximum height. True, a higher impedance part of the antenna touches the tree, but you can't have everything free! While not having managed to work any DX with the few watts available on 160 meters, the author can *hear* Central European ham stations quite well on this band for the first time ever at this location. Of course, a different stub or tuning arrangement is necessary for getting a low SWR on 160 meters. The writer just unshorted the stub and put a tapped coil (about 14 microhenries) in series with the coax feed line to tune it.

Construction Details

These antennas were all built with no. 14 soft-drawn copper wire. The elements are pulled over the top limbs of trees that have approximately the correct spacing and tied to nylon strings at about shoulder level. If your trees aren't a full quarter wave high, the extra wire can be stretched out parallel to the ground at each end as shown in Fig. 14. Note that you still get the maximum current point at the highest support point you have.

The best technique found for getting a line over the highest tree is to use a fisherman's casting reel (I use a Zebco 202) with an 8-pound nylon line and a 1- or 1-1/2-ounce sinker. The reel is put in the RELEASE position and a sling shot is used to shoot the sinker and line over the desired tree. This is a very accurate technique and you can easily put a line over 90-foot (27.43-m) trees. The light fishing line is used to pull over a heavier line. It's a good idea to paint the sinker a bright color since the line is almost invisible. You can waste a lot of time looking for the sinker at the end of the line.

Be sure to leave enough droop in the horizontal section of the antenna to allow for tree motion in high winds. The open-wire feed line used for the harness and stub sections is made with the same no. 14 wire spaced 3-1/2 inches (89 mm).

The avid Field Day or camping fan will notice the obvious advantages of the half-square antenna for that type of operation. It uses minimum materials, goes up very quickly, will really put out a booming signal on 80 meters, and can be taken down in just a few minutes. You can also make good use of the fishing rod in between your turns at the key. This material was originally presented in *QST* by Ben Vester, K3BC.

References

[1] [Editor's Note: While complete cancellation of the horizontally polarized component occurs at a point broadside to the array, some radiation from the flat-top portion will exist at other angles. This is because of incomplete cancellation of the components from the opposite ends of the flat top.]

[2] *The ARRL Antenna Book*, 12th Edition, p. 46.

The HRH Delta-Loop Beam

Here's an antenna that should excite the interest of the amateur fraternity. The Delta-Loop beam has all the features of a quad without many of the disadvantages.

The evolution and design of the antenna shown in the photographs came about primarily because of a lack of confidence in the mechanical durability of quads. Anyone who has ever built and used quads knows how difficult it is to keep them up in icing and high wind conditions.

The author being an old-time mechanic, felt there had to be a better design for an antenna that used full-wave elements in a parasitic array, so some years ago a start was made to find that better design. One thing that nearly all readers will agree is that Mother Nature is about the best designer there is. Along these lines, just observe the growth of branches on tree limbs. You'll find that many of the smaller branches will grow up in a semivertical pattern from the limb; see Fig. 15. It takes a considerable amount of icing and wind to break such branches, indicating the soundness of the design. This same idea should apply to elements mounted on a boom.

To test the idea, an element of aluminum tubing was mounted on a boom. The element was mounted as shown in Fig. 17. The boom was then secured to the family car with the element in a vertical position. The assembly was then tested at 65- to 75-mph. Even though the element was only secured to the boom as shown in Fig. 17, there was no "set" or permanent bending. This test opened the door to several possibilities in beam design.

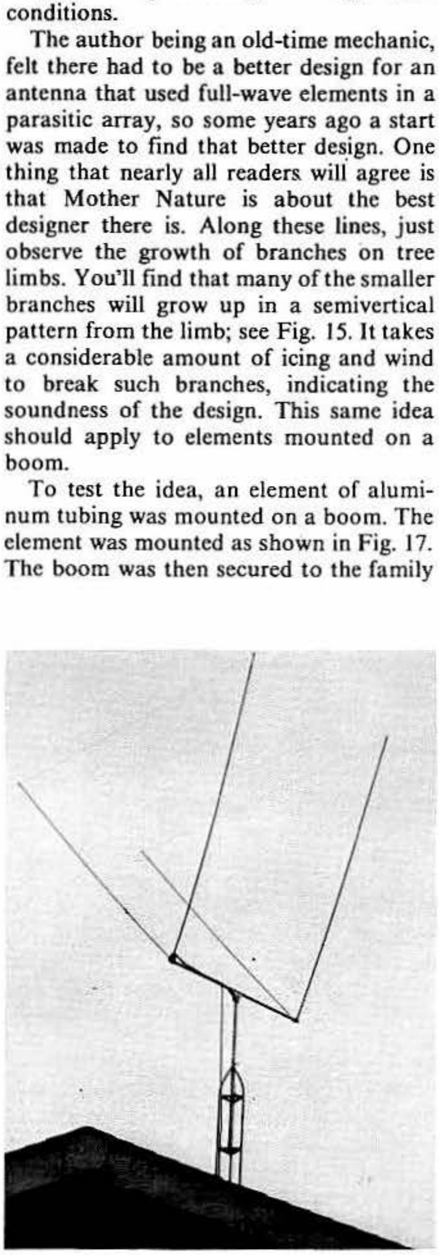

This is a 10-meter, 2-element Delta Loop that is very similar in construction to the antenna described in the article.

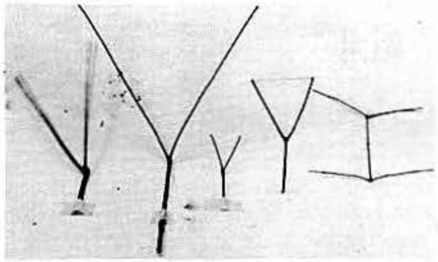

Fig. 15 — The basic idea for the Delta-Loop beam was conceived from the growth of tree limbs and branches. Here are just a few of nature's configurations that lend themselves to antenna design!

A loop aerial can take many shapes. Anyone who studies the history of the quad will find that many different configurations have been used. Why not use a triangle?[1] It should be possible to achieve a very good design using a vertical triangle. See Fig. 16. This shape appears ideal for a beam with full-wave elements, and has most of the advantages of the quad without many of the disadvantages. This led to the construction of several antennas of the type shown in the photographs, and some of these advantages become quickly apparent.

First, the entire antenna is *above* the boom. Second, the antenna is constructed primarily of aluminum tubing, which provides extra strength as compared with wire elements. Third, the antenna has Plumber's Delight type construction, meaning that the antenna is at ground potential for lightning protection, plus the fact that this type construction lends itself to gamma matching of the feed point, eliminating the need for a balun. Last, but not least, we find the antenna is very attractive in its symmetry. Of course, the important point is how well does the antenna perform?

Several 10-meter models have been tested and the unit shown in the roof-mounted position is 20-meter Delta Loop. Using a gamma match with 50-ohm coaxial feed, a 10-meter model was matched at 28.8 MHz. The antenna consisted of a driven element plus reflector with the elements spaced 6-1/2 feet (2 m) apart, or approximately 0.2 wavelength. Surprisingly, the antenna was extremely flat across the entire 10-meter band, the worst mismatch being *less* than 1.2 to 1.[2]

The author has no means of measuring gain or front-to-back ratios. However, with the spacing and element sizes the same as a quad, the gain should be the same, or at least so close the difference would be insignificant. On-the-air tests have shown the antenna to be as good as or better than quads the author has had up. Front-to-back and front-to-side reports have been outstanding, both on ground wave and skip.

Element Lengths

Many tests have been made on driven elements to determine the effect of element lengths on SWR. It was found that the flattest curve was obtained with the formula

$$\frac{1005}{f_{MHz}} \text{ (feet)}$$

for the driven element. The reflector should be about three percent longer, or

$$\frac{1030}{f_{MHz}}$$

If directors are desired it would appear that a length three percent shorter than the driven element would be adequate.

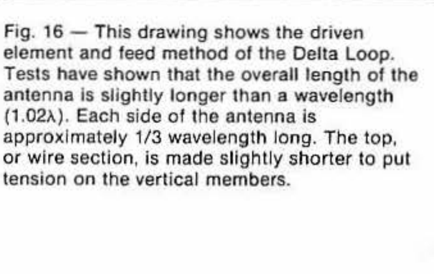

Fig. 16 — This drawing shows the driven element and feed method of the Delta Loop. Tests have shown that the overall length of the antenna is slightly longer than a wavelength (1.02λ). Each side of the antenna is approximately 1/3 wavelength long. The top, or wire section, is made slightly shorter to put tension on the vertical members.

Table 1

Bill of Materials

Tubing used is 6061 T6.

4 — 12-foot lengths of 3/4-inch OD, 0.035 wall (elements).

1 — 12-foot length 1-3/4-inch OD, 0.065 wall (boom).

8 — 3/4-inch diameter stainless steel hose clamps.

1 — 9-inch length of 3/8-inch OD by 5/16-inch ID.

1 — 9-inch length of 1/2-inch OD to slide over 3/8-inch section.

25 feet of copper wire no. 12 or 14, or equivalent in stranded wire.

Feet X 0.3048 = m.
Inches X 25.4 = mm.

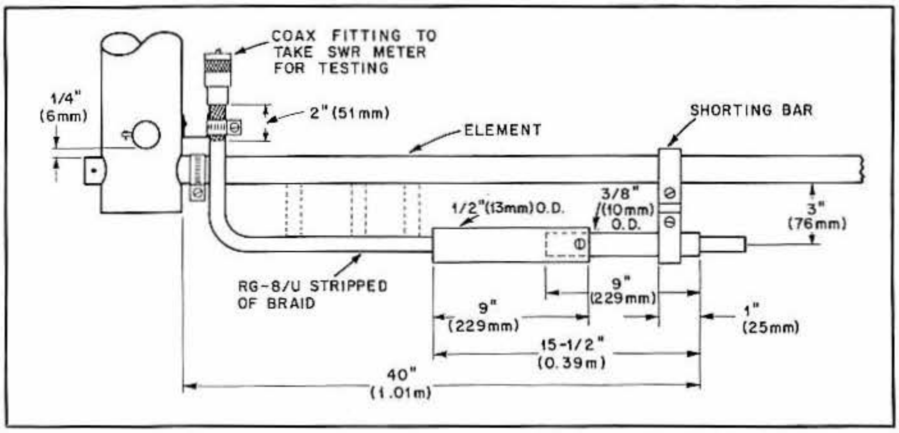

Fig. 18 — Here are the details on the gamma-matching section. As mentioned in the text, the coax fitting can be a female type SO-239, mounted on a small bracket which is mounted on the boom. Spacing insulators for spacing the gamma line from the element can be made up from pieces of Lucite or Plexiglas.

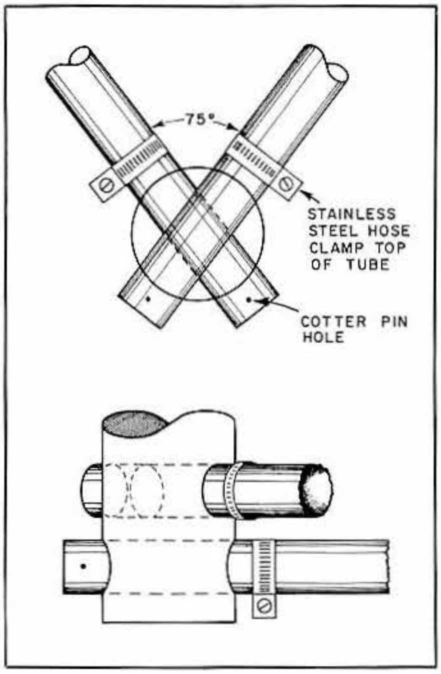

Fig. 17 — This drawing shows the method of mounting the element to the boom. A cotter pin is used on the under side of the boom and a hose clamp on the top to hold the element securely to the boom.

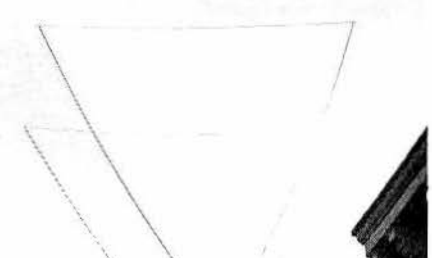

Here is our roof-mounted 20-meter Delta-Loop beam that is now undergoing a series of tests.

As to element spacing, the author has long been an advocate of wide spacing (0.17 to 0.2 wavelength) whenever possible. In the 10-meter tests 0.2 wavelength spacing was used.

For readers who are interested, complete construction information on a 10-meter model, including a list of the required tubing, follows.

Construction Notes

Fig. 17 shows the method of mounting the element tubing to the boom. The angle between the semi-vertical elements is shown as 75 degrees but this can vary a few degrees either way without any appreciable effect on the performance of the array. It is difficult to drill and line up the element support holes in the boom and come out to *exactly* 75 degrees.

There are a couple of methods of making the holes in the boom to hold the elements. One of the simplest is to use a Greenlee-type chassis punch. This makes a clean hole in the boom. Another method is to drill holes large enough to take the end of a 3/4-inch (19-mm) reamer and then ream out the holes to the necessary diameter.

One not accustomed to working with angles might find it difficult to drill holes at a 75-degree angle. A simple method is to first drill holes to take a single element. Next, make a jig from a piece of stiff cardboard or similiar material, using a protractor to get the 75-degree angle. Then insert the single element into the boom and lay the jig along the element; this will give you the correct alingment for the other element.

The elements are inserted through the boom just far enough to take a cotter pin on the underside of the boom. On the top side, a hose clamp is used to hold the element to the boom and prevent slippage. A length of copper wire can be wrapped around both elements just below the clamp to insure a good contact between the elements and the boom. Hose clamps are also used at the tops of the elements to hold the horizontal wire.

Fig. 18 shows the details of the gamma matching section. The gamma section is made up from the inner conductor of the coax, including its insulation, and two sections of telescoping aluminum tubing. The outer covering and braid is removed from a 52-inch (1.32-m) length of RG-8/U coax. In Fig. 18, a male coax fitting is shown with a short length of braid, 2 inches (51 mm) long, which is grounded to the boom via a clamp and bracket. If desired, a chassis-type coax fitting, type SO-239, could be mounted on the boom with a small metal bracket. The capacitor section is made of two lengths of tubing 3/8-inch (10 mm), and 1/2-inch (13 mm) diameter, respectively, and both 9 inches (229 mm) long. The 3/8-inch (10 mm) tubing fits over the insulation around the coax inner conductor while the 1/2-inch (13 mm) tubing slides over the 3/8-inch (10 mm) material. The larger tube is drilled and tapped to take a locking screw. Spacing of the section from the element is 3 inches (76 mm).

In adjusting the gamma, figures given in

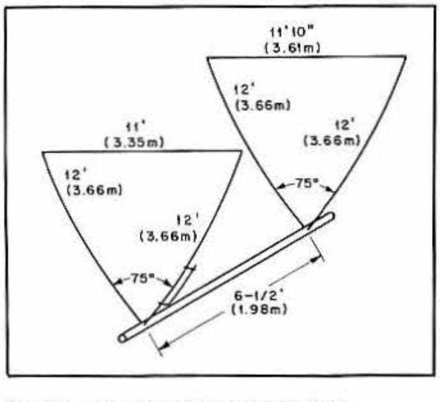

Fig. 19 — For those interested in exact figures, this sketch provides the element and boom lengths for a Delta-Loop beam for 28.8 MHz.

Fig. 19 can be used as a guide. In order to avoid matching errors, the gamma should be adjusted with the SWR bridge right at the beam. Set the shorting bar to the dimension given in the diagram and slide the 1/2-inch (13 mm) tubing over the smaller tubing, looking for a setting that gives a match. If one cannot be obtained, move the shorting bar and sliding tubing assembly a short distance, say an inch (25 mm), and then try different settings of the 1/2-inch (13 mm) tubing until you find a match. We found that the settings were not critical in getting a match. Once the beam is matched, the regular feed line to the station can be attached to the feed point. The connection should be taped to prevent moisture from getting into the fitting and coax.

As stated earlier, the antenna is extremely flat across the entire 10-meter band. The present model was designed for 28.8 MHz and matched at this frequency. The SWR at the highest point, at 29.7 MHz, was about 1.2 to 1. This dropped to 1 to 1 at about 29 MHz and stayed at 1 to 1 all the way down to 28 MHz. This feature of the antenna certainly is a help when using a transceiver designed for 50-ohm output. The beam has proved its ruggedness through winds and icing conditions. All in all, we think the Delta-Loop beam is one of the best performers we have seen. This material was originally presented in *QST* by Harry R. Habig, K8ANV.

References

[1] [Editor's Note: A triangular loop configuration also was described in *QST* in April 1968, "Technical Correspondence" by Norman Watson, W6DL, independently of K8ANV's design. Mr. Habig has been working for some years on this design and has a patent pending.]

[2] [Editor's Note: A 10-meter model of the Delta Loop beam tested at ARRL headquarters was matched at 28.8 MHz and an SWR curve was made using a Bird Wattmeter. The largest mismatch also was less than 1.2 to 1 MHz (at 29.7 MHz).]

The Delta-Loop Beam on 15

DX capability, low SWR and readily available construction materials are highlights of this 15-meter beam antenna. Similar to a quad, but rugged as a Yagi — those are the virtues of the Delta Loop.

Some of the features of the 15-meter beam described in this article include a very low SWR across the entire band, use of readily available materials, "Plumber's Delight" type construction — which is always an appealing feature of any beam antenna — and last but not least, excellent performance. On the last point, the antenna shown in the photographs was tested during a DX contest with a transmitter input power of 100 watts. The boom of the antenna was mounted only eight feet (2.44 m) above the ground, but in about five hours of contest operation 58 different countries were worked — and, believe it or not, the majority of the reports received were 59. Possibly conditions were above par, but even so the antenna shows it can do a real job under crowded band conditions.

As stated above, the beam uses Plumber's Delight type construction. For the newcomer's information, this type of construction has the antenna elements connected directly to the boom which in turn can be connected directly to earth ground without having any effect on the performance of the antenna. This is desirable because of the lightning protection offered by the system.

Fig. 20 is a drawing of the Delta-Loop driven element. The overall length around the driven element is slightly over one wavelength. Each side is approximately 1/3 wavelength long. The reason we say "approximately" is that the top section of the antenna is made of wire and is slightly shorter than 1/3 wavelength in order to put tension on the wire. A gamma-match feed system is used, eliminating the need for a balun.

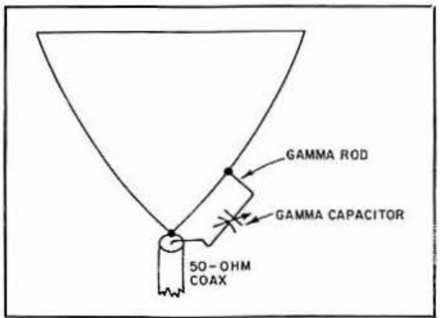

Fig. 20 — Basic configuration of the Delta-Loop driven element. Each side is approximately 1/3-wavelength long.

The formula used in figuring the driven element length is

$$\frac{1005}{f_{MHz}},$$

where f is the desired resonant frequency of the driven element. The answer will be in feet. The reflector length formula is

$$\frac{1030}{f_{MHz}}.$$

For a frequency of 21.2 MHz this figures out to a driven element 47 feet 5 inches (14.45 m) long and a reflector that is 48 feet 7 inches (14.81 m) long. Element spacing is approximately 0.2 wavelength or about 9 feet (2.74 m). Fig. 21 shows the two elements with the lengths of each segment.

Material Requirements

The beam elements were built primarily from Reynolds Do-It-Yourself aluminum tubing, which *should* be available from most hardware dealers. The reason we say "should" is because in some sections of the country hardware stores don't stock the tubing. However, most dealers will order the tubing for you. If you live in an area near or in a city of reasonable size, a look through the Yellow Pages will show any aluminum tubing dealers. The Reynolds

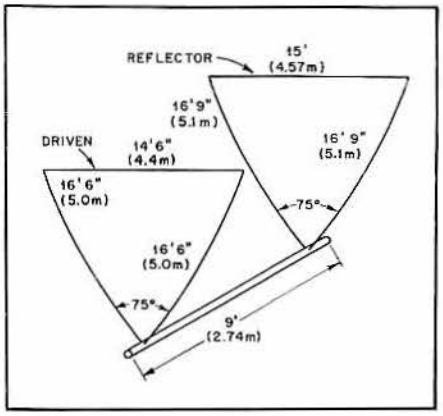

Fig. 21 — The element dimensions and the boom length. All measurements on the elements are made from where the element enters the boom; the portion of the element extending through the boom is not counted.

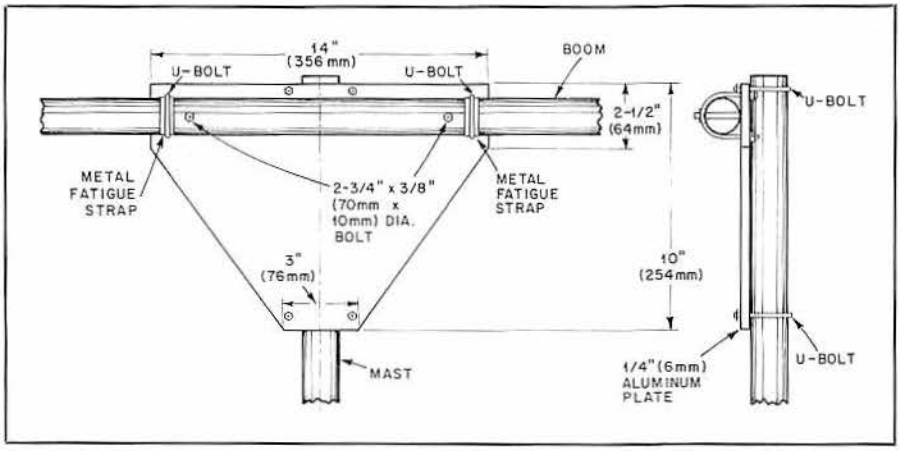

Fig. 22 — Construction details for the boom-to-mast mounting bracket. The U bolts holding the boom are 2-inch (51-mm) type. Those holding the mast will depend on the mast diameter — we used 1-1/2-inch (38-mm) U bolts for a 1-1/2-inch diameter mast.

Table 2

Shopping List

Quantity	*Length (ft)*	*Diameter (in.)*	*Reynolds no.*
4	8	1	4242
4	8	3/4	4222
2	6	7/8	4231
1	9	2	See Text.

3 feet of 3/8- or 1/2-inch diameter aluminum tubing.

12, 1-inch diameter hose clamps, stainless steel.

35 feet of no. 12 or 14 copper wire, solid or stranded.

1 variable capacitor, 100 pF maximum, 0.025 spacing or greater.

1 one-quart freezer container.

1 SO-239 Coax chassis connector.

5 1/4-inch diameter aluminum nuts and bolts, 1-1/4-inches long.

Feet X 0.3048 = m.
Inches X 25.4 = mm.

Do-It-Yourself tubing comes in 8-foot (2.44-m) lengths but most of the commercial tubing available from metal distributors comes in 12-foot (3.66-m) lengths. If you obtain the 12-foot types, be sure to specify type 6061 (61S) alloy as this material affords excellent strength for antenna elements. The boom used in this antenna was made from this alloy and is 2 inches (51 mm) in diameter with a 0.065-inch (1.65-mm) wall. Another source of aluminum or steel tubing is electrical supply houses; although this type of aluminum is a little too soft for the elements it would be suitable for a boom. Also, electrician's thin-wall steel tubing is available in various sizes, including the 2-inch diameter, for a boom. In a pinch, the boom could be made of wood, such as a length of 2 × 4. In such a case, the bottoms of the elements could be connected together with a length of wire to maintain the Plumber's Delight feature. The shopping list included here should prove of help in purchasing the parts.

Construction

The first step in building the antenna is to make up the vertical elements. The description given here is based on using the Reynolds tubing. Using a hack saw, cut the 7/8-inch (22-mm) tubing into 2-foot (0.61-m) sections. This particular type is made by Reynolds for the purpose of telescoping the 1- and 3/4-inch (25-mm and 19-mm) diameter sections together and comes in 6-foot (1.83-m) lengths. After cutting off four 2-foot (0.61 m) lengths you'll have one 4-foot (1.22-m) piece remaining. Cut this into four 1-foot (0.31-m) lengths and set these pieces aside for the time being.

Next, cut two lots, the thickness of the hacksaw blade, about 2 inches (51 mm) deep in one end of each of the 1-inch (25-mm) and 7/8-inch (22-mm) diameter lengths. Slide the 7/8-inch tubing inside the 1-inch diameter to a length of about 6 inches (152 mm) and then slip the hose clamps over the slots and tighten them. The 3/4-inch (19-mm) tubing can then be inserted into the slotted ends of the 7/8-inch (22-mm) tubing and these sections clamped with hose clamps. Two of the elements should be adjusted to the proper length for the driven element and two for the reflector, as per Fig. 21 (or whatever length you wish from the formulas). Some amateurs may prefer to cut their antennas for the cw portion of the band and others for the phone section.

The next step is to cut the 1-inch (25-mm) holes in the boom to take the ends of the 1-inch diameter elements. We used a 1-inch chassis punch to make the holes, but a 1-inch metal reamer could also be used. The aluminum is easy to cut or punch so making the holes should be no problem. What can be a problem is lining up the holes so that the 75-degree angle between the elements is obtained. We made a jig from a piece of cardboard, using a protractor. Once one set of holes for the director and reflector were made the elements were temporarily mounted in the boom. The cardboard jig was laid on the elements and the 75-degree holes were marked off and drilled.

When the holes are all drilled, the boom can be laid on the ground and the element ends inserted into the boom holes. We allowed about 1/4-inch (6-mm) extension of the base of the elements through the boom so as to allow space for cotter-pin holes. However, before drilling the cotter-pin holes, insert the four 1-foot (0.31-m) lengths of the 7/8-inch (22-mm) tubing into the elements at the boom. This will serve to give added strength to the elements at the support point. Drill each of the four ends to take the cotter pins and install the pins. Tighten down the hose clamps on the top side of the boom to secure the elements in place.

Drill the tops of the elements to take the 1/4- × 1-inch (6-mm × 25-mm) aluminum bolts. The top wires can then be installed.

The close-up view shows the gamma installation. The gamma capacitor is mounted inside a plastic freezer container which is held in place by a small metal L-shaped plate mounted on the box and the boom, using four self-tapping screws. This plate also holds an SO-239 coax chassis fitting to take the feed line. We like to install an SWR bridge directly at the beam when making SWR or matching tests to reduce matching errors, and having the fitting there simplifies the procedure. The gamma rod, made from a length of 3/8-inch (10-mm) aluminum tubing 36 inches (0.91 m) long, is flattened at one end for a length of 1-1/2 inches (38 mm).The flatted portion is bent over at right angles and drilled to take one of the 1/4- × 1-1/4-inch (6-mm × 32-mm) aluminum bolts, which is mounted through the top of the freezer box. The gamma rod is held in place by this bolt and by the shorting bar between the rod and the element. The shorting bar is made from a piece of aluminum, 1 inch (25 mm) wide and long enough to fit around the two pieces of tubing and provide a separation of 3 inches (76 mm) between the rod and the elements.

How you mount the antenna boom to the supporting mast will depend a great deal on what type of tower or supporting structure you have. The method shown in the photo is a fairly simple one, and easy to make up. Fig. 22 shows details of the construction of the mounting unit. The one shown in the photo supporting the beam was made from two pieces of 1/16-inch (2 mm) thick aluminum plate. However, we found that even with the double plate there was still too much "give" in the mounting setup, so a replacement was made from 1/4-inch (6 mm) thick aluminum. This made the mount competely steady. Either a plate of 1/4-inch aluminum or 1/8-inch (3 mm) steel should be adequate for the jobs. The 2-inch (51 mm) U bolts that hold the boom to the mounting plate were adequate, but just for added insurance, 3/8-inch (10 mm), 2-3/4-inch (70 mm) long machine bolts (two required) were installed through the boom and mounting plate for greater holding strength.

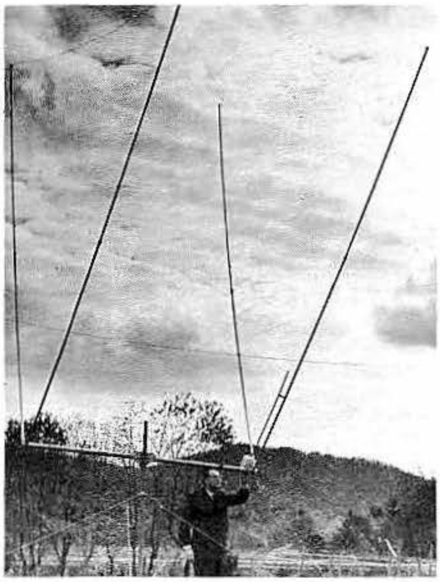

This view shows the completed antenna being matched.

During our tests, a weather front moved through the area and we had wind gusts up to 40 mph plus icing and snow loading on the antenna. The antenna and elements showed very little movement in these winds, and the entire antenna appeared very strong.

Adjustments

Adjustment of the gamma section is quite simple. Install an SWR bridge at the gamma feed point and tune up the rig on the desired frequency, Set the gamma shorting bar 24 inches (0.61 m) from the base of the gamma rod and tighten the gamma-bar nuts and bolts just enough to make electrical contact. Set the SWR bridge to read reflected power and adjust the gamma capacitor for a dip. The object is to find a setting of the capacitor and the gamma bar that gives a reading of zero in the reflected position. This may take a few tries, but you'll find the settings are not critical. If you adjust the antenna near the ground you may find it will require a new adjustment when it is up in its permanent location.

How you mount the antenna in its permanent location will depend on what you have to hold the antenna. If you use a boom mounting-plate-to-mast assembly such as already described, the U bolts holding the boom could be left slightly loose and the 3/8-inch (10-mm) bolts left out. The antenna could then be installed in an upside-down position and then rotated into place. Two men are better than one for this job, but one man *can* do the job. If a hole, say 1/2-inch (13-mm) diameter, is drilled through the boom a 1/2-inch rod could be slid through to serve as a lever to rotate the boom and antenna into an upright position and the U bolts and hardware could be then tightened down.

We think the Delta-Loop beam is an excellent antenna and offers many possibilities. Certainly the results obtained to this point prove the antenna is a real performer. Our tower is a crank-up job, making antenna testing a fairly simple project. The antenna was first matched at the 8-foot (2.44 m) level (boom height) and at the cranked-up height, about 55 feet (16.76 m) above ground. A very slight readjustment of the gamma capacitor was required at the greater height in order to get down to a 1-to-1 SWR. However, the change in SWR from the 8-foot to 55-foot (2.44- to 16.76-m) heights was very small. Also, the highest SWR was at 21,450 kHz, the top end of the band, with a mismatch no worse than 1.2 to 1. At 21,300 kHz the SWR dropped to 1 to 1 and remained at that figure all the way to the lower band edge.

The entire antenna weighs about 15 pounds (using the aluminum boom) so the antenna could be rotated with a heavy-duty TV rotator. One last note: We sprayed the cotter pins and hose clamps with a clear acrylic spray to reduce corrosion. This simple precaution should be taken in all antenna construction to prevent or reduce rusting. This material was originally presented in *QST* by Lewis G. McCoy, W1ICP.

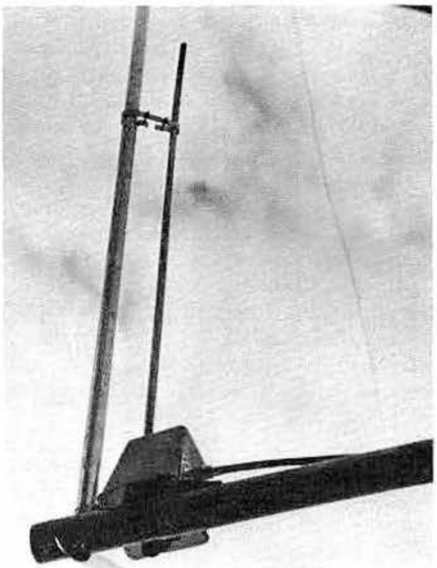

The plastic freezer box is held in place by the small metal bracket which is attached to the boom with self-tapping screws. Also on the bracket is a coax fitting to take the RG-8/U coaxial feed line. The gamma rod is mounted on top of the box and is held to the element with the shorting clamp.

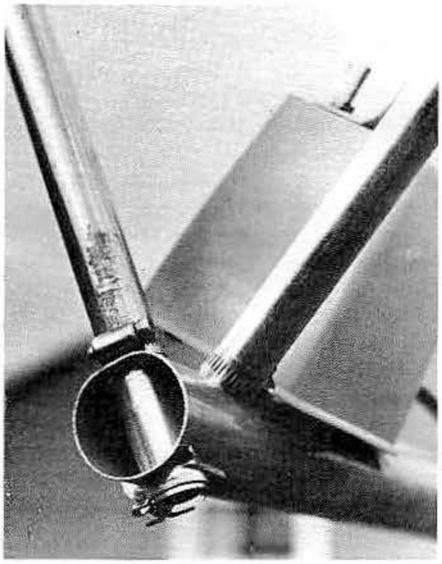

In this close-up view of the end of the boom, the mounting of the element ends are clearly shown. Cotter pins are used on the bottom end of the element and hose clamps on the top side of the boom. This secures the elements to the boom. The plastic freezer box that houses the gamma capacitor is clearly visible behind the elements.

Antennas for Travel Trailers and Campers

Pounding brass and rendering "lip service" from an RV or mobile home can be disappointing as all get-out if the antenna doesn't "play" well. Here are some tips for rig-mounted radiators that get the job done.

Who knows why so many hams like to go camping? Perhaps it's their Field Day training; perhaps it's the next logical step after mobile operation, or perhaps they just cannot leave home without their rigs. At any rate, it is nice to have your own means of communication with the outside world when you are away from home in the Maine woods, at the Grand Canyon, or even down in Mexico.

One thing all hams know is that no matter how good a transmitter you may have, you cannot get out without an antenna; and the better the antenna is, the better you get out. The antenna is usually no problem for the home station. The sky is the limit, so to speak. A 65-foot (19.81 m) crank-up tilt-over tower with a triband beam on top, and a couple of inverted Vs for 40 and 75 are not at all uncommon. But did you ever contemplate transporting such an antenna system to a state park, or some other campground, and setting it up while your XYL cooked suppper? Obviously, the mobile operator is faced with vastly different problems than those confronting the ham who stays at home, when it comes to antennas.

In 1929, the author built a 14-foot travel trailer, which was christened "Black Maria." Shortly thereafter, the writer got a ham license, and one of the first thoughts was toward a rig for the trailer. This was in the days when mobile operation was still a novelty, and the story of Black Maria was the author's first contribution to *QST*, in 1933. The rig used a pair of 33s in parallel operating from a B-battery pack.

Over the ensuing 35 years, many portable antenna designs have been tried, and much has been learned about their relative performance. The degree to which a trailer traveler is restricted in a choice of an antenna varies widely, of course, depending upon the facilities available at each stopping point. These are seldom known in advance, but to get the most out of the equipment, the trailer operator should be prepared to take maximum advantage of whatever facilities he may find on each occasion.

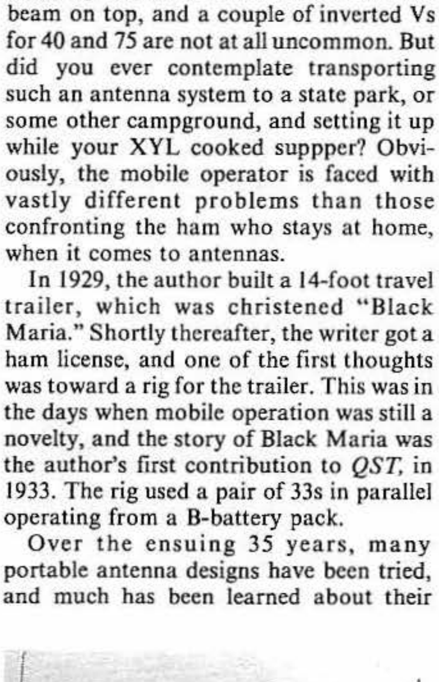

The author's 30-foot (9.14 m) Airstream trailer with the 75-meter Airstream Loop in place. The "tuning" mast is the one carrying the flag.

Dipole Antennas

Long ago, it was found that a simple dipole for 40 or 75 would far outperform the author's mobile whip, even with the dipole only a few feet above ground. While writing this, sitting in the latest Black Maria — a 30-foot Airstream job — up in Port Colburne, ON, the author was listening to the 75-meter phone band. The dipole is only 8 feet (2.44 m) above ground. When checking into the Airstream Travel Trailer Net, which meets every Sunday on 3963 kHz at 8 A.M. local time, the writer worked other trailer stations in Connecticut, New Jersey, Maryland, Pennsylvania, New York and Ohio. Several times the author switched to the mobile whip and could not be heard through the QRM. Such a dipole can be prefabricated and, when rolled up, occupies very little storage space.

When the length of the stay warrants it, and trees or other supports are available, the antenna will be even more effective when raised to a respectable height. A bow and arrow is kept on hand for shooting lines over taller trees. Incidentally, Hy-Gain makes a deluxe portable dipole consisting of two stainless-steel tapes in a double plastic housing. The tapes are calibrated in meters, and can be quickly reeled out to the proper length for use on any desired band. The unit is fitted with a coax connector, making installation very easy.

If only a single support is available, the dipole can be used in "inverted-V" fashion. A convenient way of putting up such an antenna is to use three or four 5-foot (1.52 m) lengths of 1-1/4-inch (32 mm) aluminum TV masting as a center pole, mounting it on the front of the trailer. The author has a mounting for such an arrangement clamped to the bumper, as shown in one of the photographs. The mast can be equipped with light nylon halyards for hoisting the center of the dipole. The ends of the dipole can be attached to whatever may be handy, but the higher, the better.

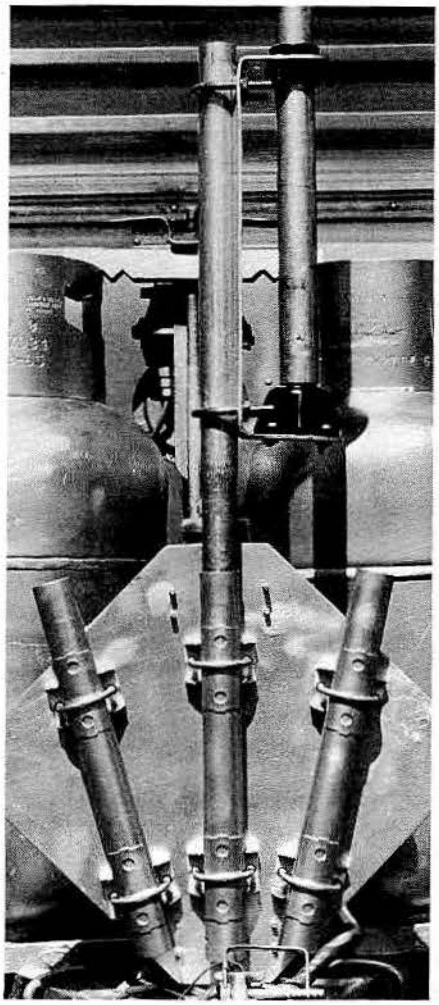

Insulated mounting at one end of the "Airstream Loop" 75-meter antenna. The "tuning" mast can be mounted directly in one of the other sockets. The tuning will vary with the spacing between masts as well as the length of the "tuning" mast. The insulated mounting can be transferred to one of the side sockets to obtain maximum spacing.

Loaded Antennas

Of course, there are many times when the surroundings, or the shortness of the stopover, make erection of a conventional dipole impractical. On these occasions, the author used one or another of several arrangements shown in the accompanying photographs and sketches. Where the

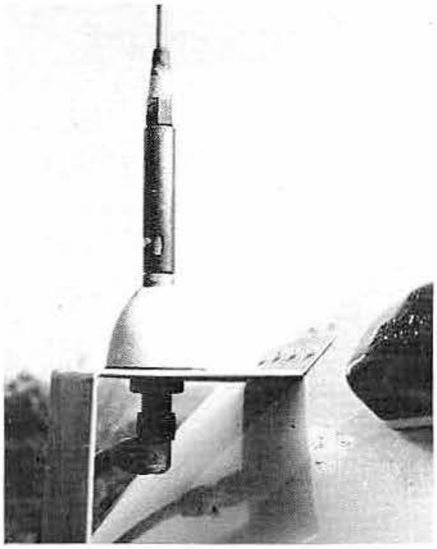

An elevated mounting for a mobile whip at a top rear corner of the trailer body. The bracket is fastened by pop rivets. This point may also be used, with the whip removed, for feeding wire dipoles.

A check with a 20-watt fluorescent lamp by W1WKZ indicates the "hot" points on the Airstream Loop 75-meter antenna system.

antenna system is one that requires a ground for operation, it was found that a quarter-wave wire counterpoise connected to the trailer, and strung out a few feet above ground, is usually a much better "ground" than the trailer alone, even if the trailer is actually grounded to a driven rod. The author carries two each of wires cut to lengths of 8, 16, 32 and 60 feet (2.44, 4.88, 9.75 and 18.29 m) for use as counterpoise on the 10-, 20-, 40- and 75-meter bands, respectively. One end of each wire is fitted with a large battery clip, while the other is terminated in an insulator. Trailer awning poles come in handy as counterpoise supports when nothing else is available.

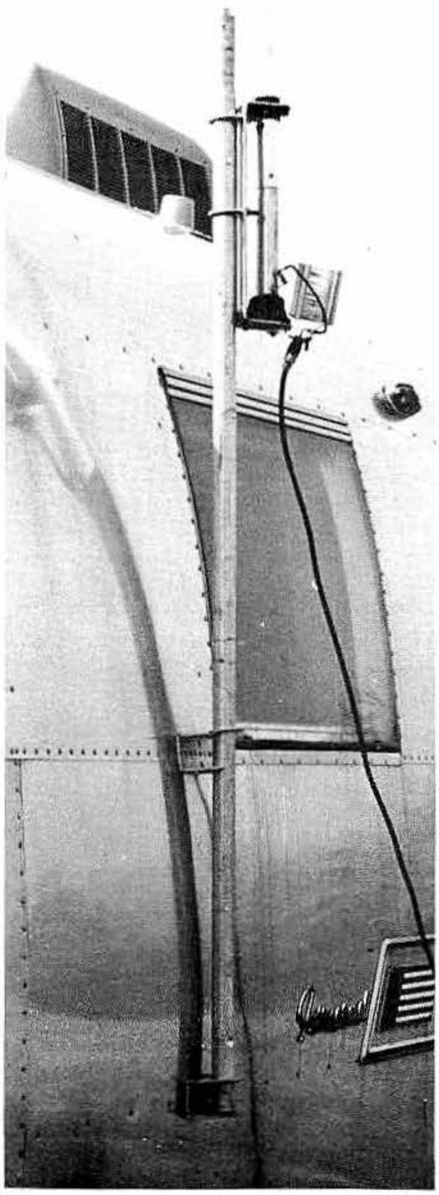

A 25-foot (7.62 m) base-loaded antenna mounted on brackets attached to the side of the trailer.

"Airstream Loop" Antenna

Of particular interest is the 75-meter folded configuration shown in the sketch of Fig. 23, which is called the "Airstream Loop," although, strictly speaking, it does not function as a true loop. With the fittings that are provided, this antenna can be put up in a few minutes, and its performance seems to be about equal to that of a low dipole. It has the advantage that it is completely supported by the trailer, and requires no external supports. It is contained entirely within your own campsite. There are no wires strung out for you or your neighboring campers to trip over or be garroted by in the dark. A 30-foot (9.14-m) length of wire is strung between, and connected to, two poles made of aluminum TV mast sections. A third mast (A), of approximately the same length as the others, and connected to the trailer, is an essential part of the system. This arrangement appears to work as a folded half-wave antenna, the 30-foot (9.14-m) wire and the two poles connected to it forming one half of the antenna, while the other half is formed by the trailer body and mast (A). Probing with a fluorescent lamp shows that the end of the trailer opposite the feed point is "hot" while the feed-point end is "cold." The system can be tuned by adjustment of the length of pole A, and/or changing the spacing between A and the adjacent pole. The spacing can be changed by a selection of the pole mounting sockets shown in one of the photographs. In this manner, the resonant frequency can be changed by almost 150 kHz. For any adjustment, the transmitter can be tuned about 30 kHz either side of the resonant point before the SWR exceeds 3 to 1. Since the "hot" end of the trailer is the front end in my case, the tow car must be unhitched, and an insulating block used under the customary supporting jack.

In a similar arragement for 40 meters, mast A is omitted, and a wire approximately 21 feet (6.40 m) long is connected to the pole at the feed-point end. The open end of the wire is supported from the pole at the opposite end of the trailer by a length of nylon rope.

40-Meter Monopole

Another 40-meter antenna that has been used with considerable success is a 34-foot (10.36 m) length of 300-ohm ribbon line connected as a folded monopole, and strung up in the same manner as the antenna just described. One conductor of the ribbon is grounded to the trailer at the feed point, while the other conductor is connected to the center conductor of the coax line. The two ribbon conductors are connected together at the far end.

Matching

Standing-wave ratios up to 5 or 6 to 1 are hardly worth worrying about at 4 MHz so

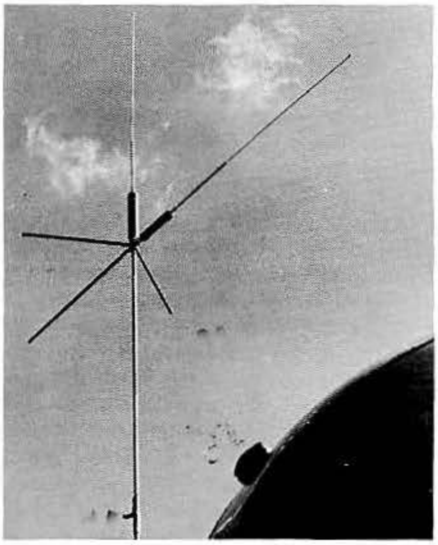

A five-band antenna system. A New-Tronics mobile antenna base section is fitted with a Waters Add-a-Band adapter. The short resonators for 10, 15 and 20 meters are Waters. The 80-meter resonator (top) is New-Tronics. A threaded stud has been added to the vacant side of the adapter to take a New-Tronics 40-meter resonator. The system is fed at the base with a single coax line.

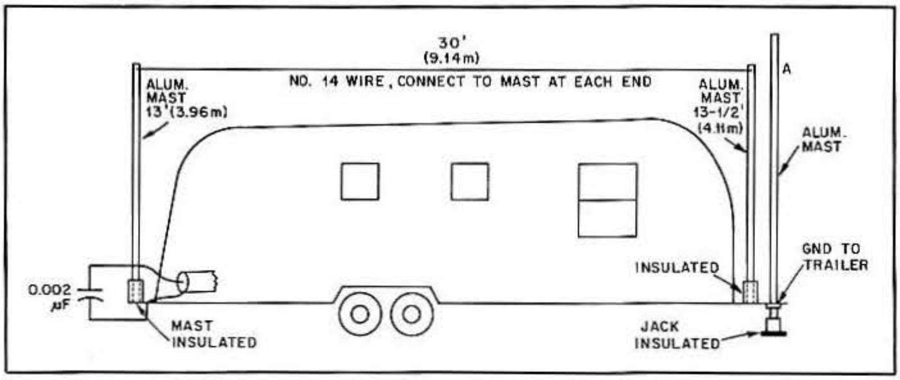

Fig. 23 — 75-meter "Airstream Loop"

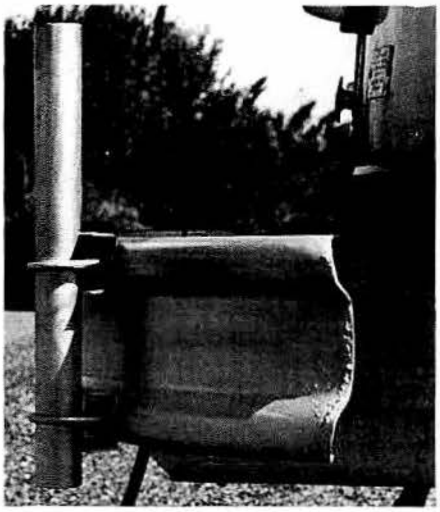

A bumper-mounted socket for TV masting.

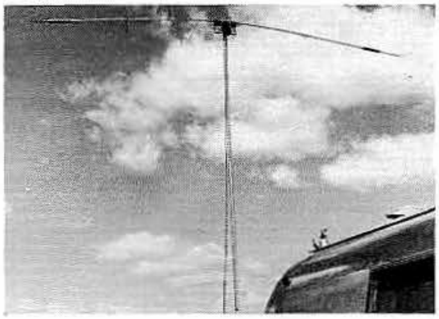

An 80-meter dipole made of two New-Tronics 75-meter mobile whip back to back. The 15-foot (4.57 m) mast is bumper-mounted. A 1-to-1 balun is mounted on the center insulating panel, which is secured to the mast with U bolts. This antenna is quite narrow-band when operated on 75 meters, making it necessary to retune the antenna for excursions of more than 30 or 40 kHz.

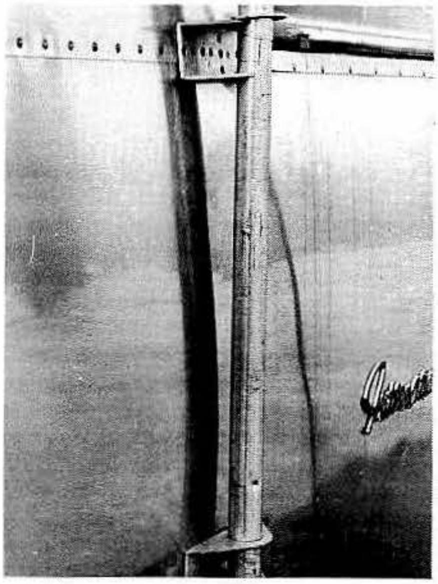

Brackets for mounting TV masting on the side of the trailer. The upper bracket is fitted with a removable U-shaped yoke bolted to the bracket. The brackets are fastened to the trailer with pop rivets.

Table 3

Hustler 75-meter mobile whip mounted vertically on top rear corner of trailer	S7
Same as above, with 60-foot counterpoise wire connected to trailer	S9
Two Hustler 75-meter mobile whips back to back as a horizontal loaded dipole	S9 + 5 dB
60-foot horizontal wire 8 feet high, using trailer as ground	S9 + 10 dB
Hustler 4BTV trap vertical with 75-meter resonator	S9 + 10 dB
120-foot dipole, 15 feet high at center	S9 + 20 dB
"Airstream Loop" antenna	S9 + 20 dB
Home-station dipole 50 feet high	S9 + 30 dB

Feet = 0.3048 = m.

far as additional power loss is concerned. However, an SWR greater than 1 to 1 causes the line input impedance to vary, depending on the length of the line. The output circuits of many transmitters and transceivers are not designed to work into loads departing significantly from 50 ohms. With these units, it may not be possible to load the final stage, or components in the output stage not designed for the voltages that develop with higher-impedance loads may be damaged. Use a Monimatch, or other reflectometer, and adjust the antenna for minimum SWR. The author carries along several mica capacitors ranging in value from 50 pF to 0.002 μF with battery clips attached, which are used to shunt the line at the feed point. Start out with the lowest value, and double the capacitance in steps until you get an SWR of 1 to 1 somewhere in the band, and then adjust the antenna length for resonance at the desired frequency.[1]

To get some idea of the relative performances, a series of checks was made at 3825 kHz on several of the antennas described. A receiver with a horizontal antenna was set up at a distance of eight miles from the trailer, and S-meter readings were taken with the results shown in the accompanying tabulation. All readings were made with identical input power to the transmitter, and with the antennas adjusted for an SWR of 1 to 1.

To sum it all up:

1) Carry a complete set of mobile whips for lunch stops, or other occasions where you just do not have the time or room for anything better.

2) If at all possible, use a 1/4-wave counterpoise clipped to the trailer when using a mobile whip.

3) If you have the time or space, clip a 1/4-wave wire, instead of a whip, to your mobile antenna mount, and run it out to a bush or tree.

4) In addition to (3), clip a 1/4-wave counterpoise to the trailer, and run it in the opposite direction.

5) If you are stopping for a short time, or for the night, rig up the 5-band mobile whip with five 1/4-wave counterpoises, one of appropriate length for each band.

6) If you want to concentrate on one band, put up the 16-foot (4.88 m) mobile-whip dipole, if your space is limited.

7) If you have plenty of room, put up a 1/2-wave dipole for your favorite band.

8) If you want to work all bands, and space for dipoles is limited, then put up your trap vertical with the 75-meter resonator tuned to your favorite frequency.

9) If you are going to spend a week in a particular campground, shoot some lines over a couple of high trees with a bow and arrow, and get your dipoles up in the air as high as possible.

10) The use of balloons and kites is good, but depends completely on the wind.

11) Next best to a high dipole, put up the "Airstream Loop" antenna for 75 meters.

12) If all else fails, bring your 2-kW PEP linear and your 2.5 kW ac gas generator along. This will at least keep the other campers awake. Don't forget a low-pass filter for TVI! This material was originally presented in *QST* by Philip S. Rand, W1DBM.

---

[1] [Editor's Note: This method is equivalent to the use of a stepdown L network. It can be used when the antenna feed-point impedance is lower than the line impedance. To obtain a match, the antenna length must be somewhat longer than a resonant length, so that the antenna shows inductive reactance at the feed point. (This reactance serves as a substitute for the reactance of the coil of a conventional L network.) For an accurate match, it is necessary to adust the antenna length and the value of the capacitor until the appropriate transformation ratio is obtained.]

The Apartment Dweller's Dilemma

Having trouble getting a signal out from "concrete city"? Nonresonant, random-length wires can provide surprising results if a Transmatch is used to make the rig happy with the load it sees.

It is apparent that many would-be amateurs living in apartment buildings believe it is impossible to get an effective signal on the air. They feel the need to have an outdoor antenna in order to make contacts. This isn't true. Of course, the apartment dweller can always go mobile, but there comes a time when the amateur wants a home station. This article treats the problem and, hopefully, will get a few more hams on the air.

Some General Considerations

There are exceptions to the following rules but, in general, they can be depended upon.

1) An outdoor antenna will work better than an indoor one.

2) An antenna inside a frame building with wood exteriors is better than the same antenna in a steel-and-concrete building.

3) The higher above ground, inside or out, the better the antenna will work.

4) The bigger (or longer) you can make an indoor antenna, the better — even if it means running the wire around corners.

5) Even a poor antenna *should* produce some contacts.

The Coupling Problem

Most transmitters are designed to work into a 50-ohm load, and contain little or no provision for adjusting the transmitter when the load is other than 50 ohms. Unfortunately, there is no random-length wire antenna that will present a 50-ohm load on all bands. What is required is a Transmatch. A Transmatch is simply an adjustable LC network that converts the unknown antenna impedance to 50 ohms. This unit, shown in Fig. 24 and the photographs, will cover the 80- through 10-meter bands and can handle 1 kW of rf power. In addition to matching the transmitter to the antenna load, the Transmatch helps reduce harmonic radiation. Many Novices get in trouble with the FCC because of 2nd-harmonic radiation from the 80-meter signal. This unit should reduce such harmonic radiation. Also, the use of a Transmatch can provide better front-end selectivity for the receiver. Some receivers are subject to severe cross modulation when operated near a broadcast station. The Transmatch should reduce this effect.

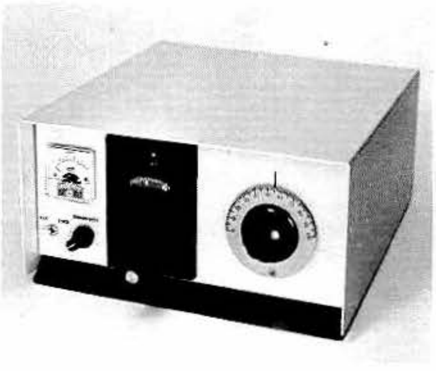

The completed Transmatch. The counter dial and knob are James Millen Mfg. Co., Inc. products. Check with G. R. Whitehouse, Amherst, NH 03031.

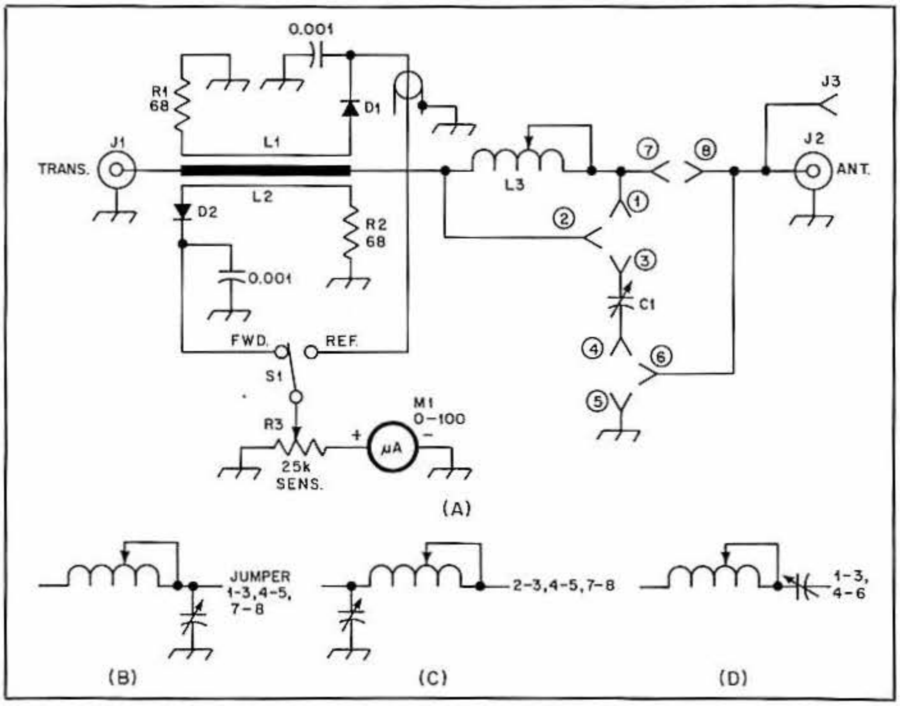

Fig. 24 — Circuit diagram of the L-network Transmatch. The eight banana jacks are E. F. Johnson type 108-900, and three dual banana plugs are required, E. F. Johnson type 108-200.
C1 — Variable capacitor. 350 pF (E. F. Johnson 154-10).
D1, D2 — 1N34A germanium diode.
J1, J2 — Chassis connector, type SO-239.
J3 — Feedthrough terminal, isolantite.
L1, L2 — See Fig. 25, part of etched circuit assembly.
L3 — Variable inductor, 28 µH (E. F. Johnson 220-203).
M1 — 100-µA meter.
R1, R2 — 68-ohm, 1/2-watt carbon or composition, *not wirewound.*
R3 — 25,000-ohm carbon control, linear taper.
S1 — Spst toggle.

Circuit Details

The unit shown in Fig. 24 is designed to be used in three configurations. They are shown at B, C and D. With one of the three hookups, it should be possible to match practically any antenna to the transmitter.

In order to get complete band coverage and avoid the complexities of bandswitching, banana and jack plugs are used to change the circuit to the configuration needed. For example, if one wanted the setup shown at B, he would jumper terminals 7 and 8, 1 and 3, and 4 and 5. Using the banana plugs makes for easy changing of the circuit.

Whenever a Transmatch is used, the operator should have a way of knowing when the unit is adjusted correctly. The answer to this need is a Monimatch or other SWR indicator. If you have an SWR meter, it can be connected in the line between the transmitter and the coupler. If

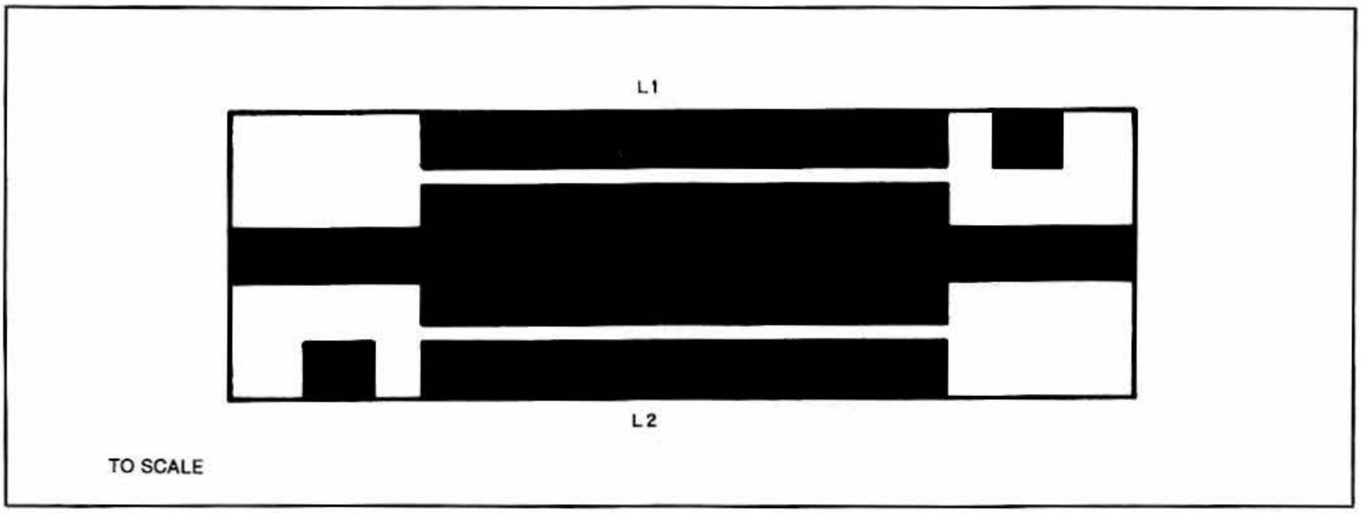

Fig. 25 — Etched circuit-board template. The foil side is shown, with etched portion shaded.

not, you can build the Monimatch[1] shown in Fig. 24, and include it in the cabinet with the Transmatch.

Construction Details

The chassis for mounting the Transmatch is made from a piece of aluminum measuring 10 × 19 inches (254 × 482 mm). The ends of the 19-inch length of aluminum are bent up to form a U-shaped chassis, the ends being 4-1/2 inches high to form a chassis 10 × 10 × 4-1/2 inches (254 × 254 × 114 mm). The back side of the U has an opening cut out, 3-1/4 inches high by 4-1/2 inches (83 × 114 mm) long. A piece of Plexiglas is mounted over this opening. The jack-plug sockets are installed directly on the plastic. Connections from the roller inductor, L3, and variable capacitor, C1, are made to the banana jacks. Be careful when drilling the holes for the jacks to insure that they will mate with the plugs. Fig. 25 shows the details for a pc-board Monimatch.

Methods for making etched circuit boards are given in detail in the construction chapter of The Radio Amateur's Handbook, so we won't treat the process here. When installing D1 and D2 on the board, be sure to use a heat sink while soldering the leads. Too much heat can ruin the diodes. Shielded leads are used from the circuit board to S1 to prevent unwanted rf pickup.

Today's equipment builder soon finds out that locating parts can be a tough problem. Fortunately, we have found a distributor who will furnish all the parts. With the exception of the chassis, the parts for the Transmatch can be purchased from Barry Electronics.[2]

Some Antenna Ideas

With this Transmatch one should be able to match any random-length antenna to his transmitter. To test the system, we strung up 25 feet (7.62 m) of wire in the ARRL lab (a steel-and-concrete building, first floor). The Transmatch provided the desired 50-ohm impedance on all bands. We didn't get any "60-over-9" reports, but we *did* make contacts.

As to your antenna installation, we suggest running the wire up to the ceiling, around the room, perhaps into the next room, making the antennas as long as possible. If the XYL objects to the wire, use an invisible antenna. Nope, we're not

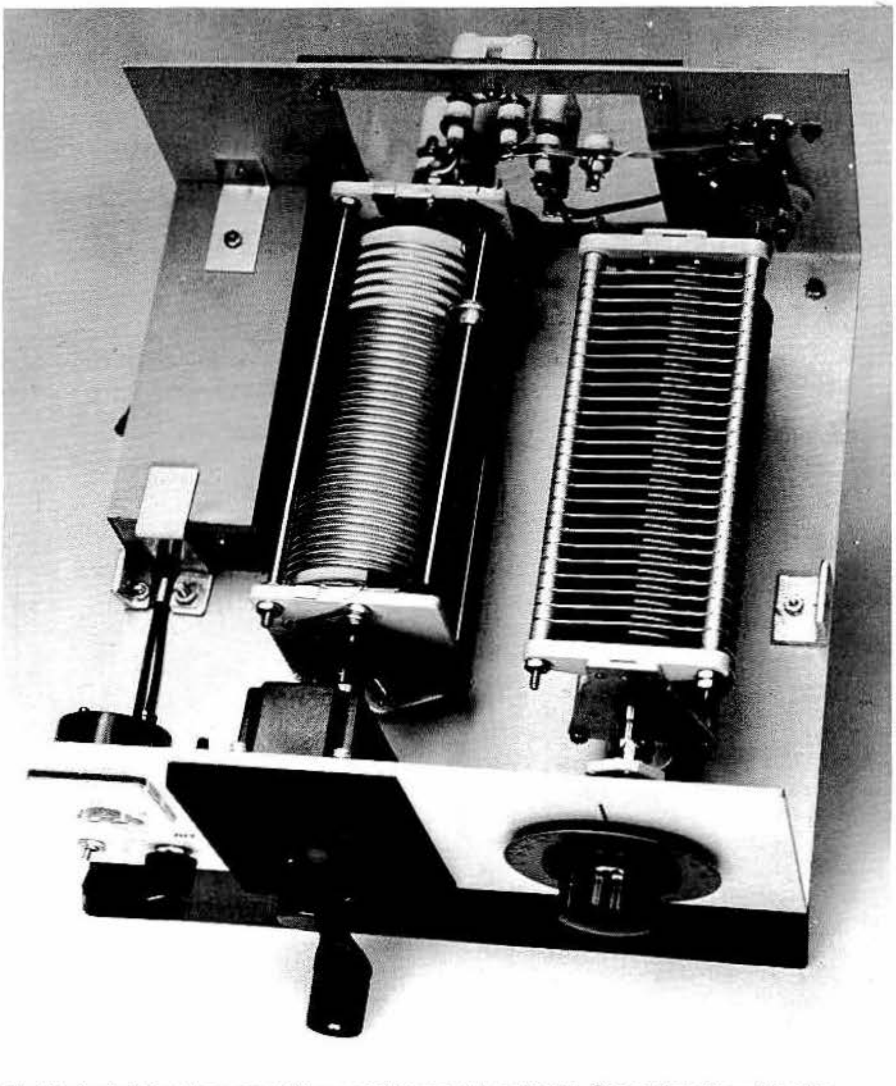

The Monimatch is at the upper left, covered by a metal enclosure. Connections from the roller inductor and the variable capacitor to the terminals on the jacks are made with thin strips of copper, although no. 12 or 14 wire can be used instead. The two antenna terminals are at the rear right. The top terminal is for use with a coax-fed antenna, if desired.

joking! You can make the antenna from no. 26 or 28 wire, which will be practically invisible. Also, you can conceal the wire behind molding. Be sure to insulate the wire if it runs near any metal. In some cases, it is impossible to get the wire outside. However, if you have a screen in one of your windows, terminate the end of the antenna at the screen. The screen *is* outside and it will help the system radiate. While it may be difficult to install, the antenna length to shoot for is approximately 120 to 130 feet (36.58 to 39.62 m) because this is a good length for multiband operation.

If you cannot put an antenna on the apartment roof, there is still a trick or two for having an outside antenna. You can drop a wire out the window and let it hang down. How long the wire should be will depend on how high up you are. You can make the wire long enough to reach nearly to the earth, but high enough to be beyond the reach of passersby. *Safety first*, always! There is no electrical law that states that a vertical must be fed at the bottom. Just connect the end of the wire to your Transmatch and tune up the system. You can use small-diameter wire. Use a lead sinker on the end of the wire and possibly imbed the sinker in a small sponge-rubber ball. (That way, you won't break your neighbor's window.)

If there is a support to which you can attach the far end of the antenna you can use the invisible-wire trick. Use rubber bands for insulators. Erect no. 26 or 28 wire. Don't be too discouraged if the wire gets broken by passing birds!

We know one ham who lived in a basement flat. He ran a wire through the wall to the bottom of a rain gutter, tuned up the system, and managed to work DXCC. (Safety first, again. The downspout should be beyond the reach of human beings!) The general idea is to use ingenuity. With a Transmatch you never can tell *what* will prove to be a useful antenna.

How to Tune Up

Using the Transmatch is not complicated. Although it takes some time to find the correct combination of settings, once determined, they can be logged for later reference. Use a short length of 50-ohm coax to connect the Transmatch to the transmitter. Attach the antenna to the Transmatch. Tune up your transmitter on the desired band, making sure that the final amplifier is resonated, but with the power output reduced. With the Monimatch in the forward-reading position, set the sensitivity control for a full-scale reading. You may have to increase the output of the transmitter to get full deflection, but be sure to keep the final amplifier tank in resonance. Switch the meter to the reflected position, and then adjust L1 and C1, until you get the lowest indication of reflected power. It should be possible to get the meter to read zero. When you have a zero reading in the reflected position, versus full scale in the forward setting, the Transmatch is correctly adjusted, and the SWR is 1. You may have to change the circuit to one of the other configurations in order to get a match, but one combination should work. Once you have the Transmatch set properly, you can then adjust the transmitter to its rated power input. One other point: It isn't always possible to get a good ground connection in an apartment. Therefore, a connection to a cold-water pipe or earth ground should be used. This material was originally presented in *QST* by Lewis G. McCoy, W1ICP.

References

[1] McCoy, "An Etched-Circuit Monimatch for Checking Your Antenna System," *QST*, October 1969. (See Chapter 5.)

[2] Barry Electronics, 512 Broadway, New York, NY 10012.

A 40-Meter DDRR Antenna

This antenna consists of a 1/4-wavelength element grounded at one end and wound into a single-turn coil, a few conductor diameters above the ground (see Fig. 26).

Shortly after the author's article on a practical DDRR Antenna appeared,[1] a number of inquiries were received regarding low-frequency versions of this antenna for operation at ground level. After a little experimenting, the author came up with this version of a 40-meter DDRR which works quite well. In early tests very fine QSOs were held with stations as far away as Phoenix, while the antenna was located inside the garage and surrounded by myriad metallic objects. When the antenna was relocated to the backyard, marked improvement was noted, even though this antenna site was marred by a sharp rise of 4 feet (1.22 m) in ground elevation, plus close proximity to house wiring and power lines. In spite of these obstacles solid contact was easily achieved with stations up and down the West Coast. Encouraged by these results, we moved the antenna to a rooftop location which put it well in the clear insofar as metallic objects above the groundplane were concerned. In this location, the DDRR really proved its worth. In spite of its low profile, a little over 1 foot (0.31 m), and its very small span, a few inches over 9 feet (2.74 m), it was more than adequate when competing with any signal on the band. The obvious conclusion from my experiments is that the 40-Meter DDRR is the apartment-dweller's dream. It is principally for that group that this article is prepared. Other interested amateurs might be those who are limited, as this author is, by too much house on too little real estate; or those who for other reasons cannot cope with high towers, masts and guy wires.

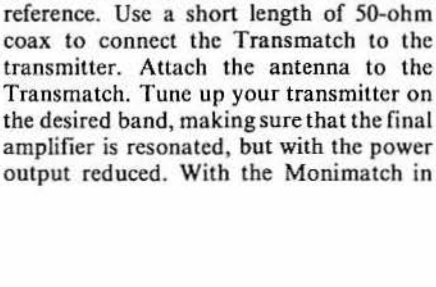

Fig. 26
RD = 0.078 λ (28°)
SP = 0.11D (2.5°)
FP = 0.25 h (See Note 1)
C = (See Note 2)
D = (See Note 3)
G = (See Table 4)
Notes: (1) Actual dimension must be found experimentally. (2) Value to resonate the antenna to the operating freq. (3) d ranges upward from 1/2" (13 mm). The larger "d" is the higher efficiency is. Use largest practical size, e.g., 1/2" (13 mm) for 10 meters, 5" (127 mm) or 6" (152 mm) for 80 or 160 meters.

Before this writer's enthusiasm sends you out to rip down your inverted V or to dismantle your beam, remember this: The DDRR, for all of its capabilities, will not supplant a full-size single-frequency antenna which is properly erected over clear terrain. What it will do is provide an

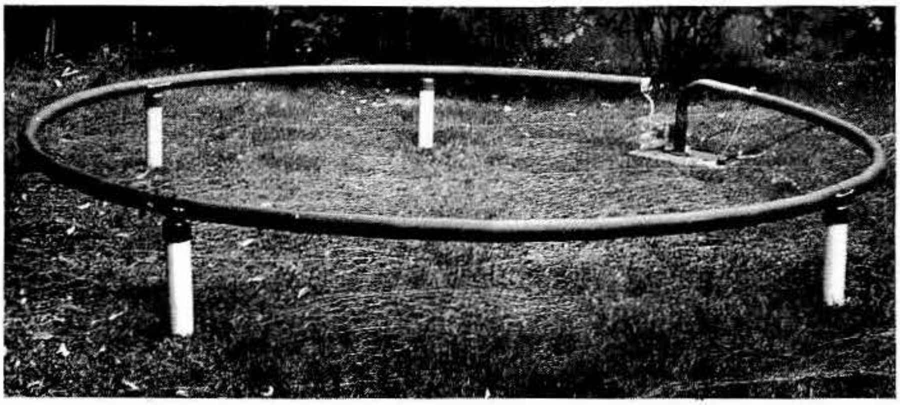

The chicken-wire groundplane is evident in the background. The base plate can be seen at lower right. Note the relative positions of the 52-ohm coaxial feed at the left end of the plate, the flange on the foot of the post, and the tuning unit at the right-hand end of the plate.

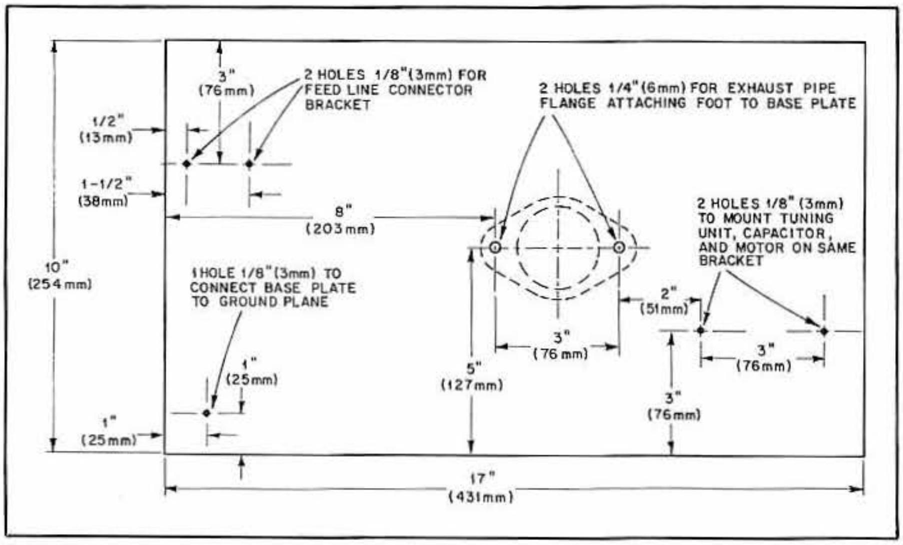
3"
(76mm)
2 HOLES 1/8"(3mm) FOR FEED LINE CONNECTOR BRACKET
2 HOLES 1/4"(6mm) FOR EXHAUST PIPE FLANGE ATTACHING FOOT TO BASE PLATE
1/2"
(13mm)
1-1/2"
(38mm)
8"
(203mm)
2 HOLES 1/8" (3mm) TO MOUNT TUNING UNIT, CAPACITOR, AND MOTOR ON SAME BRACKET
10"
(254mm)
1 HOLE 1/8"(3mm) TO CONNECT BASE PLATE TO GROUND PLANE
2"
(51mm)
3"
(76mm)
3"
(76mm)
1"
(25mm)
5"
(127mm)
3"
(76mm)
1"
(25mm)
17"
(431mm)

Fig. 27 — Drawing of the base plate which can be made from either steel or aluminum, as described in the text.

Table 4

Dimensions for 1/4 Wavelength DDRR Elements

Band (Meters)	*160*	*80*	*40*	*20*	*15*	*10*	*6*	*2*
Feed Point (FP)*	12"	6"	6"	2"	1.5"	3"	1"	1/2"
Gap (G)	16"	7"	5"	3"	2.5"	2"	1.5"	1"
Capacitor, pF (C)	150	100	70	35	15	15	10	5
Spacing (Height) (SP)	48"	24"	11"	6"	4 3/4"	3"	1 1/2"	1"
Tubing Diameter (D)	5"	4"	2"	1"	3/4"	3/4"	1/2"	1/4"
Ring Diameter (RD)	36'	18'	9'	4.5'	3'4"	2'4"	16 1/4"	6"

*See Fig. 26 for explanation of designations.

Feet X 0.3048 = m.

Inches X 25.4 = mm.

antenna which will enable communications of respectable quality, where heretofore it has been impossible because of inadequate space to erect a 40-meter antenna.

Some consideration should also be given to the fact that the high-Q nature of the DDRR and its resultant narrow-band characteristics serve to reduce the noise level. Boyer[2,3] reports that in the initial experiments it was found that DX stations which could copy signals from either a vertical or a DDRR could only be heard on the DDRR due to the reduction in background noise. So if you have a noisy location, it might be to your benefit to try the DDRR, regardless of what antenna you are presently using.

Constructing the 40-Meter DDRR

In this application, 2-inch (51 mm) diameter automobile exhaust pipe was used as the radiating element. The local muffler shop not only supplied the material, but also undertook to bend it to specifications. This was an obvious course since the material and the power bender were right at hand.

The braided lead across the flared joint is to assure electrical continuity. The screws used are self-tapping sheet-metal screws. The top end of the insulator (2-inch [51 mm] diameter PVC pipe) is a standard PVC pipe cap; the bottom is closed off with a cap from an aerosol can (2-inch [51 mm] ID).

The dimensions for 40 meters are:

Ring — 9 foot (2.74 m) diameter, center to center.

Height — 12 inches (0.31 m) from ground plant to element center.

Gap — 6 inches (152 mm) from upright post center to open end of ring.

In forming the ring to these dimensions, four 10-foot (3.05 m) lengths of tubing were used. A 10-degree bend was made at 9-inch (229 mm) intervals in three of the lengths. The fourth length was similarly treated, except for the last 18 inches (0.46 m) which were bent at right angles to form the upright leg of the ring. One end of each section was flared so that the sections could be coupled together by slipping the end of one into the flare of its mate.

The required flares are easily made at the muffler shop with the aid of the forming tools. Another task which can best be completed at the shop is to weld a flange onto the end of the upright leg. This flange is to facilitate attaching the leg to the mounting plate which provides a chassis for the tuning mechanism and the coaxial-feed coupler. After bending and flaring is complete, the ring is assembled and minor adjustments made to bring it into round and to the proper dimensions. This can best be done by drawing a circle on the floor with chalk and fitting the ring inside the circle. The circle must be slightly larger than the center-to-center diameter so that the reference line can be seen easily. For example, with two-inch (51-mm) tubing the actual diameter of the reference circle must be 9 feet, 2 inches (2.79 m). When you

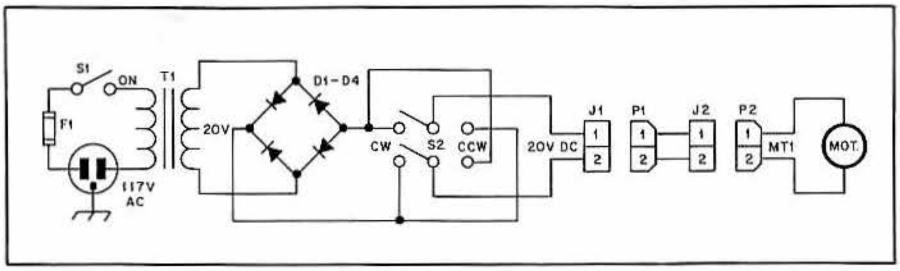

Fig. 28 — Circuit diagram of the power supply and control motor.

D1 through D4 — 1 A, 50 PRV, or equiv.
F1 — Fuse, 1 A.
J1, J2 — Two-terminal jack.
MT1 — Motor (see text).
P1, P2 — Two-terminal plug.
S1 — Spst.
S2 — Dpdt, center off.
T1 — Transformer (see text).

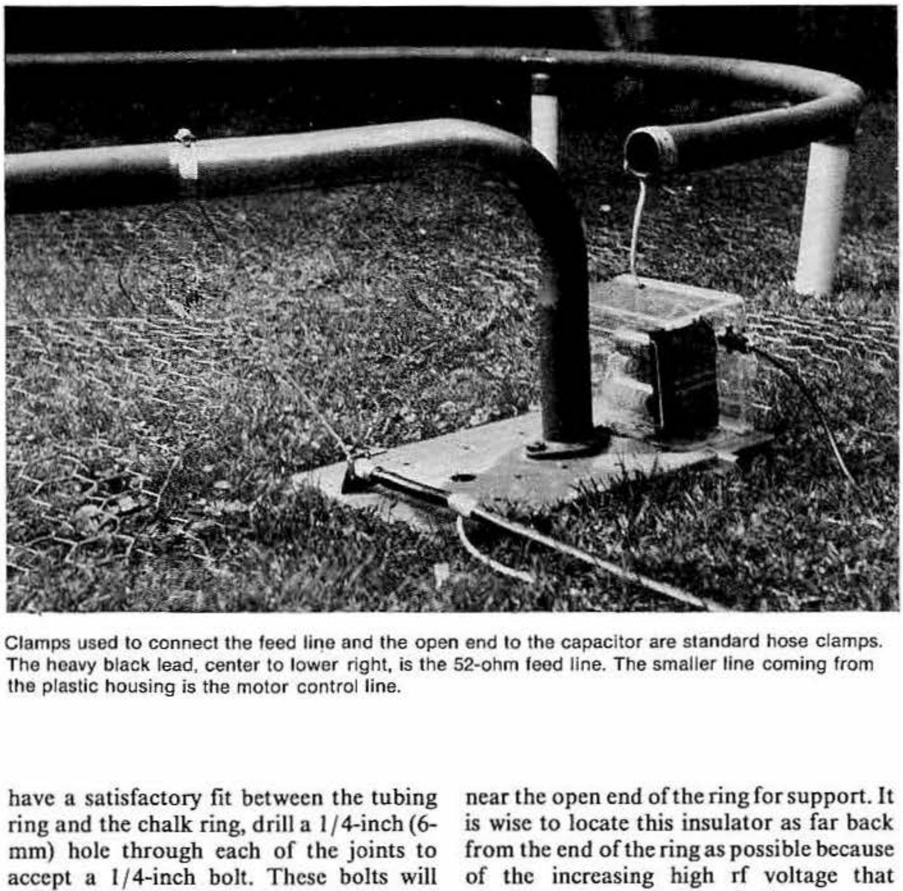

Clamps used to connect the feed line and the open end to the capacitor are standard hose clamps. The heavy black lead, center to lower right, is the 52-ohm feed line. The smaller line coming from the plastic housing is the motor control line.

have a satisfactory fit between the tubing ring and the chalk ring, drill a 1/4-inch (6-mm) hole through each of the joints to accept a 1/4-inch bolt. These bolts will clamp the sections together. Also, they can be used to attach the insulators which support the ring at a fixed height above the groundplane.

Making and Attaching the Insulators

Insulators for the antenna were made from 11-inch (279-mm) lengths of 2-inch (51-mm) PVC pipe inserted into a standard cap of the same material. The PVC caps are first drilled through the center to accept the 1/4-inch (6-mm) bolt previously installed at the joints. The caps are then slipped onto the bolts and nuts are installed and tightened to secure the caps in place. The 11-inch length of pipe, when inserted into the cap and pressed firmly until it touches bottom, results in a total insulator length of 12 inches (0.31 m). Four insulators are required: one at each of the joints and one near the open end of the ring for support. It is wise to locate this insulator as far back from the end of the ring as possible because of the increasing high rf voltage that develops as the end of the ring is approached.[4] As a final measure, the bottom ends of the insulators were sealed to prevent moisture from forming on the inside surfaces. Standard PVC caps could be used here, but we found that plastic caps from 15-ounce aerosol cans fit well.

Making the Mounting Plate

The mounting plate is required to provide good mechanical and electrical connections for the grounded leg of the radiator, the coaxial feed-line connection, and the tuning mechanism. If you are using aluminum tubing, you should use an aluminum plate, and for steel tubing, a steel plate to lessen corrosion from the contacting of dissimilar metals. Dimensions for the plate are shown in Fig. 27. The important consideration here is that good, solid mechanical and electrical connections are made between the ground side at the coaxial connector, the ring base, and the tuning capacitor.

The Tuning Unit

We found that the 9-foot (2.74 m) ring resonated easily with approximately 20 pF of capacitance between the high end of the ring and the base plate or ground. A 35-pF double-spaced variable from the junk box was pressed into service here (Cardwell NG-35-DS). Any variable which will tune the system to resonance and which will not arc under full power should be satisfactory. Remember, the rf voltage at the high impedance end of this antenna can reach 20 to 30 kV with high power, so if you are using the maximum legal limit, you would do well to consider using a vacuum variable capacitor. Since we limited our power to 500 watts PEP, the double-spaced Cardwell unit was satisfactory. To provide for full band coverage, the capacitor was coupled to a reversible, slow-speed motor which enabled the antenna to be remotely tuned from the operating position. An indicated SWR of 1.1 to 1 was easily achieved over the entire 40-meter band. The motor used was a surplus item made by Globe Industries of Dayton, OH.[5] At 20 volts dc the shaft of this motor turns at about 1 rpm which is ideal for DDRR tuning. The gears used were surplus items. If you cannot obtain gears, string and pulley drive will do almost as well, or you can mount both the motor and the capacitor in line and use direct coupling. Of course, if you operate on a fixed frequency or within a 40- to 50-kHz segment of the band, you can dispense with the motor entirely and simply tune the capacitor manually. In any case, the tuning unit must be protected from the weather. We used a plastic refigerator box to house the tuning capacitor and its drive motor.

Fig. 28 shows the electrical connection for the motor. A small train transformer or power supply for toy slot-cars will work admirably as a tuning motor power source. Standard ac zip cord was used for the connection between the control unit and the motor.

Electrical Connections and the Groundplane

The connection between the open end of the ring and the tuning capacitor is made with no. 12 wire or larger. On the end of the base plate opposite the tuning unit, and directly under the ring about 8 inches (203 mm) from the grounded post, install a bracket for a coaxial connector. The connector should be oriented so that the feed line will lead away from the ring at close to 90 degrees. Install a clamp on the ring directly above the coaxial connector. Connect a lead of no. 12 or larger wire from the coaxial connector to the clamp. This wire must have a certain amount of flexibility to accommodate the movement necessary when adjusting the match. The matching point must be found by experimentation. It will be affected by the nature and quality of the groundplane over which the antenna is operating. The antenna will function over earth ground; however, in our location we found the electrical ground to be unpredictable. A groundplane surface of chicken wire (laid under the antenna and bonded to the base plate) provided a constant ground reference and improved performance. In a rooftop location sheet metal roofing should provide an excellent groundplane. A poor ground usually results in a matching point for the feed line far out along the circumference of the circle. In our installation a near-perfect match was obtained with the feed line connected to the ring about 12 inches (0.31 m) from the grounded post. During testing, when the antenna was set up on a concrete surface without the groundplane, a match was found when the feed line was connected nearly 7 feet (2.13 m) from the post!

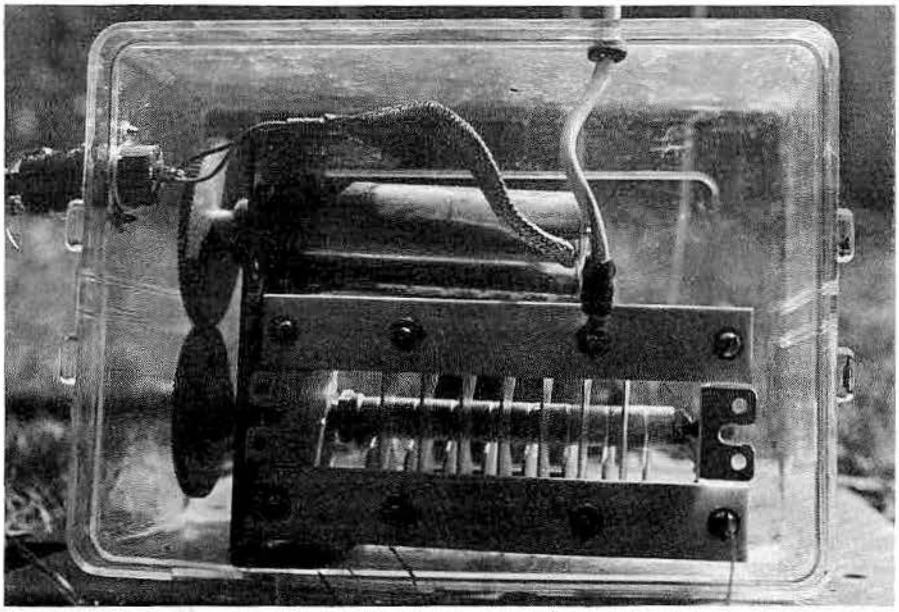

The tuning motor, above the capacitor, couples to the capacitor shaft through the gears at left. Both the motor and the capacitor are mounted on a common bracket which attaches to the base plate by two bolts through the weatherproof housing. The lead passing upward through the grommet connects the stator of the capacitor to the open end of the ring.

As shown in photos, the compactness of the antenna is readily apparent. The groundplane is made up of three 12-foot (3.66-m) lengths of chicken wire, each 4 feet (1.22 m) wide, which are bonded along the edges at about 6-inch (152-mm) intervals. In our installation the antenna, with the groundplane, could be dismantled in about 30 minutes. If portability is not important, it is best to bond all of the joints in the tubing so that good electrical continuity is assured.

After all construction is completed, the antenna should be given a coat of primer paint to minimize rust. If it suits you, there is no reason why a final coat of enamel could not be applied.

Tuning Procedures

Once the mechanical construction is completed, the antenna should be erected in its intended operating location. Coupling to the station may be accomplished with either 52-ohm or 72-ohm coaxial cable. Tune and load the transmitter as with any antenna. While observing an SWR meter in the line, operate the tuning motor. Indication of resonance is the noticeable decrease in indicated reflected power. At this point, note the loading of the transmitter; it will probably increase markedly as antenna resonance is approached. Retune the transmitter and move the feed-point tap on the antenna for a further reduction in indicated reflected power. There is interaction between the movement at the feed tap and the resonance point; therefore, it will be necessary to operate the tuning motor each time the tap is adjusted until the lowest SWR is achieved. Don't settle for anything less than 1.1 to 1. With a good ground and proper tuning and matching, this ratio can be achieved and maintained over the entire band. Once the proper feed point has been located, the only adjustment necessary when changing frequency is retuning the antenna to resonance by means of the motor. If the antenna is to be fixed tuned, provide an insulated shaft extension of 18 inches (457 mm) or so to the tuning-capacitor shaft for manual adjustment. This not only provides insulation from the high rf voltage but also minimizes body-capacitance effects during the tuning process.

Alternatives

A number of materials other than the steel tubing used here are well suited for the ring element. Standard E.M.T. or electrical conduit would work as well with a slight increase in weight. One advantage to be gained through the use of conduit is the elimination of the flaring operation since standard couplings would serve to connect each segment of the ring to its adjacent member. Another suitable material is copper tubing. This material is superior to either exhaust pipe or conduit in terms of its conductivity characteristics. Another advantage of copper is that it is available in continuous lengths and the joints could be omitted entirely.

The 2-inch (51 mm) dimension is by no means mandatory. Smaller diameter tubing has been used with satisfactory results.[6] In fact, DDRR antennas have been fabricated with wire elements. But, if the element diameter is reduced, the antenna tunes more *sharply*. Some experimenters may wish to go in the other direction and use a larger diameter. The author recommends the use of aluminum downspout with a diameter of about 4 inches (2 mm). This material does not lend itself to bending, however, and the ring must be configured as a regular polygon of eight or more sides. Because of the large number of joints involved, welding is about the only practical means of joining the segments. Unless you are equipped to do this work yourself, the cost of welding might be prohibitive. Anyone who undertakes to make a DDRR antenna of a large element diameter will be rewarded in terms of improved performance.

Performance

Results have been quite encouraging, and it is hoped that more and more hams will equip themselves with the DDRR in the future. The antenna has proved its worth and deserves more investigation by the amateur fraternity than it has been given in the past. No intensive efforts have been made to work DX with the antenna described here: however, a low angle of radiation is conducive to DX, and this antenna demonstrates a low radiation angle. It was found that distant areas, such as the East Coast, are more easily contacted than are stations nearby. All of the results could be attributed to peculiarities of individual stations or skip conditions, but since they are charateristics which can be anticipated with high-efficiency antennas having low radiation angle, we prefer that interpretation. Besides, where else will you find an antenna for 40 meters that is small enough to fit into a corner of the backyard and not protrude above the fence; or for that matter, which could be mounted on the roof of an apartment building or even a ranch-type house and not be visible from the street?

Nearly everyone who listened to the description of this new antenna was enthusiastic. The author hopes to hear many hams on the air working with the DDRR in the near future. This material was originally presented in *QST* by W. E. English, W6WYQ.

Parts List

- 4 — 10-ft (3.05 m) lengths of 2-in. (51 mm) tubing, exhaust pipe, conduit or copper tubing.
- 1 — base plate 7-1/2 X 19 X 1/8 inches (191 X 482 X 3 mm), steel, aluminum, or copper to match tubing.
- 4 — PVC pipe caps for 2-inch (51 mm) pipe.
- 4 feet (1.22 m) of 2-inch PVC pipe.
- 2 clamps for 2-inch tubing.
- 4 — 1/4 X 4-inch (6 X 102 mm) bolts with nuts.
- 1 reversible motor with 1-rpm shaft output.
- 1 wide spaced variable capacitor. 5 to 35 pF (Millen 16550 or equiv.).
- 1 coaxial connector SO-239.
- 2 sets 2-connector plug and socket for motor control.
- 36 feet (10.97 m) of 4-foot (1.22 m) chicken wire or equivalent.
- 1 flange to attach tube to base plate.
- 2 — 1/2 X 1-inch (13 X 25 mm) bolts with nuts (flange mounting).
- 8 — 3-inch (76 mm) bonding strips no. 8 braided wire.
- 4 plugs 2-inch (51 mm) ID for insulator bases.

References

Hicks, "The DDRR Antenna: A New Approach to Compact Antenna Design," *CQ*, June 1964.

Horn, "The Half Wavelength DDRR Antenna," *CQ*, September 1967.

The Radio Amateur's Handbook, 46th ed., The American Radio Relay League.

Notes

[1] English, "A Practical DDRR Antenna," *73*, June 1970.

[2] Boyer, U.S. Patent Nos. RD 26196, RE 3,151,328 (All rights assigned to the Northrop Corporation.)

[3] Boyer, "Hula-Hoop Antennas: A Coming Trend?" *Electronics*, January 11, 1963.

[4] Editor's Note: Because of the danger of rf burns, in the event of accidental contact with the antenna, precautions should be taken to prevent random access to the completed installation.

[5] Globe C-5A-1106 (available from Lectronic Research Labs Inc., 75 Arch St., Philadelphia, PA 19106).

[6] See footnote 1.

Novi-Loop

A medium-size roll of wire, some insulators and a day of good antenna "puttin' up" weather can net you a Novi Loop. What will it do for you? Well, follow these instructions and get set to work 40-meter DX. Nuff said!

One of the sad facts of life for many beginners is the matter of poor antenna performance. Many enthusiastic Novices have fallen behind in their efforts to work DX because of bitter disappointment caused by poor performance in the antenna department. All too often it is suggested that a simple end-fed hunk of wire is ample for all-around Novice band operation, or that a dipole will suffice for most Novice operation. Sure, almost any kind of radiator that can be tuned to the operating frequency and marched to the impedance of the feed line and transmitter will enable the operator of a low-power station to make some contacts from time to time. It depends on how high the system is above ground, how free and clear it is of surrounding objects that can detune the radiator and absorb the energy radiated from it, and upon the condition of the band at a given time. It should stand as a firm rule, then, that any amateur who wants good results should erect the best antenna system he can, consistent with available space, finances and structural practicality. It is possible to advance beyond the plateau of simple dipoles and random-length wires in the quest for better performance, and without going to the expense and effort of erecting large directional beams such as the cubical quad or Yagi.

What will be described here can be applied to any of the high-frequency amateur bands. However, the example given relates to only the 40-meter Novice band. The antenna is a full-wave loop of wire, mounted through necessity at W1FB in a rectangular format, and matched with a simple quarter-wavelength coaxial transformer. It has given excellent performance in DX work and for making "solid" stateside contacts. Total cost for the antenna system (wire, matching transformer and insulators) is under $12, exclusive of the feed line used to connect the antenna to the rig.

Antenna Features

Full-wave, closed-loop antennas are broadband, low-Q devices. This is a handy feature because it permits the operator to move his transmitter frequency within a particular band without need to compensate for an increase in standing-wave ratio (SWR) which might otherwise occur if other types of antennas were used. (The higher the Q of an antenna, the narrower its bandwidth, even though an SWR of 1 can be obtained at some frequency within the band for which it is built.) The foregoing rules out the need for a Transmatch with the loop antenna, provided the feed method recommended here is applied.

Another interesting feature of this kind of antenna is the theoretical gain of approximately 2 dB it exhibits over a half-wave dipole. The angle of radiation from a properly erected full-wave loop is considered to be lower than that of a dipole when both are less than one-half wavelength above ground. This feature suggests the superiority of the loop in situations where significant height above ground is not a practical goal for the chap wishing to put up a 40-meter antenna.

A full-wavelength closed loop need not be square. It can be trapezoidal, rectangular, circular, or some distorted configuration in between those shapes. For best results, however, the builder should attempt to make the loop as square as possible. The more rectangular the shape, the greater the cancellation of energy in the system, and the less effective it will be. The effect is similar to that of a dipole whose effectiveness becomes impaired as the ends of the dipole are brought closer and closer together. The practical limit can be seen in the "inverted-V" antenna, where a 90-degree apex angle between the legs is the minimum value used. Angles that are less than 90 degrees cause serious cancellation of the rf energy.

The loop can be fed in the center of one of the vertical sides if vertical polarization is desired. For horizontal polarization it is necessary to feed either of the horizontal sides at the center. At the time of this writing there have been no data compiled to provide a comparison between the performance, vertical versus horizontal. Such an experiment could be interesting (and possibly productive) for the Novice.

Erecting the Antenna

Optimum directivity occurs at right angles to the plane of the loop, or in more

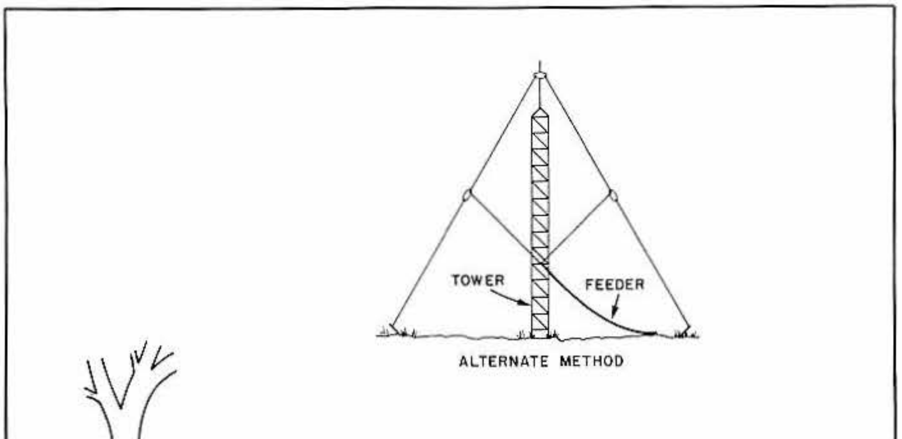

Fig. 29 — Details of the full-wave loop. The dimensions given are for operation at the low end of 40 meters (7050 kHz). The height above ground was 7 feet (2.13 m) in this instance, though improved performance should result if the builder can install the loop higher above ground without sacrificing length on the vertical sides. The inset illustrates how a single supporting structure can be used to hold the loop in a diamond-shaped configuration. Feeding the diamond at the lower tip provides radiation in the horizontal plane. Feeding the system at either side will result in vertical polarization of the radiated signal.

simple terms, broadside from the loop. Therefore, one should try to hang the system from available supports which will enable the antenna to radiate the maximum amount in some favored direction. The bidirectional pattern is maximized NE and SW at the writer's QTH. This gives good results in working Europeans and the stations "down under." Excellent signal reports have been obtained off the sides of the loop when working within the USA.

Just how the wire is erected will depend on what is available in one's yard. Trees are always handy for supporting antennas, and in many instances the house is high enough to be included in the lineup of solid objects from which to hang a radiator. If only one supporting structure is available it should be a simple matter to put up an A frame or pipe mast to use as a remaining support. (Also, tower owners see Fig. 29 inset.)

The overall length of the wire used in a loop is determined from the formula

$$\frac{1005}{f\ (\mathrm{MHz})}.$$

Hence, for operation at 7125 kHz the overall wire length will be 141 feet (42.98 m). The matching transformer, an electrical quarter wavelength of 75-ohm coax cable, can be computed by dividing 234 by the operating frequency in MHz, then multiplying that number by the velocity factor of the cable being used. Thus, for operation at 7125 kHz

$$\frac{234}{7.125\ (\mathrm{MHz})} = 32.84 \text{ feet}$$

If coax with solid polyethylene insulation is used, a velocity factor of 0.66 must be employed. Foam-polyethylene coax has a velocity factor of 0.80. Assuming we are to use the solid dielectric coax, RG-59/U, the length of the matching transformer becomes 32.84 (feet) × 0.66 = 21.67 feet, or 21 feet 8 inches (6.605 m). The transformer will convert the 100-ohm feed-point impedance of the loop to 50 ohms, a convenient value for connecting the system to the transmitter. The loop described here was measured with the K1TD Macromatcher[1] and found to exhibit an impedance of 110 ohms. Some inductive reactance was indicated. By removing 1 foot (0.31 m) of wire from the overall loop the reactance dropped to zero, though the antenna loaded up fine prior to the final trimming. The feed impedance at 21 MHz was also 110 ohms, but considerable inductive reactance was observed. Most certainly one could use the antenna on 15 meters with reasonable success, despite the SWR which might be present.

Results

During two weeks of testing the loop gave an excellent account of itself. It was compared with the writer's two-element short beam for 40 meters, and the results indicated that the loop was better than the beam on some occasions. At other times the beam provided better performance, or was at least equal to the loop. The differences noted were the result of band conditions and the time of day. High-angle signals can cause misleading results when comparing antennas for hf-band use. More meaningful tests can be made on ground-wave contacts, however, and several checks made between the loop and the beam indicated that radiation at right angles to the plane of the loop compared closely with that from the front of the beam. The beam is rotatable and installed 40 feet (12.19 m) above ground.

With 150 watts dc input to the transmitter PA, and operating at the low end of 40 meters (loop was cut for 7050 kHz), the first night of operation resulted in the logging of three Russian stations, two Polish ones, two Yugoslavians, one Italian station and two English ones. The lowest report was RST 549. The highest report was RST 599. Generally speaking, the loop has outperformed the beam during most long-haul contacts. Though no effort has yet been made to work South American stations (off the side of the loop), they have been heard loud and clear, as have some VK stations via long path at approximately 2300 UTC.

It is quite likely that the Novice who chooses to erect his own Novi-Loop will have a commanding signal on 40 meters. If backyard space permits, it should be a worthwhile venture. Don't forget, the formulas and guidelines given here are applicable to the other hf Novice bands as well. This material was originally presented in *QST* by Doug DeMaw, W1FB.

---

[1] Hall, "The Macromatcher," *QST*, January 1972.

Chapter 5

Antenna Theory and Test Methods

In order to obtain top performance from a given style of amateur antenna, it is important to understand the basic principles of operation. Additionally, some test procedures are necessary if one is to ensure that his or her antenna is functioning at its best. This chapter contains assorted data on antenna theory, plus a number of common-sense test methods.

These reprints of popular *QST* articles were chosen for all amateurs, regardless of their technical-aptitude levels. There's something to honor in the phrase, "Learn by doing." This chapter will help to make that philosophy a reality.

Quad vs. Triband Yagi

The controversy about quads working better than Yagis, or vice versa, may never end. But, we'll bet our newest roll of RG-8/U that you'll find some interesting reading on the subject in this report.

While stationed in Japan, a sort of DX crossroads of the world, this writer had the opportunity to observe, firsthand, the excellent performance of the cubical quad in competition with the Yagi, dipoles and an assortment of other antenna systems. In pursuing this undertaking certain steps were necessary to insure than any conclusions made would be meaningful. With this in mind a play emerged.

Objectives

In the many articles written on the cubical quad, it is noteworthy that only on a few occasions have the authors been privileged to compare the quad with other types of antennas on a real-time basis, and from the same operating location. Furthermore, when such comparisons were made, the authors generally compared against some type of monoband antenna system. A casual scanning of the 10-, 15- and 20-meter phone bands would lead one to conclude that the triband Yagi enjoys a rather high position of popularity among the antennas in general use. This being the case, it appeared that a worthwhile contribution to the data already available on the Yagi and quad might be made by conducting a series of controlled comparative tests, employing the triband Yagi and the quad. The test objectives were then defined: to compare various configurations of a cubical quad antenna with a representative commercial triband Yagi; such tests to be conducted over short, medium, and long transmission paths, and to arrive at conclusions regarding the relative merits of each antenna.

Table 1

	Mod. 1	*Mod. 2*	*Mod. 3*	*Mod. 4*
Reflector Element	72'3"	70'4"	72'5"	72'5"
Driven Element	69'	70'4"	70'5"	70'5"
Director Element	—	—	—	69'1"
Spacing "A"	7'6-1/2"	8'5"	13'4"	13'
Spacing "B"	—	—	—	13'
Stub	20-30"	34-38"	—	—

Feet X 0.3048 = m.
Inches X 25.4 = mm.

Test Plan and Procedure

Every effort was made to conduct the tests in a manner which would lessen the possibility of compromising the techniques employed by either the writer or participating stations:

1) The test to be performed by establishing communications with Amateur Radio stations located throughout the world on a random and scheduled basis.

2) Amateur Radio stations volunteering to assist in this effort to be briefed on conduct of test and data desired.

3) A voice ssb transmission to be made to the participating station, identifying the first antenna used as antenna "A."

4) The voice transmission to be followed immediately by an unmodulated carrier for a period of approximately 5 seconds.

5) The antennas would be switched, and a voice transmission made identifying the antenna as "B," and the procedures above repeated.

6) Participating stations will note signal strength related to each antenna, and provide a numerical value as observed on his S meter or other indicating device. These values to be logged, and the test reinitiated with another volunteer station.

Equipment Preparation

1) Antenna heights to be as nearly identical as possible.

2) Center of antenna horizontal lobe patterns to be as nearly identical as possible when pointing the antennas toward a participating station.

3) Resonant frequency of each antenna to be matched as closely as possible.

4) Transmission lines to be matched to antennas and transmitter loading to be as nearly identical as possible with each antenna.

5) Instantaneous transfer of antennas.

6) Relative power and SWR to be monitored continuously.

7) Prior to and after each data gathering session, equipment parameters will be verified. If a significant deviation in any parameter is noted, data collected will be discarded.

Antenna performance conclusions to be based on an analysis of data derived from a minimum of 50 unmodulated-carrier observations with each antenna configuration, and supplemented with data gathered during conventional ssb QSOs.

Table 2

	Mod. 1	*Mod. 2*	*Mod. 3*	*Mod. 4*
Total Observations	50	60	60	52
Less than 2100 miles	12	2	3	3
2100 to 4800 miles	33	31	33	32
Greater than 4800 miles	5	27	24	17
Signal Difference				
More than 1 S unit better	—	—	—	—
Less than 1 S unit better	—	—	7	9
No discernible difference	1	5	51	43
Less than 1 S unit poorer	27	46	2	—
More than 1 S unit poorer	22	9	—	—

Antenna Selection

This writer had been using a four-element commercial triband Yagi (boom length 24 feet [7.32 m], and 55 feet [16.76 m] above ground) for approximately 1-1/2 years, so the properties of this antenna were fairly well established. Furthermore, in on-the-air comparisons with competitive models of triband Yagis in use by other U.S. amateurs operating from the Tokyo area, the antenna appeared representative of commercial triband antennas in general use by the amateur community. Therefore, the Yagi in use at the author's station was selected as the reference antenna.

Text material concerning quad antennas, available to the author in Japan, was reviewed. It became evident that there are almost as many variations in quad design as there are writers on the subject. After much deliberation, and many discussions with amateurs throughout the world, the decision was made to test three models of the quad (a fourth model was tested as will be noted later). Since the physical characteristics of the quad are fairly standard, only the dimensions of the elements and the spacing between them was considered. The dimensions for the three models tested were obtained from a Japanese manufacturer of cubical quads, from Orr's book, *All About Cubical Quad Antennas,*[1] and from Dr., J. E. Lindsay, Jr., W0HJ.[2]

Preliminary Testing

Several days were spent "dry running" the test plan to validate the concept, and to smooth out the operating procedures and techniques. Of particular concern was the possible time required to make a valid data-gathering observation. If data were to be reasonably accurate, the transmission path had to be stable, and the signal-strength observations must be taken on each antenna during a short period of time. The dry runs were valuable in this respect.

A problem became evident during the first day of testing. It appears that those of us who speak and understand English do not always convey the same message when using the same words. As a result it was necessary to modify the verbal format, utilizing simple sentences and placing them in a logical sequence.

It also became apparent that the test could not be conducted under all transmission-path conditions; that even under ideal conditions several observations were often necessary before a conclusive report could be compiled. It was decided to conduct the tests only on 20 meters. The operating time available to the writer favored openings on 20 meters to Europe via the long path, and to Australia, the U.S. and various islands in the Pacific. It was also decided to orient the test antennas so that the topography and obstructions seen by each antenna would be essentially the same. (Physical separation between the two antennas was in the order of one wavelength.)

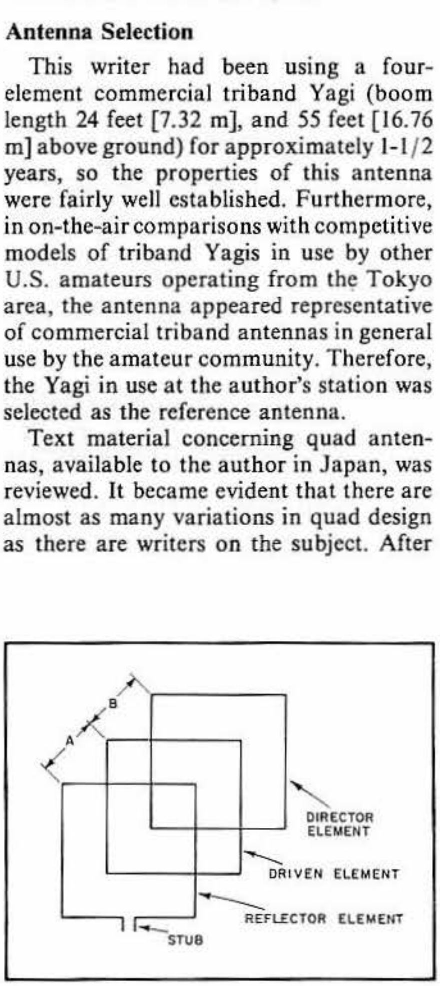

Fig. 1 — Element spacing information for Table 1.

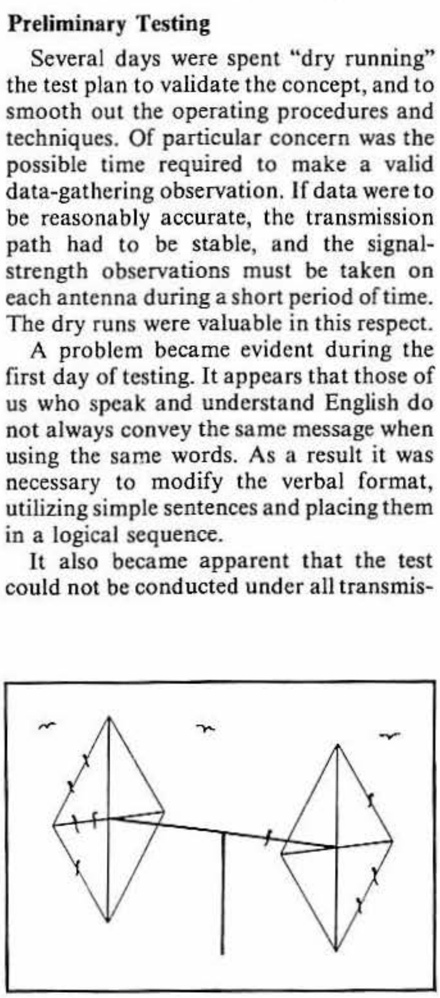

Fig. 2 — CLOSE SPACING IS FOR THE BIRDS!

Testing

Dimensions of the first quad model selected were furnished by a Japanese manufacturer of cubical-quad antennas. (see Table 1). The antenna was assembled, utilizing commercially manufactured heavy-duty hardware and fiberglass spreaders. It was tuned to a center frequency of 14,200 kHz. Testing of the first model began in November 1967 and continued for one month. The results for this period are given in Table 2.

In mid-December 1967 the first quad was replaced by a model constructed according to the formula and dimensions given in Orr's book. The results obtained with model 2 are contained in Table 2.

Construction of the third model (with wider element spacing) was carried out next. Two matching systems (Gamma and Q-section) were experimented with on this antenna. A satisfactory match could be had with either system. However, the Q-section was used for the test because it was the technique used with the previous two quad models (SWR with each antenna was never more than 1.3:1 with a difference between antennas no greater than 0.1). The results conducted with this model were most enlightening, as shown in Table 2. The model antenna was also used extensively during the first weekend of the 1968 ARRL DX Contest. Though these contacts were not used in tabulating test samplings, it is interesting to observe that openings to the U.S. (using the quad) lasted 15 to 30 minutes longer on each end of the period than with the Yagi. It is assumed that this phenomenon would also apply to each of the other quad models.

The fourth quad tested was a 3-element wide-spaced model constructed according to more dimensions furnished by W0HJ. The results of the samplings were somewhat disappointing and are given in Table 2. (Frankly, the author felt that the 3-element quad would show a substantial improvement over the Yagi in every case.) The 3-element model did appear to have a better front-to-back and front-to-side ratio than either the Yagi or the other quad models. One positive comment: The 3-element model is a monster to assemble and put up! In the author's opinion the difference in performance isn't worth the small improvement. Perhaps, on the other hand, if one accepts the 2-element model as the departure point between a simple mechanical structure and a major project, a 4-element model might be more worth the effort. However, this is purely conjecture on the part of the author.

Summary

The antenna tests indicate that:

1) One can expect to achieve the same or better results with a 2-element quad of proper dimensions than with a 3- or 4-element triband Yagi.

2) A wide-spaced quad will perform substantially better than a close-spaced quad.

3) Dollar for dollar, the quad appears to be a better investment than a Yagi.

Acknowledgements

The writer wishes to thank all of the amateurs who participated in the series of tests, particularly the VK gang, who night after night tolerated the request for observations. The support couldn't have been better, and on many occasions, upon completing a check with a particular station, several other stations would call to give their observations (which were taken during the same transmission test). This material was originally presented in *QST* by Col John H. Parrott, Jr., W4FRU.

References

[1]Orr, *All About Cubical-Quad Antennas*, Radio Publications, Wilton, CT 06897.

[2]Dimensions later published: Lindsay, "Quads and Yagis," *QST*, May 1968. (See chapter 3).

High Versus Low Antennas Performance Tests Using Identical Arrays

This section compares the performances of identical antennas mounted side-by-side at different heights, and produces some experimental evidence to support the practical importance of antenna height for most types of Amateur Radio communication.

"The higher the better" is the traditional rule of thumb for ham radio antennas. Few amateurs would disagree with this axiom, and most experienced antenna builders can recall how much better they got out after "raising the beam another 15 feet (4.57 m)." However, few empirical studies of the effect of antenna height on signal strength have been published in amateur circles.

This is unforunate, since several amateurs have studied the question methodically, including Dr. J. E. Lindsay, who has now published his definitive treatise on quads and Yagis[1] but not his excellent work in this area. And while both *The ARRL Antenna Book* and Orr's *Beam Antenna Handbook*[2] discuss the importance of antenna height in theoretical terms, neither reports the results of practical studies in the field.

With this in mind, the author set out to study the effect of height on signal strength at various distances, on several bands, and with various forms of propagation. Only amateur equipment was available for the study, but in other respects, the author strived to keep the methodology as rigorous as possible.

The Method

The author's approach was to erect identical antennas atop two towers of different heights and obtain comparative signal reports — with some safeguards to minimize reporting errors and to assure proper statistical treatment for the resulting data. For the main tests, one tower was 72 feet (21.95 m) high (the practical maximum for the author's residential backyard at the time of the tests), while the other stood 34 feet (10.36 m) high (about the minimum usable height for DX work, according to Orr). The two towers were placed as far apart as possible without trespassing — about 50 feet (15.24 m). With this separation, no interaction was evident on any band.

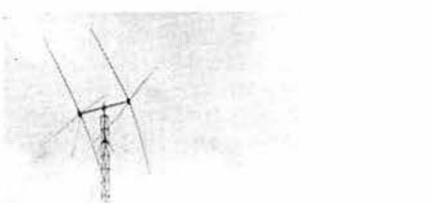

The two cubical quads used by the author to study the effects of antenna height on performance. The antenna at right is atop a 34-foot (10.36 m) tilt-over mast described by the author in June 1969 *QST*. The quad at left is on a 72-foot (21.95 m) tower.

A pair of 2-element cubical quads were selected for the 10-, 15- and 20-meter experiments. Each had an eight-foot (2.44 m) boom with all elements mounted concentrically on two sets of spreaders, and each antenna was fed with a single feed line (85 feet [25.9 m] of RG-8/U in both cases). This design obviously involves some compromises in element spacing and impedance matching, but both quads developed good SWR curves and front-to-back ratios when tuned. And, more important for our study, whatever compromises existed were essentially the same for both the high and low antennas.

It became apparent during the tests that many amateurs regard cubical quads as exceptional performers at low heights, but (unlike Yagis) not much better if raised higher. This popular impression is contrary to Lindsay's findings, which suggest that quads and Yagis respond quite comparably to changes in their height. Accepting Lindsay's conclusions, we assumed that our findings would be roughly the same if Yagis had been used for the primary experiments.

Table 3

Comparative Performance Data for High and Low Antennas

Band/path	*No. reports*	*No. favoring high antenna*	*No. favoring low antenna*	*No. reporting no difference*	*Mean high-antenna advantage*
20/DX	16	16	0	0	2.1 S units
20/U.S.	18	15	0	3	1.4 S units
15/DX	12	12	0	0	1.8 S units
15/U.S.	36	20	0	16	0.9 S units
15/ground-wave	12	12	0	0	2.2 S units
10/DX	8	7	0	1	1 S unit
6/ground-wave	5	5	0	0	2 S units
6/E-skip	10	2	7	1	−1.3 S units
2/ground-wave	16	16	0	0	3.3 S units

The author wanted to determine how well high and low antennas would perform not only on DX work, but on "Stateside" F-layer communications, ground-wave work, and E-skip as well. Thus, separate tabulations were kept for DX reports (i.e., those from stations more than 4000 miles away) and for reports from "Stateside" stations (i.e., F-layer reports over 1500—3000-mile paths) on each frequency band.[3] The author was careful to avoid seeking reports during anything that resembled an E-skip opening on 10, 15, and 20 meters, in the interest of studying E skip vs. antenna height separately on 6 meters.

To study the effect of height on vhf groundwave signals, the author replaced the cubical quads on the two towers with two pairs of seven-element 2-meter Yagis, each fed in phase with ratio-dipole driven elements. A move to a new home prevented the use of the same two towers for 6-meter experiments, but the author later set up two 3-element 6-meter Yagis — one atop a new 90-foot (27.43 m) tower, and the other on a 42-footer (12.8 m) — to continue the study.

On each band, the tests were conducted on ssb (except on 2 meters, where a-m was used), with an audio tone serving as a reference signal. The author switched between the high and low antennas repeatedly on each test, in an effort to counteract the effects of QSB.

This test procedure does introduce some sources of error, including the variability of hams' skills in observing the results, and the lack of receiver S-meter standardization. To cope with this kind of sampling fluctuation, the author obtained a number of reports on each band and then used t-distribution statistical methodology to evaluate the resulting data. Even with these efforts to make the study as rigorous as possible, however, the author makes no claim to finality for the results obtained. The findings on each band did prove to be statistically significant (at the 0.05 level of significance or better), but you don't "prove" things or arrive at exact parameters this way. The only claim the author makes for his results is that they suggest general trends that will probably be confirmed if this study is replicated elsewhere. There is no implication here that a 72-foot (21.95 m) high antenna produces "12-dB gain" over a 34-foot (10.36 m) -high antenna in 20-meter DX work just because the mean advantage in our tests was about two S units in this instance.

The Results

The results of these tests, summarized in the accompanying table, were interesting in several respects. With the exception of one case on 10 meters, every single DX station reported a substantially better signal from the high quad than the low one. This is thoroughly predicted in the literature, but we also found that the high quad produced a significantly stronger signal (taking the mean, or average, figure) on "Stateside" work. And on ground-wave work, the high antennas were absolutely consistent in outperforming lower ones. However, E-skip contacts proved to be another story, a point to be covered in detail later. The mean advantage of the high quad over the low quad in DX work was 2.1 S units on 20 meters, 1.8 S units on 15, and one S unit on the 10-meter band.

Since the author's linear amplifier (a commercial unit using two 3-400Zs) produces a signal gain of about two S units over the exciter alone on most S meters, it became apparent that many DX stations could not distinguish the kilowatt feeding the lower quad from the exciter "barefoot" with the higher array. The implications of this dramatic difference in signal strength for DX-contest operations are obvious: it's no wonder the "big guns" using antennas 130 feet (39.62 m) high get through the pileups quickly, with or without big linears.

On Stateside F-layer work, the pattern was less consistent, although the high array continued to hold a significant edge. While a majority of the reports still favored the high quad by one or two S units, more than a third of the stations dissented, reporting equal or virtually equal signals from the two quads. No one reported a stronger signal from the lower antenna, perhaps because of the author's effort to avoid E-ship contacts in this phase of the study. The mean advantage of the higher quad was 1.4 S units on 20-meter Stateside work, and 0.9 S unit on 15 — still a substantial advantage, but not enough to render a low antenna useless for portable work, as generations of Field Day operators can attest. In fact, the author previously described a portable quad[4] that stood only 19 feet (5.79 m) high, but nevertheless worked very well for "back-home" contacts during a long vacation. An antenna only 19 feet (5.97 m) high is clearly undesirable, but if higher antennas are not feasible, much successful hamming can still be done. (This is especially true on an occasion like Field Day, when almost everyone's antenna is low, and the few beams that are 40 [12.19 m] or more feet high seem to dominate the bands; if the competition is using a low antenna you can too!)

Ground-wave Tests

Although the plan was to study the effect of antenna height on ground-wave communication only on vhf bands, a number of stations within 100 miles volunteered reports on the hf bands. These reports were not tabulated except on 15 meters, but they consistently favored the higher array by at least two S units.

In the vhf experiments, height again proved decisive for gound-wave work. On both 2 and 6 meters, every station favored the high antenna. On two meters, the mean advantage was an amazing 3.3 S units. Surprisingly, the higher array appeared to be equally superior on contacts with San Bernardino-Riverside stations (some 30 miles from Beaumont, CA, where these tests were conducted) and on contacts into the Los Angeles basin (almost 100 miles away).[5]

After moving and acquiring the 90-foot (27.43 m) tower, the author put a single 7-element 2-meter Yagi up and took an S-meter reading on a nearby repeater with the new tower cranked down. As the tower ascended to its full height, the signal rose from S8 to 15 dB over S9 on a popular transceiver's S meter.

On 6 meters, only a few ground-wave tests were conducted, but the results were similar: Everyone favored the higher antenna substantially.

6-Meter E-Skip Tests

However, on 6-meter E skip an entirely different relationship developed: The *lower* antenna had the advantage in seven out of 10 tests during single-hop openings.[6] Only one station favored the high antenna, and the low array had a mean advantage of 1.3 S units over the high array!

Although this finding is surprising and contrary to every other test conducted, it is consistent with the one previous practical study of antenna height vs. signal strength the author could locate in ham radio literature. Working in the late 1940s, Leroy May[7] set up identical 6-meter Yagis at heights of 35 and 75 feet (10.67 and 22.8 m), and compared 350 received signals on the two arrays. He found that the low array was superior for distances below 1000 miles while the higher array began to pull ahead at distances exceeding 1100 miles.

The author made no attempt to fully replicate May's work, but all of our 6-meter E-skip experience supports him: For contacts into many major cities from greater Los Angeles, a high antenna is a handicap! Much to his chagrin, the author has frequently seen his 90-foot (27.43 m) - high Yagi outperformed on 6-meter E skip by a beam sitting 10 feet (3.05 m) above somebody's roof on a TV mast.

Conclusions and Recommendations

It is customary when reporting a study such as this to conclude by summarizing the findings and making some recommendations. The conclusion here must be that, for every application except E-skip work, there are worthwhile advantages of having an antenna at least 70 feet (21.34 m) high. This height advantage is most pronounced on ground-wave and long-haul DX work.

The only logical recommendation we can make, then, is that all hams should own towers 70 feet (21.34 m) high. May all of the XYLs, frightened neighbors, bill collectors, and zoning-conscious city officials of the world forgive us!

Alternately, maybe we should suggest that all hams (except the author) use 35-foot (10.67 m) towers to give the author a better chance in DX contests. Or better yet, how about an antenna-height multiplier for the DX contests? Let's see, that's 1.5 if your antenna is below 75 feet (22.86 m), 2.0 if it's below 40 feet (12.19 m), and 3.0 if you bury it in the ground . . . This material was originally presented in *QST* by Wayne E. Overbeck, N6NB.

References

[1] Lindsay, "Quads and Yagis," *QST*, May 1968 (See chapter 3).

[2] Orr, William I., *Beam Antenna Handbook*, 3rd Edition.

[3] No 10-meter "Stateside" tests are reported because the muf was too low for reliable F-layer work across the U.S. when these tests were run.

[4] Overbeck, "The 20-Minute Portable Quad," *QST*, May 1967.

[5] On the hf bands, it is normally assumed that a high antenna has a greater advantage over a low antenna on long-haul DX than it does on shorter paths, but the opposite may well be true for vhf ground-wave work. This is suggested in the *ARRL VHF Manual*, (1st Edition, pages 24,25). Thus, one would expect that an increase in height would result in greater improvement in signals over the 30-mile path than over the 100-mile path.

[6] At the time of the 6-meter tests, E skip was barely "in season." The author is continuing this study during seasons of more frequent E-skip openings.

[7] May, "Some Practical Results on the Question of Antenna Height," *CQ*, April 1953.

Limited-Space Antennas and Methods Of Coupling

When short, random-length wire antennas prove disappointing, chances are the installer didn't follow a few basic rules for good performance. These tips should lead the way to better results with "randoms."

One very popular type of antenna, probably because of its simplicity, is a wire of either a predetermined or random length, fed at one end. It is common procedure among amateurs these days to refer to such an antenna as a "long wire." Actually, this term is usually incorrect because, by definition, a long-wire antenna is a wire that is several wavelengths long at the operating frequency. But no matter what you call it, the end-fed wire can be a very simple antenna to put up and use, if certain steps are taken.

Transmitter Loads

Nearly all transmitters described these days are designed to work into 50- to 70-ohm loads, with little thought given to any load that departs very far from these figures. If the load is something different than the above figure it may be impossible to get the amplifier in a transmitter to load up. In order for the final amplifier stage to operate at its best efficiency, the load (the antenna system) must be within the design range of the amplifier.

An end-fed wire may present a load of 50 to 70 ohms, but if it does it will be a case of pure luck on the part of the user — and what's more important, that load won't stay the same across a band. In other

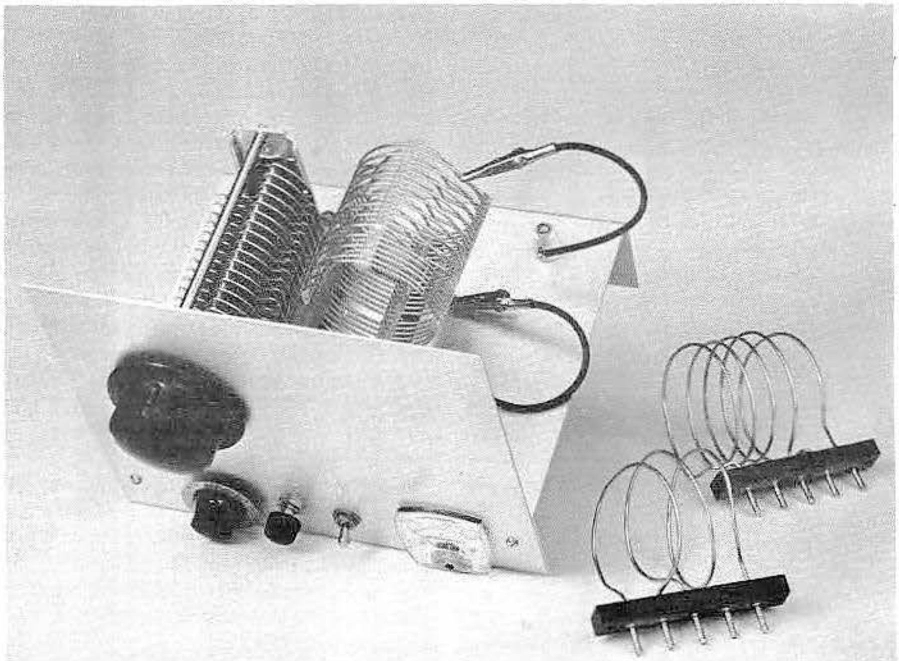

The completed transmatch with the 80/40-meter coil plugged in. At the rear is the 20-meter coil and the 10/15-meter unit is to the front. In regular operation, the antenna would be attached to the standoff insulator at the top rear of the chassis. The clip lead visible below the coil is the one attached to the jack bar.

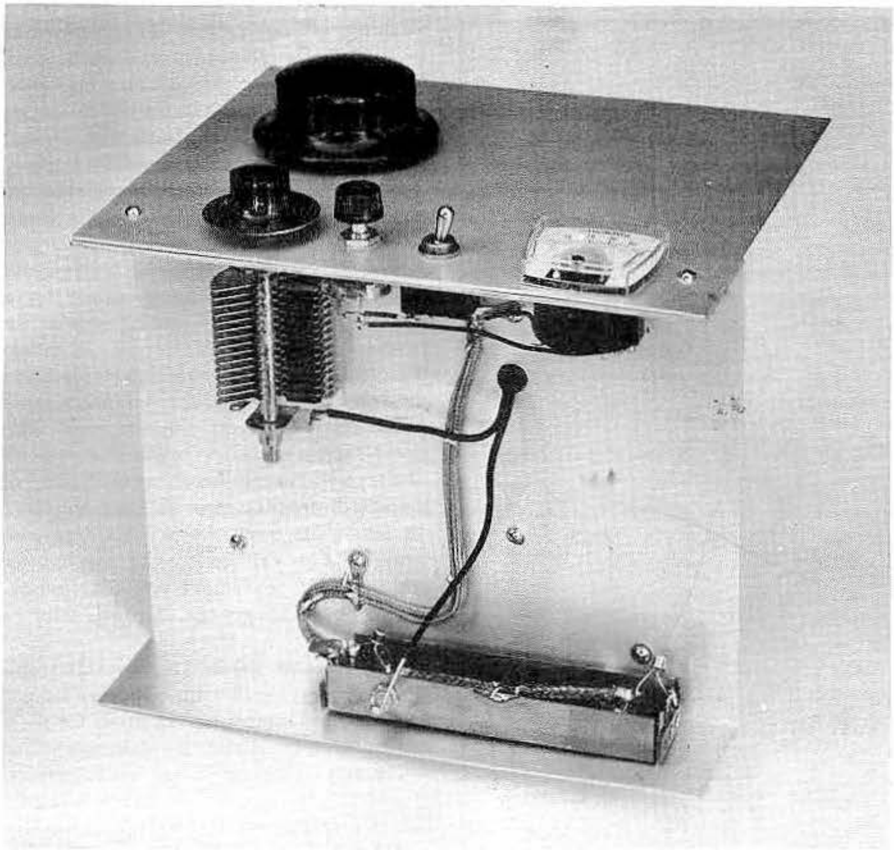

The wiring below deck. At the rear of the chassis is the Varimatcher section. C2 is at the left front of the chassis.

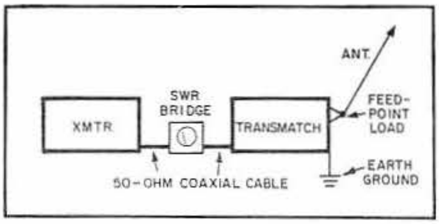

Fig. 3 — This block diagram shows the set-up for using a Transmatch. Not shown is an antenna relay which normally would be inserted in the coaxial line. If a low-pass filter is used, it should be the last item in the line before the Transmatch.

words, the antenna will present a 50- to 70-ohm load for only a very limited frequency range. Once we QSY the load may change to where it is no longer within the tuning range of the amplifier. Our problem boils down to having the transmitter *always* see a load of 50 to 70 ohms, regardless of the band or frequency in use. This in turn means we must make the antenna system look like a 50- to 70-ohm load, no matter which band we use it on.

The Transmatch, an Adjustable RF Transformer

Let's make one point clear about end-fed wires, or any antenna, for that matter: There is no way of predicting what the feed-point impedance (load) will be before making and installing the antenna. We can come close, but there are many factors that get into the act to determine what the impedance will be. The antenna's height, its proximity to nearby objects, its length, and the antenna's relation to earth ground are just a few of the factors. If we don't know what the feed-point impedance is, how can we make the transmitter always work into a 50-ohm load? The easiest way is to insert a Transmatch between the transmitter and the feed point. A Transmatch is simply an adjustable rf transformer that takes our unknown feed-point load and converts it to a desired load, 50 ohms in the case shown in Fig. 3.

Note in Fig. 3 that we show a random-length antenna and an earth ground connection. In this case, our unknown load exists between the end of the antenna and ground. On the transmitter side of the Transmatch we have 50-ohm coaxial cable and inserted in this cable is a 50-ohm reflectometer or SWR (standing-wave-ratio) bridge. The reflectometer is simply a measuring instrument that tells us when our Transmatch is properly adjusted to convert our unknown load to a 50-ohm load. If the Transmatch has adequate matching range, we can convert any feed-point load to 50 ohms. And this means that we can take any random length wire and make a working antenna out of it for any amateur band or frequency. There are some practical limits as to how short a wire we can use, but actually the antenna can be quite short in relation to a wavelength and still work.

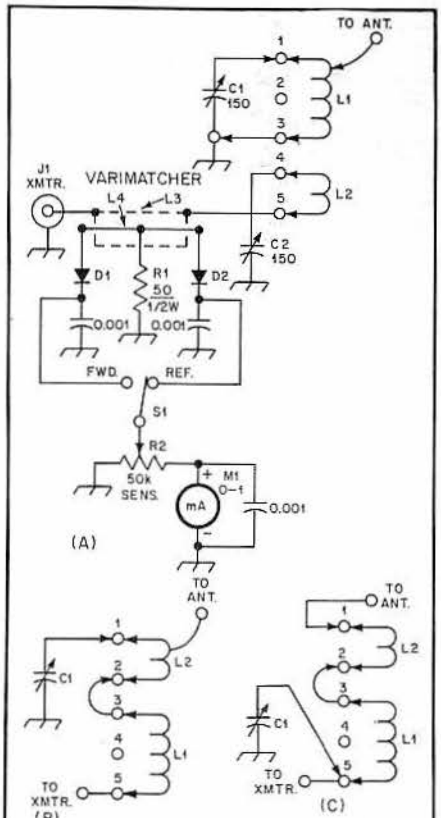

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (μF) ; OTHERS ARE IN PICOFARADS (pF OR μμF); RESISTANCES ARE IN OHMS ; k = 1 000, M = 1 000 000.

C2 — 150-pF variable (Millen 19140 or similar).
D1, D2 — 1N34A germanium diode.
J1 — Coax chassis fitting, type SO-239.
L1, L2 — See Fig. 7.
L3, L4 — See Fig. 6.
C1 — 150-pF variable, (air gap 0.077 inch [2 mm] or larger for high power, Millen 12515 or similar; air gap 0.040 inch [1 mm] for low power, Millen 19140W or similar).
M1 — 0-1 milliammeter (more sensitive meter can be used).
R1 — 50-ohms, 1/2-watt carbon or composition, not wire-wound.
R2 — 50 kΩ control, linear taper.
S1 — Single-pole, single-throw toggle switch.

Fig. 4 — Circuit diagram of the Transmatch and Varimatcher. All 0.001-μF capacitors are disk ceramic.

Limited-Space Antennas

Many hams live in apartments or locations where it is impossible to put up an "antenna farm" or even a half-wave dipole for 80 or 40 meters. Random-length wires are sometimes the best bet in these circumstances. For hams who live under such conditions there are some general rules that can be followed for best performance. If possible, always get the antenna up on the roof. If not, get the antenna as high as possible. Make the wire as long as possible, even if it has to go around corners. Don't overlook existing antenna possibilities, such as rain gutters or roof flashing. In a moment we are going to describe a Transmatch that can be used to couple practically any wire or metal structure to the transmitter, including rain gutters.

If it is impossible to get an outdoor antenna up (and don't overlook "invisible" antennas[1]). an indoor antenna will work, but not as well as an outdoor one. With the Transmatch to be described, you can run a wire around the ceiling molding and couple the wire to the rig, maintaining that 50-ohm load we mentioned earlier. When you hear hams talking about loading up the bed spring, don't laugh — many a ham has made contacts that way. All you need is some method of coupling the bed spring to the rig; that's what our Transmatch will do.

The Transmatch

Shown in Fig. 4 is the circuit for a Transmatch that will do the job we've been talking about. This unit will match practically any wire to our 50-ohm rigs on any band from 80 through 10 meters. Depending on which capacitor plate spacing you use for C1 the Transmatch can be used for Novice powers or up to the legal limit of 1 kW. Actually, the closer-spaced capacitor will handle 200 watts or so. The same type coil stock is used for either power level, high or low.

Also included in the Transmatch shown in Fig. 4 and the photographs is an SWR bridge. This is a necessary item in order to know when the Transmatch is correctly adjusted. If you already have a Monimatch or reflectometer, the unit described can be eliminated from the Transmatch.

With the random-wire antenna we'll be dealing with various loads, depending on the frequency and band in use. Shown in Fig. 5 are the three basic circuit configurations used with the Transmatch. The circuit can be changed to conform with type A, B, or C by reversing the plug-in coil in the jack bar or by changing the capacitor C1 from the output side of the L configuration, as in B, to the input side, as in C. The capacitor is changed by means of a clip lead connected to the stator.

Some loads may require the circuit setup of A. In this configuration, a link-coupled parallel-tuned circuit is used, the antenna being tapped onto L1.

Construction Details

The chassis used to hold the parts is 2 × 7 × 9 inches (51 mm × 178 mm × 229 mm) and the panel measures 7 × 9 inches (178 mm × 229 mm). Any chassis large enough to hold the parts can be used. C1 and the jack bar for the plug-in coils are mounted on top the chassis, and the Varimatcher reflectometer and C2 are mounted below the chassis top. The remaining components are mounted on the chassis front.

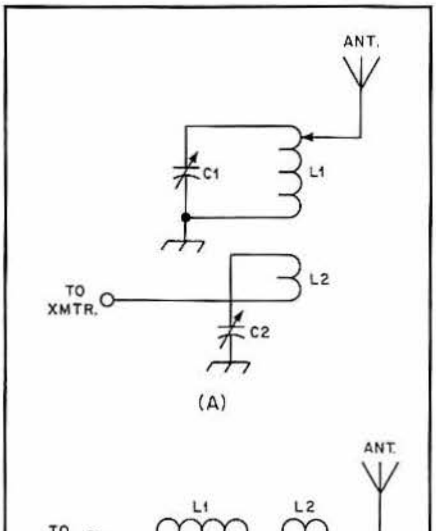

Fig. 5 — Three possible configurations with the Transmatch as described in the text. To set up types B or C the plug-in coil is reversed in the jack bar and terminal no. 3 of the jack is clipped to terminal no. 2.

A 4-inch (102 mm) length of RG-58/U coaxial cable is used for the pickup and conductor sections of the SWR bridge. Details for making this section are given in Fig. 6. Study these details and the photograph of the bottom view carefully before starting construction. The black outer covering of the coax is removed, leaving the braid and inner conductor with its insulation. The outer braid will be L3 and the inner conductor L4. You'll notice that there is an open area at the center of the coax braid. This is to allow room for the connection of one end of the 50-ohm resistor to the center of the inner part of the coaxial line. Use a probe or pointed instrument to open up the area in the braid and then carefully tin the edges of the opening with solder so as to prevent any of the hair-like wires from shorting to the resistor lead or inner conductor. The ends of the braid can be wrapped with a few turns of solid wire and tinned. The wire ends will provide a support for the connection to the feed-through terminal and the inner pin of J1.

The metal trough that houses the bridge section is made from a piece of copper flashing, although tin or aluminum could be used. It is a good idea to drill the holes for J1 and the feed-through bushing before bending up the trough into a U. Also, when soldering the leads for the diodes, D1 and D2, use a heat sink between the body of diode and the point being soldered, because too much heat can ruin the diode.

Using the Transmatch

In making the antenna, shoot for a length of at least 1/4 wavelength at the lowest operating frequency. In other words, if you plan on working 80 meters as the lowest band, then the antenna should be 60 feet (18.29 m) long, at least. This doesn't mean that shorter wires won't work — they will, but not nearly as well. A lot of hams will say they can't get up a length of 60 feet! The wire doesn't *have* to run straight. You can go around corners, up and down, or what have you. The important point is to get up as much wire as possible.

By the same token, many hams think (mistakenly) that they have to have a good earth ground connection in order for their station to work. It's fine if you have one but if you don't, forget it. If you have access to a water pipe or other earth ground connection make a connection to the rig and the Transmatch, but if you can't, don't worry. The antenna will still work.

With the antenna connected and the rig on the desired frequency, feed enough power through the system to get at least a half-scale reading on the Varimatcher meter, and have the Transmatch set up as in (A). Tap the antenna onto the first turn of the coil L1 at the "hot" end, or end opposite the ground. Incidentally, always turn the rig off when making any adjust-

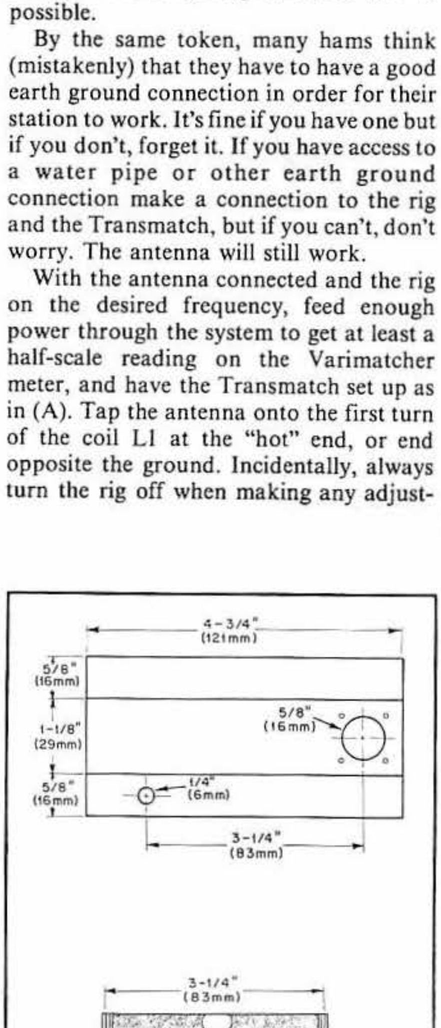

Fig. 6 — Constructional details of the Varimatcher bridge section. The braided portion of the coax is L3 and the inner portion is L4.

---

[1] Any wire that is no. 28 or smaller is practically impossible to see when strung up as an antenna. Many hams use "invisible" antennas to prevent a touchy landlord situation.

ments on the Transmatch. There may not be any dc voltages present but some very, very high rf voltages can be developed on the Transmatch coils and can cause a nasty rf burn.

With the antenna tapped on near the hot end of L1 switch S1 to read reflected power and adjust C1 and C2 for a dip or minimum reading. If you cannot get the reflected reading down to zero, or close to it, turn off the power and move the antenna tap closer to the ground end of the coil. You'll eventually find a tap point that will give a reading of zero in the reflected position versus full scale in the forward position of S1. If you find that the tap point on L1 is very close, say within 25 percent, to the ground end, you may want to try configuration B or C.

In eight of these configurations, start at the antenna end of the coil and, using another short clip lead, short out turns or portions of turns as you move the antenna tap towards the transmitter end of the circuit. In other words, the antenna is always on the end of the coil, but you may have some of the coil shorted out in order to get a match as indicated by M1.

Once you get a match for any given band, make a chart of the taps and so forth so you can return to the setup when you change bands.

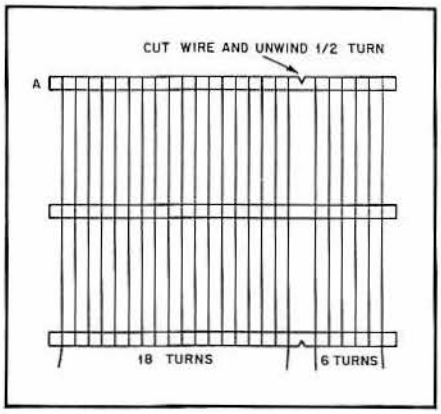

Fig. 7 — The 80/40-meter coil. To use the coil on 40 meters, short out 10 turns, counting from the "A" end of the coil. Shorting clips should be copper and the shorting leads should be insulated. Coil stock is no. 14 wire, 8 turns per inch (25 mm), 2-1/2 inches (64 mm) in diameter (Air Dux 2008T, B & W 3906-1, or Polycoils 1775).
The 20-meter coil consists of 4 turns of no. 12 solid wire, 2-1/2 inches in diameter, with the turns spaced so that L1 is 1-1/2 inches (38 mm) long. L2 is one turn of no. 12, same diameter as L1, spaced 1/2 inch (13 mm) from the end of L1. The 15/10-meter L1 coil is 2 turns of no. 12 solid wire, 2-1/2 inches (64 mm) in diameter, turns spaced to cover 1-1/2 inches (38 mm). L2 is a single turn, same diameter as L1, spaced 1/2 inch (13 mm) from L1.
All coils are mounted on Millen type 40305 plugs and the socket (jack bar) is a Millen 41305. See G. R. Whitehouse ads in *QST*.

The main reason for using configurations B and C is that if the antenna is tapped near ground on L1 in configuration A, extremely high rf voltages can develop across C1 even with Novice powers. We don't recommend playing around with "arcs" from coils but it might be wise to prove to yourself just how much voltage can be developed. With the antenna tapped near the ground end of L1, take a pencil and hold it by the wood, not at the eraser end but in the center. With the key closed and the rig loaded up to about 75 watts, touch the lead end of the pencil to the stator of C1. You'll quickly see why caution is necessary. Such an arc can make a very nasty and painful burn.

If you want to try using antennas that are *very* short for the frequency, such as 10 feet (3.05 m) long on 80 meters (similar to a mobile whip) certain additional steps may be required to obtain a match with the Transmatch. When you make the coil for 80 and 40 meters you'll have some coil stock left over. With very short antennas, this can be used as a "loading" coil. Connect the antenna to one end of the coil and connect the other end to the antenna terminal on the Transmatch. Then go through the tune-up procedure. You may have to use a clip lead to short out turns on the coil but you'll find a setup that will permit matching with the Transmatch. Again, avoid coming in contact with the series loading coil because very high rf voltages can develop in such a setup.

Some beginners may not be aware of it, but a fluorescent lamp or neon bulb will light up when in an rf field, if the rf voltage is high enough. If you have a fluorescent lamp (a burned-out one is OK), the lamp can be laid along the antenna wire where it leaves the Transmatch. This makes a good output indicator, showing when power is flowing in the antenna. This material was originally presented in *QST* by Lewis G. McCoy, W1ICP.

Off-Center-Loaded Dipole Antennas

So you can't manage a buried radial system on your city lot, eh? Then why not scrap those plans for a vertical antenna and try K1TD's version of a short dipole? Here are the design tips you need.

In these times when much of our amateur population lives in urban areas, the subject of shortened antennas for the lower frequency amateur bands is a very popular one. Physically short ground-mounted vertical antennas with lumped-constant loading to make them resonant can be quite efficient radiators, if a good radial system has been installed. This has certainly been evidenced in Sevick's series of *QST* articles.[1] To many amateurs, however, the "hitch" in constructing such a system is the installation of a good radial system. It must be admitted that for the "top" amateur bands, 160 and 80/75 meters, an efficient system of buried radials requires a sizable amount of real estate, even for a physically short radiator. On the average city-size lot, 50 or 75 by 120 to 150 feet (15.24 - 22.86 m by 36.58 - 45.72 m), it's almost impossible to install a highly efficient radial system for 80/75 meters, much less for 160 meters, when structures like a house and perhaps a separate garage exist. Or to some amateurs, just the thought of burying hundreds or maybe thousands of feet of wire is enough to turn off any enthusiasm for the project.

What's the alternative? A dipole type of antenna with lunped-constant loading. At modest heights, 30 or 40 feet (9.14 or 12.19 m), such an antenna will prove to be quite satisfactory if it is physically longer than about 0.2 wavelength. Shorter lengths may also be used, at reduced efficiency. Such an antenna can be fed directly with 50-ohm coaxial line, and it can be operated with no earth ground. (Of course the chassis of the transmitter and/or receiver should be grounded adequately for protection against shock hazard.)

Nearly all of us are familiar with the concept behind the use of inductive loading. A vertical antenna which is shorter than a quarter wave (or a dipole antenna which is

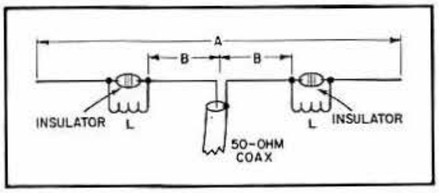

Fig. 8 — A dipole antenna lengthened electrically with off-center loading coils. For a fixed dimension A, greater efficiency will be realized with greater distance B, but as B is increased, L must be larger in value to maintain resonance.

(Eq. 1):

$$L_{\mu H} = \frac{10^6}{68\pi^2 f^2} \left\{ \frac{\left[\ln \dfrac{24\left(\frac{234}{f} - B\right)}{D} - 1\right] \left[\left(1 - \dfrac{fB}{234}\right)^2 - 1\right]}{\dfrac{234}{f} - B} - \frac{\left[\ln \dfrac{\left(24\,\frac{A}{2} - B\right)}{D} - 1\right] \left[\left(\dfrac{\frac{fA}{2} - fB}{234}\right)^2 - 1\right]}{\dfrac{A}{2} - B} \right\}$$

where

$L_{\mu H}$ = inductance required for resonance

ln = natural log

f = frequency, MHz

A = overall antenna length, feet

B = distance from center to each loading coil, feet

D = diameter of radiator, inches

shorter than a half wave) will exhibit capacitive reactance at its base (or center) feed point. To cancel such capacitive reactance, a coil having the proper inductive reactance may be connected in series with the base feed point of the vertical. The same result will be obtained through the use of two such coils for a dipole, one coil connected in series with each half. It is not necessary for the inductor to be installed at the feed point, however. In fact greater radiating efficiency results through improved current distribution if the inductor is located along the radiator some distance away from the low-impedance feed point, *viz,* in the manner of a center-loaded mobile whip antenna. Fig. 8 shows this concept extended to a dipole element, with off-center loading. The inductors resonate the antenna to the operating frequency, but do little actual radiating themselves. (This is in contrast to helically wound or continuously loaded elements, where a long thin inductor is the radiator as well as the loading element.)

In the antenna represented by Fig. 8, there are many variable factors to be considered when a practical antenna for a given frequency is being constructed. Of primary consideration from an efficiency standpoint is the overall length, shown as dimension A. Another consideration for efficiency is the distance of the coils from center, dimension B. The longer the overall length (A), up to a half wave, and the farther the loading coils are placed from the center (B), the greater is the efficiency of the antenna. However, the greater is distance B (for a fixed overall antenna size), the larger the inductors must be to maintain resonance. Theoretically, if the coils were placed at the outer ends of the dipole, they should be infinite in value to maintain resonance. Capacitive loading of the ends, either through proximity of the antenna to other objects or through the addition of capacitance hats, will reduce this requirement to a more practical value.

What Inductance Values?

As a matter of personal interest, this writer has been doing experimental work for a number of years with off-center-loaded antennas. One big drawback to such experimentation was the ever-present need for a large amount of cut-and-try work to arrive at resonance whenever a new set of dimensions was to be used. Probably the number of pruned-off turns from coil stock from such experiments, if straightened out and soldered end to end, would make up several full-sized half-wave antennas for the 160-meter band. Therefore, most of the writer's work of late in this area has been in going through paperwork exercises, looking for a way whereby at least "ball-park" values of inductance needed for a particular system could be calculated.

The equation contained in the Mobile chapter of *The ARRL Antenna Book* for determining the capacitance of a vertical antenna shorter than a quarter wavelength looked promising in early computations, and, indeed, it became the basis for the calculation procedure which finally resulted. This procedure has been found to produce results much closer than mere "ball-park" values for the necessary inductance — for wire antennas "in the clear" at moderate heights, the final inductance values found by cut-and-try pruning for lowest SWR at the desired frequency have been so close to the value from calculations that a laboratory bridge was necessary to measure the difference. The results are equally good for elements using tubing. Once the needed inductance value is determined by calculations, it is generally found sufficient to obtain coil dimensions from an ARRL L/C/F Calculator (see LaPlaca[2]) or by equation. Any significant pruning which has been found necessary could always be attributed to objects in proximity to the ends of the antenna.

The complete set of calculations is expressed in the mathematical relationship as Eq. 1, presented here primarily for mathematics buffs or those having access to electronic computers. This equation yields the inductance required, in microhenrys, for single-band resonance of a shortened antenna of a particular physical size at a given frequency, for a specific position of the loading coils from the center of the antenna. To spare the reader the task of performing some rather tedious calculations, Fig. 9 has been prepared from Eq. 1. The curves of the chart have been normalized, and may be used for any frequency of resonance. The chart is based on a half-wavelength/diameter ratio of the radiator of approximately 24,000. (This corresponds to no. 14 wire on 80 meters or no. 8 wire on 160 meters.) For "thinner" conductors, the required inductance will be somewhat greater than that determined from Fig. 9, and less inductance will be required for "thicker" conductors.

The use of the chart is as follows: At the intersection of the appropriate curve from the body for dimension A and the proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at the left. Dimensions A and B are shown in Fig. 8, and for use with the chart are expressed as percentages. Dimension A is taken as percent length of the shortened antenna with respect to the length of a resonant half-wave dipole of the same conductor material. Dimension B is taken as the percent of coil distance from the feed point to the end of the shortened antenna. For example, resonating an antenna which is 50 percent or half the size of a half-wave dipole (one-quarter wavelength overall), with loading coils positioned midway between the feed point and each end (50 percent out), would require loading coils having an inductive reactance of approximately 950 ohms at the operating frequency. If the antenna is hung "in the clear," and if the length/diameter ratio of the conductor is near 24,000, inductance values as determined from the chart will be very close to actual values required. (Eq. 1 above takes the diameter of the radiator into account, and thus may be used for any length/diameter ratio.) For practical

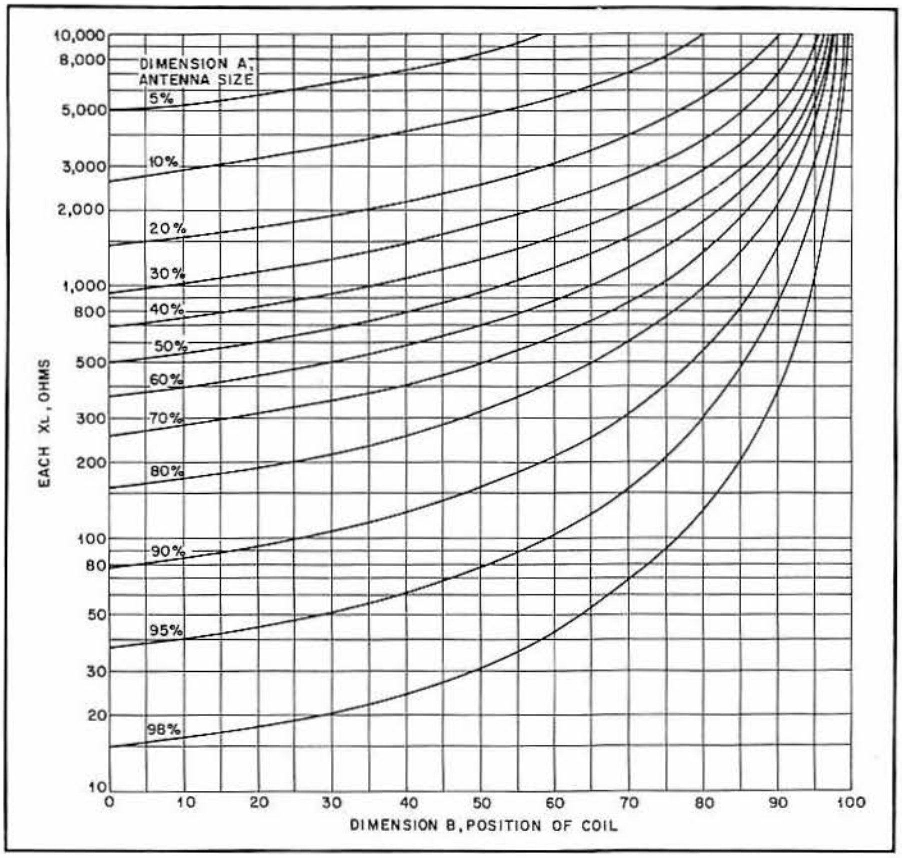

Fig. 9 — A chart for determining approximate inductance values for off-center-loaded dipoles. At the intersection of the appropriate curve from the body of the chart for dimension A and the proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at left. See Fig. 8 regarding dimensions A and B.

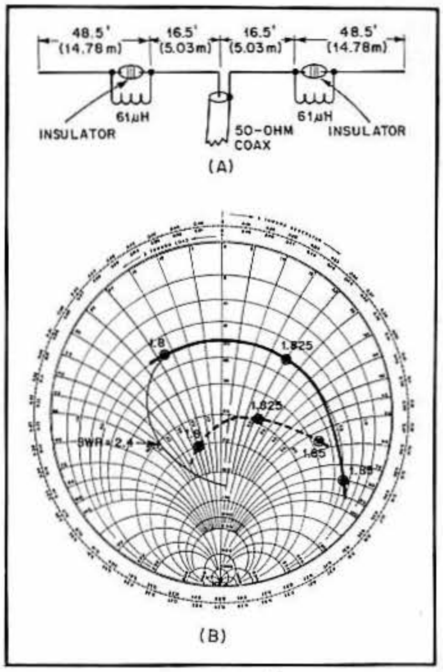

Fig. 10 — At A, an 80-meter dipole loaded for 160-meter operation. The inductors are 37 turns of coil stock having a 3-inch (76 mm) dia and 10 turns per inch (25 mm) (B & W 3035). At B, the impedance plot of this antenna installed at a height of 50 feet (15.24 m) (solid curve), and the plot of a 160-meter half-wave dipole (broken curve).

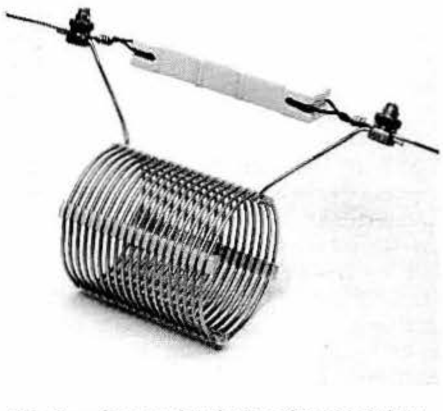

Fig. 11 — Copper electrical service connectors, sold under one trade name of Servit, provide a simple means of installing the loading coils. The antenna wire and the ends of the coil wires should be tinned to prevent corrosion. In addition, a protective coating of acrylic spray may be used at each connection.

purposes, dimension B may be taken as that distance from the center of the feed-point insulator to the inside eye of the loading-coil insulator, and dimension A as the eye-to-eye distance inside the end insulators (which are not drawn in Fig. 8).

Proximity of surrounding objects in individual installations may require some pruning of the coils, and the exact amount of final inductance required should be determined experimentally. If the antenna is hung in inverted-V style, with the ends brought near the earth, the required inductance will almost always be somewhat less than that determined from the chart or equation. A grid-dip meter, Macromatcher (see Hall and Kaufmann[3]), or SWR indicator may be used during the final adjustment procedure.

Practical Antennas

Although one might erect an inductively loaded antenna that is cut for a single amateur band, it is possible to use the antenna itself for two, three or more bands of operation, if provision is made to lower the antenna for band changes. A simple rope halyard and pulley arrangement at one of the supports will do the trick. Fig. 10A shows a three-band antenna of this nature, for 160, 80 and 20 meters. If the insulators shown are left open, with nothing bridging them, the antenna is a simple half-wave dipole cut for 14.18 MHz. (The 48.5-foot [14.78 m] lengths act merely as support wires, and have negligible effect on operation of the antenna.) If the insulators are bridged with short lengths of antenna wire, the antenna becomes a center-fed 80-meter dipole, resonant at about 3.6 MHz. For 160-meter operation the 20-meter insulators may be bridged with loading coils to resonate the antenna at 1.8 MHz, as shown in Fig. 10A. Burndy or other manufacturers' "Servit" type of electrical connectors may be used for ease in making band changes quickly, as shown in Fig. 11.

The calculation procedure for determining loading-coil values for the antenna of Fig. 10A, using the chart of Fig. 9, goes like this. If operation is desired on 1.8 MHz, the length of a full-sized half-wave dipole is found from the relationship 468/f to be 260 feet (79.25 m). The 130-foot (39.62 m) length of Fig. 10A represents 50 percent of this size, meaning that the dimension-A curve marked "50 percent" in Fig. 9 is to be used. The position of the coils is 16.5/(16.5 + 48.5) × 100, or 25 percent of the distance out from center, dimension B. From the intersection of 25 (horizontal scale at bottom) and the 50 percent curve, the required inductive reactance is read from the scale at the left of Fig. 9 to be 650 ohms. The inductance, L, is $650/2\pi f$ or 57.5 microhenrys, if no. 8 wire is to be used. For smaller diameter wire, the inductance should be somewhat larger. (Calculations from Eq. 1 for no. 12 wire indicate the required inductance is 60.99 µH.)

The radiation resistance of a shortened antenna loaded to resonance is less than that of a full-sized antenna. Further, the shortened antenna is "sharper," meaning that the change in reactance versus frequency is greater. In other words, the shortened antenna acts as a tuned circuit having a higher Q than a full-sized antenna.

To check these characteristics, the line input impedances for the antenna of Fig. 10A were measured with a laboratory bridge, and the electrical line length at the measurement frequency was then taken into account to determine the impedance at the antenna feed point. The antenna was constructed of no. 12 wire and hung at a height of 50 feet (15.24 m) as a "flat-top" radiator.

The solid curve of Fig. 10B is a plot of the feed-point impedance versus frequency for this antenna. The plot on Smith Chart coordinates is more meaningful than a simple SWR-vs.-frequency curve because the magnitudes of the resistive and reactive components are shown, as well as the sign of the reactance. (Capacitive reactance is negative, plotted to the left of the vertical center line, and inductive reactance is positive, plotted to the right.) In this presentation, a 50-ohm nonreactive impedance will appear at the exact center of the chart. The SWR in 50-ohm line for a given frequency may be determined by first noting the distance from the center of the chart to the particular impedance plot on the curve, and next measuring this same distance down the vertical center line from chart center (a drawing compass is helpful for this task), and finally dividing 50 into the value read at that point on the center line. For example, the SWR at 1.8 MHz equals 120/50 or 2.4, as indicated by the segment of the 2.4 SWR circle in Fig. 10 B. It may be seen that resonance (zero reactance) occurs at approximately 1810 kHz, where the resistance is about 22 ohms. The SWR at resonance is 2.33:1, and climbs to 3:1 at 1825 kHz. At 1850 kHz, the SWR is 10:1. Without any matching provisions the antenna is relatively sharp, as mentioned earlier. If one sets the usable bandwidth as the frequency range where the SWR is 3:1 or less, it is approximately 35 kHz, or 1.9 percent of the resonant frequency. As far as efficiency is concerned, ohmic losses are low, and the antenna is a good performer on 160 meters. Because of its horizontal polarization, it has proved to be most effective at night, and stations several hundred miles away have been worked with S 9 reports received for the 50-watt signal.

For a comparison of impedances, the broken curve of Fig. 10B is a plot of measured impedances of a full-size half-wave dipole, 260 feet (79.25 m) long overall, hung in place of the shortened antenna. From this curve it may be seen that resonance occurs at 1810 kHz, where the resistance is 59 ohms. The 3:1-SWR bandwidth for the half-wave antenna is in the order of 60 kHz, or 3.3 percent of the resonant frequency. It is interesting to note on this curve that the SWR at resonance is 1.18:1, and that it is a somewhat lower value, 1.15:1, at a frequency a few kilohertz above resonance. (Measurements were made every 5 kHz across this band, but plot points are shown only for 25-kHz increments to avoid crowding of the data.) This evidence refutes the oft-heard statement that the SWR-vs.-frequency curve is *always* lowest at antenna resonance. Points to remember are that the SWR in a transmission line is completely dependent upon the characteristic-impedance value of the line in use. Using a line of different impedance may shift the position of the SWR curve along the frequency axis in a simple SWR-vs.-frequency plot. This is definitely true in this case — if the 160-meter half-wave dipole were to be fed with 75-ohm line, the lowest SWR would occur at a frequency about 5 kHz *below* antenna resonance, whereas with 50-ohm line the lowest SWR is at a frequency slightly *above* resonance. The reason this happens is that the resistive component of the impedance, which consists of the radiation resistance plus any loss resistance, is not constant with frequency, even over a rather narrow frequency range. It must be acknowledged that the differences here are very slight, however, and for practical purposes the frequency of lowest SWR is (within a few kilohertz) the resonant frequency of the antenna.

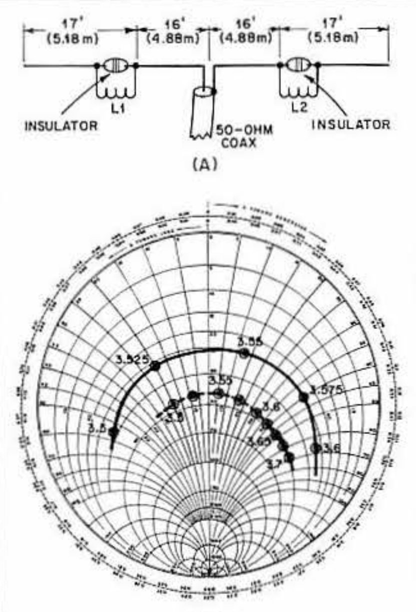

Fig. 12 — At A, a 40-meter dipole loaded for 80-meter operation. For resonance at 3.55 MHz the coils should be approximately 40 μH (27 turns of stock); 3.75 MHz, 35 μH (24 turns); 3.9 MHz, 31 μH (22 turns); and 4.0 MHz, 29 μH (21 turns). These are calculated inductance values for no. 12 antenna wire. Coil stock referenced above is 3-inch (76 mm) diameter, 10 turns per inch (25 mm) (B & W 3035). At B, the impedance plot of the 3.55-MHz version (solid curve) and of an 80-meter half-wave dipole (broken curve).

Another point concerning the SWR values bears noting. The values as determined from the plots in the manner described above are quite accurate, having been determined by measurements with laboratory equipment. In contrast, measurements with simple SWR indicators usually cannot be relied upon for anywhere near the equivalent accuracy.

For example, the author owns a commercially manufactured SWR indicator of the Monimatch type (see McCoy[4]) which, under a particular set of conditions, indicates a 2.5:1 SWR in a line where laboratory measuring equipment shows the *true SWR to be 4:1.* A significant difference! Herein lies another reason why impedance plots on Smith Chart coordinates are more meaningful than a simple SWR-vs.-frequency curve — greater accuracy may generally be expected.

A Half-Size 80-Meter Antenna

Fig. 12A shows the 3-band concept described earlier as it can be applied to 80, 40 and 20 meters. Its overall length is 66 feet (20.12 m), not a difficult length to use on a small lot. This antenna was constructed for 80-meter operation with a design-center frequency of 3.55 MHz, using no. 12 antenna wire and 40-μH loading coils — 27 turns of stock having a diameter of 3 inches (76 mm) and a pitch of 10 turns per inch (25 mm) (tpi). Feed-point impedances versus 80-meter frequency for the antenna, hung at a height of 50 feet (15.24 m), are shown by the solid curve at B of Fig. 12. Actual resonance occurred at 3.54 MHz, where the resistance was about 26 ohms. The bandwidth within which the SWR is 3:1 is 60 kHz, or 1.69 percent of the resonant frequency.

Also shown in Fig. 12B, by the broken curve, are the feed-point impedances of a half-wave dipole, 132 feet (40.23 m) overall length, hung in place of the shortened antenna. Resonance occurs at 3.54 MHz, where the resistance is 43.5 ohms and the SWR is 1.15:1. The broader nature of the half-wave antenna is exhibited by the "tighter" curve which swings closer to the 50-ohm center point of the chart than the shorter, loaded antenna. The SWR at 3.5 MHz is 1.6:1, and remains below 3:1 to 3.67 MHz.

Capacitive and Inductive Loading

One would assume that a combination of capacitive and inductive loading might provide a different feed-point impedance than would inductive loading alone, because of different current distributions in the radiators. To check out this assumption, the antenna of Fig. 12A was used as a "test bed" for comparative measurements. Capacitance hats were attached at different points along the 17-foot (5.18 m) lengths of wire outside the coils, and the coils were pruned to reresonate the antenna at about the same frequency as before. The impedance measuremnts were then repeated.

Dangling End Sections

First, "hats" consisting of 18 inches (457 mm) of no. 12 wire were affixed to the antenna ends and permitted to dangle. This lowered the resonant frequency to 3435 kHz. By calculations, this was approximately the same effect as that of extending the 17-foot portions of the antenna by the same amount as the dangling lengths, so it

Table 4

Characteristics of Various Loading Techniques, 66-Foot, 80-Meter Dipole

Loading	Approx. feed-point resistance, resonance	SWR at resonance	3:1-SWR bandwidth, % of resonant freq.
40-μH coils only	26 ohms	1.92:1	1.69
36.5-μH coils, 18" dangling ends	26	1.90:1	1.79
36" hats outside 32.5-μH coils	23	2.15:1	1.68
30-μH coils, 36" hats at ends	25	1.98:1	2.05
None (λ/2 dipole)	43.5	1.15	Greater than 3.6

Coil positions for each loaded antenna were 16 feet from antenna center. All antennas were constructed of no. 12 wire and installed at a height of 50 feet.

Feet X 0.3048 = m.

Inches X 25.4 = mm.

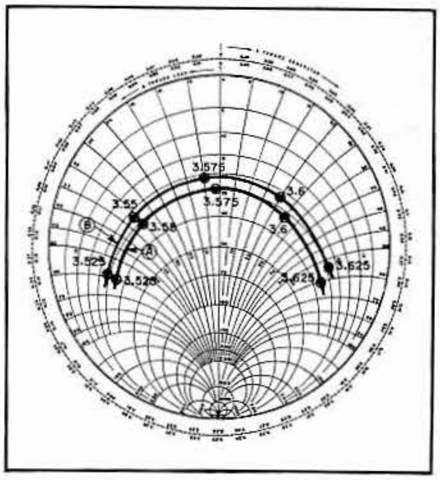

Fig. 13 — Curve A is the impedance plot of the antenna of Fig. 12A with 18-inch (457 mm) dangling end sections added and the coil trimmed to restore resonance near the original frequency. Curve B is a plot of the same antenna with X-shaped capacitance hats added at a point just outside the loading coils (dangling sections removed and coils trimmed to reestablish resonance.).

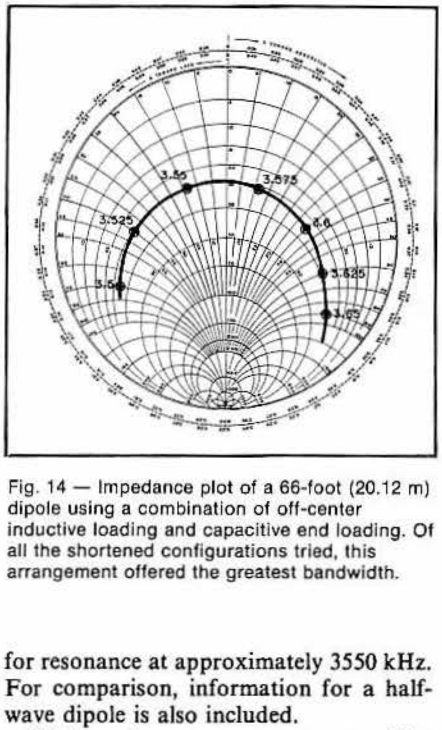

Fig. 14 — Impedance plot of a 66-foot (20.12 m) dipole using a combination of off-center inductive loading and capacitive end loading. Of all the shortened configurations tried, this arrangement offered the greatest bandwidth.

would seem to make little difference whether short sections of extra length are added inside the supporting insulators or are at the ends, suspended at right angles to the main antenna wire.

The inductors were reduced from 40 to 36.5 μH (25-turn coils replaced the original 27-turn coils), and resonance occurred at about 3575 kHz. At this frequency the resistance was 26 ohms and the SWR 1.90:1. The 3:1-SWR bandwidth, 64 kHz, is 1.79 percent of the frequency of resonance. The impedance plot for this arrangement is shown as Curve A in Fig. 13. The resistance at resonance for this antenna is identical to that with the coils alone, and the bandwidth is only 4 kHz greater, 64 kHz vs. 60. From these results, one would conclude that the main advantage offered by the "danglers" is a small saving of space over a flat-top antenna.

Capacitance Hats near Loading Coils

Next the dangling end sections were removed and a pair of capacitance hats was formed, each from two 36-inch (0.91 m) lengths of no. 12 solid wire. The two wires for a single hat were attached at their centers to the antenna wire at a point just outside one of the loading coils. The hat wires were then bent radially to form an X at right angles to the antenna wire, like four spokes of a wheel with the main antenna wire at the hub. The diameter of the X-shaped hat was thus 36 inches. The second hat was placed in a like manner just outside the second coil. Burndy connectors were used to affix the hat wires. The resonant frequency of this configuration with the original 40-μH loading coils was found to be 3290 kHz. The effect of adding the hats was about the same as that of extending the 17-foot (5.18 m) lengths to 19 feet (5.79 m).

When the inductors were replaced with 23-turn coils (32.5 μH), the antenna resonated at about 3.575 MHz, the resistance being 23 ohms. The SWR at resonance is 2.15:1, and the 3:1-SWR bandwidth for this configuration is 60 kHz, 1.68 percent of the resonant frequency. The impedance of this arrangement versus frequency is shown by Curve B of Fig. 13.

It is surprising to note that, by the standards of most amateurs, the characteristics of this antenna are not as good as those of the same length antenna with loading coils alone. The SWR at resonance for the antenna with combination capacitive and inductive loading is higher (2.15 vs. 1.92), and the 3:1-SWR bandwidths are the same, 60 kHz. Perhaps a significant factor here, though, is that the diameter of the capacitance hats used for these measurements was small, only 0.011 wavelength. Supporting much larger hats presents mechanical problems with wire antennas, however, as even these were a bit flimsy and would require reshaping after gusty weather.

Capacitance Hats at Antenna Ends

Finally, the X-shaped capacitance hats were moved to the outside ends of the antenna, just inside the end insulators. With the original 40-μH coils, resonance appeared at 3215 kHz. From calculations, it was as if the 17-foot (5.18 m) end sections were actually 21 feet (6.40 m) long. With 30-μH coils (22 turns) in place, the resonant frequency was 3560 kHz. At this frequency the resistance was 25 ohms and the SWR 1.98:1. The 3:1-SWR bandwidth is 73 kHz, or 2.05 percent of the resonant frequency. The impedance plot of this antenna is given in Fig. 14.

It is interesting to note that the position and shape of the plot for this antenna on Smith Chart coordinates is nearly identical to that for the same length antenna with loading coils only, the solid curve of Fig. 12B. For this antenna, however, the plot points for 25-kHz frequency increments appear closer together, which accounts for the increased bandwidth.

Conclusions

The measured characteristics of these various configurations of loading for the 80-meter antennas are tabulated in Table 4. Remember that the overall "flat-top" length of each antenna arrangement is 66 feet (20.12 m), and that the loading coils are always positioned 16 feet (4.88 m) each side of the center of the antenna, being pruned for resonance at approximately 3550 kHz. For comparison, information for a half-wave dipole is also included.

Of the various arrangements, capacitive end loading decidedly provides the greatest bandwidth, excepting the full-size half-wave antenna, of course. Although there are slight differences in the resistance value at resonance, all are of the same order of magnitude. These values, as well as those for the 160-meter antenna discussed earlier, tend to confirm a *broad* rule of thumb that the writer has formulated for this type of antenna: The feed-point impedance value at resonance is roughly proportional to the length of the antenna. That is, a loaded antenna which is half the size of a half-wave dipole will have approximately half the radiation resistance of the full-sized antenna.

Eq. 1 given earlier or the chart of Fig. 9 allows one to calculate loading-coil values for antennas with loading coils only. Additional capacitive loading is not taken into account. Calculating the effects of various capacitive loading arrangements appears to be difficult, and work remains to be done in this area.

Multiband Antennas with Loading Coils

All of the foregoing material has been devoted to the loading of an antenna for resonance at a single frequency. Resonated as described, the antenna is electrically a half wave in length. It will, however, operate well on higher frequencies — frequencies at which it is an odd multiple of half waves in *electrical* length . . . three half waves, five half waves, etc. Because of the lumped loading of the shortened antenna, these higher frequencies will likely not be closely related to odd-order harmonics of the fundamental frequency, as the case would be for a nonloaded radiator. (For example, it is a well-known fact that a 7-MHz half-wave dipole operates well on its third harmonic, 21 MHz.)

A loaded dipole will become an electrical 3/2-λ antenna at some frequency below that which is three times the fundamental resonant frequency. Depending upon the overall antenna length, coil value, and coil position, it is possible for an 80-meter loaded dipole to become a 3/2-λ performer on 40 meters. With such an arrangement, one would have a dual-band antenna without requiring the use of traps. The idea can be expanded upon to arrive at a loaded antenna without traps which will operate on more than two bands. This scheme offers considerable constructional simplification as compared with trap arrangements.

The multiband loading-coil concept has been recognized for better than half a century, but little use of the technique has been made by amateurs. Some years ago a very good article on the subject was published by William Lattin, W4JRW.[5] That article is recommended reading for anyone interested in more details on the concept. Supplemental information has been published by Buchanan.[6] Attempts by this writer to calculate antenna sizes and coil values for dual-band antennas have met with some success. From calculations and experiments to date, it appears that with only two loading coils (one each side of center), the antenna must always be greater than a half wave in physical length for the higher of the two frequency bands. In other words any 80/40 meter arrangement, for example, apparently would need to be longer than 66 feet (20.12 m) from tip to tip. However, much work also remains to be done in this area. This material was originally presented in *QST* by Jerry Hall, K1TD.

References

[1] Sevick, "The Ground-Image Vertical Antenna," *QST*, July 1971. (See chapter 1.) Also, "The W2FMI Ground-Mounted Short Vertical," *QST*, March 1973. (See chapter 1.) "A High Performance 20-, 40- and 80-Meter Vertical System," *QST*, December 1973. (See chapter 1.) "The Constant-Impedance Trap Vertical," *QST*, March 1974, p. 29.

[2] LaPlaca, "Using the ARRL L/C/F Calculator," *QST*, December 1973, p. 26.

[3] Hall and Kaufmann, "The Macromatcher, An Rf Impedance Bridge for Coax Lines," *QST*, January 1972, p. 14. Information also contained in *The Radio Amateur's Handbook* and *The ARRL Antenna Book*.

[4] McCoy, "Monimatch, Mark II," *QST*, February 1957, p. 38. Also, "An Etched-Circuit Monimatch For Checking Your Antenna System," *QST*, October 1969. (See chapter 4.)

[5] Lattin, "Multiband Antennas Using Loading Coils," *QST*, April 1961, p. 43.

[6] Buchanan, "An Inexpensive 40- and 80-Meter Antenna, Hints and Kinks," *QST*, September 1962, p. 62.

Efficient Short Radiators

If you're cramped for space to put up a beam antenna, you'll want to investigate these ideas as one way to shrink 'er down to usable size.

Shortened antennas have been popular with amateurs for either esthetic, space limitation or economic reasons. And most are inefficient radiators with narrow bandwidth. This section will first define some limitations and problems of conventinal short antenna designs. A design for a half-size 2-element Yagi, the performance of which is competitive with its full-size equivalent, will follow.

It has been proved that no appreciable difference exists between the directive patterns of "short" and full half-wave horizontal dipoles. Since the gain of a half-wave dipole over an isotropic radiator is determined by the dipole directive pattern and since the length of the dipole (up to λ/2) has negligible effect on its directive pattern, the length of the dipole (up to λ/2) has a negligible effect on its gain. An assumption is made that the efficiencies of both the short and half-wave dipoles are equal. The efficiency of an antenna can be expressed as

$$\mathrm{Eff} = \frac{R_r}{R_r + R_{loss}} \qquad \text{Eq. 1}$$

where R_r = radiation resistance

R_{loss} = loss resistance of conductors, coils, etc.

The radiation resistance (R_r) of a horizontal dipole is determined by its length and height above ground. In free space, the R_r of a half-wave dipole is 73 ohms. As the length of the dipole is reduced, the R_r decreases as the square of the length. Thus, for a half-size dipole (as compared with λ/2), the R_r decreases to a value one-fourth that of the λ/2 dipole or approximately 13 ohms.

The efficiency of a half-wave dipole is very high, being on the order of 95 percent. The R_r is large and R_{loss} small by comparison. If we reduce the overall length by a factor of two, the R_r decreases by a factor of four. Input impedance at the center of the dipole is comprised of a low resistance and large capacitive reactance, so inductive loading is required in order to resonate the dipole. The inductive reactance needed to center-load the antenna will be approximately 1 kΩ (depending on conductor size used for the antenna). Assuming a coil Q of 200 and computing efficiency using Eq 1: This means that 28 percent of the transmitting power applied to the antenna is dissipated in the loading coil. The author has neglected any additional losses in the matching network which would result if the above antenna were driven by a 50-ohm source.

Let's suppose we decide to build a 2-element Yagi using shortened center-loaded elements. The driven element by itself has an R_r of 13 ohms. Adding a close-

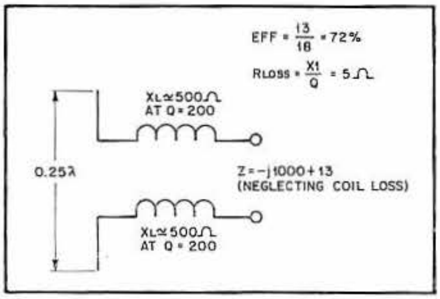

Fig. 15 — Electrical equivalent for the center-loaded dipole discussed in the text. Efficiency is 72 percent, as determined by Eq. 1.

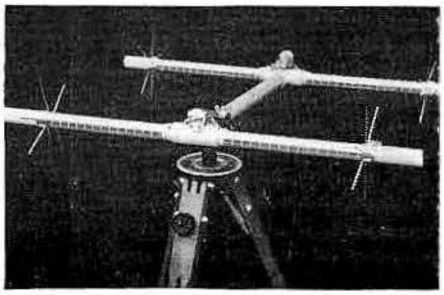

The 100-MHz "model" used by the author for testing shortened elements in a Yagi antenna design.

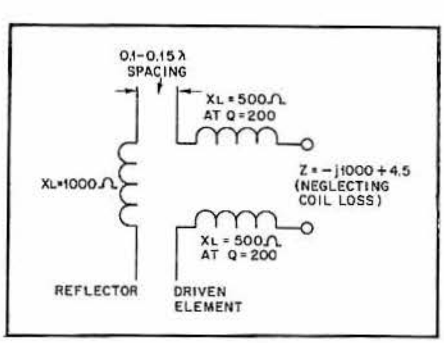

Fig. 16 — Electrical equivalent of a loaded 2-element antenna. The efficiency is 47 percent.

spaced parasitic element reduces the radiation resistance by approximately a factor of three. Thus, the driven element has an impedance composed of approximately 4.5 ohms resistive and 1 kΩ reactive. Loading both the driven and parasitic elements with the same inductors that were used for the dipole and computing efficiency

$$\text{Eff} = \frac{4.5}{4.5 + 5} = 47 \text{ percent}$$

This efficiency equates to slightly more than a 3-dB power loss in the driven element. The writer has neglected the sizable loss in the matching network since it must transform the 50-ohm source to the 9.5-ohm impedance of the antenna at resonance. In addition, the loading coil of the parasitic element will contribute approximately 1 dB of loss.

Taking into account the losses on loading coils of both the driven and parasitic elements and the loss on the driven element to feed-line matching network, we can easily account for a 5-dB power loss. Such a power loss would nullify the gain of a properly tuned Yagi, making it little better than a full-size dipole.

It should be obvious from the previous examples that the poor performance of short dipoles, and particularly Yagis, is a direct result of losses in loading inductors. If we can reduce these losses while raising the radiation resistance, the increase in efficiency will (might?) make the short dipole or Yagi competitive with its full-size counterpart.

Test Antenna

The helically wound whip antenna has always been appealing because the R_r tends to be higher than that of the base-loaded whip. This is because it is continuously loaded. In fact, using end-loading (large capacitance hats), the current distribution will be much more constant, yielding an R_r of up to four times that of a base-loaded whip. The author decided to use the helical design for a short test dipole and investigate the properties of this antenna versus that of a full-size dipole.

Conventional design for a helically wound antenna uses a form (wood, bamboo, fiberglass) wound with approximately $\lambda/2$ of no. 14 wire per $\lambda/4$ element. The writer felt that the resistive R_{loss} of this large amount of wire would be as great as that of the loading coils in the center-loaded dipole, making the helical design ineffective. As it is desirable to reduce conductor resistance, the surface area of the conductor must be substantially increased. Tubing is not mechanically suitable, and because skin depth is only on the order of 0.001 inch (0.03 mm),[1] the use of tubing is not required. Some 1/2-inch (13 mm) wide copper tape[2] was on hand and a study of its skin resistance revealed a loss per unit length of 12 percent less than that of the no. 14 wire.

A helically wound dipole was constructed (using the 1/2-inch [13 mm] wide tape) for use at 100 MHz. This frequency was chosen so that field-strength measurements could be performed using a local fm broadcast station as the signal source. A 28-inch (711 mm) length of 3/4-inch (19 mm) CPVC tubing (plastic water pipe) was wound with 38 equally spaced turns of the 1/2-inch (13 mm) wide tape. Using a grid-dip meter, resonance was measured at 104 MHz. Next, two 5-1/2-inch (140 mm) diameter six-spoke capacitance hats were attached to the ends. Resonance was again measured and found to be 84 MHz. A C-match[3] was constructed to transform the low impedance of the short dipole to 50 ohms. The C-match requires that the element be made to look inductive by lowering its resonant frequency. A variable capacitor is then shunted across the feed point and adjusted for resonance. This L/C ratio determines the impedance transformation. A 100-pF variable capacitor was shunted across the feed point of the test dipole and a 50-ohm coaxial cable was attached. Since the dipole is a balanced antenna, feed-line decoupling was necessary. The author used a quarter-wave sleeve (bazooka).

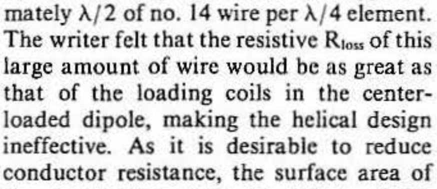

Fig. 17 — Curves depicting gain and FBR (front-to-back ratio) vs. frequency when the Yagi antenna is tuned for operation at 100 MHz.

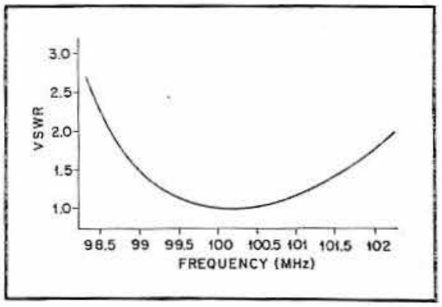

Fig. 18 — VSWR vs. frequency when the Yagi antenna is tuned for operation at 100 MHz.

Antenna Tuning

Rf power was applied to the test antenna through an SWR meter (measurements were taken in a screen room). Turns were removed, one at a time, and the shunt capacitor adjusted until a 1:1 match occurred. The capacitance hats were always installed at the tips of the element. When adjusted for a 1:1 match, the dipole consisted of 32 turns of copper tape, center fed, with 40 pF shunting the feed point and an overall element length of 21 inches (533 mm) which is 37 percent of the length of a full-size dipole. Tuning was very simple and took but 15 minutes to complete.

Measurements

The feed-point impedance was measured (with C-match disconnected) to be 20 ohms. The skin depth resistance was computed to be 0.24 ohm. Computing efficiency

$$\text{Eff} = \frac{20}{20 + 0.24} = 98 \text{ percent}$$

which is essentially the same as a full-size dipole. Bandwidth between 2:1 SWR points was measured to be greater than 6 percent of the operating frequency. Large capacity hats (for 100 MHz) and wide tape play an equally important role in reducing the Q of the antenna sufficiently to obtain a wide bandwidth. Next, a low-level signal generator and dipole were set up 5 wavelengths away from the test antenna. Using an Empire NF-105 noise and field-strength meter, a comparison was made on field strength between the test antenna and a full-size dipole. Results indicated no apparent measurable difference between the short helically wound dipole and a full-size dipole. Field-strength measurements were then taken on the signal from a local fm broadcast station. Again, there was no apparent difference; not bad for an antenna that is less than half size. The next question to be answered was whether gain could be secured by the addition of a parasitic element.

A boom and a parasitic element were added to the shortened test dipole. Although a director might provide better gain and front-to-back ratio (FBR), the decrease in radiation resistance, to a very low value (dipole by itself equals 20 ohms) because of the close spacing, was thought to decrease efficiency and bandwidth. A parasitic reflector was constructed, similar to that of the driven element. The element was split in the center and a 100-pF capacitor installed for tuning purposes. Element spacing was adjusted to 0.15 λ. With the capacitor fully meshed, the six-spoke capacitance hats were physically positioned on the element (both equidistant from element center) so that the reflector resonated at 93 MHz. Nominal reflector tuning calls for reflector resonance approximately 5-percent lower in frequency than the driven element, or 95 MHz. With the above configuration, the author was able to adjust reflector resonance, using the variable capacitor, anywhere in the range of 93 to 96.5 MHz or 3.5 to 7 percent lower in frequency than the driven element.

Using the same field-strength measurement set up as with the dipole, the reflector was aimed at the source signal and adjusted (using the variable capacitor at element center) for minimum pickup. The driven element was then readjusted for a 1:1 SWR and the process repeated until minimum rear pickup and a 1:1 SWR occurred simultaneously. Gain and FBR measurements were made at the design frequency and at points two percent of the design frequency on either side. The results are described in Fig. 17. The bandwidth measurements are given in Fig. 18.

As can be seen from the curves in Fig. 17, a gain of 4 dBd, and a 15-dB FBR can be secured quite easily with a physically short Yagi. As with other Yagis, the gain is relatively constant over a wide frequency range. However, the FBR drops quite fast especially when trying to use the antenna below its design frequency. From curves given in Fig. 18, the bandwidth between the 2:1 VSWR points is shown to be about 3.7 percent. Translated to 20 meters this bandwidth would be 500 kHz, equal to or greater than the average 3-band trap Yagi. If indeed the design were scaled to 14 MHz, the antenna would have element lengths of slightly over 12 feet (3.66 m) and a boom length of 10 feet, 6 inches (3.20 m).

Adapting Design to Ham Bands

Since it has been shown that a physically short Yagi can provide significant gain and FBR, a few design pointers are in order. First, make the element length no shorter than physically necessary for your particular situation. Bandwidth and efficiency will improve with greater lengths. Second, the use of large capacitance hats is recommended, as this reduces the helical conductor length and thus R_{loss}. Third, use large forms of good dielectric quality (fiberglass). The larger the form diameter, the greater the length reduction for a given number of turns. Also, with large diameter forms, 1.5 inch (38 mm), the width of the conductor can be increased, thus reducing R_{loss}. Fourth, regardless of the matching network used, construction of capacitance hats, etc., *solder all* joints and *seal* with silicone rubber. *Do not rely on pressure joints.* Fifth, to secure a reasonable FBR with any Yagi using a split driven element and fed with coaxial cable, some form of balun transformer is required to keep currents from flowing on the outside of the coax braid. Sixth, the use of a grid-dip meter is the only practical method to secure resonance with a fixed physical length. Construction will require some cut and try. Proximity to surrounding objects will affect resonance, and should be avoided when tuning the array.

No doubt the author has left out items, which may prove to be a problem for the reader. This article was intended to be more thought provoking than constructional, although there should be enough information here to build your own antenna with a reasonable amount of experimental effort. My current plans call for construction of a 2-element 10-meter Yagi (for reception of OSCAR signals) with 6-foot (1.83 m) elements and a 5-foot (1.52 m) boom. The author wishes to thank Mike Povlich, WB9HGS and Russ Mills for their assistance with measurements. This material was originally presented in *QST* by Ronald J. Gorski, N9AU.

References

[1] Kuecken, *Antennas and Transmission Lines*, Howard W. Sams and Company, Indianapolis, IN, 1st Edition (1969), p. 274.

[2] Copper tape available from Minnesota Mining and Mfg. Check Yellow Pages for the address of a local distributor.

[3] Orr, *Beam Antenna Handbook*, Radio Publications Inc., Wilton, CT, 3rd Edition, p. 63.

Simple Arrays of Vertical Antenna Elements

Work DX with a vengeance! Simple arrays of vertical elements will provide good low-angle radiation. DXCC may be closer than you think if you can manage a system like this one!

Because of the excellent low-angle radiation properties of vertical antenna elements and because of the benefits to be gained by horizontal-plane directivity, it is interesting to attempt to determine "optimum" arrangments of elements to produce "beam" antenna arrays. Because of the enormous variety of configurations possible, and the great difficulty of either experimental modeling or calculating exact theoretical performance, it has not been practical in the past to arrive at valid conclusions. However, it is now quite easy to simulate such antenna arrays by computer programs, and these can produce relatively quick and precise answers to a very large number of postulated configurations. After a number of computer trial runs, "optimum" configurations can be found quite readily. Such a computer program for quarter-wave elements has been written and supplied by H. Hurwitz, WA2VBW. Using a modification of this program, the author has investigated a

Table 5

Case	Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	1	0	0	0	0°	3 dB	-3 dB	-α
	2	1	-90	0.25	0				
2	1	1	0	0	0	0°	4.1 dB	5.7 dB	-11.5
	2	1	-120	0.25	0				
3	1	1	0	0	0	90°	1.1 dB	0 dB	0 dB
	2	1	0	0.25	0				

Table 6

I1	I2	$\phi 1$	$\phi 2$	Beam Gain	Back Radiation	Minor Lobes
1	0.8	0	-120	4.0 dB	-5.5 dB	-10.8 dB
1	1	0	-120	4.1	-5.7	-11.5
1	1.2	0	-120	4.1	-5.6	-12.5
1	1	0	-100	3.4	-3.8	-11.1
1	1	0	-120	4.1	-5.7	-11.5
1	1	0	-140	4.6	-6.6	-6.6

Table 7

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	0	0	0				
2	1	-120	0.25	0	0°	5.6 dB	-8.7 dB	-8.7 dB
3	1	-240	0.5	0				

Table 8

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	0	0	0				
2	1	0	0.25	0	90°	2.7 dB	0 dB	0 dB
3	1	0	0.5	0				

Table 9

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	0.5	0	0	0.145				
2	1	-110	0.25	0	0°	4.35 dB	-6.1 dB	-15. dB
3	0.5	0	0	0.145				

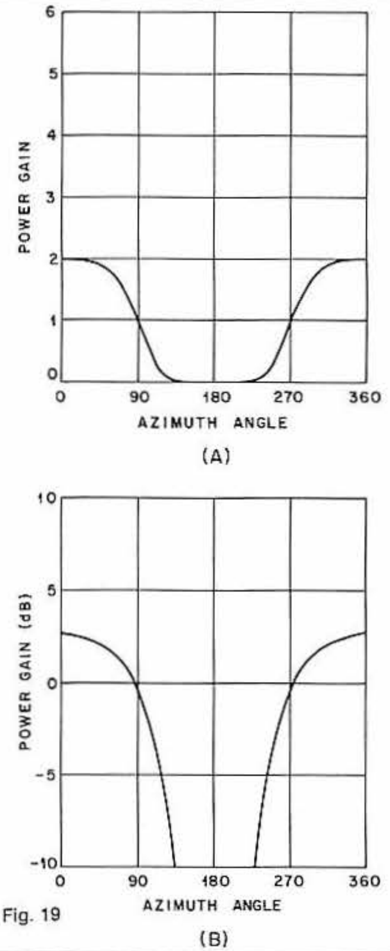

(A)

Fig. 19

(B)

large number of simple array configurations, which are potentially quite useful at the lower amateur frequencies, with the salient results given here. It is hoped that these will prove interesting to those considering construction of such arrays.

Technique for Optimizing Arrays

We must first try to define the problem and the parameters leading to an optimum design. The horizontal-plane pattern of an array of (vertical) elements depends only on the magnitudes of the individual element currents and their phases, and not on the way the currents are produced (e.g., driven or parasitic). The computer program allows one to specify the locations (x,y coordinates) of any number, n, of quarter-wave elements; also one must specify the magnitude of current and its phase in each element (usually normalized to unity current and zero phase for a given fiducial element). With these data inputs the radiation pattern is calculated and plotted out, giving the relative power gain in each azimuth angle interval, say, 5 or 10 degrees. The power gain is normalized to that power which would have been radiated at the same azimuth by a *single* vertical element using the same input electrical drive power. By using trial values of the current in all elements, in both amplitude and phase, one soon develops a feel for the behavior of the system, and can usually arrive at a "best" value for all the currents and phases. One can then try different geometrical separations of the elements and in this way arrive at the "best" spacing. Usually this best configuration is a compromise between power gain and good pattern control, e.g., front-to-back ratio, or more generally, main lobe to minor lobe (side or back) ratio.

In addition to the optimization described above it must be noted that if the array is to be used primarily to produce a "beam" in a given general direction with little emphasis on other directions, one usually finds a linear array of elements best (equivalent to an end-fire Yagi array). However, if the array is to be used for all directions, a more (circularly) symmetrical array should be considered. These points will become clear in the discussion of results.

Two-Element Array

In the case of two elements there is only one possible configuration. Consider one element at the origin of the x,y plane (coordinates 0,0) and the other element along the x axis (coordinate x,0). The parameters available are x, which we specify in units of wavelength, λ, the rf currents I1 and I2 in each element (we shall normalize one of the currents, e.g., I1, to unity) and the phases, $\phi 1$ and $\phi 2$, of the currents (we shall generally normalize the phase of the current in element 1 as zero and specify the relative phase of the current in element 2 in electrical degrees, either positive for leading currents or negative for lagging currents). It soon becomes apparent that the region of main interest is where the spacing x is in the neighborhood of 0.25, i.e. $\lambda/4$, and there the currents are nearly equal in magnitude. The reasons are that, for spacings of 0.5 or greater, other strong lobes appear in the pattern, and for grossly unequal currents relatively poor

Table 10

Element Spacing	Optimum $\phi 2$	Beam Gain	Back Radiation	Minor Lobes
0.2	-130°	4.65 dB	-6.5 dB	-18 dB
0.25	-120°	4.6	-6.5	-16
0.3	-110°	4.4	-6.2	-14
0.35	-110°	4.2	-5.8	-12

Table 11

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	0	0	0	0°	7.4 dB	-9.9 dB	-9.9 dB
2	1	-120	0.25	0				
3	1	-240	0.5	0				
4	1	0	0.75	0				

Table 12

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	0	0	0	90°	4.2 dB	0 dB	0 dB
2	1	0	0.25	0				
3	1	0	0.5	0				
4	1	0	0.75	0				

Table 13

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	0	0	0	45°	6.4 dB	-18 dB	-18 dB
2	1	-110	0.25	0				
3	1	-220	0.25	0.25				
4	1	-110	0	0.25				

Table 14

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	0	0	0	0°	5.1 dB	-11.5 dB	-11.5 dB
2	1	-120	0.25	0				
3	1	-120	0.25	0.25				
4	1	0	0	0.25				

Table 15

Square side Length (inλ)	ϕA opt.	ϕB opt.	Beam Gain	Back Radiation	Minor Lobes
0.167	-130	-260	6.6 dB	-22 dB	-22 dB
0.25	-110	-220	6.4	-18	-18
0.333	-80	-160	5.6	-18	-18

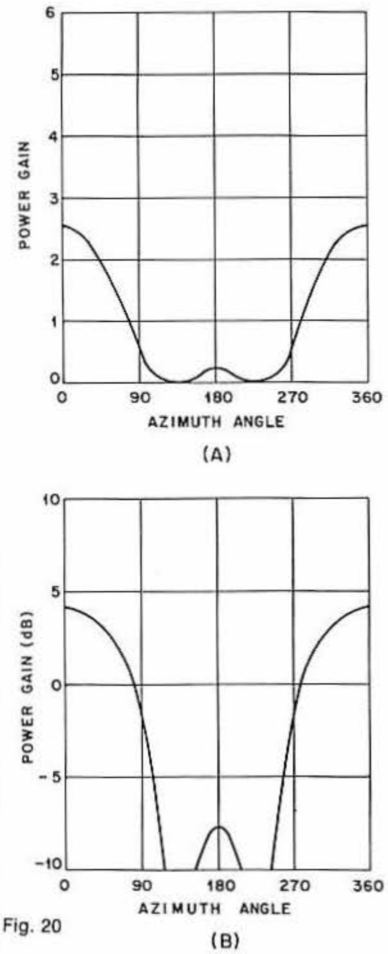

Fig. 20

peak gain and pattern discrimination result. Although the spacing x is not especially critical in the neighborhood of 0.25 (one can essentially compensate by adjustment of phase angle) we shall now examine the case where x = 0.25, and where I1 = I2 = 1. Best overall pattern discrimination (minimum back lobe) for the endfire case occurs where $\phi 1 = 0$, and $\phi 2 = 90$ degrees, and the power-gain pattern is shown in Fig. 19A.

Peak gain is 2 (or 3 dB) and occurs along the x-axis direction. Fig. 19A shows the power pattern of the array, i.e., it shows quantitatively where the radiated power goes. However, if one is looking at what an S meter would indicate, one can present exactly the same information on a vertical *logarithmic* scale as shown in Fig. 19B.

The vertical scale is shown in dB relative to the value for a single vertical radiator. This type of display is useful in judging excellence of pattern with respect to minor lobes. For all future cases discussed here, both displays will be shown.

It is possible to increase the gain of this array at the expense of some back lobe by increasing the phase lag in element 2, e.g., if $\phi 2 = -120$ degrees the peak gain, again along the x axis, is 2.59 (or 4.1 dB) and the entire pattern is shown again in Figs. 20A and B.

Note that the back lobe now shows a gain of 0.186, which is −11.5 dB relative to the forward lobe.

These two elements can be used to form either a beam along the +x axis as shown, or reversed along the −x axis (by reversing the phase $\phi 2$, i.e., making it +120 instead of −120 degrees) or if the two elements are driven in the same phase, i.e., $\phi 1 = \phi 2 = 0$, one produces a bilateral broadside beam as shown in Fig. 21.

Here the peak gain is 1.1 dB, but there is a "back" lobe of equal magnitude to the front lobe. Note that all patterns will be symmetrical around the x axis, or more generally around the line of any set of linearly-positioned elements. This example has illustrated some of the aspects of all of the results to be presented; namely, that:

a) The best element spacing will be found to be in the neighborhood of 0.25 wavelength.
b) Current ratios and phases can be adjusted to produce either a best *pattern discrimination* (minor lobe rejection) or best *main lobe gain.* It is generally not possible to get both at the same time, and a compromise must be made.

It may be useful to tabulate the salient results of each "best" case by noting the coordinates of the elements in the x,y plane, the individual drive currents, the individual phases, main beam heading in the x,y plane, main beam gain, and the

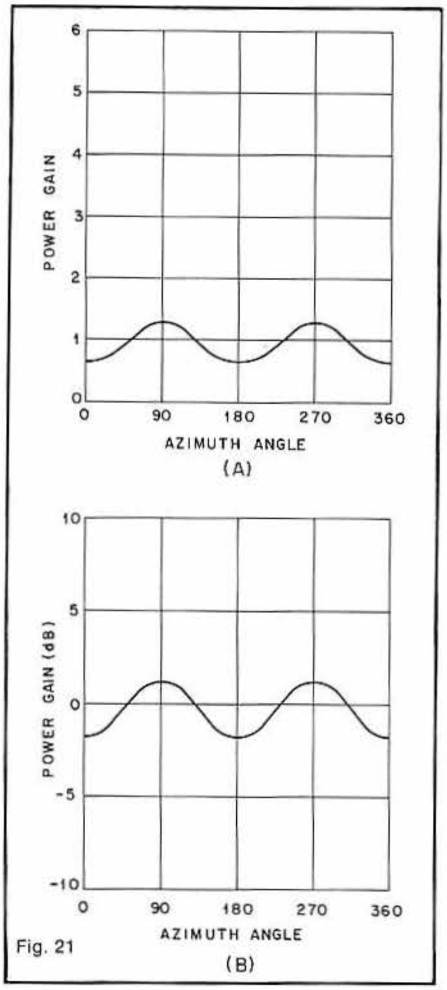

Fig. 21

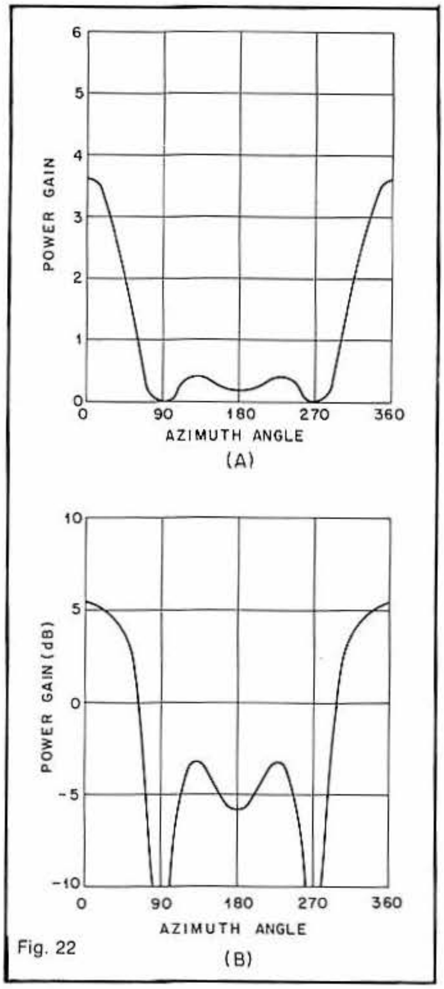

Fig. 22

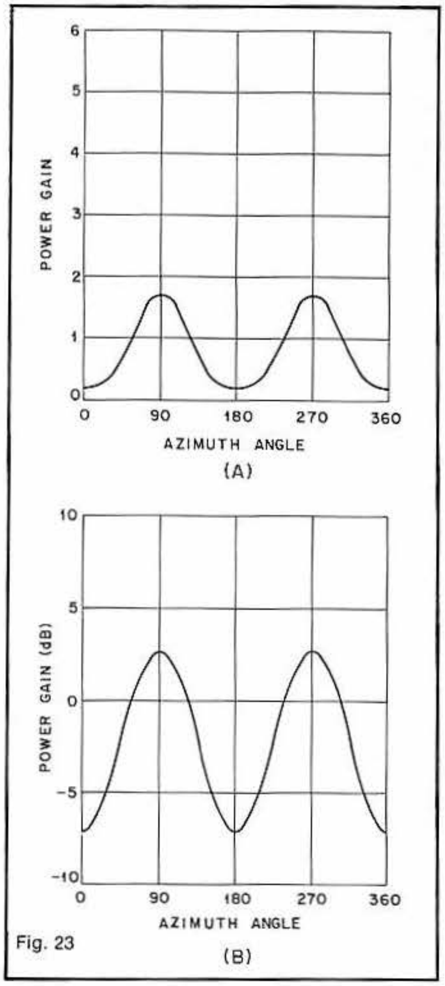

Fig. 23

largest pattern gain (such as given by either the side of the main lobe or by minor lobes) observed anywhere in the entire 180-degree sector centered just opposite in direction to the main lobe. We shall designate this quantity as "back radiation" and express it in dB relative to the main lobe gain. We shall also tabulate the largest minor lobe in this sector and label it "minor lobe," again expressed in dB relative to the main lobe. Thus for the cases just described, Table 5 is presented.

The sensitivity of these results to variations of currents and phases is not high. For example, one can summarize case 2 above with somewhat altered I2, or $\phi 2$, as shown in Table 6.

The author feels that end-fire case 2, where $\phi 2 = -120$ degrees, probably represents the "best" design with a good pattern and gain. Other more complicated cases presented below are optimized in this same way and only the "best" configuration will be shown with actual results.

Three-Element Arrays

For three elements, two different geometries suggest themselves: the linear array, best for a single preferred directional line, and an equilateral triangle configuration, the closest approach to circular symmetry. The latter, as we shall see, could be switched to give "beams" in any one of six angular positions (every 60 degrees). Let us consider first the linear array.

Linear Array — As in the 2-element linear array just discussed, the optimum arrangement seems to center around an element spacing of about 0.25 wavelength with approximately equal current drives. The optimum end-fire case is shown in Table 7; the broadside case is shown in Table 8.

The patterns for these cases are shown in Figs. 22 and 23.

Remember that for the end-fire case, which one can think of as a 3-element Yagi on its side, the rather large side lobe results because the pattern lies in a plane perpendicular to the elements. Side radiation is hard to avoid, as one cannot take advantage of the nulls off of the ends of excited elements.

Equilateral Triangular Array — For this case, the optimum situation appears to be as shown in Table 9. The behavior is very similar to the case of two elements in end-fire. See Table 5 (we have just split one of the two elements and have moved the separate pieces to the triangle corners). The pattern is shown in Fig. 24.

Reversing the phases will reverse the beam direction, and by rotating the drive currents (and phases) around the triangle, one can produce six beam directions (every 60 degrees). This is an attractive feature, but it does not come free; the maximum gain of the system is somewhat less than one obtains with the three linearly-disposed elements (which are chiefly useful in a direction along the line of the array).

The exact size of the triangular array is not critical — one must simply choose the best phase angle for the particular geometry. The "best" phase angles, $\phi 2(\phi 1 = \phi 3 = 0)$, as determined by computer runs, for different spacings (triangle side lengths expressed in terms of λ) are shown in Table 10.

Four-Element Arrays

With four elements, there are at least three configurations which are interesting to investigate. As before, the linear array with a preferred beam direction is one type, but there are two configurations which possess a measure of circular symmetry. One of these is described by three antenna elements at the corners of an equilateral triangle with the fourth element at the geometric center of the triangle. This one we shall designate as a center-filled triangle. The other type is a square array.

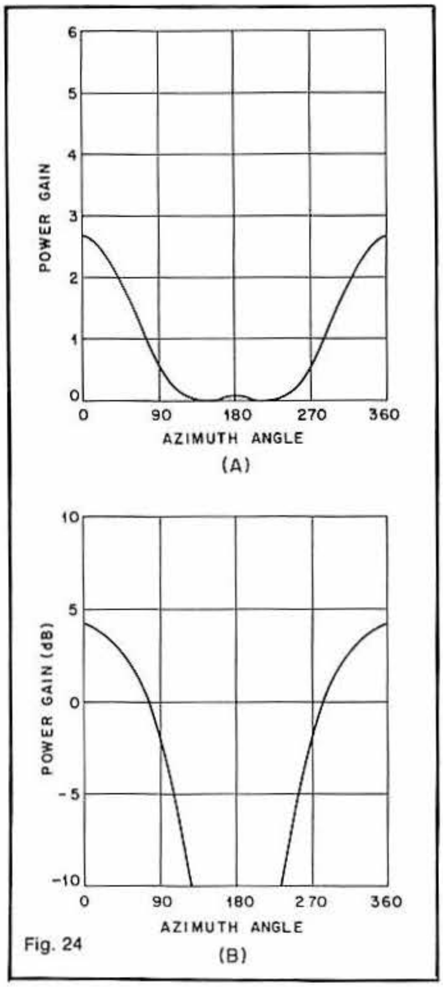

Fig. 24

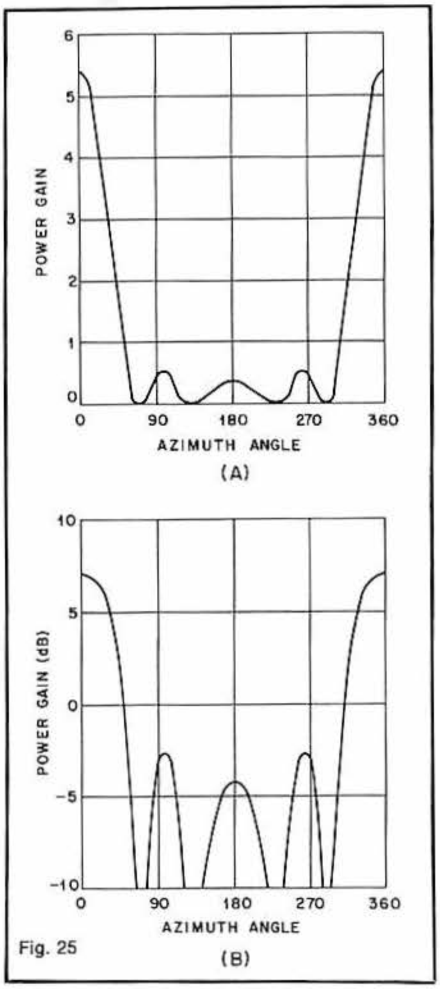

Fig. 25

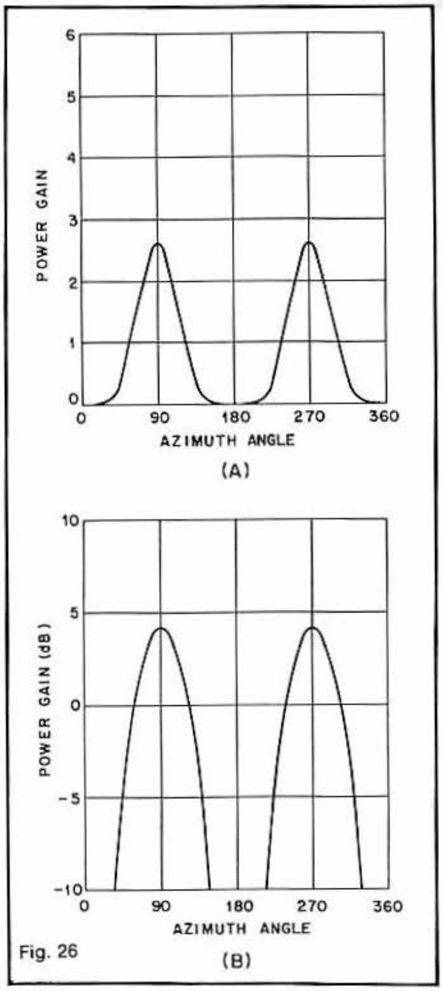

Fig. 26

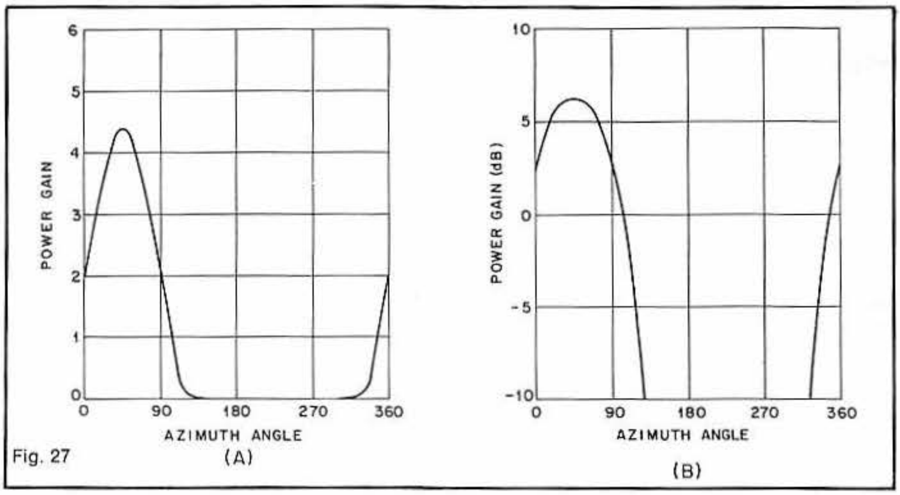

Fig. 27

Let us take these three types in order.

Four-Element Linear Array — For this case the optimum end-fire solution appears to be as shown in Table 11. The gain and major lobe shape of this configuration seems to be excellent. The broadside case is shown in Table 12. Patterns for this case are shown in Figs. 25A, B, and 26A, B.

Note that for this array there is poor coverage in some azimuthal directions (e.g., ±45 degrees) if only the two tabulated phase arrangements are used.

Center-Filled Triangle — A large number of situations were examined for this configuration, but unfortunately no combinations of spacing, currents or phases were found that looked very good. The best pattern was not particularly better than that of the three-element triangle alone, and was substantially inferior to that of the 4-element square to be discussed next.

Four-Element Square Array — For this case the best situation appears to be as shown in Table 13. Note that this arrangement fires diagonally and produces a respectable gain and excellent pattern. The "broadside" arrangement for the same square is shown in Table 14. The computed patterns are shown in Figs. 27 and 28.

The diagonal-fire pattern is exceptional, providing extremely low back radiation and minor lobes. The broadside pattern is not nearly as good, somewhat superior but reminiscent of the end-fire pattern obtainable with just 2 elements (see Fig. 20). In fact, the power-gain coverage at all angles (including the broadside angle) just using the switched four diagonal beams is large enough that it probably is not worthwhile using the broadside arrangement at all even to "fill" between diagonal beams.

This square array is superior to the center-filled triangle, and to all other simpler configurations (circular) in both gain and pattern discrimination. It is sufficiently good that one might ask what particular properties of the square have led to the nearly ideal performance. The

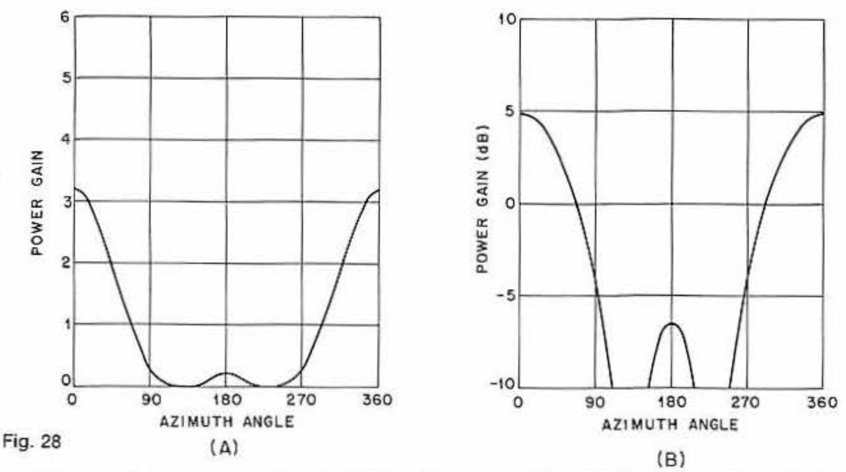

Fig. 28

author believes that there are two parameters which are noteworthy: first, the gain is respectable in diagonal fire because one can view the square along its diagonal as a three-element linear array in which the center element is split, each half being displaced to the other corners of the square. Thus the array behaves something like a 3-element (Yagi-like) linear array which is known to have a good gain. Secondly, the lobes are much smaller than the linear array due to the tendency for the phase cancellation of waves produced by the spatially-separated corner elements.

Again, as in previous examples, the actual dimensions of the square are not very critical, as long as one adjusts the phase (delay), ϕA in elements 2 and 4 and the larger phase delay, ϕB, in element 3 to the "best" value. Examples of such "best" phase values for different size squares are shown in the Table 15.

It turns out that the calculated gain figures for the smaller arrays are actually slightly larger than the gain for the larger squares, but as a practical matter this fact is probably offset by the larger (reactive) currents required for radiation. The larger currents are necessary because of the mutual coupling and phase relations between elements and the required phase relationships; in practice with the inefficiencies usually caused by ground currents when using vertical antennas, the gain differences for the various squares are probably inconsequential. The patterns for all cases are quite good, but the best pattern is obtained for the $\lambda/4$ square. Incidentally, the 0.167λ and 0.33λ square allows the possibility that a single square can be used for two amateur bands, and this has indeed become the subject of a practical antenna system.[1]

Conclusion

Computer trial runs, through a program developed by WA2VBW, have facilitated the investigation of a large number of $\lambda/4$ vertical antenna-array configurations. "Best" configurations, drive conditions, power gains and patterns have been found for the cases of two, three and four elements. Especially interesting is the case of four elements arranged in a square whose side dimension is about one-quarter wavelength.

Cases using more than 4 elements have been investigated, but the complexity, probable cost, and difficulty of installation of such configurations, reduces their utility and so they will not be reported here.

The author wishes to acknowledge the contribution of Henry Hurwitz, WA2VBW, who supplied the original computer program for this investigation. This material was originally presented in *QST* by James L. Lawson, W2PV.

References

[1]Lawson, W2PV, "75/80 Meter Vertical Antenna Square Array," *QST*, March 1971.

360-Degree Steerable Vertical Phased Arrays

Microwave power-splitting and phasing techniques were applied to the design of this 80-meter array. It has established an outstanding record for performance on long and difficult paths.

The possibility of providing effective 360-degree horizontal coverage with an array switched electrically, rather than rotated mechanically, has long intrigued many designers. Such a solution to the directional antenna problem is of particular interest to amateurs operating on the lower frequencies, where any mechanical rotation system is difficult and expensive to build. One of the authors, now W1CF, writing under his old call, W1HKK, described an 80-meter switchable phased array of four quarter-wave in-line elements more than 10 years ago.[1] The present article describes several novel array designs, using up-to-date technology, which provide greatly improved operation. Also covered are techniques used for effective power division, the beam-forming phasing system, switching, and impedance matching. Ground radials, cable and connector selection, and dc continuity testing are discussed.

One of these arrays recently constructed for 80-meter use has a single 97-degree lobe with a gain over a single vertical of approximately 7 dB at low angles. Its front-to-back ratio is 25 dB, and its front-to-side ratio is more that 12 dB, over the whole 80-meter band. It is switchable over 360 degrees in four 90-degree steps.

Computer-Modeled Arrays and Patterns

In the spring of 1975, while sitting over a cup of coffee, the three authors discussed the possibility of rebuilding the 1965 4-

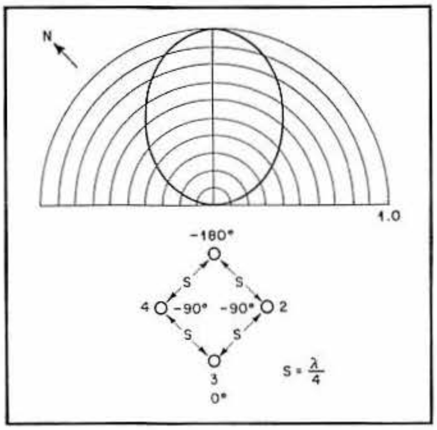

Fig. 29 — Polar plot of relative power, and planar view of the four-element diamond array showing the pattern obtained with no dc voltage on the switching relays, as in Fig. 31. Minor lobes are too far down to show on this scale.

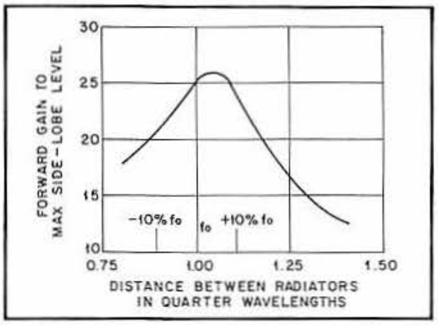

Fig. 30 — Bandwidth and side-lobe study of the four-element diamond array.

element array, using many improvements in technology that have evolved in the last 10 years. Dr. Joseph White accelerated the effort by coming up with a computer program written in Fortran IV, called ARRAY, which allows the user to plot patterns of arrays of from 2 to 99 elements in various spacings and combinations of phase and power, without ever leaving the computer terminal. Over 30 iterations using 4 elements were explored, and the computer-printed polar plots and other gain data analyzed, before we arrived at a configuration which provided good forward gain over a broad beamwidth, with no measurable unwanted lobes.

The configuration selected has four quarter-wavelength vertical elements in a square, with quarter-wave spacing between adjacent elements, as shown, with its predicted pattern, in Fig. 29. All elements are fed with equal amplitudes, the rear element at 0°, the two side elements at -90° and the lead element at -180°. The beam is transmitted along the diagonal from the rear to lead element. Gain due to horizontal beam formation alone is about 5.3 dB over a single vertical element. Front-to-back ratio is 25 dB. Front-to-side ratio is 12 dB, at 90° either side, increasing to much higher levels at 135° either side. Since most of the vertical energy is concentrated at low angles by the array, as much as 4 dB of gain additional to the predicted 5.3 dB in the horizontal plane can be achieved in theory, with perfectly conducting ground. The authors estimate that a good radial system and less-than-perfect ground yield an additional gain of 2 dB, or a total gain just over 7 dB for the system described.

The computer predicts a half-power beamwidth of 97°, so a suitable switching matrix can be used to direct the beam to four different quadrants, with only a slight loss of forward gain at the cross-over points, and virtually no deterioration of the front-to-back and front-to-side suppression. A computer study of the effect of variations in element spacing on forward-gain to maximum side-lobe level, plotted in Fig. 30, led to selection of $\lambda/4$ spacing of the corners of the square array. The 80-meter beam erected according to the computer-synthesized array design has proven to be remarkably successful. On-the-air results clearly approximate the performance anticipated by the program.

RF Power Dividers

Good power splitters are essential to the operation of phased arrays. Two-way and three-way power dividers will be discussed. Amateurs in broadcasting are well aware that tapped coils are the mainstay for power division in their industry. This reactive technique presents many problems.[2] The authors, being more versed in microwave techniques, thought it logical to scale up in wavelength the methods of power division used most successfully in the microwave region, particularly what is now known as the Wilkinson Power Divider,[3] shown in Fig. 31. Power from the transmitter is fed through a 50-ohm line of any length to a coaxial T, feeding two quarter-wavelength 70-ohm lines, W1 and W2. The two inner conductors of the 70-ohm lines are connected through a 100-ohm noninductive resistor, R1. This type of divider gives an equal power split, matches to 50-ohm loads at the two outputs, and has the unique property that any energy returning to the two outputs out of phase, due to mismatches or mutuals, is absorbed in the 100-ohm resistor.

The 90-degree phase-delay cable (DL) used in one side of each of the three power-splitting hybrids serves to assure that equal power reflections from the antennas are absorbed. Theoretically the resistors absorb none of the forward power. This technique provides approximately 30 dB of isolation from one output terminal to the other, to unwanted energy.

Reference is made in the 1965 article to difficulty in adjusting a phased array, because of element interaction due to mutual impedances between elements. The Wilkinson Power Divider, when used to feed phased arrays, reduces these problems and those resulting from imperfect match at the antenna inputs, and contributes to the realization of the predicted patterns. The "Wilkinson," is a remarkably simple and uncritical solution to the problem of power division. The 3-element vertical array described later uses a three-way "Wilkinson," shown in Fig. 33. It is constructed by splitting the 50-ohm input three ways, into three quarter-wave lengths of 93-ohm coaxial cable. The center conductors are connected together through 150-ohm noninductive resistors. An alternative approach is to return each center conductor to an ungrounded common point, through 50-ohm noninductive resistors.

Feeding, Switching and Phasing

The 4-element diamond array erected at W1CF uses three Wilkinson two-way

Terminations and switching relays at the center of W1CF array.

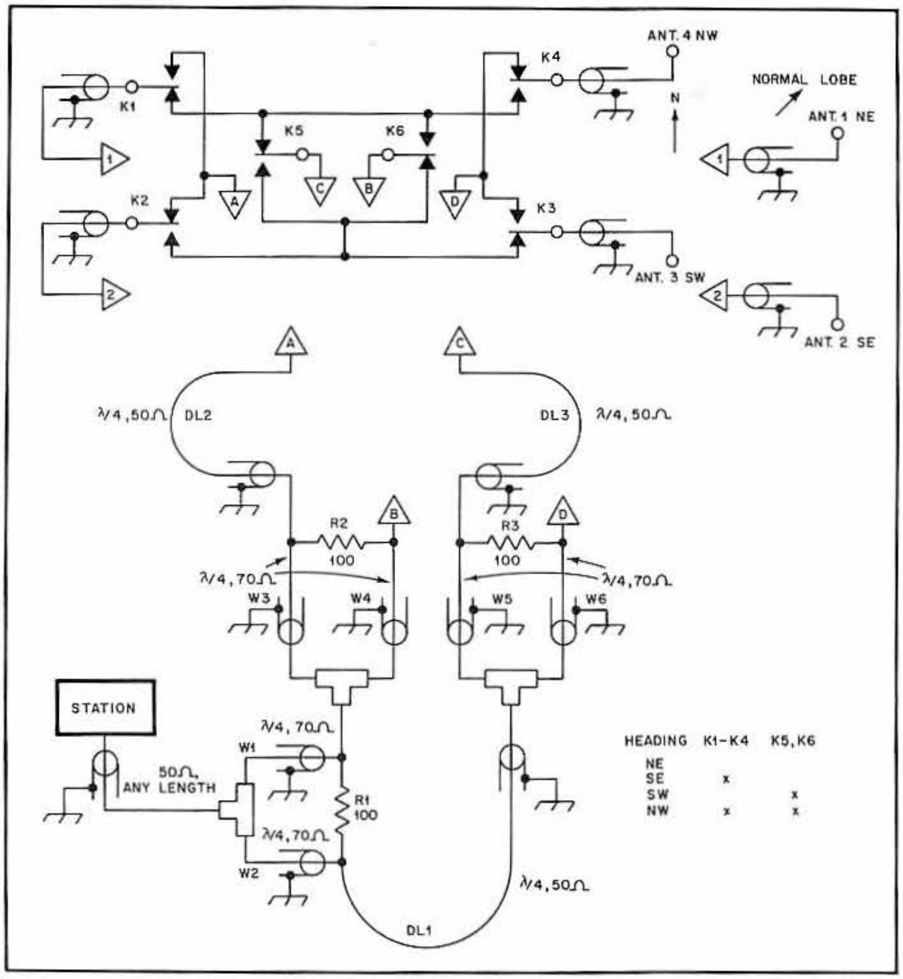

Fig. 31 — Schematic diagram of the Wilkinson Power Dividers, phasing lines, and switching relays (rf connections) for the four-element array. Dc switching commands for the four pattern headings are given at the lower right.

power dividers, as shown in Fig. 31. Phasing is accomplished with three 90-degree sections of RG-8/U cable, DL1, DL2, DL3. The four outputs are at 0°, -90°, -90° and -180° respectively in phase relationship, with power from the transmitter divided into four equal parts. Every effort was made to preserve symmetry throughout the system, in the hybrids, phase shifters, rf switching, feeds, and antenna placement, in order to have the array perform uniformly as it is switched between the four headings.

The switching can be done with six spdt coaxial relays, as in Fig. 31, or with one transfer relay and four spdt relays. With no voltage on the relay coils, the arms are in the positions shown in Fig. 31, giving northeasterly directivity (57° true, on Paris) the heading used most of the time at W1CF. The array is about 500 feet (152.40 m) from the transmitter, necessitating use of remote relays. A prospective user should consider the tradeoff between cable and relay costs. If the array is close to the station, four equal lengths of 50-ohm line can be brought into the station, and the switching done with a 4-pole, 4-position switch.

The long run of line here was dictated by the desire to take advantage of ground qualities and concealment afforded by swampland 500 feet (152.4 m) from the house. Not wanting to be electrocuted, we used 28 volts dc for operation of the switching relays. By judicious use of binary arithmetic, the switching commands require only three wires. Selection of the four beam headings is done at the operating position with a 4-wafer, 4-position switch. Relays should be selected for reliability and ease of weatherproofing. The cluster of relays and teminations shown in the photograph is at the geometrical center of the W1CF array. The connectors are protected against moisture by 3/4-inch (19 mm) heat-shrink tubing, which should extend from the bottom half inch (13 mm) of the connector 2 inches (51 mm) down the cable. In addition, wide plastic wastebaskets are used to protect these components. The expensive precautions were taken as the result of hard experience with moisture in the cables and connectors in the 1965 array.

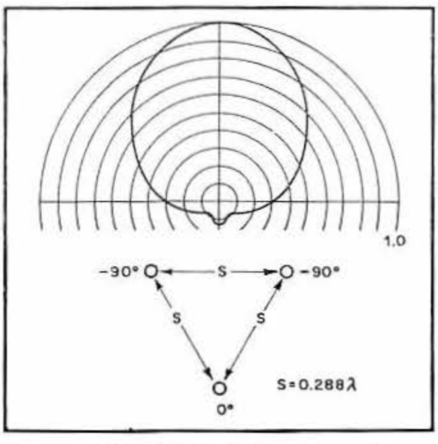

Fig. 32 — Polar plot of relative power, and planar view of the three-element triangular array.

Radiators and Radials

Relays K1 through K4 (Fig. 31) connect to their respective radiators through equal lengths of 50-ohm line, with no special attempt made for matching. The radiators are 57.5 feet (17.53 m) tall, above the base insulators. Constructional details follow the 1965 article, except as outlined below. Base insulators are 2-inch (51-mm) ID PVC pipe, 1/4 inch (6-mm) wall, 18 inches (457 mm) long, affixed to the aluminum with PVC cement and self-tapping screws. Each radiator is guyed at three levels with 3/16-inch (5-mm) galvanized cable, broken every 30 feet (9.14 m) with egg insulators.

Though 120 radials are considered to be the optimum number, only 40 per element are used here. Most are 65 feet (19.81 m) long, no. 12 galvanized wire, lying directly on the ground. This presents no hazard, as there is no foot traffic through the swamp area. The center radials are interconnected to two common busses in the central region, as recommended in reference 2. The authors feel intuitively that the square geometry of the array, in addition to providing considerable symmetry of the mutuals, allows a higher packing density for the radials, reducing ground losses.

Though the array has quarter-wavelength elements, all the feedpoints are at 50 ohms, and any type of vertical element can be used, provided an effort is made to match the inputs, and networks used for matching are *the same for all 4 elements*. Obviously, a height of 5/8 wavelength would be desirable, particularly for arrays built for higher frequencies. An indication of match would be to put rf voltmeters across the 100-ohm resistors, and tune the matching networks simultaneously for minimum voltage across the resistors. A nice thing about the λ/4 case is that the match is close enough that this step has not been necessary.

Testing

Before the array is fired up somebody should go to each element and short the input, while another person watches an ohmmeter placed across the main line at the station end. Make sure that a very low dc resistance is measured. Then, with the array in the normal position, as in Fig. 31, make sure that the resistance across the

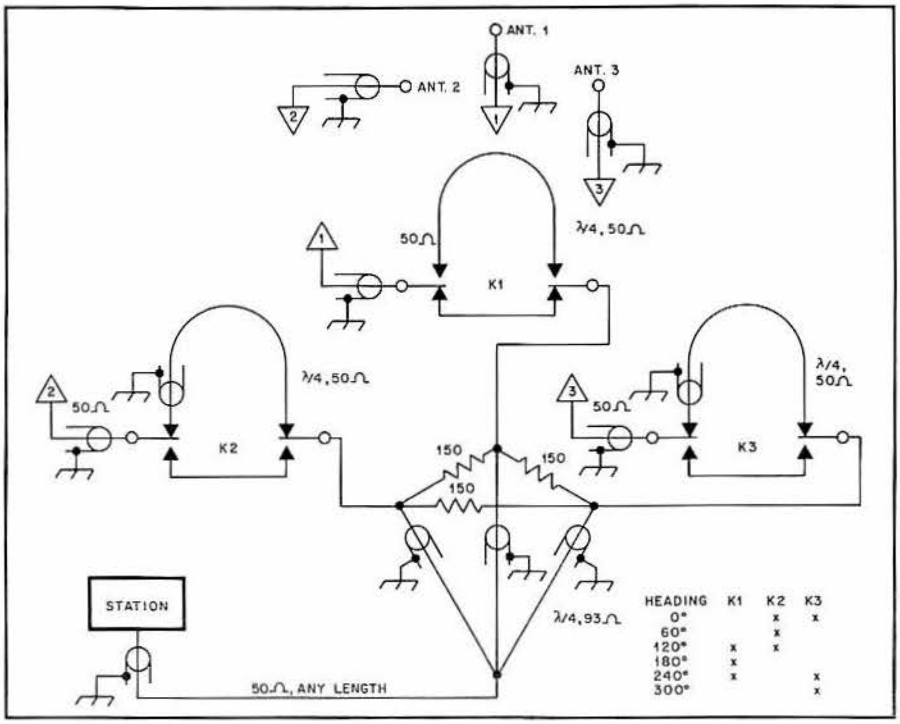

HEADING	K1	K2	K3
0°		x	x
60°		x	
120°	x	x	
180°	x		
240°	x		x
300°			x

Fig. 33 — Schematic diagram of a three-element vertical array and its directivity switching and power dividing circuitry. Interconnecting leads should be as short and symmetrical as possible. The 50-ohm lines between the relays and antenna elements should be all the same length.

input is high. If it is low, check for moisture in the cables or connectors or for other leakage resistance. Silicone grease in the connectors is a good moisture preventive measure. It is recommended that these resistance checks be repeated periodically to be sure that all is well.

When the array is ready for use, go easy at first, as any reflected energy will be dissipated in the 100-ohm resistors. If they become hot, better matching is necessary, and if they blow up, perhaps higher-wattage resistors are needed. At W1CF, resistor banks were made up of 6, 20-watt, 150-ohm surplus resistors in series-parallel to give 100 ohms. These handle the dissipated power when the system is driven from a 4CX1000A amplifier, and after a month of operation there was no evidence of resistor damage. No attempt should be made to hot-switch the directivity. This was done inadvertently, and the result was the need for extensive emergency repairs — in midwinter.

Performance

In the short time that the array has been in use, all continents have been worked on 80-meter phone, including several stations in Japan and India, tough paths from New England. Gain and side and back rejection on all signals except those arriving from very high angles are just what the computer program predicted.

In receiving in the "search" mode, one hand tunes the receiver while the other operates the lobe selector switch, to see which position "listens" best. The big W1CF rotary array for 10, 15 and 20 takes 45 seconds to rotate 360°, which tends to discourage frequent directional checks. With the phased array a complete scan takes but a few seconds. The high front and side rejection eliminates most of the interference from signals in unwanted directions, and in transmitting the clean patterns help to prevent ruffled feathers.

A surprising by-product has been the reduced atmospheric noise pickup from unwanted directions. In particular, when listening toward Europe atmospheric noise coming from electrical storms in the southwest is greatly reduced, improving the signal-to-noise ratio on signals arriving from across the pond. The array is an order of magnitude better on both transmitting and receiving than any antenna previously used on 80 at W1CF. It has more than held its own with the competition, in all four quadrants. Of special interest to operators who like to use both phone and cw on 80 is the fact that the SWR is close to unity, from 3.5 to 4 MHz.

Triangle 3-Element Array

Anticipating that some amateurs may wish to use three rather than four elements, the ARRAY program was used to explore equilateral triangle configurations. Though it was not reduced to practice, a 3-element vertical array with 0.288λ spacing between adjacent radiators looks promising. The lead element is driven at 0° and the two side elements at -90°. The forward lobe is 135° wide, with a gain of close to 5.5 dB, including an estimated 1 dB contributed by vertical focusing. Front-to-back ratio is 12 dB. With suitable switching, the pattern can be aimed in six different directions, again providing 360-degree coverage. The three-way Wilkinson mentioned earlier must be used with three 90-degree sections of 50-ohm line, to provide the proper phase relationships. Several other configurations appear promising, but involve more complex switching and power-dividing circuitry.

Applications for Other Frequencies

The diamond and triangle antenna configurations may find uses at other frequencies, as was done with the 1965 design.[4] Formulas are given below for determining the line lengths and element spacings. A 2-meter phased array can be mounted on the roof of a car and wafer-switched to the desired directions. For people who like to experiment with logic and rf switching with semiconductors, circuits can be devised so that the receiver can scan periodically, and the antenna steered to the desired signal.

The authors wish to thank the many people who have contributed to the success of this project, including W1FRR, K1GXT, W1HMV, W1FXT, K1CCK, M. E. Hines, F. Howe, H. Wells, R. Rearwin, H. P. Scott, and many others. We also wish to thank Microwaves Associates, Inc., for the use of their superb computer facilities.

Formulas for length of quarter-wave (90°) phasing lines, and spacing of radiators, are given below:

Electrical λ/4, solid-dielectric coaxial line, 50 or 70 ohms (RG-8, -11, -58, -59):

$$L_{ft} = \frac{162.36}{f_{MHz}}$$

Electrical λ/4, solid-dielectric coaxial line, 93 ohms (RG-62, -71):

$$L_{ft} = \frac{206.64}{f_{MHz}}$$

300-ohm Twin-Lead — parallel three λ/4 sections to approximate 93-ohm cable (requires 1 input and 3 output 1:1 baluns):

$$L_{ft} = \frac{201.72}{f_{MHz}}$$

Radiator spacing, diamond array:

$$S_{ft} = \frac{246}{f_{MHz}} > 0.25\lambda$$

Radiator spacing, triangular array:

$$S_{ft} = \frac{283.4}{f_{MHz}} > 0.288\,\lambda$$

This material was originally presented in *QST* by Dana W. Atchley, Jr., W1CF, Harold E. Stinehelfer and Joseph F. White, Ph.D.

References

[1] Atchley, "Switchable 4-Element 80-Meter Phased Array," *QST*, March 1965.

[2] National Association of Broadcasters, *Engineering Handbook*, 6th Edition.

[3] Wilkinson, "An N-Way Hybrid Power Divider," *IRE Transactions on Microwave Theory and Techniques*, January 1960.

[4] Botts, "A Four-Element Vertical Beam for 40/15 Meters," *QST*, June 1975.

An Etched-Circuit Monimatch for Checking Your Antenna System

Why buy an expensive SWR indicator when you can build this slick little unit at home? Not only will you save money, but you'll learn something about how SWR meters operate. Try the W1ICP circuit — you'll like it!

A Novice who reads the instruction manual that comes with his transmitter and uses a 50-ohm dummy antenna can follow the tune-up procedures fairly accurately. With the dummy antenna he will find that the settings for the tuning controls will be fairly close to those specified by the manufacturer. However, when an antenna system is attached to the rig, in many instances the adjustments are far removed from any "book" setting. When this happens the Novice finds that he cannot get proper tuning of the rig, or worse yet, actually damages the equipment by trying to "force" it to work.

Nearly every transmitter these days, whether commercial or home-built, has a final amplifier stage that is designed to work into a 50-ohm load. If the load is something other than 50 ohms it may be impossible to tune the amplifier stage correctly. Of course, an important part of the problem is finding out what the load is — or, rather, how far from 50 ohms it happens to be. The piece of measuring gear described in this article is a device for doing just this. However, before describing the Monimatch and what it can do, let's take a little closer look at antenna-system loads.

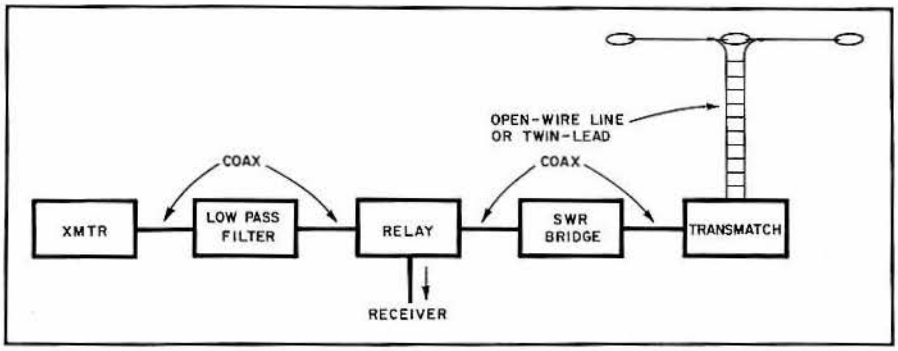

Fig. 34 — A typical setup for using a Monimatch in a multiband antenna system using a single dipole. The length of the dipole is not critical but it should be at least 1/4 wavelength overall at the lowest operating frequency for good efficiency. The feed line can be any length. The antenna relay and low-pass filter may be omitted if not needed.

The "50-Ohm" Load

The evolution of transmitter design since WW II has been influenced by several factors that have led to design that is more or less standard these days. First off, television came along right after the war and the hams quickly discovered that extremely tight shielding of a transmitter was needed to prevent undesired radiation that could cause TVI. However, when tight shielding was installed, band changing without band-switching became a real chore because there were so doggone many screws to unscrew and rescrew.

The one type of tank circuit that lent itself very well to the problem was the pi network. It was a fairly simple job to design a tightly shielded bandswitching transmitter, using the pi network, that would work into a 50-ohm load. Why 50 ohms? Simply because at this time 50-ohm coaxial cable had become a very popular type of transmission line. During the war, techniques were developed that made the manufacture of flexible coaxial cable a

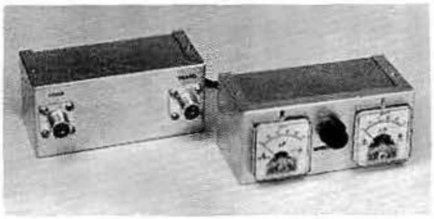

Here is the completed Monimatch with the two meters and the sensitivity control in the box at the right. The sensing unit is at the left.

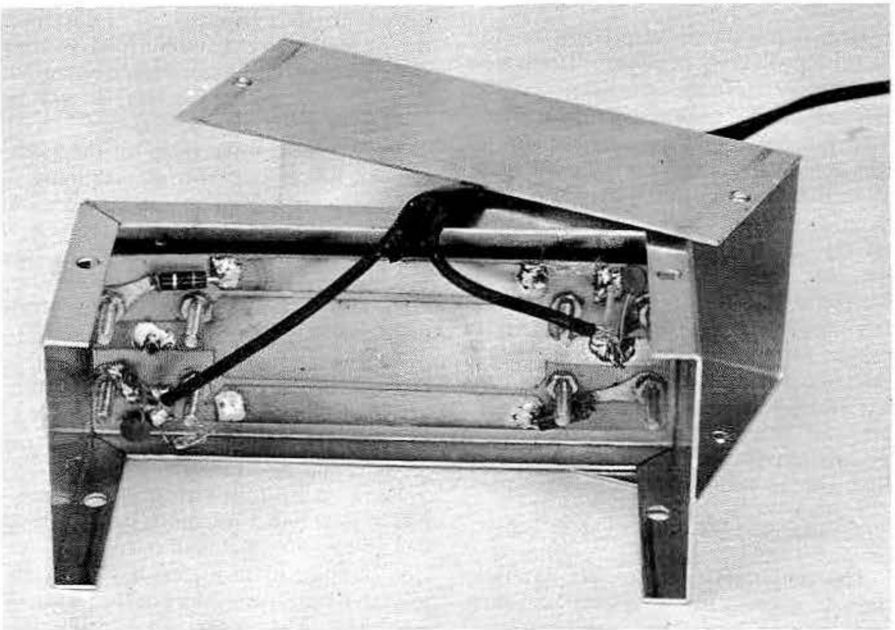

This is the sensing unit of the etched circuit Monimatch. As pointed out in the text, be sure to use a heat sink when soldering the diodes and resistors to the circuit board. The two shielded pickup leads are routed out the back of the Minibox, through a rubber grommet.

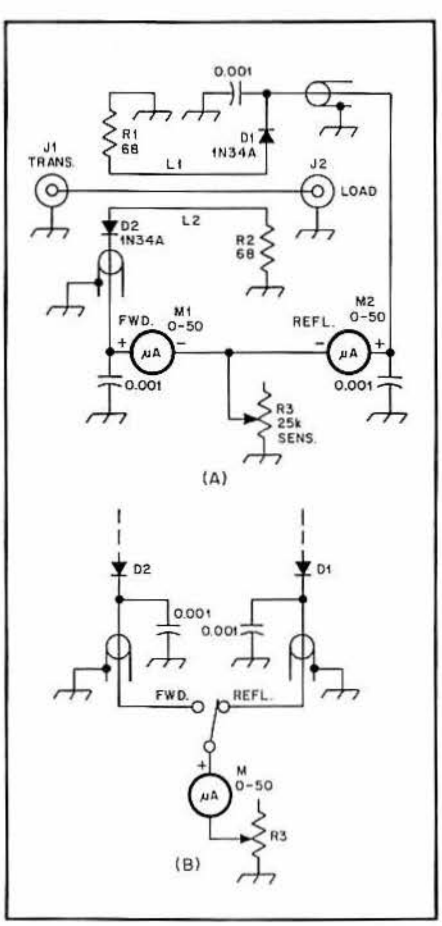

Fig. 35 — Circuit details of the etched-circuit Monimatch. The 0.001 μF capacitors are disk ceramic.
D1, D2 — 1N34A germanium diodes.
J1, J2 — Coax chassis fittings, type SO-239.
L1, L2 — See text and Fig. 36.
M1, M2 — 0-50 μA meter (Lafayette 99 H 5049).
R1, R2 — 68-ohm, 1/2-watt carbon or composition.
R3 — 25 kΩ control, linear taper.
S1 — Spdt switch.

reliable and economical process. So TVI and the availability of coax feed lines were the primary contributing factors that led to our present-day transmitter design.

If the load that is attached to the transmitter is something other than 50 ohms then the transmitter may be difficult to load, depending on a couple of other factors. While it is possible to design a pi network that will handle quite a wide variety of loads, many present day manufacturers, in order to compete in given price ranges, use a minimum number of parts in the tank circuit of the amplifier. For such rigs to operate properly the load must be between 25 and 75 ohms. This of course means that the user *must* furnish a load that will fall within this range.

Transmission Lines

The output terminal on all rigs these days is a coax fitting, which of course implies that a coaxial line must be attached to the rig. This doesn't mean that the coaxial line has to go all the way to the antenna. It could be connected to a Transmatch or a balun. What is important is that the first thing in the antenna system is the coaxial line that is attached to the rig.

Many Novices mistakenly believe that if you attach a 50-ohm coaxial cable to rig you *automatically* have a 50-ohm load. This is not true. The 50-ohm designation on the cable merely means that 50 ohms is the *characteristic impedance* of the cable. The characteristic impedance of a transmission line is determined by the size of the conductors used, the spacing of the conductors, and the dielectric material used to separate and support the conductors. The only time you would have a 50-ohm load using 50-ohm cable is when the line is terminated in its characteristic impedance. In other words, if the antenna has an impedance of 50 ohms then you will have a 50-ohm load on the rig. If the antenna has an impedance of other than 50 ohms then the load at the transmitter will be something other than 50 ohms. This in turn leads us up to a short discussion of standing-wave ratio.

SWR

If a transmission line is terminated in its characteristic impedance, all the power fed into the line from the transmitter will be delivered to the load end — in this case, the antenna. Actually, not quite all the power will reach the antenna because there is always some loss in the transmission line itself. However, what is important is that when the line is terminated in its characteristic impedance none of the power that reaches the end is reflected back toward the transmitter; all of it is used up in the antenna.

When the antenna impedance is different from the line impedance some of the power will be reflected back toward the transmitter end. Standing waves of voltages and currents will then exist on the transmission line. When this happens, the transmitter will no longer "see" a 50-ohm load. Exactly what the load will be will depend on several factors,[1] but suffice to say it will be something other than 50 ohms.

The standing-wave ratio on the transmission line is the ratio of maximum to minimum voltage or maximum to minimum current that exists along the line. If the line were matched in its characteristic impedance the voltage would be the same along the line and of course the SWR would be 1 to 1. The SWR is determined by dividing the resonant antenna impedance into the line impedance, or vice versa. For example, if the antenna impedance were 25 ohms and a 50-ohm line were used, the SWR would be 2 to 1.

For a moment, let's assume that regardless of how bad a mismatch exists, we are still able to tune and load our transmitter. The question then arises, how does the mismatch affect the losses in the transmission line? The answer to the question depends on how efficient the transmission line is.

Remember earlier we said there are always some power losses in every transmission line. If we have a mismatch at the antenna end, some of the power that reaches the end will be reflected back down the line. In traveling back, some of this power will be dissipated in the line, and the higher the SWR the higher these additional losses will be, because a higher SWR means that a greater proportion of the power will be reflected. In a transmission line that is 100 percent efficient (one that has no losses) it follows that regardless of how high an SWR exists, we wouldn't have any losses due to the SWR. Unfortunately, there "ain't no such" line, although some types of lines are much less lossy than others.

Also unfortunately, coaxial lines fall into the class that can be considered to be lossy lines. Just as an example, let's assume that you are using 100 feet (30.48 m) of RG-58/U on the Novice 15-meter band, and you are getting 50 watts out of your 75-watt Novice rig. This 50 watts is what is leaving your transmitter on the way to antenna via the 100 feet (30.48 m) of line. The loss for 100 feet (30.48 m) of RG-58/U at 21 MHz is 1.9 decibels. Translating this figure to power, we would lose about 20 of our 50 watts in the losses in the feed line, leaving only 30 watts to reach the antenna and be radiated. This is assuming the antenna impedance to be 50 ohms, the same as that of the line. If there is a mismatch the losses will be higher, as pointed out earlier. Suppose the SWR is 3 to 1, using the same setup. The *additional* loss in the system because of the SWR would be 1 dB, or a total of close to 3 dB. A 3-dB loss represents almost exactly one-half the power — that is, only 25 watts reach the antenna to be radiated. RG-8/U cable has the same characteristic impedance as RG-58/U but has less loss because it has larger conductors and more spacing between the conductors (and of course is more expensive).

The closest thing to a lossless transmission line is open-wire line. An open wire line on 21 MHz has only 0.08 dB loss per 100 feet (30.48 m). Even with a very high mismatch — for example, an SWR of 20 to 1, the additional losses are still less than 1 db!

This should not be interpreted to mean that coax is an undesirable type of line to use. For beam antennas it is difficult to beat the ease and convenience of using coax. However, for a single antenna, such as a dipole that is to be used on all bands and all frequencies, the best system is one consisting of a Transmatch and a feed line of open-wire line, such as shown in Fig. 34. With this system you can forget about line losses, SWR on the line, and mismatches between the antenna and feed line. By correctly adjusting the Transmatch, you can always give your rig a 50-ohm load regardless of what the load is on the antenna side of the Transmatch. If we use an SWR bridge in the short length of 50-ohm coaxial line that connects the rig to the

Transmatch, we can adjust the Transmatch so that the 50-ohm line shows a match, or SWR of 1 to 1, and the transmitter always sees a 50-ohm load.

The SWR bridge in the photographs and drawings is simple to build, and when installed in 50-ohm cable, will show the relative mismatch in the line, and also will indicate when you get the Transmatch properly adjusted for a match. The SWR bridge can also be used as an output indicator, which is very handy when tuning up the rig.

SWR Bridge Circuit Details

The etched-circuit Monimatch shown here is a reflectometer that samples the forward and reflected voltage in a 50-ohm line. Fig. 35 shows the circuit diagram. L1 and L2 are the pickup lines. In operation, a very small amount of power is coupled into the pickup lines and the rf voltages are rectified by D1 and D2. The rectified voltages are then fed to the two meters, M1 and M2, and the SWR then determined from the readings.

While a Monimatch is not a precise piece of measuring equipment, the SWR readings will be close enough for practical purposes. In order to determine the SWR, the forward-reading meter is set to full scale by adjusting R3, the sensitivity control, and then the reading on the "reflected" meter is noted. The formula for the SWR using this system is

$$\frac{F+R}{F-R}.$$

For example, let's assume the "reflected" reading is 5, with the "forward" reading being 10. Then 15 divided by 5 (that is, 10 + 5 divided by 10 − 5) would mean the SWR is 3 to 1. The closer the "reflected" reading is to zero, versus full-scale reading on "forward," the closer you come to 1 to 1, or a matched condition.

In the unit shown, two meters are used, one for the forward and the second for the reflected reading. However, if desired a single meter can be used and switched as shown in Fig. 35 at B. We used two meters in the indicator as this permits constant monitoring of what is happening in the line. The meters are inexpensive ones made in Japan.

Construction Details

Fig. 36 is a full-sized template of the etched circuit board. A *QST* article[2] went into detail showing simple methods for making etched circuits, so we won't treat the process here. In making this board, it is suggested that the board be covered with masking tape and then the pattern of Fig. 36 transferred to the tape. Using a sharp knife edge or razor blade and a straight edge, the masking tape can be carefully and accurately cut to the pattern.

After the board is etched, it can be positioned in the Minibox over the chassis connector holes and the board can then be marked at the drilling points for the mounting holes and the center conductor pins of the coax fittings. When installing the mounting screws, be sure they don't short to the center — conductor portion of the foil on the board.

There are a couple of other construction points that should be stressed. The lead lengths on R1 and R2 should be kept as short as possible. Also, be sure to use carbon or composition resistors, *not* wire-wound. When mounting the resistor and diode ends to the pickup sections, L1 and L2, the connections should be at the very ends of the sections. Also, use a heat sink when soldering the leads on any of the components mounted on the board, as too much heat from the iron can ruin the component.

The Monimatch and meters are mounted in separate Miniboxes, 2-1/4×2-1/4×5 inches (57 mm × 57 mm × 127 mm). The two connectors on the Monimatch sensing unit, J1 and J2, are mounted with their center pins 3-3/4 inches (95 mm) apart, center-to-center. In order to avoid an impedance "bump" in the feed line when the bridge is inserted in the line, the circuit board should be mounted 1/4 inch (6 mm) above the base of the Minibox. Quarter-inch (6 mm) spacers can be used under the circuit board at the screws holding both the board and the coax fittings to position the board accurately.

Shielded conductors should be used for the connections from the diodes to the meter enclosure. The shields should be grounded to the chassis at both boxes. These lead lengths are not critical, and the Monimatch can be remote from the meter indicator.

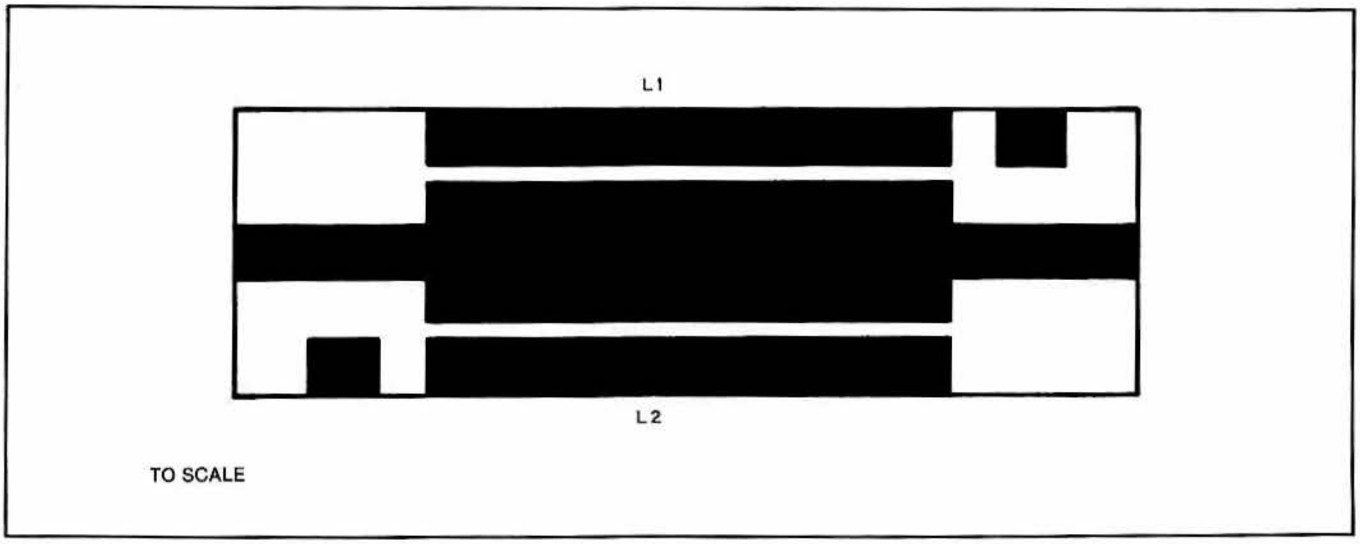

Fig. 36 — Etched circuit board template. The foil side is shown; the etched portion is shaded.

Using the Bridge

If you are using coax feed from the rig to the antenna, the bridge can be installed at any convenient spot in the line. If you are using a Transmatch, similar to the system as shown in Fig. 34, the Monimatch should be installed on the transmitter side of the Transmatch. Any relays or filters should be installed between the bridge and the transmitter, as shown.

Set R3 so that the arm of the control is at the top of the resistance — in other words, with all the resistance in series with the meter circuit. Tune up your rig in the normal fashion, and once tuned up adjust the sensitivity of the "forward" meter by moving the arm of R3 until the meter reads full scale. You can then determine the SWR by the formula mentioned earlier.

When adjusting a Transmatch, feed just enough power through the system to obtain about half-scale reading on the "forward" meter and then adjust the Transmatch controls for a match, as indicated by zero reflected power. You may have to adjust R3 as you adjust the Transmatch to keep the "forward" meter from reading more than full scale. Once you have the Transmatch adjusted for a 1-to-1 ratio as indicated by the bridge, the transmitter can be loaded up in the normal manner. We usually reduce the forward reading to about half scale, and then tune the rig for maximum output, as indicated by the meter. When doing this, you may notice that maximum output as indicated by the bridge meter occurs at some setting other than the normal transmitter plate meter "dip" reading. (Normally, the instruction manuals tell you to tune for a plate dip if the transmitter has no output meter.) However, the amplifier stage will work better if you tune for maximum output rather than the dip — keeping the plate loading within the transmitter ratings, of course.

Once you become familiar with the use of the bridge and interpreting the readings, you'll find it a very valuable device in your station. This material originally presented in *QST* by Lewis G. McCoy, W1ICP.

References

[1] [Editors Note: It is recommended that the newcomer obtain a copy of *The ARRL Antenna Book* for more detailed information on transmission line operation than can be provided in this article.]

[2] Schiebold, "Fast'n' Easy Printed Circuit Boards," *QST*, August 1969.

In-Line RF Power Metering

Problems are frequently encountered in designing and building reflected-power meters and SWR bridges. Examples of practical in-line rf wattmeters are given here, along with complete details for building a unit that will provide two power ranges, forward and reflected, for use from 3.5 to 30 MHz.

μA
REF
FWD
ANTENNA

It is neither costly nor difficult to build an rf wattmeter. And, if the instrument is equipped with a few additional components it can be switched to read reflected power as well as forward power. With the foregoing feature the instrument can be used as an SWR meter for antenna matching and Transmatch adjustments.

Perhaps the most difficult task faced by the constructor is that of calibrating the power meter for whatever wattage range he desires to have. The least difficult method is to use a commercial wattmeter as a standard. If one is not available, the power output of the test transmitter can be computed by means of an rf ammeter in series with a 50-ohm dummy load, using the standard formula, $P = I^2R$. Or, if one is not interested in obtaining power readings the bridge can be used solely as an SWR indicator, as is done with the Monimatch-style SWR bridge.[1]

The advantage of the circuits shown here over those of Monimatch bridges is that these instruments are not frequency-sensitive. Monimatch indicators become more sensitive as the operating frequency is increased, thus making it impractical to calibrate them in watts for more than one band, or for more than one portion of a given band. The units described here are more sensitive than Monimatches are. This makes it possible to calibrate them for power levels as low as 1 watt, full scale, in any part of the hf spectrum.

All of the circuits shown in this article are similar to the basic one which was described in *QST*.[2] Some of the circuits are those of commercial power meters, and are used to illustrate variations in the basic Bruene design. The reader may wish to experiment with some of these circuits.

Design Philosophy

Referring to the circuit of Fig. 37B, the circuit used by Collins Radio Company, the transmission-line center conductor passes through the center of a toroid core and becomes the primary of T1. The multi-turn winding on the core functions as the transformer secondary. Current flowing through the line-wire primary induces a voltage in the secondary which causes a current to flow through resistors R5 and R6. The voltage drops across these resistors are equal in amplitude, but 180 degrees out of phase with respect to common or ground. They are thus, for practical purposes, respectively in and out of phase with the line current. Capacitive voltage dividers, C3C7 and C4C8, are connected across the line to obtain equal-amplitude voltages *in* phase with the line voltage, the division ratio being adjusted so that these voltages match the voltage drops across R5 and R6 in amplitude. (As the current/voltage ratio in the line depends on the load, this can be done only for a particular value of load impedance. Load

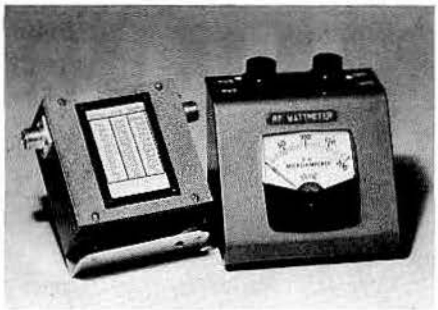

values chosen for this standardization are pure resistances that match the characteristic impedance of the transmission line with which the bridge is to be used, usually 50 or 75 ohms.) Under these conditions, the voltages rectified by D1 and D2 represent, in the one case, the vector *sum* of the voltages caused by the line current and voltage, and in the other, the vector *difference*. With respect to the resistance for which the circuit has been set up, the sum is proportional to the forward component of a traveling wave such as occurs on a transmission line, and the difference is proportional to the reflected component.

The Collins circuit uses two 8-μF capacitors, C5 and C6, to permit the meter to approach the PEP level during ssb operation. The dc voltages in the forward and reflected lines charge the capacitors to permit a near-peak reading. The discharge rate is set by the series calibrating resistors, R1 through R4, and is dependent upon which of them is switched into the metering line at a given time. The circuit of Fig. 37B uses two 43-pF capacitors, C1 and C2, to cancel the inductive reactances of R5 and R6. Such reactance may become manifest at the high end of the range for which the instrument is built. If reactance is present in that part of the circuit the meter readings may not be accurate, especially at 10 and 15 meters. The capacitors were not needed in the circuit of Fig. 38, perhaps because the resistor leads were very short when they were mounted on the etched-circuit board.

Some Design Hints

It is important that the layout of any rf bridge be as symmetrical as possible if good balance is to be had. The circuit-board layout for the instrument of Fig. 38 meets this requirement. Also, the input and output parts of the equipment should be isolated from the remainder of the circuit so that only the sampling circuits feed voltage to the bridge. A shield across the end of the box which contains the input and output jacks, and the interconnecting line between them, is necessary. If stray rf gets into the bridge circuit it will be impossible to obtain a complete zero reflected-power reading on M1 even though a 1:1 SWR exists.

Referring again to Fig. 38, resistors R1 and R2 should be selected for the best null reading when adjusting the bridge into a resistive 50- or 75-ohm load. Normally the value will be somewhere between 10 and 47 ohms. The 10-ohm value worked well with the homemade instruments shown here. It was found that half-watt resistors exhibited somewhat less inductive reactance at 30 MHz than did some 1-watt units tried. R1 and R2 should be as closely matched in resistance as possible. They need not be exactly 10 ohms, so a VTVM can be used to match them. The resistors used for the circuit of Fig. 38 were actually 10.5 ohms each, and were chosen from an assortment of "10-percenters" on hand.

Silver-mica capacitors C3 and C4 were close enough in value so that special selection was not required. There should be enough leeway in the ranges of C1 and C2 to compensate for any difference in the values of the 330-pF capacitors. Ideally, however, C3 and C4 should be matched in value.

Diodes D1 and D2 should also be matched for best results. An ohmmeter can be used to select a pair of diodes whose forward dc resistances are within a couple of ohms of being the same. Similarly, the back resistances of the diodes can be matched. The matched diodes will help to assure equal meter readings when the bridge is reversed. (The bridge should be perfectly bilateral in its performance characteristics.) Germanium diodes are used in the bridges described here, but

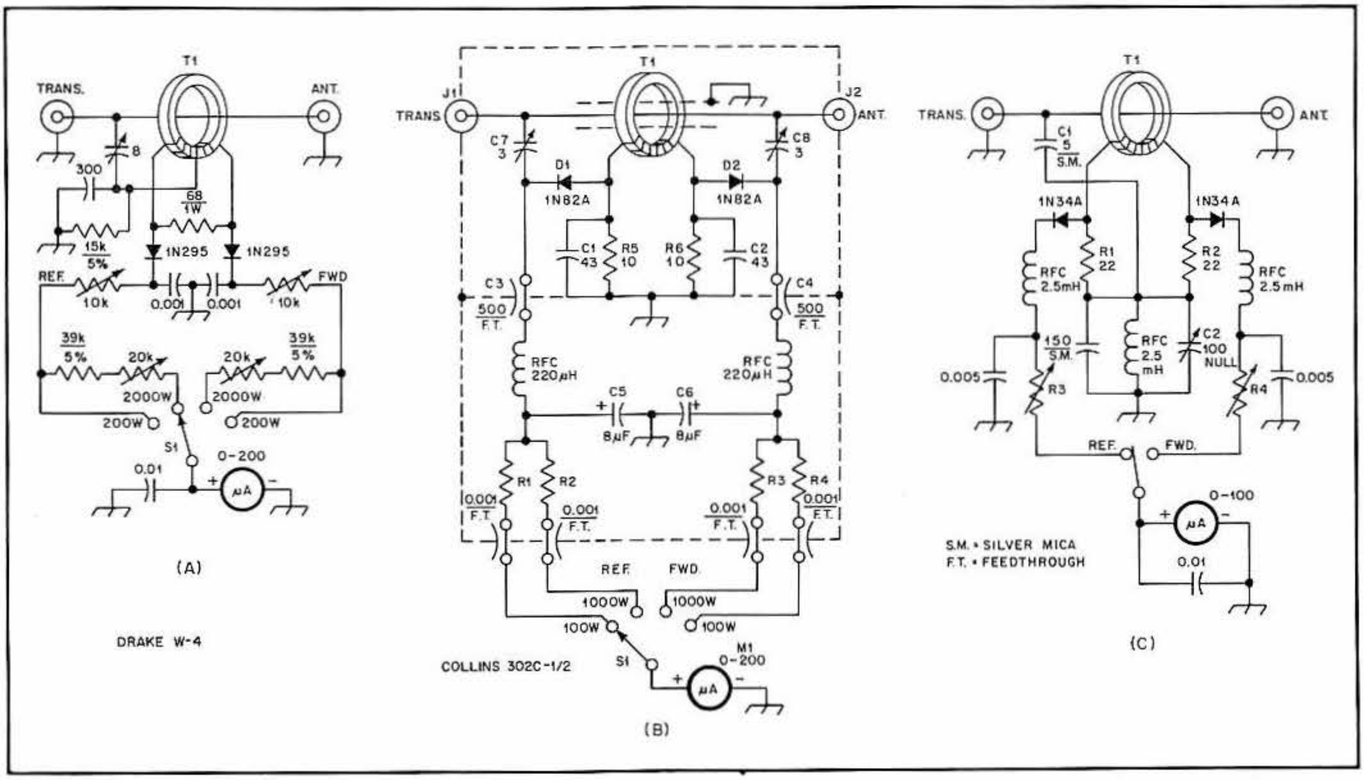

Fig. 37 — Schematic diagrams of typical in-line power meters. At A, the R. L. Drake W-4 instrument. It uses a center-tapped transformer at T1 and has but one capacitive voltage divider in the sensing circuit. The circuit at B is discussed in the text, and is used by Collins Radio Company. The capacitive voltage dividers in this circuit use two 500-pF feedthrough capacitors in place of the silver-mica capacitors specified in Fig. 38. Capacitors C5 and C6 permit a charge time that enables the meter to read near-peak power on ssb. Calibrating resistances R1 through R4 are factory selected. The circuit at C is similar to one used by Comdel in their power meter. In this circuit C1 is a fixed-value (small) capacitor, and the bridge is nulled by the larger capacitor in the divider, C2.

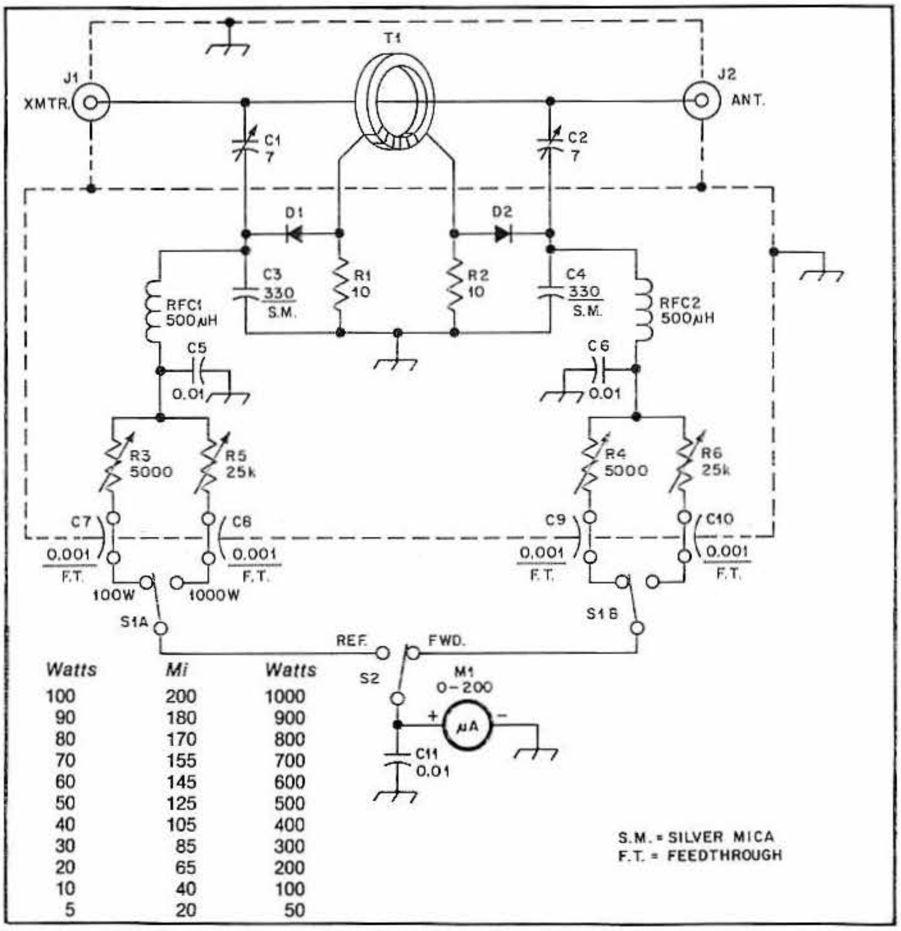

Watts	*Mi*	*Watts*
100	200	1000
90	180	900
80	170	800
70	155	700
60	145	600
50	125	500
40	105	400
30	85	300
20	65	200
10	40	100
5	20	50

Fig. 38 — Schematic diagram of a practical power wattmeter. A calibration scale for M1 is shown also. Fixed-value resistors are 1/2-watt composition. Fixed-value capacitors are disk ceramic unless otherwise noted. Decimal-value capacitances are in µF. Others are pF. Resistance is in ohms; k = 1000.

C1, C2 — 1.3- to 6.7-pF miniature trimmer (E. F. Johnson 189-502-4. Available from Newark Electronics, Chicago, IL).
C3-C11, incl. — Numbered for circuit-board identification.
D1, D2 — Small-signal germanium diode. IN34A, etc. (see text).
J1, J2 — Chassis-mount coax connector of builder's choice. Type SO-239 used here.
M1 — 0- to 200-µA meter (Triplett type 330-M used here.)
R1, R2 — Matched 10-ohm resistors (see text).
R3, R4 — 5000-ohm printed-circuit carbon control (IRC R502-B).
R5, R6 — 25,000-ohm printed-ciruit carbon control (IRC R253-B).
RFC1, RFC2 — 500-µH rf choke (Millen 34300-500 or similar).
S1 — Dpdt single-section phenolic wafer switch (Mallory 3222J).
S2 — Spdt phenolic wafer switch (Centralab 1460).
T1 — Toroidal transformer; 35 turns of no. 26 enam. wire to cover entire core of Amidon T-68-2 toroid (Amidon Assoc., 12033 Otsego St., No. Hollywood, CA 91607).

silicon diodes can also be used. Silicon diodes conduct at a higher voltage than germanium diodes do — approximately 0.7 volt — and will not work too well in low-power wattmeters. Some silicon diodes were tried, but ceased to conduct at approximately 8 watts in the circuit of Fig. 38. This effect can cause misleading results when low values of reflected power are present during antenna adjustments. The SWR can appear to be zero when actually it isn't. The germanium diodes conduct at approximately 0.3 volt, making them more suitable for low-power readings.

Any meter whose full-scale reading is between 50 microamperes and 1 milliampere can be used at M1. The more sensitive the meter, the more difficult it will be to get an absolute reflected-power reading of zero. Some residual current will flow in the bridge circuit no matter how carefully the circuit is balanced, and a sensitive instrument will detect this current flow. Also, the more sensitive the meter, the larger will have to be the calibrating resistances, R3 through R6, to provide high-power readings. A 0- to 200-microampere meter represents a good compromise for power ranges between 100 and 2000 watts.

Construction

The power meter of Fig. 38 is built in two sections. The rf circuit and the calibrating resistors are housed in a 4 × 4 × 2-inch (102 × 102 × 51 mm) aluminum utility box. All components other than J1, J2, and the feedthrough capacitors, are assembled on the etched-circuit board.[3] Switches S1 and S2, and the meter, M1, are installed in a sloping-panel utility box which measures 5 × 4 inches (127 × 102 mm). Four-conductor shielded cable — the shield serving as the common lead — is used to join the two pieces. There is no reason why the entire instrument cannot be housed in one container, but it is sometimes awkward to have coaxial cables attach to a unit that occupies a prominent place in the operating position. Built as shown, the two-piece instrument permits the rf pickup head to be concealed behind the transmitter, while the control head can be mounted where it is accessible to the operator.

Toroidal transformer T1 fits into a cutout area on the circuit board. A 1-inch (25 mm) long section of RG-8/U cable — vinyl jacket and shield braid removed — provides a snug fit in the center hole of the toroid, and is used to complete the line between J1 and J2. The inner conductor of the RG-8/U section solders to the circuit board, thus holding T1 in place.

A flashing-copper shield divides T1 and its center-conductor line from the remainder of the circuit. This partition is shown in dotted lines in Fig. 38. It is mounted on the non-foil side of the circuit board and is secured at each end to solder lugs which are mounted under the retaining screws for J1 and J2.

The circuit board is held in place, at the end near T1, by means of an aluminum L bracket. The circuit-board end nearest the feedthrough capacitors is held in place by a no. 6 spade bolt. A solder lug is mounted under the no. 6 nut (outside the case) which secures the spade bolt. The lug serves as a connection point for the common lead between the rf head and the control box. Two solder lugs are mounted under the bottom two retaining screws of each coax connector. The free ends of the lugs are soldered to the copper foil of the circuit board.

A partition is visible in the foil-side view of the rf head. It can be eliminated if desired, since it did not prove necessary when the unit was tested. Similarly, an extra shield partition is shown on the top side of the board. It too can be eliminated, for it turned out to be unnecessary. The flashing-copper shield discussed earlier is the only one required for the circuit of Fig. 38.

Checkout and Tune-up

Once the instrument is wired and ready to test it should be inspected for unwanted solder bridges between the circuit-board foils. It is usually a good idea to scrape out the rosin buildup between the foils, and this can be done with the blade of a small screwdriver. A continuity check for "opens" and "shorts" should also be made before power is applied to the unit. Make certain that the diodes are installed for the correct polarity — the banded ends (cathodes) toward C1 and C2.

Connect a noninductive 50-ohm dummy load to J2. A Heath Cantenna or similar load will serve nicely for adjustment purposes. Place S2 in the FORWARD position, and set S1 for the 100- range. An rf ammeter or calibrated power meter

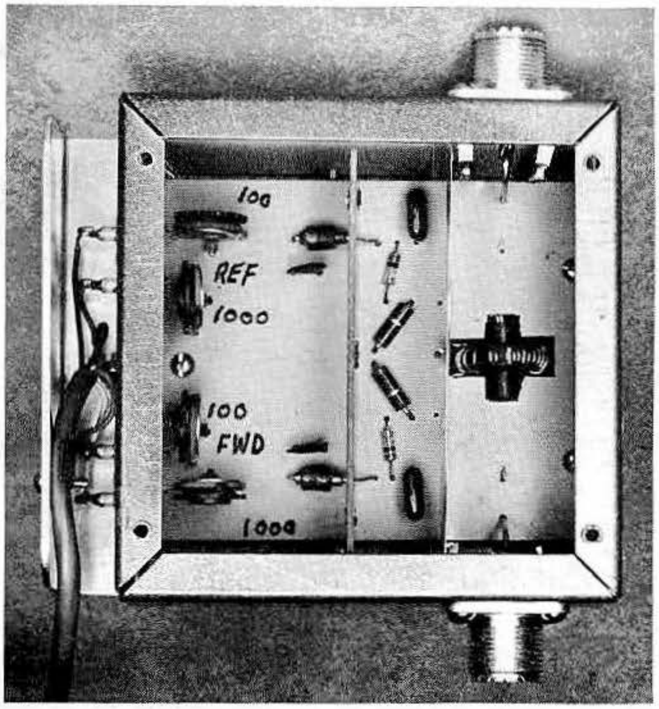

Top view of the rf head for the circuit of Fig. 38. A flashing-copper shield isolates the through-line and T1 from the rest of the circuit. The second shield (thicker) is not required and can be eliminated from the circuit. If a 2000-watt scale is desired, fixed-value resistors of approximately 22 kΩ can be connected in series with high-range printed-circuit controls. Or, the 25 kΩ controls shown here can be replaced by 50 kΩ units.

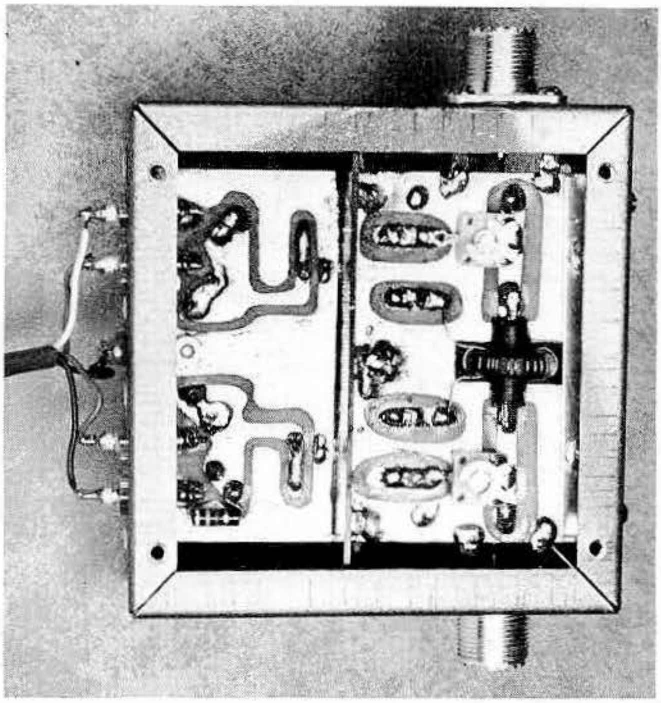

Bottom view of the rf head for the circuit of Fig. 38. The fixed-value resistor at the lower left does not belong in the circuit, but was added as a shunt for one of the calibrating controls which was too high in value — a 50 kΩ unit that was on hand. The shield partition shown here proved unnecessary and can be eliminated.

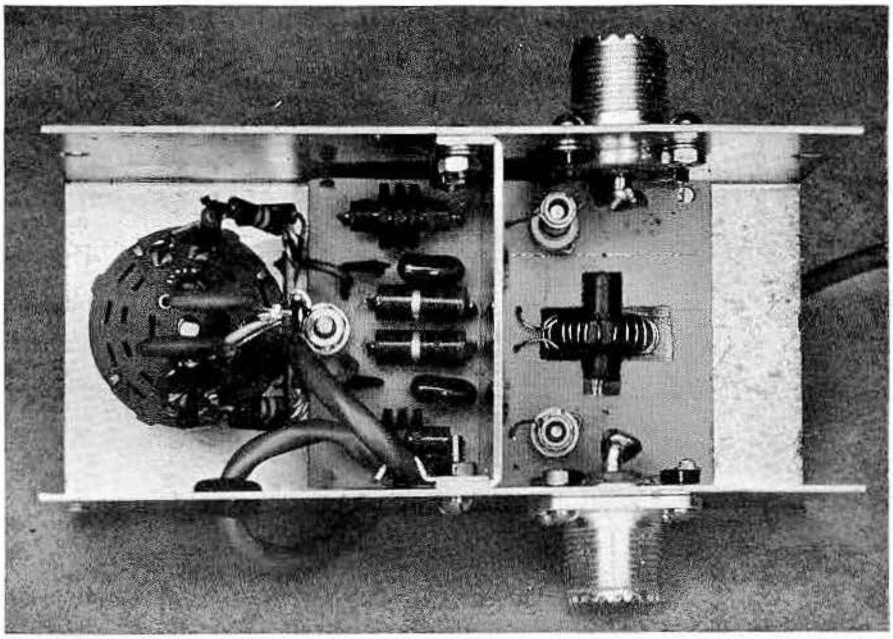

Fig. 39 — Inside view of the 5- and 50-watt power meter rf head. Component values are the same as in the circuit of Fig. 38, except for the calibrating resistances (see text). An aluminum shield isolates the through-line and toroid from the remainder of the circuit. A 50-μA meter is used in this model.

should be connected between J2 and the dummy load during the tests, providing power calibration points against which to plot the scale of M1. Apply transmitter output power to J1, gradually, until M1 begins to deflect upward. Increase transmitter power and adjust R4 so that a full-scale meter reading occurs when 100 watts is indicated on the rf ammeter or other standard in use. Next, switch S2 to REFLECTED and turn the transmitter off. Temporarily short across R3, turn the transmitter on, and gradually increase power until a meter reading is noted. With an insulated screwdriver adjust C2 for a null in the meter reading.

The next step is to reverse the coax connections to J1 and J2. Place S2 in the REFLECTED position and apply transmitter power until the meter reads full scale at 100 watts output. In this mode the REFLECTED position actually reads forward power because the bridge is reversed. Calibrating resistance R3 is set to obtain 100 watts full scale during this adjustment. Now, switch S2 to FORWARD and temporarily place a short across R4. Adjust C1 for a null reading on M1. Repeat the foregoing steps until no further improvement can be obtained. It will not be necessary to repeat the nulling adjustments on the 1000-watt range, but R5 and R6 will have to be adjusted to provide a full-scale meter reading at 1000 watts. If insufficient meter deflection is available for nulling adjustments on the 100-watt range, it may be necessary to adjust C1 and C2 at some power level higher than 100 watts. If the capacitors tune through a null, but the meter will not drop all the way to zero, chances are that some rf is leaking into the bridge circuit through stray coupling. If so, it may be necessary to experiment with the shielding of the through-line section of the rf head. If only a small residual reading is noted it will be of minor importance and can be ignored. In the circuit of Fig. 38 there remained approximately one-half a meter division when the null was reached, and this occurred only on the 1000-watt range. Since this was representative of less than 2 watts of power, it was deemed inconsequential.

With the component values given in Fig. 38 the meter readings track for both power ranges. That is, the 10-watt level on the

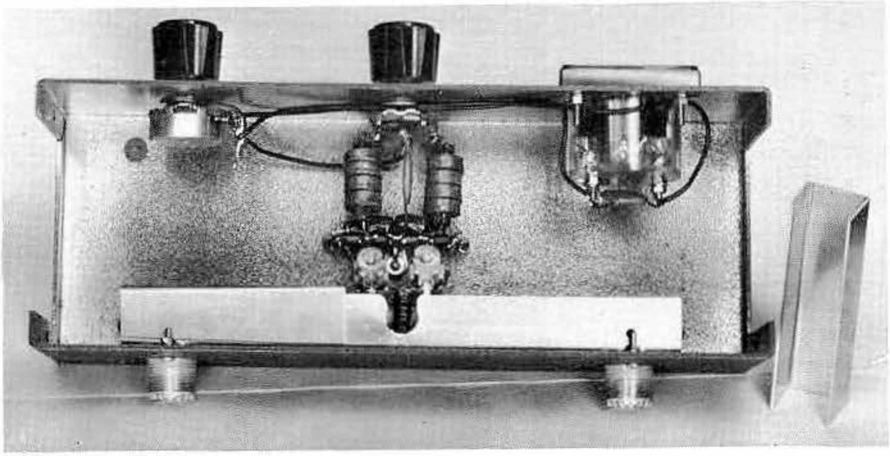

Inside view of the modified Heath HM-15 bridge. The new components are grouped at the center of the chassis on a 5-lug terminal strip. The nulling capacitors are connected between the inner line and the terminal strip. Press-fit aluminum shield covers are slipped over the trough line to aid in rf isolation. One cover is in place; the other is at the right of the photo. Improved shielding might be effected by installing an aluminum plate between the terminal strip and the two nulling capacitors.

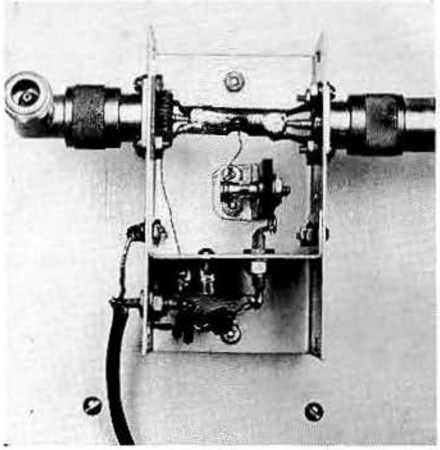

Inside view of a 2000-watt power meter built by N1RM. This bridge is patterned after the circuit of Fig. 37B. Point-to-point wiring is used throughout, thus avoiding the need for a circuit board. Two piston trimmers are used for the nulling capacitors and are mounted one above the other on a phenolic block. The two 500-pF feedthrough capacitors are part of the capacitive voltage dividers.

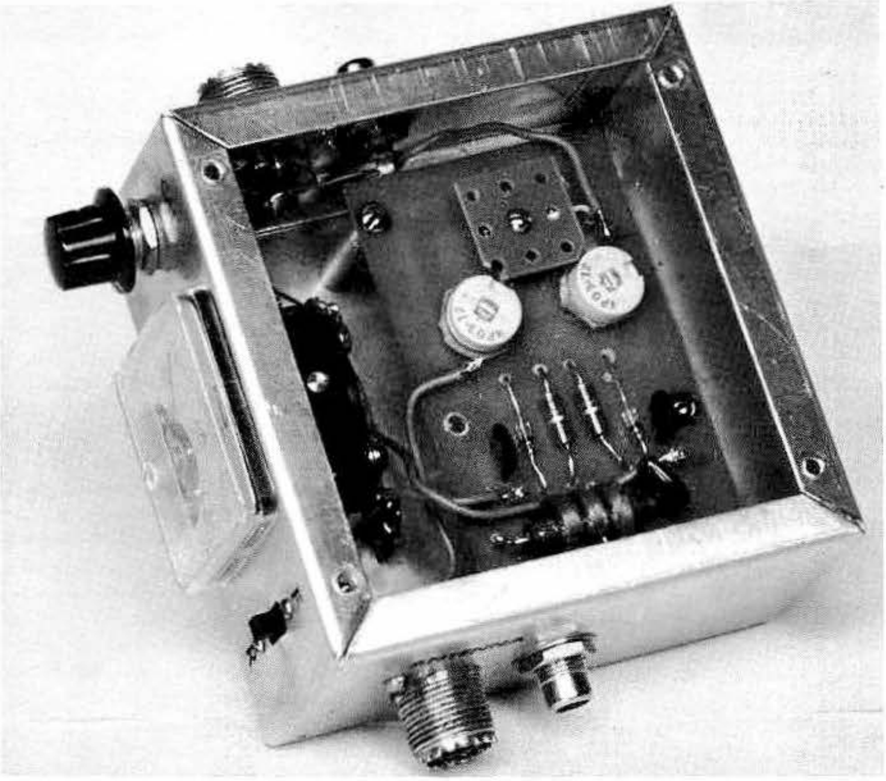

Fig. 40 — Inside view of a 3-watt power meter for QRP rigs. Its circuit is given on page 16 of June 1969 *QST*. Ceramic trimmers are used for nulling the bridge. Type SO-239 connectors are paralleled with phono jacks to add versatility. A 4 X 4 X 2 inch (102 X 102 X 51 mm) utility box houses the entire unit.

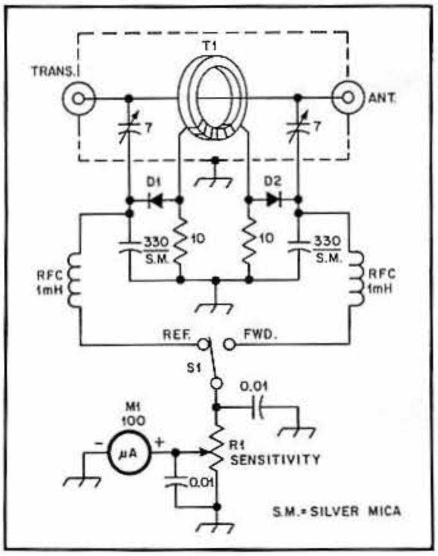

Fig. 41 — Schematic diagram of the modified Heath HM-15 SWR meter. D1, D2, M1, R1 and S1 are original components from the HM-15. T1 is the same as in Fig. 38. See text for additional details.

100-watt range, and the 100-watt point on the 1000-watt range fall at the same place on the meter scale, and so on. This no doubt results from the fact that the diodes are conducting in the most linear portion of their curve. Ordinarily, this desirable condition does not exist, making it necessary to plot separate scales for the different power ranges.

Tests indicate that the SWR caused by insertion of the power meter in the transmission line is negligible. It was checked at 28 MHz and no reflected-power could be noted on a Bird wattmeter. Similarly, the insertion loss was so low that it could not be measured with ordinary instruments.

Other Circuits

Additional circuits and photos are shown for variations in the basic design used at Fig. 38. A low-power model, having scales for 5 and 50 watts, is shown in Fig. 39. It uses fixed-value resistors for meter calibration. The required values of resistance were first determined by temporarily inserting a potentiometer in the meter line, obtaining the required full-scale reading, then substituting fixed-value resistances of the proper ohmage. The meter readings for the two power ranges do not track in this model.

A low-power meter was designed for use with the QRP transmitter described in June 1969 *QST*. It is shown in Fig. 40, and has a full-scale calibration of 3 watts. To obtain additional sensitivity, the primary of the toroidal transformer consists of a one-turn link instead of the single wire that would normally pass through the hole in the toroid core.

Some experiments were conducted to see if a Heath HM-15 SWR bridge (a Monimatch type) could be modified to work in a Bruene circuit. The results were satisfactory, and the circuit is given in Fig. 41. No attempt was made to obtain a calibration scale for the meter. The unit is being used as a simple SWR indicator, but now has better sensitivity in the lower part of the hf spectrum — 7 watts, full scale, from 3.5 to 30 MHz. Also, the instrument is no longer "frequency-conscious" as was the case before modification. The original pickup lines were discarded, the FWD-REV panel switch was rotated 180 degrees so that the labels were correct for the new circuit, and press-fit shield covers were installed on the trough line as shown in the photo. A power scale could be plotted by setting the sensitivity control in a fixed position — possibly replacing the existing control with a screwdriver-adjust type. A new 100-μA meter could be installed to provide a better scale for calibration in watts.

It was necessary to dismantle the trough line so that the toroidal transformer could be slipped over the inner line. A few wraps of mylar tape were wound over the center of the inner line to insulate the toroid winding from the line, and to provide a snug fit to keep the toroid in place. The trough was notched out with a nibbling tool to allow clearance for the toroidal transformer. Additional shielding can be added between the line and the rest of the circuit to further assure a zero meter reading in the reflected position.

It is hoped that the experimenter will find sufficient information here to enable him to build a power meter that will satisfy his specific needs. This material was originally presented in *QST* by Doug DeMaw, W1FB.

References

[1] "Monimatch Mark II," *QST*, February 1957.
[2] Bruene, "An Inside Picture of Directional Wattmeters," *QST*, April 1959.
[3] The etched-circuit board pattern and parts layout sheet for this power meter are available from ARRL hq. Send 25 cents and an s.a.s.e. Check with Circuit Board Specialists, P. O. Box 969, Pueblo, CO 81001 for negative and circuit boards.

A Resistive Antenna Bridge — Simplified

SWR indicators are neat, but when it comes to making measurements of resonant impedances, you need a true bridge type of instrument. Measure those low antenna resistances directly with this W2FMI circuit, and do it up to 30 MHz with ease!

As shown previously,[1] a bridge using a single variable capacitor can be used to measure resonant impedances of antennas and losses in coils with reasonable accuracy. With addition of an L/C network, the bridge can also be very useful in adjusting antennas for resonance conditions. Recently it was brought to the author's attention that a bridge using a variable resistance could also perform well. This is possible because the range of antenna impedances is generally low enough to allow for low values of resistance in the arms of the bridge. Under these conditions parasitic capacitances are negligible up to 30 MHz.

Resistive bridges have the advantages of small size, readily available components, and dial calibrations which can be made practically linear. Two bridges are shown with their respective calibration curves. The first is a linear device. The second expands the lower portion of the calibration curve, 0 to 30 ohms, to measure lower values of antenna impedance and coil losses more accurately.

Bridge Details

The resistive bridge with a variable resistance in one arm and equal resistances in the other two is shown in the photographs, and schematically in Fig. 42A. The nonlinear bridge is similar mechanically. The small enclosures and compact layout minimize lead lengths. With the linear version the dial practically reads directly in

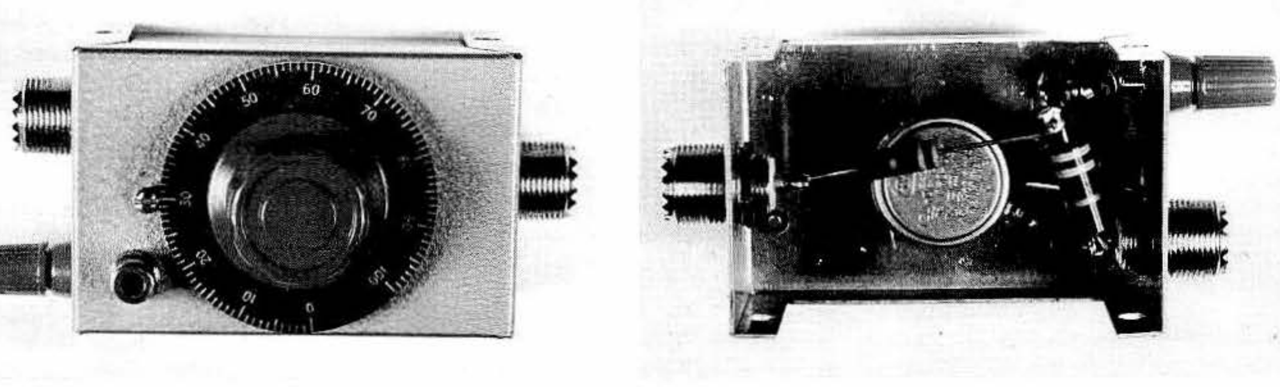

Exterior and interior view of the resistive bridge. Appearance and component arrangement are essentially the same for the two types described.

---

[1] Sevick, "Simple RF Bridges," April 1975 *QST*, page 11.

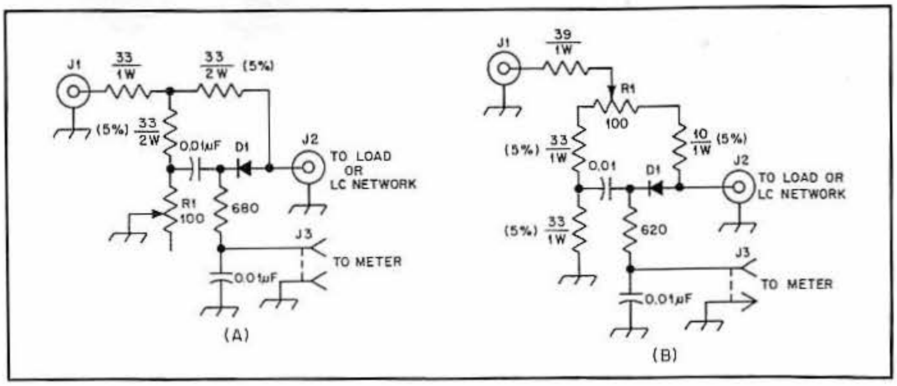

Fig. 42 — Schematic diagram and parts information for the linear (A) and nonlinear (B) resistance bridges. Resistor wattages are the minimum recommended.

D1 — Germanium diode.
J1, J2 — Coaxial connector, chassis type.
J3 — Phono jack.
R1 — 100-ohm, 2-watt linear control (Allen-Bradley, Type J).

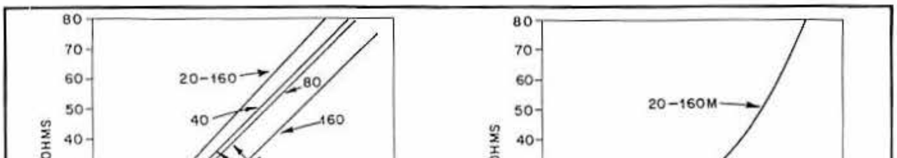

Fig. 43 — Calibration curves for the linear (A) and nonlinear (B) resistance bridges. The add-on L/C network used was described in the author's *QST* article on capacitive bridges.

ohms. In using the L/C network described in the previous article, a constant offset occurs, because of ohmic loss in the coils. This can be as high as 20 ohms, as is shown by the 160-meter curve with the L/C network, Fig. 43A.

The variable resistor used in the bridges has a parasitic (stray) capacitance of the order of 10 to 20 pF, which has little effect even at 14 MHz. Other potentiometers of similar design can probably be substituted. The calibration curves provide a good check. If the difference between the calibrations for the lowest and highest frequencies to be used is greater than an ohm or two, parasitic capacitances are probably excessive.

The nonlinear bridge uses most of the same components, but the control is connected in the differential mode, as shown in Fig. 42B. Resistance is subtracted from one arm and added to the other, tending to spread out the low end of the resistance scale for increased accuracy. In this version, the dial can be read easily to 0.5 ohm.

Results

The author has found these resistance bridges to be entirely satisfactory for measurements in the range of 1.8 to 30 MHz. Because of their greatly expanded scales, as compared with the capacitive bridges, and generally lower values of resistance in all arms of the bridge, they tend to have less marked nulls when used with very low-power signal sources. Increasing the minimum power to 20 or 30 milliwatts, when using the sensitive meter described previously, should take care of this.

In closing, the writer wishes to thank Roger A. Sykes, W1PP, for bringing the resistive bridge to the author's attention. This material was originally presented in *QST* by Jerry Sevick, W2FMI.

Why a Transmatch?

Let's cut away some of the mumbo jumbo about antennas. Here's some basic information to help the newcomer make the decision that best suits his need.

These days, nearly all transmitters have a pi-network tank circuit in the final amplifier. There are good reasons for this. With the pi tank it is possible to design a band-switching transmitter with excellent shielding (in order to prevent undesired harmonic radiation as far as TVI is concerned) and to accomplish the job with a minimum of complications. Additionally, the tank circuit easily can be set up to work into 50- or 70-ohm loads, the characteristic impedance of the popular types of coaxial feed lines. However, as we will see, this is where a "clinker" can get into the act.

When a pi network is designed to work into a 50-ohm load, the load *must* be 50 ohms in order for the transmitter to work at full efficiency. Merely connecting 50-ohm coax to the tank circuit does *not* accomplish this. For the tank to "see" a 50-ohm load, the far end of the coax — the load or antenna end — must also have the same impedance as the coax, 50 ohms. If the load is not 50 ohms the pi network will see some other value than 50 ohms. Usually, too, the tank will see reactance as well as resistance when the line is not matched. While reactance is expressed in ohms, it isn't a true resistance, and what it does is make it more difficult to put power into the line unless certain steps are taken to compensate for it.

Pi-Network Tank Circuits

In the conventional transmitter pi tank we have two controls, customarily referred to as the "tuning" and "loading" controls. The tuning control is the one that resonates the tank circuit to the desired frequency. The loading control, when adjusted properly, permits the final amplifier stage to be loaded to the desired input. To a limited extent it can also be used to compensate for the reactance in the load.

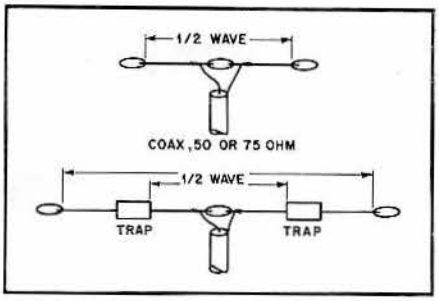

Fig. 44 — At the top is a simple half-wave dipole using coax feed. At the bottom is a drawing of a trap dipole. When a signal is fed to this antenna, for example an 80-meter one, the feed line "sees" a more-or-less resonant antenna on 80. When the transmitter is switched to 40, the traps "divorce" the 80-meter portion and the feed line sees a resonant 40-meter half-wave.

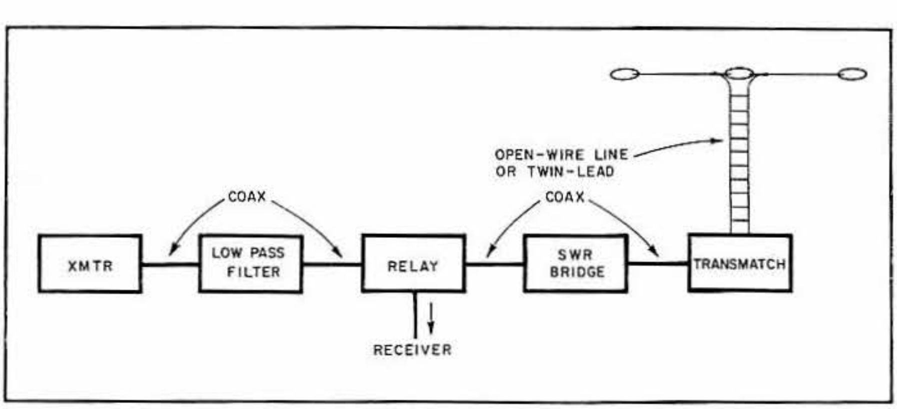

Fig. 45 — This drawing shows an ideal multiband antenna system. The units can be connected together by either 50- or 70-ohm coaxial line. The SWR bridge must match the line used. By putting the antenna relay in this section of the line, the Transmatch can be used to advantage for receiving as well as transmitting.

The clinker in this is that the reactance of many antenna-system loads may have too wide a range for the loading capacitor to handle. If you've ever been in the situation where is was impossible to load (or unload) the amplifier stage, the likely reason is that the load is too reactive to be handled by the tank circuit.

Such a condition can lead to some serious difficulties. While a correctly tuned pi network will offer good harmonic attenuation, an improperly adjusted tank may provide *no* harmonic attenuation. In fact, many Novices who receive second-harmonic violation notices for their 80-meter operation can lay the blame on poor loads and incorrectly adjusted tank circuits. This leads us to the real crux of the problem. How do we get a 50-ohm load on the various bands and frequencies desired?

Antenna Loads

As we said earlier, in order for a transmitter to see a 50-ohm load the far end of the 50-ohm coax must also see 50 ohms. The point where you feed the antenna has a certain impedance; in a half-wave dipole, for example, the impedance will be in the neighborhood of 50 ohms, depending somewhat on the dipole's height above ground. Also, at resonance there will be no reactance in the impedance. When we change frequency, the antenna is no longer resonant and reactance is introduced into the impedance. In turn, reactance is present at the transmitter end of the line. Whether or not the transmitter tank circuit can compensate for it will depend on the range of the loading capacitor.

How do we go about getting an antenna that will look like 50 ohms over a wide range of frequencies within a band — and, for that matter, on several bands? It isn't exactly easy.

For example, if we cut an 80-meter half-wave dipole for the center of the band, 3750 kHz, it is possible to obtain a fairly good match for 50-ohm coax at that frequency and approximately 50 kHz each side of 3750 kHz. However, at the band edges, 3500 and 4000 kHz, the impedance will be very reactive and the mismatch between the coax and the feed point may be as high as 10 to 1. It would be difficult, if not impossible, for our pi network to handle that type of mismatch. Actually, there is no "simple" antenna, by ham standards, that can be coax fed on 80 meters and cover the entire band with a reasonable match. The 40-, 20- and 15-meter bands are narrow enough so that the pi tank will handle a coax-fed dipole under normal circumstances. However, a coax-fed dipole is essentially a one-band system so such an antenna is not the answer if we want to work "all" bands and all frequencies with a single antenna.

Trap Antennas

One type of multiband antenna with coax feed is the trap antenna, either horizontal or vertical. The horizontal variety is essentially a dipole with traps added to divorce parts of the antenna so the coax "sees" an electrically resonant dipole on the band in use. However, the trap dipole is no "broader" than a single dipole. Fig. 44 shows such a trap dipole antenna.

There are a couple of drawbacks to trap dipoles. First, a coax-fed single-band dipole is essentially a selective circuit, in that it tends to discourage undesired harmonics from being accepted and radiated. For example, a 40-meter harmonic would have a tough time in an 80-meter coax-fed dipole because the antenna and line are badly mismatched on 40. A trap system, however, does not discriminate against such undesired signals. Also, traps introduce power loss — probably not a great deal, but nevertheless loss. But if you want to use coax line and have multiband operation, the trap dipole is an answer. On the other hand, a Novice may ask, "Is there a better multiband system, one without the trap losses and harmonic problems?" The answer is yes, but before describing it let's take a look at feed lines.

Feed Lines

Three common types of feed lines are used by amateurs: coaxial, twin-lead, and open-wire. Each has a certain amount of loss, with coax the lossiest, twin-lead next, and open-wire line the most efficient. An important point to keep in mind is that the greater the mismatch, or standing-wave ratio, on the transmission line, the greater the line loss. Also, the loss in any line increases with frequency: The loss will be least at 80 meters and will become progressively greater as we approach vhf or uhf. As an example, the loss in 100 feet (30.48 m) of RG-58/U line at 144 MHz is about 6 decibels. Translating this to power, if we had a transmitter that was putting out 100 watts only about 25 watts would reach the antenna, at the end of a 100-foot (30.48 m) line, to be radiated. The remaining 75 watts would be dissipated as heat in the line! And, most important, this figure is based on the coax being *matched* at the antenna end. The mismatch or SWR will depend entirely on the impedance of the transmission line and the impedance of the antenna. If there is a mismatch, the losses increase. On 80 meters, this same coax has a loss of less than 1 dB for 100 feet (30.48 m) of line, so we could tolerate more of a mismatch.

But don't overlook the fact that if the line and antenna are mismatched it may be impossible to make the amplifier load, even if the extra loss can be tolerated. The question that then comes to mind is if we use a low-loss line so we can tolerate a high degree of mismatch, how can we make the transmitter see a 50-ohm load? The answer to this is to use a Transmatch between the transmitter and antenna.

The Transmatch

A Transmatch is simply an adjustable rf transformer that converts an unknown load (antenna side) to a desired load (transmitter side). It consists of inductances and capacitances which can be adjusted to provide a perfect 50-ohm load for the transmitter on any band or frequency.

Let's take an example to show how a Transmatch can do the job. The impedance at the center of an 80-meter dipole is

Approximate Attenuation in Decibels/100 ft. (30.48 m)

Type	Z_0	3.5	7.0	14.0	21.0	28.0	50.	144	220	420
RG-58/U	52.5	0.68	1.0	1.5	1.9	2.2	3.1	5.7	7.2	10.4
100W. Output		8	18	28	35	38	51	72	81	92
RG-8/U	52.	0.30	0.45	0.66	0.83	0.98	1.35	2.5	3.3	4.8
RG-59/U	73	0.64	0.90	1.3	1.6	1.8	2.4	4.2	5.2	7.2
RG-11/U	75	0.38	0.55	0.80	0.98	1.15	1.55	2.8	3.8	4.9
Twin-lead	300	0.18	0.28	0.41	0.52	0.60	0.85	1.55	1.9	2.8
Open-wire	—	0.03	0.05	0.07	0.08	0.1	0.13	0.25		

This chart shows the attenuation figures in decibels for common types of feed lines. Z_0 is the nominal impedance. In the case of open-wire line, the impedance will depend on the size of the conductors and their spacing. The second line of the chart indicates the amount of power that would be lost in a 100-foot (30.48 m) run of RG-58/U when properly matched. For example, on 21 MHz, the loss per 100 feet (30.48 m) is 1.9 dB. With 100 watts output from the transmitter, 1.9 dB represents a loss of 35 watts in the line. A mismatch will increase the line loss.

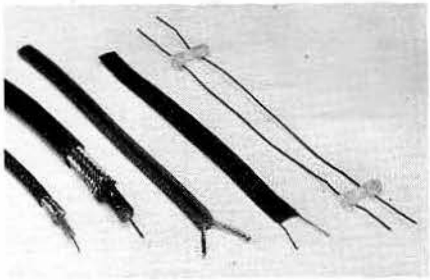

At the left are two examples of coaxial feed lines. The three lines at the right are balanced-type lines, 70-ohm, 300-ohm twin-lead, and TV open-wire line.

between 50 and 70 ohms on 80 meters. On 40 meters, the impedance at the same point on this same antenna is about 4000 ohms. If we feed this antenna with 300-ohm twin-lead the mismatch on 40 would be on the order of 13 or 14 to 1. However, bear in mind that 300-ohm line has relatively low loss, so we can probably tolerate the high SWR if the line is fairly short — less than 100 feet (30.48 m). Our only problem in making this antenna work on 40 is to take care of coupling the line to the transmitter. To do this the feed line is connected to the Transmatch and the Transmatch is then adjusted so that the transmitter sees a 50-ohm load. What we have now is an 80-meter dipole being used efficiently on 40, and our transmitter is working into the load it was designed for.

Now let's suppose that we feed this same antenna with 50-ohm coax. On 40, the mismatch would be 4000/50, or about 80 to 1! Because coax is not a low-loss line, the loss in the transmission line would be prohibitive. However, with twin-lead — or, even better, open-wire line — the loss is usually insignificant.

Also, and most important, a Transmatch adds selectivity at the output of the transmitter. It will discriminate against any undesired harmonics or spurious signals. And, if the antenna switching is set up so that the Transmatch is in the circuit on receiving, it will provide more selectivity for the receiver. In fact, a common problem for hams who live near broadcast stations is cross-modulation of 80- and 40-meter ham signals by the strong bc signal. A Transmatch will provide enough selectivity in most cases to eliminate bc cross modulation when it occurs in the receiver's front end.

Many hams are reluctant to use a Transmatch because it requires additional adjustments. However, the advantages are so great in a multiband system that any additional adjustments are more than worth the effort. With a Transmatch, the antenna doesn't have to be a half wave long to work on any band. For example, we described a complete multiband system just recently[1] using an antenna only 100 feet (30.48 m) long. Most hams can manage to get up a 60-foot (18.29-m) dipole, either horizontal or in an inverted-V configuration, and such an antenna also will work on 80 through 10. One method of putting up a multiband system is to find two supports for each end of the antenna, cut a wire long enough to run between them, put an insulator in the center of the antenna and at each end, and attach open-wire or twin-lead feeders long enough to reach the station. That's all there is to it. Bear in mind that a dipole fed this way doesn't have to be a half wave long; it can be any length, and it will work. Of course it is always a good idea to make it as long as possible and get it up as high as possible. Also, if possible, make it at least 1/4-wavelength long at its lowest operating frequency — about 60 feet (18.29 m) for 80 meters, for example.

A Few Notes

Many hams are reluctant to use open-wire line because of the problems in getting it into the shack from outdoors. One simple answer to this is to bring the line to the entrance — window sill, doorway, or whatever — and at that point attach 300-ohm twin-lead (which is insulated) and run the twin-lead in to the Transmatch. The fact that the two lines are not alike is of no concern because the load at the antenna side of the Transmatch is going to be some unknown value anyhow.

You may have read or heard that you must use coaxial lines in order to avoid TVI. Don't believe it. If you have a well-shielded transmitter and use a low-pass filter with it you won't have any harmonics coming out. Ask youself this simple question: "If there are harmonics, what difference can the kind of transmission line make?" The answer is, of course, "No difference."

One of the beauties in using a Transmatch is that the coaxial line between it and the rig can always be matched, and this matched section is where you put your low-pass filter. This eliminates any danger to the filter from excessive voltage or current caused by a mismatch.

One typical Novice question is "If my transmitter loads properly and I have no TVI problem, do I need more matching between the transmitter and antenna? I am using coax feed. What good would it do me to pay for the additional complexity of changing from band to band?" The answer to such a question depends on several factors. Is the harmonic attenuation — not TVI, but low-frequency harmonics — adequate? This could be a problem the Novice isn't aware of. As to the complexities of band changing, this will depend on how much the Novice desires to work other bands.

What we have tried to do in this article is to show the reader the problems and the recommended cure. There is nothing new about the multiband system using tuned feeders; it has been used for many years, and the interesting thing about it is that it is still the best multiband system. For further information it is recommended that the newcomer study *Understanding Amateur Radio* and *The ARRL Antenna Book* to gain a better insight as to how antennas work. This material was originally presented in *QST* by Lewis G. McCoy, WIICP.

References

[1] McCoy, "A Complete Multiband Antenna System," *QST*, November 1967.

The Ultimate Transmatch

Most hams would like to own a Transmatch that could cover 80 through 10 meters, and work during coax-to-coax or balanced-lines conditions. Here is just such a unit. It will match the proverbial bedsprings, or any other nonresonant or resonant antenna — without plug-in coils or band switching — even on 160 meters.

Some amateurs assume that because they use coaxial feed lines they don't need a Transmatch. This can be an incorrect assumption.

The Harmonic Problem

A large number of hams have found to their sorrow that the harmonic suppression in their transmitters may leave much to be desired. The FCC issues many citations to Novices each year for second-harmonic violations.

A typical Novice station might consist of a transmitter and receiver, with direct coax feed to a trap dipole or vertical antenna. In a properly adjusted transmitter of typical design, often there is insufficient attenuation of the second-harmonic energy. This can mean that the second harmonic of the signal can reach the antenna and be radiated. What is required to eliminate or reduce the second harmonic, and higher-order harmonics, is more selectivity between the transmitter and the antenna. A Transmatch is an excellent aid in taking care of the harmonic problem.

The 50-Ohm-Load Problem

Still another reason for using a Transmatch is that nearly all manufacturers of transmitters design their rigs to work into a nonreactive, 50-ohm load. Any departure from this load impedance causes difficulty in tuning and loading the final-amplifier stage. It might be pertinent to point out that simply because a ham uses 50-ohm coax, it doesn't necessarily follow that he will have a 50-ohm load. Basically, in order to have a 50-ohm load when using 50-ohm coax, the coax must be terminated in a 50-ohm resistive impedance. In other words, the antenna impedance must be 50 ohms.

When the terminating impedance of the coaxial cable is other than 50 ohms, then the transmitter will be "looking" at some value other than 50 ohms. It is practically impossible to design a single antenna that will have an impedance of 50 ohms on all of the bands a ham might want to use. One answer to the problem is to employ a device that will "match" the unknown load to the 50-ohm impedance required by the transmitter. One such circuit or device is a Transmatch, simply an adjustable rf transformer that can match the unknown load presented by the transmitter end of the feed line to the required 50-ohm transmitter impedance.

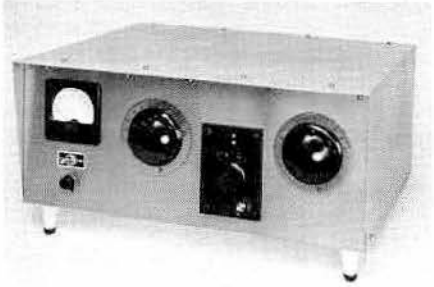

The kW version of the Ultimate Transmatch. The knobs and dial counter are standard Millen components. The front bottom of the cabinet is tilted by means of two 2-inch (51 mm) standoff insulators used as feet.

Receiver Cross-Modulation

In addition to taking care of the problems just outlined, a Transmatch can serve another very useful function at the receiver. Some receivers are susceptible to what is known as cross-modulation, or front-end overloading, by strong nearby radio stations. This is particularly true on 160 or 80 meters when a broadcast station is nearby. The bc station overloads the front end of the communications receiver, causing "birdies" and other unwanted responses across the band. A Transmatch will eliminate or greatly reduce this problem in nearly all instances.

The Ultimate Transmatch

We shall describe two Transmatches — one for the Novice power limit, and another for the legal 2-kW PEP limit. We have found that many amateurs on the verge of taking their General, or higher-class licenses, would rather invest initially in a unit that will take care of all of their future needs.

Some years ago, a Transmatch was described that was designed for use with coaxial feed lines.[1] It was called "the 50-Ohm Transmatch." This unit handles mismatches on the order of 4:1. It can transform a load of about 200 ohms, or 12 ohms, to a 50-ohm impedance. The circuit of the Transmatch described in this article, Fig. 46, is similar, but with two notable exceptions. In place of a tapped, switched inductor, a variable inductor is used. (The idea for this innovation came from Al LaPlaca, K2DDK.) The variable inductor is a coil wound on an insulated form that can be rotated so that a movable wheel shorts out and grounds the unused portion of the coil. Using the roller inductor, along with the two variable capacitors, the Transmatch provides an almost unlimited matching range, plus the feature of 80- through 10-meter coverage without band switching. With the values shown in the units, it is even possible to add one fixed capacitor to obtain 160-meter coverage.

The other exception is the use of a 1:4 (unbalanced-to-balanced) balun transformer on the output side of the Transmatch, which permits the use of balanced feed lines. This means that a Transmatch can cover 160 through 10 meters (including the MARS frequencies). It can be used to match into coaxial lines, random-length end-fed wires, or balanced feeders, and can do the job without using switches.

The two units shown in the photographs underwent extensive tests as to efficiency, power-handling capabilities, and methods of adjustment. In coax-to-coax tests, power inputs and outputs through the Transmatch were carefully measured while using Bird Thruline wattmeters and an adjustable dummy load. As carefully as could be measured, and with indicated mismatches of up to 15:1, the Transmatches showed an efficiency of over 95 percent. This small loss in efficiency more than outweighs the usefulness of the device.

Many tests were made with the high-power version to check heating of the roller inductor and the balun. Under a key-down condition, with 1200 watts indicated power through the Transmatch, the transmitter was kept on for 15-minute periods. In one test, the measured rf current through the coil was in excess of 15 amperes. The roller inductor used in the unit is rated at 5 amperes, so this was considered a fairly rugged test. The coil did get warm, but not enough to be a cause for concern. Keep in mind that in regular amateur work, 15

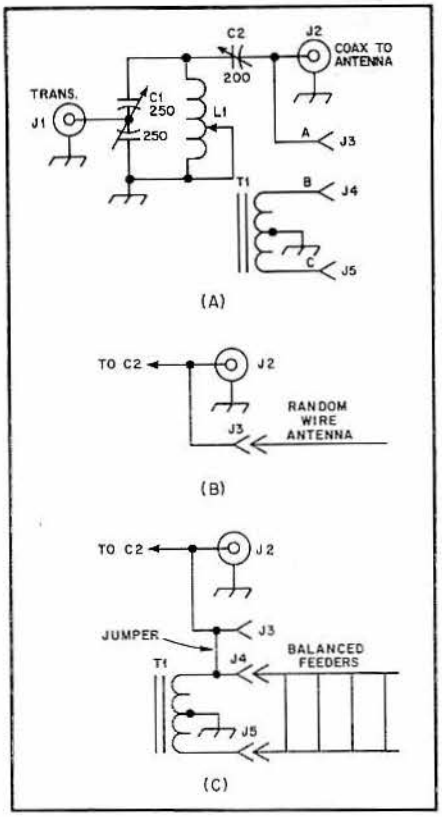

Fig. 46 — Circuit diagram of the Ultimate Transmatch.
C1 — Split-stator variable, 250 pF per section; see text. For low-power version, E. F. Johnson type 16250; or similar, for high-power version, Millen type 16250, or similar.
C2 — 200-pF variable, for low power, E. F. Johnson type 167-12 or similar, for high power, Millen type 16520 0.171-inch (0.43 cm) spacing, 16520A, 0.077-inch (2 mm) spacing, or similar.
J1, J2 — Coax chassis fitting, type SO-239.
J3, J4, J5 — Feed through insulators.
L1 — Roller inductor, see text. If 160-meter operation is desired, total inductance should be 28 μH, E. F. Johnson type 229-203; otherwise, 18 μH is adequate, E. F. Johnson type 229-202.
T1 — 1-to-4 balun; see text for details, cores are Amidon type T-200-2.

minutes would be a ridiculous figure for continuous operation.[2] In cw and ssb operation, the duty cycle would be in terms of seconds or even fractions of seconds.

The balun used for balanced operation in the high-power unit was made up from three Amidon 1-kW cores — more about that in a moment. N1RM has made extensive checks on a single-core balun. It was found that a single core would easily handle 500 watts of power during severe mismatch conditions. Failure of the core, most likely due to saturation and heating, took place at the 700-watt level. In the high-power Transmatch shown here, three cores are used. They are wrapped with glass insulating tape and then wound with Teflon covered wire. Wire insulation material can be important because of the extremely high rf voltages that may be developed with some load conditions. In many tests of long duration, using up to 1400 watts output into a dummy load, it was almost impossible to detect any heating of the balun.

Bear in mind that the balun would only be needed if balanced-line operation were desired. For coax-to-coax or random-wire feed, the balun can be omitted.

Construction Information

A look at Fig. 46, at A, will show that the circuit is a very simple one. The input line from the transmitter is connected to the rotor of C1, a split-stator capacitor. This means that the capacitor must be insulated from the chassis, using standoff pillars, and from the panel by means of an insulated shaft coupler. The same is true for C2. In the kW unit shown, we have installed a power bridge. Construction details are not given here for the bridge because such a unit was described in detail in *QST*[3]. Also, many amateurs already have a Monimatch[4] or some other type of SWR indicator. An SWR bridge or matching indicator is needed in the 50-ohm line between the rig and the Transmatch to show when the Transmatch is correctly adjusted.

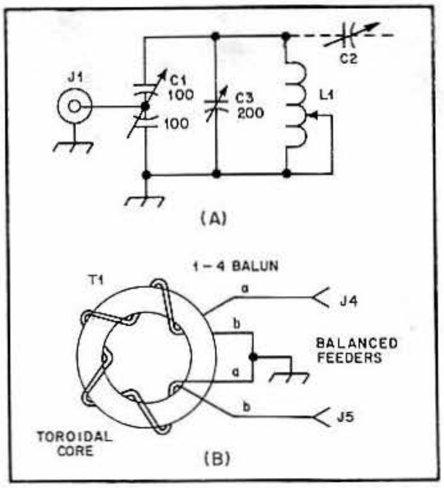

Fig. 47 — At A, details for padding a low-value variable for C1 by installing a second variable, C3, the same value as the low-power unit C2. In the case of tuning 160 meters, using an inductor of 28 μH for L1, and 250 pF per section for C1, a fixed capacitor of 100 pF can be used at C3, Centralab type 850-100N, or similar. At B, details for connecting the bifilar windings on the balun are given. The ungrounded leads a and b are connected to J3 and J4, the feedthrough insulator terminals.

Layout of the components should follow the same arrangement shown in the photographs — C1 at the input side, the roller inductor at the center, and then C2. If a balun is going to be installed, allow enough room behind the roller coil for mounting the balun. In the two units shown, the cabinets are made up from aluminum sheet stock. If desired, a commercial chassis and cabinet can be used. In the low-power Transmatch, leave at least 1/2-inch (13 mm) spacing between the edges of the capacitors and any part of the chassis and cabinet. Allow 1-inch (25 mm) spacing in the kW unit.

Balun Details

For the low-power Transmatch, a single toroid core will more than handle the power. Incidentally, the low-power unit was tested under all conditions at about 100 watts through the Transmatch. The

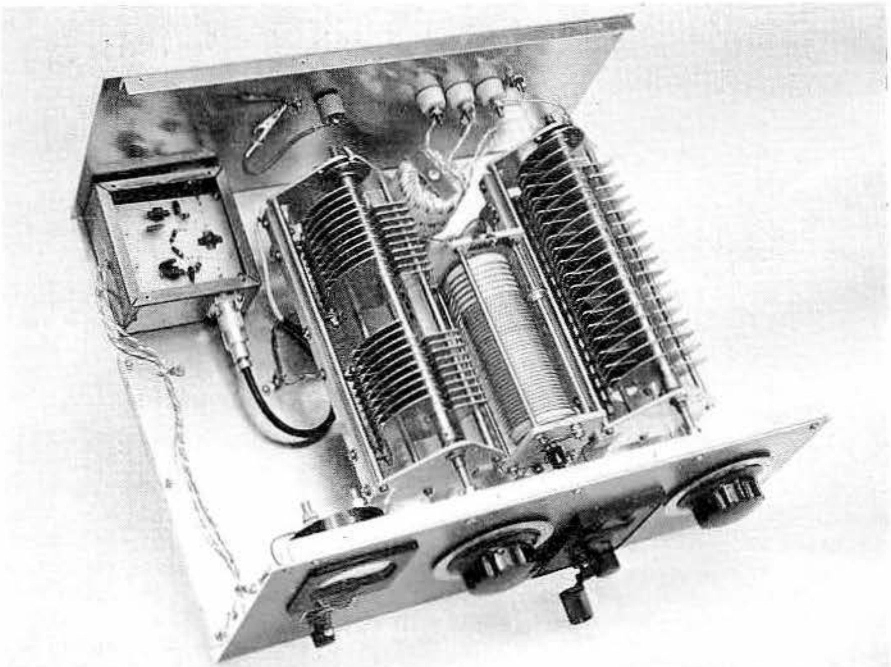

At the rear of the roller inductor is the 1:4 balun. The balun is set on top of a standoff insulator and is held in place with a piece of insulated board and a screw. The fixed capacitor mounted on the rear wall is the 160-meter unit. Both the variable capacitors and inductor are mounted on 1-inch (25 mm) high standoff insulators.

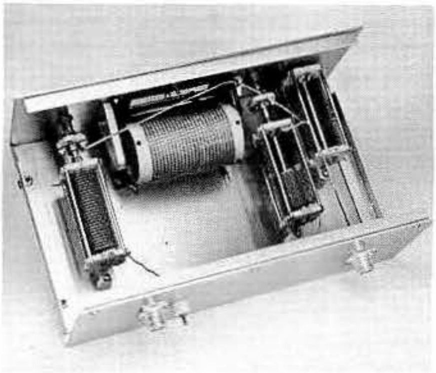

This is the low-power Transmatch. The salvaged roller inductor is mounted on the front panel, making use of the original mounting brackets.

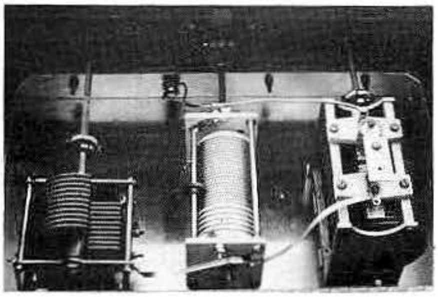

Here is an inside view of the Ultimate Transmatch as built by Al LaPlaca, K2DDK, who came up the idea of using the roller inductor. The installation on top the right-hand capacitor is the homemade 160-meter switch and fixed loading capacitor.

power-handling capabilities of the units depend primarily on the plate spacing used for C1 and C2, aside from the balun requirements mentioned earlier. At the 100-watt level, a plate spacing of 0.03 inch (0.8 mm) is adequate. A 0.045-inch (1 mm) spacing should handle 500 watts. In the kW unit shown, C1 has a 0.077-inch (2 mm) spacing. While a 0.077-inch spacing capacitor wasn't tried at C2, such plate spacing should work at the kW level.

In making the low-power balun, wind two layers of Scotch brand electrical tape around the core before putting on the wire. This will provide extra insulation, and will protect the edges of the wire to prevent the enamel from being scraped away. Fig. 47, at B, shows a sketch of the winding. The winding consists of 10 bifilar turns of no. 14 Formvar insulated wire. Connected as shown, this will provide a 1:4, unbalanced-to-balanced configuration.

Three cores are needed in the high-power balun. Each core is covered with two layers of 3-M type 27 glass-cloth insulating tape. Next, the three cores are stacked, then covered with two more layers of the tape. The winding consists of 15 bifilar turns of either no. 12 or 14 insulated wire. The insulation should have a minimum rating of 1000 volts and can be either Teflon or vinyl-nylon (or any other suitable covering with a minimum voltage rating of 1000 volts).

In making the low-power Transmatch, a roller coil from a junked 40-meter ARC-5 transmitter was salvaged and installed in the unit. Many amateurs know other hams who have old ARC-5 gear, and these units are a good source for the roller coils. However, some hams may not be able to locate used ARC-5 transmitters, so we have included some information on sources of surplus and new roller inductors. There may be surplus outlets in your area worth investigating. Look for a roller coil that has a minimum inductance of 10 μH when used with C1 having a value of 200 pF per section. Such a combination will cover a low frequency of 3500 kHz. In the unit shown, in addition to the ARC-5 roller, we used a value of 100 pF per section for C1 and padded that capacitor with another of 200 pF (single section) in order to provide the required coverage. This added capacitor is shown in Fig. 47 at A. By using this approach, it is sometimes easier to save money then when trying to locate a 200-pF-per-section variable for C1. (Incidentally, you can use a variable at C1 of more than 200 pF per section for good coverage with the roller inductance specified above.)

An even better approach to get the required capacitance for C1, because it is better to maintain the symmetry of the circuit as shown in Fig. 46 at A, would be to gang two 200-pF variables to make up your own split-stator capacitor, with the desired plate spacing. (There may be some around, but we couldn't find any manufactured split-stator with the required plate spacing and capacitance). Still another approach would be to clip fixed transmitting-type mica capacitors across each section of C1 to obtain the required capacitance for use on 160 and 80.

Finding the Parts

As most hams are finding out, it is becoming more and more difficult to buy, or find, components for construction projects. Some searching was done, and the following information is worth passing on. Barry Electronics of New York City has informed us that they will fill any order for any amateur component at the manufacturer's current amateur net price. The parts specified in Fig. 46 are currently manufactured items. Also, Fair Radio Sales in Lima, OH has some surplus roller inductors and counter dials in stock. The roller-inductor catalog number is 5950-199-6469, and the counter dial is listed as RT-45. Also, Millen components can be purchased directly from G. R. Whitehouse, Amherst, NH 03031.

Using the Transmatch

One of the first considerations in using a setup for "coax-to-coax" is exactly how much, or how high an SWR can be tolerated. As has been pointed out, this Transmatch will handle *any* mismatch on the line. Without getting into a lot of details, a few observations about coaxial lines are in order. First, the higher the SWR, the greater the losses in coaxial lines. The amount of losses depends on the line construction and the *frequency* at which the line is used — the lower the frequency, the less will be the losses. For example, RG-8/U has a loss per 100 feet (30.48 m) of line of only 0.33 -dB at 3.5 MHz with an SWR of 1 (a matched condition). At 28 MHz, this same line would have a loss of slightly less than 1 dB. However, as the SWR rises, so do the losses. With an SWR of 10:1, for example, the additional losses at 3.5 MHz would amount to about 0.6 dB, or an overall loss of about 1 dB. For the same SWR at 28 MHz, the losses would be about 2.5 dB, or a total loss of about 3.5 dB. Since a loss of 3 dB would mean half of the ouput power from the rig would be dissipated in 100 feet (30.48 m) of line, one could easily tolerate the condition at 80 meters, but the losses would be prohibitive at 28 MHz. This means that the amateur would be wise to keep the match at the antenna feed point as close to 50 ohms as possible, and thereby keep the SWR low.

With trap or multiple dipoles, the SWR should be less than 5 to 1; the Transmatch would be ideal. Keep in mind that you must provide a load for a transceiver or amplifier that is close to 50 ohms. The Transmatch will provide a means.

In a coax-to-coax system, set both C1 and C2 at maximum capacitance (plates fully meshed), then feed enough power through the Transmatch to get a full-scale forward reading on your SWR indicator. It is always important to use low power when adjusting a Transmatch. Next, switch the SWR bridge to read reflected power, then run the roller inductor slowly through its range. At some point you will notice a sharp dip in the reflected-power reading. Once this point is reached you should adjust C1 and C2, with possibly a slight adjustment of the roller coil, to get a perfect match. Power can then be increased to the desired level. You'll find that several different settings of the Transmatch may provide a perfect match. The one to shoot for is the condition that requires the most capacitance at C1 and C2 for a matched condition.

If a random-length, end-fed antenna is used, merely connect the end of the antenna to terminal J3, Fig. 46B, and go through the adjustment procedures outlined above. It is helpful to have an earth ground connection on the Transmatch. Also, if possible, use an antenna length that comes out to a quarter wavelength, or odd multiple thereof. Such an antenna will be of low impedance, and will reduce the chances of high rf voltages appearing on the Transmatch or associated equipment.

If balanced feeders are used, the feeders should be connected to terminals J4 and J5 with a short jumper connected between terminals J3 and J4. Again, the previously discussed tune-up and adjustment procedures should be followed in setting up the Transmatch.

Some Other Thoughts

Some amateurs might be concerned about the amount of voltage that can be developed in a mismatched coaxial line. In a matched condition, with 1000 watts in a 50-ohm line, the rf voltage would be about 220 volts. With a standing-wave ratio or 20:1 (!), the increase in voltage would be 4-1/2 times, or approximately 990 volts. This however, is well within the rating of the RG-8/U cable.

This Transmatch should satisfy the needs of those hams who desire a multi-band unit without switches or plug-in coils. Once the settings for the controls are determined, a chart can be made up for any given istallation. Band changing is then quickly done. This material was originally presented in *QST* by Lewis G. McCoy, W1ICP.

References

[1] "The 50-Ohmer Transmatch," *QST*, July 1961.

[2] [Editor's Note: However, teletype operation would be continuous (to say nothing of those long-winded a-m operators!).]

[3] "In-Line RF Power Metering," *QST*, December 1969. (See chapter 4.)

[4] *Understanding Amateur Radio*, p. 209.

Build a Baby Ultimate

Got a match? Not a classic "match," but one between the antenna and your QRP rig may be what you need. If so, this article is for you.

Nobody needs or wants a 50-pound Transmatch during a QRP junket to a favorite lake, campsite or other distant location. The entire point of QRP operation is negated when the low-power station can be held on the palm of one's hand, but the accessory gear needs to be transported by hand truck to the car, airport or whatever! Operators who prefer to use a random-length wire antenna for field work will require a matching network between the 50-ohm transmitter output and the end of the antenna, whose characteristic impedance can be anything from a few ohms to a couple of thousand. For low-power work (25 watts or less) it doesn't take a big Transmatch to do the job, and some novel ideas can be applied to make the matcher cover 80 through 10 meters without a roller inductor or messy tapped-coil switching arrangement. This article shows some tricks that can be applied to meet that goal.

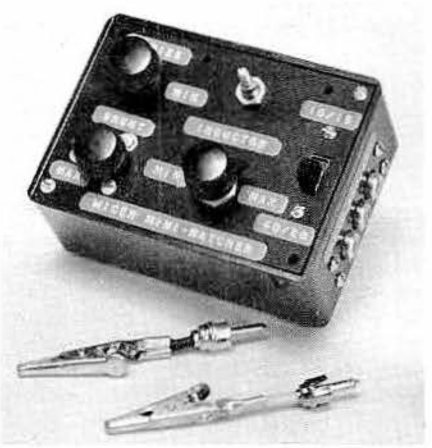

Exterior view of the Baby Ultimate. Dymo tape labels are used to identify the controls.

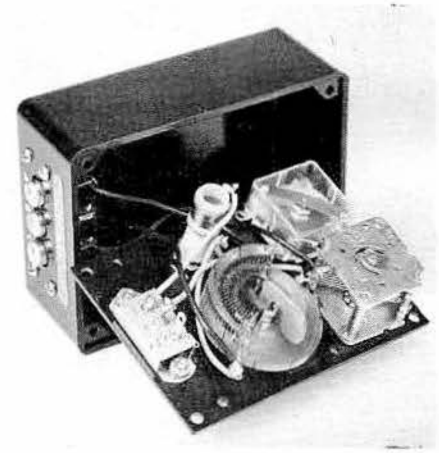

Interior view of the matcher showing the K1KLO variable inductor (lower center).

The heart of the Baby Transmatch is the half-toroid variable inductor. It is the by-product of a few "skull sessions" between Andy Pfieffer, K1KLO, and the writer. Both amateurs are QRP "freaks," and finding the proper mechanical technique for building a miniature panel-operated variable inductor of reasonable inductance range was challenging and productive. The machine-shop skill and impagination of K1KLO can be realized when viewing the accompanying close-up photograph. Early efforts along this line provided a slider type of full-toroid variable inductor (rheostat style), and dismal results were had: the inductor Q was very low, approximately 5, as a result of the shorted-turn syndrome. Finally, a powdered-iron Toroid core of Q1 material, 1-inch (25 mm) diameter, with 3/16 by 3/16-inch (5mm by 5mm) walls, was sawed in half and cemented to a plastic arm. The arm was glued to the shaft of the control so that it would enable the operator to change the coil inductance by moving the semicircular toroid section in and out of the coil.

The coil in this unit was air wound, then cemented to a semicircular piece of plastic which is affixed to a plastic disk of 1-1/2-inch (38 mm) diameter. The overall coil-and-plastic assembly is glued to the 3/8-inch (10 mm) bushing and 1/4-inch (6 mm) diameter control shaft. Although all of the metal work was done from scratch on a lathe, an old volume control or rheostat bushing and shaft can be modified to perform the function required. All that's needed here is some amateur ingenuity! The tuning range of the variable inductor is 3 to 9 μH, and the unloaded Q was measured as 150 at 7.9 MHz.

It is reasonable to conclude that the K1KLO "giz," as he calls it, could be made to cover a greater inductance range by making the coil bigger in ID and cementing both halves of the toroid core together — one atop the other — to provide greater permeability overall. Ferrite material is unsuitable for this component because of the hardness of ferrite. Without special cutting tools the latter will resist all efforts to saw through it. One-half of an Amidon 1-kW balun-kit, powdered-iron toroid should be excellent for larger units of this kind.

Other Circuit Features

As is shown in Fig. 48A, S1 is used to switch an additional variable inductor (L1) in parallel with the K1KLO unit, L2. This provides a lower minimum-inductance amount for operaion on 15 and 10 meters.

Fig. 48B illustrates some variations which will extend the Transmatch range down to 80 meters. The model shown photographically was designed for 40 through 10 meters. In the second example, S2 and L3 have been added to increase the total inductance to permit 80-meter matching. Also, C1 has been changed to a

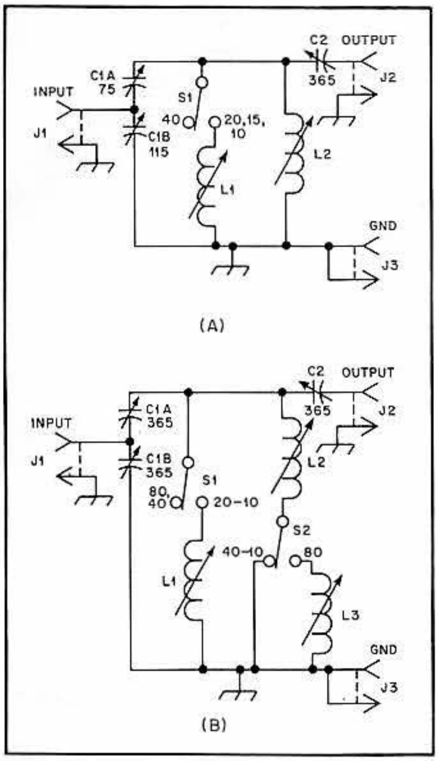

Fig. 48 — The diagram at A is for the 40- through 10-meter Transmatch. At B, a suggested circuit for coverage from 80 through 10 meters.

C1 — Dual-section air variable (Miller 2109, J. W. Miller Co., 19070 Reyes Ave., Compton, CA 90224). See text.
C2 — Calectro or Archer single-section miniature 365-pF variable.
J1-J3, incl. — Phono jack.
L1 — 3.1-to 4.8-μH slug-tuned inductor (Miller 4504 with red core).
L2 — See text. Contains 32 turns of no. 22 enam. wire, air wound, 7/16-inch (11 mm) OD.
L3 — 5.5-to 8.6-μH slug-tuned inductor (Miller 4505 with red core.)
S1, S2 — Spdt slide or toggle switch.

two-section 365-pF, bc-style variable, also for the same reason. A dual 365-pF capacitor can be used for the circuit at A, but the tuning rate will be much sharper than with the unit specified.

Three phono jacks are used for external connections. The first photograph shows a pair of clip leads which have been soldered to two phono plugs. These are used for making connections to a single-wire antenna and ground.

The slug screw of L1 protrudes from the top panel of the small plastic meter case. This feature enables the operator to have additional control of the overall inductance amount during operation on 15 or 10 meters. All internal ground connections are brought to a common point — J3, which has both of its terminals joined.

The "Ultimate" design is based on one which was popularized by W1ICP when he described in *QST* a 1-kW version of the circuit.[1] The network will match a wide range of impedances, but it should be pointed out that this circuit is not capable of providing much (if any) harmonic attenuation. Examination of the circuit will show that the network is basically a high-pass type.

Building the Transmatch

A 1-1/2 × 2-3/4 × 4-inch (38 mm × 70 mm × 102 mm) plastic meter style of case is used to house the Transmatch. The box used by the writer was made by RCA as an enclosure for a 9-volt regulated power-supply kit. Similar boxes are available at Radio Shack and other parts stores, but most have a metal cover. If the latter are used, it will be necessary for the builder to isolate the rotors of C1 and C2 from the metal plate. Adhesive-backed plastic feet are affixed to the bottom of the case. They will keep it from moving about on the operating table.

Using the Transmatch

When matching the 50-ohm transmitter to a given antenna (fed with coaxial line or single wire in nature), an SWR indicator will be necessary between the transmitter and the Transmatch.[2] C1, C2 and L2 are adjusted one at a time for the lowest reflected-power reading on the meter. There will be interaction between the controls, so it will be necessary to go over them several times to effect an SWR of 1.

Some difficulty may be encountered while working with a voltage-fed (high impedance) wire antenna. The symptom is one of hand capacitance affecting the SWR reading. In such instances the tune-and-try method will be required, moving the hands away from the box after each adjustment. For this reason the builder may wish to house the circuit in a metal box and use insulating shaft couplings on C1, C2 and L2. Of course, this will make the unit somewhat larger.

The writer used this Transmatch for two weeks in October 1975 while on a QRP sortie to Barbados (8P6EU). Maximum power used was 20 watts output, and minimum power was 2 watts ouput. The 40-, 20- and 15-meter bands were used, and all manner of antennas (end fed and coaxially fed) were employed. In all cases the SWR could be set for a 1:1 condition. Insertion loss through the Transmatch was measured at 14 MHz while delivering 20 watts to a test dummy load. The loss was approximately 0.1 dB at a matched condition. No heating of the components was noted.

One final word is in order. Always use the maximum possible amount of capacitance at C2 when adjusting the Transmatch for an SWR of 1. There can be several settings of C1, C2 and L2 which will provide a match. Best efficiency will always occur with maximum usable C at C2, and this is true of all variations of the Ultimate Transmatch.

If you're a QRP enthusiast, this matcher should be part of your setup when 50-ohm antennas aren't used. It is lightweight, small of size, and can be tucked away in the XYL's handbag when enroute, if need be! This material was originally presented in *QST* by Doug DeMaw, W1FB.

References

[1] McCoy, "The Ultimate Transmatch," *QST*, July 1970. (Included in this chapter.)

[2] DeMaw, "A QRP Man's RF Power Meter," *QST*, June 1973, p. 13.

Simple Broadband Matching Networks

This article gives complete details on constructing broadband matching transformers for ratios of 4:1, 6:1, 9:1 and 16:1.

Are you looking for a physically small DX antenna? To obtain low-angle radiation on 80 and 160 meters, many amateurs use vertical antennas. Horizontal antennas would have to be at elevations of a half-wavelength or greater to be comparable. Suitable supporting structures are beyond the reach of most amateurs at these lower-frequency bands. Quarter-wavelength verticals require some 60 feet (18.29 m) on 80 meters and 120 feet (36.58 m) on 160 meters — still very tall structures! However, short verticals over a good image plane have been shown theoretically[1] and experimentally[2] to perform practically as well as their full-size quarter-wavelength counterparts. The only problems are very low input impedances and narrower bandwidths.

This paper describes broadband matching networks which can efficiently match 50-ohm coaxial cable to verticals as short as 1/25 wavelength. Also included is information on short vertical antenna designs incorporating large top hats in order to improve bandwidth, such as a very short 40-meter vertical and a dual-band 80- and 160-meter short vertical with remote switching. In addition, loading coils using powdered-iron toroidal cores are discussed.

Broadband Matching Transformers

Broadband transformers have been widely used because of their inherent bandwidth ratios (as high as 20,000:1) from a few tens of kilohertz to over a thousand megahertz. This is possible because of the transmission-line nature of the windings. The interwinding capacitance is a component of the characteristic impedance and therefore, unlike the conventional transformer, forms no resonances which seriously limit the bandwidth.[3] At low frequencies, where interwinding capacitances can be neglected, these transformers are similar in operation to the conventional transformer. The main difference (and a very important one from a power standpoint) is that the windings tend to cancel out the induced flux in the core. Thus, high-permeability ferrite cores, which are not only highly nonlinear but also suffer serious damage even at flux levels as low as 200 to 500 gauss, can be used. This greatly extends the low-frequency range of performance. Since higher permeability also permits fewer turns at the lower frequencies, high-frequency performance is also improved since the upper cutoff is determined mainly from transmission-line considerations. At the high-frequency cutoff, the effect of the core is negligible.

Connie, W2FMI's better half, is making use of the shade under Jerry's 40-meter umbrella antenna. It is hoped that Jerry warns her about touching the antenna when he is running high power!

The applications of broadband transformers to antennas and broadband circuits have been covered by a number of authors.[4,5] This paper extends the application of these transformers to include the operation of short ground-mounted vertical antennas with impedances as low as 3 to 4 ohms. Further design considerations illustrate variable matches to vertical antennas having impedances even greater than 35 ohms.

Bifilar matching transformers for ground-mounted verticals lend themselves to unbalanced operation. That is, both input and output terminals can have a common ground connection. This eliminates the third magnetizing winding required in balanced-to-unbalanced (balun) operation. By adding third and fourth windings, as well as tapping windings at appropriate points, various combinations of broadband matching can be obtained. Fig. 49 shows a 4:1 unbalanced-to-balanced configuration. This transformer can be used with verticals of 12.5 or 200 ohms when using 50-ohm cable. No. 14 wire can be used and it will easily handle 1000 watts of power. By tapping at points 1/4, 1/2 and 3/4 of the way along the top winding, ratios of approximately 1.5:1, 2:1, and 3:1 can also be obtained. It should be noted that one of the wires was covered with vinyl electrical tape in order to prevent voltage breakdown between the windings. This was necessary when a step-up ratio was used at high power to match antennas with impedances greater than 50 ohms.

Fig. 50 shows a transformer with four windings, permitting wide-band matching ratios as high as 16:1. A picture shows a four-winding transformer with taps at 4:1, 6:1, 9:1 and 16:1. In tracing the current flow in the windings when using the 16:1 tap, one sees that the top three windings carry the same current. The bottom winding, in order to maintain the proper potentials, sustains a current three times greater. The bottom current cancels out the core flux due to the other three windings. If this transformer is used to match into low impedances, like 3 to 4 ohms, the current in the bottom winding can be as high as 15 amperes if the high side of the transformer is fed with 50-ohm cable handling a kilowatt of power. If one needs a 16:1 match like this at high power, then cascading two 4:1 transformers is recommended. In this case, the transformer at the lowest impedance side only requires each winding to handle 7.5 A. Thus even no. 14 wire would suffice in this application.

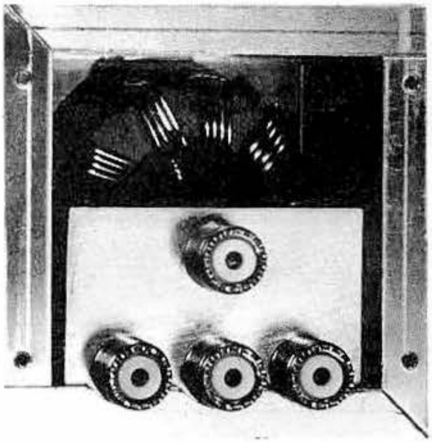

A 4-winding, wideband transformer (with front cover removed) with connections made for matching ratios of 4:1, 6:1, 9:1, and 16:1. The 6:1 ratio is the top coaxial connector and, from left to right, 16:1, 9:1, and 4:1 are the others. There are 10 (quadrifilar) turns of no. 14 enameled wire on a Q1, 2.5 inch (64 mm) ferrite core.

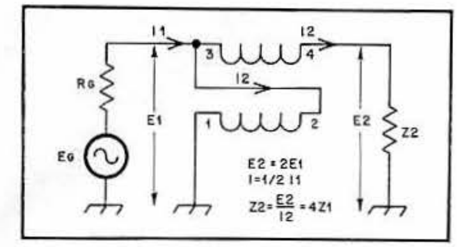

Fig. 49 — Four-to-one broadband bifilar transformer. Upper winding can be tapped at appropriate points to obtain other ratios such as 1.5:1, 2:1 and 3:1.

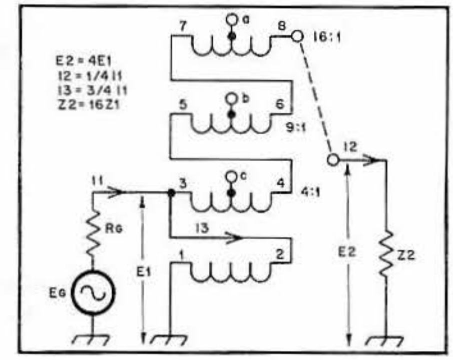

Fig. 50 — Four-winding, broadband, variable impedance transformer. Connections a, b and c can be placed at appropriate points to yield various ratios from 1.5:1 to 16:1.

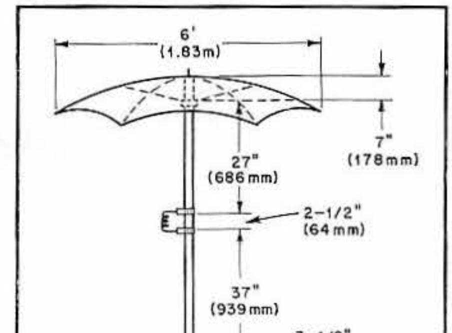

Fig. 51 — Final design and bandwidth characteristic of a short 40-meter vertical having an input impedance of 3.5 ohms and utilizing a commercial beach umbrella. This antenna established its worth as a prototype for the lower frequencies.

The popular cores used in these applications are 2.5 inch (64 mm) OD ferrites[6] of Q1 and Q2 material and powdered iron cores,[7] of 2-inch (51 mm) OD. The permeabilities of these cores, μ, are nominally 125, 40 and 10, respectively. Powdered iron cores of permeabilities 8 and 25 are also available.

In all cases these cores can be made to operate over the 10- to 160-meter bands with full power capability and very low loss. The main difference in their design is that lower permeability cores require more turns at the lower frequencies. For example, Q1 material required 10 turns to cover the 160-meter band, Q2 required 12 turns, and powdered iron (μ = 10) required 14 turns. Since the more common powdered-iron core is generally smaller in diameter and requires more turns because of lower permeability, higher ratios are sometimes difficult to obtain because of physical limitations. Although most of the author's experimentation on these transformers was done on the 40-, 80- and 160-meter bands, the results indicated that they should operate on the higher bands equally well. When working with low impedance levels, unwanted parasitic inductances come into play, particularly on 14 MHz and above. In this case lead lengths should be kept to a minimum.

All the results obtained by the author have shown that the transformers can be used to match a variety of loads and impedance ratios, they can be used over many bands, and that they exhibit negligible loss in operation. More importantly, the core materials are readily available and low in cost.

A Short Vertical Antenna Design

These transformers permit efficient matching to impedances as low as 3 to 4 ohms. This means, with some top-hat loading, antennas could be 4 to 5 feet (1.22 to 1.52 m) tall on 40, 8 to 10 feet (2.44 to 3.05 m) on 80, and 16 to 20 feet (4.88 to 6.10 m) on 160 meters. With the capability of tapping off any portion of the transformer windings, thus providing variable matching, the heights of these antennas can be increased at will depending upon the bandwidth desired.

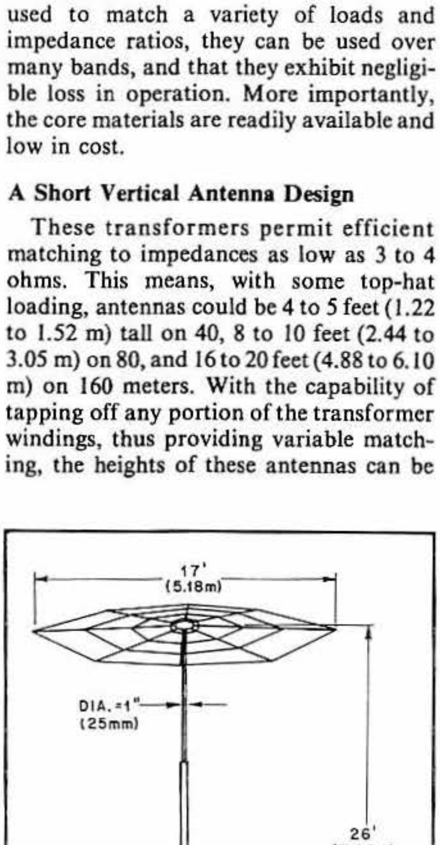

Fig. 52 — The 80- and 160-meter dual-band vertical system. The remotely switched matching network is described in the text. Not shown are three sets of polypropylene guy ropes at heights of about 9, 18 and 26 feet (2.74, 5.49 and 7.92 m).

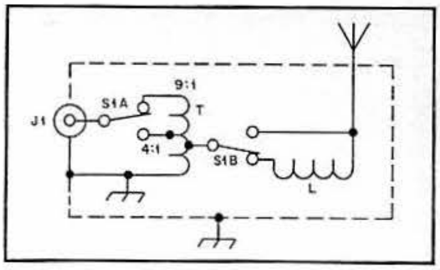

Fig. 53 — The remotely switched matching network for the 80- and 160-meter dual-band vertical antenna. The indicated switch positions are for 160-meter operation.
S1 — Dpdt antenna relay.
J1 — Coaxial connector, chassis type.
L — Load coil for 160 meters, 22-1/2 turns, no. 14 wire on 4 cores of 2-inch (51 mm) OD powdered-iron, permeability of 10 (Amidon T-200-2).
T — Broadband transformer (see text).

The first short vertical designed by the author was brought about by a trip to a local shopping center. The author saw a beach unbrella that had about the right proportions one would use in designing a short 40-meter vertical. The "antenna" was purchased (at a nominal price) and a 14-turn coil, connected across an insulated section about three feet (0.91 m) from the ground, was used to tune the vertical to resonance. The measured input impedance was 3.5 ohms and was matched easily to 50-Ω line with a 16:1 transformer. Fig. 51 shows the final design and the bandwidth characteristic. Several months of operation on 40 meters verified the outstanding capability of this short vertical antenna (with an extensive radial system) and established its worth as a prototype for the lower frequencies. A further study showed that the 6-foot (1.83 m) diameter top-section was about the equivalent of a 4-1/2-foot (1.37 m) diameter solid disk.

A Dual-Band 160- and 80-Meter System

The antenna shown in Fig. 52 is a dual-band design using a remotely switched matching network (see photograph) which not only changed a tap on a broadband transformer from 4:1 to 9:1, but also introduced a powdered-iron base-loading coil for 160-meter operation.

The circuit diagram for the remote matching network is shown in Fig. 53. The loading coil was made with four powdered-iron (μ = 10, 2-inch [51 mm] OD) cores. They were first wrapped individually with several layers of vinyl electrical tape before being combined and wrapped again. The necessary inductance was obtained with 22-1/2 turns. A slight adjustment of the loading coil and the height of the antenna was found necessary in order to put the minimum SWR at 1812 kHz. In general, a 1-foot (0.31 m) change in height of the antenna changes the 80-meter resonance by 60 kHz and the 160-meter resonance by 15 kHz. A change of 1/2 turn on the

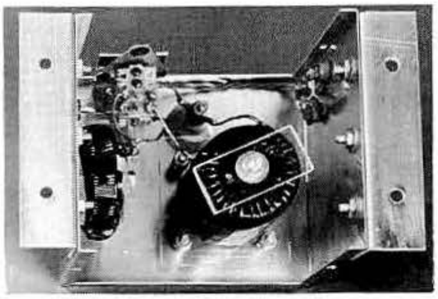

Bottom view of the matching network for the dual-band 80- and 160-meter vertical. The loading coil is on the right and the trifilar 4:1 and 9:1 matching transformer on the left. The dpdt 24-volt antenna relay switches in or out the loading coil and changes the taps on the transformer to the appropriate ratio. The enclosure is the conventional 4 X 5 X 6-inch (102 mm X 127 mm X 152 mm) aluminum Minibox.

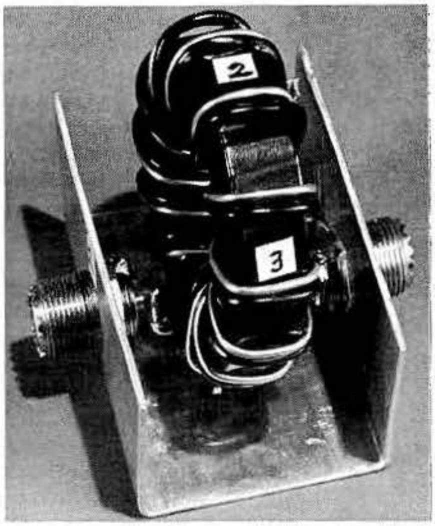

The broadband bifilar transformer shown has 12 bifilar turns, no. 14 wire, on a ferrite (Q1 material, 2.5 inch [64 mm] OD) core. Taps are also placed at 3, 6, and 9 turns, thus yielding ratios of 1.5:1, 2:1, 3:1 and 4:1. This transformer has a connection at the 1.5:1 ratio tap. Extra insulation was provided for the enameled wire in order to prevent voltage breakdown when matching into loads greater than 50 ohms.

powdered-iron core coil changes the 160-meter resonance by about 50 kHz. It should be noted that the windings had extra insulation (about one layer of vinyl tape) added on the inside part of the windings in order to prevent breakdown between them.

Low-permeability ferrite material Q2 (μ = 40, but still four times larger than the powdered iron material) was also investigated for use as a core for a 160-meter loading inductance. It was hoped that with several stacked cores of this material, the flux density would be reduced to a level where nonlinearities and permanent damage could be avoided. Two cores and about 10 turns of wire were used. The result was negative. Large changes in inductance were noted on voice peaks by the sudden large changes in the SWR. Also, small but permanent changes in the SWR were observed, indicating permanent damage. In contrast to ferrite, powdered iron maintains its linearity at high flux densities and suffers no permanent damage. As noted above, four cores of permeability 10, and 22-1/2 turns, provided the proper inductance (approx. 30 μH). The loss in this coil was about equal to the air-wound coil (33 turns, 6 tpi, no. 12 wire, 2-1/2-inch [64 mm] OD). Two stacked powdered-iron cores of permeability 25 were also investigated. This arrangement required about 17 turns, but was rejected because the loss, which was mainly in the two cores, went up to 3 ohms. This is certainly intolerable since it is of the order of magnitude of the radiation resistance of the very short vertical.

Concluding Remarks

The predominant characteristic of the short antenna is probably its very low value of radiation resistance. One must be careful of excessive ground and ohmic losses in coils and matching networks. A good image plane can eliminate the ground-loss factor. The application of a top hat eases the problem somewhat because it increases the effective length and thus the radiation resistance. The radiation resistance can be increased five-fold[8] by resorting to complete top-hat loading of a vertical. Combinations of top-hat and inductive loading can be used depending upon the radiator height and the bandwidth desired. One can also resort to a short folded unipole. With these combinations at his disposal, the designer is left with the problem of efficiently matching his transmission line to various impedance levels. The wide-band transmission-line transformer can provide the answer. It is capable of operating very efficiently over a wide range of impedances and impedance ratios.

The author would like to acknowledge the generosity of the following companies who supplied cores and information for the experiments: Indiana General Corp., Amidon Associates, and the Arnold Engineering Co.

Finally, the author would like to thank his colleague Mort Fagen, WA2EIT, for his encouragement and discussions of the antenna papers the author has written. This material was originally presented in *QST* by Jerry Sevick, W2FMI.

References

[1] King, *The Theory of Linear Antennas*, Harvard University Press, Cambridge, MA, 1956.

[2] Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, March 1973. (See chapter 1.)

[3] Ruthroff, "Some Broad-Band Transformers," *Proc. IRE*, Vol. 47, p. 137, August 1959.

[4] Turrin, "Application of Broadband Balun Transformers," *QST*, April 1969.

[5] Krauss and Allen, "Designing Toroidal Transformers to Optimize Wideband Performance," *Electronics*, August 16, 1973.

[6] Indiana General Corp., Electronics Div., Keasby, NJ 08832.

[7] Amidon Associates, 12033 Otsego Street, North Hollywood, CA 91607; The Arnold Engineering Co., Marengo, IL.

[8] See reference 2.

An RF-Sensed Antenna Changeover Relay

Flick of the switch, sir? Not with this system — it's automatic!

One subject that can be confusing for the newcomer to Amateur Radio is that of antenna-changeover systems. Several systems have been devised over the years and most can be classified into three or four main categories. Before delving into a discussion on these systems, a brief look at the simplest method of all might be in order, that is, no antenna changeover at all. With this setup two separate antennas are used, one for the transmitter and one for the receiver. This system is impractical for the most part since duplicate antennas would be required for each band of operation. There are instances, however, when separate antennas with different directive patterns can be used to an advantage, as in the case of specialized receiving antennas. Generally speaking, most stations use the same antenna for transmit and receiver purposes, so this is more the exception than the rule.

Several antenna-changeover systems, designed to transfer a single antenna from the receiver to the transmitter, are shown in Fig. 54. At A, a manual switch is used to transfer the antenna connection. This scheme is simple and works well, but is not without several drawbacks. Good-quality, high-power rf switches are often costly and

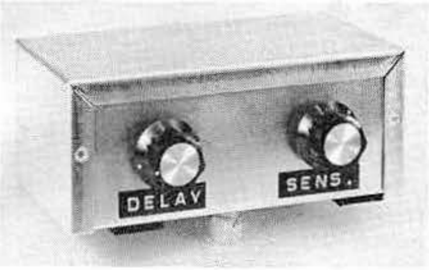

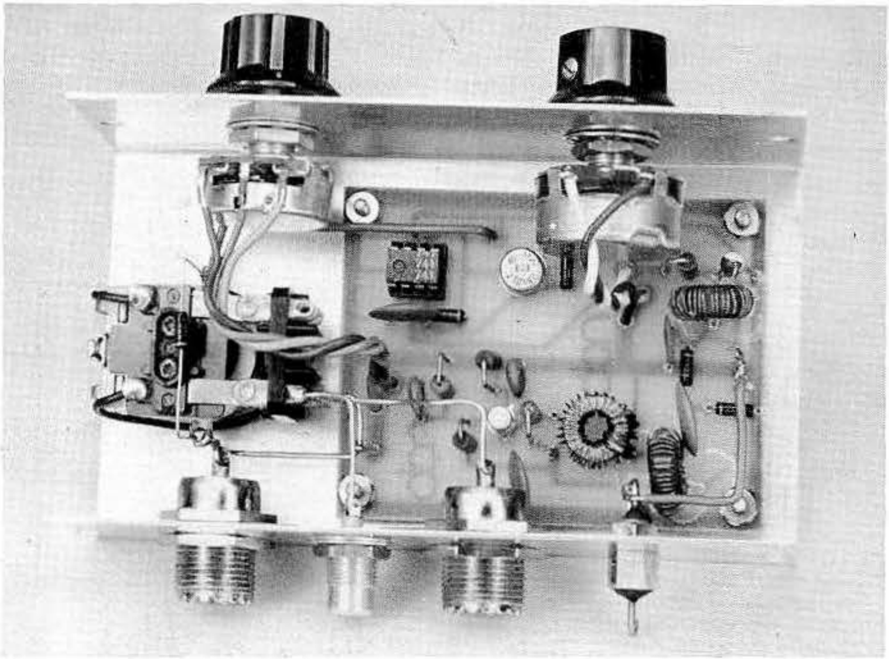

The inner layout of the antenna-changeover relay. Wiring or placement of circuit board and other components is not critical.

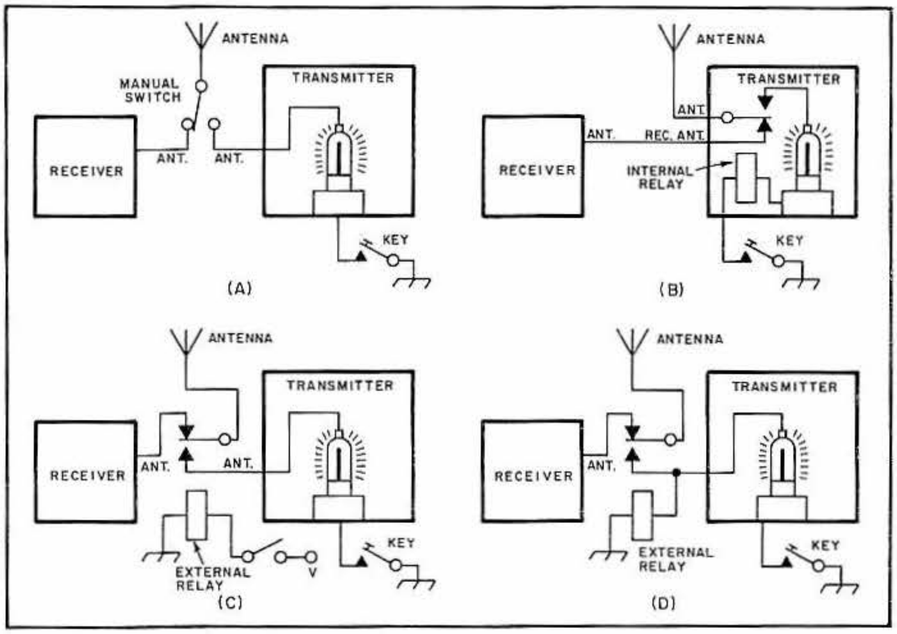

Fig. 54 — Several different antenna-changeover systems in block diagram form.

hard to locate and, of course, the operator is required to switch manually from receive to transmit.

One of these problems was eliminated with the advent of the coaxial relay. These relays were, in most cases, mounted external to the transmitter and receiver with connecting cables running to and from the unit, as shown in Fig. 54B. Many of these relays were rated for a kilowatt a-m signal, and maintained the 50-ohm line impedance. The voltage needed for the relay field was switched on or off with a low-powered toggle or foot switch. Alternatively, the transmit-receive switch might be located on the front panel of the transmitter with control voltage or a normally open or closed circuit available at a socket on the rear apron. Several extra sets of relay contacts were usually furnished with the relay to perform various functions, including that of muting the station receiver during transmit periods. Although this system of antenna transfer represented an improvement over the previous setup, it was still necessary for the operator to switch the antenna manually via the relay-control switch.

It's been said that laziness is the mother of invention and so it seemed natural that a more automatic system be developed. The relatively modern scheme at Fig. 54C has the transfer relay built into the transmitter. This system is quite popular and is used by most commercial manufacturers of amateur equipment that offer a separate transmitter and receiver. When the key is pressed the relay activates automatically and switches the antenna to the transmitter. A variable-length delay is usually included to hold the relay closed, thereby keeping the antenna connected to the transmitter during code characters. After the last group of code characters is sent, the relay will remain closed for a second or two — the amount of time controllable. When the relay deactivates, the antenna is reconnected to the receiver.

Unfortunately, some amateurs who build their own equipment or purchase transmitters and receivers of different manufacture, sometimes resort to switching methods A or B. These setups, of course, do not include the automatic features of system C. The circuit described here and shown in simplified form in Fig. 54D was designed with this in mind. Basically, this system detects the presence of rf at the output of the transmitter and immediately breaks the antenna connection to the receiver. At the same time the antenna connection to the transmitter is made. One of the features of this system is that no modifications to either the transmitter or receiver are required.

Circuit Description

R1 of Fig. 55 serves as an rf-voltage divider to permit the circuit to be used with transmitters of various power-output amounts. Rf energy is routed through C1 to the base of broadband amplifier Q1. The amplified hf-band energy is supplied to a voltage-doubler (D1 and D2) through a broadband toroidal step-down transformer, T1. The rectified rf voltage at the output of D1 and D2 is filtered by means of RFC2, C5 and C6. This prevents unwanted rf from reaching U1 and affecting its performance.

C6, R7 and R6 comprise a timing network (variable) which governs the hold-in time of the relay, K1. The smaller the resistance amount at R6, the shorter will be the time delay.

U1 functions as an inverting amplifier. When the input dc voltage at pin 2 increases, the output dc voltage at pin 6 decreases. The output voltage causes the base of relay driver Q2 to be forward biased negatively when it drops below approximately 1.4 volts. Diodes D5 and D6, by virtue of their combined barrier voltages (0.7 V each), establish the 1.4-V fixed bias level. Without the diodes, Q2 would conduct sufficiently to prevent the relay from dropping out during no-signal periods. D4 is used to suppress transients caused by the field coil of K1. When no rectified rf reaches U1, Q2 is cut off because of the high positive base voltage it receives from U1, and the relay contacts to the transmitter are open.

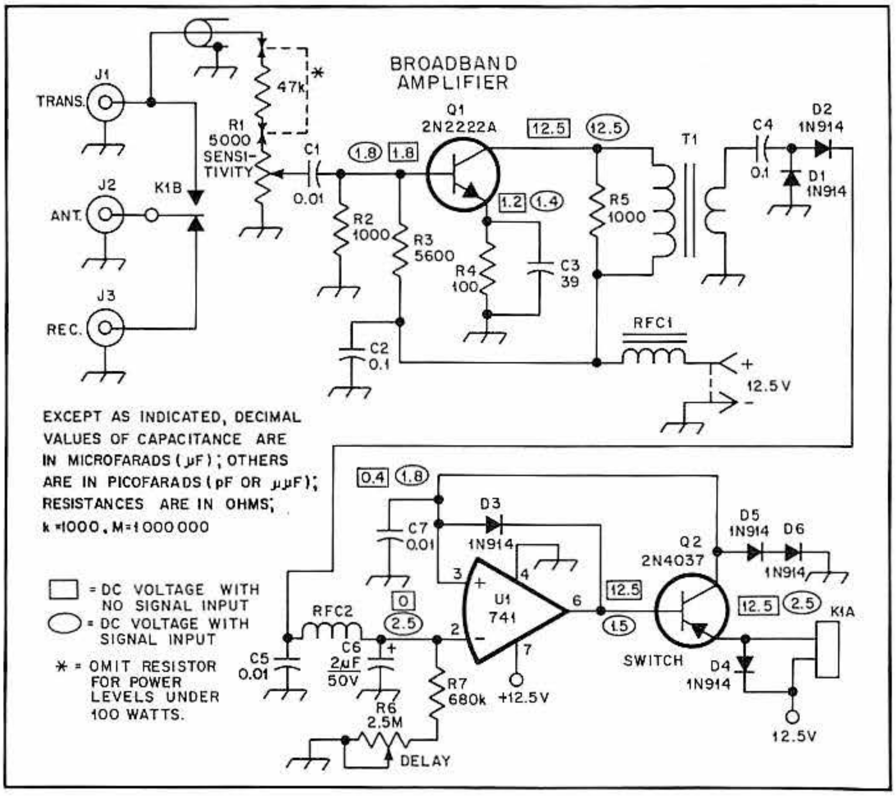

J1
TRANS.
J2
K1B
ANT.
J3
REC.
47k
*
R1
5000
SENSI-
TIVITY
C1
0.01
R2
1000
BROADBAND
AMPLIFIER
Q1
2N2222A
12.5
12.5
1.8
1.8
1.2
1.4
R3
5600
R4
100
C3
39
R5
1000
T1
C4
0.1
D2
1N914
D1
1N914
RFC1
C2
0.1
+
12.5V
-
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (μF); OTHERS ARE IN PICOFARADS (pF OR μμF); RESISTANCES ARE IN OHMS; k = 1000, M = 1 000 000
0.4
1.8
C7
0.01
D3
1N914
3
4
U1
741
6
2
7
0
2.5
+12.5V
12.5
1.5
Q2
2N4037
SWITCH
D5
1N914
D6
1N914
12.5
2.5
K1A
D4
1N914
12.5V
▭ = DC VOLTAGE WITH NO SIGNAL INPUT
◯ = DC VOLTAGE WITH SIGNAL INPUT
* = OMIT RESISTOR FOR POWER LEVELS UNDER 100 WATTS.
RFC2
C5
0.01
C6
2μF
50V
R7
680k
R6
2.5M
DELAY

Fig. 55 — Schematic diagram of the rf-sensed antenna-changeover relay. All resistors are half-watt composition type. Capacitors are disk-ceramic unless noted otherwise. Component numbers not appearing in the parts list are for identification purposes only.

J1, J2 — Connector, SO-239.
J3 — Connector, phono-type.
K1 — Dpdt relay, 5-ampere contacts, 12-volt field.
R1 — Linear potentiometer, 5000 ohms.
R2 — Linear potentiometer, 2.5 megohms.
RFC1, RFC2 — Rf choke, 42 turns no. 28 enam. wire on an Amidon FT-50-43 core.
T1 — Broadband transformer. Primary: 25 turns no. 28 enam. wire on an Amidon FT-50-43 core; secondary: 5 turns no. 28 enam. wire wound over primary winding.
U1 — Op amp, 741.

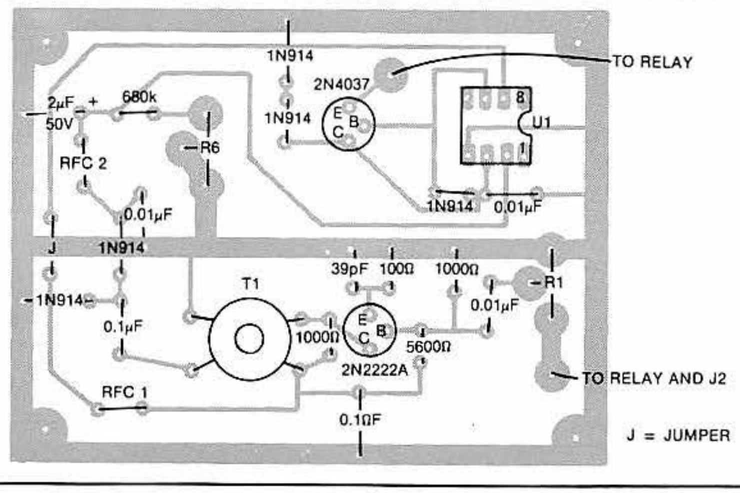

1N914
2N4037
TO RELAY
2μF
+
50V
680k
1N914
E
B
C
8
U1
1
RFC 2
R6
0.01μF
1N914
0.01μF
J
1N914
39pF
100Ω
1000Ω
R1
T1
1N914
0.1μF
0.01μF
1000Ω
E
C
B
5600Ω
2N2222A
TO RELAY AND J2
RFC 1
0.1ΩF
J = JUMPER

Fig. 56 — Circuit-board etching pattern for the changeover relay. The pattern is shown at actual size from the foil side of the board with gray areas representing copper. Decimal-value numbers alone represent capacitance in microfarads. Whole-number values with no units represent resistances in ohms; k = 1000. J = wire jumper. All components are mounted on the nonfoil side of the board. K indicates the cathode of a diode.

Construction

In Fig. 56 is a suggested printed-circuit board layout showing the foil pattern and parts placement. R1 and R6 may be circuit-board-mounted potentiometers if the builder desires, as the board layout was designed to accommodate such units. Potentiometers of this type are entirely suitable if the builder does not anticipate frequent changes in power levels and length of delay, as these controls must be readjusted for such changes. If the builder is interested in having access to these controls, panel-mounted potentiometers can be used at the front apron, in which case wires will have to be run from the circuit board to the controls.

Our finished model was housed in a 5 × 3 × 2-inch (127 mm × 76 mm × 51 mm) aluminum Minibox. R1 and R6 were mounted on the front panel and Kurz-Kasch knobs were placed on the control shafts to dress up the appearance of the finished product. D4 is attached directly to the relay terminals. The rear panel supports the terminals for the transmitter, antenna, receiver antenna and power supply. SO-239 connectors were used for the transmitter and antenna connections, and a phono connector was employed for the receiver antenna. Any type of connectors may be substituted to match those used in the builder's station. The power-supply lead is brought out through a feedthrough capacitor. This type of feedthrough is not necessary; a simple feedthrough insulator or wire could be used instead. In our case, the capacitor lead served as a convenient tie point during preliminary testing of the unit.

Operation

Connect the antenna, transmitter, receiver antenna lead and the power-supply connections to the appropriate terminals. Set R6, the time-delay control, to its minimum delay position (ccw) and the sensitivity control to its midrotation position. While keying the transmitter, decrease the setting of the sensitivity control to a point just above that where the transmitter no longer makes the relay energize. Once this has been set, it may be left in that position until a moderate change in power level has been made. The delay control should then be adjusted for the length of delay desired. The circuit shown has been tested from 1.8 to 30 MHz and at dc power input levels up to and including 1000 watts. This material was originally presented in *QST* by Doug DeMaw, W1FB, and Jay Rusgrove, W1VD.

Measuring Antenna Gain with Amateur Methods

Beat the cost of expensive test gear. Make your antenna measurements using this unique method! A VU meter, a handful of parts and this simple explanation are all you need.

Several excellent articles have been written about the techniques used at vhf conferences to measure antenna gains. Two such articles in particular, by Clark[1] and Knadle,[2] suggest very good antenna-range procedures. Both describe techniques which, if properly applied, will produce accurate relative antenna-gain data (in decibels) for vhf-uhf antennas.

However, these antenna range-measurement systems require equipment most amateurs lack — ratiometers, precisely calibrated attenuators, or at least specialized detectors and commercial grade audio or VSWR meters. Moreover, setting up a near-field antenna range is a tricky business in itself. One text[3] simply warns amateurs not to expect accurate antenna-gain measurements if they try it! The problem, of course, is that outdoor antenna ranges invite all sorts of error-causing reflections from various objects, not the least of them being the earth itself.

Ground reflections alone can completely invalidate the results on an otherwise good antenna range. At the 1973 West Coast VHF Conference, for instance, the author's 8-element quagi for 432 MHz[4] was measured at precisely the same gain as a well-built 15-element Yagi more than twice its size, despite the fact (established in other tests) that the bigger antenna had 2 to 3 dB more actual gain! The problem was that, even with the sophisticated measuring equipment in use, ground reflections were distorting the results. The best "gain" indication came with both antennas within a few inches of the earth, and the smaller quagi with its lower Q didn't mind the detuning effects of the earth's proximity as much as the bigger and higher Q array.

With results like that, wild claims are sometimes made for one vhf antenna or another. And the claims may be even wilder on the hf bands, where an amateur may well be told his signal is two of three S units louder than another ham's signal when the actual gain difference between the two antennas is insignificant. The ham who "won " the comparison by several S units, of course, is likely to go away thinking his antenna is 12 or 18 dB better than the other guy's, since S meters are supposed to show 6-dB signal difference per S point.

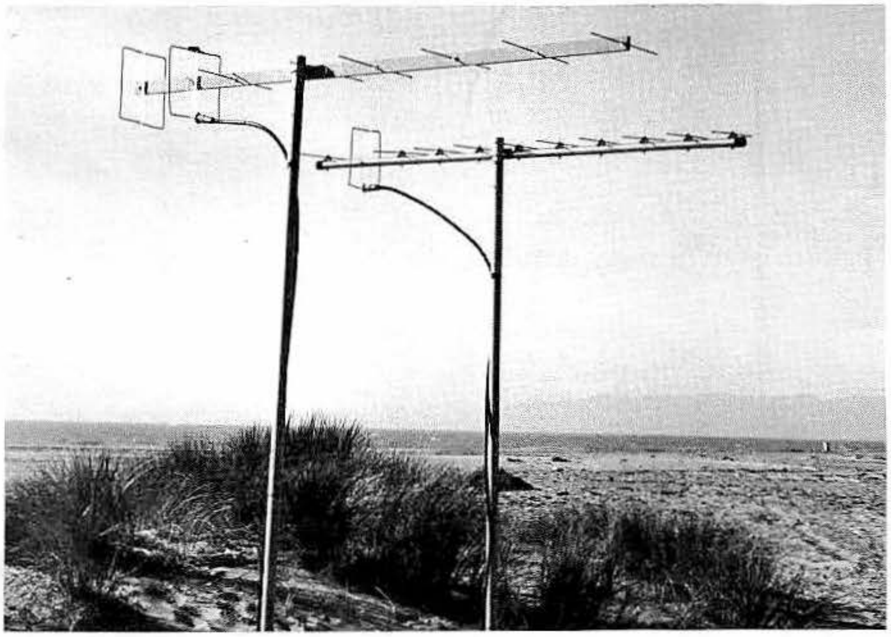

For the antenna gain-measurement technique described here, two identical masts in a clear area (here a southern California beach) are required. With the test and reference antennas at equal heights and fed with identical feed lines, the signal produced by the two antennas is compared by rapid and repeated switching. Here an 8-element quagi is tested against an 11-element Yagi that has been modified for quagi-type drive. The modified Yagi delivers 3 dB more gain than the conventional version, but 2 dB less than the quagi.

A Better Way

This article describes a more reliable and valid way to estimate antenna gain, and without expensive or hard-to-find test equipment. The technique requires only this equipment: a cooperative nearby ham with a transmitter capable of generating a few watts of steady carrier, a receiver whose agc can be turned off, an ordinary audio VU meter, and some antennas to measure.

The technique is nothing more than a more rigorous application of the test procedure radio amateurs have always used when they asked for "comparison" signal reports. It takes advantage of the fact that audio VU meters, unlike S meters, are calibrated directly and fairly accurately in decibels, with easy readout down to fractions of a dB. And with the agc turned off, most receivers will closely reproduce changes in the signal input at the antenna terminals with corresponding changes in audio output until they are badly overloaded.

This means that a nearby amateur's signal transmitted at a reasonably low power level can be used to compare the gain of various antennas by rapidly switching back and forth between them and averaging the results.

The only trick is to use a "test signal" that is coming from a point source — not a signal that is being reflected from surrounding objects. If your test antennas can "see" the signal source a few miles away and are well clear of nearby objects (and clear of the earth itself), valid results should be attainable if the procedures below are followed.

The author has spent many hours working with both local antenna ranges covering a few dozen wavelengths, and on-the-air signals from local stations. It is apparent that ground reflections and other

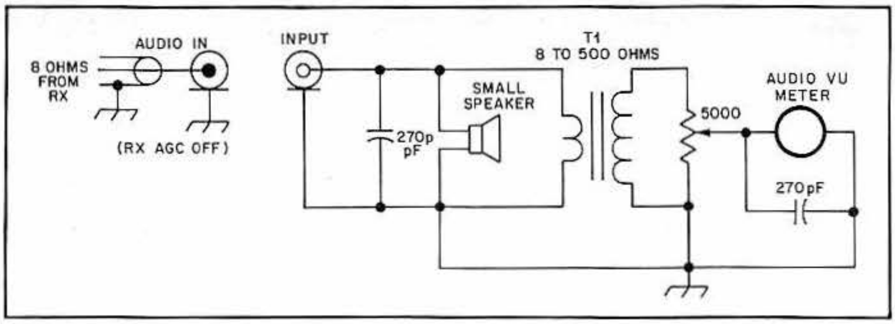

Fig. 57 — The audio VU-meter system for antenna gain measurements.

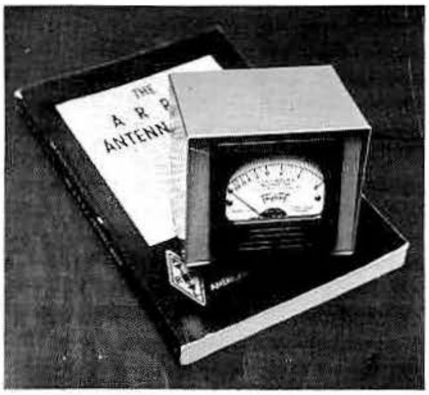

Fig. 58 — This small box houses all the test equipment required for antenna gain measurements using the technique described here. The front-panel audio VU meter is calibrated directly in decibels, with readout possible to about 0.2 dB in the most expanded part of the scale. Not visible are the meter level potentiometer (on rear), and small speaker which is mounted downward.

such error sources are less serious problems when a steady, on-the-air carrier a few miles away — and not a local signal a few dozen wavelengths away — is the signal source.

Test Procedures

Here's how the procedure works at vhf (at hf, it's a little more complicated because antennas are larger, as we'll see later):

1) Build an audio VU metering circuit like the one shown in the photo and Fig. 57. All parts should be readily available anywhere that consumer electronics parts are sold. The speaker and meter-gain control are handy to have when the receiver itself is located some distance away from the antenna test site, but they aren't absolutely essential.

2) Get a nearby amateur (perhaps one to five miles away) to generate a steady carrier into an antenna of the same polarization as your test antennas. A small beam pointed toward your test site is much better than a dipole or groundplane antenna because it is less likely to invite reflection error. A beam pointed somewhere else is worst of all!

3) Two of the antennas to be tested should be set up side by side, with identical lengths of feed line going to a coaxial relay or well-isolated switch. The two supporting poles should be high enough that the test antennas are at least two wavelengths above ground (13 feet [3.96 m] or more at 2 meters) and clear of nearby obstructions. To minimize interaction between the two antennas, the two masts should also be at least one to two wavelengths apart and standing side by side in relation to the signal source.

4) Make sure you can raise and lower your antennas and move them from side to side by at least a full wavelength or so (81 inches [2.06 m] at 2 meters) without the indicated signal varying more than about 1 dB. If you can't do that, you probably have reflection problems that will have to be solved before you can make valid readings. If this is the case, try different sites until you do achieve nonspotty reception.

5) Check the receiver agc to be sure it is off; then rapidly switch between the two antennas and record the difference in decibels. Do it several times and calculate the average figure (the mean, for statisticians).

6) Next reverse the two antennas. Put antenna no. 1 on mast no. 2 and attach it to feed line no. 2. Move antenna no. 2 over to mast no. 1 and feed line no. 1. Do not eliminate this step and expect valid results.

Now run the test again. If there are more than very slight differences from the previous test, one antenna or the other is seeing some obstruction or reflection source. Your site isn't clear enough for good results; try a different one.

Since this test procedure requires only receiving, there is no reason not to perform it somewhere other than at home — anywhere that is clear of obstructions. Use a receiver or transceiver that runs on battery power, if necessary.

7) If you have followed steps one through six, you now have an accurate indication of the relative gain of your two antennas. Now you can make one of your antennas a "reference" antenna and compare any other antennas you or

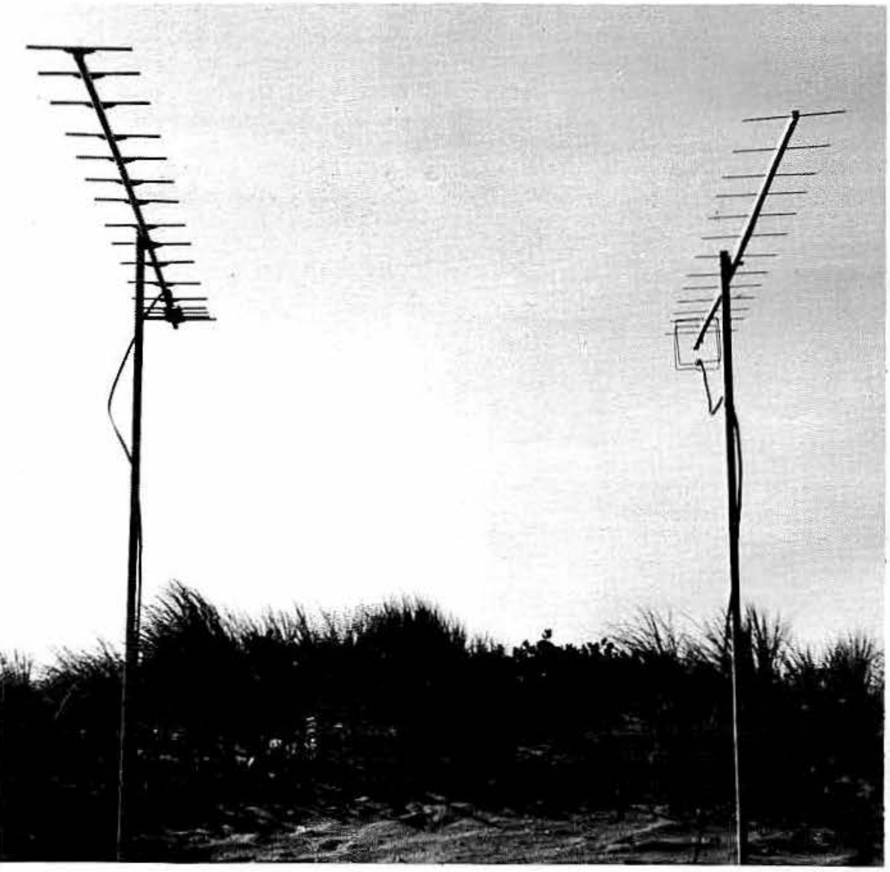

Fig. 59 — Here a 15-element long-boom quagi for 432 MHz is compared to a 16-element log-periodic Yagi, using the test procedure described in this article. The two antennas both deliver 15 dB gain over a dipole, decisively better than any other 432-MHz antenna tested here. The antennas to be tested are set up side by side and fed with identical lengths of coax. The antennas should be at least two wavelengths above ground and clear of nearby obstructions. To minimize interaction between the two antennas, the masts should also be at least one to two wavelengths apart, and standing side by side in relation to the signal source.

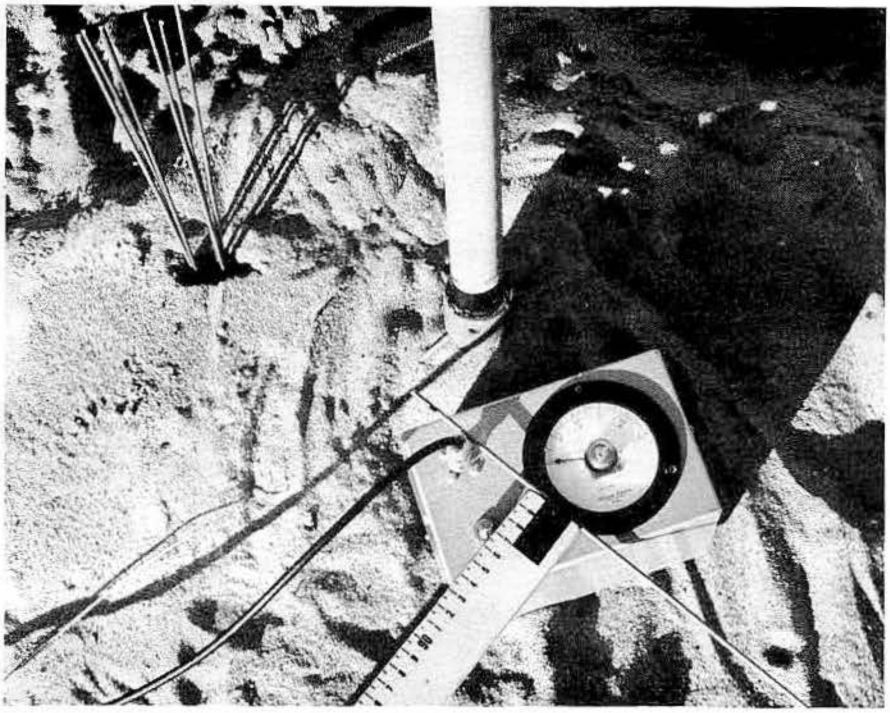

Fig. 60 — An earlier version of the metering system in Fig. 57, with assorted elements of various lengths and the yardstick used to make temporary placements on the boom.

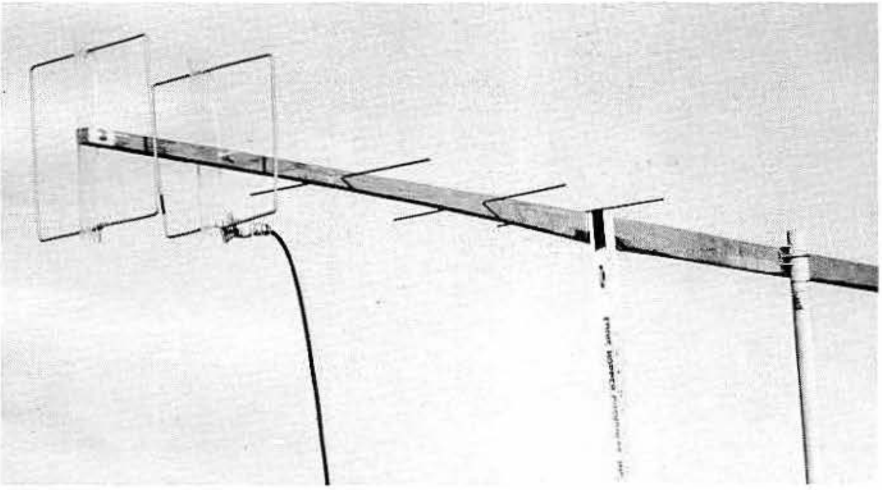

Fig. 61 — The technique used to design a new vhf or uhf antenna experimentally. The boom is shown with the first few elements already in place, and each additional element is adjusted back and forth on the boom until a point of maximum indication of the VU meter is found.

neighboring hams have against this reference.

So how do you determine how much gain your antennas have over a dipole? Some people would say, "Simple, make the reference antenna a dipole." But it may not be that easy, because dipoles are notorious for finding reflections that more directional antennas can't "see." At vhf conference gain-measurement seminars, dipoles have sometimes shown as much as 3 or 4 dB gain over a dipole!

The best reference antenna is probably a small beam of known (or approximately known) gain, assuming you don't have the National Bureau of Standards type of reference antenna that professionals use for this purpose. A 3-element Yagi or a 2-element quad built to handbook dimensions will usually deliver about 6 dB gain over a dipole at 2 meters, if the SWR is okay at the design frequency.

Thus, if you find that a particular 8-element antenna consistently produces 4.5 more dB on the VU meter than your little reference beam, you'll know you have an antenna with about 10.5 dB gain. You won't be exactly correct, but you'll be close to the right figure. Conversely, if your bigger beam doesn't show much gain over the small one, and there are no measurement errors of the sort just described, you'll know for sure that something is wrong with the big antenna.

Once you have established the integrity of your test site, you can even design your own antennas very accurately, using the audio VU-meter technique. To do that, set up your reference antenna and begin designing your new antenna by putting an empty boom on the other pole. Now start with 2 or 3 elements, get a decent SWR and try adding more elements.

As you add elements, you will be able to actually see where each new element "wants to be" on the boom, and you can see your new antenna gain over the reference antenna increase as you go along. Try all sorts of element lengths and spacings as you progress.

Antenna design professionals call this process "perturbation" and they do it with sophisticated mathematics, but you can design antennas that work just as well as theirs on your own antenna range, once you master this technique. Then, if you do go to a vhf conference with your best home-designed antennas, you'll know about how much gain they have before you enter the gain contest. And you'll never come away from such an event with an antenna that has "minus 3 dB gain!" Moreover, you'll know that your measurements, made without any of the sophisticated equipment you saw at the vhf conference, are about as accurate as anybody's. In fact, you may find your own measurements are more valid than the ones at the conference, especially if the antenna-gain figure jumps all over the place when you raise and lower it at the conference antenna range.

HF Antenna Measurements

So far, all of this has been devoted to vhf antenna-gain measurements. Many readers will say, "That's fine, but I want to measure the gain of my 5-element 20-meter beam!" Of course, an audio VU meter and a receiver without agc are a much better way than the notoriously inaccurate S meter to compare two dissimilar hf antennas. The author has seen an S meter show three or four *S units* of signal difference between two antennas that are really only 3 or 4 dB apart in gain.

However, the only accurate way to calculate the gain of an antenna is to compare it to a reference antenna at the same height and then switch the two antennas to opposite masts to validate the data. That may be impractical at hf, but perhaps not as much so as it seems at first. Most amateurs motivated enough to install a 5-element beam on a crankup tower can probably find a way to temporarily set up a smaller beam (not a dipole) on a 20-foot (6.10 m) mast above a roof. Then the two antennas at identical heights can be compared on stable, line-of-sight signals.

The results of this test can be invalidated by obstructions or even by other antennas mounted on the same tower with the test antenna, but at least this will be more valid than a comparison against another ham's readings, when the two locations and rigs may be entirely different.

And if you can switch your test reference hf antennas to opposite masts, you can measure the actual gain of one over the other. One suggestion before you put that giant, long-boom array on the big tower: Put a little beam on the tower and an identical one on the rooftop mast and perform the audio VU-meter tests to see if the two antenna sites do seem to be identical. If they are comparable, then you can put the big antenna on the tower and find out how much gain it has, if any, over the small one on your roof when the two are at the same height.

This may seem cumbersome, and in fact it is! That's why most antenna designers do their research at vhf where antennas are so nice and small, and then scale their best designs up for the hf bands. But if you really want to know your big beam's actual gain, this is probably the easiest way to find out.

Conclusions

The antenna gain-measuring technique described in this article has repeatedly been proved to be reliable and valid. The author has used it to design new antennas that work well enough for all kinds of contest and DX work, including moonbounce. For a dedicated amateur, working DX on a home-designed antenna far surpasses the thrill of working the same DX with an antenna purchased at a store! But remember, whether you're working 20 meters or 23 centimeters, a signal source, a receiver with its agc disabled and an audio VU meter are the only pieces of equipment you need to measure antenna gain with this technique. This material was originally presented in *QST* by Wayne Overbeck, PhD, N6NB.

References

[1] Clark, "Direct Methods for Measuring Antenna Gain," *Ham Radio,* July 1969.

[2] Knadle, "UHF Antenna Ratiometry: An Outdoor Range Improvement Technique," a paper presented to IEEE Intercon session L, April 1975.

[3] Brier and Orr, *VHF Handbook for Radio Amateurs,* Radio Publications, Inc., Wilton, CT p. 225.

[4] Overbeck, "The VHF Quagi," *QST,* April 1977.

[5] Knadle, "UHF Antenna Ratiometry," *QST,* February 1976.

Understanding Antennas for the Non-Technical Ham

A Book By Jim Abercrombie, N4JA
(Jim Abercrombie *n4ja@prtcnet.com*)

Illustrations by Frank Wamsley, K4EFW

Edited by Judy Haynes, KC4NOR

Copyright July 2005. Second Edition

This is a book length web article provided by the author FREE for all hams.

Note: This is copyrighted material and is the property of Hamuniverse.com and/or the article author and is to be used only for personal non-profit educational use.

Hamuniverse.com is Powered by Ham Radio!
© 2000 - 2008 N4UJW Hamuniverse.com or article author! - All Rights Reserved.

Preface

One reason for writing this book is to educate you so you can make an informed choice concerning the best antenna for you. Another reason is to dispel the many antenna myths that circulate in the amateur community. The third reason is a desire to teach basic antenna theory to the average ham. Therefore, to achieve that goal, you should read this book from cover to cover. It was written primarily for the newcomer and the non-technical old-timer.

This book is about common medium wave and high frequency (short wave) antennas, but the theory presented here relates to antennas of any frequency. It is in a condensed form and the antenna theory is explained so most hams can understand it. Realizing many hams are mathematically challenged, only simple mathematics procedures are used. If you can add, subtract, and divide using a calculator, you will not have trouble with this book.

A few principles in here are based on conclusions drawn from the Laws of Physics. Everything else in this book can be found scattered through The *A.R.R.L. Antenna Book* and nothing in here contradicts what is written there.

Contents

1. Why All The Fuss About Antennas 1

2. How Antennas Work 3

3. Polarization Of Electromagnetic Waves 5

4. Frequency 6

5. The Ionosphere and Modes of HF Propagation 7

5.1 The Ionosphere 7
5.2 Ground-Wave Propagation 8
5.3 Direct Wave or Line of Sight Propagation 8
5.4 Propagation by Refraction 9
5.5 Skywave Propagation 9
5.6 Greyline Propagation 10
5.7 Long Path Propagation 10
5.8 160-Meter Propagation 11
5.9 80-Meter Propagation 11
5.10 40-Meter Propagation 11
5.11 30-Meter Propagation 12
5.12 20-Meter Propagation 12
5.13 17-Meter Propagation 12
5.14 15-Meter Propagation 12
5.15 12-Meter Propagation 12
5.16 10-Meter Propagation 13

6. Standing Wave Ratio 14

7. Real Antenna Systems 16

8. Half-Wave Resonant Dipole Antennas 17

8.1 The Half-Wave Flat-Top Dipole 17
8.2 Inverted-V Dipole 18
8.3 Dipole Shape Variations 20

8.4 Calculating the Length of a Half-Wave Resonant Dipole 20
8.5 The Decibel 21
9. Antenna Basics 22
9.1 Resistances and Reactance 22
9.2 Feeding Dipoles Efficiently 23
9.3 The Cause of Feed-Line Radiation 25
9.4 Baluns 26
10. Other Types of Dipoles 27
10.1 A Shortened Dipole Using Loading Coils 27
10.2 All Band Dipole 28
10.3 The Sloping Dipole 28
10.4 Half-Wave Resonant Sloping Dipole 28
10.5 The Folded Dipole 29
10.6 The Double Bazooka Dipole 30
10.7 Broad-Banded Coax-Fed Fan Dipole 31
10.8 Two-Element Collinear Dipole 31
10.9 4-Element Collinear Dipole 32
10.10 Coax-Fed Dipoles Operated on Odd Harmonic Frequencies 33
10.11 All Band Random Length Dipole 35
10.12 All Band Center-Fed Random Length Dipole 35
10.13 A 2-Band Fan Dipole 35
10.14 Trapped Dipole for 75 and 40 Meters 36
10.15 The Extended Double Zepp Dipole 37
10.16 The G5RV Dipole 37
10.17 Off-Center Fed Dipoles 39
11. End-Fed Antennas 42
11.1 End-Fed Zepp 42
11.2 End-Fed Random Length Antenna 42
12. The Half Sloper 44
13. Vertical Antennas 46
13.1 Why Verticals Are Used 46
13.2 Disadvantages of Using Quarter-Wave Verticals 47
13.3 Long and Short Verticals 48
13.4 Unscientific Observations of Verticals 49

13.5 The Inverted-L Vertical 50
13.6 Vertical Mobile Antennas 51
14. One-Wavelength Single Loop Antennas 54
14.1 The Horizontally Oriented Loop 54
14.2 The Vertically Oriented Single Loop for 40 and 80 Meters 56
15. Directional Beam Antennas 59
15.1 The Monoband Yagi 59
15.2 Trapped Multi-Band Yagis 62
15.3 The SteppIR Antenna 65
15.4 The Log-Periodic Array 65
15.5 Directional Cubical Quad and Delta Loop Antennas 66
15.6 The Quagi 72
16. Gain Versus Front-to-Back 73
17. Feed-Lines Commonly Called Transmission Lines 74
18. Antenna Safety 76
18.1 Erecting Antennas on Masts 76
18.2 Tower Safety 76

1.

Why All The Fuss About Antennas

Definition: An antenna is a piece of metal, a conductor of electricity, to which you connect the radio. It radiates your signal and receives the signals you want to hear.

Definition: An antenna system consists of the antenna, the feed-line, and any matching unit. Most antennas are made of copper or aluminum, while most mobile antennas are made of stainless steel. A feed-line consists of two conductors that carry the signal to and from the radio and to and from the antenna. A matching unit can be an antenna tuner, a series matching section, or one of several different kinds of matching circuits at the feed-point.

Does the type of antenna make much difference? Here is an example: Once in 1959 two of us were involved in testing two antennas on 15 meters. The late R. Lynn Kalmbach, W4IW, using one antenna received a 30-dB better signal report on his antenna from a station in England than we did on our antenna. (Decibel or dB will be explained later). Thirty dB means his signal appeared that he was running 1000 times more transmitter power than we were. At that time, we didnt live that far apart so we couldnt blame it on propagation. We both were running about equal power. Both antennas were at 50 feet. The comparison proved that a good antenna could make a difference. Lynn used a home-built G4 ZU mini-beam; we were using a 15-meter 2-element Mosely Mini-Beam, which had short loaded elements. Evidently, it had a lot of loss.

Another example: Today we hear people breaking in to our ragchews with signals almost level with the noise. Why is that? The reason is they are using the wrong antennas. Their signals are 20 to 30 decibels below everyone elses. They are making contacts, but just barely. The first question our group asks, “What kind of antenna are you using?” Experienced amateurs know the antenna can make all the difference. The guy with the poor signal sometimes will blame his bad signal report on band conditions or his lack of a linear amplifier. He is just sticking his head in the sand.

What we are trying to prove is next to your radio, the most important part of your station is the antenna. Many years ago, an old-timer said, “For every dollar you spend on a radio, you should spend two dollars on your antenna.” That is also true today. You can do more to improve your signal strength with antennas than you can ever do by increasing your power. Having the ability to make contacts on a particular antenna doesnt mean it works well! Any antenna will make contacts, but your signals will be stronger on some antennas than on others. In addition, some antennas hear better than others.

2.

How Antennas Work

First of all to work properly the antenna system must be matched to the transmitter. That is, all modern transmitters have an output impedance of 50 ohms. Antenna systems range in impedance of a few ohms to several thousand ohms. There are several ways to match them: pruning the length of the antenna, using an antenna tuner, matching the antenna with a length of transmission line called a matching section, or the use one of several matching systems at the antenna feed-point. Antenna matching is beyond the scope of the material found in this book and it is suggested you consult a more comprehensive antenna manual. Simple half-wave dipoles eliminate the need for a matching system because a resonant half-wave dipole has an impedance near 50-ohms.

You must understand electromagnetism to understand how antennas work. If you attach the two poles of a direct current (DC) voltage source to the two ends of a coil of wire, current will flow through the coil of wire and it will become magnetized. The magnetized coil is known as an electromagnet. Its magnetism will extend out to infinity becoming weaker with distance. Remove the voltage and the magnetic field collapses back into the coil. If an alternating current (AC) is connected to the coil, the magnetism moves out and collapses into the coil in step with the frequency of the alternating current source. The north and south poles of the electromagnet reverse on each half-cycle of the AC voltage.

If voltage and current can cause a coil to become magnetized, the reverse is true: A magnetic field can produce a voltage and a current in a coil. This is known as Faradays Principle of Magnetic Induction. A voltage will be produced at the ends of the coil of wire as you move any permanent magnet close to and parallel to the coil. The difference in this case is the magnet must be kept moving. Move the magnet in one direction, and current will flow in one direction. Reverse the direction the magnet is moving and the current will flow in the opposite direction. Moving the magnet back and forth produces alternating current. An AC generator spins a coil of wire between the two poles of a magnetic field. It doesnt matter which one is moving. The coil or the magnet can be moving. Any moving magnetic field can induce current in anther coil. It doesnt have to be a piece of metal we call a magnet. Imagine a moving magnetic field produced by AC circulating in and out of a coil. If that moving magnetic field passes through a second nearby coil, it will induce an alternating current in the second coil. A transformer uses this method to work. Transformers have a continuous iron core running from the inside of one coil through the inside of the second coil to confine the magnetism inside the iron core. This makes the transformer nearly 100% efficient since only a little of the magnetic energy escapes.

A straight wire that has an AC current flowing through it also has a magnetic field surrounding it. But it is a weaker field than is produced by a coil. The magnetic field from the wire radiates out into space and becomes weaker with distance. The radiating magnetic field from a wire is known as "electromagnetic radiation" and a radio wave is one type of it. The wire that radiates becomes the transmitting antenna. Some distance away, a second wire in the path of these waves has current induced into it by the passing electromagnetic waves. This second wire will be the receiving antenna. The voltage in the receiving antenna is many times weaker than the voltage in the transmitting antenna. It may be as weak as one-millionth of a volt or less and still be useful. The receiving antenna feeds that voltage to the amplifiers in the receiver front-end where it is amplified many thousands or millions of times.

The dipole antenna is made of a wire broken in the center and where broken, each half of the wire connects to an insulator that divides the wire in two. Two wires from the voltage source, which is the transmitter, are connected across the insulator. On one side of the dipole, the current in the form of moving electrons flows first from the voltage source toward one end of the dipole. At the end, it reflects toward the voltage source. The same thing occurs on the other half of the wire on the other half cycle of alternating current. An antenna that is the right length for the current to reach the far end of the wire just as the polarity changes is said to be resonant. Because electricity travels at 95% the speed of light in a wire, the number of times the polarity changes in one second (frequency) determines how long the wire has to be in order to be resonant.

3.

Polarization Of Electromagnetic Waves

Electromagnetic waves travel away from the wire in horizontal, vertical, slanted, or circular waves. If the antenna wire runs horizontal or parallel to the earth, the radiation will be horizontally polarized. A wire or conductor that runs at right angles to the earth produces vertical radiation. A slanted wire has components of both horizontal and vertical radiation. Crossed wires connected by proper phasing lines that shift the phase from one wire to the other wire by 90 degrees will produce circular polarization. Amateurs working orbiting satellites at VHF, UHF, and microwave frequencies use circular polarization.

When your high frequency signals are reflecting off the ionosphere, it isnt important if the other stations antenna has the opposite polarization from yours (the polarization does matter for line of sight communication). The reflected polarized waves passing through the ionosphere are slowly rotated causing fading signals (QSB). The reason the polarization of antennas is most important is that it determines the angle of radiation. Horizontally polarized antennas at ordinary heights used by hams produce mostly high angle radiation and weaker low angle radiation, but this doesnt mean there is no low angle radiation. It is there but is weaker than high angle radiation. However, you must put a horizontally polarized antenna up more than one-wavelength high to get a strong low angle radiation. One wavelength is 280 feet on 80 meters, 140 feet on 40 meters, and 70 feet on 20 meters. High angle radiation works nearby stations best and low angle radiation works distant stations (DX) best. A vertically polarized antenna produces mostly low angle radiation, with its high angle radiation being weak. For this reason, vertical antennas do not work as well as horizontal antennas do at ordinary heights for working stations less than about 500 miles away.

4.

Frequency

The number of times the polarity of an AC voltage changes per second determines its frequency. Frequency is measured in cycles per second or Hertz (Hz). A thousand cycles per second is a kilohertz (kHz). One million hertz is a Megahertz (MHz). The only difference between the 60 Hz electric power in your house and radio frequencies (RF) is the frequency, but 60 Hz electricity in a wire also produces electromagnetic radiation just like radio waves. Useful radio waves start at 30 kHz and go upward in frequency until you reach the infrared light waves. Light is the same kind of waves as RF except light is at a much higher frequency. Light waves are used like radio waves when they are confined inside fiber optic cable. Above the frequencies of light are found x-rays and gamma rays.

The radio bands: The Long Wave Band (LW) starts at 30 kHz and goes to 300 kHz. The Medium Wave Band (MW) is from 300 kHz to 3000 kHz or 3 MHz. The High Frequency Band (HF) is from 3 MHz to 30 MHz. The Very High Frequency Band (VHF) is from 30 MHz to 300 MHz. The Ultra-High Frequency Band (UHF) is from 300 MHz to 3000 MHz or 3 GHz. Above these frequencies are several microwave bands which are defined as the Super High Frequency Band (SHF).

5.

The Ionosphere and Modes of HF Propagation

In the upper air around fifty miles and higher where the air molecules are far apart, radiation from the sun strips electrons from oxygen molecules causing the molecules to become ionized forming the ionosphere. The ionized oxygen molecules and its free electrons float in space forming radio-reflecting layers. Ionization of the ionosphere varies by the time of day, seasons of the year, and the sunspot cycle. The strength of ionization also varies from day to day and hour to hour. Since the height of the ionosphere varies, the higher the ionized layer becomes, the farther the skip will be.

5.1 The Ionosphere

The part of the earths atmosphere called the ionosphere is divided into three layers. The three layers are, from lowest to highest, the D layer, the E layer, and the F layer. Each layer has a different effect on HF radio propagation.

Being at a lower altitude, the D layer molecules are squeezed closer together by gravity than those in higher layers, and the free electrons reattach to the molecules easily. The D layer requires constant radiation from the sun to maintain its ionization. Radio waves at lower frequencies such as the frequencies of the AM broadcast band cannot penetrate this layer and are absorbed. The higher frequency signals are able to pass through the D layer. The D layer disappears at night causing AM broadcast stations to reflect from the higher layers. This is why AM broadcast signals only propagate by ground wave in the daytime and they can be received from great distances at night. Like the broadcast band, the D layer absorbs signals on 160 and to a lesser extent 80 meters during the day making those bands go dead. During solar flares, the D layer becomes ionized so strongly that all high frequency radio waves are absorbed, causing a radio blackout.

E-layer propagation is not well understood. Being at a lower altitude than F layer, the E layer is responsible for summertime short skip propagation on the higher high frequency bands. The skip zone is around 1000 miles, but at times when the E-cloud covers a wide area in the summer, double hops can be seen. A double hop occurs when the signal reflects from the ionosphere, then returns to the ground, reflects from the ground back to the ionosphere where it is reflected back to the ground. A double hop can propagate the signal 2000 miles or more. The E-layer forms mostly during the day, and it has the highest degree of ionization at noon. The E layer like the D layer disappears at night. Even so, sporadic-E propagation can and does form at night. There is a minor occurrence of sporadic E propagation during the wintertime. On rare occasions, sporadic E propagation can surprise you by occurring anytime regardless of the sunspot cycle or the season of the year.

The F layer is the highest layer and it is divided into two levels: F1 and F2. At night the F1 and F2 merge into one layer. During the day, the F1 layer doesnt play a part in radio propagation, but F2 does. It is responsible for most high-frequency long distance propagation on 20 meters and above. However, the F layer makes it possible for you to work DX on the lower bands at night. Sunspots are responsible for the ionization layers and in years with high sunspot numbers, worldwide contacts can be made easily on 10-20 meters by F2 layer propagation. In years of low sunspot numbers, working distant stations is difficult on those bands. Consequently, 10 and 15 meters will be completely dead most days and 20 meters will go dead at night. In years of low sunspot numbers DX contacts are easily made at night on 160, 80, and 40 meters. The sunspot numbers increase and decrease in 11-year average cycles.

Since the curvature of the earth averages about 16 feet every 5 miles, an object 5 miles from you on perfectly flat earth will be 16 feet below the horizon. Because light travels in straight lines, you cannot see objects beyond the horizon. Radio waves travel in straight lines, but there are ways to get them beyond the horizon. This is referred to as propagation.

5.2 Ground-Wave Propagation

Ground wave works only with vertical polarization. One side of the antenna is the metal vertical radiator and the other side of the antenna is the earth ground. The surface wave in the air travels faster than the part of the wave flowing through the ground. The surface of the earth is curved like the curved part of a racetrack. On the curved track, a car on the outside of the track has to travel faster than the car on the inside lane to stay even, and the two cars travel in a curved path. Although the wave in the air travels faster than the wave on the ground, the two parts of the wave cannot be separated. Because of this, the radio wave also travels in a curved path that follows the curvature of the earth.

The AM broadcast stations use ground wave propagation during the day and skywave propagation at night. Since radio waves at lower frequencies conduct better through the ground, an AM broadcast station on 540 kHz will be many dB stronger than a station on 1600 kHz, if both run the same power. This fact is important in understanding why ground mounted verticals do not work as well at high frequencies as they do on the broadcast band.

5.3 Direct Wave or Line of Sight Propagation

Antennas located on high structures can “look” over the horizon and “see” the receiving antennas. Because refraction is involved, direct waves travel 20% farther than light waves due to scattering of radio waves by the environment. Trees and other foliage are invisible to HF radio waves. Direct wave propagation is possible at all frequencies, but this mode of propagation is seldom used on our high frequency bands, but it is the usual propagation mode used by repeaters and others on VHF and UHF. If you watch TV on an outside antenna or on a "rabbit ears antenna," you are receiving the signal by direct wave propagation.

5.4 Propagation by Refraction

Refraction occurs when the lower part of a wave travels slower than the top part of the wave because the wave is passing through two media. These media can be two layers of air at different temperatures or they can be air and a solid. One form of refraction is caused by a radio wave passing over a hill or ridge being bent as it passes over the obstruction. This is known as "knife edge refraction." Another form of refraction occurs when layers of air of different temperatures bend the radio waves around the horizon. This is called tropospheric ducting. This mode of propagation makes long distance contacts possible at VHF frequencies. Tropospheric ducting does occur on 10 meters and lower frequencies and is noticeable when other forms of propagation are absent. On high frequency bands, many hams mistakenly call tropospheric ducting and direct wave "ground wave."

5.5 Skywave Propagation

Skywave propagation occurs when radio waves are reflected from the ionosphere. Practically all HF communication is done by skywave. In the ionosphere, the waves are really refracted twice, and they just appear to be reflected. The reflections are frequency sensitive, meaning each ham band reflects differently from the others. Low frequencies, such as 80 meters, reflect mainly from the lower levels of the ionosphere and the reflected signal comes nearly straight back down. This causes 80 meters to propagate to points from local out to more than a few hundred miles in the daytime. At night, when the D layer and E layer are absent, signals striking the ionosphere at lower angles may propagate many thousands of miles on 80 meters. On the bands from 20 to 10 meters, high angle signals pass straight through the ionosphere and do not reflect back down to the nearby stations. The low angle signals on these higher bands reflect from the ionosphere near the horizon and return to the Earth some miles away. The in-between region cannot hear the transmitted signals nor can you hear signals coming from this region. The in-between region is called the "skip zone." Only when the ionosphere is weakly ionized do you have a skip zone on 80 meters.

Another interesting type of skywave propagation seen on the higher HF bands is called chordal hop propagation seen frequently in trans-equatorial (TE) propagation, which is propagation crossing the equator. When this occurs, signals entering the ionosphere are trapped inside the F2 layer then they are finally refracted back to earth across the equator thousands of miles away. There is no propagation between the signal entry point and the exit point. This is skip in the extreme. On many occasions, we have worked stations far away across the equator in the southern part of South America and stations in between could not be heard. We have frequently worked VQ9LA in the Chagos Archipelago located in the Indian Ocean. The path to The Chagos Archipelago is across Europe and the Middle East and finally across the equator to his location in the Indian Ocean. One time when he was working Europe and North America at the same time, we could not hear the European stations because our path to him was via chordal hop propagation. Another way of describing chordal hop propagation is to call it ionospheric ducting.

Skywave propagation sometimes produces an effect called “backscatter.” What happens is the radio waves that strike the ionosphere, instead of only reflecting father away from the transmitting station, part of the signal reflects backwards toward the transmitting station. Stations that are too close to hear each other by direct wave can communicate by the backward reflecting waves. Both stations that communicate by backscatter must point their directional beam antennas in the same direction although their direction toward each other may be at some other azimuth. Backscatter will confuse front-to-back measurements of directional beam antennas. This is because, when you turn the back of the antenna toward the station you are hearing, you may be able to hear him on backscatter from a direction opposite from him. You will be hearing him from the ionized atmospheric cloud in the opposite direction. During intense solar magnetic storms, when aurora occurs at high latitudes, stations are able to communicate by backscatter on VHF and UHF by both stations pointing their directional beams toward the aurora. This will be due north for stations in the Northern Hemisphere and due south for stations in the Southern Hemisphere. Audio from aurora backscatter will have a “wispy” sound.

5.6 Greyline Propagation

Greyline propagation occurs when the sun is low in the sky near dawn or dusk, although we have seen greyline propagation occur as early as two hours before sunset or as late as two hours after sunrise. It is often used to work stations on the other side of the world on 160 and 80 meters. For example, at certain times of the year when it is approaching sunset here in the States, the sun will have just risen in Asia or Australia and vice-versa. At that time, radio waves propagate along the semidarkness path that encircles the Earth called the greyline. Both locations must be in the greyline in order to make 2-way contacts. The tilt of the Earth makes the position of the greyline change as the seasons change. Greyline propagation occurs between any two locations for a brief period of a few weeks. Afterwards, different places fall into the greyline. For several weeks in the fall of the year, an interesting example of greyline propagation occurs in the southeastern part of the U.S. On 3915 kHz, the BBC outlet in Singapore can be heard for about an hour before sunset coming in by greyline propagation. Stations to the east hear it before we do. Stations farther to the west can hear the fading signals after it fades out here because the greyline moves as the earth rotates. For those hearing it, the signal fades in, it peaks, and it slowly fades out.

5.7 Long Path Propagation

Long path propagation occurs when signals propagate the long way around the world. It can occur on any band. It usually occurs from stations on the opposite side of the world from you. We have worked South Africa via long path by beaming northwest early in the morning on 20 meters. When this happens, we are working him long path through the nighttime side of the earth. Since at all times half the Earth has daytime and half the Earth has night, long path propagation is determined by whether the signal is propagated through the nighttime path or daylight path. Sometimes the daylight path will bring in stations by long path propagation and at other times the darkness path provides long path propagation. One night on 20 meters, we heard a station in India coming in short path and long path simultaneously, but the short path was stronger. At the same time, California was working India by long path and they could not hear him short path. They were working him through the daylight path, and he was stronger here on the East Coast via the nighttime path.

5.8 160-Meter Propagation

Each amateur band propagates signals differently. The 160-meter (1.8-2.0 MHz) band is our only MW band and it acts similar to the broadcast band. It is primarily a nighttime and wintertime band as it suffers from high summertime static (QRN). Most hams that use this band for nearby contacts use horizontal dipoles or inverted-V antennas. Some hams use vertical antennas on this band to work distant stations (DX). These DX contacts are made in the fall and wintertime at night via F layer or greyline propagation when the static levels are low. Dipoles and inverted-V antennas do not work well for DX on this band.

5.9 80-Meter Propagation

The CW part of this band is called the 80-meter (3.5 4.0 MHz) band and the voice part of the band is known as 75 meters. Like 160 meters, eighty meters suffers from the same QRN in the summertime. Working DX on this band is a popular avocation during the fall and winter. However, 80 meters is used primarily for working nets and ragchewing. Eighty meters is primarily a nighttime band. This band can vary from being open most of the day in years with low sunspot numbers to being closed during the middle of the day in years with many sunspots. Many DX contacts have been made using dipoles and inverted-V antennas, but a vertical with many ground radials will be better.

5.10 40-Meter Propagation

The 40-meter (7.0-7.3 MHz) band has propagation that can act like either 80 meters or 20 meters. It just depends on the stage of the sunspot cycle. During the years with high sunspot numbers, nearby contacts are possible all day. At night, the skip lengthens making contacts possible to those parts of the world where it is still dark. Working DX on 40 meters is a nighttime or greyline event. When the sunspots are low, 40 meters may have long skip during the day, and nearby contacts may be impossible or they may be very weak. During the time when we suffer from low sunspot numbers, many DX contacts are made during early morning, late afternoon, and at night.

If your primary interest on 40 meters is SSB, our 40-meter voice band is a broadcast band in Regions 1 and 3. Region 1 is Europe, North Asia, and Africa and Region 3 is the Pacific, Southern Asia, and Australia. The top part of 40 meters is a voice band in Region 2, which is North and South America. To work SSB on 40 meters at night, you will have to find a frequency between broadcast stations. Strong broadcast stations heard at night begin to fade out slowly as the morning sun rises and moves higher in the sky. As the suns angle declines in the afternoon, the broadcast stations begin to break through the noise becoming stronger as the sun begins to set. It is only in the middle of the day when no broadcast stations are heard on 40 meters.

Since DX stations in region 1 and most of region 3 can only transmit below 7100 kHz, working DX on 40-meter SSB is still possible. Stations in those regions will have to transmit below 7100 kHz. (Australian and New Zealand amateurs can operate up to 7200 kHz.) They call CQ and announce where they are listening in our voice band above 7150 kHz. This is what is called “working split.”

5.11 30-Meter Propagation

The 30-meter (10.1-10.15) band has such a narrow frequency that the only modes allowed here are CW and digital modes. That means no SSB. Propagation here is much like 40 and 20 meters. Unlike 20 meters, this band stays open longer at night during years with low sunspot numbers. During the daylight hours, it has much shorter skip than 20 meters. In the United States, we are allowed only 250 Watts.

5.12 20-Meter Propagation

The 20-meter (14.0-14.35 MHz) band is the best DX band because it is open for long-skip for more hours than any other band and it does not suffer from QRN as the lower bands. In years of high sunspot numbers, short-skip and long-distance DX can be worked at the same time during daylight hours. Although DX is there most of the time, most of the DX worked is at sunrise, sunset, and all night during peak sunspot years. During the years of low sunspots, it is common to work into Europe and Africa during the day and into Asia and the South Pacific during the evening hours and early at night. Low sunspot numbers cause 20 meters to go dead for east to west contacts at night an hour or so after sunset, but there is some TE propagation. During periods of moderate sunspot numbers, the propagation on this band is a blend of propagation of low and high sunspot years.

5.13 17-Meter Propagation

The 17-meter (18.067-18.167 MHz) band propagation acts much like 20 meters except it is affected more by low sunspot numbers than 20 meters. In periods of low sunspot numbers, this band does not stay open as late as 20 meters, fading out as the sun begins to set. Yet, the 17-meter band does stay open all night when the sunspot numbers are high. The propagation on this band is like a blend of 20 meters and 15 meters, but it is closer to 20 meters. Most users of this band use dipoles and other simple antennas since triband beam antennas wont work here.

5.14 15-Meter Propagation

15 meters (21.0-21.45 MHz) is a fantastic DX band during the high sunspot years. This band may be open for 24 hours, and it is common to work more than 100 countries during a contest weekend on this band. Many have worked more than 300 different countries on 15 meters. In years of low sunspot numbers, 15 meters may be completely dead for several days in a row. When it opens during those years, you may hear only the Caribbean, South America, and on rare occasions the extreme southern part of Africa via TE propagation.

5.15 12-Meter Propagation

The 12-meter (24.89-24.99 MHz) band is much like 15 meters, but it is affected more by sunspot numbers. Because this band is little used, many hours can pass without hearing any amateur signals. Occasionally you will hear South American Citizen Band “pirates” on lower sideband. It is mostly a daytime band but openings to Asia and the South Pacific are common early at night during peak sunspot years. The reason this band is little used is that tri-band beam antennas don't cover this band.

5.16 10-Meter Propagation

The 10 meter (28.0-29.7 MHz) band that is most affected by the sunspot numbers. You may have noticed in this discussion, the higher the frequency, the more it is affected by sunspots. During peak sunspot years, 10 meters can be open some days for 24 hours. Mostly it is a daytime band. When they are at the peak, the sunspots enable you to work worldwide with power as low as 5 Watts. A 10-meter confirmed country total of over 250 is common. In the low sunspot years, the band can be closed for days. 10 meters can open for very short skip by sporadic E propagation during the summer months. Very short skip means contacts as close as 200 miles out to 1000 miles. Sporadic E propagation can suddenly occur without regard to the sunspot numbers.

6.

Standing Wave Ratio

A standing wave ratio bridge is used to measure the standing wave ratio, or SWR. SWR is an indication of how well the radiating part of an antenna is matched to its feed-line or how well the tuner is matching the antenna system. Most amateurs pay far too much attention to SWR. An SWR reading below 2:1 is acceptable, because the mismatch is so small that the feed-line loss can be ignored. If you are using a modern transceiver, its power may fold back to a lower power output above this SWR level.

When you have mismatch between the feed-line and the antenna, part of the power feeding the antenna system reflects back toward the tuner and the transmitter. The part of the power going toward the radiating part of the antenna system is called forward power. The part reflected back down the feed-line is called reflected power. The larger the mismatch the larger the reflected power will be.

If the feed-line and antenna are not matched, waves traveling toward the radiating part of the antenna system meet the waves being reflected back down the feed-line. The waves interfere with each other, and at certain points along the feed-line, the amplitudes of both waves combine. This will result in a current maximum to be found at that point; and at that point, the current will appear to be standing still. The length of feed-line and the frequency will determine where this point occurs. At another point, the forward and reflected waves interfere, and they subtract from each other. At that point, there will be a current minimum. If you could visualize this phenomenon, you would see a series of current maximums and minimums standing still along the feed-line. This is why we refer to them as standing waves. At different points along the feed-line, where you have high current, you will have low voltage, and where you have low current, you will have high voltage. At any point along the feed-line, multiplying the voltage times the current will equal the power in Watts. When the feed-line is matched to the antenna, current and voltage remain the same all along the feed-line because there is no reflected current to interfere with the forward current.

As happens with the current, the voltage will also appear to be standing still. The voltage maximums and voltage minimums will not be at the same locations as the current maximums and minimums. SWR is the ratio of the maximum voltage to the minimum voltage on the line. It is called *Voltage Standing Wave Ratio* or VSWR, but we shorten it to just SWR. There is also a current SWR or ISWR, and it is the same value as the VSWR. For example, if the standing wave voltage maximum is 200 volts and the minimum voltage is 100 volts, the VSWR will be 2:1. If the voltage maximum and voltage minimum are equal, the SWR will be 1:1. If the voltage minimum is zero, the SWR is infinite.

In measuring SWR at the transmitter, you need to realize that feed-line losses affect the SWR readings. If the feed-line losses are high, much of the power reflecting back from the antenna will be lost, and the SWR reading on the meter will indicate it is lower than it actually is. If a feed-line is so lossy that it consumes all forward and reflected power, it will measure an SWR of 1:1.

When measuring SWR on an antenna having a small amount of reflected power, the length of the feed-line between the bridge and the antenna may affect your SWR reading. An example of this is a 70-ohm antenna being fed with 50-ohm coax. Different lengths of feed-line will give you small differences in SWR readings because at certain lengths, the mismatched feed-line starts to act like a series matching section. In the case of a 70-ohm antenna fed with 50-ohm coax, if the feed-line is a half wave long, the SWR will measure 1.4:1. At some particular length of feed-line and on one frequency, the SWR will measure 1:1 because that length of that feed-line transforms the impedance to make a match. Some hams have adjusted their feed line length to get a perfect match. This is called “tuning your antenna by tuning your feed-line.” With other feed-line lengths, you will measure something different. Suppose the impedance of the feed-line and the antenna are perfectly matched. Then there is no reflected power. You will get a 1:1 reading on the SWR-bridge with any length of feed-line.

There is a myth that reflected power is burned up as heat in the transmitter. The reflected power coming back down the feed-line sees an impedance mismatch at the transmitter or tuner and it reflects back up again. The reflected power does not get back into the transmitter. Because the reflected power reflects back and forth, the radiating part of the antenna system absorbs most of the power being reflected back up each time. All of it eventually is radiated except for the power lost in the feed-line. The losses in a real feed-line will burn up some of the power on each pass. This is why the feed-line loss increases with SWR.

Built-in tuners are found in most modern transceivers. If yours doesnt have one, then you can use an outboard tuner to give the transceiver a proper load. The place you want a 1:1 SWR is between the output of a transceiver and antenna or between the transceiver and the input of a tuner in order for the transmitter to deliver its maximum power. Because built-in tuners are in most modern transceivers, many hams use them to match antenna systems having high loss.

7.

Real Antenna Systems

In this book, we will be talking about the losses that rob an antenna of its maximum performance. The ideal antenna system will radiate 100% of your transmitter power on all bands without a tuner and in the direction you want to work. Such an antenna system does not exist. Many new hams succumb to antenna advertisements making claims that are exaggerated. No antenna will have low SWR, work all bands without a tuner, and radiate efficiently at the same time. A dummy load has a low SWR and will load up on all bands, but it will not radiate a signal. A resonant coax-fed dipole antenna will have a low SWR and will radiate efficiently on the band for which it is resonant, but it will not work well on all bands. For example, if the tuning range of your tuner has a sufficient range, you will be able to load up any antenna with it, but it will not necessarily radiate a signal efficiently. It may have high tuner and feed-line losses.

When you choose an antenna, you must decide how much loss you can accept. DXers and hams that work weak signals at VHF frequencies try to eliminate as much loss as possible. If your contacts are going to be made under good band conditions and without much interference, you can get by with high losses. In that case, coax-fed antennas used on bands where they are not resonant will allow you to make contacts. You can be greatly surprised by how little radiated power can be used to make contacts under ideal conditions. If you want to make contacts regularly under changing band conditions, you will want to eliminate as much loss as possible and use antennas with gain. Lower loss will enable you to hear weaker signals.

Nothing will take the place of resonant half-wave dipoles, not because they radiate more efficiently, but because they dont require lossy tuners and dont have high coax losses. Remember that all antenna systems have compromises.

8.

Half-Wave Resonant Dipole Antennas

Most dipoles consist of two pieces of wire of equal lengths with one of the two ends connected together through an insulator. The far ends of the wires are also connected to insulators. The two conductors of a feed-line are separated and connected across the gap at the center insulator. The antenna is held up by rope that connects the insulated ends of the antenna to two supports. It is a “balanced” antenna, because equal currents flow on both halves of the antenna. Coax is an unbalanced feed-line. (The possible effect of using an unbalanced feed-line on a balanced antenna like a dipole will be discussed later.) The dipole that is stretched between two high supports is called a flattop dipole, distinguishing it from other configurations.

8.1 The Half-Wave Flat-Top Dipole

The simplest antenna system of all is the half-wave resonant dipole fed with coax and no tuner. The only reason for using a half-wave resonant dipole antenna is to eliminate the need for a matching device such as a tuner. The feed-point impedance will be near 50 ohms at ordinary heights and they can be fed directly with 50-ohm coax from the output of todays modern radios. The two halves of a dipole are fed 180 degrees out of phase, meaning when one side is fed positively, the other side is fed negatively. That is why a feed-line has two conductors. Of course, the sides swap polarity on each half cycle.

If you could visualize the current flowing on the half-wave dipole, the current will appear to be standing still. The maximum current will be seen at the center of the wire and no current will be at the ends. This occurs because the electrons flowing out to the ends reflect back toward the center where they meet the next wave and the current is reinforced there. The minimum voltage occurs at the center and the maximum voltage occurs at the ends of the half-wave resonant dipole. If you were to measure the voltage and the current at any point on the dipole wire, the voltage times the current will equal the power in Watts.

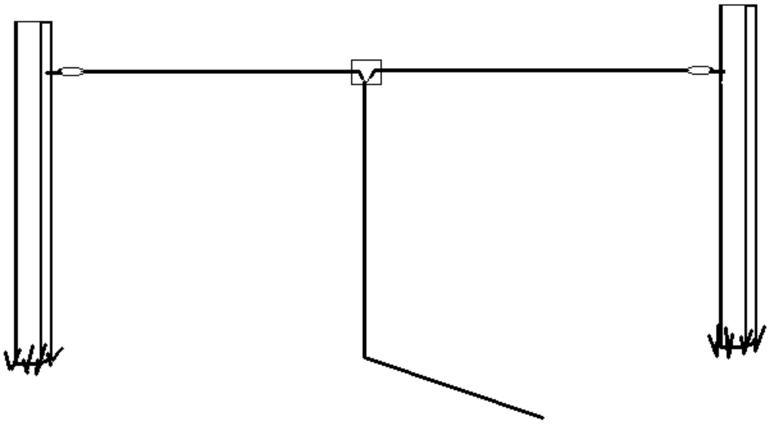

Figure 8.1: Flat Top Dipole

8.2 Inverted-V Dipole

Another configuration for the half wave resonant dipole is one having one support in the center and the ends stretched down toward the ground. The single support can be a tree, mast, or tower. The ends of a dipole have high RF voltages on them, and need to be at least 10 feet above ground for safety. This antenna is called an "inverted-V," because the shape of the dipole looks like a "V" turned upside down. Most dipoles illustrated in this book can be put up in the inverted-V configuration. This configuration works well because the current is concentrated on the middle two-thirds of the antenna at the apex. The current in an antenna is what is responsible for the radiation. The ends of the antenna have very little current in them and it doesnt matter if the ends are close to the ground. The middle of the antenna is up high where the radiation is taking place and that is the place you want the radiation to be. An inverted-V has an advantage that the horizontal space required for it is less than what is needed for a flattop dipole. The angle between the wires on an inverted-V needs to be greater than 90 degrees. The gain of an inverted -V is 0.2 dBd and it has a radiation pattern nearly omni-directional. Since it is easy to construct and works so well, the inverted-V is the most commonly used dipole. An explanation of the decibel will come later.

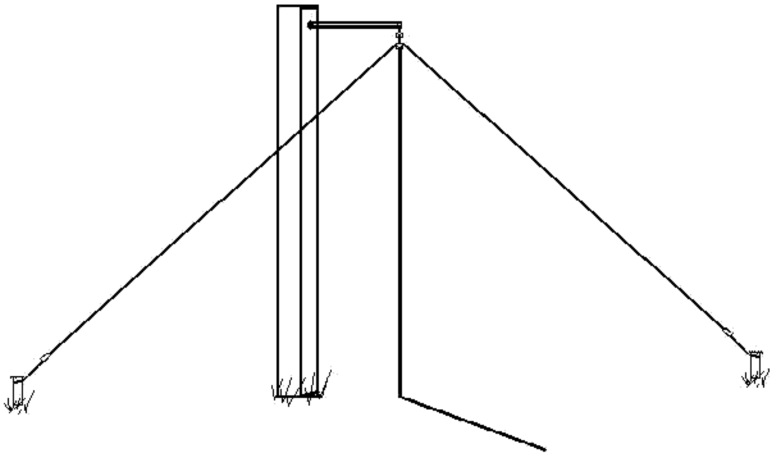

Figure 8.2: The Inverted-V Dipole

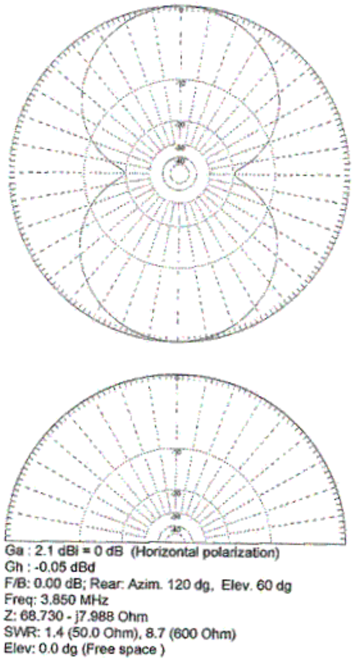

Ga : 2.1 dBi = 0 dB (Horizontal polarization)
Gh : -0.05 dBd
F/B: 0.00 dB; Rear: Azim. 120 dg, Elev. 60 dg
Freq: 3.850 MHz
Z: 68.730 - j7.988 Ohm
SWR: 1.4 (50.0 Ohm), 8.7 (600 Ohm)
Elev: 0.0 dg (Free space)

Figure 8.3: Radiation Pattern of Inverted-V for 80-Meters at 65 Feet

In *Figure 8.3: Radiation Pattern of Inverted-V for 80-Meters at 65 Feet* on page 19, the top graph shows how the radiation would appear to you, if you were situated above the dipole and you were looking down on it. The plane of the antenna runs from side to side on the top graph, and that graph demonstrates only a 5-dB null off the ends of the antenna. Therefore, it is essentially omnidirectional. The bottom graph shows how the radiation would appear if you were looking at the antenna from the end of the wire. As you can see, the pattern shows no radiation at the horizon and its maximum radiation is at about 40 degrees above the horizon, and the radiation straight up is only down 3 dB from its maximum. This antenna was modeled on 80 meters with the apex at 65 feet above ground and the ends at 35 feet.

It is a myth that a horizontal antenna orientation makes a difference on 80 meters at heights used by most amateurs. I have heard many amateurs say on 80 meters, “The reason my signal is weak to you is because you are off the end of my dipole.” The radiation pattern from a dipole is essentially non-directional until the dipole is elevated more than a half wave, that is about 125 feet on 80 meters, and it is 65 feet on 40 meters. The main reason it makes no difference regarding orientation is because propagation for signals closer than 500 miles (the distance of most 80 meter contacts) is essentially by high angle radiation nearly straight up and down. Only signals radiated and received at low angles make a difference in antenna orientation even at low heights above ground. At low heights, there are nulls about 3 to 4 dB off the dipole ends.

8.3 Dipole Shape Variations

The wire of a dipole doesnt have to be run in a straight line. A dipole does not have to be perfectly horizontal. Thats the way it is usually depicted in books and magazines, but you can bend the legs of the antenna up, down or sideways.

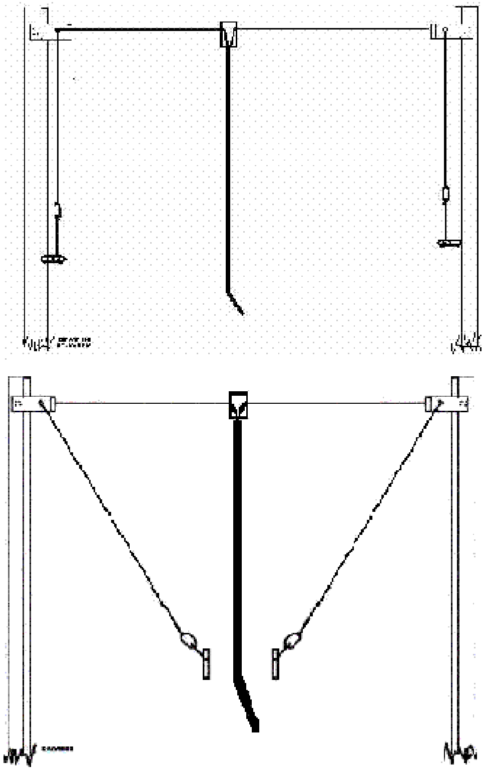

Figure 8.4: Two Dipole Shape Variations

If you make either wire one-half wavelength long and carefully prune it to resonance, you can use it without a tuner on and near its resonant frequency. Both antennas have the current part at the top where most of the radiation takes place. The vertical parts of these antennas radiate a weak vertically polarized wave. The only reason these dipoles are contorted this way is to make them full-sized and to fit in the available space. Other shapes are possible, and you can be creative at your location.

There are many more dipoles than the ones just described. We explore the other kinds of dipoles in *Section 8. Half-Wave Resonant Dipole Antennas* on page 17.

8.4 Calculating the Length of a Half-Wave Resonant Dipole

The approximate length in feet of a half-wave resonant dipole is found by dividing 468 by the frequency in MHz. The actual length of it will be determined by several factors. Using larger diameter wire will make the dipole resonate lower in frequency. Therefore, to make it resonant at the higher desired frequency, It must be shortened. Raising a dipole higher above ground will make it resonate higher in frequency. An insulated wire will make the dipole resonate lower in frequency than a bare wire.

Using the above formula, cut the antenna a little longer than the calculations say. If the SWR is best at a lower frequency than you desire, the antenna will have to be made shorter by pulling the excess wire through the end insulators, folding the ends of the extra wire back on itself. Then wrap the ends of the overlapped wire on itself so it wont come loose. This causes the excess wire to “short” itself to the rest of the antenna. If you are using insulated wire, you will need to cut off the excess wire. The reverse is true if the antenna resonates too high in frequency. The extra wire can be let out to make it resonate on a lower frequency. This is why you originally cut the wire a little longer.

8.5 The Decibel

The decibel (dB) is a unit of measurement for comparisons of the ratio of power, current, and voltage and is the term we will use in comparing antennas in this book. At one time, antenna comparisons were made using a dipole as a standard, but today most comparisons use the isotropic radiator as a reference. An isotropic radiator is an imaginary antenna that radiates equally well in all directions. It has no gain. The terms “dBi” and “dBd” are used to label which reference is being used. In this book, we will use the dipole as a standard for the most part.

How do you derive decibels from power ratios? The formula for power ratios is dB = 10 log P1/P2. For voltage and current, the values are doubled. Formulas of this type are beyond the scope of this book. Doubling the power will produce a 3 dB stronger signal. Double the power and double it again will equal a 4 times power increase and that gives 3 dB plus 3 dB or 6 dB. Double 4 and that is a power increase of 8 and that adds 3 more dB for a total of 9 dB. Increasing the power from 1 Watt to 10 watts or increasing it 10 times will give a 10-dB increase. Multiply 10-Watts times 10 give us 100 watts, which adds another 10 dB above 1 Watt for 20 dB. Therefore, increasing the power another 10 times to 1000 Watts will produce a signal 30 dB stronger than 1 Watt.

Your receiver, if modern, will have a signal strength meter or “S Meter.” That meter is calibrated in “S-Units” from one to nine and decibels over S-9. S-9 is usually calibrated using 50 microvolts (uV) from a signal generator. Each S-unit is approximately a difference of 5 or 6 dB. Therefore, a reading of S-9 is about 6 dB stronger than S-8. Therefore, from S-0 to S-9 is 54 dB. On some low cost transceivers, the S-units and dB above S-9 are only relative signal readings and actually have nothing to do with decibels.

9.

Antenna Basics

9.1 Resistances and Reactance

Two factors measurable in antenna impedance are resistance and reactance. When we refer to antenna resistance, we are referring to its radiation resistance. It is neither a resistance like the electronic component called a “resistor,” nor is it the same as the resistance found in all conductors. Those types of resistances, called “loss resistances,” change electrical energy into heat energy. Heat energy disappears by radiating out into its surroundings and it dissipates away to infinity. When we feed RF into the antenna, the energy put into the radiation resistance disappears from the antenna by radiation of electromagnetic waves, and that makes an antenna appear to have a resistor in it. Loss resistance robs power from the radiation resistance and lowers the efficiency of an antenna system, but the loss resistance in dipoles is very low if the feed-line loss is low. The efficiency of any antenna system is found from a ratio of radiation resistance and loss resistance. We can either calculate the loss resistance by the loss in the feed-line from published tables and by estimating the loss in tuning units. Feed-line loss and tuning unit loss can be measured, but that is beyond the scope of this book.

Antenna systems having reactance prevent the transmitter from delivering its full power and the reactance needs to be tuned out. There are two kinds of reactance: capacitive and inductive. Antennas have both. In antennas, reactance is a virtual reactance meaning the antenna acts as if there were a capacitor or an inductor in the antenna, but neither is there. You can only measure the sum of both reactances but not a value for either one. Using an antenna analyzer, you can determine whether the sum of the reactance is inductive or capacitive. Inductive reactance is a negative number and capacitive reactance is a positive number.

The reactance of an antenna forms the “J” factor in antenna impedance measurements. The “J” factor is measured in ohms and the reactance is expressed as + or “J” ohms depending on whether it is capacitive or inductive reactance. Capacitive reactance is expressed as +J ohms and inductive reactance is expressed as -J ohms. Capacitive and inductive reactance are opposite factors and one can cancel the other. An antenna having 6 ohms capacitive reactance or + J 6 ohms and an inductive reactance of J 5 ohms will result in an antenna with a reactance of 1 ohm capacitive or + J 1. Since one term is positive and the other term is negative, you subtract smaller value from the larger. The answer has the sign of the larger one. In antennas, the reactance and resistance together determine the overall impedance of the antenna. The J factor is mentioned here only because you may see it in other books and on the extra class examination, but it will not be used further here.

A resonant antenna has equal amounts of inductive and capacitive reactance, and the sum of the reactance equals zero. As an example, when the inductive reactance equals J 5 and the capacitive reactance equals +J 5, their sum equals zero. When the sum of the total reactance of an antenna is tuned to zero, its impedance is totally resistive. The use of an antenna analyzer will tell you if the antenna is too long or too short for resonance. The simplest way to tune out antenna reactance is to change its length. The sum of the reactance of a long antenna will be inductive, and the sum of the reactance of a short antenna will be capacitive. If an antenna is short because it wont fit your property, it can be tuned to resonance by putting an inductor (coil of wire) in each leg. These coils are called “loading coils.” An equal amount of inductive reactance will cancel the excessive amount of capacitive reactance. An antenna with loading coils is described in section “X.” When an antenna is too long, the sum of its reactance will be inductive, and a variable capacitor can be inserted in each leg to tune out the inductive reactance. This is seldom done because it is easier to shorten the antenna.

A resonant antenna may still have SWR if its radiation resistance is not exactly 50 ohms. Not many resonant antennas have a radiation resistance of exactly 50 ohms, and most real antennas have a small amount of SWR. An antenna is resonant only at one frequency per band. It will also be resonant on its harmonic frequencies, where its radiation resistance will range from high to very high. Hams talk about using resonant antennas. What is meant by this is they use an antenna on its fundamental frequency close to resonance, the resistance is near 50 ohms, and the SWR without a tuner is near 1:1.

To calculate the impedance of an antenna with both resistance and reactance requires a mathematical procedure called the Pythagorean Theorem. That type of math is beyond the scope of this book. However, you should know how to use the Pythagorean Theorem to solve impedance problems on the Extra-Class test. Otherwise, you will have to memorize the answers from the question pool.

9.2 Feeding Dipoles Efficiently

For maximum power transfer from transmitter to the antenna, the antenna system must be resonant, and the resistance of the load (antenna system) has to be equal to the internal resistance of the source (transmitter). Notice we said an antenna system, not the antenna, must be resonant. As mentioned previously, an antenna system consists of the antenna, the feed-line, and any matching networks (tuners). A tuner at the input end of the feed-line can make a non-resonant antenna system resonant, and have a resistance of 50 ohms, and that matches the internal resistance of the transmitter. A tuner will not change the SWR between the tuner and the dipole part of an antenna system, and will not remove the reactance from the dipole.

When the load of an antenna system does not match the source and the impedance is high, the load will not draw power from the source and high RF voltages will be present at the output of the final transistors. In this case, high RF voltages can damage the output transistors of the transmitter. When the impedance of the load is low, too much of the power may be dissipated across the internal resistance of the transmitter possibly destroying the output transistors. These are the two reasons why transceivers “fold back” their power when the SWR is high.

It is a myth that the dipole part of an antenna has to be resonant to be efficient. When power reaches the radiating part of the antenna system, it obeys the “The Law of Conservation of Energy.” The Law of Conservation of Energy states, “Energy can neither be created nor destroyed. Only its form can be changed.” (What is important is to get the power to the dipole itself, because in some systems power is lost in the feed-line, especially when using coax with high SWR) The miniscule amount of power in the dipole that does not radiate is changed into heat, another form of energy. Because the dipole part of an antenna system is made of conductors with low loss resistance, 99% or more of the power reaching it will radiate regardless of its length if that length is reasonable. The loss resistance of the conductors of the radiating part of most antenna system is so low it can be ignored. (Short mobile HF antennas are an exception because they may be lossy because of the very high current flowing in them.)

Not all the energy fed into an antenna system will reach the antenna itself. If the system has a tuner, part of the power is lost in the inductor of the tuner and part is lost in the feed-line. When properly tuned, tuners using T-networks lose about 10% of the power and L-network tuners lose about 5% of the power being fed to them. Notice we said properly tuned. However, improper tuning of the antenna tuner may cause you to believe the feed-line is matched, but when this happens there is a very high circulating current in the inductor causing it to get hot. This causes extremely high losses, and very little power reaches the radiating part of the antenna. In addition, so much heat is produced in the inductor that it can be damaged. We melted the plastic insulation that forms the inductor on one tuner this way. For this reason, some hams dont like tuners, preferring to use resonant antennas. Read the instructions for your tuner for proper tuning or you may wind up with a poor signal and a damaged tuner. The resistive losses in the conductors of the feed-line and the dielectric losses in the feed-line also rob power from the system. These are the reasons for you to use the best tuners and feed-lines possible.

Another loss to be considered is feed-line radiation. Any energy that radiates from the feed-line does not reach the radiating part of the antenna, and it may be absorbed by near-by objects and may not radiate in the desired direction. When coax radiates, it is called common-mode radiation. If the feed line can radiate, it can also receive signals. This can be detrimental because the coax can then pick up noise from near-by power lines, etc. Feed-line radiation will also destroy the directional pattern of a beam antenna. The causes of feed-line radiation will be described in the next section.

As we pointed out earlier, when you are using a half-wave resonant dipole fed with low-loss coax without using a tuner, almost all of the power coming out of the transmitter will radiate. On its resonant frequency, the dipole is one of the most efficient antenna systems a ham can use. However, a half-wave resonant dipole has a finite bandwidth. Why use a tuner with resonant antennas? On 160 and 80 meters the bands are wide compared to the percentage of frequency. The width of 80 meters is 500 kHz and its frequency is 3500 kHz. The width of 80 meters is 14% of the frequency. The 350 kHz of 40 meters is 5% of the frequency and most of the band can be covered without a tuner. The 350 kHz width of the 20-meter band is 350 divided by 14000 kHz, or 2.5 % of the frequency, etc. The percentage of frequency for a band will determine if a resonant dipole will work the whole band without a tuner. If you are planning to move around on 160 or 80 meter bands, it makes sense to have a tuner, because the bandwidth of resonant dipoles on those two bands is narrow. For example, the normal 2:1

SWR bandwidth of an 80-meter dipole is less than 200 kHz and the band is 500 kHz wide. However, if you have an antenna resonant for the voice portion of the band, you can still use a tuner to work the CW part of the band without inducing more than a dB of loss. Except for 40 and 10 meters, full-sized resonant dipoles on the rest of the HF bands will have enough bandwidth for them to cover the whole band.

The best place to insert a tuner is up at the antenna feed-point. However, if it is placed there, you wont be able to reach the tuners controls. Therefore, it is more practical to place it between the transceiver and the shack-end of the antenna feed-line. A piece of 50-ohm coax connects the radio to the tuner. With the tuner located in the shack, adjustments can be made. Remote automatic antenna tuners can be placed at the antennas feed-point, but the disadvantage of them is that the ones available today will not handle high power.

A coax-fed dipole and a tuner should not be used to feed an antenna on its even harmonically related bands. The even harmonics are 2, 4, 6, etc, times the fundamental resonant frequency. If an 80-meter antenna being fed with coax through a tuner is used on 40 meters, it will put out a weak signal because the SWR will be around a hundred to one. Coax has a tremendous loss with SWR this high. Only a few Watts from a hundred-Watt transmitter will reach the antenna. However, you will be able to make contacts with those few Watts. If you want to use any antenna having high SWR, ladder-line has much less loss than coax. If you feed an 80-meter dipole on 40 meters using ladder-line and a tuner, it will only be slightly less efficient than a half-wave 40-meter coax-fed resonant dipole. However, the SWR will still be high between the tuner and the antenna, but this doesnt matter since ladder-line has an insignificant loss. Since the feed-point impedance will be high, the SWR will only be about 9:1 in the ladder-line because ladder-line is a high impedance feed-line.

Extremely short antennas may not work at all because of the above mentioned reasons. To reiterate, the extremely high capacitive reactance may make it impossible for its reactance to be tuned out and reactance prevents a transmitter from delivering power to the antenna. Even if you are able to tune out the capacitive reactance, tuning it out requires an inductor and most of the power will be lost in the inductor. Do not take the statement about the Conservation of Energy to mean you can put up any piece of wire and it will radiate your entire signal.

9.3 The Cause of Feed-Line Radiation

Contrary to popular myth, SWR in a feed-line will not cause it to radiate. The cause of feed-line radiation is unequal current in the two conductors of the feed-line. What are the causes of unbalanced current in a feed-line? They are an unbalanced feed-line feeding a balanced antenna; the feed-line being brought away from and parallel to one leg of the antenna; the antenna not being fed in its center; and one leg of the antenna being close to metal objects. In coax, unbalance causes RF to travel on the outside surface of the coax shield, and the shield radiates. When everything is balanced, coax normally has current flowing on its center conductor and on the inside of its shield. The shield prevents it from radiating.

Ladder-line will also radiate when it is fed from the output of a tuner not having a balun. Baluns are discussed in the next section. Since the output of a transceivers tuner is unbalanced and feeding ladder-line directly from your transceivers tuner, the currents in the ladder-line will not be balanced. When balanced, ladder-line has equal currents with a 180-degree phase difference, which produce waves that null each other out, and no radiation takes place. Hams mistakenly refuse to bring ladder-line into the shack because of a fear of feed-line radiation, but ladder-line does not radiate when balanced. The simple cure for feed-line radiation is to use a balun at the antenna feed-point for coax and a balun at the output of the tuner when using ladder-line.

9.4 Baluns

The word “Balun” is a contraction of “balanced to unbalanced.” It is pronounced “bal un” like “bal” in “balanced and like “un” in “unbalanced”. Many hams mistakenly pronounce an “M” at the end of the word making it “balum.” A balun transforms the unbalanced transmitter output to a balance feed-line such as ladder-line. It is also used to connect an unbalanced feed-line such as coax to a balanced dipole. In the latter case, the balun is located at the antenna feed-point and is constructed so the balun takes the place of the center insulator.

There are two kinds of baluns: voltage baluns and current baluns. They both accomplish the same thing. The difference in baluns is in the way they are wound. A voltage balun produces equal voltage with opposite polarity at its output. As its name implies, a current balun provides equal currents with opposite polarity at its output.

Running the coax through ferrite beads can make a 1 to 1 current balun. In addition, you can build a 1 to 1choke current balun by winding 8 to 10 turns of coax around a two-liter soda bottle and placing the coiled coax at the antenna feed-point. Any balun is designed to “divorce” your antenna from the feed line. It is used to prevent common mode radiation of coax, which makes the coax to be part of your antenna. You want it to be able to deliver all your power to the radiator itself. A choke balun does this perfectly, without using any ferrite beads or toroids. In most cases common mode coax radiation does not occur when a balun is not used, but it is preferable to use one to be safe.

Other baluns provide a step-up or step-down impedance transformation. A 4-to-1 balun steps up the impedance four times. It will transform a 50-ohm impedance to 200 ohms. This type of balun transformer is used at the output of tuners to increase the tuning range of a tuner 4 times. If a tuner without a balun can match 500 ohms, a 4-to-1 balun will increase the range of impedances it can match to 2000 ohms. Many hams think the 4-to-1 balun is used to match 50 ohms to 450-ohm ladder-line but it is not. It would take a 9-to-1 balun to match 50 ohms to 450 ohms, and it is not important to match the impedance to ladder-line.

A balun should always be placed at the input end of ladder-line or open wire feeders to prevent feed-line radiation. When using ladder-line a step up balun is commonly used although a 1:1 balun will work.

10

Other Types of Dipoles

10.1 A Shortened Dipole Using Loading Coils

If you are unable to put up a full-sized dipole on your property, putting loading coils into the dipole could shorten the antenna. See *Section 11.1 End-Fed Zepp*. A short antenna has capacitive reactance and the capacitive reactance can be tuned out with a coil. The overall length of the shortened antenna will be determined by the amount of inductance in the coil. Pre-tuned antennas of this type are available from at least one manufacturer. The main problem with loaded antennas is they are very narrow banded. If the loading coils are wound with small diameter wire, the coils may introduce unwanted loss into the antenna. Loading coils can also be found in shortened vertical antennas for high frequency (HF) mobile use.

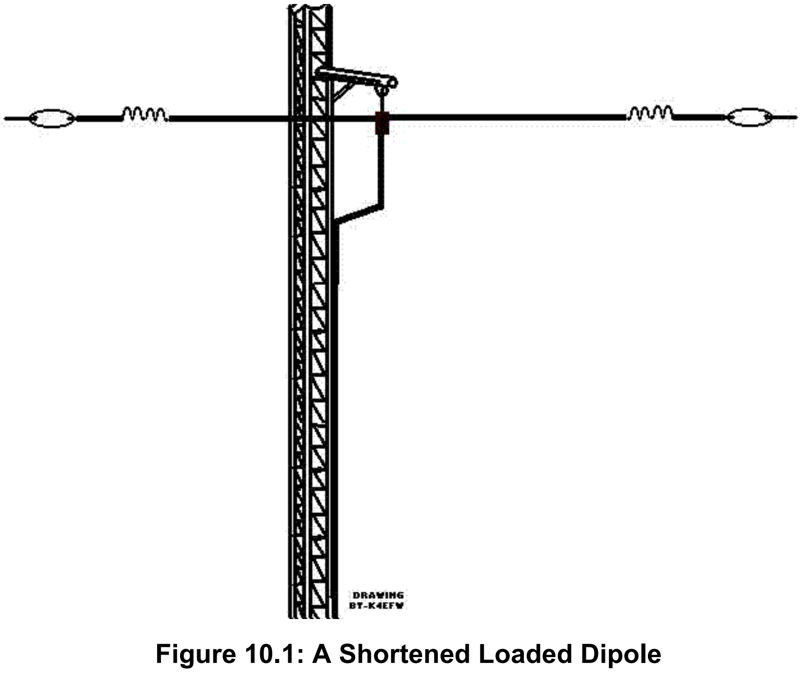

Figure 10.1: A Shortened Loaded Dipole

10.2 All Band Dipole

In *Figure 10.2: All Band Dipole*, a dipole is cut to a half wave on the lowest band you want to operate. Feeding it with ladder-line and a tuner makes it possible for you to work all the other higher bands. The only losses in this antenna system are the loss in the tuner and the very small loss in the ladder-line. This system is more than 90% efficient. As mentioned above the balun in the tuner will be used, or if your tuner doesnt have a balun, an external balun can be connected between the tuner and ladder-line with a short run of coax. Four-to-one baluns are the most commonly used ones for this arrangement.

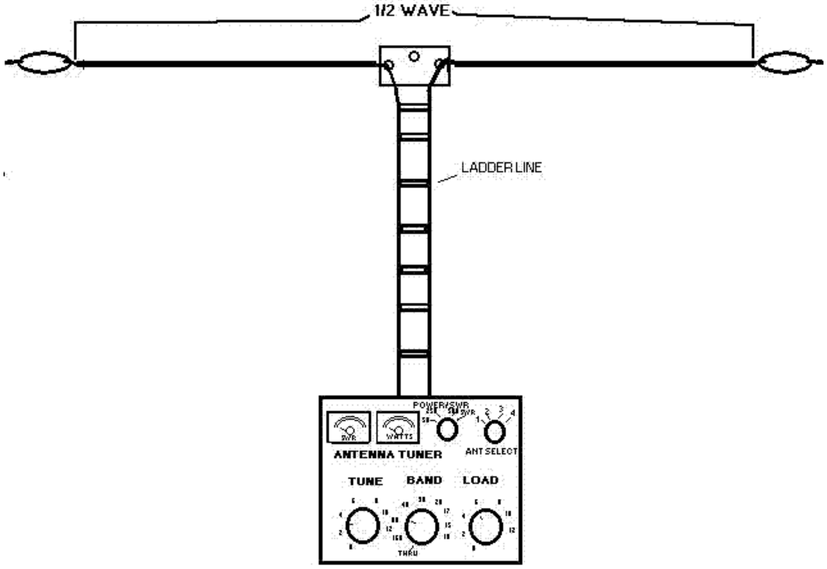

Figure 10.2: All Band Dipole

10.3 The Sloping Dipole

A lower angle of radiation can be achieved by tying one end of a half-wave dipole to a high support and the other end near the ground. It is fed with or without a balun with 50-ohm coax. The sloping dipole will show some directivity and have low angle gain in the direction of the slope. More directivity can be gained if the dipole is strung from a tower, and the tower is acting as a passive reflector. The sloping dipole is mostly a vertically polarized radiator and it works well for DX. Since the sloping dipole is fed in its center, it does not need to be grounded to the earth as a quarter-wave vertical does. Make sure the bottom end of a sloping dipole is at least 10 feet above ground because like all dipoles there is high RF voltage on its ends.

10.4 Half-Wave Resonant Sloping Dipole

In *Figure 10.2: All Band Dipole*, the field of maximum radiation is in the direction of the slope or toward the right side of the picture. The formula for the length of a sloping dipole is the same for any half-wave resonant dipole.

10.5 The Folded Dipole

The B&W Company makes a folded dipole that claims to have a good match on all bands and it does. However, on the low bands much of the power is burned up in the resistor that connects the two ends together. The power going toward the ends encounter the resistor and is consumed as heat. All that power is lost and does not radiate, and no power is reflected back to the feed point making the antenna have low SWR. On the higher bands, a large part of the power radiates before it reaches the resistor and the antenna is moderately efficient on those bands. On 80 meters the 90 foot-long dipole model will produce a signal at least 10 dB lower than that from a resonant dipole.

If you remember the single channel TV antennas used years ago, the driven element was a folded dipole. Folded dipoles are very broad-banded. That is the reason they were used for TV antennas since a TV channel is 4 MHz wide.

When constructing a folded dipole, the formula for calculating the length of it is the same as for any dipole. The folded dipole consists of two parallel conductors with the ends tied together. The conductors can be spaced from less than an inch to more than two inches apart when made from TV ribbon or ladder-line. At the ends, strip the insulation back several inches, Twist the bare wires together, solder them, and run them through insulators. The feed-point is in the center of only one of the two parallel conductors.

The feed-point impedance of a folded dipole at resonance is close to 300 ohms resistive and can be fed directly with 300-ohm TV twin-lead or a tuner with its balun. This antenna was very popular years ago when coax was expensive and 300-ohm TV twin-lead was relatively cheap. A length of 450-ohm can be substituted for the twin-lead. An alternate feed method is placing a 6:1 balun at the feed-point and then feeding it with 50-ohm coax. The folded dipole will not radiate its second harmonic, so it is not good for a multi-band tuner-fed antenna.

Another folded dipole type is the three wire folded dipole. We have seen this dipole only in books and do not know anyone who uses one. The feed-point impedance is 600 ohms resistive and is fed with home-built 600 ohm open wire feeders.

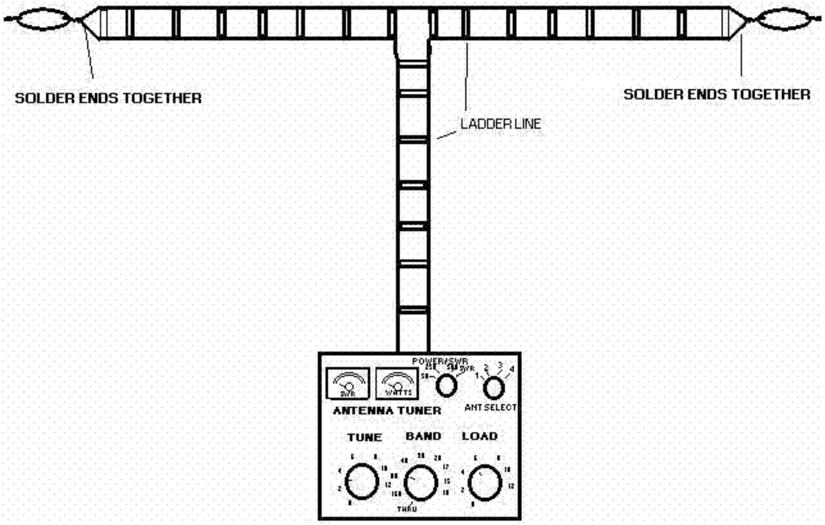

Figure 10.3: Folded Dipole

10.6 The Double Bazooka Dipole

The double bazooka is claimed by its users to be broad-banded, a quality especially interesting for those hams operating on 75/80 meters. Tests done at the A.R.R.L. have shown the double bazooka is only slightly more broad-banded than a regular dipole, probably due to the use of a large conductor (coax) for the center part of the antenna. The double bazooka will not transmit its second harmonic, and its users say it does not need a balun. Other users say it is quieter than a regular dipole.

The center of the antenna is made from RG-58 coax. To find the length of coax needed, divide 325 by the frequency in MHz. The coax forms the center part of the double bazooka and a piece of number 12 wire on each end completes the antenna. The length of each of the end wires is found by dividing 67.5 by the frequency in MHz. To increase the bandwidth some builders use shorted ladder-line in place of the number 12 wire, which makes the end pieces to be electrically larger.

The feed-point of the double bazooka is unique. At the center of the coax dipole, remove about 3 inches of the plastic covering, exposing the shield. Cut the shield in the center and separate it into two parts. Do not cut the dielectric or the center conductor. Leave the center conductor with its insulation exposed. On the feed-line strip off about 3 inches of outer insulation, separate the shield from the center conductor, and strip about 1 inches of the insulation from the center conductor. To attach the feed-line, solder the two exposed feed-line conductors to the two pieces of the separated exposed shield of the dipole center. It goes without saying: seal the feed-point to prevent water from getting in. At each of the two ends of the coax forming the center of the antenna, the coax is stripped back and the center conductor and shield are shorted together and soldered. The end wires are soldered to the shorted coax ends, run to insulators at the end of the antenna, and the soldered joints are sealed against the weather.

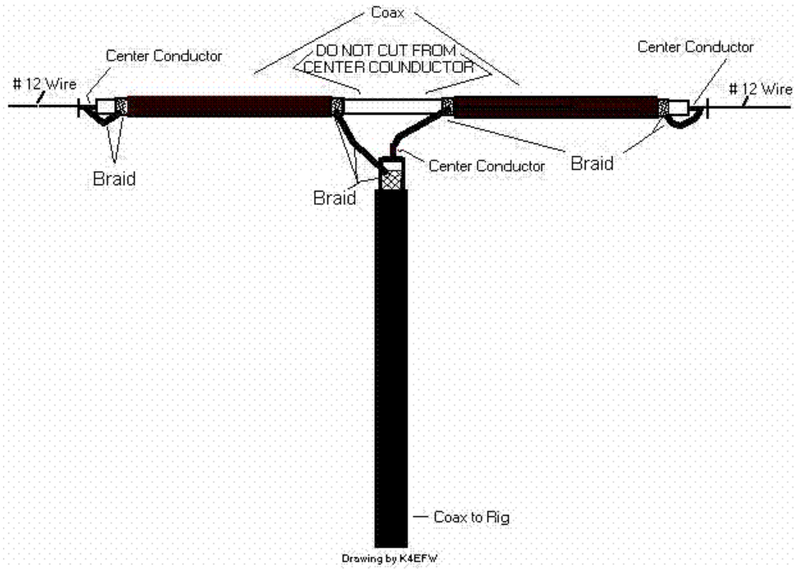

Figure 10.4: Double Bazooka Dipole

10.7 Broad-Banded Coax-Fed Fan Dipole

A broad-banded dipole for 75/80 meters can be constructed by attaching two equal length dipoles to the center feed-point and spreading the ends about 3 feet apart using PVC water pipe to separate them. The completed dipole looks like a bow tie. This makes the antenna to appear electrically to have that of a large diameter conductor. Because of this, the overall length will need to be shorter than a single wire alone. When we used the antenna, we found a length of 110 feet would cover most of the 75/80-meter band without a tuner. It is fed with 50-ohm coax. The use of a balun is optional. The antennas for most of the higher bands have enough bandwidth so they do not need broad banding.

10.8 Two-Element Collinear Dipole

The 2-element collinear dipole is an antenna that is a full-wavelength antenna having a 2-dBd gain. It can be fed with ladder-line and a tuner and used as a multiband antenna, or it can be fed with a quarter-wave-matching stub with 50-ohm coax cable to make it a single band array. In the stub matching system, a quarter wavelength of ladder-line is connected across the center insulator, and the opposite end of the ladder-line is shorted. A shorted quarter-wave piece of feed-line acts like an open circuit. Going from the shorted end of the ladder-line toward the dipole, there will be a point where a piece of 50-ohm cable will find a perfect match. The 50-ohm feed-point will have to be found empirically (trial and error).

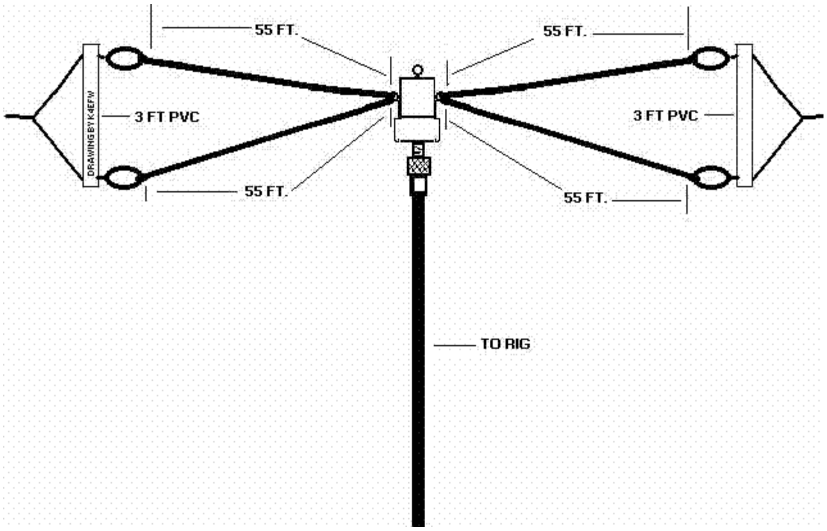

Figure 10.5: Broad-Banded Fan Dipole for 80 Meters

10.9 4-Element Collinear Dipole

The 4-element collinear dipole array consists of four half-wave segments connected end-to-end with an insulator between each two adjoining segments. The feed-point is at the center of the array. The antenna is fed with ladder-line through a tuner. A quarter wave shorted ladder-line stub hangs down vertically from the insulators between the inside and the outside half-wave segments. This stub provides a 180-degree phase shift so that all half-wave segments are fed in phase. This antenna has a 6-dBd gain and it radiates bi-directionally at an angle perpendicular or broadside to the plane of the wires.

This antenna is too long for most hams to use on 80 and 40 meters, and the stubs hanging vertically will be too close to the ground. For 20 meters, the four-element collinear array will be 97 feet long and the stubs will be 18 feet. To find the length of each half-wave segment, divide 468 by the frequency in MHz, and for the quarter-wave stubs, divide 246 by the frequency in MHz.

MFJ has begun marketing the four-element collinear monoband array. They have them for 20, 17, and 15 meters. This antenna is so easy to build that you can do it yourself. All you need is 5 insulators, antenna wire, and some ladder-line.

It will have no gain if you use it on bands for which it is not designed because the stubs are used as phasing lines. It is definitely not a multiband antenna.

It is possible to add more half-wave segments to the ends of this array to make it have 6, 8, 10, etc half wave segments. Adding more segments will add more gain and make the lobes narrower.

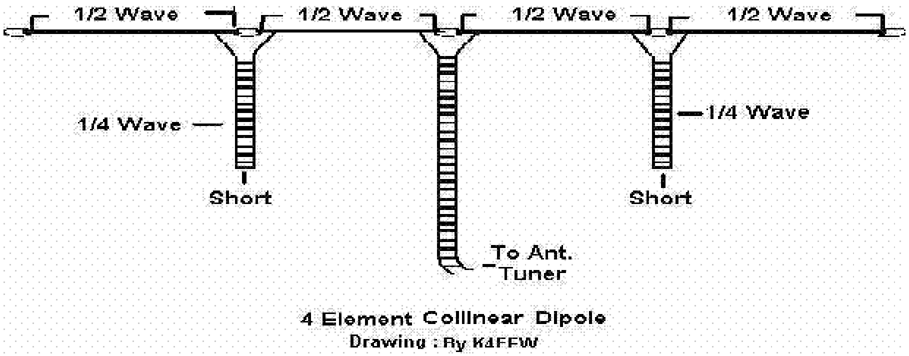

Figure 10.6: 4-Element Collinear Dipole

10.10 Coax-Fed Dipoles Operated on Odd Harmonic Frequencies

Antennas fed with 50-ohm coax can be used on other bands for which they are not cut. An 80-meter dipole will have a relatively low SWR and will be resonant at a single frequency on 10 meters and not

much power will be lost in the coax even if operated off resonance. A 40-meter dipole will work the same way on 15 meters. Using coax, a dipole will work on its fundamental frequency and on odd-harmonic frequencies and it is not necessary to use ladder-line. The fundamental frequency is the frequency for which the antenna is a half-wavelength long, and the odd harmonics are 3 times, 5 times, 7 times, etc. the fundamental resonant frequency. A frequency of 21 MHz is 3 times or the third harmonic of 7 MHz, and 28 MHz is the seventh harmonic of 4 MHz.

Antennas operated on their odd harmonics will be resonant a little higher in frequency than exact multiples of their fundamental frequencies. Since the odd harmonic antennas input impedance is higher than it is on its fundamental frequency, many amateurs use a series quarter-wave matching section of 70-ohm coax to give it a better match. The 80 meter inverted-V dipole in use here has a 2:1 SWR on 10 meters indicating it has an impedance of around 100 ohms. However, modeling the antenna for 10 meters shows the resonance to be below 28 MHz, probably because the antennas fundamental resonant frequency is 3920 instead of 4000 kHz. A quarter wave 70-ohm matching section should bring the SWR down to a much lower level.

As said earlier, if you try to use coax with a dipole on its even harmonic frequencies, the feed-point impedance will be very high, the SWR will be extremely high, and the coax will absorb most of the power. In addition, when operating a coax-fed antenna on its even harmonics, the tuner may not be able to provide a match. Operating any antenna on any of its harmonic frequencies, odd or even, will work better if it is fed with ladder-line and a tuner.

This antenna is matched by a quarter-wave 70-ohm series matching section. Three half waves will resonate higher than you would expect because the center half wave doesnt have to contend with end effects. To calculate the length of a three half-wave dipole, divide 1380.6 by the frequency in MHz. Five half waves is found by dividing 2316.6 by the frequency.

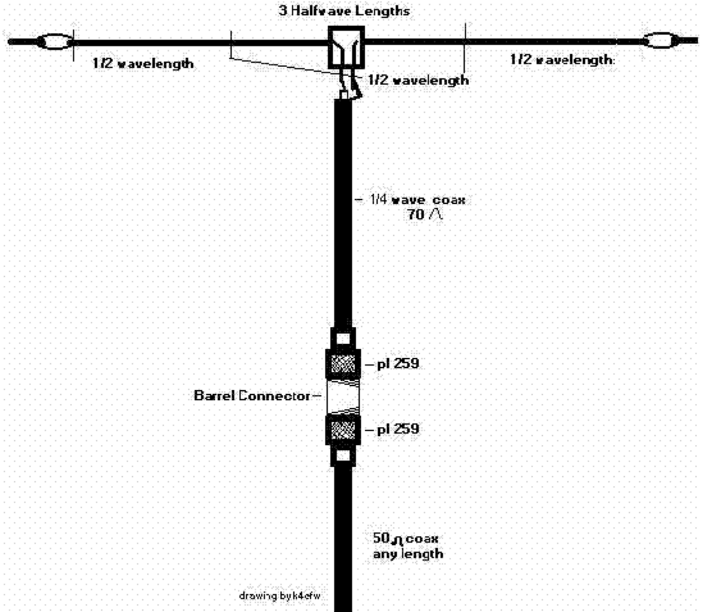
3 Halfwave Lengths
1/2 wavelength
1/2 wavelength
1/2 wavelength
1/4 wave coax 70 Ω
pl 259
Barrel Connector
pl 259
50 Ω coax any length
drawing by k4efw

Figure 10.7: 3 Half-wave Dipole

To use a 3 half-wave antenna on 15 meters, the 70-ohm matching section needs to be 7 feet 7 inches and the antenna needs to be 64 feet long for a good match. It will be just a little long on 40 meters. When using a 40-meter dipole with a 15-meter quarter-wave matching section, it will still have acceptable SWR on 40 meters.

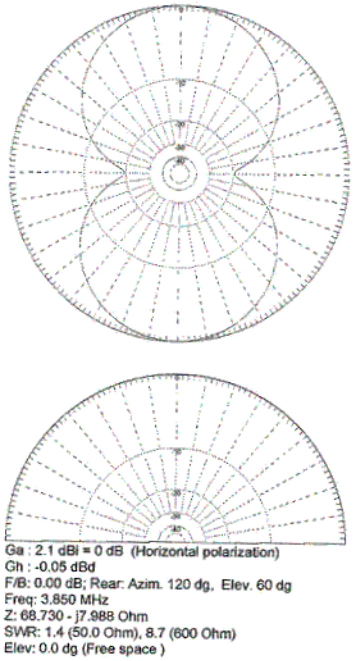

```
Ga : 2.1 dBi = 0 dB  (Horizontal polarization)
Gh : -0.05 dBd
F/B: 0.00 dB; Rear: Azim. 120 dg,  Elev. 60 dg
Freq: 3.850 MHz
Z: 68.730 - j7.988 Ohm
SWR: 1.4 (50.0 Ohm), 8.7 (600 Ohm)
Elev: 0.0 dg (Free space )
```

Figure 10.9: Radiation Pattern of a 15-Meter Three Half-Wave Dipole at 65 Feet

The pattern shows 6 lobes, 4 major lobes and 2 minor lobes. The vertical radiation pattern shows low angle radiation.

10.11 All Band Random Length Dipole

A random length of wire cut into two pieces can be used as a dipole, and it will radiate efficiently. It has to be at least a half-wave length on the lowest band you want to work. It looks the same as the all-band dipole and is the same, except it is not resonant on any band. The random length dipole is being described here to emphasize that the radiating part of an antenna doesnt have to be resonant. Because it will have a feed-point impedance that is unusual, it must be fed with ladder-line a tuner, and a balun. Since you are using a tuner, it can be used on multiple bands. If you make it very long, it can have gain over a dipole. For example, if it is four wavelengths long, it will have 3-dBd gain. As you move to higher bands, the electrical wavelength of the antenna increases, and each higher band will have more gain.

A half-wave antenna radiates perpendicularly to the plane of the wire. As you move to higher bands, this antenna begins to show some gain, and instead of two lobes of radiation, the two lobes split into four lobes and the pattern resembles a 4-leaf clover. As you make the antenna longer, the four lobes move nearer the to the ends, the gain increases, and there are minor lobes of radiation between the major lobes. These minor lobes make it possible to work in all directions. The longer the wire, the closer the antennas major lobes radiate bi-directionally toward its ends

10.12 All Band Center-Fed Random Length Dipole

The problem with using a random length of wire for this antenna is you may find that because of limitations of your tuner, you may not be able to tune a particular length of antenna on some bands. Certain lengths will tune all bands and one of those lengths is 135 feet. That particular length will be nearly resonant on all bands of 80-10 meters. Resonance only makes it easier to tune, but it has no effect on efficiency. A length of 260 feet will tune from 160-10 meters. Lengths of 260 and 135 feet have been used here successfully. Some hams use random lengths of wire without problems. Then some hams have had problems with other random lengths. The ones having the problems solved the tuner problems by changing the length of the dipole wire. If you plan to put up this antenna using a random length of wire, you will need to experiment with various lengths until you find a combination that works.

Tests were performed here using two towers of equal height and spaced 100 feet apart. On one tower, was an 80-meter inverted-V 120 feet long fed directly with coax, and running parallel to it on the other tower was a 135-foot long inverted-V fed with ladder-line and a tuner. At the resonant point of the coax-fed dipole and having tuned the ladder-line fed antenna, it was possible to switch antennas instantly and many hams were asked to look at their “S-meters” while the antennas were switched. All hams that participated in the test said the signals from both antennas were equal. The signals were measured on analog S-meters, not on segmented LCD meters found on most of todays transceivers.

10.13 A 2-Band Fan Dipole

A 2-band dipole can be constructed by connecting together the feed point two dipoles for even harmonically related bands. It is fed with 50-ohm coax with or without a balun. The best example of this is 80 and 40-meter dipoles connected together. Both dipoles are cut for half-wave resonance on each of the two bands. They are fed together and the ends of the wires are spread apart. If the ends are close together, there will be interaction between the dipoles. In such an antenna system, both dipoles must be carefully pruned for lowest SWR one band at a time. The lower band will be tuned first since the shorter dipole will not interact with the longer one. Each dipole has a low antenna resistance on the band for which it is resonant. RF energy follows the path of least resistance, and it automatically selects which dipole will receive power. The remaining antenna will have a high impedance. High impedance will block RF. Such an antenna will have a narrower bandwidth than a single band dipole, but close to the resonant frequency of each dipole, a tuner will not be needed. To connect many dipoles for multiple bands is possible, but it is not recommended because multiple wires are prone to interact and it will be impossible to achieve a low SWR on some bands. However, on the two band model, the 40-meter dipole will resonate close to 15 meters, the 80-meter dipole will resonate close to 10 meters, and working four bands with this set-up is possible. Some hams are using this antenna successfully with a tuner on all bands, although the signal on 20 meters suffers somewhat because of high SWR.

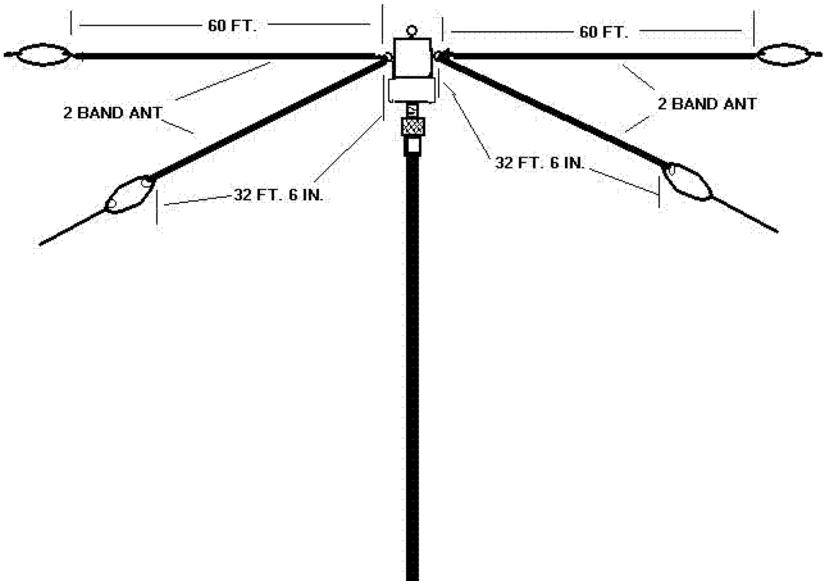

Figure 10.10: Two-Band Fan Dipole for 40 and 75 Meters

10.14 Trapped Dipole for 75 and 40 Meters

A trap is constructed from a capacitor and an inductor connected in parallel. It acts as an open switch on the frequency for which it is resonant. A trap is placed on each side of the dipole. For a 75 and 40 meter trapped dipole, the traps must be resonant on 40 meters, and each trap should be placed a quarter wave from the center insulator. The center section between the traps is electrically isolated from the ends of the dipole by the traps on 40 meters, and the center section of the antenna becomes a full-sized half wave resonant dipole for that band. This antenna is fed with 50-ohm coax and an optional balun. Wires connected to the outside of the traps are run to the end insulators and are tuned so the entire antenna resonates on 75 meters. The 75 and 40 meter trapped dipole will be shorter than a 75-meter dipole because the inductor in the 40-meter trap acts as a loading coil on 75 meters. In addition, the ends of the antenna can be tuned to operate on the 80-meter CW band instead of the 75-meter voice band. Several sets of traps can be inserted at the correct points in the dipole to make a multi-band dipole. Multi-band trapped dipoles are being sold, but in many cases they will require the use of a tuner. If a good match is found at a frequency on some bands, the bandwidth without a tuner will be very narrowThe antenna is only 108 feet long instead of 120 feet because of the loading effect of the traps on 75 meters. These dimensions are for antennas using the traps made by W2AU. If you use other brands of traps, the length of the end wires will have to be adjusted. What you do in that case is make the wire long, measure its resonant frequency on 75 meters, and prune the ends to resonance at your favorite frequency.

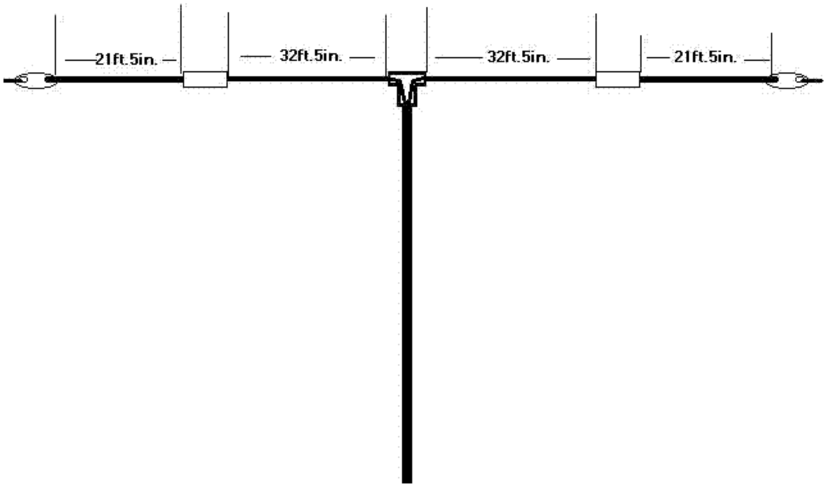

Figure 10.11: Trapped 75- and 40-Meter Dipole

10.15 The Extended Double Zepp Dipole

An extended double zepp is a long dipole with 3-dBd gain. It is the longest dipole antenna, which will radiate at right angles to the plane of the antenna. To find the overall length of an extended double zepp, divide 1197 by the frequency in MHz. Each leg of the antenna is 0.64 wavelength long and the total length is 1.28 wavelengths. An extended double zepp for 75-meters at 3.8 MHz is 315 feet. Not many amateurs have space for that antenna. The extended double zepp is mostly fed with ladder-line. Another method of matching an extended double zepp is to use tuned lengths of 450-ohm ladder-line as a series matching transformer connected between the 50-ohm coax and the dipole. The length of the matching section of 450-ohm ladder-line can be found by dividing 135 by the frequency in MHz.

10.16 The G5RV Dipole

An interesting antenna you can buy that will work somewhat on all high-frequency bands is the so-called G5RV antenna. It is named after the call letters of Louis Varney (SK) who designed it. It is a 102-foot long or three half-wavelength dipole antenna on 20 meters (14.150 MHz), and can be used with a tuner on other bands as well. In his original design, Varney calculated the length to be 102.57 feet, but chose to make it an even 102 feet since a tuner was going to be used with it anyway. It was originally fed through a 34-foot 500-ohm homebrew

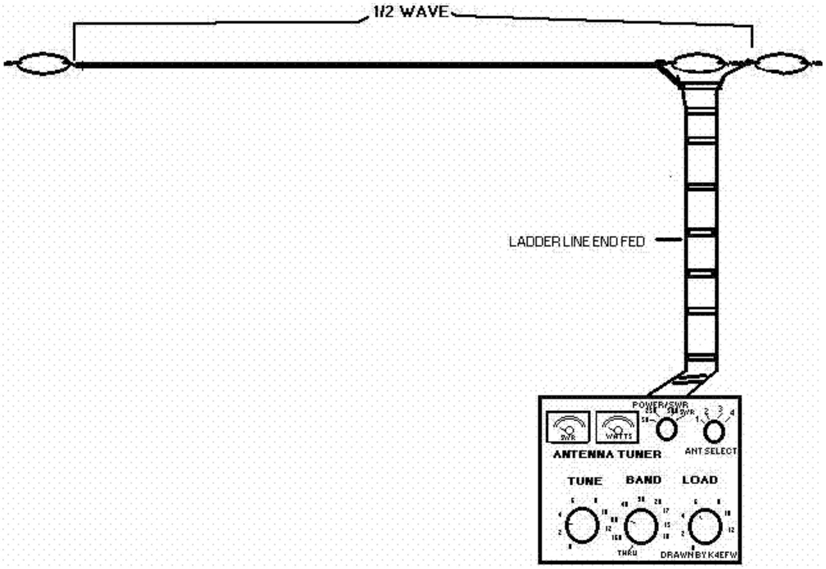

Figure 10.12: Extended Double Zepp Dipole

open wire matching section from a 70-ohm coax or parallel conductor feed-line. The 34-foot open wire line is a half wavelength on 20 meters and at the end of a half-wave feed-line, you will see the antennas impedance repeated regardless of the feed-line impedance. The ladder-line helps partly to match the antenna on the other bands.

The G5RV antenna is around 20 feet short of being a half-wave on 80 meters, and on bands on 20 meters and up, it has theoretical gain. We believe that gain is negated by losses in the coax of the feed system, except for 20 meters. At the frequency of the best match, commercially made models of the G5RV are said to have a 1.8:1 SWR on 80 meters. Where the coax joins the open wire, Varney recommended using a choke made of 8 to 10 turns of coax. He advised against using a balun, because, as he says SWR of 2:1 or higher may cause the balun to heat and possibly burn out. The SWR will be moderately high or high on bands other than 20 meters. Varney recommends using the lowest loss coax available and as short a run as practical because of feed-line losses caused by high SWR. This recommendation is very important today, as it was when Varney designed it. Some G5RV antennas put out decent signals and some others have relatively weak signals. Without further investigating, the only way to explain this is that some are using lossy coax and baluns while others are not, and the height above ground may play a part in how well it works.

The G5RV antennas being made today use small diameter 50-ohm coax, 450-ohm ladder-line, and a balun between the ladder-line and the coax, contrary to Varneys suggestions. There are several variations of the G5RV antenna being sold today because many believe they can improve the original design. If you use a G5RV antenna, a tuner will be required.

The G5RV shown in *Figure 10.13: G5RV Dipole* on page 39 is close to the original version of the antenna. This one pictured below is from an old article that K4EFW found somewhere. It is like the one he used. As you can see, it uses 300-ohm TV ribbon. The length of the parallel TV ribbon is 36 feet, but modern designs of this antenna use 34 feet of 450-ohm ladder-line. All these variations work equally well when they are used with a tuner. It is shown in the inverted-V configuration but it could be put up in the flattop configuration as is, with no modification.

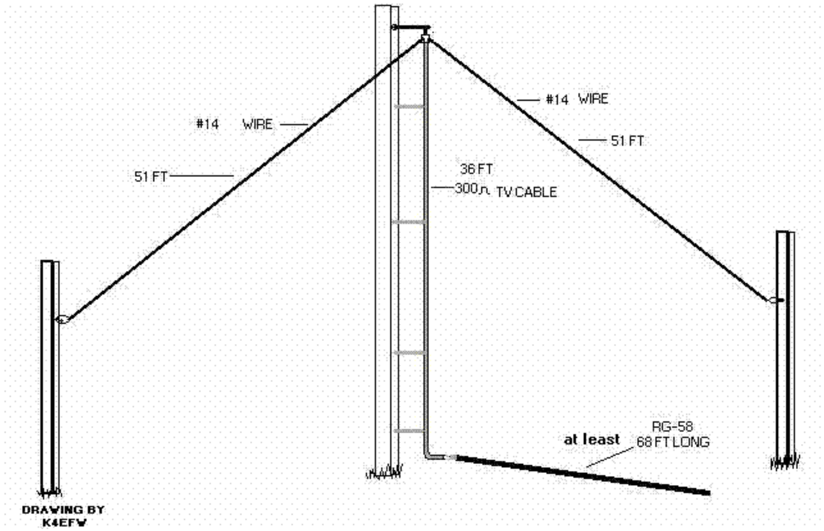
#14 WIRE
51 FT
#14 WIRE
51 FT
36 FT
300Ω TV CABLE
at least
RG-58
68 FT LONG
DRAWING BY
K4EFW

Figure 10.13: G5RV Dipole

Jeff, AI8H, in Oxford, Georgia, had a pair of G5RV dipoles oriented in different directions. Recently he put up a 75-meter half-wave inverted-V. Being able to switch antennas, he ran A-B tests on 3902 kHz and the inverted-V was 10 dB stronger than the first G5RV and 15 dB stronger than the other one. Now if we are saying the stronger signal is 40 dB over S-9 and the weaker signal is 25-30 dB over S-9, no one will notice the difference. Only under marginal band conditions will the difference be important. In addition, the G5RV antenna will work better on the other bands.

10.17 Off-Center Fed Dipoles

A long dipole consisting of multiples of equal half-wave segments is normally fed in the center using ladder-line. Dipoles do not necessarily have to be fed in the center. They can be fed in the center of any one of these half-wave segments, even fed off-center. A fair match will occur if coax is used.

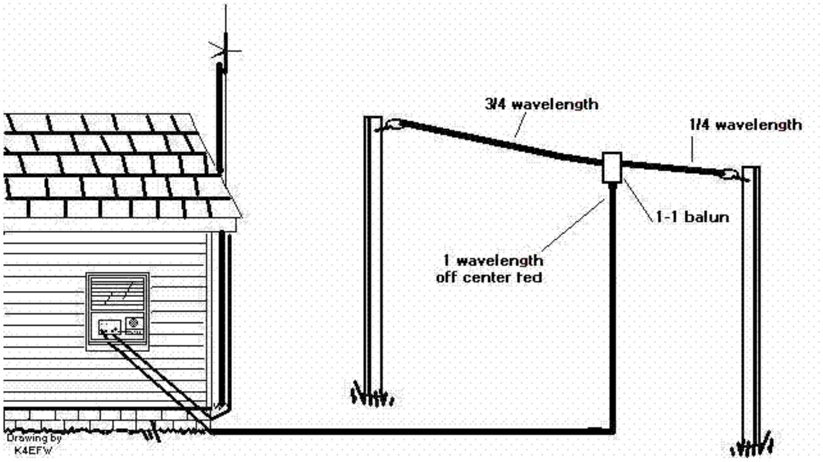
3/4 wavelength
1/4 wavelength
1-1 balun
1 wavelength
off center fed
Drawing by
K4EFW

Figure 10.14: One Wavelength Off-Center Fed Dipole

The dipole shown in *Figure 10.14: One Wavelength Off-Center Fed Dipole* on page 39 is a one-wavelength dipole. It is nothing but two half waves end to end. It is being fed in the center of one half-wave segment or a quarter wave from one end. It is possible to make it any number of half waves, and if it is fed a quarter wave from one end, it will have a fair match. The way it is shown above is an example of how to feed an antenna with even multiples of a half wave using coax. A 2:1 or 4:1 balun will improve the match on longer versions.

The windom antenna is another example of an off-center fed antenna. The original windom was fed off center with a single wire. The other side of the transmitter was connected to ground. The feed-point impedance at the transmitter was reported to be 500 ohms on all bands. The antenna was designed by William L. Everett and J.F Byrne at Ohio State University. W8GZ, whose last name was Windom, described the antenna in the September 1929 issue of *QST.*

A lot of research concerning the modern variations of the Windom antennas has been done, including the ones described by Fritzel, K4ABT, W4RNL, The Carolina Windom, and ON4BAA. The main differences in these variations are the slight differences in the position of the feed-point and the impedance of the baluns used for matching. The Windoms are sensitive to the height over ground, meaning the height above ground affects the SWR. The offset position of the feed-point will also determine the feed-point impedance. The one sold by K4ABT is a variation of the Fritzel antenna, and the one sold by Radio Works, The Carolina Windom, (shown in *Figure 10.15: Windom Dipole* on page 41) claims it has a vertical radiator.

There are two variations of Windoms, both claiming they have vertical radiators, The Carolina Windom and the one previously marketed by W4COX have two pieces of transmission line in series. The upper piece is connected to the dipole, and the lower piece is connected to the transmitter. The feed-point of the dipole is placed off center. In The Carolina Windom being marketed today, the upper transmission line is coax. The one made by W4COX had the upper piece made from ladder-line, but in either case, the principle is the same. The two pieces are connected together through a line isolator, a type of balun. The line isolator keeps the lower piece of transmission line from radiating. Because the antenna is fed off-center, the marketers of The Carolina Windom claims it causes an unbalance of current in the upper piece of transmission line. This is doubtful because there is a balun at the feed-point, which should prevent the feed-line attached there from radiating. The main difference between The Carolina Windom and the one sold by W4COX is that a 4:1 transformer is between the coax and the ladder-line, and a 1:1 line isolator is between the upper and lower coax cables. Both variations of this antenna show low SWR on several bands, but a tuner is used to match it.

Another unique variation of the Windom dipole is the Fritzel antenna, named after its inventor and manufacturer, Dr. Fritz Spillenger (SK), a German ham, call sign DJ2KY. Alpha Delta is now selling an almost exact duplicate of the original Fritzel antenna. Alpha Delta calls it an OCF antenna and it is made by Buckmaster Antennas. There are two models of the Alpha Delta antenna: one for low power and one for high power, the power rating of the balun being the limiting factor. The Fritzels short side is 0.18 wavelength long and its long side is 0.32 wavelength long. It is fed with coax and a 6:1 balun. Theoretically, the feed point impedance is 300 ohms, and the balun provides a 50 to 300 ohm impedance transformation. Modeling the antenna on its lowest resonant frequency at 35 feet, it shows about 120 ohms impedance. The original Fritzel antenna being used by K4LMS reportedly will work all bands with a tuner, but it will work 40, 20, 17, 12, and 10 meters without a tuner with an acceptable SWR. The

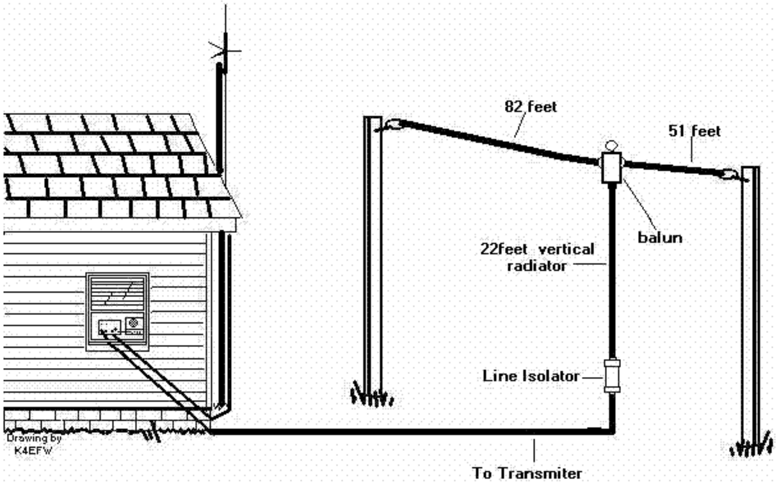
82 feet
51 feet
balun
22feet vertical radiator
Line Isolator
To Transmiter
Drawing by K4EFW

Figure 10.15: Windom Dipole

Windom being sold by K4ABT uses a 4:1 balun and the feed-point is at a slightly different location. That one is shown in *Figure 10.16: Windom Dipole (Fritzel Type)*.

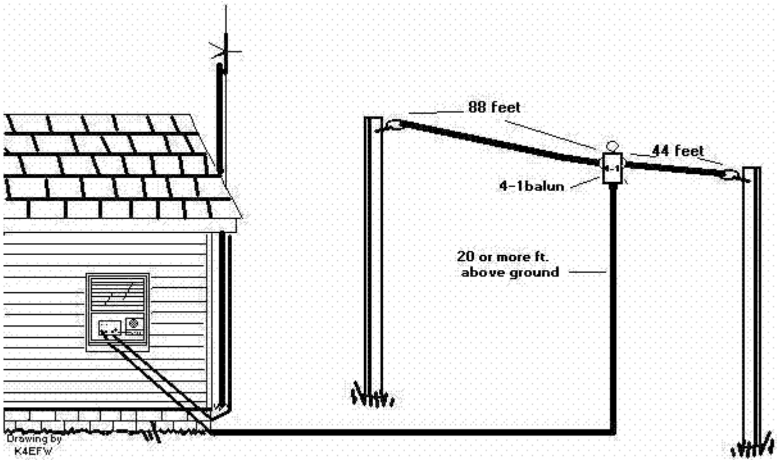
88 feet
44 feet
4-1
4-1balun
20 or more ft. above ground
Drawing by K4EFW

Figure 10.16: Windom Dipole (Fritzel Type)

The difference between the Windom antenna sold by K4ABT and the original Fritzel is the difference in the offset of the feed-point. Since the K4ABTversion uses a 4:1 balun, it appears his is fed at the 200-ohm point, and the original Fritzel is fed at the 300-ohm point. On any resonant dipole, the lowest feed-point impedance is found at the center. As you place the feed-point offset toward either end, the impedance gets higher. The highest feed-point impedance occurs at the end of the dipole.

11.

End-Fed Antennas

11.1 End-Fed Zepp

A half-wave resonant antenna can be fed from its end. When fed this way, it is also known as an end-fed zepp. An end-fed zepp will work on its fundamental frequency and on odd and even harmonic frequencies. The name “Zepp” goes back to the days of dirigibles or Zeppelins, which used trailing wire antennas that had to be fed at one end. The end of a half-wave antenna has very high impedance, and an antenna fed this way is said to be voltage fed. Feeding a half-wave resonant dipole in the center means it is current fed. The normal way of feeding the end-fed antenna is with ladder-line. One side of the ladder-line is connected to one end of the antenna and the other side of the ladder-line is connected to nothing. To secure the unconnected side of the ladder-line, it is connected to a short wire running between two insulators. Since the antenna is connected at its high impedance point, no current flows into an antenna, but there will be a large current in the center of this antenna. No current flows from the open side of the feed-line because it is at a zero current point. (See *Figure 11.1: End-Fed Zepp* on page 43.)

The end-fed zepp can be matched by cutting the ladder-line to a quarter wavelength with the bottom end of the ladder-line shorted. A certain distance above the short is a 50-ohm feet-point and it can be fed directly with coax. MFJ is marketing antennas of this type made for single bands, and they are selling the parts separately so you can build your own. You will have to find the 50-ohm point by trial and error. This method of feed makes it a single band antenna.

11.2 End-Fed Random Length Antenna

Figure 11.2: End-Fed Random length or Long Wire Antenna on page 43 shows another end-fed antenna made from a random length of wire connected to the back of the tuner. The wire then exits the shack and goes to a high support where it then runs horizontally to another high support. The tuners groundside must be connected to a good RF ground, since a poor ground causes high losses. This antenna is commonly called a “long wire.” Since the end of the antenna comes in the shack, you will be exposed to high levels of RF. In addition, this type of installation may cause RF to be picked up in the microphone, noted by distortion. The

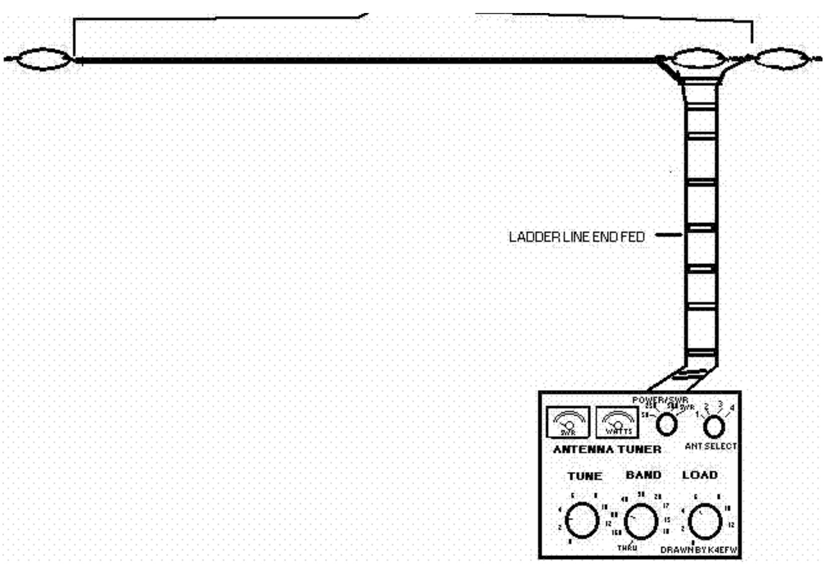
LADDER LINE END FED
POWER/SWR
SWR
WATTS
ANTENNA TUNER
ANT SELECT
TUNE
BAND
LOAD
THRU
DRAWN BY K4EFW

Figure 11.1: End-Fed Zepp

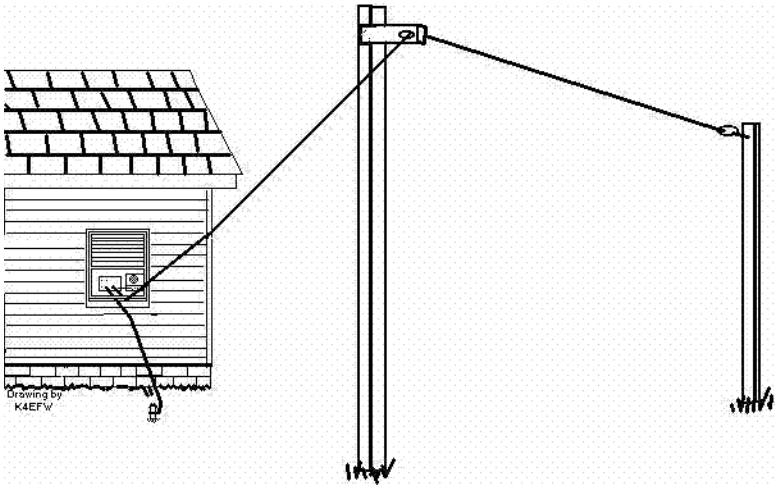
Drawing by
K4EFW

Figure 11.2: End-Fed Random length or Long Wire Antenna

feed-point of the long wire being connected directly at the output of the tuner can have an impedance of a few ohms to a thousand ohms depending on the antennas length. If the wire is cut to a multiple of a half wave at the lowest frequency, the system will be efficient since it is fed at a voltage point and very little current flows into the ground. This antenna is really a variation of an inverted-L fed directly without a feed-line from the tuner.

12.

The Half Sloper

The half sloper shown in *Figure 12.1: Half-Sloper Antenna* on page 45 is an antenna that is hard to categorize, since it is not a sloping dipole and it is not a vertical. The half sloper is half of a sloping dipole. To make one of these antennas, cut a quarter-wave radiator by dividing 234 by the frequency in MHz and tie an insulator to both ends. One insulator is tied near the top of a tower and the radiator wire is run down toward the ground. Coax is split into its center conductor and shield, and it connects across the insulator at the tower. The center conductor of the coax is tied to the quarter wave radiator and the shield is grounded to the tower. This means the tower is acting as the missing half of the dipole. It is a difficult antenna to get a good match because the height above ground of the feed-point and the angle of the slope affect the impedance. Some users of this antenna say to mount the feed-point at 45 feet up on the tower and have a beam antenna on the tower above the feed-point to use as a counterpoise. Other users say you must find the 50-ohm point on the tower, which is a tedious task. It has also been said, “Some installations work super, while others do not work well at all.” The half-sloper is used almost exclusively on 80 and 160 meters. The Alpha-Delta half sloper was tried here and its performance was disappointing. The signal from it was down a least 10 dB below a dipole and the SWR wasnt low enough. The half sloper is mostly vertically polarized and it is directional toward the slope.

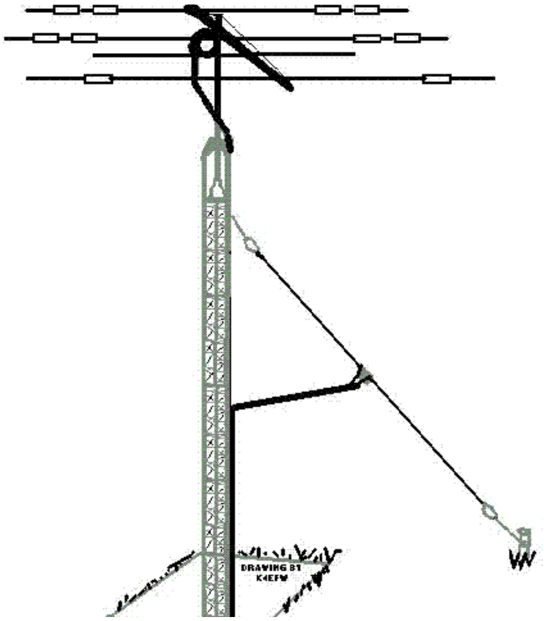
DRAWING BY
K4EFW

Figure 12.1: Half-Sloper Antenna

13.

Vertical Antennas

13.1 Why Verticals Are Used

Vertical antennas have the radiator mounted at right angles to the earth. The vertical is used whenever you desire to radiate your signals in all directions at a low angle. Low angle radiation is needed to work DX effectively. Radio waves traveling to the ionosphere where they are reflected need to hit the ionosphere at a point near the horizon in order to reflect farther around the curvature of the earth. In order to get a dipole to radiate a strong signal at low angles, it has to be more than a wavelength above ground. A low dipole is not particularly a good DX antenna for 80 and 160 meters. However, the average dipole at modest heights will outperform any ground-mounted vertical having a poor ground system. Vertical antennas work very well at low frequencies such as the broadcast band, but the ground losses increase as we move higher and higher in frequency (Refer to *Section 5.2 Ground-Wave Propagation* concerning ground-wave propagation). It is very difficult to get a good ground for a ground-mounted vertical unless you live next to salt water. Vertical antennas, because they are unbalanced antennas, do not need baluns. They are normally fed with coax.

If a ground mounted, quarter-wave vertical is all you can put up at your location (QTH), then use it. However, it will be a mistake to put up this antenna if you are not be able to have a ground radial system and are able to put up a dipole. Most ground mounted quarter-wave verticals manufactured today are trapped in order to work multiple bands.

The ground-mounted vertical also needs to be put out in the clear away from RF absorbing objects. These facts do not apply to half-wave verticals, which are in themselves different animals, nor do they apply to high quarter-wave verticals using elevated radials.

The approximate length of a full-sized resonant quarter wave vertical can be found by dividing 234 by the frequency in MHz. Note: 234 is half of 468, the number we used to calculate the length of a half-wave antenna. The actual length for resonance may be a little different from what you calculate, because of the diameter of the vertical element. Trapped verticals are physically short of a quarter wave in length because the traps load them. The vertical is fed at one end at the bottom where it is insulated from the ground. The center conductor of the coax connects to the vertical element and the shield is connected to the ground system.

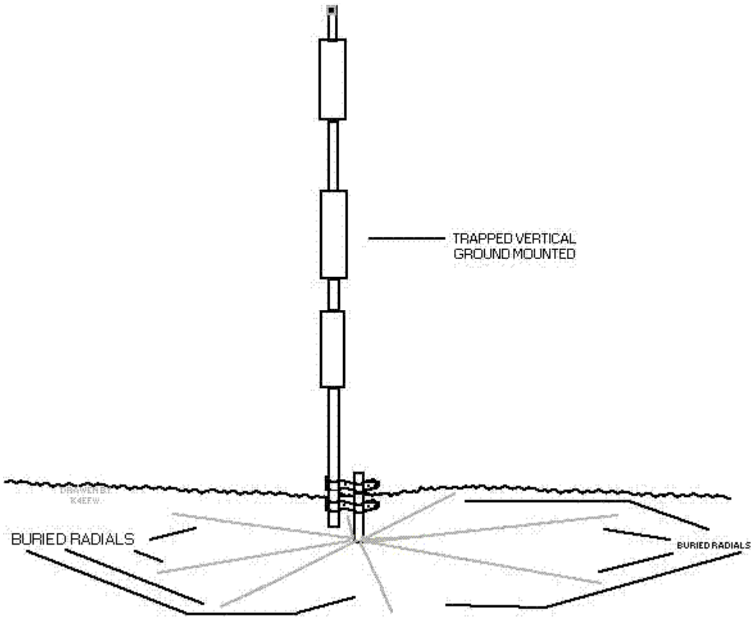

Figure 13.1: Ground Mounted Trapped Vertical

13.2 Disadvantages of Using Quarter-Wave Verticals

The most obvious disadvantage of using any vertical antenna is on 80 meters it has less than optimum high-angle radiation needed to work stations within a few hundred miles. Ground-mounted quarter-wave verticals use a ground system for the other half of the antenna and the ground system losses can be very high. The ground wave signal should eventually radiate in space at angles at the horizon, but since there are very high losses in the ground wave at amateur frequencies, a ground-mounted vertical has almost no signal down near the horizon. At angles below 10 degrees, the signal will be greatly attenuated.

A ground-mounted quarter-wave vertical with an ideal ground should have an impedance of 35 ohms resistive. If you were to measure its impedance, and it measures 60 ohms resistive on an antenna analyzer, it means it has a loss resistance of 25 ohms. Moreover, that loss resistance is mostly in the ground system. Under these conditions, only 58 percent of the power will radiate as RF, although you will have a 1.3:1 SWR. 42 percent of the power will be turned into heat by the loss resistance. With the feed-point being at ground level, some more loss comes from the radiated wave being absorbed by power lines, trees, and buildings with its associated wiring. That loss does not show up in antenna analyzer measurements.

The best ground system for a ground-mounted vertical is 120 wires, called “radials,“ radiating from the feed-point like the spokes of a wheel. These radials need to be a quarter wave long. At the feed-point, the radials are bonded together and are fed from the shield side of a coax cable. Not many amateurs have the resources to build such a ground system. Many short radials will be more effective than a few long ones. When using a ground mounted vertical, many hams drive an 8-ft. ground rod into the earth for their ground system. The ground losses are very high in that case. Using a ground rod for the ground system of a vertical antenna confirms the old adage: “Verticals radiate poorly in all directions.”

To eliminate ground losses, you can use an elevated quarter-wave vertical with an elevated ground system called a “ground plane.” The ground plane vertical, as it is called, needs to be mounted high enough to prevent the return path from coming back through the earth ground. Ground plane verticals need to be mounted above nearby objects that absorb RF. They will be nearly 100% efficient if they are high enough. The ground plane consists of two or more radials, but most ground planes have three or four. The ground plane radials do not have to be resonant, but should be at least a quarter wave long. An elevated ground-plane vertical will be more effective for working DX than a dipole.

13.3 Long and Short Verticals

Verticals can be less than a quarter wave in length. They can be loaded by coils or linear loading sections or a short vertical can be fed directly with a tuning unit at the feed-point. The loss resistance in a short vertical may be appreciable. Since the radiation resistance is very low at the feed-point of a short vertical, the current at the feed-point will be very high. The more current that flows into loss resistance, the higher the loss will be. Any coils used in the tuning unit and for loading should be made of as heavy a conductor as possible, since these can cause appreciable loss when the current is high. This is also true for the ground system. The loss described here is called “I squared R loss”, which means the loss in watts is found by multiplying the current times itself and then multiplying that answer by the loss resistance. That means if the current into a lossy antenna system is doubled, the power lost in watts is increased four times. Making a vertical very short and tuning it to resonance with an inductor will also result in an antenna with a very narrow bandwidth.

A more subtle loss of energy in very short vertical antennas is coronal discharge from the tip end of the vertical. Corona occurs when the voltage is very high at the end and electrons flow out into the air. This can be visible at night if the transmitter power is high and you are at a high altitude. Power is lost from the antenna when corona is produced because corona is a form of light and light is another form of energy.

In 1973 while we were working for radio station WWNC in Asheville, North Carolina, a trapped vertical for 10 through 80 meters was erected. The length of the antenna was only about 25 feet. A loading coil near the top made it resonant on 75 meters. The ground system was the metal body of a 75-foot long mobile home. Fair reports were received from this set-up. The reports were not bad because of our having a good ground. One night, while working 75 meter SSB, one of the neighbors came over and said, “You’re tearing up my TV.” Checking all of the inside connections proved they were tight. Our wife keyed up the transmitter while we made a trip to the antenna to check the connection there. Before arriving there, looking up, we saw blue fire coming off the end of the vertical. The corona was responsible for the television interference (TVI). It was visible because Asheville is at a relatively high altitude and the transmitter was running 700 watts. An inverted-V was put up, the TVI disappeared, and better signal reports were received.

You can realize up to a 1.5 dBd gain from a vertical antenna by making it longer than a quarter wave, but there is a limit to how long to make it and still get low angle radiation. That limit is 5/8th-wave. To find the length of a 5/8th-wave vertical, divide 585 by the frequency in MHz. For example, to calculate the length of a 5/8th-wave vertical for 20 meters (14.000 MHz) divide 585 by 14.0. It equals 41.786 feet or approximately 41 feet 9.5 inches. A tuning unit will be needed at the feed-point of this antenna, as the impedance of a 5/8th-wave antenna is low and high current will flow into it. A tuning unit will usually have enough bandwidth to cover the entire band on each band of 20 meters and higher. A tuning unit is also called a matching network. It is similar to an antenna tuner, but has fixed inductors and capacitors. Tuning units for 80-and 160-meter verticals will cover only a portion of the bands. Outside the bandwidth limits of the tuning unit, you can use the tuner at the transmitter end. Radials or ground planes are needed for a 5/8th-wave vertical and they need to be a quarter wave long.

The impedance of a half-wave antenna is high if fed at its end. An end-fed half-wave vertical will have a small amount of gain over a quarter-wave vertical. This antenna does not have the ground losses a quarter-wave vertical has because it is fed at a high impedance point and the current flowing into the ground is negligible. Commercially made resonant half-wave trapped verticals now on the market are end fed at the bottom. A built-in matching network is found at the base, and several very short radials are mounted below the feed-point to de-couple RF from the feed-line. These antennas should be mounted as high as possible away from RF absorbing objects. Because the ground losses are lower, the half-wave vertical will outperform a quarter-wave vertical by several dB and in many cases many dB.

13.4 Unscientific Observations of Verticals

At our home, an old Hy-Gain trapped quarter-wave vertical for 40-10 meters was erected in 1961. It was mounted on the roof and had two quarter-wave radials for each band. It worked, but it was never compared to another antenna. It gave the impression it was a mediocre antenna. Other antennas replaced it.

One time in 1964, a grounded 60-ft tower was shunt fed as a vertical on 75 meters. Without having any radials, the transmitted signal was 10 dB weaker on this vertical 650 miles away in New York than on the inverted-V.

In 1969, a 4-band trapped vertical was put up on the top of a 60-ft tower. A 15-meter 4-element yagi under it was used for the ground plane. It was probably the best vertical installation we ever tried. It was good because it was high and in the clear and the 15-meter yagi made a good ground plane.

While we are on 80 meters, a ham 200 miles away frequently joins in the roundtable. He uses a trapped quarter wave vertical with a chain link fence as the ground. Several of the others are also 200 miles away run the same power. His signal is 10 to 20 dB below everyone elses on the frequency. It is good there are no interfering signals or noise or he will not be copied.

Another ham uses a Hy-Gain Hy Tower vertical with 3 ground rods as the ground system. According to our S-meter, his signal is 40 dB down below those of the other guys.

Charlie, AD5TH, works 40 meters using a Hustler 5-BTV vertical ground mounted with 72 quarter-wave radials. He has an outstanding signal for a ground-mounted vertical. His installation is out in the clear away from RF absorbing objects. He says, because of antenna restrictions at his location, it is the only antenna he can put up.

Another ham friend, N2HGL, has both a dipole and a half-wave trapped vertical on 40 meters. At a location 160 miles away, he is 10 dB stronger on the dipole, but he is equal in strength on both antennas in Indiana 600 miles away. This comparison shows the superiority of the half wave vertical over the quarter-wave one because his signal with the half-wave vertical was equal to his signal from the dipole. If he were using a quarter wave vertical, we would expect his signal would be better on the dipole in Indiana. It also demonstrates the superiority of a dipole over a vertical for working short distances.

Bill, W4ZQL, runs a ground-mounted SteppIR vertical. He lives beside a salt-water river in Florida that he uses for a ground. He puts out a very good signal on 40 meters. No ground losses!

13.5 The Inverted-L Vertical

The inverted-L antenna is a wire vertical antenna with part of the top end bent horizontally. It resembles an “L” turned upside down. The inverted-L is used to reduce the height required by a vertical and still keep the antenna resonant and full sized. It is fed at the end at ground level the same way a ground mounted vertical is fed, and all the losses we described for a ground-mounted vertical apply here. Some current flows in the horizontal part of the inverted-L and for that reason, it has both strong vertical and weaker horizontal polarization. If you make it a half-wave antenna, you wont need a good ground because negligible current flows into the ground. A half-wave inverted-L antenna needs to be fed with 50-ohm coax and a tuning unit.

An inverted L for 160 meters is usually made of wire one-quarter wavelength long or about 127 feet. It runs vertically from near ground level to the top of a support, perhaps 60 or 70 feet. Then the end runs horizontally and is tied to a nearby support. The antenna is coax fed at ground level between the vertical section and ground system across some type of insulator. A matching network at the feed-poimt will be required to match it if the impedance is not equal to 50 ohms.

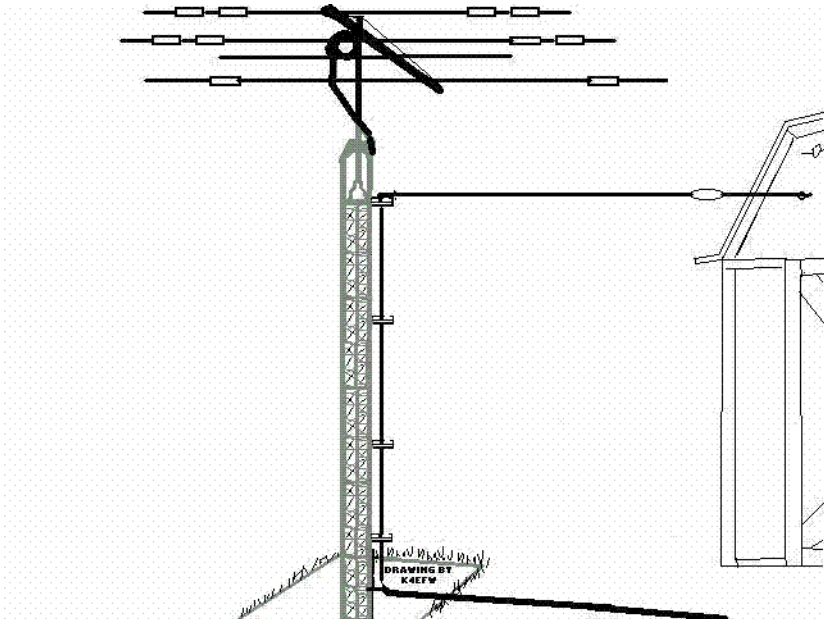

Figure 13.2: Inverted-L

The picture shows an inverted-L running up the side of a tower. The feed-point is at ground level with the center conductor of the coax attached to the bottom end of the wire. The coax shield connects to a ground system of radials. The total length of wire used in this antenna is half of what is needed for a dipole since the other half of the antenna is the radial ground system. The inverted-L is used mostly on 160 meters, but some have built them for 80 meters. The inverted-L antenna can also be cut for a half-wave to reduce ground losses.

13.6 Vertical Mobile Antennas

We have heard many good signals from mobiles, many being stronger than those from hams using ground mounted quarter-wave verticals. The mobile antenna, being so short, has a large capacitive reactance. A coil is inserted in the antenna to provide an equal amount of inductive reactance to make it resonant. As we said in the paragraph on short verticals, a coil of this type, carrying a large amount of antenna current, causes some loss resistance in the system. To reduce losses in the coil, wind it with a conductor as large as practical. (Thats exactly what some mobile antenna manufacturers have done.) The sources of loss in mobile antennas are in the coil losses, losses in the conductors making up the radiating part of the antenna, corona discharge, and the ground loss from the vehicle on which its mounted. However, because of the large amount of metal in the body of the vehicle, the ground losses are not as high as the losses from ordinary ground mounted verticals. Matching transformers are now available that step down the impedance of 50-ohm coax to the very low impedance of the loaded vertical. Good advice is to use the transformers rather than to rely on the internal tuner of the transceiver.

Some low priced single-band mobile antennas are constructed by using a polymer shaft and a small gauge wire encapsulated in polymer material running beside the shaft. The loading coil made of the same wire is also encapsulated in the polymer. The small wire, because of its size and because it carries a large RF current, will lose a lot of power by becoming hot. This type of mobile antenna is rated for 200 watts. If the wire didnt get hot, there would be no power limit.

All mobile antennas have corona loss and for such, there is no remedy. Most amateurs, because they cant see it, dont believe its there. Corona will not be visible unless you run high power and it is dark.

Ground losses from the vehicles body diminish with increasing vehicle size. This is why 18-wheeler hams have such big mobile signals. To diminish the ground losses on any mobile installation, you should use as large as a conductor as possible to bond the coax shield to the vehicle body. All metal parts of the vehicles body, fame, and drive system need to be bonded together with heavy ground straps. To make the mobile antenna system more efficient, use an antenna with an adjustable inductor and use as long a “stinger” as practical above the coil. You will increase the radiation resistance by using a longer stinger, and then the loss will be less because you will require less coil inductance. The ratio of radiation resistance to loss resistance becomes larger by raising the radiation resistance and reducing the loss resistance. As we said earlier, the efficiency of any antenna system is found from the ratio of radiation resistance to total resistance, or radiation resistance divided by total resistance times 100%. The total resistance is equal to all the loss resistances plus the radiation resistance.

The latest development in HF mobile antennas is motor driven variable inductors. These antennas are known as “screwdriver antennas.” The name refers to the electric screwdriver motors used to vary the inductance. A control cable is run from the motor to a switch at the operators position so it can be tuned from the operators seat in the front of the vehicle. Because a mobile antenna has a very narrow bandwidth, you will have to tune it often as you move frequency (QSY). It hasnt been many years since we had to get out of the vehicle to make inductor changes or make changes in the length of the stinger when the frequency was changed.

Mobile antennas for 20 through 10 meters do not require the care in installation that is needed for 160, 80, and 40 meters because the length of a mobile antenna becomes closer to a quarter wave as you move to higher bands. The radiation resistance increases on each higher band. While moving to higher bands, less inductance is needed to tune the antenna, and that lowers the loss resistance. A 96-inch mobile whip is just a couple of inches short of being a quarter wavelength on 10 meters and a loading coil is not needed there. The band that has the least mobile antenna efficiency is 160 meters. If you reach a radiation efficiency of 2% on 160 meters on your mobile installation, you will be doing well.

Below is some information concerning mobile antennas, which was received in an email. There was a 75-meter mobile “shoot-out” in California. (A shoot-out is an event where a group of hams gets together and compares signals radiated from various antennas.) Supposedly, equal power was applied to each antenna under test. Apparently, some type of field strength meter was used. A screwdriver antenna and a bug catcher, both with top hats, were used as the standard by which other antennas were compared because they put out equal signals. The other antennas are measured in how many dB they were below the standard. Here are the results of that test, and because it is hearsay, the accuracy of these figures is not guaranteed, but they do compare to what we have observed.

Table 13.1: Antenna Shoot Out Results

Screwdriver/bug catcher with top hats	0 dB reference
Screwdriver/bugcatcher without top hats	-3dB -50%
Hustler	-7 dB -80%
Outbacker	-9 dB -88%
Hamstick	-12 dB -94%
Whip with autotuner	-14 dB -96%.

The efficiency of the best 75-meter mobile antenna is from 5% to 10%. In using the best mobile antenna on 75 meters, a 100-Watt mobile rig will radiate 10 Watts at most. This means that a Hamstick being fed with 100 Watts will radiate only 0.6 watts, which is 6% of 10 Watts. Ninety-nine and four tenths Watts will be converted to heat. The person sending this information said it was published on the Internet in some news group. Again with good band conditions, it is amazing how little signal can be used to communicate.

The things that increase the efficiency of mobile antennas are:

- Place the loading coil about half way from the feed-point to the antenna tip. Efficiency decreases if you put the coil above or below this point

- Mount the antenna as high up on the vehicle as possible. This reduces the ground losses because it reduces the capacitance of the antenna to ground

- Use a loading coil with a Q as high as possible. See the ARRL Handbook for a discussion of coil Q

- Make the antenna as long as possible. Note: long antennas are prone to strike tree limbs and bridge overpasses

- Increase the size of the mast between the loading coil and feed-point

- Put a capacity hat above the loading coil. The capacity hat reduces the number of coil turns needed to resonate the antenna

- Make the coil with as large a diameter wire possible. This decreases the coil loss, which is a large part of the total loss of a mobile antenna.

Any changes made in the antenna system that raises the radiation resistance will increase the efficiency

14.

One-Wavelength Single Loop Antennas

14.1 The Horizontally Oriented Loop

To calculate the length in feet of any one-wavelength loop, divide 1005 by the frequency in MHz. Horizontally oriented one-wavelength loop antennas have become very popular on 160, 80, and 40 meters and it is one type of NVIS antenna. (NVIS stands for “near vertical incidence skywave” because of its high angle radiation pattern.) It is claimed by its users that the loop antenna is quieter than other antennas. This is because it doesnt pick up the noise from power lines, thunderstorms, etc., coming in at low angles. These antennas radiate on their fundamental frequencies with a broad pattern straight up to put a strong signal for nearby contacts. Recently published articles on this type of antenna have called them “cloud warmers.” There are other types of antennas called NVIS antennas other than loops. They are dipoles at low heights or dipoles with parasitic reflectors placed under them to cause the signal to radiate mostly straight up. The NVIS antennas have an advantage in working nearby stations because you dont get the static noise and interference from far distances. They are definitely not DX antennas. An article on NVIS antennas appears in the December 2005 QST.

On their fundamental frequencies, horizontally oriented loops take up half the horizontal distance as a half wave antenna for that band. Loops are two-dimensional antennas having depth as well as breadth. There are two loop configurations: The square loop and the triangle loop. Some hams have pulled the loops out in irregular shapes to fit where the supports are located. The only advantage in using a rectangular loop instead of a square loop is to take up less horizontal space. This is true because the gain of a rectangular loop is diminished below a square loop. The area enclosed by the perimeter of the loop determines the gain of a loop. A circular loop has the most enclosed area, but it requires an infinite number of supports. The gain of a loop comes from the loop having two maximum current points separated by a distance of one-quarter wavelength. From here on we will call a horizontally oriented loop a horizontal loop.

We also modeled the gain of the horizontal loop for the 80-meter band over real ground. The maximum gain occurs with the loop at 7 meters or about 25 feet above ground. Mind you, this gain is straight up from the loop. At that height, its gain is about 9.25 dBi and that equates to about 7 dBd in free space. The gain of the loop diminishes slightly as the antenna is raised.

The feed-point radiation resistance at 7 meters height is 35 ohms resistive and 0.0 ohms reactance and you do not need a matching section of 70-ohm coax. At a height of 10 meters or about 33 feet, the radiation resistance rises to 63.5 ohms. There the SWR will be 1.27:1, if it is fed directly with 50-ohm coax. At 15 meters or about 50 feet, the radiation resistance rises to 118 ohms and a 70-ohm matching section will be in order. The gain drops to a little less than 7 dBi at that height. These figures may or may not be applicable to your QTH, because your soil conductivity may be different from the soil we used to model it. As you can see from the above numbers, the feed-point resistance rises as the loop is raised.

The horizontal loops also are used on their harmonic frequencies. The loop with more gain and a superior pattern is a two-wavelength loop. An 80-meter loop is a two-wavelength loop on 40 meters. The two-wavelength loop has a lower angle of radiation, but is a very large antenna for 80 meters. At 3800 kHz it has a perimeter of about 530 feet. A two-wavelength loop is not an NVIS antenna. Using coax with a tuner is not an ideal way for working a loop on its harmonic frequencies. This is because of the high SWR in the coax on some bands will cause high loss. For example, an 80-meter loop fed on 40 meters will have an SWR of 8:1 and the SWR on 20 meters will be 49.5:1. There will be some hams who will say they get satisfactory results this way, however theory suggests they will have a stronger signal if they use a ladder-line because ladder-line has less loss. Feeding a loop antenna with ladder-line makes more sense when working a loop on harmonic frequencies.

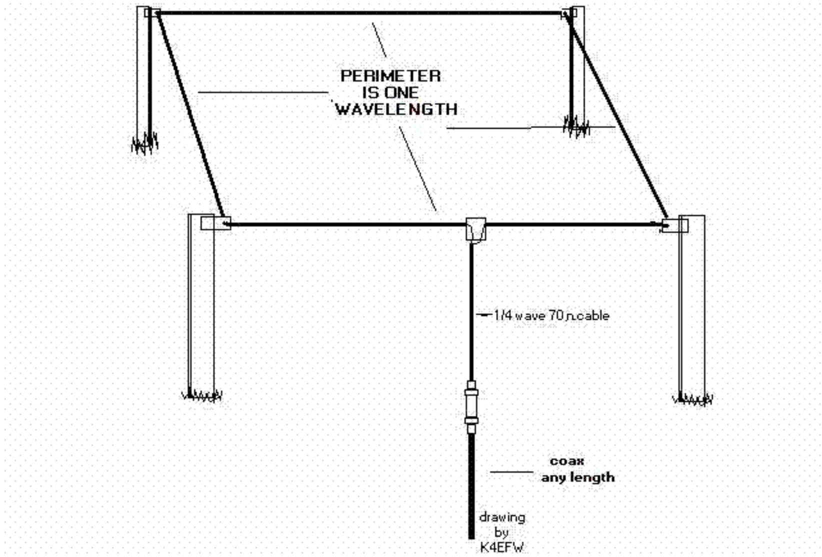

Figure 14.1: One Wavelength Horizontal Loop

To realize maximum gain, make the square and triangle have equal sides. When the sides are equal, the loop has maximum enclosed area for whatever configuration you use. Other shapes will work, but the gain will suffer.

To support a square loop, you will need four supports, one for each corner. We hope you will have trees or masts in the right places. A triangular loop will need three supports. Once you have cut the single piece of wire to the right length, run the wire through as many insulators as you have corners. At each corner of the loop, put an insulator and tie the corner to a support with a rope from the insulator. To make the feed-point, connect both ends of the loop to an insulator. Strip the insulation from the outer part of the coax. Separate the shield from the center conductor. The ends of the coax are connected to the ends of the loop across the insulator. Most hams do not feed loops with a balun at the feed-point.

14.2 The Vertically Oriented Single Loop for 40 and 80 Meters

Vertically oriented loops radiate broadside to the plane of the loop. A horizontally polarized vertically oriented loop has both vertical and horizontal wires. From here on out, we will refer to a vertically oriented loop as just a vertical loop. When using this term, we are not referring to its polarization. If the feed-point is on one of the horizontal wires, the loop radiates horizontally polarized waves. The vertical wires radiate weaker vertically polarized waves. If the feed-point is on one of the vertical wires, vertically polarized waves will be radiated. The radiation from a one-wavelength vertical loop has both high-angle and low-angle radiation. It is a good antenna for both nearby stations and for DX contacts. It is better than a dipole for DX because the vertical loop puts out a stronger low angle signal than a dipole does.

The gain of a vertical delta loop is 4.55 dBi or about 2.4 dBd. Its feed-point impedance is about 120.5 ohms. The square vertical loop has 5-dBi gain and about 2.85 dBd and the feed-point resistance is 143 ohms. They both need to be fed with a series quarter-wave matching section of 70-ohm coax.

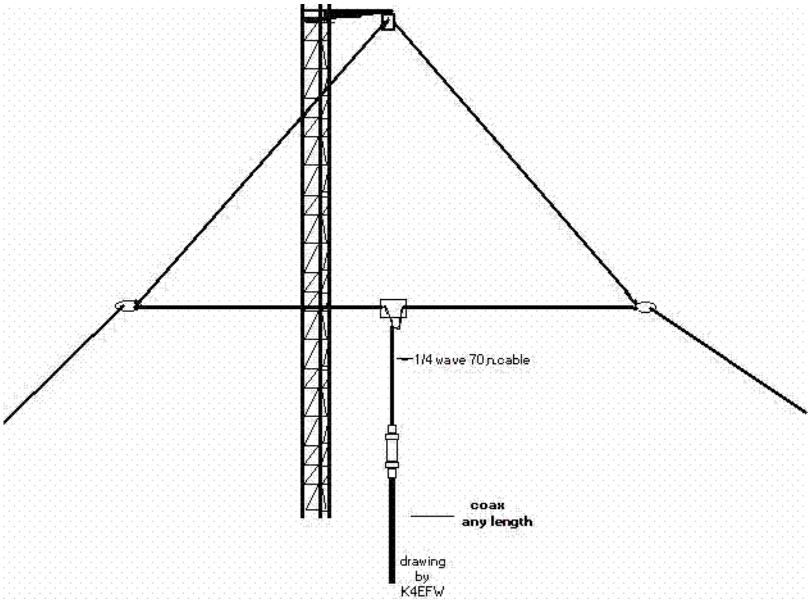

Figure 14.2: One Wavelength Horizontal LoopSingle-Element Vertical Delta Loop

Square vertical loops need two supports. The square vertical loop needs less vertical space than the delta loop. The vertical space needed for a square vertical loop for 80 meters is 92 feet. For 40 meters the vertical space is half that. It is rare to find someone using the square vertical loop these days. The vertical delta loop is more common because it needs only one high support. The apex of a delta loop for 3500 kHz needs to be 102 feet high and on 40 meters, it needs to be 62 feet. This assumes the bottom horizontal wire will be 20 feet off the ground. In order to make a vertical loop fit on a shorter support, the sides of the loop can be reduced in length while making the horizontal wires longer. This will put the two maximum current points closer together, which has the effect of reducing the gain.

Like the horizontal loop, the formulas for finding the length in feet of these loops are the same: 1005 divided by frequency in MHz. In addition, because the feed-point resistance is nearly the same as horizontal loops, quarter-wave matching sections and other methods can be used to feed the vertical loops. The vertical loop is not as sensitive to height as the horizontal loop. Both vertical square loops and vertical delta loops can be operated on harmonically related bands. (See *Figure 14.4: Radiation Pattern of a 30-meter Delta Loop on 15 Meters* on page 58).

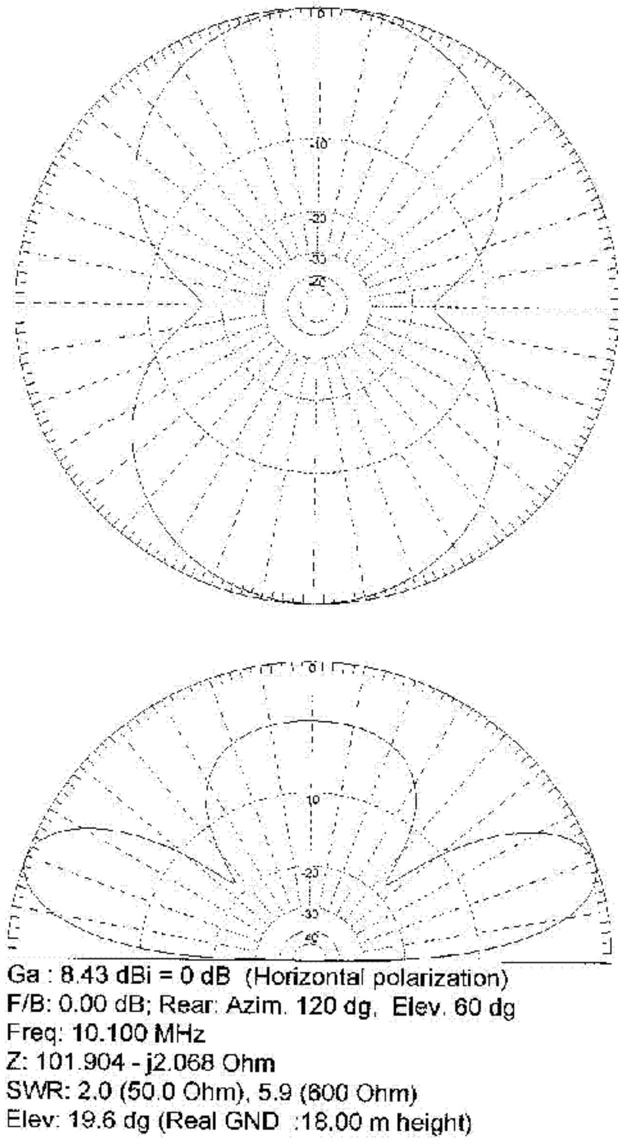

Ga : 8.43 dBi = 0 dB (Horizontal polarization)
F/B: 0.00 dB; Rear: Azim. 120 dg, Elev. 60 dg
Freq: 10.100 MHz
Z: 101.904 - j2.068 Ohm
SWR: 2.0 (50.0 Ohm), 5.9 (600 Ohm)
Elev: 19.6 dg (Real GND :18.00 m height)

Figure 14.3: Radiation Pattern of a 30-Meter Delta Loop on 30 Meters

The Bottom Wire is at 18 Meters above Ground.

The horizontal pattern shown above demonstrates that the 30-meter delta loop has a bi-lobal pattern broadside to the plane of the loop. The vertical pattern below the horizontal pattern shows both high angle and low angle radiation. The angle of maximum radiation is at 35 degrees above the horizon. The angle of radiation straight up is only down about 1.5 dB. This is pattern demonstrates the vertical delta loop is good for both nearby stations as well as DX.

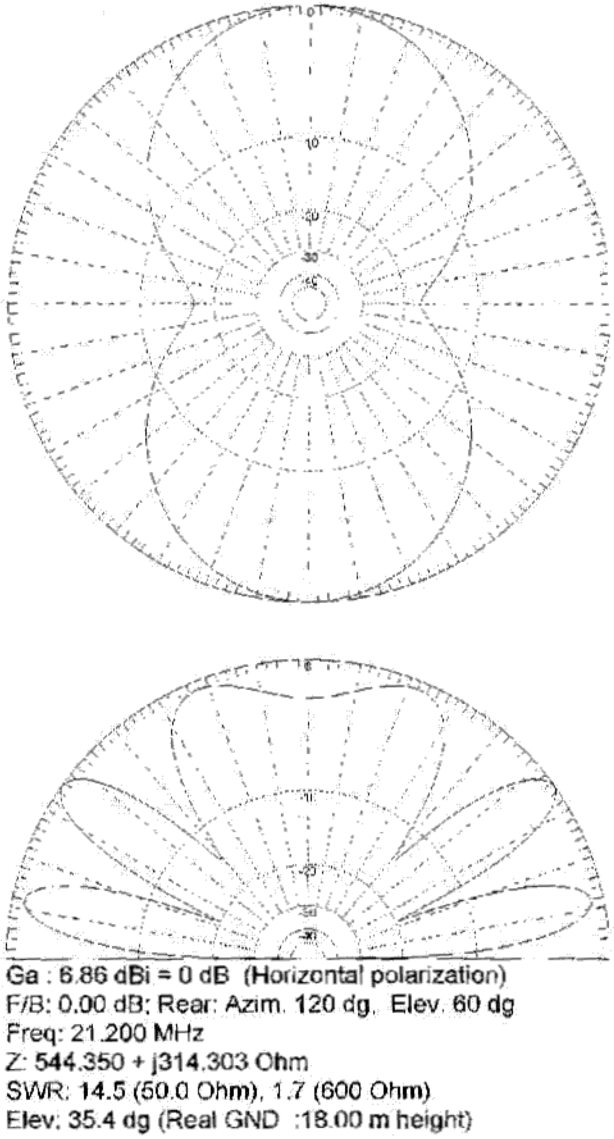
Ga : 6.86 dBi = 0 dB (Horizontal polarization)
F/B: 0.00 dB; Rear: Azim. 120 dg, Elev. 60 dg
Freq: 21.200 MHz
Z: 544.350 + j314.303 Ohm
SWR: 14.5 (50.0 Ohm), 1.7 (600 Ohm)
Elev: 35.4 dg (Real GND :18.00 m height)

Figure 14.4: Radiation Pattern of a 30-meter Delta Loop on 15 Meters

15.

Directional Beam Antennas

15.1 The Monoband Yagi

Between 1926 to 1929, Shintaro Uda and Hidetsugu Yagi developed a beam antenna that had sharp directivity and high gain. Later, work was done primarily by Mr. Yagi and yagi was the name given to the antenna until finally recognition was given to Mr. Uda. Its proper name is the Yagi-Uda Array. Most hams call it a beam.

A monoband yagi is the name given to a yagi for a single band. The performance of any commercially made monoband yagi is touted to have its dimensions tuned for maximum performance. As you will see later, this is not always the case. Monoband yagis being sold today are much improved over older designs because of computer modeling programs available.

The yagi is made of two or more aluminum elements mounted on and perpendicular to a boom. Hams use antenna rotors to turn the antenna in the direction of the station they want to work. However, there are wire beams, fixed in one direction, mainly on 80 meters, suspended between trees or other supports. Most high frequency beam antennas used by hams are in the horizontally polarized configuration, which means the elements are parallel to the ground. CB beam antennas and some 2-meter beams are vertically polarized with the elements at right angles to the ground (See *Section 3. Polarization Of Electromagnetic Waves* on page 5).

A 2-element yagi has a gain around 3 to 4 dBd. A two-element yagi will have a driven element with either a reflector or a director. The driven element is the only element receiving power directly from the transmitter. The reflector and directors are called parasitic elements because they receive power from the driven element by inductive coupling.

The 3-element yagi will have a gain of approximately 5 to 7 dBd or 7 to 9 dBi depending on its boom length. A three-element yagi has one reflector, one driven element, and one director. Because the yagi has a low radiation resistance, a matching system is located at the driven element feed-point. The ratio of the radiation off the front compared to the radiation off the back is called front-to-back ratio. Front-to-back ratio and forward gain are factors to be considered in choosing a yagi design. Both measurements are given in dB. All yagis have a good front-to-side ratio, with the signal off the side being 50 dB below the front.

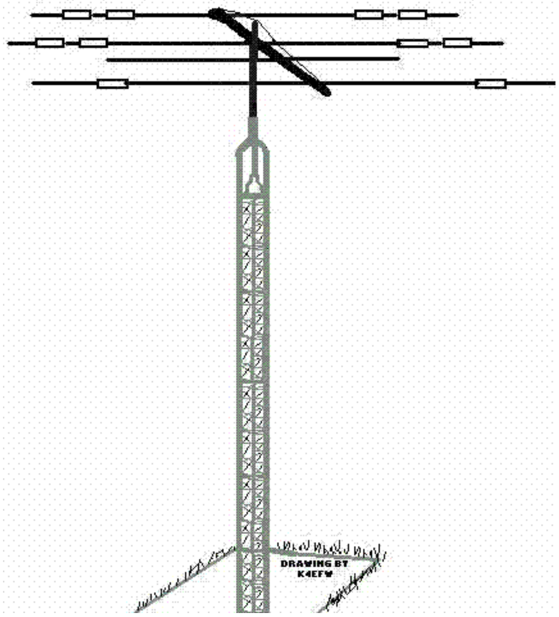

Figure 15.1: 3-Element Yagi

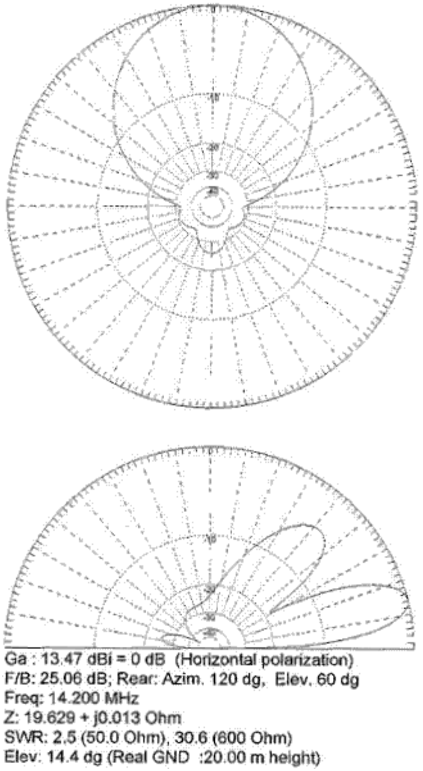

```
Ga : 13.47 dBi = 0 dB  (Horizontal polarization)
F/B: 25.06 dB; Rear: Azim. 120 dg,  Elev. 60 dg
Freq: 14.200 MHz
Z: 19.629 + j0.013 Ohm
SWR: 2.5 (50.0 Ohm), 30.6 (600 Ohm)
Elev: 14.4 dg (Real GND  :20.00 m height)
```

Figure 15.2: 3-Element Yagi Radiation Pattern

The reflector of a yagi is about 5% longer than the driven element. The reflector, being longer, will have inductive reactance. The inductive reactance shifts the phase of the re-radiated wave, which radiates and combines with the driven elements wave and reinforces it in the direction away from the reflector toward the driven element. A director is about 5% shorter than the driven element. The director, being shorter, has capacitive reactance, and this changes the phase of the reradiated wave to reinforce the wave away from the driven element opposite the reflector.

The gain of a yagi is derived from radiation being concentrated in one direction at the expense of the other directions. One hundred watts fed into a yagi with a gain of 6 dBd will have an apparent power of 400 Watts in the main lobe. Because one hundred watts put into a yagi radiates only one hundred Watts, and because that one hundred Watts of power is concentrated in the main lobe, it is equal to the power from a dipole being fed with 400 Watts. This is referred to as effective radiated power or ERP, but a yagi is not any more efficient than other antennas. Because of the Principal of Reciprocity, an antenna having a 6 dBd gain on transmitting will also have a 6 dBd gain on receiving.

Adding more directors and increasing the boom length will increase the gain of a yagi.

The front-to-back ratio ranges from 18 dB for a 2-element yagi to over 25 dB for a multi-element yagi, provided the parasitic elements are carefully tuned. The gain of a yagi is generally proportional to the boom length and not necessarily the number of elements. Doubling the boom length, while keeping the proper number of elements for that boom length, will add about three more dB of gain.

Tuning the yagi for maximum gain makes the bandwidth very narrow, and it will have a poor front-to-back ratio. For these reasons, we dont recommend tuning a yagi for maximum gain, because you will only increase the gain by a fraction of a dB at the expense of front-to-back and feed-point impedance. Tuning the yagi for maximum front-to-back will help eliminate interference coming from the rear of the antenna. The building of any yagi involves compromise spacing and element tuning.

As you make the yagi larger by adding directors, the main radiation lobe becomes narrower increasing the gain and ERP. The gain of a yagi with four elements is about 7 to 8 dBd. You used to see 3 or 4 element yagis advertised claiming a gain of more than 10 dB, but they never said if that gain was referenced to an isotropic or a dipole. That gain also involves the gain derived from signals reflected from the ground adding to the direct wave. A more realistic gain figure is the “free space gain.” Some companies, who sell monoband yagis, inflate their gain figures. Beware! Increased spacing of the elements will increase the gain of a yagi up to a point. Increasing the spacing past that point will reduce the gain. The spacing of a reflector or director needs to be in a range of 0.1 to 0.3 wavelengths. With a 3-element yagi maximum gain occurs with both parasitic elements spaced at about a quarter wavelength. Second and third directors can have wider spacing.

Most hams do not build yagis but buy them from the many companies who sell them. Ham catalogs are full of pre-cut and tuned yagis that come in boxes ready to be assembled in the back yard. Many of these are very good. However, there is a lot of satisfaction to be gained from building your own.

In 1971, we purchased, a 15-meter monobander being sold by a reputable company. Its performance was disappointing. It had only a 10-dB front-to-back ratio. That design is no longer being sold. After reading some books, we readjusted the antenna elements to some new dimensions and it performed much better. This was the beginning of our yagi building.

During the last nearly 50 years, we built many yagis. During the period of 1979 until 1986, many multi-element yagis were constructed, gain measured, formulas derived for spacing and element length, and the radiation patterns plotted on graphs. In 1986, a computer program titled *Yagi* by Dean Straw, N6BV, was bought. From that point on, that program was used to design and set the element lengths to their proper values. Not much difference in performance of the new designs was seen over what was previously used, but tuning parasitic elements and running back and forth to the field strength meter was eliminated. There are many better computer programs available today for designing yagis and other antennas.

The largest yagis we built were a 4 element 20-meter yagi on a 38-foot boom, a 5 element 15 meter one on a 27-foot boom, and a 5 element 10 meter beam on a 24-foot boom. These are modest designs compared to some of the big antennas used by contest stations. All these yagis were stacked one above the other on a 20-foot mast coming out of the top of the tower. The 20 meter one was on the bottom, next came the 15-meter, and the 10-meter yagi was on top. This method of stacking yagis for different bands one above the other makes what is called a “Christmas tree array.” These antennas worked well. Since retiring and moving back home, we use pre-tuned directional antennas because of the lack of a good place for an antenna range. Climbing is not now an option because of age and infirmity.

If you make the reflector 5% longer than the driven element and the director 5% shorter than the driven element, you will be pretty much in the ballpark. The beautiful part about a yagi is it will work reasonably well with the element lengths only in the ballpark. By carefully tuning, you will get a fraction of a dB more gain or a few more dB front-to-back, because the spacing and diameter of parasitic elements affect the length required for those elements. A yagi can be tuned for maximum forward gain, maximum front-to-back ratio, or best impedance, but you can achieve only one of these conditions at a time. Element tuning, at best, is a compromise.

Most hams who are yagi builders do not tune their antennas at all, but use published dimensions for building them. Yagi builders who do tune, tune for either gain or front-to-back and then match the driven element with a gamma match, hairpin match, a series-resonant coax matching section, or a step down balun. The feed-point of a properly tuned yagi is close to 25 ohms.

Formulas for calculating yagi element lengths will not be given in this book. Because yagi elements are made from telescoping aluminum tubing, the elements will be tapered. The diameter of the elements and the taper determine the lengths required for tuning of the elements. A tapered element will resonate higher in frequency than one not tapered. The formula to calculate the length of the tapered elements is complicated, but there are computer programs to do that.

15.2 Trapped Multi-Band Yagis

Some yagis have traps in the elements to make them into a multi-band beam. Many of these commercially made antennas are available at ham radio stores or directly from the manufacturers. In a 3-element, 3-band design, the spacing on the booms is a compromise. A 3-band beam is known as a "tribander." The spacing is close on 20 meters, optimum on 15 meters, and wide on 10 meters. You cannot tune the trapped elements for maximum performance on three bands simultaneously and have a good match on all those bands. Since a good match is important to most hams, gain and front-to-back ratio are sacrificed for a good match on triband beams

The inductors in the traps load the elements in triband beams. Therefore, the elements are shorter than the elements of a 20-meter monobander. Regardless of the compromised design, a triband-trapped beam is much better for working DX than a dipole. Many hams have achieved working over 300 entities with tribanders having short booms.

The radiation pattern from a yagi is at a lower angle than a dipole. This gives the impression a yagi has much more gain than it does. A dipole has unity gain, but that gain will be at a higher angle. The dipole puts out a weaker signal at the low angles needed to work DX, and a yagi puts a strong signal at low angles. In comparing a dipole to a yagi, the yagi may only have a 4 dBd gain in its major lobe. The gain of the yagi at a low angle may be 10 dB or so better than a dipole at that same lower angle. The gain of any antenna is always measured in its major lobe, irrespective of where the angle ar which the maximum radiation lobe occurs.

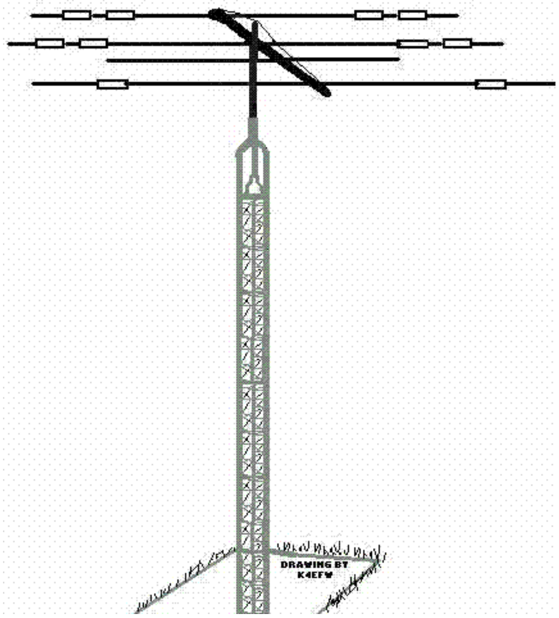

Figure 15.3: Trapped 3-Element Yagi

Figure 15.3: Trapped 3-Element Yagi shows two sets of traps in two of the elements and one set in the rear element. The front element is the director with traps for 10, 15, and 20 meters (it takes two sets of traps to make the elements work three bands). Directly behind it is the driven element with traps also for 10, 15, and 20 meters. The rear element is trapped for 15 and 20 meters (a single set of traps makes it work two bands). The entire lengths of the three longest elements are resonant on 20 meters. The short element is a reflector for 10 meters. Only the part of the antenna between the 10-meter reflector and the front director is used on 10 meters. The maximum signal is radiated in a direction coming out of the page toward you. Mosely

builds trapped antennas that have two traps in one enclosure and you can not determine the bands from the traps as you can on Hy-Gain and Cushcraft beams.

Some triband beam models as the one above are built with longer booms so they would have more gain on 20 meters, a good match on all bands, and optimum 3-band performance. They achieve this by interlacing extra monoband reflectors and directors on the boom placed between the 20-meter elements as is done with the antenna in *Figure 15.3: Trapped 3-Element Yagi* on page 63. The extra elements have no effect on 20 meters or any band for which they are not resonant. Some amateurs mistakenly think the extra elements work on all bands, but they dont. The Cushcraft A-4 shown is not a beam with four working elements on any band. The old Hy-Gain TH6DXX and Mosley Classic 36 had six elements on the boom. They both had three trapped elements and three monoband elements. They had three working elements on 20 meters, three on 15 meters, and four elements on 10 meters. The trapped reflector worked on 15 and 20-meters. The trapped driven element worked on all three bands. The trapped director worked on 10 and 20 meters. On the boom was a resonant reflector for 10 meters and one each resonant directors for 10 and 15-meters. When using one of them, we have often heard amateurs saying they were using a six-element beam. This gave the other station the mistaken idea they were working someone with an antenna with six working elements. Other beam antennas interlace additional elements of different lengths to make the tribander into a 5-bander covering 20, 17, 15, 12, and 10 meters. Hy-Gain makes a 5-band yagi for 20 through 10 meters that has 11 trapped and monoband elements. It is the Hy-Gain TH-11. Mosely makes a 6-bander that includes two elements for 40 meters. It is the Pro-67.

In order to achieve better SWR curves over a wide bandwidth, some triband yagis have two driven elements spaced 3 to 5 feet apart. The front driven element is shorter than the rear driven element. Both driven elements are trapped. This double driven element scheme is called a log-cell. A log cell, by itself, has a small gain and may slightly increase the overall gain of the tribander. The KLM KT-34 and the HY-Gain TH-7 are examples of this kind of antenna.

Is a monobander better than a tribander? We dont know if our tests can be duplicated and no one else has ever said he has actually compared the two antennas. It is “common knowledge” that traps have loss. Therefore, the ham fraternity believes a monobander has to be better. From the tests we performed here, we believe it is a myth a monobander is significantly better than a tribander having an equal boom-length. We believe the traps do not have enough loss to make enough difference to matter. However, monobanders having very long booms and many directors will outperform any tribander.

Having two towers, both having the same height and being 100 feet apart, made it possible for us to do the experiment described here. The result is useful information because it was made in a real world situation that would be comparable to the average hams location. Both antenna element lengths were set to Hy- Gain specifications. The constants were terrain, antenna height, antenna boom length, frequency, coax length, and power level. The only variable in the tests was the two antennas being tested. The test was performed to see how much loss antenna traps have. Had there been more than one variable, the tests would not have been valid, because in any scientific experiment, the test is valid only when one variable is being tested. In addition, more than one test has to be made in order to average out the collected data errors. In this case, many tests were made.

On one tower was a 20-meter four-element Hy-Gain 204-BA monobander with a boom-length of 26 feet. This antenna is arguably not one of the best monobanders made, but it is what we had and it was about the same size as our tribander. On the other tower was a trapped 6-element Hy-Gain TH-6 DXX tribander having a 24-foot boom. The entire tribander boom-length was used on 20 meters, so both boom-lengths were comparable.

The transmitted signal strength of the two antennas was compared on 20-meters. This test involved many DX stations and one local amateur 5 miles away. With both antennas pointing toward the receiving station, a carrier power of 10 watts was fed from the transmitter, and held constant while the antennas were “hot” switched several times. (The power level was unimportant as long as it was held constant on both antennas). None of the many DX stations involved in this test could see any difference in either antenna, and, yes, their analog meters could discern a difference of one dB. These tests by themselves were not conclusive because of the possibility of fading signals (QSB). A second series of tests was performed with a local ham when 20 meters was dead. Testing with him was done to eliminate QSB from spoiling the results. He could also measure no difference on his S-meter. He could also see a one-dB difference on his analog S-meter. As a third series of tests, the antennas were switched while we looked at the signals on the S-meter from distant stations and the local station. No differences in received signals were noted. Maybe the difference was a monobander has only a few tenths of a dB less loss, such a small amount of difference no one was able to see it on receiver S-meters. Certainly, the difference in the two antennas was less than one dB.

Conclusion: The Hy Gain TH6DXX and the 204-BA antennas perform equally well on 20 meters at a height of 56 feet.

15.3 The SteppIR Antenna

The latest developments in yagi designs are found in the ones being sold by SteppIR Antennas. There are two, three, and four element versions. All these versions are frequency agile and cover continuously from 13.5 to 54 MHz. The MonstIR adds three very long elements for 6.9 to 13.5 MHz. The elements are made of fiberglass tubes with beryllium-copper ribbons inside. Each element has stepping motors to wind and unwind the copper ribbons to change their lengths inside the tubes. A multi-wire control cable connecting the control box to the stepping motors accomplishes this. The proper element lengths for all frequencies in its range have been calculated by a computer and stored in the control boxs computer. As you move from frequency to frequency, the control box in the shack readjusts each element length. Thus, the antenna is configured into a properly tuned monobander for any frequency in its range. These antennas are expensive, but the hams who own them say they are worth the money.

15.4 The Log-Periodic Array

Another beam antenna that looks like a yagi is the log-periodic antenna. It is configured using many elements with each element being shorter than the one behind it. This means the longest element is at the rear of the array and the shortest element is at the front. All elements are divided in the center and insulated from the boom, and all elements are driven. On both sides of the insulator at the center of each element, wires run from the front element of the array to the rear element. Each wire criss-crosses the other ones but they do not touch. That makes a 180-degree phase reversal from one element to the next one behind it. The feed-point is across the insulator at the shortest element. The feed-point impedance is about 200 ohms and a 4:1 balun is used to feed it.

The advantage of the log-periodic antenna is, it that it is very broad banded and it can cover all frequencies with an SWR below 2:1in its design frequency range. The disadvantage is the gain of a log-periodic antenna is lower than a yagi with an equal boom length. There are designs being sold today that cover continuously from 14 to 30 MHz. In Fort Gordon, Georgia, there used to be a monster log-periodic at the MARS station that covered from 2 to 30 MHz. The boom length was 120 feet and the antenna was rotatable.

15.5 Directional Cubical Quad and Delta Loop Antennas

We built a number of quads at various times and with them on the test stand and with the bottom wire a foot or so above the ground, worked many DX stations. When we built yagis and they were on the same test stand nine feet above the ground, we could hardly get a signal out of the back yard. Since the vertical beam-width of a quad is narrower than a yagi and the radiation angle is lower, the quad will work better at low heights. Because of its lower angle radiation, many quad users claim a quad “opens and closes” the band.

The two-element cubical quad is a square-or diamond-shaped loop antenna that has a second loop acting as a parasitic element. The quad configuration has all loops in the vertically oriented plane as *Figure 15.3: Trapped 3-Element Yagi* on page 63 demonstrates. Feeding it in one of the horizontal wires results in horizontal polarization, and feeding it in one of the vertical wires makes it vertically polarized. Every two-element quad being sold today uses a reflector for the parasitic element, although it is possible for it to have a director. The theory of operation is the same as that of a yagi.

Some quad builders believe a diamond-shaped quad has more gain than a square-shaped one. Their logic is that since the maximum current points of both wires are spaced farther apart than with a square quad, the increased spacing of the current points should produce higher gain. To find out if this was true, we built both a diamond-shaped and a square-shaped quad for two meters. Using a commercial field-strength meter connected to a receiving antenna, we fed equal amounts of power to both antennas and measured the radiated field in each ones major lobe. Field strength measurements were made a few wavelengths away and many wavelengths away from the quads. No difference in the radiated field of either could be found.

According to Bill Orr in his book about cubical quads, a two-element cubical quad is equal to a pair of 2-element beams; one is stacked over the other a quarter-wavelength. The ends of the beams bottom driven element are bent up and the top element has its ends bent down where the ends of the top and bottom elements are joined together on the side. When they are joined, this forms the square we call a quad. The bottom element is then feeding the top element from its ends. The parasitic elements have the same configuration except the wire loop has the ends bonded together to form a continuous square.

There are multi-element quad designs that use one or more parasitic directors in addition to the reflector. Adding a director will lower the feed point impedance. The wire of the reflector is about 3% longer than the driven element, and each director has about 3% less wire than the driven element. Adding directors to a two-element quad makes the horizontal beam width narrower, producing more gain.

Another quad design “the delta loop” uses triangular-shaped driven elements. One or more triangular-shaped parasitic elements make the antenna complete. Theoretically, the delta loop antenna will have slightly less gain than the cubical quad, because there is less enclosed area in the triangular loop. We believe that there are no instruments available to hams to be able to measure the difference.

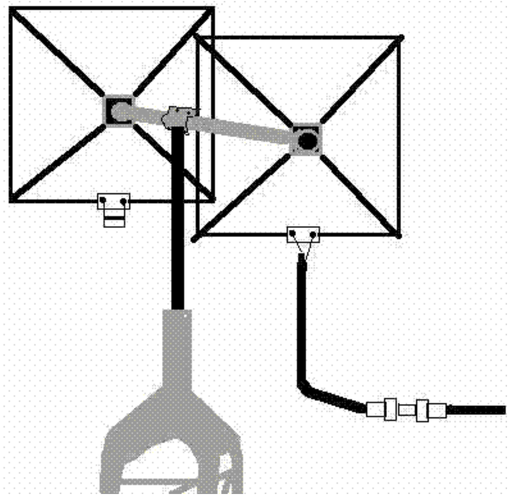

Figure 15.4: Single Band Cubical Quad

Most cubical quad and delta loop antennas that can be rotated are used on 20 meters and higher. A few ambitious hams have built rotatable quads for 40 meters. Others have made 80-meter quads, supported between trees, fixed in one direction.

In order to make the quad smaller, adding loading coils or linear loading sections in its wires has been suggested, but that will defeat the purpose of using a quad. Because the quads gain is produced by the enclosed area inside the loop, reducing the enclosed area will result in less gain.

The wires for quads for 20 through 10 meters are strung around the perimeter of an “X shaped frame made of fiberglass poles or bamboo. Each element has its own X-shaped frame. A smaller X-shaped metal structure, called a “spider,” attaches the poles to the boom. The poles are referred to as “spreaders.” The four spreaders attached to the spider form the “X.” The “X” can be rotated 45 degrees on the boom to form a diamond-shaped quad instead of a square quad. A few have tried with limited success to make the spreaders out of PVC or aluminum.

A wire is attached near the ends of the spreaders to form a loop around them. The two ends of the wire are connected to an insulator to attach the feed-line as is done on a dipole. The quad loop has a theoretical feed-point impedance of 100 ohms. To match it, you can use a quarter-wave matching section of 70-ohm coax, a gamma-match, or a 2:1 balun. More on this is in another paragraph. The delta loop is made much the same way, but it requires only three spreaders to form an equilateral triangular loop. It is matched the same way since the feed-point impedance is about the same.

The reflector and director are formed the same way as the driven element except the two ends are shorted together to form a continuous loop. In order to get maximum performance from a quad you need to tune the reflector for either maximum gain or best front-to-back ratio. For tuning purposes, the wire of the reflector is cut a little shorter than calculated and the ends of the loop are connected to an insulator. A shorted stub, consisting of two parallel wires, is connected to the loop ends and hangs down from the insulator. Another wire is shorted across the two parallel stub wires. The shorting wire is moved up and down the stub to tune the reflector. The stub is a means of adjusting the total length of the reflector. See *Figure 15.4: Single Band Cubical Quad* on page 67. A field strength meter is needed to do this and you need a large area and two people. The field strength meter needs to be placed several wavelengths away from the antenna. Low power is fed into the antenna while it is tuned. One person tunes the reflector while the other person reads the field strength meter. Tuning the reflector involves tuning the stub for minimum signal off the back. Once the shorting wire has found its proper position, it is soldered in place. Quads made by the formulas work satisfactorily without tuning. Tuning for maximum front-to-back ratio instead of maximum gain will do more for the performance of the quad.

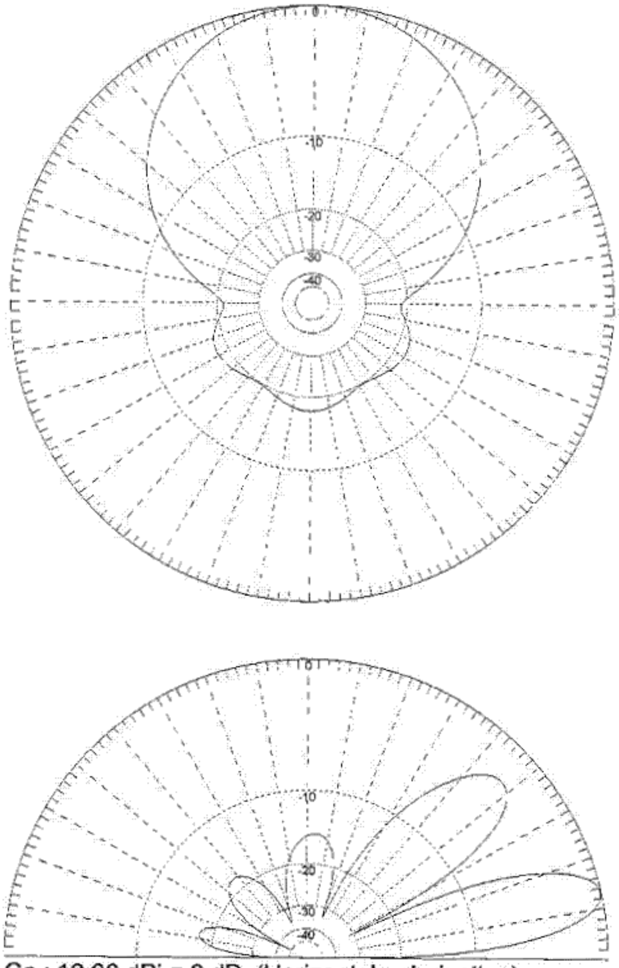

Ga : 13.36 dBi = 0 dB (Horizontal polarization)
F/B: 17.62 dB; Rear: Azim. 120 dg, Elev. 60 dg
Freq: 14.150 MHz
Z: 86.586 - j2.470 Ohm
SWR: 1.7 (50.0 Ohm), 6.9 (600 Ohm)
Elev: 13.1 dg (Perfect GND :20.00 m height)

Figure 15.5: Radiation Pattern of a Two Element Cubical Quad at 65 Feet

Some believe you cannot stack another antenna above a quad. They assume that because the quad has both vertical parts of the loop, a metal vertical mast will couple to the vertical part and detune the quad. We believe a vertical mast will have to be resonant at the operating frequency to detune a quad. Using MMANA, the quad was modeled with a metal vertical mast going through the plane of the quad. The only difference observed was that the resonant frequency was changed by a couple of kHz. The gain and front-to-back remained the same.

The gain of a two-element quad is nearly the same as an average 3-element yagi. The best part of the gain of a quad is the vertical beam width, or H-plane of the major lobe logically should be narrower due to it being equal to two stacked beams. For this reason, the 2-element quad has a lower angle of radiation. A horizontally polarized quad should have a slight advantage over a yagi. A lower angle is better for working DX. While operating using both quads and yagis, we have noticed that the horizontal beam width or E-plane of a quads pattern is wider than a 3-element yagi. We believe the horizontal beam-width of the quad is the same as a two-element yagi. This is why the 2-element quad is not as directional as a 3-element yagi.

Modeling our 2-element quad in free space on 20 meters, we found its gain to be 5.49 dBd. The boom-length of the quad is 8 feet. A three-element yagi with a boom-length of 16 feet will have 6.4-dBd free- space gain on 20 meters. The free-space gain of a 20-meter optimum spaced monoband yagi on a 25-foot boom will only have slightly more gain.

The compromise spacing for a 2-element multi-band quad for 20 through 10 meters is 8 feet. This spacing is 0.115 wavelength on 20 meters, 0.175 wavelengths on 15 meters, and 0.23 wavelengths on 10 meters. These spacings are within acceptable limits. For a single-band 20-meter quad, space the elements 12 feet apart. If you want to build a 12 and 17-meter dual band quad, the spacing will be 8 feet, the same as it should be for 15 meters. Eight feet is also a satisfactory spacing for a 10-meter quad, but it can be as close as four feet.

With smaller perimeter requirements, loops for the higher bands can be strung inside and parallel to the lower band loops to make a multi-band quad. It is easier to make a multi-band quad than a multi-band yagi. Quad kits for triband and 5-band quads are available. These kits cost less than a multi-band beam.

In the construction of most quads, an insulator is put in the bottom horizontal part of the wire on the driven element so it can be fed like a dipole. A 2-to-1-balun transformer will match the feed-point to 50-ohms, then you can tie all the feed-points of a multi-band quad together. The Lightning Bolt Antennas 32MCQ/WB quad feeds five loops this way and the SWR is 1.4:1 or less on all five bands. The person manufacturing the Lightning Bolt quad went out of business on December 12, 2005.

With other more complicated schemes, each quad loop is fed separately, and each loop uses a 70-ohm odd multiple of a quarter-wave series matching section placed between the 50-ohm coax and the feed-point. Used this way, the quarter-wave matching section will match 50 ohms to 100 ohms. A remote antenna switch will have to be mounted close to the feed-point to select the desired loop. Other builders use a separate gamma-match on each driven element to get a perfect match to a 50-ohm coax but this method would also require a remote antenna switch. Without the switch, several pieces of coax, one for each band, would have to be run into the shack.

If you are going to build a monoband quad, you need to use the following formulas to cut the wire loops to these approximate lengths:

- For the driven element, you divide 1005 by the frequency in MHz.
- For the reflector, you divide 1030 by the frequency in MHz.
- For the director, you divide 968 by the frequency in MHz.
- Make any additional directors the same length as the first one

These formulas were derived experimentally from tests run here. The exact measurements will be determined by the element spacing, but the lengths cut by these formulas will be very close for any reasonable spacing.

After giving you the advantages of a quad, here are the disadvantages: The two-element quad for 20 meters is large vertically and horizontally. When the 20-meter quad is on the ground, the boom is 8 feet high and most people arent tall enough to maneuver it by holding the boom. Some quads, which are made from lightweight materials, are flimsy, and they will suffer during wind and ice storms. The best-constructed quads have their spreaders made of heavy fiberglass. Those quads, although they are heavier, stand up well under adverse weather conditions. The Lightning Bolt quad used here has stood up very well during three ice storms in the past two winters.

Here is some information we discovered after originally writing this book. The MMANA antenna-modeling program does not perform very well when modeling a quad. On 20 meters, the modeling program says the front-to-back ratio of our quad is at a maximum at 14525 kHz, but actually, it occurs at 14050 kHz. The measurements of actual front-to-back were made using a field strength meter. We reduced the power levels off the front to give the same field strength reading we got off the back. The front-to-back ratio in dB was calculated from the two power levels. What was interesting was the maximum front-to-back ratio occurs at a single discrete frequency and the front-to back deteriorates somewhat on either side of that frequency. While looking at the radiated field off the front, the field strength does not vary one dB across the whole band. Maximum gain and maximum front-to-back was very close to the same frequency. Not having tested them in this way, we believe yagi beams perform the same way regarding front-to-back ratio and gain. In working stations, the gain is the important parameter. Front to-back ratio is important in reducing interfering signals from behind the antenna. We decided to tune our quad for maximum front-to-back rather than for maximum gain. The next step is to lower the quad and carefully tune the reflector for each band. After running tests to determine the frequency where maximum front-to-back occurs, we found the maximum measured front-to-back ratio was 22 to 23 dB.

Good news! After writing the above paragraph, we lowered the multi-band quad and reduced the reflector element lengths on the two-element Lightning-Bolt Quad. The original reflector lengths were too long according to the field strength readings we made. The formulas that were originally used to cut the reflector lengths were anywhere from 1029 to 1036 divided by the frequency in MHz. We derived from field-strength measurements that the maximum front-to-back ratio occurred when the reflector length was cut by dividing 1022 by the frequency in MHz. While searching the Internet, we discovered EI7BA in Ireland used 1019 divided by the frequency on his multi-band quad. We decided to use his formula and we could lengthen the reflector by adjusting the stubs if necessary. After cutting the reflectors to the new dimensions, we made new field strength readings. The front-to-back ratio occurred near the frequency of our calculations. In addition, the frequency of maximum field strength from the front also occurred inside each band. As an example, today we were listening to GD4PTV on the Isle of Man on 17 meters. On the front of the quad, he was S7. With the quad 180 degrees from him, he was inaudible. We also found that other stations were down by at least six S-units off the back. Originally, stations off the back of the quad were down only two S-units.

Several months later it was discovered that a multi-band quad tunes differently from a single-band quad because the interlaced elements react to detune each other. That may be the reason we found the reflectors of a multi-band quad needed to be different lengths than the 1030/frequency formula.

Here are some words of wisdom about using a field-strength meter in trying to tune a quad:

Tuning the quad reflector stubs can give you misleading data. If you look only for the minimum signal from the back of the quad, that may not be the point of best front-to-back because you may have detuned the quad so that the signal from the front may have also deteriorated.

Using a field strength meter, keep the receiving antenna as short as possible to prevent the receiving antenna from being nearly resonant. Certain receiving antenna lengths seem to be frequency sensitive, that is, as you change frequency toward the receiving antennas resonant point, the field strength meter will give a false higher reading. The only problem with using a very short receiving antenna is the meter may not have enough sensitivity to make measurements from the back of the antenna.

It will be impossible to achieve a high front-to-back ratio on certain bands on a multiband-quad because the reflector wires for adjacent bands affect the tuning by interacting with each other. When the Lightning Bolt Quad was designed, we are not certain which parameter was used in its design (gain, SWR, or front-to-back). From the field strength readings made with that design, it was impossible to draw any conclusions. If you are going to build a quad, what we said about designing a yagi is also is true for the quad: you can tune for best gain, best front-to-back, or best impedance match. You cannot tune for more than one of these parameters at a time. To tune a quad for maximum gain is relatively easy using a short antenna on the field-strength meter.

While we were trying to measure the frequency of the highest forward gain on 17 meters, we found the maximum field strength occurred at the high end of the band on one receiving antenna. Subsequently, it was strongest on the low end of the band on another receiving antenna having a different length. No changes had been made to the dimensions of the quad in either case. Trying to move the frequency of the maximum field to the middle of the band, we adjusted the length of the reflector stubs and it made no difference to the frequency where the maximum field occurred. What caused the error was we were trying to measure the field strength 100 feet in front of the quad. The long receiving antenna connected to the field strength meter was acting like a parasitic element and was not accurately measuring the signal being radiated from the quad.

15.6 The Quagi

A variation of the quad and the yagi is a marriage of the quad and the yagi called the quagi. The quagi has a quad driven element, quad reflector, and yagi directors. Hams who have built the quagi report the yagi directors work better than quad directors, but we have never compared the two types of directors.

At one time we converted a 2-meter yagi to a quagi and compared the field strength readings from both configurations. By changing the driven element and reflector to quad loops, we measured a signal increase of 1.8 dB. We also experimented to see what effect the quad reflector had on the signal. While using the quad driven element we changed the reflector back to a yagi reflector. What was surprising to us was the configuration of the reflector had no affect on the radiated signal. Only by changing the driven element from a yagi element to a quad element made any change in the field strength. All these field strength readings were made using a commercially manufactured field strength meter. To insure our readings were valid, the power being fed to the antenna was measured and kept constant.

16.

Gain Versus Front-to-Back

As we have said before the front-to-back ratio of a multi-band cubical quad can be maximized by careful tuning to achieve about 18 to 20 decibels front-to-back ratio. A properly designed yagi can achieve a front-to back ratio of better than thirty decibels. A two-element quad has about the same gain as a three-element yagi. You can tune a yagi or quad to either maximum front-to-back or maximum gain. You can also tune them to compromise settings somewhere in between. The question arises as to which maximum should either antenna be tuned? It is our opinion that either antenna should be tuned for maximum front-to-back ratio. In that case the maximum gain will be deteriorated by only a fraction of a decibel Let us explain why we reached that conclusion with an example.

Today we were on 17 meters to work VP8TD on Pitcairn Island in the South Pacific Ocean. He is a visitor to the island and will be there for about two more months as of this writing. A resident of the island, VP6TC, Tom Christian hasnt been heard from in months. I suspect he is getting elderly and doesnt get on much anymore. Anyway VP6TD had an enormous pileup going. We were using a three-element SteppIR yagi up sixty-five feet on our tower. Also, the amplifier puts out about 1490 watts on 17 meters. We make up for a lack of antenna forward gain with the amplifier. We worked him with one call through the pileup. The SteppIR replaced the two-element Lightning Bolt Quad about 10 months ago. When VP6TD answered us I could hear him over the pileup. From the rear of the antenna were several very loud Italian hams calling him, one of which continued to call even when VP6TD answered someone. The Italians were 180 degrees from the front of our antenna, or directly off the back of where we were beaming. Because of the superior front-to-back of the yagi, I could hear the Pitcairn Island station over the Italians. Had we been using the quad, the Italians would have been at least 10 dB louder and we could have found it impossible to make the contact.

Today, we were in contact with N4XPZ, Joe, on 75 meters while several more hams were talking about the VP6TD on 17 meters last evening. Joe said he tried to work the VP6TD station using a single wire antenna. He complained he could not copy the VP6 because of the Italians who continued to call even when the VP6 answered someone. That illustrated the point we are making in this section. The old adage is true: “You can’t work ‘em if you can’t hear ‘em!”

17.

Feed-Lines Commonly Called Transmission Lines

Always use the best feed-line you can afford. Resist the urge to be penny wise and pound-foolish. This is particularly true of coax. Better (less lossy) coax will cost more. This cable is carrying your precious RF signal to and from your antenna.

The most common feed lines used by amateurs are 50-ohm coaxial cables. There are many types of 50-ohm coax such as RG-174, RG-58, RG8-X, RG-8, RG-213, RG-8 foam, and 9913. In this book we will only discuss these types. A suffix letter such as an "A" or "U" may be attached to the "RG" numbers such as RG-8U or RG-58A. All these cables have a center conductor surrounded by a plastic insulating material, called the dielectric, and a copper braided shield covering it. There is a plastic covering on the outside of the shield to protect the conductors from water. The center conductor and the shield carry RF currents.

These are the common 50-ohm cables:

- RG-174 has a very small diameter, 0.101 inches. This cable is used to carry small amounts of RF between circuits in equipment. RG-174 has the highest loss and the least power handling capability of any coax. It is useless as an antenna feed-line because of its loss and low power handling ability.
- RG-58 is larger coax having a diameter of 0.195 inches. It can handle low power and can be used on the lower bands to feed antennas a one hundred feet or so away. It is not recommended to use RG-58 on 10 meters because it has a loss of 3dB per hundred feet on that band and half your power will be lost in the coax.
- The next larger cable is RG-8X, sometimes referred to as mini-8. Its diameter is 0.242 inches. The dielectric surrounding the center conductor is foam rather than the solid dielectric used in the most coax. Making cables with foam insulation can reduce the loss. Some hams are successfully feeding a kilowatt of power into RG8-X on the lower bands. You will lose 2 dB of power by using one hundred feet of RG-8X on 10 meters. On 80-meters the loss of this cable is negligible.

You will want to use RG-8 or RG-213 if you are planning to use a kilowatt or more of power from 160 to 10 meters or for short runs on VHF and UHF. RG-213 is RG-8 made to military standards. Both have diameters of 0.405 inches. This cable has lower loss and can handle higher power because it has larger conductors and a larger diameter dielectric. RG-8 can handle 4000 watts peak envelope power on the broadcast band. RG-8 has only about 1dB loss on 10 meters per 100 feet. The loss becomes greater and the power handling rating of any coax decreases as the frequency of RF is increased.

There is a lower-loss version of RG-8. It is called RG-8 foam. Beldens number for this product is 8214. Because of the dielectric being foam, a larger center conductor has to be used to keep the impedance 50 ohms. The loss resistance of the larger conductor is less than the smaller conductor used in regular RG-8. In addition, the foam having many air pockets has less dielectric losses. Other manufacturers also make RG-8 foam. One hundred feet of RG-8 foam has a loss of 0.9 dB on 10 meters. Many amateurs will not use RG-8 foam because they mistakenly believe the foam will soak up water. Cut off a piece of this foam material and put it into a container of water. It will continue to float ad infinitum, because it doesnt soak up water. Most of the water seen in coax gets between the dielectric and the plastic outer covering and within the braid shield. Water has also gotten into the strands of the center conductor. Water will get into any coax if the ends are not properly sealed.

Solid conductors have less loss at radio frequencies compared to stranded conductors. Braid has more loss than a solid conductor used for the coax shield. A much lower loss coax, especially for higher frequencies, is available. The Belden 9913 is this product. This coax has a solid center conductor and the shield consists of a coating of aluminum foil covered with braid. The aluminum foil is a solid conductor. The braid over the foil is used to make a good solder connection because you cant solder aluminum. The mostly air dielectric material used in this product requires the center conductor to be larger to make the impedance 50 ohms. Air dielectric also has less dielectric loss than solid. There are a few manufacturers making 9913 look-alike products. One hundred feet of 9913 will have a loss of about 0.66 dB on 10 meters. There is a coax that looks like 9913 but has a stranded center conductor to make it flexible. It has a little more loss. If you are going to use 9913 on an antenna that rotates, flexing the cable as the antenna turns will cause the center conductor to break. Run the 9913 to the top of the mast, and using a barrel connector, connect the 9913 to a short run of RG-8. Run the RG-8 across the rotor to the antenna.

Coax cables of other impedances are available such as 70-ohm cable. RG-59 and RG-11 are common 70 ohm cable. Hams, except to make quarter wave matching sections, do not use these cables much anymore. There are many other types of cable other than the ones described here.

Open wire feeders, ladder-line, or window-line have much lower loss than coax. The three types are essentially the same except for the method of insulating the two wires from each other. When making open wire feed-line, you should use solid conductors, as large a conductor as possible, and as little dielectric as possible. These factors make open wire have less loss. There is so-called ladder-line for sale, which is really window line, which is made with 16 gauge solid conductors. The solid conductors make for low loss. There is another grade of the same feed-line that has 14 gauge-stranded conductors.

18.

Antenna Safety

18.1 Erecting Antennas on Masts

Erecting antennas pose some danger especially if they encounter power lines. Never erect an antenna near a power line. Make sure to leave enough clearance so if the antenna supporting structures fall they will clear the power lines. There are many cases of metal masts being raised accidentally encountering power lines, electrocuting the person or persons raising the masts. To raise a mast can expose you to a large force called leverage, which appears to increase the weight of the mast. Exerting oneself to raise a heavy mast can result in painful muscle and back injuries. Never try to raise a mast without sufficient help.

18.2 Tower Safety

A tower is a wonderful device for supporting wire and beam antennas, but a person who has never put up one should seek advice of people who have experience in erecting towers. The obvious danger is falling off the tower. It should never be climbed without a climbing belt. Most people falling off a tower do so because of some kind of equipment failure or the tower collapses because of overload.

In erecting a tower, a gin pole strong enough to support the weight of the tower section being raised should be used. Do not use improvised gin poles, as the strength of them may not support the weight of the tower section and the force from the other end of the rope being pulled by the ground crew. To hold a 50-pound tower section stationary requires a hundred pounds of force, which is the weight of the tower section and 50 pounds of force of the ground crew. The ground crew must exert more than 50 pounds of force to cause the section to be raised. There would be no greater tragedy than the gin pole breaking dropping the tower section on the ground crew. Then there is the possibility of the person on the tower being knocked loose by the falling, broken gin pole.

Another problem can arise if under-sized guy cables and clamps are used to support the tower. We have seen tower failure when guy cables broke in a windstorm, or an insufficient number of clamps holding the guy cable allowed the cable to pull through the clamps. Professional tower people do not use cable clamps. They use “preformed tie-wraps” that grip the guy cables tighter as the force in the guy cables increases. Preformed tie-wraps are available from Texas Towers. Never tie the ground end of a guy cable to a tree. A tree swaying in a heavy windstorm can put enough force on the cable to cause it to break or to pull the tower over. Screw-in anchors available from mobile home suppliers make adequate anchors. Do not anchor a guy cable where a tree can fall across the guy cable. This could break the guy cable and cause the tower to fall. Never place a tower near a house, where if it falls, it could hit the house. Remember *Safety First!*

These lengths are for coax having a solid dielectric with a velocity factor of 0.66 and foam dielectric with a velocity factor of 0.78. , You can use odd multiples like 3, 5, 7, etc. of the lengths below if those lengths are too short for your installation

Table 18.1: Quarter Wave Matching Sections of 70-ohm Coax

		Foam dielectric cable
160 meters	85ft 6in	106ft 4in
80 Meters	43ft 4in	51ft 9 in
40 Meters	22ft 6in	27ft 6 in
30 Meters	16ft 0in	19ft 6 in
20 Meters	11ft 5in	13ft 10 in
17 Meters	8ft 11 in	10ft 9 in
15 Meters	7ft 7 in	8ft 4 in
12 Meters	6ft 6in	8 ft 0in
10 Meters	5ft 8 3/8 in	6ft 11 in

Questions? Email Jim Abercrombie *n4ja@prtcnet.com*.

Hamuniverse.com is Powered by Ham Radio!
© 2000 - 2008 N4UJW Hamuniverse.com or article author! - All Rights Reserved.

$14

More Wire Antenna Classics

Volume 2

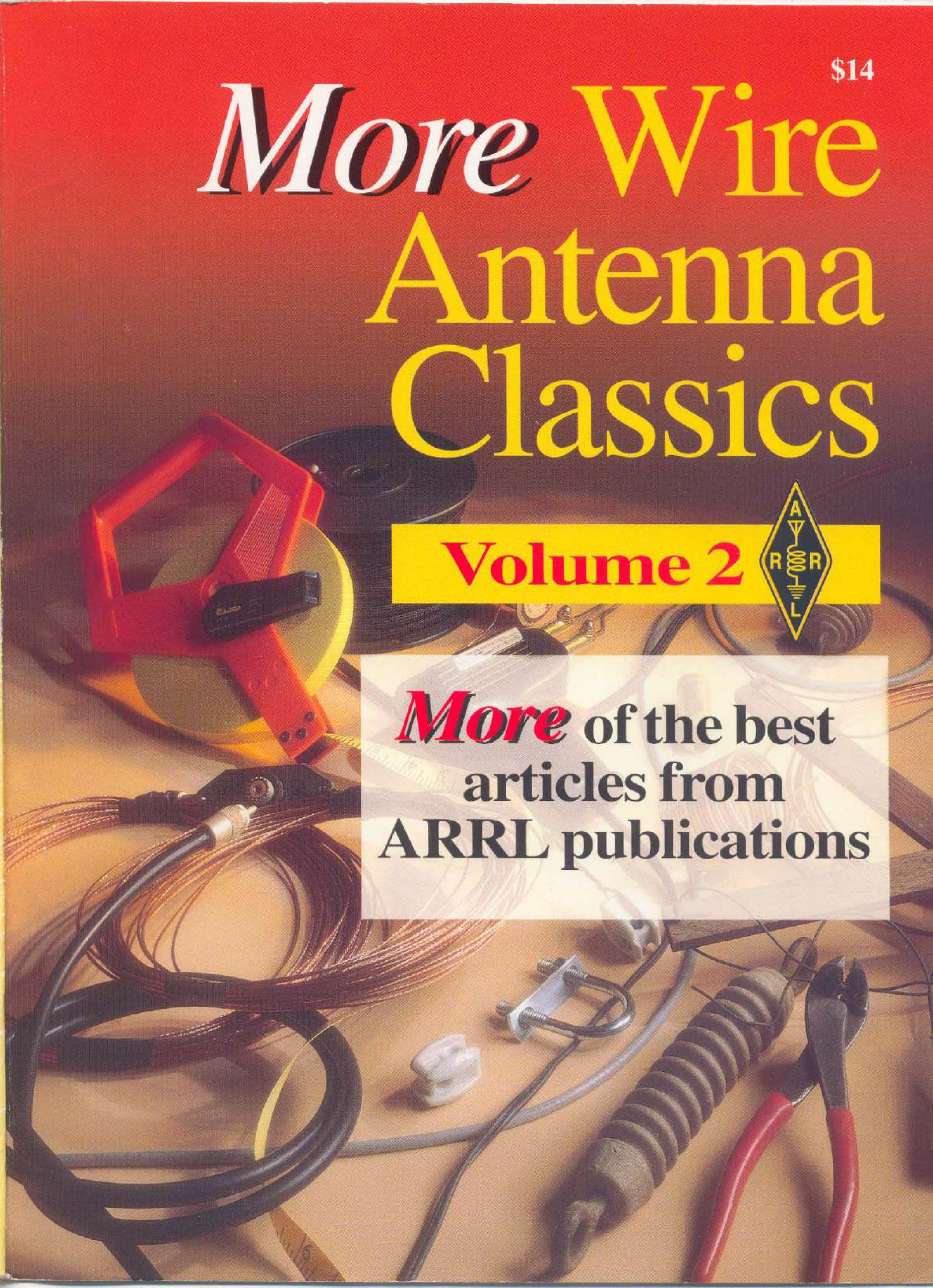

More of the best articles from ARRL publications

More Wire Antenna Classics Volume 2

More Wire Antenna Classics

Volume 2

Compiled by Chuck Hutchinson, K8CH

Production:
Paul Lappen

Cover Design:
Sue Fagan

Published by:

ARRL *The national association for AMATEUR RADIO*

225 Main Street
Newington, CT 06111-1494

Contents

1 Dipoles

2 Off-Center-Fed Dipoles

3 Multiband Antennas

4 Loop Antennas

5 Collinear Antennas

6 Wire Beams

7 Vertically Polarized

8 Receiving Antennas

9 Miscellanea

CHAPTER ONE

DIPOLES

1-1 **An Inverted V for Field Day**
From June 1968 *QST*

1-4 **E-Z-Up Antenna for 75 and 40**
From October 1961 *QST*

1-6 **A Zip-Cord Special Antenna**
From May 1972 *QST* (Hints & Kinks)

1-7 **Zip-Cord Special Antennas–Do They Work?**
From March 1979 *QST*

1-9 **Piggyback Antenna for the Ten-Meter Band**
From April 1973 *QST* (Hints & Kinks)

1-10 **RFD-1 and RFD-2: Resonant Feed Line Dipoles**
From August 1991 *QST*

1-14 **Off-Center-Loaded Dipole Antennas**
From September 1974 *QST*

1-20 **A Simple Broadband Dipole for 80 Meters**
From September 1993 *QST*

1-25 **Once More With the 80-Meter Broadband Dipole**
From June 1985 *QST* (Technical Correspondence)

1-26 **Fat Dipoles**
From *The ARRL Antenna Compendium, Vol 2*

By Frank Gue, VE3DPC, ex-VE6BH

From *QST*, June 1968

An 80-Meter Inverted V for Field Day

It's that time of year again when portable operation is a delight. Here is a full-sized low-frequency antenna that you can put up in the field or at home without any help.

A Field Day antenna for 80 meters should be light, portable, inexpensive, capable of easy erection without any special equipment, and it should not contain any fragile parts.

Of course, the antenna must perform adequately.

The popular inverted V suggests itself as an answer to these requirements. It can be constructed so that it guys itself, minimizing the inevitable exasperating tangle of rope and wire, and it requires only one mast. Like all dipoles the inverted V can be fed at the center with low impedance line, and elements for several bands can be connected in parallel at the feed point of the antenna.

The inverted V shown in the photographs and sketches is simple, easily managed (20 minutes to erect, 10 minutes to dismantle), and costs under $15 for all components. Not only is it a convenient antenna for portable operation, but it can serve as a fine permanent antenna for the home station as well. In fact, that's what has happened to the inverted V pictured; it's been in use in the back yard of VE3DPC ever since Field Day, 1967.

Assembly

As shown in Figure 1, the antenna support consists of three 10-foot sections of 1¼-inch TV mast. One end of each of the two lower sections is swaged, permitting all three sections to be force-fitted together. Hooks of music wire were installed in the top section, so that the element wire and feed line could be easily coiled for storage.

The center insulator (Figure 2) is made of maple; however, if this wood cannot be found, birch or fir can be used instead.

Figure 2 shows how the center insulator was drilled and Figure 3 shows how the feed line and antenna element were attached. Dimensions are obviously not critical, and the sketches are intended to show in a general way how the insulator was made, not to give precise detail. Just be careful where you drive the screws!

The inverted V as it looked this past winter in the author's backyard.

Upper portion of the 80-meter inverted V as it appeared at the Field Day site.

Softwood (pine) was picked for the end insulators (Figure 4) because it is adequate for the application and easier to work with than maple. Fragile ceramics should not be used in place of the wood insulators.

The 80-meter element (Figures 3 and 5) is made from small-gauge, Teflon-insulated, stranded Copperweld that was bought at a surplus outfit. This wire is worth looking for; it's tough, durable, light, and easy to handle. If operation on both 80 and 40 is desired, the builder can connect a 40 meter element to points A and B in Figure 3.[1]

The feed line can be 50-ohm coax, 72-ohm coax or 72-ohm Twin-Lead. I chose the latter because it is low cost, lightweight, very flexible, easy to store, and because it lends itself to such outrageous techniques as being tied in knots; however, since most transmitters have coax output, either a balun or an antenna coupler must be used between the feed line and the transmitter to maintain balance.[2] If elements for more than one band are to be used, the antenna coupler is preferred because it will reject harmonics of the transmitting frequency that might fall in the range of one or more of the paralleled antennas.

Raising and Lowering

Transporting the antenna to and from the Field Day site is easy. Tape the three mast sections together, wrap them with cloth to protect the car's finish, and tie the works between one of the door handles and the rear bumper. If the antenna extends past the back of the car, don't forget to tie a red flag on the end of the bundle to alert other drivers.

Upon arrival at the Field Day site, unpack the antenna and join the TV masts together. The three-section support is light enough to be "walked up" by one man, if the wind is under 10 miles per hour; however, prior to the antenna raising, you must put stakes into the ground, determine the

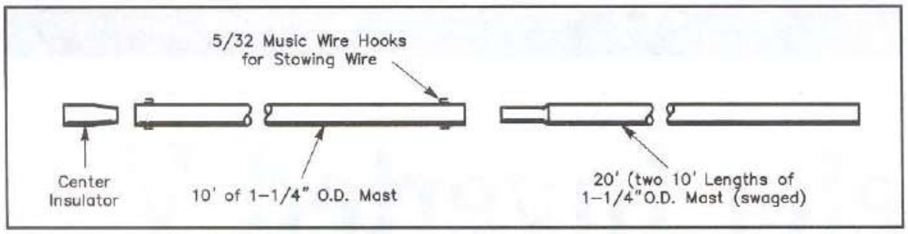
5/32 Music Wire Hooks for Stowing Wire
Center Insulator
10' of 1-1/4" O.D. Mast
20' (two 10' Lengths of 1-1/4" O.D. Mast (swaged)

Figure 1—The makings of the inverted-V center support.

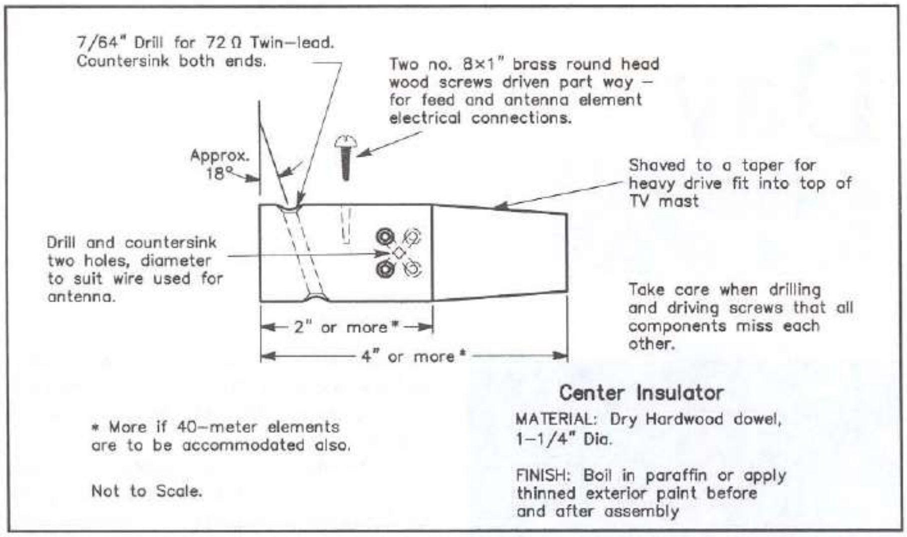
7/64" Drill for 72 Ω Twin-lead. Countersink both ends.
Two no. 8×1" brass round head wood screws driven part way – for feed and antenna element electrical connections.
Approx. 18°
Shaved to a taper for heavy drive fit into top of TV mast
Drill and countersink two holes, diameter to suit wire used for antenna.
Take care when drilling and driving screws that all components miss each other.
2" or more*
4" or more*
Center Insulator
MATERIAL: Dry Hardwood dowel, 1-1/4" Dia.
FINISH: Boil in paraffin or apply thinned exterior paint before and after assembly
* More if 40-meter elements are to be accommodated also.
Not to Scale.

Figure 2—An X-ray view of the center insulator.

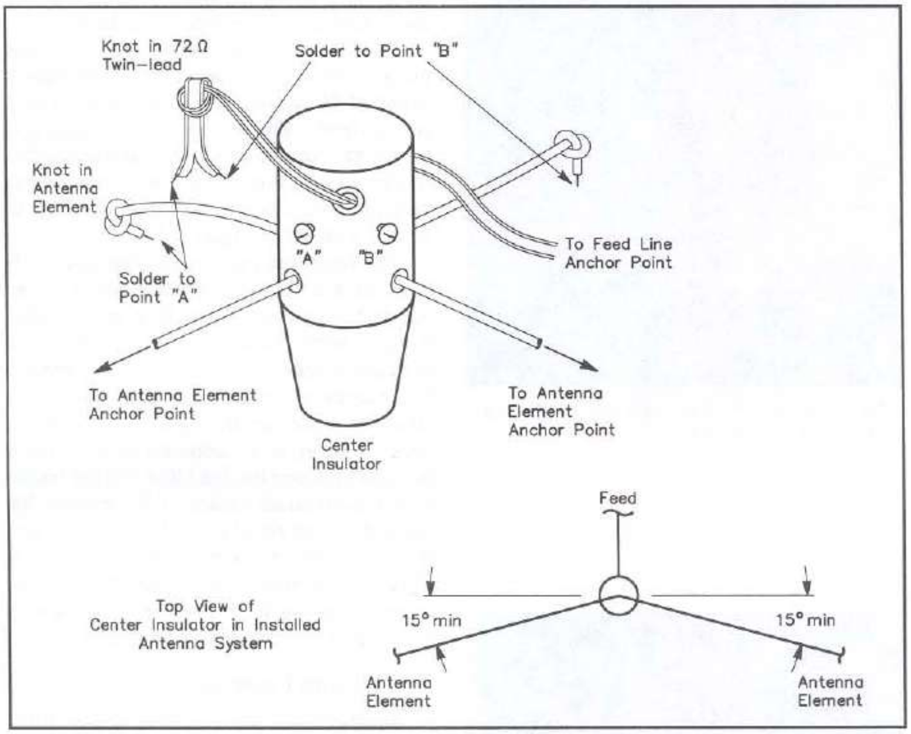
Knot in 72 Ω Twin-lead
Solder to Point "B"
Knot in Antenna Element
"A"
"B"
To Feed Line Anchor Point
Solder to Point "A"
To Antenna Element Anchor Point
To Antenna Element Anchor Point
Center Insulator
Feed
Top View of Center Insulator in Installed Antenna System
15° min
15° min
Antenna Element
Antenna Element

Figure 3—Details of how the feed line and element wires should be attached to the center insulator, and a top view of the installed insulator.

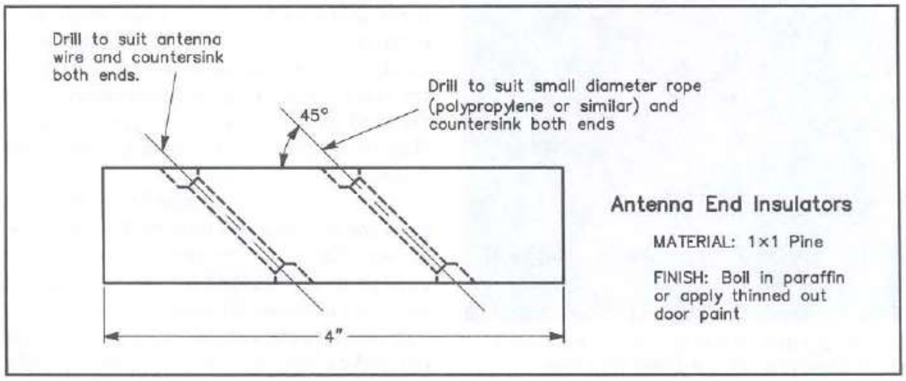
Drill to suit antenna wire and countersink both ends.
45°
Drill to suit small diameter rope (polypropylene or similar) and countersink both ends
Antenna End Insulators
MATERIAL: 1×1 Pine
FINISH: Boil in paraffin or apply thinned out door paint
4"

Figure 4—A sketch of one of the wooden end insulators. Two insulators are required for each dipole used.

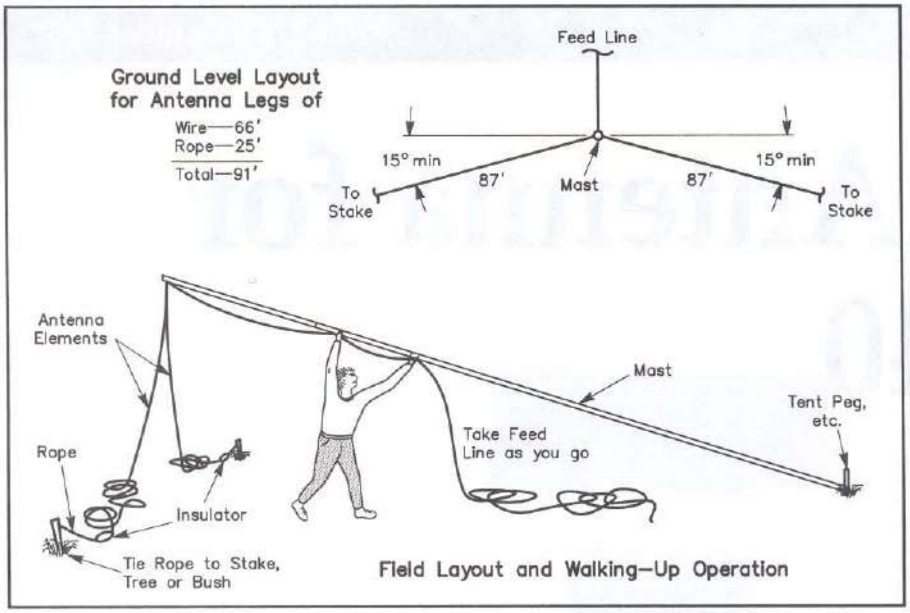

Figure 5—By using the dimensions and layout shown in the upper right corner of the sketch—so that ground stakes could be installed and proper length guys attached between them and the mast—the author was able to raise his 80-meter inverted V without any help. For other element lengths than shown, one man can easily put up the antenna by using simple geometry to determine the amount of rope required and the locations of the stakes.

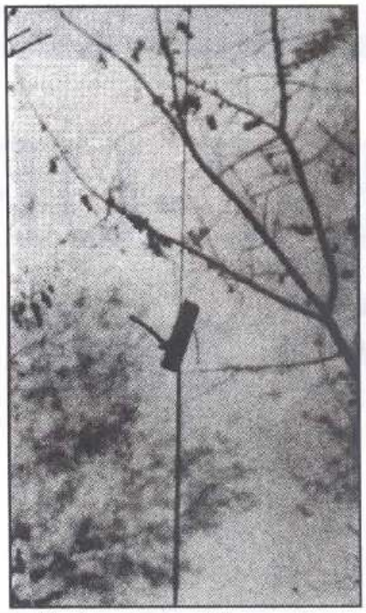

A sideview of one of the wooden end insulators.

exact length the guys will be when the antenna is in its operating position, and attach the guys to the stakes (Figure 5).

If help is available, it won't be necessary to install stakes or measure guys prior to putting up the antenna. Have the helpers pay out each half of the antenna element and keep slack out of the wire as the mast goes up. Station the men about 45 degrees off the line of the mast as it lies on the ground, and then, once the mast is up, have the helpers walk around to the final positions shown in Figure 5. The last few degrees of "walking up" is where your control is poorest.

With the rear guy (feed line) in your hand, move quickly and smoothly past top dead center and promptly and firmly apply back tension on the antenna element. Once the feed line is snubbed, you can easily correct the mast for lean by lifting the base and putting it where it suits you. If you are fortunate enough to have three helpers, there is no problem at all. Give the third man the feed line, station him as far away as possible from the mast and antenna element, and have him take up the slack as the mast goes up.

Base support can be little or nothing. The mast, which is now permanently erected in my back yard, rests directly on the grass. It was moved a couple of inches during each mowing operation, and no apparent harm came to the lawn.

Lowering the antenna is brutal but quick. Release the back guy (feed line) and let the whole system fall. If the ground is reasonably flat, this will be OK. But if the ground is very rough, the pipe may take a set as it strikes the ground. Under such conditions, it would be better to use helpers to reverse the erecting procedure.

The antenna has been left up for a period of several months, during which there have been many autumn storms and a severe ice storm. It has shown no signs of distress, and it has performed normally when wet or loaded with ice. The small size of the inverted V's parts has led to little wind and ice loading and is probably the main reason for the antenna's durability.

[1] Although dimensions for an 80-meter inverted V are given in Figure 5, these aren't necessarily the optimum dimensions for every installation: there are too many variables involved. If the antenna is not to be used with a transmatch and open-wire line, it is best to experimentally determine the length of the inverted V. This can be done by starting with an overall length equal to about 525 divided by the frequency in megahertz. The ends of the inverted V should then be trimmed until the SWR is minimum at the operating frequency. A simpler method is described in footnote 2.

[2] An easier arrangement, especially for multiband operation, can be achieved by using a transmatch and either 300-ohm Twin-Lead or 450-ohm Ladder-Line. Only one dipole—about 100 feet long (length not critical)—is all that is necessary to cover all the amateur bands between 3.5 and 30 MHz.

By John C. Allred, W5LST **From *QST*, October 1961**

E-Z-UP Antenna for 75 and 40

Full-size dipoles for the 75- and 40-meter bands occupy more space than is conveniently available on the 75 by 113-foot lot at W5LST. The increasing popularity of the "drooping," or inverted-vee, dipole antenna among amateurs led us to investigate it for our somewhat crowded conditions. Based on the electrical design of Glanzer,[1] this system has performed meritoriously at W5LST. Requiring only one support, it was surprisingly easy to erect, gives a satisfactory s.w.r. over the phone bands and, importantly, the cost was less than thirty dollars complete. A plan-view sketch is shown in Figure 1.

The Mast

The mast is a telescoping Channel-master, capable of 50-foot height, but extended only to 35 feet. Extending the upper sections to less than their full lengths gives rigidity to the mast, and has apparently eliminated the need for guys on each section, as recommended by the manufacturer. To date this mast has withstood gusts of 50 miles per hour without a shudder; it is yet to be tested in a real gale, however.

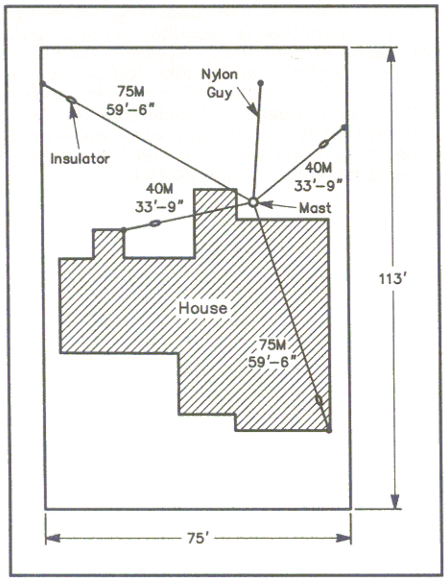

Figure 1—W5LST's layout for effective 40- and 80-meter antennas on a small lot. The two "drooping" dipoles are fed in parallel with a single coax line. Nylon-line extensions are used to reach convenient anchorages.

As shown in Figure 2, a 9-foot length of 3-inch pipe is cast in concrete with 6 feet of its length extending above ground. Three pairs of clamps, such as those used on chain-link fence, secure the mast to the upright standard. During erection, these clamps are loosened and all three lie at the base of the standard, so that the mast need only be lifted about 6 inches to be put in place. When the mast is in place, the clamps are raised and tightened.

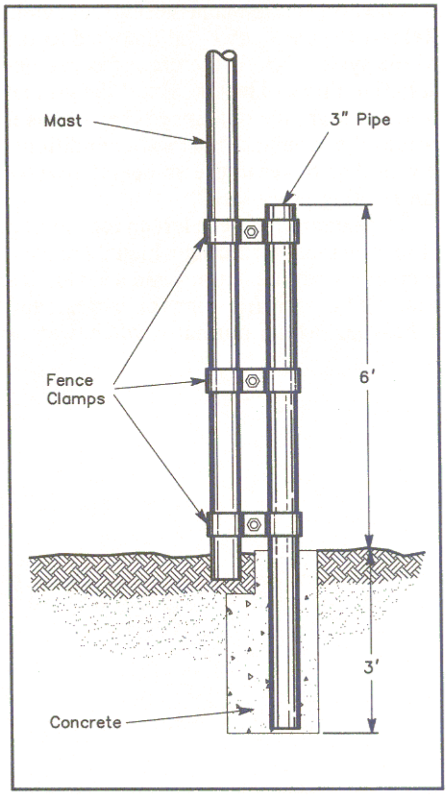

Figure 2—The mounting for the antenna mast. The 9-foot pipe is guyed temporarily while the concrete is poured.

Rigging

Except for the antenna conductors, all rigging is of nylon line of 500-pound test. A halyard is reeved through a pulley of suitable size which is wired securely to the top of the mast. The two ends of the halyard are made fast to a harness snap which, in turn, supports the center of the antennas. Provision of this halyard has proved to be a great convenience in permitting inspection of the antenna connections and the adjustment of tension in the wires without the necessity for lowering the mast.

Nylon line has a tendency to ravel at its ends but this problem is easily solved. Most fastenings were made with two halfhitches, followed by sewing the end of the line to itself with thread, as shown in Figure 3A, and doping with one of the quick-drying model-airplane cements.

Some weeks after the initial installation, it became apparent that some additional stabilization of the mast against occasional strong northerly winds would be desirable. Accordingly a nylon line was run from the harness snap at the top of the mast to a convenient anchor in the back yard, which happened to be the top of the children's swing set. Experience seems to show that the antenna wires, together with the additional nylon line, stabilize the mast against aerodynamically-excited vibration, without any appreciable strain on the antennas.

The Antennas

As shown in Figure 3, the two antennas are connected in parallel at the top of the mast. The lower ends are connected to convenient tie points so that the two legs of a given antenna are more or less in a straight line. To our great surprise, very little effect is produced by moving the ends of the an-

---

[1] Glanzer, "The Inverted V-Shaped Dipole," *QST*, August, 1960.

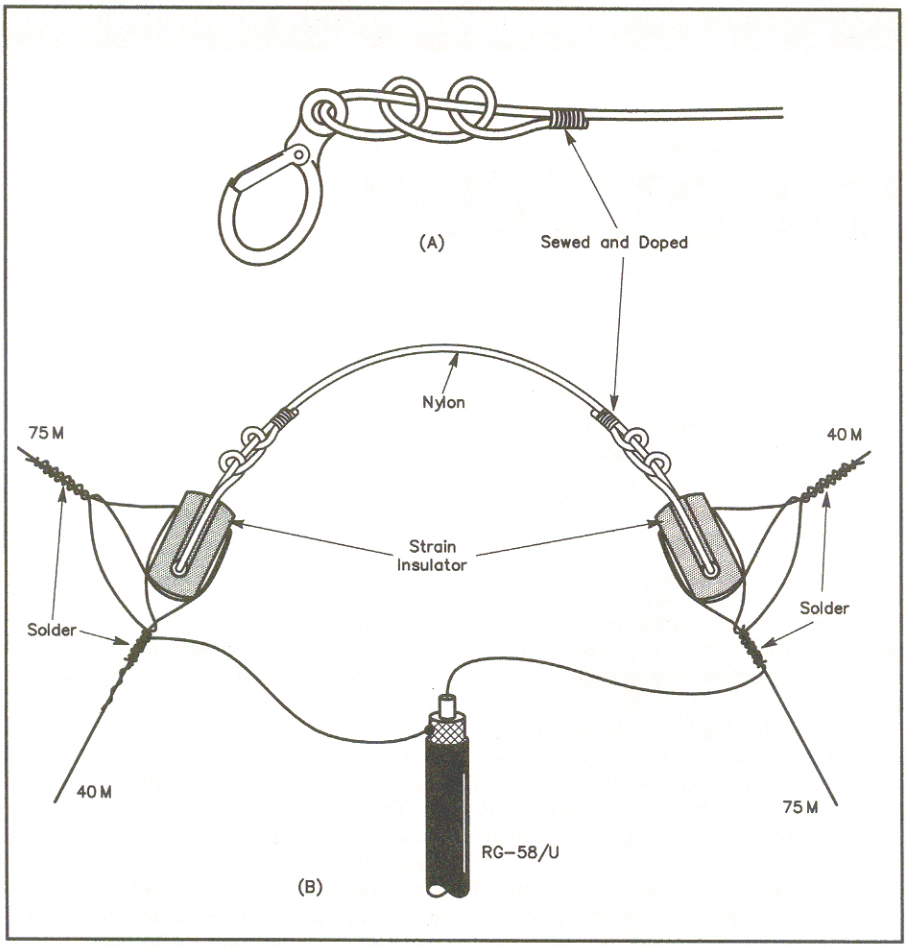

Figure 3—The halyard and feeder arrangement. The coaxial cable is sealed securely with polyethylene tape after the connections are made. The harness snap of A at the end of the hoisting halyard engages the bridle between the two insulators in B.

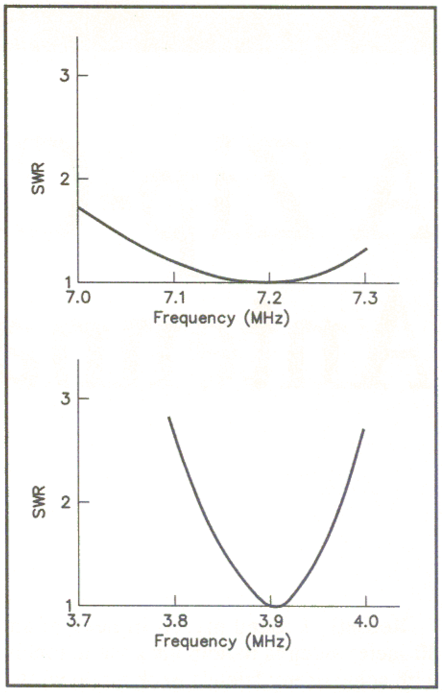

Figure 4—SWR curves as indicated by a "Monimatch"-type SWR indicator.

tennas either horizontally or vertically. There is apparently negligible electrical interaction between them as indicated by the s.w.r. bridge.

Pruning of the antennas is, as always, desirable. In our case, the optimum lengths of each leg turned out to be 33 feet 9 inches for 40 meters and 59 feet 6 inches for 75 meters. Figure 4 shows the performance of the antennas on 40 and 75 as measured by a Heath s.w.r. bridge. Although the antennas tune sharply, they are usable over the entire phone band in each case, and it would be difficult to imagine a better performer on the lower-frequency bands than this simple antenna system.

A Zip-Cord Special Antenna

Recently I found myself in need of an 80-meter antenna to help me keep in touch with some of my friends back home while operating portable in Arizona. The antenna and feed line were not readily available, but I remembered that one of my friends had recommended zipcord as a substitute feed line for a dipole, provided low power was used. Going one step further, I felt the entire antenna could be made of this low-priced material. I purchased 80 feet of zip cord at a local hardware store (at 4 cents a foot) and peeled it apart for 60 feet. This gave me the required 120 feet for a dipole, plus a 20-foot feed line which was taped to prevent further splitting. The feeder was fastened to a 2×4 which was used as a center support. The ends were extended as far as possible, then tied with heavy cord to a fence that was in the trailer park where I was staying. This made a slight inverted-V antenna.

To reach the desired frequency of 3935 kHz it was necessary to shorten the antenna somewhat. I did this by folding three feet on each end of the dipole back upon itself. This enabled me to keep skeds with many "7s" at home.

Having met with success on 80-meters I immediately began to eye the 40-meter band for possible use. I then measured out 32 feet from the center and cut the wire there, then tied a knot in the two pieces, leaving a couple of inches of bare wire dangling. Now for forty meters, I merely disconnect the wires and reconnect them for 80 meters.

As the *ARRL Antenna Book* states, a 40-meter dipole will often present a good match at 15-meters, and this one is no exception; I have worked many stations across the country on 15 meters with low power while using this simple antenna.

The *Handbook* or *Antenna Book* has not recommended this type of feed line in many years, though using a twisted pair at one time was usual practice in feeding a doublet. But, when the need is there for a quick and easy antenna, I don't think you can find an easier one as long as you remember to use low-power transmitters (100 watts or less) and trim the wire for a low value of SWR. — *Frank S. Wise, W7QYC*

By Jerry Hall, K1TD

From *QST*, March 1979

Zip-Cord Antennas–Do They Work?

Basic Amateur Radio: Parallel power cord is readily available and is easy to work with. How efficient is it when used at radio frequencies? Well, that depends.

Zip cord is readily available at hardware and department stores, and it's not expensive. The nickname, zip cord, refers to that parallel-wire electrical cord with brown or white insulation used for lamps and many small appliances. The current price is about 6 cents a foot in hundred-foot rolls in most parts of the U.S. The conductors are usually no. 18 stranded copper wire, although larger sizes may also be found. Zip cord is light in weight and easy to work with.

For these reasons, zip cord is frequently used as both the transmission line and the radiator section for an emergency dipole antenna system. The radiator section is obtained simply by "unzipping" or pulling the two conductors apart for the length needed to establish resonance for the operating frequency band. The initial dipole length can be determined from the equation $\ell = 468/f$, where ℓ is the length in feet and f is the frequency in megahertz. (It would be necessary to unzip only half the length found from the formula, since each of the two wires becomes half of the dipole.) The insulation left on the wire will have some loading effect, so a bit of length trimming may be needed for exact resonance at the desired frequency.

For installation, many amateurs like to use the electrician's knot shown in Figure 1 at the dipole feed point to keep the transmission-line part of the system from unzipping itself under the tension of dipole suspension. This way, if zip cord of sufficient length for both the radiator and the feed line is obtained, a solder-free installation can be made right down to the input end of the line. Granny knots (or any other variety) can be used at the ends with cotton cord to suspend the system. You end up with a lightweight low-cost antenna system that can serve for portable or emergency use.

But just how efficient is a zip-cord antenna system? Since it is easy to locate the materials and simple to install, how about using such for a more permanent installation? Upon casual examination, zip cord looks about like high-quality 72-ohm balanced feed line. Does it work as well? Ask several amateurs these questions and you're likely to get answers ranging all the way from, "Yes, it's a very good antenna system!" to "Don't waste your time and money – it's not worth it!" Myths and hearsay seem to prevail, with little factual data. The intent of this article is to rectify that situation.

Zip Cord as a Transmission Line

In order to determine the electrical characteristics of zip cord as a radiofrequency transmission line, we purchased a 100-foot roll and subjected it to tests in the ARRL laboratory with an rf impedance bridge. Zip cord is properly called parallel power cord. The variety purchased was manufactured for GC Electronics, Rockford, IL, being 18-gauge, brown, plastic-insulated type SPT-1, GC cat. no. 14-118-2G42. Undoubtedly, minor variations in the electrical characteristics will occur among similar cords from different manufacturers, but the results presented here are probably typical.

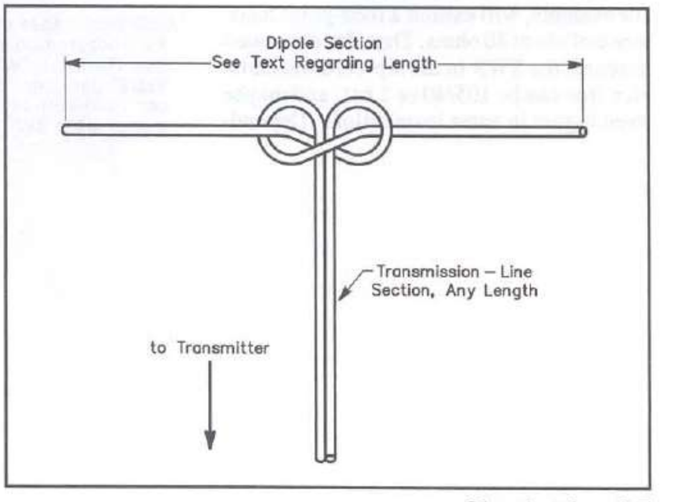
Dipole Section
See Text Regarding Length
Transmission – Line Section, Any Length
to Transmitter

Figure 1—This electrician's knot, often used inside lamp bases and appliances in lieu of a plastic grip, can also serve to prevent the transmission-line section of a zip-cord antenna from unzipping itself under the tension of dipole suspension. To tie the knot, first use the right-hand conductor to form a loop, passing the wire behind the unseparated zip cord and off to the left. Then pass the left-hand wire of the pair behind the wire extending off to the left, in front of the unseparated pair, and thread it through the loop already formed. Adjust the knot for symmetry while pulling on the two dipole wires.

As the first step, we checked the physical length of the wire. GC generously provided nearly 4 feet over the specified 100-ft length, and the extra length was promptly lopped off.

We wanted to avoid measurement errors that might arise from coiling the wire on the supply spool (inductance between turns) or laying it out on the vinyl-tiled cement-slab floor (capacitance to a large surface), so the second step was to suspend the wire in a long hallway, about a foot or so below the false ceiling. Cotton twine was used for the supporting material. Undaunted by snide comments from fellow staff members about cheating the telephone company out of installation work and facetious suggestions that we use two paper cups and a string for our intercom, we continued with preparation for the tests.

With a General Radio 1606A rf impedance-measuring bridge, we made measurements at the input end of the line on 10, 15 and 29 MHz, while the far end of the line was terminated first in a short and then in an open circuit. We did notice during measurements that our readings changed somewhat as people walked in the hallway, especially during the 29-MHz tests. After a seemingly endless period of shooing away would-be helpers who disrupted measurements by grabbing the line, hanging objects on it, and so forth, we finally arrived at a set of readings considered to be satisfactory for our purposes. From there, a bit of work with an electronic calculator and a Smith Chart told us what we wanted to know — the electrical characteristics at radio frequencies.

The results? Well, as the saying goes, I have some good news and some bad news.

Figure 2—Attenuation of zip cord in decibels per hundred feet when used as a transmission line at radio frequencies. Measurements were made only at the three frequencies where plot points are shown, but the curve has been extrapolated to cover all high-frequency amateur bands.

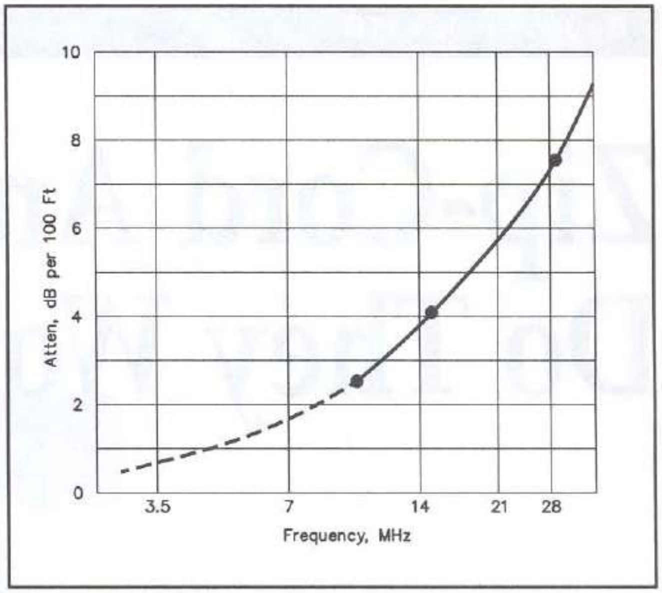

First, the Good News

If ever you need a balanced 105-ohm transmission line, then zip cord is the stuff for you! Its characteristic impedance was determined to be 107 ohms at 10 MHz, dropping in value to 105 ohms at 15 MHz and to a slightly lower value at 29 MHz. The nominal value is 105 ohms at hf. The velocity factor of the line was determined to be 69.5 percent.

This reported change in impedance with frequency may raise a few eyebrows, but it is a fact that most lines do not exhibit a constant characteristic impedance across a range of frequencies. However, this writer has not previously encountered any line which was not "flat" between 3 and about 50 MHz. Zip cord may be exceptional from this standpoint, especially considering that the insulating material was not chosen for its qualities at rf.

And Now the Bad News

Who needs a 105-ohm line, especially to feed a dipole? Most of us know that a dipole in free space exhibits a feed-point resistance of 73 ohms, and at heights above ground of less than 1/4 wavelength the resistance is even lower. An 80-meter dipole at 35 feet, for example, will exhibit a feed-point resistance of about 40 ohms. Thus, for a resonant antenna, the SWR in the zip-cord transmission line can be 105/40 or 2.6:1, and maybe even higher in some installations. Depending on the type of transmitter in use, the rig may not like working into the load presented by the zip-cord antenna system.

But the really bad news is still to come–line loss! Figure 2 is a plot of line attenuation in decibels per hundred feet of line versus frequency. Chart values are based on the assumption that the line is perfectly matched (sees a 105-ohm load as its terminating impedance). For lengths other than 100 feet, multiply the figure by the actual length in feet and then divide by 100.

In a feed-line, losses up to about one decibel or so can be tolerated, because at the receiver a 1-dB difference in signal strength is just barely detectable. But for losses above about 1-dB, beware. Remember that if the total losses are 3 dB, half of your power will be used up just to heat the transmission line.[1]

Based on this information, we can see that a hundred feet or so of zip-cord transmission line on 80 meters might be acceptable, as might 50 feet on 40 meters. But for longer lengths and higher frequencies, look out! The losses become appreciable.

What About Zip Cord Wire as the Radiator?

For years, amateurs have been using ordinary copper house wire as the radiator section of an antenna, erecting it without bothering to strip the plastic insulation. Other than the loading effects of the insulation mentioned earlier, no noticeable change in performance has been noted with the insulation present. And the insulation does offer a measure of protection against the weather. These same statements can be applied to single conductors of zip cord.

The situation in a radiating wire covered with insulation is not quite the same as in two parallel conductors, where there may be a leaky dielectric path between the two conductors. In the parallel line, it is this current leakage which contributes to line losses. The current flowing through the insulation on a single radiating wire is quite small by comparison, and so as a radiator the efficiency is high.

Now back to the original question: How efficient *is* a zip-cord antenna system? Well, that *does* depend, on the length of the wire used for the feed-line section and on the frequency. In a pinch on 160, 80, 40 and perhaps 20 meters, communications can certainly be established with this kind of antenna. For higher frequencies, especially with long line lengths for the feeder, you're on your own.

[1] Additional losses over those charted in Figure 2 will occur when standing waves are present. See Gibilisco, "What Does Your SWR Cost You?" January 1979 *QST*. Trouble is, you can't use a 50- or 75-ohm SWR instrument to measure the SWR in zip-cord line accucately.

From *QST*, April 1973 (Hints & Kinks)

Piggyback Antenna For The Ten-Meter Band

Most amateurs from Novices on up use dipole antennas for the eighty- and forty-meter bands. In many cases, these two antennas were all the Novice needed since the fifteen-meter band could also be covered with the forty-meter dipole.

However, with the advent of the new ten-meter subband for Novices, a simple and inexpensive method of getting the needed coverage is to use a "piggyback" antenna. It can be easily added to an existing dipole, eliminating the need for additional supports or feed lines. Also, it will have no effect on either an eighty- or forty-meter dipole.

Of course, performance will not be as good as that obtained with a beam, but it should not be sold short either. A KH6 was worked during a recent contest with a power output of 1.8 watts. — *Robert M. May, II, WA4DEG*

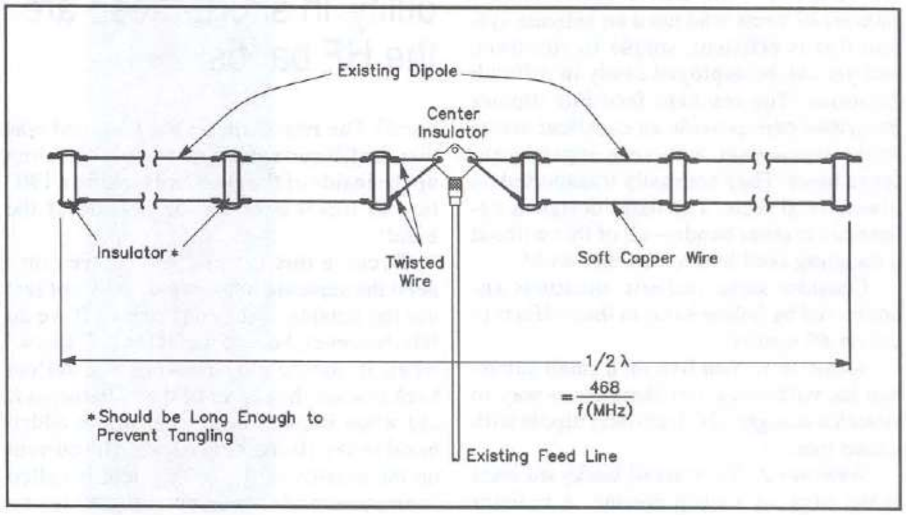
Existing Dipole
Center Insulator
Insulator *
Twisted Wire
Soft Copper Wire
$1/2\,\lambda = \frac{468}{f(\text{MHz})}$
* Should be Long Enough to Prevent Tangling
Existing Feed Line

By James E. Taylor, W2OZH

From *QST*, August 1991

RFD-1 and RFD-2: Resonant Feed-Line Dipoles

This unique design offers simplicty and general utility. In short, these are "reel" great antennas for the HF bands.

During several decades of operation on the 80-meter band, I have encountered a number of hams who need an antenna system that is efficient, simple to construct, and yet can be deployed easily in difficult locations. The resonant feed-line dipoles described here provide an excellent match to the transceiver without a separate antenna tuner. They are easily transported on plastic cord reels. The basic design is extensible to other bands—all of this without a dangling feed line to contend with!

Consider some realistic situations encountered by fellow hams in their efforts to get on 80 meters:

Scenario 1: You live on a small suburban lot with trees, but there is no way to stretch a straight 120-foot-long dipole with a feed line.

Scenario 2: Your small backyard ends at the edge of a steep decline. A radiator wire could be snaked through the trees, but running a separate feed line is out of the question.

Scenario 3: You live in a high-rise building and could secretly stretch a wire upward or downward, but there is no way to install a feed line.

Scenario 4: You're going on vacation to the mountains and you know that your cottage is nestled among hundred-foot-high pines You can take along a bow and arrow and cord, but then what would you do?

In response to these problems, I began to search for a universal solution under the following ground rules. (1) No separate antenna tuner would be required, (2) the antenna could be deployed with no more difficulty than stringing up a length of coax and (3) the antenna could be easily stored and unwound from a cord reel without a tangle of cable and wire.

Several years ago I used a 10-meter vertical linear coaxial sleeve antenna, as shown in Figure 1. A vertical dipole is constructed from a quarter-wave whip and a quarter wavelength of shielding braid. Its feed line passes through the braid, yielding a simplified geometry. Although this concept could be adapted to 80 meters, who wants to deal with 60 feet of shielding braid? The important lesson I learned was that the RF current has no trouble traveling up the inside of the coax and making a 180° turn to travel back on the outside of the braid!

Because this is true, perhaps we don't need the separate outer braid. Why not just use the outside of the coax itself? If we do this, however, how do we let the RF "know" when it should stop flowing and reflect back toward the center of the radiator, as it did when it came to the end of the added braid in the 10-meter vertical? The current on the outside of the coax shield is called "common-mode" current—there is no counteracting equal and opposite current as there is inside the coax. One design approach utilizes the primary function of a balun transformer: to place in unbalanced reactance in the path of this common-mode current without affecting the desired balanced transmission line currents. The development of this concept is discussed in the Appendix.

The common-mode current on the coax shield is transformed by a quarter-wavelength stub to maximum near the transceiver. It was not surprising to note experimentally that a coil of a few turns placed at the fed end of the dipole decreases the resonant frequency of the system substantially (because of the added inductive reactance). However, an unexpected dividend of this is that, at the resonant frequency, an almost perfect impedance match to the 50-ohm source is realized! Now we have the design for our simplified antenna system! We can increase the resonant frequency by moving the coil along the cable away from the current maximum while retaining the perfect match. And at the same time we can decrease the common-mode current on this part of the line, because of the coil's inductive reactance. For the values chosen, the coil is near self-resonance from the distributed capacitance of the coil windings. The equivalent parallel-resonant circuit serves to increase the reactance at this point in the antenna, which assures a reflection when the RF reaches this virtual end of the dipole. To my knowledge, this configuration and method of resonating is unique and novel, and I refer to it as a T choke.

Constructon, Installation and Adjustment

The simple arrangement of the resonant coaxial linear dipole is shown in Figure 2. The dimensions and test results are for a nominal

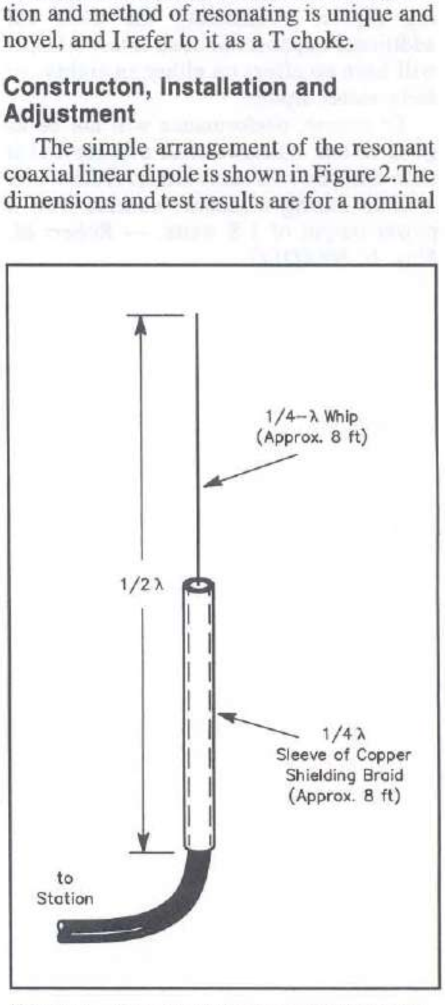

Figure 1—A vertical linear coaxial sleeve antenna for 10 meters.

frequency of 3.95 MHz. To dramatize the simplicity of the antenna, I list all required parts in Table 1.

The coaxial cable connector is assembled at the input end of the coax for connecting to the transceiver. The center conductor of the far end of the cable is connected to the antenna wire to form a hook and eye, and is securely soldered. This junction is potted in the center of a short length of PVC pipe to form a robust center insulator assembly. At the outset, 13 turns of cable can be wound on the cord reel at a point approximately 59 feet plus 25 inches from the center of the dipole. The 13 turns can then be taped in place on the reel with duct tape or equivalent. The remaining coax and wire can now be wound on the reel for ease of transporting. To avoid kinks when unwinding, secure the end of the wire and rotate the reel, keeping the wire and coax taut.

Ground losses are a great enemy of an HF antenna, especially at the lower frequencies, so place the dipole for 80 meters as far above the ground as possible! After installation, check the resonant frequency with a noise bridge or an SWR meter. If these construction details are followed, the resonant frequency should be approximately 3.95 MHz. If the resonance indication is indefinite or if the resistance is not close to 50 ohms, adjust the self-resonance of the coil by moving the turns slightly on the reel. This alters the inter-turn capacitance, permitting adjustment of the reflection of RF at the end of the dipole. During this adjustment, remember that the greatest effect is between the input and output turns, where the voltage difference is greatest. If you want to lower the resonant frequency of the antenna, remove the tape and rotate the reel to increase the 25-inch distance, thereby increasing the length of the dipole. This is done in a manner to retain the 13-turn coil—you are winding and unwinding equal lengths of cable. When the desired resonant frequency is attained, you are ready to operate! This is a broadband antenna, and not much adjustment is required.

The 13 turns is mentioned only as a nominal coil size. I have used both 11-turn and 13-turn coils. Self-resonance can be adjusted as mentioned earlier. As the number of turns is increased, the initial 25-inch distance will be decreased. For example, in my test installation a change from 11 turns to 13 turns altered this distance required for resonance from 68 inches to 25 inches. A greater impedance at this point serves to improve the isolation of the dipole from ground.

The RFD-2 for Two-Band Operation

The preceding construction information is for an antenna dedicated to a single amateur band, the RFD-l. With a slightly different configuration, referred to as the RFD-2, the RFD-1 design can be extended to cover the 40-meter band without severing the coaxial cable. Conversion from 80 to 40 meters and back requires only a few minutes. The RFD-2 design allows you to change operation to 40 meters for a few days without permanently altering the 80-meter lengths. In other words, the total dimensions of the 80-meter coaxial feed-line dipole are retained but operation is adapted to 40 meters. This is done simply by winding coaxial cable on reels. This is readily achieved by altering the winding of the coax on the original reel and then adding a second coil near the far end of the dipole. Once the values have been established, this band change is accomplished simply.

Table 1

RFD-1 Parts List

1–118-foot length RG-8X (minifoam) coaxial cable

1–59-foot length no. 12 stranded or Copperweld wire

1–PL-259 male coaxial cable connector

1–10-inch-diameter cord reel (Doscocil model no. 32500 or equiv.)

1–3-inch length of 1/2-inch OD PVC pipe (for center insulator)

1–Epoxy potting compound (sufficient to fill pipe)

The RFD-2, as it has evolved at W2OZH, is shown in Figure 3. Total dimensions are given in Figure 3A. The length of the feed line has been increased from that of the RFD-1 to 143 feet to make allowance for the cable used in the 13-turn coil. This assures an isolating stub which is approximately 1/4 λ on 80 meters, and at the same time permits a stub length that is close to 1/4 λ on 40 meters, thereby achieving the desired high impedance for isolation on both bands. Also, the no. 12 terminating antenna wire has been replaced by a length of coaxial cable—the outer jacket provides insulation so the coax can now be coiled to provide the desired dipole length on 40 meters.

As pointed out earlier, a self-resonant coil is used to assure a high coefficient of reflection at the fed end of the 80-meter dipole. For other bands it is necessary to calculate the approximate number of turns required to approach self-resonance. A suitable winding for 80 meters was determined to be 13 turns on a 6-inch diameter reel, so the numbers for other bands can be obtained by a simple scaling calculation. Details are given in the Appendix. For guidance in adapting the design for the other HF bands, the approximate number of turns for the terminating coils is shown in Table 2.

For wavelengths shorter than 40 meters it may be desirable to use a radiating section longer then 1/2 λ. The length either side of center can be any odd number of quarter wavelengths. For example, on 20 meters the radiator could be 3/2 λ (three

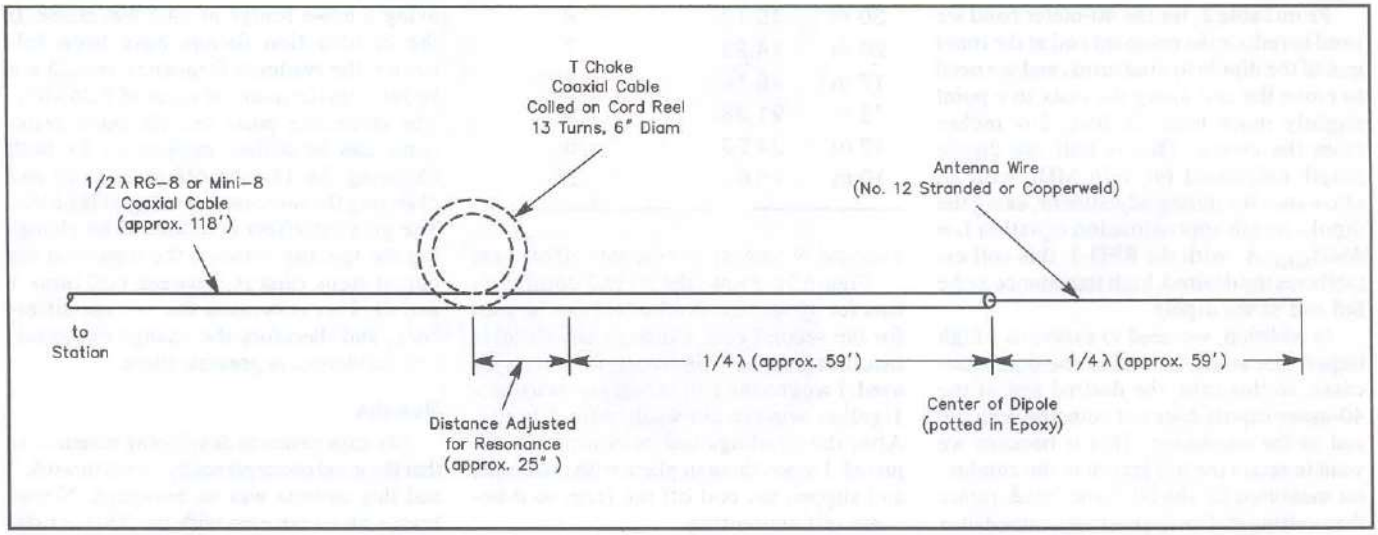

Figure 2—The RFD-1 (resonant feed-line dipole) antenna for 80 meters.

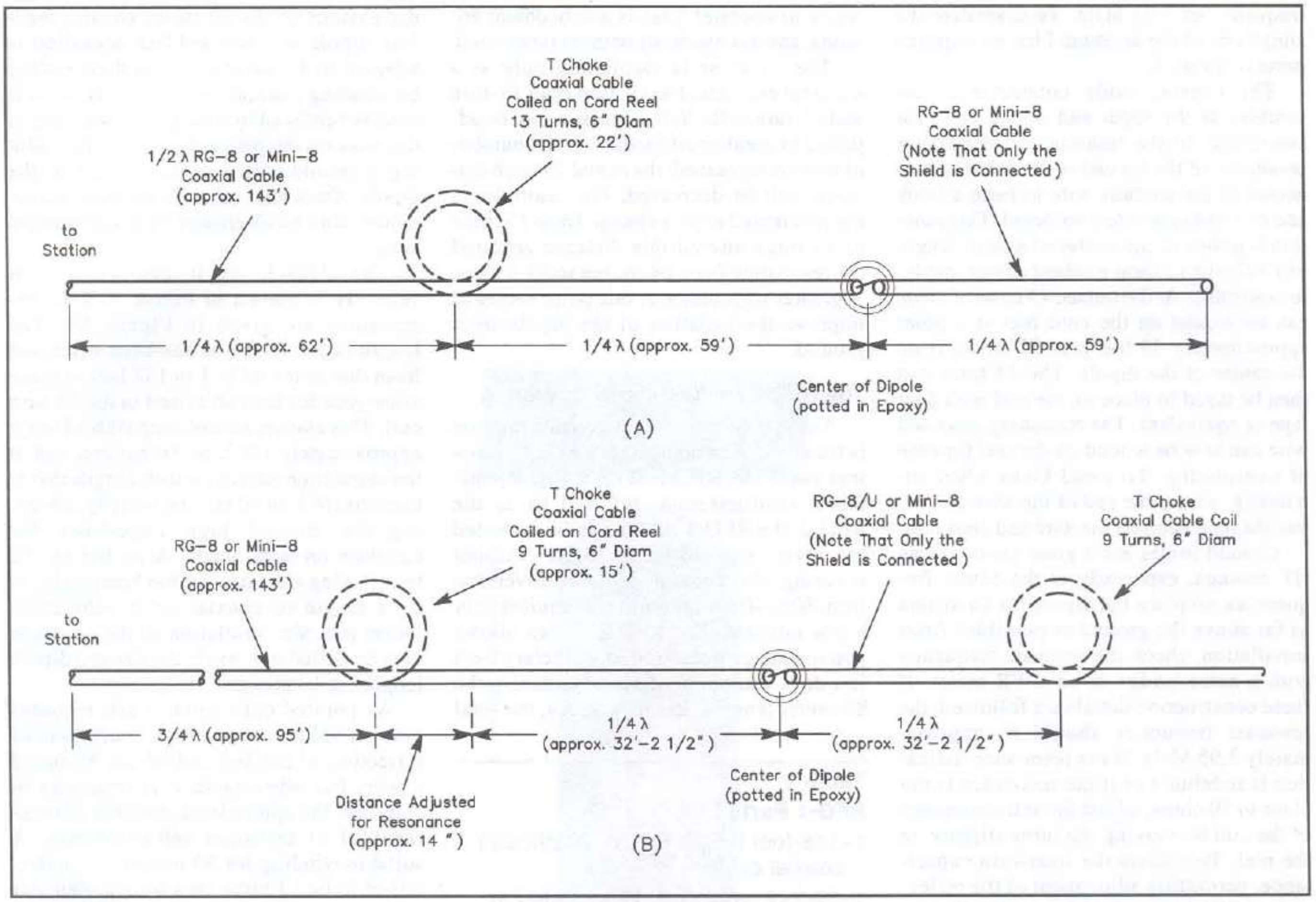
T Choke
Coaxial Cable
Coiled on Cord Reel
13 Turns, 6" Diam
(approx. 22')
1/2 λ RG-8 or Mini-8
Coaxial Cable
(approx. 143')
to
Station
RG-8 or Mini-8
Coaxial Cable
(Note That Only the
Shield is Connected)
1/4 λ (approx. 62')
1/4 λ (approx. 59')
1/4 λ (approx. 59')
Center of Dipole
(potted in Epoxy)
(A)
T Choke
Coaxial Cable
Coiled on Cord Reel
9 Turns, 6" Diam
(approx. 15')
RG-8 or Mini-8
Coaxial Cable
(approx. 143')
to
Station
RG-8/U or Mini-8
Coaxial Cable
(Note That Only the
Shield is Connected)
T Choke
Coaxial Cable Coil
9 Turns, 6" Diam
3/4 λ (approx. 95')
1/4 λ
(approx. 32'-2 1/2")
1/4 λ
(approx. 32'-2 1/2")
Distance Adjusted
for Resonance
(approx. 14")
Center of Dipole
(potted in Epoxy)
(B)

Figure 3—The RFD-2 antenna can be configured for either of two bands. Shown at A is the setup for 80 meters, and at B for 40 meters.

quarter waves either side of center). Depending upon the antenna orientation, the additional radiation lobes may be advantageous. Also, the unused conductor lengths at the ends are less apt to cause trouble because of parasitic excitation from the main radiator.

Arrangement and Construction

FromTable 2, for the 40-meter band we need to reduce the resonant coil at the input end of the dipole to nine turns, and we need to move the coil along the coax to a point slightly more than 32 feet, 2¾ inches from the center. This is half the dipole length calculated for 7.26 MHz with an allowance for tuning adjustment, using the dipole-length approximation equation $L = 468/f_{MHz}$. As with the RFD-1, this coil establishes the desired, high impedance at the fed end of the dipole.

In addition, we need to establish a high impedance at the far end of the dipole because, in this case, the desired end of the 40-meter dipole does not coincide with the end of the conductor. This is because we want to retain the full length of the conductor measured for the 80-meter band, rather than cutting it. The desired high impedance reflecting termination is achieved by use of a second, 9-turn, self-resonant coil of coax.

Table 2
Scaling of RFD-1 to Other Bands

Band	*Frequency*	*Turns*
160 m	1.90	19
80 m	3.95	13
40 m	7.26	9
30 m	10.12	8
20 m	14.29	7
17 m	18.14	6
15 m	21.38	6
12 m	24.96	5
10 m	28.65	5

Figure 3B shows the RFD-2 configuration for 40 meters. A cord reel can be used for the second coil, although any suitable insulating 6-inch diameter form can be used. I wound the coil in two layers using a 1-gallon windshield-washer-fluid bottle. After the windings had been properly adjusted, I taped them in place with duct tape, and slipped the coil off the form so it became self-supporting.

As with the RFD-1, it is desirable to place the antenna as high above ground as possible to reduce ground losses and to belp achieve low-angle radiation, if that is desired. The dipole does not have to be installed in a straight line, and if you can gain appreciably greater height by bending it, this may be a desirable compromise.

After the antenna with its two coils is erected, measure the resonant frequency using a noise bridge or an SWR meter. If the construction details have been followed, the resonant frequency should not be far from the nominal value of 7.26 MHz. The resonance point and the input resistance can be shifted moderately by both changing the 14-inch offset distance and changing the inter-turn spacing of the coils. The greatest effect is produced by changing the spacing between the input and the output turns (that is, between coil turns 1 and 9). This is because the voltage difference, and therefore the change in capacitive influence, is greatest there.

Results

My experience in developing antennas is that the good concepts really "want to work," and this antenna was no exception. Noise-bridge measurements with the RFD-1 indicated an input resistance at resonance of

49 ohms, and from a practical standpoint the SWR is 1:1. An H-field antenna probe was used to evaluate the power radiated at the center of the dipole compared with that at the current loop near the feed-point. The current ratio was 5.5 to 1, which corresponds to a power ratio of 30 to 1, or 15 decibels. This indicates that the coil is very effective in attenuating the common-mode current flowing back toward the feed point.

For the RFD-2 with the 40-meter dimensions shown in Figure 3B, the initial resonance point was within the 40-meter phone band and the SWR was essentially 1:1. Slight adjustment brought the resonant frequency up to the desired value, and the input resistance was very close to 50 ohms. A salient characteristic of the RFD antennas is the ease with which an impedance match is attained. It appears always to be easy to get a reflected power indication of zero. This is probably because the common-mode current on the shield of the coax is indistinguishable from the desired radiating current—in other words, the common mode is used rather than avoided.

Appendix

When the separate outer braid of the antenna shown in Figure 1 is removed, there must be some way to let the RF "know" when it should stop flowing on the shield of the coax and reflect back toward the center of the radiator. A balun box that I made up for another purpose is one approach: 30 turns of bifilar winding on an Amidon T-200-2 iron-powder toroidal core. The turns formula for such a balun transformer is

$$T=\sqrt{\frac{\text{desired}\,L(\mu H)}{A_L}}$$

where

T = no. of turns

A_L = inductance index (microhenries per 100 turns)

From this, the inductive reactance for unbalanced current is

$$X_L=2\pi fL=\frac{2\pi fT^2A_L}{10^4}$$

At a frequency of 4 MHz, this 30-turn coil has an unbalanced reactance of only 270 ohms. We would need about ten of these in series to support the RF field at the end of a dipole antenna!

An alternative method involves placing a toroidal isolation transformer in the coaxial line at the fed end of the linear dipole. I actually wound such a transformer on two stacked T-200-2 cores, to provide a 1-kW power capability. This configuration worked, after a fashion, but the impedance match was less than desired, probably because of excessive capacitance between secondary and primary windings. However, before I took steps to control this, a simpler approach came to mind.

We need to isolate the transceiver from the fed end of the dipole, so why not cut the coaxial feed line to be a quarter wavelength long? (This length is measured in free-space, not in the line.) This serves to tranform the high impedance of the fed end of the dipole to the low impedance at the grounded transceiver. Any unused portion of this approximate 60-foot length of coax can be wound in a coil and used for further isolation and, as it turns out, for tuning the system to resonance.

Scaling for a Second Band

For coverage of a second band, we would like to have the same inductive reactance that we had for the 80-meter coil.

From handbooks, the inductance of a coil of assumed dimensions is

$$L = AN^2$$

where

A = a constant determined by the coil geometry

N = number of turns

Thus, we have for the reactance

$$X_L = 2\pi fL = 2\pi fAN^2$$

For equal reactances we can calculate N_2, the new number of turns, at frequency f_2, from N_1, the known number of turns at frequency f_1, by the equation

$$N_2=N_1\sqrt{\frac{f_1}{f_2}}$$

By Jerry Hall, K1PLP **From *QST*, September 1974**

Off-Center-Loaded Dipole Antennas

In these times when much of our amateur population lives in urban areas, the subject of shortened antennas for the lower frequency amateur bands is a very popular one. Physically short ground-mounted vertical antennas with lumped-constant loading to make them resonant can be quite efficient radiators, if a good radial system has been installed. This has certainly been evidenced in Sevick's series of recent *QST* articles.[1] To many amateurs, however, the "hitch" in constructing such a system is the installation of a *good* radial system. It must be admitted that for the "top" amateur bands, 160 and 80/75 meters, an efficient system of buried radials requires a sizable amount of real estate, even for a physically short radiator. On the average city-size lot, 50 or 75 by 120 to 150 feet, it's almost impossible to install a highly efficient radial system for 80/75 meters, much less for 160 meters, when structures like a house and perhaps a separate garage exist. Or to some amateurs, just the thought of burying hundreds or maybe thousands of feet of wire is enough to turn off any enthusiasm for the project. What's the alternative? A dipole type of antenna with lumped constant loading. At modest heights, 30 or 40 feet, such an antenna will prove to be quite satisfactory if it is physically longer than about 0.2 wavelength. Shorter lengths may also be used, at reduced efficiency. Such an antenna can be fed directly with 50-Ω coaxial line, and it can be operated with no earth ground. (Of course the chassis of the transmitter and/or receiver should be grounded adequately for protection against shock hazard.)

Nearly all of us are familiar with the concept behind the use of inductive loading. A vertical antenna which is shorter than a quarter wave (or a dipole antenna which is shorter than a half wave) will exhibit capacitive reactance at its base (or center) feed point. To cancel such capacitive reactance, a coil having the proper inductive reactance may be connected in series with the base feed point of the vertical. The same result will be obtained through the use of two such coils for a dipole, one coil connected in series with each half. It is not necessary for the inductor to be installed at the feed point, however. In fact, greater radiating efficiency results through improved current distribution if the inductor is located along the radiator some distance away from the low-impedance feed point, *viz*, in the manner of a center-loaded mobile whip antenna. Figure 1 shows this concept extended to a dipole element, with off-center loading. The inductors resonate the antenna to the operating frequency, but do little actual radiating themselves. (This is in contrast to helically wound or continuously loaded elements, where a long thin inductor is the radiator as well as the loading element.)

In the antenna represented by Figure 1, there are many variable factors to be considered when a practical antenna for a given frequency is being constructed. Of primary consideration from an efficiency standpoint is the overall length, shown as dimension A. Another consideration for efficiency is the distance of the coils from center, dimension B. The longer the overall length (A), up to a half wave, and the farther the loading coils are placed from the center (B), the greater is the efficiency of the antenna. However, the greater is distance B (for a fixed overall antenna size), the larger the inductors must be to maintain resonance. Theoretically, if the coils were placed at the outer ends of the dipole, they should be infinite in value to maintain resonance. Capacitive loading of the ends, either through proximity of the antenna to other objects or through the addition of capacitance hats, will reduce this requirement to a more practical value.

What Inductance Values?

As a matter of personal interest, this writer has been doing experimental work for a number of years with off-center-loaded antennas. One big drawback to such experimentation was the ever-present need for a large amount of cut-and-try work to arrive at resonance whenever a new set of dimensions was to be used. Probably the number of pruned-off turns from coil stock from such experiments, if straightened out and soldered end to end, would make up several full-sized half-wave antennas for the 160-meter band. Therefore, most of the writer's work of late in this area has been in going through paperwork exercises, looking for a way whereby at least "ball-park" values of inductance needed for a particular system could be calculated.

The equation contained in the Mobile chapter of *The ARRL Antenna Book* for determining the capacitance of a vertical antenna shorter than a quarter wavelength looked promising in early computations,

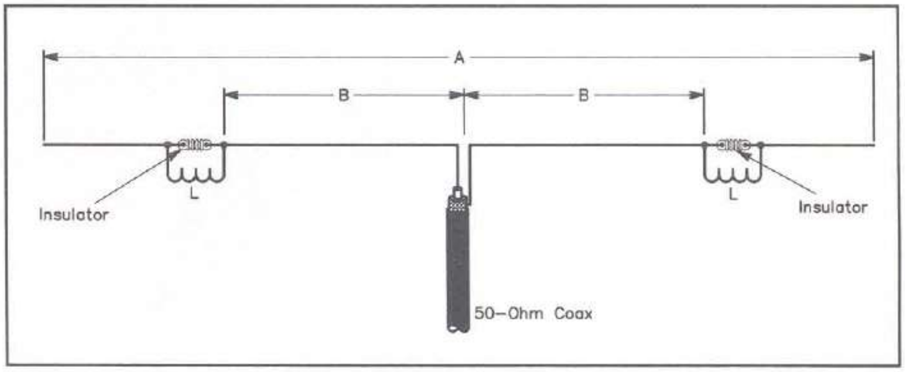

Figure 1—A dipole antenna lengthened electrically with off-center loading coils. For a fixed dimension A, greater efficiency will be realized with greater distance B, but as B is increased, L must be larger in value to maintain resonance.

and, indeed, it became the basis for the calculation procedure which finally resulted.

This procedure has been found to produce results much closer than mere "ball-park" values for the necessary inductance – for wire antennas "in the clear" at moderate heights, the final inductance values found by cut-and-try pruning for lowest SWR at the desired frequency have been so close to the value from calculations that a laboratory bridge was necessary to measure the difference. The results are equally good for elements using tubing. Once the needed inductance value is determined by calculations, it is generally found sufficient to obtain coil dimensions from an ARRL L/C/F Calculator (see LaPlaca*) or by equation. Any significant pruning which has been found necessary could always be attributed to objects in proximity to the ends of the antenna.

The complete set of calculations is expressed in the mathematical relationship below as Eq. 1, presented here primarily for mathematics buffs or those having access to electronic computers.

This equation yields the inductance required, in microhenrys, for single-band resonance of a shortened antenna of a particular physical size at a given frequency, for a specific position of the loading coils from the center of the antenna. To spare the reader the task of performing some rather tedious calculations, Figure 2 has been prepared from Eq. 1. The curves of the chart have been normalized, and may be used for any frequency of resonance. The chart is based on a halfwavelength/diameter ratio of the radiator of approximately 24,000. (This corresponds to No. 14 wire on 80 meters or No. 8 wire on 160 meters.) For "thinner" conductors, the required inductance will be somewhat greater than that determined from Figure 2, and less inductance will be required for "thicker" conductors.

The use of the chart is as follows: At the intersection of the appropriate curve from the body for dimension A and the proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at the left. Dimensions A and B are shown in Figure 1, and for use with the chart are expressed as percentages. Dimension A is taken as percent length of the shortened antenna with respect to the length of a resonant half-wave dipole of the same conductor material. Dimension B is taken as the percent of coil distance from the feed point to the end of the shortened antenna. For example, resonating an antenna which is 50% or half the size of a half-wave dipole (one-quarter wavelength overall), with loading coils positioned midway between the feed point and each end (50% out), would require loading coils having an inductive reactance of approximately 950 ohms at the operating frequency. If the antenna is hung "in the clear," and if the length/diameter ratio of the conductor is near 24,000, inductance values as determined from the chart will be very close to actual values required. (Eq. 1 above takes the diameter of the radiator into account, and thus may be used for any length/diameter ratio.) For practical purposes, dimension B may be taken as that distance from the center of the feed-point insulator to the inside eye of the loading-coil insulator, and dimension A as the eye-to-eye distance inside the end insulators (which are not drawn in Figure 1).

Proximity of surrounding objects in individual installations may require some pruning of the coils, and the exact amount of final inductance required should be determined experimentally. If the antenna is hung in inverted-V style, with the ends brought near the earth, the required inductance will almost always be somewhat less than that determined from the chart or equation. A grid-dip meter, Macromatcher, (see Hall and Kaufmann*) or SWR indicator may be used, during the final adjustment procedure.

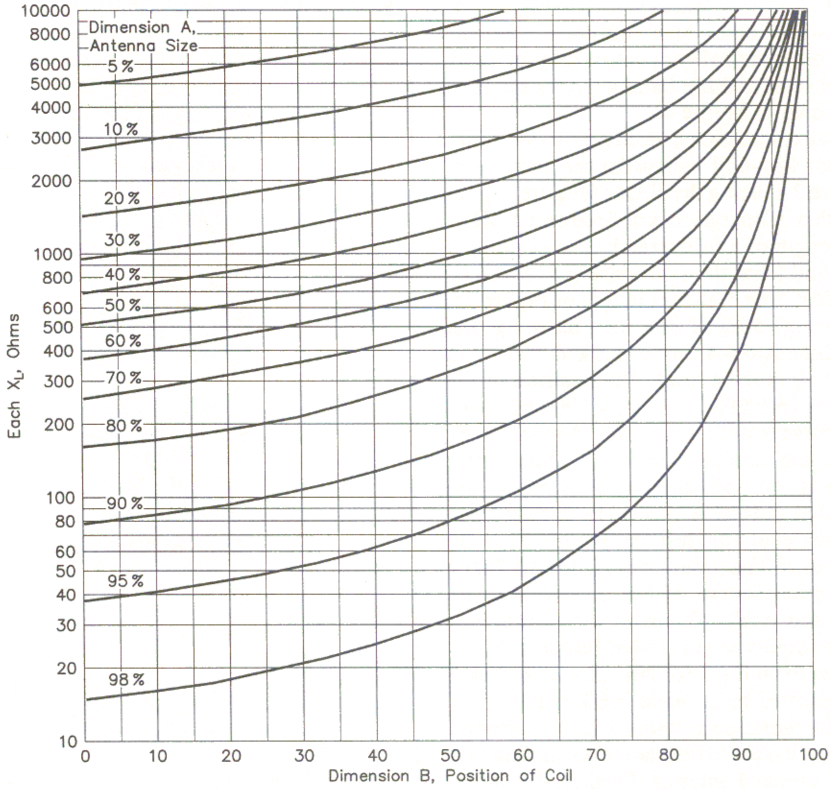

Figure 2—Chart for determining approximate inductance values for off-center-loaded dipoles. At the intersection of the appropriate curve from the body of the chart for dimension A and the proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at left. See Figure 1 regarding dimensions A and B.

(Eq. 1):

$$L_{\mu H} = \frac{10^6}{68\pi^2 f^2} \left\{ \frac{\left[ln \frac{24\left(\frac{234}{f} - B\right)}{D} - 1 \right] \left[\left(1 - \frac{fB}{234}\right)^2 - 1 \right]}{\frac{234}{f} - B} - \frac{\left[ln \frac{24\left(\frac{A}{2} - B\right)}{D} - 1 \right] \left[\left(\frac{\frac{fA}{2} - fB}{234} \right)^2 - 1 \right]}{\frac{A}{2} - B} \right\}$$

where
$L_{\mu H}$ = inductance required for resonance
ln = natural log
f = frequency, megahertz
A = overall antenna length, feet
B = distance from center to each loading coil, feet
D = diameter of radiator, inches

Practical Antennas

Although one might erect an inductively loaded antenna that is cut for a single amateur band, it is possible to use the antenna itself for two, three, or more bands of operation, if provision is made to lower the antenna for band changes. A simple rope halyard and pulley arrangement at one of the supports will do the trick. Figure 3A shows a 3-band antenna of this nature, for 160, 80, and 20 meters. If the insulators shown are left open, with nothing bridging them, the antenna is a simple half-wave dipole cut for 14.18 MHz. (The 48.5-foot lengths act merely as support wires, and have negligible effect on operation of the antenna.) If the insulators are bridged with short lengths of antenna wire, the antenna becomes a center-fed 80-meter dipole, resonant at about 3.6 MHz. For 160-m operation the 20-meter insulators may be bridged with loading coils to resonate the antenna at 1.8 MHz, as shown in Figure 3A. Burndy or other manufacturers' "Servit" type of electrical connectors may be used, for ease in making band changes quickly, as shown in Figure 4.

The calculation procedure for determining loading-coil values for the antenna of Figure 3A, using the chart of Figure 2, goes like this. If operation is desired on 1.8 MHz, the length of a full-sized half-wave dipole is found from the relationship $468/f$ to be 260 feet. The 130-foot length of Figure 3A represents 50% of this size, meaning that the dimension-A curve marked "50%" in Figure 2 is to be used. The position of the coils is $16.5/(16.5 + 48.5) \times 100$ or 25% of the distance out from center, dimension B. From the intersection of 25 (horizontal scale at bottom), and the 50% curve, the required inductive reactance is read from the scale at the left of Figure 2 to be 650 ohms. The inductance, L, is $650/2\pi f$ or 57.5 microhenrys, if No. 8 wire is to be used. For smaller diameter wire, the inductance should be somewhat larger. (Calculations from Eq. 1 for No. 12 wire indicate the required inductance is 60.99 µH.)

The radiation resistance of a shortened antenna loaded to resonance is less than that of a full-sized antenna. Further, the shortened antenna is "sharper," meaning that the change in reactance versus frequency is greater. In other words, the shortened antenna acts as a tuned circuit having a higher Q than a full-sized antenna. To check these characteristics, the line input impedances for the antenna of Figure 3A were measured with a laboratory bridge, and the electrical line length at the measurement frequency was then taken into account to determine the impedance at the antenna feed point. The antenna was constructed of No. 12 wire and hung at a height of 50 feet as a "flat-top" radiator.

The solid curve of Figure 3B is a plot of the feed-point impedance versus frequency for this antenna. The plot on Smith Chart coordinates is more meaningful than a simple SWR-vs.-frequency curve because the magnitudes of the resistive and reactive components are shown, as well as the sign of the reactance. (Capacitive reactance is negative, plotted to the left of the vertical center line, and inductive reactance is positive, plotted to the right.) In this presentation, a 50-ohm nonreactive impedance will appear at the exact center of the chart. The SWR in 50-ohm line for a given frequency may be determined by first noting the distance from the center of the chart to the particular impedance plot on the curve, and next measuring this same distance down the vertical center line from chart center (a drawing compass is helpful for this task), and finally dividing 50 into the value read at that point on the center line. For example, the SWR at 1.8 MHz equals 120/50 or 2.4, as indicated by the segment of the 2.4 SWR circle in Figure 3B. It may be seen that resonance (zero reactance) occurs at approximately 1810 kHz, where the resistance is about 22 ohms. The SWR at resonance is 2.33:1, and climbs to 3:1 at 1825 kHz. At 1850 kHz, the SWR is 10:1. Without any matching provisions the antenna is relatively sharp, as mentioned earlier. If one sets the usable bandwidth as the frequency range where the SWR is 3:1 or less, it is approximately 35 kHz, or 1.9% of the resonant frequency. As far as efficiency is concerned, ohmic losses are low, and the antenna is a good performer on 160 meters. Because of its horizontal polarization, it has proved to be most effective at night, and stations several hundred miles away have been worked with S-9 reports received for the 50-watt signal.

For a comparison of impedances, the broken curve of Figure 3B is a plot of measured impedances of a full-size half-wave dipole, 260 feet long overall, hung in place of the shortened antenna. From this curve it may be seen that resonance occurs at 1810 kHz, where the resistance is 59 ohms. The 3:1-SWR bandwidth for the half-wave antenna is in the order of 60 kHz, or 3.3% of the reso-

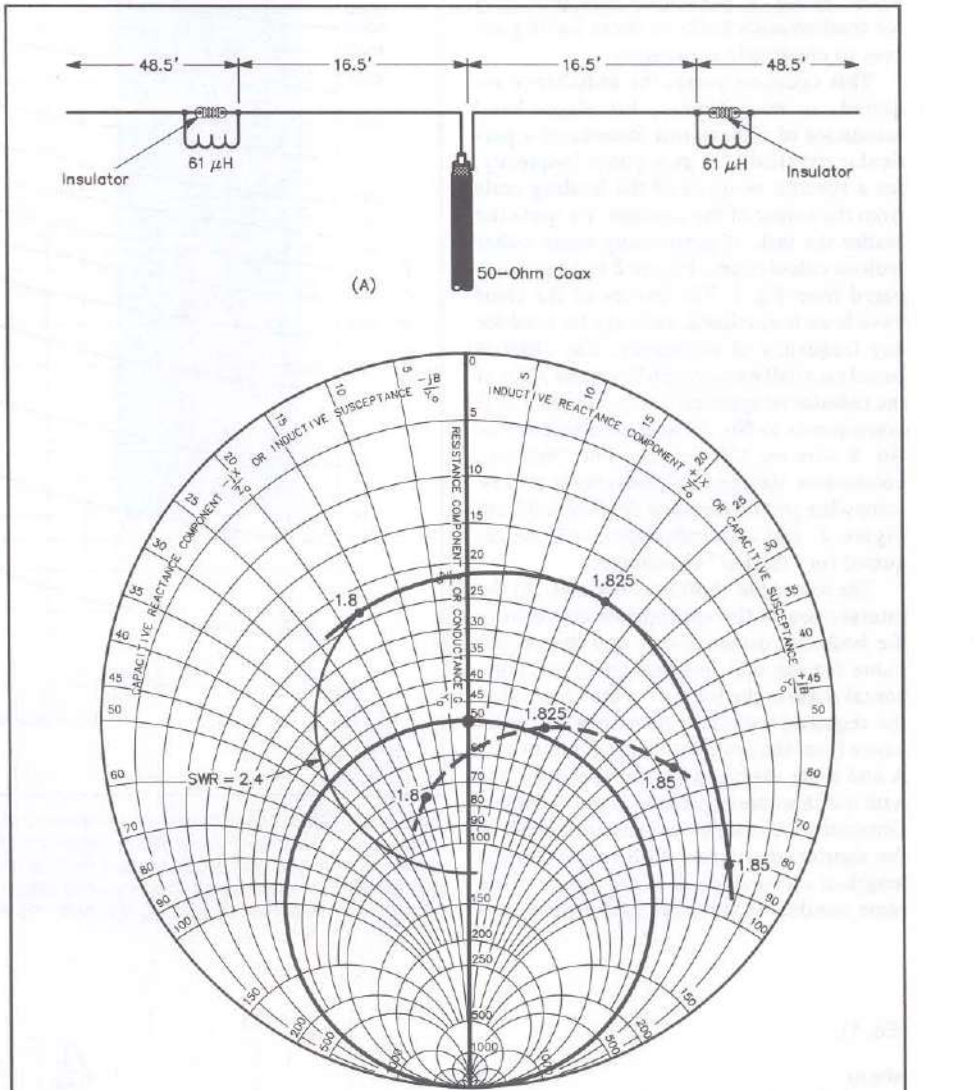

Figure 3—At A, an 80-meter dipole loaded for 160-meter operation. The inductors are 37 turns of coil stock having a 3-inch diameter and 10 turns per inch (B&W 3035). At B, the impedance plot of this antenna installed at a height of 50 feet (solid curve), and the plot of a 160-m half-wave dipole (broken curve).

nant frequency. It is interesting to note on this curve that the SWR at resonance is 1.18:1, and that it is a somewhat lower value, 1.15:1, at a frequency a few kilohertz above resonance. (Measurements were made every 5 kHz across this band, but plot points are shown only for 25-kHz increments to avoid crowding of the data.) This evidence refutes the oft-heard statement that the SWR-vs.-frequency curve is *always* lowest at antenna resonance. Points to remember are that the SWR in a transmission line is completely dependent upon the characteristic-impedance value of the line in use. Using a line of different impedance may shift the position of the SWR curve along the frequency axis in a simple SWR-vs.-frequency plot. This is definitely true in this case — if the 160-meter half-wave dipole were to be fed with 75-ohm line, the lowest SWR would occur at a frequency about 5 kHz *below* antenna resonance, whereas with 50-ohm line the lowest SWR is at a frequency slightly *above* resonance. The reason this happens is that the resistive component of the impedance, which consists of the radiation resistance plus any loss resistance, is not constant with frequency, even over a rather narrow frequency range. It must be acknowledged that the differences here are very slight, however, and for practical purposes the frequency of lowest SWR is (within a few kilohertz) the resonant frequency of the antenna.

Another point concerning the SWR values bears noting. The values as determined from the plots in the manner described above are quite accurate, having been determined by measurements with laboratory equipment. In contrast, measurements with simple SWR indicators usually cannot be relied upon for anywhere near the equivalent accuracy.

For example, the author owns a commercially manufactured SWR indicator of the Monimatch type (see McCoy*) which, under a particular set of conditions, indicates a 2.5:1 SWR in a line where laboratory measuring equipment shows the *true SWR to be 4:1*. A significant difference! Herein lies another reason why impedance plots on Smith Chart coordinates are more meaningful than a simple SWR-vs.-frequency curve — greater accuracy may generally be expected.

A Half-Size 80-Meter Antenna

Figure 5A shows the 3-band concept described earlier as it can be applied to 80, 40, and 20 meters. Its overall length is 66 feet, not a difficult length to use on a small lot. This antenna was constructed for 80-m operation with a design-center frequency of 3.55 MHz, using No. 12 antenna wire and 40-μH loading coils—27 turns of stock having a diameter of 3 inches and a pitch of 10 turns per inch (tpi). Feed-point impedances versus 80-m frequency for the antenna, hung at a height of 50 feet, are shown by the solid curve at B of Figure 5. Actual resonance occurred at 3.54 MHz, where the resistance was about 26 ohms. The bandwidth within which the SWR is 3:1 is

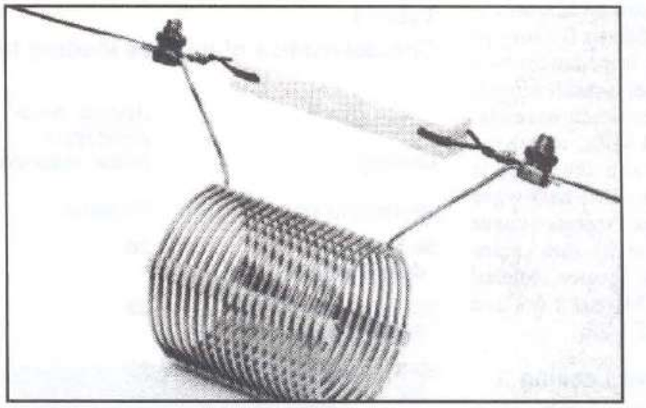

Figure 4—Copper electrical service connectors, sold under one trade name of Servit, provide a simple means of installing the loading coils. The antenna wire and the ends of the coil wires should be tinned to prevent corrosion. In addition, a protective coating of acrylic spray may be used at each connection.

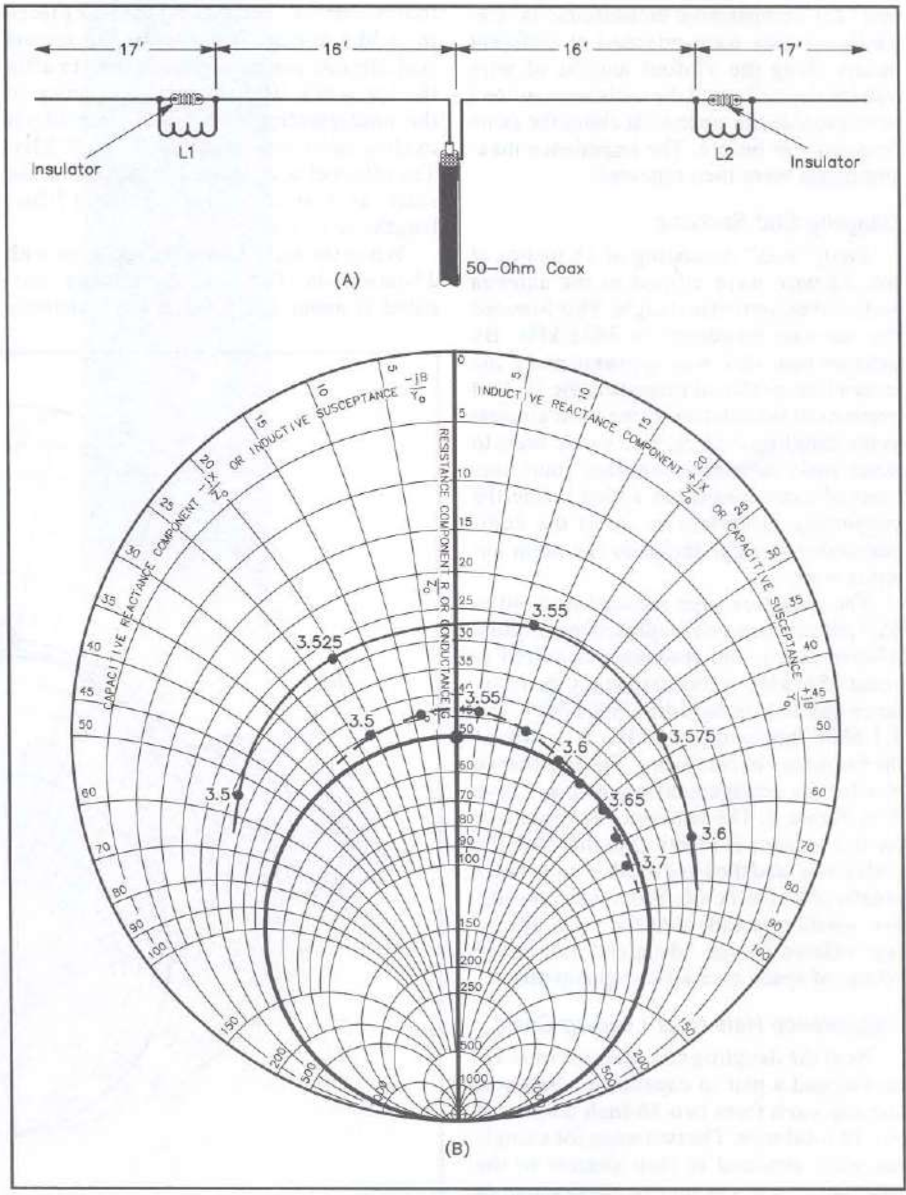
17'
16'
16'
17'
L1
L2
Insulator
Insulator
50–Ohm Coax
(A)
INDUCTIVE REACTANCE COMPONENT +jX/Z0 OR CAPACITIVE SUSCEPTANCE +jB/Y0
CAPACITIVE REACTANCE COMPONENT −jX/Z0 OR INDUCTIVE SUSCEPTANCE −jB/Y0
RESISTANCE COMPONENT R/Z0 OR CONDUCTANCE G/Y0
3.5
3.525
3.55
3.575
3.6
3.5
3.55
3.6
3.65
3.7
(B)

Figure 5—At A, a 40-meter dipole loaded for 80-meter operation. For resonance at 3.55 MHz the coils should be approximately 40 μH (27 turns of stock); 3.75 MHz, 35 μH (24 turns); 3.9 MHz, 31 μH (22 turns); and 4.0 MHz, 29 μH (21 turns). These are circulated inductance values for No. 12 antenna wire. Coil stock referenced above is 3-inch diameter, 10 turns per inch (B&W 3035). At B, the impedance plot of the 3.55-MHz version (solid curve) and of an 80-m half-wave dipole (broken curve).

60 kHz, or 1.69% of the resonant frequency.

Also shown in Figure 5B, by the broken curve, are the feed-point impedances of a half-wave dipole, 132 feet overall length, hung in place of the shortened antenna. Resonance occurs at 3.54 MHz, where the resistance is 43.5 ohms and the SWR is 1.15:1. The broader nature of the half-wave antenna is exhibited by the "tighter" curve which swings closer to the 50-ohm center point of the chart than the shorter, loaded antenna. The SWR at 3.5 MHz is 1.6:1 and remains below 3:1 to 3.67 MHz.

Capacitive and Inductive Loading

One would assume that a combination of capacitive and inductive loading might provide a different feed-point impedance than would inductive loading alone, because of different current distributions in the radiators. To check out this assumption, the antenna of Figure 5A was used as a "test bed" for comparative measurements. Capacitance hats were attached at different points along the 17-foot lengths of wire outside the coils, and the coils were pruned to reresonate the antenna at about the same frequency as before. The impedance measurements were then repeated.

Dangling End Sections:

First, "hats" consisting of 18 inches of No. 12 wire were affixed to the antenna ends and permitted to dangle. This lowered the resonant frequency to 3435 kHz. By calculations, this was approximately the same effect as that of extending the 17-foot portions of the antenna by the same amount as the dangling lengths, so it would seem to make little difference whether short sections of extra length are added inside the supporting insulators or are at the ends, suspended at right angles to the main antenna wire.

The inductors were reduced from 40 to 36.5 µH (25-turn coils replaced the original 27-turn coils), and resonance occurred at about 3575 kHz. At this frequency the resistance was 26 ohms and the SWR 1.90:1. The 3:1-SWR bandwidth, 64 kHz, is 1.79% of the frequency of resonance. The impedance plot for this arrangement is shown as Curve A in Figure 6. The resistance at resonance for this antenna is identical to that with the coils alone, and the bandwidth is only 4 kHz greater, 64 kHz vs. 60. From these results, one would conclude that the main advantage offered by the "danglers" is a small saving of space over a flat-top antenna.

Capacitance Hats Near Loading Coils:

Next the dangling end sections were removed and a pair of capacitance hats was formed, each from two 36-inch lengths of No. 12 solid wire. The two wires for a single hat were attached at their centers to the antenna wire at a point just outside one of the loading coils. The hat wires were then bent radially to form an X at right angles to the antenna wire, like four spokes of a wheel with the main antenna wire at the hub. The diameter of the X-shaped hat was thus 36 inches. The second hat was placed in a like manner just outside the second coil. Burndy connectors were used to affix the hat wires. The resonant frequency of this configuration with the original 40-µH loading coils was found to be 3290 kHz. The effect of adding the hats was about the same as that of extending the 17-foot lengths to 19 feet.

When the inductors were replaced with 23-turn coils (32.5 µH), the antenna resonated at about 3.575 MHz, the resistance being 23 ohms. The SWR at resonance is 2.15:1, and the 3:1-SWR bandwidth for this configuration is 60 kHz, 1.68% of the resonant frequency. The impedance of this arrangement versus frequency is shown by Curve B of Figure 6.

It is surprising to note that, by the standards of most amateurs, the characteristics of this antenna are not as good as those of the same length antenna with loading coils alone. The SWR at resonance for the antenna with combination capacitive and in-

Table 1

Characteristics of various loading techniques, 66-foot 80-m dipole.

Loading	*Approx. feed-point resistance, resonance*	*SWR at resonance*	*3:1-SWR bandwidth, % of resonant freq.*
40-µH coils only	26 ohms	1.92:1	1.69
36.5-µH coils, 18″ dangling ends	26	1.90:1	1.79
36″ hats outside 32.5-µH coils	23	2.15:1	1.68
30-µH coils, 36″ hats at ends	25	1.98:1	2.05
None (λ/2 dipole)	43.5	1.15	Greater than 3.6

Coil positions for each loaded antenna were 16 feet from antenna center. All antennas were constructed of No. 12 wire and installed at a height of 50 feet.

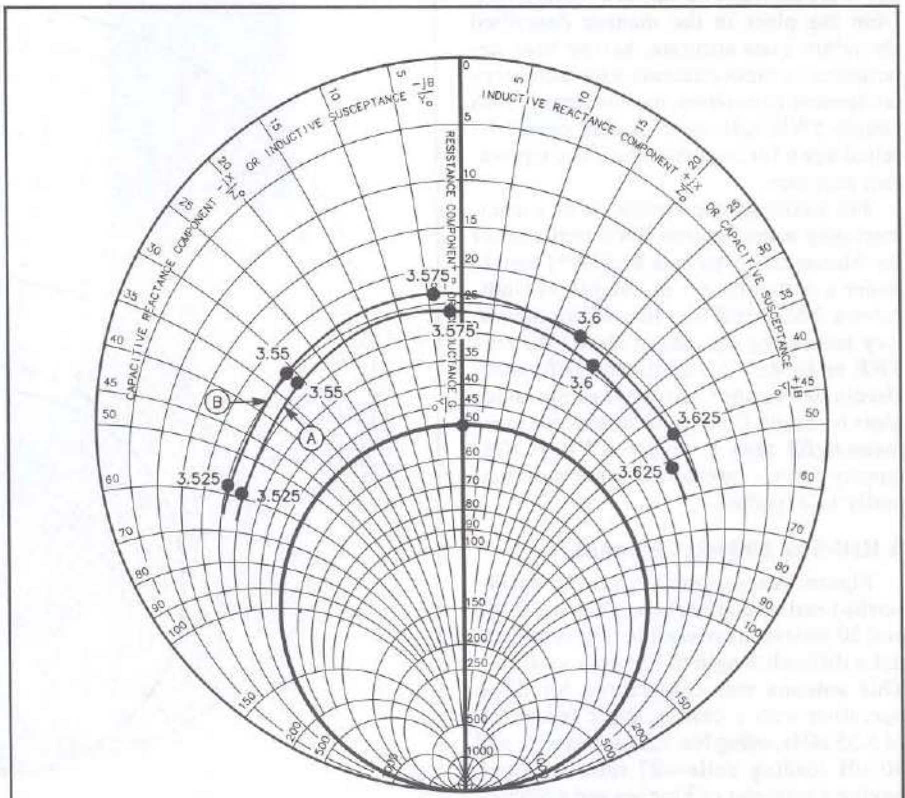

Figure 6—Curve A is the impedance plot of the antenna of Figure 5A with 1.5-foot dangling end sections added and the coil trimmed to restore resonance near the original frequency. Curve B is a plot of the same antenna with X-shaped capacitance hats added at a point just outside the loading coils (dangling sections removed and coils trimmed to reestablish resonance).

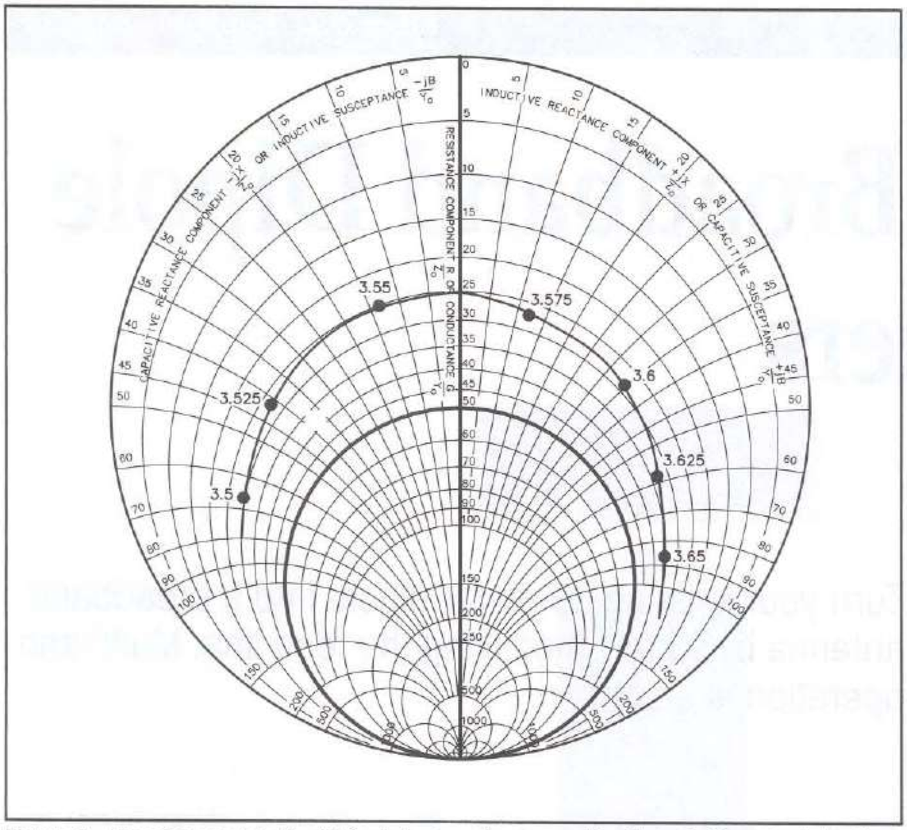

Figure 7—Impedance plot of a 66-foot dipole using a combination of off-center inductive loading and capacitive end loading. Of all the shortened configurations tried, this arrangement offered the greatest bandwidth.

ductive loading is higher (2.15 vs. 1.92), and the 3:1-SWR bandwidths are the same, 60 kHz. Perhaps a significant factor here, though, is that the diameter of the capacitance hats used for these measurements was small, only .011 wavelength. Supporting much larger hats presents mechanical problems with wire antennas, however, as even these were a bit flimsy and would require reshaping after gusty weather.

Capacitance Hats at Antenna Ends:

Finally, the X-shaped capacitance hats were moved to the outside ends of the antenna, just inside the end insulators. With the original 40-µH coils, resonance appeared at 3215 kHz. From calculations, it was as if the 17-foot end sections were actually 21 feet long. With 30-µH coils (22 turns) in place, the resonant frequency was 3560 kHz. At this frequency the resistance was 25 ohms and the SWR 1.98:1. The 3:1-SWR bandwidth is 73 kHz, or 2.05% of the resonant frequency. The impedance plot of this antenna is given in Figure 7.

It is interesting to note that the position and shape of the plot for this antenna on Smith Chart coordinates is nearly identical to that for the same length antenna with loading coils only, the solid curve of Figure 5B. For this antenna, however, the plot points for 25-kHz frequency increments appear closer together, which accounts for the increased bandwidth.

Conclusions:

The measured characteristics of these various configurations of loading for the 80-meter antennas are tabulated in Table 1. Remember that the overall "flat-top" length of each antenna arrangement, is 66 feet, and that the loading coils are always positioned 16 feet each side of the center of the antenna, being pruned for resonance at approximately 3550 kHz. For comparison, information for a half-wave dipole is also included.

Of the various arrangements, capacitive end loading decidedly provides the greatest bandwidth, excepting the full-size half-wave antenna, of course. Although there are slight differences in the resistance value at resonance, all are of the same order of magnitude. These values, as well as those for the 160-m antenna discussed earlier, tend to confirm a *broad* rule of thumb that the writer has formulated for this type of antenna: The feed-point impedance value at resonance is roughly proportional to the length of the antenna. That is, a loaded antenna which is half the size of a half-wave dipole will have approximately half the radiation resistance of the full-sized antenna.

Eq. 1 given earlier or the chart of Figure 2 allows one to calculate loading-coil values for antennas with loading coils only. Additional capacitive loading is not taken into account. Calculating the effects of various capacitive loading arrangements appears to be difficult, and work remains to be done in this area.

Multiband Antennas with Loading Coils

All of the foregoing material has been devoted to the loading of an antenna for resonance at a single frequency. Resonated as described, the antenna is electrically a half wave in length. It will, however, operate well on higher frequencies — frequencies at which it is an odd multiple of half waves, in *electrical* length . . . three half waves, five half waves, etc. Because of the lumped loading of the shortened antenna, these higher frequencies will likely not be closely related to odd-order harmonics of the fundamental frequency, as the case would be for a nonloaded radiator. (For example, it is a well-known fact that a 7-MHz half-wave dipole operates well on its third harmonic, 21 MHz.)

A loaded dipole will become an electrical 3/2-λ antenna at some frequency *below* that which is three times the fundamental resonant frequency. Depending upon the overall antenna length, coil value, and coil position, it is possible for an 80-meter loaded dipole to become a 3/2-λ performer on 40 meters. With such an arrangement, one would have a dual-band antenna without requiring the use of traps. The idea can be expanded upon to arrive at a loaded antenna without traps which will operate on more than two bands. This scheme offers considerable constructional simplification as compared with trap arrangements.

The multiband loading-coil concept has been recognized for better than half a century, but little use of the technique has been made by amateurs. Some years ago a very good article on the subject a was published by William Lattin, W4JRW.* That article is recommended reading for anyone interested in more details on the concept. Supplemental information has been published by Buchanan.* Attempts by this writer to calculate antenna sizes and coil values for dual-band antennas have met with some success. From calculations and experiments to date, it appears that with only two loading coils (one each side of center), the antenna must always be greater than a half wave in physical length for the higher of the two frequency bands. In other words: any 80/40 meter arrangement, for example, apparently would need to be longer than 66 feet from tip to tip. However, much work also remains to be done in this area.

References

Buchanan, "An Inexpensive 40- and 80-Meter Antenna," Hints and Kinks, *QST*, September, 1962, p. 62.

Hall and Kaufmann, "The Macromatcher, An RF Impedance Bridge for Coax Lines," *QST*, January, 1972, p. 14. Information also contained in the 50th and subsequent editions of *The Radio Amateur's Handbook*, and the 13th edition of *The ARRL Antenna Book*.

LaPlaca, "Using the ARRL L/C/F Calculator," *QST*, December, 1973, p.26.

Lattin, "Multiband Antennas Using Loading Coils," *QST*, April, 1961, p.43.

McCoy, "Monimatch, Mark II," *QST*, February, 1957, p. 38. Also, "An Etched-Circuit Monimatch For Checking Your Antenna System," *QST*, October, 1969, p.29.

Sevick, "The Ground-Image Vertical Antenna," *QST*, July, 1971, p. 16. Also, "The W2FMI Ground-Mounted Short Vertical," *QST*, March, 1973; "A High Performance 20- 40- and 80-Meter Vertical System," *QST*, December, 1973, p. 30; and "The Constant-Impedance Trap Vertical," *QST*, March,1974,p.29.

By Frank Witt, AI1H **From *QST*, September 1993**

A Simple Broadband Dipole for 80 Meters

Turn your existing 80-meter dipole into a broadband antenna by simply modifying the feed line. Multiband operation is an option.

A conventional coax-fed, half-wave dipole doesn't provide a low SWR over the entire 80-meter band—an inconvenience for those of us who like to operate phone and CW on that band. Several approaches to overcoming this limitation, short of an antenna tuner in the station, have been described.[1,2] The antenna system described here is simpler than any of its predecessors and has the following features:

- A 2:1 SWR or better is achieved over all or most of the 80-meter band.
- Antenna length and appearance are the same as those of a conventional half-wave dipole. Consequently, it's lightweight and has small wind and ice loading.
- The antenna configuration permits multiband operation with a single feed line.
- The losses due to broadband matching are acceptable.
- The cost is about the same as a conventional half-wave dipole.

All the SWR data given in this article were measured at the transmitter end of the feed line. The reference impedance is 50 Ω, since most equipment is designed for this impedance. The term *antenna system* as used throughout this article includes not only the radiating wire, but also the feed line, balun (if used), any lightning-protection measures, antenna tuner and so forth.

The dipole antenna itself is not broadband; the system uses a broadband *match*. The key broadbanding element of this antenna system is the *transmission-line resonator*: Part of the transmission line compensates for the reactance presented by the dipole away from its resonant frequency. This part of the line is a multiple of an electrical half wavelength. Another part of the line presents an appropriate source impedance to the transmission-line resonator (TLR).

First I'll describe a version of the broadband antenna system, along with some practical results. Then I'll cover the important matter of antenna-system loss. Following that are some variations to suit specific requirements, and a method for using the antenna for several bands. I'll also compare transmission-line-resonator broadbanding to other broadbanding methods.

The 80-Meter Broadband Antenna System

Figure 1 shows the simple broadband antenna system as used at my station. The antenna proper is a center-fed half-wavelength dipole. The transmission line is segmented into one electrical wavelength of 50-Ω coax and an electrical quarter wavelength of 75-Ω coax. The calculated and actual lengths are shown in Table 1. Lengths were calculated using the formulas given later in this article, using a center frequency (F_0) of 3.75 MHz and VF (velocity factor) of 0.66. The actual lengths resulted after I performed the tuning procedure described later. Manufacturing variations from the published cable velocity factors, and some stretching of the coax, contributed to the differences between actual and measured values. (The actual lengths were measured on untensioned cable.) The antenna is installed as an inverted V with a 140° included angle and an apex height of 60 feet. The wire size is #14, but is not critical.

Table 1
Calculated and Actual Lengths of the Broadband Dipole Antenna at AI1H

	Calculated	*Actual*
1/4-λ Coax	43.3 feet	43.3 feet
1-λ Coax	173.1 feet	170.5 feet*
Dipole	124.5 feet	122.7 feet

*Includes 11 inches for balun.

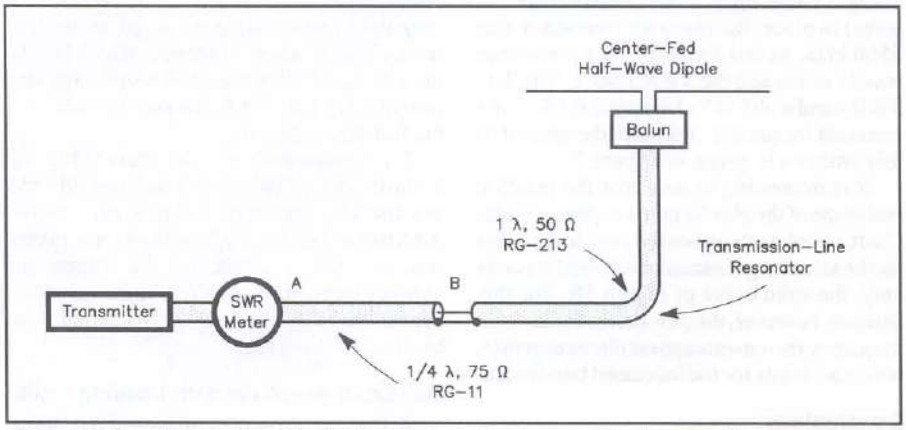

Figure 1—One form of the simple broadband antenna system. It resembles a conventional dipole except for the 1/4-wavelength, 75-Ω segment. Points A and B are discussed in the text.

This system's SWR (at the transmitter) as a function of frequency is shown in Figure 2. For comparison, the SWR for the same dipole fed with about 5/4 wavelengths (214 feet) of RG-213 coax is also shown. (This is the same total length as the RG-213 and RG-11 segments used in the broadband system.) The broadband system's 2:1 SWR bandwidth is 2.2 times that of the conventional system—and the *only* difference is the feedline configuration!

The radiating properties of the broadband antenna over the 80-meter band are essentially identical to those of a dipole cut for any specific frequency in the band. Also, since the antenna system is designed for a 50-Ω transmitter, the feed-line length may be extended by adding the required length of 50-Ω coax between the transmitter and the quarter-wave segment (point A in Figure 1).

A 1:1 current balun should be installed at the antenna's feed point. I use the balun on general principles. Often, it provides no visible difference in operation, but the balun does minimize feed-line radiation. You can determine whether your antenna needs a balun by measuring the SWR versus frequency with and without a balun installed. If the balun is not needed, the two sets of data will be identical.

Antenna-System Losses

It's important to know the losses in any antenna system. This is especially true for broadband antennas, because loss alone can broadband an antenna system. As the next section shows, the configurations presented in this article do not yield a significant loss penalty. Although other loss contributors exist in antenna systems, we will focus on the primary ones: *feed-line loss* and *mismatch loss*. Other losses, such as ohmic loss in the antenna wire, are the same for both the conventional and broadband systems described here.

Feed-line loss is the easiest to understand. It is unavoidable, and is lowest when the feed line is *flat* (when the line SWR is close to 1:1). At HF, feed-line loss results primarily from ohmic losses in the copper conductors.

Mismatch loss occurs when the impedance seen by the transmitter is not the complex conjugate of the transmitter's impedance (when the line SWR at the transmitter is not 1:1). For a 50-Ω transmitter, the mismatch loss is 0 dB when the load impedance is 50 Ω. When the load impedance is not 50 Ω, the mismatch loss can be made to be 0 dB if a transmitter with a tunable output stage (such as a conventional tube-type linear amplifier) is tuned for a conjugate match. An antenna tuner can also provide this match. In this case, however, the antenna-tuner loss (perhaps as much as 1 dB) replaces the mismatch loss in the total-loss equation. That subject isn't discussed here.

If you don't use an antenna tuner and the transmitter has a fixed-tuned 50-Ω output, loads that present the transmitter with an SWR under 2:1 are highly desirable. The impact of high SWR on mismatch loss will become clear in the next section.

Loss must be kept in perspective. All of the broadband antenna systems described here have a worst-case total loss of less than 3 dB—not enough to notice in many 80-meter QSOs. (If the loss is 3 dB, half of the transmitter's output power is radiated and half is lost elsewhere.) The main effect of loss is stress on system components: that on the transmitter due to the mismatched load, and that on the transmission line due to heating.

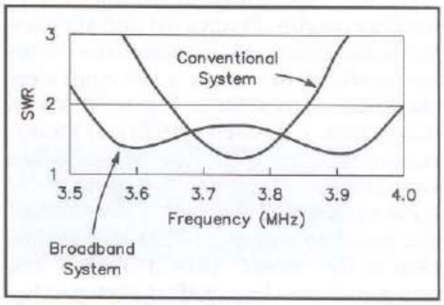

Figure 2—Measured SWR versus frequency for the broadband and conventional antenna systems.

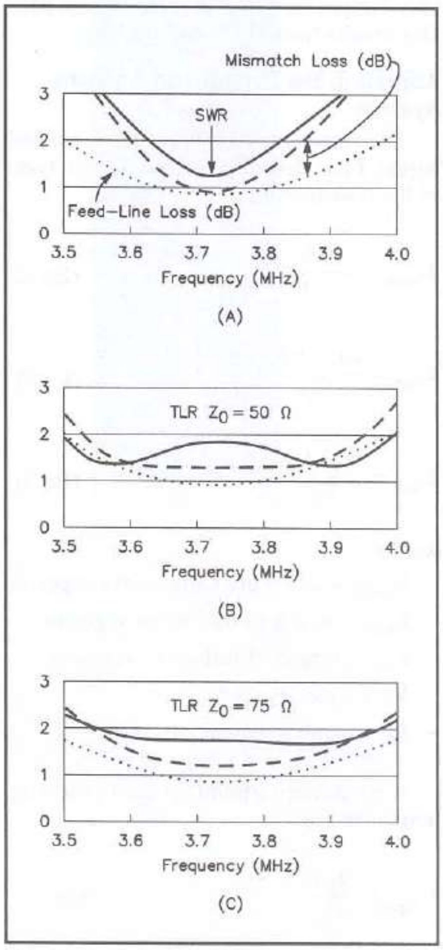

Figure 3—Antenna-system configurations for long feed-line runs. The solid lines are SWR; the dotted lines are feed-line loss; and the dashed lines are feed-line loss plus mismatch loss. At A, a conventional system using an RG-213 feed line; at B, the feed line is a 1/4-λ section of RG-11 (75 Ω) followed by 1 λ of RG-213; and at C, a 1/4-λ segment of RG-11 is followed by 1 λ of RG-11 (one 5/4-λ piece of RG-11). The total feed-line length in each case is 216.4 feet.

Variations

The broadband antenna system described above is well-suited for the installation at my station, where the distance between the shack and the antenna is rela-

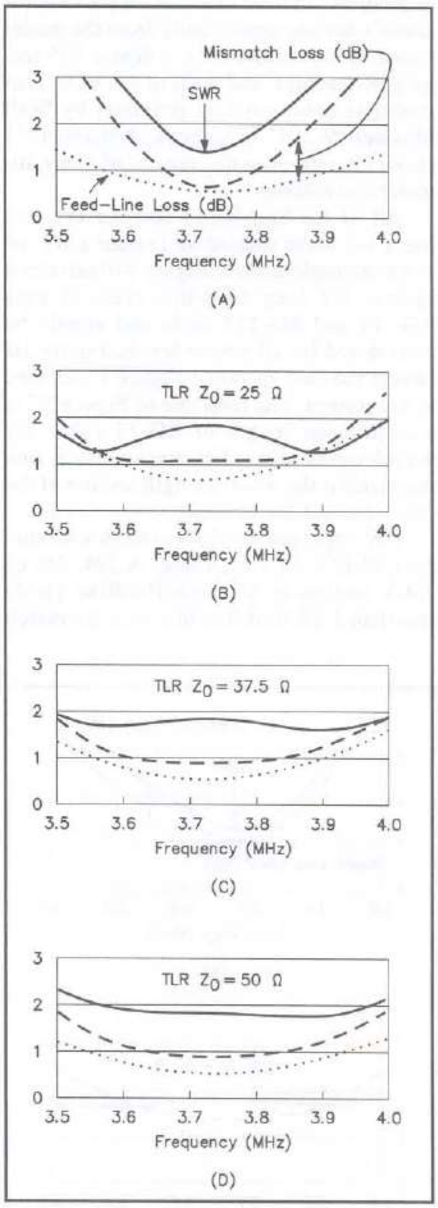

Figure 4—Antenna systems for high power and shorter feed-line runs. The solid lines show SWR; the dotted lines represent feedline loss; and the dashed lines show feed-line loss plus mismatch loss. At A, the feed line is RG-213; at B, it's a 1/4-λ section of RG-11 followed by two paralleled 1/2-λ lengths of RG-213; at C, a 1/4-λ segment of RG-11 is followed by two paralleled 1/2-λ lengths of RG- 11, and at D, 1/4-λ of RG-11 is followed by 1/2-λ of RG-213. The total feed-line length in each case is 129.8 feet.

tively long (more than 200 feet) and because I use a 1-kW amplifier. Other feedline combinations are better suited to other installations. Some of these are shown in Figures 3 through 5, along with calculated SWR and loss data. From this information, you can select an appropriate feed-line combination for your needs.

The figures also show the characteristics of conventional dipole antenna systems. If you compare them, you'll see that the transmission-line resonator provides broadbanding without a significant loss penalty. I haven't tried all these combinations, but based on my experience, they should perform as predicted in most situations if the radiator doesn't deviate significantly from the model I used in my calculations: a dipole 125 feet long, 40 feet high, and made of #14 wire. This model is based on data provided by Walt Maxwell, W2DU, in his book, *Reflections*.[3] I chose his data since it is typical of many 80-meter installations.

All of the broadband antenna systems use a 1/4-wave section and either a 1/2- or 1-wavelength section. Figure 3 illustrates a system for long feed-line runs. It uses RG- 11 and RG-213 cable and should be considered for all power levels. Figure 3B covers the case shown in Figure 1 and used at my station. The feed line of Figure 3C is a continuous length of RG-11 cable 5/4 wavelengths long. The transmission-line resonator is the 1-wavelength section of the cable nearest the antenna.

This approach would also work with surplus 75-Ω CATV Hardline. A 3/4, 5/4 or 7/4-λ section of 1/2-inch Hardline yields less than 2 dB feed-line loss plus mismatch loss over the entire band, and less than 1 dB total loss over any 300 kHz of the band. This configuration is particularly attractive to contesters and DXers, because even a fairly long line7/4 λ is 372 feet of 1/2-inch CATV Hardline—gives low loss and a very good match over, say, the 3.5- to 3.8-MHz range.

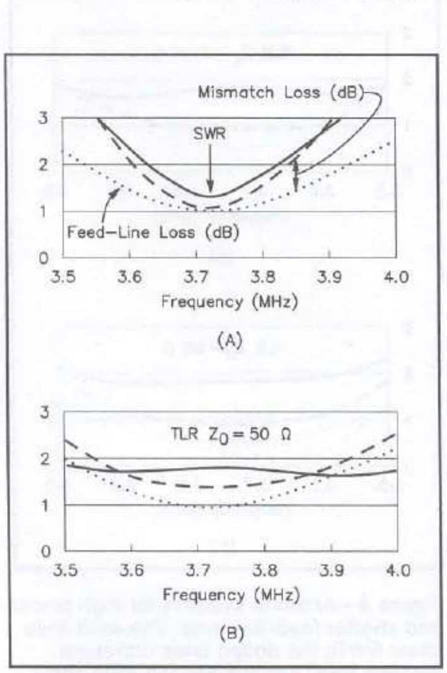

Figure 5—Antenna systems for low power and shorter feed-line runs. The solid lines show SWR; the dotted lines represent feedline loss; and the dashed lines show feed-line loss plus mismatch loss. At A, the feed line is RG-58; at B, it's a 1/4-λ section of RG-59 followed by 1/2 λ of RG-58. The total feedline length in each case is 129.8 feet.

Three broadband antenna systems are shown in Figure 4. All of these are candidates for applications requiring shorter feedline lengths. Figures 4B and 4C show the performance realized when coax cables are paralleled to achieve a low equivalent characteristic impedance. Figure 3B, which results from a 1-wavelength RG-213 transmission-line resonator, and Figure 4B, are very similar. The latter system uses the same amount of cable, but it's cut in half and parallel-connected. This will become clear in the sidebar, "How It Works." The configuration in Figure 4D is attractive because of its simplicity.

Lower-power applications without long feed-line runs can use RG-58 and RG-59 coax. Figure 5B shows how excellent broadbanding is achieved with a remarkably simple feed line. Again, no loss penalty results from the broadbanding.

Adjusting the Broadband Antenna System

The antenna system is easy to build and adjust. First calculate the lengths (in feet) of the transmission-line segments:

$$L_{\text{quarter}} = \frac{245.9\,\text{VF}}{F_0} \qquad \text{(Eq 1)}$$

$$L_{\text{half}} = \frac{491.8\,\text{VF}}{F_0} \qquad \text{(Eq 2)}$$

$$L_{\text{full}} = \frac{983.6\,\text{VF}}{F_0} \qquad \text{(Eq 3)}$$

where

L_{quarter}=length of quarter-wave segment

L_{half} = length of half-wave segment

L_{full} = length of full-wave segment

VF = velocity factor

F_0 = center frequency in MHz

A good starting point for the dipole wire length (in feet) is:

$$L_{\text{dipole}} = \frac{467}{F_0} \qquad \text{(Eq 4)}$$

For the 80-meter application, I suggest using an F_0 of 3.75 MHz. It's a good idea to cut the wires so that the overall length is 4 feet longer than necessary, in case you need to lengthen the wire during tuning. Pass 2 feet of the extra wire through each end insulator and wrap it back around the antenna wire.

To tune the antenna system, you'll change only the dipole and transmission-line-resonator lengths. The best approach is to build the antenna system as I have outlined here and to measure the SWR at the transmitter end of the system. Any tilt or frequency offset in the SWR characteristic can be removed by increasing or decreasing the dipole or transmission-line resonator length. Start by changing the length of the dipole. To improve the SWR at the high end of the band, the dipole must be shortened; to improve the SWR at the low end of the band, the dipole must be lengthened. Progressively add or subtract 6 inches from both legs of the dipole until the SWR curve is symmetrical about the center frequency.

Frequency offset may be required to center the SWR characteristic in the 80-meter band. You can move the entire curve along the frequency axis without causing asymmetry by changing both the dipole and transmission-line resonator lengths using the following equation:

$$L_{\text{New}} = L_{\text{Old}}\left(\frac{3750-\Delta F}{3750}\right) \qquad \text{(Eq 5)}$$

ΔF is the required frequency offset in kilohertz. Shortening the dipole and resonator moves the curve center up in frequency, and lengthening them moves the center down. The length of the quarter-wave segment need not be changed, since the SWR characteristic is not very sensitive to its length.

Lightning Protection

Every antenna system should be designed to minimize the likelihood of a lightning strike. One part of this is keeping all parts of the antenna proper at ground potential. The grounding should be done *outside the shack*, by means of a good ground rod.

I recommend that you install a coaxial lightning protector, which bleeds any static charge from the center conductor, at point B of Figure 1. The protector (and therefore the feed-line shield) should be connected to a high-quality ground rod (the kind electricians use) driven 8 feet into the ground.

Conversion of Existing 80-Meter Dipoles

A study of the cases shown in Figures 3B, 4D and 5B suggests that it's possible to easily convert many existing 80-meter half-wave dipole antennas. Because the most popular way to feed an 80-meter dipole is with a 50-Ω coaxial feed line, the conversion to a broadband antenna system is straightforward. First trim the dipole for resonance at about 3.75 MHz. Then cut the 50-Ω feed line at a multiple of an electrical half-wavelength (at 3.75 MHz) from the antenna. Calculate this length using Eq 2 or Eq 3. Add the 75-Ω quarter-wave section, then complete the run to the shack (if necessary) with 50-Ω coax. Then use the tuning procedure described earlier to optimize the system.

Multiband Operation

Most broadband 80-meter antenna systems are usable only on the 80-meter band, because the broadbanding elements do not allow efficient power transfer on other bands. This is not true with the approach described here, since the structure consists only of a center-fed dipole and a transmission line. Moreover, the transmission-line segments are close to multiples of an electrical half-wavelength near 40 meters and other bands. This opens the possibility for paralleling other half-wave dipoles with the 80-meter dipole and sharing the feed line.

To minimize their interaction, the various dipoles should be spaced from each other away from the feed point. Of course, some interaction will occur and you must tune the multiband system to meet your requirements. I recommend first tuning the 80-meter broadband system and then the next-highest-frequency dipole, and so forth. Only the 80-meter antenna will be broadband, but such broadbanding is not required on the other bands. Figure 6 shows the result of adding a 40-meter dipole to the Figure 1 antenna. Each dipole leg is 34.4 feet long. Note that the SWR on 80 meters changes very little compared to Figure 2. No change was made to the 80-meter dipole or the transmission line.

The multiple-dipole approach described above achieves resonance on several bands and eliminates the need for an antenna tuner on those bands. Of course, if you use an antenna tuner, operation on all HF bands should be possible, but this arrangement is usually not as effective as the multiple-resonance antenna system described here because the feed-line loss is much higher.

Comparison with the Coaxial-Resonator Match

How does the simple broadband dipole described here stack up against other approaches for achieving a good match over the entire 80-meter band? The coaxial resonator match broadband dipole[4,5] represents one of the more efficient designs published to date. It achieves broadband matching at the antenna by the integration of 1/4 wavelength of coaxial cable as a part of the antenna.

Since the coaxial-resonator match achieves a good match at the antenna, the SWR on the feed line is low and the feedline loss is about the same as its matched loss. However, the coaxial cable in the match itself increases the system loss. The net result is that the total loss is about the same with the coaxial-resonator match, but the SWR at the transmitter is lower, never exceeding about 1.6:1 between 3.5 and 4 MHz. Once the SWR is less than 2:1, however, a lower SWR has little value unless you're using a transmitter that significantly reduces power at such SWRs.

Note that the approach described in this article uses a thin wire for the antenna. Most other broadbanding approaches use additional wires or radiators made partly from coaxial cable and are vulnerable to damage from wind and ice loading. Their additional weight and complexity are also limitations.

From the above comparison, the simple broadband antenna system has, by its very simplicity, an edge over the coaxial resonator match, at least in applications where the simpler approach is feasible. Because of the limitations of available coaxial cables, the opportunity for a satisfactory design is constrained. On the other hand, the coaxial resonator match has more adjustment parameters, is useful over a much broader range of applications and yields the lowest SWR over the band.

How It Works

A fundamental way of achieving a broadband match to a resonant dipole antenna involves a parallel-tuned LC network and an appropriate source resistance. In an *RF Design* article,* I described the method for designing such networks, even with lossy resonators. The top of Figure A shows the equivalent circuits of the antenna and matching network. The bottom of Figure A illustrates the corresponding elements in the antenna system.

The role of the resonator is played by the transmission-line segment nearest the antenna. It must be a multiple of an electrical half-wavelength. The quarter-wavelength "Q"-section, made from 75-Ω coax, transforms the 50-Ω transmitter reactance to 112.5 Ω $(75^2/50=112.5)$. I won't go into the design details here; they're the subject of another article, "Broadband Matching Using the Transmission-Line Resonator," in preparation for *The ARRL Antenna Compendium, Volume 4.*

For the structure of Figure A to yield a broadband match, the characteristic impedance of the transmission-line resonator and the transmitter resistance must be within a range of values. Fortunately, commonly used transmission lines, which are available in 50- and 75-Ω characteristic impedances, work well in this application. The broadband systems of Figures 3 through 5 show the usefulness of this approach.

Figure B makes another significant point. For this application, the network parameters of a one-wavelength transmission-line resonator (top) are similar to those of a half-wavelength resonator (bottom) with half the characteristic impedance of the upper resonator. Parallel-connecting two identical cables is a convenient way of achieving lower characteristic impedances. This explains the similarity of Figures 3B and 4B and the similarity of Figures 3C and 4C.—*AI1H*

*F. Witt, "Optimum Lossy Broadband Matching Networks for Resonant Antennas," *RF Design*, Apr 1990, pp 44-51 and Jul 1990, p 10.

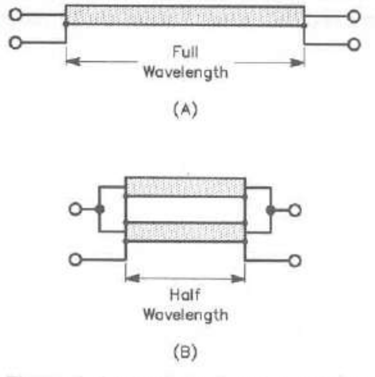

Figure A–Lossy broadband-matching-network equivalent circuit (top), and corresponding simple broadband antenna system elements (bottom).

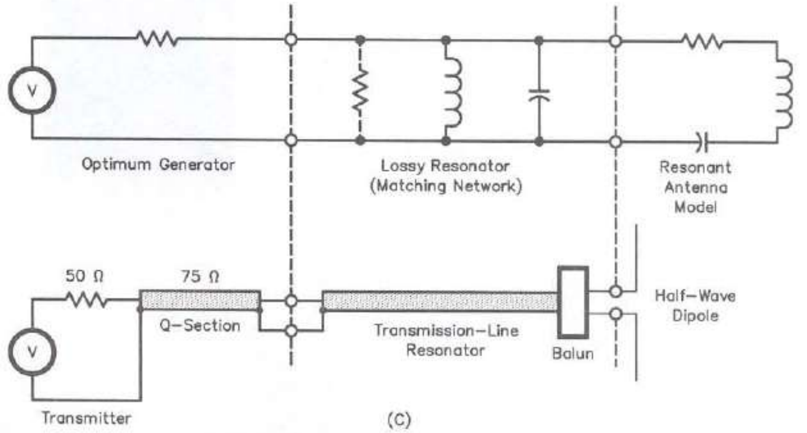

Figure B–These two transmission-line resonators behave essentially the same in this application. The characteristic impedance of each cable segment is the same, making the characteristic impedance of the lower resonator half that of the upper one.

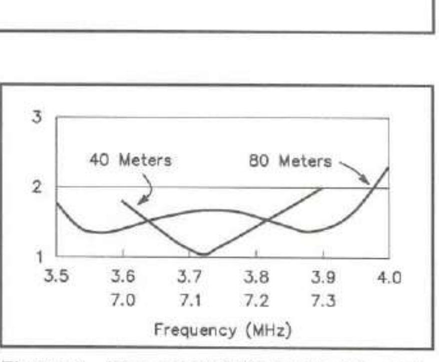

Figure 6—Measured SWR for the 80- and 40-meter multiband antenna system.

Summary

The simple broadbanding technique I've described here capitalizes on the common availability of coaxial cables that fit the application. It overcomes the narrow bandwidth limitations of a conventional 80-meter, half-wave dipole without significant disadvantages. Even parallel dipoles for other bands may be fed with the same feed line.

The limitation of available coaxial cable parameters can be overcome by using the transmission-line resonator as a resonant transformer. Applying this technique is described in an upcoming *ARRL Antenna Compendium* article, "Broadband Matching Using the Transmission-Line Resonator."

This work has benefited from the support and encouragement of my wife, Barbara, N1DIS. Also, I must credit Andrew Griffith, W4ULD, for helping to turn my attention to the approach described here. After reading my *QST* article on match bandwidth of resonant antenna systems,[6] Andy noted that antenna systems should be viewed from their match to a 50-Ω transmitter, even if the feed line does not have a 50-Ω characteristic impedance. He showed examples of the narrowing of match bandwidth to make his point. In my response, published with Andy's letter in *QST*,[7] I pointed out that match bandwidth of an antenna system may actually be *increased* by selecting the right cable length and characteristic impedance. As an example, I showed in Figure 3 of that correspondence the large match bandwidth of a dipole fed with a 5/4-wavelength, 75-Ω RG-11 cable. Note that this is the same case shown in Figure 3C of this article. Thank you, Andy!

Notes

[1] G. Hall, ed, *The ARRL Antenna Book*, 16th ed (Newington: ARRL, 1991), pp 9-1 through 9-12.

[2] M.W. Maxwell, *Reflections: Transmission Lines and Antennas* (Newington: ARRL, 1990), pp 18-1 through 18-6.

[3] See Note 2, p 15-19. I frequency-scaled Walt's data, which is equivalent to changing the wire length. This antenna has a Q of 13 and a resonant resistance of 65-Ω. I took into account the fact that the antenna's radiation resistance increases with frequency.

[4] F. Witt, "The Coaxial Resonator Match and the Broadband Dipole," *QST*, Apr 1989, pp 22-27.

[5] F. Witt, "The Coaxial Resonator Match," *The ARRL Antenna Compendium, Volume 2* (Newington: ARRL, 1989), pp 110-118.

[6] F. Witt, "Match Bandwidth of Resonant Antenna Systems," *QST*, Oct 1991, pp 21-25.

[7] A. Griffith and F. Witt, "Match Bandwidth Revisited," *QST*, Technical Correspondence, Jun 1992, pp 71-72.

From *QST*, June 1985 (Technical Correspondence)

Once More With the 80-Meter Broadband Dipole

With reference to earlier items by K1TD and W7ZOI in *QST*, it *is* possible to design a simple, adjustment-free matching network that will make an 80-meter dipole have a reasonable SWR across the band.[1,2] When using Hayward's RLC model of the dipole, a simple parallel-tuned circuit across the antenna feed point gives less than 2:1 SWR across all but the bottom 10- to 20-kHz of the band, as shown in Figure 1.

This method *does* require a source impedance of 140 ohms, so a broadband matching transformer will be needed at the input. I see no reason why you couldn't just tap the input down on L1 and use L1 as an autotransformer to accomplish the impedance matching function. In fact, although I have not tried it, it may be possible to include the balanced-to-unbalanced transformation in the same coil, as shown in Figure 2. To tune this network, use a dip meter with no antenna connected. Then install the network on the antenna and adjust the input tap for the best compromise SWR across the band.

By choosing a lower characteristic impedance, you can get much better SWR over a narrower bandwidth (Figure 1). Retune both the antenna and the network to center the response on either the phone or CW portion of the band.

You can get better SWR in both the fullband and narrow-band case by adding a series-tuned circuit to the input (Figure 3). The improvement is marginal, however, and complicates the impedance step-up arrangement, so it is probably not worth the effort. It would be better to take Hayward's suggestion and make the antenna itself inherently more broadband. One method is to "fatten" the elements by using a wire cage arrangement or similar. Of course this would change the values of Zo, L1 and C1.

The original graphs for Figures 1 and 3 were drawn with an HP-87. I would be happy to send a copy of the rogram to anyone who sends me an s.a.s.e.—*Alan Bloom, N1AL*

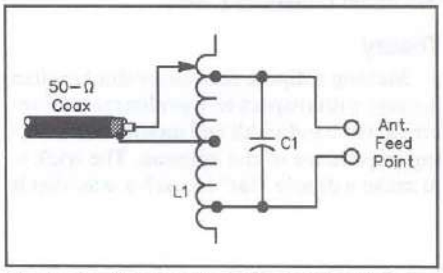

Figure 2—Proposed method of feeding a balanced antenna with unbalanced line. The shielded conductor of the coax line must be connected at the exact center of the coil, which is an electrically neutral point in the antenna system.

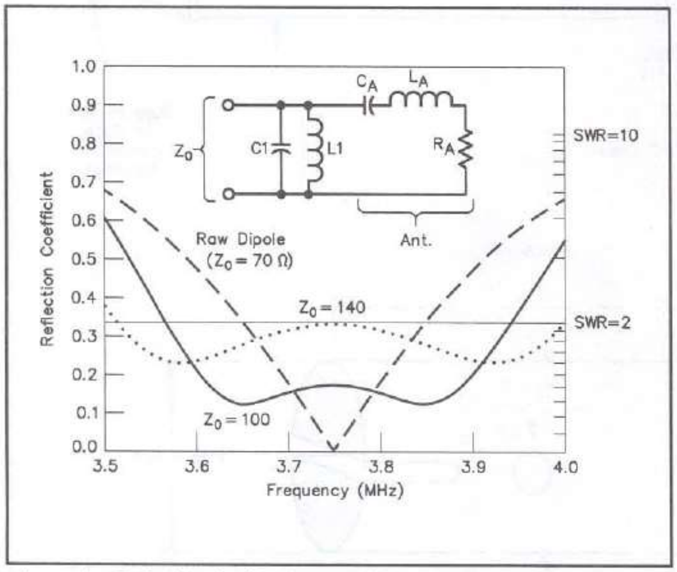

Figure 1—Calculated reflection coefficient and standing-wave ratio vs. frequency for a single-wire dipole and dipoles with lumped constants (C1 and L1) at the feed point.

For the curve where Zo is 140 ohms:
C1 = 2000 pF
L1 = 0.9006 μH

For the curve where Zo is 100 ohms:
C1 = 3500 pF
L1 = 0.5146 μH.

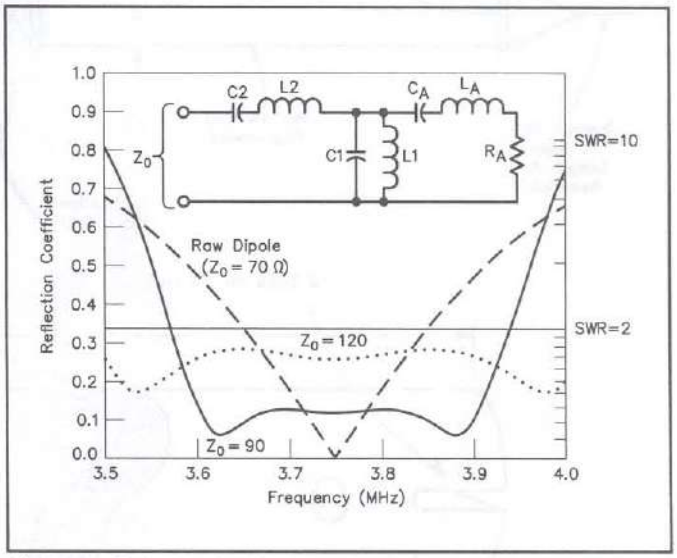

Figure 3—Calculated reflection coefficient and standing-wave ratio vs. frequency for an antenna with shunt and series lumped constants at the feed point (C1-L1 and C2-L2, respectively).

For the curve where Zo equals 120 ohms:
C1–3500 pF
C2–60 pF
L1–0.5146μH
L2–30.0211 μH

For the curve where Zo equals 90 ohms:
C1–7000 pF
C2–60 pF
L1–0.2573 μH
L2–30.0211 μH

---

[1] J. Hall, "The Search for a Simple, Broadband: 80-Meter Dipole," *QST*, April 1983, p. 22.

[2] W. Hayward, "Limitations to Broadband Impedance Matching," Technical Correspondence, *QST*, July 1984, p. 45.

By Robert C. Wilson From *The ARRL Antenna Compendium, Vol 2*

Fat Dipoles

Antennas without problems make radio communications enjoyable. I design overseas radio stations for a living, so I'd rather not have to fight my own ham station when just relaxing and rag chewing. Fat dipoles do the things I want. They match the coax line well over a wide band, and they launch the signal remarkably well.

Theory

Making a dipole conductor thicker than normal with respect to wavelength will increase the bandwidth and modify the working impedance of the antenna. The trick is to make a dipole "fat" in such a way that it may be easily constructed from cheap materials, be highly efficient and at the same time arrange things so that it will match the transmission line from the lower band edge to the upper band edge.

I started with the assumption that my band of interest would be the 80/75-meter band. From end to end, this requires a 13% bandwidth to the 2:1 SWR points for my broadband-solid-state final. I also assumed that this antenna was going to be at a nominal height of 30 feet or 0.11 wavelength above ground. The calculations indicated that a dipole built of four quarter-wavelength no. 14 wires (0.064 inch) with a spacing of 0.0114 wavelength would produce the necessary results. The correct length would have to be 0.45 wavelength to match a 50-ohm line.

Length = 442.5/f feet
Width = 11.25/f feet
Height = 112.5/f feet
where f = center frequency in MHz

Construction

Very few problems will be encountered in building this simple fat dipole if you follow the drawing (Figure 1). First you will need five good insulators. I prefer egg type

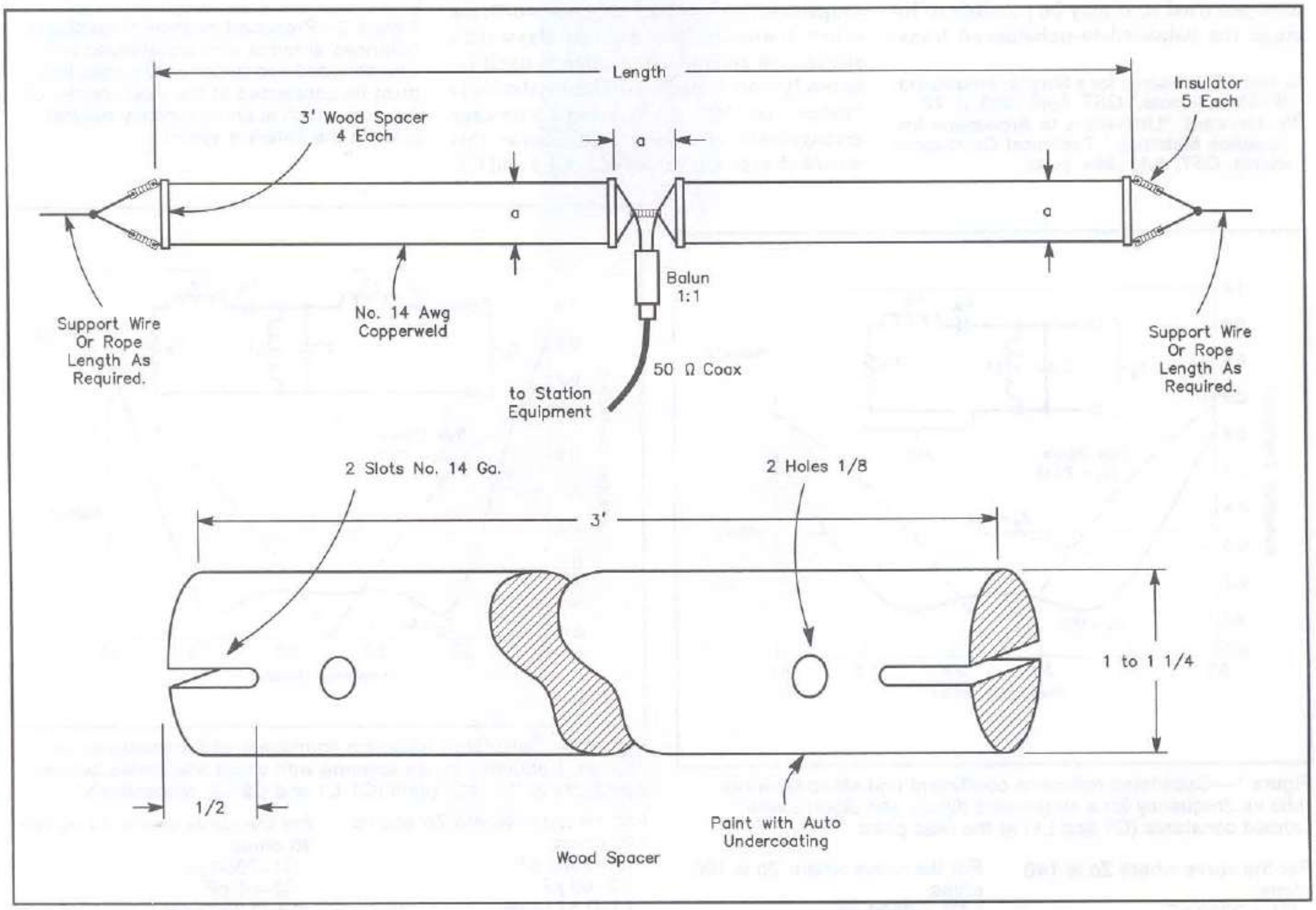
Length
3' Wood Spacer
4 Each
a
Insulator
5 Each
a
a
Balun
1:1
Support Wire
Or Rope
Length As
Required.
No. 14 Awg
Copperweld
50 Ω Coax
to Station
Equipment
Support Wire
Or Rope
Length As
Required.
2 Slots No. 14 Ga.
2 Holes 1/8
3'
1 to 1 1/4
1/2
Paint with Auto
Undercoating
Wood Spacer

Figure 1—The fat dipole with construction details of the spreaders. See Table 1 for the length and spreader dimensions (A).

insulators but there is no critical problem here. One insulator is for the center of the dipole and the others are for the four ends. You will need four 3-foot-long broom sticks or 1-inch wood or plastic rods with good weatherproofing. I painted mine with auto undercoating but outdoor paint or varnish should also work. Copperweld wire is very desirable because it won't stretch and change the tuning of your antenna. This type of wire is available through advertisements in *QST*. The same source may also be able to supply the essential wide-band balun transformer and coax. Either the RG-58 or RG-8 types of coax cable are satisfactory but the latter requires more support because of its greater weight.

Measure your wire carefully and leave enough extra so that the insulators can be attached. The final length will need to be the calculated value, from insulator wire end to insulator wire end. After building the four-wire section, attach dowel rods as shown to act as spreaders. Fasten each rod in place with pieces of wire threaded through the holes and then wrapped around the antenna wire. Wrap these spreader wires tight enough so that the rods will not slip out of place. Snip off projecting wire ends wherever they occur to prevent RF corona power loss. Then, using either wire or rope, make a bridle to hold the ends of the antenna.

Table 1

Dimensions and Bandwidths on Various Bands

Frequency, MHz	*Length, ft*	*Spacing (A), ft*	*Bandwidth, MHz*
1.9	233	5.9	0.252
3.75	118	3	0.5
7.15	62	1.6	0.951
14.175	31.2	0.8	1.885
21.225	20.85	0.53	2.823
28.860	15.34	0.39	3.837

Last, solder the balanced end of the balun transformer to the dipole. Each wire from the balun should go to the pair of wires on the same side of the dipole. The solder job should be of the best quality and permanent because it is hard to repair later. The coax needs to be connected to the unbalanced side of the balun. If you use large-diameter coax (3/8 inch) then think about ways to support the weight. Perhaps a piece of nylon rope from the dipole center insulator to the coax will help take the load, but I'll leave the details of the problem up to you. After this final construction step, haul the antenna up in the trees, using care that no twists are allowed.

Operation

For once I had a 75-meter antenna that worked better than predicted. The SWR was 1.6:1 or better from 3.5 to 4.0 MHz. Better yet, reports received were excellent with my old 100-watt solid-state transceiver. Moving up and down the band gave no loading problems from the broadband final. The fat dipole is just what I needed for a good, relaxing rag chew after a hard day with the 500-kW rig at the office.

CHAPTER TWO

OFF-CENTER-FED-DIPOLES

2-1 **The Off-Center-Fed Dipole Revisited: A Broadband, Multiband Antenna**
From August 1990 *QST*

2-9 **Improved Feed for the Off-Center-Fed Dipole**
From May 1996 *QST* (Technical Correspondence)

2-11 **Off-Center-Fed Dipole Comments, Part 2**
From October 1996 *QST* (Technical Correspondence)

2-13 **Off-Center-Fed Dipole Comments, Part 3**
From October 1996 *QST* (Technical Correspondence)

2-14 **Choke the OCFD**
From September 1997 *QST* (Technical Correspondence)

By John Belrose, VE2CV and Peter Bouliane, VE3KLO **From *QST*, August 1990**

The Off-Center-Fed Dipole Revisited: A Broadband, Multiband Antenna

The search for a simple broadband, multiband antenna with acceptable SWR over the entire 80/75meter band continued with the publication of the article by Witt, AI1H, in April 1989 *QST*. This inspired author Belrose, VE2CV, to reopen his files on the "Windom antenna," or, more properly, the off-center-fed dipole. Particularly since the opening of the 10, 18 and 24-MHz bands, this time-honored, somewhat controversial antenna is attracting a revival in interest and usage because of its multiband characteristics. Thus, the off-center-fed dipole deserves an update. Certainly, most North American radio amateurs have not heard about the German version of this antenna.

Even though it is one of the simplest antennas, the drooping dipole (a dipole with drooping ends, popularly referred to as an inverted V) is effective on 80 meters. For good DX performance, the apex of an 80-meter drooping-dipole antenna should be as high as possible above ground (at least 15 meters [50 feet]), because the antenna's vertical radiation angle decreases with height above ground. The arms of an 80-meter drooping dipole should not be too close to the ground (say, greater than 3 m [10 ft]). The input impedance ($R_a \pm jX_a$) of a drooping dipole depends on the operating frequency, the length of the dipole, the angle between its arms, the dipole's height above ground, and–particularly the resistive component–on the conductivity of the earth beneath the antenna. The included angle between the arms of an 80-meter drooping dipole should be about 127°, since, in this configuration, the antenna's pattern is dipole-like, and its input impedance is about 50 Ω for antenna heights typically employed by radio amateurs. The principal disadvantage of such an antenna is that its SWR bandwidth–less than 200 kHz–is too narrow to cover the 80/75-meter band.

The broadband performance of an antenna can be improved through the use of a matching network at the feed point. This network can comprise discrete components (see Hall,[1] Hayward,[2] Bloom,[3] Hately,[4] and Li, et al[5]); or transmission-line stubs (Snyder,[6] Hansen,[7] and Witt,[8]). Authors Li, et al give a microcomputer program for the design of LCR networks for broadband matching.

Another way to improve broadband performance is to use two dipoles fed in parallel: one dimensioned for the middle of the lower half of the band, and the other for the middle of the upper half of the band. The drooping-dipole configuration is ideal for this arrangement, since the ends of the two dipoles can be fanned (for angular separation) in either the vertical or horizontal plane. An alternative is to use an off-center-fed dipole, which, in addition to broadbandedness, has multiband performance characteristics. This article is concerned with such an antenna.

The "Windom Antenna" and Single-Wire Feed

The original Windom antenna (devised in 1928-29), named after Loren Windom, W8GZ, the amateur who wrote a comprehensive article about it,[9] employed single wire feed at a point of 1/6 to somewhat over 1/7 of the antenna length from the center (see Figure 1A). Windom reported on a detailed experimental study by colleagues John Byrne, W8DKZ, Edward Brooke, W2QZ, Jack Ryder, W8DKJ, and Prof W. L. Everitt of the Ohio State University Dept of Electrical Engineering. They found that if ammeters were placed on the antenna, with the single-wire feeder at the position just described, the current distribution on the dipole was sinusoidal and symmetrical, with no discontinuity in the vicinity of the feed point, and no standing wave on the feeder. Clearly, the feeder was terminated in its characteristic impedance.

This article did not discuss pattern, and it was many years before the computational tools for predicting pattern were available. Parfitt and Griffin[10] have recently analyzed the single-wire-fed dipole (the Windom antenna), and their results show, as anticipated, that radiation from the feed wire does

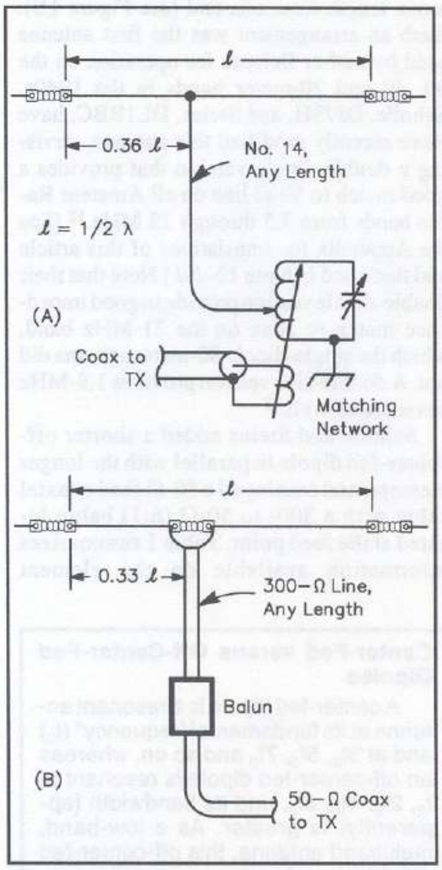

Figure 1—Two versions of the off-center-fed dipole. At A, the original "Windom antenna"–actually a Hertz (1/2 λ) element excited via a single-wire feeder; at B, a dipole off-center-fed via balanced line. The single-wire-fed version dates from the late 1920s; the balanced-line-fed version (sans balun) dates from the 1940s, and the balanced-line-and-balun version appeared in ARRL publications beginning in the late 1950s.

modify the antenna pattern, primarily at low elevation angles with the horizon. The feed line for their modeled antenna was vertical.[11] The pattern was basically dipole-like, with a squint away from the broadside. Current on a wire, whether uniform (a traveling wave) or sinusoidal (a standing wave) leads to radiation, unless the radiation that results from this feedwire current is canceled by an equal amplitude, out-of-phase current on an adjacent parallel wire (as in a balanced transmission line). For the Windom antenna, feeder radiation could be significant, depending on the arrangement and the length of the transmission line.

The Windom antenna was widely used during the 1930s and into the 1940s. For example, in author Belrose's experience, the British Columbia Forest Service were using the single-wire, off-center-fed dipole in the mid-1940s. It was a simple antenna that was easy to tune and match with the π network employed in the output circuit of their transmitter.

The Off-Center-Fed Dipole and Balanced Transmission Line

A later version (1940s) of an off-center fed dipole (*mis*called Windom) employed 300-Ω ribbon feed at a point 1/3 of the antenna length from one end (see Figure 1B). Such an arrangement was the first antenna used by author Belrose for operation on the 80, 40 and 20-meter bands in the 1940s. Scholle, DJ7SH, and Steins, DL1BBC, have more recently modified this antenna, devising a double-dipole version that provides a good match to 50-Ω line on all Amateur Radio bands from 3.5 through 28 MHz.[12] [See the Appendix for translations of this article and that cited in Note 13–*Ed.*] Note that their double-dipole version provides a good impedance match to coax on the 21-MHz band, which the single-dipole 80-meter antenna did not. A double-size version provides 1.8-MHz coverage as well.[13]

Scholle and Steins added a shorter off-center-fed dipole in parallel with the longer element, and employed a 50-Ω feed coaxial cable with a 300- to 50-Ω (6:1) balun located at the feed point. Table 1 summarizes information available on the element lengths for single and double off-center-fed dipoles. Since the dipole must be resonant on harmonic frequencies, its length must be somewhat longer than optimum for the 75-meter band, as noted by Scholle and Steins. It is apparent from their SWR curves that the resonant frequency of their antenna is less than 3.5 MHz.

Table 1

Summary of Element Lengths for Published Off-Center-Fed-Dipole Designs

Figure 1B Version	13.5 m (44.29 ft)	27.94 m (91.67 ft)
cq-DL double-dipole version	13.8 m (45.28 ft) 4.69 m (15.39 ft)	27.7 m (90.88 ft) 9.38 m (30.77 ft)
cq-DL double-sized, double-dipole version	25.88 m (84.91 ft) 4.69 m (15.39 ft)	51.77 m (169.85 ft) 9.38 m (30.77 ft)
Author's double-dipole version, (balanced coaxial-line feed, Figure 4)	23.3 m (76.44 ft) 6.78 m (22.24 ft)	48.2 m (158.13 ft) 13.58 m (44.55 ft)

Center-Fed versus Off-Center-Fed Dipoles

A center-fed dipole is a resonant antenna at its fundamental frequency* (f_0) and at $3f_0$, $5f_0$, $7f_0$ and so on, whereas an off-center-fed dipole is resonant at f_0, $2f_0$, $4f_0$, $8f_0$, and its bandwidth (apparently) is greater. As a low-band, multiband antenna, this off-center-fed dipole has advantages over the center-fed dipole, but a disadvantage (or advantage, depending on the antenna configuration and one's point of view) is that feeder radiation can contribute to the antenna's resultant radiation pattern. The use of a current balun minimizes feeder radiation.—*VE2CV and VE3KLO*

*The frequency at which it is 1/2 λ long.

In the arrangement employed by Scholle and Steins, the longer elements were horizontal and the shorter elements sloped (the included angle between the arms of which was Λ≈100°). The authors' version was a drooping-dipole configuration (Λ≈127°) with the angle between the arms of the two dipoles approximately 45°. This antenna was fed with two 15-meter (50-foot) lengths of RG-62A (foam dielectric) coaxial cable configured as a balanced transmission line.[14]

We fabricated a *single* off-center-fed dipole (dimensionally identical with the longer element of the Scholle and Steins "double dipole") and used it as a Field Day antenna (in 1985) for 75/80 in. The stimulus for this, and for carrying out a detailed study, was a search for a simple all-band antenna conducted for the Canadian National Institute for the Blind.[15] We experimented with 4:1 and 6:1 baluns, primarily because 4:1 baluns are easier to come by (more on baluns later). The antenna was operated in a drooping-dipole configuration (Λ≈127°), with the apex (which, in our case, was the feed point) at about 12 meters (40 feet). (In retrospect, we should have positioned the center of the antenna at apex height, since in this configuration the current maximum on the antenna [for 80-meter operation] would be at maximum height.) Extensive impedance and SWR measurements were made. To facilitate this, we employed a low-loss (foam-dielectric) coaxial feeder consisting of a 32.3-meter (106-foot) length of Belden 8214 (velocity factor, 0.8.) This feeder was 1/2 λ long at 3.5 MHz, 1 λ at 7 MHz, and so on. High-quality commercial baluns (made by Antenna Engineering, Australia[16]) were used between the feed line and the antenna.

Figure 2 shows our SWR measurements for the 80/75-meter band. With the 6:1 balun in line, the 2:1 SWR bandwidth was 3.47 to 3.93 MHz—broadbanded, indeed. The SWR was even lower on 40 and 20 meters, and less than 2:1 for the lower half of the 10-meter band; see Figure 3. Clearly, these bandwidths are not in accord with simple

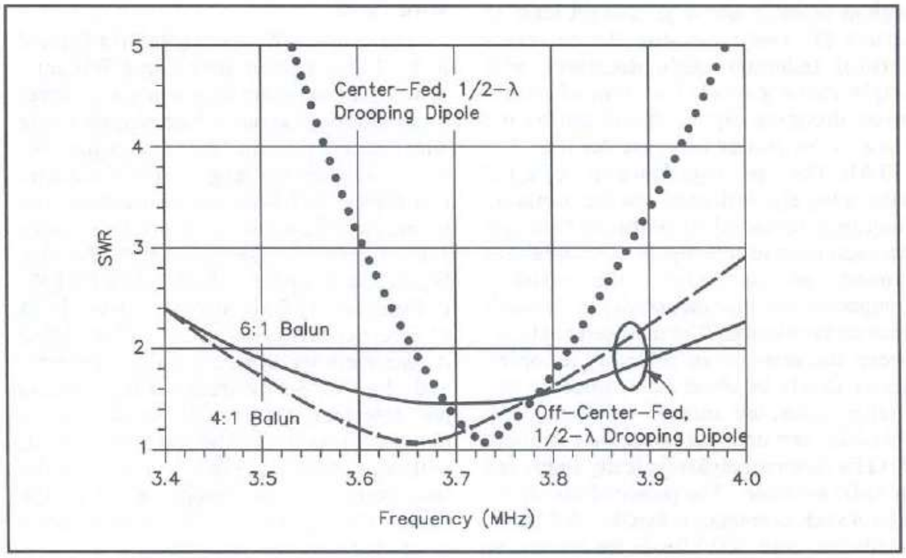

Figure 2—SWR vs frequency for a single off-center-fed dipole employing different baluns (4:1 and 6:1) at the feed point, compared with a center-fed 1/2-λ dipole. Both antennas were operated in a drooping-dipole configuration.

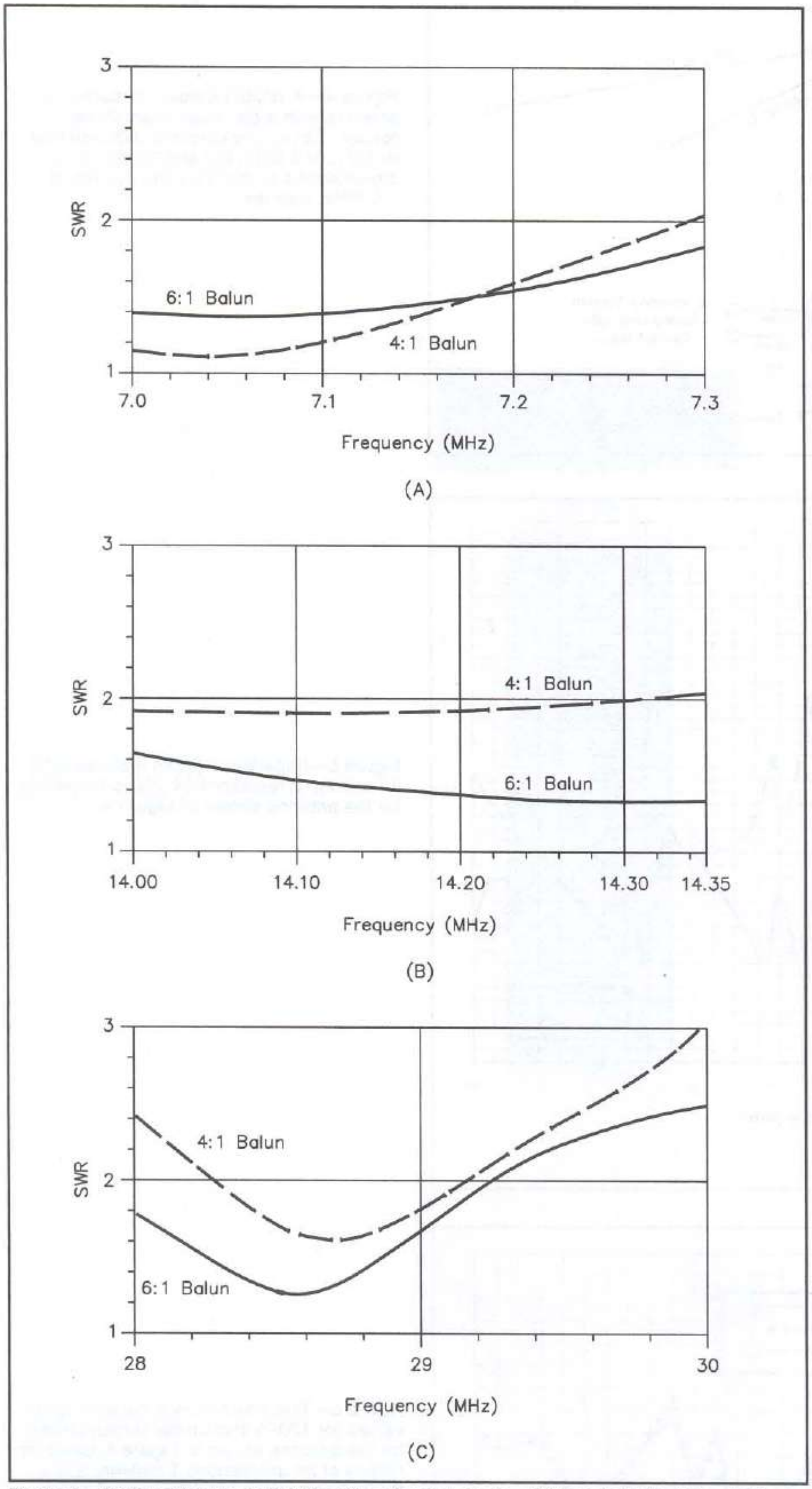

Figure 3—SWR vs frequency for the off-center-fed dipole of Figure 2 at 40, 20 and 10 meters.

theory.[17] Fifteen-meter data are not shown because this single off-center-fed dipole was not resonant on the 15-meter band.

The 4:1 balun provided somewhat sharper resonances and lower SWR at resonance in the 80/75- and 40-meter bands, but a low SWR was not obtainable in the 20-meter band. The measured impedance at resonance in the 80/75-meter band with this balun was exactly 50 Ω; hence the antenna's effective feed-point impedance was 200 Ω. This explains why the minimum SWR with the 6:1 balun was 1.5:1; the 6:1 balun, however, seems to be a better compromise if operation on other bands is wanted.

As noted, we employed high-quality commercial baluns (from the AEA Model 250 series). Radio amateurs have several alternatives in this connection; these include using a 4:1 balun and 75-Ω coax; or purchasing or winding your own 6:1 balun (see Orr, W6SAI[18]),and using 50-Ω coax.

Aside from impedance matching, however, there are other factors to consider in feeding an off-center-fed dipole and selecting a balun for this service. Next, we address some of these aspects.

Balun and Feeder Considerations

In the autumn of 1989, we decided to have another look at off-center-fed dipoles. Because of the multiple-resonant-frequency response of the double off-center-fed dipole, we decided to explore the potential of this antenna system for broadband frequency coverage. If used with a suitable antenna-system tuning unit (ASTU), a double off-center-fed-dipole system could perhaps be used on any frequency from 1.8 to 30 MHz. Since conventional baluns do not perform satisfactorily into reactive loads, we decided to eliminate the balun, at least insofar as the antenna and its feeder were concerned. We fed our antenna with a balanced 190-Ω transmission line consisting of two 15-meter (50-foot) lengths of RG-62A (foam-dielectric) coaxial cable as shown in Figure 4 and described in the work cited in Note 15. Teflon-dielectric cable would have been preferable because of Teflon's superior insulating characteristics but 95-Ω, Teflon-dielectric coax is unavailable.

We started out with a single dipole element dimensioned to 1/2 λ at 2 MHz. This antenna exhibited an antiresonant response (very high input impedance, measured at the transmitter end of the transmission line) at 7.6 MHz. Next, we installed the shorter dipole and dimensioned it to *minimize* the system's impedance at 7.6 MHz. Figure 5 shows the system's impedance v frequency. Except for a narrow band of frequencies near 6 MHz, the input resistance fell in the range of 20 to 400 Ω; and the input reactance fell in the range $+j100$ to $-j200$ Ω. This double off-center-fed dipole system is rather easily tuned and matched, since its reactance is low.

Figure 6 graphs antenna currents for a transmitter output power of 100 W. The antenna system was tuned using an unbalanced T network, with a ferrite-bead current balun of the type described by Maxwell[19] between the ASTU and the antenna. A balanced T network patterned after that described by Belrose was initially used since this is the type of ASTU that, in principle, should be used to tune and match a balanced antenna system. Unfortunately, however, this tuner employed a standard toroidal-core balun—a *voltage* balun, which applies almost-equal *voltages* to the wires of a balanced feeder. Even though the balun was on the "tuned side" of the ASTU—a practice we recommend—and the ASTU it-

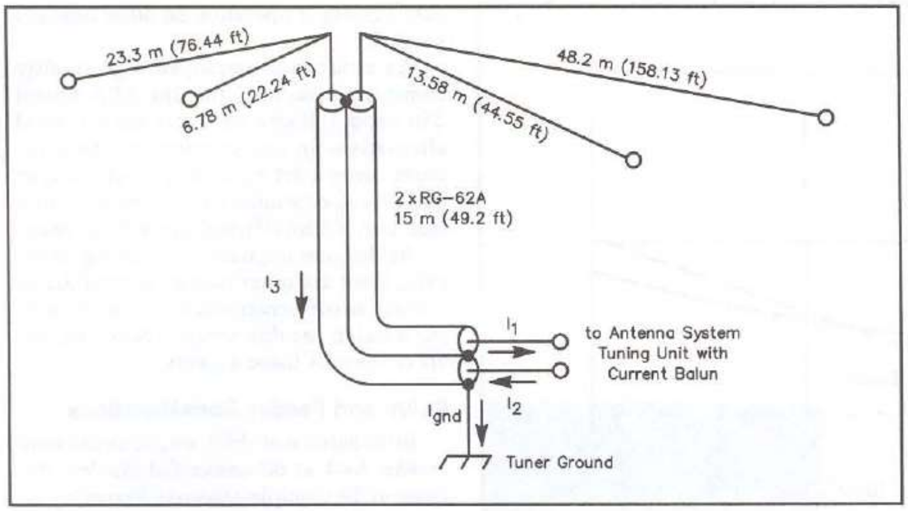
23.3 m (76.44 ft)
6.78 m (22.24 ft)
13.58 m (44.55 ft)
48.2 m (158.13 ft)
2 x RG–62A
15 m (49.2 ft)
I3
I1
I2
Ignd
to Antenna System
Tuning Unit with
Current Balun
Tuner Ground

Figure 4—A double-dipole, off-center-fed antenna with a balanced coaxial-line feeder. The longer dipole is dimensioned to 1/2 λ at 2 MHz; the shorter dipole is dimensioned to minimize the system at 7.6 MHz. See text.

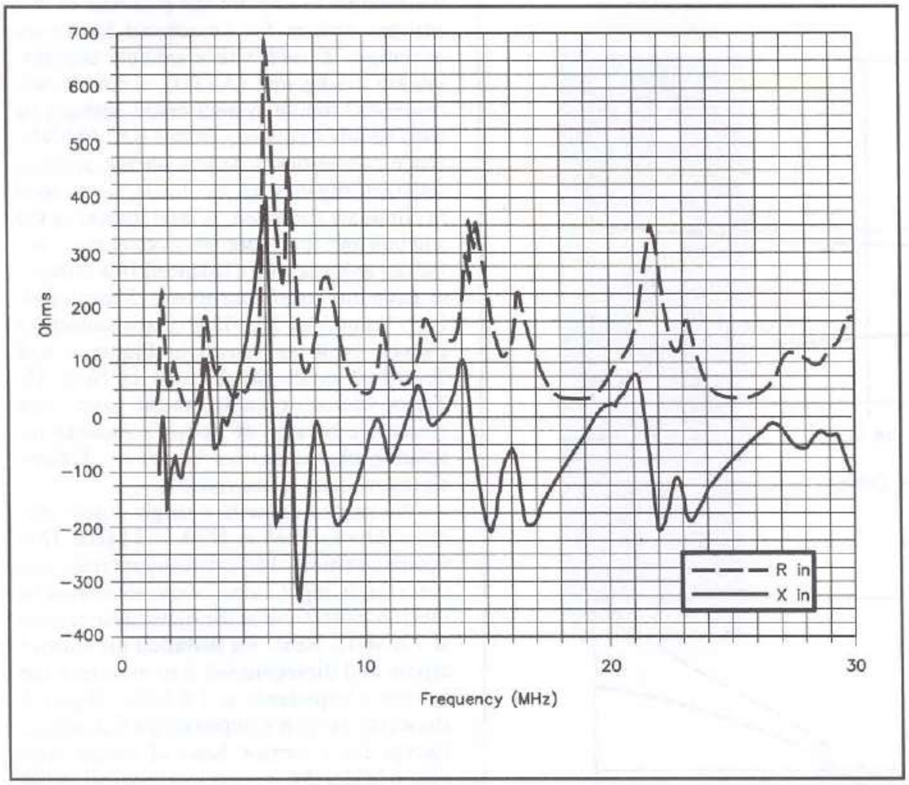

Figure 5—Impedance (input resistance [R in] and input reactance [X in]) vs frequency for the antenna shown in Figure 4.

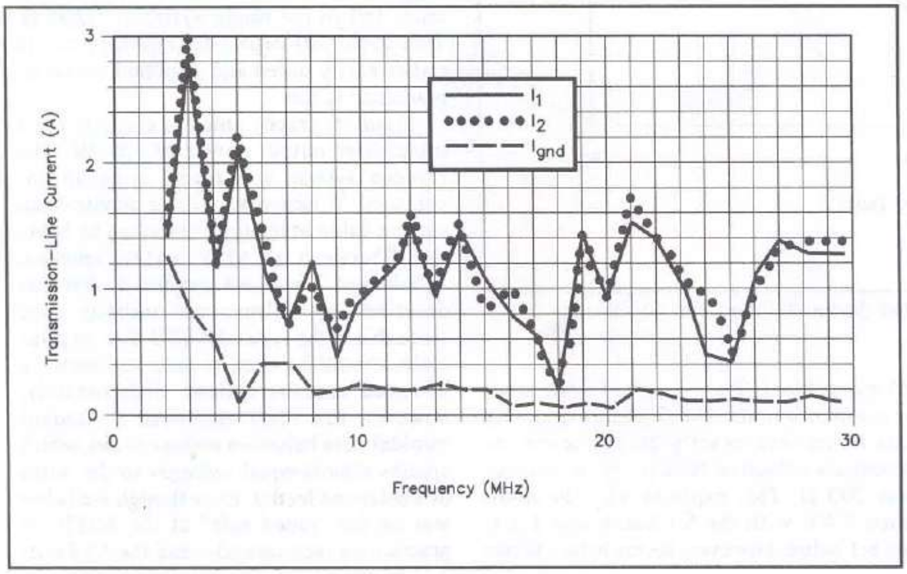

Figure 6—Transmission-line currents (peak values for 100-W transmitter output power) for the antenna shown in Figure 4, tuned by means of an unbalanced T network and a ferrite-bead-choke (W2DU) current balun.

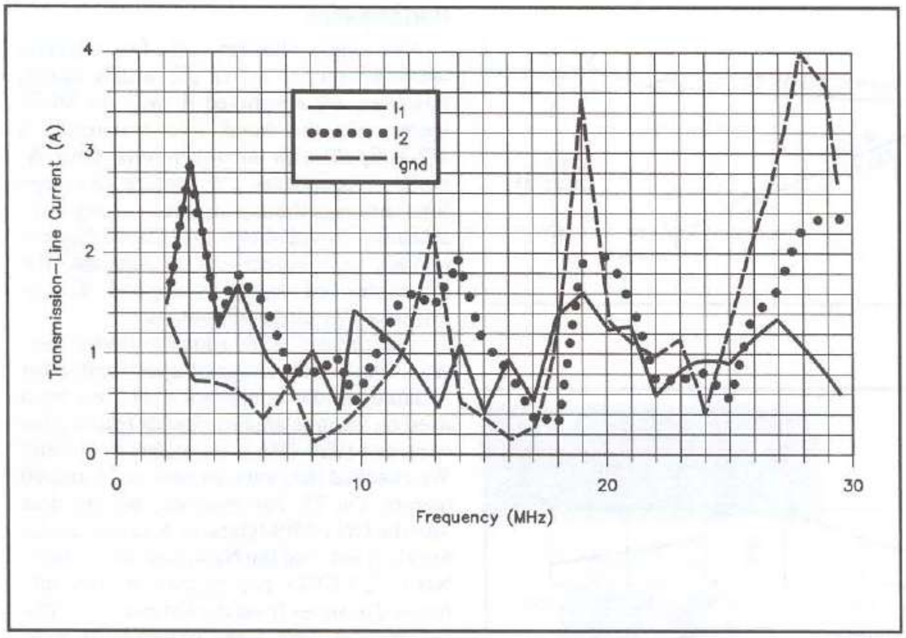

Figure 7—Transmission-line currents (peak values for 100-W transmitter output power) for the Figure 4 antenna, tuned by VE2CV's balanced tuner, which employed a voltage balun.

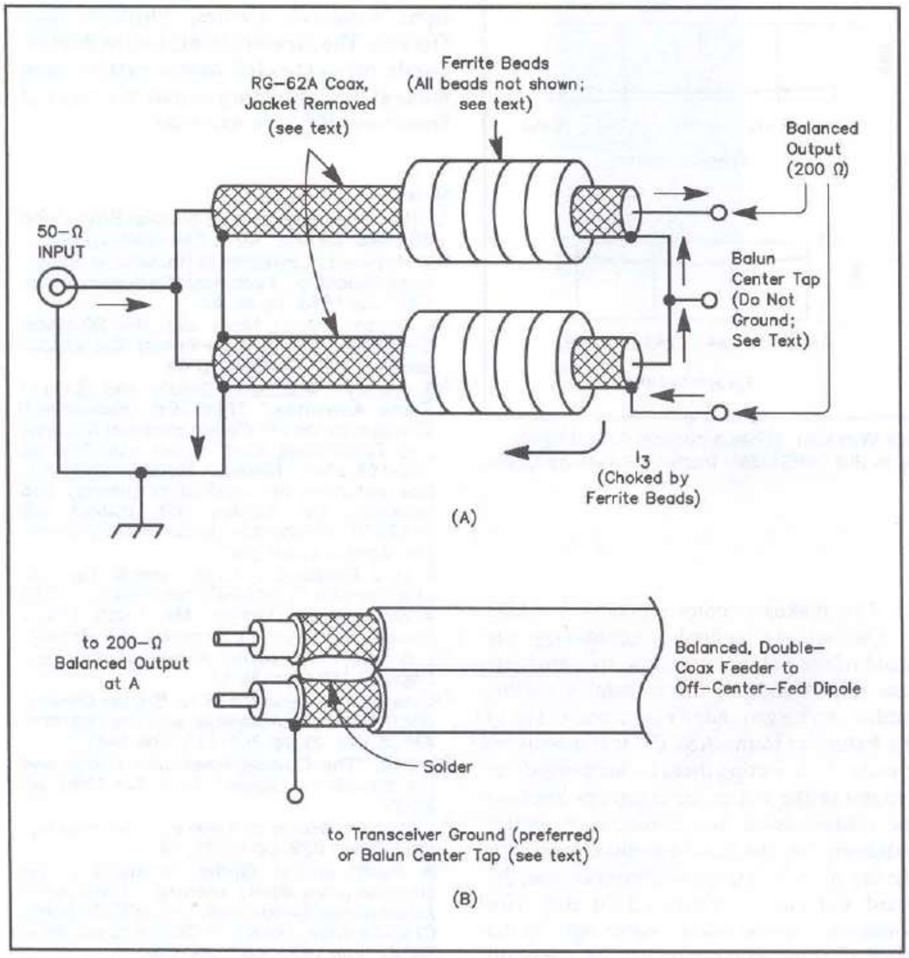

Figure 8—At A, a 4:1, ferrite-bead current balun; at B, connections to the balun end of the dual-coax-balanced feed line.

self was balanced, we found that the transmission-line currents (I_1 and I_2 in Figure 4) differed greatly in amplitude at some frequencies. The difference may surprise you; it did us (see Figure 7).

For an antenna that is asymmetrical with respect to its feeder–such as an off-center-fed dipole–a *current* balun must be used, since this type of balun forces almost equal *currents* into each conductor of the balanced line. This is necessary if transmission-line radiation is to be minimized. If these currents are exactly equal, there will be no difference in current flowing in the ground lead (I_{gnd}), which connects the braids of the coax to the tuner ground. Any current on this lead is then due to radiation-coupled current (I_3) induced on the *outside* surface of the coaxial shield. Clearly, we want this current to be small (to minimize radiation from the transmission line), and indeed it *is* small except at frequencies below 3 MHz (see Figure 6). (Our antenna was suspended from a bracket at the 12-meter [40-foot] level on a 21-meter [70-foot] aluminum-lattice mast. In the final analysis, reradiation by this mast should be considered because current may have been induced on the mast surface.)

Off-Center-Fed Dipoles for the Amateur Radio Experimenter

The off-center-fed dipole used by author Belrose in the 1940s was fed with 300-Ω twin lead via a balanced ASTU that was link coupled to a balanced (push-pull) power amplifier. There was no concern about balun losses and what type of balun to use because the system contained no balun.[20]

The authors' double-dipole, off-center-fed antenna employed a balanced coaxial feed line. This antenna was attractive for the authors because it could be used throughout the HF range if fed via a suitable matching network. If you decide to fabricate and try an off-center-fed dipole system, we suggested that you dimension it in accord with the Scholle and Steins versions, which are optimized for the amateur bands. Furthermore, we suggest that you feed such an antenna with a balanced transmission line. We have used dual RG-62A cables to make a balanced 190-Ω line, but one could use paralleled RG-63 (125-Ω) coaxial cable, which would make a balanced transmission line more in accord ($Z_o \approx 250$ Ω) with the traditional 300-Ω-twin-lead feeder. In such a system, the balun, and the ASTU (if required), can be in the shack to allow experimentation in achieving balanced current feed and reducing losses in the balun. Whatever method you use, feed your off-center-fed antenna via a current balun.

A shield-choke current balun can be constructed by slipping ferrite beads (–43 or –73 material) over a length of coaxial cable. (Depending on the beads obtainable and the diameter of the cable you use, you may need to remove the cable's outer jacket and install the beads directly around the

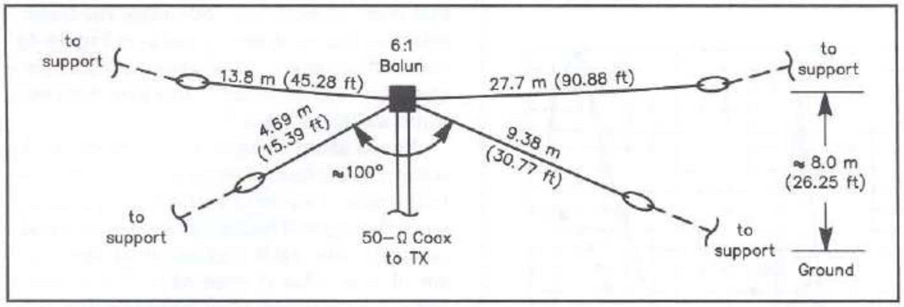

Figure A—Double Windom.

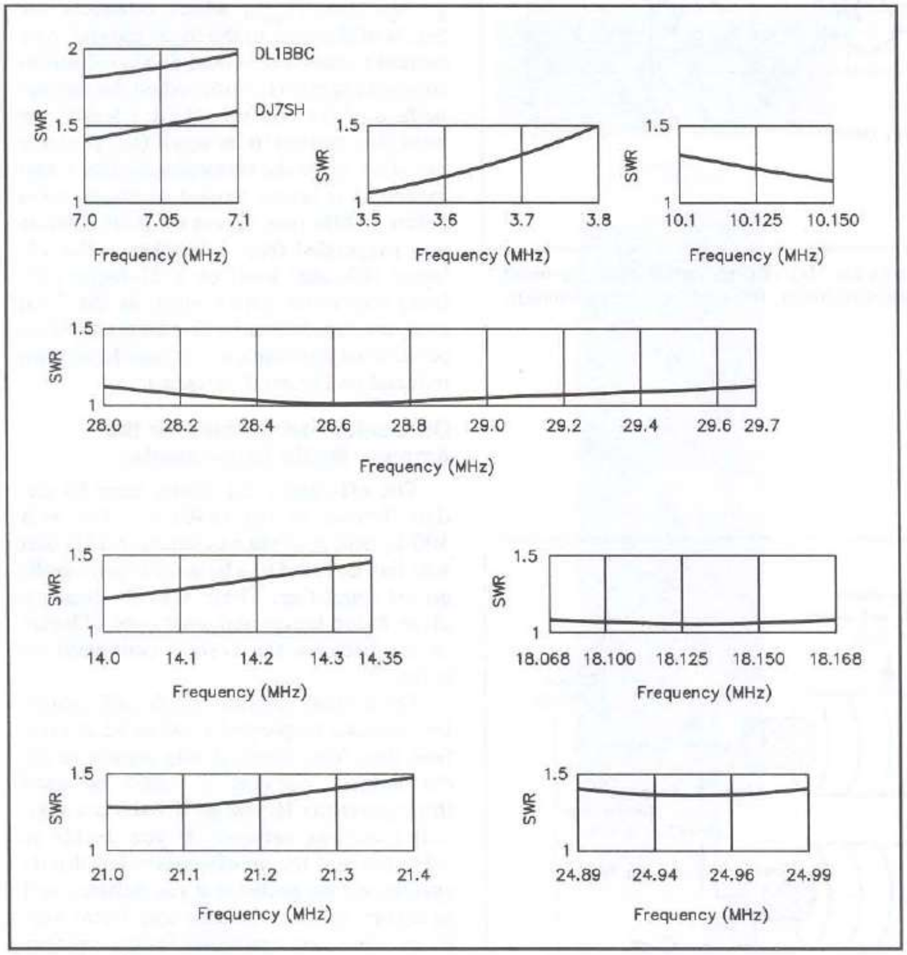

Figure B—SWR curves for the eight-band double Windom. [These curves do not cover all US amateur frequencies because allocations in the FRG differ from those in the US—*Ed.*]

shield.) As described in the work cited at Note 19, Walt Maxwell, W2DU, used 50 beads of no. 73 ferrite (Amidon no. FB-73-2401, Palomar FB-24-73 or equiv) on a piece of Teflon-dielectric cable to make a practical, low-loss, 1.8-to-30-MHz balun about 12 inches long.[21] Such a balun, however, is a 1:1 transformer—not very useful for the present application, where impedance transformation is also required. A 4:1 balun of similar type (see Note 20) can be constructed by using two equal lengths of RG-62 (95-Ω) coaxial cable, each fitted with ferrite beads. The inputs to these two coaxial-cable baluns are connected in parallel, and the outputs in series (see Figure 8). This makes a center-tapped 4:1 balun.

Our antenna feeder has a center tap: the braid of the balanced coaxial transmission line. The braids of the transmission-line cables can be grounded to the center tap of the balun, or to the ASTU (or transceiver) ground. Connecting them to the transceiver ground is the better arrangement because the transmission-line currents are better balanced, and the braid current is less. (For the braid-to-ASTU-ground connection, the braid current is measured in the wire connection to the balun center tap; in the braid-to-transceiver-ground case, the braid current is measured in the connection to the equipment ground.)

Performance

Our single-element Field Day antenna worked well for us; as previously noted, however, we employed it only on 80/75 meters. On this band, it is essentially a 1/2-λ dipole with unconventional feed. At higher frequencies, the pattern develops lobes because the dipole acts as a long-wire antenna. Provided that this directivity coincides with directions of interest, the off-center-fed dipole is a good, simple broadband/multiband antenna.

The authors' 1989 (double-dipole) version, while not designed specifically for Amateur Radio communication, has been used on various amateur bands (during the winter of 1989-90). It works, but how well? We checked into various nets on 75 and 40 meters. On 75, for instance, we checked into the ONTARS (Ontario Amateur Radio Service) net, and the Newfoundland Phone Net (1730 EST), two regions at quite different distances from the Ottawa area. The reports received were comparable with those given to other stations by net control. On the 160-meter band, we found that we could work stations we could hear, provided that they were running comparable power (100 W) and their local noise levels were reasonable. We have not yet determined the antenna's gain and pattern.

Acknowledgment

The authors work at the Communications Research Centre, Shirleys Bay, Ontario. The measurements for the double-dipole off-center-fed dipole system were made at the Laboratory as part of a study of broadband HF wire antennas.

Notes

[1] J. Hall, "The Search for a Simple, Broadband 80-Meter Dipole," *QST*, Apr 1983, pp 22-27.

[2] W. Hayward, "Limitation to Broadband Impedance Matching," Technical Correspondence, *QST*, Jul 1984, pp 45-47.

[3] A. Bloom, "Once More with the 80-Meter Broadband Dipole," Technical Correspondence, *QST*, Jun 1985, p 42.

[4] M. Hately, "Multiband Dipole and Ground Plane Antennas," *Third IEE International Conference on HF Communication Systems and Techniques*, Conf Series Pub 245, pp 102-106,1985. Maurice Hately, GM3HAT, has patented his method of feeding and matching HF dipoles (UK Patent GB 2112579), and he manufactures and markets his "dipoles of delight."

[5] S. Li, J. Rockwell, J. Logan, and D. Tan, *Microcomputer Tools for Communications* (610 Washington St, Dedham, MA 02026: Artech House, 1983), Ch 14: Broadband Matching.

[6] R. Snyder, "The Snyder Antenna," *rf design*, Sep/Oct 1984, pp 49-51.

[7] R. Hansen, "Evaluation of the Snyder Dipole," *IEEE Trans on Antennas and Propagation*, AP-35 (No. 2), pp 207-210, Feb 1987.

[8] F. Witt, "The Coaxial Resonator Match and the Broadband Dipole," *QST*, Apr 1989, pp 22-27.

[9] L. Windom, "Notes on Ethereal Adornments," *QST*, Sep 1929, pp 19-22, 84.

[10] A. Parfitt and D. Griffin, "Analysis of the singlewire-fed dipole antenna," *IEEE AP-S International Symposium, Vol III* (IEEE cat no. CH2654-2/89, Library of Congress no. 89—84327, Jun 1989, pp 1344-1347.

[11] Such an arrangement is an extreme case because the feeder is perpendicular to the

ground—a configuration probably atypical of singlewire feed in Amateur Radio installations, and one that affords practically zero cancellation between feeder and ground currents. At the other extreme, one of the amateur systems described in Windom's "Notes on Ethereal Adornments" (see Note 9) employed a *1200-foot-long* (365.8-m) feed wire that likely more nearly approached parallelism with the ground. Clearly, "modeling a Windom" is a fuzzy procedure.–*Ed.*

[12] H. Scholle and R. Steins, "Eine Doppel-Windom Antenne für Acht Bänder," *cq-DL*, Sep 1983, p 427.

[13] H. Scholle, "Eine Doppel-Windom Antenne für Neun Bänder," *cq-DL*, Jul 1984, pp 332-333.

[14] J. Belrose, "Tuning and Constructing Balanced Lines," Technical Correspondence, *QST*, May 1981, p 43.

[15] J. Swail, VE3KF, and W. Loucks, VE3AR, private communications, Apr 1985.

[16] Antenna Engineering, Australia Pty Ltd, PO Box 191, Croydon, Victoria 3136, Australia.

[17] W. Wrigley, "Impedance Characteristics of Harmonic Antennas," *QST*, Feb 1954, pp 10-14.

[18] W. Orr, Ham Radio Techniques, *ham radio*, Jan 1983, pp 68-69.

[19] M. Maxwell, "Some Aspects of the Balun Problem," *QST*, Mar 1983, pp 38-40.

[20] J. Belrose, "The Balun Saga Continued," *QST*, in preparation.

[21] As a less-costly alternative, 10 ≈1-inch-long mix-77 beads—with an inner diameter to pass the coax used—*may* serve as a substitute. (The authors do not have experience with beads of this material, however.) Consult your ferrite-bead supplier on the availability of beads of this type.–*Ed.*

APPENDIX

A Double Windom Antenna for Eight or Nine Bands

This work, which originally appeared as two articles in *cq-DL*, was translated from the German by Dr. George Elliott Tucker, WA5NVI.

Part 1: A Double Windom Antenna for Eight Bands

By Hubert Scholle, DJ7SH, and Rolf Steins, DL1BBC

The asymmetrical dipole antenna developed and described by Windom (W8GZ) in 1929 has been used by many amateurs for many years as the FD4. This has also been the case in Germany.

We discovered in an older periodical (*QRV*) the explanation by F. Spillner (DJ2KY) that this antenna, with the addition of a small one-band Windom for 15 m, can be used as a five-band Windom. After the installation of the additional elements, this antenna worked very well for two years at DL1BBC.

With the opening of new bands (10, 18 and 24 MHz), the thought occurred to try out a new extension of the FD4 to eight bands (3.5 to 29.7 MHz).

What worked for 21 MHz must also be possible for 10 MHz.

So we took off the 21-MHz extension to my antenna and hung two elements of 4.69 and 9.38 m (15.39 and 30.77 ft), respectively, on the FD4 and stretched these downwards from insulators as an inverted V (Figure A).

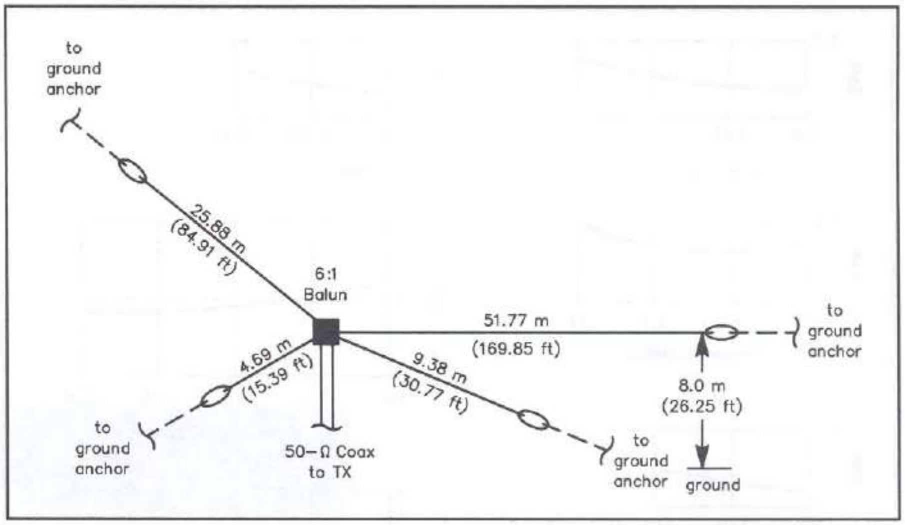

Figure C—A double Windom antenna for nine bands.

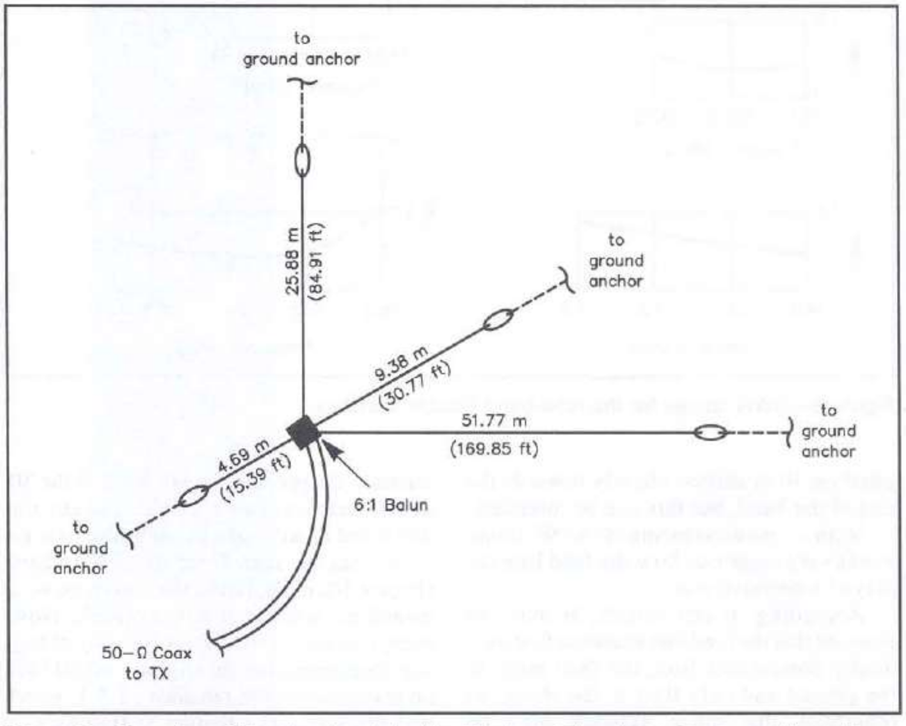

Figure D—Top view of the nine-band double Windom. [The elements are positioned to reduce coupling between the antenna's two off-center-fed dipoles.—*Ed.*]

To calculate the length we used the formula:

$$L/2 = 142.5 \div f \qquad \text{(Eq A)}$$

Whatever would work for 30 m should also work on 15 m.

As suspected, it worked.

As a by-product, it turned out in the measurements that this double Windom resonated just as well on 18 MHz and 24 MHz. So our eight-band Windom came into being with really simple means.

Construction

Thanks to our neighbors, we were able to extend the basic antenna (FD4) to its full length.

At DL1BBC it was installed about 6.9 m (22.63 ft) above the ground, rising to about 8 m (26.25 ft) at each support point. At DJ7SH it hung about 5 m (16.4 ft) above the ground and partly ran over a garage roof. Both extension legs were stretched downwards as an inverted V with an angle of about 100°. Changing this angle allows the whole antenna to be easily tuned during final adjustments.

After construction, the first measurements showed that because of the length of the 30-m elements, the 80, 40 and 20-m bands each had a resonance point that was shifted towards the low end of the band. This effect was eliminated by lengthening slightly the 30-m section, so the resonance points fell more in the middle of the bands.

With this adjustment, the resonance

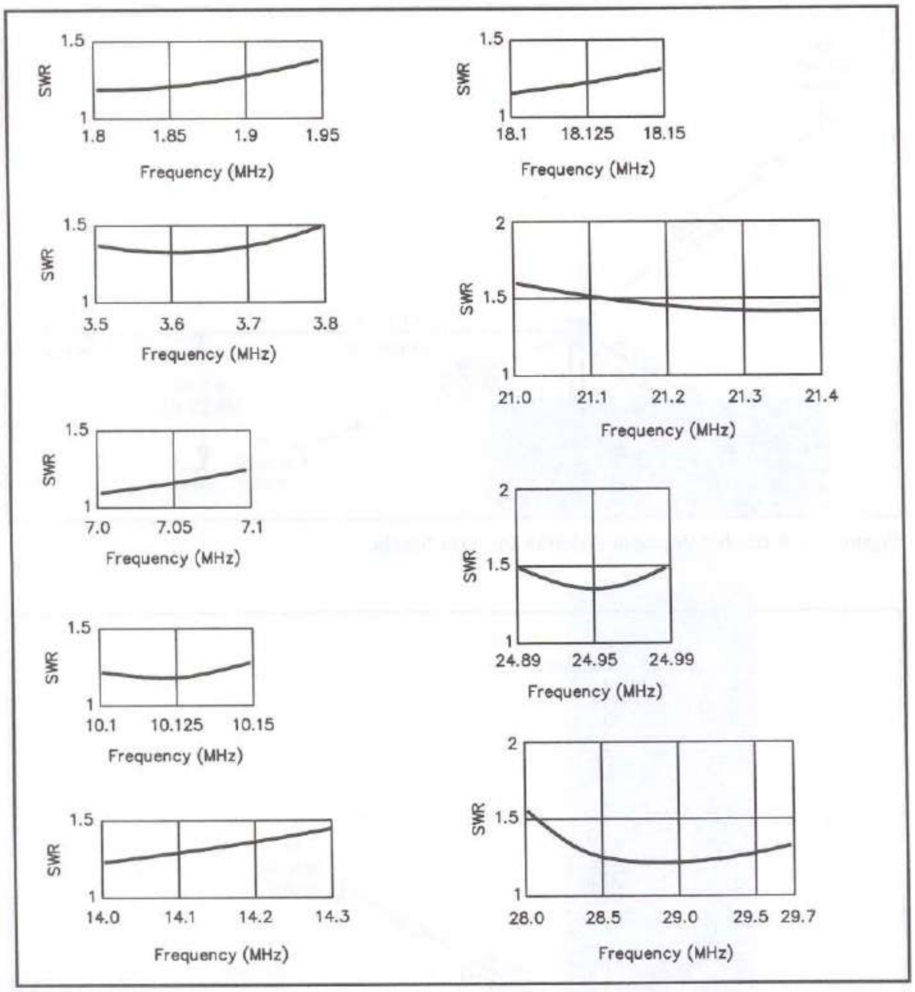

Figure E—SWR curves for the nine-band Double Windom.

point on 30 m shifted slightly towards the end of the band, but this can be tolerated.

With all measurements of the Windom, it was very clear that how the feed line ran played a decisive role.

According to our results, it must be stressed that the feed line must run first vertically downwards from the feed point to the ground and only then to the shack, as otherwise the entire antenna may be detuned. This is especially the case when the height of the antenna is under 10 m (32.8 ft). The 50-ohm-coax feed at DL1BBC was pulled through an old garden hose and then buried under the lawn.

The lower antenna height at DJ7SH had the result that, with the first construction attempt, the precalculated length of the 30-m elements was exactly right. The antenna delivered on all eight bands at the first go.

As can be seen from the SWR charts (Figure B), at DL1BBC the match on 40 m turned out somewhat less favorable. However, this was immediately fixed by changing the antenna height slightly. At DJ7SH, no resonance curve ran above 1.5:1, which was the goal since neither station uses an antenna tuner.

Performance

First contacts were made with both antennas. These showed that the antennas had a good degree of performance for a long wire. Especially the downwards sloping extension elements have a clear advantage over the horizontal basic antenna for DX.

With the first try on 30 m, many contacts were made with the US (East and West coasts), with signal reports between S6 and S7 while running 100 W.

At present, we cannot make a concrete statement about contacts within Europe.

This article makes no scientific claims, but intends to stimulate the long-wire enthusiast, and especially the friends of CW.

Part 2: Adding Another Band

Because the response was unexpectedly great to the publication of the above in *cq-DL*, we went to work again on an extension, as it was worthwhile to add 160 m.

With a half wavelength at 1.835 MHz, we calculated the basic length of the antenna to be 77.65 m (254.75 ft). We tapped the antenna at 25.88 m (84.9 ft) from one end and fed it with 50-ohm coax through a 6:1 balun. The basic antenna of this length was installed horizontally as a reclining L at DL1BBC. The additional elements, with lengths of 4.69 and 9.38 m (15.39 and 30.77 ft), were attached at the balun. This additional Windom for 10 and 21 MHz was again stretched downwards as an inverted V with an angle of about 100°. Here the additional Windom was mounted so that its elements were not extended in the same direction as those of the reclining L, which gave sufficient decoupling (Figures C and D).

For the feed, the Fritzel company made available for testing a new 6:1 balun, series 83, which can also handle high power. The SWR charts (Figure E) were obtained with the wire lengths given in the preceding paragraph. In case builders experience slight resonance shifts, these can be balanced out by lengthening or shortening the additional Windom.

Performance

First contacts were made with the antenna installed at DL1BBC. Here it was once again shown that the antenna has a good degree of performance for a long wire, especially for 1.8 and 3.6 MHz within Europe. The additional Windom again had the degree of performance described in the first part of this article.

The authors welcome questions and exchange of information. (When writing, please include return postage.)

From *QST*, May 1996 (Technical Correspondence)

Improved Feed for the Off-Center-Fed Dipole

By Richard A. Formato, K1POO

Theoretical data suggest that the commonly used feedpoint for the off-center-fed dipole (OCFD) may not be the best. The OCFD is an attractive multiband antenna because it's simple, inexpensive and requires no antenna tuner. Improving its performance simply by moving the feedpoint makes the antenna even more attractive. This note illustrates how the feedpoint influences antenna performance by analyzing computer-modeled SWR data for three different feedpoint locations.

The OCFD (shown schematically in Figure 1), consists of a single wire radiator of length L, fed off center a distance D from one end. The usual implementation uses a "1/3 feed," that is, the RF source is located one-third of the way from the end, so that D ≈ L/3. Why the feedpoint should be located there is not exactly clear. The ninth edition of *The ARRL Antenna Book,*[1] for example, observes that there is not much theoretical justification for this choice. Nevertheless, the 1/3-feed is accepted practice for building an OCFD.

Design details for a 1/3-feed three-band OCFD (80, 40 and 20 meters) appear in the 17th edition of *The ARRL Antenna Book.*[2] A 4:1 current balun at the feedpoint matches this antenna to any length of 50-Ω coax. More recently, Bill Wright, G0FAH, described a four-band, 1/3-feed OCFD (40, 20, 15 and 10 meters) fed with 300-Ω ladder line.[3] Matching 50-Ω coax requires a 4:1 balun on 40, 20 and 10 meters, and a 1:1 balun on 15 meters. Four-band operation, therefore, requires switching baluns. Another minor limitation is that the ladder-line length can be only an odd multiple of the wavelength at 21 MHz because the line is used as an impedance transformer. A simpler approach to achieving four-band operation is to feed the OCFD at a different point along its length.

Table 1

SWR Versus Feedpoint Placement (see Figure 1)

Feedpoint Distance D (m)	*SWR*
40 Meters	
8.65	2.5
6.98	2.1
3.65	2.0
20 Meters	
8.65	2.0
6.98	1.7
3.65	1.7
15 Meters	
8.65	1.2
6.98	21.3
3.65	1.9
10 Meters	
8.65	2.2
6.98	2.2
3.65	2.4

I computer-modeled a 21.03-meter (69-foot) long, 0.2053-cm-diameter (#12 AWG) OCFD in free space. The dimensions are the same as those in the G0FAH design. Free-space results are a good approximation for antennas high enough above the ground (typically a significant fraction of a wavelength). The band-center SWR was computed on 40, 20, 15 and 10 meters at the antenna-input terminals for a feed system impedance of 200 Ω. The theoretical values of input resistance and reactance were used to calculate SWR (the antenna was not assumed to be tuned). Because the feedpoint impedance is 200 Ω, a 4:1 balun is required to feed the antenna with 50-Ω coaxial cable. The results for three different feedpoints appear in Table 1.

For the conventional 1/3-feed (D = 6.98 m), the 40 and 10-meter SWR values are slightly over 2, while the 20-meter SWR is about 1.75. In marked contrast, the 15-meter SWR is off the scale (the actual value > 20). It is this behavior that makes a special feed system necessary on 15 meters, a complication that can be avoided by moving the feedpoint.

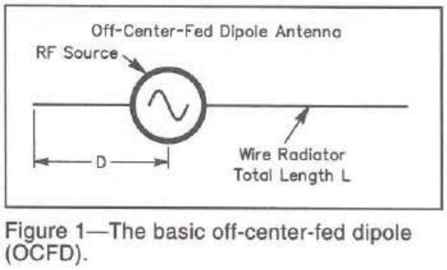

Figure 1—The basic off-center-fed dipole (OCFD).

When the OCFD feed is located 8.65 meters from one end, the 40, 20, and 10-meter SWRs are somewhat higher than they are with the 1/3 feed, but the 15-meter SWR is very low (≈ 1.2). Moving the feedpoint 1.67 meters closer to the antenna's center of the antenna results in a much better average SWR. And, more importantly, special matching is not required to achieve SWR ≤ 2.5 at the antenna terminals on all bands. Balun and coaxial cable losses, which are inevitable, reduce the SWR at the coax input to even lower levels. For most installations, it is probably reasonable to expect SWR at the transmitter to be less than 2 on all bands.

With the feedpoint located 3.65 meters from one end, the SWR on 40, 20 and 15 meters is excellent. The 40-meter SWR is only slightly above 2, and the 20 and 15-meter SWRs are below 2. The highest SWR occurs on 10 meters, where it's approximately 2.4. Because the SWR is reduced by feed-system losses, it will be less than 2.4 at the coax input. And, because balun and cable losses increase with frequency, the SWR reduction will be greatest on 10 meters where it is needed most. Feeding the antenna 3.65 meters from one end may well provide the best overall four-band performance.

In a specific implementation, the OCFD, like any antenna, must be tweaked for optimum SWR. This is accomplished by adjusting the feedpoint location. Other antennas, nearby metallic objects, and the earth are typical factors that influence antenna performance. Since these factors are not included in the computer model, they must be dealt with empirically by adjusting the antenna on-site. The data presented here provide a starting point for experimenting with different feedpoints. Depending on the total antenna length L, height above ground, earth electrical parameters, and feed system Z_0, it should be possible to operate a single OCFD on four or more bands without an antenna tuner or special feed arrangement. It's apparent that the OCFD's SWR varies dramatically as the feedpoint is moved, and that the commonly used $^1/_3$-feed is not necessarily the best. Other feedpoints may therefore produce a better antenna.

---

[1] *The ARRL Antenna Book*, 9th Edition (Newington: ARRL, 1960), pp 191 to 192.

[2] R. Dean Straw, N6BV, Editor, *The ARRL Antenna Book*, 17th Edition (Newington: ARRL, 1994), pp 7-20 to 7-21.

[3] Bill Wright, G0FAH, "Four Bands, Off Center," *QST*, Feb 1996, p 65.

Dean Straw, N6BV, Senior Assistant Technical Editor, comments:

I've modeled what Richard describes and find that he's basically correct in his assertions. However, I would add a couple of caveats:

- The SWR across an individual band will vary. For example, the 40-meter SWR will rise above the level he describes at the ends of the band, because it is a rather large band, percentage-wise. This antenna is no different from an ordinary center-fed 40-meter dipole in that sense.
- The azimuth patterns for this OCFD will not be symmetrical, as is the case for any such off-center-fed antenna. This is nothing new, but should still be mentioned, particularly for those hopeful folks who want a "one-thing-to-all-people" antenna.
- The ham who tries this approach on 30, 17 or 24 meters may burn out the 200:50-Ω balun transformer. The SWR is very high indeed on those bands.

From *QST*, October 1996 (Technical Correspondence)

Off-Center-Fed Dipole Comments, Part 2

By Richard A. Formato, K1POO

In an earlier correspondence,[1] I suggested that the conventional "1/3-feed" used for the off-center-fed dipole (OCFD) is not the best choice. This letter provides additional information, and responds to comments made by Dean Straw, N6BV, at the end of that correspondence. Before addressing the comments, it is important to examine in more detail the SWR data which are the basis of my earlier letter. Figure 1 plots computer-modeled free-space SWR for the prototype 21.03-meter long, 0.0253-cm-diameter OCFD. SWR for a 200-Ω feed system impedance was calculated at the antenna-input terminals every 50 kHz for source frequencies from 5 to 30 MHz. Three feed-point locations were modeled, 3.65, 6.98 (1/3 feed), and 8.65 meters from the end of the antenna; the curves are labeled accordingly.

The most important features of the SWR curves are the locations of minima and maxima, and the corresponding SWR.

Table 1

SWR Minima

3.65 m		*6.98 m (1/3-feed)*		*8.65 m*			
Freq (MHz)	*SWR*	*Freq (MHz)*	*SWR*	*Freq (MHz)*	*SWR*	*Band (Meters)*	*Fc*
6.95	1.27	7.05	2.03	7.05	2.49	40	7.15
14.15	1.70	14.05	1.62	14.00	1.61	20	14.20
21.20	1.91	—	—	21.15	1.14	15	21.20
28.25	1.18	28.30	1.30	28.25	1.41	10	28.85

Table 1 lists minimum SWRs and the frequencies of the minima for the three feed points. For reference, the right-hand columns list the corresponding amateur band and its approximate center frequency.

The main point of my previous correspondence was that the 3.65 and 8.65-meter feeds *might* permit four-band operation without a matching network, because these feed points each result in *four* SWR minima between 5 and 30 MHz (see Figure 1). By contrast, the 6.98-meter SWR curve has only *three* minima in that range. The 1/3-feed antenna cannot possibly operate on more than three bands without some sort of matching, which is why the G0FAH design[2] requires two baluns and a specific transmission line length for 15 meters.

How feasible four-band operation is depends only on where the minimum SWRs occur relative to the ham bands, and how low the SWR is across the bands. Table 1 shows that the prototype OCFD has SWR minima very close to the 40, 20, 15 and 10-meter bands. With some on-site tweaking,

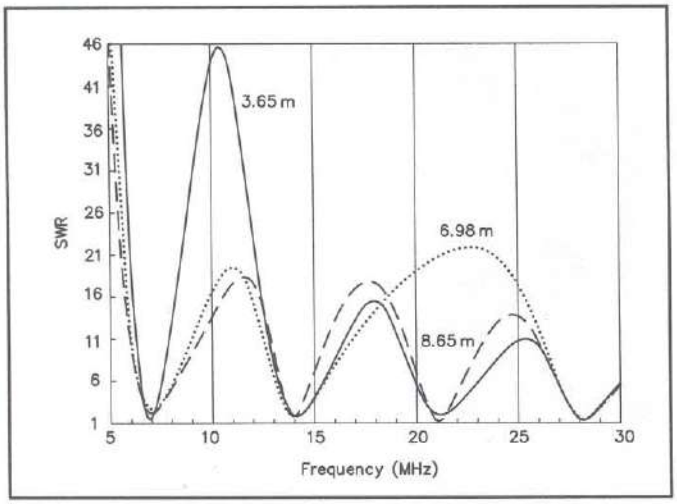

Figure 1—A plot of computer-modeled free-space SWR for the prototype 21.03-meter long, 0.0253-cm-diameter OCFD. SWR for a 200-Ω feed system impedance was calculated at the antenna-input terminals every 50 kHz for source frequencies from 5 to 30 MHz.

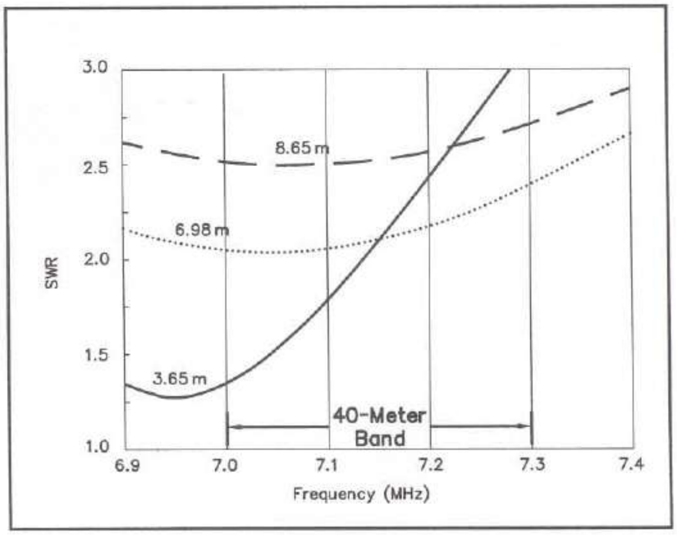

Figure 2—A plot of the prototype OCFD's SWR from 100 kHz below to 100 kHz above the 40-meter band in 5-kHz steps.

---

[1] Richard A. Formato, K1POO, "Improved Feed for the Off-Center-Fed Dipole," *QST*, May 1996, page 76.

[2] Bill Wright, G0FAH, "Four Bands, Off Center," *QST*, Feb 1996, page 65.

it is reasonable to expect that this antenna, fed either 3.65 or 8.65 meters from the end, could provide good SWR across all four amateur bands without a matching network.

The prototype OCFD dimensions were chosen because they correspond to the G0FAH design, thus permitting direct comparison with that antenna, not because they are optimized in any way. In fact, the prototype dimensions are *not* optimum for a four-band OCFD. In free space, the frequencies at which SWR minima occur, and the depth of the minima, are determined by three parameters: radiating-element length and diameter, and feed-point location. Changing any one of these changes both the frequencies of the SWR minima and the minimum SWR values.

Optimizing a four-band, free-space OCFD comes down to determining a set of antenna parameters that produces an acceptably low SWR (typically less than 2) across the 40, 20, 15, and 10-meter bands. Although I have not determined optimum parameter values, the very good predicted performance of the prototype antenna suggests that still better performance is almost certainly achievable. The purpose of my first letter was to encourage experimentation with OCFD designs, which would hopefully advance the state of the art by producing near-optimum designs.

Turning next to the comments at the end of my first correspondence, they are addressed as follows:

SWR Variation Across a Band

In terms of SWR behavior, the OCFD is much different from an ordinary center-fed dipole (CFD). The CFD is intentionally cut (tuned) to place minimum SWR within the band, which is why SWR increases toward the ends of the band (moving away from the minimum). An OCFD may or may not exhibit this behavior, depending on where its SWR minima occur. To illustrate, Figure 2 plots the prototype OCFD's SWR from 100 kHz below to 100 kHz above the 40-meter band in 5-kHz steps. The SWR does *not* increase toward each end of the band. If the SWR minimum is either outside or at one end of the band, as it is in Figure 2, the SWR will increase in one direction (up or down band), but decrease in the other, quite unlike an ordinary CFD. The OCFD's SWR will increase at both band edges *only* when its SWR minimum is inside the band, which generally is not the case.

Azimuth Pattern

The OCFD does indeed have an asymmetrical azimuth pattern, because the radiating element is not symmetrical about the feed point. But feeding the OCFD 8.65 meters from the end provides a *higher* degree of symmetry than the conventional $^1/_3$-feed. Locating the feed 8.65 meters from one end should result in a more symmetrical azimuth pattern than the conventional feed, not less. Of course, feeding the OCFD 3.65 meters from the end increases its asymmetry compared to the $^1/_3$-feed, so that the pattern for this implementation would be expected to be less symmetrical. Even so, pattern asymmetry is not necessarily undesirable. Many operators may want to take advantage of the OCFD's pattern by orienting the antenna to radiate in a preferred direction. This consideration applies to any antenna, even to the single-band CFD, which is an extremely poor radiator in the direction of the antenna axis at low to moderate take-off angles.

Operating the OCFD Out-of-Band

Regardless of where the feed is placed, or how it is implemented, certainly no attempt should be made to operate a 40, 20, 15, 10-meter OCFD on any other band. High SWR conditions may very well result in balun damage.

The data presented here provide additional insight into the advantages of feeding the OCFD at points other than $^1/_3$ of its length from the end. With computer-models for wire antennas widely available, it should be possible to optimize the OCFD in free space and over typical ground so that multiband operation is achievable without a matching network.

From *QST*, October 1996 (Technical Correspondence)

Off-Center-Fed Dipole Comments, Part 3

By Roy O. Hill, Jr, W4PID

QST's May 1996 Technical Correspondence column article "Improved Feed for The Off-Center-Fed Dipole" contained a statement that surprised me: "Why the feed point should be located [$^1/_3$ of the way from the end] is not exactly clear. The ninth edition of *The ARRL Antenna Book*, for example, observed that there is not much theoretical justification for this choice." It seems to me that there is plenty of theoretical justification. I am not an electrical engineer, and I am going to keep this simple. I won't say anything about computer simulations, ground effects, reactance, unbalanced feed-line currents, or adding 15-meter coverage; and I will use the *lowest* frequencies on the 80, 40, and 20-meter bands with full awareness that the frequencies actually used will be higher than that.

Thirty or 40 years ago, it was more or less generally accepted that the impedance of a half-wave antenna was about 4 kΩ at the ends, about 72 Ω at the center, and that the impedance along the antenna could be accurately represented by a straight line on semi-log graph paper. Using these premises, I plotted the impedance along a half-wave 3.5-MHz antenna on the graph of Figure 3. Then I added 7.0 MHz and 14.0 MHz plots for the same antenna.

All three lines cross at two points at about the 280-Ω impedance mark. These two points are $^1/_3$ of the way (within 0.1 percentage point as I measure it) from the ends of the antenna. The antenna can be fed with 300-Ω line at either of the triple-crossing points and it will work on all three frequencies. You could miss the exact crossing point (or change the frequency) somewhat and still have an impedance between 150 and 600 Ω, which would present an SWR less than 2[4]—an allowable value for most transmitters. All this was pretty widely known in the 1950s. What happened to it? Was all the old low-tech knowledge thrown out when the modern high-tech stuff came along?

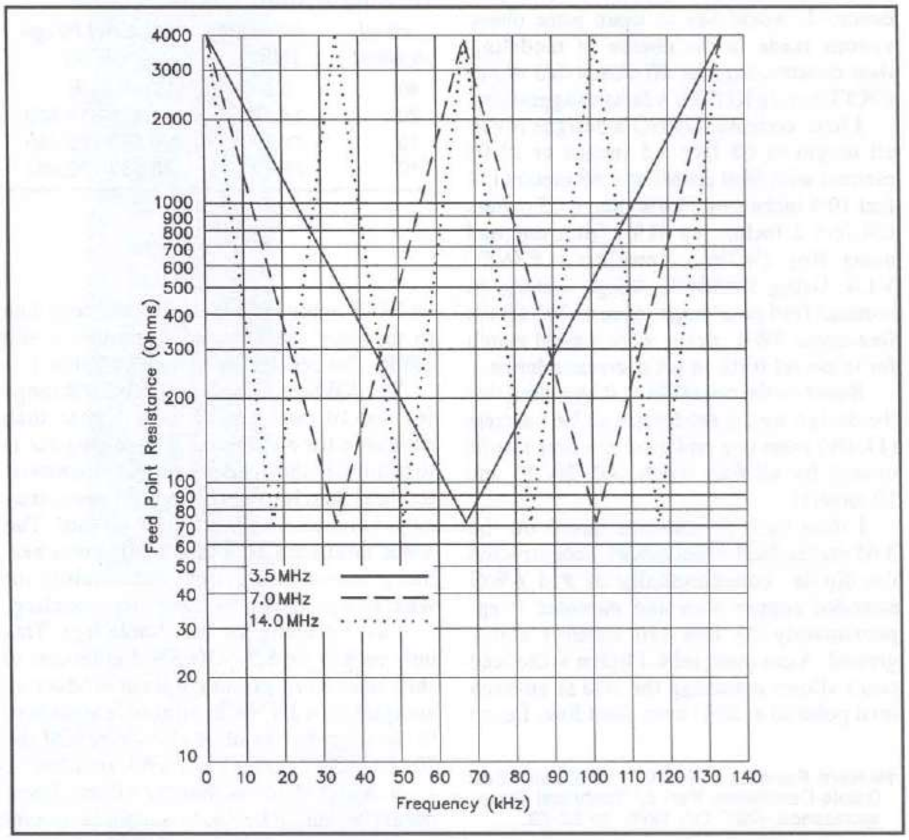

Figure 3—Impedance plot of an OCFD for 3.5, 7 and 14 MHz as a function of distance (in feet) from either end.

Also, the article mentioned that the "$^1/_3$-feed" antenna requires no antenna tuner. That is true, of course, for loading the antenna, but back in the days I've been talking about, the ARRL was stressing that you should *not* use a multiband antenna without an antenna tuner because of the danger of harmonic radiation. Do we not worry about harmonic radiation any more?

A little over 30 years ago, using as a guide a graph like the one shown in Figure 3, but using frequencies more like those on which I would be operating, I put up an antenna fed with 300-Ω ribbon 80 feet from one end and 44 feet from the other. I have used this antenna ever since. I do use an antenna tuner with it. The system works fine and causes no RFI or TVI. I used similar antennas for about 12 years before that, at other locations.

I just couldn't pass up that "not much theoretical justification" quoted from an ARRL publication without comment. The $^1/_3$ feed has, not only plenty of theoretical justification, but plenty of practical justification—from use.

---

[4] This is only accurate if the impedance is purely resistive. With reactive impedances, it is entirely possible to have an SWR higher than 2.—*Zack Lau, KH6CP*

From *QST*, September 1997 (Technical Correspondence)

Choke the OCFD

By Dale Gaudier M0AOP/K4DG

After reading with interest Richard (K1POO) Formato's technical correspondence,[3] I would like to share some observations made in the course of modeling, then constructing, an off-center-fed dipole (OCFD) using K1POO's design suggestions.

I first modeled K1POO's designs (overall length of 68 feet 5.5 inches or 21.03 meters) with feed points at 3.65 meters (11 feet 10.5 inches, or 17.4%) and 8.65 meters (28 feet 2 inches, or 41%) from one end using Roy (W7EL) Lewallen's *EZNEC* V1.0. Using K1POO's design criteria (a nominal feed point impedance of 200 Ω) the free-space SWR curves were a good match for those set forth in his correspondence.

Based on the model data, it appeared that the design with a feed point at 3.65 meters (17.4%) from one end gave the best results overall for all four bands (40, 20, 15 and 10 meters).

I then built an antenna based on the 3.65-meter-feed-point model. I constructed the dipole conventionally of #14 AWG stranded copper wire and mounted it approximately 33 feet (10 meters) above ground. A commercial 4:1 balun at the feed point allows matching the 200 Ω antenna feed point to a 50 Ω coax feed line. I used an SWR analyzer attached to the feed line to measure the antenna's resonance and SWR. The results are shown in Table 1.

Table 1
Results without the RF Choke

Band (meters)	*Minimum (MHz)*	*2:1 SWR Range (MHz)*
40	~8.5	(SWR >>3)
20	14.17	13.250 - 15.500
15	21.58	20.420 - 22.350
10	29.69	28.550 - 30.480

The SWR minimum and 2:1 SWR range for the 10 meter band were higher than desirable for an antenna whose purpose is to minimize the need for a tuner. However, the real puzzle was the lack of resonance anywhere near the 40 meter band! The SWR minimum around 8.5 MHz was neither pronounced nor deep, and certainly not what was predicted by computer modeling.

I tried pruning the two dipole legs. This only caused the 8.5-MHz SWR minimum to shift somewhat. Pruning did not produce an acceptable (< 2:1 SWR) minimum anywhere in the 40 meter band; it also decreased the 20, 15 and 10-meter band SWR minima.

It occurred to me that the effects I was observing might be due to unequal currents flowing at the feed point (possibly due to the asymmetrical design of the OCFD). I made a simple RF choke by coiling several turns of the RG-8X feed line close to the feed point and balun. After some initial adjustments of the choke diameter and number of turns, I had my 40-meter resonance! (See Table 2.) The coil is 5½ turns of RG-8X with an inside diameter of 9½ inches (24 cm). Note that the 40-meter band minimum and 2:1 SWR range can be raised by removing part of a turn from the RF choke.

Table 2
Results with the RF Choke

Band (meters)	*Minimum (MHz)*	*SWR Minima*	*2:1 SWR Range (MHz)*
40	7.00	1.1	6.850 - 7.220
20	14.10	1.2	13.520 - 14.420
15	21.22	1.1	20.420 - 22.340
10	28.34	1.4	27.460 - 29.050

As is apparent, the OCFD design of K1POO is surprisingly broadband, and with a little adjustment, will cover virtually all of the 40, 20, 15 and 10-meter bands with an SWR of 2:1 or less. However, an RF choke appears to be necessary for this design to work on its fundamental frequency.

---

[3] Richard Formato, K1POO, "Off-Center-Fed Dipole Comments, Part 2," Technical Correspondence, *QST*, Oct 1996, pp 72-73.

CHAPTER THREE

MULTIBAND ANTENNAS

3-1 **A Trap Collinear Antenna**
From August 1963 *QST*

3-3 **Antenna Traps of Spiral Delay Line**
From November 1972 *QST*

3-6 **Build a Space-Efficient Dipole Antenna for 40, 80 and 160 Meters**
From July 1992 *QST*

3-8 **Trap Construction Information for Al (W8NX) Buxton's July 1992 Dipole**
From September 1992 *QST* (Technical Correspondence)

3-9 **Two New Multiband Trap Dipoles**
From August 1994 *QST*

3-13 **A Center-Fed "Zepp" for 80 and 40**
From May 1966 *QST*

3-16 **All-Band Antenna**
From March 1967 *QST* (Technical Correspondence)

3-17 **Multiband Antennas Using Decoupling Stubs**
From December 1960 *QST*

3-20 **"All-Band" Antenna**
From December 1954 *QST* (Hints & Kinks)

3-21 **Three-Band Matching System for a Forty-Meter Doublet**
From January 1975 *QST* (Hints & Kinks)

3-22 **A "Z" Antenna for the 10-160 Meter Bands**
From December 1979 *QST* (Technical Correspondence)

3-23 **Compact Multiband Antenna Without Traps**
From November 1981 *QST*

3-25 **Five Bands, No Tuner**
From June 1995 *QST*

3-26 **The NRY: A Simple, Effective Wire Antenna for 80 through 10 Meters**
From March 1993 *QST*

By Wesley M. Bell, W7QB **From *QST*, August 1963**

A Trap Collinear Antenna

Simple 3-Band Radiator with In-Phase Elements

This antenna covers the 15, 20, and 80-meter bands. On the two higher frequency bands the antenna operates with two extended half waves in phase thereby realizing some gain over the dipole operation of a conventional trap antenna.

As your *ARRL Handbook* tells you, broadside gain over a dipole approximately equivalent to doubling transmitting power may be obtained by using a center-fed antenna about 1¼ wavelengths long (extended double Zepp). Advantage of this is taken in the three-band trap antenna shown in Figure 1. The basic antenna is a dipole for 80 meters. The traps isolate sections of approximately 1¼ wavelengths for 20 and 15 meters. Since the center of a 1¼ wavelength wire is not at a current loop, wire is added in the form of a short open-wire feeder to make the total length about 1½ wavelengths, thereby bringing a current loop at the point where the system is fed by coax line. A balun is used to couple the unbalanced line to the balanced antenna system.

Trap Construction

The coil and capacitor specifications given under Figure 1 should be adequate for transmitters running at 100 watts input or less. For higher power, the inductance and capacitance values should be the same, but coils should be wound with heavier conductor, and capacitors should be of the transmitting type, such as the Centralab 850SL type. I made my own coils by wrapping a 2¼-inch form with waxed paper and winding the turns with double strands of No. 18 wire, unwinding one strand and cementing the remaining turns with strips of model-airplane glue. When the glue was dry, the completed coil was slipped off the form. At least one full extra turn should be wound to allow for pruning.

As shown in Figure 2, the capacitor is placed inside the coil, and the terminals of both capacitor and coil soldered to the heads of brass machine screws. These screws serve to hold the assembly central in a plastic waterproof container as shown in Figure 3. The container is the 1-pint size commonly found filled with chip dip, ice cream, potato salad, oysters and whatnot in grocery stores. The screws are fastened in the top cover and bottom of the container with nuts. The container is suspended from an insulator at the appropriate point in the antenna by short lengths of wire dropped from the insulator to the mounting screws where they are secured by a second set of nuts.

After mounting the traps in the containers, they should be resonated to the designated frequencies by carefully pruning the coils while checking with a grid-dip meter. The strain insulator with its wire wraps must be included, as shown in Figure 3, since the insulator capacitance is in parallel with the trap capacitor. (See *ARRL Handbook*.) As resonance is approached, final adjustment can be made by forming what is left of the last turn into a hairpin, and bend-

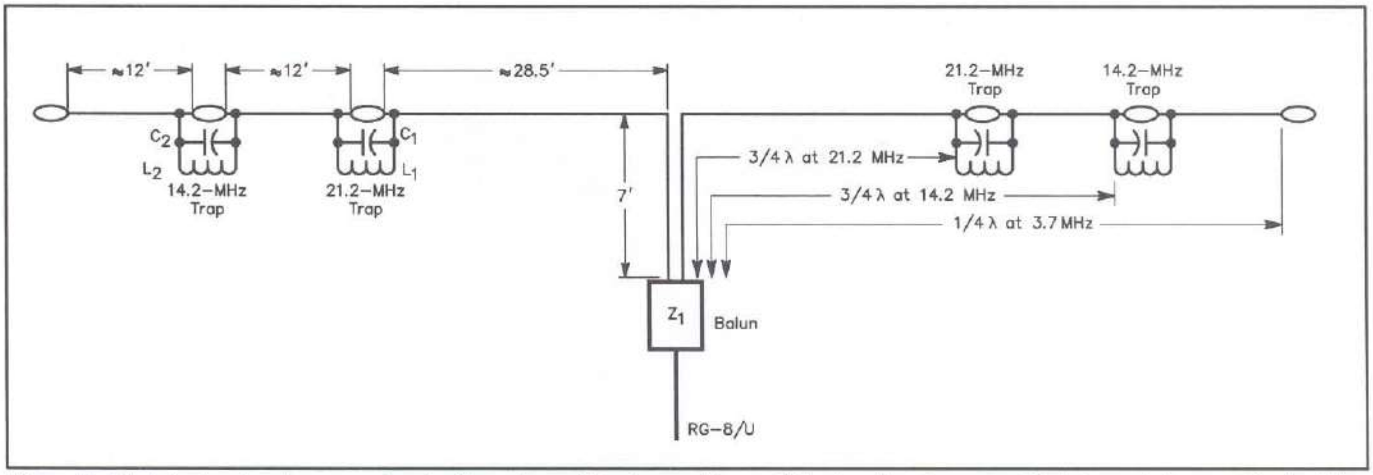

Figure 1—Sketch showing the approximate dimensions of the trap collinear. Wire lengths shown on left side are in terms of feet, while the approximate wavelength equivalents (electrical length) are shown on the right side. Frequencies and approximate wire lengths are for the centers of the three bands as a compromise for full-band coverage. It may be desirable to increase frequency for phone-only operation or lower frequency for c.w.-only operation.

C1, C2—25-pf. 6000-volt disk ceramic. See text.

L_1—Approx. 2 μH–4¾ turns No. 18, 2¼-inches diam., ⅜ inch long, or 5 turns No. 18, 2-inch diam., 16 tpi. See text.

L_2—Approx. 5 μH.—8 turns No. 18, 2¼-inch dia., ⅝ inch long, or 9 turns No. 18, 2-inch diam. 16 tpi. See text.

Z_1—1 to 1 balun.

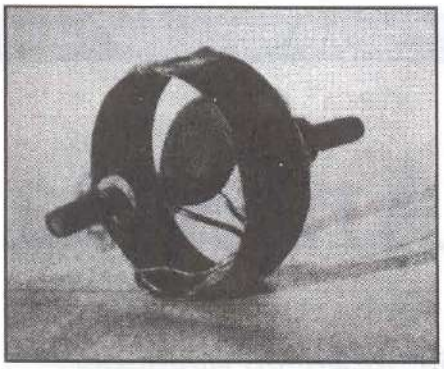

Figure 2—Trap components ready to be mounted in weather-proof container.

ing or twisting the hairpin to alter its inductive relationship to the main part of the coil.

Antenna Adjustment

Antenna resonance can be checked by shorting the ends of the 7-ft. open-wire line and coupling to a grid-dip meter. Initially, the wire lengths should be made a foot or so longer than the lengths shown in Figure 1. Start out with the 21-MHz sections only, anchoring the outer end of each wire section to one side of a 21-MHz trap insulator and connecting it to one side of the trap. An additional insulator should be attached temporarily between the other side of the trap insulator and the antenna-supporting rope. Then gradually shorten the wire until the grid-dip meter shows the desired resonant frequency.

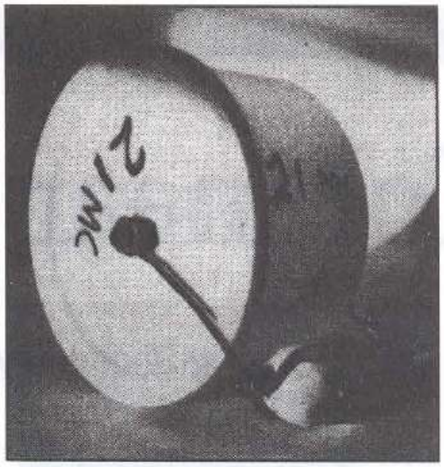

Figure 3—The completed 21-MHz trap and its supporting insulator ready for checking with a grid-dip oscillator.

Then add the second sections and traps and adjust similarly for the desired frequency in the 14-MHz band. The end sections of wire are then added and adjusted to show resonance at he desired frequency in the 3.5-MHz band.

In making the antenna adjustments, do not adjust the traps after they have once been set with the g.d.o.; change only the lengths of the wire sections. Devote plenty of time and patience to the adjustments. The job just can't be done correctly in a few minutes.

I use a center supporting pole, and the balun is enclosed in a weather-proof box mounted at the top of the pole.

Results with this antenna have been good. Using it in " inverted-vee" fashion, with one pole at the center and the ends attached to bushes, fences, clotheslines, or whatever else might be handy, I changed directions by simply walking the ends around to different positions. Without half trying, I worked 44 countries with a DX-60. I'm moving to the country soon where I plan to put up several of these antennas. I hope to add reflectors too. (Well, a guy can dream, can't he?)

By William J. Lattin, W4JRW **From *QST*, November 1972**

Antenna Traps of Spiral Delay Line

Most "traps" used in amateur radio multiband antennas are made of lumped inductance and capacitance in parallel. These consist of inductors made of coil stock of No. 12 or No. 14 wire and ceramic capacitors having voltage ratings up to 15,000 volts dc, which are relatively expensive. Vacuum capacitors would be the best, of course, but are also rather expensive.

Another type of trap has a capacitor made of two pieces of aluminum tubing arranged with a small-diameter tube inside a larger tube. Some have polystyrene dielectric, others air dielectric. The *ARRL Handbook* has a very complete description of these types.

Quarter-wave stubs of transmission line can be used for isolating sections of an antenna.[1] Loading coils can be used to modify the harmonic responses of a doublet to the second, third, fifth, seventh, and so on, to obtain a multiband antenna.[2] However, the use of loading coils is quite complicated if more than two bands of frequencies are desired. Traps tuned to the desired resonant frequencies make it much easier to adjust the lengths of the antenna sections, and also to obtain closer spacing between bands than can be obtained with loading coils.

Since quarter-wave sections of transmission line can be used as decoupling stubs for isolation of sections of an antenna, the idea occurred that perhaps quarter-wave sections of spiral delay line (SDL) might be used to make a very simple trap, without lumped capacitance. Spiral delay line is coaxial line with a helical inner conductor.

Construction of SDL Traps

Figure 1 shows the coil which is the helical inner conductor of the spiral delay line, along with the polystyrene tubing and end pieces for the coil, and the aluminum tube. The completed assembly is shown in the title photograph and Figure 2.

For a 28.5-MHz trap, a coil of No. 12 magnet wire was wound on a ½-inch rod, 37 turns, close wound. The coil was removed from the ½-inch rod and it sprang out to about ¾-inch OD. The coil was 3¼ inches long. Enough wire was used to allow end wires straightened out to be 2 inches long on each end of the coil. Two pieces of ¾-inch OD polystyrene rod cut ½-inch thick were drilled in the center to fit over the No. 12 wire ends and one was slipped over each end of the coil. This coil was placed inside a piece of 1-inch OD × ¾-inch ID polystyrene tube 4¼-inches long, and a piece of 1⅛-inch OD × 1-inch × 4¼-inch long aluminum tube slipped over this. The assembly was held together with No. 6-32 × 5/16-inch screws in holes drilled and tapped just far enough into the polystyrene end pieces to hold the screws in place. The short for this quarter-wave section of spiral delay line was made with a solder lug under one screw with a wire soldered between it and one end of the coil, as shown in the photographs.

The curve in the graph, Figure 3, shows the number of turns of No. 12 wire required for quarter-wave sections of the above construction. A close-wound coil of No. 12 magnet wire has approximately 12 turns per inch. The length of the assembly for a particular frequency can be determined approximately by dividing the number of turns on the coil by 12 to get the length in inches.

The traps were adjusted to frequency through the use of a grid-dip meter (checked on a receiver for accuracy of each frequency). The coil can be changed quite easily to the desired frequency by trimming turns if an extra turn or two is put on for this purpose. The coil can also be wound with spacing between turns and compressed or expanded to get the trap exactly on frequency.

After the assembly is completed and tuned to frequency, the coil can be sealed in the polystyrene tube with polystyrene ce-

Figure 1 — The parts used in the make-up of a W4JRW SDL trap.

H1, H2 — End pieces of ½-inch length of ¾-inch OD Polystyrene rod with center hole for No. 12 wire.

H3 —1-inch OD × ¾-inch ID polystyrene tube, length one inch greater than that of coil turns of L1

H4 —1⅛-inch OD × 1-inch ID aluminum tube, length equal to that of H3.

L1— Close-wound coil of No. 12 enam. or magnet wire, ¾-inch OD to fit inside H3. See Figure 3 and text for turns information.

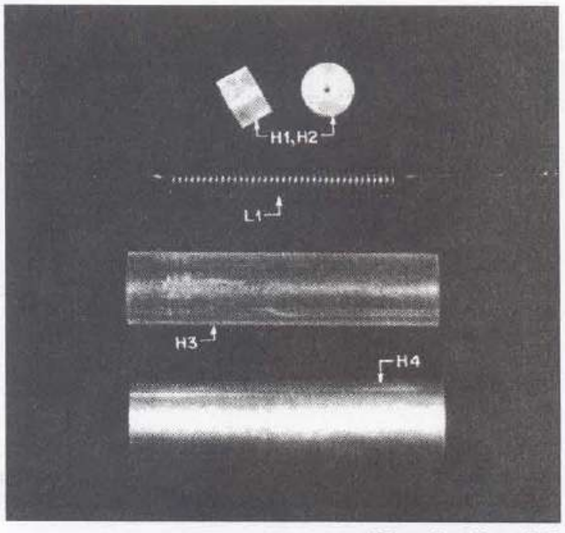

---

[1] Lattin, "Multiband Antennas Using Decoupling Stubs," *QST*, December, 1960, p. 23.

[2] Lattin, "Multiband Antennas Using Loading Coils," *QST*, April, 1961, p. 43.

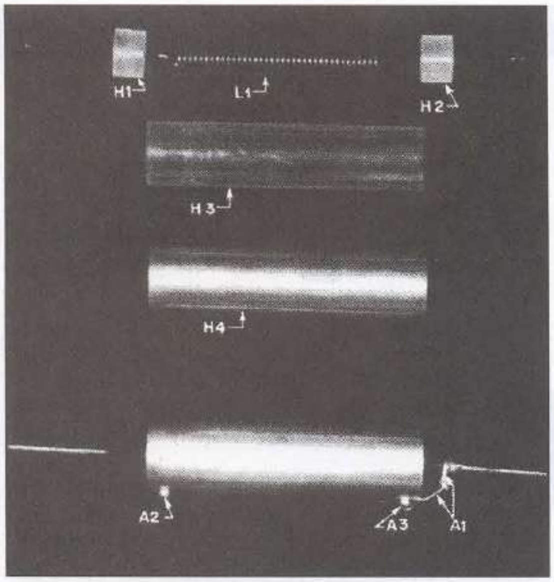

Figure 2—Assembly of the SDL traps. See Figure 1 and text for identification of parts not listed below.
A1—Solder lug and short length of wire assembled and soldered to short one end of H4 to one end of L1.
A2, A3—See text. The screw at A2 must not contact the wire of L1.

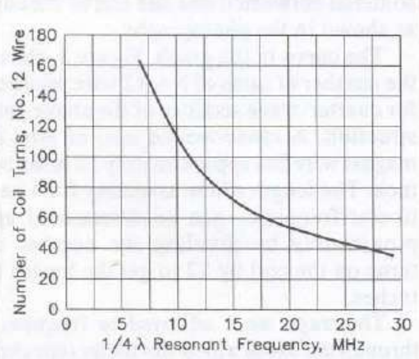

Figure 3—Resonant frequency of spiral-delay-line trap versus number of turns in coil.

ment or coil dope. An inert gas could be sealed inside quite easily, but there seems to be no particular advantage to this.

Trap Ratings and Performance

The thickness of the polystyrene tube used was 1/8 or 0.125 inch. The average voltage rating for polystyrene is given in various handbooks as 500 volts per mil (.001 inch). This would be 62,500 volts for the thickness of 1/8 inch. The maximum power rating of these traps has not been determined. They have been used with a 2-kW PEP ssb transmitter without any failures from either voltage breakdown or heating. Larger wire, polystyrene tubing, and aluminum tubing can be used, of course, but the curve of Figure 3 will be different. Formulas for characteristics of spiral delay lines can be found in radio handbooks and textbooks in which this type of line is described.

The resonant impedance of the traps was measured and found to be approximately 100,000 ohms. For comparison, a trap made of No. 12 wire and a ceramic capacitor gave about the same resonant impedance. Several lumped-constant 15-meter traps borrowed from triband beams were measured and values from 16,000 to 28,000 ohms were found!

The SDL traps can be used in beam antennas, of course, with suitable mechanical modifications to fit the aluminum tube used in the beam. Spiral delay line can be used for other purposes, such as matching transformers, phasing, and any place where coax line is used but short dimensions are needed. The Zo (characteristic impedance) is a function of wire size, diameter, and spacing of the helical coil, dimensions of the insulator and external aluminum tube. Measurement with an rf bridge indicated a Zo of about 250 ohms for the construction used in these traps.

Figures 4 and 5 give the dimensions of two doublet antennas experimented with here. The antenna of Figure 4 has two 7.2-MHz traps and resonates at 3.9 and 7.2 MHz. The antenna of Figure 5 has eight traps, two each for 10 meters, 15 meters, 20 meters, and 40 meters. Resonances are at 3.9, 7.2, 14.3, 21.3, and 28.6 MHz. The flat-top portions were made of No. 12 solid copper wire. The feeder used was RG-8/U. No balun was used in our experiments. It was found that if a trap was not tuned exactly to frequency, it could still be used by changing the wire lengths in the antenna adjacent to the trap to get the desired antenna resonance. An antenna shortened and using traps is sharper in resonance than a full-length doublet. This is generally very well known, but perhaps bears repetition.

SWR curves are also shown in Figures 4 and 5 for these two antennas. During measurements, the antennas were supported in the center about 30 feet high and were 20 feet high at the ends. Measurements were made at the transmitter with 100 feet of RG-8/U coax between the transmitter and the antenna.

It is advisable to support doublet antennas at the center as well as at the ends, with strain relief at the ends – a simple arrangement of a screw eye, plastic rope, and a sash weight or a brick will do.[3] Since RG-8/U coax is fairly heavy, the center support is helpful to reduce the strain on the antenna. The breaking load of No. 12 copper wire is given in handbooks as 197.5 pounds for soft or annealed wire and 261.6 pounds for medium hard-drawn wire. In a high wind any type of support such as trees, towers, push-up masts, and so on, may move a few inches, putting thousands of pounds of tension on a wire stretched between them. Is there any ham who hasn't broken a wire antenna stretched between two trees when no strain reliefs were used?

One spiral-delay-line trap was tested with a hoist and concrete blocks for weight and didn't break at 200 pounds. As the No. 12 wire used in the antenna was softer than

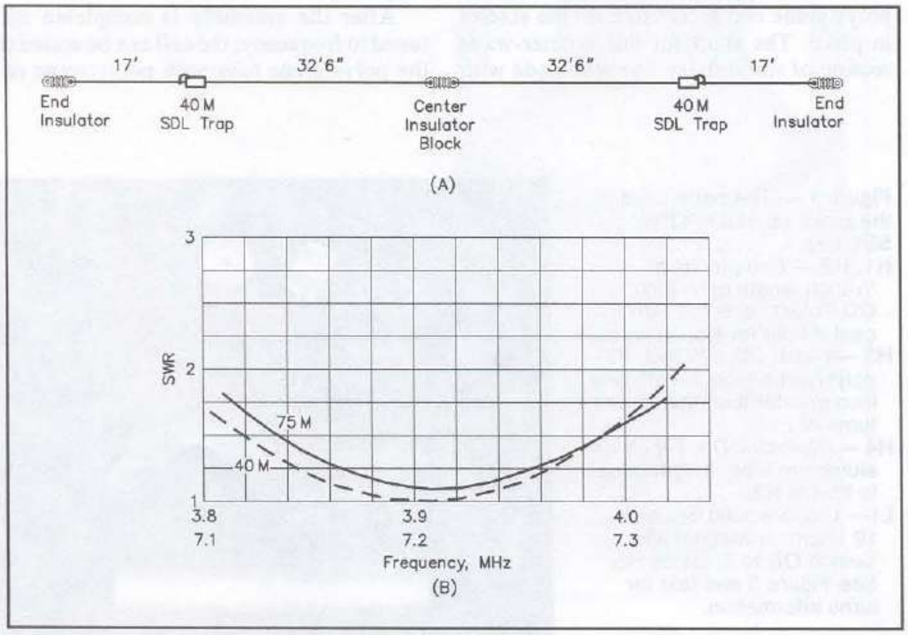

Figure 4—At A, dimensions for a 2-band spiral-delay-line antenna, resonant at 3.9 and 7.2 MHz (not drawn to scale). At B, the measured SWR values with this antenna.

---

[3] [EDITOR'S NOTE: To reduce wear and eventual breaking of the plastic-rope halyard, a pulley should also be used; large-diameter types sold in hardware stores as clothesline pulleys are economical and quite satisfactory.]

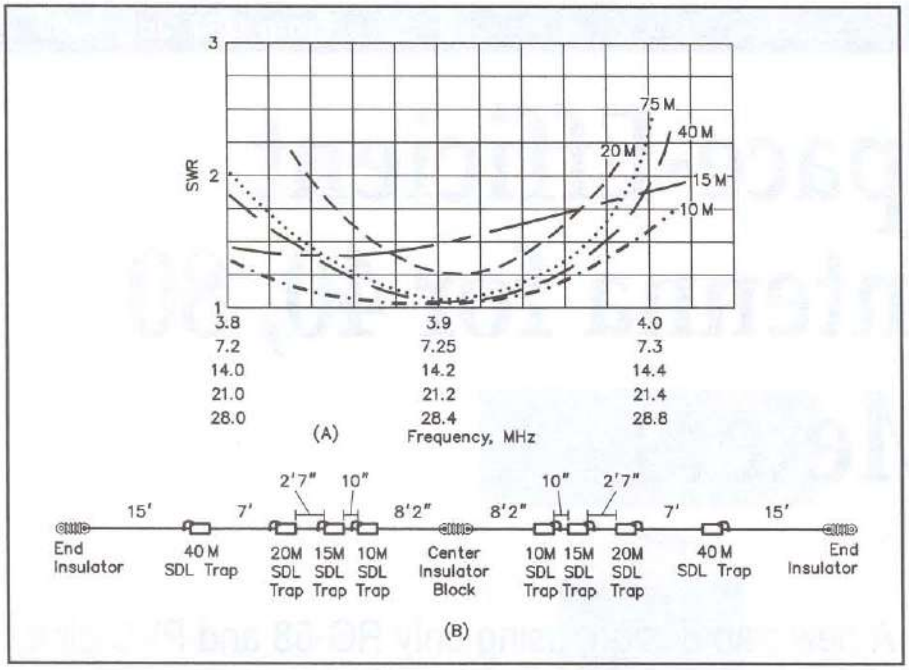

Figure 5—At A, the measured SWR values for a 5-band SDL antenna and at B the dimensions for this antenna.

the No. 12 magnet wires in the SDL traps, it appeared that the antenna wire would probably stretch before the wire in the traps. We have had the antenna up for almost two years, supported by three large trees, with strain reliefs at the ends. The antenna was not damaged at all by an 85-mph wind during a storm which bent the top section of our guyed crank-up tower into an inverted U shape with the triband beam hanging down. We did have one ice storm, but it wasn't severe enough to lift the strain-relief weights at the ends of the antenna.

If one desires, he can make the traps stronger by using two screws 180 degrees apart at each end, or even three screws at 120-degree spacing. Also the ends could be made of copper-weld wire soldered to the inner coil. The materials are not difficult to obtain. Most cities now have plastic supply distributors, and also aluminum tubing suppliers. The magnet wire can be obtained at a motor repair shop or from an electrical supply distributor.

By A. C. Buxton, W8NX From *QST*, July 1992

Build a Space-Efficient Dipole Antenna for 40, 80 and 160 Meters

A new trap design, using only RG-58 and PVC pipe, yields better space efficiency than conventional coaxial traps.

These days more than ever before, many hams who want to work the low bands need an effective antenna that fits on a small lot. I'll show you how to build a shortened dipole for 160, 80 and 40 meters using improved coaxial-cable traps that I call Super Traps. The antenna, which covers the three ham bands below 7.3 MHz, is about the same length as a fullsize 80-meter dipole. If you install the antenna as an inverted V with a 90° included angle, the baseline length is 88 feet. The antenna uses traps that are easily constructed, rugged and weatherproof. They use no exposed capacitors or inductors.

You can feed the antenna directly with balanced 75-Ω line or via a 1:1 balun with either 50- or 75-Ω coaxial cable. Feed-line length is not critical. The antenna resonates at 1.865, 3.825 and 7.225 MHz. I installed such an antenna on my lot as an inverted V, with the apex 38 feet high and the ends at about 15 feet.

As part of this project, I developed a BASIC-language computer program[1] for trap design; a listing is available from the ARRL.[2] You can use this program to design these traps for frequencies of your choice, but you don't need a computer to make the antenna described here.

Figure 1 shows the antenna layout. The antenna is made of #14 stranded wire and two pairs of coaxial traps. Construction is conventional in most respects, except for the high inductance-to-capacitance (L/C) ratio that results from the unique trap construction. Two recent *QST* articles give tips on dipole construction and feeding.[3,4]

The traps use two-layer windings of the core (dielectric and center conductor) of RG-58 coaxial cable. Coaxial cable with flexible, rugged stranded-wire center conductors is preferable to that with a more brittle solid-wire center conductor. Figure 2 shows the traps. The 3.8-MHz trap is shown with the weatherproofing cover of electrical tape removed to show the construction details.

Precautions and Trap Specifications

With this trap-winding configuration, there are two thicknesses of core dielectric material between adjacent turns, which doubles the breakdown voltage of the traps. The transformer action of the two windings gives a second doubling of the trap-voltage rating. Thus, the trap voltage rating is 5.6 kV (four times RG-58's 1.4-kV rating). Conventional coaxial-cable traps made of RG-58 have a rating of 2.8 kV.

The 7-MHz traps have 33 μH of inductance and 15 pF of capacitance, and the 3.8-MHz traps have 74 μH of inductance and 24 pF of capacitance. The trap Qs are over 170 at their design frequencies, as measured on a Boonton Q meter.

These traps are suitable for operation at the 1-kW power level. When making the traps, *do not* use RG-8X or any other foam-dielectric cable. Winding such cables on small-diameter forms causes the center conductor to migrate through the dielectric toward the inside, decreasing the breakdown rating and compromising trap performance. The core diameter also differs from that of RG-58.

Construction

Although these traps are similar in many ways to other coaxial-cable traps, the shield winding of the common coax-cable trap has been replaced by an outer winding that fits snugly into the grooves formed by the inner layer. Capacitance is reduced to 7.1 pF per

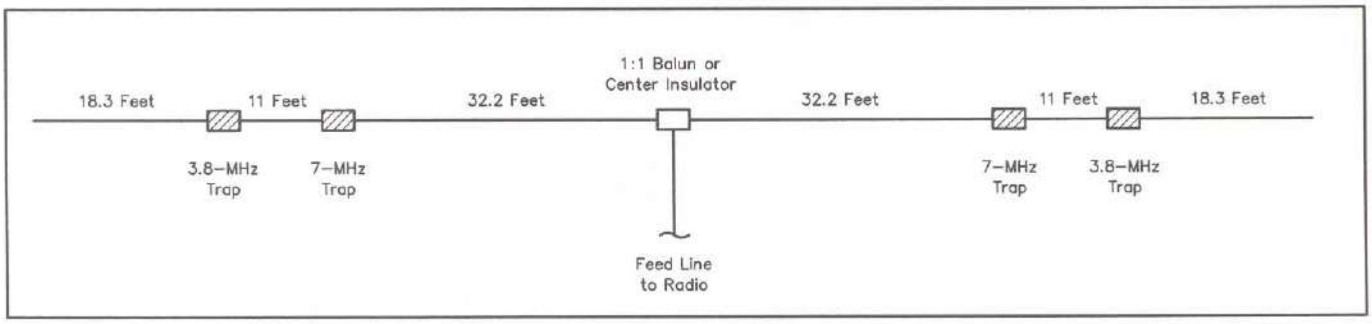

Figure 1—The shortened dipole resonates in the SSB portions of the 40, 80 and 160-meter bands. The antenna is 124 feet long.

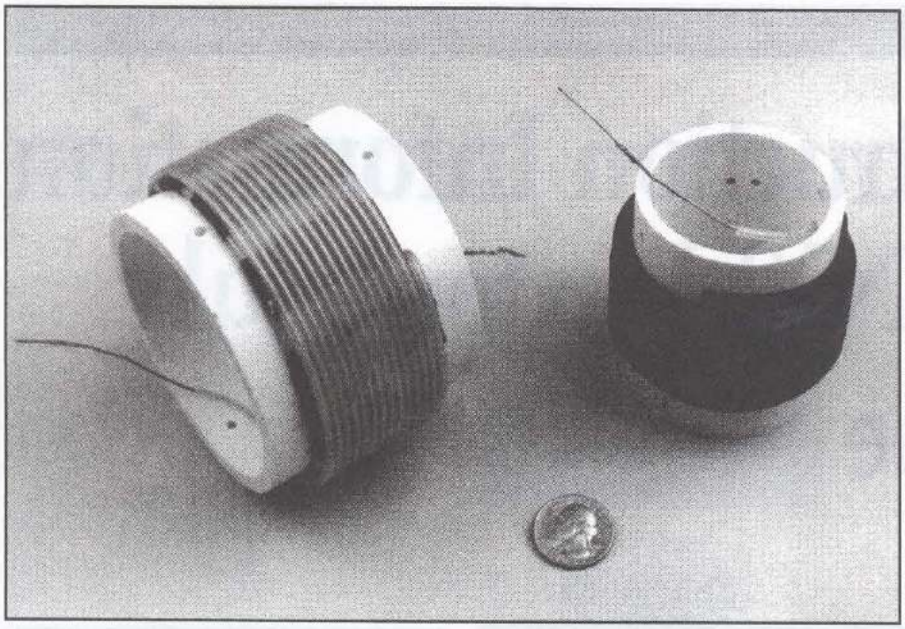

Figure 2—The improved coaxial-cable traps use two layered windings to provide an unusually high inductance-to-capacitance ratio, higher Q, and twice the breakdown voltage of single-layer traps. The 3.8-MHz trap is shown without its protective electrical-tape wrap to show the details of trap construction. This construction method makes for simple, lightweight, rugged and weatherproof traps.

foot, compared to 28.5 pF per foot with conventional coax traps made from RG-58. Trap reactance can be up to four times greater than that provided by conventional coax-cable traps.

The coil forms are cut from PVC pipe available at plumbing-supply stores. The 7-MHz trap form is made from 2-inch-ID pipe with an outer diameter of 2.375 inches. The 3.8-MHz trap form is made from 3-inch pipe with an outer diameter of 3.5 inches. The 7-MHz trap uses a 12.3-turn inner winding and an 11.4-turn outer winding. The 3.8-MHz trap uses a 14.3-turn inner winding and a 13.4-turn outer winding. All turns are closewound. The inner-trap frequency is 7.17 MHz and the outer-trap frequency is 3.85 MHz.

If you are unable to get PVC forms of exactly the same diameters as those called for here, compensate for the effect of form-size differences by taking advantage of the fact that the number of turns varies inversely with the form diameter. Thus, if the form diameter you use is, say, 5% larger than mine, reduce the number of turns by 5%. If necessary, add or remove fractions of a turn at the end of the outer winding. If you have a computer, you can use the BASIC programs[5] to calculate the exact number of turns for other form diameters. Stay as close as you can to the prescribed diameters because too much deviation changes the loading effect of the traps. A small change in trap loading may require a change in the lengths of the tip segments beyond the traps.

Use a #30 (0.128-inch diameter) drill for the feed-through holes in the PVC coil forms. The start and end holes of the 7-MHz traps are spaced 1.44 inches center to center, measured parallel to the trap center line. The holes in the 3.8-MHz traps are 1.66 inches apart. Wind the traps with a single length of coax core. The unspliced lengths are 17.55 feet for the 7-MHz traps and 28.45 feet for the 3.8-MHz traps. These lengths include the trap pigtails and a few inches for fine tuning.

Strip the jacket from the coax. This is easily done using a wood vise with wide jaws to hold the cable while cutting the jacket longitudinally with a sharp knife or razor. The coax outer conductor (braid) is best removed by pushing (not pulling) it off.

Use electrical tape to keep the turns of the inner-layer winding closely spaced during the winding process. This counteracts the tendency of the tension in the outer-layer winding to spread the inner–layer turns. Stick the tape strips directly to the coil form before winding and then tightly loop them over and around the inner layer before winding the outer layer. Use six or more tape strips for each trap.

If possible, check the resonant frequencies of your traps with a dip meter. Try to maintain an accuracy of 50 kHz or better.

For low-noise reception, erect the antenna as close to horizontal as possible. If you let the ends of the antenna droop toward the ground, as I have done with my inverted-V installation, you may have to accept a somewhat higher noise level in the interest of structural simplicity and reduced baseline length. Some feel that the inverted V configuration is better for DXing than a horizontal dipole at the same height. For an inverted V with a 90° included angle (legs that slope downward at 45°), you'll need a minimum apex height of about 55 feet and a baseline length of 88 feet. Get the apex as high as you can and keep the ends at least 10 feet above the ground for safety.

Configuration and Performance Trade-Offs

You seldom get something for nothing. This antenna proves no exception to that rule. As with all trap dipoles, this one has less-than-ideal bandwidth due to the loading effect of the traps. This is the price paid for multiband coverage and physical shortening. This antenna covers 65 kHz of 160 meters, 75 kHz of 80 meters and the entire 40-meter band with SWRs under 2:1. The bandwidth limitations on 160 and 80 meters can be largely offset with an antenna tuner.

It is also important to recognize that the traps are used in low-current portions of the antenna, minimizing I^2R trap losses. A relatively high radiation resistance is therefore also retained.

Good luck with your low-band antennas!

Notes

[1] I wrote the program in GWBASIC 3.2. It uses generic BASIC commands and can be easily converted for use with computers other than IBM PCs and compatibles.

[2] For a copy of the BASIC program, send a business-size SASE to the ARRL Technical Department Secretary, 225 Main St, Newington, CT 06111-1494. Request the July 1992 *QST* BUXTON BASIC PROGRAM listing.

[3] J. Healy, "Antenna Here is a Dipole," *QST*, Jun 1991, pp 23-26.

[4] J. Healy, "Feeding Dipole Antennas," *QST*, Jul 1991, pp 22-24.

[5] See note 2.

From *QST*, September 1992 (Technical Correspondence)

Trap Construction Information for Al (W8NX) Buxton's July 1992 Dipole

Several readers have written asking for clearer instructions on how to build the antenna traps described by A. C. Buxton, W8NX, in his July 1992 *QST* article, "Build a Space-Efficient Dipole Antenna for 40, 80 and 160 Meters." For those of you who didn't write, but aren't sure how to build the traps, here's what you need to know.

The trap-winding technique is deceptively simple. Each trap is simply comprised of one winding atop another, wound in the same direction. First, drill four holes in the form for wire entry and exit, as described in July *QST*. Starting from inside the form, pass the wire through the hole labeled "1" in Figures 1 and 2. Wind a layer of the inner conductor on the form. Then, run the wire end down through a second hole into the form (just below the one labeled "EXIT" in Figure 2) and, inside the form, bring the wire back to a point adjacent to where the first winding started. Bring the wire up through the hole marked "2" and wind another layer on top of the first, in the same direction as the first winding. Pass the wire down through the hole labeled "EXIT" and you're done. The tape you see in the photos, described in the article, helps hold the bottom winding's turns together and the top winding in place as you're assembling the trap.

The wires passing through holes in the form are strain-relieved, but use the two mounting holes you drilled in the form to support the trap, just to be on the safe side.
–Rus Healy, NJ2L

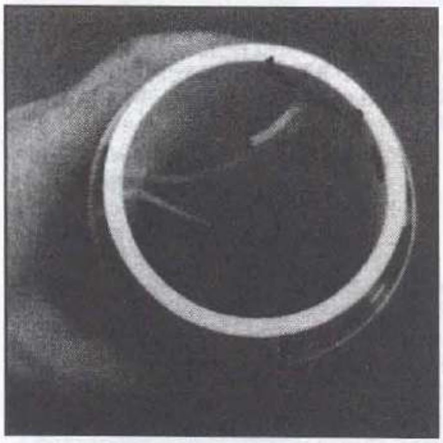

Figure 1—An inside view of a W8NX two-layer trap shows how the windings enter and exit the form. Two holes at each end of the form pass the windings into and out of the form.

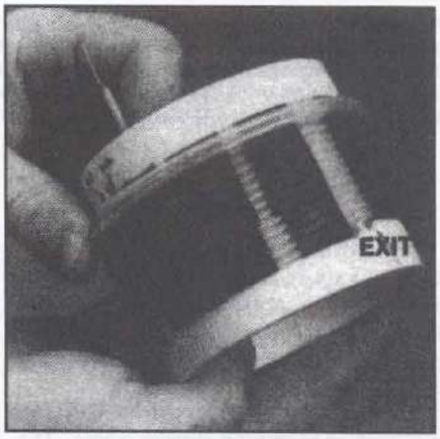

Figure 2—An outside view of a partially assembled W8NX trap. The bottom winding starts at hole "1" and reenters the form just below the "EXIT" hole. The wire then comes back through the inside of the form to hole "T," and is used to make a second winding atop the first. It reenters the form at the "EXIT" hole.

By Al Buxton, W8NX **From *QST*, August 1994**

Two New Multiband Trap Dipoles

W8NX details a new coax trap design used in two multiband antennas; one covering 80, 40, 20, 15 and 10 meters, and the other covering 80, 40, 17 and 12 meters.

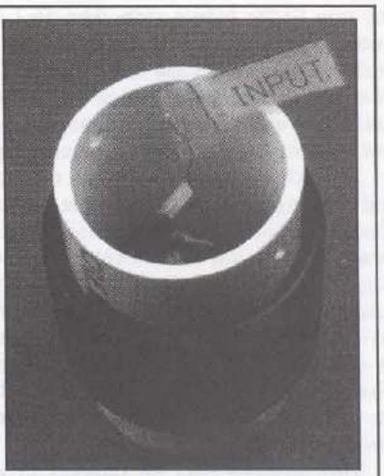

Over the last 60 or 70 years, amateurs have used many kinds of multiband antennas to cover the traditional HF bands. The availability of the 30, 17 and 12-meter bands has expanded our need for multiband antenna coverage. A fortunate few have the space and resources for multiband antennas like rhombics or long Vs, but many hams have employed inverted-L long wires or parallel dipoles. Old-timers will recall the offcenter-fed *Windom* of the '30s–the first version using a single-wire transmission line, and the later design using two-wire feed line. Over the years, random-length dipoles with open-wire feeders and associated tuners have been used successfully as multiband antennas. The G5RV multiband antenna is a specialized example of this approach.[1]

The *log periodic array* represents a kind of brute-force approach to the goal of achieving coverage of multiple HF ham bands. It seems inefficient because of the large gaps between our relatively narrow amateur HF bands.

Over the last few decades, two factors have affected the development of multiband antennas—the popularity of low-impedance (usually 50-Ω) coaxial feed lines, and the appearance of untuned, 50-Ω solid-state amplifiers. The impedance of an antenna is relatively low only at its fundamental frequency and at odd-order harmonics. Although antenna tuners are often necessary to resonate an antenna system, the quest for expanded multiband coverage with simple antennas continues.

At the end of the 1930s, a different technological approach appeared in the form of resonant traps in antennas. The *Mims Signal Squirter* is the grandfather of modern day tribanders.[2] This article discusses in detail an innovative trap design employed in two multiband dipoles.

One W8NX Trap Design—Two Multiband Dipoles

Two different antennas are described here. The first covers 80, 40, 20, 15 and 10 meters, and the second covers 80, 40, 17 and 12 meters. Each uses the same type of W8NX trap—connected for different modes of operation—and a pair of short capacitive stubs to enhance coverage. The new W8NX coaxial-cable traps have two different modes: a high- and a low-impedance mode. The inner-conductor windings and shield windings of the traps are connected in series in the conventional manner for both modes. However, either the low- or high-impedance point can be used as the trap's output terminal. For low-impedance trap operation, only the center conductor turns of the trap windings are used. For high-impedance operation, all turns are used, in the conventional manner for a trap. The short stubs on each antenna are strategically sized and located to permit more flexibility in adjusting the resonant frequencies of the antenna.

Figure 1 shows the configuration of the 80, 40, 20, 15 and 10-meter antenna. The radiating elements are made of #14 stranded copper wire. The element lengths are the wire span lengths in feet. These lengths do not include the lengths of the pigtails at the balun, traps and insulators. The 32.3-foot-long inner 40-meter segments are measured from the eyelet of the input balun to the tension relief hole in the trap coil form. The 4.9-foot segment length is measured from the tension relief hole in

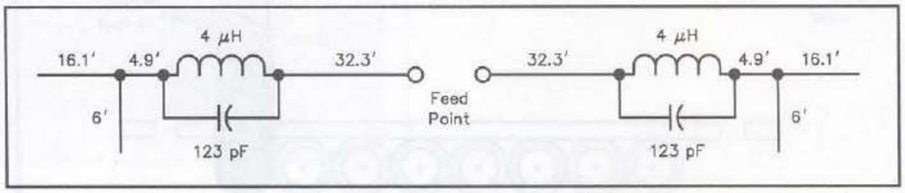

Figure 1—A W8NX multiband dipole for 80, 40, 20, 15 and 10 meters. The values shown (123 pF and 4 µH) for the coaxial-cable traps are for parallel resonance at 7.15 MHz. The low-impedance output of each trap is used for this antenna.

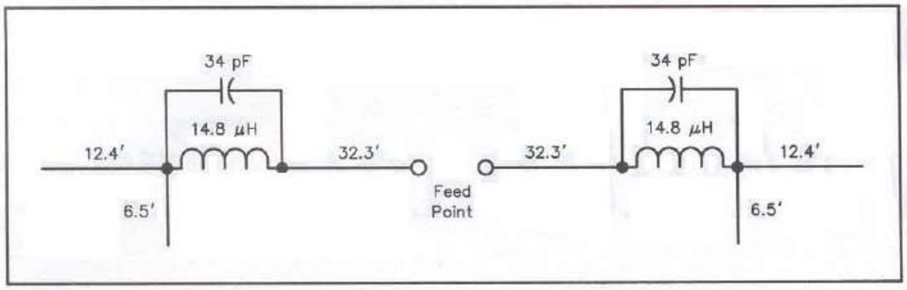

Figure 2—A W8NX multiband dipole for 80, 40, 17 and 12 meters. For this antenna, the high-impedance output is used on each trap. The resonant frequency of the traps is 7.15 MHz.

the trap to the 6-foot stub. The 16.1-foot outer-segment span is measured from the stub to the eyelet of the end insulator. The coaxial-cable traps are wound on PVC pipe coil forms and use the low-impedance output connection. The stubs are 6-foot lengths of 1/8-inch stiffened aluminum or copper rod hanging perpendicular to the radiating elements. The first inch of their length is bent 90° to permit attachment to the radiating elements by large-diameter copper crimp connectors. Ordinary #14 wire may be used for the stubs, but it has a tendency to curl up and may tangle unless weighed down at the end. I recommend that you feed the antenna with 75-Ω coax cable using a good 1:1 balun.

This antenna may be thought of as a modified W3DZZ antenna[3] (shown for many years in various ARRL publications) with the addition of capacitive stubs. The length and location of the stub give the antenna designer two extra degrees of freedom to place the resonant frequencies within the amateur bands. This additional flexibility is particularly helpful to bring the 15 and 10-meter resonant frequencies to more desirable locations in these bands. The actual 10-meter resonant frequency of the W3DZZ antenna is somewhat above 30 MHz, pretty remote from the more desirable low frequency end of 10 meters.

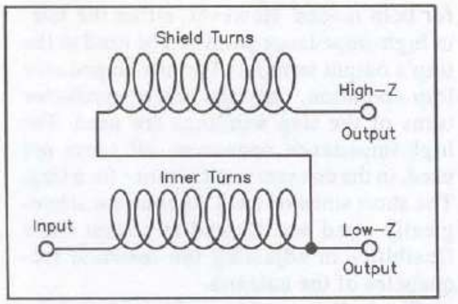

Figure 3—Schematic for the W8NX coaxial-cable trap. RG-59 is wound on a $2\frac{3}{8}$-inch OD PVC pipe.

Figure 2 shows the configuration of the 80, 40, 17 and 12-meter antenna. Notice that the capacitive stubs are attached immediately outboard after the traps and are 6.5 feet long, 0.5 foot longer than those used in the other antenna. The traps are the same as those of the other antenna, but are connected for the high-impedance output mode. Since only four bands are covered by this antenna, it is easier to fine tune it to precisely the desired frequency on all bands. The 12.4-foot tips can be pruned to a particular 17-meter frequency with little effect on the 12-meter frequency. The stub lengths can be pruned to a particular 12-meter frequency with little effect on the 17-meter frequency. Both such pruning adjustments slightly alter the 80-meter resonant frequency. However, the bandwidths of the antennas are so broad on 17 and 12 meters that little need for such pruning exists. The 40-meter frequency is nearly independent of adjustments to the capacitive stubs and outer radiating tip elements. Like the first antennas, this dipole is fed with a 75-Ω balun and feed line.

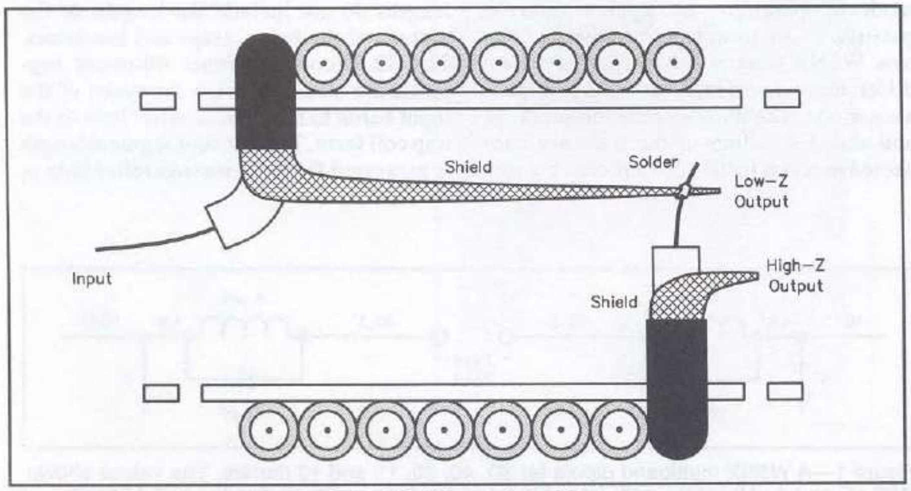

Figure 4—Construction details of the W8NX coaxial-cable trap.

Figure 3 shows the schematic diagram of the traps. It explains the difference between the low and high-impedance modes of the traps. Notice that the high-impedance terminal is the output configuration used in most conventional trap applications. The low-impedance connection is made across only the inner conductor turns, corresponding to one-half of the total turns of the trap. This mode steps the trap's impedance down to approximately one-fourth of that of the high-impedance level. This is what allows a single trap design to be used for two different multiband antennas.

Figure 4 is a drawing of a cross-section of the coax trap shown through the long axis of the trap. Notice that the traps are conventional coaxial-cable traps, except for the added low-impedance output terminal. The traps are $8\frac{3}{4}$ close-spaced turns of RG-59 (Belden 8241) on a $2\frac{3}{8}$-inch-OD PVC pipe (schedule 40 pipe with a 2-inch ID) coil form. The forms are $4\frac{1}{8}$ inches long. Trap resonant frequency is very sensitive to the outer diameter of the coil form, so check it carefully. Unfortunately, not all PVC pipe is made with the same wall thickness. The trap frequencies should be checked with a dip meter and general coverage receiver and adjusted to within 50 kHz of the 7150 kHz resonant frequency before installation. One inch is left over at each end of the coil forms to allow for the coax feed-through holes and holes for tension-relief attachment of the antenna radiating elements to the traps. Be sure to seal the ends of the trap coax cable with RTV sealant to prevent moisture from entering the coaxial cable.

Also, be sure that you connect the 32.3-foot wire element at the start of the inner conductor winding of the trap. This avoids detuning the antenna by the stray capacitance of the coaxial-cable shield. The trap output terminal (which has the shield stray capacitance) should be at the outboard side of the trap. Reversing the input and output terminals of the trap will lower the 40-meter frequency by approximately 50 kHz, but there will be negligible effect on the other bands.

The title-page photos show a coaxial-cable trap. Details of the trap installation are shown in Figure 5. This drawing applies specifically to the 80, 40, 20, 15 and 10-meter antenna, which uses the low-impedance trap connections. Notice the lengths of the trap pigtails: 3 to 4 inches at each terminal of the trap. If you use a different arrangement, you must modify the span lengths accordingly. All connections can be made using crimp connectors rather than by soldering. Access to the trap's interior is attained more easily with a crimping tool than with a soldering iron.

Antenna Patterns

The performance of both antennas has

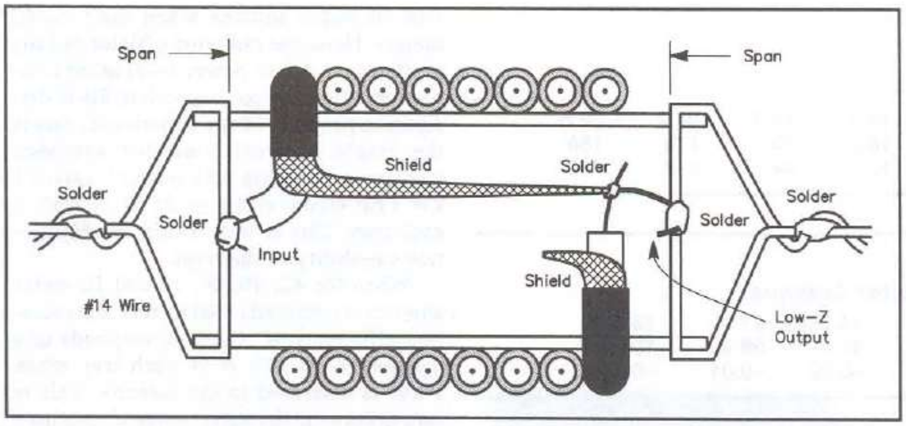

Figure 5—Additional construction details for the W8NX coaxial-cable trap.

been very satisfactory. I am currently using the 80,40, 17 and 12-meter version because it covers 17 and 12 meters. (I have a tribander for 20, 15 and 10 meters.) The radiation pattern on 17 meters is that of 3/2-wave dipole. On 12 meters, the pattern is that of a 5/2-wave dipole. At my location in Akron, Ohio, the antenna runs essentially east and west. It is installed as an inverted V, 40 feet high at the center, with a 120° included angle between the legs. Since the stubs are very short, they radiate little power and make only minor contributions to the radiation patterns. The pattern has four major lobes on 17 meters, with maxima to the northeast, southeast, southwest, and northwest. These provide low-angle radiation into Europe, Africa, South Pacific, Japan and Alaska. A narrow pair of minor broadside lobes provides north and south coverage into Central America, South America and the polar regions.

There are four major lobes on 12 meters, giving nearly end-fire radiation and good low-angle east and west coverage. There are also three pairs of very narrow, nearly broadside, minor lobes on 12 meters, down about 6 dB from the major end-fire lobes. On 80 and 40 meters, the antenna has the usual figure-8 patterns of a half-wave-length dipole. I have some pattern distortion and input impedance effects from aluminum siding on my house. Nevertheless, DX is easily workable on either of these antennas using a 100-W transceiver, when the high-frequency bands are open.

Both antennas function as electrical half-wave dipoles on 80 and 40 meters with a low SWR. They both function as odd harmonic current-fed dipoles on their other operating frequencies, with higher, but still acceptable, SWR. The presence of the stubs can either raise or lower the input impedance of the antenna from those of the usual third and fifth harmonic dipoles. Again, I recommend that 75-Ω, rather than 50-Ω, feed line be used because of the generally higher input impedances at the harmonic operating frequencies of the antennas.

The SWR curves of both antennas were carefully measured. A 75 to 50-Ω transformer from Palomar Engineers was inserted at the junction of the 75-Ω coax feed line and my 50-Ω SWR bridge. The transformer prevents an impedance discontinuity, with attendant additional undesired line reflections appearing at the 75 to 50-Ω junction. The transformer is required for accurate SWR measurement if a 50-Ω SWR bridge is used with a 75-Ω line. No harm is done to any equipment, however, if the transformer is omitted. Most 50-Ω rigs operate satisfactorily with a 75-Ω line, although this requires different tuning and load settings in the final output stage of the rig or antenna tuner. I use the 75 to 50-Ω transformer only when making SWR measurements and at low power levels. The transformer is rated for 100 W, and when I run my 1-kW PEP linear amplifier the transformer is taken out of the line. (I hope my absent-mindedness doesn't catch up with me some day!)

Figure 6 gives the SWR curves of the 80, 40, 20, 15 and 10-meter antenna. Minimum SWR is nearly 1:1 on 80 meters, 1.5:1 on 40 meters, 1.6:1 on 20 meters, and 1.5:1 on 10 meters. The minimum SWR is slightly below 3:1 on 15 meters. On 15 meters, the stub capacitive reactance combines with the inductive reactance of the outer segment of the antenna to produce a resonant rise that raises the antenna input resistance to about 220 Ω, higher than that of the usual 3/2-wavelength dipole. An antenna tuner may be required on this band to keep a solid-state final output stage happy under these load conditions.

Figure 7 shows the SWR curves of the 80, 40, 17 and 12-meter antenna. Notice the excellent 80-meter performance with a

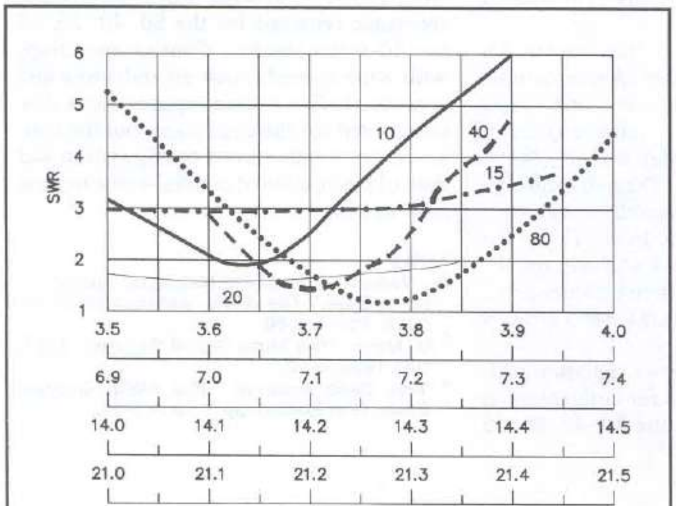

Figure 6—Measured SWR curves for an 80, 40, 20, 15 and 10-meter antenna, installed as an inverted-v with 40-ft apex and 120° included angle between legs.

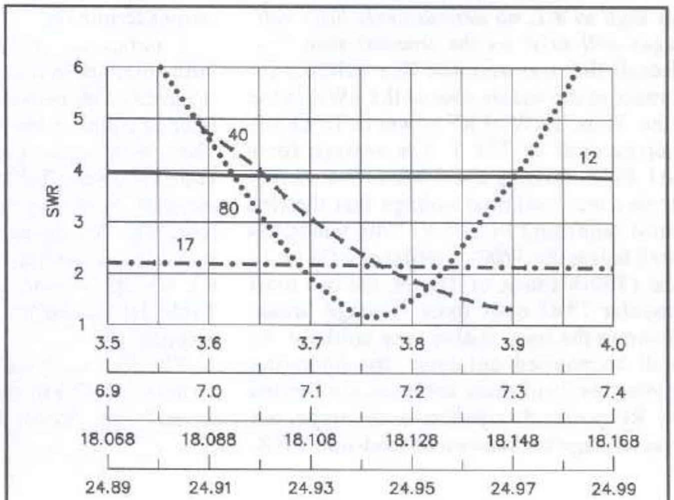

Figure 7—Measured SWR curves for an 80, 40, 17 and 12-meter antenna, installed as an inverted-v with 40-ft apex and 120° included angle between legs.

Table 1
Trap Q

Frequency (MHz)	3.8	7.15	14.18	18.1	21.3	24.9	28.6
High Z out (Ω)	101	124	139	165	73	179	186
Low Z out (Ω)	83	103	125	137	44	149	155

Table 2A
Trap Loss Analysis: 80, 40, 20,15, 10-Meter Antenna

Frequency (MHz)	3.8	7.15	14.18	21.3	28.6
Radiation Efficiency (%)	96.4	70.8	99.4	99.9	100.0
Trap losses (dB)	–0.16	–1.5	–0.02	–0.01	–0.003

Table 2B
Trap Loss Analysis: 80, 40, 17, 12-Meter Antenna

Frequency (MHz)	3.8	7.15	18.1	24.9
Radiation Efficiency (%)	89.5	90.5	99.3	99.8
Trap losses (dB)	–0.5	–0.4	–0.03	–0.006

nearly unity minimum SWR in the middle of the band. The performance approaches that of a full-size 80-meter wire dipole. The short stubs and the very low inductance traps shorten the antenna somewhat on 80 meters. Also, observe the good 17-meter performance, with the SWR being only a little above 2:1 across the band.

But notice the 12-meter SWR curve of this antenna, which shows 4:1 SWR across the band. The antenna input resistance approaches 300 Ω on this band because the capacitive reactance of the stubs combines with the inductive reactance of the outer antenna segments to give resonant rises in impedance. These are reflected back to the input terminals. These stub-induced resonant impedance rises are similar to those on the other antenna on 15 meters, but are even more pronounced.

Too much concern must not be given to SWR on the feed line. Even if the SWR is as high as 9:1, *no destructively high voltages will exist on the transmission line.* Recall that transmission-line voltages increase as the square root of the SWR in the line. Thus, 1 kW of RF power in 75-Ω line corresponds to 274 V line voltage for a 1:1 SWR. Raising the SWR to 9:1 merely triples the maximum voltage that the line must withstand to 822 V. This voltage is well below the 3700-V rating of RG-11, or the 1700-V rating of RG-59, the two most popular 75-Ω coax lines. Voltage breakdown in the traps is also very unlikely. As will be pointed out later, the operating power levels of these antennas are limited by RF power dissipation in the traps, not trap voltage breakdown or feed-line SWR.

Trap Losses and Power Rating

Table 1 presents the results of trap Q measurements and extrapolation by a two-frequency method to higher frequencies above resonance. I employed an old, but recently calibrated, Boonton Q meter for the measurements. Extrapolation to higher frequency bands assumes that trap resistance losses rise with skin effect according to the square root of frequency, and that trap dielectric losses rise directly with frequency. Systematic measurement errors are not increased by frequency extrapolation. However, random measurement errors increase in magnitude with upward frequency extrapolation. Results are believed to be accurate within 4% on 80 and 40 meters, but only within 10 to 15% at 10 meters. Trap Q is shown at both the high- and low-impedance trap terminals. The Q at the low-impedance output terminals is 15 to 20% lower than the Q at the high-impedance output terminals.

I computer-analyzed trap losses for both antennas in free space. Antenna-input resistances at resonance were first calculated, assuming lossless, infinite-Q traps. They were again calculated using the Q values shown in Table 1. The radiation efficiencies were also converted into equivalent trap losses in decibels. Table 2A summarizes the trap loss analysis for the 80, 40, 20, 15 and 10-meter antenna and Table 2B for the 80, 40, 17 and 12-meter antenna.

The loss analysis shows radiation efficiencies of 90% or more for both antennas on all bands except for the 80, 40, 20, 15 and 10-meter antenna when used on 40 meters. Here, the radiation efficiency falls to 70.8%. A 1-kW power level at 90% radiation efficiency corresponds to 50-W dissipation per trap. In my experience, this is the trap's survival limit for extended key-down operation. SSB power levels of 1 kW PEP would dissipate 25 W or less in each trap. This is well within the dissipation capability of the traps.

When the 80, 40, 20, 15 and 10-meter antenna is operated on 40 meters, the radiation efficiency of 70.8% corresponds to a dissipation of 146 W in each trap when 1 kW is delivered to the antenna. This is sure to burn out the traps—even if sustained for only a short time. Thus, the power should be limited to less than 300 W when this antenna is operated on 40 meters under prolonged key-down conditions. A 50% CW duty cycle would correspond to a 600-W power limit for normal 40-meter CW operation. Likewise, a 50% duty cycle for 40-meter SSB corresponds to a 600-W PEP power limit for the antenna.

I know of no analysis where the burnout wattage rating of traps has been rigorously determined. Operating experience seems to be the best way to determine trap burn-out ratings. In my own experience with these antennas, I've had no traps burn out, even though I operated the 80, 40, 20, 15 and 10-meter antenna on the critical 40-meter band using my AL-80A linear amplifier at the 600-W PEP output level. I have, however, made no continuous, keydown, CW operating tests at full power purposely trying to destroy the traps!

Summary

Some hams may suggest using a different type of coaxial cable for the traps. The dc resistance of 40.7 Ω per 1000 feet of RG-59 coax seems rather high. However, I've found no coax other than RG-59 that has the necessary inductance-to-capacitance ratio to create the trap characteristic reactance required for the 80, 40, 20, 15 and 10-meter antenna. Conventional traps with wide-spaced, open-air inductors and appropriate fixed-value capacitors could be substituted for the coax traps, but the convenience, weatherproof configuration and ease of fabrication of coaxial-cable traps is hard to beat.

Notes

[1] L. Varney, "The G5RV Multiband Antenna... Up-to-Date," *The ARRL Antenna Compendium, Vol. 1*, p 86.

[2] M. Mims, "The Mims Signal Squirter," *QST*, Dec 1939, p 12.

[3] "Five Band Antenna," *The ARRL Antenna Book*, 16th Edition, pp 7-10 to 7-11.

By William C. Gann, W4NML From *QST*, May 1966

A Center-Fed "Zepp" for 80 and 40

Fast QSY for the Phone-C.W. Operator

The center-fed "Zepp" antenna is reviewed by W4NML, showing how complete coverage of a single band is made easy by using old concepts. Although the author shows how to use the Zepp on 80 and 40, only, the system can be used from 80 through 10 meters by employing an all-band transmatch.

Multiband antennas fed with resonant feeders were very popular in the pre-coax cable days. This article is presented to review a good, but seemingly forgotten system. This antenna should be of interest to traffic and contest operators, and to the casual operator who likes to use both the c.w. and phone portions of the 80- and 40-meter bands.

Our section s.s.b. net meets on 3965 kHz and the c.w. net meets on 3575 kHz. Many schemes were tried to make one antenna usable on both ends of the band so that a low s.w.r. could be maintained while securing efficient operation at the different frequencies. None of the antennas tried would permit an excursion of more than 300 kHz without a serious s.w.r. problem between the transmitter and the line. Getting from 80 to 40 meters with such an antenna was even more perplexing. The writer's dilemma was finally solved by the installation of the old reliable center-fed Zepp antenna.

Choosing the Dimensions

In order to use the antenna on 40, 75, and 80 meters, tuned feeders are required[1]. So that the feeders can be matched to the transmitter, a transmatch is used at the "shack" end of the line. Parallel tuning is used to minimize the complexity of the transmatch. This requires that the transmission line presents a high impedance to the transmatch on both bands.

The charts in the handbooks did not give a set of Zepp antenna dimensions that were suitable for the author's installation. Because of the existing tower, which would permit the antenna to be supported at the 50-foot level, and because the ham shack was adjacent to the tower, the prescribed feeder lengths were not practical. A graph was plotted to show the frequency extremes to which the antenna would be tuned, showing the minimum and maximum impedance points across the bands. It was determined that the combined length of one leg of the feed line and one section of the dipole would be 114 feet.[2] A length of 53 feet was used for the feed line and each leg of the driven element was cut to 61 feet.

To broaden the antenna's response, the driven element's effective area was made larger by paralleling two lengths of No. 12 copper wire as shown in Figures 1 and 2. With this arrangement, the Q of the antenna is lower, permitting the operator to QSY approximately 200 kHz without readjusting the transmatch.

Construction Notes

The driven element and the feed line are made from No. 12 copper wire. The 4-inch wide ceramic spreaders used to hold the feeder wires apart are made by the E. F. Johnson Co. Light-weight poly spacers are used to spread the driven-element wires and are sold as TV "clothespins" by the Telco Co. (Figure 2). All of the spreaders are attached to the No. 12 wire by short pieces of No. 18 copper wire. The distance between the spreaders is 4 feet for both the driven element and the feed line.

Sections of 1 × 4-inch lumber are used to hold the feeders away from the steel tower (Figure 3). Each piece is 24 inches long, notched at one end, and is fastened to the tower with U-bolts. Porcelain telephone-type insulators are attached to the feed-line end of each board, offering low-loss anchor

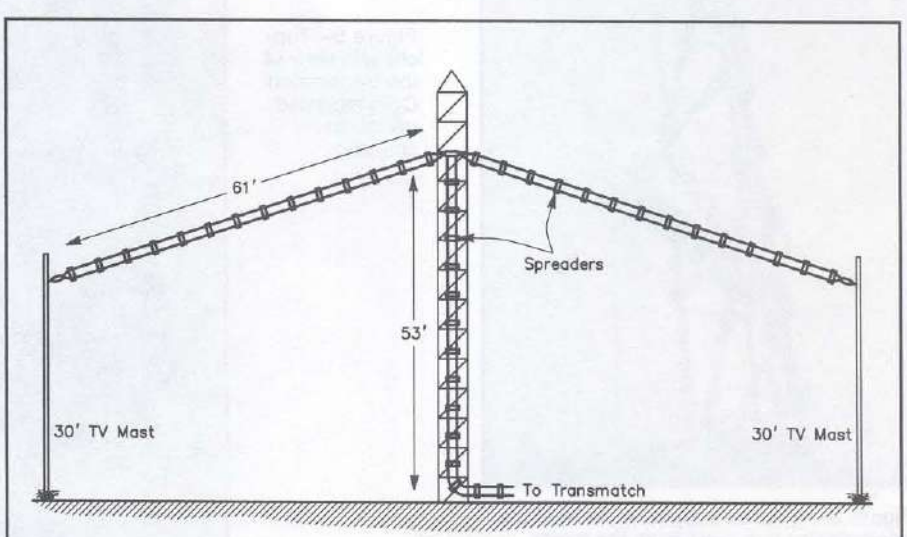

Figure 1—Layout of the 2-band Zepp antenna. Dimensions for each part of the antenna are shown. Feed point is anchored to one of the wooden support arms (See Figure 3).

---

[1] Center-fed Antennas, *A.R.R.L. Antenna Book*, Chapter 6.

[2] This length is generally 145 feet for operation in the 3.5- to 30-MHz range when the antenna is mounted horizontally (no droop), away from steel towers, and with a singlewire driven element.–*Editor*

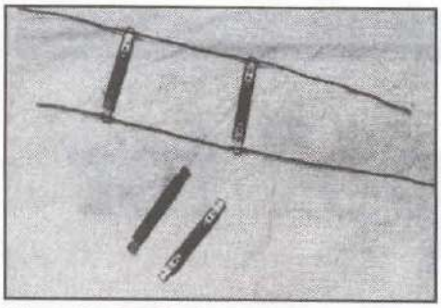

Figure 2—Details showing how the driven element spreaders are attached to the No. 12 wire. No. 18 copper wire is wrapped above and below each spreader to hold it in place.

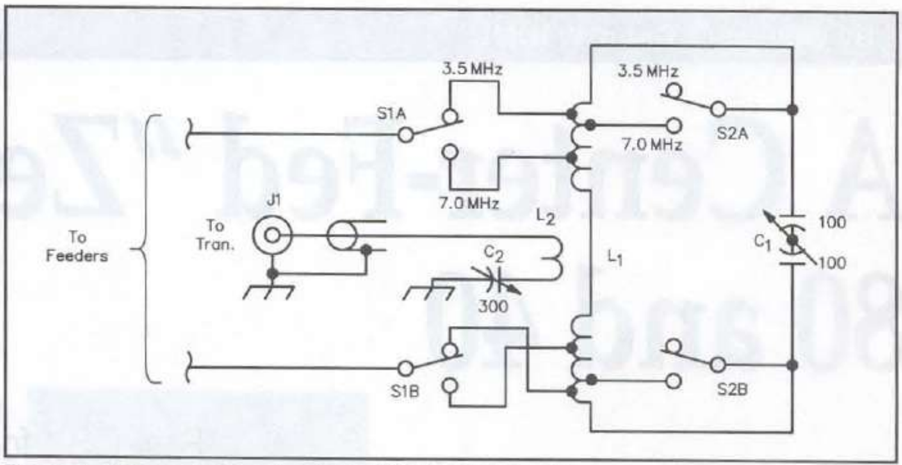

Figure 4—Schematic diagram of the W4NML transmatch.

C_1—100 pf. per section transmitting variable. (Split stator type with 0.175-inch spacing between plates.)

C_2—300-pf. variable capacitor(0.078 spacing or greater).

J_1—SO-239 coax connector.

L_1—56 turns, 3-inch diam., 8 turns per inch coil. S_1 taps are 9 turns from ends of coil for 80 meters, and are 22 turns from ends of coil for 40 meters. S_2 taps are 5 turns from ends of coil for 80 meters and 17 turns from ends of coil for 40 meters. (Air Dux 2408T or Polycoils 1779 usable.)

L_2—8 turns of Air Dux 2408T (center portion of L_1).

S_1, S_2—Ceramic rotary, 2 poles, 2 positions, 2 sections.

points for the transmission line. The uppermost support arm, at the 50-foot level, is used as a mount for the center of the driven element.

The far ends of the antenna are supported by 30-foot TV masts. A pulley and halyard arrangement is used for raising and lowering the ends of the antenna. Because the end supports are not as high as the feed point of the antenna, the dipole has a slight droop, but this does not seem to impair the performance.

The transmission line is brought into the operating position by means of feed-through insulators, mounted on a plywood strip which fits under a partially raised window. Insulated No. 12 house wire is used between the feed-through insulators and the transmatch.

Transmatch

Ideas for the author's tuner (Figure 4) were taken from the excellent transmatch article by McCoy.[3] Band changing is made possible by a large ceramic switch (origin unknown) which was obtained at a hamfest. An identical switch is used for selecting the taps for the feed line (Figure 5). Coil L_1 contains 56 turns of No. 14 wire, is 3 inches in diameter, and has 8 turns-per-inch (Air Dux 2408T). A stationary link, L_2, at the center of L_1, contains 8 turns of No. 14 wire and is a part of the Air Dux coil from which L_1 is made. The link is tuned with a 300-pf. variable capacitor. The author did not have a unit of the correct type, so two 150-pf. capacitors were parallel-connected (mounted under the chassis). Capacitor C_1 is a 100-pf-.per-section variable

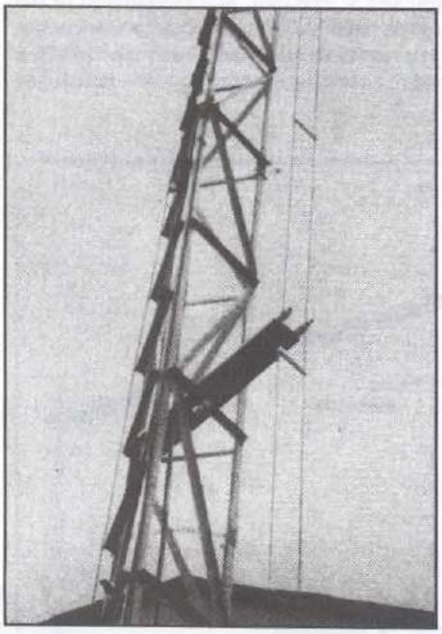

Figure 3—Wooden support arms hold the transmission line away from the tower. Telephone-type insulators are mounted at the end of each board to make the feed line secure.

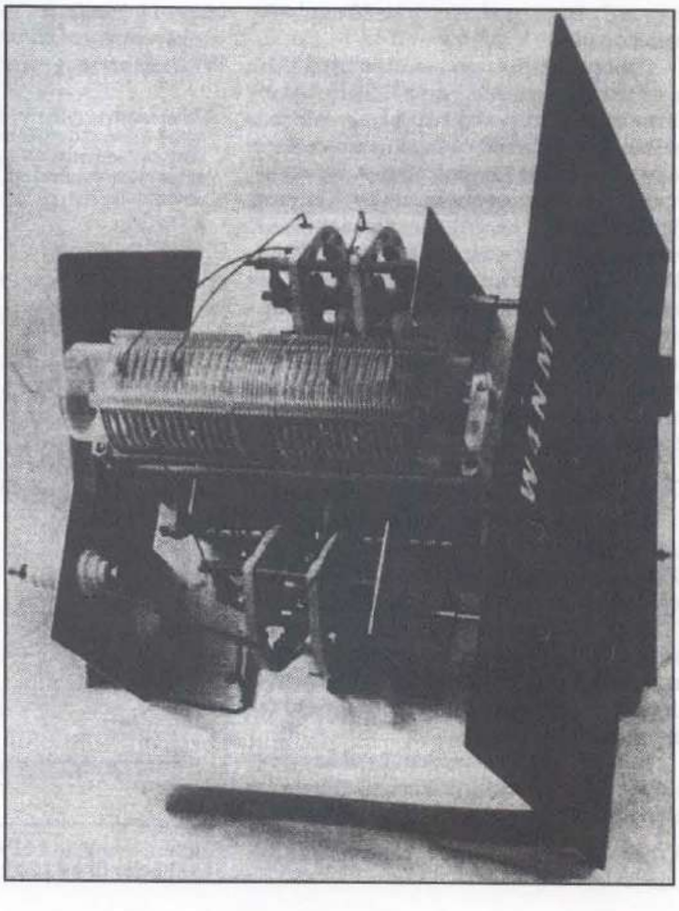

Figure 5—Top-chassis view of the transmatch. C_2 is mounted under the chassis.

---

[3] *QST*, July 1965.

with wide spacing. To give L_1 some rigidity, it is mounted on a plexiglass tube which is supported by the frame of C_1 with standoff insulators.

Results

While using clip leads, the correct tap points for the feeders were found by operating the transmitter through a Collins wattmeter and tuning C_1 and C_2 for zero reflected power. The transmatch permitted the transmitter to "see" 50 ohms in any part of either band. After establishing the correct tap points for the feed line, permanent connections were made between L_1 and the switches.

Next, the tuner was used with the 30L-1 amplifier at an output level of 700 watts. After a 30-minute QSO, no evidence of coil heating could be detected.

When compared to other antenna systems used by the author, the new skywire showed improved performance. It was believed that some sacrifice in efficiency would result from changing to the new antenna. Happily, it was found that we could have our cake and eat it too! Extended use indicated that the performance was, indeed, better than with previous antennas used.

I wish to thank three friends for their help in making this article possible: K4WWN for his tower climbing and photography work, K4ADK for building the transmatch cabinet, and Roy LeCrone for additional darkroom and photographic assistance.

Although this antenna system is an old standard, it may be the answer to your QSY problems. The cost is nominal and the results are most rewarding.

From *QST*, March 1967 (Technical Correspondence)

All-Band Antenna

With reference to the article on a Center-fed Zepp for 80 and 40 in May 1966 *QST*:

I set out to accomplish several things with an antenna to be installed on a California lot which runs east-west:

1) One pole.
2) No guys.
3) Good for short skip up and down the West Coast on 80, 75 and 40 meters.
4) Throw lobes across populated DX areas on 20 and 15 meters with a fairly low radiation angle.
5) Allow some onmidirectional DX on 40 meters.
6) Keep away from anything with critical antenna length or critical tuning.
7) Minimum cost.

Figure 1 shows the arrangement I ended up with. Results have been exceptional for a simple system of low height.

I use No. 14 wire for the antenna. The feeders are also No. 14. The center is mounted on an unguyed wooden pole about 34 feet high. Each section of the antenna is 65 feet long and the ends are only 14 feet high. Thirty-foot feeders are used with series tuning on 80 and 40 and parallel tuning on 20, 15 and 10 meters. Loading from 3.5 to 30 MHz is excellent and not at all critical in tuning. The fact that the feeders are less than 1/8 wave on 80 allows reactance to be tuned out in the feeder-tuning arrangement on that band. The antenna is a bit long for the high end of 75 meters, but tuning there is good (this length was picked because of the slightly longer physical length required on the upper bands for end effect).

I have the antenna itself running east-west, giving some directivity north-south for QSOs with short skip up and down the west coast on 80 and 40 meters (it was found in an earlier antenna . . . vertical . . . that a vertical was not satisfactory for high-radiation-angle short-skip operation). On 20 and 15 the lobes tend to cut across major population DX areas. The tilt of the wire, which lowers the vertical radiation angle plus apparently some lobe addition, seems to give better results in the desired DX directions on 20 and 15 meters than 33-feet-high half-wave horizontal antennas oriented in the correct directions. Quite a bit of omnidirectional DX has been worked on 40 meters, undoubtedly because of the antenna tilt. DX operation on 15-meter c.w. has been really exceptional. Quite often I hook a DX station through the pileup when local beam stations miss (power output is about 150 watts). Since 10 has opened up I have used the antenna quite a bit on that band with very good results for both North American and DX contacts.

As a result of playing around on 160 meters with the bottom of the feeders connected together and working the antenna as a "T" against ground, I decided to see what happened when it was operated as a top-loaded vertical on 80 and 40 meters. The feeders were tied together and the antenna worked against a ground consisting of two 8-foot rods in water-soaked earth. On 80 this places the maximum-current point directly at the top of the vertical section (1/8 wave long) and on 40 gives the effect of a 1/4-wave vertical with maximum current at the bottom. This arrangement gave much better results than the Zepp where low-angle radiation was required, and less effective results than the Zepp where medium- and high-angle radiation was required. An exception is directly off the ends on 80 meters, where the vertical and Zepp seem to give the same results. The Zepp arrangement is therefore now used for short and medium skip on 80 and 40 and the vertical arrangement for long skip or DX. In receiving, the signal-to-noise ratio decreases greatly with the vertical arrangement (vs horizontal), thus somewhat offsetting the overall advantage of the vertical for DX operation. If this condition is extreme, I use the Zepp for receiving and the vertical for transmitting, for DX operation. It appears that the vertical transmitter and horizontal receiver is by far the best DX arrangement for metropolitan areas, but most likely the vertical for both would be best for rural areas where the QRN is lower. – *Dave Hardacker, W6PIZ*

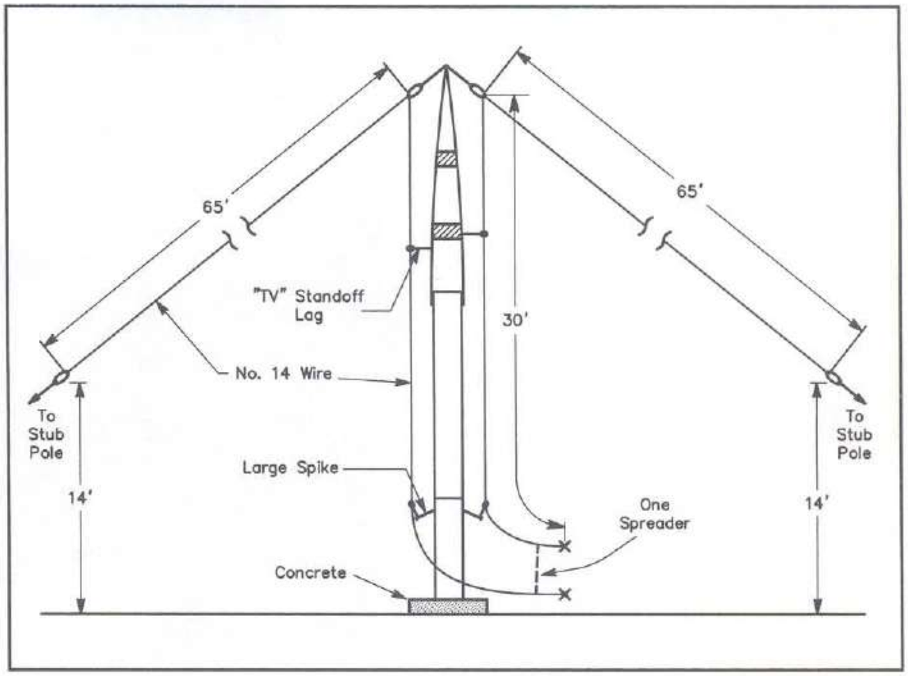

Figure 1–The W6PIZ all-band antenna. Power is applied through a series- or parallel-tuned link-coupled matching circuit at X-X for operation on 3.5 through 30 MHz (see text). For 2 MHz, and also for certain types of work on 3.5 and 7 MHz, points X-X can be connected together and the antenna worked against ground.

By William J. Lattin, W4JRW **From *QST*, December 1960**

Multiband Antennas Using Decoupling Stubs

Substituting Transmission Line Sections for Lumped-Constant Traps

Since W4JRW obtained a patent on this multi-frequency antenna system nearly ten years ago we can't call it "new," but at least it should be welcome news to those seeking a simple way to get good radiation on several bands, Shorted 1/4-wavelength stubs provide r.f. insulation and also serve as part of the antenna.

Since amateurs usually desire to operate on more than one band, several methods have been devised to use a single antenna on several bands. The earliest arrangements employed various combinations of feeder lengths, antenna lengths, and series or parallel tuning of the coupling circuit. Later on, the use of parallel-tuned "traps" with lumped constants which act as insulators at a particular frequency was invented.[1] A practical arrangement of this system for amateur use was developed[2] and is in rather wide use.

It is well known that the parallel-tuned circuit and quarter-wavelength shorted stub of Figure 1 are very similar electrically. Both configurations show a high impedance across points A and B. However, if a stub is connected to an antenna in this manner it does not act as an insulator but rather as a phase changer. The collinear antenna uses such stubs to operate a series of halfwave sections in phase.

There is a different connection possible for the stub, that is from A to C, which will result in insulator action or decoupling in an antenna.[3] For instance, shorted stubs a quarter wavelength long at 28 MHz can be attached to the ends of a 28 MHz dipole as in Figure 2. The 28 MHz dipole is effectively isolated or decoupled from the balance of the antenna which can be made long enough to resonate at 14, 7 or 3.5 MHz. If another pair of stubs is added for 14 MHz,

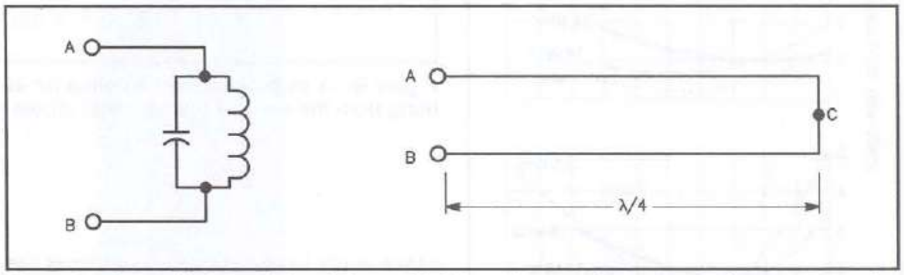

Figure 1—A parallel-tuned circuit has a high impedance at its resonant frequency, and so does a 1/4-wavelength shorted transmission line.

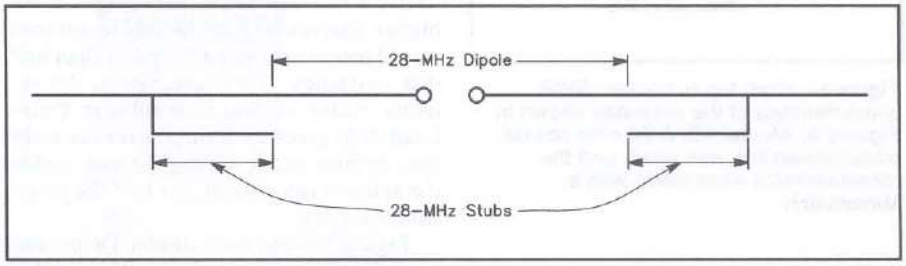

Figure 2—A two-band antenna for 28 MHz and some lower frequency. The center portion is an ordinary 10-meter dipole. The shorted stubs are 1/4 wavelength long at 28 MHz and look like an open circuit at that frequency when connected to the dipole as shown. Extensions on the ends of the stubs can be used to resonate the antenna at any frequency less than half of 28 MHz.

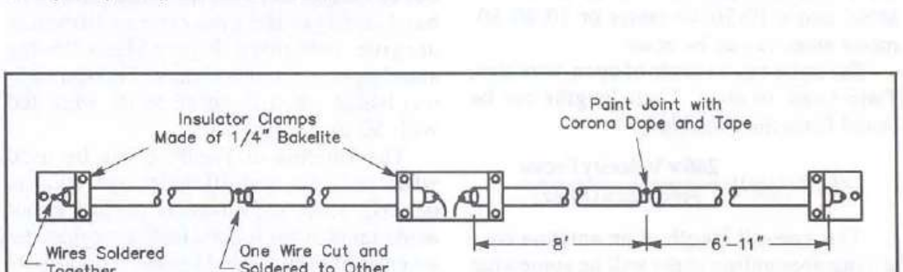

Figure 3—Construction and dimensions of an antenna for 10 and 20 meters using 300-ohm tubular Twin-Lead for both the dipole and stubs. Either a 50- or 75-ohm transmission line can be connected at the center of the dipole.

---

[1] Morgan, "A Multifrequency Tuned Antenna System," *Electronics*, August, 1940.

[2] Buchanan, "The Multimatch Antenna System," *QST*, March, 1955.

[3] Lattin, Patent No. 2,535,298.

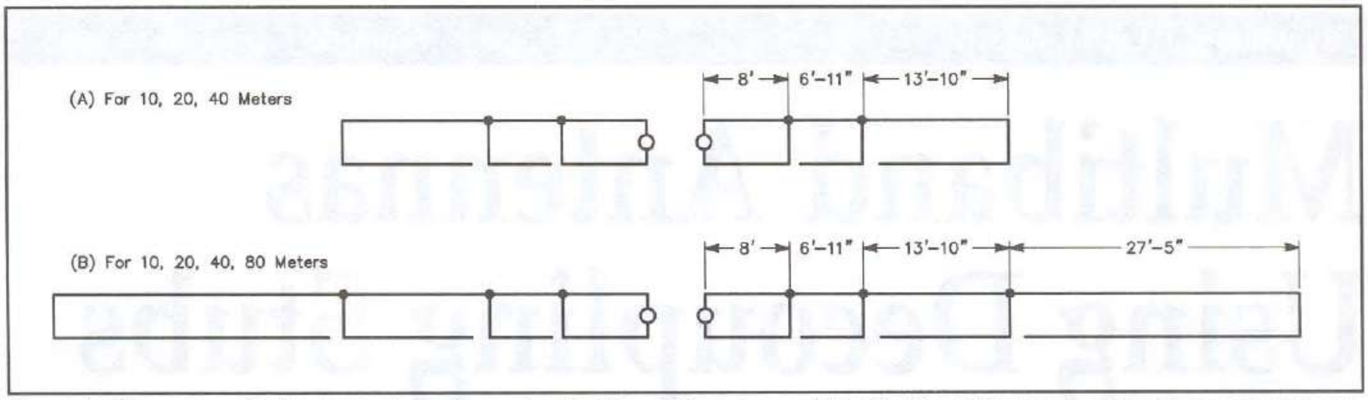

Figure 4—Dimensions of stub-decoupled antennas for 10, 20 and 40 meters and 10, 20, 40 and 80 meters made of tubular Twin-Lead. Either antenna can also be used on 15 meters where the 40-meter section is 3/4 wavelength long.

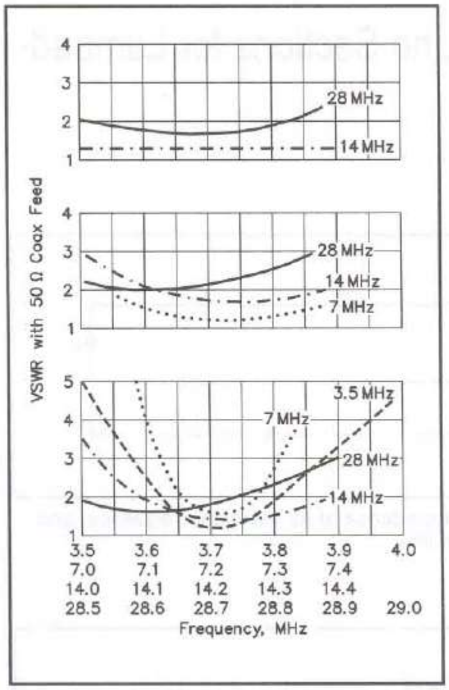

Figure 5—From top to bottom, SWR characteristics of the antennas shown in Figures 3, 4A and 4B. A 50-ohm coaxial transmission line was used, and the measurements were made with a Micromatch.

there will be isolation at both 28 and 14 MHz, and a 10-20-40-meter or 10-20-80-meter antenna can be made.

The stubs can be made of open-wire line, Twin-Lead, or coax. Their lengths can be found from the formula

$$\text{Length (feet)} = \frac{246 \times \text{Velocity Factor}}{\text{Frequency (MHz)}}$$

The over-all length of an antenna containing decoupling stubs will be somewhat less than the figure given by the usual formula for a half-wavelength dipole — *Length* (feet) = 468/*Frequency* (MHz). For instance, an antenna for 10 and 20 meters must be 29 feet, 10 inches long for resonance at the lower frequency, whereas the formula gives a length of 33 feet.

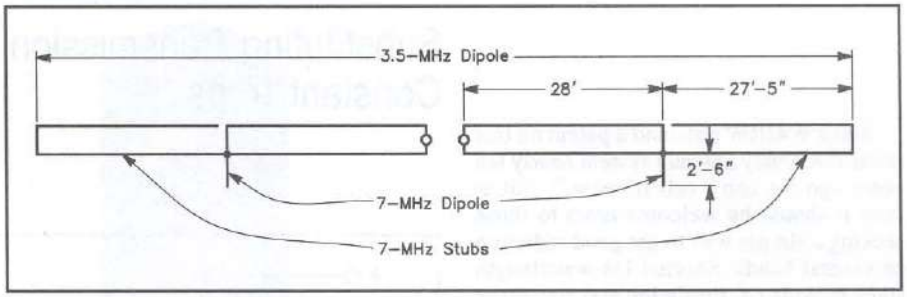

Figure 6—A stub-decoupled antenna for 40 and 80 meters. In this case wires must be hung from the ends of the 40-meter dipole to resonate the antenna in that band.

If open line with a velocity factor of nearly unity is used for the stubs, the overall length of a two-band antenna would be nearly a full free-space wavelength at the higher frequency and the whole antenna would resonate at something *less* than half that frequency. Very fortunately, the velocity factor of 300-ohm tubular Twin-Lead (0.8) gives such lengths for the stubs that, in most cases, adding the stub makes the antenna resonate at just half the original frequency.

Figure 3 shows how tubular Twin-Lead can be used for the antenna itself as well as the stubs and includes dimensions for 10- and 20-meter operation. The foam-filled type of Twin-Lead is recommended to keep out moisture. Lengths for three- and four-band antennas using the same construction are given in Figure 4. Figure 5 indicates the standing-wave ratios observed across various bands when these antennas were fed with 50-ohm coax.

The antenna of Figure 6 can be used when only 40- and 80-meter operation is desired. Since the 40-meter portion is not made up of stubs it must be longer than the antenna of Figure 4A. However, the isolating stubs must still be 1/4 wavelength long (allowing for velocity factor), and the whole antenna would resonate at a frequency below 3.5 MHz if the stubs were simply added to the ends of the 7-MHz dipole. To get around this, the dipole is shortened until the whole antenna tunes to 80 meters. Then resonance at 40 meters is restored by adding extra lengths of wire at the stub junctions. These wires are short and can just hang down from the antenna as shown.

Any of the antennas which will operate on 40 meters can be used on 15 meters as the 40-meter stubs will be approximately 3/4 wavelength long and will provide decoupling. The result is equivalent to operating a 7-MHz dipole at three times its resonant frequency, and we have found the s.w.r. is usually not lower than 3 to 6 when using 40-meter antennas of any type on 15 meters[4].

The power rating of the antenna will depend on the insulation at the stub junctions. These junctions can be painted with corona dope and covered with vinyl tape. It has been our experience over several years that the insulation will not break down with a kilowatt-input transmitter, 100 percent modulated, except when wet or very damp. In this case, the input should be reduced to

[4] Theoretically, a center-fed antenna working on its third harmonic shouldn't be more than about 50 percent higher in resistance than on the fundamental. One would expect an s.w.r. on the order of 2 to 1 rather than such high figures. – Ed.

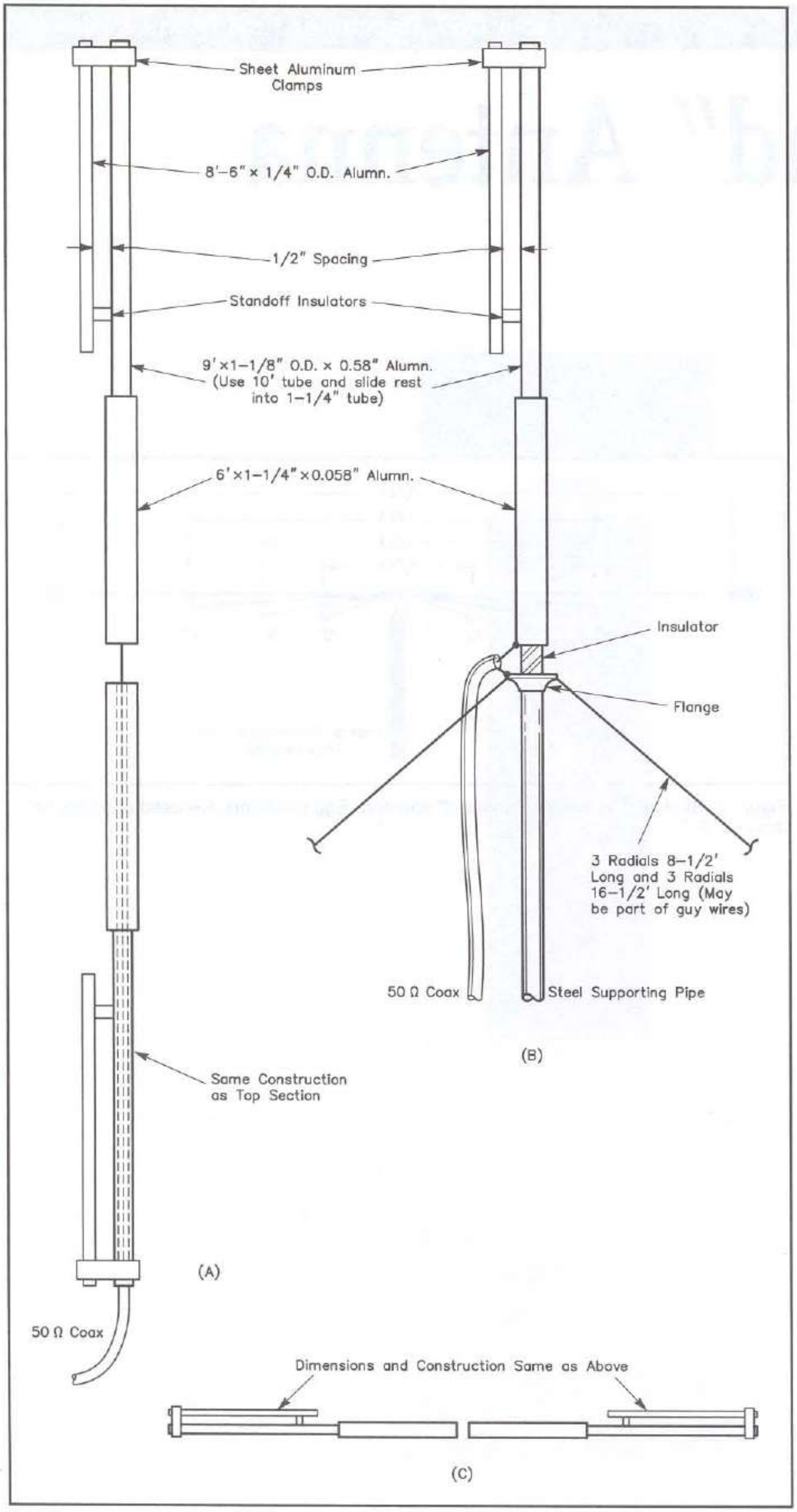
Sheet Aluminum Clamps
8'-6" × 1/4" O.D. Alumn.
1/2" Spacing
Standoff Insulators
9'×1-1/8" O.D. × 0.58" Alumn. (Use 10' tube and slide rest into 1-1/4" tube)
6'×1-1/4"×0.058" Alumn.
Insulator
Flange
3 Radials 8-1/2' Long and 3 Radials 16-1/2' Long (May be part of guy wires)
50 Ω Coax
Steel Supporting Pipe
(B)
Same Construction as Top Section
(A)
50 Ω Coax
Dimensions and Construction Same as Above
(C)

Figure 7—Dimensions and suggested construction for coaxial, ground-plane and tubing dipole antennas for 10 and 20 meters. The arrangement in A might be mounted with standoff insulators attached to the 1¼-inch sections near the center of the antenna. The dipole in C could be closed at the center and fed with a gamma or "T"-matching system. Similarly-constructed parasitic elements could be added to make a multiband beam.

perhaps 500 watts unless special precautions have been taken to seal up the junctions at the open ends of the stub. Of course, on the lowest band for which the antenna is designed the stubs do not have voltage across them and will not be subject to breakdown or flashover. The high voltage across the open end of a stub occurs only at the resonant frequency of that stub.

Figure 7 shows the construction of several 10 and 20-meter antennas which have been built and the dimensions required for resonance in these bands. The spacing between the rods forming the shorted stubs is not critical—the same lengths were obtained with 1-inch instead of 1/2-inch spacing. Insulators should be made of low-loss material. Reflectors and directors for a multiband beam could be made up the same way.

"All-Band" Antenna

Figure 1 is a sketch of an "all-band" antenna system that I have been using with success for some time. The idea is not a new one, having appeared in *QST* at least 10 years ago. However, I feel that there are many newcomers since that time who would be interested in a simple system that can be fed with a single 70-ohm transmission line.

The arrangement consists of dipoles, cut for each band and connected in parallel at the center. Although I have not checked standing-wave ratios, the results seem to indicate that it gets out as well as a bunch of individually-fed doublets. If you haven't tried it, you're in for some surprises.– *R. L. Cope, W8MOK*

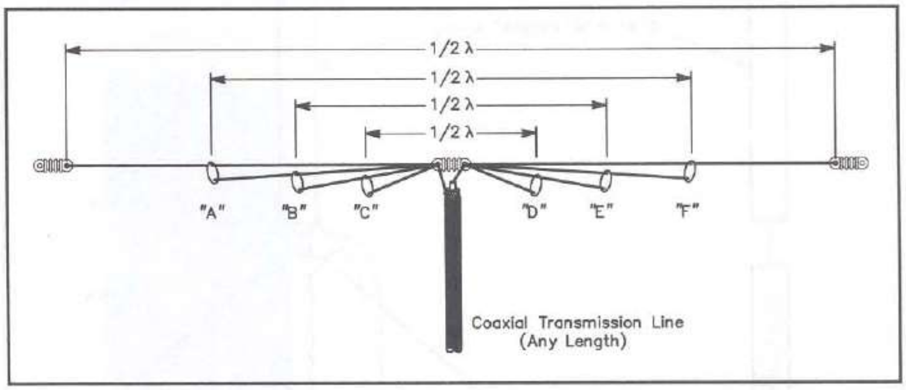

Figure 1–Sketch of W8MOK's "all-band" antenna. Egg insulators are used at points "A" through "F."

Three-Band Matching System for a Forty-Meter Doublet

A common method for energizing a half-wave antenna is to feed it at the center with parallel-conductor TV lead-in, or Twin-Lead as it is usually called, and to use an open stub for matching the 50- to 70-Ω antenna resistance to the 300-Ω impedance of the line. However, this technique, as described in the latest edition of *The ARRL Antenna Book*, generally gives proper matching on only one band.

After a number of trial-and-error calculations on a Smith Chart, along with lots of cut-and-try experimenting, I devised a three-stub matching scheme so that I could operate my 40-meter doublet on 40, 20 and 15 meters. Figure 1 illustrates this method and gives the lengths of the stubs and their positions along the feed line. The dimensions shown are for standard Twin-Lead, with a velocity factor of 0.82. All of the stubs are open at the ends and are made from the same type of line as the feed line. Note that the two lower stubs are connected at the same point on the feed fine.

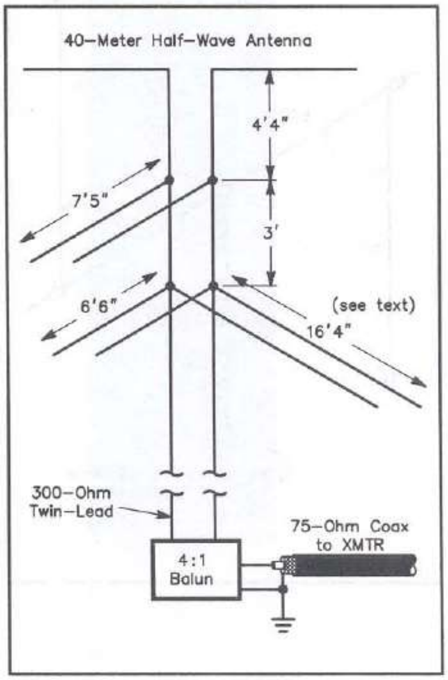

The length of the longest stub is fairly critical. It should first be cut to 17 feet and then trimmed no more than two inches at a time, until the SWR is minimum in the center or the desired portion of the 20-meter band. The feed fine can then be matched with a 4:1 broad-band balun to a 75-Ω coaxial cable from the transmitter. With my antenna, the described matching system gives an SWR of less than 2.5 to 1 over all of the three bands, with minimum values of 1.3 to 1 on 40 and 15 meters, and 1.7 to 1 on 20 meters. As with any multiband antenna, one must guard against harmonic radiation. – ***Frank Stuart, K7UUC***

From *QST*, December 1979 (Technical Correspondence)

A "Z" Antenna For the 10-160 Meter Bands

One of my interests in ham radio is designing and constructing antennas for both general amateur use and for the Army MARS system. For the amateur who has limited space, I have designed a "Z" antenna that covers the bands from 10 through 160 meters. It is easily constructed from wire. Spreaders for the transmission line are fabricated from Lucite strips or Plexiglas rods. (Refer to the yellow pages of telephone directories for the names of dealers who handle Plexiglas or the equivalent.) For two no. 14 wires, a 2-inch (51-mm) spacing is adequate.

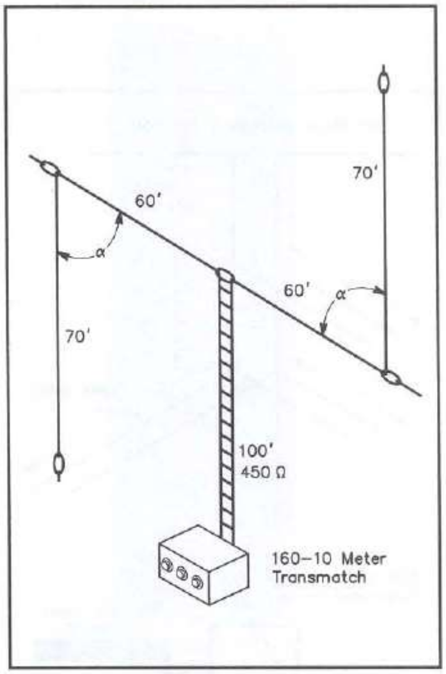

A 10- to 160-meter horizontal Z antenna. Inexpensive no. 14 copper-clad (electric fence) wire may be used. For complete information on open-wire transmission lines, see *The ARRL Antenna Book*, any recent edition.

Although a height of 100 feet (30 meters) is indeed desirable for this antenna, hams who settle for elevations between 30 and 50 feet (9 and 15 meters) will still obtain good results. The angles α between the wire segments will depend on individual situations such as the placement of trees or other supports. Generally, the wider the angle, the better the performance.

W1NH, the New Hampshire SCM, who is really into antennas, says my design is "FB." My evaluation of the antenna is that the "aerial" is great.–*John N. MacInnes III, WB1FPD*

By Taft Nicholson, W5ANB/AAR6AG

From *QST*, November 1981

Compact Multiband Antenna Without Traps

Looking for an inexpensive, easy-to-build, portable antenna to use with your tube-type transmitter? This first cousin of the G5RV may be the aerial of your dreams.

Do you need an antenna for portable operation? I do! My preference is for making things as simple as possible. I don't like traps and matching units. The antenna shown in Figure 1 has no traps, needs no matching unit for the 10-, 20- and 40-meter bands (when used with vacuum-tube PAs), is lightweight and installs easily. What more could you ask for?

Construction

Construction is simple. Cut two lengths of stranded copper wire (such as Radio Shack 278-1292) to 44 feet, 2 inches each (m = ft × 0.3048). Attach 36 feet, 8 inches of 300-ohm twin lead as shown in Figure 1. Coaxial cable attaches to the other end of the twin lead at the points marked A and B in the diagram. I wound 7 feet, 2 inches of RG-58/U coaxial cable into an rf choke to minimize problems with rf flowing on the outside of the coaxial cable. This length of cable in the choke evolved from an attempt to match the antenna to the transmitter for operation on 15 meters.

Alternatively, you could use open-wire feeders in place of the twin lead. If you choose that method, you will need to make the section 42 feet, 6 inches long. Or you could attach the twin lead or openwire conductors to a matching unit. When using a matching unit, there is no particular merit in the lengths given in the diagram.

SWR

The SWR of the antenna is less than 3:1 on 10, 20 and 40 meters. I have no difficulty loading transmitters with tube-type PAs (e.g., Galaxy V, Swan 350 or Drake T4X). This antenna will work with some transmitters on 15 meters, but tuning is quite critical.

However, 80-meter operation is a problem. I have not been able to obtain full output power with these transmitters on frequencies below 3.750 MHz. The SWR measures between 5:1 and 8:1 for the lower part of the band. I constructed a loading coil to go between the choke and the twin lead (Figure 2). The coil consists of 44 inches of twin lead wound on a 7/8-inch diameter form (my left thumb). Once the coil was wound, I removed it from my thumb and used electrical tape to secure it. For 80-meter operation, attach ends C and D to the choke and ends E and F to the twin lead. Remove the coil for operation on the other bands. Banana plugs and sockets can be used to facilitate the insertion and removal of the coil.

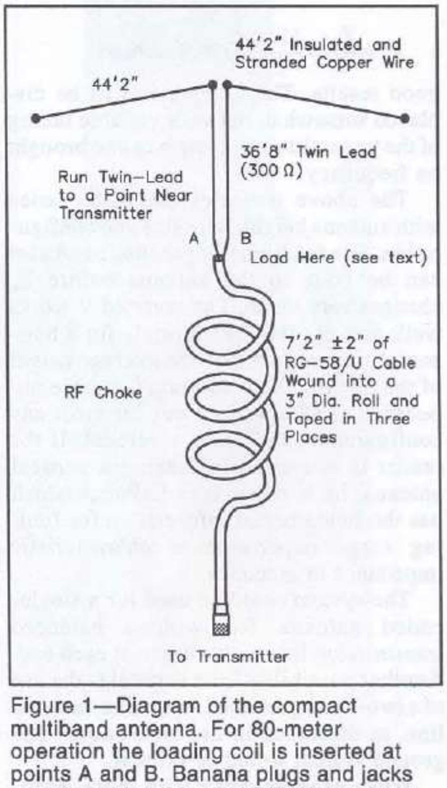
44'2" Insulated and Stranded Copper Wire
44'2"
36'8" Twin Lead (300 Ω)
Run Twin–Lead to a Point Near Transmitter
A
B
Load Here (see text)
7'2" ±2" of RG–58/U Cable Wound into a 3" Dia. Roll and Taped in Three Places
RF Choke
To Transmitter

Figure 1—Diagram of the compact multiband antenna. For 80-meter operation the loading coil is inserted at points A and B. Banana plugs and jacks may be added here to facilitate insertion and removal of the loading coil.

Installation

The antenna should be as high as practical. I've had satisfactory results with the center of the antenna only 25 feet above ground, with the ends tied to fences or other convenient supports. Telescoping TV mast sections make a good support if nothing else is available. The legs of the antenna serve as two of the guy wires. One or two additional guy supports should be added (nonconducting material such as nylon rope is best).

This compact multiband antenna works satisfactorily on all bands from 20 through 80, and 10, meters. It has no traps and requires no matching unit when used with tube-type equipment. I have used it for portable operation in and out of the country. It is easy to pack, carry and erect. Perhaps you might want to try one. I think you'll like it! A brief discussion of the theory of operation follows in the appendix.

[Editor's Note: A description of the G5RV appears in the RSGB *Radio Communications Handbook*. Gray described it in the June issue of *Ham Radio Horizons*. A similar design was depicted in the Collins Radio manuals of the 1930s.]

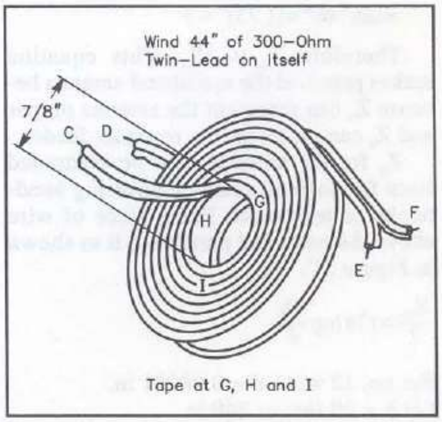

Figure 2—Loading coil for 3500-3750 kHz operation. Points E and F on the coil are connected to the coaxial cable. Points C and D are connected to the twin lead.

Appendix

This multiband antenna evolves from two connected transmission lines with critical length and ratios of surge impedances. The system is self-resonant at a fundamental frequency and at most of the even harmonics and several of the odd harmonics. The first five are 2nd, 4th, 5th, 7th and 8th.

Consider the transmission lines in Figure 3.

The two lines are of equal length, "ℓ" and different surge impedances, Z_o and Z_s.

Looking into the lines:

$$X_o = \frac{Z_o}{\tan(2\pi\ell/\lambda)} \text{ and } X_s = Z_s \tan(2\pi\ell/\lambda)$$

where

λ = wavelength

X_o = reactance looking into open line

X_s = reactance looking into shorted line

From the theory of resonant circuits, we know if we connect the lines the system will be resonant at all frequencies where $\overline{X}_0 = \overline{X}_S$ provided the two reactances are of equal value and opposite signs. The open and shorted line provide this condition except at some harmonics.

Joining the lines as depicted in Figure 4 we find that

$$\frac{Z_o}{\tan(2\pi\ell/\lambda)} = Z_s \tan(2\pi\ell/\lambda)$$

$$\frac{Z_o}{Z_s} = \tan^2(2\pi\ell/\lambda) = \tan^2(360°\ell/\lambda)$$

If the angle $2\pi\ell/\lambda$ is made 60°, then the amplitude of the tangent at 120° or second harmonic will be the same. This will be true for 240° (4th harmonic) and 300° (5th harmonic). Similarly, these harmonic responses will continue at discrete angles above 360°, e.g., 7th, 8th, 10th, 11th, 13th, 14th and so on. The signs of the tangents wash out when squared.

The angle $2\pi\ell/\lambda$ becomes 60° by making $\ell = \lambda/6$ (1/6 of a wavelength) at the fundamental frequency.

$$\frac{Z_o}{Z_s} = \tan^2(360° \times 1/6)$$

$$= \tan^2 60° = (1.73)^2 = 3$$

Therefore $Z_o = 3Z_s$. This equation makes practical the multiband antenna because Z_o can represent the antenna proper and Z_s can represent the resonant feeder.

Z_o for the antenna may be computed from formulas in radio engineering handbooks or textbooks. For a piece of wire above the earth and parallel to it as shown in Figure 5C.

$$\frac{Z_o}{2} = 138 \log \frac{4h}{d}$$

For no. 12 wire, d = 0.08081 in.

Let h = 20 feet or 240 in.

$$\frac{Z_o}{2} = 138 \times \log\left(\frac{960}{0.08081}\right)$$

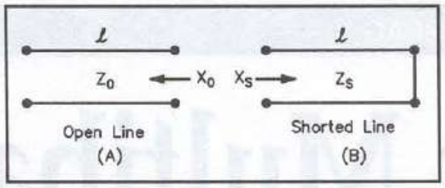

Figure 3—Two identical lengths of transmission line. Looking into the lines, the opposite ends are open and shorted at A and B, respectively.

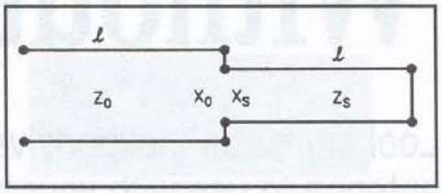

Figure 4—The two transmission lines from Figure 3 are joined to form one line.

138 log (11879) = 562 ohms

This value is not critical. One can use 300-ohm twin lead or 400-ohm open line with

If the wire size had been no. 18,

$$\frac{Z_o}{2} = 605 \text{ ohms}$$

If we use no. 12 wire, then we would have 562 ohms on one side and + 562 ohms on the other side.

$$Z_o = \frac{Z}{2} + \frac{Z}{2} = 1124 \text{ ohms}$$

From the formula $\frac{Z_o}{Z_s} = 3$

$$Z_s = \frac{Z_o}{3} = \frac{1124}{3} = 374.6 \text{ inches}$$

good results. The harmonics will be displaced somewhat, but with variable tuning of the transmitter the system can be brought on frequency.

The above indicates that $Z_o/2$ varies with antenna height, wire size and configuration. The function is logarithmic and a lot can be done to the antenna before Z_o changes very much. The inverted V works well; just use the $Z_o/2$ formula for a horizontal wire and let h be the average height of the inverted V. A formula for surge impedance can be worked out for most any configuration, including a vertical. If the reader is interested in feeding a vertical antenna, he is referred to LaPort,[1] which has the fundamental information for finding surge impedance or characteristic impedance of antennas.

The system could be used for a single-ended antenna fed with a balanced transmission line with a balun at each end. Another possibility for a vertical is the use of a two-wire, grounded, open transmission line, as discussed in LaPort's book.[2] The ground system would be critical.

When experimenting with these multiband lines, it is convenient to have some "stock" numbers to apply to the lines (see Figure 6). One-sixth of a wavelength is one-third of a half wavelength. A convenient length for a half wave on 80 meters is 135 feet. One-third of that is 45 feet or λ/6 for 80 meters. One-sixth of a wavelength on 40 meters is 22-1/2 feet. When you are designing an antenna, these lengths need to be multiplied by the propagation constant of the line. After construction and testing, the dimensions can be pruned for end effect, etc.

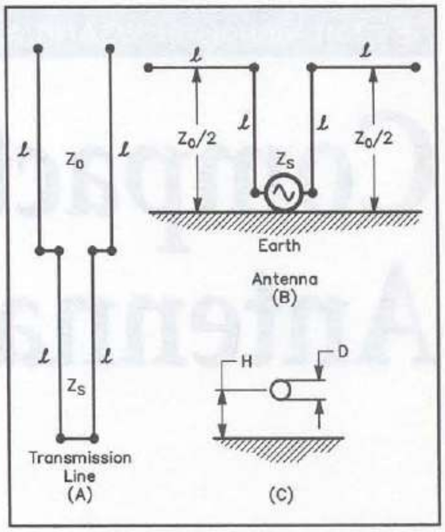

Figure 5—At A and B, the open portion of the transmission line evolves into the flat-top portion of the antenna. At C, diagram illustrating the formula for calculating Z_o for the antenna.

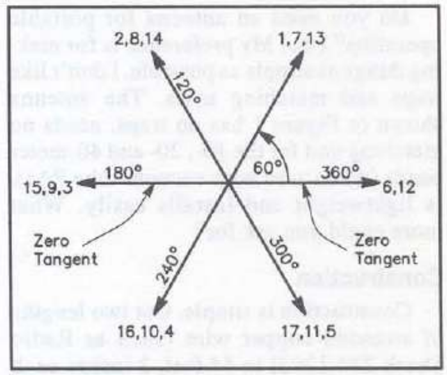

Figure 6—Angular position chart useful for determining "stock numbers" to apply to the chart.

When operated as a transmission line, the system as described may have application in end-feeding half-wave antennas, especially two half waves in phase. The system transforms a high impedance to a low impedance as a quarter-wave line will; however, it will do this at several even harmonics, in contrast to the quarter-wave line that is only responsive to odd quarter wavelengths.

The author wishes to thank Walt Maxwell, W2DU, for his detailed analysis of the theory section of this article.

Notes

[1] E. LaPort, *Radio Antenna Engineering* (New York: McGraw-Hill Book Co., 1952).

[2] See note 1.

By Bill Wright, G0FAH **From *QST*, June 1995**

Five Bands, No Tuner

Enjoy some of the advantages of a multiband, ladder-line-fed antenna without an antenna tuner.

Reading "The Doctor is IN" (*QST*, January 1995) reminded me that the search continues for a simple backyard antenna. A wire *dipole* antenna fed at the center with 450-Ω ladder line is a good choice. The ladder line keeps your losses low—even at moderately high SWRs. All you need is an antenna tuner and you're in business. No coils or traps necessary.

But can you do away with the tuner and still keep the ladder line? That would certainly make life simpler. To achieve this, your transceiver needs to "see" an impedance that looks reasonably close to 50 Ω on as many bands as possible. Without an antenna tuner acting as the middleman between the 450-Ω ladder line and your 50-Ω radio, this could be a problem.

Table 1

Calculated SWRs for a 94-foot Dipole Fed with 41 Feet of 450-Ω Ladder Line

Note: Although the antenna is cut for the CW portions of the band, expect similar results at other frequencies.

Frequency	*SWR*
3.56	7.6:1
7.1	2.4:1
14.2	1.5:1
18.1	2:1
24.9	1.5:1
29	2.4:1

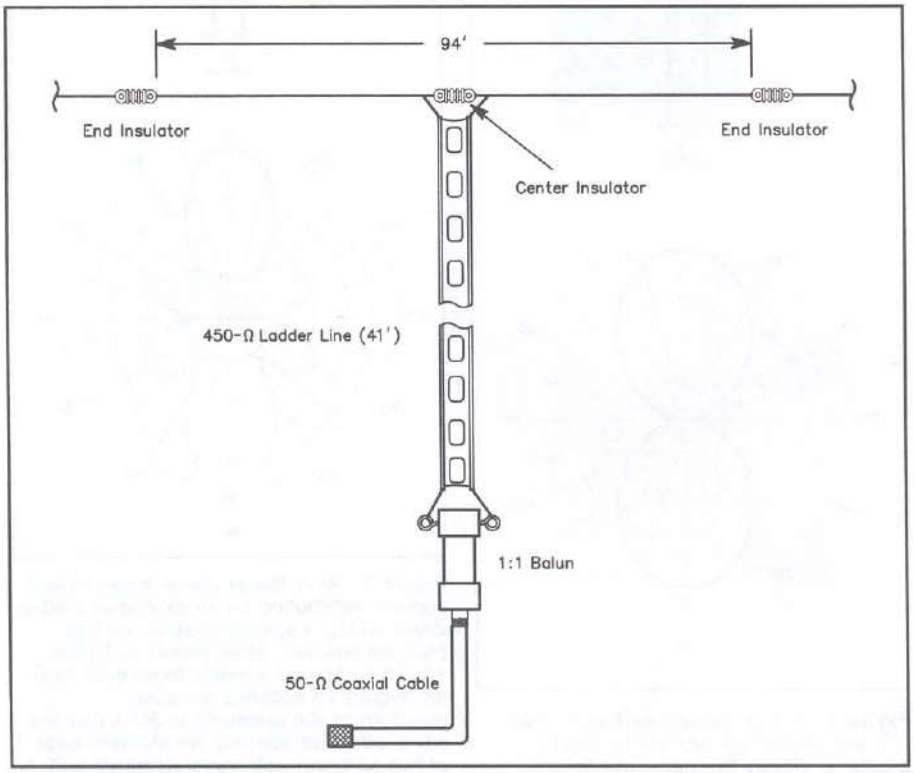

Figure 1—The wire dipole antenna is 94 feet in length. If you don't have 94 feet of open space, don't hesitate to droop the ends of the dipole to make it fit. Feed the antenna with 41 feet of 450-Ω ladder line that is connected to a 1:1 balun. From the balun to your radio, use 50-Ω coax.

Can It Be Done?

A few years ago I attempted to design an antenna that would work on several HF bands from 80 to 10 meters. Full details were published in the spring 1992 edition of *SPRAT*, the journal of the G-QRP club.

My inspiration was the venerable G5RV. I took a 94-foot-long dipole and fed it with ladder line (see Figure 1). By cutting the ladder line to a specific length and using a 1:1 balun to make the transition to coaxial cable, I found that I could get close to 50-Ω (and thus achieve reasonably low SWRs) on at least five bands: 40, 20, 17, 12 and 10 meters (see Table 1).

The on-air results were better than I expected. My radio was happy and I didn't need to meddle constantly with an antenna tuner.

Of course, you'll need an antenna tuner to work the bands where the SWR exceeds 3:1.

A simple tuner will do the job, though. Because you're using *unbalanced* coax ahead of the balun, you won't need one of the more expensive tuners designed for *balanced* feed lines.

For best results, put your antenna as high as possible. If the ends must bend downward to accommodate the size of your lot, don't worry. Run the ladder line to your balun and take your coax from there to your radio. Keep the coax portion as short as possible.

Conclusion

By eliminating the antenna tuner completely, you lose the flexibility of loading your ladder-line-fed antenna on virtually any band. In return, however, you gain the convenience of operating on several bands without making tuner adjustments each time you change frequency. Your losses are held to a minimum, which means that most of the power your radio generates is radiated by your antenna. Not a bad compromise! Of course, the results I achieved will vary when used at other locations. Still, it's a simple, fun project in the experimental spirit of Amateur Radio.

By Rick Olsen, N6NR **From *QST*, March 1993**

The NRY: A Simple, Effective Wire Antenna for 80 through 10 Meters

Known as a *broadside collinear curtain array*, this antenna is simple to build, rakes in DX signals and has gain over a dipole on all the bands it covers!

Hams' thoughts ultimately turn to antennas. So it was with me a while ago. I had just moved to the emerald beauty of the Pacific Northwest and began pondering the 160-foot-tall Douglas firs in my backyard. I had considered putting up a crank-up tower, but decided that neighbors, architectural committees and crank-ups don't mix. Then I thought about topping a tree. Nope–I moved here because I love these trees.

I started sniffing around for alternatives. I obtained a copy of fellow ARRL Technical Advisor Roy (W7EL) Lewallen's antenna modeling program, *ELNEC*, along with a copy of *The ARRL Antenna Book*, and began to incubate the seeds of what would become my most enjoyable antenna project since I built a four-element quad back in 1974.

Hanging ropes from big trees is not a new idea.[1,2] I began by pacing off the distances between the monsters in my yard so I would know how much space I had to work with and how I could optimally direct my signal to places like Europe, Africa and the Caribbean.

Next came the selection of a configuration. Because the trees are approximately 120 feet apart, my first choice was a multielement collinear array. But a good friend, Terry Conboy, N6RY (in *NRY*, I'm the *NR* and he's the *RY*), suggested stacking a pair of *extended double Zepps*. I filed his suggestion away in my memory for the time being. I started my analysis.

The ARRL Antenna Book, on page 8-32, describes the basic function of the collinear array. In sum, it concentrates energy perpendicular to the wire, hence the term *broadside array*. This antenna's physical layout and pattern are shown in Figure 1. Such an antenna's gain can be increased by varying the end-to-end spacing of the elements.[3]

Chapter 8 goes on to explain three- and four-element arrays. Disappointment set in when I sized a four-element in-line array for 20 meters: It wouldn't fit between the trees! What *would* fit, however, was a "two-over-two." In fact, depending on the vertical separation, this array might even have some gain over the four-element collinear arrangement.[4]

My choice was made. I wanted to maximize performance on 20 meters and take what I could get on the other bands. I chose two 1-λ, center-fed wires vertically spaced 5/8 λ apart. According to *ELNEC*, with the array 100 feet high at the top and modeled

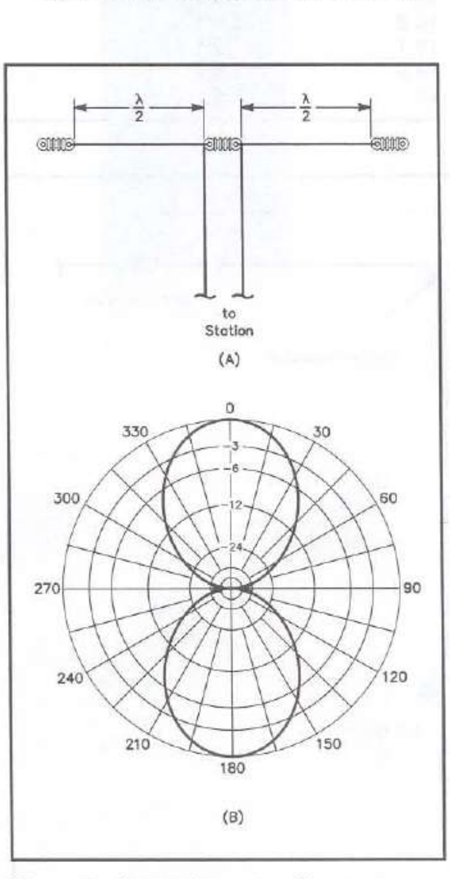

Figure 1—A two-element collinear array (A), also known as *two half-waves in phase*, is merely two half-wave wire elements placed end to end. This antenna has about 1.5 dB gain over a dipole in the same surroundings, and B shows its free-space pattern.

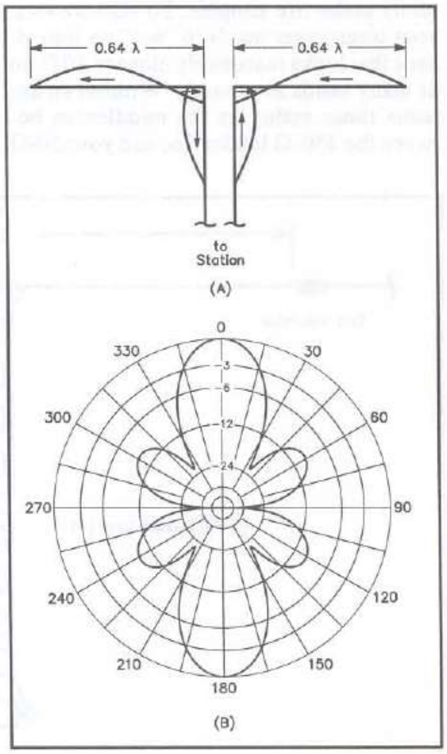

Figure 2—At A, the physical layout of and current distribution on an *extended double Zepp* (EDZ), a special case of the two-element collinear array shown in Figure 1A. This antenna exhibits more gain than the Figure 1A antenna because lengthening the elements to 5/8 λ has the same effect as spacing the element ends of two collinear half-wave elements 0.28 λ apart. (See Note 3.) At B, the antenna's azimuth-plane pattern in free space. This antenna has about 3 dB gain over a dipole in the same surroundings.

over lossy ground, this antenna's gain is approximately 12 dBi.[5] Not bad! The only problem is that this antenna's 20-meter feedpoint impedance is very high.

Reenter N6RY and the extended double Zepp. Take a look at Figure 2A. The extended double Zepp is simply a 1.25-λ, center-fed wire (two end-to-end 5/8-λ elements). There's a hidden message here! Have you found it?

Well, at first, I didn't either. Terry explained to me that the extended double Zepp (EDZ) is essentially two half-wave collinear elements spaced just over 1/4 λ apart. The extended double Zepp's added gain over two half waves in phase comes *not* from the extra 0.28 λ of wire between them, but from the *increased end-to-end spacing of the outermost pair of halfwave elements.* (Radiation from the inner 0.28-λ actually works *against* that from the outer halfwaves, but the gain increase from the new element spacing more than overcomes this disadvantage.) Figures 1B and 2B are similar. The EDZ has broader lobes and useful smaller lobes, both of which are attractive for a fixed array.

Sure enough, it was right in front of me all the time in the *Antenna Book* (page 8-34, fifth paragraph). In fact, the *whole darned array* was right in front of me in Chapter 8! I just needed some help piecing it all together. I modeled this array every way I could imagine, looking for the best possible performance. My wife didn't see me in the evenings for two weeks! Figure 3 shows the final configuration.

Figures 4, 5 and 6 show some of the results of my modeling. I used a design frequency of 14.2 MHz, a height of 100 feet above ground (for the top wire), elements made of #14 copper wire, and ground coefficients typical of the Pacific Northwest. All the elements in this array are fed in phase, as described in a bit.

As the figures show, the antenna's gain and patterns are impressive. On the 80- and 40-meter bands, the antenna has remarkably dipole-like patterns and gains and seems to work at least as well as a dipole in the same height range. On the higher bands, it works *much* better than a dipole. Its predicted gains in my installation are 9, 11 and 14 dBi at 7, 10.1 and 14 MHz, respectively. At 18.1 MHz (Figure 6) and above, the array begins to take on long-wire properties. It has several major lobes in directions closer to the wire axis, which the lower-frequency patterns don't provide.

To help you evaluate the NRY's patterns, Figures 4-6 superimpose the patterns of halfwave dipoles cut for 10.1, 14 and 18.1 MHz, respectively, on the NRY's patterns. These comparisons assume that the dipole is the same height as the NRY's feed point (halfway between the elements, or about 22 feet lower than the top wire), and over the same ground.

You may wonder about this antenna's effectiveness as a function of height. According to the *Antenna Book*, arrays like this have the best gain when the lower element is at least 1/2 λ above the ground (33 feet at 14 MHz). And, of course, as with all horizontally polarized antennas, this array's radiation angle decreases as you raise the antenna, which makes for better DX performance. The bottom line is that for DX work, you don't have to get the array up 100 feet for it to work well—you just need to put it up as high as you can. For closer-in coverage, it will work well at lower heights.

Another advantage of this array is that its feed-point impedances are manageable. To maintain my goal of broadband use, I decided to feed the antenna as shown in Figure 3. This is the "lazy-H" configuration described starting on page 8-37 of *The ARRL Antenna Book.* The attractiveness of this configuration is that it doesn't rely on specific phasing-line lengths to work properly. To the extent made possible by the different mutual impedances between each element and the ground, the phase relationship between elements remains constant regardless of frequency.

Final Dimensions

The antenna's physical dimensions (86 feet, 10 inches long and 43 feet, 5 inches high, not counting insulators) are easy to remember because the height is half the length. Because the antenna is basically 1.25 (5/4) λ long, center-fed and vertically spaced 5/8 λ apart, you can easily scale it to whatever band you like. For instance, to scale it to 10 meters, divide all the dimensions by two (28 MHz/14 MHz = 2).

Construction

Now for the really fun stuff: putting this thing up in the air! Building the antenna is a breeze. It took me all of about an hour to cut and solder the wires and attach the tethers. The harder part was preparing the trees.

If you have a friend like Bernie Olshausen, N6RUX, you have it made. He's a crack shot with a slingshot, similar to that described by Wade Calvert, WA9EZY, in *QST* a couple of years back.[6]

Once Bernie landed the fishing lines over the right tree limbs, the next job was to pull up some lightweight twine (we used seiner twine—the stuff fish nets are made of) with the fishing line. This is important; rope sometimes snags as it goes over limbs, and you don't want to try to pull it over with 6-pound-test fishing line! With the twine, we pulled up some 800-pound-test nylon rope with a pulley on one end, using another piece of nylon rope through the pulley to act as a catenary for the array.

It's important to use a strong, weather-resistant catenary rope to hang the array, because the antenna loses its desirable properties quickly when it takes on the shape of a V. You can expect a 2-dB gain

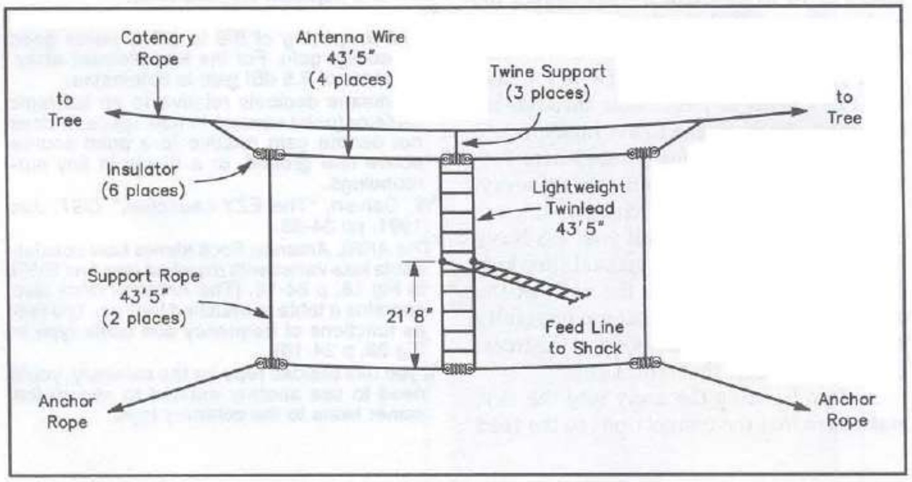

Figure 3—The broadside collinear curtain array—the NRY—in its final configuration. This antenna is cut for 20 meters, but works well on all the HF ham bands. The array hangs from a catenary rope via three short lengths of twine. The lower end insulators are supported by ropes to those of the top wire, and the center insulators are joined by the phasing line, which is 43 feet, 5 inches of parallel-wire feed line. The feeder to the radio is made of the same type of line connected at the halfway point on the phasing line.

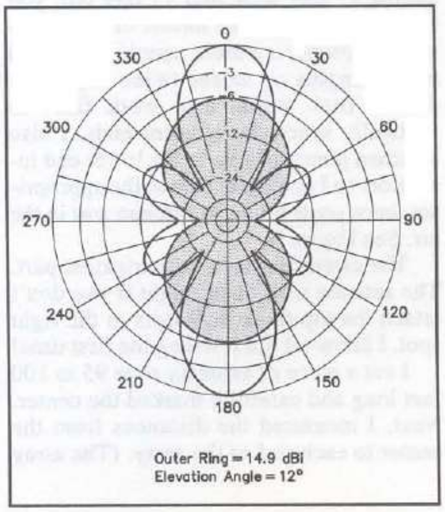

Figure 4—The NRY's 14-MHz azimuth-plane pattern with the bottom wire at 56.5 feet and top wire at 100 feet. For comparison, the shaded part shows the pattern of a half-wavelength, 14-MHz dipole at 78 feet (the same height as the NRY's feed point).

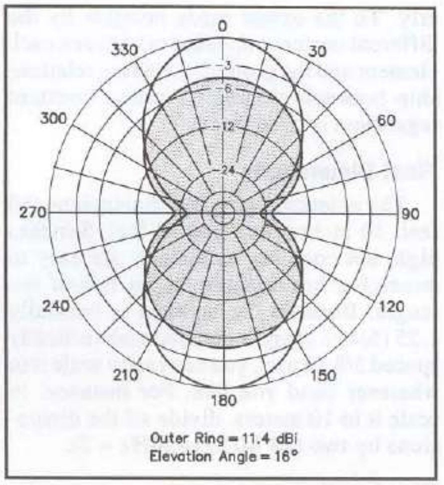

Figure 5—The NRY's 10.1-MHz azimuth-plane pattern with the bottom wire at 56.5 feet and top wire at 100 feet. For comparison, the shaded part shows the pattern of a half-wavelength, 10.1-MHz dipole at 78 feet (the same height as the NRY's feed point).

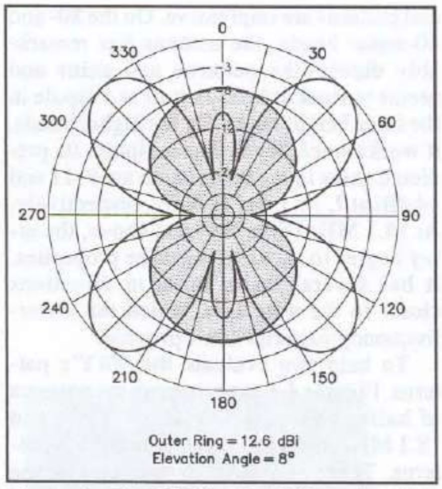

Figure 6—The NRY's 18.1-MHz azimuth-plane pattern with the bottom wire at 56.5 feet and top wire at 100 feet. For comparison, the shaded part shows the pattern of a half-wavelength, 18.1-MHz dipole at 78 feet (the same height as the NRY's feed point).

decrease at 20 meters with a 20° variance from the flat-top configuration. The rope, therefore, must be strong enough to minimize droop and maintain integrity in the sun and wind. We'll get back to the rope in a minute.

For the phasing harness I used low-loss, 300-Ω twinlead. Any parallel-wire feeder should work fine. Don't use coaxial cable in the feed system. It's too heavy, unbalances the array and has very high loss when operated at a high mismatch.[7]

To assemble the feed system, I used porcelain insulators and equal lengths of 300-Ω line. I then cut the #14 antenna wire to the proper length and soldered it to the feed system in a standard dipole configuration.

Next came the tethers. For an antenna almost 87 feet wide and 43 feet tall, you need to provide some means of maintaining the proper element spacings. I used equal lengths of seiner twine (nylon or dacron rope would also work fine) to vertically space the element ends. I also attached plenty of line to the lower end insulators so I could tie them to the appropriate fence posts once the antenna was in the air. See Figure 3.

The catenary rope is the trickiest part. The antenna won't hang right if you don't attach the top three insulators in the right spot. I know—I did it wrong the first time!

I cut a piece of catenary rope 95 to 100 feet long and carefully marked the center. Next, I measured the distances from the center to each end of the array. (The array itself is a handy measuring line.) Then I added about 10 more inches and marked the line. I'll tell you why in a minute.

Then I cut three pieces of twine, each about 24 inches long. At the center of the catenary rope, I carefully opened the rope weave and passed about 6 inches of twine through it, then let the catenary rope retake its natural shape. Wrapping the twine around the rope several times on either side of the joint relieved the tension placed on the rope introduced by passing the twine through it.[8] I then tied the other end of the twine to the top center insulator so that the spacing between the rope and the insulator is 12 inches.

After doing that, I attached the twine at the end marks on the catenary rope in the same manner. The distance to the end insulators is 18 to 20 inches. This causes the ends to hang off the rope at an angle and provides strain relief.

If you use a braided rope (with a weave you can't open to pass twine through it), one good alternative is to use insulators in the catenary rope (at the points where I attached the twine) as tie points for the array.

Now you're ready to haul the antenna up on the lanyards. Consult your US Navy *Marlinspike Seamanship* manual (just kidding, of course) and join the ends of the catenary to the lanyards that are hopefully by now through the pulleys, up in the trees. Use a strong knot that won't slip.

Before hoisting the array into the sky, *make sure* that the connections to the feed harness at the top and bottom are the same (ie, that the left elements both attach to the same wire in the twinlead). This thing does not work right if the top and bottom are fed 180° out of phase!

Now you're ready to haul it up and connect the feed line to the antenna tuner. Check to make sure that the symmetry and top-to-bottom wire spacings are within reasonable limits, then tie the bottom tethers off so that the bottom wire is as flat as you can make it.

You're Done!

This is quite an effective antenna. I've installed two of them: One is boresighted on the Middle East and works like gangbusters into Europe as well; the other favors the Caribbean and most of the US. In the first three months after putting them up, I worked 105 countries on CW, including some rare ones, with a 100-watt transceiver. I'm surprised at how easily I can bust pileups—especially on 30 and 20 meters.

I hope you have an opportunity to try this array on for size. Don't limit yourself to just the backyard, either: It's a great Field Day antenna, to say the least. It also comes in handy for emergency work where you need a gain antenna.

You don't have to be a rocket scientist to design effective antennas! Your old friend *The ARRL Antenna Book*, and other ARRL publications, like Walt (W2DU) Maxwell's *Reflections*, the *Handbook*, the *Antenna Compendium* series and Wilfred Caron's *Antenna Impedance Matching*, are serious reference guides and sources of numerous ideas. It doesn't hurt to have your favorite antenna-modeling software loaded on your PC, either!

Notes

[1] D. Brede, "The Care and Feeding of an Amateur's Favorite Antenna Support—the Tree," *QST*, Sep 1989, pp 26-28, 40.

[2] J. Hall, ed, *The ARRL Antenna Book*, Chapter 22 and pp 4-3 and 15-3.

[3] *The ARRL Antenna Book*, Fig 8-38, p 8-32, illustrates how gain varies with the spacing of the two adjacent element ends.

[4] The *Antenna Book*'s Fig 8-45, p 8-35, suggests that a spacing of 5/8 to 2/3 λ yields good broadside gain. For the four-element array, upwards of 8.5 dBi gain is obtainable.

[5] dBi means decibels relative to an *isotropic radiator* (point source) in free space. It *does not* denote gain relative to a point source above real ground, or a dipole in any surroundings.

[6] W. Calvert, "The EZY-Launcher," *QST*, Jun 1991, pp 34-35.

[7] *The ARRL Antenna Book* shows how coaxial-cable loss varies with matched loss and SWR in Fig 18, p 24-18. (The *Antenna Book* also contains a table of matched loss per 100 feet as functions of frequency and cable type in Fig 26, p 24-18).

[8] If you use braided rope for the catenary, you'll need to use another method to secure the seiner twine to the catenary rope.

CHAPTER FOUR

LOOP ANTENNAS

4-1 **Dual Full-Wave Loop Antennas**
From March 1982 *QST*

4-2 **A Gain Antenna for 28 MHz**
From July 1994 *QST*

4-3 **10-Meter "Lazy Quad"**
From July 1968 *QST* (Technical Correspondence)

4-4 **The Full-Wave Delta Loop at Low Height**
From October 1984 *QST*

4-7 **A Triband 75/40/30-Meter Delta Loop**
From *The ARRL Antenna Compendium Vol 5*

4-9 **A Two-Band Loop for 30 and 40 Meters**
From March 1989 *QST* (Hints & Kinks)

4-10 **Feeding an 80-Meter Delta Loop at 160 Meters**
From April 1990 *QST* (Hints & Kinks)

By John Griggs, W6KW **From *QST*, March 1982**

Dual Full-Wave Loop Antenna

Achieving gain from an antenna at 7 and 3.5 MHz normally requires a rather large piece of real estate, a high tower or both. To obtain significantly improved performance over a dipole at these frequencies, and to do it inside an average city lot, is a goal worth pursuing! With this in mind I decided to replace my inverted-V dipoles on 7 and 3.5 MHz with a dual full-wave loop antenna, one inside the other.

I had noticed the strong signals on 7 MHz from Pat Kearins, W7UI (now a silent key), and from Carl Winter, W6OAW, on 3.9 MHz, and was interested to find that both were using full-wave loops. Both signals would pin my S-meter and my QTH is well over 200 miles from either station. The antenna at W7UI was a loop suspended by a very high supporting structure. This loop resembled a square, with one corner facing up, one facing the ground, and the other two corners pulled outward by guy wires.

My attention was attracted to a horizontal loop antenna used by another ham on 3.9 MHz that produced extremely strong signals. It was a square loop, 65 feet on a side, with each leg parallel to the ground but only 14 feet high. Fed by a tuned line, it functioned well, but it was most effective for relatively short ranges, up to 200 miles or so.

Of more interest to me, however, was the 75-meter rectangular loop used at W6OAW. I learned that this was a dual antenna, i.e., a full-wave rectangular loop for 3.9 MHz suspended about 40 feet above ground, with another full-wave loop for 7 MHz inside the first, in the same plane. The 3.9-MHz loop is a closed circuit, whereas the 7-MHz loop is an open circuit. This permits operation on 10, 15 and 20 meters as well as 40 meters when used with open-wire feed line and a matching network. The outer loop is fed with coaxial cable.

Design Planning

In considering loop antennas for my lot, which is 75 feet wide by 125 feet deep, I found that I could use two 40-foot high pipe masts, 130 feet apart along a diagonal across my lot. I duplicated the W6OAW arrangement, except that I hung the antenna vertically and made both loops closed-circuit designs. I could use them only on the bands for which they were cut, 7.2 MHz and 3.8 MHz. Each is fed with a quarter-wave matching section of RG-59/U, which provides a 50-ohm match for the RG-8/U cables leading in to my operating position.

Construction

See Fig. 1 for construction details and dimensions. Matching the 104-ohm antenna impedance to the 50-ohm line impedance requires a quarter-wave coaxial line having a characteristic impedance of 72 ohms. The formula for a ¼-wave section is:

$$\frac{246}{f(\text{MHz})} = \ell(\text{ft}) \qquad \text{(Eq. 1)}$$

This result must be multiplied by the velocity-factor of the coaxial cable, 0.66 for RG-59/U 73-ohm cable. The quarterwave section was determined to be 42.7 feet for 3.8 MHz and 22.6 feet for 7.2 MHz. These connect to RG-8/U cables for a 50-ohm match to the transceiver.

The top section of the 3.8-MHz loop is fed at the center because of the length of the quarter-wave line, but the shorter matching section for the 7.2-MHz loop allows me to feed the bottom portion of that antenna.

The inner loop is supported from the corners of the outer loop by means of nylon rope and suitable insulators. It is best to use lightweight, high-quality insulators and wire no larger than no. 14 to reduce the weight of the array. Pulleys (and rope) will be required at the tops of the 40-foot poles and also at the 18-foot level. The fact that the bottom horizontal section of the antenna narrowly misses the top of my house roof seems to have no deleterious effect on its operation.

Measure the antenna sections carefully, and cut them to length. Pay particular attention to locate the feed point at the exact center of the horizontal wire sections. This provides horizontal polarization. Feeding the loop at the center of either vertical section will provide vertical polarization.

Tuning

If trimming is needed to resonate the antenna to your favorite operating frequency, cut or add equal length pieces at each vertical section. Do not shorten the sides to the point at which the top and bottom sections are less than 0.1 wavelength apart. Careful pruning of the horizontal sections would be required if such a problem develops.

Information given to me by those using this antenna seems to confirm my results. A typical comment is "Boy are you ever loud!" A gain of 1.8 dB over a dipole is claimed for this type of antenna. The principal difference appears to be a lower angle of radiation. While it is possible to work stations off the ends of the antenna, maximum radiation occurs broadside to the element. What I like about the antenna is the ability to hear weak stations. It is outstanding in this regard.

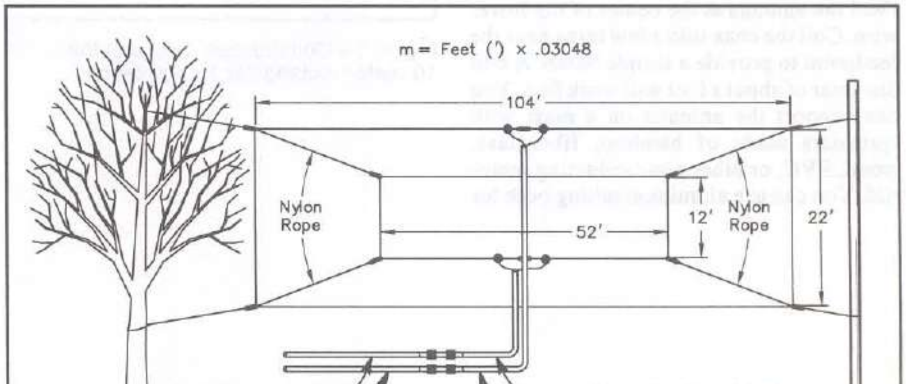

Fig 1—Dimensions for the dual full-wave loop antenna. Input impedance is on the order of 104 ohms. A quarter-wave matching section of RG-59/U is used to provide a match to 50 ohms.

By Brian Beezley, K6STI **From *QST*, July 1994**

A Gain Antenna for 28 MHz

Give your 10-meter signal a boost with this simple antenna.

Although in coming years the 10-meter band won't provide the excitement it did at the peak of sunspot cycle 22, DX openings will still occur, especially during spring and fall. And who knows? With the large number of 10-meter operators and wide availability of inexpensive, single-band radios, the band may be the hangout for local ragchews that it was before the advent of 2-meter FM.

I'd like to present a simple antenna for 10 meters that provides gain over a dipole or inverted V. Similar designs have been published before, but because I never hear them on the air, I think they must not be fully appreciated. The antenna is a resonant, rectangular loop with a particular shape. It provides 2.1 dB gain over a dipole at low radiation angles when mounted well above ground. This represents a power increase of 62%. The antenna is simple to feed–no matching network is necessary. When fed with 50-Ω coax, SWR is close to 1:1 at the design frequency. SWR is less than 2:1 from 28.0 to 28.8 MHz for an antenna resonant at 28.4 MHz.

The antenna is made from #12 wire (see Figure 1). For horizontal polarization at 28.4 MHz, the loop is 73 inches wide and 146 inches high (just larger than 6×12 feet). Feed the antenna at the center of the lower wire. Coil the coax into a few turns near the feedpoint to provide a simple balun. A coil diameter of about a foot will work fine. You can support the antenna on a mast with spreaders made of bamboo, fiberglass, wood, PVC, or other nonconducting material. You can use aluminum tubing both for support and conductors, but you'll have to readjust antenna dimensions for resonance.

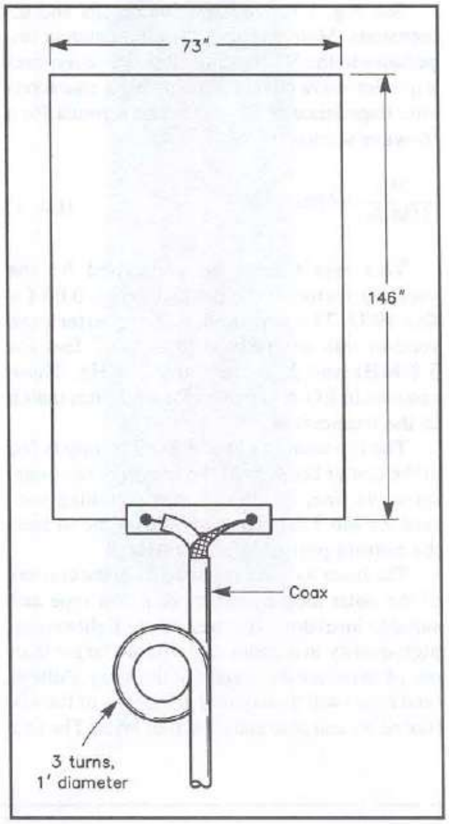

Figure 1—Construction details of the 10-meter rectangular loop antenna.

This rectangular loop has two advantages over a resonant square loop. First, a square loop has just 1.1 dB gain over a dipole. This is a power increase of only 29%. Second, the input impedance of a square loop is about 125 Ω. You must use a matching network to feed a square loop with 50-Ω coax. The rectangular loop achieves gain by compressing its radiation pattern in the elevation plane. The azimuth pattern is slightly wider than that of a dipole (it's about the same as that of an inverted V). A broad pattern is an advantage for a general-purpose, fixed antenna. The rectangular loop provides bidirectional gain over a broad azimuth region.

You should mount the loop as high as possible. To provide 1.7 dB gain at low angles over an inverted V, the top wire must be at least 30 feet high. The loop will work at lower heights, but its gain advantage disappears. For example, at 20 feet the loop provides the same gain at low angles as an inverted V.

A small, 3-element Yagi can provide 6 dB gain over a dipole and great rejection of signals to the rear. If you can install a beam and rotor, you'll find it much more effective than a loop. But for a simple, cheap, gain antenna that can be thrown together quickly, the rectangular loop is hard to beat.

Note: I used the AO 6.0 Antenna Optimizer program to automatically optimize the dimensions of a rectangular loop for maximum forward gain and unity SWR. I used NEC/Wires 1.5 to verify the design with the Numerical Electromagnetics Code.

From *QST*, July 1968 (Technical Correspondence)

10-Meter "Lazy Quad"

Among your readers that like to tinker with antennas there may be some that would like to try the antenna shown in Figure 1. The basic idea came from a station using a somewhat similar configuration on 15 meters, and I make no claim of originating the idea. However, I have never heard of anyone using the antenna on 10 meters. Since it is currently nameless, and has features similar to both a Lazy H and a quad, perhaps it should be known as a "Lazy Quad."

The major advantages of the antenna are: (1) extreme simplicity, (2) feasibility of installation on a light unguyed pole, (3) small horizontal space requirements (as compared to a horizontal dipole), (4) low QRN in receiving (as compared to a vertical dipole), (5) ability to withstand high wind loads and (6) broadband operation.

When looking at loop 1 we see a horizontally-polarized full-wave loop radiating broadside, with maximum currents along the top and bottom horizontal wires H_1 and H_2. When looking at loop 2 we see a similar horizontally-polarized loop with maxiinum current on wires H_2 and H_3. When looking at wires H_1 and H_3 we notice two in-phase horizontal wires spaced a half wave, both carrying maximum current. Minimum currents appear on the vertical portions of the loops where the tuning stubs are inserted.

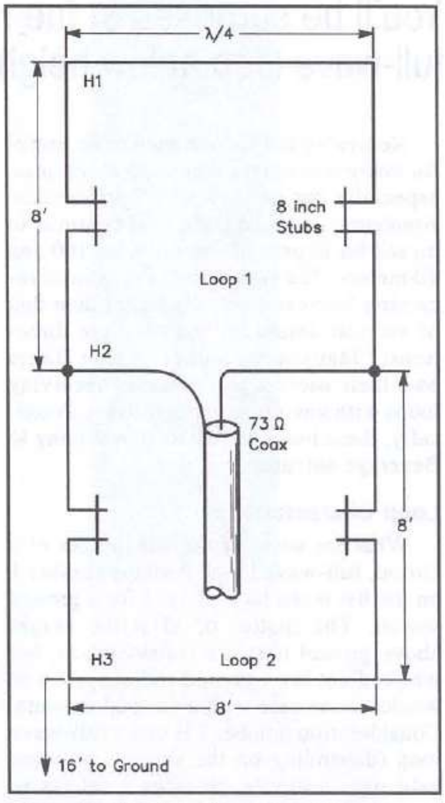

Figure 1—W6PIZ's "Lazy Quad" for 10 meters. The loops are No. 14 wire, with the horizontal sections mounted an 8-foot lengths of 1×2 wood. The adjustable stubs are of the same type wire and are self-supporting since the length is only 6 inches. The antenna at W6PIZ is mounted on an unguyed wooden pole 32 feet high.

I operate mostly on 10-meter c.w. and have the antenna peaked at 28,050 MHz. In my specific case the length of each stub is 6 inches for this frequency. Loading is almost constant from 28 to 30 MHz.

The coax feed line is run straight down the pole (wood) to the ground, and there is little antenna effect on the feeder.

Results have been consistently better than had been obtained with vertical or horizontal dipoles previously installed at the same effective height at the same location. Indications are that it outperforms some beams of the same approximate height, perhaps because of the broad vertical pattern (which allows longer QSOs under critical skip conditions), as compared to the beams.

Perhaps some of your readers might be urged to try this basic idea with a reflector of the same basic configuration as the driven element. It would appear that the Double Lazy Quad would be anything but lazy in operation. — *Dave Hardacker, W6PIZ*

By Doug DeMaw, W1FB and Lee Aurick, W1SE **From *QST*, October 1984**

The Full-Wave Delta Loop at Low Height

Property size and antenna-support height are ever-present concerns of the urban amateur. Many good antennas are untried because the radio amateur is unable to imagine how a large wire antenna could be squeezed onto a small lot. Certainly, this is typical in the case of full-wave loop antennas. But there is no rule that dictates using a symmetrical loop. It can be distorted rather severely without spoiling the performance. The same philosophy is appropriate with regard to height above ground and the plane in which the antenna is erected. In most instances, a less-than-optimum full-wave loop will outperform a dipole or inverted V antenna that is close to the ground in terms of wavelength. It is possible that such a loop will give comparable or better performance than a vertical antenna that is less than 90 degrees (with respect to ground), or one with a substandard ground screen.

We want to discuss the Practical considerations of loops that can be supported from low supports on small pieces of property. The results we have obtained are noteworthy with respect to all-around "solid" communications within and outside the USA. Perhaps you will be inspired to unroll some wire and try a loop at your QTH.

Some Loop History

Loops were used first as receiving antennas. While single- and multiturn small loops worked well for receiving, they were not satisfactory for transmitting: They were inefficient in terms of gain, and the feed impedance was generally a fraction of an ohm, making them difficult to match. The losses were significant. But, it was possible to use a compact loop (less than 0.5 wavelength) for receiving in place of a fullsize version that could require thousands of feet of conductor. One of us owned a portable broadcast-band receiver in the 1930s. The loop antenna was stored in the lid of the cabinet, and needed to be mounted atop the radio during reception periods! The radio was heavy: it weighed 91 pounds, including the various dry batteries.

You'll be surprised at the results you'll get from a full-wave loop at low heights.

Receiving loops continued to be useful for many years in the commercial services, especially for LF and VLF applications. Amateurs also used them (and continue to do so) for improved reception on 160 and 80 meters. The signal-to-noise ratio of receiving loops is markedly better than that of vertical antennas, and they are directional.[1] Many successful 160-meter DXers owe their success to the use of receiving loops with low-noise preamplifiers. Practically, these loops are the next best thing to Beverage antennas.[2]

Loop Characteristics

What are some of the advantages of a closed, full-wave loop? Perhaps number 1 on the list is the lack of need for a ground screen. The matter of effective height above ground is still a consideration, but we need not lay a ground-radial system as would be the case with a vertical antenna. Consideration number 2 is that a full-wave loop (depending on the shape) has some gain over a dipole. Number 3 relates to noise factor. A closed loop is a much "quieter" receiving antenna than are most vertical and some horizontal antennas.

To illustrate this point, the 160-meter antenna at W1FB is a $^3/_8$-wavelength inverted-L with twenty $^3/_8$-wave radials. Since this is essentially a vertically polarized antenna, it is noisy (man-made and atmospheric noise). There are times when an S9 signal is unreadable because of the ambient noise being S9 or greater in strength. Upon switching to the 75-meter Delta loop, the same signal will rise above the noise by 1 or 2 S units, while the noise and signal will drop well below S9. For example, the received signal may drop to S6 on the loop, but the noise will decline to S4.

Feed-point selection will permit the choice of vertical or horizontal polarization. Various angles of radiation will result from assorted feed-point selections. The system is rather flexible when we want to maximize close-in or faraway communications (high angle versus low angle). Figure 1 illustrates various configurations that can be used. The arrangement at C is used at W1SE, and the shape at D is being applied

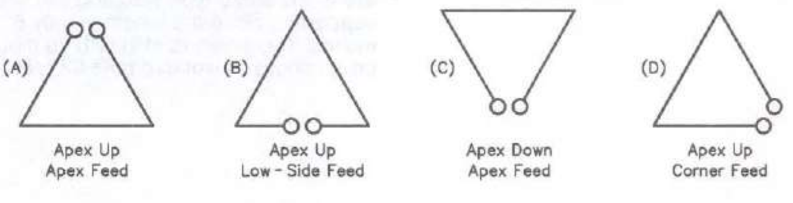

	Polarization	Radiation Angle
A	Horizontal	Moderately High
B	Horizontal	High
C	Horizontal	Moderately High
D	Vertical	Low

Overall Loop Size in feet ≈ $\frac{1005}{1\ \mathrm{MHz}}$

Feed Impedance ≈ 100 Ω

Figure 1—Various configurations for a full-wave Delta loop. Radiation angles and polarization are affected by the feed-point placement and location of the apex.

at W1FB. Both antennas are cut for 80-meter operation. The bandwidth at resonance is on par with that of a dipole. A Transmatch is used for matching the system to the transmitter in those parts of the band (75 and 80 meters) where the SWR is too high to deal with.

Our loops are not deployed in a vertical plane, owing to the lack of tower height. A 60-foot tower and 50-foot tree support the W1SE antenna. A single 50-foot tower is used at W1FB. Both loops are tilted away from the supports at roughly 45 degrees (Figure 2). This shows the present W1FB system. The loop is broadside northeast and southwest for maximum radiation in those directions at 80 meters. More on this later.

When these low-to-the-ground experiments began in the summer of 1983, we were joined by Bill Martinek, W8JUY, near Traverse City, Michigan. Bill experimented with various loop configurations so that he and W1FB could make signal comparisons locally and afar. He finally adopted the W1SE format with the apex down (Figure 1C, with the flat top strung between two 50-foot trees). In order to keep the loop completely vertical (not sloping), he chose a triangle that was not equilateral. The upper side of his triangle is substantially longer than the two downward sides. His signal on 75 meters is consistently 10 to 20 dB stronger than with his inverted V. The point of this discussion is that you need not use an equilateral triangle if it will not fit on your property. Erect whatever you can, then give it a try!

Feed Methods

A Q section is used for feeding the W1SE loop. A Q section is a quarter-wave-length line with an impedance that is somewhere between the antenna feed impedance and that of the feed line. Calculation is a simple matter:

$$Z\ (\text{Q section}) = \sqrt{Z1\ Z2}\ \text{ohms} \qquad \text{(Eq. 1)}$$

where Z1 is the antenna impedance, and Z2 is the feeder impedance in ohms. In this case, assuming approximately 100 ohms for the antenna feed impedance, we would have $\sqrt{100 \times 50} = 70.7$ ohms for the Q-section impedance. This represents a close match to 52-ohm coaxial cable. The Q-section length (made from RG-59/U) can be determined from L(feet) = 246 V/f(MHz), where V is the velocity factor of the coaxial line for the matching section. (The length should be verified using a dip meter.) For operation at the W1SE-chosen frequency of 3.825 MHz, the calculation calls for a Q section of 42 feet 5 inches (Figure 3).

Open-wire feed is used at W1FB (Fig 6B) to permit multiband operation through 10 meters. Unfortunately, a short run of RG-8/U was needed to bring the feed line to the ham station–under the driveway. The coaxial cable was buried in the ground for this reason. A homemade 4:1 toroidal balun transformer (two stacked T200-2 Amidon

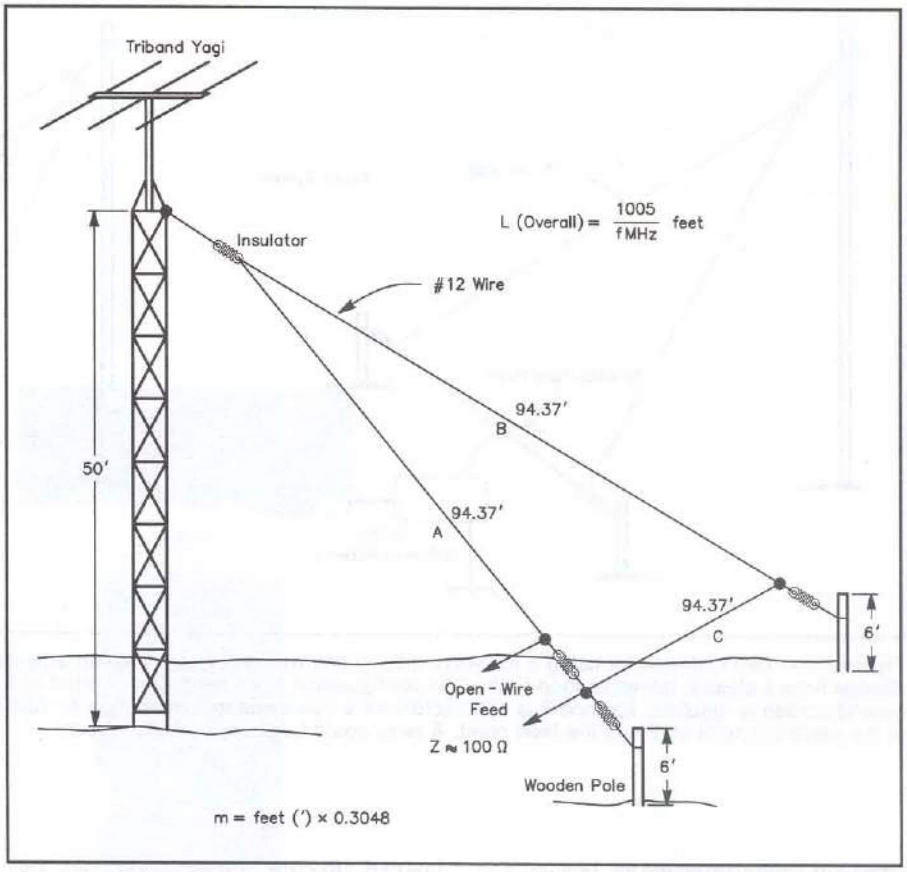

Figure 2—A tilted Delta loop for 80 meters is used at W1FB. The tower height is only 50 feet. Homemade open-wire line is used as the feeder to permit multiband use with vertical polarization and a low radiation angle.

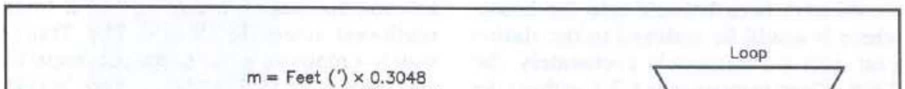

Figure 3—At A is the feed method used at W1SE. A coaxial Q section closely matches the 100-ohm feed impedance to a 52-ohm coaxial line. Illustration B shows the W1FB feed arrangement. Open-wire line, a balun transformer and a short length of RG-8/U cable permit multiband use with a Transmatch. Ideally, the open-wire line would continue all the way to the Transmatch, and the balun transformer would be located at the Transmatch.

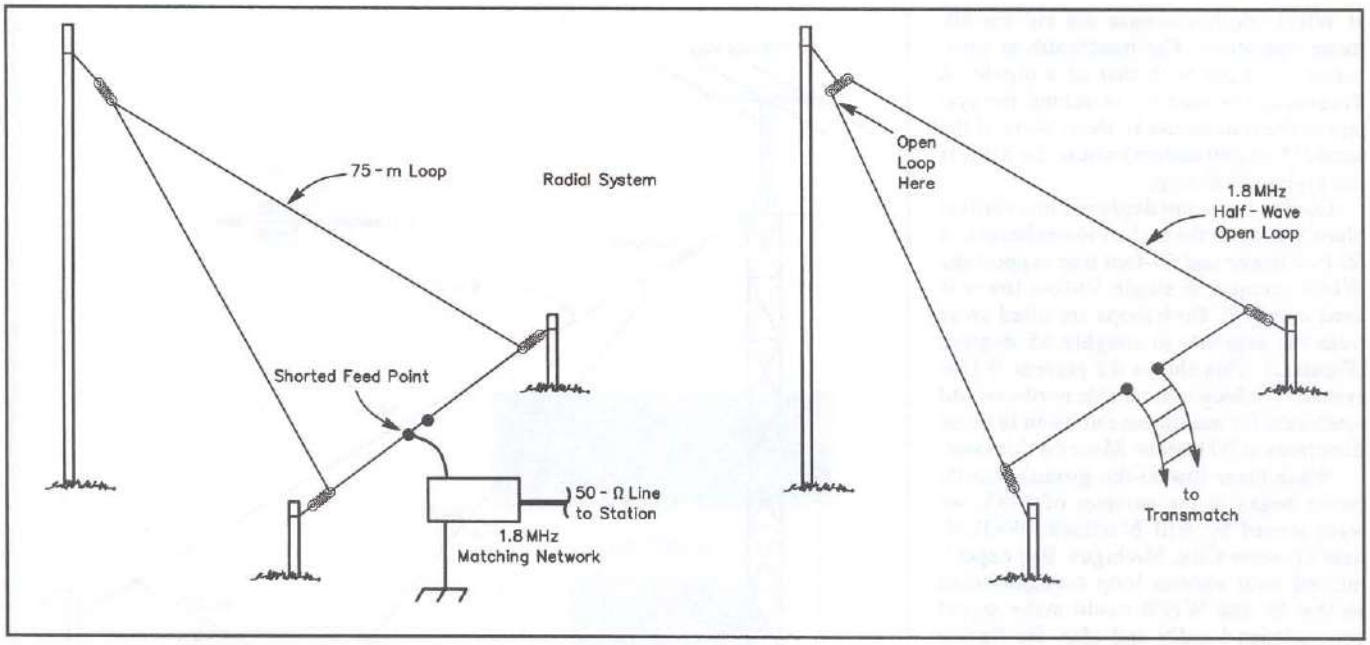

Figure 4 — Two methods for using a full-wave loop at half frequency. A switching arrangement could be applied at the feed point to change from a closed, full-wave loop to the first configuration seen here. The method at A performs as a ¼-wavelength radiator, but a ground screen is required. Method B is satisfactory as a ½-wavelength open loop for half-frequency use. It requires opening the loop at the electrical point opposite the feed point. A relay could be used for this purpose.

cores and Teflon-insulated no. 14 wire) was enclosed in a weatherproof box and mounted on one of the support poles for the 450-ohm open-wire line. The RG-8/U was run underground from that location (about 25 feet). Ideally, the openwire line would have been brought into the house, where it would be matched to the station gear with a Transmatch. Fortunately, the SWR at loop resonance is 1.3:1 without the Transmatch in use.

Performance

This is the part of our article that many of you have been waiting to read. Well, the W1FB results have been entirely gratifying. The loop replaced an inverted V with an apex height of 50 feet. This led to a pronounced improvement in all-around communications on 75 and 80 meters out to 500-600 miles. But, the loop proved to be very effective also for DX communications to Europe on 80 meters. The first version was that of Figure 1B. Although the antenna was outstanding for close-in 75- and 80-meter work, it offered dismal DX performance. The configuration at D of Figure 1 seems to offer a good compromise in performance for local and DX work. The theoretical launch angle to the horizon at the loop fundamental frequency is 10 degrees, as reported by VE2CV in a letter to W1FB. This assumes that the loop is erected vertically and at a reasonable height above ground.

Harmonic operation of the loop, as depicted in Figure 1D, is superb. At times it outperforms the trap tribander atop the tower during DX operation to Europe and Africa. The loop shows an average 6-dB signal increase on 20 and 15 meters in the favored direction, owing to the gain and lower radiation angle of the loop. Radiation at the harmonics is in the plane of the loop rather than broadside to it. This makes it ideal for contacts into Africa. It is perhaps the most effective 40-meter DX antenna that has been used at W1FB from northwest lower Michigan. The Transmatch is required on all harmonic frequencies other than 18.111 MHz, where W1FB has been conducting propagation studies with Bill Orr, W6SAI, Prose Walker W4BW, Bob Haviland, W4MB and Stu Cowan, W2LX, under special experimental/research licenses (KM2XQV). The loop has worked very well on 24.9 MHz as well during these tests. At 18.111 MHz, the SWR is 1.4:1.

The operating results at W1SE also indicate that a tilted loop, close to the ground, functions quite well. With loop resonance at 3825 kHz, the 2:1 SWR points occur at 3734 and 3934 kHz, respectively. This 200-kHz bandwidth spectrum can be shifted up or down the band by lengthening or shortening the loop conductor and Q section accordingly. From the W1SE location in Newington, the loop has delivered impressive performance for local and DX work.

A 40-meter Delta loop was constructed for use at W1SE after noting the fine performance of the 80-meter system. It was cut for resonance at 7016 kHz. This model was erected in a completely vertical format, using 143 feet 3 inches of wire. The Q section is 23 feet 2 inches long. The apex (feed point) is 4 feet above ground. The SWR on 40 meters is less than 2:1 across all of the band. The 80- and 40-meter W1SE loops showed resonance slightly apart from the design frequency, perhaps because of the proximity of the antennas to ground. Resonance on 40 meters was checked as 7050 kHz. Both loops are performing better for local and DX contacts than any of the many antenna types tested at W1SE. We would be even more impressed if we could elevate our Delta loops so the lower portions were a half wavelength or greater above ground.

In Conclusion

There is no rule that dictates the shape of a full-wave loop. The triangular format is convenient for mounting the radiator. If the apex is at the top, only one high support structure is needed. You may have one or more tall trees that can be used as supports. Circular, square or rectangular shapes have been used by many amateurs and the results were good. Certainly, a loop is an impressive receiving antenna, in terms of noise reduction. In some urban locations, that may be more important than transmitting a "death-ray" signal! There is something to be said about the age-old expression, "If you can't hear 'em, You can't work 'em."

An 80-meter Delta loop can be used on 160 meters by adopting one of two simple methods (Figure 4). A closed loop does not, however, offer good results when the overall length is a half wavelength. Either of the techniques in Figure 4 will work, but the method at A requires a ground radial system for best results.

Notes

[1] D. DeMaw, "Beat the Noise with a Scoop Loop," *QST*, July 1977, and "Maverick Trackdown," *QST*, July 1979.

[2] H. H. Beverage and D. DeMaw, "The Classic Beverage, Revisited," *QST*, Jan. 1982.

By David J. Crockett, WB4DFW **From *The ARRL Antenna Compendium Volume 5***

A Triband 75/40/30-Meter Delta Loop

WB4DFW wanted to use his Delta Loop on other bands besides 75 meters, so he installed some coaxial traps. He also reveals an effective technique to prevent breakage when support trees sway in the wind.

I've been using horizontally mounted, coax-fed Delta Loop antennas on 75 meters for many years. I've found them to be superior to dipoles—they are generally quieter on reception (perhaps because they are dc-grounded), exhibit broader bandwidth (typically less than 2:1 SWR over more than 200 kHz using a coaxial matching section), and are often more easily adaptable to available tree supports. They also don't fall down very often when installed correctly!

Outside of using an antenna tuner, a 265-foot-long 75-meter loop doesn't lend itself to working other bands–and I don't have a tuner. So while listening to a virtually dead 75-meter phone band one day and wishing I had a way to hit 30 meters for a long-overdue dose of CW, I had a brainstorm. Why not use traps? I'd homebrewed coaxial traps based on Sommer's article[1] several times in the past, and I knew they were easy to make and adjust.

The Traps

A quick examination of the latest incarnation of my 75-meter loop found it to be a reasonably equilateral triangle (a little over 85 feet per leg). I calculated that a 30-meter dipole should be something over 46 feet in total length (465/frequency in MHz). It appeared to be a simple enough matter to station some homebrewed coaxial traps 23 feet on each side of the feedpoint. A little more calculating found that it might even be practical to put in a pair of 40-meter traps as well.

Sommer gives a useful set of nomographs for computing trap dimensions on PVC pipe forms, using either RG-58 or RG-174 coax (the former capable of handling a kW with few problems). I opted for the larger cable. The 30-meter traps each require 46.8 inches of coax on a 1.62-inch OD section of PVC (6.75 turns). The 40-meter versions each use 61 inches of coax on a 2.25-inch OD PVC section (6.0 turns). Both PVC sections can be any convenient length. I used four-inch long pieces.

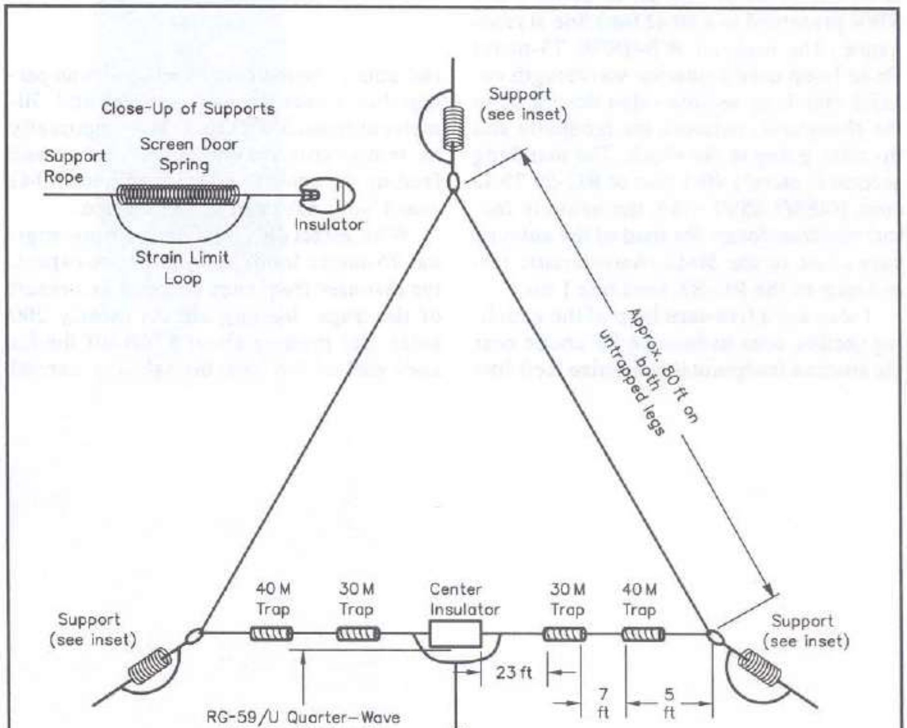
Close-Up of Supports
Support Rope
Screen Door Spring
Insulator
Strain Limit Loop
Support (see inset)
Approx. 80 ft on both untrapped legs
Support (see inset)
40 M Trap
30 M Trap
Center Insulator
30 M Trap
40 M Trap
Support (see inset)
23 ft
7 ft
5 ft
RG-59/U Quarter-Wave Matching Section for 75 Meters (40.1 ft, including choke coil)
Choke Coil (5 turns approx. 1 ft diameter)
RG-8X (any length) to shack

Figure 1—Overhead view of the Triband 75/40/30-meter Delta Loop, which WB4DFW mounted 40 feet off the ground. The traps were constructed with RG-58 coax. See text for winding details.

I wound the 30-meter traps in less than an hour and inserted them in the loop antenna 23 feet either side of the feedpoint. A quick SWR check found them resonant well below the bottom of the 30-meter band, but spacing the turns slightly apart and securing them with electrical tape raised the resonance easily enough, with the SWR bottoming out at 1.2:1 at 10.125 MHz.

The 40-meter traps went together with similar ease. Placement in the existing antenna was a little more challenging than with the first set of traps. The loading caused by the 30-meter traps made determining the physical location for the 40-meter traps uncertain. Grebenkemper notes that traps in a half-wave antenna tend to "pull" an antenna toward resonance,[2] even if they are not placed at exactly the optimum location. So, I simply took the formula length for 40-meter resonance, generously subtracted the physical length of the coax in each of the 30-meter traps and installed the 40-meter traps accordingly.

The approach was unscientific, but SWR was 1.3:1 at 7.1 MHz on the first try. Dumb luck, I guess, and I opted to make no further adjustments. The final antenna layout is shown in Figure 1, and a close-up photo of a trap is shown in Figure 2.

The Matching Section

The ARRL Handbook notes that the typical impedance of a full-wave loop antenna is on the order of 100 Ω, or about a 2:1 SWR presented to a 50-Ω feed line at resonance. The original WB4DFW 75-meter Delta Loop used a quarter-wavelength coaxial matching section (also described in the *Handbook*) between the feedpoint and the coax going to the shack. The matching section is merely 40.1 feet of RG-59 75-Ω coax [(465/3.85)/2 ×.66, the velocity factor] and transforms the load of the antenna very close to the 50-Ω characteristic impedance of the RG-8X feed line I use.

I also use a five-turn loop of the matching section coax to form an RF choke near the antenna feedpoint to minimize feed-line radiation. The matching section has no perceptible effect on the new 40 and 30-meter antenna SWRs (checked empirically by temporarily removing the section and feeding the antenna directly with the 50-Ω coax), so it has been left in service.

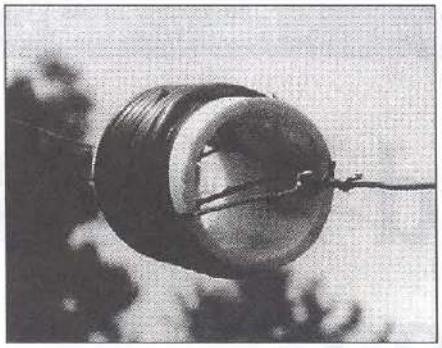

Figure 2—Close-up photo of one of the 40-meter traps. The ends of the coax cable used in each trap have been sealed against moisture with silicone caulking. See text for construction details.

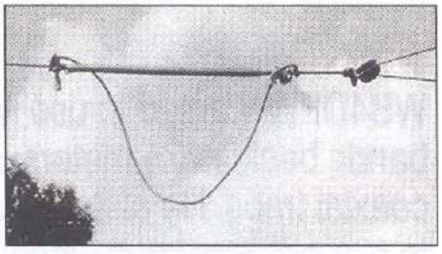

Figure 3—Photo of one of the support system springs, with rope used to limit total stretch. The rope is twice the length of the unstretched spring.

What effect did all this have on my original 75-meter loop? As one might expect, the resonant frequency dropped as a result of the traps' loading effects (nearly 200 kHz). But pruning about 8 feet off the far apex end of the loop brought my normal 1.2:1 resonant SWR back to my regular 3.842-MHz stomping grounds.

Keeping It Up in the Air

Here in South Carolina, pines are a great source of antenna support but their swaying during summertime thundershowers can wreak havoc on long, unprotected wire antennas. Over the years, I've found a simple shock-absorber system using a run-of-the-mill screen-door spring placed between a support rope and each antenna insulator can save a lot of grief. See Figure 3 for a closer look at the details.

A special aspect of my system is the employment of a "strain limit rope" in parallel with each spring. This piece of rope, approximately twice the length of the unstretched spring, keeps swaying pines from overly taxing the spring and gives a nice visual reference for gauging how much tension the antenna is operating under. It also provides a fail-safe for the inevitable corrosion-caused failure of the spring after several years aloft. It makes a fine bird perch, too.

So, How Does It Play?

Performance on all three bands has been most satisfactory. The 2:1 SWR bandwidth of the 30-meter section far exceeds the band edges. Similar 40-meter bandwidth is about 100 kHz (noticeably narrower than dipoles I've used, and I suspect partially attributable to my guesswork on the 40-meter trap placement). The 75-meter bandwidth is now approximately 150 kHz (slightly narrower than before the traps were installed). Signal reports both ways on all three bands are comparable to any other non-directional HF antennas I've used.

Notes and References

[1] Robert C. Sommer, "Optimizing Coaxial Cable Traps," *QST*, Dec 1984, pp 38-42.

[2] John Grebenkemper, "Multiband Trap and Parallel HF Dipoles-A Comparison," *QST*, May 1985, pp 29-30.

From *QST*, March 1989 (Hints & Kinks)

A Two-Band Loop for 30 and 40 Meters

After trying to find a way to place a 30-m delta loop inside an existing 40-m loop, I remembered an article in *All About Cubical Quad Antennas*[1] describing a $1\frac{1}{2}$-λ, or "Mini X-Q," loop. The gain of this antenna was said to be about 1 dB more than a 1-λ loop. I installed a large, ceramic SPST knife switch in the center of the delta-loop's bottom leg (see Fig 11). With this switch open, the full-wave, 40-m loop becomes a $1\frac{1}{2}$-λ, 30-m loop! The resonant frequency of this arrangement was 10.5 MHz. By adding 18-inch wires to the loop at both sides of the switch, I obtained resonance at 10.125 MHz.

Since the bottom of the loop is only 12 ft above ground, it's a simple matter to reach the band switch from ground level. (Caution: High RF voltage appears at the switch when the antenna is used for transmitting on 30 m.) Incidentally, the loop also works well on 15 m (SWR under 2:1 across the band) when set for 40 m, and I have used the 30-m configuration successfully on 80 m with the help of an antenna tuner.—*James Brenner, NT4B*

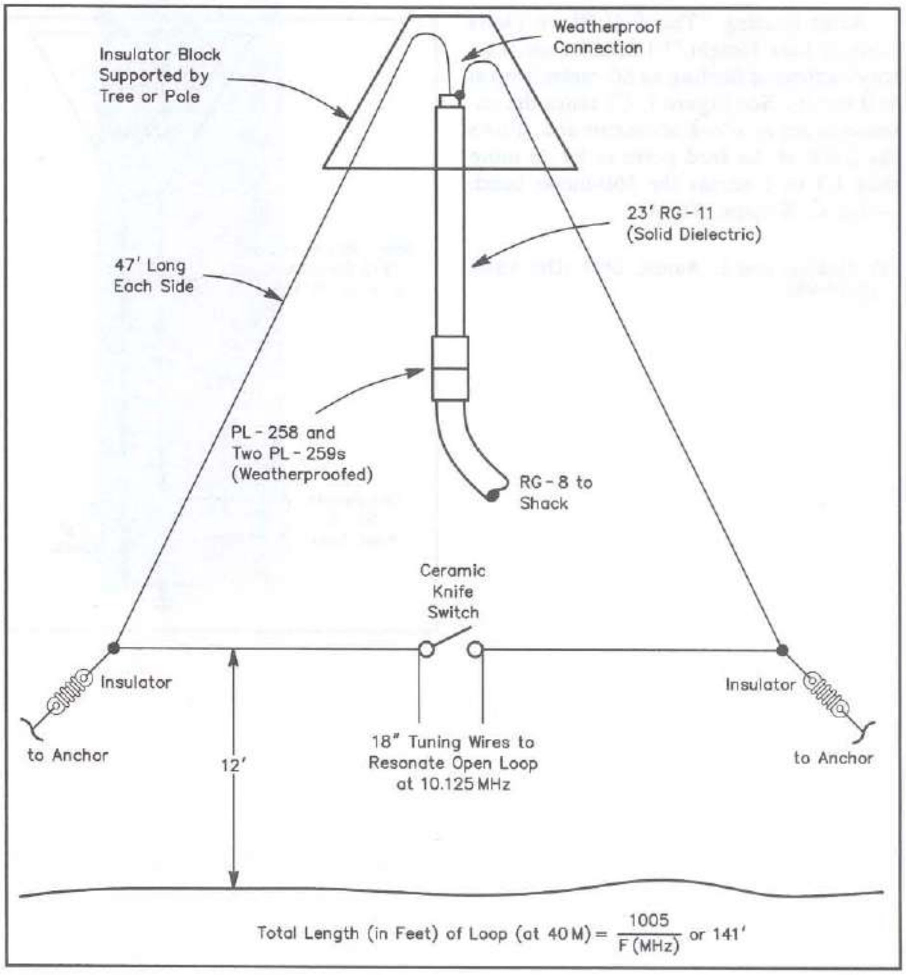
Weatherproof Connection
Insulator Block Supported by Tree or Pole
23' RG-11 (Solid Dielectric)
47' Long Each Side
PL-258 and Two PL-259s (Weatherproofed)
RG-8 to Shack
Ceramic Knife Switch
Insulator
Insulator
to Anchor
to Anchor
12'
18" Tuning Wires to Resonate Open Loop at 10.125 MHz
Total Length (in Feet) of Loop (at 40 M) = $\frac{1005}{F\,(MHz)}$ or 141'

Fig 11—Jim Brenner's 30- and 40-meter loop. Note the 18-inch tuning wires used to lower the antenna's 30-m resonance from 10.5 to 10.125 MHz. The antenna is top-fed via a $\frac{1}{4}$-λ 40-m matching section. See text.

---

[1] W Orr and S. Cowan, *All About Cubical Quad Antennas* (Wilton, CT: Radio Publications, 1970).

From *QST*, April 1990 (Hints & Kinks)

Feeding an 80-Meter Delta Loop at 160 Meters

After reading "The Full-Wave Delta Loop at Low Height,"[1] I found a satisfactory method of feeding an 80-meter loop at 160 meters. See Figure 1. C1 tunes the antenna to act as a $^{3}/_{4}$-λ resonator and allows the SWR at the feed point to be no more than 1.1 to 1 across the 160-meter band. *—Roy C. Koeppe, K6XK*

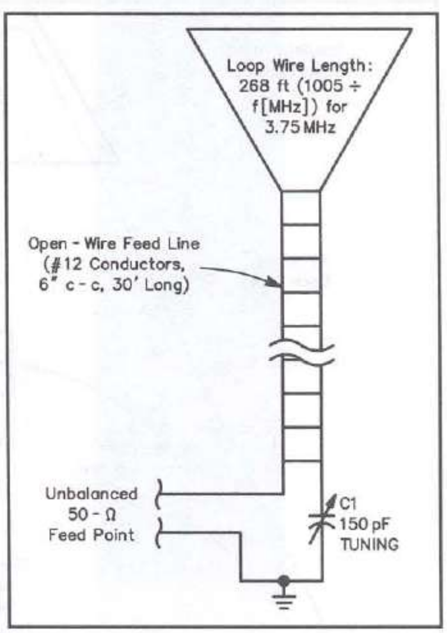
Loop Wire Length: 268 ft (1005 ÷ f[MHz]) for 3.75 MHz
Open - Wire Feed Line (#12 Conductors, 6" c - c, 30' Long)
Unbalanced 50 - Ω Feed Point
C1
150 pF
TUNING

Figure 1—Roy Koeppe enjoys satisfactory 160-meter operation with an 80-meter delta loop by feeding the loop as shown here. On 80 through 10 meters, Roy takes C1 out of the circuit and feeds the loop via its open-wire feed line and a balanced tuner.

---

[1] D. DeMaw and L. Aurick, *QST* (Oct 1984, pp 24-25).

CHAPTER FIVE

COLLINEAR ANTENNAS

5-1 **The Extended Double-Zepp Antenna**
From June 1938 *QST*

5-7 **Try an Extended Double-Zepp Antenna**
From February 1992 *QST* (Hints & Kinks)

5-8 **Notes on Wire-Antenna Construction**
From March 1992 *QST* (Hints & Kinks)

5-10 **Collinear Arrays**
From *The ARRL Antenna Book* 18th edition

By Hugo Romander, W2NB **From *QST*, June 1938**

The Extended Double-Zepp Antenna

Simple Antenna Structures Having Improved Gain and Horizontal Diversity

The questions of antenna directivity and antenna "gain" are becoming increasingly popular in discussions on antenna systems among amateurs, and it would therefore seem proper to preface an article dealing with directive antenna arrays with a warning, or perhaps a reminder, that one cannot have high gain and radiate in all directions at the same time. To most amateurs this is an obvious fact, but it may not be as well known that this principle is almost equally important in the *vertical* plane. Those of us fortunate enough to have available large open spaces in which to hang wires will find long-wire antennas, such as the "V" and the horizontal diamond, the easiest way to obtain high gain, but the radiation from such antennas is restricted in the vertical plane fully as much as in the horizontal plane. The result may well be that in the very direction such an antenna is supposed to work best, a simple horizontal doublet will put in a far better signal at certain distances.

It would seem, therefore, that the most practical antenna for a variety of distances is one with a fairly wide radiation pattern in the vertical plane.[1] The "stacking" of elements or the use of long wires is not recommended for distances under 1000 miles, except for the very short distances normally reached by the ground wave. The most universal high-gain antenna must restrict its radiation in the horizontal plane only and its height above ground must also be considered to obtain the best compromise in the vertical plane. It is the purpose of this article to discuss the merits of a simple antenna array which combines these desirable features.

THE DOUBLET

Let us discuss, first, the simple doublet, since this antenna will serve admirably as our basis of reference or comparison. To be more specific, consider this doublet as suspended horizontally and fed at one end in the time-honored fashion of the Zepp feeder, as in Figure 1A. Ignoring, for the moment, the fact that the open-ended feeder wire will be at somewhat higher potential than the other feeder wire and will therefore radiate, the horizontal radiation pattern about this doublet, if you're lucky, will be about as shown on Figure 2. Of course, it is assumed the antenna is sufficiently remote from power-line wires and house plumbing to be unaffected by such linear conductors, since our problem is complicated enough without having to consider the mutual impedance between our doublet and the neighbor's b.c.l. antenna!

This pattern of the horizontal doublet is, no doubt, familiar to most amateurs. Less familiar, perhaps, is the vertical radiation pattern (in the direction of maximum horizontal radiation) as shown in Figure 3. Here it is assumed that the height of the doublet above ground is one-half wavelength and that the earth has perfect conductivity. Fortunately, even the relatively low-conductivity soil and sand on Long Island reflects a high percentage of the horizontally polarized waves radiated at angles less than 50° to the earth's surface. Such reflected

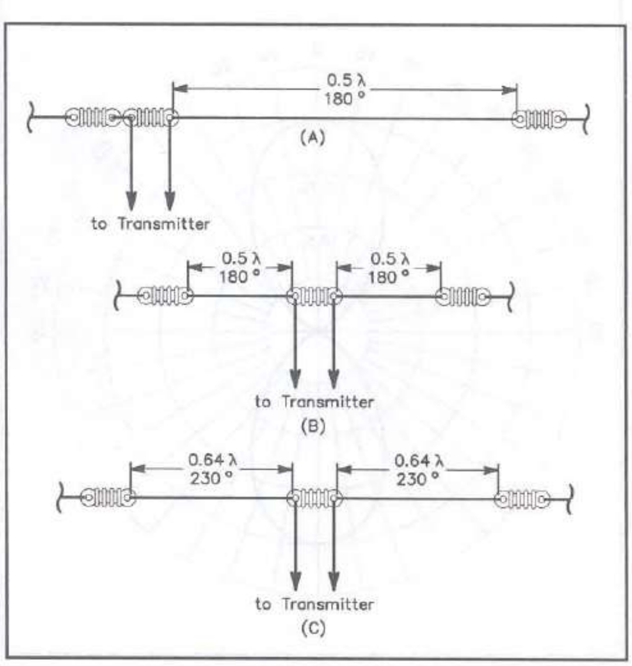

Figure 1—Basic Antenna Configurations (A) The half-wave Zepp; (B) Double Zepp or "two half-waves in phase"; (C) Extended double-Zepp.

---

[1] Although this viewpoint appears to be opposed to that expressed in the article "Simple Directional Arrays Using Half-Wave Elements," by N. C. Stavrou, in May, 1938 *QST* there is no actual conflict. As the present author points out, ihe broad vertical characteristic is to be preferred when the antenna is to give optimum results over short as well as long distances; the former article was concerned with long-distance transmissions, where the lowest possible angle gives best results under nearly all conditions. The type of work to be carried on naturally will be a determining factor. In any event, the simpler structures are not likely to be too sharp in either plane for satisfactory general work. —EDITOR

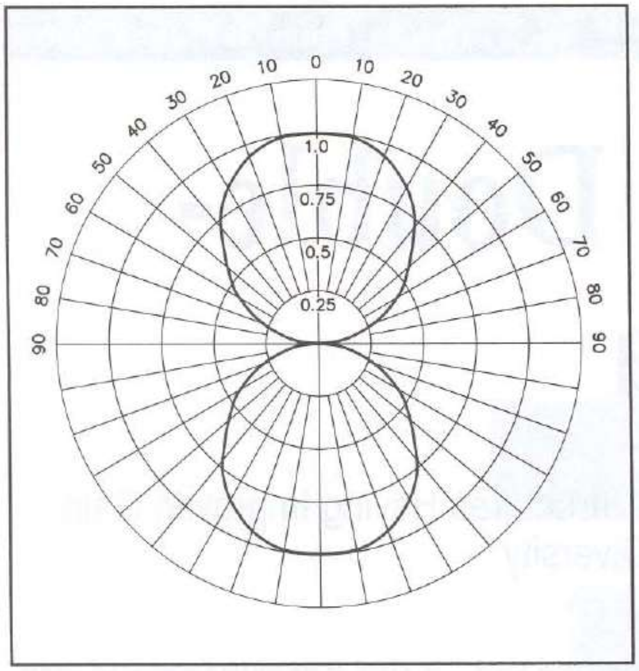

Figure 2—Horizontal pattern (relative field strength) of a doublet half-wave antenna.

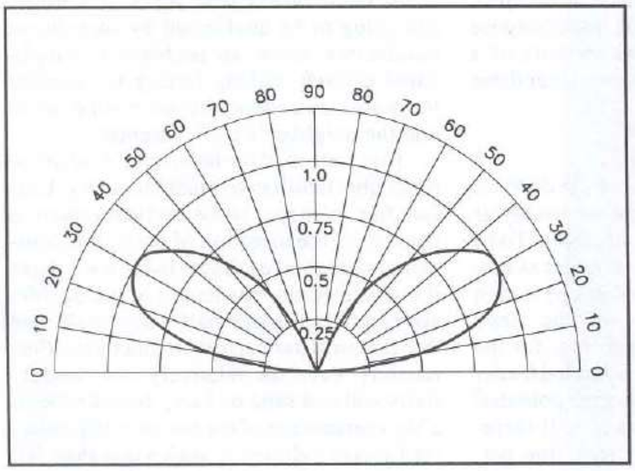

Figure 3—Vertical-plane pattern for a horizontal antenna, ½-wave above perfect earth in direction of maximum radiation.

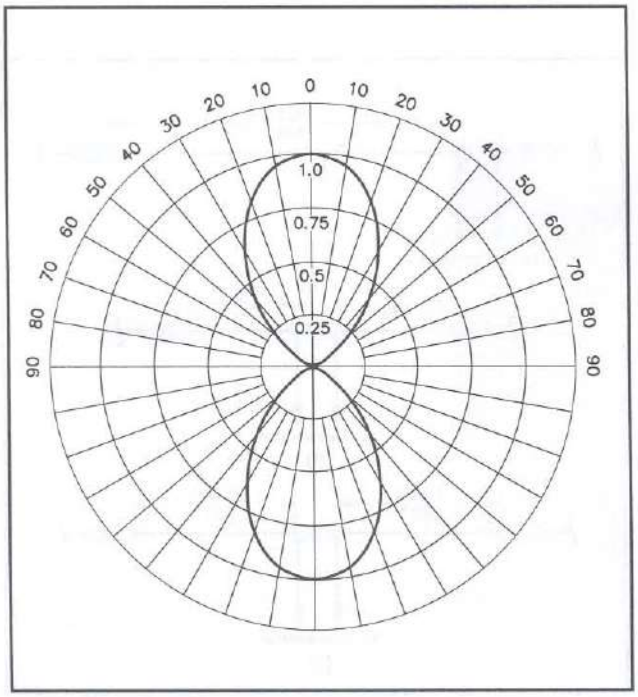

Figure 4—Horizontal pattern of a double-Zepp or two half-waves in phase.

waves, combining with the direct radiation from the antenna within the range of vertical angles in which we are most interested, are responsible for a maximum gain of nearly 6 db over the same doublet in "free space"; that is, without the presence of the earth.

This gift of 6 db must not be taken too much for granted, however. Consider the fact that most amateur communications at high frequencies utilize vertical angles ranging from 10 degrees to 50 degrees. That 6 db must, therefore, come from radiations reflected from the earth's surface quite some distance from the doublet, and in the desired direction of transmission. If the ground slopes sharply upward, or houses or wires are in this area, it becomes questionable, indeed, whether any great portion of the available 6 db is realized. This is especially true for the lower angles of radiation, but usually nothing can be done about this situation, so let us see if we can make up for our ground reflection losses by increasing the horizontal directivity.

THE DOUBLE-ZEPP

By the simple expedient of attaching another doublet to the open-ended terminal of the Zepp feeder, as in Figure 1B, and hanging this doublet parallel and coaxial with the original doublet, an appreciable gain may be obtained. The horizontal radiation pattern will be as shown in Figure 4. The gain should measure about 1.9 db, corresponding to a 55 per cent increase in power. This antenna is widely used among amateurs and is popularly known as the "double-Zepp" antenna, or "two half-waves in phase." A fair amount of gain and a reduction in interfering signals from end-wise directions is easily realized, but the gain is rather disappointing, since it would seem one is entitled to 100 per cent gain in power when doubling the number of radiating members.

The reason why only 1.9 db is realized with the double-Zepp is made clear only after a mathematical study of the situation. Briefly, the close proximity of the two doublets causes a mutual coupling between them, and this coupling (mutual impedance) has an adverse effect on the radiation resistance insofar as gain in the broadside direction is concerned. Obviously, the thing to do is to move the doublets further apart, but this complicates the method of feed. A much simpler way of obtaining increased gain was evolved by Mr. A. A. Alford, of Mackay Radio and Telegraph Company, from the principles discussed in Mr. G. H. Brown's article on broadcast antennas in the Proceedings of the I.R.E. for January, 1936. Mr. Alford presented this idea in a paper delivered at an I.R.E. meeting in Washington.

EXTENDED DOUBLE ZEPP

The gain of the double-Zepp may be increased from 1.9 to 3.0 db by the simple expedient of increasing the length of each doublet until its electrical length is 0.64

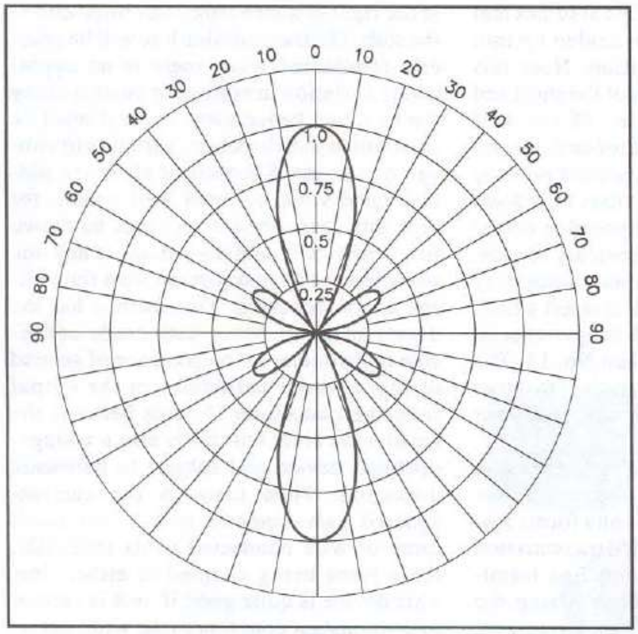

Figure 5—Horizontal pattern of double 230-degree Zepp

wavelength instead of 0.5 wavelength. In electrical degrees the double-Zepp consists of two 180-degree elements; the *extended* double-Zepp for maximum gain should consist of two 230-degree elements. See Figure 1C. In this way the power gain may be increased from 55 per cent to 100 per cent. The gain decreases rapidly for extensions beyond 230 degrees, and therefore, when operating over a band of frequencies, each of the two elements should not exceed 240 degrees for the highest frequency. The horizontal pattern for the double 230-degree Zepp is shown in Figure 5. The vertical pattern in a plane perpendicular to the antenna will be the same as for the simple doublet.

ANTENNA IMPEDANCE

The impedance of this antenna at the termination of the transmission line is of interest, since it has an important bearing on the standing-wave ratio of current or voltage in the line, and it will be compared with that of the ordinary double-Zepp. The double 180-degree antenna presents an impedance of approximately 4400 ohms of almost pure resistance as a termination for the transmission line. This value will be slightly affected by the size of wire used in the antenna and, to a moderate extent, by the height above ground or the influence of nearby conductors, and so the "free space" value is given for No. 14 wire. With this antenna and a 600-ohm surge-impedance line the ratio of maximum to minimum current along the line will be 4400 divided by 600, or 7.3. Incidentally, the terminating resistance of the simple Zepp-fed doublet is about 12,000 ohms, resulting in a standing wave ratio of 20 on our 600-ohm line. For calculation of the surge impedance of the line the reader is referred to *The Radio Amateur's Handbook.*

The impedance at the center of the double 230-degree antenna is not a pure resistance, and hence its effect upon the transmission line be such that maxima or minima of voltage and current along the line will not be odd multiples of a quarter wave from the antenna, as with the ordinary Zepp or double-Zepp antennas. As might be expected, the current or voltage maxima will be shifted towards the antenna, since the two doublets are longer than normal, and this shift is approximately 0.13 times the wavelength. At any rate, the antenna impedance is such that the equivalent pure resistance at any voltage maximum will be about 6000 ohms; that is, the standing-wave ratio will be 10 on a 600-ohm line.

Knowing the standing-wave ratio, it becomes an easy matter to calculate the input resistance to the transmission line if it was cut, let us say, at any current maximum. Thus, for the simple Zepp-fed doublet, this resistance would be 600 divided by 20, or 30 ohms. For the ordinary double-Zepp antenna this resistance would be 600 divided by 7.3, or 92 ohms. For the double 230 degree antenna this resistance would be 600 divided by 10, or 60 ohms. Those who do not use correcting stubs or other methods to obtain a "flat" line (standing wave ratio of 1), but who cut their transmission lines at the maximum current point after the manner of the "tuned feeder," will find these figures useful in calculating the power output of their transmiitters.

FEED LINES

However, the use of flat lines, in amateur circles usually referred to as untuned lines, is becoming increasingly popular as their merits are more generally known, and the figures given are essential in calculating the length and position, for example, of a correcting stub, the latter being one of the simplest devices for reducing the standing-wave ratio to unity. A discussion of the use of the correcting stub may be found on page 307 of the *Handbook* (15th edition) where a table indicates the lengths and positions of the stub for various standing-wave ratios. The shorted stub or "loop" will generally be most practical, and some idea of dimensions may be obtained from the fact that when using the double 230-degree Zepp on 14,200 kHz the stub should be 3.5 feet long; that is, it should be composed of the same kind of wire as is used in the transmission line, this wire being 7 feet long and bent to form a U-shaped rectangle with a width equal to the transmission line. At the frequency mentioned, where each extended doublet is 43 feet long, the correct location for the stub will be about 8 feet from the antenna.

Ordinarily the best place to introduce power into an antenna is at a point symmetrically located with respect to the opposite ends of the system. This will not always be practical, because of space limitations, and hence it may become necessary to feed the extended double-Zepp antenna at one end in true Zepp fashion. Some sacrifice in gain will result, due, for the most part, to unequal distribution of current between the two doublets and to radiation from the feeder because it has an unbalanced load. Physical dimensions of feeder, extended doublets, and phasing stub are shown in Figure 6. Note that the phasing-stub length is shorter than a quarter wavelength, but by an amount not exactly

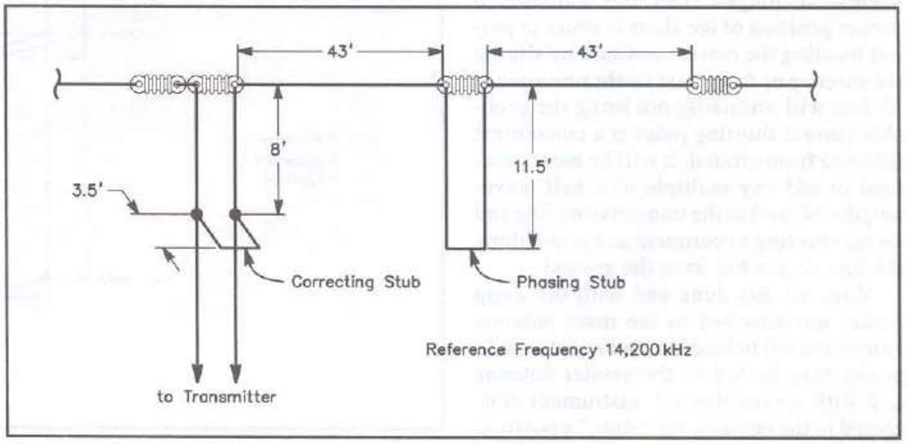

Figure 6—End-fed double 230-degree Zepp with 600-ohm non-resonant transmission line.

equal to the extension of each doublet beyond a half wavelength. This discrepancy is due to interaction between the doublet wires and the wires of the phasing stub. Again a reference frequency is given to simplify calculation of lengths at other frequencies by inverse proportion. The figures given should be regarded as approximations, and tests made individually to determine dimensions and positions of important elements.

ANTENNA ADJUSTMENTS

One of the most effective methods of adjusting an antenna system is to use the transmitter to excite, at the desired frequency, another antenna stretched out, perhaps temporarily, at least 8 or 10 feet from the ground. This "exciter" antenna should preferably be at least one-half wavelength away from the antenna to be adjusted. If the two are parallel to each other, so much the better. Using the end-fed extended double-Zepp antenna as an example, the first consideration is the length of each doublet, and this brings up the problem of "end-effect." We have approximately solved this by cutting our doublets 5 per cent short of the theoretical length. Actually the end effect, which normally exists only at the ends of the antenna, depends on the wire size and the size of the insulator fitting (if metal caps are used), as well as the frequency. Moreover, the end-effect does not vary quite as rapidly as a direct function of the wavelength, so that if an end-effect of 1 foot is correct at 14,200 kHz, it is more nearly 1.5 feet than 2 feet at 7100 kHz.

The theoretical "uncorrected" length of each half of an extended double-Zepp antenna at 14,200 kHz is 44.3 feet. The end-effect may be no more than 1 foot, but since it is better that the antenna be a little too short than too long, let's make the correction 1.3 feet, so that each half will be exactly 43 feet long. At the center insulator which separates the two halves a "stub" must be connected, if the system is to be end-fed. Now, this stub will, take the form of a transmission line shorted at the end opposite from the antenna, and for the frequency under consideration its shortest length will be about 12 feet. Since it is our intention to make this transmission line somewhat longer than the anticipated proper position of the short in order to permit locating the correct position by sliding the short up or down, and furthermore since 12 feet will ordinarily not bring the probable correct shorting point at a convenient distance from ground, it will be more practical to add any multiple of a half wavelength (34 feet) to the transmission line and do the shorting experiment at a point along the line accessible from the ground.

With all this done and with the Zepp feeder *not* attached to the main antenna (why not use it to feed the exciter antenna?), power may be fed to the exciter antenna and with a sensitive r.f. instrument connected in the short on the "stub," a position of the short may easily be found where maximum current flows. This assumes that the main antenna has been hauled up into its normal operating position. Note this maximum-current position of the short and measure in exact multiples of one-half wavelength from this point towards the antenna, thus arriving at the nearest point to the antenna at which a short may be placed. This will give the shortest possible center stub. Of course, it is not necessary that the center stub be made that short, since very little loss will be incurred if it is left a half-wave or even a wavelength longer–that is, if the wire is not smaller than No. 14. The Zepp feeders may now be attached to either end of the antenna, as it is now ready for operation.

The method of adjusting the center stub just outlined is also useful where it is desired to feed any of the antenna forms discussed above with an untuned transmission line—that is, a transmission line terminated by its surge impedance. Using the example cited in the previous paragraph, let us assume the point of maximum current through the short has been found and, after soldering a piece of wire across the line at this point, the extra length of transmission line is chopped off. The two-wire line from the transmitter may now be tapped in above this shorted end of the stub, and it is only a question of position of the feeder tap from the shorted end to terminate the line properly. See Figure 7. This point will be approximately 3.3 feet from the short for a 600-ohm transmission line. Of course, when all this is done, we will have the same transmission line with correcting stub discussed earlier in this article.

LINE-CURRENT MEASURING DEVICES

The same principle holds true for the double Zepp and for center feeding the simple doublet, the only ambiguous point being that of knowing when the transmission line from the transmitter is connected at the right distance from the closed end of the stub. The transmission line will be properly terminated when there is no appreciable variation in voltage or current along the line, and hence some method must be used which will detect any variation in voltage or current. Simplest of all is the old-fashioned wood-covered lead pencil, for with this device bright arcs may be drawn from either wire and the voltage at any one point judged by comparison with the voltage at another point. This method has the disadvantage of being both crude and liable to error due to the presence of second or higher-order harmonics in the output from the transmitter. A little better is the small neon bulb, but this is also a voltage-operated device and subject to harmonic distortion. Then there is the current-squared galvanometer with a few small turns of wire connected to its terminals, these turns being coupled to either line. This device is quite good if one is careful to use constant coupling to the wire, but its disadvantages are that coupling between the meter and line is both inductive and capacitive, causing the meter to read differently according to how it is held up to the line, and that the meter will also be influenced by harmonics.

The most satisfactory device seems to be a sensitive galvanometer in a miniature tuned circuit coupled to the line. Such a device combines sensitivity with freedom from harmonic distortion, but the effects of capacity are still there to some extent. The tuned circuit should therefore be carried along the line with its position relative to the plane of the two wires maintained constant. See Figure 8. Checking along one of the two wires only is usually sufficient, but if one suspects an unbalance exists, the other line should also be checked. If the antenna system is reasonably symmetrical, unbalanced currents will most likely be due to improper coupling to the transmitter, but that is another story. At any rate, if a cur-

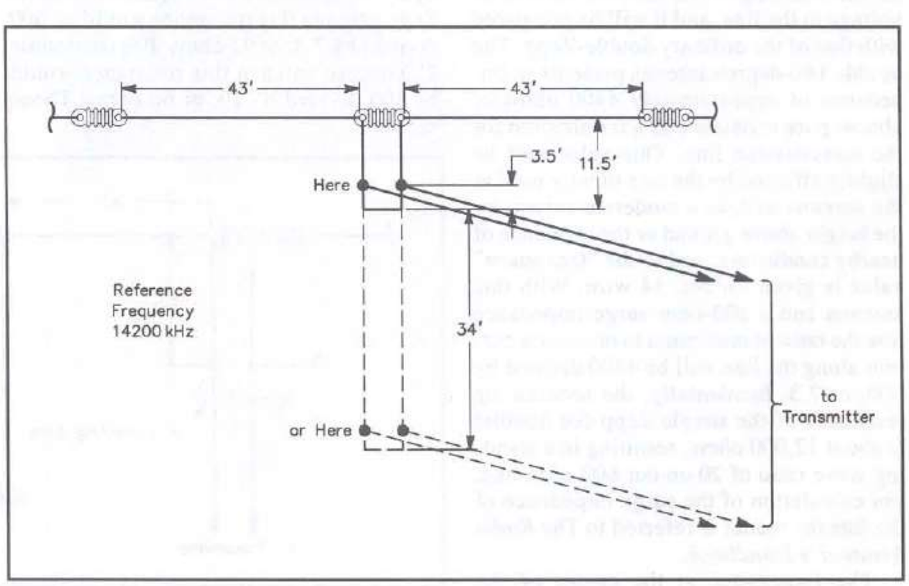

Figure 7—Center-fed double 230-degree Zepp with 600-ohm transmission line.

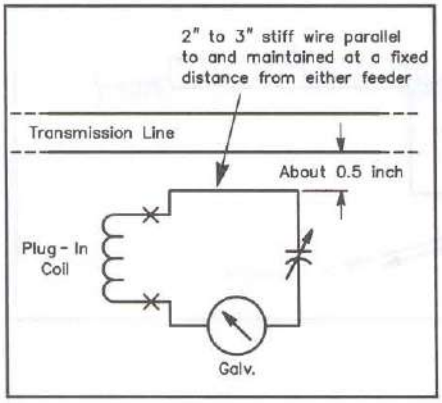
2" to 3" stiff wire parallel to and maintained at a fixed distance from either feeder
Transmission Line
About 0.5 inch
Plug - In Coil
Galv.

Figure 8—Portable line-current indicator.

rent maximum (voltage minimum) appears at the junction of transmission line and stub, the connection to the stub is too near the shorted end, and if the current is a minimum the opposite, of course, will be true. If the current maximum or minimum (assuming standing waves actually exist) does not occur near the stub, there has been a slip-up somewhere in the procedure, for this will be evidence that the stub itself did not tune the antenna to resonance.

PARASITIC ELEMENTS

The use of spreaders several feet in length to support each end of a "flat-top" antenna composed of two wires is again becoming popular, and for a good reason. Gains of around 5.5. db for a bi-directional array and around 7 db for a unidirectional array are thus made possible. One of the most promising arrangements is that of two double-230-degree antennas supported parallel to each other only 0.2 wavelength apart, as illustrated in Figure 9. Antenna A is excited by the transmitter using the same technique as when exciting a single double-230-degree antenna. Antenna B may be tuned by proper positioning of a short on its stub to become either a director or a reflector. In fact, this stub might be extended by multiples of a half wavelength to enter the operating room where, by means of a single switch the shorting position may be changed from that corresponding to a director to the position corresponding to a reflector, thus reversing the directivity of the antenna system. Information as to the gain of such an antenna system when one antenna is used as a director is not available, but when it is used as a reflector and adjusted to give a *minimum* signal to the rear, the signal forward will be 7 db better than a simple doublet, and the signal backwards will be 7 db less than from a doublet, resulting in a front-to-back ratio of 14 db. Adjustment of the reflector shorting bar should be on the basis of minimum backward signal from the antenna, or minimum signal received from a station in that direction when using the antenna system for reception, since this adjustment can be made more accurately than one resulting in maximum front signal. This will minimize QRM for both transmitting and receiving. The same principles hold true when adjusting the auxiliary antenna for use as a director.

If adjustments of the auxiliary antenna are to be made when transmitting and an ammeter is temporarily inserted in the shorting bar, minimum backwards signal will occur for the reflector when the short is moved slightly further from the antenna than that position corresponding to maximum current through the short. For the director, optimum conditions will be obtained when the short is slightly closer to the antenna than the maximum shorting current position. From this it can be seen that the proper adjustment of the auxiliary antenna is quite critical and certainly very important. Since field strength instruments are not generally available, a receiver fitted with a signal strength indicator of some sort located within a few miles of the antenna and in the desired direction should prove to be the next best thing. Connecting your own receiver to the antenna, with the other fellow transmitting, is perhaps even more practical in search for that minimum-signal adjustment.

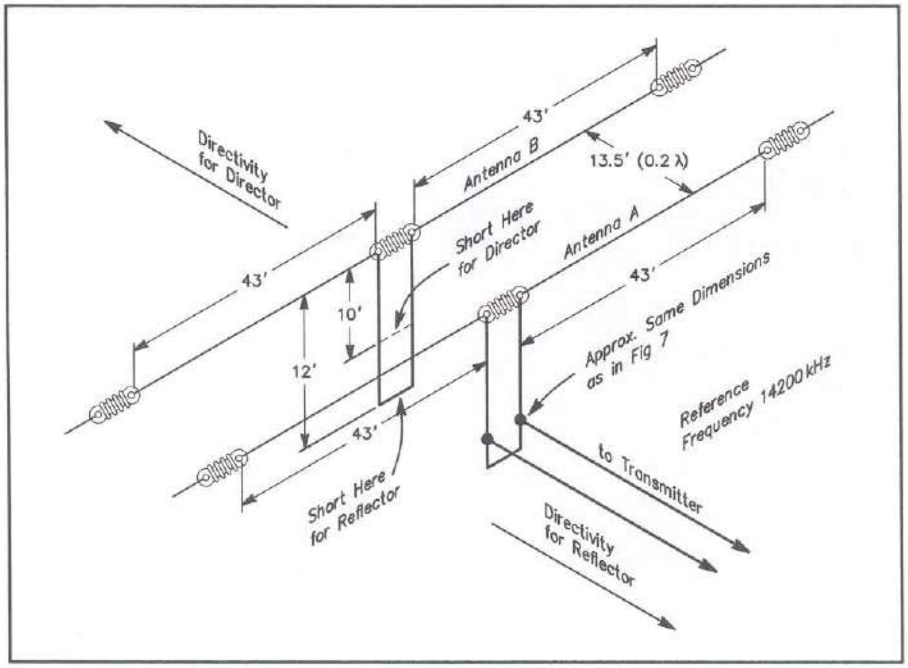

Figure 9—Double 230-degree Zepp with parasitic reflector-director.

The presence of the auxiliary antenna only 0.2 wavelength away from the driven antenna and adjusted properly for maximum forward or backward radiation will obviously affect the radiation resistance and impedance of the driven antenna. As a result, if the auxiliary antenna is left open-circuited and the stub on the driven antenna is adjusted in accordance with the method previously given, this adjustment will not be correct when the auxiliary antenna is, in turn, properly tuned. Moreover, the impedance of the driven antenna will differ slightly according to whether the auxiliary antenna is tuned as a director or as a reflector. If quick change from reflector to director is contemplated, some sort of compromise is indicated in the adjustment of the stub on the driven antenna. From a practical standpoint, however, the adjustment of the auxiliary antenna should be made first, be it reflector or director, with the driven antenna excited by the two-wire transmission line without a correcting stub. This may require temporary adjustment of the transmission-line length, either physically or by means of series coils or condensers, to bring a low-impedance point at the coupling coil to the transmitter in order to load the latter satisfactorily. In fact, if the line length from antenna to transmitter is not more than a wavelength or so, the reduction in line losses resulting from the use of a correcting stub is hardly worth the trouble of installing it. The point is, if a "flat" line is desired, adjustments of the stub at the driven antenna should not be made until after the auxiliary antenna has been tuned to give the desired radiation pattern.

LARGER COLINEAR ARRAYS

Now let us suppose our backyard is big enough to hang up more than two doublets end to end. An interesting possibility would be a four-element array of 230-degree elements. But here the principles of the extended double-Zepp must be carefully considered in designing the length of each element and the length of the phasing stub separating them. The actual arrangement will be as shown in Figure 10. Both degrees and dimensional designations are used to indicate the electrical and physical length of each element. The length of the stubs indicated is approximate only and must be adjusted for best results. The lengths of the various antenna elements as shown, however, may be assumed correct. The transmission line can be tapped on to any of the stubs, but connection to the center one will give the greatest gain.

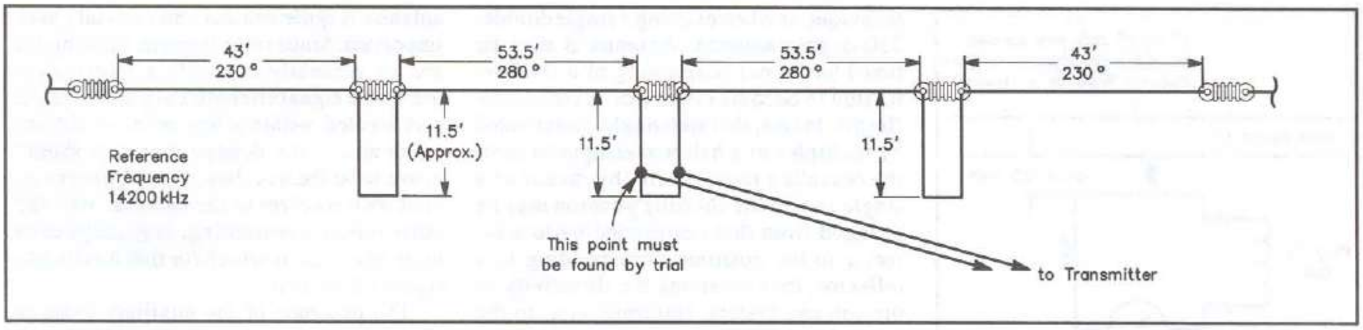

Figure 10—Four-element array with 230-degree elements.

The principle of the design shown for four elements is to provide the same separation in space between each doublet center as is provided between the two doublet centers in the double-230-degree antenna. Each phasing stub will, of course, be considerably shorter than a quarter wavelength, and when adjusting their length to tune the whole antenna to resonance, all three stubs must be made the same length; that is, if one is shortened two inches, the other two must also be shortened by two inches, assuming they were all the same length to start with.

This may prove to be an awkward and tedious method of adjustment, but the procedure may be simplified by adjusting the length of the center stub only as a first approximation. The use of an "exciter" antenna temporarily rigged nearby is assumed. If it is found the center stub must be shorter than the other two stubs by 3 feet, for instance, in order to obtain maximum current through its shorting bar, the other two stubs should be shortened by 1 foot and a new position found for maximum current through the center stub. This new position should correspond closely to equal length for all three stubs. This procedure of tuning the four-element array should be followed even though a stub at the center will not be used, finally, its place being taken by the transmission line itself. Such a connection will result in standing waves along the line and it will have to be "tuned" to permit easy coupling to the transmitter or receiver.

The horizontal pattern of the four-element array just described will have two major lobes at right angles to the antenna and several minor lobes of small amplitude. The gain in actual practice has proved to be greater than anticipated, and probably is more than 7 db. The major lobes will be much narrower than for the double-230-degree antenna, thus requiring more careful "aim" in erecting the array, or provisions for swinging it about. It is assumed this array, as well as the others described, is hung horizontally. It may be hung vertically, of course, but the possible restrictions due to sharp directivity in the vertical plane should then be considered. In general, the horizontally polarized antenna will prove most practical, chiefly because the ground reflection usually encountered with vertically polarized radiations is inferior, and because high masts are not often available.

The adjustment of the antennas described in this article may seem a bit involved as compared with the simple doublet or the double-Zepp antenna, but the extra gain thus made available should make their use worthwhile. For that matter, the double-230-degree antenna need not be tuned at all, its construction being made in accordance with the dimensions given and the transmission line tuned to fit the transmitter. In this respect the simplest of all directive antennas, size and gain also taken into consideration, is the extended double-Zepp.

From *QST*, February 1992 (Hints and Kinks)

Try an Extended Double-Zepp Antenna

Although the extended double Zepp (EDZ) antenna (Figure 1) has been in just every antenna handbook since the year one, hams seldom use it. Its overall length is 1.28 wavelengths (1.28 λ), and it's bidirectional broadside. Fed with open-wire line and a balanced antenna tuner, an EDZ also makes a fine multiband antenna. Let's look at an extended double Zepp for 17 meters. We can calculate the overall length of its two wire elements with the formula

$$\frac{984}{f(\text{MHz})}\times 1.28 = \text{length in feet} \qquad \text{(Eq 1)}$$

Using this formula, an 18.15-MHz EDZ works out to be 69.4 feet long. At this frequency, the EDZ exhibits, 3 dBd gain [*AB* page 8-34][1] in a figure-8 pattern with two major and four minor lobes. It still performs usefully, when operated on several bands lower in frequency than 17, however. At 20 meters, a 17-meter EDZ acts as two slightly long half waves in phase, exhibiting between 1.6 and 2 dBd gain [*AB* page 8-32]. At 40 meters, it's a slightly long ½-λ dipole [*AB* page 2-16,2-17,3-11 and 3-12]. All of these modes are directional broadside if the EDZ is positioned at least ½-λ high at 40 meters [*AB* pages 3-8 and 3-9]. At 15 meters, it exhibits a four-leaved-clover pattern, with minor lobes broadside; at 10 and 12 meters, it's close to two full waves in phase and produces a pattern similar to that at 15 meters. It can even be used as a short 75-meter dipole–not bad for a 70-foot piece of wire.

Scaled for 28.7 MHz–43 feet, 10 inches long—the extended double Zepp gives a 3-dBd-gain, figure-8 pattern at 10 meters and a similar pattern with a bit less gain at 12 meters. It acts as two halfwaves in phase at 15 meters, with about 1.6 dBd gain and a figure-8-pattern. At 17 and 20 meters, it's somewhat long for a half-wave dipole; a tuner can make it work at 40. On all of these bands, assuming that it's at least ½ λ high at its lowest band of operation, the 10-meter EDZ is directional broadside.

There's nothing magic about the extended double Zepp. It's a tried-and-true dipole that offers useful gain at its design frequency and good multiband performance. *–Bob Baird, W7CSD*

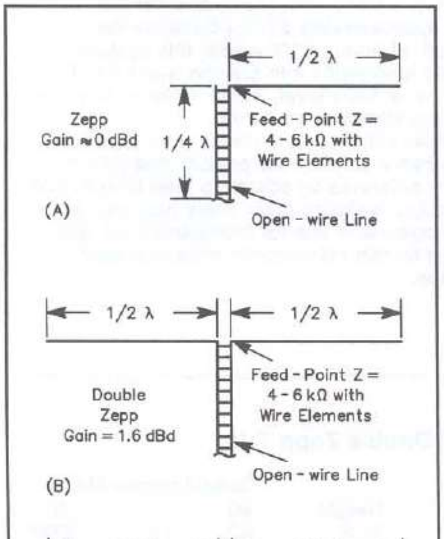

Figure 1—Evolution of the extended double Zepp antenna. At A, the classic Zepp—a ½-λ wire end-fed via an open-wire feeder operated with one of its two wires unterminated. (Antenna lore tells us that this feed method received its name from its use in zeppelin radio installations.) Commonly, Zepp systems use a feeder length of ¼ λ (or an odd multiple thereof) to transform the antenna's high feed-point impedance to a low impedance at the feeder terminals. Assuming efficient feed, this system exhibits 0 dBd gain—that is, *no gain* relative to a half-wave dipole cut for the same frequency and erected in the same position.

B shows a double Zepp, also known as *two half waves in phase* or (incorrectly) a center-fed Zepp. This is a form of collinear antenna because each ½-λ wire acts as an element, and because both elements lie along one line. It's also a *dipole*—a *full-wave* dipole that exhibits about 1.6 dB gain relative to a half-wave dipole in the same position. As with the Zepp, feeders cut to an odd multiple of ¼ λ make for least hassle feed with the double Zepp, but a good balanced antenna tuner should be able to handle whatever impedance appears at the feed-line input terminals.

A double Zepp becomes an *extended* double Zepp when you make its wires 0.64 λ long instead of 0.5 λ. Its gain increases, too—to about 3 dB relative to a half-wave dipole in the same position and cut for the same frequency.

The rough feed-point impedances shown here are for wire elements (fatter elements, such as those made of aluminum tubing, will lower these values somewhat) and convey only the resistive portion of the antennas' impedance. For more on the extended double Zepp antenna, see *The ARRL Antenna Book* and J. Reh, "An Extended Double Zepp Antenna for 17 Meters," *QST*, December 1987, pp 25-27.

---

[1] This and the other AB references in this item refer to pages in G. Hall, ed, *The ARRL Antenna Book*, 16th ed (Newington: ARRL, 1991).

From *QST*, March 1992 (Hints and Kinks)

Notes on Wire-Antenna Construction

Robert J. Zavrel, W7SX, seconds last month's extended-double-Zepp hint as he reports on how he keeps three EDZs high in trees for world-encompassing coverage (Figure 1) at 80, 40 and 20 meters—*Ed.*

What is a terminal DX hound to do when he suddenly moves to a rural area and has access to the stuff dreams are made of: several 100-foot-plus ponderosa pine trees atop a gentle rise in a reasonably clear area, only 500 feet from his house? The answer: *Wire antenna arrays!* A few months spent confronting the practical problems of installing three extended-double-Zepp antennas in these trees has yielded some solutions that I'd like to share.

Counterweights. Although my pines are wonderful supports for wire antennas, they have a few minor disadvantages. Among these is the fact that they sway in the wind. Because too much tension can snap antenna wires and supporting cords in strong winds, a wire-antenna engineer must compensate for variable tension in a horizontal wire antenna strung between two of these majestic plants. The textbook solution to this problem involves installing pulleys where the wires meet the trees. Ropes attached to the wire-end insulators pass over the pulleys and connect to counterweights hanging near the ground. Installed with antenna-to-counterweight lines of sufficient length, such a system also allows the antenna to be raised and lowered without further climbing.

My system includes three extended double Zepps, each fed with 450-Ω, low-loss TV twinlead. Including its supporting lines, each EDZ spans 165 feet. Counterweighting the ends of these antennas serves two purposes: It keeps tree movement from snapping them, and it keeps their wires reasonably horizontal. But how much weight is required, and what should the weights consist of? Charts exist for computing tension necessary to keep a singlewire span taut, but the presence of a twinlead feeder at the span's center renders such charts useless.

Enter experimentation! I strung the extended double Zepp, minus feed line and plus pulleys, between the two trees at about 15 feet above ground. Next, I weighed the feed line. Then I located an object of simi-

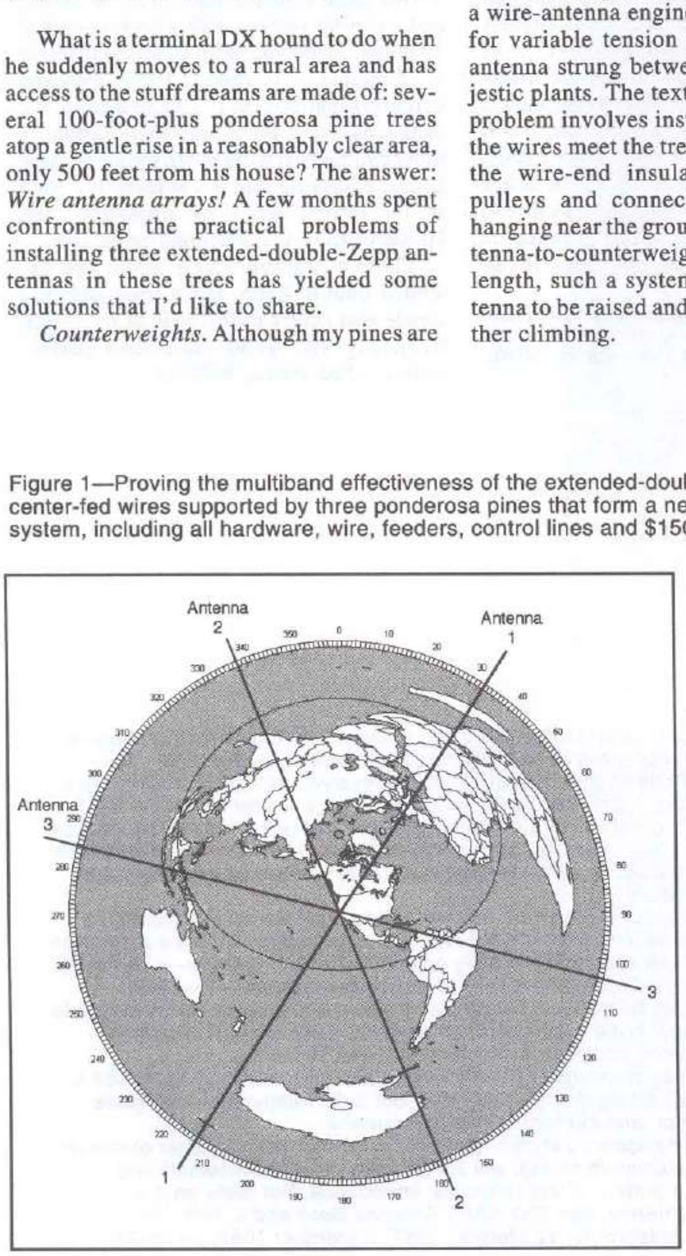

Figure 1—Proving the multiband effectiveness of the extended-double-Zepp antenna, "The W7SX Wire Farm" consists of three center-fed wires supported by three ponderosa pines that form a nearly equilateral triangle atop a small rise. (Total cost for the system, including all hardware, wire, feeders, control lines and $150 fee for the tree climber, approximately $700.) Because the ground slopes to the south at about a 10° angle, this system works best for VK, ZL and long paths into Europe and Asia. To the north, the horizon is essentially level. Table 1 shows how the antennas function on the system's three bands.

You're not a second-class citizen just because your best antenna is a dipole. You can easily set the pattern and gain of center-fed horizontal wire antennas by adjusting their length. Get the wires as high as possible, keeping them away from nearby objects, use a tuner and open-wire line for multiband feed, and have a blast! Surrounding terrain may greatly influence your antenna's DX performance.

Table 1
The W7SX Extended Double Zepp Trio

			Band/Antenna Mode		
	Length	*Height*	*80*	*40*	*20*
Antenna 1	82 ft	98 ft	SD	LD	EDZ
Antenna 2	82 ft	98 ft	SD	LD	EDZ
Antenna 3	164 ft	82 ft	LD	EDZ	LW

SD = short dipole; LD = long dipole; EDZ = extended double Zepp, LW = long wire (dipole legs 1 λ or more in length). Because a dipole longer than $^1/_2$ λ exhibits gain over a half-wave dipole in the same position, these three antennas exhibit gain over a half-wave dipole in all but the SD mode.

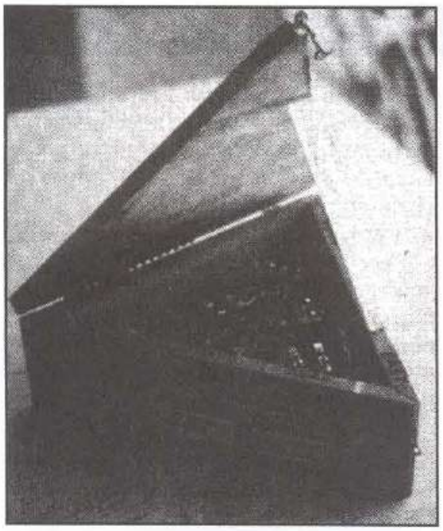

Figure 2—Bob Zavrel uses a triangular relay box to switch feeders between antennas mounted in a triangular pattern; a square box in this situation twists feeders, en-couraging antenna sway to break feeder wires. *(W7SX photo)*

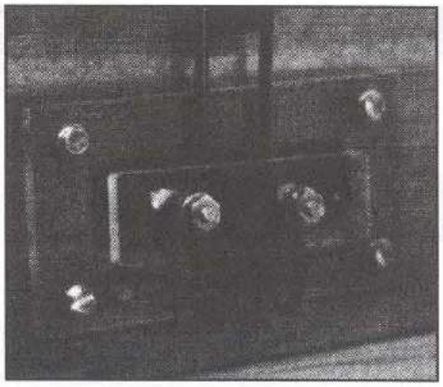

Figure 3—A combination feedthrough insulator/strain relief made of fasteners and two plexiglass plates.

lar weight, put it in a plastic bag, tied the bag at the antenna feed point, and experimented with various weights. Two 15-lb weights kept the antenna wire horizontal without overtension.

Next, I needed to construct multiple 15-lb weights at minimum cost—weights that had to be long and narrow to avoid collisions with the supporting trees and other nearby objects. I decided that slumpblock bricks inside a 2 × 4 wood frame would be satisfactory. (Slump blocks are long, narrow and about 12 lbs each; some are just as wide as a standard 2 × 4.) Including bricks, wood and nails, each weight cost me about $1.50, so the price was right—and 2 × 4s can be easily drilled for attaching ropes.

The relay box. A square relay box centered between three antennas twists the antennas' open-wire feeders at the feed points. I solved this problem for a triangular, relay-switched vertical array[1] and my three-EDZ system by building triangular relay boxes (Figure 2). (Use redwood, cedar or cypress and weatherproof the box with sealant.[2]) Plexiglass plates and screws insulate the feeder wires from the box, continue the feeders to the relays and provide strain relief (Fiureg 3).

Results with the system. Using a 600-W amplifier, I often get through pileups first on long path. Short path to Europe is more difficult because the ground slopes slightly upward in that direction. My blissful encounters with nature's gift to the RF enthusiast continue!–*W7SX*

[1] See Volume 3 of *The ARRL Antenna Compendium*.

[2] In my work with wire antennas, I discovered a reference on wood that may be of interest to amateurs: *Wood Handbook: Wood as an Engineering Material* (Washington, DC: US Department of Agriculture, rev 1987), 466 pages (for sale by the Superintendent of Documents, US Government Printing Office, Washington, DC 20402). This text describes, among many mechanical and structural considerations, electrical qualities of wood that come into play when wood supports are used in RF fields.

Collinear Arrays

Collinear arrays are always operated with the elements in phase. (If alternate elements in such an array are out of phase, the system simply becomes a harmonic type of antenna.) A collinear array is a broadside radiator, the direction of maximum radiation being at right angles to the line of the antenna.

Power Gain

Because of the nature of the mutual impedance between collinear elements, the feed point resistance is increased as shown earlier in this chapter (Figure 9). For this reason the power gain does not increase in direct proportion to the number of elements. The gain with two elements, as the spacing between them is varied, is shown by Figure 38. Although the gain is greatest when the end-to-end spacing is in the region of 0.4 to 0.6 λ, the use of spacings of this order is inconvenient constructionally and introduces problems in feeding the two elements. As a result, collinear elements are almost always operated with their ends quite close together–in wire antennas, usually with just a strain insulator between.

With very small spacing between the ends of adjacent elements the theoretical power gain of collinear arrays is approximately as follows:

2 collinear elements—1.6 dB
3 collinear elements—3.1 dB
4 collinear elements—4.2 dB

More than four elements are rarely used.

Directivity

The directivity of a collinear array, in a plane containing the axis of the array, increases with its length. Small secondary lobes appear in the pattern when more than two elements are used, but the amplitudes of these lobes are low enough so that they are not important. In a plane at right angles to the array the directive diagram is a circle, no matter what the number of elements. Collinear operation, therefore, affects only E-plane directivity, the plane containing the antenna. At right angles to the wire the pattern is the same as that of the individual 1/2-λ elements of which it is composed.

When a collinear array is mounted with the elements vertical, the antenna radiates equally well in all geographical directions. An array of such "stacked" collinear elements tends to confine the radiation to low vertical angles.

If a collinear array is mounted horizontally, the directive pattern in the vertical plane at right angles to the array is the same as the vertical pattern of a simple 1/2-λ antenna at the same height (Chapter 2).

TWO-ELEMENT ARRAY

The simplest and most popular collinear array is one using two elements, as shown in Figure 39. This system is commonly known as "two half-waves in phase." The manner in which the desired current distribution is obtained is described in Chapter 26. The directive pattern in a plane containing the wire axis is shown in Figure 40.

Depending on the conductor size, height, and similar factors, the impedance at the feed point can be expected to be in the range from about 4 to 6 kΩ, for wire antennas. If the elements are made of tubing having a low λ/dia (wavelength to diameter) ratio, values as low as 1 kΩ are representative. The system can be fed through an open-wire tuned line with negligible loss for ordinary line lengths, or a matching section may be used if desired.

THREE- AND FOUR-ELEMENT ARRAYS

When more than two collinear elements are used it is necessary to connect "phasing" stubs between adjacent elements in order to bring the currents in all elements in phase. It will be recalled from Chapter 2 that in a long wire the direction of current flow reverses in each 1/2-λ section. Consequently, collinear elements cannot simply

Figure 38—Gain of two collinear 1/2-λ elements as a function of spacing between the adjacent ends.

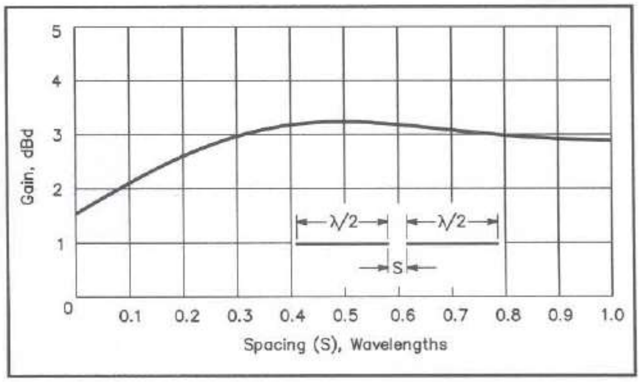

Figure 39—A two-element collinear array (two half-waves in phase). The transmission line shown would operate as a tuned line. A matching section can be substituted and a nonresonant line used if desired.

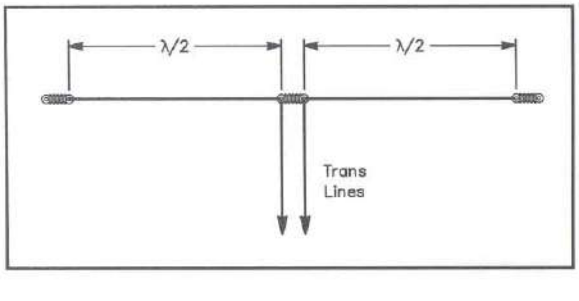

be connected end to end; there must be some means for making the current now in the same direction in all elements. In Figure 41A the direction of current flow is correct in the two left-hand elements because the transmission line is connected between them. The phasing stub between the second and third elements makes the instantaneous current direction correct in the third element. This stub may be looked upon simply as the alternate $^1/_2$-λ section of a long-wire antenna folded back on itself to cancel its radiation. In Figure 41A the part to the right of the transmission line has a total length of three half wavelengths, the center half wave being folded back to form a $^1/_4$ λ phase-reversing stub. No data are available on the impedance at the feed point in this arrangement, but various considerations indicate that it should be over 1 kΩ.

An alternative method of feeding three collinear elements is shown in Figure 41B. In this case power is applied at the center of the middle element and phase-reversing

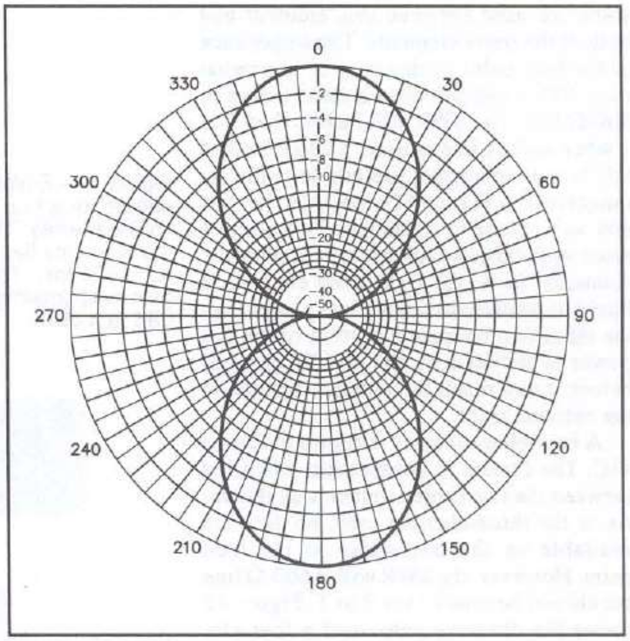

Figure 40—Free-space E-plane directive diagram for the two-element collinear array of Fig 39. The axis of the elements lies along the 90°-270° line. This is the horizontal pattern at low wave angles when the array is horizontal. The array gain is approximately 1.6 dBd (3.8 dBi).

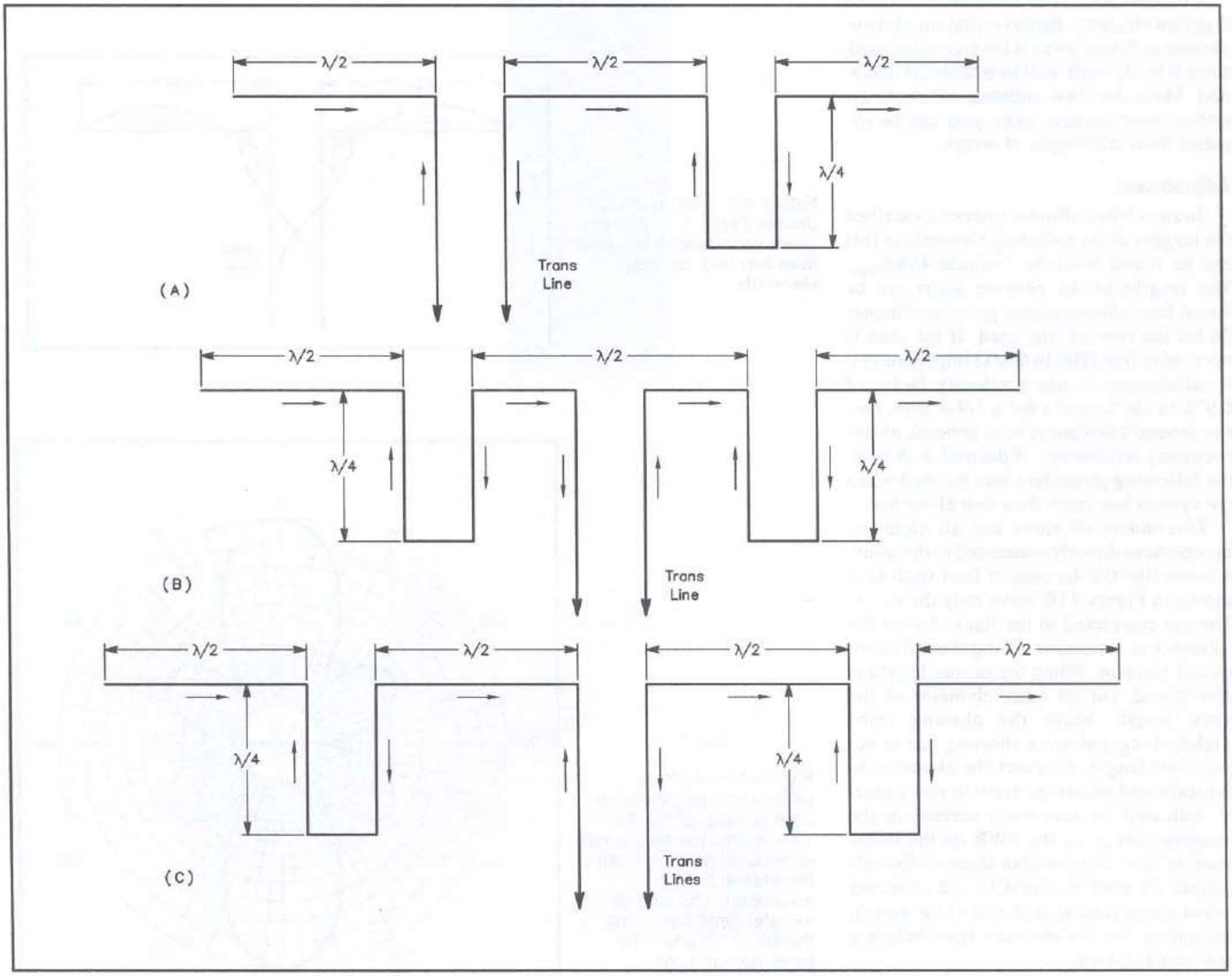

Figure 41—Three and four-element collinear arrays. Alternative methods of feeding a three-element array are shown at A and B. These drawings also show the current distribution on the antenna elements and phasing stubs. A matched transmission line can be substituted for the tuned line by using a suitable matching section.

stubs are used between this element and both of the outer elements. The impedance at the feed point in this case is somewhat over 300 Ω and provides a close match to 300-Ω line. The SWR will be less than 2 to 1 when 600-Ω line is used. Center feed of this type is somewhat preferable to the arrangement in Figure 41A because the system as a whole is balanced. This assures more uniform power distribution among the elements. In A, the right-hand element is likely to receive somewhat less power than the other two because a portion of the fed power is radiated by the middle element before it can reach the element located at the extreme right.

A four-element array is shown in Figure 41C. The system is symmetrical when fed between the two center elements as shown. As in the three-element case, no data are available on the impedance at the feed point. However, the SWR with a 600-Ω line should not be much over 2 to 1. Figure 42 shows the directive pattern of a four-element array. The sharpness of the three-element pattern is intermediate between Figures 40 and 42, with four small minor lobes at 30° off the array axis.

Collinear arrays can be extended to more than four elements. However, the simple two-element collinear array is the type most used, since it lends itself well to multiband operation. More than two collinear elements are seldom used because more gain can be obtained from other types of arrays.

Adjustment

In any of the collinear systems described the lengths of the radiating elements in feet can be found from the formula $468/f_{MHz}$. The lengths of the phasing stubs can be found from the equations given in Chapter 26 for the type of line used. If the stub is open-wire line (500 to 600 Ω impedance) it is satisfactory to use a velocity factor of 0.975 in the formula for a 1/4-λ line. On-the-ground adjustment is, in general, an unnecessary refinement. If desired, however, the following procedure may be used when the system has more than two elements.

Disconnect all stubs and all elements except those directly connected to the transmission line (in the case of feed such as is shown in Figure 41B leave only the center element connected to the line). Adjust the elements to resonance, using the still-connected element. When the proper length is determined, cut all other elements to the same length. Make the phasing stubs slightly long and use a shorting bar to adjust their length. Connect the elements to the stubs and adjust the stubs to resonance, as indicated by maximum current in the shorting bars or by the SWR on the transmission line. If more than three or four elements are used it is best to add elements two at a time (one at each end of the array), resonating the system each time before a new pair is added.

THE EXTENDED DOUBLE ZEPP

An expedient that may be adopted to

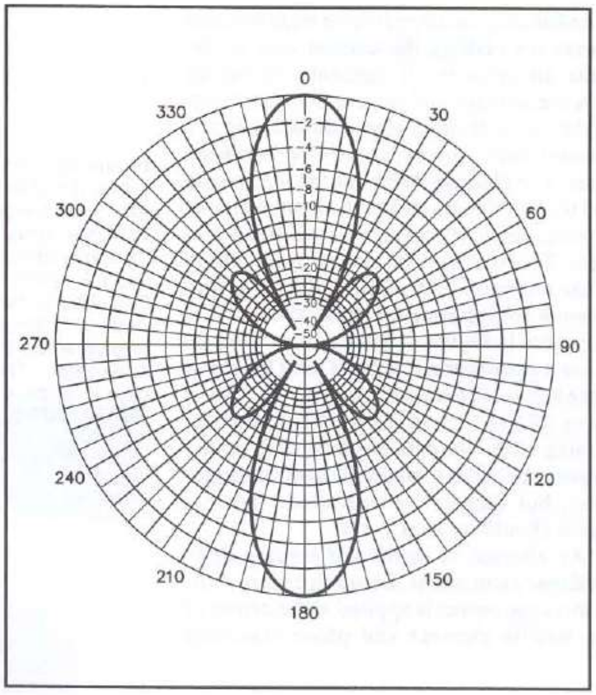

Figure 42—E-plane pattern for a four-element collinear array. The axis of the elements lies along the 90°-270° line. The array gain is approximately 4.2 dBd (6.4 dBi).

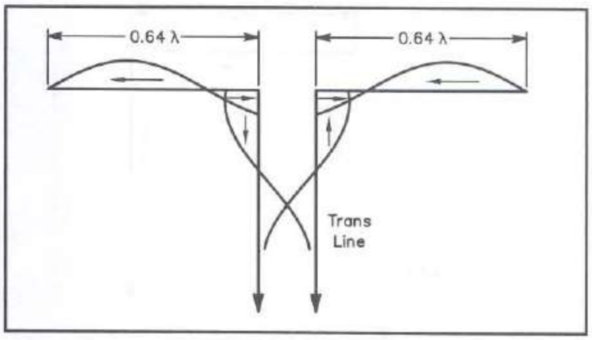

Figure 43—The extended double Zepp. This system gives somewhat more gain than two 1/2-λ collinear elements.

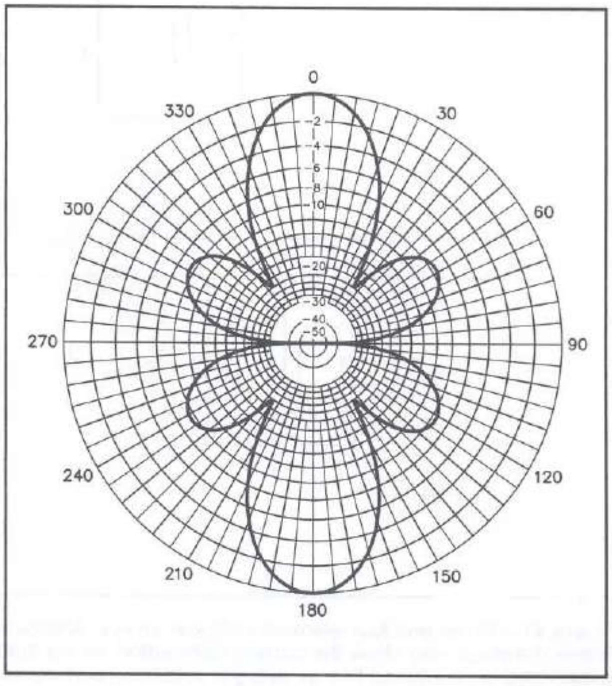

Figure 44—E-plane pattern for the extended double Zepp of Fig 43. This is also the horizontal directional pattern when the elements are horizontal. The axis of the elements lies along the 90°-270° line. The array gain is approximately 3 dBd.

obtain the higher gain that goes with wider spacing in a simple system of two collinear elements is to make the elements somewhat longer than $^1/_2$ λ. As shown in Figure 43, this increases the spacing between the two in-phase $^1/_2$-λ sections at the ends of the wires The section in the center carries a current of opposite phase but if this section is short the current will be small; it represents only the outer ends of a $^1/_2$-λ antenna section. Because of the small current and short length, the radiation from the center is small. The optimum length for each element is 0.64 λ. At greater lengths the system tends to act as a long-wire antenna, and the gain decreases.

This system is known as the "extended double Zepp." The gain over a $^1/_2$-λ dipole is approximately 3 dB, as compared with approximately 1.6 dB for two collinear $^1/_2$-λ dipoles. The directional pattern in the plane containing the axis of the antenna is shown in Figure 44. As in the case of all other collinear arrays, the free-space pattern in the plane at right angles to the antenna elements is the same as that of a $^1/_2$-λ antenna—circular.

CHAPTER SIX

WIRE BEAMS

6-1 **The K6NA 80-Meter Wire Beam**
From *The ARRL Antenna Compendium Volume 5*

6-10 **Construct a Wire Log-Periodic Dipole Array for 80 or 40 Meters**
From August 1986 *QST*

6-14 **Input Impedance of LPDA Antennas**
From October 1986 *QST* (Technical Correspondence)

6-15 **The WA1AKR 40- and 75-Meter Slopers**
From August 1980 *QST* (Technical Correspondence)

6-16 **Inexpensive 30-Meter Beam Antenna**
From June 1983 *QST* (Hints & Kinks)

6-17 **Try the "FD Special" Antenna**
From June 1984 *QST*

6-21 **Build a 4X Array for 160 Meters**
From February 1985 *QST*

6-25 **The 'BRD Zapper: A Quick, Cheap and Easy "ZL Special" Antenna**
From June 1990 *QST*

6-27 **Combination Driven Arrays**
From *The ARRL Antenna Book* 18th edition

By Glenn Rattmann, K6NA **From *The ARRL Antenna Compendium Volume 5***

The K6NA 80-Meter Wire Beam

Ever get crunched in a DX pileup on 80/75 meters by K6NA? Lots of folks have! K6NA reveals the secrets behind his two-element "killer" Yagi.

This article describes a practical design for a two-element, horizontal wire Yagi for the 75/80-meter band using just two support towers. Designed with DXing and contesting in mind, this radiator-and-reflector combination features moderate gain and good front-to-back ratio, combined with instant beam reversal and band-segment (mode) switching. It uses tuned feeders and a balanced-output antenna tuner at the operating position for matching at the transmitter. Remote relays at ground level handle reversal and parasitic-tuning requirements, taking advantage of the convenient properties of half-wave transmission lines.

I designed this antenna empirically, before PC-modeling programs were readily available. I've confirmed that it really works, through years of successful 80-meter operation. Later, I verified this with computer modeling. Figure 1 shows the general layout of my system.

Background

At my former location I had two towers, 70 and 100-feet high, fairly close together on a half-acre of property. There was an assortment of quads and Yagis for 10-40 meters,[1] but as is often the case with all-band DXers, the 80-meter antenna was an afterthought. Through the years I experimented with many types, looking for the "silver bullet" antenna for 80 that would finally make this challenging band a pleasure to operate.

Each season numerous magazine articles and endless discussions by experts extolled the virtues of this or that magic 80-meter DX antenna. I even tried some of them: a diamond quad loop, a delta loop (bottom-fed or corner-fed), a pair of parasitic deltas, phased quarter-wave slopers (tilted ground planes), a $\lambda/4$ vertical with lots of radials. I always came back to a basic truth: with the tower heights available, and faced with the reality of rough, rocky, lossy earth, my 95 foot high inverted-V dipole was usually as good as–or better than—any of the "trick" antennas I had spent hundreds of hours building. As a bonus, usually the dipole was quieter on receive.

Even Jim Lawson, W2PV, in an early *QST* article about broadbanding an 80-meter antenna, mentioned almost in passing that his high (110-foot) inverted-V was superior to his four-element phased vertical array, at least on transmit.[2] What I observed about 80-meter DXing in the 1970s was that most attempts using vertically polarized systems to achieve low-angle radiation were not satisfactory. This was due to excessive ground losses (both near-field return-current losses and far-field reflective losses); in addition, verticals were generally noisy on receive.

Neither vertically polarized antennas nor relatively low horizontal antennas were really getting thejob done. I resolved to construct my next station thinking about the 80-meter band from the start."High and horizontal" would be the goal.

Design Criteria

My new station would have two main towers, each 140 feet high. Because I had all-band contesting in mind, there would be numerous other antennas on the towers. I wanted to put a three-element 40-meter beam on top of one tower and a large 20-meter Yagi on the top of the other. Though it likely would be an excellent performer, a rotating beam of some type for 80 meters was ruled out because I did not want to dedicate the top of one tower to it. Also, there was the near-certainty of difficult maintenance problems with such a gigantic antenna. Instead, I settled on a design for a fixed-wire beam. These would be the main criteria for reliable, day-to-day operation:

- Oriented properly for best results in both DX and domestic contests
- Moderate gain, with a wide lobe for good azimuth coverage
- Instant beam reversal with a decent front to-back ratio
- Near-instant band-segment (phone or CW mode) switching, especially for casual, daily DXing
- Easy, low-loss matching anywhere in the band
- Horizontally polarized to minimize both ground-reflection losses and noise on receive
- High enough to produce reasonably low angle radiation
- Reliable and easy to maintain without disrupting other antennas on the towers.

Design Discussion

Some discussion of these interrelated design criteria will aid in understanding the trade-offs and choices I made while planning this antenna. For example, from San Diego the typical short path to Europe on 80 meters *is not* the expected true Great-Circle path of about 25°. Rather, the short path toward Europe is almost always a "bent" or skewed path, where signals pass to the south or southeast of the highly absorptive auroral oval, propagating by means of a scatter mode that uses ionized patches over the central Atlantic Ocean. Consequently, our pseudo short-path heading toward Europe is commonly 50° to 90° in azimuth. Occasionally, European signals on 80 meters arrive in California from the southeasterly direction. Rarely, around the time of the Equinox, they may even arrive by scatter path from straight south, as reported by some rotary beam users in California.

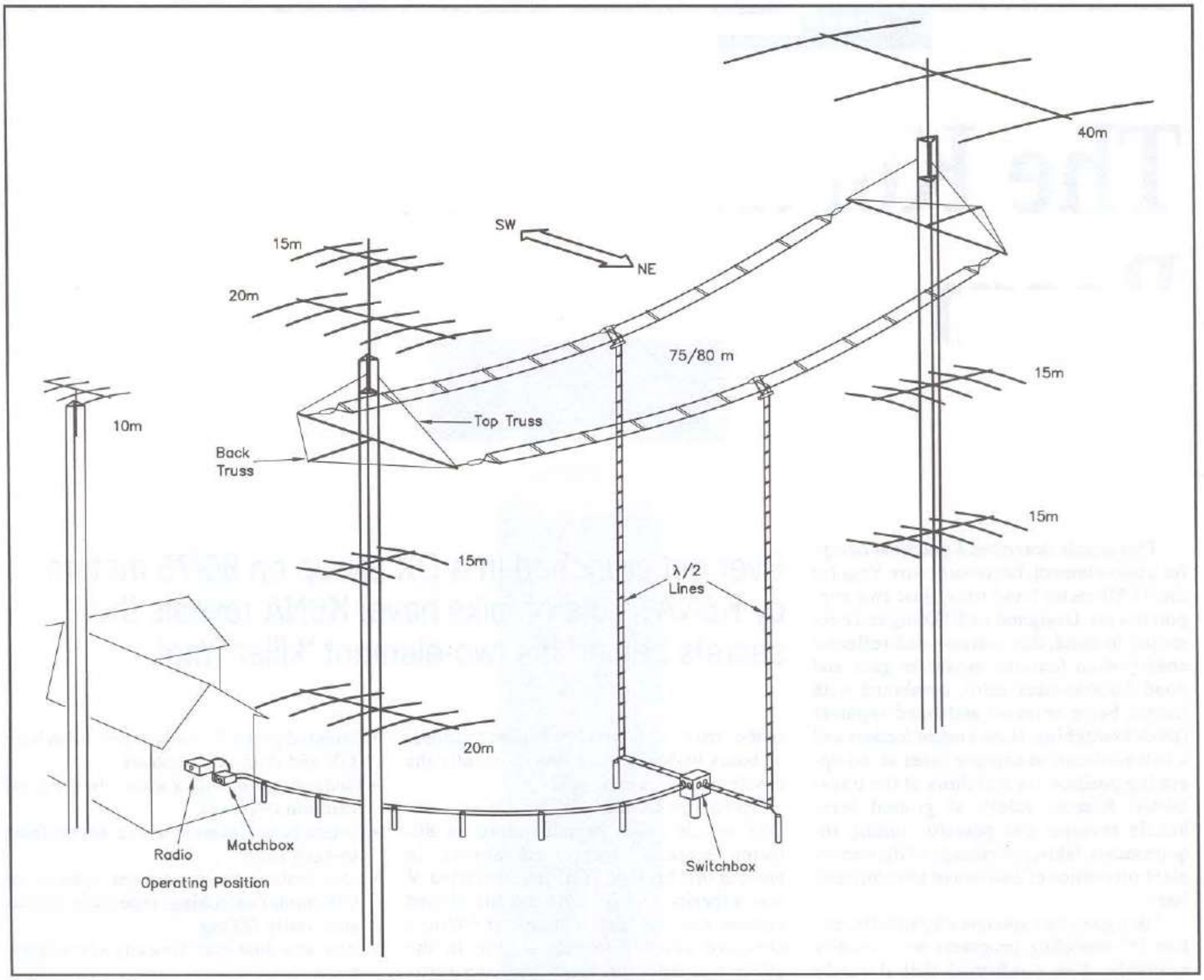

Figure 1—The antenna farm at K6NA, showing the 80-meter wire beam installed on the pair of 140-foot towers, using 35-foot long crossarms made of aluminum tubing and fiberglass. The rearward truss and upper truss wires make each crossarm rigid. For clarity, the tower guy wires are not shown.

I decided to orient the two support towers so that the reversible wire beam, slung between the towers on a pair of horizontal crossarms, would point at 60°/240°. This would provide good coverage across the USA and into Europe (both short and long path), Africa, the Caribbean and South Pacific areas, which together make up about 85% of my annual QSOs on 80 meters.

I could build a simple two-element parasitic design, using λ/2 elements on a reasonable boomlength of about 35 feet, or λ/8.[3] Figure 2B shows that such a design has a half-power azimuth beamwidth of about 74° in the azimuth plane. I was willing to accept the slightly lower gain and front-to-back ratio compared to a three-element design (which would be more difficult mechanically), or a two-wire phased array using double-extended Zepp elements. The Zepp would require the towers to be placed about 350 feet apart. Either of these antennas conceivably would provide more gain, but at the price of reduced azimuthal coverage. A reversible, all-driven array has the additional disadvantage that its phasing and matching networks become overly complicated to make it work properly at both ends of the band.

To achieve the twin requirements of easy mode-switching and beam reversal, I decided to construct a symmetrical, two-element Yagi beam. The elements are identical dipoles, self-resonant near 3.675 MHz. This frequency is about 3% to 4% lower than the phone operating area around 3.800 MHz, meaning that each element could be used as a reflector in the phone band with no additional loading. Each dipole is fed with λ/2 of open-wire line. These lines come together in a relay box at ground level. Power is applied to one or the other driven element, and a short is applied across the base of the remaining feeder to force the remaining element to act as a parasitic reflector.

To operate near 3.5 MHz, the reflector simply is retuned to about 3.4 MHz, using an inductor in place of the direct short in the relay box. Supplying RF to the relay box is accomplished using low-loss, open-wire feeders. A balanced-output tuner in the shack is simply touched up to move between phone and CW subbands, and to accommodate the change in SWR. There is no requirement for an exact match at the line input to the driven element because of the low-loss transmission lines.

With the crossarm supports mounted at 130 ft, the dipoles are about λ/2 high and thus produce fairly low-angle radiation. A horizontally polarized antenna is less responsive to man-made noise sources, which have a large vertical component. The height allows λ/2 lines to reach down to the switchbox near ground level. All switching relays are in this weatherproof box, so the system is reliable and easy to maintain. No inaccessible relays or matching networks are up at the elements, where failures might occur.

Construction

No doubt, the 80-meter wire beam can be built and integrated into many existing

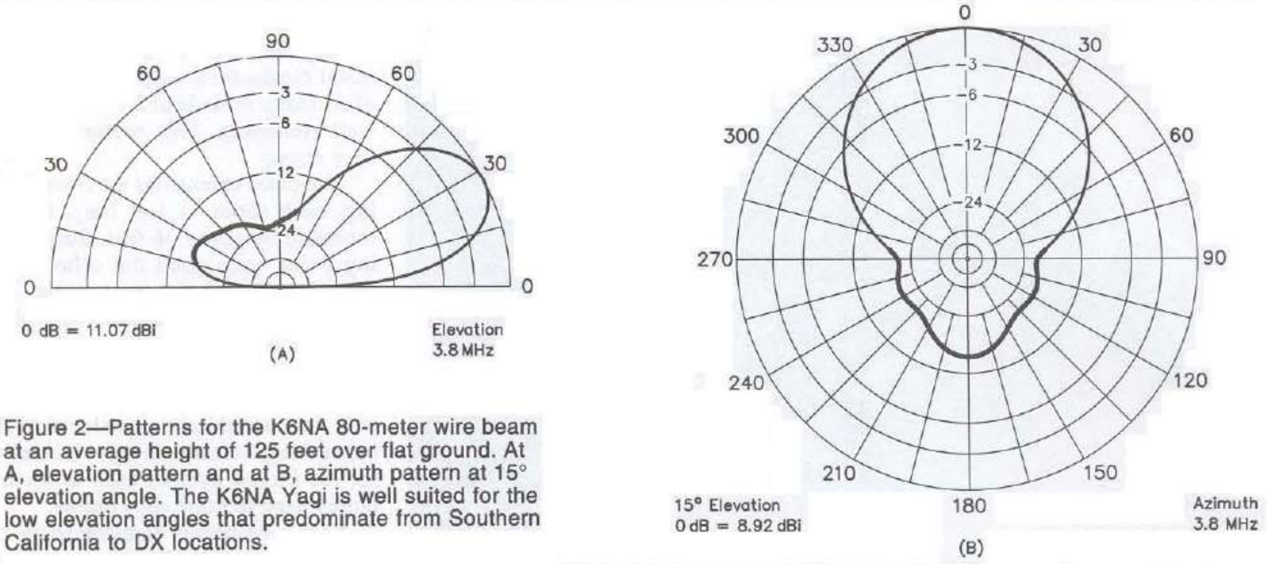

Figure 2—Patterns for the K6NA 80-meter wire beam at an average height of 125 feet over flat ground. At A, elevation pattern and at B, azimuth pattern at 15° elevation angle. The K6NA Yagi is well suited for the low elevation angles that predominate from Southern California to DX locations.

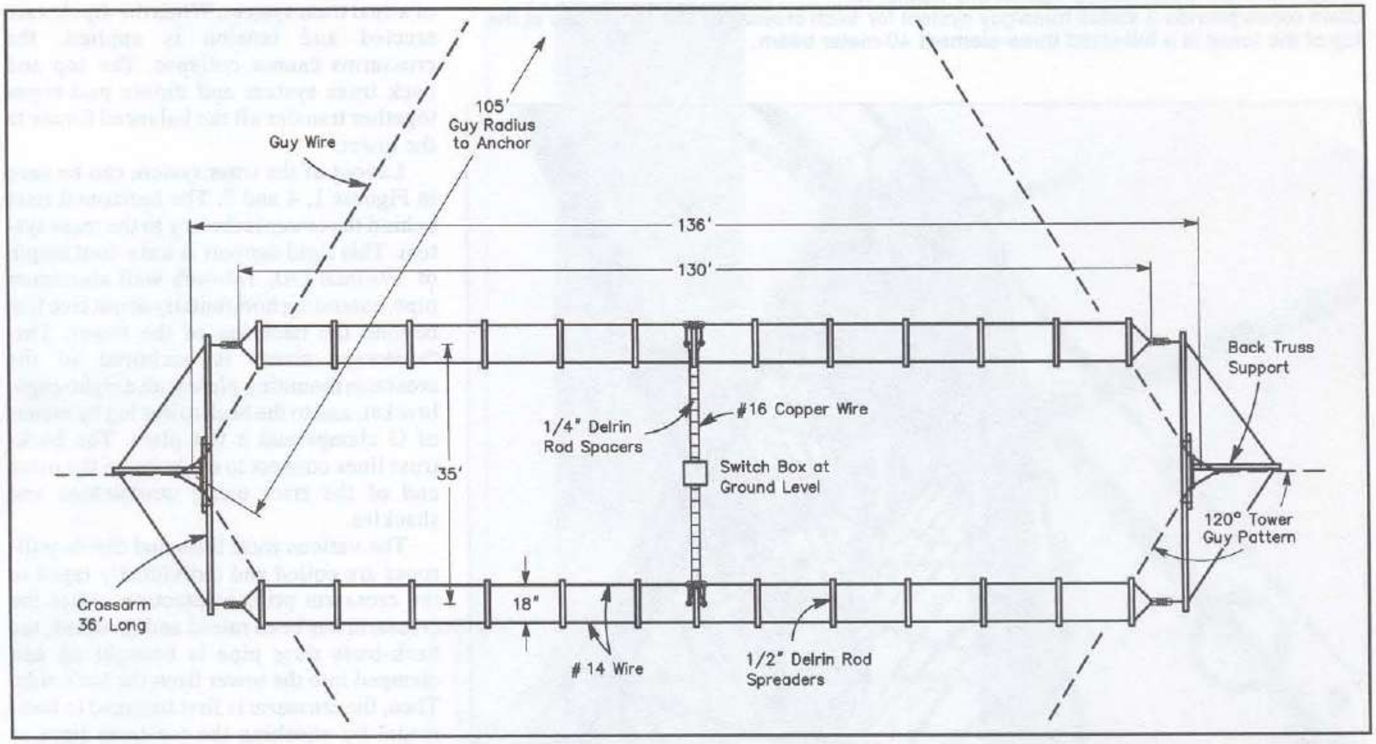

Figure 3—Top view of the 80-meter wire beam mounted on the towers, with details of the crossarms and rearward truss system that prevents the tensioned antenna wires from bending the crossarms inward. The tower-guy placement geometry provides several advantages for antenna-farm management! See text.

multi-tower systems. I had the luxury of planning the tower and guy arrangement well in advance of building the antenna. This helped make erection and maintenance simpler. Such planning contributes to convenient erection of other antennas too!

Figure 3 shows the recommended two-tower layout that has served very well for this and other projects. Through symmetry of guy-anchor placement, any forces due to loading by wire antennas or temporary trams are balanced out. Note that each tower has an opposing, backward-facing guy exactly in line with its counterpart and parallel with the centerline of the towers.

When the 36-foot crossarm is mounted about 10 feet below the top of the tower and the topmost guy bracket, the ends extend out beyond and above the sloping, upper guys. This arrangement keeps the dipoles in the clear, allowing them to be raised and lowered using ropes and pulleys with minimal guy-wire interference, without having to remove or move the crossarms.

This tower/guy layout also presents two parallel tower faces on which to mount the booms and to correctly orient the wire antenna. Non-parallel tower faces can work, but the mounting brackets would be more complex and some advantages of symmetry would be lost. For safety, significant sideloads should be distributed across *at least two* tower legs. In addition, guys passing through the central area between the towers would complicate the raising and lowering of the wire antennas, so this should be avoided if possible.

Another advantage of the layout in Figure 3 concerns the overall use and maintenance of other antennas in the system. Note that the 140-foot towers are 136 feet apart, and the guy radius (distance from tower to each of its guy anchors) is 105 feet. These facts, together with the generally uncluttered work area between the towers, make

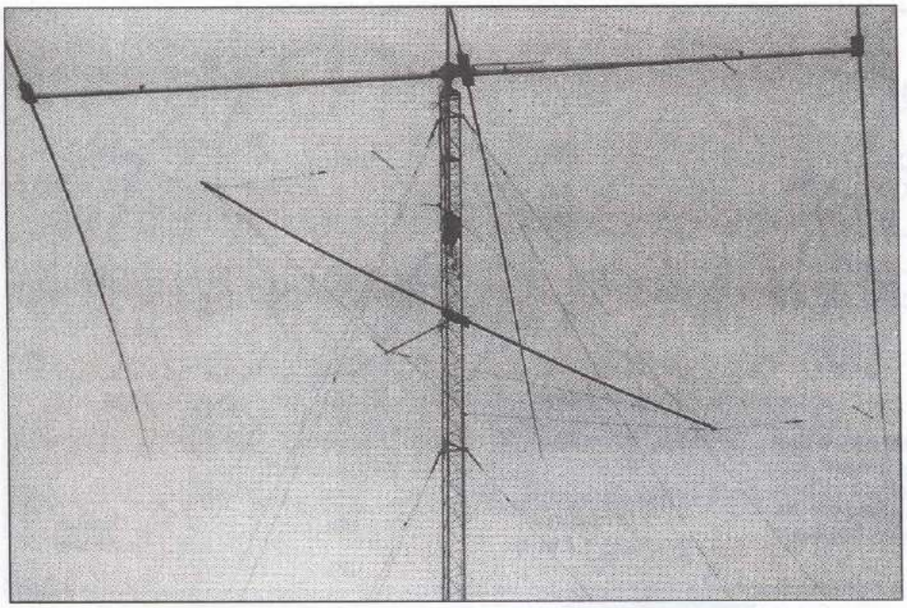

Figure 4—Photo of one of the crossarms (booms) mounted on the tower. The truss system and 80-meter folded dipoles are visible. Note that the truss, dipoles and pull-down ropes provide a stable truss/guy system for each crossarm. The large Yagi at the top of the tower is a full-sized three-element 40-meter beam.

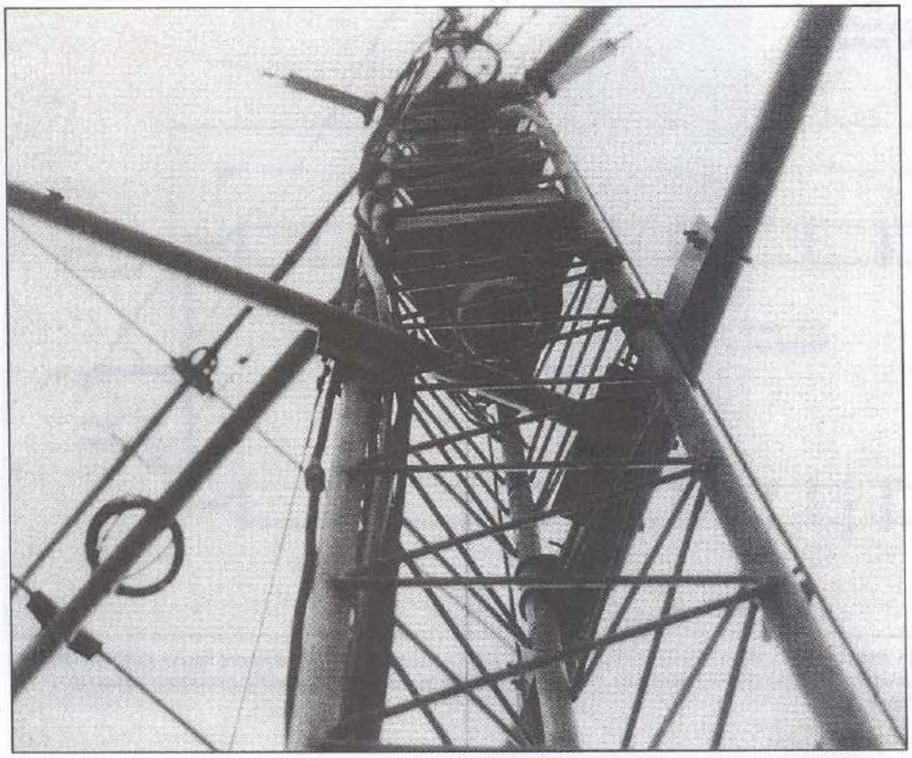

Figure 5—Photo showing mounting method for installing the 80-meter crossarm and associated horizontal back-truss pipe. The horizontal riser prevents the crossarm from collapsing when the side load (antenna tension) is applied. Note how a plate with U-bolts is used to secure the back-truss pipe to one of the tower legs, while the other end is secured with an L-shaped bracket and U-bolts to the mounting plate for the crossarm.

it convenient to run a temporary tram wire anchored at the base of one tower to the top (or anywhere along the length) of the other tower. The tram can be used to raise the 80-meter crossarms themselves, or to move other HF Yagis. If an upper-mast-mounted beam is to be erected, the 80-meter wires first can be lowered easily to the ground. If the tram has to reach only to the area of a side-mounted beam, the wire antenna may stay in place above it. I can detach the 80-meter feeders from the feed-through insulators at the switchbox, and pull them away from the vicinity of the tram wire, if necessary.

Crossarms and Trusses

I made each crossarm (boom) using aluminum and fiberglass. In order to eliminate the possibility of "boom resonance" that might interfere with nearby HF Yagis and to keep the 80-meter wire ends away from metal conductors, only the middle 14 feet of crossarm is made of 2-inch OD, 1/4-inch wall aluminum. This section is stiff and very strong.

The outer extensions are fiberglass tubing, each about 11 feet long, for a total crossarm length of 36 feet. Surplus vaulting poles were used but other types of fiberglass would also be suitable. The fiberglass should be wrapped with high-quality black tape (such as Scotch 33 or 88) to prevent decomposition due to ultraviolet radiation. At the end of each fiberglass pole, I mounted a six-inch sleeve of aluminum tubing, before drilling a clearance hole for the forged eyebolt used to secure the pulley used to hoist each dipole.

Each crossarm is clamped to the tower face using an aluminum plate and muffler clamps. See Figures 4 and Fig 5. Once in place, the crossarm is made rigid by means of a dual truss system. When the dipoles are erected and tension is applied, the crossarms cannot collapse. The top and back truss system and dipole pull-ropes together transfer all the balanced forces to the tower.

Layout of the truss system can be seen in Figures 1, 4 and 5. The horizontal riser behind the tower is the key to the truss system. This rigid support is a six-foot length of $1^1/_2$-inch OD, 1/4-inch wall aluminum pipe, extending horizontally about five feet beyond the back leg of the tower. This "sideways riser" is anchored to the crossarm mounting plate with a right-angle bracket, and to the back tower leg by means of U clamps and a flat plate. The back-truss lines connect to eyebolts on the outer end of the riser using turnbuckles and shackles.

The various truss lines and dipole pull-ropes are coiled and individually taped to the crossarm prior to erection. After the crossarm has been raised and installed, the back-truss riser pipe is brought up and clamped into the tower from the back side. Then, the crossarm is first trimmed to horizontal by attaching the top-truss lines to the tower legs and adjusting the turnbuckles until the crossarm looks good, much like a Yagi boom truss is adjusted. Then the backtruss lines are untaped from the crossarm, one at a time, and connected to the far end of the horizontal back-truss pipe. This is an interesting maneuver, which should be accomplished only by an experienced climber/rigger! The turnbuckles are adjusted initially for moderate tension. Properly adjusted, the result is a crossarm that doesn't sag downward or fold inward when antenna tension is applied.

After the truss system has been rigged, the installer untapes one of the coiled-up dipole pull-ropes, which has been pre-rigged to extend along the arm to the outer crossarm pulley and back again to the vicinity of the tower. By attaching either a weight or tag line to the thimble, the dipole-end of the rope is released to hang vertically from the pulley. The rigger then pays out the coiled rope until a ground man can reach the thimble end to attach the dipole insulator. The pull-ropes ultimately are brought to the tower legs through pulleys and thence brought down to ground level. The angle of the pull-ropes completes what is essentially a four-point guy on each end of the crossarm (antenna tension, back-truss, up-truss, and angled pull-down rope). This configuration presents a stable structure when the dipoles are raised and tension applied.

I must emphasize that all subassemblies should first be fabricated and fitted to the tower at ground level. Make sure that all plates, clamps and hardware mate to the tower, referenced to diagonal braces, welds, or leg bolts. Premark the tower legs up at the target location ahead of time to preclude surprises on the tower!

Construction of the Dipoles

Two-wire folded dipoles were selected to help bring the feedpoint impedance up closer to the characteristic impedance of the feeders and to improve bandwidth. This helps minimize tuner adjustments when moving within a subband, but there is no need to go overboard trying to effect a perfect match at the load.

Each folded dipole is built with #14 stranded, non-stretch copper wire. The two parallel wires are held apart by seven horizontal spreaders located about every 20 feet along the wire length. These spreaders of black Delrin are 1/2-inch diameter rod, 18 inches long (See Figure 6). They are held in position by tie wires, as are the ribbed end insulators. I used an unbroken length (about 260 feet) of wire for each two-wire dipole, making the free ends come together at the center-feed insulator. The free ends of the wires and feeders are tied to the center insulator for strain-relief, and then the loose ends are soldered. This technique prevents breakage, because soldered connections (which are weaker due to wire heating) are not under strain, or flexing. After each dipole is assembled on flat ground, it should be stretched temporarily at low level between the towers in order to check the symmetry and mechanical behavior. When adjustments are complete, the feeder can be soldered to the antenna.

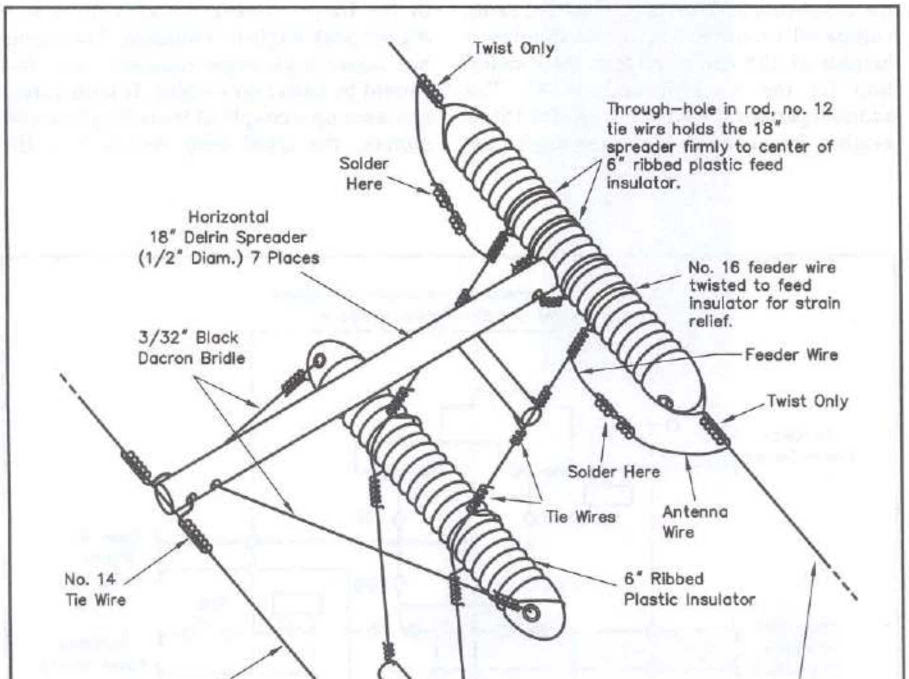
Twist Only
Through–hole in rod, no. 12 tie wire holds the 18" spreader firmly to center of 6" ribbed plastic feed insulator.
Solder Here
Horizontal 18" Delrin Spreader (1/2" Diam.) 7 Places
No. 16 feeder wire twisted to feed insulator for strain relief.
3/32" Black Dacron Bridle
Feeder Wire
Twist Only
Solder Here
Tie Wires
Antenna Wire
No. 14 Tie Wire
6" Ribbed Plastic Insulator
3" Delrin spacer (1/4" Diam.)
No. 18 Tie Wire
260' continuous length of antenna wire, passes through each end insulator (not shown) and terminates at feedpoint shown here.
No. 14 Stranded Antenna Wire
Feedline to Ground Level
To End Insulator (not shown)

Figure 6—Folded-dipole center feedpoint with stabilizing cradle. The cradle encourages stability in the large, two-wire dipoles, preventing twist. See text.

There is little point in making any sort of analyzer measurements at five feet off the ground because the results will be misleading. When the dipole is installed in the clear at a half-wavelength in height, it will perform as predicted. However, just to be sure, I grid-dipped one dipole at its final height through the half-wave line. I had first built the $\lambda/2$ transmission line and had checked for resonance near 3.650 MHz with a grid-dip meter. When the first dipole and its feeder are on frequency, the second dipole and feeder was cloned and the array was erected. The dipoles at K6NA are about 130 feet long.

Feed Line Construction

Each feed line was constructed using old techniques but modern materials. Minimum weight is very important. The lines were made about 130 feet long, using #16, prestretched solid copper wire. Spacers are 1/4-inch Delrin rods, three inches long and spaced every four feet along the feeder. Spacers were notched and drilled, and held in position using #18 copper tie wires through the holes. The result is a very lightweight, low-loss feeder with a characteristic impedance of about 575 Ω.[4] To minimize sag in the dipole, avoid using heavy ceramic spacers.

A good technique for constructing the long, open-wire feeder is to stretch two wires tightly between the towers about four feet above the ground. Space the wires three inches apart in a plane parallel to the earth. Precut all the tie wires and make a simple, four-foot long, interval-measuring rod from a piece of wood or tubing. With all the spacers, tie wires, rod and long-nose pliers in a tool bag, the installer begins at one end, moving easily along the line, installing spacers at the correct intervals. The feeder wires sit in the notched ends, while the tie wires pass through the holes and are twisted in place over the feeder wires. When complete, the suspended line is in a good position for checking resonance with a grid-dip meter, assuming the line is insulated from the towers. The final length of each feeder described here is about 128 feet and each weighs less than three pounds.

Some additional comments about feedpoint construction will be helpful to those planning assembly of such a large, wide-spaced folded dipole and associated feeder. Initially, the dipole was built with upper and lower wires, as depicted in most antenna books. An 18-inch Delrin rod connected the lower feed insulator to the upper wire in order to distribute the weight of the feeder onto both wires and to maintain correct spacing. This worked, but with unforeseen side effects. When the dipole was raised under tension, the result was a half-twist in the parallel wires about halfway back toward the end insulator. This twist was not due to "live" or "kinky" wire! The twist was due to a catenary sort of effect.

The solution to the twist problem is to avoid the upper wire/lower wire configuration, and allow all the spreaders (including the center spreader) to lay over into the horizontal plane. To encourage this, a simple cradle was constructed to distribute the weight of the feeder at the center of the two antenna wires. This forces the wires (and all the other balanced spacers) to form a plane parallel to the earth. Fig 6 shows the centering cradle below the Delrin rod. The feeder passes over the cradle insulator (tie wires take the load) and continues up to the feedpoint. Small Dacron lines tie the cradle insulator to the other end of the rod, keeping the load centered. These large folded dipoles have been well behaved, even in high winds.

Remote Switching

A plywood "doghouse" was constructed under the antenna center, and mounted about four feet above ground for convenience. Fig 7 is a schematic of the switching network housed in the box. Two DPDT relays control the direction of the beam pattern. The main feeder connects through K1 and K2 to one dipole, while the other dipole becomes a phone-band reflector due to the short across K3. In the shack, switch S1 activates K1/K2 for reversal. S2 is a phone/CW selector switch that allows K3 to retune the reflector element to about 3.4 MHz by means of an inductor. Relays are large, ceramic-base open-frame types.

The center-loading inductor for a CW reflector is best chosen experimentally. A coil is connected across K3, S2 is set for CW, and the tuner in the shack is adjusted for a match low in the CW band. From the tuner, a coax cable is routed temporarily back to the switchbox. At the box, the temporary coax feeds a receiver, which is used to monitor signals arriving off the back of the antenna around 3.515 MHz. A variable tap is moved along the coil until the best front-to-back ratio is obtained. After a period of testing, the tap can be soldered. With the dipoles self-resonant around 3.650 MHz, a CW reflector coil of about 17 μH was required. This consisted of a 2-inch OD coil, $2\frac{3}{4}$-inches long, on which were wound about 25 turns of #16 wire.

Operation and Maintenance

The 80-meter wire beam performs as expected. Typical F/B is 11 to 14 dB on both phone and CW modes. For high-band operators accustomed to quiet bands with low atmospherics, this level of F/B may not seem like much. However, on a consistently noisy band like 80 meters, this kind of rejection on receive is very important. The high wire beam makes it possible to work long-path Europeans (southwest heading, before sunrise in San Diego) on a regular basis in the winter months, due to effective rejection of storm noise over North America. The antenna performance is roughly comparable to that of a nearby, two-element rotary beam (195 feet high) at N6ND—we hear and work many of the same weak stations during DX sessions. In domestic contests, reports from the East Coast are excellent on both modes.

The wire beam has been very reliable for over 10 years. The first version of the antenna used # 16 solid-copper wire for the elements, and there was a broken wire or two after perhaps six years. The antenna was rebuilt in 1992 with #14 stranded wire and has been perfect since then. The original dipole pull-ropes of black, 1/4-inch polypropylene lasted about seven years before they were replaced. Black Dacron, readily available from several *QST* advertisers, would be an excellent replacement for the pull-ropes, but it is more expensive than polypropylene. Do not use any color except black, and do not use polyethylene lines for permanent antenna supports.

Antenna Comparisons and Modeling

Recently, I wanted to investigate this 80-meter wire beam using computer modeling. By comparing the predicted performance of this antenna with that of several other common types of 80-meter DX antennas, the relative merits of each can be described briefly. Example antennas were modeled over average earth, with a dielectric constant of 13 and a ground conductivity of 5 mS/m. Gain figures listed are referenced to an isotropic antenna in free space.

Fig 8A shows the elevation pattern of the wire beam 125 feet high over real earth, compared to single horizontal dipoles at heights of 125 and at 70 feet. Peak radiation for the Yagi occurs at 28°. The azimuth pattern shown in Fig 8B is at 15° elevation, because low elevation angles are required for success on most DX paths from Southern California. For example, the W6 short path to Europe supports signals between 8° and 18° elevation, essentially 100% of the time.[5] The Yagi has about 5 dB of gain over the high dipole.

Note that a single high dipole is an excellent DX performer, and likely will beat any sort of beam antenna on a 70 to 80-foot tower. In addition, a high dipole has better rejection of local (high-angle) signals, so it will be a superior DX receiving antenna. Inspection of Figure 8A indicates that the high wire beam probably will outperform a 70-foot high dipole by nearly 10 dB at low elevation angles.

Some people might say: "With 140-foot towers, you should be using quad loops!" Not so fast. Figure 9A shows that with the top of a diamond-like, full-wave loop at 135 feet, the $\lambda/2$ dipole at 125 feet is superior by at least 1 dB at all angles lower than 40°, in spite of the small stacking gain expected from the loop. The expected gain from the bigger loop cannot be realized fully because the upper and lower wires are at drastically different heights (in terms of fractional wavelengths) relative to each other and the earth. This prevents proper phase addition. Further, the average height of the loop is lower, which results in a higher peak angle of radiation. The dipole has better high-angle rejection, too, and should be better on receive. If both antennas were up a couple of wavelengths on 80 meters, the quad loop would beat the

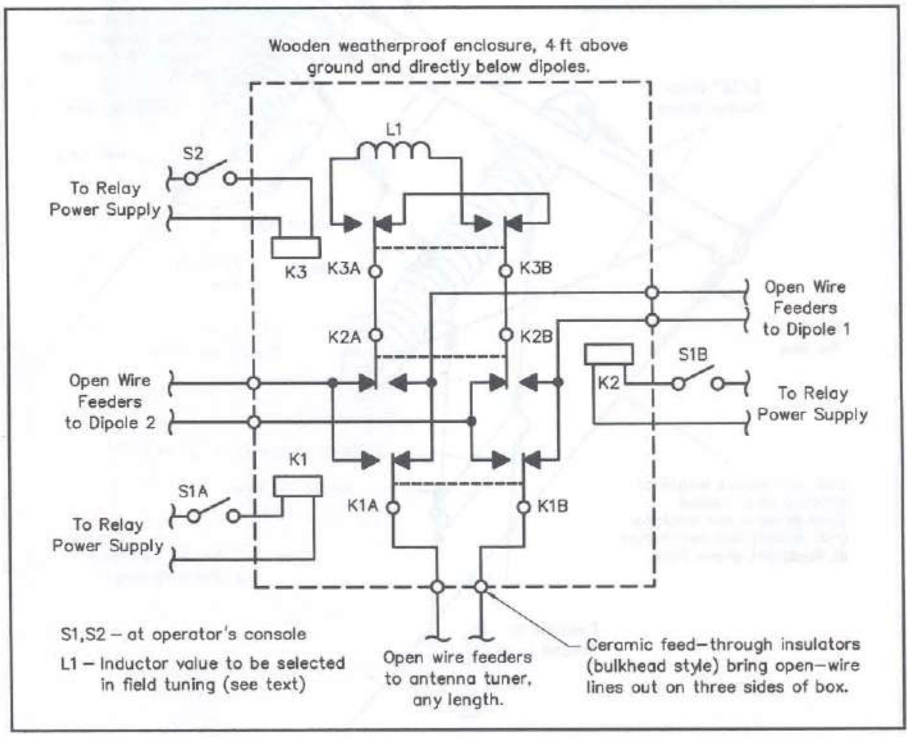
Wooden weatherproof enclosure, 4 ft above ground and directly below dipoles.
L1
S2
To Relay Power Supply
K3
K3A
K3B
Open Wire Feeders to Dipole 1
K2A
K2B
S1B
K2
Open Wire Feeders to Dipole 2
To Relay Power Supply
K1
S1A
K1A
K1B
To Relay Power Supply
S1,S2 – at operator's console
L1 – Inductor value to be selected in field tuning (see text)
Open wire feeders to antenna tuner, any length.
Ceramic feed–through insulators (bulkhead style) bring open–wire lines out on three sides of box.

Figure 7—Schematic of the feed system for the reflector-type, 80-meter wire Yagi. With two identical dipoles tuned at 3.650 MHz, only three relays and one inductor are needed to shift the antenna instantly from the phone to the CW subband, and to switch the beam direction. The relays are open-frame DPDT relays with ceramic insulation. S1 is a DPST switch used to reverse the beam direction by activating K1 and K2. S2 activates K3 to tune the parasitic element as a CW reflector.

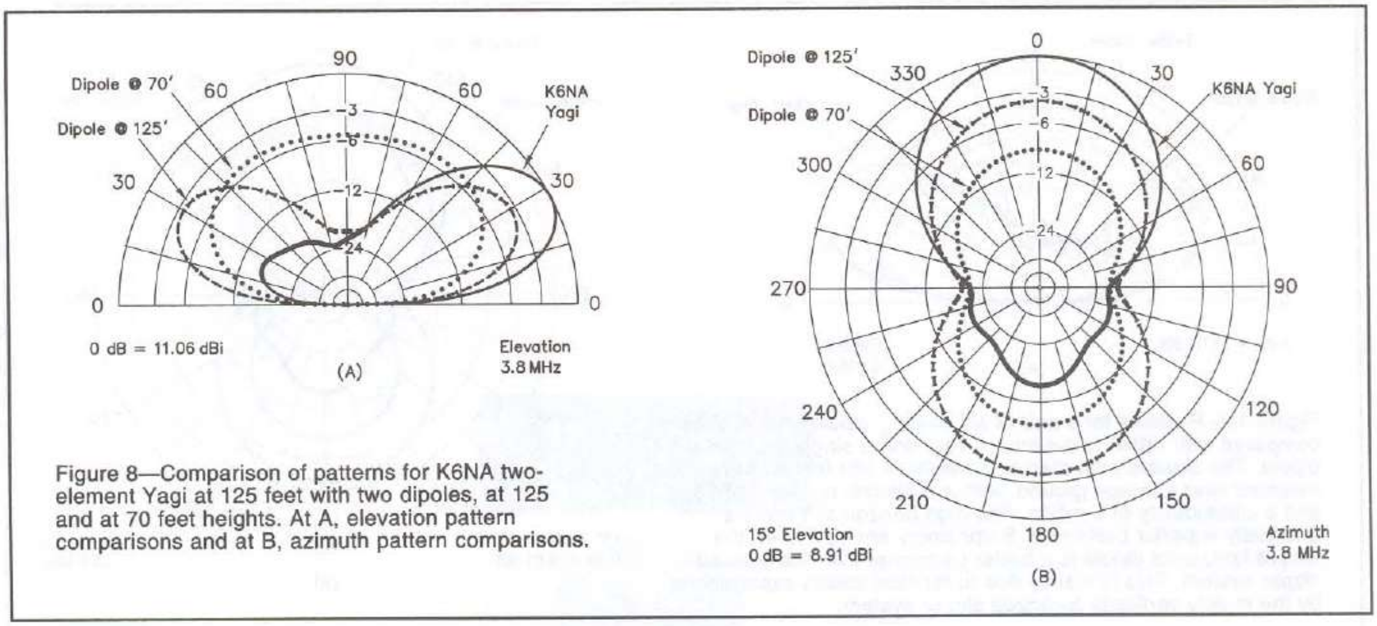

Figure 8—Comparison of patterns for K6NA two-element Yagi at 125 feet with two dipoles, at 125 and at 70 feet heights. At A, elevation pattern comparisons and at B, azimuth pattern comparisons.

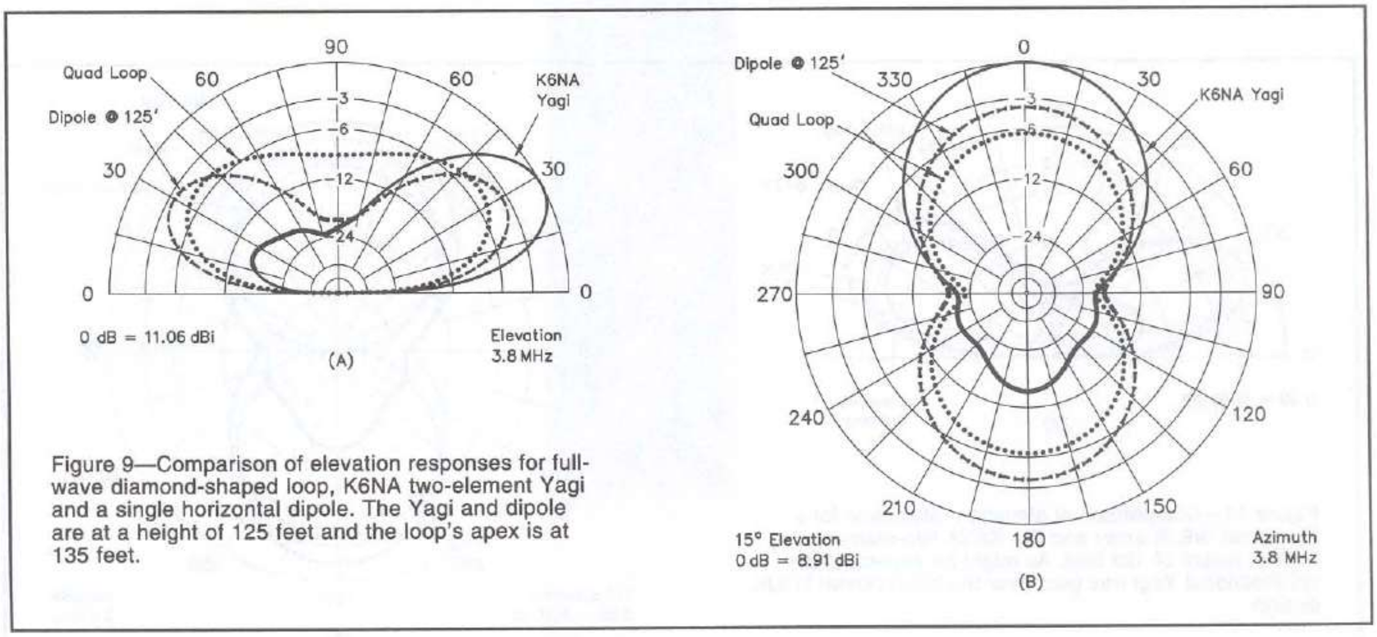

Figure 9—Comparison of elevation responses for full-wave diamond-shaped loop, K6NA two-element Yagi and a single horizontal dipole. The Yagi and dipole are at a height of 125 feet and the loop's apex is at 135 feet.

dipole...but who has a 550-foot tower?

How about a pair of half-wave slopers (that is, tilted vertical dipoles) tied from the tops of the 140-foot towers one λ/2 apart, and fed in phase for broadside gain? See Figure 10A. All the antenna books show the phased-sloper array to be a 4 dB gain, low-angle monster. The author had this exact antenna up for a few months in 1980. It performed fairly well, but was very responsive to powerline noise. For comparison, a single horizontal dipole was hung between the towers a few feet above the top ends of the slopers. With this single dipole, my local noise level dropped by 8 to 10 dB, and stations in New Zealand reported a consistent transmit advantage over the sloper array. This was a real eye-opener. As seen in Fig 10A, the high wire beam *buries* the phased slopers by 7 dB at the peak angle of 28°, and by 5 dB at 15° elevation!

The famous W8JK is a bi-directional, all-driven wire beam made from two dipoles fed 180° out of phase. At 125 feet in the air, this antenna has excellent rejection of high-angle signals, and is only slightly down from the peak gain of the Yagi configuration. See Figure 11A. It requires no relay box. For DXing, it is inferior to the Yagi because there is no way to reject signals (noise) off the back. Years ago, the author fed the two new folded dipoles as a W8JK for a few months prior to building the switchbox. It performed much like the patterns show, but once the switchbox was available to configure the array as a Yagi with good F/B ratio, the W8JK configuration was abandoned. However, a W8JK antenna 100-feet high in Kansas City would be an outstanding Sweepstakes antenna.

Figure 12 compares the horizontal wire beam to a similar antenna with inverted-V elements with a 90° included angle. The V-style gain is down a bit and the pattern is not as clean. But if you have a single tall tower and install this antenna (only one crossarm is required), you will be in the 95th percentile of effective 80-meter DXing antennas.

In Figure 13 the K6NA Yagi is compared to a quarter-wave vertical mounted over average ground. Since there are less losses due to ground reflection characteristics, horizontal polarization allows the Yagi to outperform substantially the verti-

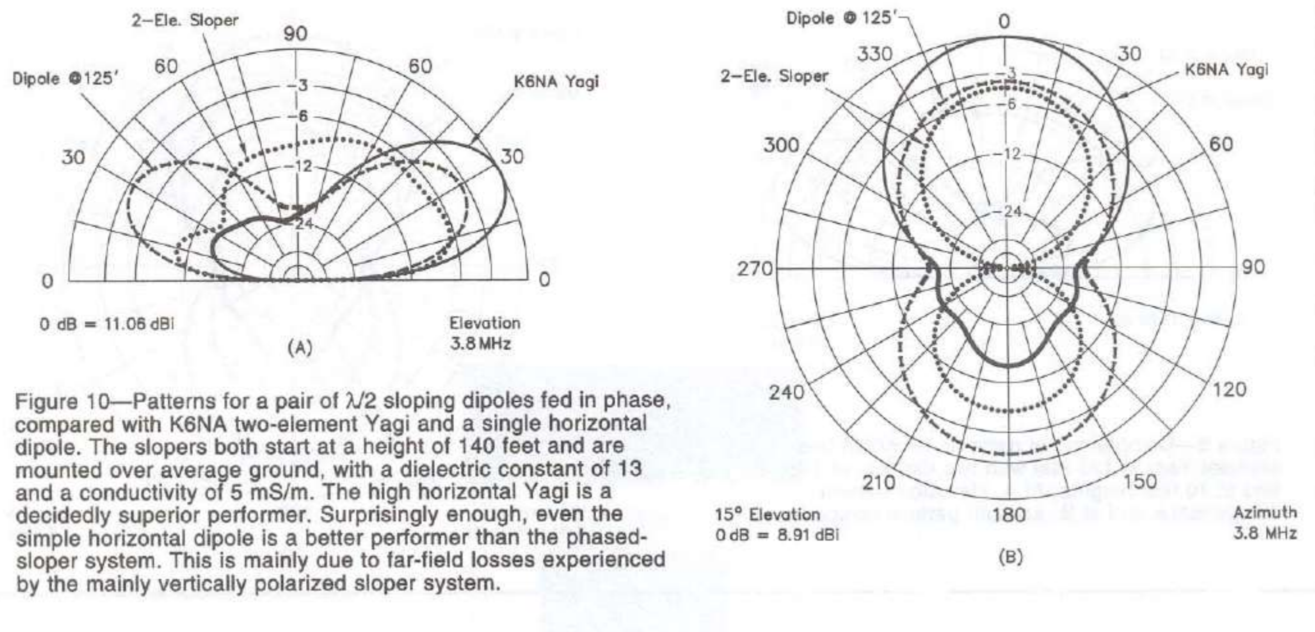

Figure 10—Patterns for a pair of λ/2 sloping dipoles fed in phase, compared with K6NA two-element Yagi and a single horizontal dipole. The slopers both start at a height of 140 feet and are mounted over average ground, with a dielectric constant of 13 and a conductivity of 5 mS/m. The high horizontal Yagi is a decidedly superior performer. Surprisingly enough, even the simple horizontal dipole is a better performer than the phased-sloper system. This is mainly due to far-field losses experienced by the mainly vertically polarized sloper system.

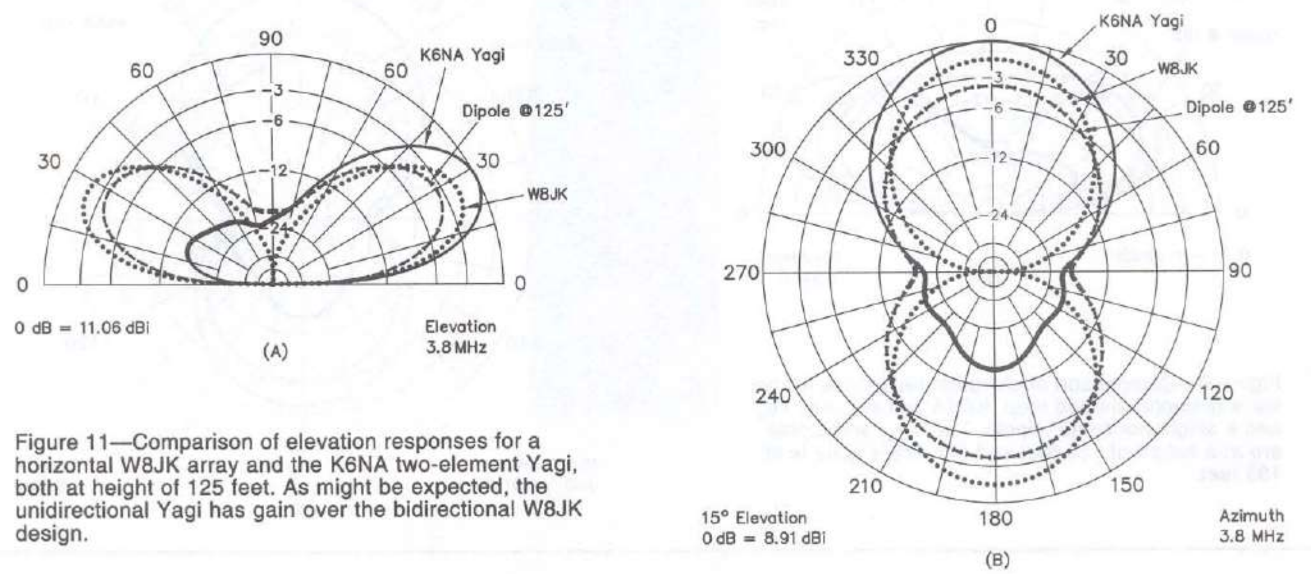

Figure 11—Comparison of elevation responses for a horizontal W8JK array and the K6NA two-element Yagi, both at height of 125 feet. As might be expected, the unidirectional Yagi has gain over the bidirectional W8JK design.

cal at all angles of interest, even down to 10°. In most locations, the vertical will be noisy on receive. *An extensive local ground screen (or elevated radials) will not make any significant change in the pattern relationship depicted here.* Only by installing the vertical in, say, a saltwater marsh extending perhaps 100 λ would its low-angle performance be improved substantially.^{6}

The "Four-Square" has become popular in recent years. Figure 13 also shows the beautiful pattern of this array, along with that of the two-element beam and a single vertical. In a quiet location this vertical system with a switching matrix makes a terrific receiving antenna, allowing full azimuth coverage in four directions. However, the horizontal wire Yagi still has superior gain at all elevation angles of interest—note the 4 dB advantage at 15°—and the Yagi likely will be quieter on receive in all but the most remote locations. The remarks in the paragraph above about attempts to improve the lowangle performance of the λ/4 vertical also apply to the Four-Square array.

Conclusions

Someone once said: "All antennas work... some more so." The sometimes-large differences seen during model comparisons can be misleading, as we know from our day-to-day operating. Propagation, pileup dynamics, and operator skills (one skill is picking a quiet location ...) are important factors that determine 80-meter DXing success. Virtually every antenna discussed in the comparison section of this report is a "good" antenna. The author's first 180 countries on 80 meters were worked with an inverted-V dipole at 70 feet.

If *tall* supports are available, a horizontally polarized, gain antenna will certainly provide a statistical increase in performance over the others. In lieu of really tall

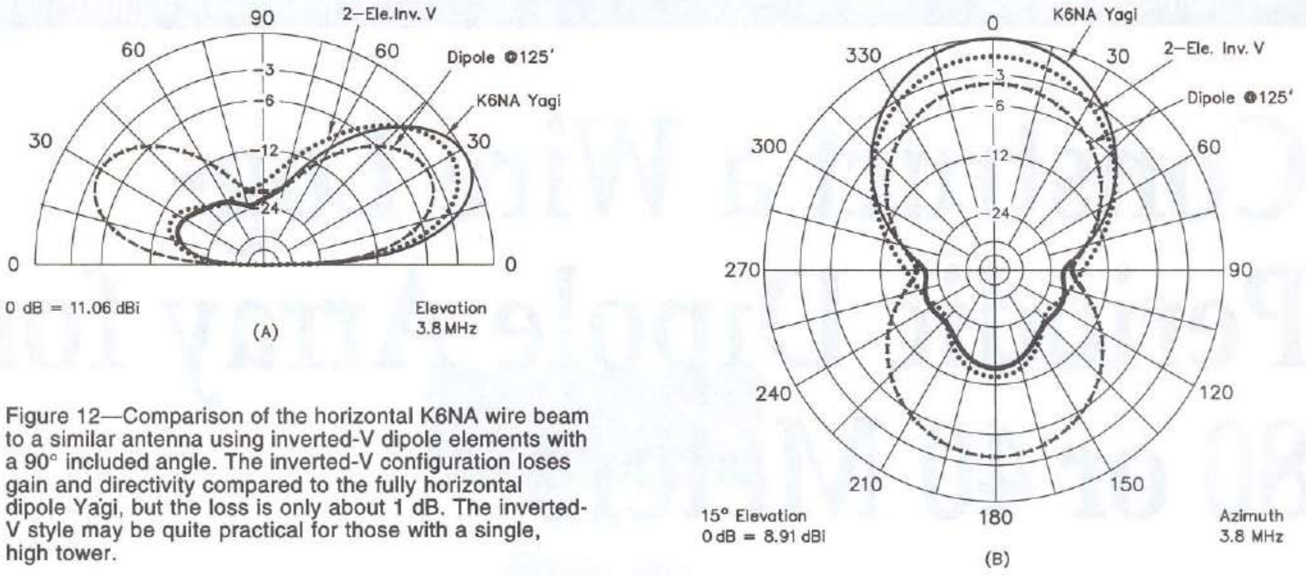

Figure 12—Comparison of the horizontal K6NA wire beam to a similar antenna using inverted-V dipole elements with a 90° included angle. The inverted-V configuration loses gain and directivity compared to the fully horizontal dipole Yagi, but the loss is only about 1 dB. The inverted-V style may be quite practical for those with a single, high tower.

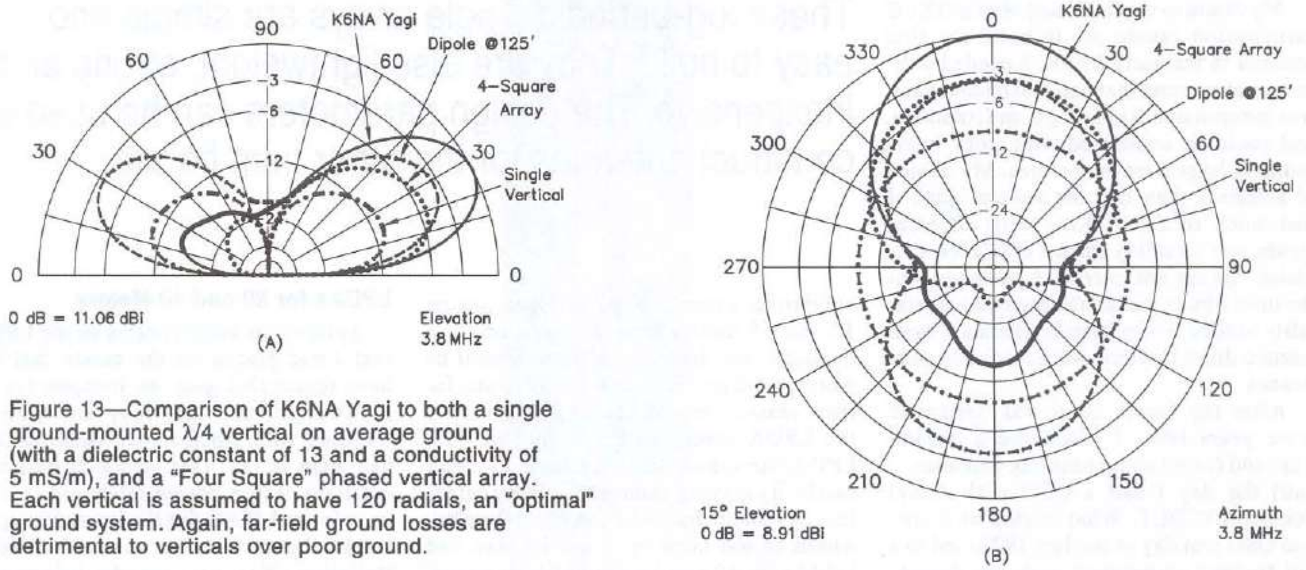

Figure 13—Comparison of K6NA Yagi to both a single ground-mounted λ/4 vertical on average ground (with a dielectric constant of 13 and a conductivity of 5 mS/m), and a "Four Square" phased vertical array. Each vertical is assumed to have 120 radials, an "optimal" ground system. Again, far-field ground losses are detrimental to verticals over poor ground.

towers, a well-installed vertical system can be an excellent antenna, especially in a quiet receiving location. As the computer models show, height above ground is the most significant variable when designing a lowband station using horizontally polarized antennas. [Hills help too!–*Ed.*]

It is hoped that some of the construction techniques for the crossarms, wire antennas, and feed lines described in this report will stimulate others to build even better horizontal arrays. You will enjoy special satisfaction using homemade open-wire lines, and a 40-year old Johnson Matchbox tuner makes a nice addition to your modern shack.

The author would like to thank H. Shepherd, W6US, for his valuable advice regarding wire antennas and open-wire feeder construction; R. Craig, N6ND, for mechanical suggestions; E. Andress, W6KUT, for manuscript review; and N6ND and J. McCook, W6YA, for modeling assistance.

Notes and References

[1] Wayne Overbeck, N6NB, "Quads vs. Yagis Revisited," *Ham Radio*, May 1979, p 12.

[2] James L. Lawson, W2PV, "160/80/75-Meter Broad-Band Inverted-V Antenna," *QST*, Nov 1970, p 17.

[3] James L. Lawson, W2PV, *Yagi Antenna Design*, W. Myers, K1GQ, editor (Newington: ARRL, 1986), p 2-2.

[4] Doug DeMaw, W1FB, editor, *ARRL Electronics Data Book* (Newington: ARRL, 1976), p 82.

[5] Frank Witt, AI1H, "Optimizing the 80-Meter Dipole," *The ARRL Antenna Compendium, Vol 4* (Newington: ARRL, 1995), p 39.

[6] Charles J. Michaels, W7XC, "Some Reflections on Vertical Antennas," *QST*, Jul 1987, p 18.

By John J. Uhl, KV5E **From *QST*, August 1986**

Construct a Wire Log-Periodic Dipole Array for 80 or 40 Meters

These log-periodic dipole arrays are simple and easy to build. They are also lightweight, strong and inexpensive. The design parameters can be used to construct antennas for the other ham bands.

My desire to work DX and obtain DXCC certification caused me to build my first antenna in the early 1960s. I needed a directional antenna that had reasonable gain, was inexpensive, lightweight and rotatable, and could be assembled with stock items found in large hardware stores. My choice of antennas then was the cubical quad. I had much success DXing with different quads, and I quickly earned DXCC certification. Quads are excellent antennas, but the ones I built lacked the mechanical stability needed in southern Louisiana. I soon learned this when they were ruined by hurricanes.

After my fourth quad was destroyed some years later, I purchased a triband Yagi and forgot about building antennas ... until the day I had a QSO with Ansyl Eckols, YV5DLT. What started as a normal QSO that day in the late 1970s led to a full-fledged experiment with the design, construction, erection and use of log-periodic dipole arrays made of wire. At that time, YV5DLT was using a triband log-periodic dipole array (LPDA) for 20, 15 and 10 meters. What immediately piqued my interest was that his beam was made of wire, and that his signal had outstanding quality and strength.

During the QSO, I asked Ansyl for construction details of his antenna. His response was generous. He mailed me diagrams, schematics and photographs of the LPDA that he had named Telerana. (He subsequently published his design in *QST*.[1]) After reading and studying all of his data, I was convinced that his design had the mechanical stability to withstand hurricanes, and I began plans in my mind to build a copy of Telerana.

I began a search of the literature, reading all of the LPDA articles that I could find.[2-8] By the time I gathered and read several references, three years had passed. Sunspot activity had diminished and band conditions weren't as good. Openings on 10 and 15 meters were few and of short duration, and future conditions would be worse. I did not duplicate the Telerana for these reasons, but decided instead to apply the LPDA theory to the design of wire LPDAs for use on the 160, 80 and 40-meter bands. By making some preliminary calculations I found that an LPDA for 160 meters would be too large to fit my lot size, but LPDAs for 80 and 40 would fit. However, it would not be possible for me to rotate these LPDAs.

LPDAs for 80 and 40 Meters

I placed the same criteria on the LPDAs that I had placed on the quads–that they have reasonable gain, be inexpensive and lightweight, and that they could be assembled with stock items found in large hardware stores. This article is written to detail the design, construction, erection and use of wire LPDAs for the lower frequency bands. Figure 1 shows one method of installation. You can use the information presented here as a guide and point of reference for building similar LPDAs.

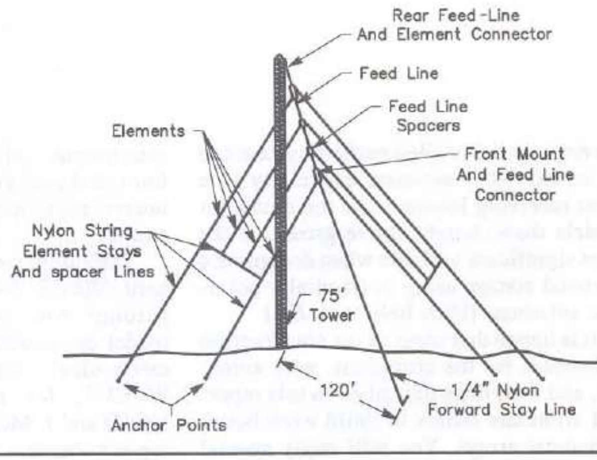
Rear Feed-Line And Element Connector
Feed Line
Feed Line Spacers
Elements
Front Mount And Feed Line Connector
Nylon String Element Stays And spacer Lines
75' Tower
120'
1/4" Nylon Forward Stay Line
Anchor Points

Figure 1—Typical 4-element log-periodic dipole array erected on a tower.

If space is available, the antennas can be "rotated" or repositioned in azimuth after they are completed. A 75-ft tower and a clear turning radius of 120 ft around the base of the tower are needed. The task is simplified if only three anchor points are used, instead of the five shown in Figure 1. Omit the two anchor points on the forward element, and extend the two nylon strings used for element stays all the way to the forward stay line.

For the design procedure, refer to *The ARRL Antenna Book*.[9] By using the formulas given there and other data in the text, all of the dimensions can be obtained and the LPDAs will take shape on paper. The design results are summarized in the concluding section of this article. The next step is to fabricate the fittings; see Figure 2 for details. Cut the wire elements and feed lines to the proper sizes and mark them for identification. After the wires are cut and placed aside, it will be difficult to remember which is which unless they are marked. When you have finished fabricating the connectors and cutting all of the wires, the antenna can be assembled. Use your ingenuity when building one of these antennas; it isn't necessary to duplicate my LPDAs exactly.

The elements are made of standard no. 14 stranded copper wire. The two parallel feed lines are made of no. 12 solid copper-coated steel wire, such as Copperweld. This will not stretch when placed under tension. The front and rear connectors are cut from 1/2-in-thick Lexan sheeting, and the feed-line spacers from 1/4-in Plexiglas sheeting.

Study the plans carefully and be familiar with the way the wire elements are connected to the two feed lines, through the front, rear and spacer connectors. Details are sketched in Figure 4. Connections made this way prevent the wire from breaking. All of the rope, string and connectors must be made of materials that can withstand the effects of tension and weathering. Use nylon rope and strings, the type that yachtsmen use. Figure 1 shows the front stay rope coming down to ground level at a point 120 ft from the base of a 75-ft tower. It may not be possible to do this in all cases. In my installation I put a pulley 40 ft up in a tree and ran the front stay rope through the pulley and down to ground level at the base of the tree. The front stay rope will have to be tightened with a block and tackle at ground level.

Putting an LPDA together is not difficult if it is assembled in an orderly manner. It is easier to connect the elements to the feeder lines when the feed-line assembly is stretched between two points. Use the tower and a block and tackle. Attaching the rear connector to the tower and assembling the LPDA at the base of the tower makes raising the antenna into place a much simpler task. Tie the rear connector securely to the base of the tower and attach the two feeder lines to it. Then thread the two feed-line spacers onto the feed line. The spacers will be loose at this time, but will be positioned properly when the elements are connected. Now connect the front connector to the feed lines. A word of caution: Measure accurately and carefully! Double-check all measurements before you make permanent connections.

Connect the elements to the feeder lines through their respective plastic connectors, beginning with element 1, then element 2, and so on. Keep all of the element wires securely coiled. If they unravel, you will have a tangled mess of kinked wire. Check that the element-to-feeder connections have been made properly. (See Figure 4.) Once you have completed all of the element connections, attach the 4:1 balun to the underside of the front connector. Connect the feeder lines and the coaxial cable to the balun.

You will need a separate piece of rope and a pulley to raise the completed LPDA into position. First secure the eight element ends with nylon string, referring to Figures 1 and 3. The string must be long enough to reach the tie-down points. Connect the front stay rope to the front connector, and the completed LPDA is now ready to be raised into position. While raising the antenna, uncoil the element wires to prevent their getting away and balling up into a mess. Use care! Raise the rear connector to the proper height and attach it securely to the tower, then pull the front stay rope tight and secure it. Move the elements so that they form a 60-degree angle with the feed lines, in the direction of the front, and space them properly relative to one another. By adjusting the end positions of the elements as you walk back and forth, you will be able to align all the elements properly. Now it is time to hook your rig to the system and make some QSOs.

Performance

The reports I received using the LPDAs were compared with an inverted-V dipole. All of the antennas are fixed; the LPDAs radiate to the northeast and the dipole to the northeast and southwest. The apex of the

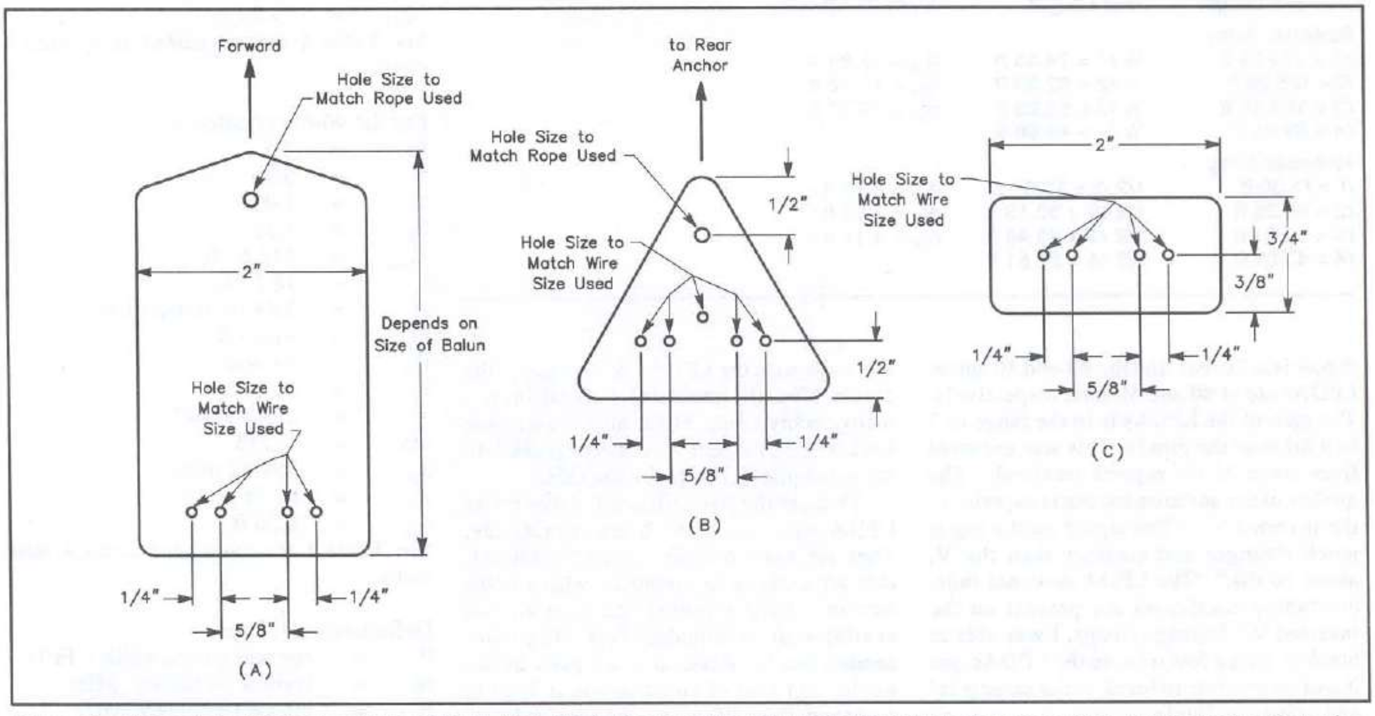
Forward
Hole Size to Match Rope Used
2"
Depends on Size of Balun
Hole Size to Match Wire Size Used
1/4"
5/8"
1/4"
(A)
to Rear Anchor
Hole Size to Match Rope Used
1/2"
Hole Size to Match Wire Size Used
1/2"
1/4"
5/8"
1/4"
(B)
Hole Size to Match Wire Size Used
2"
3/4"
3/8"
1/4"
5/8"
1/4"
(C)

Figure 2—Pieces to be fabricated for the LPDA. At A, the forward connector, made from 1/2-inch Lexan. At B, the rear connector, also made from 1/2-inch Lexan. At C is the pattern for the feed-line spacers, made from 1/4-inch Plexiglas. Two of these spacers are required.

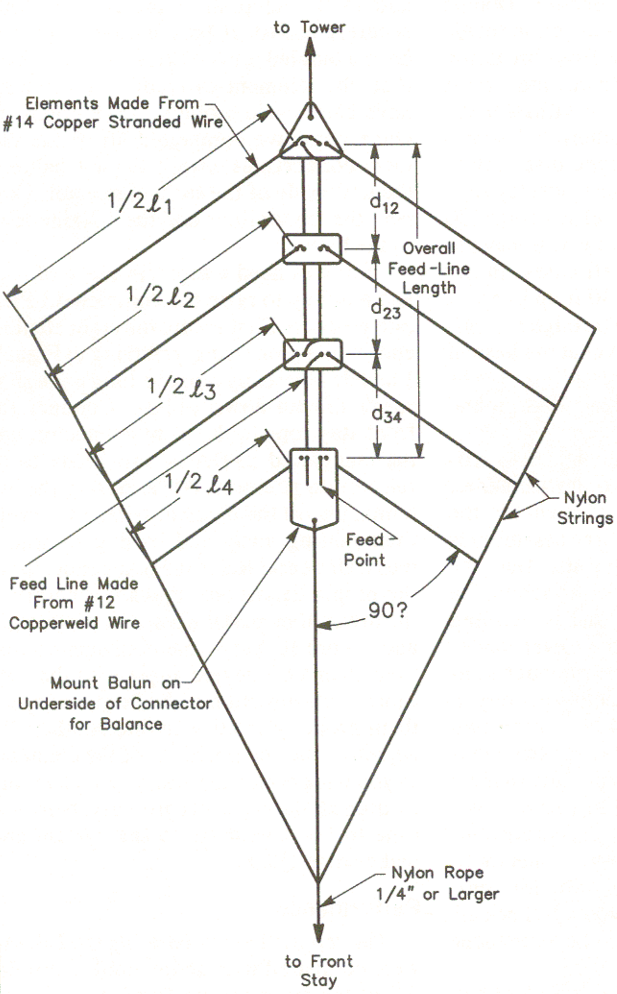
to Tower
Elements Made From #14 Copper Stranded Wire
1/2ℓ1
1/2ℓ2
1/2ℓ3
1/2ℓ4
d12
d23
d34
Overall Feed-Line Length
Nylon Strings
Feed Point
90?
Feed Line Made From #12 Copperweld Wire
Mount Balun on Underside of Connector for Balance
Nylon Rope 1/4" or Larger
to Front Stay

Figure 3—Typical layout for the LPDA. Use a 4:1 balun at the point indicated. See Table 1 for dimensions.

Table 1
Calculated Array Dimensions

Element Length	*Half Length*	*Element Spacing*
80-Meter Array		
ℓ1 = 149.09 ft	½ ℓ1 = 74.55 ft	d_{12} = 17.89 ft
ℓ2= 125.98 ft	½ ℓ2 = 62.99 ft	d_{23} = 15.12 ft
ℓ3 = 106.45 ft	½ ℓ3 = 53.23 ft	d_{34} = 12.77 ft
ℓ4 = 89.95 ft	½ ℓ4 = 44.98 ft	
40-Meter Array		
ℓl = 71.30 ft	1/2 ℓl = 35.65 ft	d_{12} = 8.56 ft
ℓ2 = 60.25 ft	1/2 ℓ2 = 30.13 ft	d_{23} = 7.23 ft
ℓ3 = 50.91 ft	1/2 ℓ3 = 25.46 ft	d_{34} = 6.11 ft
ℓ4 = 43.02 ft	1/2 ℓ4 = 21.51 ft	

dipole is at 70 feet, and the 40-and 80-meter LPDAs are at 60 and 50 feet, respectively. The gain of the LPDAs is in the range of 7 to 9 dB over the dipole. This was apparent from some of the reports received: "The quality of the audio on the log is superior to the inverted V." "The signal on the log is much stronger and steadier than the V, about 10 dB." "The LPDA does not fade, but fading conditions are present on the inverted V." During pileups, I was able to break in with a few tries on the LPDAs, yet it was impossible to break in the same pileups using the dipole.

During the CQ WW DX Contest I was able to break into some *big* pileups after a few calls with the LPDAs. Switching to the dipole, I found it impossible to break in after many, many calls. Then, after I switched back to the LPDA, it was easy to break into the same pileup and make the QSO.

Think of the possibilities that these wire LPDA systems offer hams worldwide. They are easy to design and to construct, real advantages in countries where commercially built antennas and parts are not available at reasonable cost. The wire needed can be obtained in all parts of the world, and cost of construction is low! If damaged, the LPDAs can be repaired easily with pliers and solder. For those who travel on DXpeditions where space and weight are large considerations, LPDAs are lightweight but sturdy, and they perform well. They'll even withstand a hurricane!

Calculations for Log-Periodic Dipole Arrays

Design constants and the results of design procedures follow. (Terms are defined at the end of this section.)

τ = 0.845
σ = 0.06
σ_{opt} = 0.152
cot α = 1.548
α = 32.86°
Gain = 7.5 dBi (5.35 dBd). [By sloping the elements forward, the gain may be increased 3 to 5 dB over this figure. *–Ed.*]
σ' = 0.065
B_{ar} = 1.39; see note 10.
R_0 = 70

Fed with 50-ohm coaxial cable and a 4:1 balun

For the 80-meter antenna,
fn = 4.1
f1 = 3.3
B = 1.24
B_S = 1.72
λ_{max} = 298.18 ft
L = 48.42 ft
N = 4.23 (rounded to 4)
ℓ1 = 149.09 ft
R_0 = 70 ohms
h = 62.4
a = 2.667×10^{-3}
h/a = 23400
Z_{av} = 937.26 ohms
Z_0 = 80.72 ohms
d_{12} = 17.89 ft
See Table 1 for calculated array dimensions.

For the 40-meter antenna,
fn = 7.5
fl = 6.9
B = 1.09
B_s = 1.51
λ_{max} = 142.61 ft
L = 18.57 ft
N = 3.44 (rounded to 4)
ℓ1 = 71.30 ft
R_0 = 70 ohms
h = 32.727
a = 2.667×10^{-3}
h/a = 12273
Z_{av} = 859.82 ohms
Z_0 = 81.76 ohms
d_{12} = 8.56 ft
See Table 1 for calculated array dimensions.

Definitions of Terms
B = operating bandwidth = fn/f1
fn = highest frequency, MHz
fl = lowest frequency, MHz
τ = design constant
σ = relative spacing constant

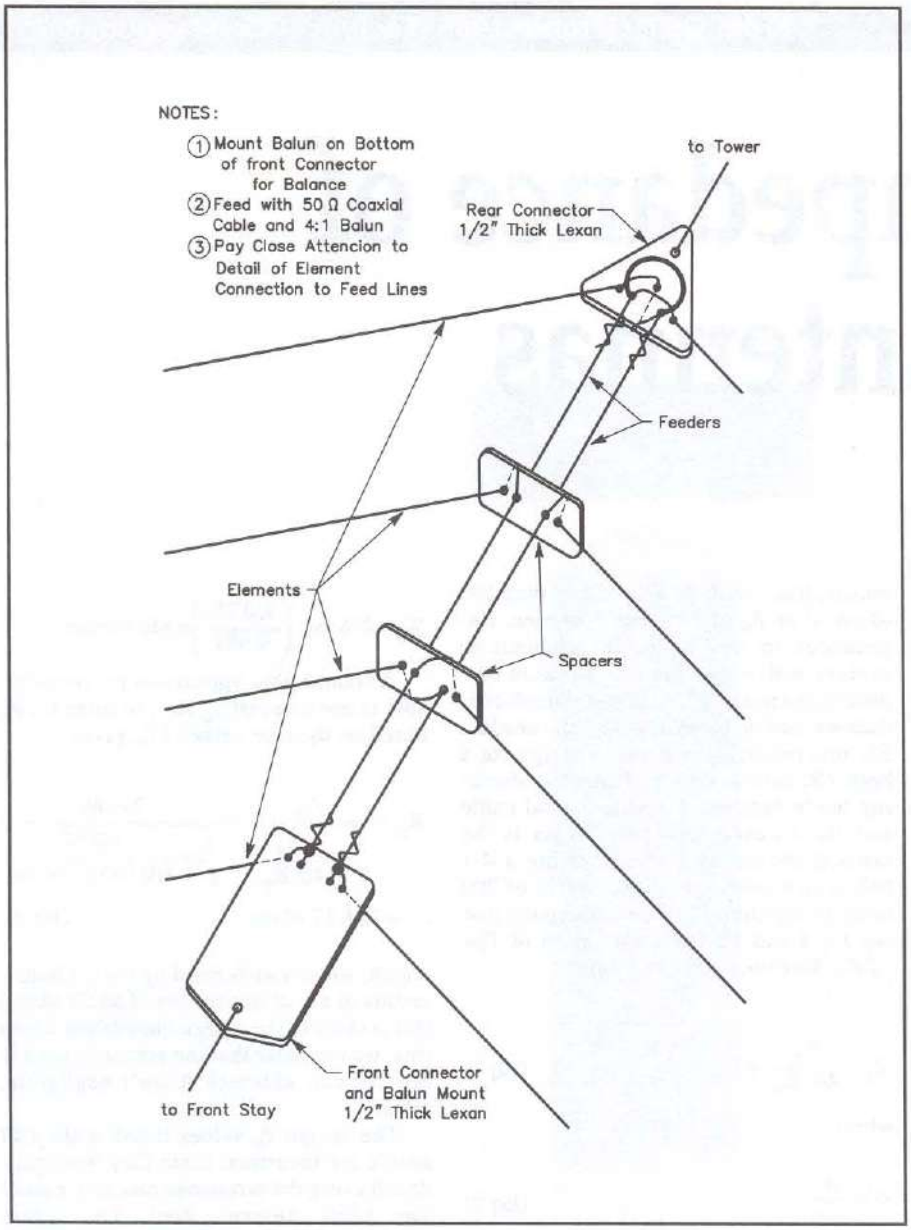
NOTES:
① Mount Balun on Bottom of front Connector for Balance
② Feed with 50 Ω Coaxial Cable and 4:1 Balun
③ Pay Close Attencion to Detail of Element Connection to Feed Lines
to Tower
Rear Connector 1/2" Thick Lexan
Feeders
Elements
Spacers
Front Connector and Balun Mount 1/2" Thick Lexan
to Front Stay

Figure 4—Details of electrical and mechanical connections of the elements to the feed line. Knots in the nylon stay lines are not shown.

σ_{opt} = value of σ for optimum gain
σ' = mean spacing factor
α = apex half-angle
B_{ar} = bandwidth of the active group. See note 10.
B_s = structure (array) bandwidth
L = boom length for N elements
N = number of elements
$\ell 1$ = longest element = 492/f1
λ_{max} = longest free-space wavelength = 984/f1
Z_0 = characteristic impedance of feeder
R_0 = mean radiation resistance level of required input impedance of active region
Z_{av} = average characteristic impedance of a dipole

$$= 120\left(\ln\frac{h}{a}-2.25\right)$$

h = element half length
a = radius of element
ℓ = length of elements
d = spacing between elements

Notes

[1] A. Eckols, "The Telerana—A Broadband 13- to 30-MHz Directional Antenna," *QST*, Jul 1981, pp 24-27.

[2] C. T. Milner, "Log Periodic Antennas," *QST*, Nov 1959, p 11.

[3] D. E. Isbell, "Log-Periodic Dipole Arrays," *IRE Transactions on Antennas and Propagation*, Vol AP-8, No. 3, May 1960, pp 260-267.

[4] P. D. Rhodes, "The Log-Periodic Dipole Array," *QST*, Nov 1973, pp. 19-22.

[5] *The GIANT Book of Amateur Radio Antennas* (Blue Ridge Summit, PA: Tab Books, 1979), pp 55-85.

[6] C. A. Balanis, *Antenna Theory, Analysis and Design* (New York: Harper and Row, 1982), pp 427-439.

[7] G. L. Hall, ed., *The ARRL Antenna Book* (Newington, CT: The American Radio Relay League, Inc, 1982), pp 6-24 to 6-26, 9-12 to 9-14.

[8] D. A. Mack, "A Second-Generation Spiderweb Antenna," *The ARRL Antenna Compendium Vol 1* (Newington, CT: The American Radio Relay League, Inc, 1985), pp 55-59.

[9] See note 7.

[10] See Ref 6, p 435. B_{ar} is found from the following equation:

$$B_{ar} = 1.1 + 7.7(1-\tau)^2 \cot \alpha$$

From *QST*, October 1986 (Technical Correspondence)

Input Impedance of LPDA Antennas

The log-periodic dipole array (LPDA) has intriguing possibilities, some of which have been explored by Uhl.[1] Unfortunately, some misleading information in *The ARRL Antenna Book* has kept the subject from being more accessible to the amateur.[2] The design procedure called out by *The Antenna Book* and by Rhodes refers to graphs of R_0, the mean antenna input resistance, versus the log-periodic design parameters τ and α.[3] It would appear that R_0 is dependent solely on these parameters; such is not the case. As shown by Eq 2 and 4, R_0 is dependent on τ, σ, Z_{av} and Z_0. This relationship is demonstrated by Carrel.[4] The graphs in question are from the seminal experimental study of the LPDA by Isbell, and are used to demonstrate the effect on R_0 of changing τ and σ; Z_{av} and Z_0 were constant.[5] Since τ and σ are usually chosen to provide the desired antenna geometry and directivity (gain), R_0 should be controlled by either Z_{av} or Z_0. Z_{av} is usually constrained by mechanical considerations, which leaves Z_0 as the easiest parameter to adjust for desired R_0.

The erroneous procedure determines R_0 from a graph. The correct procedure is to calculate the necessary Z_0 based on the *desired* R_0. For most practical LPDA antennas, this results in a Z_0 of less than 100 ohms, if an R_0 of 50 ohms is chosen. Impedances in this range are difficult to achieve with open-wire line, the kind normally used in an LPDA, because thick conductors and/or close spacing are needed. For this reason, it is usual to design for a higher R_0 and use an impedance-transforming balun between 50-ohm coaxial cable and the antenna feed point. This is the method chosen by Uhl. Assuming a 4:1 balun, an antenna-input impedance of 200 ohms is required. Z_0 is then calculated (using Eq 9 and 11 from page 6-25 of *The ARRL Antenna Book*) as follows:

$$Z_0 = \frac{R_0^{\,2}}{8\sigma' Z_{av}} + R_0\sqrt{\left(\frac{R_0}{8\sigma' Z_{av}}\right)^2 + 1} \quad \text{(Eq 1)}$$

where

$$\sigma' = \frac{\sigma}{\sqrt{\tau}} \quad \text{(Eq 2)}$$

$$Z_0 = \frac{200^2}{8(0.065)\,(937.26)} + 200\sqrt{\left(\frac{200}{8(0.065)\,(937.26)}\right)^2 + 1}$$

$$= 298.26 \text{ ohms}$$

The actual impedance of the feed line used by Uhl can be computed from the formula on p 3-16 of *The ARRL Antenna Book*:

$$Z_0 = 276 \log\left(\frac{2S}{d}\right) \quad \text{(Eq 3)}$$

where S is the spacing between conductors, 0.75 inch, and d is the conductor diameter, 0.081 inch for no. 12 AWG wire. Therefore:

$$Z_0 = 276 \log\left(\frac{2(0.75)}{0.081}\right) = 349.86 \text{ ohms}$$

Although this appears to be considerably in error, recasting Eq 1 to solve for R_0 based on the true value of Z_0 gives:

$$R_0 = \frac{Z_0}{\sqrt{1 + \dfrac{Z_0}{4\sigma' Z_{av}}}} = \frac{349.86}{\sqrt{1 + \dfrac{349.86}{4(0.065)(937.26)}}}$$

$$= 224.17 \text{ ohms} \quad \text{(Eq 4)}$$

which, when transformed by the 4:1 balun, results in a load impedance of 56.04 ohms; this is close to the design impedance. From this, we can infer that the actual Z_0 used is not critical, although it isn't negligible, either.

The design Z_0 values listed in the Uhl article are incorrect, since they were produced using the erroneous procedure from *The ARRL Antenna Book*. The values should be 298.26 and 308.56 ohms for the 80- and 40-meter antennas, respectively. In addition, R_0 could be made closer to the design value of 200 ohms by using a different feeder spacing to change the feeder impedance. From Eq 2:

$$S = \left(\frac{d}{2}\right)10^{(Z_0/276)} = \left(\frac{0.081}{2}\right)10^{(298.26/276)}$$

$$= 0.488 \text{ inch}$$

Rounding this value to 0.5 inch results in a Z_0 of 301.26 ohms and an R_0 of 201.46 ohms for 80 meters and 196.62 ohms for 40 meters. These translate to load impedance of 50.37 and 49.15 ohms, respectively.—*Jon Bloom, KE3Z*

---

[1] J.J. Uhl, "Construct a Wire Log-Periodic Dipole Array for 80 or 40 Meters," *QST*, Aug 1986, p 21.

[2] G.L. Hall, ed., *The ARRL Antenna Book* (Newington: ARRL, 1982), pp 6-24 to 6-26.

[3] P.D. Rhodes, "The Log-Periodic Dipole Array," *QST*, Nov 1973, pp 16-22.

[4] R.L. Carrel, "The Design of Log-Periodic Dipole Antennas," *1961 IRE International Convention Record, Part 1, Antennas and Propagation*, pp 61-75; also PhD thesis, University of Illinois, Urbana, 1961.

[5] D. E. Isbell, "Log-Periodic Dipole Arrays," *IRE Transactions on Antennas and Propagation*, Vol AP-8, No. 3, May 1960, pp 260-267.

From *QST*, August 1980 (Technical Correspondence)

The WA1AKR 40- and 75-Meter Slopers

Several amateurs have suggested that I submit a description of my sloper antenna system for publication in "Hints and Kinks." Other amateurs may be interested in this adaptation of the 8JK beam. Construction information is shown for both the 75- and 40-meter bands.

As shown in the accompanying diagram, the array has two half-wave sloping elements joined by a 1/8-wave, 300-ohm phasing line. Transposing the phasing line should bring the element currents into phase. I find the antenna is broadbanded. There appears to be no need for a Transmatch.

If one desires to suspend an additional sloper from the tower for a directional change, installation of remote switching at the top of the tower will permit the use of a single transmission line. Otherwise, separate transmission lines will be required.

Ends of the antenna are suspended by ropes with the tops placed roughly 1 foot away from the tower. An angle of 45 degrees between the antenna and ground should be maintained. Do not use an angle greater than 50 degrees. Resonance with the dimensions shown should occur near 3.8 MHz for the 75-meter sloper and 7.150 for the 40-meter antenna.

How well do my antennas work? I have contacted stations "across the pond" while competing with the big boys who sport three and four-element beams. I have also experienced little difficulty in working VKs and ZLs. – *Carl Bissonnette, WA1AKR*

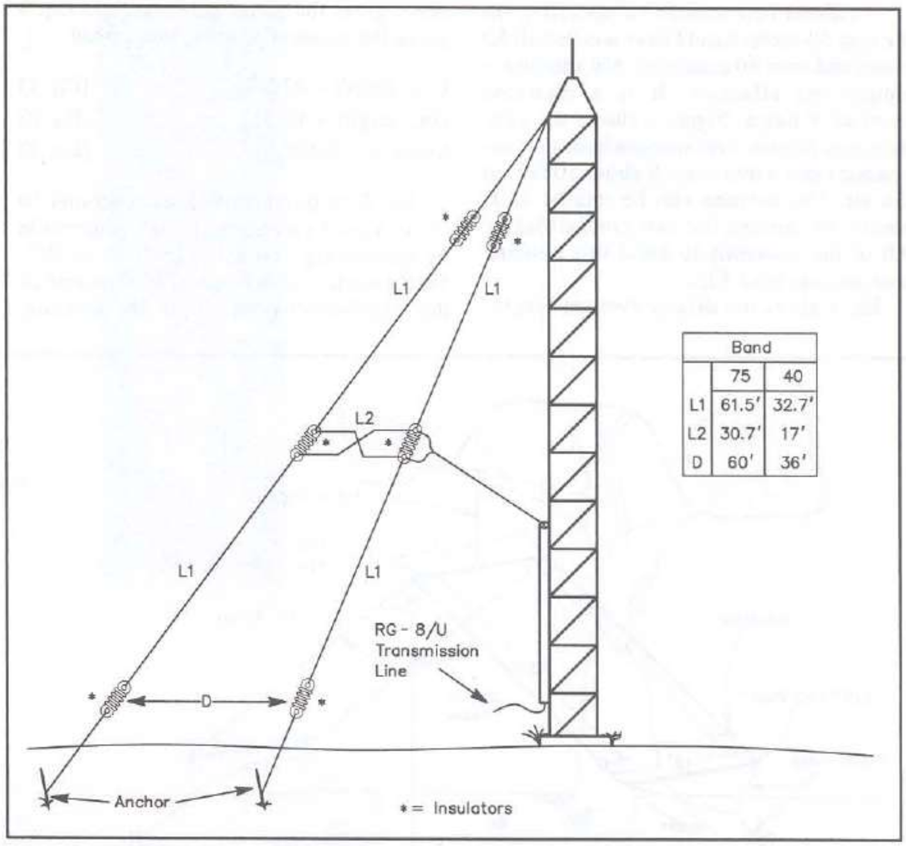

	Band	
	75	40
L1	61.5'	32.7'
L2	30.7'	17'
D	60'	36'

When Carl Bissonnette, WA1AKR, chases DX he uses a sloper like the one illustrated. Carl's arrangement is fashioned after the famous 8JK beam. The feed system resembles that of the ZL Special.

From *QST*, June 1983 (Hints & Kinks)

Inexpensive 30-Meter Beam Antenna

In about two months of operating on the new 30-meter band I have worked all 50 states and over 50 countries. My antenna is simple but effective. It is a rotatable inverted V beam. Figure 1 shows the construction details. The antenna boom is suspended from a tree branch about 50 feet in the air. The antenna can be rotated 360° simply by moving the two ground stakes. All of the materials to build this antenna cost me less than $25.

Eq. 1 gives the driven element length, Eq. 2 gives the director length, and Eq. 3 gives the element spacing that I used.

D.E. length = $476/f_{MHz}$ (Eq. 1)

Dir. length = $450/f_{MHz}$ (Eq. 2)

Spacing = $120/f_{MHz}$ (Eq. 3)

The feed-point impedance is around 30 ohms. I used a matching transformer made by connecting two 1/4-λ sections of RG-59/U coaxial cable in parallel. One end of the transformer connects to the antenna, and the other end goes to 50-ohm cable to the shack. Figure 1B shows how this is wired. Perhaps the easiest method to join the two pieces of 75-ohm cable is to use coaxial T connectors. You should use a balun at the antenna feed point to prevent rf from flowing on the outside of the shield braid.

I also built an antenna of this type for 40 meters, and it works great. I guess the key word is rotatable! —*Jon Ferrara, N9DWR, Chattanooga, Tennessee*

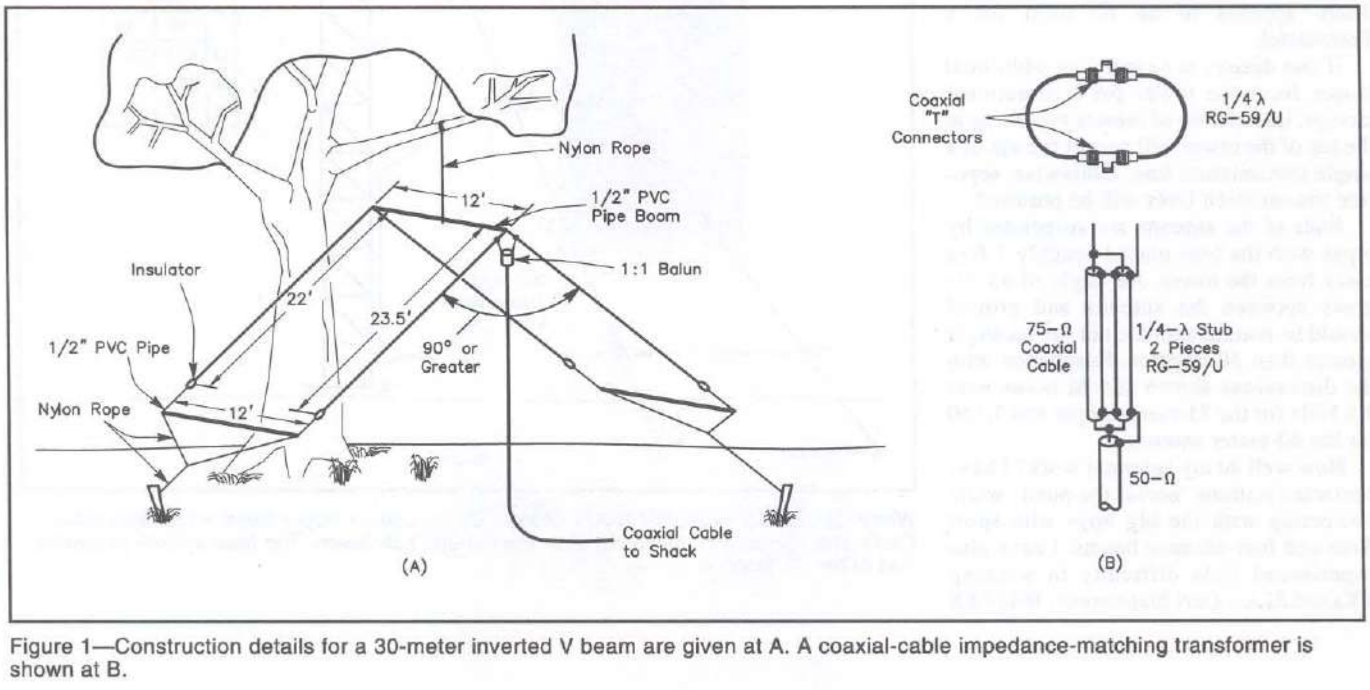
Nylon Rope
12'
1/2" PVC Pipe Boom
1:1 Balun
Insulator
22'
23.5'
1/2" PVC Pipe
90° or Greater
Nylon Rope
12'
Coaxial Cable to Shack
(A)
Coaxial "T" Connectors
1/4 λ RG-59/U
75-Ω Coaxial Cable
1/4-λ Stub 2 Pieces RG-59/U
50-Ω
(B)

Figure 1—Construction details for a 30-meter inverted V beam are given at A. A coaxial-cable impedance-matching transformer is shown at B.

By Roy W. Lewallen, W7EL

From *QST*, June 1984

Try the "FD Special" Antenna

Looking for an antenna that's simple, inexpensive, lightweight and easy to install? Here's one that fits the description.

While the Field Day "juices" were flowing last year, my Field Day partners and I decided to become more competitive without compromising our general philosophy: Use all homemade gear, pack it in to the site, and don't let operating interfere with watching the scenery. Since most of our gear already ran the accepted QRP power limit of 10-W dc input and had been designed to provide high output efficiency, the antenna seemed like the reasonable point of attack (our high quality superheterodyne receivers have not been the limiting item). The constraints dictated that an antenna be lightweight, portable and easy to put up. It should also have substantial gain.

Because of our West Coast location, the front-to-back ratio was not a concern. But having a reasonably wide lobe toward the East was. We desired a low SWR because of the relatively high loss of our RG-58/U and/or RG-174/U feed line. We were interested only in the CW portion of the band, but this antenna works well over all of the band. It seemed that 20 meters would be our main "money-maker," so we designed the antenna for that band. It can be scaled for other bands, too.

The Research

Although a number of antenna types might have done the job, I settled quickly on a horizontal, close-spaced, driven array. Experience has shown that driven arrays are generally more tolerant of imperfect construction and erection than are parasitic arrays. Experience and much measurement have convinced me that horizontal arrays outperform vertical ones in the high-frequency bands, except perhaps from an exceptional location. In addition, we didn't want the nuisance of establishing a decent ground system — which most vertical arrays require.

The theoretical gain and front-to-back ratio of 2-element arrays with 1/8-wavelength spacing between the elements are shown in Figures 1 and 2. Note the lower curve of Figure 1. It shows the effect of losses on the gain (losses don't affect the front-to-back ratio, and change only the scaling of the pattern). Figure 3 shows the patterns of arrays with 135, 160 and 180-degree relative spacing. All are drawn to the same scale. The 1/8-wavelength spaced, 135-degree-fed array is frequently called the "ZL Special."[1] The close-spaced, 180-degree-fed array is known as an "8JK."[2] From 135 to 160 degrees, phasing was chosen because of the combination of relatively high insensitivity to loss, reasonable gain and wide forward lobes. Note that the gain stays about the same in this range, ensuring good performance if the phasing isn't exactly as predicted. Actually, it's much easier to generate and maintain precise 180-degree phasing than the angles I've chosen — particularly over a wide frequency range.

There's one major flaw (usually fatal) in a simple analysis like the one presented here: It assumes that equal-magnitude currents are flowing in the elements. This is not easy to realize, for even in arrays with elements spaced 1/2 wavelength or greater, mutual coupling has a profound effect on element impedances. This changes them dramatically and unequally, as a rule. This

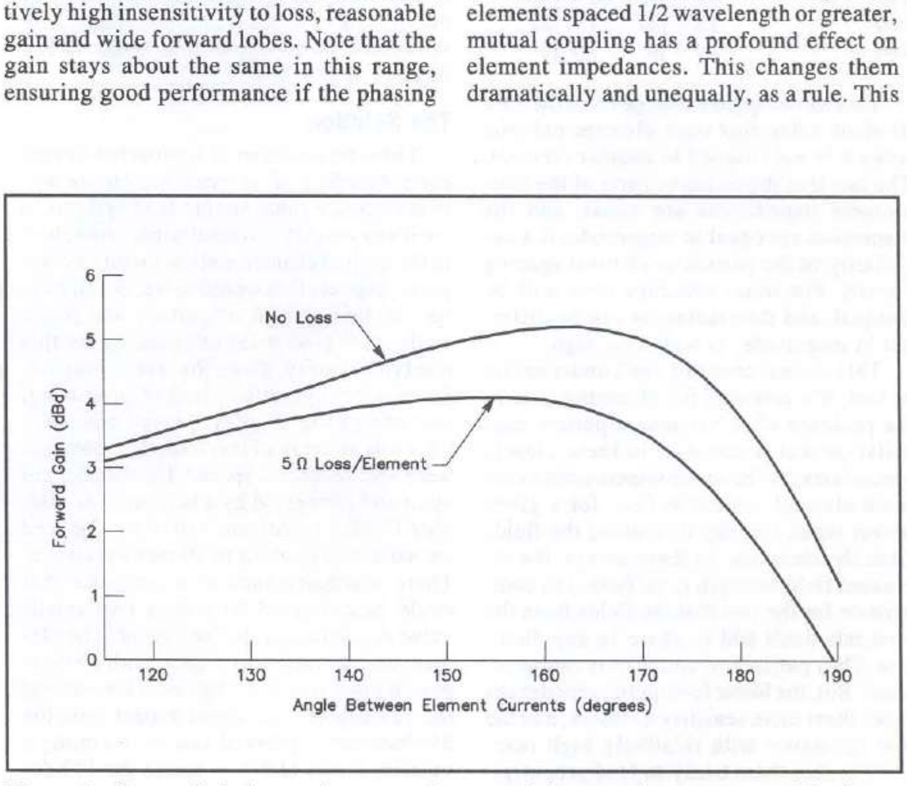

Figure 1—Curves that show gain versus phase angle for two-element arrays with 1/8-wavelength spacing.

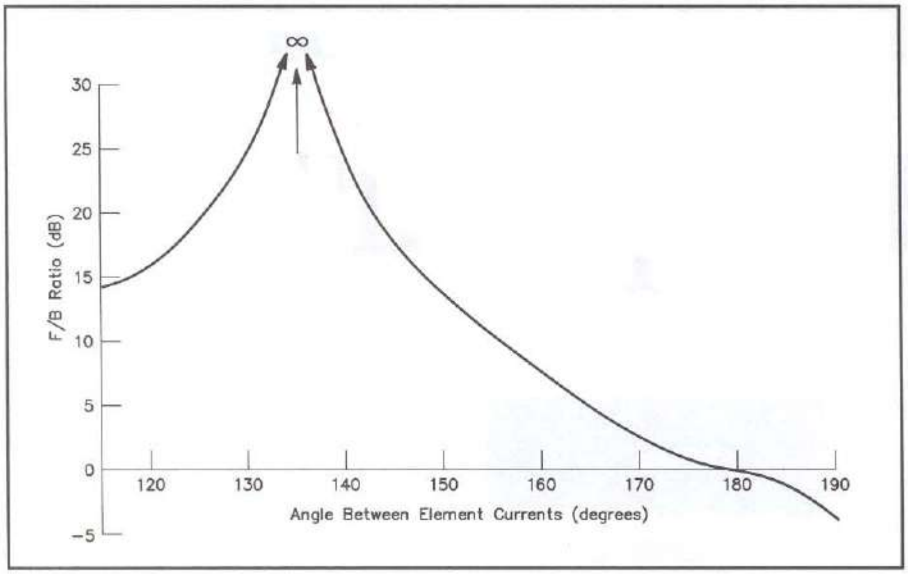

Figure 2—Theoretical F/B ratio for a two-element array with 1/8-wavelength spacing.

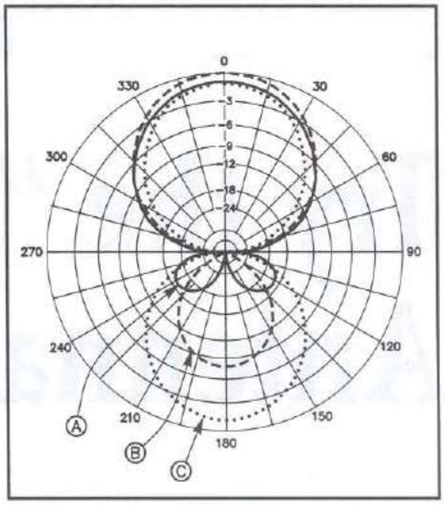

Figure 3—Dipole array patterns for 135, 160 and 180-degree relative phasing at λ/8 spacing. Curves A, B and C, respectively, represent these conditions. Add 5 dB for dBd. These curves are based on the array being fed with equal currents.

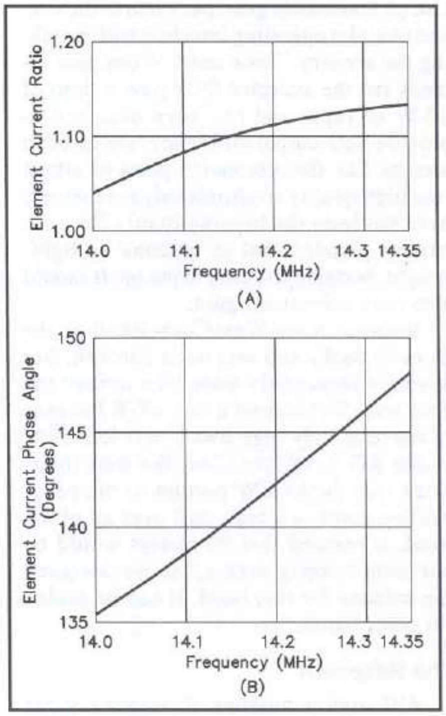

Figure 4—Element current ratio (A) and phase angle (B) as a function of frequency.

impedance change is a function of not only "mutual impedance," but also the relative magnitudes and phases of the currents flowing in the elements. In an array as closely spaced as these, coupling is so intimate that it could be argued that the term "driven array" is a misnomer. For example, the feed-point impedances of the elements in a 1/8-wavelength-spaced array, assuming equal currents can be made to flow, are

Phase Angle Between Currents (Degrees)	*Loading Element (Ohms)*	*Lagging Element (Ohms)*
135	28 – j46	28 + j46
160	13 – j22	13 + j22
180	9 + j0	9 + j0

This shows quite a change from the 74 + j0 ohms value that each element exhibits when it is not coupled to another element. The fact that the resistive parts of the two-element impedances are equal, and the reactances are equal in magnitude, is a peculiarity of the particular element spacing chosen. For other spacings they will be unequal, and the reactances can be different in magnitude, as well as in sign.

This mutual coupling isn't undesirable; in fact, it's essential for obtaining gain in the presence of rather severe pattern cancellation that is common in these closely spaced arrays. The lower impedances cause more element current to flow for a given power input, thereby increasing the fields from the elements. In these arrays, the increased field strength is sufficient to compensate for the fact that the fields from the elements don't add in phase in any direction. They partially or completely cancel instead. But, the lower feed-point impedances make them more sensitive to losses, and the low resistance with relatively high reactance makes them tricky to feed properly.

Why do these different and reactive feed impedances make feeding the arrays so difficult? The first problem is that, with few exceptions, the magnitude of current out of a line not terminated in its characteristic impedance won't be equal to the current into the line. In classic "If you can't fix it, feature it!" fashion, this impedance-transforming property is put to good use in the form of the 1/4-wavelength Q section.[3] The second (and almost always overlooked) difficulty is that, again with only a few exceptions, the phase delay of current in an imperfectly terminated transmission line doesn't equal the electrical length of the line. This effect isn't minor: The phasing of a casually designed array can easily be off by tens of degrees. In one design I investigated, an 80-degree line produced 139 degrees of phase shift.

The Solution

There are a number of approaches toward correct feeding of an array. My choice was to investigate some simple feed systems to see if any would yield results that came close to the desired characteristics. I wrote a computer program that would solve, iteratively, for element-current magnitude and phase angle, plus feed-point impedances for this one type of array, given the array specifics. Several configurations looked promising, and one of the simplest proved adequate. This was an array of two folded dipoles that were self-resonant, spaced 1/8 wavelength apart and connected by a taut piece of 300-ohm TV ribbon with one half twist. The feed impedance was close to 50 ohms resistive. There was some inductive reactance that could be corrected by adding two small-value capacitors at the feed point. The element current ratio was 1.13:1, with element phasing that was 154 degrees. This was not the 124 degrees one might expect from the 56 electrical degrees of line — assuming a velocity factor of 0.8 — minus the 180 degrees caused by the half twist.

It was this array that we used for Field Day, with very good results (see section on performance). However, when the array was reconstructed at the home QTH, a dramatic rise in SWR was noticed when operating the antenna at other than the low end of the band. Computer analysis showed that, above the design frequency, the phase angle increased. This caused a substantial lowering of the element feedpoint impedance, plus narrowing of the forward lobe. The analysis also showed the antenna to be well-behaved *below* the design frequency. Consequently, a similar array was designed (figuratively speaking) for 14.5 MHz. It gave good results over the 20-meter band.

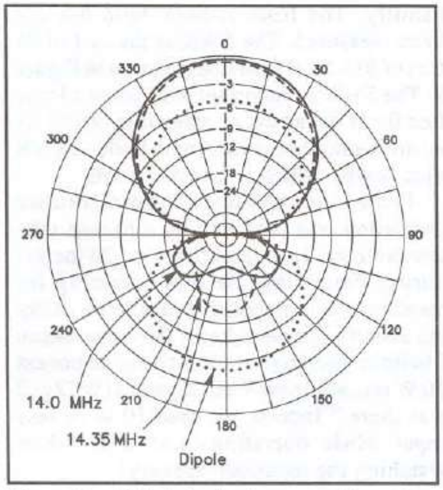

Figure 5—Calculated antenna patterns for the high and low ends of the 20-meter band. Add 5 dB for dBd.

Element phasing varies from 135 degrees at 14.0 MHz to 148 degrees at 14.35 MHz, with current ratios from 1.04 to 1.13:1 (see Figure 4). The gain can be calculated as fairly constant from 4.5 to 4.6 dBd across the band. Again, the array feed-point impedance can be corrected easily to provide a low SWR. The calculated patterns for the antenna at the top and bottom ends of the 20-meter band are shown in Figure 5. These take into account the changes in element phasing, spacing, current magnitude and element self-impedance with frequency.

Construction

The antenna is made from quality 300-ohm TV line to the dimensions given in Figure 6. Sketches of the insulators are provided in Figure 7. They are made from scrap pieces of epoxy-glass PC-board material. This results in ruggedness and minimum weight. The spreaders are readily available 10-foot lengths of "1 inch" (1$\frac{5}{16}$ in OD)[4] schedule 40 PVC pipe. The capacitors are used only to provide a good match to 50-ohm feed line: They don't otherwise affect the performance of the array. Small 500-V mica or monolithic ceramic units may be used for power levels up to a few hundred watts, since they are at a relatively low-voltage part of the system. Open-ended stubs could probably be substituted for the capacitors, if desired. I recommend that a balun transformer be used with this antenna. Attempts to measure the impedance of one element of this array resulted in a unique experience — the first substantial evidence of the need to use a balun transformer. The antenna-bridge readings varied greatly as the measuring equipment was moved, or as I placed my hand around the feed line. This ceased when I added a balun transformer. The phenomenon is explained by Maxwell in a recent paper.[5] Nearly any style of balun transformer will prevent the unwanted flow of current on the coaxial cable outer conductor. I use a choke type of balun transformer. It consists of 10 turns of small-diameter coaxial cable

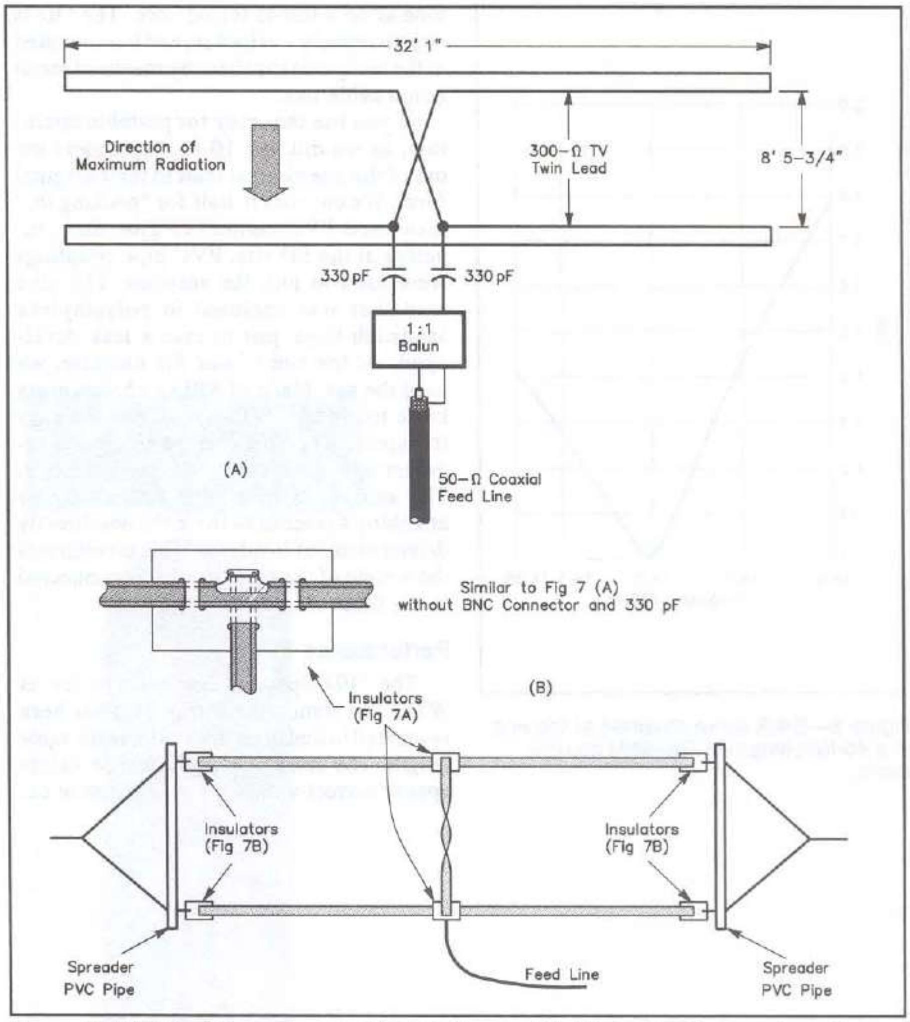
32' 1"
Direction of Maximum Radiation
300-Ω TV Twin Lead
8' 5-3/4"
330 pF
330 pF
1:1 Balun
(A)
50-Ω Coaxial Feed Line
Similar to Fig 7 (A) without BNC Connector and 330 pF
Insulators (Fig 7A)
(B)
Insulators (Fig 7B)
Insulators (Fig 7B)
Spreader PVC Pipe
Feed Line
Spreader PVC Pipe

Figure 6—Electrical dimensions for the W7EL array (A). Illustration B shows how the antenna is assembled on spreaders of PVC pipe.

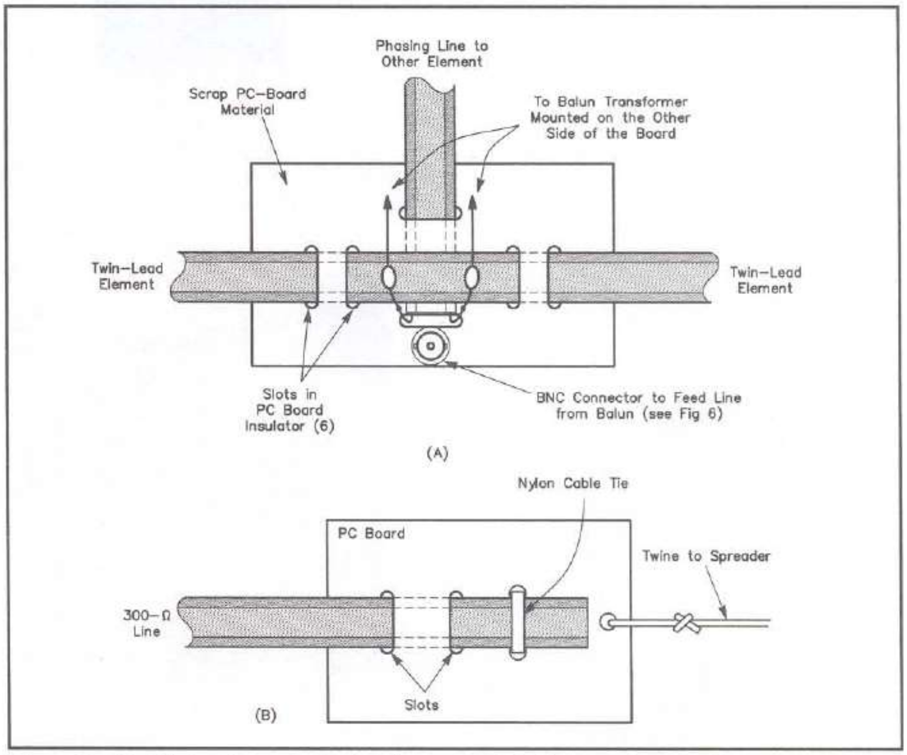
Phasing Line to Other Element
Scrap PC-Board Material
To Balun Transformer Mounted on the Other Side of the Board
Twin-Lead Element
Twin-Lead Element
Slots in PC Board Insulator (6)
BNC Connector to Feed Line from Balun (see Fig 6)
(A)
Nylon Cable Tie
PC Board
Twine to Spreader
300-Ω Line
Slots
(B)

Figure 7—Details for the antenna insulators used in the two-element 20-meter array. The drawing at A shows how the TV ribbon is affixed to the feed-point/phasing-line insulating block. The example at B provides details for the end insulators.

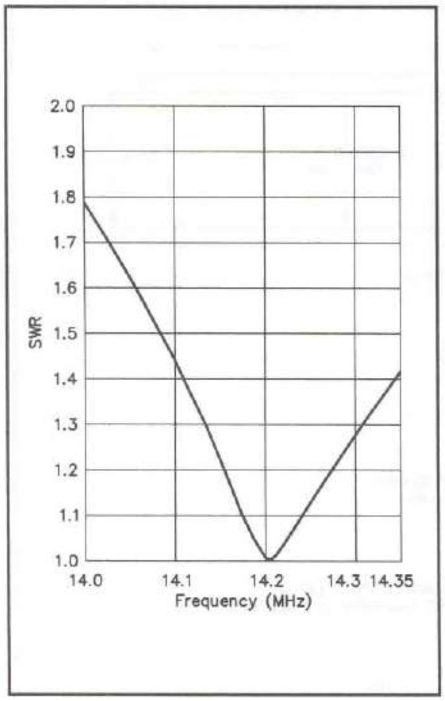

Figure 8—SWR curve obtained at the end of a 45-foot length of RG-58/U coaxial cable.

wound on a ferrite toroid core. The OD is approximately 1 1/4 inches, and it is mounted at the feedpoint insulator by means of small nylon cable ties.

If you use the array for portable operation, as we did, the 10-foot spreaders are out of the question, at least in their original form. We cut ours in half for "packing in," then used PVC cement to glue them together at the FD site. PVC pipe couplings were used to join the sections. The glue container was enclosed in polyethylene sandwich bags, just in case a leak developed. At the end of our FD exercise, we used the saw blade of K0ED's Swiss army knife to cut the PVC pipes again for easy transport. The spreaders were glued together again for use at the home station. The antenna is held horizontal easily by attaching a piece of twine to the nondirectly driven element insulator. This counteracts the weight of the feed line that is connected to the other element.

Performance

The "FD Special" has been in use at W7EL for some time. Array gain has been compared to that of an inverted V at the same height. The calculated performance values appear correct within the measurement capability. The front-to-back ratio has not been measured. The SWR at the end of 45 feet of RG-58/U feed line is shown in Figure 8. The SWR is important only when a lossy line feeds the array, or when it is driven by an intolerant transmitter (with built-in SWR shut down), which is now the norm.

Perhaps the most revealing performance indication was provided by a person who encountered us several times on 20 meters during Field Day. He was operating for another, very competitive local club. After the exercise he remarked, "The only reason I believe you guys were running an honest 10 W is that I know Wes Hayward (W7ZOI) was there." Indeed, we used 10 W or less input while operating—and 0 W while watching the mountain scenery!

Notes

[1] My apologies to the first person who described or named this antenna. I don't know its history. (See L.A. Moxon, "Two-Element Driven Arrays," *QST*, July 1952, p 28.)

[2] Named after W8JK, Dr. John Kraus, "Antenna Arrays with Closely-Spaced Elements," *Proc. IRE*, Feb. 1940.

[3] G. Hall, ed., *The ARRL Antenna Book* (Newington: ARRL, 1982).

[4] mm = in x 25.4; m = ft x 0.0348.

[5] W. Maxwell, "Some Aspects of the Balun Problem," *QST*, March 1983.

By Riki Kline, 4X4NJ **From *QST*, February 1985**

Build a 4X Array for 160 Meters

Low-angle radiation and electrical rotation of directivity are the features of this vertically polarized top-band antenna. If you are interested in 160-meter DX, this system could be your secret weapon.

Is your lament a common one — no room for an effective DX antenna on 160 meters? This complaint is voiced frequently by amateurs who live in urban areas, or who are programmed toward horizontal wire antennas. But, a number of successful top-band operators have adopted the philosophy, "If you can't go out, go up!" It is no secret that a physically short vertical antenna is generally more effective than a horizontal antenna that is close to the ground electrically, at least for DX work.

My 4X array is electrically rotatable. It is compact and is effective as a low-angle radiator. Let's examine how my antenna evolved from some basic designs. I will also cover the practical details of construction and system performance.

The Tilted Ground-Plane Look

The tilted ground plane is almost identical to the usual vertical. The physical format of this antenna resembles a four-conductor ground plane. The major difference is that the radiating elements tilt up toward the supporting structure. The 4X array contains four sloping ground planes. Each of the slope wires is 100 feet long.[1] They are supported at the high end by an 80-foot tower. A four-element, 20-meter Yagi antenna is atop the tower.

Each of the sloping wires is fed separately near ground by means of a tapped-coil matching device (Figure 1) that is returned to radial wires and ground rods. In effect, each radiator is a ground-plane vertical antenna that is slightly less than 0.25 wavelength. The matching inductor provides resonance and effects an impedance match to the coaxial feed line.

I believe that the metal tower and 20-meter antenna may possibly be functioning as a reflector because the Yagi antenna and tower combined with ground wires are resonant slightly below 1.8 MHz.

Two Tilted Ground-Plane Verticals in Phase

I had excellent results with one sloping vertical. Next, I installed a second system in the opposite direction. Switching between the two antennas (north-south sloping radiators) showed considerable front-to-back ratio (a relative reading of 15-20 dB). Subsequently, I connected the two antennas in phase. This gave a bidirectional pattern, east and west. Although I did not gather extensive data on the performance, I observed a 6-dB signal improvement with stations about 700 miles to the

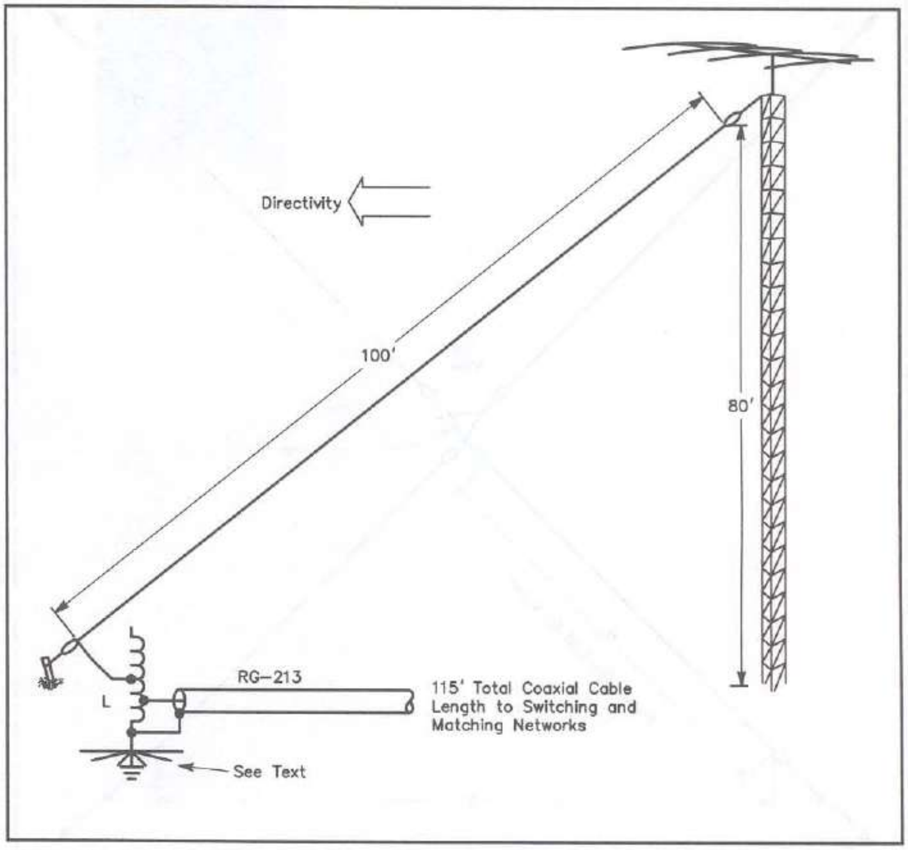
Directivity
100'
80'
RG-213
L
115' Total Coaxial Cable Length to Switching and Matching Networks
See Text

Figure 1—Basic tilted ground-plane vertical. L has 25 turns of heavy conductor (see text); length is 5½ inches, and diameter is 3 inches. Radiator is tapped 15½ turns above ground.

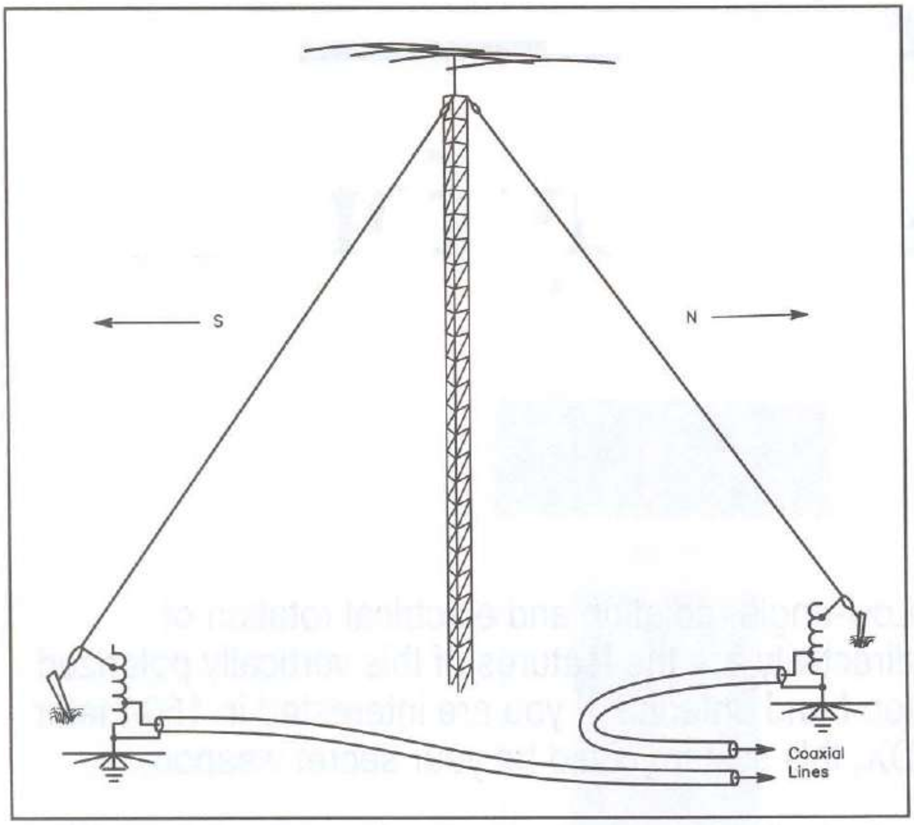

Figure 2—Two tilted ground-plane antennas that can be fed separately or in phase. The feed method is shown in Figure 1, with the method of Figure 7 used for feeding the antennas in phase.

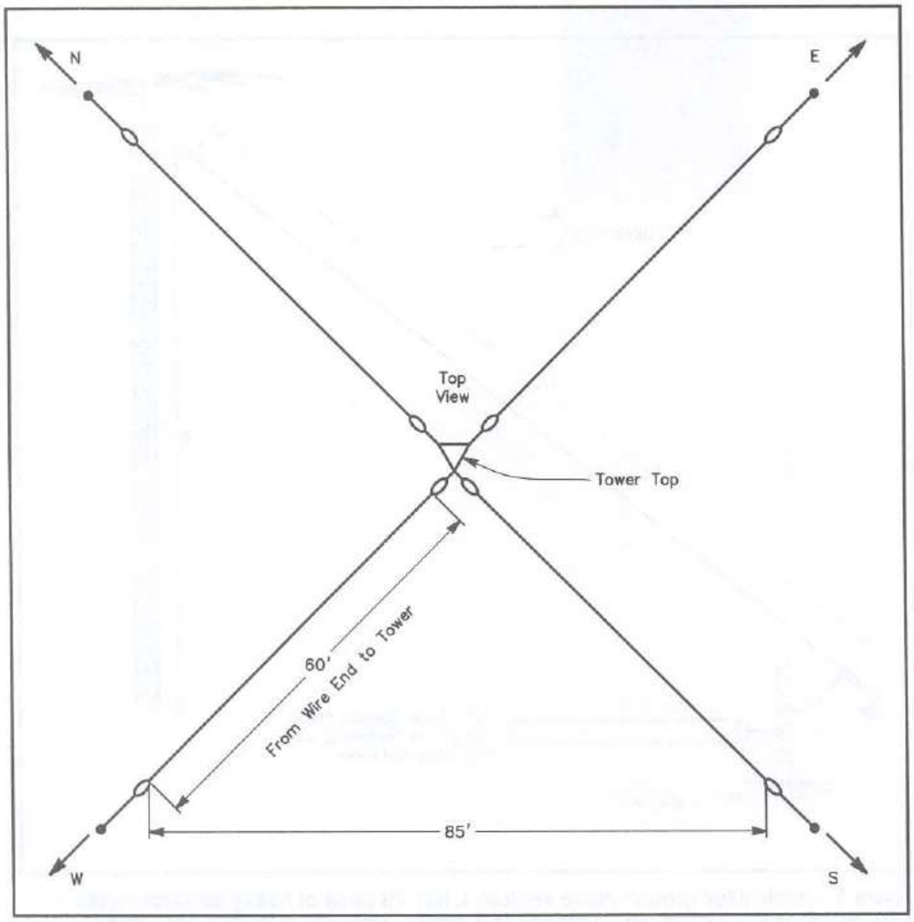

Figure 3—The 4X array as viewed from above the tower.

east. Some of you may want to explore the possibilities further. Figure 2 shows the details of the two-element phased system.

The 4X Configuration

Two more tilted ground-plane verticals were added, thereby providing east-west sloping radiators (Figure 3). A switching and phasing arrangement was added to my 4X array. It allows me to feed any of the slope wires separately, adjacent pairs in phase, or all four wires in phase. When using adjacent pairs in phase, maximum radiation is along a line that bisects the angle between the two antennas (NE, SE, SW or NW directions). When I feed all four wires in phase I note that the radiation is essentially omnidirectional. All of the unfed radiators are resonated to serve as reflectors. This concept is described in *The ARRL Antenna Book.*[2] My switching network is shown in Figure 4.

Phasing Networks

Most phasing methods call for long lengths of non-50-ohm coaxial cable.[3] I found this economically prohibitive. This negative factor inspired the approach I am using.

Each of my radiators is fed by means of 115 feet of RG-213 coaxial cable (formerly RG-8A/U 50-ohm line). The coil at the base of each wire is adjusted for the same resonance and SWR as the remaining three coils. The verticals to be fed in phase have their transmission lines connected in parallel through a suitable network for changing the reflected impedance back to 50 ohms. My networks are shown in Figures 7 and 8.

Tapped-Coil Matching

By using inductance and no intentional parallel capacitance for my matching coils, I am able to obtain greater effective antenna bandwidth because of reduced Q. Stray capacitance and antenna capacitance to the tower and ground are present, however. All electrical connections are soldered. A simple rain cover is used over each coil to protect it from moisture and dirt. The absence of switches, variable capacitors and rotary inductors enables construction of a highly reliable matching system without the need for weatherproof boxes.

This system is relatively easy to tune to obtain nearly identical performance from each antenna branch. This becomes a necessity when using "brute-force" parallel feed in the phased-pair and omnidirectional modes. Otherwise, the power distribution and phasing would be disturbed. This would distort the radiation pattern. Large-diameter, heavy-conductor, air-wound coils are best for this job. Two of my coils are made from silver-plated 1/4-inch-diameter copper tubing. The two remaining coils are made from large, flat conductor material of the kind found in some rotary inductors.

I used a dip meter to adjust the coils for resonance (coaxial cables disconnected). My coaxial cables were tapped initially one

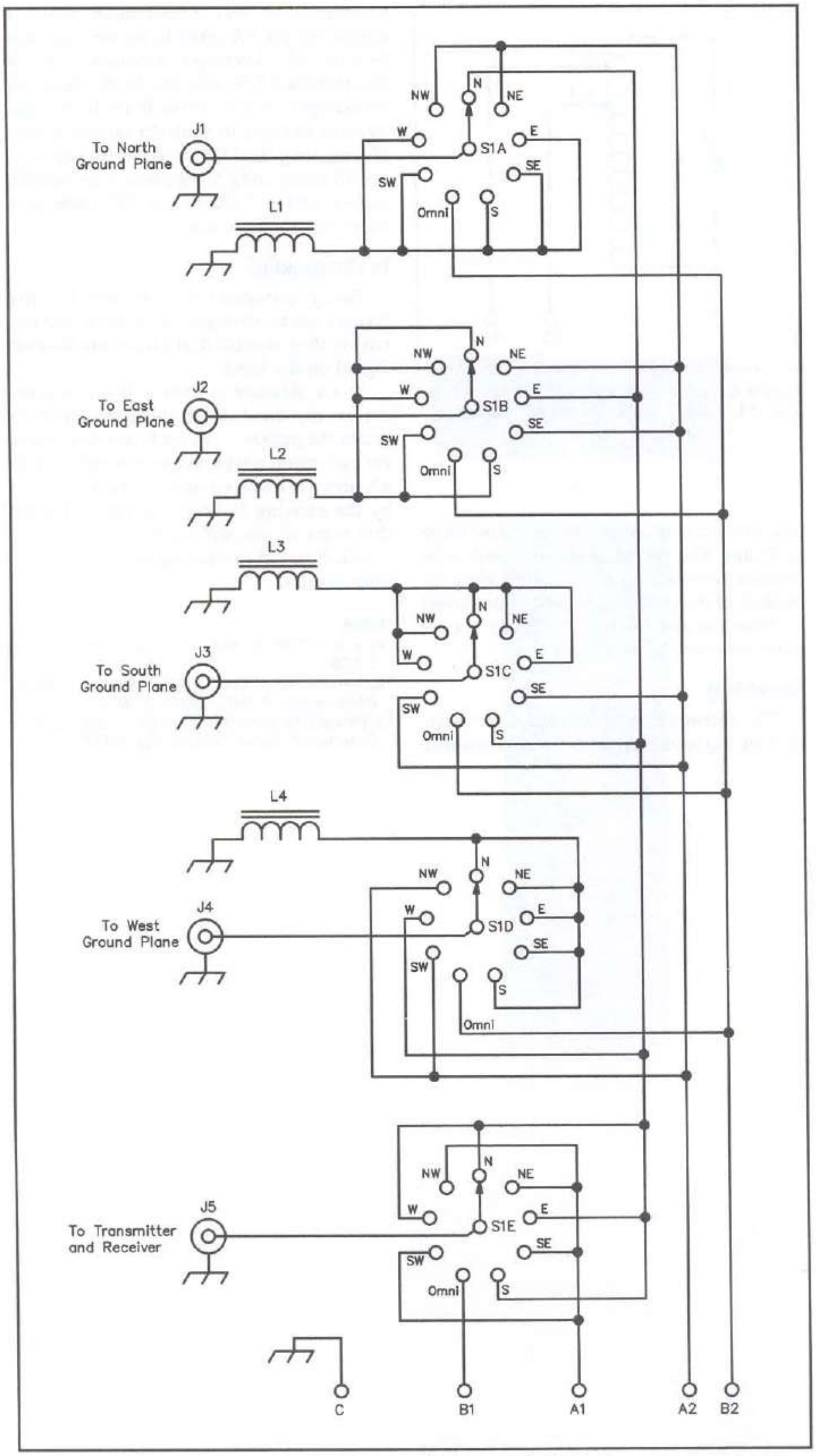

Figure 4—Switching system for the 4X array. L1 through L4, inclusive, are described in the text. A1, A2 and C are connected to the matching circuit shown in Figure 7. B1, B2 and C are connected to the matching circuit shown in Figure 8. S1 is a heavy-duty ceramic rotary switch, five poles, nine positions. J1-J5, inclusive, are coaxial connectors of the builder's choice.

Direction	SWR vs. Frequency				
	1800 kHz	1825 kHz	1850 kHz	1875 kHz	1900 kHz
N	1.4	1.0	1.2	1.8	2.6
NE	1.5	1.1	1.0	1.1	1.2
E	2.2	1.1	1.1	1.4	1.9
SE	1.2	1.0	1.0	1.0	1.1
S	1.4	1.0	1.1	1.4	2.3
SW	1.5	1.1	1.0	1.0	1.1
W	1.4	1.1	1.0	1.2	1.6
NW	1.2	1.0	1.0	1.0	1.2
Omni	1.4	1.1	1.0	1.0	1.0

Figure 5—Chart that shows SWR versus frequency in kilohertz.

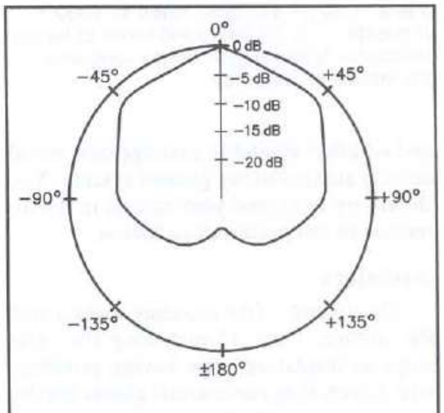

Tabulated Data	
Direction	Relative Strength
Forward	0 dB
±45°	−3 dB
±90°	−8 dB
±135°	−14 dB
±180°	−19 dB
Omni	−7 dB

Figure 6—Directivity pattern of the array. The pattern is a composite average of measurements made while receiving 10 different stations. There is no apparent difference between the two directive modes—single radiator or phased pair.

third of the way up from the ground ends of the coils. Final tap placement is made while feeding power to the antenna and observing an SWR meter. Alligator clips make this an easy matter to accomplish. When the SWR bottoms out at the same frequency for all four radiators, remove the alligator clips and solder the coil taps in place. Some interaction between the four antennas will occur, so make certain that all of the taps are where they belong before soldering them.

Reflector Tuning

The radiators not being fed are used as reflectors. This is done by switching small inductors in parallel with the ends of the coaxial feed lines. My inductors contain three or four turns of no. 16 wire wound around the center part of a 3/8-inch-diameter ferrite rod from a built-in AM broadcast receiver antenna. The coils are adjusted to give a resonance that is four percent lower than the resonant frequency of the radiators.

Ground Conditions

The efficiency and performance of the antenna depends on the quality of the ground system. Note 3 provides a good reference for ground systems, and a bibliography. Each of my radiators is worked against a counterpoise that contains two or three 1/4-wavelength wires, bent to fit in the boundaries of my property. In addition, I use a 10-foot rod in the ground. Water pipes

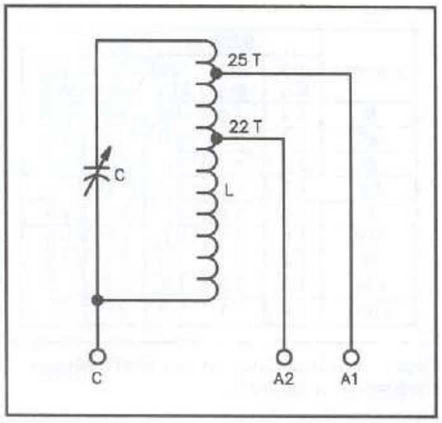

Figure 7—Phased-pair matching network. C is a 1000-pF variable, rated at 1000 V or greater. L is 30 airwound turns of heavy conductor (see text), 7 inches long and 3½ inches in diameter.

and all other available underground metal objects are tied to my ground system. You should try to extend your radials in the direction of the preferred radiation.

Insulators

The top ends of the radiators contain high RF voltage. I use 15-inch-long Plexiglas strips as insulators, after having problems with 8-inch-long commercial plastic insulators. Moisture and air pollutants caused these problems. The present insulators need to be cleaned periodically. High-quality glass insulators of the type used aboard ships should be excellent and, with luck, should not require periodic cleaning.

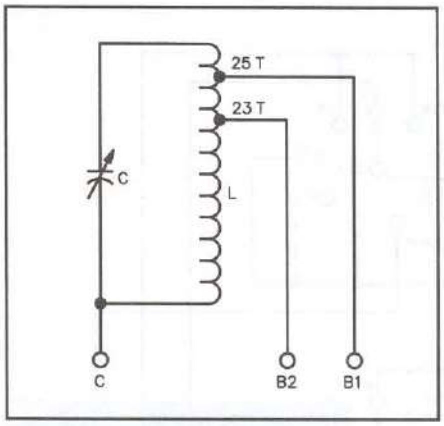

Figure 8—Network for omnidirectional use. C and L are the same as for Figure 7.

Receiving

The antenna directivity enhances reception by rejecting signals from unwanted directions. I was encouraged when I compared my 4X array to an 800-foot unterminated Beverage antenna that is bidirectional NW and SE. In the past, the Beverage wire showed an 8-10 dB S/N advantage over the tilted north vertical alone. During long-haul QSOs to North America, the 4X array, used in the phased-pair mode, comes within 3 dB of the S/N ratio provided by the Beverage.

In Conclusion

Many European stations tell me my signals are as strong as local ones. On occasion they remark that I have the loudest signal on the band.

It's a pleasure to have a directional array on top band. I simply turn a knob to rotate the pattern — much faster than a motor can rotate a typical beam antenna! The 4X array is compact and can be supported by the existing HF-antenna tower. I hope that some of you will try this antenna, and I look forward to hearing from you about your results.

Notes

[1] m = ft × 0.3048; mm = in × 25.4; km = mi × 1.609.

[2] G. Hall, ed., *The ARRL Antenna Book* (Newington: ARRL, 1984), p. 8-12.

[3] J. Devoldere, *80-Meter DXing* (Greenville, NH: Communications Technology, 1978).

By Rod Newkirk, W9BRD **From *QST*, June 1990**

The 'BRD Zapper: A Quick, Cheap and Easy "ZL Special" Antenna

Here's a multipurpose directional wire antenna that has its origins in Kraus, Windom and Newkirk *(who?–Ed.)*.

Working Europe from Chicago with low power on 21 MHz, using an indoor, bent-ends W8JK bidirectional wire beam (Figure 1)[1], is fairly straightforward fun. Voltage feed via a λ/4 stub, gamma-matched to coax, is hard to beat for spartan simplicity. No worry about balancing element currents, no nit-picking with element lengths. Just make the whole system symmetrical and dip it to your favorite frequency by adjusting the stub-shorting point.[2] The 8JK has a similar pattern and good gain all the way from the fundamental to the second harmonic, so it's really a multiband antenna. You can roll the whole thing up in three minutes when it's not in use.

One problem: The bidirectional characteristics of the 8JK cause my ears to be flattened regularly by undesired signals from the direction opposite Europe. Unless you're in the geographical center of a three-way QSO, the unused lobe of this antenna can be a nuisance. Question: Is it feasible to convert easy 180° bidirectionality to tricky 135° *uni*directionality without resorting to clumsy center-feeding, multi-wire-dipole elements, balanced-gammas, etc?[3]

Yes—because there's a *uni*directional ZL Special lurking in our little 8JK. The principal difference between the W8JK and the ZL Special lies in the phasing: In the 8JK, the elements are fed 180° out of phase; in the ZL Special, they're driven 135° out of phase. Therefore, instead of feeding the antenna near the stub's shorting point, we'll need to feed it λ/16 from the stub's shorted end. That's where the feed path to one element is λ/8 (45°) longer or shorter than the other, which, after the stub's 180° phase reversal, produces the 135° phase shift that were looking for between the elements.

You can find the proper feed point on the stub by "sniffing" signals of known origin along *one side* of the stub with the insulated center conductor of some coax hooked to a receiver. (Start looking for this point by measuring λ/16 up from the bottom of the stub.) It's refreshing to hear Europeans rolling in while most of the murderous rearward signals now barely budge the S meter! Directivity is reversed at the opposite stub point (Figure 2).

So he's in there all right, that ZL Special, but coaxing the wily rascal out for an honest day's work is a challenge. The ZL

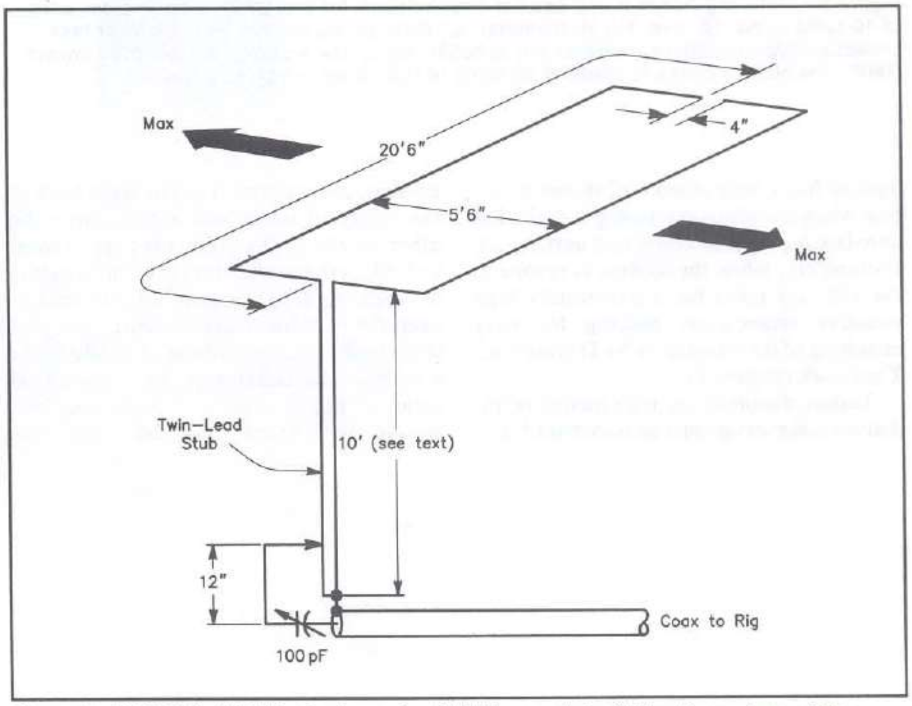

Figure 1—W9BRD's W8JK wire beam for 21 MHz, made to fit the dimensions of the townhouse bedroom that it occupies. Its dimensions aren't critical, although the element lengths should be close to λ/2 (total, each), the stub should be λ/4 or odd multiples, and the element spacing should be close to λ/8. At the bottom of the stub, connect both stub wires and the coax braid together, after resonating the system as outlined in Note 2. The 12-inch tap distance and 100-pF variable capacitor constitute a gamma match that brings the stub impedance to 50 Ω.

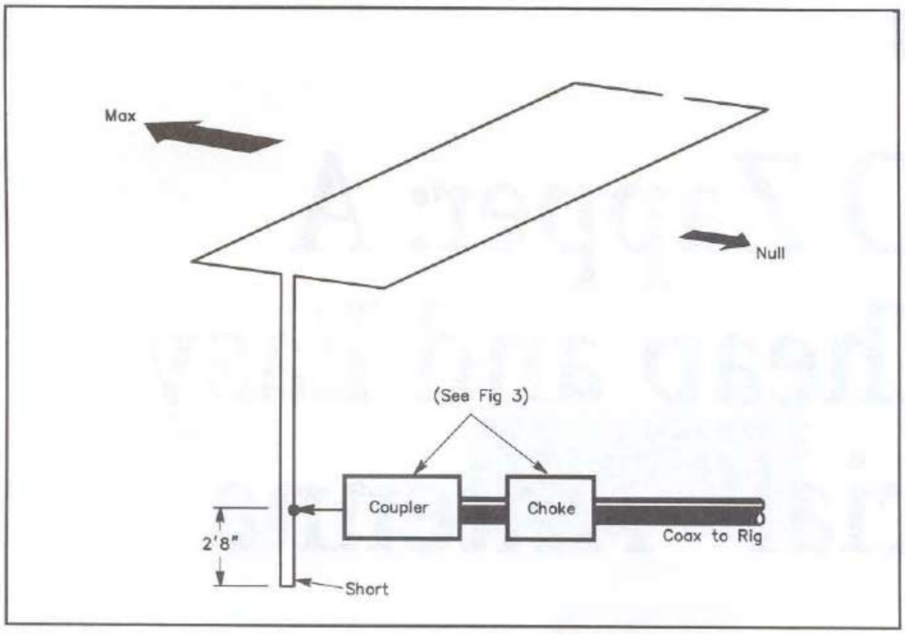

Figure 2—The antenna of Figure 1 configured for unidirectional operation with 135° element phasing. Resonate the system as described in Note 2 and connect the coupler via a very short wire to one leg of the stub about $\lambda/16$ above the stub-shorting point. The coaxial choke is needed to eliminate feed-line radiation, which can ruin the rearward null that this system provides.

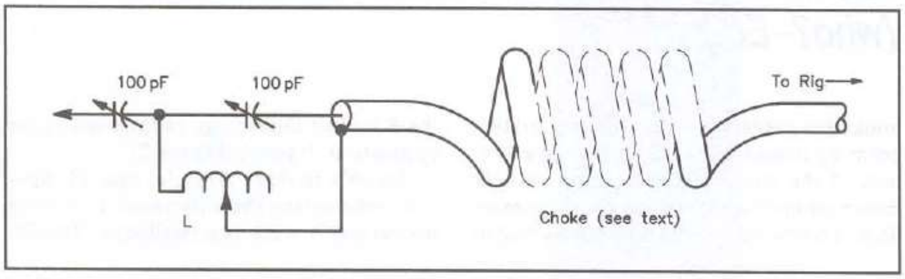

Figure 3—Matching-network and coaxial-choke details for the 'BRD Zapper. L consists of 10 turns of no. 16 wire, 1½ in. diameter, air core, space wound. For 100 W or less, broadcast-type variable capacitors are suitable. Adjust the inductor-tap point for lowest SWR. The coaxial choke is made of 30 turns of RG-58 wound on a ferrite rod.

Special has a very sharp null in one direction when the element phasing is right, but unbalancing the stub could cost us that null. Fortunately, when the system is resonant, the 135° tap point has a moderately high resistive impedance, making for easy matching of the antenna to 50-Ω coax via a T network (Figure 3).

Pattern distortion through incidental radiation and pickup must be minimized. Excessive leakage would put us right back at the mercy of those loud signals from the other direction. Therefore, the coupler must be built in the most compact form possible, mounted right at the stub, and isolated to keep the feed line from distorting the pattern. Such isolation is done at W9BRD via a home-brew coaxial choke made of 30 turns of the antenna's RG-58 feed line wound on a ferrite rod just before the matching network.

Because feeder radiation could upset the stub balance—and thus spoil the antenna's rearward null—the single-wire run from the matching network to the antenna had better be no more than an inch or two. Here, I'm borrowing on the single-wire-feed theme by Windom. Any circuitry above the choke will be hot with RF, so a bulky meters-bells-and-whistles matching unit will not do. It's best to use a compact, barebones, dedicated network. Run the stub all the way down from the flat top if possible. Incidentally, the tap point on the stub corresponding to the best rearward null is higher on the stub for element spacings wider than $\lambda/8$. System Q is also lower with wider spacings, but gain is maximum at $\lambda/8$.

As for the stub itself, TV twin lead is okay indoors, but close-spaced, end-fire elements mean high voltages and currents. If the stub is to be longer than a quarter wavelength (odd multiples only) or outside in the weather, low-loss, open-wire line is a must. You can run the stub of an outdoor version directly into the shack for handy directivity reversal.

Though we've been talking horizontal, this setup should be equally interesting as a vertical. By the way, when the elements are placed horizontally, one above the other, 135° phasing accentuates radiation and reception at high angles. Neat for local traffic nets, Sweepstakes and Field Day!

Notes

[1] Dr John D. Kraus, W8JK, first described this antenna in *QST* in "Directional Antennas with Closely Spaced Elements," *QST*, Jan 1938, pp 21-23, 37. Kraus also discussed this antenna in three articles in *Radio* magazine in 1937 and 1939, and in *Proceedings of the IRE*, Feb 1940.

[2] The easiest way to do this is with a dip meter. Jab a straight pin into each leg of the stub near the bottom, short the pins together and couple the dip meter into the system by bringing the dip motor close to the shorted pins. Locate resonance using the dip meter and measure the dip-meter frequency on your receiver. Move the pins up or down the stub and repeat the process until the antenna is resonant at or near your frequency of interest. When you've done this, cut the stub, strip and solder a short in the wires where the pins were located at resonance.

[3] Of course, this antenna can be center fed in either its bidirectional or unidirectional modes. Such flat tops are described in J. Hall, Ed., *The ARRL Antenna Book*, 15th ed, (Newington: ARRL, 1988), Chapter 8, and J. Devoldere, *LowBand DXing* (Newington: ARRL, 1987), pp 2-101 through 2-102.

Combination Driven Arrays

Broadside, end-fire and collinear elements can readily be combined to increase gain and directivity, and this is in fact usually done when more than two elements are used in an array. Combinations of this type give more gain, in a given amount of space, than plain arrays of the types just described. The combinations that can be worked out are almost endless, but in this section are described only a few of the simpler types.

The accurate calculation of the power gain of a multi-element array requires a knowledge of the mutual impedances between all elements, as discussed in earlier sections. For approximate purposes it is sufficient to assume that each set (collinear, broadside, end-fire) will have the gains as given earlier, and then simply add up the gains for the combination. This neglects the effects of cross-coupling between sets of elements. However, the array configurations are such that the mutual impedances from cross-coupling should be relatively small, particularly when the spacings are 1/4 λ or more, so the estimated gain should be reasonably close to the actual gain.

FOUR-ELEMENT END-FIRE AND COLLINEAR ARRAY

The array shown in Figure 1 combines collinear in-phase elements with parallel out-of-phase elements to give both broadside and end-fire directivity. It is popularly known as a "two-section W8JK" or "two-section flat-top beam." The approximate gain calculated as described above is 6.2 dB with 1/8-λ spacing and 5.7 dB with 1/4-λ spacing. Directive patterns are given in Figures 2 and 3.

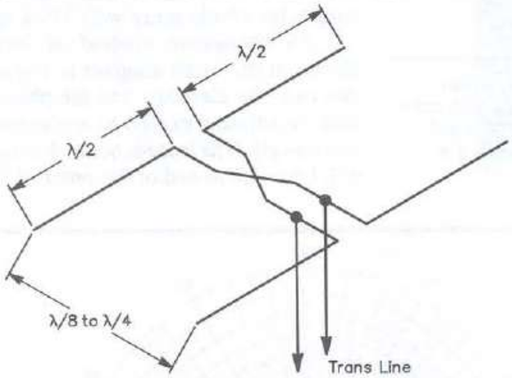

Figure 1—A four-element array combining collinear broadside elements and parallel end-fire elements, popularly known as the W8JK array.

The impedance between elements at the point where the phasing line is connected is of the order of several thousand ohms. The SWR with an unmatched line consequently is quite high, and this system should be constructed with open-wire line (500 or 600 Ω) if the line is to be resonant. With 1/4-λ element spacing the SWR on a 600-Ω line is estimated to be in the vicinity of 3 or 4 to 1.

To use a matched line, a closed stub 3/16 λ long can be connected at the transmission-line junction shown in Figure 1, and the transmission line itself can then be tapped on this matching section at the point resulting in the lowest line SWR. This point can be determined by trial.

This type of antenna can be operated on two bands having a frequency ratio of 2 to 1, if a resonant feed line is used. For example, if designed for 28 MHz with 1/4-λ spacing between elements it can be operated on 14 MHz as a simple end-fire array having 1/8-λ spacing.

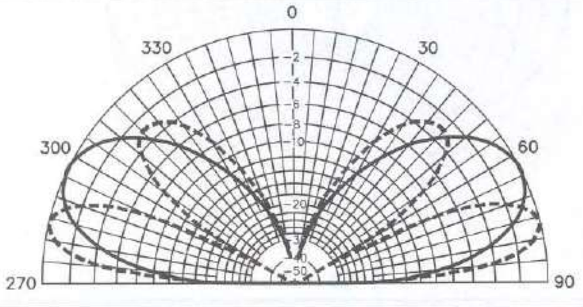

Figure 3—Vertical pattern for the four-element antenna of Figure 1 when mounted horizontally. Solid curve, height 1/2 λ; broken curve, height 1 λ above a perfect conductor. Figure 2 gives the horizontal pattern.

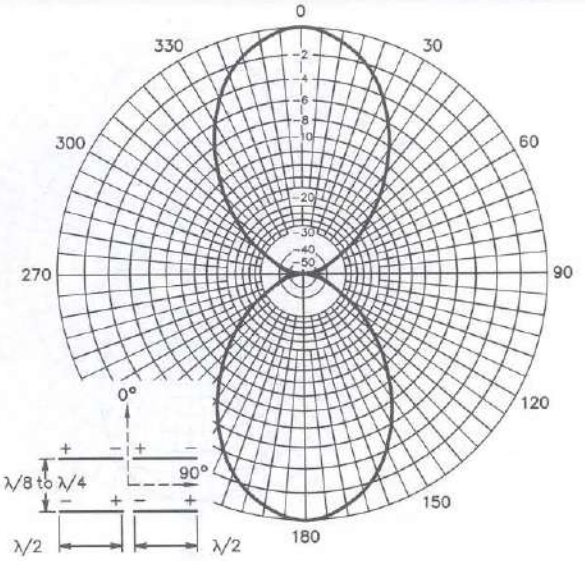

Figure 2—E-plane pattern for the antenna shown in Figure 1. The elements are parallel to the 90°-270° line in this diagram. Less than a 1° change in half-power beamwidth results when the spacing is changed from 1/8 to 1/4 λ.

FOUR-ELEMENT BROADSIDE ARRAY

The four-element array shown in Figure 4 is commonly known as the "lazy H." It consists of a set of two collinear elements and a set of two parallel elements, all operated in phase to give broadside directivity. The gain and directivity will depend on the spacing, as in the case of a simple parallel-element broadside array. The spacing may be chosen between the limits shown on the drawing, but spacings below 3/8 λ are not worthwhile because the gain is small. Estimated gains are as follows

3/8-λ spacing—4.4 dB
1/2-λ spacing—5.9 dB
5/8-λ spacing—6.7 dB
3/4-λ spacing—6.6 dB

Half-wave spacing is generally used. Directive patterns for this spacing are given in Figures 5 and 6.

With 1/2-λ spacing between parallel elements, the impedance at the junction of the phasing line and transmission line is resistive and is in the vicinity of 100 Ω. With larger or smaller spacing the impedance at this junction will be reactive as well as resistive. Matching stubs are recommended in cases where a nonresonant line is to be used. They may be calculated and adjusted as described in Chapter 26.

The system shown in Figure 3 may be used on two bands having a 2-to-1 frequency relationship. It should be designed for the higher of the two frequencies, using 3/4-λ spacing between parallel elements. It will then operate on the lower frequency as a simple broadside array with 3/8-λ spacing.

An alternative method of feeding is shown in the small diagram in Figure 3. In this case the elements and the phasing line must be adjusted exactly to an electrical half wavelength. The impedance at the feed point will be resistive and of the order of 2 kΩ.

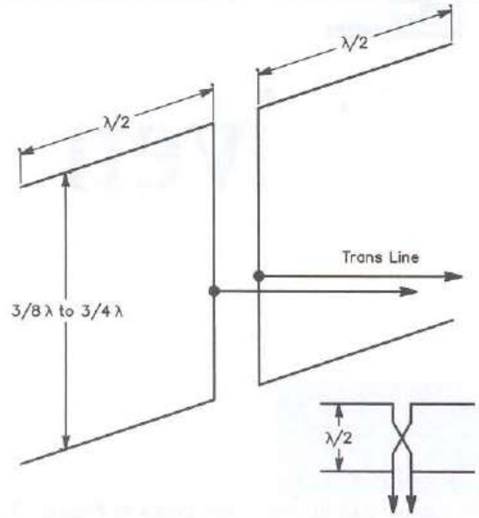

Figure 4—Four-element broadside array ("lazy H") using collinear and parallel elements.

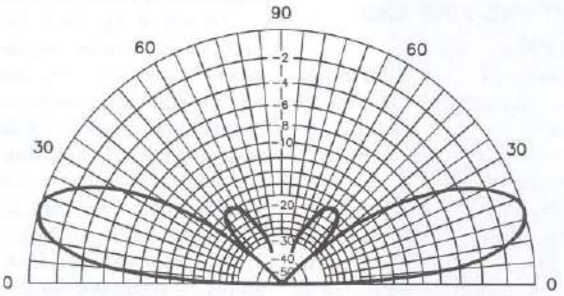

Figure 6—Vertical pattern of the four-element broadside antenna of Figure 3, when mounted with the elements horizontal and the lower set $^1/_2$ λ above a perfect conductor. "Stacked" arrays of this type give best results when the lowest elements are at least $^1/_2$ λ high. The gain is reduced and the wave angle raised if the lowest elements are close to ground.

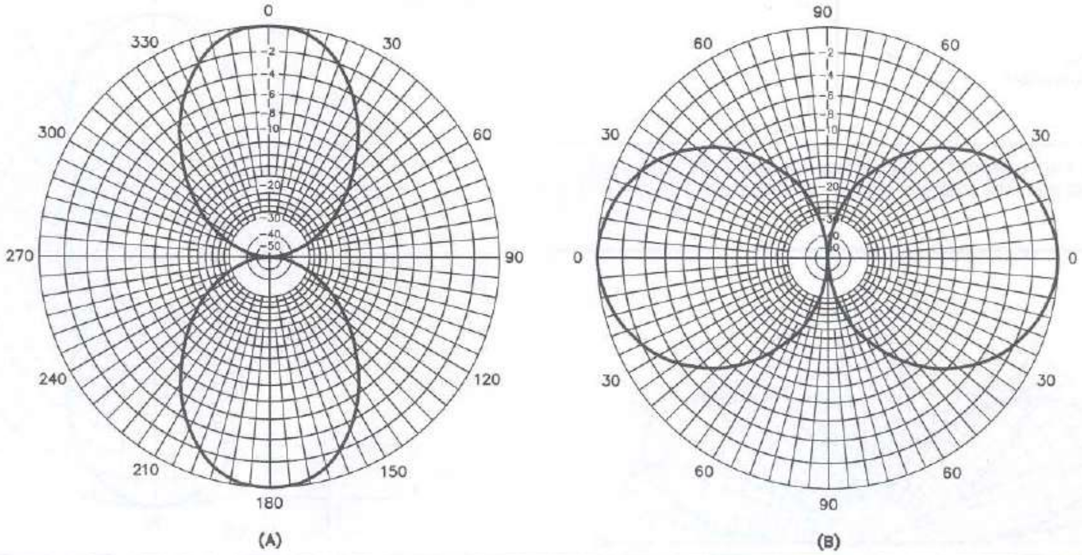

Figure 5—Free-space directive diagrams of the four-element antenna shown in Figure 4. At A is the E-plane pattern, the horizontal directive pattern at low wave angles when the antenna is mounted with the elements horizontal. The axis of the elements lies along the 90°-270° line. At B is the free-space H-plane pattern, viewed as if one set of elements is above the other from the ends of the elements.

CHAPTER SEVEN

VERTICALLY POLARIZED

7-1 **More on "Half Slopers" and Half-Sloper Variation for 160 Meters**
From December 1979 *QST* (Technical Correspondence)

7-3 **A Two-Band Half-Sloper Antenna**
From June 1980 *QST*

7-7 **The KI6O 160-Meter Linear-Loaded Sloper**
From April 1986 *QST*

7-8 **The Super Sloper**
From December 1995 *QST*

7-12 **The Inverted L Revisited**
From January 1983 *QST*

7-14 **160-Meter Antennas**
From July 1991 *QST* (Technical Correspondence)

7-16 **A Truly Broadband Antenna for 80/75 Meters**
From April 1986 *QST*

7-19 **Improved Broadband Antenna Efficiency**
From August 1987 *QST* (Technical Correspondence)

7-20 **Simple, Effective, Elevated Ground-Plane Antennas**
From June 1994 *QST*

From *QST*, December 1979 (Technical Correspondence)

More on "Half Slopers"

The July 1979 *QST* article on quarter-wave slopers has prompted me to write you on the subject. I have had considerable experience using them on 160 and 80, and I have tried them on 40.

The first one I know who used this type antenna was W3AU. His success with the system on 160 caused him to recommend it to W7RM and me. The original configuration used more than one sloping element fed in parallel against the tower. It resembled a top-loaded vertical with a simulated ground system. All my results have supported the view that this is an omnidirectional, vertically polarized radiating system, even when only one sloping element is used.

All the antennas of this type I have tried have been at my father's station, W6UA, and have been on a single tower. The tower is 82 feet (25 m) high, approximately 1½ feet (0.45 m) in uniform triangular cross section, and has a single set of guys at the 70-foot (21.3-m) level. The guys are broken with insulators at about 20-foot intervals, and the top section of each guy is electrically connected to the tower. The base has whatever insulation the concrete provides and there are 10-, 15- and 20-meter Yagis on top of the tower. (This is an AB105 tower, the same as is used by W3AU and W7RM.)

The first antenna of the type in question was tried on 80 meters. It was attached at the 60-ft level, so that top-loading effects of the guys, as well as the rotaries, were probably effective. The attached wire came off at 30° to vertical and was trimmed for minimum VSWR. The resultant length was within 2 feet of the formula length, and the antenna then had a VSWR less than 1.5 from 3.5 to 4.0 MHz. It worked well on transmit but was extremely noisy on receive. The original wire sloped east, so I put up an identical system, attached to the same tower level, sloping west. Electrical characteristics were nearly identical, and no directivity difference between the two could be detected either on transmit or receive.

This is logical considering how little of the antenna current in the sloping wire has any horizontal component, other antenna systems used on 80 at W6UA have been dipoles, inverted Vs, and a delta loop. The quarter-wave sloper is better than any of the others for transmitting, but it is poor on receiving. I have taken down all but the two slopers. They are used for transmitting and receiving, with frequent use of a 40-meter collinear antenna on receiving weak signals.

I also tried this system on 40, with a wire attached to the tower at 30 feet, coming off at 30°. Perhaps because there was so much tower electrically above the attachment point – at least half a wavelength – this one was much harder to tune. The wire length, which was very critical, was some feet shorter than expected, and the VSWR behavior across the band less benign. Not much time was spent on this particular antenna because it seemed no better than a λ/2 vertical. Both worked well, but neither was the equal of the collinear types available.

The most recent quarter-wave sloper system I have used has been on 160 meters. My first attempt was a single sloping element from the 75-foot level. This places whatever contribution there is from the top guy sections below the point of attachment. I couldn't get this antenna to match by altering the length of the sloping wire, even while making rather severe adjustments. There did seem to be some promise for the system, so the next season I used two sloping elements of about the "correct" length and fed the arrangement with a matching network at the bottom of the tower, with 300-ohm twin-lead going up the tower (Figure 1). This was the best antenna I have used on 160 at W6UA. Previously I had used an inverted V, a center-fed 80-meter dipole at 90 feet and a λ/2 inverted L. Since the wires must come off the tower at about 45-50° to vertical, the rationale of using two wires instead of one was to attempt to cancel the horizontal current component. As a matter of curiosity I tried adding a third wire and could find no change in setting of the matching components or in the

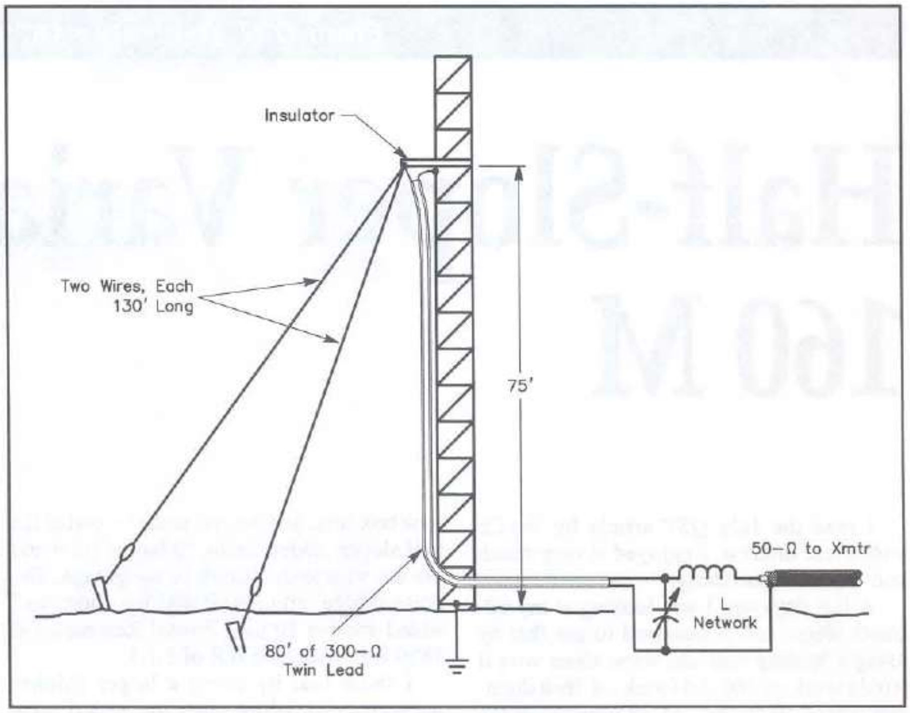

Figure 1—Illustration of the arrangement used at W6UM for a 160-meter half sloper.

overall effectiveness. Using some rough estimates of component values required for matching, length of feed line, etc., I calculate the feed impedance of the antenna itself to be approximately 100–*j*300 ohms.

My estimate of how well this antenna performs must be tempered by the realization that conditions during the past two years have been much worse than the period before. During the 160-meter contests, openings to W1, W2 and W3 have been marginal. I have worked nothing with this antenna I hadn't worked before, but the impression remains that the antenna is better than those used earlier. It has done a consistent job into the Caribbean, SA, Pacific and Japan, as well as across the USA. Receiving remains a problem because of the noise. I have taken down the 160-meter version of this antenna. If I put it up again I will definitely install some auxiliary receiving antennas.

The net result seems to be that this is an effective antenna. Getting a good match directly into coax appears easier if the sloping wire is more nearly vertical and if there is not too much loading above the attachment point. The system should be more efficient, up to a point, as there is more vertical structure above the feed point. There is no compelling reason to insist upon using a resonant configuration. The next attempt I make on 160 might well be with the existing 80-meter system, fed at 60 feet and with 80-meter sloping elements, but using a transmission line and matching network at the bottom of the tower as before.

Given that the antenna does work, the areas I feel need further investigation are better characterization and description of it for different parameters, e.g., measurement of the feed impedance as a function of frequency for different sloper attachment points, lengths, angles and conditions of top loading. I would also like to see the current flow on all of the structure mapped via a probe. If some of the results are forthcoming I hope to see them in an early issue of *QST*. — *Charles Weir Jr., W6UM*

From *QST*, December 1979 (Technical Correspondence)

Half-Sloper Variation For 160 M

I read the July *QST* article by W1CF with great interest. I enjoyed it very much and learned something.

A few days ago I was looking at my 80-meter sloper and it occurred to me that by using a loading coil and some more wire it would work on 160. So I took a 4-inch diameter coil of 10 turns of no. 10 wire out of the junk box, attached the coil to the far end of the half sloper, added about 30 feet of wire and ran the wire to the corner of my garage. The noise bridge indicated it was too short, so I added another 10 feet. Presto! Resonance at 1830 kHz with an SWR of 1.1:1.

I think that by using a larger (higher inductance) loading coil, the added wire could be shortened considerably without seriously affecting the 2-to-1 SWR points on the curve.

I have not made too many contacts on 160 yet, as there is very little activity this early in the season around here. I have had several contacts though, always getting at least as good a report as I have given. *—Philip True, W7AQB.*

By Gary E. Myers, K9CZB

From *QST*, June 1980

A Two-Band Half-Sloper Antenna

When an off-the-wall empirical design like this works well, suspicion and skepticism are warranted. Maybe you'll agree that this sloper idea is an exception!

The popularity of the half-sloper antenna seems to be increasing, as evidenced by recent articles in *QST*.[1,2,3] This type of antenna has some worthwhile advantages, particularly for the lower frequency operator – low-angle radiation for antennas of modest height, compactness and simplicity of construction. On the minus side, narrow bandwidths and difficulties in resonating the system have been reported.[3]

The antenna system to be described here evolved from a simple-minded attempt to design a two-band half sloper for 80- and 40-meter operation. A trap-type of antenna was selected as the design basis because of previous experience with trap antennas and because the inductive loading of the trap on the lower-frequency band allows a somewhat shorter overall length. The same inductive loading, however, was also expected to increase the antenna Q and thereby further decrease the bandwidth. For this reason I was prepared to experiment.

It is well that I was so prepared because the final form of the antenna bears little resemblance to the initial concept. But most interesting — and exciting — is the impressive bandwidth on 80 meters.

The Design

A diagram of the two-band half-sloper antenna system is shown in Figure 1. It can be seen that there is no obvious quarterwave dimension in the entire system. In fact, the radiator itself is a nonresonant device.

Initial attempts to prune the wires to resonance resulted only in a mound of wire clippings and one frustrated amateur. After many hours of cut-and-try experimentation, accompanied by a growing, gut-level appreciation for what apparently was happening, the magic combination of wire lengths and trap component values was found. Impedances of 40 $-j80$ ohms at 7.2 MHz and 60 $-j40$ ohms at 3.6 MHz were measured with a noise bridge. It was then a simple matter to cancel out these capacitive reactances with an inductor.

For convenience, the inductor was placed in the transmission line, rather than at the feed point – final "tweaking" of the system is more easily performed on the ground than at 40 feet (12.2 m) in the air. A Smith Chart exercise shows that the SWR on the transmission line between the feed point and inductor L is 5:1 at 7.2 MHz and 2:1 at 3.6 MHz. When RG-8/U is used, the additional loss incurred because of these SWRs is less than 0.5 dB at 7.2 MHz and is almost nonexistent at 3.6 MHz.

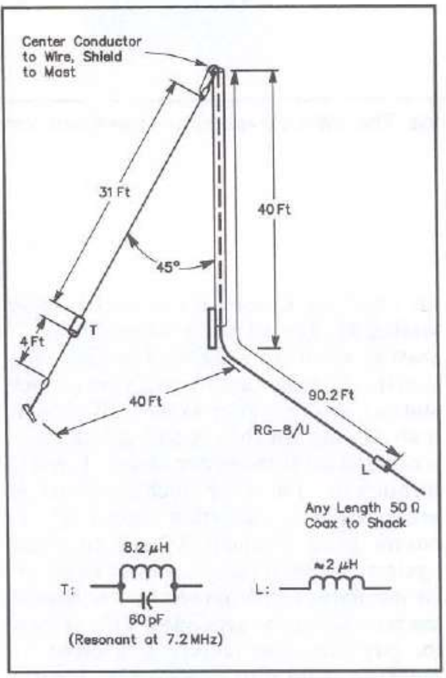

Figure 1—The two-band half sloper. To shield the transmission line from the antenna field, the line is routed up the inside of the support mast, which is grounded. Inductor L is needed to resonate the system.

Performance

This antenna performs very well. Operation at K9CZB is primarily 80-meter cw, with some 40-meter ssb. Power output is nominally 100 watts. Signal reports on 80 have been uniformly good, with comments such as U R LOUDEST 9 ON BAND AND VY FB SIG, VY STRONG. Voice operation on 40 has also resulted in good signal reports, although the praise has not been so lavish. This is my only 80/40-meter antenna, so direct comparisons were not possible. However, it appears to greatly outperform two previous antennas, a 160-foot end-fed wire and a trap dipole, both strung 30 feet above ground. All in all, it is about what I expected from a half sloper.

This kind of performance is nice, but nothing to write an article about, since that has already been done. The real performance story about this antenna can be summed up in one word: bandwidth. A glance at the SWR curves in Figure 2 will open the eyes of any 80-meter operator. As far as I know, this is unheard-of bandwidth for such a simple and compact antenna. It is a real treat to QSY 400 kHz And see the SWR meter needle barely move. This isn't a low-Q antenna —it's a *no*-Q antenna! The bandwidth on 40 is far less impressive and is, in fact, similar to what has been reported previously for half slopers.

Construction

During experiments with prototypes of this antenna, I noticed some sensitivity to feed-line placement and length. Therefore,

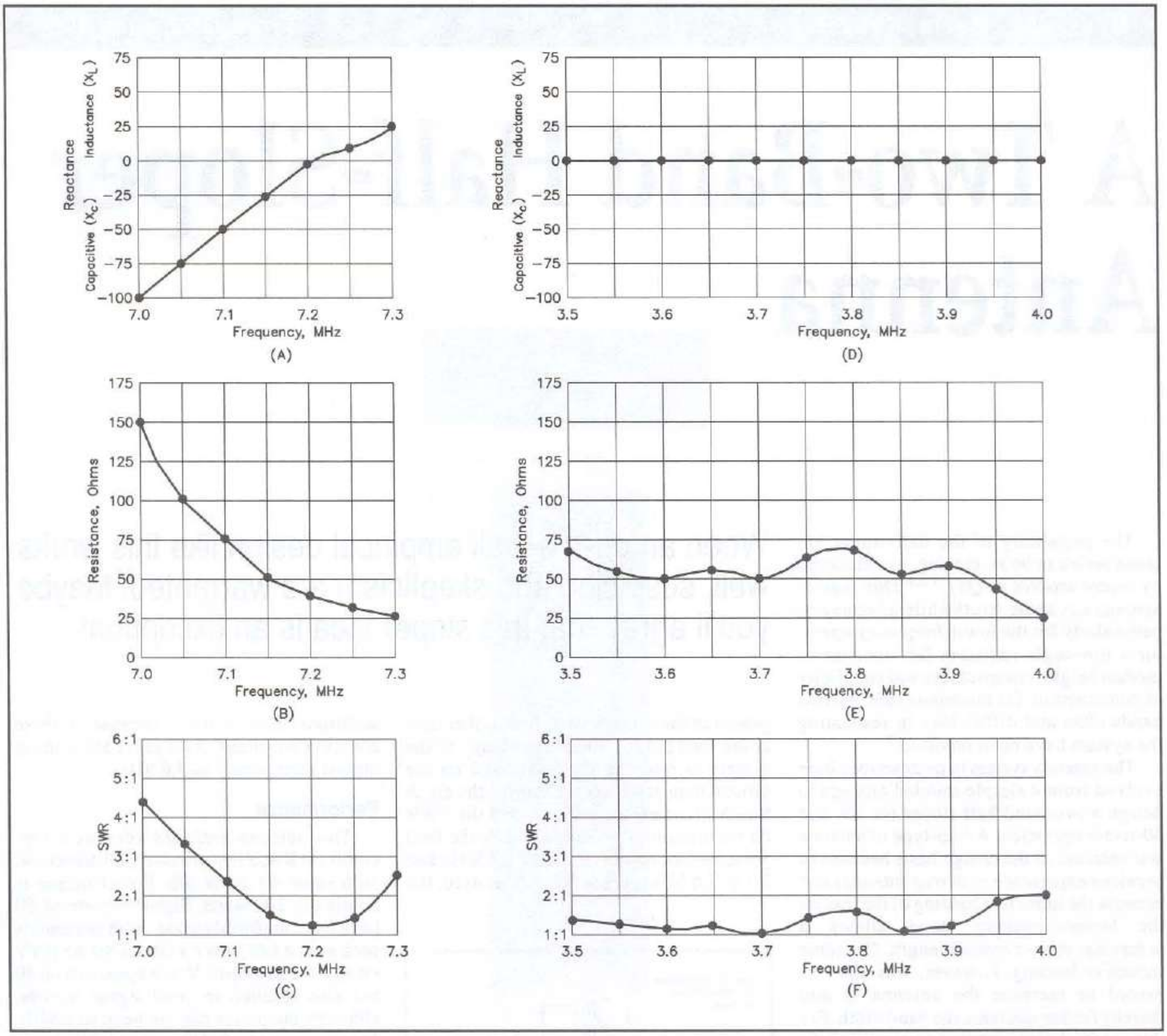

Figure 2—Loading characteristics of the antenna. The SWR curves were determined from a Smith Chart. Impedance values were measured with a noise bridge.

in later versions I ran the feed line up the *inside* of the support mast to shield it from the antenna field. This precaution seems to be effective, for no such sensitivity has been observed since. (I can't help but wonder if this might improve the behavior of any cantankerous half sloper.) If a nonmetallic support is used, or if there is no possibility of placing the feed line inside the support, double-shielded coaxial cable should serve equally well, provided the outer braid is connected to the inner braid at the top of the tower and a ground is connected to the outer braid at the bottom of the tower.

I used standard 10-foot TV mast sections for my support, simply because they were on hand. This results in a very flimsy and flexible mast in a 40-foot length, though. The first 20 feet must be doubled up with long U-bolts if the mast is to be walked up. The price of six sections of TV mast is about the same as a 36-foot telescoping push-up mast, but the latter is much sturdier and far easier to erect. If support materials are not already on hand, the telescoping mast is the better choice. I should mention that I took the trouble to bond all sections of mast together electrically to ensure good conductivity and to guard against TVI from rectification at joints after inevitable corrosion sets in. The base of the mast should be grounded. Effects from the guy wires can largely be avoided by breaking them into nonresonant lengths with strain insulators placed at the mast and every 19 feet thereafter.

The inductance and capacitance values shown in Figure 1 must be used for the trap. Construction techniques for the trap are covered in *The ARRL Antenna Book*.[4] A novel and inexpensive method of trap construction has been described by WB9OQM.[5] I built my trap using the method shown in Figure 3. Traps made in this fashion are much stronger than they appear. I've never had one break, even in high winds that caused property damage. In this antenna, however, the radiator also serves as one of the top guys. For that reason the trap was reinforced. Two 3/16-inch thick pieces of plastic were used with three layers of glass cloth and epoxy sandwiched between them. A rotary wire brush serves well to rough up the inner surfaces of the plastic to ensure good adhesion. However, one may use coarse sandpaper for that purpose. Glass cloth and epoxy are sold as a repair kit in many hardware stores.

Before the antenna wires are connected,

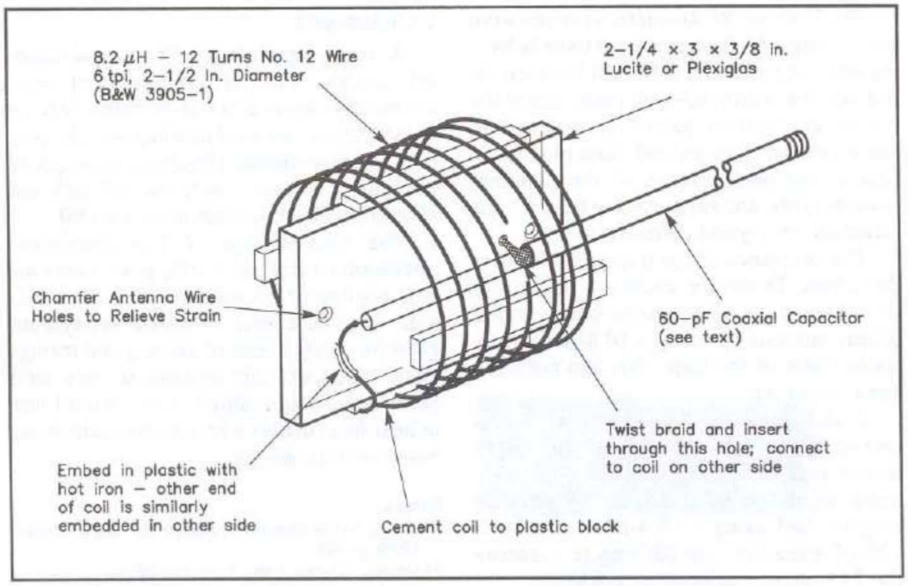

Figure 3—A simple, sturdy trap. The coaxial capacitor should be taped to the antenna wire after installation. It is not necessary to enclose the trap.

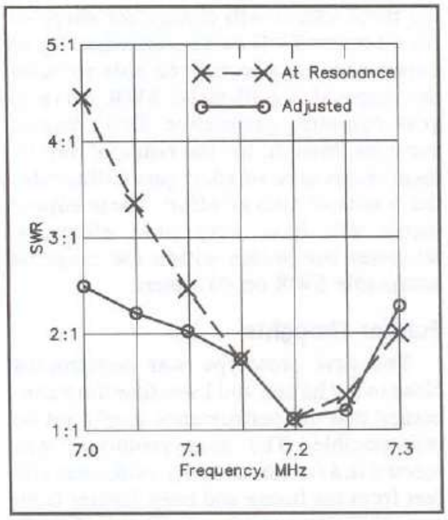

Figure 4—SWR curve for 40 meters, showing the effect of tapping L and adjusting feed-line length between L and the transmitter to obtain the best SWR curve. Such adjustments have little effect on 80-meter characteristics.

the trap must be tuned to resonance. A dip meter or noise bridge can be used to measure the resonant frequency. Start with about 30 inches of RG-8/U for the coaxial capacitor. After connecting it to the coil, as shown, 26 or 27 inches of braid will remain. At this point, the resonant frequency should be below 7.0 MHz. Trim the braid at the far end, a little at a time, snipping off the center conductor as you go. Recheck the resonant frequency each time. As 7.2 MHz is approached, continue trimming the braid, but stop cutting the center conductor. To increase the leakage path, the polyethylene dielectric should extend beyond the braid 1/8 to 3/16 inch (3.2 to 4.8 mm) when the trap is resonated at 7.2 MHz. Very close to 24 inches of braid should remain at completion. Tightly tape this end with several layers of plastic electrical tape.

The component values for this trap are exactly the same as those used in the W3DZZ trap dipole,[4] so there are several commercially made traps that may be suitable for this antenna. Traps made for a five-band, two-trap dipole, 108 feet long should have the proper values of capacitance and inductance.

In any antenna system, the radiator feed point impedance repeats itself every half wavelength along the transmission line. Inductor L must be inserted in the transmission line at a half-wave point in order to exactly cancel the capacitive reactance of this antenna system. It is, of course, advantageous to place L as close to the feed point as possible in order to minimize losses. A half wavelength at the lower frequency is as close as you can get without going to the feed point itself. The 90.2-foot length of RG-8/U shown in Figure 1 is an electrical half wavelength at 3.6 MHz for solid polyethylene dielectric coaxial cable *only*. If cable having a velocity factor other than 0.66 (e,g., foam-dielectric coaxial cable) is used, this length will have to be recalculated from the equation

$\ell = 492\ V/3.6$ feet

(Feet × 0.305 = m)

In this equation, V is the velocity factor of the cable to be used. Only RG-8/U or a similar type such as RG-213/U should be employed for this section of the transmission line in order to keep the losses low. If you don't mind a dB or so of loss on 40 meters, RG-58/U is acceptable. The loss on 80 meters will be negligible in any case.

The value of inductor L should be 1.75 μH, but I recommend that a coil having about 3-μH inductance be used to allow some latitude for final tune-up of the system. I mounted 12 turns of a no. 3018 Miniductor ($1^1/_4$ inch in diameter, 8 turns per inch) in a small Minibox with SO-239 coaxial connectors placed at each end. After tapping the coil for the best SWR curve on 40 meters, the entire assembly was sealed and waterproofed with bathtub caulk.

Tune-up

As seems to be characteristic of half slopers, this antenna can be very touchy to tune up. If the length of transmission line between the radiator and L is not an exact integral multiple of a half wavelength at 3.6 MHz, tune-up can be a real "can of worms." Since the oft-quoted value of 0.66 for the velocity factor of standard RG-8/U is only a nominal value and can vary appreciably from brand to brand (in cheap cable from lot to lot), this length should be determined with a noise bridge.

If a noise bridge is not available, the following procedure may be tried. Cut this section of cable about 6 feet shorter than the calculated length. Prepare a section of RG-58/U, 12 feet long, with solderless connectors on each end. Connect it to both the shortened feed line, using a PL-258 double female connector, and to L. Tap L to obtain the best combination of SWRs at 3.6 and 7.2 MHz. Record the SWR figures and tap position.

Now shorten the RG-58/U by 6 inches and repeat — and repeat — until you are certain you have passed through the point where the SWR values simultaneously bottom out at both frequencies. Prepare a length of RG-58/U (from the same lot) exactly as long as the best experimental length, using permanent coaxial connectors. Seal and waterproof all connections. RG-58/U is recommended for relative ease of pruning. If you don't mind unsoldering a PL-259 each time, RG-8 could be used. However, the additional loss from such a short section of RG-58/U will be infinitesimal at these frequencies.

This procedure is obviously tedious, but it is necessary to obtain good performance on 40 if a noise bridge is not available.[6] In fact, some adjustment of this section of transmission line may be necessary even if a noise bridge is used to measure the electrical length to obtain optimum two-band performance. The exact half wavelength should always be used as a starting point, in any case.

Strangely enough, the above procedures are necessary only to optimize 40-meter performance. My experience has been that merely cutting the half-wavelength section of transmission line to the calculated length, then tapping L to obtain the best SWR at 3.6 MHz, is sufficient to obtain a ratio of 2:1 or less over the entire 80/75-meter band. So tune for 40, and 80 should take care of itself.

Once the antenna system has been resonated, it may pay to experiment with the value of L and the length of transmission line between L and the transmitter. Changing these values will change the shape of the 40-meter SWR curve somewhat. By so experimenting, you may be able to tailor the shape of the 40-meter SWR curve to your operating preference. Don't expect miracles, though, for the range of adjustment seems to be small. Figure 4 illustrates the results of such an effort. These adjustments will have very little effect on 80-meter bandwidth within the range of acceptable SWR on 40 meters.

Further Thoughts

The first prototype was constructed close to my house, and I was therefore concerned that the performance might not be reproducible. The next prototype was erected in a far corner of my yard, over 100 feet from the house and even further from any other structures or conductors. The final version was similarly located. Except for final tune-up parameters, all three behaved almost identically, even though a number of physical changes was made each time. As a final test of the soundness of the design, I built scaled versions for 40/20 and 20/10 meters. They exhibited very similar characteristics, although all parameters of the system seem to become very critical as the design frequency is increased.

The first two 80/40-meter systems were built using a 30-foot mast, yet their behavior was not markedly different from the final version with a 40-foot mast. Since the mast is an electrical part of the system, and since proximity to ground must play some role in the performance of the antenna, other heights, and supports that have beams attached, may yield different results.

The reactance of the trap at 3.6 MHz is 245 ohms. Therefore, there is a possibility of constructing an 80-meter-only version of this antenna by using a 10.8-µH inductor in place of the trap. This has not been tried, however.

Scaling up to 160/80 meters is an attractive possibility. Conceivably, a mast height as low as 50 feet could be used. A starting point would be the doubling of all wire lengths, and using a 16.4-µH coil with a 120-pF capacitor for the trap to preserve the 245-ohms reactance at 1.8 MHz.

There also seems to be a possibility that a slight increase in trap capacitor value, to resonate the tap at 7.1 or 7.15 MHz, might allow better coverage of 40 cw, but probably at the expense of the phone portion of the band. Such a change might have little effect on 80-meter bandwidth, but another round of cut-and-try could prove necessary.

Conclusions

At this point, this is still an experimental design. Further development may eventually allow a cut-to-formula type of construction, but until that happens, be prepared to experiment. The dimensions given in Figure 1 will put you in the ball park and should yield immediate results on 80.

The convenience of Transmatchless operation overall of 80/75, plus a reasonable portion of 40, coupled with an excellent radiated signal, is ample repayment even for many hours of cutting and trying. Once tuned up, this antenna is very well behaved and enjoyable to use. I would like to hear from others who construct antennas based on this design.

Notes

[1] Hopps, "A 75-Meter DX Antenna," *QST*, March 1979, p. 44.

[2] Atchley, "Putting the Quarter-Wave Sloper to Work on 160," *QST*, July 1979, p. 19.

[3] DeMaw, "Additional Notes on the Half Sloper," *QST*, July 1979, p. 20.

[4] *The ARRL Antenna Book*, 13th edition, 1974.

[5] Mathison, "Inexpensive Traps for Wire Antennas," *QST*, February 1977.

[6] [Editor's Note: A third alternative is to use a dip meter to determine an exact half wavelength of line. See Downs, "Measuring Transmission-Line Velocity Factor," *QST*, June 1979.)

By Deane J. Yungling, KI6O **From *QST*, April 1986**

The KI6O 160-Meter Linear-Loaded Sloper

No room for a top-band antenna? Try this one on for size!

After having good success using my linear-loaded, inverted-L antenna over the winter 1984-85 160-meter season, I decided to try linear loading on a different type of 160-meter wire antenna.[1] I have had considerable success with a quarter-wave sloper on 80 meters, but my city-sized lot isn't deep enough to accommodate a full-sized sloper for the 160-meter band. About 120 feet would be required for the sloper to be 10 feet off the ground at the low end; I have about half that distance to work with. Linear loading solved the problem.

This antenna provides an effective bandwidth of about 70 kHz with an SWR of 2:1 or less. At the design frequency of 1.840 MHz, SWR is 1.1:1 with 50-ohm feed line and no matching network or tuner.

Although the individual dimensions are not critical, the sloping wire and the ladder line must resonate at the desired frequency. If the sloping wire is less than 65 feet long, the ladder must be longer, and vice versa, If you use a different length for the sloping wire, you will need to experiment a bit to see how much to add or remove from the ladder length.

Construction

My sloper is hooked onto the tower at about the 55-ft level using a strain insulator (see Figure 1). The coaxial-cable feed fine is securely taped to one tower leg, and the shield is connected to the tower leg with a radiator hose clamp. The tower is grounded at the base with several ground rods. The center conductor of the coaxial cable is soldered to the sloper at the strain insulator and is taped for weather protection where the center conductor and shield separate.

The ladder portion of the antenna is made with the same type of wire as the sloping portion. The ladder spacers are made of 3/8-inch hardwood dowels that have small holes drilled 1 inch from each end to hold the wires. The dowels should be soaked or sprayed with a wood preservative, prior to assembly, for weather protection. Plastic spreaders could also be used, if desired. The wires are firmly tied to the dowels with waxed lacing twine or similar material where the wires pass through the holes. The dowel spacing can be adjusted, with some difficulty, after the dowels are tied.

One end of the ladder is tied to the tower with heavy monofilament fishing line, with the other end tied to a tree or whatever else is handy. The tie-off lines are fanned vertically to keep the ladder from twisting and should be tensioned enough to eliminate any sagging in the ladder. Monofilament fishing fine provides some cushioning if the supporting tree sways or if the wind is strong.

Adjustment

Initially, the ladder should be a few feet longer than shown, to allow for adjustment. The resonant frequency should be checked, and a few inches removed at a time from the tower end of the ladder until the desired frequency is achieved. At this point, the wires should be soldered together and then taped to the end dowel. It appears that there must be an BF beam, or something similar, on the tower for any quarter-wave sloper to work properly. This sloper is no different in this respect. Finally, the antenna, particularly the sloping section, should be kept as far as possible from surrounding guy wires and other objects.

This antenna requires no ground system, balun or matching network over its operational bandwidth. The linear-loaded sloper is simple to construct and easy to adjust, and its performance is superior to my linear-loaded, inverted L.

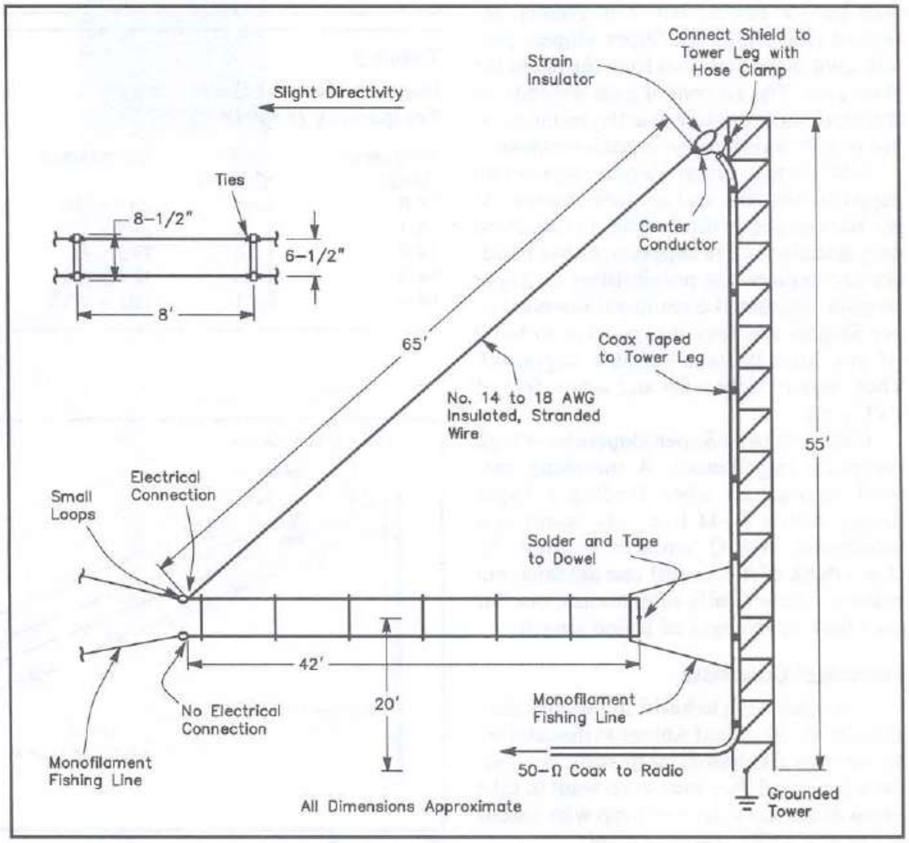

Figure 1—Construction details for the KI6O 160-meter linear-loaded sloper.

---

[1] D.J. Yungling, "The KI6O Top-Linear-Loaded 160 Meter Inverted 'L' Antenna," *CQ*, Apr 1985, pp 38-39.

By Roger Sparks, W7WKB

From *QST*, December 1995

The Super Sloper

Dramatically improve the front-to-back of a sloper.

You can imagine my surprise when *ELNEC*[1] revealed that a parasitic element can be combined with a long wire to form a whole family of directional antennas. There on my screen, I was looking at useful combinations not found in that bible of antennas, *The ARRL Antenna Book*. I had been exploring combinations of full-wave parasitic elements with wire arrays. How far off center can I slide a parasitic element, I wondered? "A lot," was the answer coming from *ELNEC* [and confirmed by *EZNEC—Ed.*]. As I offset a closely spaced (0.015 to 0.046 λ) parasitic element by 1/8 to 5/8 λ, a whole family of antennas appeared (see Figure 1).

I call the configuration a Super Sloper because the pattern resembles that of the well-known sloper, but it is greatly enhanced (see Figure 2). Super Slopers provide gain in the direction from the tall to the short pole. The amount of gain depends on antenna length, or more exactly, on the number of half wavelengths in each element.

Like slopers, Super Slopers require two supports, one tall and another shorter. At my station, and in this article, I considered only nonconductive supports. Other builders can explore the possibilities of Super Slopers suspended from metal towers. Super Slopers are very inexpensive to build (if you already have suitable supports). They require only wire and a few feet of PVC pipe.

Unlike slopers, Super Slopers have high feedpoint impedances. A matching network is required when feeding a Super Sloper with a 50-Ω line. The result is a broadband, low-Q antenna (Figure 3). Don't think of this as just one antenna, but rather a whole family of antennas, one for each half wavelength of added length.

Technical Concepts

If you just want to build an antenna, skip ahead to the Practical Antennas discussion. Those with a technical inclination can continue here, and they may even want to take a look at the discussion of long-wire antennas in *The ARRL Antenna Book*.

An antenna is called a long wire if it is one wavelength or longer. The radiation pattern can be described as the surfaces of two opposed cones that are coaxial with the wire and have their apexes meeting at the feed-point. The apex angle becomes smaller and the lobes grow stronger as antenna length increases. An azimuth plot of radiation pattern for single long-wire antennas over ground shows four principal lobes at low radiation angles. There are two principal lobes, *in one direction*, when the antenna is terminated in a matched, resistive load.

ELNEC shows that a parasitic element added to a single long-wire antenna creates a *unidirectional* radiation pattern similar to that created by a resistive termination. Energy eliminated from the back goes into useful forward gain. The parasitic element can be tuned as either a director or a reflector.[2]

Experience with Yagis (and many other applications of parasitic elements) led me to believe that all parasitic elements must be nearly a half wavelength long and located within the span of the driven element. That concept is completely *incorrect*. Parasitic elements can be *any* resonant length and offset from the driven element, as long as there is adequate coupling between the elements.

Here's how Super Slopers produce front-to-back ratio (F/B) and gain. The best

Table 1
Spacing and Phase Angle for Various Antenna Lengths

At the spacings listed, the current in the resonant parasitic element is equal to driven-element current ±10%. (Derived from *ELNEC*, under free-space conditions, 0.25-λ offset, at 7.1 MHz.)

Antenna Length (λ)	*Spacing (λ)*	*Phase Angle (degrees)*
0.5	0.046	−156
1.0	0.030	−147
1.5	0.023	−144
2.0	0.018	−142
2.5	0.015	−142
3.0	0.015	−144

Table 2
Impedance and SWR versus Frequency (Figure 5)

Frequency (MHz)	*SWR (200-Ω)*	*Impedance*
14.0	1.54	293 + *j*48
14.1	1.25	248 + *j*13
14.2	1.16	172 + *j*0
14.3	1.75	127 + *j*54
14.4	2.61	107 + *j*113

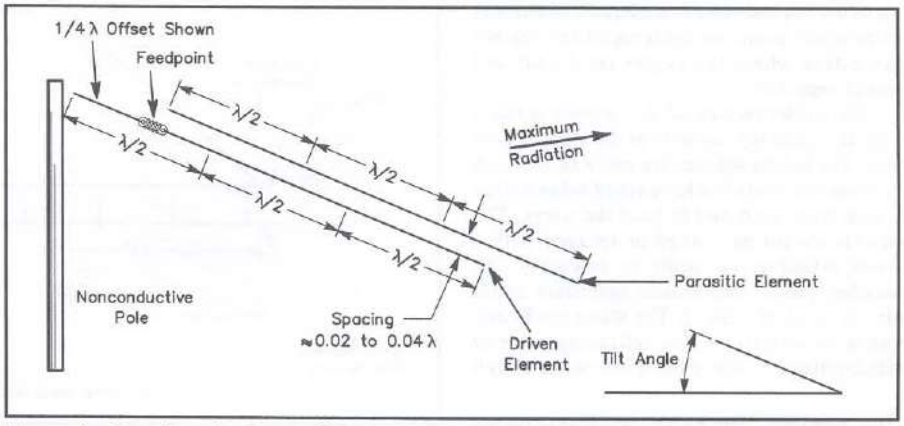

Figure 1—Drawing of a Super-Sloper antenna showing nomenclature. This is a 3-λ/2 model. See Figure 7 for construction details.

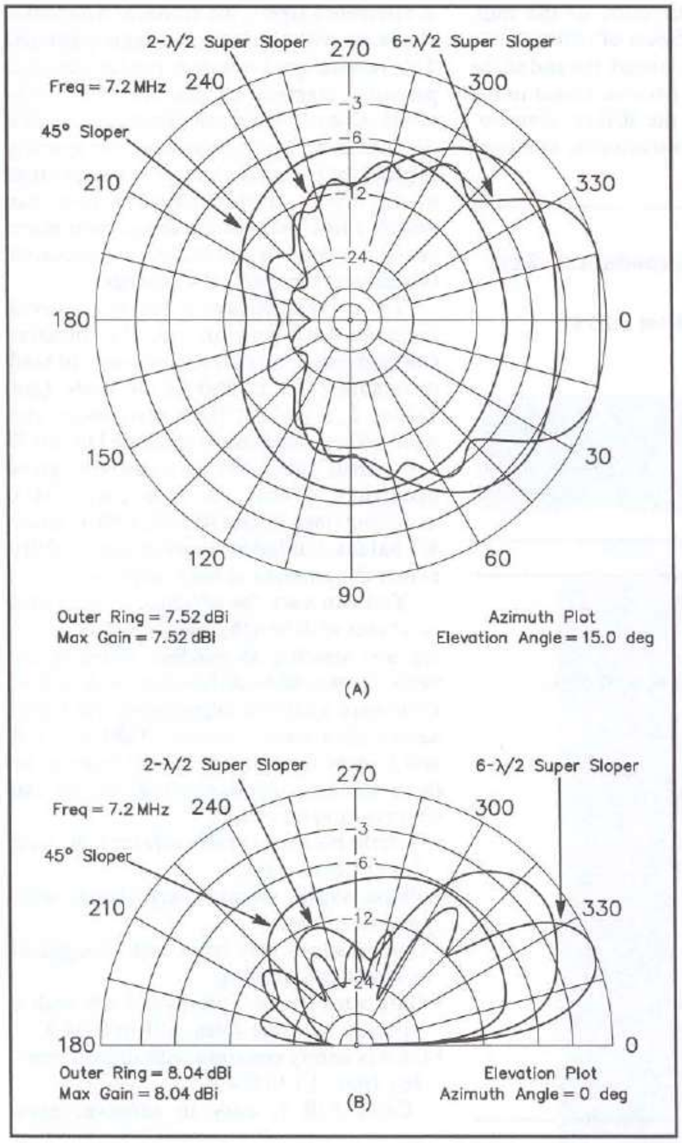

Figure 2—Azimuth (A) and elevation (B) plots for a 45° sloper, 2-λ/2 (0.25-λ offset, spaced 0.04 λ, low end 0.119 λ above ground) and 6λ/2 (0.25-λ offset, spaced 0.015 λ, low end at 0.066 λ above ground) Super Slopers. The high end of each is λ/2 high.

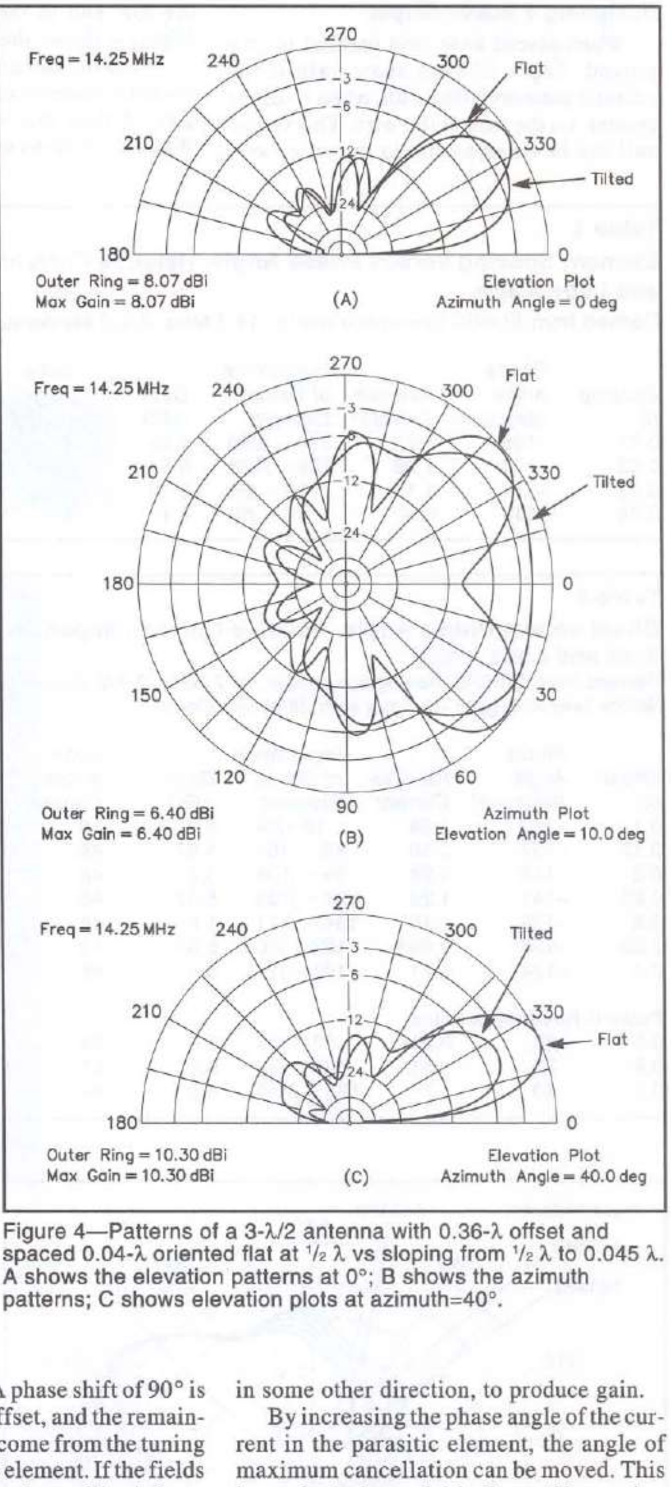

Figure 4—Patterns of a 3-λ/2 antenna with 0.36-λ offset and spaced 0.04-λ oriented flat at 1/2 λ vs sloping from 1/2 λ to 0.045 λ. A shows the elevation patterns at 0°; B shows the azimuth patterns; C shows elevation plots at azimuth=40°.

F/B results from a 180° phase shift between the currents in the driven and parasitic wires, as seen from a distant location where a pattern null is desired. A phase difference of 180° means that the signals completely cancel each other, if each element delivers equal signal strength. A phase shift of 90° is possible with a 1/4-λ offset, and the remaining 90° phase shift can come from the tuning (length) of the parasitic element. If the fields cancel in one direction, they will reinforce in some other direction, to produce gain.

By increasing the phase angle of the current in the parasitic element, the angle of maximum cancellation can be moved. This is particularly useful in Super Slopers because the principal lobes of long wires lie at some small angle to the direction of the wire. From Table 1, notice that the current in a resonant parasitic element has about –144° to –156° phasing, which is about right for correct reverse-lobe canceling.

If we want each element to deliver equal signal strengths to a distant location, nearly equal currents must flow in both the driven and parasitic elements: The coupling between elements must be very close. Table 1 suggests the approximate spacings for currents to be equal within ±10%. (The antenna also has gain at closer spacings because gain is affected less by unequal currents than is F/B.)

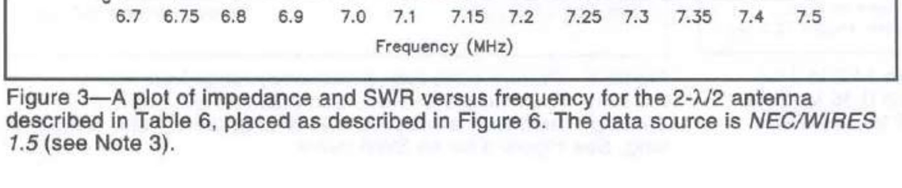

Figure 3—A plot of impedance and SWR versus frequency for the 2-λ/2 antenna described in Table 6, placed as described in Figure 6. The data source is *NEC/WIRES 1.5* (see Note 3).

Designing a Super Sloper

When placed near (and parallel to) the ground, Super Slopers show twin-lobed azimuth patterns with a null, often 10 dB or greater, on the axis of the wire. This center null can be filled by tilting the wire, with the low end in the direction of the null. Figure 4 shows the effects of tilting.

When the antenna is tilted, the end of the parasitic element can become closer to the ground than that of the driven element. When considered as two separate antennas at different heights, the driven and parasitic elements will not have the same patterns. This reduces performance, but locating the parasitic element so that its end height equals that of the driven element can solve the problem. That is, increase the spacing to place the parasitic element end above that of the driven element. Models give this design 1 to 2-dB gain advantage over more closely spaced antennas, but construction is a greater mechanical challenge.

The driven element is fed at a current antinode, 1/4 λ from an end. The antennas can be made to exhibit a wide range of feed resistances: 20 to 300 Ω, or more (see Tables 2, 3 and 4). High resistances and wide offsets combine to produce low SWR and useful gain over an unusually great bandwidth (Table 2). I've used 1/4-λ matching lines (using RG-62, a 93-Ω line), 4:1 baluns and ladder lines to successfully match the antenna at high impedances.

You can vary the antenna length (gain increases with length), height, offset, phasing and spacing as needed. These many variables would be difficult to work with if each were a critical adjustment, but fortunately, they are not critical. Tables 1, 3, 4 and 5 show the tolerant design features of these antennas. From these tables, we can observe several trends:

- Current balance greatly affects F/B, with less effect on gain.
- Phase angles change very slowly with antenna length.
- Gain changes very little with changes of current and phasing.
- Gain varies by a little over 1 dB with a spacing increase from 0.01 to 0.04 λ.
- Gain is nearly constant with offsets moving from 0.1 to 0.4 λ.

Good F/B is easy to achieve, even

Table 3

Element Spacing versus Phase Angle, Relative Current, Impedance, Gain and Lobe Angle

Derived from *ELNEC* free-space model, 14.2 MHz, 2-λ/2 elements, offset 0.25 λ

Spacing (l)	*Phase Angle (degrees)*	*Relative Current*	*Impedance of Driven Element*	*Gain (dBi)*	*Lobe Angle (degrees)*
0.01	−136	1.62	205 − j230	4.98	48
0.02	−141	1.29	128 − j138	5.54	48
0.03	−144	1.11	103 − j94	5.78	48
0.04	−149	0.97	80 − j59	6.1	48

Table 4

Offset versus Phase Angle, Relative Current, Impedance, Gain and Lobe Angle

Derived from *ELNEC*, free-space model, 14.2 MHz, 2-λ/2 elements, spaced 0.02 λ. Notice how little gain changes with different offsets.

Offset (λ)	*Phase Angle (degrees)*	*Relative Current*	*Impedance of Driven Element*	*Gain (dBi)*	*Lobe Angle (degrees)*
0.1	−167	1.08	19 −j54	5.7	48
0.15	−157	1.18	48 − j101	5.67	48
0.2	−148	1.26	89 − j135	5.6	48
0.25	−141	1.29	128 − j138	5.57	48
0.3	−139	1.16	134 − j111	5.6	48
0.35	−138	0.963	123 − j70	5.58	48
0.4	−134	0.71	108 − j14	5.4	48
Pattern Reversal Begins					
0.5	−26	0.25	83 + j46	3.0	54
0.6	39	1.31	126 − j57	5.2	61
0.7	43	1.71	192 − j180	4.0	64

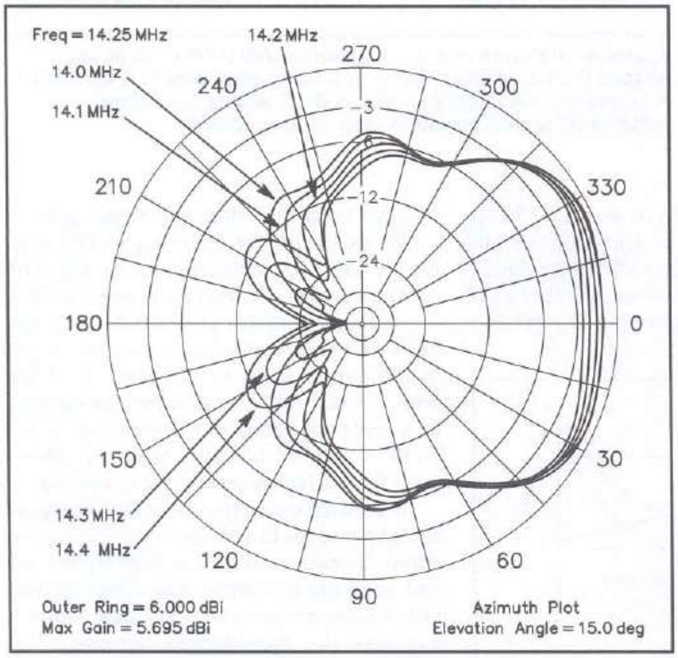

Figure 5—A radiation pattern over a range from 14.0 to 14.4 MHz. The pattern is that of a 3-λ/2 antenna with 0.36 λ offset, with the ends at 1/2 λ and 0.045 λ. See Table 2 for the SWR tabulation.

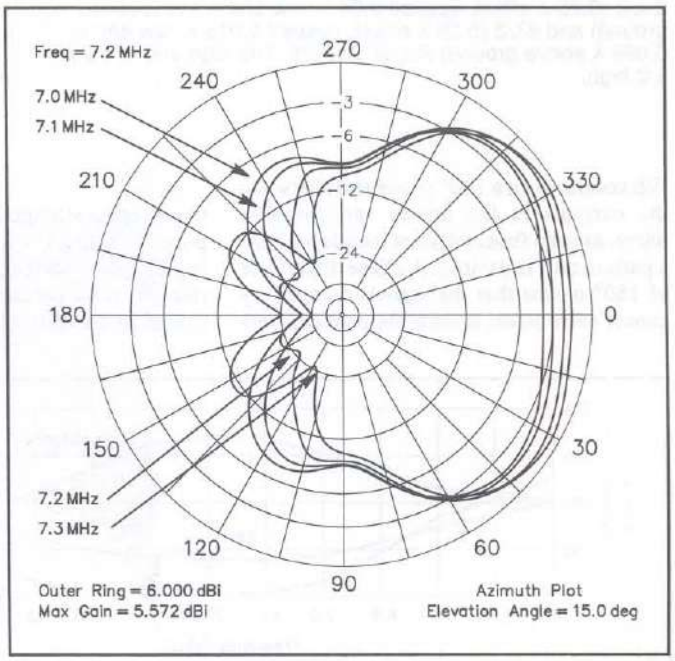

Figure 6—Pattern plots over a frequency range from 7.0 to 7.3 MHz. The antenna is 2-λ/2, with 0.25-λ offset and 0.04-λ spacing. The driven element is 1.02-λ long; the director is 0.98-λ long. See Figure 3 for an SWR curve.

though the antenna is designed for best gain or easy construction. Improved F/B ratios and reverse-oriented patterns are possible from designs optimized at specific frequencies. Computer modeling is the best way to optimize reverse patterns. Figures 5 and 6 show some possible patterns and how the patterns vary with frequency. Table 2 and Figure 3 show predicted SWR.

Practical Antennas

Table 6 describes Super Slopers constructed at my station (refer to Figure 1 for nomenclature). The principal difference between the two 3-λ/2, 20-meter antennas in Table 6 is the increased offset from 17½ feet to 22 feet. This changes the phasing, resulting in a slightly improved F/B ratio and slightly improved gain. The trade-off is a narrower pattern and a longer antenna.

Table 5
Parasitic Element Length versus Phase Angle, Relative Current, Gain and Lobe Angle

Derived from *ELNEC*, under free-space conditions with 14.2 MHz, 2-λ/2 elements, spaced 0.02 λ apart.

Percent Short (λ)	*Phase Angle (degrees)*	*Relative Current*	*Gain (dBi)*	*Lobe Angle (degrees)*
0	−140	1.30	5.54	48
1	−125	1.47	5.33	49
2	−107	1.53	5.02	51
3	−89	1.43	4.70	52
4	−75	1.23	4.44	53
5	−65	1.05	4.19	53

Table 6
Working Antenna Dimensions

Use a 4:1 step-down transformation to match 50-Ω line.

Antenna Number	*1*	*2*	*3*	*4*	*5*	*6*
Band	40 m	20 m	20 m	20 m	15 m	10m
Length*	2	3	3	2	2	2
Driven (ft)	139.3	106.5	105.6	70	46	34.5
Director (ft)	134.3	104.0	102.2	68	45	32.8
Offset (ft)†‡	34.3	17.5	22	17.5	11.5	8.6
Spacing (ft)	6	1.5	1.5	1.5	1.0	1.0
Feedpoint (ft)‡	34.3	17.5	17.25	17.5	11.5	8.6

*Length expressed as a multiple of λ/2.
†Offset is λ/4, except for antenna 3, where the offset is 0.31 λ.
‡Feedpoint and offset are both measured from the high end of the driven element.

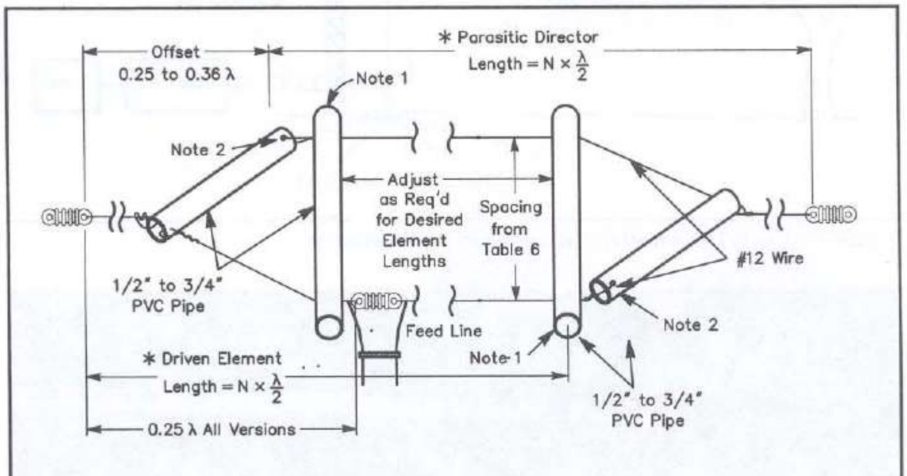

Note 1. Spreader length is 2" longer than spacing. Wires are threaded through holes drilled in spreaders.

Note 2. Tension member is 2" longer than 0.60 × spacing. Wires are threaded through holes drilled 1" from each end.

* Length must be adjusted for optimum gain or F/B ratio. Begin with lengths given in Table 6.

λ = Wavelength

Figure 7—A method of constructing the long-wire parasitic antenna. Additional center spacers reduce the tendency of the wires to twist in the wind.

A 60-foot mast supported the 3-λ/2 and 2-λ/2 20-meter versions. They were mounted back-to-back and tilted to fill the center null. Another 3-λ/2, 20-meter unit was mounted from a 38-foot mast. The 40-meter version sloped from 60 feet down to 6 feet. All antennas performed as predicted. Additional height at either end definitely increases low-angle radiation, but it reduces high-angle radiation.

Figure 7 shows how I assembled and spread the wires. The departure from straight lines (as depicted in Figure 1) has no practical effect on the antenna. Be sure to place the director over the driven element when erecting the antenna. A director placed at the side will skew the pattern, favoring the side of the director. Keep the Super-Sloper support lines tight to minimize sagging. Severe sagging leads to improper phasing and degraded results. I put a support under my 40-meter Super Sloper at midspan, but I just pull the 20-meter antenna support lines tight. The two elements have a tendency to twist and wrap together in the wind. You can prevent this by using two support ropes at the low end or by using spacers as on transmission lines. Both methods work well. Super Sloper performance suffers from excessive ground losses when the low end is at ground level. I strive for an antenna slope (tilt angle) of 10° to 20° and a minimum height of 6 feet. Higher is better.

A 200-Ω feedpoint impedance is easily transformed to 50 Ω with a 4:1 balun. Ladder line and a tuner are another option. Refer to *The ARRL Antenna Book* for other methods of transforming high feed impedances to values acceptable for modern transceivers. Do not feed the Super Sloper directly with coax unless the coax is part of an impedance-matching section.

Results and Conclusions

I've built and used six Super Slopers very successfully. The F/B is often dramatic. A difference of six S units is not uncommon when switching between two antennas built to favor opposite directions.

This family of antennas has not yet been well researched and studied. Use an antenna modeling program before building designs that are substantially different (eg, longer or made from tubing) from those described in Table 6.

Notes

[1] *ELNEC*, and its successor *EZNEC*, are computer antenna modeling programs available from Roy Lewallen, W7EL. The description files for the antenna plots shown in this article are available as a self-extracting archive file named SUPSLOPE.EXE. This file can be found on the Internet (FTP to OAK.OAKLAND.EDU, directory pub/hamradio/arrl/qst-binaries).

[2] Parasitic elements offset more than 1/2 λ from the driven element, become reflectors.

[3] *NEC/WIRES 1.5* is a computer antenna modeling program available from Brian Beezley, K6STI.

By John F. Lindholm, W1XX From *QST*, January 1983

The Inverted L Revisited

City dwellers, don't despair. Here is a *good* 160-meter antenna that should fit on your lot!

I'll never know what inspired me to make a few contacts in the ARRL 160Meter Contest. My antenna was made by tying together the open-wire feeders of my 80-meter dipole. The performance was not fantastic, but it was the first step in getting me "hooked" on the "gentlemen's band." Working three dozen European stations from a friend's house got my interest up. My friend has a good 160-meter antenna system – and *lots* of property to fit it on!

Returning to my 60 × 150-foot lot made me feel depressed.[1] I suffered all winter while listening to the others working VKs, ZLs and even JAs at daybreak. What could I do to improve *my* signal?

Many hours the following summer were spent trying to figure out how to cram a 160-meter antenna within the confines of my small lot. Space restrictions dictated that my wire be no longer than a standard 80-meter dipole. I began to consider alternative antennas.

Shunt feeding my 50-foot tower was investigated but dismissed for various reasons. This arrangement would require disconnecting the shunt feed when cranking down the tower, and all guy wires would have to be broken up into nonresonant lengths with insulators. Additionally, my cables would have to be rerouted to ground level. I then considered a full-length dipole originating in a neighbor's yard three houses to the east, crossing my property and finally terminating in the yard two houses to the west! This idea was rejected. The legal negotiation fees would have run into six figures! After much head scratching, I settled on the inverted L, an antenna made popular by the grand master of 160-meters, Stew Perry, W1BB. I credit Stew for coming to my rescue!

The Inverted L

The inverted L was selected because it requires no more space than an 80-meter dipole and I could utilize my 50-foot crank-up tower for attachment. The vertical part is 50 feet long, and the horizontal part measures 130 feet, for an overall length of 180 feet (Figure 1). This makes the antenna approximately 3/8 λ, with the horizontal part providing top-loading. W1BB advises making the vertical section as long as possible (depending on tower height), and that an overall length of 160 to 180 feet works well.

Construction

With a bow and arrow, I successfully attached the far horizontal end of the antenna to the top of a 60-foot fir tree. From there, I ran the wire back to the top of the tower, where I bracketed a 30-inch-long two-by-four with an insulator screwed in the end (Figure 2). No. 14 Copperweld wire is used for the horizontal section, and no. 10 copper wire is used for the vertical section, which is spaced about 2 feet from the tower. These two wires are soldered to-

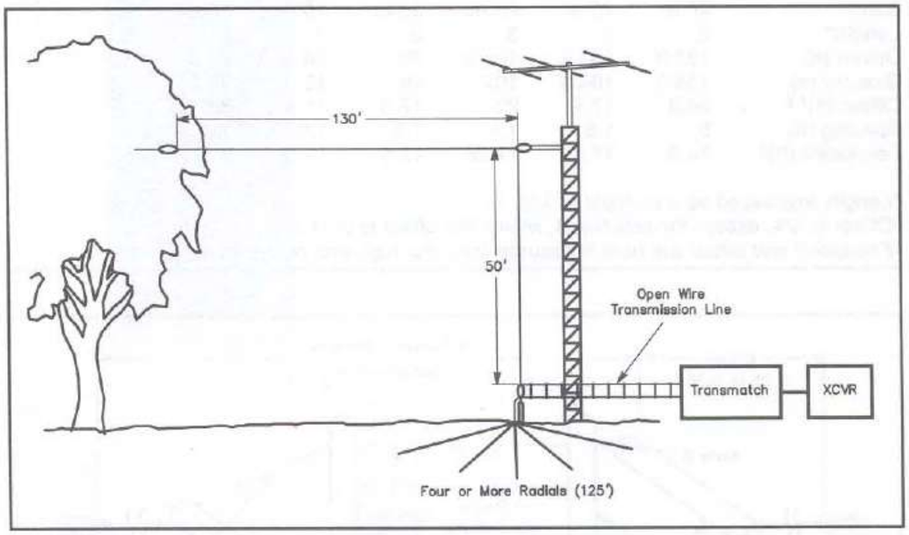

Figure 1—The W1XX inverted L is arranged in this manner.

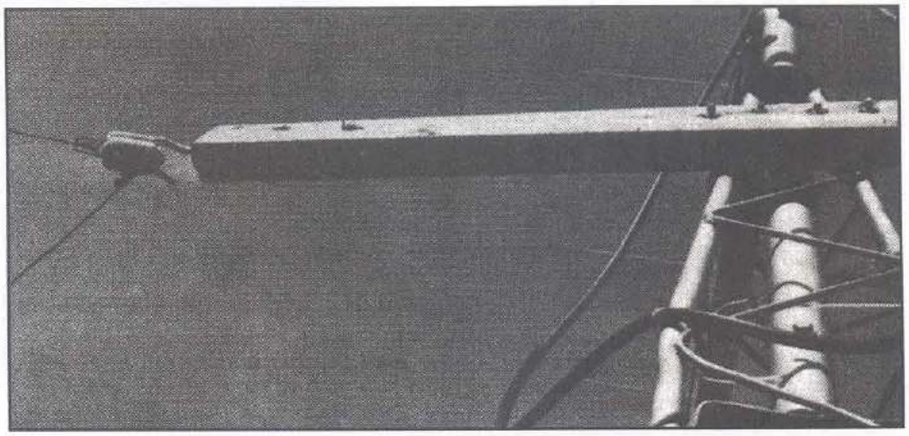

Figure 2—Detail of how the antenna is mounted to the top of the tower. TV-mast clamps are used to secure the wooden insulator.

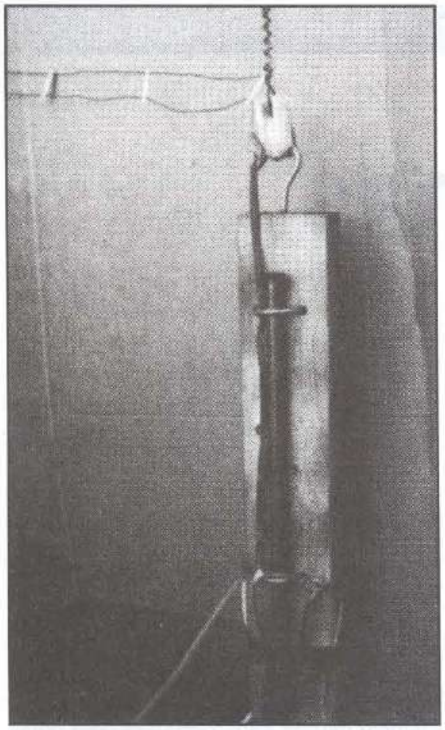

Figure 3—A ground rod at the tower base serves as a physical support for the insulator, and as an rf ground. Radial wires are connected to the rod by means of a ground bus, as described in the text.

gether at the insulator.

A plumb line was used to locate a point on the ground directly below the 90° bend in the "L." At this point, a 10-foot copper-clad ground rod was pounded into the earth, leaving 18 inches sticking out. A two-by-four is clamped to the ground rod by means of TV-mast U bolts. The lower end of the vertical wire is attached to an insulator that is screwed into the top of the wooden block (Figure 3). Next, I stripped some coaxial cable (RG-8/U) of its outer braid and used this to make a ground bus around the antenna base. One side of the open-wire transmission line is soldered to the base of the vertical antenna element, and the other side to the bus, which is attached to the ground rod with a clamp. This bus also serves as a connection point for the radials. So far, less than an afternoon of work had been invested.

Radials — the More the Better

The next day, my objective was to install radial wires, which are necessary because the inverted L is essentially a top-loaded vertical radiator. Previous meditation convinced me that several radials would fit on my lot. I'd *make* them fit! All radials were cut to 1/4 λ (125 feet), using scrap wire. About 300 feet of surplus telephone ground wire provided a good start. Stripping some old coaxial cable with a single-edged razor blade produced two radials from one length of wire (outer and center conductors). My technique for radial installation consisted of creasing the earth with a spade and shoving the wire in; afterwards the turf was pressed back into place with my heel.

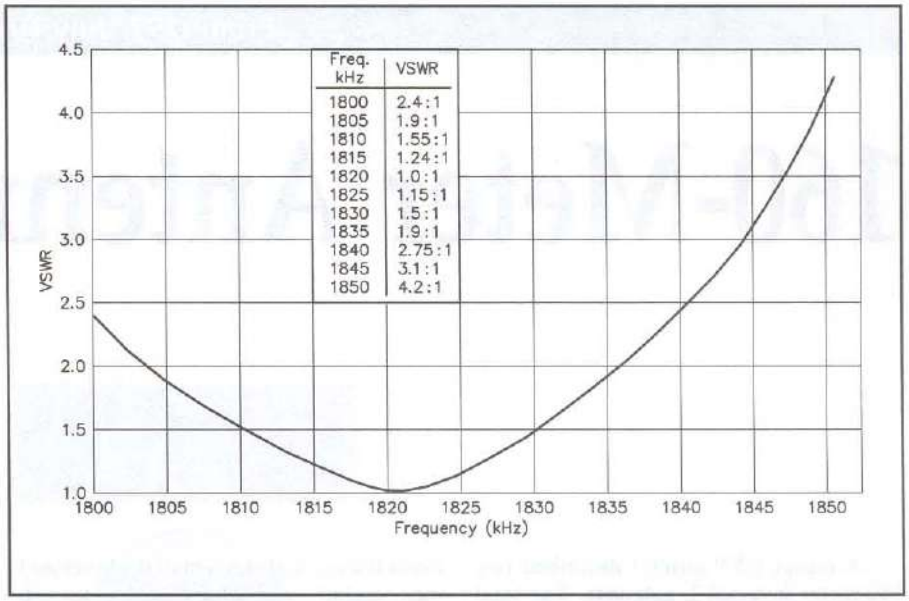

Freq. kHz	VSWR
1800	2.4:1
1805	1.9:1
1810	1.55:1
1815	1.24:1
1820	1.0:1
1825	1.15:1
1830	1.5:1
1835	1.9:1
1840	2.75:1
1845	3.1:1
1850	4.2:1

Figure 4—SWR curve for the inverted L, with the Transmatch adjusted for an SWR of 1 at 1820 kHz.

With space a problem, it may be impossible to place all your radials in a straight line; don't worry, because it is not necessary. My installation followed a zigzag path to avoid fixtures like the house and driveway. W1BB advises putting some radials under the horizontal part of the antenna. Unfortunately, the location of my garage prevented this. Initially, only four radials were planted, but more were added later. As with all vertical antennas, the more radials you can put in the ground, the better the performance!

Matching System

Voltage-fed antennas approximately $^{1}/_{2}$-λ long, such as the inverted L, will have a fairly high feed-point impedance. They will also exhibit inductive or capacitive reactance, depending on whether the antenna is slightly longer or shorter than $^{1}/_{2}$ λ. Since the inverted L does not have a 50-ohm impedance, a matching system is needed. To make tuning adjustments easier, I opted for locating the Transmatch *in* the shack, using open-wire line to the antenna. Being able to adjust the antenna match conveniently is recommended, since the SWR climbs rapidly as you shift frequency. For example, adjusting the Transmatch for a 1:1 SWR at 1820 kHz produces an SWR of 2.4:1 at 1800 kHz (see Figure 4). With the Transmatch located in the shack, you can easily adjust for a 1:1 SWR no matter where you operate in the band. My Transmatch consists of a plug-in, link coupled coil and variable capacitors – all scrounged at flea markets. Any of the configurations found in the *ARRL Antenna Book*[2] should work well.

Performance

Does the antenna work? Having no comparison antenna, my conclusions are subjective. But I've been on the air enough to know when, as they say, "it plays." With only four radials in place, my first night of operation yielded plenty of U.S. contacts, plus a Caribbean DXpedition on the first call. Subsequently, many European stations have been worked from my northeast location with good signal reports. Contest activity has yielded some respectable scores, including many QSOs with the Caribbean, and South and Central America — even Antarctica!

By adjusting my Transmatch, I made limited tests with the L on 75-meter ssb. Comparisons were made to a 75-meter dipole at 50 feet. For signals close in, the L was down by some 3 to 5 dB, but equal or superior to the dipole for signals from eastern Europe. Apparently, the 160-meter inverted L also provides a low angle of radiation on this band too. On bands higher than 75 meters, the radiation angle will be tilted *upwards*, rendering the antenna inefficient for DX work. This phenomenon is explained in *The ARRL Antenna Book.*[3]

Giving up the 160-meter band for lack of sufficient real estate is unwarranted. With the inverted L, you can work Top Band from your urban lot. Installation is a breeze, and the performance is admirable. *Now* what's your excuse for missing out on the excitement of 160 meters?

Notes

[1] m = ft × 0.3048.

[2] *The ARRL Antenna Book*, 14th ed. (Newington: ARRL, Inc., 1982), pp. 4-1 through 4-8.

[3] *Ibid.*, pp. 2-23 through 2-24.

From *QST*, July 1991 (Technical Corespondence)

160-Meter Antennas

A recent QST article[1] described two 160-meter inverted-L antennas. The total length of one of the antennas (the vertical element and the horizontal arm) is approximately 1/4 λ; the other antenna is 1/2 λ long. The current-driven 1/4-λ antenna requires a fairly extensive ground screen to realize good radiation efficiency. On the other hand, the voltage-driven 1/2-λ antenna requires only a ground stake-or no ground at all. Here's some additional information to aid you in selecting between these alternatives.

Because the current distribution of these two antennas is quite different, the radiation characteristics are also different. That explains why some signals are ". . . very loud or very weak," depending on station distance, propagation conditions and which antenna is being used.

In the captions for the figures illustrating the two antennas, it's stated that the support poles could be metal or wood. Because the vertical element of the inverted L runs parallel—and rather close to—the support pole, one may wonder whether, if the pole is metal, this proximity will influence antenna performance.

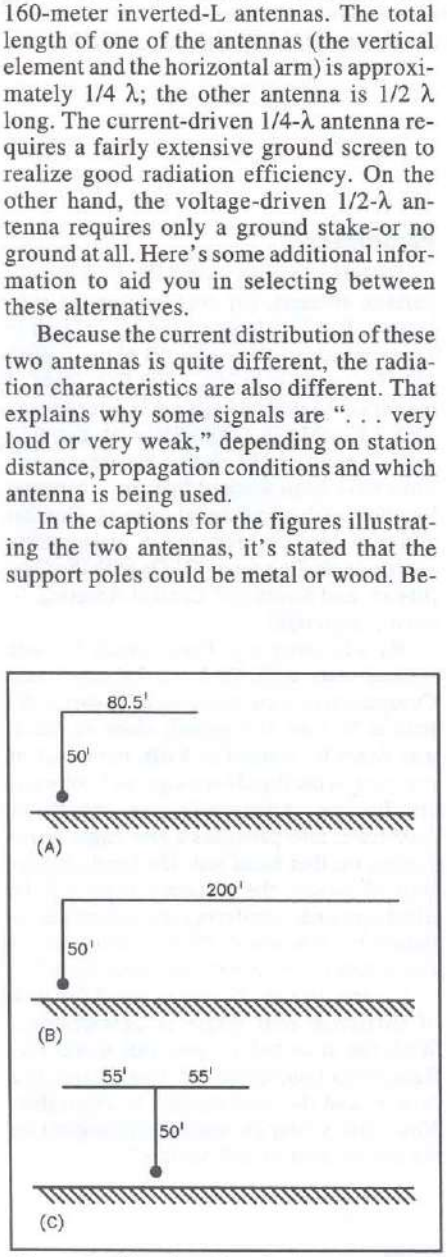

Figure 1—Sketches for simple wire antennas for 160 meters (1.9 MHz); A, 1/4-λ inverted L; B, 1/2-λ inverted L; and C, 1/4-λ T antenna.

The three antenna types I'll discuss are illustrated in Figure 1; they are (A) the 1/4-λ inverted L; (B) the 1/2-λ inverted L; and (C) the T antenna. Figures 2 and 3 show the calculated[2] radiation patterns for the two inverted-L antennas, assuming average ground ($\sigma = 3$ mS/m, $\varepsilon = 13$). The support poles are trees (or other wooden supports) 50 feet tall. Note that the horizontal arm of the antenna is close to the ground, in terms of wavelength (0.1 λ), and that, in this case, *ELNEC* overestimates the gain of horizontally polarized antennas at a height of 0.1 λ by about 3 dB.[3]

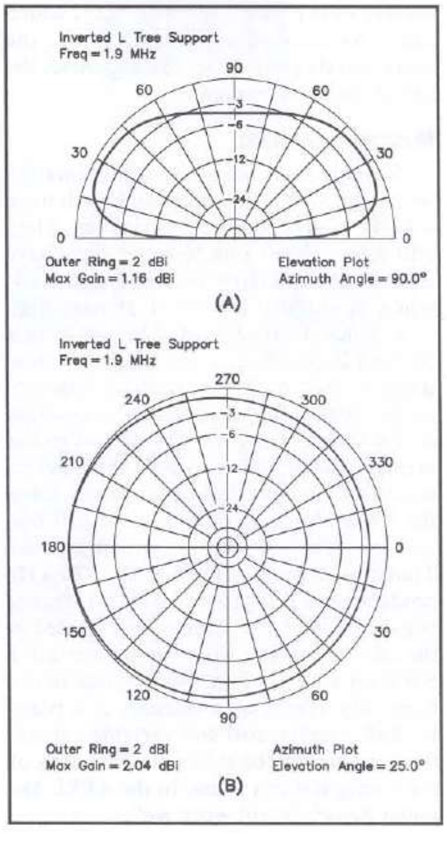

Figure 2—Radiation patterns for Figure 1A's antenna over average ground. The radiated field is dominantly vertically polarized—a monopole-like pattern. The absence of an overhead null in the vertical-plane pattern (compare Figure 2A with Figure 4A) is due to the horizontally polarized component of the field radiated by current on the antenna's horizontal arm.

The reason the patterns are so different is because the vertical antenna element carries a heavy current when the antenna length is 1/4 λ, whereas the vertical antenna element carries only a small current when the antenna length is 1/2 λ. Hence, there is a great difference between the vertically and horizontally polarized components of the radiation field for these two inverted Ls. The former antenna has a monopole-like pattern, whereas the latter antenna has a dipole-like pattern.

The lengths given in Figure 1 are those for resonance at 1.9 MHz. DeMaw's dimensions give wire lengths that are a bit too short for resonance (according to *ELNEC*). (In practice, one would trim the antenna length for resonance.)

If you want to work DX, it is advantageous to deploy an antenna that has a deep overhead null in its radiation pattern. This minimizes high-angle sky-wave signals, noise and interference, and improves the signal-to-noise ratio of distant weak signals. The T antenna exhibits this feature. Figure 1C provides dimensions for a T antenna with a resonant frequency of 1.9 MHz; the radiation patterns are given in Figure 4. This antenna's radiation field is almost entirely vertically polarized.

For these plots, the antenna was in the X-Z plane (Z is the vertical axis and 0° azimuth is in the + X direction). The 0° azimuth for the inverted Ls is the direction that the horizontal arm points (away from the feed). There is a slight azimuthal pattern asymmetry for these antennas because the antenna structure is not symmetrical

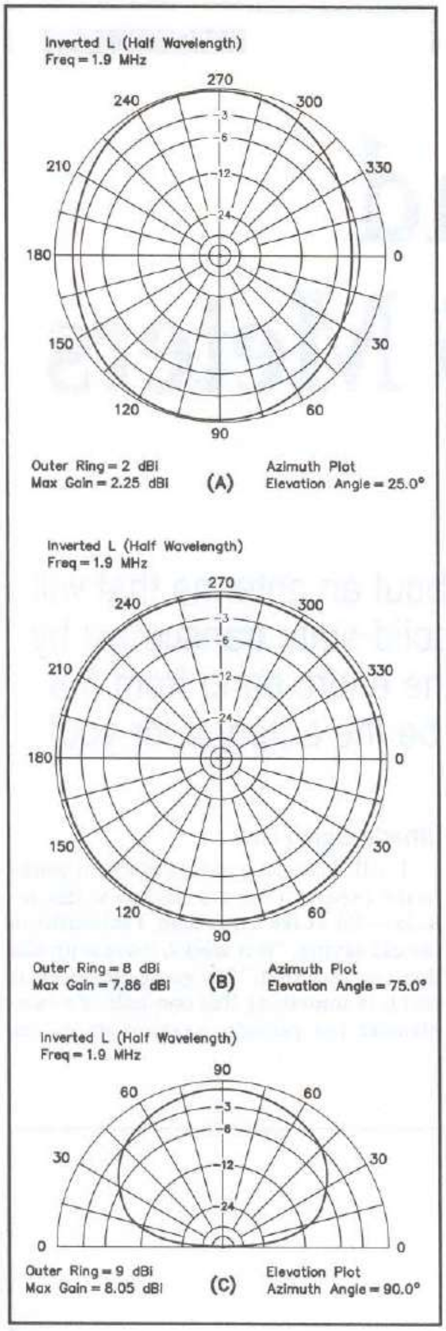

Figure 3—Radiation patterns for Figure 1B's antenna over average ground. The radiated field is dominantly horizontally polarized in the plane broadside to the antenna, and vertically polarized in the plane of the antenna—a dipole-like pattern. Hence the almost circular pattern at the high elevation angle, 75° (B), and the broadside directivity at the lower elevation angle, 25° (A).

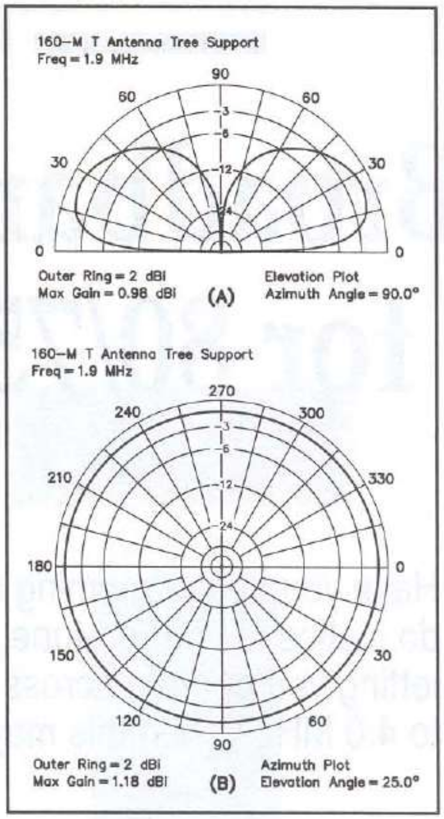

Figure 4—Radiation patterns for Figure 1C's antenna over average ground. The radiated field is almost entirely vertically polarized. The horizontally polarized component of the field radiated by current on one horizontal arm of the T is canceled by that radiated by the other arm.

with respect to the feed.

The effect of a well-grounded metal support tower was also studied. Surprisingly (to me), the effect is rather small, particularly for the $1/4\ \lambda$ inverted L and $1/4\ \lambda$ T antennas. The effect on the impedance of the $1/2$-λ inverted L looked at first to be significant, but this is only because a small change in the resonant frequency of the antenna (the metal tower couples reactance into the antenna system and so changes the system's resonant frequency) results in a significant impedance change at frequencies near the anti-resonant frequency. Only a slight retuning of the matching circuit is necessary; the radiation patterns are affected relatively little. A $6^{1}/_{2}$-foot nonconductive rope connecting the top ends of the horizontal arm to the support towers is assumed. For 50-foot support poles, it matters little (for 160-meter antennas) whether the poles are metal or wood.

But this is not the case if the support tower is also the mount for a 20-meter Yagi, because the tower and Yagi are much more nearly resonant than the tower alone. In this case, the tower carries a heavy current (0.8 A compared with the 1-A base current in the $1/4$-λ inverted L). Because the phase of the tower current is +145° with respect to the phase of the current on the vertical element of the inverted L, the antenna's vertically polarized field is almost canceled. Also, the inverted L is no longer resonant: In fact, it is far from resonant; its impedance is $1547 - j2635\ \Omega$, compared with the inverted L's impedance of $14\ \Omega$ with tree support.—*John S. Belrose, VE2CV*

Notes

[1] D. DeMaw, "The 160-Meter Antenna Dilemma," *QST*, Nov 1990, pp 30-32.

[2] *ELNEC* was used for all calculations. This program is available from Roy Lewallen, W7EL.

[3] *ELNEC* is a version of *MININEC*. *MININEC* assumes that the antenna is over a perfect earth for the purpose of calculating the antenna's impedance and the current distribution. The result of this approximation is that for horizontally polarized antennas, *MININEC* overestimates the gain of the antenna by an amount that increases as the height of the dipole decreases below a height of about $0.2\ \lambda$. The effect of real ground on the shape of the radiation pattern is correctly predicted, however, since for the calculation of the far-field pattern the program employs Fresnel reflection coefficients for the specified earth type.

By Brian L. Wermager, K0EOU **From *QST*, April 1986**

A Truly Broadband Antenna for 80/75 Meters

Have you been dreaming about an antenna that will do justice to your no-tune, solid-state transceiver by letting you operate across the entire band from 3.5 to 4.0 MHz? Then this may be the antenna for you!

With declining sunspots and poor conditions on the higher HF bands, 80 meters has suddenly become very popular. But, unfortunately, many hams are not able to use this band to its full potential. It offers every kind of ham activity from CW to phone, from nets and ragchewing to great DXing, but many hams are too limited by the frequency range of their antennas to enjoy this band completely.

Antenna-matching networks are one answer, but they spoil the advantage of the no-tune feature of modern transceivers. Matching networks also are often less effective than many hams think; they introduce losses. The losses can be significant at some settings which provide a match. With these things in mind, I decided to try some ideas that might give me a more broadbanded antenna. The prime requirement was that it be fed with common 50-ohm coaxial cable, with no traps, coils or capacitors.

First, I tried a quarter-wave sloper. This antenna worked very well, with a bandwidth of 300 kHz between the 2:1 SWR points. It still, however, limited me from operating CW DX at the bottom of the band and the phone nets at the top of the band. There had to be a better antenna.

Antennas can be broadbanded by using large-diameter elements. With this in mind, I began experimenting with two-wire slopers, attached to a common feed point, but with the wire ends fanned out from each other. (See Figure 1.) This seemed to help, but not as much as I had hoped. It did, however, shorten the length required for the sloper. For those with a short tower, this idea could make an 80 or 160-meter sloper possible when a single-wire sloper would be too long.

The Fickle Finger of Fate Strikes!

While I was trying one of these two-wire antennas at a low height on my tower, the SWR was less than 2:1 from 3.5 to 4.0 MHz! After several attempts to get it to work the same way at the top of the tower, I discovered that these results could be attained only when my old quarter-wave sloper was at the top of the tower *and grounded* to the tower. (See Figure 2.) The two antenna elements were obviously interacting with each other, broadening the bandwidth tremendously. Further pruning of the lengths of both the sloper and the two-wire element resulted in the amazing SWR curves shown in Figure 3.

What's Going On?

I will leave the question of *why* it works to the experts. (See the sidebar to this article.—Ed.) Like a true ham, I subscribe to the old saying, "If it works, leave it up and don't mess with it." My guess, however, is that it is something like one-half of a two-element log periodic seeking its mirror

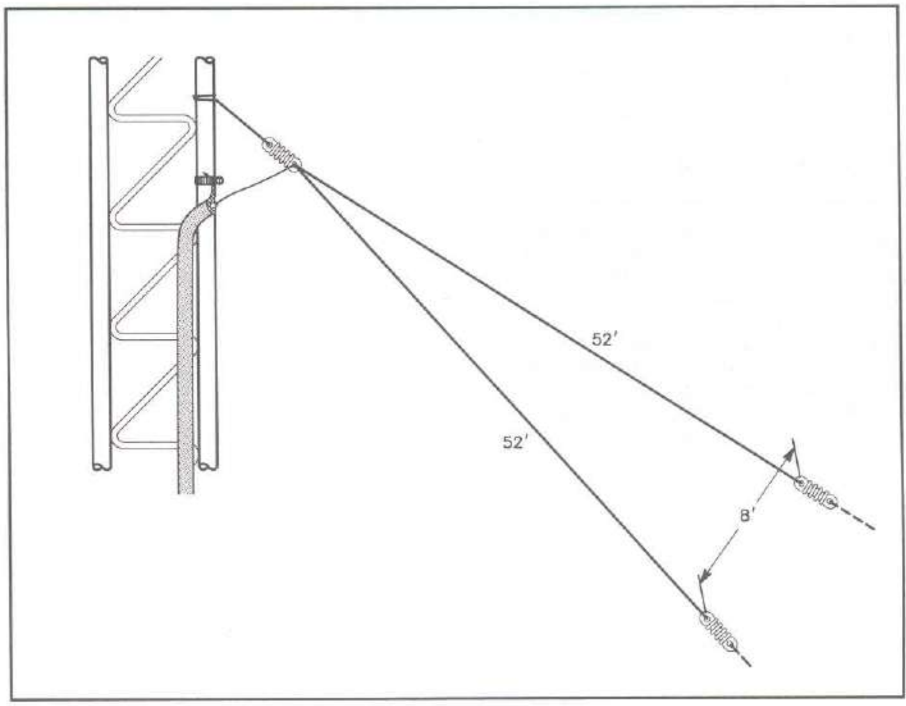

Figure 1—Arrangement of the original K0EOU experiment. The coaxial cable shield is connected to the tower, with both wires of the two-wire element connected to the center conductor. The wires are spread approximately 8 feet at the ends and are each approximately 52 feet long.

The MININEC Analysis of the KØEOU Three-Element Half Sloper

The Mini-Numerical Electronics Code (MININEC) analyzes thin-wire antennas, solving an integral equation representation of the electric fields using a method-of-moments technique. MININEC solves for the currents, impedance and patterns for antennas composed of wires in arbitrary orientations in free space and over perfectly conducting ground.

The impedance at the feed point calculates to be 79.4 + *j*1859.8 ohms. This assumes a perfect ground beneath the structure and simply a 3-foot extension of the tower above the connection of the upper wire. In practice, the impedance will be affected by both the ground conductivity and the top-loading effect of a beam antenna atop the tower.

A fair amount of current flows in the top section above the upper wire connection point. Top loading will affect the phase of this current, which will be reflected as a change in impedance at the feed point. In other words, the calculated data is not absolute. Use it as an approximation only.

A relatively high current flows at the base of the tower to ground—more than in any other part of the system. This indicates that a good earth connection, and even a radial system, would offer highest efficiency.

The antenna patterns, Figures A, B and C, are also approximations. Polarization is predominantly vertical–at low angles it may be considered to be almost completely vertical. Broadside to the direction of the wires, the polarization becomes horizontal at high radiation angles, ie, above 75 degrees. At 60-degrees elevation, the vertical component is almost 16 dB greater than the horizontal. The vertical component increases significantly at lower elevation angles, being in excess of 30 dB above the horizontal component at a 5-degree elevation. These figures all apply in a direction broadside to the wires. In the direction of the wires, both "front" and "back," the radiation is entirely vertically polarized.–*Gerald L. Hall, K1TD*

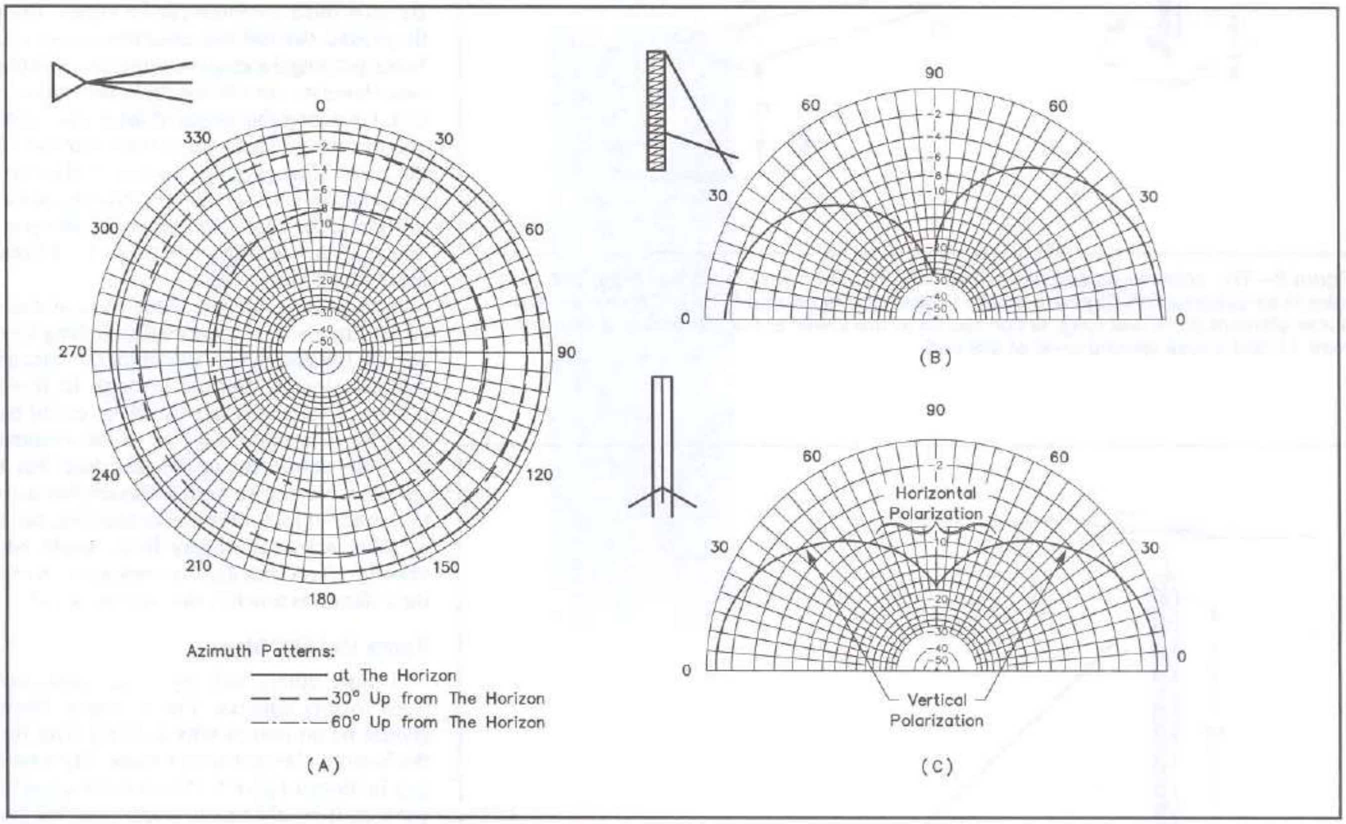

Figure A–Antenna azimuth radiation pattern for the KØEOU three-element half sloper antenna. Values are in dBi. Add 6.0 dB to the values shown.

Figure B—Antenna elevation radiation pattern, in the direction of the wires, for the KØEOU antenna. Values are in dBi. Add 6.0 dB to the values shown.

Figure C—Antenna elevation radiation pattern, in a direction broadside to the wires, for the KØEOU antenna. Values are in dBi. Add 6.0 dB to the values shown.

image in the grounded tower. The top element is tuned for the lower portion of the band and the two-wire element for the upper portion. In fact, there is a little SWR "bump" in the middle of the band that seems to give further evidence of this.

How Well Does It Work?

Although I have no way of scientifically plotting the antenna pattern, it does seem to be vertically polarized. Good DX performance from the antenna seems to verify this. Because many contacts have been made in all directions, the antenna probably has a fairly omnidirectional pattern. On-the-air comparisons with a quarter-wave sloper across town show that the antenna performs at least as well as the sloper. It also seems to have a little less noise on receive than the sloper.

Getting One Up for Yourself

If you have a tower over 40 feet high, you should be in business. The element dimensions will vary according to the height of your tower. My friend Kelly Davis, KD7XY, constructed one of these antennas on his 50-foot tower so we could see how the dimensions would change.

Measurements of his antenna are shown in Figure 4.

It is interesting to note that the height of the feed point on the tower does not appear to be critical at all. The angle of the wires in the two-wire element does not seem to be critical either. The sloper element, however, should come down *between* the two wires of the lower element. The sloper angle is about 45 degrees. As you attempt to get the

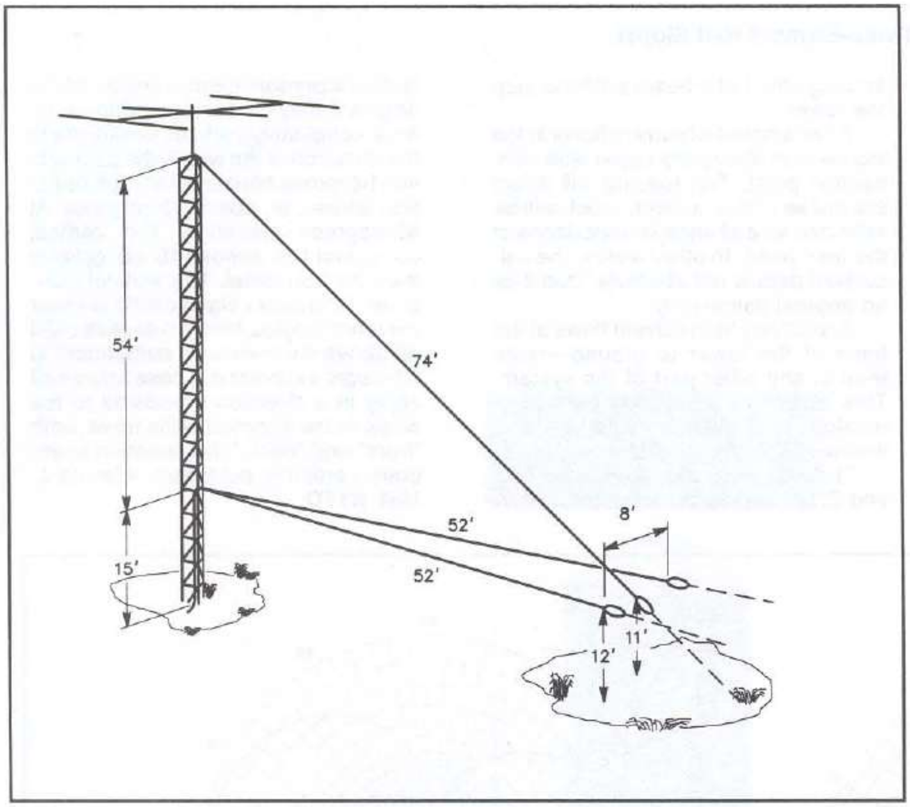

Figure 2—The antenna as constructed at K0EOU. The tower is 70 feet high. The feed point is as described in Figure 1 and is 15 feet above ground level on the tower. The sloper element is 74 feet long, is connected to the tower at the top end and slopes to a point 11 feet above ground level at the end.

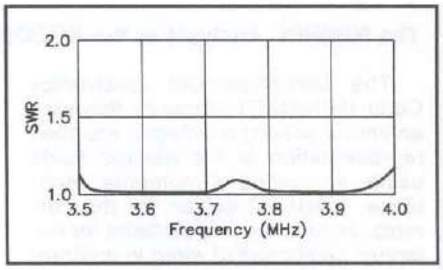

Figure 3—SWR measurements for the antenna at K0EOU. The highest SWR measurement between 3.5 and 4.0 MHz is 1.2:1.

lowest possible SWR from the antenna, remember that the angle of the sloper to the tower and its distance from the ground at the end will have an effect on the bandwidth. Because these antennas can be pruned from the ground, the trial-and-error method is easy. When pruning the antenna, remember that the two elements are cut for different frequencies. Changing the length of the top element changes the performance at the lower part of the band. Changing the lengths of the two-wire elements changes the performance at the top part of the band. Don't give up until your antenna SWR is, at the very most, 1.5:1 from band edge to band edge.

Use good insulators at the ends of each of the three wires. I used nylon fishing line at first, but one foggy, wet night the wire in the top element burned through in three places. It surprised me that there could be such high currents in a part of the antenna not even connected to the feed line, but I should have known better. Remember also that other objects around the antenna, such as other antennas or guy lines, could adversely affect antenna performance. Keep the antenna as much in the clear as possible.

Some Untried Ideas

I hope others will try some modifications to this antenna. For example, there should be no reason why a single wire for the bottom element won't work. My small city lot doesn't give me room to try a single wire, as it would surely need to be longer than the two-wire element. I would also like to see someone try cutting the top element for the high end of the band, and the bottom element for the low end. This could shorten the sloper element for someone with a shorter tower and may even give the antenna some gain in the direction of the two-wire element. It might also be possible to construct an antenna for another band inside of this one (40 meters, for example). The same feed point could be used, but with another sloper element for the second band. Another possibility is to construct the antenna with two dipoles. It is my guess (and only a guess), that it is the merging of the ends of the elements that causes the 50-ohm impedance of the antenna.

Winter gives us the best conditions on the 80/75-meter band. You can be ready to use the whole band with this simple-to-construct and very broadband antenna.

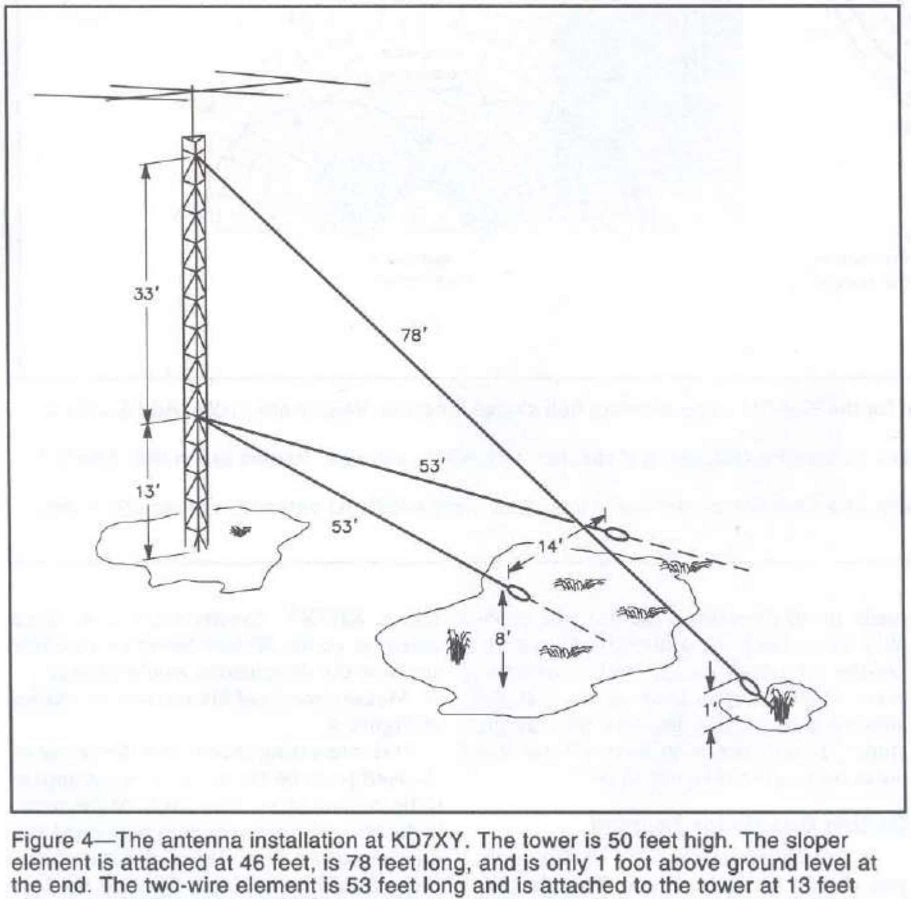

Figure 4—The antenna installation at KD7XY. The tower is 50 feet high. The sloper element is attached at 46 feet, is 78 feet long, and is only 1 foot above ground level at the end. The two-wire element is 53 feet long and is attached to the tower at 13 feet above ground level. Like K0EOU's antenna, it is virtually "flat" across the band; highest measured SWR is 1.4:1.

From *QST*, August 1987 (Technical Corespondence)

Improved Broadband Antenna Efficiency

When Brian Wermager's article. "A Truly Broadband Antenna for 80/75 Meters," arrived at HQ, I wanted to see if the antenna worked as well as claimed before publishing the information in *QST*.[1] I stopped tuning the antenna when the SWR was less than 1.6 across the entire band.

For working DX the antenna seemed to work as well as, or better than, a dipole at 50 feet. Casual contacts were not difficult to make, but in contests, it was a different matter. The results were satisfactory, but could "satisfactory" be changed to "outstanding"?

The key to improving the efficiency of this antenna is found in the sidebar accompanying the article. The MiniNumerical Electronics Code (MININEC) computer analysis done by Gerald Hall, K1TD, shows that: "A relatively high current flows at the base of the tower to ground—more than in any other part of the system. This indicates that a good earth connection, and even a radial system, would offer highest efficiency."

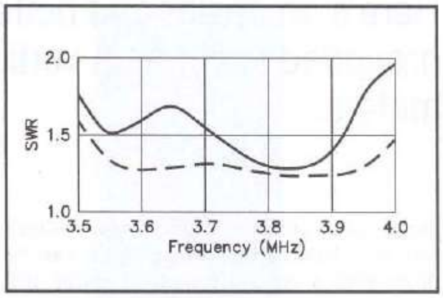

Figure 1–Graph showing SWR vs frequency for the K0EOU 80/75-meter antenna. The lower curve depicts results before adding ground radials; the upper curve was plotted after adding radials and retuning the antenna.

A system of radial wires improves antenna efficiency by cutting ground resistance losses. The decreased resistance in ground losses will also narrow the antenna bandwidth. The question now becomes: Is the narrowing of the bandwidth too much? I decided to investigate that question next.

With help from my brother-in-law, I installed 42 radials around, and bonded to, the tower base. As expected, SWR increased across the entire band. After I retuned the antenna, the SWR was below 2:1 across the entire band (see Figure 1). As for results, DX stations are typically 2 to 6 S units stronger on the K0EOU antenna than on the dipole at 50 feet. Contest QSOs are now a reality. I am not the loudest East Coast station on the band, but I am very pleased with the performance of the broadband antenna. (The dipole has been removed and stored!)—*Chuck Hutchinson, K8CH*

---

[1] B. Wermager. "A Truly Broadband Antenna for 80/75 Meters," *QST*, Apr 1988, pp 23-25.

By Thomas Russell, N4KG **From *QST*, June 1994**

Simple, Effective, Elevated Ground-Plane Antennas

Here's an easier and better way to use your grounded tower as a vertical antenna on 160 or 80 meters.

This article describes a simple and effective means of using a grounded tower, with or without top-mounted antennas, as an elevated ground-plane antenna for 80 and 160 meters.

Grounded towers have been used as shunt-fed verticals on the low-frequency amateur bands for many years. Generally. they required a gamma- or omega-type matching network and an extensive radial system for efficient operation. Recent computer studies reveal that simple elevated radial systems consisting of only four wires can produce results equivalent to 120 buried radials. Typically, these antennas are modeled as isolated monopoles. Presumably, grounded towers could be used with an appropriate shunt-fed matching network.

I've found an even easier method!

From Sloper to Vertical

Recall the quarter-wave length sloper, also known as the half-sloper. It consists of an isolated quarter wavelength of wire, sloping from an elevated feedpoint on a grounded tower. Best results were usually obtained when the feedpoint was somewhere below a top-mounted Yagi antenna. You feed a sloper by attaching the center conductor of a coaxial cable to the wire and the braid of the cable to the tower leg. Now, imagine four (or more) slopers, but instead of feeding each individually, connect them together to the center conductor of a single feed line. *Voila*! Instant elevated ground plane.

Now, all you need to do is determine how to tune the antenna to resonance. With no antennas on the top of the tower, the tower can be thought of as a fat conductor and should be approximately 4% shorter than a quarter wavelength in free space. Calculate this length and attach four insulated quarter-wavelength radials at this distance from the top of the tower. For 80 meters, a feedpoint 65 feet below the top of an unloaded tower is called for, The tower guys must be broken up with insulators for all such installations. For 160 meters, 130 feet of tower above the feedpoint is needed.

That's a lot of tower to dedicate to a single-band antenna, especially for someone with limited real estate. What can be done with a typical grounded-tower-and Yagi installation?

A top-mounted Yagi acts as a large capacitance hat, top loading the tower. Fortunately, top loading is the most efficient means of loading a vertical antenna. The amount of loading can be approximated by using an empirical formula developed by John Devoldere, ON4UN.[1]

Devoldere found that the electrical height of a top-loaded tower can be approximated by:

$$L=0.38F(H+\sqrt{2S-H/500}) \qquad \text{(Eq 1)}$$

where

L is the approximate electrical length in degrees

F is the frequency in MHz

H is the height of the tower under the Yagi in feet

S is the area of the Yugi in square feet

Table 1

Effective Loading of Common Yagi Antennas

Antenna	*Boom Length (feet)*	*S (area, ft²)*	*Equivalent Loading (feet)*
3L 20	24	768	39
51 15	26	624	35
4L 15	20	480	31
31 15	16	384	28
5L 10	24	384	28
4L 10	18	288	24
3L 10	12	192	20
TH7	24	—	40 (estimated)
TH3	14	—	27 (estimated)

To check Eq 1, consider the case of no antenna on top, where S = 0. Then L = 0.38 × 3.6 × 65 = 88.9°, which is very close to the desired 90° quarter wavelength.

The effective loading of a Yagi is the portion of the equation under the radical. The examples in Table 1 should give us an idea of how much top loading might be expected from typical amateur antennas. The term H/500 is ignored as insignificant compared with 2S.

The values listed in the Equivalent Loading column of Table 1 tell us the approximate vertical height replaced by the antennas listed in a top-loaded vertical antenna. To arrive at the remaining amount of tower needed for resonance, subtract these numbers from the nonloaded tower height needed for resonance. Note that for all but the 10-meter antennas, the equivalent loading equals or exceeds a quarter wavelength on 40 meters. For typical HF Yagis, this method is best used only on 80 and 160 meters.

Construction Examples

Consider this example: A TH7 Yagi mounted on a 40-foot tower. The TH7 has approximately the same overall dimensions as a full-sized 3-element 20-meter beam, but has more interlaced elements. I estimate its equivalent loading to be 40 feet. At 3.6 MHz, 65 feet of tower is needed without loading. Subtracting 40 feet of equivalent loading, the feedpoint should be 25 feet below the TH7 antenna.

I ran 10 quarter-wavelength (65-foot) radials from a nylon rope tied between tower legs at the 15-foot level, to various

---

[1] J. Devoldere, *Antennas and Techniques for Low Band DXing* (Newington: ARRL, 1994).

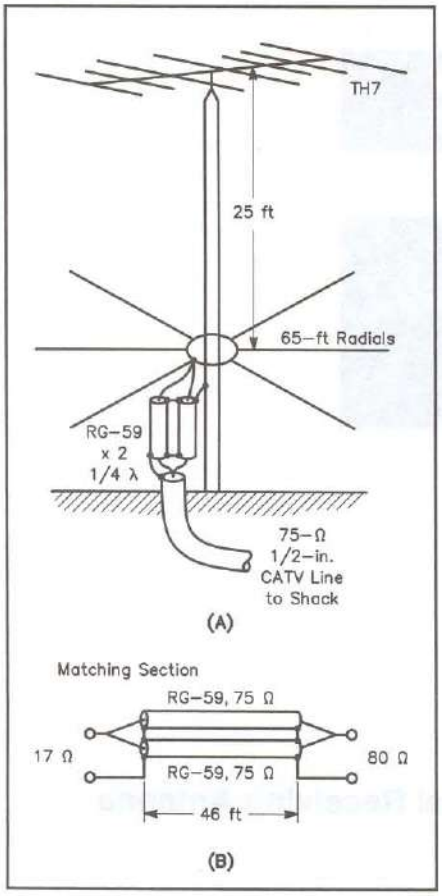

Figure 1—At A, an 80-meter top-loaded, reverse-fed elevated ground plane, using a 40-foot tower carrying a TH7 triband Yagi antenna. At B, dimensions of the 3.6 MHz matching network, made from RG-59.

supports 10 feet high. I tied nylon cord to the insulated, stranded, 18-gauge wire, without using insulators. The radials are all connected together and to the center of an exact half wavelength (at 3.6 MHz) of RG-213 coax, which will repeat the antenna feed impedance at the other end. Figure 1 is a drawing of the installation. I used a Hewlett-Packard low-frequency impedance analyzer to measure the input impedance across the 80-meter band.

An exact resonance (zero reactance) was seen at 3.6 MHz just as predicted. The radiation resistance was found to be 17 Ω. The next question is how to feed and match the antenna.

My approach to 80-meter antennas is to tune them to the low end of the band, use a low-loss transmission line, and switch an antenna tuner in line for operation in the higher portions of the band. With a 50-Ω line, the 17-Ω radiation resistance represents a 3:1 SWR, meaning that an antenna tuner should be in-line for all frequencies. For short runs, it would be permissible to use RG-8 or RG-213 directly to the tuner. Since I have a plentiful supply of low-loss 75-Ω CATV rigid coax, I took another approach.

I made a quarter-wave (70 feet × 0.66 velocity factor = 46 foot) 37-Ω matching line by paralleling two pieces of RG-59 and connecting them between the feedpoint and a run of the rigid coax to the transmitter. The magic of quarter-wave matching transformers is that the input impedance (R_i) and output impedance (R_o) are related by

$$Z_o^2 = R_i \times R_o \qquad \text{(Eq 2)}$$

For $R_i = 17\ \Omega$ and $Z_o = 37\ \Omega$, $R_o = 80\ \Omega$, an almost perfect match for the 75-Ω CATV coax. The resulting 1.6:1 SWR at the transmitter is good enough for CW operation without a tuner.

The Proof is In the Log

How effective is this antenna? Well, I used to install a 60-foot aluminum tower and 100 radials, 100 feet long in a clear one-acre field every winter, and remove it every spring for mowing. The top-loaded reverse-fed elevated ground-plane antenna has replaced that antenna with no regrets. My only other 80-meter antenna is a dipole at 110 feet broadside to Europe and the South Pacific,

I use the elevated ground plane for South Africa, South America. the Caribbean, parts of the Pacific and Asia, both long and short path. With it I have worked everything I can hear, with 1200 W output, including HL (rare in Alabama); HS; UA9; UA0; UI; UJ; UL; most of the VK9s; XV; ZS1; ZS8MI; ZS9; 3Y5X; 8Q; 9M2; and 9V1. While running 5 W output. I have even worked two JAs with this antenna. which may say more for its effectiveness than anything else.

Will It Work on 160 Meters?

You bet it will, but it takes another tower. For the 160-meter band, a resonant quarter-wavelength requires 130 feet of tower above the radials. That's a pretty tall order. Subtracting 40 feet of top loading for a 3-element 20-meter or TH7 antenna brings us to a more reasonable 90 feet above the radials. Additional top loading in the form of more antennas will reduce that even more.

Recently, a friend moved to the country, and he needed a 160-meter antenna in a hurry for an upcoming contest. He had stacked TH6s on a 75-foot tower. I suggested he try four elevated radials at 10 feet above ground, with a tuner if necessary. He connected four radials about 120 feet long and a piece of RG-58. The SWR measured under 2:1 and he worked everything he heard in the contest. Figure 2 is a drawing of this installation.

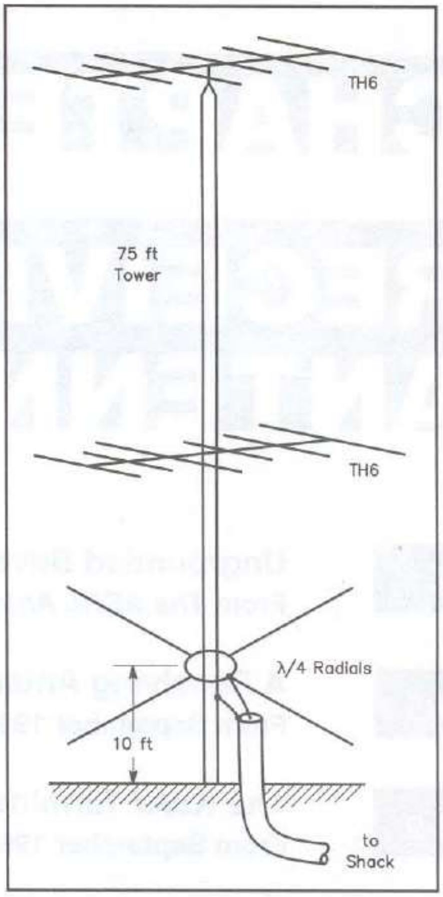

Figure 2—A 160-meter antenna using a 75-foot tower carrying stacked triband Yagis.

Another friend had a 120-foot tower with no antennas on it. He ran four elevated radials at 10 foet and obtained an SWR below 1.5:1 with a 50-Ω feed line. During the contest, he even beat out some big guns.

Elevated ground-plane antennas work! This simple, reverse-feed system makes it possible to feed grounded towers easily and efficiently.

References

A. Christman, "Elevated Vertical Antenna Systems," *QST*, Aug 1988, pp 35-42 (Feedback, *QST*, Oct 88, p 44).

A. Christman, "More on Elevated Radials," *QST*, Mar 1993, p 72 (Technical Correspondence).

CHAPTER EIGHT

RECEIVING ANTENNAS

8-1 **Ungrounded Beverage Antennas**
From *The ARRL Antenna Compendium Volume 4*

8-5 **A Receiving Antenna that Rejects Local Noise**
From September 1995 *QST*

8-9 **The K9AY Terminal Loop–A Compact, Directional Receiving Antenna**
From September 1997 *QST*

8-13 **Hum Problems When Switching the K9AY Loops**
From May 1998 *QST* (Technical Correspondence)

Ungrounded Beverage Antennas

So you've heard how Beverages can help you hear better on the lower bands, but you're still not convinced they're worth the bother. K6STI analyzes the Beverage, especially an easy-to-construct "ungrounded" version.

Beverages are amazing receiving antennas. A good Beverage can turn a signal barely audible on your transmit antenna into solid, Q5 copy. A Beverage antenna (named after 1920s inventor H. H. Beverage) is a long, low wire used for receiving. They're most popular on the 160, 80, and 40-meter bands. Although occasionally pressed into service, Beverages are much too lossy for regular use as transmitting antennas.

Conventional Beverages are fed against ground. Although very long wires may develop a few dB of front-to-back ratio, unterminated Beverages are essentially bidirectional. Terminating a Beverage through a resistance to ground at the far end creates a highly unidirectional pattern. Beverages typically are 1 to 15 feet high and 1 to 4 wavelengths long.

Beverages are broadband, traveling-wave antennas. Power traveling toward the feedpoint is absorbed by the receiver, while that traveling in the opposite direction is absorbed by the termination resistance. Standing waves can't develop. Because waves traveling toward the feedpoint accumulate substantially in phase, while those arriving from other directions tend to cancel, the azimuth pattern of a long Beverage is highly directive. Beverages are nonresonant antennas and can be used over a wide frequency range.

A very useful Beverage variation uses two parallel wires, with one grounded at the far end. Waves traveling toward the far end reflect and return to the feedpoint differentially as transmission-line currents. With suitable termination and switching circuitry at the feedpoint, you can receive signals from either direction. A two-wire Beverage thus can replace two single-wire antennas.

For simplicity, I'll consider just single-wire Beverages here.

Eliminating Ground Connections

Beverages normally connect to ground at the feedpoint and far end. See Figure 1A. However, obtaining a good, reliable ground can be difficult. A short ground rod may exhibit thousands of ohms resistance to ground when installed in dry soil or when corroded. Since Beverages require a termination resistance of just several hundred ohms, a poor far-end ground can grossly misterminate an antenna and destroy its pattern. A poor feedpoint ground can reduce signal levels and encourage coax-shield pickup. You can use multiple ground rods, a ground screen, or a ground-radial system to lower ground resistance, but this greatly increases the work required to install a Beverage.

A ground connection isn't essential for Beverage operation. You can place the termination resistance about one-quarter wavelength from the far end of the wire. Since the impedance of a quarter-wave wire is low and resistive, the remainder of the wire just sees the termination resistor in series with a low resistance. Similarly, you can feed a Beverage about one-quarter wavelength from the near end. Both ends of the wire are left unconnected. See Figure 1B.

Ungrounded Beverages are simple to install and have stable properties. Their radiation patterns often are superior to those of their grounded counterparts because they don't suffer unwanted response from vertical wires. (This response can be reduced by lowering wire height. One way to do this is to slant the feed and termination wires over a long distance.)

The only real disadvantage of ungrounded Beverages is that they're frequency-dependent. For example, the low-impedance, quarter-wave wire sections become high-impedance, half-wave sections at twice the design frequency. I'll show you how to multiband an ungrounded Beverage later.

Antenna Modeling

While you can use a *MININEC*-based antenna- analysis program to model Beverage antennas, you'll have to settle for approximate results at best. The *MININEC* algorithm uses perfect-conductivity ground during current calculations. It takes lossy ground into account only later when calculating radiation patterns. Because Beverages are so close to ground, their current distribution is strongly affected by ground characteristics. *NEC*'s Sommerfeld-Norton ground model takes ground dielectric constant and conductivity into account when calculating wire currents. This lets you accurately analyze antennas in close proximity to earth. I used the *NEC/Wires 1.0* program to obtain all results presented here. I modeled antennas over average-quality earth (dielectric constant= 13, conductivity = 5 mS/m).

Wire Height

Figures 2A and 2B shows the azimuth and elevation patterns for a 1000-foot long, 10-foot high, ungrounded Beverage on 160 meters. The antenna uses # 18 wire and a 650-Ω termination resistance. It's fed and terminated 120 feet from each end. The remarkable pattern is better than that of many HF Yagis.

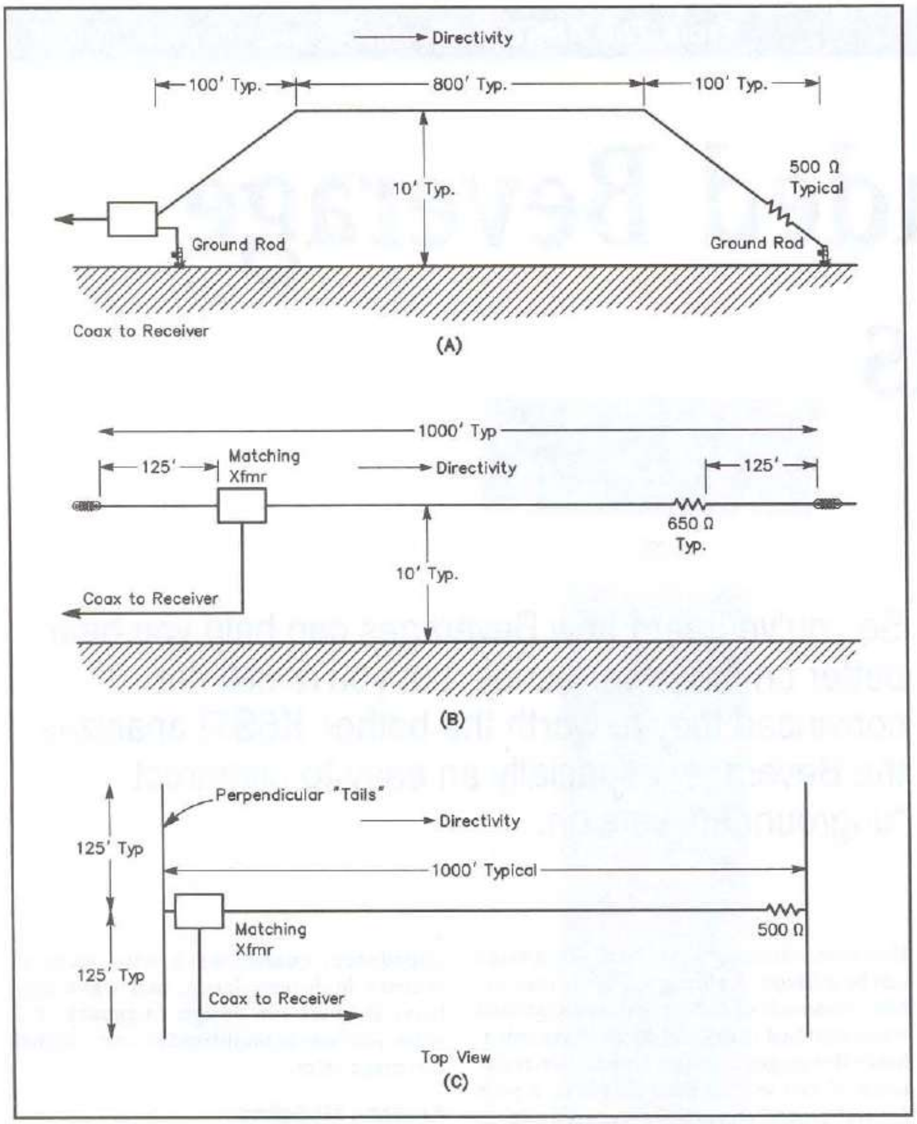

Figure 1—Single-wire Beverage antenna feeding and termination configurations. At A, conventional method is shown using slanted-wire terminations. At B, ungrounded Beverage is shown with matching transformer and terminating resistor inserted λ/4 in from either end of wire. At C, ungrounded Beverage using full span of wire by employing perpendicular "tails" at both ends. The tails function as elevated counterpoise radials at 160 meters.

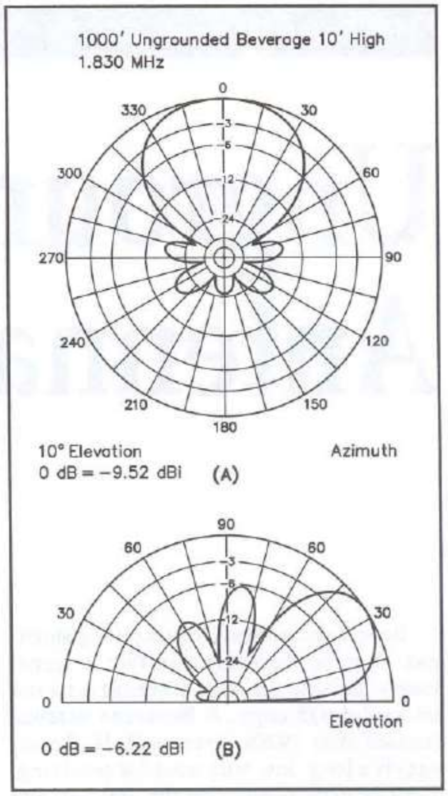

Figure 2—Azimuth and elevation patterns at 1.83 MHz for 1000-foot long ungrounded Beverage mounted 10 feet off flat ground with average conductivity and dielectric constant. The worst-case rearward lobes are better than 18 dB down compared to peak response. Note that the gain is only –9.52 dBi at 10° elevation, peaking at –6.22 dBi at about 30° elevation. If the transmission line to the receiver is very long, a preamplifier may be needed.

Figure 3 shows the azimuth pattern of a similar Beverage 1 foot above ground. The termination resistance for this antenna is 750 Ω and it's fed and terminated 125 feet from the wire ends. The pattern is even sharper than that of Figure 2. Output is about 1 dB lower.

If you're tempted as I was to just lay a Beverage on the ground, take a close look at Figure 4. This is the Beverage of Figure 3 but just 1 inch off ground (*NEC* provides accurate results for wire heights down to within several wire radii of ground). While still useful, the pattern has degraded considerably. Signal output is quite a bit lower. (Nevertheless, some users report excellent results from on-the-ground Beverages. Don't hesitate to try one if it's your only option.)

In the West you can just lay an ungrounded Beverage in the chaparral a few feet off ground. In the East you can unroll one on top of a foot or two of snow (snow has very low dielectric constant and conductivity and should be transparent to RF at low frequencies). Beverages like these are very easy to install (no support poles or ground rods). You can roll one out just before a contest.

Occasionally a Beverage may have to clear a height-sensitive area. Figure 5 shows what can happen when wire height varies. The first 200 feet of this antenna is 3 feet high (chaparral height). Next, a 50-foot section slopes from 3 to 15 feet. This is followed by a 200-foot span 15 feet high, a 50-foot downslope, and finally the remainder of the wire at 3 feet. This model is similar to a Beverage used by K6TQ which must clear an agricultural area with produce trucks. While the rear part of the pattern is relatively unaffected, the height variation causes a large sidelobe bulge. It doesn't matter much if your Beverage is

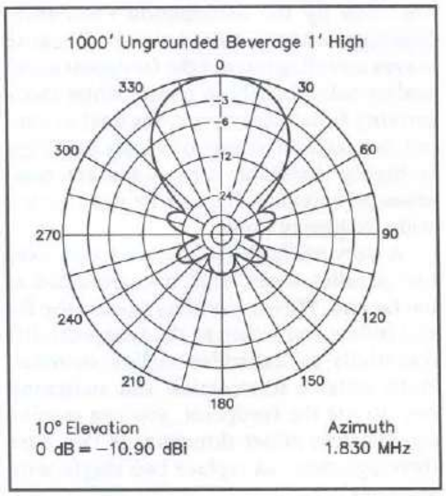

Figure 3—Azimuth response for 10° elevation angle at 1.83 MHz for 1000-foot long ungrounded Beverage mounted only 1 foot off flat ground. Compared to the higher antenna in Figure 2, this low Beverage has about 1 dB less gain, but the worst-case backlobes are suppressed even better.

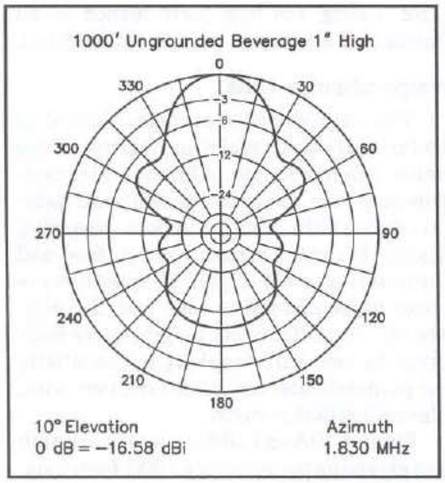

Figure 4—Azimuth response for same antenna as in Figure 3, except that the height is now only 1 inch off ground. The rearward pattern suffers and the gain fails about 5.4 dB compared to the Beverages in Figures 2 and 3.

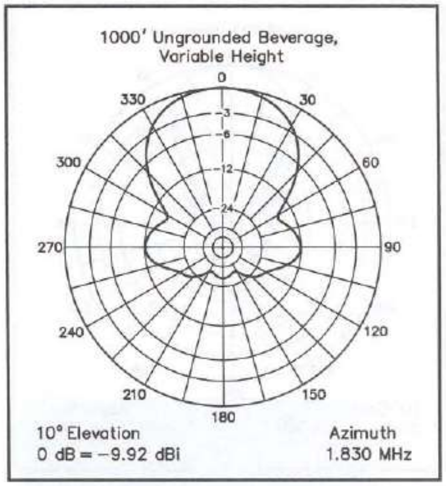

Figure 5—Azimuth response at 1.83 MHz for 1000-foot long, variable height ungrounded Beverage. The sidelobes at 90° and 270° are larger than the antenna in Figure 2, but the pattern and gain are still quite acceptable.

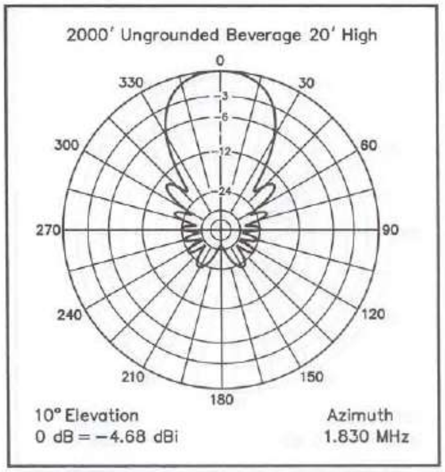

Figure 6—Azimuth response at 1.83 MHz for 2000-foot long ungrounded Beverage mounted 10 feet over flat ground. The gain is almost 5 dB higher than that of Figure 2, and the backlobes are all suppressed in excess of 23 dB. This is a great 160-meter receiving antenna, but must be oriented carefully at desired receiving bearing because of narrow frontal lobe.

high or low, but try to keep the height as constant as possible.

Wire Length

Figure 6 shows the azimuth pattern of a 2000-foot ungrounded Beverage. The termination resistance is 550 Ω. The pattern of a very long Beverage like this may be too narrow for complete coverage of a general direction.

Figure 7 shows the pattern of a 500-foot ungrounded Beverage. The termination resistance is 1100 Ω. While much broader than the patterns of longer antennas, it's still quite directive and can provide a substantial receive improvement.

Feeding Ungrounded Beverages

Conventional Beverages usually are fed with an autotransformer having a 3:1 turns ratio. This matches a 450-Ω impedance to 50 Ω. Ungrounded Beverages typically have an input resistance of 500 to 1000 Ω (roughly equal to the termination resistance) and an input reactance of up to 100 Ω. An exact impedance match isn't important; a better match just delivers a higher signal. Although usually not required to improve signal-to-noise ratio, a receive preamplifier can equalize Beverage and transmit-antenna signals for easier direct comparisons.

Both grounded and ungrounded Beverages can be sensitive to unwanted coupling at the feedpoint. If you connect coax through an autotransformer to a Beverage, the outside of the coax shield can become part of the antenna system if it exhibits an impedance similar to or less than whatever connects to it. It's the same thing that can happen with a coax-fed dipole. To prevent this, use a matching transformer with separate primary and secondary windings and low interwinding capacitance. This will isolate the outside of the coax shield from the antenna. Alternatively, you can use an autotransformer followed by a current-type balun. (Because the feed tail of an ungrounded Beverage is likely to provide a lower-impedance termination than a typical ground rod, these Beverages should be less sensitive to coax-shield pickup.)

Terminating Ungrounded Beverages

Optimal termination resistance depends on wire length, height, diameter, and ground constants. The termination resistance normally just affects the rear part of the pattern. It's easy to find the optimal value by modeling the exact geometry of your antenna with *NEC*. If you use long spans of wire with considerable sag, model the sag using multiple wires.

You can determine optimal termination resistance experimentally using a signal source to the rear of the antenna. Simply adjust the termination resistance for minimum signal.

(Actually, minimum response directly to the rear isn't quite optimal for most designs. This condition usually causes the pattern to degrade slightly over a broad region elsewhere. Nevertheless, adjusting for minimum response is close enough for all but the most fanatical Beverage enthusiast.)

Multibanding Ungrounded Beverages

If you provide a termination resistance one-quarter wavelength from the far end for each band of interest, you can use a Beverage on several bands. At the near end use a separate quarter-wave tail for each band and connect them in parallel at the feedpoint.

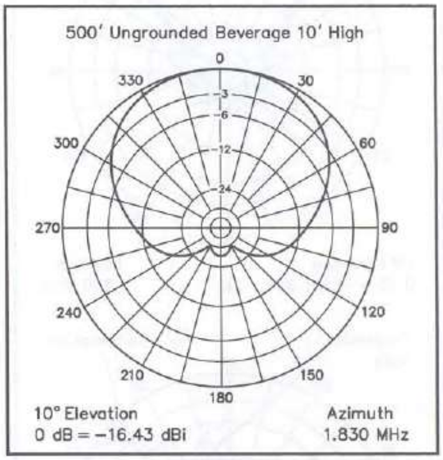

Figure 7—Azimuth response at 1.83 MHz for 500-foot long ungrounded Beverage mounted 10 feet over average ground. The gain is almost 7 dB less than Figure 2 at 10° elevation. The side lobes are large, even though the rear lobes are well suppressed.

Figure 8 and Figure 9 show 160-meter and 80-meter patterns of a 1000-foot-long, 10-foot-high, dualband Beverage. This antenna uses a 700-Ω resistor 63 feet from the far end and a 350-Ω resistor 125 feet out. The two termination resistances provide near-optimum performance on 160. The backlobe of the 80-meter pattern is suboptimal but the sidelobes are still well down. I didn't experiment much with termination-resistance values or positions. I'll bet both can be improved.

Another alternative is to use a single termination resistance with multiple quarter-wave tails. The resistor value will be optimal on just one band since antenna length isn't constant in wavelengths. But the value won't be far off on any band if the

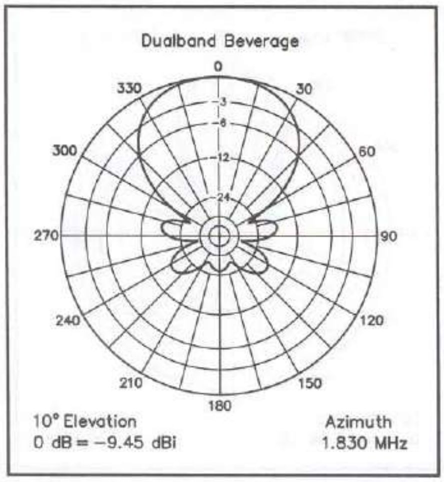

Figure 8—Dualband Beverage on 1.83 MHz, using a 700-Ω resistor 63 feet from the far end, and a 350-Ω resistor 125 feet out.

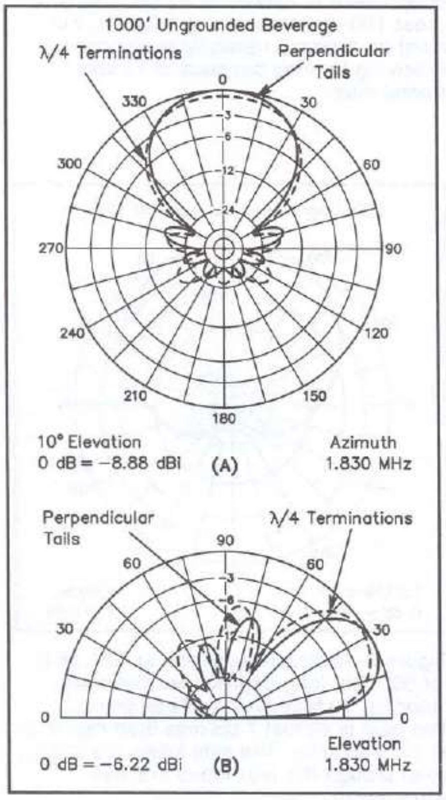

Figure 10—Ungrounded Beverage, 1000 feet long with two λ/4 perpen-dicular termination "tails." Pattern from ungrounded Beverage in Figure 2 is overlaid for comparison. At A, the azimuthal patterns at 10° elevation are shown; at B, the elevation patterns are shown.

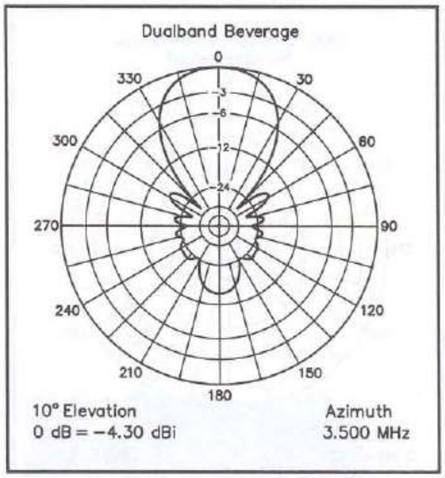

Figure 9—Dualband Beverage on 3.5 MHz, using same terminations as in Figure 7. The termination is not optimal for this band, but still yields desirable directivity.

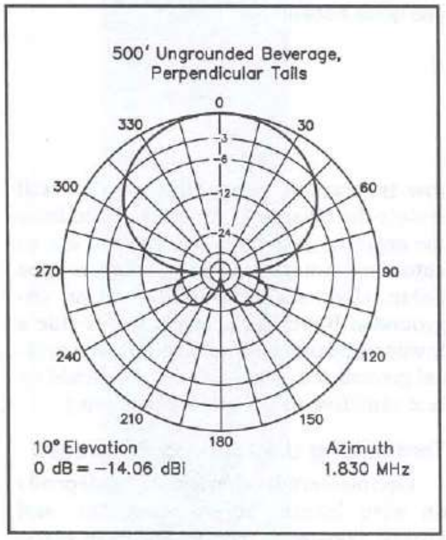

Figure 11—Azimuthal pattern for ungrounded Beverage, 500 feet long with two λ/4 perpendicular termination tails.

wire is long. For best performance on all bands, use a separate resistor for each tail.

Perpendicular Tails

The ungrounded Beverages described so far really don't make optimal use of the entire antenna length. Although they contribute power, the near- and far-end quarter-wave tails don't support traveling waves. If space permits you can feed and terminate against a pair of quarterwave wires perpendicular to each end. See Figure 1C. This effectively lengthens the Beverage by one-half wavelength. Essentially the perpendicular tails form two, two-wire, elevated radial systems.

Figures 10A and 10B shows the azimuth and elevation patterns for a 1000-foot-long, 10-foot-high Beverage with perpendicular tails, with the patterns from Figure 2 overlaid for comparison. This antenna uses a 500-Ω termination. The pattern is somewhat narrower than that of Figure 2 and the backlobes are even smaller. The elevation plot shows that overhead response is about 10 dB less with perpendicular tails and that the forward lobe cants at a lower angle.

Since the quarter-wave tails of a 500-foot Beverage comprise half its length, you might expect a more dramatic improvement for perpendicular tails on this antenna. Figure 11 shows the pattern of a 500-foot-long, 10-foot-high Beverage with perpendicular tails. This antenna also uses a 500-Ω termination. Side response is down quite a bit from Figure 7 and forward response is 2 dB greater.

While they provide better performance, designs with perpendicular tails are more complex and occupy additional space. You can make better use of the additional space by phasing two Beverages with in-line tails to obtain a much better azimuth pattern . . . but that's another story.

By Brian Beezley, K6STI **From *QST*, September 1995**

A Receiving Antenna that Rejects Local Noise

Simplicity and performance combine to give birth to a compact antenna you'll want to have!

Noise can make a ham's life miserable on any amateur band. As we approach the minimum of the sunspot cycle, many hams are discovering that noise can be particularly frustrating on the low bands. In summer, static crashes caused by thunderstorm lightning can totally mask weak signals on the 160, 80, and 40-meter bands. During other seasons, power-line noise, noise from household appliances, and incidental radiation from home electronic products often limits reception.

A recent *QST* article by Floyd Koontz, WA2WVL,[1] describes a small receiving antenna for the low bands that provides a cardioid directional pattern. This pattern can reduce noise and QRM from the rear. As I marveled at the elegance and simplicity of Floyd's design, I realized that the antenna did have one shortcoming: Because it is vertically polarized, the antenna responds strongly to local noise propagated by ground waves. I wondered whether it was possible to devise a receiving antenna to better reject local noise.

The Ground Wave

Most hams who operate HF are familiar with the sky wave (or space wave) that's responsible for long-distance ionospheric propagation. See Figure 1A. The space wave has two components: The *direct wave* propagates along a straight line from the transmit antenna toward the ionosphere. The *ground-reflected wave* bounces off the earth's surface and heads in the same direction.

The space wave also exists for local propagation, as shown in Figure 1B. The direct wave travels in a straight line between the transmit and receive antennas, while the ground-reflected wave takes a midpoint bounce. But when the antennas are close to ground, the direct and reflected waves nearly cancel, leaving a very small residual space wave. When the antennas are right at the earth's surface, the waves cancel completely. So what makes local communication possible? Answer: A third wave, called the *surface wave*, that exists for antennas close to ground. This wave diminishes in intensity as you increase antenna height. The surface wave exists *only near the surface of the earth*. The combination of the direct wave, the ground-reflected wave, and the surface wave is called the *ground wave*.[2]

Surface-wave intensity varies with frequency and ground conductivity. It's stronger at low frequencies and for highly conductive ground. But the most important property of the surface wave is its polarization sensitivity. The surface wave is much weaker for horizontal fields. For example, on 80 meters the broadside ground-wave response of a short piece of wire is about 34 dB lower when oriented horizontally. The difference is about 42 dB at 160 meters.[3]

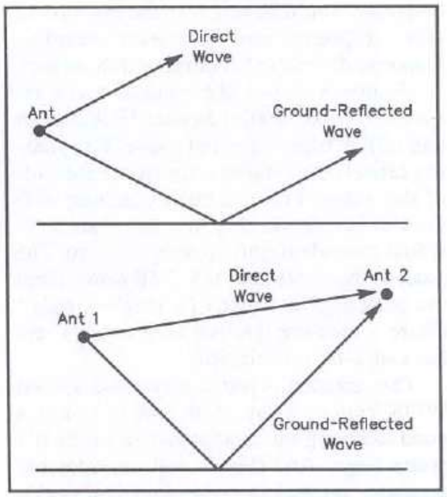

Figure 1—At A, the direct and ground-reflected components of the space wave for ionospheric propagation. At B, the space-wave components for local propagation.

The surface wave makes local AM radio broadcasting possible. Because of the poor propagation of horizontally polarized ground-wave signals, AM broadcasters universally use vertical polarization. The same phenomenon causes vertical antennas to pick up much more local noise than horizontal antennas do. Even if a noise source has a stronger horizontal component, by the time the field reaches the receive antenna, the vertical component almost always dominates.

These facts suggest that the first requirement of a receiving antenna with low response to local noise is insensitivity to vertically polarized radiation, the dominant component of the ground wave. Surprisingly, simply avoiding the use of vertical wires isn't enough. An antenna composed only of horizontal wires can still respond to vertical fields.

A Low Dipole

Figure 2 shows the ground-wave response of an 80-meter dipole 10 feet high. An easy-to-install, inconspicuous dipole like this is sometimes used for receiving when the transmit antenna is vertically polarized. The pattern shows the electric field strength 10 meters above ground at a distance of 1000 meters for an input power of 1 kW (the dipole exhibits the same pattern on receive). This geometry might be representative of that for a noisy power pole. Although the pattern may look similar to that of a free-space dipole, I think

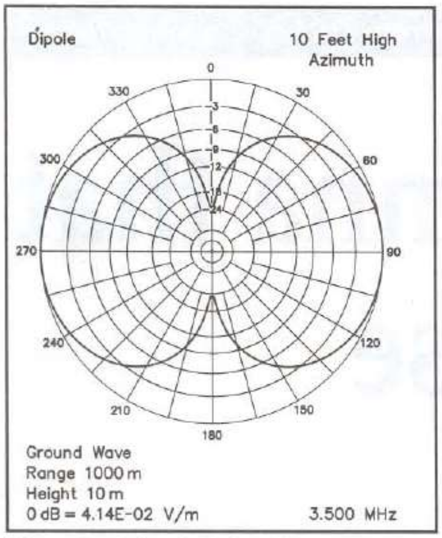

Figure 2—Azimuthal plot of the ground-wave response of a 10-foot-high 80-meter dipole. The input power to the antenna is 1 kW. The peak electric field is shown.

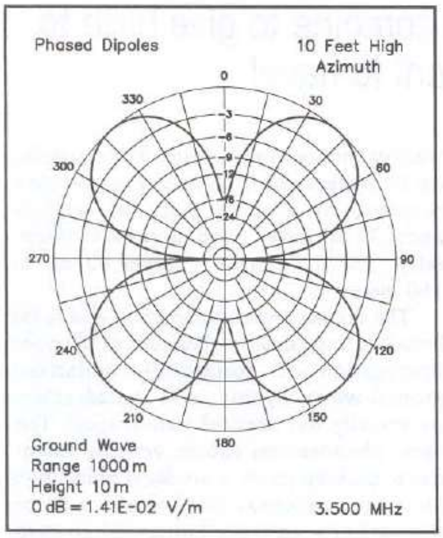

Figure 3—Azimuthal plot of the ground-wave response of two parallel dipoles 10 feet apart, 10 feet high and fed out of phase. The input power to the antenna is 1 kW. The peak electric field is shown.

you'll be surprised to know that the wire is oriented broadside to the pattern null. A low dipole actually responds to ground-wave fields best off its ends!

Here's an explanation for this peculiar behavior: The dipole has no response to the vertical component of a broadside ground wave because the electric field is perpendicular to the wire. The antenna responds only to the weak horizontal component. The vertical component also is perpendicular to the mirror image of the antenna formed by the ground-reflected wave. But away from broadside, the dipole, its image, or both, have a nonvanishing projection in the vertical plane. This enables the dipole to respond to the vertical component of a ground wave. In addition, lossy earth causes the surface wave to develop a radial component in the direction of propagation. Away from broadside, the radial component also projects onto the wire and induces current. The vertical and radial components induce maximum current when the surface wave arrives in line with the wire.

When compared to a quarter-wave vertical with four radials elevated 10 feet above ground, the low dipole has a 12.5 dB lower ground-wave response in its most sensitive direction in line with the wire. For sky-wave signals arriving at 20° elevation from their weakest direction (also off the ends), the dipole has 10 dB lower response than the vertical. Therefore, the low dipole has a signal-to-noise ratio advantage of 2.5 dB for signals and noise arriving from their worst-possible directions. In the most favorable directions broadside to the wire, the S/N advantage peaks sharply at 26.1 dB.

Raising the dipole broadens the broadside S/N peak and improves S/N off the ends. For example, for signals arriving at 20° elevation, a dipole at 50 feet has an S/N advantage over the reference vertical of 6.3 dB in line with the wire and 25.6 dB broadside. If you have just a single noise source and you can rotate a high dipole, you should be able to come within a few decibels of the latter figure most of the time. But when multiple noise sources in different directions arise (typical for power-line noise in times of low humidity), rotating the antenna won't help much. The S/N advantage of the high dipole then is likely to be near the worst-case figure.

These numbers illustrate the advantage of horizontal receiving antennas and substantiate the notion long held by amateurs that "verticals are noisy." But you can reduce ground-wave noise much more effectively if you don't rely on a simple horizontal wire.

Canceling Ground-Wave Components

Although a horizontal wire responds only weakly to a broadside ground wave, its response is substantial off the ends. If you could somehow eliminate the end response, you'd be left with the low broadside response and whatever residual response developed at intermediate angles.

Figure 3 shows the ground-wave response of two parallel dipoles 10 feet apart and 10 feet high fed out of phase. The phasing cancels everything arriving off the ends of the wires. The end nulls combine with the low broadside response to create a residual cloverleaf ground-wave pattern. The peak of the cloverleaf is 9.3 dB down from the peak end-response of a single dipole.[4] (Wire losses are ignored here to illustrate the cancellation principle.)

This antenna is just a very-close-spaced W8JK endfire array. Although it makes a good receiving antenna for the low bands, it's pretty large. And there's still considerable ground-wave pickup in the cloverleaf peaks. If the wires somehow could remain parallel for all directions, it might be possible to achieve complete cancellation of the vertical and radial components of the ground wave.

In some sense, the sides of a circular loop are parallel everywhere. Current amplitude and phase vary little in small loops of regular shape. Therefore, the currents in opposite sides of such loops are nearly equal and out of phase. Unlike a W8JK array, a small horizontal loop does not have a null anywhere along the ground. But its ground-wave response is uniformly low in all directions because the antenna responds only to the weak horizontal component. Everything else cancels out (or nearly so).

A small loop usually is defined as one with a total conductor length of less than 0.1 λ. But unless you use large-diameter conductors to minimize RF resistance, loops this small are inefficient. A preamplifier may be needed to overcome receiver noise. You can increase the output of a loop by increasing its size, but the larger you make it, the less constant the current becomes. This reduces groundwave cancellation.

The antenna of Figure 4 overcomes this difficulty by using two feedpoints, each on opposite sides of the loop to force current balance. One of the phasing lines is twisted to maintain proper phase.[5] This loop can be made quite large and still exhibit very low response to ground-wave noise.

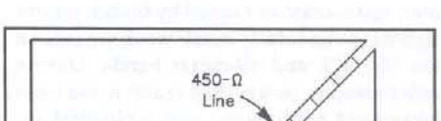

Figure 4—Basic diagram of the 80-meter low-noise loop antenna showing detail of the feedpoint arrangement. The antenna measures 25 feet on a side, is 10 feet high, and made of #14 wire. It's fed at opposite corners with phasing lines made of #14 wire spaced 1.5 inches. A small ferrite transformer at the junction of the phasing lines matches the antenna to 50 Ω coaxial feed line and also functions as a balun. The trimmer capacitor (a capacitance of about 40 pF is required) in series with the antenna-side winding resonates the loop at 3.5 MHz.

A Practical Design

The 80-meter loop of Figure 4 has a perimeter of 0.36 λ. It's 25 feet on a side, 10 feet high, and made of #14 wire. It's fed at opposite corners with phasing lines made of #14 wire spaced 1.5 inches. A small ferrite transformer at the junction of the phasing lines matches the antenna to 50 Ω and also functions as a balun. A trimmer capacitor (about 40 pF is needed) in series with the antenna-side winding resonates the loop at 3.5 MHz.

The ground-wave response of this particular loop is shown in Figure 5 and its sky-wave response in Figure 6. For signals arriving at 20°elevation, the worst-case sky-wave response is 20.5 dB below that of the reference vertical. This signal level is quite usable on 80 meters without a preamp. Because the loop reduces ground wave noise at least 45.1dB, S/N improvement is 24.6 dB for the worst-case combination of signal and noise directions. if you use a preamp and adjust for equal signal levels, ground-wave noise will be four S units lower on the loop *no matter what direction it comes from!* For signals arriving at higher angles, S/N enhancement approaches 30 dB.

Both the ground-wave noise pattern and sky-wave signal pattern are very uniform in azimuth. The antenna is essentially omnidirectional with an overhead null, just like a vertical. An overhead null is useful for reducing near-vertical-incidence skywave signals from nearby stations. (The loop rejects their ground-wave signals along with local noise.)

Although the S/N performance of this loop is not particularly sensitive to height, you can increase output level substantially by raising the antenna well above ground. For example, if you raise the loop to 50 feet, the output increases 15.2 dB for signals arriving at 20° elevation. At this height, the output level is only 5.3 dB below that of the reference vertical. You'll never need a preamp with a loop this high. The S/N advantage drops 0.1 dB. At a height of 20 feet, output increases 7.2 dB and S/N drops 0.6 dB.

You can shrink the loop to 10 feet on a side. The S/N advantage increases 0.5 dB, but the signal level drops 6.5 dB. You can raise the signal level 3.4 dB by doubling the loop side lengths to 50 feet, but the S/N advantage then drops 7.9 dB. You can thus trade-off loop size, S/N enhancement, and signal level. If you use smaller wire, output drops. For example, it's 1.7 dB lower for #22 wire.

Although this loop is a narrowband device and must be carefully resonated, the resonance is much broader than that of a typical small loop. You can get away with a single capacitor setting for both 3.5 and 3.8 MHz if you're willing to accept somewhat lower signal levels. The capacitor setting does not affect the patterns or skywave-to-ground-wave ratio—it simply alters output level.

The input resistance at the junction of the phasing lines varies over a wide range with loop size, height, and phasing-line characteristic impedance. The input resistance is about 40 Ω for the loop of Figure 4 at a height of 10 feet. A transformer using a type-77 ferrite core (such as an FT-82-77, FT-114-77 or FB-77-1024) with 9 turns of any size enameled wire on the coax side and 8 wire turns on the antenna side provides a good match to 50-Ω coax.[6] The input resistance drops to about 20 Ω when the loop is raised to 20 feet and to about 15 Ω at 50 feet. Use 13 turns of wire on the coax side for 20 feet and 15 for 50 feet.

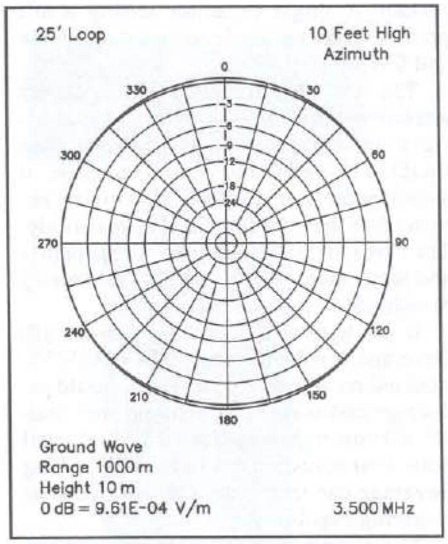

Figure 5—Azimuthal plot of the ground-wave component of a 25-foot-square, 80-meter loop at a height of 10 feet. The input power to the antenna is 1 kW. The peak electric field is shown.

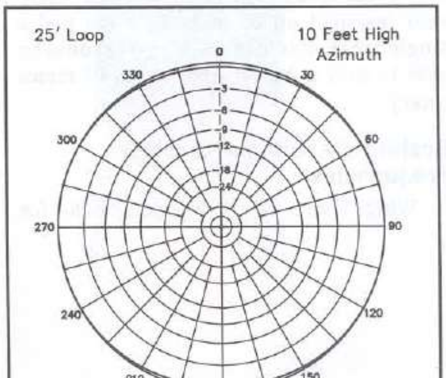

Figure 6—The sky-wave responses of the 10-foot-high, 25-foot-square, 80-meter loop. At A, the azimuthal plot; at B, the elevation plot.

Although the optimum transformer turns ratio varies with antenna height, little output is sacrificed if you use a fixed ratio.

For example, output is only 1 dB less than optimal when a loop designed for 10 feet is used at 50 feet. If you're determined to obtain the best possible match, use 16 turns of wire on the coax side and tap the winding for lowest SWR (alternatively, a switch can be used to select any of two or more taps). Use the lowest possible power when measuring SWR. It's easy to puncture the dielectric of a small trimmer capacitor with a momentary blast.

The antenna's resonant frequency shifts when the phasing lines get wet. If you use true open-wire line with plastic spacers, the frequency shift will be less than 100 kHz. But if you use 450-Ω line with segmented polyethylene dielectric, the resonant frequency decreases more than 200 kHz when the line becomes thoroughly damp. Although this won't affect signal-to-noise ratio, output drops in the desired frequency range. You may be tempted to try phasing lines of 300-Ω twinlead routed inside PVC tubing to avoid moisture effects, but line impedances this low work only for smaller loops.

This antenna should be constructed as symmetrically as possible to maximize cancellation of the vertical and radial components of the ground wave. Make the loop perfectly square and accurately align it in the horizontal plane. Cut the phasing lines to the same length. Although these loops perform well near houses, fences, and towers, try to install the antenna as far from other conductors as possible to maximize current balance. Use the shortest possible leads to interconnect the matching components. Although it's probably unnecessary, I like to split the antenna-side transformer winding and put the tuning capacitor in the center to promote equal currents in the phasing-line conductors.

To minimize the number of supports, a loop about 17 feet on a side can be constructed using a 20-meter quad spreader. You can mount the spreader well up on a tower to increase output. Alternatively, you may be able to eliminate supports altogether by stringing a loop in your attic or garage. However, the current balance of indoor loops may be degraded by electrical wiring, plumbing, heating ducts, or other nearby conductors.

On-the-Air Performance

Ed Andress, W6KUT, located in the San Diego suburb of Poway, constructed a loop 21 feet on a side and 10 feet high. Ed's location is subjected to strong, chronic power-line noise. On 80 meters, Ed uses a pair of phased quarter-wave verticals for transmitting. We used the verticals as a reference when evaluating the loop. A 20-dB preamp was available. With the preamp, signals near loop resonance were about equal to those from the verticals.

The loop performed as expected. It

dramatically enhanced the signal-to-noise ratio of most sky-wave signals. Sometimes the verticals did better during a momentary change in propagation; occasionally a particular noise arose that the loop didn't attenuate much. But overall, the loop was far superior. It made little difference on strong signals. It made listening to moderately strong signals much more pleasant. It let us copy weak signals that were buried in the noise and unreadable on the verticals.

During times of no detectable powerline noise, we often noticed a curious effect: The loop still enhanced signal-to-noise ratio by one or two S units, making copy of moderately weak signals more pleasant. On these occasions, we were unable to hear the telltale, raspy buzz of power-line noise (or any other noise signature) when we listened with the transceiver's AM detector. Unless noise happens to arrive at low angles and signals at high, there's no reason for the loop to enhance skywave S/N. We believe that the unidentified noise is local and propagates by the ground wave. We speculate that it may be the sum of hundreds of weak man-made noise sources in the densely populated suburb. (The superposition of a large number of noise sources tends to be characterless even when the individual sources aren't.)

Total Ground-Wave Cancellation

If you stack two of these loops vertically and bring both feed lines into the shack, you can form a deep null on the horizon for all azimuth angles by combining the signals with a fixed amplitude and phase offset. Except for azimuth-response irregularities caused by nearby conductors and small elevation-angle differences due to range, this system will cancel all ground-wave components. This noise canceler was to have been the original subject of this article. However, a single component loop worked so well in practice that I decided total ground-wave cancellation was overkill.

Comparison with Other Antennas

A conventional, single-feedpoint, small loop about two feet on a side—oriented horizontally—yields roughly the same S/N enhancement as the loop of Figure 4. However, output will be down about 46 dB from the reference vertical. You'll need a low-noise preamp with a loop this small unless your receiver has a very low noise figure. Still, even with a preamp, a conventional small loop makes an attractive, low-profile alternative. You must construct a low-output antenna like this carefully to avoid stray pickup. A single capacitor setting won't provide good output levels on both phone and CW.

The WA2WVL cardioid antenna attenuates thunderstorm static and ground-wave noise to the rear. If you're seldom troubled by omnidirectional local noise, it should make a more effective receiving antenna than the loop described in this article. The cardioid also requires fewer supports and less space. As a bonus, it's inherently broadband.

If you have room for a two-wavelength Beverage, it will outperform the WA2WVL cardioid on sky-wave noise and should reduce ground-wave noise arriving more than 45° off boresight by at least 15 dB. If local noise near boresight isn't a problem, a long Beverage can tremendously improve your receiving capability.

The easiest way to improve reception on 80 or 160 meters is to use the most sensitive horizontal antenna available at your antenna switch. Many hams with 80-meter verticals find that switching to a 40-meter dipole or beam improves copy of weak signals even though the antenna is nowhere near resonant on 80 meters. When just a single noise source is active, you should be able to null it by broadsiding a 40-meter rotary.

Scaling the Antenna to Other Frequencies

While I've used the 80-meter band for illustration in this article, it's easy to scale the design to other frequencies. Simply multiply lengths, heights, transformer turns and capacitor values by the number you get when you divide 3.5 by the target frequency in MHz.

That said, I don't recommend this antenna for use above 40 meters. If you're using a vertical antenna on the upper HF bands, do yourself a favor and replace it with the highest horizontally polarized antenna you can manage. Not only will your receive noise decrease, your transmit signal almost certainly will improve due to higher ground-reflection gain.[7] (Exception: If your vertical radiates over saltwater, keep it!) If you're already using a horizontal wire on upper HF, I think your next antenna project should be a rotary beam rather than a receiving loop.

Notes

[1] Floyd Koontz, WA2WVL, "Is this Ewe for You?" *QST*, Feb 1995, pp 31-33.

[2] Frederick Terman, *Radio Engineers' Handbook*, 1st ed (New York: McGraw-Hill, 1943), pp 674-709.

[3] *MININEC*-based antenna-analysis programs do not compute the surface wave, nor can they accurately model antennas close to ground. I used *NEC/Wires 2.0* with its surface-wave and Sommerfeld-Norton ground options for all antenna models in this article. All models assumed average ground characteristics (dielectric constant 13, conductivity 5 mS/m).

[4] A small cloverleaf peak requires very close wire spacing. For example, the peak is down only 5.4 dB for a spacing of 50 feet.

[5] The feed arrangement is like that of an Alford Loop. In fact, the low-noise receiving antenna really is just a variant of the VHF/UHF loop devised by Andrew Alford for a different purpose in 1940. See Terman, pp 814-815.

[6] Don't be tempted to eliminate the transformer and connect 50-Ω coax directly to a loop. The transformer functions as a current balun. It keeps noise current induced on the coax shield from entering the receiver. Just a little pickup by the shield can pollute the low-noise output of a loop.

[7] For example, on 20 meters, a horizontal dipole only 35 feet high has 3 to 6 dB broadside gain over a vertical dipole at low angles over average ground.

By Gary Breed, K9AY **From *QST*, September 1997**

The K9AY Terminated Loop—A Compact, Directional Receiving Antenna

Wish you had enough room for an effective low-band receiving antenna? You do! This four-direction system fits in a 30-foot circle!

Low-band operators are always looking for ways to improve their hearing. As a low-band fan, I was impressed with the EWE antenna developed by Floyd Koontz, WA2WVL.[1,2] Koontz shows us how to build a compact, directional antenna—a design that quickly became very popular. But when I sat down at the computer to figure out the best way to install my own EWEs, a surprising new design emerged from my modeling experiments.

Ladies and Gentlemen...

Allow me to introduce you to the *terminated loop*, a concept that further shrinks the space required for a good receiving antenna *without sacrificing performance*! Figure 1 compares the real estate requirements of my four-direction loop system to that of an equivalent EWE array. The new system is not only smaller, it's easier to install, needing only one support instead of five.

The terminated loop (see Figure 2) is physically and electrically quite simple. It consists of a wire loop of any convenient shape (diamond, delta, etc), hung from a single support and with a ground rod at the bottom. A 9:1 impedance-matching transformer connects from one end of the loop to ground; a terminating resistor connects the other end of the loop to ground. This antenna is directional, favoring signals arriving from the feed point end, rejecting by several S units any signals arriving from the end connected to the terminating resistor, R_{TERM}.

Here's the real news: A very small terminated loop maintains its directional pattern. As the loop's size gets smaller, however, the desired signal strength is also reduced. The antenna described here is small enough to fit into a corner of almost *any* backyard. The loop could be made smaller, but I wanted to collect enough signal energy so that

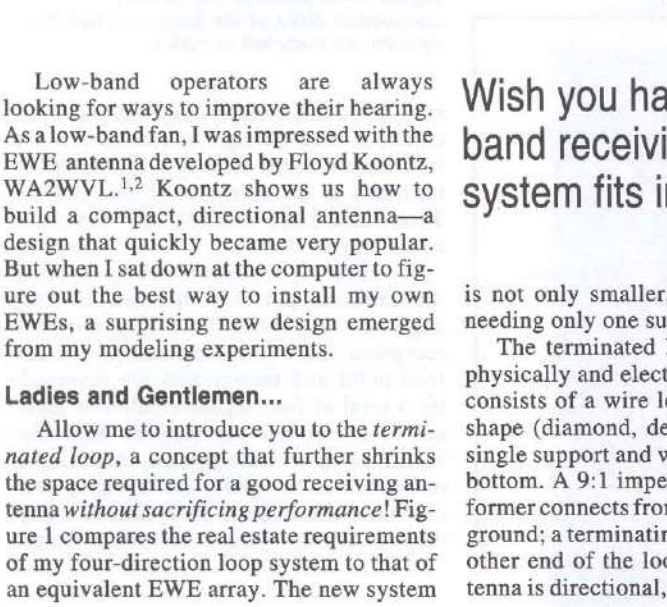

Figure 1—A comparison of the real estate needed for four EWEs and the K9AY Loop system shows that the loops need only 1/7 the area of the EWEs, yet they provide the same directional patterns.

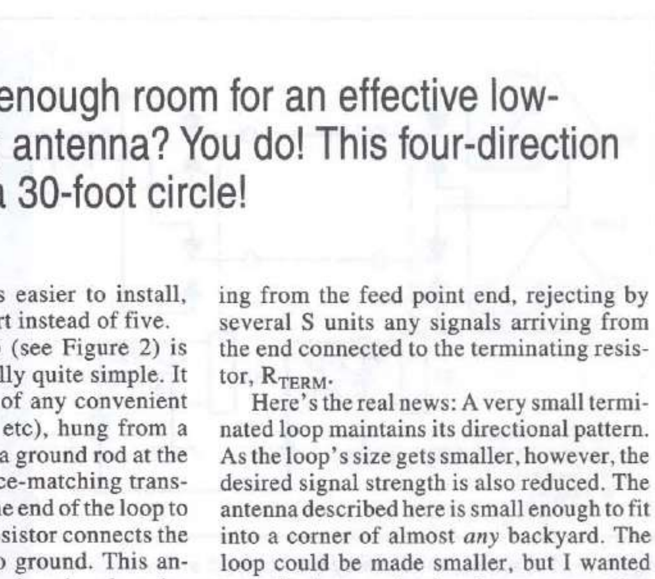

Figure 2—The basic design of a single loop in an easy-to-construct quasi-delta loop configuration. Exchanging the feed point and termination reverses the pattern. A four-direction system uses two of these loops installed at right angles to one another (as shown in Figure 1) and a relay controlled switching system.

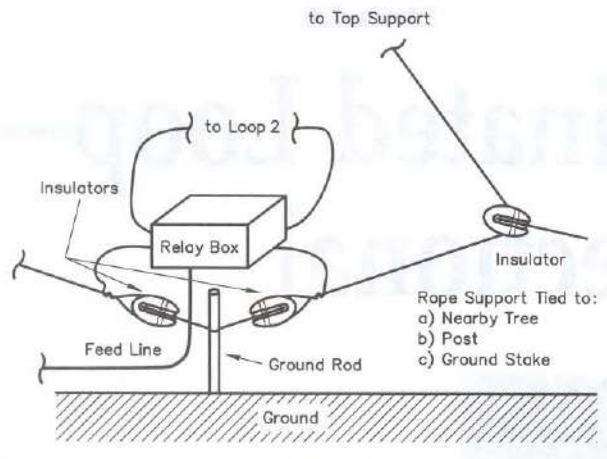

Figure 3—Construction details of a terminated loop, showing the base connections and corner supports of one of the loops. The relay box also houses the 9:1 matching transformer.

Figure 4—A photo of the central connection point at the base of a two-loop system, as installed at K9AY.

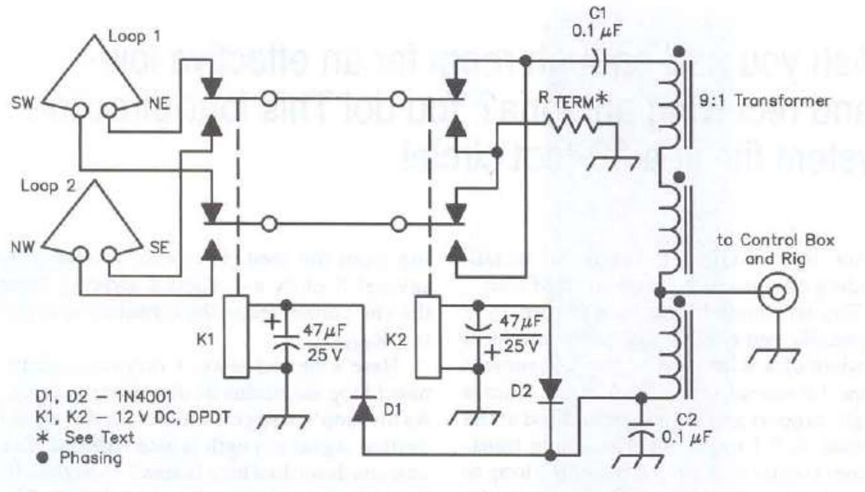

Figure 5—Schematic of the relay box located at the base of the system. The relays switch between the two loops, reversing the feed point and termination connections.

even a modest preamp (such as those included in most HF rigs) can be used. More information on the antenna's operation can be found in the sidebars "How Does the Terminated Loop Work?" and "Summary of Characteristics."

To cover all directions, two loops using the same support are oriented at right angles to each other. Each loop provides reception from two directions when the feed point and termination are reversed, for a total of four separate patterns. Here are all the details you need to build the K9AY Loop antenna system, including a relay-controlled pattern-switching system that uses the coaxial feed line to carry the switching system control voltage.

How Does the Terminated Loop Work?

The terminated loop was developed after examining the behavior of the EWE. In theory, the EWE is a *terminated half-loop*. High-frequency directional couplers such as those used in the well-known Bird Thruline wattmeters are constructed similarly, just much smaller.* My analysis determined that a *terminated full loop,* fed and terminated to a single ground point at bottom center, behaves the same way as a half loop.

How does a single loop achieve a directional pattern? By the way it responds to the electric (E) and magnetic (H) field components of the arriving electromagnetic wave. Let's say we have a signal arriving at the loop from one end. As the wave passes, the loop's wire intercepts the E-field like a short vertical antenna, creating a *voltage* at the feed point. As expected from a vertical, the E-field response is omnidirectional.

The magnetic field works differently. The H field is at right angles to the E field and induces a *current* as it passes through the loop. The voltage developed across the terminating resistor by that current is combined with the E-field voltage. If the wave arrives from the feed point end, the voltages add in phase. If the wave arrives from the opposite direction, the voltage is 180° out of phase with (and subtracted from) the E-field voltage. To maximize the front-to-back ratio, the terminating resistor must have a value that balances the voltages created by the two field components so that their sum is close to zero.

The resulting pattern is a cardioid with a single null. With an optimum terminating resistor, the null can reach 40 dB or more in depth—that's more than six S units! The null is not at ground level, but at 20 to 55° elevation, depending on the shape of the loop and local ground conditions. Unless you build a short, wide loop or a tall, skinny loop, the null will be at 30 to 40° elevation, very convenient for reducing QRM from in-country stations.

Ground is an essential part of the antenna—its resistance is part of the circuit. If ground conductivity is known at your location, it can be included in the computer modeling parameters. However, ground conditions can change over a short distance and vary with seasonal changes and moisture content. If you find that the antenna does not perform as modeled, adjustment of the terminating resistor value may be needed, as noted in the article.

Ground does not need to be lossy, as is the case with a Beverage antenna. The loop has directivity even with perfect ground. This means that you can install it over *any* type of ground, including a buried radial system (as long as it's not too close to another structure).—*Gary Breed, K9AY*

**Instruction Book*, Model 43 Wattmeter, Bird Electronic Corp, Solon, OH.

Let's Build It!

First, gather the following materials to construct the loops:

- Two lengths of **wire** about 85 feet each. Although almost any wire size will do, #14 copper is probably the best for a long-lasting installation.
- Ten simple **insulators**—anything from old toothbrush handles to fancy porcelain insulators will do.
- A three or four-foot length of **copper pipe**—the ground rod doesn't need to be driven deeply—a depth of three to four feet is sufficient. Because copper-plated steel rods eventually rust, I prefer to use 1/2 or 3/4-inch-diameter copper water pipe.
- One **support** 25 feet above ground—my antenna hangs from a limb of a Georgia Pine, but anything that gets the top of the loops up about 25 feet will work. The wooden **A** frame support described in *The ARRL Handbook* and *Antenna Book* is an excellent choice. A metal mast can be used, but make it only as long as necessary and insulate it from ground. Large metal objects affect the antenna's performance, so install it in the clear.
- You'll also need some **rope**—nothing fancy, just something strong enough to keep things in place.

Now, grab your tools and head for the backyard. Attach the midpoint of each wire length to an insulator that will be positioned at the top of the loop. The two loops must not touch each other, so leave some room between their insulators. I separated my insulators with about a foot of rope.

Aligned directly under the tops of the loops, drive a ground rod, leaving a foot or so above ground as an attachment point. Using short ropes or wires, connect four insulators a few inches away from this post to receive the lower ends of the loops. Leave a foot of wire for a pigtail after twisting the ends of the loop wires around the insulators. Figure 3 sketches the important installation details, and the photo of Figure 4 shows how my system is installed.

The lower corners of the loops are supported by insulators and ropes tied off to nearby trees, fence posts, or stakes driven into the ground. Pull out the corners with enough tension to maintain the loop shapes. At this point, the major mechanical work is finished and you should have something resembling an eggbeater, as shown in Figure 1.

The Relay Box and Controller

Figure 5 shows the relay circuit used to switch the feed and termination points of the loop wires. House the components in a weatherproof box having external connection points for the four ends of the loops, a coax connector for the feed line and a ground-wire attachment.

K1 is a DPDT relay that switches between the two loops. K2, another DPDT relay, swaps the connections of the terminating resistor and matching transformer, thereby reversing the pattern. Relay power is supplied via the coaxial feed line. C1 keeps the control voltage from reaching the antenna, and C2 provides an RF ground for the transformer.

Most builders will likely choose to orient the loops for northeast/southwest and northwest/southeast directions. The relay box has four switching modes, one for each direction: 1) northeast, when *neither* relay is energized; 2) southeast, when K1 *only* is energized; 3) southwest, when K2 *only* is energized and 4) northwest, when *both* relays are energized. Switching is accomplished using a single power connection through steering diodes D1 and D2. D1 allows K1 to operate when +12 V is applied, while D2 blocks the voltage from reaching K2. When −12 V is applied, K2 operates, but not K1. When 12 V ac is applied, the diodes rectify it to operate *both* relays.

The matching transformer is a 9:1 impedance, 3:1 turns ratio type that should be familiar to many readers. Five trifilar turns of ordinary hookup wire are wound on a 3/4-inch-diameter (0.825 inch) 43 material toroid. The terminating resistor (R_{TERM}) value will be between 390 to 560 Ω, depending on your band preference. With average ground conductivity, a value of 390 Ω provides the optimum F/B at 160 meters, while 560 Ω optimizes the loops for 80 meters. A value of 470 Ω splits the difference for "pretty good" performance on both bands. I chose to optimize the antenna for 160 meter operation, so used a 390 Ω resistor. Use an R_{TERM} power rating of at least 1 W in case some transmitter power ends up being coupled to the loops. I use a 2-W carbon resistor, but two parallel 1/2-W or four 1/4-W units of appropriate ohmic value can also be substituted.

Figure 6 is a diagram of the control box that is located in the shack. One 12 V ac transformer provides the ac relay power and feeds two half-wave rectifiers and filter capacitors to generate ±12 V. One SP4T switch selects the proper voltage to apply to the coax. The RF choke and 0.1 µF capacitor keep the RF and control voltage separated at the shack end.

Once the control unit and relay box are built and operating properly, mount the relay box at the ground rod and connect the four ends of the loops to their proper terminals. A length of 50 Ω coax carries the signal and power between the antenna and the switch box in the shack. I highly recommend keep-

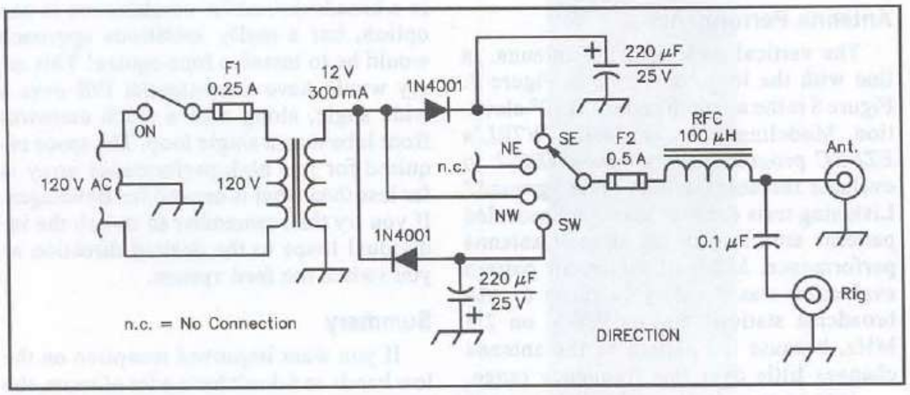

Figure 6—Schematic of the control box located in the shack. The control voltage to operate the relays is delivered via the coaxial line.

Summary of Characteristics

When choosing a specific configuration or location for your K9AY Loop system, consider this list of characteristics:

- Like all antennas, this one can be affected by nearby metallic objects or structures. Do your best to keep the antenna in the clear, away from your tower, house and power lines.
- Noise reduction in this antenna is achieved mainly by the directional pattern, although being grounded offers some reduction of wind, rain and snow static. Compared to the typical omnidirectional vertical or low inverted **V**, the reduction in overall noise and interference can be dramatic.
- The maximum circumference of the loop is a little over $1/4\ \lambda$ at the highest frequency of operation. If the loop is larger, the E and H-field responses of the antenna can no longer be balanced. Smaller loops (or same-size loops at lower frequencies) retain the directional pattern, which makes this an excellent antenna for AM broadcast reception. Unfortunately, the received signal voltage is proportional to the area enclosed by the loop, so sensitivity decreases rapidly as the antenna becomes smaller. Unless you have a very good preamp, keep the loop sizes near the maximum.
- The T-Loop is not just a low-frequency antenna! It can be used at high frequencies for shortwave listening, or to provide improved reception over an omnidirectional antenna such as a vertical. Just scale the size to the desired frequency using the guidelines presented in the accompanying article.—*Gary Breed, K9AY*

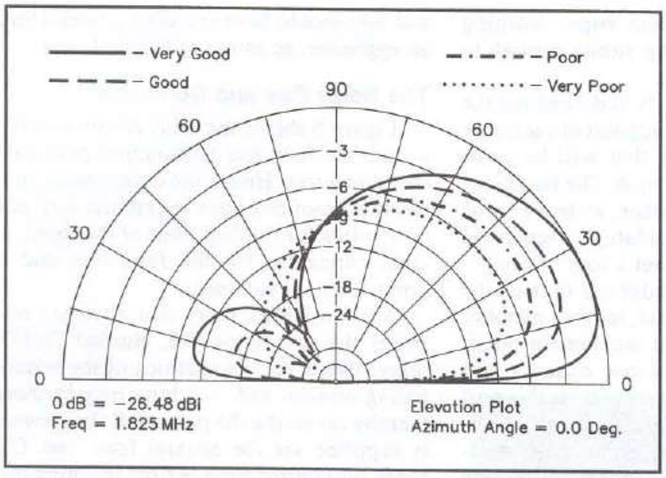

Figure 7—Vertical radiation pattern of the loops along the plane of the loop.

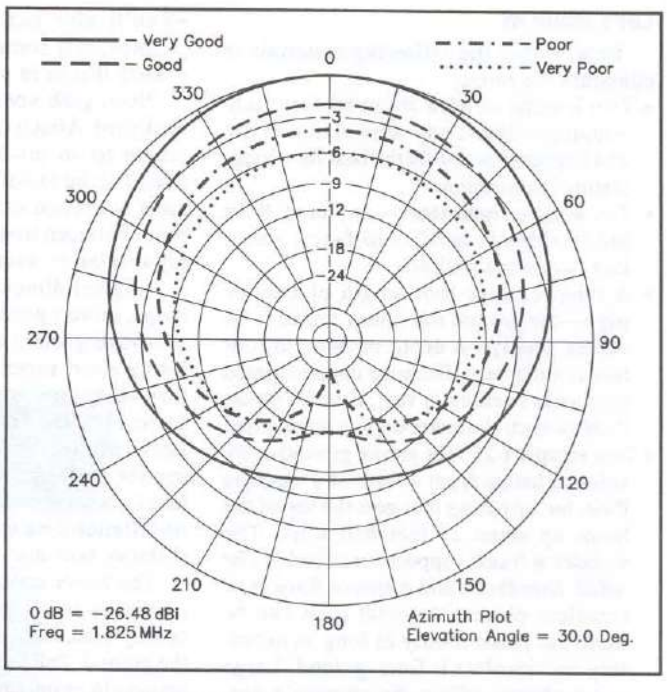

Figure 8—Horizontal radiation pattern of the loops at 30° elevation.

ing the coax on the ground (buried is better) to minimize pickup of noise or energy coupled from your transmit antenna. A receiving preamplifier is almost certainly needed, either your rig's internal preamp or an external preamp.

Evaluation and Adjustment

Next, verify that the antenna is working correctly. Some listening is probably what you'll do first, but fading makes it almost impossible to determine the antenna's actual performance. At best, you will be able to confirm that the antenna has reasonable directivity.

More accurate listening tests can be performed several ways. The best way is to enlist the aid of a nearby ham whose station is very close to being in line with one of the two loops. If such help is not available, the next-best option is to identify a local AM radio station high in the band (1400-1600 kHz) and use that as your test signal. Switch the loops to their various directions. If the test station is directly in line with one loop, you should see a front-to-back (F/B) ratio of about 2 to 3 S units as the antenna is switched toward and away from the station. You won't see a *huge* F/B because the deepest null is up at 30 to 40°.

Local ground conductivity can affect performance. You might not get an optimum pattern at your particular location with the "normal-value" terminating resistor. If you aren't getting the expected performance, substitute a 1 kΩ potentiometer for the terminating resistor and adjust it for best F/B while listening to your test station. Then, measure the pot's resistance and install a fixed-value resistor of the same value.

Antenna Performance

The vertical pattern of the antenna, in line with the loop, is shown in Figure 7. Figure 8 is the azimuth pattern at 30° elevation. Modeling was done using W7EL's *EZNEC* program,[3] which uses *NEC-2* to evaluate the antenna over "real" ground.[4] Listening tests confirm that these modeled patterns are close to the as-built antenna performance. Much of the on-air pattern evaluation was done by listening to AM broadcast stations and to WWV on 2.5 MHz, because the pattern of the antenna changes little over this frequency range. Stations at exactly the right distance and directions have verified the deep null off the back. For example, New York City area radio stations are exactly in line with my northeast/southwest loop, and show 40 dB F/B ratio when the arrival angle is just right!

The front of the pattern is quite broad, with maybe one S unit F/S. The advantage of this antenna is its rearward null, which reduces local noise and distant QRM. I installed my system just 2½ weeks before the 1996 ARRL 160 Meter Contest, so it got a thorough evaluation in a short time. During the contest, changing the pattern direction often made the difference between Q5 copy and a busted contact. For example, pointing southeast to hear Caribbean stations reduced stateside QRM by 2 or 3 S units, enough to easily hear the DX through unruly pileups.

Higher-Performance Ideas

The compact size of the K9AY Loop makes it easy to build more than one for a multielement array. Installing two of them in a broadside/endfire combination is one option, but a really ambitious approach would be to install a four-square! This array would have a substantial F/B over a wide angle, along with a much narrower front lobe than a single loop. The space required for this high-performance array is far less than what is needed for Beverages. If you try this, remember to switch the individual loops to the desired direction as you switch the feed system.

Summary

If you want improved reception on the low bands and don't have a lot of room, the terminated loop is an excellent choice. It is small, easy to build and its directional pattern makes DX much easier to hear.

Notes

[1] Floyd Koontz, WA2WVL, "Is This Ewe for You?" *QST*, Feb 1995, pp 31-33.

[2] Floyd Koontz, WA2WVL, "More EWEs for You," *QST*, Jan 1996, pp 32-34

[3] *EZNEC* 1.0 by Roy Lewallen, W7EL, PO Box 6658, Beaverton, OR 97007; tel: 503-646-2885, fax: 503-671-9046.

[4] The modeling accuracy of antennas with a direct connection to ground is somewhat uncertain, even with the power of *NEC-2*. The high accuracy ground characterization of the Sommerfield method gives answers that vary with the number of segments chosen for each wire. The *MININEC*-type ground model gives more consistent results, but on-air tests show that the deepest null is obtained with a lower value terminating resistor than this model indicates. The plots presented here are the "best fit" between modeled and observed performance.

Photos by the author.

From *QST*, May 1998 (Technical Correspondence)

Hum Problems When Switching the K9AY Loops

By Gary Breed, K9AY

I am pleased to report that many hams have successfully built the receiving antenna described in the September 1997 issue of *QST*.[1] I had some concern that variations in local ground conditions and nearby structures could reduce the antenna's performance, but the performance I obtained seems to be readily duplicated.

There is one problem that needs to be addressed. When the loops are switched to the northwest direction, an ac voltage is sent down the coax to the relay box. A few hams have reported hum or distortion when the antenna is switched in this direction. Two explanations are possible: a ground loop due to widely separated antenna and station grounds, or modulation of the core of the matching transformer. The presence of hum was not evident in my prototype, but to avoid either cause of the problem, I recommend using a *separate* three-conductor control wire to operate the relays. The power is solely 12 V dc, and no current flows through the transformer. The modifications are shown in the accompanying figures, which can replace Figures 5 and 6 in the original article.

Secondary benefits of this arrangement include the option of using an existing 12 V power supply to provide the operating voltage, and the ability to use an additional conductor to carry power to an antenna-mounted preamplifier.

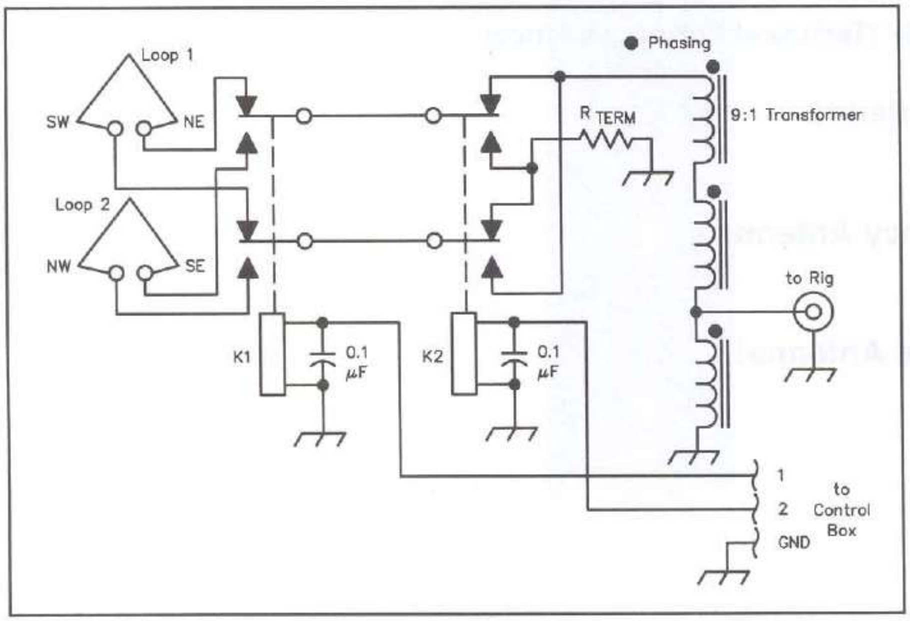
Loop 1
SW
NE
Loop 2
NW
SE
● Phasing
R TERM
9:1 Transformer
to Rig
K1
0.1 µF
K2
0.1 µF
1
2
GND
to Control Box

Figure 1—A revision of the schematic shown in Figure 5 of the September 1997 *QST* article.

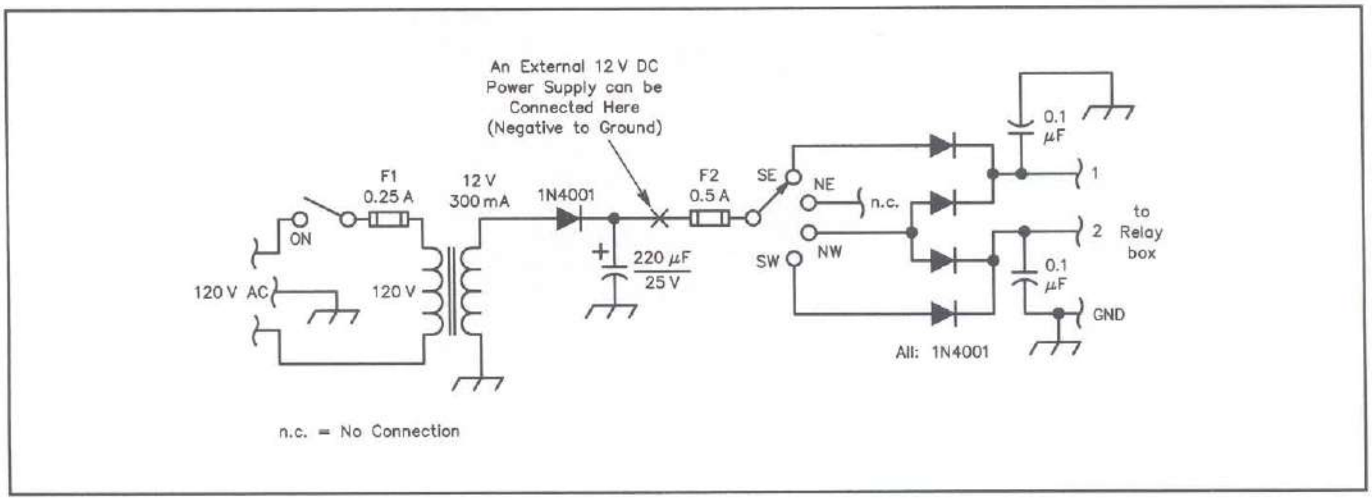
An External 12 V DC Power Supply can be Connected Here (Negative to Ground)
F1 0.25 A
ON
120 V AC
120 V
12 V 300 mA
1N4001
220 µF 25 V
F2 0.5 A
SE
NE
n.c.
SW
NW
0.1 µF
0.1 µF
1
2
GND
to Relay box
All: 1N4001
n.c. = No Connection

Figure 2—Changes to Figure 6 of the original article.

---

[1] Gary Breed, K9AY, "The K9AY Terminated Loop—A Compact, Directional Receiving Antenna," *QST*, Sep 1997, pp 43-46.

CHAPTER NINE

MISCELLANEA

9-1 **A Skyhook for the '90s**
From May 1997 *QST*

9-4 **Balloon Skyhooks**
From August 1997 *QST* (Technical Correspondence)

9-5 **The Clothesline Antenna**
From July 1998 *QST*

9-8 **A Variable-Frequency Antenna**
From July 1949 *QST*

9-12 **Honey, I Shrunk the Antenna!**
From July 1993 *QST*

By Don Daso, K4ZA **From *QST*, May 1997**

A Skyhook for the '90s

Is your group looking to put up a BIG vertical antenna for Field Day? The crew at N4ZC started their own air corps. Their full-size 160-meter quarter-wavelength vertical is *not* supported by hot air!

Some History, as Introduction

By 1936, the Zeppelin company was without rival in the design, construction and operation of rigid airships. Flying successfully around the world, providing regular passenger service between Germany and South America, the *Graf Zeppelin* airship was nonetheless unsuited for the hardship of crossing the North Atlantic. The *Hindenburg* was designed and built for this route. This huge airship was one of the largest flying machines the world has ever known. Her crash and destruction at Lakehurst on May 6, 1937, ended the era of passenger transport by airship.[1]

Later, growing up on a farm near Akron, I sometimes saw the Goodyear airships passing silently overhead. Twenty-odd

PHOTO BY SCOTT DOUGLASS, K2SD

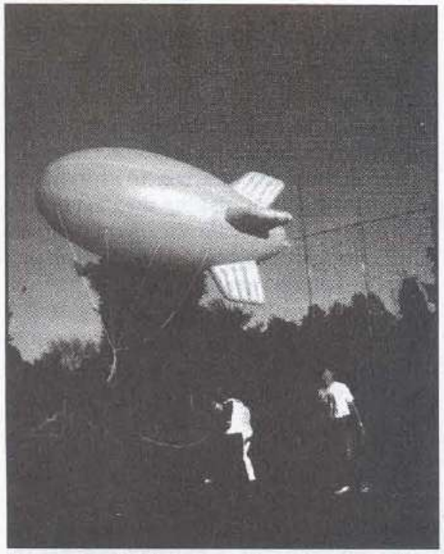

Figure 1—K4ZA fastens the antenna and tether to the miniature blimp prior to launch.

PHOTO BY DON DASO, K4ZA

Inspiration, on a grand scale.
What if we had one of *these* to play with?

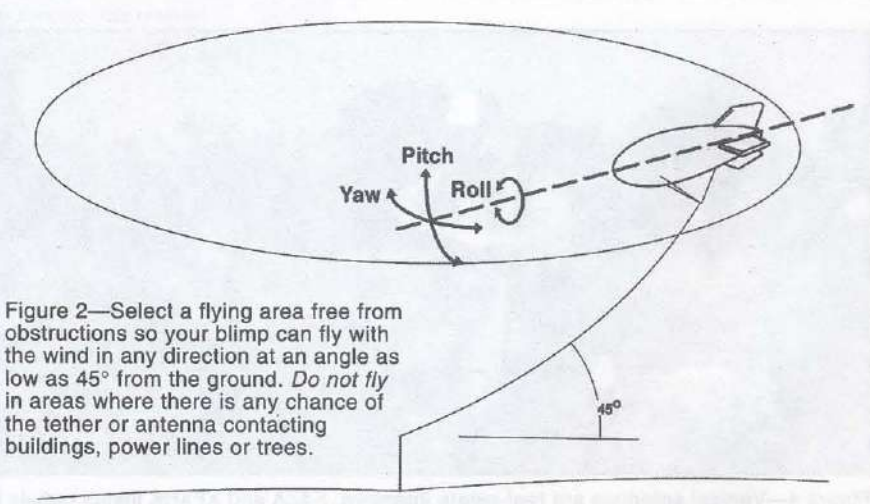

Figure 2—Select a flying area free from obstructions so your blimp can fly with the wind in any direction at an angle as low as 45° from the ground. *Do not fly* in areas where there is any chance of the tether or antenna contacting buildings, power lines or trees.

---

[1] The terms concerning lighter-than-air (LTA) craft are often misapplied. *Webster's Ninth Collegiate Dictionary* indicates that *dirigible*, *airship* and *blimp* all denote an LTA with *steering and propulsion* capabilities. Hence, the terms do not strictly apply to the kytoon, which has neither steering nor propulsion systems. Nonetheless, *blimp* is so commonly used by manufacturers, sellers and users to describe the kytoon that we continue that usage here, as well. —*Ed.*

PHOTO BY DON DASO, K4ZA

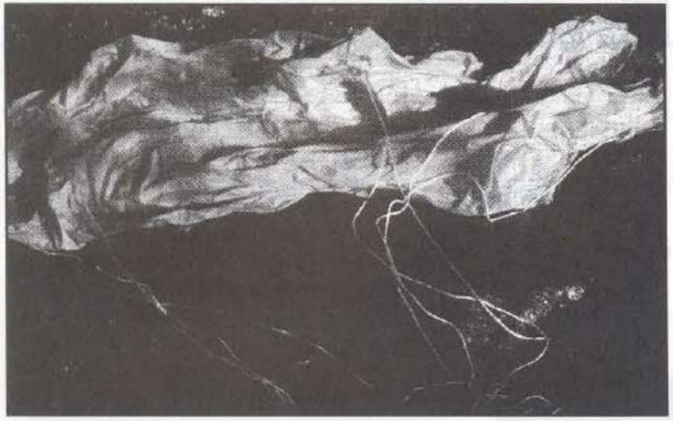

Figure 3—Here's our miniature version of the real thing—fresh out of the box and ready for inflation.

years after the *Hindenburg* disaster, I began to experiment with balloons myself—originally using surplus weather balloons to support wires, and more recently, with small blimps holding contest antennas. Many hams have considered using balloons or kites for antenna supports. Indeed, they come closer than anything else to that "skyhook" we all jokingly refer to on occasion. Yet, each one has limitations.

The Kytoon

Balloon-supported antennas tend to lose altitude in the wind, however slight. Kites cannot fly without wind. Satisfactory service from either support requires a limited range of wind and weather conditions—something that's not too common. Combining features from each—the balloon and kite—solves several of these limitations. Indeed, the Kytoon is a small, inflatable blimp-shaped balloon with kite-like surfaces that cause it to fly into the wind. An early Kytoon—with a rigid skeleton—is registered with the US Patent Office and enjoyed some success as a skyhook. (See Bibliography for further information.)

Today, the early Kytoon's bladder-and-frame construction has been replaced by completely inflatable balloons, made from 2 to 3-mil polyurethane film (see Figure 1). Brilliant skin colors are common because these miniature blimps are intended for advertising purposes. These small blimps are available from various sources, with prices varying according to size—typically from 10 to 30 feet long and starting at $300. [Beware: Many suppliers specialize in advertising Kytoons and sell them for twice the price of those sold for scientific uses.—*Ed.*]

The plastics used in their construction have been specially formulated for high elasticity, helium retention and paint application. Helium (a safe, inert gas) is available from party suppliers or welding supply companies. One tank will easily support an antenna for the typical contest weekend. If you have some room, but maybe not the resources to put up a full-sized radiator, consider a tethered Kytoon. As a modern antenna support, these miniature blimps are the skyhook of the '90s.

Tethered Flight

Most hams considering these small blimps as antenna supports concentrate on their antennas. This natural inclination can create problems. Most hams have little knowledge or experience with the basic principles of LTA flight, knots and rigging rules, with FAA regulations, with weather and especially with how these all relate to tethered flight. It's a complex process and serious business to lift an antenna aloft with a small balloon or blimp.

These blimps consist of a closed envelope of gas—usually helium—that is lighter than the atmosphere. Gas balloons can only descend by losing lift (that is, by venting gas) or when some force overcomes their lift (retrieval by tether or powered flight). As a gas balloon rises, it will reach an equilibrium altitude, where it will remain until the lift-to-weight ratio changes. These principles—which are mostly associated with larger, cargo-carrying blimps—are important, especially because our balloon will be tethered. As with any tower or antenna project, use common sense and make safety your highest priority!

Because space is three-dimensional, any "skyhook" can turn in the air—rotating in three axes: a longitudinal roll, a lateral yaw and a vertical pitch. The shape of these blimps allows them to "fly" in the wind, countering these motions, and the tether allows the blimp to move easily enough within those three axes. Generally, you will experience something like Figure 2.

A 10 to 12-foot blimp is sufficient for most antennas suitable for low-band use. Most blimp applications will be for 160 meters, because the sheer size of effective antennas for this band makes them difficult (and costly) to erect. A 10 to 12-foot-long blimp of 3 to 4 feet in diameter holds about 70 cubic feet of helium and is capable of lifting two pounds. This is sufficient for an effective $\lambda/4$ wire vertical for 160 meters.

Setup, Antenna and Ground

You'll need a working area on the ground to inflate your blimp. A 20×20-foot area should suffice. It's important to protect the thin skin of the balloon. I usually lay the balloon out on a tarp to prevent any stubs or twigs from poking the plastic skin (see Figure 3). While the plastic seems sturdy and elastic, I'm always careful and treat the balloon gently. I hold the helium tank upright, and I usually lash it securely to a two-wheeled dolly, so I can move it to the field easily. Follow the manufacturer's inflation directions carefully (each will be slightly different).

Inflation temperature is important: If possible, inflate the balloon at the same temperature it will experience in flight. Helium comes out of the tank cold. With rising temperatures, it expands quickly inside the balloon. Sunlight and warm air cause further expansion. (In very hot weather, a balloon loses some of its lifting power, but this is usually not a problem.) Do not overinflate the balloon. Finger pressure will barely dent the surface of a properly inflated polyurethane blimp. Pay special attention while closing the neck of the blimp. With many blimps, you can simply fold the neck back on itself several times and then fasten it in place with heavy rub-

PHOTO BY SCOTT DOUGLASS, K2SD

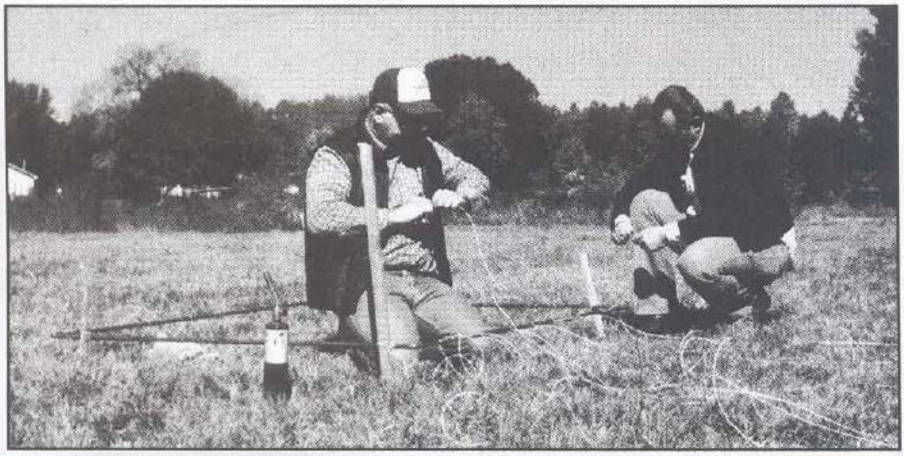

Figure 4—Vertical antennas are real-estate intensive. K4ZA and KF4HK install radials in the horse pasture at N4ZC's contest station.

PHOTO BY SCOTT DOUGLASS, K2SD

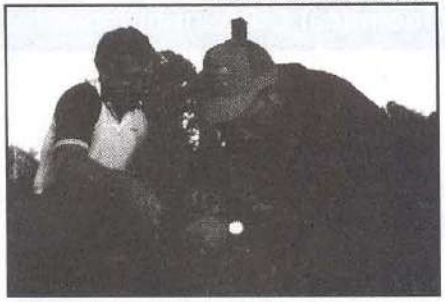

Figure 5—WA4UNZ and K4ZA work inside the large copper ground ring installing the feed system.

ber bands (folding them, if necessary) to provide a tight, leak-free fit. Refills are much easier when rubber bands, rather than knotted rope, seal the balloon.

Use a quality ball-bearing swivel between the end of the tether and the balloon's attachment line. This allows the blimp to easily turn into the wind. Fishing line works well as a tether; so do some kite lines and other synthetic ropes. I've used 100-pound fishing line and 1/8-inch nylon rope.

Stranded wire is a lightweight antenna conductor. Aluminum welding wire is light and strong enough to use. (I use phosphor-bronze wire that was originally intended for use with the "Gibson Girl" rescue unit's kite- or balloon-supported antenna. It comes on a small reel. The irony, weight and low cost appeal to me.) To save time, measure and mark the antenna length before construction begins. I have found no instances where the addition of a tether created a problem, in launch, retrieval or during flight. The antenna has never broken; and our balloon has never broken free. Obviously, it's simple insurance, and you don't want your antenna wire touching power lines or other hazards, which might happen with a runaway blimp.

The original installations at N4ZC were all within a working pasture (see Figure 4). On-ground radials were easier to install than elevated radials, albeit more labor-intensive (easier because they need no supports, labor-intensive because it requires more radials—4 versus 60). A large diameter ring of copper water pipe served as the radial connection point and made soldering easier. The ring also created a work space to stand in during set up and launch (see Figure 5). (Future installations will use a high, central wooden post, with attachment points for the antenna, tether and four elevated radials.) Our installation is over 150 feet from the nearest tower (the 20-meter array). RG-213 runs from the feed point to the base of that tower, where it connects to Hardline running to the shack.

Hints

Always wear gloves when launching and retrieving the blimp. Let the blimp rise slowly; retrieve it slowly. The lift can be terrific, and the antenna wire and tether can easily cut your hands. Always have helpers coil up the antenna and tether on spools as you walk the balloon down. (I recommend reeling in the blimp during daylight hours. It's a tempting target, and we hear gunshots in the woods all the time at N4ZC's somewhat country location. Make certain winds cannot push the blimp into nearby objects, even when it's near the ground.)

FAA regulations for flying tethered balloons—those under five pounds in weight—are clear: No person may operate a kite or balloon in a manner that creates a hazard to persons, property or other aircraft.

Avoid all dangerous situations: operation near airports, in wet or stormy weather, near electric power lines, over public streets or areas congested with people or at extreme heights (which is probably the first thing a ham will think of). Use common sense and play it safe. Using less than the legal limit with a single 1/4 vertical supported by a 10-foot blimp, I've been able to work any station heard on 160 meters.

Sources

Here are names and contact information for two suppliers of these inflatable blimps:

Toy-Tex Novelty Company, 7315 N Linder, Skokie IL 60077.

The Blimp Works, 156 Barnes Airship Dr, Statesville NC 28677.

Annotated Bibliography

There's a wide variety of applicable information and material available, but not much has appeared in Amateur Radio publications. This is new territory, in many ways, that's worthy of further experimentation. I recommend the following books and articles for further reference and study—before buying, building and flying:

R. Carleton Greene, W8PWU, "More On Balloon Supported Antennas," *QST*, Nov 1940, pp 38, 39 and 82. One of the earliest articles on using meteorological balloons as antenna supports, complete with lifting charts.

David T. Ferrier, W1LLX and William G. Baird, W9RCQ, "A New Kind of Skyhook," *QST*, Oct 1946, pp 24-25. The original use, as best I can determine, of the Kytoon as an antenna support. Although focused on emergency or Field Day use, the final line of the article presents this challenge: "Perhaps long-wire vertical antennas can now be exploited with outstanding results."

Stan Gibilisco, W1GV, "Balloons as Antenna Supports," *The ARRL Antenna Compendium*, Volume 2, (Newington: ARRL, 1989). A brief overview of the topic, with good descriptions and several safety tips.

Maxwell Eden, *Kiteworks*, (New York: Sterling Publishing Company, 1989). An excellent book about kites of all kinds—including building and flying them. It's filled with tips and techniques.

Will Hayes, *The Complete Ballooning Book*, (Mountain View, California: World Publications, 1977). Although focused on large hot-air balloons, the history, basics of flight and safety chapters are well worth your time. It also includes the FAA regulations.

John Belrose, VE2CV, "A Kite-Supported 160- (or 80-) Meter Antenna," *QST*, Mar 1981, pp 40-42. A unique application of the Parafoil—a special type of kite. Interesting tips and techniques applicable to blimp-supported antennas.

"160-Contest Results," *QST*, Jun 1976, pp 71-74. Pictures and text from W8LT, the Ohio State University Amateur Radio Club, using an original Kytoon to support a vertical antenna.

Balloon Skyhooks

By Jack M. Hughes, WB6SOI

I appreciated and enjoyed "A Skyhook for the '90s."[1] The company I work for, TCOM, L.P., has been doing this sort of thing commercially since 1981, but on a much larger scale. Since 1972, TCOM, L.P. has been an authority on and builder of aerostats.[2] We manufacture and market an aerostat system called the Tethered Aerostat Antenna Platform (TAAP).

VLF and LF transmitters and receivers are used by the military and governments worldwide for high-reliability communication of strategic information. A typical antenna system for VLF/LF communication is a complex array of wires suspended on insulated towers that take months to erect and place into operation. A unique scheme for the rapid deployment of a complete VLF/LF (up to 100 kW) communications system uses the TAAP.

An aerostat is employed to elevate an encapsulated vertical antenna/tether to a height approaching one quarter wavelength for the frequency in use. At VLF, this technique enables the system to operate with a high antenna-radiation efficiency unobtainable by other land or sea-based antennas.

The radiated-power efficiency is further enhanced by the installation of a simple groundplane consisting of eight radial wires emanating from beneath the mooring-system trailer. The radial ends are secured by ground rods and copper wire. The entire system—antenna/tether, aerostat, mooring system, transmitter and equipment shelter—is designed to be transportable by land, air or sea and to be rapidly erected at a selected site.

The aerostat is a helium-filled, aerodynamically stable tethered balloon typically 32 meters long. Its antenna/tether is composed of Kevlar with optional fiber optics and two concentric outer sheaths of aluminum braid. The Kevlar construction provides up to five times the strength of a similar-weight steel cable. A fiber-optic core provides for secure communication between the ground and the aerostat, and the double-aluminum sheaths act as the radiating elements—as well as a path for the 400 Hz electrical power to the aerostat.

The aerostat is launched and retrieved at a rate of 200 feet per minute and it does not vent helium. In addition to the helium chamber, it uses an air-filled ballonet system that expands and contracts depending on the outside air pressure to keep the aerostat rigid at all times. It can operate in winds of up to 50 knots and will survive in winds of greater than 70 knots. The system is designed to fly continuously for up to 21 days before replenishing it with helium.

I commend author Don Daso for his creativity and resourcefulness. He seems to have just as much of a thrill as we do in handling LTA vehicles.

---

[1] Don Daso, K4ZA, "A Skyhook for the '90s," *QST*, May 1997, pp 31-33.

[2] Aerostats are essentially helium-filled, aerodynamically shaped balloons that are tethered to ground. They're not cheap—a 15-meter-long Aerostat system costs about $175,000.

By Robert Victor, VA2ERY **From QST, July 1998**

The Clothesline Antenna

Dry your laundry or work DX. Could this be the first dual-purpose antenna?

Every once in a while, you run across an idea that seems so simple, so obvious, you can't believe it hasn't been done before. Surely (you say to yourself), you're not the first person in the universe to have thought of this...

That's the case with the Clothesline antenna and me. While mulling over a variety of ideas for an antenna suited to my apartment, I started drawing some graphs of sine waves at various frequencies. One thing led to another and I wound up with a terrific solution for my antenna needs—and it was one that I haven't found any references to anywhere. As far as I'm concerned, *I've* invented the Clothesline. Still, I'm not going to be surprised if someone shows me that it's been done before!

But even if this antenna design has been around since hydrogen, it may still be new to you. And even if you have seen something similar to the Clothesline, this design may be worth a second look. For hams in a variety of situations, this could be just what the doctor ordered to cure your DX dilemma.

What It Is ... And Isn't

I will tell you what the Clothesline is, but first I'll tell you what it is not. The Clothesline isn't some dubious trick for loading up an actual clothesline with 12 cubits of RG-213 looped four turns to the foot around your washing machine. This is an antenna, it works all the HF bands, and it gets great results. It doesn't need a tuner because it's dead-on resonant on the 160, 80, 40, 20, 15 and 10 meter bands. A tiny tweak will bring in 12 and 17 meters, too.

It's remarkably easy to build. There are no traps, no stubs, no loading coils, no variable or fixed capacitors, no screws, no clamps, and you don't have to drill anything. This antenna is so simple to put up, it hurts. The Clothesline consists of little more than a piece of wire, a center insulator, some feed line, and a couple of $2 hardware-store fittings.

Too good to be true, you say? What's the catch, you ask?

Well, you *do* have to adjust this antenna for most band changes. But before you get too excited by the word "adjust," I'm not talking about unwrapping yards of black tape to get at a loading coil, or tweaking the bare shaft (whoops, hot side—sorry!) of a 50-year-old fleamarket variable capacitor. No, the Clothesline makes band changes a piece of cake. When you see how easy it is, you're going to chuckle!

BEASLEY
KG6BJH

THAT LAST PAIR OF LONG JOHNS YOU REELED OUT MOVED MY FEEDPOINT

The Concept

The easiest way to explain how the Clothesline works is to lead you through the simple reasoning I used to come up with the thing. I started with a drawing like the one in Figure 1. It shows a simple dipole for 80 meters. It's 132 feet long and fed in the middle with coax. Next I drew the graph of the voltage distribution along the antenna (at its fundamental resonant frequency). Note the voltage is at a maximum (with reference to ground) at the antenna ends. That's important for any dipole radiator. The ends of the antenna have to correspond to the high voltage points of the curve.

I already knew that the curve the voltage wave describes is a quarter of a full cycle—a quarter wavelength—on each half of the antenna. And that the distance between the two places where the voltage is highest, or in other words, from one end of the antenna to the other, is $^1/_4$ plus $^1/_4$—a half wavelength. That's why they call it a half-wave dipole.

I also understood why the feed point is in the middle. It's located precisely where the voltage curve is lowest with respect to ground. To me, an antenna is just like a transformer—it transforms the low impedance coming out of the coax up to the impedance needed to couple to the cosmos. Come in low, go out high. So much for the plain dipole.

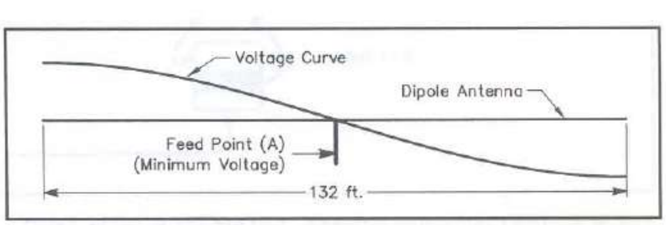

Figure 1—The voltage distribution along an 80-meter dipole antenna. Note that the voltage is at maximum (with respect to ground) at the ends of the antenna.

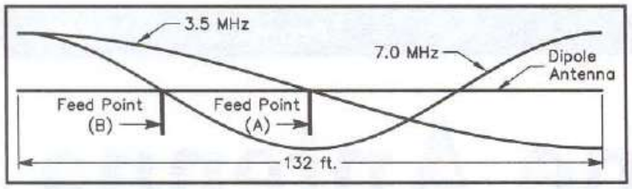

Figure 2—Take the 80-meter dipole shown in Figure 6 and overlay the voltage distribution for a 7-MHz signal. Note how the 7-MHz feed point (B) has shifted to the left of the 3.5-MHz feed point (A).

I then dropped a 7-MHz voltage curve over the original 3.5 MHz length (see Figure 2). Because the new frequency is a whole multiple of our original frequency ($3.5\,\text{MHz} \times 2 = 7\,\text{MHz}$), there is still a whole number of half wavelengths from one end to the other, so the voltage peaks are still at the ends. This means this same length of wire will resonate at 7 MHz, which means it will radiate perfectly well at this frequency too.

But I noticed right away that something had changed from our earlier situation—the point where the voltage minimum crossed ground was no longer in the middle. There were now two such points, neither of them anywhere near the middle. So the antenna was resonant—the voltage peaks were at the ends—but my original feed point was now useless! Alright, I reasoned. If I shift the feed line to the new minimum I'll have the perfect setup—high voltage at the ends, low voltage at the feed point. Look out DX, here I come!

I thought about how far I'd moved the feed point. Looking at the curve, you can see that from (A) to (B) is a quarter wavelength at the new frequency, right? We doubled the frequency, which means we cut the old wavelength in half. So what used to be a quarter wave at the old frequency—from the middle to one end—will be twice that, or one-half wavelength, at the new frequency. And I wanted to get the feed point over to the new location at point (B), so I move it half that half—a quarter wavelength—over to the new minimum at (B).

I also realized that I'd added a quarter wave to the right half of the antenna, making the right-hand side ½ plus ¼ wavelengths, or a total of ¾ wavelengths long. Since ¾ on the right plus ¼ on the left equals 1, I now had a full-wavelength antenna with some gain compared to a dipole, fed one-quarter wavelength from one end.

I kept going. I doubled the frequency again, and the same thing happened. The voltage peaks would stay at the ends, but I'd see the feed point moving further to the left by one-quarter wavelength at the new frequency for each time I doubled frequency.

I tried it for other multiples—by 3, 5, 6, 7 and so on. Just as I expected, I always wound up with some whole number of half waves, and voltage peaks at the ends. All I ever had to do was to move the feed point to the new quarter-wave point of the antenna. And for that matter, for any band other than 80, I had a choice of feed points. Each higher-band antenna had more than one voltage minimum point in its length, and I could feed at any one!

Now I drew a bunch of these curves like the ones shown in Figure 2 for 80, 40, 20, 15 and 10 meters. I dropped reference lines down from each voltage minimum to show where the feed points would lie along this single piece of wire, depending on the frequency. Because there are multiple feed points for each of the higher bands, I could see that any time I want to change frequency, I would just shift to the closest feed point for the next frequency, and away we go!

But so what? How was I going to make a simple antenna with all these feed points? For that matter, what was simple about an antenna with all these feed points anyway? The answer was, "nothing." What about a *single* feed point that somehow travels or slides along the antenna to anywhere I want it to go? Could I make that happen? I was looking out the window at my 40-meter folded dipole when I came up with the answer.

Other than the fact that it's fed with 300-Ω line rather than coax, the folded dipole is almost identical to the standard dipole. It resonates over the same lengths at the same frequencies, so my marvelous (but yet to be invented) multi-band, sliding feed-point antenna would be just as comfortable with a folded dipole 132 feet long as it would be with a plain dipole. But what about the sliding feed point?

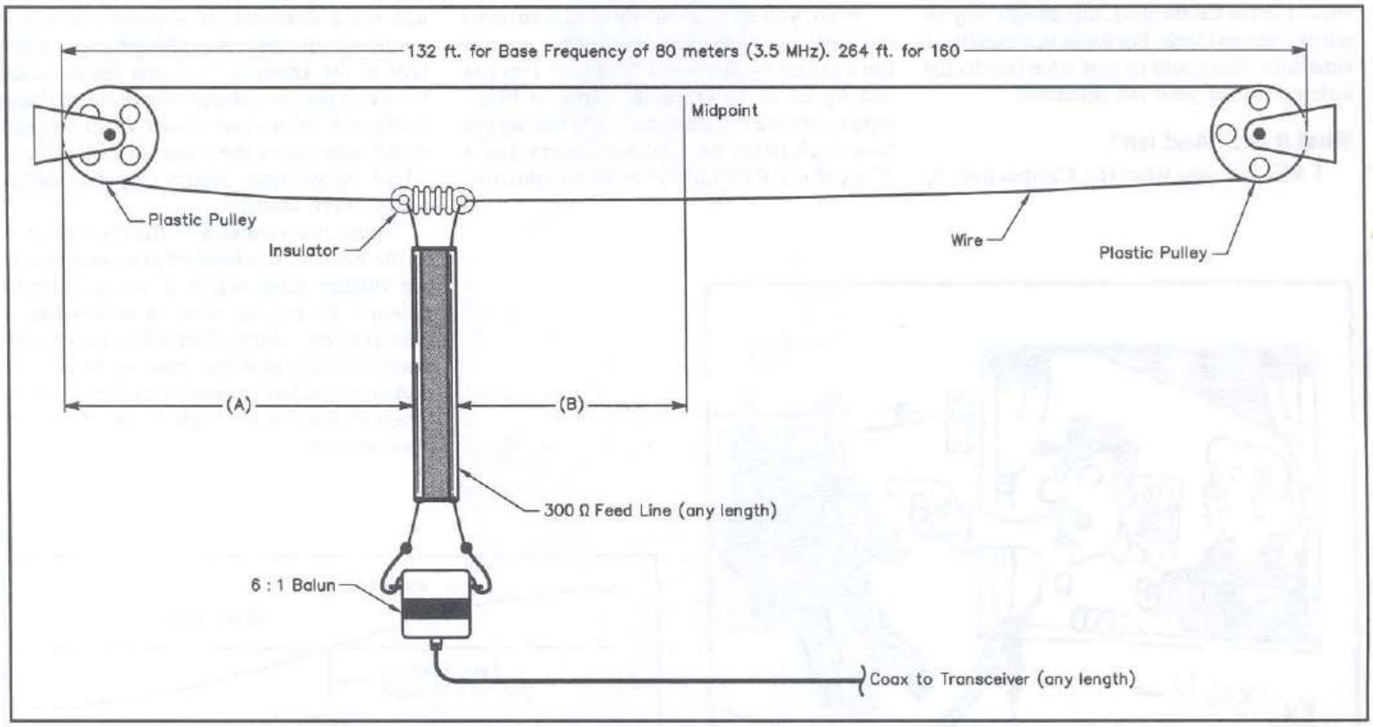
132 ft. for Base Frequency of 80 meters (3.5 MHz). 264 ft. for 160
Midpoint
Plastic Pulley
Insulator
Wire
Plastic Pulley
(A)
(B)
300 Ω Feed Line (any length)
6 : 1 Balun
Coax to Transceiver (any length)

Figure 3—The Clothesline antenna. Just set (A) or (B) to equal a quarter wavelength at your operating frequency. As long as it is a multiple of your base frequency (most HF ham bands are), your antenna is tuned. You can measure out your favorite bands, peak them by observing the SWR, and put color marks or tags on the bottom wire. Roll up to the mark and you're ready to couple to the cosmos!

What would happen with my folded dipole if I just grabbed the feed line and tried to slide it along toward one end? It hit me: If the endpoints were run over pulleys the feed point would effectively slide! It would just run along the bottom of the antenna to anywhere you cared to put it. As far as the electrical nature of the antenna is concerned, absolutely nothing has changed—except the position of the feed point! And there it was—the birth of the multi-band, sliding feed-point antenna, otherwise known as the Clothesline! See Figure 3.

But wait a minute, I thought. It's a clothesline. It can't possibly work. It's too simple!

Yet no amount of poking about on paper could find fault with it. Still very skeptical, I strung up a miniature version (for 10 and 20 meters) running down my hallway. It worked! It loaded like a breeze, with the feed about 8 feet from the end on 10 meters and about 16 feet from the end on 20 meters, just where those graphs said it should. I bought some decent wire, a couple of plastic pulleys, 75 feet of 300-Ω twin lead, and threw one up on the roof to cover 40 through 10 (66 feet is all the horizontal space I have up there). Within a couple of days I had racked up dozens of DX contacts on every band. Most remarkable was the apparent gain on the higher bands. I consistently surprised other stations (and myself) with my signal strength, usually trading equal signal reports and getting lush praise from guys running far greater power and fancier antennas. My 100 W and a Clothesline are reaching the world!

How to Build and Use the Clothesline

Measure out enough wire for twice the lowest frequency. If this happens to be 80 meters, you'll need 2 × 132 feet (plus a bit for trimming). String it up like a clothesline, using nylon or rope leaders to attach the plastic pulleys to their supports. Just make sure that all the wire is used, so that the antenna is full-length. You can use long leaders if you want to hang one end off a distant support. Run each end of the wire through the pulleys top-to-bottom, attach the two ends to a center insulator and hook up your feed line. That's it!

If you're building a version for 80 or 160 meters, you'll want to install a third little pulley to ride along the top run, attached by a short piece of cord to the center insulator, to help keep the top and bottom runs of the antenna roughly parallel. Build it for 40 as the lowest frequency and you won't even need this little extra.

There is a full set of feed points between the middle of the antenna and half-way out to either end, so if you want to keep the feed line short, you can cut it to provide just enough slack to move within this range. Alternatively, because different feed points for the same frequency can affect the gain pattern, you may want the flexibility of having a choice of feed points for each frequency, and decide to leave more line slack. Losses in twin lead are so much lower than in regular coax, you can use all you want (within reason). Do be careful to stand it off from any metal it encounters on the way into the shack.

Inside, the simplest way to match the twin lead to your rig is with a balun. I use a 6:1 balun to get the impedance down to the 50 Ω my radio likes. A 4:1 balun works fine too, though the 6:1 usually presents a better match. Both baluns are sold by a number of *QST* advertisers. Ditto for the 300-Ω twin lead.

Speaking of the shack, if yours is on an upper level at your home, consider setting the Clothesline up with one end near a shack window. You'll be able to reach out, retune (that is, haul in the line), then go back to your rig to check your SWR meter.

If you don't have this luxury, don't sweat it. I calibrated my Clothesline by setting my rig on tune-up power for each band, and going up to the roof of my apartment building to peak the antenna with a field-strength meter. I confirmed each setting by checking the SWR back in the shack. On every band I tuned for, it was well below 1.5:1. As I found each feed-point setting, I marked them with various colored indelible markers for each band. Now when I want to change bands, I just go up to the apartment roof, run the Clothesline out to the right color, and I'm tuned! It takes seconds.

I can imagine other settings where you'd run the Clothesline right outside a window, out to some convenient support. Or vertical, up or down the side of a building, a tree, or a flagpole (tie a flag to it, if you like). If you're camped out in suburbia with difficult, anti-antenna neighbors, a detachable feed line using alligator clips makes it the perfect disguised antenna. Who's going to suspect a clothesline? You could even use it during the day to… wait for it… *dry clothes!* You could even make it out of your standard, garden-variety hardware-store clothesline kit—plastic pulleys, vinyl-covered wire, the works. This makes the Clothesline the only ham antenna kit I know of that you can buy complete (minus the feed line) from your local hardware store. (In a perfect world, your local hardware store would stock twin lead.) Yes, I know I said at the beginning that this wasn't a trick for loading up a real clothesline and I've been true to my word. This is an antenna that happens to *look* like a clothesline.

I mentioned covering 17 and 12 meters. If you cut the original length to resonate just a little up from the bottom of 80, at 3.615 MHz, the fifth harmonic is smack dab on 17 meters. Cut it to 3.55 MHz and harmonic number seven is on 12 meters. In fact, these are such minor variations that you can surely find a convenient center frequency on 80 that'll put you right where you want to be on all bands. Though you don't need it, a tuner (either in or out-board) can get you right down to zero reflected power. Remember, feeding with twin lead or other balanced lines keeps your losses way down, so a little elevated SWR at the band edges is no big deal.

I'm now thinking about some kind of motor drive that'll permit me to tune up from inside the shack. My first reaction to this idea was, if I'm going to get into motor drives and the like, is that any simpler than a beam? I now think the answer is "yes." The drive would be simpler, in that I wouldn't need any position feedback—all I need to do is watch my reflected power meter to know when I've hit resonance. It doesn't need a whole lot of travel because a full range of feed points is available over just a quarter-length of the antenna. As well, few beams are all-band. And since the antenna you can put up is always simpler than the antenna you can't, for those of us in apartment settings where a beam or other, more complex antenna systems are out of the question, a motorized Clothesline might well be worth the effort. I've got a sneaking suspicion that, somewhere out there, there's a cheap, off-the-shelf drive unit with "Clothesline" written all over it!

By Ralph O. Williams, W8AJC

From *QST*, July 1949

A Variable-Frequency Antenna

One to Ten Meters with a Single Antenna System

Multiband transmitters and band-switching receivers make it easy for us to jump from band to band, but the antenna has not kept pace. Many of us are prevented from operating on several bands by the thoughts of the multiplicity of antennas demanded by the conventional approach to multiband work. Here, to go with the VFO, is the VFA — tunable to resonance from the operating position. It covers 11, 10, 6 and 2, and all the television and f.m. bands in between, with the optimum performance all along the line. Ideas for the lower frequencies are included, too.

It has been the writer's lifelong ambition as a radio amateur to have a universal antenna; one that would not only work on several bands, but also tune within a band, providing optimum operating conditions on any frequency. The memory of endless trips to the rooftop or out to the mast to lower the antenna and cut off or splice on a few inches of wire to hit a special spot in the band is still fresh in mind. The old Zepp was pretty good but it required spaced feeders and tuning at the transmitter end. The half-wave aerial split in the center for a 72-ohm transmission line required no tuning at the station end but necessitated a different antenna for every band. Often the resonant frequency of these antennas varied widely from the values indicated by the formulae because of conditions not always apparent, and under some circumstances it was difficult to get adequate loading over an entire band. As the years passed matching systems were introduced, and they, too, are usually one-band devices. With all their disadvantages the antenna systems were not too bad back in the days of crystal control, but now we have VFO and often operate anywhere within the band. It goes without saying that what we need to go with VFO is a good VFA!

The group, other than amateur, most in need of a VFA, is that vast population trying to receive the various television and f.m. channels on a single antenna. A good half-wave dipole will outperform most of the existing elaborate receiving antennas providing, of course, it is possible to adjust the antenna accurately for each channel.

For years I have been giving thought to ways and means of feeding out and retrieving wire to make an adjustable center-fed half-wave antenna. The increased use of the VFO and the advent of television and f.m. broadcasting have made such an antenna practically a necessity. The folded dipole makes it a possibility, with rigging less complicated than the dial drives on some broadcast receivers! It is the purpose of this discussion to describe a remotely-controlled center-fed antenna capable of continuous adjustment from band to band and within the bands, using an untuned transmission line.

The Folded Dipole as a VFA

The folded dipole lends itself admirably to rigging with cords and pulleys so that the length of the flat top may be varied simply by pulling down on the feeders (see Figure 1). If the spacing between the center pulleys is adjusted so that the characteristic impedance of the two parallel conductors is 300 ohms the feeders will match 300 ohm ribbon regardless of their length and the length of the antenna. In this way, we have a tuned antenna with a flat line for any frequency, accomplished without resorting to sliding contacts at any point in the system. Figure 2 shows a practical hand-operated rigging that will enable the experimenter to make a set-up and observe its characteristics. Antennas of this type may be made large or small but it is suggested

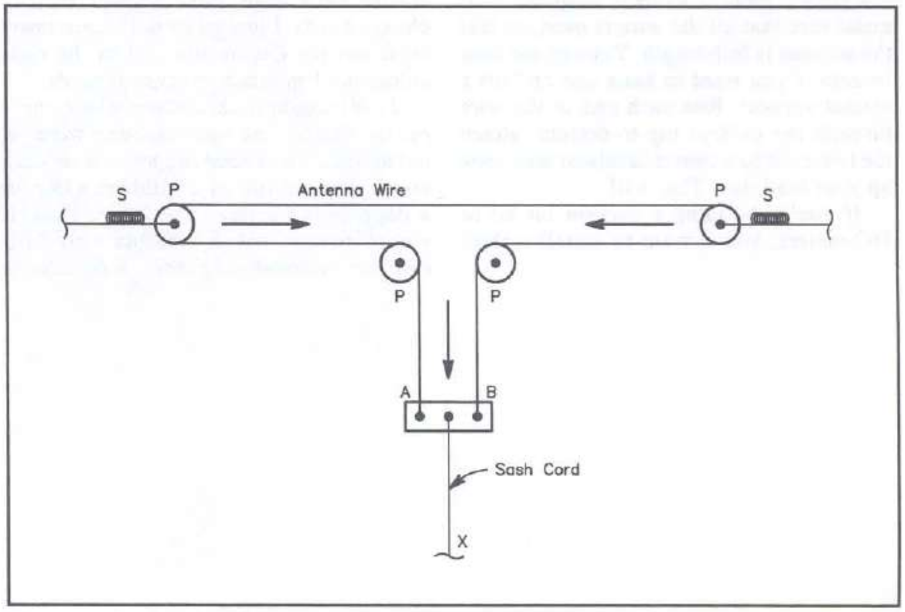

Figure 1—Basic principle of the adjustable folded dipole. Dipole and feeder section are made of one piece of flexible wire. Antenna length is changed by pulling down on the feeder at point X. Spacing of the pulleys at the center is such that the characteristic impedance of the feeder section is 300 ohms. Twin-Lead is connected at points A and B.

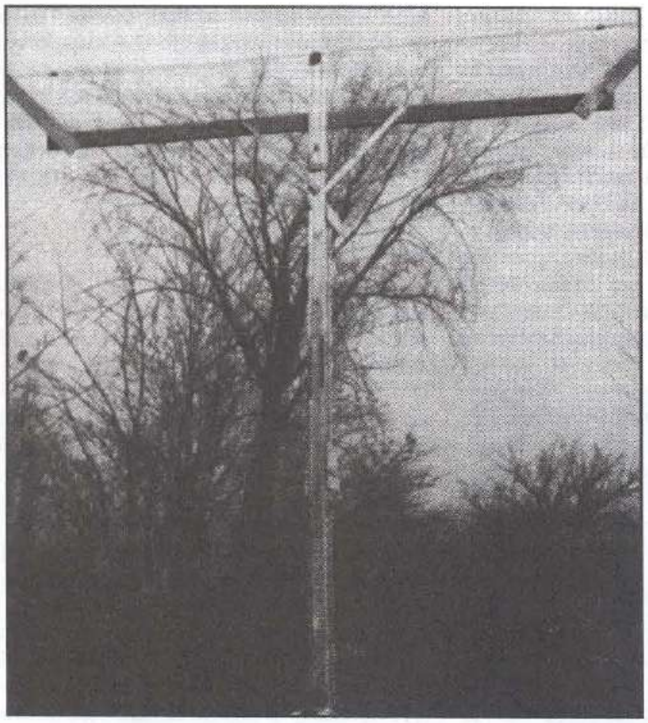

Practical working model of the folded-dipole VFA of Figure 3 used at W8AJC. The system is operated by the servo motor at the base of the mast, and is controllable from the operating position.

that the beginner make up small models for 2 and 6 meters, or simply for the f.m. and television bands, to prove the merit of the antenna.

The photograph and Figure 3 show the details of an experimental working model, made long enough to tune to 10 meters, but designed so it could be pulled down to a flat-top length of only a few inches. The tuning was fairly sharp and the results over conventional antennas for the reception of f.m. broadcasting were gratifying. A considerable improvement was noticed even within the 88-108 MHz f.m. band when the receiver was tuned to different stations and the antenna adjusted for maximum response. It was a real thrill to couple the feeders to the 10-meter transmitter and watch the plate milliammeter go up as the antenna came into resonance, and then pass through, and return for maximum; then without changing antennas, to switch on the f.m. broadcast receiver, run the antenna down to about four and a half feet to pick up a Detroit station. It was interesting to observe the effects on the received signal strength as the antenna was shortened from resonance at 10 meters to the proper length for the f.m. band, with the receiver tuned to a station on 98.5 MHz. Reception was possible with the long antenna and became good as the flat top hit fourteen feet (three half waves), falling off to a very sharp null at 121.5 inches (critical), after which it returned to full signal strength at 56.25 inches, approximately a half wave for the received signal. This ability to tune to an extremely critical null might find application in the elimination of an undesired, strong nearby signal under certain receiving conditions.

The uses to which a continuously-variable antenna may be put are limited only by the operator's imagination. Once the mechanical details have been worked out they may be used singly or in multiple, as antennas or reflectors, driven by common or separate servo motors and in various phase relations. Such antennas may be used for transmitting or receiving, or for special applications such as field operations covering a wide band of frequencies, signal-strength measurements, target transmitters for lining up rotary beams, and antenna studies. Their greatest commercial application will no doubt be in the f.m. and television fields where simplified versions, adjustable from the receiver, should find wide acceptance.

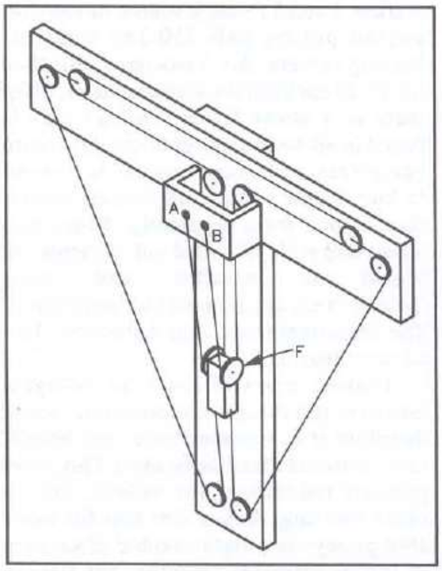

Figure 2—Hand-operated version of the variable-frequency antenna mounted on a wooden frame. To minimize mounting space, the feeders are folded back over ratio pulleys, F. The 300-ohm line connects at A and B. With this arrangement the vertical movement of pulley F is equal to the end movement of the antenna pulleys. The supporting fishline or sash cord must be taut.

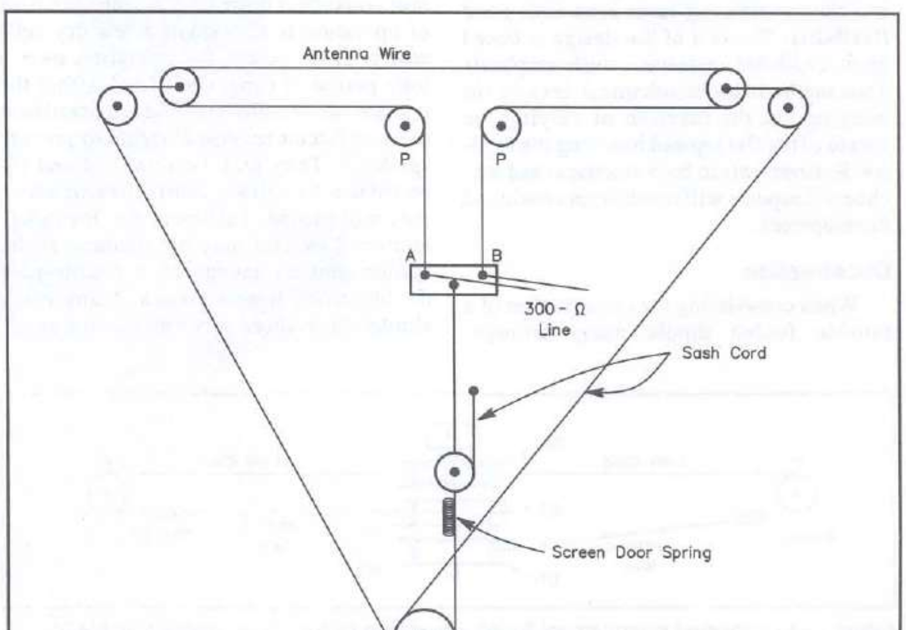

Figure 3—Diagram of the remotely-controlled antenna shown in the photograph. This system is now in use at W8AJC for 28 MHz and up. It may be reduced to a flat top of a few inches.

Other Types of Adjustable Antennas

Although the folded doublet lends itself most easily to continuous adjustment it might also be desirable to have a variable center-fed half-wave antenna suitable for use with coaxial or other low-impedance feeders. This may be done by the use of pulleys, insulators and wire, but requires a sliding contact at the center where the wind-up drums are located, and unless operated frequently it is subject to all the ills of exposed slip rings. This type of antenna is shown in Figure 4 and although at present, untried at W8AJC, it may prove worth-while on the lower frequencies, filling the long-felt need for a tunable 80-meter antenna that will work on 40 and 20 also!

Folded-Dipole Design

Useful information relative to the design of folded dipoles may be found in the ARRL *Handbook* for 1948, War Department *TM11-466*, RCA's little book *A Practical Analysis of U.H.F.* and in a paper by W. Van B. Roberts appearing in the *RCA Review* for June, 1947, page 289. The article by Roberts, in which he tells of the work done by his group at Princeton during the war under Government contract for the study of the folded dipole, is very helpful in providing the reader with a mechanism for analysis of the antenna. The literature indicates that spacing of the conductors composing the flat top should be close, in the order of 1/100 wavelength. Since our antenna is to have flexible conductors it is desirable to work with a type of wire that has a large diameter for low r.f. resistance, and at the same time has a high degree of flexibility. Such a conductor may be composed of insulated wire with a braided copper shield, the shield acting as the antenna and thus combining large area with good flexibility. The rest of the design is based upon available materials, with emphasis upon methods and mechanical devices for carrying out the function of varying the length of the flat top and handling the feeders. Refinements in both electrical and mechanical aspects will result from continued development.

Construction

When considering the construction of a tunable folded dipole many arrangements using springs or cords and pulleys will come to mind and it is up to the individual to select the method best suited to his particular use. In the beginning I tried a variety of springs, shock cords and weights to hold out the ends of the flexible antenna wire, but each of these methods had its own drawbacks and they all had the disadvantage of having to pull against a spring to shorten the antenna and depend upon the spring to pull it back out again. Metal springs come into resonance at certain frequencies. I tried metal-spring sash supports which would extend about 40 inches but they came into resonance in their extended positions and were not very smooth in operation.

It was finally decided to use cords and pulleys so arranged as to be in mechanical equilibrium and use the servo motor or other means only for the purpose of adjustment. This required less power in the servo and while it calls for more pulleys the result was smoother adjustment. I have found it convenient to support the antenna from the ends by means of insulated pulleys on a wood or other nonconducting structure. If the dipole is to be operated in a horizontal position there is no objection to using a vertical metal support pipe, but horizontal metal rods or pipes should not be used. For long antennas a center support must be provided for the feeder pulleys and the wind-up mechanism. The ends may be supported by poles, trees or buildings.

Servo Motors

The servo motor shown in the photograph is from a surplus Azon bomb tail assembly. It has plenty of power and may be reversed at will. It requires a 4-wire cable to the battery or other d.c. source. The current consumed is small and since the time of operation is also small a few dry cells will provide power for operation over a long period of time. Contained within the unit are two selenium rectifiers placed there to short circuit reverse currents to prevent sparking. They may be removed and inserted in a 30-volt a.c. line to the unit where they will provide sufficient d.c. for its operation. Reversal may be obtained at the station end by means of a double-pole double-throw toggle switch. Many other similar slow-speed servo motors are available on the surplus market, most of them reversible, and varying in size and power requirements. In some cases where d.c. is not available advantage may be taken of the gear train by connecting a universal coupling to the motor end and driving with a reversible universal fan or vacuum-cleaner motor operated from 115 volts a.c.

Sources of Materials

Antenna wire should be light, durable, flexible, of large diameter and a good conductor. For ease of adjustment it should pull around a one-inch pulley readily. Super-flexible stranded copper wire of large diameter would be quite heavy whereas an insulated stranded wire, if size 20 or so and covered by a braided tinned-copper shield, would be light in weight, adequately flexible and of sufficient diameter. Belden No. 8885 shielded grid wire having an o.d. of 0. 1 inch has been found satisfactory. Too-stiff wire will make the system unwieldy.

Pulleys must be free-running for smooth operation and have as little friction as possible. A number of different kinds of pulleys normally available at hardware stores were tried and all had mechanical imperfections. Usually, though they seemed free-running when tried at the store, they turned out to have prohibitive friction when loaded and in the system. Since the number of pulleys required is fairly large and the accumulated friction may be excessive, ball-bearing pulleys are recommended. The first ones used here were homemade and turned out of fiber and used small ball bearings in the center. The ball bearings had 1/4-inch holes and a 5/8-inch o.d. and were obtained from disassembly of surplus gear trains, bomb sights, computers and other equipment so plentiful on the surplus market. Later I found a source of one-inch aircraft pulleys (AN-210-1A) with ball-bearing centers. Air Associates sells them for $1.25 each but the surplus market offers them at a lower figure. Pulleys may be found in all sorts of surplus aircraft control equipment and sometimes it is cheaper to buy a unit containing several pulleys than to buy them separately. Homemade hardwood pulleys turned out of maple and boiled in paraffin and using 1/4-inch brass axles should be satisfactory. The important thing is to have good low-friction bearings.

Twisted rope will cause the pulleys to turn over and twist and short out the aerial; therefore it is recommended that braided sash cord or clothesline be used. This, when properly fed through the pulleys, will not cause twisting. A nice size that fits available pulleys is a light braided clothesline 1/8-inch in diameter. Of course any flexible insulating line may be used such as dial cable, fishline or upholsterer's twine. Spring loading to prevent slipping because of stretching is advisable in some cases.

Limit Switches: If the antenna is not visible from the operating point Micro-Switches may be so placed that when the end of travel of the antenna is reached the

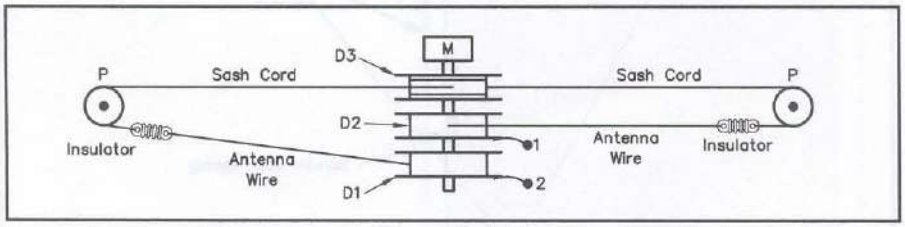

Figure 4–A suggested arrangement for an adjustable motor-driven dipole suitable for use on the lower amateur bands. Reels D_1 and D_2 reel in the antenna wire, while D_3 plays out braided sash cord. The assembly is made from a Signal Corps reel, Type RL-42-B, fitted with three reels, Type M-235, all available on the surplus market. The antenna reels are fastened to the cord reel by means of stand-off bushings secured to the center insulation. Brushes 1 and 2 connect to a 72-ohm line running to the transmitter. Success of this system depends upon maintenance of good contact at the brushes.

circuit will be opened and the motor will stop, it being possible then to reverse it and run it to the other limit where another switch will furnish protection from overtravel. The switches should not be connected in the common lead to the motor but in the circuit controlling that direction only. (Switches and associated wiring must be placed so as not to interfere with the electrical operation of the antenna.)

Conclusion

The examples shown are but a few of many possible ways of setting up remotely-controlled variable-frequency antenna systems. It is hoped that this article will serve as a basis for further development of adjustable antennae for amateur and commercial use.

By Rod Newkirk, W9BRD **From *QST*, July 1993**

Honey, I Shrunk the Antenna!

Think small! Communicate with multiconductor miniature loop antennas.

Great energy and ingenuity go into efforts to boost the effectiveness of compact transmitting antennas—*compact* meaning configurations with circumferences of $^1/_8$ wavelength or less. By this definition, a 40-meter design could be a square with 4-foot sides. In the following discussion, I depart slightly from the square. I prefer $3^1/_2$ feet high by $4^1/_2$ feet wide, suspended in the vertical plane, for convenience in my antenna's location.

The traditional small-loop approach is to insert capacitance in series with a one-turn loop, tune it to the desired frequency, and then attempt to feed power to it as efficiently as possible–no simple matter. The antenna's Q is astronomical, current and voltage are monstrous, bandwidth is razor sharp, and its radiation resistance is ridiculously low. This lossy situation can be improved by reducing the resistance of the loop material through the use of piping, foil, etc, but the resulting plumbers' nightmares are hardly worth the pains.

A more promising avenue toward practical, simple compact antennas would appear to be in the direction of varied configuration. The goal is to "cool" the system with better power distribution, and at the same time simplify the feed. For example, vast improvement in the key parameters is possible by application of the venerable "folded dipole" technique. This can be done by adding one more turn, tuned identically, connecting the two turns *in series*, and re-resonating the antenna. Now, looking into one of the twin current nodes, the input impedance will have roughly quadrupled. The one-turn loop's skimpy 5 or 6 Ω rises to approximately 25 Ω. Adding yet another such turn doubles this to about 50 Ω–a convenient match to common coax. See Figure 1. As with the folded dipole, more than four turns are unprofitable; diminishing returns raise the input impedance. little faster than ohmic resistance. Note that the tuning capacitances, each about 40 pF, must be kept nearly equal for overall system balance.

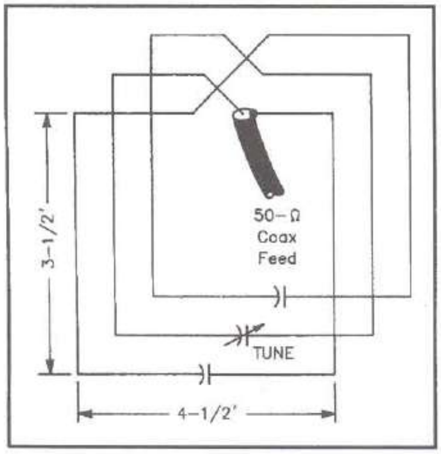

Figure 1—The three-turn loop antenna for 40 meters. Each capacitor is about 40 pF to resonate and match the antenna to a 50-Ω feed line.

Increasing Multiturn Loop Efficiency

Lots of hams use small loops in situations where they can't erect a larger antenna. Making contacts with these antennas largely depends on high radiation efficiency. Here are some things you can do to increase efficiency and effectiveness of such antennas:

- Use large-diameter copper wire. Keep the copper surface clean. Tin-plated wire is noticeably worse than copper in this application because of skin effect–most of the RF travels in the tin, which is far lossier than copper at RF.
- Use low-loss capacitors. The key here is minimizing dielectric loss, which is most favorable in vacuum- and air-dielectric capacitors. You can also parallel several capacitors to decrease losses.

Increasing the wire size used in these loops raises the antenna Q, narrowing the bandwidth and increasing the voltages and currents present. For this reason, you may have to use higher-quality capacitors than those used in the experimental versions if you increase the wire size.

Remember that high voltages can exist on a multiturn loop antenna's wires at power levels of a few tens of watts or more. Thus, it's best to keep the antenna well away from other objects.—*Roy Lewallen, W7EL*

Construction techniques, conductor material, insulation and components now become less critical because voltage peaks and current nodes are distributed throughout the system. I use ordinary zip cord, two lengths entwined, giving four available turns.[1] For proper continuity, the four wires on each side of the loop are clearly tagged at alligator-clip connectors, top and bottom centers. For 7 MHz, one turn is left open (at top *and* bottom), but will become useful on the lower bands as discussed later. A more classy construction might feature braided hook-up wire of contrasting colors. Small cardboard spreaders at the clips keep adjacent capacitors separated.

Midget receiver-type capacitors are adequate for CW power output up to 100 watts. One of these should be conveniently variable, preferably a double-gang or splitstator type, for tweaking the system to the desired center frequency as indicated by a 1: 1 SWR. At 7 MHz, each capacitance is about 40 pF in my antenna, so the old war surplus Hammarlund APC midgets do nicely. Bandwidth can approach that of a normal linear dipole if a slightly higher center frequency SWR is accepted by stagger-tuning.

Excellent 40-meter results indicated that the same loop, dimensions unchanged, could have a shot at 80 meters. At 3.5 MHz, it becomes truly a miniloop, only $^1/_{16}$ wavelength in circumference. Adding the fourth turn (Figure 2) raises the feed impedance at a slightly higher rate than ohmic resistance, so I included it. Armed with a bargain bag

The Ingredient Called *Belief*

Even though my busy family life keeps hamming time rare, the time I spend hamming has to be spent *right*. The idea isn't just to get on the air and contact anyone anywhere—that's too *easy*. The idea is to do something a hair off-the-wall and see if anyone shares the same orbit. So I'm one of those guys who calls CQ QRP for ten minutes at 10149.5 kHz even as game clogs the band's low end—someone who looks for (and *expects*) ragchews on 3560 at noon. And so I naturally was the guy to call a 5-watt, 3 × 2 × 1 CQ at 7138 kHz at around 10 PM one night last fall when the broadcasters had long since muscled straight-thinking NovTechs off to safer terrain. And so I naturally was *of course* immediately answered by W9BRD.

It figures. We've worked on schedule many times before, but he and I—he's the father; I'm the son—have never had to prearrange each new "first contact" after changes of station setup or locale. *Chance* is too facile a word for this, but then I haven't looked too hard for a better word, either. I don't need to; we just do our individual quirky things and click.

This time, we *both* had new setups: I'd just relocated the armful of gear constituting my "shack" to the bedroom and was talking to a low dipole through 100+ feet of mongrel coax; *he* was trying out an indoor, largely below-ground loop antenna—one of the versions shrunken in this article—with just a few tens of watts, tunneling past those steamroller broadcast signals as soundly as any medium-scale ham signal from Chicago.

Enjoying the magic of yet another first contact, I hardly gave his setup a second thought. But you may wonder if such antennas really work. Gooch's Paradox explains it thus: "RF gotta go *somewhere*." Seeking to soothe the linearless, towerless masses who toil with "*only* 100 watts," Newkirk's Tiresome Chant puts it, "Successful MF/HF radio communication may proceed at astoundingly low received-signal strengths." Reduced to generic essentials, it's this: The most important single factor in success with *whatever* you use is overcoming your own disbelief.—*David Newkirk, WJ1Z*

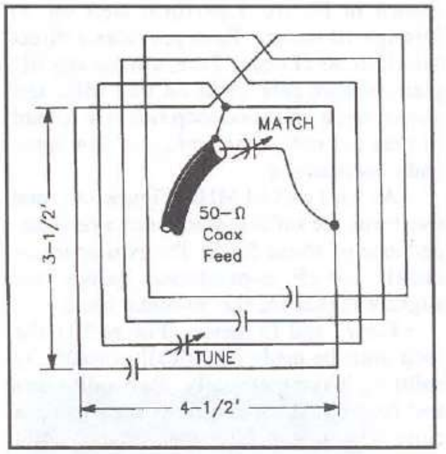

Figure 2—Extending the Fig 1 concept to 80 and 160 meters requires adding another turn to the antenna and increasing the capacitances. See the text for capacitor values.

of small 1-kV ceramic capacitors, I reached 3.6-MHz resonance with a value of 100 pF for each of the four capacitances. One capacitor was then replaced with an old broadcast-style, two-gang 300-300-pF variable for tweaking. I measured the feed impedance, at the center of one of the four turns, at about 18 Ω–well within the range of simple gamma matching. So another junk-box broadcast variable, three 350-pF gangs paralleled, was suspended at the loop's top center. Only about 650 pF was needed for a 1:1 match to coax after a 30-inch gamma lead was dangled to a tap at the center of one side of the feed turn. Then the system was tweakable over the entire 80-meter CW range. The remaining ceramic capacitors get slightly warm at 80 watts, which indicates significant power loss. These capacitors should be replaced with low-loss units for best performance, but I left them alone for my experiments. I easily contacted the East Coast and as far west as Arizona in the following few nights on the air.

My next inclination, as you may surmise, was to try the little gem on 160 meters. A four-foot-square transmitting loop for 1.8 MHz? I wasn't overly optimistic. Years of tinkering at W9BRD had failed to produce a decent indoor compact antenna, even of much larger size, for top-band work. But I dug into my bag of ceramics and gave it a go. The four-turn loop, now only $^1/_{32}$ wavelength in circumference, resonated at about 1.8 MHz with four 350-pF capacitors. One of these was replaced with an old broadcast-type 500-pF variable for tweaking the center frequency between 1.8 and 1.85 MHz. I double- and triple-checked the measured feed impedance—16 Ω. Now I knew it would work. The gamma-match SWR dipped to 1:1 at about 800 pF using the same lead, tap and broadcast capacitor as on 80 meters. Immediate solid 40-watt CW QSOs with WN9W, KC4WWV, WK0B and W4VZB were most gratifying.

Variations and the Higher Bands

If such a $^1/_{16}$-wavelength loop can function at 80 meters, the same should be true at 20 meters. So I wound a 1-foot-square, three-turn model on a cardboard box, hung it on a wooden bulkhead, and resonated it to 14.050 MHz with three 50-pF ceramic capacitors. This time, instead of the usual fudging capacitor, I spaced the turns $1^1/_2$ inches apart. Center-frequency tweaking thus could be done by uniformly edging the turns away from or toward each other. This spacing approach works fine for box-wound loops, but for the lower bands, I prefer closewound zip cord. Input impedance turned out to be about 11 Ω—I should have used a fourth turn—and gamma coupling worked okay, as on 80 and 160. The 1:1 feed-line dip occurred with a 120-pF series capacitor, with a gamma lead of 13 inches. In my installation, this antenna is no low-angle DX radiator, to be sure, but I've received S9 reports during many stateside QSOs. Anyone for a 3-inch-square 2-meter version?

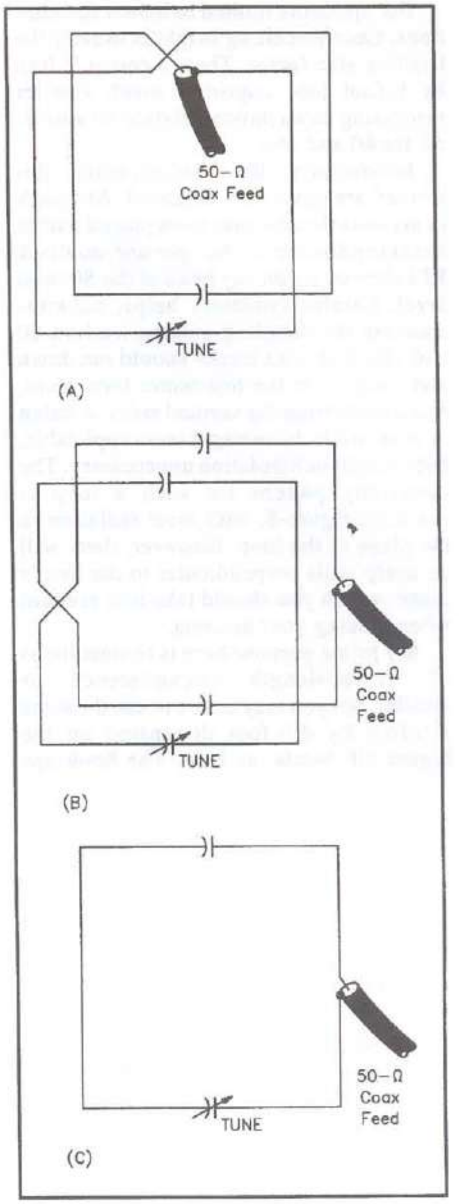

Figure 3—At A, the 30/20-meter loop; at B, the 15/17-meter version, and at C, the 10-meter variant.

Antennas smaller than necessary are interesting stunts that help prove a design. However, for the lowest HF bands I recommend the largest practicable circumference.

For operators limited to indoor installations, floor-to-ceiling height is usually the limiting size factor. Thus, a cooler 7-foot by 9-foot job, requiring much smaller resonating capacitances, should be a natural for 80 and 160.

Incidentally, the configurations described are quite well balanced. Although in my tests the antennas were placed within tweaking distance of the operator, no shack RFI showed up on any band at the 80-watt level. Careful symmetry helps, notwithstanding the dangling gamma lead on 80 and 160. The coax feeder should run down and away from the top-center feed point, equidistant from the vertical sides. A balun or coaxial RF choke might seem applicable, but I found such isolation unnecessary. The directivity pattern for such a loop is the usual figure-8, with most radiation in the plane of the loop. However, there will be sharp nulls perpendicular to the loop's plane, which you should take into account when placing your antenna.

My prime purpose here is to treat loops of $^1/_8$-wavelength circumference or smaller, but you may desire to use the same $3^1/_2$-foot by $4^1/_2$ foot dimension on the higher HF bands, as I do. The hook-ups shown in Figure 3 perform well on 30 through 10 meters. Each provides a direct match to 50-Ω coax. True, almost any old piece of wire gets QSOs on 10.1 MHz and above, but a balanced loop radiator is hard to beat for indoor hamming. A few hints and observations:

- At 10. 1 and 14 MHz (Figure 3A), just two turns are sufficient to reach a feed impedance of about 50 Ω. The two approximately 20-pF capacitance values are slightly higher for the 30-meter band.
- For 17 and 15 meters (Figure 3B), the loop must be made electrically smaller by splitting it symmetrically. Two split turns and four equal capacitances then bring a close 50-Ω match. Note that the feed point is moved to the center of one of the four vertical sides. Capacitances are in the 15-pF range.
- On 12 and 10 meters (Figure 3C), one split turn will suffice for a near-50-Ω match, with two capacitors of about 12 pF. Here, splitstator midgets can ease adjustment by minimizing hand capacitance, but it's a fancy junk box that includes them. The old APC midgets, whose rotors have little more mass than their stators, will maintain enough system balance.

Cautions and Conclusions

Like any indoor antenna, the loops I describe here generate substantial electromagnetic fields in operation. Thus, they have considerable potential to generate RFI. For this reason and to prudently avoid placing yourself or others in large RF fields, you should keep all antennas as far from consumer electronic devices and people as possible, and use the least RF power necessary to conduct the desired communications. The current editions of *The ARRL Handbook* and *The ARRL Antenna Book* cover this subject in more detail.[2]

A final comment: For overall results with compact antennas, like almost any skyhook, the higher above ground the better. In my case, radiation and reception strongly favor higher skywave propagation angles. All antenna configurations I describe here were tested and operated in a cellar ham shack, where half of the system is below ground level!

Notes

[1] See the sidebar, "Increasing Multiturn Loop Efficiency," for some ideas on optimizing such an antenna's performance.

[2] See Chapter 9 of the *Handbook* and Chapter 1 of the *Antenna Book*.

Radio and Electronics Cookbook

Edited by
Dr George Brown, CEng, FIEE, M5ACN

RSGB

Newnes
OXFORD AUCKLAND BOSTON JOHANNESBURG MELBOURNE NEW DELHI

Contents

Preface ix

1. A medium-wave receiver 1
2. An audio-frequency amplifier 4
3. A medium-wave receiver using a ferrite-rod aerial 9
4. A simple electronic organ 12
5. Experiments with the NE555 timer 17
6. A simple metronome 21
7. What is a resistor? 24
8. Waves – Part 1 27
9. A beat-frequency oscillator 31
10. What is a capacitor? 34
11. Waves – Part 2 38
12. An LED flasher 41
13. Waves – Part 3 44
14. Choosing a switch 46
15. An aerial tuning unit for a receiver 49
16. A simple 2 m receiver preamplifier 52
17. Receiving aerials for amateur radio 54
18. The Colt 80 m receiver – Part 1 58
19. A crystal radio receiver 62
20. The varactor (or varicap) diode 64
21. A portable radio for medium waves 65
22. The Colt 80 m receiver – Part 2 70

23. A simple transistor tester 73
24. An introduction to transmitters 77
25. The Colt 80 m receiver – Part 3 81
26. A two-way Morse practice system 88
27. The Colt 80 m receiver – Part 4 91
28. A simple crystal set 95
29. A crystal calibrator 100
30. A simple short-wave receiver – Part 1 104
31. A fruit-powered medium-wave radio 106
32. A capacitance bridge 109
33. A simple short-wave receiver – Part 2 113
34. A basic continuity tester 117
35. A charger for NiCad batteries 119
36. An 80 metre crystal-controlled CW transmitter 123
37. A solar-powered MW radio 129
38. A receiver for the 7 MHz amateur band 133
39. Diodes for protection 137
40. An RF signal probe 140
41. An RF changeover circuit 142
42. A low-light indicator 146
43. A J-pole aerial for 50 MHz 149
44. Measuring light intensity – the photometer 153
45. A 70 cm Quad loop aerial 156
46. A UHF field strength meter 160
47. Christmas tree LEDs 162
48. An audio signal injector 166
49. Standing waves 168
50. A standing-wave indicator for HF 170
51. A moisture meter 174
52. Simple aerials 177
53. A breadboard 80 cm CW transmitter 182

54. A 7-element low-pass filter for transmitters 186
55. Radio-frequency mixing explained 189
56. A voltage monitor for a 12 V power supply 192
57. A 1750 Hz toneburst for repeater access 196
58. A circuit for flashing LEDs 201
59. Digital logic circuits 205
60. A resistive SWR indicator 210
61. An audio filter for CW 213
62. An electronic die 215
63. The absorption wavemeter 222
64. An HF absorption wavemeter 224
65. A vertical aerial for 70 cm 228
66. A UHF corner reflector aerial 230
67. A switched dummy load 234
68. A simple Morse oscillator 238
69. A bipolar transistor tester 240
70. The ‘Yearling’ 20 m receiver 245
71. Adding the 80 metre band to the Yearling receiver 251
72. How the Yearling works 255
73. A field strength meter 258
74. Preselector for a short-wave receiver 261
75. An audible continuity tester 265
76. An experimental 70 cm rhombic aerial 268
77. Water level alarm 272
78. A delta loop for 20 metres 275
79. A simple desk microphone 279
80. Morse oscillator 284
81. A simple 6 m beam 287
82. An integrated circuit amplifier 291
83. A novice ATU 293
84. CW QRP transmitter for 80 metres 297

85. An audio booster for your hand-held 303
86. A grid dip oscillator 306
87. A CW transmitter for 160 to 20 metres 312
88. Matching the end-fed random-wire aerial 315

Preface

Although we are surrounded by sophisticated computerised gadgets these days, there is still a fascination in putting together a few resistors, capacitors and the odd transistor to make a simple electronic circuit. It is really surprising how a handful of components can perform a useful function, and the satisfaction of having built it yourself is incalcuable.

This book aims to provide a wide variety of radio and electronic projects, from something that will take a few minutes to a more ambitious weekend's worth. Various construction techniques are described, the simplest requiring no more than a small screwdriver, the most complex involving printed circuit boards.

Originally published by the Radio Society of Great Britain, the projects were all chosen to be useful and straightforward, with the emphasis on practicality. In most cases the workings of the circuit are described, and the projects are backed up by small tutorials on the components and concepts employed. In the 21st century it may seem strange that few of the published circuits use integrated circuits (chips). This is intentional as it is much easier to understand how the circuit works when using discrete components.

Anyone buying the *Radio and Electronics Cookbook* will find that it will lead to hours of enjoyment, some very useful and entertaining gadgets, and increased knowledge of how and why electronics circuits work, and a great sense of satisfaction. Beware, electronic construction is addictive!

WARNING: This book contains construction details of transmitters. It is illegal to operate a transmitter without the appropriate licence. Information on how to obtain an Amateur Radio Licence can be obtained from the Radiocommunications Agency, tel. 020 7211 0160.

1 A medium-wave receiver

Introduction

Let us start off with something that is really quite simple and yet is capable of producing a sense of real satisfaction when complete – a real medium-wave (MW) radio receiver! It proves that receivers *can* be simple and, at the same time, be useful and enjoyable to make. To minimise the confusion to absolute beginners, *no* circuit diagram is given, only the constructional details. The circuits will come later, when you have become accustomed to the building process. In the true amateur spirit of ingenuity and inventiveness, the circuit is built on a terminal strip, the coil is wound on a toilet roll tube (as amateur MW coils have been for 100 years!), and the receiver is mounted on a piece of wood.

Putting it together

Start by mounting the components on the terminal strip as shown in **Figure 1**, carefully checking the position and value of each one. The three capacitors are all the same, and so present no problem. They (and the resistors) may be connected either way round, unlike the two semiconductors (see later). The resistors are coded by means of coloured bands. You can refer to Chapter 7 if you have difficulty remembering the colours and their values.

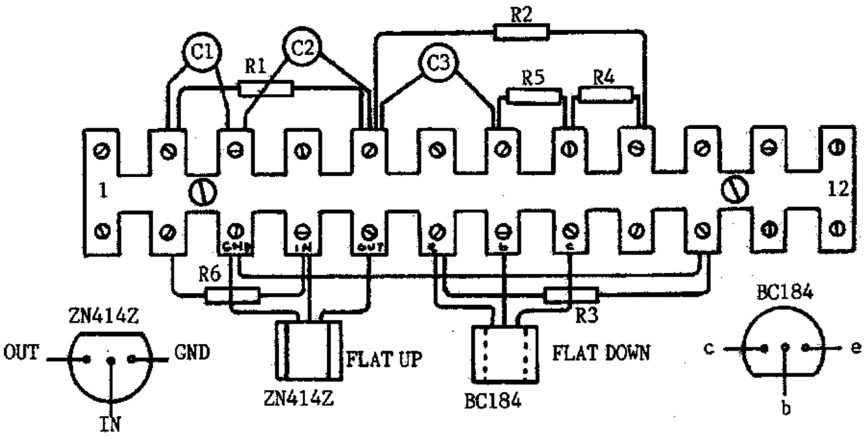

Figure 1 Terminal strip – position of components

1. Brown, Black, Yellow	100 000 ohms	(R1, R5, R6)
2. Green, Blue, Brown	560 ohms	(R2)
3. Red, Violet, Brown	270 ohms	(R3)
4. Brown, Black, Orange	10 000 ohms	(R4)

The integrated circuit (the ZN414Z) and the transistor (the BC184) *must* be connected correctly. Check Figure 1 carefully before fitting each device.

Now wind the coil. Most tubes are about 42 mm diameter and 110 mm long. Don't worry if your tube is slightly different; it shouldn't matter. Make two holes, about 3 mm apart, about 40 mm from one end, as shown in **Figure 2**. Loop your enamelled wire into one hole and out of the other, and draw about 100 mm through; loop this 100 mm through again, thus anchoring the wire firmly. Now wind on 80 turns, keeping the wire tight and the turns close together but not overlapping. After your 80th turn, make another two holes and anchor the wire in the same way as before. Again, leave about 100 mm free after anchoring. Using another piece of enamelled wire (with 100 mm ends as before), loop one end through the same two holes which contain the end anchor of the last winding, wind two turns and anchor the end of this short winding using another pair of holes. Figure 2 shows the layout.

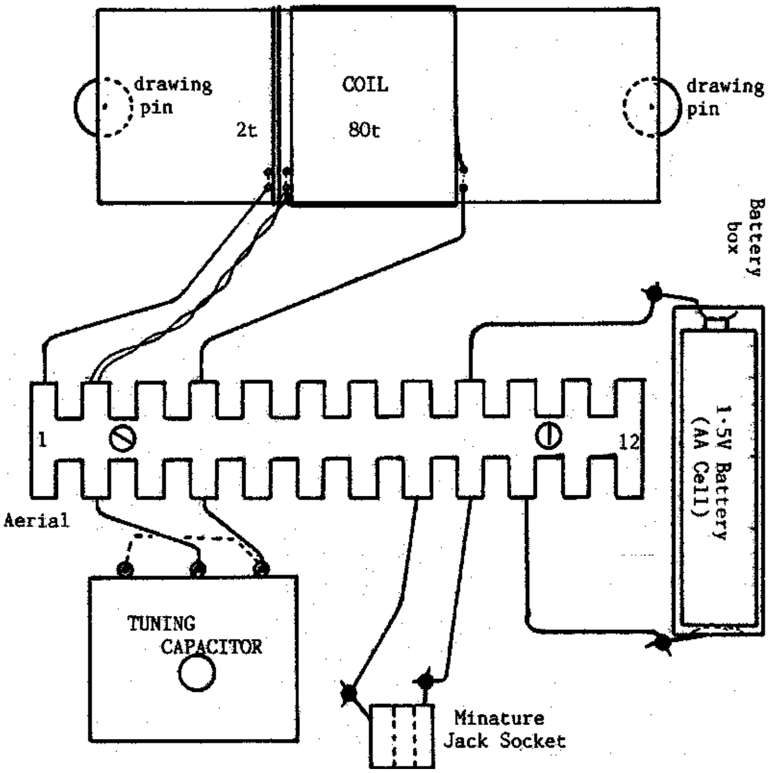

Figure 2 The layout of the parts on the wooden base

With some glass paper, remove the enamel from the ends of both pieces of wire which go through the same holes (i.e. the bottom of the large coil and the top of the small coil), then twist these bare ends together. Remove the enamel from the remaining ends of the coil. The coil is now finished!

The baseboard can be any piece of wood about 150 mm square. Fix the coil near the back edge using drawing pins and connect the wires from the coils to the terminal strip as shown in Figure 2. Using short pieces of PVC-insulated wire (and with assistance if you have never soldered before), solder one piece across the two outer tags of the variable capacitor, shown by the dotted line in Figure 2, and then two longer pieces to the centre tag and one outside tag. Connect these to the terminal strip. Then solder two more insulated wires on to the jack socket (into which you will plug your crystal earpiece), the other ends going to the terminal strip. The last two wires (one must be red) need to be soldered on to the battery box, their other ends going to the terminal strip also. Make sure the red wire goes to the *positive* terminal on the battery, and is connected to terminal 9. The other connection to the battery goes to terminal 10.

Attach the terminal strip to the baseboard with small screws or double-sided sticky tape. The other parts can be mounted the same way.

Listening is done ideally with the recommended crystal earpiece. Don't be tempted to use your Walkman earpieces; they are not the same and will not perform anything like as well. The receiver should work without an extra aerial, but one can be attached to terminal 1 if necessary. A long piece of wire mounted as high as possible is ideal. The *Audio-frequency Amplifier* project will enable you to use a loudspeaker with your receiver, using the signal from the jack socket. No circuit modifications will be needed!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R5, R6	10 kilohms (kΩ)
R2	560 ohms
R3	270 ohms
R4	10 kilohms (kΩ)

Capacitors

C1, C2, C3	100 nanofarads (nF)
	500 picofarads (pF)

Semiconductors

ZN414Z, BC184

Additional items
- 12-way 2 A terminal strip
- 22 metres of 28 SWG enamelled copper wire
- A few short pieces of coloured PVC-insulated wire
- Crystal earpiece
- 3.5 mm jack socket
- 1.5 V AA-size battery and box
- Toilet roll tube
- Double-sided sticky tape or selection of screws

Tools required

Small screwdriver, soldering iron.

2 An audio-frequency amplifier

Introduction

This simple amplifier can be built by anyone who is able to solder reasonably well. It doesn't require any setting up and, provided our instructions are followed exactly, will work very well. The circuit diagram is included for the benefit of our more advanced readers, but it is not needed in the construction process. Please practise your soldering before you start, and don't use a printed circuit board (PCB) until you are confident that your soldering is up to scratch.

The amplifier can be used with other projects; it will provide plenty of sound from the MW Radio or from the Morse Sounder projects. It will usually be built into other pieces of equipment, so a box is not supplied with the kit. There is no reason why it shouldn't be put into a box and used as a general-purpose amplifier to help test other projects.

The components

Before you start, you should check that you have *all* the components to hand. A list and some helpful hints are given below.

1. PCB. The plain side is the *component side* and the soldered side is the *track side*. **Figure 1** shows the track side full size. Make the PCB from the pattern given in Figure 1. Otherwise, build the circuit on a matrix board.
2. Three resistors. Locate the gold or silver band around the resistor, and turn the resistor until this band is to the right. There are three coloured bands at the left-hand end of the resistor. Find the resistor whose colours are YELLOW, VIOLET, RED, and look at the resistor colour code chart which you will find in Chapter 7. From this, you will see that YELLOW indicates the value 4, VIOLET the value 7, and RED the value 2. The first two colours represent real numbers, and the last value is the number of zeros (noughts) which go *after* the two numbers. So, the value is 47 with two zeros, i.e. 4700 ohms. In this way, the resistor coloured BROWN, GREY, BROWN has a value of 180 ohms, and the last one, BROWN, RED, GREEN, has a value of 1 200 000 ohms. The ohm (often written as the Greek letter omega (Ω)) is the unit of resistance. If you do not yet feel confident in identifying resistors by their colours, use the Resistor Colour Codes given in Chapter 7.
3. Four capacitors. The two small 'beads' are tantalum capacitors and will be marked 4.7 μF or 4μ7, with a '+' above one lead. A tubular capacitor with wires coming from each end should be marked 220 μF, with one end marked '+' or '–'. This is called an *axial* capacitor because the wires lie on the axis of the cylinder. This is in contrast to the final capacitor, where both wires emerge from the same end. This is a *radial* capacitor, and will be marked 47 μF. Again, one lead will be marked '+' or '–'. Capacitors marked like this are said to be *polarised*, and it is vital that these are placed on the PCB the right way round, so take notice of those signs!
4. Two diodes. These are tiny glass cylinders with a band around one end, and may be marked 1N4148; this is their type number. Like polarised capacitors, they *must* be put on the PCB the correct way round!

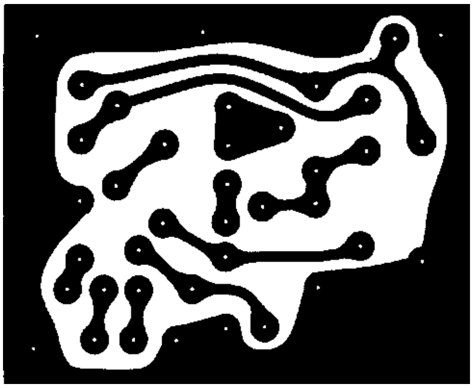

Figure 1 The toil pattern of the PCB – looking from the track side

5. Three transistors. One should be a BC548 (or a BC182), the other two should be BC558 (or BC212).
6. One volume control with internal switch.
7. One loudspeaker. This is quite fragile – don't let anything press against the cone.
8. One PP3 battery clip with red and black leads.

Putting it together

Lay the PCB on a flat, clean surface with the track side downwards. It is always useful to compare the layout with the circuit diagram, given here in **Figure 3**. Although you can't see it, the D-i-Y Radio sign should be at the top. Compare the hole positions with those shown in **Figure 2**. Bend the resistor wires at right angles to their bodies so that they fit cleanly into the holes in the PCB. Push each resistor towards the board so that it lies flat on the board. Then supporting each one, turn the board over and splay out the wires *just* enough to prevent the resistor falling out. Then, solder each wire to its pad on the PCB, and cut off the excess wire. When you have more confidence, you can cut of the excess wire *before* soldering; it often makes a tidier joint.

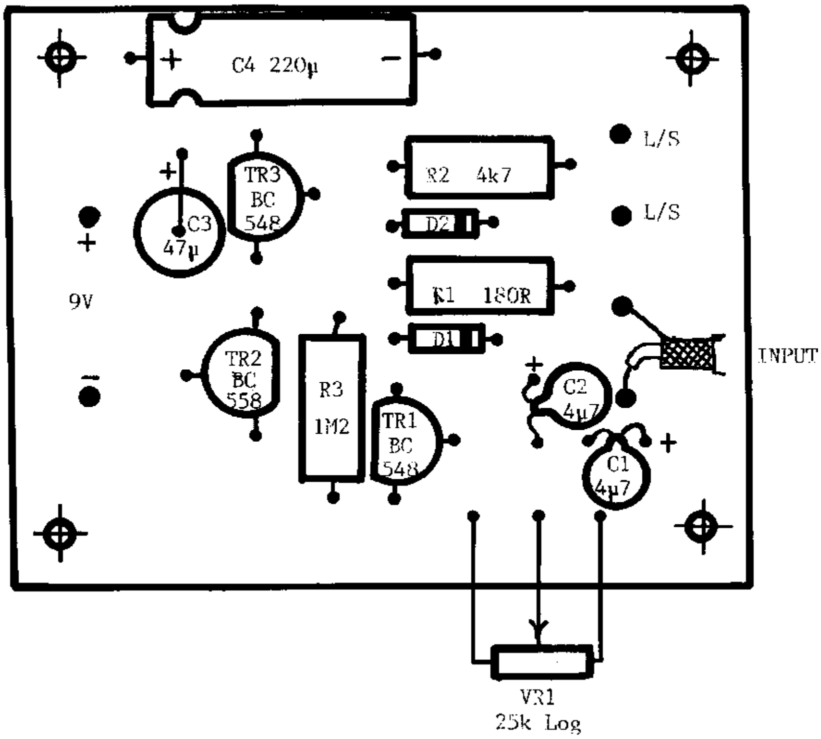

Figure 2 Positions of the components on the printed circuit board (PCB)

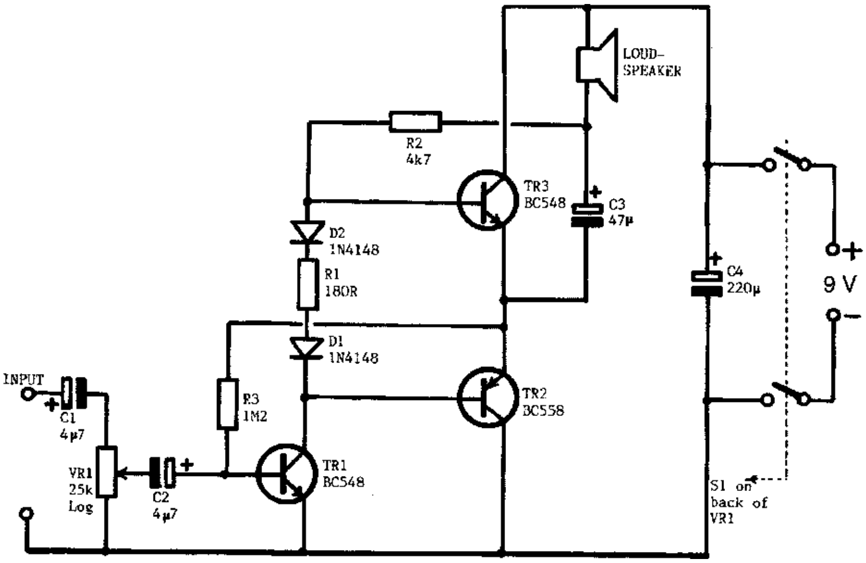
LOUD-
SPEAKER
R2
4k7
TR3
BC548
C3
47µ
D2
1N4148
R1
180R
C4
220µ
+
9 V
−
D1
1N4148
INPUT
C1
4µ7
R3
1M2
TR2
BC558
VR1
25k
Log
C2
4µ7
TR1
BC548
S1 on
back of
VR1

Figure 3 The amplifier's circuit diagram

Now fit the four capacitors. Each must be connected the right way round, so look at each component, match it up with the diagram of Figure 2, bend its wires carefully and repeat the soldering process you performed with the resistors, making sure that the components are close to the board and not up on stilts! Fit the two diodes the correct way round, and solder then as quickly as you can – they don't like to be fried!

Mount the transistors about 5 mm above the PCB. Make sure the correct transistors are in the correct places, and that the flats on the bodies match up with those shown in Figure 2.

Mount the volume control so that the spindle comes out from the front of the board. Use a piece of red insulated wire to the pad marked + on the PCB, and a black piece to the pad marked –, and solder these to the tags on the back of the control, as shown in **Figure 4**. Connect the two leads from the battery clip to the other tags on the switch; Figure 4 will help you. Finally, use two pieces of insulated wire about 100 mm long, twisted together, to connect the loudspeaker to the PCB.

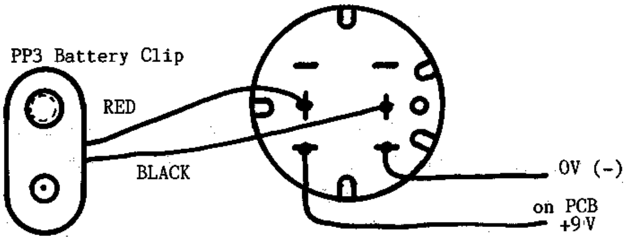
PP3 Battery Clip
RED
BLACK
−
−
+
+
0V (-)
on PCB
+9V

Figure 4 Connections to switch on back of VR1

Box clever!

If you wish to put the amplifier into a box, there is no problem; almost any box that is big enough will do. All that is needed is one hole big enough to accept the bush of the volume control; the PCB will be supported by the volume control. The prototype was not fitted into a box, but mounted on an odd piece of aluminium, bent into an L-shape and screwed on to a wooden base. The loudspeaker was mounted on the aluminium panel by two small pieces of aluminium with 3 mm holes drilled in them, which acted as clips around the edge of the speaker. Drill a few holes in the panel in the position of the speaker to let the sound get out!

Your input signal can be connected to the amplifier with two short pieces of wire, but if the connection needs to be long, use screened cable, with the braid connected as shown in Figure 2.

If you decide to use a different loudspeaker, make sure that its impedance (the resistance value marked on the back of the magnet) is at least 35 ohms. Anything lower may damage TR2 and TR3, and will certainly run down your battery very quickly. You will be surprised at the uses you can find for this little amplifier!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	180 ohms (Ω)
R2	4.7 kilohms (kΩ)
R3	1.2 megohms (MΩ)
VR1	25 kilohms (kΩ) log with DPST switch

Capacitors: all rated at 25 V minimum

C1, C2	4.7 microfarads (μF)
C3	47 microfarads (μF)
C4	220 microfarads (μF)

Semiconductors

TR1, TR3	BC548 npn
TR2	BC558 pnp
D1, D2	1N4148

Additional items
- PCB
- Speaker >35 ohms
- PP3 battery clip and battery

3 A medium-wave receiver using a ferrite-rod aerial

Introduction

This design came from the Norfolk Amateur Radio Club, and enables you to build a simple Amplitude Modulation (AM) receiver for frequencies between 600 kHz and 1600 kHz. It should take you around 2 hours to build, and can be used with Walkman-type earpieces. **Figure 1** shows the circuit diagram.

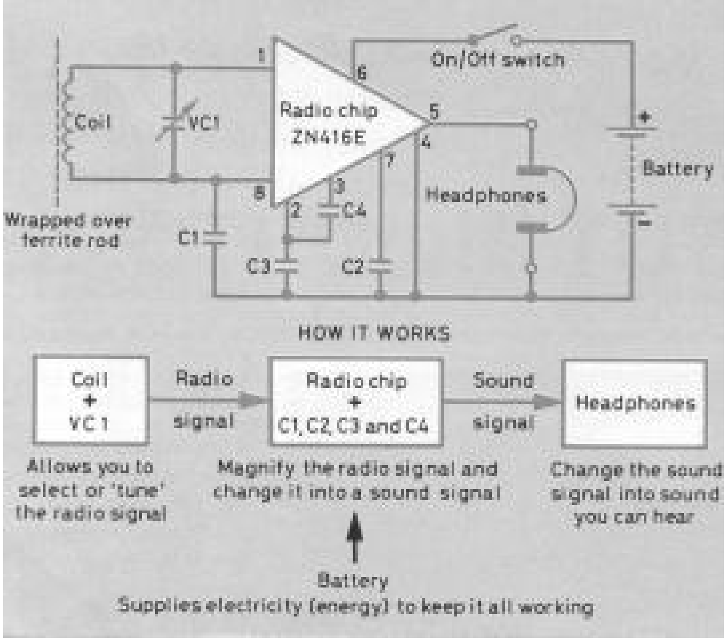
1
6
On/Off switch
Coil
VC1
Radio chip
ZN416E
5
4
7
+
Battery
−
8
3
2
C4
Headphones
Wrapped over
ferrite rod
C1
C3
C2
HOW IT WORKS
Coil
+
VC 1
Radio
signal
Radio chip
+
C1, C2, C3 and C4
Sound
signal
Headphones
Allows you to
select or 'tune'
the radio signal
Magnify the radio signal and
change it into a sound signal
Change the sound
signal into sound
you can hear
Battery
Supplies electricity (energy) to keep it all working

Figure 1 Circuit and block diagrams of the radio

Description

The whole circuit is built on a 50 mm by 50 mm printed circuit board (PCB) designed to fit on the inside of the lid of a plastic box, and is stuck there using sticky pads, the shaft of the variable capacitor going through a hole in

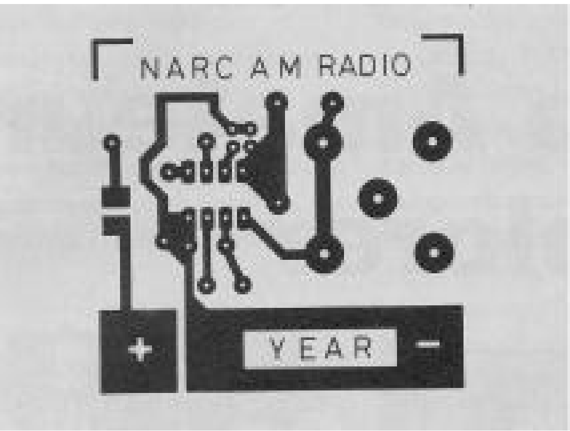

Figure 2a The PCB, solder side

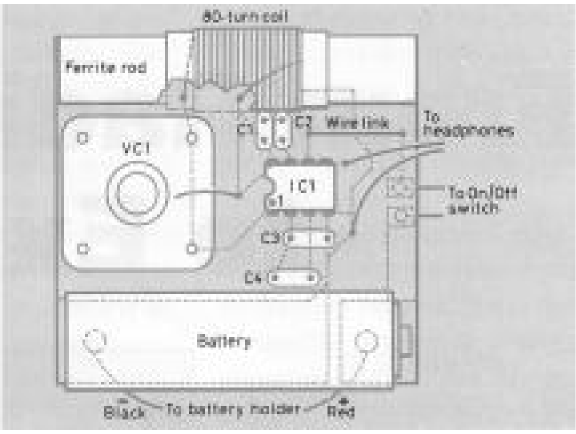

Figure 2b The PCB, component side

the lid. Only two pairs of leads are soldered to the board – one pair goes to the 1.5 V battery in its holder, and the other to the earphone socket. **Figures 2a** and **2b** show the printed circuit and the component layout double size for clarity. You are not obliged to build the circuit on a PCB.

Building it

1. **Check and identify components**. Tick the parts list.
2. **Carefully unwind the wire.** Use paper to make an insulating tube (called a 'former') around the centre of the ferrite rod and secure it with Sellotape. Now, close-wind *all* the wire (leave no gaps between adjacent turns) around the paper former. Secure the winding with more Sellotape, leaving 50 mm of wire free at each end for connection to the circuit. See **Figure 3a.**
3. **Solder in VC1.**
4. **Solder in the integrated circuit holder.** There is a notch in one end of the holder; this should face VC1. Solder also the wire link and the capacitors. Be careful to avoid solder 'bridges' between adjacent tracks on the PCB.
5. **Solder the battery leads.** These must be connected properly – the red battery lead to the + (positive) area and the black lead to the – (negative) area.
6. **Strip bare 1 cm of insulation from the ends of two wires**. Solder them between the PCB and the headphone socket (see **Figure 3b**). Use the end tabs on the socket. Using another pair of insulated wires connect the ON/OFF switch to the PCB tabs shown in Figure 2b.

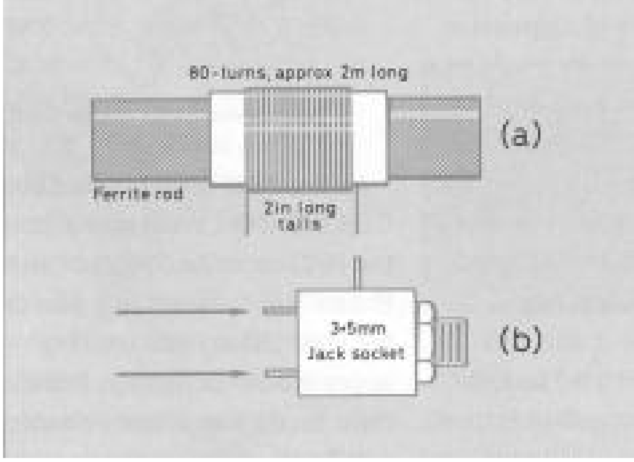

Figure 3 Details of coil and headphone socket

7. **Fix the elastic band**. This goes through the holes at the top of the PCB, with the ferrite rod being slipped through the two end loops. (Note: although the coating on the copper wire is designed to melt away during soldering, it is quite common for difficulty to be experienced in obtaining a good soldered joint; to be on the safe side, remove the coating *before* soldering (with a small piece of sandpaper).) Carefully place the wire ends of the coil through the PCB just above VC1, and solder on the track side.
8. **Fit IC1 into its holder**. This should be done carefully, making sure that *all* the pins are located above their respective clips *before* applying any pressure! Make sure also that the notch on the IC (as shown in Figure 2b) matches the notch in the holder, and faces VC1.
9. **Put battery in its holder**. Listen for some noise in the headphones as VC1 is rotated. Make sure the headphone plug is fully inserted into its socket.
10. **Fix the working board to the lid.** Use the sticky pads and apply *gentle* pressure. Fit the tuning knob, the ON/OFF switch and the earphone socket.
11. **Test again.** If all is still working, fit the lid screws and admire your completed radio!

Parts list

Capacitors	
C1, C2	0.01 microfarad (μF)
C3, C4	0.1 microfarad (μF)
VC1	500 picofarads (pF)
Semiconductor	
IC1	ZN416E

Additional items
Plastic box (recommended size 76 × 64 × 50 mm internal)
8-pin DIL socket for IC1
Printed circuit board
Tuning knob for VC1
Wire link for PCB
2 m of 30 SWG copper wire, self-fluxing
Piece of paper 25 × 50 mm, to make the coil former
Ferrite rod 70 mm long by 10 mm diameter, approximately
Battery, AA size 1.5 V, with holder and attached wires
Miniature earphone socket (3.5 mm stereo jack)
ON/OFF switch (push-button SPST latched or slide switch)
4 off 100 mm insulated connecting wires, for jack socket and ON/OFF switch
Pair Walkman-type earphones
Elastic band, to attach ferrite rod to PCB
4 off sticky pads for securing PCB to box lid

Kits

Ready-made PCBs may be available from Alan J. Wright, G0KRU, Hewett School, Cecil Road, Norwich NR1 2PL.

4 A simple electronic organ

Introduction

This project has nothing to do with radio but, let's admit it, *any* electronics project is good experience! Why not build this little organ – it will keep the children amused at least! It uses the popular NE555 integrated circuit, which contains a circuit which will periodically switch the voltage on the output pin between the supply voltage and zero. Just how frequently this switching occurs depends upon the components external to the integrated circuit. If this switching occurs several hundred or thousand times a second, the change in voltage produced will generate a musical note when connected to a small loudspeaker. The circuit is shown in **Figure 1**.

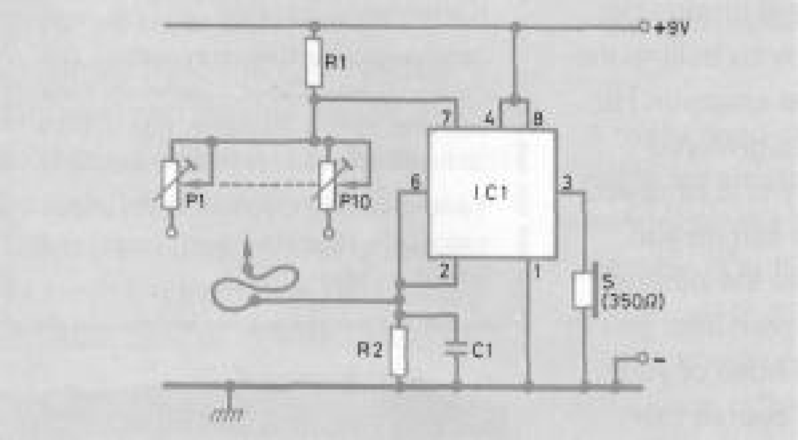
+9V
R1
7
4
8
6
IC1
3
P1
P10
2
1
S
(350Ω)
R2
C1
−

Figure 1 Circuit diagram

Putting it together

(a) Using a PCB. The job is very simple. The placement of components on the unsoldered side of the board is shown in **Figure 2** and the design on the copper track is illustrated in **Figure 3**. Put each component, in turn, on the board, making sure that it lies flat on the board with its tags or wires going cleanly through the holes provided for it; then, solder the wires to the board, cropping them before or after the soldering,

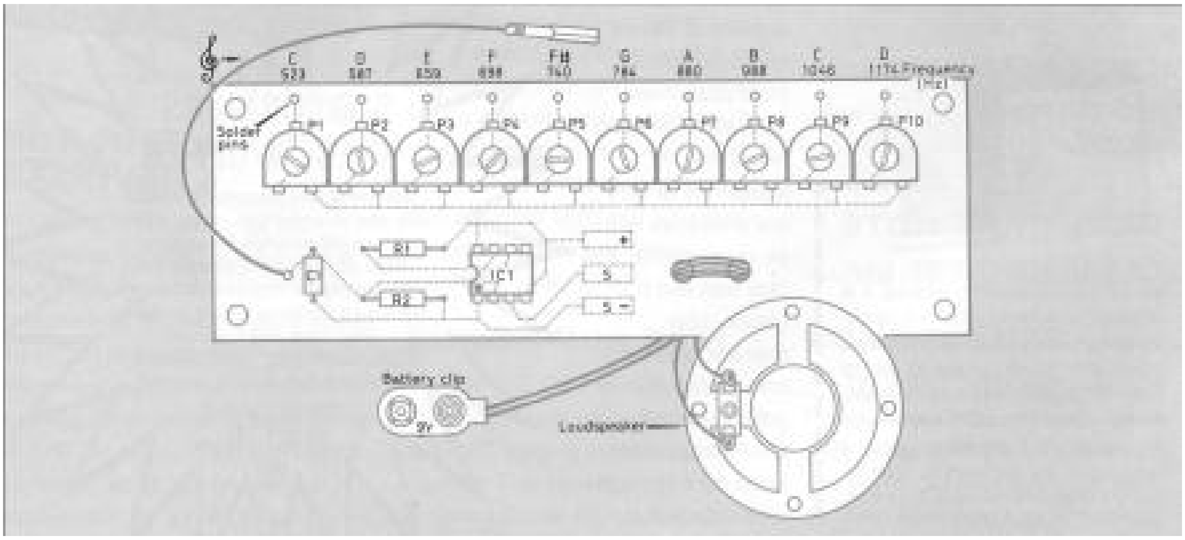
C
523
D
587
E
659
F
698
F#
740
G
784
A
880
B
988
C
1046
D
1174 Frequency (Hz)
Solder pins
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
R1
R2
C1
IC1
+
S
S −
Battery clip
9V
Loudspeaker

Figure 2 Position of components on the printed circuit board (PCB)

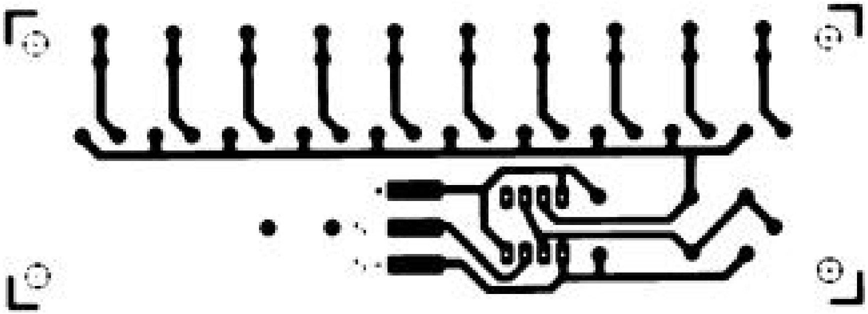

Figure 3 The connections

depending on your preference. If you choose to use a holder for your integrated circuit (highly recommended if your soldering is less than perfect), make sure that the end with a notch in it faces R1 and R2, as shown in Figure 2. Solder the two leads to the speaker to the tabs marked S (either way round), having looped them through the two holes to the right of the tabs in Figure 2. Looping them through the holes acts as a strain relief, ensuring that the soldered joints are not subjected to pulling and bending as you move the wires about. Do the same with the battery leads, the red lead going to the + tab and the negative lead to the – tab (which also has one speaker lead already attached to it). **Figure 4** shows this in detail. Treat the loudspeaker with care – the cone is quite fragile and must not be touched.

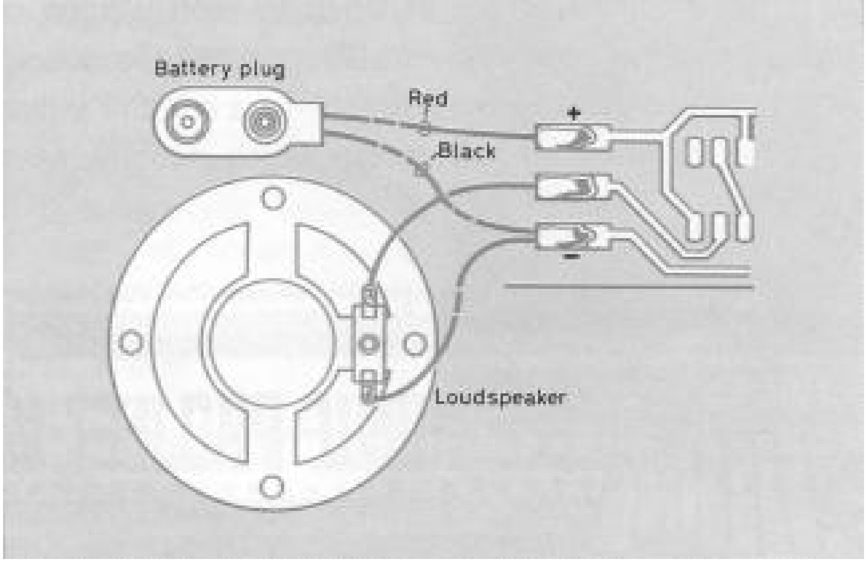

Figure 4 Battery plug and loudspeaker connections

(b) Without a PCB. This is more difficult, and you may need to enlist some help. Using some matrix board (such as Veroboard) is probably the best way of replacing the PCB. You could arrange your circuit in exactly the same way as in the PCB in Figure 3, using wires to replace the copper track.

A simple ‘keyboard’

The keyboard is a row of solder pins along the rear edge of the PCB, one for each note covering the range shown in Figure 2. A flying lead with a small spade on it is provided to touch any of the pins in turn, producing any one of ten different notes.

Testing

Check first that each component is in the correct place. When inserting the NE555 chip, first make sure that the end carrying the notch lies over the end of the holder with the notch; then, make sure each pin of the chip lies directly above the hole into which it fits, before pressing *gently* to insert the chip into the socket. Make sure the battery connections are correct, and insert the battery into the clip. Nothing should happen, except for a click from the loudspeaker; touching the spade on any of the pins should produce a coarse note from the speaker. If nothing happens, check everything again; don’t *assume* that wires go where you think they go!

After you get the first note, all the others should work, too, but they will sound off-tune at first. The organ needs tuning up by adjusting the 10 preset variable resistors P1 to P10. The approximate frequency to which each note should be tuned is given in Figure 2; if you can beg, borrow or steal a frequency counter, setting up is easy. If you have a piano, the organ can be tuned by comparison of the notes with those on the piano. The frequencies are given in Hertz (abbreviation Hz), and represent the number of times the IC switches on and off every second. If the sound coming from the loudspeaker is too loud or very distorted, then try putting an 330 Ω resistor (colour code orange, orange, brown) in series with the loudspeaker. This is done by taking the resistor and cutting its leads to about 5 mm; then, disconnect one speaker lead from the tab on the PCB (it doesn’t matter which). Solder one end of the resistor to the vacated speaker tab, and the free speaker lead to the other end of the resistor. This will limit the volume of sound from the speaker, and lengthen the life of your battery. If it is still too loud, try a resistor of a larger value, or use a smaller resistor to make it louder.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	2.7 kilohms (kΩ)
R2	1 Megohm (MΩ)
P1–P5	Preset resistor 100 kilohms (kΩ)
P6, P7	Preset resistor 50 kilohms (kΩ)
P8	Preset resistor 25 kilohms (kΩ)
P9, P10	Preset resistor 10 kilohms (kΩ)

Capacitor

C1	100 nanofarads (nF) or 0.1 microfarad (μF)

Integrated circuit

IC1	NE555 timer chip

Additional items

S	Loudspeaker >60 ohms (Maplin)
	1 off battery clip (for PP3 battery)
	1 off spade terminal
	12 off solder pins 'Veropins'
	3 off 10 cm lengths of 'hook-up' wire

(This article is based on projects originally designed by Radio Scouting, Netherlands.)

5 Experiments with the NE555 timer

Introduction

Several of the projects in this book use the NE555 timer, an integrated circuit which is at the heart of many circuits whose processes are determined by time intervals. **Figure 1** shows the circuit diagram of an audio oscillator using the 555. The timing voltages (governing the frequency of oscillation) are produced by R1, R2 and C2; a voltage appears at pin 3 which ‘switches’ at this frequency between zero and a voltage close to the supply voltage, which in this case can be anywhere between 6 V and 14 V. The output current, when applied through R3 to a small loudspeaker, produces an audible tone, provided that there is a DC path between the two test leads.

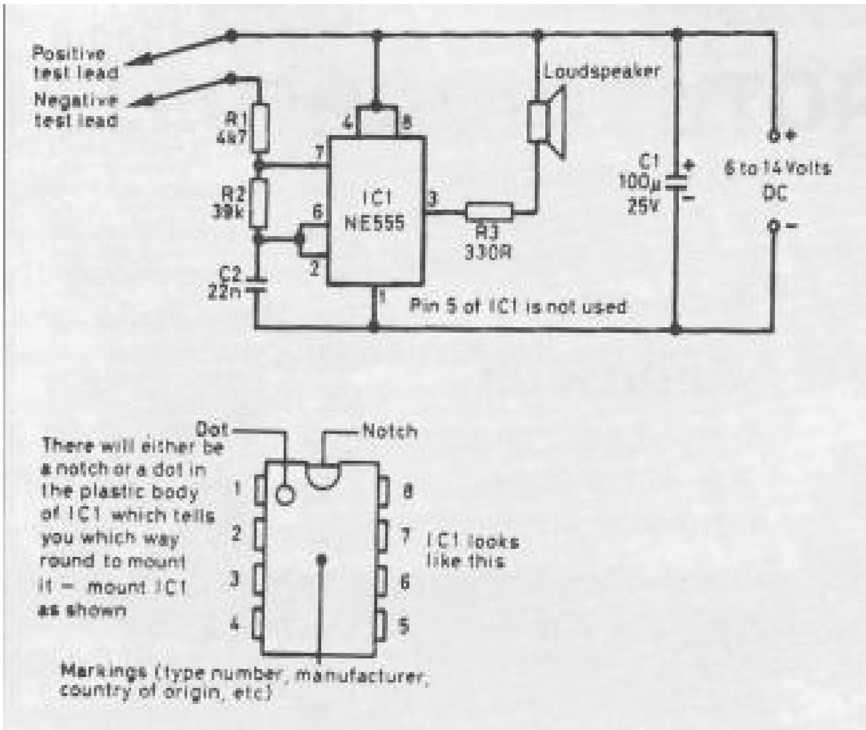
Positive test lead
Negative test lead
R1 4k7
R2 39k
C2 22n
4
8
7
6
2
1
3
IC1 NE555
R3 330R
Loudspeaker
C1 100μ 25V
+
−
6 to 14 Volts DC
+
−
Pin 5 of IC1 is not used
Dot
Notch
There will either be a notch or a dot in the plastic body of IC1 which tells you which way round to mount it – mount IC1 as shown
1
2
3
4
8
7
6
5
IC1 looks like this
Markings (type number, manufacturer, country of origin, etc)

Figure 1 The circuit diagram

Construction

The simplest way to mount the components is on a piece of matrix board (Veroboard), available from any of the good suppliers. The prototype of this circuit used the type of board with copper strips along the underside; these strips are used like the copper tracks on a PCB, to join components together. Firstly, cut the four strips between the positions of the pins of the IC socket, as shown in **Figure 2**. You can buy a tool for this purpose, but a small twist drill (about 3 mm diameter) is just as good. Turn it between your fingers – if you use a drill you will end up with holes right through the board! Then solder in the IC socket (with the notch in the position shown), followed by the four links made with single-conductor insulated wire. Put in each component as shown, ensuring that C1 (an *electrolytic* or *polarised* capacitor) is connected correctly. When all the components have been soldered in, take the 555 chip and lay it on its socket, with its own notch lying above that of the holder. Then, making sure that each pin lies *directly* above its corresponding socket, press down *gently* on the chip, with the board supported on a flat surface.

Testing

Connect the circuit to a battery or small power supply, ensuring that the positive and negative leads are the right way round. Always use red and black leads here, then you are less likely to get it wrong! Switch on. Nothing

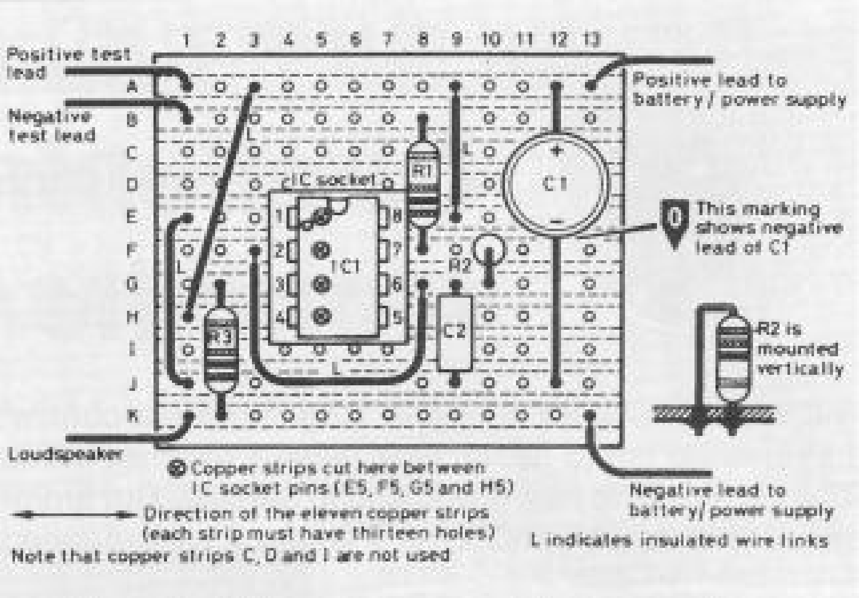
1 2 3 4 5 6 7 8 9 10 11 12 13
A B C D E F G H I J K
Positive test lead
Negative test lead
Loudspeaker
Positive lead to battery / power supply
Negative lead to battery/ power supply
IC socket
IC1
1 2 3 4 5 6 7 8
R1
R2
R3
C1
C2
+
−
L
This marking shows negative lead of C1
R2 is mounted vertically
Copper strips cut here between IC socket pins (E5, F5, G5 and H5)
Direction of the eleven copper strips (each strip must have thirteen holes)
Note that copper strips C, D and I are not used
L indicates insulated wire links

Figure 2 Veroboard layout. If you can read a circuit diagram, the project can be built using other methods

should happen until you short together the two test leads, when there should be a note from the loudspeaker. If this doesn't happen, switch off, disconnect the circuit and check your wiring and soldering. Is it exactly like Figure 2? Are the soldered joints round and shiny? If any are dull, then 'sweat' them briefly with a hot soldering iron until the solder runs, remove the iron, and check that they are as shiny as the rest. Check that there are no solder 'bridges' between adjacent tracks by holding your board up to a strong light. Then, reconnect, switch on and touch the test leads together. All should now work!

Uses of your circuit

1. As a Morse practice oscillator. Simply connect the two test leads to your key and, each time the key is pressed, you should hear a note from the speaker. The frequency of the note may be altered by putting a resistor in series with the key. To do this, remove one test lead from the key and select a resistor; connect one end of the resistor to the free test lead and the other to the empty terminal on your key. Selecting the value of resistor that you need will be a useful experiment in itself.
2. As a continuity tester. You can check fuses and lamp bulbs by connecting them across your test leads. If the speaker remains silent, the fuse or bulb has blown.
3. To indicate changes of resistance. Hold the ends of the test leads in each hand; you should hear a low note, because of the high resistance of your body. Squeeze the ends harder, and the frequency of the note should rise, because you are now making better contact. Repeat this with damp hands and the frequencies will be higher still.
4. As a thermometer. Connect the test leads to a thermistor (a device whose resistance changes with temperature) and warm it with a hair-dryer, or even in your hands, and you will hear the pitch changing with the temperature of the thermistor. A suitable 'bead' thermistor is available from Maplin (order code FX21).
5. As a diode tester. Use any diode, and connect the negative test lead to the end of the diode marked with the ring. This is the cathode of the diode. The other end, the anode, should be connected to the positive test lead, and a note should be heard from the speaker. This does not necessarily mean that the diode is working – yet. Reverse the connections and *nothing* should be heard. If this is the case, the diode *is* working.
6. As a light meter. Use a photoconductive cell (a device whose resistance changes with light intensity) connected between the test leads. A note should be heard. Shading the device with your hand will increase its resistance and the note should decrease in frequency. A suitable device is the ORP12 cell from Maplin (order code HB10).

There are many more applications. **Do not connect the test leads to other circuits that are switched on.** Your circuit, or the circuit you are connecting it to could be damaged. Think of a *passive* device or circuit (i.e. one *not* requiring a power supply or battery) where changes of resistance occur, and you have found another application!

Parts list

Resistors: all 0.25 watt carbon film types

R1	4.7 kilohms (kΩ) – yellow, violet, red
R2	39 kilohms (kΩ) – orange, white, orange
R3	330 ohms (Ω) – orange, orange, brown

Capacitors

C1	100 microfarads (μF) 25 V radial electrolytic
C2	22 nanofarads (nF) or 0.022 microfarad (μF) polyester type with 10 mm lead spacing

Integrated circuit

IC1	NE555

Additional items
- Miniature loudspeaker (preferably 35, 40 or 80 ohm)
- 8-pin DIL socket for IC1
- 0.1 inch Veroboard (‘stripboard’), size 11 strips by 13 holes
- PVC-covered stranded wire for test leads, loudspeaker and battery connections
- PVC-covered solid wire for links on the board
- A power source of between 6 V and 14 V, such as a 9 V battery (PP3)

Component sources

Cirkit
Tandy – many high street shops

6 A simple metronome

Introduction

A metronome is a device used by musicians to indicate the tempo of a piece of music. Until electronics came on the scene, this 'beating of time' was achieved in much the same way as a clock keeps time, i.e. with a pendulum device, the clicking of the escapement indicating the beats of the music.

Those of you who have already built the Morse Key and Buzzer from the designs in this book, will recognise the circuit of this metronome – it is exactly the same as was used to produce the note of the buzzer. This circuit is shown in **Figure 1**.

The circuit

Three components determine the speed at which the circuit oscillates – the speaker (LS), the resistors (VR1 + R1) and the capacitor (C1). VR1 is a variable resistor, so that the speed at which the oscillator operates can be varied. Compared with the component values of the Morse Buzzer (which operated at around 800 Hz), these components now give an oscillation frequency of around 1.25 Hz, which is far too low to be heard as a note. What we *do* hear, however, is a series of clicks, as the voltage across the speaker changes quickly from 0 to 9 V and back again.

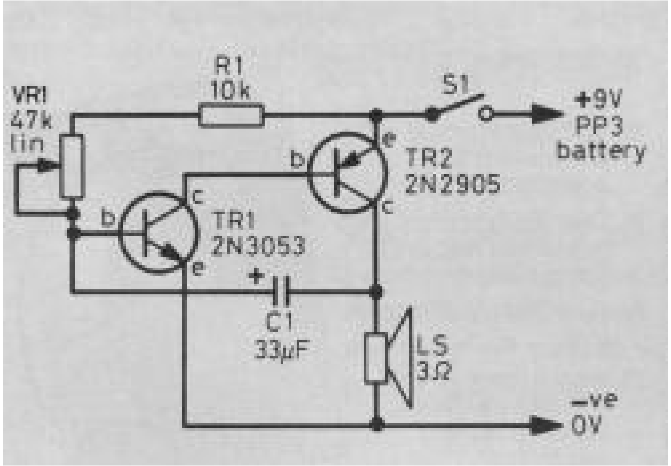

Figure 1 The metronome circuit is rather like the Morse oscillator

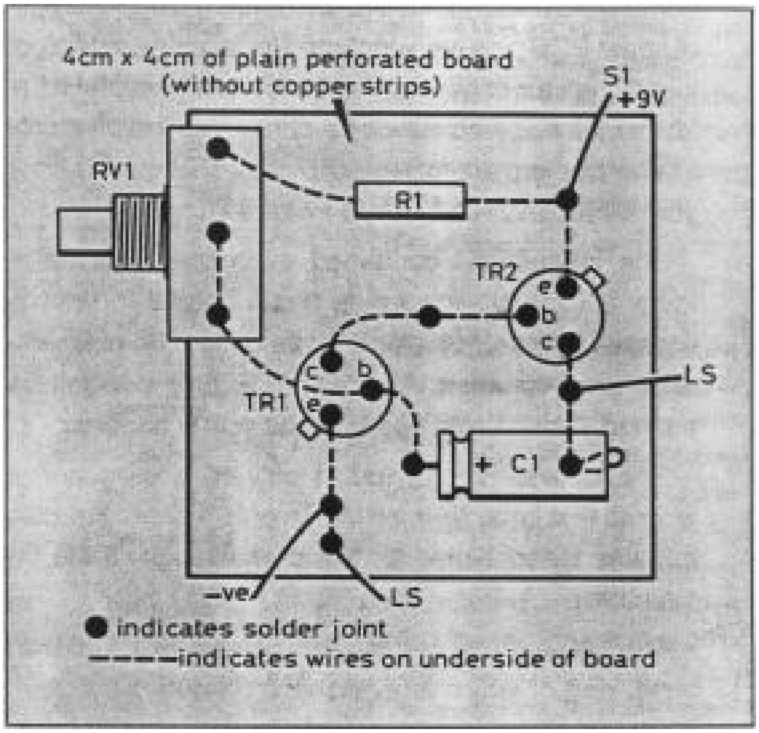
4cm x 4cm of plain perforated board
(without copper strips)
S1
+9V
RV1
R1
TR2
e
b
c
LS
TR1
c
b
e
+ C1
–ve
LS
● indicates solder joint
– – – –indicates wires on underside of board

Figure 2 The component wires are pushed through holes in the circuit board and joined together underneath

Variation of speed could be achieved by varying resistance or capacitance. However, as you may already know, variable capacitors have values in the picofarad range, not the tens of microfarads used here, so it is very simple to employ a variable resistor (potentiometer) to control the oscillator. You could use a multi-way switch to switch in one of several capacitors, as well as having the variable resistor, but this was found to be an unnecessary complication. This design operates between about 100 clicks per minute and 200 clicks per minute.

Making the prototype

A single piece of plain matrix board (no copper strips) measuring about 40 × 40 mm is sufficient to hold all the components except the potentiometer and switch (see later). The case can be plastic or aluminium, and one measuring 65 × 100 × 50 mm is about right. Make sure there are holes in the case beside the speaker cone to let the sound out, and larger holes for the potentiometer and switch. If a potentiometer is used with a combined ON/OFF switch, then the extra hole for the switch is not necessary! It is advisable to construct the circuit *before* putting it in the box, so that it can be tested to ensure that everything is working. If it is, then you can exercise your ingenuity in mounting the speaker, battery and board inside the box. A final test can be made before starting the calibration process.

Calibration

There is no 'easy' way to do this. The frequencies involved are too low to be measured with the average frequency counter, so you will need to resort to using a stopwatch and counting the number of clicks per minute.

Parts list

Resistors: 0.25 watt, 5% tolerance

R1	10 kilohms (kΩ)	
VR1	47 kilohms (kΩ) linear potentiometer	

Capacitor

C1	33 microfarads (μF) electrolytic	

Transistors

TR1	2N3053	npn
TR2	2N2905	pnp

Additional items

S1	SPST	ON/OFF switch
LS	3 ohms (Ω)	loudspeaker
	Knob with pointer for VR1	
	PP3 battery and connector	
	Aluminium case, 65 × 100 × 50 mm	
	Matrix board (plain), 40 × 40 mm	

7 What is a resistor?

Introduction

Materials that carry electricity easily are called **conductors**. They include all metals and salt water, for example. We use wire as a conductor, and the ease with which it passes an electric current depends upon the material, its thickness and its length. Silver (symbol Ag), gold (Au), copper (Cu) and aluminium (Al) are the best metallic conductors. Most wires are made of copper, although the best conductor, weight for weight, is aluminium.

Materials that *don't* carry current (or, at least, do so very badly) are called **insulators**, and they include dry wood, rubber, plastic and glass among their number. Wires are often coated with a layer of insulator to prevent adjacent wires touching and causing an accident.

Resistors

If there wasn't such a thing as resistance, the subject of electronics wouldn't exist; only infinite currents would flow and voltages wouldn't exist either! We need to reduce the flow of current if we are to make current do something useful for us. Components that resist the flow of current are called **resistors**, and they are said to have a **resistance** which is measured in ohms (Ω), named after Georg Ohm, who formulated the law (also named after him) by which the voltage and current through a conductor are related. His law gave rise to the formula *everyone* remembers:

$$I = \frac{V}{R},$$

where I is the current flowing, measured in amps,
V is the voltage across the conductor, and
R is the resistance of the conductor, measured in ohms.

From this equation, you can see that, for a constant value of voltage, V, if the resistance goes up, the current will go down, and vice versa. The circuit symbols for resistors are shown in **Figure 1**. You will find the upper symbol in older magazines; it is still preferred by many engineers. The lower symbol is the prevalent standard symbol.

Resistors are made in several ways, the cheapest using carbon; another type is usually made from a ceramic cylinder (used only as a support) on which

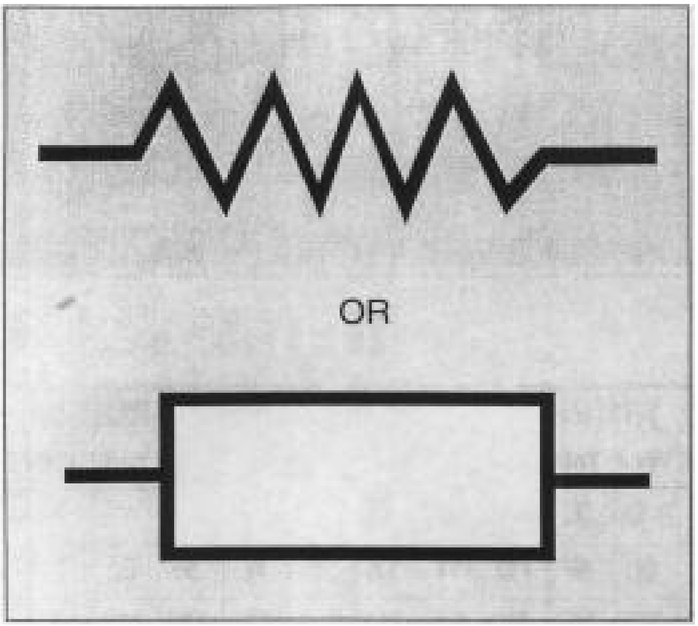

Figure 1 Circuit symbols for resistors

is placed a very thin film of metal – the thinner the film, the greater the resistance. All resistors are coated with a thin film of insulation, for the same reason we discussed earlier.

The colour code

Each resistor has coloured bands on it which enable us to see what value of resistance it has. There are normally three (but sometimes four) at one end, and a single one at the other (see **Figure 2**). The colours indicate figures, according to the list below.

Colour	*Value*	*Colour*	*Value*
Black	0	Green	5
Brown	1	Blue	6
Red	2	Violet	7
Orange	3	Grey	8
Yellow	4	White	9

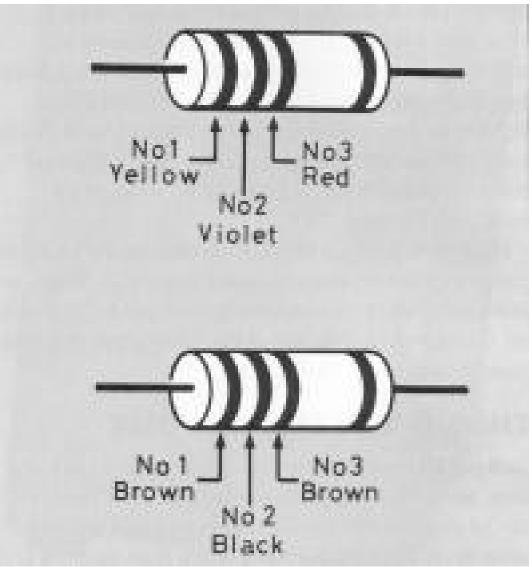

Figure 2 Some examples of resistor colour codes; top 4700 Ω (4.7 kΩ) and bottom 100 Ω

Using the colour codes is easy, once you see the logic behind it. Hold the resistor so that the single band is towards the right. The three colours on the left are read in the normal order from left to right. The first two bands *always* indicate numbers; the third band gives the number of zeros to add to the right of these two numbers. So, looking at the top resistor in Figure 2, yellow, violet, red means 4, 7, and two zeros, giving 4700 ohms! Looking at the lower resistor, brown, black, brown means 1, 0, and one zero, giving 100 ohms.

Remembering the order of the colours may be difficult at first. The colours from red to violet are the colours of the rainbow, in order, so if you know those, you're almost there! Around those colours are black and brown below the red, and grey and white above the violet, which you can imagine as getting brighter from black to white (well, almost!). It won't be long before you don't need to remember them at all.

The isolated band on the right-hand side is not part of the resistor's value; it indicates its *tolerance*, i.e. how close it might be to the indicated value. A brown band indicates ±1%, a red band ±2%, a gold band ±5% and a silver band ±10%. For example, a resistor marked as being 100 ohms with a ±5% tolerance will have an *actual* value somewhere between 95 ohms and 105 ohms.

8 Waves – Part 1

Introduction

Waves are responsible for most of the processes in life where energy is transferred from one place to another. Heat and light energy from the sun, for example, come to us as *electromagnetic waves*. Sound travels through the air as a wave; it is not the same sort of wave as light or heat, but it obeys many of the same properties. Damage is caused to coastal margins by the waves of the sea – again, another type of wave, but still obeying many of the same properties. Radio waves are of the same type as heat and light waves, as are gamma rays, X-rays, ultra-violet waves and infra-red waves. So, once we begin to understand what radio waves do, we are also learning about a huge chunk of physics at the same time! All these waves are part of the *electromagnetic spectrum*. The word ‘spectrum’ simply means a ‘range’, so what we have is a range of electromagnetic waves – that’s all!

Sensing things

Light waves are invisible, but our eyes can detect the *effect* they have on different materials because the waves produce an effect on the retina of the human eye which the brain can interpret. We cannot see heat waves either, but we can feel the *effect* they have on our skin. Gamma rays and X-rays are also invisible, but their detrimental *effects* on human tissue are well known. It is not surprising, then, that we cannot see radio waves. We cannot sense them, either, until we produce a device upon which they have an *effect*. That device is a **radio receiver**; it is able to process certain characteristics of radio waves, and make these characteristics audible by generating sound waves from the loudspeaker or headphones. Other characteristics of the same waves may be turned into light as a TV picture on a cathode-ray tube, or as a fax image on a sheet of paper.

Visible waves

Let’s start our description with some waves that we can actually see! When a small stone is thrown into a pond, we see circular water waves *radiating* from the point where the stone fell into the water, as **Figure 1** shows. (Notice that we use the word *radiating*, even with water waves; it is not a radio term, but one which describes any motion where the *radius* of a circle is increasing. In this case, it is the radius of the circular waves which is

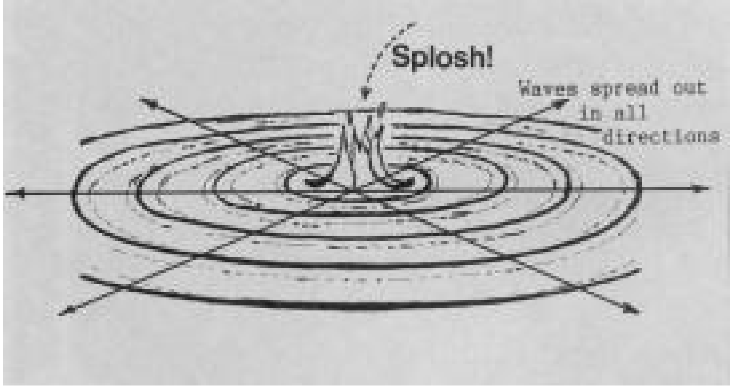

Figure 1 The waves move out from the point where the stone landed

increasing.) If you were to watch the water waves down at the water level, perpendicular to the direction in which they are travelling, you would see something like the illustration in **Figure 2**.

The horizontal line represents the water level before the wave started, and the vertical line represents the direction in which the water is displaced at any instant.

All waves are described in the same way

'Freezing' the motion of the water in this way allows us to define two very important characteristics of a wave, characteristics which we talk about every day – *wavelength* and *amplitude*. The wavelength of a wave is simply the distance (measured in metres) from any point on one wave to the same point on the adjacent wave. Look at the diagram and you will see what is meant by 'the same point on the adjacent wave'. The amplitude of a wave

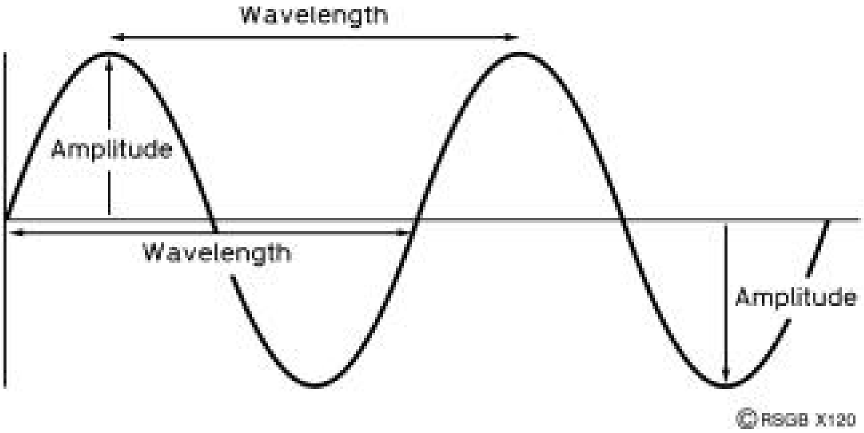

Figure 2 A water wave, viewed in cross-section

is always measured from the centre (undisturbed) position of a wave to the peak (or the trough) of the wave. Both these positions are shown, the arrow indicating that the measurement is taken *from* the centre *to* the peak or trough. The amplitude of a wave is defined as the maximum displacement of the wave from the centre position – the direction (up or down) of that displacement does not matter. Waves of greater amplitude carry more energy with them.

If we now 'unfreeze' the wave, we will see it travel from left to right (or right to left, depending on where we are looking). The speed at which it moves is called its *velocity*, and is measured in metres per second.

Another useful word is *propagate*; it means *travel*. We talk about radio waves propagating from a transmitter to a receiver. This velocity of propagation (for electromagnetic waves) is very fast indeed – they will cover 300 million metres in one second. This is virtually incomprehensible, so think of a radio wave travelling around the earth – it can travel 7½ times round the earth in one second! We use the symbol *c* for the velocity of radio waves (which is the same as the velocity of light, of course – all electromagnetic waves travel at this speed through air and space).

The last thing we need to know about the wave is its *frequency*. Imagine a cork floating on the water in the path of the wave; it will bob up and down. If we were able to count the number of times it went through its highest position in one second, then that number would be its frequency. Any periodic motion like this is said to go through one *cycle* each time one complete wave passes a point (in this case, our cork). We are thus counting the number of cycles per second of the cork's motion. The unit of frequency is thus 'cycles per second'; this unit is now named after Hertz, a radio pioneer, and is abbreviated to Hz.

Our description of the wave is now quite simple – we need only four quantities:

(a) **Frequency** symbol f – unit, hertz (Hz)
(b) **Wavelength** symbol λ (Greek letter lambda, pronounced 'lamb-da') – unit, metre (m)
(c) **Amplitude** symbol a – unit depends on application
(d) **Velocity** symbol c – unit, metres per second (m/s or ms^{-1})

The basic formula

Whatever may happen to a wave while it travels through different media (vacuum, air, brick, wood, etc.), one thing and **only** one thing remains constant – its frequency. Its wavelength, amplitude *and* velocity may change, but its frequency never does. Three of the four characteristics already identified are connected by the simple relationship

$$c = f \times \lambda.$$

Remember that c is constant if the wave travels in air or in a vacuum. This means that waves having higher frequencies (f large) must have smaller wavelengths (λ small) and vice versa. You can imagine frequency and wavelengths being on opposite ends of a see-saw!

Divisions of units

Because the frequencies of radio waves are so high (despite them having the lowest frequencies in the electromagnetic spectrum!) we have a problem with writing them down. Do you write in your log book that you have just heard a station on 14 100 000 Hz? Of course not, you write it as 14.1 MHz, knowing that the prefix mega (M) means ‘one million’. The prefixes which you need to know (when applied to frequency) are:

kHz	kilohertz	meaning	1000 Hz
MHz	megahertz	meaning	1 000 000 Hz
GHz	gigahertz	meaning	1 000 000 000 Hz.

Notice that the ‘k’ in kilohertz is a lower case letter. It is **incorrect** to write it as an upper case letter. ‘K’ is a computer-related prefix meaning **not** 1 000 but 1 024!

When we come on to discuss heat and light waves, we will use wavelengths rather than frequencies, because of the see-saw effect – as the frequencies get larger and larger, the wavelengths get smaller, and hence are numbers which are more manageable, both to talk about and to write down!

Bands

Gamma rays, X-rays, ultra-violet waves, light waves, infra-red waves are all part of the electromagnetic spectrum, but we divide them up because they have different properties. This is why we divide up our radio frequencies into different bands. The radio waves of top-band signals (around 2 MHz) have completely different properties compared with those in the 20 metre band, so we are dividing up the *radio spectrum* in the same way – by property.

9 A beat-frequency oscillator

Introduction

Many readers will know that, although they have a short-wave radio which covers at least one of the amateur bands (e.g. 7 MHz or 14 MHz), they are unable to listen to SSB or Morse signals. This is because the receiver lacks a Beat-Frequency Oscillator (BFO). We need the ‘carrier’ frequency of a BFO to replace the carrier that has been removed from the signal at the transmitter. When listening to Morse signals, the BFO signal ‘beats’ with the incoming signal to produce a note in the loudspeaker. If you are a musician, you will be familiar with the method of using ‘beats’ to tune one musical instrument from another; in the BFO, the beat frequency produced is the tone signal you hear.

In the more complex amateur radio receiver, a BFO is incorporated as part of the whole system. In our model, it is an external circuit that sits alongside your radio. The circuit diagram is shown in **Figure 1**.

Construction

Built on a small piece of matrix board about 80 × 50 mm, the circuit can be fitted inside a small plastic box. For once, we *don’t* want to screen the circuit to prevent it interfering with other equipment; we *want* it to interact

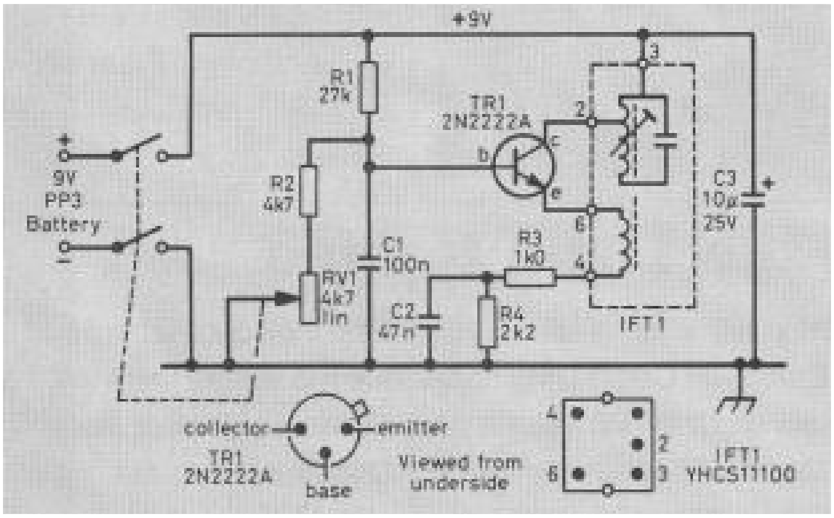
+9V
R1
27k
TR1
2N2222A
b
c
e
+
9V
PP3
Battery
−
R2
4k7
C1
100n
RV1
4k7
lin
C2
47n
R3
1k0
R4
2k2
3
2
6
4
IFT1
C3
10µ
25V
+
collector
emitter
base
TR1
2N2222A
Viewed from
underside
4
2
6
3
IFT1
YHCS11100

Figure 1 Circuit diagram of the BFO

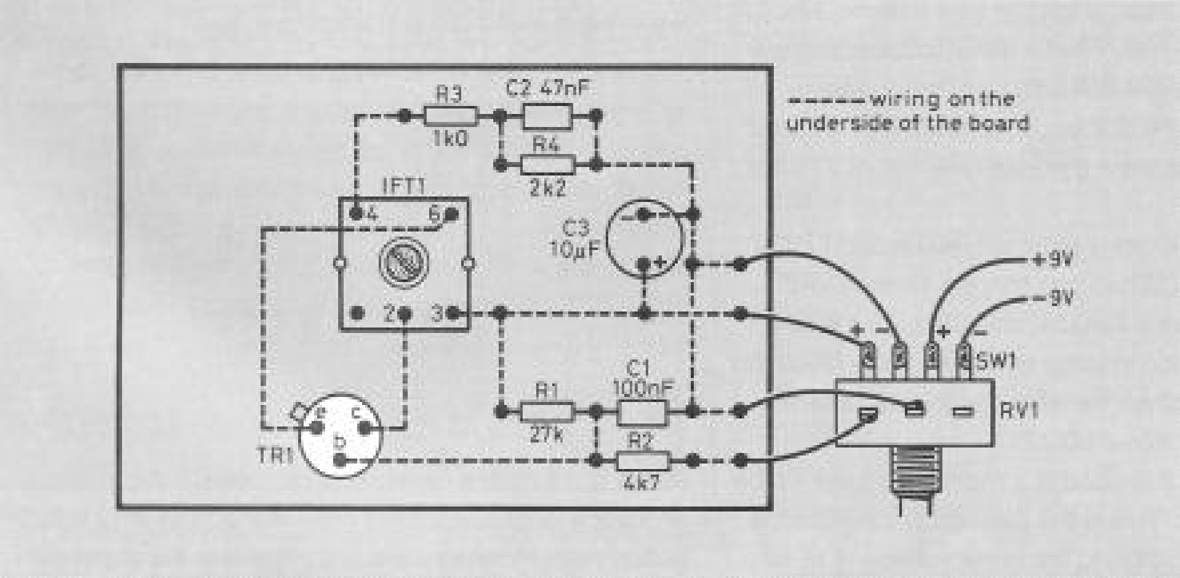

Figure 2 Matrix board layout shown from the component side. Adjust IFT1 carefully for the best results

with our receiver! This is why we use a plastic box. Maplin Electronic Components supply a suitable box, complete with the matrix board to fit inside (order code YU46).

Look at **Figure 2** carefully before you start to build the circuit, so that you can position the components correctly. Firstly, inspect the intermediate frequency (IF) transformer, IFT1, and remove its metal screening can very carefully. Again, this is to allow some signal to escape from our circuit and enter our radio. Having done this, solder the components, using the matrix board as a support. Underneath the board, the components are linked by single-conductor, insulated wire. Take particular care with the polarity of the electrolytic capacitor, C3, and the connections to the transistor, TR1 (see Figure 1).

The variable resistor, VR1, has a switch mounted behind the control itself, and the insulated leads to it from the battery should be about 10 cm long. Connect these before fitting VR1 into the case, so that the BFO can be calibrated (adjusted) correctly.

Calibration

After a final check that all the components have been fitted and soldered correctly, connect the battery, switch on, and hold the transistor between your fingers, to check that it is not getting hot. Place the circuit close to your receiver, and set RV1 to mid-position. Tune your receiver to find an amateur SSB transmission; the frequencies listed below will help you in knowing where to look. It may sound very strange, but don't worry. *Slowly* turn the core of IFT1 with a small, non-metallic screwdriver or with the correct 'trimming tool'. The core into which the blade fits is very fragile, so attempt this process with care. When the speech sounds as natural as you can get it, leave the core at this position, and use RV1 to make the speech sound natural.

Using the BFO

For best results, you may have to move the BFO nearer or further away from your radio. At the lower end of most bands (for instance just above 7.000 MHz or 14.000 MHz) you should be able to resolve Morse code (CW) signals. If you find that the BFO signal is a little weak, solder a 15 cm length of insulated wire to pin 2 of IFT1, and place it alongside your radio. This should improve signal intelligibility. When you are happy with the performance, switch off, drill a 10.5 mm hole in the box and fit RV1, followed by the matrix board assembly. Screw the base to the box, fit the knob, and the BFO is complete!

Where to listen

Band	**Frequencies (MHz)**
15 m	21.000–21.450
17 m	18.068–18.168
20 m	14.000–14.350
30 m	10.100–10.150
40 m	7.000–7.100
80 m	3.500–3.800

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	27 kilohms (kΩ)
R2	4.7 kilohms (kΩ)
R3	1 kilohm (kΩ)
R4	2.2 kilohms (kΩ)
VR1	4.7 kilohms (kΩ), linear, with DPST switch

Capacitors

C1	100 nanofarads (nF) or 0.1 microfarad (μF), ceramic
C2	47 nanofarads (nF) or 0.047 microfarad (μF), ceramic
C3	10 microfarads (μF), 25 V radial, electrolytic

Additional items

TR1	2N2222A npn
IFT1	Toko type YHCS11100
Box	plastic, approximate size 100 × 70 × 45 mm
Board	matrix, to fit inside the box
Connector	for PP3 battery
Knob	for RV1

10 What is a capacitor?

Introduction

A capacitor is a device that stores energy in the form of electricity. Much less energy than a battery, and for a shorter time, however. The simplest form of capacitor takes the form of two flat metal sheets separated by air; connections are made to each plate, as **Figure 1** shows. If you imagine a pair of these plates, 30 cm square and separated from each other in air by 1 mm, the capacitance of this device would be almost exactly 80 picofarads (pF), i.e. 80 million-millionths of the unit of capacitance, the farad. Now this is quite a small value, you will agree, and it comes about because the farad is such a large unit. Nevertheless, as you will probably know, we may have capacitors of value 10 000 microfarads (μF) in our radio equipment, and they can be smaller than your little finger, so they are obviously not made the same way!

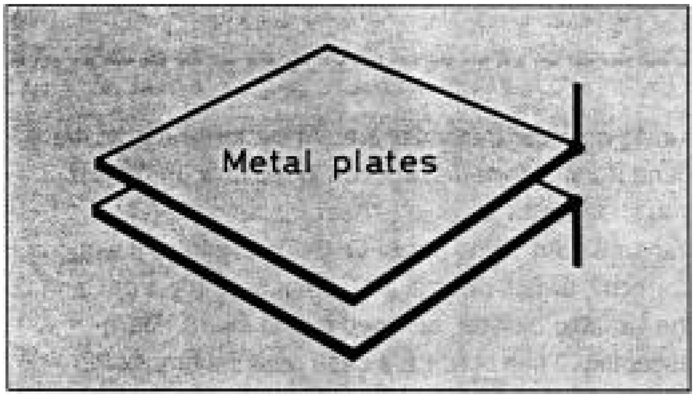

Figure 1

Large and small

We cannot go into the manufacture of capacitors here – after all, we are users of the devices, not the designers! First of all, beware of incorrect statements; the words 'capacitor' and 'capacitance' are *not* the same. For example, a large capacitor would be a description of one the size of a toilet roll. It need not have as large a capacitance as one the size of your little finger. A 'large capacitor' is one which is physically *big*, a 'large capacitance' refers to a capacitor which can store a larger amount of energy when a certain voltage is applied between its plates. The capacitors in a mains power supply are usually big **and** have large capacitances. High-power RF amplifiers may have large capacitors with small capacitances!

Electrolytics . . .

Electrolytic capacitors usually have capacitances of 1 μF or above. They differ from other capacitors in that they **must** be connected the right way round (i.e. they have positive and negative connections, just like a battery). They may explode if the connections are reversed!

. . . and the others

Other capacitors may be connected either way round, despite their names. We have *polystyrene*, *ceramic*, *silver-mica* and *tantalum*. Each has its own advantages and disadvantages, and the parts list for a project will always tell you which type is best.

Storing energy

If you were to connect a large capacitance across a 12 V power supply, nothing would appear to happen. Removing the capacitor from the supply and connecting it to a voltmeter would show that the capacitor has 12 V between its ends. This shows that, while the capacitor was connected to the supply, energy flowed from the supply into the capacitor. We say that the capacitor was *charged up* by the supply. If you are using an *analogue* voltmeter (i.e. one with a meter and pointer), you will notice that the indicated voltage slowly drops until, eventually, there is no voltage across the capacitor. This is because the capacitor has *discharged* its energy into the voltmeter. If you had used a smaller capacitance, the same would happen, except that the voltage would drop to zero more quickly – the capacitor stores a smaller amount of energy because its capacitance is smaller. Capacitors behave like other things in life – a small car can move more quickly than a large bus – a small piccolo emits a higher note than a flute – the voltages in a circuit containing a small capacitance will change more quickly than those in a circuit with large capacitance.

Varying the capacitance

Some capacitors are capable of having their capacitance changed manually; these are called *variable capacitors*. They work like the basic capacitor of Figure 1. Imagine moving the top plate of the pair a little to one side; the capacitance is determined, not just by the size of the two plates, but by their *area of overlap*. As this decreases, so does the capacitance. Such devices are limited in their capacitance, about 500 pF being the maximum value.

AC and DC

Because the plates of a capacitor do not touch each other, a direct current (DC) cannot pass between them. However, an alternating voltage on one plate can induce an identical alternating voltage on the opposite plate, and thus a capacitor appears to pass an alternating signal, even though currents as such, do not pass between the plates. This property of passing AC and not DC is very important, and a capacitor used in this way is called a *DC blocking capacitor* or, simply, a *blocking capacitor.* A blocking capacitor can be used at the same time, to couple a signal from one circuit to the next; here it would be known as a *coupling capacitor. Decoupling capacitors* are to be found where the capacitor is employed to remove an AC signal while retaining a DC component.

Finally. . .

Unlike resistors, the manufacture of capacitors renders them susceptible to excess voltage, so if you find a capacitor labelled 10 μF 16 V, it means that operating it above 16 V may fatally damage the device (and the circuit around it). This voltage is called the *working voltage* of the capacitor; on some electrolytics, you may find it expressed as *volts working* (i.e. 8 μF 450 V WKG).

Many smaller capacitors have their properties marked on them in a colour code, like resistors. **Figure 2** shows these codes, and their meaning, and the table below summarises the values of the colours.

Table 1

Colour	*Value*	*Voltage (tantalum capacitor)*	*Voltage (polyester capacitor)*
Black	0	10	–
Brown	1	–	100
Red	2	–	250
Orange	3	–	–
Yellow	4	6.3	400
Green	5	16	–
Blue	6	20	–
Violet	7	–	–
Grey	8	25	–
White	9	3	–

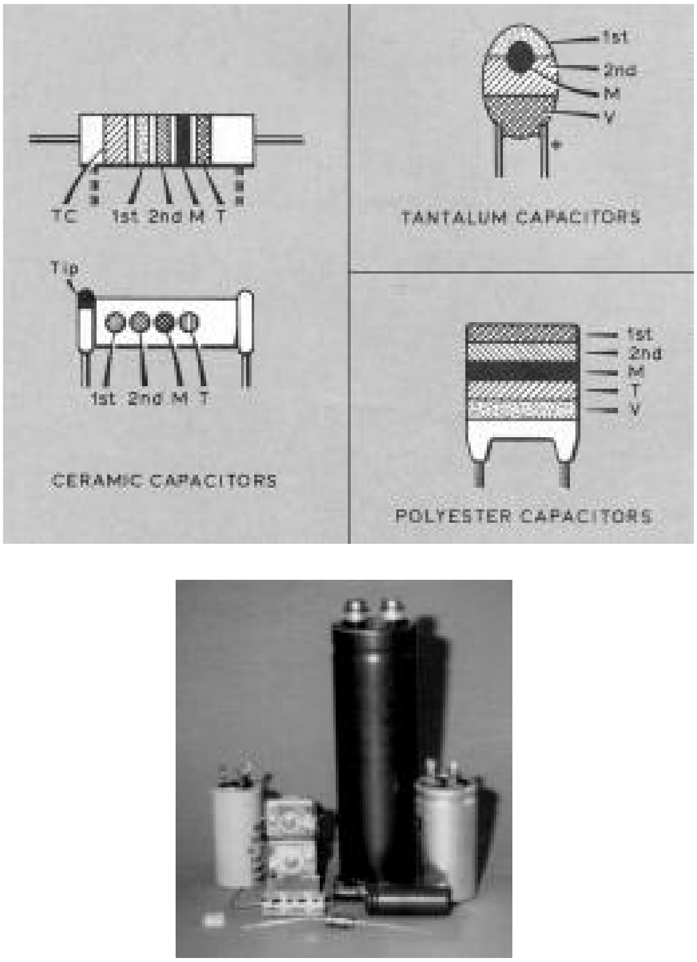
TC
1st
2nd
M
T
TANTALUM CAPACITORS
1st
2nd
M
V
+
Tip
1st
2nd
M
T
CERAMIC CAPACITORS
1st
2nd
M
T
V
POLYESTER CAPACITORS

Figure 2 Some capacitors have coloured bands or stripes, rather like resistors. The colour code, which is the same as the resistor code, is shown in Table 1. The band shown on the chart as ‘1st’ is the first number of the capacitor’s value in pico-Farads, ‘2nd’ is the second number and ‘M’ is the Multiplier or number of noughts. For example, a capacitor reads from the top: Brown, Black, Yellow, Black, Red. Its value is One, then Nought, then Four more noughts = 100 000 pF (also referred to as 0.1 μF or 100 nF). Its tolerance (Black) is 20% and the working voltage (Red) is 250 V. The ‘V’ means the maximum working voltage. The band marked ‘T’ shows the tolerance, just like resistors, and the one marked ‘TC’ is only used on special capacitors designed to change their value with temperature

11 Waves – Part 2

Introduction

We left Part 1 with the concept that radio waves are divided up into bands which have different properties. Not *all* the properties are different, though. We need to discuss several wave properties, so we will start with what happens to waves as they propagate over long distances.

Getting weaker

Imagine a torch battery connected to a bulb by wires about 1 metre long. The bulb lights normally. If we now take the bulb 100 metres away from the battery and wire it up, we would expect the bulb to be somewhat dimmer, which is exactly what would happen. It happens because of the resistance of the wires – the wires do not form a perfect conductor. A similar situation occurs with radio waves.

All waves suffer from *attenuation* – they get weaker the further they travel. In cases of extreme attenuation, we need to apply some *amplification* before the attenuated wave can be used in a receiver.

Carrying information

When we speak over the telephone, the range of frequencies in our voices extends from very low frequencies up to about 15 or 20 kHz. In audio terms, this is quite a large *bandwidth* (meaning a wide band of frequencies). For communications purposes, however, most of this bandwidth is not needed, and in the telephone system (and in our transceivers), this is cut down so that it extends from about 200 Hz to 3 kHz, a reduction in bandwidth from 20 kHz to about 3 kHz. A bandwidth of 3 kHz has been found to be sufficient to convey speech intelligibly which, after all, is just what we need!

The radio waves coming from an amateur transmitter convey our speech signals over long distances. By themselves, the speech signals do not travel very far, so they have to be combined with a radio signal that *will* travel long distances. This extra signal is called the *carrier wave* (or just the *carrier*), because its job is to *carry* the speech signals along with it! The process of combining the speech (or Morse code) with the carrier is called *modulation*.

Wider and wider

Sending Morse code is achieved simply by switching (or *keying*) the carrier on and off. The bandwidth of the transmitted signal is only about 100 Hz. Speech, with its reduced bandwidth of 3 kHz, will produce a *single-sideband* (SSB) transmitted signal with a bandwidth of 3 kHz. If the same speech signal were used to produce an *amplitude-modulated* (AM) signal from the transmitter, it would have a bandwidth of about 6 kHz. Perhaps you can now understand why the bandwidths needed to produce hi-fi broadcasts need to be so large. TV signals need bandwidths running into tens of megahertz!

Waves need aerials

Radio waves are produced whenever changing currents flow through a wire, and when that wire is made in such a way as to maximise the *radiation* from the wire, it is called an *aerial* or *antenna*. The same piece of wire will receive radiation from other aerials; an aerial will transmit and receive. This is an important property of the aerial: when a current flows through it, electromagnetic waves are launched into the air; when electromagnetic waves in the air encounter the aerial, currents are produced in it.

From the simplest transmitting aerial, waves travel in all directions, like the waves on the pond that we considered in Part 1. They will travel a long way through air and space before they become too week to be received. They won't travel very far into the earth, however! The earth will *reflect* some of the wave and will *absorb* the rest. That portion of the wave which is reflected will again travel through air and space until it is totally attenuated.

Look at **Figure 1**; A represents a radio transmitter, with B1, B2 and B3 being receiving stations. The two arrows pointing 'downwards' from A represent two of the waves from A which just graze the earth's surface. Waves above

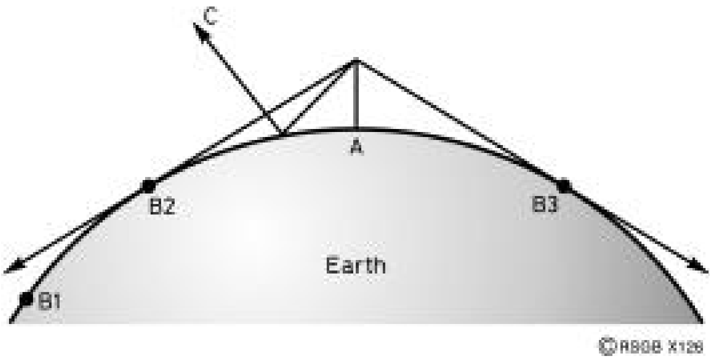

Figure 1 Stations between B2 and B3 receive the *ground wave*

these will travel on into space; waves below them will either be absorbed and reflected by the earth or received by aerials. B1 will not receive any signals from A, because it is below A's horizon. B2 and B3 can receive A's signals because they are *just* on A's horizon. Any stations between B2 and B3 will also receive A's signals, which are known as *ground-wave signals*. The wave at C represents one which is reflected by the ground and travels into space.

This description begs the question of how signals are received from stations well beyond the ground-wave range.

Mirrors in space

Suppose that there was something, out in space, that would reflect radio waves. Waves from A that travel out into space could be reflected off it and return to earth, enabling stations such as B1 to receive A's signals. The situation just described is illustrated in **Figure 2**.

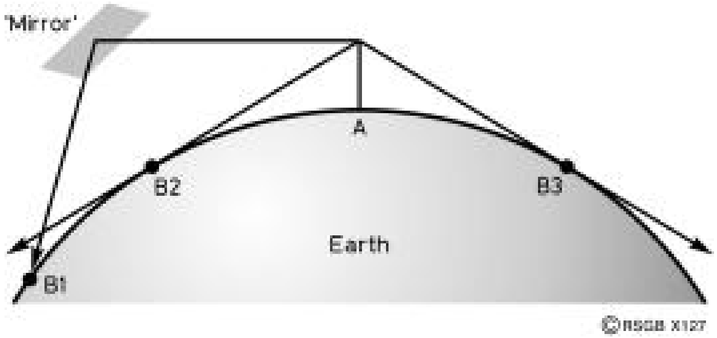

Figure 2 Reception of A's signals at B1 using the *sky wave*

Just such a mirror in the sky really *does* exist. It is not man-made, of course! The earth is surrounded by the atmosphere, a mixture of many gases, such as nitrogen and oxygen. The energy contained in the radiation from the sun is more than sufficient to *ionise* these gases, thus making them into electrical conductors. When a gas is ionised, some of its electrons are physically stripped out of the atoms and are free to move about, just as the electrons of a metal do in a wire. Consequently, we can regard this part of the atmosphere (the part illuminated by the sun) as acting like a sheet of metal, which *reflects* radio signals! It is not a perfect reflector, but is sufficient to produce long-range (DX) radio propagation via the *sky wave* under the right conditions. (A more down-to-earth example of ionised gases conducting electricity can be found in the fluorescent tube and the neon sign – many gases glow when they are continuously ionised.)

This conducting region at the extremity of the atmosphere is called the *ionosphere*, and it exists in layers between 60 km and 700 km above the earth's surface. When the ionosphere is sufficiently ionised, it glows; this is the natural phenomenon known as the *aurora borealis*, or the *northern lights*.

The property of the ionosphere that enables radio waves to be reflected does not act in a uniform way; it is very selective about which waves it reflects, and which waves go straight through it and into outer space. In general, it reflects only those waves with frequencies below about 30 MHz – the HF bands!

In Part 3 we will look at families of waves.

12 An LED flasher

Introduction

The LM3909 is an integrated circuit (IC) which will flash a light-emitting diode (LED). Using only two extra components and a battery, the circuit is cheap and has a very low current drain from a 1.5 V cell. The circuit can be used as a novelty flasher, an indicator for a dummy alarm bell box, or it could be attached to a torch so that it could be found easily in the dark! The simple circuit is shown in **Figure 1**.

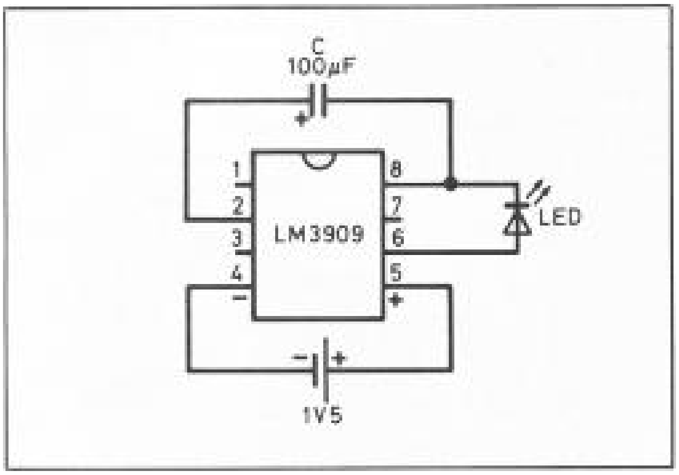

Figure 1 Circuit diagram of the LED Flasher. Pins 1, 3 and 7 of the IC are not used

Assembly

The circuit can be built on a small piece of Veroboard (the piece shown in **Figure 2** measures 15 holes by 10 strips). Using such a board, follow these instructions.

1. Depending on how far away you want the LED from the circuit board, solder a length of insulated wire to each lead of the LED. Use different colours of insulation – say, red and black, connecting the *red* lead to the *anode* (a) lead (the longer one) of the LED, and the black one to the *cathode* (k). Figure 2 shows these leads.
2. Cut the copper tracks as shown in Figure 2, using a 3 mm (⅛ inch) diameter drill, rotated between thumb and forefinger, or use the proper tool. Make *absolutely* sure that the tracks are completely broken!
3. Fit the IC holder in the correct position, using the cut tracks as guides, and make sure the small notch is facing towards the top of the board. Solder the pins to the copper tracks.
4. Mount the capacitor, positive end to the left, so that the positive lead is soldered to track F, which connects it to pin 2 of the IC; the negative lead is soldered to the right-hand side of track E, this being connected to pin 8 of the IC.
5. Solder on the battery leads, positive to the right, and the extended LED leads, positive downwards.

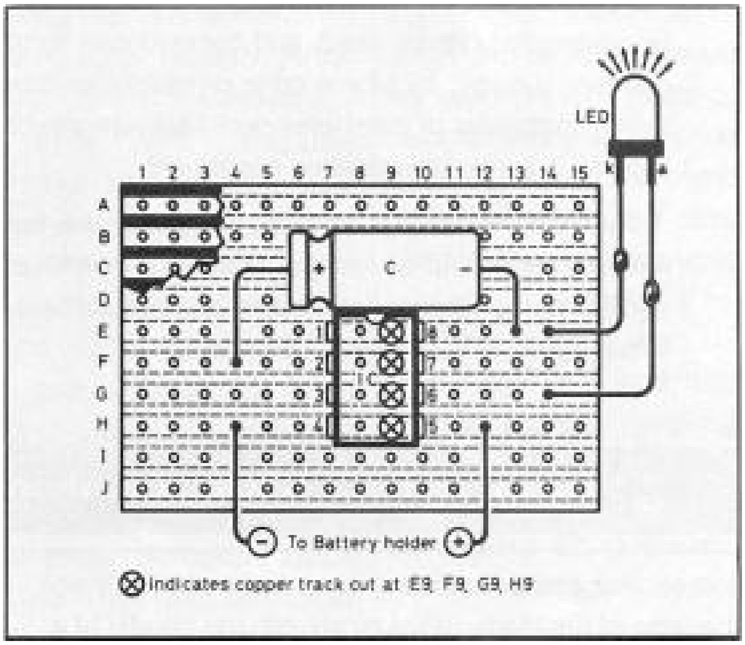

Figure 2 Board layout viewed from the component side. The tracks are cut under the board where shown

6. Check the circuit, and hold up the board to a bright light and look carefully for solder bridges between the tracks and pieces of copper swarf which may have escaped your inspection in 2 above! Remove whatever you find.
7. When all seems well, put the IC into the socket, ensuring that the notch or dot on the upper surface of the IC lines up with the notch on the holder. Line up each pin on the IC with the hole below it before pressing *gently* on the IC with the board supported on a firm surface.
8. Connect the battery; the LED should start to flash. The circuit is complete and working!

If you prefer, the whole circuit (battery included) can be mounted in a small plastic box, with the LED mounted on a clip and protruding through the panel. There are many other possibilities, and it is up to you to find an application for your own use.

Parts list

		Maplin code
LM3909	Integrated circuit	WQ39N
IC socket	8-pin DIL	BL17T
LED	5 mm diameter	WL27E
100 microfarad (μF)	Electrolytic capacitor (10 V)	FB48C
Battery holder	For AA-size cell	YR59P
Battery	1.5 V AA cell	

Small piece of Veroboard (15 holes by 10 strips)
Small plastic box (if required)
LED clip (if required)
Two lengths of coloured, insulated wire for LED (as required)

Availability

All parts can be obtained from Maplin Electronics Ltd.

13 Waves – Part 3

Introduction

When we talk about the spectrum being divided up into bands, this is just for our convenience; there are no natural divisions although, as we have seen, some of the properties of different bands really *are* different! Let's look at **Figure 1**, and see how the frequencies are divided up.

The divisions

- **Very low frequencies** (VLF) cover the range from a few kilohertz up to 30 kHz. Very long-range communication is possible, but at *very* small bandwidths. It is used for special purposes.
- **Long waves** (LW) are used for medium-distance commercial broadcasting and have frequencies from 30 kHz to 300 kHz.
- **Medium waves** (MW) are used for commercial broadcasting, and use frequencies from 300 kHz to about 1.5 MHz (1500 kHz). Typical range is about 200 km.
- **Short waves** (SW) encompass both the low-frequency (LF) and high-frequency (HF) amateur radio bands. There are nine narrow amateur bands in the SW spectrum between 1.8 MHz and 30 MHz. Some of these bands give round-the-world communication.
- **Very high frequencies** (VHF) span the range between 30 MHz and 300 MHz. Relatively short-range communication is possible. They were once used for broadcast TV before it moved to UHF. There are now three

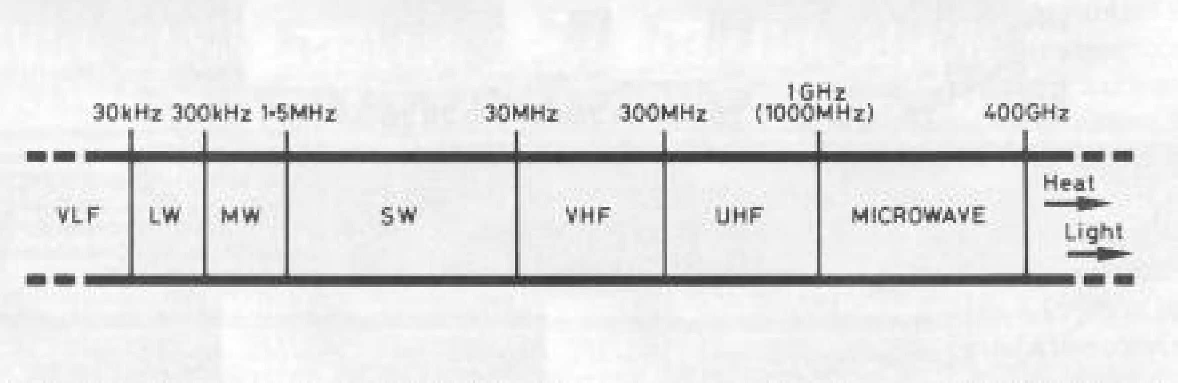

Figure 1 Diagram of radio frequency spectrum

amateur bands here – 6 m, 4 m, and 2 m. Repeaters are used to extend the usable range of mobile stations. VHF waves are not usually reflected by the ionosphere, but when they are, ranges of several thousand kilometres are possible. Weather affects these waves on a regular basis, however. In addition to amateur users, the VHF part of the spectrum is also used by the police, the fire and ambulance services, weather satellites, and many others,

- **Ultra-high frequencies**, sometimes called *centimetre waves*, cover the range from 300 MHz to 1000 MHz (or 1 GHz). The only amateur band in this range is the 70 cm band, and we share it with radar, TV and cellular telephone users as well.
- **Microwaves** begin at 1 GHz and extend to about 400 GHz. They are *never* reflected by the ionosphere, are partially attenuated by buildings, and are reflected from aircraft and cars. Microwave absorption in the atmosphere is quite significant, and rain and fog can attenuate microwaves quite heavily.
- **Heat, light** . . . Above 400 GHz we run into the infra-red bands and on into the visible light and ultra-violet bands. We generally take 400 GHz as being the limit of what we class as radio waves.

Bandwidth again

Complex signals need more bandwidth than simple signals. Even if it were possible, we would not be able to transmit a single television channel in the whole of the MW broadcast band! When TV used part of the VHF band, only five channels were possible in the range from 45 MHz to 68 MHz. By moving TV to the UHF band, we now have 47 channels between 470 MHz and 855 MHz!

It's your choice!

The number of permutations you have amongst all the modes and all the bands is enormous! Only you can decide what you are interested in and what you want to learn about. That is the attraction of amateur radio!

14 Choosing a switch

Simple, but be careful!

A switch is the simplest electronic component, but care is needed in choosing the correct one for the job. Here is a list of the main characteristics of a switch, which will help you to select what you want.

1. **Rating**. This gives the maximum voltage and current that a switch can handle. For example, a 250 V 1.5 A switch will switch mains voltage at a current not exceeding 1.5 A. If the current is greater than 1.5 A, the switch may get hot and fail. If the voltage is too great, the switch may arc each time it is switched off, thus wearing away the contacts.
2. **Number of poles**. A single switch can control many circuits; the number of poles tells you *how many* different circuits it can handle. See **Figure 1**.
3. **Number of throws**. This tells you the number of positions each pole can have. This is best illustrated in Figure 1. The simplest ON/OFF switch is a Single-Pole, Single-Throw (SPST) switch. A Single-Pole,

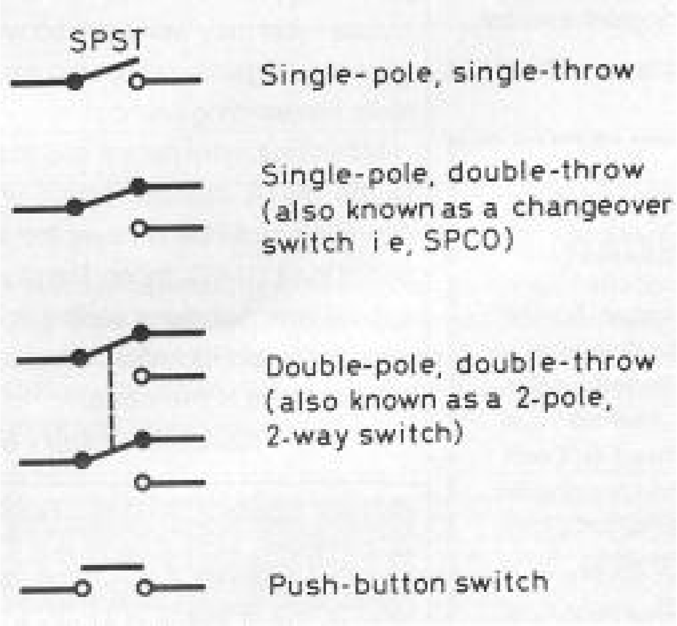
SPST
Single-pole, single-throw
Single-pole, double-throw (also known as a changeover switch i e, SPCO)
Double-pole, double-throw (also known as a 2-pole, 2-way switch)
Push-button switch

Figure 1 Some circuit symbols for different types of switch

Double-Throw (SPDT) switch may also be used as an ON/OFF switch, but is used mainly to change between two parts of a circuit, and is commonly known as a Single-Pole Change-Over (SPCO) switch. Two or more SPCO switches can be operated at once; Figure 1 shows an example. The two switches are said to be *ganged*. See also 'Types of switch' below.

4. **Number of ways**. When a switch has more than one throw, we tend to use the word 'ways' instead. This means that if a switch has one pole and six throws, we would normally call it a '1-pole 6-way' switch. Such switches tend to be rotary switches, as are described below.

Types of switch

(a) **Push-button**. These are found on calculators, telephones, electronic games and most equipment with a digital display.

(b) **Rotary**. These are switches controlled by a knob, and are turned instead of moved up and down. **Figure 2** shows the rear of such a switch and its circuit symbol. The commoner types of rotary switch are: 1-pole, 12-way; 2-pole, 6-way; 3-pole, 4-way; 4-pole, 3-way; 6-pole, 2-way. All these switches have 12 click positions, as you may have guessed, but each one comes with an adjustable end-stop so that you can set the correct number of ways according to the contacts on the switch.

(c) **Slide**. This switch is common on the cheaper types of radio, mainly as an ON/OFF or band-changing switch. They are not very rugged, but are small and cheap to produce. Very small types are manufactured for use on PCBs.

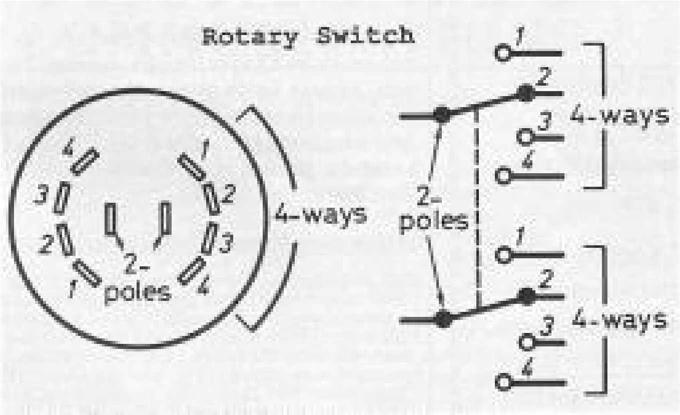

Figure 2 The rotary switch can select several different circuits at once

(d) **Toggle**. Available as two-way types or three-way, with a centre-off position. Can be manufactured to handle very high currents.

(e) **Micro-switch**. These are devices which are usually operated indirectly, such as when the cover is removed from a high-voltage power supply, or when the door of a fridge is opened. The 'micro' part of the name doesn't refer to the size of the switch, but to the *small* movement that is required to activate it.

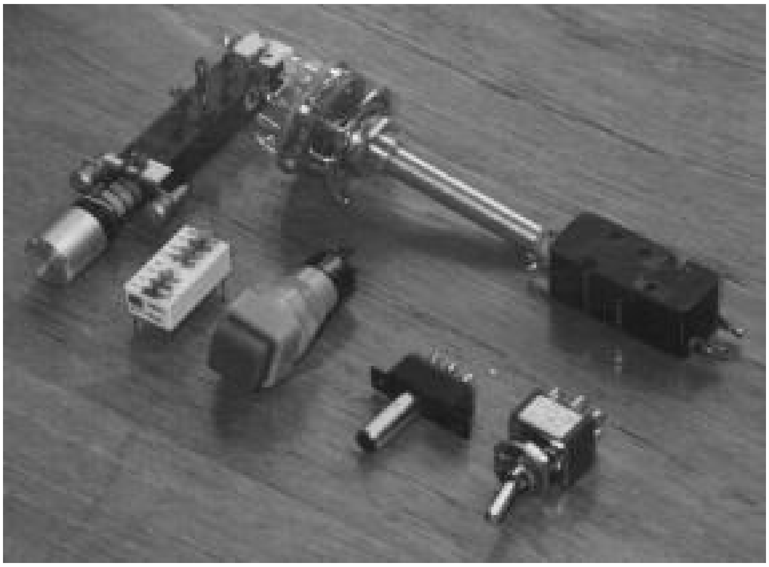

15 An aerial tuning unit for a receiver

Introduction

Any length of wire will act as an antenna (or aerial) but, to get the best results from a transceiver or receiver, an aerial tuning unit (ATU) is required. This matches the *impedance* of your aerial to the impedance at the aerial socket of your radio. Impedance is like resistance, and is measured in ohms, but it is used for alternating currents, and hence is common in audio and RF engineering. Most receivers have an impedance at the aerial socket of about 50 ohms (Ω); aerial impedance, on the other hand, can be anywhere between 20 Ω to over 1000 Ω, depending on its length and its height above ground.

Fortunately, you don't *need* to be able to calculate your aerial's impedance; all you need is a device that will perform the matching operation for aerials with a large range of impedances, and this is exactly what is described here! The subject of *matching* is a complex one, but all you need to know is that most signals will become clearer, and that there will be less noise and interference. Stations will become louder, so you will probably be able to reduce the setting of your RF gain control (always a good thing to do).

This design of ATU covers all amateur and broadcast bands from 10 m (28 MHz) to 80 m (3.5 MHz), and is very easy to build. The circuit is shown in **Figure 1**.

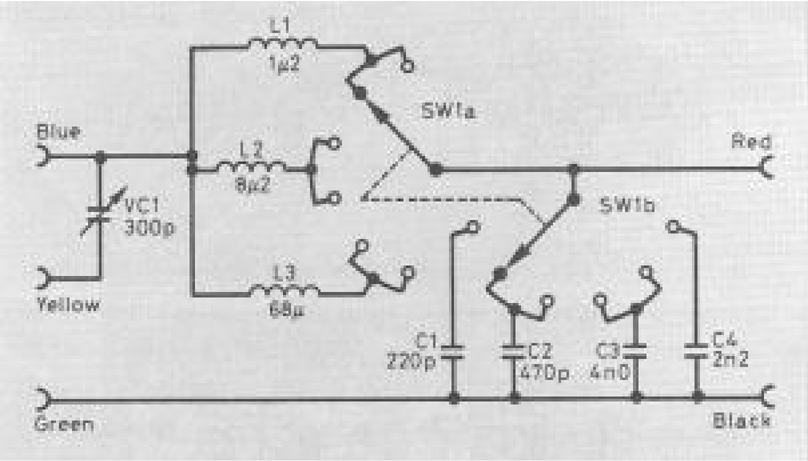

Figure 1 Circuit diagram of the Antenna Tuning Unit, showing the use of a 2-pole, 6-way rotary switch to select inductors (L) and capacitors (C)

Construction

1. Firstly, you will need a simple plastic case in which to house the ATU. The size should be approximately 85 × 145 × 50 mm.
2. Start by drilling two 10.5 mm holes in the front of the case; these are for the 6-way switch and the tuning capacitor.
3. Drill three 8 mm diameter holes in the left-hand side of the box, for the three sockets, coloured blue, yellow and green.
4. On the right-hand side of the box, drill two 8 mm diameter holes for the red and black sockets.
5. Now, fit the 6-way switch (SW1), the tuning capacitor (VC1), and all the sockets to the case. Check that the vanes of the capacitor rotate smoothly when the shaft is turned.
6. Wire up the inductors (coils). Figure 1 and the wiring diagram of **Figure 2**, will help with this. As you can see, each one side of each coil is connected to two switch connections, the other end going to VC1.
7. Solder in the fixed capacitors. One end of each goes directly to the ground socket (black), and the other end goes to the switch.
8. Solder a wire between the green and black sockets. The output from the ATU comes from the red socket, and this is connected to the two tags in the centre of the switch, as Figure. 2 shows clearly.
9. Finally, connect the blue and yellow sockets to the tuning capacitor, and the ATU construction is complete.

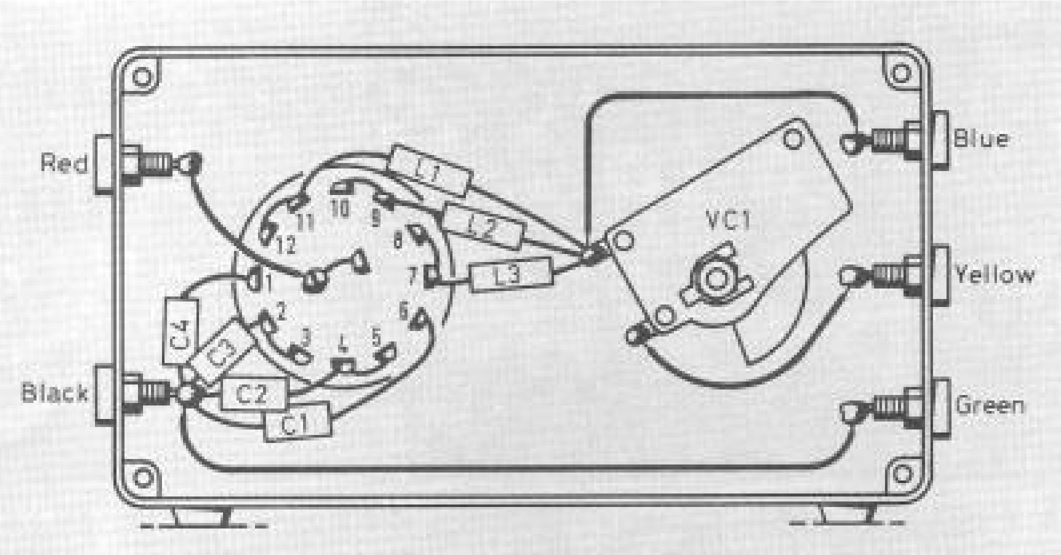

Figure 2 The internal view of the case shows the main tuning capacitor, VC1. This is a solid dielectric type, which has adjustable brass plates

In practice . . .

Figures 3 and **4** show two different ways of connecting your aerial to your ATU. In each case you will need to select each switch position in turn, and rotate the tuning capacitor through its full range while listening to a station. You should find that one switch position enables VC1 to produce a peak in the signal strength in the loudspeaker. At this point, your aerial and receiver are said to be *matched*. Stations in the same band will probably peak with VC1 at the same setting of SW1, but different bands will almost certainly require different positions of SW1.

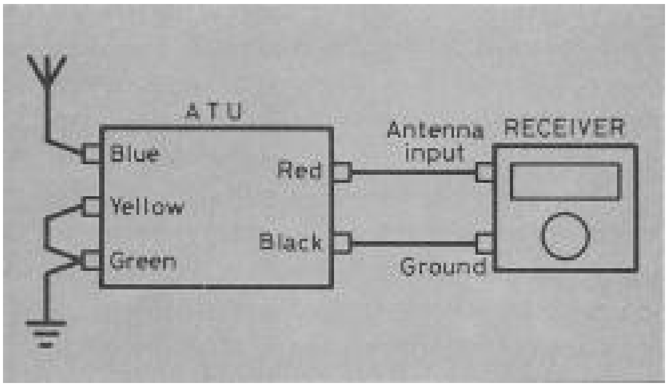
ATU
Blue
Yellow
Green
Red
Black
Antenna input
Ground
RECEIVER

Figure 3 You may find that parallel tuning gives best results with your antenna

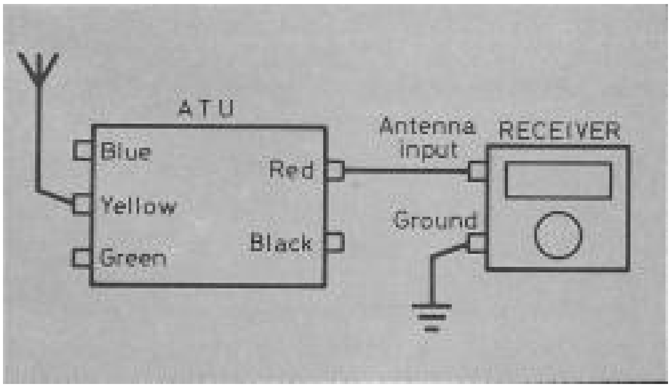
ATU
Blue
Yellow
Green
Red
Black
Antenna input
Ground
RECEIVER

Figure 4 In other cases, series tuning could be the most effective arrangement

Parts list

Capacitors (all rated at 16 V or more)

C1	220 picofarads (pF) polystyrene
C2	470 picofarads (pF) polystyrene
C3	1000 picofarads (pF) or 1 nanofarad (nF) polystyrene
C4	2200 picofarads (pF) or 2.2 nanofarads (nF) polystyrene

Inductors

L1	1.2 microhenries (μH)
L2	8.2 microhenries (μH)
L3	68 microhenries (μH)

Switch

SW1	2-pole 6-way rotary

Sockets
  4 mm type, one each of red, black, yellow, blue, green

Additional items
  Plastic or metal case, e.g. Maplin type YU54
  Two large knobs for SW1 and VC1

16 A simple 2 m receiver preamplifier

Introduction

Designed specifically to complement the modified air-band portable (also described in this series), this can be used with some success on many receivers suffering from ‘deafness’ on VHF.

The circuit

An RF preamplifier is a device which improves the input signal to an existing receiver, enabling it to work more effectively. Because of the noise which is added to the signal by the preamplifier, *very* weak signals may not be usefully enhanced, but stronger signals will be improved considerably.

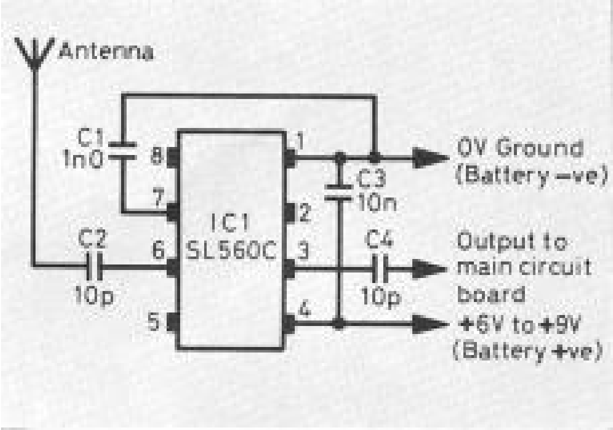

Figure 1 The integrated circuit is mounted upside down. Make sure you identify pin 1 which has a small dot next to it

This little circuit (shown in **Figure 1**) uses a GEC/Plessey integrated circuit type SL560C. With the addition of four capacitors, it is used between your aerial and the aerial input of the radio's PCB.

Putting it together

1. Use a small piece of prototype (matrix) board about 25 mm square. Use an 8-pin DIL socket for the integrated circuit (don't risk soldering the chip – it is seldom a risk worth taking). Figure 1 shows the connections to the socket, *looking from underneath*.
2. Make special note of the pin numbers, so that you know how to put the chip into the socket when you have finished. The positive and negative connections to the circuit are taken from the main PCB *after* the ON/OFF switch – so that the switch operates the preamplifier, too.
3. Unsolder the lead to the radio's telescopic aerial and connect it to the free end of C4, as shown. Then solder a short lead between the telescopic aerial and C2.
4. Sometimes it is possible to cajole your little preamplifier into the radio's plastic case, provided there is room and that you make sure that none of the soldered joints on your little board touch any of the metal inside the case when you replace the back and screw it on again.
5. If there is not enough room inside, then put the preamplifier into its own box, with battery and switch, and its own aerial. Keep the lead from the preamplifier to the aerial connection of the radio as short as possible – perhaps using screened cable.

Because your preamplifier is untuned, you will find not only that it helps with reception on 2 metres, but also that reception on the FM broadcast band is improved!

Parts list

Integrated circuit

IC1	GEC/Plessey SL560C

Capacitors

C1	1 nanofarad (nF) ceramic
C2, C4	10 picofarads (pF) ceramic
C3	10 nanofarads (nF) ceramic

Additional item

Prototype broad approx. 25 × 25 mm

17 Receiving aerials for amateur radio

Introduction

For any radio receiver to work well, it must have some form of antenna, or aerial. In almost all domestic transistor radios, the aerial is built into the set, either as a ferrite rod (which looks like a rod of dark grey metal) on which are wound coils of wire, or as a chromium-plated telescopic metal rod. Some radios have both forms of aerial, using the ferrite rod aerial for long waves (LW) and medium waves (MW) and the metal rod for very high-frequency (VHF) stations using frequency modulation.

There's broadcast reception . . .

No aerial is perfect, and these two types are *far* from perfect! As in most mass-produced equipment, *they serve their purpose*, which is not critical, and they allow the radio to be carried around easily, because they are not big and bulky. The broadcast stations are very powerful, usually quite close, and the circuits in the radio are quite sensitive, so the need for large aerials disappears.

. . . and there's amateur DX

Here we have an application which, in comparison with the broadcast situation, couldn't be much more different. The stations do not use much power, they may be half the world away, and the requirement for good aerials and very sensitive receivers is paramount. The aerials must be large and they must be as high as possible, away from buildings and trees, which cause reduction in signal strength, and away from man-made sources of interference.

You may not have thought about this but, in general, the larger an aerial becomes, the longer the cable (or 'feeder') must be in order to reach your shack and the receiver inside it. Cables reduce the received signal, so what your aerial gains by being large, the feeder (if you're not careful) will lose!

A simple aerial . . .

We now know that a 'good' aerial is essential. But what is a 'good' aerial? It depends on your purse, your property and your enthusiasm! One of the simplest (and, incidentally, one which is *not* subject to the cable loss problem discussed above) is the **Long Wire**, shown in **Figure 1**. This is, quite literally, a long piece of wire going from a chimney stack to a tree or pole

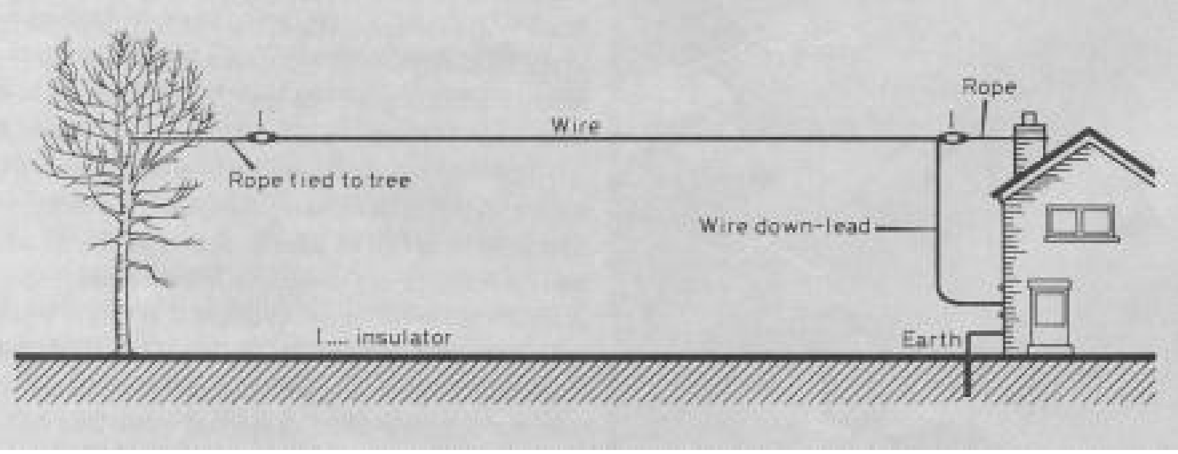

Figure 1 A tree at the bottom of the garden can provide a useful antenna support. Insulators can usually be obtained from Tandy stores or TV aerial suppliers

at the end of the garden. The longer it is, the better. Notice that the wire itself is *not* used to loop around the tree or chimney. Rope is used for both, and is secured to the end of the wire using an insulator of the 'dog bone' or 'egg' variety, to be found in profusion at rallies.

The wire is brought into the house through a window into the shack where the receiver is situated. The long wire aerial is best used with an aerial tuning unit (ATU), which is also described in this book. One of the advantages of the long wire is that it can be used on several frequency bands.

. . . and more complicated ones

Next up the ladder of complexity is the **dipole** (meaning two poles, or two elements). One form of dipole for lower frequencies would take the same basic form as the long wire, except that the feed to the receiver is taken, not from the end, but from the centre. The wire is essentially cut into two halves, and the two ends at the centre are connected to one end of a coaxial cable, which is then taken to the receiver. A smaller form of this is shown in **Figure 2**, which is conveniently mounted in a house loft. Aerials should always be mounted outside for best results, but will work when mounted inside, and the loft space is the logical situation.

The total length of the dipole should ideally be one-half of the wavelength of operation – hence the term 'half-wave dipole'. For example, a dipole for use on 20 m should be about 10 m long. Dipoles are thus 'single-band' aerials. They can be modified for use on several bands, and then become known as 'trapped dipoles', having coils and capacitors at certain points along their lengths.

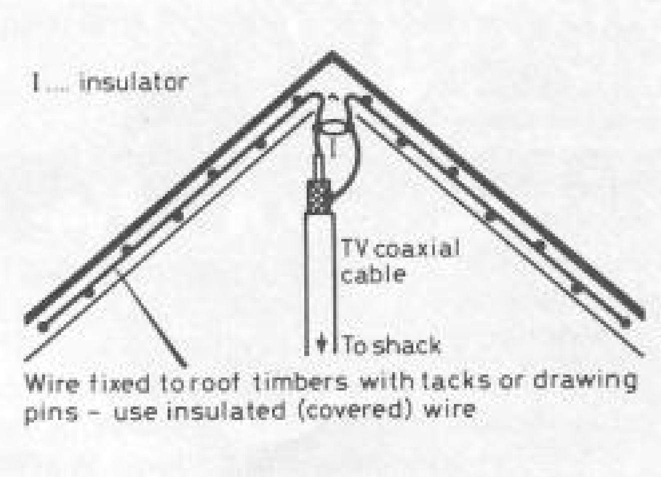
I .... insulator
I
TV coaxial cable
To shack
Wire fixed to roof timbers with tacks or drawing pins – use insulated (covered) wire

Figure 2 An indoor antenna should be mounted as high as possible

Indoor aerials

For flat- or apartment-dwellers, an indoor aerial is often the only solution, in which case the dipole of Figure 2 may be of great use. Other popular types include the **loop aerial**, of which several types are shown in **Figure 3**. The shape of the loop is of secondary importance, but it should be as large as possible.

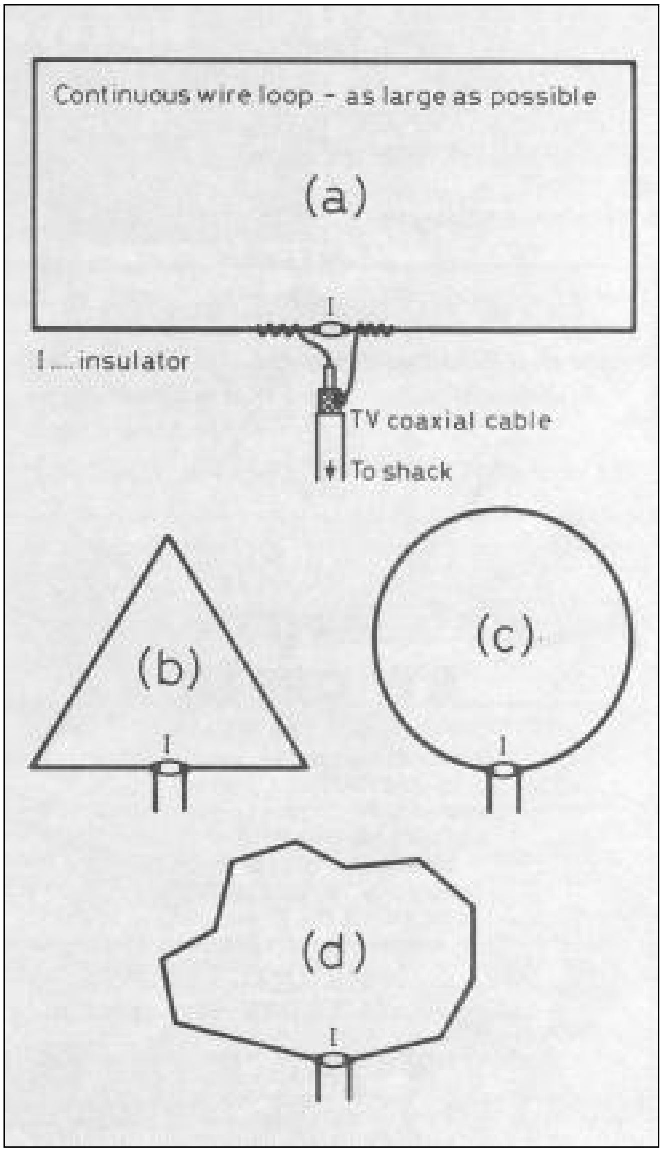
Continuous wire loop - as large as possible
(a)
I
I.... insulator
TV coaxial cable
To shack
(b)
I
(c)
I
(d)
I

Figure 3 Try different shapes of loop antenna to see which works best

Ingenuity is called for when considering indoor aerials. A dipole can be mounted along a picture rail or pelmet. A vertical loop of aluminium foil can be stuck to an inside wall and covered with wallpaper! The designs are limited only by your imagination, and will give you hours of listening pleasure

Outdoor aerials

If you want to venture beyond the simple long-wire discussed earlier, then the erection of outdoor aerials can be a bit tricky; they can be more than a one-man job, and can involve seeking planning permission. Get some help advice from someone who has done it before – what about your local radio club? Plenty of designs for aerials appear in *RadCom*, so you should never be short of ideas!

18 The Colt 80 m receiver – Part 1

Introduction

This is the first of four parts detailing the design and construction of a simple radio capable of good reception of amateur radio signals on the 80 metre band. If you can solder and have some basic hand tools, at the end of Part 4 you will have a working receiver, in which you can take pride, and start some serious listening on 80 m! For testing, all you will need is another receiver and a multimeter.

Description

The radio will be built in three modules, or stages, as illustrated in **Figure 1**. Each of these is built on a printed-circuit board (PCB) or matrix board and can be tested in its own right. The case will need some holes drilled, and care will be needed to mount the PCBs in the case.

Get to it!

Surprisingly enough, when you are building a radio section by section (which is *always* the best way), it is easiest to start at the output and work

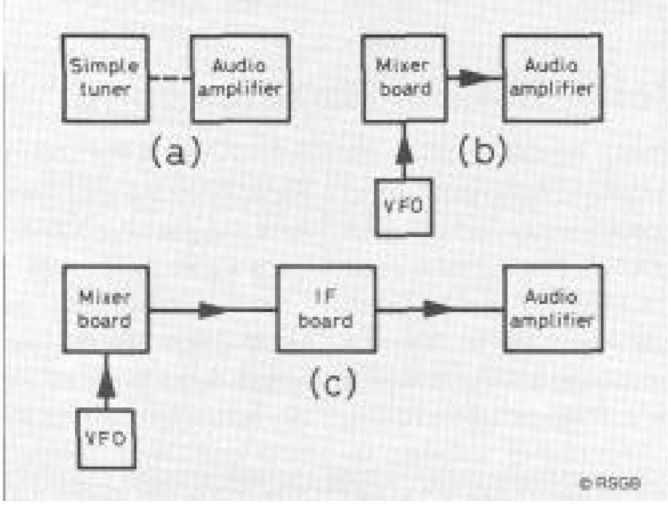
Simple tuner
Audio amplifier
(a)
Mixer board
Audio amplifier
VFO
(b)
Mixer board
IF board
Audio amplifier
VFO
(c)
© RSGB

Figure 1 Build your Colt and watch it grow. A simple crystal set (a) becomes a direct conversion receiver (b) and finally an 80 metre amateur band superhet (c)

towards the input, and you can test what you have done stage by stage. You will see what this means as you progress with your construction.

Every receiver needs some audio frequency (AF) amplification to make the sound signals big enough for you to hear. The circuit for the AF amplifier is shown in **Figure 2**. It uses the TDA7052 integrated circuit (IC) plus a handful of extra components. R1 in conjunction with C2 and C3 *decouple* the battery supply, preventing any audio signals getting through to it and affecting other parts of the radio, when they are connected. C1 acts to prevent high frequencies (above the 3 kHz bandwidth) going into the amplifier input. A volume control, VR1, is connected across the amplifier input, so that the amplifier can accept signal inputs over a wide range. **Figure 3c** shows the connections, which are made with screened cable. The

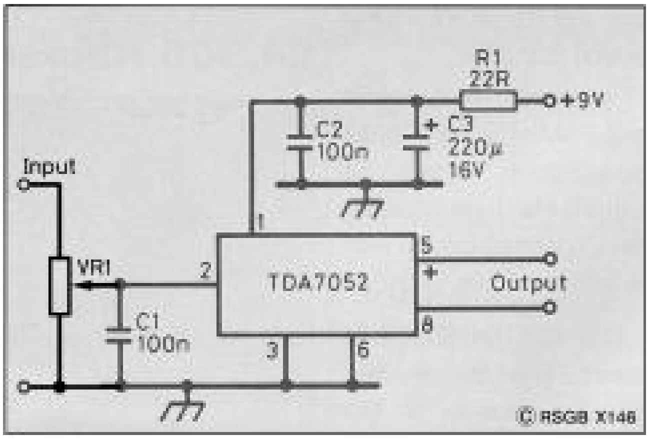
R1
22R
+9V
C2
100n
C3
220µ
16V
Input
VR1
C1
100n
1
2
3
6
5
8
TDA7052
+
Output
© RSGB X146

Figure 2 The Philips TDA7052 integrated circuit (IC) used in the audio amplifier needs very few extra components. It has a signal voltage gain of 100 times and the output is suitable for a loudspeaker or headphones

centre conductor of the left lead goes to the point marked 'input' in Figure 2, the braid being connected to the amplifier earth (0 V) tag. The right lead to VR1 goes to whatever signal source you have for testing – see later. VR1 is *not* mounted on the PCB.

The circuit is constructed on a small PCB or matrix board. Make sure that the electrolytic capacitor, C3, is soldered into the board the correct way – its positive and negative connections are shown in **Figure 3b** for reference. If you are **at all** concerned about soldering the IC into the board, enlist some help, *or*, obtain an 8-pin DIL socket, which you can solder in and then carefully insert the chip into the socket, making sure that is the correct way round. The markings on the chip are shown in **Figure 3a**.

The output leads from the amplifier go to a plastic 6.3 mm (¼ inch) mono jack socket, so that *neither* output lead is connected to the metal case. The amplifier will drive a pair of headphones or a small 8 Ω loudspeaker.

The battery leads must be the right way round also; the battery itself can be a PP3 or PP9, or you can use a small DC power supply.

Testing

First, check that all the components are in the right places, that your soldering is good, and that you have headphones or a loudspeaker connected. Set VR1 about halfway along its travel. Connect the battery.

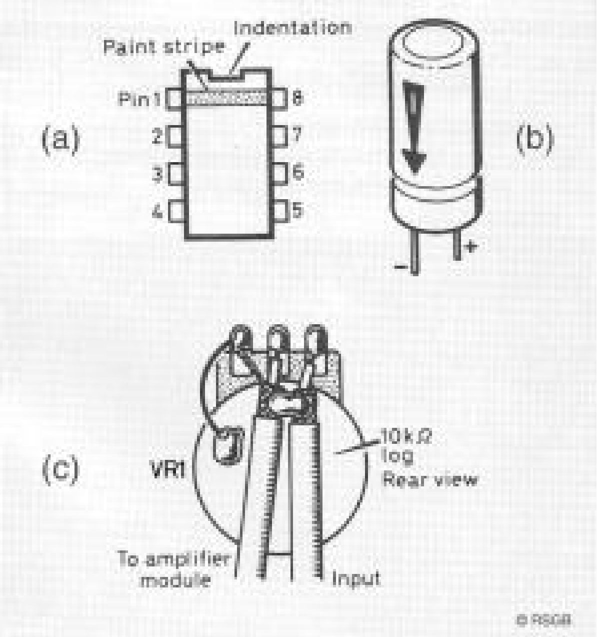

Figure 3 It is important to check the component connections carefully. The diagram shows (a) top view of the IC, (b) electrolytic capacitor and (c) volume control (VR1) connected across the input of the amplifier

A slight hissing noise should be heard; touching the input lead to the amplifier (the centre of the three connections on VR1) should produce a loud buzz. Touching the shaft at the same time will make the buzz quieter. This is the quickest way of confirming that your amplifier *seems* to be working. The only real test is to give it something meaningful to amplify! See the design of our *Crystal Radio Receiver* for full details.

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	22 ohms (Ω)
VR1	10 kilohms (kΩ) log
Capacitors	
C1, C2	0.1 microfarad (μF)
C3	220 microfarads (μF) electrolytic 16 V
Integrated circuit	
IC1	TDA7052 audio amplifier
Additional item	
PCB	(see below)
Component suppliers:	
Maplin	

The next part . . .

The metal case will be marked out ready to receive the completed modules.

19 A crystal radio receiver

Introduction

Here's a quick project to fill in a winter's evening! It was originally designed as a piece of test gear for the *Colt* receiver, but can be successfully used by anyone as a first radio project. Another use for it would be as test gear for any audio amplifier project requiring an audio input for test purposes; this is ideal, as it does not require any power supply!

Details

The initial rough-and-ready test for the audio amplifier of the *Colt* receiver will have whetted your appetite; you will want to prove more conclusively that your amplifier works, and in a way that others in the household will appreciate. Buzzing noises are not convincing in this respect! You are going to put together a simple crystal set – the simplest type of radio that there is – and use it as a signal injector for your amplifier. In this way, you build a real medium-wave (MW) receiver which drives a loudspeaker, as an intermediate product of the construction of an 80 metre amateur band receiver!

The circuit diagram is shown in **Figure 1**. It has a 60-turn coil mounted on a small piece of paper or card wrapped round a ferrite rod. The coil has a connection made to its centre-tap (the middle turn of the coil). The tuning capacitor, VC1, is the most expensive part of the circuit but don't worry, it will be used in the final design of the receiver also! Connect the diode, D1, from the centre-tap to the input to the potentiometer of the amplifier circuit of the *Colt*. Be careful to connect the aerial to the *vanes* of VC1, and not to its *frame*, or you will experience some strange effects when you are tuning. If your amplifier is working correctly, you should be able to receive local stations on medium-waves quite well.

If you think the crystal set will be of use to you in the future as a signal injector, all you will need will be another variable capacitor! If you do not intend to use the amplifier, a small crystal earpiece will allow you to listen. Walkman-type headphones will *not* work!

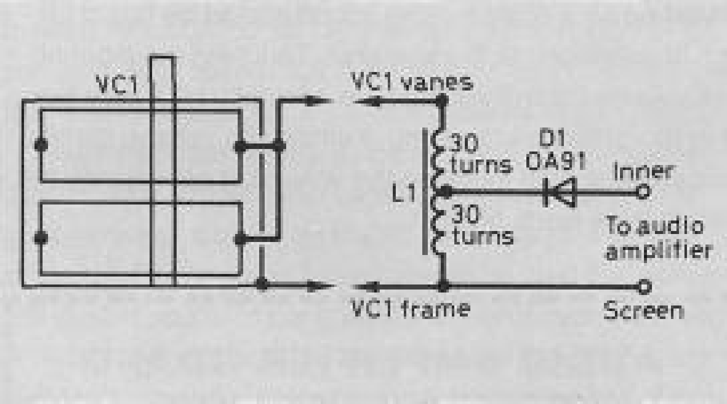
VC1
VC1 vanes
30 turns
D1
OA91
Inner
L1
30 turns
To audio amplifier
VC1 frame
Screen

Figure 1 Both sets of moving vanes are joined as shown. L1 is wound on a ferrite rod with 32 SWG wire and centre-tapped. A single winding (no tap) can be used, joining D1 to the top of the winding

Parts list

Variable capacitor

VC1	125 + 125 picofarads (pF) twin-gang

Ferrite rod

About 100 mm to 140 mm long

Diode

D1	OA91 germanium diode

Additional items

Wire	2 m of enamelled copper wire, between 23 SWG and 32 SWG
Earpiece	High-impedance crystal type (only needed if you are not using the amplifier)

20 The varactor (or varicap) diode

Introduction

Many of the circuits for receivers and transmitters presented in this series rely upon the variable capacitor as a means of tuning. Another method of varying capacitance (without any moving parts) is provided by the *varactor diode*, sometimes called a varicap diode. This is a component which changes its capacitance as the voltage across it is varied.

The details

Figure 1 shows how a varactor diode might be connected to demonstrate its operation. Its symbol is that of an ordinary diode, with a capacitor symbol next to it. A variable voltage is applied across it in such a way that the diode is *reverse-biased*. This means that virtually no current passes through it – the positive voltage is applied to the cathode. Varactors are cheaper than variable capacitors, and they are tiny in comparison, very suitable for today's miniature circuits. If A and B were connected across the tuning coil in a simple receiver (with a series capacitor to block the DC from the battery reaching the coil), the tuning operation would be accomplished by turning the knob on the 10 kilohm potentiometer.

Varactors are available with different values, from less than 20 picofarad (pF) for VHF applications to 500 pF for medium-wave radios. They are

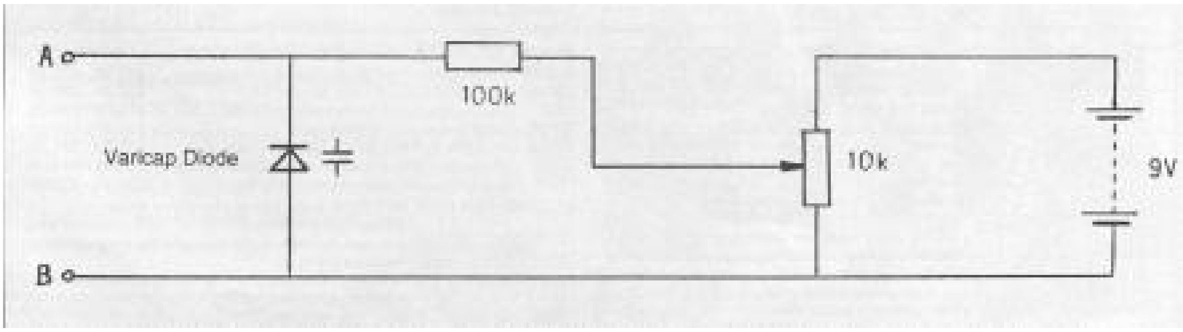

Figure 1 The capacitance of the varicap diode (between A and B) increases as the voltage is reduced, using the variable resistor

tuned usually by voltages between 2 V and 9 V. For a real application of varactors, you should consult the circuit diagram of the *Yearling* 20 metre receiver, elsewhere in this book.

In some circuit designs, several circuits are all tuned to the same frequency in order to improve the overall *selectivity* (the ability of the circuit to reject signals very close in frequency to the wanted signal). Special dual- and triple-varactors are available for circuits like this. Having been made at the same time from the same materials makes their individual characteristics virtually identical. Like all other diodes, they must be correctly wired into the circuit – their *polarity* is important.

Changes in temperature will cause the capacitance to change which, if it were part of an oscillator circuit, would cause the oscillator to drift – you would have to keep retuning the radio! This can be corrected by using a special integrated circuit called a *phase-locked loop* (PLL). Modern TV sets and satellite receivers use varactors and PLLs in this way.

Some useful varactor types

Type No.	*Tuning range*		*Description*
	pF/V	*pF/V*	
BB204B	42/2.0	15/12	Dual VHF
BB212	560/0.5	22/8	AM tuning
KV1235	450/2.0	30/8.5	Triple AM
KV1236	450/2.0	30/8.5	Dual AM
MV1404	120/2.0	9/10.0	HF tuning

21 A portable radio for medium waves

Introduction

The ZN415E integrated circuit (IC) can be used to make a very efficient AM portable MW broadcast radio with a built-in loudspeaker. Here's how!

The circuit

Figure 1 shows the circuit diagram of the portable radio. It's not as complicated as it may appear, especially after you have got started. L1 is a coil of wire mounted on a ferrite rod, acting as an aerial; VC1 is a variable capacitor which works, with L1, to tune in different stations. IC1 contains circuits of its own which boost the selected signal and it includes a *detector* which extracts the audio signal from the incoming RF signal. Earphones could be connected to the output of IC1 (between pins 4 and 5), but the output would not be powerful enough to drive a loudspeaker.

More sound

This is where IC2, an LM386 comes in. This is a small audio *power* amplifier which produces audio signals with enough power to drive a small

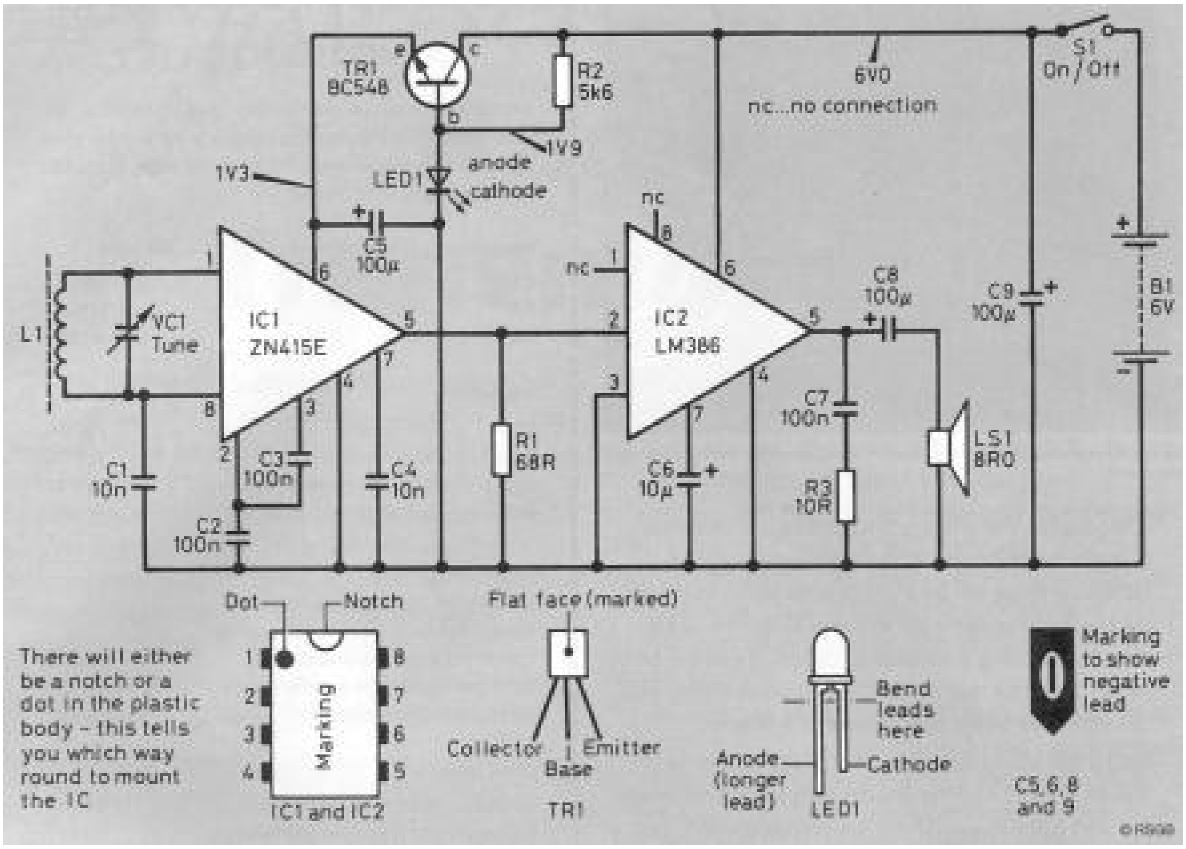
TR1
BC548
e
c
b
R2
5k6
1V9
6V0
nc...no connection
S1
On / Off
1V3
LED1
anode
cathode
C5
100µ
nc
IC1
ZN415E
IC2
LM386
L1
VC1
Tune
C8
100µ
C9
100µ
B1
6V
C7
100n
LS1
8R0
C1
10n
C3
100n
C4
10n
R1
68R
C6
10µ
R3
10R
C2
100n
Dot
Notch
Flat face (marked)
There will either be a notch or a dot in the plastic body – this tells you which way round to mount the IC
Marking
IC1 and IC2
Collector
Base
Emitter
TR1
Bend leads here
Anode (longer lead)
Cathode
LED1
Marking to show negative lead
C5,6,8 and 9
© RSGB

Figure 1 The circuit diagram of our easy-to-build portable radio. Take care to mount the ICs and LED the correct way round

loudspeaker, LS1. The radio uses a 6 volt battery, which is made by connecting four 1.5 volt AA cells in series (4 × 1.5 V = 6 V) using a battery holder designed for this purpose. Although a 6 V supply is ideal for IC2, it is far too great for IC1, which needs only about 1.3 V. This lower voltage is provided from the 6 V supply by TR1 (an npn transistor), R2 and LED1 (a light-emitting diode). When current passes through an LED (see the description of the LED in this series) a reasonably constant voltage of 1.9 V appears between the anode and the cathode. Because of the voltage (0.6 V) that *always* exists between the base and emitter of a working transistor, the voltage on the emitter is about 1.9 V – 0.6 V = 1.3 V, and this is used as the power supply for IC1.

To keep the radio as simple as possible, no volume control has been fitted. Instead, you can use the directional properties of the ferrite rod aerial (see the information on ferrites in this book) to reduce the volume by rotating the set about a vertical axis using the handle provided.

Putting it all together

1. Start by covering the ferrite rod with Sellotape, or alternatively wrap a piece of paper tightly around it, and secure it with Sellotape. Then, with at least 2 metres of 24 SWG enamelled copper wire, wind 75 turns tightly around the rod. To be safe, leave about 50 mm of wire at the ends of the coil, then wrap the whole coil with Sellotape to hold the turns in place, leaving only the ends free. Then, using a small piece of sandpaper, remove the enamel from the last centimetre of each end of the coil.
2. Most of the components are mounted on a piece of Veroboard (the type with parallel copper strips on one side). The piece used on the prototype measured 32 holes by 10 strips, as **Figure 2** shows. Before you start fitting components, cut the copper strips as shown. It is easier to do it now than when the board is littered with components! The strips may be cut with a 3 mm (⅛ inch) twist drill rotated between thumb and forefinger. Resist the temptation to use a hand drill – the idea is just to cut the copper, **not** to drill right through the board!
3. Solder the IC sockets and the other components on the board as shown in Figure 2. Make sure that the IC sockets are fitted with their notches towards the *top* of the board, as viewed in Figure 2. Do not insert the chips yet. Always keep the wires left over from cropping resistors and capacitors, they will come in handy at times like this: make the wire links that are clearly shown in Figure 2. Connect the electrolytic capacitors (C5, C6, C8 and C9), the transistor and the LED the correct way round; then check it again when you have done it!
4. Finally, solder lengths of stranded insulated wire to act as ‘flying leads’ for future connection to L1, VC1, LS1, S1 and the battery connector.
5. Apart from the battery holder, everything is mounted on the case *lid*. This makes assembly and testing much easier, and eases fault-finding if

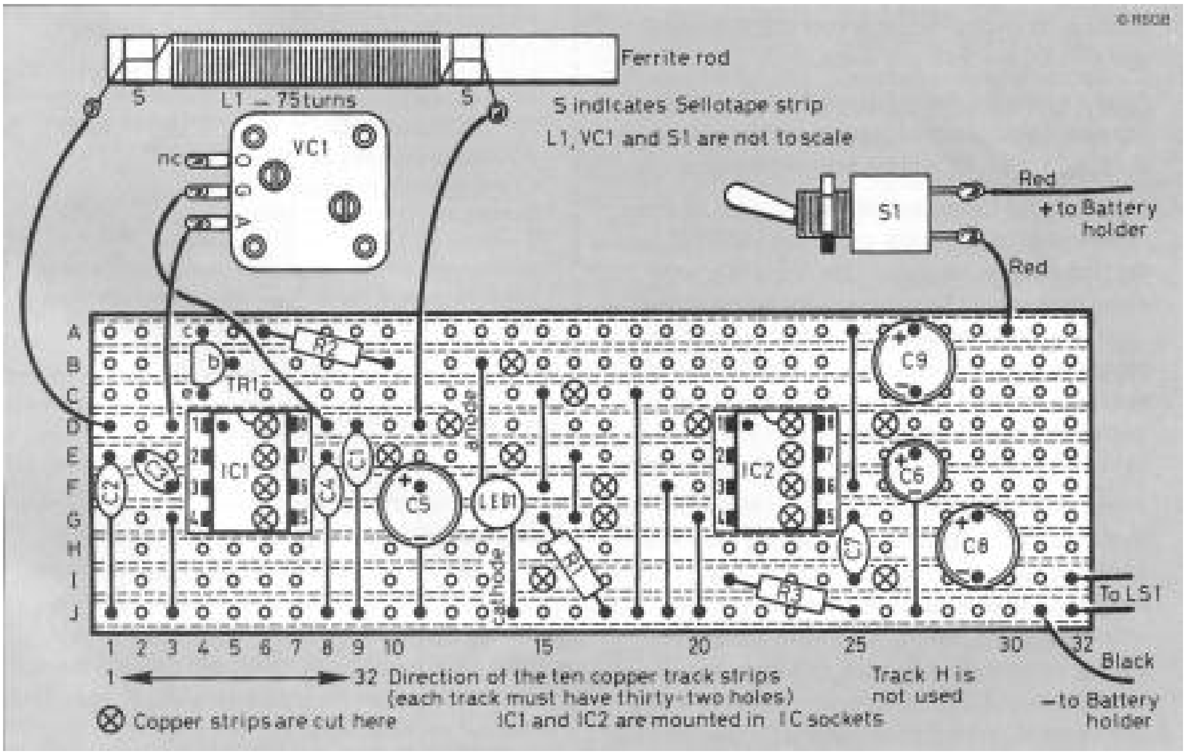
© R508
Ferrite rod
S
L1 — 75turns
S
S indicates Sellotape strip
L1, VC1 and S1 are not to scale
VC1
nc
O
G
A
S1
Red
+to Battery holder
Red
A
B
C
D
E
F
G
H
I
J
c
b
e
TR1
R2
IC1
IC2
C1
C2
C3
C4
C5
C6
C7
C8
C9
LED1
anode
cathode
R1
R3
To LS1
1 2 3 4 5 6 7 8 9 10 15 20 25 30 32
Black
1 ◄— —► 32 Direction of the ten copper track strips (each track must have thirty-two holes)
Track H is not used
–to Battery holder
⊗ Copper strips are cut here
IC1 and IC2 are mounted in IC sockets

Figure 2 Veroboard layout for the Portable Radio. Make sure that all the wire links are included

the need arises! At the speaker position, make one large hole or a series of small holes to let the sound out. The ferrite rod may be stuck to the lid, as may the loudspeaker. Drill holes of the correct size to fit the particular types of variable capacitor (VC1) and switch (S1) that you are using. The Veroboard may be held in position by Blu-Tack or double-sided sticky tape.

6. Before inserting IC1 and IC2, connect the battery and switch S1 on. The LED should glow dimly (you may have to shield it with your hand in order to see it). If you have a test meter, check that there is about 1.3 V between pins 6 and 4 of IC1. If the reading is around 6 V or there is no glow, you may have connected the LED the wrong way round! When everything seems normal, switch off and disconnect the battery. Insert IC1 and IC2, making sure that the pins are straight and lie immediately above their corresponding holes in the sockets, and that the notches line up with the notches in the holders. Then push *gently* downwards on each IC in turn until the chip is firmly seated in its socket.
7. Switch on! By rotating the tuning capacitor, VC1, you should now be able to tune in many stations, rotating the radio to give you some volume control.

Final touches . . .

The handle was made with part of an old leather belt, secured to the case with ‘number plate’ nuts and bolts from Halfords. The loudspeaker grille is the lid from a *pot-pourri* container, and some extra holes were drilled in the back to improve the sound. See what you can find to finish off your radio!

Parts list

Maplin order codes are given for most of the parts, but you should get used to using the ‘beg, borrow or steal’ technique, or to use your ever-expanding junk box.

		Maplin code
Resistors: all 0.25 watt, 5% tolerance		
R1	68 ohms (Ω)	M68R
R2	5.6 kilohms (kΩ)	M5K6
R3	10 ohms (Ω)	M10R
Capacitors		
C1, C4	10 nanofarads (nF) or 0.01 microfarad (μF) ceramic	BX00A
C2, C3, C7	100 nanofarads (nF) or 0.1 microfarad (μF) ceramic	YR75S
C5, C8, C9	100 microfarads (μF) electrolytic, at least 10 V	FF10L
C6	10 microfarads (μF) electrolytic, at least 25 V	FF04E
Semiconductors		
IC1	ZN415E radio chip	
IC2	LM386 audio power amplifier	UJ37S
LED1	3 mm green LED	WL33L
TR1	BC548 npn transistor	QB73Q
Additional items		
LS1	Miniature 8 ohm loudspeaker	WB08J
S1	Miniature SPST toggle switch	FH97F
Ferrite rod	Length approx. 100 mm	YG20W
	24 SWG enamelled copper wire	BL28F

Additional items (*continued*)

VC1	Tuning capacitor 140 to 300 picofarads (pF)	FT78K
	Tuning knob	FK41U
	8-pin DIL IC sockets (two required)	BL17T
	4 × AA-size battery holder (long)	HF94C
	PP3-type clip for battery holder	HF28F
	Plastic box approx. 158 × 95 × 54 mm	LH51F
	0.1 inch Veroboard, min. size 32 holes × 10 strips	JP46A

Plus
- Stranded insulated conductor for flying leads
- Multicore solder
- Materials for handle and speaker grille
- Double-sided sticky tape or Blu-Tack
- Sellotape
- Glue
- Four AA-size 1.5 V batteries

22 The Colt 80 m receiver – Part 2

Introduction

In Part 1 we constructed the audio amplifier module for the system and tested it in a very simple way. If you did what was suggested and built the simple Crystal Set to use as a signal source, you will know just how well the amplifier works.

The case

Metal cases for the project are available from Maplin, telephone 01702 554 161 (code XB67). From the photograph on p. 72 you can see the way the components are mounted. The audio amplifier is seen at the top right of the base, to the right of the tuning capacitor VC1. The next in this series will deal with the variable-frequency oscillator (VFO) and VC1. The current part deals with preparing the case to receive the components.

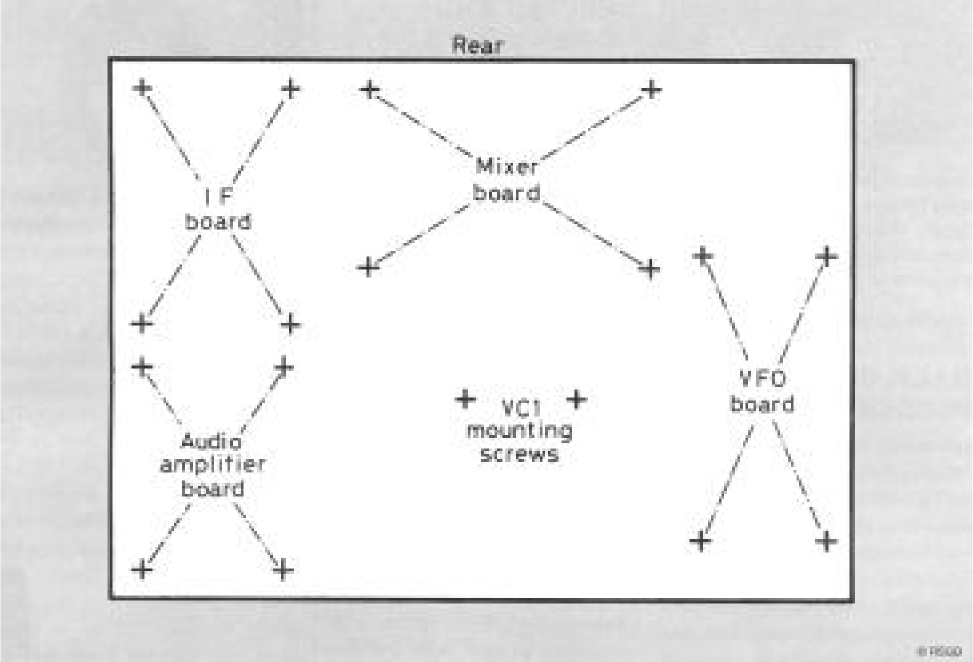
Rear
I F board
Mixer board
VFO board
Audio amplifier board
VC1 mounting screws
© RSGB

Figure 1 Fixing holes for each module are best measured from each printed circuit board or matrix board

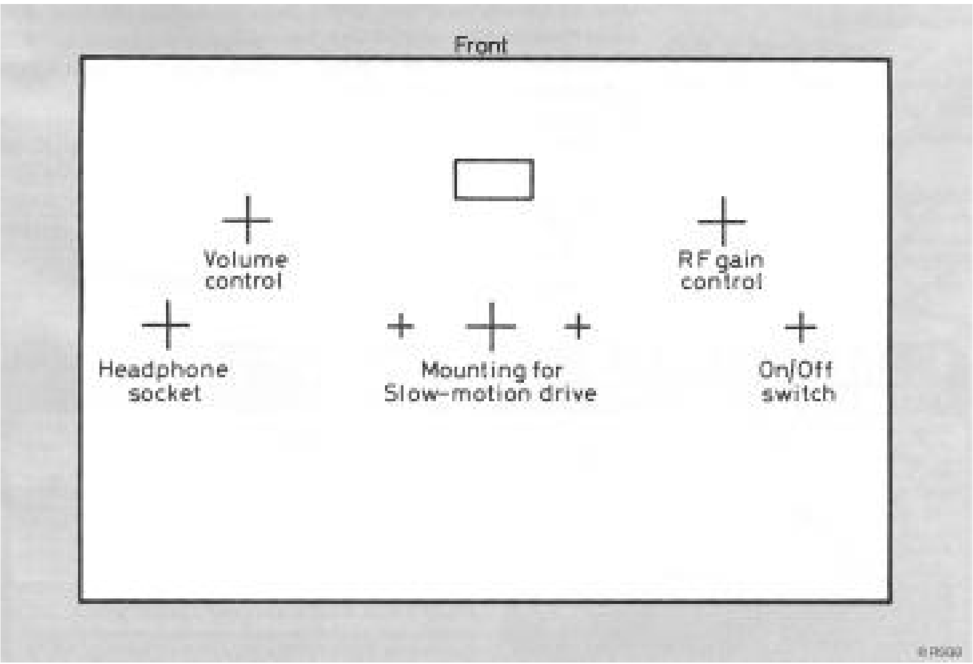
Front
Volume control
RF gain control
Headphone socket
Mounting for Slow-motion drive
On/Off switch
© RSGB

Figure 2 Position the slow motion drive to allow viewing of the tuning dial

Figures 1 and **2** show the markings for preparing the front panel and base. All the circuit boards and the tuning capacitor are mounted on the base using 10 mm stand-off pillars with 6BA bolts. The board locations are shown in Figure 1. The front panel control positions are shown in Figure 2, together with the small rectangular hole for viewing the tuning dial.

The best way to mark out the holes for the boards is to lie the boards flat on the base (before you've started soldering the components in) and marking the base through the holes in the boards. This minimises the scope for errors!

A *reduction drive* is used between the tuning knob and the capacitor shaft. This is simply a gear mechanism that slows down the capacitor shaft by a factor of six compared with the tuning knob, and makes tuning very much easier. The recommended variable capacitor also has a pulley wheel mounted on the shaft. Glued to this wheel will be a scale marked with frequency and is visible through the rectangular hole in the front panel.

The next part . . .

The variable-frequency oscillator and mixer will be added to the project.

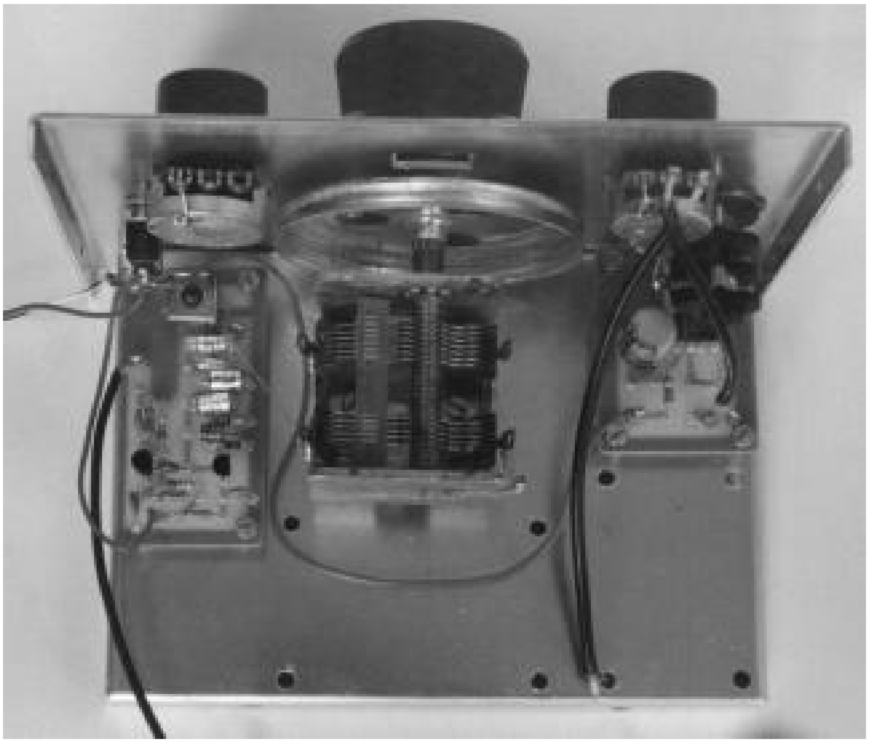

23 A simple transistor tester

Introduction

Although transistors aren't used as much as they were before integrated circuits came along, a transistor tester is still a useful piece of test equipment to have around the shack. This design is about the simplest possible and will produce an indication of whether a transistor is giving any current gain; this does not necessarily mean that the transistor is perfect but that it is working to a certain extent. This tester will **not** test field-effect transistors (FETs). If you buy a bag of transistors at a rally, this tester is useful for giving a yes/no indication of which ones go straight in the bin and which are kept for further use.

How it works

Figure 1 shows the simple wiring circuit. In order to explain the working of the circuit, a circuit diagram is shown in **Figure 2**, with the npn transistor, TR1, under test shown as part of the circuit. The pnp/npn selector switch, SW1, is omitted for clarity.

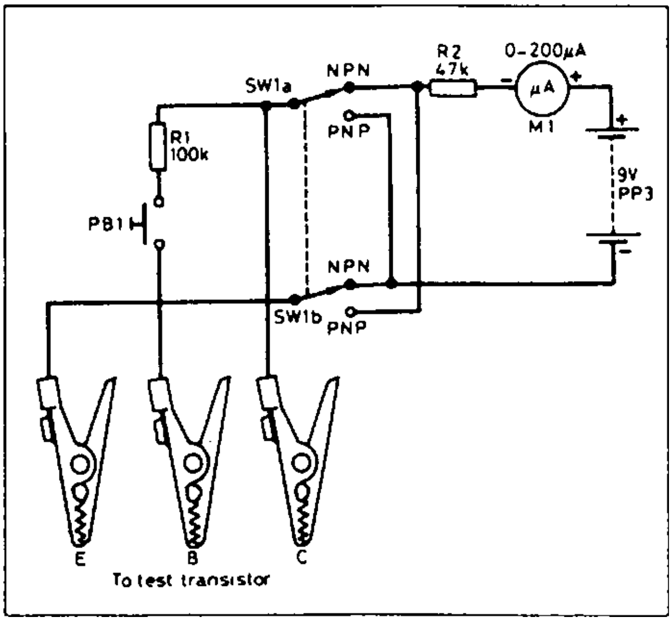

Figure 1 Circuit diagram of the transister tester

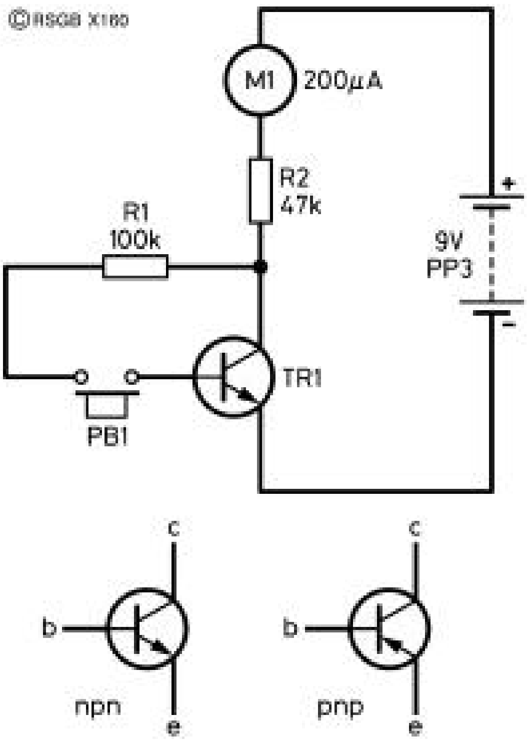

Figure 2 Testing an npn transistor

Any current that flows through the transistor, TR1, must flow from the battery, through the meter, M1, and through the protective resistor, R2. R2 prevents excessive current flowing through the meter and damaging it. Even if there is a short-circuit between emitter and collector, the maximum current that will flow is given by the simple equation

$$I = \frac{V}{R},$$

where I is the current flowing in amps,
V is the battery voltage, and
R is the total circuit resistance in ohms.

Putting in the correct values, gives

$$I = \frac{9}{47\ 000} = 0.000191 \text{ A, or 191 microamps } (\mu\text{A}).$$

The resistance of the meter, M1, will cut this down a little more, but it is within the indicating range (200 μA) of the meter.

As shown, with the push-button switch, PB1, *open*, a good transistor will not draw any current from the battery, and M1 will thus remain at zero. When PB1 is *pressed*, a *very small* current is injected into the base of the transistor. If the transistor is working, it will produce a much larger current between the collector and emitter, and this current will also flow through M1 and R2, giving a significant reading on M1, showing that all appears to be well. If an appreciable current flows when PB1 is *open*, then your transistor is suspect.

Don't be put off by the apparently complicated switch, SW1. It is there to allow the other type of transistor, the pnp type, to be tested. All it does is reverse the battery connections, so that the emitter goes to the *negative* battery terminal for testing an npn transistor, and to the positive terminal for a pnp type!

Most transistors in common use are of the npn type, which is why Figure 2 shows the testing of an npn type. The connections to the different transistor *encapsulations* (shapes) are given in any good component catalogue. Avoid the trial-and-error method to discover the connections to a transistor. This is unscientific, and can be very frustrating, particularly if the transistor is faulty in the first place!

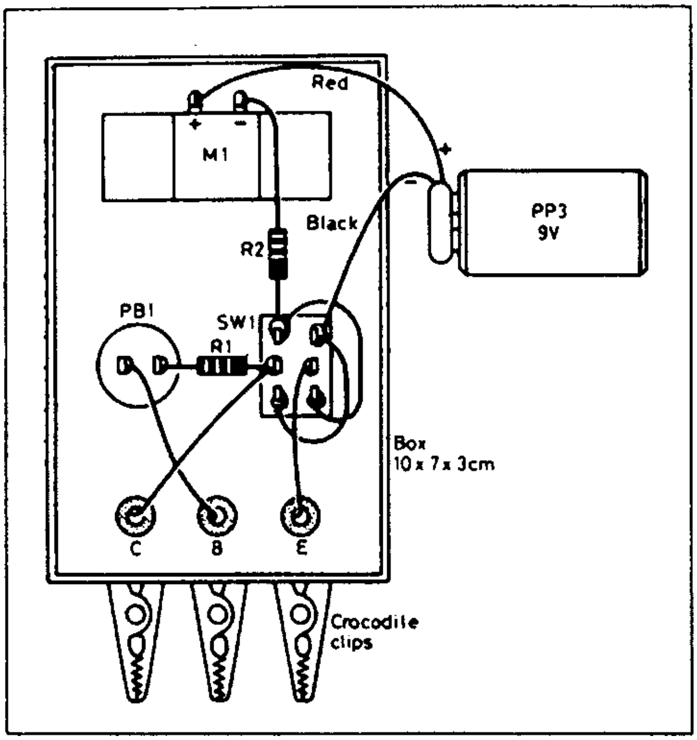

Figure 3 This shows how to wire up the tester

Construction

The unit can be built into a small plastic box, the components being soldered directly to the fixed terminals of the meter and the two switches; no circuit board is necessary! Use different colours of wire for the three test leads, and make sure you know which is which! Check the connections against the wiring diagram of **Figure 3**. When you are confident that all is correct, connect the battery and make sure there is no reading on the meter when nothing is connected to the crocodile clips!

Find any transistor for which you know the connections and the type (npn or pnp). Set the npn/pnp switch accordingly. Connect the three clips, making sure that they do not touch each other. A small reading on the meter at this stage means the transistor is suspect; a large reading means it is not working and should be thrown away! Press PB1 and watch the meter; a reading greater than half of full-scale indicates a good transistor. If it is less than half, you may have a transistor with 'low gain'; it may be usable for non-critical applications, but if you are in any doubt – throw it away!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	100 kilohms (kΩ)
R2	47 kilohms (kΩ)

Additional items

PB1	Push-button switch
SW1	DPDT (double-pole double-throw) switch
M1	Micro-ammeter – *not* less than 200 µA full-scale deflection
	Crocodile clips (3 needed)
	PP3 battery and connector
	Plastic box about 10 × 7 × 3 cm

Common types of transistor

BC108, BC109, 2N2369A
These are small-signal npn types, used in audio amplifiers. They are in metal cases (called TO18) and have a tab next to the emitter lead. Common types have a B or C suffix (e.g. BC109C). The C suffix indicates a higher current gain than those with a B suffix. The 2N2369A is specially designed for radio use at high frequencies.

2N3703, BC212L, BCY71
These are pnp transistors, and so must be used with the collector and base *negative* with respect to the emitter. These three types are used in small amplifiers and audio oscillators. The first two have plastic encapsulations (TO92), while the BCY71 has a metal case (TO18).

BFY50, BFY51, BFY52
For slightly higher powers, these are ideal. They have been used in novice transmitters up to 600 milliwatts (mW). The TO5 case is a scaled-up version of the TO18 case. They are all npn types.

2N3055, 2N3773, TIP35C
These are high-power transistors in bigger encapsulations. The thick metal TO3 case of the 2N3055 and 2N3773 is designed to bolt to a heat sink, a large piece of metal which conducts the heat into the air more rapidly than the transistor itself can. The TIP35C is made of plastic but has a thick metal tab by which it, too, can be bolted to a heat sink.

24 An introduction to transmitters

Introduction

We usually think of a transmitter as being a 'black box'. However, that is the form a transmitter takes for our use on the amateur bands. Many electrical circuits are transmitters, even though transmitting may not be their primary function!

Anything that emits electromagnetic energy at *any* frequency is a transmitter, from radio at the low-frequency end of the spectrum, to gamma rays at the high-frequency end. We all know that a magnet will attract certain metals and that a comb rubbed on your coat sleeve will pick up small pieces of paper. The former is an example of the effect of a magnetic field, the latter of an electric field. Electromagnetic fields are combinations of both types of field, and are produced whenever an electromagnetic wave is transmitted.

What frequency?

Many everyday objects have a natural frequency of oscillation. This is called their *resonant frequency*. A wine glass will ring when struck gently; an empty wine bottle will sound if you blow across the top; a guitar string will vibrate when plucked. These are all examples of resonance, and the resonant frequencies will **not** change unless the objects themselves are changed physically in some way. These are resonances in *sound*; we are primarily interested in electrical resonances.

The basic electrical resonant circuit is the combination of an inductor (coil) and a capacitor, as shown in **Figure 1**.

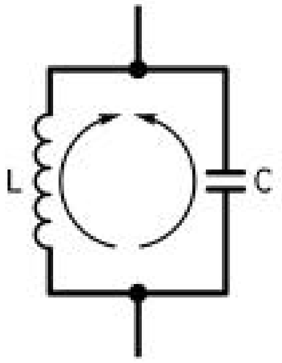

©RSGB X162

Figure 1 The basic electrical resonator. The energy in the circuit alternates between the inductor and the capacitor

A pulse of energy applied to this *tuned circuit* will make it *ring* (oscillate) at its resonant frequency. The energy in the circuit transfers between the inductor and the capacitor every cycle of the oscillation. Just like the wine glass, its oscillation dies away because it is losing some energy to its surroundings – it is transmitting! The frequency of the resonance depends on the values of L and C.

Keeping it going

If we want to keep the circuit oscillating, rather than having it die away, we must supply the circuit with just enough energy to replace the energy lost both by radiation and by losses in the circuit itself. Because of this, you will find in all oscillator circuits, a transistor, valve or FET working with the tuned circuit to provide this extra energy.

As it stands, of course, even with its transistor, our oscillator will not radiate very far. Connecting an aerial to it, and a Morse key to interrupt the power supply, it would become a very low-power CW transmitter. Add a couple more transistors to form a radio-frequency (RF) amplifier, and you have the basis of a simple low-power (QRP) transmitter.

Resonant circuits can also be made using quartz crystals; these work at the crystal frequency only, and this is marked on the crystal case.

A tiny spark transmitter

This is a simple piece of test gear that will increase your knowledge and understanding of resonance. You can use it to estimate the resonant frequency of most of the inductor/capacitor (LC) tuned circuits that you build. The circuit is shown in **Figure 2**. It operates around a relay. Any relay that operates from a 6 V to 9 V source and has contacts which are normally closed (i.e. closed when the battery is not connected). Fit the relay, a toggle switch and the battery in a metal box, and connected up as shown in the diagram. Some foam rubber inside the box may help to reduce the escaping noise of the relay. A small hole in the side of the box enables the 2-turn loop to emerge. This should be about 40 mm diameter, made with insulated wire. Switch on; there should be a loud buzzing noise from the relay. If not, you have probably chosen the wrong contacts on the relay!

When it is working, bring the loop close to the aerial of a radio – it should produce a loud noise from the speaker!

How it works

When you switch on, current flows through the relay contacts and through the relay coil. The relay operates and opens the contacts, causing the relay to 'drop out'. When it does, the circuit is completed again and the contacts are opened, and the cycle repeats. Each time the relay contacts open, there is a small spark between them, causing very rapid current surges through the wire loop. This makes the loop transmit RF energy, very briefly. In the early days of radio, this type of circuit was known as a *spark transmitter.*

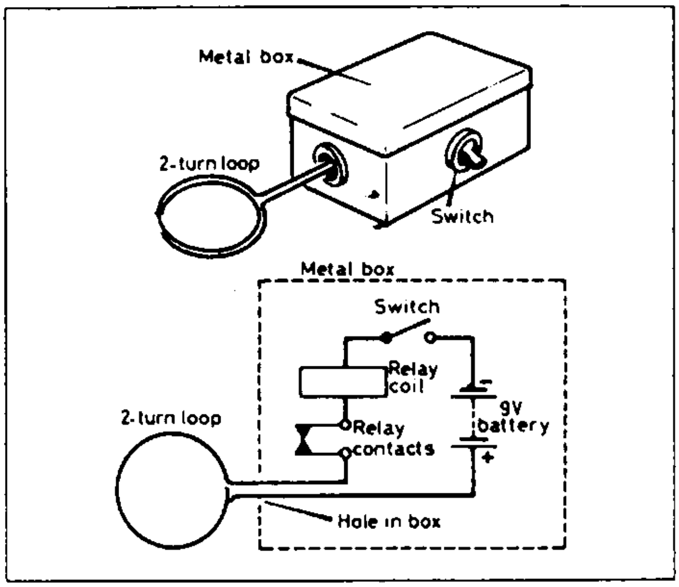

Figure 2 The current path is interrupted when relay is energised as shown above

Make a tuned circuit

Use a discarded toilet-roll centre, and wind about 10 turns of enamelled copper wire round it, keeping each turn close to the next. Scrape off the enamel for about 1 cm at each end, and solder a 100 picofarad (pF) capacitor (or a variable capacitor of about the same value) between the ends. The resonant frequency should be about 10 MHz. If you have used fewer turns or a smaller capacitor, the frequency will be higher.

Measuring the resonant frequency

Set up the buzzer as shown in **Figure 3**, with the loop around one end of your coil. Then make a similar loop, solder it to the end of a piece of coaxial cable going to the aerial socket of a calibrated receiver. Set the buzzer going, tune the receiver around 10 MHz, and search for the maximum noise level from the speaker. When you have found it, move the two loops as far away as possible from the main coil. This is called reducing the *coupling* between the coils, and it may result in a slightly different, but more accurate, resonant frequency.

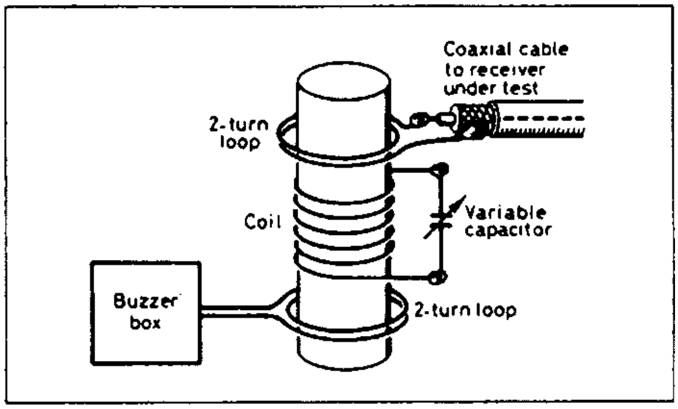

Figure 3 Experiment with the spacing between the loop and the tuned circuit

Parts list

Any small relay which operates between 6 V and 9 V
Metal box – do not use a plastic box!
9 volt battery and connector
On/off (SPST) toggle switch
Plastic foam, as required

25 The Colt 80 m receiver – Part 3

Introduction

In Part 2 we marked out the case ready for installation of the modules when they are completed. In this part, the building of the variable-frequency oscillator (VFO) and mixer will be described. This will produce a type of receiver known as *direct-conversion*, because it converts the radio-frequency (RF) signal directly into an audio-frequency (AF) signal which we can hear in a loudspeaker after amplification. A block diagram of the system is shown in **Figure 1**. The modifications needed to make a full superheterodyne receiver will be left until later.

The direct-conversion receiver covers the 80 metre amateur band and will receive both Morse (CW) and speech (SSB) signals. The audio amplifier was covered in Part 1, so your *Colt* is rapidly taking shape! By the time your construction has reached the end of this part, you will have a receiver ready to use, even if the project is not yet complete!

The direct conversion process

Like most things in radio, the principles of direct conversion are not difficult. From the aerial, the signal we want to hear is selected by the *tuned filter*, which rejects the signals we don't want. The signal then enters the

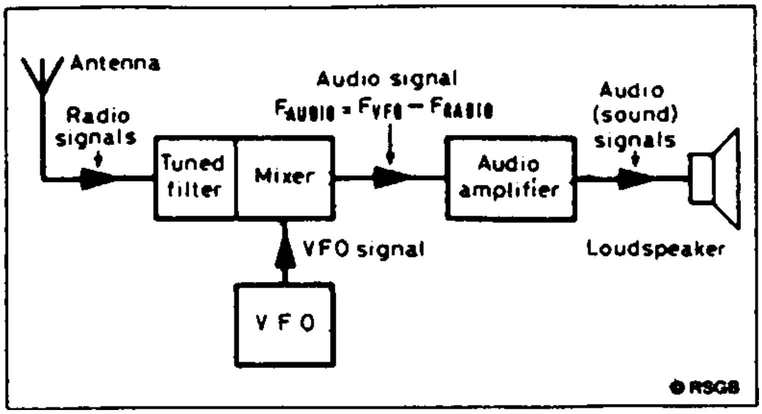
Antenna
Radio signals
Tuned filter
Mixer
Audio signal
$F_{AUDIO} = F_{VFO} - F_{RADIO}$
Audio amplifier
Audio (sound) signals
Loudspeaker
VFO signal
VFO
© RSGB

Figure 1 Stages of direct-conversion receiver

mixer, along with the signal from the *VFO*. The VFO produces a sine wave whose frequency can be varied across the whole of the 80 metre amateur band (3.5–3.8 MHz), by turning the knob on the tuning capacitor, VC1. The mixer produces, at its output, *two* signals; one signal is at the frequency of the **sum** of the signal and VFO frequencies, the other is at the **difference** of the two frequencies. It is the latter that we want. Let's look at the numbers involved. If the signal is at 3650 kHz and the VFO is at 3651 kHz, then the sum frequency is 7301 kHz, and the difference frequency is 1 kHz. If we feed the output of the mixer into our audio amplifier, the 7301 kHz signal is automatically removed (it is far too high to be considered an audio signal!) and the resulting 1 kHz signal is amplified and fed to the speaker, producing a note which we can hear!

Building the VFO

Figure 2 shows the circuit of the VFO. It is a tried and tested circuit, and should work first time. It uses a *field-effect transistor* (FET) for TR1, the oscillator itself. RFC is a *radio-frequency choke*, a coil of wire which will pass a direct current (DC) but which will prevent radio-frequency (RF) signals getting through.

All VFOs have a tuned circuit which, in this case, is formed by the coil L1 and the capacitors C1 and VC1. The frequency will also be affected to some extent by C2, C3, C4 and C5. Transistor TR2 is an *emitter follower*, a stage which gives no voltage gain but provides a good *buffer* stage, isolating the VFO from the effects of the stages that follow it. When building a VFO, the

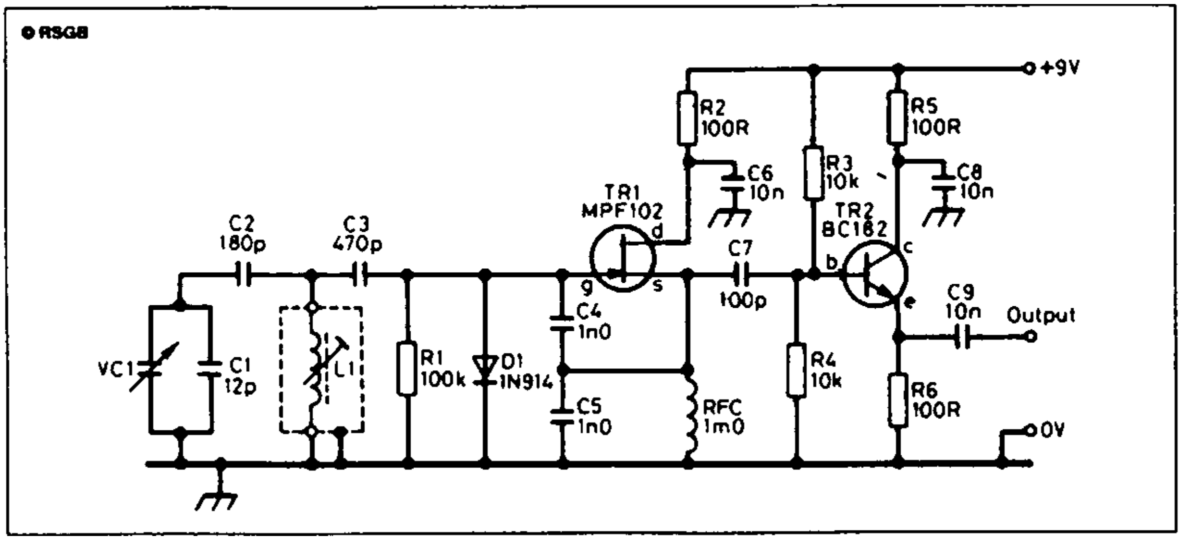

Figure 2 The variable-frequency oscillator uses a field-effect transistor (FET)

parts must be securely mounted. If components move, so does the frequency! At worst, the oscillation will become unstable and the VFO will be useless. Keep the component leads as short as possible – this improves their mechanical stability as well as their electrical stability!

Mount the parts on the printed-circuit board (PCB) or matrix board and, when completed, the VFO should look like the one shown on the left in the photograph on p. 72. Make sure that TR1, TR2 and D1 are the right way round.

On completion, check the component positions then mount it in the case as shown in the photograph, to the left of the tuning capacitor when viewed from the rear. The VFO coil, L1, will need some adjustment, but that will have to wait until the mixer is built. Connections to the other boards are made with screened cable.

The mixer board

So far, we have an audio-frequency (AF) amplifier and a VFO; the addition of a mixer board gives us a complete direct-conversion receiver for 80 m. The mixer circuit diagram is shown in **Figure 3**. Let's follow the signal path.

- From the aerial, the RF signal goes to the gain control potentiometer, RV1. This reduces very strong signals, to prevent them overloading the mixer.
- To select the required band of signals, a *bandpass* filter is made up of RF transformers T1 and T2; these are tuned by C1 and C3, and are coupled together by C2.

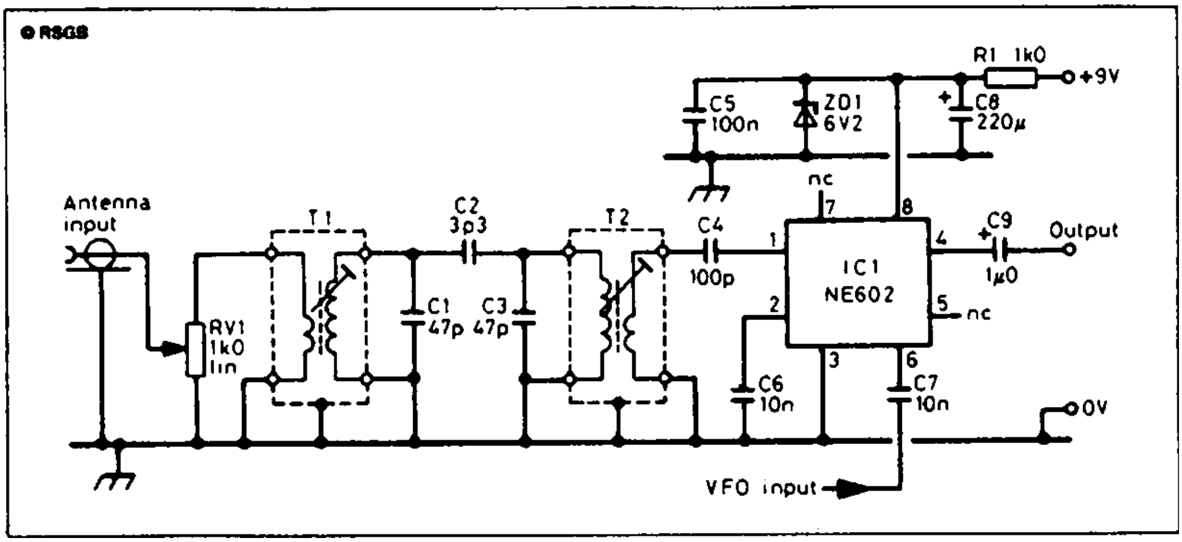

Figure 3 The mixer board has a bandpass filter and stabilised supply

- After the filter, the signal is coupled into the integrated circuit (IC) mixer type NE602, by the capacitor C4. Capacitors C5 and C8 *decouple* the supply line, to prevent unwanted signals on the supply from disturbing the VFO operation. The use of the term *decouple* is exactly the opposite of *couple*; when two circuits are *coupled* together, the signal passes from one to the other; when two circuits are *decoupled*, signals **cannot** pass from one to the other.
- The NE602 works with a 6 V supply; it is produced here from the 9 V supply by the *Zener diode* ZD1 and the resistor R1. ZD1 operates at 6.2 V, and gives a steady output for the mixer. The audio output from the mixer appears at pin 4 or IC1. This is taken to the audio amplifier board via C9 and the volume control (see Part 1).

Care must still be taken to insert some components the right way round. These are the electrolytic capacitors, C8 and C9, the Zener diode, ZD1, and the integrated circuit, IC1. Check all component positions and make sure all your soldered joints are bright and shiny.

Putting it together

The interboard wiring, shown in **Figure 4**, uses screened cable; ideally, this should be thin coaxial cable, but screened microphone cable is suitable. The diagram shows how the two controls, the RF Gain and Volume, are connected to the boards. The leads marked '+9 V' are all connected to the battery supply via a miniature on/off toggle switch. Double check all connections before connecting the battery.

Setting the VFO

Very little adjustment is needed to get the receiver going. Firstly, the VFO must be adjusted to cover the required band, in this case 3.5–3.8 MHz.

If you have a frequency counter, connect it to the output of the VFO. If you haven't, read this part anyway so you understand the process, then another means of setting the VFO will be given especially for you! Rotate VC1 until the vanes are fully meshed. Very carefully, adjust the core of the VFO coil (L1) with a plastic trimming tool so that the frequency approaches and settles at 3.500 MHz. When you rotate VC1, the frequency should increase to **at least** 3.800 MHz at the far end of its travel.

In the absence of a frequency counter, borrow a communications receiver, set it for SSB reception (USB or LSB) on exactly 3.500 MHz. Set VC1 with the vanes fully meshed and turn the core of L1 in both directions until you hear a whistle in the communications receiver. Rotate the core so that the whistle reduces in frequency. It will eventually fade out at around 200 Hz;

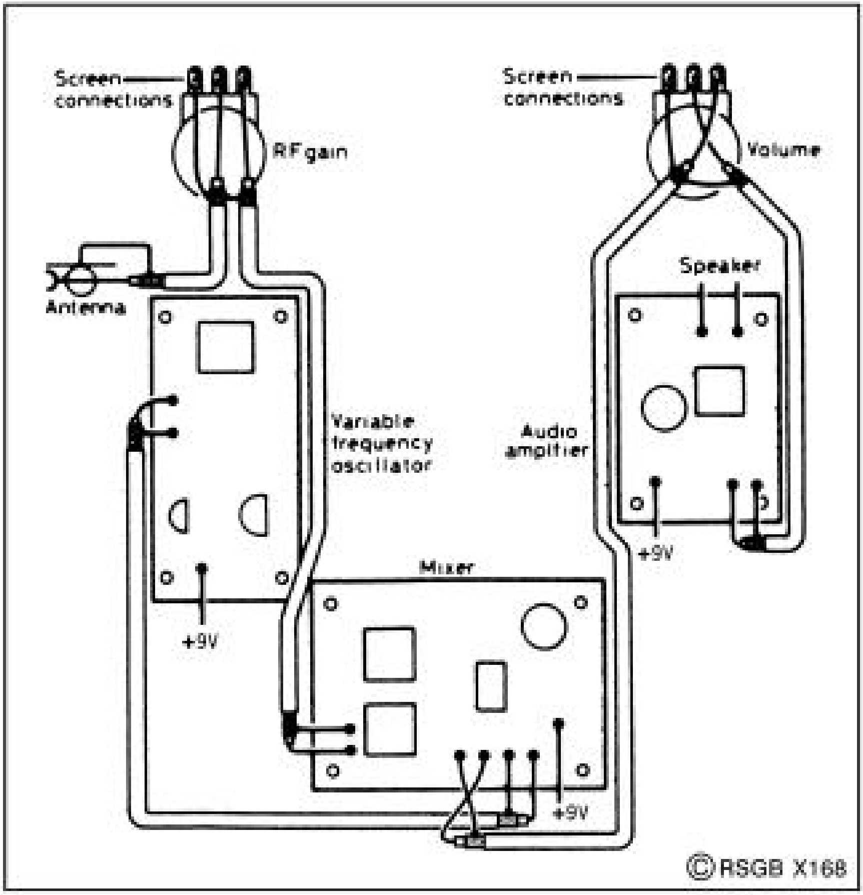
Screen connections
RF gain
Antenna
Screen connections
Volume
Speaker
Variable frequency oscillator
Audio amplifier
+9V
Mixer
+9V
+9V
© RSGB X168

Figure 4 Make sure that interconnections between the boards are correct, including cable screens

turn the core a little further and then leave it at that position. Rotate VC1 right to the other end of its travel, and search for the whistle with the tuning knob of the communications receiver. Check that the VFO frequency is at least 3.800 MHz.

Then, whichever method you are using for frequency measurement, mark the dial with frequency steps of 50 kHz. Setting and calibration are finished!

Setting the mixer

Again, there are various ways of doing this. If you have a signal generator, inject a signal at a frequency within the 80 m band, and adjust the cores of T1 and T2 sequentially for maximum output.

If you haven't a signal generator, connect an aerial to the mixer input, set the RF Gain to maximum (fully clockwise), and find a consistent signal. Adjust the volume control to a comfortable level. Rotating the core of T2 with your trimming tool, maximise the output. Then do the same with T1, although this will have much less effect. Find another station, and check that the positions of the cores aren't too different for a maximum signal.

You may find that your receiver benefits from the insertion of an aerial tuning unit (ATU) between the aerial and the input, to compensate for the impedance of your aerial not being 50 Ω. A design for such an ATU is presented in another part of this series. If the signals are still weak, connect the ATU to the junction of C1 and C2 via a 100 pF capacitor.

Try listening!

Remember that 80 m is a variable band. During daylight hours, your will hear Morse signals at the lower end of the band, and some British and closer continental stations between 3.7 and 3.8 MHz. In the evenings, stations up to 1000 miles away should be heard. Look for Novices around 3.7 MHz!

Parts list – VFO board

Resistors: all 0.25 watt, 5% tolerance		
R1	100 kilohms (kΩ)	
R2, R5, R6	100 ohms (Ω)	
R3, R4	10 kilohms (kΩ)	
RV1	1 kilohm (kΩ) linear	
Capacitors		
C1	12 picofarads (pF) polystyrene	
C2	100 picofarads (pF) polystyrene	
C3	470 picofarads (pF) polystyrene	
C4, C5	1 nanofarad (nF) polystyrene	
C6, C8, C9	10 nanofarads (nF) polystyrene	
C7	100 picofarads (pF) min. ceramic	
VC1	140 + 140 picofarads (pF) variable	
Semiconductors		
TR1	MPF102	FET
TR2	BC182	npn
D1	1N914	silicon
Inductors		
L1	Toko KANK3334	
RFC	1 mH	RF choke

Parts list – mixer board

Resistors: 0.25 watt, 5% tolerance

R1	1 kilohm (kΩ)
VR1	1 kilohm (kΩ) potentiometer (linear)

Capacitors

C1, C3	47 picofarads (pF) min. ceramic
C2	3.3 picofarads (pF) min. ceramic
C4	100 picofarads (pF) min. ceramic
C5	100 nanofarads (nF) min. ceramic
C6, C7	10 nanofarads (nF) min. ceramic
C8	220 microfarads (μF) electrolytic 16 V
C9	1 microfarad (μF) electrolytic 16 V

Integrated circuit

IC1	Philips NE602 or NE602A

Additional items

T1/T2	Toko KANK3333
	On/off switch
	Miniature toggle switch

The next part . . .

The IF amplifier and the Beat-Frequency Oscillator will be added to convert the *Colt* into a superheterodyne receiver.

26 A two-way Morse practice system

Introduction

It is said that the pleasure of sending and receiving Morse code more than compensates for the learning effort needed. Transmitters for Morse code can be much simpler than those needed for any of the speech modes.

Many people choose to learn sending and receiving on their own. It can be much more fun if you have someone to learn with you, and it is for this reason that the following project arose. It comprises just one small circuit, and you build an identical circuit for each of the people who want to learn with you. All the individual circuits are connected with their output leads in *parallel*, as the two-way circuit of **Figure 1** shows.

A simple circuit

This diagram shows two identical circuits, as would be used if two people wanted to learn together. The circuit centres around our old friend the NE555 integrated circuit (IC). As it is connected here, it works as an *astable multivibrator*, a daunting name for what is essentially an oscillator. Each circuit is self-contained, having its own battery, Morse key, sounder and

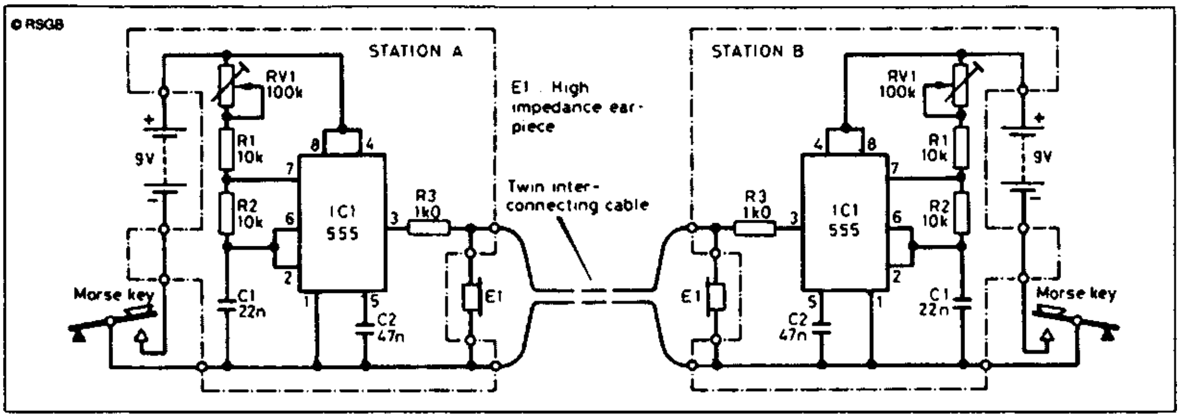
© RSGB
STATION A
STATION B
RV1
100k
R1
10k
R2
10k
9V
+
IC1
555
8
4
7
6
2
1
5
3
R3
1k0
C1
22n
C2
47n
Morse key
E1
E1 . High impedance ear-piece
Twin inter-connecting cable

Figure 1 The circuit diagram shows two stations, but you could connect several more if needed

plastic box. The Morse key makes and breaks the power supply to the circuit, thus turning on and off the note emitted by the earpiece. The frequency of the note is controlled by the variable resistor RV1, resistors R1 and R2, and the capacitor C1. The range given by RV1 is approximately from 500 Hz to 2200 Hz.

R3 protects the output of the IC against accidental shorting either to the 0 V or to the 9 V supply rails. For best results, a crystal earpiece or high-impedance headphones should be used. You could try Walkman-type headphones, but the volume may be too low for you.

The units may be interconnected with thin twin cable (the sort used for wiring doorbells) and can be comfortably separated by 25 m, more than enough to communicate between rooms, or even with a neighbour!

In use

When an operator presses his key, the note is heard by himself and by everyone else who is connected. If more than one key is pressed at the same time, two or more notes are heard by everyone! It helps if each person adjusts his own potentiometer, RV1, to give a note which is different from the others. When one operator stops sending, the other can start immediately, without pausing to change from transmit to receive. In practice, this technique is known as *full break-in*.

The current drawn by the circuit is only 10 milli-amps (mA) when the key is down, meaning that the life of a typical PP3 battery will be virtually its shelf-life! No switch is needed, because the key acts as the switch. The current drain will be even less if the NE555 is replaced by an ICM7555 IC. It has exactly the same pin connections as the NE555, so no circuit modifications are needed.

The circuit board

The circuit is constructed easily on a small piece of Veroboard of the copper-strip type. **Figure 2** shows the layout of the prototype, the board measuring 18 holes by 10 strips. **Be aware that there is no row 'I' when you are transferring mental images of where the parts are to real positions on the board!** Break the copper tracks where shown with a 3 mm (⅛ inch) twist drill, rotated carefully between thumb and forefinger. Hold the board up to a bright light to make sure that the tracks are completely broken and that there are no fragments of copper swarf shorting adjacent tracks. Then, solder in the wire links shown in Figure 2, followed by the resistors, RV1, and the capacitors, C1 and C2. Solder the battery connector leads to tracks A and G, G being positive. Use insulated wire to connect from the board to the jack sockets used as connectors for the key and the earphones. If you use

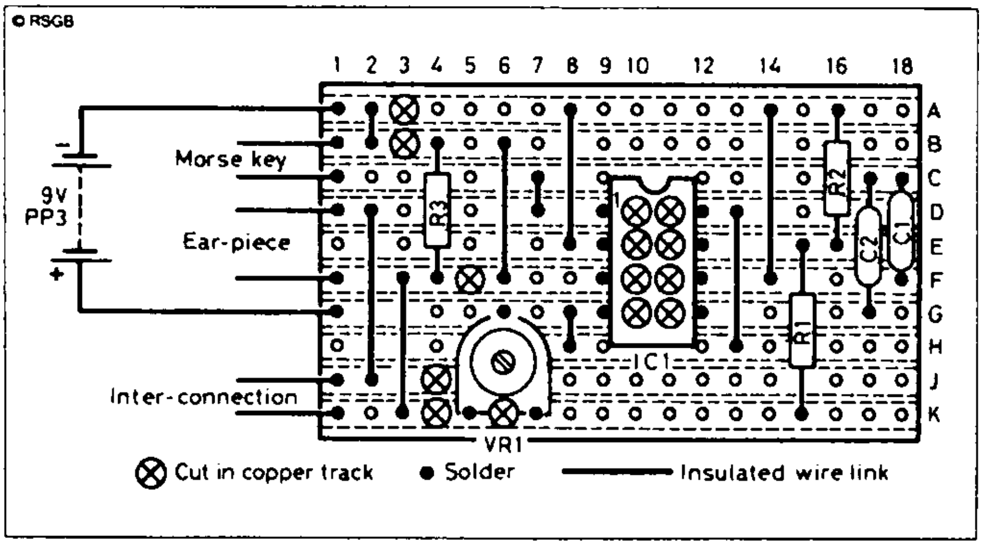
© RSGB
1 2 3 4 5 6 7 8 9 10 12 14 16 18
A B C D E F G H J K
9V
PP3
Morse key
Ear-piece
Inter-connection
R1
R2
R3
C1
C2
IC1
VR1
⊗ Cut in copper track
● Solder
Insulated wire link

Figure 2 The Morse Duet is an easy project to build on stripboard (Veroboard)

different sized jack sockets (i.e. 2.5 mm and 3.5 mm) you won't plug the earphones into the key socket! A two-screw terminal strip is useful to connect the cable runs between operators.

The key

The circuit is designed to be used with a *straight key* (one that moves up and down). It is a requirement of the UK Morse test that a straight key is used, so it is very sensible to learn sending with a straight key before you try anything more complex! You will find many to choose from at rallies.

Parts list

For each board, you will need:		**Maplin order codes**
Resistors: all 0.25 watt, 5% tolerance		
R1, R2	10 kilohms (kΩ)	M10K
R3	1 kilohm (kΩ)	M1K0
RV1	100 kilohm (kΩ) min. preset (horizontal mounting)	UH06G
Capacitors		
C1	22 nanofarads (nF) ceramic, 25 V	WX78K
C2	47 nanofarads (nF) ceramic, 25 V	RA47B

Integrated circuit	
IC1 NE555	QH66W
Additional items	
8-pin DIL socket	BL17T
Jack plug and socket 2.5 mm	HF76H, HF78K
Jack plug and socket 3.5 mm	HF80B, HF82D
Battery type PP3	
Battery connector	HF28F
Plastic box with lid 114 × 76 × 38 mm	LH14Q
Veroboard 0.1 inch with copper strips 18 holes × 10 strips	JP46A
Terminal strip (two section)	FE78K
Crystal earpiece	LB25C
Morse key	

27 The Colt 80 m receiver – Part 4

Introduction

In Parts 1 to 3 of this series, the design of an 80 metre direct-conversion receiver has been described. In this final part, we are going to change the circuit to operate as a superheterodyne receiver, or *superhet*. Most radio receivers are superhets, and they change the incoming signal to another frequency, known as the *intermediate frequency*, or IF, before producing an audio signal. The use of a superhet in a good receiver is mandated by the requirements for good *sensitivity* and *selectivity.*

Sensitivity and selectivity

Figure 1 shows the block diagram of the receiver. If you look closely, you will see that it is two similar circuits, one after the other. The incoming

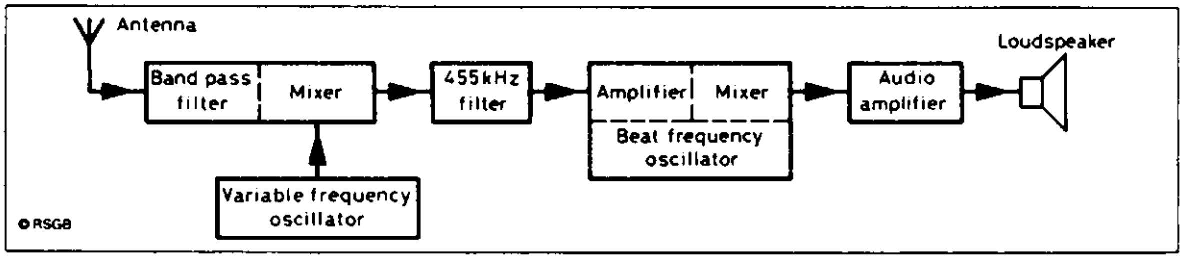

Figure 1 This is how the different stages of the Colt go together to make a superhet receiver

signal is filtered and fed to a mixer where it is combined (mixed) with the signal from the variable-frequency oscillator (VFO). This oscillator operates at a frequency which is 455 kHz higher than the incoming signal from the aerial, and the mixer output is therefore at a frequency of 455 kHz. If you are not sure about this, please refer to the section 'The direct conversion process' in Part 3 of this project. This new frequency is called the *intermediate frequency*, or IF. This frequency doesn't change; the tuning is accomplished by the VFO, and the mixer output is *always* at 455 kHz. The extraction of the audio signal from the IF signal is identical with the direct conversion process which is used in your existing receiver.

This may seem a long-winded way of doing things, but it has its advantages. A receiver must have a good *sensitivity*, or gain, so that it can receive very weak signals. In very general terms, it is easier to handle low-frequency signals than it is to handle high-frequency signals. We are changing our signal frequency from around 3.6 MHz down to 0.455 MHz (455 kHz), which is much lower and can be filtered and amplified relatively easily. A second advantage is that it is easier to provide gain at a fixed frequency than at a variable frequency. Remember that the IF is fixed, and providing gain is, again, relatively simple.

Our receiver also needs good *selectivity*, the ability to separate (or select) one station from another very close to it in frequency. This requirement is significantly simplified by the fact that the IF is fixed, and a good filter in the IF circuits can do wonders for the rejection of adjacent-frequency stations! Several stages of IF amplification and filtering are possible in more adventurous designs.

The filtered signal at 455 kHz passes to another mixer which has an associated oscillator, usually called a *beat frequency oscillator* (BFO). When receiving CW (Morse) signals, the BFO is usually tuned about 1 kHz above or below the IF (i.e. at 454 or 456 kHz) to produce a 1 kHz beat note as the audio signal. This signal is then amplified and fed to a loudspeaker or headphones.

The circuit

Figure 2 shows the circuit of the IF section shown in the photograph on p. 94; this section is added to the existing circuit to make it a superhet. The existing mixer board (described in Part 3) is used as the first mixer. Between it and the audio amplifier is connected the new IF board.

The signal output from the first mixer (at 455 kHz IF) is fed into a *crystal filter*, the most expensive part in the whole receiver. It provides the selectivity which makes the Colt such a good receiver. Another NE602 mixer/oscillator chip follows the filter. The oscillator section is controlled by the tuned circuit in T1, the frequency of which can be altered by rotating the core inside the coil. Once set, it remains fixed.

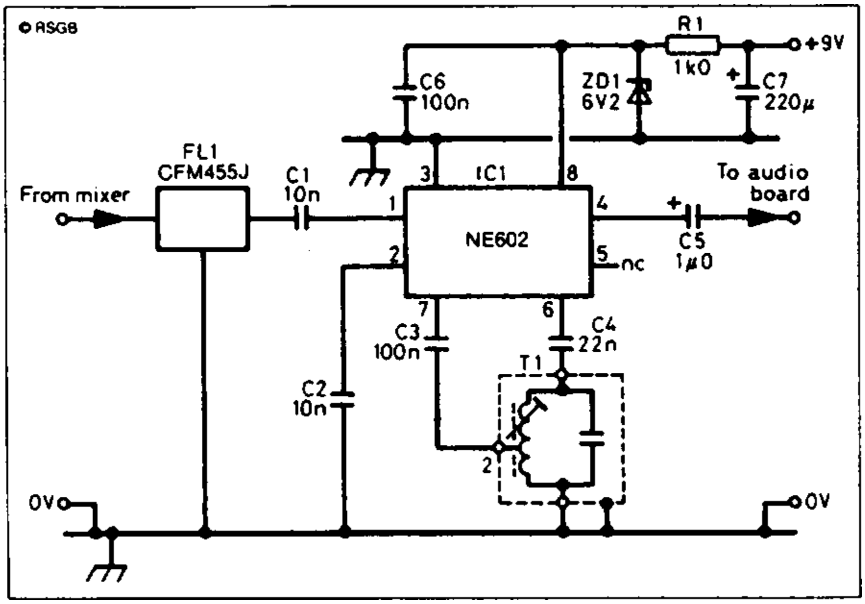

Figure 2 The IF board is connected between the mixer and audio amplifier

At this stage you should have built the first mixer and the audio amplifier, and proved that they *both* work by using the circuit as a direct-conversion receiver. When you have finished constructing and checking the IF board, you will need to add it to your existing circuits.

Adjusting and testing

There are three pairs of connections to the IF board – the 9 V supply leads, the input leads and the output leads. The IF board should be mounted on the metal baseplate, along with the other boards. **In making the following connections, make absolutely sure that the braid of each piece of coaxial**

cable is soldered to the correct connection on each board. The same applies to the polarity of the 9 V leads. Disconnect the screened lead which presently goes from the first mixer board to the AF amplifier at the amplifier end. Connect it instead to the *input* of the IF board. Connect a new piece of screened cable from the IF output to the vacated AF input. Now connect the 9 V supply leads to the same points as the supply leads from the other boards. Check these newly made connections.

You should have confidence at this point that things should be right. After all, you *have* tested the direct-conversion process and you *know* it works. All that you are now testing is the IF board. This is the attraction of building a receiver in modules, working from the speaker backwards, and testing as you go!

The VFO and BFO need to be correctly adjusted. Mesh the vanes of the tuning capacitor and, using the same frequency measurement method as you did originally, set the VFO frequency to 3.955 MHz (which is 3.500 MHz + 0.455 MHz, if you hadn't guessed!). The BFO can be set using a frequency counter, but it is just as good to set it by listening to SSB or CW signals. A high-pitched hissing sound should be heard in the speaker. As you rotate the core, the pitch should reduce, go through a minimum, then increase again. Set the core at the minimum pitch position. You may want to readjust the two cores a little as your listening skills improve but, once you are happy, they will never need to be altered again!

In conclusion . . .

You should now have built a superhet receiver capable of excellent results. It uses the same type of circuit as that found in far more advanced receivers.

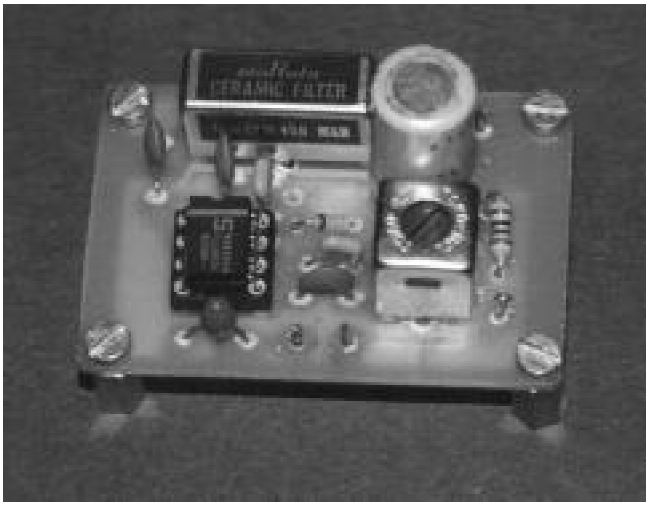

The superhet is far more sensitive than the direct-conversion type, and can weed out those elusive DX stations. You may have found that a station will appear at two places on the dial of the direct-conversion receiver; you will have no such problem with the superhet.

Parts list

Resistor	
R1	1 kilohm (kΩ), 0.25 watt, 5% tolerance
Capacitors	
C1, C2	10 nanofarads (nF) min. ceramic
C3, C6	100 nanofarads (nF) min. ceramic
C4	22 nanofarads (nF) min. ceramic
C5	1 microfarad (μF) 16 V electrolytic
C7	220 microfarad (μF) 16 V electrolytic
Integrated circuit	
IC1	NE602
Additional items	
ZD1	Zener, 6.2 V 0.5 watt
FL1	Crystal filter, Murata CFM455J
T1	Tuned inductor, Toko YHCS11100AC

28 A simple crystal set

Introduction

There are many designs of crystal set – they are all ‘simple’, and even the most seasoned radio amateur will build one of these every so often because it is something that never ceases to amaze! Although using modern components, the design is a period piece!

Design and construction

The circuit diagram of **Figure 1** shows the simplicity of the circuit. The prototype was constructed on a wooden baseboard with an aluminium front panel, although this could be made of wood if you prefer.

Winding the coil always causes the most groans but when finished, gives the most satisfaction. The construction of the coil is detailed in **Figure 2**. A discarded toilet roll centre is an ideal *former* on which to wind the coil. The coil will take up about 7 cm of the length of the tube, so assuming the tube is about 11 cm long, the ends of the coil will be about 2 cm in from each end of the tube.

The aerial coil, the *primary winding* of this radio-frequency (RF) transformer, is made from about 30 turns of 26 SWG enamelled copper wire, wound on the matchstick ribs. These ribs should be about 2 cm long, and are glued on **top** of the end of the secondary coil *after* the secondary coil has been wound. The *secondary winding* has 140 turns with *taps* every 10 turns for 70 turns. This coil covers most of the length of the cardboard tube. To make your coil-winding easier, here are some tips for completing the coil and still having some hair left at the end!

- Support the reel of wire on a dowel rod or pencil, clamped in a vice or held rigid in a vertical position by some other means. Don't leave it to trail around on the floor.
- Have a small piece of sandpaper handy to remove the enamel from the copper wire, starting with the end coming from the reel. Remove about 1 cm only, until the shiny copper is visible all round the end of the wire.

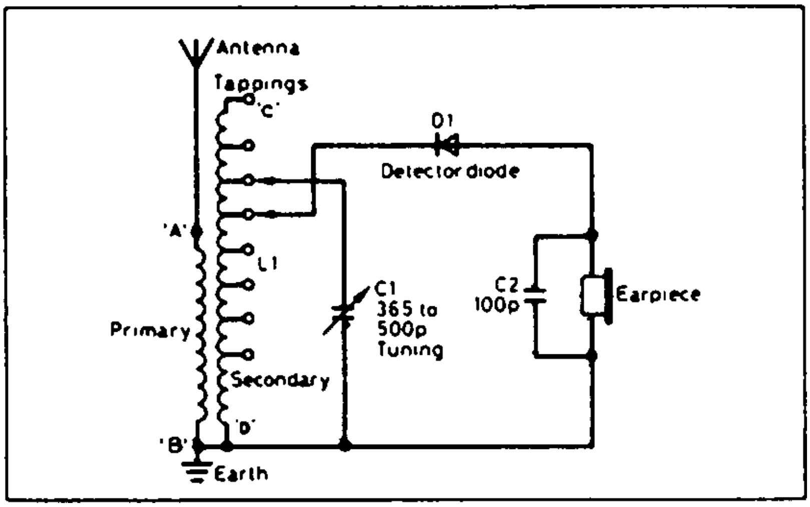

Figure 1 Circuit diagram of the crystal set

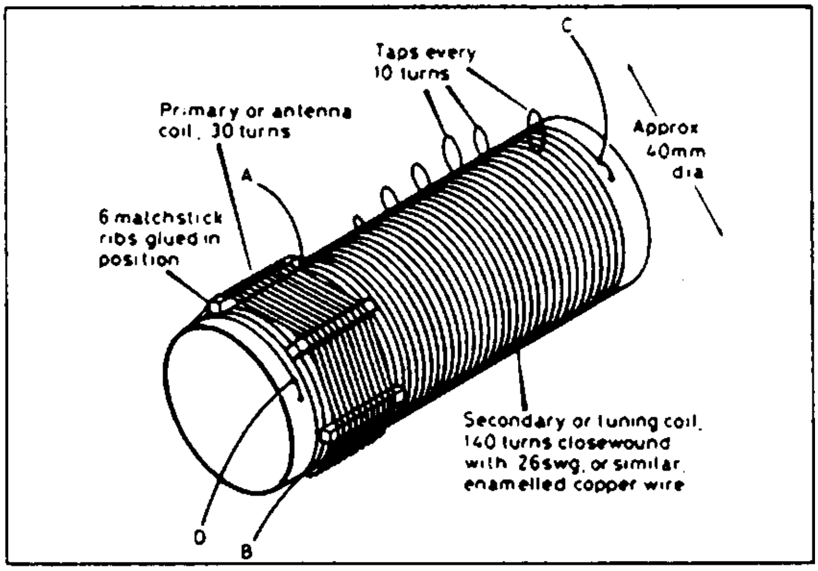
C
Taps every 10 turns
Primary or antenna coil, 30 turns
A
Approx 40mm dia
6 matchstick ribs glued in position
Secondary or tuning coil, 140 turns closewound with 26swg, or similar, enamelled copper wire
D
B

Figure 2 The coil is mounted on a cardboard tube as shown. Exact size is not too important. Try different taps for best results

- This part is the most difficult of all. Seasoned coil winders may have their own favourite ways of doing this, but for first-timers, play it safe and do it this way! Have a second bobbin (or a piece of stout card) ready, on which you will need to wind the wire loosely while you prepare the taps. Then, starting from the free end of the wire, measure 125 cm and, at this point, remove about 2 cm of the enamel with the sandpaper. Bend the wire firmly back on itself in the middle of the bared area, and *tin* (cover with solder) the exposed copper. This is your first tap. Then repeat this six more times at 125 cm intervals, thus giving you seven taps; wrap the wire as you go, on to the second bobbin or piece of card. After the seven taps, you are half-way down your coil so, to begin winding it in earnest, wind the wire carefully from the second bobbin or card **back** on to the original reel of wire.
- Looking at Figure 2, make two small holes about 3 mm apart, about 2 cm down from the top of the tube, where wire C will be entering. Poke the end of the wire into one hole, then bring it out again through the other; leave about 10 cm of wire on the free end. Loop this wire in and out once more, thus anchoring the wire firmly.
- Then, begin winding; keep the wire tight and make adjacent turns touch neatly. As you pass each tap, make sure it sticks outwards while avoiding flexing the wire too much; the wire is inherently weak at each tap.
- When you reach the last 10 cm of wire, stop. Put two more holes, like the first, beside your stopping point, and anchor the end, D, of the coil in the same way as you did with end C. The secondary winding is complete. Take a break *after* you have completed the next step!

- Prepare the six matchsticks for supporting the primary coil and glue these every 60 degrees around the end of the coil at D. Remove the enamel from the free end of the new copper wire on your reel and, using your favourite quick-drying cement, put a drop over the wire on one of the matchsticks, leaving 10 cm free as before. If you are using 26 SWG enamelled copper wire, its diameter is 0.46 mm, so 30 turns should occupy 14–15 mm of the 20 mm matchstick length. This will allow you to position the glued point such that the coil will be roughly in the centre of the matchsticks. When the glue has set, wind the 30 turns and anchor this end of the coil in the same way. Then remove the enamel at the end of the 10 cm lead.

Assembly

The coil former can be screwed to the baseboard at each end, *before* the front panel is screwed to the baseboard! As the photograph on the next page shows, the tuning capacitor is mounted in the centre of the front panel. Five solder tags (or drawing pins as a last resort) should be screwed into the baseboard. The two to the left of the photograph are the connections to the earphones, the two along the rear edge have the diode, D1, soldered to them. One end of the diode (and, for once, you can connect the diode either way round!) is connected to one earphone tag. The other end has a flying lead of about 10 cm attached to it and terminated in a crocodile clip which is used to connect to one of the taps on the coil. The fifth tag secures the end, A, of the primary coil, to which you will attach the aerial. The earth tag on the tuning capacitor, C1, serves as an anchor point for the ground connections to B and D and to the other earphone tag.

Use

The longest piece of wire you have available to use as an aerial should be connected to the aerial tag just mentioned. If you can connect the earth tag of the tuning capacitor to a good electrical earth, this will help also. You should be able to hear something as you turn the tuning knob. Try adjusting the tapping connections on the coil – change only one at a time, or you will never find the optimum positions. You now have a fully operational crystal set!

Simple it may be, but this circuit illustrates some important principles which are used even in the most expensive receivers. Firstly, the coil, L1, and the capacitor, C1, form a *tuned circuit* which *resonates* at the frequency of the station you have tuned in. This *selects* the signal you want to hear. The *modulation* on this signal is removed by the *detector*, D1, and fed to the earphones, which act as *transducers*, turning the electrical energy into sound energy which you can hear. It doesn't need a battery or other power supply either!

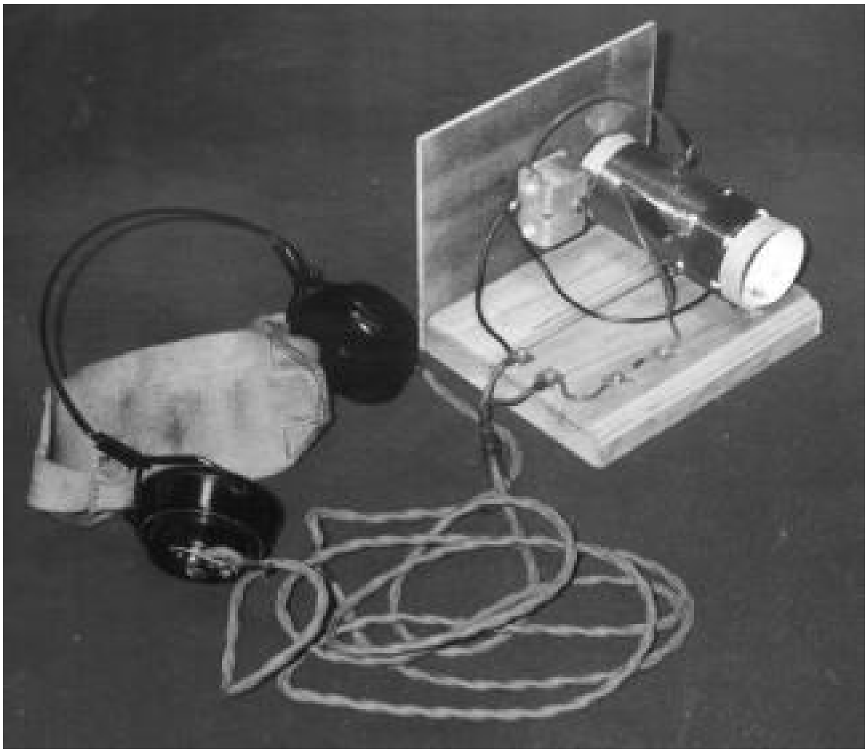

Parts list

Capacitors

C1	Variable capacitor of between 200 and 500 picofarads (pF) maximum
C2	100 picofarads (pF) min. ceramic

Additional items

D1	Germanium, type OA47 or OA90
	High impedance crystal earphone (not Walkman type)
	Tuning knob
	Wooden base, approx. 110 mm square
	Aluminium (or wood) front panel, approx. 80 × 100 mm
	Cardboard tube (toilet roll centre), about 110 mm long, 40 mm diameter
	Reel of 26 SWG enamelled copper wire for both windings of L1

29 A crystal calibrator

Introduction

A crystal calibrator is a device that produces oscillations of a precise frequency which are rich in *harmonics*. A harmonic is an integer multiple of the fundamental frequency; for example, if our fundamental frequency was 100 kHz (as it is in this circuit), there would be harmonics of this frequency at 200 kHz, 300 kHz, 400 kHz, and so on. These harmonics can be used, when picked up by any receiver, to calibrate that receiver, as the harmonics are quite accurate in frequency (see later for an assessment of accuracy). However, a gap of 100 kHz between harmonics is rather wide for most purposes, so we reduce this frequency to 25 kHz, so that the harmonics are then 25 kHz apart, thus producing a much more useful set of *marker points*. A calibrator such as this is often called a *crystal marker*, producing these marker points from 25 kHz to beyond 30 MHz.

The circuit

The circuit diagram is shown in **Figure 1**. The circuit around TR1 is the fundamental oscillator, and its frequency is controlled by the quartz crystal, X1. Even crystal oscillators are not 100% accurate, and the small trimmer capacitor, TC1, is able to ‘pull’ the frequency to one which is nearer the

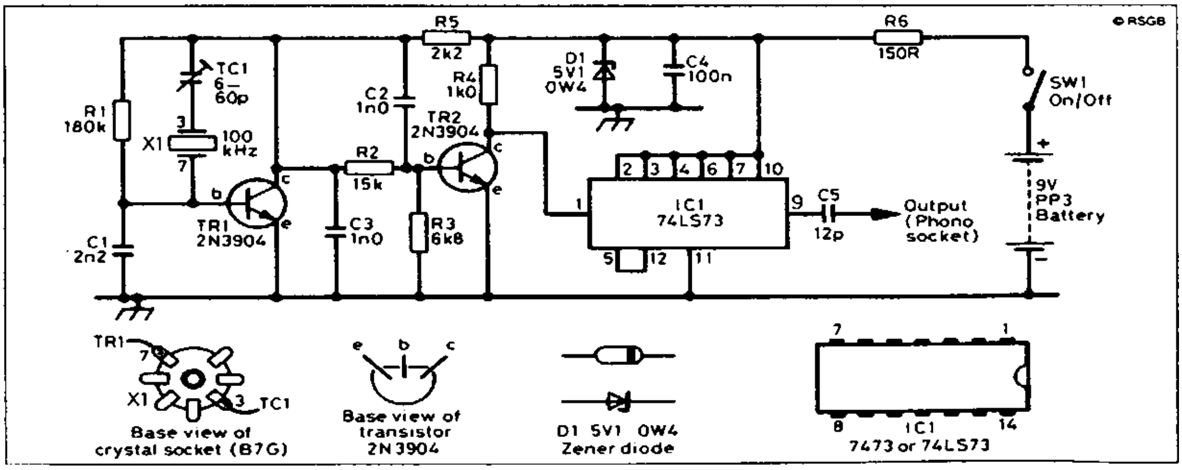
© RSGB
R5
2k2
R6
150R
TC1
6–
60p
R1
180k
X1
3
7
100
kHz
C2
1n0
R4
1k0
D1
5V1
0W4
C4
100n
SW1
On/Off
TR2
2N3904
b
c
e
R2
15k
TR1
2N3904
C1
2n2
C3
1n0
R3
6k8
2
3
4
6
7
10
1
IC1
74LS73
9
C5
12p
Output
(Phono
socket)
5
12
11
+
9V
PP3
Battery
−
TR1
7
X1
3
TC1
Base view of
crystal socket (B7G)
e
b
c
Base view of
transistor
2N3904
D1 5V1 0W4
Zener diode
7
1
8
14
IC1
7473 or 74LS73

Figure 1 Circuit diagram of the calibrator which uses easily obtainable components

correct one. TR2 is a *buffer* stage, which isolates the oscillator from the rest of the circuit. It acts as a switch, and applies a good signal (switching between 0 and 5 V) to the input of IC1.

IC1 is an integrated circuit which can be connected to do several things. It is connected here to divide the incoming frequency by a factor of 4, producing on pin 9 an output frequency of 25 kHz.

The combination of R6, C4 and D1 produces a supply of 5 V for IC1; it would be damaged if the battery voltage were applied to it. You will no doubt be ready to assemble the circuit, so here is some information for you.

Construction

If possible, always build a circuit in individual stages, which you can test as you go along. It is not always easy in small projects like this, but even the crystal calibrator can be split into two for construction and testing.

It is an ideal project for assembly on Veroboard of the copper-strip type; the prototype layout is shown in **Figure 2**, on a board measuring 11 holes by 24 strips. First of all, remove the copper strip at the locations shown, using a 3 mm drill rotated between the fingers. Hold the board up to the light to ensure that there are no pieces of copper swarf bridging adjacent strips, and that the copper is completely removed where it should be! Assemble the circuit from left to right, but do not wire up anything around IC1. Be aware that the diagram shows the board *from the component side*. Although you

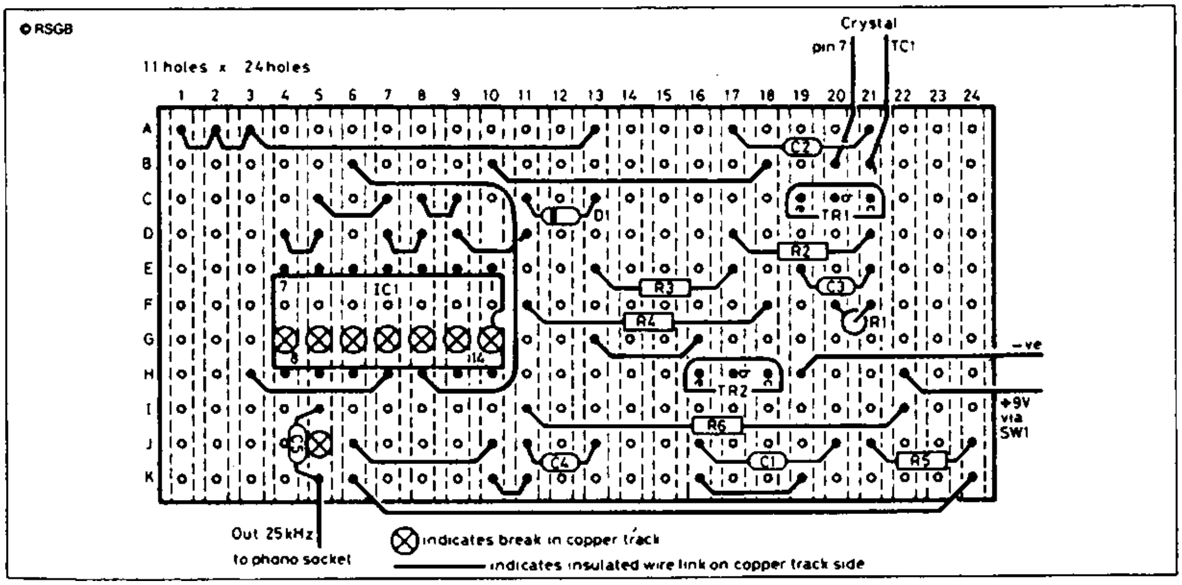
©RSGB
Crystal
pin 7
TC1
11 holes x 24holes
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
A B C D E F G H I J K
IC1
7
8
14
D1
C2
TR1
R2
C3
R3
R4
R1
TR2
R6
C1
C4
R5
C5
−ve
+9V via SW1
Out 25kHz to phono socket
indicates break in copper track
indicates insulated wire link on copper track side

Figure 2 Stripboard, such as Veroboard, should be cut as shown to approx. 62 × 28 mm

can use virtually any kind of crystal, you will need to mount it firmly. If you use a small one, it can be soldered directly in position on the board. With one in a valve envelope, you will need a B7G valveholder. Be prepared to mount the circuit in a **metal** container; the prototype used an (empty) tin of tuna! The valveholder mounts well on any metal case.

Disconnect the aerial of your receiver, and replace it with about 30 cm of wire laid near your oscillator circuit. Connect the battery and switch on. Tune the receiver around until you can hear a whistle on SSB. Rotate the tuning knob to reduce the frequency of the whistle and, as the whistle becomes too low to hear, you have reached one of the calibration markers, and your frequency will be an integer multiple of 100 kHz. If you are already using a calibrated receiver, you will be able to verify this. Going up or down in frequency should locate another marker 100 kHz away, and so on right through the receiver's tuning range. If the signals coming from the circuit are very weak, switch off the calibrator and connect another 30 cm piece of insulated wire to the collector of TR2, and lay it close to (but *not* touching) the wire from the aerial connector. Switch on an try again; you should not now have a problem! If you have access to a multimeter, check that the voltage at the collector of TR1 is very close to 5 V. If it is close to 9 V, you have connected the diode, D1, the wrong way round!

Having verified that the oscillator is working, you can now wire up the integrated circuit socket, being careful to put the notched end of the socket pointing towards R4, the collector resistor for TR2.

Check your connections around IC1, and when you are satisfied that they are correct, line up IC1 with its socket, making sure that the notched ends are together, and press down *gently* to insert the IC into its socket. Insert the 30 cm piece of insulated wire into the output socket, connect the battery and switch on. You should still hear whistles in your receiver, but now they should be 25 kHz apart, rather than 100 kHz apart.

All that now remains to be done is to mount the circuit rigidly inside whatever casing your have chosen. Make sure that none of the connections under the board touch the metal case, and secure the valveholder, on/off switch and output connector to the case. You now have a completed crystal calibrator.

Calibration

The simplest way to calibrate your circuit is with a frequency counter. Most clubs will have one of these and, if not, will know someone who has! Connect it to the collector of TR2, where the frequency should be 100 kHz. Do *not* connect it to any part of the circuit around TR1, or you may alter the frequency you are trying to measure! If the frequency is not exactly 100.000 kHz, rotate TC1 until it is (or is as close as you can get it). Now your calibrator is as accurate as the counter with which you have calibrated it.

Accuracy

Despite your best efforts at calibration, by whatever means, your crystal will *never* have a constant frequency. Such a thing is a scientific impossibility. It is usual to express the accuracy of a crystal in parts-per-million (ppm), and it is governed by many things, principally its temperature. You will be *very* fortunate if your circuit maintains an accuracy better than about ±10 ppm. Expressed in figures, it means that the true frequency can be anywhere between 99.999 kHz and 100.001 kHz, i.e. within 1 Hz of the correct frequency.

Although this may seem more than adequate, it is as well to remember that the accuracy degrades as the frequency increases. At 1 MHz the error will be ±10 Hz and at 10 MHz it will be ±100 Hz. At 30 MHz it will be 300 Hz. Even so, this should be acceptable for most non-critical applications.

Parts list

Resistors: all 0.25 watt, 5% tolerance		**Maplin code**
R1	180 kilohms (kΩ)	M180K
R2	15 kilohms (kΩ)	M15K
R3	6.8 kilohms (kΩ)	M6K8
R4	1 kilohm (kΩ)	M1K0
R5	2 kilohms (kΩ)	M2K0
R6	150 ohms (Ω)	M150R
Capacitors		
C1, C2, C3	1 nanofarad (nF) or 1000 picofarads (pF)	WX68Y
C4	100 nanofarads or 0.1 microfarad (µF)	YR75S
C5	12 picofarads (pF)	WX45Y
TC1	60 picofarads (pF) trimmer	WL72P
Semiconductors		
TR1, TR2	2N3904 npn	QR28F
IC1	7473 or 74LS73N dual JK flipflop	YF30H
D1	5.1V 500 mW Zener	QH07H
Additional items		
100 kHz crystal		
B7G valveholder (found at most rallies)		
14-pin DIL IC socket		BL18U
Phono socket		YW06G
PP3 battery connector		HF28F
PP3 battery		
On/off switch		FH00A
Veroboard		JP47B
Metal case as required		

30 A simple short-wave receiver – Part 1

Introduction

No apologies for yet another design of short-wave radio. This is the beauty of our hobby – there is always something new to try. Not all circuits operate in the same way; not all circuits work equally well; sometimes a simple design suits the operator better than a complicated design. So here is another one for your consideration; it's a good project for a novice, *and* a good one-evening project for someone more experienced.

A basic description

Figure 1 gives the block diagram of the receiver, which employs a *regenerative detector*, one of the earliest techniques by which excellent *selectivity* (the ability to separate two stations very close in frequency) could be combined with good *sensitivity* (the ability to pick up very weak stations) using a very simple circuit. The receiver falls into the category known as *tuned radio frequency* (TRF), meaning that the whole circuit (prior to the extraction of the modulation from the carrier) operates at the incoming radio frequency. In other words, it is *not* a superhet.

As you can see, the regenerative detector has what is called a *feedback loop*, which feeds a small amount of the output signal back to the input. You have heard the effects of feedback with a public address system, when the microphone gain is too great – everything becomes very loud and then bursts into a deafening squeal! This is exactly what the feedback does here, except that it is carefully controlled, thus providing both gain and selectivity. The resulting audio is then amplified for use with headphones.

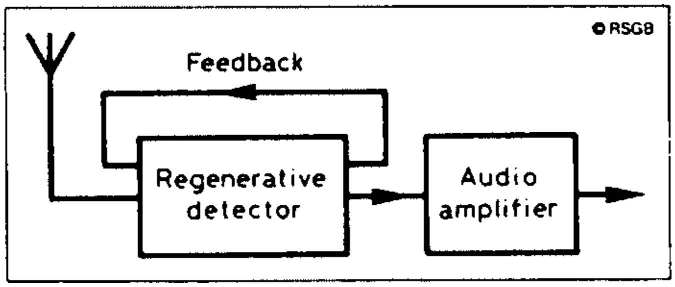

Figure 1 Just two simple stages make up the circuit

The circuit

The circuit of **Figure 2** shows the complete system. TR1 is an untuned *field-effect transistor* (FET) stage, and is used to match the aerial to the next stage. Occasionally, a regenerative detector produces unwanted signals, and TR1 also prevents them from reaching the aerial and being transmitted! The smaller of the two windings on T1 (the *primary* winding) will match a low-impedance aerial, the capacitor input matching a high-impedance aerial.

Ignore TR2 for the moment – the next stage in the signal path is TR3, the detector, using another FET. The tuned circuit is formed by T2, VC1 and VC2 (remember TR1 is untuned). The primary winding on T2 taking the output from TR1. The reason for having two variable tuning capacitors, one large, the other small, is that the large capacitor is the main tuning capacitor, while the small one is used as a *bandspread* control (for very fine tuning). The tuned signals are then detected (converted to audio frequencies) by TR3.

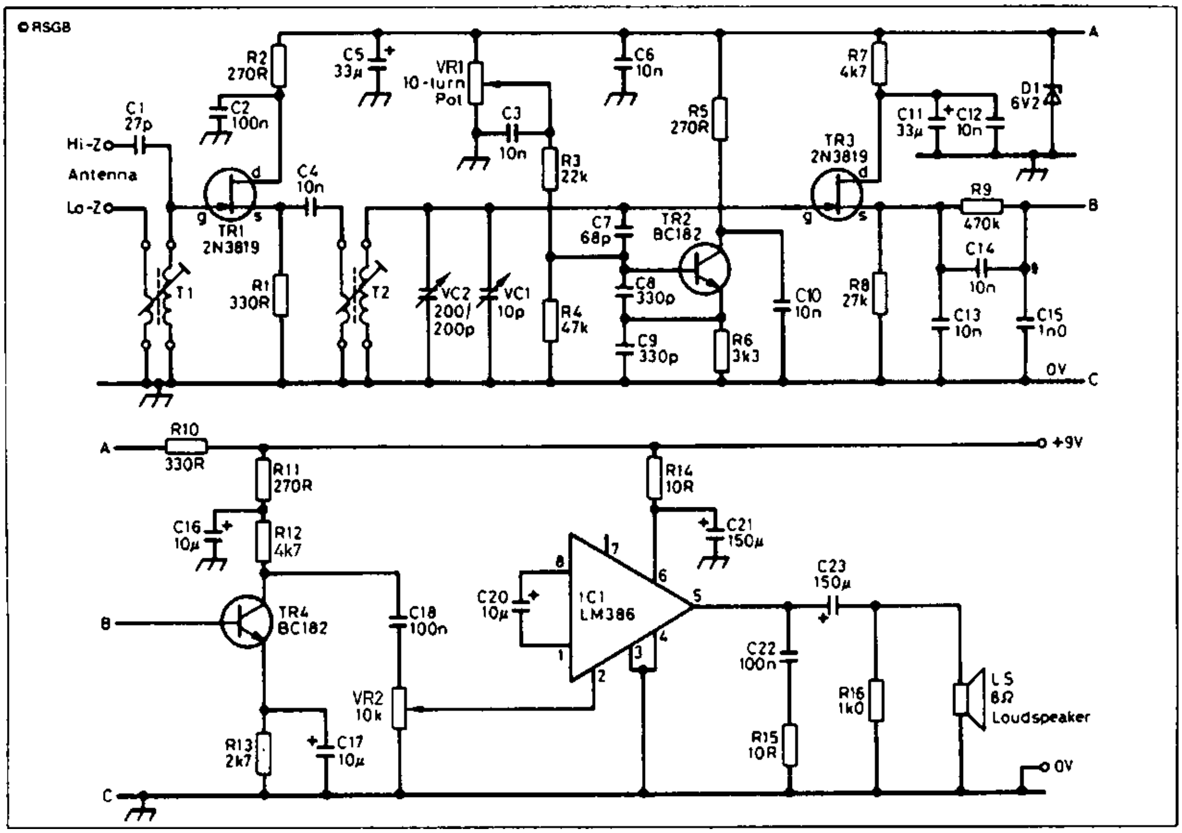
©RSGB
R2 270R
C5 33µ
VR1 10-turn Pot
C6 10n
R7 4k7
A
C1 27p
Hi-Z
Antenna
Lo-Z
C2 100n
C3 10n
R5 270R
C11 33µ
C12 10n
D1 6V2
TR3 2N3819
d
g
s
C4 10n
R3 22k
R9 470k
B
TR1 2N3819
T1
R1 330R
T2
VC2 200/200p
VC1 10p
C7 68p
TR2 BC182
C8 330p
C9 330p
R4 47k
R6 3k3
C10 10n
R8 27k
C14 10n
C13 10n
C15 1n0
0V
C
A
R10 330R
+9V
R11 270R
R14 10R
C16 10µ
R12 4k7
C21 150µ
8
7
6
5
C23 150µ
C20 10µ
IC1 LM386
B
TR4 BC182
C18 100n
1
2
3
4
C22 100n
VR2 10k
R16 1k0
LS 8Ω Loudspeaker
R13 2k7
C17 10µ
R15 10R
0V
C

Figure 2 The circuit diagram showing the 10-turn potentiometer VR1. This is the regeneration control for the circuit

TR2 is a *Q-multiplier* stage. It is a Colpitts-type oscillator which uses C8 and C9 to give feedback and produce the oscillation. C7 couples the tuned circuit (and the input to TR3) to TR2. The tuned circuit thus controls the frequency of oscillation of TR2 and the signal passed on to TR3.

The secret of the ease of operation of this receiver is VR2, a 10-turn potentiometer (or *helipot*). The resistance wire is wound in the form of a helix, giving a much greater wire length than in a normal potentiometer, and the shaft must be turned ten times to cover the whole length. Helipots are very useful when very fine adjustments have to be made. Here, VR2 sets the regeneration (or *reaction*) level, depending on the type of signal you are receiving, as will be discussed in Part 2.

TR4 provides the first stage of audio amplification and, after the volume control, VR2, the audio amplifier integrated circuit, IC1, will drive a small loudspeaker or headphones.

In Part 2, the construction will be discussed, together with the choice of aerial, the parts list, and advice on using the receiver.

31 A fruit-powered medium-wave radio

Introduction

This is a one-evening project that will result in a working medium-wave (MW) radio, and will also teach you a little about the way electricity can be generated from the right metals and a little (safe) acid. All you need are three lemons or other citrus fruit, three pieces of copper and three pieces of zinc (or galvanised metal) for your power supply.

Construction

Figure 1 shows the circuit and **Figure 2** its layout on a simple ‘plug-in’ prototype board. The six pieces of metal are connected as shown, to wire the three lemons in series; use ordinary wire between each lemon and the next. If you have a meter to measure the total voltage, it should be about 1.8 V. Use a standard ferrite rod, and wind on it about 40 turns of single-conductor PVC-insulated wire.

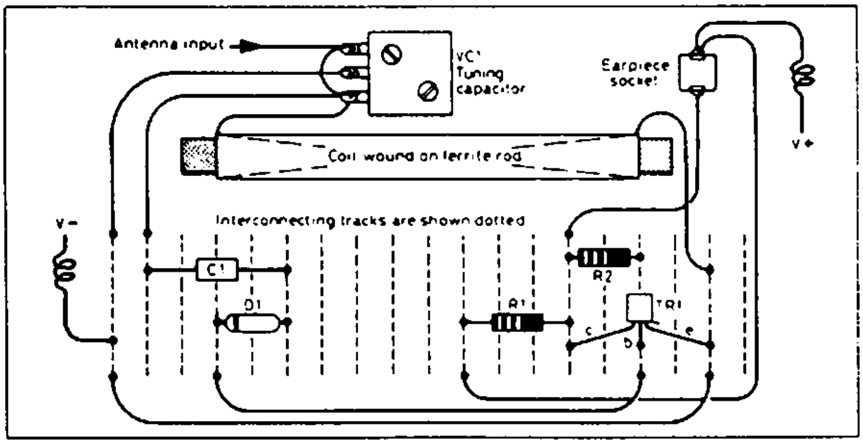
Antenna input
VC1
Tuning
capacitor
Earpiece
socket
V+
Coil wound on ferrite rod
Interconnecting tracks are shown dotted
V−
C1
D1
R2
R1
TR1
c
b
e

Figure 1 Most parts are plugged into the board as shown – soldering is not required

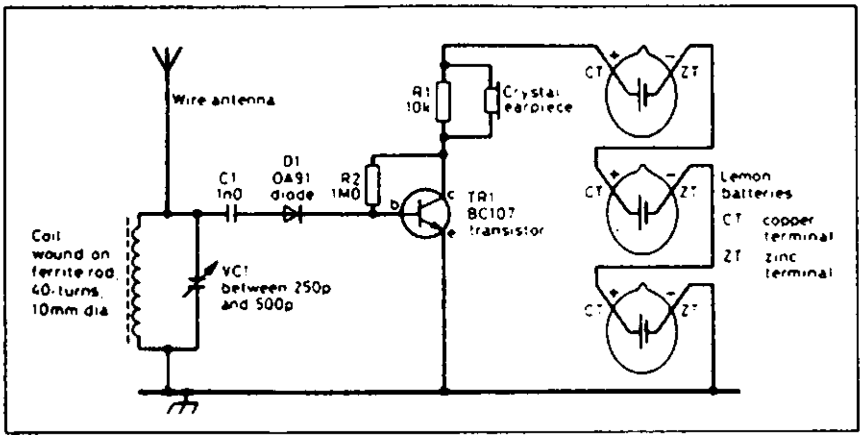
Wire antenna
R1
10k
Crystal
earpiece
C1
1n0
D1
OA91
diode
R2
1M0
b
c
e
TR1
BC107
transistor
Coil
wound on
ferrite rod
40 turns
10mm dia
VC1
between 250p
and 500p
CT
ZT
Lemon
batteries
CT copper terminal
ZT zinc terminal

Figure 2 Three lemons power the radio, which gives good results for such a simple circuit

Wire up the circuit on the board as illustrated in Figure 2. Soldering is not required with this type of board – just plug in the components and the wires. Only one transistor is needed. The tuning capacitor, VC1, selects the station you want to hear, and D1 helps to remove the carrier from the RF signal. The resulting audio signal is fed to TR1, a small transistor audio amplifier, which makes the signal big enough to drive a crystal earpiece comfortably. Walkman-type earphones will not work, so invest in a crystal earpiece which you can use in several other projects, too! If you use a smaller capacitor than that specified for VC1, you will need more turns on the aerial coil.

Tests on the prototype indicated that the radio will run for about a week on three lemons!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	1 megohm (MΩ)
R2	10 kilohm (kΩ)

Capacitors

C1	1 nanofarad (nF) min. ceramic
VC1	250–500 picofarad (pF) variable

Semiconductors

TR1	BC107 npn (or BC108, BC109C)
D1	OA90, OA91 germanium (not silicon)

Coil

L1	2 metres of single-conductor insulated wire on a standard ferrite rod

Additional items
- Plug-in prototype board, e.g. Maplin YR84F
- Wire aerial at least 3 m long
- Crystal earpiece
- Three juicy fruits

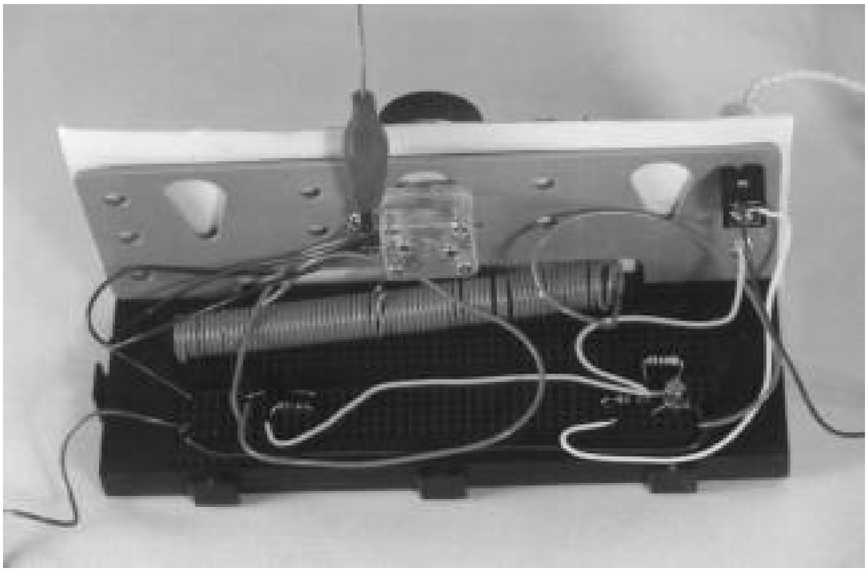

32 A capacitance bridge

Introduction

It is always valuable to have a means of measuring capacitors amongst your test gear. Universal LCR bridges (i.e. systems that can measure inductance, capacitance and resistance) can be found at most rallies, and you can work your way up to owning one of these. In the meantime, a piece of home-made equipment gives you experience as well as resulting in a useful measuring instrument.

The circuit described here is called a *capacitance bridge*, because it balances the effects of one resistor/capacitor pair against another; if one capacitor has an unknown value, then the other can be calculated. The basic bridge circuit is shown in **Figure 1**. To avoid having calculations to perform, this instrument will be calibrated by using capacitors of known values. The bridge is a useful way of performing measurements, because a knob is turned until there is a null in the signal from an earpiece or loudspeaker. The ear is very precise in being able to perceive nulls, which makes the bridge easy to use and reasonably accurate. At the null, the bridge is said to be *balanced*.

How does it work?

Figure 2 is the circuit diagram for this capacitance bridge. Transistors TR1 and TR2 form an oscillator. This is the audio oscillator shown in Figure 1, and produces an alternating voltage which is fed to the bridge. RV1 (in the collector lead of TR2) replaces both R1 and R2 in Figure 1 – that part of

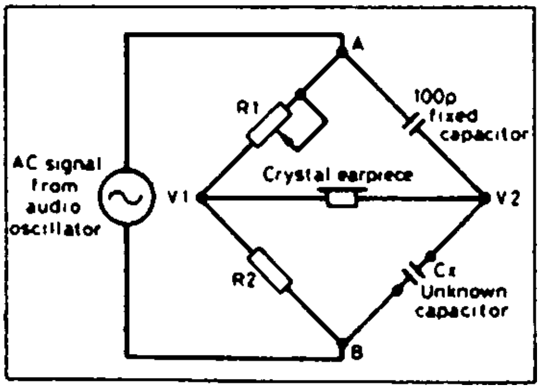

Figure 1 Simplified circuit of a capacitance bridge. R1 is adjusted for minimum sound

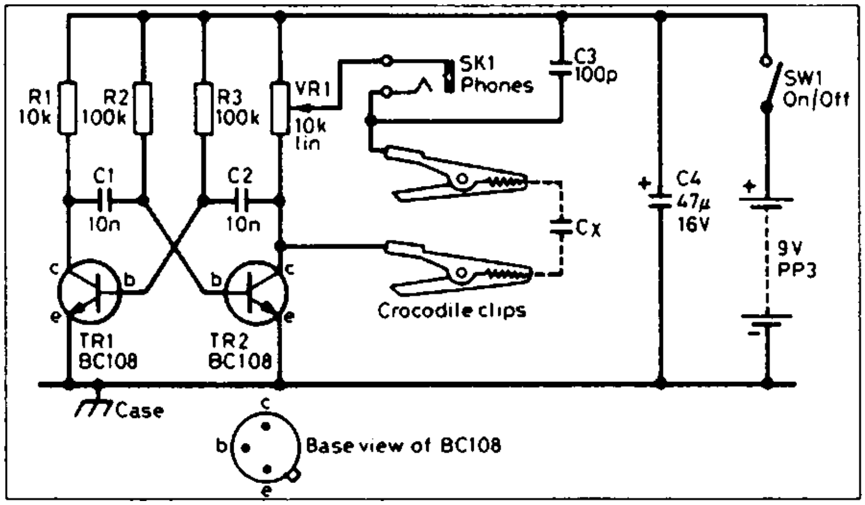
R1
10k
R2
100k
R3
100k
VR1
10k
lin
SK1
Phones
C3
100p
SW1
On/Off
C1
10n
C2
10n
Cx
C4
47µ
16V
9V
PP3
c
b
e
TR1
BC108
TR2
BC108
Crocodile clips
Case
Base view of BC108

Figure 2 Transistors TR1 and TR2 give an audio signal which is adjusted by variable resistor RV1

RV1 *above* the wiper represents R1, while that part *below* the wiper represents R2. The voltage on one side of the earpiece is determined by the ratio of these values, and is adjusted by rotating RV1. The voltage on the other side of the earpiece is determined by the ratio of C3 to Cx, where Cx is the unknown-valued capacitor. When these two voltages are the same, the bridge is balanced and there is no current through the earpiece.

Figure 3 shows the layout of the components on the matrix board of the prototype. It measures 10 holes by 14 holes, and is of the plain type, i.e. no copper strips. All earth connections are taken to a single solder tag. When

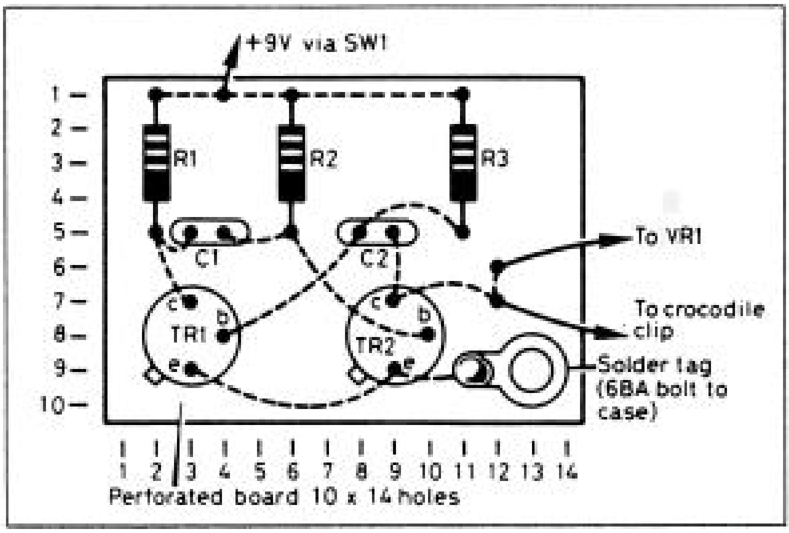
+9V via SW1
1
2
3
4
5
6
7
8
9
10
R1
R2
R3
C1
C2
TR1
TR2
c
b
e
To VR1
To crocodile clip
Solder tag (6BA bolt to case)
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Perforated board 10 x 14 holes

Figure 3 Components are soldered together on a small prototype board

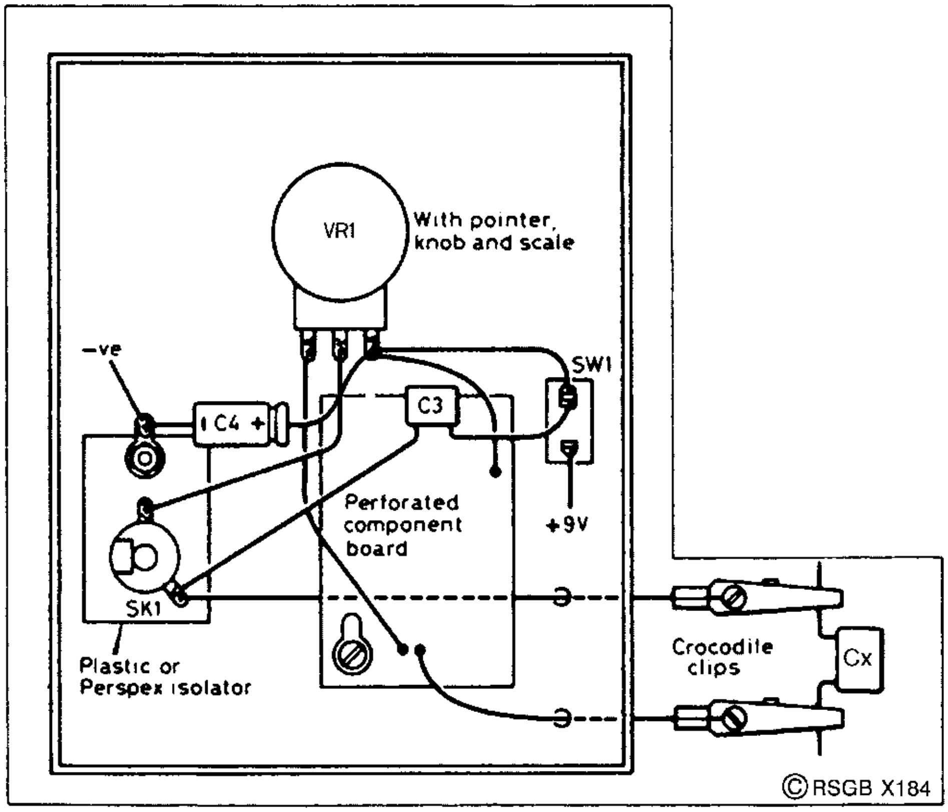
VR1
With pointer, knob and scale
−ve
C4
C3
SW1
+9V
Perforated component board
SK1
Plastic or Perspex isolator
Crocodile clips
Cx
©RSGB X184

Figure 4 Layout of the parts inside an aluminium box. The earpiece socket is insulated from the case

mounting the board inside a metal box, a long screw is used with extra nuts, to earth the board to the case while providing a stand-off, thus preventing any unwanted short-circuits between the board connections and the case. Be aware that a *crystal* earpiece must be used; low-impedance headphones of the Walkman type are not satisfactory. Solder two flying leads as shown in **Figure 4**, about 15 cm long, terminated in small crocodile clips for attaching to the unknown capacitor.

Calibration

After checking the circuit carefully, attach the battery and switch on; a buzzing noise should be heard in the earpiece. This is the first sign that everything should be OK. If there is no buzz, switch off and recheck the connections.

You will now need a range of close-tolerance (1%) silver-mica capacitors covering the range 10 to 1000 picofarads (pF). Arrange the capacitors in ascending order and connect them to your bridge in sequence. After having prepared a neat piece of card or paper mounted behind the knob on the front panel, mark the dial at the positions of the nulls for all the capacitors. You have now calibrated your capacitance bridge. If you are more likely to want to measure larger capacitors, replace C3 by a 1 nanofarad (nF) capacitor, and the bridge will measure up to 10 nF approximately.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	10 kilohms (kΩ)
R2, R3	100 kilohms (kΩ)
RV1	10 kilohms (kΩ) linear

Capacitors

C1, C2	10 nanofarads (nF) ceramic
C3	100 picofarads (pF) silver mica or polystyrene
C4	47 microfarads (μF) 16 V electrolytic

Semiconductors

TR1, TR2	BC108 npn

Additional items

SW1	SPST on/off switch
	Battery connector, PP3 type
	Earpiece, crystal type
SK1	3.5 mm jack socket for earpiece
	Battery, 9 V PP3
	Matrix board, approx. 10 holes by 14 holes
	Aluminium case, approx. 10 × 8 × 5 cm

33 A simple short-wave receiver – Part 2

Introduction

In Part 1, the design of this receiver was discussed in some detail. Now we are going to put it all together and see how it works. The receiver is laid out on a printed-circuit board (PCB) or on a matrix board.

Construction

The layout of the components is shown in **Figure 1**. Identify each part separately, insert it into the holes in the board and solder carefully. Long leads may be cropped before or after soldering, depending on your skill and preference. All electrolytic capacitors, T1, T2 and IC1, must be connected correctly. The front-panel controls are connected to the PCB terminals shown in Figure 1.

The layout of the controls and the placing of the board inside the case are matters of personal preference. The size of the prototype front panel is shown in **Figure 2**. The prototype had a slow-motion drive fitted to VC2, the main tuning capacitor. This required the capacitor to be fitted on its own small panel. The bandspread control does not need any form of

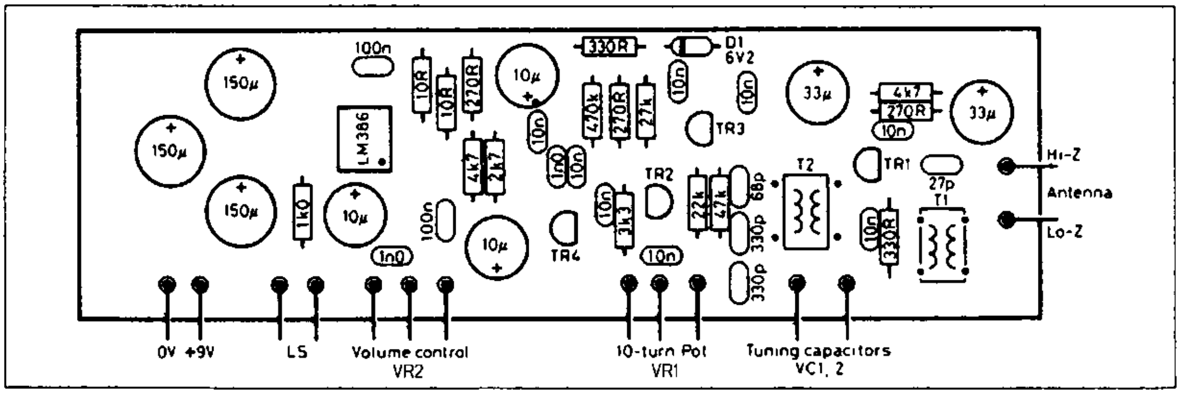
150µ
150µ
150µ
100n
LM386
1k0
10µ
100n
1n0
10R
10R
270R
4k7
2k7
10µ
10µ
10n
1n0
10n
TR4
330R
D1
6V2
470k
270R
27k
10n
10n
TR3
10n
3k3
TR2
10n
22k
47k
68p
330p
330p
T2
33µ
4k7
270R
10n
33µ
TR1
27p
T1
10n
330R
Hi-Z
Antenna
Lo-Z
0V
+9V
LS
Volume control
VR2
10-turn Pot
VR1
Tuning capacitors
VC1, 2

Figure 1 Layout of the receiver circuit board

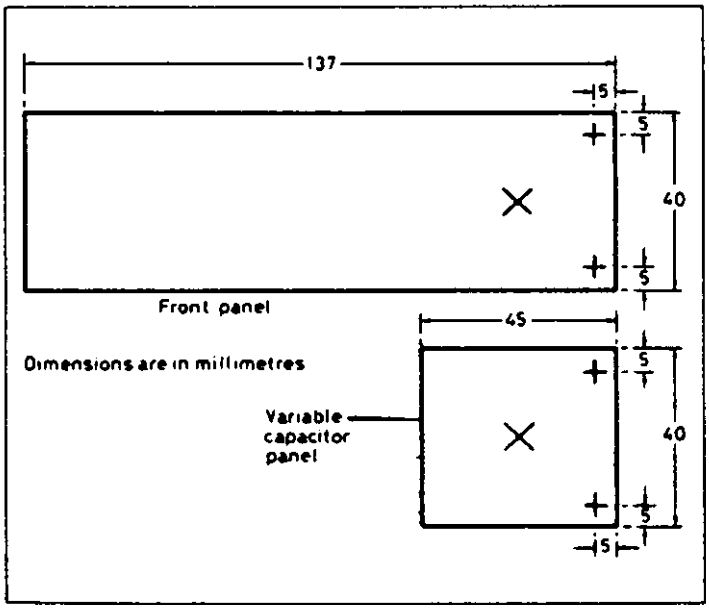

Figure 2 The case can be metal or made from PCB material

slow-motion drive. A tuning dial may be fitted behind the tuning knob on VC2 if required. The volume control, VR2, has its own on/off switch for the battery, which can be mounted behind the back panel, on which are mounted the aerial input sockets and the headphone/speaker socket. Check your soldering carefully, together with all the wiring between the board, the potentiometers and variable capacitors. When you are confident that everything is perfect, connect the battery and switch on.

The first tests

If you have a good aerial and are using an aerial tuning unit (ATU), use the low-impedance input. Excellent results are possible, though, with about 3 metres of wire connected to the high-impedance input. Set the volume control to give a gentle hiss in the headphones or speaker. Advance the reaction control to give a definite hiss. As you tune in an AM station, the hiss will change to a whistle; back off the reaction until the oscillation *just* stops. The receiver is now correctly set for AM reception.

On the amateur bands, the stations will be SSB or CW, and the reaction needs to be set *just above* the oscillation point, and will need slight adjustment from time to time for different qualities of signal. Juggling with the volume, reaction and tuning is part of the pleasure of using regenerative receivers!

In action

Practice is needed for best results. The regenerative receiver is renowned for its versatility in being able to be set up *exactly* for all types and strengths of signal. The basic receiver tunes from 6.5 MHz to 11 MHz approximately; this includes two amateur bands at 7.0 MHz and 10.1 MHz and two broadcast bands.

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1, R10	330 ohms (Ω)
R2, R5, R11	270 ohms (Ω)
R3	22 kilohms (kΩ)
R4	47 kilohms (kΩ)
R6	3.3 kilohms (kΩ)
R7, R12	4.7 kilohms (kΩ)
R8	27 kilohms (kΩ)
R9	470 kilohms (kΩ)
R13	2.7 kilohms (kΩ)
R14, R15	10 ohms (Ω)
R16	1 kilohm (kΩ)
VR1	10 kilohm (kΩ) linear 10-turn potentiometer
VR2	10 kilohm (kΩ) log potentiometer
Capacitors	
C1	27 picofarads (pF)
C2, C18, C22	100 nanofarads (nF)
C3, C4, C6, C10, C13, C14	10 nanofarads (nF)
C5, C11	33 microfarads (μF) 16 V electrolytic
C7	68 picofarads (pF)
C8, C9	330 picofarads (pF)
C15	1 nanofarad (nF)
C16, C17, C20	10 microfarads (μF) 16 V electrolytic
C21, C23	150 microfarads (μF) 16 V electrolytic
VC1	10 picofarads (pF) variable
VC2	200 picofarads (pF) variable
Semiconductors	
TR1, TR3	2N3819
TR2, TR4	BC182
IC1	LM386
D1	6.2 V 0.5 W Zener

Inductors

T1	3 turns primary, 15 turns secondary, wound on 2-hole ferrite bead, with 28 SWG wire
T2	2 turns primary, 17 turns secondary, wound on a T68-2 toroidal former, with 28 SWG wire

Additional items

- Printed-circuit board
- Battery connector
- PP3 battery
- 8-pin DIL socket
- 8 ohm speaker or headphones (Walkman type)
- Case to suit

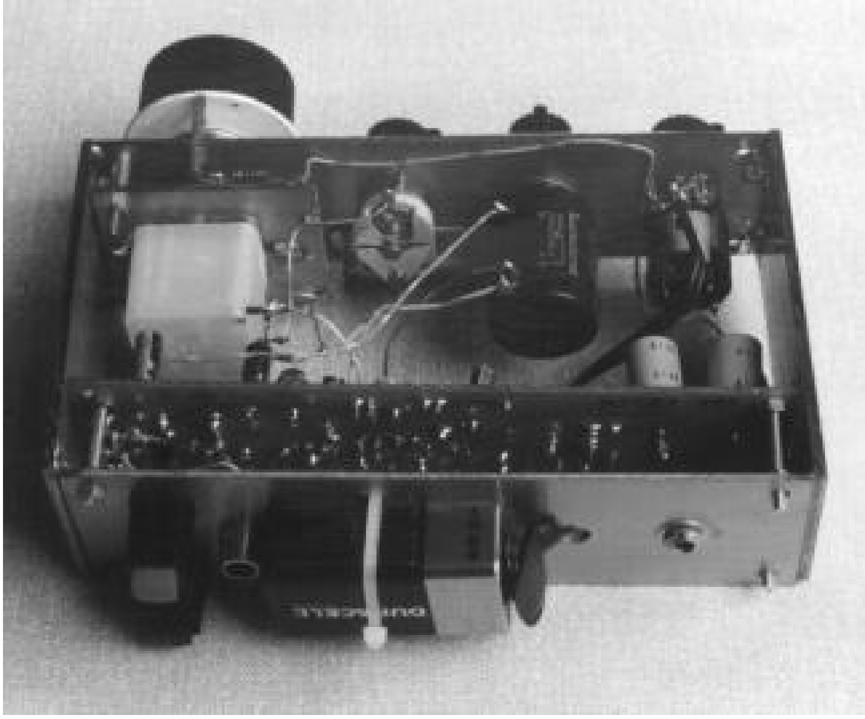

34 A basic continuity tester

Introduction

This little device can be built on a plug-in breadboard, and is ideal for testing fuses and cables, as well as doubling as a signal source for testing amplifiers, etc. It can even be used for testing npn transistors by replacing either of the transistors in the circuit, and seeing if the circuit still works!

Simple and quick to build

Using a plug-in breadboard, this circuit is so simple it could almost be built when you need it, and then dismantled again! If you want the circuit to make a different sound, the components to change are R2, R3, C1 and C2. Always make sure that R2 = R3 and C1 = C2, or the sound may be excessively 'edgy' and lacking in volume.

Basically, the circuit is an oscillator which drives a little loudspeaker directly. The fuse or cable you are testing is connected between the crocodile clips. If there is a current path between the two clips, the current also flows through the circuit, thus operating the oscillator and producing a sound from the loudspeaker.

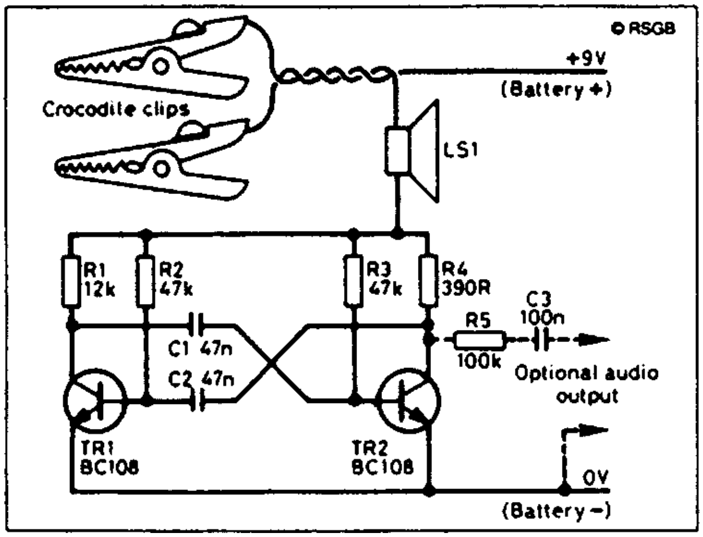
© RSGB
+9V
(Battery +)
Crocodile clips
LS1
R1
12k
R2
47k
R3
47k
R4
390R
C3
100n
R5
100k
C1 47n
C2 47n
Optional audio output
TR1
BC108
TR2
BC108
0V
(Battery −)

Figure 1 The circuit diagram of the continuity tester

If you want to use the circuit to produce a signal to test an amplifier, for instance, connect the croc clip which comes directly from the battery to the junction of R1, R2, R3 and R4. This supplies current to the circuit while, at the same time, bypassing the loudspeaker. Use the two components R5 and C3 between the oscillator and your amplifier. The loudspeaker is bypassed because you will want to listen to the output from your amplifier, *not* from the oscillator!

Remember, NEVER test any equipment which is still connected to the mains electricity supply. Avoid testing anything which is switched on and has its own power supply. You may damage both your tester ***and*** the circuit you are 'testing'!

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	12 kilohms (kΩ)
R2, R3	47 kilohms (kΩ)
R4	390 ohms (Ω)
R5	100 kilohms (kΩ) optional
Capacitors	
C1, C2	47 nanofarads (nF) ceramic
C3	100 nanofarads (nF), 0.1 microfarads (μF) ceramic optional
Semiconductors	
TR1, TR2	BC108 npn
Additional items	
LS1	Loudspeaker (8 Ω)
	2 crocodile clips
	Prototype circuit board
	PP3 battery connector
	PP3 battery

35 A charger for NiCad batteries

Introduction

The cost of replacing dry batteries can be alleviated to a great extent by the use of rechargeable batteries such as those containing nickel and cadmium (called *NiCads*). They are more expensive than the batteries they replace, but they can be charged hundreds of times and thus prove cheaper in the long run.

NiCads produce only 1.2 V per cell, compared with the 1.5 V of ordinary cells, so before you rush out and buy lots of these, please make sure that the equipment on which you plan to use them will work on the reduced voltage! For example, if you are using four cells to give you a 6 V supply, NiCads will give you only 4.8 V, which is quite a reduction. Using six cells to replace a 9 V battery will give you only 7.2 V. Not all equipment is happy with these reductions!

However, we all use them when we can, and they save substantial amounts of money. Here is the circuit of a charger to keep them in prime condition.

Charging NiCads – the ampère-hour

NiCads require charging at constant current, which means that connecting one across a normal power supply (constant *voltage*) is useless and can destroy it. They need pampering to the extent of needing a long charge (around 16 hours) at a rate dependent upon the *capacity* of the battery. By the capacity of a battery, we mean how much energy it can store. You will probably know that energy is measured in joules. For the purposes of storing energy in batteries, the joule is *not* the ideal unit, so we use one that is! This unit is the ampère-hour (Ah), and must be interpreted with some realism. For example, if the battery is rated at 2 Ah, it will deliver a current of 0.5 amp for 4 hours, or 0.25 amp for 8 hours. Provided the current is not too high, the product of the current (in amps) and the time for which it will flow (in hours) before the battery is flat will always be around 2 Ah. A workable ‘rule of thumb’ for calculating the charging current is that its value should be around one-tenth of the numerical value of the capacity; so, for our 2 Ah battery, a charging current of around 200 mA (2 ÷ 10 = 0.2 A or 200 mA) would be used.

Constant voltage to constant current

Many integrated circuit (IC) chips are available for use as voltage regulators, i.e. they supply a constant voltage. Most of these can be persuaded to become constant current supplies with one external resistor!

The voltage regulator IC usually has only three connections – 'input', 'output' and 'common'. It is designed (in the case of the LM7805) to produce a constant 5 V output between the 'output' and the 'common' connections, at currents of up to 1 A. If a resistor is connected between these, the IC will maintain 5 V across it. If you look at **Figure 1**, you will see the circuit performing this conversion.

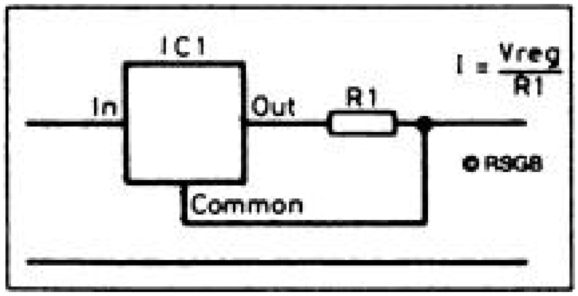

Figure 1 Voltage regulator arranged to produce a constant current

For the previous example, we derived a charging current of 200 mA, so we now need to calculate the value of resistor that will produce this current. Using the equation which is derived from Ohm's law:

$$I = \frac{V}{R}, \text{ from which } R = \frac{V}{I},$$

where R = R1, the resistance in ohms that we are calculating,
V is the voltage across R1 (5 V), and
I is the current flowing (200 mA).

So,

$$\text{R1} = \frac{5}{0.2} = 25\,\Omega.$$

25 ohms is not a 'common' or 'preferred' resistor value, so we must choose the next *largest* value, which is 27 Ω. This reduces the current, but only slightly – it is now 185 mA. When calculating resistor values in power supply circuits, we must *always* check the power that they dissipate and make sure we specify and fit a suitable resistor.

Power (in watts) is the product of the voltage across and the current through a device, so in this case it is given by:

$$\text{Power} = V \times I = 5 \times 0.185 = 0.925\,\text{W}.$$

Rather than use a 1 watt resistor operating very near its limit, it is safer to use a 2 watt resistor operating well within its limits.

Looking again at Figure 1, we now have a constant-current source producing 185 mA, when R1 is a 27 Ω, 2 W resistor. For use with NiCads requiring charging currents other than 200 mA, you will need to repeat the two equations above, using a new value for *I*.

The full circuit and its assembly

This is shown in **Figure 2**, and can be broken down into two parts. The first is the AC to DC conversion produced by the mains transformer, T1, the bridge rectifier, BR1, and the smoothing capacitor, C1. The second is the constant-current section already discussed.

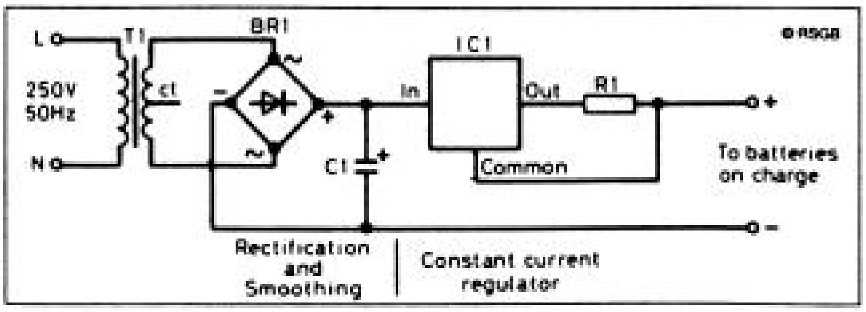

Figure 2 The complete circuit including mains components

The prototype was assembled on matrix board measuring 18 holes by 12 strips, although, as **Figure 3** shows, this is much larger than is strictly necessary. No strip cutting is needed, but make sure that IC1 and C1 are inserted correctly.

Warning! Before you attempt to wire up the transformer and the bridge rectifier, be aware that you will eventually be connecting the circuit to the mains supply, so there are three possibilities for you: (1) get a qualified friend to supervise your completion and testing of the circuit; (2) get your qualified friend to complete and test the circuit for you; (3) replace the transformer and bridge rectifier with a mains adapter.

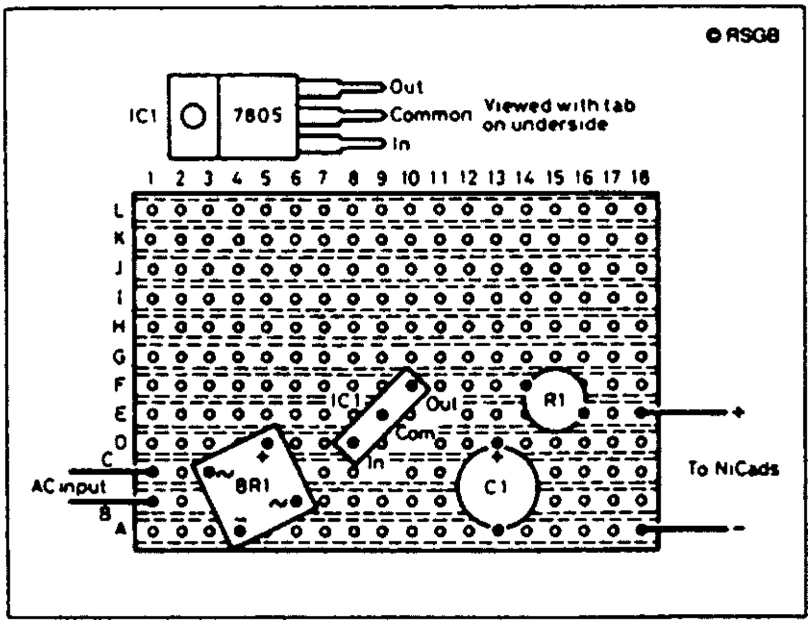

Figure 3 The charger can be built on Veroboard

If you decide to use the mains adapter, its output is connected directly across C1, because T1 and BR1 are now no longer needed. Make sure the polarity (positive and negative) is correct and that the adapter output is set to 12 V.

A quick test

If you have built the mains version, make sure all connections are correct, and that there are no soldered joints which will touch other parts of the circuit. The box must be securely closed before tests begin. The RSGB cannot be held responsible for damage to equipment or batteries! The version using the mains adapter need not be closed during tests.

Switch on. With nothing connected to the output, the unit should run cold. If this is not the case, switch off and recheck your circuit. If all is well, switch on again and connect a DC multimeter (on the current range) across the output. It should indicate only a slight difference from the calculated value of 185 mA. You can now charge your NiCads!

Other charging currents can be set by having different values for R1, perhaps selectable by a rotary switch. Remember to make sure that the values of both resistance *and* power dissipation are correct, and don't exceed the 250 mA rating of the transformer (or the 1 amp rating of the IC if you are using a bigger transformer).

Parts list

		Maplin code
Resistor		
R1	As required – see text	
Capacitor		
C1	1000 microfarads (μF) 35 V electrolytic	FF18U
Semiconductors		
IC1	LM7805 5 V, 1 A regulator	CR14Q
BR1	W005 50 V, 1 A full-wave rectifier	QL37S
Transformer		
T1	9–0–9 V, 250 mA sub-min. transformer	YN15R
Additional items		
Case		FG41U
Veroboard, 18 holes by 12 strips		
Plug to suit NiCads		
Double-sided sticky tape as required		
Insulated wire for battery connections		

36 An 80 metre crystal-controlled CW transmitter

Introduction

A simple transmitter is ideally suited to anyone venturing into our marvellous hobby for the first time. If you are put off by the complexities and prices of the 'black boxes' then this is the transmitter you've been looking for!

The circuit

The complete circuit is shown in **Figure 1**. It is a single-transistor crystal oscillator which is *keyed* (switched on and off) by the Morse key in the +12 V supply rail. The circuit first appeared in *QST* (the American equivalent of *RadCom*) in 1982, and has since appeared in a modified form in the *ARRL Handbook*. The circuit can produce about 2 watts on 80 m, and can be constructed on a piece of plain matrix board; the prototype board measured 7 cm by 3 cm, and its layout is shown in **Figure 2**.

Note that a bare copper wire runs along the bottom edge of the board to act as an earth wire for the relevant components. The only 'difficult' part of the construction is the winding and wiring of T1. The main winding is 38 turns of 26 SWG enamelled copper wire, and there are two link windings of four turns each. Make sure that all windings are wound the correct way round the toroid – Figure 1 shows this and should be studied carefully. If the windings do not have the correct *sense* (i.e. a clockwise coil has been wound anticlockwise, or vice versa), or have been connected incorrectly to the rest of the circuit, the oscillator will not work!

When putting components on the board, wire in the crystal socket *without the crystal in it*. Crystals do not like to be subjected to the horrors of a soldering iron, so keep your crystal to one side during the construction process!

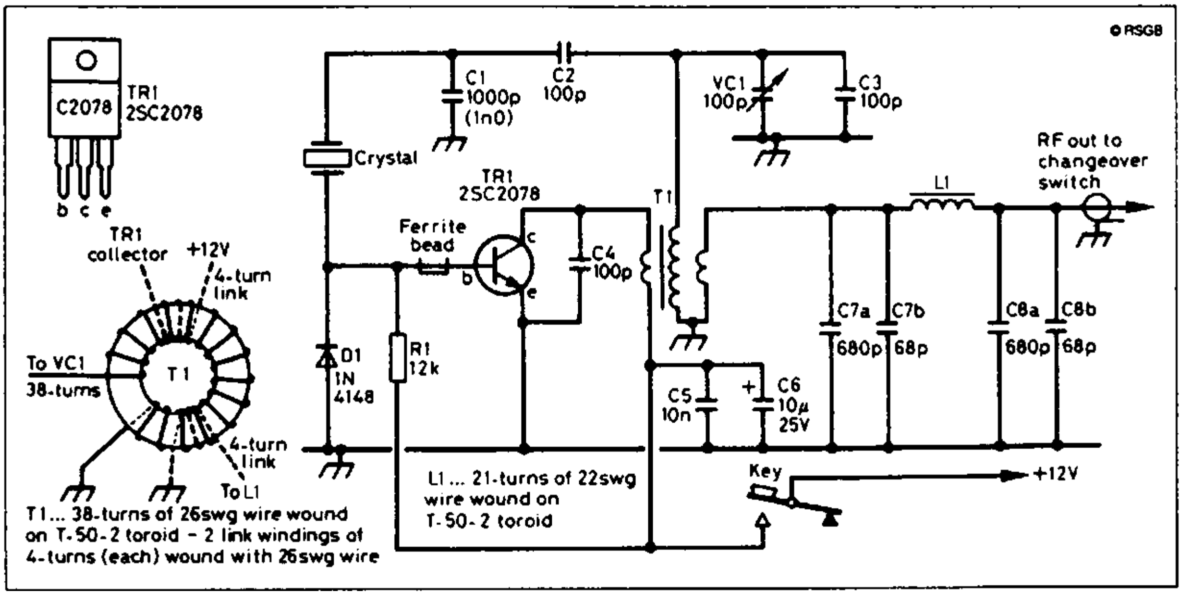
© RSGB
C2078
TR1
2SC2078
b c e
TR1
collector
+12V
4-turn
link
To VC1
38-turns
T1
4-turn
link
To L1
T1... 38-turns of 26swg wire wound
on T-50-2 toroid – 2 link windings of
4-turns (each) wound with 26swg wire
C1
1000p
(1n0)
C2
100p
Crystal
TR1
2SC2078
Ferrite
bead
b
c
e
C4
100p
D1
1N
4148
R1
12k
VC1
100p
C3
100p
T1
L1
RF out to
changeover
switch
C5
10n
C6
10µ
25V
C7a
680p
C7b
68p
C8a
680p
C8b
68p
Key
+12V
L1 ... 21-turns of 22swg
wire wound on
T-50-2 toroid

Figure 1 The simplest form of transmitter is a keyed crystal oscillator. Note that L1, C7 and C8 make up a *low-pass filter* which reduces unwanted *harmonics* (outputs at the transmitter frequency multiplied by 2, 3, 4, etc.)

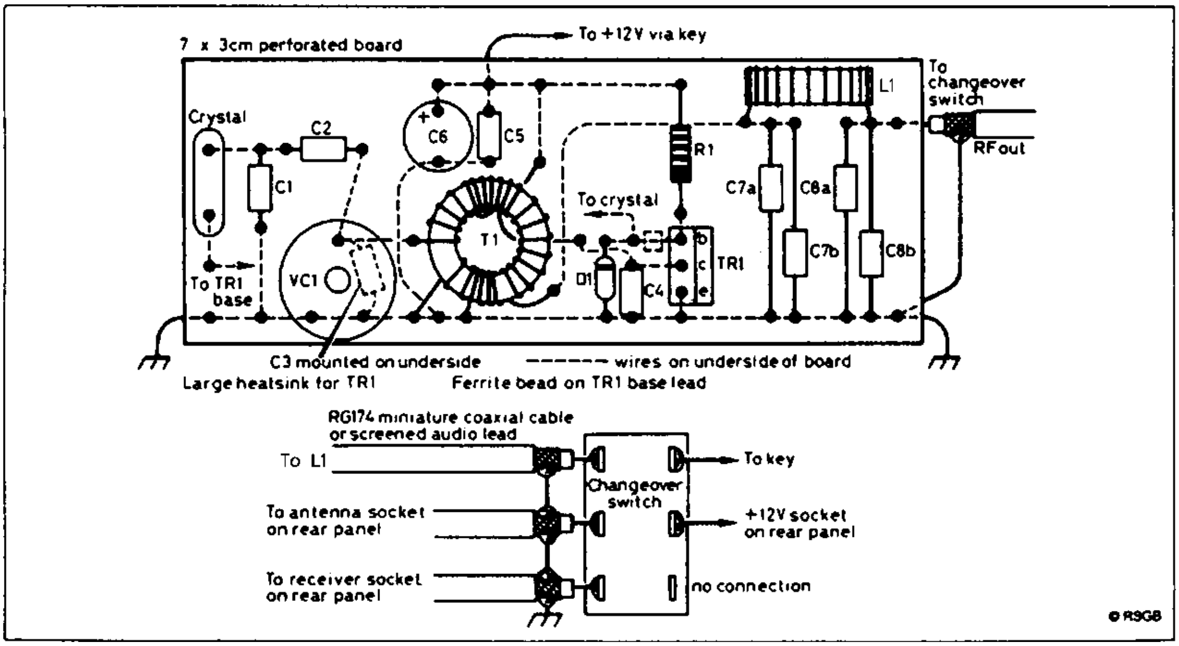
7 x 3cm perforated board
To +12V via key
To changeover switch
L1
Crystal
C2
C6
C5
R1
C1
C7a
C8a
To crystal
T1
VC1
b
c
e
TR1
C7b
C8b
D1
C4
To TR1 base
RFout
C3 mounted on underside
wires on underside of board
Large heatsink for TR1
Ferrite bead on TR1 base lead
RG174 miniature coaxial cable or screened audio lead
To L1
To key
Changeover switch
To antenna socket on rear panel
+12V socket on rear panel
To receiver socket on rear panel
no connection
© RSGB

Figure 2 Component layout is straightforward on a 'matrix board' (it has holes but no copper strips). The dotted lines are the connecting wires. The lower part shows the connections for the transmit/receive switch

The output filter, which comprises C7, C8 and L1, is a *low-pass* filter, which helps reduce any *harmonics* present in your signal. Harmonics are integral multiples of your transmitter frequency, so if you are transmitting on a frequency, *f*, harmonics will be present at frequencies 2*f*, 3*f*, 4*f*, . . . and so on. L2 is another inductance using 22 SWG enamelled copper on a ferrite toroid. The changeover switch is external to the transmitter board, and is used to switch your aerial between the transmitter and the receiver; its wiring is shown in Figure 2.

Use a dummy load

A dummy load enables you to test your circuit without actually transmitting a signal. If you haven't such a thing already, it is easy to construct one to use with this transmitter. Don't use it for transmitters of more than 2 watts output, though. Use two 100 ohm, 1 watt resistors, connected in parallel across the end of a short piece of coaxial cable, terminated in a BNC, PL259 or N-type free plug, as shown in **Figure 3**. Plug this into the aerial socket on your transmitter, plug in your crystal and connect the transmitter to a 12 V supply. Have another receiver switched on and tuned to the crystal frequency. Although the radiation from your dummy load is minimal, it will

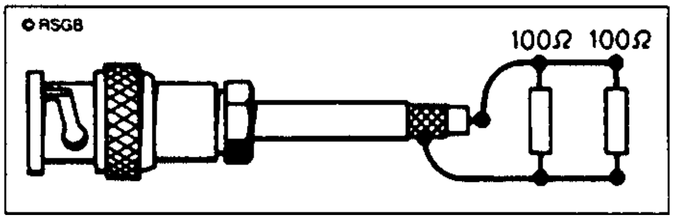

Figure 3 A simple 2 W dummy load can be made from two 100 Ω (ohm) resistors in parallel. The plug should match the socket on your transmitter

be enough to be picked up by a receiver in the same room. Send dashes with the Morse key, and adjust VC1 until the received note is *clean*. It should not sound rough, or have a *chirp* (change its frequency during a dash or dot). Avoid tuning for maximum power; this is seldom the correct setting!

You will need to put your completed transmitter in a metal box, using sockets for the power supply, aerial, receiver and Morse key. The sockets can be chosen to match your existing equipment.

Figures 4–7 are taken from the RSGB book *Practical Antennas for Novices*, and may give you some ideas on the type of aerial to be used with your transmitter.

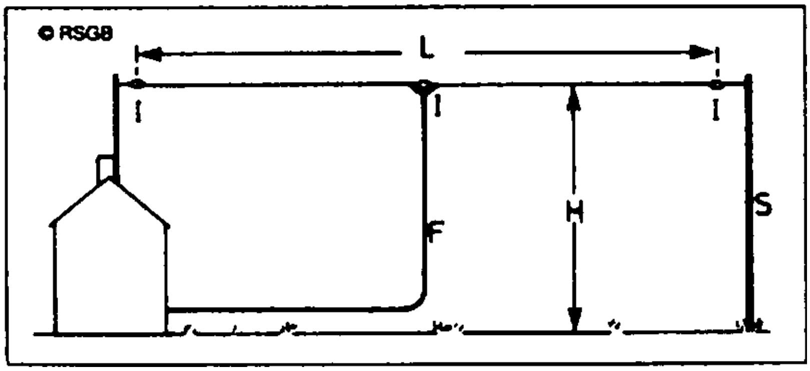

Figure 4 A simple dipole can be very effective. For the 3.5 MHz band, length L is 40 metres and height H should be as large as possible. The far support S can be a tree, pole or building. Insulators I may be home made from strong plastic and the feeder F should be 50 Ω (ohm) coax cable

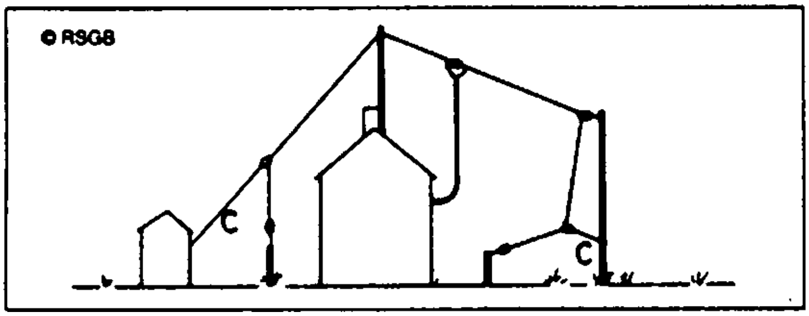

Figure 5 Your signal is radiated mostly from the centre of the dipole so the ends can droop or even be bent but the length may need shortening by a few centimetres because the ground and the bends will detune the dipole. Cords C are best made from strong plastic rope from a sailing or camping shop

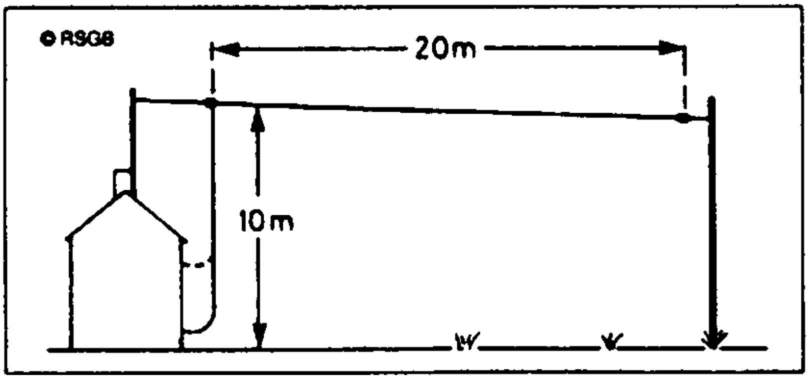
© RSGB
20m
10m

Figure 6 An 'inverted-L' takes up less space than a dipole and doesn't need coax cable. Like the dipole, the end can droop or be bent to save space as in this case most of the radiation comes from the area around the top of the vertical part

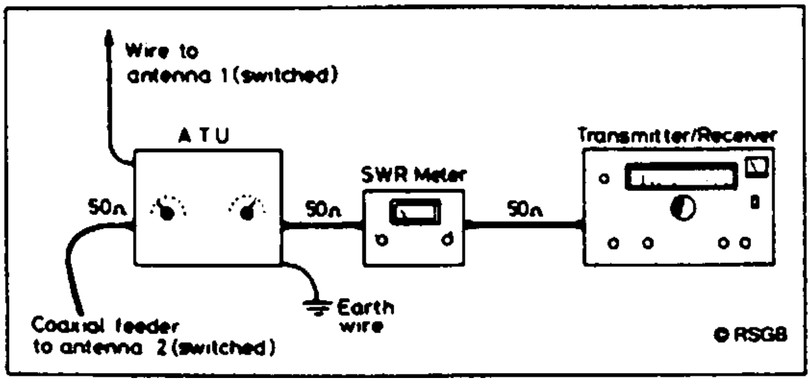
Wire to
antenna 1 (switched)
ATU
50Ω
50Ω
SWR Meter
50Ω
Transmitter/Receiver
Earth
wire
Coaxial feeder
to antenna 2 (switched)
© RSGB

Figure 7 Almost any length of wire more than 10 m or so long will work (though it will work better the longer and higher it is) but an Aerial System Tuning Unit (ASTU or ATU) will be needed

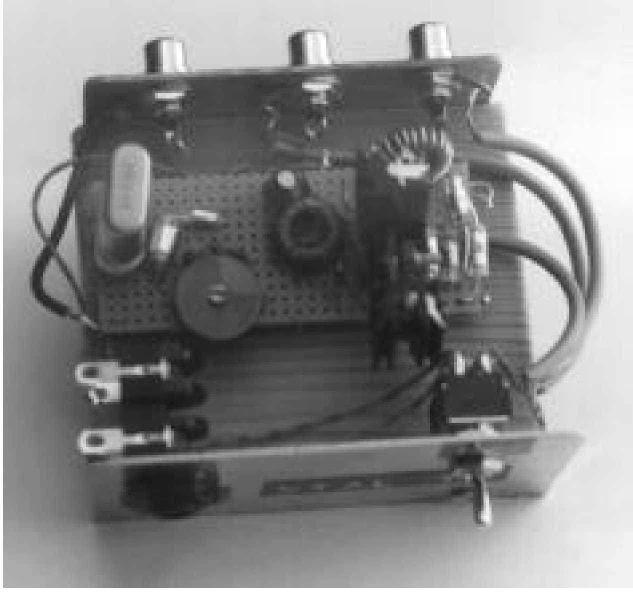

Parts list

Resistor

R1	12 kilohms (kΩ), 0.25 watt, 5% tolerance

Capacitors

C1	1000 picofarads (pF) polystyrene
C2	100 picofarads (pF) polystyrene
C7a, C8a	680 picofarads (pF) polystyrene
C7b, C8b	68 picofarads (pF) polystyrene
C3, C4	100 picofarads (pF) ceramic
C5	10 nanofarads (nF) polyester
C6	10 microfarads (μF) electrolytic 25 V
VC1	100 picofarads (pF) trimmer

Semiconductors

D1	1N4148
TR1	2SC2078 (see sources list)

Inductors

T1	38 turns 26 SWG enamelled copper on T–50–2 toroid, with two link windings of four turns
L1	21 turns 22 SWG enamelled copper on T–50–2 toroid

Additional items

- Ferrite bead
- Crystal (e.g. 3.579 MHz) and holder
- Metal box
- Socket for Morse key.
  This must be totally isolated from the metal of the box, as *both* connections can be at +12 V.
- Sockets for 12 V supply, aerial and receiver
- Switch – DPDT
- Heat sink for TR1
- RG174 miniature coaxial cable for signal leads (see Figure 2)

Component sources

Special components

2SC2078	Cricklewood Electronics Ltd, 40 Cricklewood Broadway, London, NW2 3ET.

37 A solar-powered MW radio

Introduction

What could be more ecologically friendly than a radio powered by the sun's energy? This design is quite standard, and if you have built any of the other medium-wave radios in this series, then this one should present few problems.

The solar panel

The solar panel is to the solar cell as the battery is to the cell; in other words a solar panel is several solar cells connected in series. The solar panel quoted for this radio will generate about 9 V at a current of around 30 mA on a sunny day. The circuit will operate on a supply of around 2 V, so bright sunshine is *not* necessary for satisfactory operation. The volume will be less, of course.

The circuit

The radio uses the ZN415E integrated circuit (IC), connected as shown in **Figure 1**, the complete circuit diagram. The signal is tuned in by the combination of L1 and VC1. L1 is made by winding about 35 turns of 24 SWG enamelled copper wire on a ferrite rod. A standard ferrite rod of 10 cm length and 1 cm diameter is used.

Signals selected by the tuned circuit are passed to IC1, which amplifies the signals and removes the audio component, which is then amplified further by IC2 for driving a small loudspeaker. The removal of the audio component is the process we call *detection*. In addition to this, IC1 provides *automatic gain control* (AGC), which helps to keep the audio signal constant, even when the incoming RF signal may vary due to fading.

The prototype board

Veroboard (also known as matrix board or stripboard) is ideal for the construction of the radio. The layout is shown in **Figure 2**. The board size is 11 strips by 30 holes (**please note that there is no row 'I', so take care with your counting!**). Using a 3 mm (1⁄8 inch) twist drill rotated between

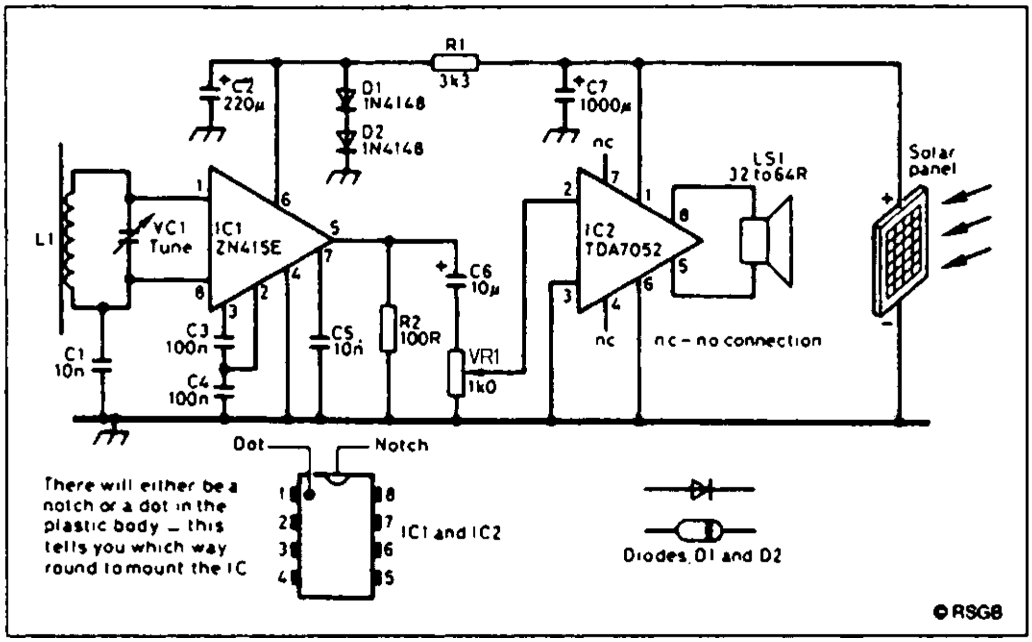
R1
3k3
C2
220µ
D1
1N4148
D2
1N4148
C7
1000µ
nc
LS1
32 to 64R
Solar panel
L1
VC1
Tune
IC1
ZN415E
IC2
TDA7052
C6
10µ
C3
100n
C4
100n
C1
10n
C5
10n
R2
100R
VR1
1k0
nc
nc – no connection
Dot
Notch
There will either be a notch or a dot in the plastic body – this tells you which way round to mount the IC
IC1 and IC2
Diodes, D1 and D2
© RSGB

Figure 1 The radio uses just two integrated circuits (chips) and operates at any voltage from 2 to 9 V

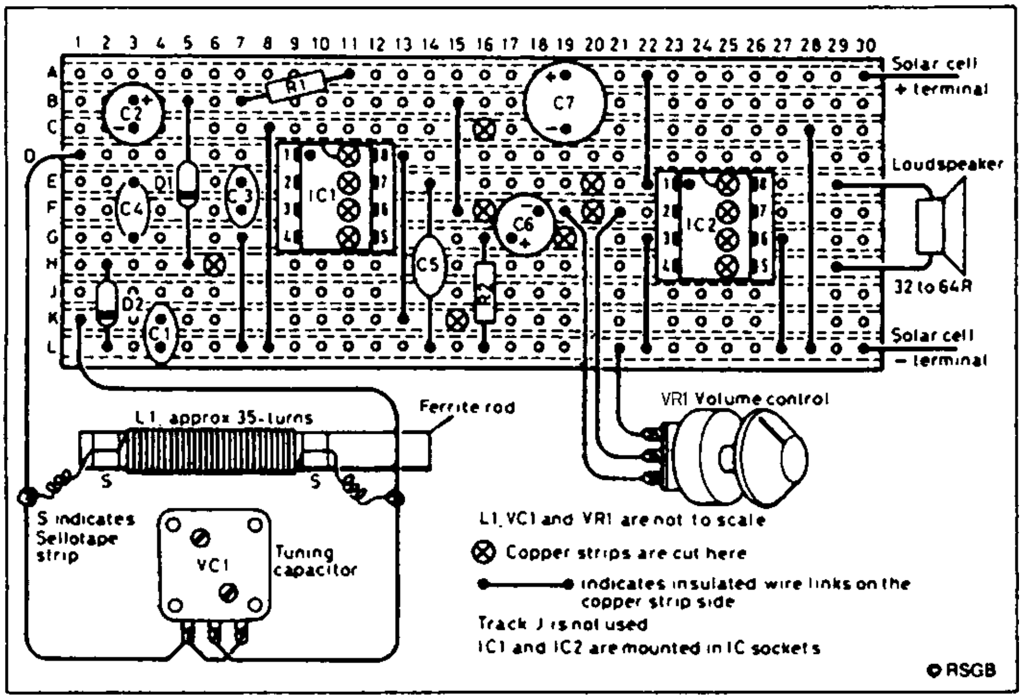
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
A B C D E F G H J K L
C2
R1
C7
D1
C4
C3
IC1
IC2
C5
C6
R2
D2
C1
Solar cell + terminal
Loudspeaker
32 to 64R
Solar cell – terminal
VR1 Volume control
Ferrite rod
L1 approx 35-turns
S
S indicates Sellotape strip
VC1
Tuning capacitor
L1, VC1 and VR1 are not to scale
⊗ Copper strips are cut here
indicates insulated wire links on the copper strip side
Track J is not used
IC1 and IC2 are mounted in IC sockets
© RSGB

Figure 2 Follow the layout carefully, making sure that all connections are neatly soldered

your thumb and forefinger, break the strips at the points marked with a cross inside a circle. Hold the board up to the light to ensure that the tracks *have* been broken and that there is no copper swarf shorting tracks together.

Firstly, solder in the 8-pin DIL sockets for the ICs, making sure that the notches in the sockets are facing upwards, as shown on the diagram. Then, solder in the wire links, resistors, capacitors and diodes, making sure that the electrolytic capacitors (C2, C6 and C7) and the diodes (D1 and D2) are connected the correct way round. Use different colours of wire for the connections to the volume control, VR1, to avoid incorrect connections. Note the wiring of the tuning capacitor (VC1) shown in Figure 2; a two-section type is used, and both sections are wired in parallel to give twice the capacitance of a single section.

There is no on/off switch – just turn the volume down when you are finished using the set! The solar panel can be mounted parallel to the top of the case, or angled to receive the maximum energy from the sun, as shown in the photograph. You could have a battery available as a standby source to use the radio after dark; any battery of between 6 V and 9 V will do. Wire it with a simple changeover switch, so you can switch between solar and battery power. Ask a friend for help with this if you are not sure how to do it.

You may need to adjust the number of turns on L1 to get the best results, but it should be possible to receive at least five stations at good volume with your sun-powered radio!

SOLAR PANEL

Parts list

		Maplin code
Resistors: 0.25 watts, 5% tolerance		
R1	3.3 kilohms (kΩ)	M3K3
R2	100 ohms (Ω)	M100R
VR1	1 kilohm (kΩ) log	FW21X
Capacitors		
C1, C5	10 nanofarads (nF) ceramic	BX00A
C2	220 microfarads (μF) electrolytic (10 V)	FB60Q
C3, C4	100 nanofarads ceramic	YR75S
C6	10 microfarads (μF) electrolytic (10 V)	FB22Y
C7	1000 microfarads (μF) electrolytic (10 V)	FB81C
VC1	140–300 picofarads (pF)	FT78K
Semiconductors		
IC1	ZN415E (or ZN416E)	UR70M
IC2	TDA7052	UK79L
D1, D2	1N4148	QL80B
Solar panel	9 V at 50 mA	RK23A
Additional items		
LS1	32–64 ohm miniature loudspeaker	YT28F
	Ferrite rod	YG20W
	24 SWG enamelled copper wire	BL28F
	Plastic box, approx. 220 × 140 × 70 mm	YN39N
	Veroboard, cut to size	JP47B
	8-pin DIL sockets, 2 required	BL17T
	Knobs, 2 required	FD67X
	Material for speaker grille	
	Connecting wire	

38 A receiver for the 7 MHz amateur band

Introduction

Listening on the 40 metre band (from 7.0 to 7.1 MHz) can be very rewarding – it is a popular haunt for HF Special Event stations, and at night there are signals to be heard from all over Europe. This receiver is designed purely for the 40 m band, and is ideal for those who have built the simpler receivers and are looking for something a little more challenging. The more experienced constructor may prefer to build this on prototype board.

The circuit and its construction

Figure 1 shows the circuit diagram. The receiver will work well with headphones or loudspeaker. Walkman-type headphones and speakers are ideal for use here.

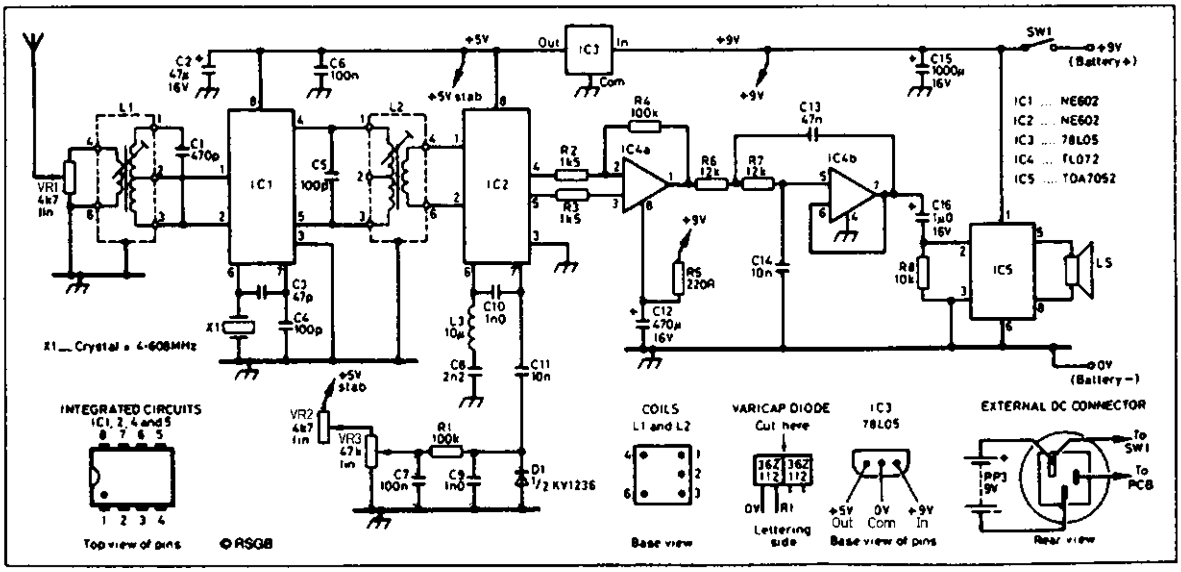
+5V
Out
IC3
In
Com
+9V
SW1
+9V (Battery+)
C2 47µ 16V
C6 100n
+5V stab
C15 1000µ 16V
L1
L2
C1 470p
C5 100p
IC1
IC2
VR1 4k7 lin
R4 100k
C13 47n
IC4a
IC4b
R2 1k5
R3 1k5
R6 12k
R7 12k
+9V
R5 220Ω
C14 10n
C16 1µ0 16V
R8 10k
IC5
LS
C12 470µ 16V
C3 47p
C4 100p
X1
X1 — Crystal = 4·608MHz
C10 1n0
L3 10µ
C8 2n2
C11 10n
0V (Battery −)
IC1 ... NE602
IC2 ... NE602
IC3 ... 78L05
IC4 ... TL072
IC5 ... TDA7052
INTEGRATED CIRCUITS IC1, 2, 4 and 5
8 7 6 5
1 2 3 4
Top view of pins
© RSGB
+5V stab
VR2 4k7 lin
VR3 47k lin
R1 100k
C7 100n
C9 1n0
D1 1/2 KV1236
COILS L1 and L2
Base view
VARICAP DIODE
Cut here
362 112
0V
R1
Lettering side
IC3 78L05
+5V Out
0V Com
+9V In
Base view of pins
EXTERNAL DC CONNECTOR
To SW1
To PCB
PP3 9V
Rear view

Figure 1 The receiver gives good performance on the 7 MHz amateur band as well as being simple to construct and align

Signals arriving at the aerial are coupled into IC1 via gain control VR1, which also functions as the on/off switch. Tuning is provided by varying the voltage on the *varactor diode* (or varicap), D1. VR3 is the main tuning control, and VR2 is the *bandspread* (fine tuning) control. The varactor diode is supplied as a dual device, which must be cut down the middle *carefully* with a sharp knife; with the lettering upwards, the ground lead (0 V) is on the left-hand side, as Figure 1 illustrates.

Solder in the IC sockets first, followed by the coils. After this come the links, resistors, capacitors and varactor diode. Ensure that IC3, the voltage regulator, is wired correctly, and check the polarity of the electrolytics. The crystal, X1, is very fragile, so take extra care with it. The wiring of the three controls is shown in **Figure 2**.

Before putting the ICs in their sockets, connect up the battery and check the following voltages with the negative voltmeter lead connected to the negative terminal of the battery:

Pin 8	IC1	5 V
Pin 8	IC2	5 V
Pin 8	IC4	9 V
Pin 1	IC5	9 V

When all these have been found to be correct, switch off and put the ICs carefully into their sockets. Use wire of different-coloured insulation to wire up the front-panel controls.

The case can be a small plastic box of size 22 cm by 15 cm by 8 cm, with three 10.5 mm holes drilled in the front and two 8 mm holes in the side for the aerial and earth connections. On one side are a 6 mm hole for the speaker socket and an 11 mm hole for the optional external power supply.

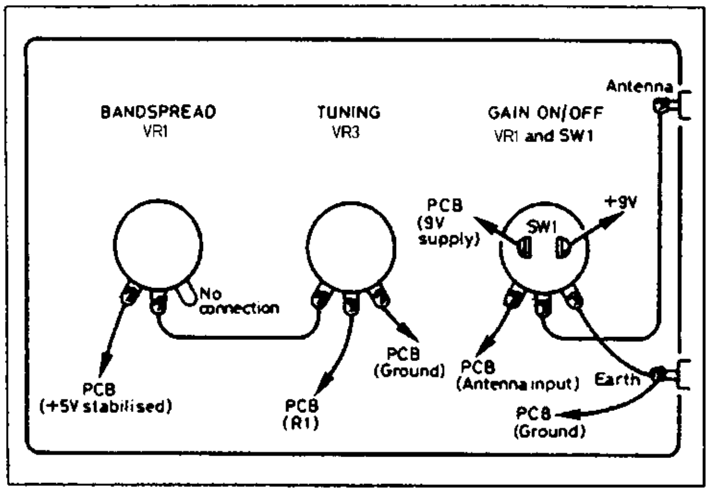

Figure 2 Rear view of the variable resistors. Check the connections carefully to make sure the wires are fitted correctly

Testing and tuning

The aerial for the receiver should be between 30 and 70 feet of wire, mounted as high as you can make it, away from trees and buildings if possible. Connect the battery and switch on. Adjust L1 and L2 for the best results. Tune slowly with VR3; you should find CW stations at the lower end of the band (anticlockwise) and SSB stations at the upper end (clockwise). You may find that it is easiest to make these adjustments *before* mounting the board in the case with double-sided sticky tape or pads. If you are planning to use an external DC supply, make sure it is a safety approved stabilised 9 V type, and **disconnect the battery before you use such a supply!**

If you suspect that the tuning doesn't quite cover the lower CW end of the band, try increasing C9 to 1200 pF. If it is the upper SSB end which is missing, decrease C9 to 820 pF.

It is always advisable to use an aerial tuning unit (ATU) between your aerial and the receiver. A suitable design of ATU is included as a project in this book.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R4	100 kilohms (kΩ)
R2, R3	1.5 kilohms (kΩ)
R5	220 ohms (Ω)
R6, R7	12 kilohms (kΩ)
R8	10 kilohms (kΩ)
VR1	4.7 kilohms (kΩ) linear, with SPST switch
VR2	4.7 kilohms (kΩ) linear
VR3	47 kilohms (kΩ) linear

Capacitors: all rated 16 V or more

C1	470 picofarads (pF) polystyrene 5%
C2	47 microfarads (μF) electrolytic
C3	47 picofarads (pF) polystyrene 5%
C4, C5	100 picofarads (pF) polystyrene 5%
C6, C7	100 nanofarads (nF) ceramic
C8	2.2 nanofarads (nF) polystyrene 5%
C9, C10	1 nanofarad (nF) polystyrene 5%
C11, C14	10 nanofarads (nF) ceramic
C12	470 microfarads (μF) electrolytic
C13	47 nanofarads (nF) ceramic
C15	1000 microfarads (μF) electrolytic
C16	1 microfarad (μF) electrolytic

Inductors

L1	Toko KANK3335R
L2	Toko KANK3333R
L3	10 μH 5%, e.g. Toko 283AS–100

Semiconductors

IC1, IC2	NE602 or NE602A
IC3	78L05 5 V, 100 mA
IC4	TL072
IC5	TDA7052

Additional items

D1	Toko KV1236 cut into two sections (one half used)
X1	4.608 MHz (available from Cirkit)
	3 × silver knobs, one with pointer
	Plastic case approx. 22 × 15 × 8 cm
	Speaker 8–32 Ω, or headphones
	4 × 8-pin DIL sockets for IC1, IC2, IC4, IC5
	2 × 4 mm sockets (red and black) for aerial and earth
	3.5 mm chassis-mounting jack socket for speaker
	DC power socket for external supply (if required)
	Prototype board

Kits

A complete kit is available from JAB Electronic Components.

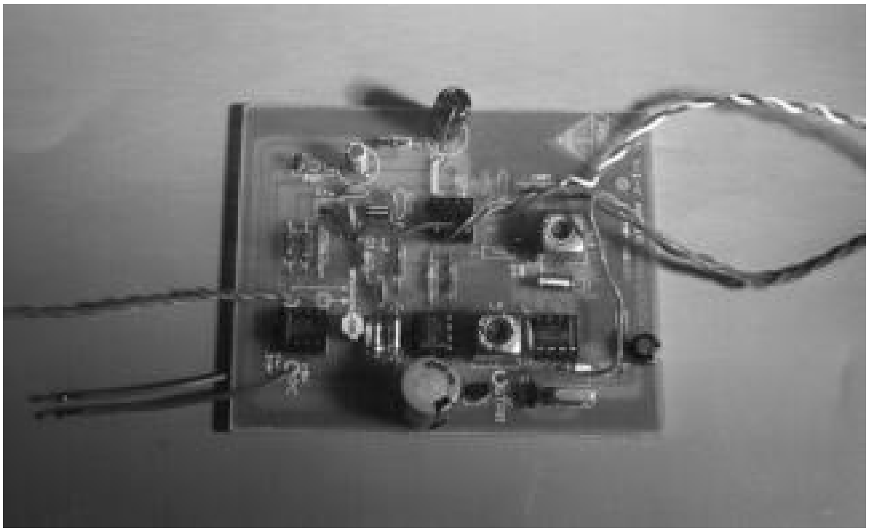

39 Diodes for protection

Introduction

Many semiconductor devices can be destroyed in an instant if their supply is reversed. With the use of batteries as power sources, it is quite common for the battery to be connected ‘the wrong way round’, and the scope for damaging equipment is significant and very real. Diodes can protect equipment in several ways, and you may do worse than to consider one of these approaches to protect your next expensive project.

Choose wisely!

All the diode circuits given here are so simple as to invite calamity. The circuits are not foolproof but, with a little care, will work first time. Remember that a semiconductor diode has a forward voltage drop of between 0.5 V and 0.7 V, depending on its type and the current flowing. This will be mentioned later.

The first thing you need to do is to insert a good multimeter *in series* with the circuit you want to protect; set the range to Amps DC, and switch on. Check that the circuit works properly. Then, decrease the current range on the meter until a good reading is indicated. Make a note of this current, as it is the normal running current of your circuit.

To choose a diode, you must consult the catalogues or data sheets and find one where the quoted *maximum forward current* exceeds the current you have measured; preferably it should be at least twice your measured current. Secondly, the diode will have a *peak inverse voltage* (PIV); this is the maximum voltage it can withstand when the cathode is made positive with respect to the anode, i.e when it is *reverse-biased* and not conducting. This voltage must be greater than your battery voltage, again by a factor of about 2. Except in the case of the bridge rectifier (see later), these criteria will enable the selection of a suitable diode to be made easily.

The series diode

The simplest and most obvious way to protect equipment is to insert a diode in the positive supply lead, as shown in **Figure 1**, with the diode passing current *only* when the supply is of the correct polarity. Because of the 0.6 V that exists across the diode, your equipment will normally operate on a

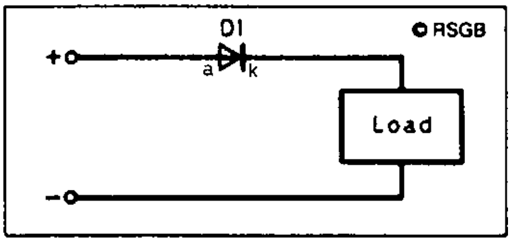

Figure 1 Series diode protection

slightly lower voltage. If you imagine reversing the supply to this circuit, you will see that the negative terminal of the battery is connected to the anode of the diode; the diode becomes reverse-biased and will not conduct any current, thus protecting the load. Your equipment will not operate when the battery connections are reversed.

The parallel diode

The circuit of **Figure 2** overcomes this voltage drop. It places the diode in parallel with the load (your equipment) but in a normally reverse-biased condition so that it draws no current when the battery is correctly connected. Reverse the battery connections, however, and a very large current will flow through the diode, thus blowing the fuse! For this technique to work successfully, the current drawn by the diode when the battery connections are reversed *must* be much greater than the maximum current drawn by the equipment, in order to blow the fuse. This is usually not a problem, however.

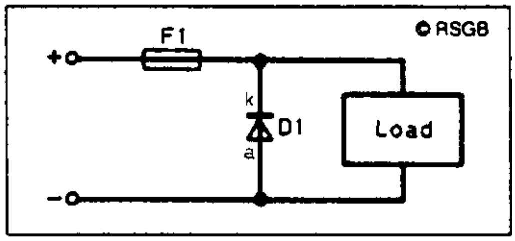

Figure 2 Parallel diode and fuse

The diode bridge

For sheer elegance, the circuit of **Figure 3** takes the biscuit! It uses four diodes connected in the form known as a *bridge rectifier.* Such rectifiers exist, and do not have to be made up from four discrete diodes.

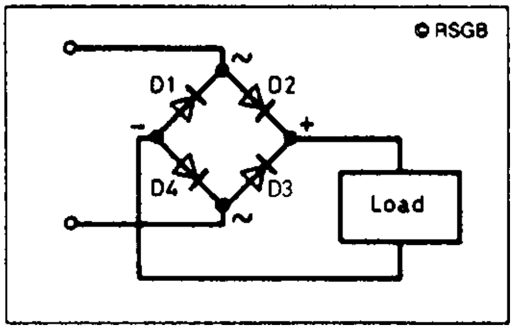

Figure 3 The diode bridge

Follow the current round the circuit from the supply, assuming initially that the top wire is positive and the bottom wire negative. It flows from the positive supply

(a) through D2
(b) through the load (top to bottom)
(c) through D4

and back to the negative of the supply.

Now assume that the bottom supply lead is positive and the top lead negative. The current flows from the positive supply

(a) through D3
(b) through the load (top to bottom)
(c) through D1

and back to the negative of the supply.

So, whichever way round the battery is connected, the current will *always* flow the same way through your equipment!

The circuit does have a drawback, however. Whichever way round the battery is connected, there are always **two** diodes conducting the current at any time. In the first case it is D2 and D4; in the second case, D1 and D3. This means that there is a total voltage drop of about 1.2 V. If your equipment can tolerate that reduction in voltage, then you will not have a problem.

Decoupling

Whenever the supply rail to a piece of equipment, or even to an individual stage of a circuit, is broken for the insertion of a device that will drop voltage, strange things can happen. This is because the supply for any circuit is assumed to have a low resistance to DC and a low *impedance* to AC. (Impedance is the AC equivalent of DC resistance.) These two are not the same, and the insertion of a diode or diodes is certain to make a big difference to them both. To overcome any irregularities in the operation of the circuit you are protecting in any of the ways described previously, the protected supply needs to be *decoupled* with a capacitor.

If the circuit handles audio frequencies only, placing an electrolytic capacitor directly across the load in all three circuits should solve the problem. The parallel circuit of Figure 2 is less at risk than the other circuits. The size of the capacitor will be determined by the current taken by the circuit, and may need to be chosen within the range 100 μF to 10 000 μF, with a working voltage greater than the supply voltage.

If the circuit is mainly handling RF currents, placing a capacitor of 0.01μF across the load should prevent any problems. A second capacitor, also across the load, of between 10 μF and 100 μF may be needed. Again, the parallel circuit of Figure 2 is less at risk than the other two.

Don't be afraid to experiment, but confine your experimenting (at first) to small equipment and low currents, until you get a 'feel' for the technique.

40 An RF signal probe

Introduction

A radio-frequency (RF) diode probe is a simple device which, when used with a conventional multimeter, enables the measurement of RF voltages in a circuit. When constructed, this will be one of the most useful pieces of test equipment for the experimenter who revels in the construction of transmitters and receivers.

The circuit

Figure 1 shows the simple circuit diagram. It is *almost* the same as a common diode rectifier circuit, but with a simple change to make it more sensitive. The circuit is known as a *voltage-doubler*, and is often found in high-voltage supplies, with beefier capacitors and diodes, of course! Because we are dealing with high frequencies and smaller voltages, the diodes and capacitors can be physically very small. The diode circuit of D1 and D2 *rectifies* or *detects* the RF from the probe, and any remaining AC is removed (short-circuited to ground) by C2. This produces, at the output, a constant voltage proportional to the peak-to-peak RF voltage present at the input; the output voltage is fed to an ordinary multimeter (on a voltage range).

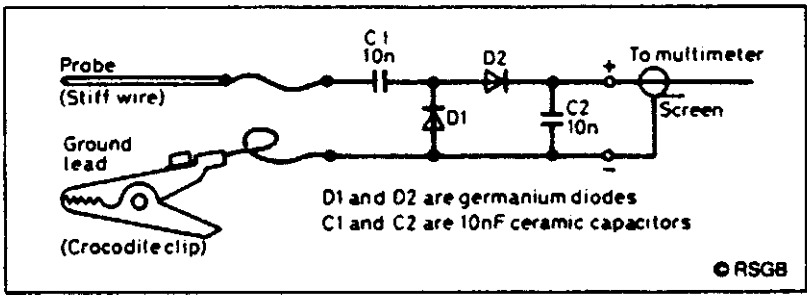
Probe
(Stiff wire)
Ground
lead
(Crocodile clip)
C1
10n
D2
D1
C2
10n
+
−
To multimeter
Screen
D1 and D2 are germanium diodes
C1 and C2 are 10nF ceramic capacitors
© RSGB

Figure 1 The diode probe can be made from four electronic components

Construction

Although the circuit layout is not critical, a description of the prototype is given here for your information. **Figure 2** has the details. The components are soldered to square pieces of copper-clad printed-circuit board (PCB) glued to a larger piece of PCB. The larger piece serves as a ground connection. Use a stiff copper wire as the probe, and an insulated flexible wire with a crocodile clip to connect the probe to the ground of the circuit under test.

Cut a piece of plain PCB, 30 mm by 45 mm and another of 15 mm by 45 mm. Cut the smaller piece into three measuring about 15 mm square. Stick the three small pieces to the larger piece, ensuring that there are small gaps between each, as Figure 2 shows. Solder the components in place. The connection to the multimeter should be thin coaxial cable or screened microphone cable. If you use unscreened cable, there may be RF pickup here which can lead to false readings.

Simple to use

Using the probe is simple. Connect it to the multimeter and set the meter to around 10 V DC – you may need to reduce this, depending upon the magnitude of the RF voltage you are trying to measure. Hold the probe by

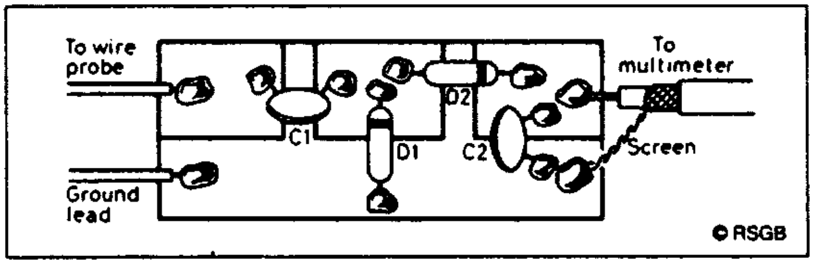
To wire
probe
Ground
lead
C1
D1
D2
C2
To
multimeter
Screen
© RSGB

Figure 2 The components are mounted on a copper-clad PCB

its base, being careful *not* to touch any of the components. Connect the croc clip to a ground point on the circuit and touch the probe on the circuit point to be tested. If an RF signal is present, there should be a reading on the multimeter. That's all there is to it!

Parts list

Capacitors

C1, C2	10 nanofarads (nF), or 0.01 microfarad (μF) ceramic

Semiconductors

D1, D2	Any *germanium* signal diode, e.g. OA91 or AA119

Additional items

- Stiff copper wire
- Insulated flexible wire
- Copper-clad PCB
- Crocodile clip
- Coaxial or screened cable

Source

All components are available from Maplin.

41 An RF changeover circuit

Introduction

The simplest entry route to operating an HF station is to buy or build a receiver, and to add a simple CW (Morse) transmitter. In theory, this sounds so simple, yet in practice there is one significant hurdle to be overcome. How can they both share the same aerial? A manually operated changeover switch is the obvious solution, but it takes time to perform the switching operation, by which time someone else has squeezed in before you and contacted the DX station! This small circuit accomplishes the changeover automatically, as soon as the transmitter is keyed.

How it's done

The circuit is shown in **Figure 1**, and uses the signal from the transmitter to operate a relay. A relay is a switch which is operated electrically in the following way. A coil of wire with an iron core is used as an electromagnet. When current flows through the coil, it produces a magnetic field which, in turn, is used to pull a set of switch contacts. These contacts will be used to switch the aerial from transmitter to receiver and vice versa. The relay used here has a double-pole changeover (or DPDT, or DPCO) switch. If you are confused by the different types of switch, look at the basic descriptions elsewhere in this book. The *pole* of a switch is the part that doesn't move; in Figure 1, one pole is connected to L1 and the other to the aerial output. In a circuit diagram, the switch contacts are always shown in their *normal* state, i.e. when *no* current flows in the relay coil. The changeover switch is normally in the receive position, so we say the receive switch is *normally closed*, and the transmit switch is *normally open*. We choose to have the circuit in the receive position normally, because everyone spends a lot more time receiving than transmitting. The relay is energised only when you are transmitting.

There are three RF sockets – one for the transmitter, one for the aerial, and one for the receiver. When the transmitter is not in use, there is a direct connection (via the Rx normally closed contacts of the relay) between the aerial and the receiver. The transmitter (which is not keyed at this time) is connected to a 50 ohm dummy load, R1.

When the transmitter is activated by pressing the Morse key, an RF signal appears at the transmitter socket. The wire carrying the signal to the relay

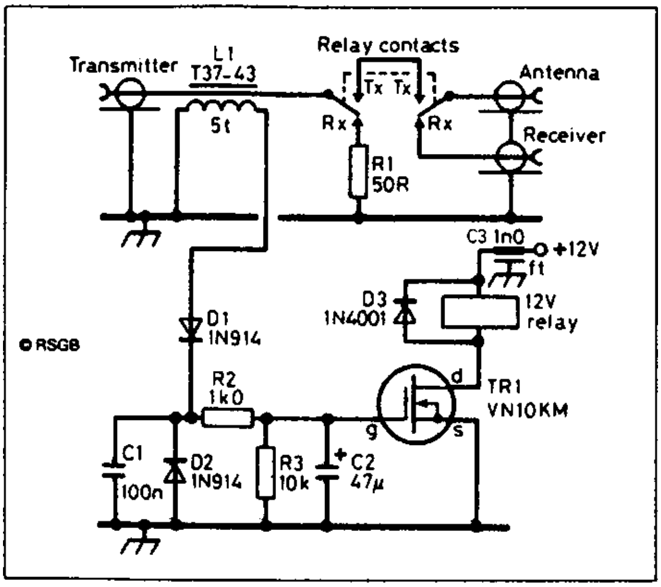

Figure 1 RF activated relay

Tx contacts passes through a toroid (a ferrite coil former, looking like a black Polo mint), which has a five-turn coil wrapped around it. This five-turn coil picks up the signal, which is then detected by D1 and D2, and converted to a steady voltage. C1 removes any remaining RF and C2 provides *hang*. Without it, the circuit would detect the gaps between every dot and dash being sent, and switch the aerial over very rapidly and very frequently! This is *not* what we want. We need the relay to remain in the transmit condition at normal Morse keying speeds, and then return to receive when the sending is complete. The combination of C2 and R3 achieves this.

The voltage appearing across C2 is fed to a VMOS field-effect transistor (FET), TR1, which acts as an electronic switch. When a voltage appears across C2, the FET switches on and passes a current through the relay coil, changing the contacts from receive to transmit. The diode, D3, across the relay winding protects the FET from being damaged by the large reverse-voltage spike which occurs across the coil when the FET switches off.

Construction

This is very simple – a circuit board is not required. The prototype was built into an old 50 g tobacco tin. All ground leads are soldered directly to the tin, thus supporting the components automatically. Phono sockets were used for the aerial, receiver and transmitter; you could use whatever connectors matched the rest of your station. The 12 V supply is fed through the tin using a 1000 pF feed-through capacitor, C3.

The ‘dummy load’, R1, must be able to withstand the RF output power of the transmitter. For novice use, 4 watts is adequate. Two suggestions for making up R2 from standard resistors are shown in **Figure 2**. Six 330 ohm, 1 watt resistors in parallel have a combined resistance of 330/6 = 55 Ω at 6 W. Two 100 ohm, 2 watt resistors in parallel have a combined resistance of 100/2 = 50 Ω at 4 W.

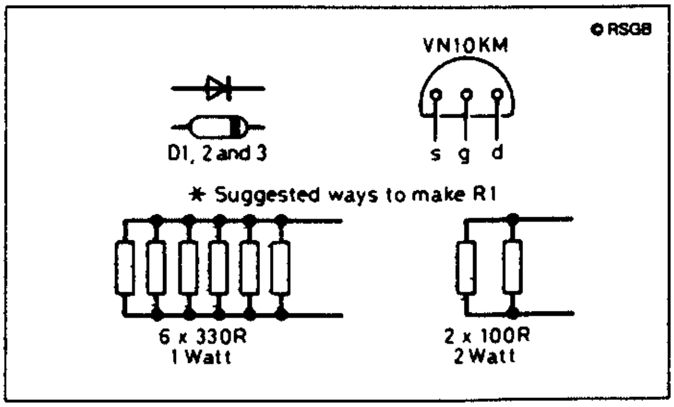

Figure 2 Component details

Make sure all the contacts to the relay are correct; also ensure that TR1, C2, D1, D2 and D3 are correctly wired.

Winding the toroid, L1, is very simple. It is made from PVC-insulated hook-up wire. Each time the wire passes through the centre of the toroid counts as one turn. The wire from the transmit socket to the relay simply passes through the centre! Use thin 50 Ω coax for the leads to the three sockets.

What happens to the receiver?

On transmit, the receiver is not connected to the aerial, but is only the separation of the switch contacts (about 0.5 mm!) away from the transmit lead. The receiver *will* pick up the transmitted signal, and there is a possibility that this signal will be enough to damage the receiver's sensitive input circuits. This can be prevented with the simple addition shown in **Figure 3**. Two diodes are connected back to back across the receiver socket. These act as a *limiter*, reducing the amount of signal that can enter the receiver. Solder these directly between the receiver socket and the tin.

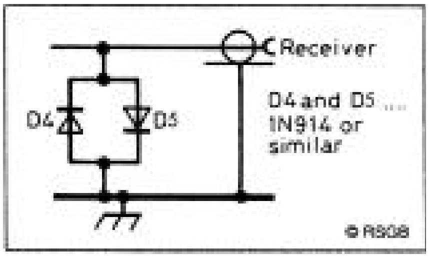

Figure 3 Receiver input protection

Warning

The wiring **must** be thoroughly checked. If the relay switch connections are wrong, there is a **great probability** that the full RF power will be applied to your receiver! You can check these connections as follows. With the 12 V supply connected, check that the relay energises (clicking sound heard) when the FET drain (d) is shorted to earth. Check the aerial and receiver connections with an ohmmeter or continuity tester. Disconnect the 12 V supply, and check that there is continuity between the centre pins of the aerial and receiver sockets. Reconnect the 12 V supply, short the FET drain to earth again (this will not damage the FET) and check that there is continuity between the centre pins of the aerial and transmitter sockets. If these tests show correct operation, disconnect the drain shorting wire (*most* important!) and your aerial auto-changeover is ready for use! Avoid using the device without the lid fitted.

Parts list

Resistors: 0.25 watt, 5% tolerance, unless otherwise stated

R1	50 ohms (Ω) – see text and Figure 2
R2	1 kilohm (kΩ)
R3	10 kilohms (kΩ)
Capacitors	
C1	100 nanofarads (nF), 0.1 microfarad (μF)
C2	47 microfarads (μF) 15 V electrolytic
C3	1000 picofarads (pF) feedthrough
Semiconductors	
D1, D2, D4, D5	1N914
D3	1N4001
TR1	VN10KM
Additional items	
L1	T37–43 toroid
	Relay 12 V DPDT (DPCO) relay
	Three sockets to suit (phono, SO239, etc.)
	Thin coax cable and hook-up wire

42 A low-light indicator

Introduction

This is a simple one-evening project that can be built for the pure fun of it, or to use as the basis of a more complex project to switch your shack lights on when it gets dark! In its prototype form, it simply flashes an LED when the ambient light level drops to a preset point.

Operation

The heart of the circuit shown in **Figure 1** is a *photo-conductive cell*, also called a *light-dependent resistor* (LDR), a device whose resistance changes according to the amount of light falling on it. In bright light, the resistance is low (about 1 kΩ), whereas in the dark, its resistance is very high (up to 10 MΩ). The cell is made from a semiconducting material known as cadmium sulphide (CdS), and is enclosed in a small plastic container. The semiconductor is laid on a flat insulating surface in the form of a small flat ribbon. The ribbon construction gives a good area of surface for a given

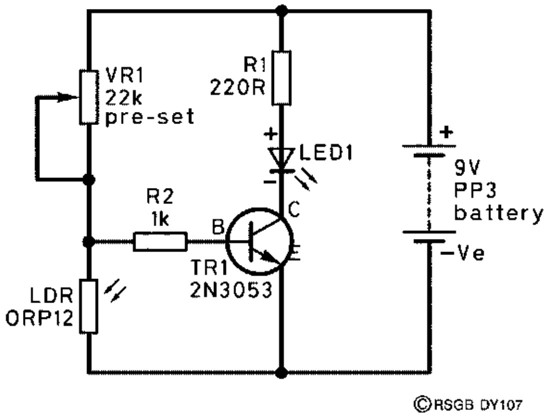
VR1
22k
pre-set
R1
220R
+
LED1
–
9V
PP3
battery
+
–Ve
R2
1k
B
C
E
TR1
2N3053
LDR
ORP12
©RSGB DY107

Figure 1 Circuit diagram

length of ribbon, and the length of the ribbon is maximised by laying it out in a zig-zag pattern, as can be seen diagrammatically in **Figure 2**. In the dark, CdS is an insulator; when light falls on it, electrons are released inside the CdS, making it conduct. The more light there is, the more electrons there are, and the resistance falls.

In this circuit, the LDR is connected across the 9 V supply in series with a variable resistor, VR1. In this arrangement, the voltage that exists across the LDR will be determined by the light level. As the light intensity increases, the resistance of the LDR falls, dropping a smaller voltage across it. The

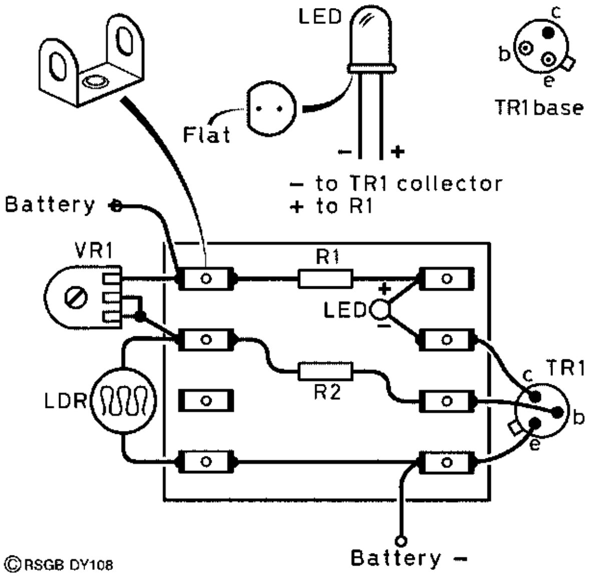
LED
Flat
– +
– to TR1 collector
+ to R1
c
b
e
TR1 base
Battery +
VR1
R1
+
LED
–
LDR
R2
c
TR1
b
e
Battery –
©RSGB DY108

Figure 2 Component layout on tagstrip

reverse happens when the light intensity falls – the voltage across the LDR increases. This voltage is used to drive an npn transistor, TR1, connected as an electronic switch. As the voltage on the base (B) rises, it will reach a point where TR1 will suddenly start to pass current, just as an ordinary switch does when pressed. The current flowing through TR1 also flows through a flashing light-emitting diode (LED), D1 and its series resistor, R1. D1 can be a steadily glowing type, if preferred. If you want to make the circuit switch the LED on at a different ambient light level, adjust VR1.

Construction

The prototype was made on an eight-tag tagboard (Figure 2). The resistors R1 and R2 can be laid on the tagboard for soldering, the rest of the components lying above or to the side of the board. Check the circuit after you have soldered everything on, then connect the battery. The LED should flash if you put your hand over the LDR, and VR1 can be adjusted to vary the point at which the LED lights. The project can be housed in a small plastic box with holes provided for the LDR and LED and an on/off switch if you want one.

Parts list

Resistors: all 0.25 ohm, 5% tolerance

R1	220 ohms (Ω)
R2	1 kilohm (kΩ)
VR1	22 kilohms (kΩ) linear pre-set

Semiconductors

LDR	ORP12
TR1	2N3053 npn
LED1	Flashing LED

Additional items

- Plastic box
- Tagboard with 8 tags
- PP3 battery connector

43 A J-pole aerial for 50 MHz

Introduction

For FM communication (i.e. voice and data) on the VHF bands, a vertical aerial is used to give all-round (non-directional) coverage. This is a half-wave aerial which can be fed at the end, thus removing the principal problem with the conventional vertical centre-fed half-wave dipole, which is that the feeder should leave the dipole at right angles. This is no problem when the dipole is horizontal, but can be difficult for the vertical dipole.

Basic facts

A feeder must be connected to an aerial at a point where the *impedance* (AC 'resistance') of the aerial closely matches that of the feeder. The difference between the two impedances gives rise to the voltage standing-wave ratio (VSWR), which is unity only when the two impedances are the same. With 50 Ω feeders, the feed point of a half-wave aerial is at the centre, where the aerial impedance is around the same value. At the end of a half-wave aerial, the impedance is high, so it is *not* a suitable point to connect a 50 Ω feeder.

Connection at this point can be effected using an RF transformer. RF transformers act in the same way as ordinary transformers, except that they are much smaller, and usually comprise wires of particular lengths adjacent to each other. **Figure 1** is a good starting point. It shows the aerial in its diagrammatic form. Notice that the aerial is in the form of an elongated letter 'J'; this shape gives rise to its nickname – the *J-pole*. The quarter-wave RF transformer is the lower 'U' section below the half-wave element. At the bottom of the U section, the impedance is zero (this may become clearer later) and at the top of the U section it is high, thus matching the aerial impedance. The coaxial feeder cable is connected part-way up the U section, where the impedance is around 50 Ω.

The practicalities

Figure 2 shows the aerial as constructed. As it is about 4.5 metres high, it may be too high for the average house loft, but is ideal for mounting outside, supported by a non-metallic pole or hung from a tree branch. The upper half-wave section is made from 1.5 mm insulated copper wire, as used in domestic mains wiring. The quarter-wave transformer below the half-wave section is made from 300 Ω balanced line ('ribbon cable'). The

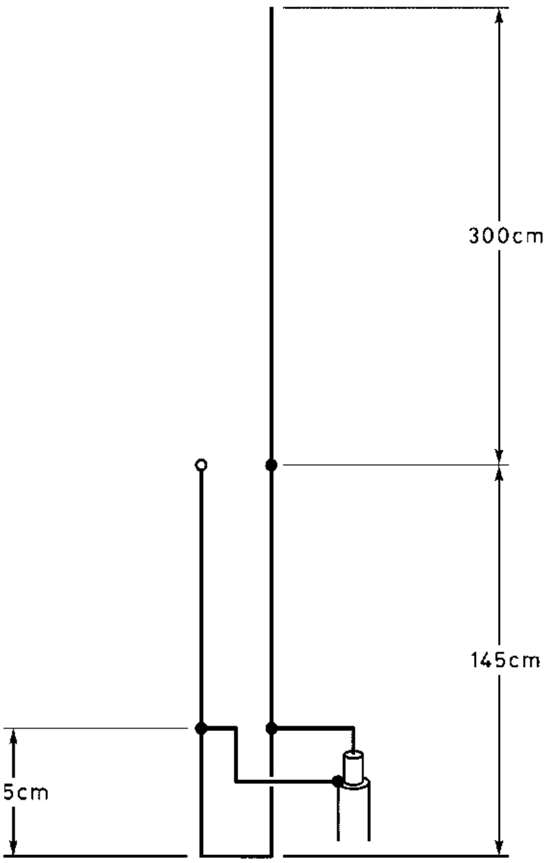

Figure 1 Overall dimensions of the 6 metre J-antenna

wires at the bottom of the transformer section are stripped of their insulation, twisted and soldered. At the upper end, only one wire of the balanced line is soldered to the bottom of the half-wave section. The other wire of the pair is not connected and is left insulated.

At the feed point of the transformer, the insulation needs to be carefully stripped from the balanced line. You will need a standing-wave meter (VSWR meter) in the coaxial line between your transmitter and the aerial, and you will need to adjust the position of the feed point. 45 cm from the bottom was the best point on the prototype, but this position is dependent upon the immediate surroundings of the aerial, and *must* be done when the aerial is in its final operating position. **Warning: *Never* make adjustments to the feed point when the transmitter is on. Make a VSWR measurement, switch off, move the feed point, switch on again, make another measurement, and so on.** You will need to aim for the lowest VSWR you can – certainly better than 2:1. Having found the best position, wrap all the exposed wires with self-amalgamating tape, to seal them against the ingress of moisture.

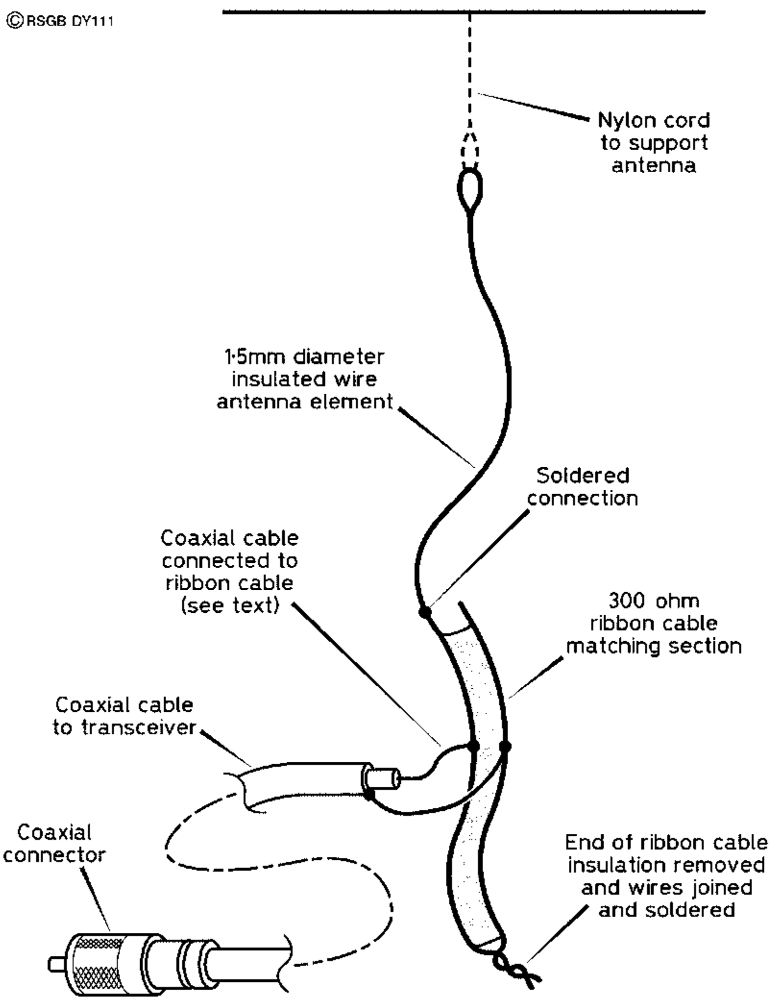
©RSGB DY111
Nylon cord to support antenna
1·5mm diameter insulated wire antenna element
Soldered connection
Coaxial cable connected to ribbon cable (see text)
300 ohm ribbon cable matching section
Coaxial cable to transceiver
Coaxial connector
End of ribbon cable insulation removed and wires joined and soldered

Figure 2 Construction details

How it performs

Figure 3 shows a computer prediction of how the J-pole radiates. It is called a *polar diagram*, and shows the distribution of your transmitted power when viewed 'from the end of your garden'. Most of your signal is sent at a fairly small angle to the horizontal; very little signal goes upwards, which is a good thing, of course. This also shows why the J-pole (or any other vertical aerial) should not be called 'omnidirectional', which means it radiates in all directions. It is omnidirectional *only* in the horizontal plane.

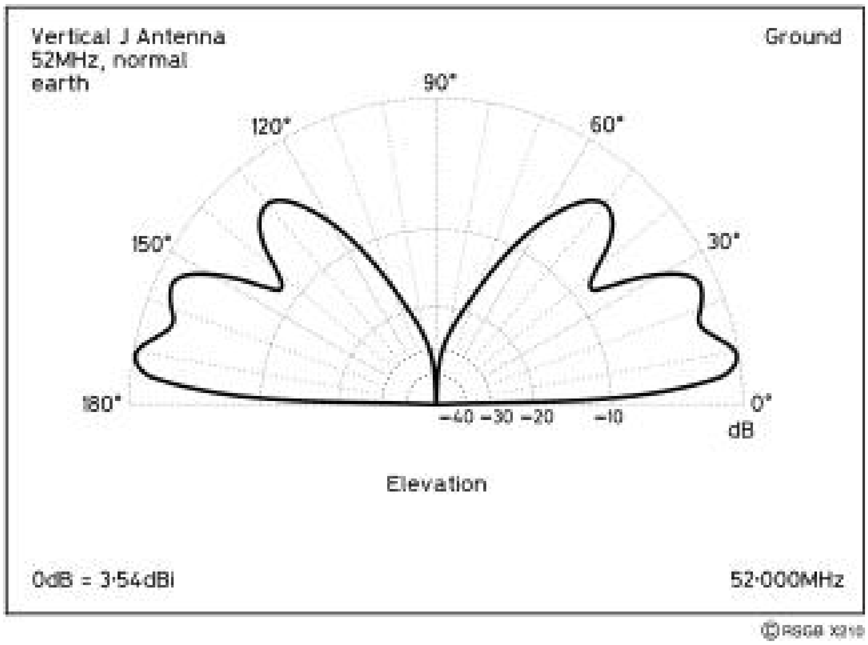

Figure 3 This polar diagram computed for the J-pole

Safety

Where you mount your aerial is a matter of personal preference and the restrictions of height and space, but the following safety rules *must* be applied.

1. Never fix an aerial where it may come into contact with power lines or telephone lines.
2. When climbing a ladder to put up an aerial outside, make sure the ladder is safe and that it is secured.
3. Don't do this alone. Preferably have someone with you. If this is not possible, make sure someone knows where you are.

Parts list

3.00 metres	1.5 mm insulated copper wire
1.50 metres	300 ohm balanced line
As required	50 ohm coaxial cable
As required	Self-amalgamating tape

44 Measuring light intensity – the photometer

Introduction

Before the days of automatic 'point-and-shoot' cameras, a photographer would use a light meter or *photometer* to measure the light level, then manually convert this reading into shutter speed and lens aperture settings to ensure a correctly exposed negative. Modern cameras have quite sophisticated photometers, which control the shutter speed and iris settings automatically.

A short explanation

A simple photometer circuit is shown in **Figure 1**, and is based on a device called a *light-dependent resistor*, or LDR. As its name tells us, its resistance depends upon the amount of light falling on it. In bright light, the resistance is relatively low (about 1 kΩ), whereas in the dark, its resistance is very high (up to 10 MΩ). The cell is made from a semiconducting substance known as cadmium sulphide (CdS), and is enclosed in a small plastic container. The semiconductor is laid on a flat insulating surface in the form of a small flat ribbon. The ribbon construction gives a good area of surface for a given length of ribbon, and the length of the ribbon is maximised by laying it out in a zig-zag pattern. In the dark, CdS is an insulator; when light falls on it, electrons are released inside the CdS, making it conduct. The more light there is, the more electrons there are, and the resistance falls.

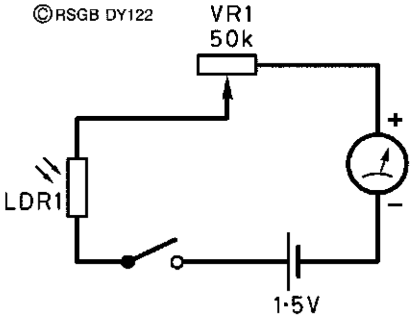

Figure 1 Circuit diagram of photometer

The photometer circuit

The circuit is simply a series connection of four things – the battery, the LDR, a variable resistor and a meter. A switch is also provided. The combination of the LDR resistance and that of VR1 determines the current flowing, which is indicated on the meter. Altering the resistance of VR1 sets the *sensitivity* of the photometer – you may want full-scale deflection of the meter needle for a bright light, or for a dim light.

It is simple to make, and a plug-in type matrix board is ideal to test the circuit, so that you can decide if you want to make a permanent version. Connect all the components in series; the only change you may want to make is the connection to VR1. If you find that the sensitivity control seems to work 'backwards', simply unsolder the wire from the end tag of VR1 and solder it to the opposite end tag. Problem solved!

In use

As soon as you connect up the battery, you will probably have a meter reading because of the daylight falling on the LDR. Shading it with your hand should reduce the reading. If the meter needle is hard over against the end-stop, turn VR1 until it indicates about half-scale. The LDR is very sensitive, and will read zero only in a dark room. If you put on a torch to see what the meter reading is, the LDR will detect the torch light, and will give a reading!

Here is a simple project where you can use the photometer in an experiment which has an analogy in radio. Draw a circle on a large (A3 or bigger) sheet of paper and divide it up into 30-degree sectors, as shown in **Figure 2**. Draw a series of smaller circles which divide the maximum radius into five. Look at the figure if you're not sure about this. Bring the LDR away from the circuit by using two long, flexible wires. The experiment must be performed in a darkened room (preferably in total darkness). Prepare a table with two columns, the left-hand one headed 'Angle (degrees)' and the right-hand one 'Meter reading'. Fill in the left-hand column 0, 30, 60 . . . and so on up to 360°.

Place the torch in the position shown, with its lens at the centre of the circle and pointing along the 0° line; switch it on. Place the LDR facing the torch and adjust VR1 until you have full-scale deflection of the meter needle. Suppose the meter indicates 10 units at this point. Enter this into your table in the 0° row. Keeping the torch the same distance from the circle centre, and pointing at it, move the LDR round all 30° positions and record the meter readings. Switch off the torch and take the sheet of paper and your tabulated results into daylight!

Lay the large sheet of paper on a table with your results beside it; then, at each 30° interval, plot the point along the radius corresponding to the meter

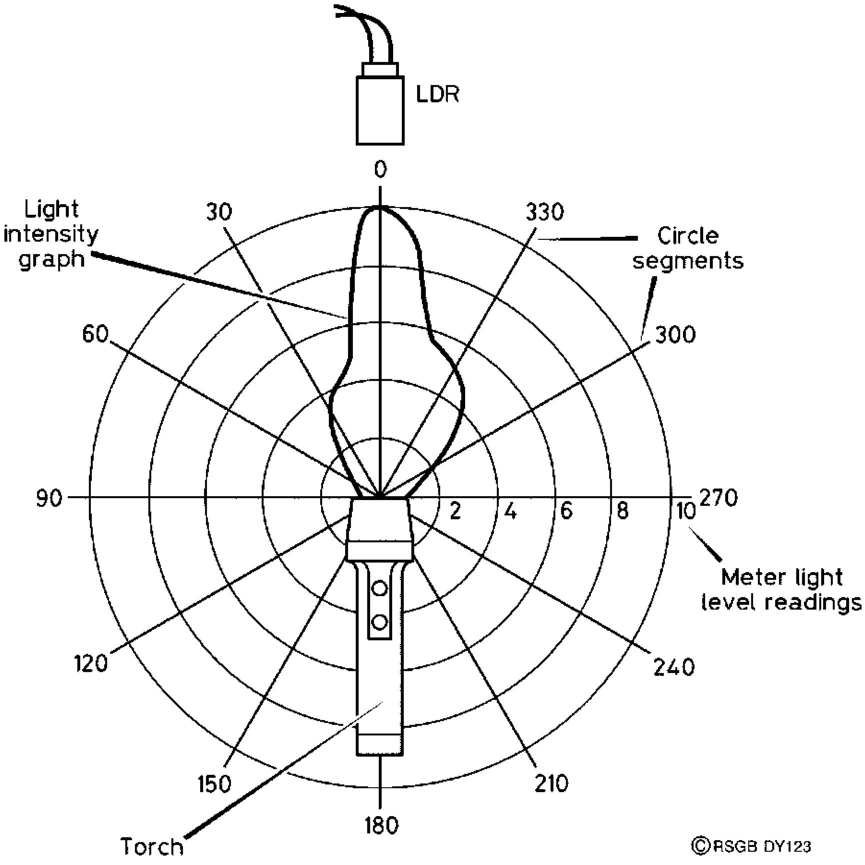

Figure 2 Torch light intensity pattern plotted at RSGB HQ

reading. Then, join up all the points and you have what is called a *polar diagram* of light intensity. The use of the word 'polar' implies that the readings have been taken in a circle and plotted that way.

Light waves and radio waves are both examples of *electromagnetic radiation*. The torch is designed to 'beam' its radiation in a particular direction, just like an aerial does. Hence the use of the word 'beam' for a directional aerial. If a similar polar diagram is drawn for a Yagi-type aerial, it will show the same general characteristics as does Figure 2, namely a main direction (or 'lobe') where most of the energy is concentrated, with evidence of *sidelobes*, indicated by 'lumps' on the otherwise smooth main lobe.

Also in this book you will find a project for the construction of a UHF field-strength meter, which you could use to carry out a measurement of the polar diagram of a UHF aerial.

Parts list

Resistor

VR1	50 kilohms (kΩ) linear

Semiconductor

LDR1	ORP12

Additional items
- Meter, 50 or 100 μA
- Battery clip for single AA cell
- AA cell

45 A 70 cm Quad loop aerial

Introduction

This is a description of how to make an aerial for the 70 cm band which has gain compared with the 'rubber duck' or the dipole aerial. It can easily be dismantled and reassembled, making it ideal for contest use.

The principles

Most aerials which comprise several similar *elements* arranged along a *boom* are variations of the design originated by Yagi and Uda, and which takes the name (for historical reasons) of the former, and is know as the Yagi array. Let us suppose we have a Yagi aerial beaming left to right in front of us. The elements get progressively shorter from the left (look at almost any TV aerial to see this). All the elements on a Yagi aerial are classified as follows:

- The *reflector* – the leftmost element as we look at the array. It is the longest. Next to it is:
- The *driven element* – this is the element connected to the feeder, which in turn runs down to the transceiver. It is slightly shorter than the reflector.
- All the elements beyond the driven element are called *parasitic elements*, or *directors*. They are shorter than the driven element and usually get progressively shorter as we progress to the right along the boom. The directors are mainly responsible for the *directivity* (or *beamwidth*) of the array.

The progression from a simple dipole (a driven element in isolation) to a Yagi array is simple, but is nevertheless important. To make an aerial of two elements, a reflector (*not* a director) is added to the driven element. For three or more elements, directors are added to the two-element design. Adding more and more directors soon becomes impractical, the reduction in beamwidth (such as it is) does not warrant the extra expense, weight and wind resistance that is incurred.

Instead of using linear (straight) elements, as in the generic Yagi, this design uses loops. Designs using squares of wire instead of loops are known as *Quad* aerials, and HF designs require large X-shape frames to support the large squares of wire. At 70 cm, however, the use of wire loops is easier, and they are self-supporting.

Construction

This is quite simple. Any type of material (metal, plastic, wood) can be used for the *boom* (the support for the elements) and for the mast. The elements are made from 14 SWG enamelled copper wire. 16 SWG hard-drawn aerial wire, which is not enamelled, has also been used with success. Thinner wire might result in a rather ‘floppy’ aerial! The separate parts of the aerial are held together with *jubilee clips* (hose clips).

The driven element is secured to the boom with a jubilee clip and a three-connector plastic connector block as shown in detail in **Figure 1**, and in the photograph. Cut the wire for the driven element 70 mm longer than the 700 mm indicated in Figure 1. Then, using sandpaper, remove the enamel from one end to a distance of 20 mm, and from the other to a distance of 50 mm. After forming the loop of the driven element, bend both stripped ends through 90°, and insert them into the first to holes of the plastic connector block (Figure 1). Do not tighten the screws yet. Push both ends into the block as far as they will go, then bend the 50 mm end back on itself and pull the ends back through the connector block so that the end you have just bent goes into the third hole in the block. Now tighten the screws in the block and in the jubilee clip.

Each director and the reflector should be made 40 mm longer than the circumferences shown in Figure 1. Strip the enamel, as before, from the last 20 mm at each end. Form the wire into the loop, slip the ends under the jubilee clip (Figure 1) and tighten it. You may find that it helps to solder the stripped ends together before securing the jubilee clip.

The boom is fixed to the mast using jubilee clips and wire, as shown in Figure 1. Solder the feeder cable to the driven element, with the braid soldered to the end of the driven element which is connected to the boom (this applies to metal booms; with plastic or wooden designs, the feeder connections are not critical).

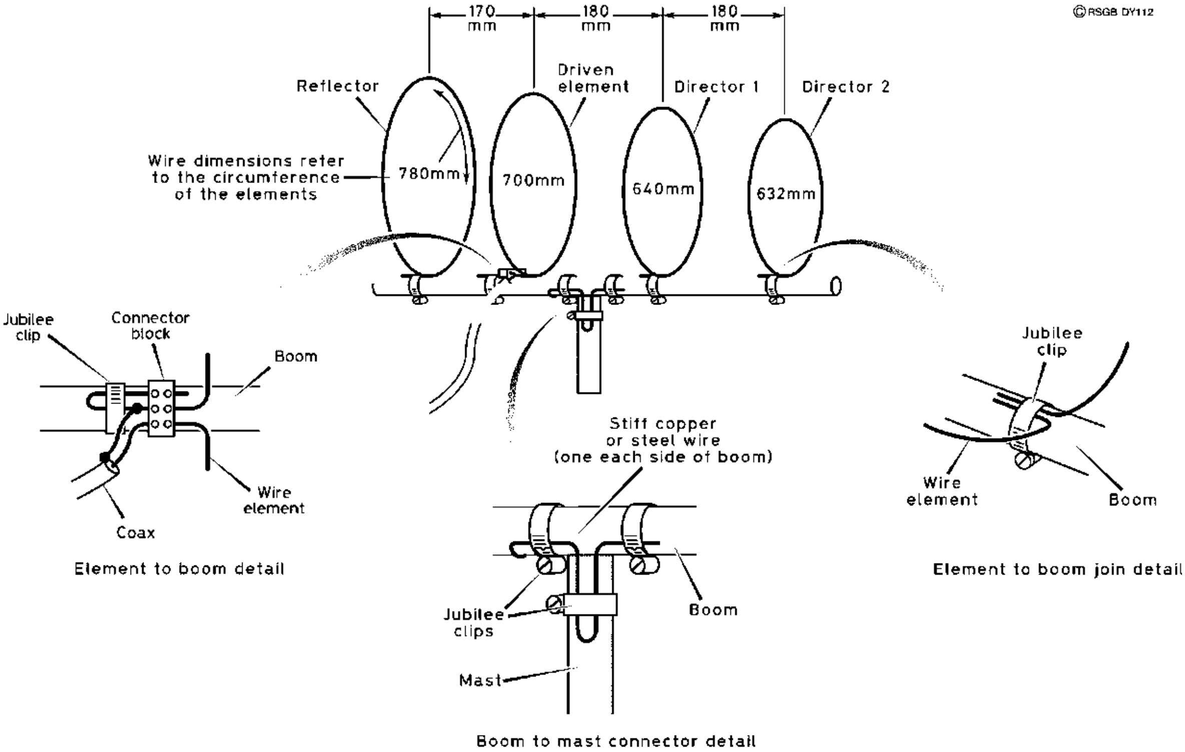
170 mm
180 mm
180 mm
©RSGB DY112
Reflector
Driven element
Director 1
Director 2
Wire dimensions refer to the circumference of the elements
780mm
700mm
640mm
632mm
Jubilee clip
Connector block
Boom
Wire element
Coax
Element to boom detail
Stiff copper or steel wire (one each side of boom)
Jubilee clips
Boom
Mast
Boom to mast connector detail
Jubilee clip
Wire element
Boom
Element to boom join detail

Figure 1 70 cm 4 element quad construction, with detail of how hose clamps are used in the construction

You will need a plug on the shack end of your feeder to suit the transceiver, aerial tuning unit (ATU) or the standing-wave meter (SWM) you are using. Always connect these in the order: transceiver – SWM – ATU – aerial.

Testing

Always test aerials outside and away from buildings (if possible!). This avoids getting misleading results.

Use a rubber duck, or whatever aerial you usually use, and tune around to find a repeater or beacon signal which is consistent. Note the reading on the S-meter. Then, connect your new Quad loop. Rotate it to give the strongest S-meter reading (don't forget it is directional). Verify that the meter reading varies as you rotate it. How does the S-meter reading compare with the original reading? Much depends on the siting of your original aerial; if you are comparing your Quad loop at ground level with a vertical on the chimney, you wouldn't expect your new aerial, even with its gain, to outperform a vertical which is well elevated!

Now a test on transmit is called for. The use of an SWR meter is essential here. Find a clear frequency and check that it *really is clear* before announcing your presence and carrying out the test. A reading of 1:1 is excellent, but any value less than about 1.8:1 would be acceptable. You can measure the directivity of your aerial using the field strength meter, also described in this series.

Materials

4 metres of 14 SWG enamelled copper wire
Material for boom and mast
15 amp connector block
7 Jubilee clips

The enamelled copper wire is available from AA&A Ltd, Sycamore House, Northwood, Wem, Shropshire SY4 5NN. Everything else is available from most hardware stores.

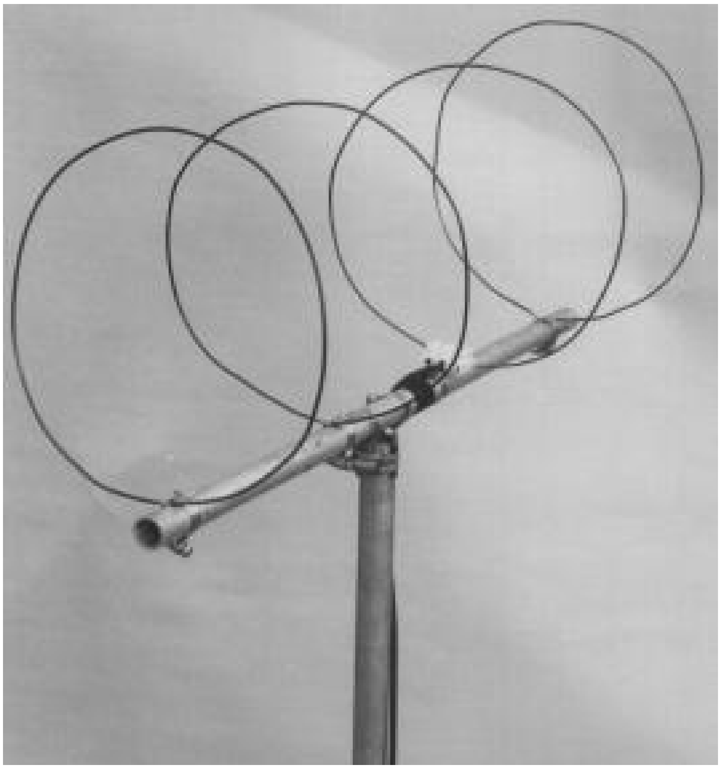

46 A UHF field strength meter

Introduction

It is always interesting, and often useful, to know just *where* the radiation is going from your aerial. How much of your radiated energy is going in the general direction of the station you are in contact with, and how much is being effectively wasted? Some of these questions can be answered with the use of a *field strength meter*. A field strength meter is simply a receiver, stripped down to its bare essentials, such that it responds only to the magnitude of the carrier. The use of a field strength meter assumes that the aerial under test is radiating a continuous carrier. Don't forget to find a clear frequency and identify your transmissions at least every quarter-hour, in order to comply with the terms of your licence.

Description

Two types of field strength meter are shown in **Figure 1**. You will recognise both circuits (Figure 1b particularly) as being types of 'crystal set' with a meter replacing the headphones. Figure 1a is a broad-band HF design (there is no tuning provision) and Figure 1b is tuned in the same way as the crystal set; with a loop of wire as an aerial, it will perform well in the VHF/UHF range.

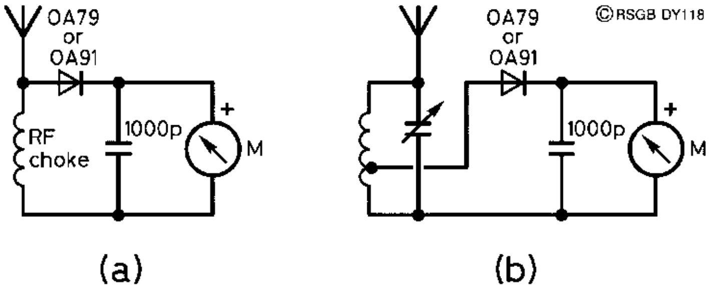

Figure 1 Construction of the UHF field strength meter

A tuned field strength meter also doubles as an *absorption wavemeter* if it is well calibrated. Such devices are useful for detecting transmitter harmonics also. This design is intended for use in the UHF band, so it will have to be sensitive around 432 MHz.

Construction

The field strength meter comprises a loop of wire, 600 mm long, which acts both as aerial and as the tuning inductance, a diode, a capacitor, a connector block, a meter and a length of twin wire. All the components, with the exception of the meter, are fixed to a pole with a jubilee clip, as shown in **Figure 2**. The meter should have a sensitivity of between 50 µA and 100 µA, or a multimeter can be used. The multimeter is more flexible, as you can select different current ranges, giving you a range of sensitivities.

Using the field strength meter

Connect your meter to the ends of the twin wire from the pole. Place a hand-held transceiver about 2 metres away, and press Transmit. If there is no meter reading switch off the transmitter and check the wiring. If the needle attempts to go negative, simply reverse the wires to the meter. If the reading is too high, either move the transmitter further away, or increase the current range on the multimeter. Try changing the orientation of the transmitter aerial, and note how the signal varies.

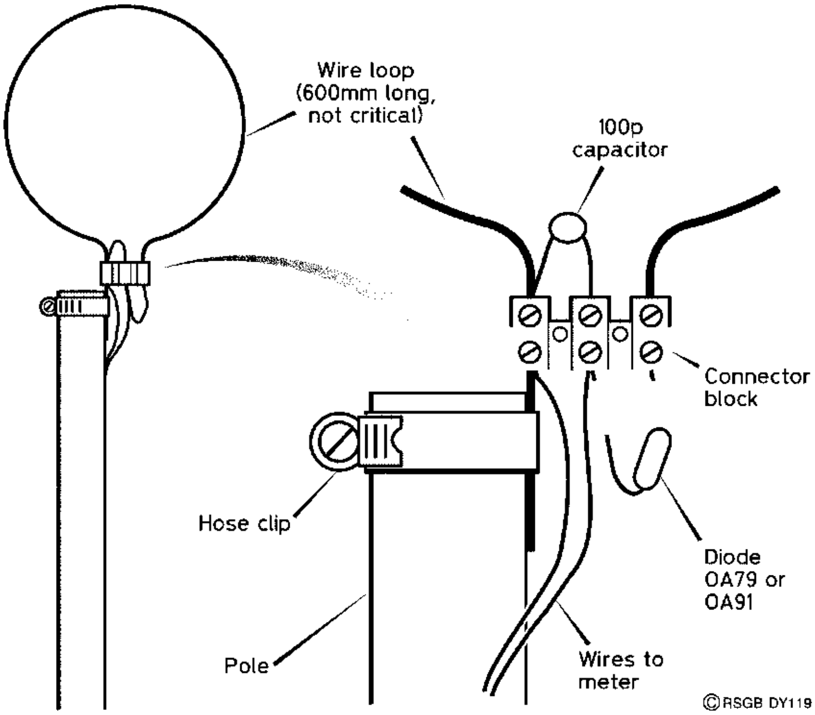
Wire loop
(600mm long,
not critical)
100p
capacitor
Connector
block
Hose clip
Diode
OA79 or
OA91
Pole
Wires to
meter
©RSGB DY119

Figure 2 Assembly of the field strength meter

To measure the *polar diagram* (a graph of the signal strength against the angle between the aerial boom and the field strength meter, plotted on polar axes) of a beam aerial, mount the loop as far away from the aerial as you can, preferably at the same height, and rotate the aerial, say 15 degrees at a time, and record the signal strength, until the aerial has been turned through 360 degrees. If you need help plotting the graph, enlist the help of a friend who has done it before, or one who knows about polar graphs! If you have already built the Photometer project, you will have measured and plotted the polar diagram of the light intensity from a torch. Now you will see the great similarity!

Parts list

Capacitor	100 picofarads (pF)
Diode	Germanium, OA79 or OA91
Connector block	10 A, 3-way
Wire	600 mm of 16 SWG enamelled copper
	Length of twin cable for meter connection
Clip	Jubilee (hose) clip

47 Christmas tree LEDs

Introduction

A novelty ideal for the festive season, this circuit causes one LED at a time to light up around a small cardboard Christmas tree.

Warning

This circuit uses members of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which each IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains-earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**
3. **Repeat the process with the second IC.**

The circuit is safe from destruction while it is connected to the battery. However, when the battery is removed, the same care should be exercised with its handling, because there are no supply decoupling capacitors across each IC.

Description

Figure 1 shows the layout, and how the wires are connected from the circuit board to the LEDs around the tree. The circuit, shown in **Figure 2**, is quite complicated, so you need to be confident in your logical approach to circuit-building before you attempt this one! It uses two common integrated circuits (ICs). IC1 is simply an oscillator which provides timing pulses for IC2, which 'counts' up to a maximum of 10. The outputs of the counter are indicated by light-emitting diodes (LEDs); red, orange, yellow and green are common colours which you can use.

The circuit is built on a single piece of Veroboard measuring 29 holes by 12 strips. **Be aware that there is no strip labelled 'I', so don't make mistakes in your counting!**

First of all, cut the tracks in the positions shown in Figure 1 using a 3 mm ($\frac{1}{8}$ inch) twist drill rotated between thumb and forefinger. Then, solder in the IC sockets, the notches facing row M. Wire up and solder in the links and then the Veropins for the connections to the LEDs and battery. Having done this, the resistors and the capacitor should be fitted. When wiring the LEDs, each cathode (the lead adjacent to the 'flat' on the LED encapsulation) is connected to R4, and the anodes go to separate pins on IC2.

Testing

Hold up your circuit board to the light and check carefully for solder bridges between adjacent tracks. Then check again that the wiring is correct. Place the ICs in their holders, with the notched ends lining up with those on the holders. Connect the battery, and the LEDs should illuminate in sequence.

If you have no success, you are now wishing you had checked the circuit more carefully! Learn something from your mistake and it will not have

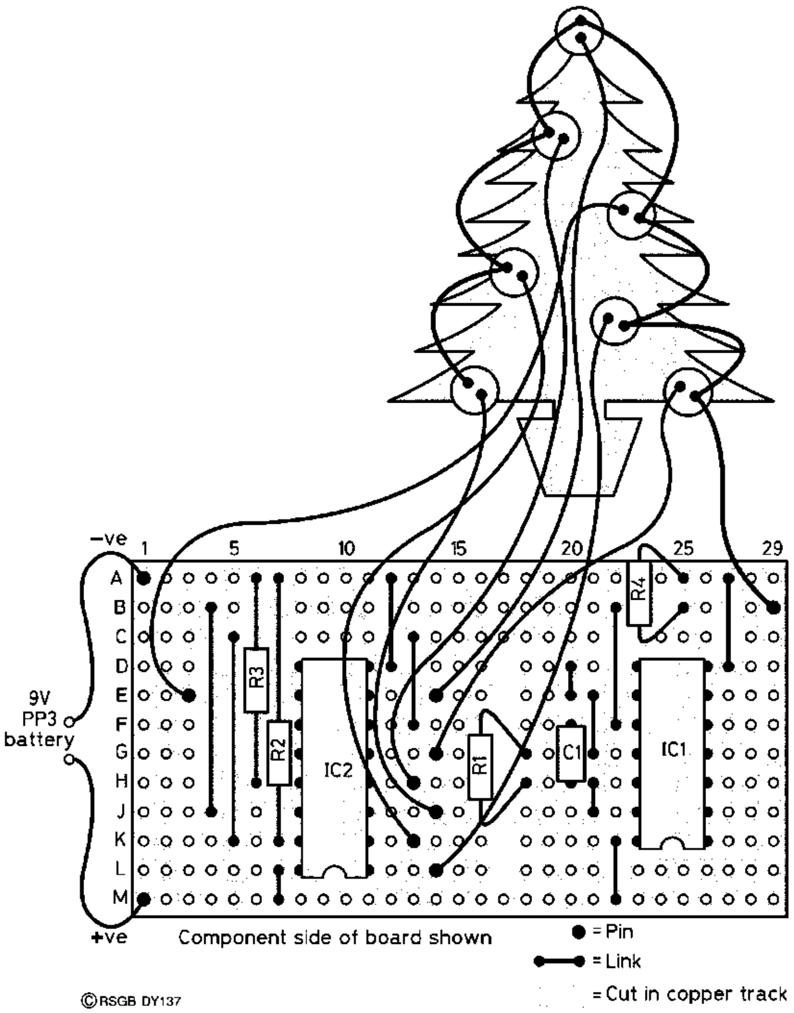
−ve
1
5
10
15
20
25
29
A
B
C
D
E
F
G
H
J
K
L
M
9V
PP3
battery
R3
R2
IC2
R1
C1
IC1
R4
+ve
Component side of board shown
● = Pin
●—● = Link
= Cut in copper track

©RSGB DY137

Figure 1 Christmas tree, component layout

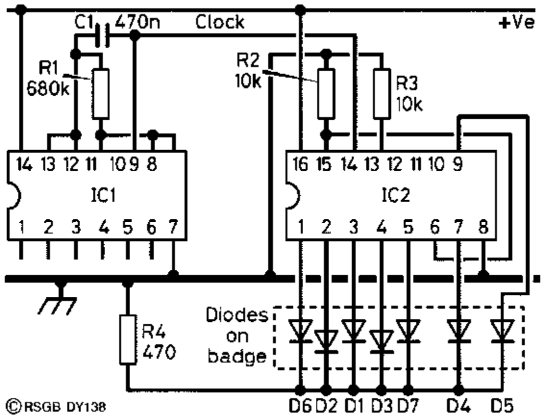
C1 470n
Clock
+Ve
R1
680k
R2
10k
R3
10k
14 13 12 11 10 9 8
IC1
1 2 3 4 5 6 7
16 15 14 13 12 11 10 9
IC2
1 2 3 4 5 6 7 8
R4
470
Diodes on badge
D6 D2 D1 D3 D7 D4 D5

©RSGB DY138

Figure 2 Christmas tree, circuit diagram

been in vain! The first thing to check is that the LEDs are the correct way round. If that is OK, then check the positions where the tracks are broken (intentionally!). After that, are the wire links in the right places – had you forgotten there is no row ‘I’? Check *all* the wiring, then check *again* for solder bridges, and switch on again. One of these tests should have revealed a fault. If it *still* refuses to work, perhaps one (or both) of your ICs were damaged by static electricity, despite your precautions – or did you choose to ignore them?

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	680 kilohms (kΩ)
R2, R3	10 kilohms (kΩ)
R4	470 ohms (Ω)

Capacitor

C1	0.47 microfarad (μF) min. metallised polyester film

Semiconductors

IC1	4011
IC2	4017

Additional Items

D1–D7	5 mm LEDs in choice of colours
	14-pin DIL socket for IC1
	16-pin DIL socket for IC2
	PP3 battery clip and battery
	Veroboard, Veropins
	Insulated wire for links

48 An audio signal injector

Introduction

An audio signal injector is a device used to test audio frequency circuits. It is simply an oscillator running at a frequency in the audio range, so that when its output is fed to the input of an amplifier, it will produce a sound in the loudspeaker if the amplifier is working. The oscillation is so rich in harmonics that the signal can also be heard (sounding rather different) when injected into an RF circuit.

The design

The circuit is shown in **Figure 1**, and is a basic *astable multivibrator*, a free-running oscillator producing a roughly rectangular-wave output. The two transistors, TR1 and TR2, operating as switches, switch on and off alternately at a frequency around 500 Hz. The prototype was constructed on plain matrix board (no copper strips), as illustrated in **Figure 2**.

Both transistors are type BC108, which are only a few pence each new, and can be found at almost any rally. You can add an on/off switch, or simply disconnect the battery when you are finished using it. To make the unit in one piece, the battery can be taped to the board, as the diagram shows.

The probe itself is made from a short piece of stiff insulated wire, soldered to a tag on the board; an earth lead is also soldered to the board, and terminated in a crocodile clip to attach to the ground lead of the equipment under test.

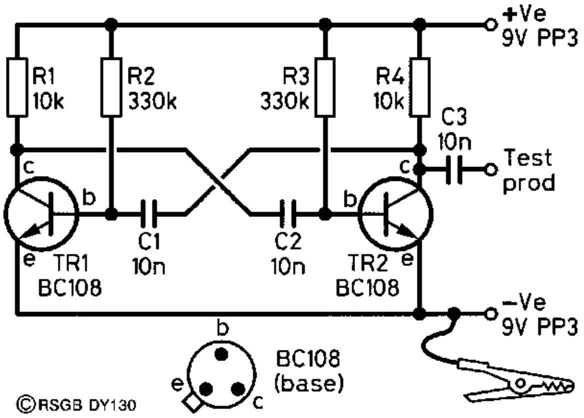

Figure 1 Circuit diagram of the signal injector

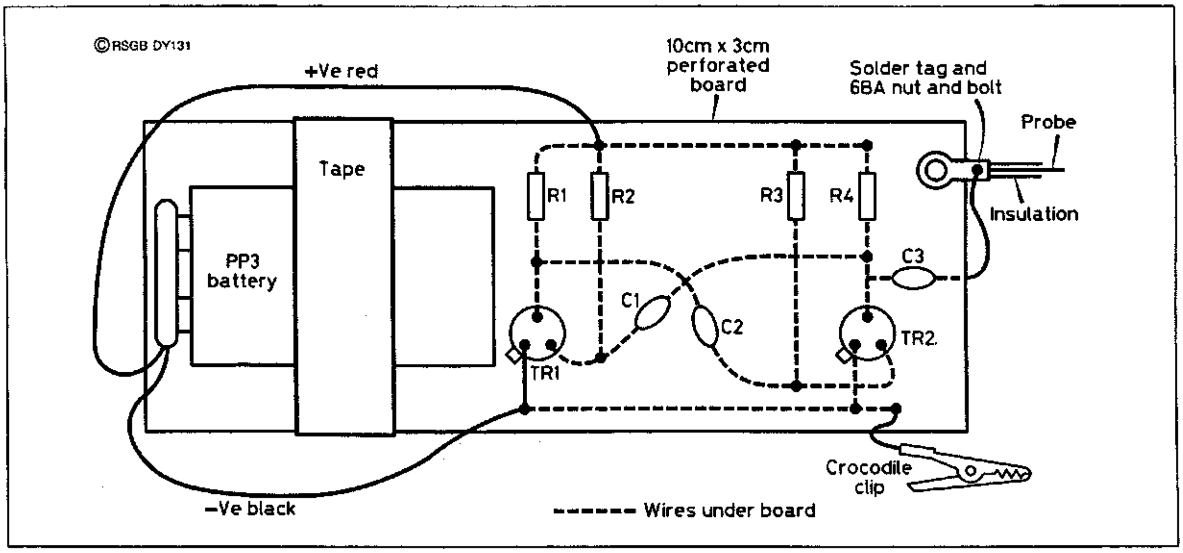
©RSGB DY131
+Ve red
10cm x 3cm perforated board
Solder tag and 6BA nut and bolt
Probe
Tape
PP3 battery
R1
R2
R3
R4
Insulation
C3
C1
C2
TR1
TR2
Crocodile clip
−Ve black
Wires under board

Figure 2 Component layout of the signal injector

Using it

After checking the wiring of the circuit, connect the battery. Find a transistor radio that is known to work. Carefully remove the case, switch on and apply the probe to the centre tag of the volume control. A *very* loud buzz should be heard from the speaker, indicating that the audio circuits of the radio are working. Using the injector to fault-find equipment you have made yourself is rather more instructive and rewarding, because you *know* where to inject the signal, and you should *know* what to expect when you do.

Warning

Do *not* work on any equipment connected to the AC mains. Work only on battery-powered circuits, for your own safety.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R4	10 kilohms (kΩ)
R2, R3	330 kilohms (kΩ)

Capacitors

C1, C2, C3	10 nanofarads (nF) or 0.01 microfarad (μF) ceramic

Semiconductors

TR1, TR2	BC108

Additional items
- PP3 battery and connector
- Matrix board 10 cm by 3 cm
- 6BA solder tags
- Thick insulated wire for probe
- Crocodile clip

49 Standing waves

Introduction

Everyone has, or *should* have, a standing-wave ratio (SWR) meter as part of his/her array of test gear. Most people know how to use it, but what does it really do?

Before attempting to answer that question, we need to look at some aerial fundamentals. An aerial is a *transducer*, the word meaning 'to lead across'. We use it whenever one form of energy is converted into another form. A bulb is a transducer; it converts electrical energy in the filament to radiated heat and light energy. **Figure 1** shows the situation.

An aerial, or antenna, is also a transducer; it converts radio-frequency (RF) energy in the feeder (or *transmission line*, to give it its proper name) into radiated electromagnetic energy, in the manner shown in **Figure 2**. The aerial has resistance, just like the bulb filament, and if the filament had zero resistance, there would be no radiated energy. In the same way, if the aerial had no resistance, there would be no radiation from it.

Transmission lines should convey RF energy from transmitter to aerial with the minimum power loss. A common form of transmission line is the *coaxial*

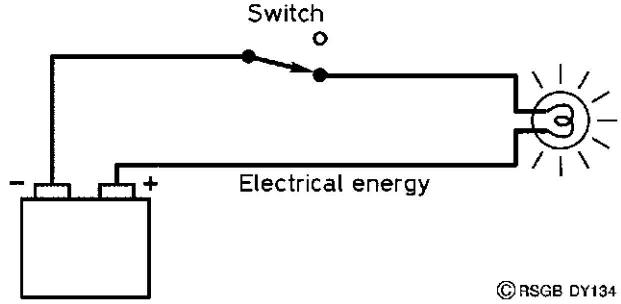

Figure 1 The light bulb converts electrical energy to electromagnetic light energy

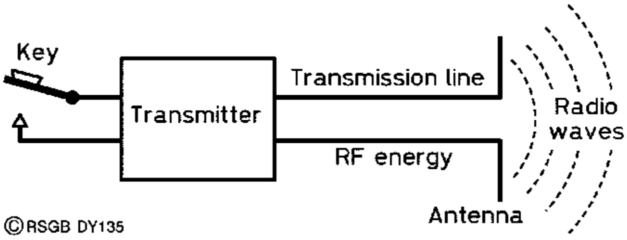

Figure 2 The antenna converts RF current to electromagnetic radio waves

cable. As you might expect, any feeder cable has a DC resistance caused by the resistance of the copper wire from which it is made. It also has an AC resistance, caused by the capacitance between the centre conductor and the braid, and by the inductance of the cable itself. This means that the feeder has an *impedance*, which is constant for the particular type of cable. This is what we call the *characteristic impedance* and, for the cables used in most amateur radio applications, it is 50 Ω.

If this impedance can be made the same as that of the aerial (it is already the same as the impedance of the transmitter output), then the transfer of energy will occur with minimum loss. If the aerial and cable impedances are *not* the same, then there is a *mismatch*, which causes some of the RF energy to be reflected back towards the transmitter. We now have a situation where RF energy is flowing along the cable from the transmitter to the aerial (the *forward* wave) and, at the same time, flowing from the aerial to the transmitter (the *reflected* wave). The two waves interact along the cable and form a stationary pattern of voltage and current. The pattern is known as a *standing wave*, and can be visualised from the waveforms in **Figure 3**. The ratio of the maximum voltage to the minimum voltage on a given wave defines the *voltage standing-wave ratio* (VSWR), or just *standing-wave ratio* (SWR) for short.

The SWR meter is easy to use. It is positioned in the feeder between the transmitter and the aerial tuning unit (ATU) if there is one, or between the transmitter at the aerial otherwise. Most meters have a single meter and a four-position switch. The transmitter is first keyed and, with the switch in

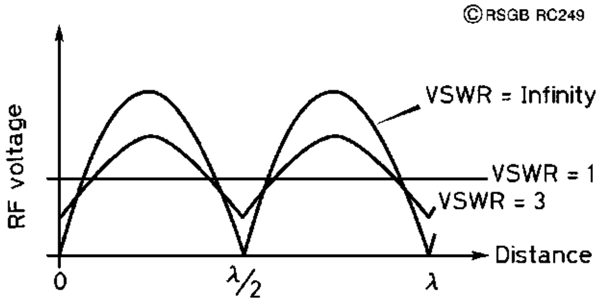

Figure 3 Standing waves on a transmission line

the 'calibrate' position, the sensitivity control is adjusted to give full-scale deflection on the meter. The switch is then changed to 'forward' to read the forward power, to 'reflected' to read the reflected power, and to 'SWR' to read the value of the standing-wave ratio. When there is no reflection (see Figure 3), the meter should read 1:1 or, simply, 1.

Most SWR meters remain in the feeder line while the transmitter is operating, so the condition of the aerial and feeder can be constantly monitored. Problems with the aerial (such as water entering the feeder at its junction with the driven element) are immediately shown up. Without the use of the SWR meter, the situation would slowly deteriorate over several months and you would be left wondering why so few stations were answering your calls!

50 A standing-wave indicator for HF

Introduction

The standing-wave ratio (SWR) meter shows how well the aerial system, including the feeder, is *matched* to the output of the transmitter. This design does not *measure* SWR, but it gives an indication of when the SWR is minimum for a given system of aerial and feeder. The design is usable on the HF bands from 1.8 to 28 MHz, and can be used at 50 MHz with reduced sensitivity.

How it works

There are two types of wave in any feeder: the *forward* wave, which travels from the transmitter to the aerial; the *reflected* wave, which travels back to the transmitter from the aerial. The presence of a reflected wave is evidence that some of your transmitted power is not being radiated, but is being returned to the transmitter to be lost as excess heat. When aerial and feeder are perfectly matched, there is *no* reflected wave, and all the power from the transmitter is radiated.

Referring to the circuit of **Figure 1**, a tiny fraction of the signal is removed by the transformer, T1, and by the capacitors, VC1 and C1. It is then detected by the germanium diodes, D1 and D2, and any residual RF removed by the capacitors, C2 and C3. The currents through the diode and meter (depending on the position of switch, S1) represent the forward and reflected signals. VR1 acts as a sensitivity control for the meter.

It pays to shop around for a suitable meter. Surplus types from tape recorders and hi-fi equipment are usually ideal for this purpose. A new one would cost several pounds. The more sensitive the meter, the more sensitive your indicator will be. Meter sensitivity is given by the current required to give full-scale deflection (FSD) of the pointer. One with an FSD of between 50 and 200 micro-amps (μA) is suitable for this circuit. The higher the FSD, the less sensitive the circuit.

Construction

The meter circuit and the sampling transformer (see **Figure 2**) are built and mounted on Veroboard of the copper-strip variety. It simplifies construction

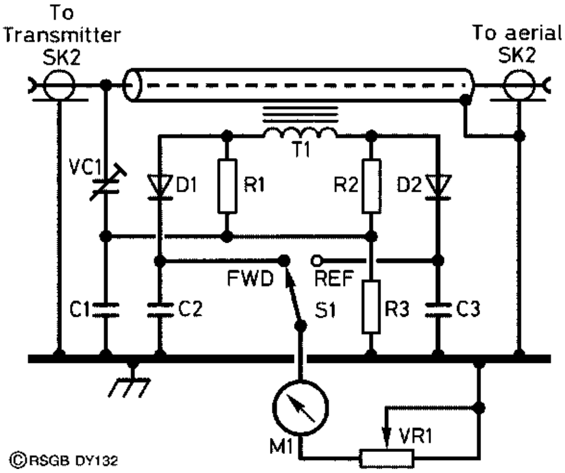

Figure 1 Circuit diagram of the SWR meter

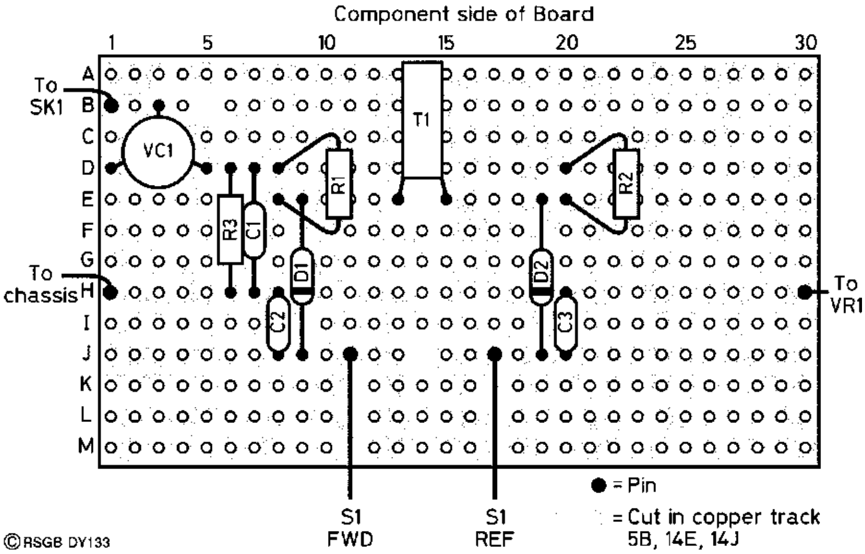

Figure 2 Component layout of the SWR meter

but reduces the operational range of the meter to below 30 MHz because of the capacitive coupling between strips. The board has 13 strips by 30 holes, although you can reduce this if you have a smaller case.

Firstly, cut the tracks at the three points shown. Then insert and solder Veropins for connections to the external components, the switch, variable resistor and the meter. Solder in the components starting with the resistors and followed by the capacitors and the diodes, ensuring that the diodes are inserted correctly.

Now you have to wind the transformer, T1, on a small toroidal ferrite core. Wind the secondary with 15 turns of 36 SWG enamelled copper wire, spaced evenly over about two-thirds of the former. The turns should not overlap, and considerable care must be taken; the wire is very thin, will kink easily and will break if you apply too much tension. The ‘primary’ is an 8 cm length of 50 Ω coaxial cable which passes through the toroid on its way between the input and output connectors. The braid of the cable is connected to the case at only *one* of the connectors (see Figure 1); this prevents the screen and the metal case between the two sockets forming a single, shorted turn.

The ends of the secondary winding must be carefully stripped of their enamel with sandpaper, before attaching the toroid to the board with cotton or nylon fishing line. On no account must wire be used for this!

Solder the secondary connections of T1 to the board and thread through the coaxial cable ready for soldering to the connectors.

The case used was an aluminium box (Maplin LF02C), but any suitable metal box could be used. Aluminium is preferable, as it is easily drilled with simple tools. Use standoff insulators to mount the board in the case. Once this has been done, the leads from the board to the chassis-mounted components can be soldered. So can the coaxial cable passing through the toroid. Make the lead from the input socket to VC1 as short as possible.

Setting up

You will need a 50 Ω dummy load and a transmitter to set up your indicator. Connect the transmitter to SK1 and the dummy load to SK2. Set the toggle switch, S1, to *forward* and the sensitivity control, VR1, to mid-travel. Switch on the transmitter, and set VR1 for maximum meter deflection. Switch to *reflected* and adjust VC1 until the reading is minimum (ideally zero). This completes the setting up!

Using the indicator

For setting up an aerial, connect your circuit between the transmitter and the cable leading to the aerial. With S1 in the *forward* position, key the transmitter and adjust VR1 for maximum reading on the meter. Switch to *reflected*, and then adjust your ATU to give minimum reflected power. If your adjustments are to be made to the aerial itself, to give minimum reflected power, you must make a note of the reflected reading, switch off the transmitter, make a change to the aerial, key the transmitter, and note whether the reflected power is greater or less than before. Then, make more changes to the aerial. **Never adjust your aerial with the transmitter on**. Make your adjustments on an unused frequency, and do it as quickly as possible, thus avoiding (or minimising) interference to other stations.

Parts list

Resistors: all 0.25 watt, carbon 5% tolerance
(or Maplin 0.6 watt metal film)

R1, R2	27 ohms (Ω)
R3	2.2 kilohms (kΩ)
VR1	10 kilohms (kΩ) linear

Capacitors

C1	220 picofarad (pF) disc ceramic 50 VDC
C2, C3	0.1 microfarad (μF) disc ceramic 50 VDC
VC1	20 picofarad (pF) trimmer

Semiconductors

D1, D2	OA91 germanium

Additional items

S1	Single-pole changeover (SPDT or SPCO)
SK1, SK2	Coaxial sockets to suit station standards
	Veroboard – 13 strips by 30 holes
	Veropins (7 off)
	Amidon FT 50–43 ferrite toroid
	Meter, less than 200 μA FSD
	36 SWG enamelled copper wire
	Short length of UR43 or RG58 coaxial cable
	Insulated stranded wire
	Aluminium box
	Standoff insulators for mounting the board
	Knob for the sensitivity control

51 A moisture meter

Introduction

Dry rot (*Merulius Lacrymans*) can strike havoc in buildings, causing the timbers to decay and crumble to dust – hence the term *dry rot*. Wood is attacked only if its moisture content rises above 20%.

Construction

The circuit of the moisture meter is shown in **Figure 1**. The two probes touch the wood, and the current that flows between them depends on the moisture content of the wood. If the moisture is sufficiently high, the current, after amplification, will be enough to light the LED.

The meter can be made on a piece of plain matrix board (no copper strips), as **Figure 2** shows. The board is big enough (10 cm by 2.5 cm) to accommodate the PP3 battery, taped on. No case is needed, unless you want

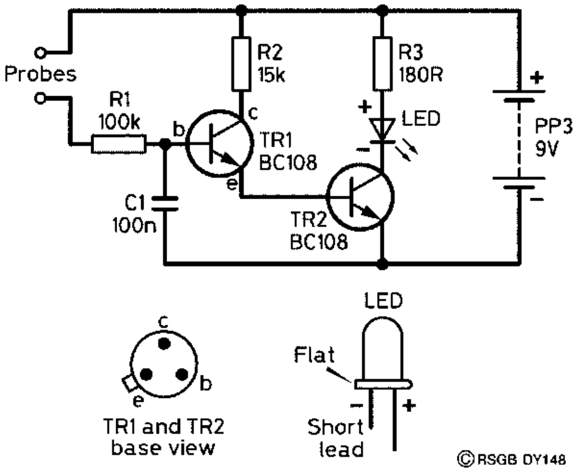
Probes
R1
100k
b
c
e
TR1
BC108
R2
15k
R3
180R
+
LED
−
TR2
BC108
C1
100n
+
PP3
9V
−
LED
c
b
e
TR1 and TR2
base view
Flat
Short
lead
−
+
©RSGB DY148

Figure 1 Circuit diagram of a moisture meter

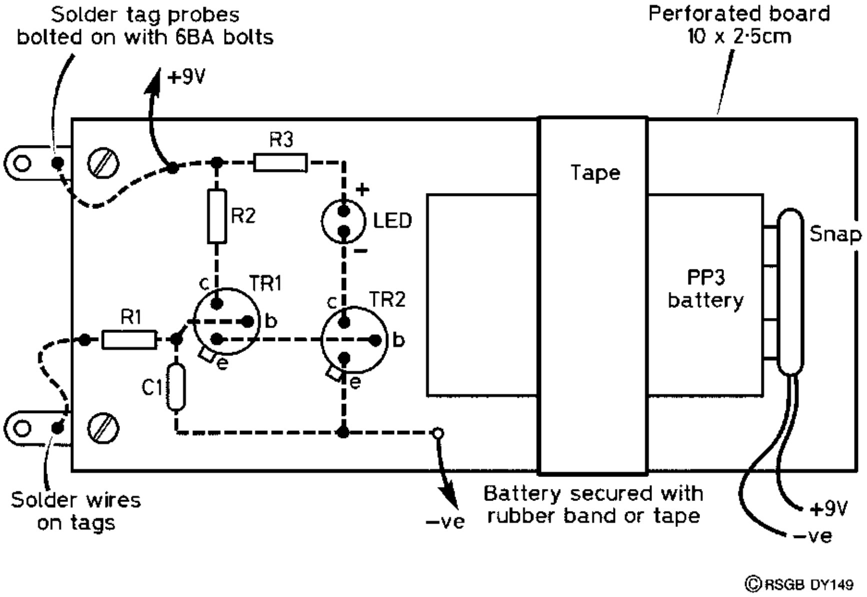
Solder tag probes
bolted on with 6BA bolts
+9V
R3
R2
+
LED
−
TR1
c
b
e
TR2
c
b
e
R1
C1
Tape
Perforated board
10 x 2·5cm
PP3
battery
Snap
Solder wires
on tags
−ve
Battery secured with
rubber band or tape
+9V
−ve
©RSGB DY149

Figure 2 Moisture meter, component layout

to leave the meter in a damp location for a protracted period! Make sure that the transistors and LED are mounted correctly. In Figure 2, the connections as shown to TR1 and TR2 are illustrated as if the transistors were transparent. An on/off switch is not really necessary, as only a very small current flows when the probes do not touch anything. Use solder tags, screwed to the board, to act as probes.

After the assembly is completed, check your circuit one final time, and then connect the battery. Nothing should happen at first. If you lick your forefinger and hold it across the probes, the LED should light.

Using it

The prototype was compared with a commercial moisture meter, and the LED lit when the moisture was around 20%. This was quite fortuitous, as the point at which the LED lights depends both on the separation of the probes and on the gain of the two transistors.

In addition to searching for dry rot, the instrument may be used to monitor the moisture in the soil of household plants. In this case, probes made of 16 SWG copper (*not* enamelled) should be soldered on to the two tags, and should penetrate the soil to a depth of several centimetres, and R1 may require adjusting so that the LED extinguishes if the soil is too dry, and lights if the soil is sufficiently moist. If you wanted to leave the meter with the probes in the soil, an on/off switch *would* be necessary.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	100 kilohms (kΩ)
R2	15 kilohms (kΩ)
R3	180 ohms (Ω)

Capacitor

C1	0.1 microfarad (μF) polyester

Semiconductors

TR1, TR2	BC108
LED	Any shape or colour will do

Additional items
- PP3 battery and connector
- Solder tags (2 off) for probes
- Matrix board 10 cm by 2.5 cm

52 Simple aerials

Introduction

The performance of *any* receiver or transceiver, no matter how expensive it is, is limited by the aerial that feeds it. Two of the most frequently asked questions are:

- Which is the best sort of aerial to use?
- Where is the best place to locate an amateur radio aerial?

To answer these questions, you must ask yourself what sort of operation you want to do. Are you interested in local, chatty contacts on the lower bands or VHF, or are you more disposed towards long-distance (DX) contacts, and on what band?

A house with a moderately sized garden is assumed in the diagrams here, to illustrate the configurations of some simple aerials. You would not need all these aerials festooned around your house, because one or two would be sufficient for your needs. The problems incurred by properties with more restricted space will be covered later.

VHF aerials

For VHF operation, the aerial should be mounted as high as possible, either on a mast or on a chimney. For all-round coverage on FM and the local repeaters, a vertical colinear is a good choice. For SSB and CW DX operation, a horizontal rotatable beam is needed. If satellite working is envisaged, you will need to contemplate mounting an elevator on top of your rotator, so that your beam can point in *any* direction, including vertically upwards! An advantage of satellite working is that the aerials do not *necessarily* have to be up in the air, provided you have a relatively uncluttered site. Your rotator and elevator can be at ground level, which is good!

If the VHF aerial is mounted on the chimney, use a double mounting bracket, particularly if you have a beam and rotator. Keep the TV, broadcast FM and amateur aerials as far apart as possible, and keeping the feeders separated is also a good plan.

The dipole aerial

One of the simplest types of aerial for single-band operation is the half-wave dipole. (The name ‘dipole’ simply means ‘two poles’ or ‘two elements’, and

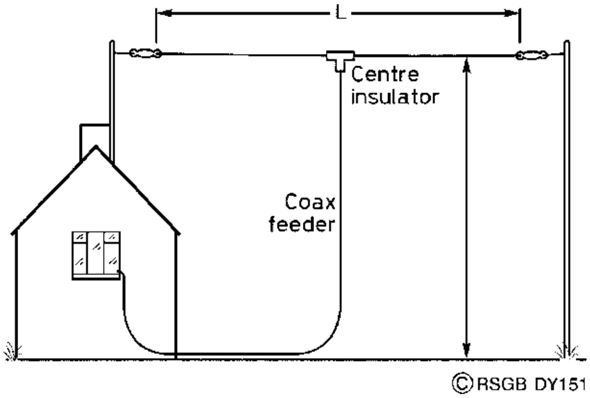

Figure 1 Layout for a dipole aerial

in this case the total length of the dipole is approximately half a wavelength at the operating frequency.) It is usually fed in the centre by coaxial cable as shown in **Figure 1**. The length of the dipole for the lowest frequency in each band is shown in **Table 1**. Normally, the length of the aerial will be 'trimmed' to be tuned to the centre frequency of the part of the band in which you will operate. This is done using the data in the right-hand column of Table 1. As an example, suppose you wanted your aerial to be resonant at 3.7 MHz. The table gives an overall dipole length of 42.86 m for 3.5 MHz. To resonate the aerial 200 kHz *higher*, then this length must be shortened by 2 × 0.595 m = 1.190 m. Your dipole would thus be 41.67 m long. Remember to allow extra wire for fixing the dipole ends to the insulators.

Table 1 Dipole lengths for lowest frequency of each band and the length to be trimmed from each to raise the resonant frequency by 100 kHz

Band (MHz)	*Dipole length (m)*	*Trim each end (mm/10 kHz)*
1.8	83.33	2190
3.5	42.86	595
7	21.43	150
10	14.85	70
14	10.71	35
18	8.33	20
21	7.14	15
24	6.03	12
28	5.36	10
50	3.00	6

On the lower-frequency bands, the lengths become rather large. In this case, you can 'bend' your dipole, as illustrated in **Figure 2**. The length of wire required to give an acceptable value of SWR (less than 2:1 on transmit) may need to be different from the calculated value, so be prepared to experiment!

Dipoles are single-band aerials, although they will often work acceptably on the third harmonic of their design frequency: a 7 MHz dipole often operates reasonably well on 21 MHz. It is possible to operate several dipoles in parallel, as **Figure 3** shows. Interaction between the elements can occur if the spacing between them is less than about 10 cm. A multi-band dipole, as shown in Figure 3, has the elements separated with plastic spacers, and drooping ends to produce maximum spacing between the elements' ends.

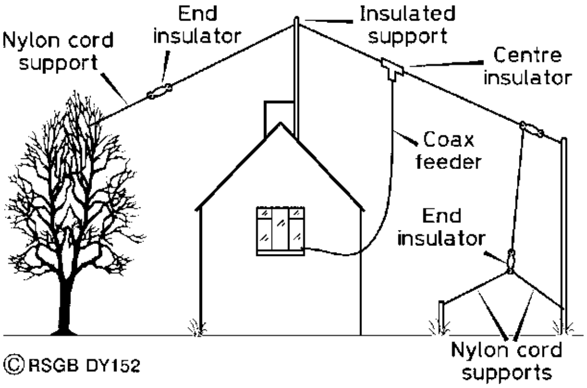
End insulator
Insulated support
Nylon cord support
Centre insulator
Coax feeder
End insulator
Nylon cord supports
©RSGB DY152

Figure 2 Possible layout for a dipole aerial in a confined space

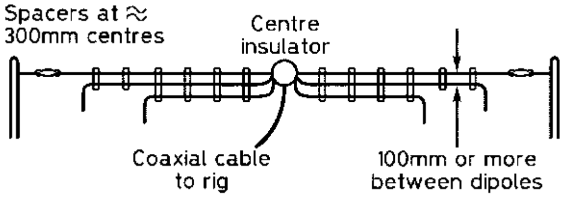
Spacers at ≈ 300mm centres
Centre insulator
Coaxial cable to rig
100mm or more between dipoles

©RSGB DY153

Figure 3 Multi-band dipole aerial

The long-wire aerial

This aerial is simple, cheap, easy to erect, and suits most houses and gardens, as **Figure 4** shows. Using an aerial tuning unit (ATU), an end-fed long wire can function on several bands when used with a set of radials or a counterpoise. **Figure 5** illustrates the setup. The length of the aerial will determine the bands which will be covered.

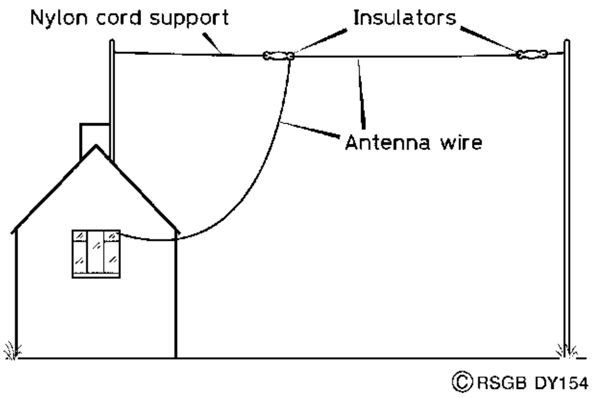
Nylon cord support
Insulators
Antenna wire
©RSGB DY154

Figure 4 Long-wire or inverted-L aerial

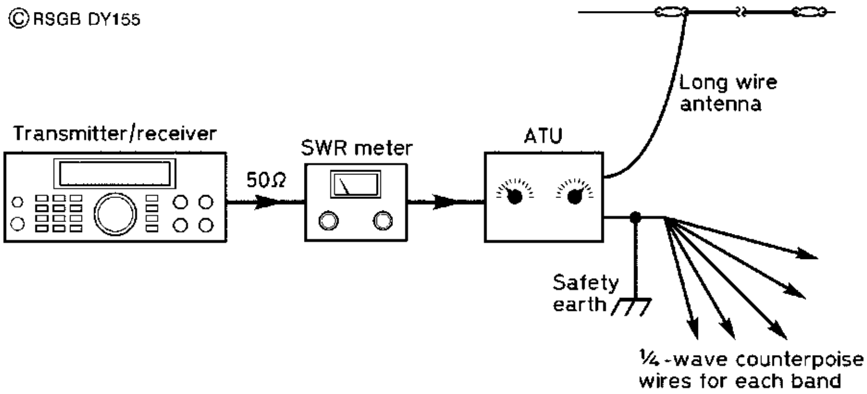
©RSGB DY155
Long wire antenna
Transmitter/receiver
SWR meter
ATU
50Ω
Safety earth
¼-wave counterpoise wires for each band

Figure 5 How to connect a radio to a long-wire aerial

A wire length of 10.5 m will work on the 40, 30, 17, 15 and 12 m bands.

A wire length of 15.5 m will work (with an ATU) on the 80, 40, 20 and 12 m bands and possibly (depending on your ATU) on the 17 and 15 m bands.

A wire length of 26.5 m will operate on all bands, but may be difficult to load on 10 m.

The wire lengths given here may need some adjustment because of the geometry of your particular house and garden. For receive-only purposes, the lengths are far less critical.

In general, you cannot get a good radio-frequency (RF) earth from a first-floor (or higher) shack. Unless a good RF earth exists within a small fraction of a wavelength of the transceiver, an *artificial ground* comprising a single $\frac{\lambda}{4}$ radial or counterpoise will be needed. You will need one counterpoise for each band you intend to use, and the wire can be concealed around the skirting-board of the shack, or under the carpet. Make sure that the free end

Table 2 Lengths of elements for vertical antennas, radials for verticals and counter-poises for end-fed long wire antennas

Band (MHz)	*Element length (m)*
1.8	39.66
3.5	20.40
7	10.20
10	7.14
14	5.1
18	3.96
21	3.4
24	2.95
28	2.55

of each counterpoise is well insulated; this point can carry a *very high voltage* when you transmit; anyone coming into contact with this can suffer very severe RF burns. Counterpoise lengths can be read from **Table 2**.

The vertical aerial

The single-band vertical aerial is sometimes used by DX operators because it has a low angle of radiation, which favours long-distance propagation. However, it must be sited clear of obstructions and must have a *good* counterpoise or radial system. Illustrations of the vertical aerial are shown, and the lengths of the vertical and radial sections are given in Table 2. The centre of the coaxial feeder is connected to the vertical section, and the braid to the counterpoise or radial system, which is made up of four or more wires buried just below the surface and joined together near the base of the aerial.

Cable entry to the house

Bringing coaxial cable into the house by an open window must be regarded as a temporary measure. Wooden window frames can be drilled, one hole for each feeder. Make the holes slope downwards from inside to out to prevent rain entering, and treat these with wood preservative. Leads from long-wire and inverted-L aerials should be kept separate from other cables.

Alternatively, a plastic pipe large enough to take all your feeders could be fitted into the brickwork (again, sloping downwards towards the outside). You may want to let a friendly builder do this for you.

53 A breadboard 80 m CW transmitter

Introduction

In the early days of radio, many circuits were built on a wooden baseboard, the parts being screwed down on the board. This was called breadboard construction, because it was a breadboard that was frequently commandeered for the process! Wives have always been generous in this respect, it appears!

This circuit was originally designed by GM3OXX, and became known as the *Oner*, because it was built on a circuit board one inch square! The circuit appeared in the G QRP Club journal *Sprat* and, since that time, many hundreds of *Oner* circuits have been built and used on the air. It is a well-proven circuit.

The transmitter has no tuned circuits in the power amplifier (PA) and thus has a rather high *harmonic* content. It **must** be used with the ***low-pass filter*** described elsewhere in this book. Without the low-pass filter, interference will be caused to other stations.

Simple aerial changeover switching is provided, which allows this circuit to be used with any of the 80 m receivers, such as the *Colt*, described in this book. It can also be used with any kit or commercial receiver for the 80 metre band.

The circuit

The transmitter circuit is shown in **Figure 1**. TR1 is a crystal oscillator, the frequency of which is controlled by crystal X1. A small trimmer capacitor, TC1, is added to allow the frequency of X1 to be varied by a small amount. If adjustment of this trimmer is made possible from the front panel, it is useful to adjust the transmit frequency to avoid other stations already on the crystal frequency. The collector load resistor, R2, of the oscillator transistor, TR1, determines the power output; a value of 3.3 kΩ seems to work well in producing an output of 3 watts.

TR1 is directly coupled to TR3, a VMOS transistor (a type of field-effect transistor (FET)). This acts as the power amplifier (PA) stage. TR3 should

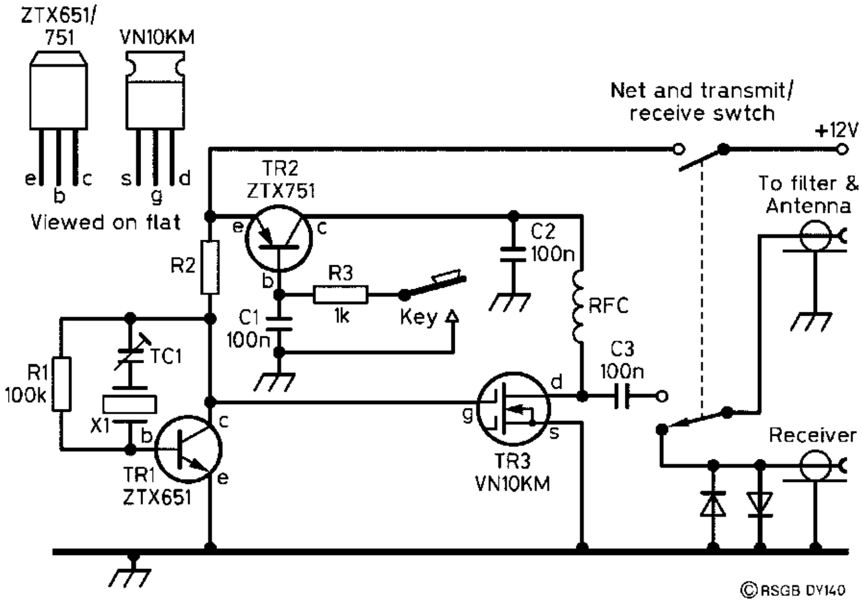
ZTX651/
751
VN10KM
e
b
c
s
g
d
Viewed on flat
Net and transmit/
receive swtch
+12V
TR2
ZTX751
e
c
b
C2
100n
To filter &
Antenna
R2
R3
1k
Key
RFC
C1
100n
R1
100k
TC1
X1
b
c
e
TR1
ZTX651
C3
100n
d
g
s
TR3
VN10KM
Receiver
©RSGB DY140

Figure 1 Circuit diagram of the breadboard transmitter

give about 3 W output, which is then coupled to the output by C3. The radio-frequency choke (RFC) providing the drain load of TR3 is simply a few turns of wire on a ferrite bead.

TR2 is an interesting addition to the circuit. It is used as a switch to 'key' the PA, TR3. The transmitter *could* be built without TR2, just placing the Morse key between the top of the RFC and the 12 volt supply. Adding TR2 is helpful, because it means that one side of the Morse key can be grounded (always a good thing), and some degree of *shaping* of the output RF waveform is provided by R3 and C1. This makes the transmission sound a little better and reduces the possibility of spurious frequencies being generated and transmitted. TR2 is a pnp transistor; note that it is the *emitter* of this transistor which is connected to the positive side of the supply.

Some form of changeover switching is needed for the aerial. A double-pole changeover toggle switch can be used. See the chapter on switches, later on in this book. One pole is used to switch the aerial between transmitter and receiver; the other pole is connected in the 12 volt supply line, and is labelled RECEIVE/TRANSMIT – NETTING. Its use will be described later. In this simple circuit, the PA cannot work when the key is open, because the key breaks its supply (via the RFC). When the key closes, TR2 switches on and applies the 12 volt supply to the top of the RFC. C2 is a *decoupling* capacitor, which prevents any residual RF signals at the top of the RFC reaching TR2.

Building

The prototype was put together as follows. Take a piece of plain printed-circuit board (PCB) measuring 5 cm by 4 cm. Then, with a *new*, *sharp blade* in a junior hacksaw, draw the blade horizontally across the surface of the copper in order to make a pattern of 6 squares along the 5 cm side and 5 squares along the 4 cm side. No more pressure should be applied than is necessary to cut through the copper! All the parts will be soldered on these pads in a form of surface-mount construction. To do this, each active pad (i.e. one that is going to have a component soldered to it) needs to be *tinned*. This means coating the pad's surface with solder, and is carried out as follows. Place the hot tip of the soldering iron on to the pad, and hold it there for a second or so. Then, with the tip still in place, touch the end of your reel of solder *on the pad*, not the tip of the iron. The solder should flow evenly all over the pad, and you can remove the iron. The solder should solidify in a rounded, shiny blob! This provides a good surface for making soldered joints.

To join component leads to the pads, cut each lead about 1 cm long, and then bend the last 2 mm at right angles to the rest of the lead. As you did before, tin the 2 mm length of each lead. Place the tinned portion on to the pad, and place the tip of your iron *on the pad*, close to the lead. The solder on the pad and on the lead will melt and run together; remove the iron and *hold the component still* until the solder solidifies. When the joint has cooled, give the lead a gentle tug to make sure you have a good joint (a good *mechanical* joint is usually a good *electrical* joint, too!). Each transistor straddles three pads, so the centre lead will need to be shorter than the other two. Take care here to get the lead lengths right – if you do, you will be surprised how much more firmly the transistor is held than if you just botched the lead lengths by bending them to fit! Make sure the connections to the transistors are correct.

Winding the RFC is quite simple. Seven (or more) turns of thin (32 SWG) enamelled copper wire are threaded through a small ferrite bead. This requires care, because the bead is small and the wire is thin. Trim the ends to within about 1 cm of the bead, remove the enamel carefully with sandpaper and tin the bare ends, prior to soldering the choke to the board.

After completion of the wiring, check the circuit against Figure 1. Breadboarding a circuit like this has its advantages, but it can have disadvantages, too. One of these disadvantages is that it can make circuit checking difficult. For a simple circuit like this, it is not too bad! Check that no solder has run between the pads. Plug in your crystal for the 80 m band, and connect up the 12 V power supply. Do not connect the Morse key yet and do not switch on the power.

Testing and operating

Clear your workbench of all metallic objects, slivers of copper, bits of wire, etc., switch on your power supply. With an external receiver, listen on and around your crystal's frequency for a signal. Remember that the oscillator runs all the time and, because you haven't yet connected the Morse key, your receiver is close enough to pick up the signal from the oscillator. This confirms that your oscillator is running. Switch off.

Connect the station aerial to the transmitter's aerial socket, and the receiver to the transmitter's receiver socket. Connect the Morse key, put the Receive/Transmit–Netting switch in the receive position and switch on. You should be able to hear stations in the normal way. Now put the switch in the Transmit–Netting position. Signals in the receiver should almost disappear, as the circuit has disconnected the receiver's aerial.

Tune the receiver until you can hear your own crystal oscillator signal. This is known as *netting*, tuning your receiver and transmitter to the same frequency. Pressing the key will now transmit your signals when the switch is in the Transmit–Netting position; switch back to the receive position to listen for stations answering your call. As soon as you are happy that your circuit is functioning properly, **you must build the low-pass filter circuit** before using the transmitter regularly.

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	100 kilohms (kΩ)
R2	See text
R3	1 kilohm (kΩ)
Capacitors	
C1, C2, C3	100 nanofarads (nF), or 0.1 microfarad (μF)
TC1	3–60 picofarads (pF) trimmer
Semiconductors	
TR1	ZTX651
TR2	ZTX751
TR3	VN10KM
Additional items	
RFC	7 turns of 32 SWG enamelled copper on a ferrite bead
Switch	Double-pole changeover (DPDT or DPCO) toggle
Crystal	For 80 m band
Crystal holder	HC25 type
Sockets	According to station fittings

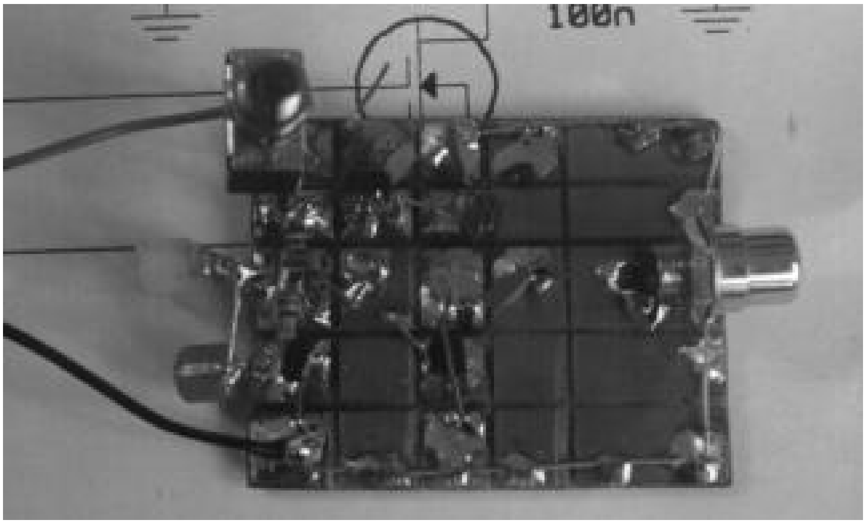

54 A 7-element low-pass filter for transmitters

Introduction

The simpler the transmitter, the more likely it is to radiate *harmonics* of its fundamental frequency. Harmonics are integral multiples of the frequency on which the transmitter is *designed* to operate. If you think you are transmitting on a frequency, *f*, for instance, you will also be radiating the harmonics of *2f*, *3f*, *4f*, . . . and so on. This results in your signals being heard on *several frequencies*, spread over a very wide frequency range. You are also contravening the terms of your licence. To avoid this, it is always advisable to use a *low-pass* filter between your transmitter and aerial. This is a filter which will *pass* your signal frequency, *f*, and all frequencies *below* it, but will *not pass* (attenuate) frequencies above *f* to any significant extent. The value of *f* is known as the *cutoff frequency* of the filter.

A design of 7-element low-pass filter

A 7-element low-pass filter (LPF) is so called because it has seven components, as the circuit diagram of **Figure 1** shows. Filters containing any odd number of elements are possible: a 3-element filter would

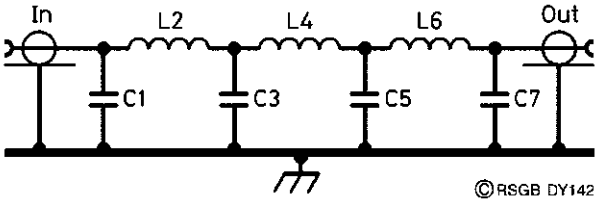

Figure 1 Circuit diagram of filter

comprise C1, L2 and C3 only, and is sometimes called a pi-network because the element disposition resembles the Greek letter pi (π); a 5-element filter would comprise C1, L2, C3, L4 and C5 only, and so on. In general, the more elements the filter has, the more effectively it attenuates signals above *f*.

The circuit of Figure 1 is designed to have an input and an output impedance of 50 Ω, which means that it can be placed in the aerial feed of any common transmitter. Filter design is a very complex business, and is best left to the experts. One such expert is W3NQN, who produced a number of computer designs of LPF using commonly available (*preferred value*) capacitors, and aimed specifically for use on amateur frequencies. The results of this work are condensed into **Table 1**. The inductors are wound on standard toroidal cores, and their details are included in the table.

Table 1 Filter component values for each brand

Band metres	*C1,7 pF*	*C3,5 pF*	*L2,6 turns*	*L4 turns*	*Core type*	*Wire SWG*
80	470	1200	25	27	T37-2	28
40	270	680	19	21	T37-2	26
30	270	560	19	20	T37-6	26
20	180	390	16	17	T37-6	24
15	82	220	12	14	T37-6	24
10	56	150	10	11	T37-6	22

Making the filter

The filter was made originally as an adjunct to the *Breadboard 80 m CW transmitter*, which you will also find in this book. It uses the same constructional technique, based on a single piece of plain, copper-clad

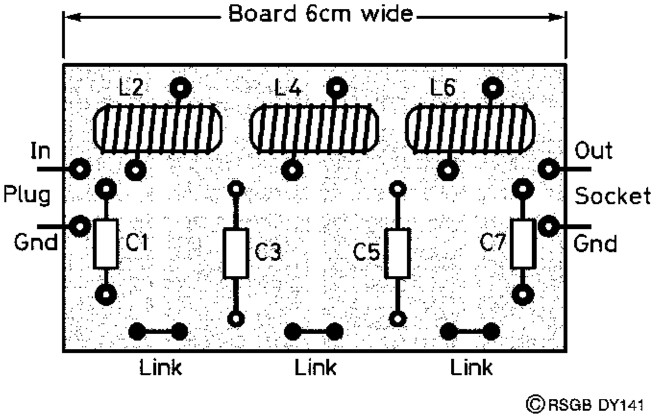

Figure 2 Layout and construction of filter

PCB, with ‘pads’ created by using a sharp blade in a junior hacksaw. The cuts in the copper are shown in the layout diagram of **Figure 2**. There are two ways of mounting the components: the first way is to drill small holes in each pad, as shown in Figure 2, and mount the components through the holes in the normal PCB manner; the second way is to solder the components directly to the pads, in the way that was described for the *Breadboard transmitter.*

Winding the inductors is quite simple. All you need to remember is that each time the wire passes through the core counts as one turn. Cut off the spare wire at each end of each coil to about 1 cm, scrape off the enamel with sandpaper, and tin the exposed copper. See the transmitter description if you are unsure of how to do this. Note the wire links between each of the lower pads, forming a solid ‘ground’ for the elements. The prototype had a plug and socket on the ends, to match the transmitter and aerial terminations.

The type of capacitor used in the design is not critical; the polystyrene type works well.

55 Radio-frequency mixing explained

Introduction

Mixers find widespread use in electronic circuitry. Many of the projects in this book, together with every TV set and radio in the home, contain mixer circuits – a good indication of their usefulness.

Confused?

Audio mixers (as used in recording studios and radio broadcast stations) are used to **add** or ‘balance’ the signals from various sources such as microphones, CD players, etc. These have nothing whatsoever to do with radio-frequency (RF) mixers, and should never be confused with them.

RF mixers and beat frequencies

Instead of **adding** signals (as in the audio mixer), the RF mixer **multiples them together.** As you might expect, this has an entirely different effect. The two signals entering the mixer *beat* or *heterodyne* with each other to produce *signals on other frequencies*. One example of this occurs in sound, when two musical notes of almost the same frequency are heard together. Instead of hearing two separate notes, the listener hears one note whose *intensity* (*loudness*) appears to increase and decrease. This intensity variation is called a *beat*, and its frequency is equal to the difference in the frequencies of the two original notes. The technique is used by musicians to tune their instruments. If one note is known to be a correct frequency, the other can be tuned to it by making the beat frequency as close to zero as is possible.

Multiplying together

The process of mixing presupposes that we have a device which will automatically multiply two signals together. Fortunately, this is easy; so easy, in fact, that it often occurs when we do not want it! Multiplying is achieved by any device which is *non-linear*; this means a device whose output is not a constant factor larger than its input, something that can be achieved by many electronic devices and circuits.

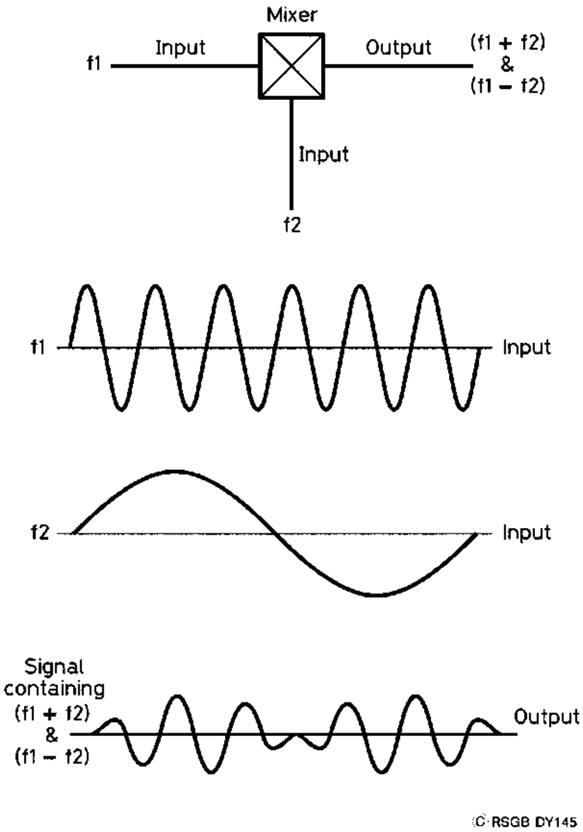

©RSGB DY145

Figure 1 The effect of multiplying (or mixing) two signals together

Let us look now at what a mixer does in concrete terms. Suppose two signals, of frequencies *f*1 and *f*2 go into our mixer. These signals are shown in **Figure 1**. Putting numbers in, to make the situation clearer, suppose *f*1 is 1.000 MHz and *f*2 is 160 kHz. The beat frequency is the *difference* of these: 1.000 MHz – 0.160 MHz = 0.840 MHz, or 840 kHz. A mixer *also* produces an output at the *sum* of these frequencies; in this case the new frequency would be 1.000 MHz + 0.160 MHz = 1.160 MHz.

Suppose you fed the output of your mixer, operating with these input frequencies, into a receiver and tuned around to find what frequencies were present. You would find two signals, one at 840 kHz and one at 1.160 MHz, showing that the two 'new' frequencies were very real!

In addition to drawing out the waveform of the resultant signal, as in Figure 1, we can draw the inputs and outputs on a frequency axis, to form a *spectrum* of the signal components. This is done in **Figure 2**. The top two diagrams show the input signals at *f*1 and *f*2. The bottom diagram shows the output signals in relation to the input signals. Depending on the type of mixer used, one or both of the input signals would be removed.

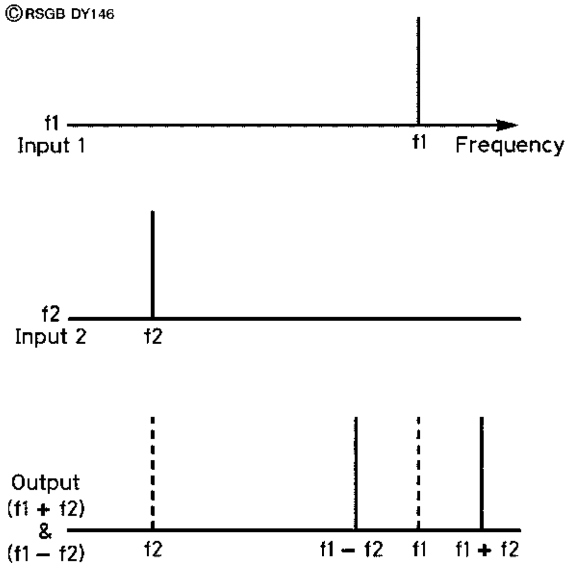

Figure 2 The result of multiplying or mixing two signals together as seen on a spectrum analyser

A mixer in every radio

Basically, a mixer is used to change a signal from one frequency to another, something it does without altering the characteristics of the incoming signal. If the incoming signal is amplitude modulated (AM), then the frequency-changed signal would be AM also. The same applies to FM, SSB, CW, and all other modulation forms you can think of. This explains why mixers are often called *frequency changers*.

Frequency changing is the key process in the type of radio known as a *superheterodyne* (or *superhet*). By mixing the incoming signal with a variable-frequency *local oscillator* as **Figure 3**, shows, the signal can be converted to the fixed frequency of a filter and amplifier. This is useful

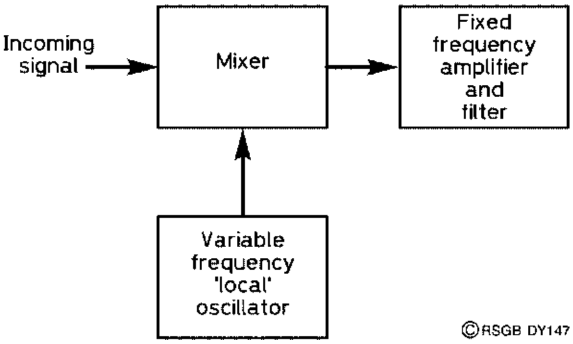
Incoming signal
Mixer
Fixed frequency amplifier and filter
Variable frequency 'local' oscillator
©RSGB DY147

Figure 3 The basic idea of a superhet receiver

because it is easier to make a very high-quality filter and amplifier at a single fixed frequency, than at a variable frequency.

All TV receivers and virtually all radio receivers (and transmitters) use mixers. Both the *Yearling* and *Colt* receivers (see the relevant projects) are superhets, and use mixers.

56 A voltage monitor for a 12 V power supply

Introduction

If for any reason, the stabilisation of your main 12 volt power supply unit (PSU) breaks down, it is possible that a voltage much higher then the nominal 13.8 V will be applied to your precious equipment. If you would like to know the instant that this occurred, and hence be able to switch things off before it was too late, then this circuit is what you need. It will give audible *and* visual indications if the voltage rises above 14.4 V, and a visual indication only if the voltage is reduced.

The circuit

The circuit uses three ICs and is shown in **Figure 1**. The circuit is powered by the PSU whose output is being monitored, and the circuit's immunity to supply line variations is secured by the 6 volt regulator, IC1. The heart of the circuit is IC2, an LM3914; it is a *bargraph driver*, which operates ten LEDs in a display resembling a thermometer – the string of lit LEDs increases in length as the voltage on pin 5 increases.

The input voltage range on pin 5 is 1.2 V maximum, making each LED correspond to one-tenth of this, which is 0.12 V, the *step size*. R1 and R2 act as a voltage divider, so that voltages of up to the maximum of 14.4 V may be applied to R1 without exceeding 1.2 V at pin 5. R3 sets the brightness of the LEDs and R4 determines the step size.

IC3 is an *opto-isolator*, a device containing an LED and a phototransistor in one package. This enables the piezoelectric sounder to operate without affecting the operation of the bargraph driver. The input to IC3 is provided by the voltage on D8, so that if any of the LEDs at or above D8 are lit, the

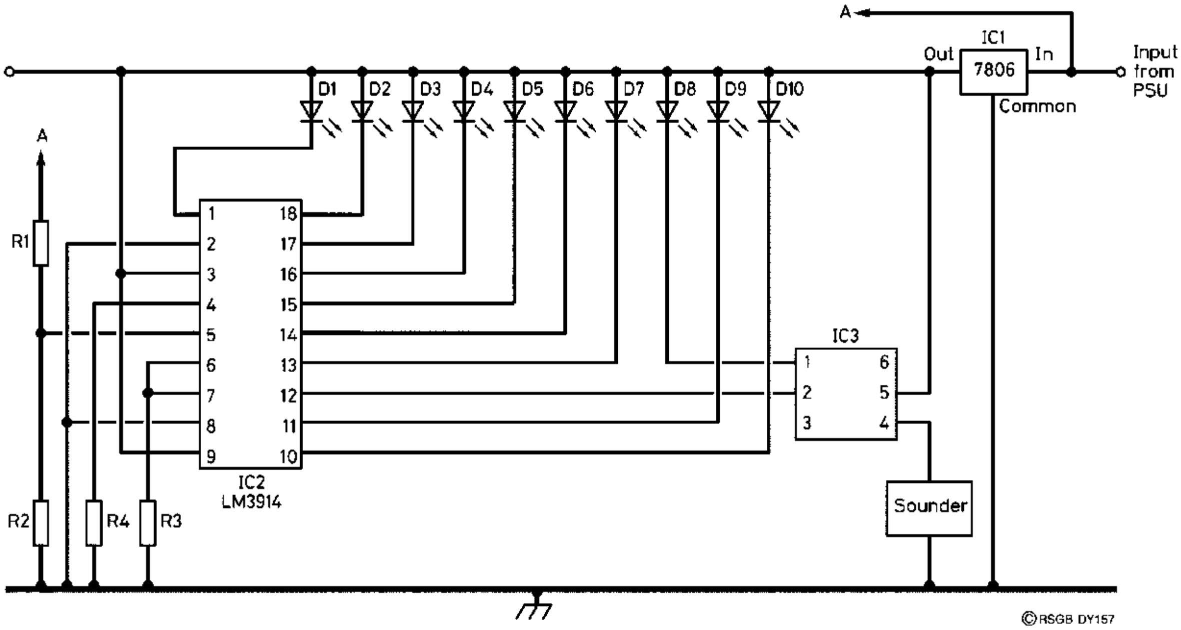
A
IC1
Out
7806
In
Input from PSU
Common
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
A
R1
R2
R4
R3
1
2
3
4
5
6
7
8
9
18
17
16
15
14
13
12
11
10
IC2
LM3914
IC3
1
2
3
6
5
4
Sounder
©RSGB DY157

Figure 1 PSU monitor, circuit diagram

sounder will operate, indicating overvoltage. Despite the fact that some of your other equipment might be damaged by this overvoltage, the monitor circuit itself is unaffected.

Construction

The prototype was built on two pieces of Veroboard of the copper-strip variety. The main (circuit) board measured 15 strips by 25 holes, and the display (LED) board measured 4 strips by 30 holes. Track cuts are necessary in each board. The correct places are shown for the main board in **Figure 2**, but there is sufficient flexibility in the layout of the display board for a prescriptive layout not to be needed. All the anodes of the LEDs are connected to the same strip, which makes things comparatively simple!

For the main board, insert the Veropins and the wire links first, and solder them to the copper strips. Then fit the IC holders and the resistors. Fit the IC holders with their notches towards the top of the board. When fitting the voltage regulator, IC1, note than the centre lead (the 'common' lead in Figure 1) does *not* go to position A3; the track should be cut at A3, and the common lead soldered at B3. This is indicated in Figure 2.

Wire up the LED board with its ten LEDs and 11 connecting wires, each about 10 cm long. This length depends on how far away from the main

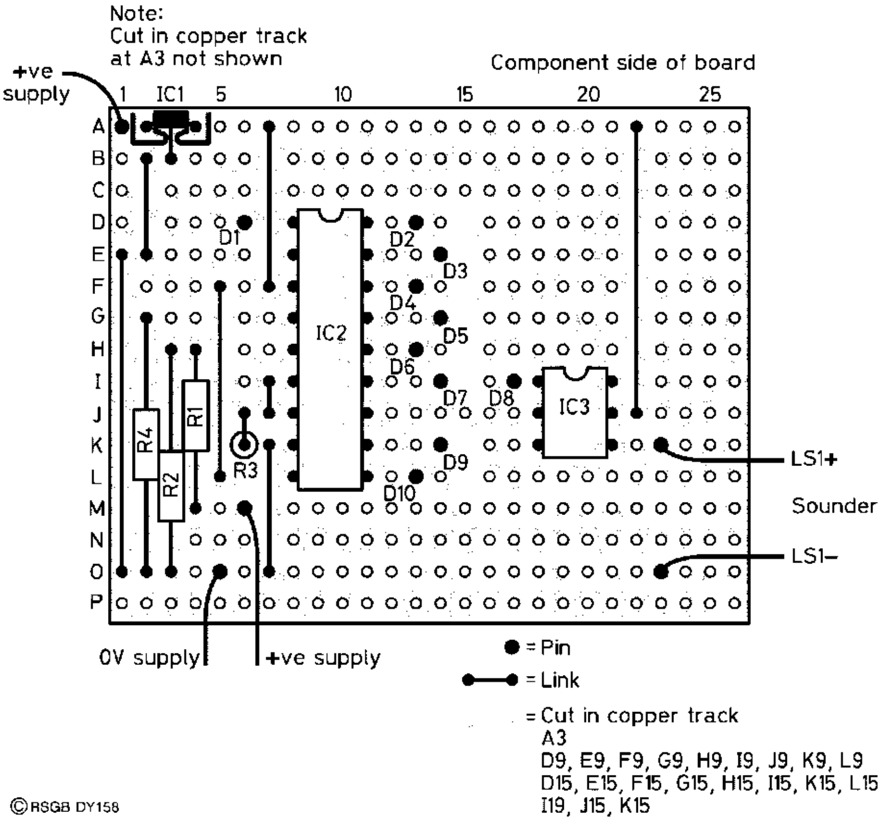
Note:
Cut in copper track
at A3 not shown
Component side of board
+ve supply
1
IC1
5
10
15
20
25
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
IC2
IC3
R1
R2
R3
R4
LS1+
Sounder
LS1−
0V supply
+ve supply
● = Pin
●—● = Link
= Cut in copper track
A3
D9, E9, F9, G9, H9, I9, J9, K9, L9
D15, E15, F15, G15, H15, I15, K15, L15
I19, J15, K15
©RSGB DY158

Figure 2 PSU monitor, PCB layout

board you are planning to mount the display. The LEDs should have different colours: three orange, four green and three red, to indicate ‘low’, ‘medium’ and ‘high’ voltage.

Now solder the other ends of these leads to the Veropins on the main board, making sure that the order is correct. Finally, connect the piezoelectric sounder to the main board; the polarity is important, so make sure the red lead goes to pin 4 of IC3 and the black lead to the ground rail. If you were careful to fit the IC holders with their notches in the correct positions, match these up with the notches on the ICs before pushing home the ICs *gently*. Check your circuit for solder bridges and unwanted pieces of copper swarf before screwing the small heat sink to the voltage regulator, IC1.

Testing

(a) **With a variable-voltage PSU**. Set the PSU for 10 V, connect the circuit and switch on. Increase the voltage slowly, and check that each LED lights up after the one before it. As the voltage exceeds about 14.4 V, the first red LED should light and the sounder should operate. If an LED does not illuminate, you should immediately suspect either a dry joint or an incorrect LED polarity. The voltage at which the first red LED lights can be adjusted by varying R1; increase R1 if the LED comes on too early; decrease R1 if the LED comes on too late.

(b) **Without a variable-voltage PSU**. For your 'variable supply', you can use several AA-type 1.5 V cells (or 1.2 V NiCad cells) in series. The voltages produced by a range of cells is shown in **Table 1** – because of the lower voltage of NiCad cells, more of them are needed to produce a given voltage.

Once the operation of the circuit has been checked, it can be fitted into a plastic or metal box. Only two connections are needed for the PSU. You may want to drill some holes in the case to increase the apparent loudness of the sounder.

Table 1 Test voltages available from batteries in series

No. of batteries	*Voltage, nicads*	*Voltage, dry cells*
7	8.4	10.5
8	9.6	12
9	10.8	13.5
10	12	15
11	13.2	–
12	14.4	–
13	15.6	–

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	11 kilohms (kΩ) – see text
R2	1 kilohm (kΩ)
R3	1.2 kilohms (kΩ)
R4	18 kilohms (kΩ)

Semiconductors

IC1	L7806
IC2	LM3914
IC3	Opto-isolator – Maplin code WL35Q
D1–D3	3 mm LED, orange
D4–D7	3 mm LED, green
D8–D10	3 mm LED, red

Additional items

LS1	Piezoelectric sounder, wire leads
	6-pin DIL socket for IC3
	18-pin DIL socket for IC2
	Veroboard – two pieces for main and display boards, see text for sizes
	Veropins
	Heat sink for IC1
	Single-core insulated wire for links
	Insulated stranded wire for interconnecting the boards
	Case as required

57 A 1750 Hz toneburst for repeater access

Introduction

Repeaters across the UK and much of Europe need an access tone to switch the transmitter from standby ready for use. Commonly, this is a 1750 Hz tone of duration no greater than half a second. Although many UK repeaters may now be accessed using the *continuous tone-coded squelch system* (CTCSS – see the *RSGB Yearbook*), you may wish to access a repeater whose CTCSS frequency you don't know; in this case, using the universal 1750 Hz tone will gain you access. Commercial transceivers are usually fitted with an automatic toneburst, but if you are using a home-made design, then you may want to incorporate this little circuit.

Warning

This circuit uses a member of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current

and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from destruction while it is connected to the battery.

Circuit description

The circuit is shown in **Figure 1**. The tone is generated by an integrated circuit oscillator (IC1), whose frequency is controlled by a *ceramic resonator*, XL1. Its frequency is very high, and is divided down to the 1750 Hz needed by the same chip. The ceramic resonator is designed to operate at 455 kHz, the intermediate frequency of many receivers. Because all divider circuits use powers of 2, we need the oscillator to run at 448 kHz so that when it is divided by 256 ($256 = 2^8$), we end up with 1750 Hz. Try it on your calculator:

$$\frac{448\ 000}{256} = 1750.$$

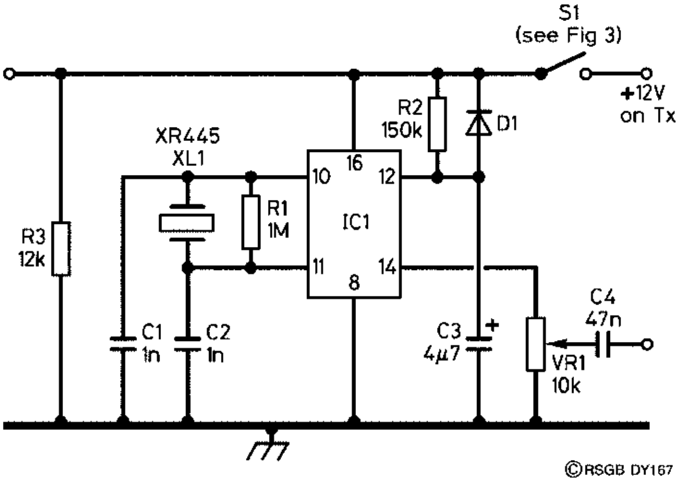

Figure 1 Toneburst module, circuit diagram

We use C1, C2 and R1 to *pull* the frequency of the oscillator away from 455 kHz to 448 kHz. The divider chain has eight *counters* in it, and each counter divides the frequency of the signal it sees by two, giving the final division of 256.

This counting process can be stopped at any time by taking the voltage on the *reset* pin (pin 12) up to the supply voltage. When the circuit is switched on by closing S1, pin 12 is at 0 V because C3 is discharged. The oscillator runs, producing the output frequency of 1750 Hz. As time progresses, C3 charges up through R2 and the voltage on pin 12 rises. When this has risen sufficiently, and in a time determined by the values of C3 and R2, the counter resets and stays in the reset state; no division takes place and there is no output. The duration of the toneburst is thus governed by C3 and R2.

When S1 is opened, the circuit is switched off, and C3 is discharged through D1 and R3, ready for the next toneburst. If you have used a repeater, you will know that a toneburst is needed **only** to activate a repeater in the standby condition; it is not needed once a contact has been established.

VR1 adjusts the amplitude of the tone fed to the microphone, and C4 prevents any voltage that may be present on your microphone connector from damaging the integrated circuit.

Construction

The prototype circuit was built on Veroboard of the copper-strip variety, measuring 20 holes by 14 strips. The layout is shown in **Figure 2**. Make the track cuts first, and check that there are no slivers of copper wedged between adjacent tracks. Then, solder in the IC socket (with the notched end

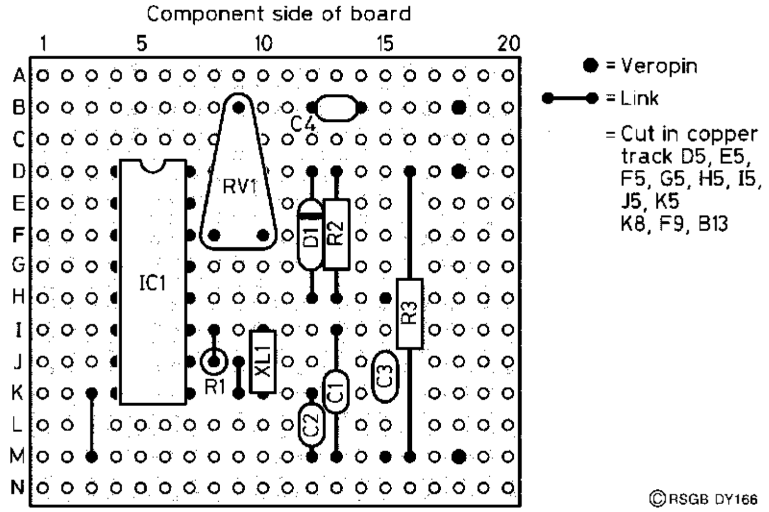

Figure 2 Veroboard component layout

facing towards track A), the wire links and the three Veropins. Having done this, solder in the resistors, capacitors and diode, making sure that D1 and C3 are the right way round! Using your best soldering technique, solder in the ceramic resonator quickly, to prevent heat damage. Recheck your circuit, check for solder splashes and bridges, and then gently insert IC1 into its socket, matching up its notch with that of the socket.

Testing

Set VR1 to half-way and connect a crystal earpiece to the output; apply power to the circuit. You should hear the tone, lasting for about half a second. If there is no tone, disconnect your circuit from the power supply, and check for dry joints in the vicinity of pin 12. Is the diode, D1, the correct way round? Is C3 the correct way round? Did you choose to ignore the CMOS safety precautions given earlier?

Once the circuit is working, you need to decide how you are going to connect it to your transmitter. Two options are shown in **Figure 3**. If you have access to a point in your transmitter circuit that has between 9 V and 12 V positive on it during transmit, you can use this to power your circuit. As the toneburst is needed only for repeaters, the switch, S1, disconnects it when not needed, as shown in Figure 3a. If you want the circuit to be self-powered, then a 9 volt PP3 battery may be used; Figure 3b shows this configuration.

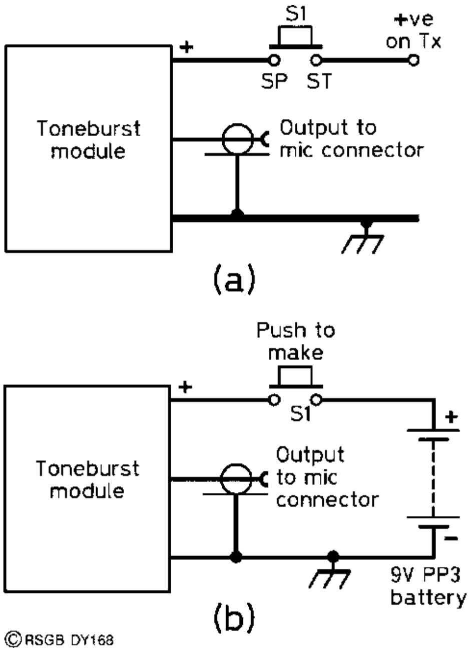

Figure 3 Toneburst module: two alternative switching arrangements

The output from the circuit board is fed directly into the microphone socket, in parallel with the microphone itself; use thin coaxial or screened cable for this lead, or you may induce hum into the microphone circuit and suffer from RF breakthrough into the audio circuits. To adjust the setting of VR1, start with it at the zero output position and connect a dummy load to your transmitter. Slowly, increase the output while monitoring your transmitted signal on another nearby receiver. Make sure you do **not** increase the output so far that the signal sounds distorted. If you would prefer that the tone was on continuously while you made this adjustment, simply connect a wire across C3 remembering, of course, to remove it as soon as you have completed the test!

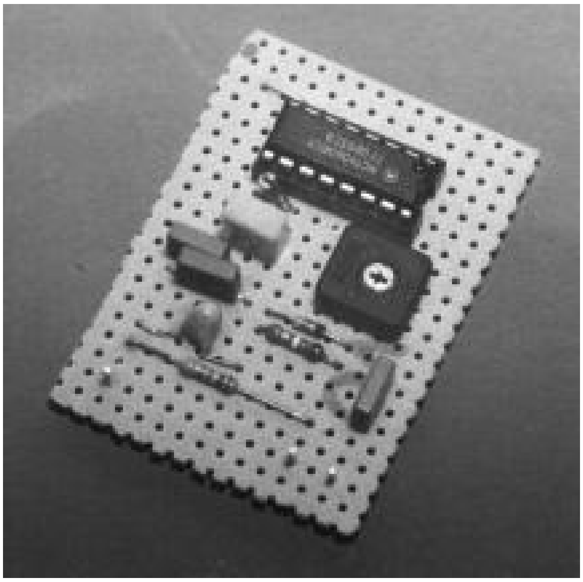

Parts list

Resistors: all 0.25 watt, 10% tolerance (or better)

R1	1 megohm (MΩ)
R2	150 kilohms (kΩ)
R3	12 kilohms (kΩ)
VR1	10 kilohms (kΩ) horizontal preset

Capacitors: all 16 V WKG or higher

C1, C2	1 nanofarad (nF) disc ceramic
C3	4.7 microfarad (μF) tantalum bead
C4	47 nanofarads (nF) disc ceramic

Semiconductors	
IC1	4060
D1	1N4148
Additional items	
XL1	XR455
	Veroboard (see text for size)
	Veropins (3)
S1	Switch (momentary action push-to-make SPST)
	16-pin DIL socket for IC1
	Single-core insulated wire for links
	Coaxial or screened cable for microphone connection

58 A circuit for flashing LEDs

Introduction

There are many occasions when one's attention needs drawing to the fact that something important has happened. A single red light coming on is seldom sufficient to attract attention, particularly if it is surrounded by other lights and indicators. The eye is known to be very sensitive to *changes* in its peripheral vision; such changes can be brought about by movement or by differences in light level – a flashing light, for example. So, a circuit that flashes a single LED or a pair of LEDs finds plenty of uses in the amateur station.

Warning

This circuit uses a member of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton

destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from destruction while it is connected to the battery. However, when the battery is removed, the same care should be exercised with its handling, because there is no supply decoupling capacitor across the IC.

Basic description

LEDs can be made to flash (switch on and off) by driving them from sources that switch on and off. Such a source is an *astable multivibrator*. If you have built or read about *A basic continuity tester*, elsewhere in this book, you will have come across such a beast before. That circuit used an astable multivibrator made from two transistors. This new circuit achieves the same behaviour from a single integrated circuit, the CMOS 4011. To give it its full description, the 4011 is a *quad 2-input NAND gate*. Quite a mouthful, but all it means is that inside the chip are four NAND gates, each with two inputs.

A NAND gate needs a positive voltage (known as a *logic 1*) on *both* inputs in order to produce zero volts (known as *logic 0*) at the output. Two NAND gates can be connected, as are A and B in **Figure 1**, to make our astable multivibrator. The combination of A and B has been described as the most perverse circuit in electronics; as soon as the output goes to logic 1, the circuit decides that it would prefer to have a logic 0 there, and switches over. With logic 0 at the output, the circuit now prefers to have logic 1 there, and so it goes on! We are going to use this continuous switching backwards and forwards to flash two LEDs. The rate at which A and B 'change their minds' is the frequency at which our LEDs will flash, and is controlled by the charging and discharging times of C1 through R2 and C2 through R1. As the values of R1 and R2 are the same, and those of C1 and C2 are the same, the ON and OFF states of the circuit are the same.

Gates C and D do not contribute to the flashing action; they act as *buffers* to isolate the LEDs from the multivibrator circuit itself. You will find in electronics that an oscillator is seldom used to drive another device *directly*; there is usually a buffer between it and the stage it drives.

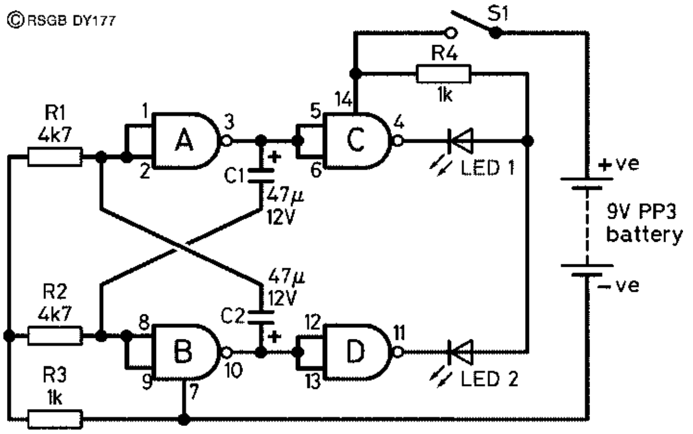

Figure 1 Flashing LEDs, circuit diagram

A characteristic of all multivibrators, astable or not, is that they have two outputs. In this case, those outputs are at pins 3 and 10, which are then buffered and appear at pins 4 and 11, respectively. When one output is at logic 1, the other is at logic 0, and vice versa. This means that LED1 is off when LED2 is on, and LED1 is on when LED2 is off, the two states switching backwards and forwards at the frequency of the oscillator.

Construction

Read the warning at the beginning of this article again. It is not intended to scare you off from building this, but is a genuine piece of advice which can save you time and irritation when all your labours result in a circuit that doesn't work! That extra bit of care can make all the difference!

Veroboard (the copper strip type) is used for the layout, shown in **Figure 2**. It measures 20 holes by 12 strips. **Be aware that there is no row 'I' in the layout, so don't miscount when you are placing components on the board!**

Firstly, cut the tracks using a 3 mm (⅛ inch) twist drill held between thumb and forefinger; check that there are no slivers of copper bridging any of the tracks, and that the tracks have been completely cut by the drill. Solder in the components carefully. Leave the IC in its carrier for the time being, and solder in the IC socket, with the notched end facing row A. On completion, check the circuit carefully. If you are happy that it is correct, follow the instructions given earlier and fit IC1 into its socket, matching up the two notches. Connect the battery and switch on. The two LEDs should flash on and off alternately. If only one LED flashes, you have probably connected the other one the wrong way round. Switch off, check and correct if necessary. If neither LED flashes, you must have a significant error in your circuit, which will need checking *again*! Or did you choose to ignore the handling precautions for the CMOS chip?

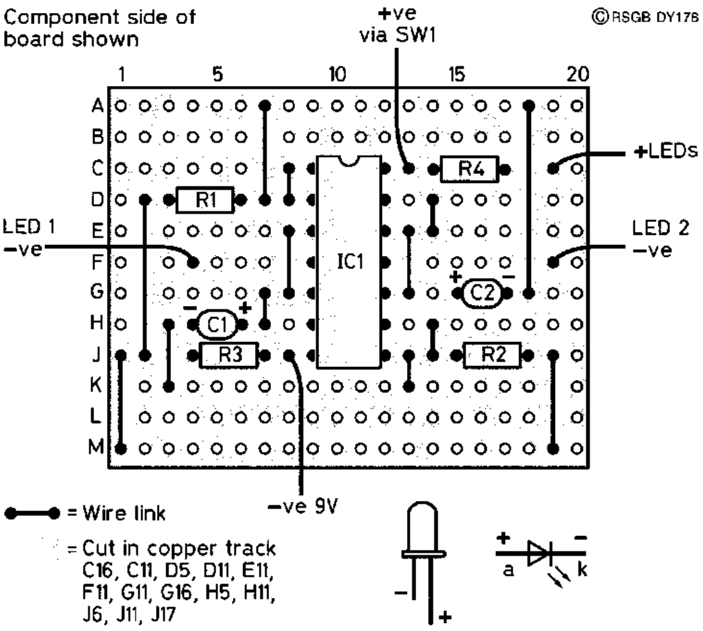
Component side of board shown
+ve via SW1
©RSGB DY17B
1
5
10
15
20
A
B
C
D
E
F
G
H
J
K
L
M
R1
R4
IC1
C1
C2
R3
R2
+LEDs
LED 1 −ve
LED 2 −ve
−ve 9V
= Wire link
= Cut in copper track C16, C11, D5, D11, E11, F11, G11, G16, H5, H11, J6, J11, J17
−
+
a
k

Figure 2 Flashing LEDs, component layout

The board can be mounted near to the point where you want your flashing LEDs to be seen, although long leads to the LEDs are acceptable. The LEDs can be different colours – it's all up to you now!

Parts list

Resistors: all 0.25 watt, 10% tolerance or better

R1, R2	4700 ohms (Ω)
R3, R4	1 kilohm (kΩ)

Capacitors

C1, C2	47 microfarads (μF) 12 V WKG

Semiconductors

IC1	4011

Additional items

LED1, LED2	Any size of LED, any colour
S1	SPST on/off
	Plastic box if needed, 8.5 by 5 by 2.5 cm

Source

Components are available from Maplin.

59 Digital logic circuits

Introduction

Logic circuits form the backbone of even the most advanced computer, yet their basic operation can be demonstrated by a couple of switches, a battery and a bulb.

Logic using switches

Everyone reading this article will look at **Figure 1** and know immediately how it works and be able to write down something like 'When switch A and switch B are closed, the light will come on'. Without knowing it, you have written down a logic statement involving the so-called **AND** operation; the light comes on only when switches A **AND** B are ON. Below the circuit in Figure 1, is a table showing the only possible positions of the two switches and the state of the bulb for each position. This is called a *truth table*, and is frequently used in logic analysis.

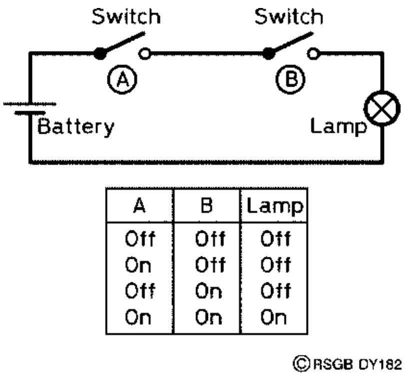

A	B	Lamp
Off	Off	Off
On	Off	Off
Off	On	Off
On	On	On

©RSGB DY182

Figure 1 Switches and lamp AND gate

Figure 2 shows a different circuit. Here, the two switches are in parallel rather than in series, as was the case in Figure 1. Again, if you analyse the circuit in words, you would say that the light will be on when switch A **OR** switch B is ON. This is an example of the **OR** operation, and its truth table is shown in Figure 2. The statement above is not complete, however; can you see why? The truth table will show you. The light comes on if A is ON, **OR** if B is ON, ***OR*** if A **AND** B are both ON. That third condition is easy to miss, but don't worry about it!

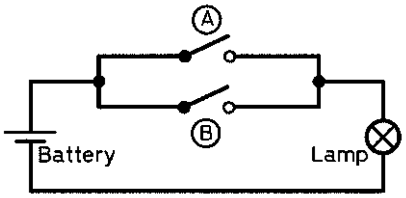

A	B	Lamp
Off	Off	Off
Off	On	On
On	Off	On
On	On	On

Figure 2 Switches and lamp OR gate

©RSGB DY184

Believe it or not, some very complicated logic is possible (in theory) using switches and lights, but it is highly impractical and would be very slow. This is where electronic logic circuits come in.

Switches with no moving parts

You may have come across projects in this book where a statement is made such as '. . . the transistor is being used as a switch . . .'. Transistors *can* be used as switches, as were thermionic valves in the world's first programmable computer *Colossus*, at Bletchley Park. However, technology has moved on from valves, through transistors to *logic gates*, combinations of electronic switches designed specifically to perform logic functions.

These act on voltage levels as their inputs and produce changes in voltage levels as their outputs. A positive voltage is called *logic 1*, and corresponds to a switch being ON in our previous descriptions; a zero voltage is called *logic 0*, and corresponds to a switch being OFF. The output from a logic gate (normally labelled Q) is also logic 1 or logic 0, corresponding to our light being ON or OFF, respectively, in our switch analogy.

Many logic devices operate from a stabilised 5 V supply, and this determines the *ideal* voltages corresponding to the two logic states:

logic 0 ≡ 0 V,

logic 1 ≡ 5 V.

The world isn't an ideal place, so the *real* voltage ranges used by the logic gates are:

logic 0 ≡ 0.0 to 0.4 V,

logic 1 ≡ 3.0 to 5.0 V.

The **AND** circuit of Figure 1 is now called an **AND gate**, and requires logic 1 inputs on A **AND** B to produce a logic 1 at the output. **Figure 3** shows this. The truth table is identical with that of Figure 1 – logic 1 replaces ON and logic 0 replaces OFF. Now compare Figure 2 with **Figure 4** – circuits and truth tables for the **OR** function. Again, we have exact similarity.

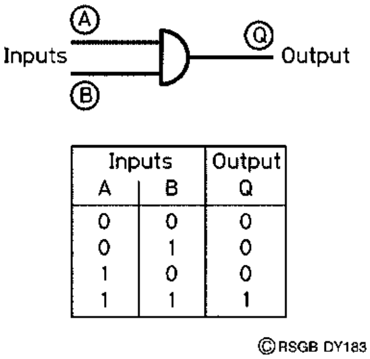

Inputs		Output
A	B	Q
0	0	0
0	1	0
1	0	0
1	1	1

©RSGB DY183

Figure 3 Electronic AND gate

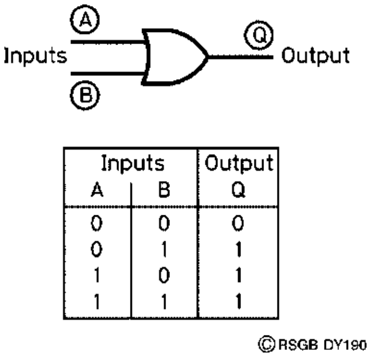

Inputs		Output
A	B	Q
0	0	0
0	1	1
1	0	1
1	1	1

©RSGB DY190

Figure 4 Electronic OR gate

There is another very common logic gate, which performs the **NOT** function. It is easy to understand. Just ask yourself the question 'What is **NOT** logic 0?', and the answer is obviously 'logic 1'. Similarly, logic 0 is **NOT** logic 1. A **NOT** gate simply changes the logic state of the input; it is also known (because of this behaviour) as an *inverter*. Its symbol and truth table can be found in **Fig 5**. Note the little circle on the output of the gate in Figure 5. In logic circuits, this symbol always implies inversion, or the presence of a **NOT** gate. Keep an eye open for it!

So far, the logic functions we have discussed have all been words which we use in everyday language, which has made the electronic interpretation of them relatively easy. Now we must introduce a function for which there is *no* analogy in normal speech – the **NAND** function. This means a

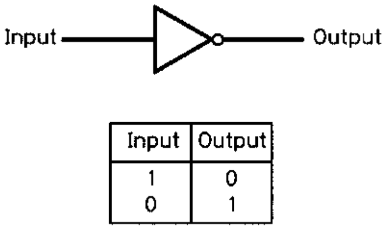

Input	Output
1	0
0	1

©RSGB DY185

Figure 5 Electronic NOT gate

combination of an **AND** gate and a **NOT** gate, and the sharp-eyed reader will have spotted the little circle added on to the normal **AND** gate symbol in **Figure 6**!

To make things easier to understand, the truth table in Figure 6 has four columns, not three as in previous tables. The third column is the standard **AND** output – compare it with the third column in Figure 3. That is the output from the **AND** gate *before* it encounters the little circle that inverts it, so the final output from the **NAND** gate is the output of the **AND** gate, inverted! The third and fourth columns are the inverse of each other.

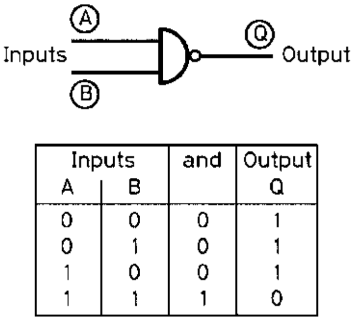

Inputs		and	Output
A	B		Q
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

©RSGB DY186

Figure 6 Electronic NAND gate

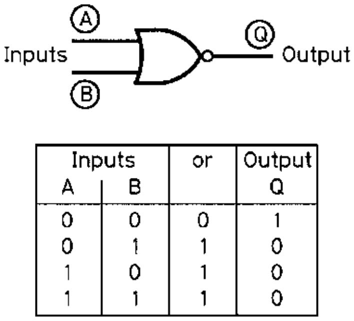

Inputs		or	Output
A	B		Q
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

©RSGB DY187

Figure 7 Electronic NOR gate

We can add the inverting operation to the output of an **OR** gate also, producing a **NOR** gate! The symbol and its truth table are shown in **Figure 7**. Again, notice that the final output is that of the ordinary **OR** gate, inverted!

These new functions of **NAND** and **NOR** are used more than the **AND** and **OR** functions because it makes other circuits easier to design using combinations of these gates.

A taste of Boolean algebra

The design of circuits using combinations of logic gates usually begins with a little mathematics, where the functions to be implemented are analysed. The mathematics used is surprisingly simple, and is a slightly changed version of ordinary algebra called *Boolean algebra*, which allows manipulation of logic functions to be made. Normal algebra has operations in it such as addition, subtraction and multiplication and division. The mathematician Boole found that the logical **AND** operation could be handled by the algebraic operation of *multiplication* (symbols × or •), and the **OR** operation by the algebraic operation of *addition* (symbol +). The **NOT** operation involved a new symbol, that of a bar over the input being inverted, such as $\overline{A}$.

So, our five basic logic operators can now be written in a mathematical form:

AND	$Q = A \times B$
OR	$Q = A + B$
NOT	$Q = \overline{A}$
NAND	$Q = \overline{A \times B}$
NOR	$Q = \overline{A + B}$.

Using logic operations in a mathematical form enables the most complex logic to be designed, simplified and converted into circuit diagram form in a very efficient and rapid way. There are more logic operators than the five we have considered here but, in general, they can all be broken down into combinations of these gates alone!

60 A resistive SWR indicator

Introduction

When a transmitter produces some output power, we want to make sure that as much as possible of this power is radiated by the aerial. This often requires the use of an *aerial tuning unit* (ATU), which matches the aerial impedance to that of the transmitter.

How do you know when this matching has been achieved? The most usual way is to use a *standing-wave ratio* (SWR) indicator. If the impedance of the aerial does not match that of the transmitter output, some of your transmitter power (also known as the *forward* power) is reflected back along the aerial feeder and back into the transmitter, where it causes excess heating. The forward and reflected waves interact along the feeder to produce a wave whose position remains constant, and which is therefore called a *standing wave* or a *stationary wave*. An SWR meter simply indicates forward power and reflected power, and adjustments are made to your ATU until the reflected power is as small as possible (ideally zero, of course). If there is no reflected power then, by a process of elimination, all your forward power is reaching the aerial!

Sampling the RF

Whatever type of SWR indicator you use, it must use some sort of sampling circuit to pick up the forward and reflected waves. The project *A standing-wave indicator for HF*, elsewhere in this book, uses a toroidal transformer to separate the readings for the forward and reverse waves. This design differs in that it measures the voltages across resistors through which the RF current is passing. Its advantages are:

(a) it uses cheap parts – four resistors, two capacitors and a diode, together with a rotary switch, a surplus meter, a preset potentiometer and two sockets;
(b) within this SWR indicator, there is always a resistive path for the RF current from the transmitter, formed by R1, R2 and R3; this can prevent damage to simpler home-made transmitters, which may be damaged during adjustment of the ATU when using more conventional SWR indicators.

The only disadvantage of this form of indicator is that it must be switched out of circuit once the ATU has been adjusted for a particular band.

Construction

The SWR indicator is very simple to build, as most of the components can be mounted on the back of the 3-way rotary switch. This is shown in **Figure 1**. The switch is a 4-pole, 3-way rotary type, of which only two poles are used.

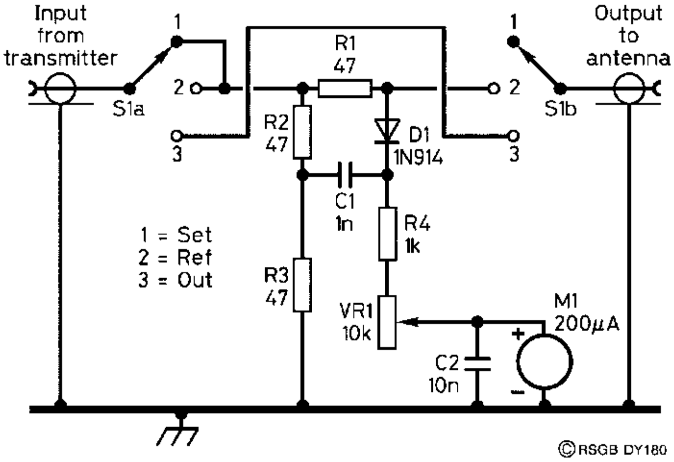

Figure 1 Resistive SWR meter, circuit diagram

Because the other switch contacts are not being used, they can be employed as support tags for other components. The ground wires are all soldered on to the metal frame of the switch. If your switch frame is of all-plastic construction, then a 12 SWG copper wire run around the switch will make a good earth connection to the metal case for the leads shown in Figure 1. The preset potentiometer used to control the sensitivity of the circuit can be mounted directly on the meter tag.

Resistors R1, R2 and R3 handle the RF power during the tuning-up process. If you have them, use 1 watt resistors; otherwise, you can use two 100 ohm half-watt resistors in parallel for each of R1, R2 and R3. The meter, M1, can be any DC type of sensitivity around 200 μA.

In use

First, find a clear frequency, and without the indicator in circuit, check that the frequency really *is* clear by asking and identifying yourself. If it is, connect the indicator between the transmitter and the ATU which, in turn, is connected to your aerial. Turn SW1 to the SET position and key the transmitter. Adjust VR1 until the meter reads full scale. Switch off the transmitter. Turn SW1 to the REF position and key the transmitter again. Adjust the ATU until the lowest reading is obtained on the meter. Switch off the transmitter. For the chosen frequency, you have adjusted your ATU for minimum reflected power and hence the lowest SWR. You will need to repeat the process when you change bands, and possibly when you change frequency within the same band. Switching SW1 to the OUT position, you are ready to transmit. You may have noticed that it is good practice to switch the transmitter off when operating SW1. Get into that habit!

Parts list

Resistors	
R1, R2, R3	47 ohm (Ω) 1 watt (or 2 × 100 Ω, ½ watt, see text)
R4	1 kilohm (kΩ) 0.25-watt
VR1	10 kilohms (kΩ) preset
Capacitors	
C1	1 nanofarad (nF)
C2	10 nanofarads (nF)
Semiconductors	
D1	1N914 or similar
Additional items	
SW1	4-pole 3-way rotary switch, of which only 2 poles are used
M1	200 μA DC
	Case

61 An audio filter for CW

Introduction

This is a simple passive circuit (it has no power supply) that adds some audio selectivity for Morse code reception and also includes a very simple noise limiter that gives a visual indication of when noise spikes are being removed!

The circuit

Figure 1 shows the complete circuit. The tuned circuit of C1 and L1 resonates very close to 800 Hz, so initially you will have to tune a signal in carefully until it sounds loudest – you will soon be able to do this without thinking. The two LEDs connected back to back across the signal path act as a noise limiter, reducing the amplitudes of static crashes and noise from car ignition systems, etc. The LEDs blink when they conduct – this is not necessary to the operation of the circuit, but adds a little colour to your listening! The noise limiter *does* make listening more comfortable, though.

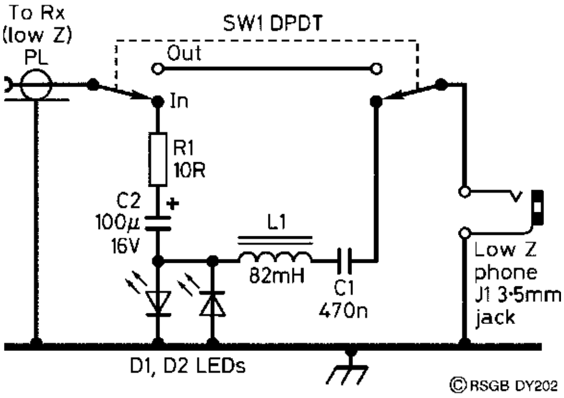

Figure 1 CW filter, circuit diagram

Construction

The circuit layout is shown in **Figure 2**. Point-to-point wiring is used, with a small tag-strip being the only item used for the extra support of C2 and R1. The LEDs and SW1 support the other components. An aluminium box (**Figure 3**) is used to make the circuit tidy and usable.

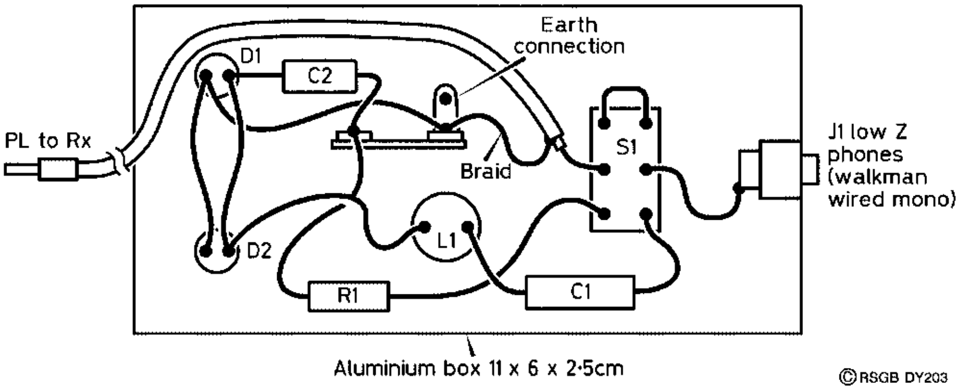
Earth connection
D1
C2
PL to Rx
S1
Braid
J1 low Z phones (walkman wired mono)
L1
D2
R1
C1
Aluminium box 11 x 6 x 2·5cm
©RSGB DY203

Figure 2 CW filter, component layout

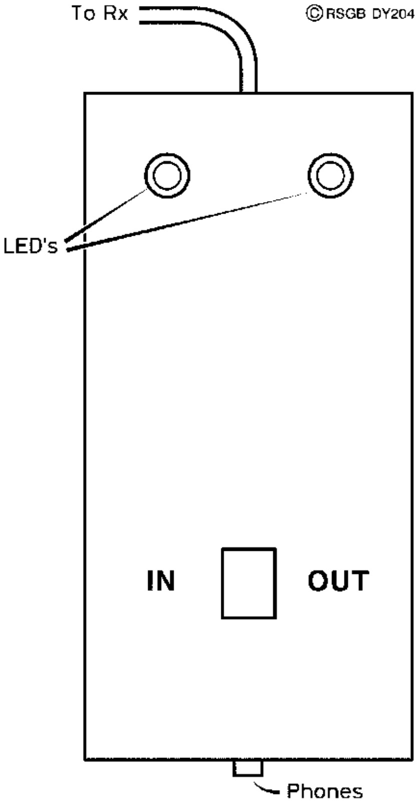
To Rx
©RSGB DY204
LED's
IN
OUT
Phones

Figure 3 CW filter, front panel

You must use low-impedance headphones for the circuit to perform properly. Plug the completed unit into the headphone socket of your receiver and adjust the volume so that the LEDs are just *not* blinking on normal audio. The unit should be switched out of circuit for speech listening; this is the purpose of the toggle switch, SW1.

Parts list

Resistor	
R1	10 ohms (Ω), 0.25 watt, 10% tolerance
Capacitors	
C1	470 nanofarads (nF)
C2	100 microfarads (μF) electrolytic, 16 V WKG
Inductor	
L1	82 millihenries (mH)
Semiconductors	
D1, D2	LEDs
Additional items	
SW1	DPDT
J1	3.5 mm jack socket
PL	Jack plug to suit receiver
	Aluminium box, approx. 11 by 6 by 2.5 cm

Source

Components are available from Maplin.

62 An electronic die

Introduction

Throwing a die is a venerable way of generating a ‘random’ number between 1 and 6. A die is easy to lose, so here is an electronic die-throwing circuit which brings the technique up to date and serves as another application of logic circuits. If you need a reminder of the basics of logic, refer to *Digital logic circuits*.

Warning

This circuit uses members of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the

magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains-earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from damage while it is connected to the battery.

Description

The circuit is shown in **Figure 1**. IC1a and IC1b form an *astable multivibrator*, similar to that used in 'An LED Flasher' also in this book. It runs constantly, and its output is fed to IC2 via a single NAND gate, IC1c. The other input of IC1c is held at 0 V by R3 and C3. Whenever one input of a NAND gate is logic 0 (or 0 V), there can be *no output* from the gate, irrespective of what is happening at other inputs. So, despite the fact that the oscillator (or *clock*) is running all the time, its square-wave output never

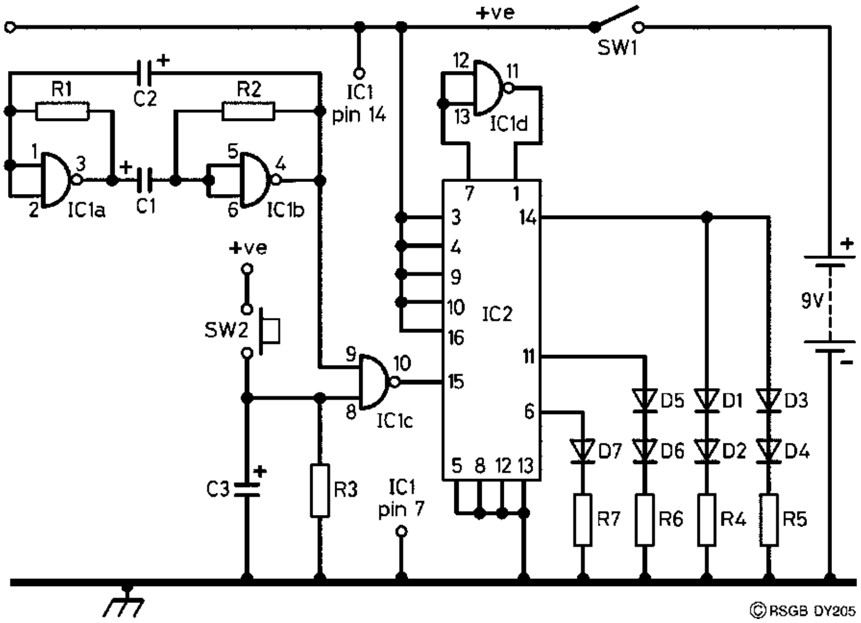
+ve
SW1
R1
C2
R2
IC1 pin 14
12
11
13
IC1d
1
3
5
4
IC1a
C1
2
6
IC1b
7
1
3
14
4
9
10
IC2
16
+ve
SW2
9
10
15
11
8
IC1c
6
D5
D1
D3
D7
D6
D2
D4
5 8 12 13
C3
R3
IC1 pin 7
R7
R6
R4
R5
9V
©RSGB DY205

Figure 1 Electronic dice, circuit diagram

reaches IC2 until the ‘other’ input of IC1c is brought away from logic 0. This happens when SW2 is pressed. Pressing SW2 momentarily, as is normally done, charges up C3 to 9 V, putting a logic 1 on the second input of IC1c and allowing the clock signal through to IC2. As the switch is released, C3 begins to discharge through R3, gradually lowering the voltage on pin 8 of IC1c. As this voltage crosses 4.5 V (half the supply voltage), IC1c then treats this input as logic 0, which cuts off the clock signal from IC2 again. Thus, when SW2 is momentarily pressed, the clock signal is fed to IC2 for a short interval of time, before being blocked again.

IC2 is a binary counter, and the outputs that concern us are from pins 14, 11 and 6. Pin 14 is the *most significant bit* and pin 6 the *least significant bit*. See the panel for an explanation of what happens here.

To understand these outputs, you should be able to count in binary, using three bits. At the start of the counting process, all the bits have zero values, i.e. 000. The left-most bit is called the *most significant bit*, and the right-most bit is called the *least significant bit*. (When we write numbers normally, a number 1 in the left-hand position represents one hundred, 1 in the middle column represents ten, while the number 1 in the right-hand position means one. One hundred is a more significant number than one, hence the nomenclature.)

The three bits of our binary number have values, from left to right, of 4 (=2^2), 2 (=2^1) and 1 (=2^0). This means that a number 1 in the left-hand column signifies the normal number four, in the middle column it would represent two, and in the right-hand column, it would represent one. This should help you understand the patterns of bits which emerge as the clock waveform is counted, as the next paragraph explains.

As each cycle of the clock enters IC2, it increments its internal counter and the values of that counter are shown by the states of pins 14, 11 and 6. After the first pulse, these three pins would have states corresponding to 001; after the second, 010; after the third, 011; after the fourth, 100; after the fifth, 101; after the sixth, 110. This sequence of 3-bit numbers represents a binary count from 1 to 6 in ‘normal’ parlance. These six states are used to illuminate the conventional pattern of dots on a die, using LEDs.

As the clock cycles are counted, the LEDs flicker as the die is ‘rolled’. Resistors R4 to R7 are used to limit the current through the LEDs. When the counter stops, it can be in any of the positions shown in **Table 1**. Because of

Table 1 Dice number relative to counter output

Step	*Pins*			*Dice number*
	14	*11*	*6*	
1	0	0	1	4
2	0	1	0	2
3	0	1	1	6
4	1	0	0	1
5	1	0	1	5
6	1	1	0	3

the way in which the dots are grouped on the faces of a die, the wiring of the LEDs is simpler than it might otherwise be. You can see from Figure 1 that there are really only three sets of connections to the LEDs – one from each bit of the counter output. When pin 14 is at logic 1, four LEDs are lit, corresponding to the number four. For the number five, pins 14 and 6 will be at logic 1. For the number one, only pin 6 is at logic 1. For three, pins 11 and 6 are at logic 1, for six, pins 14 and 11 are at logic 1, and for two, only pin 11 is at logic 1. These conditions are summarised in Table 1.

Construction

Two pieces of Veroboard (of the strip type) are needed – one for the main circuit (**Figure 2**) and the other for the display LEDs (**Figure 3**). The display board is easier to build, so we will do that first. It measures 20 holes by 18 strips.

Cut the tracks as shown in Figure 3, using a 3 mm (1/inch) twist drill rotated between thumb and forefinger. Insert and solder the Veropins, resistors and the wire links. Then solder the LEDs in place, making sure their polarities are correct. The LEDs are mounted proud of the board (**Figure 4**, which will help when you come to fix the board into a case.

The main board measures 32 holes by 18 strips. As before, remove the tracks in the places shown in Figure 2. Note that track 'L' is *not* broken under the position for IC2. Solder in the IC holders and the links. Solder in R1 and R2 vertically, and R3 horizontally. Then solder in C1, C2 and C3; these all being electrolytics, check their polarities. Solder in the Veropins and prepare the connections to those components not on the main board, using stranded, insulated wire. Check the circuit carefully, and read the handling precautions given earlier before carefully inserting the two ICs.

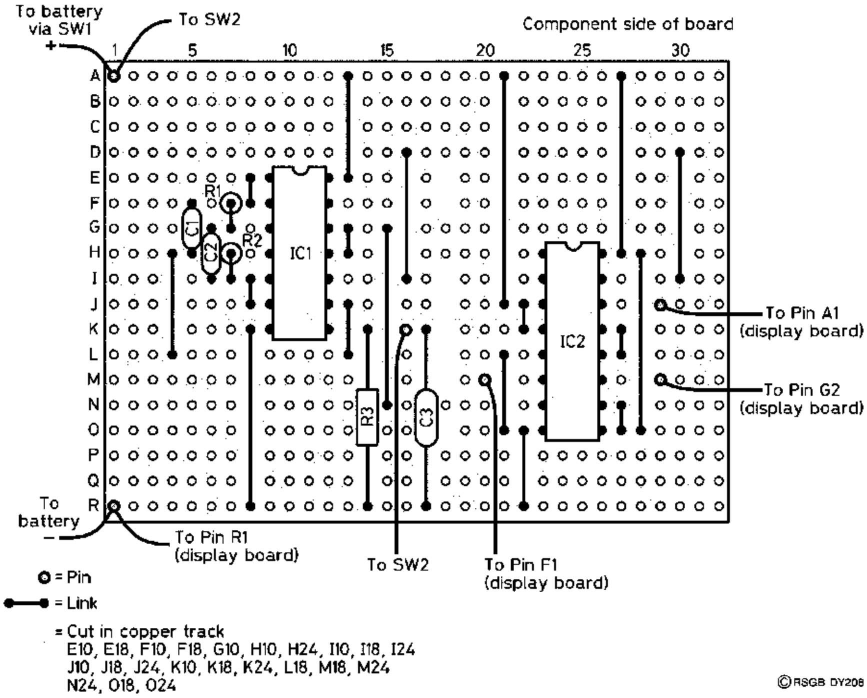
To battery via SW1
+
To SW2
Component side of board
1
5
10
15
20
25
30
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
R1
R2
C1
C2
IC1
IC2
R3
C3
To Pin A1 (display board)
To Pin G2 (display board)
To battery
−
To Pin R1 (display board)
To SW2
To Pin F1 (display board)
= Pin
= Link
= Cut in copper track
E10, E18, F10, F18, G10, H10, H24, I10, I18, I24
J10, J18, J24, K10, K18, K24, L18, M18, M24
N24, O18, O24
©RSGB DY208

Figure 2 Main board, component layout

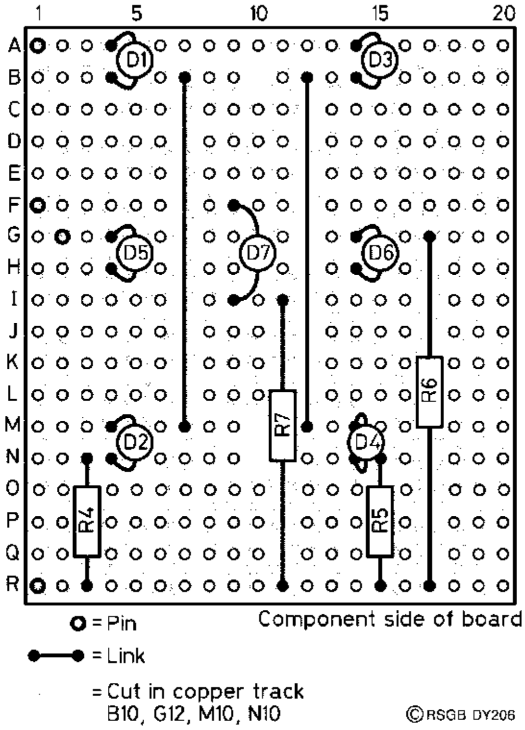
1
5
10
15
20
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
D1
D3
D5
D7
D6
D2
D4
R4
R5
R6
R7
Component side of board
= Pin
= Link
= Cut in copper track
B10, G12, M10, N10
©RSGB DY206

Figure 3 Board for the six LEDs

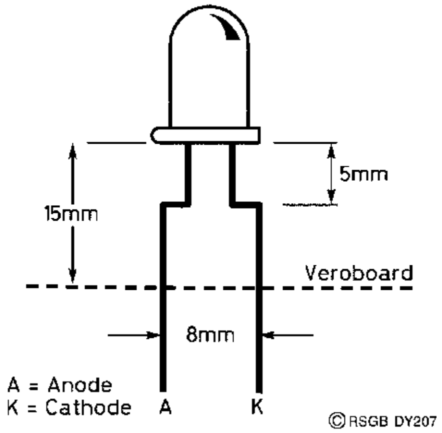

Figure 4 D7 lead bending

Testing and use

When the circuit is switched on, it is likely that the centre LED only will be lit. Press SW2, the 'throw' switch. All the LEDs should flicker, even after the switch is released, but only for a second or so. Then a standard pattern should show, signifying numbers between one and six. If this does not happen, switch off and check your circuit again. Are there any missing links? Are there any obvious dry joints (usually dull instead of shiny)? Can you see any solder bridges between tracks? Did you choose to ignore the handling precautions? If none of these results in a working circuit, try some fault-finding.

- Disconnect the lead to D7 on the main board (marked 'to pin F1' on Figure 2). Touch the lead on to the positive supply rail. D7 should light. Now transfer this lead on to pin 4 of IC1. If it flickers, then the oscillator is running (which it should).
- Transfer this lead to pin 10 of IC2. D7 should flicker when SW2 is pressed, but should be on permanently when it is released. This means that the correct signals are reaching IC2 from IC1.
- If one of the chains of LEDs (i.e. D1 and D2, D3 and D4, D5 and D6) does not light, you may have one *or both* LEDs the wrong way round.

When the die is working, you may care to experiment with the values of R1 and R2, but you should keep their values the same, i.e. R1 = R2. The larger the values of these, the more slowly the die will appear to 'roll'.

Finishing touches

On completion of the project, you will want to mount it in a smart case; any plastic box is suitable for this, with the display board mounted so that the

LEDs protrude through the top and are fixed to it using LED clips. The way in which you mount both boards to the case is entirely up to you.

Parts list

Resistors: all 0.25 watt, 10% tolerance, or better

R1, R2	5600 ohms (Ω)
R3	15 kilohms (kΩ)
R4–R7	470 ohms (Ω)

Capacitors

C1, C2	1 microfarad (μF) electrolytic, 16 V WKG
C3	68 microfarads (μF) electrolytic, 16 V WKG

Semiconductors

IC1	4011
IC2	4029
D1–D7	LEDs, any size and colour

Additional items

SW1	On-off switch SPST toggle
SW1	Push-to-make, non-latching
	Veroboard, 2 pieces, see text for sizes
	Veropins, 10
	Stranded insulated wire for general wiring
	Single-core insulated wire for links
	PP3 battery clip
	PP3 battery
	Plastic box to suit
	LED mounting clips (7)
	Means of mounting boards to box

Source

Components are available from Maplin.

63 The absorption wavemeter

Introduction

The purpose of the absorption wavemeter is to check that a transmitter is radiating within the correct waveband, and to detect any spurious *harmonic* emissions. A harmonic frequency is an integral multiple of the carrier frequency – e.g. if the carrier is at a frequency *f*, the harmonics are at frequencies 2*f*, 3*f*, 4*f* . . . and so on.

It can also be employed as a relative field-strength indicator, being used in experimentation with aerials. What it *cannot* do is to measure the transmitted frequency accurately; all it can do is to place the signal within a particular band, say the 7 MHz or the 18 MHz band (or their harmonics).

How it works

You may recognise the circuit of **Figure 1** as a crystal set with a meter replacing the headphones. It is simply the parallel combination of an inductance with a variable capacitance, which constitutes a resonant circuit with a variable frequency. A dial on the variable capacitor can be roughly calibrated with frequency. The fixed capacitor across the meter helps to improve the meter reading by making it read the peak value of the carrier wave rather than the average value.

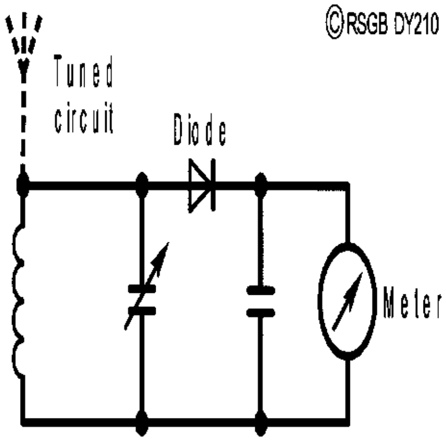

Figure 1 Absorption wavemeter circuit

When the wavemeter is situated in the vicinity of a relatively strong radio-frequency (RF) field, the inductor absorbs a small amount of energy from that field. If the circuit is resonant at the same frequency as that of the transmitter, an RF voltage is produced across the coil and is proportional to the strength of the RF field. The RF voltage is *rectified*, or *detected* (turned from AC into DC) by the diode and the second capacitor, and is displayed by the meter.

The tuned circuit

The formula for calculating the resonant frequency of a coil and capacitor is relatively simple. It is

$$f = \frac{1}{2\pi\sqrt{LC}},$$

where f is the resonant frequency (Hz),
L is the inductance (H), and
C is the capacitance (F).

If this formula puts you off, there is this slightly easier version:

$$f = \frac{159}{\sqrt{LC}},$$

where f is the frequency in **MHz**,
L is the inductance in **microhenries** (**μH**), and
C is the capacitance in **picofarads** (**pF**).

When a variable capacitor is specified, it is usual to quote its maximum capacitance, i.e. the value when the vanes are fully meshed. When the value of the capacitor is small, it is usual to quote its minimum capacitance also, the value when the vanes are fully open. If you use a variable capacitor of 10–100 pF with a 100 μH inductor, you will use the value of 10 pF to calculate the maximum resonant frequency, and the value of 100 pF to calculate the minimum resonant frequency, giving a tuning range of 1.6 MHz to 5 MHz. Check it yourself!

64 An HF absorption wavemeter

Introduction

An absorption wavemeter is basically a simple tunable detector circuit, such as would be used in a crystal set, but with the headphones replaced by a meter, in order to indicate the strength of the received signal.

The circuit

Figure 1 shows the circuit of the wavemeter. It is designed to cover all the HF bands from 1.8 MHz to 28 MHz in four switchable ranges. There is no built-in method of amplification, so a sensitive meter is needed in order to indicate sufficiently using the available absorbed energy. The meter does not need calibrating, it serves to produce only an indication of the absorbed power, *not* its absolute value.

An absorption wavemeter does not have outstanding selectivity, mainly because of the loading effect of the meter. This problem is ameliorated here (but only slightly) by the inclusion of R1 in series with the meter. It gives

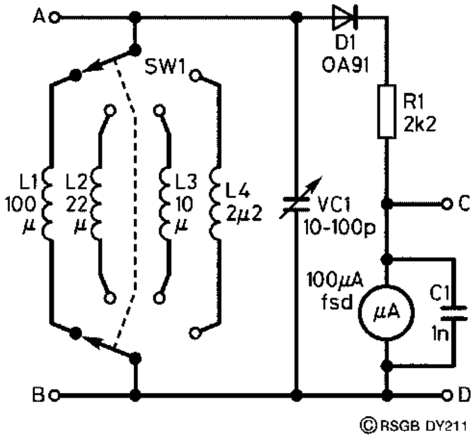

Figure 1 HF absorption wavemeter, circuit diagram

improved selectivity at the expense of a slight loss of sensitivity. The appearance of capacitor C1 across the meter is discussed in *The Absorption Wavemeter*, in the previous section.

Coils

These can be purchased (see the parts list) or hand-wound on short lengths of PVC water pipe or conduit, and the coil-winding details are given in **Table 1**. In theory, the 100 pF tuning capacitor would cover all the HF bands with only three coils, omitting the 22 μH coil. Sample calculations are explained in the previous article. However, its inclusion avoids the common problem of having some bands cramped at the extreme edges of the tuning scales.

Table 1 Close-wound coil values

L (μH)	No of turns	Span (mm)	SWG
PVC FORMER OF OUTSIDE DIAMETER 25 mm			
2.2	9	9	20
10	24	24	20
22	44	44	20
100	110	60	26
PVC FORMER OF OUTSIDE DIAMETER 20 mm			
2.2	10	10	20
2.2	9	5	26
10	32	32	20
10	24	12	26
22	63	62	20
22	41	21	26
100	134	70	26
PVC FORMER OF OUTSIDE DIAMETER 19 mm			
2.2	11	11	20
2.2	9	5	26
10	35	34	20
10	25	13	26
22	68	67	20
22	44	22	26
100	157	79	26

Construction

Hand-wound coils, using 19–25 mm PVC formers, have single-layer windings, with their ends secured by threading the wires through two small holes drilled in the formers at each end of each winding.

You may have wondered about the use of the two-pole switch SW1, as all the lower ends of the coils are grounded anyway. The answer lies in the ease of construction. Each coil can be mounted on the tags of the switch for support. If you are using the commercial coils, they are small enough to be accommodated on a miniature rotary switch; a larger rotary switch is required for hand-wound coils. The specification of a 6-way switch allows for the addition of extra frequency ranges at little extra cost. Rotary switches such as the 2-pole, 6-way variety have an adjustable end-stop which allows the number of ways to be set anywhere between 2-ways and 6-ways. Remove the nuts from the switch, and a 'washer' will fall out. It is not a washer, really, and has a little piece of metal bent over at right angles to the washer. This fits into one of 12 holes in the body of the switch, and prevents the shaft from turning through 360° and selecting the number of ways.

Calibration

This process needs a commercial amateur-bands transceiver, if it is to be done quickly and accurately.

- Connect the transceiver to a dummy load, and connect a piece of flexible wire from point A in Figure 1 and wrap two or three turns around the cable between the transceiver and the dummy load.
- Turn the transceiver power level control to minimum, and set the frequency at 1.81 MHz and the mode to FM or AM. Switch to transmit.
- With the wavemeter set to its lowest frequency range (i.e. 1.8–3.5 MHz), rotate the tuning capacitor until maximum deflection is obtained from the meter needle. Write the frequency on the inner ring of the wavemeter scale at that point.
- Switch off and retune the transmitter to 3.5 MHz. Repeat the above process.
- Repeat, using switch range 2 for 3.5, 7 and 10 MHz, range 3 for 7, 10 and 14 MHz, and range 4 for 14, 18, 21, 24 and 28 MHz.

An alternative source of calibration signal could be an HF signal generator, with a single-turn loop of wire at the remote end of its cable, and the wire from the wavemeter brought close to the loop.

Remember that the marks on the scale represent frequency *bands*, not precise frequencies.

Extending the range

If you would like to experiment with increasing the range to the lower VHF band, then 50 MHz should be achievable with an additional self-supporting coil (no former). Try three turns of 20 SWG enamelled copper wire of 25 mm ID (internal diameter). Wind it on some 25 mm PVC pipe, as before, then slide out the pipe! The same result could be achieved with four turns of 19 or 20 mm ID, or 12 turns of 7.1 mm ID.

Parts list

Resistor	
R1	2200 ohms (Ω) 0.25 watt, 10% tolerance
Capacitors	
C1	1 nanofarad (nF) min. ceramic
VC1	100 picofarad (pF) variable
Inductors	
L1	100 microhenries (μH)
L2	22 microhenries (μH)
L3	10 microhenries (μH)
L4	2.2 microhenries (μH)
Semiconductors	
D1	Germanium OA91 or 1N4148
Additional items	
SW1	Rotary 2-pole, 6-way
	Moving-coil meter, 50 or 100 μA FSD (any meter of this sensitivity will do)
	Plastic box, approx. 150 by 80 by 50 mm
	Connecting wire, coloured

65 A vertical aerial for 70 cm

Introduction

If the range of your hand-held transceiver is very limited, what you need is a vertical aerial mounted outside. This design is a half-wave dipole fed at the end instead of the middle. Because the impedance of the dipole is high at its ends (and low in the centre), a matching circuit is needed so that this high impedance can be matched to that of the low-impedance coaxial cable. All that is needed is a coil, which increases the *electrical length* of the aerial to ⅝-wavelength. (Note that the *electrical* length (i.e. the length as it appears to an RF signal) is not necessarily the same as the *actual* length.)

Construction

The aerial element and the coil are made from a single piece of 1.5 mm welding (brazing) rod, and the dimensions are given in **Figure 1**. Wind the coil around a 4 mm rod or the shank of a twist drill. The lower end is filed to a point and then soldered into the centre conductor of a 4-hole panel-mounting BNC socket. (Try to obtain a good-quality BNC socket with PTFE insulation – the insulation of cheaper sockets is easily damaged.) Trim the element to 427 mm (top of element to top of coil) *after* the wire has been soldered to the socket.

The base coil causes the aerial to be rather ‘whippy’, so a piece of 5 mm plastic knitting needle can be cut to the length of the coil and then forced into it.

The radials are made from four lengths of 3 mm welding rod. These are bent and soldered into the four mounting holes of the socket, and then cut to the lengths shown in Figure 1.

Testing

Using a standing-wave-ratio (SWR) meter connected between your aerial and the transceiver, measure the SWR (or obtain some indication of the reflected power) on transmit. If it is greater than 1.5, switch off the transmitter, trim about 3 mm off the end of the element, and try again. Repeat the process until the measured SWR (or the reflected power) is as low as you can get it.

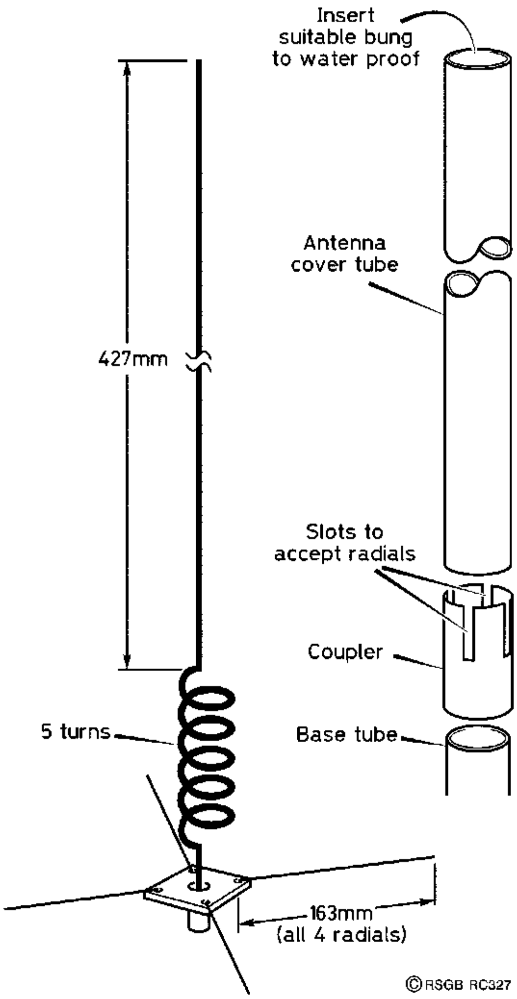
Insert suitable bung to water proof
427mm
Antenna cover tube
Slots to accept radials
Coupler
5 turns
Base tube
163mm (all 4 radials)
©RSGB RC327

Figure 1 Construction of the 70 cm antenna

Installation

The aerial is made waterproof by enclosing it in a 22 mm diameter PVC waste water pipe. It is 'weldable', and available at plumbers' merchants, usually by the metre. A coupler is slotted to take the radials (Figure 1). File the BNC socket as required, so that it slides inside the coupler until the radials poke out of the slots. Cut a length of plastic tubing which is 30 mm longer than the aerial, and push it into the coupler. You will then need a plastic bung or screw-top to waterproof the top end.

Plastic welding solution is now applied to the joints, sealing the aerial inside the tube. The coupler has been weakened as a result of making the slots, so it is worthwhile applying PVC tape around this joint until the welding solution sets.

A support for the aerial is made from an off-cut of tube pushed into the lower end of the coupler and held with a self-tapping screw.

Parts list

1	4-hole panel-mounting BNC socket
1 m	3 mm brazing rod (may be available from a small garage)
1 m	1.5 mm brazing rod
1 m	22 mm PVC waste water pipe
1	22 mm straight coupler
	Plastic welding solution

66 A UHF corner reflector aerial

Introduction

The corner reflector is a well-known design and is capable of good performance on the VHF and UHF bands. At UHF, the practical implementation of the corner reflector is an ideal constructional project.

Some details

Expressed quite simply, the aerial consists of a $\frac{\lambda}{2}$ dipole (where λ is the standard symbol for wavelength, making a '$\frac{\lambda}{2}$ dipole' a half-wave dipole).

Nothing new in that, you might say. However, the interesting feature is the reflector, which is not the usual single element, but a 90° metal 'corner', acting rather like a parabolic dish as used for satellite signal reception. The wind resistance of this type of reflector makes it impractical so, to reduce the 'windage' quite significantly, we make the 'corner' from closely spaced rods, as illustrated in **Figure 1**.

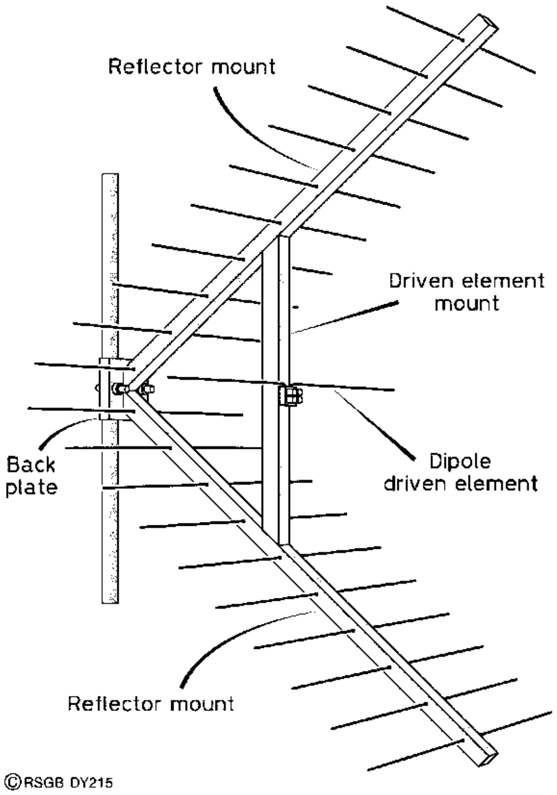

Figure 1 70 cm corner reflector antenna

The reflector consists of a number of 0.6λ rods, spaced from each other by 0.1λ. The aerial frame can be made of metal or wood, but wood is easier to work with, and mounting the elements to the frame is simpler. The prototype was made with wood of 20 mm by 15 mm cross-section, as **Figure 2** shows. The wood was varnished for protection. The elements were made from 1.5 mm diameter copper wire, because a large reel of the wire happened to be available. The wire diameter is not critical; tubing could be used just as successfully. 14 SWG hard-drawn copper aerial wire would be even better than that used in the prototype.

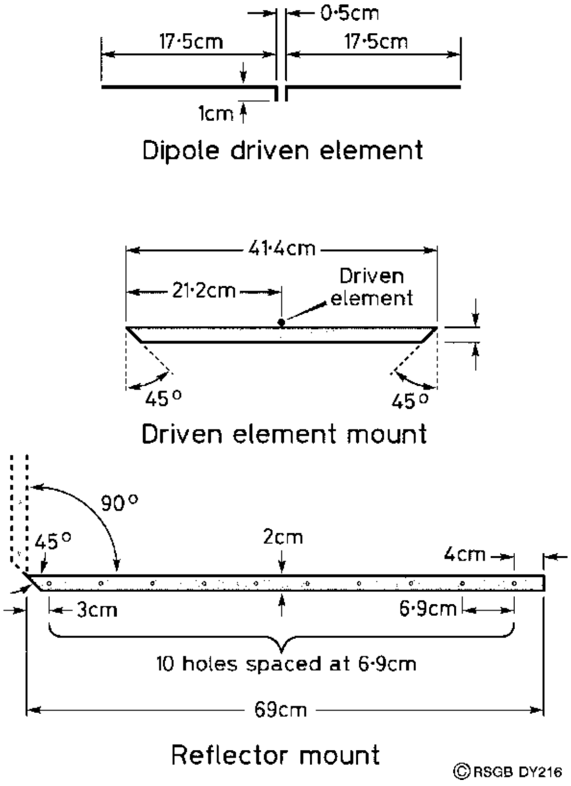
0·5cm
17·5cm
17·5cm
1cm
Dipole driven element
41·4cm
21·2cm
Driven element
45°
45°
Driven element mount
90°
45°
2cm
4cm
3cm
6·9cm
10 holes spaced at 6·9cm
69cm
Reflector mount
©RSGB DY216

Figure 2 Driven element dimensions, together with boom dimensions for driven element and reflectors

Construction

This project is just as much a woodworking project as a radio project! Follow the instructions carefully, and you should have little trouble.

- Cut the booms for the reflectors, as shown in Figure 2. A mitre block is invaluable here in producing the 45° corners.
- Using the dimensions given on the diagram, mark the hole positions for the reflector elements, and then drill holes of a size which holds the elements firmly.
- Cut the driven element boom according to the diagram, and mark the point midway along the longer side, which will assist you later in positioning the driven element.
- Cut the back plate to size (about 120 mm by 80 mm). You may need to alter this size depending on the size of the U-bolt you will be using to clamp the aerial to the mast.

- If you want to be extra cautious in your construction, use the belt-and-braces approach, commonly known as ‘screw-and-glue’ to fix the booms to each other and to the back plate.
- Fix the reflector booms to the back plate first, then slide in the driven-element boom until it will go no further, then apply the wood glue and screw the two ends tightly to the reflector booms. Leave for the period prescribed by the glue manufacturers for the glue to harden.
- Varnish the whole structure.
- Cut the driven element to the correct size plus a couple of centimetres (the reason for this will be evident in the *Testing* section), and fix it to the centre of its boom (at the position you marked earlier) with a ‘chocolate block’ connector to which the coaxial feeder cable will eventually be connected.
- Cut and fix the reflector elements in place. If you find that these are a loose fit in the holes then, *for each element*, drill a pilot hole through the boom to intersect the hole for the element. File off the point of a woodscrew, and screw it gently into the pilot hole until it meets the element and grips it in place. You will now see why the point was filed off! Alternatively, you can glue the elements in place.

Testing

Place the aerial on a mast, clear of obstructions. Connect it to a transceiver with a length of coaxial cable, with a standing-wave-ratio (SWR) meter in circuit. Find a clear frequency, identify your transmission and ask if the frequency *really is* clear. If so, key the transmitter again and note the SWR. **Do not stand in front of any aerial when it is radiating!** The length of the driven element must be adjusted to obtain an SWR of less than 2. If you have to shorten the dipole, bend the ends over rather than cut them off. That way, if you go too far, your can lengthen them again! The dipole was initially cut too long intentionally, to allow for adjustment here! Bending the ends over also reduces the risk of physical damage to clothing, skin and eyes. You may like to consider applying the same technique to the reflector elements for that reason alone.

Moving on . . .

Once you have warmed to the idea of the corner reflector as an aerial, you might like to ring the changes regarding the reflector. How about replacing the 20 reflector elements with a wire mesh, such as garden centres sell as ‘chicken wire’? Choose the finest mesh if there is a choice. Some extra support may be needed around the edges of the mesh, but you could go on to make a comparison of aerial gain between the two types, using the *UHF Field Strength Meter* described elsewhere in this book.

Materials

Stiff wire or thin-walled tubing for dipole and reflector
Frame – wood, 15 mm by 20 mm cross-section, lengths given in text
Back plate – stout plywood, dimensions given in text
U-bolt to suit mast
Wood screws
Wood glue
50 W coaxial cable for feeder
2-terminal ‘chocolate block’ for dipole connection to feeder
Varnish

67 A switched dummy load

Introduction

A dummy load is a pure resistor of value 50 ohms which can replace your transmitting aerial and enable you to operate the transmitter for test purposes without radiating a signal. It sounds simple enough, but there are two main problems. Firstly, it is impossible to delve into your junk box and emerge with a resistor that will dissipate 100 W PEP and still retain its marked resistance value. Secondly, a ‘pure resistance’ is very difficult to achieve. A pure resistance is a device which has resistance but no reactance. All common resistors have significant reactance at radio frequencies, particularly the wire-wound varieties, which have a helical (i.e. wound like a coil) construction. This is particularly annoying, because wire-wound construction is normally used for large-wattage resistors.

Although all resistors have *some* reactance, not all are quite as bad as the wire-wound type. Carbon film resistors are made by depositing a thin film of carbon on the surface of a small, hollow ceramic cylinder, the thickness of the film of carbon determining the value of the resistor. Provided the lead lengths are kept short, these resistors have a tolerably small reactance, and will be used in this project.

Bearing the load

A 2 W carbon film resistor is hardly going to withstand our 100 W PEP of SSB, so it is obvious that the design of our dummy load must be a little more complex than a single resistor and a switch, despite what **Figure 1** might suggest! In fact, it uses 20 resistors, each of value 1000 Ω (1 kΩ). How does this solve our problem?

Perhaps a little theory is in order here, but no more than is required by the Radio Amateurs' Examination.

When two *equal* resistors of value r are combined in parallel (i.e. side by side), the total resistance, R_T, is given by:

$$\frac{1}{R_T} = \frac{1}{r} + \frac{1}{r}.$$

Adding $1/r$ to $1/r$ gives $2/r$, therefore:

$$\frac{1}{R_T} = \frac{2}{r}, \quad \text{i.e. } R_T = r/2.$$

So, by connecting *two* equal resistors in parallel, we get a combined resistance which is *half* the individual resistances. If we combine *three* in parallel, we get a *third* of the resistance, and so on.

Here, we are connecting *20* resistors of 1 kΩ in parallel, so we will produce an overall resistance of *one-twentieth* of the individual resistance, i.e. 1000/20 = 50 Ω, which is what we set out to achieve!

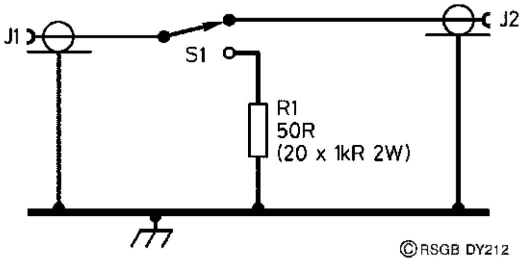

Figure 1 Basic circuit diagram

This is not the only advantage, though. Each resistor is capable of dissipating 2 W, so 20 of them will safely dissipate 40 W, for short periods at least. This power dissipation is approximately the same as 100 W PEP of normal speech, so the design should be capable of use in an 'average' amateur station. The power-handling ability of any dummy load can be improved by providing a 'heat sink' which helps to conduct the heat energy away from the resistors, thus lowering their temperature. One popular heat sink is a can of transformer oil, into which the resistors are immersed. This design uses a rather more mundane heat sink, but which is adequate for the job in hand.

Figure 2 PCB resistor supports

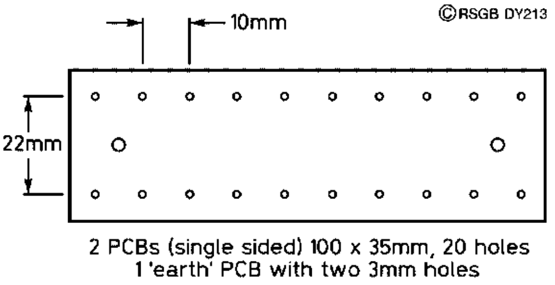
10mm
©RSGB DY213
22mm
2 PCBs (single sided) 100 x 35mm, 20 holes
1 'earth' PCB with two 3mm holes

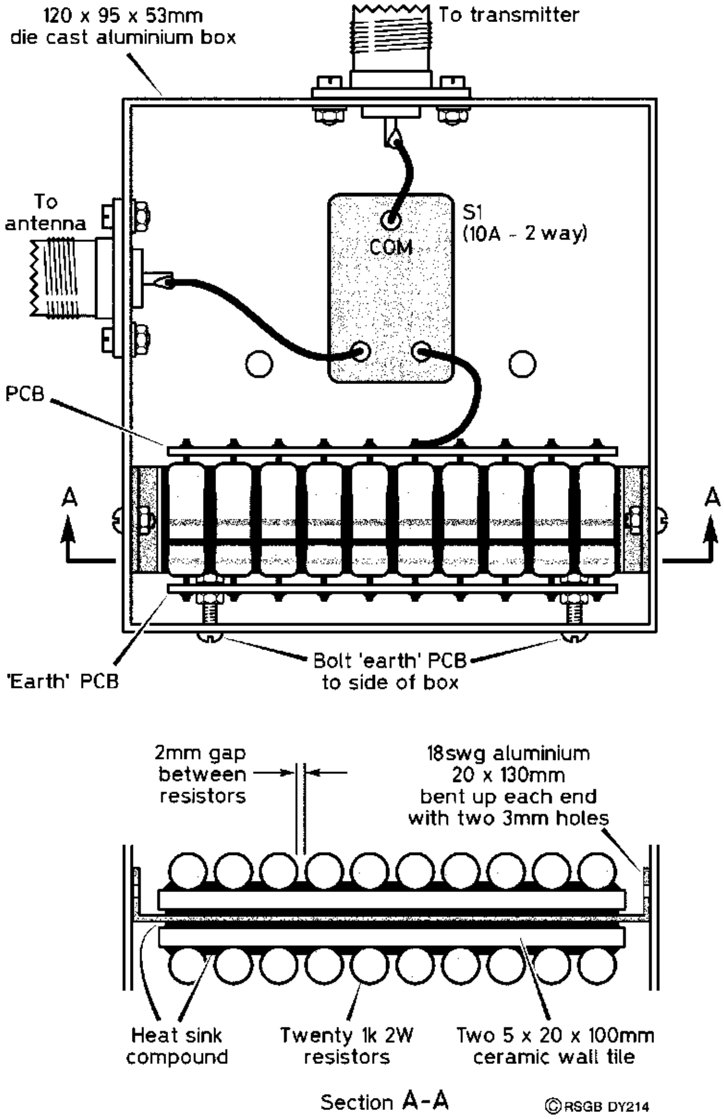
120 x 95 x 53mm
die cast aluminium box
To transmitter
To
antenna
S1
(10A - 2 way)
COM
PCB
A
A
'Earth' PCB
Bolt 'earth' PCB
to side of box
2mm gap
between
resistors
18swg aluminium
20 x 130mm
bent up each end
with two 3mm holes
Heat sink
compound
Twenty 1k 2W
resistors
Two 5 x 20 x 100mm
ceramic wall tile
Section A-A
©RSGB DY214

Figure 3 Switched dummy load, component layout

Pouring the heat away

Two basic PCBs are needed, the details being given in **Figure 2**. Both measure 11 × 35 mm. Each has a set of 20 holes of diameter about 1 mm to take the ends of the resistors. Only one of them has two 3 mm holes which are used for mounting the completed load. **Figure 3** shows how the heat sink is assembled. Solder one end of each resistor into the 'earth' PCB, and then mount this to the box using the two bolts as shown in the diagram. Between each rank of resistors is a 'sandwich' consisting of an aluminium strip and two pieces of ceramic wall tile, to act as a heat sink for the resistors. Heat sink compound is used to provide good thermal contact between the resistors and the tiles, and between the tiles and the aluminium strip. This is shown in the lower part of Figure 3. Thread the loose wires of the resistors through the holes in the second PCB, solder into place, and crop the protruding wires.

Switching

A changeover switch must be used so that the transceiver can be switched between the dummy load and the aerial without unscrewing connectors. For most purposes, an ordinary 10 A 230 V changeover switch will suffice, as found in many electrical shops and DIY stores.

Wire this into the circuit as shown in Figure 1 and Figure 3. Check your wiring. Put S1 in the 'dummy load' position. If you have a multimeter which includes an ohmmeter, measure the resistance across the socket, J1, before connecting it to the transceiver. It should be very near 50 Ω. With S1 in the 'aerial' position, there should be an infinite resistance across J1. Move your ohmmeter to read the resistance across J2. It should be infinite for both positions of S1. If all seems correct, close the box and your dummy load is ready for use!

Parts list

Resistors
  1000 Ω (1 kΩ) 2 W carbon film, 20 required
Additional items
  10 A 230 V changeover (SPDT) switch
  SO–239 sockets, 2 required
  Aluminium box 120 × 95 × 53 mm or similar
  Ceramic wall tile cut as required
  Aluminium strip, 18 SWG
  Nuts and bolts as required

68 A simple Morse oscillator

Introduction

This is an excellent project which uses the 'junk box' as its source of components. If you have trouble in finding the bits for this one, a good source of the components for this and many other similar projects is to be found with the parts list at the end of the project.

The circuit

This is shown in **Figure 1**, and uses an 'unknown' Plessey chip, which makes the overall circuit extremely easy to build. A 0.1 μF capacitor is connected between pins 7 and 8, a speaker (in the popular 8 to 25 Ω impedance range) is connected between pins 8 and 9. If a 9 V battery is connected with its positive terminal to pin 8 and its negative terminal to pin 1, 3 or 5, a tone will be produced in the speaker.

To make this circuit into a good Morse practice oscillator, it is necessary only to insert a Morse key into the supply rail from the battery.

However, there is another refinement which you may care to build into the circuit. The tone from the loudspeaker is different, depending upon which of pins 1, 3 or 5 you use. In the prototype, a single-pole changeover switch was used to select the tones from pin 1 or pin 3, and the Morse key would be connected to the circuit via a small jack socket. There is no need for an on/off switch, as the Morse key performs that function. The switch and the jack socket can be seen in the photograph.

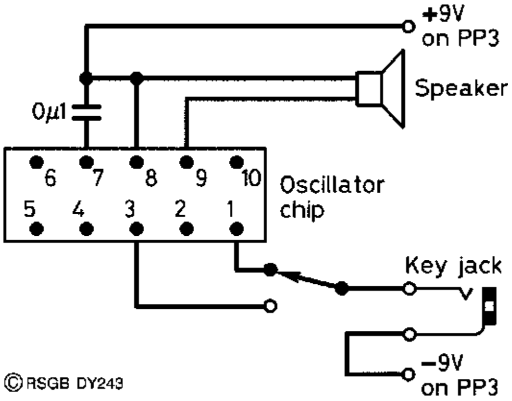

Figure 1 Morse oscillator, circuit diagram

The case

Almost any small loudspeaker will do; the higher the impedance the better. The 'impedance' is the figure usually printed on the rear of the speaker magnet. It is a value given in ohms (Ω). For simple circuits like this, you will usually find that the higher the impedance (within reason), the louder the sound it will produce. Speakers from old transistor radios will work, although their impedances can be rather low sometimes.

Any case big enough to house the components and the battery will be suitable. The prototype used a 'Walkman'-type speaker and case, and is shown in the photograph.

To use the circuit, simply attach a 3.5 mm jack plug to your Morse key, and insert it into the socket. Nothing should happen until you press the key, when a tone should be heard from the loudspeaker.

Another application

Try soldering the two wires of a 'twisted pair' to the 3.5 mm jack plug. Touching the wires together produces a tone from the speaker. This simple circuit can then be used as a 'continuity tester'. Touch the two wires to the ends of a fuse. If there is no sound, then the fuse is blown. There are many other such tests you could perform with this device – to check whether there really *is* a connection between one end of a wire and the other, for example.

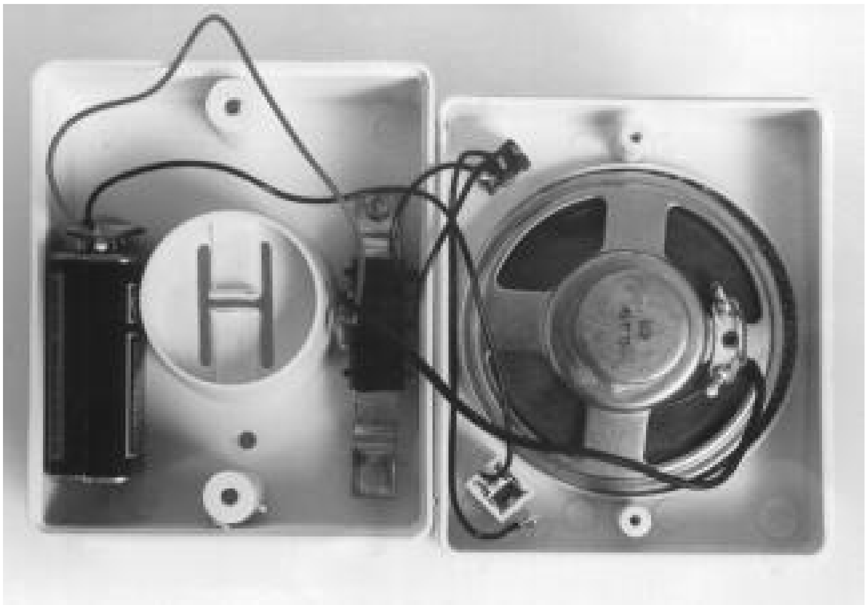

Warning

You must **never** make such tests on any equipment which is connected to the mains supply, even if it is switched off. If you want to make any such tests, make sure you are supervised by someone who understands what you are doing and is competent to advise and supervise you.

Parts list

Plessey oscillator chip
0.1 μF capacitor (disc ceramic)
Loudspeaker 8 to 25 Ω
3.5 mm jack socket (and 3.5 mm jack plug if needed)
Single-pole changeover switch

The Plessey oscillator chip is available from J. Birkett, 25 The Strait, Lincoln LN2 1JF, tel: 01522 520 767.

John Birkett may also provide a kit of parts (the chip, capacitor and loudspeaker).

69 A bipolar transistor tester

Introduction

This is a circuit which will test normal transistors, i.e. npn or pnp. It has the advantage of being able to test devices while they are still connected in their original circuits. However, when such tests are made, **the circuit containing the transistor under test must not be switched on**.

The circuit and how it works

The circuit runs from a 9 V battery such as a PP3 or six AA-type 1.5 V cells. Alkaline cells are to be preferred, as their electrolyte leakage properties are better. The circuit shown in **Figure 1** uses a single CMOS integrated circuit type 4001 or 4011. CMOS circuits require special handling precautions which are described in the project *Christmas Tree LEDs*, elsewhere in this book.

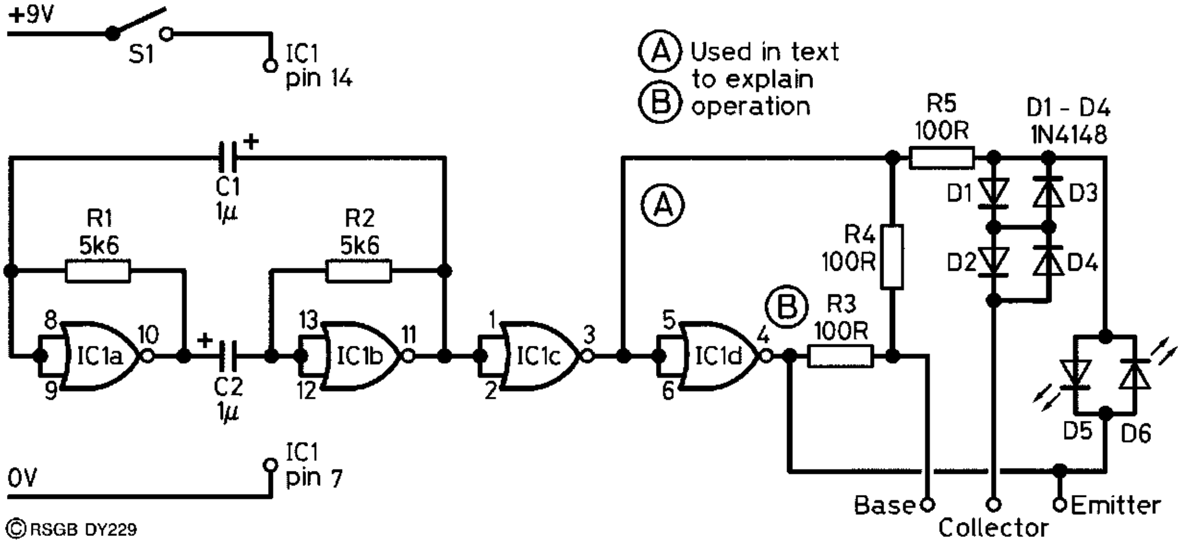
+9V
S1
IC1 pin 14
Ⓐ Used in text to explain Ⓑ operation
C1 1μ
R1 5k6
R2 5k6
IC1a
IC1b
IC1c
IC1d
C2 1μ
8
9
10
13
12
11
1
2
3
5
6
4
Ⓐ
Ⓑ
R3 100R
R4 100R
R5 100R
D1 - D4 1N4148
D1
D2
D3
D4
D5
D6
IC1 pin 7
0V
Base
Collector
Emitter
©RSGB DY229

Figure 1 Transistor tester, circuit diagram

Inside the IC are four logic gates (see *Digital Logic Circuits*) which are all connected as inverters, which means that the output signal is always the logical ‘opposite’ of the input. The first two gates are connected as an oscillator; the circuit being the same as that used in *An Electronic Die*.

The output of the oscillator, at pin 11, is connected to the input of a buffer stage, IC1c, which helps to isolate the oscillator from the circuit that follows it. The buffer output appears on pin 3, which we shall label as *test point A* for future use. Another inverter, IC1d, follows this, its output at pin 4 being labelled *test point B*.

There are two LEDs connected back to back at the circuit output. These are D5 and D6, D5 being red and D6 being green. Whatever the output of the oscillator at any instant, one of the LEDs must be lit and the other unlit. With point A positive and point B zero, the red LED is lit, when A is zero and B is positive, the green LED is lit. Because the oscillator output is repeatedly switching from one polarity to the other, the lit LED is alternately red and green. They switch between the two colours much faster than we can see, so what we think we see are both LEDs lit together.

The two 100 ohm resistors, R3 and R4, provide the bias to the base of the transistor under test. When A is positive and B is zero, the base-emitter junction of the transistor will be forward-biased, and the transistor will switch on (if it is a working npn type). When the transistor is on, it effectively short-circuits D5 (the red LED) and it extinguishes. When point A is zero and B is positive, an npn transistor will be switched off and the green LED (D6) will light. Thus, for a working npn transistor, only the green LED is lit.

If the transistor under test is a pnp type, it will switch on when A is zero and B positive, thus short-circuiting the red LED (D6). When A is positive and B zero, the transistor is off and the red LED (D5) is lit. Thus for a working pnp transistor, only the red LED is lit.

To summarise, the states of the LEDs indicate the following conditions:

- Both LEDs apparently lit: no transistor connected, or transistor permanently open circuit.
- Neither LED lit: a collector-emitter short-circuit is almost certain.
- Red LED alone: pnp transistor in working order.
- Green LED alone: npn transistor in working order.

Construction

The prototype tester was built on a piece of Veroboard measuring 15 strips by 18 holes. Cut the tracks using a track-cutting tool or a 3 mm (⅛ inch) twist drill, as shown in **Figure 2**. Notice that, in this diagram, there is no

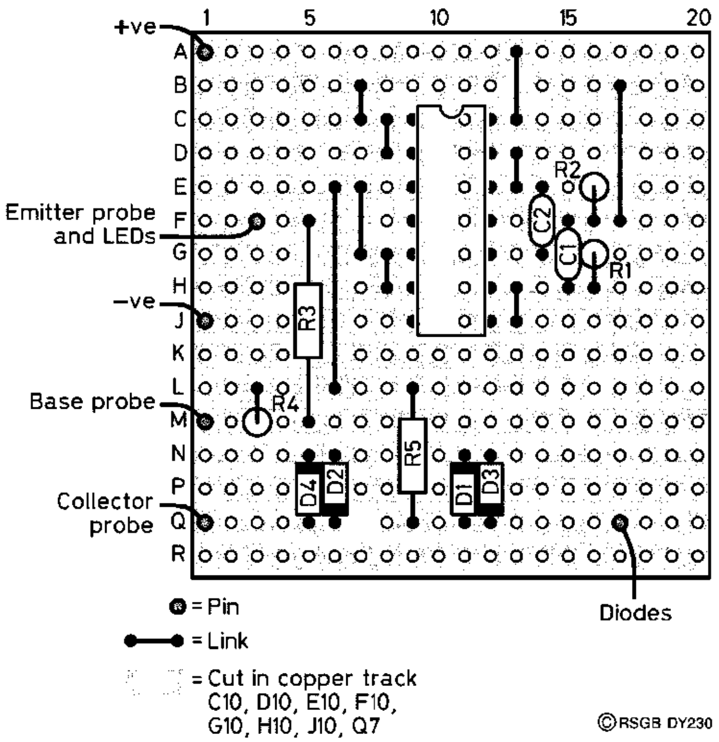
+ve
1
5
10
15
20
A
B
C
D
E
F
G
H
J
K
L
M
N
P
Q
R
Emitter probe and LEDs
–ve
Base probe
Collector probe
R1
R2
R3
R4
R5
C1
C2
D1
D2
D3
D4
Diodes
= Pin
= Link
= Cut in copper track C10, D10, E10, F10, G10, H10, J10, Q7
©RSGB DY230

Figure 2 Transistor tester, component layout

track ‘I’, so try to avoid miscounting when you translate diagram positions to real positions on the board. Hold the board up to a strong light to ensure that there is no copper swarf shorting adjacent tracks together, and that you have made the cuts in the correct places.

Having done this, insert and solder Veropins for all the connections to be made to components not on the board itself. Then solder in the IC holder and the wire links. Then insert the other components in the order resistors, capacitors and diodes D1–D4. Some resistors are mounted vertically so that their connections are on adjacent tracks. Double check the diode polarities – it is easy to make a mistake when wiring diodes in anti-parallel! Next, connect up the off–board components, again making sure that the LEDs have the correct polarity. The probe leads for the emitter, base and collector should be made from different colours of wire and terminated in probe clips (small insulated crocodile clips).

Check carefully for dry joints and errant blobs of solder. Plug the IC into its holder, ensuring that it is inserted the right way round, as shown in Figure 2.

Testing

Without a transistor in circuit, and the battery connected, both LEDs should be lit. Connect a known good npn transistor and verify that the green LED lights. Now simulate two transistor faults: disconnect the base lead and both LEDs should light; remove the transistor and connect the emitter and collector leads together. Neither LED should light.

Repeat the tests with a known good pnp transistor. The results should be the same, except that the correct indication should now be a lit red LED. On your computer, make a small label of the bulleted list above, to fit on your tester showing the states of the LEDs and what they mean. It will act as a useful *aide mémoire* when you use the tester in future.

Using

The circuit will test transistors in isolation or in an existing circuit, i.e. prior to use. You can check the lead identifications in component catalogues such as the Maplin catalogue. The tester is ideal for going through the large bags of unmarked transistors that you can buy for a song at rallies. You can sort them into three piles – npn, pnp and dud!

Parts list

Resistors: all 0.25 watt, 10% tolerance or better

R1, R2	5600 ohms (Ω)
R3, R4, R5	100 ohms (Ω)

Capacitors

C1, C2	1 microfarad (μF) electrolytic, 16 V WKG

Semiconductors

D1–D4	1N4148 general-purpose silicon diodes
D5	Red LED
D6	Green LED

Integrated circuit

IC1	CMOS 4001 or 4011

Additional items
- Veroboard, 15 strips by 18 holes
- Veropins
- PP3 battery and connector (or 6 × AA cells in PP3 clip holder)
- Switch, SPST
- Connecting wire
- One each of known working npn and pnp transistors for test purposes.

Source

Components are available from Maplin.

70 The ‘Yearling’ 20 m receiver

Introduction

Published to celebrate the first anniversary of *D-i-Y Radio*, this excellent receiver design forms a suitable ‘second receiver project’ for those who have successfully completed the MW receivers earlier in this series. The receiver is powered from a PP3 battery or from a mains adaptor, and can be built with the help of an experienced constructor, on a prototype board. The circuit diagram and some of the components used are shown in the separate diagram. Headphones or a loudspeaker can be used and, once the radio is completed, a few simple adjustments will make the *Yearling* spring to life!

Building the receiver

Before starting the constructional process, start by identifying all the parts. One by one, tick them off against the parts list. Are their values correct? The varactor diode is a twin type (see circuit diagram overleaf), and must be cut *carefully* down the middle, producing two devices, D1 and D2, with two wires each.

First, solder the IC sockets, followed by the coils (inductors); L1 is pink inside the top, and L2 is yellow inside. Then, solder in the varactors; the lettering on D1 should be next to coil L1, and the lettering on D2 should face resistor R7. After those, the capacitors, wire links and resistors should be soldered to the board. Take care to wire the voltage regulator, IC3, correctly. Solder in the crystal X1 as quickly and deftly as you can – crystals do *not* take kindly to having their leads bent and being fried with a soldering iron! Make sure that the electrolytic capacitors C2, C12, C15 and C16 are fitted the right way round. Most electrolytic capacitors have only the negative lead marked.

Figure 1 shows the rear of the front panel, illustrating the connections from the board and antenna socket to the controls. All normal connecting wires are 22 SWG or thereabouts, with insulation. Their lengths should be about 15 cm, except for the battery lead to the switch, which is about 8 cm. It is recommended that you use different-coloured wire for each connection to a control. **Figure 2** shows this. The variable resistor section of VR5, the AF gain control, uses single screened cable connected to 0 V (ground) at the

Amateur Radio

D-i-Y

Year

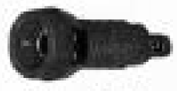
Antenna Socket

Variable Inductor

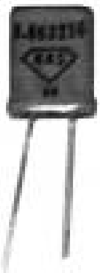
Crystal

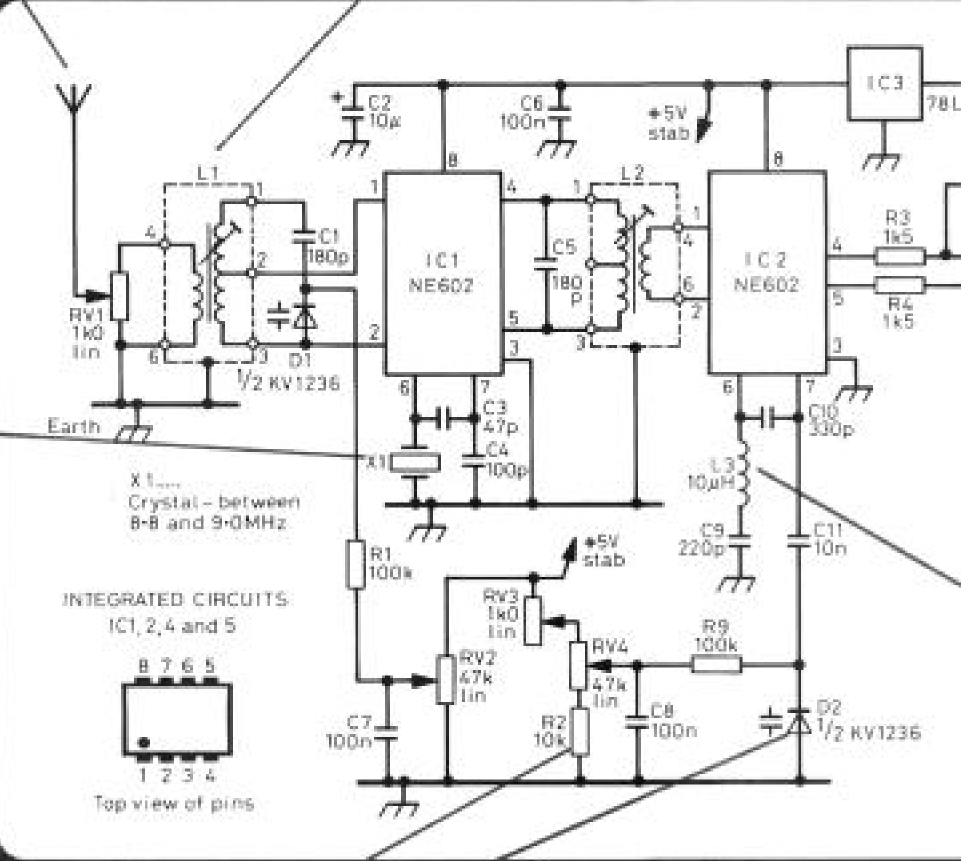

C2
10µ
C6
100n
+5V
stab
IC3
78L
L1
4
1
2
6
3
C1
180p
RV1
1k0
lin
D1
1/2 KV1236
Earth
IC1
NE602
8
4
5
2
1
3
6
7
C3
47p
C4
100p
X1
L2
C5
180
P
IC2
NE602
R3
1k5
R4
1k5
C10
330p
L3
10µH
C9
220p
C11
10n
X1....
Crystal - between
8-8 and 9-0MHz
R1
100k
+5V
stab
RV3
1k0
lin
RV4
47k
lin
R9
100k
RV2
47k
lin
C7
100n
R2
10k
C8
100n
D2
1/2 KV1236
INTEGRATED CIRCUITS
IC1, 2, 4 and 5
8 7 6 5
1 2 3 4
Top view of pins

Resistor

Varicap Diode

Most radio amateurs enjoy building par
is Do-it-Yourself (D-i-Y) Radio, and it is
successfully built yourself. License
allowed (by the Government) to use ho
can see how radio components are dra
circuit diagram for the

D-i-Y Radio, RSGB, Lambda House, Cra

– Do It Yourself

RADIO

rling

Polyester Capacitor

Electrolytic Capacitor

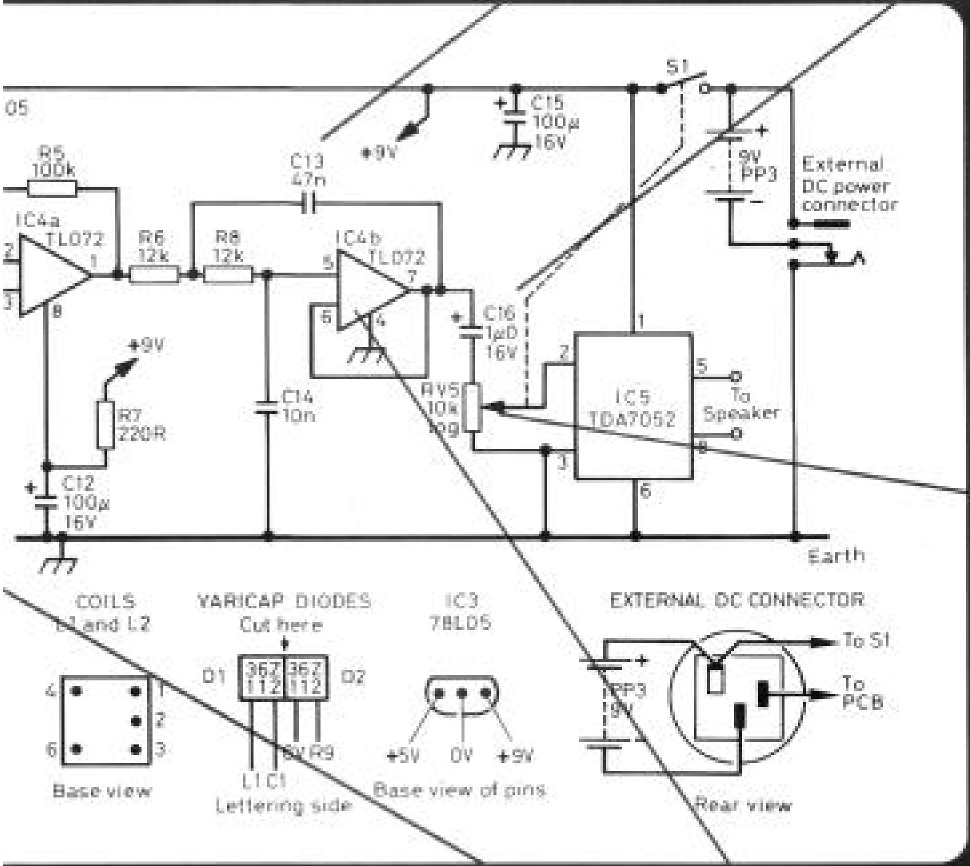
05
R5
100k
IC4a
TL072
2
3
1
8
R6
12k
R8
12k
C13
47n
+9V
C15
100µ
16V
S1
9V
PP3
External
DC power
connector
IC4b
TL072
5
6
7
4
C16
1µ0
16V
+9V
R7
220R
C14
10n
RV5
10k
log
IC5
TDA7052
1
2
3
5
6
8
To
Speaker
C12
100µ
16V
Earth
COILS
L1 and L2
4
1
2
6
3
Base view
VARICAP DIODES
Cut here
D1
367
112
367
112
D2
L1 C1
0V R9
Lettering side
IC3
78L05
+5V
0V
+9V
Base view of pins
EXTERNAL DC CONNECTOR
To S1
To
PCB
PP3
9V
Rear view

AF Gain Control

of their radio stations themselves, this
very satisfying to use a radio you have
radio amateurs are the only people
me-built transmitters. On this page you
wn on a circuit diagram. This is the full
D-i-Y Radio Yearling.

Fixed Inductor

Integrated Circuit

nborne Road, Potters Bar, Herts. EN6 3JE

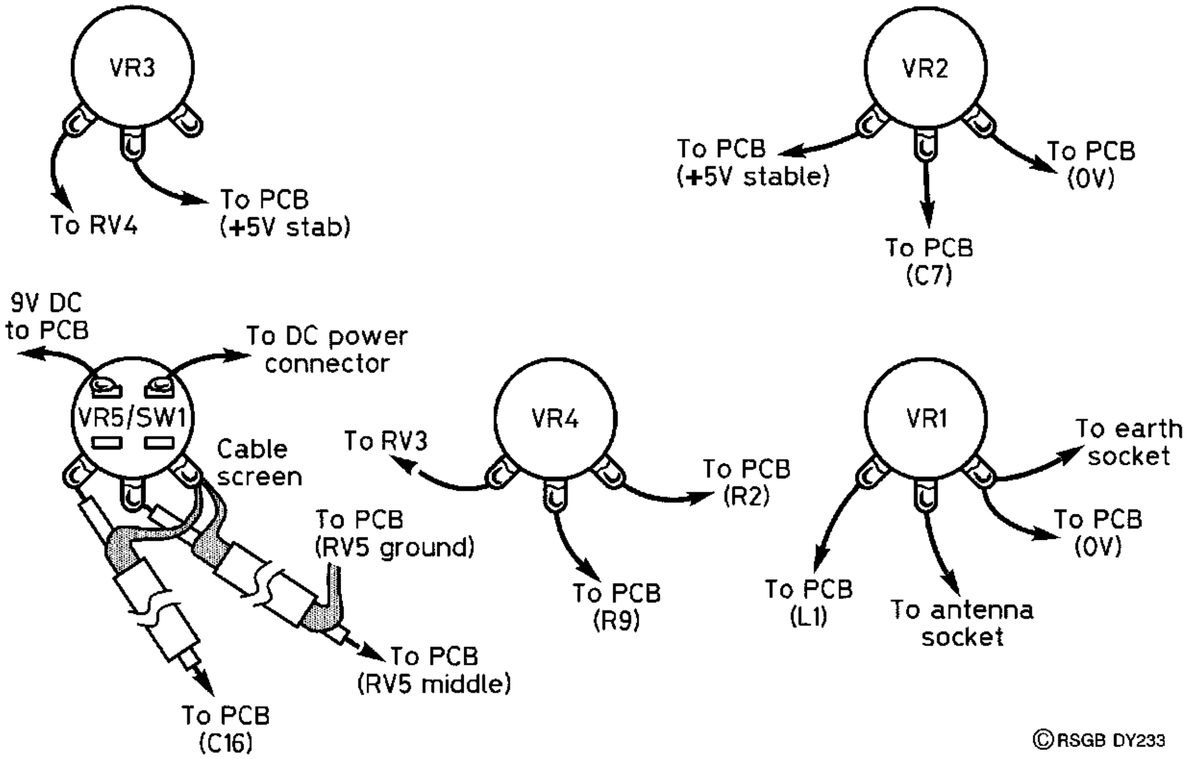
VR3
To RV4
To PCB
(+5V stab)
VR2
To PCB
(+5V stable)
To PCB
(0V)
To PCB
(C7)
9V DC
to PCB
To DC power
connector
VR5/SW1
Cable
screen
To RV3
VR4
To PCB
(R2)
VR1
To earth
socket
To PCB
(RV5 ground)
To PCB
(0V)
To PCB
(R9)
To PCB
(L1)
To antenna
socket
To PCB
(RV5 middle)
To PCB
(C16)
©RSGB DY233

Figure 1 Rear view of the variable resistors. Check the connections carefully to make sure the wires fit the correct holes

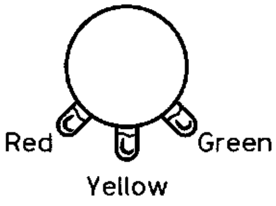
Red
Green
Yellow

©RSGB DY234

Figure 2 It is helpful to use different colour insulated connecting wires. Wires between each variable resistor and the board should be twisted together to give a neat wiring outfit.

board end. Now fit the ICs into their sockets. Make sure that they are the right way round (see circuit diagram) and that each pin lies directly above its corresponding socket before applying *gentle* pressure with the back of the board firmly supported.

Check that all the connections are correct (don’t *assume* this – check the ends of each wire against the circuit diagram) and that all your soldered joints are shiny. Lastly, drill the five 10.5 mm diameter holes for the main controls. On the side of the case, a 6.3 mm diameter hole is needed for the speaker socket, and 8 mm holes for the antenna and earth sockets. The external power socket requires an 11 mm diameter hole.

Adjusting it

Before you can make adjustments, and in order to hear *anything* on your *Yearling*, you will need to connect an antenna (aerial) to the antenna socket. About 8 metres of wire, preferably outdoors and as high as possible, is all you need to connect to the socket. Connect the 3.5 mm jack socket for the speaker, and a 3.5 mm jack to your speaker leads. Connect a battery and switch on.

1. Using a very small screwdriver or, better still, a non-metallic ‘trimming tool’, *gently* screw in the core of L2 as far as it will go, but *don’t force it*. Then unscrew it by three turns anticlockwise.
2. Set VR1, VR2 and VR4 to mid-position and rotate the core of L1 anticlockwise until the hissing noise you hear reaches a maximum intensity. Then adjust L2 for maximum noise.
3. If you now tune carefully with the main tuning control, VR4, you should hear some amateur Single-Sideband (SSB) speech signals. You may have to adjust the bandspread (fine tuning) control, VR3, to make the speech sound normal.
4. Having verified that everything is working, switch off and mount the controls on the front panel and the sockets on the side. To do this, it is much safer to disconnect all the controls and sockets, mount them in their final positions, and then wire them up again.
5. Fit the front panel knobs, connect your aerial and switch on again, checking that everything is working. Then, locate the cluster of CW (Morse) signals to be found at the bottom of the 20 metre band, set the bandspread control to mid-position and slacken off the main tuning knob. Turn the knob (but *not* the control!) until the pointer lies a little clockwise of 14.0 MHz. The SSB signals should now lie roughly between the dial centre and 14.35 MHz. Tighten up the knob.
6. Finally, fix the board to the rear panel, and secure the battery (if you are using one). Attach the rear panel to the back of the box, and you are finished!

Listening!

Make a habit of keeping the bandspread control in its centre position when searching for stations; then you can adjust it either way to make the signals readable. Unless you have a very big aerial, it is best to have the 'RF Gain' control, VR1, at maximum. Use the 'Antenna Tune' control, VR2, to give the best signal, and control the volume with the 'OFF/AF Gain' control. You will find some excellent DX stations with your *Yearling* receiver, and it will serve you well.

Parts list

Resistors (all 0.25 W, 5%)

R1, R5, R9	100 kilohms (kΩ)
R2	10 kilohms (kΩ)
R3, R4	1.5 kilohms (kΩ)
R6, R8	12 kilohms (kΩ)
R7	200 ohms (Ω)

Capacitors (all rated at 16 V or more, tolerance *at least* what is quoted)

C1, C5	180 picofarads (pF) polystyrene 5%
C2	10 microfarads (µF) electrolytic
C3	47 picofarads (pF) polystyrene 5%
C4	100 picofarads (pF) polystyrene 5%
C6, C7, C8	100 nanofarads (nF) or 0.1 microfarad (µF) ceramic
C9	220 picofarads (pF) polystyrene 2%
C10	330 picofarads (pF) polystyrene 2%
C11, C14	10 nanofarads (nF) or 0.01 microfarad (µF) ceramic
C12, C15	100 microfarads (µF) electrolytic
C13	47 nanofarads (nF) or 0.047 microfarad (µF) polyester, 5%
C16	1 microfarad (µF) electrolytic

Variable resistors

VR1, VR3	1 kilohm (kΩ) linear
VR2, VR4	47 kilohms (kΩ) linear
VR5	10 kilohms (kΩ) log with switch

Inductors

L1	Toko KANK3335R
L2	Toko KANK3334R
L3	10 microhenries (µH), 5%

Semiconductors	
IC1, IC2	Philips/Signetics NE602 or NE602A
IC3	78L05 5 V 100 mA regulator
IC4	TL072 Dual Op-Amp
IC5	Philips TDA7052 audio amplifier
Additional items	
D1, D2	Varactor diode Toko KV1236 (cut into two sections – see text)
X1	Crystal 8.86 MHz type (from Maplin, etc.)
4 off	8-pin DIL sockets for IC1, IC2, IC4 and IC5
2 off	4 mm sockets aerial (red) and earth (black)
1 off	3.5 mm chassis-mounting speaker jack socket
1 off	DC power socket for external supply (if required)
4 off	Red knobs with pointers
1 off	Tuning knob with pointer (e.g. 37 mm PK3 type)
1 off	Printed-circuit board or prototype board
1 off	Plastic case approx 170 × 110 × 6 mm (e.g. Tandy number 270-224)
1 off	Speaker 8–32 Ω impedance (or headphones)

71 Adding the 80 metre band to the Yearling receiver

Background

You will have noticed that your *Yearling* receiver has a dial which shows coverage of the 80-metre amateur band (3.5–3.8 MHz). This band is used for local contacts during the day, and contacts up to about 1600 miles in darkness. Longer distances are possible, particularly in the middle of winter.

The modifications

Only a few extra parts are required, as you may have noticed from the parts list. A low-pass filter, FL1 (one which passes low frequencies and rejects higher frequencies), is switched into the circuit on 80 m. The circuit of the

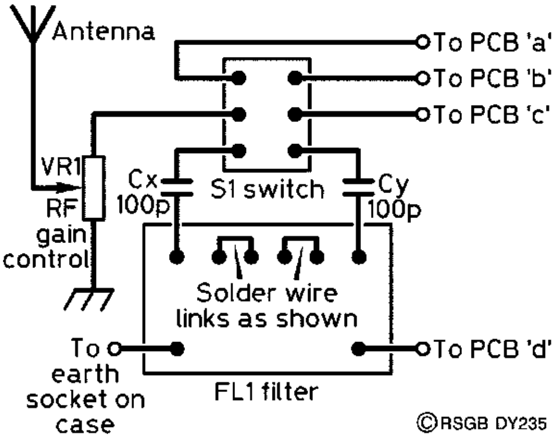
Antenna
To PCB 'a'
To PCB 'b'
To PCB 'c'
VR1
RF gain control
Cx 100p
S1 switch
Cy 100p
Solder wire links as shown
To earth socket on case
To PCB 'd'
FL1 filter
©RSGB DY235

Figure 1 The circuit diagram shows the extra components for 80 m operation. Note the connections to the PCB 'a', 'b', 'c' and 'd'

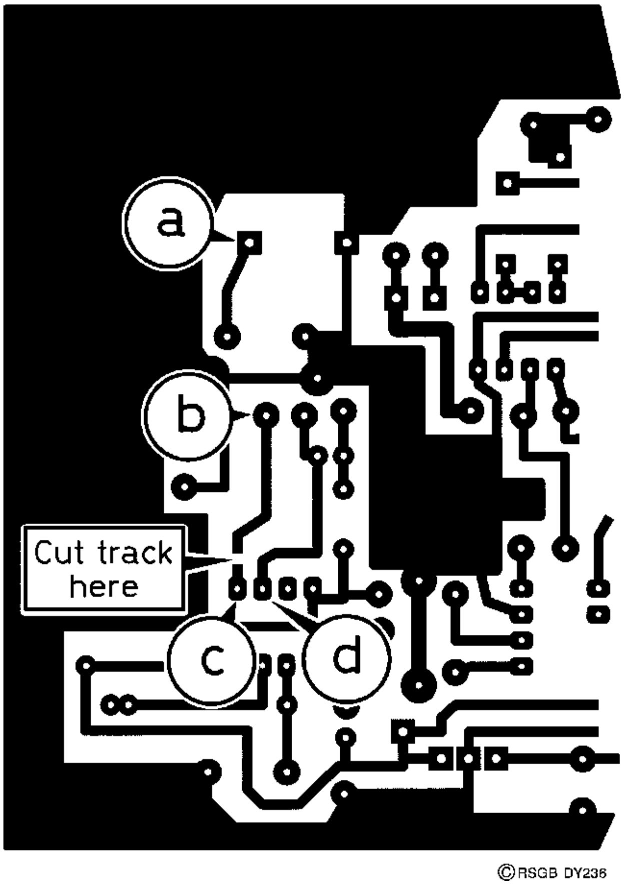
a
b
c
d
Cut track here
©RSGB DY236

Figure 2 The underside of the PCB. Wires are connected from the switch and filter as shown

switch and its connections is shown in **Figure 1**. Before making the modifications, dismantle the receiver so that you have easy (and safe) access to the case and the track side of the PCB.

1. The first thing to do is to drill a 6.5 mm diameter hole in the side of the case into which the switch fits.
2. Then, using a sharp Stanley knife or scalpel, carefully cut the track on the PCB as shown in **Figure 2**, making a gap about 1 mm wide.
3. Using 10 cm lengths of different-coloured insulated wire, make the four connections, a, b, c and d, to the PCB, as shown in Figure 2.
4. Solder the two links on the filter, and then make the connections to the switch, capacitors and PCB. You will have to disconnect the existing wire between the RF gain control, VR1, and pin 4 of L1 on the PCB.
5. Lastly, check your new connections carefully, then mount the switch in the new hole and fix the filter to the bottom of the case with a little glue, as shown in **Figure 3**. Reassemble the circuit, and replace the back of the case.

More testing!

Firstly, switch your new switch, S1, into the 20 m position, to check that the original circuit still works! If you find that the ‘Antenna Tune’ control peaks at a slightly different position, don’t worry.

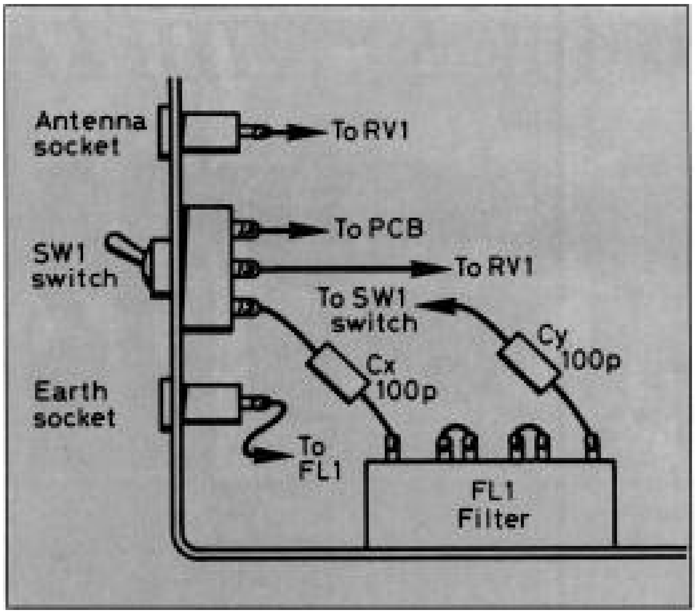

Figure 3 Internal view of the Yearling case. The filter FL1 is attached to the base with glue

Now switch to 80 m, and tune around the anticlockwise end of the dial; you should hear some SSB stations, particularly in the evenings and at weekends, when may people are on the air. At the other end of the travel of the tuning control, you should hear CW (Morse) stations.

The Radio Society of Great Britain broadcasts amateur radio news every Sunday morning on or about 3.65 MHz; the table below has the details. Finally, a good antenna is more important than ever for 80 metre reception – aim for more height and length, and then consider the project concerned with making an Antenna Tuning Unit (ATU)!

Parts list

Capacitors (all rated at 16 V or more, tolerance 10% or better)

Cx, Cy	100 picofarads (pF) polystyrene

Filter

FL1	Toko 237LVS1110 low-pass filter

Additional items

SW1	2-pole 2-way (changeover) toggle switch
7 off	Short lengths of insulated wire of different colours

The RSGB news broadcasts, GB2RS – Sunday mornings

Frequency (MHz)	*Local time*	*Reception area*
3.650	0900	SE England
3.650	0930	Midlands
3.650	1000	SW England
3.650	1100	Yorkshire
3.640	1130	Aberdeen
3.660	1130	Glasgow

The Midlands transmission is repeated at 1800 (6pm) local time on 3.650 MHz. All frequencies are approximate in order to avoid interference, and use lower sideband (LSB). If you also have a 40 metre receiver, there are GB2RS news broadcasts on 7.048 MHz at 0900 local time from Northern Ireland and from 1100 local time from the north midlands.

72 How the Yearling works

Introduction

The *Yearling* was designed to provide an introduction to *Amateur Radio* on the 20 m amateur band. Let's look at how the different sections (or 'stages') of a radio work, and how they fit together to form a complete receiver. **Figure 1** shows a block diagram which you can follow and compare with the circuit diagram of your *Yearling* receiver.

The antenna (or aerial)

Connected to your receiver, it will pick up not only amateur signals, but all other signals as well! This means that the receiver has to select the one signal that interests you, while rejecting all the others. The following stages do just that.

The RF filter

This stage (centred around L1) selects the band of radio frequencies (RF) containing the signal you want, in this case, those having wavelengths around 20 m. Signals from the 40 m band, for example, would not get through.

The crystal oscillator

This is an oscillator circuit designed around a quartz crystal (X1), and has a very stable frequency. It produces a single, very pure frequency to feed into the mixer. A crystal having a frequency between 8.800 MHz and 9.000 MHz is suitable for this circuit. The oscillator and mixer functions are both carried out inside IC1.

The first mixer

Yes, this stage 'mixes' two signals together. In this case, the two signals are (i) from the aerial via the RF Filter, and (ii) from the crystal oscillator. Two

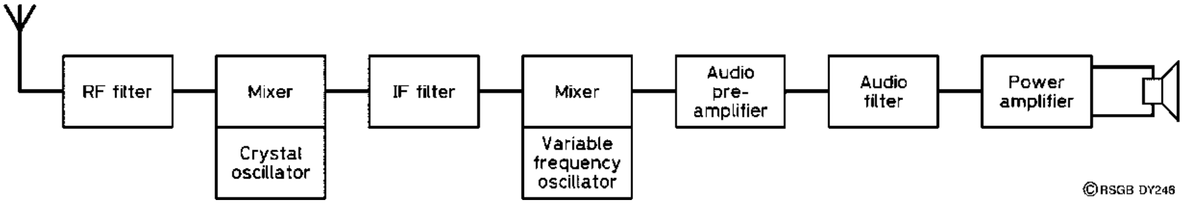

Figure 1 Block diagram of the Yearling, showing how the various stages fit together to make a complete radio receiver

bands of signals emerge from the mixer. The first is centred upon a frequency equal to the incoming signal frequency **added** to the crystal frequency, and the second is centred upon a frequency equal to the incoming signal frequency **subtracted** from the crystal frequency. Look at an example – if the signal is at 14 MHz and the oscillator at 9 MHz, then the mixer outputs will be 14 + 9 = 23 MHz and 14 – 9 = 5 MHz.

Intermediate frequency (IF) filter

It is the purpose of the IF filter (centred around L2) to select only one of these two bands of frequencies emerging from the mixer. In this case, it is the lower band of frequencies (around 5 MHz) which we select. This is because, in general, lower frequencies are easier to handle than higher ones.

Variable-frequency oscillator (VFO)

The VFO (part of IC2) enables us to tune into a particular station, and operates over a band of frequencies between 5 MHz and 5.35 MHz in this receiver. You will notice that IC1 and IC2 are the same type of chip, so that you will be expecting another mixer stage to be associated with the VFO. You are quite right!

The second mixer

This mixer obeys exactly the same rules as those of mixer 1. Sum and difference frequencies are produced, like this. Mixing is between the incoming IF signals (around 5 MHz) and the VFO signals (around 5 MHz), producing output frequency bands centred upon 10 MHz and 0 MHz. The use of the words ‘band of frequencies’ throughout this explanation is intentional. If all the signals were pure, there would be no bands; the bands are produced because of one thing – the modulation imposed on the pure frequencies at the transmitter. So, the 'bands' contain the one thing that we want to extract from the signal, and that is the speech or Morse code that the signal contains. The band of frequencies at 0 MHz is just that – the audio frequencies we want in the loudspeaker. Because of this, the audio frequency output of the second mixer is selected and passed on for amplification.

The audio preamplifier

Preliminary amplification of the minute audio signal which emerges from the second mixer is provided by IC4a, which will respond only to audio signals, automatically rejecting the 10 MHz signal.

The audio filter

The bandwidth of normal speech when transmitted by an amateur station is around 3 kHz, so there is no advantage to be gained in amplifying frequencies greater than this. IC4b is known as a low-pass filter, because it passes (lets through) lower frequencies and rejects higher ones.

The power amplifier

IC5 produces the final audio amplification and provides enough power (about 350 milliwatts (mW) to drive a small speaker.

How does it work on 80 m?

If you have fitted the 80 m modification to your receiver, you are probably wondering how the circuit works at this different frequency. Firstly, the filter which you fitted selects the 80 m band instead of the 20 m band. The only other slight difference lies in the way the first mixer stage works. Its job is to produce sum and difference frequencies from the incoming signal and crystal frequencies. On 20 m, it did this by subtracting the crystal frequency (9 MHz) from the incoming frequency (14 MHz) to produce an IF output of 5 MHz. On 80 m, the incoming frequency (3.5 MHz) is subtracted from the crystal frequency (9 MHz) to produce an IF output of 5.5 MHz, which is still within the tuning range of the VFO in the next stage.

73 A field strength meter

Introduction

Having a reliable field strength meter is always useful. There are two basic types – the tuned variety (also known as an absorption wavemeter) and the broad-band variety. A project centred on the tuned variety can be found in this book (*An HF absorption wavemeter*), and the way it works is also presented (*The absorption wavemeter*). This project describes an untuned or broad-band type which is every bit as useful as the tuned type, when it is the level of RF in which you are principally interested.

With the possible exception of the meter, you may have all the parts needed in your junk box. The prototype was distinguished by its meter, found at a rally, and marked 'Safe', 'Dangerous' and 'Explosive' on its scale.

How it works

The basic field strength meter uses the circuit of a crystal set, but with a meter replacing the headphones. A better design, which is used here, is that of a voltage doubler, giving more sensitivity. **Figure 1** illustrates the voltage doubler circuit.

The AC input shown will be our RF input, which will be explained soon. Diode D1 will pass the positive half of the signal and use it to charge up C1 to the peak value of the signal. D2 uses the negative half of the signal to charge up C2 to the same value. Because C1 and C2 are in series, the peak voltage appearing across *both* of them (which is the DC output voltage) is equal to *twice* the peak input voltage, hence the name voltage doubler. If you're wondering why DC is present at the output when AC comes in at the

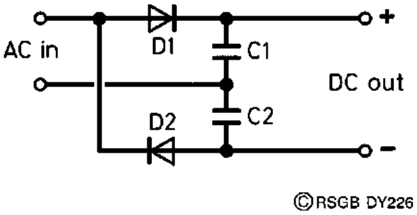

Figure 1 Voltage doubler circuit

input, remember that capacitors pass AC (RF) but block DC; thus the RF is shorted out through C1 and C2, but the steady voltage (DC) remains across the two capacitors. The voltage is thus proportional to the size of the RF signal applied at the input.

The circuit

The voltage doubler is converted into a field strength meter using the circuit of **Figure 2**. A piece of wire serves as the aerial to provide an RF signal across the radio-frequency choke (RFC). A choke is an inductor which is large enough to prevent the RF passing through it – it 'chokes' the RF. This produces the maximum RF signal at the input to the voltage doubler, and the DC output from it is measured on the meter.

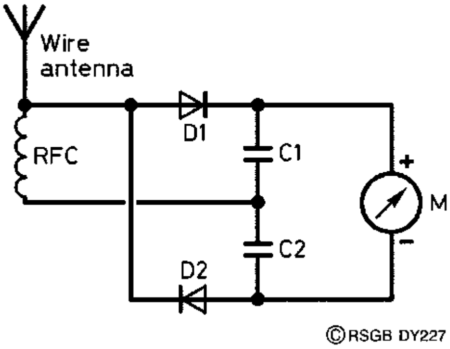

©RSGB DY227

Figure 2 Field strength meter

The parts used are not very critical. The RF choke can have any value between about 1 mH and 2.5 mH. Almost any common diode such as the 1N914 or 1N4148 can be used for D1 and D2. The two capacitors could be any value between 1 nF and 100 nF (0.001 μF and 0.1 μF). The meter should be reasonably sensitive, with a full-scale deflection (FSD) in the range 50 μA to 100 μA. Look for VU meters at rallies – these are ideal.

Construction

The prototype circuit was made on matrix board. If you don't want to use pins with the board, simply push the component leads through the holes and make the connections on the underside of the board, either with the excess component leads themselves, or with ordinary connecting wire.

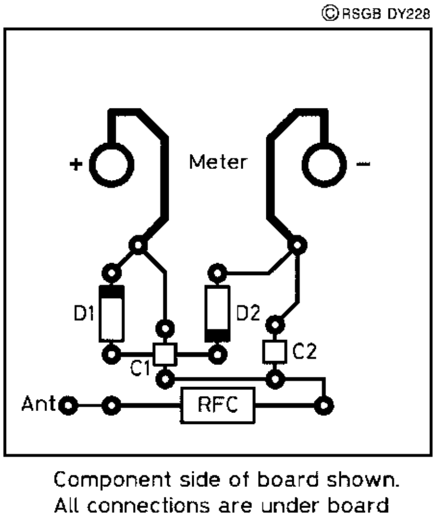
©RSGB DY228
Meter
+
−
D1
D2
C1
C2
Ant
RFC
Component side of board shown.
All connections are under board

Figure 3 Board layout and interconnections

Figure 3 shows how the parts are placed and connected. The matrix board can be mounted directly on the meter using its terminal bolts!

In use

Needing no power supply other than an RF signal, just connect it all up and leave it to work! A short length of insulated wire is enough to pick up some RF and display it on the meter. Using a 200 μA meter, about 3 metres of wire gave a good deflection on the meter. To increase deflection, the wire can be wrapped around the aerial lead, provided that the wire is PVC covered and doesn't come into contact with the aerial wire.

This requires a little experimentation. Try a long piece of wire first and adjust its position until the meter needle kicks whenever there is a transmission. It is very reassuring to see the meter moving during a transmission. Although SWR meters also indicate power, they are usually set to read reverse power, and show little or no movement during transmission.

Parts list

D1, D2	1N914, 1N4148 or similar
C1, C2	10 nF disc ceramic
RFC	Miniature axial choke (1 mH)
Meter	Surplus VU meter or similar
Matrix board or similar	

Components are available from Maplin.

74 Preselector for a short-wave receiver

Introduction

A preselector is a simple RF tuned amplifier which is inserted between the aerial and the receiver. It provides some extra gain and may improve the overall performance of the receiver. This project uses a Field-Effect Transistor (FET) amplifier in grounded-gate mode.

The design has a tuned circuit at both the input and output which, with excessive gain and poor construction, would produce only one thing – oscillation! So, to avoid this happening, we will have only a low gain, and use a circuit which provides good isolation between input and output. The grounded-gate FET amplifier fulfils both these criteria. It will also cover a frequency range from about 7 to 30 MHz, which includes most of the HF amateur bands.

The circuit

This is shown in **Figure 1**. The signal from the aerial arrives at an RF transformer, the secondary of which is tuned with capacitor VC1a. The output from the tuned circuit is taken from a tap on the secondary to the source of the FET. The gate is grounded (earthed) and the amplified signal appears at the FET drain, which is then fed to the primary of another RF transformer, which is tuned by VC1b. The output to the receiver comes from the secondary of the RF transformer.

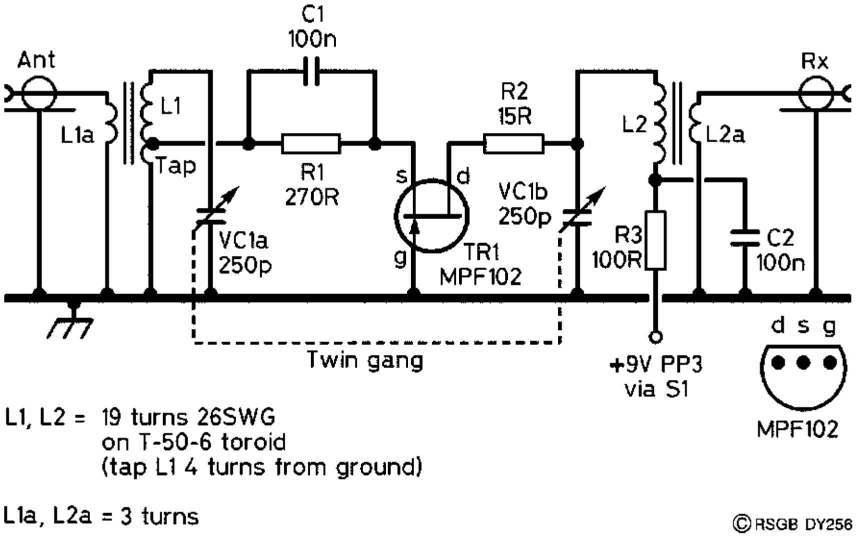
C1
100n
Ant
Rx
R2
15R
L1
L1a
Tap
R1
270R
s
d
g
VC1b
250p
L2
L2a
VC1a
250p
TR1
MPF102
R3
100R
C2
100n
d s g
Twin gang
+9V PP3
via S1
MPF102
L1, L2 = 19 turns 26SWG
on T-50-6 toroid
(tap L1 4 turns from ground)
L1a, L2a = 3 turns
©RSGB DY256

Figure 1 HF preselector, circuit diagram

Notice that the two RF transformers are identical, but they are used 'back to back', with the secondary of the first and the primary of the second being tuned. They are tuned with identical capacitors, fitted on the same shaft of a variable capacitor. We say that the two capacitors are 'ganged'. Because L1 and L2 are the same, and VC1a and VC1b are the same, both RF transformers should be resonant at the same frequency, no matter what that frequency is.

Construction

The final layout should look something like that shown in **Figure 2**. The external connectors and controls being two SO-239 sockets for connection to your receiver and aerial, a tuning control and its associated scale, and an on/off switch.

The circuit can be put together on a plain matrix board, using pins to anchor the components, or simply by pushing the component leads through the board and making connections on the underside. The layout of the prototype is shown in **Figure 3**. Mounting the board to the aluminium box is accomplished with bolts, solder tags and stand-off insulators.

Check your construction against the circuit diagram and against the layout diagram. Wire in the PP3 battery clip, put the switch in the 'off' position, and fit the battery. Testing can be carried out without fitting the top of the box.

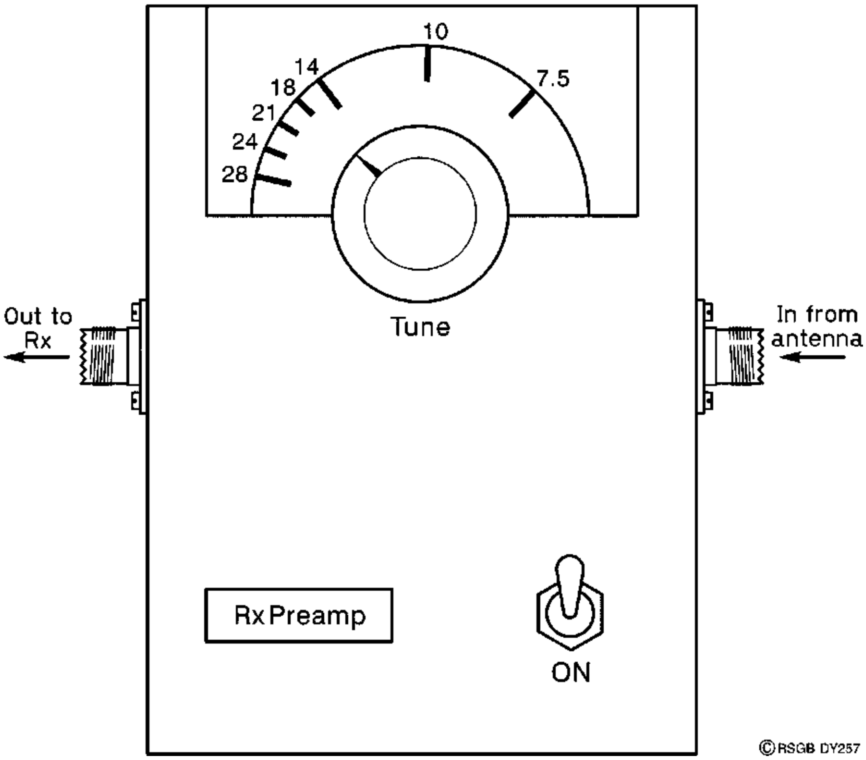

Figure 2 HF preselector, front panel layout

Testing

Don't fit your preselector yet. Tune your radio to a broadcast station, preferably a fairly weak one. Disconnect the aerial and fit your preselector between the aerial and receiver. Switch it on and rotate the tuning knob slowly. You should find a position where your original station is received more clearly than before. If it doesn't work at all, recheck your wiring. Is there a positive voltage on the drain of the FET? If not, work back towards the positive battery terminal. Is there a voltage at the junction of L2 and R2? Is there a voltage at the junction of L2 and R3? Is there a voltage at the junction of R3 and the battery lead? If there isn't a voltage at that point, then you have probably mounted your switch upside down, and it is off, not on! It's a common mistake.

Calibration

This is not obligatory, surprisingly enough. However, if having a frequency scale appeals to you, then using an RF signal generator (or using the services of a friend who has one) is the simplest solution. Feed in a weak modulated

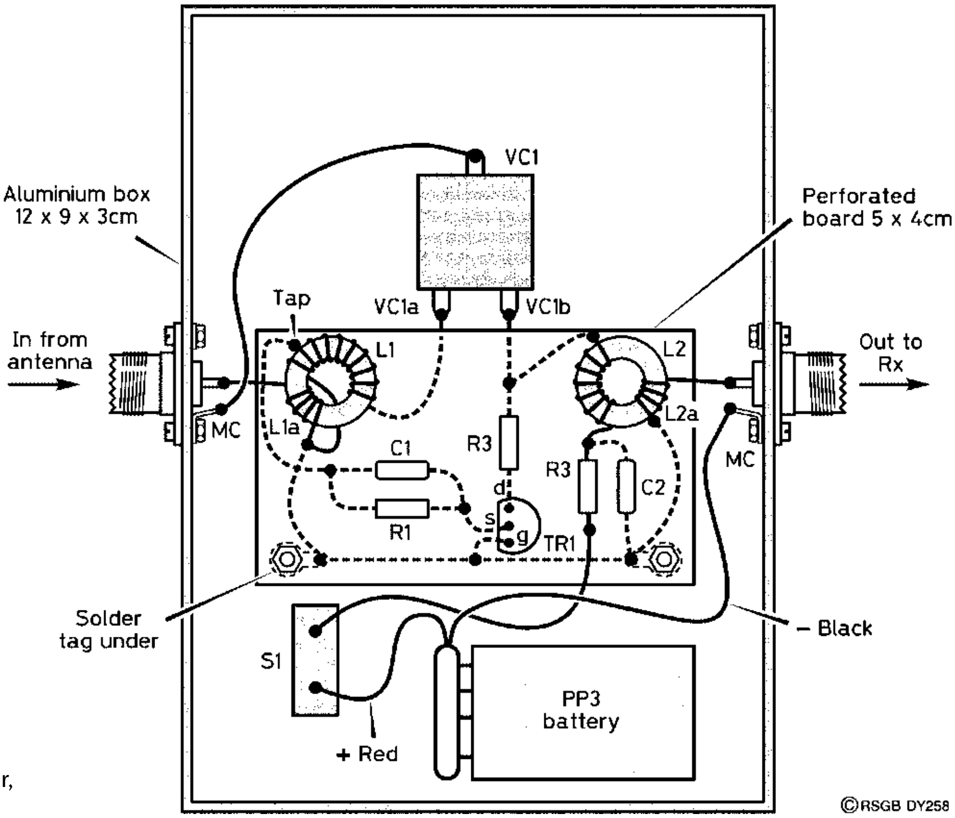
VC1
Aluminium box 12 x 9 x 3cm
Perforated board 5 x 4cm
Tap
VC1a
VC1b
In from antenna
L1
L2
Out to Rx
MC
L1a
L2a
C1
R3
R3
C2
MC
d
s
g
R1
TR1
Solder tag under
– Black
S1
PP3 battery
+ Red
©RSGB DY258

Figure 3 HF preselector, component layout

signal from the generator to the preselector, and rotate VC1 until the signal is maximum. Mark this frequency on your dial. Repeat the process for the frequencies shown on the dial in Figure 2.

If you have a commercial transceiver, feed its output into a dummy load (such as the type described in *A Switched Dummy Load*, elsewhere in this book) using a distinctive modulating signal such as an idling RTTY signal. Set the receiver to the same frequency with the preselector out of circuit. Insert it into the aerial lead, and search for the signal with VC1 until it gives the maximum deflection on your S-meter, then mark the frequency on your dial.

Parts list

Resistors: all 0.25 W carbon film or better

R1	270 Ω
R2	15 Ω
R3	100 Ω

Capacitors

C1, C2	100 nF (0.1 μF) ceramic
VC1	250/250 pF polyvaricon

Inductors

L1, L2	19 turns 26 SWG enamelled copper on T.50.6 toroid L1 has a tap 4 turns from ground end L1a and L2a – 3 turns wound over previous winding

Semiconductors

TR1	MPF102 FET

Additional items

Matrix board	To fit aluminium box (see Figure 3)
Aluminium box	12 × 9 × 3 cm
Battery and connector	PP3 9 V
S1	SPST on/off switch
Knob	As required
SO239	Sockets – 2 required

75 An audible continuity tester

Introduction

This is not the only continuity tester in this book. This alone attests to their use, so you may well want to experiment with several designs, then come up with one of your own! The very simplest form of continuity tester is probably a battery and a bulb in series, with the circuit being closed by connecting it to a fuse or other object being tested for continuity. The bulb could be replaced by a buzzer to give an audible indication. The current taken by the buzzer could damage some components, however. An ohmmeter can also be used, and is very popular for the purpose, as it indicates whether the circuit is low, medium or high resistance. This project has the advantages of indicating whether there is no continuity, some resistance or good continuity, and making an audible sound, so that you don't have to move your eyes from the circuit while making the test.

The circuit

Using only one integrated circuit, four components and a battery, this is a particularly simple circuit, as **Figure 1** shows. IC1 is usually used to flash an LED from a 1.5 V source, and to have a low current consumption. By changing the component values, IC1 is made to oscillate at audio frequencies, and we can hear these through the loudspeaker, LS1. Low-impedance speakers can be used, but result in an increased current drain, so the higher the impedance, the better.

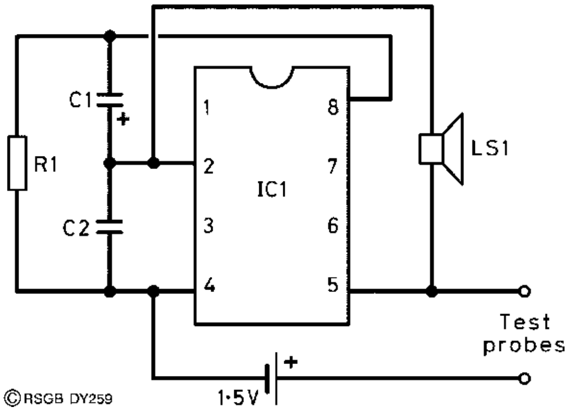

Figure 1 Continuity tester, circuit diagram

Construction

The layout of the circuit on Veroboard measuring 7 strips by 15 holes is shown in **Figure 2**. Start by making the four track cuts which will lie underneath the IC. Use a track cutter or a 3 mm (⅛ inch) twist drill rotated between thumb and forefinger for this. Solder the Veropins in place, followed by the wire link. Then solder in the IC holder, the resistor and the capacitors, making sure that the electrolytic capacitor, C1, is connected the correct way round.

Check, with the board against a bright light, that there are no shorted tracks, either by large blobs of solder or by copper swarf from the track-cutting process. Then insert IC1 into its holder the right way round. The probes can be ordinary connecting wire, the free ends being tinned with solder to prevent wire whiskers from touching components other than the one you are testing.

If you are happy that the circuit and the wiring appear to be correct, fit the battery into its clip. Nothing should happen until the probes are touched together, when you should hear a note from the loudspeaker. If nothing happens, all you can do is to recheck your circuit, as there is nothing else to go wrong!

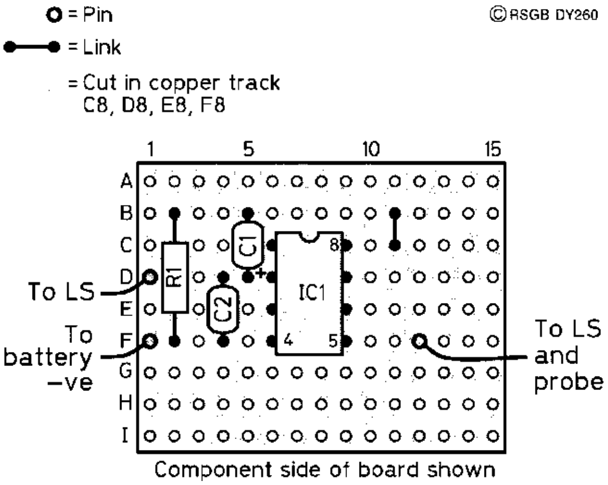
= Pin
= Link
= Cut in copper track
C8, D8, E8, F8
©RSGB DY260
1
5
10
15
A
B
C
D
E
F
G
H
I
R1
C1
C2
IC1
+
8
4
5
To LS
To battery −ve
To LS and probe
Component side of board shown

Figure 2 Continuity tester, component layout

Any box can be used; there is no justification for a metal case, unless you want to make use of something, which is to hand, such as a tobacco tin. Alternatively, any suitable plastic box will do.

Use

The tester will give different pitch notes for resistances of different values. The higher the resistance, the higher the pitch from the speaker. Try it with small inductors, and you will learn to recognise the different tones produced by the IC.

Safety notice

Using the continuity tester on components *in situ* is not advisable, as the results could be misleading. It can be dangerous to make measurements *in situ* on equipment, which is operating. If you must make such tests, always disconnect the equipment form its power source first.

Parts list

Resistor

R1	1000 Ω (1 kΩ), 0.25 W carbon, 10% or better

Capacitors

C1	10 μF electrolytic, 16 V
C2	0.1 μF subminiature polyester or ceramic

Integrated circuits

IC1	LM3909N

Additional items

	Veroboard 7 strips by 15 holes
LS1	64 Ω miniature loudspeaker
	IC holder
	Veropins
	Connecting wire
	Test leads
	1.5 V AA alkaline battery

76 An experimental 70 cm rhombic aerial

Introduction

Most commonly used aerials can be classed as *resonant* or *standing-wave* aerials. There is another class known as *non-resonant* or *travelling-wave*. Resonant aerials, such as the dipole, are narrow-band; this occurs because resonance occurs only over a narrow band of frequencies. Travelling-wave aerials, on the other hand, can operate over a wide band of frequencies.

The *rhombic* is an example of a non-resonant or travelling-wave aerial. It is often employed for fixed commercial and military short-wave radio links. Made with wire, it has a diamond shape when looking down on it from above. The four corners are supported on four masts. It is a very effective aerial, and has good gain, a quality which can be judged from the *polar diagram* shown in **Figure 1**. This has been obtained from a computer program, and so is the perfect shape for a rhombic aerial.

Theory (but only a little)

A polar diagram shows graphically the ability of an aerial to radiate (or receive) more effectively in one direction at the expensive of the radiation in other directions. Figure 1 shows the polar diagrams of two aerials, a simple 70 cm dipole and the rhombic described in this project. The dipole has the

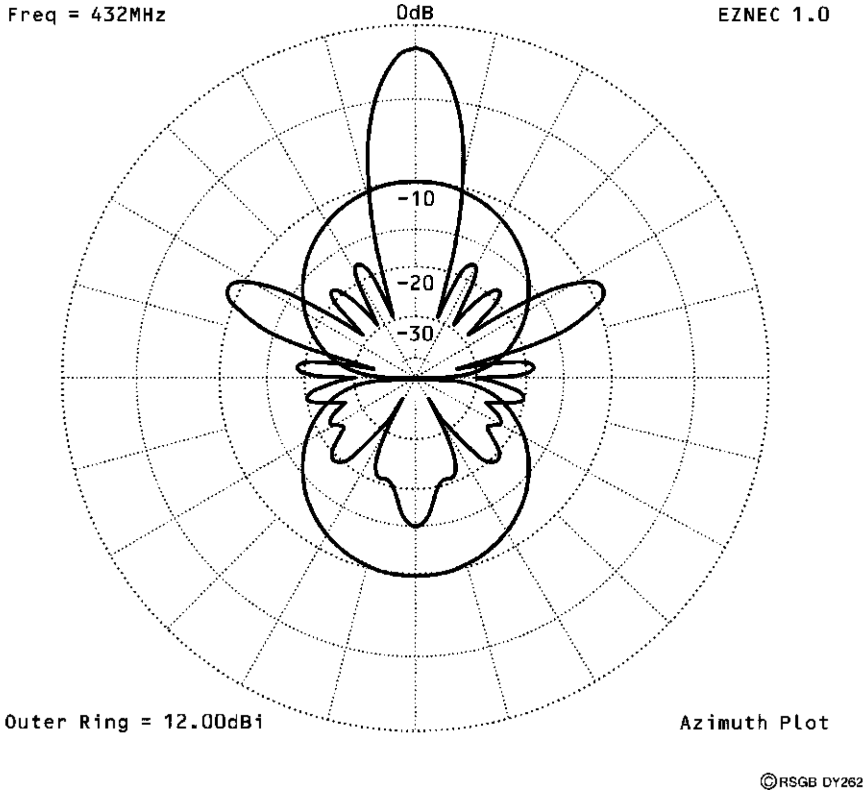

Figure 1 The horizontal polar diagrams of the small rhombic antenna and a dipole compared

well-known figure-of-eight shape, showing a symmetrical angular behaviour about the direction of the aerial and about a direction perpendicular to the aerial. That of the rhombic, on the other hand, is quite irregular by comparison, but is still symmetrical about one axis only, not two. It's this asymmetry that gives the rhombic its gain, by virtue of its *front-to-back ratio*. This is the ratio of the power radiated forwards to that being radiated backwards. Notice the large *lobe* (lump) at the top of the polar diagram; this is the direction in which the aerial transmits best. The lobe in the opposite direction has been reduced significantly, allowing more power to be directed forwards, not backwards

Problems

The rhombic is not found in every amateur's back garden, despite its attractions. To work best each edge of the diamond shape should be about two wavelengths long. Hence, a rhombic for the 20 metre band could be about 80 m from tip to tip! Another disadvantage is that, because of its size, it cannot be rotated.

It will operate well over a range of frequencies. An HF rhombic could work on the 7, 10, 14, 18, 21 and 28 MHz bands. One for the lower VHF frequencies could operate on the 50, 70 and 144 MHz bands.

Its large size is less of a problem at UHF. A portable rhombic can be made for 70 cm which can be used with a hand-held transceiver. The aerial design to be described here will give a gain of up to 9 dB relative to a dipole. This is written as 9 dBd, the second 'd' meaning 'relative to a dipole'. This is equivalent to improving your signal by 1.5 S-points at the receiver or (and you may be surprised by this) by fitting a linear amplifier to your transceiver that would take 5 W input and produce 40 W output to a normal dipole! Consider the relative costs of the two approaches to producing the same received signal. You will also receive everyone else's signals 1.5 S-points better than before!

Unlike the popular Yagi aerial, this design has no critical dimensions and can be folded up for transport by car or bicycle. It has a fairly high input (or feed) impedance, being fed usually by balanced twin feeder. To feed it with standard 50 Ω coaxial cable, a matching transformer in the form of a *balun* is required. A balun will convert the *bal*anced (symmetrical) aerial impedance to the *un*balanced and lower impedance of the coaxial cable. The details of how to build the balun, which in this case is a half-wave transformer, are included in the constructional details.

Construction

The aerial frame is made up of 1 cm × 2 cm strips of wood fixed to a plywood centre using 30 mm long M4 bolts, as shown in **Figure 2a**. The outer bolts fixing the front and side supports can be removed for folding prior to transportation, as shown in **Figure 2b**.

The wires are fixed to the front and rear supports using screw connectors, sometimes called 'chocolate block' connectors. Detail X and Y of Figure 2 show how this is done. The side supports have holes in the ends through which the wire is threaded.

The aerial must be mounted in the horizontal plane using a small shelf bracket attached to the centre plate, using the same bolts that hold the rear support. The other half of the bracket may be mounted to a vertical mast using screws or jubilee clips (hose clips).

The balun

Cut a 23 cm length of coaxial cable; this is to be our half-wave transformer. Cut and remove 2 cm from each end of the sheath. Make the braid into a pigtail at each end. Cut and remove 1 cm of the inner insulator from each

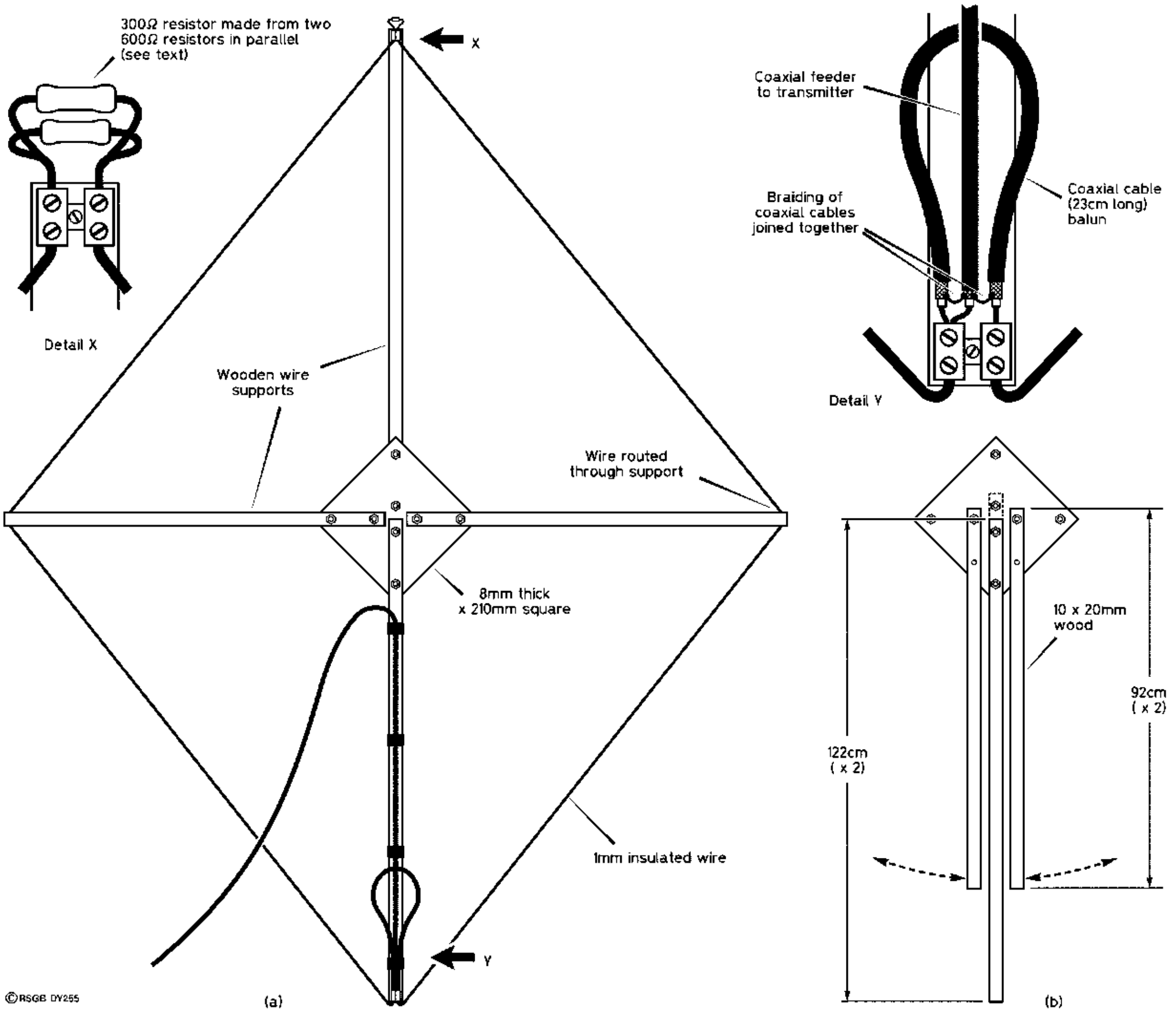
300Ω resistor made from two 600Ω resistors in parallel (see text)
X
Detail X
Wooden wire supports
Wire routed through support
8mm thick x 210mm square
1mm insulated wire
Y
©RSGB DY255
(a)
Coaxial feeder to transmitter
Coaxial cable (23cm long) balun
Braiding of coaxial cables joined together
Detail Y
10 x 20mm wood
92cm (x 2)
122cm (x 2)
(b)

Figure 2 Construction of the UHF rhombic antenna

end, leaving 1 cm of the centre conductor exposed. Fasten this piece of cable to the chocolate block connector together with the coaxial feeder cable as shown in Detail Y of Figure 2. The braids of all three prepared ends are soldered together, but are not connected to anything else. The inner conductor of the feeder from the transceiver is soldered to one end of the centre conductor of the 23 cm piece, and is connected to one end of the rhombic by one side of the chocolate block. The other end of the half-wave transformer also goes to the chocolate block, where it is connected to the other side of the rhombic loop.

The opposite corner of the rhombic, as shown in Detail X of Figure 2, shows that the loop is broken at the chocolate block, and is 'terminated' by two 600 Ω resistors in parallel. This makes the aerial a broad-band travelling-wave device, and gives it its directivity and gain. There is a rule of thumb, which includes a safety factor, which says that the terminating resistor must be able to absorb one-half of the maximum transmitter output power. So, if you use two 2 W 600 Ω carbon resistors (*not* wire-wound resistors) in parallel, you can use a transmitter with an output of 8 W, which is more than adequate for 70 cm hand-helds.

Using the rhombic

Fit the aerial to a pole or mast in the horizontal plane, if you intend to use CW or SSB, but in the vertical plane if you want to concentrate on FM work. In the latter case, a wood or fibreglass pole is mandatory. Tune to a local repeater, whose signal strength you know. Rotate the aerial to face the repeater, and you should see that the signal strength is greatly improved! Verify the directional properties of the aerial by rotating it and observing the changes in signal strength on your S-meter.

77 Water level alarm

Introduction

This is a simple device which, when you have built it, you begin to wonder how you ever managed without it! It will sound a buzzer when the level of water reaches a particular point, which you can set and vary at will.

Detecting water

Water is not too difficult to detect electrically, because it is a conductor of electricity. Not a good one, but sufficiently so to carry enough current to control a warning circuit.

The circuit is shown in **Figure 1**. If you imagine a resistor placed between the ends marked 'water detector', the circuit is then recognisable as a transistor used as an electronic switch. Whether the switch is off or on depends on the value of this resistor. If the resistance is low, the transistor switches on and the buzzer sounds. If the resistance is high, the transistor switches off and the buzzer stops.

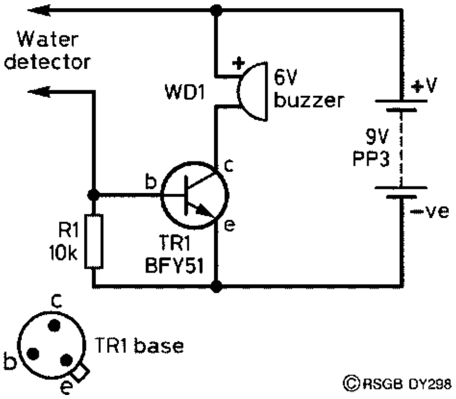
Water detector
WD1
6V buzzer
+
+V
9V PP3
-ve
b
c
e
TR1 BFY51
R1 10k
TR1 base
©RSGB DY298

Figure 1 Circuit diagram of the water level detector

Construction

The unit is made in two separate sections – the detector board (**Figure 2**) and the main board (**Figure 3**). The main board is plain perforated board (without copper strips). The components are pushed through holes in the main board and connected together on the underside. The battery is held on with sticky tape. No on/off switch is used, as the circuit consumes virtually no current when it is not sounding the buzzer.

Unlike some buzzers, this type is polarised, and the red lead must go to the supply as shown in Figs 1 and 3.

The detector board is a piece of standard matrix board – the type with copper strips along one side. It measures 9 strips by 15 holes, and is wired in such a way that alternate strips are connected together, as Figure 2 shows. This gives a greater surface area for the water to touch, thus increasing the sensitivity of detection.

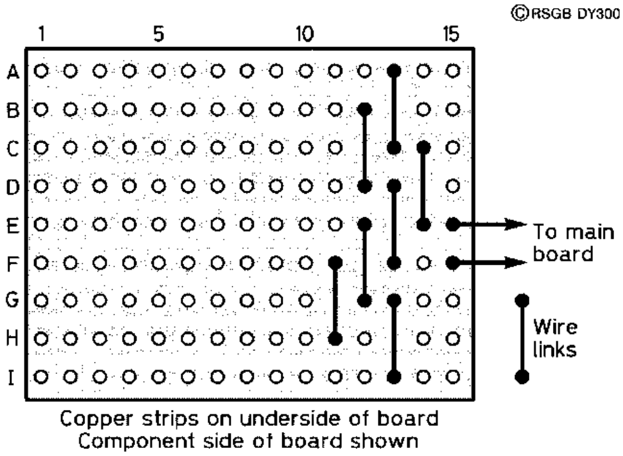
©RSGB DY300
1
5
10
15
A
B
C
D
E
F
G
H
I
To main board
Wire links
Copper strips on underside of board
Component side of board shown

Figure 2 Layout of the main board

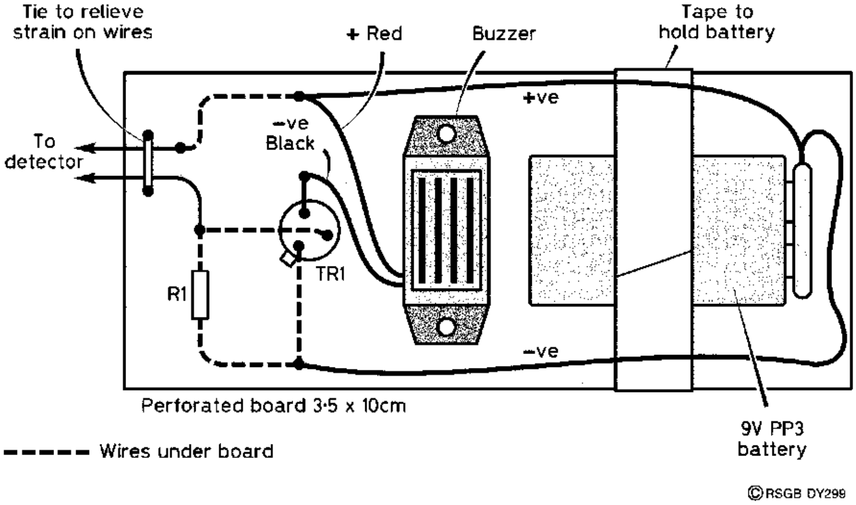
Tie to relieve strain on wires
+ Red
Buzzer
Tape to hold battery
To detector
−ve Black
+ve
TR1
R1
−ve
Perforated board 3·5 x 10cm
9V PP3 battery
Wires under board
©RSGB DY298

Figure 3 Water detector constructed from stripboard

The two boards are connected by ordinary twisted pair, enabling the detector to hang over the edge of the water container, with the buzzer circuit safely out of the way. When positioning the detector board, the left-hand edge (as shown in Figure 2) should face downwards, with the connecting-wire edge facing upwards. When the water level rises, the water causes current to flow between adjacent strips, and the higher the water, the lower the resistance between them. Eventually, the resistance will be low enough to turn on the transistor and sound the alarm.

Parts list

Resistor

R1	10 000 Ω (10 kΩ), 0.25 W

Semiconductor

TR1	BFY51

Additional items

WD1	6 V miniature buzzer (polarised)
	Battery PP3
	Battery connector
	Strip board 9 strips by 15 holes
	Perforated board 3.5 × 10 cm

78 A delta loop for 20 metres

Introduction

HF aerials take up a lot of room when they are straight. Space can be saved by bending them and, at the same time, giving them properties which are quite different from their original linear forms. The delta loops shown in the accompanying diagrams have the shape of the Greek upper-case letter delta (Δ) upside down. It can be upright, except that it is more practical to have the feed point at the bottom than at the top.

Putting the loop together

This is the design for a loop to be used on the 20 metre band. It is a very popular band and carries most of the amateur radio DX traffic. The aerial is light in weight and operates best if the top section is at least 30 ft above the ground, with the feeder point being about 6 ft above ground. All you need is some good wire, some polypropylene rope, some insulators and a method of matching the output of this 75 Ω balanced aerial to the 50 Ω unbalanced coaxial input of your transceiver or receiver.

Each of the three sides must be 7.2 m long. Using tent pegs or six-inch nails, mark out the three corners on the lawn, making sure that each pair of pegs or nails is exactly 7.2 m apart. Put the wire around the nails, together with two insulators on the wire for the two top corners. These insulators can be of the 'dog bone' variety, or of the home-made type, using a piece of flat plastic with holes in each end. The use of these is shown in **Figure 1**. Located at the positions of two of the tent pegs or nails, they must be secured at those points by any convenient means.

Now attach a support rope to the free end of each of the insulators, ready for suspension, as shown in the diagrams, from any convenient supports (house, tower, tree, etc.). The top section of the loop could be taped to a continuous piece of rope between the two supports. You might like to try this if you think it is easier. Then, when you find some dog-bone insulators, you can change the design and see if there is any noticeable change of performance.

Connecting to the radio

If you are the proud possessor of an aerial tuning unit (ATU) with a balanced output, all you will need is a length of twin cable soldered to the

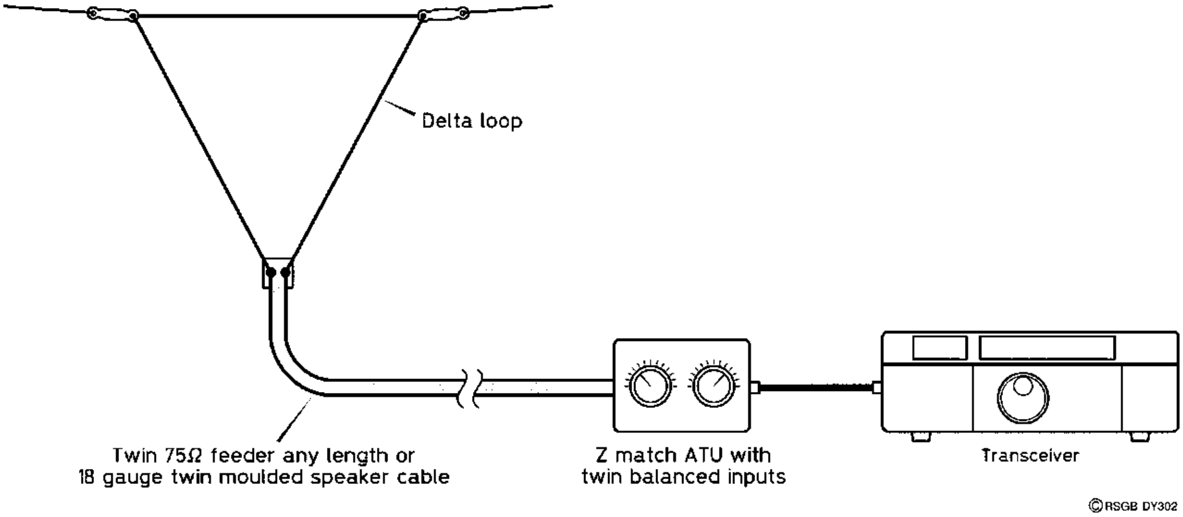
Delta loop
Twin 75Ω feeder any length or
18 gauge twin moulded speaker cable
Z match ATU with
twin balanced inputs
Transceiver
©RSGB DY302

Figure 1 Multiband version of the delta loop connected to the ASTU using 75 Ω twin feeder

ends of your delta loop and connected to the balanced output terminals on the ATU, as Figure 1 shows.

Another way is to buy a ferrite-cored 1:1 balun, and use it as indicated in **Figure 2**. This produces an aerial which will operate over the whole range of amateur and commercial short-wave frequencies, when used with an ATU.

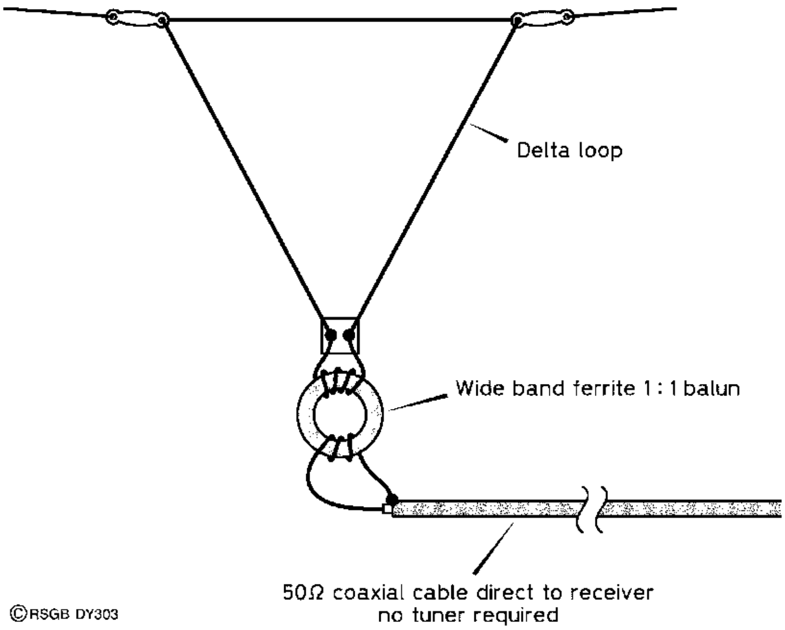
Delta loop
Wide band ferrite 1 : 1 balun
50Ω coaxial cable direct to receiver
no tuner required
©RSGB DY303

Figure 2 Multiband version of the delta loop using 75 Ω coaxial cable. This arrangement requires an ASTU if used with a multiband antenna

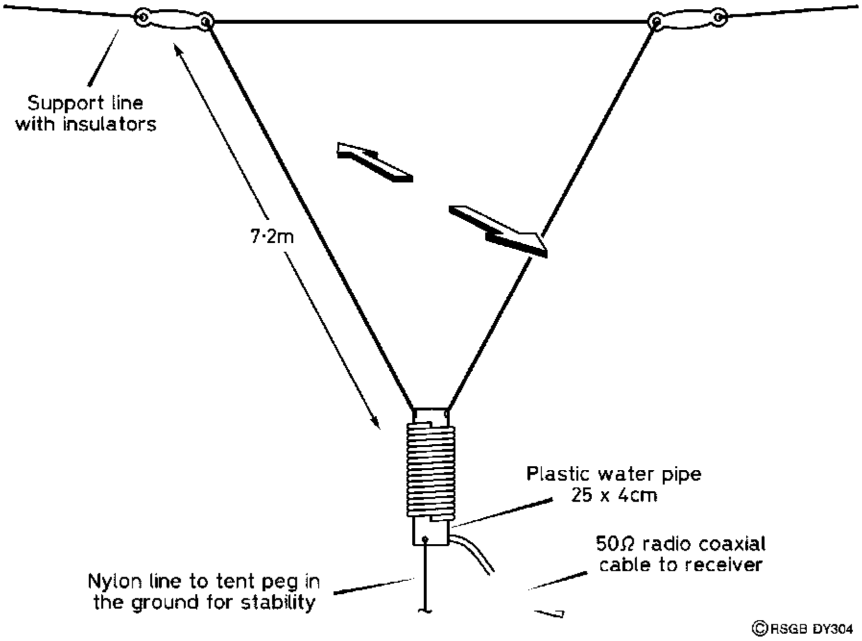
Support line with insulators
7·2m
Plastic water pipe 25 x 4cm
50Ω radio coaxial cable to receiver
Nylon line to tent peg in the ground for stability
©RSGB DY304

Figure 3 Single band version of the loop antenna using a coaxial balun. The direction of maximum signal strength is indicated by the arrows

However, since this aerial is designed to be a single-band type, there is a simpler answer to the problem of matching the aerial to your 50 Ω coaxial cable – **Figure 3** shows this. Make a tuned balun with a length of plastic water pipe, 25 cm long and 4 cm diameter and a piece of good-quality TV aerial cable. TV aerial cable has a characteristic impedance of 75 Ω, compared with the 50 Ω of the coaxial feeder that we usually use with amateur radio equipment. Use the type with a brown sheath and a closely knit earth braid, not the type having an earth foil inside the sheath.

The length we need for the balun is 3.8 m, but always allow between 3 and 4 cm extra for preparing the ends. Drill two small holes diametrically opposed in the top of the tube; these will be used to anchor the two ends of the delta loop, as shown in Figure 3. Drill another single hole in the bottom end of the tube, which will be used to anchor a nylon line going to the ground to add stability to the loop. Then, after drilling a pilot hole, drill a 5 or 6 mm hole near the top end of the tube, as shown in **Figure 4**. Prepare both ends of the coaxial cable, then feed one end into the tube, far enough for its ends to be soldered to the ends of the loop when the assembly reaches that stage.

Now, close-wrap the cable around the tube until only about 3 cm remain. Holding the cable tightly, drill another 6 mm hole beside the free end of the cable and feed it into this hole. At this point, feed the two ends of the loop into the top two holes and twist it back on itself. Figure 4 shows how this is done. Then, solder the ends of the coaxial balun to the ends of the loop.

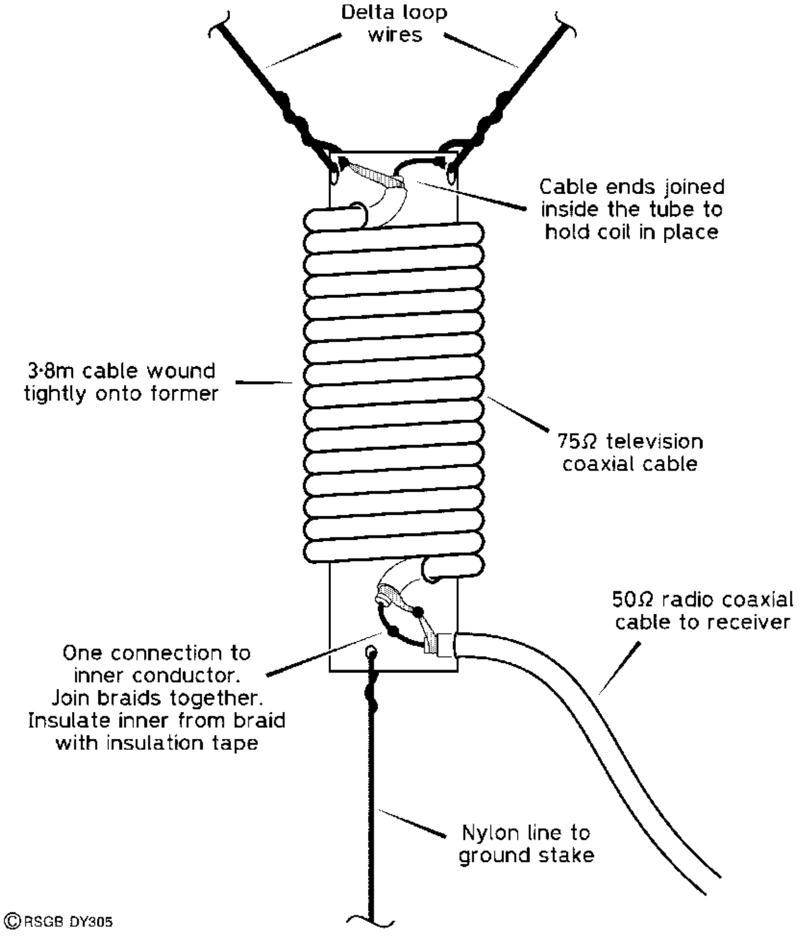
Delta loop wires
Cable ends joined inside the tube to hold coil in place
3·8m cable wound tightly onto former
75Ω television coaxial cable
50Ω radio coaxial cable to receiver
One connection to inner conductor. Join braids together. Insulate inner from braid with insulation tape
Nylon line to ground stake
©RSGB DY305

Figure 4 Construction of the balun

The bottom ends of the balun are then soldered to the 50 Ω coax which goes to your shack and to the transceiver.

Hoist the aerial into position carefully, being careful not to pull too hard on the support lines. Then, take the nylon line from the bottom of the balun to a peg in the ground. This adds stability to the aerial.

Using the delta loop

It is a directional aerial, as Figure 3 shows. It produces maximum power (and has maximum receive sensitivity) along a direction perpendicular to its own plane. Don't be too concerned with which direction it is pointing at first. Give it a try 'on the air' and see how it performs. Then you can contemplate how to point it in your favourite direction, to the USA, or Australia, for example.

Experimenting

You may want to enclose the balun in some sort of weatherproof container. Plastic ice cream containers are favourites for this sort of job. Seal all the holes where wires enter it with silicone sealant or self-amalgamating tape.

If you have used ordinary single-strand or multi-strand wire for your aerial, it will stretch over time under its own weight and that of the balun. Its operating frequency will fall slightly as a result. If you notice a significant difference, then dismantle it, remeasure and fix the sides and erect it again, perhaps facing a different direction. You can buy pre-stretched or hard-drawn wire for such purposes, if you feel that periodic tweaking of your aerial is a chore.

You can make a delta loop for different frequencies simply by scaling the lengths of wire for the loop and for the balun according to the design frequency. If your maths isn't quite up to this, enlist the help of someone well versed either in maths or aerial design!

79 A simple desk microphone

Introduction

A hand-held transceiver is usually the first type used by most novices. It is ideal when used out of doors, in the shack or mobile. However, using one in the shack with an outside aerial is not the best of solutions.

This article describes the design of a desk microphone which can be used either for transceivers using a common microphone and push-to-talk (PTT) line, or for those using a separate PTT line.

It is a simple design, using the barrels of two plastic pens, the lid of an old coffee jar, and the plastic sheath of a DIN plug.

Construction

The end caps and the ink tubes should be removed from two ball-point pens, leaving only the two plastic shells. One of them is used for the vertical microphone support, while the other is used to hold the microphone element.

The vertical support should have the pointed end shaped so that it holds the microphone support at a convenient angle, as **Figure 1** shows.

The base should be prepared next – see **Figure 2**. Four holes are required. Two, A and B, are small holes for the cable to be fed through. A third, C, is slightly larger and is used to take one of the plastic end caps. The fourth hole, D, is determined by the type of switch that you want to use for the PTT function.

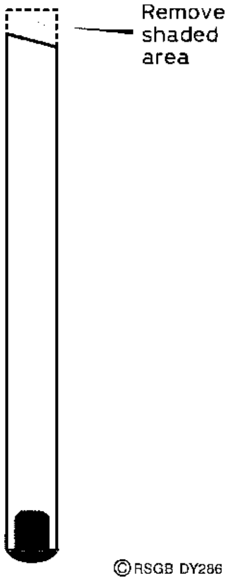
Remove
shaded
area
©RSGB DY286

Figure 1 Vertical support made from ball-point pen shell

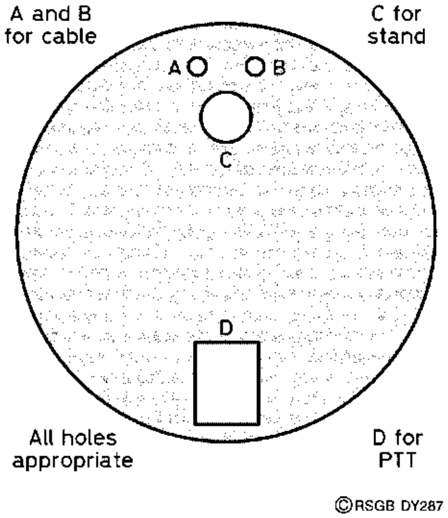
A and B
for cable
C for
stand
A
B
C
D
All holes
appropriate
D for
PTT
©RSGB DY287

Figure 2 The microphone stand base

Some epoxy resin such as Araldite Rapid is used to mount the separate parts. Mix only a small amount at a time, and glue each part of the base assembly and allow to harden before moving on. Be sure to read and follow the instructions for the use of the glue to prevent accidents.

Firstly, glue the end cap to the base, feeding it underneath and pushing it up through hole C. Then, the vertical pen body is placed over the protruding end cap and glued into place, making sure that the top (shaped) end is pointing in the right direction to hold the other plastic tube carrying the microphone. While the glue hardens, place the DIN plug sheath on the second pen body. Place and glue this pen body on the shaped top of the vertical support, making sure that the whole assembly is supported while the glue sets. What you should have now ought to resemble **Figure 3**, but without the microphone and wiring. Leave everything to set for 24 hours.

Any switch of the press-to-make variety can be used for your PTT switch, but choose one which does not apply too much force to operate! If you find the whole assembly is too light and slides over your table, some plasticine can be pushed into the base when all the wiring has been completed.

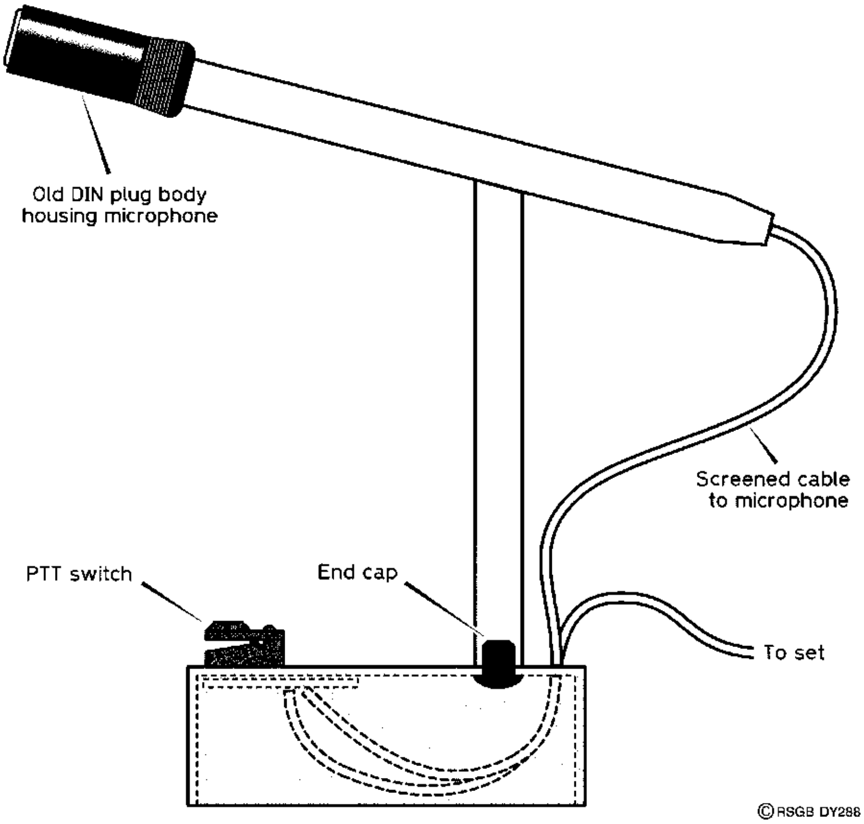

Figure 3 Complete microphone stand assembly

The circuit

A small electret condenser microphone is used, powered by a DC supply. Most hand-held radios use a single screened lead for both the PTT line and the microphone audio lead. In such cases, the circuit of **Figure 4** will operate the PTT function. Note that the PTT switch is in series with the audio lead from the microphone. If you find that the PTT switch does not switch your radio into transmit, reduce R1 from its normal value of 33 kΩ to 27 kΩ. You should find that the original value works with most hand-helds.

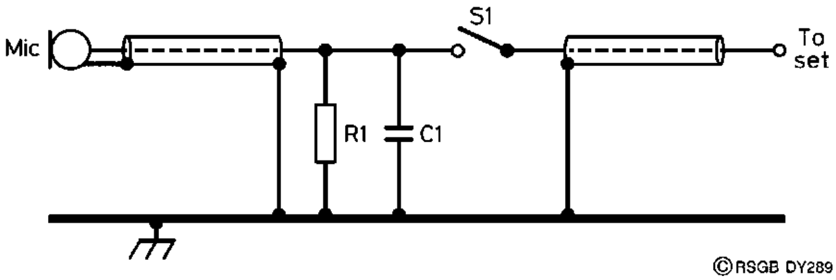

©RSGB DY289

Figure 4 Wiring diagram for hand-held radios

For radios with a separate PTT lead, the circuit of **Figure 5** is used. However, a power supply is needed for the electret microphone. This can be a PP3 battery, or can (in most cases) be derived from the microphone socket on the transceiver. You will need to consult the makers' handbook for this information. Figure 5 shows the PTT switch wired in a 'ground-to-transmit' configuration, which is correct for most base-station transceivers. If you're in any doubt about what your radio needs, consult an experienced friend.

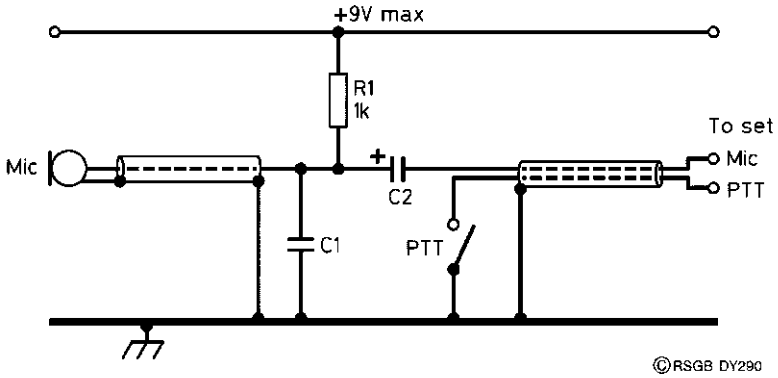

©RSGB DY290

Figure 5 Wiring diagram for radios which use separate PTT lines

The last connections

The microphone element should be connected with screened cable. Tape (masking or insulating) should be wound around the electret insert until it is a snug fit inside the DIN plug sheath. Feed the screened cable through the barrel of the pen until it emerges from the far end. Poke it through hole B and make the connections under the coffee jar lid. The components can all be mounted on the PTT switch or on a small piece of Veroboard mounted under the top surface. The output lead emerges through hole A and is of sufficient length to reach your transceiver. A suitable plug needs to be fitted to it.

Parts list

Resistors: all 0.25 W carbon, 5% tolerance

R1 (Figure 4)	33 000 Ω (33 kΩ)
R1 (Figure 5)	1000 Ω (1 kΩ)

Capacitors

C1 (Figure 4 and 5)	0.1 microfarads (0.1 μF) disc
C2	10 microfarads (10 μF) electrolytic

Additional items

S1	Push-to-make
Mic	Electret microphone (Maplin type FS34W)
Screened cable	As required
Plug	As required by your radio
Veroboard	If required
PP3 battery and clip	If required

Two old plastic ball-point pens
Lid of coffee jar
Plastic sheath from DIN plug
Araldite or similar epoxy resin glue

80 Morse oscillator

Introduction

This is not the simplest Morse oscillator to build, but it differs from the simple circuits in that it produces a pure note, not a coarse, rasping sound. People who have practised Morse using a non-sinusoidal oscillator sometimes find that they have trouble copying Morse code with a pure tone. As the pure tone is the correct way to receive Morse code, it is important that you should learn to listen to the code from a pure oscillator – so here's one!

The twin-T

There are many oscillator circuits, and there are many variations of the twin-T oscillator that we are going to use. **Figure 1** shows one version of a very useful circuit. All oscillators must have positive feedback in order to work. The feedback determines the frequency of the note produced by the oscillator.

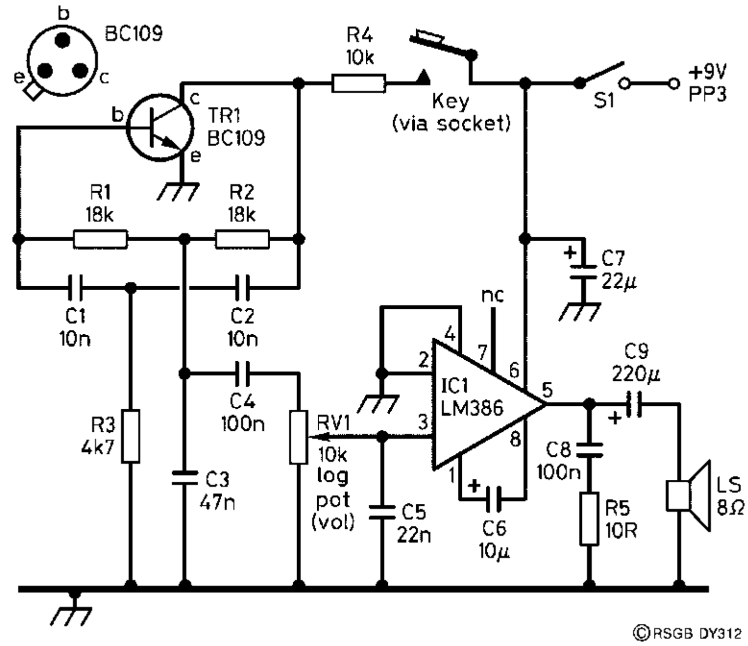
b
e
c
BC109
R4
10k
Key
(via socket)
S1
+9V
PP3
c
b
TR1
BC109
e
R1
18k
R2
18k
C1
10n
C2
10n
+
C7
22μ
nc
4
7
2
6
IC1
LM386
5
C9
220μ
+
C4
100n
R3
4k7
RV1
10k
log
pot
(vol)
3
8
C8
100n
1
+
C3
47n
C5
22n
C6
10μ
R5
10R
LS
8Ω

Figure 1 Circuit diagram of the sinewave oscillator and amplifier

©RSGB DY312

Here, the feedback circuit looks like two letter Ts. If you look at Figure 1, one T is formed by R1, R2 and C3, the other by C1, C2 and R3 – hence the name 'Twin-T'. Notice that the two Ts are connected in parallel between the collector and base of TR1, so any signal appearing at the collector is fed into both Ts. What emerges is then fed back into the base, producing in turn a signal at the collector. And so it goes on, producing a sine wave output.

The oscillator output is fed into an integrated circuit amplifier for output via a small loudspeaker.

Putting it together

The prototype was constructed on plain matrix board (the type *without* the copper strips), as shown in **Figure 2**. The components have their leads pushed through the holes in the board, and connections are made underneath.

Build the amplifier circuit first, using a socket for IC1. Connect the 9 V supply and touch pin 3 with an ordinary piece of wire. If a buzz is heard from the speaker, all should be well. If not, check your circuit and make changes until it does.

Build the oscillator circuit, and connect its output to the volume control VR1 via C4. Set VR1 half-way along its travel and switch on. A note should be heard from the speaker when the Morse key is depressed. The component values making up the twin-T determine the frequency of the note. Try varying them if you think your note is too high or too low. Whatever changes you make, either to the resistors or the capacitors, always ensure that R1 = R2 and that C1 = C2.

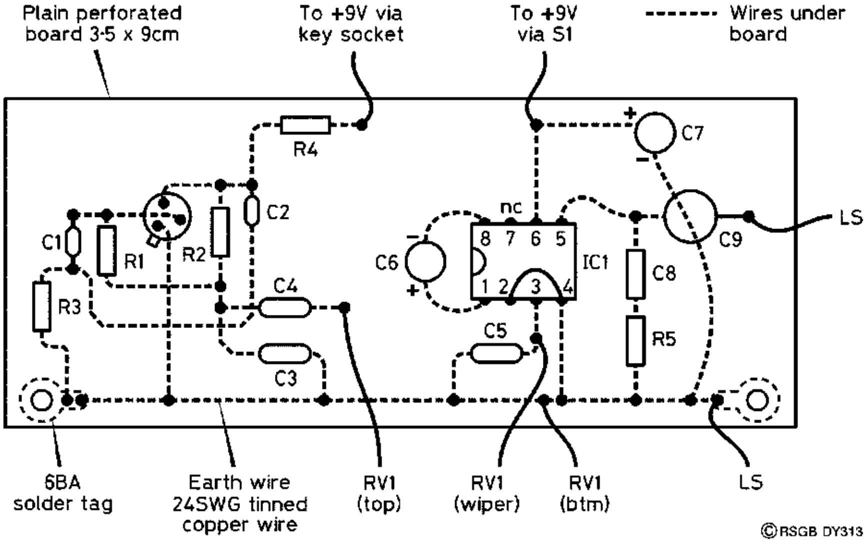
Plain perforated board 3·5 x 9cm
To +9V via key socket
To +9V via S1
------- Wires under board
R4
C7
C2
C1
R1
R2
C4
R3
C3
nc
8 7 6 5
1 2 3 4
IC1
C6
C5
C8
R5
C9
LS
6BA solder tag
Earth wire 24SWG tinned copper wire
RV1 (top)
RV1 (wiper)
RV1 (btm)
LS
©RSGB DY313

Figure 2 Component layout and interconnection diagram

Parts list

Resistors: all 0.25 W carbon film

R1, R2	18 000 ohms (18 kΩ)
R3	4700 ohms (4.7 kΩ)
R4	10 000 ohms (10 kΩ)
R5	10 ohms (10 Ω)

Capacitors

C1, C2	10 nF (0.01 μF)
C3, C8	47 nF (0.047 μF)
C4	100 nF (0.1 μF)
C5	22 nF (0.022 μF)
C6	10 μF electrolytic, 16 V WKG
C7	22 μF electrolytic, 16 V WKG
C9	220 μF electrolytic, 16 V WKG

Semiconductor

TR1	BC109

Integrated circuit

IC1	LM386

Additional items

VR1	10 kΩ log potentiometer with switch
LS	8 Ω loudspeaker

Matrix board 3.5 × 9 cm
Small jack socket for key input
Box
PP3 battery and clip

81 A simple 6 m beam

Introduction

The attraction of building your own aerials is an abiding feature of our hobby. You can *buy* almost any shape or size of aerial, but one you have made yourself can often work every bit as well as a commercial device costing ten times as much.

The design

This aerial, designed for use on the 6 m band, is essentially a two-element Yagi, with the elements bent in order to reduce the physical size. It is known

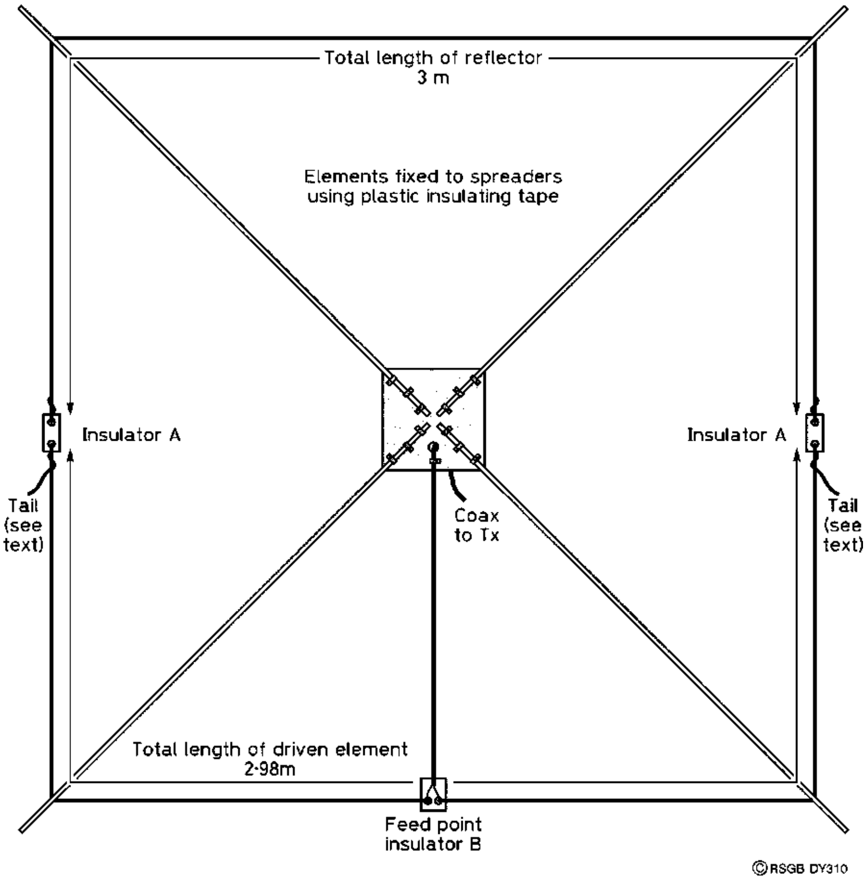
Total length of reflector
3 m
Elements fixed to spreaders
using plastic insulating tape
Insulator A
Insulator A
Tail (see text)
Tail (see text)
Coax to Tx
Total length of driven element
2·98m
Feed point
insulator B
©RSGB DY310

Figure 1 Plan view of the complete 50 MHz VK2ABQ antenna

as the VK2ABQ beam, and was designed originally for the 20, 15 and 10 m bands, principally because of its space-saving qualities. It is made using a wooden frame and wire elements, and is ideal for portable operation.

Tools ready?

The beam is shown in **Figure 1**. The driven element is the one whose centre is fed by the coaxial cable, and lies between the two insulators marked A and the feedpoint at B. The reflector is also anchored at the points A, and lies over the upper half of the frame.

The wooden centrepiece is used to support the cross-pieces and to mount the aerial on the mast, using a common shelf bracket. The cross-pieces, known as *spreaders*, can be wooden canes or dowelling, and are mounted to the centrepiece using cable clips and adhesive. **Figure 2** shows how this is done.

If the aerial is to be a permanent installation, the spreaders should be weatherproofed using a good-quality exterior varnish. The wire elements are PVC covered and fed through holes in the spreaders.

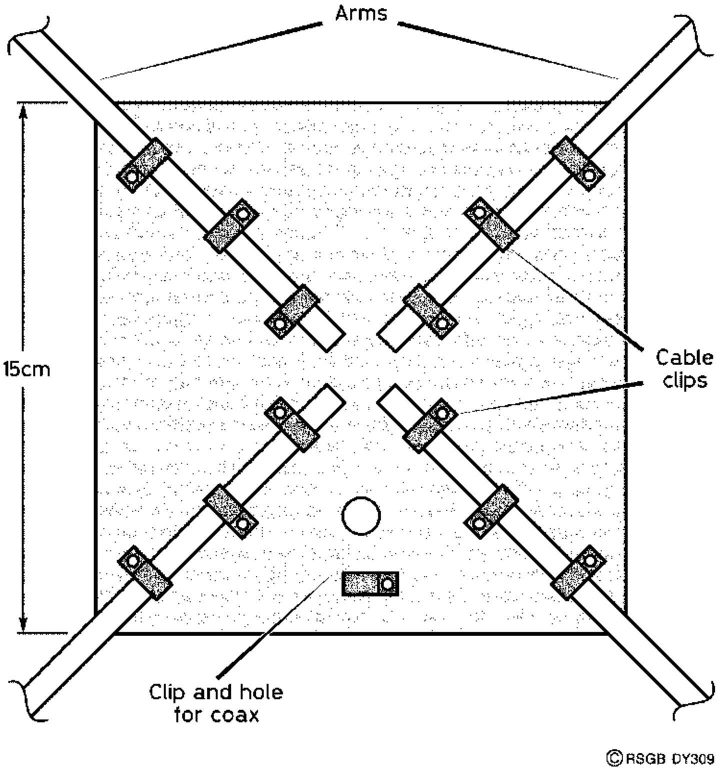

Figure 2 Centre support piece

The end insulators are made of drilled perspex, and the wire passed through the two holes and twisted, as shown in **Figure 3**.

Note: if you have not drilled holes in perspex before, take care! The drill bit must be well-lubricated because it generates a lot of heat (enough to melt the perspex and jam the drill). Never turn the drill for more than a few seconds at a time, and moisten the bit in between. Start with a pilot hole and use bits of increasing size until the hole is the size you want.

Another perspex insulator is used to secure the feeder to the aerial, as shown in Figure 3. The feeder then passes directly to the centrepiece, where it is fastened with a cable clip and then passes down the mast.

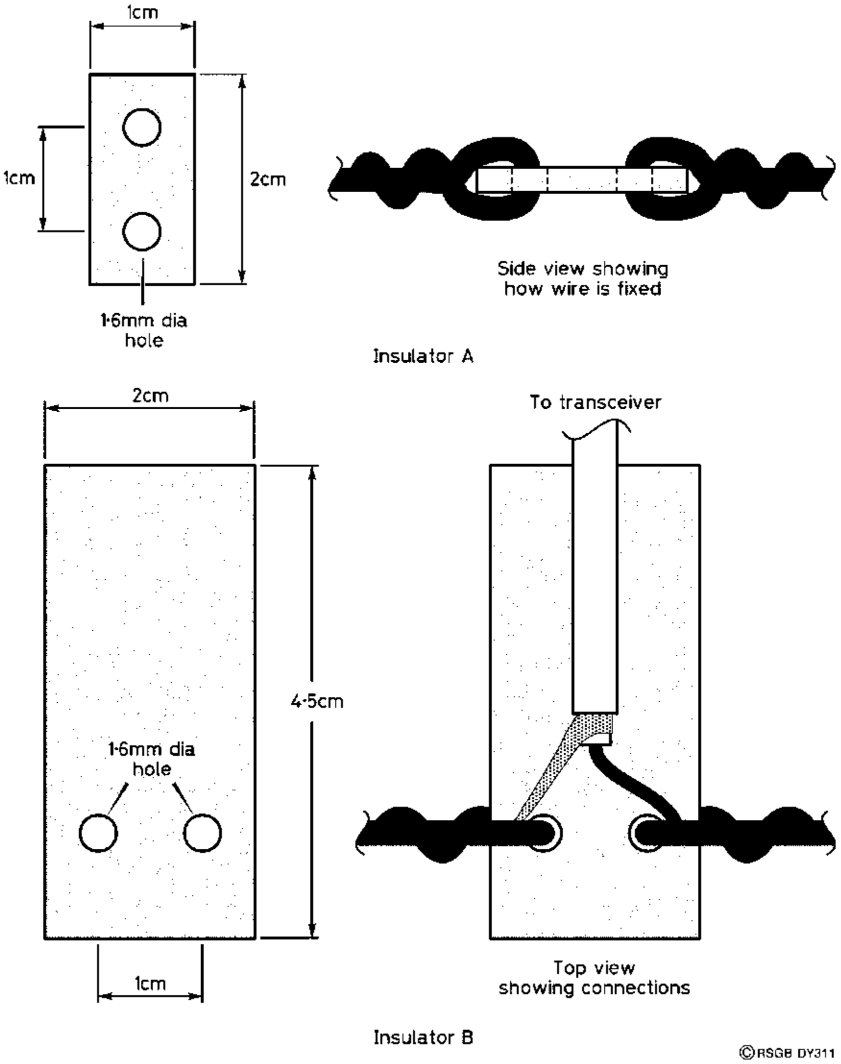
1cm
1cm
2cm
1·6mm dia hole
Side view showing how wire is fixed
Insulator A
2cm
To transceiver
4·5cm
1·6mm dia hole
1cm
Top view showing connections
Insulator B
©RSGB DY311

Figure 3 Details of insulators

Adjustment

The driven element (A–B–A) in Figure 1 is fixed to the end insulators in such a way as to have 'pigtails' which are about 10 cm long. Using an SWR meter between the aerial and the transmitter, trim the pigtails equally at each end for minimum SWR in that portion of the 6 m band in which you plan to work. Make sure that the transmitter is off when you trim the ends, as high voltages can be present there. Always listen on the frequency before you transmit and, when you do, ask if the frequency is in use and identify yourself. Use as little power as possible.

The prototype had an SWR of 1.2 at 50.2 MHz and performed well. If you look at Figure 1, which is a view of the aerial from above, you will immediately see that when the aerial is horizontal, it radiates with horizontal polarisation in a direction from the top of the page to the bottom. A metal pole or mast can be used for horizontal polarisation, but if you intend to use the aerial for vertical polarisation, it is much better to use a wooden or fibreglass pole.

Portable use

If you plan to operate portable with this aerial, the only real modifications you need are to the centrepiece and how it supports the spreaders. Instead of using glue and cable clips, nuts and bolts through the spreaders and centrepiece would allow the spreaders to be 'hinged' closed for transport.

Materials

Centrepiece	Hardwood, 15 × 15 × 25 mm
4 spreaders	110 cm long (cane or 6 mm dowel)
Wire	1.5 mm PVC-covered copper
50 Ω coaxial cable	RG58 or similar
Cable clips	6 mm (12 off)
Varnish	Polyurethane for waterproofing
Tape	Self-amalgamating, to waterproof all soldered joints
Insulators	See text

82 An integrated-circuit amplifier

Introduction

A simple audio-frequency amplifier is a very useful building-block in many more advanced electronic circuits. It is also a very useful piece of test equipment. To have one spare in the shack can be a life-saver at times.

Planning

Get into the habit of planning your project. How do you want it to look when it is finished. Do you want it in a box? Do you want it 'open-plan', with all the components on view?

The 'minimalist' approach to any project is simply a front panel and a baseboard, on to which all the components are fitted. In this case, the front panel would accommodate the loudspeaker, the volume control and the input and output jack sockets, and the baseboard would support the circuit board. Once you have decided these things, and know the size of the board, speaker and volume control, you can decided how big the panel and baseboard should be.

The amplifier

Instead of building the amplifier from discrete components (i.e. transistors), as was done in the project *An audio-frequency amplifier* (which you will find elsewhere in this book), we are going to use an integrated-circuit (IC) amplifier. External resistors and capacitors are still needed but, compared with the number of components *inside* the chip, these are very few indeed!

The IC we are going to use is the TBA820M. The circuit may look complicated, but with the use of a matrix circuit board it becomes quite simple. **Figure 1** shows the circuit diagram and **Figure 2** the layout on the board. The external connections to the PCB are shown in **Figure 3**.

To avoid having a separate switch to switch off the power supply when the amplifier is not being used, a volume control which incorporates a switch is used.

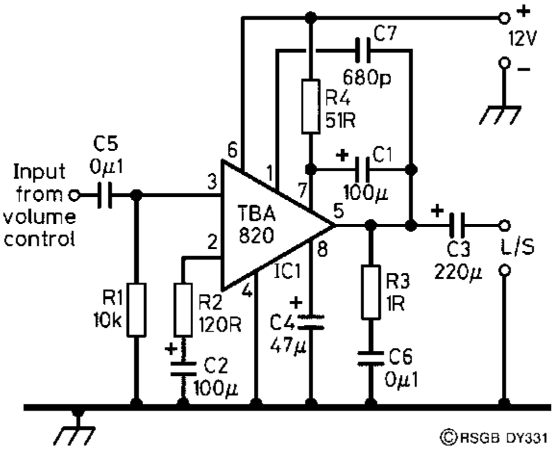
+
12V
−
C7
680p
R4
51R
C5
0µ1
Input
from
volume
control
6
1
3
7
TBA
820
5
2
8
IC1
4
C1
100µ
R1
10k
R2
120R
C2
100µ
C4
47µ
R3
1R
C6
0µ1
C3
220µ
L/S

©RSGB DY331

Figure 1 Audio amplifier, circuit diagram

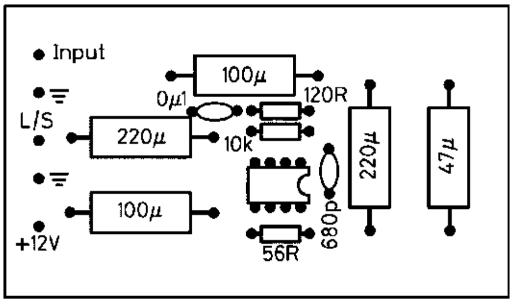
Input
L/S
+12V
100µ
0µ1
120R
220µ
10k
100µ
56R
680p
220µ
47µ

©RSGB DY332

Figure 2 Audio amplifier, component layout

©RSGB DY333

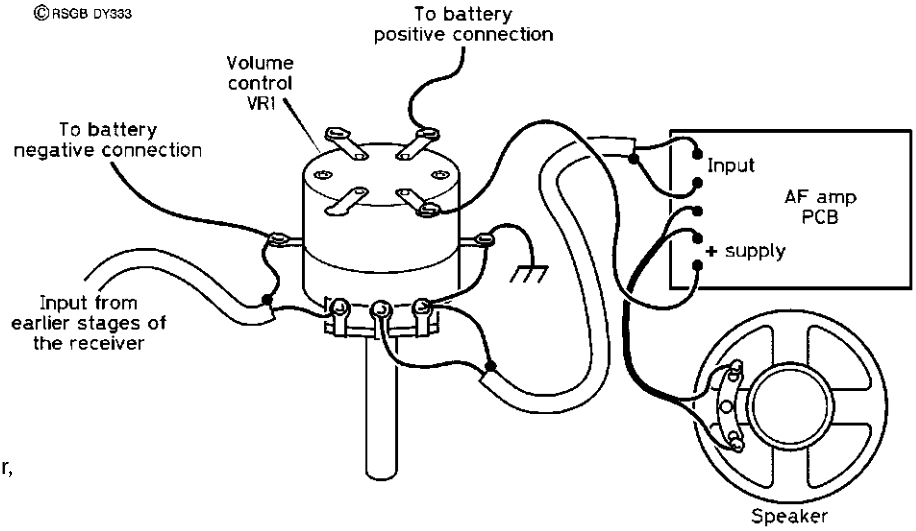
To battery
positive connection
Volume
control
VR1
To battery
negative connection
Input from
earlier stages of
the receiver
Input
AF amp
PCB
+ supply
Speaker

Figure 3 Connecting the circuit board to the speaker, volume control/switch and battery

What power supply?

If the amplifier is to be used only for short periods, as a test instrument, for example, then it can be run from a PP3 battery, which can be mounted on the baseboard. If you intend to use it often, then a connection to an external power supply is preferable. For this option, you may want to consider fixing two terminals to the baseboard for this connection.

83 A novice ATU

Introduction

Having an aerial tuning unit (ATU) is always useful. It is used for adjusting the aerial impedance, as 'seen' by the transceiver, to be the same as that of the transceiver itself, usually 50 Ω. This process is called *matching*, and ensures that the receiver and the power amplifier (PA) stage of the transmitter work efficiently.

This design is due to the late Doug DeMaw, W1FB, and uses readily available components. It will handle up to about 5 W, and operates over the frequency range 1.8 to 30 MHz.

Circuit evolution

The basic circuit of one type of ATU is shown in **Figure 1**. On transmit, the input signal at L1 is coupled into L2, which forms a resonant circuit with C1. The signal across the resonant circuit is fed to the aerial by C2. The combination of L2 and C1 helps to remove signal harmonics, because they are not at the resonant frequency and are shunted to earth. On receive, only those signals which are within the pass-band of L2 and C1 will pass into the receiver. This improves receiver performance by rejecting many out-of-band signals which the simpler receiver doesn't like.

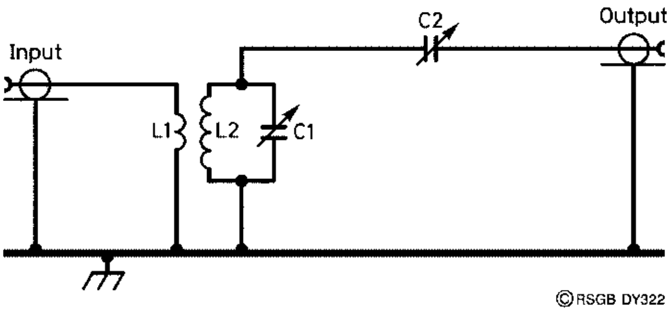

Figure 1 Basic circuit of the ATU

For this design, the all-important resonant frequency is chosen to be in the 80 metre band. If we want to make the circuit operate on other bands, we need to change either L2 or C1. It is easy to change L2 by using a set of switched inductors (L3–L5) as shown attached to S1 in **Figure 2**. Forget about S2 and L6 for the moment.

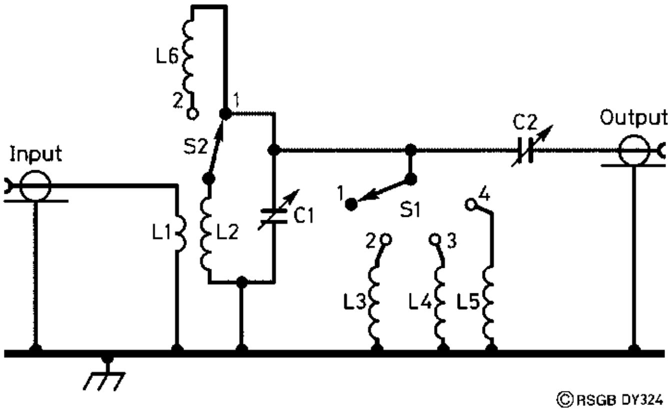

Figure 2 L6 added for 160 m

When S1 is in position 1, the circuit reverts to that of Figure 1. S2 should be brought into circuit only when S1 is in position 1. Looking at the way the circuit is drawn, you can see that when S2 is in position 2, we have two coils in series, which is equivalent to adding more turns to L2. More turns means more inductance, which lowers the resonant frequency still more, and gives coverage of the 160 m band.

Rather than having to remember to flick S2 to the correct position *and* move S1 to position 1 when we want to operate on 160 m, both functions can be combined into one rotary switch with two wafers. This circuit is shown in **Figure 3**.

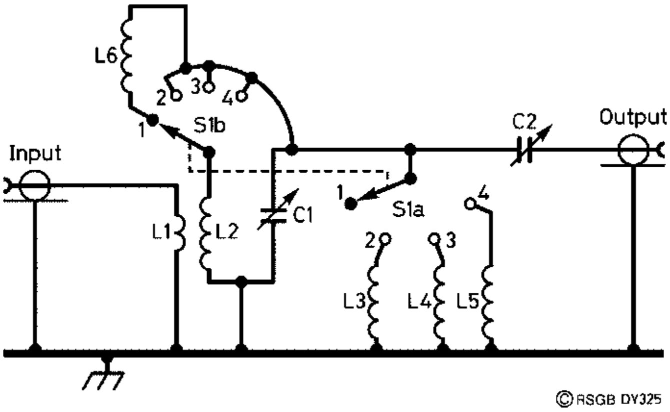
L6
2
3
4
1
S1b
Input
C2
Output
1
S1a
4
L1
L2
C1
2
3
L3
L4
L5
©RSGB DY325

Figure 3 Two-pole switch for all bands

In all positions of S1a *except* position 1, the extra coil, L6, is shorted out. The two halves of the switch, S1a and S1b, move together, as the shaft is turned, so both halves are in position 1 at the same time, position 2 at the same time, and so on. The two switches, S1 and S2, are said to be *ganged*.

Construction

L1 is formed by winding four turns of 22 SWG enamelled copper wire over L2, as in **Figure 4**. L2 already exists on the purchased former. After scraping the enamel off the ends of the wire (with a sharp knife or sandpaper), one end of L1 must be soldered to one end of L2, as shown in Figure 4. The free end of L1 goes to the transceiver aerial socket.

All components except the capacitors are assembled on the switch. Note that the rotor of C1 is earthed, but neither side of C2 is earthed. This means that the metal shaft of C2 is not earthed, and touching it will detune the ATU. Using a plastic knob for C2 will minimise this effect. If you decide to use a metal box, take precautions to ensure that no part of C2 is in electrical contact with the box.

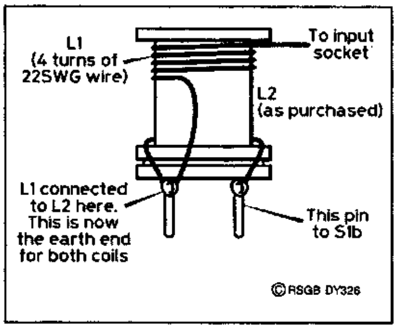
L1
(4 turns of
22SWG wire)
To input
socket
L2
(as purchased)
L1 connected
to L2 here.
This is now
the earth end
for both coils
This pin
to S1b
©RSGB DY326

Figure 4 L1 is wound over L2 as shown

The capacitors are mounted using M2.5 screws. Make sure that the screws do not foul the vanes of the capacitor. If your screws are too long, a few washers between the box and the capacitor will solve the problem!

In use

The best indication of a good match is obtained with a standing-wave meter between the ATU and the transceiver. The controls are adjusted alternately to ‘feel’ your way to a better and better match.

For receive-only use, the same alternate adjustments are used, watching for the maximum signal strength on the S-meter or, for a very weak signal, making the signal from the loudspeaker as large as possible.

Parts list

Capacitors	
C1, C2	350 pF variable
Inductors	
L1	See text
L2	27 μH
L3	10 μH
L4	2.2 μH
L5	1 μH
L6	65 μH
Additional items	
SO239 sockets (2 off)	
S1	2-pole 6-way rotary
Box	as required
Plastic knobs (2 off)	
Stick-on feet (4 off)	
M2.5 screws for capacitors	
Screws, nuts and washers for mounting the input and output sockets.	

84 CW QRP transmitter for 80 metres

Introduction

This is a relatively simple transmitter design having an output of 1 W. The design is not new, having been described before in other amateur radio publications. The components are all available new and the total cost should not exceed £15.

The circuit

Like other simple transmitters (see *An 80 Metre Crystal-Controlled CW Transmitter* and *A Breadboard 80 m CW Transmitter* elsewhere in this book) this one is crystal controlled. This assures frequency stability, but limits the usefulness of the transmitter. The key to increased frequency coverage without a conventional Variable Frequency Oscillator (VFO) is the use of a low-cost 3.58 MHz *ceramic resonator*. The 'pulling' range of a 3.58 MHz ceramic resonator covers the UK novice 80 m sub-band and some of the CW segment below 3.525 MHz.

A ceramic resonator is like a crystal, but not quite as stable in frequency. Its main advantage is its large pulling range.

The block diagram is shown in **Figure 1**. It is very similar to a crystal-controlled transmitter, and includes an oscillator, buffer and final amplifier.

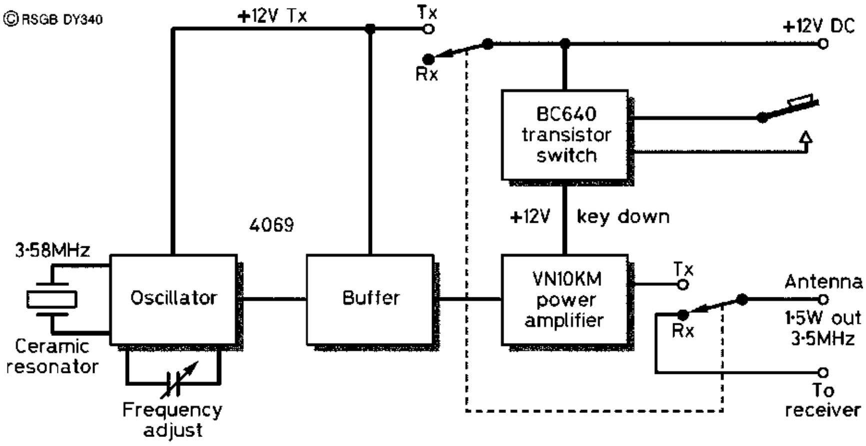
©RSGB DY340
+12V Tx
Tx
Rx
+12V DC
BC640 transistor switch
+12V
key down
4069
3·58MHz
Oscillator
Buffer
VN10KM power amplifier
Tx
Rx
Antenna
1·5W out 3·5MHz
Ceramic resonator
Frequency adjust
To receiver

Figure 1 Block diagram of the ceramic resonator controlled 80 m CW transmitter

This amplifier is keyed, the oscillator remaining running all the time. This improves frequency stability because the oscillator is not being continuously stopped and started by the key. It is switched off while receiving, though, to avoid interference with the received signal. Transmit/receive switching is accomplished by a panel-mounted switch controlling both the aerial, oscillator and buffer switching.

Figure 2 shows the transmitter circuit diagram. An unusual aspect of this transmitter is the use of a digital CMOS integrated circuit (IC) type 4069 for the buffer and oscillator stages. The IC houses six inverters, four of which are used in the circuit. One is used as the oscillator, two are used for the buffer stage, and the fourth provides an output for a direct-conversion receiver, should one be added at a later date.

The frequency of the oscillator is changed by varying the capacitance in the ceramic resonator circuit. This is provided by VC1.

The power amplifier (PA) is a small MOSFET (metal oxide semiconductor field-effect transistor), TR1. This is capable of providing an output power of 2 W but, in this circuit, it is run conservatively to give 1.5 W. The output can be varied by changing the resistance (R5 + R6) in the gate circuit. Attempts

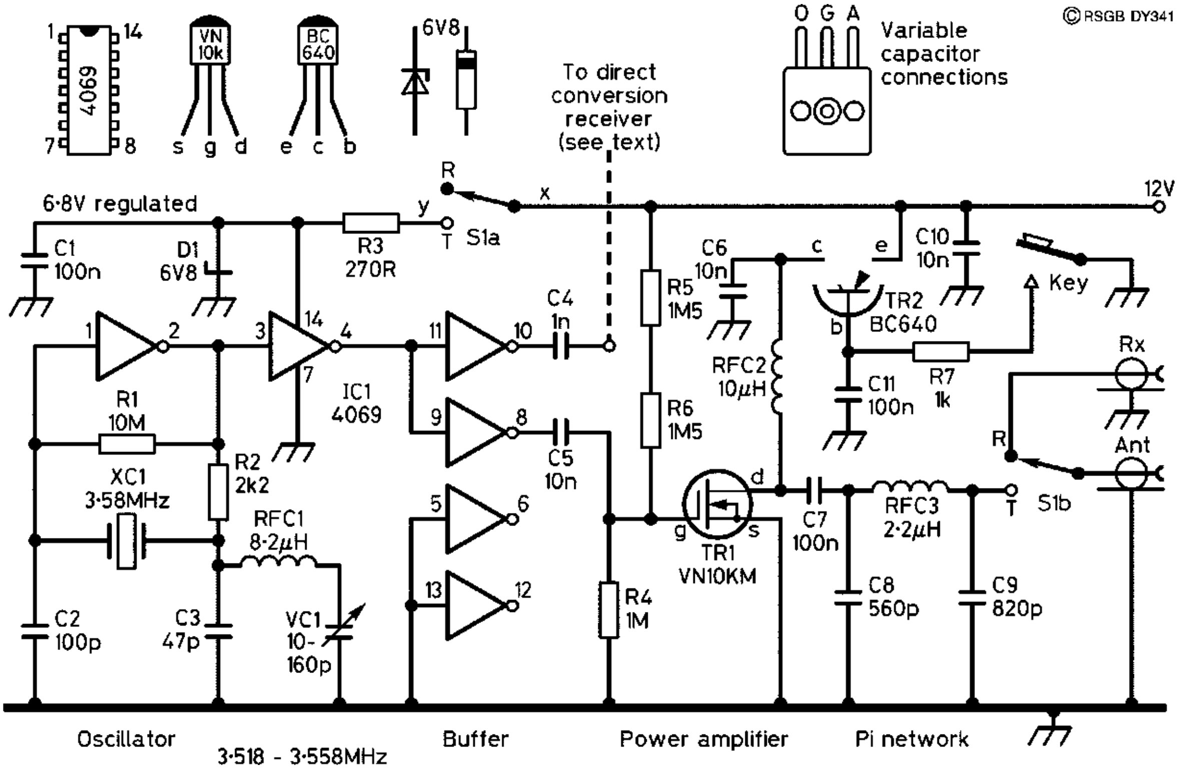
1
14
4069
7
8
VN 10k
s g d
BC 640
e c b
6V8
To direct conversion receiver (see text)
O G A
Variable capacitor connections
©RSGB DY341
R
x
12V
6·8V regulated
y
T
S1a
C1 100n
D1 6V8
R3 270R
C6 10n
c
e
C10 10n
Key
R5 1M5
TR2 BC640
b
1
2
3
14
4
7
IC1 4069
11
10
C4 1n
9
8
5
6
13
12
RFC2 10μH
C11 100n
R7 1k
Rx
R1 10M
R6 1M5
R
Ant
XC1 3·58MHz
R2 2k2
C5 10n
d
S1b
RFC1 8·2μH
g
s
TR1 VN10KM
C7 100n
RFC3 2·2μH
T
C2 100p
C3 47p
VC1 10-160p
R4 1M
C8 560p
C9 820p
Oscillator
3·518 - 3·558MHz
Buffer
Power amplifier
Pi network

Figure 2 Circuit diagram of 80 metre transmitter

to raise the output power by decreasing the values of these resistors may result in immediate MOSFET failure.

A pi-network (C8, RFC, C9) provides impedance matching to 50 Ω, together with harmonic suppression. Like all inductors in this transmitter, the pi-network inductor is a pre-wound RF choke. A pi-network is so called because the components are arranged in the shape of the Greek letter pi (π).

Keying is carried out by a pnp transistor switch, TR2. Closing the key earths the base and supplies 12 V to the collector of TR2 and to the drain of the MOSFET, TR1, allowing the PA to operate.

Construction

You must house your transmitter in a metal box, to avoid hand-capacity effects and the radiation of spurious frequencies. Size is not important, provided it is large enough to accommodate the transmitter without cramping the components. You may want to allow space for future additions such as a direct-conversion receiver, break-in keying, sidetone or a small power amplifier. A good size is 5 × 15 × 15 cm. You can make your own box, buy it, or even use a biscuit tin!

Front and rear panel connectors can be fitted first. The choice of these is a personal matter, but a good working choice would be:

(a) Power socket – 2.1 mm panel socket – centre pin positive.
(b) Key socket – ¼ inch jack socket.
(c) Aerial and receiver connectors – panel-mounting SO239 type.

Particular attention must be paid to the mounting of the variable capacitor, VC1. Make sure the hole for the shaft is amply big enough, and if you use screws to mount the capacitor on the front panel, then make sure they are not too long, otherwise they will touch the vanes of the capacitor! Mounting can be by means of glue, sparingly applied and kept well away from the shaft.

A board size of about 6 × 10 cm is adequate. Component layout on the board is suggested in **Figure 3**. The prototype used ordinary matrix board, which is preferable to stripboard for a design like this; stripboard has undue capacitance between adjacent strips. Component leads are fed through holes in the board and are connected underneath. Make sure that leads and connections are rigid because, if they can move, there is always the danger of short-circuiting, and capacitance changes.

To facilitate construction, servicing and testing, it is advisable to use Veropins for connections to the variable capacitor, transmit/receive switch,

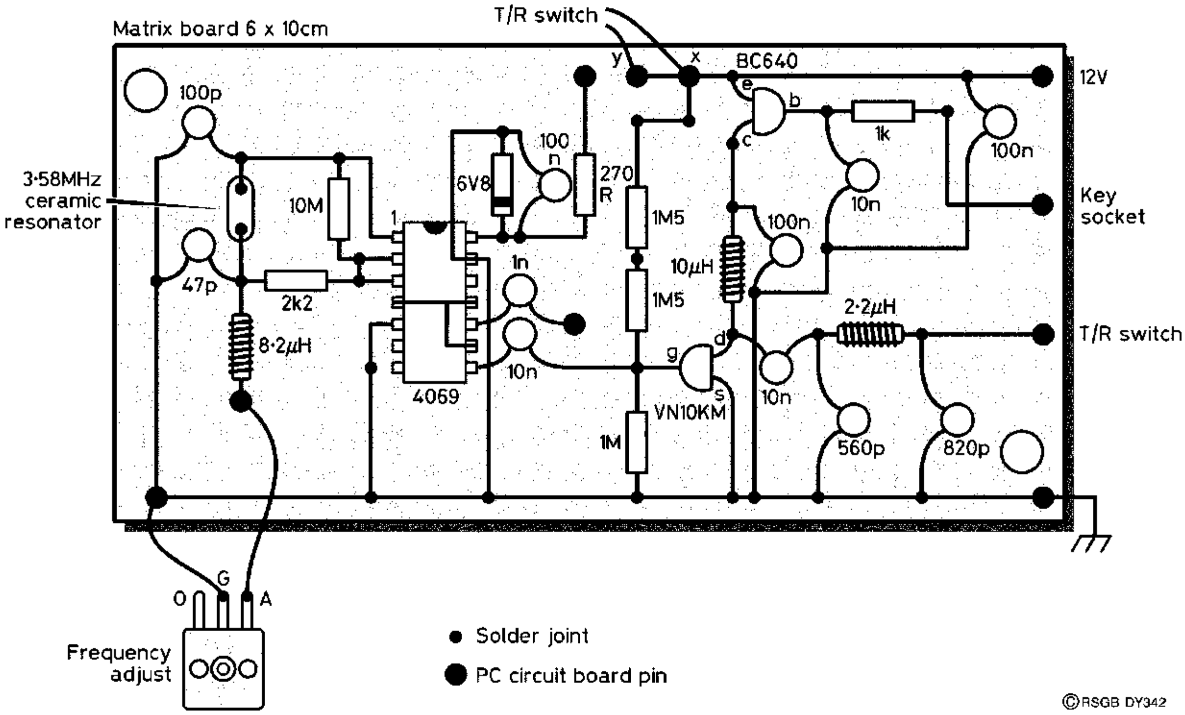
Matrix board 6 x 10cm
T/R switch
y
x
BC640
e
b
c
12V
100p
3·58MHz ceramic resonator
10M
1
6V8
100 n
270 R
1M5
1k
100n
10n
Key socket
100n
10µH
47p
2k2
1n
1M5
2·2µH
T/R switch
8·2µH
10n
g
d
s
10n
4069
VN10KM
560p
820p
1M
G
O
A
Frequency adjust
Solder joint
PC circuit board pin
©RSGB DY342

Figure 3 Component layout of the 80 metre transmitter. The transmit/receive switch is not mounted on the board and is not shown in this diagram

aerial and power sockets. Use screws and spacers to mount the circuit board to the box. Mounting the board horizontally assists troubleshooting.

Use a socket for the IC, and observe the CMOS handling precautions given in *An Electronic Die*, elsewhere in this book. When soldering the leads to the ceramic resonator, do it quickly – excessive heat damages the device. The earth lead running acoss the bottom of the board must be connected to the metal case by a short length of stout wire.

Testing

After carefully checking your wiring, both against the circuit diagram and the layout diagram, it is time to test your circuit. You will need a multimeter, an 80 m SSB receiver and a 50 Ω dummy load. A good design of dummy load can be found in the project *A Switched Dummy Load*, also in this book. An RF power meter and frequency counter will also be useful, although if your receiver has a digital frequency readout and S-meter, the latter two items are not really necessary. You will also need a 12 V 1 A power supply unit (PSU) to power the transmitter.

Switch the transmitter to receive and switch on the transmitter. No current should be consumed. Switch to transmit and check that pin 14 of IC1 is 6.8 V positive. With the dummy load connected to the aerial socket, press the key. The voltage on TR2 collector should now be 12 V, dropping to zero when the key is released.

Now check the operation of the oscillator. In transmit mode, you ought to be able to find a strong carrier signal with the receiver, even though the dummy load is connected. Adjusting the variable capacitor should change the frequency. At the lower end of the frequency range, you may find that the oscillator is unreliable in starting, because the circuit is attempting to pull the resonator too low in frequency. If this is the case, set the trimmer at the back of VC1 to minimum capacitance. If your version of VC1 has two trimmers, and you don't know which one to set, set them *both* to minimum capacitance. If there is still a problem, reduce the value of RFC1 to 6.8 or 4.7 μH.

In all probability, the unmodified circuit of Figure 2 will not require any of the changes outlined here. A coverage of 3.518 to 3.558 MHz should be possible, while preserving good frequency stability and reliable oscillation.

A signal probe (see *An RF Signal Probe*, elsewhere in this book) is useful for checking the operation of the oscillator and PA. Alternatively, an RF power meter or the receiver's S-meter can be used. With the PA running, the unit should draw between 200 and 300 mA. If TR1 becomes too hot to touch after a few seconds of transmitting, increase R5 or R6 to limit the transistor's heat dissipation. A small 6.3 V bulb connected across the aerial output is a simple way to check that the PA is working. An orange/white glow when the key is pressed is indicative of correct operation.

The final test is to monitor keying 'quality'. With your dummy load connected, press the key and listen to the note on the receiver's loudspeaker. Then operate the key, sending a string of dits, for example. What you hear should be free of chirps and clicks, as well as being stable in frequency. This test is sometimes better performed with no aerial connected to the receiver, thus preventing receiver overload and its associated plops. No problems should be encountered here.

Frequency tuning

A peculiarity of ceramic resonators is that, every now and again, their frequency changes abruptly by 100 Hz or so, then remains stable for some time. This is certainly noticeable in the received signal, but does not detract from the QSO and no characters are lost as a result. Try to keep the area around the ceramic resonator cool, to avoid temperature variations.

Parts list

Resistors: all 0.25 W, 5% tolerance

R1	10 megohms (MΩ)
R2	2200 ohms (2.2 kΩ)
R3	270 ohms (270 Ω)
R4	1 megohm (1 MΩ)
R5, R6	1.5 megohms (1.5 MΩ)
R7	1000 ohms (1 kΩ)

Capacitors

C1, C7	100 nF
C2	100 pF
C3	47 pF
C4	1 nF
C5, C6, C10	10 nF
C8	560 pF
C9	820 pF
VC1	10–160 pF variable

Inductors

RFC1	8.2 μH
RFC2	10 μH
RFC3	2.2 μH

Semiconductors

IC1	4069
TR1	VN10KM
TR2	BC640
D1	6V8 Zener

Additional items

Matrix board (see text)
14-pin DIL socket
Pointer knob
Sockets (see text)

85 An audio booster for your hand-held

Introduction

The audio output from many hand-held transceivers and receivers usually leaves much to be desired, so this little amplifier was designed to increase the output at minimal expense.

All that is needed is a separate amplifier and bigger loudspeaker. This is accomplished using a single integrated circuit (IC), a few components, and a loudspeaker from the junk box. This circuit will enable the output from your hand-held to be heard easily in a car.

The circuit

This is shown in **Figure 1**. It uses only those components necessary to operate the IC amplifier. VR1 is the preset volume control, and varies the signal coming from the 'External speaker' jack on the hand-held before feeding it into the IC for amplification. C1 blocks any constant voltage present on the input.

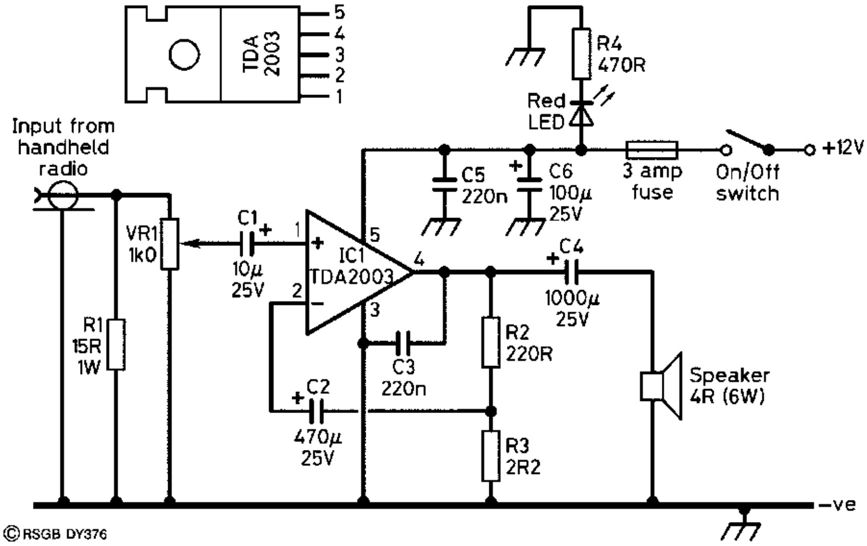
TDA 2003
5
4
3
2
1
R4
470R
Red LED
Input from handheld radio
+12V
C5
220n
C6
100µ
25V
3 amp fuse
On/Off switch
VR1
1k0
C1
10µ
25V
IC1
TDA2003
C4
1000µ
25V
R1
15R
1W
R2
220R
C3
220n
C2
470µ
25V
Speaker
4R (6W)
R3
2R2
–ve
©RSGB DY376

Figure 1 Circuit diagram of the audio amplifier. The power is derived from the cigarette lighter socket and the fuse is in-line with the lead

The IC output comes from pin 4 and is fed via the electrolytic capacitor, C4, to the loudspeaker. The circuit is provided with an on/off switch, fuse and LED to indicate when the circuit is switched on.

Construction

The box is made of aluminium. This is necessary to help to dissipate some of the heat generated by the IC. Do not build the circuit inside a plastic box unless you take special precautions! The IC has a metal mounting tab with a hole, specifically designed to be mounted to a metal box or other metal heat sink. Apply plenty of heat sink compound between the tab and the box, tighten the nut and bolt, and then wipe off any excess compound. The box will get slightly warm in operation.

The size of the box is not specified. You may want to decide on this when you find a loudspeaker. Choose one which will be able to handle 6 W

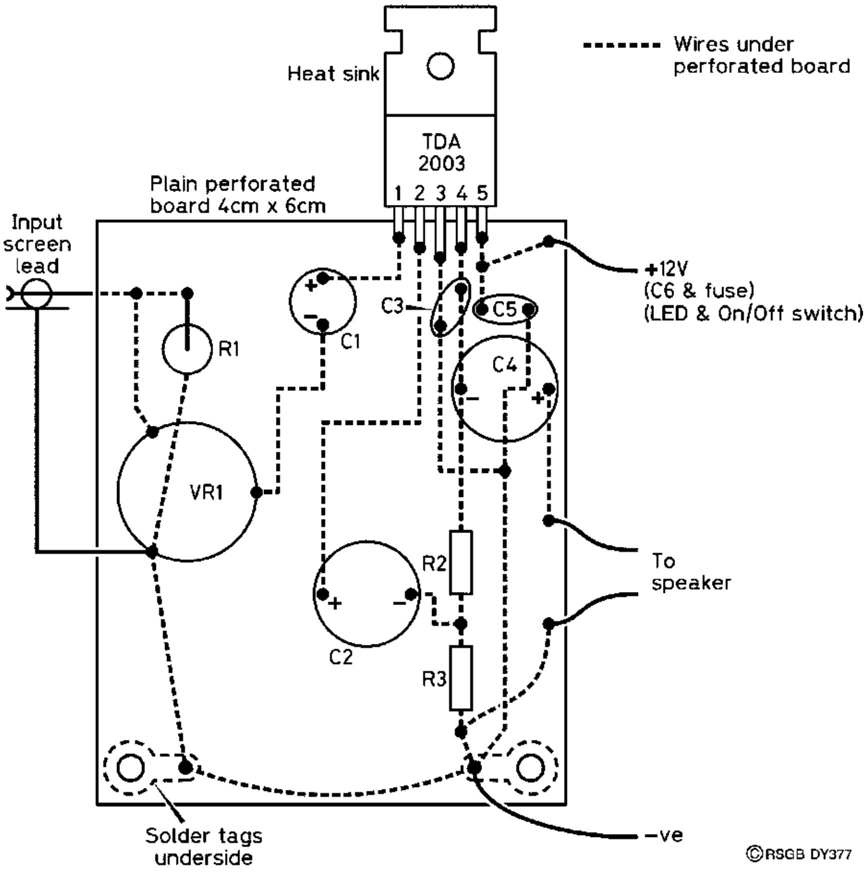
Heat sink
TDA 2003
1 2 3 4 5
Wires under perforated board
Plain perforated board 4cm x 6cm
Input screen lead
R1
C1
C3
C5
C4
VR1
R2
C2
R3
+12V
(C6 & fuse)
(LED & On/Off switch)
To speaker
Solder tags underside
−ve
©RSGB DY377

Figure 2 Layout of the components within the box

output. Drill all the holes in the box first. Holes for the speaker, input phono socket and the LED. The amplifier can be constructed on ordinary matrix board, which can be mounted inside the box with screws and spacers.

The layout is shown, for your guidance, in **Figure 2**. The components are mounted by pushing their leads through the holes in the board and making connections on the underside. The preset volume control, VR1, is set such that the hand-held's volume control is sufficient to control the final output over a good volume range. Use a screened lead from the 'External speaker' jack socket to the phono plug.

An external power supply is needed for this circuit. The normal dry battery which we usually use for small projects in this book will *not* work here, so you will need a proper mains power supply producing a stabilised 12 V. If you are going to use the unit principally in a car, then the cigar lighter socket can supply this voltage easily. Do make sure that the polarity is correct before you switch on!

When you plug the jack plug from your booster into the 'External speaker' socket on your hand-held, its internal speaker will be muted, so don't think that something dire has gone wrong! Adjust VR1 for a good volume range on your booster, when the volume control is turned on the hand-held.

Parts list

Resistors: all 0.25 W, 5% tolerance, unless otherwise stated

R1	15 ohms (15 Ω) 1 W
R2	220 ohms (220 Ω)
R3	2.2 ohms (2.2 Ω)
R4	470 ohms (470 Ω)
VR1	1000 ohms (1k Ω)

Capacitors

C1	10 μF 25 V
C2	470 μF 25 V
C3, C5	220 nF (0.22 μF) Mylar
C4	1000 μF 25 V
C6	100 μF 25 V

Semiconductor

LED	5 mm Red

Integrated circuit

IC1	TDA2003

Additional items
  Heat sink compound
  Nuts and bolts
  Loudspeaker 4 Ω 6 W
  3 A fuse
  On/off switch (SPST)
  Matrix board 4 × 6 cm
  Solder tags
  Plugs and screened cable for connecting lead
  Aluminium box

86 A grid dip oscillator

Introduction

When an inductor and a capacitor are connected, whether in series or parallel, they form a circuit with a natural (or *resonant*) frequency. The circuit stores energy, and this energy is being constantly shifted from the inductor to the capacitor and back again.

The dip oscillator is a simple instrument used to measure the resonant frequency of a tuned circuit without having to make any direct connection to the circuit. The circuit is more commonly known as the *grid dip oscillator* (GDO), from the days when the active device in the circuit was a valve. The FET or Field-Effect Transistor operates in a way which is very similar to that of the valve, so it is not quite a misnomer to call this instrument a grid dip oscillator, too.

The circuit

The GDO uses a calibrated, tunable FET oscillator in the circuit of **Figure 1**. It has a frequency range of 1.6 to 35.2 MHz in four ranges using a set of plug-in coils, shown in **Figure 2**. When the oscillator coil, L1, is placed near an external resonant circuit, some of its RF energy is coupled into the external circuit. A gain in energy of the external circuit must mean a loss of energy in the GDO circuit, resulting in a change of current through TR1, which is measured by the meter, M1.

The current through TR1 is of the order of 5 to 8 mA, but the change of current may be only a few microamps. To measure a very small

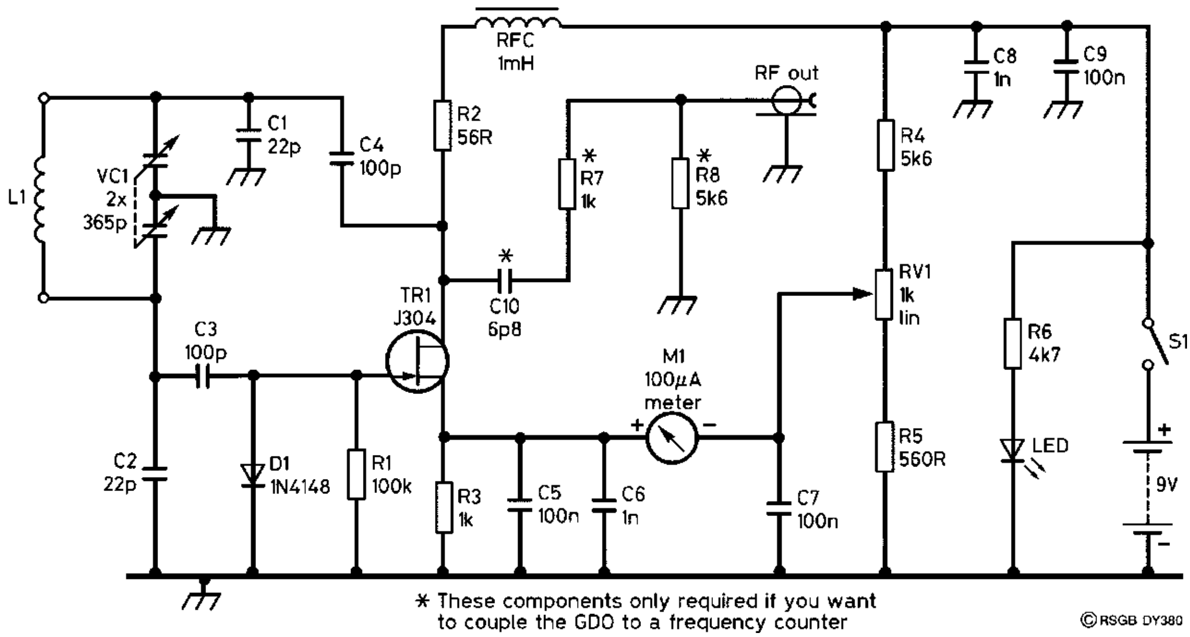
RFC
1mH
RF out
C8
1n
C9
100n
L1
VC1
2x
365p
C1
22p
C4
100p
R2
56R
R7
1k
R8
5k6
R4
5k6
C10
6p8
TR1
J304
C3
100p
RV1
1k
lin
R6
4k7
S1
M1
100μA
meter
C2
22p
D1
1N4148
R1
100k
R3
1k
C5
100n
C6
1n
C7
100n
R5
560R
LED
9V
*
+
–

* These components only required if you want to couple the GDO to a frequency counter

©RSGB DY380

Figure 1 Circuit of an FET GDO. The coils are wound on DIN speater plugs, which provide both a plug-in base and a coil former

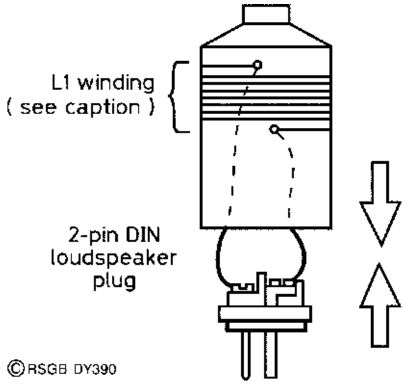
L1 winding
(see caption)
2-pin DIN
loudspeaker
plug

©RSGB DY390

Figure 2 Details of coil construction:
Range 1: 1.6–4.0 MHz 55 turns of 30 SWG
Range 2: 3.3–7.9 MHz 27 turns of 30 SWG
Range 3: 6.3–4.0 MHz 14 turns of 26 SWG
Range 4: 11.9–4.0 MHz 7 turns of 24 SWG.

change superimposed on a much larger standing current, the method of *offset* can be used.

One connection to the meter goes to the source of the FET, while the other goes to a variable offset voltage set by VR1. M1 has a full-scale deflection (FSD) of 100 μA. If the current through TR1 changes, the voltage across R3 changes. When there is no resonance, the voltage at the wiper of VR1 is set to be very slightly greater than that across R3, and there is a 75% FSD meter deflection. When the voltage across R3 decreases very slightly, due to external circuit resonance, a significant ‘dip’ in the meter deflection is produced, hence the name of the instrument.

The circuit is not difficult to make on standard matrix board. Provided you can follow a circuit and translate it into a good component layout, then this project is probably only an evening’s work.

The most important part of the GDO is the tuning capacitor and its associated frequency-calibrated dial. New, air-spaced tuning capacitors can cost you up to £20, so it is worth delving around in junk boxes, or scouring the tables at a local boot sale or rally. The tuning capacitor from an old transistor radio should be perfect. It may even have a slow-motion drive and a dial which can be remarked for the project.

Choose a coil plug and socket arrangement that is practical. Think about crystal holders or phono plugs and sockets. The prototype shown in Figure 1 and Figure 2 used 2-pin DIN plugs, with the coil wound on the outside of the plastic plug cover. Figure 2 shows the coil construction and the winding details.

If you use a variable capacitor, VC1, with a value different from that shown in the parts list, then the frequency ranges will be different. This does not matter, as it will be taken into account during calibration.

Position VC1 so that the dial will be easy to see and to operate, while locating the coil socket as close to it as possible. **Figure 3** shows the traditional layout of the GDO.

Calibration

Because the GDO also radiates a very small amount of energy, a general coverage receiver can be used to calibrate the dial. Don’t try to aim for great accuracy and clutter the dial with marks and figures! If you include C10, R7 and R8 in your circuit, you can connect a frequency counter directly to the GDO and leave it in circuit all the time.

The GDO in use

Always try to place the external tuned circuit with its coil coaxial with the plug-in coil, as shown in Figs 1 and 3. If the coils are at right angles, the GDO may not produce any resonance. Set the offset control to give about 75% FSD and slowly tune L1 through its whole range. If no dip occurs, you may have the wrong coil plugged in. When you eventually find a dip, move the external coil further away until only a minute dip is seen. You may have to retune the dip meter as you do this, but it gives a much more accurate reading of frequency.

Remember that you cannot ‘dip’ a coil by itself – there must always be a capacitor present.

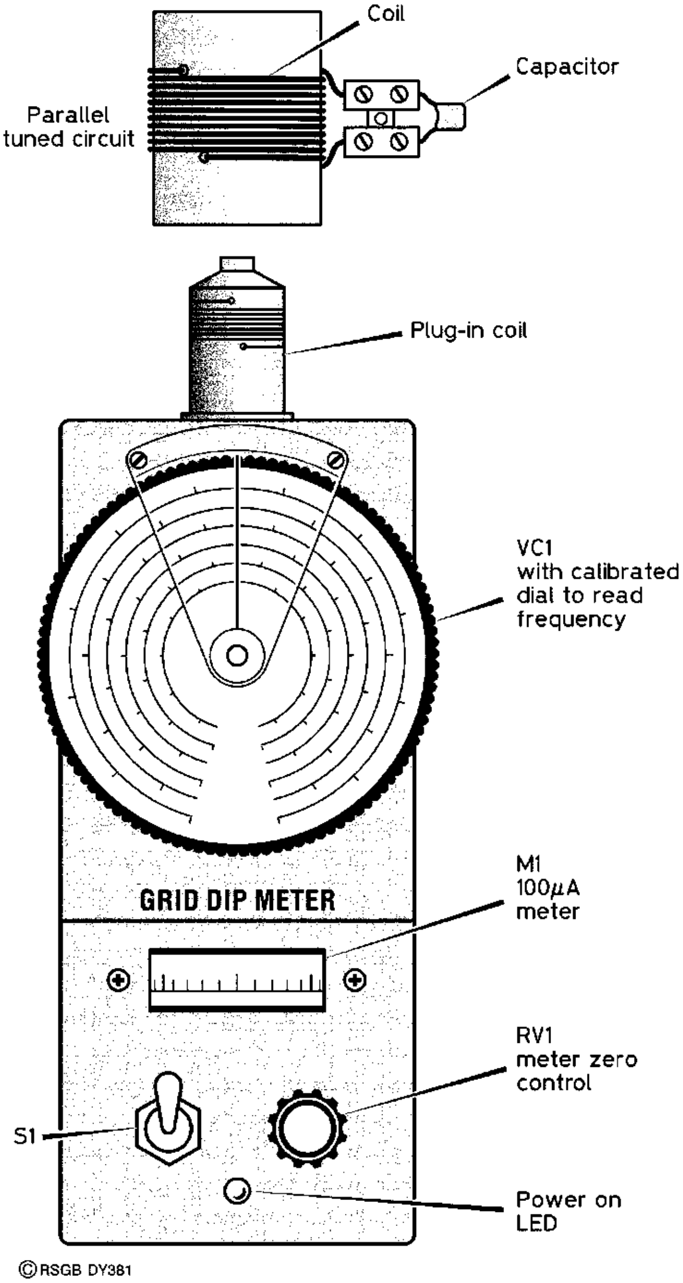
Coil
Capacitor
Parallel tuned circuit
Plug-in coil
VC1 with calibrated dial to read frequency
GRID DIP METER
M1 100μA meter
RV1 meter zero control
S1
Power on LED
©RSGB DY381

Figure 3 Layout of a typical GDO. The dial, meter and the location of the coil to the circuit under test can all be viewed at the same time. It is shown measuring resonance of a tuned circuit

Aerial resonance

Figure 4 shows how to check the resonance of a dipole aerial. Disconnect the coax feed at the aerial and place a short piece of wire, terminated with crocodile clips, across the centre insulator to short together the two ends of the aerial. By placing the GDO close to the shorting link, a dip should be seen on the meter while VC1 is turned. Alternatively, a loop in the element can be made around the coil, as in Figure 4, or the shorting link can be made long enough to loop over the coil. The latter method does not require tampering with the mounting and tensioning of the dipole wires.

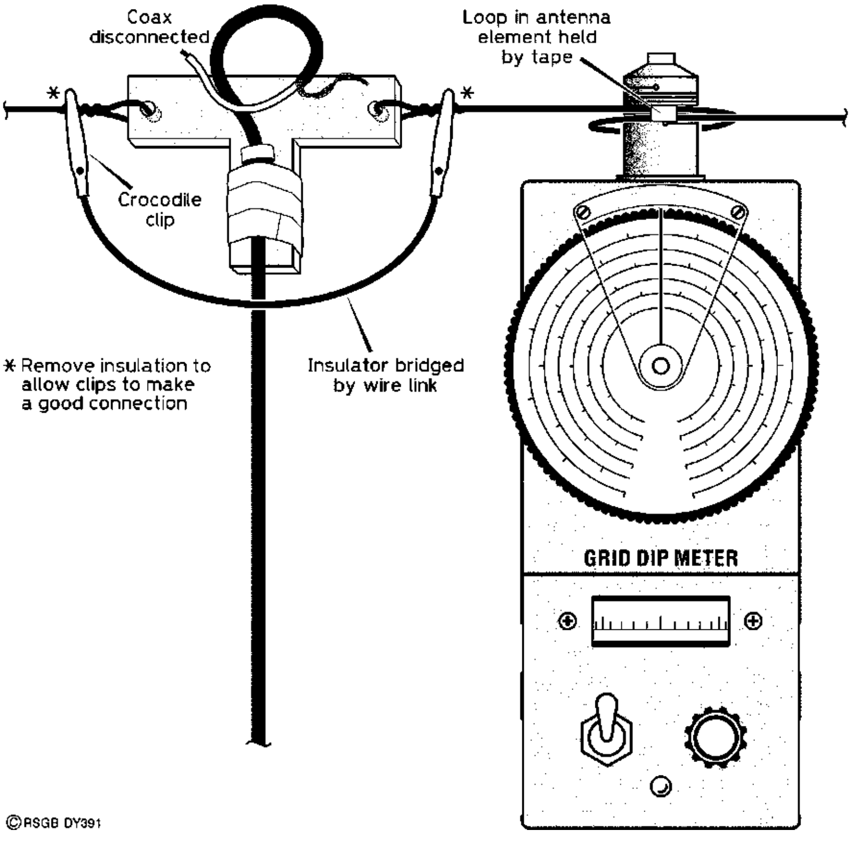
Coax disconnected
Loop in antenna element held by tape
*
*
Crocodile clip
* Remove insulation to allow clips to make a good connection
Insulator bridged by wire link
GRID DIP METER
©RSGB DY391

Figure 4 A wire antenna element can be looped into a single turn coil for increased coupling to the GDO

Parts list

Resistors: all 0.25 W, 5% tolerance

R1	100 000 ohms (100 kΩ)
R2	56 ohms (56 Ω)
R3	1000 ohms (1 kΩ)
R4	5600 ohms (5.6 kΩ)
R5	560 ohms (560 Ω)
R6	4700 ohms (4.7 kΩ)
R7	1000 ohms (1 kΩ)
R8	5600 ohms (5.6 kΩ)
VR1	1000 (1 kΩ) linear

Capacitors

C1, C2	22 pF
C3, C4	100 pF
C5, C7, C9	100 nF (0.1 μF)
C6, C8	1 nF (1000 pF)
C10	6.8 pF
VC1	2 × 365 pF

Semiconductors

TR1	J304 or similar
D1	1N4148
LED	

Additional items

L1	See Figure 2
RFC	1 mH
S1	SPST
M1	100 μA

Source

Components are available from Maplin.

87 A CW transmitter for 160 to 20 metres

Introduction

This very small transmitter is designed to work on any band from top band (160 m) to 20 m, with an RF output of 1 W. It will work on higher frequencies but with a reduced output.

The circuit

The three-transistor circuit is shown in **Figure 1**. It comprises a crystal oscillator using a BC182 transistor which drives a 2N3866 power amplifier (PA) keyed by a ZTX750 PNP transistor. The oscillator and PA are coupled by a capacitor and resistor; this provides a very small amount of positive bias to the PA.

The oscillator can be used as a basic crystal oscillator but, by including a variable series capacitor as shown in Figure 1, the crystal frequency can be 'pulled' slightly, making the oscillator a *variable crystal oscillator* (VXO).

Construction

The PCB layout is shown in **Figure 2** and in the photograph. Although the prototype was built around a PCB, this circuit is equally amenable to

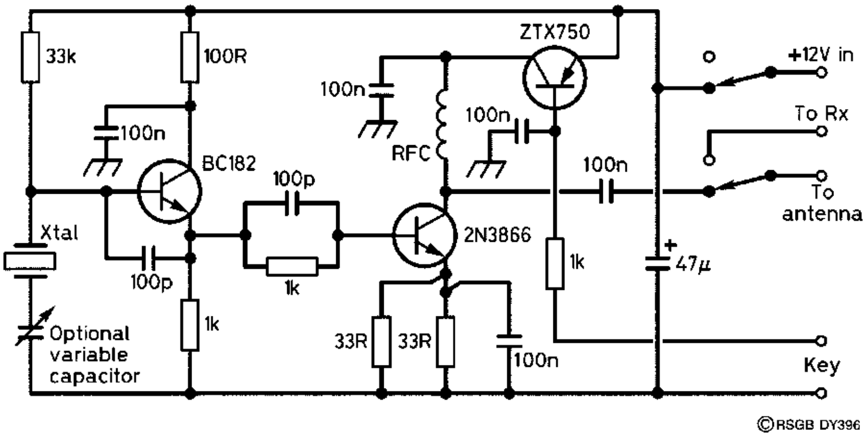
33k
100R
100n
BC182
100p
Xtal
100p
1k
Optional
variable
capacitor
1k
100n
RFC
ZTX750
100n
2N3866
33R
33R
100n
100n
1k
47µ
+12V in
To Rx
To
antenna
Key
©RSGB DY396

Figure 1 Transmitter circuit diagram

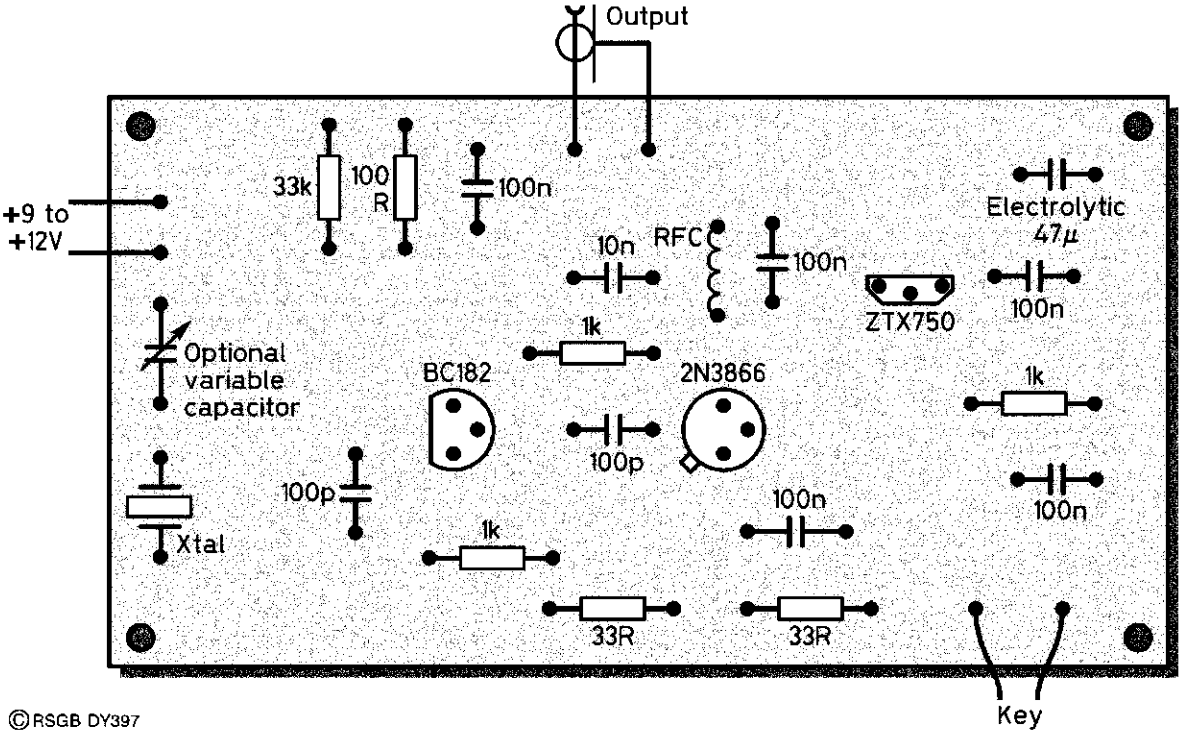
Output
+9 to
+12V
33k
100
R
100n
10n
RFC
100n
Electrolytic
47µ
100n
ZTX750
1k
Optional
variable
capacitor
BC182
2N3866
1k
100p
100p
Xtal
100n
100n
1k
33R
33R
Key
©RSGB DY397

Figure 2 Component placing on the PCB. The external connections are also shown

construction on a matrix board. Populating the PCB is very simple, and you can expect to be able to do this in about one hour. The radio-frequency choke (RFC) is made by winding 10 turns of 33 SWG wire on a ferrite bead. The enamel coating of the wire is intended to vaporise when soldered into the board, thus obviating the need to remove the enamel manually with a knife or sandpaper. However, if you do have problems with the PA either not working or keying intermittently, it is suggested that you investigate the RFC connections immediately!

If you decide to make the VXO version, you will have to cut the track between the crystal and earth, and connect the variable capacitor (250 pF) across the break.

In use

After performing the usual checks on the accuracy of your circuit building and the wiring of the external components, it is time to connect a 12 V battery between the points shown in Figure 2. Do not switch on yet. An aerial needs to be connected to the output via an ATU and a crystal, of frequency matching that of the aerial, fitted. The variable capacitor should give you a tuning range from about 14.058 to 14.064 MHz. 'Netting' (the process of tuning your transmitter to the same frequency as that of a received station) is achieved simply, because the oscillator is always running, and the leakage of the signal (despite the fact that the PA is not powered) is sufficient to bring the two frequencies to zero beat.

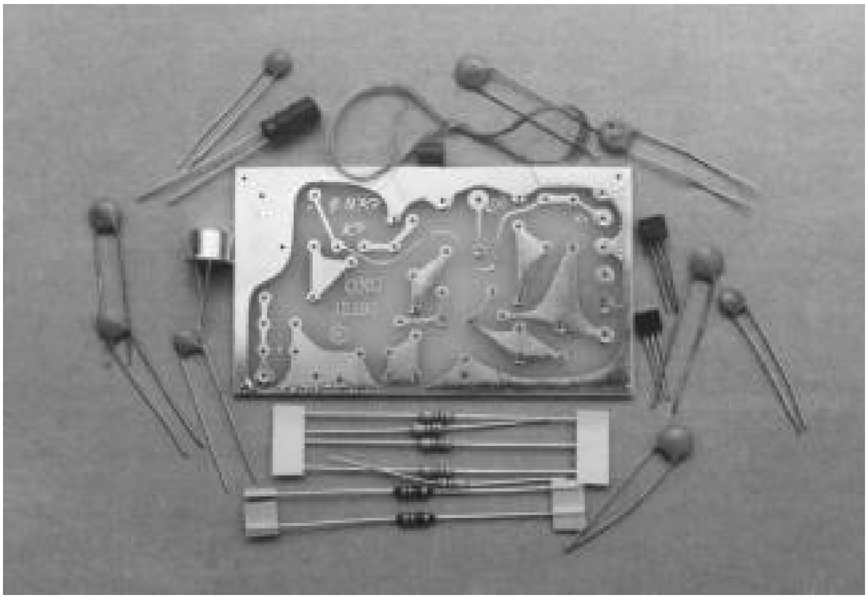

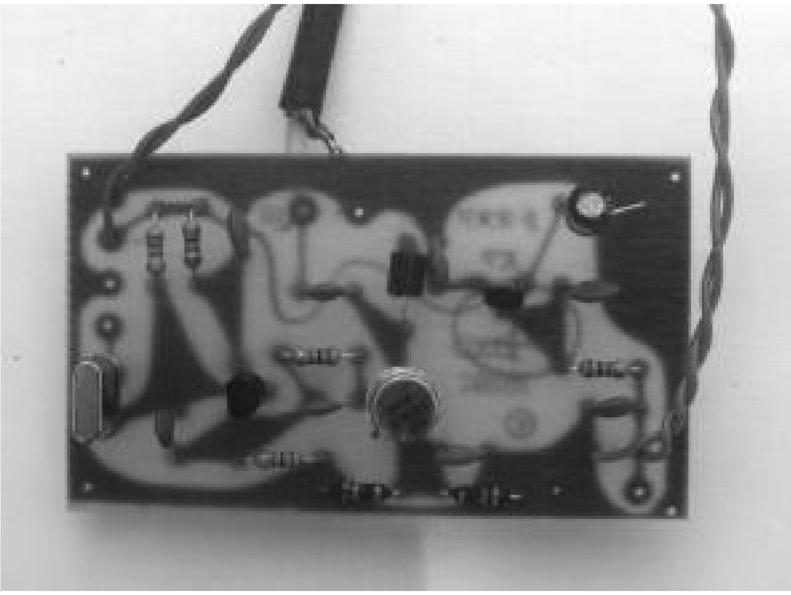

Warning

Note that the transmitter has no filtering; harmonics are not suppressed. It is strongly recommended that you use this transmitter in conjunction with the excellent low-pass filter described in the project *A 7-element low-pass filter for transmitters*, described elsewhere in this book.

88 Matching the end-fed random-wire aerial

Introduction

Many amateurs who do not have the space (or money) for a multi-band beam aerial, make use of the simplest possible alternative – the longest piece of wire that they can erect, with its end connected to the transceiver or receiver by an aerial tuning unit (ATU).

The length of the wire is not of major importance. Any length between 10 m and 80 m, with bends if necessary, will suffice. A good earth connection to the radio is just as important. Bends in the aerial wire can have some interesting effects on the directional properties of the aerial; V- or L-bends, or even a square shape are permitted. The only thing *not* permitted is to fold the wire back on itself in a tight hairpin bend!

Longish wire aerials

The term ‘long wire’ is usually used (incorrectly) to describe an end-fed aerial. How long is a piece of string? It depends what you mean by ‘long’. In aerial parlance, it means ‘long with respect to one wavelength’. Again, this depends on the band you are using. A long wire at 20 m is somewhat different from a long wire at 160 m. However, if you have sufficient real estate for a long wire on the 160 m band it must, by definition, be a long wire on all the other bands, too!

This should put you on your guard when analysing published data about feed-point impedance and the directional properties of a long wire aerial. Such theoretical data relate to a *real*, *ideal* long wire which is straight, horizontal, very high above perfect (conducting) ground, and not obstructed in any way. So your aerial doesn't quite match these criteria? Join the club!

Don't let this dampen your ardour when it comes to evaluating what the longish wire can do for you. The following should explain why.

Feed-point impedance

The impedance at the end of a longish-wire aerial can vary from a few tens of ohms to several thousand ohms, depending on the frequency in use and the wire length. It is also affected by factors such as bends, height above ground, proximity to buildings and wire diameter.

The actual value doesn't matter, provided we can make the aerial *appear* to have a 50 Ω impedance at the aerial socket of the transmitter. This process is what we call *impedance matching*, or simply *matching*. It maximises the power transfer from the transmitter to the aerial, and from the aerial to the receiver.

That is why an aerial tuning unit, or ATU, is almost (but not necessarily) obligatory.

The ATU

Many commercially produced HF receivers and transceivers have single 50 Ω coaxial sockets as their one and only means for connecting an external aerial. This means that an external aerial should have a 50 Ω feed impedance if it is to work efficiently, and it rules out most of the aerials being used by amateurs on the HF bands. Some means is necessary to change the aerial feed impedance to 'match' that of the transceiver. Such impedance-matching, or Z-matching (because Z is the symbol for impedance, just as L is the symbol for inductance) is the rôle of the ATU.

These can be bought and will accommodate either an end-fed or a coax-fed aerial. They can be bought ready for use or in kit form. Whether you want one for receiving only, or for use with a low-power (QRP) or high-power (QRO) transmitter, will determine what you need and how much you pay.

A simple single-band ATU

The simplest form of ATU is shown in **Figure 1**. It is simply a parallel LC (coil and capacitor) circuit, resonant at the chosen frequency, with taps on the coil for the aerial and the coaxial feed to the transceiver. If we assume the circuit is resonant, a high impedance exists at the top of the coil, and a low impedance at the bottom.

We said earlier that the end-fed longish wire presented an impedance which was high (or at least higher than 50 Ω). This explains why the aerial is tapped to the coil near the top, where the impedance is high, and the 50 Ω coax is tapped near the bottom, where the impedance is low.

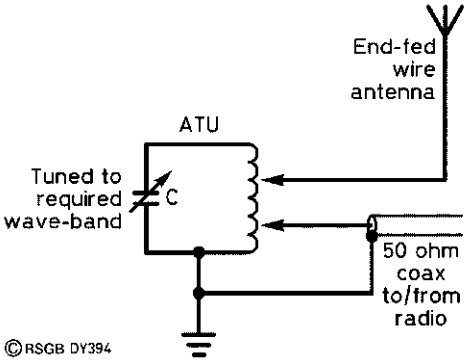

Figure 1 Parallel tuned circuit as single-band ATU for end-fed longish-wire antenna

Because the feed-point impedance of the aerial changes with frequency, so must the point at which the aerial is tapped to the coil to achieve impedance matching. The value of C must be changed also, to ensure that the circuit is resonant, and the 50 Ω tap will require tweaking also.

Setting up an ATU is quite simple. Make up an LC parallel-tuned circuit consisting of 50 turns of enamelled copper wire on an empty 35 mm film plastic container (or similar), tuned with a 500 pF variable capacitor. Make sure the enamel is removed from the ends of the wire before soldering.

Solder the inner wire of the coaxial cable from the radio to the first or second turn of the coil from its grounded end. Then solder the braid to the grounded end. Connect the aerial about one-third of the way down the coil from the top, removing the enamel at the connection point.

Adjust the variable capacitor for maximum noise or signal strength in the receiver. Then, try different tapping points from the aerial, to maximise the signal again. This matches the aerial impedance to that of the tuned circuit. Repeat the process with the coax tap, thus matching the impedance of the radio and the cable to that of the tuned circuit.

You will no doubt find that the tapping process on the coil was not easily accomplished, especially when the enamel must be removed at each tapping point without shorting adjacent turns together. It is therefore logical to produce a design where the taps have been prepared during the winding of the coil, and are selected with a rotary switch. To this end, the following multi-band design is described.

A multi-band ATU

The same type of coil design around a 35 mm film container is used (see **Figure 2**). The tapping points can be prepared in advance by a little judicious planning. Dismantle your original coil and measure the length of wire on it. It will be a little more than a calculation of $50 \times \pi D$ would suggest (where D is the diameter of the film container), due to the lack of tension in the coil and the wire diameter itself. You will need aerial tapping points at turns 1, 2, 3, 5, 10, 15, 20, 25, 30, 35, 40 and 45, corresponding to all the bands from 28 MHz to 1.8 MHz. The coaxial cable tap is fixed at turn 2 (an acceptable compromise). All the turns are counted from the earthy end.

Cut another piece of enamelled copper the same length as you used originally. Then, with the aid of an ordinary calculator, work out the positions of the points where the enamel must be removed for the taps. For example, turn 15 will have to be made $\frac{15}{50}$ of the way along the wire, turn 20 tap made at $\frac{20}{50}$ of the way along. So, if the length of wire is, say 4.7 metres, the two taps in question will be made at $\frac{15}{50} \times 4.7 = 1.41$ m and $\frac{20}{50} \times 4.7 = 1.88$ m from one end. This must be repeated for each of the tap positions, and the enamel removed ready for the wire to be soldered to it. With a

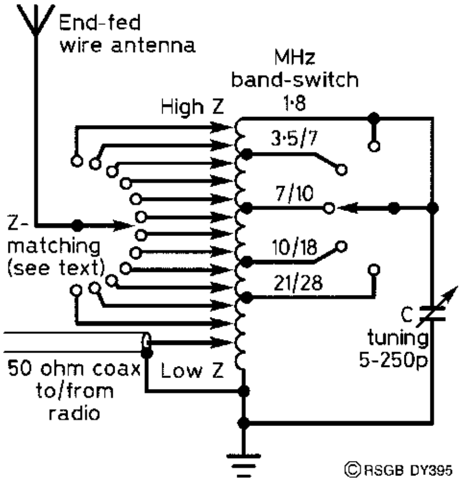

Figure 2 Simple multi-band ATU for end-fed longish-wire antennas

soldering iron, tin each tap point while the copper is shiny, thus ensuring a good, low-resistance connection. Do not solder on the taps yet.

The coil can now be wound as before. The taps can be soldered on, taking the lead from each one to the wafer of a single-pole 12-way rotary switch, the pole being connected to the aerial.

The tuning of the ATU is carried out by the same 250 pF capacitor, with a single-pole 5-way rotary switch used to select the band. Its tapping points will need to be chosen manually, using the method described earlier. Don't attempt to make new tapping points on the coil for this – use the taps available on the wafer of the other rotary switch, and find which is optimum for each band.

Notes

- For the aerial, use PVC-covered stranded tinned copper, of size 16/0.2 mm or 24/0.2 mm.
- Make the wire as long as possible, but anything over 10 m should be OK.
- Keep the wire as high as you can, in the clear and away from obstructions.
- Don't worry about bends, but don't use hairpin bends.
- Use a good insulator to attach your aerial – anything plastic will do.
- Anchor the wire near the point of entry to the building, but use a U-bend to prevent ingress of water.
- The wire can be brought in through the corner of a window, the PVC acting as an insulator. If you must drill a hole in the brickwork, make sure it slopes upwards from outside, so that water is deterred from entering.
- Use a good RF earth (as opposed to an electrical earth) such as half a dozen bare copper wires buried under the lawn in a fan shape. They should be joined together at the point of the fan and strapped to the earth connector of your ATU.

ARRL's

$19.95

RF Amplifier Classics

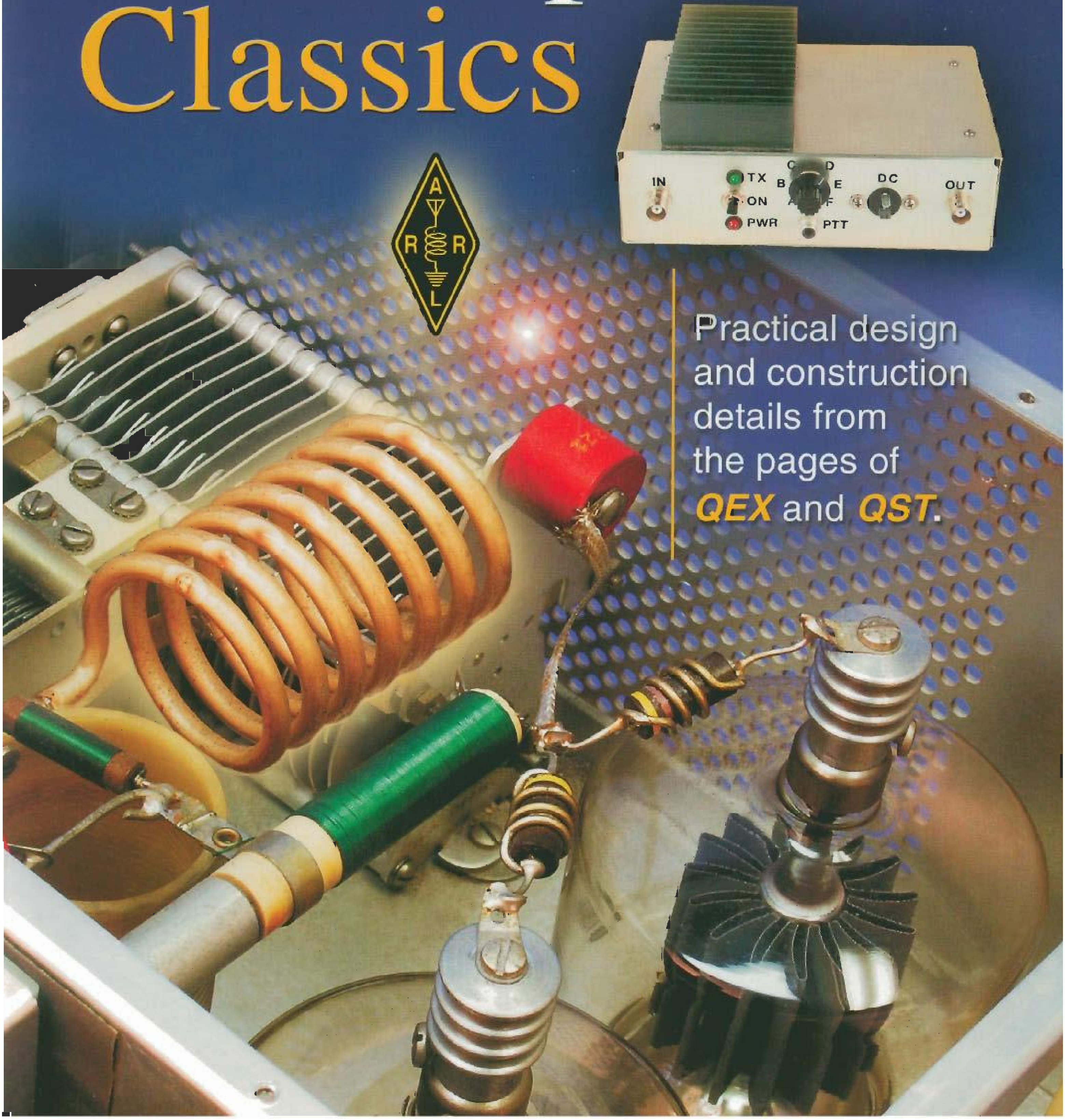
IN
TX
ON
PWR
A
B
C
D
E
F
PTT
DC
OUT

Practical design
and construction
details from
the pages of
QEX and ***QST***.

Contents

HF/MF/VHF Amplifiers (1 to 54 MHz)

An Easy-to-Build 25-Watt MF/HF Amplifier	Gary Breed, K9AY
A Compact 1-kW 2-50 MHz Solid-State Linear Amplifier	H.O. Granberg, K7ES/OH2ZE
A Broadband HF Amplifier Using Low-Cost Power MOSFETs–Parts 1 and 2	Mike Kossor, WA2EBY
A 1.8 to 54 MHz 5-Watt Amplifier	Zack Lau, KH6CP
An Experimental Solid-State Kilowatt Linear Amplifier for 2 to 54 MHz	Joel Paladino, N6AMG
An All-Band, 1500-Watt-Output 8877 Linear Amplifier–Parts 1 and 2	Jerry Pittenger, K8RA
High-Efficiency Class-E Power Amplifiers–Parts 1 and 2	David Rutledge, KN6EK, et al.
A 100-W MOSFET HF Amplifier	William E. Sabin, W0IYH
The FARA HF Project	Jim Valdes, WA1GPO

VHF/UHF Amplifiers

A Compact "Brick" for 6 Meters	Pat Bunn, N4LTA
A 300-W MOSFET Linear Amplifier for 50 MHz	Richard Frey, K4XU
A No-Bandswitch, Dual-Band VHF Desktop Amplifier	Paul Hewitt, WD7S
An 8-Watt, 2-Meter "Brickette"	Bob Larkin, W7PUA
903-MHz Linear Amplifiers–Parts 1 and 2	Dave Mascaro, WA3JUF
25-Watt Linear Amplifiers for 144 and 220 MHz	Dave Mascaro, WA3JUF
A High-Power 2-Meter Amplifier Using the New 3CX800A7	David Meacham, W6EMD
A Quick Powerhouse	Russ Miller, N7ART
A Cathode-Driven Tetrode for 6 Meters	David Munyon, W7DVB
A UHF Amplifier—from Scratch	John Reed, W6IOJ
A Solid-State 6-Meter Linear Amplifier You Can Build	Tarmo Tammaru, WB2TMD
Build a 6-Meter "Mini-Lini"	Wilson Hoag, WA5OLT
A Grounded-Grid Kilowatt Amplifier for 432 MHz	Stephen J. Powlishen, K1FO

Microwave Amplifiers

A Quarter-Kilowatt 23-cm Amplifier–Parts 1 and 2	E.R. "Chip" Angle, N6CA
A 2-W 13-cm Amplifier	Zack Lau, KH6CP
1296-MHz Solid-State Power Amplifiers	Al Ward, WB5LUA

Amplifier Maintenance

Amplifier Care and Maintenance	Ward Silver, N0AX

Foreword

If antenna gain alone is not sufficient to bridge the path between two stations, the alternative is to increase RF power. Hams have been aware of this fact since the earliest days, and that is why RF power amplifier projects have always been popular.

In *RF Amplifier Classics* we have assembled a collection of articles published in *QST* magazine, and its sister technical journal, *QEX*. The collection spans the early 1980s through 2003 and includes many prominent authors. In those few instances where an author provided a design revision (or correction) after the article was published, that revision is *included* in the article as presented in this book.

See the leatest issue of *QST* for other ARRL RF design-related publications, or visit our on-line bookstore at **www.arrl.org/catalog**. Please take a few minutes to give us your comments and suggestions on this book. There's a handy Feedback Form for this purpose at the back, or you can send e-mail to **pubsfdbk@arrl.org**.

Our thanks to the many authors whose work appears in this book. Without their willingness to share their knowledge with the amateur community, *RF Amplifier Classics* would not be possible.

Dave Sumner, K1ZZ
Executive Vice President
Newington, Connecticut
August 2004

By Gary Breed, K9AY **From *QST*, February 1994**

An Easy-to-Build 25-Watt MF/HF Amplifier

Do you need a medium-power linear amplifier for SSB or CW? Congratulations—you just found it!

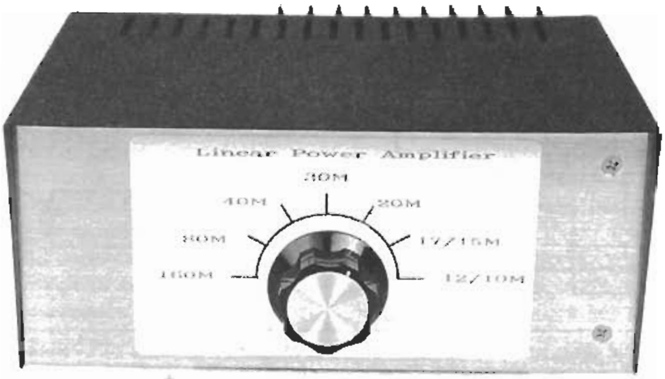

Here's a 25-W, 1.8- through 30-MHz, class-A linear power amplifier that's simplicity itself. What makes it simple is the use of a self-biased transistor module requiring few external components. To control harmonic output, a set of five-section low-pass filters is included. Power-supply requirements are +28 V at 2.5 A and —5 V at 200 mA.[1] With a gain of about 13 dB, a 1- to 1.4-W driving signal is all that's needed to deliver 25 W output. Gain is flat within ±0.75 dB across the covered frequency range.

If 25W isn't enough for you, it's easy to directly apply the design information to build a 50-W amplifier—all you do is use a larger transistor module! Another step toward project simplicity is the availability of kits. Each kit contains all the major components for either a 25- or 50-W version.[2]

Amplifier Design

When designing a power amplifier, the first step is to select the right transistor(s). Excellent bipolar-junction transistors (BJTs) and field-effect transistors (FETs) are available from well-known companies such as Motorola, M/A-COM PHI, SGS-Thomson, Philips, Mitsubishi and others. A number of smaller companies also make power transistors, usually for more-specialized applications. MicroWave Technology, Polyfet RF Devices, and Directed Energy may be company names unfamiliar to you, but they all make power transistors for MF and HF applications.

In this amplifier, I use the SLAM-0111 from MicroWave Technology.[3] I didn't choose it because of its gain, its efficiency, or even its price; I selected it because it's very easy to use. The device consists of two power JFETs (the particular specialty of MicroWave Technology), operating in push-pull. Since JFETs behave similarly to triode vacuum tubes, the company dubbed them *Solid State Triodes*. SLAM (Solid-state-triode Linear Amplifier Module) devices include thick-film bias resistors in the package with the transistors. These resistors set the gate bias for class-A operation, and establish a 50-a input impedance. At the rated power and supply voltage, the push-pull output impedance is also 50 Ω!

With such convenient input and output impedances, matching the devices to a 50-Ω system merely requires 1:1 balun transformers at the input and output. Because the bias voltage is internally generated, the only other external circuitry required is a suitably bypassed and isolated 28-V power supply.

Circuit Description

The amplifier schematic is shown in Fig 1. The balun driving the gates of the push-pull transistors is a conventional transformer. The primary and secondary windings are each three turns of #28 wire, wound on a two-hole ferrite balun core of 73 material ($\mu_i = 2500$). These transformers are broadband enough to provide 1.8- to 30-MHz operation and offer dc isolation with no additional components. The input-transformer primary is center-tapped and bypassed to provide access to the gates for external dc bias (more on this later).

The output transformer is constructed in the same manner as the input transformer—it's just larger. Two ferrite beads of 77 material ($\mu_i = 2000$) make a two-hole core, with primary and secondary windings of three turns each, using #24 hookup wire. The primary (transistor side) is center-tapped and bypassed to provide dc voltage to the drains. Feeding dc through a center-tapped transformer eliminates the need for the usual bifilar RF choke seen in push-pull amplifiers—another reduction in the component count. Multiple bypass capacitor values (0.01, 0.1 and 10 µF) are used to cover the MF/HF range. That's the basic amplifier block: two transformers, a SLAM device, and a few bypass capacitors!

Class-A Operation Notes

By definition, transistors operating in class A conduct over the entire 360 degrees of the signal (that's all the time, of course). This operational mode assures that the transistor is always operating in the linear region of its input-to-output transfer characteristic. To do this, the device must be biased to handle the maximum signal at all times.

Obviously, this class of operation is pretty inefficient, since full current is drawn whenever the amplifier is on. A "perfect" transistor operating class A can only be 50-percent efficient, and real transistors

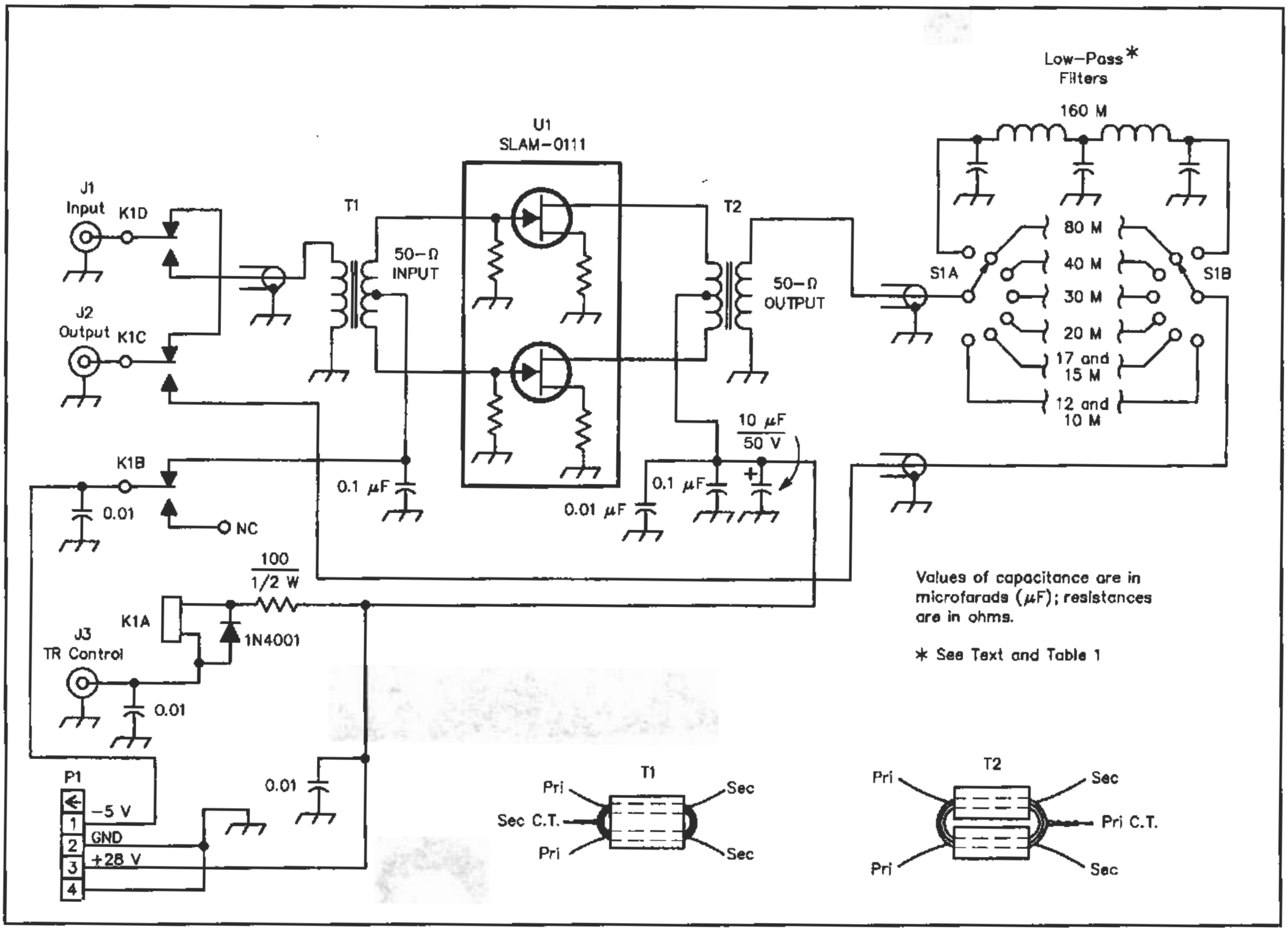

Fig 1—Schematic diagram of the 25-W class-A amplifier. Unless otherwise specified, resistors are ¼-W, 5%-tolerance carbon-composition or film units. Equivalent parts can be substituted.

J1—Panel-mount BNC socket.
J2—SO-239 connector.
J3—Phono jack.
P1—4-pin male Jones plug.
K1—3PDT relay with a 24-V dc coil. A surplus Potter & Brumfield KHP series 4-pole relay is shown in Fig 6; one pole is unused. (All Electronics catalog number 4PRLY-24N [$4] or Ocean State Electronics R12-17D3-24 [$10.90] are suitable. See the Part Suppliers List on pp 35-40 of *The 1994 ARRL Handbook* for addresses and telephone numbers. —Ed.)
S1—2-pole, 7-position ceramic rotary switch. My switch is made from two surplus CRL 11-position switch wafers and an indexing assembly providing selectable stops. The wafers are spaced about 1½ inches apart. CAL PA-200 series switch wafers and PA-300 series shaft and indexing assemblies are suitable (switches are available from Newark Electronics; tel 312-784-5100, fax 312-784-5100, ext 3107, to locate your nearest Newark distributor).
T1—Primary: 3 turns #28 AWG; secondary, 3 turns #28 AWG, center-tapped. Core: Fair-Rite #2873002402 balun (Amidon BN 73-2402).
T2—Primary, 3 turns #24, center-tapped; secondary, 3 turns #24. Core: two Fair-Rite #2677006301 beads (Amidon FB 77-6301).
U1—SLAM-0111 ultralinear 25-W, class-A, self-biased power FET module or SLAM-0122, 50-W version (Microwave Technology, 4268 Solar Way, Fremont, CA 94538, tel 510-651-6700, fax 510-651-2208).
Misc: RG-174 coax, enclosure (3½ x ⅞ x 5¼ inches [HWD]), heat sink (3 x 4⅛ x 1⅜ inches [HWD]), PC-board material, knob, mounting hardware.

do no better than about 40 percent. This amplifier draws 2.5 A from a 28-V power supply for an input power of 70 W. When it is providing 25 W, it's 36-percent efficient. (When there is no input, it's 0-percent efficient!)

To help reduce the heat generated by an amplifier that requires 70 W, a negative bias can be applied to the gates when not transmitting. A bias of —5 V results in a 0.25-A standby drain current instead of the full 2.5 A. The internal bias resistors are about 50 ~ on each gate, and dissipate a maximum of I W. Under these biasing conditions, the resistors each dissipate 0.5 W. Don't try to cut off the transistors completely with greater bias voltage! You'll risk burning out the resistors.

Some may ask, "If class A is this power hungry, why use it?" In a word: linearity. If you want excellent linearity (which means minimum distortion caused by harmonics or intermodulation), class A is the way to go. For example, all small-signal amplifiers for receivers and low-level transmitter stages operate class A because they must handle signals without distortion. However, they operate at very low power, so power dissipation is rarely an issue. This power amplifier further minimizes distortion by using push-pull operation, which cancels even-order distortion products in the output and makes the next part of the design easier than usual.

Harmonic Filter Design

As mentioned previously, the amplifier uses several low-pass filters to cover the nine MF/HF amateur bands. Each filter was initially designed for a cutoff frequency 20 percent higher than the upper end of their respective 160, 80, 40, 30, 20, 15 and 10-meter ham bands. The 15-meter filter is also used for 17 meters, and the 10-meter filter for 12 meters.

With no filtering, even-order harmonics (2nd, 4th, etc) are more than 40 dB below the carrier, the result of good push-

Table 1

Filter Circuit and Comparison of Ideal and Final Component Values.

	Ideal Filter Values			*Actual Filter Values*		
Cutoff Freq. (MHz)	*C1, C5 (pF)*	*C3 (pF)*	*L2, L4 (µH)*	*C1, C5 (pF)*	*C3 (pF)*	*L2, L4 (µH)*
2.40	1521	2620	4.55	1470 (1000 + 470)	2880 (2200 + 680)	4.41 (30 t on T50-2)
4.80	761	1310	2.27	830 (560 + 270)	1430 (1000 + 430)	2.37 (22 t on T-50-2)
8.76	417	718	1.25	430	820	1.25 (16 t on T-50-2)
12.18	300	516	0.900	300	560	0.960 (14 t on T-50-2)
17.22	212	365	0.634	220	370 (270 + 100)	0.706 (12 t on T-50-6)
25.74	142	244	0.424	150	240	0.460 (10 t on T-50-6)
35.64	102	176	0.306	100	180	0.314 (8 t on T-50-6)

In some cases it is necessary to parallel two smaller-value capacitors to obtain the proper values of capacitance for C1, C2 and C5

The inductors are wound on T-50-2 or T-50-6 cores. Inductors for the 160- and 80-meter filters are wound with #26 AWG wire in order to fit all turns on the cores; the other inductors are wound with #22 wire.

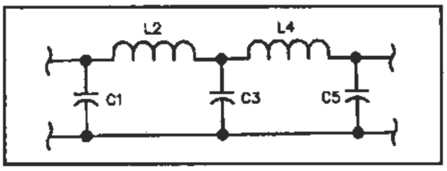

Fig 2—Schematic of the filter used for each band.

pull balance using factory-matched transistors. The 3rd and 5th harmonics are more than 15 dB down. To reduce the 3rd harmonic to at least 50 dB below the carrier, a five-section Chebyshev filter with low passband ripple is an appropriate choice. This type of filter has a good SWR in the passband, and a smooth roll-off characteristic. The design process began by creating ideal designs using a public-domain filter design program.[4]

Ideal designs rarely correspond to standard capacitor or inductance values that can be realized with a discrete number of turns on common toroid cores. Using a circuit analysis program,[5] the ideal designs were analyzed to see the effects of such realworld limitations on harmonic rejection and SWR performance.

First, the ideal component values were entered into the program, and varied ±20 percent to see which ones had the greatest effect on performance. Cl and CS (see Fig 2 and Table 1) were found to be least sensitive to variations, L2 and L4 were moderately sensitive; varying C3 had the greatest effect on both passband and stopband performance. The ideal capacitor values were then replaced with standard capacitor values or—in some cases— parallel combinations of two common capacitor values. Inductors were given the nearest value available for coils wound on either T-50-2 or T-50-6 toroid cores. The final filter designs are the result of trade-offs between inductance, capacitance and filter performance. Table 1 shows the filter topology, along with a comparison of the original ideal filter component values and the values selected for the finished unit.

Construction

I built my amplifier and low-pass filter modules on single-sided PC boards, using pads to mount the components. No holes are drilled (except for mounting screws) and all leads are attached by soldering them to the pads. The PC-board patterns for the amplifier and filters are available (see Note 2).

Fig 3 shows the amplifier-assembly parts. This assembly is mounted to a heat sink (see Figs 4 and 5) capable of dissipating more than 40 watts without excessive temperature rise. (This assumes a worst case of 50-percent transmitting time, and 7-watts dissipation in standby.) A cutout in the middle of the amplifier board allows placement of the SLAM device. The PC board leaves a conducting path around the ends of the SLAM to maintain a ground potential across the entire board. Four mechanical components make up the amplifier assembly. The first is a 0.1875-inch-thick aluminum base plate to which the SLAM is mounted. Next are two aluminum 0.1-inch-thick spacers, which are placed between the base plate and the circuit board. These spacers set the proper distance from the base plate to SLAM leads. The SLAM is installed through the top of the PC board, and its leads are soldered to the traces on top of the board.

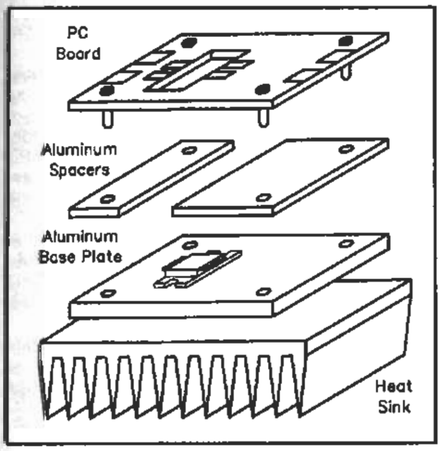

Fig 3—Mechanical assembly of the amplifier-module PC board, aluminum spacers and heat sink.

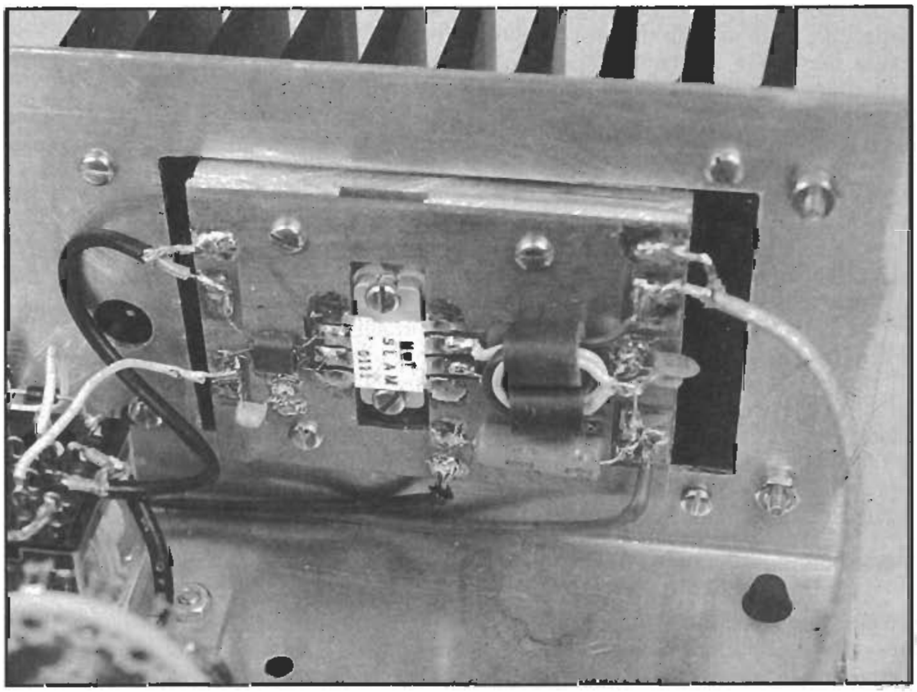

Fig 4—The assembled amplifier-module PC board in position and secured to the heat sink (see Fig 5, next page).

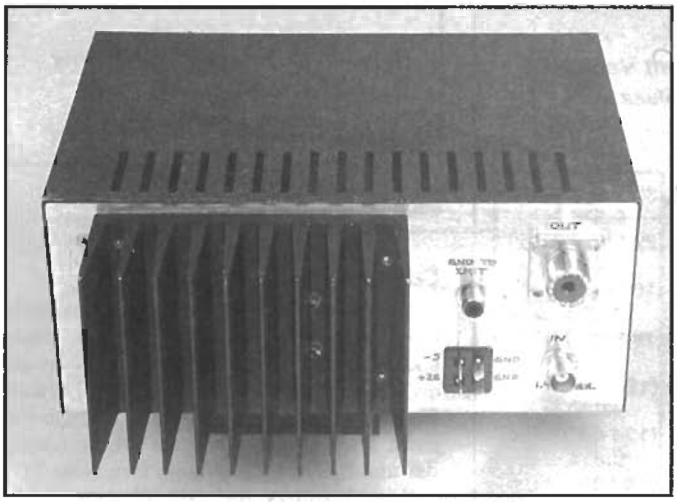

Fig 5—Rear view of the completed amplifier showing the hefty heat sink.

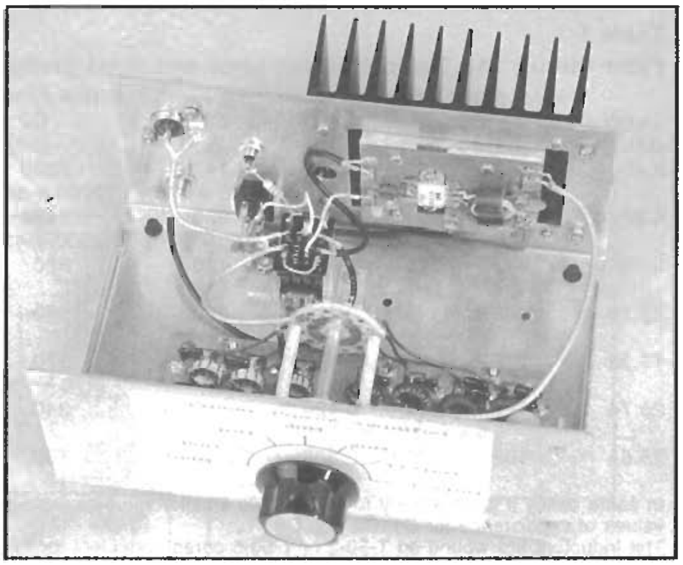

Fig 6—This interior view of the amplifier shows its simple and clean layout. The band switch is centered on the front panel. Immediately beneath the band switch is the filter assembly. Behind the switch and to the left is the TR relay, K1. A four-pin Jones plug power connector is mounted on the rear panel behind and to the left of the relay. On the bottom, near the outside lip of the rear panel, is J1. Above it is J2, with J3 to its right. Most of the rear panel—from its middle to the right lip—is occupied by the SLAM IC PC board and the aluminum spacers secured to the heat sink mounted on the rear panel's exterior. Rubber feet on the cabinet bottom help prevent scratching the supporting surface beneath and keep the amplifier from sliding. The band-switch knob center section is 1⅛ inches in diameter; the skirt flares to a diameter of 1½ inches.

Construction is easiest if the transformer connections to the SLAM are not soldered until after the SLAM is installed. This eliminates the possibility that the transformer connections will get in the way when you try to solder the SLAM into place. As with any power device, place a thin coating of thermal compound between the SLAM and the base plate, and between the base plate and the heat sink. Solder bypass capacitors directly to the transformer center tap and to the ground plane, with the minimum possible lead lengths.

The low-pass filter board is constructed one filter at a time. First, install the capacitor at the center (C3), then the inductors L2, L4), and finally the end capacitors (C1, C5). All inductors are wound with even spacing over three-quarters of the core circumference. Simply solder the capacitors to the pads and ground plane. Silver-mica capacitors were used in the prototype because they were on hand. Ceramic-disc capacitors with 200- to 500-V ratings will work equally well.

If the band switch is located close to the filter board (see Fig 6), short lengths of hookup wire can connect the filters to the switch wafers.

A spacious box houses the filter and amplifier assemblies, along with a TR relay that also switches the standby bias. Power and relay control leads are bypassed where they enter the enclosure.

Before final assembly, I gave the panels of the case a brushed look using a sanding block with oiled sandpaper. Band markings for the switch (see the title-page photo) are drawn on a large, adhesive-backed label attached to the front panel.

Performance

Amplifier gain ranges from 12.5 to 14 dB between 1.8 and 30 MHz. The gain flatness is basically a function of the input and output transformers. (It's possible to make the amplifier gain flat within 1 dB from 1 MHz to 100 MHz using transmission-line transformers and frequency compensation.) The required drive power for 25 watts output is 1.0 to 1.4 watts.

On-the-air performance is excellent. Besides low distortion in the SSB mode, a small advantage of linear amplification is a complete absence of rise and fall distortion of a CW waveform, which sometimes occurs in class-C amplifiers.

Summary

This project shows how new RF products can make home construction of amateur equipment very easy. Home-brewers can benefit from a growing trend in RF product engineering: reducing development time by using "super components" that require few external components and little engineering time to design them into a product.

A secondary purpose of this project is to show how even simple software tools can be used to speed up design. The programs used to design the amplifier's low-pass filters are inexpensive, and accurate at frequencies in the MF/HF bands. In this case, they made it possible to examine tradeoffs among standard-value components for seven different filters, without having to build, measure and tweak each one.

The result is a linear power amplifier with good gain and performance. Its uncomplicated design leaves little room for error, and no fancy test equipment is needed to successfully build it. Projects this easy can make an old-timer forget about the "simpler" days of vacuum tubes!

Notes

[1] Power supplies are available from Marlin P. Jones & Assoc, mc, PO Box 12685, Lake Park, FL 33403-0685, tel 407-848-8236; fax 1-800-432-9937.

[2] Parts kits for this project are available from Crestone Engineering, PO Box 3702, Littleton, CO 80161, tel 303-770-4709. Each kit includes all electronic and mechanical components for the amplifier module and lowpass filter assembly, including circuit boards, heat sink and rotary band switch. Kits do not include an enclosure, connectors, or TR relay. A 25-W kit using the SLAM-0111 is $115; a 50-W kit using the larger SLAM-0122 is $190. Add $6 per kit for shipping. Payment may be made by check, money order, VISA, MasterCard or American Express.

PC board patterns for the amplifier and filters are availablefreefrom the ARRL. Send your request to the Technical Department Secretary, ARRL, 225 Main St, Newington, CT 06111. With your request for the BREED AMPLIFIER PC-BOARD TEMPLATE, enclose a business-size envelope with one First-Class stamp.

[3] Microwave Technology, 4268 Solar Way, Fremont, CA 94538. Their products are distributed by Richardson Electronics, 40W267 Keslinger Road, LaFox, IL 60147, tel 708-208-2200.

[4] Mike Ellis, "A Comprehensive Filter Design Program," RF Design, July 1991. The program is available from the RF Design Software Service, PO Box 3702, Littleton, CO 80161-3702, tel 303-770-4709 (part #RFD0791, $15 postpaid).

[5] NOVA, a shareware program by Robert Stanton, also available from the RF Design Software Service (part #RFD-0391 ,$15 ppd).

By H.O. Granberg, K7ES/OH2ZE **From *QEX*, July 1990**

A Compact 1-kW 2-50 MHz Solid-State Linear Amplifier

Solid-state high-power linear amplifiers are becoming more and more popular in the field of ham radio as the prices of HF power transistors continue to fall. 250-W devices are now available for almost half the price they were selling for a few years ago. RF power FETs are still more expensive, but eventually their prices will also fall, although not as fast since they are still novelty items and the manufacturing yields are low due to ESD problems and requirement for cleaner facilities for wafer processing.

General

It is much easier to design wideband power amplifiers with FETs than bipolar transistors mainly due to their higher input impedances at least up to VHF. Their input impedance also varies less with frequency than that of bipolar devices and changes in the output load line are reflected back to the input to a lesser degree because of the much lower value of feedback capacitance (collector to base vs drain to gate). Practically all RF power FETs on the market today are of the enhancement MOS type, meaning that positive voltage at the gate in respect to the source is required to turn the device on.

The 1-kW amplifier described here would be difficult, if not impossible, to design to cover four and a half octaves with comparable performance using inexpensive bipolar transistors. In addition, a series of power splitters and combiners would be required to reach high power levels. Biasing to class AB linear operation is also much simpler with FETs since the gate does not draw any dc current, whereas a current equal to I_C(peak)/h_{FE} must be supplied to the base of a bipolar device. One example of this and the splitter-combiner complexity is presented in the Application Note AN-758 by Motorola, Inc.

This article features a state of the art extremely compact design using a pair of FETs rated for 600 W of power output each. It would be capable of a power output of 1.2 kW as a push-pull circuit, but with the output matching employed, which is optimized at around 800 W, the unit starts saturating at around 1 kW at a 50-V dc supply, resulting in high IM distortion. Similarly at a 40-V supply, it would be usable up to 800 W. The type output matching transformer employed allows only integers as 1:4, 1:9, 1:16, etc. The 1:16 impedance ratio transformer would make the output matching optimized at 1400 W, which would result in a poor efficiency at 1200 W and lower power levels. The only way to compensate for this would be to adjust the supply voltage accordingly, in this case 45-46 V. However, the 1:16 ratio transformer of this type is physically much more difficult to fabricate than the lower ratio ones, and may not be available in the commercial market.

The Bias Regulator

The gate bias regulator (IC1 in Fig 1) allows the main supply voltage to be varied or the use of an unregulated supply while keeping the gate bias voltages and the FET idle currents constant. Since the maximum operating voltage of the regulator IC is only 40 V, a Zener diode (D1) is employed to keep it at a safe level. The regulator supply terminals are separated from the main power supply permitting the use of a separate bias supply if desired. There is also an option for a thermistor connection to stabilize the idle currents against temperature changes. The thermistor should be in a physical contact preferably with a mounting flange of one of the FETs. The gate voltages are individually adjustable (R1, R2) making gate threshold voltage matching of the devices unnecessary. In case of a device failure, such as a drain-gate short, D2 and D3 block the full supply voltage from being fed back to destroy the regulator. R10, R11 and C3, C4 are merely AC filters to protect the regulator from possibly strong RF fields. To set the idle currents, R1 and R2 must be adjusted to minimum. R3 is then adjusted for a regulator output voltage of about double the FET gate threshold voltages (IC1, pin 3). The current is monitored at the main supply voltage point while adjusting R1 for a desired idle current, typically 800 mA-1.0 A. R2 is then advanced until the current is doubled, resulting in equal idle currents for both devices. After this procedure, the settings of R1 through R3 should remain until one or both FETs must be replaced.

The RF Path

The amplifier is designed to operate into the industry standard 50-ohm input and output interface. The impedance matching to the low impedance levels of the FETs is accomplished with broadband RF transformers. Both the input transformer (T1) and the output transformer (T2) are of the so-called conventional type in contrast to transmission line transformers.[1,3,4] Both employ only one turn in the low impedance winding. T2 is far more critical than T1 because it determines the efficiency and the high frequency end gain characteristics, plus it must be able to handle a large amount of RF power. For increased bandwidth characteristics, its low impedance, one turn winding consists of three paralleled 10-ohm coaxial cables, resulting in a tight and controllable coupling between the primary and secondary. According to formulas given in Reference 2, approximately twice the present $4.7\ cm^2$ ferrite cross sectional area would be required in order for the core not to saturate with the calculated 127 gauss flux density. The saturation mainly occurs at the lowest frequencies, in this case at 2-3 MHz. Unfortunately most ferrite manufacturers do not give information on saturation flux densities that applies to applications such as this. However, it is known that high permeability ferrites, in

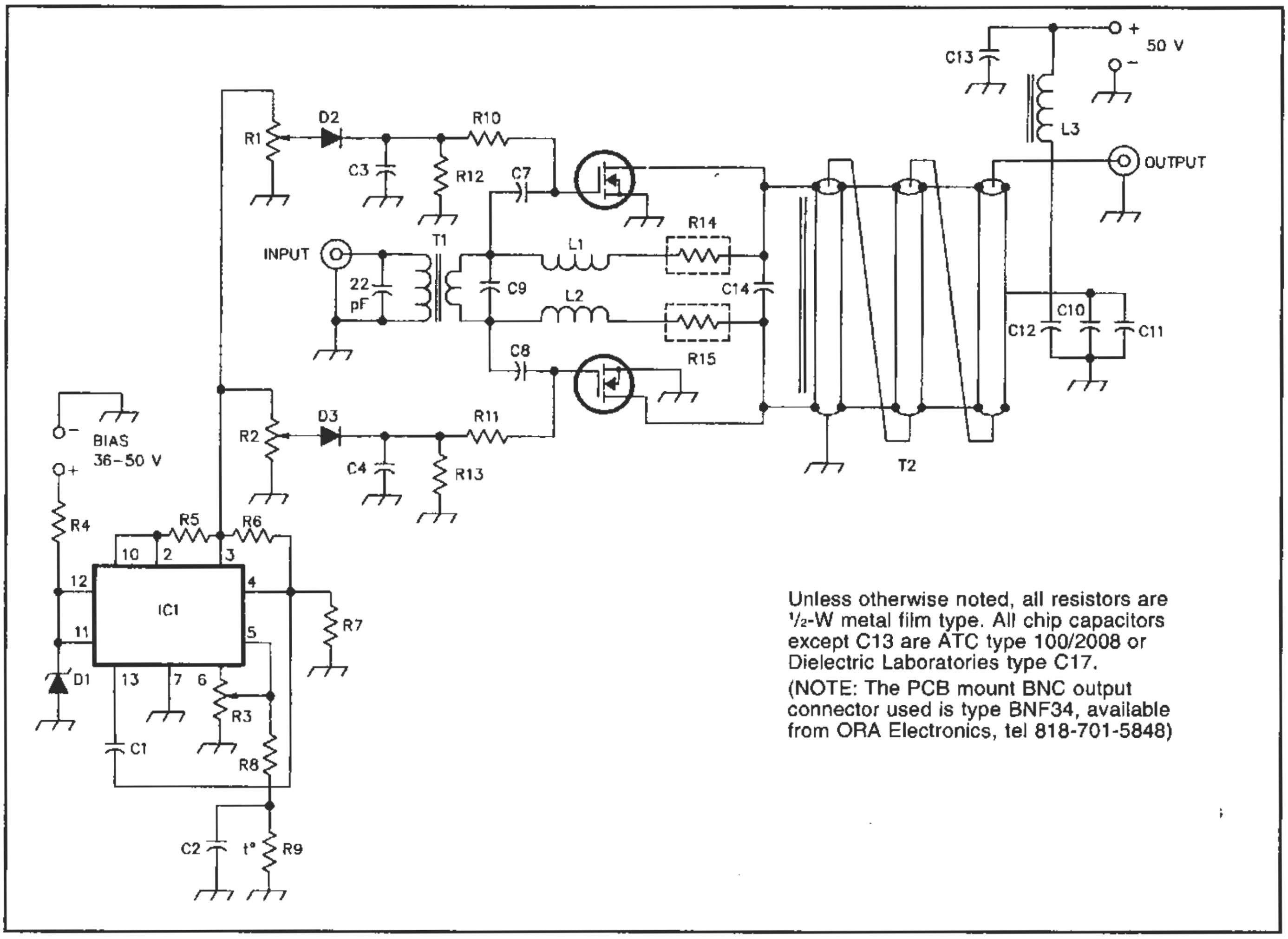

Unless otherwise noted, all resistors are ½-W metal film type. All chip capacitors except C13 are ATC type 100/2008 or Dielectric Laboratories type C17.

(NOTE: The PCB mount BNC output connector used is type BNF34, available from ORA Electronics, tel 818-701-5848)

Fig 1—Circuit Diagram—2 to 50 MHz Amplifier 2-50 MHz Amplifier Components List

R1,R2—1 kΩ single-turn Trimpots
R3—10 kΩ single-turn Trimpot
R4—470 Ω, 2 watts
R5—10 Ω
R6,R12,R13—2 kΩ
R7—10 Ω
R8—Exact value depends on thermistor R9 used (typically 5-10 kΩ)
R9—Thermistor, Keystone RL1 009-5820-97-Di or equivalent
R10,R11—100 Ω, 1 W carbon
R14,R15—EMC Technology model 5308 or KDI Pyrofilm PPR 870-150-3 power resistors, 25 Ω
D1—1 N5357A or equivalent
D2,D3—1N4148 or equivalent
IC1 —MC1723 (723) voltage regulator
C1—1000-pF ceramic disc capacitor
C2,C3,C4—0.1-µF ceramic disc capacitor
C5—0.01-µF ceramic chip capacitor
C6,C12—0.1-µF ceramic chip capacitor
C7,C8—Two 2200-pF ceramic chip capacitors In parallel each
C9—820-pF ceramic chip capacitor
C10,C11—1000-pF ceramic chip capacitor
C13—0.47-µF ceramic chip capacitor or two smaller values in parallel
C14—Unencapsulated mica, 500 V. Two 1000-pF units is series, mounted under T2.
L1,L2—15 ηH, connecting wires to R14 and R15, 1.5 cm each #20 AWG
L3—10 µH, 10 turns #12 AWG enameled wire on Fair-Rite Products Corp ferrite toroid #5961000401 or equivalent
T1,T2—9:1 and 1:9 impedance ration RF transformers, types RF800-3 and RF2067-3 R, respectively (RF Power Systems, 3038 E Corrine Dr, Phoenix, AZ 85032)

general, saturate easier than low permeability materials. Thus, the lowest permeability material should be selected that will satisfy the minimum inductive reactance requirement at the lowest frequency of operation. The formula to calculate this is $NX_L = 2R_{S(L)}$, where: X_L=inductive reactance for one turn, N = number of turns, $R_{S(L)}$ = source or load impedance. Low permeability material is also less lossy at high frequencies, resulting in less heat generated in the transformer. T1, which must handle only 8-12 W of power, is made of higher permeability ferrite. This makes it possible to make the unit physically small as well. In Ti, the secondary consists of metal tubes (see Ref 1), where three turns of the primary wire is threaded through. Metal tubes are also used in T2, but only to hold the structure mechanically together.

At high-power levels generated with solid-state devices, which operate at relatively low voltages, the impedance levels automatically become low. This creates a problem for finding passive components, especially capacitors to handle the high RF currents involved. In vacuum tube circuits a similar problem exists, but in the form of high voltages. In this design, C14 gets the roughest treatment. It must be able to carry RF currents in excess of 10 amperes at the higher frequencies, although the voltage across it is only 75 V rms. At first, several good quality ceramic chip capacitors were tried in parallel, but temperature excursions caused them to crack resulting in AF arcs that burned the circuit board in the area as well. Finally, two unencapsulated mica capacitors (brand names such as Unelco, Underwood, Standex, Elmenco and Semco) were soldered in series by attaching the terminal tabs together, making it a sym-

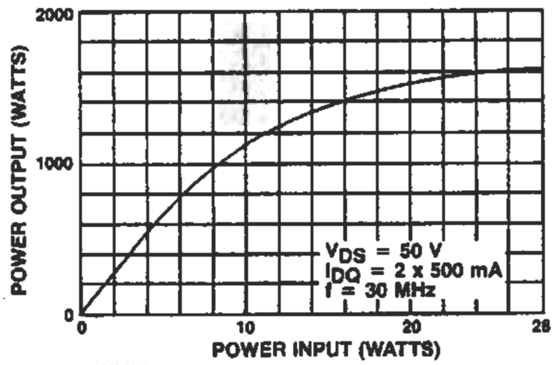

FIGURE 2 — POWER OUTPUT VERSUS POWER INPUT

Fig 2

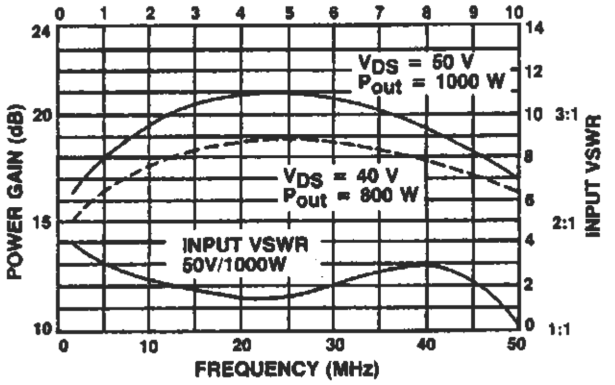

FIGURE 3 — POWER GAIN AND INPUT VSWR VERSUS FREQUENCY

Fig 3

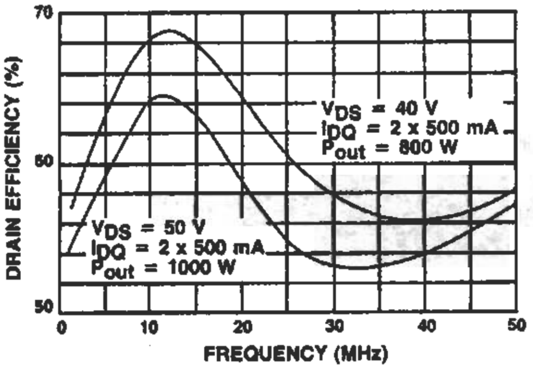

FIGURE 4 — DRAIN EFFICIENCY VERSUS FREQUENCY

Fig 4

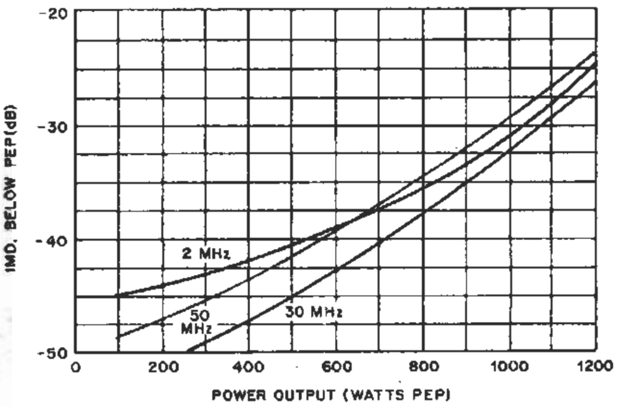

Fig 5

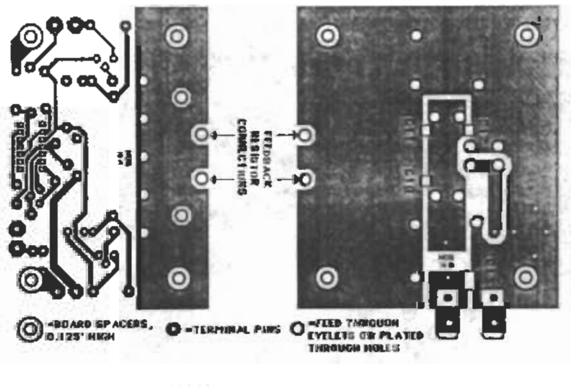
FEEDBACK RESISTOR CONNECTIONS
=BOARD SPACERS, 0.125" HIGH
=TERMINAL PINS
=FEED THROUGH EYELETS OR PLATED THROUGH HOLES

COMPONENT LAYOUT — BOTTOM SIDE

Fig 6

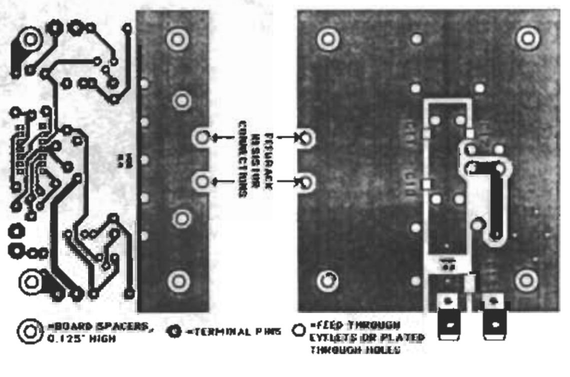
FEEDBACK RESISTOR CONNECTIONS
=BOARD SPACERS, 0.125" HIGH
=TERMINAL PINS
=FEED THROUGH EYELETS OR PLATED THROUGH HOLES

COMPONENT LAYOUT — BOTTOM SIDE

Fig 7

metrical structure. Since each is double the total value required and with double the number of plates, this increases the RF current carrying capability and provides a larger area to be soldered to the board metal foil to make the cooling more efficient. The low impedance winding terminals are then soldered to the tops of the capacitor metal casings, leaving the effective capacitance across the winding. For further fine tuning, an Arco (Elmenco) #469 or Sprague #GM-40900 compression mica trimmer can be soldered to the fronttop terminals of the transformer. Slot openings in the metal foil (Fig 7) located on each side of the output transformer, next to the drain terminals, were provided to increase the series inductance for certain highfrequency narrowband applications. This tunes out some of the FET output capacitance, resulting in increased efficiency. At lower frequencies (below 80 MHz) however, they only add to the IR loss and should be shorted. The location of C9 is also critical and should be placed approximately as shown in Fig 7. This will affect the input VSWR at frequencies above 30 MHz.

Bypass capacitors C10 through C12 must also be of good quality. The center tap of T2 should be free of AF if the circuit is balanced. This may not always be the case, in which case these capacitors will aid this function. L3 and C13 form an additional filter, ensuring that no RF energy is being fed back to the power supply. Switchmode power supplies especially are sensitive against RF and may actually get damaged from it.

Negative feedback is provided through the networks L1-R14 and L2-R15. Its purpose is to produce a relatively flat power gain versus frequency response. It also

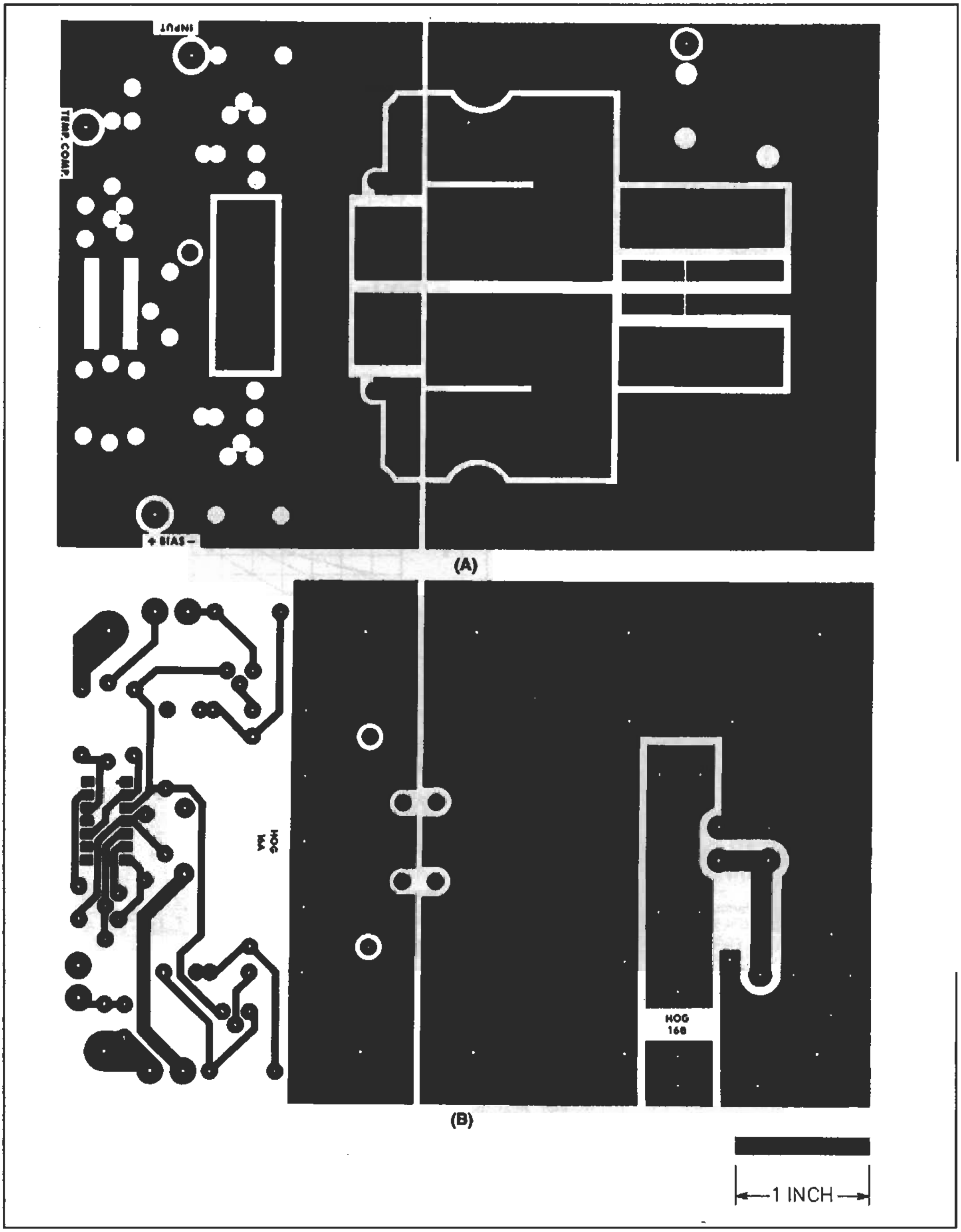
INPUT
TEMP. COMP.
+ BIAS −
(A)
HOG 16A
HOG 16B
(B)
1 INCH

Figs 8A and 8B–Circuit Board–Top Side and Bottom Side

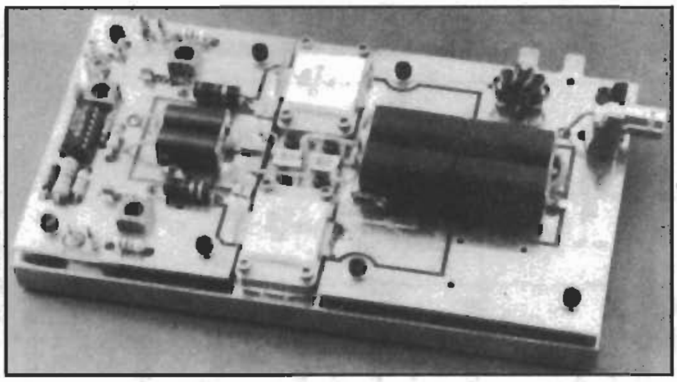

Fig 9—Amplifier mounted to the Heat Spreader.

improves the input return loss and helps to stabilize the amplifier at low frequencies, where the power gain would be 25-30 dB without it. The feedback is at its minimum at the high frequency end and at maximum at low frequencies, where most power is dissipated in R14 and R15. This power is roughly the difference in power input without the feedback between 2 and 50 MHz assuming a constant power output (in this case 25-30 W). A simple formula for calculating the feedback resistor values as well as their dissipation ratings is given in Reference 5. Reference 5 also includes information on physical construction of RF transformers such as used here.

Thermal Aspects

Assuming a 50% worst case efficiency for the unit, each FET dissipates 500 W of heat in an area of 1 x 1.5 inch. It is imperative that the transistors are mounted on the surface of a material with low thermal resistance such as copper. This is called a heat spreader as it is then attached to a heat sink made of material with poorer thermal resistance. It should extend about one inch beyond the edges of the FET mounting flanges at least on three sides. It is even more practical to make the heat spreader as large or larger than the amplifier itself. This would allow all circuit-board spacers to be an equal height of 0.125 inch. The thickness of the heat spreader should be a minimum of 0.375 inch. The heat spreader is then separately attached to the actual heat sink, which can be a 12-inch length of Wakefield Engineering type 4559 extrusion or equivalent[6]. Heat sink compound must be applied to all thermal interfaces and the recommended transistor mounting procedure should be followed, including the screw torque. Fig 9 shows the amplifier mounted to the heat spreader. Although the heat sink is not shown, one must be used for continuous operation and for test periods longer than a couple of minutes. For continuous operation, two 5-inch muffin fans under the heat sink will suffice. They will keep the device case temperature at below 80°C, and the die temperature, which equals to device thermal resistance x power dissipation + case temperature = 0.13 x 500 + 80, at less than 145°C, which is well below the 200-degree maximum recommended value. We must realize that the 500-watt dissipation is only valid when the unit is operated into a 50-ohm load. Under mismatched conditions, depending on the phase angle, the dissipated power may be lower or higher than this value.

Performance

Some of the amplifier performance characteristics are shown in Figs 2 through 5. Although at 30 MHz and above all harmonics are 25 dB or more below the fundamental, an output filter is required to comply with FCC regulations. However, it can be a simpler one than required for the low frequencies, where the third harmonic may be only attenuated 12-15 dB. In push-pull amplifiers, the even harmonics are not usually a problem since they are attenuated by the balanced operation of the circuit. Information on high power low-pass filters for applications as this can be found in Reference 7. These filters are automatically relay switched with BCD code available in most modern transceivers.

References

The circuit boards and other components for this design are available from Communication Concepts, Inc, 508 Millstone Drive, Xenia, OH 45385, tel 513-429-3811/220-9677.

[1] Motorola, Inc, Semiconductor Sector Application Notes AN-749 and AN-i 035.

[2] Hilbers, A.H., "Design of HF Wideband Power Transformers," Amperex (Philips) Application Laboratory Report EC06907 and ECO7213.

[3] Blocksome, Aoderick K., "Practical Wideband RF Power Transformers, Combiners and Splitters," *Proceedings of RF Expo*, February 1986.

By Mike Kossor, WA2EBY

From *QST*, March 1999

A Broadband HF Amplifier Using Low-Cost Power MOSFETs

Part 1—With only 1 W of drive, you'll get over 40 W out—from 160 through 10 meters!

MOSFET
POWER
80M 40M 30M 20M-17M
160M
15M-10M
BAND
WA2EBY HF AMPLIFIER

Many articles have been written encouraging experimenters to use power MOSFETs to build HF RF amplifiers. That's because power MOSFETs—popular in the design of switching power supplies—cost as little as $1 each, whereas RF MOSFET prices start at about $35 each!

Over the years, I tucked away several of these articles, waiting for an opportunity to experiment with them. That opportunity came when I received a call from Al, W2OBJ. Al wanted a low-cost linear amplifier to use with his 5 W QRP transmitter when band conditions got poor. Ideally, the amplifier would generate at least 25 W on all the HF bands. Al's inquiry renewed my interest in the topic and provided the motivation I needed to get my project underway.

Al provided me with an extensive list of RF-amplifier construction articles that use power MOSFETs.[1-8] These articles provided useful information about MOSFETs and general guidelines for working with them, including biasing, parasitic-oscillation suppression, broadband impedance-matching techniques and typical amplifier performance data. It was clear from the performance data that Al's desire to get 25 W output from power MOSFETs on 1.8 to 30 MHz was going to be a challenge! The RF output power of most of the amplifiers described in the articles drops off to 10 W or less as frequency increases just to 14 MHz.

An Idea Brews

After hundreds of hours of experimentation, I came up with a design that exceeds our original objective: One watt of input power produces over 40 W of output (after harmonic filtering) from 160 through 10 meters. To the basic amplifier, I added an RF-sensed TR relay and a set of low-pass filters designed to suppress harmonic out-

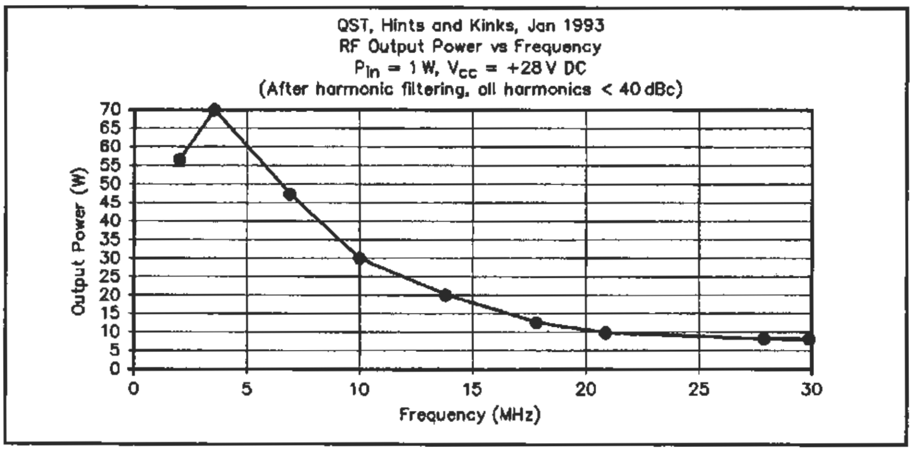

Figure 1—Jim Wyckoff, AA3X, "1 W In, 30 W out With Power MOSFETs at 80 M," Hints and Kinks, *QST*, Jan 1993, pp 50-51.

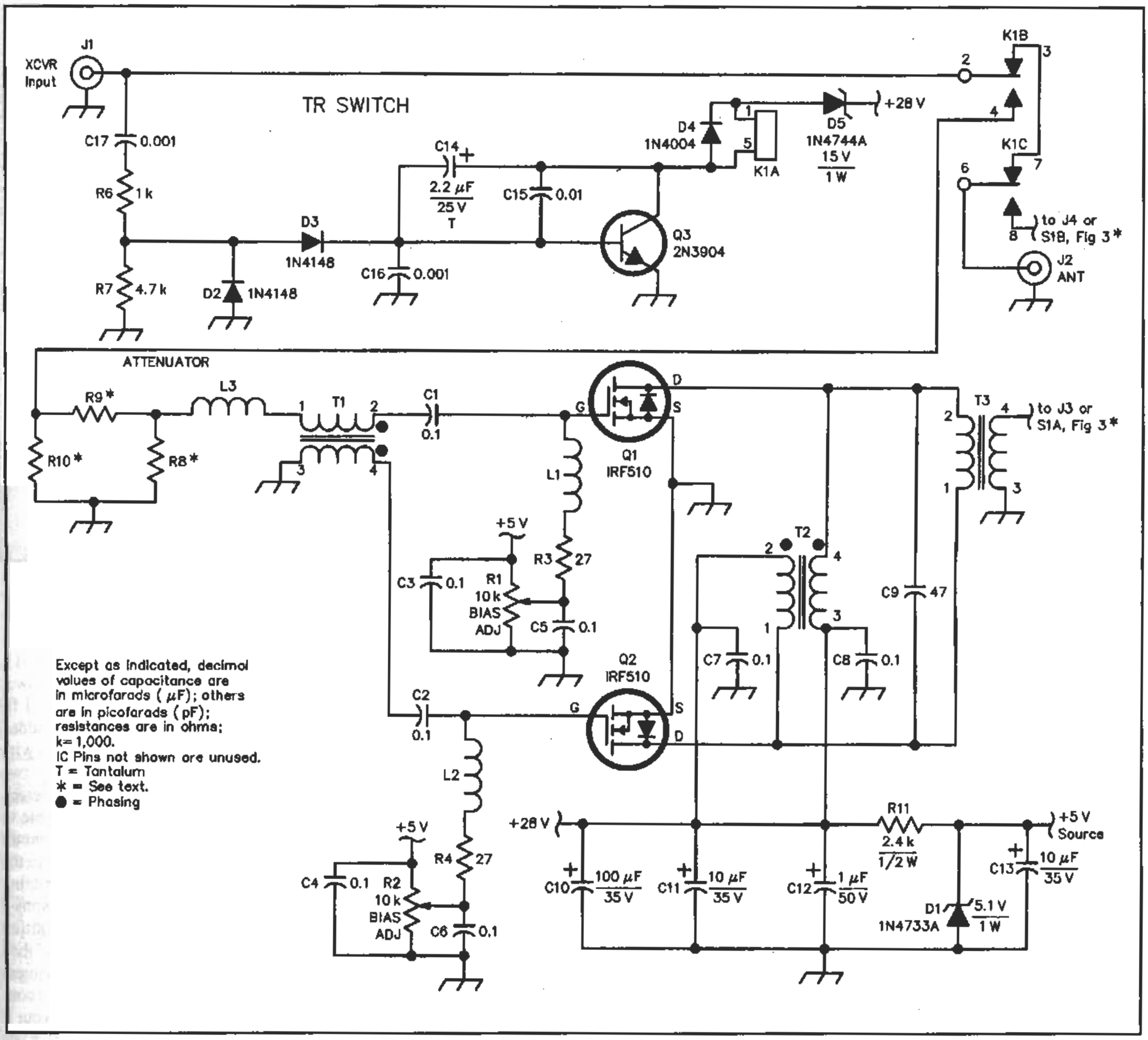

J1
XCVR Input
TR SWITCH
C17 0.001
R6 1 k
R7 4.7 k
D2 1N4148
D3 1N4148
C14 2.2 µF 25 V T
C15 0.01
C16 0.001
Q3 2N3904
D4 1N4004
K1A
D5 1N4744A 15 V 1 W
+28 V
K1B
K1C
to J4 or S1B, Fig 3*
J2 ANT
ATTENUATOR
R9*
R10*
R8*
L3
T1
C1 0.1
L1
Q1 IRF510
G
D
S
+5 V
R3 27
C3 0.1
R1 10 k BIAS ADJ
C5 0.1
Q2 IRF510
C2 0.1
L2
R4 27
C4 0.1
R2 10 k BIAS ADJ
C6 0.1
T2
C7 0.1
C8 0.1
C9 47
T3
to J3 or S1A, Fig 3*
+28 V
R11 2.4 k 1/2 W
+5 V Source
C10 100 µF 35 V
C11 10 µF 35 V
C12 1 µF 50 V
C13 10 µF 35 V
D1 1N4733A 5.1 V 1 W

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k= 1,000.
IC Pins not shown are unused.
T = Tantalum
* = See text.
● = Phasing

Figure 2—Schematic of the MOSFET all-band HF amplifier. Unless otherwise specified, resistors are 1/4 W, 5% tolerance carbon-composition or film units. Equivalent parts can be substituted. Part numbers in parentheses are Mouser (Mouser Electronics, 968 N Main St, Mansfield, TX 76063; tel 800-346-6873, 817-483-4422, fax 817-483-0931; **sales@ mouser.com**; **http://www.mouser.com**); see Note 9.

C1-C8—0.1 µF chip (140-CC502Z104M)
C9—47 pF chip (140-CC502N470J)
C10—100 µF, 35 V (140-HTRL35V100)
C11, C13—15 µF, 35 V (140MLR35V10)
C12—1 µF, 50 V (140-MLRL50V1.0)
C14—2.2 µF, 35 V tantalum (581-2.2M35V)
C15—0.01 µF chip (140-CC502B103K)
C16, C17—0.001 µF chip (140-CC502B102K)
D1—1N4733A, 5.1 V, 1 W Zener diode (583-1N4733A)
D4—1N4004A(583-1N4004A)
D2, D3—1N4148 (583-1N4148)
D5—1N4744A, 15 V, 1 W Zener diode (583-1N4744A)
J1, J2—SO-239 UHF connector (523-81-120)
K1—12 V DPDT, 960 Ω coil, 12.5 mA (431-OVR-SH-212L)
L1, L2—9½ turns #24 enameled wire, closely wound 0.25-in. ID
L3—3½ turns #24 enameled wire, closely wound 0.190-in. ID
Q1, Q2—IRF510 power MOSFET (570-IRF510)
Q3—2N3904 (610-2N3904)
R1, R2—10 kΩ trim pot (323-5000-10K)
R3, R4—27 Ω, ½ W (293-27)
R6—1 kΩ chip (263-1K)
R7— 4.7 kΩ chip (263-4.7K)
R8—130 Ω, 1 W (281-130); for 7 dB pad (5 W in, 1 W out)
R9—43 Ω, 2 W (282-43); for 7 dB pad (5 W in, 1 W out)
R10—130 Ω, 3 W (283-130); for 7 dB pad (5 W in, 1 W out)
R8, R10—300 Ω, ½ W (273-300); for 3 dB pad (2 W in, 1 W out)
R9—18 Ω, 1 W (281-18); for 3 dB pad (2 W in, 1 W out)
R11—2.4 kΩ, ½ W (293-2.4K)
T1—10 bifilar turns #24 enameled wire on an FT-50-43 core.
T2—10 bifilar turns #22 enameled wire on two stacked FT-50-43 cores.
T3—Pri 2 turns, sec 3 turns #20 Teflon-covered wire on BN-43-3312 balun core.
Misc: Aluminum enclosure 3.5∞8∞6 inches (HWD) (537-TF-783), two TO-220 mounting kits (534-4724), heat-sink compound (577-1977), amplifier PC board (see Note 9), heat sink (AAVID [Mouser 532-244609B02]; see text), about two feet of RG-58 coax, #24 enameled wire and #20 Teflon-insulated wire.

put and comply with FCC requirements. The amplifier is built on double-sided PC board and requires *no tuning*. Another PC board contains the low-pass filters. Power-supply requirements are 28 V dc at 5 A, although the amplifier performs well at 13.8 V dc.

Several of these amplifiers have been built and exhibit similar performance. Al has been using his amplifier on each of the HF bands, logging well over 500 contacts in 18 months. Signal reports indicate a noticeable improvement in readability (about two S units on average) over his 5 W rig. No indications of in-stability, CW key clicks or distortion on SSB have been reported. To make it easy for you to duplicate this project, PC boards and parts kits are available, all at a cost of about $100![9]

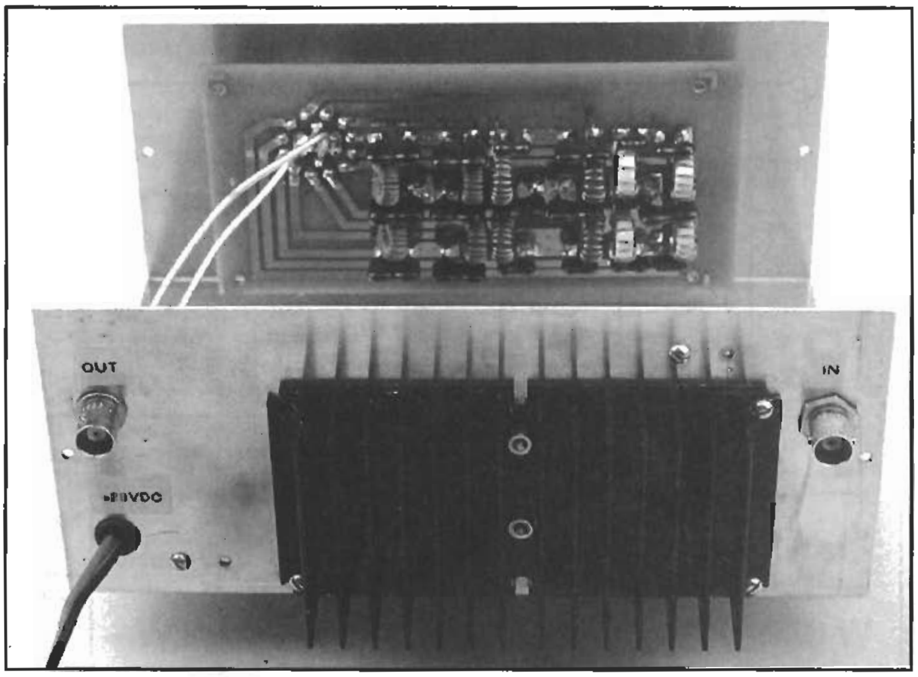

A rear panel view showing the heat sink.

An Overview of MOSFETs

MOSFETs operate very differently from bipolar transistors. MOSFETs are voltage-controlled devices and exhibit a very high input impedance at dc, whereas bipolar transistors are current-controlled devices and have a relatively low input impedance. Biasing a MOSFET for linear operation only requires applying a fixed voltage to its gate via a resistor. With MOSFETs, no special bias or feedback circuitry is required to maintain the bias point over temperature as is required with bipolar transistors to prevent thermal runaway.[10] With MOSFETs, the gate-threshold voltage increases with increased drain current. This works to turn off the device, especially at elevated temperatures as transconductance decreases and $R_{DS(on)}$ (static drain-to-source *on* resistance) increases. These built-in self-regulating actions prevent MOSFETs from being affected by thermal runaway. MOSFETs do not require negative feedback to suppress low-frequency gain as is often required with bipolar RF transistors. Bipolar transistor gain increases as frequency decreases. Very high gain at dc and low frequencies can cause unwanted, low-frequency oscillation to occur in bipolar transistor RF amplifiers unless negative feedback is employed to prevent it. Low-frequency oscillation can damage bipolar transistors by causing excess power dissipation, leading to thermal runaway.

MOSFET Limitations

Of course, MOSFETs do have their limitations. The high gate impedance and the device structure make them susceptible to electrostatic discharge (ESD) damage. Some easily applied precautions prevent this: Use a soldering iron with grounded tip; use a wrist strap connected to ground through a 1 MΩ resistor to bleed off excess body charge while handling MOSFETs and do all work on an anti-static mat connected to ground via a 1 MΩ resistor.

The sensitivity of a MOSFET's gate to static and high-voltage spikes also makes it vulnerable to damage resulting from parasitic oscillation. This undesired self-oscillation could result in excessive gate-to-source voltage that permanently damages the MOSFET's gate insulation. Another MOSFET limitation is gate capacitance. This parameter limits the frequency at which a MOSFET can operate effectively as an RF amplifier. I recommend reviewing the referents of Notes 1-3 if you are interested in more detailed information about MOSFETs.

Power MOSFET RF Amplifiers

Of the several power MOSFET amplifiers I built to check their performance, the one providing the best performance is the push-pull design described by Jim Wyckoff, AA3X, in *QST* (see Note 3). I used IRF510 power MOSFETs rather than the IRF511s specified. The performance of this power MOSFET amplifier design is summarized in Figure 1; its basic design is very similar to another amplifier described in the referent of Note 4, written 10 years earlier. That amplifier uses a pair of more-expensive MRF138 MOSFETs designed specifically for RF applications.

As Figure 1 shows, the Hints and Kinks amplifier performance is excellent from 1.8 MHz to 7 MHz and far exceeds the published figure of 30 W output on 3.5 MHz. As frequency increases above 10 MHz, however, output drops off rapidly, falling below 10 W above 21 MHz. (These levels were measured after harmonic filtering.)

Although the amplifier is identified as stable, my first attempt at duplicating the amplifier resulted in oscillations that destroyed one of the IRF510s. I was puzzled by this. At first, I thought the problem was caused by my substitution of the slightly more robust IRF510 MOSFETs for the called-for IRF511s. That idea proved wrong when my second attempt to power up the amplifier with IRF511 MOSFETs installed also resulted in a blown IRF511. (Thank goodness these are $1 power MOSFETs, not $35 RF MOSFETs!). I finally achieved good stability when I added a small amount of inductance in series with the MOSFET source to ground (just two turns of #24 wire, 0.125 inch diameter). With this added inductance, I was able to remove the ferrite beads from the circuit without any sign of instability. I believe the substitution of the IRF510 and minimizing source lead inductance are the reasons I obtained significantly higher RF output power and wider bandwidth than described in the referent of Note 3. This experiment underscores the need to observe *exact* construction techniques and physical layout if similar performance is to be expected. Even though I used PC board construction, I got significantly different results because my layout was not the same as the author's.

Modifying the Design

Although the amplifier performed better than expected, its bandwidth was significantly less than desired. Considerable experimentation (and I do mean considerable!) resulted in the circuit shown in Figure 2. This amplifier consists of two power MOSFETs operating in push-pull and employs an RF-sensed TR relay.

During receive, TR relay K1 is deenergized. Signals from the antenna are connected to J2 and routed through K1 to a transceiver connected to J1. (This path loss is less than 0.3 dB from 1.8 MHz through 30 MHz.) In transmit, RF voltage from the transceiver is sampled by C17 and divided by R6 and R7. D2 and D3 rectify the RF voltage and charge C16. Q3 begins conducting when the detected RF voltage across C16 reaches approximately 0.7 V. This energizes K1, which then routes the transmitted RF signal from J1

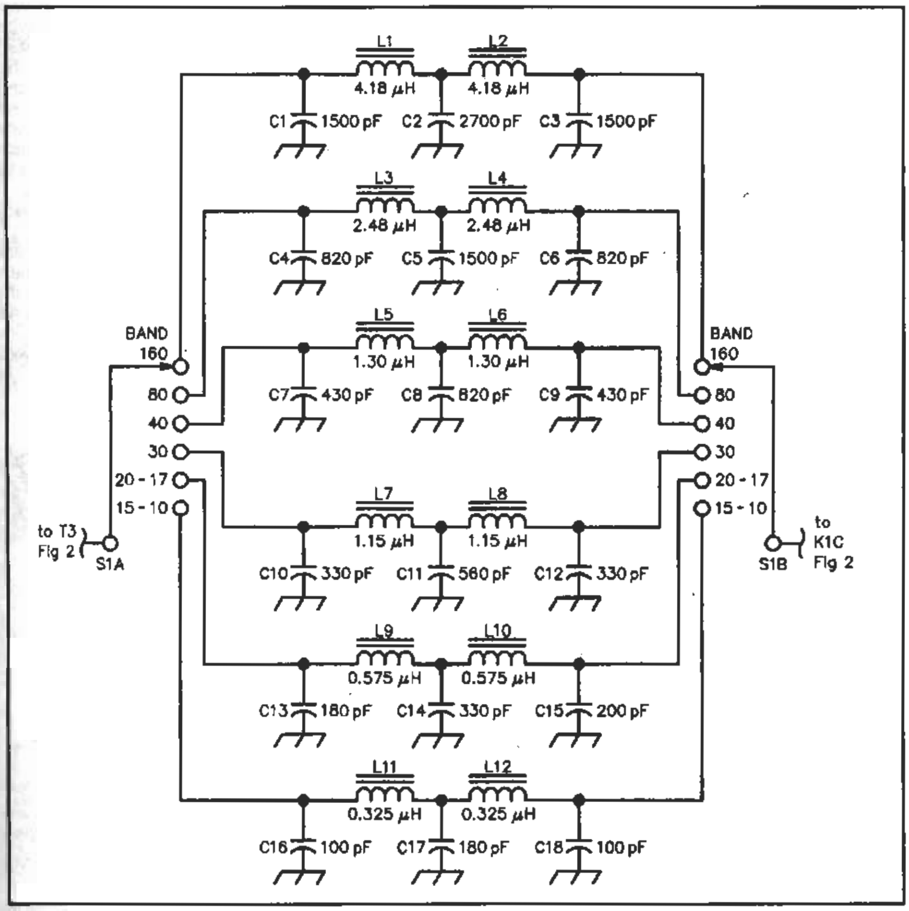

Figure 3—Low-pass filter schematic. In some cases, the actual filter component values differ from the calculated values of a standard 50 Ω-input filter. Such differences improve the impedance matching between the amplifier and the load. Capacitors are all dipped mica units.

C1, C3, C5—1500 pF (5982-19-500V1500)
C2—2700 pF (5982-19-500V2700)
C4, C6, C8—820 pF (5982-19-500V820)
C7, C9—430 pF (5982-15-500V430)
C10, C12, C14—330 pF (5982-19-500V330)
C11—560 pF (5982-19-500V560)
C13, C17—180 pF (5982-15-500V180)
C15—200 pF (5982-15-500V200)
C16, C18—100 pF (5982-10-500V100)
S1—2 pole, 6 position rotary (10YX026)
Misc: low-pass filter PC board (see Note 9)

to the input of the amplifier and sends the output of the amplifier to the antenna at J2. RF-sensed relay response is very fast. No noticeable clipping of the first CW character has been reported.

I made provisions to include an RF attenuator (consisting of R8, R9 and R10) to enable adjusting the amplifier input power to 1 W. (The parts list contains resistor values to reduce the output of 2 or 5 W drivers to 1 W.) The 1 W signal is then applied to the primary of T1 via an input impedance-matching network consisting of L3. T1 is a 1:1 balun that splits the RF signal into two outputs 180 degrees out of phase. One of these signals is applied by C1 to Q1's gate. The other signal is routed via C2 to Q2's gate. The drains of Q1 and Q2 are connected to the primary of output transformer T3, where the two signals are recombined in phase to produce a single output. T3 also provides impedance transformation from the low output impedance of the MOSFETs to the 50 Ω antenna port. Dc power is provided to the drains of Q1 and Q2 by phase-reversal choke, T2. This is a very effective method to provide power to Q1 and Q2 while presenting a high impedance to the RF signal over a broad range of frequencies. The drain chokes for Q1 and Q2 are wound on the same core, and the phase of one of the chokes (see the phasing-dot markings on T2) is reversed. C9 increases the bandwidth of impedance transformation provided by T3, especially at 21 MHz.

The 5 V bias supply voltage is derived from 28 V by Zener diode D1 and current-limiting resistor R11. Bypass capacitors C3, C4, C5, C6 and C13 remove RF voltages from the bias supply voltage. Gate bias for Q1 and Q2 is controlled independently. R1 adjusts Q1's gate-bias voltage via R3 and L1. R2 works similarly for Q2 via R4 and L2.

At low frequencies, the amplifier's input impedance is essentially equal to the series value of R3 and R4. L1 and L2 improve the input-impedance match at higher frequencies. The low value of series resistance provided by R3 and R4 also reduces the Q of impedance-matching inductors L1 and L2, which improves stability. Dc blocking capacitors C1 and C2 prevent loading the gate bias-supply voltage.

C14 keeps transistor Q3 conducting and K1 energized between SSB voice syllables or CW elements. Without C14, K1 would chatter in response to the SSB modulation envelope and fast keying. Increasing the value of C14 increases the time K1 remains energized during transmit. The reverse voltage generated by K1 when the relay is deenergized is clamped to a safe level by D4. D5 drops the 28 V supply to 13 V to power 12 V relay K1. D5 can be replaced with a jumper if K1 has a 28 V dc coil or if you intend to operate the amplifier with a 13.8 V dc supply.

Harmonic Filtering

Although biased for class AB linear operation, this amplifier (like others of its type) exhibits some degree of nonlinearity, resulting in the generation of harmonics. This push-pull amplifier design cancels even-order harmonics (2f, 4f, 6f, etc) in the output transformer, T3. Odd-order harmonics are not canceled. Second-order harmonics generated by the amplifier are typically less than 30 dBc (30 dB below the carrier) whereas third-order harmonics are typically only 10 dBc. FCC regulations require all HF RF-amplifier harmonic output power to be at least 40 dBc at power levels between 50 to 500 W. To meet this requirement, it is common practice for HF amplifiers to use low-pass filters. Separate low-pass filters are needed for the 160, 80, 40 and 30 meter bands. The 20 and 17 meter bands can share the same low-pass filter. So, too, the 15, 12 and 10 meter bands can share a common low-pass filter; see Figure 3.

Switching among the six filters can be a messy wiring problem, especially on the higher-frequency bands where lead lengths should be kept short for optimum performance. This problem is solved by mounting all six low-pass filters on a PC board. A two-pole, six-position rotary switch (S1) mounted directly on the same PC board manages all filter interconnections. One pole of S1 connects the amplifier output to one of the six filter inputs, while S1's other pole simultaneously connects the corresponding filter's output to the TR relay, K1. Only two coaxial-cable connections are required between the RF amplifier and the low-pass filter board.

Next Month

In Part 2, I'll wrap up with amplifier construction and adjustment, and discuss the amplifier's overall performance. See you then!

Notes

[1] Doug DeMaw, W1FB, "Power-FET Switches as RF Amplifiers," *QST*, Apr 1989, pp 30-33. See also Feedback, *QST*, May 1989, p 51.

[2] Wes Hayward, W7ZOI, and Jeff Damm, WA7MLH, "Stable HEXFET RF Power Amplifiers," Technical Correspondence, *QST*, Nov 1989, pp 38-40; also see Feedback, *QST*, Mar 1990, p 41.

[3] Jim Wyckoff, AA3X, "1 Watt In, 30 Watts Out with Power MOSFETs at 80 Meters," Hints and Kinks, *QST*, Jan 1993, pp 50-51.

[4] Doug DeMaw, W1FB, "Go Class B or C with Power MOSFETs," *QST*, March 1983, pp 25-29.

[5] Doug DeMaw, W1FB, "An Experimental VMOS Transmitter", *QST*, May 1979, pp 18-22.

[6] Wes Hayward, W7ZOI, "A VMOS FET Transmitter for 10-Meter CW," *QST*, May 1979, pp 27-30.

[7] Ed Oxner, ex-W9PRZ (SK), "Build a Broadband Ultralinear VMOS Amplifier," *QST*, May 1979, pp 23-26.

[8] Gary Breed, K9AY, "An Easy-to-Build 25-Watt MF/HF Amplifier," *QST*, Feb 1994, pp 31-34.

[9] Parts for this project are available in five modular kits. The following three kits are available from Mouser Electronics (Mouser Electronics, 958 N Main St, Mansfield, TX 76063; tel 800-346-6873, 817-483-4422, fax 817-483-0931; **sales@mouser.com**; **www.mouser.com**): Amplifier components (Mouser P/N 371-HFAMP1) consisting of the amplifier PC board and all PC-board-mounted components (except for the ferrite cores). Price: $35, plus shipping. Amplifier hardware kit (Mouser P/N 371-HFAMP2) consisting of the aluminum enclosure, two UHF connectors, two TO-220 mounting kits, AAVID heat sink and one container of heat sink compound. Price: $30 plus shipping. Low-pass filter kit (Mouser P/N 371-HFAMP3) consisting of the low-pass filter PC board, rotary switch and all PC-board-mounted capacitors (inductor cores are *not* included). Price: $35, plus shipping. Part-placement diagrams accompany the PC boards.

PC boards only are available from Mouser Electronics: HF amplifier board (#371-AMPPWB-2); filter PC board (#371-LPPWB-2). Price $15 each, plus shipping.

The following two kits are available from Amidon Inc (Amidon, Inc, 240 Briggs Ave, Costa Mesa, CA 92626; tel 1-800-898-1883, 714-850-4660, fax 714-850-1163): Amplifier ferrite kit (Amidon P/N HFAFC) containing the ferrite cores, balun core and magnet and Teflon wire to wind the transformers for the HF amplifier. Price: $3.50 plus shipping. Low-pass filter cores kit (Amidon P/N HFFLT) containing all iron cores and wire for the low-pass filters. Price: $4.50 plus ship.

[10] See Motorola Application Reports Q1/95, HB215, *Application Report AR346*.

Thermal runaway is a condition that occurs with bipolar transistors because bipolar transistors conduct more as temperature increases, the increased conduction causes an increase in temperature, which further increases conduction, etc. The cycle repeats until the bipolar transistor overheats and is permanently damaged.

A Broadband HF Amplifier Using Low-Cost Power MOSFETs

Part 2—Let's put the finishing touches on this all-band HF amplifier!

Last month,[11] I covered the history and development of this 40 W (average) amplifier. I'm sure you're anxious to get your amplifier finished and on the air, so let's get going!

Amplifier Construction

The amplifier is constructed on a double-sided PC board with plated through holes to provide top-side ground connections. I used chip resistors and capacitors to simplify construction, but leaded capacitors may work if lead lengths are kept short. First, assemble all chip capacitors and resistors on the PC board. Tweezers help to handle chip components. Work with only one component value at a time (chip caps and resistors are very difficult to identify!). Chip capacitor and resistor mounting is simplified by tinning one side of the PC board trace with solder before positioning the capacitor or resistor. Touch the soldering iron tip to the capacitor or resistor to tack it in place. Finish mounting by soldering the opposite side of the component. *Don't apply too much heat to chip capacitors.* The metalized contacts on the capacitor can be damaged or completely removed if too much heat is applied. Use a 15 to 20 W soldering iron and limit soldering time to five seconds.

Mount axial-leaded resistors, diodes and remaining capacitors next. To avoid damaging them, mount inductors and transformers last. L1 and L2 are wound on a 0.25-inch drill-bit shaft. By wrapping the wire around the shaft 10 times, you'll get 9½ turns. The last turn arcs only a half-turn before entering the PC board. L3 is wound on a 0.190-inch diameter drill bit with 3½ turns wound the same way as L1 and L2. Mounting K1 is simplified by first bending all its leads 90° outward so it lies flat on the PC board. Use a wrist strap connected to ground through a 1 MΩ resistor to bleed off static body charge while handling MOSFETs, and do the work on an anti-static mat connected to ground via a 1 MΩ resistor. The gate input can be damaged by electrostatic discharge!

When winding T3, wind the primary first and add the secondary winding over the primary. Be sure to use Teflon-insulated wire for T3's windings; the high operating temperatures encountered will likely melt standard hook-up wire insulation.

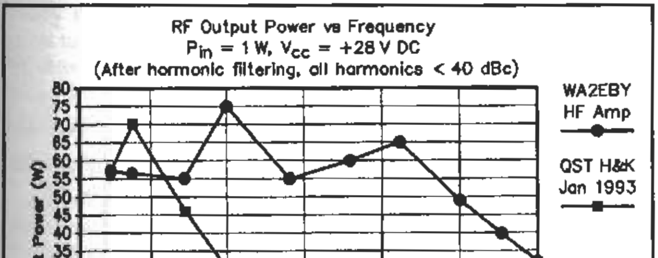

Figure 4—RF output power comparison of the Hint and Kink amplifier and this design.

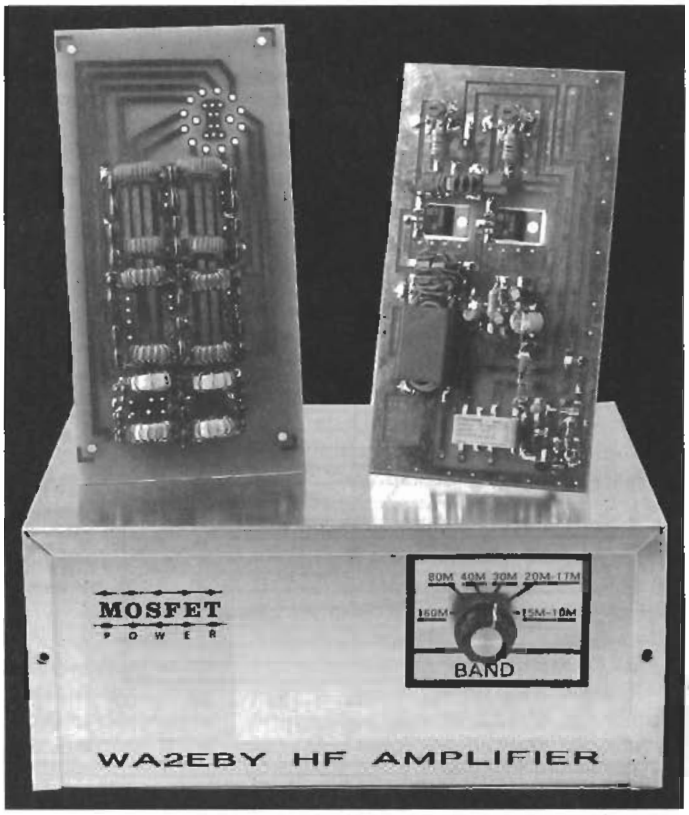

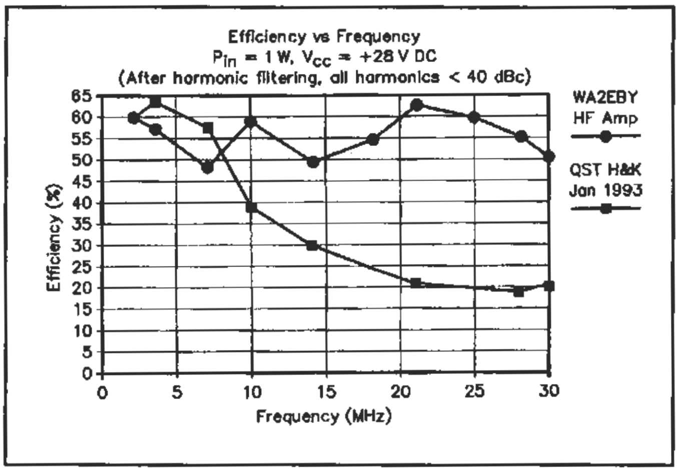

Figure 5—Efficiency comparison of the Hint and Kink amplifier and this one.

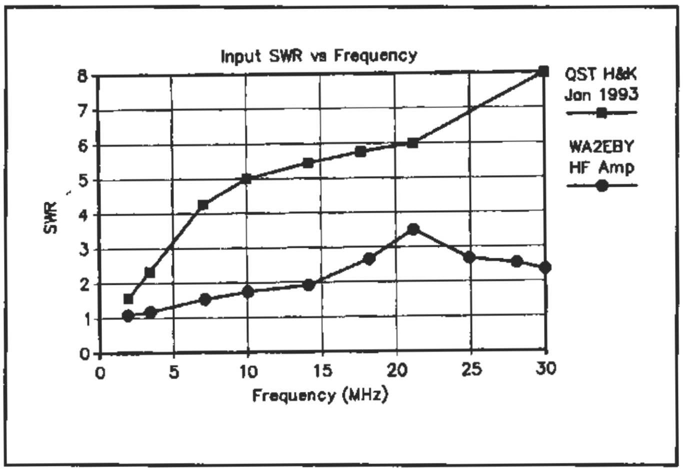

Figure 6—Input SWR comparison of the two amplifiers.

Heat Sinking

Together, Q1 and Q2 dissipate up to 59 W. A suitable heat sink is required to prevent the transistors from overheating and damage. I used an AAVID 244609B02 heat sink originally designed for dc-to-dc power converters. The amplifier PC board and heat sink are attached to an aluminum enclosure by two #4-40 screws drilled through the PC board, enclosure and heat sink at diagonally opposite corners. A rectangular cutout in the enclosure allows Q1 and Q2 direct access to the heat sink. This is essential because of the large thermal impedance associated with the TO-220 package (more on this topic later). Mark the locations of the transistor-tab mounting-hole location in the center of the heat sink in between the cooling fins. Disassemble the heat sink to drill 0.115 inch holes for #4-40 mounting screws, or tap #4-40 mounting holes in the center of the heat-sink fins.

Use mica insulators and grommets when mounting Q1 and Q2 to prevent the #4-40 mounting screws from shorting the TO-220 package drain connections (tabs) to ground. Coat both sides of the mica insulator with a *thin* layer of thermal compound to improve the thermal conduction between the transistor tab and the heat sink. Be sure to install the mica insulator on the heat sink *before* assembling the amplifier PC board to the enclosure and heat sink. The mica insulators are larger than the cut outs in the PC board, making it impossible to install them after the PC board is mounted.

Low-Pass Filter Construction

Inductor winding information for the low-pass filters is provided in Table 1.

Single Band

A PC-board trace is available on the amplifier PC board next to amplifier output (J3) to allow the installation of a single-band low-pass filter between the terminals of J3 and K1's input, J4. This is handy if you intend to use the amplifier on one band only. The input inductor of the low-pass filter connects from J3 to the single PC trace adjacent to J3. The output inductor connects in series between the single PC trace to J4. The three filter capacitors connect from J3, J4 and the PC-board trace near J3 to ground. *This single trace is not used when multiple filters are required.* Remember to remove the single trace adjacent to J3 on the amplifier PC board before attaching the amplifier board between the RF connectors on the enclosure's rear panel.

Multiple-Band Filters

Using the amplifier on more than one band requires a different approach. A set of

Table 1
Low-Pass Filter Inductor Winding Information

(Refer to Figure 3 in Part 1))

Inductor Number	*No. of Turns*	*Core*
L1, L2	30 turns	T-50-2
L3, L4	22 turns	T-50-2
L5, L6	16 turns	T-50-2
L7, L8	14 turns	T-50-2
L9, L10	11 turns	T-50-6
L11, L12	8 turns	T-50-6

Note: All inductors are wound with #22 enameled wire except for L1–L4, which are wound with #24 enameled wire.

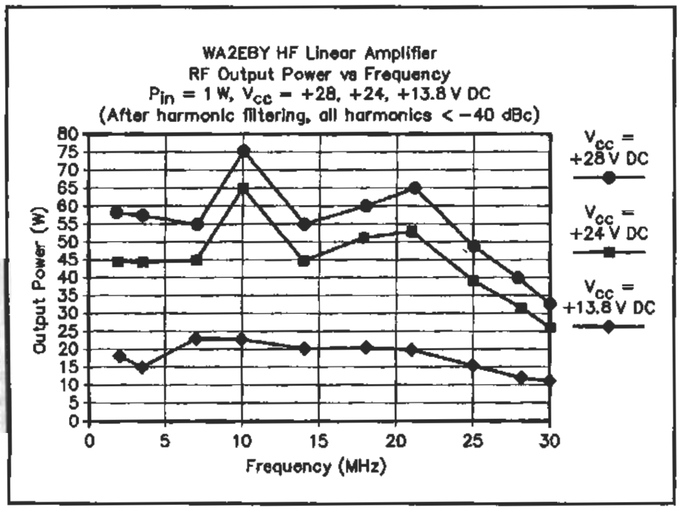

Figure 7—RF output power versus supply-voltage of this amplifier.

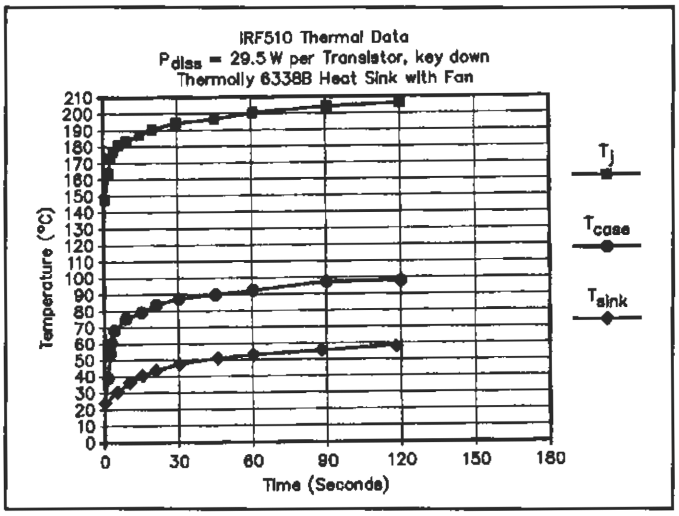

Figure 8—Thermal performance of the amplifier during key-down conditions.

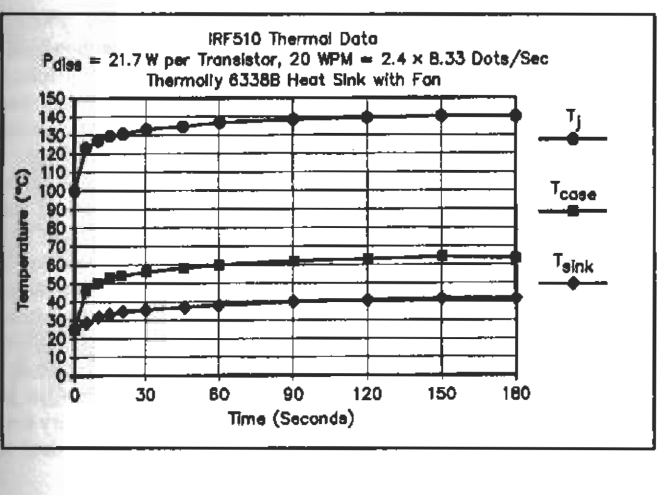

Figure 9—Thermal performance of the amplifier during simulated CW conditions.

six low-pass filters is built on a double-sided PC board with plated through holes to provide top-side ground connections. A PC-board mount, two-pole, six-position rotary switch does all low-pass filter selection. Silver-mica, leaded capacitors are used in all the filters. On 160 through 30 meters, T-50-2 toroids are used in the inductors. T-50-6 toroids are used for inductors on 20 through 10 meters. The number of turns wound on a toroid core are counted on the toroid's OD as the wire passes through the core center (*The ARRL Handbook* [12] provides complete details for winding toroids). Assemble one filter section at a time starting with the 160, 80, 40-meter filter, then the 30-meter filter. With the switch mounting position at your upper left, the filter input (C1) is near the top edge of the board and the filter output (C3) is near the bottom edge. *The last two filters are out of sequence;* the 15-10 meter filter comes *before* the 20-17 meter filter) and the inputs/outputs are reversed to simplify the PC-board layout. The input capacitors, C13 and C16, are mounted on the board *bottom edge*, and output capacitors, C15 and C18, are on the *top edge*.

Use care when assembling the rotary switch. All 14 terminals must fit through the PC board without damaging or bending the pins. Make sure there are no bent pins before you attempt assembly. Insert the rotary switch into the PC board. Do *not* press the rotary switch all the way into the PC-board holes flush with the ground plane! If you do, the top flange of the signal pins may short to the ground plane.

Bias Adjustment

The biasing procedure is straightforward and requires only a multimeter to complete. First, set R1 and R2 fully counterclockwise, (0 V on the gates of Q1 and Q2). Terminate the RF input and outputs with a 50 Ω load. Next, connect the 28 V supply to the amplifier in series with a multimeter set to the 0-200 mA current range. Measure and record the idling current drawn by the 5 V bias supply. The value should be approximately 9.5 mA (28 – 5.1 V) / 2.4 kΩ = 9.5 mA). Set Q1's drain current to 10 mA by adjusting R1 until the 28 V supply current increases by 10 mA above the idling current (9.5 + 10 = 19.5 mA). Next, adjust R2 for a Q2 drain current of 10 mA. This is accomplished by adjusting R2 until the 28 V supply current increases by an additional 10 mA (to 29.5 mA).

Amplifier Performance

With a 28 V power supply and 1 W of drive, the RF output power of this amplifier exceeds 40 W from 1.8 MHz through 28 MHz. Peak performance occurs at 10 MHz, providing about 75 W after filtering! A performance comparison between this amplifier and my modified version of the Hint and Kink amplifier mentioned earlier is shown in Figure 4.

As shown in Figure 5, this amplifier achieves an efficiency of better than 50% over its frequency range, except at 7 MHz where the efficiency drops to 48%. In contrast, the Hint and Kink amplifier delivers greater efficiency between 1.8 and 7 MHz, but it drops rapidly to only 20% as frequency is increased.

Figure 6 compares the input SWR of the two amplifiers. The Hint and Kink amplifier's SWR is acceptable (< 2:1) only at 1.8 MHz. This amplifier is better, however it, too, exceeds 2:1 above 14 MHz. The input SWR of this amplifier can be improved to better than 2:1 on all bands by adding a 3 dB pad (R8-R10 of Figure 2) at the input and supplying 2 W to the pad input. This keeps the amplifier drive at 1 W.

Figure 7 graphs this amplifier's RF output power as a function of drain supply voltage. During this test, the amplifier RF drive level was kept constant at 1 W. As you can see, even when using a 13.8 V dc supply, the amplifier provides over 10 W output (a gain of more than 10 dB) from 1.8 to 30 MHz.

Operation

The amplifier requires no tuning while operating on any HF amateur band. You must, however, *be sure to select the proper low-pass filter prior to transmitting*. If the wrong low-pass filter is selected, damage to the MOSFETs may result. Damage will likely result if you attempt to operate the amplifier on a band with the low-pass filter selected for a lower frequency. For example, driving the amplifier with a 21 MHz signal while the 1.8 MHz low-pass filter is selected will likely destroy Q1 and/or Q2.

The amplifier can also be damaged by overheating. This limitation is imposed by the TO-220 packages in which Q1 and Q2 are housed. The thermal resistance from junction to case is a whopping 3.5°C/W. This huge value makes it virtually impossible to keep the junction temperature from exceeding the +150°C target for good reliability. Consider the following conditions: key down, 1 W input, 53 W output on 7 MHz (worst-case band for efficiency). The amplifier consumes 28 V × 4 A = 112 W, of which 53 W are sent to the antenna, so 59 W (112 W – 53 W = 59 W) are dissipated in Q1 and Q2. Assuming equal current sharing between Q1 and Q2, each transistor dissipates 29.5 W. To keep the transistor junction temperature below +150°C requires preventing the transistor case temperature from exceeding 46.8°C (150 – [3.5 × 29.5]) while dissipating 29.5 W. Also, there is a temperature rise across the mica insulator between the transistor case and heat sink of 0.5°C/W. That makes the maximum allowable heat-sink temperature limited to 46.8 – (0.5 × 29.5) = 32°C. In other words, the heat sink must dissipate 59 W (29.5 from each transistor) with only a 7°C rise above room temperature (25°C). Even if the junction temperatures were allowed to reach the absolute maximum of 175°C, the heat sink temperature must not exceed 57°C. Accomplishing this requires a heat sink with a thermal resistance of (57 – 25) / 59 = 0.54°C/W. This is far less than the 1.9°C/W rating of the AAVID 244609B02 heat sink I used. The situation may seem bleak, but all is not lost. These calculations make it clear that the amplifier should not be used for AM, FM or any other continuous-carrier operation. The amplifier should be used only for CW and SSB operation where the duty cycle is significantly reduced.

Thermal performance of the amplifier is illustrated in Figure 8. Data was taken under dc operating conditions with power-dissipation levels set equal to conditions under RF operation. A RadioShack brushless 12 V dc fan (RS 273-243A) blows air across the heat sink. Key down, the maximum rated junction temperature is reached in as little as five seconds as illustrated in Figure 8. Prolonged key-down transmissions should be avoided for this reason.

Under intermittent CW conditions, the situation is very different. Transistor-case temperatures reached 66°C after operating four minutes under simulated CW conditions at 20 WPM (60 ms on, 60 ms off). The corresponding junction temperature is +141°C (based on an equivalent RMS power dissipation of 21.7 W per transistor). This keeps the junction temperature under the 150°C target (see Figure 9). One simple way to reduce power dissipation is to reduce the power-supply voltage to 24 V. RF output power will decrease about 10 W from the maximum levels achieved with a 28 V supply.

From a thermal standpoint, the IRF510 power MOSFET is a poor choice for this RF amplifier application. Although I must say I am impressed with the robustness of these devices considering the times I spent testing them key down, five minutes at a time, without failure. Q1 and/or Q2 may need to be replaced after a year or so of operation because of the compromise in reliability. Considering their low cost, that is not a bad trade-off.

Stability

High gain, broad bandwidth and close input/output signal routing (within the TR relay) all work against stability. With a good load (< 2:1 SWR) the amplifier is stable from 1.8 MHz through 39 MHz. Oscillation was observed when the transmitter frequency was increased to 40 MHz. The output load match also affects stability. Oscillation was observed on 27.5 MHz when the load SWR was 3:1. This should not be a problem since the frequency is outside the ham bands. I spent a great deal of time trying to make this design unconditionally stable even with loads exceeding 3:1 SWR without sacrificing output power (gain) at 28 MHz without success. I did identify some reasonable compromises.

One of the easiest ways to improve stability and the input SWR seen by the RF source is to add an RF attenuator (pad) at the amplifier input. An attenuator is absolutely required if the transmitter (driver) provides more than 1 W to the amplifier. R8, R9 and R10 form an RF attenuator that attenuates the transmitter drive level, but does not attenuate received signals because it is only in the circuit when K1 is energized. To drive this amplifier with a 2-W-output transmitter requires use of a 3-dB pad. The pad improves the amplifier input SWR and the isolation between the amplifier's input and output. The drawback is that 1 W is wasted in the pad. Likewise, a 5-W driver requires use of a 7-dB pad, but 4 W are wasted in the pad. (Values for R8, R9 and R10 to make a 3-dB pad and a 7-dB pad are given in the parts list.) Installing a pad requires cutting the PC-board trace *under R9*, otherwise R9 would be shorted out by the trace. Make a small cut (0.1 inch wide) in the trace under R9 before soldering R9 in position. R8 and R10 have the same values, but may have different power ratings. Connect R10 between the RF input side of R9 and ground. Install R8 between the amplifier side of R9 and ground.

An impedance mismatch between the output of a 1-W-output driver and the amplifier input can be a source of instability. (Obviously, if the driving transmitter's output power is only 1 W, you can't use a pad as described earlier.) If you encounter stability problems, try these remedies: Place a resistor in parallel with L1 and L2 to decrease the Q of the amplifier matching network (try values between 50 and 220 Ω). Try reducing the value of L3 or eliminating L3 entirely. Both of these modifications improve stability, but reduce the amplifier's output power above 21 MHz.

Summary

This project demonstrates how inexpensive power MOSFETs can be used to build an all-band linear HF power amplifier. Frequency of operation is extended beyond the limits of previous designs using the IRF510 and improved input-impedance matching. Long-term reliability is recognized as a compromise because of the poor thermal performance of the low-cost TO-220 package.

If you have been thinking about adding an amplifier to your QRP station, this project is a good way to experiment with amplifier design and is an excellent way to become familiar with surface-mount "chip" components. I made arrangements with Mouser Electronics and Amidon Inc to provide parts kits for this project at a discounted price (see the parts list in Part 1). These parts kits make it very easy to get started and more economical to "homebrew" this project.

Acknowledgments

I want to thank the following individuals associated with this project: Harry Randel, WD2AID, for his untiring support in capturing the schematic diagram and parts layout of this project; Al Roehm, W2OBJ, for his continued support and encouragement in developing, testing, editing and publishing this project; Larry Guttadore, WB2SPF, for building, testing and photographing the project; Dick Jansson, WD4FAB, for thermal-design suggestions; Adam O'Donnell, N3RCS, for his assistance building prototypes; and my wife, Laura, N2TDL, for her encouragement and support throughout this project.

Notes

[11] Mike Kossor, "A Broadband HF Amplifier Using Low-Cost Power MOSFETs—*Part 1*," *QST*, Mar 1999, pp 40-43.

[12] R. Dean Straw, N6BV, *The 1999 ARRL Handbook for Radio Amateurs*, (Newington: ARRL), 76th ed, pp 25-23*ff*.

By Zack Lau, KH6CP/1

From *QEX*, May 1992

A 1.8 to 54 MHz 5-Watt Amplifier

Need a rugged and stable amplifier for your multiband QRP rig? Not only has the design been optimized on a pricey computer program called *Touchstone* (by eesof) for unconditional stability, it has actually survived a variety of poor loads—it was used to sweep filters with 5 W of RF. The gain of the two stage amplifier was measured to be between 28 and 30 dB in the amateur bands, though there is another dB of gain around 37 MHz.

For ruggedness and ease of design, a Motorola MRF 137 was selected as the final transistor. While the MRF 138 may be more linear, insufficient design information was available to ensure a stable design. While some amateurs will balk at the high cost of these devices ($24 in November 1991), such savings are easily lost if the cheaper device has a habit of blowing up. Also, one picks up a real clean SSB signal—the high-order IMD products are way down compared to typical bipolar amplifiers. For instance, the worst IMD on 3.5,7,14 and 28 MHz was –38 dB on 28 MHz, with the 5th order products 61 dB down (relative to PEP). The device was putting out 5-W PEP while being biased at 0.5 A (28-V supply).

Perhaps the biggest flaw is the power requirement— these FETs really like to see high voltages for best performance, and the MRF 137 is no exception. I biased the MRF 137 for 0.55 A and 28.2 V. It drew 0.6 A when putting out 4.6 W at 28 MHz. The driver runs off your normal 12-V supply.

The input amplifier shown in Fig 1a is pretty straightforward—a bipolar 2N5109 with the feedback networks adjusted to compensate the gain of the MRF 137. A series network of a 470-a resistor and a 12-pF capacitor was tacked between the collector and ground to ensure stability at all frequencies. The MRF 137 rolls off a few dB at 54 MHz, but the bipolar amplifier adequately compensates for this gain deficiency. The input return loss is better than 18 dB between 1.4 and 29.9 MHz, but degrades to 12 dB at 50 MHz. The input SWR was not tested with poor loads.

By itself, the MRF 137 amplifier stage shown in Fig 1b (see next page) makes an excellent 16-dB gain block between 1 and 32 MHz, having less than 0.5 dB of gain variation. The transmission line transformer on the input seems to help the input return loss/SWR, keeping these numbers above 18/below 1.3 to 1 between 1 and 50 MHz. I suppose that putting another transmission line transformer on the output could be used to get a more powerful amplifier with less gain over a similar frequency range, but this variation has not been investigated.

The simplest circuit board I could think of was used— I cut two pads in a piece of double-sided circuit board for the gate and drain leads. Then I wrapped the edges of the board with copper tape and soldered it down for good grounding. After making holes for the MRF 137 transistor and the mounting screws in the board and a spacer, made of 0.050-inch aluminum, I attached the spacer, the circuit board, and the MRF 137 to a heat sink tapped with 4-40 screw holes. Standard ground-plane construction was used to attach the other parts. The 2N5109 amplifier was built on its own ground plane—RF amplifiers work better if there isn't too much gain in one place. Three additional amplifiers were built by Mike Gruber, WA1SVF, for use in the lab. He noted that R8 had to be changed from 4.7 kΩ to 1 kΩ to bias the amplifiers at 0.5 amps. Apparently, the MRF 137s he used have a higher gate threshold voltage. Otherwise, performance was as expected.

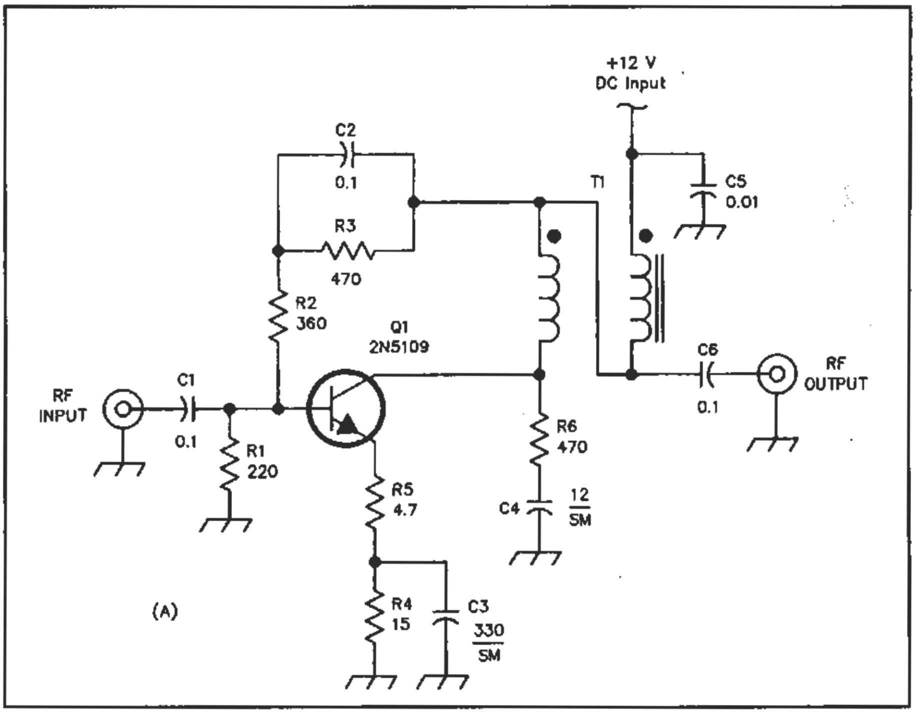

Fig 1a—Low-level amplifier designed to compensate for the gain rolloff from the power amplifier.
Q1—2N5109, 2.5-W heat-sinked RF transistor, f_T is 1200 MHz.
T1—15 turns bifilar #28 on FT-37-43 toroid core.

Fig 1b—5-W TMOS power amplifier.
L1—26 turns no. 26 enameled wire on T-44-2 toroid, 3.9 µh.
Q2—MRF 137 transistor.
R9—10-kΩ turn potentiometer for bias setting.
RFC1—21 turns of no. 26 enameled wire on FR-37-67 toriod.
T2—4 turns 25-Ω coax on FT-50-43 toroid core. The 25-Ω coax is actually two 50-Ω coax run side-byside. The prototype used RG-196/U coax.
U1—78L05 5-V regulator.

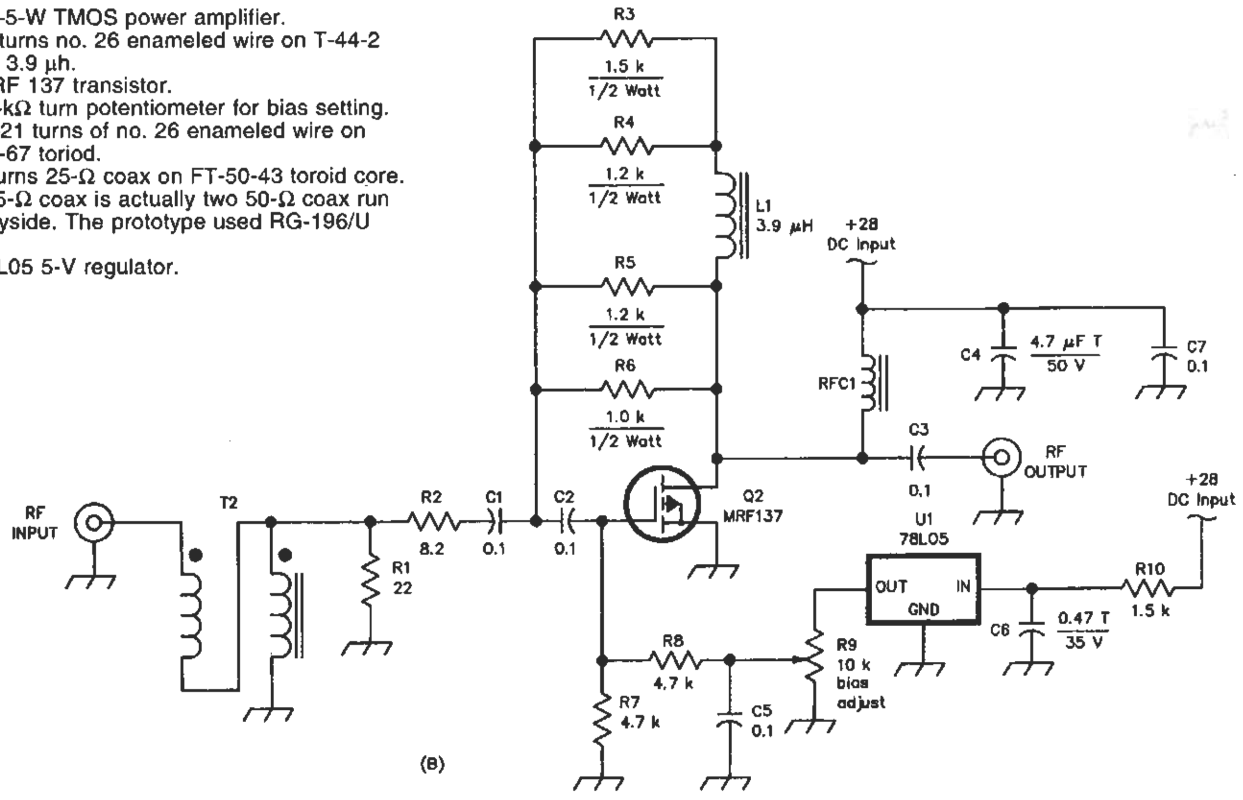

R3
1.5 k
1/2 Watt
R4
1.2 k
1/2 Watt
L1
3.9 µH
+28
DC Input
R5
1.2 k
1/2 Watt
R6
1.0 k
1/2 Watt
C4
4.7 µF T
50 V
C7
0.1
RFC1
C3
0.1
RF
OUTPUT
RF
INPUT
T2
R2
8.2
C1
0.1
C2
0.1
Q2
MRF137
+28
DC Input
R1
22
U1
78L05
OUT
IN
GND
R10
1.5 k
C6
0.47 T
35 V
R8
4.7 k
R9
10 k
bias
adjust
R7
4.7 k
C5
0.1
(B)

By Joel Paladino, N6AMG From *QST*, September 1992

An Experimental Solid-State Kilowatt Linear Amplifier for 2 to 54 MHz

Many kilowatt amplifiers could anchor a small boat, and don't cover 6 meters. Combined with its power supply, *this* kilowatt weighs less than 35 pounds.

This article describes progress toward achieving a well-defined goal: building the smallest possible MF/HF/VHF amplifier capable of at least 1000 watts output. Nicknamed the Solid State Kilowatt (SSKW), the project has its roots in an article by Helge Granberg of Motorola in October 1986 *RF Design*.[1] Two water-cooled amplifiers of that design, built by Mike Staal (K6MYC) of M2, had many problems. Some of them were device-related; others, power-supply related. (More about these issues later.) In February 1990 I rebuilt one of the water-cooled units using the old design and two new transistors. Optimized for 50 MHz, this amplifier could just reach the 1000-watt level. After using it for three weeks of South Pacific DXpeditioning, I decided to construct an air-cooled unit. I spent the spring, summer and fall of 1990 building the SSKW and getting it ready for another DXpedition in the fall.

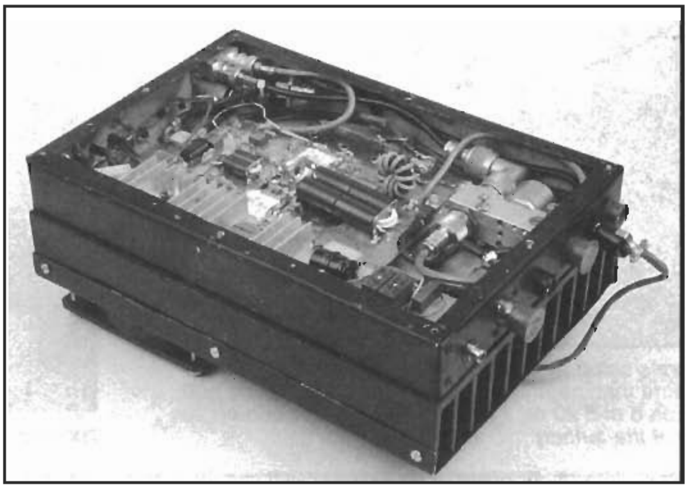

Circuit Description

Motorola's article reprint AR-347 describes the basic amplifier design. July 1990 QEX also carried an article about it.[2] Fig 1 shows the amplifier schematic. Briefly, the circuit consists of two MRF154 RF power MOSFETs in push-pull. Each of these transistors is capable of 600 watts output up to 100 MHz. The input and output transformers, 9:1 and 1:9, respectively, use cores of #67 ferrite material. The output transformer's 1:9 ratio is a compromise that is optimum at about 800 watts.

Incorporating the Motorola building block into a DXpedition-ready package required experimentation and problem-solving, as I described at the 1991 Central States VHF Conference.[3] Here's where the system stands today.

TR Switching

The SSKW includes TR relays that bypass the Fig 1 circuitry in receive mode. I designed their control circuitry to eliminate the possibility of the high RF fields in the amplifier compartment causing relay falsing problems, and to simplifies field repair by keeping parts count low. The relays are sequenced to allow the output relay to close before the amplifier puts out power. Because relays take a few milliseconds to operate, sequencing is necessary to keep the relay from hot-switching the amplifier output. (Hot-switching 1 kW will destroy a relay rapidly!)

This design has two small problems. For the short time it takes the amplifier's input relay to close on switching from receive to transmit, the exciter operates without a load. This can be remedied by delaying the exciter keying. The other problem is the relay's closure time changes as the relay heats. This can be solved two ways: Use external electronics to do the delay, or temperature-compensate for the resistance change.

SWR Protection

I haven't yet built SWR protection into the amplifier. In January 1983 *QST*, Helge Granberg described an SWR-protection circuit for a 2- to 30-MHz amplifier.[4] A phone conversation with Helge indicated that this circuit will work to 54 MHz, so it should suffice. How much mismatch the MRF154s can tolerate is unknown, however, so I don't

6-Meters: DXpeditioning with a Difference

Making the most of 6 meter DXpeditioning depends on knowing where and when to go If you want to work Europe from the Canary Islands summertime is best because sporadic E propagation is best then If you want to work the US from EA8 go when fall shifts into winter to take advantage of east/west F2 layer propagation It's always more exciting to go to a rare place but not necessarily more fun'

Once you've decided on a DXpedition site plan your travel on the assumption that you'll want to stay on site for at least two weeks This gives you about a 50% better chance of having good openings within a given month

Pay particular attention to the RF prospects of your accommodations Ideally you'd be able to see ocean in all directions If that's not possible all's not lost A direct oceanic path isn't always the best propagation may be better over the backscatter path You'll want to get your antenna up as high as possible to minimize radiation angle and AFI And consider EMI and NI possibilities carefully Some areas of the world use NTSC channel 2; others, PAL on channel 1—right in the middle of the band! If your RF gets into a resort's CATV system you'll drive 60 televisions crazy rather than one or two.

Chapter 6 of *The ARRL Operating Manual* covers DXpedition power, health, licensing and logistical concerns in detail so see that book for more about those topics I'll add one thing though It helps to find a local ham with FAX capability as well as HF who can help you with questions on licensing and living accommodations

This can take 80% of the surprises out of the expedition!

Equipment

What I bring and use on 6-meter DXpeditions is based largely on the experiences of Jim Treybig, W6JKV, in his many years of 6 meter DXpeditioning. The core of the approach is pretty much this: Run as much power as permitted and bring as big an antenna as possible. There are practical limits to this, but let your ingenuity direct your thoughts. I consider 100 W as a minimum, 500 watts good and 1000 watts as optimum. High power lets you take advantage of scatter paths that simply won't work at 100 watts. It's very frustrating to hear a well equipped station that you cannot work!

Which Antenna?

A DXpedition antenna must be compact, light and reproducible The Yagi I use has six elements on a 30 foot boom, Optimized by Brian (K6STI) Beezley's Yagi optimization program *YO* it's fed with a T match and a half wave balun. Its boom folds down to two 40 inch sections, each 2 inches in diameter. Most of the element pieces fit inside the two 40-inch sections. The antenna is light enough to lift with one hand.

A rotatable antenna is *mandatory*. You'll need to turn your antenna to find the best direction of propagation. The antenna mast, 20 feet long overall, consists of 40 inch sections of 2-inch aluminum tubing joined with internal sleeves. The rotator is at the mast bottom and there's a slip ring at the mast top just below the Yagi (Having the rotator at the bottom makes putting the antenna up easier by minimizing weight at the upper end of the mast.)

I use Dacron rope for guying. How much rope should you

The author's setup at CN2JP, Rabat, Morocco. Efficiency counts more than neatness while you're busy making 900 contacts on 6 and 20 on 2-meter EME! *(DXpedition photos courtesy of the author)*

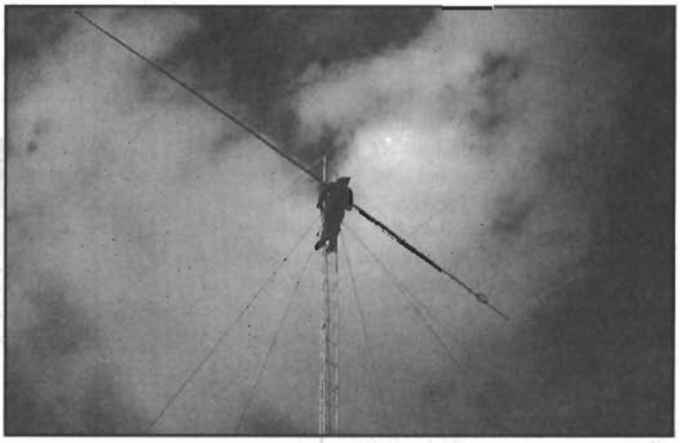

DXpeditioning would be no fun at all without antenna work!

want to use the SSKW to calibrate its own SWR-protection circuitry!

The amplifier's output can be reduced two ways: by turning down the exciter power via amplifier-generated ALC, or by reducing the amplifier bias. (These are enhancement-mode FETs, so positive bias is necessary to turn them on.) The ALC option is better because it does not affect the amplifier's linearity. A no-output-load condition must shut down the amplifier as quickly as possible.

Overdrive Protection

Overdrive can destroy power MOSFETs instantly, so a drive limit control is essential. A threshold detector should be incorporated. The SSKW's input includes a 5-dB attenuator so 100-watt exciters can drive the amplifier without damaging it. If the 5-dB pad fails to attenuate, or if the exciter puts out too much power, the amplifier must be shut down rapidly.

The amplifier must also be shut down if the transistors' flange temperature exceeds 40°C. Shutdown can be done two ways, depending on how the amplifier is used. The first way is to just turn down the bias, which changes the amplifier's linearity. The second way is to totally turn off the amplifier. This must be done gracefully so the amplifier does not shut down while producing full power. The following shutdown sequence is essential: (1) Turn down the bias; (2) open the input relay; and (3) open the output relay.

Bias

The SSKW's bias circuitry is straightforward. R8's value must be tailored to the particular MRF154s used. If the bias decreases too fast with rising temperature, increase the value of R8. The LM723's input voltage must never exceed 40 volts.

Output Filtering

Filtering must be added to make the amplifier comply with FCC signal-purity regulations. The reactance of these filters away from their intended passbands must be taken into account. Absorptive low-pass filters should be considered. Such filters dissipate harmonic energy as heat rather than reflecting it back to the amplifier transistors.

Packaging

I intended to make this amplifier as small and light as possible. Most of the amplifier's weight is its aluminum heat sink take? Simple: Estimate what you need, multiply by 2.5 and you might make it! I usually bring two diameters, $^{3}/_{16}$ and $^{1}/_{8}$ inch.

Radios

Bring at least two radios: one for HF liaison (usually at 10 meters) and another for 6 meters. I recommend maximizing redundancy by taking two radios that both cover 10 and 6. At least one of the 6-meter rigs should include an excellent noise blanker for power-line problems and motorbikes. It must also be able to drive a kilowatt amplifier to full output on 6. (I use one Kenwood TS-680S and one ICOM IC-575.) Receive coverage between 10 and 6 meters is another plus. Where it's allowed, listening to nonamateur services between 45 and 50 MHz can be very important in determining the direction of propagation.

Additional Gear

Bring 12-volt supplies capable of operating at the line voltage and frequency at your DXpedition destination. The supplies should also be able to handle wide line-voltage variations around nominal. Carry two of them—one for each radio. I modify mine by adding line filtering and fusing. Parallel a standard receptacle of some kind with your supplies' rig plug or cable. Two-prong, polarized Jones plugs work well. You can use them on everything for quick setup and breakdown.

Bring a memory keyer for use as a beacon and in normal CW operation. Also consider taking a laptop computer. I now use a computer for logging, taking notes, calculating beam headings, and determining the footprint of the sun.

Operating

Several operating subjects are of prime importance. Ten-meter liaison comes first. You must use 10 meters to get up-to-date propagation information and find out who's being heard. When you're feeling your way through a pile of stations, you can sometimes miss very rare propagation opportunities if you don't check 10 meters often.

Six-meter contact procedures come next. It's very difficult making everyone happy on 6 meters when you're the DX. You must set rules and stick to them. Pick a frequency and do your best to keep it when the band gets busy. Try to not work the same stations over and over again, especially during pileups. Your goal is to give as many people as possible a new country. Work each station as quickly as possible. Don't discuss names and grid squares—the time it takes just gives fewer people a chance to work you. Leave that info to the QSL cards! Which should you operate—phone or CW? The answer is "Probably both." CW penetrates weak conditions better than phone, and it gives more people at the edge of propagation a chance at working you. Then, when you go to phone, everyone who can will worK you again. I favor CW DXpedition operation because of my experiences on the non-DX end.

Should you operate on a single frequency or split? Go to split when there's a massive opening and you're being QRMed by people calling you. This usually happens during intense backscatter openings. Hams on the 10-meter liaison frequency can tell you when QRM gets too heavy.

DXpeditioning is also about meeting people. Here's Tarik Skiredj, CN8ST, and friends.

Conclusion

Six-meter DXpeditioning is a blast. And the SSKW makes it even more fun: I can devote more weight to antennas!—*N6AMG*

and copper heat spreader. How much heat sinking is needed depends on the duty cycle required. I decided that a duty cycle of 50% or less was acceptable. The resulting weight for the package, 14 pounds, is very acceptable. The SSKW is 8 inches wide, 12 inches long and 5 inches high.

The key to making the package light was to use the smallest possible heat spreader. I decided to use enough ¼-inch-thick copper to surround each MRF154 with at least 1 inch of spreader. I machined the spreader flat and attached it to the heat sink, also machined flat. After attaching the copper to the aluminum, I resurfaced the copper again at the transistor contact points to insure flatness. I spread a thin layer of heat-sink compound between the copper and aluminum, and another layer between the copper and the transistors.

With two high-pressure, 24-volt fans blowing on the sink, the resultant duty cycle limit is a little less than 50%. These fans are in series across the amplifier's 50-volt supply. The small fan in the compartment is a 12-volt unit that just barely fits. (The compartment needs to be a little deeper.) A fan is necessary here, however, because compartment airflow dramatically keeps the output transformer's ferrite from getting hot.

A heat sink with more fins and a thicker base, but the same fin depth, would raise the duty-cycle limit. (The copper spreader should be made larger as well.) These changes would also make the amplifier heavier.

Construction Techniques

A milling machine is required for surfacing all of the amplifier's heat-conductive interfaces. Physical flatness is essential for maximum heat conductivity and to avoid warping the transistor cases. The Motorola bulletin specifies the torque of the screws on the MRF154s' flanges. Tightening these screws uniformly assures that the flanges won't warp during temperature cycling.

R14 and R15, the feedback resistors, are flange-mounted. Their beryllium copper leads cannot be flexed very many times before they break. (According to their manufacturer, that may be only once!) These two resistors just barely fit. So, make sure the two MRF154s are separated properly.

Power Supply

To keep the amplifier small and light, I use a switching power supply. A Lambda

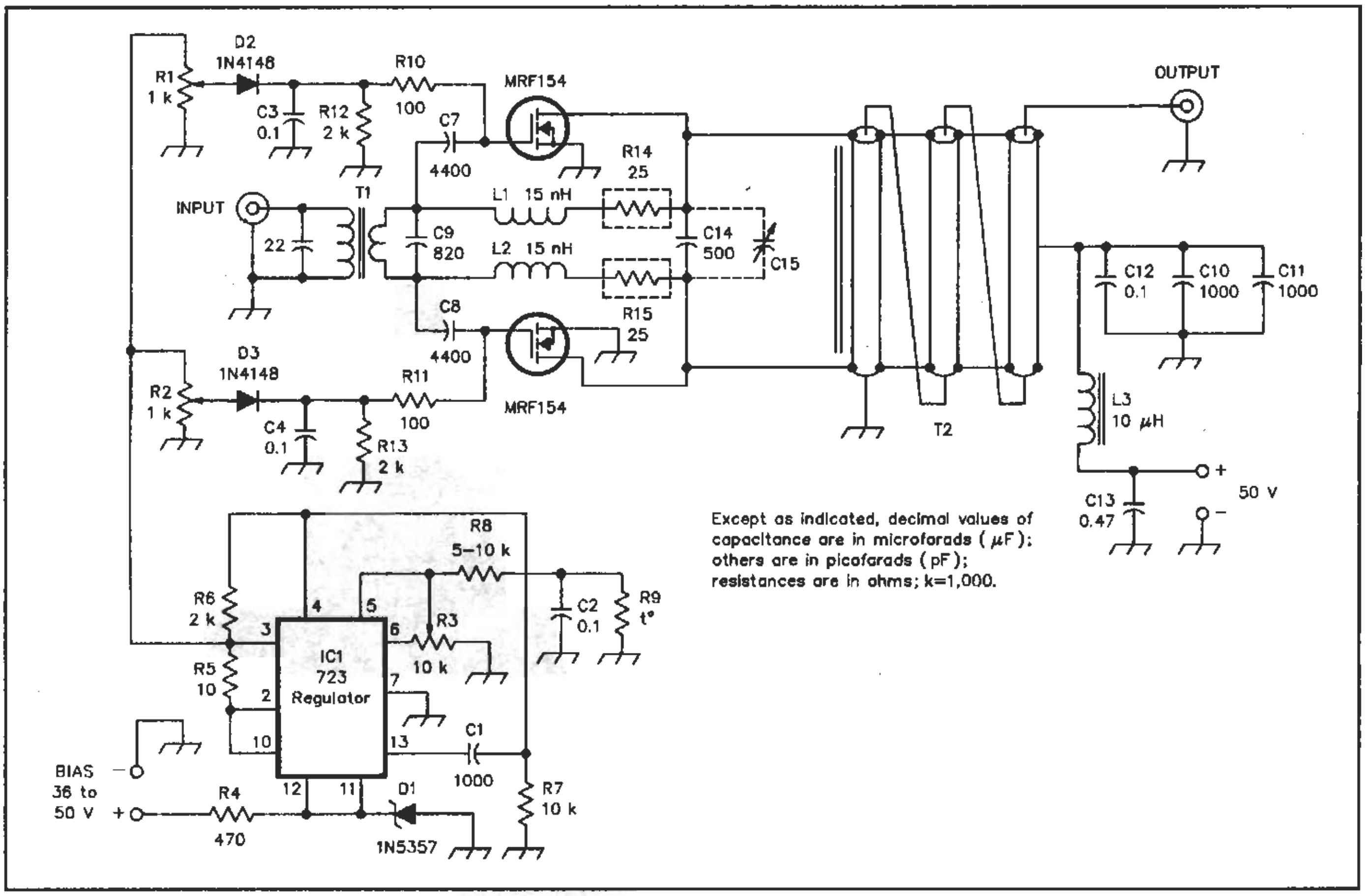
D2
1N4148
R1
1 k
C3
0.1
R12
2 k
R10
100
C7
4400
MRF154
R14
25
OUTPUT
INPUT
T1
22
C9
820
L1 15 nH
L2 15 nH
C14
500
C15
C8
4400
R15
25
MRF154
D3
1N4148
R2
1 k
C4
0.1
R11
100
R13
2 k
T2
C12
0.1
C10
1000
C11
1000
L3
10 μH
C13
0.47
+
50 V
−
R8
5–10 k
R6
2 k
R5
10
IC1
723
Regulator
R3
10 k
C2
0.1
R9
t°
C1
1000
R7
10 k
BIAS
36 to
50 V
R4
470
D1
1N5357

Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k=1,000.

Fig 1—The basic SSKW circuit can produce at least 1 kW from 2 to 54 MHz. Duplicating this circuit requires additional information not given here. Output filtering is also necessary to ensure compliance with FCC emission-purity rules. See the text, Motorola Application Note AN-287, July 1990 *QEX* and the *1991 Central States VHF Conference Proceedings* for details. The *Central States* write-up includes SSKW performance graphs and spectrograms in addition to a diagram of the amplifier's relay-control circuitry.

LFS-50-48 (net weight, 20 pounds) does the job, with good results. Sized at 15 inches long, 7.5 inches wide and 5 inches high, it can source 50 amperes at 50 volts. This power supply produces significant radio noise up to 30 MHz, most of it radiating from the supply's ac-line leads. I EMI-filtered the supply's leads with good success. (We need better, RF-quieter switching supplies!) Because linear supplies generate little or no RF EMI, they are superior to switching supplies for fixed-station use.

Overshoot is a regulated power supply characteristic that's particularly troublesome in switching supplies. Overshoot occurs when a regulator responds too slowly to keep its output voltage down in response to short-duration, high-current loading. In a 50-volt switching supply, and depending on the load, overshoot transients of more than *100 volts* may result. I've seen overshoot destroy expensive transistors! The SSKW supply must be able to safely handle the variable power-supply demand that occurs during SSB and CW transmission.

Power-supply RF sensitivity is another consideration. Some power supplies are sensitive to RF. Their output voltages may vary with the presence and amplitude of RF on their input and output leads. The SSKW's supply must be free of such effects.

Hints and Kinks

Be careful when applying dc to the amplifier for the first time. Do not use a high-current supply for initial tests. Use a

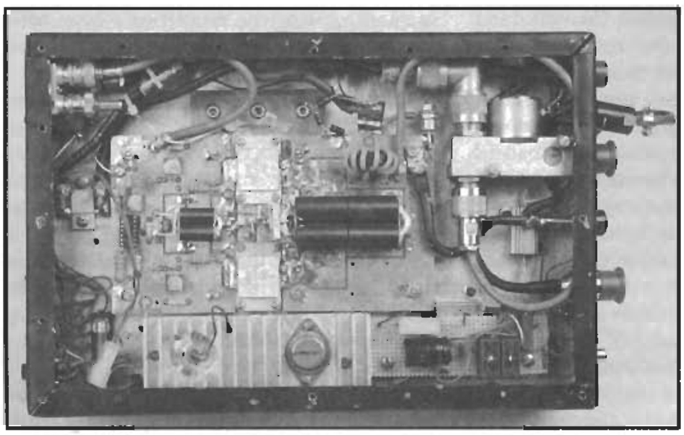

Two Motorola MRF154 RF power MOSFETS in push-pull make the SSKW perk. A 20-watt wxciter drives the amplifier to full output. (amplifier photos by Kirk Kleinschmidt, NT0Z)

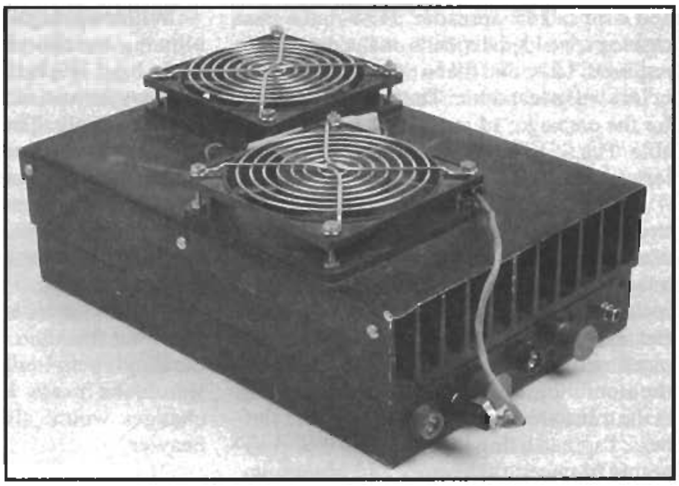

Without its two 24-volt fans, the SSKW's heat sink and spreader would have to be much larger.

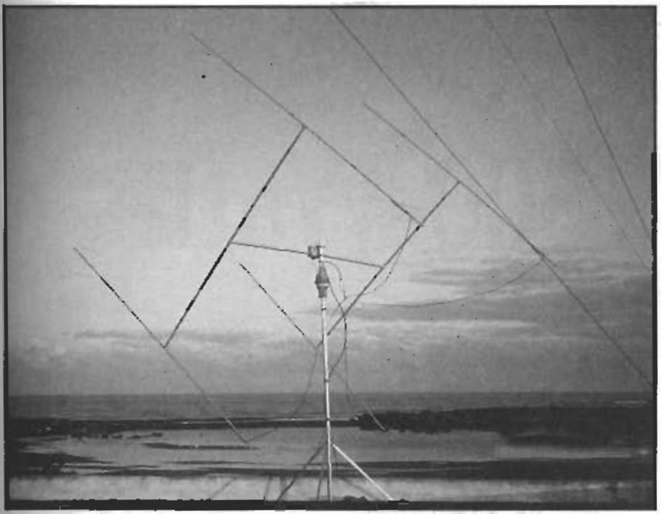

What luggage weight the SSKW amplifier conserves can be put to other uses—2-meter moonbounce gear, for instance. Operating as CN2JP, the author completed 20 EME contacts—Morocco's first—in November 1991. On the same trip, the SSKW performed like a heavyweight through 900 contacts with 56 countries on 6 meters.

current-limited, 3- or 4-ampere supply to check the regulator and set the bias. The amplifier should have a reverse polarity protection diode rated at a voltage appropriate to the supply.

Make sure that you have a 50-ohm load on the amplifier output. The bias circuitry may not act correctly if there is no load. Slowly increase drive. Watch the output power and the current drain to make sure they are in line with efficiency. If the input match is bad, check C7 and C8. If the amplifier does not achieve its efficiency capability, RF that doesn't make it to the load may try to leave via the dc input line. This can cause C12 to explode off of the underside of the board. Also, under normal operation, the single unit Motorola specifies for C12 can barely handle the RF current that passes through it. I recommend paralleling 100-volt, 0.05- or 0.1-µF chip capacitors instead of using a single 0.1-µF chip.[5] This problem is difficult to analyze because C12 is on the underside of the board.

As mentioned earlier, the SSKW's bias circuitry may have to be tailored to the particular MRF154s used. One thing not mentioned on the schematics is the addition of an Arco 365 variable capacitor (C15 in Fig 1) across C14 to cancel some of the inductive reactance in the output transformer at 6 meters.

Performance

With a 50-volt power supply, the SSKW can produce over 1.1 kW from 2 to 54 MHz. At this power level, the MOSFETs' drain current runs at around 40 amperes, depending on the operating frequency. Table 1 shows amplifier performance data taken at six different frequencies. The SSKW can produce up to 1.5 kW below 30 MHz.

Especially below 30 MHz, the SSKW's harmonic output rises with output power. It may be possible to use only four low-pass filters to cover the 2 to 54 MHz range if a second harmonic of –40 dB can be tolerated. Since the transistors operate in push-pull, the second harmonic is usually not a problem.

Without its 5-dB input pad and operating with a 50-volt supply, the amplifier exhibits an input SWR of less than 2:1 (referred to 50 ohms) throughout its frequency range. Only below 5 MHz does its input SWR exceed 1.6:1. Inserting the 5-dB pad adds 10 dB of return loss and keeps the input SWR below 1.3:1 through the amplifier's operating range. I have not yet measured the amplifier's two-tone, thirdorder IMD performance.

Table 1
SSKW Performance Versus Frequency

Freq (MHZ)	*Drive Power (W)*	*Power Output (W)*	I_D *(A)*	*Gain (dB)*	*Efficiency (%)*	*Input SWR*	V_{cc} *(V)*
2	24.5	1000	39.1	16.1	51.2	1.41	50
4	21.0	1000	33.4	16.8	59.9	1.39	50
8	17	1000	32.4	16.8	59.9	1.39	50
15	15.5	1000	32.3	17.7	61.9	1.36	50
30	19.5	1000	41.5	17.1	48.2	1.87	50
50	29	1000	39.7	15.4	50.4	1.51	50

Final Notes

For now, I can say that my goal of a compact, lightweight amplifier capable of at least 1 kW output is a reality. The SSKW has served me well in its intended DXpeditionary application. Nonetheless, the amplifier in its present form is still experimental. Before the SSKW could be acceptable for general use, output filtering and failureproof protection circuitry would have to be added. The next generation is on its way. The next project? Use the same 50-volt supply with power FETs capable of covering 144 to 432 MHz, and build it into a similar package!

Acknowledgment

I thank Helge Granberg for spending considerable time answering questions about this project.

Notes

[1] H. Granberg, "New MOSFETs Simplify High Power RF Amplifier Design," RF Design, Oct 1986, pp 43-48, 50, 52.

[2] H. Granberg, "A Compact 1-kW 2-50 MHz SolidState Amplifier," QEX, Jul 1990, pp 3-8.

[3] J. Paladino, "A Rapid Deployment Prototype Solid-State Kilowatt Linear Amplifier for 2 to 50 MHz Using MRF-154s," *1991 Central States VHF Conference Proceedings* (Newington: ARRL, 1991), pp 61-70. Available from the ARRL Bookshelf as #3614.

[4] H. Granberg, "MOSFET RF Power—An Update," Part 2, *QST*, Jan 1983, pp 30-33; also see Feedback, *QST*, Mar 1983, p 44. (Part 1 appeared in *QST*, Dec 1982, pp 13-16; also see Feedback, *QST*, Jan 1983, p 48.)

5American Technical Ceramics (ATC) has some new chip caps (the 900 series) that have higher voltage and current ratings than the capacitor Motorola specifies. I have samples but have not tried them yet.

By Jerry Pittenger, K8RA **From *QST*, September 1986**

An All-Band, 1500-Watt-Output 8877 Linear Amplifier

Part 1—This rock crusher, rated for continuous full-legal-limit output, can be built at home. It is, however, a major project requiring dedication and commitment.

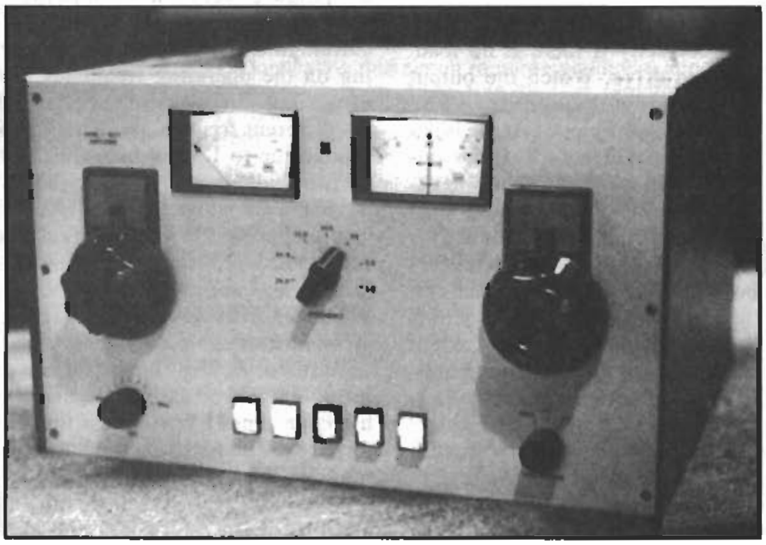

This article is the result of a 10-month project to build a legal-limit linear amplifier. The amplifier uses the popular EIMAC 8877 (3CX1500A7) high-mu power triode that can provide a continuous RF output of 1500 W to the antenna.

In recent years, I have built several different linear amplifiers, and I must admit that this previous experience was necessary to obtain the results achieved with this project.[1-4] I hope that by sharing this experience, others will benefit from it. Any amplifier design depends on the various components used and individual preferences. Therefore, you may not want—or be able—to duplicate this amplifier exactly.

The comment I receive most often from the amateur fraternity is about the high cost to build an amplifier like this. The criticism is valid. This amplifier is not inexpensive to build. Plan to spend from $1000 to $1200 for the RF deck, and another $500 to $600 on the power supply. If you really think about it, though, these costs are a bargain when you consider the performance and quality of the final product and the cost of an equivalent commercial unit.

This article is presented in two parts. In this part, I will describe the RF deck and power supply in general terms. Schematic diagrams and parts considerations are included. Part 2 gives detailed instructions for constructing the two units and considerations for the final testing and operation.

Preliminary Thoughts

Finding Parts

Finding parts can be a big task. Even the most difficult parts to find, such as the vacuum variable capacitors, vacuum relays and door-knob capacitors are available, however, and appear for sale in the ads (*QST* Ham-ads and the Yellow Sheets), or at hamfests and flea markets.[5] Probably the best source of parts is other hams who are actively building equipment. Go talk to these people and let them know what you are looking for. It's amazing how others will help, and even let you into their personal stores. There are people, like myself, who like to build amplifiers. Once you learn who these individuals are, keep in touch with them. They can help find the key parts.

Table 1
Recommended Tools

- Drill press or drill fixture (with set of highspeed bits)
- Band saw capable of cutting 1/16-inch-thick metal
- Chassis punches (5/8 inch to 1 inch)
- Fly cutter, 2-inch radius
- Vise
- Set of taps
- Common handtools (screwdrivers, pliers, soldering iron and gun)
- Volt-ohmmeter
- Variable power supply (5-26 V, 1 A)
- Dip oscillator

Parts that are not available in the surplus market can be purchased new. This will be necessary for some parts, such as

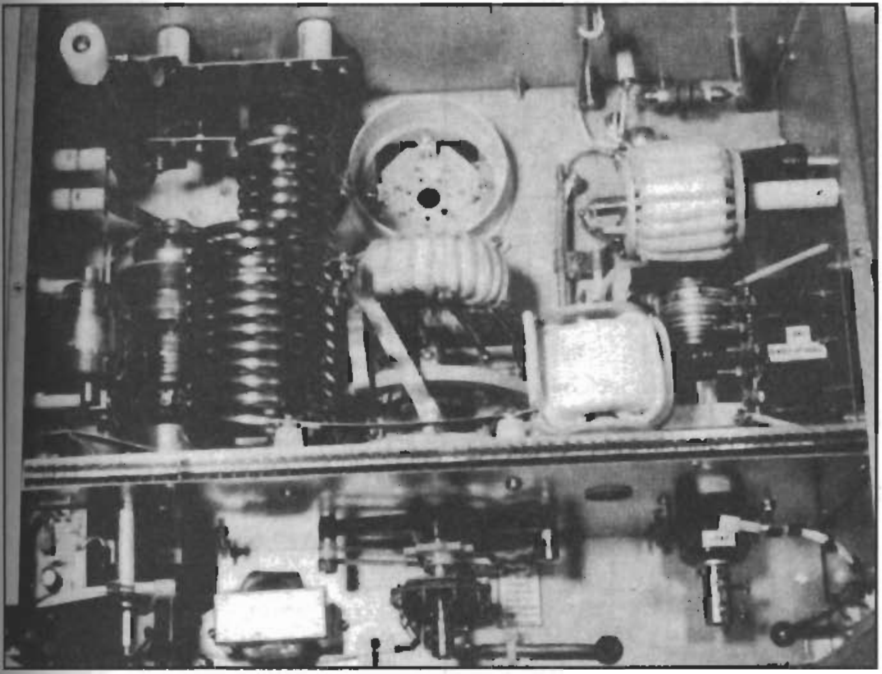

Fig 1—Top interior view of the 8877 linear amplifier RF deck.

cabinets, etc. Just remember that when you buy a new commercial amplifier, you pay the new price for every component.

Tools

A good assortment of hand tools, as well as some power tools, are necessary to complete this project. Table 1 shows the tools I recommend. In particular, I recommend that a drill press and band saw be available. You can do the job without all of the tools listed, but the job will be much more difficult.

Time

Time is probably the most valuable resource for most of us, and the one that may prove hardest to find. This project took well over 250 hours to complete. The key is to do each step right, and not hurry. Build the amplifier in a place where you can leave the project on the table and walk away. Plan each step and build in discrete modules. Work an hour or so whenever possible, and slowly, but surely, the modules will take shape. It is amazing how much you can accomplish using these small time segments. Also, great strides can be made on a Saturday or a Sunday. Commitment and consistency are the virtues required to finish the job.

RF Deck Circuit Description

The RF deck is designed to be a tabletop unit (see title photo). The power supply is remotely controlled and can be located almost anywhere. The amplifier design is based on proven circuitry. Included are all circuits required to provide a clean signal as well as adequate protection devices for the metal-ceramic 8877 tube.

Control Circuitry

Fig 2 shows the schematic diagram for the amplifier control circuitry and low-voltage power supply. The 117-V ac input from the high-voltage power supply enters the RF deck through a 5-conductor interconnecting control cable. Each control line is terminated in a pi-section filter as it enters the RF deck, to prevent RF from getting into the control cable and power supply. The pisection filters are constructed as an independent module.

The amplifier is powered up by the FIL ON/OFF switch, S1. Engaging S1 turns on the blower, filament power and 26-V dc power supply. The current inrush to the tube is limited by Ri, in series with the filament transformer primary. After approximately 1 second, K1 energizes and K1A shorts Ri thus providing full filament voltage to the tube. The K1 delay is controlled by R2 and C1 across the relay coil. R3, in series with the other leg of the filament transformer primary, is adjusted to provide the proper filament voltage (4.85 V ac) to the tube under load.

The 8877 requires a 3-minute warmup period to reach proper operating temperature. A solid-state timing circuit, formed by Q1 and Q2, locks the amplifier out of operation until the warmup period has elapsed. When the 26 V dc comes on, C2 charges through the 500-kilohm time-delay adjust and 1.2-megohm resistors. Q1 and Q2 form a high-impedance Darlington circuit, and the emitter of Q2 follows the voltage rise on C2. The high-impedance Darlington circuit is required to prevent the capacitor charge from draining through the transistors. After approximately three minutes, the potential at the emitter of Q2 reaches 18 V at which point the 4PDT relay, K2, engages. K2A applies 26 V dc to the K2 relay coil, removing the relay current load from Q2. The voltage also turns on the TIME pilot light located on the amplifier front panel to indicate that the warmup period is over. The same line also applies 26 V dc to S2B of the HV-ON push-button switch, which, when engaged, sends 26 V dc to the RF input/output relay circuits. K2B connects a 100-kilohm resistor across C2 to drain the charge from C2. This resets the 3-minute timer should the amplifier be turned off and immediately back on.

K2C and K2D are wired in parallel and apply 117 V ac to HV-ON switch S2A to energize the high-voltage power supply. The high-voltage power supply can't be turned on even if the HV-ON switch is engaged until after the 3-minute warmup period has ended. IN/OUT switch S3 allows the amplifier to be put in the standby mode with the amplifier turned on. Both HV-ON and IN/OUT front-panel push-button switches must be engaged to key the amplifier, thereby making it impossible to operate the amplifier without high voltage on the tube.

The amplifier is keyed by grounding the base of Q3 through the exciter TR-relay contact. A transistor is used to limit the current switched by the exciter VOX relay. This avoids a potential problem if the exciter VOX relay sparks on closure, which could damage the relay contacts. The "grid trip" break in the relay line causes the relays to drop out if the grid trip circuit actuates from too much grid current (approximately 120 mA). During normal operation, the grid trip break is shorted by a normally closed set of contacts on K3 (see Fig 3).

When the amplifier is keyed, the output RF relay must be closed before drive is applied to the tube—otherwise the tube will transmit for a brief period without a 50-ohm antenna load. This would not only be harmful to the tube, but also cause the grid-trip circuit to actuate. Therefore, a timing circuit, comprised of a 50-ohm resistor and i00-~iF capacitor, is included across the RF input relay K4 to allow vacuum relay KS time to close. The capacitor value depends on the relay used. Do not make the delay too long, since during the delay time, the exciter does not have a proper 50-ohm load. Check the time delay by placing a low voltage across the relay contacts and monitoring the contact closure on a dual-trace scope. I used a delay of about 20 ms.

RF Amplifier Circuit Design

The RF amplifier circuit is shown in Fig 3. The amplifier uses a tuned input network to minimize distortion products and provide a proper impedance match between the exciter and the tube. The input network is remotely switched, using small DPDT relays, to connect the correct pi-section for the selected band. A homemade switch deck is mounted on the band-switch shaft, in front of the subpanel, to ground the 12-V dc line for the proper input relay as selected by the main band switch. On 160 meters, the switch also controls a solenoid relay to add a 160-pF capacitance in parallel with the TUNE vacuum variable capacitor.

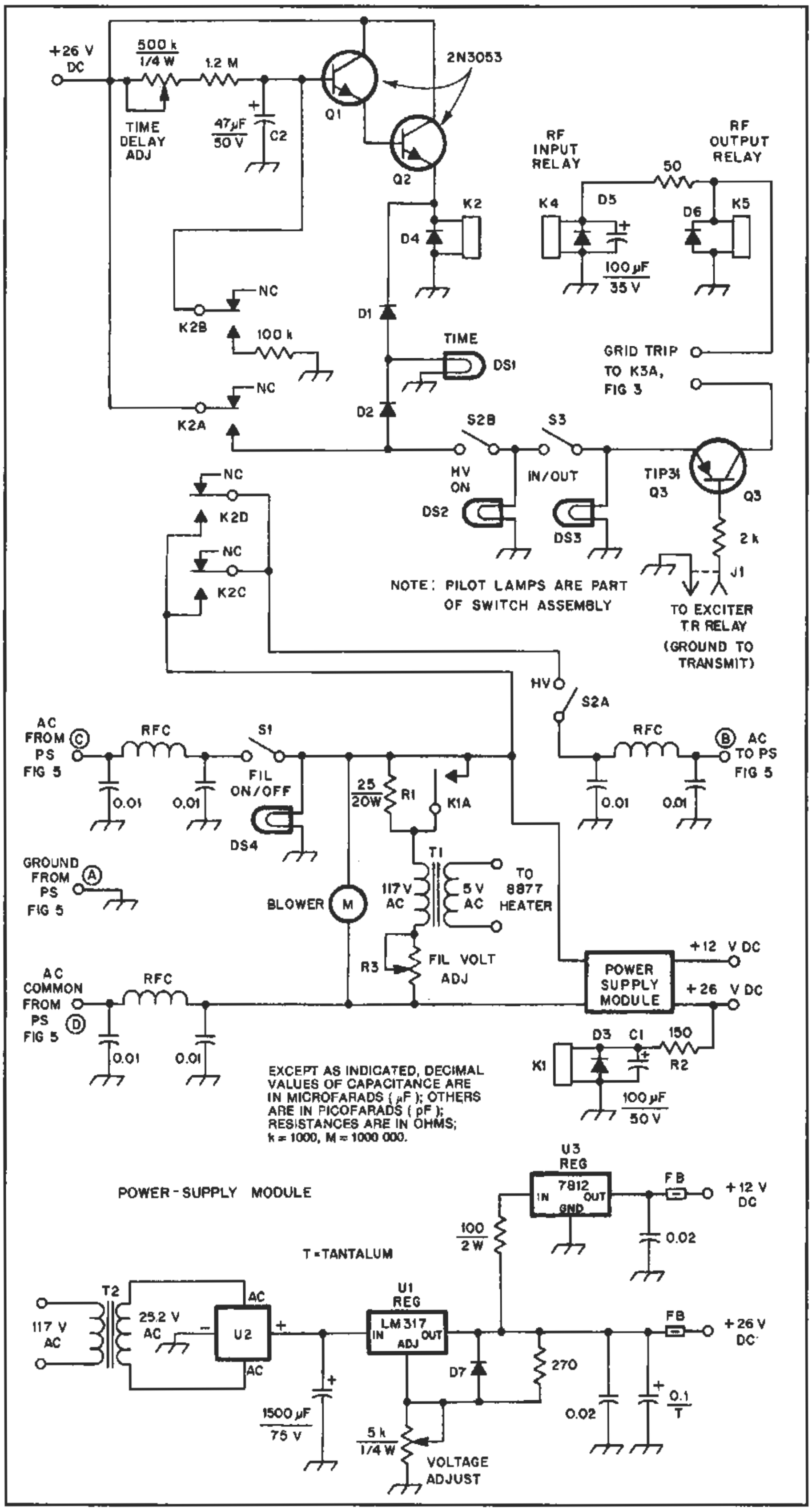

+26 V DC
500 k 1/4 W
1.2 M
TIME DELAY ADJ
47µF 50 V
C2
Q1
2N3053
Q2
RF INPUT RELAY
RF OUTPUT RELAY
50
K2
D4
K4
D5
100µF 35 V
D6
K5
NC
K2B
100 k
D1
TIME
DS1
GRID TRIP TO K3A, FIG 3
NC
K2A
D2
S2B
S3
HV ON
IN/OUT
TIP31 Q3
Q3
NC
K2D
DS2
DS3
2 k
NC
K2C
J1
NOTE: PILOT LAMPS ARE PART OF SWITCH ASSEMBLY
TO EXCITER TR RELAY (GROUND TO TRANSMIT)
HV
S2A
AC FROM PS FIG 5
RFC
S1
FIL ON/OFF
0.01
0.01
DS4
25 20W
R1
K1A
RFC
AC TO PS FIG 5
0.01
0.01
GROUND FROM PS FIG 5
T1
117V AC
5V AC
TO 8877 HEATER
BLOWER
M
R3
FIL VOLT ADJ
+12 V DC
POWER SUPPLY MODULE
+26 V DC
AC COMMON FROM PS FIG 5
RFC
0.01
0.01
D3
C1
150
R2
K1
100µF 50 V
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF); RESISTANCES ARE IN OHMS; k = 1000, M = 1000 000.
U3 REG
7812
IN
OUT
GND
FB
+12 V DC
POWER-SUPPLY MODULE
100 2W
0.02
T=TANTALUM
T2
117 V AC
25.2 V AC
AC
U2
AC
U1 REG
LM317
IN
OUT
ADJ
FB
+26 V DC
D7
270
1500µF 75 V
5 k 1/4 W
VOLTAGE ADJUST
0.02
0.1 T

Fig 2—Amplifier control circuit and low-voltage power-supply schematic diagram. Part numbers shown in parentheses are Radio Shack.

B—Blower, Dayton 4C004-i.
D1-D6—Diode, 1 kV, 2.5 A.
FB—Ferrite bead.
K1-K4—4PDT 24-V dc relay, Potter & Brumfield KHU17D11.
K5—SPDT vacuum relay, 26-V dc coil.
Q1,Q2—2N3053 NPN transistor.
Q3—TIP31 NPN transistor (276-2017).
R1—25 Ω, 20 W.
R2—150 Ω, 2 W.
R3—25 Ω, 25 W variable.
RFC—10 turns no. 14 enam wire on ¼-in-diam ferrite rod.
S1,S3—Alco 16TL5-11 SPST.
S2—Alco 16TL5-22 DPDT.
S4—Alco 16TZ pilot light.
Ti—Filament transformer, 5.0 V ac, 10 A, Peter Dahl Co.
T2—25.2 V ac, 1.0 A, Stancor P6469.
U1—50-V, 4-A bridge rectifier.

An effective ALC circuit, adjustable from a front-panel control, is included to avoid overdriving the tube. This feature is essential in this amplifier because the drive requirement is only about 80 W for 1500-W output. The ALC circuit samples the RF drive level through a 27-pF mica capacitor to generate a dc voltage that is fed back to the exciter for drive-power control.

The grid-trip-protection circuit shuts down the amplifier if grid current exceeds 120 mA. This protects the tube from tuning errors or other problems such as losing the antenna, or a tube flashover during operation. Although grid current flows through all paths from ground to the B – line, most of the grid current goes through R1. The current passing through R1 develops a voltage drop. For example, if 100 mA of grid current is drawn through R1, 1 volt is developed (E = IR = 0.100 x 10). This voltage is used to turn on the transistor switch, Qi. When Qi turns on, the grid-trip relay, K3, energizes and opens the grid trip break in the RF relay control line to shut the amplifier down. R2 sets the current level at which Qi turns on. The front-panel GRID TRIP lamp goes out if the trip circuit is activated. The switch is reset by pressing S3.

The plate tank circuit uses a pi-L configuration because this design provides approximately 20-dB better harmonic suppression than the conventional pi design. The TUNE and LOAD capacitors are vacuum variable types to minimize space requirements and also optimize performance on 12 and 10 meters where small capacitance values are needed to achieve an acceptable tank-circuit Q. The 10- to 40-meter tank coil is homemade from ¼-inch copper tubing that is silver plated to minimize skin resistance. The 80-meter, 160-meter and L-coils are toroid designs to minimize space. Using a toroid for the L coil also helps isolate the L network from the rest of the tank circuit because of the toroid's self-shielding characteristics.

Metering circuits monitor plate and grid current, as well as filament voltage. Plate current is monitored by placing a meter in series with the B – line. Therefore, only a small dc voltage is across the meter. An additional position can be included on the FIL/GRID meter for plate voltage, but one is not shown in this design because a separate high-voltage meter is included in the power supply. It would be a good idea to include a high-voltage scale on the meter in case the RF deck is ever used with a different highvoltage supply. Grid current is monitored by measuring the voltage drop across R1. R3 is adjusted to give the cor-

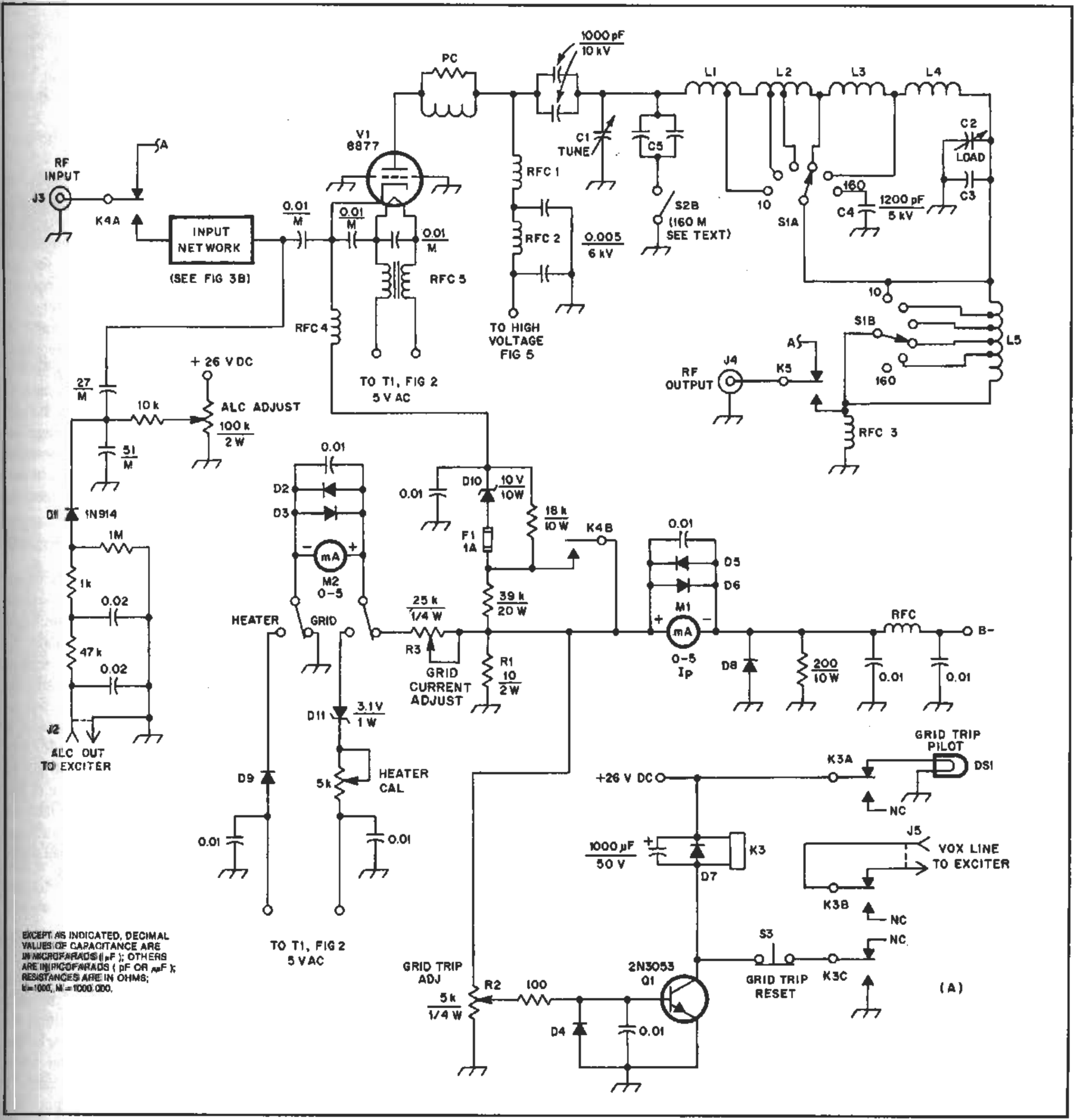
1000 pF
10 kV
PC
L1
L2
L3
L4
C2
LOAD
C3
V1
8877
C1
TUNE
C5
RFC 1
RF
INPUT
J3
A
K4A
INPUT
NETWORK
(SEE FIG 3B)
0.01
M
0.01
M
0.01
M
RFC 5
RFC 2
0.005
6 kV
S2B
(160 M
SEE TEXT)
10
S1A
160
1200 pF
5 kV
C4
10
S1B
L5
TO HIGH
VOLTAGE
FIG 5
RFC4
TO T1, FIG 2
5 V AC
+ 26 V DC
27
M
10 k
ALC ADJUST
100 k
2 W
51
M
RF
OUTPUT
J4
K5
A
160
RFC 3
0.01
D2
D3
M2
0-5
mA
0.01
D10
10 V
10W
18 k
10 W
F1
1A
K4B
0.01
D5
D6
D1
1N914
1M
1k
0.02
47 k
0.02
HEATER
GRID
25 k
1/4 W
R3
GRID
CURRENT
ADJUST
39 k
20 W
R1
10
2 W
M1
mA
0-5
Ip
D8
200
10 W
RFC
B-
0.01
0.01
J2
ALC OUT
TO EXCITER
D11
3.1 V
1 W
D9
5 k
HEATER
CAL
0.01
0.01
GRID TRIP
PILOT
K3A
DS1
NC
+26 V DC
1000 µF
50 V
K3
D7
J5
VOX LINE
TO EXCITER
K3B
NC
S3
NC
TO T1, FIG 2
5 VAC
GRID TRIP
ADJ
5 k
1/4 W
R2
100
2N3053
Q1
GRID TRIP
RESET
K3C
(A)
D4
0.01

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k = 1000, M = 1000 000.

Fig 3—RF amplifier schematic diagram (see Part B on next page).

C1—Vacuum variable capacitor, 375 pF, 10 kV.
C2—Vacuum variable capacitor, 1000 pF, 10 kV.
C3—Mica transmitting capacitor, 100 pF, 5 kV.
C4—Mica transmitting capacitor, 3 × 400 pF, 5 kV.
C5—Fixed vacuum capacitor, 2 × 80 pF, 20 kV.
D1—Diode, 600 V, 1 A.
D2-D9—Diode, 1 kV, 2.5 A, HEP 170.
D10—Zener diode, 10 V, 1 W.
D11—Zener diode, 3.1 V, 1 W.
K1-K7—DPDT 12-V dc DIP relay (275-213).
K8—SPST 12-V dc relay (275-241).
RFC—10 turns no. 14 enam wire on 1/4-in-diam ferrite rod.
RFC1—Plate choke, 2 A, Peter Dahl Co.
RFC2—Air-wound coil, 15 turns, 1/2-in diam.
RFC3—Choke, 1 mH, 800 mA.
RFC4—110 turns no. 20 enam wire on 1/2-in diam fiber rod.
RFC5—Filament choke, 18 bifilar turns no. 14 enam wire on 1/2-in-diam ferrite rod, 6 inches long.
PC—Three 150-ohm, 2-W carbon resistors in parallel with 2-inch horseshoe loop of 1/2-inch silver-plated strap.
L1-L5—See Table 2.
M1,M2—Simpson Wide-Vue panel meter, 01253 bezel and 01165 lighting kit (See text).
S1—9-position, 2-pole switch, Radio Switch model 88, 13-kV, 30-A.
S2—Solenoid-controlled switch; see text.
S3—SPST normally closed momentary switch, Alco 16TL-11 with 6T-2 red lens.

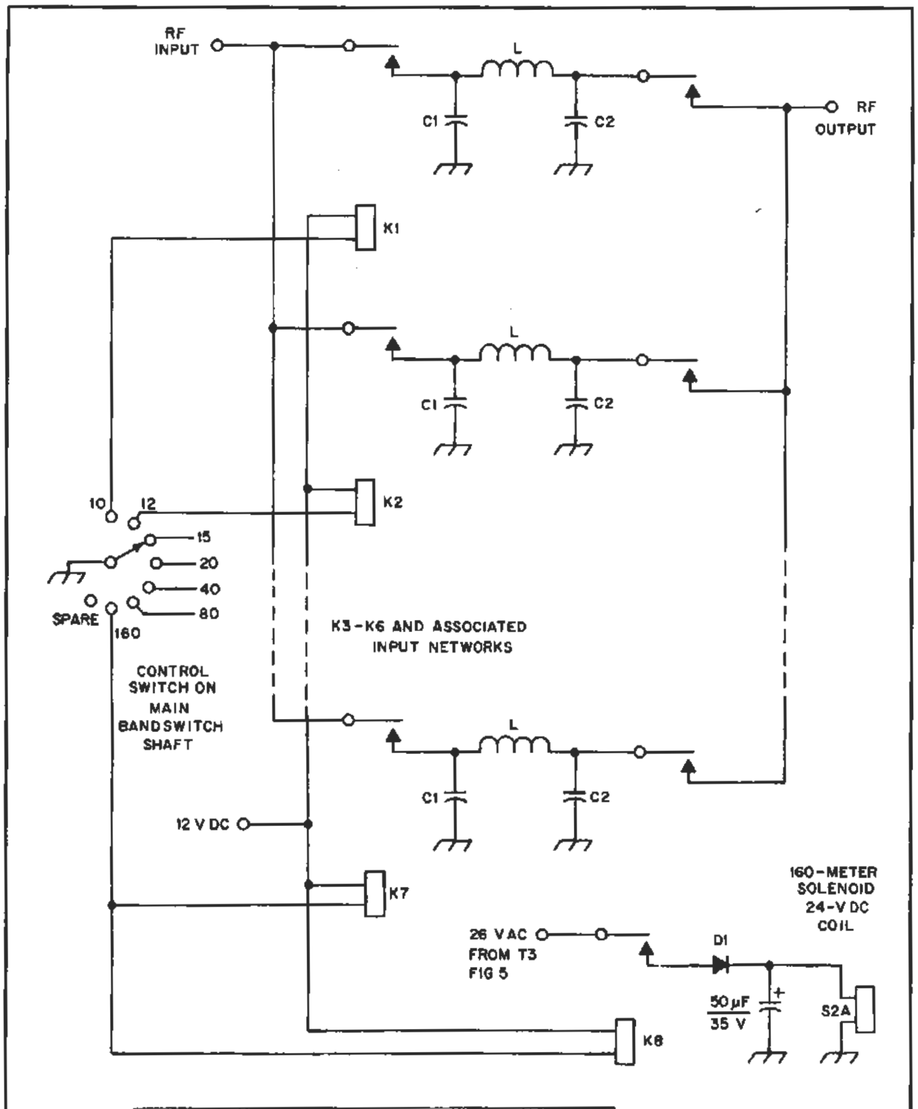
RF INPUT
L
C1
C2
RF OUTPUT
K1
K2
10
12
15
20
40
80
SPARE
160
CONTROL SWITCH ON MAIN BANDSWITCH SHAFT
K3-K6 AND ASSOCIATED INPUT NETWORKS
12 V DC
K7
26 VAC FROM T3 FIG 5
D1
50µF 35 V
S2A
K8
160-METER SOLENOID 24-V DC COIL
(B)

BAND	C1 (pF)	L (µH)	C2 (pF)	Q	CORE	NO. OF TURNS	WIRE SIZE
160	4000	3.31	4000	2.35	T94-2	19	22
80	1240	2.06	1240	1.50	T94-2	15	22
40	680	1.03	680	1.67	T68-2	14	18
20	390	0.50	390	1.80	T68-2	9	18
15	270	0.32	270	1.90	T68-2	7	18
12	270	0.25	270	2.20	T68-2	6	18
10	160	0.27	160	1.50	T68-2	6	18

Fig 3B—see previous page.

rect gridcurrent meter reading. Filament voltage is measured by converting the ac voltage to dc and displaying the dc voltage on M2. The 3.1-V Zener diode expands the meter scale by not allowing conduction until the voltage reaches 3.1 V.

A vacuum relay is used for the amplifier output. The relay is small in size, quiet and capable of handling large RF currents.

RF Deck Parts Selection

Finding all the parts for the RF deck is a major task. If you are planning to build an amplifier, begin collecting parts as soon as possible. It is the first step because the physical layout of the amplifier will depend on the components available. Don't try to exactly duplicate the components I used. For example, vacuum variable capacitors come in many different shapes and sizes, with different mounting provisions. Actually, the parts you find may be better than the parts used in my RF deck. As an example, a 1500-pF vacuum variable LOAD capacitor would be much better than the 1000-pF unit I used.Therefore, use whatever resources you have available to acquire the parts— but a word of caution! Do not compromise too much when gathering components. If you cannot find what you need on the surplus market, buy the parts new. It may cost a little more, but if the project is not done right, you will never be happy with the final result.

Vacuum Variable Capacitors

Vacuum variable capacitors are often difficult to locate at reasnable prices. Plan to spend about $50 for the TUNE capacitor and $75 to $100 for the LOAD capacitor if vacuum capacitors are used. The TUNE capacitor should be at least 300 pF at 7 kV, and the LOAD capacitor should be at least 1000 pF at 3 kV.

An air variable capacitor can be used for the LOAD control, if desired. The minimum capacitance for the LOAD capacitor is 112 pF for 10 meters, which is not difficult to obtain with an air variable type. A rating of 1 kV, minimum, is recommended. However, it is a different story for the TUNE capacitor. The minimum required capacitance is 26 pF. The direct inter-electrode capacitance of the 8877 tube in grounded-grid service is 10 pF; therefore, the TUNE capacitor must have a minimum value of not more than 16 pF for 10 meters.

This is nearly impossible with a 300-pF air variable. In addition, the voltage requirements for the TUNE capacitor make any air variable rather large. For these reasons, a vacuum variable is recommended for the TUNE capacitor.

Meters

Good-quality meters with bezels are essential for good appearance. The Simpson Wide-Vue® meters I used were purchased at a hamfest. The bezels were ordered directly from Simpson because they seldom appear on the surplus market. Actually, almost any meter movement can be used, so don't pass up a good meter just because it reads 50 V or 100 mA on the scale. Any meter with a movement from 100 µA to 5 mA can be used. This allows use of approximately 90% of the meters available on the surplus market. I will give instructions later for calibrating any meter to read whatever current or voltage is required.

RF Band Switch

Good RF band switches are very difficult to locate. More problems are experienced with arcing band switches than with any other amplifier component. If the band switch selected has insufficient voltage insulation, it will arc to the wiper rotor on the high-impedance 10-meter position when operating on the lower-frequency bands. I obtained the band switch for my amplifier from Radio Switch Corp.[6] The model 88 switch is a 2-pole, 9-position unit with a 13-kV peak flashover/30-A contact rating. This switch will not arc! Its list price is currently $107, and it is well worth the money!

Miscellaneous Parts and Materials

Many of the small parts (capacitors, relays and resistors) can be purchased at Radio Shack. Their parts selection is good, and continues to increase. You can usually find a store around the corner in almost any city. Pioneer Electronics is also a good source for commercial-grade components.[7] Good-quality PC-board material can be found at almost any hamfest. Don't compromise here—use G10 glass-epoxy board. As for coils? Make them. Complete "how-to" instructions are given later.

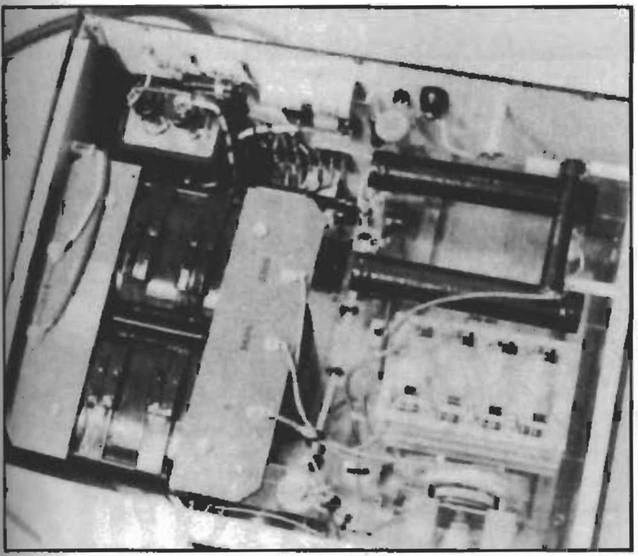

Fig 4—High-voltage power supply top interior view.

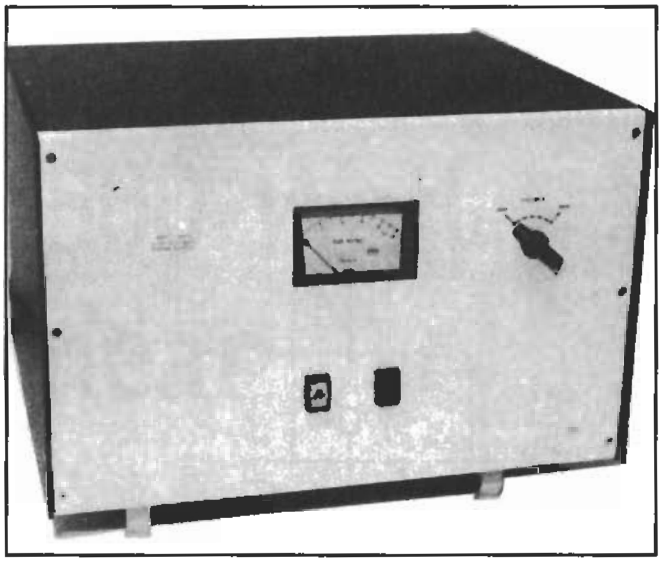

High-Voltage Power-Supply Circuit Description

The key to continuous duty in a high-power linear amplifier is the power supply. It must be able to deliver the required voltage and current on a continuous basis. Power supplies are usually the limiting factor in commercial linear amplifiers.

A WORD OF CAUTION IS IN ORDER. The power supply is a very dangerous piece of equipment! Give it proper respect. One mistake can be fatal. Use proper precautions in the construction and testing of this unit, and be careful to build a safe unit.

I recommend that the power supply be built first. The construction is not complex and it can serve as a training ground for amplifier building techniques, particularly for the first-time builder.

Power-Supply Design

The power supply is shown in Fig 4, and the schematic diagram is shown in Fig 5. The hypersil power transformer has a 234-V ac primary, and a 3300-V ac secondary that is tapped at 2600 V. This selection of two output voltages allows for a high- and lowpower capability. An alternative to this approach is to include a Variac® or Powerstat® autotransformer on the transformer primary.

The primary circuit of the power transformer includes a step-start circuit to protect the diode bank during the initial charge of C1, the 53-µF filter capacitor, when the power supply is turned on. Two 50-ohm, 25-W resistors, one in each leg of the primary, are shorted by time-delayed relays approximately 3 to 4 seconds after application of power. The more current drawn through the resistors at start up, the more voltage drop realized and this, in turn, protects the diode bank. The delay is provided by the time constant of the 500-ohm resistor and 100-µF capacitor. The relays must be dc types. Those I used have 90-V dc coils which allows power to be supplied from one 117-V leg of the primary. If 90-V relays can't be obtained, 24-V dc relays can be substituted. A 24-V dc power source must be provided if this is done.

The rectifier unit is a full-wave bridge with eight diodes in each leg. A 470-kilohm resistor and a 0.01-µF, 1-kV capacitor are wired in parallel with each diode to equalize the voltage and protect the diodes from voltage spikes.

The power supply is controlled remotely from the RF deck. A test switch has been incorporated to allow the supply to be energized without the RF deck. A shorted Cinch-Jones plug must be inserted into a socket in the rear of the supply for test switch S1 to operate.

Two pilot lights are mounted on the front panel. One pilot light is on whenever 234 V ac is present in the supply. The other lights when the power supply is activated.

A high-voltage meter is included on the front panel. The metering is done across a 25-ohm, 5-W resistor in series with the bleeder resistor. This voltage divider keeps the total high voltage off the meter. A 50-ohm, 50-W resistor in series with the high-voltage B+ circuit protects the tube and power supply from any current surge resulting from a tube flashover or other cause. In addition, a 0.6-ohm, 1-W resistor in series with the B+ line acts as a fuse resistor. A large current surge will cause the resistor to explode—an inexpensive protection device should a problem occur.

High-Voltage Power Supply Parts Selection

Transformer

It is important to find a good power transformer that can provide the proper operating voltages for the tube. Remember that some voltage drop will occur when current is drawn from the transformer. The voltage drop depends largely on the quality of the transformer (core and wire size), and can range from 200 V to over 1 kV. The transformer should have a 234-V ac primary. Transformers with 117-V primaries are usable only if two identical units can be wired in series to provide a 234-V primary. The secondaries can by wired in series or parallel, depending on the voltage requirements. Remember that the transformers must be identical.

The required transformer secondary voltage depends on the final voltage requirement of the tube and the power-supply circuitry. If a bridge rectifier is used, the power-supply high voltage will be about 1.4 times the secondary voltage. If a voltage doubler is used, the high voltage will be about 2.8 times the secondary voltage. A voltage doubler requires two filter capacitors, or more, so if a single oil-filled filter capacitor is to be used, the design can't be a voltage doubler. The *ARRL Handbook* contains circuits for both types of power supplies.[8]

The power-handling capability of a transformer can usually be estimated by its weight. As a rule, the heavier the transformer, the greater the power capability. The transformer for a 1500-W, continuous-duty amplifier will weigh 60-80 lb. The transformer used in this power supply was obtained from Peter Dahl Co.[9] The hypersil design provides a good ratio of power capability to size and weight. I have used several Peter Dahl transformer designs in the past and found them to be of excellent quality and reasonably priced.

Filter Capacitor

Enough filter capacitance is required to obtain good voltage regulation. What is enough? I have used as little as 18 µF and as much as 100 µF in power supplies. The required capacitance can be obtained with a single oil-filled capacitor or with a series string of computer-grade electrolytics. Either way, I recommend at least 25 µF be used, with at least a 10% voltage safety factor. The filter capacitor used in this power supply is a single oil-filled unit rated at 53 µF at 5kV dc. The capacitor was obtained

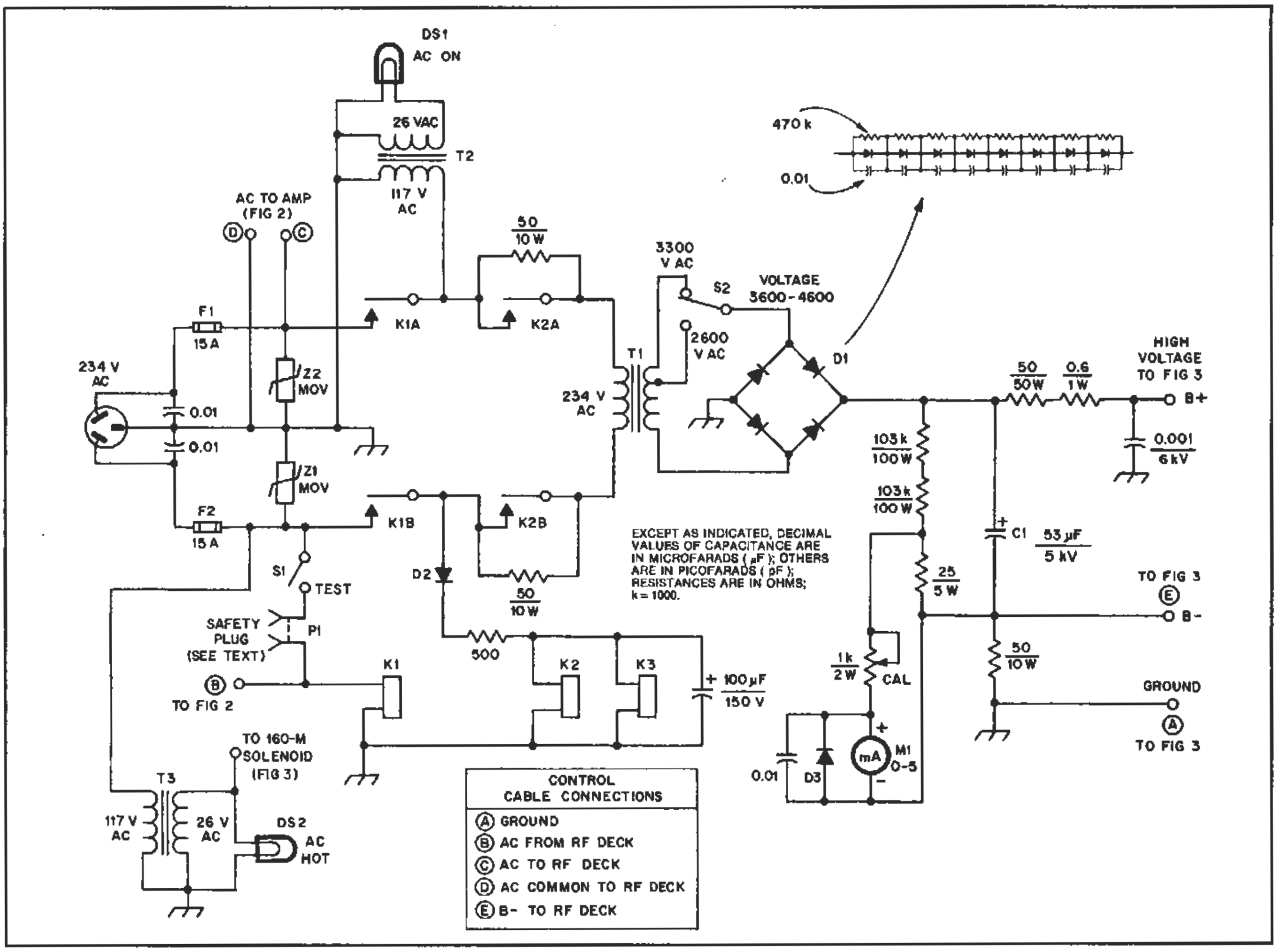
DS1
AC ON
26 VAC
T2
117 V
AC
470 k
0.01
AC TO AMP
(FIG 2)
Ⓓ
Ⓒ
50
10 W
3300
V AC
S2
VOLTAGE
3600-4600
F1
15 A
K1A
K2A
2600
V AC
T1
D1
HIGH
VOLTAGE
TO FIG 3
234 V
AC
Z2
MOV
234 V
AC
50
50W
0.6
1W
B+
0.01
0.01
0.001
6kV
103k
100W
Z1
MOV
103k
100 W
F2
15 A
K1B
K2B
C1
53 µF
5 kV
S1
TEST
D2
25
5 W
TO FIG 3
Ⓔ
B-
50
10 W
SAFETY
PLUG
(SEE TEXT)
P1
500
K1
K2
K3
+ 100 µF
150 V
1k
2W
CAL
50
10 W
GROUND
Ⓑ
TO FIG 2
Ⓐ
TO FIG 3
TO 160-M
SOLENOID
(FIG 3)
M1
0-5
mA
0.01
D3
T3
117 V
AC
26 V
AC
DS2
AC
HOT
CONTROL
CABLE CONNECTIONS
Ⓐ GROUND
Ⓑ AC FROM RF DECK
Ⓒ AC TO RF DECK
Ⓓ AC COMMON TO RF DECK
Ⓔ B- TO RF DECK

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF); RESISTANCES ARE IN OHMS; k = 1000.

Fig 5—High-voltage power supply schematic diagram. Part numbers in parentheses are RadioShack.

C1—Oil-filled capacitor, 53 µF, 5 kV, Peter Dahl Co.
D1—Diode bridge rectifier assembly, Peter Dahl Co, see text.
D2,D3—Diode, 1 kV, 2.5 A.
F1,F2—Fuse, 15 A.
K1—2PDT mercury plunger relay, Dayton 6X598-3.
K2,K3—SPDT relay, Potter Brumfield PRD1 DYO/9OVDC.
M1—High-voltage meter, 3½-inch Simpson Wide-Vue, 01253 bezel and 01165 lighting kit.
Z1,Z2—MOV transient suppressor, 117 V ac (276-568)
P1—Two-pin Cinch-Jones socket and plug (274-201 and 274-202)
P2—Eight-pin Cinch-Jones connector.
S1—SPST switch (275-690)
S2—Modified 6PST switch, Fair Radio Sales.
T1—Power transformer, 2600/3300-V ac sec, Peter Dahl Co.
T2,T3—Transformer, 26 V ac, 300 mA (273-1386)
Miscellaneous
Pilot lamp—Alco 16TZ, 6T-4 (yellow) and 6T-2 (red) lenses.
Cabinet—CTS model MCLS 10-17-14 black and white, SPP 10-14 black side panels.

from Peter Dahl Co, and is physically very small for the voltage and capacitance rating.

Diode Bridge Rectifier

The full-wave, diode-bridge rectifier is made up with 1000-PIV diodes rated at 3 A. The unit is a commercial module sold by Peter Dahl Co. Each diode string is built on a separate glass-epoxy board. The module is supplied with 1-inch angle brackets on each end, but because of space restraints, the angle was removed and the module was mounted in a vertical position using two Nylon bolts.

Should you decide to build the rectifier assembly, use good-quality diodes, such as HEP-170s or 1N5408s. Be sure to parallel each diode with a 470-kilohm resistor and a 0.01-µF, 1-kV capacitor.

High- Voltage Switch

The transformer has two taps on the secondary to provide a high- and low-voltage capability. The front-panel VOLTAGE 3600-4600 switch is fabricated from a 6-position, heavy-duty ceramic switch (Radio Switch Corp p/n 65). The switch detent and all but the second and fifth contacts are removed. New stops are fabricated from glass-epoxy board. Full high voltage appears across this switch, and therefore, it must be well insulated. The switch is mounted on two pieces of ¾-inch Plexiglas® to provide 2-inch spacing from any chassis or panel ground. A fiber shaft protrudes from the switch through the front panel. To protect the contacts, this switch must never be actuated when the power supply is on.

Construction Details

Next month, I will describe the unique construction details for building this high-power linear amplifier and power supply. In the meantime, should you be so inclined, get out there and find the parts! Remember that you should build the power supply first, so concentrate on those components.

Notes

[1] J. L. Pittenger, "Modular Linear Amplifier for the High-Frequency Amateur Bands," *Ham Radio*, Jan 1981.

[2] J. L. Pittenger, "A 3CX800A7 Linear Amplifier," *Ham Radio*, Aug 1984.

[3] J. L. Pittenger, '3CX1200A7 10 to 80 Meter Amplifier," *Ham Radio*, Aug 1985.

[4] W. I. Orr, ed, Radio Handbook (Indianapolis: H.W. Sams, 1975), 20th edition, Sect 22-4, "A Modern 3-1000Z Linear Amplifier for 80-10 Meters."

[5] Ham Trader Yellow Sheets, PO Box 2057, Glen Ellyn, IL 60138-2057.

[6] Radio Switch Corp, Rte 79, Marlboro, NJ 07746, tel 201-462-6100.

[7] Pioneer Standard Industrial Electronics, 1900 Troy St, Dayton, OH 45404, tel 513-236-9900.

[8] M. Wilson, ed, *The 1986 ARRL Handbook* (Newington: ARRL, 1985).

[9] Peter Dahl Transformer Co, 4007 Fort Blvd, El Paso, TX 79930, tel 915-566-5365.

By Jerry Pittenger, K8RA **From *QST*, October 1986**

An All-Band, 1500-Watt-Output 8877 Linear Amplifier

Part 2—Here's what you've been waiting for: detailed instructions for building a 1500-W RF deck and power supply.

Last month I described the circuitry and parts required for the 8877 linear amplifier and high-voltage power supply. This month I will cover the construction of both units. The power supply construction details are given first. Assuming that we have all the required parts in hand, we can determine the physical design of the power supply. To avoid costly mistakes, it is important to do adequate upfront planning before the first hole is drified.

Power Supply Construction

Meter Selection and Labeling

Any meter with a movement from 100 µA to 5 mA can be used for the high-voltage meter. This meter measures the voltage across the 25-ohm resistor at the bottom (B–) end of the bleeder string (see Fig 5 in Part 1). The more sensitive the meter, the higher the resistance setting of the variable 1-kilohm calibration resistor. The setting of the resistor is a simple Ohm's law problem. The maximum power-supply voltage on the meter scale causes a current equal to the meter-movement rating to flow through the meter, and thus give a full-scale reading. For example, the maximum scale on the meter in this supply is 5 kV dc. The meter has a 5-mA movement; therefore, if the supply is at 5 kV, 5 mA must flow through the calibration resistor and the meter for a fullscale reading. Looking at the complete bleeder string, the total resistance is the sum of the two 103-kΩ bleeder resistors plus the 25-ohm resistor, or 206,025 ohms. At 5 kY, approximately 24 mA (5000/206,025) flows through the string. The value of the calibration resistor, therefore, was selected to allow 5 mA through the meter and 19 mA through the 25-ohm resistor. Using Ohm's law, the value of the resistor should be approximately 95 ohms (19 mA x 25 ohms/5 mA). The required wattage rating of the resistor is 0.23 W ($I^2R = 0.005 \times 0.005 \times 95$), and a 1-kΩ, 2-W potentiometer was used. The meter used originally had a 0-50 scale. The scale was changed to read 0 to 5 kV.

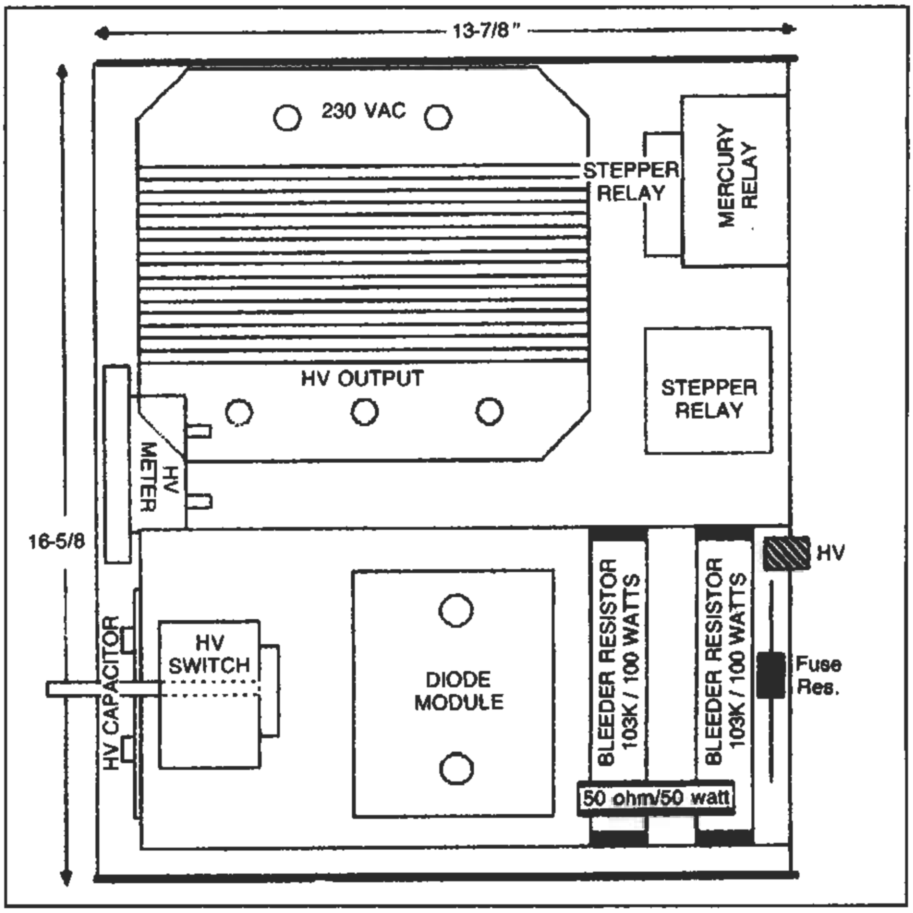

Fig 6—Typical computer-prepared layout drawing. Some component labels have been enlarged for legibility.

Panel/Chassis Layout

I do all my layouts with an Apple® Macintosh computer. An example of this layout is shown in Fig 6. A manual method may be used, instead. Cut out a piece of poster board the size of the front panel or chassis, as well as all the major components. Shuffle the pieces until you get an acceptable layout. This may seem like a lot of extra work, but "one picture's worth 1000 words." You will save time in the long run by going through this procedure, and not have to correct errors that otherwise are certain to occur. The panel layout should be symmetrical. Align switches and center meters. Make sure that components are properly spaced to accommodate the physical size of the parts behind the panel.

The computer-generated or manual layouts serve as a guide during construction, but are by no means sacred. Once you start putting the parts in place, you probably will make minor changes. Go ahead and make the changes, but always update the documentation.

Front-Panel Assembly

You probably will not be ready to cut metal until you are 3 to 4 months into the project. Do the front panel first, since the parts locations are fixed for symmetry. Parts behind the front panel can be moved to accommodate the front-panel design.

Cover the front panel with 3-inch-wide masking tape. The tape not only protects the panel from scratches, but also provides a way to lay out the panel with a pencil or pen. Before any holes are drilled, place all the major parts in their proper place in the cabinet to be sure that nothing obstructs the area behind the panel. Remember you have only one chance. Miss and it means a new panel, or cabinet!

Remove the front panel from the cabinet and center punch the panel where holes are to be drilled, as marked on the tape. Carefully drill a very small hole at each punch mark to serve as a guide, then cut the holes to final size. Holes up to approximately 7/16 inch can be drilled with either a hand drill or a drill press. Holes larger than 7/16 inch should be made with chassis punches. The meter hole is rectangular and is cut with a nibbler after drilling an access hole. Cut the meter hole about 1/32 inch smaller than needed and finish with a large file to straighten the edges. Be careful using the file—it is very easy to let the file slip out of the hole and make a big scratch in the panel. The meter hole doesn't have to be perfect since a meter bezel is used.

After the front-panel metal work is complete, carefully remove the masking tape. The front panel should be labeled before mounting the components. Labeling is done with dry-transfer lettering available from art stores or Radio Shack. Apply the labeling by laying the letter or figure on the panel in the proper position and rubbing over it with a soft pencil. The character will be transferred to the panel. If a mistake is made, the character can be removed with masking or Scotch® tape. There is a mattefinish spray available to protect the lettering. I do not recommend using this spray. It will peel if bumped and does not work well.

The parts can now be mounted on the panel. The cabling to the front panel is connected through nylon multipin connectors (available from Radio Shack) to allow easy panel removal. The same technique is used in the RF deck.

Rear-panel Assembly

Placement of parts on the rear panel is not as critical as on the front panel. Attention should still be paid to symmetry, however. I have found that it is best to mount rear-panel parts after the major parts have been mounted inside the cabinet. The rear panel is drilled and labeled in a manner similar to the front panel. The black rear panel requires white lettering. White lettering kits are available at art supply stores and some electronics suppliers.

Fig 7 shows the inside rear panel of the power supply. The control and power cables enter the rear panel and are routed directly to barrier strips. Each terminal is labeled for clarity. The rear panel is accessible by removing the front panel and the transformer. It sounds like a big job, but it can be accomplished in about 10 minutes. It is necessary to provide good strain relief for the cables. Immediately inside the rear panel, a piece of 3/4-inch aluminum angle stock is mounted, to which each cable is clamped. Large rubber grommets are used in the holes for cable protection.

Major Chassis Assembly

Power-supply components are extremely heavy and a good supporting structure is required. The 1/16-inch-thick bottom cover supplied with the cabinet is replaced with a 1/4-inch-thick base plate to provide an adequate foundation. The heavy plate is cut to size on a commercial metal shear. After the transformer, relays and filter capacitor are mounted, rubber-wheeled casters are bolted to the bottom so that the supply can be rolled, rather than carried from place to place.

The cabinet sides are formed by two removable panels. Parts can be mounted to the inner panel using countersunk screws and then covered with the 1/8-inch painted cover plate to provide a professional appearance. A 1/4-inch-thick sheet of Plexiglas® is mounted above the filter capacitor to support the high/low voltage switch, the diode bank and the bleeder resistors. The Plexiglas is supported by drilled and tapped holes for no. 6-32 countersunk screws in the rear and side panel. A post of 1/2-inch aluminum bar stock supports the front-left corner of the Plexiglas.

The mercury-wetted power relay is mounted on the rear panel and must be positioned vertically. Mercury-wetted relays have a tendency to buzz if mounted on a solid surface. Use a rubber grommet to make a bushing in each mounting hole, or mount the relay on a rubber pad. The stepstart relays, K2 and K3, are also mounted on 1/4-inch-thick rubber sheet to minimize noise.

Once the major components are mounted, wire them together, performing as much testing as possible along the way.

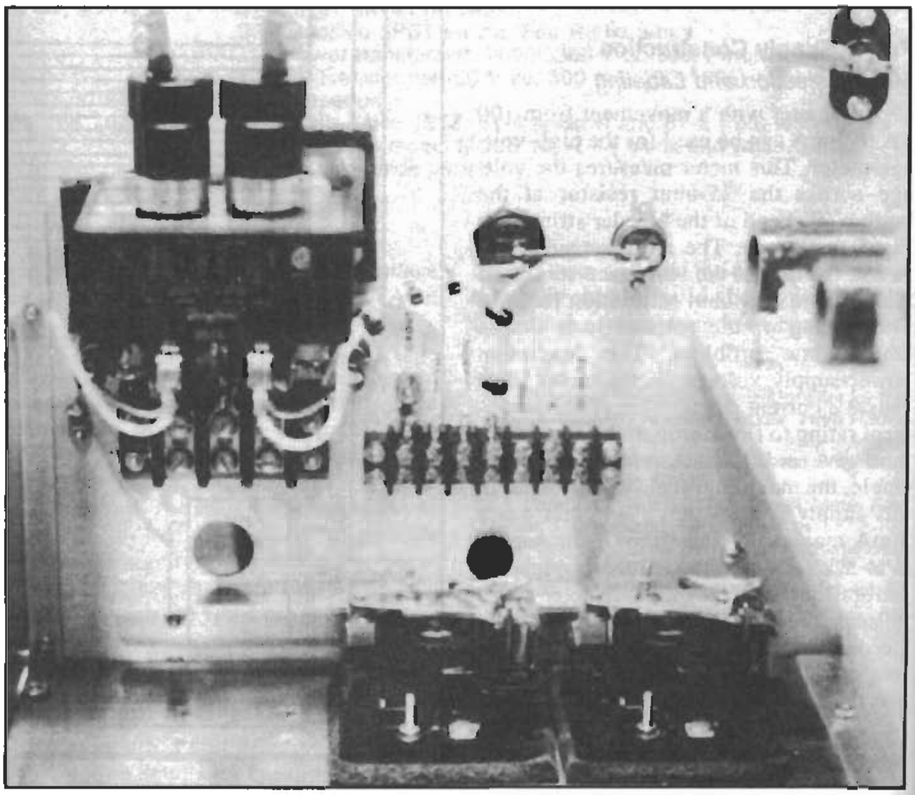

Fig 7—Interior view of rear panel. Note the mercury plunger relay mounting. Relays K2 and K3 are at the lower right.

Front- and rear-panel components are mounted and wired before those that are mounted to the cabinet sides and base plate.

Power-Supply Testing

When you are satisfied that the power supply has been correctly wired and carefully checked, go have a cup of coffee and come back later when your mind is fresh and check the wiring one more time. *Remember that this power supply can be a lethal device. One wrong move could be fatal.*

Final testing of the power supply is accomplished in three steps. Remove the wiring from the primary of the transformer and place a temporary line cord with a small variable autotransformer directly from a 117-V ac line to the primary. Turn the 117 V on and slowly run the variable autotransformer up. At 117 V, the power supply should be reading half-scale voltage. This test verifies that the diode bank and filter capacitor are correctly wired. Now is a good time to calibrate the front-panel meter. Using the variable auto-transformer, set the output voltage at a level that another VOM can measure accurately. For example, most VOMs can measure 1 kV. Adjust the calibration resistor so that the front-panel meter reads the same as the VOM.

The second test checks the primary 234-V ac circuitry. With the 234-V lines still disconnected from the primary, plug the power-supply line cord into a 234-V source. Remember that the shorted two-contact plug must be inserted into the rear-panel socket. Turn the power supply on with the test switch and listen for a 3- to 4-second time delay for the step-start relays. Check that 234 V appears across the two wires disconnected from the primary.

The final test is to try the entire power supply with 234 V applied to the primary. Check the high- and low-voltage indications on the front-panel meter.

Power-Supply Performance

The design results in a husky supply that delivers about 4.6 kV and 3.8 kV (no load) on the high- and low-voltage positions, respectively. In either position the power supply drops less than 400 V under full load. This is the performance you need to achieve full-power capability and good linearity.

RF Deck Construction

The physical design of the RF deck requires planning long before construction begins. Again, I used the Macintosh computer for my initial “paper design,” but the same process can be done using paper and pencil. The important thing is to lay out the major components so that everything fits properly before starting to drill holes and cut metal. You must have all the major components in hand before doing the physical design, so you know what you have to work with.

Fig 8 shows the scaled Macintosh designs for the front, top and bottom views of

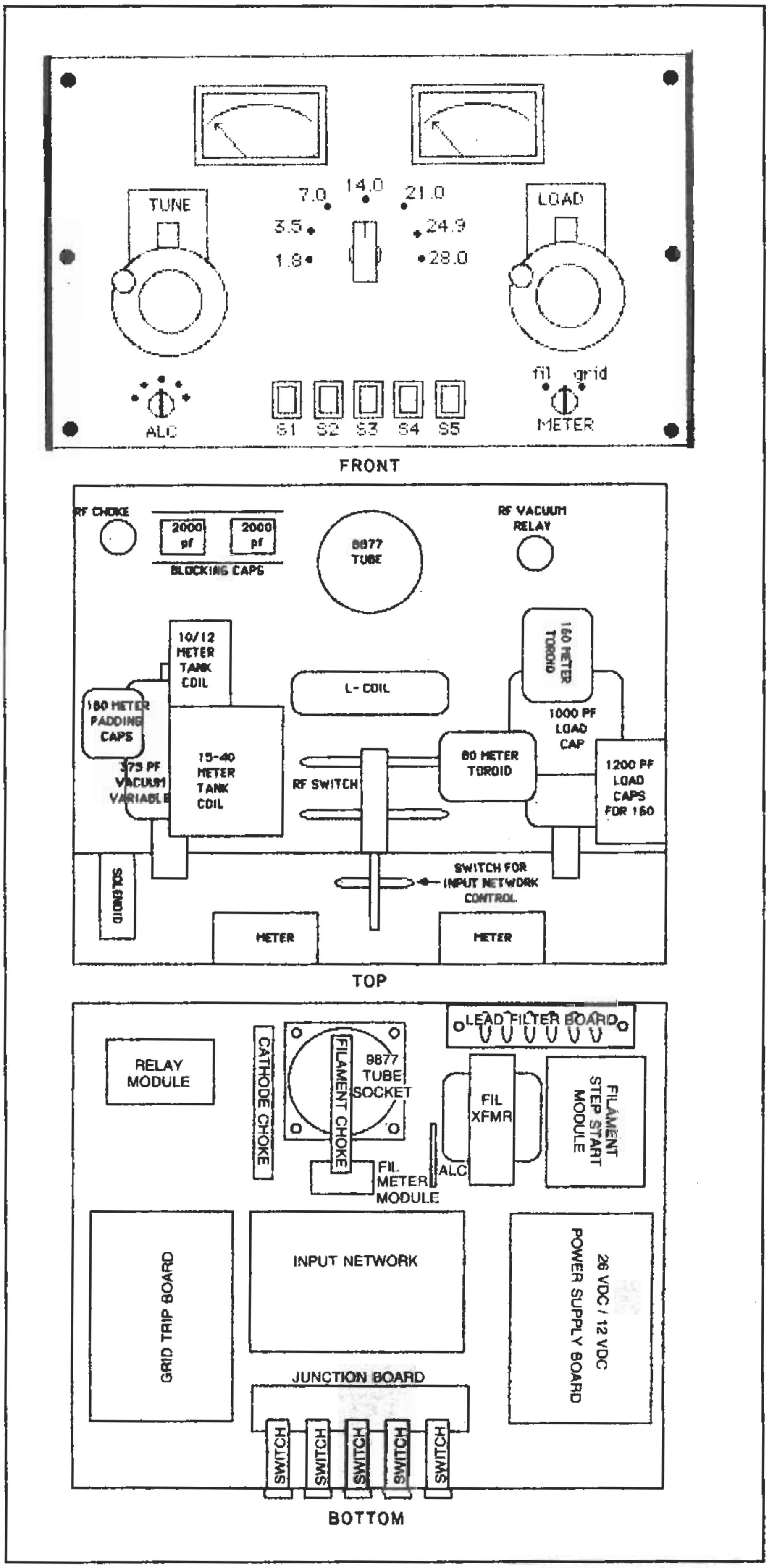

TUNE
7.0
14.0
21.0
3.5
24.9
1.8
28.0
LOAD
fil
grid
ALC
S1
S2
S3
S4
S5
METER
FRONT
RF CHOKE
2000 pf
2000 pf
BLOCKING CAPS
9877 TUBE
RF VACUUM RELAY
10/12 METER TANK COIL
160 METER TOROID
L- COIL
160 METER PADDING CAPS
1000 PF LOAD CAP
375 PF VACUUM VARIABLE
15-40 METER TANK COIL
RF SWITCH
80 METER TOROID
1200 PF LOAD CAPS FOR 160
SOLENOID
SWITCH FOR INPUT NETWORK CONTROL
METER
METER
TOP
RELAY MODULE
CATHODE CHOKE
9877 TUBE SOCKET
FILAMENT CHOKE
LEAD FILTER BOARD
FIL XFMR
FILAMENT STEP START MODULE
FIL METER MODULE
ALC
GRID TRIP BOARD
INPUT NETWORK
26 VDC / 12 VDC POWER SUPPLY BOARD
JUNCTION BOARD
SWITCH
SWITCH
SWITCH
SWITCH
SWITCH
BOTTOM

Fig 8—Scale layouts of the RF deck front, top and bottom views, as prepared on the Macintosh computer. Some component labels have been enlarged for legibility.

the amplifier. It is necessary to match the designs so that the front-panel controls end up at the right place on the panel. Everything must be drawn to scale to obtain the relative positions of the components. When designing the bottom and top layouts, only the major circuit boards and components like the grid trip, input network, low-voltage power supply, coils, capacitors and filament transformer are considered. The smaller components can be fitted in later. It is important that the unit be designed with maintainability in mind. Every component, large and small, must be accessible after the unit is completed. The amplifier was constructed in the following steps. Each major step is discussed in more detail later.

- Fabricate, build and test all printed circuit boards for the RF deck.
- Cut the holes in the front panel.
- Perform the major metal work on the subpanel and chassis plate.
- Mount the vacuum capacitors and band switch to the subpanel.
- Mount the PC boards, filament transformer and tube socket to the chassis plate.
- Wire and test the under-chassis control circuits.
- Fabricate and install the RF tank circuit.
- Label and calibrate meters.
- Complete the front panel and mate it to the chassis and cabinet.
- System test the amplifier with the power supply.

Printed-Circuit Board Fabrication

Making PC boards can be tedious, but with a little practice, good results can be obtained. Grouping interconnected circuits on the same PC board minimizes the cable harness between modules in the amplifier. In this design, nine PC board modules are required.

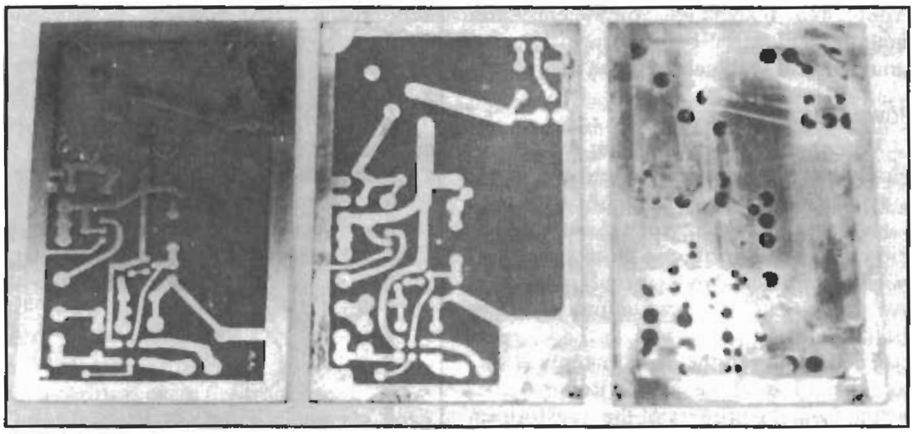

Fig 9—Things don't always work out. Three attempts were made to fabricate the grid-trip and bias PC board. The good board at left was subsequently damaged, requiring a fourth effort.

	Board Size
1) Input network	4" × 6"
2) Low-voltage power supply and timer circuit	4" × 6"
3) Grid-trip and bias circuit	4" × 6"
4) Switch bank harness interface to front panel	1" × 6"
5) Filament step-start circuit	2¾" × 3¾"
6) RF input/output relay timing circuit	2¾"
7) Line filters for control cable and blower	1¾" × 5¼"
8) Filament-voltage-meter circuit	1⅛" × 2"
9) ALC circuit	1¾" × 2¼"

All boards, except nos. 4, 7, 8 and 9, are mounted on ½-inch channel to the chassis plate. Therefore, the PC board layouts include a ground strip on the edges of each board to mount the boards to the channel.

The Macintosh computer, with a software package called "Draw," was used to lay out the circuit boards. The design of PC boards with the computer is beyond the scope of this article. PC boards can also be laid out using pencil and paper. Once the design is laid out to scale, the layout must be transferred to the board. First wash the board with a mild detergent to remove all grease and dirt. Using tape and special dry transfers (Radio Shack p/n 276-1577), copy the design onto the board. The transfers are not exact, and the hand drawing only serves as a guide. Submerge the board into an etching solution to remove the exposed copper. I use ferric-chloride etchant. A flood lamp over the etching tray warms the solution and speeds the etching process. Agitate the solution occasionally. The final result should be a nicely etched board ready for drilling the holes for mounting parts. Believe me, however, it doesn't always go as planned.

Fig 9 shows three tries at making the grid-trip and bias board. Actually the third board, on the left, was damaged, and a fourth was necessary. However, once you get the hang of it, it usually goes well.

After etching the board, polish it with

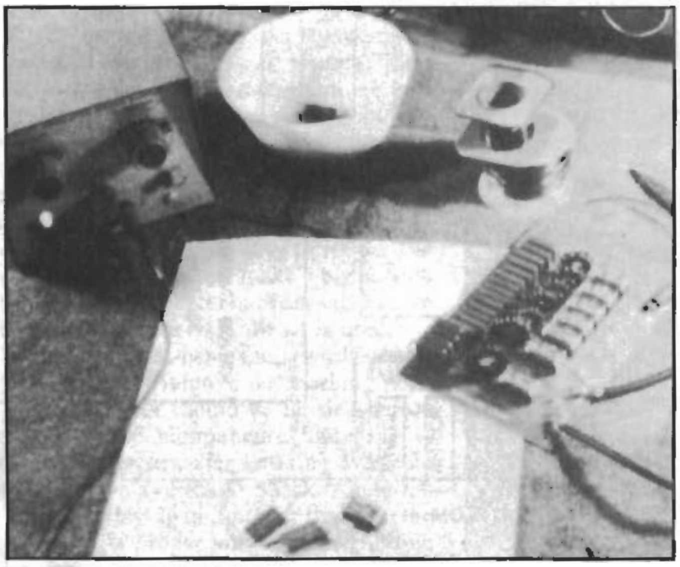

Fig 10—In-process testing of the input-network PC board. Each coil was trimmed to produce a 1:1 SWR into a 50-ohm load. Note the power supply required to actuate relays.

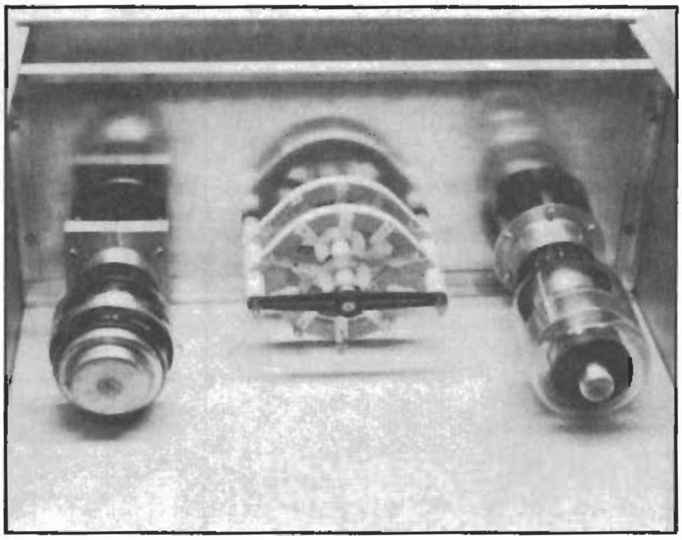

Fig 11—Subpanel installation in the cabinet. Note position of chassis plate, vacuum variable capacitors and band switch. Aluminum angle stockis used to mount the subpanel to the chassis plate and cabinet sides.

fine steel wool and drill the component mounting holes with a small drill. Use dry transfers to label the connections required to the board, before mounting parts. This will avoid wiring mistakes when the wiring harness is installed. Finally, mount and solder the components onto the board.

Test each board as much as possible before final assembly. For example, Fig 10 shows the input network being tested by hooking up coaxial cable and running 100 W through each section into a 50-ohm dummy load. The coils were adjusted at this time to give a flat 1:1 SWR. A power supply is needed to energize the relay for the pi-section being tuned.

Front-panel Fabrication

Cover the front panel with masking tape to protect it from scratches, and mark the mounting positions of the components. Take your time with this step since it is hard to recover if an error is made. Mark the front panel where the holes are to be cut and carefully drill a small pilot hole followed by the correct hole size. The holes for the band switch, capacitor control shafts, ALC control and multimeter switch are 3/8-inch diameter and can be made with a regular drill bit. The holes for the power switches are 5/8-inch diameter and require a chassis punch. The meter-mounting holes are the most difficult to make. The rectangular cutouts are marked about 1/32-inch smaller than required to protect against overcutting. Use a nibbler to cut the holes carefully, then file them to the exact size of the meter bezels. Be careful that the file doesn't slip and scratch the panel (disaster!). Using a large file will help avoid this tragedy.

I recommend doing all the front-panel metal work at one time to ensure proper component layout. When the metal work is complete, leave the masking tape on the panel for protection as it will be used as a template to locate the parts mounted behind the front panel.

Chassis and Subpanel Metal Work

This amplifier design has very little metal work that requires more than a hacksaw and a file. The chassis plate was purchased with the cabinet. The only other panel, except for front and rear panels, is the subpanel that mounts perpendicular to the chassis plate, and 3 inches behind the front panel. The subpanel shields the meter compartment from RF and serves as a mounting support for the band switch, vacuum capacitors and 80-meter toroid coil.

The chassis is mounted 3½ inches from the bottom of the cabinet to allow room under the chassis plate for the filament transformer. Therefore, the subpanel is cut to 6¾ × 16½ inches. Referring to Fig 11, ½-inch aluminum angle is attached with screws to each edge of the subpanel to provide a mounting flange to the chassis plate and cabinet side walls. The top piece of angle stock provides a mounting surface for a piece of gold-plated finger stock that seals the subpanel to the cabinet top plate.

Two large holes are required. A 3-inch-diameter hole is required in the chassis plate for the tube socket. A 5-inch-diameter hole is cut in the cabinet top plate and aligned directly above the tube socket to vent the air flowing from the tube and chimney. Use a large fly-cutter, available at most hardware stores. For safety, the fly-cutter should be used only on a drill press—never with a hand drill. Therefore, if you don't have a drill press, find a friend who has one. Before cutting the large holes, use 3-inch masking tape to cover the chassis plate and top panel to mark where the holes are to be drilled, and to protect the surfaces. For safety, clamp the panels onto a board and the base of the drill press before drilling.

Cut a piece of perforated aluminum stock to be slightly larger than the hole in the cabinet top panel. Clean the perforated metal well, and spray with paint to match the cabinet color. Fasten the perforated piece to the inside top panel with several small countersunk screws painted to match the cabinet.

Mounting Vacuum Variable Capacitors and Band Switch

The vacuum variable capacitors and band switch are mounted to the subpanel, but their control shafts must be aligned with the front panel design. Allow enough slack in the positioning of these components to perform precise alignment with the front-panel holes when the front panel is installed. To mark the hole positions on the subpanel, slide the subpanel against the rear of the front panel while both the front panel and chassis plate are bolted into place in the cabinet. Cover the subpanel with masking tape and mark the exact centers of the holes for the components. Drill the two holes in the subpanel for the vacuum variable capacitors with hole saws. Mount the vacuum variable capacitors and band switch on the subpanel.

Mounting Major Components to the Chassis Plate

The major components (PC boards, filament transformer and tube socket) are mounted toihe chassis plate as shown in Fig 12. The PC boards are first mounted to 1/2-inch aluminum channel using sheet-metal screws for easy removal. Cover the bottom of the chassis plate with masking tape. Position and mark the PC boards, filament transformer and tube socket according to the planned physical layout. Drill the mounting holes in the chassis plate, then redrill the holes from the top with a countersink bit to allow flat-head countersunk screws to be used for mounting. This retains the flat surface on top of the chassis plate. With the 1/2-inch channels on the PC boards, mark the hole positions on the bottom of each channel with a pencil, using the predrilled chassis-plate mounting holes as a template. Remove the channels from the PC boards, mount the channels on the chassis plate, and remount the PC boards on the channels.

Under-Chassis Wiring

Complete the under-chassis wiring, according to the schematic diagram, using Teflon®-insulated wire. The nylon connectors near the front of the chassis connect to the front-panel power switches, the ALC potentiometer and the multimeter switch. This allpws the front panel to be easily removed for rear-panel access. The small board in the front center of the chas-

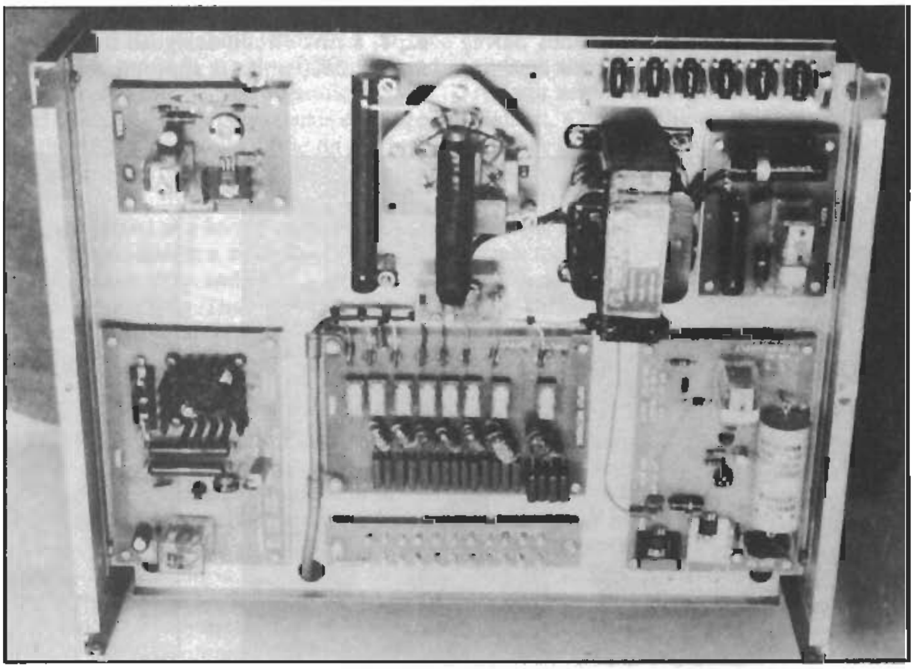

Fig 12—Bottom view of the RF deck showing placement of PC boards, transformer and tube socket.

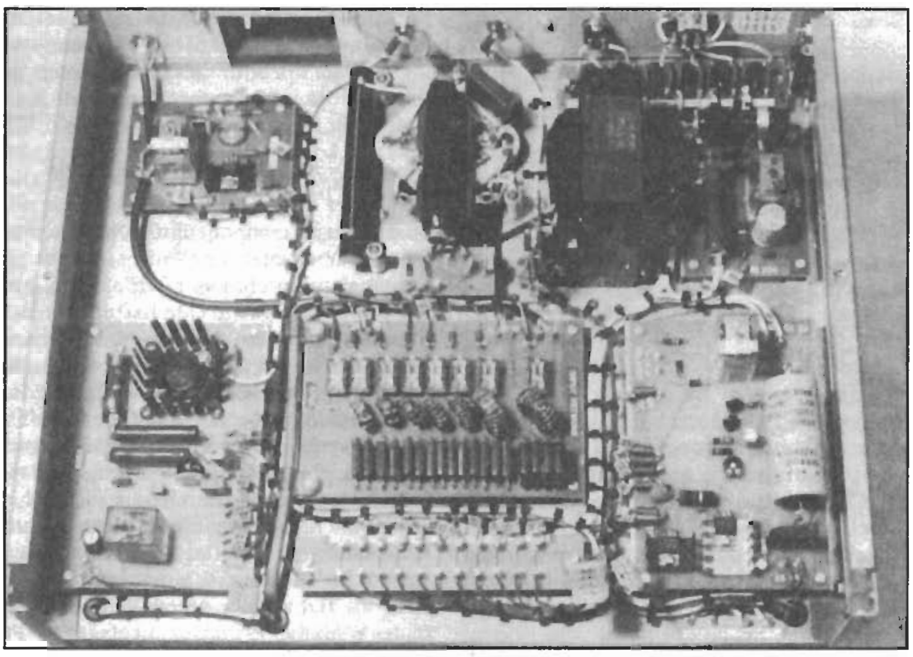

Fig 13—Bottom view of the RF deck showing harness wiring in place.

sis provides an easy way to mate the 12-pin nylon connector going to the power switches with the wiring harnesses under the chassis plate (see Fig 13). When making the PC boards, try to put all the connections to the board on one side. This allows access to the boards for maintenance without removing the wiring. Just unscrew the board from the channel and fold the board upward.

Each wire is labeled at each end with numbered tags, because the Teflon wire used is mostly the same color. A version of the schematic diagram was maintained with the wire numbers noted for easy wire tracing. The final wiring is cabled into harnesses with plastic cable ties. Don't be afraid to use plenty of ties, but at first only put a tie every inch or two. The ties will undoubtedly be cut several times during wiring to put in missing wires that are overlooked. When all wiring is complete and tested, put a cable tie every ½ inch on the major harnesses.

Amplifier Tank Circuit

The plate tank-coil set is shown in Fig 14. Before making the coils, determine the tankcircuit parameters for the given tube plate impedance. The plate impedance can be determined from Eq 1.

Plate impedance = plate voltage/ (1.57 × plate current) (Eq 1)

Assuming the amplifier runs at 60% efficiency, the input power required for 1500-W output is approximately 2500 W. With 3200 V on the 8877 plate, plate current will be approximately 781 mA. From Eq 1, a tank circuit designed for approximately 2600 ohms is appropriate. A table of pi-L network component values is contained in Hoff's article, and the values used in this amplifier are summarized in Table 2.[10]

The 80- and 160-meter coils are wound on an assembly of three T225-2 toroid cores taped together with Scotch no. 27 glass-cloth tape. Use plenty of tape to provide good voltage insulation from the cores. The 80-meter coil is wound with 11 turns of no. 10 wire covered with Teflon sleeving. The 160-meter coil is wound with 19 turns of no. 12 wire, also covered with Teflon sleeving. The cores of both coils are mounted from 1 to 2 inches from the mounting wall on ceramic insulators. Each coil is sandwiched between two pieces of fiberglass material held together by a ceramic standoff running through the middle of each toroid (see Fig 15). Wind all coils in the same direction (clockwise or counterclockwise), to avoid a "bucking" action between coils.

The 10- through 40-meter coil is made from ¼-inch soft copper refrigerator tubing. Clean the tubing using fine steel wool (000) until the surface is smooth and bright. Use a piece of pipe with an OD equal to the desired coil ID and carefully wind the copper tubing around the pipe. After the smaller diameter part of the coil has been wound, change the pipe to a larger diameter, and wind the larger part of the coil. The inductance of the coil can be easily checked using a known fixed-mica capacitor with low internal inductance and a dip meter. After determining the resonant frequency, calculate the inductance using Eq 2.

Inductance = $1/(4\ \pi\ f^2C)$ (Eq 2)

Always wind the inductor with a few extra turns and then remove turns until the desired inductance is achieved. The induc-

---

[10] I.M. Hoff, "Pi Network Design for High Frequency Power Amplifiers," *Ham Radio*, Jun 1978.

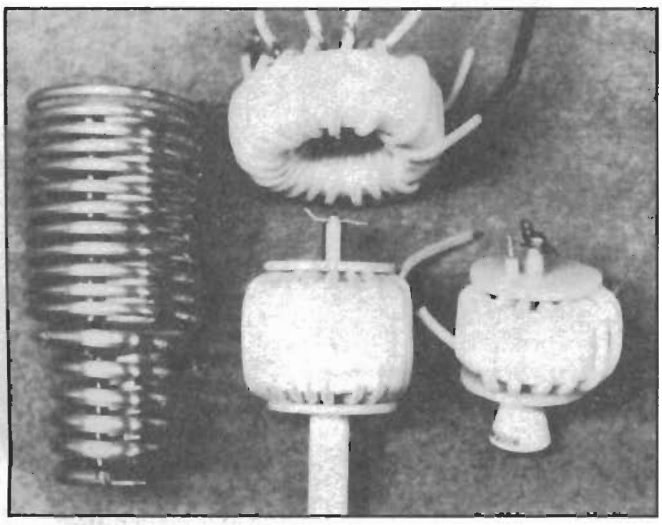

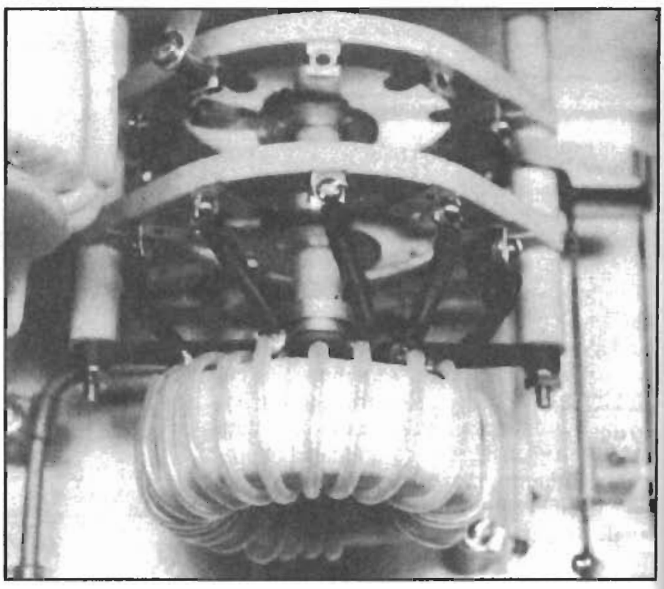

Fig 14—Plate tank coil set. Above: the various coils of the tank circuit. Right: the L toroid mounted on the rear of the band switch.

Table 2

Pi-L Network Design Parameters (See Fig 3, Part 1)

Freq (MHz)	*C1 (pF)*	*L1-L4 (μH)*	*C2 (pF)*	*L5 (μH)*	*Q*
1.8	462	22.02	2121	8.90	13.1
3.5	244	10.99	1132	4.45	13.4
7.0	113	6.03	503	2.44	12.4
14.0	55	3.08	245	1.24	12.2
21.0	37	2.05	164	0.83	12.2
24.9	30	1.70	130	0.70	12.0
29.7	26	1.48	112	0.60	12.0

L1, 10-12 meters—6 turns 1/4-inch silver-plated copper tubing, 2-inch outside diameter, 2 inches long. 10-m tap at 3 turns.

L2, 15-40 meters—12 turns 1/4-inch silver-plated copper tubing, 3 1/4-inch outside diameter, 4 inches long. 15-m tap at 2 turns, 20-m tap at 4 turns, 40-m tap at 12 turns.

L3, 80 meters—11 turns of no. 10 wire in Teflon sleeving over three T225A-2 iron-powder toroidal cores taped together with Scotch no. 27 glass-cloth electrical tape.

L4, 160 meters—19 turns of no. 12 wire in Teflon sleeving over three T225A-2 toroidal cores taped together with Scotch no. 27 glass-cloth electrical tape.

L5, L Coil—20 turns of no. 20 wire in Teflon sleeving over two T300-2 iron-powder toroidal cores taped together with Scotch no. 27 glass-cloth electrical tape. Tap as follows:

Band	10-12	15	20	40	80	160
Turn no.	2	3	4	6	10	20

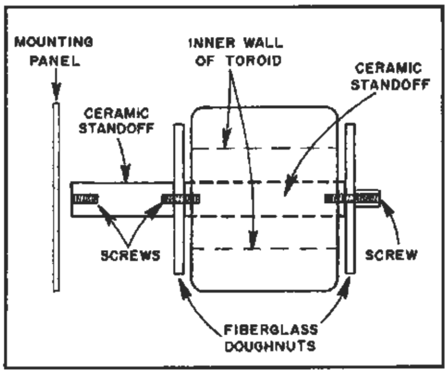

Fig 15—Mounting details of 80- and 160-meter toroids.

tance measurement is only an approximate value because there is always stray inductance introduced by the interconnecting leads and the band switch.

The copper tubing is silver plated only after the ends are configured to match the mounting lugs in the amplifier. Silver plating is simple with the right materials. Go to a photographic lab and get a couple of gallons of used fixer solution. Fixer solution is not usable when it becomes saturated with silver. The labs usually sell the spent fixer for silver recovery, so you may have to pay for it. Clean the coil by giving it another brushing with steel wool, wash it thoroughly in a mild detergent and rinse it well before plating. Put the fixer solution in a large plastic container and submerge the coil in it. A bright silver plate will form within seconds. The silver plate is not very thick, but it is sufficient to keep the copper from tarnishing. The coil taps are small wrap-around clamps made from silver-plated 1/2-inch-wide copper strap. Plate about 2 inches of copper strap, then drill a hole in one end for a no. 8-32 screw. Using a short piece of 1/4-inch tubing as a jig, bend the end sharply around the tubing with a pair of duckbill pliers. Drill a second, matching hole, but squeeze the clamp in to provide a tight fit around the coil stock. Install the clamp with a flat washer on each side to provide a compression fit. The clamp should fit tightly. Don't solder the clamps until after the amplifier is tested. When the tap location is verified, heat the clamp and coil with a large solder gun and flow solder into the connection. After soldering the coil taps, spray the coil with a thin coat of clear plastic to retain the bright silver finish.

The L coil is wound on a pair of T300-2 toroid cores taped together with glass-cloth tape. The coil is mounted on the back of the band switch and supported by the connections to the coil. The leads are not soldered to the band switch, but are held in place with no. 6-32 screws.

Other major considerations in the RF tank circuit are the TUNE and LOAD capacitors. The TUNE capacitor is a 375-pF vacuum variable. Table 2 shows that though this capacitor will easily cover 10 through 80 meters, it is too small for 160 meters. One solution is to use a 500-pF vacuum variable, but units in this size class are not readily available in the surplus market, and you probably don't want to buy a new one at approximately $500. I solved the problem by switching in an additional 160-pF capacitor in parallel with the vacuum variable for 160 meters.

A 24-V dc solenoid, mounted on the subpanel, grounds a pair of 80-pF fixed vacuum capacitors. The solenoid is powered by a small 12-V relay that is energized by the 160-meter control line on the input network control switch. The grounding strap is made from a stiff piece of brass stock obtained at a hobby shop. A hinge is made by soldering a small piece of brass tubing to the strap and putting a solid rod through the tubing. The rod is supported on each end by a small piece of Plexiglas and allowed to pivot, thus forming a hinge. A piece of flexible braid is connected from the brass strap to ground for a good connection. The strap is held in the open position by a small spring pulling the strap back toward the subpanel. The contacts on the strap and the vacuum capacitors are from a 25-A power relay and soldered in place. A piece of 1/4-inch brass bar stock is used to mount the contact on the vacuum capacitor side and to absorb the shock from the closure. A rubber grommet on the brass strap, where the rod from the solenoid hits the strap, also absorbs some of the shock. This scheme works smoothly and once thought out, took about one evening to install.

The LOAD capacitor is a 1000-pF, 3-kV unit with a fixed 100-pF, 5-kV fixed mica transmitting capacitor in parallel. This provides a total of 1100 pF for 80 meters. On 160 meters, the band switch adds an additional 1200 pF in parallel to provide a total of 2300 pF. The 1200-pF capacitor consists of three 400-pF mica transmitting capacitors in parallel.

Marking and Calibrating Meters

Any meter with a movement from 100 μA to 5 mA can be used for the front-panel meters. Use two identical meters, if possible, for esthetic reasons. In a very clean environment, begin by removing any internal meter shunts (on ammeters), or any series resistors inside the meter (on voltmeters) to obtain only the basic meter movement. Nothing else need be done for the multimeter, since the calibration resistors are a part of each metered circuit. A shunt resistor must be made for the plate-current meter. Wind about 2 feet of no. 22 enameled wire on a 2-W resistor of any value over 50 ohms, soldering the wire ends to the resistor terminals. Mount the resistor across the meter terminals to form a shunt. Connect the meter in series with another meter of known calibration and an adjustable power source. Trim the enameled wire 1 inch at a time until the meters read the same. If the meter being calibrated reads too low, the wire is too short. When the meters read the same, slip a piece of heat-shrink tubing over the resistor and seal the shunt.

Carefully remove the plastic face cover and the calibrated face plate from the meter. Do this in a clean environment because the magnet in the meter will attract metal shavings that could damage the movement. With a pencil eraser, rub off the original meter lettering, but not the analog scale tick marks. With small dry-transfer letters put the new numeric scale on the meter. Be careful not to bend the pointer when reassembling the meter.

Mounting the Front Panel

Remove the masking tape from the

panel, and carefully label the front panel with dry-transfer lettering before mounting any components. Mount the meters, potentiometer and switches in place, and complete the wiring. Wire nylon connectors to the power switches, ALC control, multimeter switch and meters to mate with the connectors in the RF deck. This will result in a totally removable panel.

When mounting the front panel to the cabinet, give careful attention to aligning the vacuum-capacitor and band-switch shafts. Loosen the component mounting screws and mate the shafts, then retighten the screws. Continue the alignment process until the controls operate smoothly.

Rear-Panel Assembly

Up to now, nothing has been done to the rear panel. Mount the panel and decide where the components should mount to obtain short connections. Lay the rear panel out in the same way as the front panel, and cut the necessary holes.

Mount the blower off the rear panel, located to allow good circulation of air up to the front of the under-chassis area and back to the tube socket. Positioning of the blower is not critical as long as air is not directly blown across the tube socket, which could cause backpressure.

Mount all components to the rear panel, and mount and wire the panel to the RF deck. The rear panel is not easily removable like the front panel.

Testing the Amplifier and Power Supply

Testing is the *big* moment and the climax of several months' work. First hook up the power-supply control cable and the ground cable. Leave the high-voltage cable off. Test the control circuits to ensure that the power supply can be turned on from the RF deck and the 3-minute time delay works. Make sure that the tube filaments are on. This isn't as easy as with a glass tube, which allows you to see the filaments glow. Let the tube run for about 10 minutes, then turn the power off. Immediately remove the tube and feel if the base is hot.

Test the amplifier first on 20 or 40 meters, since these bands use the midrange of the TUNE and LOAD capacitors. With high voltage applied to the tube, first key the amplifier with no input drive power. The resting idle current should be between 100 and 200 mA on the plate-current meter.

Now apply a little drive power to the tube. The plate current should rise. Move the TUNE and LOAD controls until power output is indicated on a wattmeter. Increase the drive while adjusting the TUNE and LOAD controls to achieve about 20 mA grid current and 650 to 700 mA of plate current, to realize 1500 W output. If the amplifier works properly on this band, proceed to the other bands and repeat the tests.

Efficiency should be at least 60% on all bands, except perhaps 10 and 12 meters, where efficiency may drop to 55%. If efficiency is poor, try moving the coil taps, but recognize that moving the taps also changes the Q of the coil. Decreasing the inductance and increasing the capacitance in the tank circuit will increase the Q. The amplifier may provide more power output over a larger frequency range with a lower tankcircuit Q, but harmonic suppression will also decrease and the amplifier could start generating interference or not meet FCC standards.

I must warn you one last time. *This device could kill you in one instant if you get tied into the high-voltage circuit. Put a good ground on both power supply and RF deck, and treat the equipment with proper respect!*

Conclusion

I have spent many enjoyable hours operating with this amplifier/power supply combination, with nothing but good signal reports. It was an exhausting task, but now that it's finished, I'm glad I did it. Try building one and you'll see what I mean.

By David Rutledge, KN6EK, et al. From *QST*, May 1997

High-Efficiency Class-E Power Amplifiers

Part 1—With 3 to 12 W of drive, you can push 300 to 500 W CW out of an $11 transistor! The trick is to use Class E.

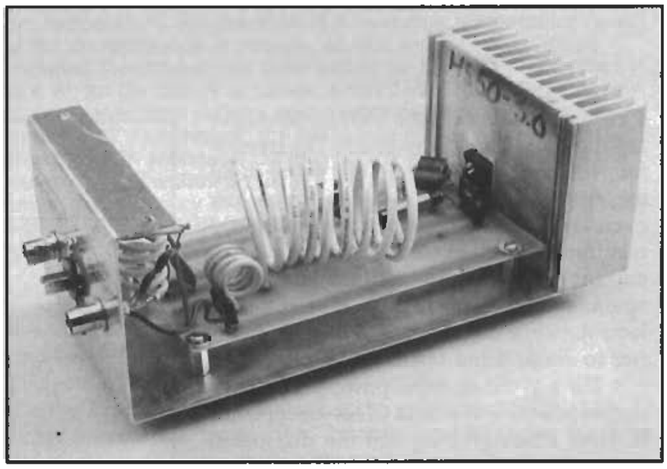

These 300 and 500-W 40-meter amplifiers evolved from a series of undergraduate student projects at Caltech. Our goal was to design an inexpensive amplifier that amateurs can easily duplicate.[1] The amplifiers use inexpensive, readily available power MOSFETs that can be driven by a QRP transceiver; in our case, the NorCal 40A. A block diagram of our setup is shown in Figure 1. The components of the station are shown in Figure 2.

Our 300-W amplifier uses an International Rectifier (IR) IRFP440; the 500-W amplifier employs an IRFP450. These transistors are widely used in switching power supplies, but we have not seen them previously reported for use as RF amplifiers. The MOSFETs have a maximum drain voltage of 500 V, with maximum RMS drain currents of 8.8 A for the IRFP440, and 14 A for the IRFP450. Both transistors are available from Digi-Key: The 440 costs $8; the 450 costs $11.

Class-E amplifiers are extremely efficient—about 90%. Because of the low loss, no cooling fan is required. No TR switch is needed; the received signal is piped through the amplifier itself. Even without an external filter, the amplifiers meet the FCC requirements for spurious emissions. They can be built and tuned up with an RF power meter, a multimeter, oscilloscope and a dummy load. Tune-up consists of adjusting an input coil for matching and an output coil to set the power level.

Operational Classes

In addition to the well-known Class A, B, and C operational modes, there are Class D and E.[2] In Class D and E, the devices operate as switches, half the time completely on, and the other half completely off. But transistors are not perfect switches. The MOSFETs have a resistance of about 1 Ω when on, and a capacitance of several hundred picofarads when off. Losses are greatly reduced in switching amplifiers, but there is a penalty: The output power no longer depends on the *drive power*, but rather on the *supply voltage*. This means that switching amplifiers are *not* linear amplifiers, and they are *not* suitable for SSB without additional limiting and modulating circuits. However, they are fine for CW, FSK and FM.

In a Class-D amplifier, a pair of transistors switch on and off, out of phase across an output transformer. Fred Raab, WA1WLW, recently developed a Class-D power amplifier that produces 250 W on 40 meters with an efficiency of 75%.[3] However, Class-D amplifiers are relatively complex. On the other hand, the Class-E circuit has significant advantages for the homebrewer because only one transistor, without gate bias or output transformer, is needed, and it can be driven by a low-power transceiver.

The Class-E Amplifier

The Class-E amplifier was invented and patented by Nathan Sokal, WA1HQC, and Alan Sokal, WA1HQB, in 1975.[4] It minimizes heat loss by having as little overlap as possible between voltage and current. Figure 3 shows an idealized Class-E circuit. As the switch opens and closes, the current alternately flows in the switch and in the load network. The switch voltage and current waveforms are shown in Figure 4. It may be easiest to understand the waveforms by starting at the beginning of the off-time interval. When the transistor turns off, the current flows into the resonant load network, and there is a transient voltage that rises and falls. With a properly de-

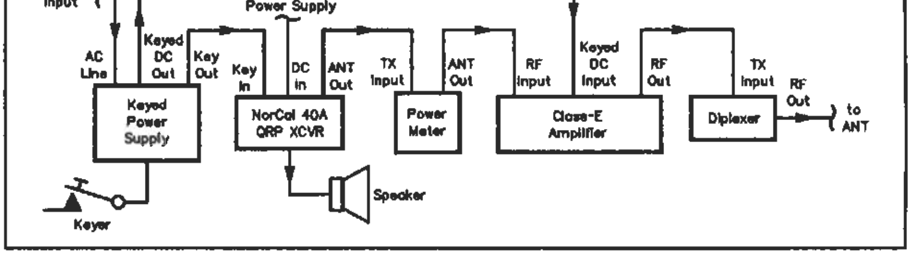

Figure 1—Block diagram of the Class-E amplifier station.

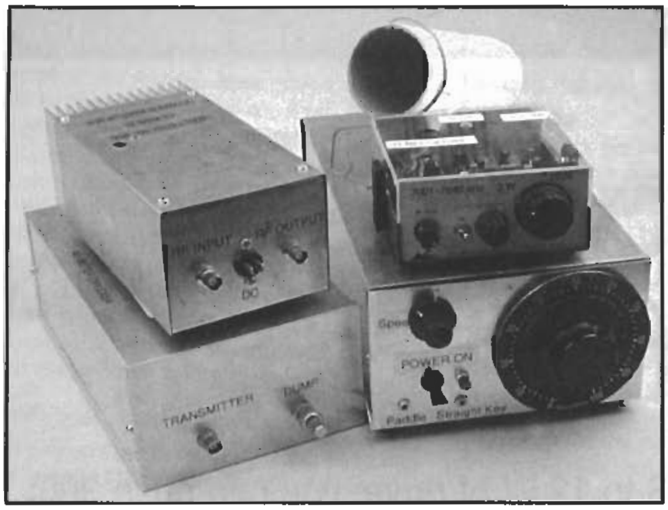

Figure 2—The components of the Class-E amplifier station. Clockwise from bottom left: the diplexer, amplifier, resonant speaker, NorCal 40A and keyed power supply. The latter houses the keyer and a pulse-stretching and shaping circuit.

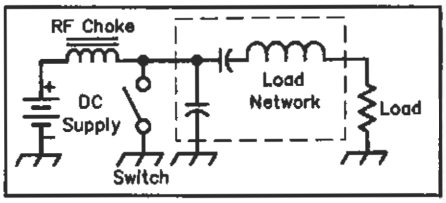

Figure 3—An idealized Class-E amplifier. The transistor is represented by a switch that opens and closes at RF.

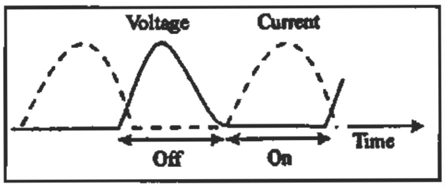

Figure 4—Class-E voltage and current waveforms. Class-E amplifiers reduce loss by keeping the overlap between voltage and current low.

signed load network, the voltage returns to zero smoothly with zero slope. The transistor switches on when both the voltage and the current are small, keeping losses low even if the switching is slow or slightly mistimed. Once the transistor turns on, the current rises smoothly until it switches off again, and the cycle repeats. The resonant load network does limit a Class-E amplifier to single-band operation.

For a given dc input power, the power output of a 90% efficient Class-E amplifier is three times greater, and the dissipated power is seven times smaller, than that of a 30% efficient Class-A amplifier. This means that for a given dissipated power, we can get *21 times* more power from a Class-E amplifier.

The Caltech Power Amplifiers

Both amplifiers have a common diagram (Figure 5) and PC board, but use components with different values. Several components are added to the basic Class-E circuit for matching and filtering. The MOSFET's gate impedance is rather low, and primarily capacitive, with a reactance of about 4 Ω. There is also a resistive component of about 2 Ω from parasitic series resistance in the gate itself, and the drain *on* resistance that is capacitively coupled to the gate. T1 reduces the 50-Ω impedance of the drive circuit to about 2 Ω to match the low resistance of the gate and sets the dc bias to 0 V. L1 is adjusted to cancel the gate capacitance; the input SWR is typically 1.5:1. C1 shunts L1 at high frequencies to reduce ringing in the VHF range. The 0-V gate bias ensures that the transistor is off when it is not driven, because this is far below the threshold voltage, which is about 4 V. We have never seen oscillations in these amplifiers.

C3 and L2 form a resonant network that produces the rising and falling voltage waveform needed for the Class-E ampli-

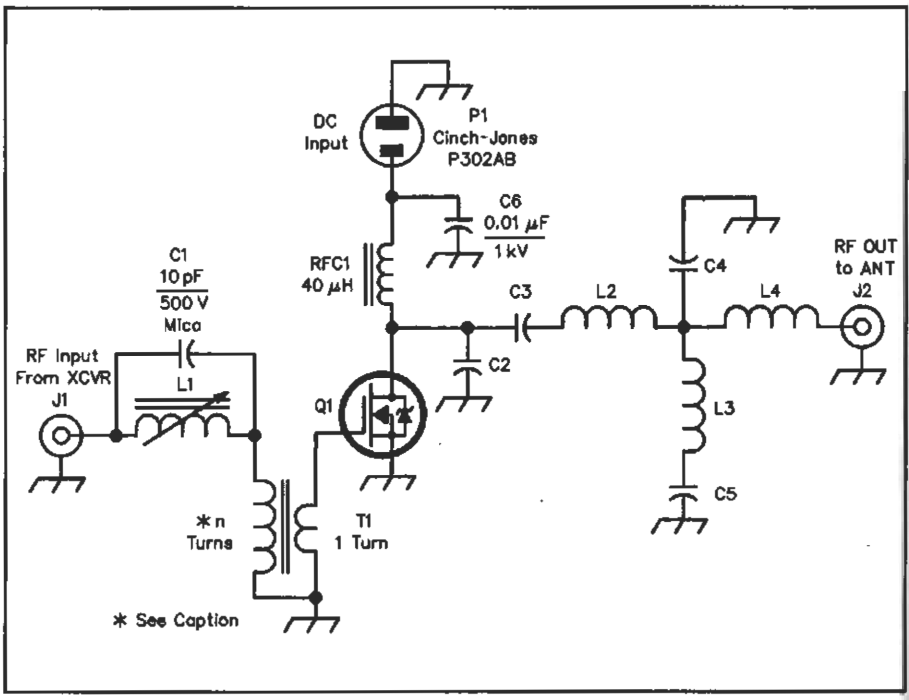

Figure 5—Circuit diagram and parts list for the amplifiers. Mica capacitors are available from Newark. L2, L3 and L4 are made with #10 THWN solid, insulated house wire sold in hardware stores. T1's core is an RF400-0, available from Communications Concepts (see Note 1 for supplier information). L1 is a Toko 10K (2.2 µH), available from Digi-Key. RFC1 is a J. W. Miller type 5240, available from Newark.

C1—10 pF, 500 V mica
C2—300 W, 270 pF; 500 W, 390 pF; (For C2 through C5, Cornell Dubilier mica capacitors, type CDV19, 1 kV, 5%
C3—300 W, 1500 pF; 500 W, 2000 pF
C4—300 and 500 W, 100 pF
C5—300 W, 680 pF; 500 W, 820 pF
C6—0.01 µF, 1 kV ceramic disc
J1, J2—BNC or SO-239 connectors
L1—2.2 µH, Toko 10K; available from Digi-Key.
L2—300 W, 9 turns; 500 W, 8 turns #10 THWN wound on 1¼-inch-OD plastic-pipe form; spread turns to fit holes in PC board.
L3—300 W, 4 turns; 500 W, 3 turns #14 THWN wound closely spaced on a ½-inch-diam drill bit
L4—300 W, 5 turns; 500 W, 5 turns #14 THWN wound on a ½-inch-diam drill bit form, turns spaced to occupy 1 inch
P1—Cinch-Jones P302AB
Q1—IRFP440 for 300-W amplifier; IRFP450 for 500-W amplifier; (use International Rectifier transistors only)
RFC1—40 µH, 3 A
T1—Pri: 300 W, 5 turns; 500 W 6 turns #26 stranded hook-up wire; wound on RF400-0 core, available from Communications Concepts
Misc: Berquist K10-104 insulating pad (thermal resistance of 0.2-K/W, 6-kV breakdown rating); available from Digi-Key.

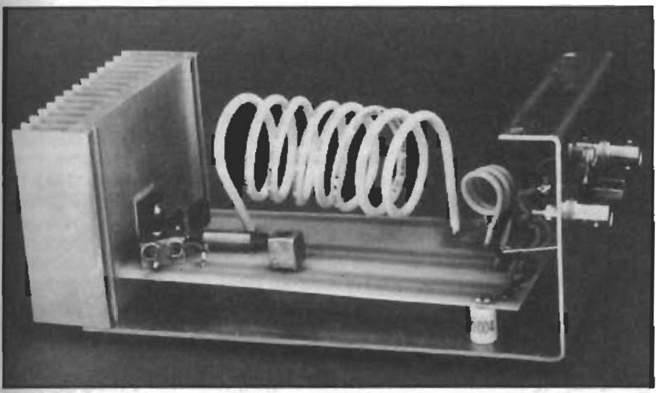

Figure 6—Inside the 500-W amplifier. The large coil is L2; L3 is the small coil. The circuit board is mounted on $^1/_2$-inch standoffs. BNC connectors are used for the RF input and output. No harm is done if the RF input and output are accidentally interchanged.

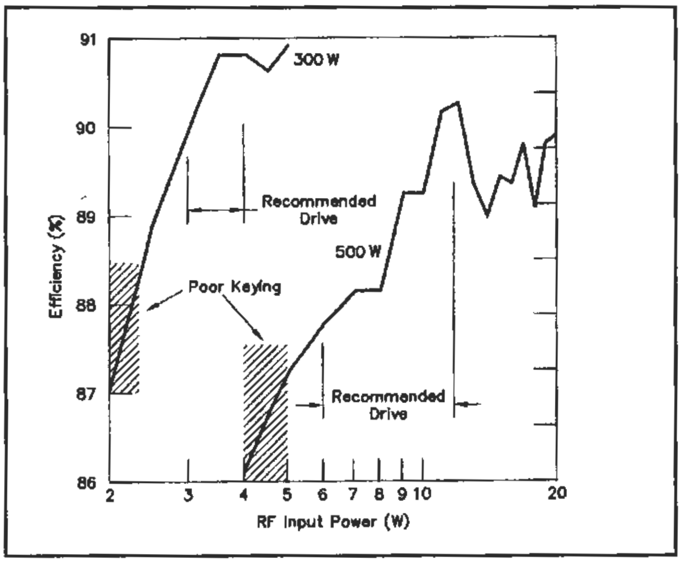

Figure 7—The measured efficiency plotted against RF input power. These measurements were taken with the heat sink at room temperature. In operation, the heat-sink temperature rises and the *on* resistance increases, so that the efficiency drops somewhat. Recommended drive power for the 300-W amplifier is 3 to 4 W; for the 500-W amplifier, 6 to 12 W. The dc input power was calculated from voltage and current measurements made with Fluke 87 multimeters; the RF output power was determined with a Bird 43P wattmeter using a 500-W element. Bird lists the accuracy of this wattmeter as ±25 W, but we were able to improve the accuracy of the measurement to an estimated 1% by thermal calibration. RF input power was measured with a Diamond SX-200 wattmeter.

fier. C5 and L3 act as a notch filter for the second harmonic. Without the notch filter, the second harmonic is typically between −25 and −30 dBc instead of the −40 dBc that the FCC requires on HF. In addition, C5 and L3 transform the 50-Ω antenna impedance to about 10 Ω, the appropriate load for a Class-E amplifier. RFC1 converts the 0 to 120-V dc input from the power supply to a current source, and C6 helps keep RF energy out of the power supply.

C4 and L4 form a low-pass filter to remove VHF harmonics. Without the filter, there are several harmonics at levels from −40 to −60 dBc in the frequency range from 130 to 210 MHz. With the filter, the VHF harmonics are reduced to the −70 to −80-dBc level.

Amplifier Construction

The 500-W amplifier is shown in Figure 6 with its cover removed. The transistor is mounted on a 3∞4$^1/_8$-inch heat sink with 1-inch fins (type HS50-3.0 from RF Parts, with a thermal resistance of 2 K/W with no fan) with a #6-32 bolt and nut. The transistor generates most of the heat (about 10% of the dc power) in the amplifier, so it must have good thermal contact to the heat sink. Because the transistor's case reaches high voltages, it must be electrically isolated from the heat sink. We use a Kapton pad manufactured by Berquist that has a thermal resistance of 0.2 K/W and a breakdown voltage of 6 kV. The heat-sink surface must be free of burrs, and the transistor should lie flat on the surface with minimal stress on the leads. If a torque screwdriver is available, International Rectifier recommends a mounting torque of 10 inch-pounds. The heat-sink baseplate is an aluminum L bracket bent from 0.050-inch-thick aluminum sheet. A U-shaped enclosure cover is made of 0.016-inch-thick aluminum sheet. A hole in the cover allows insertion of a plastic screwdriver for tuning L1.

Solder Q1 and C2 flush to the PC board to reduce VHF ringing on the gate signal. Good electrical contact is needed between Q1's source lead and the heat sink. Use a #6-32 bolt and nut, with washer spacers so that the source lead does not bend. Scope-probe pigtails soldered to Q1's gate and drain leads poke through holes in the baseplate. Rubber grommets in the holes prevent the pigtails from shorting to the chassis.

L4, L3 and L2 are made by winding solid insulated wire on pipe and drill-bit forms. Orient the coils at right angles to each other to reduce coupling between them. L2 has a 7-MHz Q of 350; L3 and L4, a Q of about 170. For C2, C3 and C5, use only 1-kV mica capacitors—even 500-V capacitors fail spectacularly with a burst of flame. If a 1-kV, 100-pF capacitor is not available for C4, substitute a series-connected pair of 500-V, 200-pF mica capacitors. For best filtering, mount C4 and L4 directly on the center pin and ground lug of J2. Mount L4 with its axis vertical to reduce coupling to the other coils.

The NorCal 40A Driver

For a driver, we use a NorCal 40A,[5] but its 2-W output is not enough to drive these amplifiers. Fortunately, the NorCal 40A can be modified to deliver greater power output.[6] We recommend 3 to 4 W drive for the 300-W amplifier, and 6 to 12 W for the 500-W amplifier. These drive levels give an efficiency in the 90% range (Figure 7). Drive levels lower than these give poor efficiency; higher drive levels increase the dissipated power without improving efficiency. Don't drive the 300-W amplifier with less than 2.3 W, and the 500-W amplifier with less than 5 W. At these low power levels, the transistor may not turn on fully at all supply voltages, subharmonic spurious components may be generated, and the amplifier may not key properly.

The recommended drive powers produce peak gate voltages of between 15 and 20 V. International Rectifier specifies a maximum peak gate voltage of 20 V to avoid rupturing the gate. Although the drive levels are close to this limit, experience shows them to be quite safe. The 20-V limit is more appropriate for the low frequencies used in power supplies than for RF voltages. In controlled tests, we've pushed the 300-W amplifier to 60-V gate-voltage peaks, *three times* the manufacturer's voltage limit, without damage.

Next month, we'll discuss the keyed power supply, keying waveform shaper and tune-up. Join us!

Notes

[1] A package of amplifier parts *only*, including PC board, components, connectors, heat sink and chassis is available at cost from Puff Distribution, Department of Electrical Engineering, MS 136-93, Caltech, Pasadena, CA 91125. Price: $50 for US orders, $60 for foreign orders. The price includes tax and shipping by surface mail. Make payment by check or money order only to "Caltech-Puff Distribution." Foreign checks must be drawn on a bank with a US branch office. Please provide your Amateur Radio call sign and specify which amplifier (the 300 or 500 W unit) you want. For more information, contact Dale Yee by e-mail at **yee@systems.caltech.edu**, fax 818-395-2137, or you can download an order form from: **http://www.systems.caltech.edu/EE/Faculty/rutledge/poweramp.html**. We do not offer power-supply components. Those parts are available from Communication Concepts, 508 Millstone Dr, Beavercreek, OH 45434-5840, tel 513-426-8600; Digi-Key, PO Box 677, Thief River Falls, MN 56701-0677, tel 800-344-4539, **http://www.digikey. com**;

Newark Electronics, (many branches throughout the US; check your telephone book for a branch near you); main office: 4801 N Ravenswood Ave, Chicago, IL 06040-4496, tel 800-463-9275, 312-784-5100, fax 312-907-5217; RF Parts, 435 South Pacific St, San Marcos, CA 92069, tel 800-737-2787; Mouser Electronics, 2401 Hwy 287 N, Mansfield, TX 76062, tel 800-346-6873, 817-483-4422, fax: 817-483-0931 e-mail **sales@mouser.com; http://www.mouser.com**.

A template package is *not* available from the ARRL.

[2] Historically, amateurs have built high-power amplifiers with vacuum tubes rather than transistors to avoid complicated power-combining networks and many low-power transistors. See Dick Ehrhorn, W4ETO, "RF Power Amplifiers and Projects," *The 1996 ARRL Handbook*, Chapter 13. Exceptions to this are the elegant designs by Helge Granberg, K7ES/OH2ZE (SK). *RF Application Reports*, published by Motorola Inc, in 1995, contains over 20 Application Notes and Engineering Bulletins written by Helge Granberg on amplifiers with output powers of 20 to 1200 W. Communication Concepts sells the boards and components for these amplifiers. *Radio Frequency Transistors*, by Norm Dye and Helge Granberg, published by Butterworth-Heineman, Boston, 1993, is recommended reading. Joel Paladino, N6AMG, adapted one of Granberg's transistor amplifiers (see "An Experimental Solid-State Kilowatt Linear Amplifier for 2 to 54 MHz," *QST*, Sep 1992, pp 19-23). However, the power transistors alone cost $900! This led us to look for a less-expensive way to make transistor power amplifiers.

[3] Fred Raab, WA1WLW, "Simple and Inexpensive High-Efficiency Power Amplifier," *Communications Quarterly*, Winter 1996, pp 57-63.

[4] Nathan Sokal, WA1HQC, and Alan Sokal, WA1HQB, "Class-E, A New Class of High-Efficiency Tuned, Single-Ended Switching Power Amplifiers," *IEEE Journal of Solid-State Circuits*, Vol SC-10, Jun 1975, pp 168-175. This paper by father and son is a classic in the radio engineering literature, and is still the basic reference for the Class-E amplifier.

[5] The NorCal 40A designed by Wayne Burdick, N6KR, is available in kit form for $129 from Bob Dyer, KD6VIO, at Wilderness Radio, PO Box 734, Los Altos, CA 94023-0734, tel 415-494-3806, **http://www.fix.net/jparker/wild.html**. The NorCal 40A Web address is: **http://www.fix.net/~jparker/norcal.html**.

[6] A number of NorCal 40 modifications were published in *QRPp*, the magazine of the Northern California QRP club that first made the NorCal 40 kit available. For information on subscriptions to *QRPp* and back issues, contact Jim Cates, 3241 Eastwood Rd, Sacramento, CA 95821. Wayne Burdick, N6KR, suggests increasing the supply voltage for more power. If you do this, increase the voltage rating of Zener diode D12 to accommodate a larger peak collector voltage. For 3 W output, we use an MRF237 at Q7, the PA. To further increase the power to 7 W, we reduced L7 from 18 turns to 11 turns, and L8 from 18 turns to 14 turns. We replaced C45 and C47 with 560-pF, 300-V mica capacitors, and C46 with a 1500-pF, 100-V mica capacitor. For other approaches to raising the output power, see Dave Meacham, W6EMD, "5 Watts from your NorCal 40A," *QRPp*, Mar 1995, pp 6-7, and Ron Manabe, KN6VO, "Increasing the Output Power of the NorCal 40," *QRPp*, Jun, 1994 pp 42-45. Ron reports an output power of 7 W.

By David Rutledge, KN6EK, et al. From *QST*, June 1997

High-Efficiency Class-E Power Amplifiers

Part 2—Class-E operation permits low-cost MOSFETs to develop considerable power.

Last month[7] we talked about Class-E amplifier fundamentals and began construction of a 40-meter unit. Now we'll tackle the power supply, keyed-waveform shaper and develop some power.

A Keyed Power Supply

Nonlinear Class-E operation sharpens the CW keying envelope, causing annoying key clicks. To prevent this, we key the power supply to shape the supply voltage. A separate 4×7×12-inch (HWD) enclosure houses the dc supplies, a stretcher circuit that delivers a stretched pulse to the driver and a shaper that produces the shaped pulse for the amplifier. Figure 8 shows how these circuits connect.

So that the RF drive does not end before the shaping pulse, the keying pulse to the NorCal 40A driver is stretched a few milliseconds. The stretcher (Figure 9) takes a keyer's CMOS logic signal and provides a buffered keying waveform to the shaper and a stretched keying waveform to the NorCal 40A driver. Dc supplies (Figure 10) provide 12 V dc to run the ICs and 0 to 120 V dc for the amplifier. A wave shaper (Figure 11) gives this 0 to-120 V dc supply voltage a controlled rise and fall time to avoid key clicks. Figure 12 shows the keyed power supply with its cover removed.

There are other advantages to using a keyed power supply to control the output power. The amplifier power dissipation is low at all supply voltage levels, so that loss is kept low throughout a keying pulse. Because the keyed power supply also acts as a solid-state TR switch, a relay is not needed. This is because the supply voltage is zero except during key down. (This feature works well with the NorCal 40A, because it, too, does not use a relay for switching.) At zero voltage, the drain-to-gate capacitance in a MOSFET is quite large, and the signal from the antenna is fed through the amplifier with a loss of only about 7 dB. The NorCal 40A receiver sensitivity is excellent,[8] and a 7 dB signal loss does not hurt reception at all. A 7-dB loss degrades the MDS to –130 dBm, still far below typical 40-meter antenna noise levels of –90 to –110 dBm. On the positive side, with 7 dB attenuation, the receiver is less susceptible to intermodulation distortion from other signals in the 40 meter band. In addition, the amplifier reduces AM broadcast signals by about 20 dB. The 7-dB loss does need to be made up at the audio end. For this, we mount a 2-inch-diameter speaker in a cardboard mailing tube cut to resonate at 650 Hz for CW reception. This gives a sound level that is quite adequate.[9]

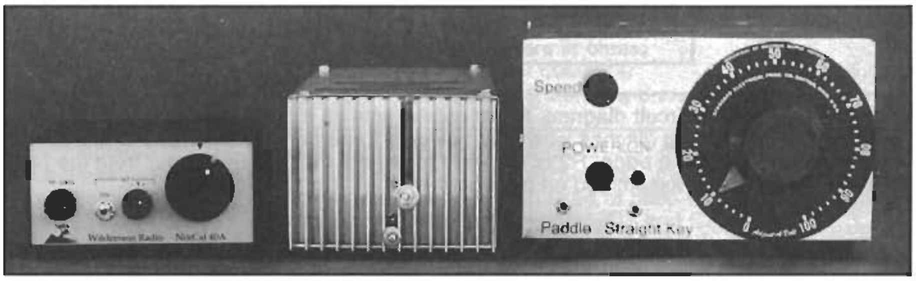

Transceiver, amplifier, and power supply for a 40-meter, 500-W station. The NorCal 40A transceiver driver is on the left, a 500-W amplifier is in the center and the power supply on the right. The amplifier's heat sink and transistor mounting screws are visible. The large dial controls the variable autotransformer, varying the RF output power.

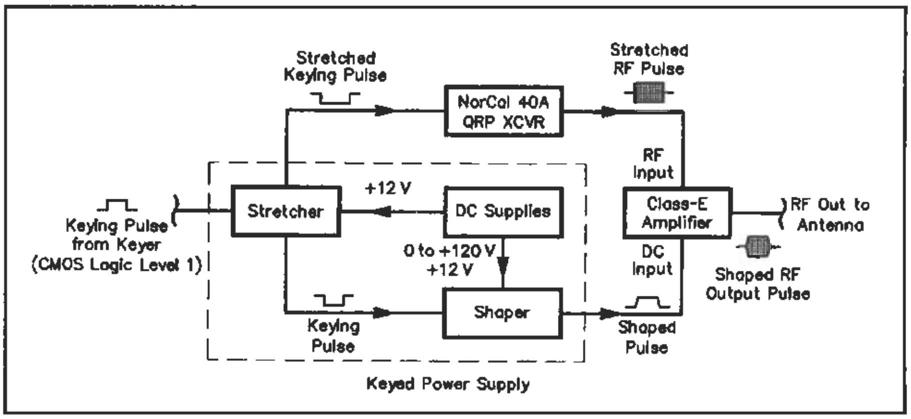

Figure 8—Block diagram showing the connections between the stretcher, dc supplies, shaper, NorCal 40A and the Class-E Amplifier. The stretcher, the dc supplies, and the shaper are in the keyed power supply. Power supply keying is done by a Curtis keyer IC (not shown) that provides a CMOS logic-level 1 during key down. The Curtis keyer IC and application note are available from MFJ Enterprises Inc, Box 494, Mississippi State, MS 39762, tel 800-647-1800, 601-323-5869, fax 601-323-6551; e-mail **mfg@ mfjenterprises.com**; **http://www.mhjenterprises.com**.(Keyer circuits using Curtis ICs can be found in recent editions of *The ARRL Handbook*.—*Ed*.)

The Diplexer

For greater reduction of spurious emissions, we recommend following the amplifier with a band-pass diplexer (see Figure 13) to terminate out-of-band spurious components in a 50 Ω load.[10] Our diplexer (in a 3.5×6×10-inch [HWD] box) uses the equivalents of a 100-pF series capacitor, an 1800-pF shunt capacitor and air-wound inductors. Stretch or compress L2 to achieve minimum SWR. The measured loss of 40-meter signals was extremely low, only 4%. Our experience shows that a diplexer can reduce all spurious components to more than –55 dBc.

Tune-Up

Refer to Figure 3 in Part 1. There are two amplifier coil adjustments. First, with the cover on and the dc input off, L1 is set for minimum input SWR with full RF input. Typically, the SWR can be reduced to 1.5:1. If it cannot be brought below 2:1, try adding or subtracting a turn from L1.

Output power is peaked by stretching or squeezing L2. Note: For safety, the amplifier's cover should always be attached when the RF drive is applied. The cover also significantly lowers the inductance of L2. Be sure to turn off the supply voltage before you touch *any* amplifier parts! A high RF voltage will burn the skin. RF burns are deep and heal slowly. Having a keyed power supply helps here, because the amplifier supply voltage is zero except during key down.

Attach a dummy load and power meter to J2. With the RF input applied continuously, slowly increase the dc input voltage, while monitoring the gate and drain voltages using an oscilloscope with 10× high-impedance probes. (You should see waveforms similar to those in Figure 15, although the peak drain voltage should be about half of

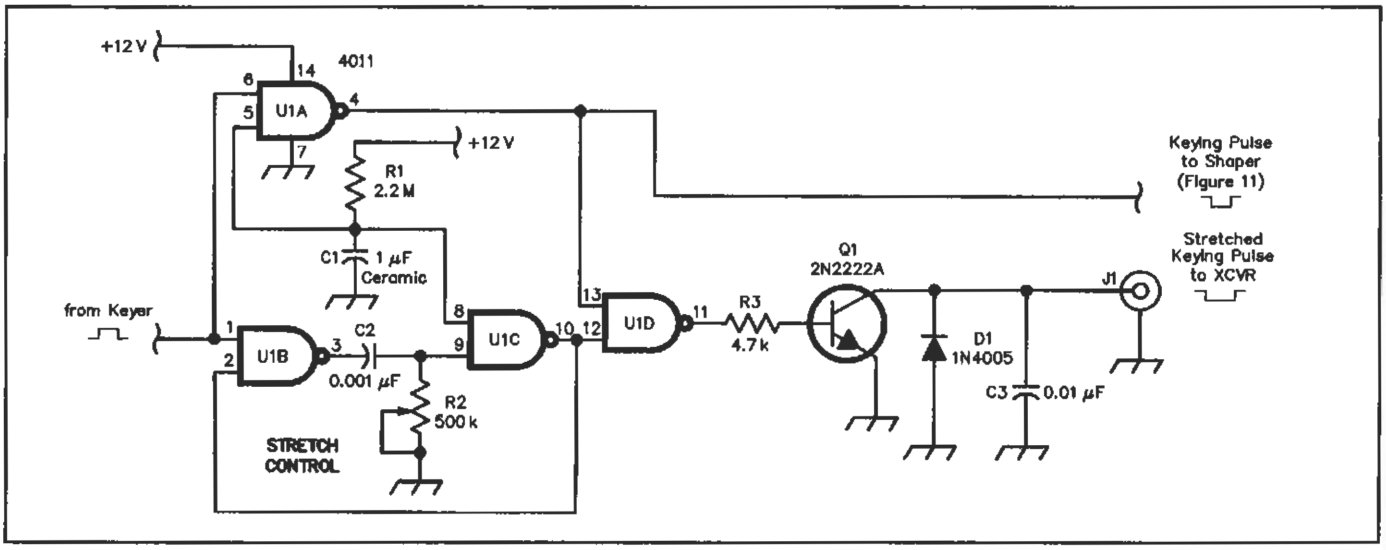
+12 V
4011
14
6
5
U1A
4
7
+12 V
R1
2.2 M
C1
1 µF
Ceramic
from Keyer
1
2
U1B
3
C2
0.001 µF
R2
500 k
STRETCH
CONTROL
8
9
U1C
10
12
13
U1D
11
R3
4.7 k
Q1
2N2222A
D1
1N4005
C3
0.01 µF
J1
Keying Pulse
to Shaper
(Figure 11)
Stretched
Keying Pulse
to XCVR

Figure 9—Stretcher circuit diagram. **NAND** gate A acts as a buffer and produces a keying pulse for the shaper. One input has an RC delay to prevent keying glitches when the power is turned on. Gates B and C are connected by an RC network that causes a pulse to be triggered on a falling edge. R2 is adjusted to ensure that the RF output from the NorCal 40A lasts longer than the shaping pulse. The components are assembled on perfboard.

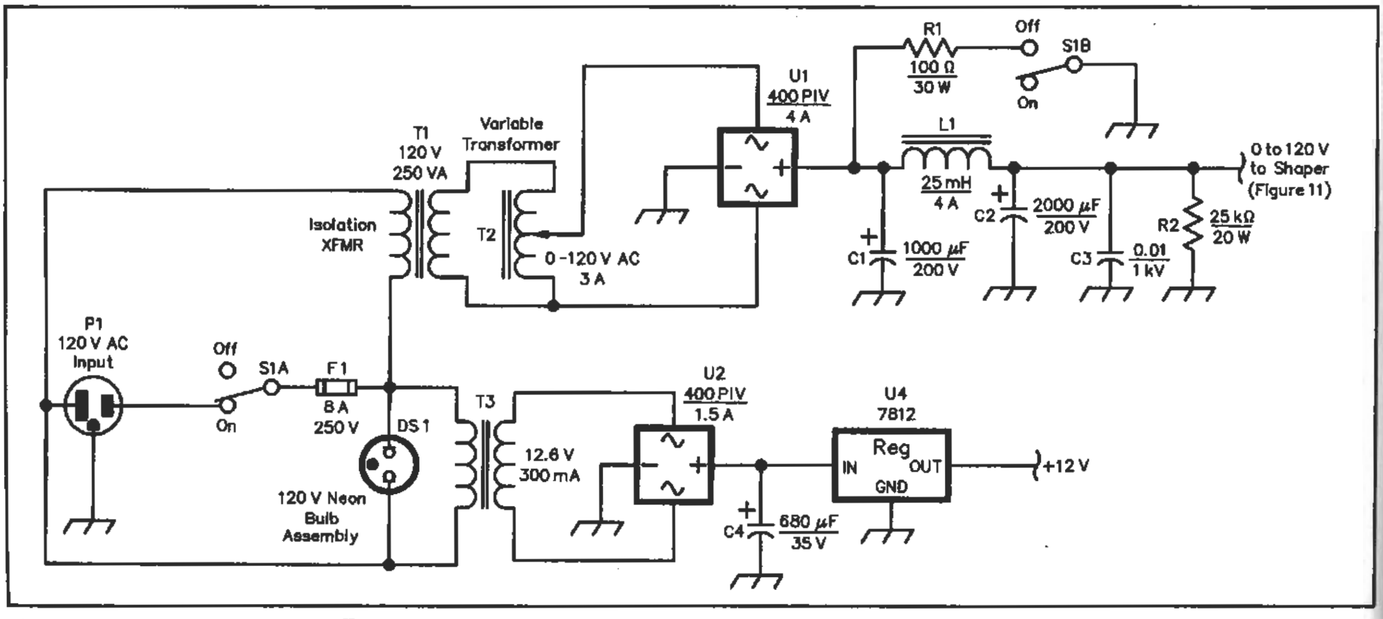
T1
120 V
250 VA
Isolation
XFMR
Variable
Transformer
T2
0-120 V AC
3 A
U1
400 PIV
4 A
R1
100 Ω
30 W
Off
S1B
On
L1
25 mH
4 A
C1
1000 µF
200 V
C2
2000 µF
200 V
C3
0.01
1 kV
R2
25 kΩ
20 W
0 to 120 V
to Shaper
(Figure 11)
P1
120 V AC
Input
Off
S1A
On
F1
8 A
250 V
DS1
120 V Neon
Bulb
Assembly
T3
12.6 V
300 mA
U2
400 PIV
1.5 A
C4
680 µF
35 V
U4
7812
Reg
IN
OUT
GND
+12 V

Figure 10—Dc supplies diagram. There are two isolated supplies: one with an unregulated 0 to 120-V output controlled by an autotransformer, and another with a +12-V regulated output. Two bleeder resistors help keep capacitor voltages at safe levels when the supply is turned off or its output reduced. The 4-A bridge rectifier is an RS404LR by Diodes, Inc, available from Digi-Key. The Panasonic 1000 and 2000-µF electrolytics in the filter are available from Digi-Key. The Stancor C-2686 25-mH choke is available from Newark. We scrounged the isolation transformer and variable autotransformer from old equipment. Look for them at swap meets, or purchase equivalents from Newark (Magnetek N55M 250-VA isolation transformer and the Staco 291 3-A variable autotransformer).

that in the figure.) The RF output should begin to rise. Increase the dc input until the forward RF output power is 25% of full power. Make sure the output SWR is 1.5:1 or less. Measure the dc input voltage, and adjust L2 to give 25% power at 60 V dc. Stretching L2 reduces its inductance and increases the output power, but usually lowers the amplifier's efficiency. Squeezing L2 reduces the output power and usually increases amplifier efficiency.

Increase the dc input voltage until the amplifier reaches maximum output power. The dc voltage should now be between 115 and 120 V, and the peak drain voltage of the RF-stage MOSFET should be between 380 and 420 V. Larger peak drain voltages run the risk of transistor failures. Lower voltages may indicate an excessive drain current, which can lead to a failure. If the voltage is too high, stretch the coils a bit more. If the voltage is too low, squeeze the coils. Check the RF drive and input SWR again. You may find that they have changed somewhat and that readjustment is needed. Measure the dc supply current and voltage, and calculate the amplifier efficiency to ensure that it is 85% or above. Use a 0.001-µF capacitor to bypass the voltmeter terminals because an RF voltage there can cause a significant measurement error. (In addition, you should realize that RF power meters often have error factors as high as 10%.)

Keyed-Waveform Shaping

The keying envelope is controlled by the three potentiometers in the shaper and stretcher circuits. R3 in the shaper circuit sets the rise time, R4 sets the fall time. R3 also helps control the power supply droop as the amplifier is keyed. R2 of Figure 9 determines the stretch in the pulse that keys the NorCal 40A, so that it does not stop transmitting before the end of the shaped

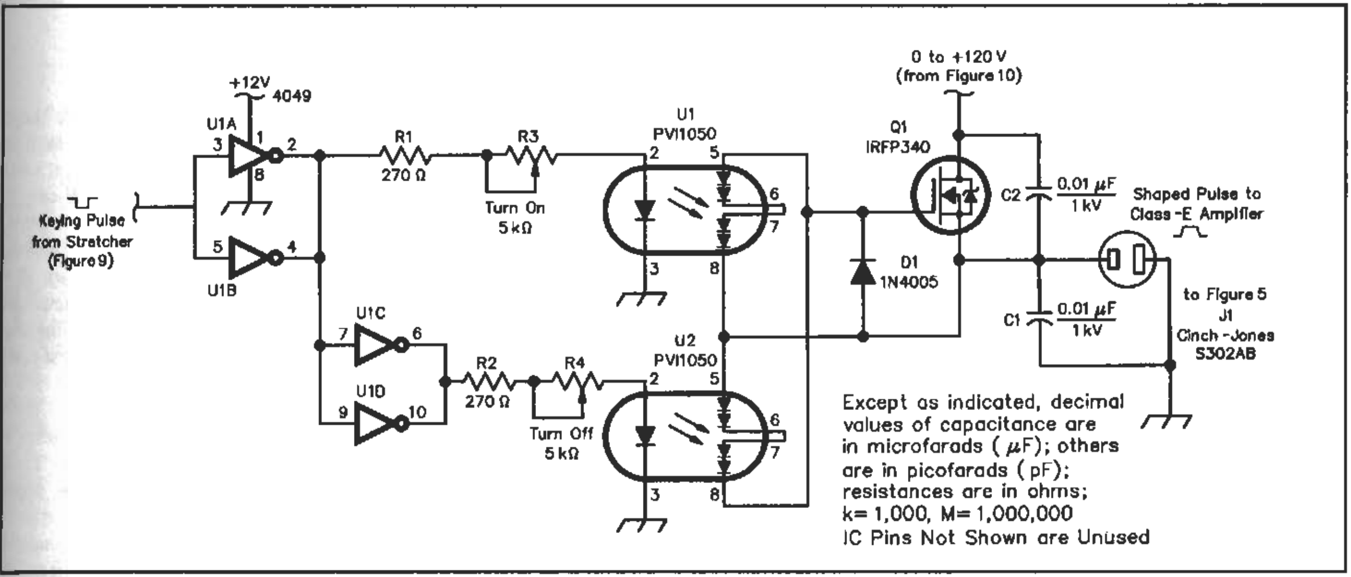

+12V
4049
U1A
3
1
2
8
5
4
U1B
Keying Pulse from Stretcher (Figure 9)
R1
270 Ω
R3
Turn On 5 kΩ
U1
PVI1050
2
5
6
7
3
8
U1C
7
6
U1D
9
10
R2
270 Ω
R4
Turn Off 5 kΩ
U2
PVI1050
2
5
6
7
3
8
0 to +120 V (from Figure 10)
Q1
IRFP340
D1
1N4005
C2
0.01 µF 1 kV
C1
0.01 µF 1 kV
Shaped Pulse to Class-E Amplifier
to Figure 5
J1
Cinch-Jones S302AB
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k= 1,000, M= 1,000,000
IC Pins Not Shown are Unused

Figure 11—Shaper circuit diagram. A pair of optically isolated MOSFET drivers (type PVI1050) turn Q1 on and off to bring the supply voltage up and down. Keying waveform rise and fall times are adjusted by R3 and R4, which control the current to each MOSFET driver. The 270-Ω resistors limit the LED current to a safe level. The TURN-ON potentiometer is also used to keep the overshoot from the unregulated power supply to a reasonable level. D1 prevents Q1's gate voltage from drifting negative. C1 and C2 are RF bypasses. To keep the supply voltage from appearing across exposed contacts, a female socket is used at J1. All components are available from Digi-Key. Point-to-point wiring on perfboard is used.

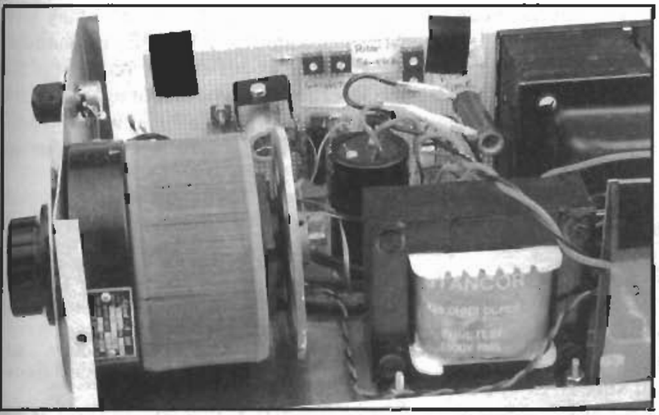

Figure 12—The keyed power supply with the cover removed, viewed from the side. T1 is mounted on the back plate, along with the fuse and connectors for the dc output, a keying line to the NorCal 40A and the ac line input. The **TUNE** switch keys the amplifier for testing. The variable autotransformer is on the left, mounted to the front panel. In the center foreground, mounted to the base plate, is L1, the 25-mH choke. Mylar sheets are taped to the perfboards for electrical isolation. The large components are attached directly to the box. The IRFP340 is mounted on the bottom of the box using a Berquist K10-104 Kapton insulating pad, #6-32 screw and a torque of 10 inch-pounds. The other shaping circuit components are mounted on perfboard. Layout is not critical.

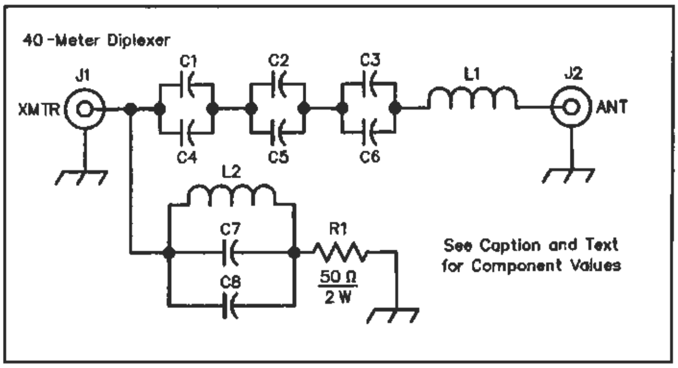

40-Meter Diplexer
J1
XMTR
C1
C2
C3
C4
C5
C6
L1
J2
ANT
L2
C7
C8
R1
50 Ω 2 W
See Caption and Text for Component Values

Figure 13—Schematic of our diplexer. It uses six 150 pF capacitors, two 910-pF capacitors and air-wound inductors made of $^{3}/_{16}$-inch-wide copper tape (see Note 1).

C1-C6—150 pF, 1 kV mica 5%-tolerance (Cornell Dubilier type CDV19; available from Newark Electronics, see Note 1)

C7, C8—910 pF (same type as above)

L1—24 turns of $^{3}/_{16}$-inch-wide copper tape, $1^{3}/_{8}$ inch ID; see Note 1.

L2—4 turns of $^{3}/_{16}$-inch wide copper tape, 1 inch ID; see Note 1.

R1—50-Ω, 2-W (approx) termination, Jameco part number 71458; available from Jameco Electronics, 1355 Shoreway Rd, Belmont, CA 94002, tel 415-592-8097; fax 415-592-2503 and 415-595-2664)

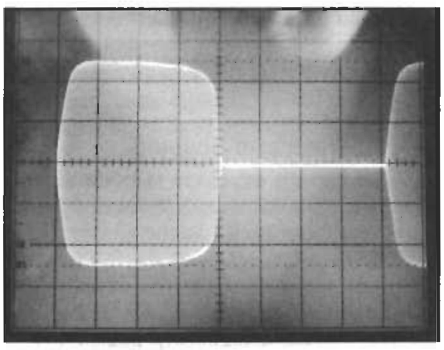

Figure 14—Keying waveform of the 500-W amplifier at 30 WPM. The horizontal axis is 10 ms per division. The rise and fall times are about 3 ms.

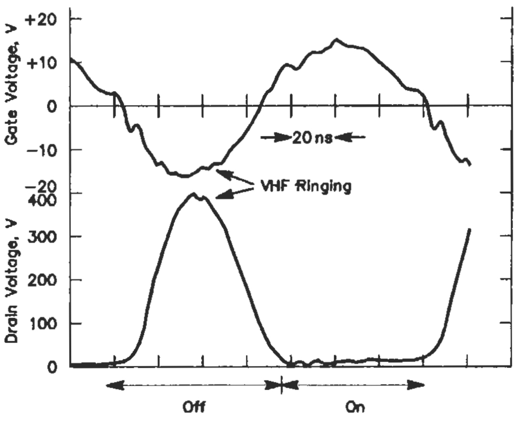

Figure 15—Drain waveform and ringing. Oscilloscope trace of the gate and drain voltages of the 300-W amplifier with 3 W drive. The dc supply voltage is 120 V; the input SWR is 1.6:1. Time scale is 20 ns per division; the transistor off and on times are shown. The transistor is off when the gate voltage is below the threshold, typically 4 V, and on when the gate voltage is several volts above the threshold. Peak gate voltage is 16 V, and the peak drain voltage is 400 V, safely within the manufacturer's ratings of 20 and 500 V, respectively.

pulse. We recommend using a keyer with adjustable weighting to offset the pulse stretching. The potentiometer settings interact somewhat, and there are variations at different sending speeds, so it is best to set them at the speed you commonly use. Adjust the controls for rise and fall times between 2 and 5 ms, and for a smooth keying envelope. Figure 14 shows keying at 30 WPM with rise and fall times of about 3 ms.

VHF Ringing

In many Class-E amplifiers, ringing in the VHF range can be seen on the gate and drain waveforms (Figure 15). This ringing can be quite pronounced, with bumps several volts high on the gate or drain or both. The bumps disappear when the RF input is removed, and that is why we refer to this as *ringing* rather than *oscillation*. We have compared measured spectral plots with *PSPICE* simulations and believe that the waves are driven by the sudden turn-on and turn-off of the transistor, acting rather like the gong of a bell.

We notice two distinct time periods and frequency ranges for the ringing. During the time the transistor is on, the ringing frequency is about 80 MHz. This appears to be a resonance of the external drain capacitor combined with the internal inductance of the capacitor and the transistor and may indicate a mismatched load. The on-ringing is usually small if the load is matched so that the drain voltage comes smoothly to zero before the transistor turns on.

The ringing while the transistor is off covers a broad range of frequencies—from 130 to 210 MHz. If the output low-pass filter is removed, the ringing can be seen easily on a spectrum analyzer at levels between −40 and −60 dBc. The off-ringing appears to be caused by a resonance of the external drain capacitor and its internal inductance, together with the transistor's internal drain capacitance and its inductance. The internal drain capacitance varies greatly with the drain voltage, so that the frequency is modulated as the drain voltage rises and falls. The low-pass filter reduces these harmonics to the −70 to −80 dBc range.

On the Air

The amplifiers meet the FCC requirements for spectral purity [confirmed in the ARRL Lab—*Ed.*]. NorCal 40A designer Wayne Burdick, N6KR, emphasizes that it is important to correctly tune the bandpass filter following the transmit mixer to minimize spurious emissions from the NorCal 40A.[11]

These amplifiers are excellent for chasing DX, schedules and "ragchews," particularly at this low point of the sunspot cycle. The amplifiers require no warm-up, no tune-up, and produce no fan or relay noise. We can vary the power from 1 W to full power via the variable autotransformer. The antenna SWR should be 1.5:1 or better because the amplifier is not protected against large mismatches. With a high SWR, the transistor will probably overheat. Check the dc voltage when the amplifier is delivering full power to the antenna to ensure that it remains between 115 V and 120 V. Readjust the coils if the voltage is too high or too low.

Our most common problem has been a poorly mounted PA transistor. If the transistor is not flat against the heat sink, heat transfer is poor and the transistor becomes quite hot; efficiency suffers (becoming usually less than 85%) and the amplifier may not reach full power. The output power may also drift downward, a sign that the transistor temperature is increasing and that the transistor is under stress.

Component temperature can be a good diagnostic tool. For the 500-W amplifier, our experience is that for CW QSOs longer than 30 minutes, the temperature of C3, C5 (Figure 5) and the heat sink rises to about 60× C. For the 300-W amplifier, the heat-sink temperature is about 50× C. This is hot to the touch, and it can be checked with a lab thermometer. The temperature will vary according to your operating style, and if the temperature is higher than you like, you can add a fan.

The Future

We see room for improvements: protection against antenna mismatches and employment of an inexpensive keyed *switching* power supply that is as lightweight as the amplifiers. Finally, it would be interesting to develop Class-E amplifiers for the other bands. We have built a 250-W amplifier for the 20 meter band that exhibits an efficiency of 88% with 10 W drive. We believe that Class-E amplifiers provide amateurs with good building challenges and operating fun at modest cost.

Acknowledgments

This work began as senior theses by Joyce Wong and Meng-Chen Yu at Caltech, and many of their ideas appear here. John Davis was supported by a scholarship from the James Irvine Foundation and the Army Research Office. Eileen Lau, Kai-Wai Chiu, and Jeff Qin received Caltech Summer Undergraduate Research Fellowships funded by the ETO Corporation and the Caltech Gates Grubstake Fund. We appreciate the help from all the people at the ETO Corporation, particularly Dick Ehrhorn, W0ID; Don Fowler, W1GRV; Tim Coutts; Chip Keen and Frank Myer. Mitsu Sakamoto, JA4FVE, of Kurashiki, Japan, sent us off-the-air recordings over the Internet. We would like to thank Wayne Burdick, N6KR; Wes Hayward, W7ZOI; and Bill Bridges, W6FA, for their advice.

Notes

[7] Eileen Lau, KE6VWU, et al, "High-Efficiency Class-E Amplifiers—*Part 1*," *QST*, May 1997, pp 39-42. The amplifier enclosures are $3^{1}/_{8}\times 4^{1}/_{8}\times 7^{1}/_{8}$ HWD.

[8] Rick Lindquist, KX4V, "Low-Power Transceiver Kits You Can Build," Product Review, *QST*, Jun 1996, pp 45-50.

[9] See J. B. and R. V. Heaton, "An Electro-Acoustic CW Filter," *QST*, Apr 1983, pp 35-36; Wally Millard, K4JVT, "A Resonant Speaker for CW," Hints and Kinks, *QST*, Dec 1987, p 43; and Richard Clemens, KB8OAB, "More on Resonant Speakers," Hints and Kinks, *QST*, Jan 1989, p 37.—*Ed.*

[10] David Newkirk, WJ1Z (now W9VES), and Rick Karlquist, N6RK, "Mixers, Modulators, and Demodulators," *The 1996 ARRL Handbook*, Chapter 15, p 15.22.

[11] Telephone conversation with Wayne Burdick. This tuning is done by carefully adjusting C50 to peak the output power.

By William E. Sabin, W0IYH

From *QEX*, November 1999

A 100-W MOSFET HF Amplifier

We had the power supply in Mar/Apr[1] and the diplexer filters in Jul/Aug,[2] here's the main event: a reliable FET power amplifier that needs only 10 dBm of drive to produce a pristine 100 W output.

The two-stage amplifier described in this article and shown in Fig 1 is intended for SSB/CW/Data operation on all nine HF amateur bands. An input ($R_{in} \approx 50\ \Omega$) of about 10 mW (+10 dBm) is amplified to 100 W (PEP or average), continuous duty, with a gain of 40 1.0 dB from 1.8 to 29.7 MHz. Third-order, two-tone intermodulation distortion (IMD) products are 35 to 40 dB below 100 W, and higher-order products are also within the high-quality range for amateur SSB equipment, as shown in Fig 2. The main goal for this amplifier is to operate as a driver, with low adjacent-channel interference, for a legal-limit 1500-W linear amplifier. At this power level, adjacent-channel reduction is especially important. And, of course, it is used in the barefoot mode as well.

The power supply (40 V at 8 A) for the push-pull, class-AB MRF150MP (matched-pair) MOSFET output stage was described previously (see Note 1). Six diplexer filters (see Note 2) also known as invulnerable filters provide more than adequate harmonic attenuation for all nine HF bands. They present a broadband load impedance to the MOSFETs that helps to assure freedom from regeneration and oscillation, and good IMD performance. A resistive load impedance between 45 Ω and 55 Ω is recommended for best performance.

The MRF 150 was chosen because it is designed for linear, class-AB SSB operation and because it has high gain (gm) at the 30-MHz end of the HF spectrum. A desirable feature of the MOSFET power transistor is its ability to achieve low values of the higherorder IMD products.[3,4] As mentioned previously, these products contribute to adjacent channel SSB interference. The first stage uses high-gain, class-A push-pull MRF426 (matched-pair) BJTs that require 13.5 V at about 1.0 A from a separate supply. This supply is also the main supply for other system components. Matched pairs of both transistors are available for a small extra fee from at least two sources.[5,6] They are both listed in the current Motorola manual.[7]

One main idea for this amplifier is to operate it in a very low-stress manner that helps assure a low probability of failure for a very long time, which offsets the initial cost of the high-quality transistors. The MRF150 is a 50-V transistor operated at 40 V; the MRF426 is a 28-V transistor operated at 13.5 V. The required input level is low enough that most of the amplification at the signal frequency occurs in one gain block. Because of the good layout, circuit design and decoupling, the 40-dB gain value does not result in any stability problems.

The balanced amplifier greatly reduces even-order harmonics—especially the second—prior to any output filtering, as shown in Fig 3 for a 7.0-MHz signal. It should be 40 dB or more below a 100-W CW signal for each amateur band. This reduction has been found reliable, once achieved. The lowlevel signal source that drives this amplifier must have at least 50 dB of second-harmonic attenuation, since this amplifier will not suppress that harmonic. This is easy to accomplish, but must be considered during the equipment system design (see Fig 9) and while bench-testing the amplifier as shown in Fig 4.

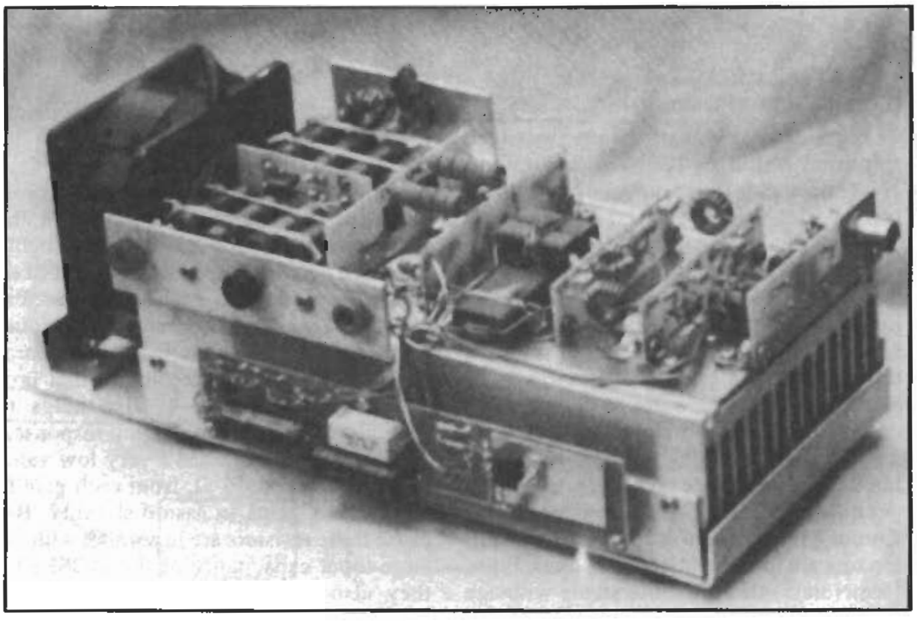

Fig 1—The 100-W broadband amplifier.

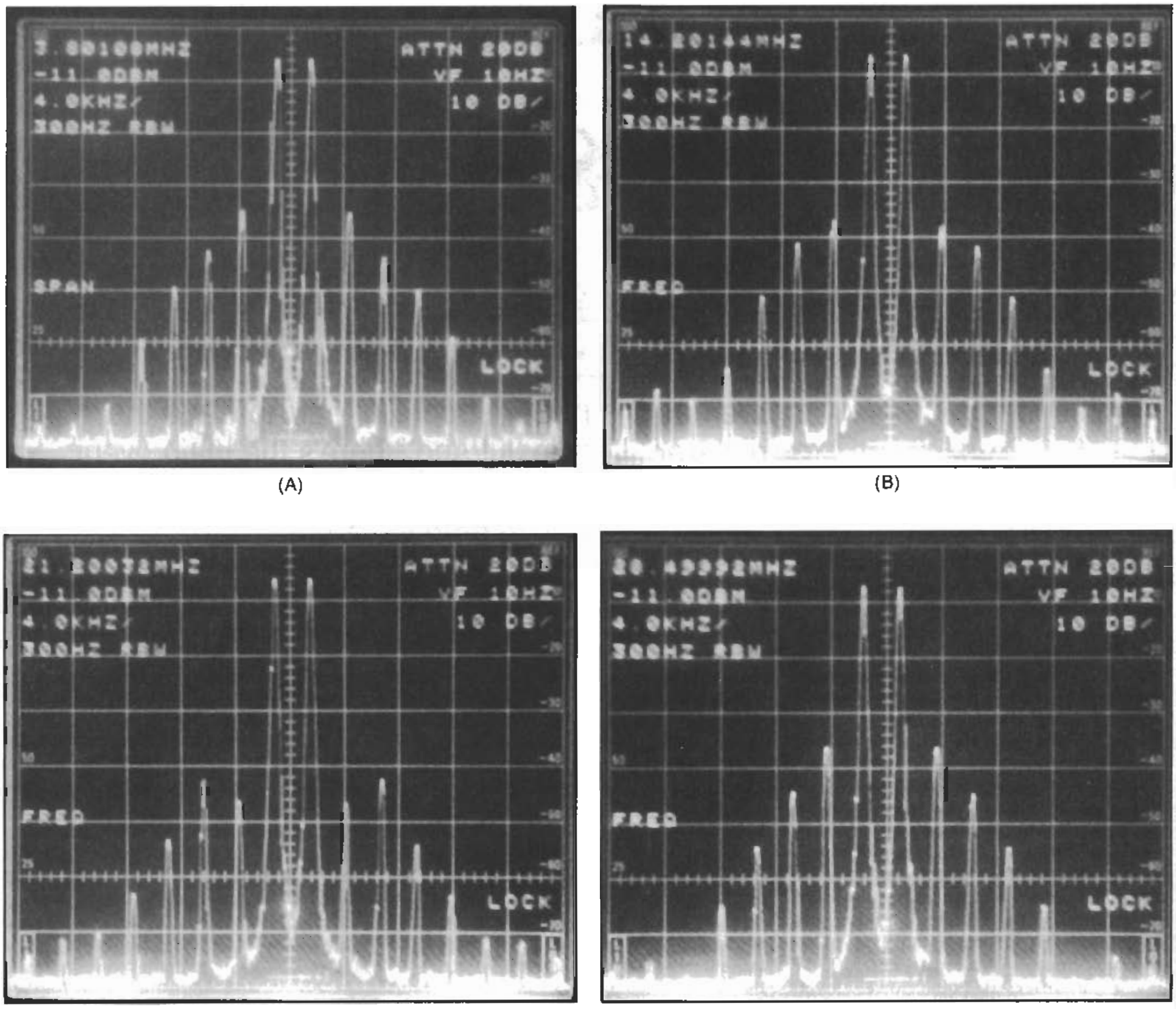

(A)

(B)

(C)

(D)

Fig 2—Two-tone IMD products: (A) 3.8 MHz, (B) 14.2 MHz, (C) 21.2 MHz, (D) 28.5 MHz.

Circuit General Discussion

Fig 5 is the schematic of the two-stage amplifier. It utilizes 1:1 choke baluns and 1:4 (impedance) step-up and step-down transmission line (Guanella) transformers. The choke baluns significantly improve the balance of the input and output stages. Notice also that T2, T3 and T4 have floating center taps rather than bypasses to ground. This is recommended to improve the even-harmonic balance[8] (verified). T4 and T5 run slightly warm as compared to conventional transformers that get quite hot. The 5-W feedback resistors get warm, but their large surface area limits their temperature rise. They also receive cooling air from the fan. All power resistors in the RF circuits are the excellent metal-oxide types (see Note 5) that are quite stable with age and temperature, and have very low reactance at 30 MHz (measured). Metal-film 1% resistors are used in several critical locations.

The first stage has resistance and inductance loading from collector to collector. A powerful free-running oscillation in the first stage at about 32.5 MHz was being triggered occasionally at the moment of dc supply turn-on while a fairly large 28.0 to 29.7 MHz input was present. The loading reduces stage gain and suppresses the parasitic collector circuit resonance that produced the oscillation. It also helps to flatten the amplifier frequency response.

The second stage has a very low value of RF resistance, 15 Ω, from each gate to ground that helps to assure stability. Because these resistors are in parallel with the large input capacitance of the MOSFETs, they also help to flatten the frequency response. Both stages have negative feedback networks that further assure stability and flatness of frequency response.

Two biasing networks are employed. The LM317 provides a highly regulated gate voltage for the second stage. The FET dissipation and linearity are very sensitive to this voltage; the bias voltage for SSB is fine-tuned for best signal purity using two-tone tests and a spectrum analyzer. Note the four 562-Ω resistors. If the LM317 fails short circuit, the voltage on the FET gates does not exceed 6.8 V, which will not damage the gates. The large drain current that results from this failure also does no damage because the 40-V power supply has current limiting and voltage fold-back that prevent harmful FET dissipation. The FETs are thus kept well within the safe-operating-area (SOAR) as defined in data sheets[9] (verified). It is important that each 562-Ω resistor from gate to ground be permanently attached directly between the gate and source tabs of the FETs themselves so that the gates are never floating. Wrap the resistor leads around the tabs so that they cannot come loose. This avoids accidental static charges that might ruin them. On the other hand, it is better to tack and not wrap the base-collector and gate-drain signal leads so that they can be easily disconnected. We want to be able to test individual segments of the circuitry easily. In my experience, the MRF150 has proven to be a rugged transistor, much more so when operated conservatively and with the power-supply safeguards that I mentioned.

The LM317 provides about 5.8 V for the particular pair of FETs that I used. Individual pairs of FETs will probably require an adjustment of this voltage for idling current and for IMD products that resemble Fig 2. The adjustment procedure is described later. This bias value is set to emphasize the reduction of the higher-order products rather than third- and fifth-order products. These lower-order products do not contribute as much to adjacent channel interference; in fact, an SSB speech processor will make them worse anyway. The higher-order products need to be reduced and the MRF150 has this capability.

The 2N3906 PNP transistor is a bias-current source. The value of this current is determined by the 4.7-Ω resistor and the base-to-ground voltage of the 2N3906. The two 562-Ω resistors force equal base currents into the MRF426s. This helps to assure equal performance. The 0.47-Ω resistor is part of a negative-feedback bias loop. If collector current increases, the voltage across this resistor increases and this reduces the base current. Thermal runaway is avoided by this strategy because the reduced base bias restricts the current increase to a small value. Because of the low dissipation of the MRF426s and the good heat sink, this method is very effective. These transistors are also well inside their SOAR requirements. This stage is very linear and contributes almost nothing to the overall IMD products. To verify this, it is necessary to turn off the drain voltage to the MRF150s and connect a spectrum analyzer across one of the 15-Ω resistors.

The diplexer filter method was decided uponafter a lot of experimentation with other low-pass filter methods as a solution that is free of problems caused by complex interactions between the filters and the output transistors. A special peculiar problem is discussed later. The MRF150 has high gain into the VHF region; the diplexers eliminated all problems associated with this fact. This approach is recommended as a simple way to assure correct operation for HF amateur-band operation of these high-frequency MOSFETs. I was able to get good enough operation with the more conventional low-pass filters, but this approach was by far the most satisfactory for an amateur-band amplitier, as confirmed by swept-frequency tests at all power levels into a 50-Ω load. More about complex loads later.

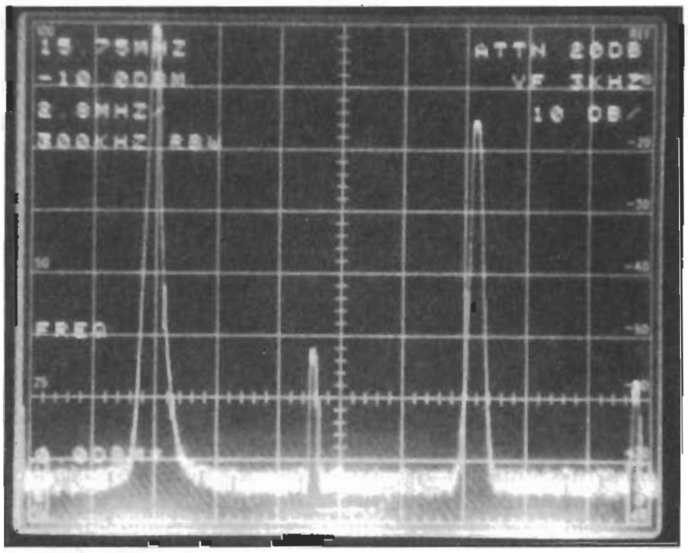

Fig 3—Wideband spectrum for a 7-MHz signal.

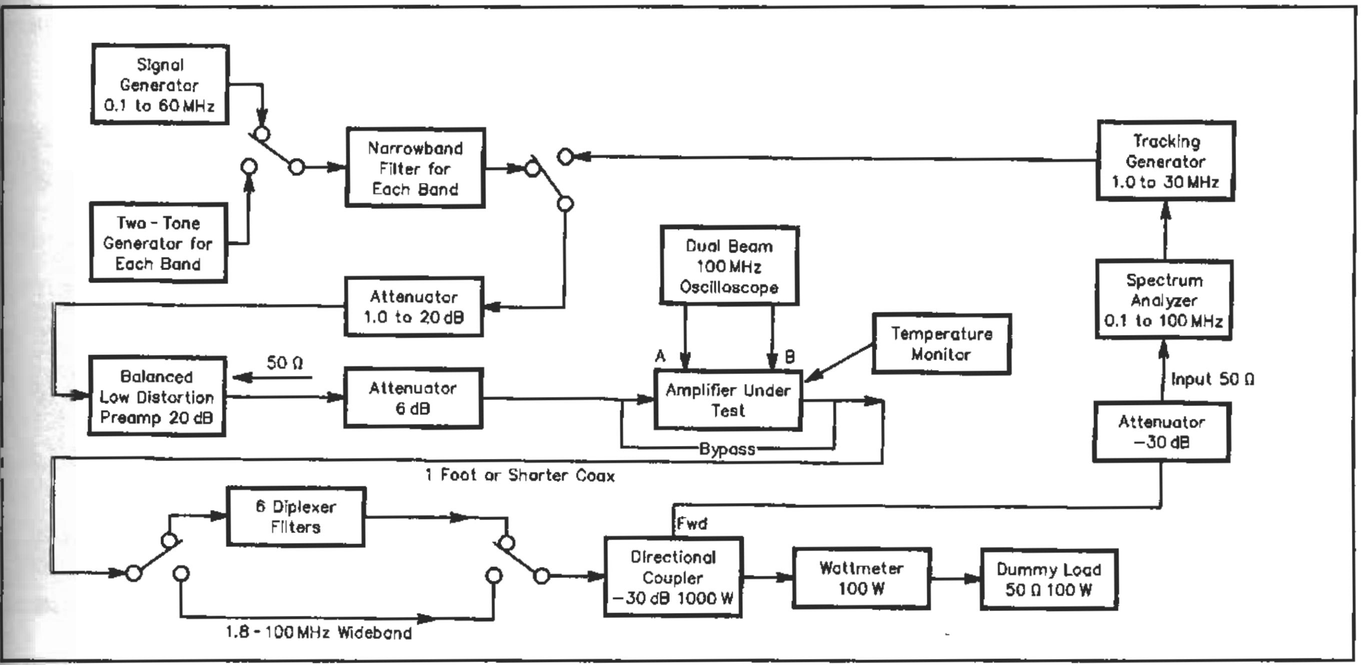
Signal Generator 0.1 to 60 MHz
Two-Tone Generator for Each Band
Narrowband Filter for Each Band
Tracking Generator 1.0 to 30 MHz
Attenuator 1.0 to 20 dB
Dual Beam 100 MHz Oscilloscope
Spectrum Analyzer 0.1 to 100 MHz
Temperature Monitor
A
B
50 Ω
Balanced Low Distortion Preamp 20 dB
Attenuator 6 dB
Amplifier Under Test
Input 50 Ω
Attenuator −30 dB
Bypass
1 Foot or Shorter Coax
6 Diplexer Filters
Fwd
Directional Coupler −30 dB 1000 W
Wattmeter 100 W
Dummy Load 50 Ω 100 W
1.8 - 100 MHz Wideband

Fig 4—Block diagram for equipment setup, design and testing of the MOSFET amplifier.

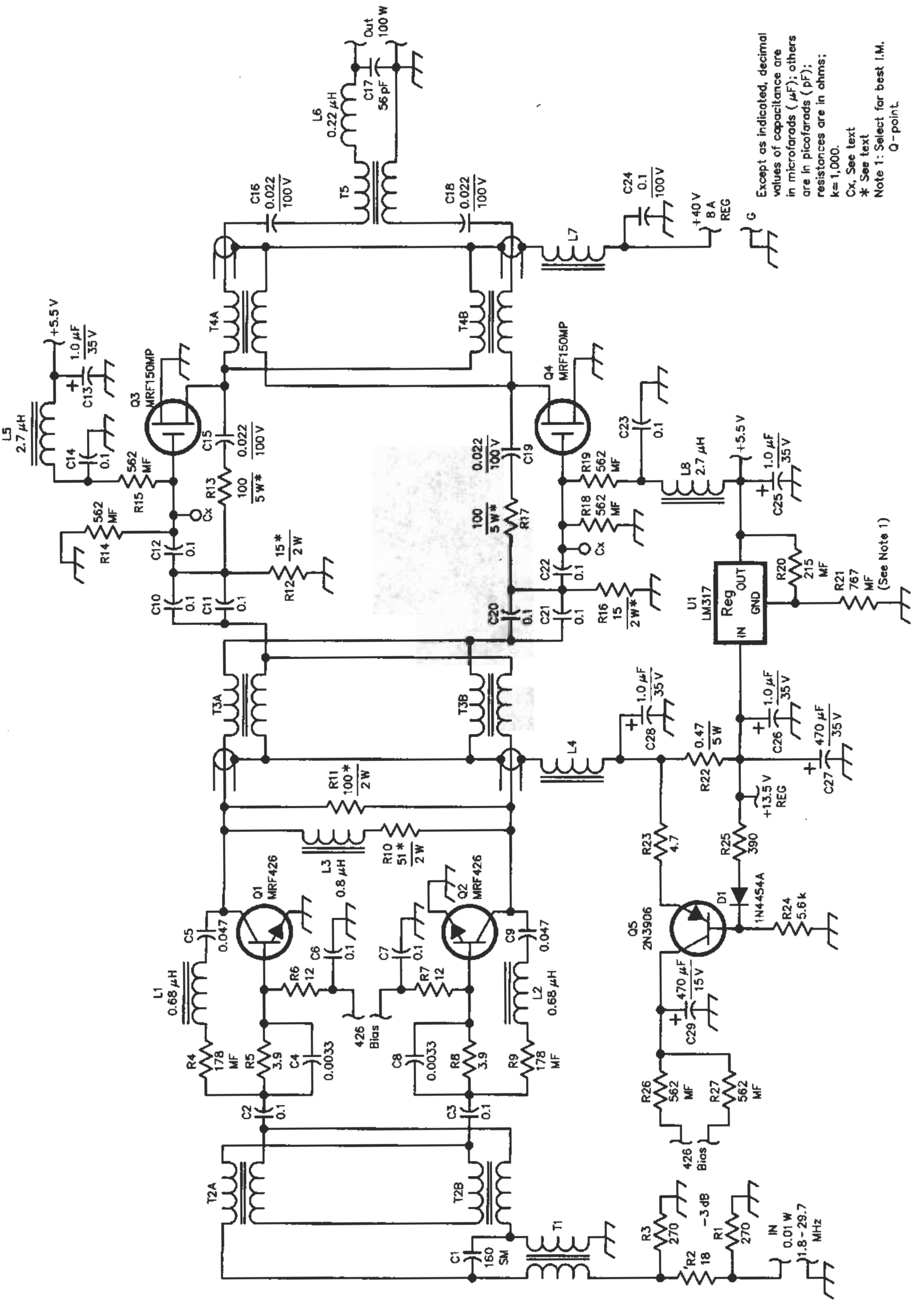
Out
100 W
C17
56 pF
L6
0.22 µH
T5
C16
0.022
100 V
C18
0.022
100 V
C24
0.1
100 V
+40 V
8 A
REG
G
L7
+5.5 V
1.0 µF
35 V
C13
L5
2.7 µH
C14
0.1
Q3
MRF150MP
T4A
T4B
Q4
MRF150MP
R15
562
MF
R14
562
MF
C15
0.022
100 V
R13
100
5 W*
Cx
C12
0.1
R12
15*
2 W
C10
0.1
C11
0.1
C19
0.022
100 V
R17
100
5 W*
C20
0.1
C21
0.1
C22
0.1
R16
15
2 W*
Cx
R18
562
MF
R19
562
MF
C23
0.1
L8
2.7 µH
+5.5 V
C25
1.0 µF
35 V
U1
LM317
Reg
IN
OUT
GND
R20
215
MF
R21
767
MF
(See Note 1)
T3A
T3B
L4
C28
1.0 µF
35 V
R22
0.47
5 W
C26
1.0 µF
35 V
C27
470 µF
35 V
+13.5 V
REG
R11
100*
2 W
L3
0.8 µH
R10
51*
2 W
Q1
MRF426
Q2
MRF426
C5
0.047
C9
0.047
L1
0.68 µH
L2
0.68 µH
C6
0.1
C7
0.1
R6
12
R7
12
426
Bias
R4
178
MF
R5
3.9
C4
0.0033
C8
0.0033
R8
3.9
R9
178
MF
C2
0.1
C3
0.1
R23
4.7
R25
390
D1
1N4454A
R24
5.6 k
Q5
2N3906
C29
470 µF
15 V
R26
562
MF
R27
562
MF
426
Bias
T2A
T2B
T1
C1
160
SM
R3
270
R2
18
−3 dB
R1
270
IN
0.01 W
1.8-29.7
MHz

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000.

Cx, See text

* See text

Note 1: Select for best I.M. Q-point.

Fig 5—(left) Schematic diagram of a two-stage, 100-W amplifier. Unless otherwise specified, use 1/4 W, 5%-tolerance carbon composition or film resistors. Resistors marked with an asterisk are metal-oxide units. MF indicates metal-film resistors. Both the MRF426 and MRF150 devices are matched pairs. All capacitors are 50 V, unless otherwise indicated.

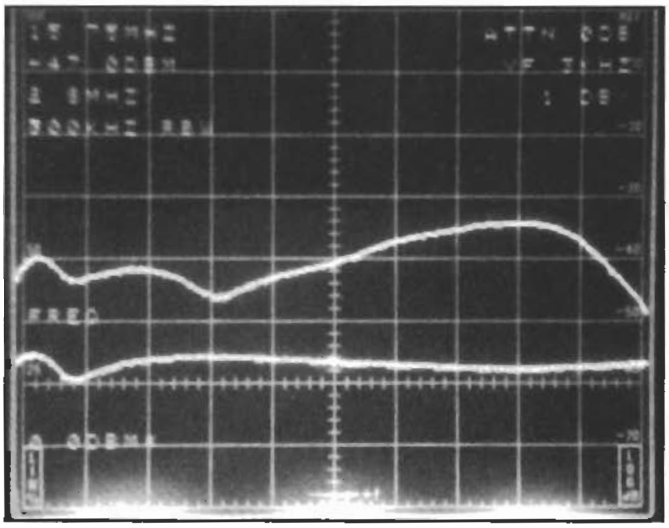

Fig 6—Gain variation, 1.8 MHz to 29.7 MHz, 1.0 dB per division vertical scale—The lower trace is a reference sweep that bypasses the 100-W amplifier—The response bump at low frequency is an artifact of the tracking generator.

Frequency Compensation

The bipolar transistors have gain values that decrease as frequency increases. The MOSFETS have large capacitances that also affect frequency-response roll-off. It was a major exercise to design networks that flatten the response from 1.8 to 29.7 MHz. Fig 5 shows the approach. The idea is to compensate smoothly from input to output in such a way that neither of the two stages is over-driven at any frequency. The criterion for this is to check IMD products and harmonics during the design process, which involves approximate analysis (see the MOSFET Stage Simulation sidebar) and negative feedback.[10] Beyond about 32 MHz, the gain falls off fairly rapidly (but not too rapidly). This is also desirable.

When testing the frequency response using the test setup of Fig 4, it is necessary to measure the frequency response of the signal path from tracking generator to spectrum analyzer while bypassing the 100-W amplifier. This reference response is then compared with the response with the amplifier in the path, as shown in Fig 6. Note the vertical scale: 1.0 dB per division. For the most credible results and ease of measurement, it is very desirable that the impedance looking back from the input be 50 Ω.

Signal Level Testing

We want to verify that the first stage is operating normally by measuring its RF voltages at 4.0 MHz. The class-A first-stage input impedance is close to 50 Ω, and the input level is +10 dBm. Temporarily disconnect the signal leads to the gates of the FETs. The voltage at each output of T3 is about 1.35 V. Reconnect the gate signal leads. The final output over the entire range into an unfiltered, wide-band 50-Ω load is then observed using the setup in Fig 4. A spectrum analyzer with tracking generator is very valuable for this. The accuracy of the wattmeter should be verified or calibrated by some means at both the 100-W and the 25-W levels (the power of each tone of a two-tone, 100-W PEP signal) in each amateur band.

If the second stage is working correctly, the output power is 100 W, or 71 V RMS across 50 Ω. These procedures assure that both stages are working properly and amplifying as intended. Because of the variations in the fabrication of the transistors, the total gain can vary a decibel up or down despite the use of negative feedback in each stage. I suggest the 3-dB attenuator at the input not be modified for simplicity reasons.

The drain-to-drain load impedance of the class-AB output stage is 12.5 Ω and the CW output power is 100 W, so the drain-to-drain ac voltage is 35.4 V. Fig 7 shows the dual-trace scope waveforms (chop mode) on each drain, superimposed on the 40-V supply, and also the drain-to-drain waveform that is confined to the linear region. The third-harmonic content is visible. It is interesting to note that although these waveforms show considerable non-linearity, the fundamental component is quite linear with respect to the gate signal level. I also found that a 50-V dc supply created more heat, but at 100W, did not improve linearity enough to make it worthwhile. The 35.4 V ac also appears across the two series-connected 100-Ω feedback resis-

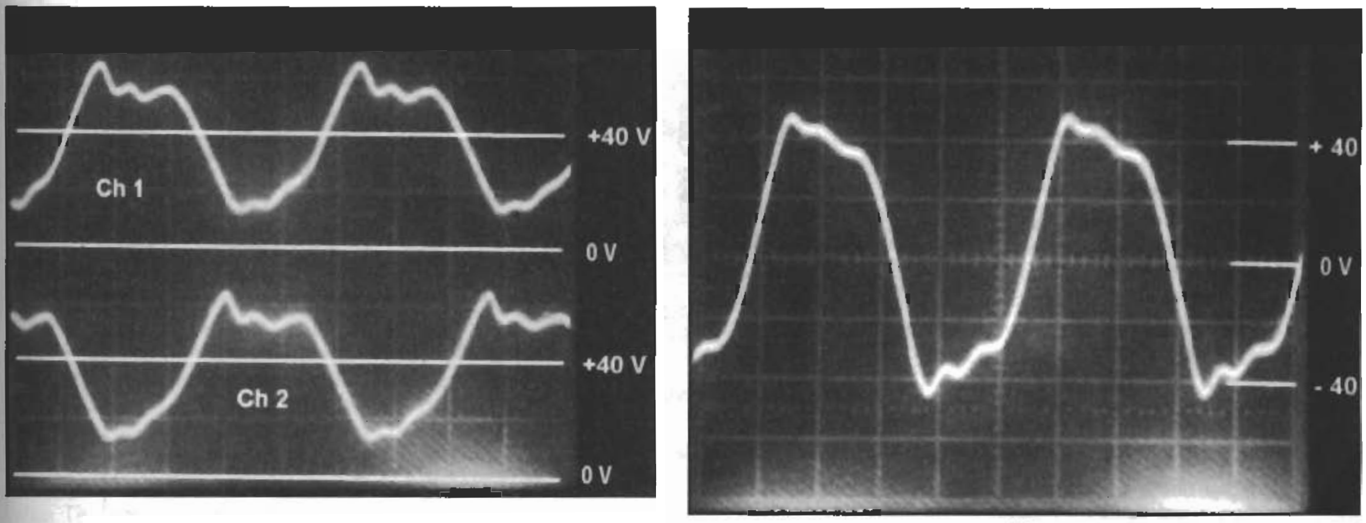

(A)

(B)

Fig 7—(A) drain and (B) drain-to-drain oscilloscope waveforms.

tors; they dissipate about (35.42)/200 = 6.3 W, or 3.2 W per resistor (64% of the 5 W rating).

A broadband, untuned power amplifier with flat frequency response and low distortion is not, by necessity, especially energy-efficient (ratio of RF output power to dc power). The output stage is less than 40% efficient for this reason. For SSB use, where the average power is not more that 25 W, even with speech processing, this is no problem at all. For continuous key-down at 100 W, the cooling fan is more than adequate.

Here is an important caution about using oscilloscope probes at the FET drains and the output connector: A 10:1 probe could be damaged (it happened to me) if used directly at this RF voltage level, especially at the upper end of the HF range. Use instead the homemade probe described previously (see Note 2) with a 50-Ω terminating resistor.

The quality of balance is checked by looking at the second harmonic on a spectrum analyzer using the setup in Fig 4 with the diplexer filters out of the signal path. At 100 W output, check the harmonics in each of the nine amateur bands. Capacitor C_x in Fig 5 is used (if needed) to improve the secondharmonic phase balance on the higher amateur bands. If the value is not well below 40 dBc, use C_x at one gate or the other to achieve 45 dBc. Extra pads are provided on the PC board for this capacitor. A value of between 22 pF and 56 pF should be adequate. Overkill is neither necessary nor desirable: The output filters will do the rest. For each decibel of power output below 100 W, the second harmonic normally drops about two decibels.

Construction Notes

Fig 1 shows how my version of the amplifier is constructed. A 0.125-inch (or 0.062-inch) double-clad PC board, 3.25 × 8.0 inches, is firmly attached to the heat sink. I suggest using this compact size for best reproducibility. The heat sink shown may not be presently available from RF Parts (see Note 5), but I also purchased a Model 99 sink from CCI (see Note 6) that is 6.5×12 inches. This can be easily tailored to the appropriate size with a band saw that has a metal-cutting blade. Pieces of angle and sheet aluminum can then be creatively fashioned to accommodate the fan (RadioShack #273-242) and the PC board that contains the bias circuitry. The transistors are bolted directly to the heat sink (through cutouts in the PC board) using heat-sink compound and tapped (and carefully deburred) #4-40 holes. Careful mounting of the transistors is essential. The main PC board has small sections of copper removed underneath the base and collector tabs of the MRF426s and underneath the gate and drain tabs of the MRF150s, so that accidental grounding is avoided. Use a hobby knife to define the areas and a hot soldering iron to peel off the copper.

A bottom cover is advised, as shown, so that air is funneled through the heat-sink fins efficiently. For a 100-W, continuous, single-tone output, the exhaust air temperature reaches 50°C. I used a RadioShack #22-174 multimeter with its temperature probe mounted inside the fins to monitor temperature during the design and testing phase.

The components are mounted on a set of seven small PC boards, six of which are mounted vertically (as shown) and bolted to the drilled-andtapped heat sink through the PC board, using #4-40 screws and small angle brackets. I used stiff, right angle #6 solder lugs that worked out very nicely. The seven PC boards are cut from a single 4×6-inch two-sided PC board, shown in Fig 8. The bias circuit board has a ground plane; the others do not. This set of boards is available from FAR Circuits.[11]

I chose this method because it is easy, makes the amplifier more compact, reduces stray L and C that can degrade the wideband frequency response and reduces stray couplings that can impair stability and harmonic balance. It also allows the ground plane to be one continuous surface, which is a plus factor. This approach worked out very well and I recommend it as a simple approach.

Temperature Rise

The cooling fan is important. This amplifier, as designed, should have the fan running, and I have found that a simple and reliable way to keep everything safe. It has been tested at 100 W continuously, for sev-

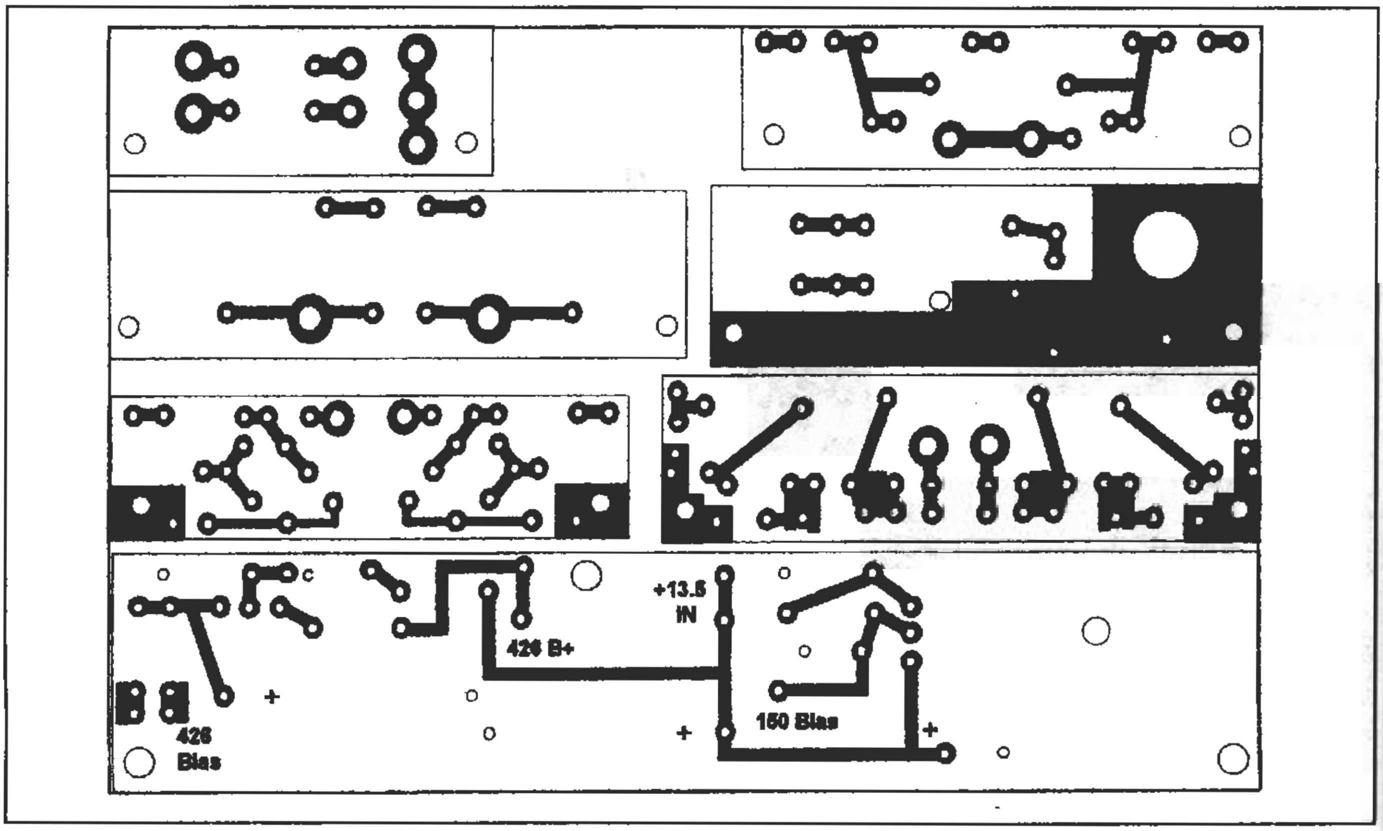

Fig 8—Etching pattern for a 4×6 inch (finished size) PC board that provides the seven individual boards.

MOSFET Stage Simulation

A simplified analysis of the second MOSFET stage is presented (Fig A) to illustrate the effect of the resistive negative feedback of the 100-Ω resistors and the 40-pF gateto-drain capacitance. The simulation diagram shows the voltagecontrolled current sources with a G_m of 6.0 S, which is assumed constant over frequency. The gate-to-source capacitance is 360 pF and the drain-to-source capacitance is 200 pF. These numbers are from the MRF150 data sheets.

The frequency plot (Fig B) obtained from the ARRL Radio Designer program shows the gain, MS21 (dB), with (lower trace) and without (middle trace) the 100-Ω feedback resistors. The gain variation is reduced from about 6 dB to about 1.5 dB. In this simplified model (good enough for this illustration), the gain drop is caused by the capacitors, especially the 40-pF capacitor whose influence is greatly magnified by the Miller effect. If this 40-pF capacitance is eliminated from the simulation, the frequency response variation—even without the feedback resistors—is less than 1.0 dB, as seen on the upper trace.

If the power output at 15 MHz is to be held constant, the turns ratio of the output transformer and the value of the feedback resistors can be manipulated to achieve that end. The simulation shows that a turns ratio of 1: $\sqrt{2}$ and a resistor value of 120 Ω does this. The FET load impedance is now 25 Ω instead of 12.5 Ω. The increased voltage gain makes the Miller effect greater and increases the gain variation (lower trace) to 2.3 dB. The drain-to-drain voltage is now 50 V instead of 35 V, and this results in higher drain efficiency for the MOSFETs. However, the push-pull 1: $\sqrt{2}$ -turns-ratio transmission-line transformer is not as simple as the 1:2 that uses 25-Ω coax, and for me, that was the determining factor.

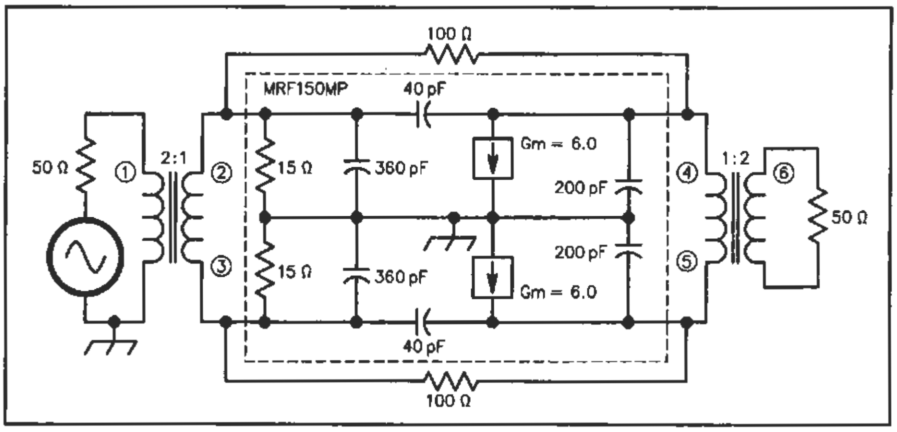

Fig A—A circuit used to analyze the PA output stage with ARD.

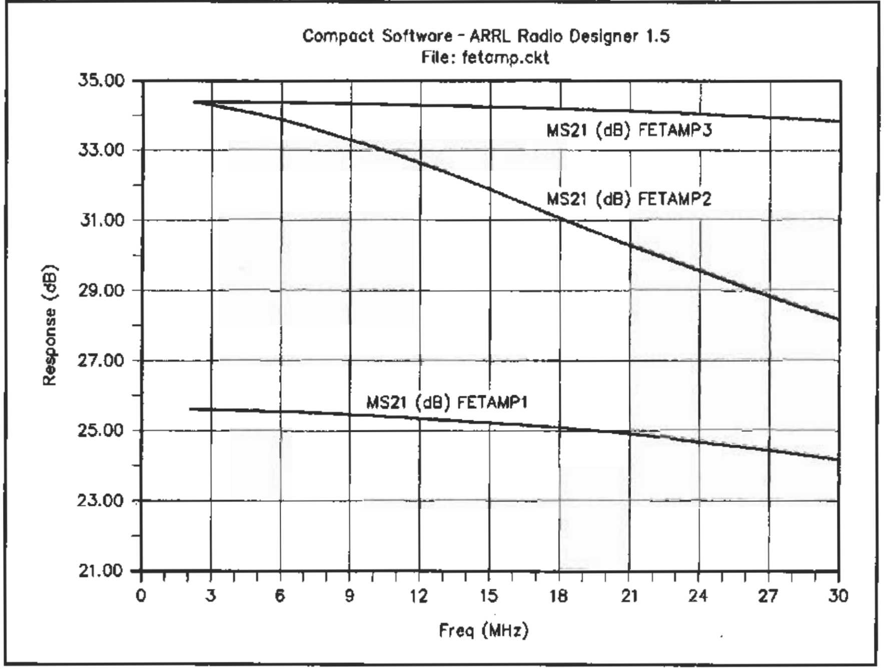

Fig B—The analysis results.

eral hours. The MOSFETswell separated on a good heat sink and with efficient airflow do not have a temperature-controlled gate-bias arrangement because I did not find it necessary at this power level. Because of the spread in threshold voltage of individual matched pairs, a one-time adjustment of gate bias is needed. The following simple procedure is used:

- Replace R21 with a resistor decade box set at 750 11.
- Place a 0-10 A meter in the +40-V line.
- With *no* signal input, switch on the amplifier and let the current reach its final value, which should reach about 1.5 A as the FETs warm up.
- Adjust the resistor value until DD is 1.5 A. After each adjustment, allow time for the FETs to reach a steady current value. This final value is not critical, but should be within 10%.
- Check the two-tone IMD at 100 W PEP (50 W average) and make further small changes, if needed, to resemble the IMD patterns in Fig 2.

In this amplifier, we are able to use the self-limiting feature of the MOSFET. This approach would not be appropriate in higher-power, higher-temperature amplifiers because of thermal runaway possibilities, where the decrease with temperature rise of the gate threshold voltage exceeds the decrease of the dc transconductance.[12,13] If fan reliability is a concern, add a thermal cutout switch to the +40-V line.

Load Impedance

Swept- and fixed-frequency tests at many signal levels and voltage valuesusing various values of parallel conductance and capacitive/inductive susceptancedid not reveal instability problems at SWR values less than 2:1. A tendency to oscillate in a peculiar manner occurred in the 20.0 to 29.7 MHz range, but only with input signal applied (a driven oscillation) and certain values of drive level, supply voltages less than 30 V, SWR of 3:1 or more, and a certain range of shunt capacitance at the output of the amplifier itself or at the output of the 12/10 meter diplexer. The instability shows up as a Christmas tree pattern in which discrete, uniformly-spaced 2- to 4-MHz sidebands appear on the main carrier. It is due to a parametric effect[14] involving the voltage-variable capacitances of the MOSFET transistors that is somehow enhanced by the shunt test capacitance. Shunt inductance did not produce the problem. The amplitude modulation of the carrier by the low-frequency oscillation is clearly visible on an oscilloscope. The low-frequency oscillation itself can be seen on a spectrum analyzer. At SWR values less than 2:1 and drain voltages greater than 35 V (40 V is recommended), the safety margin is large. The 0.22-µH, 56-pF network at the output greatly assisted with this problem. A coax cable from the PA to the diplexer assembly that is a foot or less duplicates my test setup. Under these conditions, there is no stability problem on any band caused by load-impedance values that my testing could identify. Actually, a resistive load between 45 Ω and 55 Ω is recommended for best SSB linearity, as mentioned previously. While driving a high-power linear amplifier to its rated two-tone SSB output, adjust its input impedance to 50 Ω resistive as closely as possible on each band, using a 50-Ω directional wattmeter. A broadband (untuned) solid-state driver amplifier such as this one requires such attention to load value for best results as compared to pi-network vacuum-tube PAs that can transform a fairly wide range of complex load impedance values to the correct tube plate load resistance at resonance.

It is difficult to achieve stability in a broadband, transistor power amplifier. Having done so, it is still wise to operate into the correct 50-Ω load as closely as possible. The monitoring of forward and reflected power in Fig 9 is helpful for this and assures clean, stable and reliable operation. This circuitry will cut back the drive level to a value that protects the output stage from excessive drain current or gate drive. Incidentally, the first stage goes into saturation far below the level that would damage the MOSFETs.

System Design

Fig 9 suggests a system implementation of the amplifier. For the flattest frequency response of the complete system from 1.8 to 29.7 MHz, the circuitry that drives this amplifier should also have a flat response and a 50-Ω output resistance—both easy to achieve. If it is not perfectly flat or not exactly 50 Ω, it may be necessary to slightly adjust the drive level on each band, which is not difficult. As mentioned before, the second-harmonic production of the circuitry preceding this amplifier must be of sufficiently low level (50 dBc) that balanced, push-pull operation is necessary. Narrowband resonator filters are also very commonly used for amateur-band harmonic reduction in low-level exciter circuitry.[15,16]

The directional coupler detects excessive forward and reflected power. Both of these are displayed on a panel meter. The directional couplers forward port is also a

Parts List

C1—160 pF SM
C2, 3, 6, 7, 10-12, 14, 20-23—0.1 µF, 50-V CK05
C4, 8—0.0033 µF, 50-V CKO5
C5, 9—0.047 µF, 50-V CKO5
C13, 25, 26, 28—1.0 µF, 35-V tantalum
C15, 16, 18, 19—0.022 µF, 100-V CK05
C17—56 pF SM
C24—0.1 µF, 100-V CKO5
C27—470 µF, 35-V aluminum
C29—470 µF, 15-V aluminum
D1—1N4454A or equivalent
L1, 2—0.68 µH, T50-2 core, 10 turns #26 AWG
L3—0.80 µH, T50-2 core, 12 turns #26 AWG
L4—BN-43-3312, 4½ turns #22 hookup wire
L5, 8—2.7 µH molded
L6—0.22 µH, T50-2 core, 5 turns #26
L7—2 FB-43-5621 cores, 1½ turns #12 stranded
Q1, 2—MRF426 matched pair
Q3, 4—MRF150 matched pair
Q5—2N3906 PNP
R1, 3—270 Ω, ¼ W, 5% tolerance
R2—18 Ω, ¼ W, 5% tolerance
R4, 9—178 Ω, ⅛ W metal film, 1% tolerance
R5, 8—3.9 Ω, ¼ W, 5% tolerance
R6, 7—12 Ω, ¼ W, 5% tolerance
R10—51 Ω, 2 W metal oxide
R11—100 Ω, 2 W metal oxide
R12, 16—51 Ω, 2 W metal oxide
R13, 17—100 Ω, 5 W metal oxide
R14, 15, 18, 19, 26, 27—562 Ω, ⅛ W metal film, 1% tolerance
R20—215 Ω, ⅛ W metal film 1% tolerance
R21—767 Ω, ⅛ W metal film 1% tolerance, test selected
R22—0.47 Ω, 5W wire-wound
R23—4.7 Ω, ¼ W 5% tolerance
R24—5.6 kΩ, ¼ W, 5%
R25—390 Ω, ¼ W, 5%
T1—BN-43-202, 2½ turns #32 AWG, bifilar
T2A, B—BN-43-202, 2½ turns #32 AWG, bifilar
T3A, B—BN-43-3312, 2½ turns 25-Ω miniature coax*
T4A, B—(2) FB-43-5621, 1½ turns 25-Ω miniature coax*
T5—(2) FB-43-5621, 2½ turns 50-Ω miniature coax
U1—LM317 adjustable voltage regulator

Notes

*Microdot D260-4ii8-0000 available from Communication Concepts, Inc.
All cores are available from Amidon.
Closely matched transistor pairs are from RF Parts (see Note 5).

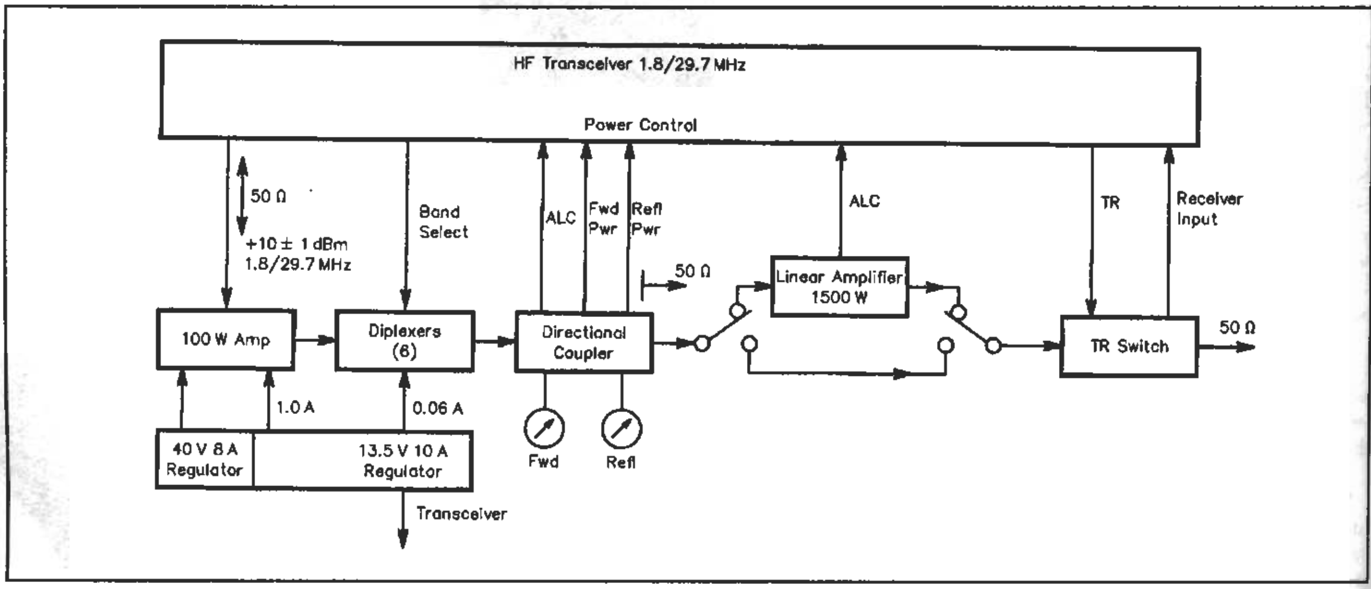

Fig 9—System implementation of the 100-W amplifier.

source of ALC voltage. These control voltages are fed back to the appropriate gain-controlled stages in the exciter. Gain control of the transistors in the 100-W amplifier is not recommended because changes in base or gate bias will degrade amplifier linearity, IMD products and possibly the flat frequency response. There are better ways to accomplish the gain-control task, such as preferably in a lowlevel IF amplifier. With respect to IMD, we are trying to control the odd-order curvature of the MOSFET transfer characteristic—especially for the higher-order products—by setting the gate-bias point. The bias value is found by looking at IMD on each amateur band and selecting the best compromise value. This was discussed previously in this article.

The diplexer filters are linked to the other band-switch circuitry so that the correct filter is always switched in. The TR relay is at the output of the diplexers. Do not hot-switch the diplexer filters because this might damage the inexpensive relays.

Conclusion

In conjunction with the power supply (see Note 1) and the diplexer filters (see Note 2), the amplifier described here is a basic module for homebrew equipment that should satisfy the requirements for a 100-W power level, a 1.8 to 29.7 MHz bandwidth and a high-quality signal. It should run trouble-free for a very long time. The initial cost of the four transistors is about $180, but they will last indefinitely if cared for properly. The ones that I use were abused considerably during the experimentation and continue to work perfectly. The approach that I suggest is best implemented with diplexer filters and a preamplifier that has very low second-harmonic output and low IMDboth easy to get at +10 dBm. The payoff for me is excellent performance and reliability, once the design was completed.

This project is suggested for the homebrew enthusiast who has at least some part-time access to lab-quality test equipment. Others who do not care to build may find the article, together with numerous other sources, interesting background information regarding MOSFET power-amplifier design and test methods.

Any attempt at building and testing this amplifier should also use the 40-V power supply in the referent of Note 1 or something similar. The automatic current limiting at 8 A, the automatic reduction of drain voltage, the short-circuit limiting to 4 A and the manual control down to 24 V help to protect the MOSFETs from mishaps that are bound to occur.

Notes

[1] W. Sabin, W0IYH, "Power Supply for a MOSFET Power Amplifier," *QEX*, Mar/Apr 1999, pp 50-54.

[2] W. Sabin, W0IYH, "Diplexer Filters for an HF MOSFET Power Amplifier," *QEX*, Jul/Aug 1999, pp 20-26.

[3] N. Dye and H. Granberg, *Radio Frequency Transistors* (New York: Butterworth-Heinemann, 1993), pp 42-43.

[4] W. Sabin and E. Schoenike, *HF Radio Systems and Circuits* (Tucker, Georgia: Noble Publishing Co, 1998), Chapter 12, Solid State Power Amplifiers, by R. Blocksome. This chapter is highly recommended regarding solid-state SSB power amplifiers. This book is also ARRL Order No. 7253. ARRL publications are available from your local ARRL dealer or directly from the ARRL. Mail orders to Pub Sales Dept, ARRL, 225 Main St, Newington, CT 06111-1494. You can call us toll-free at tel 888-277-5289; fax your order to 860-594- 0303; or send e-mail to **pubsales@arrl.org**. Check out the full ARRL publications line on the World Wide Web at **http://www.arrl.org/catalog**.

[5] RF Parts Co, 4355 Pacific St, San Marcos, CA 92069; tel 760-744-0700, 800-737-2787 (orders only), fax 760-744-1943; **rfp@rfparts.com**, **http://www.rfparts.com/**.

[6] Communication Concepts, Inc. (CCI), 508 Millstone Dr, Beavercreek, OH 45434; tel 937-426-8600, fax 937-429-3811; **cci.dayton@pobox.com**, **http://www. communication-concepts.com/**; ask for a catalog.

[7] *Motorola Wireless Semiconductor Solutions, Device Data*, Rev 9, Vol. II; Motorola #DL110/D. The data book is available from RF Parts (see Note 5). (A Motorola Selector Guide is available at **http://www.motsps.com/books/sguides/pdf/sg46rev18.pdf**. There is a MRF150 page at **http://www.mot-sps.com/products/rf_and_if/rf_transistors/high_power/power_mosfets/mrf150.html**. A MRF426 datasheet (MRF426/D)is available at **http://www.zettweb.com/CDROMs/cdrom013/pdf/mrf426rev0.pdf**, which is not a Motorola site—*Ed.*)

[8] Dye and Granberg, pp 113-115.

[9] *Motorola Wireless Semiconductor Solutions*; see pp 4.2-133 for the MRF151, similar to the MRF150. (See Note 7 for an MRF150 URL—*Ed.*)

[10] Dye and Granberg, Chapter 12.

[11] The set of boards costs $12 plus shipping and handling. FAR Circuits, 18N640 Field Ct, Dundee, IL 60118-9269; tel 847-836-9148 (Voice mail), fax 847-836-9148 (same as voice mail); **farcir@ais.net**, **http://www.cl.ais.net/farcir/**.

[12] Dye and Granberg, Chapter 4.

[13] H. Granberg, Wideband RF Power Amplifier, *RF design*, Feb 1988; also *Motorola RF Application Reports*, AR313, p 424.

[14] Dye and Granberg, similar top 124, Fig 7.7.

[15] *The ARRL Handbook* (Newington, CT: 1995-2000 editions), ARRL Order No. 1832; pp 17.60–17.66 and Fig 17.69. See Note 4 for purchasing information.

[16] W. Sabin, W0IYH, "Designing Narrow Band-Pass Filters with a BASIC Program," *QST*, May 1983, pp 23-29.

By Jim Valdes, WA1GPO From *QST*, June 2003

The FARA HF Project

Would you like an economical and relatively easy to build 30 W HF amplifier? For about $130 worth of parts you can give that QRP rig an extra 12 dB of muscle!

One of the larger and more active radio clubs in the Cape Cod, Massachusetts area is the Falmouth Amateur Radio Association (FARA). Some of its members are affectionately referred to as the "hackers"–those who enjoy the construction phase of Amateur Radio. The renewed interest in low power (QRP) radios in the 2 to 5 W output range resulted in the desire for a 10 to 12 dB gain RF linear amplifier producing 30 to 40 W of output power. Commercial amplifiers capable of being driven with less than 50 W below 30 MHz are prohibited by FCC regulations.[1] As a radio amateur, you are permitted to construct one amplifier per year for your personal use. The amplifier described here is relatively easy to fabricate, given a basic knowledge of electronics and some familiarity with hand tools. This article describes in detail the construction of a 12 dB, nominal 30 W output, 1.8 to 30 MHz RF power amplifier. It is intended for 12 to 14.7 V dc operation, making it ideal for mobile use.

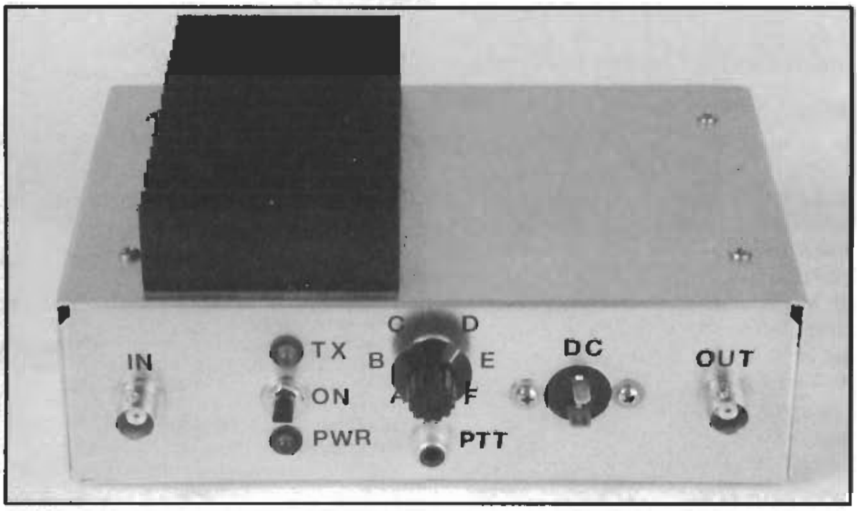

Figure 1—The completed 30 W amplifier. The band switch takes care of output low-pass filter switching.

Amplifier Description

The design of the amplifier is based upon common engineering practices. Ideas gathered from researching the handbooks published by the ARRL and The Radio Society of Great Britain (RSGB), as well as articles published in *QST* and other journals formed the basis of the design. This amplifier uses readily available components...many of the earlier designs were based on the Motorola MRF series of RF devices that are no longer available or are prohibitively expensive.

The amplifier is housed in a 5 × 7 × 2 inch aluminum box. It consists of two stacked circuit boards—an RF amplifier and a low-pass filter. The completed amplifier is shown in Figure 1. Figure 2 is the schematic diagram of the RF assembly together with the amplifier and low-pass filter parts list and Figure 3 is the low-pass filter schematic. The amplifier can be driven by 2 to 5 W at the RF input; an input attenuator consisting of R1, R2 and R3 must be selected (as noted in Figure 2) to ensure the proper drive level (2 W) for the push-pull (2SC2312C device types) class AB amplifier stage. T1 is wound on a small binocular core with a 4:1 ratio as detailed on the schematic. The low impedance secondary is a single turn center tapped; it carries the bias voltage to the output transistors. The bias voltage is derived from the LM317 regulator, which is operated as a switched current source. The LM317 is switched on when the internal PTT line is activated. Bias voltage is developed across the FES8J diode, which is in intimate thermal contact with the output transistors. The output transformer (T2) is also wound with a 4:1 ratio; it has a single-turn sense winding. The single-turn feedback winding provides some degree of negative feedback to flatten the gain and stabilize the input impedance over the HF frequency range. The RC network in the input base circuit establishes the overall gain. The 6.8 W series resistors determine the gain below 14 MHz, while the 4700 pF capacitors are effective above 14 MHz.

Figure 4 shows the output power versus frequency characteristics of the amplifier, including the second and third harmonic response. All measurements were made with an IFR1600S Communications Service Monitor. In all cases, harmonics were greater than 40 dB down referenced to the fundamental frequency (–40 dBc), and the amplifier meets current FCC requirements for spectral purity.[2] Note that the gain starts to decrease above 21 MHz, rolling off from a nominal 30 W to 20 W at 29 MHz. On 10 meters this still represents 10 dB of gain, a worthwhile improvement. The dc voltage to the output transistors is decoupled by the pi network at the center-tapped primary of the output transformer, T2. This winding carries substantial current and should be wound with #18 Teflon covered wire. Any IR drop in the dc circuit will severely degrade the amplifier performance. TR switching is provided by relay K1. The PTT line can be operated manually (pulled to ground) or RF activated switching circuitry can be installed to eliminate the need for external keying controls. The value of C12 (3.3 µF) determines the SSB time constant for the RF activated switch.

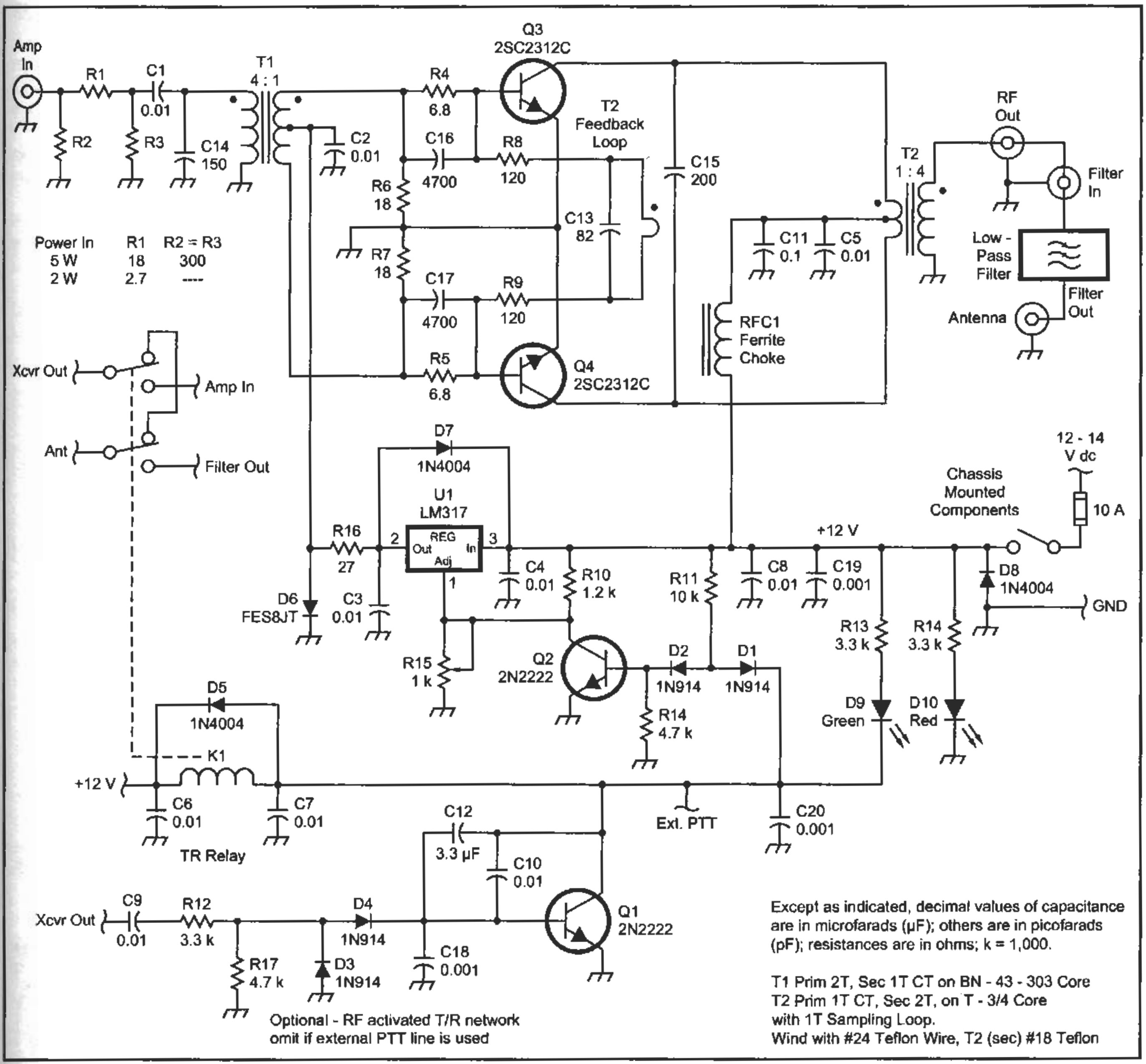
Amp In
R1
C1 0.01
T1 4 : 1
R2
R3
C14 150
Q3 2SC2312C
R4 6.8
C2 0.01
C16 4700
R8 120
T2 Feedback Loop
R6 18
R7 18
C13 82
C15 200
C17 4700
R9 120
R5 6.8
Q4 2SC2312C
RF Out
T2 1 : 4
Filter In
Low - Pass Filter
Filter Out
Antenna
C11 0.1
C5 0.01
RFC1 Ferrite Choke
Power In R1 R2 = R3
5 W 18 300
2 W 2.7 ----
Xcvr Out
Amp In
Ant
Filter Out
D7 1N4004
U1 LM317
REG Out In Adj
R16 27
C3 0.01
D6 FES8JT
C4 0.01
R10 1.2 k
R11 10 k
C8 0.01
C19 0.001
+12 V
12 - 14 V dc
Chassis Mounted Components
10 A
D8 1N4004
GND
R13 3.3 k
R14 3.3 k
D9 Green
D10 Red
R15 1 k
Q2 2N2222
D2 1N914
D1 1N914
R14 4.7 k
D5 1N4004
K1
+12 V
C6 0.01
C7 0.01
TR Relay
C12 3.3 µF
C10 0.01
Ext. PTT
C20 0.001
Q1 2N2222
Xcvr Out
C9 0.01
R12 3.3 k
D4 1N914
R17 4.7 k
D3 1N914
C18 0.001
Optional - RF activated T/R network omit if external PTT line is used

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000.

T1 Prim 2T, Sec 1T CT on BN - 43 - 303 Core
T2 Prim 1T CT, Sec 2T, on T - 3/4 Core
with 1T Sampling Loop.
Wind with #24 Teflon Wire, T2 (sec) #18 Teflon

Figure 2—The RF amplifier schematic and parts list. (M) denotes Mouser Electronics, 1000 N Main St, Mansfield, TX 76063; tel 800-346-6873; **www.mouser.com**. (RF) denotes RF Parts Co, 435 S Pacific St, San Marcos, CA 92069; tel 800-737-2787; **www.rfparts.com**. (A) denotes Amidon, Inc, 240 Briggs Ave, Costa Mesa, CA 92626; tel 800-898-1883; **www.amidon-inductive.com**. (F) denotes FAR Circuits, 18N640 Field Ct, Dundee, IL 60118; tel 847-836-9148; **www.farcircuits.net**.

RF Amplifier Board

C1-10—0.01 µF capacitor, (M) 140-100Z5-103Z.
C11—0.1 µF capacitor, (M) 80-CK06BX104K.
C12—3.3 µF capacitor, (M) 80-C340C335M5U.
C13—82 pF capacitor, (M) 5982-15-500V82.
C14—150 pF capacitor, (M) 5982-15-500V150.
C15—200 pF capacitor, (M) 5982-15-500V200.
C16, 17—4700 pF capacitor, (M) 140-50P5-472K-TB.
C18-20—0.001 µF capacitor, (M) 140-100Z5-102Z.
D1-D4—1N914 diode, (M) 610-1N914.
D5, D7, D8—1N4004 diode, (M) 583-1N4004.
D6—FES8JT diode, (M) 625-FES8JT.
D9—LED, green, with mount, (M) 512-HLMP4719.
D10—LED, red, with mount, (M) 512-HLMP4700.
K1—Relay, 12 V dc coil, DPDT, (M) 551-MR-12USR.
Q1, Q2—Transistor, switching, 2N2222, (M) 511-2N2222A.
Q3, Q4—Transistor, RF, 2SC2312C, (RF) 2SC2312C.
R1—300 Ω, 1 W, (M) 281-300.
R2, R3—18 Ω, 1 W, (M) 281-18.
R4, R5—6.8 Ω, ¼ W, (M) 30BJ250-6.8.
R6, R7—18 Ω, ¼ W, (M) 30BJ250-18.
R8, R9—120 Ω, ¼ W, (M) 30BY250-120.
R10—1.2 kΩ, ¼ W, (M) 30BJ250-1.2K.
R11—10 kΩ, ¼ W, (M) 30BJ250-10K.
R12-14—3.3 kΩ, ¼ W, (M) 30BJ250-3.3K.
R15—1 kΩ, potentiometer, (M) 531-PTC10H-1K.
R16—27 Ω, 1 W, (M) 281-27.
R17—4.7 kΩ, ¼ W, (M) 30BJ250-4.7K.
RFC1—RF choke, (RF) VK-200-3R.
T1—Transformer core, (A) BN-43-303.
T2—Transformer core, (RF) T-¾ core.
U1—IC, LM317T, (M) 512-LM317T.

Misc

2—TO-220 mounting kit, (M) 534-4724.
2—TO-220 thermal insulator pad, (M) 526-NTETP0006.
PC board, FARA RF amplifier, (F).

Low-Pass Filter Board

(see Figure 3 for component delineation)
2—100 pF capacitor, (M) 5982-15-500V100.
3—180 pF capacitor, (M) 5982-15-500V180.
3—330 pF capacitor, (M) 5982-15-500V330.
2—430 pF capacitor, (M) 5982-15-500V430.
1—560 pF capacitor, (M) 5982-15-500V560.
3—820 pF capacitor, (M) 5982-15-500V820.
3—1500 pF capacitor, (M) 5982-19-500V1500.
1—2700 pF capacitor, (M) 5982-19-500V2700.
12—0.01 µF capacitor, (M) 140-100Z5-102Z.

[continued on next page]

[continuation of Figure 2]

12—Relay, 12 V dc coil, DPDT, (M) 655-T7NS5D1-12.
1—Switch, 1 pole, 6 pos, (M) 10-YXX026.
HF Filter Kit, (A) HFFLT. Contains the cores and wire necessary to build the low-pass filters.
PC board, FARA LP Filter, (F).

Chassis Parts

Chassis box, Bud, (M) 563-AC-402.
Chassis cover, Bud, (M) 563-BPA-1589.
Heat sink, (M) 532-244609B02.
4—4-40 hex-type standoff, (M) 534-2201.
9—4-40 × ¼ pan head screw.
4—4-40 × ½ pan head screw.
6—4-40 flat washer.
9—4-40 lock washer.
2—4-40 nut.
2—8-32 × ¼ pan head screw.
2—8-32 lock washer.
2—BNC socket, chassis mount, (M) 161-9323.
RCA-type socket, female, chassis mount, (M) 161-1005.
Switch, toggle, SPDT, (M) 10TC320.
Jones-type socket, chassis mount, male, 2 pin, polarized, (M) 538-13023.

The low-pass filter assembly utilizes relays to select the proper filter network for the various frequency ranges. Relays were chosen to simplify the RF switching and to minimize cost. The six filters cover the nine amateur bands from 1.8 to 30 MHz; the frequency ranges and circuit constants are as noted on the LPF schematic diagram. The inductor cores and wire to wind the coils are available as a kit of parts from Amidon.[3] The L/C constants are the same as those recommended by WA2EBY for the MOSFET RF amplifier in *The ARRL Handbook*.[4] The filters are not used when the amplifier is off or when the PTT line is not activated—this permits multi-band listening and limited VHF use when a wide frequency transceiver is in use (like the Yaesu FT-817). There is no provision for ALC feedback, so caution must be exercised so as not to overdrive the amplifier.

Construction Hints

Although no step-by-step instructions are provided, a few hints will ease the assembly process. The circuit boards pictured are the prototype assemblies; they are not solder-plated. However, the available circuit boards (from FAR Circuits) are plated but do not have plated through-holes, so through-holes must be pinned and soldered.[5] Detailed drawings of the PC boards, the parts layout, coil-winding data and chassis templates can be found at **www.arrl.org/files/qst-binaries/faraamp.zip**. Saul, K1BI, the FARA Webmaster has also set up a site for the project. It can be found at **www.falara.org/tektalk/tektalkfs.html**. The circuit boards as they appear before wiring can be seen in Figure 5.

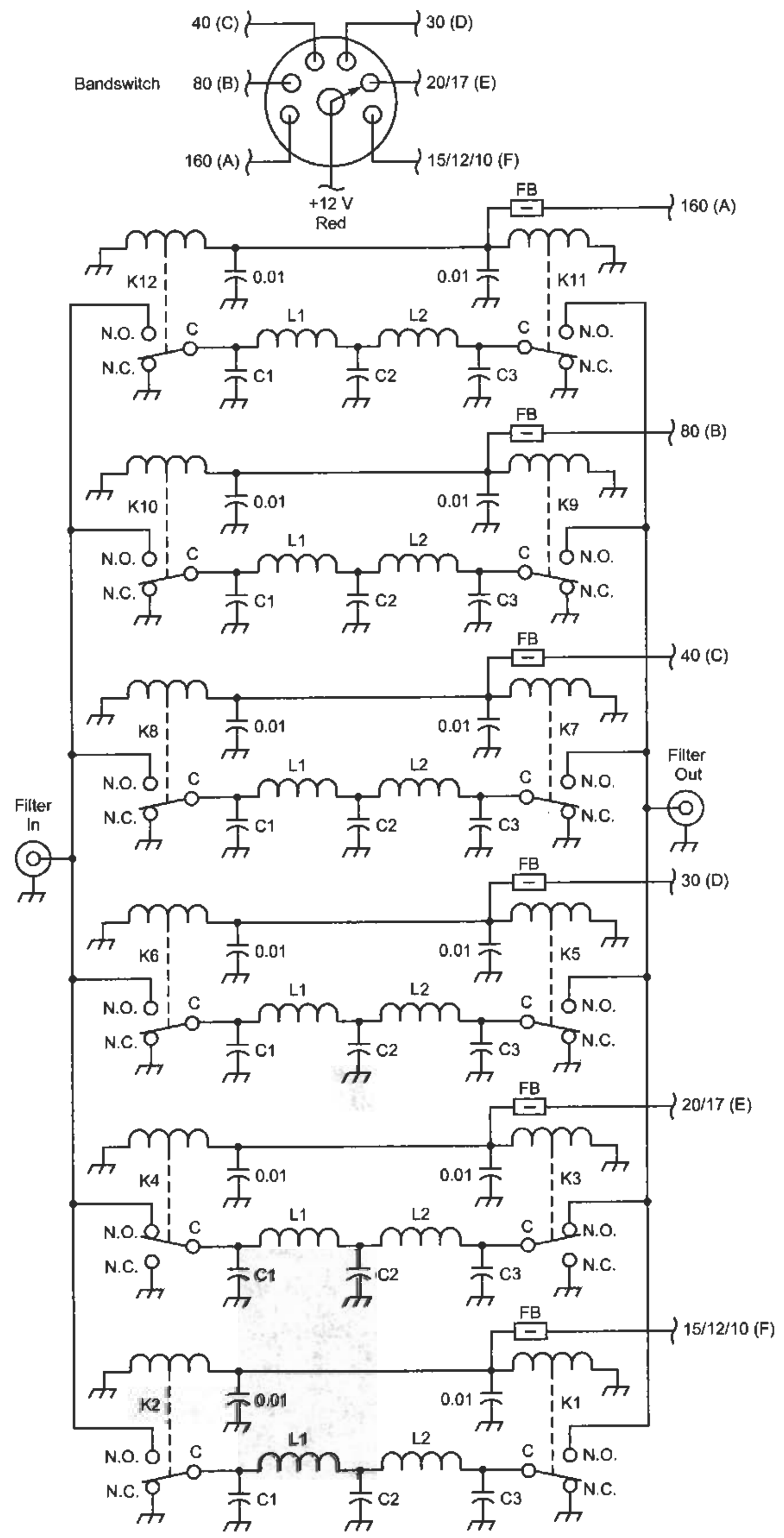

Cores T - 50 - X Amidon
L in number of turns (#22)
C in pF
Band (E) Shown Selected.
K3, K4 Energized.

Band	L1/2	C1/3	C2	Core
160 (A)	30	1500	2700	2
80 (B)	22	820	1500	2
40 (C)	16	430	820	2
30 (D)	14	330	560	2
20 - 17 (E)	11	180	330	6
15 - 10 (F)	8	100	180	6

Figure 3—The low pass filter (LPF) schematic.

Circuit Board Preparation

- Given the large ground plane area, the boards must be clean or you will experience difficulty when soldering to the foil. Plated boards are best; they are easier to solder. Some solder flux may improve the solderability of the board, but be sure to use *only* rosin core solder and a *non-corrosive* flux.
- Carefully inspect all soldered connections for cold solder joints. Good soldering technique is crucial to the performance of the amplifier.
- Periodically, the flux should be removed from the board during the construction phase with a suitable chemical cleaner.
- There are a number of holes to be drilled and pinned on each board; these are noted on the parts placements diagrams. Wires should be inserted through the board, bent into the shape of a "Z," formed flat against the board, soldered and cut.
- The four corner holes on each board must be sized as a clearance hole for the 4-40 mounting hardware.
- Components should be mounted flush to the board—fixed capacitors should be mounted as close to the foil as possible.
- The large rectangular blocks on the RF board must be trimmed in order to mount the RF output transistors and the bias diode.

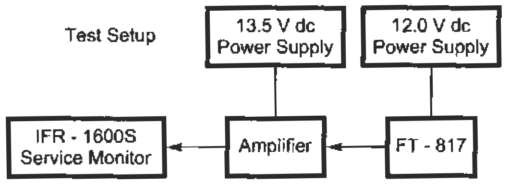

FT - 817, Power Output = 4.6 Watts ± .2 W (with −3 dB Pad)

FARA Amplifier Output Characteristics

Frequency (MHz)	RF Output (W)	2nd Harmonic (dB)	3rd Harmonic (dB)
1.8	23	-44	-43
3.5	33	-55	-63
7.2	33	-60	-50
10.1	28	-62	-45
14.2	31	-60	-50
18.1	31	-60	-52
21.2	30	-50	-47
24.9	25	-60	-53
28.4	21	-60	-56

Figure 4—Amplifier output data, including second and third harmonic response.

RF Amplifier Circuit Board

A view of the completed RF board can be seen in Figure 6. The following suggestions pertain to the amplifier board.

- Wind the secondary winding on T2 and mount it to the circuit board first.
- The emitters of the 2SC2312s (Q3, Q4) are intended to be grounded through a hole on the pad. The following modification is advised—bend a thin brass or copper strap into a "U" shape and solder it on both sides of the board. This lowers the impedance to ground.
- Next, mount the smaller fixed components, the resistors and capacitors.
- The semiconductors and the relay mount last.
- Do not mount D6, Q3 and Q4 until the assembly is fixed in the chassis and positioned relative to the heat sink.

Low-Pass Filter Circuit Board

This is a double-sided board; the reverse side is a ground plane with clearance etches for the various components. The LPF board can be seen in Figure 7. It may be necessary to form the capacitor leads slightly to conform to the hole spacing.

- Mount the filter components first, followed by the bypass capacitors and the jumper wires. Refer to the schematic for component values by frequency range.
- The relays mount last. Do not overheat their mounting pins when soldering.
- Do not mount the LPF assembly until the initial tune-up is completed. It is recommended that you pre-wire the band switch. It mounts between the two circuit boards when they are installed in the chas-

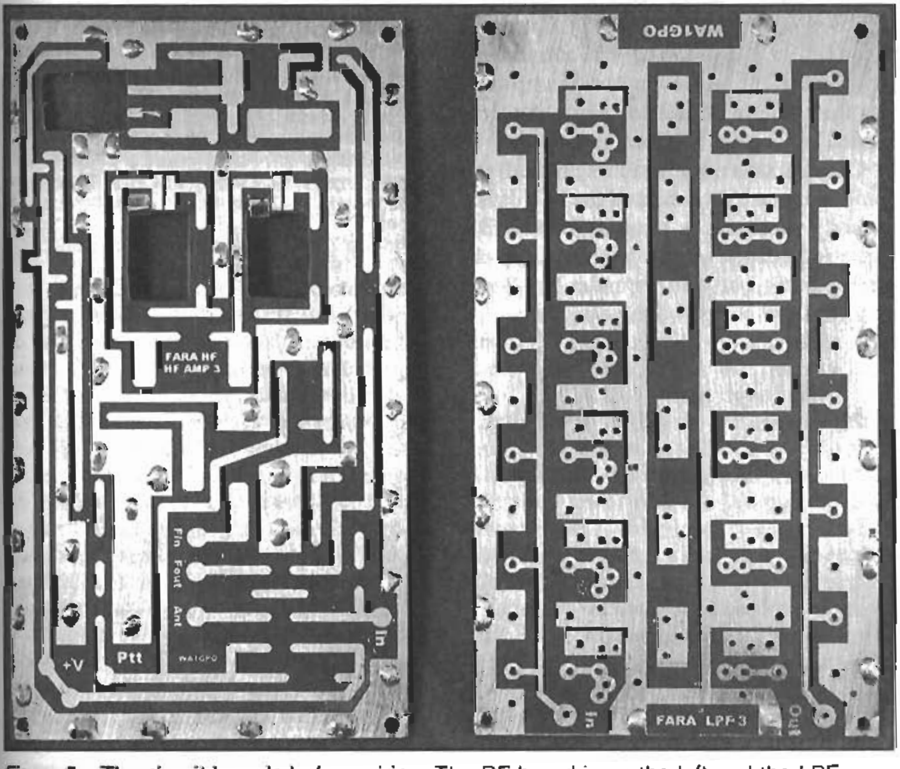

Figure 5—The circuit boards before wiring. The RF board is on the left and the LPF board is to the right. Note the pinned and soldered holes.

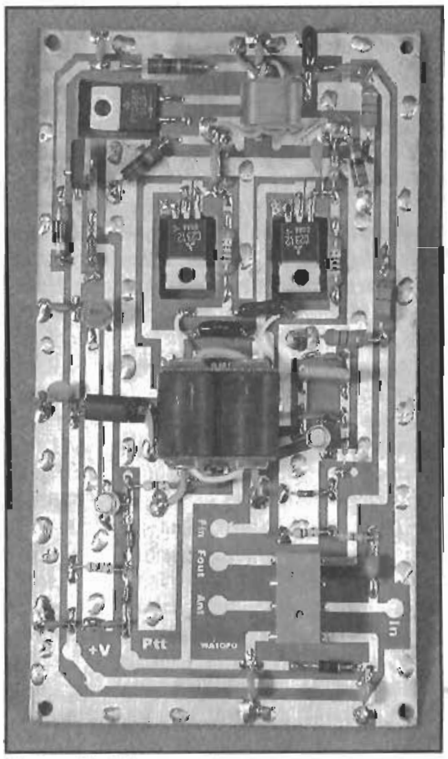

Figure 6—The completed RF assembly board.

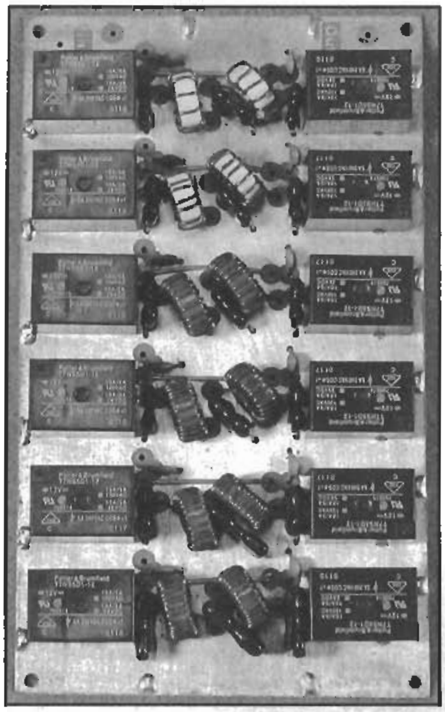

Figure 7—The completed LPF board. Note the relays that are used as filter switches–they are selected by the band switch.

sis and it is difficult to get to.

Chassis, Panel and Heat Sink

Full size templates for the chassis and panel can be found on the Web site. Trim the templates to size and fold and tape them to the aluminum box and heat sink. It is important to center punch the holes and drill a small pilot hole at each location. Enlarge the holes to size according to the dimensions. Letter the panel with dry transfer lettering available at office supply stores. Spray on several light coats of clear lacquer to protect the lettering. Mount the heat sink and the panel components, but do not mount the bandswitch at this time. Mount the RF amplifier board using a couple of 4-40 flat washers as spacers between the chassis and the circuit board at each corner. Install the 4-40 standoffs to hold the assembly in place. Mount D6, Q3 and Q4, using the TO-220 thermal pads and hardware to isolate the transistor mounting tab from the chassis. It is not necessary to isolate D6. Use RG-174 miniature coaxial cable for the internal RF connections.

Tune-up and Testing

As this is a broadband design there is no tune-up; only the bias adjustment needs to be set. A 12 to 14 V dc current limited variable supply is recommended for initial adjustment and testing. Use a dummy load at the amplifier output. Drive levels refer to the attenuator output, if used.

- Connect a temporary jumper between the F_{in} and F_{out} pads on the RF board.
- Preset R15 (1 kΩ potentiometer) so that the wiper is at ground.
- Apply 12 V dc and ground the PTT line (the relay should pull in).
- Briefly drive the input with 1 W at 14 MHz and note the power output.
- Again apply 1 W and increase bias (R15) until the output increases about 15%.
- Increase drive to 2 W and note the power output (about 25 W).
- Increase voltage to 14.7 V dc; note the output with 2 W of drive (about 35 W).

Remove power; remove the temporary jumper; mount the band switch; connect and install the LPF board. Verify that the power output over the range of 1.8 to 29 MHz is as expected. A view inside the bottom of the amplifier, with the LPF board visible, is shown in Figure 8.

Final Comments

The FCC has placed strict limitations on power amplifiers used below 30 MHz. Please review the appropriate FCC regulations before constructing this amplifier.[6] Kits offered for sale, even partial amplifier kits that require additional parts, are prohibited by current FCC regulations.[7]

Construction time, assuming that all the components are on hand and the boards are properly prepared, is about 4 hours. SSB operation results in the heat sink barely getting warm, and cooling under key-down conditions is more than adequate.

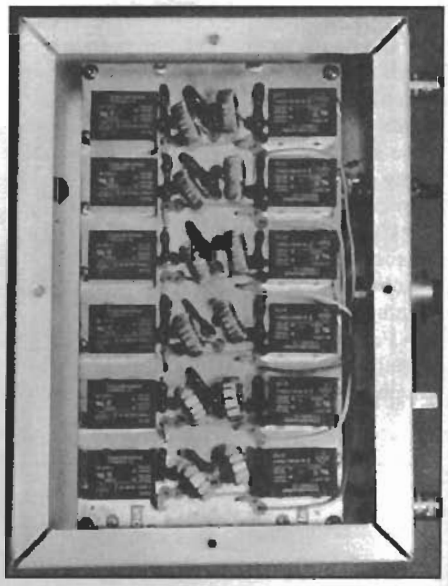

Figure 8—A bottom view inside the completed amplifier. The LPF board is mounted below the RF assembly.

A project of this scope is more fun when others participate. Harry, W2RKB, provided the necessary prodding to get it started. He also fabricated the circuit boards and the filter assemblies for the prototype amplifiers. Dave Hosom assisted with the photographs. A generous thank you is extended to them both. Give the FARA amplifier a try… it's a practical and rewarding project!

Notes

[1] Federal Communications Commission, Sec 97.315.

[2] See Note 1.

[3] Amidon Associates, Inc, 240 Briggs Ave, Costa Mesa, CA 92626; 800-898-1883; **www.amidon-inductive.com**.

[4] *The 2003 ARRL Handbook*, pp 17.91-17.97.

[5] FAR Circuits, 18N640 Field Court, Dundee, IL 60118; tel 847-836-9148; **www.farcircuits.net**.

[6] Federal Communications Commission, Sec 97.3 (19).

[7] See Note 6.

By Pat Bunn, N4LTA

From *QST*, October 1990

A Compact "Brick" for 6 Meters

Does your six-meter signal need some extra kick? This inexpensive amplifier bridges the gap from 10 to 100 watts.

Most transceivers available for the 6-meter band provide 10 W output. This power level provides satisfactory communications during good sporadic-E openings, but is not adequate for making long-haul contacts on meteor scatter or during F-layer openings. For the average station, an output level of 100 W seems to be a reasonable compromise between a 10-W exciter and a kilowatt amplifier. Unfortunately, 50-MHz amplifiers in this power class are not readily available.

I began operating on the 6-meter band a few years ago, near the lowest point in sunspot cycle 21. Sporadic E kept my interest high, and soon it became evident that I needed an amplifier. Being a builder at heart, I decided to "roll my own" and began to search the Amateur Radio literature for design examples. I found very few solid-state amplifiers, and only one in the 100-W class.[1]

I built the amplifier as specified, and I got good results. My 6-meter effectiveness made a quantum leap! One highlight was working Europe with a 12-foot-boom Yagi. The amplifier performed so well that several of my friends asked me to build them duplicates of it. With each revision, I tried to add design innovations, including on-board TR relays, a carrier-operated-relay (COR) circuit and transistor biasing, as well as low-pass filtering and more compact packaging. The product of these efforts is described in this article. At least 25 of these units have been constructed and are now in use. Several of these amplifiers have been used on DXpeditions, in beacon service, and as drivers for kilowatt-class amplifiers, and results have been excellent.

If you are looking for a low-cost 100-W amplifier, this proven design would make a nice addition to your 6-meter station. This circuit can also be adapted for use with an MRF140 FET, which produces over 160 W output with less than 10 W drive. See the sidebar for more details.

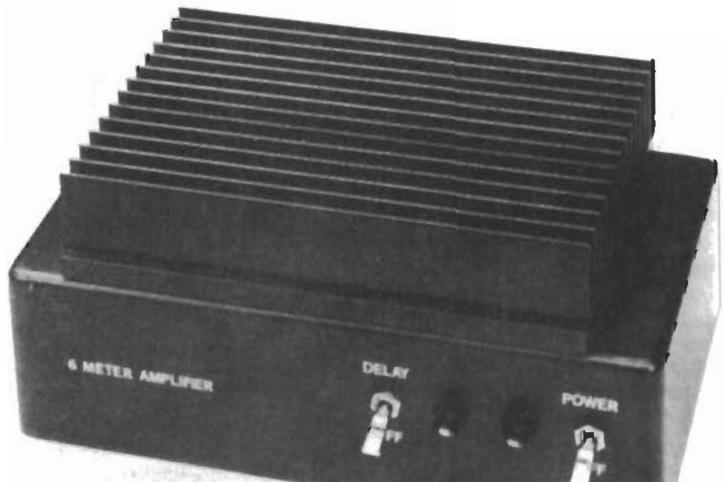

Circuit Highlights

The amplifier circuit, shown in Fig 1, is a single-stage, 10-dB-gain, class-B design using the popular Motorola MRF492 transistor. The RF circuitry is based on a test circuit in Motorola's RF Device Data book.[2] Input and output impedance matching is done with low-Q lumped circuits rather than with broadband transformers. The amplifier's bandwidth is about 1.5 MHz, which covers the DX portion of the 6-meter band with plenty to spare. And, the output network has a secondary dip at about 53 MHz, allowing CW/SSB work and FM operation without the need to retune the amplifier. The narrowband nature of the matching networks helps reduce harmonics. A half-wave low-pass filter in the output circuit further reduces harmonic content to meet current FCC specifications (see Fig 2).

The COR circuit is based on a high-gain operational amplifier that provides the sensitivity required for SSB operation. This circuit is much more effective than the Darlington-transistor COR circuits found in most commercial amplifiers. It works well with low-power SSB rigs that do not have easily accessible PTT lines, although a separate PTT input is also included.

The amplifier PC board[3] is mounted in a die-cast aluminum box (Hammond 1590R, Bud CU-247 or Jameco® part no. B5006) with an extruded-aluminum heat sink bolted to the bottom. (I used a 6-inch length of Wakefield no. 1527 heat sink. Any aluminum heat sink with similar dimensions should work fine.) The completed amplifier is operated inverted, with the removable lid serving as the bottom. The enclosure is painted flat black and labeled with white rub-on lettering. Stick-on rubber feet are attached at the corners of the bottom cover. The finished product

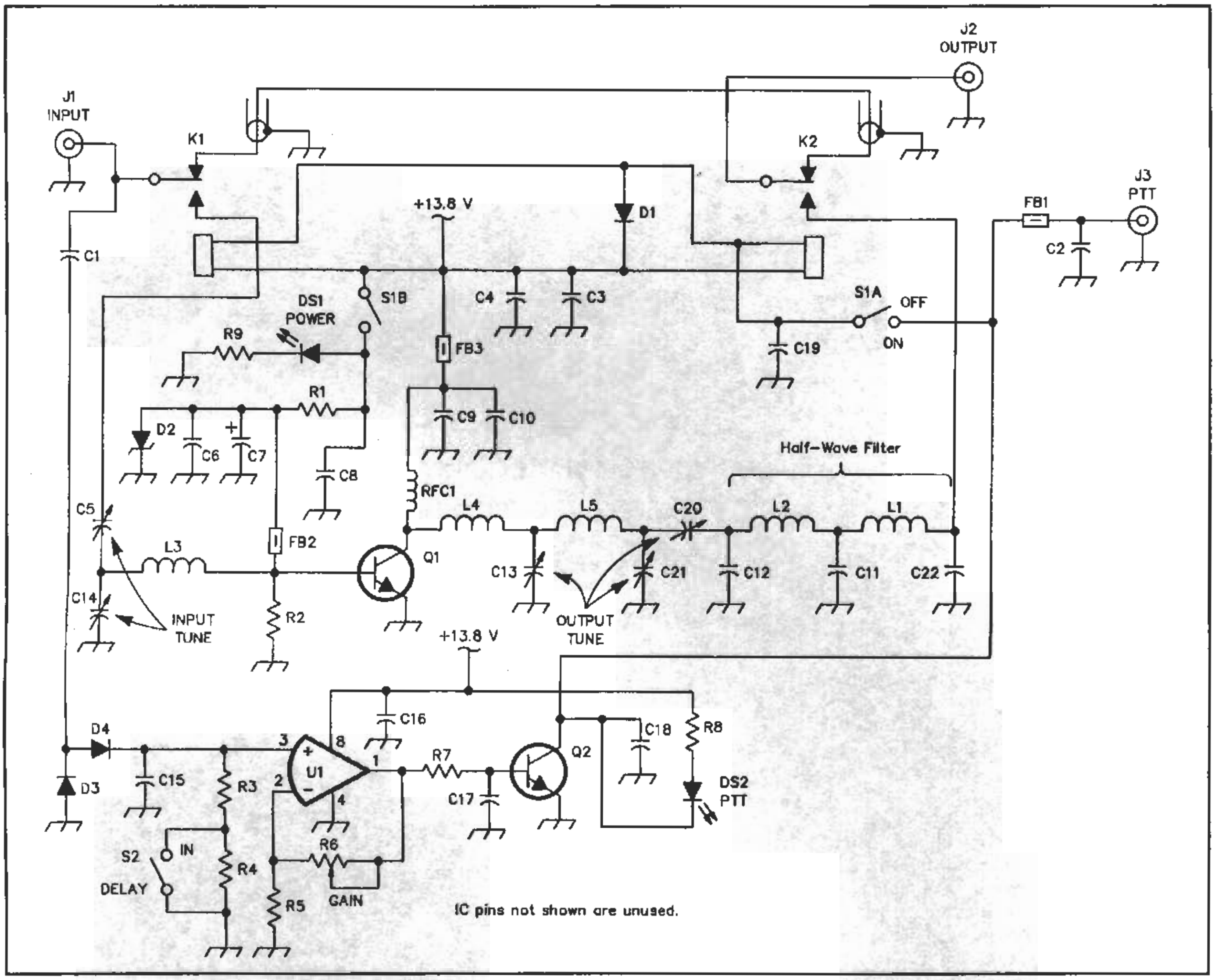
J1
INPUT
K1
J2
OUTPUT
K2
J3
PTT
FB1
C2
+13.8 V
D1
C1
DS1
POWER
S1B
C4
C3
S1A
OFF
ON
C19
R9
FB3
R1
C9
C10
D2
C6
C7
C8
Half-Wave Filter
RFC1
L4
L5
C20
L2
L1
C5
L3
FB2
Q1
C13
C21
C12
C11
C22
C14
INPUT
TUNE
R2
OUTPUT
TUNE
+13.8 V
C16
C18
R8
D4
3
8
U1
1
R7
Q2
D3
C15
R3
2
4
C17
DS2
PTT
R6
S2
IN
DELAY
R4
R5
GAIN
IC pins not shown are unused.

Fig 1—Schematic of the 100-W 6-meter amplifier. Liberal use of bypass capacitors and ferrite beads helps ensure stability and keeps RF out of the power supply.

C1—10 pF, ceramic disc.
C2—0.01 µF, polyester film.
C3, C10—0.001 µF, ceramic disc.
C4, C6, C8, C9, C15-C19—0.1 µF, polyester film.
C5—9- to 180-pF mica compression trimmer, Arco no. 463.
C7—1000 µF, 6.3 V, aluminum electrolytic.
C11—120 pF, 100 V, silver mica.
C12, C22—62 pF, 100 V, silver mica.
C13, C14, C20—50- to 380-pF mica compression trimmer, Arco no. 465.
C21—25- to 280-pF mica compression trimmer, Arco no. 464.
D1—1N4002.
D2—1N1200 stud-mount diode.
D3, D4—1N4148.
DS1—Green LED.
DS2—Red LED.
FB1, FB3—Ferrite bead, Amidon no. FB43-901 or similar.
FB2—VK-200 wide-band choke, 2½ turns no. 24 solid wire on Amidon no. FB43-5111 ferrite core.
J1, J2—Female RF connector (UHF, BNC, N, etc).
J3—Phono jack.
K1, K2—SPDT relay, 12-V dc coil, Omron no. G5L112P-PS-DC1 2. Available from Digi-Key.
L1, L2—4 turns no. 14 enam wire, 7/16 in. diam, 3/8 in. long.
L3—2 turns no. 14 enam wire, 3/8 in. diam, 1/2 in. long.
L4—no. 14 U-shaped wire loop, 3/8 in. diam, 9/16 in. finished length.
Q1—MRF492.
Q2—2N2222A.
R1—5-W bias resistor (see text).
R2—10 Ω, 1/2 W, carbon comp.
R3, R5—10 kΩ, 1/4 W, carbon comp.
R4—1 MΩ, 1/4 W, carbon comp.
R6—100-kΩ PC-board potentiometer.
R7-R9—1 kΩ, 1/4 W, carbon comp.
S1—DPDT miniature toggle.
S2—SPST miniature toggle.
U1—LM358.

Miscellaneous

Suitable die-cast-aluminum enclosure and heat sink.
Pair of binding posts or other suitable dc-supply connector.
Two 1- x 3-in. strips of double-sided foam tape.
Two LED holders.
Seven no. 4-40 × 1/4 in. machine screws.

looks professional, and also does a good job of cooling the power transistor and bias diode.

Construction

The amplifier is assembled on a double-sided printed circuit board. All RF components are surface-mounted on the component side of the PC board (see Fig 3). The COR components, including the relays, connect to traces on the other side of the board. The remaining non-component-side foil serves as a ground plane, and is electrically connected to the component-side ground plane via platedthrough holes. If you make your own PC board, you can replace these plated-through holes with no. 18 AWG wires soldered to both sides of the board, or, preferably, copper rivets such as those available from Frontier Microwave and Down East Microwave.[4] The relays also mount in plated-through holes. The PC-board patterns and part-placement overlay are available from the ARRL Technical Department Secretary.[5]

Drill the case for the switches, connectors and indicators you'll use. Begin PC-

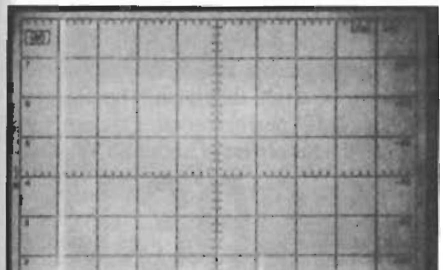

Fig 2—Worst-case spectral display of the 100-W 6-meter amplifier. Vertical divisions are each 10 dB; horizontal divisions are each 10 MHz. Output power was approximately 100 W at 50.2 MHz. All harmonics and spurious emissions are at least 60 dB below peak fundamental output when the amplifier is tuned as described in the text. The 100-W amplifier complies with current FCC specifications for spectral purity for equipment in this power-output class and frequency range. Better spectral purity can be obtained by using an external output filter (see the referent of note 1 for a suitable filter circuit).

board construction by mounting the parts that are soldered to traces on the other side of the board. This includes the COR parts and relays, the RG-174 jumper, and diode D1. Make sure that the diodes and Q2 are properly oriented.

The solder-side parts are surface-mounted. Keep lead lengths to a minimum (no more than 1/4 inch) on these parts. The bypass capacitors should have the shortest possible leads, and should be surface-mounted on the PC board.

Wind the coils using 1/4- and 3/8-inch drill bits as temporary forms. Cut the coil leads to 1/4 inch, bend them 90°, scrape the insulation off the ends and solder the coils to the board. Some of the coils may need to be compressed or spread during tune-up to get maximum output power, so allow for this when mounting them. Use Fig 3 as a guide. Mount the mica trimmers by bending the solder tabs parallel to the board and surface-mounting them. (Some trimmers have extra tabs that may short to the PC board; I cut these off with side cutters to eliminate this risk.) When surface-mounting the larger components, a high-wattage soldering gun is useful, but be careful: The traces may separate from the PC board if excessive heat is used.

When the PC board is complete, temporarily place the MRF492 in its mounting hole in the PC board, making sure to orient the transistor properly (the beveled lead is the collector). Place the PC board inside the box assembly. Carefully mark the transistor-mounting holes and the single PC-board mounting hole on the enclosure, through the PC board. Remove the PC board from the case, then drill and tap these holes for no. 4-40 hardware. Mark and drill the heat sink using the enclosure as a guide.

Drill and tap four more no. 4-40 holes, near the corners of the case, through the case and into the heat sink. Apply heat-sink compound between the die-cast box and the heat sink, then use no. 4-40 screws in these holes to secure the heat sink to the case.

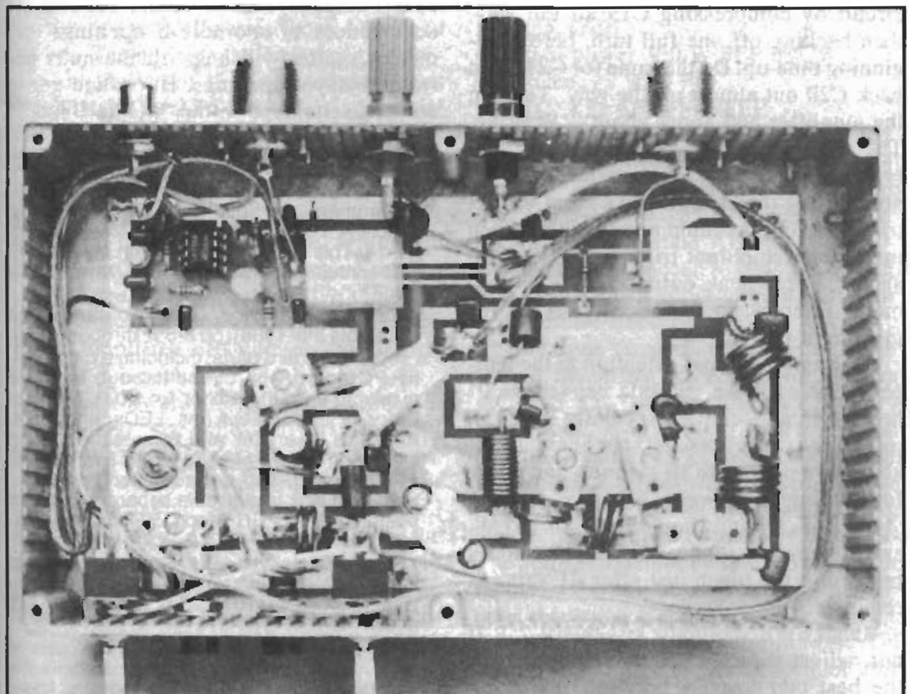

Fig 3—This inside view of the 100-W amplifier shows the circuit's simplicity. On the PC board, counterclockwise from upper left: the COR circuit; input-matching and bias networks, switches and indicators (lower left); MRF492 (lower center); collector choke and output network (lower right). The TA relays and connectors are at the upper center, and bias resistor R1 is just to the left of center.

Place the PC board into the case/heat-sink assembly again, and mark the location of the stud-mounted bias diode. Remove the board and drill and tap a 3/8-inch-deep no. 10-32 hole in the case/heat-sink assembly. The diode's threaded stud is more than 3/8 inch long, and should be carefully cut off (with a hacksaw) to 5/16 inch.

The rear of the PC board must be insulated from the box assembly, because many of the COR-circuit traces on the underside of the board are not at ground potential. I use 1/8-inch-thick double-sided foam tape to cover the solder-side traces and to support the board away from the case. Make sure the component leads are cut off closely at the rear of the PC board so that they don't punch through the foam tape and short to the box assembly. Mount the board inside the case, then screw the bias diode snugly into its mounting hole in the case/heat sink assembly.

Solder the MRF492 onto the board as the last step, *after* the transistor case has been mounted in the board and screwed into place. (This eliminates the possibility of cracking and ruining the $15 device!) Be sure to use a liberal amount of silicone heat-sink compound between the transistor and the box.

Mount the front-panel LEDs, switches and the RF connectors of your choice to the box. Any suitable RF connectors can be used. The connectors are wired to the PC board with no. 18 tinned bus-wire jumpers. Make these as short as possible. I used no. 24 insulated hookup wire to connect the switches and LEDs to the PC board. Finally, mount the dc-power binding posts to the rear panel. These are wired to the PC board with no. 18 bus wire run through FT-43-801 ferrite beads.

Initial Testing

After the switches, LEDs and connectors have been wired, carefully check all connections. Make certain that the jumper from the PC board to bias diode D2 is connected! If all looks good, connect the amplifier to a 13.8-V power supply capable of sourcing 1 A. (If you have one, a small, current-limited power supply is best for testing.) If your power supply does not have reliable current metering, a meter capable of measuring 0-500 mA should be placed in series with the positive supply lead.

With the amplifier turned off, turn on the power supply. The meter should indicate 0. Turn on the amplifier and note the total indicated current. It should be between 50 and 300 mA. If it is much greater than 300 mA, quickly turn off the power supply and look for problems.

If the current is within these limits, you can go about setting the operating bias. This involves changing the value of R1 until the collector current is between 75 and 175 mA. Note that this is not total circuit current, which includes the diode

Using a MOSFET to Get More Power from the Basic MRF492 Amplifier Circuit

The MRF492 amplifier circuit described in the main text is easily modified for use with an MRF140 VMOS power FET, which has about 15 dB gain and over 160 W output—substantially more than the MRF492. Even the device cases are the same, so the changes required to operate this circuit with an MRF140 are quite straightforward. Specific details of the modifications required, including a schematic diagram, are available from the author for an SASE.

The main differences between the devices are that the MRF140, a MOSFET, needs different biasing and a higher supply potential (28 V) than the MRF492, and is more expensive ($88 versus $15 for the MRF492, at this writing). But if you have a 28-V supply and need more than 100 W or have a 5-W exciter, the change is probably worthwhile. The MRF140 is easily biased with a three-terminal regulator (an LM317T works well), and requires very few component changes other than those in the bias circuit. An additional capacitor and inductor are also required at the MRF140's gate.

Because the modified amplifier requires a 24- to 28-V supply, different TR relays must be used, or, alternatively, 12-V-coil relays can be wired in series. (The COR circuit needs no modification because the LM358 works fine at 28 V.)

Good heat sinking is more critical to amplifier performance at the higher power level, so I used a 3- × 5- inch slab of ⅛-inch-thick copper between the device and the case as a heat spreader. A larger heat sink and/or forced-air cooling are also desirable.—*N4LTA*

current and POWER LED current. The best way to measure collector current is to lift FB3 and insert a milliammeter in series with the collector-supply lead at that location. An alternative method is to use the following equation to approximate collector current:

$$I_C = I_{TOTAL} - [(V_{R1} \div R1) + ([V_{CC} - 1] \div 1000)]$$

where

I_C = collector current in milliamperes
V_{R1} = measured voltage across R1
V_{CC} = measured supply voltage

This equation lets you calculate collector current by subtracting the current through R1 ($V_{R1} \div R_1$) and the current through the **POWER** LED ($[V_{CC} — 1] \div 1000$) from the total supply current.

Using either method, substitute different resistors at R1 until the collector current falls into the 75- to 175-mA range. In almost every case, a 5-W resistor of 82, 91, 100, 110, 120 or 130 Ω will bring the current into that region. (I have never had to substitute more than three resistors to arrive in this bias-current range.) These 5-W resistors can be purchased for less than $0.40 each, so, although a 200-Ω, 5-W wire-wound potentiometer could be used for R1, all six resistors can be purchased for far less. And, these days, high-wattage potentiometers are hard to find.

Tune-Up

When the bias current has been set, the amplifier can be initially tuned. For this procedure, the amplifier should be connected to a power supply capable of delivering at least 15 A at 13.8 V. A power meter capable of indicating at least 100 W at 50 MHz should be connected at the amplifier output, and the meter's output to a 50-Ω load. It is also desirable to measure input power from the exciter, as well as amplifier-input SWR, but these can be measured after initial tune-up if two meters aren't available.

Set the input-circuit trimmers (C5 and C14) near midrange. Because the output circuit can be tuned several ways for the proper impedance transformation, it's important to preliminarily adjust the output circuit by compressing C13 all the way, then backing off one full turn, before beginning tune-up. Do the same for C21, then back C20 out almost all the way. Turn on the amplifier and apply about 2 W drive. The COR should sense the input RF and switch immediately into transmit. Using a nonmetallic tool, tune the two input-circuit trimmers for maximum power output, then tune the three output trimmers (C13, C20 and C21) for peak output. These adjustments interact, so continue tweaking for maximum power.

During tuning, pay attention to the temperatures of these variable capacitors. If heating is noticeable (especially in C13), reset these capacitors and start again. Limit key-down periods to 10 seconds, and allow for cooling in between. Increase drive to about 9 W. Output should be 95 to 105 W, and the output-tuning capacitors should be cool to the touch just after RF drive is removed.

Input SWR should be less than 1.2:1. If not, adjust the input-circuit trimmers for the best compromise between low input SWR and maximum power output. The amplifier is now ready for operation. Some MRF492s seem to be "hotter" (exhibit more gain) than others; I have seen some devices produce 115 W easily and others strain to put out 95 W. On the air, the difference is so slight that it can't be detected. For best IMD performance, I recommend running the amplifier at about 85 W output on SSB.

FM Operation

This amplifier also works well on FM, but the power output should be limited to about 75 W because of FM's more demanding duty cycle. As mentioned earlier, the amplifier's bandwidth characteristic is double-humped; if it's tuned up at 50.1 MHz, the second hump falls in the 53-MHz area, allowing for SSB and FM operation without retuning.

If the amplifier is to be used for beacon, CW or FM operation exclusively, I suggest you configure it for class-C operation. Efficiency in class-C operation is in the 72% range, versus 60% for class B. To configure the amplifier for class-C use, connect the inductor using FB2 (the multi-hole ferrite bead) from Q1's base to ground, and eliminate all of the other bias components, including D2. If you like, you can add a panel switch to change between class-C and class-B operation.

The amplifier should never be operated class C on SSB. Doing so results in severe distortion.

Conclusion

I hope that you enjoy building this project, and will be motivated to try 50-MHz operation. As I write this, Japanese stations are being worked on the East Coast, and all continents are frequently heard. In the years to come, the regular occurrences of sporadic-E openings and meteor scatter will help fill the nulls between F-layer openings. Hurry and get in on the excitement on the "Magic Band"!

Notes

[1] T. Tammaru, "A Solid-State 6-Meter Linear Amplifier You Can Build," *QST*, May 1982, pp 11-14. This circuit was based on an MRF492 test circuit in Motorola's *RF Device Data* book.

[2] *Motorola RF Device Data*, Fourth edition (Motorola Inc: 1986), pp 3-552 through 3-555.

[3] A set of PC-board parts, including a high-quality PC board with plated-through holes, is available from the author for $100 (UPS delivery included). Switches, LEDs, enclosure, terminals, heat sink and connectors are not included. The PC board alone is available for $17 postpaid, and a black anodized-aluminum heat sink is available separately for $12 postpaid. (The ARRL and *QST* in no way warrant this offer.)

[4] Frontier Microwave, AD 1, Box 467, Ottsville, PA 18942, evening tel 215-795-2648; Down East Microwave, RR 1, Box 2310, Troy, ME 04987, tel 207-948-3741.

[5] Send a business-size SAE with return postage for 1 ounce with your request. Specify the information package for the 6-meter linear amplifier from October 1990 *QST*. This package includes full-size etching patterns for the front and rear sides of the amplifier PC board and a part-placement overlay.

By Richard Frey, K4XU

From *QEX*, May 1999

A 300-W MOSFET Linear Amplifier for 50 MHz

Everyone is looking to go QRO on 6 meters. Build this "almost a brick" to boost your signals from 15 to 300 W.

In an earlier article, I described a 50 MHz, 125 V, 250 W, class-C amplifier using ARF448AJB highvoltage MOSFET devices.[1] This paper describes an improved version of an amplifier that is capable of class-AB linear operation. The design changes required and the procedures involved are explained and demonstrated. Complete descriptions of the amplifier and its construction are presented, as well as the measured performance.

High-voltage, high-power MOSFETs have been shown to be very capable RF power amplifiers.[2] The metal-gate architecture of the ARF series from Advanced Power Technology has raised the frequency limits for this type of device to 100 MHz. The APT448A/B is typical of the series. It has a 68,000square-mil die with a breakdown voltage rating, BV_{dss} of 450 V. The device is provided in the inexpensive TO-247 plastic package, and is available in common-source symmetric pairs. Like all MOSFETs, the gate threshold voltage, V_{th}, has a negative temperature coefficient. This makes operation as a linear amplifier difficult or impossible without compensation.

When forward biased with a constant gate voltage, the quiescent drain current rises as the temperature of the die increases. Operating at the typical drain voltage for these parts, about one third of the rated BV_{dss}, the power dissipation caused by the increasing 'dq results in "hot spotting" and subsequent thermal runaway. This is an unstable system. The dissipation increases so rapidly that the outside surface of the case does not follow internal junction temperature well. As a result, a bias-compensation scheme that uses temperature sensing cannot keep up with the V_{th} shift, and the device is destroyed.

The power dissipation within the die is a direct function of the operating voltage. By lowering the operating voltage, the thermal loop gain can be reduced to a point where the gatethreshold shift can be compensated. Thermal stability can be achieved by sensing the case temperature. Linear operation thus becomes practical at 100 V and below. While this is less than 25% of the rated BV_{dss} and produces less gain, a very rugged and useful linear amplifier is the result.

Amplifier Description

The following were the design goals for the amplifier:

- Frequency range 50 to 51 MHz
- Input SWR < 1.5:1
- Gain> 13dB
- Output power 300 W PEP or CW
- Efficiency> 50%
- IMD3 > 25 dB below PEP

A push-pull topology was chosen for best output power and minimum harmonic content. The previously reported class-C design used V_d = 125 V. Since this is too high for reliable class-AB operation, 80 V was eventually chosen for this design. This is a compromise between gain, efficiency and thermal stability.

Since the gate input impedance is very low, it magnifies the effects of any stray inductance in the gate matching circuit. In a push-pull design, it is critical to maintain absolute symmetry between the two sides. This fact was demonstrated during the initial design work. One preliminary design had a slight asymmetry in the PCB artwork. The amplifier exhibited low efficiency, hot ferrite in the output transformer balun and poor distortion characteristics with asymmetrical IMD product amplitudes. This clearly demonstrates the benefit of symmetrically packaged devices.

A multiple-aperture ferrite bead was chosen for the input transformer. Brass tubing was used for the secondary, and the primary was wound inside the brass tubes to provide a very broadband balanced transformer design with minimum leakage reactance. Several cores and construction methods were evaluated, and this well-tried design proved best.

The typical HF push-pull amplifier employs a bifilar choke to decouple the drain-voltage feed. As frequency and power increase, this feed method becomes less practical because the need to reduce the number of windings to offset the stray capacity and the need to prevent core saturation conflict. In this design, a powdered-iron toroid was chosen for the feed choke core. This proved far superior to any of the ferrite cores evaluated. It is inexpensive and easy to reproduce.

Design

Three main areas must be addressed. The input matching must provide a balanced feed to a pair of low-impedance gates. The output must be matched to a suitable load impedance, and then transformed to a 50-Ω unbalanced load. The bias must be thermally compensated to track the negative temperature coefficient of the gate bias threshold.

The input is reasonably straightforward. Each gate input is 0.2 +*j*0.5, as estimated from the data sheet. The push-pull topology puts these two impedances in series, which makes the matching less difficult. A Smith Chart[3] program makes the actual design easy. The program used here is *WinSmith*.[4] The two gate impedances are added in series, and a network synthesized to transform the resulting impedance up to 50 Ω. The Smith Chart program only works with single-ended circuits, so the center tap was added later. See Fig 1.

The input-transformer design was chosen for its simplicity and relative ease of construction. Of several attempts using different material permeabilities, multiple beads and different conductor types, this proved to be the best and most consistent performer. The core is a Fair-Rite[5] "multi-

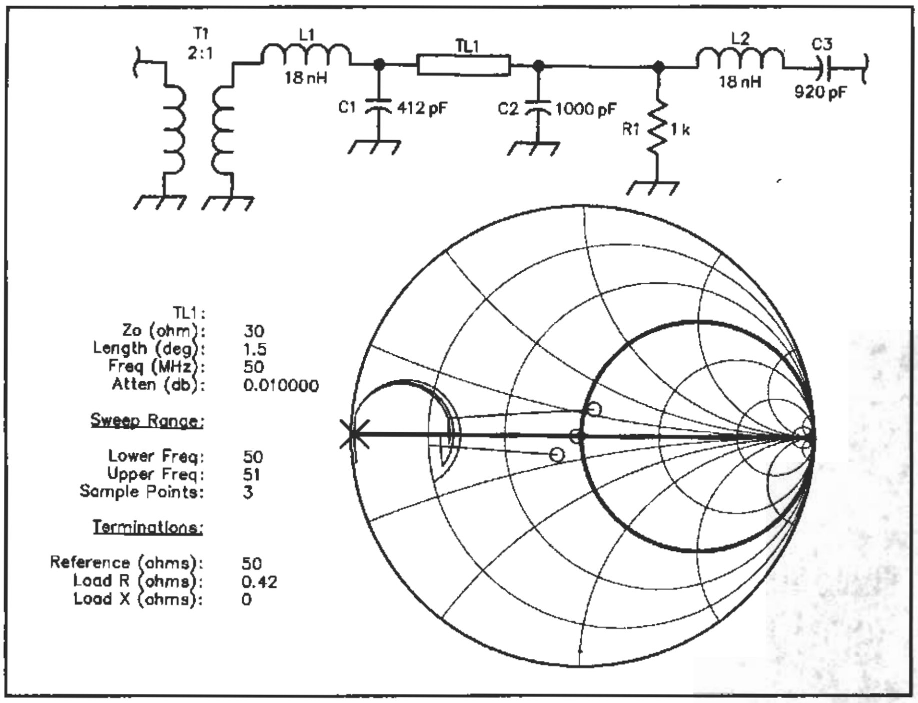
T1
2:1
L1
18 nH
TL1
C1 412 pF
C2 1000 pF
R1 1 k
L2
18 nH
C3
920 pF
TL1:
Zo (ohm): 30
Length (deg): 1.5
Freq (MHz): 50
Atten (db): 0.010000
Sweep Range:
Lower Freq: 50
Upper Freq: 51
Sample Points: 3
Terminations:
Reference (ohms): 50
Load R (ohms): 0.42
Load X (ohms): 0

Fig 1—Input matching network and calculation.

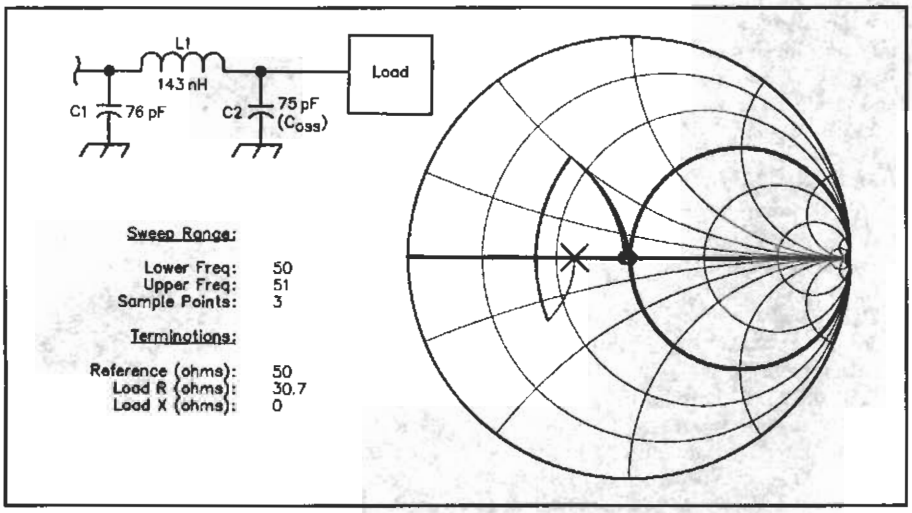
L1
143 nH
C1 76 pF
C2 75 pF (Coss)
Load
Sweep Range:
Lower Freq: 50
Upper Freq: 51
Sample Points: 3
Terminations:
Reference (ohms): 50
Load R (ohms): 30.7
Load X (ohms): 0

Fig 2—Output matching network and calculation.

aperture" core, part number 2843010402. The type-43 material has a μ_i of 850. At 50 MHz, type-61 material (μ_i of 125) would also be satisfactory. This transformer is essential to provide balanced drive to the gates of the MOSFETs. The secondary winding is 3/16-inch brass tubing. Copper shim stock forms the connections to the brass tubing at each end of the transformer secondary. The two-turn primary is wound inside the tubing. This construction provides a very reproducible transformer with minimum leakage reactance and a very broad frequency response. It would be a suitable input transformer for a broadband amplifier covering 1 to 100 MHz.

The leakage reactance of the input transformer—referred to the secondary—is about 18 nH and is represented as Li on the simplified input schematic in Fig 1. The gate load is represented by the "Load R," L2 and C3. Using all three parts of the gate impedance allows proper evaluation of the network bandwidth. A pi network consisting of C1, TL1 and C2 is used to step-up the gate load to the 12.5 Ω needed by the transformer and compensates for T1's leakage reactance. Notice that the net stray inductance of the gate is almost enough to effect a match with a single shunt capacitor. This has actually been done, but it was not easy to fit all the parts in the available space; so it was judged unacceptable here. To transform the network into the required balanced configuration, the series TL1 is split into two equal parts; the shunt capacitors remain the same, and a neutral center tap is provided at the transformer secondary.

Because of the high currents circulating in the input network, it is imperative that C2 be a larger-sized, class-1 dielectric (C0G or NP0) capacitor. It must be a leadless, surfacemount chip type, or the value will need to be adjusted. The input tuning capacitor, a 900 pF mica compression trimmer, is mounted directly to the end of the input transformer.

The output network is also straightforward. The proper load impedance for class

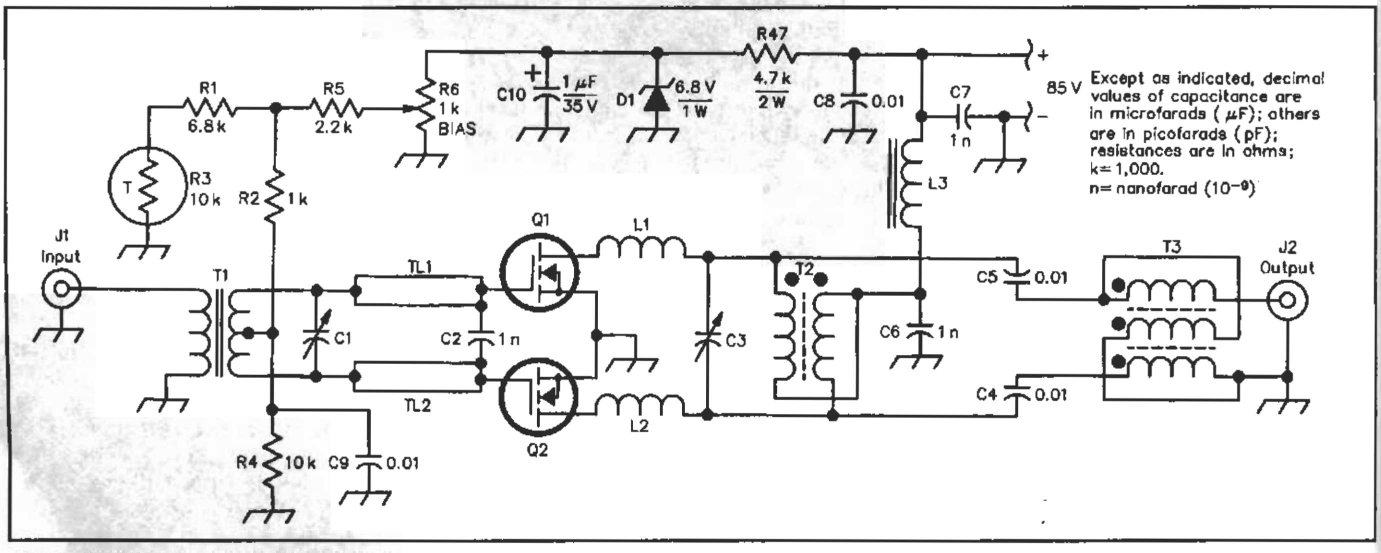
R47
4.7 k
2 W
R1
6.8 k
R5
2.2 k
R6
1 k
BIAS
C10
1 µF
35 V
D1
6.8 V
1 W
C8
0.01
C7
1 n
+
85 V
−
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k= 1,000.
n= nanofarad (10−9)
T
R3
10 k
R2
1 k
L3
J1
Input
T1
TL1
C1
C2
1 n
TL2
Q1
Q2
L1
L2
C3
T2
C5
0.01
C6
1 n
C4
0.01
T3
J2
Output
R4
10 k
C9
0.01

Fig 3—50 MHz amplifier schematic diagram.

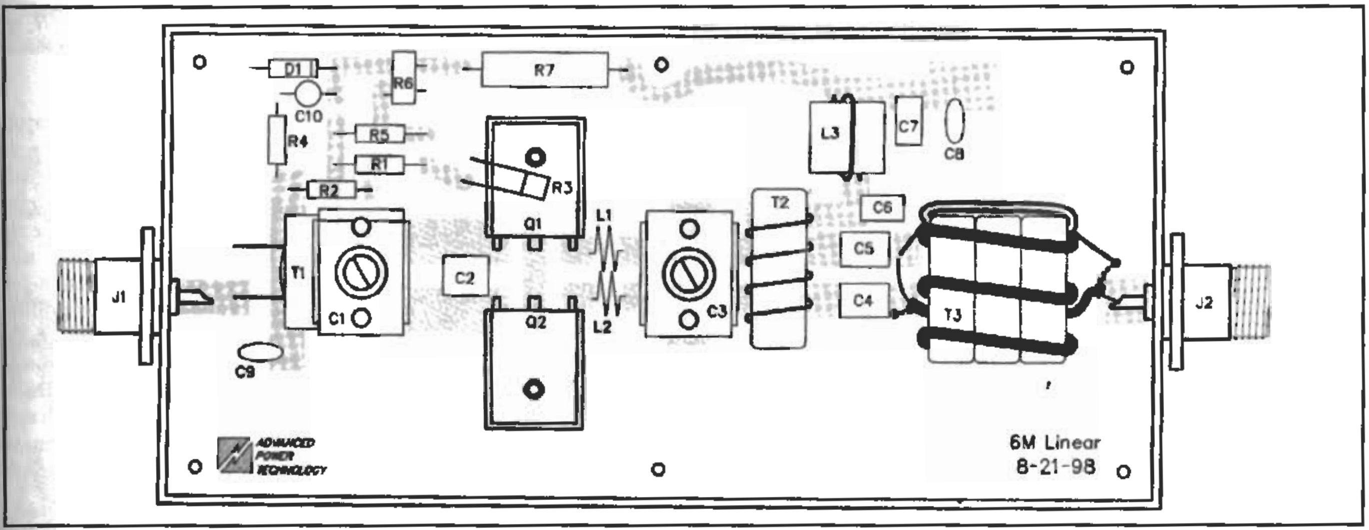

Fig 4—Amplifier parts layout.

AB is calculated from the formula:

$$R_L = \frac{0.98\,V^2_{dd}}{2P_o} \qquad \text{(Eq 1)}$$

This is the load for each device, and P_O is one half of the total in a push-pull circuit. It is shunted by the output capacitance, C_{oss}. As was done for the gate circuit, both output impedances are series-connected to represent the total output impedance. The result for both devices in push-pull is 30.7 Ω in parallel with 75 pF, half the output capacitance of a single device. Though the design goal was 300 W PEP, the amplifier was actually designed for a 400 W load line. This gives a good compromise between efficiency and linearity.

In a classical design, a suitable transformer would be used to set the load impedance, and either the output power or the operating voltage would be adjusted to fit the available turns ratio. Normally, in a low-voltage HF design, the output capacitance is ignored because it is shunted by a much smaller load resistance. At 50 MHz, the effects of the output capacitance must be compensated, so a slightly different approach was taken in this circuit.

WinSMITH was used again to design the output matching. See Fig 2. The output impedance of 30 Ω is rotated south by the effect of the shunt output capacitance, C2. Two options present themselves for compensation. Some additional shunt capacitance could be added to further reduce the equivalent series real part to 12.5 Ω, a series L used to resonate the resulting series C, then a 4:1 transformer used to go to 50 Ω. However, building a reproducible low-loss 4:1 balanced transformer was very difficult, and compensating its leakage reactance further complicates the design. The second option was used: The equivalent series output capacitance was resonated first, then more series inductance was added to rotate the load all the way up to the 1/50-mho conductance circle. Finally, a shunt capacitor was used to resonate the added X_L. The extra L and shunt C form an L-network, which transforms the 20-Ω equivalent series output impedance up to 50 Ω. This results in an easily duplicated design with a smooth, low-Q match.

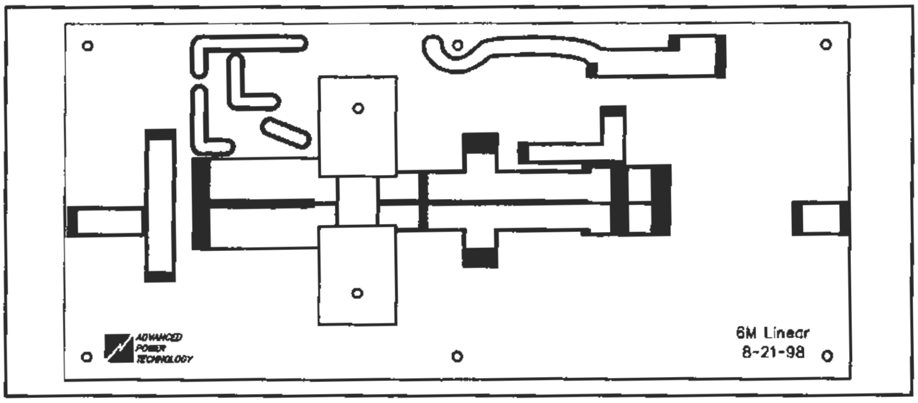

Fig 5—PCB artwork (not to scale, original size 3.35 (7.00 inches).

The dc feed to the drains is provided through a shunt bifilar choke. At this frequency, most ferrite materials exhibit too much loss to be used at this impedance level. A powdered-iron core works famously here.

The balun-transformer function is provided by a simple coax and wire transformer. Two of the windings are provided by 50-Ω Teflon coax, and the third balancing winding by an additional single wire.

The bias network requires some explanation. Power MOSFETs have normal lot-to-lot variations in gate threshold voltage, V_{th}, forward transconductance, G_{fs}, and other parameters. A number of devices were checked for V_{th}, and they were all very close. They were all from the same die lot. The die lot number is marked on the package. For comparison, devices from another lot were checked and were uniformly a half a volt lower. Were this the case for the devices to be used in the amplifier, a dc block would need to be added to each side at the transformer, and the bias-feed network duplicated for each device. Since these devices were uniform, the additional complication of individual gate-bias adjustments was omitted in this design.

Because the gate-bias voltage required to maintain a particular value of idling drain current decreases as the temperature of the die increases, it is necessary to thermally compensate the gate-bias source, or the devices will "run away." A commonly available NTC resistor tracks the temperature of the case. (Refer to Fig 3.) This bias circuit has been in the literature for many years.[6] The ratio of R1 to R3 in part determines the degree of compensation. A smaller value of R1, or a larger value of R5, will increase the thermal sensitivity. A drop of thermally conductive glue keeps the thermistor in contact with the case. Proper operation is indicated when the set value of I_{dq} does not change after the heat sink gets hot from prolonged operation.

Construction

Refer to Fig 4 for the parts layout. A photo master of the artwork is shown in Fig 5. The original size of the artwork is 3.35x7 inches. The circuit board is 1-ounce-copper, double-sided 1/16-inch G-10 PCB material. All four edges of the board and the three sides of the two rectangular cutouts for the transistors are wrapped with copper-foil tape that was soldered in place to

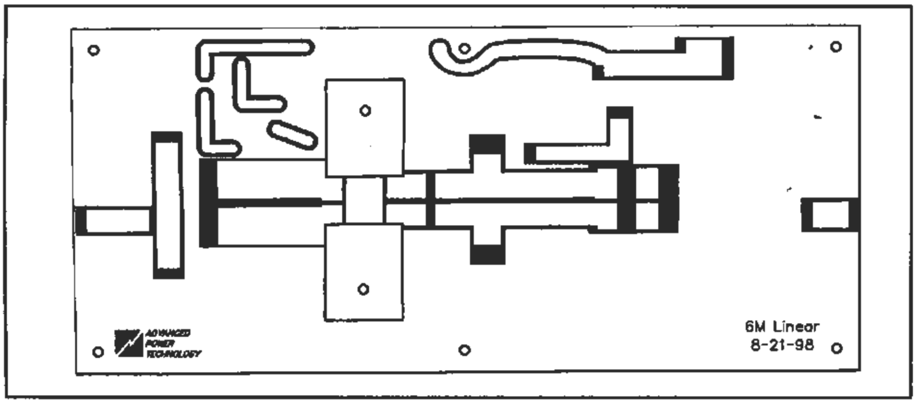

Fig 6—SSB IMD performance.

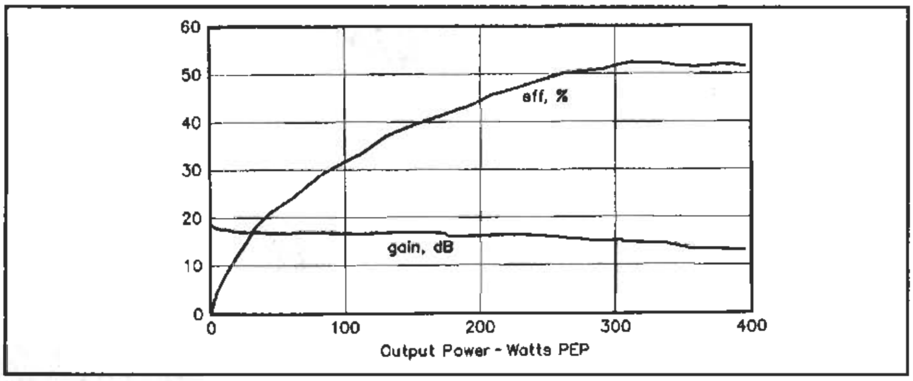

Fig 7—Amplifier efficiency and gain.

Table 1—Amplifier Parts List

C10—1μF, 35 V electrolytic capacitor
C1—215-790 pF Arco[8] #469 mica compression trimmer
C2, C6, C7—1000 pF, 500 V NPO chip cap, KD9 #2020N102J501P
C3—20-180 pF Arco #463 mica compression trimmer
C4-C5, C8-C9—0.01 μF, 500 V chip capacitor
D1—6.8 V, 1 W Zener diode
L1-L2—≈70 nH, 3t #18 AWG enameled wire 0.31-inch diameter 0.25-inch long
L3—2t #20 AWG on Fair-Rite #284301 0402 bead $\mu_i = 850$
Q1 —ARF448A
Q2—ARF448B
R3—10 kΩ NTC Fenwal[10] #140-103LAG-RB1
R6—1 kΩ, 0.5W 10-turn trimmer
T1—Primary 2t #20 PTFE, Secondary 3/16-inch brass tube on Fair-Rite #2843010402 balun core
T2—6t bifilar #20 PTFE on Amidon #T-94-2 toroid $\mu_i = 10$
T3—3t RG-316 coax, 3t #20 PTFE on three Fair-Rite #5961001801 toroids $\mu_i = 125$
TL1, TL2—30-Ω printed line, 0.6 inches long

provide a low-impedance continuous ground plane. The two cutouts for the transistors and the six mounting holes are the only holes in the board. All of the parts are surface mounted, which permits the board to be mounted directly to the heat sink.

This amplifier was built on a 7-inch length of AAVID #60765 heat-sink extrusion.[7] It is 3.5 inches wide, 1.5 inches deep and has nine fins. With 50 CFM of air blown across it, the devices will easily maintain thermal stability in a 30°C environment. The heat sink is not big enough for anything but very intermittent use without a fan to assure adequate airflow across the fins. The input and output connectors are each secured by two #4-40 screws and holes tapped into the base of the heat sink. A cover is recommended for safety; fairly high RF voltages are present.

Power Supply

Power for the amplifier needs to be fairly well regulated, since any ripple will show in the output signal as undesired 120-Hz AM. For on-air testing at the author's home, a very simple power supply was constructed using a 500 W, 120 to 240 V isolation transformer to drive a full-wave, centertapped rectifier circuit with 50,000 μF of filtering. Under SSB conditions, this is adequate. For CW, it needs some better regulation or the output power will sag over the length of a dash, and there will be some detectable hum. A regulated supply capable of providing 80 V at 6 A is needed.

Performance

This is the first-known, class-AB application of the ARF448 parts. Until now, the only other linear application is in a pulsed-mode linear amplifier for magnetic-resonance imaging. The SSB performance was encouraging because these devices were developed to serve the ISM plasma-generation market and no attention to linear performance was given in their design. The IMD performance with 200 mA of quiescent bias was better than expected. (See Fig 6.) The amplifier was tested with up to 0.5 A of I_{dq}. While the IMD performance improves somewhat at this level, the efficiency degrades significantly.

The gain and efficiency objectives have been met, as shown in Fig 7. The gain is 14.3 dB at 300 W PEP. The efficiency peaks at 51% at the same power. Under single-tone conditions, the drain efficiency is 61% at 250 W. The bandwidth of the amplifier is determined by the input network. The Smith-Chart plot of the input impedance shows the tracks for 50, 50.5 and 51 MHz. With the network adjusted for best match at 50.5 MHz, the SWR at the ± 0.5 MHz bandwidth points is 1.3:1. It would be difficult to increase this SWR bandwidth enough to cover the full 4 MHz of the amateur 6-m band without resorting to resistive loading, which would then reduce the available gain.

Conclusion

This paper has presented a 50 MHz, 300-W PEP linear amplifier using plastic-packaged, high-voltage MOSFET transistors. This is the first-known implementation of a full-duty-cycleclass-AB amplifier using these transistors. The design challenges, approaches to their solution and the resulting amplifier performance are shown. The parts, construction and mechanical layout all have been described in sufficient detail to permit duplication. The new line of plastic-packaged RF power transistors from APT offers designers a new cost-effective solution for efficient layout and performance.

Notes

[1] R. Frey, *A 50 MHz, 250 W Amplifier Using Push-Pull ARF448A/B*, APT9702, Advanced Power Technology Inc.

[2] R. Frey, "A Push-Pull 300 Watt Amplifier for 81.36 MHz," *Applied Microwaves and Wireless*, April 1998.

[3] Smith Chart is a trademark and property of Analog Instruments Co, New Providence, New Jersey.

[4] *WinSMITH*, copyright Eagleware Corp, 1995, available through Noble Publishing, Inc.

[5] Fair-Rite Products Corp, PO Box J, One Commercial Row, Walikill, NY 12589.

[6] H. Granberg, "Wideband RF Power Amplifier," *RF Design*, February 1988.

[7] AAVID Thermal Technologies Inc, Box 400, Laconia, NH 03247.

[8] Arco Electronics, 5310 Derry Aye, Agoura Hills, CA 91301.

[9] KD Components Inc, 2151 Challenger Way, Carson City, NV 89706.

[10] Fenwal Electronics, Inc, 450 Fortune Blvd, Milford, MA 01757.

By Paul Hewitt, WD7S From *QEX*, July 2000

A No-Bandswitch, Dual-Band VHF Desktop Amplifier

An old design idea together with new tubes yields legal-limit output on 6 and 2 meters from a small box.

Dual-band VHF amplifiers are by no means new in Amateur Radio. Several designs have appeared over the years, both commercial and homebrew.[1,2] The most recent commercial design was the Henry Tempo 6N2, now out of production for more than 20 years. There has never been a desktop capable of the new 1500-W PEP amateur power limit for both the 6- and 2-meter bands, until now. The amplifier I describe was designed primarily for meteor-scatter and weak-signal work, but the conservative design of the power supply and cooling system makes it well suited for EME (moonbounce) also. The amplifier uses a pair of Svetlana 3CX800A7s, operating in class AB2-push-pull on 2 meters and AB2-parallel on 6 meters. Features include: legal-limit output, no key-down time limit, no band switching, either-or operation, 50-60 W of drive for 1500 W PEP output on both bands, small size and light weight. In an effort to make this design easier to duplicate, off-the-shelf com-ponents and materials were used wherever possible. If you prefer not to "roll your own," Alpha Power should have a commercial version available soon.

Output Tank Circuit and Construction Details

Since this was a new tank design, I decided to use a box-in-a-box approach for the layout to make adjusting and debugging much easier. This should also make duplication of the tank circuits much easier for those of you who wish to use a separate, outboard power supply. The tank compartment (See Figs 1, 2 and 7) starts with a 7×7×2-inch Bud chassis with two surplus 11-pin ceramic sockets[3] mounted three inches apart (center-to-center) and four inches from the rear of the chassis. The sockets are mounted ¼ inch below the chassis in 2½-inch-square holes with each socket oriented at 45° to the hole (Fig 3). This is similar to an Eimac 2216 socket for the 8877. The Eimac #1906 chimneys were used in the normal manner. This is probably the upper frequency limit for this mounting scheme due to possible instability problems at UHF. One problem encountered while mounting the sockets this way is the small amount of tube exposed above the chimney to grasp for removal. This is easily solved by connecting several ¼-inch-wide hose clamps together end to end, then clamping them to the anode cooler to serve as a temporary handle.

The front, top and sides of the tank com-

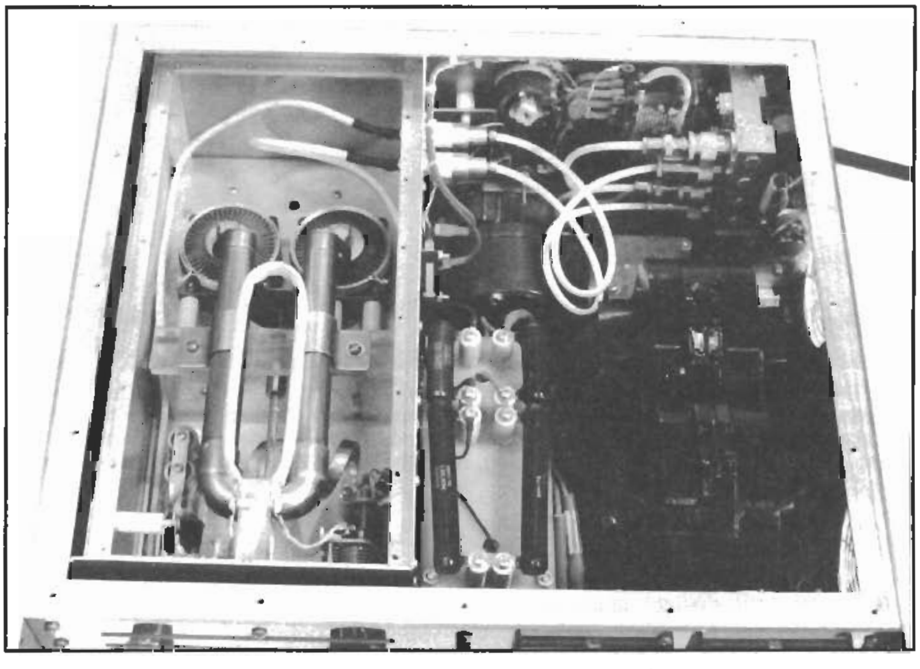

Fig 1—A top view of the amplifier. The plate transformer in the front right and high-voltage power assembly next to it occupy the majority of the control side with the blower, filament and control transformers sitting just behind. The tank compartment with the cover removed shows the plate line on its Plexiglas stand.

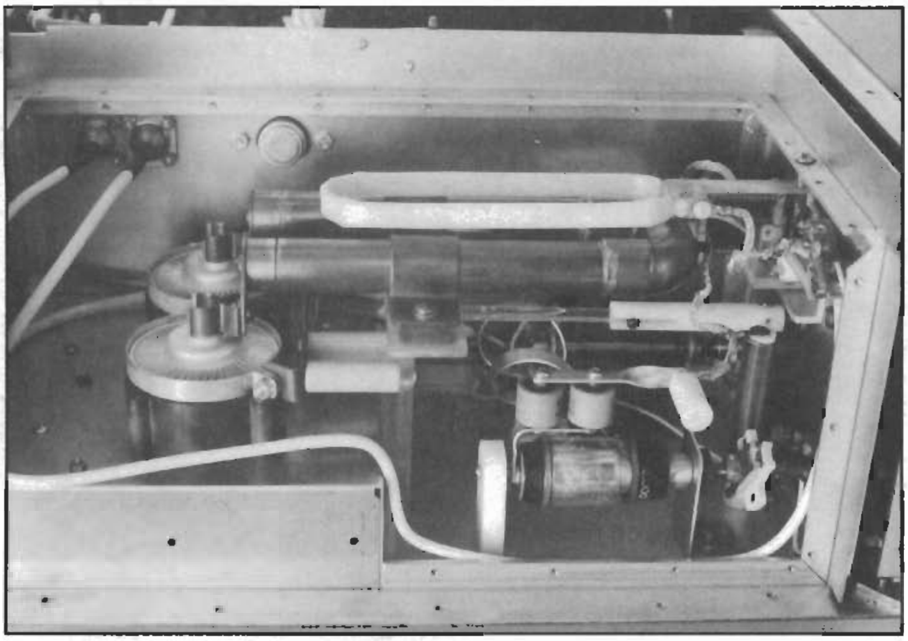

Fig 2—Side view of tank compartment. The 2-meter load capacitor on its ceramic standoffs mounted to the Plexiglas stand is in the center with the 6-meter tank coil just visible below the spline shaft assembly. The 6-meter tune capacitor on its L bracket is in the lower right front.

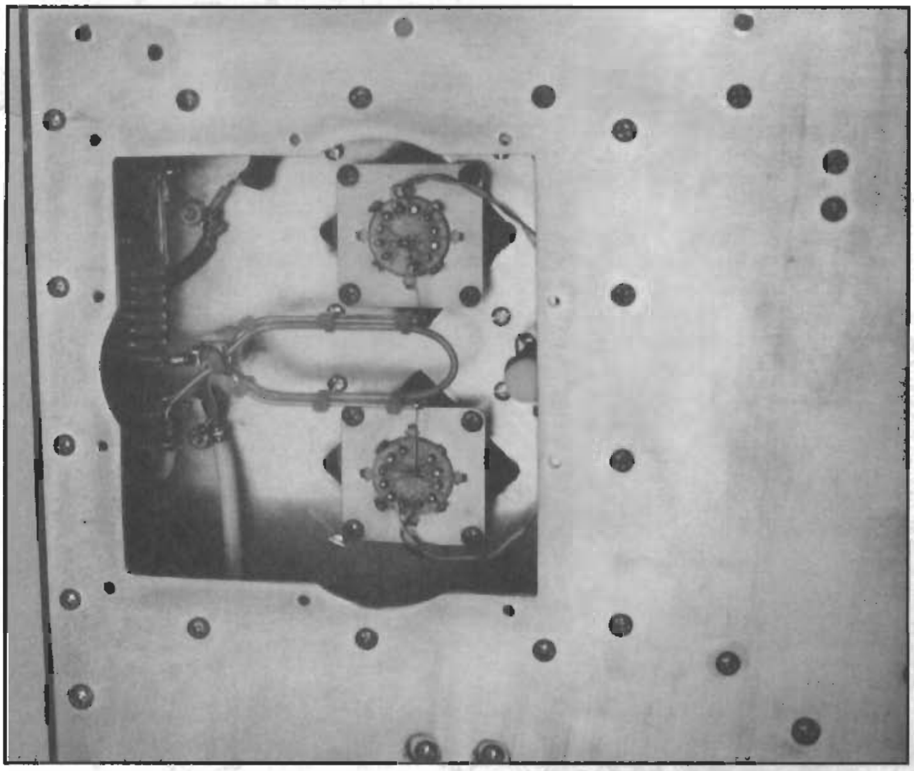

Fig 3—Bottom view of amplifier with cathode compartment cover removed The filament choke is to the right with the 2-meter input network visible centered between the tube sockets. Both input tuning capacitors are mounted on the rear panel and one of the 6-meter input coils is just visible on the left. Note the use of coax for exciter lines to keep feedback paths minimized.

partment are all made from 1/16-inch aluminum with 1/2×1/2×1/8-inch aluminum angle brackets at all corners. The completed compartment measures 7×6 1/4×14 1/2 inches (WHD). The front and right-side compartment panels extend to the top of the main cabinet cover to keep inlet and outlet air separated. The compartment cover (not shown) has 176 0.2-inch holes above the tubes for cooling exhaust. When this box is attached to the floor of the main cabinet, it leaves a 4×7-inch hole at the back before the rear panel is attached. This opening serves as access for a dip meter when checking the 2-meter tank resonance and tuning range.

The 2-meter tank circuit consists of a shorted $\lambda/4$ balanced line section, tuned by a homebrew split-stator capacitor at the tube end of the line. Designing this type of tank circuit is made very pain-less with information and design exam-ples provided in references.[4] There are several advantages of using a push-pull tank circuit instead of a parallel arrangement at 144 MHz. In the push-pull arrangement, C_{out} and C_{in} of the tubes are in series, allowing a total C_{out} and C_{in} that is one quarter of that with the parallel arrangement. This allows lower values of loaded Q, resulting in higher efficiency and reduced compo-nent heating, minimizing thermal drift. The low value of C_{out} in this case allows the use of a $\lambda/4$ line in place of a $\lambda/2$ line, increasing bandwidth and decreasing size.

The line section, L2 of Fig 4, is made from 3/4-inch type-K copper pipe; the shorted end is made from one standard 90° copper elbow and one "street" 90° copper elbow. The open ends of the line are closed with brass plugs of the same outside diameter as the pipe, with a portion machined to fit inside the pipe. Standard copper pipe caps could be substituted for these plugs. The total length of the line section is 7 3/4 inches, from the ends to the inside of the shorted end. The plate-to-anode connections are made with 1/16-inch copper plate fastened between copper fuse clips on the anode connectors and threaded holes in the ends of the brass plugs.

The line section is supported by a Plexiglas stand and held in place by two copper clamps. The stators of the 2-meter **TUNE** capacitor, C3, are sup-ported by 1 1/2-inch ceramic standoffs attached to the Plexiglas stand (Fig 5). A 3/8×5/8-inch ear from each stator attaches to a 3/8-inch-wide strap around each of the tubes' anode coolers. The rotor of C3 is a two-inch disk of 1/16-inch copper, which is mounted to a piece of 1/4×28 brass all-thread rod. The threaded rod is held by a brass fitting (tapped for 1/4×28) that is attached to the Plexiglas stand. The threaded rod is coupled to a sliding spline shaft consisting of 1/4-inch Plexiglas rod inside a 1/2-inch piece of Teflon rod that has been drilled through the center. The Teflon tube is attached to a short piece of 1/4-inch stainless-steel rod that exits the tank compartment through a 1/4-inch panel bushing. This shaft is coupled to the turns counter with a synchronous gear belt (Fig 6). This shaft relocation via the gear belt serves only to improve the front-panel appearance; it may be omitted if desired.

The 2-meter tank circuit is coupled to the antenna by an adjustable reso-nant link. The link position is adjusted by a #10-24 screw through a threaded Plexiglas block, to which the link is attached (Fig 2). One end of the adjusting screw is supported by the top-cover angle bracket and the other end by a 2-inch-long piece of 1/2-inch angle mounted to the front panel of the compart-

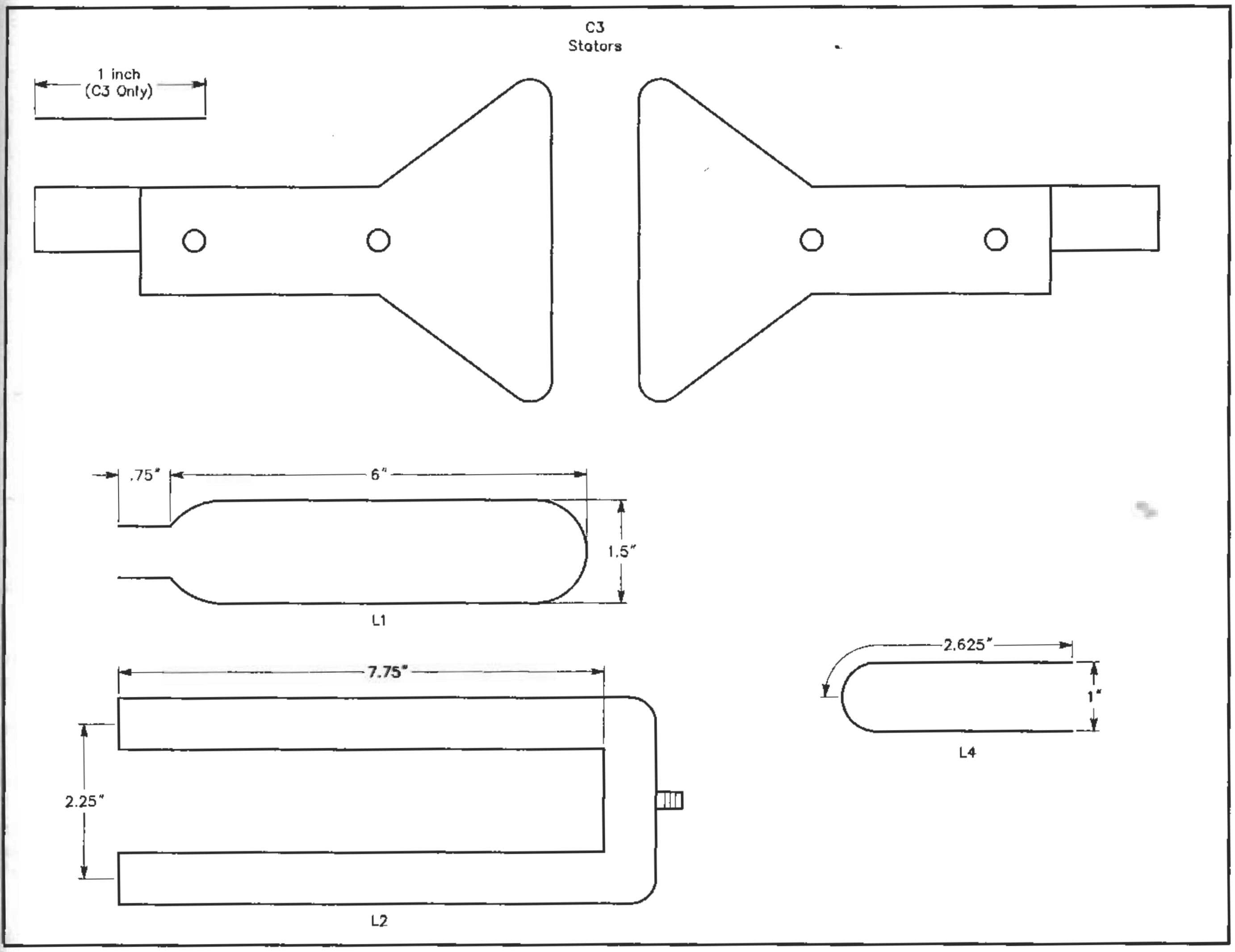

Fig 4—Details of 2-meter input and output tank indicators and the 2-meter TUNE capacitor stators.

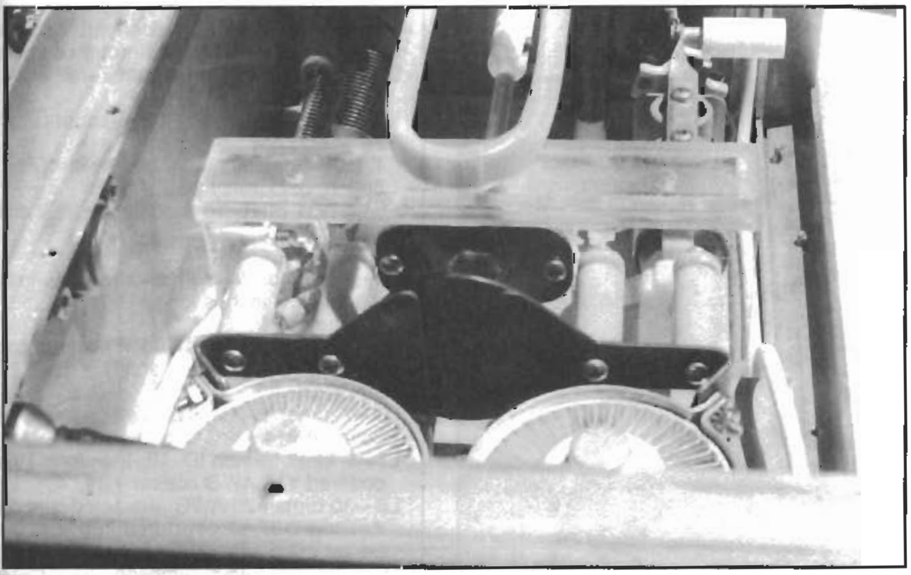

Fig 5—Rear view with the plate line removed showing the 2-meter TUNE capacitor.

ment. The link is made from $^{3}/_{8} \times ^{1}/_{16}$-inch silver-plated copper strap, covered with a $^{3}/_{8}$-inch Teflon sleeve; it is attached to the Plexiglas block with four #10-32 nylon screws. Silver-plated coax braid attaches the link to the 2-meter **LOAD** capacitor, C2, and the isolation relay, K3. The link-adjustment screw is accessed through $^{3}/_{8}$-inch hole in the main cabinet top cover during initial setup, then the hole is covered with a plastic plug.

As with all push-pull circuitry, the key to successful operation is symmetry. All stray capacitance must be divided equally between both sides of the line and tubes. Also, try to keep ferrous materials out of the tank compartment to prevent imbalance.

The 6-meter tank is a normal π network designed for a plate-load resistance of 1150 W with a loaded Q of approximately 22. The design parameters are shown in Table 1. The input of the π network is connected to the center of the shorted end of the balanced line section where a #10-32 brass stud has been soldered in place. This is the low-impedance point of the line section, and anything but a dead short can be attached here without affecting the performance of the line.[5] To verify that the physical center of the line is also the RF center, couple a dip meter to the line and tune for a dip around 144 MHz. Then without moving the dip meter (DM), touch the tip of a lead pencil at points along the line until you find the spot that has the least effect on the DM. This is the attachment point for the plate choke and the 6-meter tank.

To keep the loaded Q of the π network as low as possible, full 51-54 MHz opera-

Table 1—Print out of operating parameters for the 6-meter input and output networks

```
The values for the PI matching network were calculated with
Elmer (W5FD) Wingfield's new formulas found in the more recent
ARRL handbooks.
FREEWARE Courtesy of KD9JQ
Triode Amplifier Program  Version 2.0
For Grounded Grid Operation

(2) USER Biased 3CX800A7 at 50.0 MHz Rated for FORCED AIR
DC Plate Volts        =  2150.0   V (2500 V Max)
Max Plate Voltage     =  2150.0   V
Peak Plate Swing      =  1900.0   V
Min Plate Voltage     =   250.0   V
Plate Current Peak    =     3.368 A
Plate Current DC      =     1.119 A
Grid Current DC       =     0.063 A
Cath Current Peak     =     3.651 A
Design Plate RL       =  1128.1   Ω
RL for Matching       =  1154.8   Ω
Plate Dissipation     =   805.5   W (1600 W Max)
Grid Dissipation      =     3.2   W (8 W Max)
Cathode Bias          =     8.2   V

Conduction Angle      =   187.8   Degrees
Peak Grid Voltage     =    36.8   V
Zin @ Cathode         =    24.6   Ω
Pin Drive (PEP)       =    41.0   W
P0 @ Plate (PEP)      =  1600.0   W
P0 to Load (PEP)      =  1637.9   W
DC Power Input        =  2405.5   W
Efficiency            =    66.5   %
Power Gain            =    16.0   dB
Cath to Grid Cap      =    52.00  pF
Plate to Grid Cap     =    12.20  pF

T Match Input                     Pi Match Output
RS  =  50.0 Ω                     RP  =  1154.8 Ω
L2  =   0.521 µH                  CI  =    50.0 pF
CI  =  44.924 pF                  L1  =     0.2377 µH
L1  =   0.422 µH                  C2  =   232.0 pF
RFC =   0.195 µH                  RL  =    50.0 Ω
Zin =  22.6 –j 9.9 Ω              QL  =    22.0
QL  =   5.0
```

tion was not attempted. The 6-meter **TUNE** capacitor, C1, operates very near its minimum capacitance because of strays and the C_{out} of both tubes. These strays account for a large portion of the 6-meter tune capacitance, which causes the majority of the tank circulating current to flow through the blocking caps, C7-C8. I used parallel Centralab 858 "door-knobs" for the blocking capacitors and I haven't experienced any problems. Capacitors with larger current ratings (such as HT-57s) would be a better choice for longer duty cycles.

The 3-30 pF vacuum-variable capacitor is mounted to the floor of the tank compartment with an L-bracket bent from 1/16-inch aluminum. Use the bearing retainer nut to attach the capacitor to the bracket, but be sure to add shims to the bearing to make up for space lost by the thickness of the bracket, or backlash will occur. The other end of the capacitor is supported by a #6-32 brass screw in the end of the capacitor that passes through a piece of Teflon bar stock mounted to the floor of the cabinet. A #6-32 hex nut holds the end of L3 and a piece of 3/8-inch strap (which supports C7 and C8) tight to the end of the vacuum cap. The other end of L3 is supported by a 1/2×1-inch Teflon standoff. Silver-plated braid ties this end to the 6-meter **LOAD** cap, which is mounted on the sidewall of the tank compartment.

Setting the Output Tanks

The 6-meter tank can be checked for tuning range with an SWR analyzer at the output connector. Connect a 1.1 kΩ, noninductive resistor from one of the tube anodes to ground to simulate the 1100-W plate-load resistance. You should be able to achieve a 1:1 match at both ends of the desired tuning range. Some stretching or compression of L3 may be required to bring the tuning range to 50-51 MHz. Be sure the top cover of the tank compartment is in place, as the stray capacitance it introduces accounts for a portion of the tuning capacitance for 6-meters.

The 2-meter tank was set up entirely

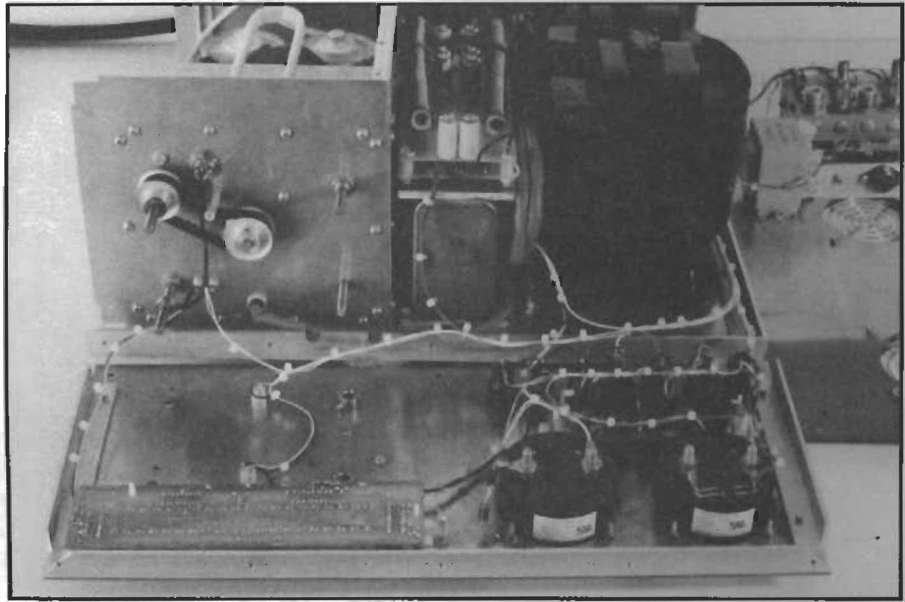

Fig 6—Front view of the amplifier with front panel folded down showing the location of high-voltage power-supply components at right and the gear-belt layout on the left. The bar-graph display PC board is visible in the lower left with the control board just beyond the two panel meters.

Fig 7—RF deck schematic. Unless otherwise specified, use 1/4 W, 5%-tolerance carbon composition or film resistors. See Table 3 for part-supplier contact information.

C1—3-30 pF vacuum-variable capacitor
C2—150 pF air-variable, 1400 V capacitor
C3—Split-stator air-variable (see text and Fig 4)
C4—325 pF air-variable, 1400 V capacitor
C5, C6—4.3-75 pF APC-style air-variable trimmer capacitor
C7, C8—1000 pF, 5 kV doorknob capacitor
C9—2500 pF, 2.5 kV feedthrough capacitor
D1—8.2 V, 50 W stud-mount Zener diode, 1N2806B
K3—Jennings RJ1A SPDT vacuum relay, 26.5 V dc coil
K4, K5—DowKey model 260B with "C" option, coax relay 26.5 V coil
K6—SPDT relay, 10 A contacts, 24 V dc coil
L1—See Fig 4
L2—See Fig 4
L3—2 turns of 3/8×1/16-inch strap, 1 1/2-inch diameter, 2 inches long
L4—See Fig 4
L5—8.5 inches of #14 Teflon covered copper wire with 6 inches tied tightly inside L4
L6—7 turns, 1/2-inch ID×1 1/8-inch long Teflon covered #14 AWG copper wire
L7—9 turns 1/2-inch ID×1 1/2-inch long Teflon covered #14 AWG copper wire
L8—30 turns #26 AWG enameled wire on an Amidon T-37-17 powdered-iron core over center conductor of coax
M1—0-2 A, 3.5-inch Simpson panel meter
M2—0-100 mA, 3.5-inch Simpson panel meter
RFC1—36 turns #18 AWG enameled wire, tight-wound on 1/2-inch-diameter Teflon rod
RFC2, RFC3—Ohmite Z-50
RFC4—Ohmite Z-144
V1, V2—Svetlana 3CX800A7

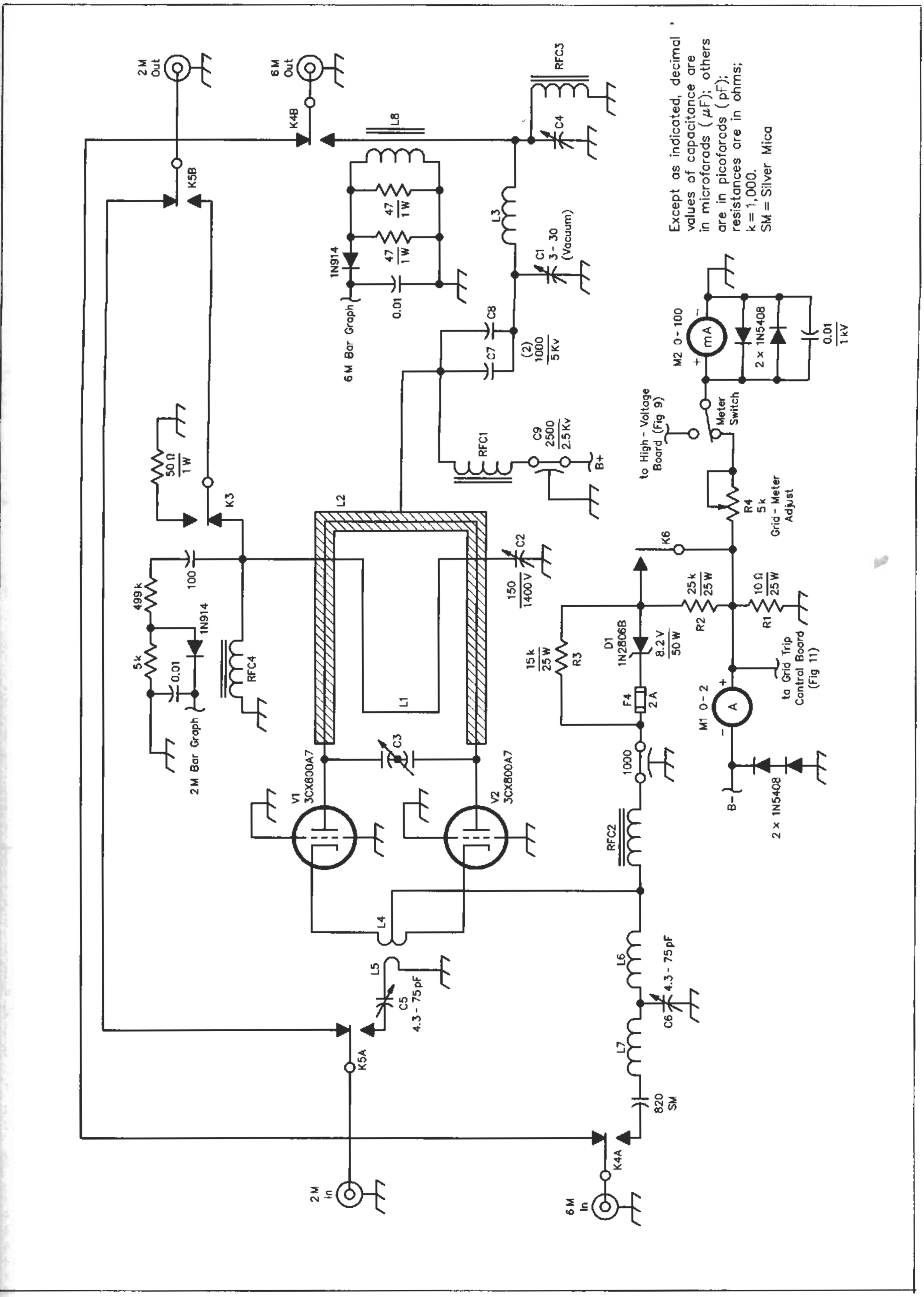
2 M Out
6 M Out
K4B
K5B
L8
RFC3
C4
1N914
47 1 W
47 1 W
0.01
6 M Bar Graph
L3
C1 3 - 30 (Vacuum)
C8
C7
(2) 1000 5Kv
RFC1
C9 2500 2.5Kv
B+
50 Ω 1 W
K3
L2
499k
100
1N914
5k
0.01
2 M Bar Graph
RFC4
C2
150 1400 V
L1
C3
V1 3CX800A7
V2 3CX800A7
L4
L5
C5 4.3 - 75pF
K5A
2 M In
6 M In
K4A
820 SM
L7
C6 4.3 - 75pF
L6
RFC2
1000
F4 2 A
D1 1N2806B 8.2 V 50 W
15k 25 W R3
25k 25 W R2
10 Ω 25 W R1
K6
R4 5k Grid - Meter Adjust
to High - Voltage Board (Fig 9)
Meter Switch
M2 0 - 100
mA
2 × 1N5408
0.01 1 kV
M1 0 - 2
A
to Grid Trip Control Board (Fig 11)
B-
2 × 1N5408

Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms;
k = 1,000.
SM = Silver Mica

with a dip meter inserted through the opening in the back of the compartment. Set the output link about $^{3}/_{8}$-inch from the top plane of the line; this is close to the final position and adds stray capacitance. To keep loaded Q low, try to set the line length so it doesn't take very much tune capacitance to resonate the line at 144 MHz. Start with a line section a little longer than shown and carefully trim it to resonance. Remember that 1 inch of line length equals approximately 10 MHz of tuning range! The tank should resonate at 144 MHz with about $^{1}/_{4}$ inch of space between the plates of C3. When the desired tuning range is found, install a $^{1}/_{4}$×28 lock nut on the threaded shaft of C3 to stop travel at the upper frequency limit. This will prevent accidental tuning of the tank to the third harmonic of 50 MHz, which could potentially damage the front end of a 2-meter exciter. Remember that the isolation of most coax relays is very poor at 144 MHz: It was only 35 dB with the DowKey 260B used here. For this reason, a Jennings RJ1A vacuum relay, K3, was added in series with the 2-meter transfer relay. The vacuum relay disconnects the 2-meter link from the coax and terminates the coax in a resistive load during 6-meter operation. Make sure the electrical length of the coax between the relays is more than $\lambda/10$ and less than $3\lambda/8$ at 144 MHz. With these safety measures in place, third-harmonic energy at the 2-meter input port was measured at –17 dBm during 1500-W 6-meter operation.

Input Networks

The 2-meter input network consists of a single tuned-link, air-core transformer (Figs 3 and 4). This coupling method requires slightly more drive than a link-

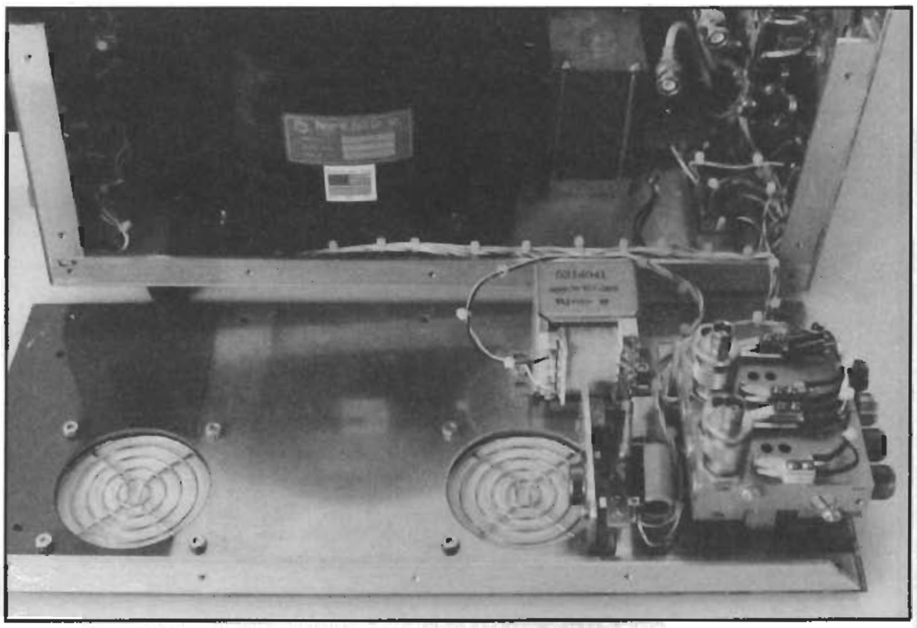

Fig 8—View of the right-hand side panel folded down showing the two transfer relays and the low-voltage power supply. The filament transformer is visible to the right of the plate transformer. Note the use of 80-mm fan covers for the cooling air inlets.

Table 2—Operating parameters of the amplifier

Plate voltage no load, 2350 V
Plate voltage @ 1 A, 2150 V
Zero signal plate current, 35 mA
6-Meter drive, single tone, 49 W
6-Meter power output,1420 W
6-Meter plate current, 980 mA
6-Meter grid current, 75 mA
Apparent efficiency, 67.4% (feed-through power not subtracted)
2-Meter drive, Single tone, 50 W
2-Meter power output, 1275 W
2-Meter plate current, 950 mA
2-Meter grid current, 68 mA
Apparent efficiency 62.4% (feed-through power not subtracted)

Table 3—Parts suppliers

Fair Radio Sales Co, Inc
1016 East Eureka St
PO Box 1105
Lima, OH 45804
tel 419-227-6573, 419-223-2196
fax 419-227-1313
e-mail **fairadio@wcoil.com**
URL **http://www.fairradio.com/**

Mouser Electronics
2401 Hwy 287 N
Mansfield, TX 76063
tel 800-346-6873
fax 817-483-0931
e-mail **sales@mouser.com**
URL **http://www.mouser.com/**

Newark Electronics
4801 N. Ravenswood Ave
Chicago, IL 60640-4496
tel 800-463-9275, 773-784-5100
URL **http://www.newark.com/**

Peter W. Dahl Co, Inc
5869 Waycross Ave
El Paso, TX 79924
tel 915-751-2300
fax 915-751-0768
e-mail **pwdco@pwdahl.com**
URL **http://www.pwdahl.com/**

RF Parts Co
435 S Pacific St
San Marcos, CA 92069
tel 760-744-0700, 800-737-2787
(orders only)
fax 760-744-1943
e-mail **rfp@rfparts.com**
URL **http://www.rfparts.com/**

Surplus Sales of Nebraska
1502 Jones St
Omaha, NE 68102-3112
tel 402-346-4750, 800-244-4567
(Orders only)
fax 402-346-2939
e-mail **grinnell@surplussales.com**
URL **http://www.surplussales.com/**

Svetlana Electron Devices
8200 S Memorial Pkwy
Huntsville, AL 35802
tel 256-882-1344, 800-239-6900
fax 256-880-8077
e-mail **sales@svetlana.com**
URL **http://www.svetlana.com/**

coupled, half-wavelength resonant line, but saves a lot of space. Two references[6, 7] give formulas and rules of thumb for designing this type of transformer, but it still required a lot of "cut and try" before suitable sizes for L4 and L5 were found. The loaded Q of the resonant input link is approximately 3.5 before the coupled-in resistance from the secondary modifies it. The degree of mutual coupling of air-core transformers is an elusive value that makes the final loaded Q difficult to calculate. The input tuning capacitor, C5, could be changed to a 25-pF unit since the one used here turned out to be much larger than needed. Both sides of C5 are above chassis ground; it requires a nonmetallic screwdriver to adjust.

Six-meter drive and B– are applied to the center tap of L4, providing parallel cathode drive for 50-MHz operation. A common **T** network, with a loaded Q of five, matches the 50-W line to the 24.6-Ω input impedance of the parallel tubes. All the cathode pins of the sockets are tied together with buss wire in a star pattern. One-inch-long bus-wire leads connect the ends of L4 to the center of the bus-wire stars on the sockets. Again, symmetry is all-important in balanced operation.

Both input networks were set up with an antenna analyzer and two 50-W carbon resistors to simulate the input load impedance. Tack-solder the resistors from cathode to ground on each socket with the shortest-possible leads. Short leads are very important on 2 meters because lead inductance and stray capacitance become quite significant. The resistors present

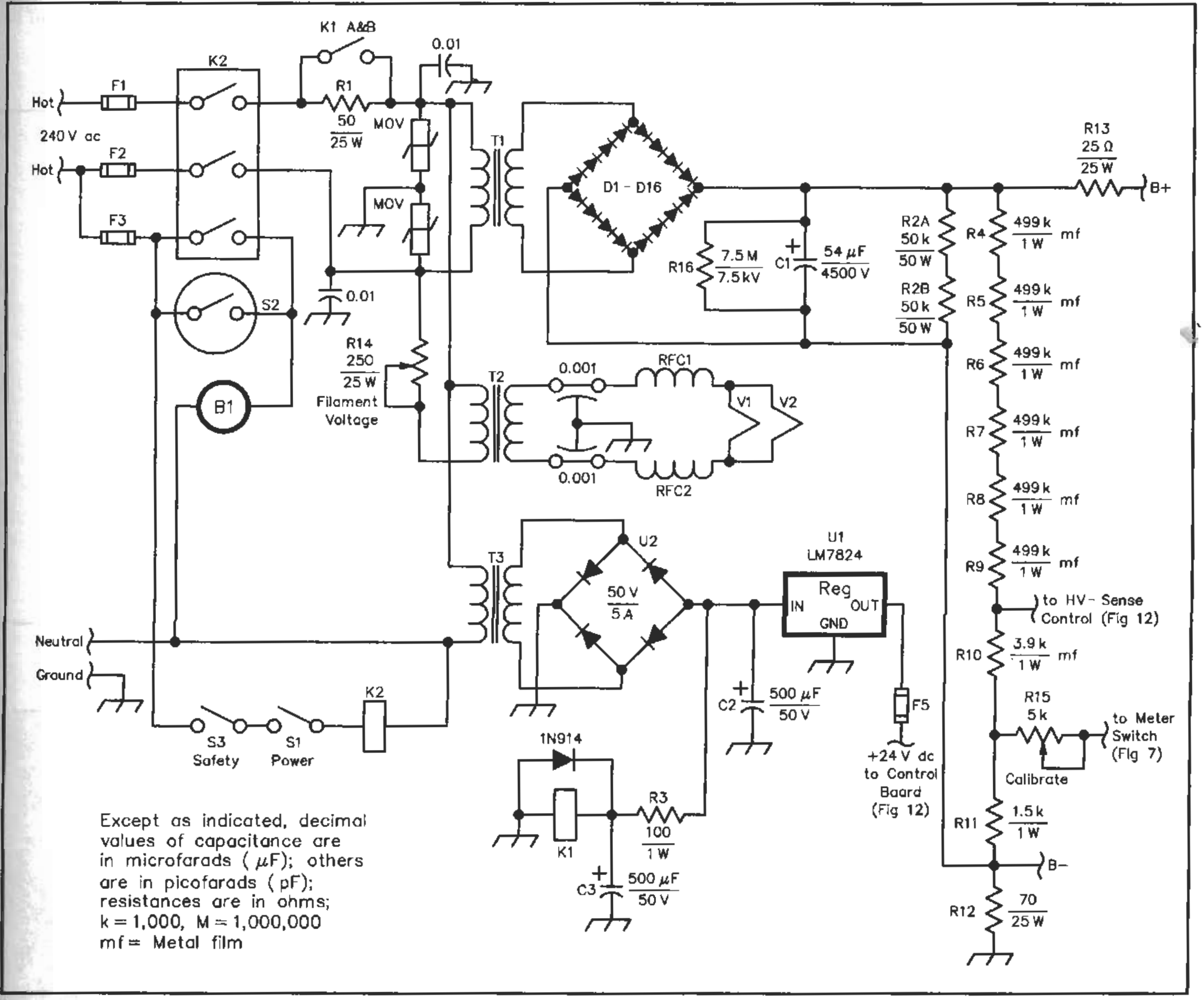

Hot
240 V ac
Hot
F1
F2
F3
K2
K1 A&B
R1
50
25 W
0.01
MOV
MOV
T1
D1 - D16
R16
7.5 M
7.5 kV
C1
54 µF
4500 V
R2A
50 k
50 W
R2B
50 k
50 W
R13
25 Ω
25 W
B+
R4
R5
R6
R7
R8
R9
499 k
1 W
mf
to HV- Sense Control (Fig 12)
R10
3.9 k
1 W
R15
5 k
to Meter Switch (Fig 7)
Calibrate
R11
1.5 k
1 W
B-
R12
70
25 W
S2
0.01
R14
250
25 W
Filament Voltage
B1
T2
0.001
0.001
RFC1
RFC2
V1
V2
T3
U2
50 V
5 A
U1
LM7824
Reg
IN
OUT
GND
C2
500 µF
50 V
F5
+24 V dc to Control Board (Fig 12)
Neutral
Ground
K2
S3
Safety
S1
Power
1N914
K1
R3
100
1 W
C3
500 µF
50 V

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000, M = 1,000,000 mf = Metal film

Fig 9—Power-supply schematic. Unless otherwise specified, use 1/4 W, 5%-tolerance carbon composition or film resistors. See Table 3 for part-supplier contact information.

B1—Dayton #4C761 squirrel-cage blower
C1—54 µF, 4500 V oil-filled electrolytic capacitor
C2, C3—500 µF, 50 V electrolytic capacitor
D1-D16—1N5409 1200 PIV, 3 A Silicon rectifiers, four diodes per string
F1, F2—20 A, 250 V ceramic fast-blow fuse
F3—3 A, 250 V fuse
F5—2 A slow-blow fuse
K1—DPST relay, 25 A contacts, 24 V dc coil
K2—3PST relay, 25 A contacts, 120 V ac coil
MOV—130 V metal oxide varistor, V130LA5
R1—150Ω 25 W wire-wound resistor
R2A, R2B—50 kΩ 50 W wire-wound resistor
R3—100 Ω, 1 W carbon resistor
R4-R9—499 kΩ, 1 W metal film resistor
R10—3.9 kΩ, 1 W metal film resistor
R11—1.5 kΩ, 1 W metal film resistor
R12—70 Ω, 25 W wire-wound resistor
R13—25 Ω, 25 W wire-wound resistor
R14—250 Ω, 25 W rheostat
R15—5 KΩ, 10-turn potentiometer
RFC1, RFC2—12 bifilar turns of #18 AWG enameled wire on 1/2-inch-diameter Teflon rod
S1—SPST 5 A lighted panel switch
S2—Temperature snap-disc control Grainger # 2E245
S3—SPDT 5 A microswitch
T1—Peter Dahl plate transformer: 240 V primary, 1800 V, 1 A CCS secondary
T2—240 V primary, 16 V, 5 A secondary
T3—120 V primary, 20 V, 2 A secondary
U1—LM7824 TO-3 case
U2—5 A, 50 V rectifier bridge

25 W to the parallel 6-meter network and 100 W to the 2-meter push-pull network. Connect the analyzer to the respective input ports and adjust each network for lowest SWR. Some adjustment of the 6-meter input coils (L6 and L7) may be required to achieve a 1:1 match. If the 2-meter network does not present a good match, some adjustment in length of L4 and L5 may be necessary. Again, the covers must be in place and tubes must be in their sockets when measuring the match. Both filament leads and the filament choke must be in their final positions because the cathode-to-filament capacitance affects input tuning at 144 MHz. Remember, this analyzer method only gets you close to the final tuning points; settings will be different under live conditions. Finally, don't forget to remove the temporary resistors from the cathode and tank circuits after the initial adjustments are done.

AC Mains and Low-Voltage Power Supplies

Several protective measures are built into the amplifier to protect the tubes and the operator. The ac mains are brought into the amplifier with four-conductor cord to keep the neutral and ground separated per the *NEC* (Fig 9). Three fuses are used to keep the blower's ac source separate from the fused plate transformer. This was done so the blower's cool-down delay still functions after a high-voltage fault, which removes ac from the entire amplifier in dual-fuse designs. A thermal snap switch in the tank compartment keeps the blower running after shutdown or a high-voltage fault if the exhaust air has been above 110°F and is not yet below 90°. This only happens after several minutes of continuous operation. You may want to use a switch with slightly higher ratings (120°F on, 110°F off) if your shack is often warmer than the 90° off point used here. The 3CX800A7's filaments dissipate only 20 W each during cutoff and natural convection is more than enough to cool the tubes when shut off.

AC voltage for K2, the main control relay, is also taken from F3, the blower fuse, to make sure blower voltage is available before the amplifier can be powered up. A safety switch that closes when the top cover is in place supplies ac to S1, the main power switch. Power-supply inrush protection is provided by K1, which closes approximately one second after K2 and effectively removes R1 from the transformer primaries. The delay period is set by R3 and C3. There are two MOVs across the ac lines after the contacts of K2 and K1. Arcing at the relay contacts can and will produce voltage spikes and spike protection on the fuse side of K2 will not always save the high-voltage diode strings and other components. Filament voltage is adjusted at the primary of T2 by R14, a 250-Ω, 25-W rheostat. Filament voltage is measured at the filament choke via two leads brought to the back panel of the amplifier. Purists may want to measure voltage at the filament pins, but the low filament-current demand of these tubes makes the voltage drop between the sockets and the choke negligible. The rheostat also provides passive filament-inrush protection.

The low-voltage supply—consisting of T3, U1 and U2—supplies regulated 24 V at 1 A. The bridge rectifier and 24-V regulator are mounted to a piece of ¾×1-inch aluminum angle that is placed directly in the cooling-air inlet path (Fig 8).

High-Voltage Power Supply

The high-voltage power supply consists of a Peter Dahl Hypersil plate transformer and a full-wave bridge rectifier with capacitor input filter. The supply is capable of producing 2150 V at 1 A CCS. The entire supply is assembled as one piece, then installed on the floor of the main cabinet next to the plate transformer (Fig 6). The oil-filled filter capacitor is sandwiched between two 3×6½-inch pieces of Plexiglas, in turn held together with four #10-24×5-inch-long flat-head screws. The rectifier strings and high-voltage-meter multiplier resistors are mounted to a 3×6½-inch PC board supported by ¼-inch spacers on the same four screws. A 3×6½-inch piece of fiberglass board holding the two bleeder resistors, B+ current limit-ing resistor and B– float resistor tops off the

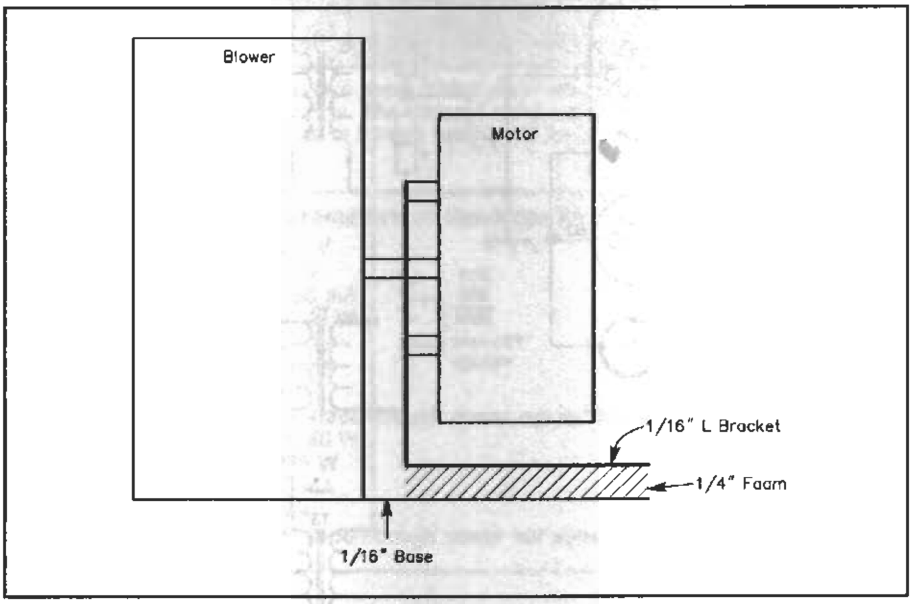

Fig 10—Layout of the blower modifications for sound reduction.

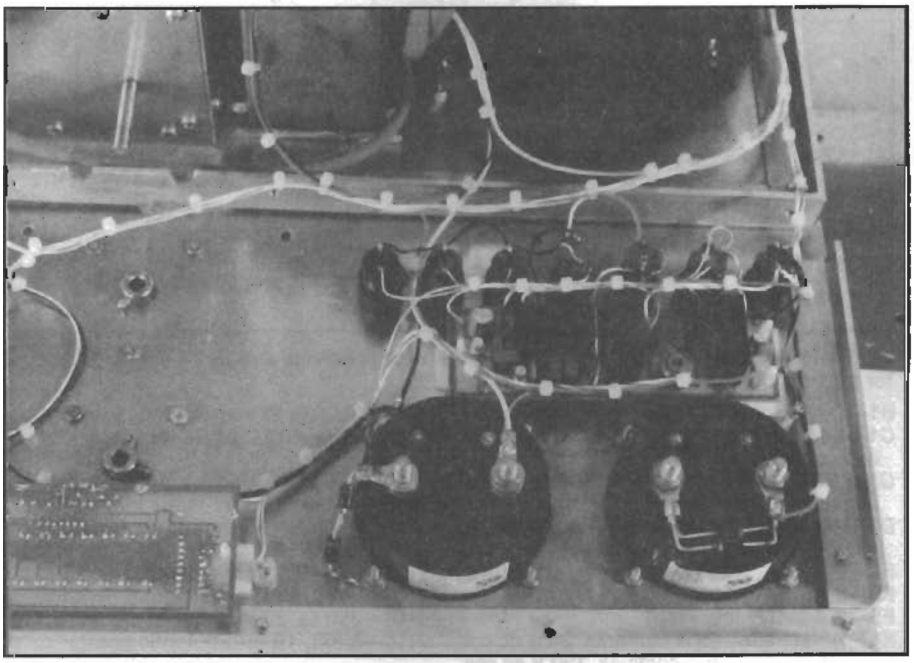

Fig 11—View of control board.

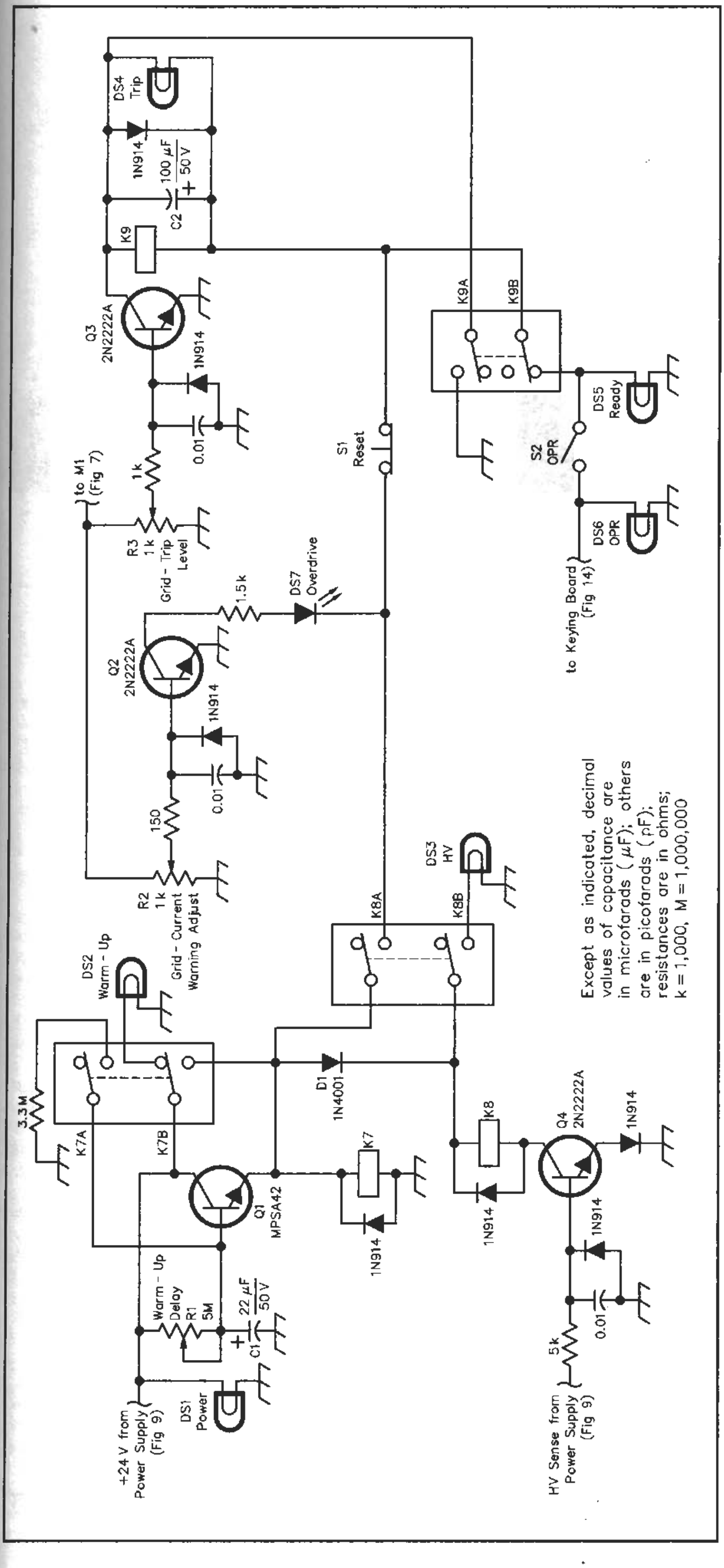
DS4 Trip
1N914
100 µF 50 V
C2
K9
Q3 2N2222A
1N914
0.01
1k
to M1 (Fig 7)
R3 1 k Grid - Trip Level
S1 Reset
K9A
K9B
DS5 Ready
S2 OPR
DS6 OPR
to Keying Board (Fig 14)
DS7 Overdrive
1.5k
Q2 2N2222A
1N914
0.01
150
R2 1 k Grid - Current Warning Adjust
DS3 HV
K8A
K8B
DS2 Warm - Up
3.3M
K7A
K7B
D1 1N4001
K8
Q4 2N2222A
1N914
K7
1N914
Q1 MPSA42
1N914
1N914
Warm - Up Delay R1 5M
22 µF 50 V
C1
+24 V from Power Supply (Fig 9)
DS1 Power
HV Sense from Power Supply (Fig 9)
5k
0.01
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000, M = 1,000,000

Fig 12—Control schematic. Unless otherwise specified, use ¼ W, 5%-tolerance carbon composition or film resistors. See Table 3 for part-supplier contact information.

C1—22 µF, 50 V electrolytic capacitor
C2—100 µF, 50 V electrolytic capacitor
DS1—24 V lamp part of S1 (see Fig 9)
DS2, DS3, DS5—24 V panel light
DS4—24 V lamp part of S1
DS6—24 V lamp part of S2
K7, K8, K9—DPDT 5 A DIP relay, 24 V dc coil
R1—5 MΩ PC-mount potentiometer
R2, R3—1 kΩ PC-mount potentiometer
S1—Normally closed momentary-contact lighted panel switch
S2—SPST lighted panel switch

stack. All of the power resistors on the top board are mounted on one-inch ceramic standoffs. High voltage is routed to the tank compartment with test-lead wire and a high-voltage feed-through capacitor. The diode strings do not use equalizing resistors or capacitors as they are from the same batch and there is plenty of PIV headroom. Spike protection for the string is in the transformer primary, where it belongs. The high-voltage supply has three bleed-down paths for the filter capacitor: the bleeder resistors, the high-voltage meter multipliers and a 7.5 MΩ, 7.5-kV resistor at the terminals of the capacitor itself. Even with these redundant safety measures, *never* assume that they are working. *Always follow standard safety procedures when working with any high voltage, including the ac mains: It can kill you.*

Cooling

Whole-cabinet cooling is accomplished via a Dayton model 4C761 squirrel-cage blower. Cooling air is drawn into the cabinet at the right-hand side through two 2¾-inch holes, one on each side of the plate transformer. This removes heat from the bleeder resistors and other components before it enters the blower inlet. The Svetlana 3CX800A7 datasheet recommends airflow of 11 cfm at a back pressure of 0.2 inches (of water), for 600 W dissipation at sea level and 25°C inlet air temperature per tube. For two tubes, this is 22 cfm at 0.2 inches for 1200 W of dissipation. At a 1500-W output level with 60% efficiency, the tubes only dissipate 960 W.

The socket sub-mounting method I used resulted in a back pressure of 0.2 inches with this blower, as measured on my bench with a home-brew manometer. Since the blower is rated for 43 cfm at 0.2 inches (sea level assumed), there should be plenty of headroom for different elevations and inlet air temperatures. Remember that the above calculations are for continuous dissipation, while SSB and CW operation rarely approaches 50% of these values.

The mounting flange at the bottom of the blower was cut off so the outlet can align with the 2-inch-tall cathode chassis. The blower motor was taken off its squirrel cage and mounted to an aluminum **L** bracket. The squirrel cage was mounted to a 4×4-inch piece of 1/16-inch aluminum that supports the **L** bracket through a piece of 1/4-inch foam (Fig 10). This sound-isolation method was borrowed from Alpha Power Inc, and it works very well.

Control Board and Metering

The control PC board (Fig 11) is mounted on nylon standoffs just below the panel meters and above the row of switches and panel lights. As with most indirectly heated, oxide-coated cathodes, a warm-up period is required; for the 3CX800A7, a minimum of three minutes is recommended. When the amplifier is turned on, regulated 24 V is applied to the warm-up

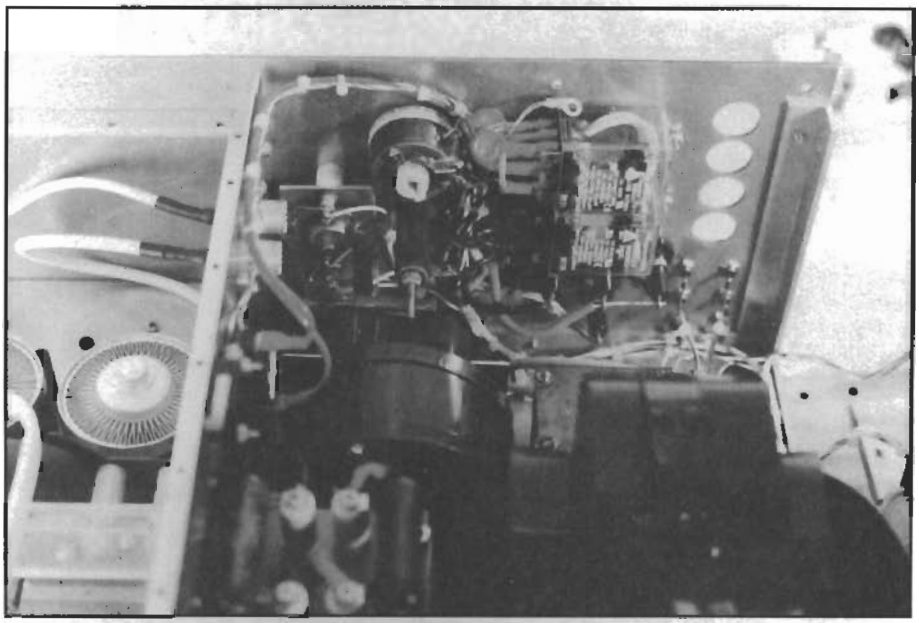

Fig 13—(right) View of panel showing K1 and K2. The bias heat sink and assembly is visible below the filament rheostat.

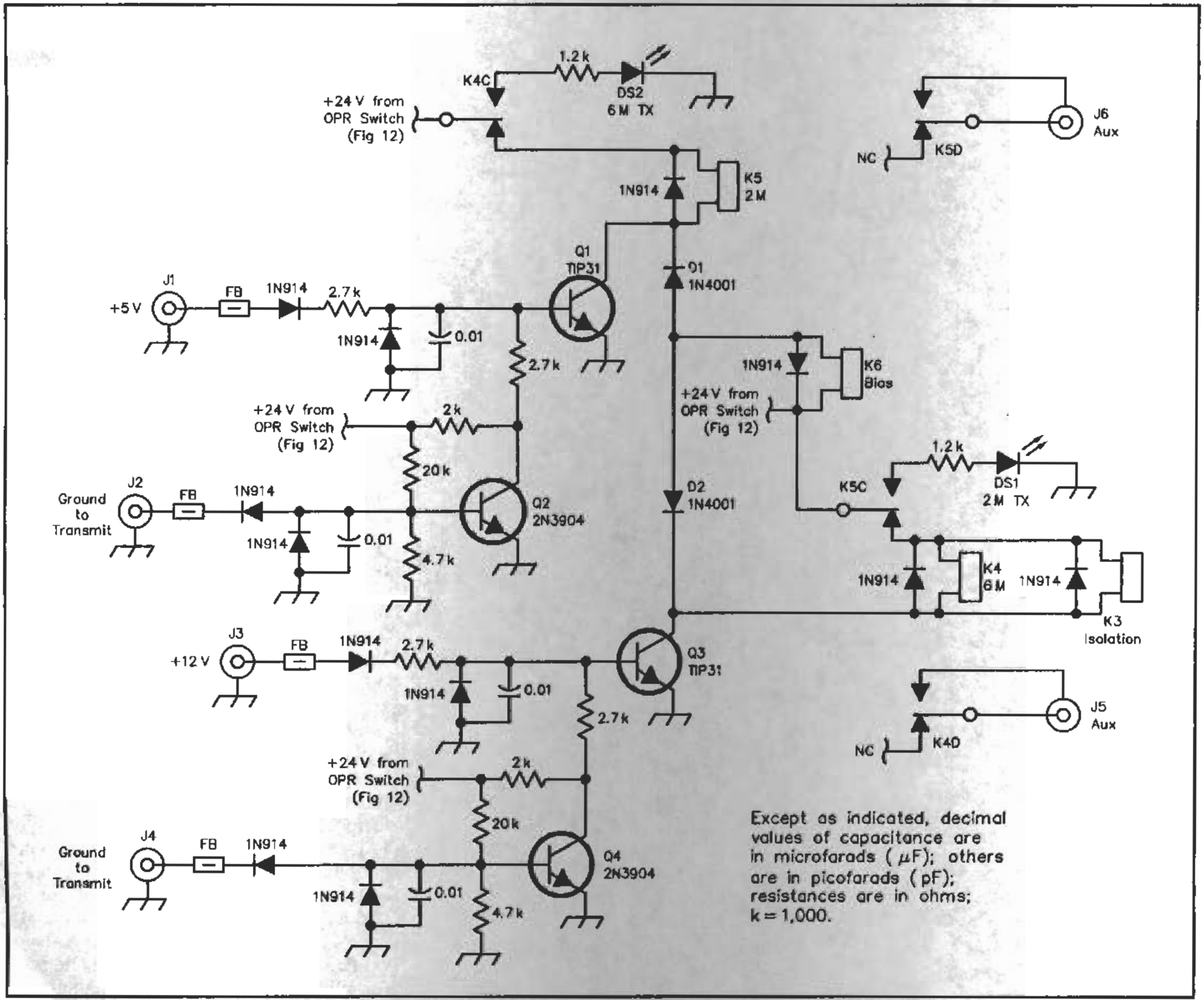
K4C
1.2 k
DS2
6 M TX
+24 V from
OPR Switch
(Fig 12)
K5
2 M
1N914
K5D
J6
Aux
NC
Q1
TIP31
D1
1N4001
J1
+5 V
FB
1N914
2.7 k
1N914
0.01
2.7 k
1N914
K6
Bias
+24 V from
OPR Switch
(Fig 12)
+24 V from
OPR Switch
(Fig 12)
2 k
20 k
J2
Ground
to
Transmit
FB
1N914
1N914
0.01
4.7 k
Q2
2N3904
D2
1N4001
1.2 k
DS1
2 M TX
K5C
1N914
K4
6 M
1N914
K3
Isolation
J3
+12 V
FB
1N914
2.7 k
1N914
0.01
Q3
TIP31
2.7 k
J5
Aux
NC
K4D
+24 V from
OPR Switch
(Fig 12)
2 k
20 k
J4
Ground
to
Transmit
FB
1N914
1N914
0.01
4.7 k
Q4
2N3904
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000.

Fig 14—Schematic of the dual keying-circuit PC board. Unless otherwise specified, use 1/4 W, 5%-tolerance carbon composition or film resistors. See Table 3 for part-supplier contact information.

delay circuit consisting of Q1, R1 and C1 (Fig 12). When sufficient voltage appears at the base of Q1, relay K7 closes, with one set of contacts latching the relay on while the other set of contacts resets C1. If high-voltage is present, Q4 is switched on and K8 closes, lighting DS3 and supplying 24 V to the rest of the amp-lifier. Diode D1 was added between the field of K8 and the field of K7 to make sure K7 closes first. The proximity of the control board to the magnetic field of the plate transformer affects the timing of K7 slightly.

With the low grid-dissipation rating of these tubes, grid over-current protec-tion is necessary. Grid current is mea-sured as voltage drop across R1 of Fig 7, a 10-Ω, 25-W wire-wound resistor. R3 on the control board is set to fire K9 with Q3 at a grid current of 120 mA. One set of contacts latches the relay on; the other set interrupts the 24-V line to the **OPR** (operate) switch. Capacitor C2—across the field of K9—stops relay chatter on voice peaks. Lamp DS4 is part of the grid-reset switch S1, and it lights while K9 is energized. A second grid-current-sensing circuit, using Q2, is set to light DS7 on the front panel at 75 mA. This LED is used as a grid-current warning indicator and can be set at the grid-current level you prefer. The trip points of these two circuits and the grid-current meter reading are ad-justed while the amplifier is discon-nected from the ac supply.

To do so, connect a 24-V external supply to the 24-V bus of the amplifier after the timing circuit. Connect the positive lead of a variable-voltage dc supply to the ungrounded side of R1 of Fig 7 through an accurate dc milliammeter. Connect the external-supply negative lead to the chassis. With the grid-current panel meter disconnected, slowly increase the voltage of the bench supply until the milliam-meter reaches the set points of 120 mA for grid trip and 75 mA for the warning LED. Adjust R3 and R2 (Fig 12), respectively, to set the trip points. After the trip points are set, reconnect the grid-current meter and adjust R4 (Fig 7) for a full-scale reading of 100 mA using the same method as for the trip points. This "cold" method of setting grid trip points is a lot safer for the tubes than removing B+ and then applying drive to induce grid current.

The panel meter, M2, serves as the high-voltage meter as well as the grid-current meter. The calibration of the high-voltage portion was "roughed in" with the amplifier off and all voltages removed. A 1.5-V dc supply was connected to the multiplier side of R15 on the high-voltage rectifier board, then R15 is adjusted for a full-scale reading on M2. This corresponds to a full-scale reading of 3000 V. Meter calibration was then checked with a high-voltage probe and an accurate DMM, after the amplifier was turned on.

Plate current is measured directly in the B– line of the high-voltage supply, which is held slightly above ground by R12 on the power-supply assembly. Both meters are protected by 1N5408 diodes; two series-connected diodes on the B– bus keep it below approximately 1.4 V and two more are connected back-to-back from the multimeter to ground.

There are relative-output bar-graph displays for each band above the tune and load controls. These displays use two LM3914 bar-graph drivers per 20-segment display. A current-transformer pickup unit is used for 6 meters and a voltage-sensing pickup for 2 meters. The dual-bar-graph-display schematic can be found in Forrest Mims' book[8] and the 6- and 2-meter pickups are shown in Fig 7. The displays look appealing but turned out to be more trouble than they are worth, since I still use external wattmeters while I tune. The bar-graph displays' rapid response time would be better suited for plate- and grid-current meters.

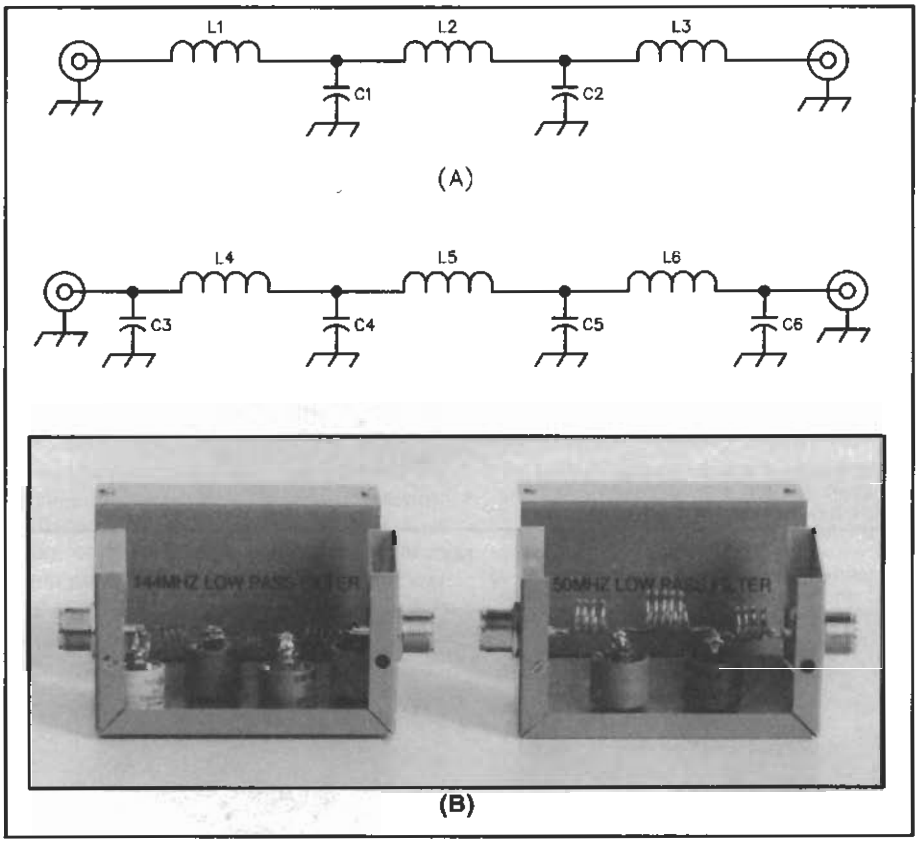

Fig 15—Details of the 2- and 6-meter low-pass filters. See Table 3 for part-supplier contact information.

50 MHz
C1, C2—50 pF doorknob capacitor
L1, L3—4 turns #14 AWG copper, 5/16-inch ID×1/2-inch long
L2—5 turns #14 AWG copper, 5/16-inch ID ×5/8-inch long

144 MHz
C3, C6—25 pF doorknob capacitor
C4, C5—40 pF doorknob capacitor
L4-L6—3.5 turns #16 AWG silver-plated copper wire, 1/4-inch ID×5/16-inch long

Bias and Keying

Operating bias is developed across D1 of Fig 7, an 8.2-V, 50-W stud-mounted Zener diode. The Zener, along with K6, R2 and R3, is mounted on a 2×3-inch piece of 0.100-inch aluminum that serves as a heat sink (Fig 13). The entire assembly is mounted on two 1/2-inch Teflon stand-offs on the back panel of the cabinet. The 2-A cathode fuse is mounted on the back panel alongside the ac-line fuses.

The keying PC board is mounted on the rear floor of the main cabinet (Fig 8). The PC board contains keying circuits for both 6 and 2 meters (Fig 14). Because they are identical, I'll only describe the operation of one.

At rest, Q2 is biased on, holding the base of Q1 low. Grounding J2 will take the base of Q2 low, turning it off. This allows voltage to appear at the base Q1, turning it on and energizing K5. Alternatively, applying +12 V to J1 will also turn Q1 on, keying the amplifier. Both keying inputs are logic-compatible. The bias relay is energized through D1 or D2, which comprise a two-input **OR** gate. Each transfer relay is a DowKey #260B with the "C" option, which is a pair of DPDT signaling switches. Each of the relays' supply voltages is routed through one of the opposite relay's normally closed signaling contacts. This mechanical **EXCLUSIVE-OR** gate keeps the amplifier from being simultaneously keyed on both bands. The normally open contacts are used to light transmit LEDs on the front panel; these are mounted between their respective tune and load controls to help eliminate confusion while tuning.

With this type of transfer relay, input and output relay sequencing is obviously

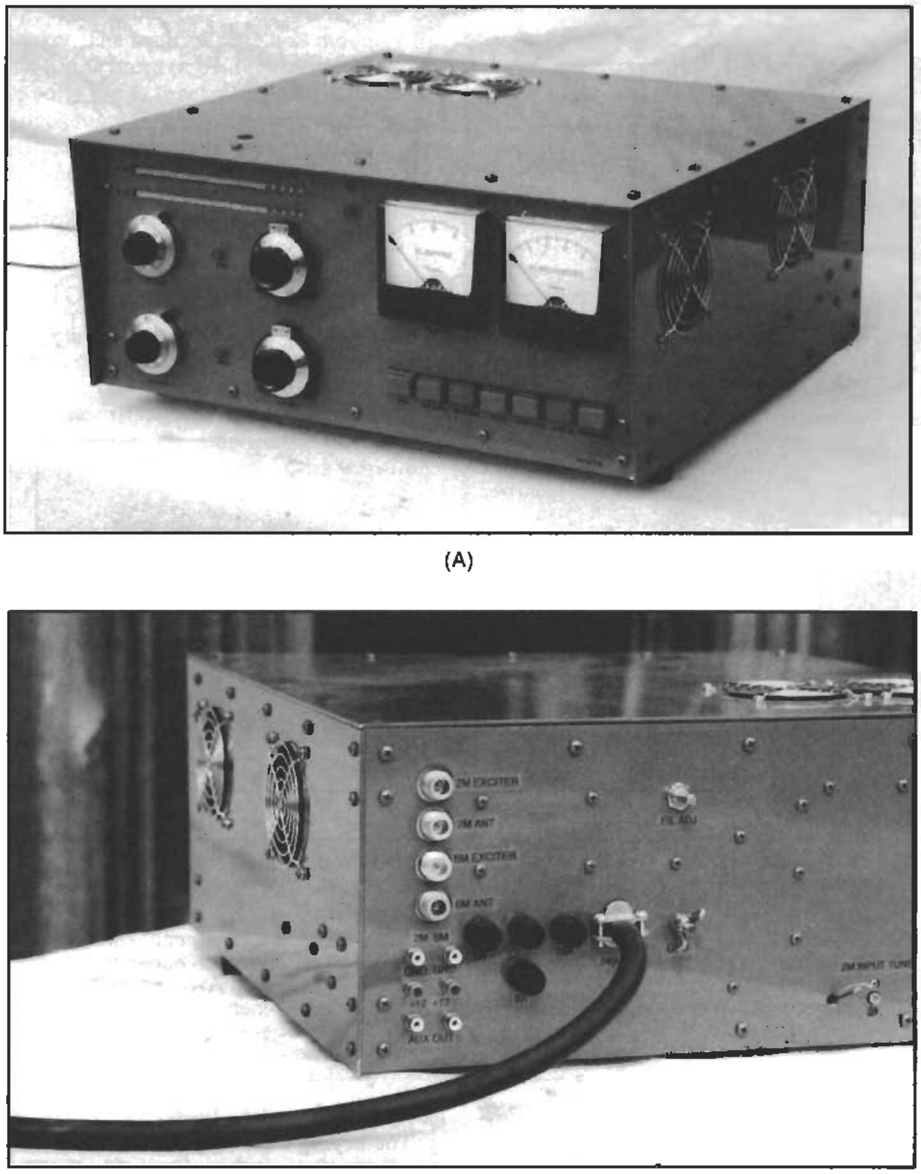

(A)

(B)

Fig 16—The completed amplifier ready for its ride to the mountain top. (A) Front view, (B) back panel.

not an option, and hot switching of the relays will result unless preventive measures are taken. One option is to key the amplifier and let it key the exciter. Each relay's spare set of signaling contacts was brought to the back panel for this possible use. Another option, which I employ, is to use an outboard keying sequencer. The sequencer takes care of my mast-mounted preamplifiers as well as the exciter-amplifier timing. A suitable sequencer can be found in the references.[9] When using separate exciters for each band, the spare set of signaling contacts can be used for audio muting of the unkeyed exciter while transmitting.

General Construction Notes

The entire amplifier cabinet was built using common hand and power tools. All aluminum was cut with carbide saw blades in a radial arm saw and table saw. Blades with a 5° negative hook angle seem to work best for aluminum. The main amplifier cabinet measures $18\times16\times7\frac{1}{2}$-inches (WDH). The plate transformer determined the cabinet height. The entire cabinet is built from 0.100-inch aluminum sheet and $\frac{3}{4}\times\frac{3}{4}\times\frac{1}{8}$-inch aluminum angle stock. All outside corners are secured with #8-32 panhead screws tapped into the angle stock. The area under the plate transformer and high-voltage supply assembly was beefed up with a second layer of 0.100-inch aluminum on the cabinet floor. A square hole cut in the cabinet floor provides access the cathode compartment; it is normally covered with a piece of $\frac{1}{16}$-inch aluminum. All round holes were cut with chassis punches. The cabinet was painted with Dupont Chromaclear, which is a two-step (color coat/clear coat) automotive paint. After the base coat was applied over the primer, dry transfer lettering was applied to the front panel controls and Brothers P-touch labels were applied to the rear panel. Two coats of clear were applied over the base coat to finish the job.

All interior wiring was done with scrap Teflon-coated wire. The wiring harnesses were arranged so that all side panels can be removed and laid flat next to the chassis for ease of alignment and debugging. The internal coax runs were made with Teflon-dielectric coax to handle these high power levels, most other coax falls short, particularly at 144 MHz.

Adjustment and Operation

Use an ohmmeter to check the ac paths and RF deck for possible shorts and wiring mistakes. Be sure to blow out the entire cabinet with compressed air to remove any hidden debris. Set the filament-voltage rheostat to maximum resistance. If you haven't already done so, apply 24 V from an external supply to the 24-V bus and check the timing, control and keying circuitry for proper operation. Connect the amplifier to a suitable ac mains supply and connect an accurate DMM to the filament-voltage test points. Connect a 2-meter exciter and dummy load. Turn on the main power switch and quickly set the filament voltage to 13.5 V, or slightly less. *Never* operate the filaments below the 12.9-V recommended minimum. Be sure to check the filament voltage again after 5 to 10 minutes to detect any thermal drift in the rheostat. Verify that the plate is at approximately 2375 V.

Both tank circuits should be set to the tuning values found during initial setup. After the three-minute delay has elapsed, key the amplifier with no drive and check for a cathode idling current of approximately 30-50 mA. Apply a little 144-MHz drive and tune C5 in the input circuit for minimum reflected power. The 2-meter output link should be preset approximately $\frac{1}{2}$ inch above the top plane of the balanced line. Still with little drive, adjust the 2-meter tuning capacitor for maximum power output. Adjust the 2-meter load capacitor for peak output power and then leave it there until you are finished. Next, adjust link spacing as you would a load capacitor. Keep increasing the drive while increasing the link coupling for maximum power output, without exceeding grid-current limits and while also touching up the main tune capacitor, C3. Once the desired output power level is reached, check the link coupling by slowly decreasing the coupling until a slight decrease in output power is observed. Then increase the coupling slightly past the point where output power peaks and grid current is reduced to within operating limits. This should be very close to critical coupling, which will result in maximum efficiency. Go back and touch up the input network for minimum reflected power. The link is now set for operation and any further tuning adjustments can be made with C2 and C3.

Tune the 6-meter tank circuit the same way you would any other p net-work: by slowly increasing drive and tuning for maximum power output and best efficiency while touching up the input network for lowest reflected power. Then increase loading until a 2% decrease in power output is observed.

With values of loaded Q below 30 in the output tanks, there is no way to keep harmonics below the 60-dB-down figure required by the FCC without outboard filtering. Harmonic filters are a very small price to pay for the higher efficiency, lower drive require-ment and reduced thermal drift that the lower loaded Q provides. I've included descriptions of two suitable filters from the references;[10,11] they are very easy to build. The silver-mica capacitors originally in the filters have been replaced with surplus door-knob capacitors for better power-handling capability. With both filters installed all harmonics and spurious signals are more than 70 dB down from the fundamental.

Final Thoughts

This amplifier was completed just before the summer E-skip season and was immediately put on the air. On-the-air signal reports were all good, and they compared well to the 8877 amplifier I normally use. Close in IMD testing was done with the help of NN7DX who is located nine miles away, over flat ground. Results showed no excessive signal bandwidth while our antennas were positioned to bring signals down to S-9 levels during 1500-W output testing. The amplifier runs cool to the touch even after several hours.

The biggest problem encountered with the operation of the amplifier had to do with interfacing with certain exciters. Some of the new multiband exciters provide multiple antenna ports that can be configured for HF, 6- and 2-meter outputs, but provide only one buffered keying line. The ICOM-746 is built this way, but it does provide two unbuffered keying lines: one for HF and 6 meters and one for VHF. The ICOM-706-series radios also provide two unbuffered lines but only two antenna ports. Both of these radios require some sort of external switching for the keying line—and the antenna line in the case of the '706—if this amplifier and an HF amplifier are both used. ICOM does provide band logic as a variable voltage at the radio's ACC plug. I've designed a simple decoder/buffer that uses this logic and the radio's internal power supply to automatically switch the keying line between HF, 6- and 2-meter amplifiers. The buffer will sink up to 3 A of relay current. It also selects from three separate ALC input lines for those of you that employ sequencers using the ALC line for transmit inhibit. The decoder is described in "Automatic Amplifier Selection for the ICOM IC-746, -736 and -706MKII Transceivers," *QST*, May 2000, pp 33-36. With the decoder in place, amplifier selection with an IC-746 becomes totally automatic when changing bands between HF, 6 and 2 meters. For the IC-706 series, an additional coaxial relay can switch one of the antenna ports between HF and 6 meters to allow automatic selection. The decoder also works with the ICOM IC-736, choosing between HF and 6 meters. Although not verified, I've been told that the Yaesu FT-847 does have separate buffered keying lines for each of its four antenna ports, making interface to this amplifier easy. Performance figures for the amplifier are listed in Table 2 and the completed amplifier is shown in Fig 16. Enjoy!

Acknowledgements

Thanks to Marv Gonsier, W6FR, and Roy Scanlon, NN7DX, for help in reviewing the text.

Notes

[1] C. M. Maer Jr, W0IC, "The Perseids Powerhouse," *QST*, Oct 1959, p 32.

[2] R. M. Richardson, W4UCH, "A Kilowatt Amplifier for 6 and 2 Meters," *QST*, June 1973, p 16.

[3] The surplus sockets have a built-in 0.005 µf grid/screen bypass capacitor. RF Parts has a limited supply of these sockets. Substitutes include Johnson #124-0311-110 (with a grid collet) and Eimac SK1900.

[4] G. R. Jessop, G6JP, *VHF/UHF Manual*, 4th edition, (Hertfordshire, England: Radio Society of Great Britain, 1983) Chapter 3.

[5] E. P. Tilton, W1HDQ, *The Radio Amateur's VHF Manual*, (Newington: ARRL, 1972) p 77.

[6] R. D. Straw, Ed., *The 1999 ARRL Handbook* (Newington: ARRL, 1998), pp 6-46 and 6-47.

[7] R. Myers, W1FBY, *The 1975 ARRL Handbook* (Newington: ARRL, 1974), p 46.

[8] F. M. Mims III, *The Forrest Mims Engineer's Notebook*, (San Diego, California: Hightext Publications Inc, 1992) p 107.

[9] *The 1999 ARRL Handbook*, pp 22-53 through 22-56.

[10] I. White, G3SEK, Editor, *The VHF/UHF DX Book*, Vol 1 First Edition (Buckingham, England: DIR Publishing, 1992) p 12-35.

[11] R. Schetgen, KU7G, Ed. *The 1993 ARRL Handbook*, (Newington: ARRL, 1992) p 31-37, Fig 95.

By Bob Larkin, W7PUA

From *QST*, June 2000

An 8-Watt, 2-Meter 'Brickette'

Put 20 mW into this little amplifier and get a 26-dB increase in transmitted power! Although the amplifier was designed initially for use with the DSP-10 transceiver, any 20-mW-output 2-meter exciter can enjoy the boost!

The DSP-10 transceiver[1] can be used with UHF/microwave transverters or as a very-QRP all-mode rig on 2-meters. The rig's 20-mW power output is sufficient for "barefoot" QSOs with locals, but that's hardly a big signal at any distance. Adding this little amplifier increases your fun quotient by raising the power level to 8 W, the high end of the QRP plateau. This brickette can also be used to drive an even higher-power, cascaded amplifier. Being a linear amplifier, it's suitable for use on all modes, including SSB, CW and FM.

The amplifier's front panel is quite simple, consisting of five monitoring LEDs. Knowing that everything is working correctly justifies the small amount of circuitry needed to operate these lights. The functions monitored include dc power on, transmit/receive status, power output, SWR and dc overvoltage. All amplifier control is done at the transceiver. Two RF cables connect to the transceiver output and receiver input, avoiding the need for an input relay in the amplifier.

This amplifier isn't limited to use with the DSP-10; any 2-meter transmitter that can deliver an output of 20 mW to drive the amplifier should work fine. In some cases, it might be desirable to use an input TR relay; there is sufficient room in the amplifier for adding one.

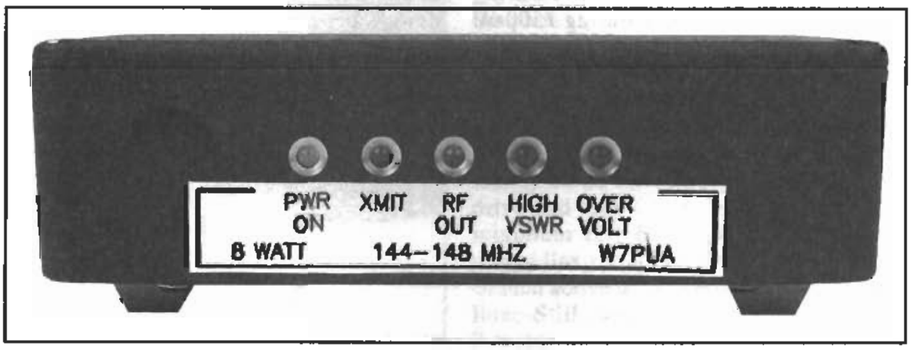

The five LEDs on the Brickette's front panel signal the amplifier's operating status. The power-output (**RF OUT**) and SWR (**HIGH VSWR**) LEDs vary in brightness depending on the forward and reflected power levels.

Circuitry

At the heart of this amplifier is an integrated power-amplifier module, manufactured by Mitsubishi. Such modules are used in many commercially manufactured transceivers and using one here makes amplifier construction and alignment simple. Within the module are two cascaded linear-amplifier transistors, along with their associated matching networks and biasing circuits for class AB operation. The module has 50-Ω input and output impedances. All we need to do is add a circuit to turn on the bias supply during transmit, install a low-pass filter for harmonic control and include an antenna-switching relay. To monitor amplifier operation, we tack on some simple circuitry.

Refer to the schematic in Figure 1. At the input pin of the RF module, U1, we apply 20 mW of drive. Blocking capacitors for the RF input and output ports are included within the module. Bypass capacitors on the three power leads are external, however. Ferrite beads, L6 to L8, prevent problems that might occur if RF gets on the power leads.

An **L** network consisting of L1 and C4 improves the impedance match to the module. Adding the network increases the output power by about 0.25 dB. Following the amplifier is a directional coupler (discussed later). Next is a five-pole low-pass filter. For simplicity, it is configured with the same coil and capacitor types as are used in the directional coupler. This filter attenuates the second and higher-order harmonics.

Separate connectors are available on the DSP-10 for the transmitter output and the receiver input. This simplifies adding the antenna relay for the amplifier since no switching is required at the amplifier input. A miniature relay is adequate at the amplifier output. Providing an isolation of about 30 dB, the relay, along with the PIN attenuator that is part of the DSP-10, provides plenty of protection for the receiver.

A lumped-element directional coupler, consisting of L2, L3 and C5 to C10, delivers power samples of the forward and reflected output signals. This directional coupler works quite well, providing a coupling of about −28 dB and a directivity[2] of 20 dB, but only over a narrow (12-MHz) bandwidth. For our application, this is adequate. Two diode detectors, built around D6 and D7, generate low-level dc signals that indicate the forward and reflected powers. These signals, in turn, are amplified by two sections of op amp U2. The resulting voltages are displayed on two LEDs, green for **RF OUT** (forward) power and red for **HIGH VSWR** (reflected) power. The LEDs serve as rough indicators of proper amplifier operation.

The **OVERVOLTAGE** indicator lights when the dc supply voltage exceeds 14.5 V, but does not automatically shut down the amplifier. The idea here is to supply a warning mechanism. When RF is not applied to the amplifier, it is quite resistant to supply overvoltage. As long as the supply overvoltage condition sets an alarm light, an observant operator will defer applying RF until the voltage is reduced.

Amplifier control is handled by the DSP-10. That transceiver has software-controlled relay sequencing, providing +5 V during transmit. Q2 and Q3 drive the antenna relay from this control. A ferrite bead, L11, and C24 keep RF at the antenna relay from getting back into the control circuits. D10 shunts the inductive kick from the deactivated relay coil. An LED, D2 (**XMIT**), across the relay displays the amplifier TR status.

Q4 provides an output that can control a follow-on amplifier. This output is an open-collector, ground-on-transmit type, compatible with many commercial and homebuilt amplifiers. For added driver-circuit protection, a reverse-voltage shunting diode, D11, is used.

Building the Amplifier

You could build the amplifier on a scrap of PC board. In fact, my first version is built that way (see the sidebar). For many builders, however, a PC board[3] is a more convenient way to assemble the project and that is shown in the photos.

This PC board is double-sided, with plated through holes. The backside is a solid ground plane allowing the board to be fastened directly to the aluminum enclosure. Such construction works very well for RF boards because a low-inductance ground path can be maintained throughout, reducing any interactions between the various circuits. However, this mounting method does not allow component leads to extend beyond the bottom surface of the board. I dealt with this primarily by using surface-mount parts[4] and by carefully bending the component leads so they behave like surface-mount parts.

All of the chip resistors, chip capacitors, diodes, surface-mount ICs and ferrite-bead chips are soldered to their board pads conventionally. Install the amplifier module *after* the board is mounted in the box. The antenna relay, fuse clips and the variable resistor all have their leads bent away from their host. Avoid making any bends close to the component body; make the bends at a point about 1/16 inch away from the component body. This approach eliminates mechanical stress on the lead attachment. These components end up about 1/16 inch above the board's top surface after they are soldered in place.

Figure 2 shows the construction of the four toroidal inductors, L2 through L5. Experience shows the inductor values are quite repeatable if the turns are always distributed in the same manner around the core. As the turns are pushed closer together, the inductance increases considerably. The matching-network coil, L1, is noncritical and it should be wound as shown in the part list and Figure 2; it should need little, if any, adjustment.

The LEDs all have long leads and are soldered to the board after bending the lead ends by about 1/8 inch. Be sure to keep track of the LEDs' longer (anode) leads: Those leads connect to the current-limiting resistors.

Putting the Board in a Box

Once the PC board is assembled, use it as a template for marking the hole locations in the enclosure, a standard Hammond die-cast box. Mount the board flat against the box bottom using #4-40 hardware. Because there will likely be some mold marks and box-identification letters where the board and RF module lie, make the enclosure's inner-bottom surface reasonably flat; you can do this with 60-grit sandpaper. Bend the leads of the five LEDs to apply a slight forward pressure on the lights as they slide into the holes in the box front. That holds the LEDs in alignment without needing adhesive.

Take care when tightening the PA module mounting screws to be sure that no pressure is applied to the ends of the module cover. The leads from the module need some trimming. They are above the board and need forming to get to the board level for soldering. Don't apply pressure at the edge of the module cover when doing this. You may need to hold each lead with needle-nose pliers to keep from damaging the case. It should be possible for the leads to reach the board surface within 1/8-inch of the cover.

Three short pieces of 50-Ω coax attach the board to J1, J2 and J3. Solder lugs under the jacks provide for ground connections at one end and PC-board pads take care of the other end of the coax.

Three leads run between connectors J4, J5 and J6 and the PC board. Each of these leads has a bypass capacitor (C28, C6 and C27, respectively) at the connector. Short, low-inductance leads are important on these capacitors. To help keep the RF signals inside the box, a small inductor wound on a ferrite core (L14, L15 and L16) is placed on each wire. Position these inductors close to their connectors. If #22 or #24 stranded hook-up wire with thin insulation is used, it's possible to wind the coils with the hook-up wire.

Turning on the Amplifier

Now, to see it work! First, connect a 50-Ω noninductive dummy load to the amplifier output.[5] If you have a variable-voltage power supply to use for initial testing, slowly raise the dc supply voltage from 0 V to 13.8 V. Otherwise, you'll need to rely on the power-supply fuse as protection from any serious construction errors when apply-

Did the Amplifier Always Look Finished?

No, it didn't. Some people may be able to put a finished amplifier on a PC board and have it work fine. I don't seem to be able to do that!

My first step in designing the amplifier was to search for suitable RF modules, mostly via the Internet. Once I selected the M57732L as having suitable power, gain and dc operating voltage, I drew a schematic in my notebook. Originally, the circuit had automatic shutdown for high SWR and full voltage regulation to deal with overvoltage. This looked too complex. It was necessary to cut off the drive to the amplifier, and the regulator had to be a low-dropout type (for these, the difference in voltage between the input and the output need be only a fraction of a volt). I couldn't find an integrated regulator that met my requirements and the thought of building one using op amps and transistors used more parts than I could justify.

I opted for simplification, putting alarm lights on for SWR and overvoltage and turned the automatic part over to the operator! A new schematic resulted, not unlike that of Figure 1. I went to the *ARRL Radio Designer*† and simulated the directional coupler and low-pass filter. A little playing around with the simulation showed that a single inductor value could be used for both subcircuits.

Next, I built the "breadboard" version using scraps of PC board. Testing showed that almost everything worked as expected. The **overvoltage** light did not have snap action, though. That was caused by using a resistor where the ferrite bead, L13, is now. I ran the input voltage temporarily up to 16 V and everything continued to function. Next, I left the amplifier running for an hour and nothing overheated. I checked the intermodulation products and harmonics and found them to be at satisfactory levels.

It was now time to layout the PC board. The experience of putting the breadboard together allowed a smarter final layout. The first version was on two boards and some of the connectors ended up on the front panel. This was all worked out for the final design.
—Bob Larkin, W7PUA

†ARRL order no. 6796. ARRL publications are available from your local dealer, or directly from the ARRL. See the ARRL Bookcase elsewhere in this issue, or check out the ARRL Web site at: **http://www.arrl.org/catalog/**.

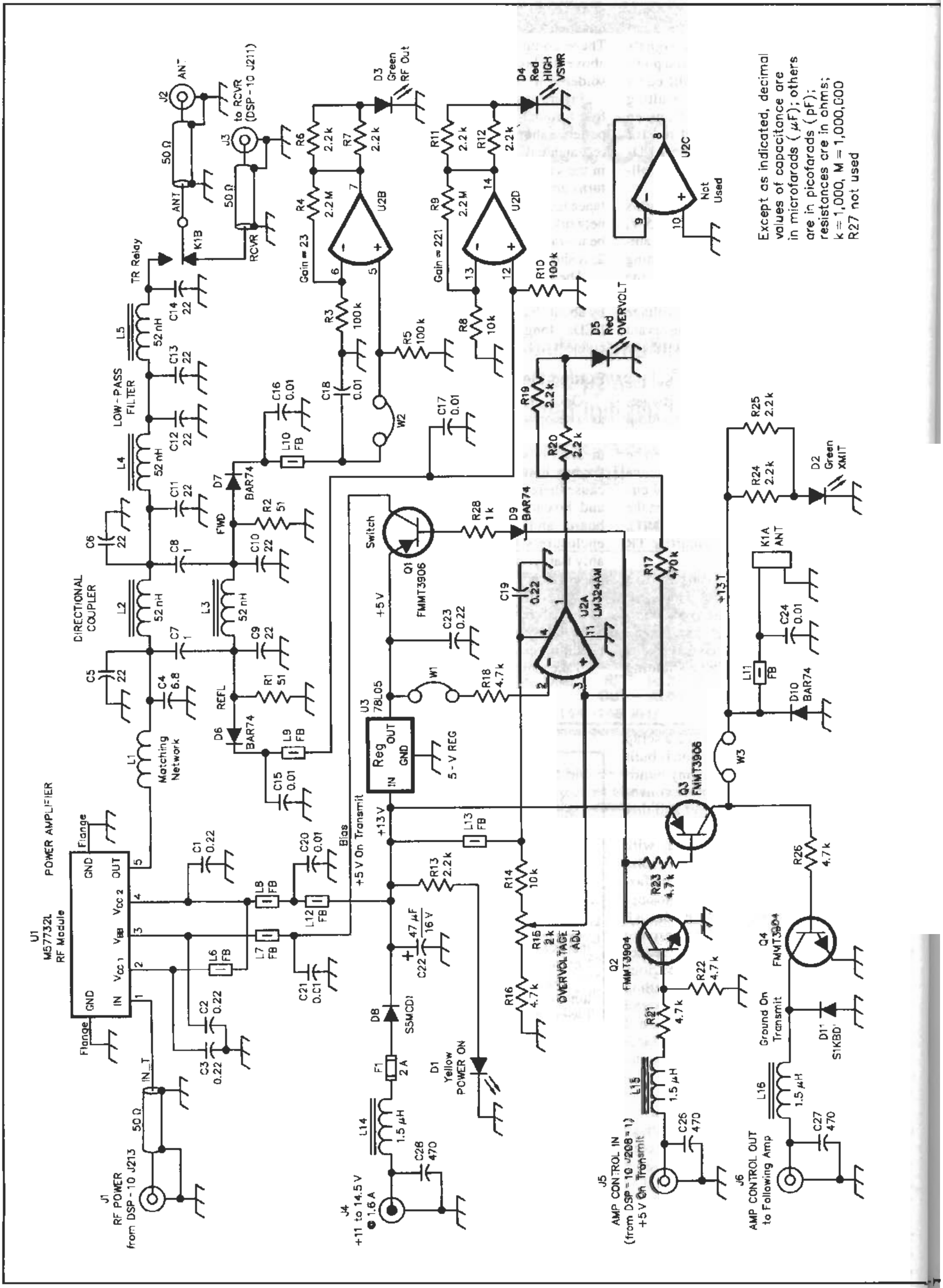
J1
RF POWER
from DSP-10 J213
50 Ω
IN_T
U1
M57732L
RF Module
POWER AMPLIFIER
GND
IN
Vcc 1
VBB
Vcc 2
OUT
GND
Flange
C1 0.22
C2 0.22
C3 0.22
L6 FB
L7 FB
L8 FB
L12 FB
C20 0.01
C21 0.01
Bias
+5 V On Transmit
+13 V
L13 FB
R13 2.2 k
47 µF 16 V
C22
D8 S5MCDI
F1 2 A
L14 1.5 µH
C28 470
J4
+11 to 14.5 V @ 1.6 A
D1
Yellow
POWER ON
L1
Matching Network
C4 6.8
C5 22
DIRECTIONAL COUPLER
L2 52 nH
L3 52 nH
C7 1
C9 22
C8 1
C10 22
C6 22
REFL
FWD
R1 51
R2 51
D6 BAR74
D7 BAR74
L9 FB
L10 FB
C15 0.01
C16 0.01
C11 22
L4 52 nH
C12 22
LOW-PASS FILTER
C13 22
L5 52 nH
C14 22
TR Relay
K1B
ANT
RCVR
50 Ω
J2 ANT
J3
to RCVR
(DSP-10 J211)
U3 78L05
Reg
IN
OUT
GND
5-V REG
W1
R18 4.7 k
+5 V
C23 0.22
Q1 FMMT3906
Switch
R28 1 k
D9 BAR74
C17 0.01
W2
C18 0.01
R3 100 k
R5 100 k
Gain = 23
R4 2.2 M
U2B
R6 2.2 k
R7 2.2 k
D3
Green
RF Out
R8 10 k
Gain = 221
R9 2.2 M
U2D
R10 100 k
R11 2.2 k
R12 2.2 k
D4
Red
HIGH
VSWR
D5
Red
OVERVOLT
R19 2.2 k
R20 2.2 k
C19 0.22
U2A
LM324AM
R17 470 k
R14 10 k
R15 2 k
OVERVOLTAGE ADJ
R16 4.7 k
U2C
Not Used
+13 T
K1A ANT
C24 0.01
L11 FB
D10 BAR74
W3
Q3 FMMT3906
R23 4.7 k
R24 2.2 k
R25 2.2 k
D2
Green
XMIT
R26 4.7 k
Q2 FMMT3904
R22 4.7 k
R21 4.7 k
L15 1.5 µH
C26 470
J5
AMP CONTROL IN
(from DSP-10 J208-1)
+5 V On Transmit
Q4 FMMT3904
Ground On Transmit
D11 S1KBDI
L16 1.5 µH
C27 470
J6
AMP CONTROL OUT
to Following Amp

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000, M = 1,000,000
R27 not used

Figure 1—Schematic of the 8-W amplifier. All resistors are 5% 1206 chips. These are available in small quantities from Mouser Electronics (Xicon). Unless otherwise noted all capacitors are either 1206 or 0805 chips (either size fits on the board). Capacitors less than 470 pF are NP0, while values from 470 pF and up are any general-purpose ceramic, such as X7R or Z5U. Component sources listed here are generally only a few of several that manufacture equivalent parts. If the amplifier is built without a pre-made PC board, change the chip components to leaded types. Component designations for the LEDs differ from *QST* style. Parts used in the amplifier are available from one or more of the following sources. Source abbreviations used in the parts list precede the company name: (DK) Digi-Key Corp, 701 Brooks Ave S, Thief River Falls, MN 56701; tel 800-344-4539, 218-681-6674; **http://www.digikey.com**; (ME) Mouser Electronics, 958 N Main, Mansfield, TX 76063; tel 800-346-6873, 817-483-4422; **http://www.mouser.com**; (RS) RadioShack—see your local distributor; **http://www.radioshack.com**; (RP) RF Parts, 435 South Pacific St, San Marcos, CA 92069; tel 800-737-2787, 760-744-0700; **http://www.rfparts.com**.

C22—47 µF, 16 V surface-mount electrolytic (DK PCE3033CT)
C26, C27, C28—470 pF, 50 V, leaded ceramic (ME 140-50P2-471K)
D1—Yellow T1 LED (DK160-1079)
D2, D3—Green T1 LED (DK 160-1080)
D4, D5—Red T1 LED (DK 160-1078)
D6, D7, D9, D10—BAR74 diode (DK BAR74ZXCT)
D8—5-A, 50-V power rectifier, SMC package (DK S5ACDICT)
D11—1-A, 100-V power rectifier, SMB package (DK S1BBDICT)
K1—SPDT 12-V dc miniature relay, Omron G5V-1-DC12 (DK Z774)
L1—2t, #20 or 22, ⅛-inch ID; see Figure 3.
L2-L5—52 nH; 5 turns, #26 enameled wire on a T-25-17 toroid; see Figure 3.
L6, L7, L9, L10, L11, L13—Ferrite SMT bead, 1206, 600 Ω at 100 MHz, Stewart HZ1206B601R (DK 240-1019-1)
L8, L12—Ferrite SMT bead, 3 A, 1206, 100 Ω at 100 MHz, Stewart HI1206N101R (DK 240-1008-1)
L14-L16—1.5 µH, 5 turns #22 or #24 enameled wire on an FT-23-43 core; see text.
F1—2-A, 5×20-mm fuse (DK F948) with two clips (DK F058)
J1, J5, J6—Phono jacks (RS 274-346)
2, J3—BNC jack (RS 278-105)
4—5.5-mm OD, 2.1-mm ID power connector (RS 274-1563)
Q1, Q3—FMMT3906 PNP transistor, SOT23 (DK FMMT3906CT)
Q2, Q4—FMMT3904 NPN transistor, SOT23 (DK FMMT3904CT)
R15—2 kΩ adjustable resistor, Bourns 3329H-1-202 (DK3329H-202)
1—RF amplifier module, Mitsubishi M57732L, (RP M57732L)
2—LM324AM dual op-amp (DK LM324AM)
3—78L05 5-V positive regulator, SO-8 package (DK LM78L05ACM)
isc: PC board, enclosure 3.7×4.7 ×1.3 inches, Hammond 1590BB (DK HM152), hardware.

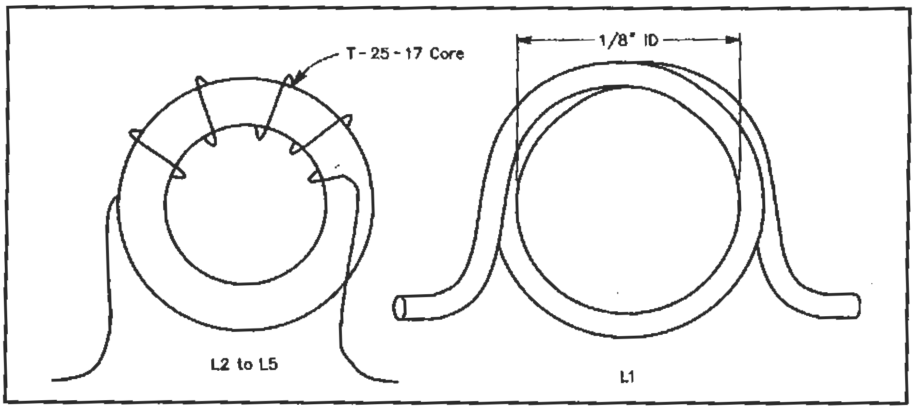

Figure 2—Coil details. Arrange the turns on the four toroids as shown here. L2 and L3 might need to have their turns spacing adjusted during tune-up.

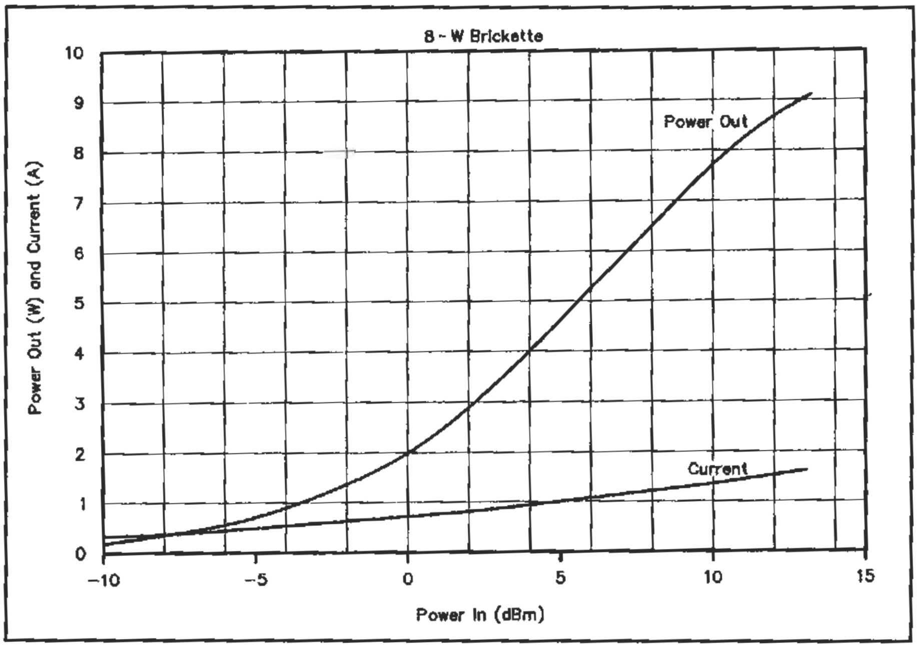

Figure 3—Measured power output and dc-supply current as a function of drive power. Notice that the output-power scale is not in decibels. This makes the output appear less compressed than it really is. The 1-dB compression point is at about 4 W.

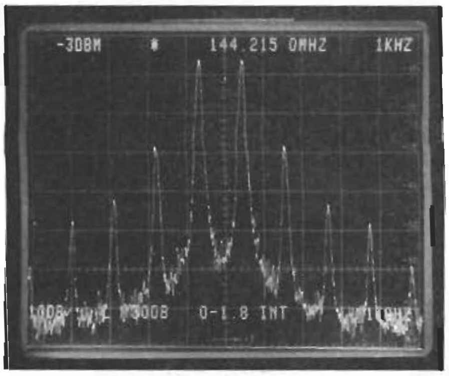

Figure 4—Measured intermodulation distortion with a PEP output of 8 W. Third-order products are down about 28 dB from peak power, fairly typical of this amplifier type. Perhaps more important is that the fifth and higher-order products continue to drop off as the order increases. These higher-order products are farther from the operating frequency and thus generally more disruptive to nearby stations. The current drawn during this two-tone test was 670 mA.

ing the full 13.8 V. At full voltage, the current drawn by the amplifier during receive should be about 10 mA. Next, set the **OVERVOLTAGE** alarm. Adjust the voltage at the wiper terminal of R15 to 4.68 V.[6] If you can vary the dc supply voltage, the **OVERVOLTAGE** LED should light at about 14.5 V.

When making the following adjustments, it is necessary to adjust the turns spacing on the toroidal inductors. For safety, *do this only when no RF drive is applied.*

Connect the TR control to the DSP-10 and connect J1, the amplifier's RF input, to the DSP-10 antenna connector. In CW mode, increase the DSP-10's RF power slowly until the **RF OUT** LED (D3) glows. Then adjust the turns spacing on L2 and L3 to extinguish the **HIGH VSWR** light (D4). This indicator is about 10 times more sensitive than the **RF OUT** LED, but it should be possible to extinguish the **HIGH VSWR** LED completely. Continue this adjustment process while increasing the drive power. You can measure the voltage at the output of the SWR op amp

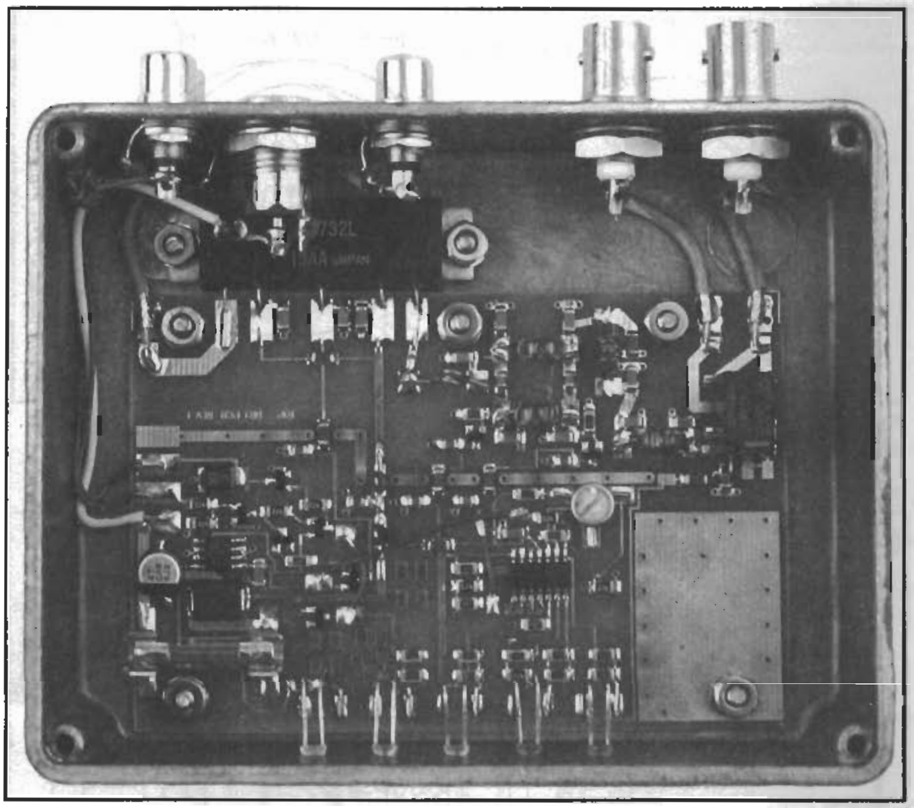

Pretty outside and pretty inside, the Brickette's construction reflects a caring hand. At the top-left rear the RF-amplifier module can be seen fastened to the base of the enclosure, which serves as a heat sink for the module. The relay is at the upper-right and a kludge area is visible at the lower right behind the front panel.

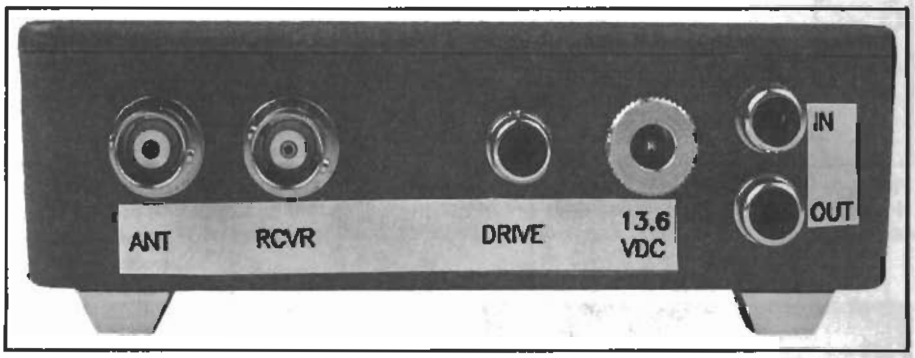

This rear-panel view of the 8-W Brickette shows compact but uncrowded I/O connectors clearly labeled.

(U2 pin 14) and continue to minimize this voltage, although the **VSWR** LED is not lit.

Optimize the amplifier matching by measuring the power output as indicated by the dc voltage at U2 pin 7 and adjusting L1's turn spacing (and perhaps changing the value of C4) for maximum output at high drive. Don't expect major changes with these adjustments as the amplifier is inherently quite well matched.

While in transmit, check that the bias voltage at pin 3 of U1 is about 4.9 V and that the idling current with no RF drive applied is about 140 mA.

Performance

Figure 3 shows the power output and current level for various drive levels. The CW output is over 8 W for the two devices that I tested. When tested with a two-tone input signal and an output of 8 W PEP, the IMD level is down 28 dB for third-order products and 43 dB for fifth-order products (see Figure 4). The strongest harmonic is the second (at about 292 MHz) and this is 65 dB down from peak output, more than enough to meet FCC 2002 requirements.[7]

Concluding Thoughts

Several areas on the PC board have either a ground plane section or pads for mounting transistors, resistors and other components. These areas are available to you to use for modifications or additions to the amplifier (perhaps you have an application that needs an input TR relay or you need a different control circuit).

Now you're ready to get on the air with a medium-sized signal. You can increase power even further by cascading another amplifier. Or, use some of the signal processing in the DSP-10 to work deeper into the noise. QRP power levels have pushed the idea of "working smarter instead of harder."

Notes

[1] Bob Larkin, W7PUA, "The DSP-10: An All-Mode 2-Meter Transceiver Using a DSP IF and PC-Controlled Front Panel," *QST*, —*Part 1*, Sept 1999, pp 33-41, —*Part 2*, Oct 1999, 34-40, —*Part 3*, Nov 1999, pp 42-45. Additional information on that project is available on the author's Web site, **http://www.proaxis.com/~boblark/dsp10.htm**. Any future information about this amplifier project will be placed at this Web site.

[2] The directivity of a coupler is the difference, in decibels, between the forward signal and the reflected signal, when the coupler is properly terminated in its designed load, in this case, 50 Ω. Couplers constructed from transmission lines have high directivity over a very wide frequency range. These lumped-element couplers are restricted in the frequency range where their isolation is high, but offer simplicity instead.

[3] Gerber files for making the PC board can be obtained from the author. Alternatively, unpopulated PC boards are available from Mashell Electric, PO Box 5, Eatonville, WA 98328. Price: $20 each in the US, $21 in Canada (air mail) and $22.50 (air mail) elsewhere. These boards have plated through holes, a solder mask on the component side and a silk-screened legend. Check their Web site **http://members.aol.com/w7slb/w7slb.htm** for details.

[4] See Sam Ulbing, N4UAU, "Surface Mount Technology—You Can Work with It!", *QST*—*Part 1*, Apr 1999, pp 33-39; —*Part 2*, May 1999, pp 48-50; —*Part 3*, Jun 1999, pp 34-36; —*Part 4*, pp 38-41. Additional comments from Avery Davis, WB4RTP, can be found in the *QST* Technical Correspondence, Feb 2000, p 70. Surface-mount techniques are used for the DSP-10 transceiver. Some people were initially concerned about dealing with the tiny components, but after constructing the board, most builders felt that it went well.

[5] A satisfactory dummy load for 144 MHz can be constructed from four 200-Ω, 3-W metal-oxide resistors. Arrange them in parallel around a coax connector. Short leads are important. *Do not use wire-wound resistors.*

[6] The **OVERVOLTAGE** alarm triggers when the dc supply voltage is greater than 14.5 V. If an adjustable dc supply is available, you can set this directly. Alternatively, set the voltage at the wiper of R15 to 0.360×(V – 0.6), where V is the dc supply voltage.

[7] Larry Price, W4RA, and Paul Rinaldo, W4RI, "WARC97, An Amateur Radio Perspective," *QST*, Feb 1998, pp 31-34. By these rules, an 8-W 2-meter transmitter must suppress the harmonics by at least 52 dB. Greater amounts of suppression are required for higher power levels.

By Dave Mascaro, WA3JUF From *QST*, June 1990

903-MHz Linear Amplifiers

Part 1—Looking for a gain block or two for your 903-MHz station? Here are *eleven* of them to cover just about any need you may have, from a simple receiving preamplifier to one with 23 dB gain and over 4 W output!

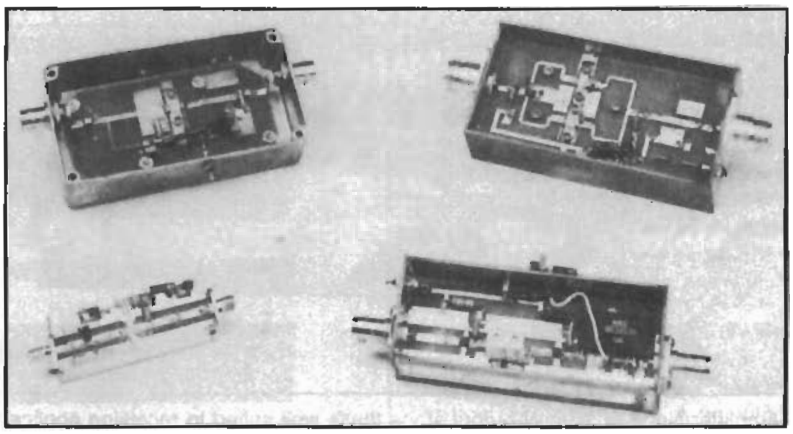

The 33-cm Amateur Radio band (902-928 MHz) is becoming well populated in many areas of the country. Propagation on 33 cm has traits similar to both 432 MHz and 1296 MHz, but has characteristics all its own at times, making it a very interesting band. It's pleasantly surprising what you can do with a few watts on 903 MHz with high-gain loop-Yagi antennas.

Articles have been published on transverters, amplifiers and receivers for 33 cm. I described a 759-MHz local oscillator[1] intended to be used with a 144-MHz IF for operation on 903 MHz. If you've built the transmit section of a transverter that gets you to the 10-mW level, or if you've purchased a commercial transverter such as Down East Microwave's low-power, no-tune unit,[2] you'll need to bring your transmitter power up to a usable level. Instead of showing a typical transmit-amplifier chain, many of which have been described in Amateur Radio literature, I will discuss eleven different gain blocks, for several different gains and power levels, in this two-part series.

I suggest that you build each amplifier in its own enclosure, rather than trying to build more than one amplifier in one box to eliminate connectors or reduce size. If you are building your own equipment, you probably aren't too concerned about size and compactness, and having separate enclosures makes for much easier tuning and troubleshooting. Many transistors used in 1296-MHz projects are suitable for use at 903 MHz; some even have more gain at 903.

Transistors with SD prefixes used in the following designs are manufactured by SGS-Thomson Microelectronics.[3] Some of these transistors don't operate at 12 V. Again, in building your own equipment, this shouldn't be a major problem. Gains listed are averages for several different devices tried. All of the amplifiers can be driven harder than indicated—for CW or FM operation only—but should be used at or below their rated power output for SSB work. All the designs can be used at lower power output; they are all linear amplifiers.

Construction

All of the amplifiers are built using similar techniques. Each design uses common $^1/_{16}$-inch-thick, G-10, double-sided, fiberglass-epoxy PC board for the microstrip circuitry. The ground-plane side of each PC board is unetched.

The microstrip boards can be made using what I call the "X-ACTO®-etch" method. This involves using a piece of clear tape as the resist. (Four-inch-wide clear tape is available at stationery stores.) After drawing the pattern on the board with a pencil or fine-tipped marker, cover it on both sides with clear tape and cut the pattern in the tape with an X-ACTO knife. Then, remove the tape from the areas to be etched and etch the boards in ferric chloride in a crock pot on low heat. Do this in a well-ventilated area! Etching takes about half an hour without agitation. The crock pot is a no-mess way to etch boards. Only an inch or so of ferric-chloride solution is needed. Of course, you can also use a photographic method to make boards, as I do for multiple or complicated boards.

After the PC board is etched, clean it with steel wool and drill holes at all dc- and RF-ground points using a no. 50 drill. No holes are needed for component leads, as all components are mounted on the microstrip side of each PC board. RF grounds must be located as close to the areas to be grounded as possible, to ensure low-inductance ground paths. There are at least two ways to do this. One is to install a rivet in each hole, flare it with an awl or ice pick, then flatten it by tapping with a small

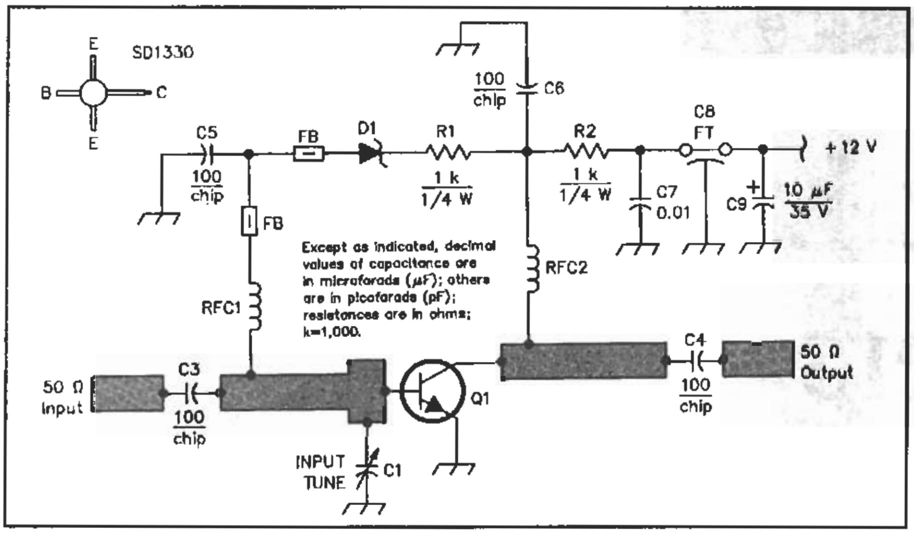
E
SD1330
B
C
E
C5
100
chip
FB
D1
R1
1 k
1/4 W
100
chip
C6
R2
1 k
1/4 W
C8
FT
+12 V
C7
0.01
C9
10 µF
35 V
FB
RFC1
RFC2
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000.
C4
100
chip
50 Ω
Output
C3
100
chip
50 Ω
Input
Q1
INPUT
TUNE
C1

Fig 1—Schematic of the SD1330 receive preamplifier/low-level transmit stage. RFC1 and RFC2 consist of 8 turns of no. 24 enam wire closewound on a 0.1-inch-ID form (such as a 0.1-inch drill). D1 is a 9.1-V, ½-W Zener. C1, a 12.5-pF trimmer, is available from Mouser (part no. 24AA071).

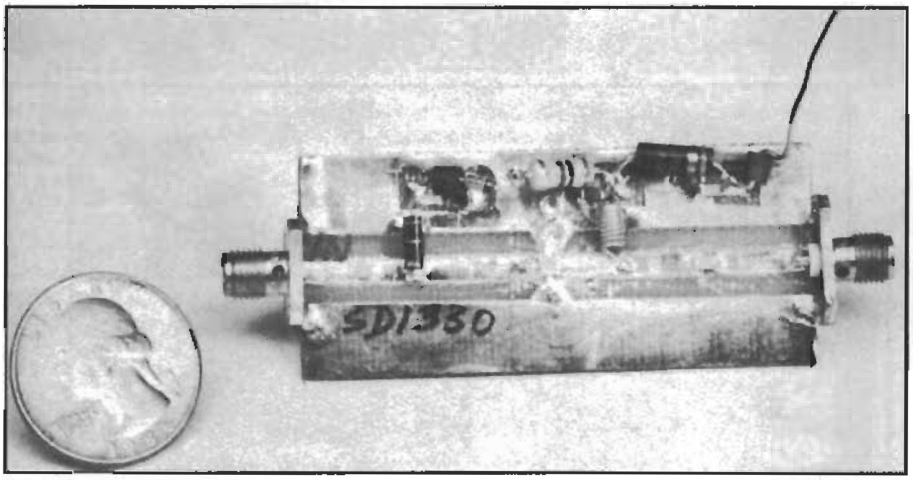

Fig 2—Amplifier No. 1, a 10-mW-output stage that's well suited to receiving applications, in prototype form.

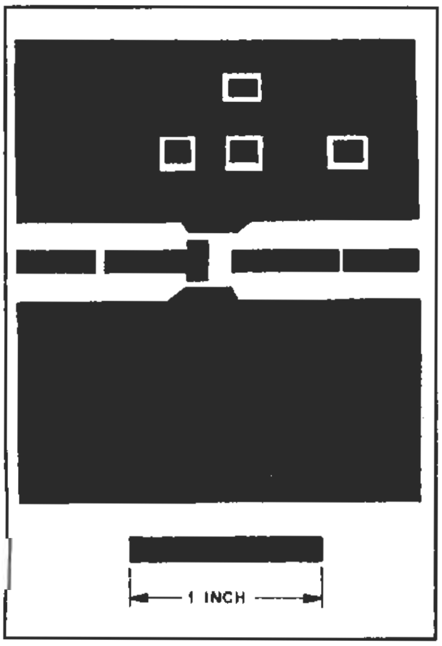

Fig 3—Full-scale PC-board artwork for the 10-mW SD1330 amplifier. Black areas represent unetched copperfoil. All parts are mounted on the foil side of the board. Crop the finished boards as necessary.

hammer on an anvil or other hard surface. Then, solder the rivets on both sides of the PC board. Alternatively, pieces of bus wire can be used for grounding.

Because parts placement is quite straightforward, I haven't provided parts-placement drawings for all of the designs. The schematics and PC-board artwork should provide all the guidance you'll need to assemble the boards. I can help with any questions that you might have when building any of the amplifiers.

All of the amplifiers described, except those operating under the 300-mW level, require some heat sinking. A small brass sheet is sufficient for the studded parts, but a finned heat sink an inch or two square is needed for the flanged-device amplifiers. Thermally conductive compound is required between the devices and their heat sinks. For studless devices, a small piece of brass shim stock soldered from the bottom of the device to the ground-plane side of the PC board should work fine.

An easy way to enclose an amplifier like those described is to make a housing using double-sided PC board or brass sheet for the four walls and bottom cover. The amplifier PC board makes the top cover. Input and output connectors (N, BNC or SMA) can be bolted or soldered to the end walls with their center pins soldered directly to the microstrip input/output lines in an end-launch configuration. When using PC-board material for housings, mount the side walls in such a way that the grounding is continuous from the connectors to the ground-plane side of the amplifier board.

Various die-cast aluminum boxes, such as Bud CU-123 and CU-124 and the Hammond 1590 series, also work great for enclosing these designs. Die-cast boxes usually provide sufficient heat sinking for amplifiers operating at less than 2 W output. Device studs or flanges should be attached directly to the boxes in these cases. All component leads must be kept as short as possible. This also applies to trimmer caps. Mount them flush to the PC board so that they act as capacitors—not as inductor/capacitor combinations. Use multiple rivets (or bus wires) to ground the trimmers.

Components

Use ceramic chip capacitors for the dc blocks and high-frequency bypassing. Some of the larger-value bypass capacitors, like 0.001 µF and 0.01 µF, are available in chip form also. These epoxy types are good for all the designs described here. I use inexpensive chip caps from Mouser Electronics.[4] Chip caps and Johanson piston trimmers are available from Microwave Components of Michigan.[5] Transistors are available from RF Parts[6] and RF Gain, Ltd.[7] Components, rivets and PC boards are available from Frontier Microwave.[8]

Tune-Up

In most of these amplifiers, the quiescent collector current I_{cq}) initially will have to be checked and set with no drive applied.

To check bias current, disconnect the cold end of the collector choke and insert a milliammeter in series with the choke. Adjust I_{cq} by changing the value of the collector-bias resistor or by changing the collector voltage. All amplifiers requiring 12 V can be powered directly from low-current, 12-V dc sources (a three-terminal, 12-V regulator, for example). Amplifiers running on 14-18 V dc or 21-23 V dc can be powered by an LM317T (or LM317K) adjustable regulator. Set the regulated voltage to a minimum, and adjust it upward from minimum while monitoring I_{cq}. Idling collector current can be adjusted to suit your gain and power-output requirements.

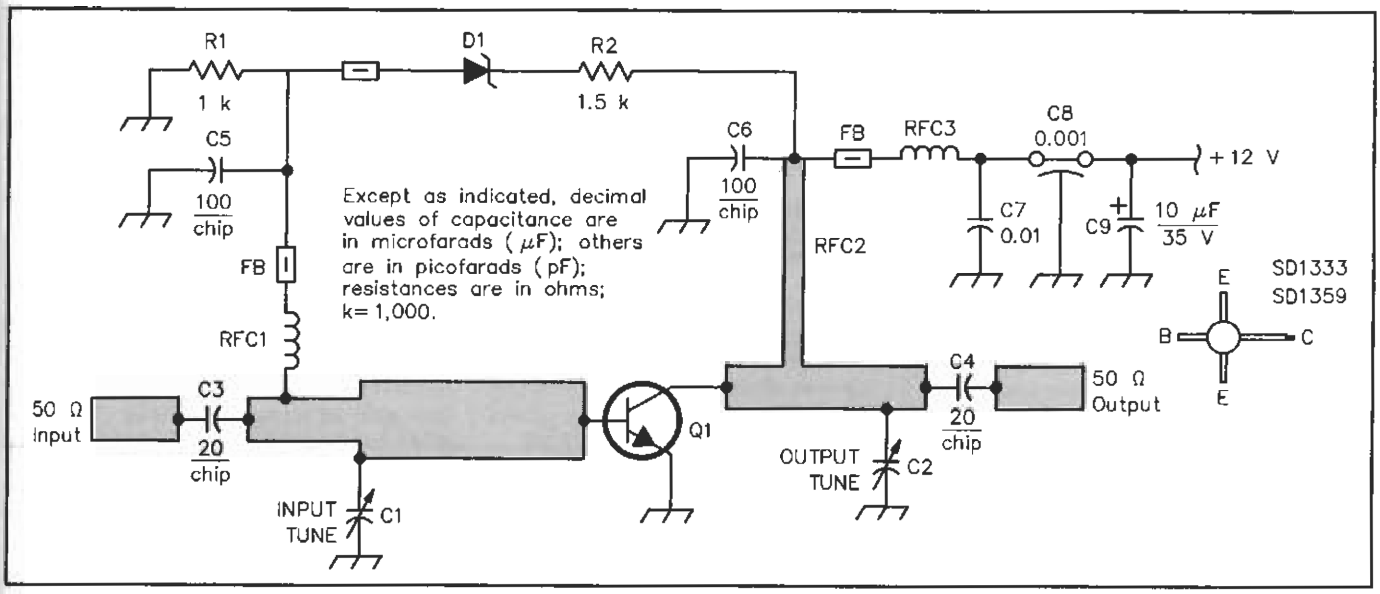

Fig 4—Schematic of the SD1333 and SD1359 amplifiers (Amplifiers No. 2 and No. 3). All components except Q1 are common to both designs (see text).

C1, C2—12.5-pF trimmer, Mouser no. 24AA071.

C7—0.01-μF disc ceramic.

D1—9.1-V, ½-W Zener.

RFC1, RFC3—8 turns of no. 24 enam wire, 0.1 inch ID, closewound.

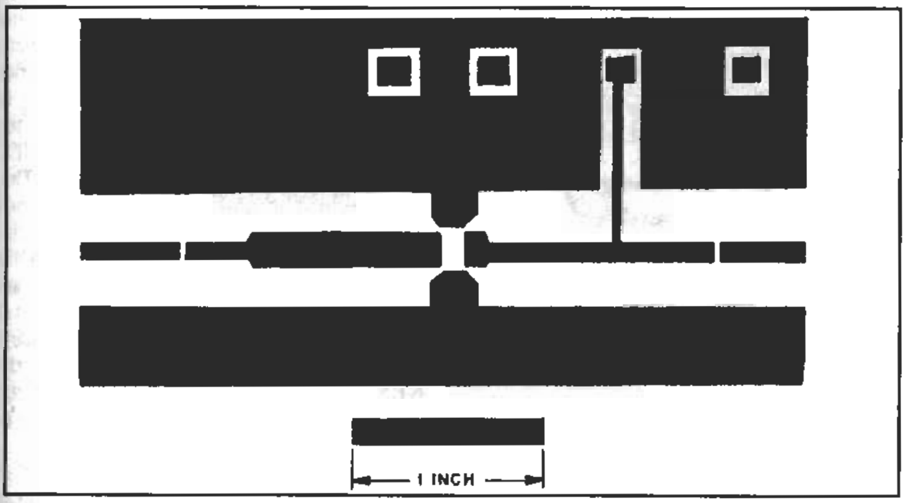

Fig 5—Full-scale PC-board network for the 100- and 250-mW amplifiers. Black areas represent unetched copper foil. All parts are mounted on the foil side of the board.

Amplifier No. 1: A Receive Preamplifier or 10-mW-Output, 13-dB-Gain Transmit Stage

The amplifier shown in Figs 1 and 2 is a low-power stage using an SD1330, Motorola MRF901 or NEC NE64535. This stage is best suited to the receiving side of a transverter. With a noise figure (NF) of 2 dB, it makes a fine front end by itself. With a good, low-noise GaAsFET preamplifier in front of it, you'll have all the sensitivity you need in front of the receive mixer. For transmitting applications, where a low noise figure isn't important, an MMIC amplifier would be better because it is a lot easier to build for the same results.

The SD1330 amplifier is built on 0-10 double sided board; the artwork appears in Fig 3. After etching the board, install rivets at the RF and dc grounds. To mount the device, drill a hole in the PC board the size of Q1's macro-X case. This allows all four leads to be soldered to the microstrip without bending them.

The idling current for this amplifier should be set to 5 mA or less for good noise figure. The idling current is kept constant for different transistors and over different voltages and temperatures by current-limiting resistors Rl and R2 and Zener diode Dl. Tune the input trimmer for best NF or maximum gain if an NF-measurement setup isn't available.

Amplifier No. 2: 10 mW In, 100 mW Out

The SD1333 transistor, a macro-X plastic-packaged device used in this design, delivers 100-125 mW. This device is capable of a reasonable noise figure (2-3 dB) and good dynamic range, which allows it to be used as a second stage (following a GaAsFET preamplifier) in a receiving system. (Motorola's BFR96/MRF961/MRF962 also work well in this circuit; the BFR96 has slightly less gain.) Fig 4 shows the amplifier schematic and Fig 5 shows the full-scale artwork.

Zener-diode bias is used for simplicity and some temperature compensation. Set the quiescent current to 40-50 mA (not critical) by varying R1. A pot can be used initially, then the pot can be replaced by a fixed-value resistor. Or the collector voltage can be varied slightly.

After the rivets are installed in the PC board, drill a hole the diameter of Q1's molded package in the PC board. Mount the device in the board so its unbent leads are soldered directly to the microstrip board. Use the small variable capacitors to tune for maximum gain.

Amplifier No. 3: 9 dB Gain, 250 mW Out

This unit is a slightly higher-power version of the previous amplifier, using an SD1359 or Motorola TRF559 plastic-packaged device. The artwork and schematic are identical to that of Amplifier No. 2. Idling current, 40-50 mA, can be optimized by adjusting the collector supply between 10 and 13 V to set the stage gain.

After the PC-board rivets are installed, drill a hole in the PC board to accept Ql. Mount the device in the board so its unbent leads are soldered directly to the microstrip board. Tuning is as discussed with Amplifier No. 2.

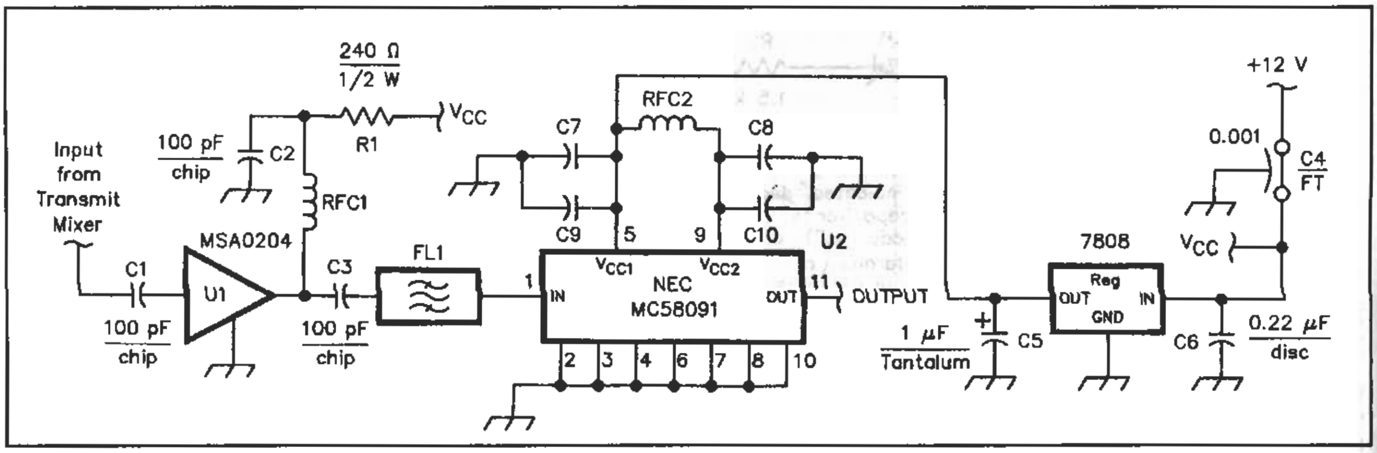

Fig 6—Schematic of the MSA 0204/MC5809L amplifier.

C7, C9—100-pF chip.
C8, C10—0.01-µF ceramic disc.
FL1—Two-stage Toko helical filter (Digi-Key no. TK23318).
RFC1—8 turns of no. 24 enam wire, 0.1 inch ID, closewound.
RFC2—1-µH molded choke.

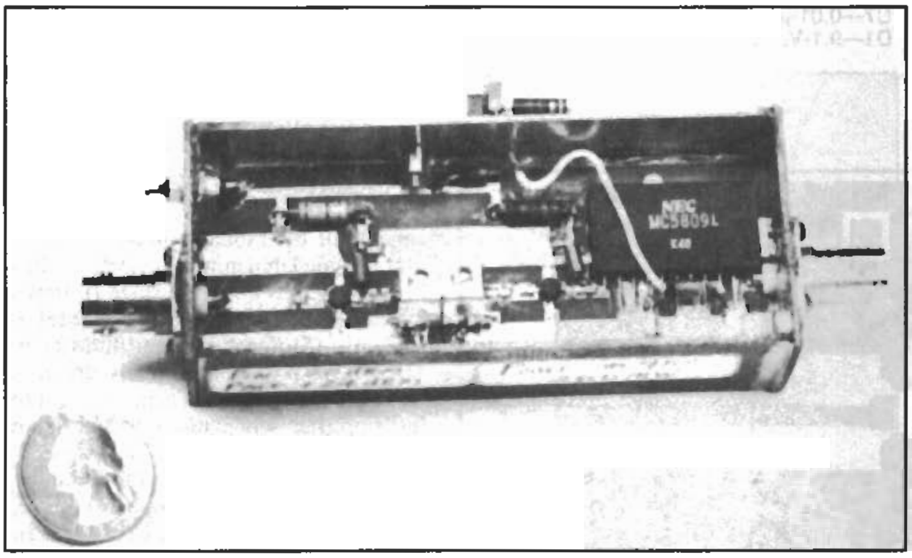

Fig 7—The 200-mW-output MSA 0204/MC5809L amplifier. In this prototype, I used MSA 0204 stages before and after the helical filter to provide addiional gain. The biasing components (the 7808 and associated parts) are mounted on the outside of the enclosure for convenience.

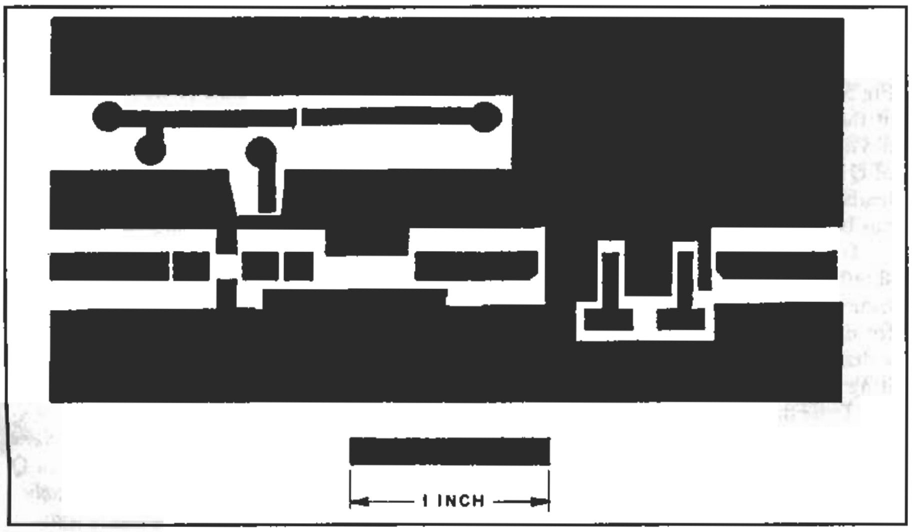

Fig 8—Full-scale PC-board artwork for the 200-mW amplifier. Black areas represent unetched copper foil.

Amplifier No.4:0.8 mW In, 200 mW Out

The amplifier shown in Figs 6 and 7 is a high-gain unit that can be used to follow the output of your 903-MHz transmit mixer. An Avantek MSA 0204 or Mini-Circuits MAR-2 MMIC is used for the first stage, followed by a Toko two-stage helical filter.[10] (In the prototype amplifier shown in Fig 7, I used two MMIC stages, in addition to the helical filter, before the final stage.) The output stage is an NEC MC5809L thick-film hybrid module. This NEC hybrid module is one of four such units intended for hand-held cellular telephones operating in the 800- to 960-MHz range. The MC5809-series amplifiers are rated at 150 mW minimum power output at 7.5 V dc. Using a more readily available power source (a 7808 regulator), the NE5809s I tried were linear at over 200 mW output. These hybrid amps are easy to use and are stable into any load.

The full-scale PC-board artwork is shown in Fig 8. The helical filter is used between stages to clean up the transmitter signal by filtering out the local-oscillator and image frequencies, but isn't required for applications where such filtering is done in other stages. A three-stage Toko filter[11] will also fit on the board, and could be used for better filtering. The filter leads are bent to the side to allow soldering to the PCboard traces. See Fig 9 for parts placement.

This amplifier is easy to build and get working. Because both active stages and the filter are designed for 50 Ω in and out, the amplifier uses 50-Ω microstrip throughout. No trimmer capacitors are needed and no bias adjustments are necessary—just apply the dc voltages and the drive signal, and peak the helical filter for maximum power output at 903 MHz!

Because more gain is available from this design than I needed, I tried mixing two signals at the input of the first stage. Using a T connection at the input of the first stage, I used the circuit to mix a 759-MHz local oscillator with a 144-MHz IF signal. The 903-MHz output signal was about 175 mW and was fairly clean. Using MMIC amplifiers for gain blocks, almost anything can be used as a mixer!

Amplifier No. 5: 35 mW In, 350 mW Out

The amplifier shown in Fig 10 is practically as simple as an MMIC amplifier. The transistor, an SD1598, produces more than 350 mW output when mounted in a 50-Ω line. The SD1598 (originally another part number) was designed by Bill Olson, W3HQT, years ago, when he was employed by Solid State Microwave (now SGS-Thomson Microelectronics). I've used this transistor, which I refer to as a "hot 2N3866," in amplifiers and frequency multipliers at frequencies from 144 MHz to 3.3 GHz. The SD1598 is in a studded package and the SD1598-1 is a studless package; either style is suitable for use at 903 MHz. I've built two-stage SD1598 amplifiers with 50-Ω lines that work from 400-1300 MHz. Such a two-stage amplifier fits nicely into a 1.5 × 3.6 × 1-inch Bud CU-123 die-cast box. Gain is 10 dB per stage at 903 MHz and 8-9 dB per stage at 1296 MHz.

PC-board artwork isn't necessary for this amplifier, as the 50-Ω line can be easily made on a piece of G-10 PC board using an X-ACTO knife and a straight edge. A 50-Ω line on 0-10 PC board is about 0.1-0.11 inches wide. After cutting the edges of the 50-Ω line, cut another line about a quarter inch from, and parallel to, the 50-Ω line on each side of the 50-Ω line. Using a soldering iron, heat the quarterinch-wide strips between the 50-Ω line and the outer sections and remove them with the knife or a needle-nose pliers. Leave some copper near the transistor for connecting Ql's emitter leads to the microstripline.

Drill a hole in the board for the transistor such that the leads of the transistor can be soldered flush to the microstripline and ground foils. This is important: Excessive lead lengths drastically reduce the gain of this amplifier. Attach a small heat sink to the stud of the device.

The SD1598 can be powered by a 12- to 18-V supply, depending on the gain and power output required. Use an LM317T voltage regulator in the power supply, and vary the voltage to set the gain of this lowlevel stage. This is an easy way to match the drive levels needed by the following stages.

Set Ql's quiescent current, I_{cq}, to 30-50 mA. Vary Rl to get the correct idling current. This bias current is not critical, and depends on the device and the collector voltage used. Tune up is simple: Adjust C8 for maximum power output.

Amplifier No. 6: 100 mW In, 1 W Out

A pair of SDl598s and a Wilkinson power divider/combiner combination are used in this design. The schematic for the combined amplifier is shown in Fig 11, and the PC-board artwork is shown in Fig 12. I didn't make an effort to terminate the 75-Ω Wilkinson divider/combiner; the amplifier works fine as is. Power sharing between the two devices is excellent, and a 3-dB improvement in power output is available over a single device. Saturated power output is well over 1 W.

The supply voltage and idling currents are the same as single-device Amplifier

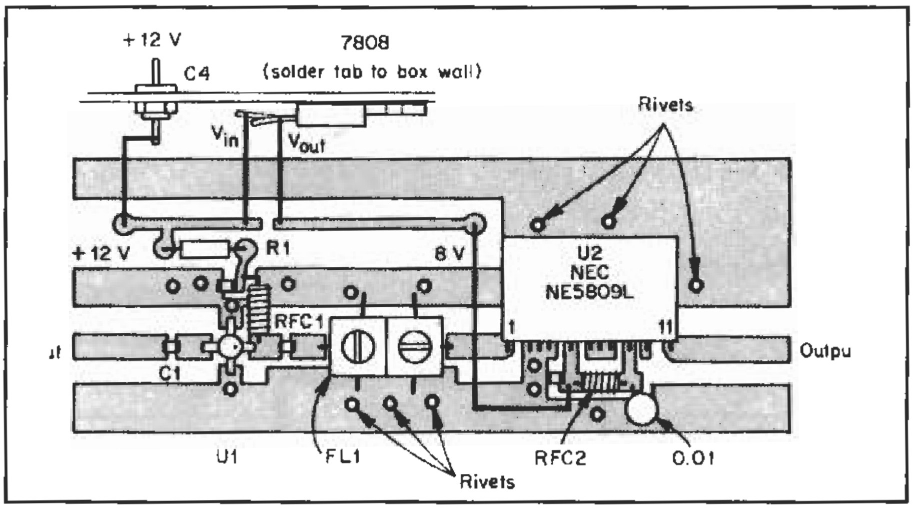

Fig 9—Parts-placement diagram for the 200-mW amplifier. All parts are mounted on the foil side of the board. Note the locations of rivets. At each V_{cc} lead on the MC5809L (pins 5 and 9), install a 100-pF chip capacitor and a 0.01-µF disc to ensure proper bypassing.

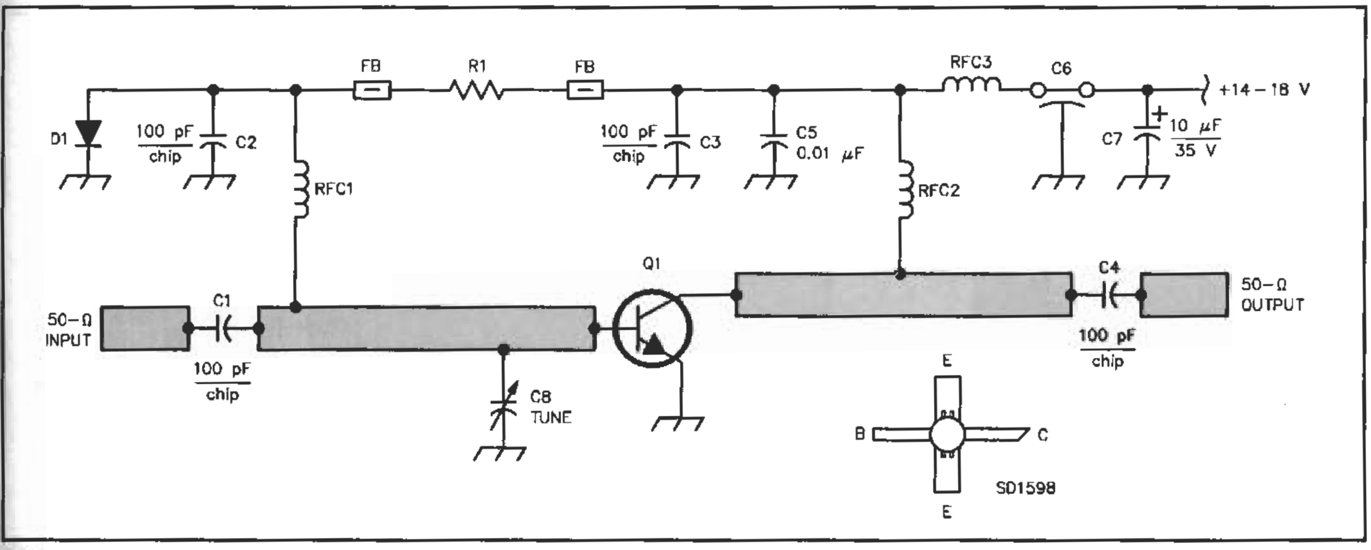

Fig 10—The 350-mW amplifier. Glue D1 to Q1's ceramic body for thermal composition.

C6—0.001-µF feedthrough.
C8—0.6- to 8-pF Johanson piston trimmer.
D1—1N4001.
Q1—SD1598.
R1—1.5-1.6 kΩ, ½ W.
RFC1, RFC2—8 turns of no. 26 enam wire, closewound, 0.1 inch ID.

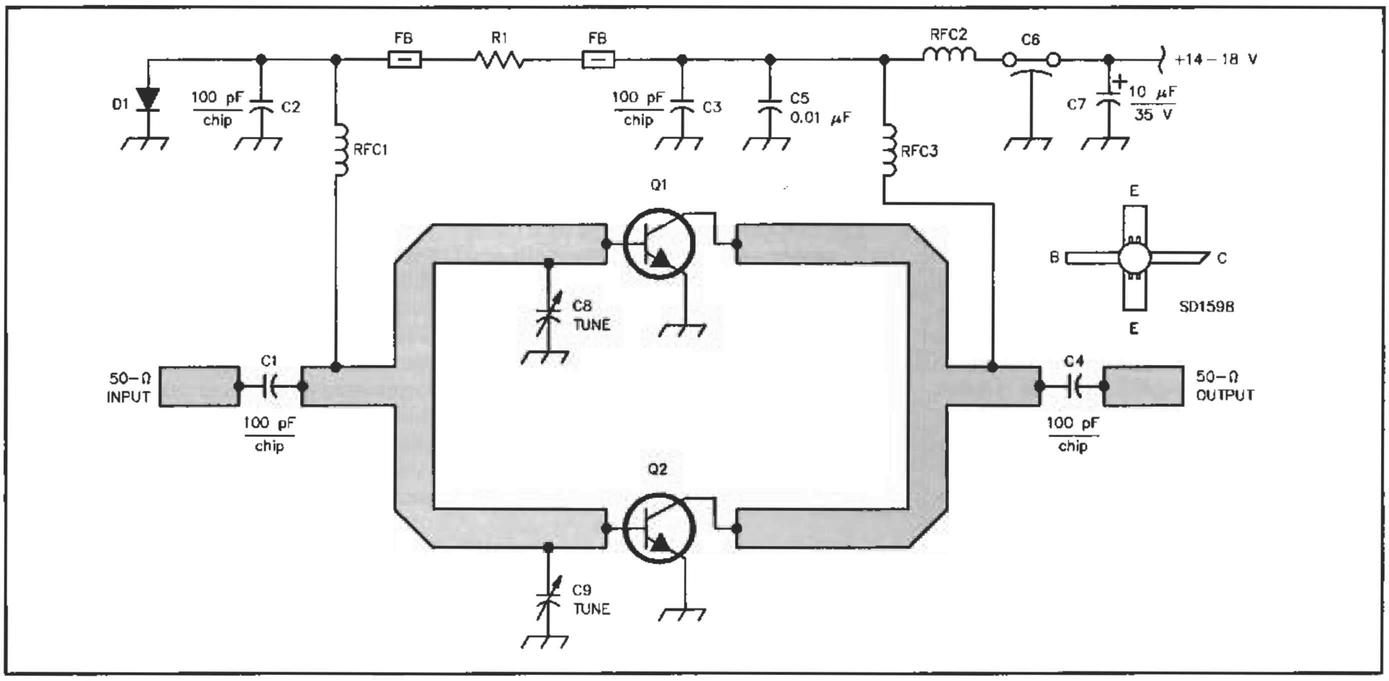
FB
R1
FB
RFC2
C6
+14 – 18 V
D1
100 pF chip
C2
100 pF chip
C3
C5 0.01 µF
C7
10 µF 35 V
RFC1
RFC3
Q1
E
B
C
SD1598
E
C8 TUNE
50-Ω INPUT
C1
100 pF chip
C4
100 pF chip
50-Ω OUTPUT
Q2
C9 TUNE

Fig 11—The 2x SD1598 amplifier. Glue D1 to Q2's ceramic body for thermal composition.

C8, C9—0.6- to 8-pF Johanson piston trimmer.
D1—1N4001.

Q1, Q2—SD1598.
R1—1.5-1.6 kΩ, ½ W

RFC1, RFC3—8 turns of no. 26 enam wire, closewound, 0.1 inch ID.

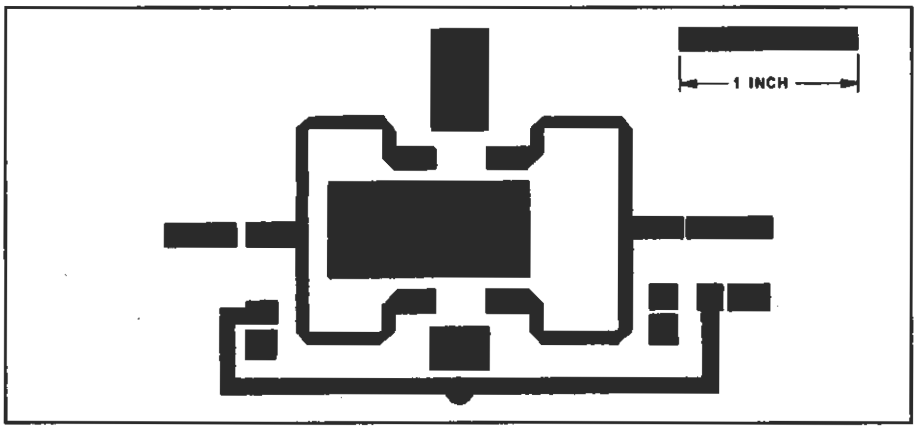

Fig 12—Full-scale PC-board artwork for the 2x SD1598 amplifier. Black areas represent unetched copper foil.

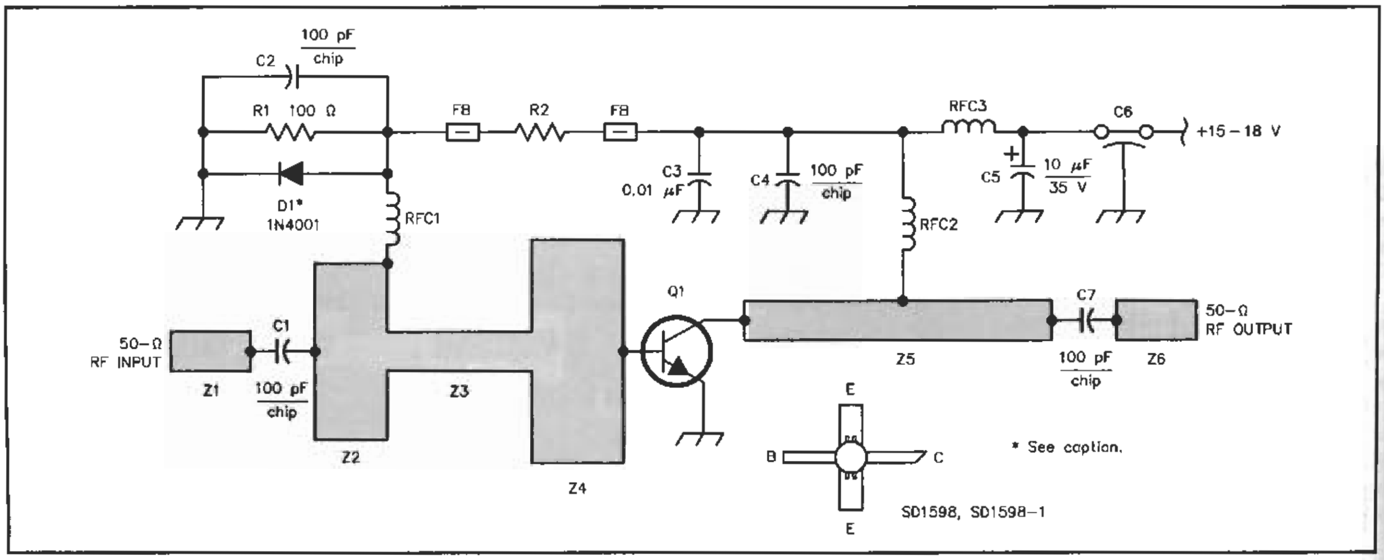
100 pF chip
C2
R1 100 Ω
FB
R2
FB
RFC3
C6
+15 – 18 V
C3 0.01 µF
C4 100 pF chip
C5 10 µF 35 V
D1* 1N4001
RFC1
RFC2
Q1
C7
50-Ω RF OUTPUT
50-Ω RF INPUT
C1
Z1
100 pF chip
Z3
Z5
100 pF chip
Z6
E
B
C
* See caption.
Z2
Z4
SD1598, SD1598-1
E

Fig 13—The 500-mW broadband amplifier. For power levels over 500 mW output, glue D1 to Q1's ceramic body for thermal compensation. RFC1 and RFC2 are implemented as PC-board traces.

Q1—SD1598 or SD1598-1.
R2—1.5-1.6 kΩ, ½ W

RFC3—8 turns of no. 28 enam wire, closewound, 0.1 inch ID.

Z1-Z5—Microstriplines. See text and Fig 14.

No. 5 (Fig 10). If this amplifier is mounted in a die-cast box, no additional heat sinking is required; If not, a small heat sink should be attached to the stud of each device. Adjust C8 and C9 for maximum power output.

Amplifier No.7: 12dB Gain, 500mW Out

The amplifier shown in Fig 13 is a broadband design that's been around for about 6 years. Bill Olson, W3HQT, designed it as a 1296-MHz amplifier. (The 1296-MHz version differs in that it has a small piston trimmer on the output circuit, about an inch from Qi's ceramic cap.)

This amplifier is a very versatile design; it works from 900-1300 MHz and has a power gain of 10-13 dB over this range. I built two test amplifiers for this design: one for the studded transistor and one for the flanged version. Over a dozen different devices were tried in these test circuits, both at 903 MHz and at 1296 MHz. Classes A, AB, C and pulsed-class-C operation were all tried, and all worked well.

The studded SD1598 is used in this design. Full-scale artwork appears in Fig 14; Fig 15 shows parts placement. As with the other designs, rivets are used at all dc and RF grounds. After the rivets are soldered in place, drill a hole in the board for the device. Trim Q1's leads with scissors, then solder them directly to the microstriplines and ground foils. If the amplifier is to be mounted in a die-cast box, the box will provide sufficient heat sinking. If you use another mounting method, attach a small heat sink to Q1's stud.

Power output for SSB is 500 mW, and

Fig 14—Full-scale PC-board pattern for the broadband 500-mW SD1598 amplifier. Black areas represent unetched copper foil.

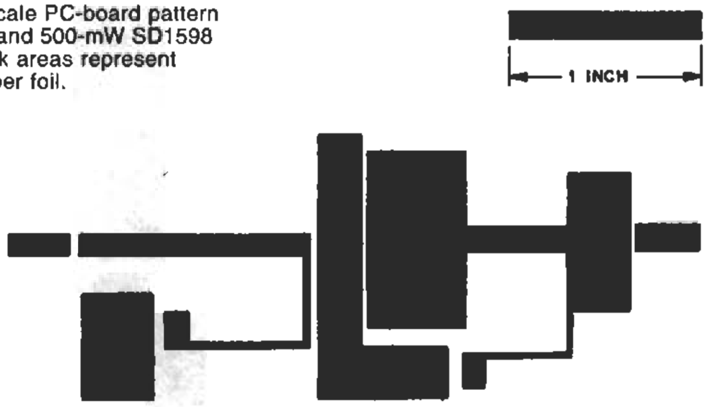

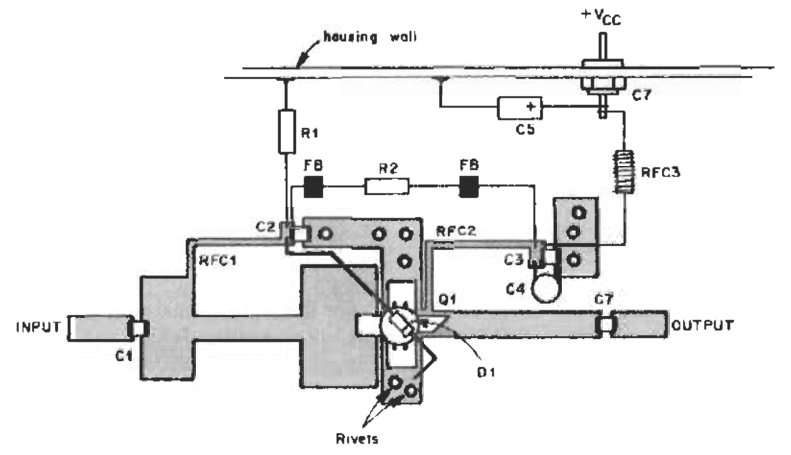
+VCC
housing wall
C7
R1
C5
FB
R2
FB
RFC3
C2
RFC2
RFC1
C3
C4
Q1
C7
INPUT
OUTPUT
C1
D1
Rivets

Fig 15—Parts-placement diagram for the broadband 500-mW SD1598 amplifier. The PC-board edges are not shown. All components are mounted on the foil side of the PC board (except those mounted to the enclosure).

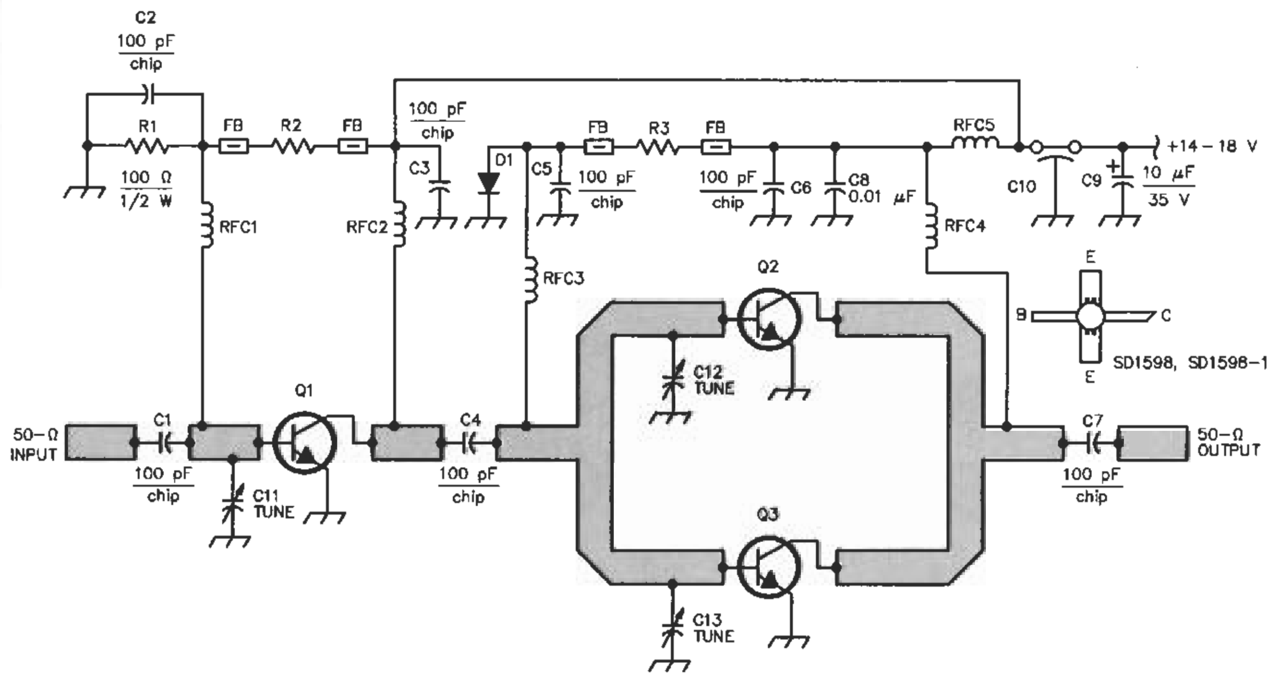
C2
100 pF
chip
R1
100 Ω
1/2 W
FB
R2
FB
RFC1
RFC2
100 pF
chip
C3
D1
C5
FB
R3
FB
100 pF
chip
100 pF
chip
C6
C8
0.01 µF
RFC4
RFC5
C10
C9
10 µF
35 V
+14 – 18 V
RFC3
Q2
E
B
C
E
SD1598, SD1598-1
C12
TUNE
Q1
C1
C4
C7
50-Ω
INPUT
100 pF
chip
C11
TUNE
100 pF
chip
100 pF
chip
50-Ω
OUTPUT
Q3
C13
TUNE

Fig 16—The two-stage, 1-W SD1598 amplifier. Glue D1 to Q2's ceramic body for thermal compensation.

C10—0.001-µF feedthrough.
C11, C13—0.6- to 8-pF Johanson piston trimmer.
Q1—SD1598 or SD-1598-1.
Q2, Q3—SD1598.
R2, R3—1.5-1.6 kΩ, ½ W.
RFC1-RFC5—8 turns of no. 24 enam wire, closewound, 0.1 inch ID.

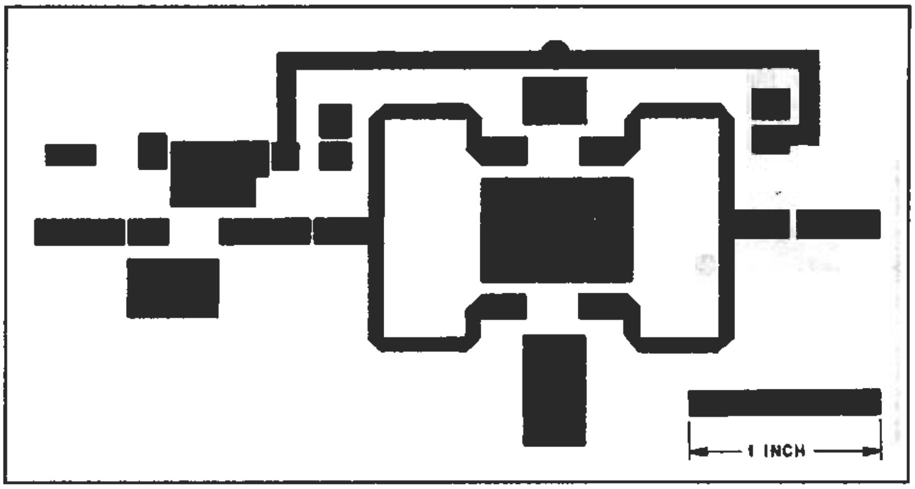

Fig 17—Full-scale PC-board pattern for the two-stage, 1-W SD1598 amplifier. Black areas represent unetched copper foil.

up to 1 W for CW and FM, depending on the device. In applications where the device will deliver over 500 mW, a IN4001 (DI) diode can be glued to Q1 for thermal stabilization, as shown pictorially in Fig 15. Power gain at 1 W output is 7 dB or so. By varying R2, set Q1's I_{cq}, for 30-50 mA for power levels up to 500 mW, and 15-20 mA for higher levels.

For low-level stages (up to 200mW out), you can use the SD1598-1 studless part in this circuit. Solder the studless device to the microstrip, then solder a piece of hobby brass to the device's gold-plated bottom area and to the ground plane. This provides enough heat sinking to dissipate a few hundred milliwatts.

At 903 MHz, no output trimmer capacitor is required. Some improvement in the output match can be achieved by trimming the width of the output line by a few thousandths of an inch with an X-ACTO knife.

Amplifier No. 8: 10 mW In, 1 W Out

A second 1-watt amplifier, this one using a pair of SD1598's driven by a single SD 1598, is shown in Fig 16. This design is similar to that of Amplifier No.6, with the addition of a driver stage. The artwork appears in Fig 17.

For the driver transistor, you can use an SD1598-1 studless device, as Q1 must provide only a hundred milliwatts or so. The driver is mounted in a hole in the PC board. A small piece of brass shim stock is soldered to the bottom of the device and to the ground-plane side of the PC board for heat sinking.

This amplifier runs on 14-18 V dc. The idling current for each device is around 30-50 mA. Again, the supply voltage can be provided by an LM317T regulator. Adjust R2 and R3 to individually set each device's idling current. Then, tune the amplifier by adjusting C11, C12 and C13 for maximum output. Vary the supply voltage to optimize the required power output or gain.

The 2.4 × 4.4 × 1-inch Bud CU-124, or an equivalent die-cast box, is a suitable enclosure for this amplifier. Mount the connectors on the ends of the box, and place the amplifier PC board inside the box body, instead of on the inside of the cover. The box provides sufficient heat sinking for this amplifier.

Next month, I'll describe three more 903-MHz linear amplifiers, including two 4-W units.

Notes

[1] D. Mascaro,"A 759-MHz Local Oscillator," *QEX*, May 1988, pp 12-15.

[2] No-tune transverters, as well as antennas and microwave components, are available from Down East Microwave, Box 2310, RR 1, Troy, ME 04987, tel 207-948-3741. Catalog available.

[3] The amplifiers in this article use transistors manufactured by SGS-Thomson Microelectronics, 211 Commerce Dr, Montgomeryville, PA 18936, tel 215-362-8500.

[4] Mouser Electronics, 2401 Hwy 287 N, Mansfield, TX 76063, tel 817-483.4422. Catalog available.

[5] Microwave Components of Michigan, 17141 Merriman, Romulus, MI 48174, evening tel 313-753-4581. Parts list available.

[6] RF Parts, 1320 Grand Ave, San Marcos, CA 92069, tel 619-744-0728.

[7] SGS-Thomson Microelectronics transistors are available through RF Gain Ltd, 100 Merrick Rd, Rockville Center, NY 11570, tel 800-645-2322. $50 minimum order.

[8] Frontier Microwave, RD 1, Box 467, Mink Rd, Ottsville, PA 18942, evening tel 215-795-2648.

[9] D. Mascaro,. "A Transverter Band-Switching Display and Universal Power Supply," *QEX*, Aug 1987, pp 8-14.

[10] Available from Digi-Key Corp, 701 Brooks Ave South, PO Box 677, Thief River Falls, MN 56701-0677, tel 800-344-4539. Catalog available.

[11] Toko filters and other components are available from Steve Kostro, N2CEI, Box 341A, RD 1, Frenchtown, NJ 08825, evening tel 201-996-3584. Parts list available.

By Dave Mascaro, WA3JUF

From *QST*, July 1990

903-MHz Linear Amplifiers

Part 2—Did you like the projects in Part 1? Here are three more amplifiers to suit your 903-MHz, 50-ohm gain-block needs.

In Part 1, I covered construction methods for building the 903-MHz linear amplifiers described in this twopart series. The amplifiers described this month cover a higher power range than those in Part 1, starting with a 2-W-output design, and finishing with a pair of 4-W output units (one with 13 dB gain, the other with 23 dB gain).

Amplifier No. 9: 100 mW In, 2 W Out

The transistor used in this amplifier, an SD1853, is a class-A device in a "strip-pac" flanged package. The PC-board artwork is shown in Fig 18, and the schematic appears in Fig 19. The parts-placement

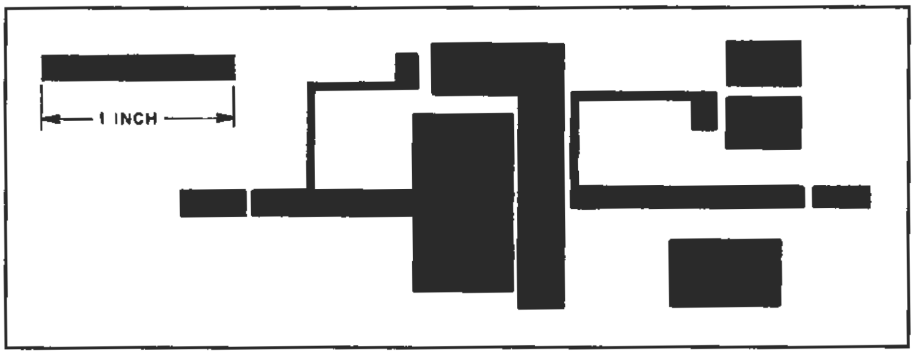

Fig 18—Full-scale PC-board pattern for the 2-W SD1853 amplifier. Black areas represent unetched copper foil.

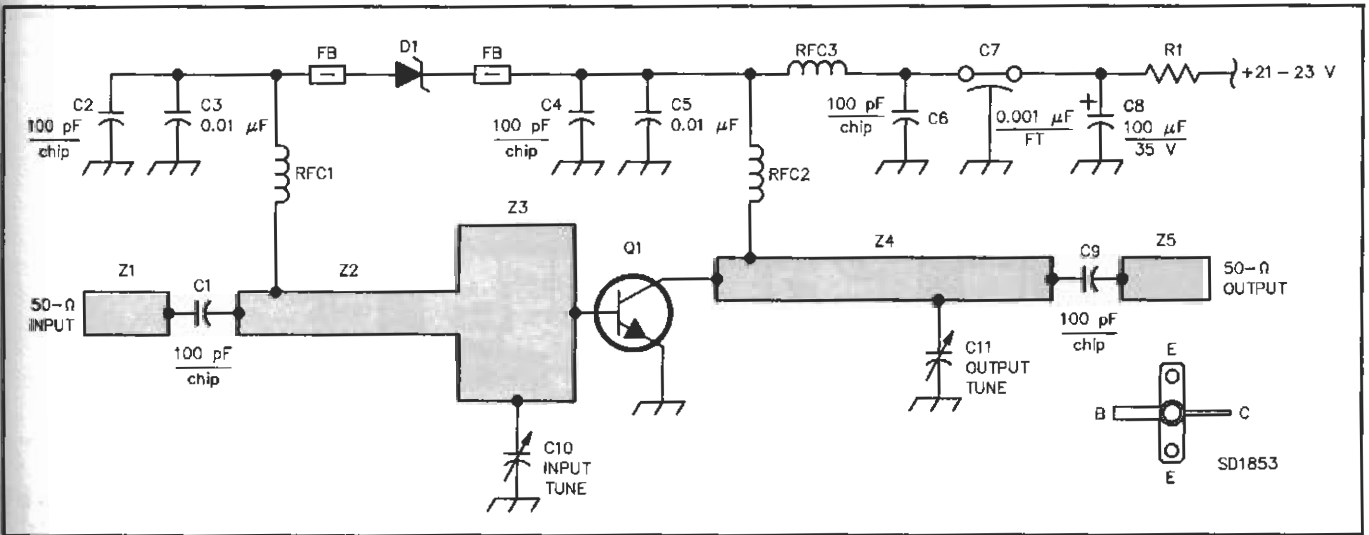
FB
D1
FB
RFC3
C7
R1
+21 – 23 V
C2
100 pF
chip
C3
0.01 µF
C4
100 pF
chip
C5
0.01 µF
100 pF
chip
C6
0.001 µF
FT
C8
100 µF
35 V
RFC1
RFC2
Z3
Z1
C1
Z2
Q1
Z4
C9
Z5
50-Ω
INPUT
100 pF
chip
50-Ω
OUTPUT
100 pF
chip
C11
OUTPUT
TUNE
E
B
C
E
SD1853
C10
INPUT
TUNE

Fig 19—The 2-W SD1853 amplifier. RFC1 and RFC2 are implemented as PC-board traces.

C10, C11—0.3- to 3-pF Johanson piston trimmer.
D1—1N4747A 20-V, 1-W Zener.
Q1—SD1853.
R1—2-3 Ω, 1 W.
RFC3—8 turns of no. 26 enam wire, closewound, 0.1" ID.
Z1-Z5—Microstriplines. See text and Fig 18.

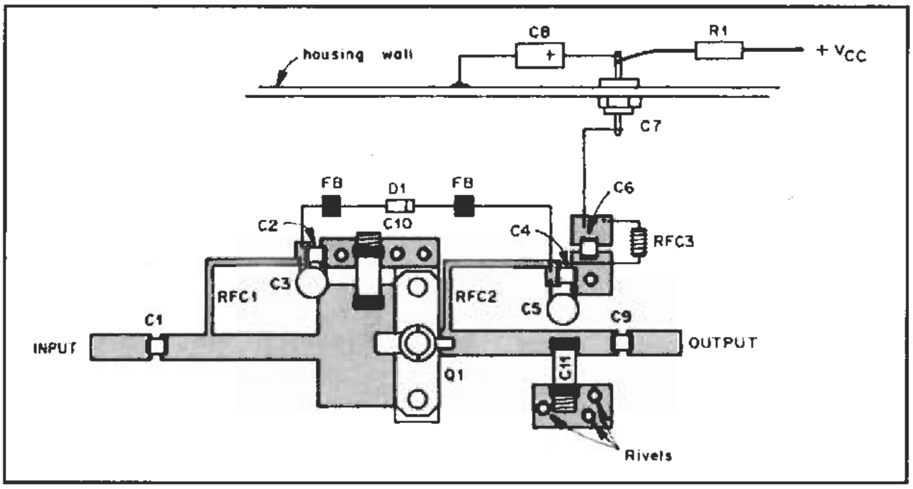
housing wall
C8
R1
+ V_{CC}
C7
FB
D1
FB
C6
C2
C10
C4
RFC3
RFC1
C3
RFC2
C5
C1
C9
INPUT
OUTPUT
C11
Q1
Rivets

Fig 20—Parts-placement diagram for the 2-W SD1853 amplifier. The PC-board edges are not shown. All components are mounted on the trace side of the PC board (except those mounted to the enclosure).

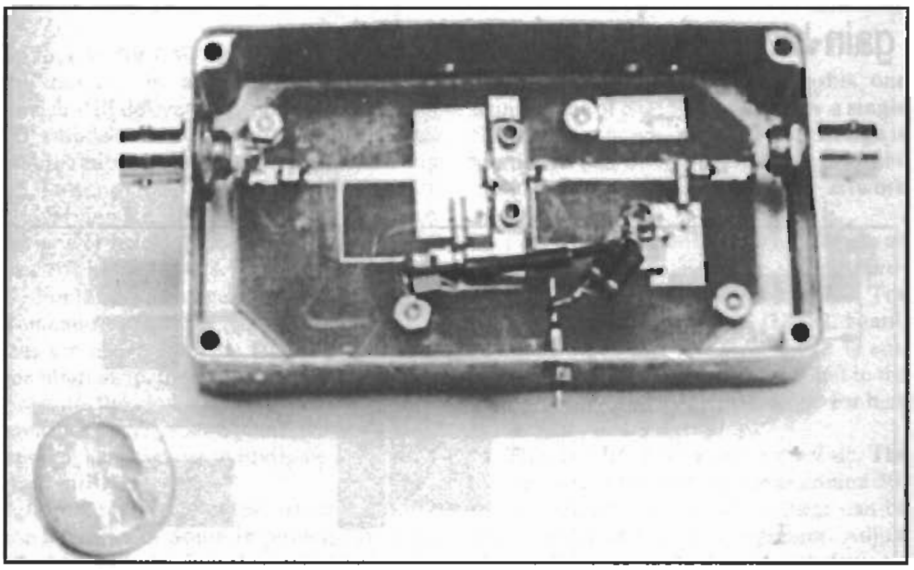

Fig 21—The prototype 2-W amplifier.

diagram is shown in Fig 20, and a photo of my prototype appears in Fig 21.

After the board is etched and rivets are installed, cut a hole and file its edges to accept Q1's flange. The SD1853 must be mounted to a heat sink. (A Bud CU-124 or equivalent die-cast box is adequate.) Use a piece of G-10 PC board between the amplifier board and the heat sink, to allow the device to be mounted flush to the microstrip. The device flange is the emitter, and a low-inductance ground connection is a must. Make this connection by using a piece of copper-foil tape soldered to the microstrip ground plane and placed between Q1's flange and the heat sink. Mount the transistor to the heat sink with no. 4-40 screws. Use heat-sink compound between Q1 and the heat sink.

An LM317T (or LM317K) adjustable regulator can be used to supply the voltage required by this amplifier. Use a 10-turn pot for the ADJUST control; smooth adjustment of the supply voltage to this amplifier is a must. Insert a milliammeter in series with current-limiting resistor R1, and slowly adjust the regulator output up from minimum until the idling current is 200-225 mA without RF drive applied. The exact supply voltage depends on the beta of

Fig 22 (below)—The 4-W, 2 x SD1853 amplifier.
C4-C6—0.3- to 3-pF Johanson piston trimmer.
D1—1N4747A 20-V, 1-W Zener.
Q1, Q2—Sd1853.
R1—2-3 Ω, 1 W.
RFC1, RFC2—8 turns of no. 26 enam wire, closewound, 0.1" ID.
Z1-Z5—Microstriplines. See text and Fig 23.

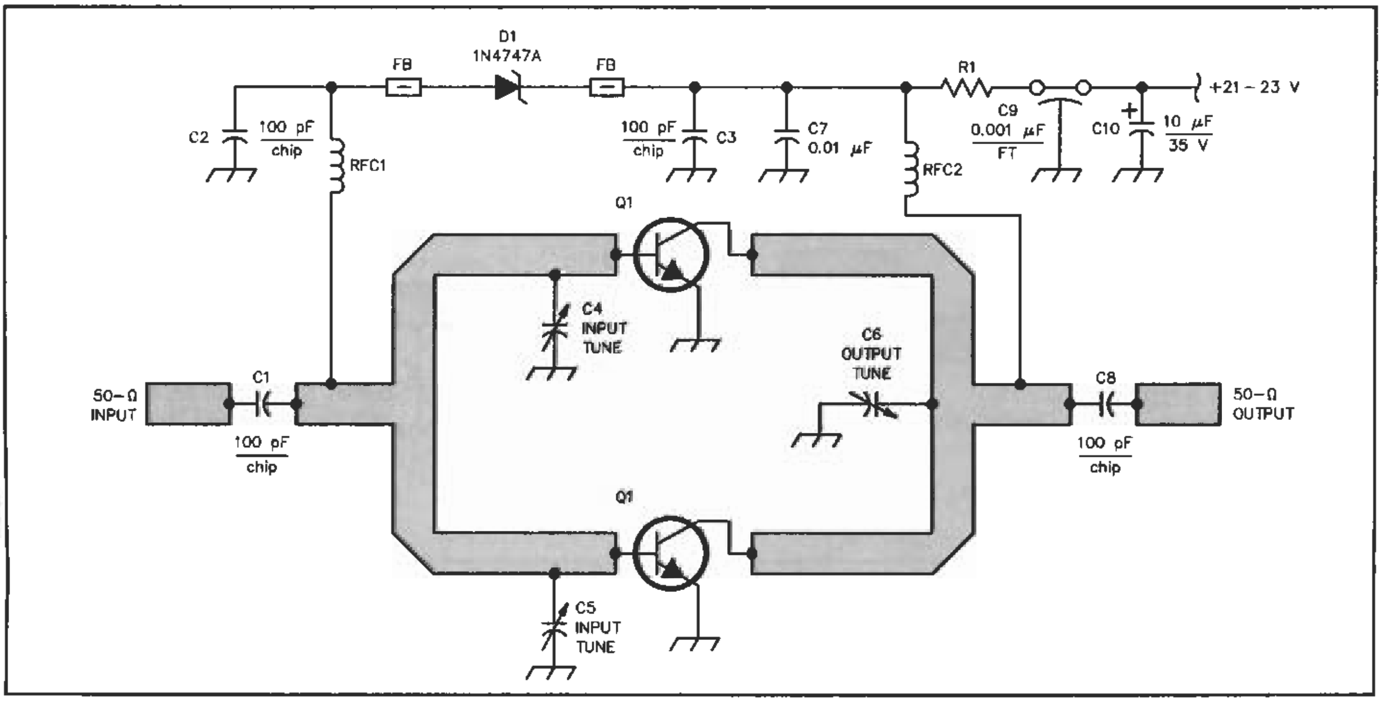
D1
1N4747A
FB
FB
R1
+21 - 23 V
C2
100 pF
chip
RFC1
100 pF
chip
C3
C7
0.01 µF
C9
0.001 µF
FT
C10
10 µF
35 V
RFC2
Q1
C4
INPUT
TUNE
C6
OUTPUT
TUNE
50-Ω
INPUT
C1
100 pF
chip
C8
100 pF
chip
50-Ω
OUTPUT
Q1
C5
INPUT
TUNE

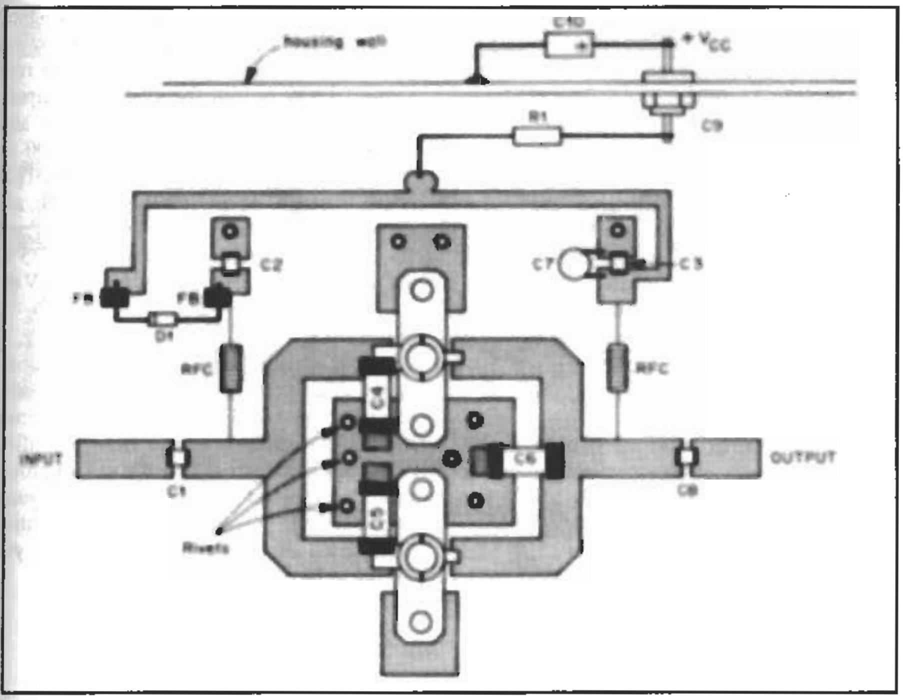
housing wall
C10
+Vcc
R1
C9
C2
C7
C3
FB
FB
D1
RFC
RFC
C4
C5
C6
INPUT
OUTPUT
C1
C8
Rivets

Fig 23—Parts-placement digram for the 2 x SD1853 amplifier. The PC-board edges are not shown. All components mount to the trace side of the PC board (except those mounted to the enclosure).

Q1, the value of R1, and D1's breakdown voltage. Idling current can be adjusted slightly to optimize gain and/or power output. The supply voltage must be removed during receive to minimize device heating.

Tune-up is simple. Apply drive and tune C10 and C11 alternately for maximum power output. If you have access to a return-loss bridge or network analyzer, you may want to trim the width of Z2 to improve the input return loss. Verify stability (as indicated by no output and no change in supply current) by tuning the trimmers through their ranges with no input signal applied.

Amplifier No. 10: 200 mW In, 4 W Out

This high-gain design consists of a pair of SD1853s combined in a Wilkinson power divider/combiner. As before, I made no effort to terminate the 75-Ω Wilkinson divider/combiner with balancing resistors. It worked just fine without them. Even the 1296-MHz version, which uses the same board layout, worked well in this configuration.

Fig 22 shows the schematic of this class-A amplifier, and Fig 23 shows parts placement. The PC-board pattern, the same

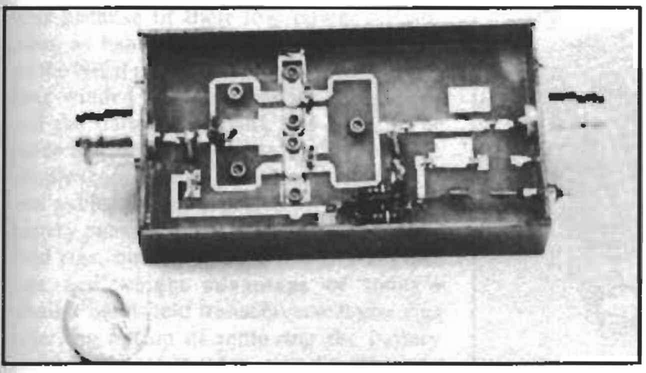

Fig 24—This prototype 4-W amplifier was built on a PC board with a spare driver stage that's bypassed with a brass strip.

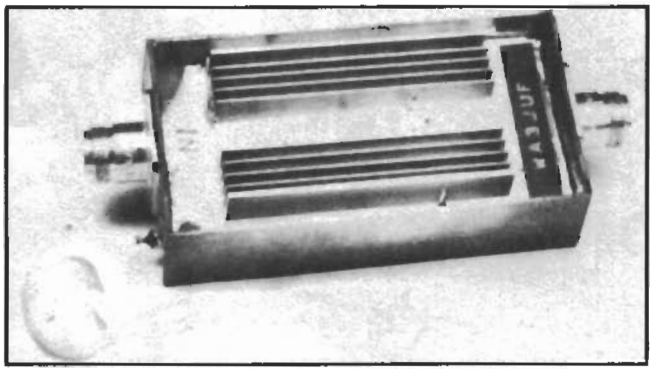

Fig 25—The 4-W amplifier's heat sink, bolted directly to the devices, must be positioned such that cooling air can flow over it during amplifier operation.

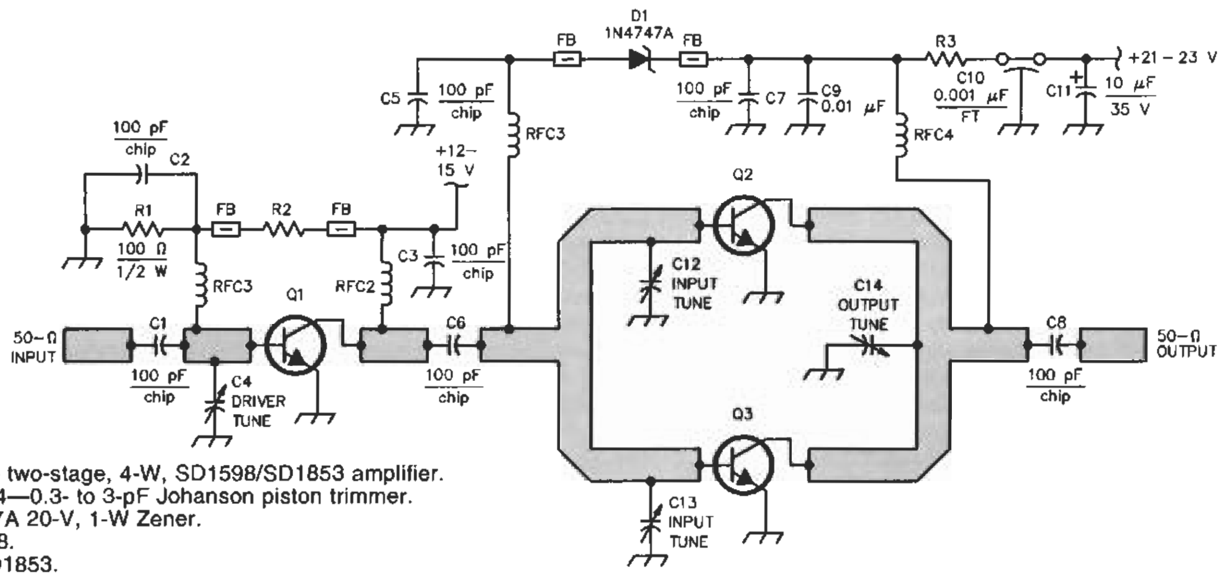
D1
1N4747A
FB
FB
R3
+21 – 23 V
C5
100 pF
chip
C7
100 pF
chip
C9
0.01 μF
C10
0.001 μF
FT
C11
10 μF
35 V
RFC3
RFC4
100 pF
chip
C2
+12–
15 V
Q2
R1
100 Ω
1/2 W
FB
R2
FB
C3
100 pF
chip
C12
INPUT
TUNE
C14
OUTPUT
TUNE
RFC3
Q1
RFC2
50-Ω
INPUT
C1
100 pF
chip
C4
DRIVER
TUNE
C6
100 pF
chip
C8
100 pF
chip
50-Ω
OUTPUT
Q3
C13
INPUT
TUNE

Fig 26—The two-stage, 4-W, SD1598/SD1853 amplifier.
C4, C12-C14—0.3- to 3-pF Johanson piston trimmer.
D1—1N4747A 20-V, 1-W Zener.
Q1—SD1598.
Q2, Q3—SD1853.
R2—1.5-1.7 kΩ, ½ W. See text, Part 1, June 1990 *QST*, p 24.
R3—1-2 Ω, 2 W.
RFC1-RFC4—8 turns of no. 26 enam wire, closewound, 0.1" ID.

as that for Amplifier No. 8, is shown in Fig 17 (shown in Part 1). The board's driver-stage traces are by-passed with a brass strip in this amplifier. Fig 24 shows my prototype mounted in a PC-board enclosure. A single base-bias source, using a current-limiting resistor and Zener diode, feeds both devices. Total idling current, measured in series with R1, is 400-450 mA.

The same mounting arrangement used with the single-SD1853 amplifier (Amplifier No. 9) applies to this unit. The device flanges must have low-inductance ground connections. Mount the transistors to a heat sink about 2 × 3 inches with ½ inch or taller fins, as shown in Fig 25. Again, the supply voltage is switched on only during transmit. To tune this amplifier, adjust C4, C5 and C6 alternately for maximum output.

Amplifier No. 11: 20 mW In, 4 W Out

A second 4-W amplifier, using a pair of SD1853s driven by a single SD1598, is shown in Fig 26. The final stage is Amplifier No. 10, and the driver is the same as that used in Amplifier No. 8. The PC-board artwork is shown in Fig 17; the parts placement can be done using Figs 23 and 16 as guides.

The power-supply and bias adjustments are the same as those for the previously described amplifiers using these devices. Transistor mounting is also the same as before, but extra heat sinking is required with this unit because of the class of operation and the power output.

Higher Power Output

Higher-power devices usable at 903 MHz are available from several manufacturers. NEC's NEL1320 (2SC3542) produces 18-20 W linear output at 903 MHz. These devices also work on the 1296-MHz band. These transistors, as well as single- and dual-stage amplifiers based on them, are available from Down East Microwave (DEM). (The DEM two-stage 3318PA runs around 18 W output for about 1 W input).

SGS-Thomson Microelectronics also makes high-power devices for the 800- to 960-MHz range. For example, the SD1423 is a 4-W-in, 30-W-out class AB device, and the SD1660 is a class-AB transistor that runs a whopping 120 W output for about 35 W drive. Both of these devices run on 28 V dc.

Summary

As I've shown in this two-part series, there's really nothing difficult about building your own 903-MHz equipment—from the local oscillator to the final amplifier. This ease of construction is a vast improvement over what it was just a few years ago—and the 903-MHz band's population has grown as a reflection of that. See you on the microwaves!

By Dave Mascaro, WA3JUF From *QST*, August 1988

25-Watt Linear Amplifiers for 144 and 200 MHz

Do you need more punch from your hand-held or portable rig for 2 meters or 220 MHz? These little amplifiers can supply it!

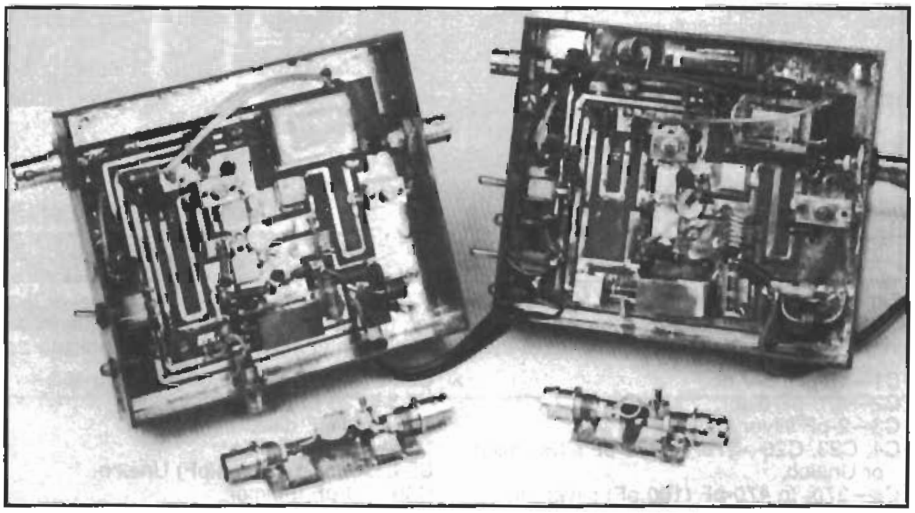

Many hams have 2-meter and 220-MHz hand-held rigs and low-power portable SSB rigs these days. Unfortunately, many of these radios have limited usefulness because of their low power output. Also, as hand-held radios get smaller, so do the NiCd packs that power them—a few long-winded transmissions on high power and the battery is dead! The low-power mode is usually good only for short-range simplex operation, or repeater use when you're close to the repeater. High-capacity battery packs are available for most hand-held rigs, but using them takes away the size and weight advantage of today's smaller hand-held transceivers. Some rigs offer the option of removing the battery pack and plugging the unit directly into 12-V dc, which gives you a small, lightweight rig, but you're still stuck with relatively low power.

The solution to this problem is the addition of a linear amplifier. An amplifier after the hand-held transceiver or portable rig that can give you 25 to 30 W of output power (depending on the output of the driving rig) without imposing large weight, cost and current-drain constraints is a great addition to your VHF station. An amplifier also allows you to use the same low-powered rig in the house and in the car with high power output. You can mount an amplifier under the seat or in the trunk of your car, and minimize the possibility of theft by taking the radio with you when you leave the car.

Amplifier Design

The amplifiers described in this article are capable of 25 to 30W of RF output for the drive power available from common handheld and portable rigs. The 2-meter and 220-MHz designs are so similar that we can describe them both in one discussion. The differences in the designs are covered in detail. Design, assembly, tuning and use of the amplifiers is virtually identical for both versions. Both versions even use the same PC board! (Note that two different boards are shown in the lead photo—they were prototypes of the fmal board design.)

Receiving preamplifiers are included in both units. They, too, are quite similar, differing only in some component values.

This power amplifier uses a single SD1274 bipolar transistor manufactured by Thomson Components/Mostek Corp.[1] The device is operated class AB for all-mode operation. Nominal dc power-supply voltage is 13.8. The amplifier will operate on any dc supply voltage between 12 and 14.5 (the typical automobile supply voltage range).

The main amplifier PC board is a microstripline design on standard $^1/_{16}$-inch-thick, double-sided, G-10 fiberglass-epoxy board. Input and output tuning capacitors are provided for maximizing gain and power output in a given band segment. If desired, the amplifiers can be tuned for broadband operation with only a slight reduction in power output across the operating frequency range.

The 2-meter and 220-MHz amplifiers are narrow-band designs. The 2-meter version gives 25 W output for 2 W of drive from 138 to 150 MHz, with a single tuning setting. When tuned for narrow-band operation, the amplifier gives about 20 W output for 1 W of drive. With the Kenwood TH-21AT on high power, my 2-meter amplifier puts out 25 W. On low power, I can adjust the amplifier for a maximum narrow-band gain, which gives about 7 to 8 W output. Efficiency is 50 to 60%, depending on tuning and power output.

The tuning range of the 220-MHz amplifier is 200 to 230 MHz. When tuned for narrow-band operation, the amplifier gives about 16 W output for 1 W of drive. Driving the amplifier with my ICOM IC-3AT on high power (about 1.5 W), the amplifier puts out 20 W. With the IC-3AT on low

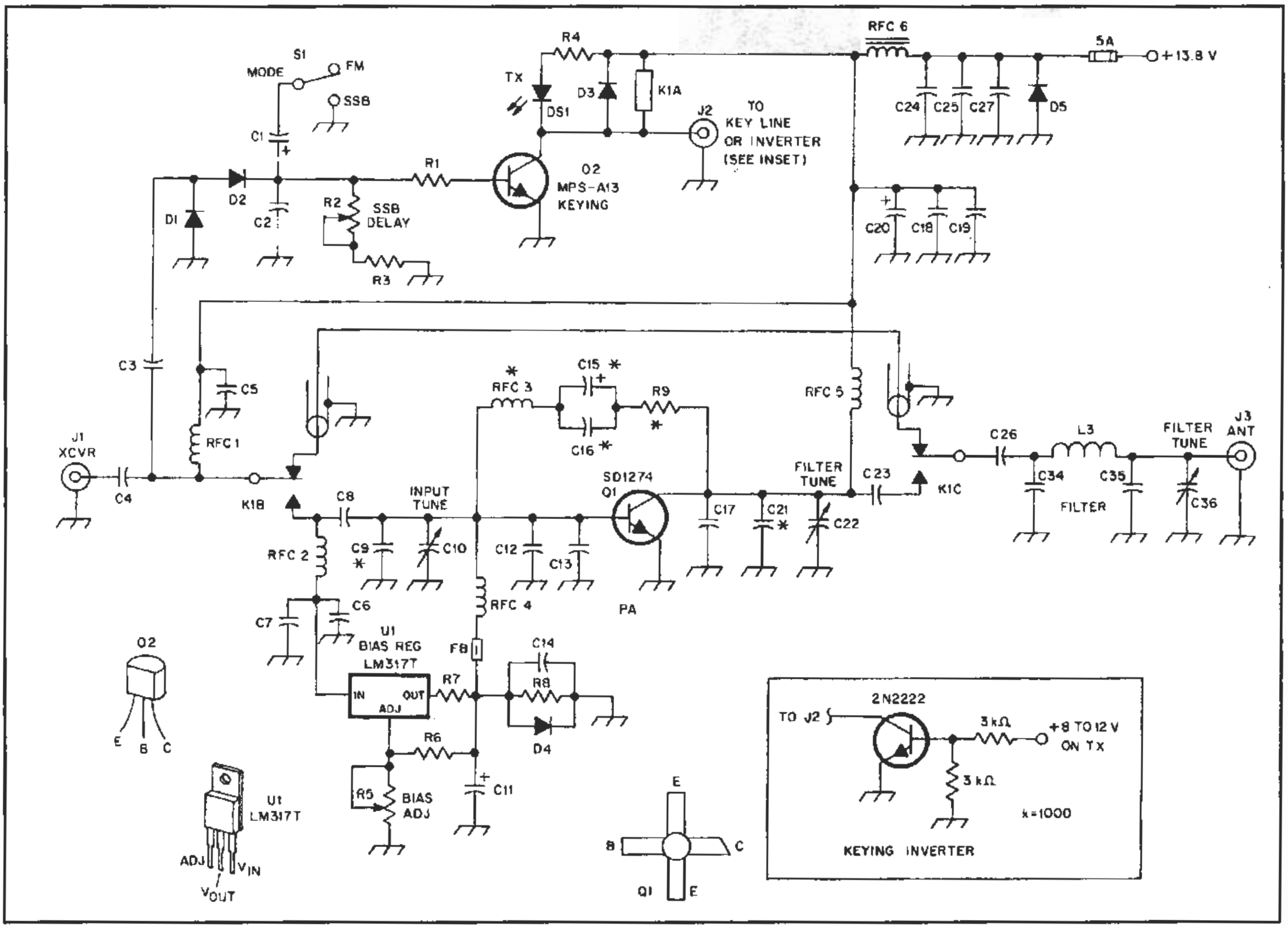
S1
MODE
FM
SSB
C1
R4
TX
DS1
D3
K1A
J2
TO KEY LINE OR INVERTER (SEE INSET)
RFC 6
5A
+13.8 V
C24
C25
C27
D5
R1
Q2 MPS-A13 KEYING
D2
D1
C2
R2
SSB DELAY
R3
C20
C18
C19
C3
C5
J1 XCVR
RFC 1
C4
K1B
C8
INPUT TUNE
RFC 2
C9
C10
RFC 3
C15
C16
R9
SD1274 Q1
C12
C13
RFC 4
PA
C17
C21
C22
FILTER TUNE
RFC 5
C23
K1C
C26
L3
C34
C35
FILTER
FILTER TUNE
C36
J3 ANT
C7
C6
U1 BIAS REG LM317T
IN
OUT
ADJ
FB
R7
C14
R8
D4
R6
R5
BIAS ADJ
C11
Q2
E B C
U1 LM317T
ADJ
VIN
VOUT
E
B
C
Q1
E
2N2222
TO J2
3 kΩ
+8 TO 12 V ON TX
3 kΩ
k=1000
KEYING INVERTER

Fig 1—Schematic of the 25-W VHF amplifiers. The 5-A fuse is included in the power lead (external to the amplifier). Be extremely careful when mounting the power transistor—see text for precautions. In the parts list below, values for the 220-MHz version are given in parentheses where they differ from those in the 2-meter version. Asterisks indicate parts not used in the 220-MHz version.

C1—15-µF, 6-V dc.
C2—470-pF disc.
C3—2-pF silver mica.
C4, C23, C26—270- to 470-pF silver mica or Unelco.
C8—270- to 470-pF (100 pF) silver mica or Unelco.
C5, C7, C14, C16, C19, C24—0.01-µF disc.
C6—0.22-µF disc.
C9—43-pF silver mica or Unelco.
C10—Arco 404—4 to 60 pF (Arco 403—4 to 40 pF).
C11, C15—1-µF, 15-V tantalum.
C12—180-pF (120-pF) Unelco.
C13, C18—220-pF Unelco.
C17—68-pF (51-pF) Unelco.
C20—10-µF, 35-V electrolytic.
C21*—33 pF silver mica or Unelco.
C22—Arco 404—4 to 60 pF (Arco 402—1.5 to 20 pF)
C25—0.1-µF disc.
C27—0.001-µF disc.
C34, C35—25-pF (18-pF) Unelco.
C36—10-pF trimmer.
D1, D2—1N4148 or 1N914.
D3, D4—1 N4001 or equiv.
D5—ECG 581 or equiv.
DS1—LED.
FB—Ferrite bead.
K1—Omron LZN2O3-UA-DC12 DPDT relay.
L3—3 turns (1 turn) no. 18, closewound, 0.2-in. (¼-in.) ID.
Q1—Thomson/Mostek SD1 274.
Q2—MPS-A13 NPN Darlington.
R1, R4—1.5-kΩ, ¼ W.
R2—500-kΩ, 10-turn potentiometer.
R3—10-kΩ, ¼ W.
R5—100-Ω miniature potentiometer.
R6—270-Ω, ¼ W.
R1, R8—10-Ω, ½ W.
R9*—15-Ω, ¼ W.
RFC1, RFC2, RFC4—0.47-µH molded choke.
RFC3*—0.15-µH molded choke.
RFC5—6 turns (5 turns) no. 16 enam, ¼-in. ID.
RFC6—VK200/4B ferrite choke.
S1—Miniature SPDT toggle.
U1—LM317T voltage regulator

Miscellaneous
2 BNC or N connectors (see text).
4 × 4 × 1½-in, heat sink.
5-A fuse and in-line holder.
2 × 1-in, scrap of thin sheet brass.

power (about 150 mW), I can tune the amplifier to give 4W output. Efficiency is about the same as the 2-meter version.

The saturated power output for both versions is more than 35 W in FM operation. The maximum FM power input is 4 W. Minimum power input for proper RF sensed keying operation is 100 mW.

The Power-Amplifier Circuit

See Fig 1. K1 is similar to relays used in VCR RF circuits. I have switched up to 40 W at 2 meters and 35 W at 220 MHz with these relays. They present a good match to 50 Ω. The loss through the amplifiers (when not in use) resulting from these relays is less than 1 dB. This is typical for 2-meter and 220-MHz amplifiers.

In addition to switching the transmitted and received signals, the relays also switch the dc supply voltage. In the transmit mode, K1 switches 13.8 V dc to the input of the bias regulator, U1. The regulator IC and its associated resistors (R7, R8) and bias diode (D5) supply a stiffer bias voltage to the base of Q1 than the more-common voltage-divider bias networks. This low-impedance bias source keeps the bias voltage constant over the range of RF drive levels. With a voltage-divider bias network, bias voltage can be upset by the base-emitter rectification (self bias) developed by the RF driving signal. In such a case, over-driving the amplifier causes the base bias to decrease, resulting in non-linear amplification. This effect is limited with the regulator-type bias

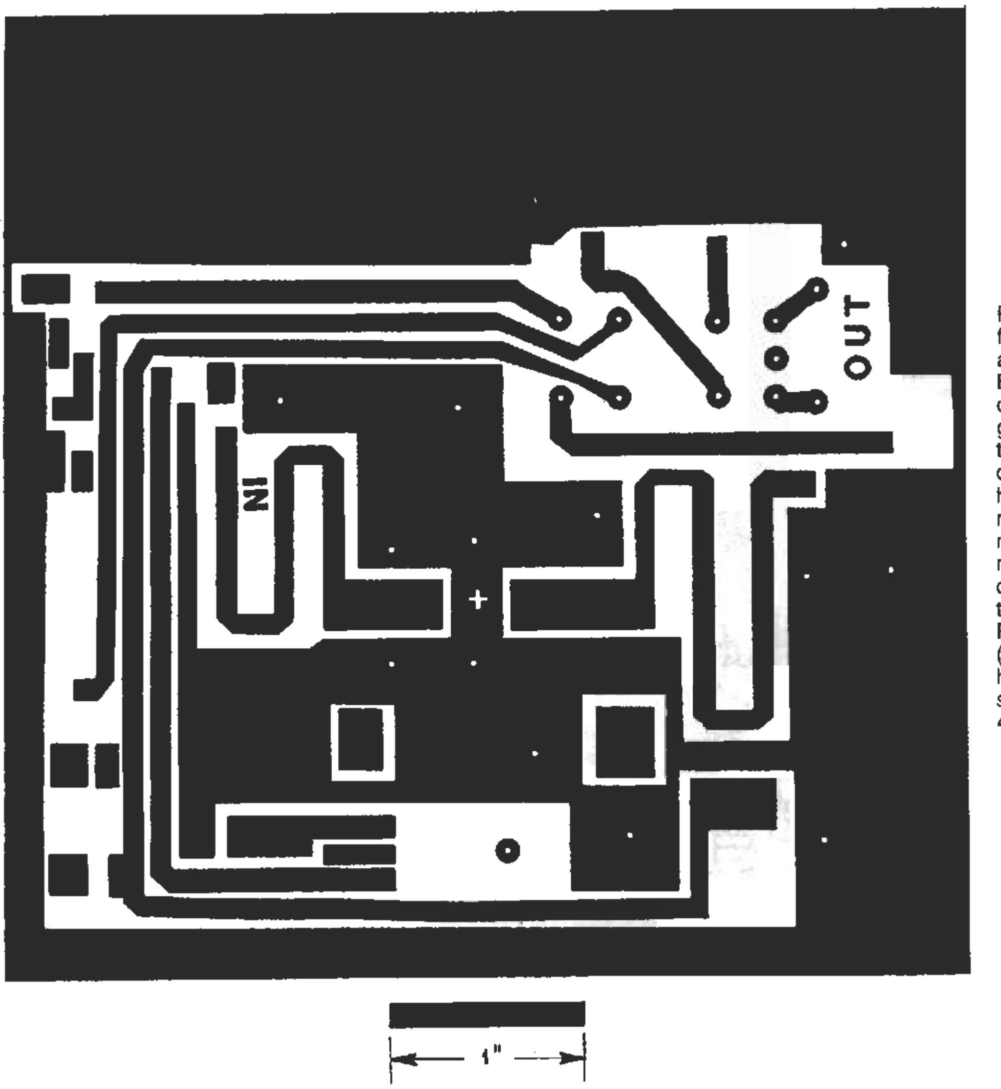

Fig 2—Main PC-board layout for the 2-meter and 220-MHz amplifiers. Artwork is full scale. Because this is a microstripline design, do not change the geometry or routing of board traces. The board must be drilled for rivets, mounting hardware and transistor mounting at the locations marked. The transistor-mounting hole is 0.38-in. diam and should be drilled at the + near the center of the PC board. All other holes (except for the relay-mounting holes near the output microstripline) are for rivets and no. 4-40 mounting hardware.

circuit used in these amplifiers.

TR switching is accomplished by an RF sensing circuit. A small amount of RF is sampled by C3 and rectified by D1 and D2, which turns on keying transistor Q2. Q2 pulls in K1, which switches the amplifier into the line.

When S1 is in the SSB position, a short drop-out delay is added in the RF-sensed-keying circuit to keep the amplifier keyed during brief pauses in speech. This delay is adjustable by varying R3. In the FM position, no delay is needed. S1 does not change the class of operation of Q1—it merely switches in the delay circuit.

The switching relays are wired so the amplifier can remain in-line at all times, with or without the supply voltage connected. Applying the supply voltage allows you to use the amplifier and preamp. Without the supply voltage connected, all transmitted and received signals pass through the amplifier.

The spectral output of these amplifiers is quite good, but to ensure clean signals, I've added filters to the output of each unit. Although there isn't much extra room for the filter components inside the amplifier cabinet, you can mount them on the back cover or in another small enclosure. The filters, shown in the title photo, are very simple to build and tune.

Construction

Each amplifier consists of two PC boards: the power-amplifier board and the preamp board (if used). After the PC boards are etched, holes are drilled with a no. 50 bit for installation of tinned grounding rivets at all RF and dc grounds (see Fig 2).[2] Install the rivets on the board as follows: After inserting the rivets, flare the inserted end with an awl. Next, flatten the rivet by tapping it lightly with a hammer, using an anvil or other solid surface under the PC board as a support. Solder the rivets on both sides of the PC board. Alternatively, pieces of wire can be soldered through the board—but because the wires will not be flush with the board, mounting the heat sink may be difficult.

Drill the holes for Q1 and the board-mounting screws as indicated in Fig 2. No holes are needed for component mounting because all parts are mounted on the trace side of the PC board. All the components except the preamp board are mounted next. (The preamp board is mounted after the amplifier has been tested, to prevent possible damage to the preamplifier.) Connect a piece of miniature 50-Ω Teflon® or RG-58 coaxial cable between K1 and the amplifier-input microstripline (see Fig 3).

After drilling or milling holes for Q1 in a suitable heat sink (see Fig 4), tap the mounting holes in the heat sink for no. 4-40 hardware, and mount the PC board to the heat sink. Trim the four leads of Q1 to about half their original length to make mounting easier.

Q1 can be mounted in one of several ways. The distance between the underside of the leads of Q1 and the heat-sink mounting area of Q1 is larger than the thickness of $^1/_{16}$-in. G-10 board material. Because of this, the heat sink must be milled (see Fig 4), or small pieces of copper or brass must be soldered under each lead of Q1 to make up the difference in height. Alternatively, a second piece of G-10 material can be cut to the same size as the amplifier PC board

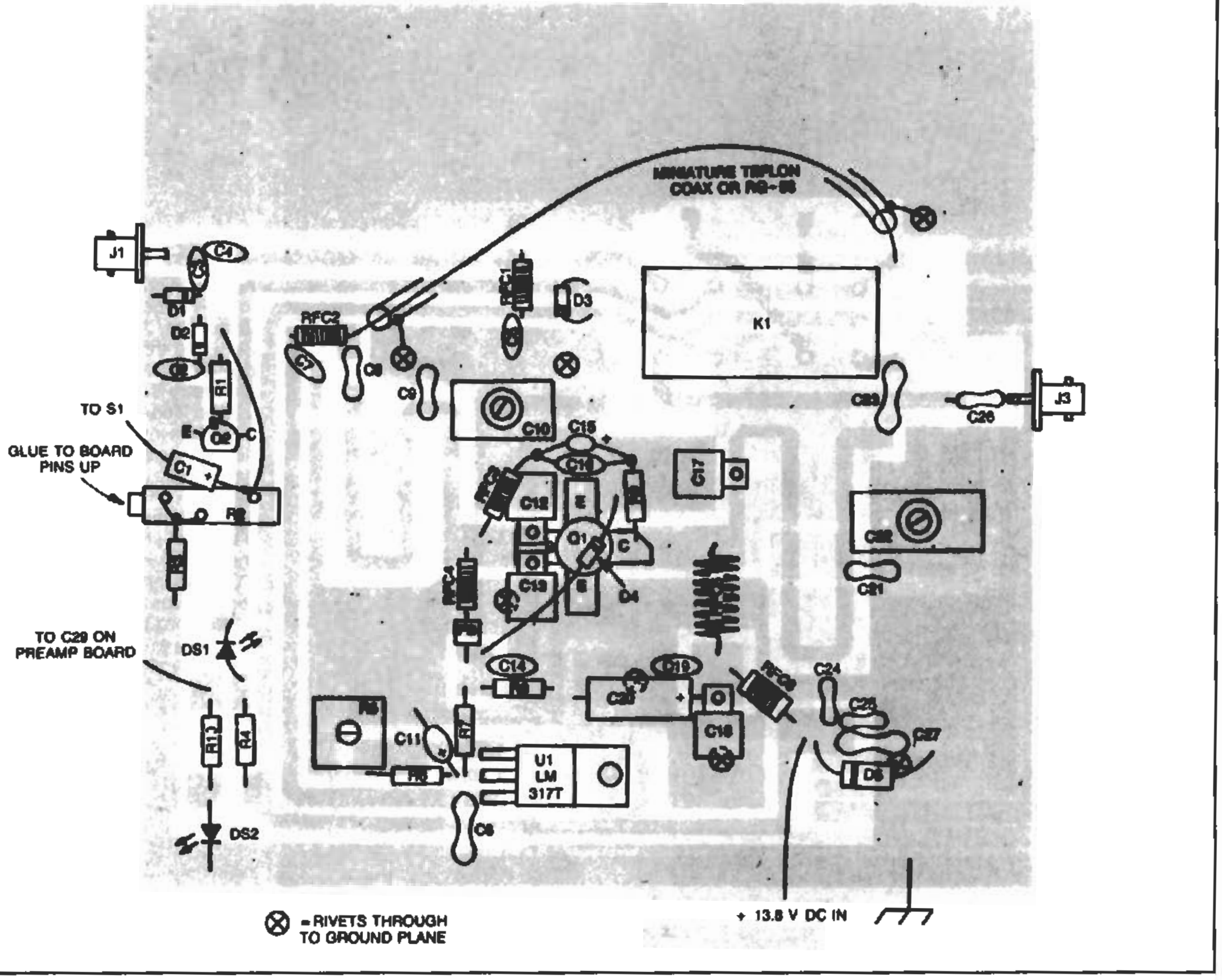

Fig 3—Parts-placement diagram for the VHF amplifiers. All parts are mounted *on the foil-trace side* of the PC board. The power-transistor mounting stud passes through the board and into the heat sink. D5 is epoxied to the ceramic cap of Q1. R2 (10-turn potentiometer) is epoxied to the PC board as indicated, with the pins facing up.0

and used as a spacer between the PC board and heat sink.

Be careful not to crack the body of Q1 where the leads of the device meet it—this can release beryllium-oxide (BeO) dust, *which is lethal*. The leads of the transistor can be bent down slightly to the PC board without affecting RF performance, but never force them in the opposite direction after the transistor has been tightened to the heat sink. Mount the transistor to the heat sink *first*, then solder its leads to the board, and *not vice versa*. Use thermally conductive compound between Q1's mounting surface and the heat sink.

Cut a piece of hobby-store sheet brass about as wide as the bias regulator IC and twice as long as the body of the device. This will be a heat sink for the bias regulator. When mounting the regulator IC, use thermally conductive compound between it and the brass sheet, and between the brass and the main PC board. After tightening the IC hardware to the PC board, bend the brass sheet in a U shape over the IC, leaving enough clearance for heat to escape between the IC and the brass heat sink. This heatsink arrangement can be seen near the bottom center of each amplifier in the lead photograph.

The Preamplifier Circuit

The high gain and low noise figure (NF) of a GaAsFET preamp are not necessary in amplifiers in this power class. I designed the preamp circuit (Fig 5) to use an inexpensive U310 FET. It has more than enough gain to overcome the losses of the amplifier switching circuits and the feed line to the amplifier. The preamp has more than 12 dB gain and noise figure of about 2 dB.

When S2 is closed, K2 closes, activating the preamplifier by applying the dc supply voltage to the drain circuit of Q3 and to the preamp ON LED, DS2. In the transmit mode, K1 switches off the supply voltage to K2, allowing K2 to drop out.

The preamp circuit is optional, and I didn't include it on the main amplifier PC board. If you're not going to build the preamp, solder foil-tape or wire jumpers between the two points on the main amplifier PC board where the preamp connects during assembly.

Building the Preamplifier

The schematic of the preamplifier circuit is shown in Fig 5. As mentioned earlier, although some component values differ between the 2-meter and 220-MHz versions, the two are essentially the same in all other respects. Components that have different values for the two versions are marked with asterisks on the schematic and the parts list. Unlike the main amplifier PC board, most preamp components are mounted through the board and soldered on the bottom side. The PC-board layout and parts-placement diagrams are shown in Figs 6 and 7, respectively.

As with most FET VHF preamplifiers, a shield is necessary between the input inductor and the active device. Solder a piece of scrap brass sheet to the ground foil between these circuit elements. (This preamp shield can be seen in the right-hand amplifier in the lead photo, just behind L1.)

After building the preamplifier, tune it

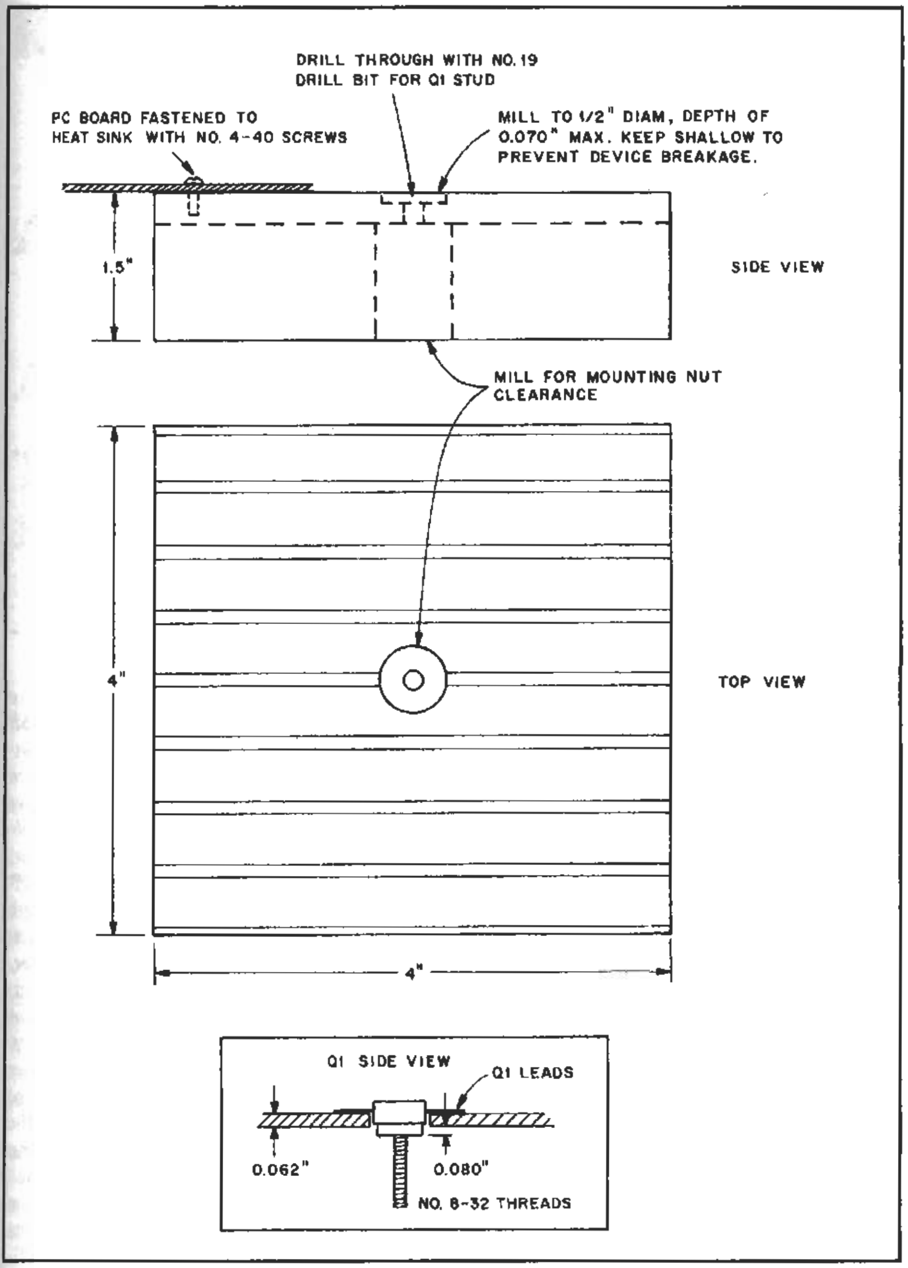

Fig 4—Heat-sink drilling dimensions for the VHF amplifiers. Overall heat-sink size isn't critical, but should be close to the dimensions shown. As an alternative to milling the heat sink to accept Q1, a piece of $^1/_{16}$-in. double-sided PC-board material can be used as a shim. Be careful not to stress the transistor lead-to-body connections during mounting.

for maximum gain with a signal generator or an on-the-air signal. If one is available to you, use a noise-figure meter and tune the preamp for best NF.

Building the Output Filters

I built the output filters for each amplifier on scraps of PC-board material. The construction is shown in the title photo. After building the filters, it's a good idea to use some silicone sealant on the coil (especially in the 2-meter filter) to hold the coil turns in place. Install the filter in a suitable enclosure, or in the amplifier cabinet.

To tune the filters, you'll need a receiver capable of receiving the second harmonic of the fundamental (288 MHz for the 2-meter filter, and 440 MHz for the 220-MHz filter), and a signal generator or an on-the-air signal. Connect the filter between the antenna (or signal generator) and the receiver, and tune C36 for minimum second-harmonic signal level.

Enclosures

I used the main amplifier PC board as the top cover for the unit (with the heat sink mounted on it). I made the sides and bottom cover of the enclosure from PC-board material. The side walls are soldered to the amplifier board, and the bottom cover is attached to the side walls with small brackets (also made of PC-board material) soldered to the bottom cover. Alternatively, the amplifier can be mounted in an aluminum enclosure.

Although I used BNC connectors for RF input and output, N connectors can also be used. Make the connections from the input connector to the PC board with a piece of no. 16 wire or a $^1/_8$-inch-wide strip of brass to the input microstripline. At the output, connect dc-blocking capacitor C26 between the microstripline and the antenna connector. Check the board for proper component placement and good solder joints, and get ready to tune up the amplifier!

Amplifier Tune-Up

As with the amplifier-circuit design and description, the tune-up procedure is very similar for both. The only difference is the exciter you use. Here's how to tune either version.

Disconnect RFC4 from the base of Q1. Connect a voltmeter to the free end of RFC4. Apply 13.8 V dc to the amplifier supply voltage leads. Using the voltmeter, verify that R5 is mounted so that output of U1 increases with clockwise rotation of R5. Turn R5 fully counterclockwise (minimum U1 output voltage). Reconnect RFC4.

Set the quiescent (no-drive) current to Q1 as follows: Disconnect one end of RFC6 and connect an ammeter in series with it. Apply 13.8 V dc to the amplifier through a 5-A fuse. Turn the preamp off. Do not apply RF drive during this adjustment. Using a clip lead, ground the collector of Q2. This should actuate K1. Check the TX LED, DS1, for operation. (If K1 actuates and the TX LED does not light, the LED may be installed in reverse.) Slowly adjust R5 for an idling current (through RFC6) of 75 to 100 mA. The amplifier should be stable; instabilities are indicated by erratic variation of Q1's quiescent current as R5 is adjusted.

Disconnect the ground lead of Q2. Q1's collector current should drop to zero. If the collector current does not drop to zero, the amplifier is unstable. If all is well, remove the ammeter and reconnect RFC6. If the amplifier is unstable, check all bypass capacitors and solder connections.

Apply about 100 mW of drive to the amplifier and check that the COR and delay circuits work properly. Adjust C10 and C22 for maximum power output. Increase drive power and retune for maximum output. After final assembly, tune the amplifier for the desired frequency and power level.

Final Assembly

The preamplifier board can now be installed on the main PC board. Mount the preamp on its edge (refer to the title photo and Fig 3) and solder the input and output microstriplines in place. Solder the ground foil of the preamp board to the ground foil of the main board in a few places to support the preamp board. Connect S2 and DS2 to the preamp board with hook-up wire (see Figs 5 and 6 for connection points). Apply power and check for proper operation of S2, K2 and DS2.

The enclosure can now be painted if you wish. Use masking tape to cover the connectors and heat sink. (Mount the LEDs and switches after painting.) You may want to

Fig 5—Schematic of the preamplifier circuit. K2 is identical to K1 on the main amplifier board. S2 and DS2 are mounted on the front panel of the amplifier. In the parts list below, values for 220-MHz version are given in parentheses where they differ from those of the 2-meter version.

C28, C29, C33—0.01-µF disc.
C30—2-pF silver mica.
C31, C32—20-pF trimmer.
DS2—LED.
D6—1 N4001.
L1, L2—5 turns (4 turns) no. 16 enam, 0.3-in. ID.
K2—Omron LZN2O3-UA-DC12 DPDT relay.
Q3—U310 FET.
R10—1.5-kΩ, ¼ W.
R11—91-Ω, ¼ W.
RFC7, RFC8—0.47-µH molded choke.
S2—Miniature SPDT toggle.

Miscellaneous
1 × ½-in. strip of thin sheet brass

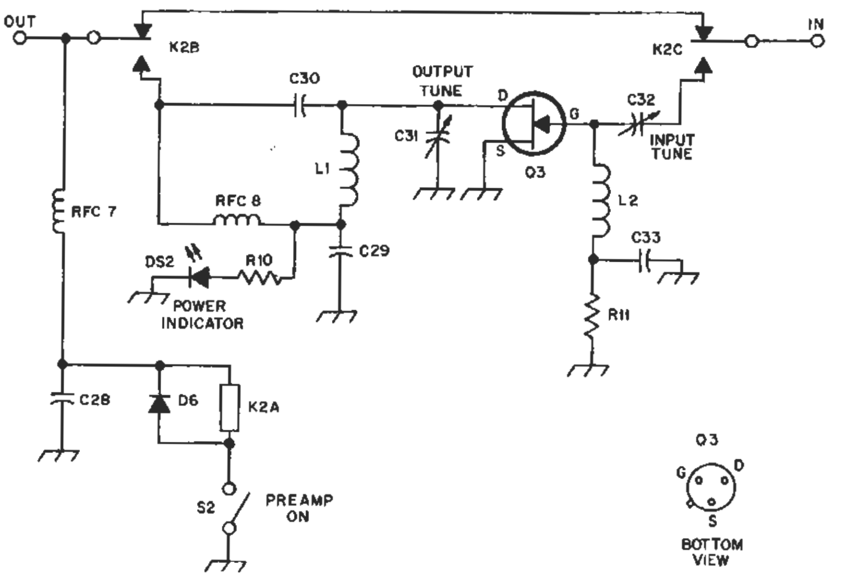

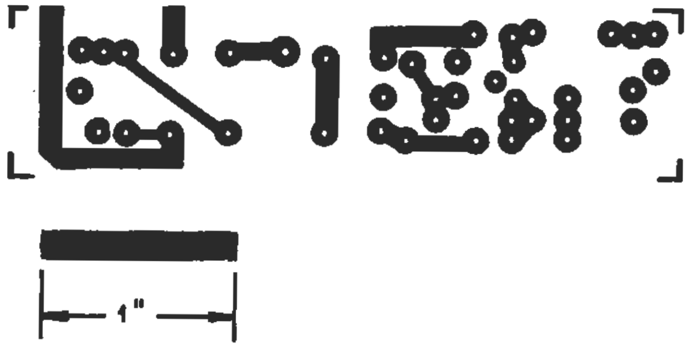

Fig 6—Full-size PC-board layout for the preamplifier circuit. Unlike the power-amplifier PC board, most of the preamplifier components are mounted on the non-trace side of the board.

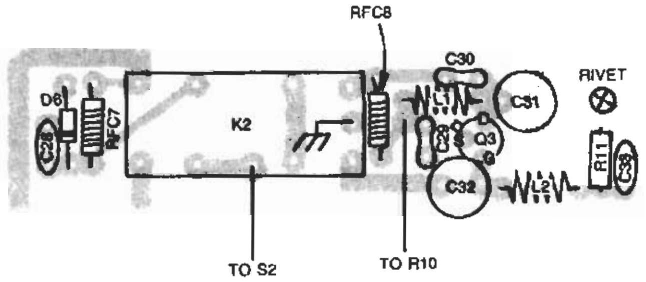

Fig 7—Parts-placement diagram for the VHF preamplifier. Mount all components on the unetched ground-plane side except for C30 and C31. Use a drill to clear the foil away from all mounting holes on the ground-plane side of the board (except for through-board ground connections). Solder all ground connections on both sides of the PC board (see text for rivet-installation instructions). Solder a brass-strip shield to the ground foil between L1 and Q3 on the component side of the PC board.

add stick-on rubber feet to the bottom cover to keep it from sliding around and being scratched.

Amplifier Operation

The 2-meter amplifier works very well with my Kenwood TH-21AT hand-held and ICOM IC-202S SSB/CW transceiver. I use one in my shack and one mobile—both at more than 25 W output. I drive the 220-MHz version of the amplifier with an ICOM IC-3AT (in its high-power mode).

Limit the amplifier power output to 25 W during linear operation. *All SSB amplifiers have a rated linear power output that should not be exceeded, even though the amplifier may be driven above that level. To keep your signal clean, do not overdrive any amplifier.* For exciters with fixed power outputs that would overdrive the amplifier, use the loss of a length of RG-58, or a discrete attenuator, between the radio and the amplifier to prevent overdrive. The gain of the preamplifier (if used) will overcome the attenuator or cable loss during receiving.

When using any amplifier or transverter in the shack, it is a good practice to "hard key" it. Hard keying is simply forcing the amplifier or transverter into the transmit mode with a switch closure or an applied voltage. To do this, run a keying line from the exciter to the amplifier. (Ground the collector of Q2 to hard key these amplifiers.) I added a phono jack to the 220-MHz amplifier to facilitate hard keying. If a positive voltage is used for keying, a 2N2222 transistor inverter can be used between the transceiver and the collector of Q2 (see the inset in Fig 1).

Summary

These amplifiers have served me well in the car and at home. After building one of these amplifiers, you'll probably find what I did: They're so handy to have around, and so easy to build, that you can't build just one!

Notes

[1] Thomson Components/Mostek Corp, Semiconductor Division, Commerce Dr, Montgomeryville, PA 18966. Thomson transistors are available through RF Gain, Ltd, 100 Merrick Rd, Rockville Center, NY 11570, tel 516-536-8868 or 800-645-2322.

[2] The following parts are available from Frontier Microwave, RD 1 Box 467, Ottsville, PA 18942: 100+ tinned rivets: $2; relays (K1, K2): 2 for $8. Prices include shipping. The ARRL and *QST* in no way warrant this offer.

By David D. Meacham, W6EMD **From *QST*, April 1984**

A High-Power 2-Meter Amplifier Using the New 3CX800A7

Tired of hearing exotic DX on 2 meters but not being able to work it? Build this amplifier, and you'll transform your pip-squeak signal into a real rockcrusher.

The 3CX800A7 triode recently announced by Varian EIMAC has a plate dissipation rating of 800 W and modest cooling requirements. Its oxide cathode gives high emission with only 20 W of heater power. With full ratings up to 350 MHz, this tube is a scaled-up version of the 8874 — a proven performer in amateur and commercial gear.

The amplifier described here is based on a design presented by Raymond F. Rinaudo, W6ZO, in January 1972 QST. Most of the changes in the new amplifier (Fig. 1) are designed to accommodate the larger size of the 3CX800A7 and its attendant higher capacitances and currents. A plate-current meter is not included because my station power supply is already well metered. However, there is room on the front panel of the amplifier for a plate-current meter if another builder wishes to add one. Other changes to Rinaudo's design are minor.

Construction

The amplifier chassis is mounted on a standard 19-inch wide, 5¼-inch-high aluminum rack panel (Bud No. SFA-1833).[1] A 5 × 13 × 3-inch aluminum chassis (Bud No. AC-422) is spaced 1¾ inches behind the panel by two aluminum end brackets. A 3½ × 4½ × 1-inch aluminum chassis (Bud No. AC-1402) houses the input circuitry and is mounted on the larger chassis between it and the front panel. The right-hand end bracket has a large lip on the rear for mounting the heater transformer and connectors. Fig. 2 shows mechanical details of the chassis and end brackets. The cabinet chosen is a lightweight aluminum unit made by TenTec (No. 19-0525).

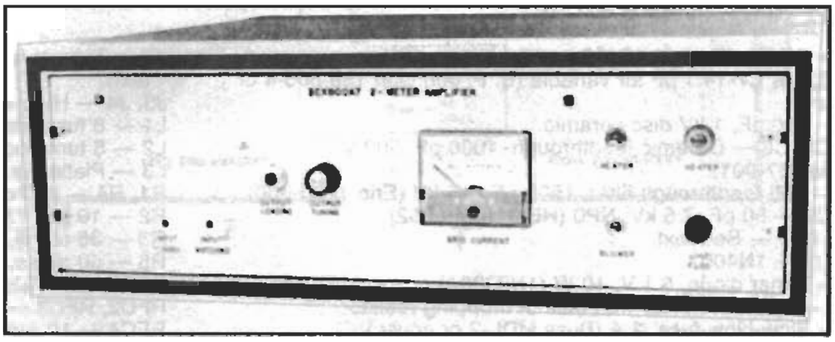

Input Circuit

In the cathode-driven configuration, the input impedance of the 3CX800A7 appears as a nominal capacitance of 26.5 pF in parallel with a resistive component that varies with operating conditions but is typically about 49 ohms. The computerdesigned input circuit of this amplifier operates at a loaded Q of about 2.1. It can be set anywhere in the 2-meter band for a VSWR of less than 1.3:1. C1 and C2 are predominantly matching and tuning controls, respectively; there is some interaction between them, however. When the input is tuned at 144.5 MHz, the input VSWR will he less than 1.4:1 from 144 to 145 MHz.

The tube socket is an EIMAC SK-1900 or a Johnson part No. 124-311-100. Its center is mounted 2 inches from the end of the input box, adjacent to the mounting bracket. The tube pins and input circuit are cooled by a small amount of air admitted to the input box from the pressurized output box. Three holes, made with a no. 50 drill, provide adequate flow. These holes are spaced in a close triangle and are located diagonally across from the variable capacitors. Air exhausts through the tuning holes. Fig. 3 shows details of the input circuitry.

The heater circuit includes two chokes, feedthrough capacitors, the filament transformer, a voltage-dropping resistor and a switch. The chokes are wound with a turn-to-turn spacing of about one-half the wire diameter. They are self-resonant (parallel resonant) just above the 2-meter band. Nominal heater voltage for the 3CX800A7 is 13.5 V. The closest available commercial transformer has a 14.0-V secondary, so R3 is used in the primary. Switching is set up so that the blower must be on before the heater can be activated. Conversely, this arrangement allows the blower to be left on after switching off the heater—a highly recommended practice.

Cathode bias is provided by a 5.1-V Zener diode, D3. R1 prevents the cathode voltage from soaring if the Zener fails. F2 will blow if excessive cathode current is drawn. R2 nearly cuts off plate current on receive. Because the grid is at dc ground, the negative supply lead must be kept above

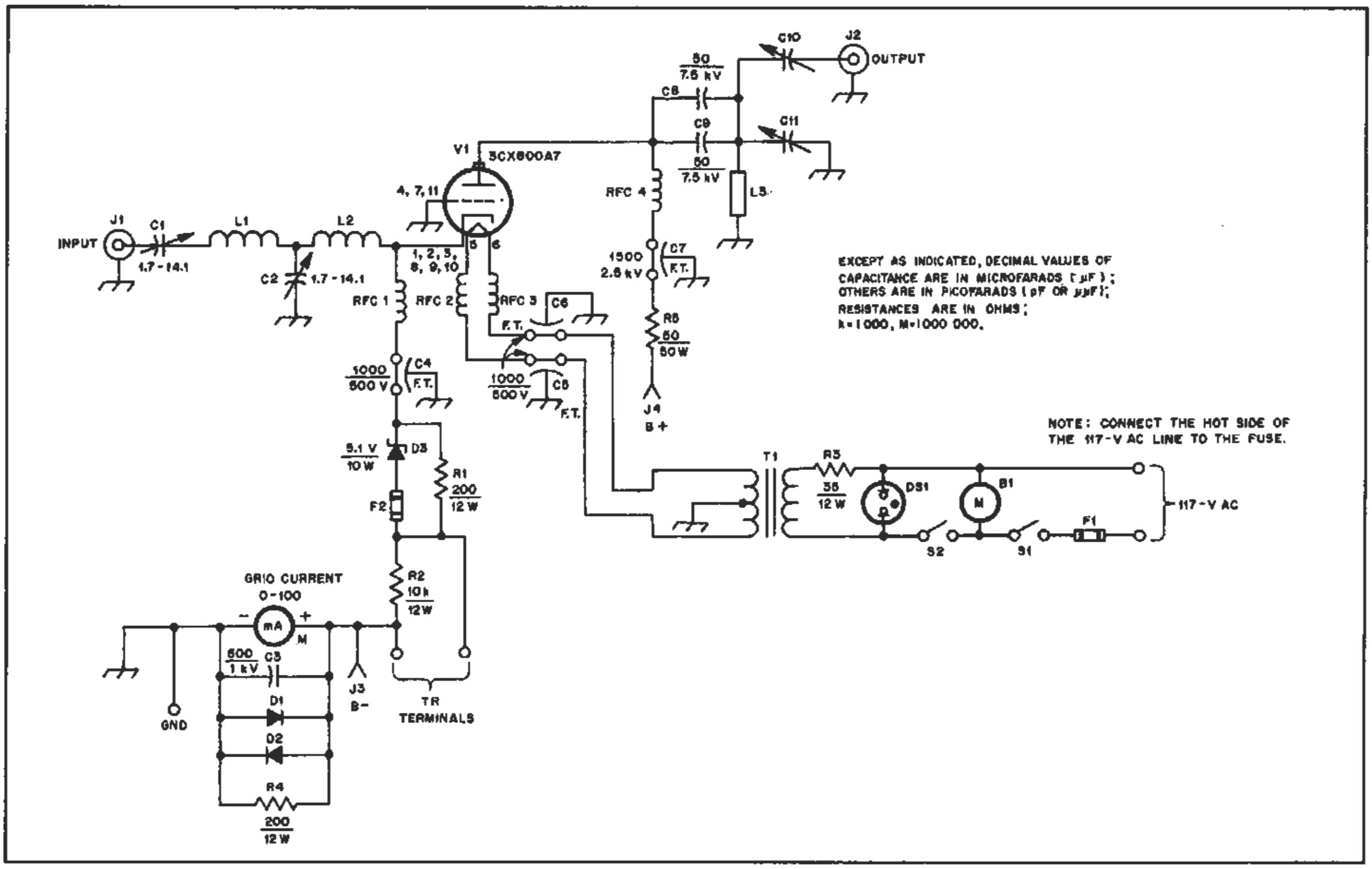
C10
J2
OUTPUT
50
7.5 kV
C8
C9
C11
V1
3CX800A7
50
7.5 kV
RFC 4
L3
4, 7, 11
J1
INPUT
C1
1.7-14.1
L1
L2
C2
1.7-14.1
RFC 1
RFC 2
1, 2, 3, 8, 9, 10
5
6
RFC 3
C6
F.T.
1500
2.5 kV
C7
F.T.
R5
50
50 W
J4
B+
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.
1000
500 V
C4
F.T.
1000
500 V
C5
F.T.
NOTE: CONNECT THE HOT SIDE OF THE 117-V AC LINE TO THE FUSE.
5.1 V
10 W
D3
R1
200
12 W
F2
T1
R3
35
12 W
DS1
B1
M
117-V AC
F1
S2
S1
GRID CURRENT
0-100
mA
M
R2
10 k
12 W
500
1 kV
C3
J3
B-
TR
TERMINALS
GND
D1
D2
R4
200
12 W

Fig. 1 — Schematic diagram of the 3CX800A7 amplifier.

B1 — Blower (Dayton 4C012 or equiv.); see text.
C1, C2 — 1.7-14.1 pF air variable (E. F. Johnson 189-505-4 or equiv.).
C3 — 500.pF, 1.kV disc ceramic.
C4, C5, C6 — Ceramic feedthrough, 1000 pF, 500 V (Erie 357-001).
C7 — EMI feedthrough filter, 1500 pF, 2.5 kV (Erie 1280-060).
C8, C9 — 50 pF, 7.5 kV, NP0 (HEC HT-50 852).
C10, C11 — See text.
D1, D2 — 1N4001.
D3 — Zener diode, 5.1 V, 10 W (1N3996A).
DS1 — NE51 in holder with built-in dropping resistor.
F1 — Slow-blow fuse, 2 A (Buss MDL-2 or equiv.).
F2 — 1.5 A fuse (Buss AGC 1$^{1}/_{2}$ or equiv.).
J1 — BNC bulkhead feedthrough connector (UG-492 A/U).
J2 — Type N female connector (part of C10 assembly — see text).
J3, J4 — High-voltage connector (Millen 37001 or equiv.).
L1 — 8 turns no. 16 tinned wire, $^{3}/_{8}$-in ID, $^{11}/_{16}$ in long.
L2 — 6 turns no. 16 tinned wire, $^{3}/_{8}$-in ID, $^{11}/_{16}$ in long.
L3 — Plate line, 1$^{5}/_{16}$ in wide, 8$^{3}/_{16}$ in long. See text and Fig. 4.
R1, R4 — 200 ohms, 12 W.
R2 — 10 kΩ, 12 W.
R3 — 35 ohms, 12 W.
R5 — 50 ohms, 50 W.
RFC1 — 1.0 µH choke, 300 mA (Miller 4602).
RFC2, RFC3 — 11 turns no. 20 HF wire on $^{3}/_{8}$-in Noryl rod.
RFC4 — 10 turns no. 16 tinned wire, $^{1}/_{2}$-in ID, 1$^{3}/_{16}$ in long.
T1 — Filament transformer, 14.0-V, 2-A secondary (Triad F-215X or equiv.).

ground for grid-current metering by M1. R4 keeps the negative side of the plate supply from rising if M1, D1 and D2 all open up. D1, D2 and C3 protect the meter from transients and RF voltages.

Output Circuit

The output tank circuit is a silver-plated quarter-wave strip line (Fig. 4) foreshortened by the tube, loading and tuning capacitances at, and near, its open end. It operates at a loaded Q of approximately 20. The silver-plated anode collet (Fig. 4) is made of 0.062-inch-thick brass sheet with Tech-Etch 134B finger stock soldered on the inside. It is supported by two Teflon® standoffs $^{1}/_{2}$-inch in diameter and 1 inch long. At the far end of the line, a silver-plated shorting block contacts the chassis and the strip line. This block (Fig. 4) is made from $^{7}/_{16}$-inch-thick brass 1$^{15}/_{16}$ inches wide and 1 inch high. EIMAC CF-800 finger stock is screwed and then soldered on both top and bottom. The chassis and stripline are both slotted to allow $^{5}/_{16}$ inch of shorting-block travel to set the tuning range of the capacitive-tuning paddle. A 1-inch-long Teflon standoff supports the center of the line.

The tuning (C11) and loading (C10) paddles are 1$^{3}/_{16}$-inch-diameter discs made from silver-plated, $^{1}/_{16}$-inch-thick brass. They are spaced 1$^{3}/_{8}$ inches center-to-center. The output loading paddle is nearest the tube, and its center is 1$^{13}/_{16}$ inches from the tube cooler surface. Spacing between the paddle and line during operation is about 0.135 inch. The output tuning boss, EIMAC part No. 720362, is tapped for a $^{1}/_{4}$-28 threaded rod. The output loading control and output connector are two separate EIMAC assemblies combined into one unit. They are available from EIMAC as Support Assembly No. 720361 and Sliding Probe Assembly No. 720407. For this amplifier, the outer conductor was lengthened to reach the front panel. The grid collet is also available from EIMAC as part No. 720359.

In addition to RFC4 and feedthrough capacitor C7, an energy-absorbing resistor (R5) is wired in series with the dc plate supply. This resistor will protect the tube and power supply in the event of a highvoltage arc. It provides necessary protection while dissipating only 12.5 W at 500-mA dc plate current.

Cooling

The blower specified provides a measured 25 CFM airflow through the output box and tube cooler at 0.42 inch of water-column static pressure. This amount of cooling is sufficient for 800 W of plate dissipation at sea level with inlet air temperatures up to 35° C. It is adequate for the same dissipation at 5000 feet of altitude with inlet-air temperatures up to 25° C.

Not evident from the photographs is the care taken to separate outgoing hot air from incoming cool air. This is accomplished by the addition of a simple dividing wall inside the cabinet running from the rear chassis cover of the output box to the rear of the cabinet. The material used is rigid fiberglass insulation — the kind you cut with a saw. It provides a measure of blower-noise suppression in addition to assuring cool inlet air.

Hot air leaving the anode cooler is directed to a homemade "honeycomb" RFI filter through a chimney made from rolled-up Teflon sheet. The filter is made of brass tubing and sheet with the honeycomb material soldered in place. See Fig. 5 for construction details. This filter acts as a waveguide-beyond-cutoff having high attenuation at 144 MHz.

The equation for calculating attenuation by this type of filter is

$$A_a = 32\ D/d \qquad \text{(Eq. 1)}$$

where A_a = aperture attenuation (dB), D = length of pipe, and d = inside diameter of pipe. In this case, the "pipe" is each cell of the honeycomb. The basic material for the honeycomb is cadmium-plated brass heat-radiator core.[2] Each hexagonal-shaped cell has a width between flat sides (diameter) of 0.100 inch. The material is ½-inch thick. Plugging these values into the equation yields an A_a of 160 dB—more than enough! From a practical standpoint, it is usually sufficient to make cell length at least three times the nominal cell diameter.

Tune-up

Initial work may be done with a dip meter, particularly on the output circuit. With the tube in place, but with no voltages applied, the shorting block on the end of the plate line should be set to give a paddle-tuning range that straddles the desired operating frequency. Bear in mind that when the tube is "hot," the resonant frequency will be somewhat lower than it is without electron flow.

The input-tuning capacitor can be dip-meter resonated with the matching capacitor set at one-half mesh for a start. Further work here must be done "hot" with a VSWR measuring device on the input.

Set the initial spacing between the output paddle and the plate line to ⅛-inch. Connect the output to a 50-ohm dummy load capable of handling at least 700W at 144 MHz through an accurate VHF wattmeter, such as a Bird model 43. Connect a driver capable of delivering about 20 W to the input through a VSWR-measuring device.

The heater of the 3CX800A7 should be run for at least three minutes before applying plate voltage. After the warm-up, short the TR terminals and apply about 1000 V to the plate. This should result in a small amount of idling plate current. Next, apply enough drive to produce a rise in plate current, and adjust the plate tuning for a peak in output power. Now tune and match the input circuit for minimum input VSWR. Next, adjust the output loading paddle for maximum power output while keeping the plate current dipped with the plate tuning paddle. Apply full plate voltage and higher drive power, and then repeak the output tuning and loading controls. Touch up the input tuning and matching, and the amplifier is ready for service.

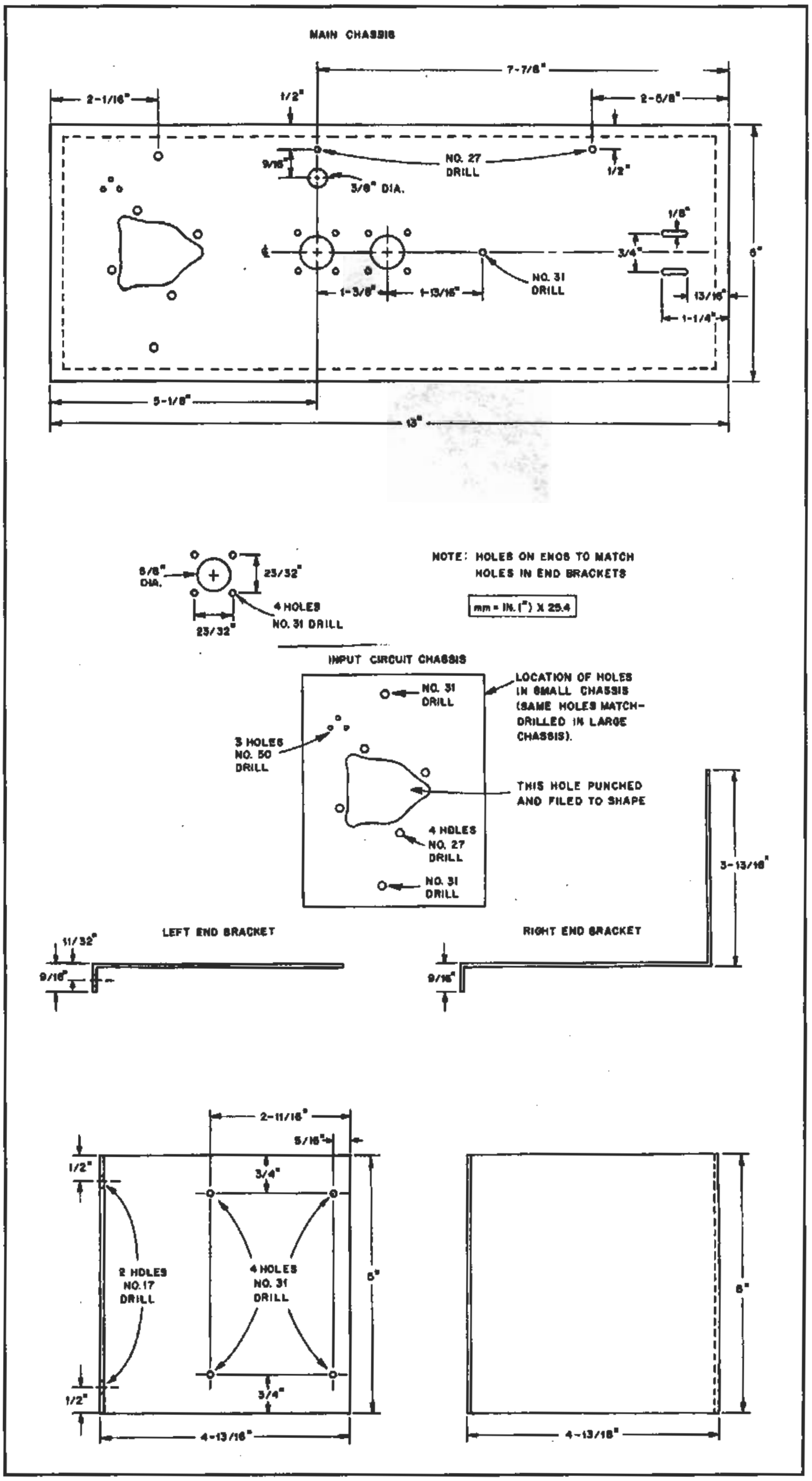
MAIN CHASSIS
7-7/8"
2-1/16"
1/2"
2-5/8"
NO. 27 DRILL
9/16"
1/2"
3/8" DIA.
1/8"
3/4"
6"
NO. 31 DRILL
1-3/8"
1-13/16"
13/16"
1-1/4"
5-1/8"
13"
5/8" DIA.
23/32"
4 HOLES NO. 31 DRILL
23/32"
NOTE: HOLES ON ENDS TO MATCH HOLES IN END BRACKETS
mm = IN. (") X 25.4
INPUT CIRCUIT CHASSIS
NO. 31 DRILL
LOCATION OF HOLES IN SMALL CHASSIS (SAME HOLES MATCH-DRILLED IN LARGE CHASSIS).
3 HOLES NO. 50 DRILL
THIS HOLE PUNCHED AND FILED TO SHAPE
4 HOLES NO. 27 DRILL
NO. 31 DRILL
3-13/16"
LEFT END BRACKET
RIGHT END BRACKET
11/32"
9/16"
9/16"
2-11/16"
5/16"
1/2"
3/4"
2 HOLES NO.17 DRILL
4 HOLES NO. 31 DRILL
5"
6"
1/2"
3/4"
4-13/16"
4-13/16"

Fig 2—Mechanical details of the chassis and end brackets. The end brackets are made from 0.032-in aluminum, 5052 alloy or softer.

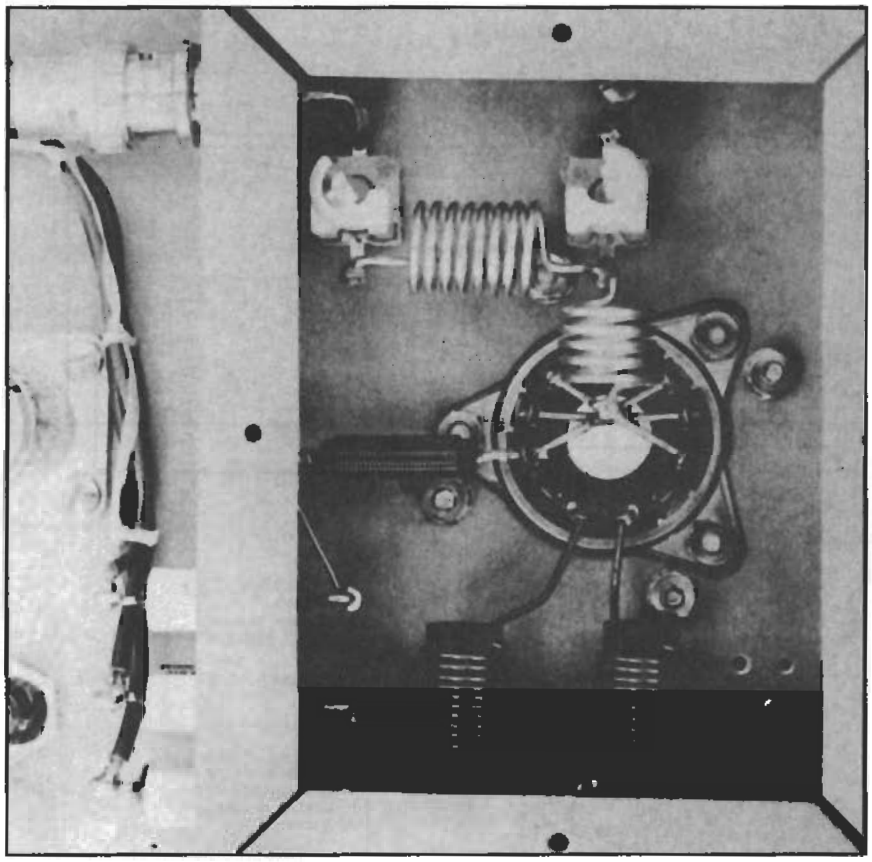

Fig 3—Close-up of the input circuitry and tube-socket wiring. Keep all leads as short as possible.

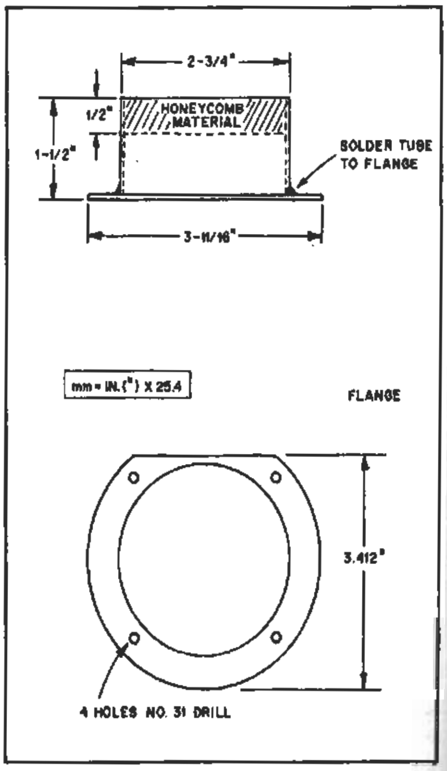
2-3/4"
1/2"
HONEYCOMB MATERIAL
1-1/2"
SOLDER TUBE TO FLANGE
3-11/16"
mm = IN.(") X 25.4
FLANGE
3.412"
4 HOLES NO. 31 DRILL

Fig 5—Mechanical details of the exhaust air RFI filter. The honeycomb material is soldered inside a 2¾-in OD, 0.065-in wall brass tube. The flange, which bolts to the chassis, is made from 0.125-in-thick brass sheet.

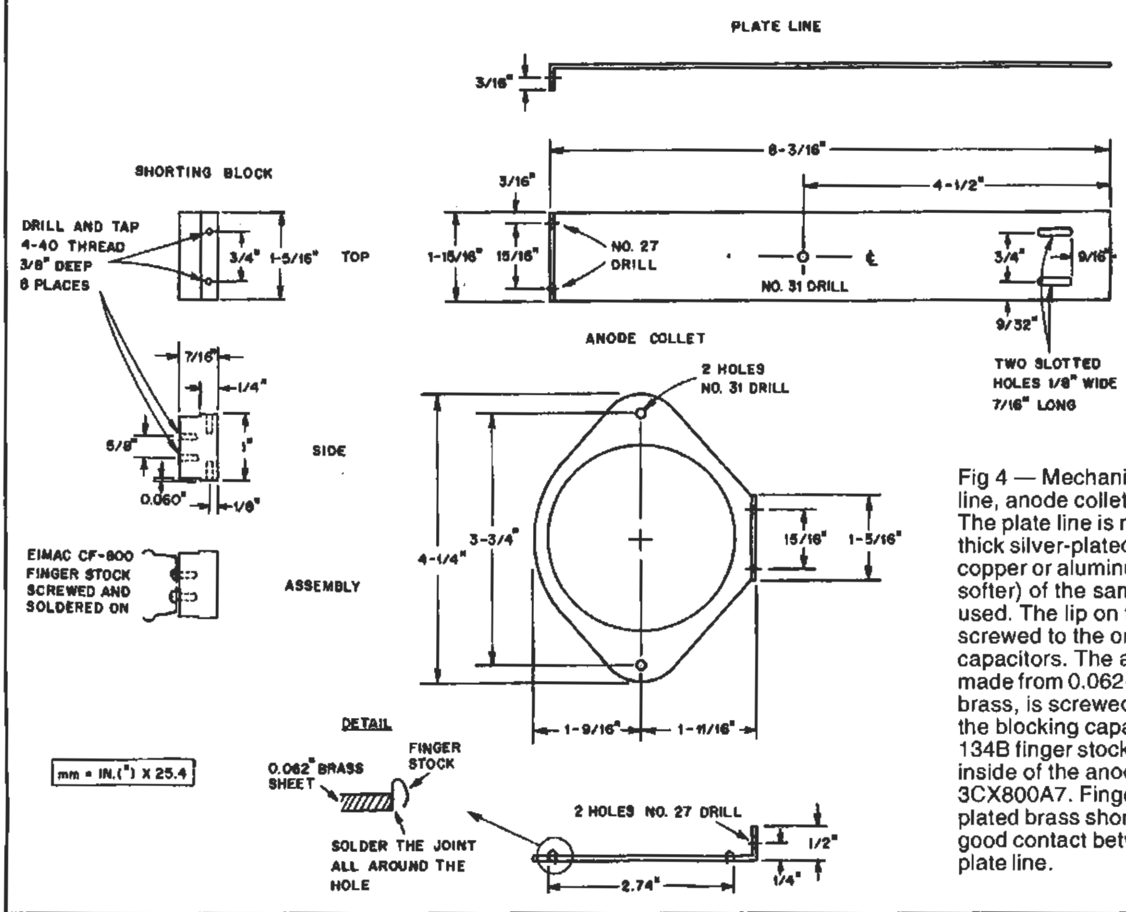
PLATE LINE
3/16"
6-3/16"
4-1/2"
3/16"
1-15/16"
15/16"
NO. 27 DRILL
NO. 31 DRILL
℄
3/4"
9/16"
9/32"
TWO SLOTTED HOLES 1/8" WIDE 7/16" LONG
SHORTING BLOCK
DRILL AND TAP 4-40 THREAD 3/8" DEEP 8 PLACES
3/4"
1-5/16"
TOP
7/16"
1/4"
5/8"
1"
0.060"
1/8"
SIDE
EIMAC CF-800 FINGER STOCK SCREWED AND SOLDERED ON
ASSEMBLY
ANODE COLLET
2 HOLES NO. 31 DRILL
4-1/4"
3-3/4"
15/16"
1-5/16"
1-9/16"
1-11/16"
DETAIL
0.062" BRASS SHEET
FINGER STOCK
SOLDER THE JOINT ALL AROUND THE HOLE
2 HOLES NO. 27 DRILL
1/2"
1/4"
2.74"
mm = IN.(") X 25.4

Fig 4 — Mechanical details of the plate line, anode collet and shorting block. The plate line is made from 0.032-in-thick silver-plated brass, although copper or aluminum (5052 alloy or softer) of the same thickness may be used. The lip on the plate line is screwed to the one end of the blocking capacitors. The anode collet, which is made from 0.062-in-thick silver-plated brass, is screwed to the other ends of the blocking capacitors. Tech-Etch no. 134B finger stock soldered around the inside of the anode collet grips the 3CX800A7. Finger stock on the silver-plated brass shorting block ensures good contact between the chassis and plate line.

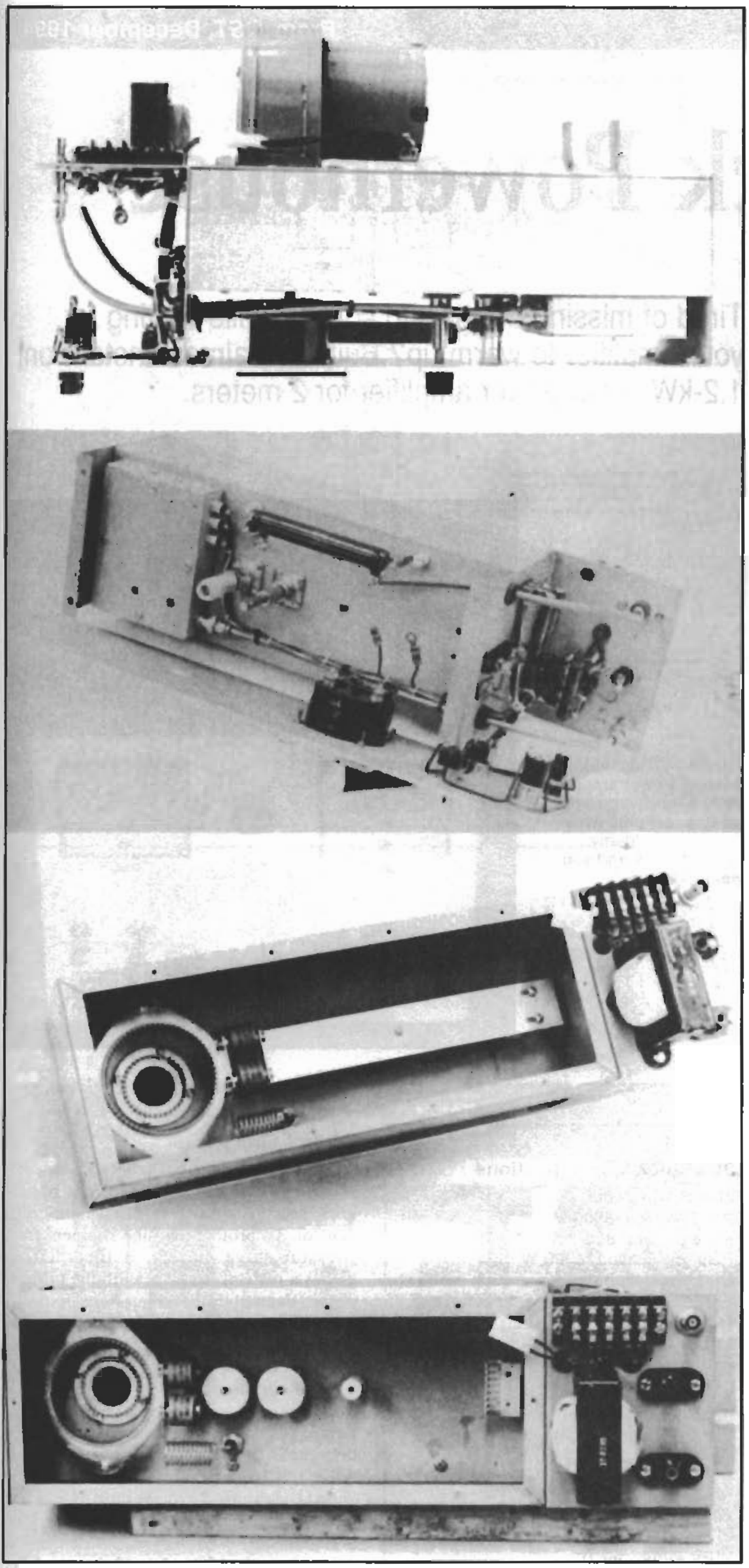

Fig 6—Various views of the completed 3CX800A7 amplifier.

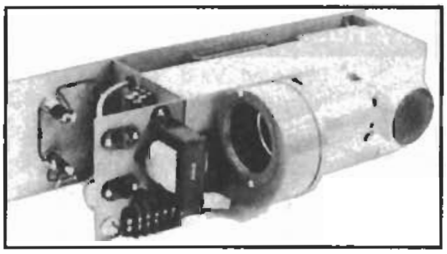

Fig 7—Rear view of the amplifier with the plate-compartment cover in place. Note the honeycomb RFI filter at the hot-air exhaust.

Table 1

Operating Conditions for the 3CX800A7 Amplifier

Plate supply voltage	2200 V
Zero-signal plate current*	65 mA
Single-tone plate current	500 mA
Grid bias (Zener bias)	–5.1 V
Single-tone grid current*	40 mA
Driving power	18.5 W
Output power	707 W
Gain	15.8dB
Efficiency**	64%

*Values may vary considerably from tube to tube.

**Actual tube efficiency is about one percent higher because of power loss in the 50-ohm series resistor in the plate lead (R5).

A few words of caution are in order. Remember that the heater voltage must never be applied without the blower running, and that the heater must warm up at least three minutes before applying plate voltage. Never exceed 60-mA dc grid current, even during tune-up. Also, because of the relatively low grid dissipation of the 3CX800A7, RF drive must never be applied unless plate voltage is applied to the tube and a suitable load is connected to the output. Following these simple rules will substantially increase tube life and amplifier reliability.

Typical Operating Conditions

With Zener-diode bias, the 3CX800A7 is best operated in Class AB2 for linear service. The data in Table 1 represent measured performance in linear service at 144 MHz. Complete data sheets are available from EIMAC. This amplifier is easily capable of conservative operation at 700-W output. Two of these tubes will run at the 1500-W output legal limit.

This amplifier uses straightforward construction techniques and is easily duplicated. Any builder will be rewarded with a great deal of satisfaction and a reliable amplifier.

Notes

[1] mm = in x 25.4; m = ft x 0.3048.

[2] One source for the honeycomb radiator core material is Technit West Division, 320 North Nopal St., Santa Barbara, CA 93103.

By Russ Miller, N7ART From *QST*, December 1994

The Quick Powerhouse

Tired of missing a new grid square while waiting for your amplifier to warm up? Build this almost-instant-on, 1.2-kW linear power amplifier for 2 meters.

It's that typical, early summer afternoon, when, out of the blue comes co-channel interference to the low-band (channels 2 through 6) TV channels. Time to turn on the 2-meter receiver. Whoops! There's a sporadic-E opening, and stations are booming in from over 1000 miles away. Quickly, I hit the switches to warm up my linear amplifier. Time marches on, and now the amp is ready—only took 3 or 4 minutes. But that juicy, far-away clatter via sporadic E is gone! Woe is me!

Having had this happen more than once, I was delighted to find that Eimac was producing a new tube with an almost-instant-warm-up filament that functions well at 2 meters—the 3CX1200Z7.[1] The Z7 is different from the 3CX1200A7 by virtue of its external grid ring, redesigned anode assembly and a 6.3-V ac filament. One advantage to the 3CX1200Z7 is the wide range of plate voltages that can be used, from 2000 to 5500. This amplifier looks much like the easily duplicated W6PO design,[2] except for the plate collet and the addition of some control circuitry. The plate collet is adapted from the W6PO 222-MHz design.[3] The RF deck is a compact unit, designed for table-top use (Figure 1). Table 1 gives some data on the 3CX1200Z7; Table 2 lists CW operating performance for this amplifier.

Input Circuit

The tuned-filament T network matches the 50-Ω drive source to the filament input impedance, providing a very low input SWR. Tuning is easy and docile. Grid bias is provided by an 8.2-V, 50-W Zener diode. Cutoff bias is provided by a 10-kΩ, 25-W resistor. A relay on the control board shorts out the cutoff-bias resistor, to place the amplifier in the TRANSMIT mode.

I didn't use a tube socket. Instead, I bolted the tube directly to the top plate of the subchassis, using the four holes (drilled to clear a #6 screw) in the grid flange. Connections to the heater pins are via drilled and slotted brass rods. The input circuit is contained within a 3½×6×7¼-inch (HWD) subchassis (Figure 3).

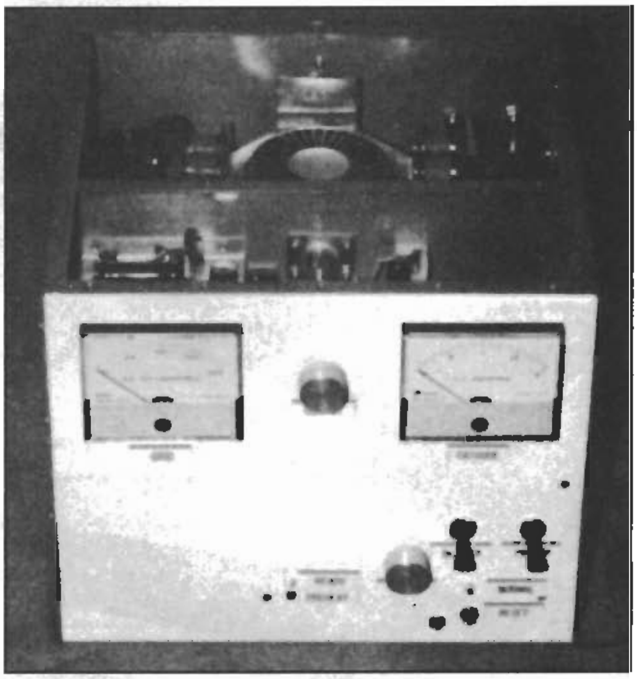

Figure 1—This table-top 2-meter power amplifier uses a quick-warm-up tube, a real plus when the band suddenly opens for DX and you want to join in.

Table 1
3CX1200Z7 Specifications

Maximum Ratings

Plate voltage: 5500 V
Plate current: 800 mA
Plate dissipation: 1200 W
Grid dissipation: 50 W

Table 2
CW Operating Data

Plate voltage: 3200 V
Plate current (operating): 750 mA
Plate current (idling): 150 mA
Grid current: 165 mA
DC Power input: 2400 W
RF Power output: 1200 W
Plate dissipation: 1200 W
Efficiency: 50%
Drive power: 85 W
Input reflected power: 1 W

Control Circuit

The control circuit (Figure 4) is a necessity. It provides grid overcurrent protection, keying control, and filament surge control. To protect the tube filament from stressful surge current, a timer circuit places a resistor in series with the primary of the filament transformer. After four seconds, the timer shorts the resistor, allowing full filament voltage to be applied. C2 and R4 establish the time delay.

Another timer inhibits keying for a total of 10 seconds, to give the internal tube temperatures a chance to stabilize. C1 and R3 determine the time constant of this timer. After 10 seconds, the amplifier can be keyed by grounding the keying line. When the amplifier is not keyed, it draws no plate current. When keyed, idle current is approximately 150 mA, and the amplifier only requires drive to produce output. A

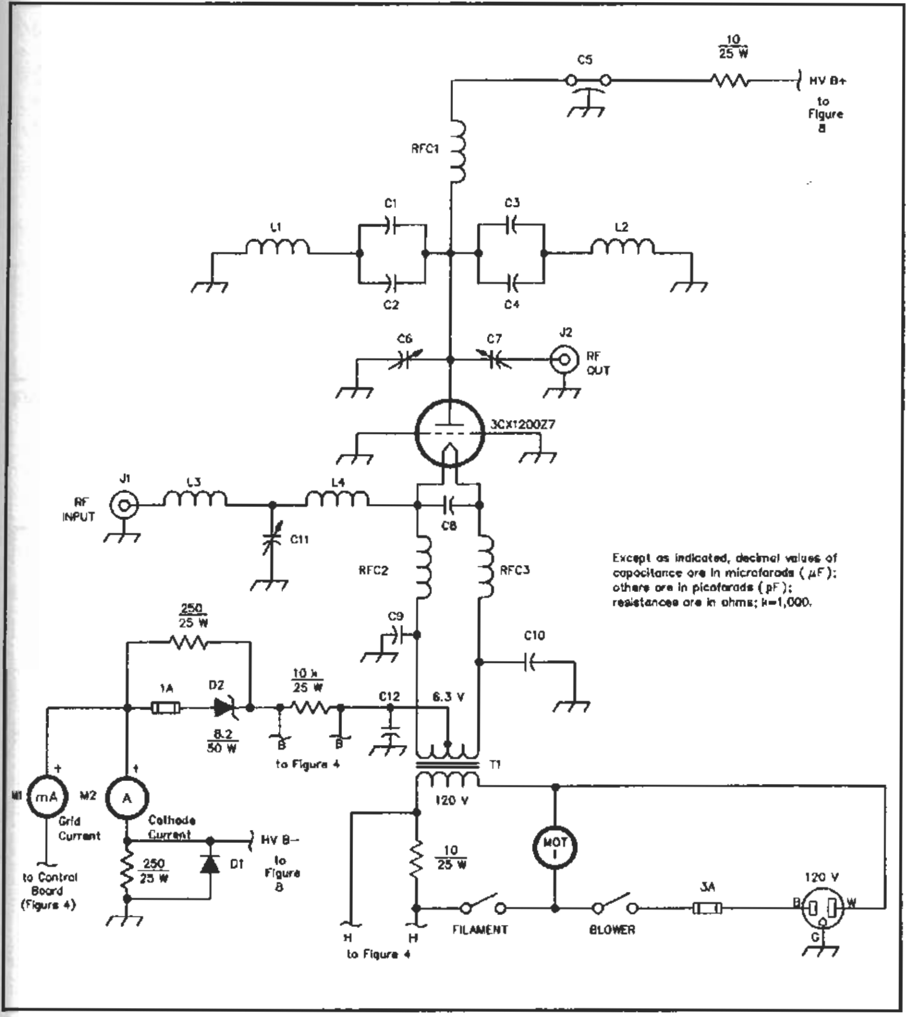
C5
10
25 W
HV B+
to
Figure
8
RFC1
C1
C3
L1
L2
C2
C4
C6
C7
J2
RF
OUT
3CX1200Z7
J1
L3
L4
RF
INPUT
C8
C11
RFC2
RFC3
250
25 W
C9
C10
10 k
25 W
1A
D2
C12
6.3 V
8.2
50 W
B
B
to Figure 4
T1
M1
mA
M2
A
Grid
Current
Cathode
Current
HV B−
to
Figure
8
250
25 W
D1
to Control
Board
(Figure 4)
120 V
MOT
1
10
25 W
3A
120 V
B
W
G
H
H
to Figure 4
FILAMENT
BLOWER

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000.

Figure 2—Schematic diagram of the 2-meter amplifier.

C1-C4—100 pF, 5 kV, type 850
C5—1000 pF, 5 kV
C6—Anode-tuning capacitor; see text and Figure 5 for details
C7—Output-loading capacitor; see text and Figure 7 for details
C8-C10—1000-pF silver mica, 500 V
C11—30-pF air variable
C12—0.01 µF, 1 kV
D1—1000 PIV, 3-A diode, 1N5408 or equiv
D2—8.2-V, 50-W Zener diode, ECG 5249A
J1—Chassis-mount BNC connector
J2—Type-N connector fitted to output coupling assembly (see Figure 7)
L1, L2—Plate lines; see text and Figure 6 for details
L3—5 t no. 14, $^{1}/_{2}$-inch diameter, close wound
L4—3 t no. 14, $^{5}/_{8}$-inch diameter, $^{1}/_{4}$-inch spacing
RFC1—7 t no.14, $^{5}/_{8}$-inch diameter, $1^{3}/_{8}$ inch long
RFC2, RFC3—10 t no. 12, $^{5}/_{8}$-inch diameter, 2 inches long
T1—Filament transformer. Primary: 120 V; secondary: 6.3 V, 25 A, center tapped Available from Avatar Magnetics (Ronald C. Williams, W9JVF, 240 Tamara Trail, Indianapolis, IN 46217, 317-783-1211); part number AV-539
M1—Grid milliammeter, 200 mA dc full scale
M2—Cathode ammeter, 2 A dc full scale
MOT1 —140 free-air cfm, 120-V ac blower, Dayton 4C442 or equivalent.

Sources for some of the "hard to get parts" include:

Fair Radio Sales, 1016 E Eureka, Lima, OH 45802, tel 419-227-6573;
Surplus Sales of Nebraska, 1502 Jones Street, Omaha, NE 68102, tel 402-346-4750.

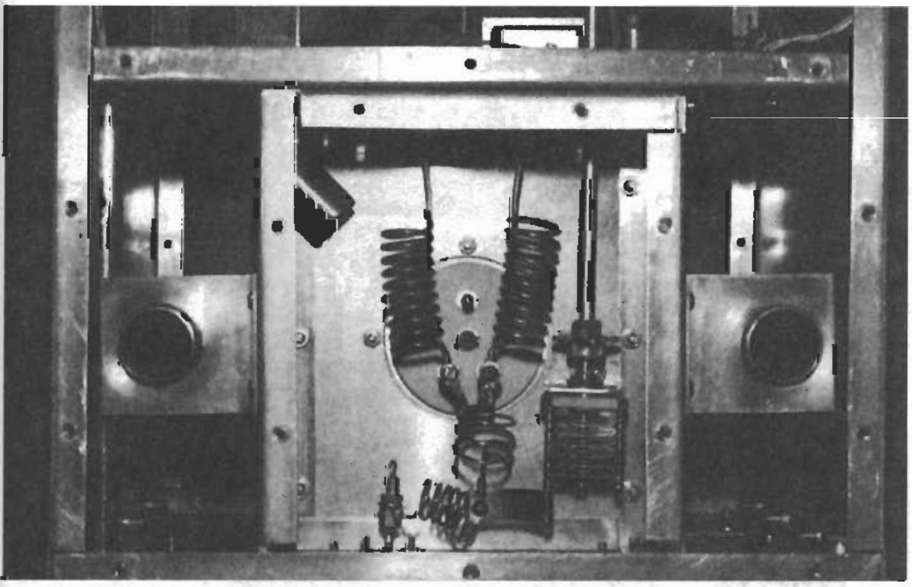

Figure 3—This view of the cathode-circuit compartment shows the input tuned circuit and filament chokes.

safety factor is built in: the keying circuit requires +12 V from the high-voltage supply. This feature ensures that high voltage is present before the amplifier is driven.

The grid overcurrent circuit should be set to trip if grid current reaches 200 mA. When it trips, the relay latches and the NORMAL LED extinguishes. Restoration requires you to press the RESET switch.

Plate Circuit

Figure 5 shows an interior view of the plate compartment.The anode collet is patterned after the one used in the W6PO 222-MHz amplifier. The differences are small. A 4×$2^{1}/_{4}$-inch tuning capacitor plate and a 2×2-inch output coupling plate are centered on the collet. These parts are the same size and shape as those used on the 2-meter W6PO amplifier. The remaining difference is the diameter of the hole for the 3CX1200Z7 anode. Sufficient clearance must be left for the fingerstock. The hole diameter will be approximately $3^{5}/_{8}$ inches.

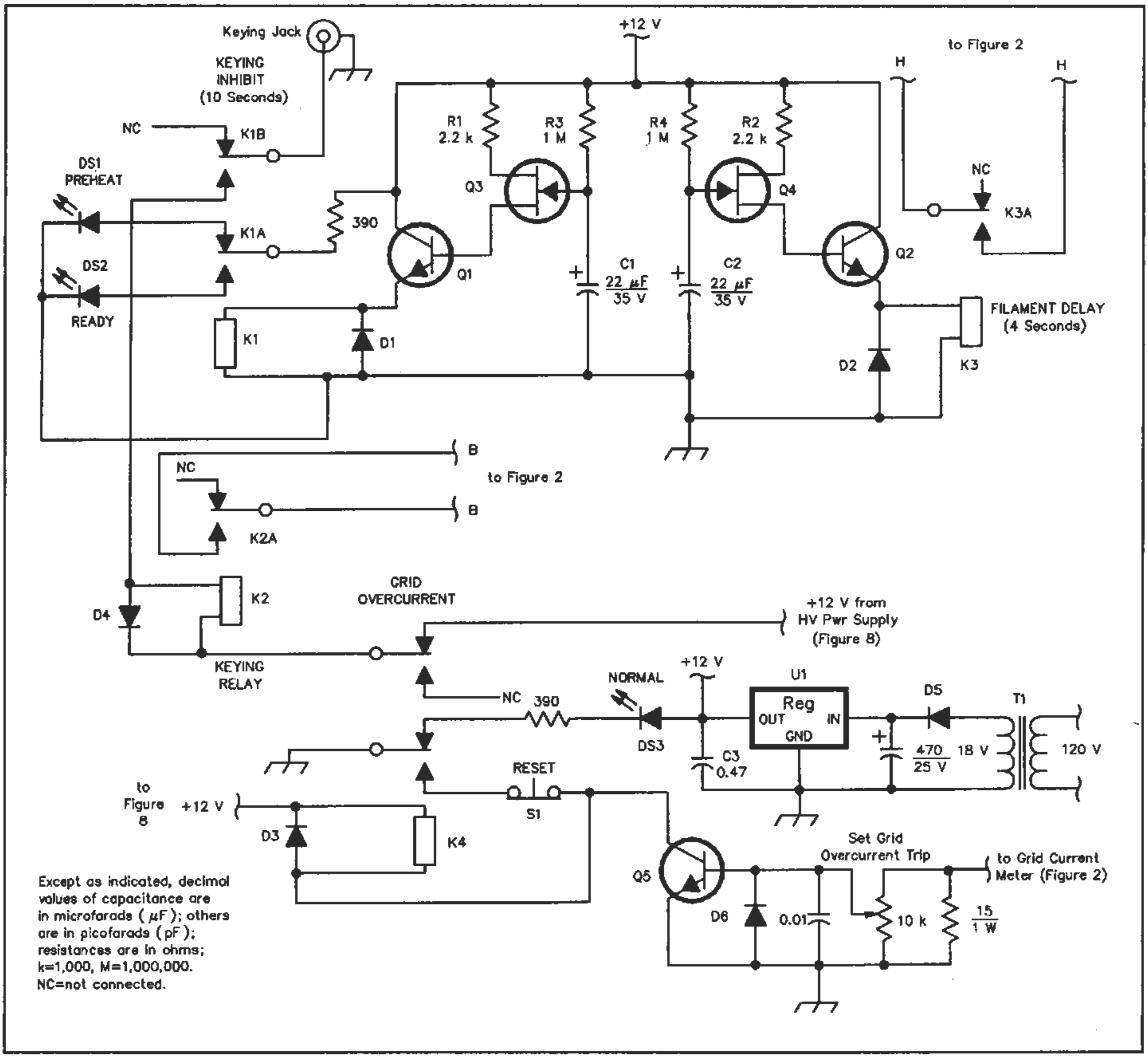
Keying Jack
KEYING INHIBIT (10 Seconds)
+12 V
to Figure 2
H
H
NC
K1B
DS1 PREHEAT
R1 2.2 k
R3 1 M
R4 1 M
R2 2.2 k
Q3
Q4
NC
K3A
390
K1A
DS2
Q1
C1 22 µF 35 V
C2 22 µF 35 V
Q2
READY
K1
D1
FILAMENT DELAY (4 Seconds)
D2
K3
B
NC
to Figure 2
B
K2A
GRID OVERCURRENT
D4
K2
+12 V from HV Pwr Supply (Figure 8)
KEYING RELAY
+12 V
U1
NORMAL
NC
390
Reg
OUT
IN
GND
D5
T1
DS3
C3 0.47
470 25 V
18 V
120 V
RESET
S1
to Figure 8
+12 V
D3
K4
Set Grid Overcurrent Trip
to Grid Current Meter (Figure 2)
Q5
D6
0.01
10 k
15 1 W
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000. NC=not connected.

Figure 4—Schematic diagram of the amplifier-control circuits.

C3—0.47-µF, 25-V tantalum capacitor
D1-D5—1N400i or equivalent
D6—1N4007 or equivalent
DS1—Yellow LED
DS2—Green LED
DS3—Red LED
K1—Keying-inhibit relay, DPDT, 12-V dc coil, 1-A contact rating (RadioShack 275-249 or equivalent)
K2—Amplifier keying relay, SPDT, 12-V dc coil, 2-A contact rating (RadioShack 275-248 or equivalent)
K3—Filament delay relay, SPST, 12-V dc coil, 2-A contact rating (RadioShack 275-248 or equivalent)
K4—Grid-overcurrent relay, DPDT, 12-V dc coil, 1-A contact rating (RadioShack 275-249 or equivalent)
Q1, Q2, Q5—2N2222A or equivalent
Q3—MPF1O2 or equivalent
Q4—2N38i9 or equivalent
S1—Normally closed, momentary pushbutton switch (RadioShack 275-1549 or equivalent)
T1—Power transformer, 120-V primary, 18-V, 1-A secondary
U1—+12 V regulator, 7812 or equivalent

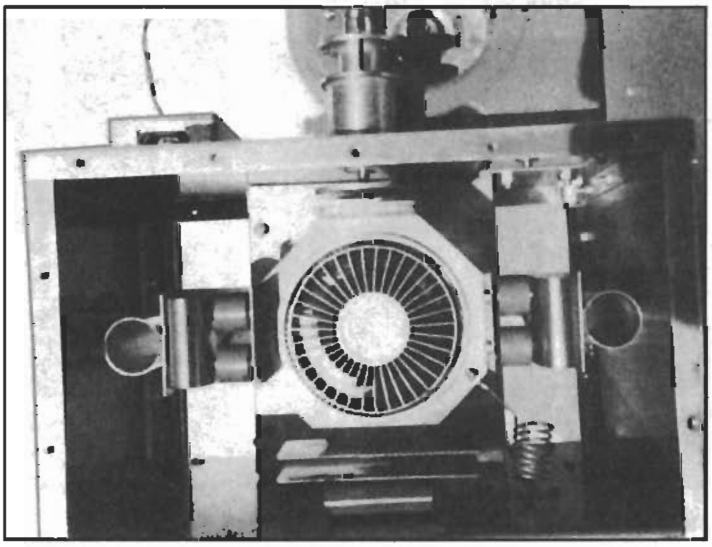

Figure 5—This top view of the plate compartment shows the plate-line arrangement, C1-C4 and the output coupling assembly.

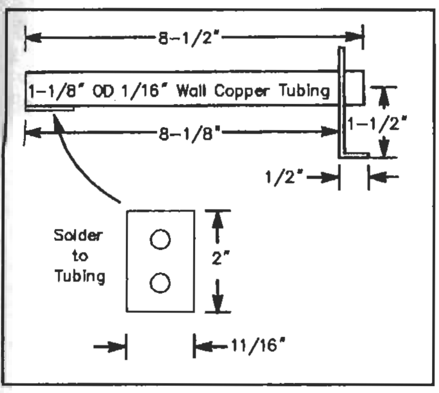
8-1/2"
1-1/8" OD 1/16" Wall Copper Tubing
8-1/8"
1-1/2"
1/2"
Solder to Tubing
2"
11/16"

Figure 6—Plate line details.

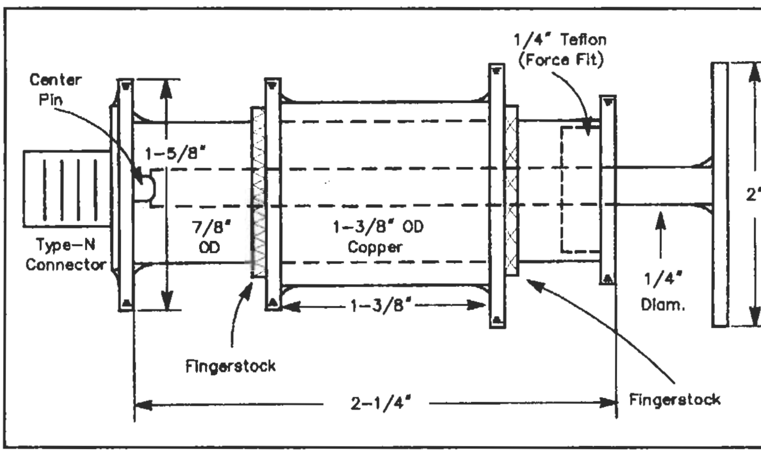
1/4" Teflon (Force Fit)
Center Pin
1-5/8"
Type-N Connector
7/8" OD
1-3/8" OD Copper
2"
1/4" Diam.
1-3/8"
Fingerstock
2-1/4"
Fingerstock

Figure 7—Details of the output coupling assembly.

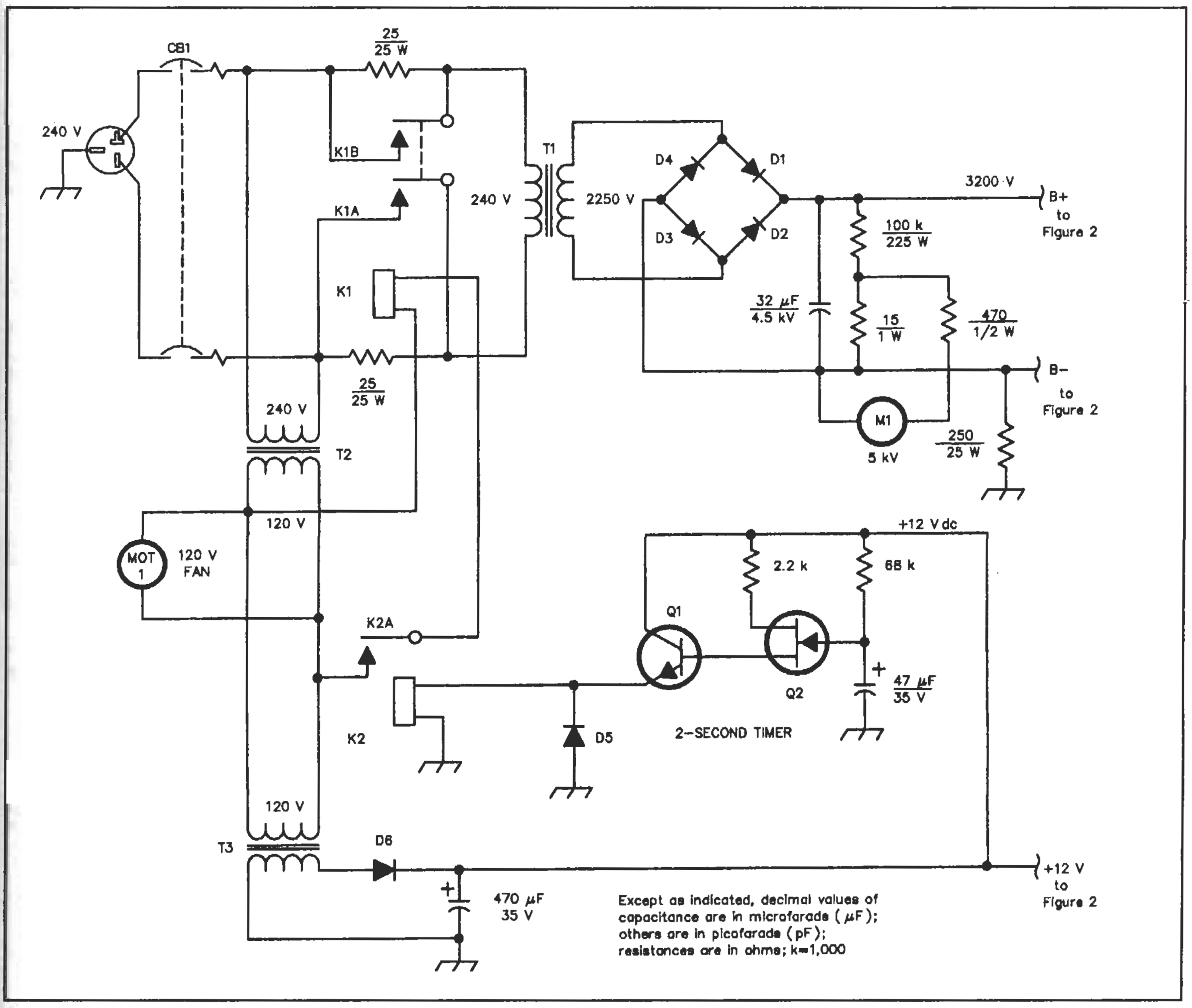
CB1
25
25 W
240 V
K1B
K1A
T1
240 V
2250 V
D4
D1
D3
D2
3200 V
B+ to Figure 2
100 k
225 W
K1
32 μF
4.5 kV
15
1 W
470
1/2 W
25
25 W
B- to Figure 2
240 V
T2
M1
5 kV
250
25 W
120 V
MOT 1
120 V FAN
+12 Vdc
2.2 k
68 k
Q1
K2A
Q2
47 μF
35 V
K2
D5
2-SECOND TIMER
120 V
T3
D6
+12 V to Figure 2
470 μF
35 V
Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k=1,000

Figure 8—Schematic diagram of the high-voltage power supply recommended for use with the power amplifier.

D1-D4—Strings of 4 each, 1000-PIV, 3-A diodes, 1N5408 or equivalent

K1—DPST relay, 120-V ac coil, 240-V-ac, 20-A contacts (Midland Ross 187-321200 or equivalent)

K2—SPDT miniature relay, 12-V dc coil (Radio Shack 275-248 or equivalent)

M1—High-voltage meter, 5 kV dc full scale (1-mA meter movement used with series resistors shown in drawing)

MOT1—Cooling fan, Torin TA-300 or equivalent

Q1—2N2222A or equivalent

Q2—MPF102 or equivalent

S1—20-A hydraulic/magnetic circuit breaker (Potter and Brumfield W68X2Q12-20 or equivalent)

T1—High-voltage power transformer, 240-V primary, 2250-V, 1.2-A secondary (Avatar AV-538 or equivalent)

T2—Stepdown transformer, Jameco 112125, 240-V to 120-V, 100 VA

T3—Power transformer, Jameco 104379, 120-V primary; 16.4 V, 1-A secondary (half used)

Figure 6 is a drawing of the plate line and Figure 7 is a drawing of the output coupling asssembly.

Cooling

The amplifier requires an air exhaust through the top cover, as the plate compartment is pressurized. You can fashion a chimney from a $3\frac{1}{2}$-inch waste-water coupling (black PVC) and a piece of $\frac{1}{32}$-inch-thick Teflon sheet. The PVC should extend down from the underside of the amplifier cover plate by $1\frac{1}{8}$ inches, with the Teflon sheet extending down $\frac{3}{4}$ inch from the bottom of the PVC.

The base of the 3CX1200Z7 is cooled by using bleed air from the plate compartment, which is directed at the tube base, through a $\frac{7}{8}$-inch tube set into the sub-chassis wall at a 45° angle.

The recommended blower will supply more than enough air for any temperature zone. A smaller blower is not recommended, as it is doubtful that the base area will be cooled adequately. The 3CX1200Z7 filament draws 25 A at 6.3 V! It alone generates a great deal of heat around the tube base seals and pins, so good air flow is critical.

Construction

The amplifier is built into a 12×12×10-inch enclosure. A 12×10-inch partition is installed $7\frac{1}{4}$ inches from the rear panel. The area between the partition and the front panel contains the filament transformer, control board, meters, switches, Zener diode and miscellaneous small parts. Wiring between the front-panel area and the rear panel is through a $\frac{1}{2}$-inch brass tube, located near the shorted end of the right-hand plate line.

High voltage is routed from an MHV jack on the rear panel, through a piece of RG-59, just under the shorted end of the left-hand plate line. The cable then passes through the partition to a high-voltage standoff insulator made from nylon. This insulator is fastened to the partition near the high-voltage feed-through capacitor. A 10-Ω, 25-W resistor is connected between the insulator and the feed-through capacitor.

The plate lines are connected to the dc-blocking capacitors on the plate collet with $1\frac{3}{4}$×2-inch phosphor-bronze strips. The bottom of the plate lines are attached to the sides of the subchassis, with the edge of the L-shaped mounting bracket flush with the bottom of the subchassis.

When preparing the subchassis top plate for the 3CX1200Z7, cut a $2\frac{11}{16}$-inch hole in the center of the plate. This hole size allows clearance between the tube envelope and the top plate, without putting stress on the envelope in the vicinity of the grid flange seal.

Table 3
Power Supply Specifications

High voltage: 3200 V
Continuous current: 1.2 A
Intermittent current: 2 A
Step/Start delay: 2 seconds

Exercise care in placing the movable tuning plate and the movable output coupling disc, to ensure they cannot touch their fixed counterparts on the plate collet.

Operation

When the amplifier is first turned on, it cannot be keyed until:

- 10 seconds has elapsed
- High voltage is available, as confirmed by presence of +12 V to the keying circuit

Connect the amplifier to a dummy load through an accurate power meter capable of indicating 1500 W full scale. Key the amplifier and check the idling plate current. With 3200-V plate voltage, it should be in the vicinity of 150 mA. Now, apply a small amount of drive and adjust the input tuning for maximum grid current. Adjust the output tuning until you see an indication of RF output. Increase drive and adjust the output coupling and tuning for the desired output. Do not overcouple the output; once desired output is reached, do not increase loading.

When you shut down the amplifier, leave the blower running for at least three minutes after you turn off the filament voltage. I found the 3CX1200Z7 to be an excellent tube. I tried it with exessive drive, plate-current saturation, excessive plate dissipation—all the abuse it's likely to encounter in amateur applications. I had no problems, but I don't recommend you repeat these tests!

A Companion Power Supply

A good, solid-state high-voltage power supply is a necessity to ensure linearity in SSB operation. Specifications of the power supply I built are given in Table 3. Figure 8 is a schematic diagram of the supply. A power supply for a high-power linear amplifier should operate from a 240-V circuit, for best line regulation. I have specified a special, hydraulic/magnetic circuit breaker that doubles as the main power switch. I don't recommend you substitute a regular switch and fuses for this breaker, as fuses won't operate quickly enough to protect the amplifier in case of an operating abnormality. The bleeder resistor dissipates about 100 W, so I included a small fan to remove the excess heat.

Power Supply Construction

The power supply can be built into a 17×13×10-inch cabinet. The power transformer is quite heavy, so use $\frac{1}{8}$-inch aluminum for the cabinet bottom, and reinforce it with aluminum angle for extra strength. The diode bridge consists of four legs, each containing five diodes.

Power Supply Operation

When the front-panel breaker is turned on, the two, 25-Ω resistors in the primary circuit limit inrush current as the filter capacitor charges. After two seconds, K1 activates, shorting both resistors and allowing full line voltage to be applied to the transformer.

As with all high-voltage power supplies, you must be extremely careful! Before opening the cabinet, remove the ac-line plug from its receptacle, and confirm that the filter capacitor is discharged before working on the supply.

Conclusion

This amplifier is a reliable and cost-effective way to generate a big 2-meter signal—almost as quickly as a solid-state amplifier.

To ensure that the output of my amplifier meets current spectral purity requirements, I use a high-power version of the half-wave output filter that appears as Figure 16 on page 39-10 of the 1993 and 1994 editions of *The ARRL Handbook*. Although I did not make spectral measurements of the output, I can run full output while my wife Mary Lou watches TV in a nearby room of our home.

Another suitable filter is the one that appears in the 1990 *ARRL Handbook* (Figure 150, on page 31-72) as part of the description of "A Legal-Limit 2-Meter Tetrode Amplifier."

Notes

[1] Suggested retail price of the 3CX1200Z7 is $625. You can obtain it from: Henry Radio, 2050 S Bundy Dr, Los Angeles, CA 90025, tel 310-820-1234; Richardson Electronics, 40 W 267 Keflinger Rd, La Fox, IL 60147, tel 708-208-2200; RF Parts, 435 South Pacific St, San Marcos, CA 92069, tel 619-744-0700.

[2] W. Orr, Editor, *Radio Handbook*, 23rd ed. (Indianapolis: Howard W. Sams and Co., 1987), pp 18-2 through 18-7.

[3] This project is also described in the *Radio Handbook*, pp 18-11 through 18-15.

By David Munyon, W7DVB From *QST*, September 1984

A Cathode-Driven Tetrode for 6 Meters

Stability, power, low cost: Is this what you're looking for in a 6-meter do-it-yourself project? If so, look what's here!

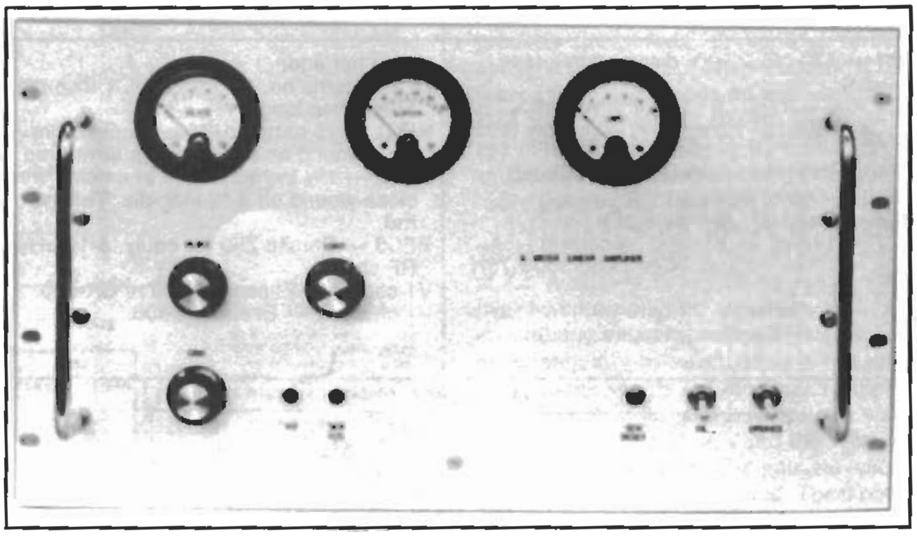

The advantages of grounded-grid amplifiers are numerous and have been proven over the years. They include: simplicity of design, good thirdorder IMD characteristics, inherent stability (generally without the need for neutralization) and noncritical tuning. The only requirement is abundant driving power. Usually, this is a small price to pay, unless the driver you intend to use is one of the many popular, low-power (10- to 25-W-output) VHF transceivers. Also, with the tremendous increase in the use of 100%-duty-cycle modes such as RTTY and SSTV, many of us would rather not operate our exciters at half to full bore while driving an amplifier. If you're willing to sacrifice only one of the aforementioned advantages—simplicity—the remaining attributes of grounded-grid operation can be made available at a considerable reduction in driving power.

Enter the Tetrode

For many years, tetrodes (most often connected as triodes) have been used in grounded-grid circuits. But we have good reasons to consider the tetrode as a cathode-driven performer in its own right. First, the drive requirements are reduced. A 4-400A requires 40W of drive power for full output when triode connected. Only half that amount is needed to drive its class-AB1 counterpart. The second reason is cost. Tetrodes are still available as hamfest and surplus items at a fraction of their original cost. (Some of the newer triodes cost more than the exciters that drive them!)

But not all tetrodes can be triode connected. All of the external anode family of tubes, and a few others, have internal geometry that allows the control grid dissipation to be greatly exceeded if the tube is operated as a triode. Therefore, those tubes should be cathode driven only as tetrodes. The usual way to do this is to ground the control and screen grids to RF with low-inductance bypass capacitors, and then place operating voltages on them. The cathode can be driven in a normal fashion, and the tube operated in any class consistent with linear service. Class-AB1 has the lowest drive requirements and is the easiest with which to work.

Tetrode Vs. Triode

I concede that the addition of a screen element complicates the overall circuit, especially if the screen-circuit provisions are made properly. But, the circuit need be no more complex than that of its grid-driven equivalent. The cathode-driven tetrode requires grid bias and a well-regulated screen supply. Also, there is no correct way to load a tetrode without a meter to monitor the screen-grid current. This device is invaluable and should not be omitted for the sake of economy.

No discussion of the screen-grid circuit would be complete without some mention of an overcurrent-protection circuit. Several solutions have been offered in the past (including the use of no protection circuit!), and each has its disadvantages. Sensitive relays are expensive and their use can have an adverse effect on screen-voltage regulation. Current-limiting supplies do an excellent job of protecting the tube, but do not inform you when something is wrong. Since screen current is extremely sensitive to minor plate-voltage excursions and plateloading conditions, the builder has to be absolutely certain that these are correct, and that plate voltage is always present with screen voltage, if the screen-protection circuit is to be eliminated. The protection circuit is well worth the small cost involved. This amplifier incorporates an inexpensive and simple circuit that contains none of the disadvantages mentioned earlier.

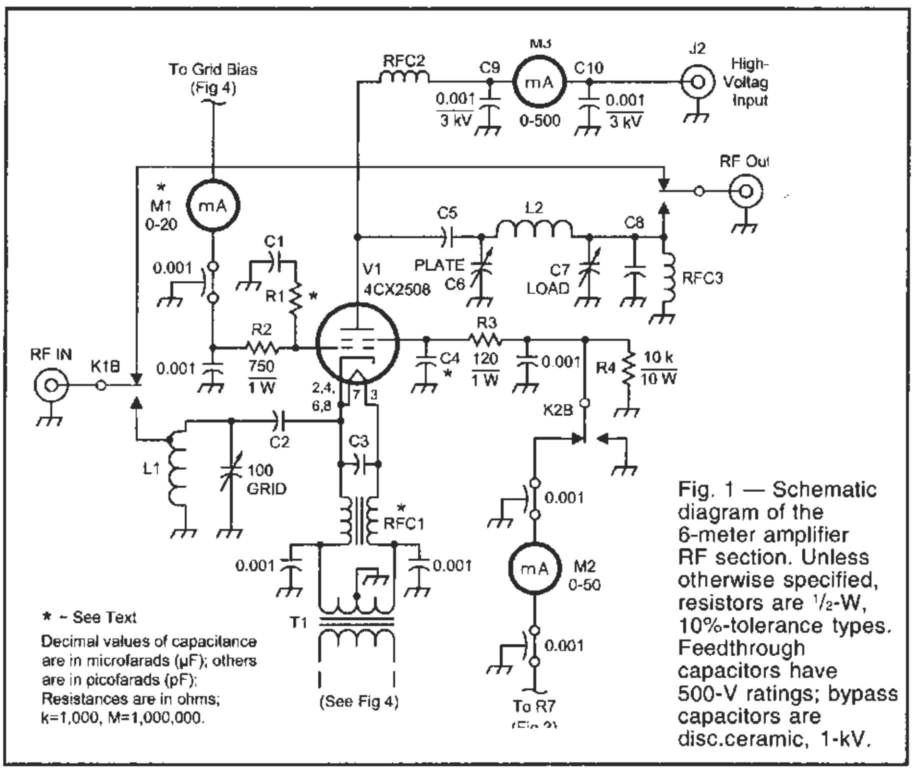
To Grid Bias
(Fig 4)
RFC2
C9
M3
mA
C10
J2
High-
Voltag
Input
0.001
3 kV
0-500
0.001
3 kV
RF Out
*
M1
mA
0-20
C5
L2
C8
C1
0.001
V1
4CX2508
PLATE
C6
C7
LOAD
RFC3
R1
*
R2
R3
RF IN
K1B
0.001
750
1 W
C4
*
120
1 W
0.001
R4
10 k
10 W
2,4,
6,8
7
3
K2B
C2
C3
L1
100
GRID
0.001
*
RFC1
0.001
0.001
mA
M2
0-50
T1
(See Fig 4)
0.001
To R7
(Fig 3)

* - See Text

Decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); Resistances are in ohms; k=1,000, M=1,000,000.

Fig. 1 — Schematic diagram of the 6-meter amplifier RF section. Unless otherwise specified, resistors are ½-W, 10%-tolerance types. Feedthrough capacitors have 500-V ratings; bypass capacitors are disc.ceramic, 1-kV.

C1 — 470-pF, 2.5-kV disc-ceramic (see text).
C2, C3 — 0.001-µF, 500.V silver mica.
C4 — Part of tube socket assembly (see text).
C5 — 500-pF, 5-ky ceramic (Centralab 858S-500 or equlv.).
C6 — See text (approx. 6 pF).
C7 — 140-pF air variable, receiving type.
C8 — 75-pF ceramic (Centralab 850S-75N or equiv.).
L1 — 4½ turns no. 14 bare wire, ½-inch-dia, 1½ inches long, tapped one turn from hot end.
L2 — 6 turns no. 10 bare wire, 1-inch-dia, 1½ inches long.
RFC1 — 15 turns no. 16 enameled wire closewound on a ½-inch-dia ferrite rod.
RFC2 — 1⅝ inches no. 22 enameled wire close-wound on a ⅜-inch-dia. Teflon® rod.
RFC3 — Ohmite Z50 (or equiv. 5-10 µH RF choke).
V1 socket — Eimac SK-600 or SK-620.
V1 chimney — Elmac SK-606.

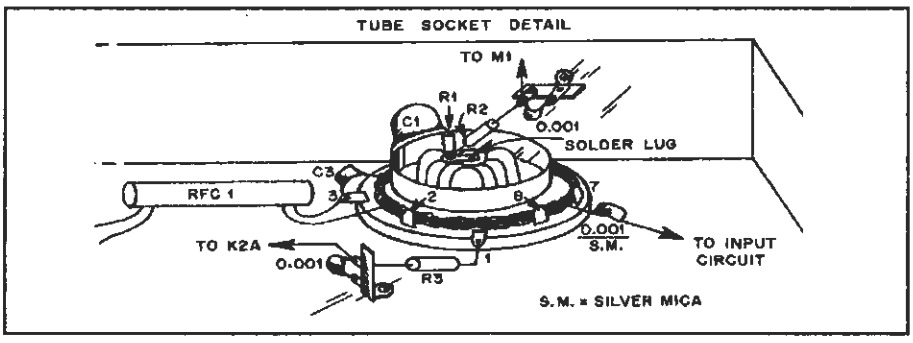
TUBE SOCKET DETAIL
TO M1
R1
R2
C1
0.001
SOLDER LUG
C3
RFC 1
3
2
8
7
0.001
S.M.
TO INPUT
CIRCUIT
TO K2A
0.001
1
R3
S.M. = SILVER MICA

Fig 2—Tube socket detail. See text.

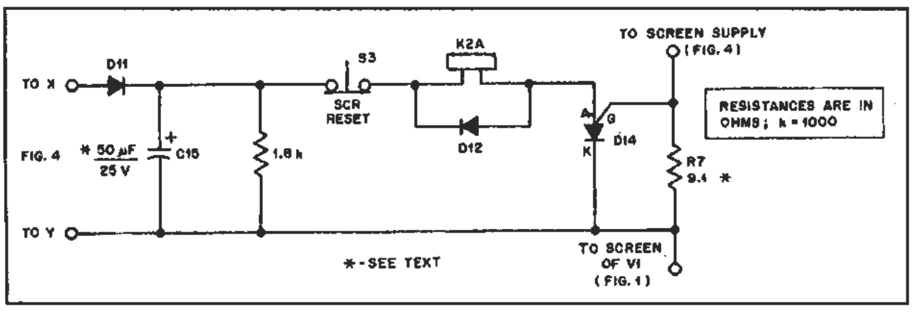
TO SCREEN SUPPLY
(FIG. 4)
D11
TO X
S3
K2A
SCR
RESET
D12
A
G
K
D14
RESISTANCES ARE IN
OHMS ; k = 1000
FIG. 4
* 50 µF
25 V
C15
1.8 k
R7
9.4 *
TO Y
TO SCREEN
OF V1
(FIG. 1)
* - SEE TEXT

Fig 3—Screen overcurrent-protection circuit. Resistors are 1/2-W, 10%-tolerance types.

D11, D12—1-kV, 2.5-A silicon diodes.
D14—4-A, 200-V SCR (Jameco C106B1, ECG 5455 or equiv.).
K2—DPDT 12-V dc relay, 1-A contacts (see text).

A Cathode-Driven 4CX250B

I became aware of the cathode-driven tetrode years ago while trying to get a 4-1000A to wotk on 50 MHz. The results were so good that, when it came time to build an amplifier for my new solid-state transverter, I decided that building a grid-driven amplifier would be a giant step backward.

The 4CX250B is an excellent tube for 6-meter use. This tube has a high plate-dissipation rating for its small physical size, has reasonable power-supply requirements, and requires approximately 10 W of drive in a cathode-driven, class-AB1 configuration. (If much more drive is available, an appropriate attenuator must precede the input circuit.) The only differences between this design and its grid-driven equivalent are the filament/cathode and control-grid circuits.

Filament/Cathode Circuit

In Fig. 1, notice that the cathode and one of the filament connections are tied together. Normally, the isolation created by the physical separation of the two elements is sufficient to ensure stability, and a filament choke would not be required. I felt that this would not be the case on 50 MHz, and a call to Eimac confirmed it. At Bill Orr's recommendation, I tied the cathode and filament together, and fed the filament through a choke. The cathode impedance is approximately 120 ohms. In the interest of best linearity and minimum, drive requirements, a tuned input circuit is used. The tank coil is tapped to present an input impedance of 50 ohms.

There are many sockets designed for use with this family of tubes. I recommend that you use one that incorporates a built-in screen bypass capacitor (1100 or 2700 pF); either one will do. A few of these types of sockets are so constructed that the four cathode pins are grounded internally; do not use one of these. An Eimac SK-600 or SK-620 socket is recommended. The four cathode pins (2,4, 6, 8) and one of the filament pins (7) of the socket must be wired together to form one low-inductance connection. The easiest way to do this is to carefully bend pins 1 and 3 parallel to the chassis, and then strap together pins 2, 4, 6, 7 and 8. Make certain that this strap forms a complete circle around the socket. For strap material, I use some stretched and flattened shield braid removed from RG-58 coaxial cable. All connections made to the cathode, or the filament (pin 7), can be made anywhere along this ring. See Fig. 2 for details.

Grid Circuit

The tube grid is accessed through the socket center connector and is bypassed to ground by C1. The capacitance of C1 is not as important as its physical size. C1 should be a high-voltage, disc-ceramic type (2 kV or more), whose body nearly spans the distance between the connection points. This provides for an absolute minimum lead

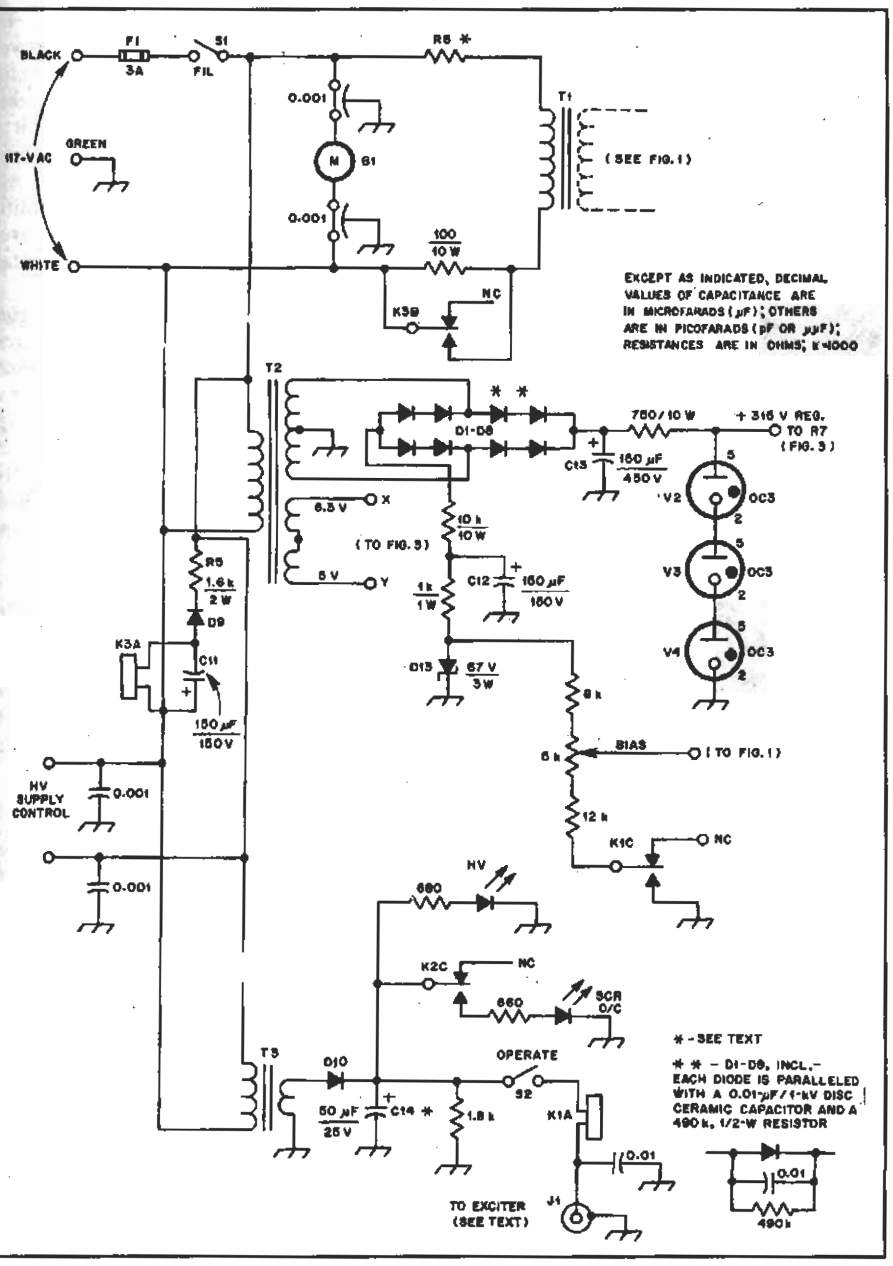
BLACK
F1
3A
S1
FIL
R6 *
117-VAC
GREEN
WHITE
0.001
M
B1
0.001
100
10 W
T1
(SEE FIG. 1)
K3B
NC
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (μF); OTHERS ARE IN PICOFARADS (pF OR μμF); RESISTANCES ARE IN OHMS; k=1000
T2
* *
D1-D8
750/10 W
+ 315 V REG. TO R7 (FIG. 3)
C13
150 μF
450 V
V2
OC3
V3
OC3
V4
OC3
6.3 V
X
(TO FIG. 3)
5 V
Y
10 k
10 W
C12
150 μF
150 V
1 k
1 W
R5
1.6 k
2 W
D9
K3A
C11
150 μF
150 V
D13
67 V
3 W
8 k
5 k
BIAS
(TO FIG. 1)
12 k
K1C
NC
HV SUPPLY CONTROL
0.001
0.001
HV
680
K2C
NC
680
SCR O/C
* - SEE TEXT
* * - D1-D8, INCL.- EACH DIODE IS PARALLELED WITH A 0.01-μF/1-kV DISC CERAMIC CAPACITOR AND A 490 k, 1/2-W RESISTOR
T3
D10
50 μF
25 V
C14 *
1.8 k
OPERATE
S2
K1A
0.01
TO EXCITER (SEE TEXT)
J1
0.01
490 k

Fig. 4 — Power-supply section of the 6-meter amplifier. Unless otherwise specified, resistors are ½-W, 10%-tolerance types; feedthrough and bypass capacitors have 500-V ratings.

B1 — Squirrel-cage blower, 56 CFM (minimum), 117-V ac motor.
D1-D10, incl. — 1-kV, 2.5-A silicon diodes.
D13 — 87-V, 5-W Zener diode.
K1 — 3PDT 12-V dc relay, 5-A contacts.
K3 — SPST 234-V ac relay, 3-A contacts (see text).
T1 — 117-V primary; 6-V. 2.6-A secondary (see text).
T2 — 117-V primary; 580-V c.t., 75-mA secondary with 6.3- and 5-V filament windings (see text).
T3 — 117-V primary; 12-V. 1-A secondary.

length. Use a solder lug on the grid pin, and solder the other capacitor lead directly to the inside of the air duct on the tube-socket bottom.

When my amplifier was first tested, I found that the plate current idled at about 40 mA, and was completely independent of grid voltage. This self-biasing was traced to a parasitic in the grid circuit, and was eliminated by the addition of a series-connected resistor, R1. If you wish to experiment, R1 should be a ½-W, carbon-composition resistor with a resistance value that is the minimum required for complete stability. R1 should be placed as shown in Fig. 2; I found that a 3.9-ohm resistor kept my amplifier unconditionally stable.

In the interest of stability, don't use chokes in place of R2 and R3 in the grid and screen circuits. M1, the grid-current meter, can have a range of from 0-1 mA to 0-20 mA or so, and is included only to ensure that grid current is never drawn, and the operation remains class-AB 1.

Screen Circuit

Relay contacts K2B are part of the screen overcurrent-protection circuit, the operation of which will be discussed later. Because the 4CX250B draws negative screen current on occasion, the screen-current meter should ideally be of the zero-center type. As these meters are expensive and not readily found, a standard meter can be pressed into service by using bleeder string, R4. Notice that 15 mA of bleeder current is drawn through the meter whenever K2 is open (normal). This offsets the actual zero reading of the meter to the 15-mA position and allows a negative current of 15 mA to be metered. This is more than adequate for the 4CX250B. If a zero-center meter is used, R4 may be eliminated unless the screen supply shunt regulator is replaced by a series regulator. In that case, the bleeder would be necessary to offset the effects of secondary emission.

For the overcurrent-protection circuit (Fig. 3), an SCR (D14) seems to be a natural. A very small gate voltage, developed across R7 by the screen current, turns on the SCR and causes K2 to close. This removes screen voltage from the tube and grounds the screen, preventing further abuse. A second set of relay contacts (K2C) is used to light a front-panel-mounted LED. D14 continues to lock out the screen voltage until the SCR is reset by removing its anode voltage. This is done by momentarily pushing the RESET switch, S3.

D11, C15 and the filament windings of T2 provide dc power for a 12-V relay at K2. A relay with a different voltage rating can be used if this supply is modified accordingly. The value for R7 was found experimentally by substituting a variable, low-voltage supply for the screen supply. A 0-100 mA meter and load resistor were placed in series with R7, and the voltage, slowly increased until the gate threshold was reached and the SCR fired. The threshold current was noted, and R7 varied until the desired current would trigger the SCR. With R7 equal to 9.1 ohms, the SCR fired consistently at 45 mA.

Since 15 mA of the current drawn through R7 will be bleeder current, this allows a maximum of 30 mA for screen current, which is well below the 12-W screen dissipation rating for the tube. Note that the entire overcurrent-protection circuit is hot with screen voltage, so proper precautions must be taken during construction. I built the protection circuit on a PC board, but it could have just as easily been chassis-mounted using terminal strips.

Construction

The amplifier is built on a 3 × 8 × 17-inch chassis; a 3 × 5 × 7-inch chassis houses the output network components. A 10½- × 19-inch rack panel is used for the amplifier front panel. The general physical layout should be followed, as it is well dictated by the flow of the circuit. All RF leads should be kept as short as possible, especially the bypass capacitor leads. Keep the dc leads cabled and

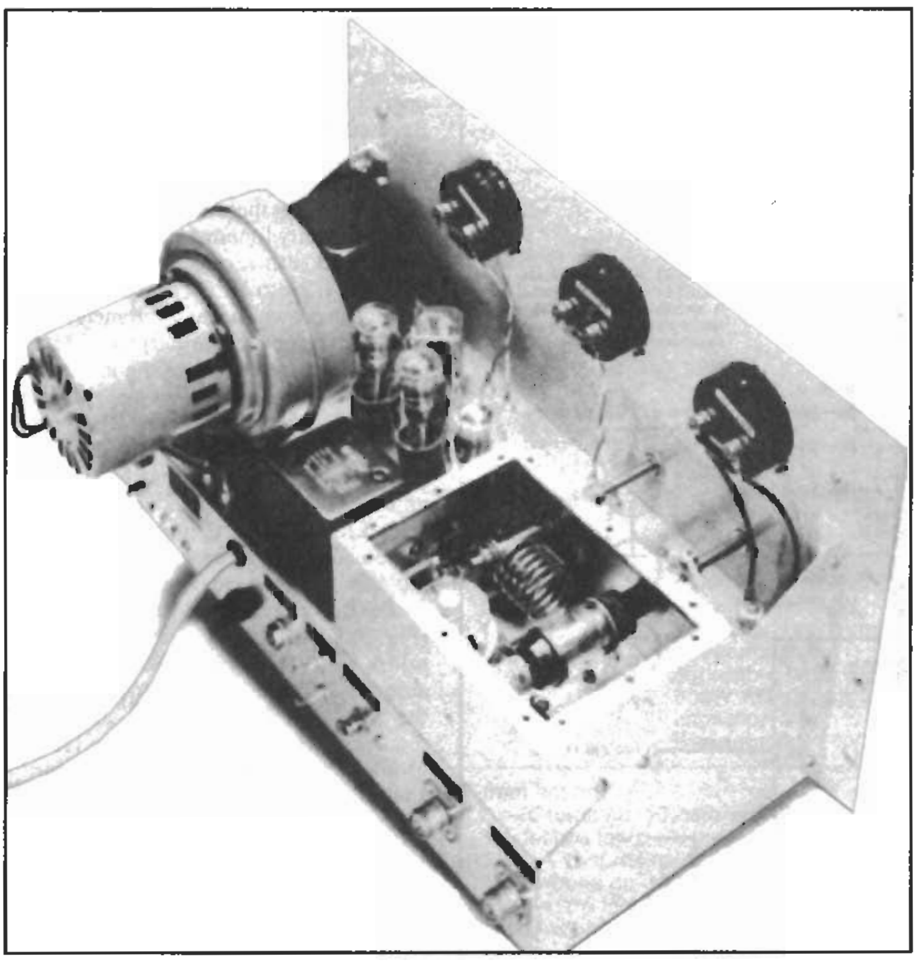

Fig 5—The final-amplifier-compartment cover has been removed for this photo.

as far away as possible from the RF leads. Most power-supply components are mounted on a fiberglass terminal board that is secured against the side of the chassis.

K3, D9, C11 and R5 (Fig. 4) constitute a time-delay circuit for filament inrush-current protection, and can be included if you feel that this is important. As it stands, the delay is approximately 1 second, which is more than ample. The 234-V ac relay is used on dc, and closes at about 70 V. The charging rate of C11 is controlled by its capacitance and the ohmic value of R5, and can be varied by changing their values.

Ideally, the input and output relays should be coaxial types; however, I used a single open-frame, 3PDT type and encountered no problems. The center pole is used for control (K1C), as that increases the isolation between the input and output circuits. In grid-driven amplifiers, I have never been able to use an open-frame relay for input and output switching. Because of the lower gain of this cathode-driven amplifier, I find no evidence of feedback or instability. Don't use plug-in type relays in their original form for K1. Remove the case, plug and connecting wires, and bolt the relay directly to the chassis. Use coaxial-cable braid to make the RF connections. The power supplies (Fig. 4) are of standard design. They may be replaced by any supplies yielding similar voltages, and need not reside on the same chassis as the amplifier. The high-voltage supply should produce 2-kV dc under load. Si turns on all supplies, including the external high-voltage supply. S2, the STANDBY/OPERATE switch, supplies power to K1, which allows the amplifier to be placed in operation; otherwise, the amplifier is simply bypassed.

If a 6.3-V ac filament transformer is used for T1, R6 will have to be.included to drop the filament voltage to 6 V. The ohmic value of R6 will have to be determined experimentally. Use a resistor with the highest wattage rating practical for good voltage stability. With an accurate voltmeter, measure the filament voltage at the tube socket with the filament choke in place, as there will be a slight voltage drop across the choke. Nothing will cause a 4CX250B to "go south" quicker than high filament voltage, and that means anything in excess of 6 V.

The filament choke (RFC1) consists of 15 bifilar turns of no. 16 enameled wire wound on half of an Amidon $^1/_2$-inch-diameter ferrite rod. The rod can be cut to a proper length by filing a small groove around its circumference, and then breaking it clean. Cover the rod with heat-shrink tubing or electrical tape before and after winding the choke. This will keep the rod and winding together, and make it self-supporting by its four leads.

The meters I used all have 0-1 mA movements and homemade shunts. The original meter-face calibration marks were carefully erased, and new ones applied using dry-transfer labels. Most such labels are ideal for panel marking, but are a bit too large for meter-face use. Your local stationery store should have transfers with smaller-sized numbers. During amplifier operation, care should be taken to avoid accidental contact with the plate-voltage meter, as it has high voltage on it.

Most VHF amplifiers have a higher-than-necessary Q in the output circuit because of the use of high minimum-capacitance air-variable capacitors. This is acceptable providing the rest of the output components are heavy-duty types and can handle the higher circulating currents. (Fhe small vacuum variable I used is admittedly first class, and it may not be used by many other builders.) Rather than using an air-variable capacitor, a homemade two-disc system is a preferred choice. If you can find a neutralizing capacitor with plates of 2-inch diameter or so, that would be ideal. If you must use an air-variable capacitor, use one with the least minimum capacitance; remove all unnecessary plates leaving just enough to do the job.

Tune-Up

Before the tube is placed in the socket, close S1 and check that all of the operating voltages are correct. Ground the bottom end of the bias string at K1C, and check that the BIAS potentiometer range will supply –50 to –60 V; then set it at –55 V. J1 should be connected to the external-circuit control jack of your exciter so that it is grounded when your exciter is keyed. Open S1 and make certain all voltages have bled down, then install the tube. Close S1 and S2. After the tube warms up, with no drive applied, key the exciter and adjust the BIAS potentiometer so that the plate idling current is exactly 100 mA.

While the tube is idling, rotate all the tuning controls throughout their respective ranges. If there are no sudden plate-current increases, the amplifier is stable. With a small amount of drive applied, tune the GRID and PLATE controls for maximum output. Once you are satisfied that both of these controls will provide resonance, apply full drive and retune the plate to resonance. Increase loading slowly, retuning the PLATE control each time, until a screen current of 5 mA (the meter actually reads 20 mA) is indicated. Observe how sensitive the screen current is to the operation of the LOAD control, and that it is by far the best indicator of plate resonance. If grid current is indicated, you're overdriving the amplifier.

Adjust the drive level and the plate TUNE and LOAD controls so that the platecurrent meter indicates 250 mA and the screen-current meter indicates 5 mA at resonance. Under these conditions, this amplifier develops a power output of about 310 W. With modulation, voice peaks should not exceed 150 mA on the plate-

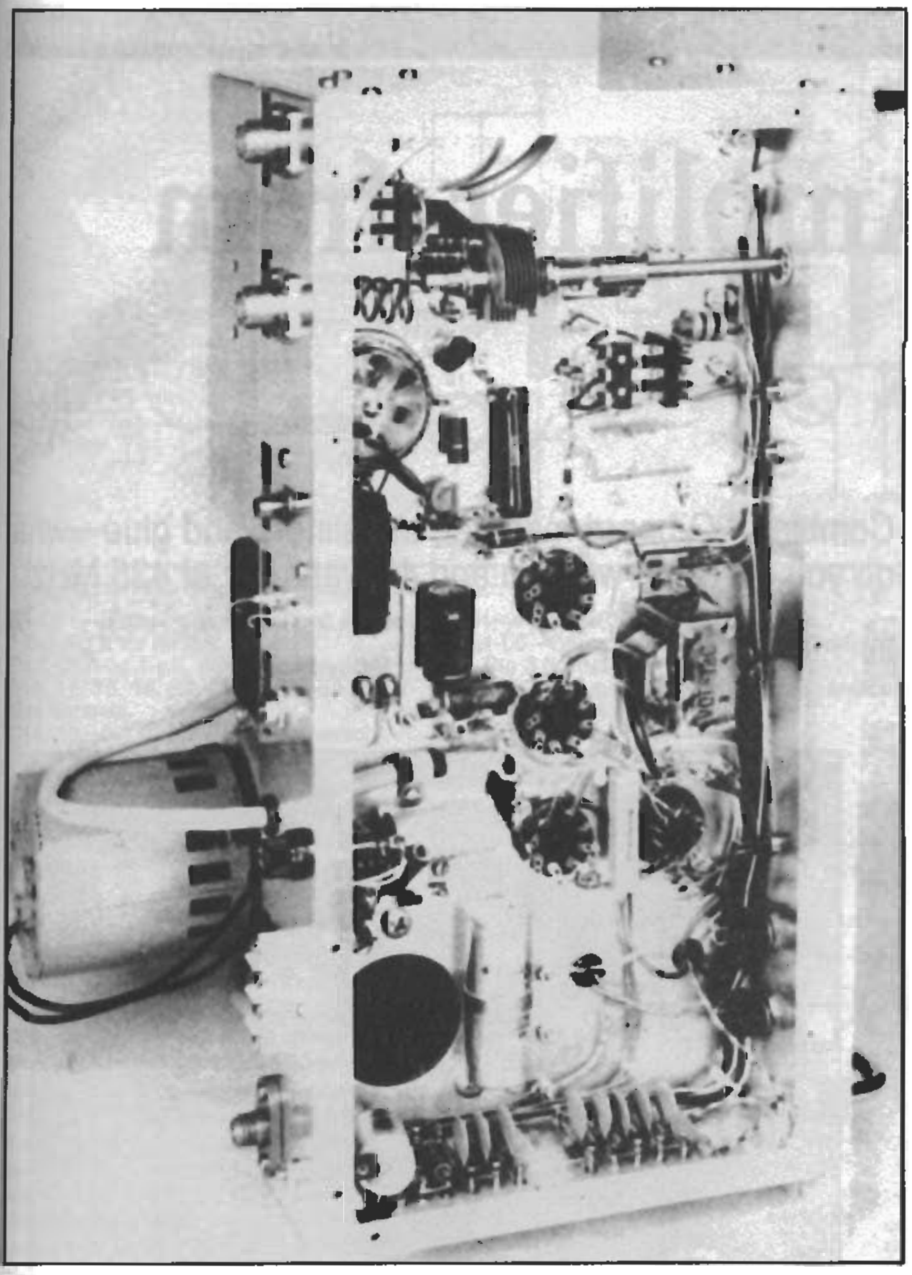

Fig 6—A bottom view of the 6-meter amplifier. The screen-protection circuit is mounted on the PC board to the right of the tube socket.

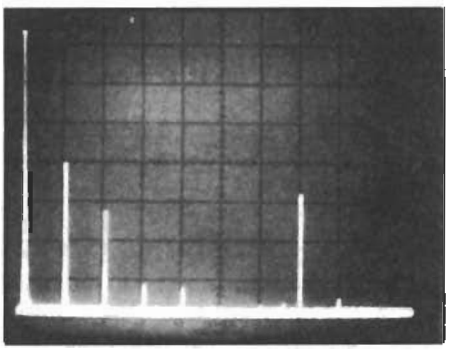

Fig 7—Spectral photo of the 6-meter amplifier without external filtering. Vertical divisions are each 10 dB; horizontal divisions are each 50 MHz. The fundamental (second pip from the left) has been notched approximately 40 dB by means of notch cavities to prevent analyzer overload. The seventh harmonic is approximately 48 dB below peak fundamental output.

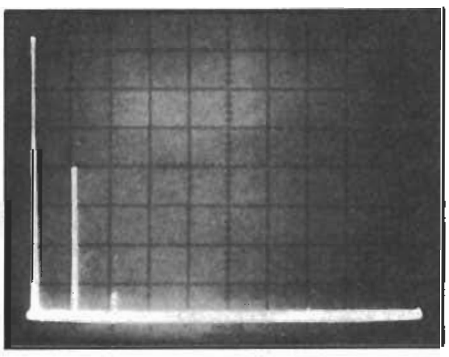

Fig 8—The addition of a simple filter (see Fig 9) at the amplifier output provides for excellent harmonic attenuation, and its use is recommended. All conditions are otherwise the same as those of Fig 7.

current meter, and a small amount of negative screen current is normal.

Summary

This amplifier has been in operation now for over a year, and all reports of its operation have been complimentary. I hope that this amplifier will be duplicated by many who had previously resigned themselves to grid-driven designs. Much of the material presented here is general enough to encourage interest in the cathode-driven tetrode as a viable alternative to the grounded-grid triode. I will be glad to answer all letters of inquiry regarding this amplifier; please include an s.a.s.e.

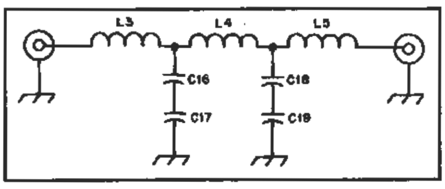

Fig 9—This simple harmonic-suppression filter, or one of similar characteristics, should be used at the output of the amplifier. The filter may be contained in a 2¼ × 2¼ × 5-in (HWD) aluminum box.

C16-C19, incl.—110 pF, 1-kV silver-mica capacitors.

L3, L5—4 turns no. 14, 5/16-in-ID, ½ in long.

L4—5 turns no. 14, 7/16-in-ID, ⅝ in long.

By John Reed, W6IOJ **From *QST*, August 1987**

A UHF Amplifier–from Scratch

Combine PC-board material, transistors and glue—what do you get? Two watts in and 45 watts out at 435 MHz!

Most UHF RF equipment designs involve undefined reactance. For example, the self-inductance of a variable capacitor, or simply the method of attaching a 50-ohm cable to the circuit may constitute reactance that is critical to the operation of the circuit. The result is that complete success in duplicating a circuit is not likely unless the components and layout are exactly the same as in the original. This is nearly always a limiting factor for the do-it-yourselfer wanting to duplicate a project. This class-C UHF amplifier project uses homemade components that anyone can assemble. Tests show that their use has resulted in little or no compromise in the performance of the amplifier over a design using far more expensive commercial parts. My amplifier shows more than 45 W output with less than 2 W input at 435 MHz. The output stage efficiency is approximately 60%, and includes a double-tuned filter for improved spectral purity.

The key to this design is the glue-down stripline technique I have used in many other UHF projects.[1] This method permits easy modification of stripline parameters for optimizing circuitry. Striplines are cut from double-sided, glass-epoxy PC board having the same dimensions that you would choose using the etched-PC-board method. One side of the stripline is smeared with glue (Radio Shack all-purpose adhesive, 64-2307) and firmly pressed against the common base PC board. Parts can be soldered to the stripline immediately (without waiting for the glue to dry). No dc connection is required between the glueline foils, and changes can be made within minutes by lifting the glue-down stripline with a knife and replacing it with one having altered dimensions. This project gives another example of how stripline parameters can be varied to optimize a circuit: homemade variable capacitors.

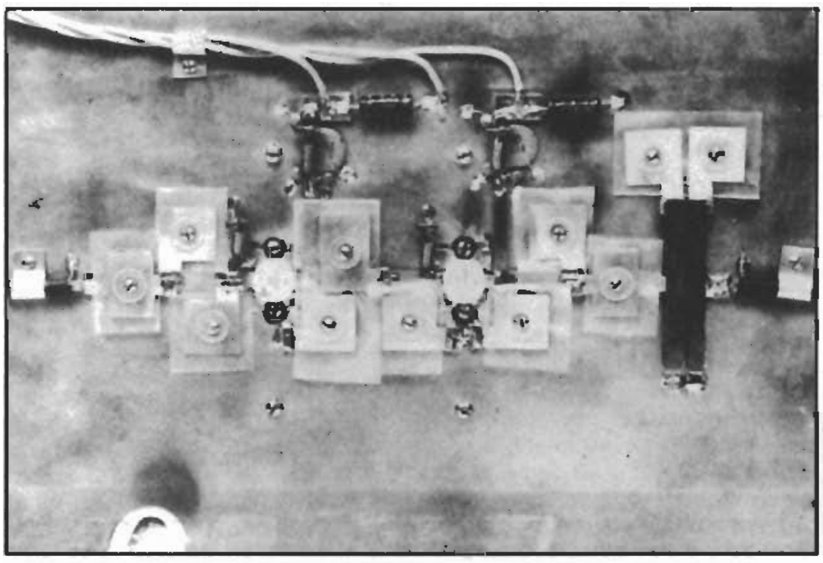

Circuit Details

See Fig 1. The two-stage amplifier uses an MRF641 (driver) and an MRF646 (final).[2] They were chosen primarily because of their availability in the surplus market. These are controlled-Q devices with a combined minimum gain of 12.6 dB. Although the output stage is rated at 45 W, typical saturated output with a 13.6-V collector supply is about 60 W (including drive power).

The home-brew compression-type variable capacitors, shown in Fig 2, dictated the overall design. I made them as small as possible (to minimize selfinductance), but kept them large enough to provide the necessary capacitance. The nominal $^{5}/_{8}$-inch-square compression plate results in a maximum capacitance of about 10 pF. The material used for the capacitor plates is 0.031-inch-thick Reynolds sheet aluminum, which is available at most hardware and home-supply stores. [Hobby store sheet brass works well in this application, is also commonly available, and is far easier to solder; it is a good alternative to aluminum—Ed.] The capacitor dielectric material is 2.7-mil polyethylene from a Dow Ziploc® heavy-duty freezer bag. Although I have tested this dielectric at a much higher voltage than it is subjected to in this application (at a high-impedance point in a vacuum-tube VHF amplifier), double thickness is used as a safeguard. The insulator between the adjustment screw and compression plate is made from plastic polymer. Most clear plastics used for miscellaneous household applications are of this type. I use plastic from a box used in packaging a tube of Grumbacher acrylic artists paint. Glass-epoxy insulation will work also, but at the higher-voltage

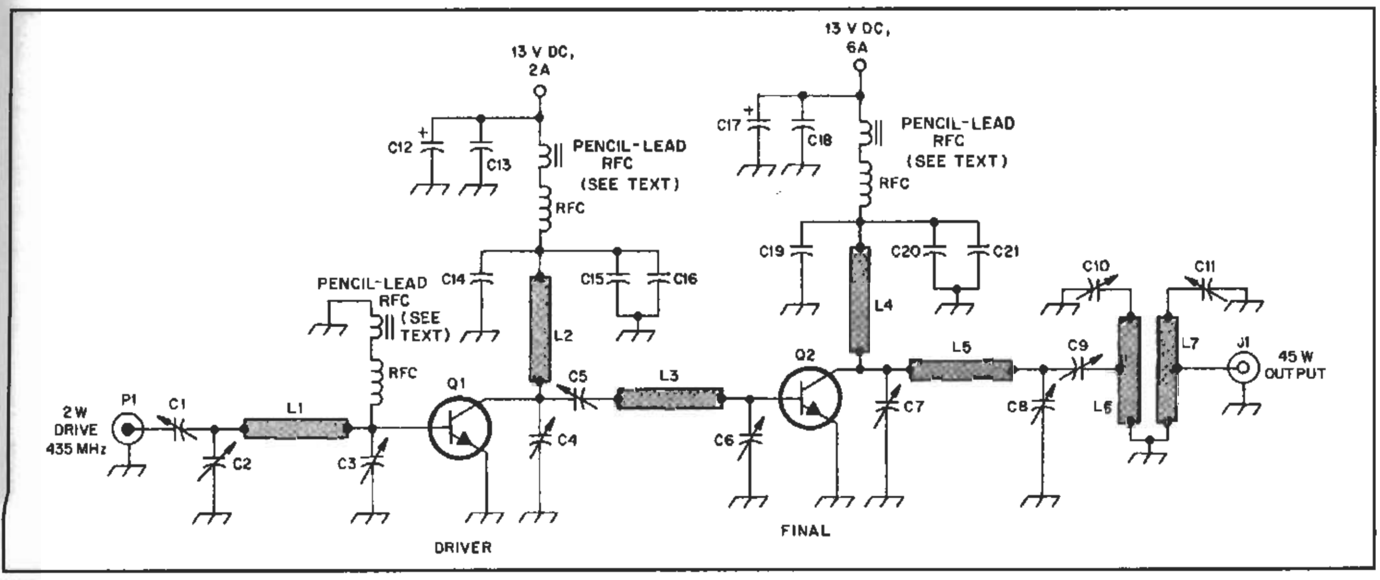
13 V DC,
2A
13 V DC,
6A
C12
C13
PENCIL-LEAD
RFC
(SEE TEXT)
RFC
C17
C18
PENCIL-LEAD
RFC
(SEE TEXT)
RFC
C19
C20
C21
C10
C11
PENCIL-LEAD
RFC
(SEE
TEXT)
C14
C15
C16
L2
L4
RFC
Q1
C5
L3
Q2
L5
C9
L7
J1
45 W
OUTPUT
2 W
DRIVE
435 MHz
P1
C1
L1
C2
C3
C4
C6
C7
C8
L6
DRIVER
FINAL

Fig 1—Schematic of the 435-MHz 45-W amplifier.

C1-C11—Homemade capacitors; see text and Fig 2 for details.
C12, C17—4.7-µF, 35-V electrolytic.
C13, 15, 16, 18, 20, 21—0.001-µF, 50-V disk ceramic.
C14, 19—0.1-µF, 50-V disk ceramic.
J1—Female BNC connector.
L1-L5—Glue-down striplines; see text and Fig 3 for details.
L6, L7—Air-gap striplines; see text and Fig 4 for details.
P1—Male BNC connector.
Q1—MRF641.
Q2—MRF646.
RFC—10 turns, no. 26 enameled copper, air-wound, 3/16-in diameter.

applications (especially at C10 and C11) there will be some loss and component heating.

Once I settled on the capacitor configuration, the input and output stripline characteristics were determined experimentally for optimum matching to Q1 and Q2. The related glue-down stripline and pad details are shown in Fig 3. The L2 L4 1/8-inch-wide striplines act as RF chokes, making the inductance associated with the subsequent disc-ceramic capacitors uncritical. The pads include a number of the capacitor stators (C1, C4, C5, C6, C7 and C9). Foils which contact stator-adjustment screws are reamed out to approximately 1/8-inch diameter to isolate the no. 2-56 screw from the stator. I did this with a large drill (3/8-inch). Screw contact with the bottom foil may also cause adjustment irregularities, so the foil was reamed appropriately. The stator plates must be completely deburred to prevent puncturing the plasticbag dielectric.

The amplifier output is coupled through C9 to a 1:1 inductively coupled, double-tuned filter. This 50-ohm filter consists of two striplines that are mounted parallel to each other and close together above the base PC board, as shown in Fig 4. Their close positioning establishes the necessary mutual coupling. Each line is tuned to resonance with a variable capacitor.

As indicated in Fig 1, dc-supply decoupling is identical for both stages. The 0.1-µF disk capacitors eliminate a bothersome low-frequency instability mode. RF chokes following these capacitors include two-turn sections with pencil-lead cores to minimize the possibility of parasitics. The 4.7-µF capacitors eliminate low-frequency power-supply instability problems.

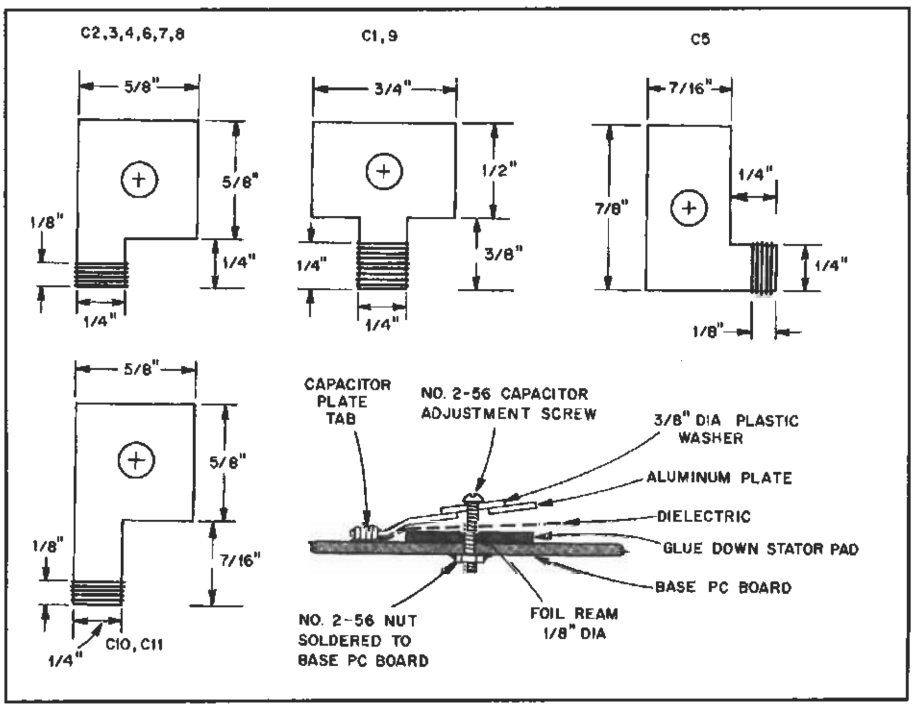
C2,3,4,6,7,8
5/8"
5/8"
1/8"
1/4"
1/4"
C1,9
3/4"
1/2"
3/8"
1/4"
1/4"
C5
7/16"
1/4"
7/8"
1/4"
1/8"
5/8"
5/8"
1/8"
7/16"
1/4"
C10, C11
CAPACITOR
PLATE
TAB
NO. 2-56 CAPACITOR
ADJUSTMENT SCREW
3/8" DIA PLASTIC
WASHER
ALUMINUM PLATE
DIELECTRIC
GLUE DOWN STATOR PAD
BASE PC BOARD
NO. 2-56 NUT
SOLDERED TO
BASE PC BOARD
FOIL REAM
1/8" DIA

Fig 2—Variable-capacitor construction details. The "rotor" plates are made from 0.031-inch Reynolds aluminum sheet stock and polished with 320-grit sandpaper to ensure that no burrs are left that may puncture the dielectric material. Plate center holes are 3/16-inch diameter. Connections are made to the capacitors by means of no. 26 tinned copper wire, tightly wrapped (and crimped in a vise) around the tab on the capacitor and soldered to the appropriate circuit board connection. Make these connections carefully to ensure good electrical contact. See text for details on the dielectric and washer materials.

Assembly

The transistor-lead configurations are slightly modified to accommodate the glue-down striplines. Both the base and collector leads are bent up into an S shape so that the lead will overlap the stripline by approximately 1/16 inch when the transistor and stripline are mounted flush with the base PC board. The leads are only 0.005-inch thick and can be bent easily with long-nose pliers. Care must be taken to minimize stress at the lead-ceramic junc-

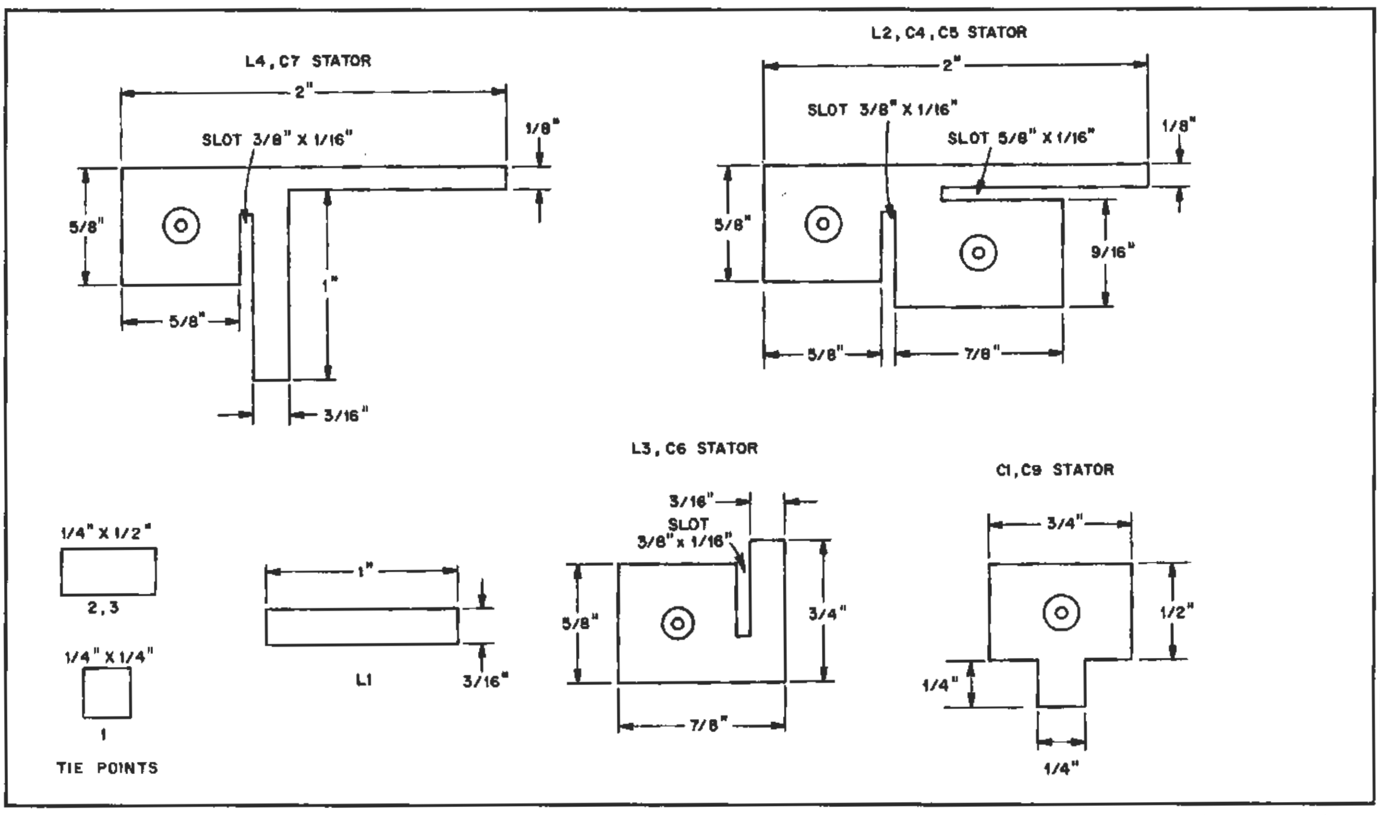
L4, C7 STATOR
2"
SLOT 3/8" X 1/16"
1/8"
5/8"
1"
5/8"
3/16"
L2, C4, C5 STATOR
2"
SLOT 3/8" X 1/16"
SLOT 5/8" X 1/16"
1/8"
5/8"
9/16"
5/8"
7/8"
L3, C6 STATOR
3/16"
SLOT
3/8" x 1/16"
5/8"
3/4"
7/8"
C1, C9 STATOR
3/4"
1/2"
1/4"
1/4"
1/4" X 1/2"
2, 3
1/4" X 1/4"
1
TIE POINTS
1"
L1
3/16"

Fig 3—Glue-down stator pad construction details. Slots are cut with a hacksaw and then polished with a pattern file. The material is 0.062-inch-thick double-sided glass-epoxy circuit board material.[6] After cutting and filing, the pads must be polished with fine steel wool to remove any burrs.

tions. Also, 3/32 inch of each emitter lead must be cut off with scissors to allow room for the glue-down pads.

Details of the mounting arrangement are shown in Fig 5. My transistor mounting method was dictated by the dimensions of an old 1/8-inch-thick aluminum panel (with many miscellaneous holes), which I pressed into service as a heat sink. First, I mounted each transistor on separate aluminum subbases. I then fastened the sub-bases to the heat sink. Thermally conductive compound is used between the transistors and the subbases and between the sub-bases and the heat sink. This arrangement permits removal of the heat sink without disturbing the transistor mountings when I need access to the bottom PC-board foil. Note that the base PC board cutout includes notches at the base and collector leads of the transistors to prevent shorting to the common foil.

Detailed assembly like that shown in Fig 6 is easy once the transistors are mounted. Simply smear glue on the pads and fit them into position. Start by positioning L1, followed by the C1 stator, then C1, C2 and C3. Drill holes in the base PC board for the no. 2-56 capacitor-adjustment screws as you go along. At the same time, solder the no. 2-56 nuts on the reverse side (using a screw to hold them in position). Bend the capacitor tabs to make certain the plates will be flush with the stator, inserting the capacitor dielectric as the final step. Make certain the dielectric has been pushed in between the capacitor plates as far as possible (maximum safety margin), and use an awl to puncture the dielectric at the adjustment-screw position. The adjustment screw can then be easily started through the dielectric.

Operation

UHF transistors are great devices, but they can be damaged permanently with just one voltage transient. For example, an accidental short by a probe during a troubleshooting exercise, or the discharge of a probe capacitor at the transistor's base lead can be disastrous. Another common problem is instability that causes selfoscillation and excessive collector voltage (no insta-

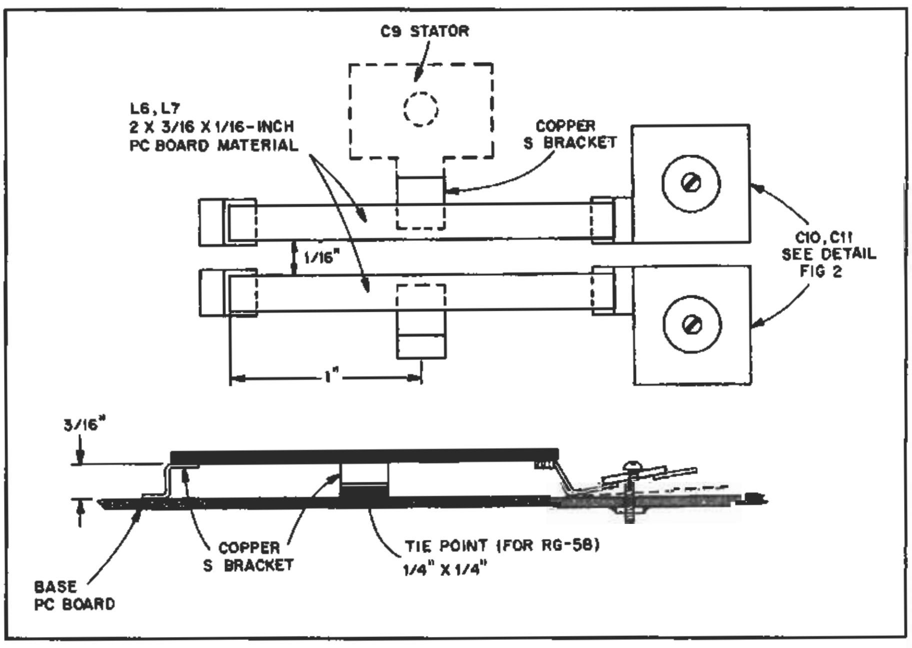
C9 STATOR
L6, L7
2 X 3/16 X 1/16-INCH
PC BOARD MATERIAL
COPPER
S BRACKET
1/16"
C10, C11
SEE DETAIL
FIG 2
1"
3/16"
COPPER
S BRACKET
TIE POINT (FOR RG-58)
1/4" X 1/4"
BASE
PC BOARD

Fig 4—Construction details of the double-tuned output filter. The assembly is supported by copper brackets and parts of the variable capacitors.

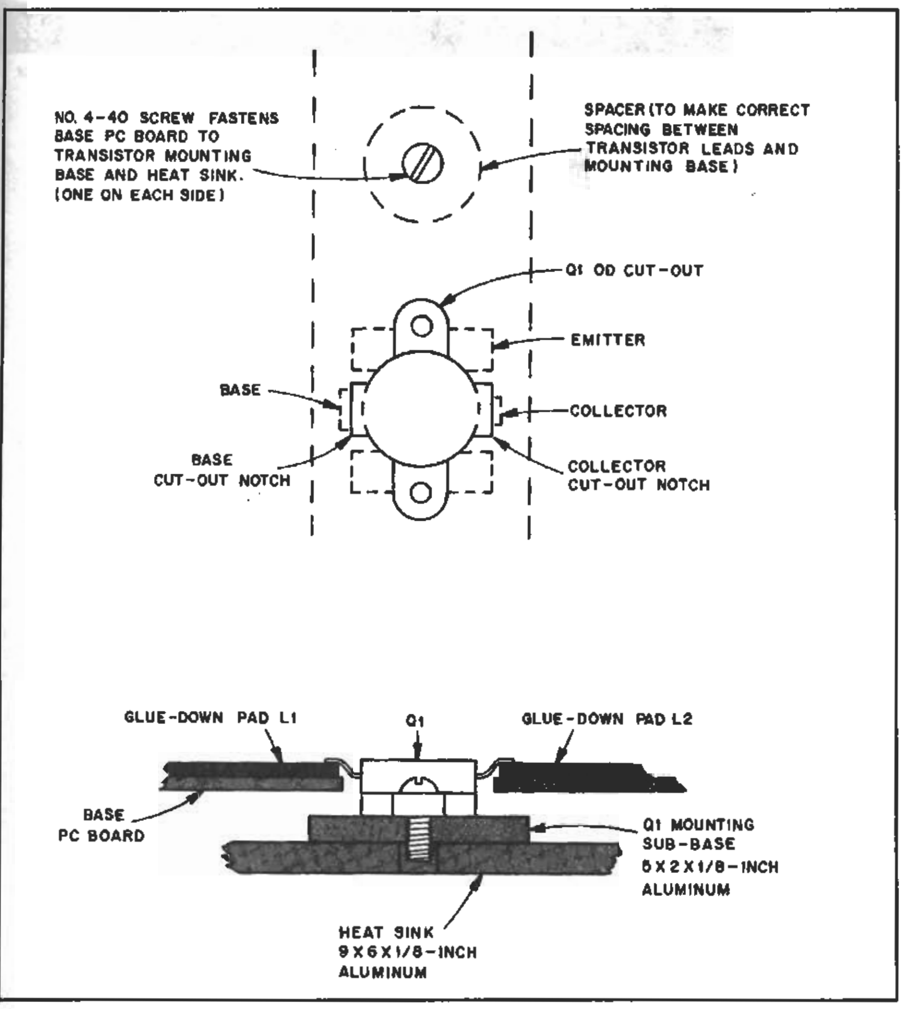
NO. 4-40 SCREW FASTENS BASE PC BOARD TO TRANSISTOR MOUNTING BASE AND HEAT SINK. (ONE ON EACH SIDE)
SPACER (TO MAKE CORRECT SPACING BETWEEN TRANSISTOR LEADS AND MOUNTING BASE)
Q1 OD CUT-OUT
EMITTER
BASE
COLLECTOR
BASE CUT-OUT NOTCH
COLLECTOR CUT-OUT NOTCH
GLUE-DOWN PAD L1
Q1
GLUE-DOWN PAD L2
BASE PC BOARD
Q1 MOUNTING SUB-BASE 5X2X1/8-INCH ALUMINUM
HEAT SINK 9X6X1/8-INCH ALUMINUM

Fig 5—Pictorial of transistor mounting method. The same scheme is used for both the driver and final transistors.

bility modes are evident in the amplifier described here). These possible failure modes can be avoided by limiting the collector supply to 5 V (or less) during tune-up, as practically all instability modes will be evident at this lower potential (particularly if the input drive is varied over a wide range). Following alignment at 5 V, it is surprising how little readjustment is required when the collector potential is increased to 13 V. (If you can't reduce your power supply voltage, you may want to use a 2N3055 as a pass transistor, driving it with an LM317T regulator [Radio Shack 276-2041 and 276-1778, respectively].) This low-voltage test procedure is mentioned in Motorola Application Note AN-548A.[3]

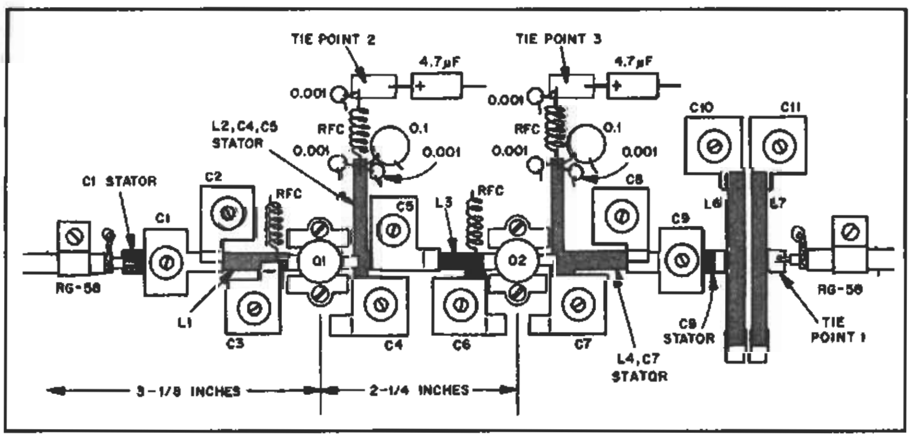
TIE POINT 2
TIE POINT 3
4.7µF
4.7µF
0.001
0.001
L2,C4,C5 STATOR
RFC
0.1
0.001
0.001
RFC
0.1
0.001
0.001
C10
C11
C1 STATOR
C2
RFC
C5
L3
RFC
C8
C1
Q1
Q2
C9
L6
L7
RG-58
L1
C3
C4
C6
C7
C8 STATOR
L4,C7 STATOR
RG-58
TIE POINT 1
3-1/8 INCHES
2-1/4 INCHES

Fig 6—Circuit board layout of the completed amplifier. Begin construction by measuring the center-to-center distance from Q1 to Q2 and then mount the transistors, working away from them as you mount the other components. Use great care when mounting and soldering the transistors into the circuit to avoid damaging them. The PC board measures 9½ in. x 6 in., and is made from double-sided, glass-epoxy material (Meshna PCB28—see note 1).

Initial Tune-Up

The dummy load/power-measuring assembly and narrow-band peak detector pick-off assembly used in tune-up are the same as the ones that I used with my 15-W transmitter.[4] The dummy load consists of ten ½-W resistors connected in parallel and arranged in a circle to minimize self-inductance. Power measurements were made by calibrating dummy-load temperature rise with a dc input (ambient air-temperature conditions, no forced-air cooling). This information is then compared to the temperature rise of the dummy load resulting from the RF output of the amplifier. The "thermistor" used for measuring the temperature rise is a 1N34A (reading the reverse resistance).

Under key-down conditions during alignment, I place a blower at the chassis end to assist the limited heat-sink capabilities of the amplifier. Also, the dummy load requires forced-air cooling, as it becomes very hot in about two minutes under keydown conditions.

Miscellaneous lengths of RG-8/M and RG-58 were used as attenuators to increase power measurement capability of this system.[5] In testing the 45-W amplifier, I used a total cable attenuation of 6.4 dB. This increases the power-measurement capability by 4.4 times.

With a 5-V collector supply and 2 W of drive, first tune the input capacitors for maximum Q1 collector current, then tune the remaining capacitors for maximum power output. The output should be about 8 W. Increase the collector supply to 13 V; minor readjustment will probably be required. My final measurements showed an output of 48 W.

Summary

This is a well-behaved amplifier. After completing an experimental version on the breadboard, I built a new assembly using the data in this article. Alignment of this final model was completed without any problems, and operation is flawless. The glue-down stripline technique is very reliable in this application, and is simple to use.

Notes

[1] J. Reed, "A Simple 435-MHz Transmitter," *QST*, May 1985, pp 14-18, 45.

[2] These transistors are available from RF Parts Company, 1320-16 Grand Aye, San Marcos, CA 92069, tel 619-744-0728.

[3] This note is reproduced in *Motorola RF Device Data* (fourth edition, first printing, 1986) on pages 6-61 through 6-65. Thus book is available from Motorola Literature Distribution, P0 Box 20912, Phoenix, AZ 85036, tel 602-994-6561. Cost is $4.75 (plus 15 percent of the total for shipping).

[4] See note 1.

[5] My 435-MHz measurements indicated that the RG-8/M I used has an attenuation of 9.0 dB per 100 feet (specified as 7.5 dB per 100 feet at 400 MHz); the RG-58 has 15.3 dB per 100 feet (specified as 12.0 dB per 100 feet at 400 MHz).

[6] Available from John J. Meshna, Jr. Inc, 19 Allerton St, Lynn, MA 01904, tel 617-595-2275.

By Tarmo Tammaru, WB2TMD From *QST*, May 1982

A Solid-State 6-Meter Linear Amplifier You Can Build

Does your 6-meter signal need a little more punch? This amplifier will solve your problem—without breaking the bank!

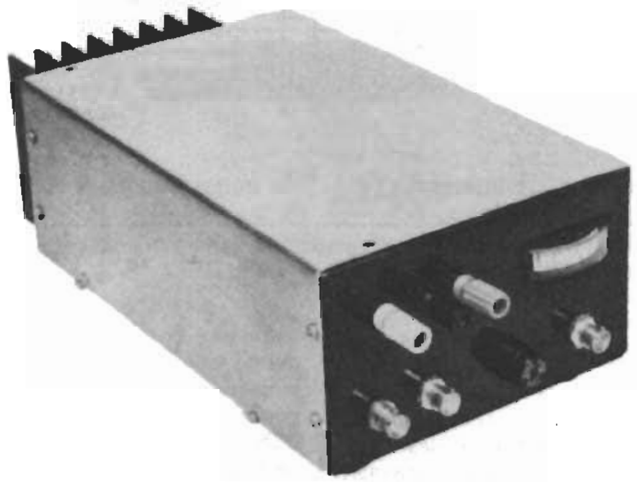

The new breed of 8-to 10-W, solid-state vhf transceivers offer compactness and relative freedom from TVI. However, there are times when 10 W will not do the job. This amplifier can correct that situation. I used it during the September vhf QSO party and was able to work every station I heard, except one.

The transistor used in the amplifier is a Motorola MRF-492, which is designed for use as an fm amplifier in the "vhf low" public-service band. In this service, the nominal output is rated at 75 W with a 12.5-V collector supply. The manufacturer's data sheet shows a maximum output of 120 W at 16 V. Hence, in ssb service, I felt it would be safe to operate at the 100-W PEP level from a 13.5-V supply. Although no numbers are given, the MRF492 is claimed to have "load mismatch capability at high line and rf overdrive."[1]

Circuit Description

The amplifier schematic diagram is shown in Fig. 1. Since I do not believe in "reinventing the wheel," the output matching network was taken directly from the Motorola data sheet. This network, incorporating L2 and L3 in the signal path, should provide excellent harmonic rejection. The network transforms the 50-Ω load impedance to the optimum transistor load of $0.6 + j1.0\ \Omega$. Note that this unit, like most rf amplifiers, does not have a 50-Ω output impedance. We simply adjust the output network to obtain the rated power into a 50-Ω load. Fig. 2A is an amplifier output spectral display. While the harmonic suppression is good, additional filtering is required before using the unit on the air. A simple, 5-pole, lowpass filter was placed at the amplifier output, and the resulting spectrum is shown in Fig. 2B. The low-pass filter circuit is given in Fig. 3.

Because the data sheet does not give values for input impedance in Class B service, the input circuit was derived basically by trial and error (the second version worked). The T matching network transforms the highly capacitive input impedance of approximately 1 Ω to a 50-Ω nonreactive load for the exciter.

In a Class B amplifier it is necessary to forward bias the base-emitter junction. There are basically two ways of doing this: the shunt diode method, and the emitter-follower method. The latter method undoubtedly results in lower inter-modulation products, but I have destroyed too many transistors using that method to try it with a $20 device. Forward bias is achieved by using the shunt regulator diode D1. This diode must be connected to the same heat sink as the MRF492, Q1. The voltage across D1 is adjusted to give a Q1 collector current of 100 mA with no drive applied. Do not attempt to set the bias point by measuring the base voltage of Q1 — use a milliammeter in the collector supply lead. I used an unmarked, surplus, 15-A rectifier diode for D1. Almost any diode will work, including power Zeners. I prefer a stud-mounted diode because it is easy to place in contact with the heat sink. Just be sure the polarity is such that the cathode is connected to the case. Several Motorola application notes specify a 1N4997 for this application. Unfortunately, this is a press-fit device. In any case, the heat sink must be grounded.

Parts Procurement

All parts, with the possible exception of an adequate heat sink, are readily available. The MRF492 has recently been advertised by Westcom[2] and Semiconductor Surplus.[3] RFC1 is a Nytronics shielded ferrite choke, designed to reduce coupling

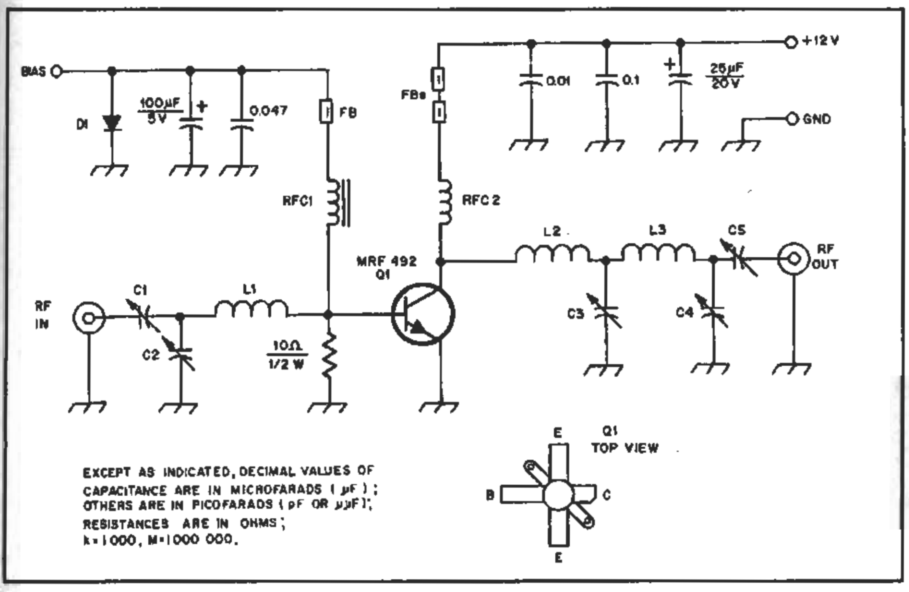
BIAS
D1
100µF 5V
0.047
FB
FBs
RFC1
RFC2
0.01
0.1
25µF 20V
+12V
GND
MRF 492 Q1
RF IN
C1
C2
L1
10Ω 1/2 W
L2
L3
C5
C3
C4
RF OUT
E
Q1 TOP VIEW
B
C
E
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.

Fig. 1 — The schematic diagram of the amplifier circuit board. Decimal value capacitors are disc ceramic and polarized capacitors are electrolytic. Resistors are 10% carbon type.

C1 — 9-to 180-pF mica compression trimmer, Arco 463 or equiv.
C2, C4, C5 — 50- to 380-pF mica compression trimmer, Arco 465 or equiv.
C3 — 80- to 480-pF mica compression trimmer, Arco 466 or equiv.
D1 — 15-A, 50-V rectifier diode (see text).
Fb — Ferrite bead, Radio Shack 273-1571 assortment or equiv.
L1 — 2 t. of no. 14 bare copper wire, $^{13}/_{32}$-inch ID, $^{3}/_{16}$-inch long.
L2— Half loop of no. 14 bare copper wire, $^{19}/_{32}$ inch high, $^{13}/_{32}$ inch long (see photograph).
L3— 2 t. of no. 14 bare copper wire, $^{13}/_{32}$-inch ID, $^{1}/_{4}$-inch long.
Q1 — Motorola rf power transistor, MRF492.
RFC1 — 6.8 µH, Nytronics SWD 6.8 or equiv. (see text).
RFC2 — No. 16 enameled wire, close wound over full length of a 330-Ω, 2-W carbon resistor.

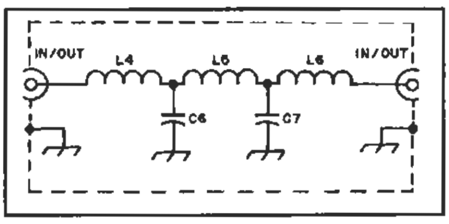

Fig. 3 — A simple low-pass filter for use with the 6-meter amplifier. The design information for this filter was taken from the 1982 ARRL *Radio Amateur's Handbook*. If the filter is mounted outside the amplifier cabinet, it should be enclosed in a metal box for shielding.

C6, C7 — 82-pF, 1000-V sliver-mica capacitor.
L4, L6 — 4 t. of no. 14 enameled wire. $^{3}/_{8}$-inch ID, $^{3}/_{4}$-inch long with $^{1}/_{8}$-inch leads (approximately 0.1 µH).
L5 — 6 t. of no. 14 enameled wire, $^{3}/_{8}$-inch ID, $^{1}/_{2}$-inch long with $^{1}/_{8}$-inch leads (approximately 0.2 µH).

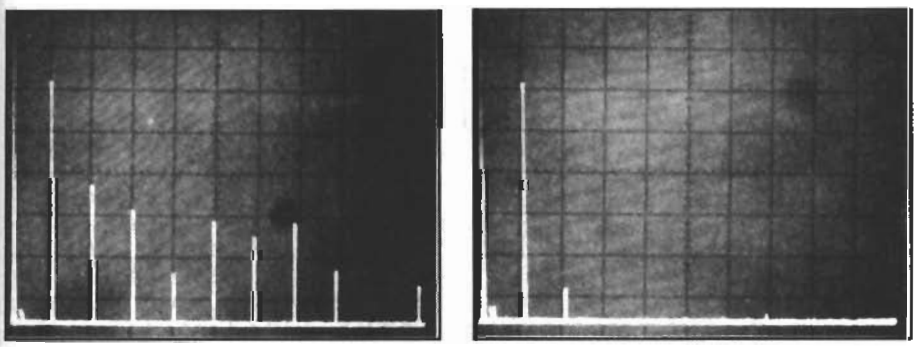

(A) (B)

Fig. 2 — The amplifier output matching network provides some harmonic suppression (A). To further reduce harmonics, a simple low-pass filter should be used (B). With the filter, all harmonics and spurious signals are more than 60 dB below the carrier, thus meeting current FCC spurious emission requirements for commercial equipment. In these displays, the carrier level has been reduced by means of a notch filter to avoid analyzer overload. Vertical divisions are each 10 dB and the horizontal divisions are each 50 MHz. These measurements were made in the ARRL lab.

between the base and collector circuit. The J. W. Miller 9250-682 should be as good. A toroidal inductor of 1 to 10 µH would also be suitable, since it is across an impedance of 1 Ω. Arco trimmers should be available from your local jobber. I used a surplus heat sink, measuring $3^{1}/_{2} \times 4^{1}/_{2} \times 1^{3}/_{4}$ inches, which is quite adequate for ssb.[4] Anything much smaller will require a fan. Remember, it must be flat on one side. A replacement heat sink for any of the 80- to 100-W 2-meter amplifiers would be ideal.

The Circuit Board

All the components shown in Fig. 1 mount on, or through, the circuit board (Fig. 4). Use double-sided glass epoxy board and cut away narrow channels to separate the various land (foil) areas. To form the channels, use a steel ruler and a sharp utility knife to cut through the foil. Next, pull a hot soldering iron along the strip of foil to be removed. You should see it curl up and away from the board. Start at one end, and pry up with the knife if it gets stuck. You will have to drill two holes, slightly more than $^{1}/_{2}$ inch in diameter, through the board to provide clearance for Q1 and D1. Use a small round file to make the two clearance areas for the flanges of Q1. Connect the ground foils on the top and bottom of the board together by placing short lengths of wire through holes drilled in the board. Place the wires close to the transistor emitter tabs and near each corner of the board. Use pieces of the same type of wire you used to make the inductors, and drill the holes just large enough to pass the wire. Solder the wire on both sides.

Assembly

Make sure you have all the parts on hand before you solder anything to the board. This will ensure that everything fits properly. To keep the inductors in place, I drilled holes through the board and stuck the ends of the wire into the holes. You will have to remove the foil from around these holes on the bottom of the board to keep from shorting the inductors to ground. A large drill bit works fine for this. Center the board on the heat sink, and drill the mounting holes for Q1 and D1. Although you may be able to use screws and nuts to attach Q1 and D1 (depending on the heat sink design), a better method is to tap the holes for the desired screw thread. Mount the board to the heat sink near the corners, shimming it with washers to the same depth as the Q1 mounting flange. Drop Q1 in from the top of the board. The tabs of Q1 should rest just above the top of the board when the flange is in contact with the heat sink. Do not put any upward pressure on the transistor tabs. Never cut or file an rf power transistor body — they contain berillium oxide. It is highly toxic in powdered form and could be inhaled or absorbed through the skin.

Fasten Q1 to the heat sink before sol-

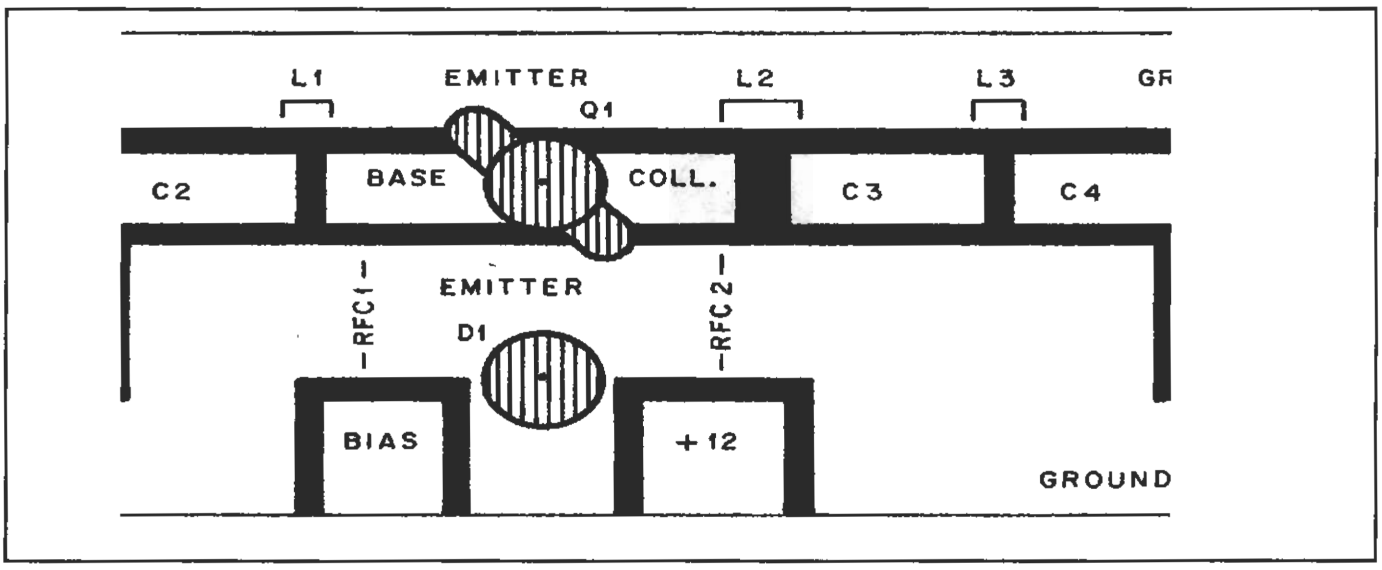

Fig 4—Full-scale circuit-board pattern and parts-placement guide for the 6-meter amplifier. Black represents those areas where copper has been removed by cutting or etching. The 10-ohm base resistor is connected between the base and ground foils, and is positioned next to RFC1.

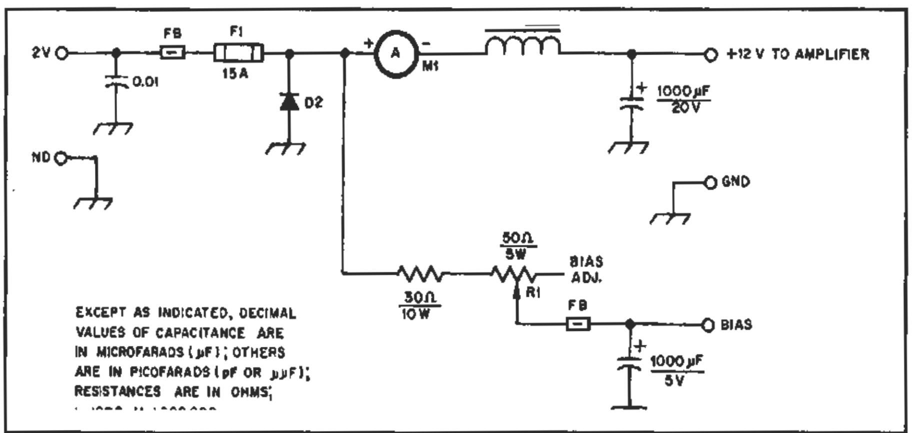
2V
FB
F1
15A
0.01
D2
ND
A
M1
+12 V TO AMPLIFIER
1000µF
20V
GND
50Ω
5W
BIAS ADJ.
R1
30Ω
10W
FB
BIAS
1000µF
5V
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS;

Fig 5—Schematic diagram of the bias-adjust circuit and other components not mounted on the circuit board. Decimal-value capacitors are disc ceramic, and polarized units are electrolytic type. Resistors are wire-wound type.

D2—20-A, 50-V rectifier diode.
RFC3—One turn of wire through a TV balun core (part of Radio Shack 273-1571 assortment or equiv.).

dering it in place. Use thermal grease when mounting Q1 and D1. Be sure to solder the full length of the tabs to the board, right up to the body of the transistor. This is especially important for the emitter tabs, because nanohenrys of inductance here translate to decibels of gain loss.

Break off the little tabs on the sides of the mica compression trimmers. Bend the ends of the mounting terminals with pliers to aid in soldering the trimmers to the board. In doing this, be sure the adjusting screw will not touch the board when it is all the way down. C2, C3 and C4 should be oriented so that the screw is connected to the ground end of the capacitor. For C1 and C5 let the screw end be the 50-Ω side.

Putting It All Together

I installed the circuit board and heat sink on the back wall of a 3½ × 6 × 10-inch Minibox. This is much larger than necessary, but does provide room for relays and a receiving preamp for an eventual remote installation. I usually discard the top (plain U-shaped) half of these boxes and bend up a new piece that fits over the other half, instead of sliding into it.

Fig. 5 shows the biasing circuit and other chassis mounted parts. Control schemes for use with a transceiver are described by Kapplin[5] and Ridpath[6] (also see the ARRL *Radio Amateur's Handbook*). D2 does not have to be heat sinked. It only has to last long enough to blow the fuse in case the wrong supply polarity is inadvertently applied. On the other hand, the 50-Ω bias adjust control (R1) gets quite warm, and should be mounted to the enclosure. In wiring the bias circuit, attach one wire from R1 to D1 and another wire from D1 to the bias point on the circuit board; all other connections are made to the circuit board. If R1 were connected to the circuit board, with a strap going to D1, you would almost certainly damage the transistor if the strap to D1 broke.

Alignment

It is necessary to set the bias before aligning the rf circuits. For an initial check, disconnect the wire between D1 and the circuit board (R1 is connected to D1). Apply power and measure the voltage across D1. It should be possible to lower it to 0.65 V, or less, by varying R1. Set R1 for the lowest voltage and connect the wire from D1 to the circuit board. Now reduce the resistance of R1 until the collector current is 100 mA. Be sure you are measuring collector current and not total supply current (you have about 300 mA flowing into D1).

For aligning the trimmers you will need, at least, an SWR indicator and a 50-Ω dummy load. You will, of course, not know how much power you are getting unless you also have a wattmeter. Connect the amplifier output to a dummy load and apply about 1 W of drive to the input through the SWR indicator. Adjust C1 and C2 for an SWR of 1:1. With C5 snug, but not tight, adjust C3 and C4 for minimum collector current. With the SWR indicator or wattmeter connected between the output and the dummy load, adjust all the trimmers for maximum output. If more than one setting gives the same power, pick the one that corresponds to the lowest collector current. Slowly bring up the drive and keep readjusting all the trimmers. With 8 W of drive you should get an output of 100 W. The collector current will be approximately 15 A. Readjust the input for a 1:1 SWR at full power. This may, or may not coincide with maximum gain and you may have to compromise.

During testing you might consider using a harmonica to generate a multitone signal to reduce the average power dissipation (keep the speech processor off). If at any time the heat sink gets so hot you cannot hold your hand on it for five seconds, let it cool off. As a rough indication of power output, I found that at 100 W, a

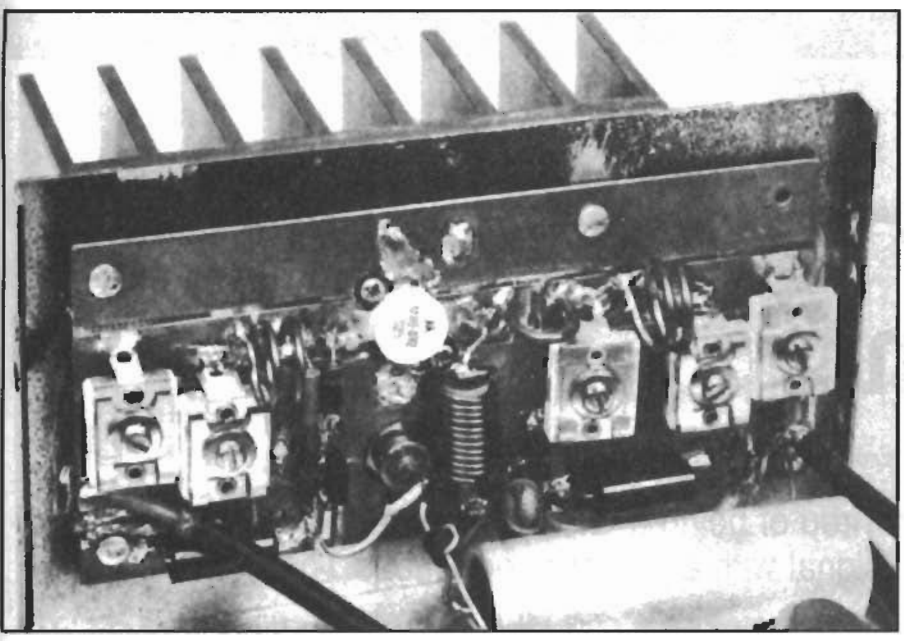

Interior view of the 6-meter amplifier showing the parts arrangement used by the author. The input matching network is at the right of the photo.

Drake DL 300 dummy load become too hot to hold after two minutes of key-down operation. To prevent rf burns, do not touch the dummy load while rf power is applied.

To maintain linearity, adjust the output circuits for 100 W, even if you are going to operate at 75 or 80 W. I evaluated the amplifier linearity by making two-tone IMD measurements with a spectrum analyzer. At a PEP output of 100 W, the third-order products were 25 dB below the PEP. At 80-W PEP they were down 30 dB. In over 100 contacts, I have not received any adverse comments on the signal quality.

Notes

[1] Motorola Inc., *Motorola Rf Data Manual*, 2nd ed. (Phoenix, AZ: Motorola Inc., 1980), p. 6-15.

[2] Westcom, 1320 Grand Ave., San Marcos, CA 92069.

[3] Semiconductor Surplus, 2822 N. 32nd St., No. 1, Phoenix, AZ 85008.

[4] mm = inches x 25.4.

[5] S. Kapplin, "Boots for QRP Rigs," *QST*, July 1981, pp. 15-20.

[6] I. Ridpath, "T-R Switching with PIN Diodes," *QST*, March 1981, pp. 19-21.

By Wilson Hoag, WA5OLT **From *QST*, April 1982**

Build A 6-Meter "Mini-Lini"

Tired of being a QRP station? Boost your low-level signal with a pair of sweep tubes.

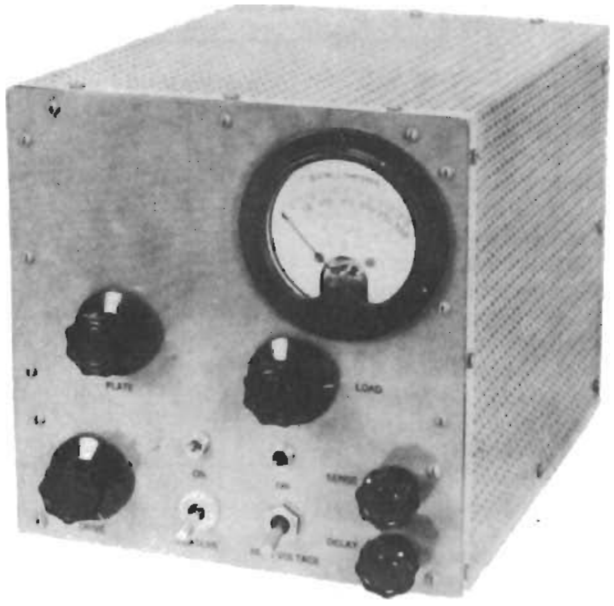

Is low power enough for equality on 6 meters? Perhaps, but it has long been my contention that some stations are more equal than others! One of the great features of 6 meters is the ability to make contacts with just a few watts when the band is open. As solar activity begins to decline, band openings will be shorter and less frequent. This amplifier will add punch to your QRP signal.

The current offerings of solid-state 6-meter transceivers and transverters provide operating ease and good signal quality, but at the expense of rf power output. This, coupled with the paucity of linear amplifiers, resulting from the infamous amplifier ban, has caused an overall reduction in the average power output on 6 meters.

When I acquired my first solid-state rig (an IC-551) some time back, my tendency was to use its scan feature to locate a signal and then work the station with a 240-watt tube-type rig. A minor panic occurred when the tube rig "rolled over and died" midway through the '79-'80 F2 season. The Mini-Lini resulted from that panic.

Design Approach

The amplifier was designed to be compatible with common power supplies, use a readily available tube, provide push-to-talk operation with the IC-551 and have approximately the same rf output as most tube-type transceivers. Fortunately, a design incorporating most of these features had been developed by Ed White, WA5RIA.[1] The major changes I made were in the methods of switching and bias adjustment. The tube used is the 6JB6, which is inexpensive, available and "happy" at 50 MHz. The biasing arrangement, an idea borrowed from Doug DeMaw,[2,3] allows the tubes to be matched for safe parallel operation. This amplifier provides about 50 watts of output when driven by a 10-watt exciter. More than one year of almost daily use has been trouble free and productive (42 states, including KH6 and KL7, plus JA and several Europeans).

Circuit Details

The circuit is conventional and typical of many low-band amplifiers. The original design used a capacitive input, but I thought that the parallel-tuned input method shown in Fig. 1 would provide more output when used with other QRP drivers. This was proven in practice: The amplifier delivers about 20 watts of output when driven by an IC-502. The potential problem of operating unmatched tubes in parallel has been avoided by biasing each tube individually. Component values shown will allow coverage of the entire 6-meter band.

The output network will not reject harmonic energy, and the second harmonic, as measured in the ARRL lab, was less than 40 dB below the fundamental (Fig. 2). This works out to a healthy 5 mW at 100 MHz. Most fm receivers in the neighborhood will detect this easily. FCC requirements for a commercial amplifier at this frequency are to have all spurious radiation at least 60 dB below the fundamental.

To meet this requirement (and to prevent a lot of RFI complaints) a 7-pole Chebyshev low-pass filter was built from data given in the 1982 *Radio Amateur's Handbook*, pages 6-11 and -12. Details of a filter with 0.01-dB of ripple and a 60-MHz cutoff frequency are given in Fig. 3. The coils can be wound on toroids, or be air-core types. Amidon T44-10 (or larger) cores should be adequate. The number of turns required will depend on core size or coil diameter, and can be calculated from data given on pages 2-12 and 2-30 of the *Handbook*. This filter reduced the second harmonic output to more than 65 dB below the fundamental (completely gone for all practical purposes). See Fig. 4.

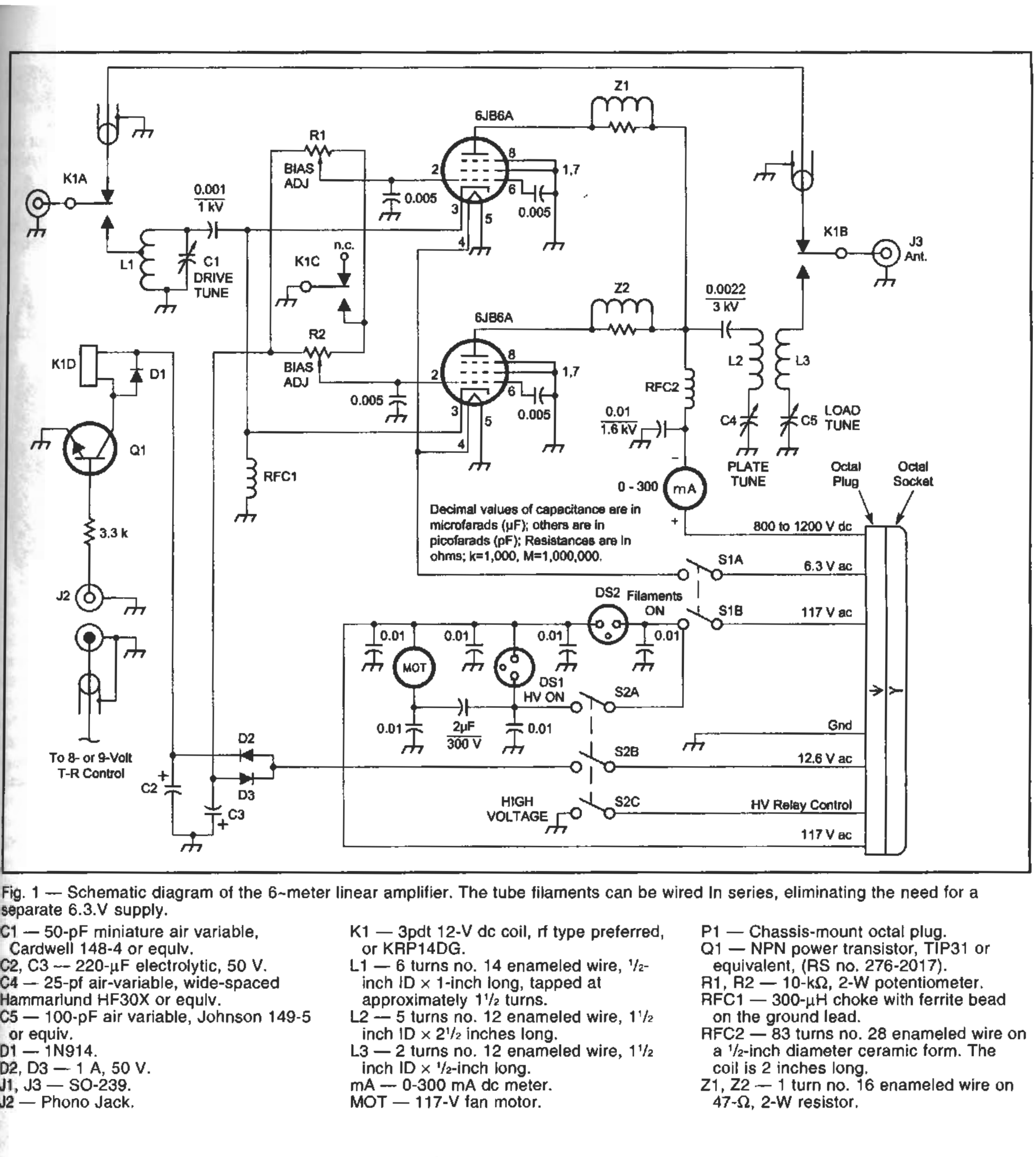
Z1
6JB6A
R1
BIAS ADJ
K1A
0.001
1 kV
0.005
1,7
0.005
L1
C1
DRIVE TUNE
n.c.
K1C
K1B
J3
Ant.
Z2
6JB6A
0.0022
3 kV
R2
BIAS ADJ
L2
L3
K1D
D1
0.005
RFC2
0.01
1.6 kV
0.005
C4
C5
LOAD TUNE
Q1
PLATE TUNE
Octal Plug
Octal Socket
RFC1
0 - 300
mA
3.3 k
Decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); Resistances are in ohms; k=1,000, M=1,000,000.
800 to 1200 V dc
S1A
6.3 V ac
J2
DS2 Filaments ON
S1B
117 V ac
0.01
0.01
0.01
0.01
MOT
DS1
HV ON
S2A
0.01
2µF
300 V
0.01
Gnd
D2
S2B
To 8- or 9-Volt T-R Control
12.6 V ac
C2
D3
HIGH VOLTAGE
S2C
HV Relay Control
C3
117 V ac

Fig. 1 — Schematic diagram of the 6-meter linear amplifier. The tube filaments can be wired in series, eliminating the need for a separate 6.3.V supply.

C1 — 50-pF miniature air variable, Cardwell 148-4 or equiv.
C2, C3 — 220-µF electrolytic, 50 V.
C4 — 25-pf air-variable, wide-spaced Hammarlund HF30X or equiv.
C5 — 100-pF air variable, Johnson 149-5 or equiv.
D1 — 1N914.
D2, D3 — 1 A, 50 V.
J1, J3 — SO-239.
J2 — Phono Jack.

K1 — 3pdt 12-V dc coil, rf type preferred, or KRP14DG.
L1 — 6 turns no. 14 enameled wire, ½-inch ID × 1-inch long, tapped at approximately 1½ turns.
L2 — 5 turns no. 12 enameled wire, 1½ inch ID × 2½ inches long.
L3 — 2 turns no. 12 enameled wire, 1½ inch ID × ½-inch long.
mA — 0-300 mA dc meter.
MOT — 117-V fan motor.

P1 — Chassis-mount octal plug.
Q1 — NPN power transistor, TIP31 or equivalent, (RS no. 276-2017).
R1, R2 — 10-kΩ, 2-W potentiometer.
RFC1 — 300-µH choke with ferrite bead on the ground lead.
RFC2 — 83 turns no. 28 enameled wire on a ½-inch diameter ceramic form. The coil is 2 inches long.
Z1, Z2 — 1 turn no. 16 enameled wire on 47-Ω, 2-W resistor.

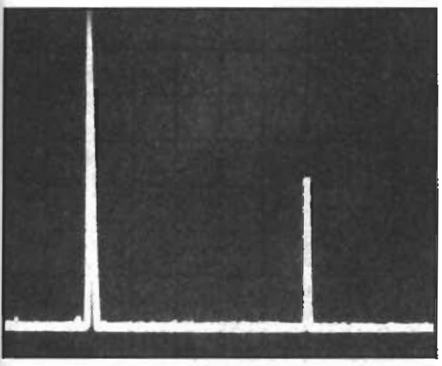

Fig 2—Spectral display of the amplifier output. Vertical divisions are each 10 dB, and horizontal divisions are each 10 MHz.

Construction

It should be possible to duplicate this amplifier for $40 or less, depending on the status of your junkbox. Oddly enough, the tube sockets proved to be the most difficult component to locate, since they are not a common catalog item. The 6JB6 uses a 9-pin NOVAR socket, and the best source turned out to be a shop that specialized in TV replacement parts. Tuning capacitor C4 is a junkbox item of questionable parentage. Any spacing greater than about 0.060 inch (mm = inches × 25.4) should be okay. RADIOKIT of Greenville, New Hampshire would be a good source for all of the variable capacitors.

My amplifier is constructed on an 8 × 10 × 2½ inch chassis (BUD AC-1418) for a base, with 8 × 8-inch end panels. The panels were flanged with ½ × ½-inch aluminum channel for attachment of a cane-metal cover for safety and TVI protection. Component placement is not critical, but all component leads should be as short as possible. The resulting layout was determined largely by some mid-project design changes aimed at making the unit compatible with a newly acquired IC-502. Visible on the lower-right front panel are potentiometers that were reserved for control of a planned rf-operated T-R relay.[4]

The tube sockets are modified slightly to ground the necessary tube elements. Copper washers, cut from flashing copper, fit inside

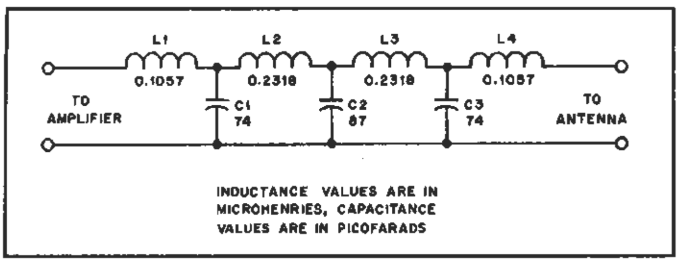
L1
0.1057
L2
0.2319
L3
0.2319
L4
0.1057
TO AMPLIFIER
C1 74
C2 87
C3 74
TO ANTENNA
INDUCTANCE VALUES ARE IN MICROHENRIES, CAPACITANCE VALUES ARE IN PICOFARADS

Fig 3—Schematic diagram of a 7-pole Chebyshev low-pass filter. Capacitors are silver-mica units, combined in parallel or series to obtain the design values. The text has information about winding the inductors.

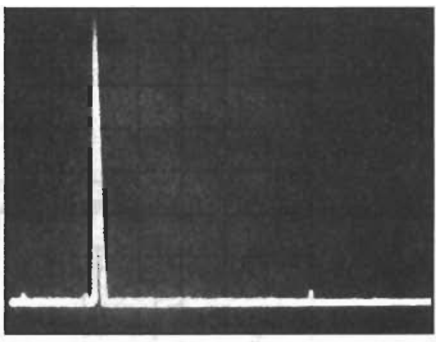

Fig 4—Spectral display of the amplifier output with a 7-pole Chebyshev filter. Vertical divisions are each 10 dB, and horizontal divisions are each 10 MHz.

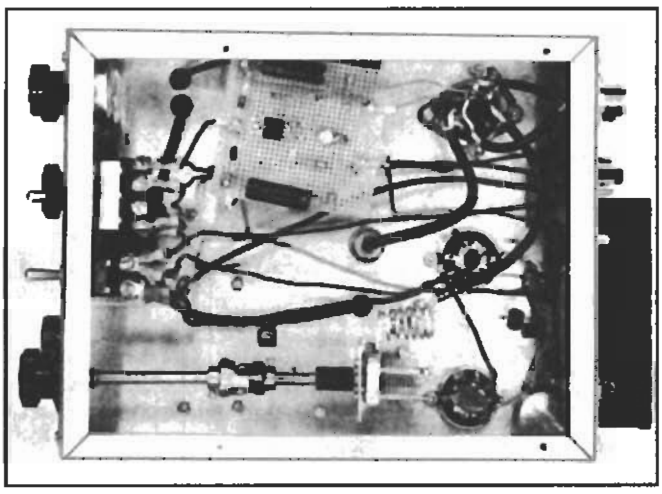

Fig 5—Bottom view of the amplifier chassis. Note the washers on the tube sockets, used to provide a ground connection for the appropriate pins.

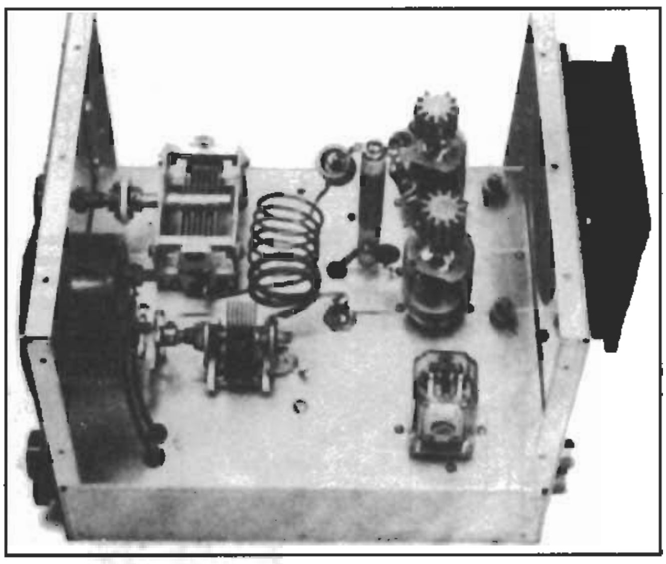

Fig 6—Top view of the Mini-Lini chassis. Large heat-dissipating plate caps and a fan help maintain cool operation of the tubes.

the pins on the underside of the sockets. Pins 1, 4, 7, 8 and 9 are bent over and soldered to the washer. The washers are grounded to the chassis with short pieces of no. 14 wire on opposite sides of the socket. Pins 2 and 6 of both tubes are bypassed to ground by connecting 0.005-µF disc-ceramic capacitors from these pins to the copper washers (Fig. 5). The control grids are raised above dc ground while providing a path for rf return. It also permits us to apply a dc bias to the control grid. This will establish the class of operation and cut the tubes off during receive, if desired.

Coils L1, L2 and L3 are wound with solid TW-insulated house wire from the local hardware store. I stripped the insulation from the wire before winding L1 and L2, but left it on L3 to prevent accidental contact with L2. The inside diameters shown are more the result of available cylindrical shapes than any electrical calculation. Coil spacing was adjusted, with the tubes in place, using a dip meter to ensure resonance at the proper frequency.

Heat-dissipating plate caps and a small fan provide cooling for the tubes. See Fig. 6. No thermal distress has been evident under any operating condition. The fan can be wired through S1 or S2, depending on whether cooling is desired during standby operation. Power connections to the fan should be isolated from the chassis.

A possible modification, shown on the schematic diagram, involves the use of a 3pdt relay for K1. The ground legs of R1 and R2 can be wired through one of the normally open contacts. This lifts the potentiometers above ground in the receive mode, applying full bias voltage and cutting off the tubes.

The bias voltage source and relay driver are "hard wired" on a small piece of perf board. The signal for T-R switching is applied to J2. This + 8-V signal is obtained from pin 6 of the IC-551 accessory socket. For use with other rigs, a 9-V battery, wired through a foot switch, works well. Current drain on this battery is low. An rf-operated T-R relay could be used in place of the directly keyed one described in this article. S2 disables the relay driver by removing the 12.6-V ac source in the standby position to permit straight-through operation.

Tune-Up and Operation

Initial tune-up is simple and ordinary. Connect a dummy load and the power supply to the appropriate jacks. Any high-voltage power supply that has an output of 750- to 1200-V dc should be satisfactory. Turn on S1 and allow the heaters to warm up for a minute or more. Switch S2 to turn on the other voltages. Actuate K1 to ground the bias resistors, R1 and R2. Adjust R1 and R2 to obtain 15 mA of idling current for each tube (30 mA total).

Connect the exciter to the amplifier through an SWR indicator. Actuate K1 and apply a small amount of drive. Adjust the position of the tap on L1 for minimum SWR. Be sure to remove all voltages each time you move the tap position!

Next, apply drive to the amplifier and adjust C1, C4 and C5 for maximum power output. For those lacking a wattmeter, an SWR indicator set in the FORWARD position can be used in the line between the amplifer and the dummy load.

Conclusion

The Mini-Lini should be ideal for those looking for more output from their solid-state rigs. Typical operating parameters are 50 watts out for 10 watts of drive, with 1050-V dc on the plates. Reports of signal quality have been complimentary. Enhance your "equality" and come join the fun on 6! I would be happy to answer any questions about the amplifier — enclose an s.a.s.e., please.

Notes

[1] The design was taken from personal correspondence with Ed White, WA5RIA.

[2] D. DeMaw, "Some Ground Rules for Sweep-Tube Linear-Amplifier Design," *QST*, July 1968, p. 30.

[3] D. DeMaw, "Some Thoughts About TV Sweep Tubes," *QST*, Feb. 1980, p. 11.

[4] D. DeMaw and J. Rusgrove, "An RF-Sensed Antenna Change-over Relay," *QST*, Aug. 1976, p. 21.

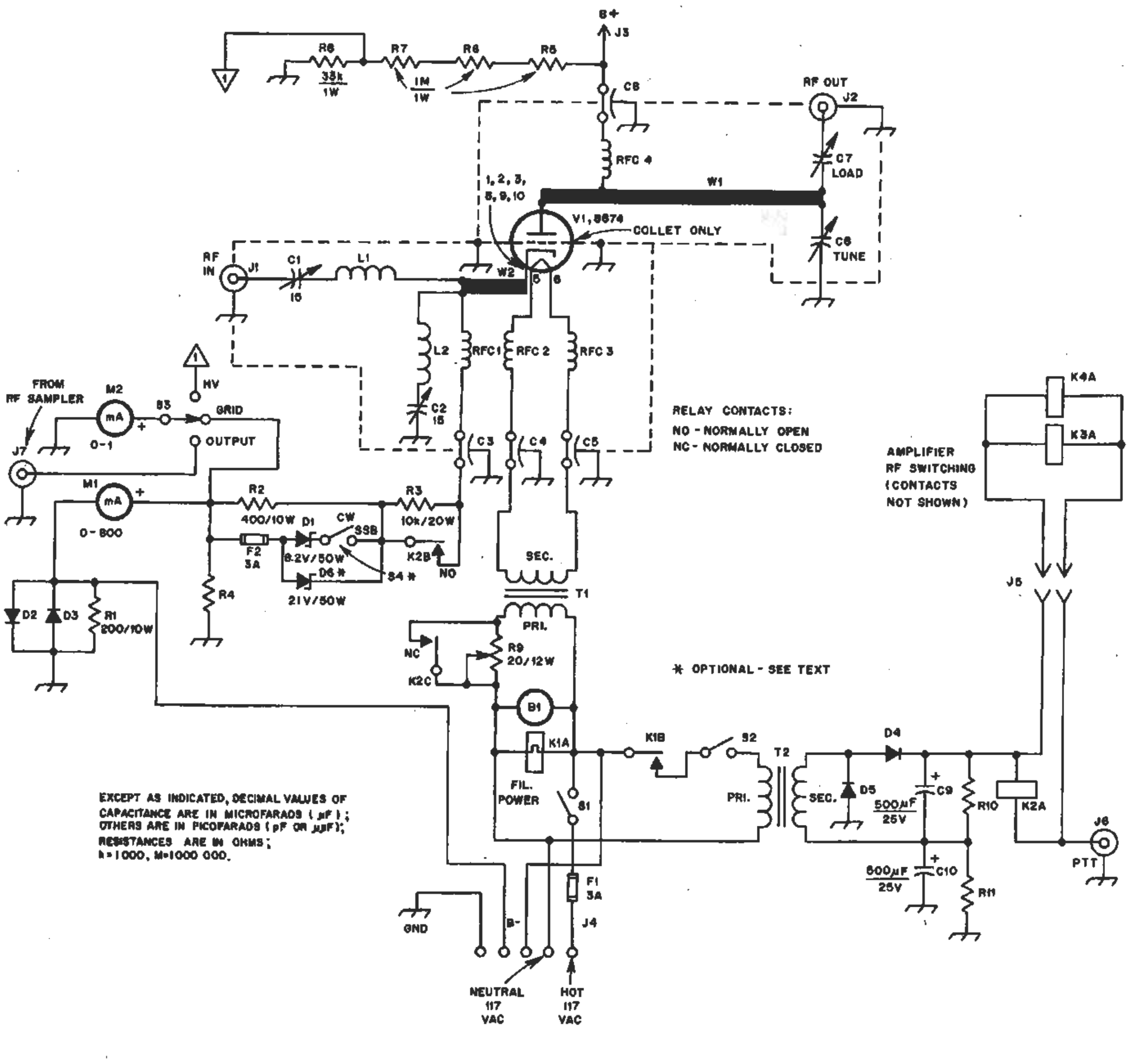
B+
J3
R8
33k
1W
R7
1M
1W
R6
R5
C8
RF OUT
J2
RFC 4
W1
C7
LOAD
1, 2, 3,
8, 9, 10
V1, 8874
COLLET ONLY
C6
TUNE
RF
IN
J1
C1
15
L1
W2
5
6
L2
RFC1
RFC 2
RFC 3
HV
C2
15
FROM
RF SAMPLER
M2
mA
S3
GRID
OUTPUT
0-1
J7
C3
C4
C5
RELAY CONTACTS:
NO - NORMALLY OPEN
NC - NORMALLY CLOSED
K4A
K3A
AMPLIFIER
RF SWITCHING
(CONTACTS
NOT SHOWN)
M1
mA
0-800
R2
400/10W
D1
CW
SSB
R3
10k/20W
F2
3A
8.2V/50W
D6 *
S4 *
K2B
NO
SEC.
R4
21V/50W
T1
D2
D3
R1
200/10W
J5
PRI.
NC
R9
20/12W
K2C
* OPTIONAL - SEE TEXT
B1
K1A
K1B
S2
T2
D4
FIL.
POWER
S1
PRI.
SEC.
D5
500µF
25V
C9
R10
K2A
J6
PTT
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.
500µF
25V
C10
R11
F1
3A
GND
B-
J4
NEUTRAL
117
VAC
HOT
117
VAC

Fig. 1 — Schematic diagram of the amplifier.

B1 — 50-ft^3 (1.4-m^3)/min blower, Ripley Sk2754-2A or equiv.
C1, C2 — Air-variable capacitor, 15 pF, E.F. Johnson 189-0565-001, 160-0107-001 or equiv.
C3-C5, incl. — Feedthrough capacitor, 500 pF, 300 V.
C6-C8, incl. — Homemade "flapper" capacitor. Details of construction in text and Fig. 3.
C9, C10 — Electrolytic capacitor, 500 µF, 25 V.
D1 — 50-watt, 8.2-volt Zener diode, IR Z-3307-C or equiv.
D2-D5, incl. — 1-A, 1000-PIV diode, 1N4007 or equiv.
D6 — 50-watt, 21-volt Zener diode (optional — see text).
F1, F2 — 3AG fuses.
J1 — Chassis mount BNC female connector, UG-1094/U.
J2 — Chassis mount N female connector, UG-58A/U.
J3 — High-voltage connector, Millen 37001.
J4, J5 — Power connectors, as available.
J6, J7 — RCA phono jacks.
K1 — Time-delay relay, 90 second, normally open contact, Amperite 115N090T.
K2 — Control relay, 28-volt coil, 1-A 4pdt contacts.
K3, K4 — Coaxial relays equipped with suitable connectors. K4 should have N connectors, K3 may be BNC or N.
L1 — 3½ turns no. 16 enam. wire, ¾-inch (19 mm) long, ¼-inch (6 mm) diameter.
L2 — 1½ turns no. 16 enam. wire, ⅝-inch (16 mm) long, ¼ inch (6 mm) diameter.
M1 — 1-mA meter movement with shunt to provide 600-mA full-scale deflection.
M2 — 1-mA meter movement with shunts to provide 90-mA (grid current) and 3-kV (plate voltage) full-scale deflection.
R4 — Grid-current shunt.
RFC1 — 10 turns no. 18 enam. wire, close wound, ¼ inch (6 mm) diameter.
RFC2, RFC3 — 10 turns no. 16 enam. wire, close wound, ¼ inch (6-mm) diameter.
RFC4 — 5 turns no. 16 wire, one inch (25 mm) long, ¼ inch (6 mm) diameter.
S1 — Toggle switch, spst.
S2 — Toggle switch, spst.
S3 — Rotary switch, single pole, three position.
S4 — Toggle switch, spst (optional, see text).
T1 — Filament transformer, 6.3-volt, 3-A, Stancor P-6466 or equiv.
T2 — Transformer, 12.6 volts, 1 A.

By Stephen J. Powlishen, K1FO **From *QST*, October 1979**

A Grounded-Grid Kilowatt Amplifier for 432 MHz

Stable, linear operation for tropo or moonbounce DX... that's the end result of this project.

In the last few years I've built several high-power 432-MHz amplifiers that used tubes from the 4CX250 family. While they worked well in Class C, their performance when biased for linear operation left something to be desired. My previous experience with grounded-grid triode amplifiers on 2 meters was so good that I decided to try the same approach on 70 cm. An Eimac 8874 high-mu triode was selected for this design and a crude prototype was built in a few evenings. After the design was verified, the amplifier described in this article was built. It is stable, compact and delivers over 500 watts output while requiring only a high-voltage supply and a source of about 25 watts of drive. The amplifier has been trouble free in over three years of heavy usage.

Circuit Description

A schematic diagram of the 432-MHz kilowatt is given in Fig. 1. W1 is a half-wavelength stripline which is tuned and loaded by C6 and C7 respectively. Plate choke RFC4 is connected at the approximate electrical center of the plate line. C8 functions as the plate-bypass capacitor. The half-wavelength cathode line is comprised of W2, L2 and C2. L1 and C1 serve to match the tube input impedance to the amplifier 50-ohm input. As the grid is grounded for dc as well as rf, D1 is used to develop operating bias at the cathode. R3 is switched in to supply near-cutoff bias during standby periods. M1 is used solely to monitor plate current in the high-voltage supply negative-return Lead. M2 is switched to read grid current, high voltage and relative output. The latter function is by means of an external line sampler.[1] With the exception of the multimeter functions the metering and bias circuits are similar to those in a 220-MHz amplifier.[2]

Separate coaxial relays attached to the

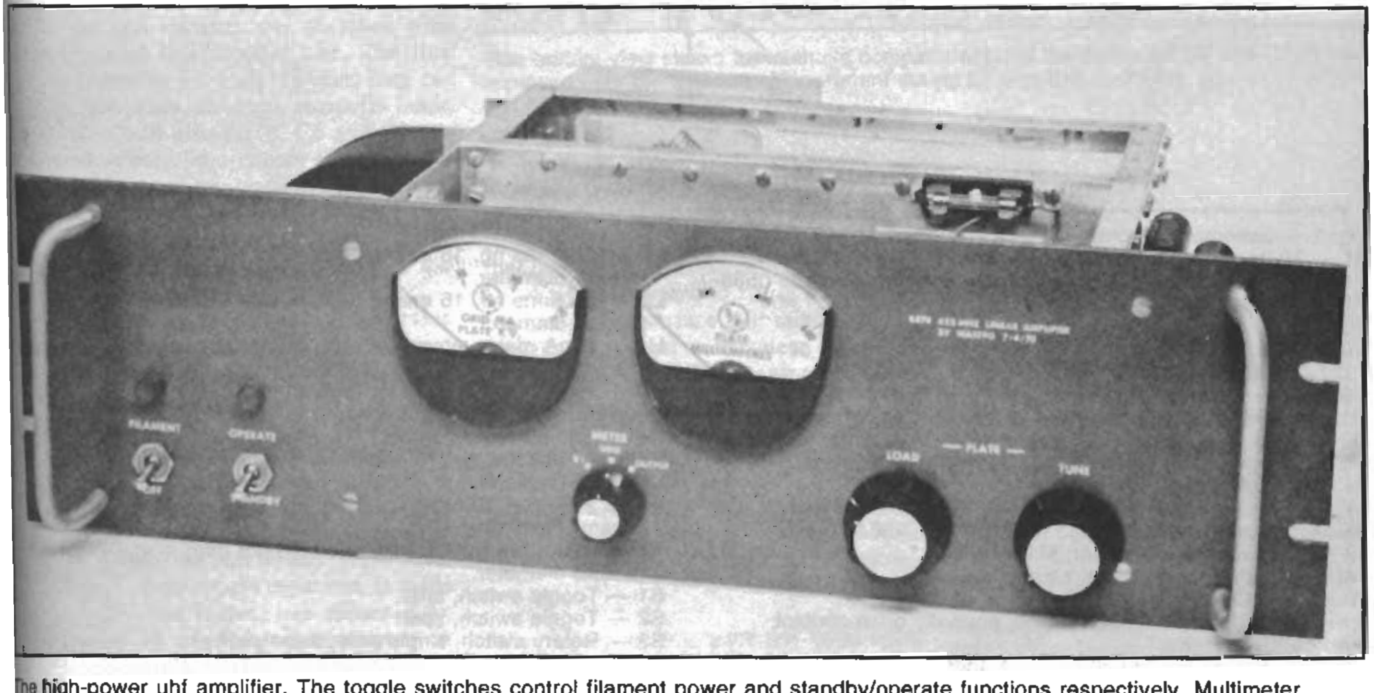

The high-power uhf amplifier. The toggle switches control filament power and standby/operate functions respectively. Multimeter function is selected with the switch located between the meters, while the plate tuning and loading controls are at the right. Modern knobs and homemade meter faces give the amplifier a commercial appearance.

input and output terminals allow the amplifier to be switched in and out of the line in a manner popular with hf amplifiers. Time-delay relay K1 prevents the amplifier from being switched into service for 90 seconds after the tube heater is energized, allowing the element to reach operating temperature. A normally closed contact of K2 applies full voltage to the heater during standby periods. The voltage is reduced during operation as recommended by the manufacturer.

Construction

Plate and cathode-compartment construction is from 0.032-inch (0.8-mm) thick aluminum sheet attached to ¹/₂-inch (13-mm) aluminum angle stock. Some angle stock may be anodized, giving the surface a dull appearance. This material must be lightly sanded to remove the anodized metal, which is a poor conductor. Holes are drilled in the angle stock to allow attachment of the covers; these are tapped for no. 4-40 screws. Details of the 10.5 × 4 × 3-inch (267 × 102 × 76-mm) plate compartment may be seen in the top view photo. Construction of the cathode compartment is similar, and may be seen in the photo of the underside. It measures 4 × 4 × 1³/₄ inches (102 × 102 × 44 mm). The aluminum brackets holding the rf enclosures to the front panel also serve as end covers for the compartments. Compartment spacing from the panel is four inches (102 mm). A 5¹/₄ × 19-inch (133 × 483-mm) rack panel is used.

The plate line was fabricated from a piece of ¹/₁₆-inch (1.6-mm) thick brass. Fig. 2 gives detailed information for making the line. In addition to brass, lines were made from copper, both unplated and silver plated, with no discernible difference in efficiency. Double-sided G-10 printed circuit board would probably work as well. Best thermal stability was obtained with the unplated solid-copper line. The line is supported by 1.5-inch (38-mm) long ceramic insulators, although standoffs made of Teflon will also serve. C6 and C1 are made from beryllium-copper sheet. Details of their construction appear in Fig. 3. These "flappers" are moved with fishing line which is tied to ¹/₄-inch (6.4-mm) fiber shafts. These shafts may be seen in the underside view.

The anode collet (Eimac no. 008294) is secured to the bottom of W1 with standard 60/40 solder. Use no. 4-40 screws and nuts to hold the collet in place during the soldering operation. The grid collet (Eimac no. 882931) is attached to the chassis with eight no. 4-40 machine screws and nuts. A poor ground connection for the grid will greatly increase the amplifier drive requirements or make the unit totally inoperative.

C8, the plate-bypass capacitor, is made from two brass plates, one mounted on either side of the plate compartment. A 0.005-inch (0.13-mm) thick piece of Teflon sheet is used for the dielectric material. While this Teflon thickness may seem inadequate, it is rated at 1000 volts per mil (0.03 mm) thickness. It is necessary to coat the dielectric with Dow Corning type DC-4 silicone grease to fill in any imperfections in the surface that might allow a leakage path and subsequent capacitor breakdown. This silicone grease has dielectric properties similar to Teflon* A no. 8 (4-mm) brass screw is used to hold the plates in place, and also acts as the high-voltage feedthrough terminal. A 3/8-inch (10-mm) diameter washer was sliced from a Teflon rod and used to center the screw in the hole. Fig. 4 gives details of the remain-

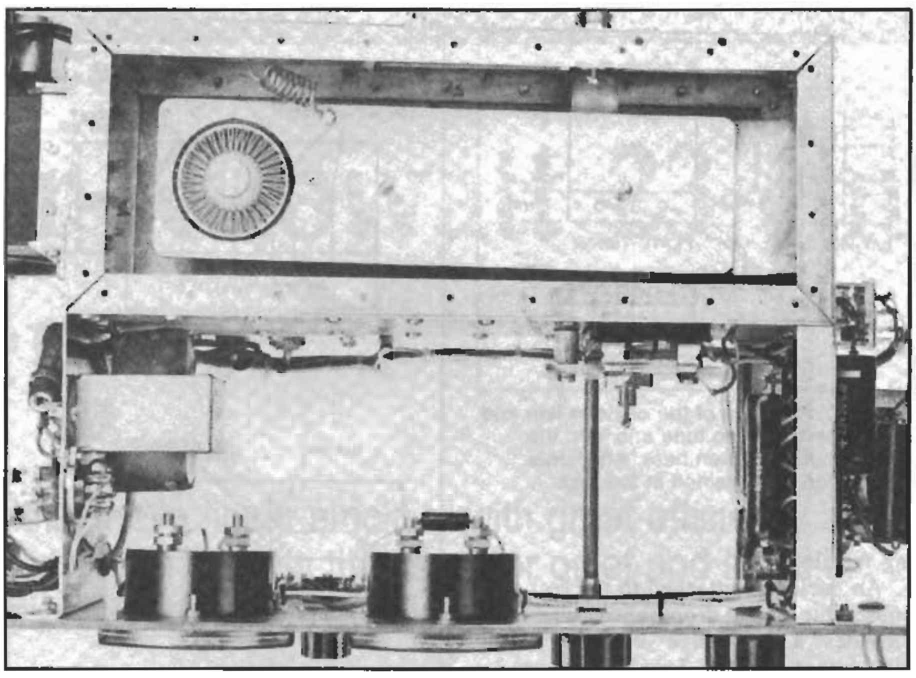

Top view of the amplifier, with the plate compartment cover removed. The tube, plate line (W2) and RFC4 may be seen at the top of the photo. Note the large number of holes drilled in the plate compartment to receive the cover hold-down screws. A tight seal is required to prevent rf and air leaks.

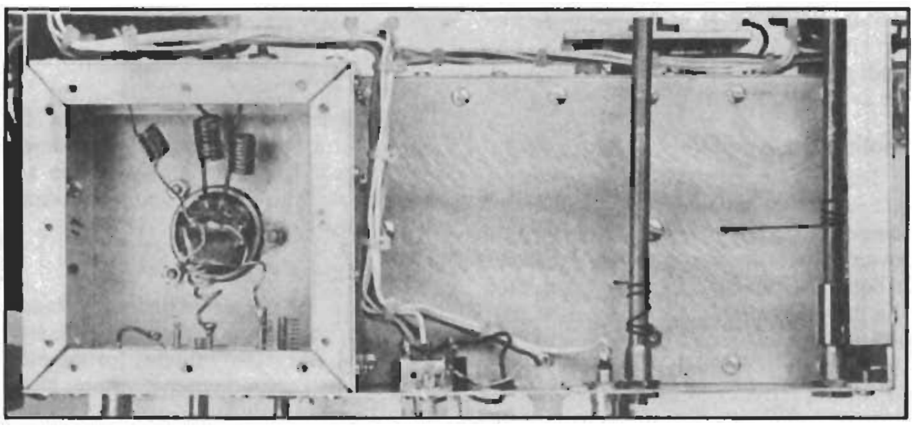

This bottom view shows the cathode compartment and the shafts for C6 and C7. A cover is placed over the cathode compartment during tuneup and operation.

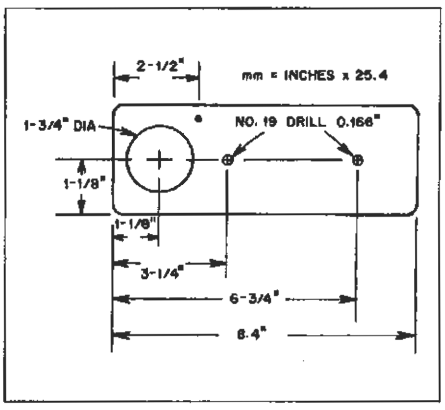

Fig 2—Dimensions of the plate line are given here. The line may be constructed from ¹/₁₆-inch (1.6-mm) thick copper or brass. Corners of the line should be filed to give a ³/₁₆-inch (5-mm) radius.

---

*Alternatively, an Erie 2498-001-X5U0-102M 1000-pF 4-kV feedthrough capacitor may be used. This component is available from ARCOS, P.O. Box 546, East Greenbush, NY 12061.

By E.R. "Chip" Angle, N6CA — From *QST*, March 1985

A Quarter-Kilowatt 23-cm Amplifier

Imagine, a linear amplifier with great efficiency and long-term stability that is super quiet and small in size. Sounds like HF? Not exactly...

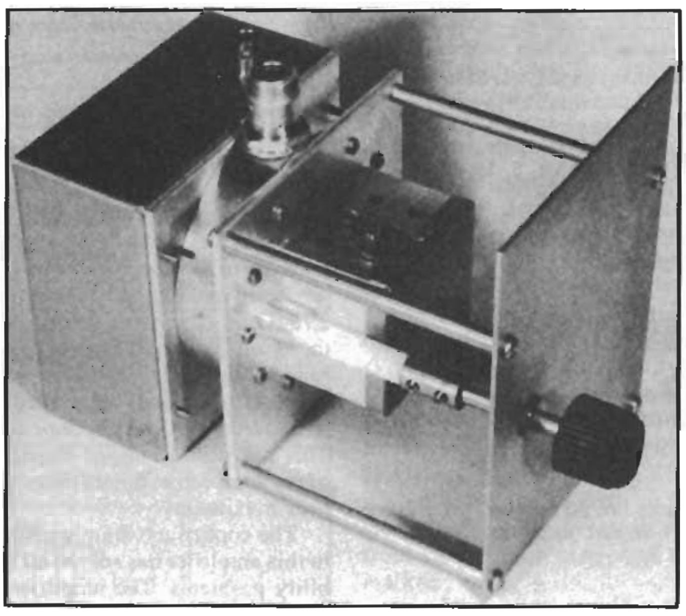

To me, there is nothing more frustrating than having to dig through a construction article to find out exactly what performance you can expect from the finished product. So here it is:

1) Grounded-grid 7289/2C39 cavity amplifier, single tube.
2) Linear operation (what you put in, you get out, only more of it).
3) Covers 1240 to 1300 MHz.
4) Power. gain ranges froth 12-20 dB depending on output power, input power, loading, anode voltage and grid bias voltage.
5) 50-ohm input and output — no stub tuner required.
6) Power output greater than 200W with about 12-W drive.

This is Part 1 of a two-part article. In this installment, I describe the design and construction of the RF deck. Part 2 describes power-supply construction, testing and operation.

This amplifier is a tried and proven design. Much development work has gone into this project. The amplifier works well, is reliable and can be duplicated. More than 50 of these amplifiers have been built to date. I have successfully worked many 1296-MHz EME (earth-moon-earth) stations with one of these amplifiers and a 384-element loop-Yagi array during the past year. Amplifiers of this design were used on both ends of the first California-to-Hawaii QSO on 1296 MHz. Another unit has logged more than 20,000 hours of continuous operation at the KH6HME beacon.

General Design Approach

A cavity amplifier is similar to a conventional amplifier designed for lower frequencies. The tube anode excites a resonant circuit, and power is in turn coupled into a load, usually 50 ohms. Instead of using coils and capacitors, as at lower frequencies, the cavity provides the resonant circuit necessary to tune the amplifier output.

The anode cavity of this amplifier is a squat cylinder. Cylinder height is set by mechanical tube requirements. The inside diameter of the cylinder sets the highest resonant frequency. Any capacitance added from the top to the bottom of the cavity will lower its resonant frequency, as will increasing the cavity diameter.

This amplifier uses 1/8-inch-thick copper plates for the cavity top and bottom, and a thick-wall aluminum ring, cut from tubing, for the walls.[1] This heavy construction virtually eliminates all resonant-frequency variations caused by thermal and mechanical changes.

Fig. 1 is a schematic diagram of the cavity amplifier. The circuit is simple. Filament voltage and cathode bias enter the RF deck through feed-through capacitors (C4, C5) and RFC1 and 2. High voltage is fed to the anode through RFC3. C8, the anode bypass capacitor, is homemade from Teflon® dielectric sandwiched between a copper plate and the chassis.

The input pi network easily tunes the entire band at any power level. it is made from two Johanson piston trimmer capacitors and a "coil" made from copper wire. An input cavity is not necessary at 23 cm.

Output coupling is through a rotatable loop that serves as a variable loading control. This allows amplifier-tuning flexibility; it may be tuned for maximum gain or for maximum power. Light loading can produce stable power gains of up to 20 dB.

Amplifier tuning is accomplished with a homemade cylindrical coaxial capacitor with Teflon dielectric (C6). There are no moving metal parts to cause erratic performance. The Teflon rod/tube screws in and

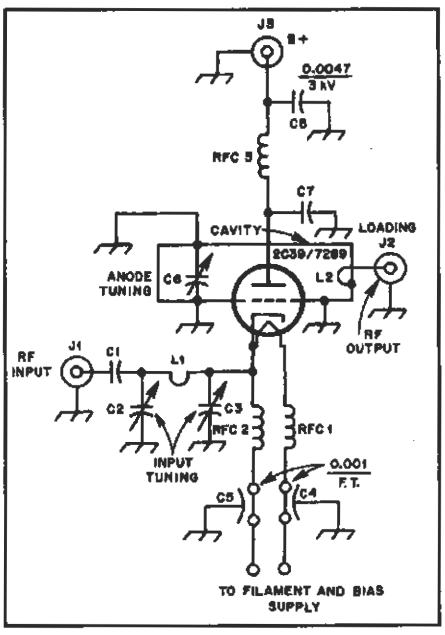
J3
B+
0.0047
3 kV
C8
RFC 3
C7
CAVITY
LOADING
J2
2C39/7289
ANODE TUNING
C6
L2
RF OUTPUT
RF INPUT
J1
C1
L1
C2
C3
RFC 2
RFC 1
INPUT TUNING
0.001
F.T.
C5
C4
TO FILAMENT AND BIAS SUPPLY

Fig. 1 — Schematic diagram of the 23-cm amplifier.

C1 — 3-pF dipped mica capacitor.
C2, C3 — 1- to 10-pF piston trimmer capacitor (Johanson no. 3957, 5201 or equiv.).
C6 — Anode-tuning capacitor. See text and Fig. 11.
C7 — Anode-bypass capacitor, 90 pF. Homemade from copper plate and Teflon sheet. See text and Figs. 5, 12 and 15.
C8 — Disc ceramic, 0.0047-µF, 3-kV capacitor.
J1 — 5-mm SMA connector, chassis mount, female.
J2 — Modified Type-N connector. See text and Fig. 7.
J3 — Female chassis-mount BNC connector.
Li — Loop of no. 18 bus wire soldered between C2 and C3. See Fig. 15.
L2 — Output-coupling loop. Part of output-connector assembly. See text and Fig. 7.
RFC1, RFC2 — 5 turns no. 20 bus wire, $^3/_{16}$-inch ID.
RFC3 — 3 turns no. 20 bus wire wound on a 20-ohm, 1-W carbon-composition resistor.

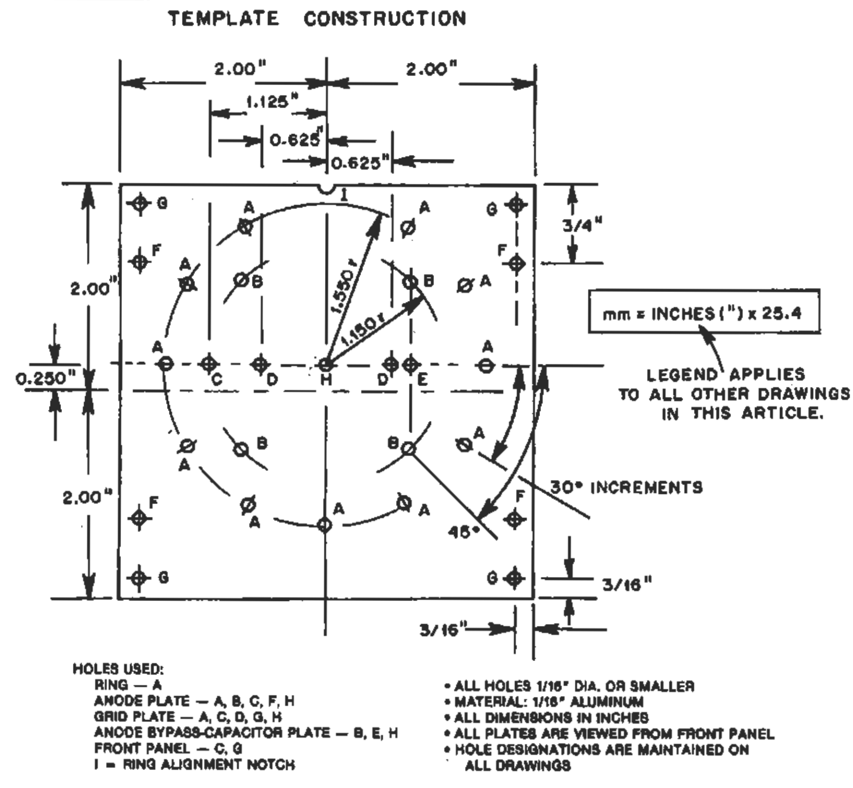
TEMPLATE CONSTRUCTION
2.00"
2.00"
1.125"
0.625"
0.625"
I
G
A
F
B
1.550 r
1.150 r
3/4"
2.00"
0.250"
C
D
H
D
E
mm = INCHES (") x 25.4
LEGEND APPLIES TO ALL OTHER DRAWINGS IN THIS ARTICLE.
30° INCREMENTS
45°
2.00"
3/16"
3/16"
HOLES USED:
RING — A
ANODE PLATE — A, B, C, F, H
GRID PLATE — A, C, D, G, H
ANODE BYPASS-CAPACITOR PLATE — B, E, H
FRONT PANEL — C, G
I = RING ALIGNMENT NOTCH
• ALL HOLES 1/16" DIA. OR SMALLER
• MATERIAL: 1/16" ALUMINUM
• ALL DIMENSIONS IN INCHES
• ALL PLATES ARE VIEWED FROM FRONT PANEL
• HOLE DESIGNATIONS ARE MAINTAINED ON ALL DRAWINGS

Fig 2—Complete dimensions for the aluminum template.

out of the coaxial capacitor, increasing or decreasing the capacitance by changing the amount of Teflon dielectric inside the cylinder. With the rod all the way in, the dielectric is all Teflon; with the rod all the way out, the dielectric is all air.

Teflon has a relative dielectric constant (relative to air = 1) of 2.05, which means that the value of the capacitor with the rod all the way in is twice the value of the capacitor with the rod all the way out. Full capacitance will pull the resonant frequency of the amplifier down to 1240 MHz. Use of only one tuning adjustment means the amplifier will have more gain because cavity shunt capacitance has been minimized.

Thermal Considerations

The cavity walls are formed by a thick-wall aluminum ring, which is sandwiched between two thick copper plates. RF and thermal properties of these two metals are reasonably close, whereas brass is rather poor in both respects. The 7289/2C39 tube used in this amplifier is being run at 2-2$^1/_2$ times its normal dissipation rating; therefore it's important to have a cavity that remains thermally stable.

Most previously described amplifiers have used sheet brass in their construction. This has usually meant constant retuning of resonance to maintain output power at or near maximum.

The copper and aluminum construction in this amplifier has solved all thermal stability problems. The amplifier can easily be run key down for over an hour at 200-W output without retuning. This, of course, is obtained only with a good tube and water cooling. A practical water cooling system will be described in Part 2 of this article.

Water cooling keeps the internal structure of the tube thermally stable. When air cooling is used for output levels of 100 to 150 W, output power fluctuations are a direct result of internal tube changes. These changes vary from tube to tube and must be tested for. In some cases, otherwise perfectly good RF tubes have had poor thermal stability. Such tubes can make good drivers at lower power levels.

"Using Simple Hand Tools Will . . ."

Hand tools are great if you are skilled and patient. Most people want to hurry up and finish their new project. If that's you, then have a machine shop make all of the parts, leaving you only the final assembly. It should cost about $200. The parts are not difficult to fabricate, but the process is time consuming. If you have the time and patience to do it yourself, this amplifier can be very inexpensive.

Gathering the Materials

All of the materials used in this amplifier are fairly common and should be available from suppliers in most metropolitan areas. Some suppliers have "short sale" racks, where they sell odd pieces cut off standard lengths or sheets at reduced prices. The parts for this project are small enough to be fashioned from cutoff stock. Surplus-metal houses have some great buys, so start there if one is nearby.

The key to successfully completing this project is careful layout work before cutting or drilling any parts. Invest in a can of marking dye, a sharp scribe, an accurate rule, vernier calipers and several center punches. These tools are available at any machinists' supply shop. The marking dye will make cutting and filing lines much easier to see. Measure all dimensions as carefully as you can and then recheck them before cutting. Mark with a sharp scribe because the sharper the scribe, the finer the marked line, and the finer the marked line, the closer your cut will be to where it should be. Remember—the accuracy of your drilled holes is only as good as your center-punching ability, so use a fine punch for the first mark and then a bigger one to enlarge the mark enough for drilling.

Access to a drill press is a must. It's extremely difficult to drill holes accurately with a hand drill. Although they are not absolutely necessary, you should have access to a lathe or milling machine.

Other tools that will aid you with this project are a nibbling tool, a set of punches, a new set of files and some sharp drill bits. If you don't already have one, purchase a file card to clean metal shavings out of your files as you work. Clean, sharp files are faster and more accurate to work with. You'll also need an assortment of sandpaper for the final finish work.

The Template Approach

I highly recommend fabrication of a single template for marking and drilling the anode plate, anode bypass capacitor, cavity ring, grid plate and front panel. The template shown in Fig. 2 has all of the holes for these parts. If you use the template, you'll only have to make the careful measurements once — after that, it's simple to mark and drill the rest of the parts.

The template approach offers several other advantages. A template makes it much easier to maintain accuracy between the anode plate, cavity ring, grid plate and front panel; these parts will fit perfectly because they were all drilled from the same master. The template approach also makes it possible to set up a small production line if you decide to build more than one of these amplifiers and combine them for higher power, or if a friend wants to build an amplifier along with you.

See Fig. 2 for complete template dimensions. Start with a piece of 1/16-inch-thick aluminum stock that is larger than you need and degrease it with soap and water. Dry it off and spray it with marking dye. Scribe a 4-inch square on the stock and cut the template to size. A shear will make this job much easier, but it can be cut with hand tools and filed to size.

Carefully measure and scribe all holes. Note that holes A and B are on the circumference of circles. Use a compass to scribe the circles, and then locate the holes. After you have marked and checked all holes, centerpunch and drill them. The holes should be drilled with a 1/16-inch or smaller bit. Recheck all measurements. If you goof, start again. The time you spend making the template as perfect as you can will save you much time and aggravation when you make and assemble the other parts.

When you finish the template, mark the front side for future reference. All plates made from the template are marked and drilled from the front side (as viewed from the front panel).

Making the Copper Plates

Once you have completed the template, it will be easy to make the copper plates. The anode plate, grid plate and anode-bypasscapacitor plate are all made from ⅛-inch-thick copper. See Figs. 3, 4 and 5 for the dimensions of these pieces.

Measure and cut the three plates to the proper dimensions. Carefully break (deburr) all sharp edges to avoid small cuts to your fingers and hands.

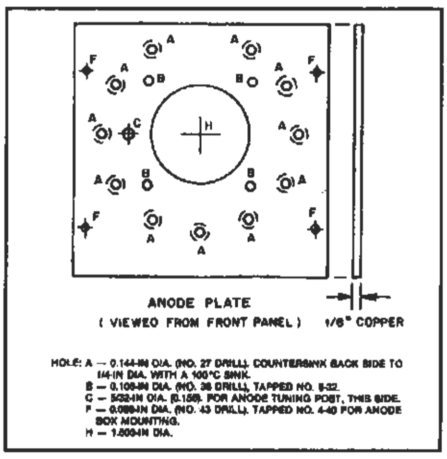
A
A
F
A
B
B
A
F
A
C
H
A
A
B
B
A
F
A
A
A
F
ANODE PLATE
(VIEWED FROM FRONT PANEL)
1/8" COPPER
HOLE: A — 0.144-IN DIA. (NO. 27 DRILL). COUNTERSINK BACK SIDE TO 1/4-IN DIA. WITH A 100°C SINK.
B — 0.106-IN DIA. (NO. 36 DRILL), TAPPED NO. 6-32.
C — 5/32-IN DIA. (0.156). FOR ANODE TUNING POST, THIS SIDE.
F — 0.089-IN DIA. (NO. 43 DRILL). TAPPED NO. 4-40 FOR ANODE BOX MOUNTING.
H — 1.500-IN DIA.

Fig 3—Drilling details for the anode plate. See Fig 2 for additional information on hole location.

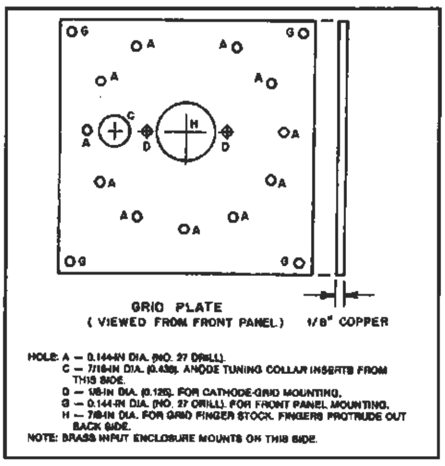
G
A
A
G
A
A
C
H
A
D
D
A
A
A
A
A
A
G
G
GRID PLATE
(VIEWED FROM FRONT PANEL)
1/8" COPPER
HOLE: A — 0.144-IN DIA. (NO. 27 DRILL).
C — 7/16-IN DIA. (0.438). ANODE TUNING COLLAR INSERTS FROM THIS SIDE.
D — 1/8-IN DIA. (0.125). FOR CATHODE-GRID MOUNTING.
G — 0.144-IN DIA. (NO. 27 DRILL). FOR FRONT PANEL MOUNTING.
H — 7/8-IN DIA. FOR GRID FINGER STOCK. FINGERS PROTRUDE OUT BACK SIDE.
NOTE: BRASS INPUT ENCLOSURE MOUNTS ON THIS SIDE.

Fig 4—Drilling details for the grid plate. See Fig 2 for additional information on hole location.

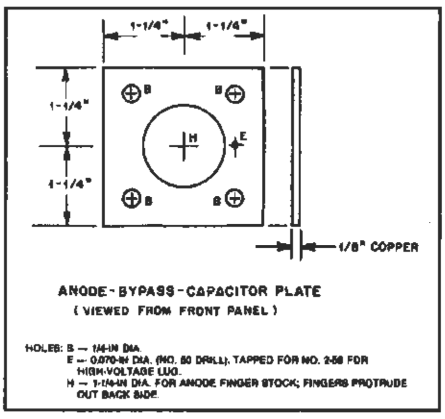
1-1/4"
1-1/4"
1-1/4"
1-1/4"
B
B
H
E
B
B
1/8" COPPER
ANODE-BYPASS-CAPACITOR PLATE
(VIEWED FROM FRONT PANEL)
HOLES: B — 1/4-IN DIA.
E — 0.070-IN DIA. (NO. 50 DRILL). TAPPED FOR NO. 2-56 FOR HIGH-VOLTAGE LUG.
H — 1-1/4-IN DIA. FOR ANODE FINGER STOCK; FINGERS PROTRUDE OUT BACK SIDE.

Fig 5—Drilling details for the anode bypass capacitor plate. See Fig 2 for additional information on hole location.

Clean the plates with alcohol and spray them with marking dye. Clamp the aluminum template to each plate, and carefully scribe the correct holes. Remember that all plates do not have the same holes. The anode plate uses holes A, B, C, F and H; the grid plate uses holes A, C, D, G and H. The anode-bypass-capacitor plate uses holes B, E and H.

Use a small center punch to punch all holes lightly. If they then look accurate, enlarge them enough for drilling.

Copper isn't the easiest metal to work with. It's very stringy, and drilling it can be frustrating. You'll need the proper drill bits for best results. Special drills can be purchased, or you can use a grinder to carefully remove the sharp points on the outer edge of the cutting surface of each side of a standard drill bit. This will eliminate any tendency for the copper to grab. Practice on an old bit and be sure to grind it symmetrically. Modified drill bits can still be used on aluminum and other metals.

Always start with a smaller drill and work up to the final hole size. It's safer and more accurate. The larger holes can be cut with a flycutter, or you can drill a series of smaller holes around the inside of a larger hole and file to finish. Either way is fine. Use lots of cutting fluid to lubricate the drill bit, and wear safety glasses and an old shirt. Remember, some cutting fluids are not to be used on aluminum.

Start with a no.50 (0.070-inch) or smaller bit and drill pilot holes at each of your punched marks. The details for finishing each hole are listed in the drawings. Some holes are countersunk or tapped. Pay attention to the details, and take your time.

When you are through drilling, you must deburr each hole. Copper is soft, so it tends to rise up around the hole during drilling and deburring. Use a flat file for the initial cut, and then remove any remaining material with a countersink. File the copper plates flat again; a flush fit on both sides of the aluminum ring is important.

When all copper work is done, you should be able to stack the plates and see all pertinent holes align correctly. Enough tolerance is included in the dimensions to accommodate minor errors. After the holes are drilled, it can be difficult to tell which side of each plate is which, so mark the front side of each plate with a permanent marker.

Machining the Ring

The aluminum ring that forms the cavity wall is cut (sliced) from a length of 3½-inch OD tubing with a ⅜-inch wall thickness. See Fig. 6. The tubing ID is about 2¾ inches. The dimensions of the ring are the most critical in this amplifier. Tolerance of the ring thickness is ± 0.005 inch to maintain full band coverage.

The ring can be hacksawed or bandsawed out of the tubing, but take extreme care to be accurate. Cutting tubing straight isn't easy. Clamp the tubing to prevent rotating on the band saw. The final finish cut is best done on a lathe or milling machine, but careful filing will work.

Once the ring is the correct thickness, deburr the sharp edges and spray it with marking dye. Notice that the outside and inside diameters are not concentric. This is normal for large tubing. Lay the ring flat and find the thickest wall section. Scribe a

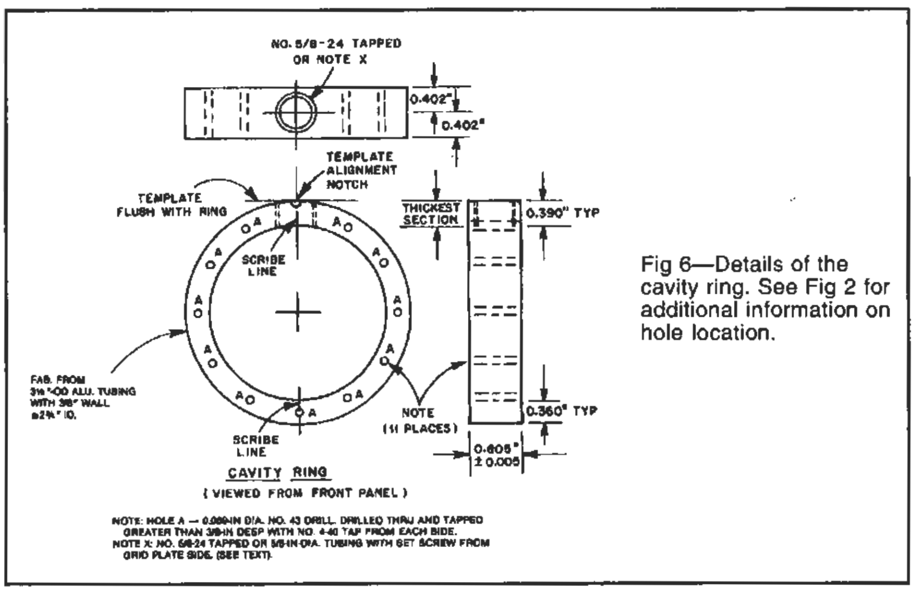
NO. 5/8-24 TAPPED
OR NOTE X
0.402"
0.402"
TEMPLATE
ALIGNMENT
NOTCH
TEMPLATE
FLUSH WITH RING
THICKEST
SECTION
0.390" TYP
SCRIBE
LINE
A
FAB. FROM
3¼"-OD ALU. TUBING
WITH 3/8" WALL
≈2½" ID.
NOTE
(11 PLACES)
0.360" TYP
SCRIBE
LINE
0.605"
± 0.005
CAVITY RING
(VIEWED FROM FRONT PANEL)
NOTE: HOLE A — 0.089-IN DIA. NO. 43 DRILL. DRILLED THRU AND TAPPED
GREATER THAN 3/8-IN DEEP WITH NO. 4-40 TAP FROM EACH SIDE.
NOTE X: NO. 5/8-24 TAPPED OR 5/8-IN-DIA. TUBING WITH SET SCREW FROM
GRID PLATE SIDE. (SEE TEXT).

Fig 6—Details of the cavity ring. See Fig 2 for additional information on hole location.

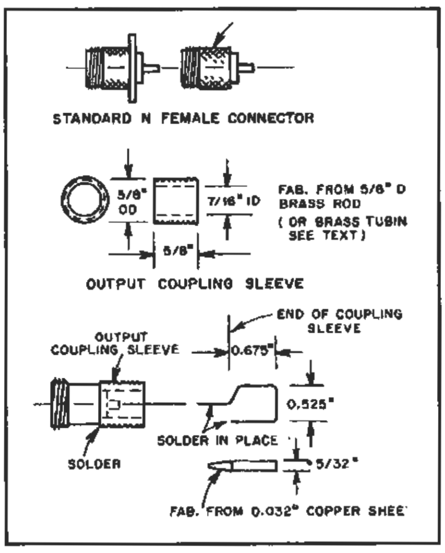
STANDARD N FEMALE CONNECTOR
5/8"
OD
7/16" ID
FAB. FROM 5/8" D
BRASS ROD
(OR BRASS TUBIN
SEE TEXT)
5/8"
OUTPUT COUPLING SLEEVE
END OF COUPLING
SLEEVE
OUTPUT
COUPLING SLEEVE
0.675"
0.525"
SOLDER IN PLACE
SOLDER
5/32"
FAB. FROM 0.032" COPPER SHEE

Fig 7—Output-probe/connector assembly details.

line across the wall at this point, across the center of the ring and across the wall on the other side. The scribed lines on each side of the ring will be used to align the template. The output connector will be placed at the thick wall section.

Carefully align notch I on the template with the line scribed on the thickest wall section on the ring. Clamp the template onto the ring. Mark each of the 11 holes labeled A on the template. After you mark the holes and remove the template, check alignment with the copper plates just in case. If everything lines up, center punch all eleven holes on one side of the ring only, and drill each hole completely through the ring. Use lots of cutting fluid. File the ring flat before and after deburring, taking care not to change the wall thickness. Tap each hole to accept no. 4-40 machine screws. Each hole will have to be tapped to a depth of at least $^3/_8$ inch from both sides because long taps don't exist. The inside of the ring doesn't need to be polished.

The hole for mounting the output connector can now be drilled. There are two ways to mount this connector, and either scheme works fme. Read ahead to the section on making the output connector for more information. The first method of mounting the connector involves tapping the ring with a no. $^5/_8$-24 tap and using a lathe to cut matching threads on the output connector coupling sleeve. Large taps are expensive, but a tap and die for Type-N connectors are handy if you do much building.

If you don't have access to a lathe or a large tap, the second method is easier. Make the output connector coupling sleeve from $^5/_8$-inch-OD brass or copper tubing, and drill the ring to just clear it. Then drill and tap the grid-plate side of the ring above the output connector to accept a setscrew. Also, drill a clearance hole in the grid plate for the setscrew. Use the setscrew to secure the output connector.

Output Connector

A standard Type-N chassis-mount female connector (silver plated) is used for the output probe/connector. See Fig. 7. First, remove the flange with a hacksaw and file flush with the connector body. Next, make the output-coupling sleeve that is right for your application (threaded orunthreaded, depending on how you fabricated the ring). The sleeve will be the same length in either case. The output-coupling loop is fashioned from a piece of 0.032-inch-thick copper sheet that is $^5/_{32}$ inches wide. Bend it to the dimensions shown in Fig. 7. We will solder the output connector together later.

Grid Compartment

The grid compartment measures 2 inches square by $1^1/_2$ inches high. See Fig. 8. It is made from brass and can be sawed out of square tubing or bent from sheet. The cover can be made from any material.

I use two small PC boards (Fig. 9) for

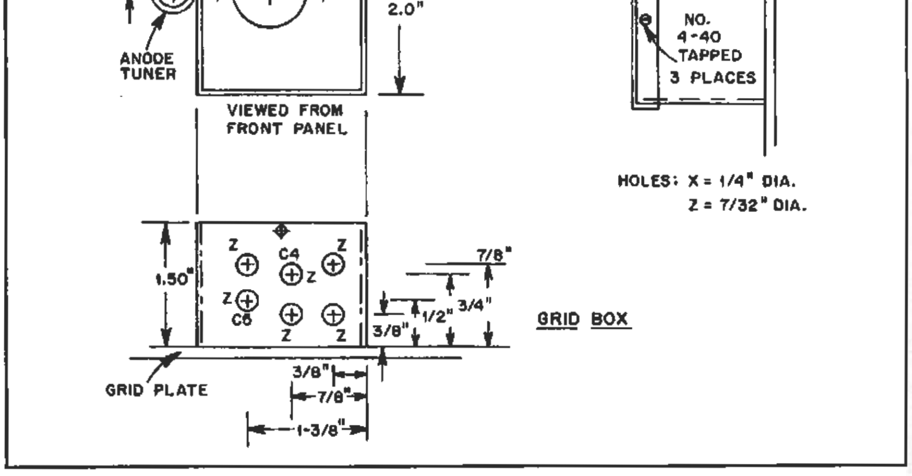
0.475"
1.05"
Z
C2
Z
X
C3
X
Z
GRID PLATE
INPUT
CONNECTOR
2.0"
0.800"
C
H
C
2.0"
ANODE
TUNER
VIEWED FROM
FRONT PANEL
COVER, FAB. FROM
0.020" ALU. OR BRASS
NO.
4-40
TAPPED
3 PLACES
HOLES: X = 1/4" DIA.
Z = 7/32" DIA.
1.50"
Z
C4
Z
Z
C5
Z
Z
7/8"
3/4"
1/2"
3/8"
GRID BOX
GRID PLATE
3/8"
7/8"
1-3/8"

Fig 8—Input-compartment details.

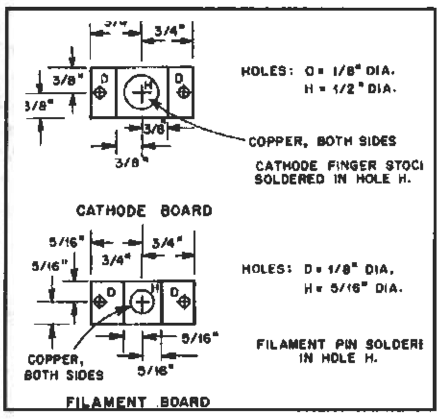
3/4"
3/8"
3/8"
3/8
3/8"
D
H
HOLES: D = 1/8" DIA.
H = 1/2" DIA.
COPPER, BOTH SIDES
CATHODE FINGER STOCK SOLDERED IN HOLE H.
CATHODE BOARD
5/16"
3/4"
3/4"
5/16"
5/16"
5/16"
HOLES: D = 1/8" DIA.
H = 5/16" DIA.
COPPER, BOTH SIDES
FILAMENT PIN SOLDERED IN HOLE H.
FILAMENT BOARD

Fig 9—Cathode and filament PC-board details.

holding the finger stock that makes contact with the filament pin and cathode ring on the 2C39 tube. These boards are cut from 1/16-inch-thick, double-sided G-10 glass-epoxy stock. The copper pattern is identical for both sides of each piece. Mark and drill or file the holes first, and then cut the boards to size. Small boards are difficult to hold while drilling them. Mark each side of each board and score the copper foil with a sharp knife.

The unwanted copper can be removed easily by heating the foil with a soldering iron and lifting it off. Use a flat file to deburr the boards. Do not use a countersink because the copper foil must be as close to the holes as possible to facilitate soldering the finger stock in place.

The input connector that I use is a 5-mm SMA type. This is an excellent RF connector, especially for low-power UHF applications. I highly recommend use of an SMA, but any small screw-on connector will do. If you really feel you have to use a BNC then do so, but it's a lousy connector at frequencies above 200 MHz. Remember to move the connector hole to accommodate its larger size.

The input connector must be as close as possible to the first input capacitor. The lead length of the input dc blocking capacitor must be as short as possible. The 3-pF capacitor is series resonant at 1200 MHz only with short (1/16-inch or less) leads.

Miscellaneous Bits and Pieces

There are still several small, but very important parts to fabricate. The front panel I use is shown in Fig. 10. It is made from a piece of 1/8-inch-thick aluminum sheet. Some builders may wish to mount the amplifier on a rack panel. Wash and dry your front-panel material and spray it with marking dye. Clamp it to the template and mark the holes. Check the hole alignment with the copper grid and anode plates. If all lines up correctly, center punch and drill the holes. The only front-panel control is for the anode tuning capacitor, which is adjusted by a 1/4-inch shaft protruding through a 3/8-inch panel bushing in hole C.

The anode tuning collar, shown in Fig. 11A, is made from a piece of 1/2-inch OD brass rod. This rod has a 3/8-inch hole drilled through its center, and it is turned down to 7/16-inch OD for half its length. The inside of the 1/2-inch-OD end is tapped to a depth of 1/4 inch to accept 3/8-24 threads. This collar will be inserted into hole C on the grid plate.

Fig. 11B also shows the anode tuning post. It is simply a length of 5/32-inch-OD brass rod that inserts into hole C on the copper anode plate. This rod will form one plate of the anode tuning capacitor.

The anode tuner (Fig. 11C) is machined from a piece of 3/8-inch-OD Teflon rod. One end of the rod is drilled out with a no. 21 drill. The outer wall of this end is threaded with a no. 3/8-24 tap. This is the end that will thread into the anode tuning collar and slip over the anode tuning post. The other end is turned down to fit inside a 1/4-inch shaft coupler.

Fig. 12 shows the remaining parts. The tuning shaft (A) is made from a piece of 1/4-inch brass rod. A coupler (B) to connect the tuning shaft to the anode tuner may be purchased or made. This also applies to the front-panel spacers (C). The Teflon dielectric for the anode bypass capacitor (D) is made from 0.010-inch-thick Teflon sheet. Use the template to locate holes B and H. Teflon washers and inserts (F) are used to insulate the mounting hardware for the anode bypass capacitor from the chassis. The inserts are made from 1/4-inch-OD Teflon rod. The washers are made from Teflon sheet. Sharpen a piece of 3/8-inch aluminum tubing and chuck it up in a drill press. This tool will cut neat, round washers from the sheet.

The box that encloses the anode compartment (Fig. 13) is fabricated from a Bud AU-1083 utility cabinet. Clean the chassis and spray it with marking dye. Secure the template to the side of the enclosure that contacts the anode plate and scribe the holes labeled F. Make sure that these holes line up with the holes on the copper anode plate. If they do, center punch and drill them to size. If air cooling is used, the blower will mount to this box.

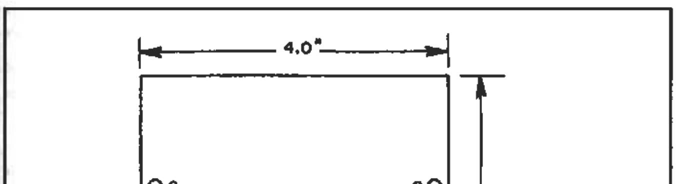
4.0"
5¼"
G
G
G
G
C
FRONT PANEL
(VIEWED FROM FRONT)
HOLES: C - 3/8" DIA. PANEL BUSHING WITH EXCESS THREADS REMOVED. H.H. SMITH NO.119 OR EQUIV.
G - 0.144" DIA. (NO. 27 DRILL)
FAB. FROM ALU., RACK PANEL OR SHEET.

Fig 10—Front-panel details.

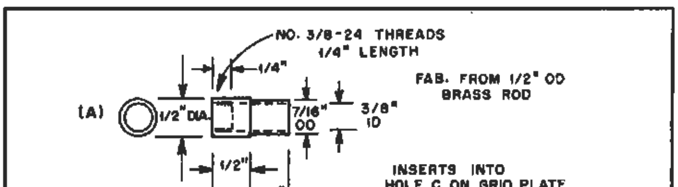
(A)
NO. 3/8-24 THREADS
1/4" LENGTH
1/4"
1/2" DIA.
7/16" OD
3/8" ID
1/2"
1/2"
FAB. FROM 1/2" OD BRASS ROD
INSERTS INTO HOLE C ON GRID PLATE
ANODE-TUNING COLLAR
(B)
1.325"
5/32" OD
FAB. FROM 5/32" OD BRASS TUBING
INSERTS INTO HOLE C ON ANODE PLATE
ANODE-TUNING POST
(C)
2.725"
COUNTER SINK
1-1/4"
1/16" HOLE TO AID IN CLEANING THREADS
NO. 3/8-24 THREADS
0.159" DIA. 1-1/4" DEEP (NO.21 DRILL)
3/8"
3/8" OD
1/4" DIA.
FAB. FROM 3/8" OD TEFLON ROD
ANODE- TUNER

Fig 11—Anode-tuning capacitor details.

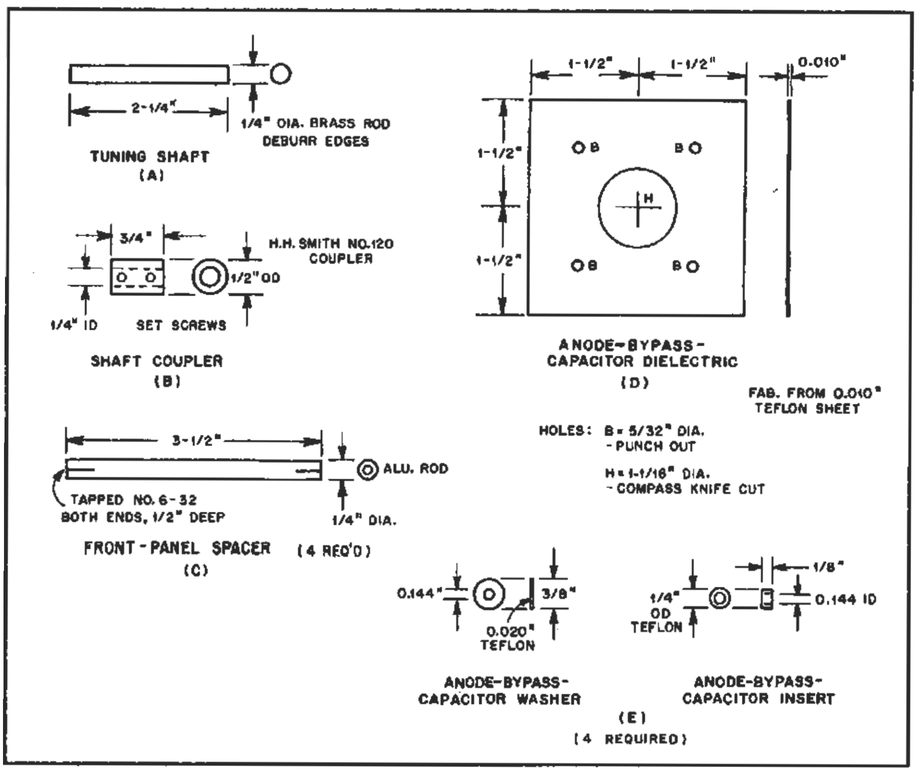
2-1/4"
1/4" DIA. BRASS ROD
DEBURR EDGES
TUNING SHAFT
(A)
3/4"
H.H. SMITH NO.120
COUPLER
1/2" OD
1/4" ID
SET SCREWS
SHAFT COUPLER
(B)
3-1/2"
ALU. ROD
TAPPED NO. 6-32
BOTH ENDS, 1/2" DEEP
1/4" DIA.
FRONT-PANEL SPACER
(4 REQ'D)
(C)
1-1/2"
1-1/2"
0.010"
1-1/2"
1-1/2"
B
B
H
B
B
ANODE-BYPASS-
CAPACITOR DIELECTRIC
(D)
FAB. FROM 0.010"
TEFLON SHEET
HOLES: B = 5/32" DIA.
- PUNCH OUT
H = 1-1/16" DIA.
- COMPASS KNIFE CUT
0.144"
3/8"
0.020"
TEFLON
1/4"
OD
TEFLON
1/8"
0.144 ID
ANODE-BYPASS-
CAPACITOR WASHER
ANODE-BYPASS-
CAPACITOR INSERT
(E)
(4 REQUIRED)

Fig 12—Miscellaneous parts necessary to complete the amplifier.

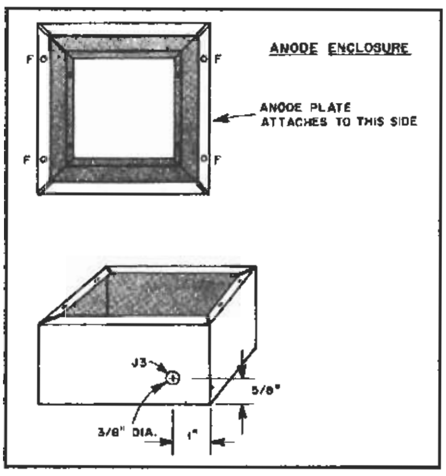
ANODE ENCLOSURE
F
F
F
F
ANODE PLATE
ATTACHES TO THIS SIDE
J3
5/8"
3/8" DIA.
1"

Fig 13—Anode-enclosure details.

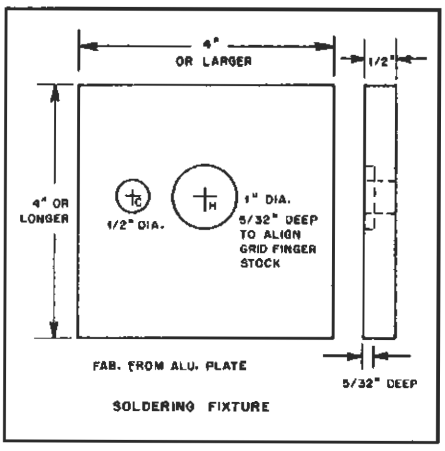
4"
OR LARGER
1/2"
4" OR
LONGER
C
1/2" DIA.
H
1" DIA.
5/32" DEEP
TO ALIGN
GRID FINGER
STOCK
FAB. FROM ALU. PLATE
5/32" DEEP
SOLDERING FIXTURE

Fig 14—Dimensions of the soldering fixture. See Fig 2 for more information on hole location.

Soldering the Subassemblies

Once all copper and brass parts are drilled and deburred, they should be cleaned with alcohol and Scotch-Brite®, a nonmetallic pot cleaner, and washed in alcohol again. Set the pieces aside and avoid touching them. Fingerprints will inhibit soldering.

I have found that the best way to solder the heavy brass and copper parts is to first build the soldering fixture shown in Fig. 14. This soldering fixture, made from 1/2-inch-thick aluminum plate, will evenly heat the entire assembly to be soldered. Even heating will allow you to do a much better soldering job than you could otherwise.

The soldering fixture should be preheated on a stove or hot plate until bits of solder placed on its surface just melt. At this point, reduce the heat slightly. Avoid excessive heat. If the copper parts placed on the fixture suddenly turn dark, it's too hot.

Solder the grid plate assembly first. You will need the copper grid plate, grid finger stock, anode tuning collar and brass input compartment.[2] Look at the drawings again to be sure that you know which parts go where. Insert the grid finger stock into hole H on the grid plate. As viewed from the front-panel side, the curved fingers will protrude out the back side, away from you. Apply liquid or paste flux and set the grid plate in the soldering fixture. The finger stock will fit in hole H in the fixture, allowing the grid plate to rest flush with the surface of the fixture. Next, apply flux to the anode tuning collar and insert it in hole C of the grid plate. Part of the tuning collar will slip into hole C in the soldering fixture. Make sure the collar seats flush with the grid plate. The flux should start to bubble.

Carefully apply solder directly to the joints of the installed parts. The solder should melt almost immediately and flow bright and smooth. Next, place the square brass input compartment in place and apply flux. In a few seconds, it can be soldered by running solder around the joints, inside and outside. If you have trouble getting it to flow on both sides, merely tap the brass box aside (1/16 inch) and return it to its original position.

Now comes the hard part — getting the soldered assembly away from the heat without disturbing the alignment. A pair of forceps is recommended, but long pliers will do. Carefully lift the assembly off the soldering fixture and set on a cooling rack. Do this without moving any part. The cooling rack can be any two pieces of metal that will allow clearance for the protruding parts. You can expedite cooling by using an ordinary hair dryer in the "cool" position to gently blow air across the assembly.

While the grid assembly is cooling, assemble the output connector. See Fig. 7. Place the modified Type-N female connector, threaded end down, on the soldering fixture. Apply flux to the top and install the output coupling sleeve. Allow both parts to heat before applying solder. Carefully remove the soldered output connector from the fixture. When it has cooled, solder one end of the loop to the center pin of the N connector and the other to the output coupling sleeve.

Now place the anode plate on the soldering fixture and allow to heat. Apply flux to hole C. Insert the anode tuning post (5/32-inch-OD brass tube) and allow to heat; apply solder. Remove the parts and cool. Next, solder the finger stock in hole H on the anode bypass capacitor plate.[3]

This completes the work with the soldering fixture. Be sure to let it cool off before handling! Save the fixture for future construction; you never know when you might want it again.

The anode plate and the anode-bypass-capacitor plate must be filed and then sanded flat on their butt surfaces to assure that there are no solder bumps or sharp points to puncture the Teflon dielectric. This must be done after soldering. The Teflon sheet is adequate insulation for many times the anode potential of this amplifier, but only if the surfaces it separates are smooth!

Next, clean the cathode and filament PC boards. Install the finger stock in hole H of the cathode board. Apply flux to both sides of the board. Heat with a hot iron and apply solder around the circumference of hole H, soldering the finger stock on both sides of the board. Use the same technique to install the filament pin.[4]

After all parts have cooled, use a spray can of flux remover to clean them. Slight scrubbing with Scotch-Brite pot cleaner will finish them nicely. Congratulations: You have finished the pieces and are now

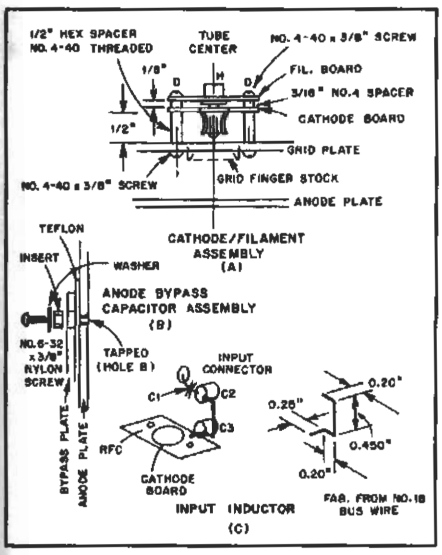
1/2" HEX SPACER NO. 4-40 THREADED
TUBE CENTER
NO. 4-40 x 3/8" SCREW
1/8"
D
H
D
FIL. BOARD
3/16" NO.4 SPACER
CATHODE BOARD
1/2"
GRID PLATE
NO. 4-40 x 3/8" SCREW
GRID FINGER STOCK
ANODE PLATE
TEFLON
INSERT
WASHER
CATHODE/FILAMENT ASSEMBLY
(A)
ANODE BYPASS CAPACITOR ASSEMBLY
(B)
NO.6-32 x 3/8" NYLON SCREW
TAPPED (HOLE B)
INPUT CONNECTOR
C1
C2
C3
RFC
BYPASS PLATE
ANODE PLATE
CATHODE BOARD
INPUT INDUCTOR
(C)
0.20"
0.25"
0.450"
0.20"
FAB. FROM NO.18 BUS WIRE

Fig 15—Assembly details for the filament and cathode boards (A), the anode-bypass capacitor (B) and the input pi network (C).

ready to bolt the amplifier together.

Silver Plating

Over the years, many people have pushed silver plating as the only way to go. You may wish to silver plate the amplifier components before soldering them together, but I do not think it's necessary. I ran several tests to prove how much various types of plating affect performance of this amplifier. Remember that the RF skin conductivity of aluminum and copper is pretty good at 23 cm; they are much better than brass.

Four amplifiers were built for this test. They were plated as follows:

1) Nickel plated
2) Tin plated
3) Silver plated
4) Unplated

There was no difference in performance among the tin-plated, silver-plated and unplated versions. The nickel-plated amplifier exhibited 3-dB less gain.

In other words, it is not necessary to silver plate this amplifier; however, it does improve appearance by making the parts a similar color. Silver does tarnish, especially with fingerprints. The decision to plate or not to plate is up to you.

Assembly

After fabrication of all parts, assembly is simple. Figs. 15 through 17 show assembly details. Loosely fasten the grid and anode plates to the ring. Mount the input connector and capacitors on the input compartment. Loosely install the cathode and filament boards and their respective spacers. See Fig. 15A.

Now insert a 7289/2C39 tube. This will center up all finger stock. Place the Teflon anode tuner in its collar on the grid plate and screw it most of the way in. Now tighten all of the screws. The 7289/2C39 tube should slide in and out snugly, and the anode tuner should screw in and out smoothly.

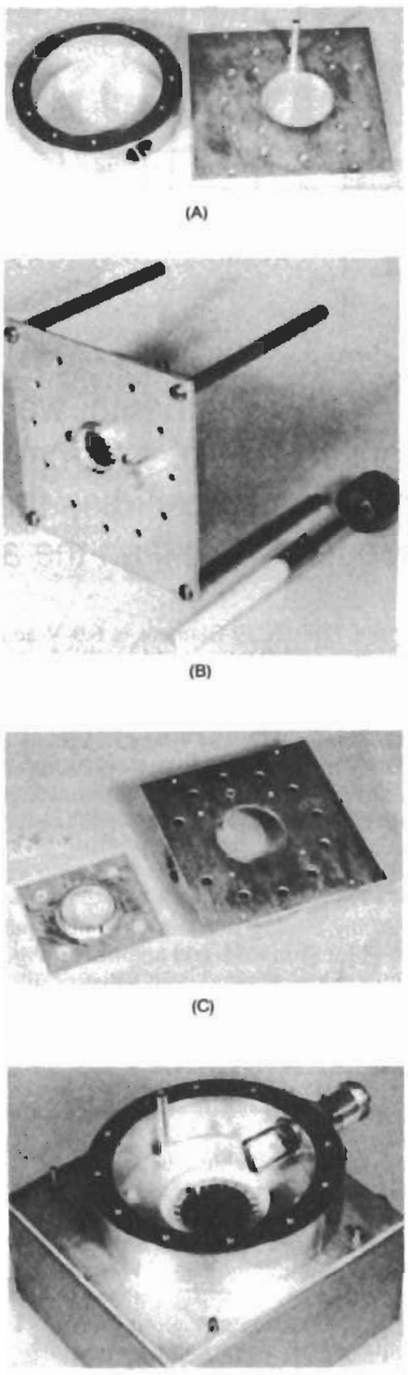

Fig. 16—The completed cavity ring and anode plate with anode tuning post soldered In place are shown at A. The photo at B shows the grid plate with finger stock, input compartment and anode tuning collar soldered in place. The completed anode tuner Is at the right. C shows the cavity ring attached to the anode plate. The anode-bypass capacitor is ready for installation. At D, the interior of the cavity as seen from the grid plate side is visible. The output probe/connector assembly is installed. The anode bypass capacitor and anode enclosure have been installed on the anode plate.

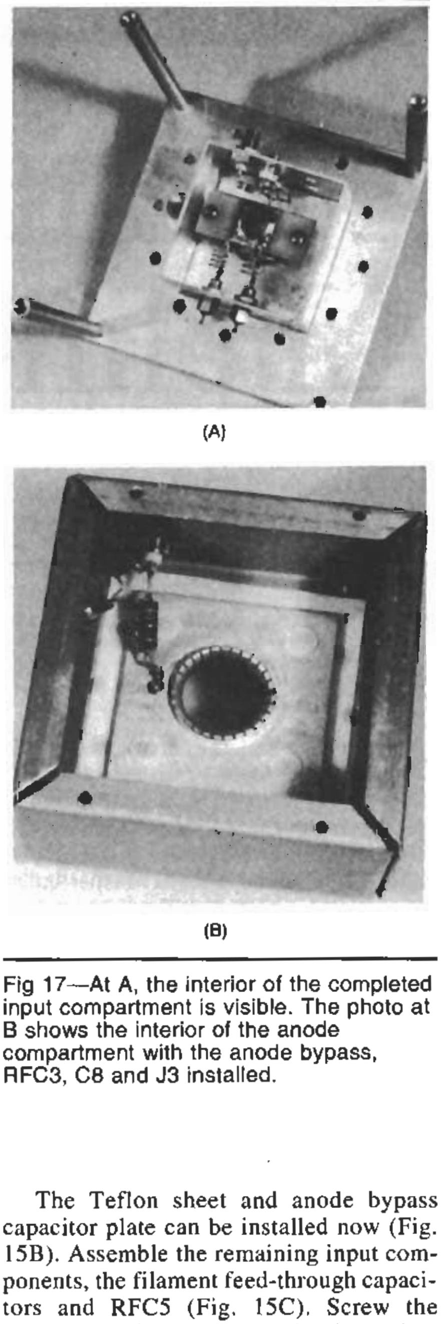

Fig 17—At A, the interior of the completed input compartment is visible. The photo at B shows the interior of the anode compartment with the anode bypass, RFC3, C8 and J3 installed.

The Teflon sheet and anode bypass capacitor plate can be installed now (Fig. 15B). Assemble the remaining input components, the filament feed-through capacitors and RFC5 (Fig. 15C). Screw the output probe into the cavity ring (or push in the probe and tighten the setscrew, depending on which method you chose). Install the high-voltage connector and other parts in the anode box. Mount the amplifier on the front panel and install the anode tuner shaft. This completes the assembly. Part 2 of this article will describe a complete power supply for the amplifier, a practical water cooling system, testing procedures, microwave radiation safety hazards, and amplifier tune-up and operation.

Notes

[1] mm = in x 25.4.

[2] The finger stock for this project is manufactured by Instrument Specialities, P.O. Box A, Delaware Water Gap, PA 18237. Contact them for the name of the closest distributor. The part numbers for this amplifier are: anode bypass capacitor plate, no. 97-70A; grid plate, no. 97-74A; cathode board, no. 97-420A; filament board, no. 97-280A.

[3] See note 2.

[4] See note 2.

By E.R. "Chip" Angle, N6CA

From *QST*, April 1985

A Quarter-Kilowatt 23-cm Amplifier

Part 2—Last month, we described the design and construction of a 23-cm cavity amplifier. This installment describes the rest of the components needed to put it on the air.

After you complete Construction of the cavity amplifier described in March *QST*, you are ready to assemble the rest of the components needed to put it on the air. This month, I will discuss the filament, bias and high-voltage supplies; a whisper-quiet, high-efficiency water-cooling system; testing and hookup; and, finally, tune-up and operation.

Power Supplies

The filament and bias supplies for the cavity amplifier are shown schematically in Fig. 1. The manufacturer's specification for the 7289/2C39 filament is 6.0-V ac at 1 A. I have found that the use of a standard 6.3-V ac, 1-A transformer only slightly increases the tube emission without much loss of tube life. The filament should be allowed to warm up before operating the amplifier, so the filament, bias and high-voltage supplies incorporate separate primary switches.

Biasing

Many biasing schemes have been published for grounded-grid amplifiers. Fig. 1 shows a bias network that satisfies all of the following operating requirements:

1) external bias supply referenced to ground

2) low-power components

3) variable bias to accommodate tube-to-tube variations

4) TR switchable with relay contact or transistor to ground

5) bias-supply protection in case of a defective or shorted tube.

U2 provides a variable bias-voltage source, adjustable by R1. The output of U2 drives the base of Q1, which is used to increase the current-handling capability of

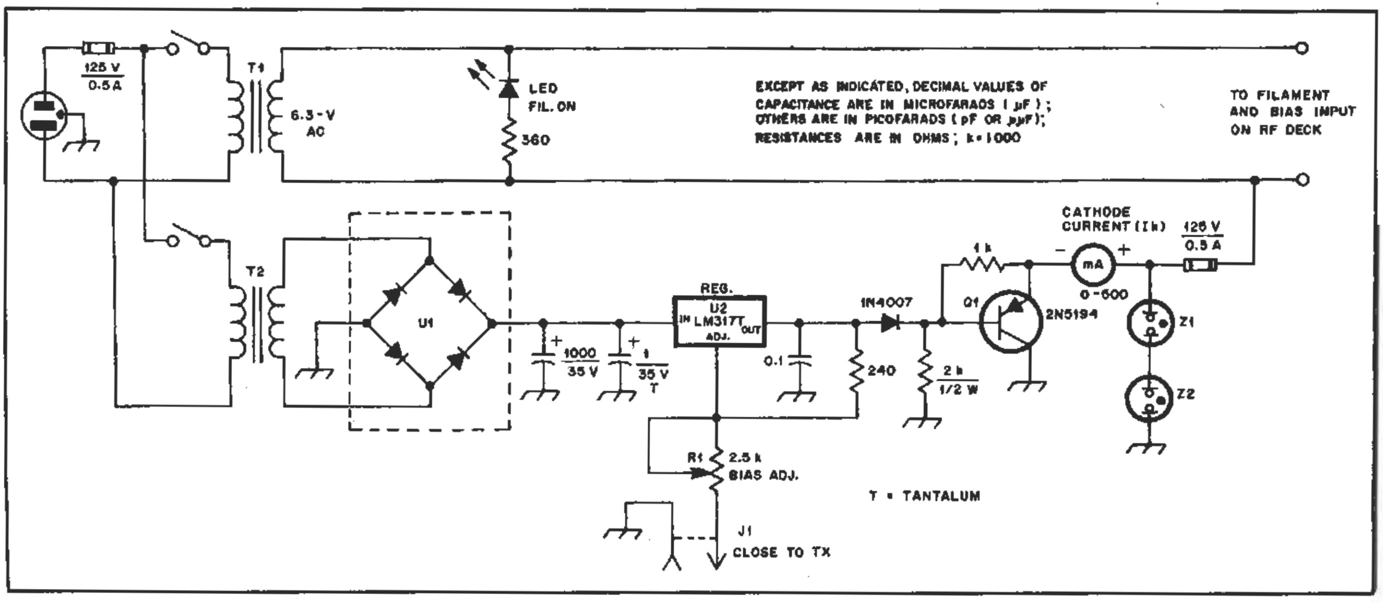
125 V
0.5 A
T1
6.3-V AC
LED FIL. ON
360
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000
TO FILAMENT AND BIAS INPUT ON RF DECK
T2
U1
CATHODE CURRENT (Ik)
125 V
0.5 A
1 k
mA
0-600
REG.
U2
IN LM317T OUT
ADJ.
1N4007
Q1
2N5194
Z1
Z2
1000
35 V
1
35 V
T
0.1
240
2 k
1/2 W
R1 2.5 k
BIAS ADJ.
T = TANTALUM
J1
CLOSE TO TX

Fig 1—Schematic diagram of the cavity-amplifier filament and bias supplies. All resistors are ¼-W carbon types unless otherwise noted.

J1—Female chassis-mount photo connector.
T1—Filament transformer. primary, 117 V; secondary, 6.3 V at 1 A.
T2—Power transformer. Primary, 117 V; secondary, 24 to 28 V at 50 mA or greater.
U1—Bridge rectifier, 50 PIV, 1 A.
U2—Adjustable 3-terminal regulator (LM317T or equiv.).
Z1, Z2—20-V unipolar metal-oxide varistor (General Semiconductor SA20 or equiv.) or two 20-V, 1-W Zener diodes.

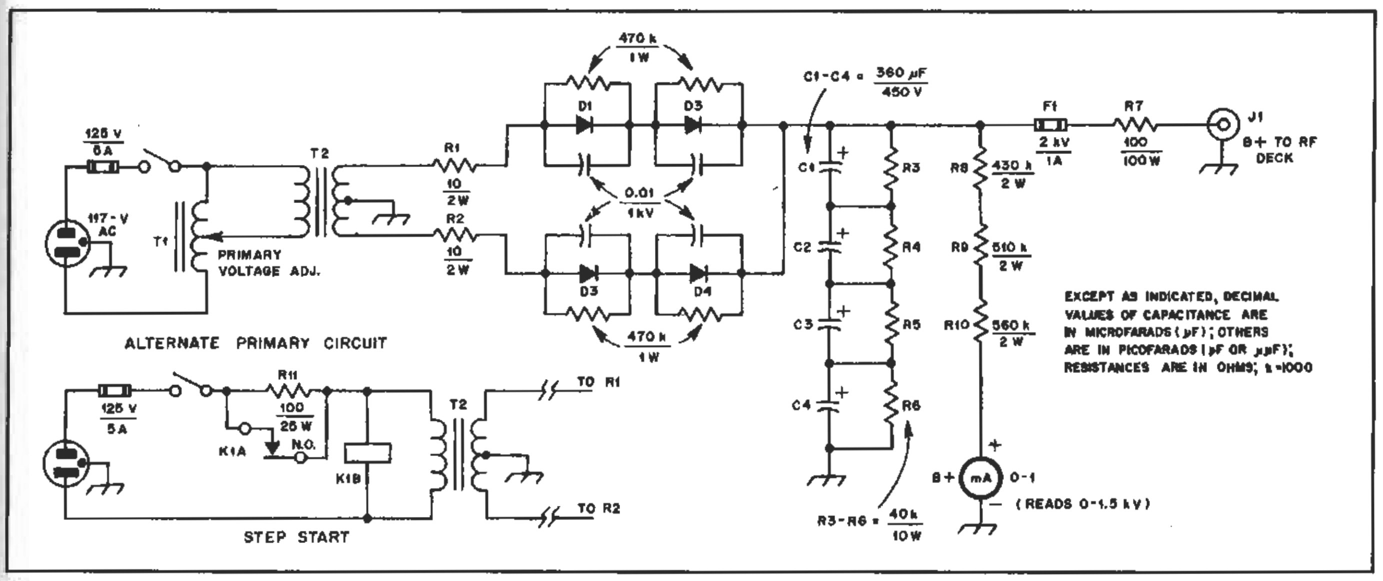

Fig 2—Schematic diagram of the amplifier high-voltage supply.

C1-C4—Electrolytic capacitor, 360 μF, 450 V.
D1-D4—Silicon rectifier, 1000 PIV, 3 A.
F1—High-voltage fuse, 2 kV, 1 A.
J1—Chassis-mount female BNC or MHV connector.
R3-R6—Wirewound resistor, 40 kΩ, 11 W.
T1—Variable autotransformer, 500 VA.
T2—High-voltage transformer. Primary, 117 V; secondary, 900 to 1050 V at 500 mA.

the bias supply. Q1 must be mounted on a heat sink. J1 is connected to the station TR switching system so that R1 is grounded on transmit and disconnected on receive. The approximate range of the bias supply is 6 to 20 V. Z1 and Z2 provide protection for Q1 in case of a shorted tube. The amplifier can be run without Z1 and Z2 if you keep the anode voltage below 1100 V.

High-Voltage Power Supply

A safe, reliable high-voltage power supply is described here. Of course, you can use any readily available HV supply; keep in mind, however, that the 7289/2C39 anode potential should never exceed 1400-V dc at full load and that the amplifier will withstand 1900-V dc at low cathode current and cut-off-bias conditions. For maximum power output, assuming adequate drive power is available, anode voltage under full load should be about 1200- to 1400-V dc.

Fig. 2 is a schematic diagram of the high-voltage supply. A power transformer (T2) that delivers 900- to 1050-V ac is ideal. The type of rectifier circuit used will depend on the type of transformer chosen. Each leg of the rectifier is made from two 1000-PIV, 3-A silicon diodes connected in series. Each diode is shunted with a 0.01-μF capacitor to suppress transient voltage spikes, and a 470-kΩ equalizing resistor.

Filtering is accomplished with a string of four 360-μF, 450-V electrolytic capacitors connected in series. R3-R6 equalize the voltage across each capacitor in the string and serve as bleeder resistors. Of course, a single oil-filled capacitor may be used here if available. Whatever type of filter you use, the total capacitance should be about 80 μF at a voltage rating of at least 1500-V dc. This value allows adequate "droop" of the anode voltage under high-current loads to protect the amplifier in case of RF overdrive or a defective tube.

Protective Circuitry

Some type of start-up protection should be incorporated in the primary. Fully discharged filter capacitors look like a dead short at supply turn-on. Initial surge current (until the capacitors charge) may be high enough to destroy the rectifiers. R1 and R2 provide some surge-current limiting, but either of the two primary configurations shown in Fig. 2 should be used. T1, a variable autotransformer (Variac and Powerstat are two common trade names), is ideal. In addition to allowing you to bring the primary up slowly (and charging the capacitors gradually), it also allows full control of amplifier output power by varying anode voltage.

The second method, a "step-start" system, uses a resistor in the T2 primary to limit the turn-on surge current. When the capacitors have charged, K1 is energized, shorting out R11 and applying full voltage to the T2 primary.

F1 and R7 protect against high-voltage arc-overs or short circuits. If sustained overcurrent is drawn, F1 will open and remove B+ from the RF deck. Use a high-voltage fuse here; standard fuses may arc when blown and not interrupt the B+ .R7 provides current limiting to protect the amplifier and power supply in case of a high-voltage arc.

Safety

An HV meter should always be used to monitor the status of the power supply. The values for R8-R10 shown in Fig. 2 will give a 1500-V dc full-scale reading on a 0-1 mA meter. RG-58 or -59 coaxial cable should be used for the high-voltage interconnection between the power supply and the RF deck. Ground the shield at both ends for safety and a good dc return.

Safety must be observed when working with all power supplies. These voltages are lethal! Always disconnect ac power and then discharge the filter capacitors before working on the power supply. Never guess or make assumptions about the status of a power supply. Assume it is hot.

Metering

Cathode-current monitoring is all that's really necessary for observing amplifier dc performance. Cathode current (I_K) is the sum of the plate (I_P) and grid (I_G) currents. Normally, when this amplifier is driven to 300- or 400-mA I_K, the grid current will be around 40 to 50 mA. The inclusion of a grid-current meter is not really necessary and only makes biasing and TR switching complicated.

Cooling

Desired output power and the level of drive power available will dictate what type of cooling to use. For intermittent duty (SSB, CW) at output levels less than 50 W, air cooling is satisfactory. Any small blower may be easily mounted to the aluminum box surrounding the tube anode. For high-duty-cycle modes and/or output levels greater than 50 W, water cooling is highly recommended. Greater than twice the normal air-cooled output power can be obtained from a water-cooled tube, and water cooling is quiet.

Tube Modification and Water Jacket

The first step is to remove the air radiator from the tube. The air radiator screws on, so it may simply be unscrewed without damage to the tube.

First, place a hose clamp around the tube anode. Secure the radiator fms in a vise and grip the hose clamp with a pair of large

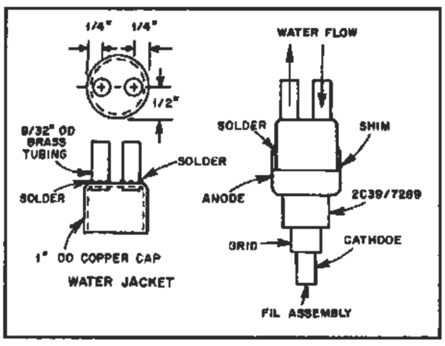
1/4"
1/4"
WATER FLOW
1/2"
9/32" OD BRASS TUBING
SOLDER
SOLDER
SOLDER
SHIM
SOLDER
ANODE
2C39/7289
GRID
CATHODE
1" OD COPPER CAP
WATER JACKET
FIL ASSEMBLY

Fig 3—Details of the solder-on water jacket.

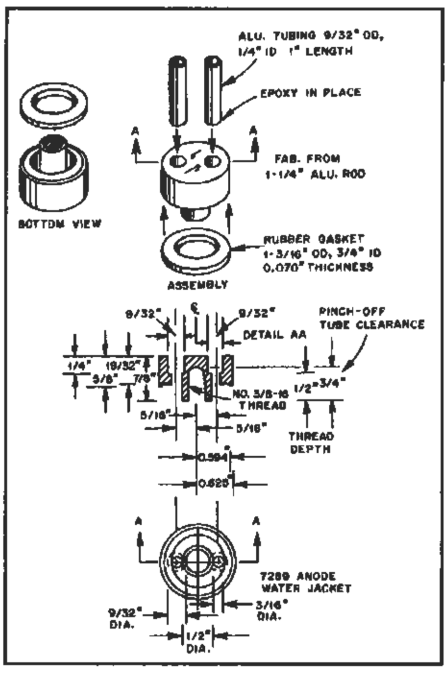
ALU. TUBING 9/32" OD, 1/4" ID 1" LENGTH
EPOXY IN PLACE
A
A
FAB. FROM 1-1/4" ALU. ROD
BOTTOM VIEW
RUBBER GASKET 1-3/16" OD, 3/4" ID 0.070" THICKNESS
ASSEMBLY
9/32"
℄
9/32"
DETAIL AA
PINCH-OFF TUBE CLEARANCE
1/4"
19/32"
3/8"
7/8"
NO. 3/8-16 THREAD
1/2"
3/4"
5/16"
5/16"
THREAD DEPTH
0.594"
0.625"
A
A
7289 ANODE WATER JACKET
9/32" DIA.
3/16" DIA.
1/2" DIA.

Fig 4—Details of the screw-on water jacket.

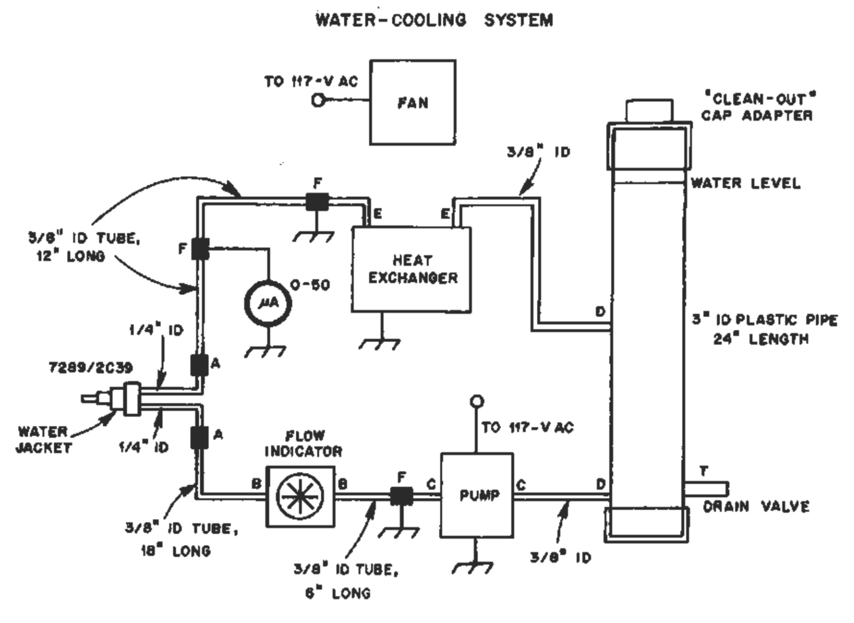
WATER-COOLING SYSTEM
TO 117-V AC
FAN
"CLEAN-OUT" CAP ADAPTER
3/8" ID
WATER LEVEL
F
E
E
3/8" ID TUBE, 12" LONG
F
HEAT EXCHANGER
0-50
μA
1/4" ID
3" ID PLASTIC PIPE 24" LENGTH
D
7289/2C39
A
WATER JACKET
1/4" ID
A
FLOW INDICATOR
TO 117-V AC
B
B
F
C
PUMP
C
D
T
DRAIN VALVE
3/8" ID TUBE, 18" LONG
3/8" ID TUBE, 6" LONG
3/8" ID

PART	DESCRIPTION		
A	1/4" TO 3/8" REDUCER.	D	3/8" OR 1/2" NPT TO 3/8" TUBE ADAPTER.
B	1/2" NPT TO 3/8" TUBE ADAPTER.	E	3/8" TUBE TO EXCHANGER ADAPTER.
C	3/8" TUBE TO PUMP ADAPTER.	F	3/8" OD COPPER TUBE W/WIRE SOLDERED TO IT. (SEE TEXT).

Fig 5—Details of the water-cooling system. Recommended pumps are: (1) Little Giant Pump Co. Model 1-42A or larger, available from most hardware stores; or (2) Calvert Engineering, Cal Pump Model 875S (160 gal/h), available from Calvert, 7051 Hayvenhurst Ave, Van Nuys, CA 91406, tel. 213-781-6029. The flow indicator (Model 15C; requires two 1/2-inch NPT adapters) is available from Proteus Industries, 240 Polaris Ave, Mountain View, CA 94043 tel. 415-964-4163.

pliers. Gently unscrew the tube from the radiator. If the hose clamp slips slightly, tighten it.

Some 7289/2C39 tubes use an air radiator that is attached with setscrews. To remove the radiator, simply remove the setscrews and pull the radiator off.

The air radiator will be replaced with a water jacket that allows water to be circulated past the tube anode and through a radiator, where it is cooled and circulated past the tube anode again. I have successfully used two different types of water jackets; both are described here.

The water jacket shown in Fig. 3 will work with any type of 7289/2C39. It is fabricated from a 1-inch-OD copper tubing cap and two short pieces of 9/32-inch-OD brass tubing. The copper tubing cap should be available from a local hardware store or plumbing supply house. Brass tubing is available from many hobby stores and metal supply houses.

Mark and drill the copper cap so that the brass tubing is a snug fit. Thoroughly clean the parts until they shine. Push the tubing into the holes in the end cap and degrease the assembly with alcohol. Using plenty of flux, solder the seam around each section of tubing. Allow the jacket assembly to cool.

Meanwhile, thoroughly clean the 7289/2C39 anode to a bright finish. Check the water jacket for fit. In some cases, you'll have to use a 0.005- to 0.010-inch-thick copper shim to fill the gap between the copper cap and the tube anode. This shim helps eliminate pin holes in the solder.

Using plenty of flux, solder the water jacket to the tube anode. Solder it quickly with a hot, high-wattage iron. Allow the tube to cool in the air after soldering to avoid thermal shock and possible breakage. After the tube has cooled, use plenty of alcohol to remove all traces of flux from the tube and water jacket.

The second type of water jacket is shown in Fig. 4. This jacket will work only with 7289/2C39 tubes that have a screw-on air radiator. It is designed to thread onto the tube anode just like the air radiator did. This jacket is machined from a piece of 1 1/4-inch aluminum rod. The water inlet and outlet tubes are made from 9/32-inch-OD, 1/4-inch-ID aluminum tubing that is epoxied in place. A rubber gasket seals the jacket against leaks.

If you have access to a lathe, you should have no trouble duplicating the jacket. You could have one made up at a local machine shop. Complete screw-on water jackets are also available from the author.

After you unscrew the air radiator from the 7289/2C39, check for and remove any burrs from the tube anode. The anode surface must be flat if the rubber gasket is to be effective. Screw the water jacket onto the tube. Tighten by hand only. Do not use any tools, or you could damage the tube or jacket! Do not use the water inlet and outlet tubes for leverage—they have thin walls and break easily.

Water System

Fig. 5 depicts the complete water-cooling system. Recommended pumps and accessories that have proven reliable and effective are listed in the caption.

Any small pump, such as a fountain pump, that can deliver 160 to 200 gallons per hour can be used here. Most inexpensive pumps are not self-priming, which means that they won't pump water if they have air in the rotor. Although water can be forced through the pump for the initial prime, my system uses gravity priming. The water reservoir is a 2-foot length of 3-inch-OD plastic pipe that is available from hardware or plumbing stores. The outlet is at the bottom, and the inlet about halfway up the column. The inlet is located here to eliminate aeration that ionizes the water and reduces its effectiveness. The outlet

Microwave Radiation Safety

Intense RF radiation concentrated on body tissues can produce heat damage; the extent and penetration will depend on the radio frequency in use and on exposure duration. You should be aware of the approximate intensity of RF radiation of the transmitting equipment and antennas you come in contact with.

RF intensity is commonly expressed in milliwatts per square centimeter (mW/cm^2), which is the power flowing away from a source through a unit sampling or interception area at some specified distance. Although the United States as yet has no federal RF protection standard, a useful interim guide is the 1982 standard of the American National Standards Institute (ANSI '82). The most stringent level in this standard Is 1 mW/cm^2 for frequencies between 30 and 300 MHz. Above 300 MHz, the protection level rises until it reaches 5 mW/cm^2 at 1500 MHz. Beyond 1500 MHz, the recommended level remains at 5 mW/cm^2. These levels represent the average power density allowed over any six-minute period and are for the sum of all polarizations from a given source.

At 1296 MHz, where one wavelength (λ) equals 23 cm, a thick resonant dipole feeding a calibrated power meter with matched coaxial cable (itself free of pickup) may be used to obtain an indication of power density. A reasonably lossless resonant dipole has an effective aperture of $\lambda^2/8$; at 23 cm this is 66 cm^2. The power meter reading in milliwatts, divided by 66, is the indicated power density. For this to be a reliable indication, the dipole must be positioned far enough from the RF source to be in its far field. For a small source, the distance should be at least $\lambda/2$, and here that would be about 12 cm (45 inches). The dipole should be oriented for alignment with the dominant polarization. Note that the power meter must be capable of readings well below 1 mW.

This arrangement would be useful for checking leaks along the coaxial route that the high power (here 250 W) takes to a load, be it dummy load or antenna. Cable connectors may not be tightly secured, or they may be faulty. For equipment operating in the SHF region, waveguide flanges may not be clamped properly.

Direct measurement of electric field strength near an antenna (with a calibrated instrument, preferably one with the indicating meter shielded and possibly positioned at the center of the sampling dipole) is another way to check for adequate protection. A field strength of 60 V per meter (V/m) corresponds to 1 mW/cm^2; 134 V/m corresponds to 5 mW/cm^2. At a distance 60 cm (2 feet) from an isolated dipole fed with 26 watts, the field strength would be about 60 V/m. This is a far-field field strength for all frequencies where the half wavelength is less than 60 cm, or for frequencies above 250 MHz. For full 250 watts applied to the dipole, the 60 V/m level occurs at a distance of 1.8 meters (6 feet), and at this distance this holds for all frequencies above 80 MHz.

With SSB or CW keying, the fields during Amateur Radio operation are highly intermittent, and usually include considerable pauses or intervals for listening. These factors reduce the average power density over the six-minute averaging period.

Further information on RF safety and protection estimates can be found in Chapter 7 of *The Satellite Experimenter's Handbook*, published by the ARRL. The following rules of good practice for RF protection are recommended:

- Never operate an RF amplifier with equipment shielding removed.
- Never handle antennas with RF power applied.
- Never guess that RF levels are safe. Take the time to consult a reliable reference for an estimate, or measure levels carefully. Allow a "cushion" of about 6 dB (factor of four in power density). If possible, borrow an RF radiation monitor (after learning how to use it), or consult with a ham who is well informed on RF protection.
- Never look into an open end of a power waveguide; never point a powered directive antenna (a beam or a paraboloid, for example) toward people. Keep all VHF and UHF transmitting antennas as high as possible, distant from humans.
- Use good-quality, well-constructed coaxial cable and connectors to avoid RF leaks.
- Think RF and electrical safety first; test later!
- Watch *QST* for news on RF measurement techniques and progression, protection standards and proposed federal and state RE regulations.—*David Davidson, W1GKM*

directly feeds the pump. The pump and the reservoir outlet port should be mounted in the same plane. The pump should be oriented so that air bubbles will rise into the impeller output port and can be blown out once the pump starts running.

Flow Indicator and Heat Exchanger

Water cooling is best described as "super quiet." There is no noisy fan to reassure you that the tube is receiving adequate cooling. If water flow is reduced or cut off during amplifier operation, tube damage is virtually assured.

Flow interlocks and switches to shut down the amplifier if water flow is reduced are hard to find and expensive. Flow indicators, however, are inexpensive and reliable. A flow indicator has a spoked rotor that turns as water passes through the unit. If the wheel is turning, there is water flow; if not, you have a problem. Changes in flow rate can be observed by watching for speed changes in the rotor. A small lamp illuminates the flow indicator, making it easy to see rotation. The flow indicator should be mounted where it can be seen from the operating position and monitored during operation.

Heat exchangers, or radiators, remove the heat from water as it passes through. For this application, a small automobile transmission-oil cooler works great. Most auto- parts stores and speed shops have a good selection. Pick one that is similar in size and aspect ratio to a whisper fan (approximately 4 × 4 × 1 inches). Some come with mounting brackets. Look for a cooler with the input and output ports on the top so air bubbles will rise to the top and move on without becoming trapped. Trapped air degrades cooler performance.

If you use the amplifier for high-duty-cycle modes such as ATV or FM, or for long, slow-speed CW transmissions (EME, for example), you should use a small axial whisper fan to increase the effectiveness of the heat exchanger. A fan isn't necessary during normal operation, or even for sustained operation at moderate power levels, but I highly recommend one if you plan prolonged operation at maximum power. Locate the fan so the warm exhaust air won't heat up other equipment.

Hoses and Fittings

Most hardware stores carry a complete line of brass fittings and adapters that can be used for this project. Brass, however, will eventually corrode and pollute the water supply. Plastic fittings are cheaper and don't corrode, but they are harder to find. Recreational vehicle suppliers are my main source for these parts. They are used extensively in drinking water systems for mobile homes and travel trailers. Procure the fittings when you have the rest of the parts in hand, as there are many variables to consider.

You can use any relatively soft, thin-wall vinyl tubing for all water lines. The main runs are made from 3/8-inch-ID hose, while 1/4-inch-ID stock is used to connect to the 7289/2C39 water jacket. The 1/4-inch-ID tubing fits snugly over the 9/32-inch-OD inlet and outlet tubes on the water jacket, so no clamps are required. All other hose connections should be secured with stainless-steel clamps to prevent leaks. Any leaks mean air in the system and deterioration of cooling performance.

Safety

The tube anode, and hence the water jacket and water, are in direct contact with the high-voltage supply, so some safety precautions must be observed. Approximately 12 to 18 inches of tubing should run between the 7289/2C39 jacket and any other component in the cooling system. This will allow enough resistance in the water to provide adequate current limiting, should the water contact any components that are grounded.

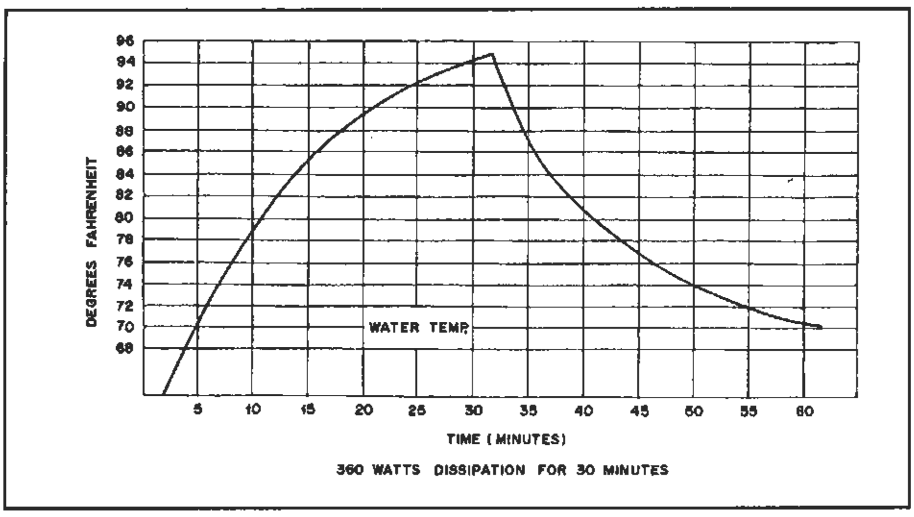

Fig 6—Performance graph of the water-cooling system.

It is best to ground the water supply at the pump. Do this by replacing a short section of the tubing that runs to the flow indicator with a piece of brass or copper tubing. Solder a wire to this metal tubing and connect the other end of the wire to your station ground. Use at least 24 inches of vinyl tubing between the anode cooling jacket and the ground point.

On the warm-water side of the 7289/2C39, run 12 inches of vinyl tubing to a small metal fitting or short section of metal tubing, and then another 12 inches of vinyl tubing to a grounded point (this can be at the heat exchanger). You can measure the water leakage current to ground by placing a microammeter between the metal fitting that connects the two vinyl hoses and ground. Leakage current should be less than 10 µA with clean water and an anode potential of 1 kV. As the water ages, the leakage current will rise; when this happens, replace the water.

Grocery stores carry distilled water for use in steam irons. It may be deionized and not truly distilled, but it works fine for about four to six months in this application. Filters can be purchased from scientific supply houses, but they're not really worth buying because deionized water is so cheap.

Do not use tap water under any circumstances! When you turn on the water system for the first time, run a gallon of water through it for half an hour to wash out fabrication impurities. Replace with clean water before using the system to cool the amplifier.

Water was chosen because it's inexpensive, nontoxic, nonflammable and easy to clean up if you have a leak. Better liquid coolants are available, but they are toxic. Don't use them!

Cooling Performance

I have used water-cooling systems for several years with no problems whatsoever. Fig. 6 is a graph of several transmit/receive cycles on a water-cooled, 500-W output, 23-cm power amplifier. For this test, I used two of the amplifiers described in this article coupled with a pair of hybrid combiners. This particular cooling system used 1 gallon of water. Experiments indicate that, during extended operation, the water temperature rises only 30° to 35°F above ambient room temperature. Typically, the tube anode and water average 10° to 15°F above ambient during casual operating.

Flow rates in this system are typically $^1/_3$ gallon per minute per tube, which is more than adequate. At this rate, more than 300 W of dissipation from a single inefficient 7289/2C39 were required to boil the water in the water jacket. The water should not be allowed to boil because this will heat the rubber gasket.

Tubes

It is not really necessary to buy a new 7289/2C39. Used tubes can be found surplus for around $1 to $5 and, in many cases, will perform as well as a new tube. Most used tubes have been sitting around for several years, so it's a good idea to run them through the dishwasher to clean them up and then run the filaments for about 24 hours. This will restore operation in many cases.

If you buy a new tube, you should be aware that the 7289/2C39 is being run far in excess of its ratings in this amplifier. The manufacturer's warranty will not cover tubes run in this application.

Contrary to popular opinion, glass tubes will work. Physically, they are not as rugged as the ceramic version, but the glass-to-metal seal seems to provide better shelf life than the ceramic seal. The glass tubes make great driver tubes and will work fine for power levels up to 100-W output. Pulse tubes (7815, 7211) are not recommended because of their poor thermal stability at high power levels. Also, they generally are 30 to 40 MHz lower in resonant frequency in this amplifier compared to the 7289/2C39. Some 7289 tubes can be as much as 30 MHz lower in frequency. Minor length adjustment of the anode-tuning post may be required to accommodate amplifier and tube differences.

Tube Insertion

Extreme care must be exercised when inserting the 7289/2C39 tube. Never force the tube in place, as damage (bending) of the cathode finger stock may result. Observe the layout of the finger stock to get an idea of how the tube inserts. Carefully position the tube so it is straight as you gently push. It should slide in snugly without any solid resistance.

Testing

After you have completed all of the parts for the amplifier, it's time to test everything before hooking it all together. Test the water-cooling system by turning it on and watching for steady water flow as indicated on the flow meter. The tube and water jacket can be removed from the cavity amplifier for this test.

Check all of the power-supply voltages first without connecting them to the RF deck. Then, without the tube in place, hook the bias and filament supplies to the cavity and check the voltages again at the tube finger-stock connections. Connect the high-voltage supply to the RF deck and bring the voltage up slowly with a variable autotransformer. Monitor the high voltage on the anode-bypass-capacitor plate, and look and listen for any possible arcing between the anode-bypass-capacitor plate and ground. Use extreme care when measuring and testing the high-voltage supply. If everything looks okay with the power supplies, shut them off and disconnect them.

You can make a safe, low-power test of the cavity resonance without applying any voltage. With the tube in place, insert a 2-inch-long coupling loop on the end of a piece of coaxial cable between the spring fingers of the anode down into the cavity. Connect the amplifier output probe/connector to a device capable of detecting low-level RF at 23 cm (for example, a spectrum analyzer or microwattmeter). Feed a signal from an L-band signal generator into cable attached to the wire coupling loop that you inserted into the cavity. Set the signal generator for various frequencies in the 23-cm band and tune the amplifier anode tuner. There will be sharp peak in output at cavity resonance.

This testing method can be used to determine cavity tuning range, anode-bypasscapacitor effectiveness and resonance of various tube types for use in this amplifier. Any cavity amplifier can be tested completely without ever applying high voltage. The better your test equipment, the easier the amplifier is to test. If all dimensions were followed strictly, the amplifier will tune as designed.

Amplifier Hookup

Installation and operation of this amplifier is relatively straightforward, but as with any amplifier, several precautions must be followed. If these are adhered to, the amplifier will provide years of reliable service.

The amplifier is designed to be operated in a 50-ohm system and should never be turned on without a good 50-ohm load connected to the output connector. Never operate it into an antenna that has not been tuned to 50 ohms!

Drive power to the amplifier should *never* exceed 15 W. *Never* apply drive power in excess of 1 W unless all operating voltages are present and the tube is biased on. Otherwise, the tube grid-dissipation rating will be exceeded and you will probably ruin it.

As in all TR-switched systems, some type of interlock or sequencing of transmit and receive functions should be incorporated. In most systems, the sequence for going into transmit is something like this: First, switch the antenna changeover relay from the receiver to the power amplifier. Next, bias the power amplifier on. Last, key the exciter and apply drive to the amplifier. To go to receive, unkey the exciter, remove operating bias from the amplifier and switch the antenna relay back to the receiver.

If the antenna relays are switched while the power amplifier is operating and putting out power, damage to the relay contacts and/or the amplifier is likely. If there is a momentary removal of the antenna while the power amplifier is biased on, oscillation may occur. This can damage the TR relay, the tube or even the receive preamplifier.

Tune-up and Operation

This is it — the big moment when you will see your project come to life! Connect an accurate UHF power meter and a 50-ohm antenna or load to the amplifier output connector. A Bird Model 43 wattmeter with a 100- or 250-W, 400-1000 MHz slug will give reasonable accuracy, depending on the purity of the drive signal. Apply filament power and tube cooling, and allow 3 to 5 minutes for the filaments to warm up. Turn on bias supply (the amplifier will draw maximum current if the anode voltage is applied without bias). Apply 300 to 400 V to the anode. There should be no current flowing in the tube as indicated on the cathode-current meter. Ground J1 on the bias supply to apply transmit bias and observe cathode current. As R1, the bias control, is turned clockwise, quiescent idling current should increase. Set for about 25 mA.

Apply 1 W of RF drive power. Turn the anode tuner while observing the RF output power meter and tune for maximum output. The output should go through a pronounced peak at'cavity resonance. Adjust C2 and C3 on the input tuning network for maximum amplifier output. If possible, use a directional wattmeter between the driver and the amplifier input to check that best input SWR and maximum amplifier output occur at roughly the same setting.

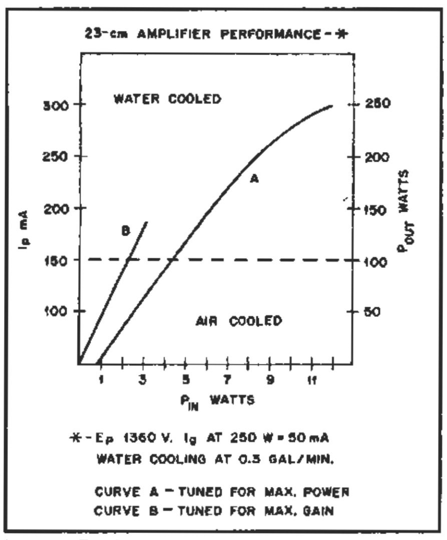

Fig 7—Performance of the cavity amplifier under different drive and plate-current conditions.

Depending on the amount of drive power available, you may want to tune the amplifier for maximum power output or maximum gain. Fig. 7 shows what you can expect from different drive levels.

Once the amplifier is tuned for best input SWR and maximum output with 1 W of drive, anode voltage and drive power can be increased. Increase both in steps; be sure to keep the anode tuner peaked for maximum output power. When you get to the 100-W output level, very carefully readjust the input circuit for maximum output. The input capacitor closest to the cathode is critical and should need to be rotated less than 90 degrees maximum. Maximum output power will be roughly coincident with best input SWR.

Increase the drive power and keep the anode tuner peaked for maximum output. Increase the drive until you reach the desired output level, but *do not exceed 400-mA* I_K*!* At 1400-V dc and 350-mA I_K, output power with a good tube should be about 230 to 250 W. At lower anode voltages, I_K will be higher for the same output power. Higher anode voltages result in higher gain, lower drive levels, lower grid current and lower plate current for a given output power.

The anode tuner's tuning rate is approximately 5 MHz per turn. Clockwise rotation of the tuner lowers the resonant frequency of the cavity. This control will require readjustment as you make large frequency excursions within the 23-cm band (for example, if you go from 1296 weak-signal work to the 1269-MHz satellite segment). You should also check the input SWR if you move more than 15 MHz. Generally, amplifier tuning does not change much after initial setup. You should be able to turn it on and use it without retuning as it heats up. Slight adjustments may be necessary, however, depending on cooling, inherent thermal differences from tube to tube and duty cycle of the operating mode. Always keep the anode tuner peaked for maximum output, and check it from time to time, especially while you are first learning how the amplifier operates.

The output loading control is the output connector and probe assembly. Loading is changed by minor rotational adjustment of the N connector. First loosen the jam-nut (or setscrew) slightly. While observing output power and keeping the anode tuner peaked, rotate the loading control ±30 degrees maximum for greatest output power. This should be done only once and should not need repeating unless another tube is installed. Even then it may not be required.

Conclusion

This cavity amplifier for the 23-cm band is capable of safe, reliable operation at output powers in excess of 200 W. More than 50 of these amplifiers are in operation, and you can build one, too. I would like to thank Mike Stahl, K6MYC, Bill Troetschel, K6UQH, William Jungwirth, AA6S, Lem Moeschler, W6KGS and Joseph Cadwallader, K6ZMW, for their help and encouragement during the development of this project.

References

Laakmann, "Cavity Amplifier for 1296 MHz," *QST*, Jan. 1968, pp. 17-22.

Laakmann, "Kilowatt Amplifier for 1296 MHz," *Ham Radio*, Aug. 1968, pp. 8-17.

"Power Amplifier for 23 cm," *VHF Communications*, April 1976, p. 222.

"Power Amplifier for 24 cm," *VHF Communications*, March 1978, p. 175.

Stefanskie, "1270 MHz ATV Power Amplifier," *Ham Radio*, June 1977, pp. 67-73.

Tech Report #6, Crawford Hill VHF Club, July 1971.

Tech Report #13, Crawford Hill VHF Club, Dec. 1972.

By Zack Lau, KH6CP/1 **From *QEX*, May 1994**

A 2-W, 13-cm Amplifier

This amplifier uses a resistively stabilized Hewlett Packard ATF-44101/AT-8140 power GaAs FET running class A to provide 14 dB of gain and a 2-W output. It wasn't too difficult to choose the device—there aren't that many linear devices that cover the 2.3 GHz band at this power level. Unfortunately, it is fairly pricey; the circa 1990 price was around $90. But since this transistor has been around a long time, it may be available surplus. Several hundred were sold at the bargain price of $3 each about a year and a half ago. I bought a few—I should have bought more.

Even if you do buy your FETs at bargain prices, you still will want to ensure that they are biased safely. As has been pointed out many times in the literature, having the negative gate bias supply fail while the drain supply is applied may result in the destruction of the device. Thus, designers have devised elaborate protection schemes to shut off the drain supply if the gate supply fails.

Fig 1 shows a simpler power-supply circuit. The drain supply is controlled by an ordinary 723 voltage regulator with current limiting. The current limiting protects the device—instead of the expensive FET getting hot and self-destructing, the cheap TIP 30 power transistor gets warm. Doing this eliminates the need for complex shutdown circuits to handle a failed bias supply. While there are more modern regulator chips than the 723, they aren't significantly better for this application. In fact, some are "improved" to the point where they are tougher to use. For example, chips with better current-sensing circuits often require special low-value resistors. While you could use the copper wire table in the *ARRL Handbook* to make your own low-value resistors from wire, it makes more sense to me to use a 723 with standard, readily available parts.

The gate bias supply is produced by an NE555 timer running as an oscillator and driving a voltage inverter. An LM337 adjustable 3-terminal voltage regulator controls the gate supply. Now, you have to be a little careful using adjustable voltage regulatorswith bias supplies. While I haven't experienced this myself, it is entirely possible that a noisy potentiometer could momentarily present an open circuit between the wiper and the resistance element. Conceivably, in some circuits, this could result in voltage spikes at the output of the voltage regulator as the voltage is adjusted. The cure for this is to wire the potentiometer so that it never presents an open circuit, even with an intermittent wiper, as I've done with R9 in Fig 1.

For some other GaAs FET circuits, it may be advantageous to use an active-bias supply which operates inside the feedback loop of the current limited supply. This allows the bias point to be unaffected by temperature variations. The current limiting acts as a failsafe, protecting the transistor if a more negative gate supply voltage is unable to turn off the transistor. I've successfully used this circuit when only a milliamp of bias current was needed. Due to the high current required by the gate circuit of the 2-W amplifier, I decided it was impractical to active-bias the FET in this application; the ICL7660 can't provide the needed current. But there are highercurrent versions of this chip, for those who don't mind the price.

RF Design

I found the AT-8140 rather difficult to get unconditionally stable. I ended up placing a 10041 chip resistor from gate to ground, which resulted in pretty good stability, except around 2.3 GHz. Unfortunately, it doesn't appear possible to dc decouple this resistor without adversely affecting stability, so quite a bit of bias current is needed. Because of this, it is conceivable that you may find a FET that needs too much gate voltage for a typical 10041 chip resistor to handle. Whether this is the case with a particular FET depends on both the threshold voltage and the transconductance of the device. The threshold voltage can vary quite a bit, particularly in older devices, though the manufacturers have tightened up this specification in recent data books. Worst case would be a large negative threshold voltage and a high transconductance. Stacking chip resistors in parallel might be a solution, though this might also invite unwanted parallel resonance effects, similar to what is seen with paralleled capacitors.

While in-band stability is not too critical, since the source and load impedance matching is usually pretty good at the operating frequency, I used a series resistor in the gate decoupling line to make the amplifier unconditionally stable, according to

Table 1—Measured Amplifier Performance

Input to MGF 1801 driver (dBm)	*Driver Output (dBm)*	*AT 8140 Output (dBm)*	*AT 8140 Gain (dB)*
−7.0	5.7	19.8	14.1
0.0	12.5	26.6	14.1
3.0	15.5	29.7	14.2
4.0	16.5	30.8	14.3
5.0	17.5	31.7	14.2
6.0	18.4	32.6	14.2
7.0	19.2	33.3	14.1
8.0	20.3	33.9	13.6
9.0	21.2	34.1	13.1
10.0	22.1	34.6	12.5
13.0	24.2	34.9	10.7

The 1-dB compression point is at 34.1/33.2 dBm output.
Comparing measurements made using an HP8563E spectrum analyzer against those made with an HP 453B/8481A power meter, a signal that measures 4.7 dBm on the spectrum analyzer reads out as 3.8 dBm on the power meter. The values listed are from the spectrum analyzer, taking into account a 30.2-dB, 25-W Bird attenuator inserted between the amplifier and the analyzer.

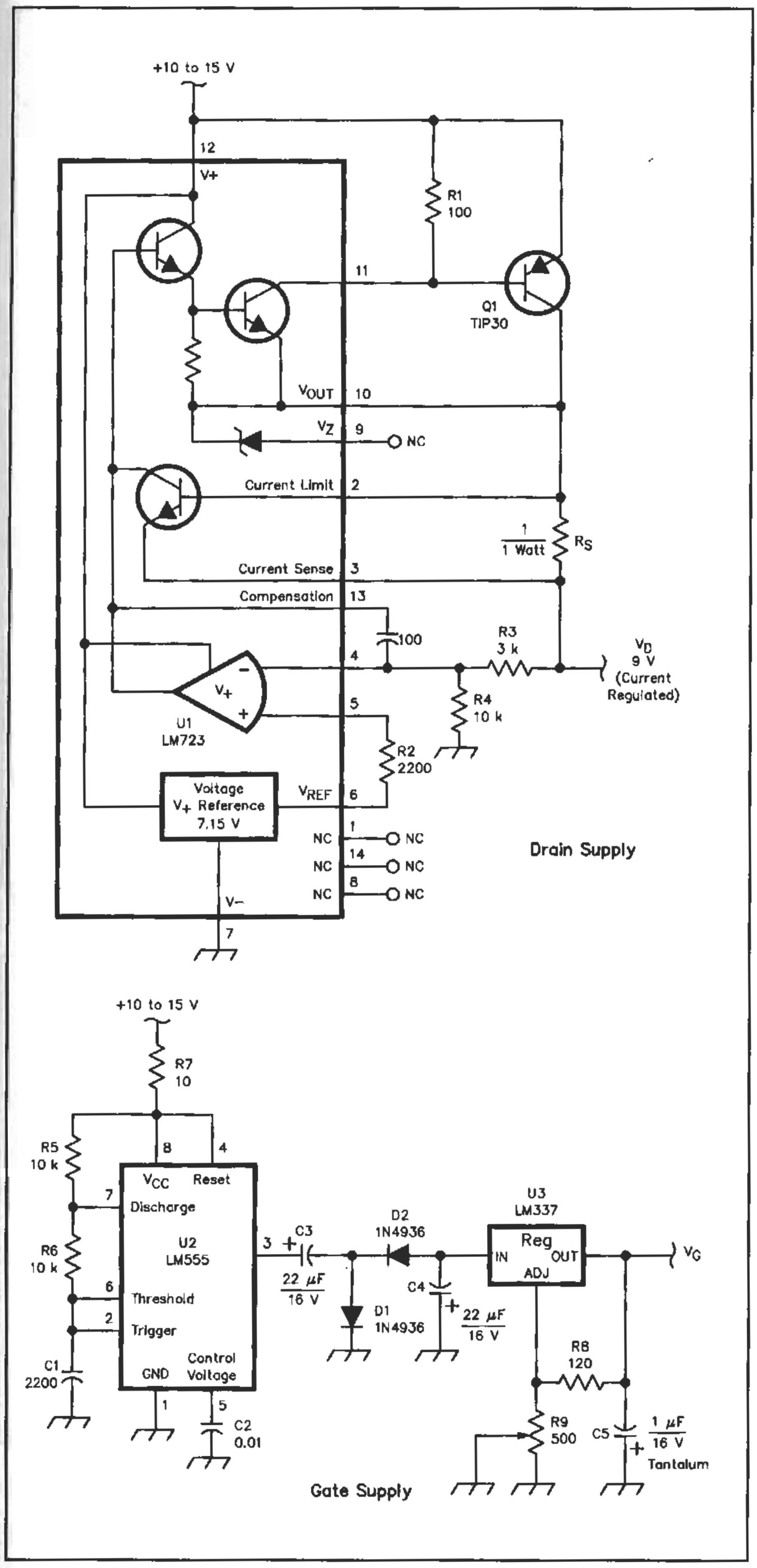
+10 to 15 V
12
V+
R1
100
11
Q1
TIP30
VOUT
10
VZ
9
NC
Current Limit
2
1
1 Watt
RS
Current Sense
3
Compensation
13
100
R3
3 k
4
VD
9 V
(Current
Regulated)
V+
U1
LM723
5
R4
10 k
R2
2200
Voltage
V+ Reference
7.15 V
VREF
6
1
NC
NC
14
NC
NC
8
NC
NC
Drain Supply
V−
7
+10 to 15 V
R7
10
R5
10 k
8
4
VCC Reset
7
Discharge
D2
1N4936
U3
LM337
R6
10 k
U2
LM555
3
C3
Reg
IN
OUT
ADJ
VG
22 µF
16 V
C4
6
Threshold
D1
1N4936
22 µF
16 V
2
Trigger
R8
120
C1
2200
Control
GND Voltage
1
5
C2
0.01
R9
500
C5
1 µF
16 V
Tantalum
Gate Supply

Fig 1—Power supply for the 2-W, 13-cm amplifier.

the computer model. This series resistor also improved the input match bandwidth, an important consideration if you don't want to tune the amplifier. Of course, improving the bandwidth with resistance also reduces the gain. The loss isn't too bad, however, seeing as this design is within a few dB of the maximum stable gain the data sheet specifies. The design was optimized with 150 Ω of series resistance, but I decided that such a high value was unwise. To accommodate such a high series resistance, it is entirely likely that the negative supply voltage would have to be more negative than the VGS limit. Then, if the 10041 resistor were to fail opencircuited, the supply could damage the gate of the transistor. For this reason, I chose to use a 5041 series resistor. I didn't reoptimize the design for 50 Ω of series resistance, since the likely improvement was small compared to the variations I've seen when optimizing the unit on the bench.

Finally, I discovered that additional gate circuit bypassing was required at audio frequencies to prevent the circuit from oscillating. This was done with a 1-µF capacitor that is not shown in the computer model. This frequency range wasn't covered by the model.

Several dB of additional saturated output power was obtained by modifying the output network with foil tabs. AT-8140s saturate at about 2.2 W output with around 200 mW of drive. An advantage to optimizing the circuit by modifying the 5041 striplines at the input and output of the amplifier is that the stability should not be adversely affected. But the board does have to be made larger to accommodate such tuning.

The original design was optimized using *Microwave Harmonica*, with the results shown in Table 2 and Fig 6. After I tweaked the output network to get maximum output power, I entered the changed circuit components into the computer and analyzed the circuit, resulting in the analysis shown in Table 3 and Fig 7. Table 1 shows the performance of the circuit in its final configuration.

Construction Notes

To ground the drain bypass capacitor and the gate resistor, I cut slots in the board and connected the pads to the ground plane with 1-mil copper foil. Copper foil was also used to connect the gate bypass capacitor, although in this case I trimmed the board so that it wasn't necessary to cut a slot.

For high-quality grounding without lots of tiny screws, I decided to try using a heat spreader made out of 1-mil copper foil. I cut a hole in the circuit board for the FET and then carefully soldered the foil across the hole in the circuit board, taking care not to put any solder where it would interfere with heat-sinking action. I didn't want any air pockets, as they are extremely poor conductors of heat. Then, after attaching the board to a piece of 0.25-inch sheet aluminum with suitably tapped holes, I used a scribe to punch holes for the 0-80 screws to

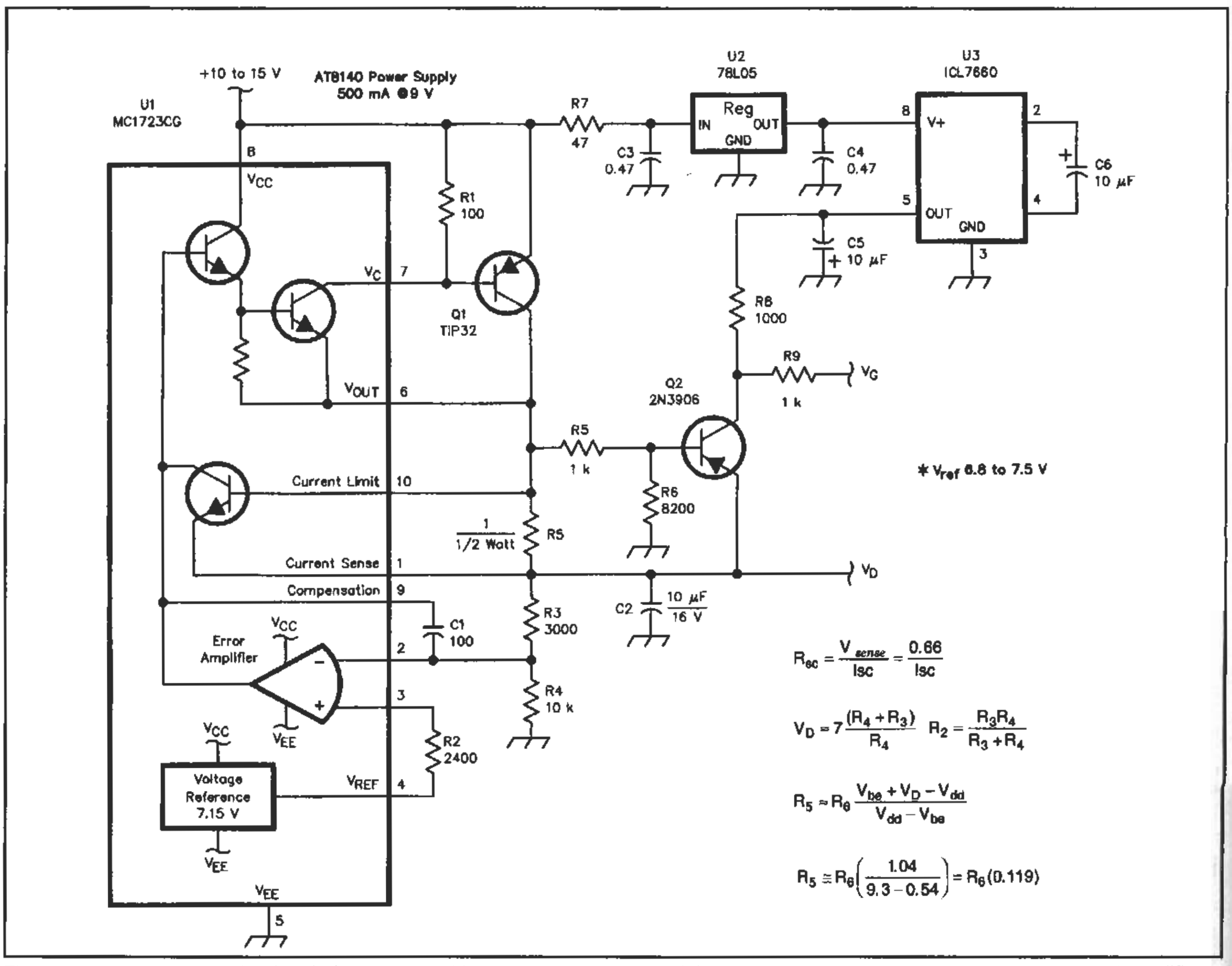
+10 to 15 V
AT8140 Power Supply
500 mA @9 V
U1
MC1723CG
8
VCC
VC
7
Q1
TIP32
R1
100
R7
47
U2
78L05
Reg
IN
OUT
GND
C3
0.47
C4
0.47
U3
ICL7660
8
V+
2
C6
10 µF
5
OUT
GND
4
3
C5
+ 10 µF
R8
1000
R9
1 k
VG
VOUT
6
Q2
2N3906
R5
1 k
R6
8200
Current Limit
10
1
1/2 Watt
R5
Current Sense
1
Compensation
9
VD
C1
100
C2
10 µF
16 V
Error Amplifier
VCC
2
R3
3000
R4
10 k
3
VCC
VEE
R2
2400
Voltage Reference 7.15 V
VREF
4
VEE
VEE
5
* Vref 6.8 to 7.5 V

$$R_{sc} = \frac{V_{sense}}{Isc} = \frac{0.66}{Isc}$$

$$V_D = 7\frac{(R_4 + R_3)}{R_4} \quad R_2 = \frac{R_3 R_4}{R_3 + R_4}$$

$$R_5 \approx R_6 \frac{V_{be} + V_D - V_{dd}}{V_{dd} - V_{be}}$$

$$R_5 \cong R_6\left(\frac{1.04}{9.3 - 0.54}\right) = R_6(0.119)$$

Fig 2—An active biasing circuit suitable for use with amplifiers that draw minimal gate current. Do *not* use this with the 2-W amplifier.

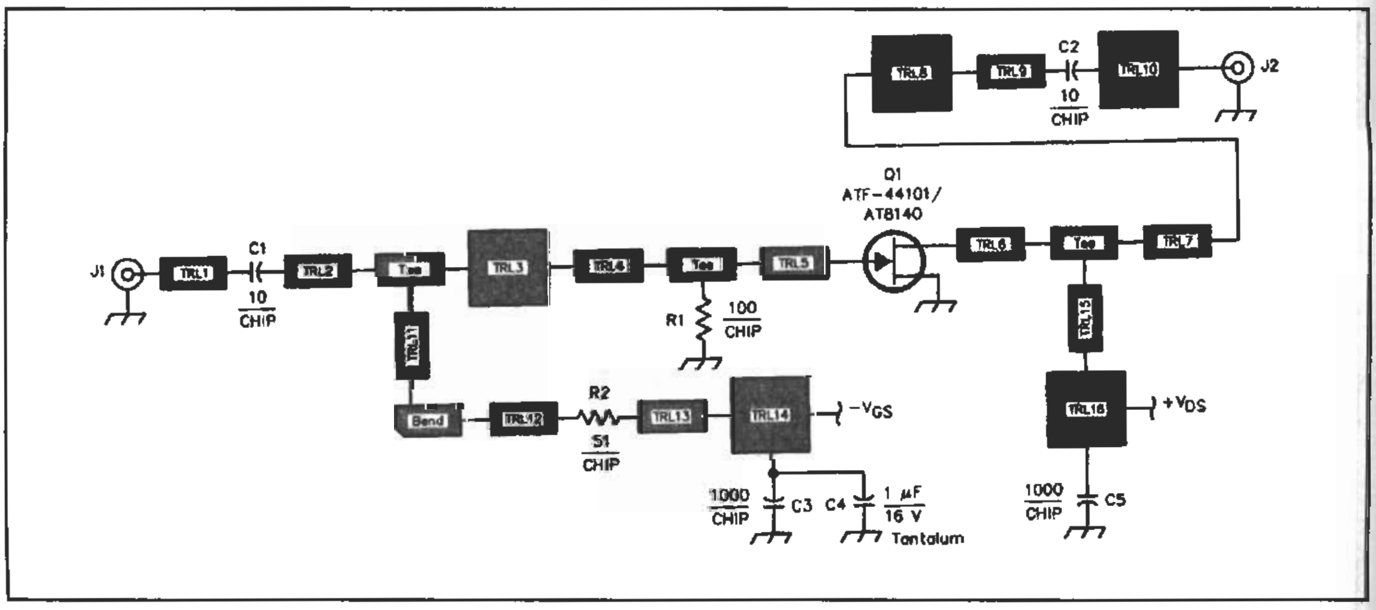
TRL8
TRL9
C2
10
CHIP
TRL10
J2
Q1
ATF-44101/
AT8140
J1
TRL1
C1
10
CHIP
TRL2
Tee
TRL3
TRL4
Tee
TRL5
TRL6
Tee
TRL7
R1
100
CHIP
TRL11
TRL15
Bend
TRL12
R2
51
CHIP
TRL13
TRL14
-VGS
TRL16
+VDS
1000
CHIP
C3
C4
1 µF
16 V
Tantalum
1000
CHIP
C5

Fig 3—Schematic diagram for the 2-W amplifier.

C1, C2—10-pF, high-quality porcelain chip capacitors. I used 55-mil capacitors, although 100-mil capacitors should work just fine.
C3, C5—1000-pF NP0 chip capacitors
C4—1-µF, 16-V tantalum capacitor.
J1, J2—SMA jacks.
Q1—Hewlett Packard ATF-44101/AT-8410 medium power GaAs FET. This device must be properly heat-sinked.
The absolute maximum channel temperature is 175°C. The case-to-junction thermal resistance is 23°C/W.
TRL1-16—Microstriplines etched on 0.031-inch-thick, $\varepsilon_r = 2.55$ Teflon circuit board.

Table 2—*Microwave Harmonica* Analysis (Original Circuit)

```
MICROWAVE HARMONICA PC V5.5  File: \mh\2304\qexamp.ckt   22-DEC-93 17:54:54

Freq    MS11    PS11   MS21    PS21   MS12    PS12    MS22    PS22    MS21      K
GHz      dB     deg    mag     deg    mag     deg      dB     deg      dB
        AMP     AMP    AMP     AMP    AMP     AMP     AMP     AMP     AMP     AMP

 1.000   -0.70 -167.5  2.084    19.3  0.017   -53.7   -6.01 -128.7    6.38    1.64
 1.100   -0.57 -177.2  1.988     0.1  0.018   -71.2   -5.35 -139.0    5.97    1.43
 1.200   -0.47  173.3  1.941   -17.8  0.020   -87.3   -5.02 -148.2    5.76    1.27
 1.300   -0.40  164.0  1.945   -35.0  0.022 -102.6   -4.96 -156.6    5.78    1.17
 1.400   -0.36  154.7  1.999   -52.0  0.024 -117.7   -5.13 -164.3    6.01    1.09
 1.500   -0.38  145.2  2.110   -69.2  0.028 -132.9   -5.59 -171.4    6.48    1.05
 1.600   -0.48  135.2  2.288   -87.3  0.032 -148.9   -6.43 -177.6    7.19    1.02
 1.700   -0.73  124.6  2.542 -106.8  0.039 -166.3   -7.84  178.8    8.10    1.01
 1.800   -1.26  113.0  2.872 -128.6  0.047  174.0   -9.91 -177.3    9.16    1.01
 1.900   -2.31  101.1  3.240 -153.5  0.056  151.3  -11.05 -156.6   10.21    1.01
 2.000   -4.14   90.9  3.547  178.8  0.064  125.8   -8.31 -138.3   11.00    1.02
 2.100   -6.75   87.5  3.682  149.7  0.070   98.9   -5.06 -141.0   11.32    1.04
 2.200   -9.11   95.3  3.655  121.2  0.073   72.6   -3.15 -153.0   11.26    1.06
 2.300  -10.01  108.4  3.587   93.9  0.075   47.6   -2.42 -167.3   11.09    1.07
 2.304  -10.02  108.9  3.585   92.8  0.075   46.6   -2.41 -167.9   11.09    1.08
 2.400  -10.08  120.9  3.583   66.5  0.078   22.4   -2.72  177.8   11.08    1.09
 2.500   -9.15  137.4  3.638   36.0  0.082   -5.9   -4.55  162.9   11.22    1.11
 3.000   -0.30   98.5  0.945 -114.5  0.025 -146.5   -1.26 -164.2   -0.49    1.17
 4.000   -0.74   28.9  0.159  140.2  0.006  122.2   -0.11  150.2  -15.96    1.91
 5.000   -2.01  -15.1  0.071   59.4  0.003   56.4   -0.08  123.6  -22.98   13.43
 6.000   -0.78    1.8  0.044  120.4  0.003  125.4   -0.07  100.6  -27.20    9.68
 7.000   -0.36  -60.0  0.158   68.8  0.014   86.8   -2.71  123.5  -16.02    8.12
 8.000   -0.77 -109.6  0.086  -76.3  0.010  -50.3   -0.74   73.6  -21.26   14.53
 9.000   -3.99  178.3  0.030  151.2  0.004 -174.8   -0.10   53.3  -30.56   50.93
10.000   -1.70 -138.8  0.016    5.4  0.003   45.4   -0.10   33.1  -36.06   75.19
11.000   -0.67  163.8  0.024  -91.2  0.006  -46.2   -0.47   12.8  -32.44   53.16
12.000   -0.95  116.5  0.031  106.0  0.010  152.0   -0.48   -0.6  -30.15   33.01
```

Table 3—*Microwave Harmonica* Analysis (Optimized for Output Power)

```
MICROWAVE HARMONICA PC V5.5  File: b:qexamp.ckt   22-DEC-93 18:04:28

Freq    MS11    PS11   MS21    PS21   MS12    PS12    MS22    PS22    MS21      K
GHz      dB     deg    mag     deg    mag     deg      dB     deg      dB
        AMP     AMP    AMP     AMP    AMP     AMP     AMP     AMP     AMP     AMP

 1.000   -0.72 -168.6  2.386    39.7  0.020   -33.3  -17.28  -86.0    7.55    1.63
 1.100   -0.67 -178.3  2.331    20.4  0.022   -50.9  -15.61 -117.3    7.35    1.42
 1.200   -0.64  172.6  2.299     2.0  0.023   -67.5  -14.05 -142.9    7.23    1.27
 1.300   -0.64  163.7  2.294   -15.7  0.026   -83.3  -12.69 -165.0    7.21    1.17
 1.400   -0.66  155.0  2.314   -32.9  0.028   -98.6  -11.48  175.0    7.29    1.09
 1.500   -0.71  146.5  2.365   -49.9  0.031 -113.6  -10.42  156.4    7.48    1.05
 1.600   -0.79  137.8  2.453   -66.8  0.035 -128.4   -9.50  138.6    7.79    1.02
 1.700   -0.91  128.9  2.588   -83.8  0.039 -143.3   -8.71  121.3    8.26    1.01
 1.800   -1.10  119.4  2.786 -101.2  0.045 -158.5   -8.08  103.9    8.90    1.01
 1.900   -1.42  108.8  3.071 -119.4  0.053 -174.6   -7.62   85.7    9.75    1.01
 2.000   -1.98   96.3  3.482 -139.1  0.063  167.9   -7.46   65.7   10.84    1.02
 2.100   -3.10   80.3  4.065 -161.7  0.078  147.5   -7.86   41.8   12.18    1.04
 2.200   -5.77   57.1  4.815  170.7  0.096  122.2   -9.64    9.1   13.65    1.05
 2.300  -14.58   10.6  5.418  135.8  0.113   89.5  -15.23  -52.9   14.68    1.07
 2.304  -15.31    6.6  5.428  134.2  0.113   88.0  -15.55  -57.2   14.69    1.07
 2.400  -10.67 -157.1  5.138   96.3  0.111   52.3  -13.32  174.5   14.22    1.09
 2.500   -4.22  165.0  4.086   61.5  0.092   19.6   -7.41  119.7   12.23    1.11
 3.000   -0.53   96.9  1.222  -35.2  0.032  -67.2   -2.33   21.6    1.74    1.17
 4.000   -0.75   29.2  0.494 -155.6  0.019 -173.6   -1.06  -79.0   -6.13    1.69
 5.000   -1.99  -15.1  0.261   93.6  0.013   90.6   -1.04 -169.9  -11.67   11.23
 6.000   -0.78    1.9  0.137  130.1  0.010  135.1   -0.55  116.4  -17.28    7.89
 7.000   -0.39  -60.0  0.101  160.7  0.009  178.7   -0.92   -3.1  -19.91    8.22
 8.000   -0.75 -109.9  0.196   -1.0  0.023   25.0   -6.65  -66.0  -14.17   14.18
 9.000   -4.01  178.4  0.126 -132.4  0.018  -98.4   -1.60 -147.6  -18.01   39.98
10.000   -1.70 -138.8  0.076   53.9  0.015   93.9   -2.18  130.1  -22.42   57.57
11.000   -0.66  163.8  0.071  -60.9  0.017  -15.9   -9.01   29.7  -22.92   50.28
12.000   -0.97  116.5  0.092  168.5  0.030 -145.5   -8.83  160.9  -20.71   31.18
```

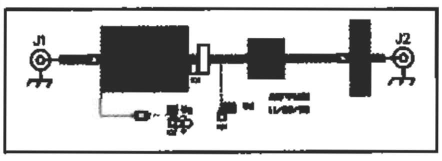

Fig 4—Parts placement diagram for the 2-W amplifier. Points labeled XX should be connected to the ground plane using copper foil.

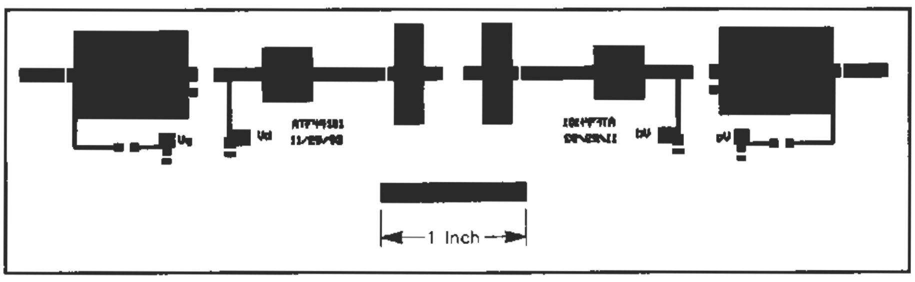

Fig 5—Etching pattern for the 2-W amplifier. Use 0.031-inch-thick, $\varepsilon_r = 2.55$ Teflon circuit board.

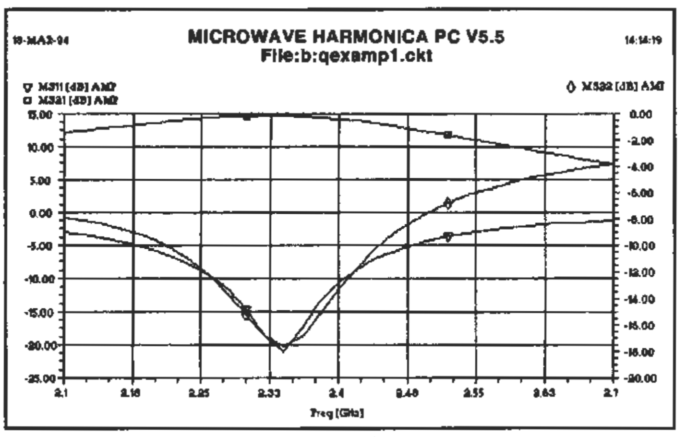

Fig 6—*Microwave Harmonica* analysis of the circuit's S parameters after computer optimization.

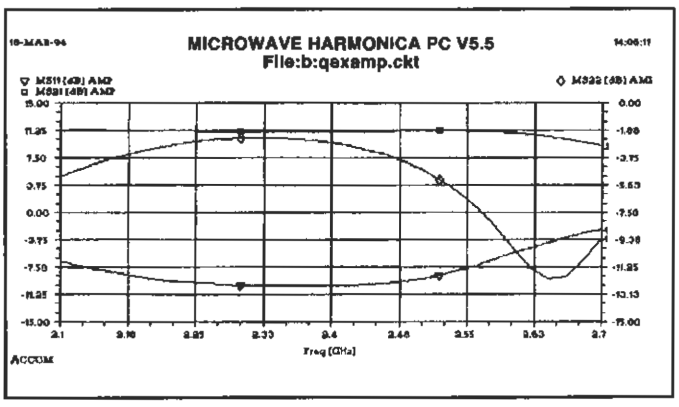

Fig 7—After the output network was optimized on the bench for maximum output power, *Microwave Harmonica* produced this analysis of the resulting circuit.

mount the FET. Ideally, clean holes would be punched into the copper, but I couldn't figure out a way to do that with the tools I have available. I used 0-80 screws, which are the largest screws I've found that will fit in the mounting holes. While I've used smaller screws in the past, I prefer using the largest screws possible since taps generally get tougher to use as they get smaller. I built one amplifier using 0.125-inch thick 6061 aluminum, but found that to be too thin to easily make the tapped holes for attaching the connectors and brass strips, although the finished mounting plate was stiff enough.

By Al Ward, WB5LUA **From *QST*, December 1985**

1296-MHz Solid-State Power Amplifiers

Explore the DX possibilities of the 23-cm band with these modern amplifiers.

With the ever-increasing number of ready-to-go 1296-MHz transverters available on the market today, there is a great demand for a simple and economical way to generate higher power than the typical 0.5 to 1 W output these transverters provide. If 1 W or less is used to drive a typical 2C39/7289 stripline or cavity amplifier, the low output power is often disappointing. A tube-type amplifier run with 1 kV on the plate typically offers a gain of 10 dB. This means that, at best, you can expect 5- to 10-W output from your 1-W or less input.

A popular way of generating higher power is to cascade two tube-type amplifiers for 50- to 100-W output. Here I will show you an alternative: two solidstate amplifiers that can replace the tubetype driver amplifier and provide 10- to 20-W output—enough to drive a two-tube amplifier to full output.

The NEC NEL1306 and the NEL1320 1300-MHz power transistors are an economical solid-state approach to generating moderate power levels (10-20W) at 1269 and 1296 MHz. These amplifiers can be used for terrestrial or satellite work. When OSCAR 10 was designed, it was thought that 10 W into a modest gain antenna (20 dBi) would produce usable signals from the satellite. Unfortunately, there were some problems, and the sensitivity of AO-10 was not as originally expected for the Mode-L uplink. These amplifiers can, however, be used as a driver for a higher-powered tube amplifier for Mode-L service. If all goes according to plan with the launch of AMSAT-OSCAR Phase IIIC, 10 to 20 W with a 20-dBi gain antenna will produce acceptable downlink signals.

The NEL1306 is rated for 6-W output at 1296 MHz at the 1-dB compression point; the NEL 1320 is rated at 20 W. These devices offer several advantages for amateur experimenters. They were designed for collector voltages of 12- to 13.6-V dc, making them ideal for portable and mobile operation. Although the price may seem high to someone familiar with HF parts, these devices are less expensive than most microwave power transistors. The NEL 1306 is in the $26 price range, while the NEL1320 costs about $42. California Eastern Laboratories makes these transistors available in single-lot quantities, so you don't have to be "in the business" to get your hands on them.[1]

The performance of the amplifiers I built and tested is shown in Table 1. The NEL1306 is a good buy. With 1.5-W input, 6- to 8-W output can be achieved. When the amplifier is tuned up at lower power levels, power gain can be as high as 10 dB. With 200-mW drive from my homemade transverter, an output power level of 2 W is attainable.[2]

Power gains as high as 17 dB are possible with a two-stage amplifier (an NEL1306 driving an NEL1320). With a mere 200 mW of drive, 10-W output is possible. When the pair of amplifiers is driven with 1 W and tuned for maximum power output, the 1-dB compression point of 18 W will be achieved.

Circuit Details

The basic design, shown schematically in Fig. 1, is an adaptation of a circuit described in the NEL1300 series data sheet. The design incorporates 30-ohm quarter-wavelength microstriplines on the input and output. C3, C4, C7 and C8, along with L1, form a pi network that matches the low input impedance of the device to 50 ohms. CS, C6, C9 and C10 and the 30-ohm transmission line (L2) form an output pi network that maximizes power transfer to 50 ohms. C10 is not always necessary, depending on variations among devices and circuit-board material.

I designed the amplifiers for 0.031-inch-thick, double-sided glass-epoxy circuit board. A 30-ohm line in this dielectric equates to a line width of 0.121 inch, which is equivalent to the width of the collector and base leads of the NEL1300 series

Table 1

Typical Operating Conditions for the 1296-MHz Solid-State Power Amplifiers

Device	*NE130681-12*	*NEL132081-12*
P_{out}(1-dB compression point)	7W	18W
Gain (1-dB compression point)	6 dB typ.	5 dB typ
Collector efficiency	40-50%	40-50%
Idling current	50 mA	150 mA
I_C @ 1-dB compression point	1.1 A	3.0 A
V_{CC}	13.5 V	13.5 V
Power input	14.9 W	40.5 W

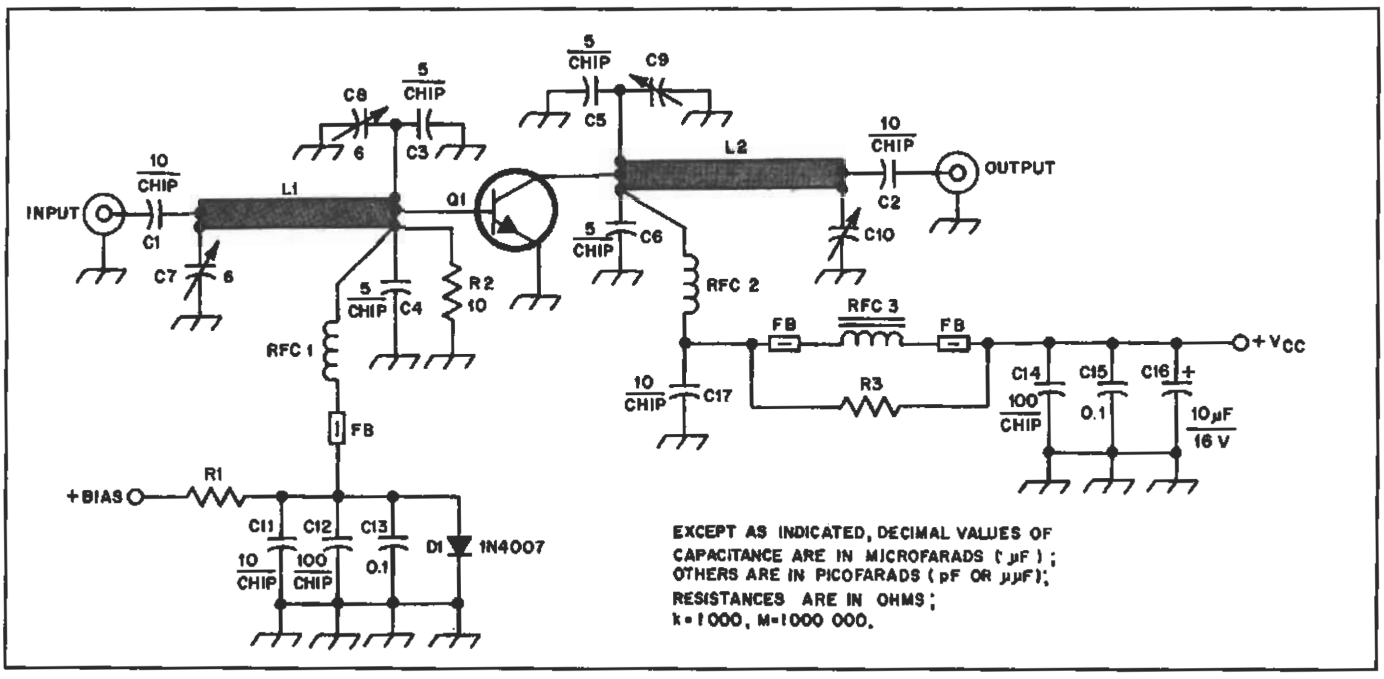
INPUT
10
CHIP
C1
C7
6
L1
C8
6
5
CHIP
C3
Q1
5
CHIP
C4
R2
10
RFC 1
FB
+BIAS
R1
C11
10
CHIP
C12
100
CHIP
C13
0.1
D1
1N4007
5
CHIP
C9
C5
L2
5
CHIP
C6
RFC 2
10
CHIP
C2
C10
OUTPUT
FB
RFC 3
FB
R3
10
CHIP
C17
C14
100
CHIP
C15
0.1
C16
+
10µF
16 V
+Vcc
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.

Fig. 1—Schematic diagram of the NEL1306 and NEL1320 1296-MHz solid-state power amplifiers. The schematic is identical for both versions. Component values are the same except as noted.

C1, C2, C11, C17—10-pF chip capacitor.
C3, C4, C5, C6—3.6- to 5.0-pF chip capacitor.
C7, C8—1.8- to 6.0-pF miniature trimmer capacitor (Mouser 24AA070 or equiv. See text).
C9, C10—Same as C7 and C8 for the NEL1306 amplifier. For the NEL1320 version, 0.8- to 10-pF piston trimmers are used (Johanson 5200 series or equiv.).
C12, C14—100-pF chip capacitor.
C13, C15—0.1-µF disc ceramic capacitor.
C16—10-µF electrolytic capacitor.
D1—1N4007 diode.
L1, L2—30-ohm microstripline, 1/4-wavelength long (see text).
Q1—NEC NEL130681-12 (6 W) or NEL132081-12 (18 W) transistor.
R1—82- to 100-Ω resistor, 2-W minimum. Vary for specified idling current.
R2—10-Ω, 1/4-W carbon-composition resistor with "zero" lead length. See text.
R3—15-Ω, 1-W carbon-composition resistor.
RFC1—3t no. 24 wire, 0.125 inch ID, spaced 1 wire diam.
RFC2—1t no. 24 wire, 0.125 inch ID, spaced 1 wire diam.
RFC3—1-µH RF choke; 18t no. 24 enam. close-spaced on a T50-10 toroid core.

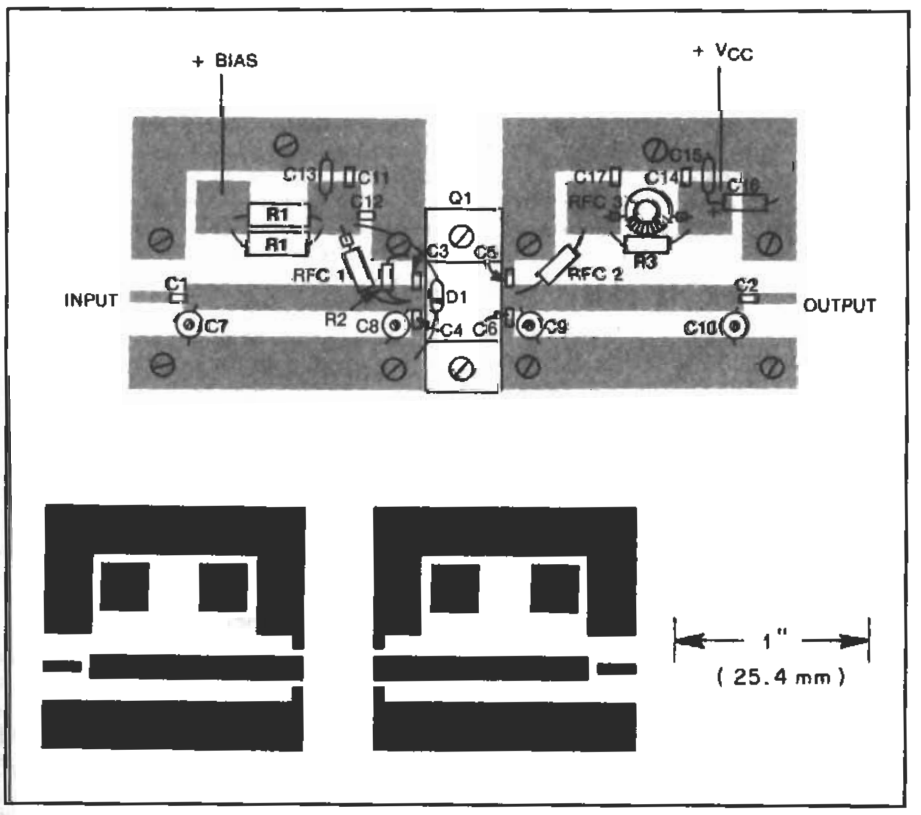
+ BIAS
+ Vcc
C13
C11
C12
R1
R1
RFC 1
R2
C8
Q1
C3
C5
D1
C4
C6
C17
C15
C14
C16
RFC 2
R3
INPUT
C1
C7
OUTPUT
C2
C9
C10
1"
(25.4 mm)

Fig 2—Parts-placement diagram and full-size etching pattern for the solid-state 1296-MHz power amplifiers. All components mount on the etched side of the board. The same PC boards are used for each version.

devices. This minimizes the discontinuity between L1, L2 and Q1.

Bias is provided by R1, R2 and D1. R1 can be optimized, if desired, to adjust the collector idling current.

I selected RFC1 and RFC2 by choosing the lowest possible reactance that will not affect power gain or output power. The RF chokes and the 10-pF bypass capacitors afford adequate decoupling at the frequency of operation. The values of RFC1 and RFC2 are purposely made different to avoid oscillations caused by bias-choke coupling.

After building several of these amplifiers, I noticed that the transistors sometimes generated low-frequency spurious signals. Although these signals were very low in amplitude and caused no problems, they were annoying. I found that I could eliminate them by keeping the high-frequency RF chokes in the collector circuit as small as possible and adding the parallel R3/RFC3 combination, as well as bypass capacitors C14, C15, C16. RFC3 must be capable of handling 1 A for the NEL1306 and 3 A for the NEL1320, so I made special RF chokes to withstand the current.

Construction

Identical construction techniques are used for both amplifiers. Most of the components are mounted to the PC board, and the board and transistor are mounted to an aluminum base plate made from ¼-inch-thick stock. PC-board layout is shown in Fig. 2. Two separate boards are used—one for the input side and one for the output—and they are mirror images of each other. The copper is retained on the bottom side and serves as a ground plane. The grounded areas on the top side must have a good connection to the bottom ground plane for low-inductance grounding of the transistor emitter leads, matching capacitors and bias circuitry. Etched PC boards and partial parts kits for this project are available from A & A Engineering.[3]

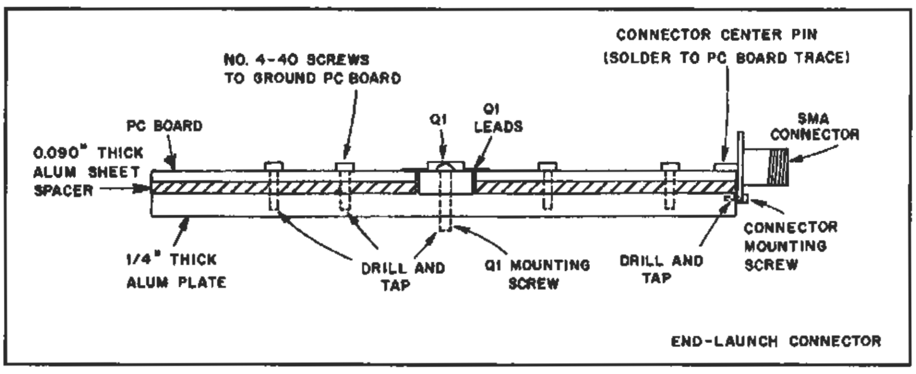

Fig 3—Construction details for the solid-state 1296-MHz power amplifiers. See text for additional information.

Several effective methods of connecting the top and bottom ground planes are summarized here.

1) Plated-through holes at the critical grounding areas mentioned earlier.

2) Use of pins or screws that penetrate through the circuit board into the aluminum base plate at the critical areas.

3) Use of "wrap-around" foils on all edges of the ground plane.

Plated-through holes are often used in the commercial and military electronic marketplace but are not so easily reproduced in the average builder's circuitboard shop. I've found that the best technique for the home builder is a combination of methods 2 and 3. First, wrap thin copper or brass foil around the board edges and then solder the foil to the top and bottom. Sometimes called "shim stock," thin sheet metal is often available from hobby shops or metal suppliers. Next, drill holes through the board and use no. 4-40 screws to tie the circuit board to the base plate at the critical areas shown in Fig. 2.

The transistor must be mounted so that its leads lie flat against the PC board. This poses a slight problem, since the transistor leads protrude from the device 0.165 inch above the bottom of the flange, and the PC boards are only 0.031 inch high. There are a couple of ways around this problem. They work equally well, so choose the method that is easiest for you.

If you have access to a machine shop, you can mill out a 0.240-inch-wide by 0.090-inch-deep slot in the aluminum base plate to clear the transistor flange. This is a simple task on a milling machine, and you can probably find a local machine shop that will do small jobs when business is slow. The other method is to mount the transistor to the base plate and use a piece of 0.090-inch-thick aluminum sheet to bring the input and output PC boards up to the right height. The 0.090-inch dimension allows a wrap-around foil and solder buildup of 0.040 inch, maximum.

I used 1-inch-long screws to secure the transistor to holes drilled and tapped in the aluminum base plate. The extra screw length that protrudes from the bottom of the base plate allows the use of an external heat sink, which is suggested if continuous operation is desired.

The clearance holes in the NEL1300 devices are for no. 4-40 hardware. I drilled out the holes to accept no. 6-32 screws to make the assembly more rugged, although this may not be necessary. Be careful if you decide to drill out the holes to accept no. 6-32 hardware; the transistor flange is soft copper, and you could damage the device. Use a small vise to hold the transistor flange during the drilling operation.

Solder the components to pads on the board using surface-mounting techniques. Silver solder (2%) is recommended for the chip capacitors, but 5N63 will work fine. Use a 15-W iron and solder quickly to avoid burning the metallization off the capacitors.

The transistor leads should be soldered in place only after the circuit boards and transistor have been firmly bolted down to the base plate. This is necessary to minimize any buildup of stress in the transistor leads. Chip capacitors C3, C4, C5 and C6 should be soldered directly onto the leads of Q1 to ensure the shortest possible lead length. Keep D1 close to Q1. Thermal compound will enhance heat transfer to D1 to ensure minimal drift in idling current with temperature changes.

In the original design of these amplifiers, Johanson piston-trimmer capacitors (5200 series or equiv.) were used for C7, C8, C9 and C10. These capacitors are fairly large, and in some instances the coupling between the bodies of C8 and C9 was enough to cause an in-band oscillation. Smaller variable capacitors should be used if at all possible. I've used a 1.8- to 6.0-pF miniature ceramic trimmer capacitor, Mouser Electronics part number 24AA070, in the input and output networks of the NEL1306 amplifier and in the input network of the NEL1320 amplifier with no performance degradation.[4] I did use the Johanson piston trimmers in the output network of the NEL1320 amplifier because of the high RF currents involved.

The PC board makes use of end-launch SMA-type connectors. Fig. 3 illustrates this technique. Four-hole, flange-mount SMA connectors can be mounted to the edge of the base plate using two of the four mounting holes. Drill and tap the base plate for no. 2-56 hardware. Be careful—it's easy to accidentally cross-thread orover-torque the no. 2-56 hardware.

An alternative approach is to mount the amplifier in an aluminum die-cast box (Bud CU-124B or Hammond 1590B) and run miniature 50-ohm coaxial cable such as RG-174 from the amplifier board to the connector. The amplifiers shown in Figs. 4 and 5 use standard SMA connectors mounted to the walls of the metal box. BNC or Type-N connectors should work equally as well. When preparing each end of the coaxial cable, try to keep the pigtail leads as short as possible (⅛ inch or less); otherwise the mismatch will be difficult to tune out.

I compared the performance of an amplifier with end-launch connectors to that of another that used the approach just described. I could measure no difference in gain or 1-dB compression point.

Considerable effort was put forth to make sure the amplifiers are stable. The devices have fairly high gain at the frequency of operation, so layout and good construction practices are very important. Here are some construction hints that can help ensure amplifier performance and stability.

1) Use the smallest (physical) size variable capacitors that will still handle the RF current.

2) Use wrap-around ground foils as noted. Grounding screws are required at the critical RF-ground areas near the shunt variable capacitors, shunt bypass capacitors and Q1 emitters.

3) Connect braids from the coaxial-cable jumpers to the same ground as the shunt variable capacitors.

4) Use as little lead length as possible on R2—less than ⅛ inch.

5) In some instances when the large piston trimmers are used, a shield approximately ¾ inch high mounted on top of Q1 and grounded via the mounting screws can improve isolation between C8 and C9.

Tune-up and Operation

Measuring RF power at 1296 MHz can be difficult. I used a calibrated 20-dB directional coupler along with enough attenuator pads to allow power to be read with a Hewlett Packard HP430C power meter and an HP477B thermistor mount. Even better is the HP431 power meter with its associated HP478A thermistor mount (a newer version of the HP430C that does not

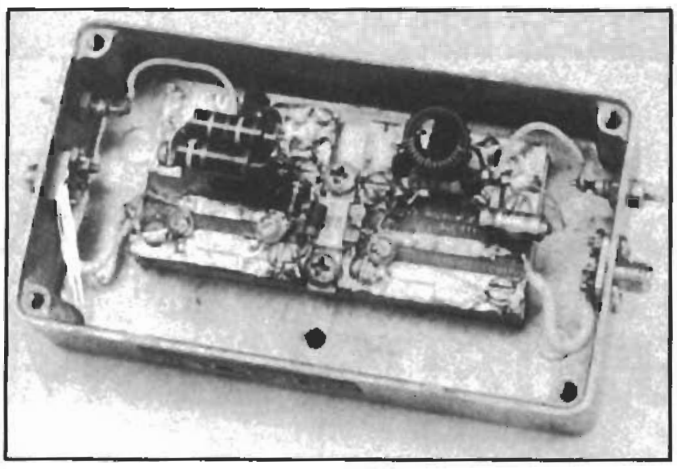

Fig 4—This NEL1306 amplifier was built inside a die-cast box. Miniature coaxial cable runs to the connectors.

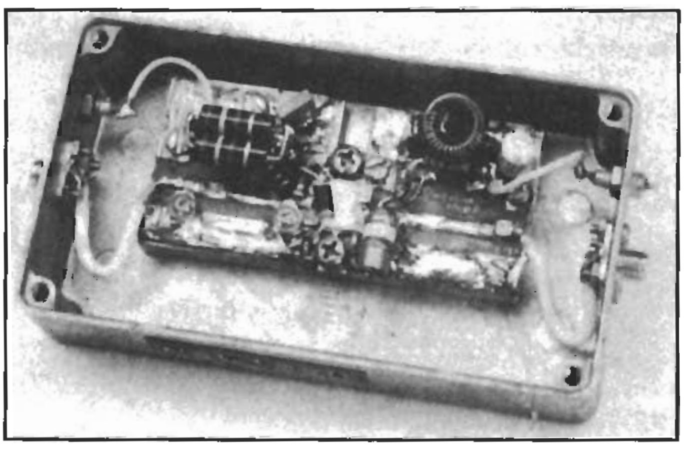

Fig 5—The NEL1320 amplifier is virtually identical to the NEL1306 version of Fig 4, except for the capacitors on the output strip line. C10 was not necessary on this version.

suffer from temperature-drift problems). Bird offers several low-power elements for the popular model 43 in this frequency range. Bird 400-1000 MHz elements are common, and they can be used with decreased accuracy.

Begin initial setup of each amplifier by terminating the input and output in good 50-ohm loads. I recommend that you use a fuse in the collector lead of Q1 to protect the device until you are sure everything is working normally. Start with all capacitors at minimum. Apply 12- to 13.5-V dc to the V_{CC} and bias terminals. The collector idling current should be as shown in Table 1. Vary the value of R1 for correct idling current.

For the NEL1306 amplifier, start out with 50 to 100 mW of drive. Adjust the output network for maximum power output and then peak the input network for maximum output power. Increase drive and repeak both matching networks for rated performance as shown in Table 1. Similarly, start out with approximately 1 W of drive for the NEL1320 and follow the same procedure. After a minute or two of operation at maximum power output, remove RF drive power and check to see that the collector idling current has not increased more than 25% over the initial setting. Keeping D1 in close contact with Q1 will minimize drift in idling current with temperature changes.

If you're going to use transistor switching to apply dc to the power amplifier stages during transmit, consider the following technique. Apply 13.5-V dc to the V_{CC} terminal during receive and transmit. Use a series transistor switch to apply 13.5-V dc to the bias terminal during transmit. A power transistor capable of carrying only a few hundred milliamperes of bias current, as opposed to several amperes of collector current, will be required. More important, the voltage drop across the transistor switch in the V_{CC} line will be eliminated. This will ensure maximum power output of the NEL1300 devices by keeping V_{CC} at 13.5-V dc.

Switching the bias port off during receive is important for another reason. Normally during receive periods, the amplifier is left unterminated. Sometimes the input port is also left open. Depending on the length of the unterminated 50-ohm cable on the amplifier ports, the unit may show signs of instability if it is drawing idling current.

Several amateurs in the Dallas area have duplicated these amplifiers with no problems. Other amplifiers of this design are in use in different parts of the country. The NEL1300 series amplifiers offer a simple and inexpensive means of generating medium power on 1296 MHz. You'll be amazed at what you can work with 18 W that you can't with 1 W.

I wish to thank everyone who offered technical advice, especially Wes Atchison, WA5TKU, for helping with the construction and evaluation of the prototype amplifiers.

Notes

[1] NEC transistors are available from California Eastern Laboratories, 3260 Jay St., Santa Clara, CA 95050, tel. 408-988-3500.

[2] Complete construction details for this transverter may be found in Chapter 32 of the *1986 ARRL Handbook*.

[3] A & A Engineering, 7970 Orchid Dr., Buena Park, CA 90620, tel. 714-521-4160.

[4] Mouser Electronics, 11433 Woodside Ave., Santee, CA 92071, tel. 619-449-2222.

By Ward Silver, N0AX **From *QST*, September 2003**

Amplifier Care and Maintenance

With few moving parts, your amplifier can be easy to overlook. Here are some ideas for taking care of amplifiers.

Amplifiers come in all shapes and sizes, large and small, light or heavy, tube or solid-state, VHF or HF. Regardless, they all need a little TLC from time to time. They can cost as much as top-of-the-line radios, so it's important that they get a little maintenance on a regular basis.

While this article focuses on amplifiers that use vacuum tubes, many of the ideas presented here can and should be applied to any amateur amplifier—HF or VHF/UHF. Solid-state amplifiers operate at lower voltages and generally have fewer points of failure, but they still need occasional maintenance.

Safety First

It is important to review good safety practices.[1] Tube amplifiers use power supply voltages well in excess of 1 kV and the RF output at full throttle can be hundreds of volts, as well. Almost every voltage in an amplifier can be lethal! Take care of yourself and use caution!

- *Power Control*—Know and control the state of both ac line voltage and dc power supplies. Physically disconnect line cords and other power cables when you are not working on live equipment. Use a lockout on circuit breakers. Double-check visually and with a meter to be absolutely sure power has been removed.
- *Interlocks*—Unless specifically instructed by the manufacturer's procedures to do so, never bypass or "rig" an interlock. This is rarely required except in troubleshooting and should only be done when absolutely necessary. Interlocks are there to protect you.
- *The One-Hand Rule*—Keep one hand in your pocket while making any measurements on live equipment. The hand in your pocket won't give current a chance to flow through you. It's also a good idea to wear shoes with insulating soles and work on dry surfaces. Current can be lethal even at milliampere levels—don't tempt the laws of physics.
- *Patience*—Repairing an amplifier isn't a race. Take your time. Don't work on equipment when you're tired or frustrated. Wait several minutes after turning the amplifier off to open the cabinet—capacitors can take several minutes to discharge through their bleeder resistors.
- *A Chicken Stick*—Make this simple safety accessory shown in Figure 1 and use it whenever you work on equipment in which hazardous voltages have been present. The ground wire should be heavy duty (12 gauge minimum) due to the high peak currents (hundreds of amperes) present when discharging a capacitor or tripping a circuit breaker. When equipment is opened, touch the tip of the stick to *every* exposed component and connection that you might come in contact with. Assume nothing—accidental shorts and component failures can put voltage in places it shouldn't be.
- *The Buddy System and CPR*—It's always a good idea to use the buddy system when working around any equipment that has the potential for causing serious injury. The buddy needn't be a ham, just anyone who can be nearby in case of trouble. Your buddy should know how to remove power and administer basic first aid. Since hams work around electrical equipment frequently, it would be a good idea to have your buddy or someone in the household know CPR, as well.[2]

Cleanliness

The first rule of taking good care of an amplifier is cleanliness. I realize that 90 percent of ham shacks have just failed the first rule. Amplifiers need not be kept sparkling new, but their worst enemy is heat. Excess heat accelerates component aging and stresses those expensive tubes and transformers. There are two areas to keep clean—the inside and the outside.

Outside the amplifier, you need to prevent dust and obstructions from blocking the paths by which heat is removed. This means keeping all ventilation holes free of the ever-present dust bunnies, pet hair and

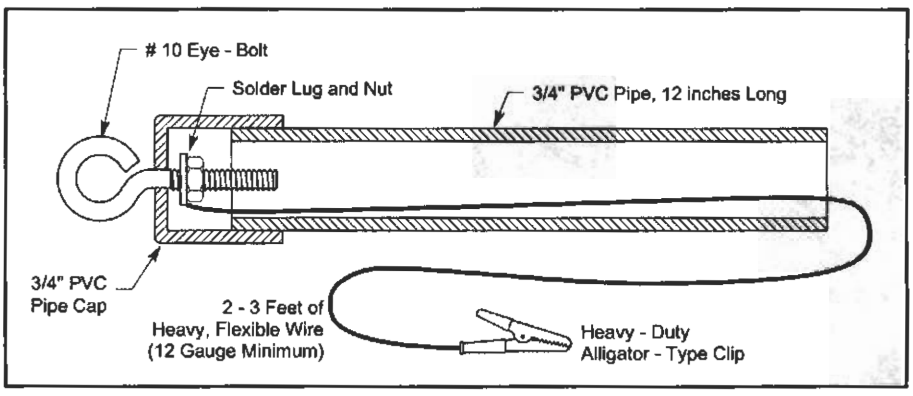
# 10 Eye - Bolt
Solder Lug and Nut
3/4" PVC Pipe, 12 inches Long
3/4" PVC Pipe Cap
2 - 3 Feet of Heavy, Flexible Wire (12 Gauge Minimum)
Heavy - Duty Alligator - Type Clip

Figure 1—The "chicken stick" is a great way to ensure that everything inside the amplifier that should be discharged actually is. It can be a life saver.

insects. Fan intakes are particularly susceptible to inhaling all sorts of "goop." Get out the vacuum cleaner and clean not only the amplifier, but the surrounding areas. Don't even think about letting liquids anywhere near the amplifier. One spilled cup of coffee can cause hundreds of dollars of damage.

Keep papers or magazines off the amplifier—even if the cover is solid metal. Paper acts as an insulator and keeps heat from being radiated through the cover. Amplifier heat sinks must have free air circulation to be effective. There should be at least a couple of inches of free space surrounding an amplifier on its sides and top. If the manufacturer recommends a certain clearance, mounting orientation or air flow, follow those recommendations.

Just as the outside needs to be kept clean, so does the inside. High voltage (HV) circuits attract dust like crazy. The dust slows heat dissipation and will eventually build up to the point where it arcs or carbonizes. Our friend the vacuum cleaner should make another appearance to remove any dust or dirt. If you find insects (or worse) inside the amp, try to determine how they got in and plug that hole. Window screening works fine to allow airflow while keeping out visitors. While you're cleaning the inside, this is a great time to perform a visual inspection as described in the next section.

Vacuuming works best with an attachment commonly known as a "crevice cleaner." Figure 2 shows a crevice cleaning attachment being used with a small paintbrush to dislodge and remove dust. Don't use the vacuum cleaner brush attachment; they're designed for floors, not electronics. Some vacuums also have a blower mechanism, but these rarely have enough punch to clean as thoroughly as a brush. Besides, that dust you blow all over is going to wind up in some other equipment, so it's best to take it out of circulation, so to speak. The brush will root dust out of tight places and off components without damaging them or pulling on connecting wires.

If you can't get a brush or attachment close enough, a spray can of compressed air will usually dislodge dust and dirt. If you use a rag or towel to wipe down panels or large components, be sure not to leave threads or lint behind. Never use a solvent or spray cleaner to wash down components unless the manufacturer advises doing so—you might leave behind a residue or damage the component.

Visual Inspection

Once the amplifier has been cleaned, it's time for a visual inspection. Remove any internal covers or access panels and...*stop!* Get out the chicken stick, clip its ground lead securely to the chassis and touch every exposed connection. Now, using a strong light and possibly a magnifier, look over the components and connections. Amplifiers have far fewer components than transceivers do, so it's quite feasible to look at every component and insulator. Look for cracks, signs of arcing, carbon traces (thin black lines), discoloration, loose connections, melting of plastic, and anything else that doesn't "look right." This is a great time to be sure that mounting and grounding screws are tight. While you're in there, does anything smell burnt? The nose can quickly detect the odors of toasted transformer, cooked capacitor or roasted resistor. Learn the smells of healthy and not-so-healthy components.

Make a note of what you find, repair, or replace—even if it's absolutely nothing. If you don't keep a shack notebook, start one. A simple spiral notebook with notes about maintenance, wiring, color coding, antenna behavior and so forth can be a big timesaver.

S. COHEN, N1SC; R. SCHETGEN, KU7G

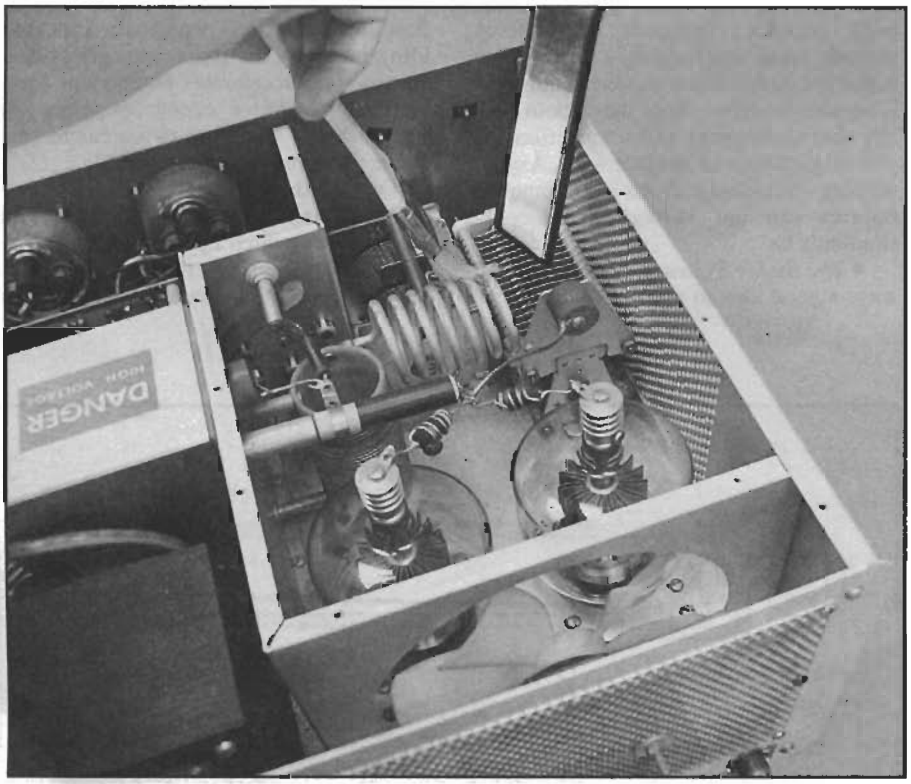

Figure 2—A small paintbrush and a vacuum cleaner crevice attachment make dust removal easy.

Electrical Components

An amplifier contains many heavy-duty HV and RF components. These can be expensive and hard to replace, so it's important that you take good care of them. Let's start with the power supply.

There are three basic parts to amplifier power supplies—the ac transformer and line devices, the rectifier/filter and the metering/regulation circuitry. Transformers need little maintenance except to be kept cool and be mounted securely. Line components such as switches, circuit breakers and fuses, if mechanically sound and adequately rated, are usually electrically okay, as well.

Rectifiers and the capacitors that filter the HV dc require occasional cleaning. Look for discoloration around components mounted on a printed circuit board (PCB) and make sure that all wire connections are secure. HV capacitors are generally electrolytic or oil and should show no signs of leakage, swelling or outgassing around terminals.

Located at the output of the filter, components that perform metering and regulation of voltage and current can be affected by heat or heavy dust. If there has been a failure of some other component in the amplifier—such as a tube—these circuits can be stressed severely. Resistors may survive substantial temporary overloads, but may show signs of overload, such as discoloration or swelling.

Amplifiers contain two types of relays—control and RF. Control relays switch ac and dc voltages and do not handle input or output RF energy. The usual problem encountered with control relays is oxidation or pitting of their contacts. A burnishing tool can be used to clean relay contacts. In a pinch a strip of ordinary paper can be pulled between contacts gently held closed. [Avoid the temptation to over-clean silver-plated relay and switch contacts, as the author points out later. It is easy to remove contact plating with excessive polishing and while silver-plated relay and switch contacts may appear to be dark in color, oxidized silver (black) is still a good conductor. Once the silver's gone, it's gone; contact erosion will then be pervasive.—*Ed.*] If visual inspection shows heavy pitting or discoloration or resistance measurements show the relay to have intermittent contact quality, it is best replaced.

RF relays are used to perform transmit-receive (TR) switching and routing of RF signals through or around the amplifier circuitry. Amplifiers designed for full break-in operation will usually use a high-speed vacuum TR relay. Vacuum relays are sealed and cannot be cleaned or maintained. When you replace RF relays, use a direct replacement part or one rated for RF service with the same characteristics as the original.

Cables and connectors are subjected to heavy heat and electrical loads in amplifiers. Plastics may become brittle and connections may oxidize. Cables should remain flexible and not be crimped or pinched under clamps or tie-downs. It's a good idea to gently wiggle cables while watching the connections at each end for looseness or bending. Connectors can be unplugged and reseated once or twice to clear oxide on contact surfaces. Carefully inspect any connector that seems loose. Be especially careful with connectors and cables in amplifiers that have RF decks that are in separate enclosures from those of their power supplies. Those interconnects are susceptible to both mechanical and electrical stress and you don't want an energized HV cable loose on the operating desk. Check both the soldered electrical integrity and the mechanical stability of those cables and make sure they are tightly fastened.

As with relays, switches found in amplifiers are either control function orientated or RF routers. Adequately rated control switches, if mechanically sound, are usually okay. Bandswitches are the most common RF switch—usually a rotary phenolic or ceramic type. A close visual inspection should show no pitting or oxidation on the wiper (the part of the switch that rotates between contacts) or the individual contacts. Arcing or overheating will quickly destroy rotary switches. Figure 3 is a photo of a heavy-duty band switch that has suffered severe damage from arcing. Slight oxidation is acceptable on silver-plated switches. Phosphor-bronze contacts can sometimes be cleaned with a light scrub from a pink pencil eraser, but plating can be easily removed, so use caution with this method and be sure to remove any eraser crumbs. Rotary switch contacts cannot be replaced easily although individual wafer sections may be replaced if an exact matching part can be obtained.

Amplifiers use all types of capacitors and resistors. When replacing them, be sure to use a part rated for the use to which it will be put. Voltage and power-handling ratings are particularly important, especially of those handling high RF currents. An RF tank capacitor replacement should be checked carefully for adequate RF voltage and current ratings, not just dc. HV resistors are generally long and thin to prevent arcing across their surfaces. Even if a smaller (and cheaper) resistor has an equivalent power rating, resist the temptation to substitute it. In a pinch, a series string of resistors of the appropriate combined value can be used to replace one HV unit. Don't use carbon resistors for metering circuits, use metal or carbon film types. The carbon composition types are too unstable.

If you are repairing or maintaining an old amplifier and manufacturer-specific parts are no longer available, the ham community has many sources for RF and HV components. Fair Radio Sales and Surplus Sales of Nebraska are familiar names.[3] Hamfests and Web sites such as **www.eham.net** or **www.k1dwu.net/hamtrader** often have amplifier components for sale. You might consider buying another amplifier of the same type in non-working condition for parts use.

Tubes

The single most expensive component in an amplifier is usually the vacuum tube that performs the amplification. Good maintenance of tubes starts with proper operation of the amplifier. Follow the manufacturer's instructions for input drive levels, duty cycles, tuning and output power level. Frequently check all metered voltages and current to be sure that the tubes are being operated properly and giving you maximum lifetime. Penta Labs has an excellent Web page on maintaining power tubes.[4]

The internal mechanical structures of tubes generally do not deal well with mechanical shock and vibration, so be gentle. The manufacturer may also specify how the amplifier is to be mounted, so read the operating manual.

Tubes generate a lot of heat, so it's important that whatever cooling mechanism employed is kept at peak efficiency. Airways should be clean, including between the fins on metal tubes. All seals and chimneys should fit securely and be kept clean. Wipe the envelope of glass tubes clean after handling them—fingerprints should be removed to prevent baking them into the surface.

On metal tubes that use finger-stock contacts, be sure the contacts are clean and make good contact all the way around the tube. Partial contact or dirty finger stock can cause asymmetric current and heating inside the tube, resulting in warping of internal grids and possibly causing harmonics or parasitics.

Plate cap connections and VHF parasitic suppressors should be secure and show no signs of heating. Overheated parasitic suppressors may indicate that the neutralization circuit is not adjusted properly. Inspect socket contacts and the tube pins to be sure all connections are secure, particularly high-current filament connections. Removing and inserting the tubes once or twice will clean the socket contacts.

Adjustments to the neutralizing network, which suppresses VHF oscillations by negative feedback from the plate to grid circuit, are rarely required except when you are replacing a tube or after you do major rewiring or repair of the RF components. The manufacturer will provide instructions on making these adjustments. If symptoms of VHF oscillations occur without changing a tube, then perhaps the tube characteristics or associated components have changed. Parasitic oscillations in high-

H. WARD SILVER, N0AX

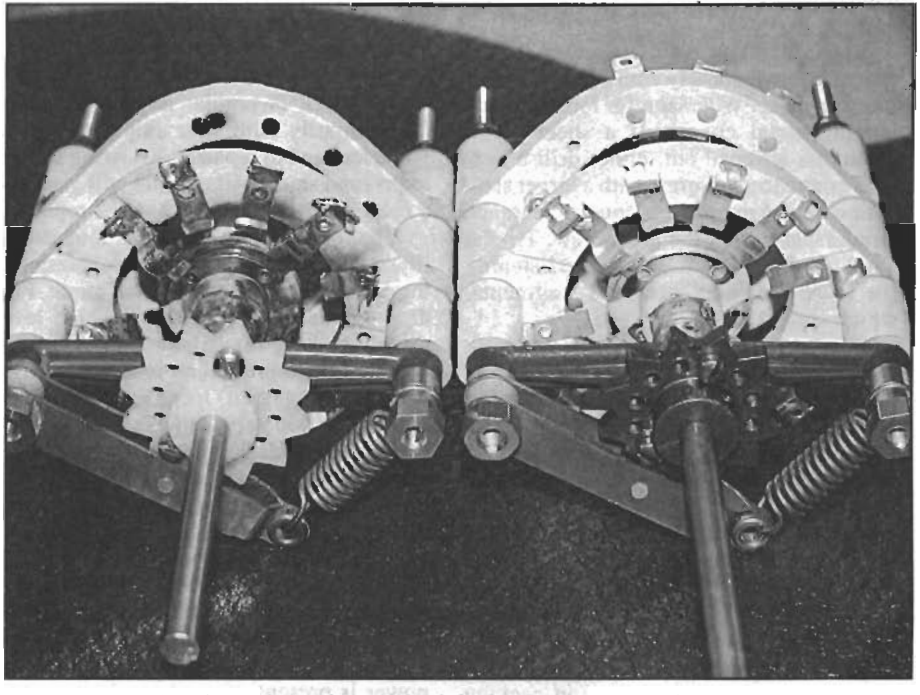

Figure 3—The band switch section on the left clearly shows the signs of destructive arcing.

power amplifiers can be strong enough to cause arcing damage. Perform a visual inspection prior to re-adjusting the neutralizing circuit.

Metering circuits rarely fail, but they play a key part in maintenance. By keeping a record of "normal" voltages and currents, you will have a valuable set of clues when things go wrong. This is perfect information for the shack notebook. Record tuning settings, drive levels, and tube voltages and currents on each band and with every antenna. When things change, you can refer back to the notebook instead of relying on memory.

Mechanical

While the amplifier is primarily an electronic beast, it has a significant number of mechanical parts that affect its well-being. Thermal cycling and heat-related stresses can result in mechanical connections loosening over time or material failures.

Switch shafts, shaft couplings and panel bearings all need to be checked for tightness and proper alignment. All mounting hardware needs to be tight, particularly if it supplies a grounding path. Examine all panel-mounted components, particularly RF connectors, and be sure they're attached securely. BNC and UHF connectors that are mounted with a single nut in a round panel hole are notorious for loosening with repeated connect/disconnections.

Rubber and plastic parts are particularly stressed by heat. If there are any belts, gears or pulleys, make sure they're clean and that dust and lint are kept out of their lubricant. Loose or slipping belts should be replaced. Check O-rings, grommets and sleeves to be sure they are not brittle or cracked. If insulation sleeves or sheets are used, check to be sure they are covering what they're supposed to. Never discard them or replace them with improperly sized or rated materials.

Enclosures and internal shields should all be fastened securely with every required screw in place. Watch out for loosely overlapping metal covers. If a sheet metal screw has stripped out, either drill a new hole or replace the screw with a larger size, taking care to maintain adequate clearance around and behind the new screw. Tip the amplifier from side to side while listening for loose hardware or metal fragments, which should all be retrieved.

A great time to clean the front and back panels and get gummy finger deposits off before they cause permanent finish damage is during maintenance. If the amplifier is missing a foot on the cabinet or an internal shock mount, replace it. A clean unit with a complete cabinet will have a significantly higher resale value than a dirty, grubby one, so it's in your interest to keep the equipment looking its best.

Shipping

When you are traveling with an amplifier or shipping it, some care in packing will prevent needless damage. Improper packing can also result in difficulty in collecting on an insurance claim, should damage occur. The original shipping cartons are a good method of protecting the amplifier for storage and sale, but they were not made to hold up to frequent shipping. If you travel frequently, it is best to get a sturdy shipping case made for electronic equipment.[5]

Some amplifiers require the power transformer to be removed before shipping. Check your owner's manual or contact the manufacturer to find out. Failure to remove it before shipping can cause major structural damage to the amplifier's chassis and case.

Tubes should also be removed from their sockets for shipment. It may not be necessary to ship them separately if they can be packed in the amplifier's enclosure with adequate plastic foam packing material. If the manufacturer of the tube or amplifier recommends separate shipment, however, do it!

Cleaning and Maintenance Plan

This discussion should have given you plenty to think about. It's easy to defer maintenance, but as with a vehicle, performance and lifetime are improved if a regular program is put into place. For amateur use, there is little need for maintenance more frequently than once per year. If there is a period of the year in which you are most active, put a note on the calendar about six weeks in advance to "open the hood," giving you time to obtain and replace any components.

Consider the maintenance requirements of your amplifier and what its manufacturer recommends. Sit down with your amplifier's manuals and make up a checklist of what major steps and tools are required. When maintenance time rolls around, you'll be prepared and be able to perform the job in the most efficient manner.

Troubleshooting

A benefit of regular maintenance will be familiarity with your amplifier should you ever need to repair it. Knowing what it looks (and smells) like inside will give you a head start on effecting a quick repair.

The following discussion is intended to illustrate the general flow of a troubleshooting effort, not be a step-by-step guide. Figure 4 shows a moderately-high-level troubleshooting flow chart. Before starting on your own amplifier, review the amplifier manual's "Theory of Operation" section and familiarize yourself with the schematic. If there is a troubleshooting procedure in the manual, follow it, of course.

You might be surprised how many "amplifier is dead" problems turn out to be simply a lack of ac power. Before even opening the cabinet of an unresponsive amplifier, be sure that ac is really present at the wall socket and that the fuse or circuit breaker is really closed. Assuming that ac power is present, trace through any internal fuses, interlocks and relays all the way through to the transformer primary terminals.

Hard failures in a high voltage power supply are rarely subtle, so it's usually clear if there is a problem and what components are involved. When you repair a power supply, take the opportunity to check all related components. If all defective components are not replaced, the failures may be repeated when the circuit is re-energized.

Rectifiers may fail open or shorted—test them using a DVM diode checker. An open rectifier will result in a drop in the HV output of 50 percent or more but will probably not overheat or destroy itself. A shorted rectifier failure is usually more dramatic and may cause additional rectifiers or filter capacitors to fail. If one rectifier in a string has failed, it may be a good idea to replace the entire string as the remaining rectifiers have been subjected to a higher-than-normal voltage.

High voltage filter capacitors usually fail shorted, although they will occasionally lose capacitance and show a rise in ESR (equivalent series resistance). Check the rectifiers and any metering components—they may have been damaged by the current surge caused by a shorted filter. Power transformer failures usually manifest themselves by insulation failures with consequent arcing of the windings. Either can result in the unmistakable aroma of overheated transformer. A failed transformer is generally not repairable.

Along with the HV plate supply, tetrode screen supplies occasionally fail, too. The usual cause is the regulation circuit that drops the voltage from the plate level. Operating without a screen supply can be damaging to a tube, so be sure to check the tube carefully after repairs.

If the power supply checks out okay and the tube's filaments are lit, check the resting or bias current. If it is excessive or very low, check all bias voltages and dc current paths to the tube, such as the plate choke, screen supply (for tetrodes) and grid or cathode circuits.

Having exhausted power supply and dc problems, you will then turn to the RF components or "RF deck." There is a natural tendency to forge ahead and swap in a known-good tube or tubes. Don't! Tubes are expensive and if the problem is elsewhere, you may damage the spares. Wait to swap tubes until you are sure that the tube is likely to be defective.

Check the input SWR to the amplifier. If it has changed (you did write down the normal SWR and drive levels, didn't you?) then you likely have a problem in the input circuitry or one or more tubes have failed. Perform a visual check of the input circuitry and the band switch, followed by an ohmmeter check of all input components.

If input SWR is normal and applying drive does not result in any change in plate current, you may have a defective tube, tube socket, or connection between the input circuits and the tube. Check the TR

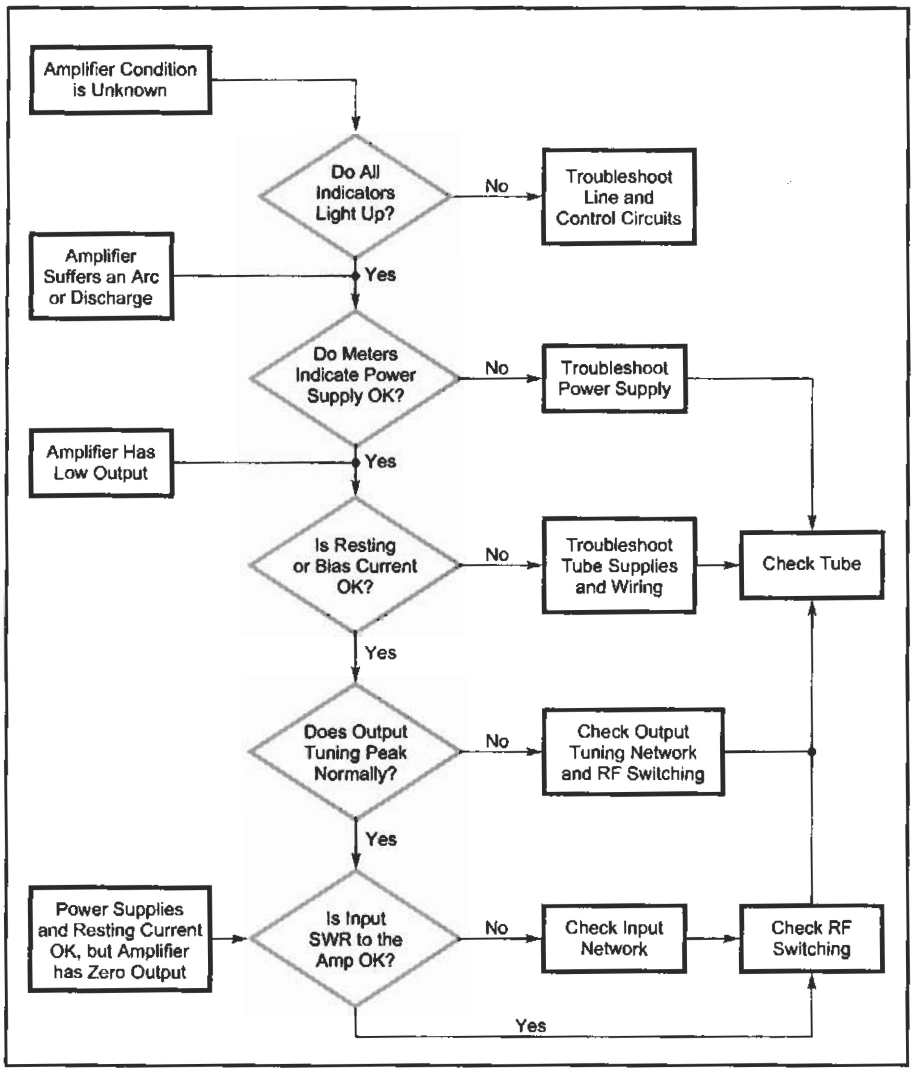
Amplifier Condition is Unknown
Do All Indicators Light Up?
No
Troubleshoot Line and Control Circuits
Amplifier Suffers an Arc or Discharge
Yes
Do Meters Indicate Power Supply OK?
No
Troubleshoot Power Supply
Amplifier Has Low Output
Yes
Is Resting or Bias Current OK?
No
Troubleshoot Tube Supplies and Wiring
Check Tube
Yes
Does Output Tuning Peak Normally?
No
Check Output Tuning Network and RF Switching
Yes
Power Supplies and Resting Current OK, but Amplifier has Zero Output
Is Input SWR to the Amp OK?
No
Check Input Network
Check RF Switching
Yes

Figure 4—This moderate-level flow chart is a good way to identify amplifier problems quickly in lieu of a misplaced or nonexistent manufacturer's troubleshooting procedure.

control circuits and relay. If plate current changes, but not as much as normal, try adjusting the output tuning circuitry. If this has little or no effect, the tube may be defective or a connection between the tube and output circuitry may have opened. If retuning has an effect, but at different settings than usual, the tube may be defective or there may be a problem in the tuning circuitry. A visual inspection and an ohmmeter check are in order.

The key to finding the trouble with your amplifier is to be careful and methodical, and to avoid jumping to false conclusions or making random tests. The manufac-turer's customer service department will likely be helpful if you are considerate and have taken careful notes detailing the trouble symptoms and any differences from normal operation. There may be helpful guidelines on the manufacturer's Web pages or from other Internet resources. Sometimes there is more than one problem—they work together to act like one very strange puzzle. Just remember that most problems are very simple and can be isolated by careful, step-by-step tests.

Summary

Amplifiers have been part of ham radio for many years. They are simple, reliable pieces of equipment that respond well to basic care and common sense. Take the time to know your amp—inside and out. If you take care of it, it will reward you with reliable service and maximum tube lifetime.

Notes

[1] Chapter 9 of the current *ARRL Handbook for Radio Communications* is an excellent source of safety information. Available from your local dealer or the ARRL Bookstore. Order no. 1921 (softcover), no. 1948 (hardcover). Telephone toll-free in the US 888-277-5289, or 860-594-0355, fax 860-594-0303; **www.arrl.org/shop/**; **pubsales@arrl.org**.

[2] Instructions for CPR can be found at: **about-the-web.com/spiritworks/web/Kaiser/html/adult.htm**.

[3] Two sources of HV and RF parts include Surplus Sales of Nebraska (**www.surplussales.com**) and Fair Radio Sales (**www.fairradio.com**). Others can be found at the ARRL Technical Information Service database (**www.arrl.org/tis/tisfind.html**).

[4] Penta Labs, "Tube Maintenance & Education" (**www.pentalaboratories.com/maintenance.asp**).

[5] Pelican (**www.pelican-shipping-cases.com**) and Anvil (**www.anvilsite.com**) make excellent shipping cases suitable for carrying amplifiers and radio equipment.

Build safe and reliable amplifiers. Produce loud and clean signals.

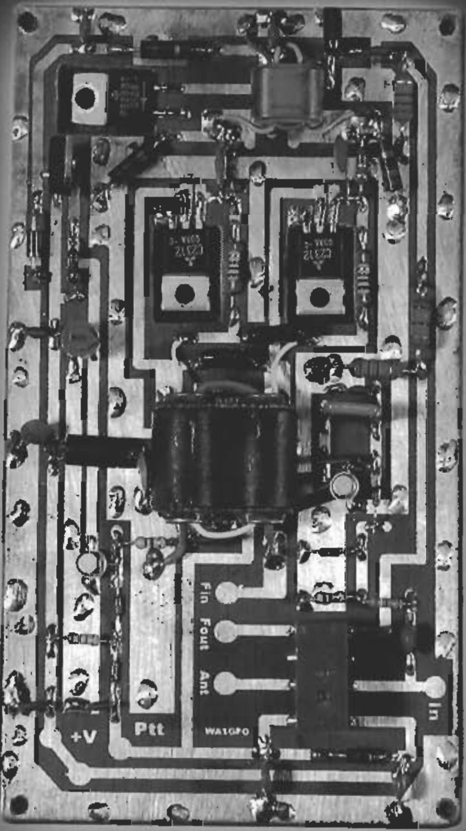

Planning to build an amplifier compares well to the work of a cook searching for a recipe and all the ingredients. Find the right design, and then locate the parts and components to begin building.

Shorten your discovery work with *ARRL's RF Amplifier Classics*! Use this book to find classic tube and solid-state designs at power levels from 5 W to 1.5 kW. Included are two-dozen amplifier projects and articles published in ARRL's popular journals between 1980 and 2003. There are amps for HF, MF, VHF and microwave. These are high quality works from respected authors such as Gary Breed, K9AY; Jerry Pittenger, K8RA; Bill Sabin, W0IYH; Al Ward, W5LUA; Dave Meacham, W6EMD and others.

Turn dreams of constructing your first amp or next brick into reality!

ISBN 0-87259-931-0
51995 >
EAN
9 780872 599314
ISBN 0-87259-931-0 ARRL Order No. 9310

ARRL *The national association for* AMATEUR RADIO
225 Main Street,
Newington, CT 06111-1494 USA
ARRLWeb: www.arrl.org/

The ARRL Guide to
Antenna Tuners

A Radio Amateur's Guide to Antenna Matching

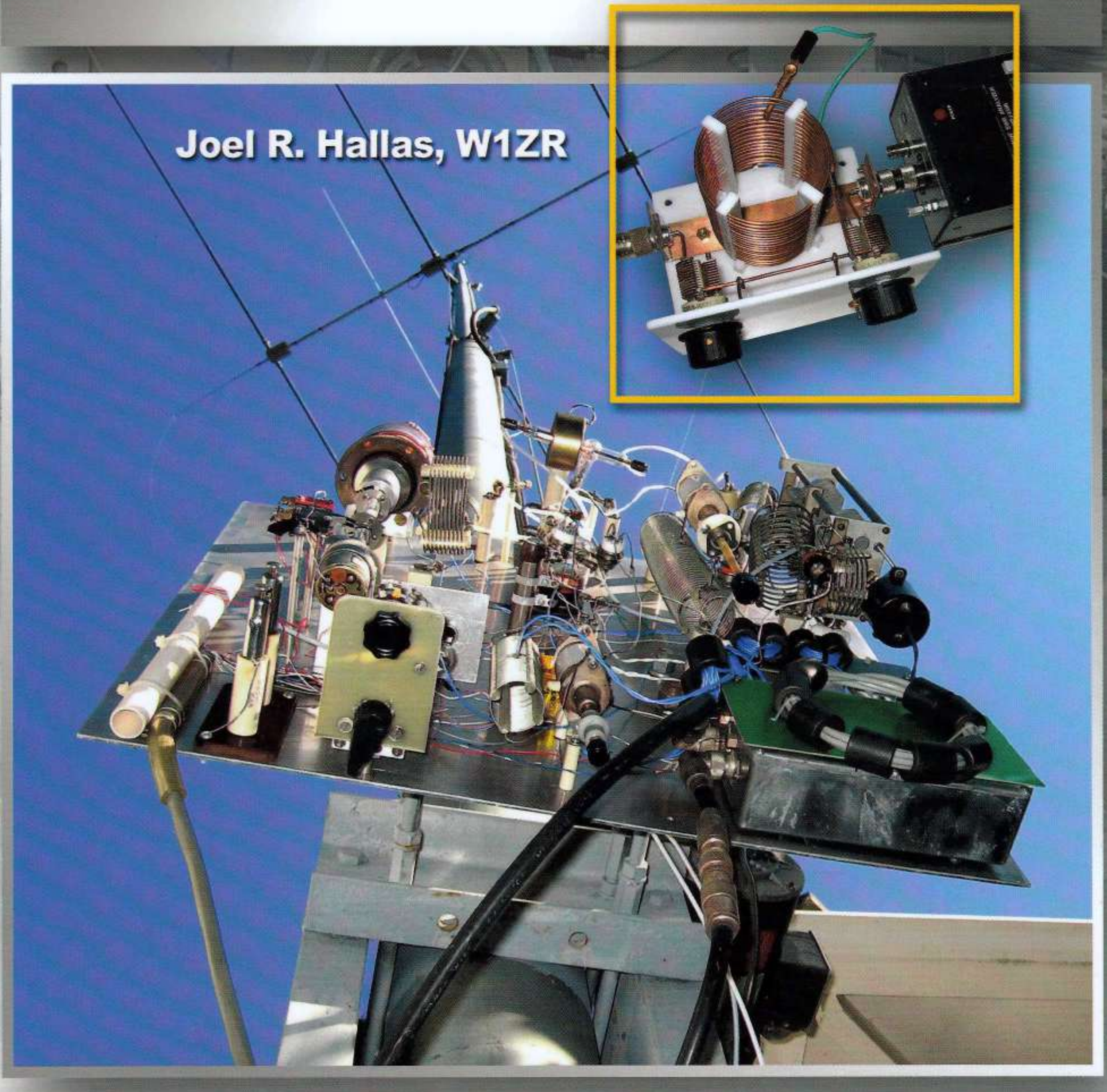

Joel R. Hallas, W1ZR

The ARRL Guide to
Antenna Tuners

A Radio Amateur's Guide to Antenna Matching

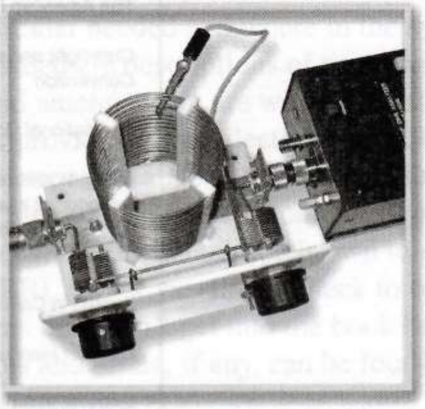

Joel R. Hallas, W1ZR

On the Cover

The group of special purpose antenna tuners made by the late George Badger, W6TC, is shown beneath his rotary antenna system. The inset shows a general purpose L-network antenna tuner built by Barry Shackleford, W6YE, who took both photos. Both are discussed in Chapter 15.

Production
Michelle Bloom, WB1ENT
Jodi Morin, KA1JPA
David Pingree, N1NAS

Cover
Sue Fagan, KB1OKW

Copyright © 2010-2012 by
The American Radio Relay League, Inc.

Copyright secured under the Pan-American Convention

International Copyright secured

All rights reserved. No part of this work may be reproduced in any form except by written permission of the publisher. All rights of translation are reserved.

Printed in Canada

Quedan reservados todos los derechos

IBSN13: 978-0-87259-098-4

1st Edition
2nd Printing

Foreword

Antenna Tuners are mystical devices that are sometimes needed between antenna systems and transmission lines or radios to make everything play the way it's supposed to. They are the subject of considerable confusion and misunderstanding by many Amateur Radio operators.

This book is intended to assist readers who have a basic knowledge of antennas, such as conveyed by *Basic Antennas — Understanding Practical Antennas and Designs*, in gaining an understanding of whether or not they need an antenna tuner in their station.[1] If it turns out one is needed, this book will help the reader decide the type of tuner needed and where in the system it should be applied. Also provided are descriptions of appropriate applications of such devices and antennas that are well suited for such use.An additional chapter provides data collected from ARRL Laboratory product reviews of representative antenna tuners to assist in tuner selection. Design methodology and specific design details are also provided for those who wish to build their own tuners.

As with all ARRL books, be sure to check to see if there are any last minute changes that didn't get into the book before it went to the printer. Updates and errata, if any, can be found at **www.arrl.org/product-notes/**.

David Sumner, K1ZZ
Executive Vice President
Newington, Connecticut
October 2010

---

[1] J. Hallas, W1ZR, *Basic Antennas — Understanding Practical Antennas and Designs,* Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9994. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/arrl.store/**; **pubsales@arrl.org**.

Acknowledgements

I would like to offer my special thanks to the ARRL production and editorial staff who made this book come together smoothly. In addition, thanks go to my long-suffering wife Nancy, W1NCY, who put up with the countless hours of my writing to make this happen.

I especially appreciated the tireless efforts of ARRL Technical Advisor Barry Shackleford, W6YE, who carefully examined every page of the manuscript in an effort to make sure there were neither technical nor grammatical errors and that all points were made clearly.

About the ARRL

The national association for Amateur Radio

The seed for Amateur Radio was planted in the 1890s, when Guglielmo Marconi began his experiments in wireless telegraphy. Soon he was joined by dozens, then hundreds, of others who were enthusiastic about sending and receiving messages through the air—some with a commercial interest, but others solely out of a love for this new communications medium. The United States government began licensing Amateur Radio operators in 1912.

By 1914, there were thousands of Amateur Radio operators—hams—in the United States. Hiram Percy Maxim, a leading Hartford, Connecticut inventor and industrialist, saw the need for an organization to band together this fledgling group of radio experimenters. In May 1914 he founded the American Radio Relay League (ARRL) to meet that need.

Today ARRL, with approximately 150,000 members, is the largest organization of radio amateurs in the United States. The ARRL is a not-for-profit organization that:

- promotes interest in Amateur Radio communications and experimentation
- represents US radio amateurs in legislative matters, and
- maintains fraternalism and a high standard of conduct among Amateur Radio operators.

At ARRL headquarters in the Hartford suburb of Newington, the staff helps serve the needs of members. ARRL is also International Secretariat for the International Amateur Radio Union, which is made up of similar societies in 150 countries around the world.

ARRL publishes the monthly journal *QST*, as well as newsletters and many publications covering all aspects of Amateur Radio. Its headquarters station, W1AW, transmits bulletins of interest to radio amateurs and Morse code practice sessions. The ARRL also coordinates an extensive field organization, which includes volunteers who provide technical information and other support services for radio amateurs as well as communications for public-service activities. In addition, ARRL represents US amateurs with the Federal Communications Commission and other government agencies in the US and abroad.

Membership in ARRL means much more than receiving *QST* each month. In addition to the services already described, ARRL offers membership services on a personal level, such as the ARRL Volunteer Examiner Coordinator Program and a QSL bureau.

Full ARRL membership (available only to licensed radio amateurs) gives you a voice in how the affairs of the organization are governed. ARRL policy is set by a Board of Directors (one from each of 15 Divisions) elected by the membership. The day-to-day operation of ARRL HQ is managed by a Chief Executive Officer.

No matter what aspect of Amateur Radio attracts you, ARRL membership is relevant and important. There would be no Amateur Radio as we know it today were it not for the ARRL. We would be happy to welcome you as a member! (An Amateur Radio license is not required for Associate Membership.) For more information about ARRL and answers to any questions you may have about Amateur Radio, write or call:

ARRL—The national association for Amateur Radio
225 Main Street
Newington CT 06111-1494
Voice: 860-594-0200
Fax: 860-594-0259
E-mail: **hq@arrl.org**
Internet: **www.arrl.org/**

Prospective new amateurs call (toll-free):
800-32-NEW HAM (800-326-3942)
You can also contact us via e-mail at **newham@arrl.org**
or check out *ARRLWeb* at **http://www.arrl.org/**

Table of Contents

Chapter 1 Why Might I Need an Antenna Tuner

Chapter 2 A Look at a Typical Configuration

Chapter 3 So Just What is an Antenna Tuner?

Chapter 4 Tuning an Antenna Tuner

Chapter 5 The Internal Tuner — How Does it Help?

Chapter 6 An External Tuner at the Radio

Chapter 7 Transmission Lines and Loss

Chapter 8 Moving the Tuner to the Back 40

Chapter 9 Transmission Line Choices for Low Loss

Chapter 10 Balanced versus Unbalanced Lines

Chapter 11 So What's a Balun, an Unun, a Choke?

Chapter 12 Balanced Antenna Tuners

Chapter 13 Antennas that Work Well with Tuners

Chapter 14 A Survey of Available Tuners

Chapter 15 Making Your Own Tuner

Appendix A Using Decibels for Loss Calculations

Chapter 1

Why I Might Need an Antenna Tuner

Amateur Radio station W1ZR, home station of the author. The unit on the far left is an early Ten-Tec 238 manual antenna tuner.

Contents

Radio Incompatibility? 1-2

Does My Antenna Have to be a Resistor?................................ 1-2

So What's the Problem? 1-3

What Happens if the Radio and Antenna Aren't Matched? 1-3

So What Can We Do?.. 1-4

What's in a Name? 1-5

Review Questions.. 1-6

To transmit radio signals, you need a transmitter, often the transmitter side of a *transceiver* — a combined transmitter and receiver — and an antenna. Of course you will need a few other things such as legal authority to transmit, from the FCC in the US, a power source and an information source, such as a telegraph key, microphone or PC. In this book we will focus on the interconnection between the transmitter and the antenna, leaving the other topics to other books.

You might notice that, in spite of the book's title, we didn't mention antenna tuners — a device that goes between the transmitter and antenna. This is because, strictly speaking, an antenna tuner isn't necessary, at least not in all cases. The antenna tuner is only required if the transmitter can't put its output power into the antenna because of an incompatibility between them.

Radio Incompatibility?

A radio transmitter comes with a set of specifications. For proper operation, the owner is responsible to ensure that the requirements listed in the specifications are met. Some are straightforward, such as "power required: 13.8 V dc at 20 A max." If we plugged such a radio into a 120 V ac outlet, we wouldn't have a right to expect it to operate properly — in fact we might expect to see smoke and flames. We would need an intermediate device, called a power supply, to transform the 120 V ac in our outlet to the 13.8 V dc our radio wants.

The compatibility issue we will be considering here is one relating to the *ANTENNA IMPEDANCE* specification. While not all transmitter specifications include an explicit antenna specification, most will say something like *ANTENNA IMPEDANCE, 50 Ω (Unbalanced)* or possibly *ANTENNA IMPEDANCE, 50 Ω (Unbalanced) with SWR of 2:1 or less.* These specifications indicate the load that the antenna system must present to the radio for proper operation. We will discuss the different parts of this specification and what they mean as we go forward. As with the power supply, if the antenna doesn't meet the specification's requirements, we might need an intermediate device — possibly an antenna tuner.

Does My Antenna Have to be a Resistor?

The indication "50 Ω" does sound a lot like a resistor. The radio will be very happy to put its power into an actual 50 Ω resistor, in fact that's how we often test a radio when we don't want the radio signal to go out over the air. Unfortunately, a resistor absorbs the transmitter power and turns it into heat — it does not make a good antenna.

In a properly designed antenna, it is often the case that the antenna's radiating properties will be converted into a load that acts like a 50 Ω resistor to the radio, but actually radiates the power as radio waves.

So What's the Problem?

It would seem that we can solve our incompatibility problem by just buying (or building) an antenna that has a compatible specification of a 50 Ω resistive impedance and connecting it to our radio. This is quite true, and can be very successful — within certain constraints (see **Figure 1-1** for the ideal case). This compatible case is referred to as a *matched* system. Unfortunately, the real world rears its ugly head in a few ways:

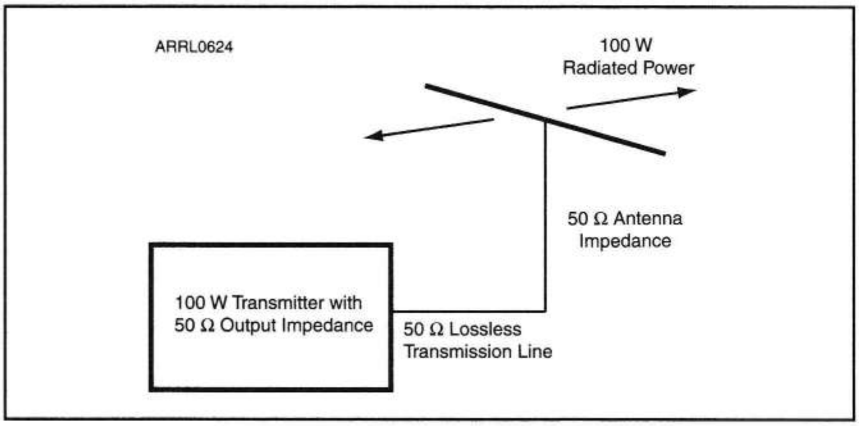

Figure 1-1 — Radio and antenna system with matched impedances.

- The biggest issue is that any antenna will have its particular design impedance on a single frequency. This is not an issue with many radio services — broadcast stations, for example — that operate on a single assigned frequency. Some services, such as the Amateur Radio service, operate within assigned bands. Shifting frequency from one end of the band to the other can result in a significant change in antenna impedance.
- A real antenna is always installed at some height above the local terrain. The electrical properties of the soil and the height above ground will have a significant impact on antenna impedance. For example, perhaps the simplest antenna, a half wave horizontal dipole has an impedance that varies from around 40 to 100 Ω as it moves from 0.1 to 0.35 λ (wavelengths) above ground.

What Happens if the Radio and Antenna Aren't Matched?

Most radios can tolerate a certain amount of mismatch from an antenna system without any problems. This is often specified in terms of standing wave ratio (SWR), a measure of mismatch. Often the allowed value is 2:1; which, for a 50 Ω system, would represent resistive values of 25 or 100 Ω (50/2, or 50 × 2).[1] Note that while the radio will operate without damage at this level of mismatch, it may not operate quite as well as if it were matched.

A mismatched load impedance in early solid state transmitters could result in damage to components in the power amplifier stages due to higher voltages or currents than it was designed to handle. Modern transceivers have *fold-back* circuitry that senses the mismatch and reduces transmitter power to avoid damage. Note that while the transmitter will not be damaged, and still can be used, it will put out less power, sometimes beginning to fold back at an SWR as low as 1.5:1. This is perhaps the most common reason that we might notice our 100 W transmitter actually putting out 25 W.

[1] In addition to these two resistive values, there is a whole set of combinations of resistive and reactive impedance components that will also result in an SWR of 2:1.

So What Can We Do?

Perhaps not surprisingly, one solution to this issue is something that is generally called an *antenna tuner*. The antenna tuner is a variable impedance transforming device that can transform the impedance of an antenna system so that it appears to the transmitter as a 50 Ω load, while causing almost all of the transmitter power to be radiated from the antenna, just as if everything were matched.

As shown in **Figure 1-2**, the antenna tuner can be placed directly at the transmitter and connect directly to an antenna. In many cases, a *transmission line* is used between the transmitter and the antenna. We will discuss transmission lines in more detail later, but for now be aware that a transmission line is frequently used if the transmitter and antenna are not in the same place. As shown in **Figures 1-3** and **1-4**, the antenna tuner can be placed at either the antenna end or the radio end of a transmission line interconnecting the radio and antenna. It could even be put at an intermediate point.

Much of this book will discuss the details of the different configurations and requirements of antenna tuners and where they are located, so you can make the most informed decision on what type of tuner to buy or build and where to install it.

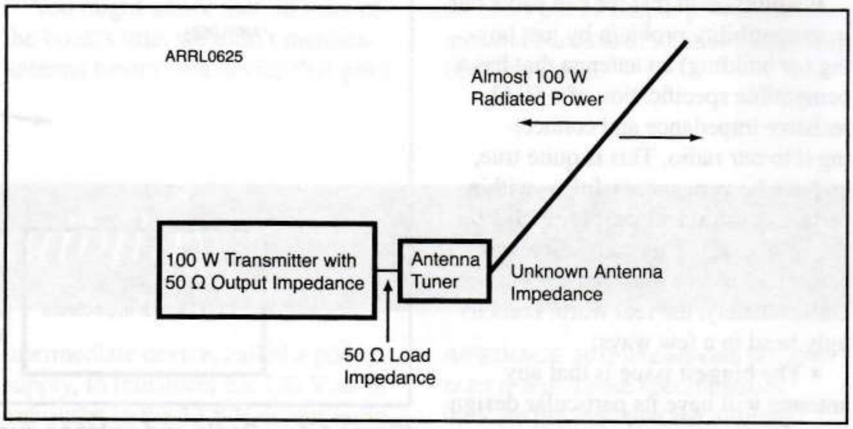
ARRL0625
Almost 100 W
Radiated Power
100 W Transmitter with
50 Ω Output Impedance
Antenna
Tuner
Unknown Antenna
Impedance
50 Ω Load
Impedance

Figure 1-2 — Antenna tuner collocated with the radio and the antenna. All the transmitter power reaches the antenna except for any loss in the tuner, generally less than 5%.

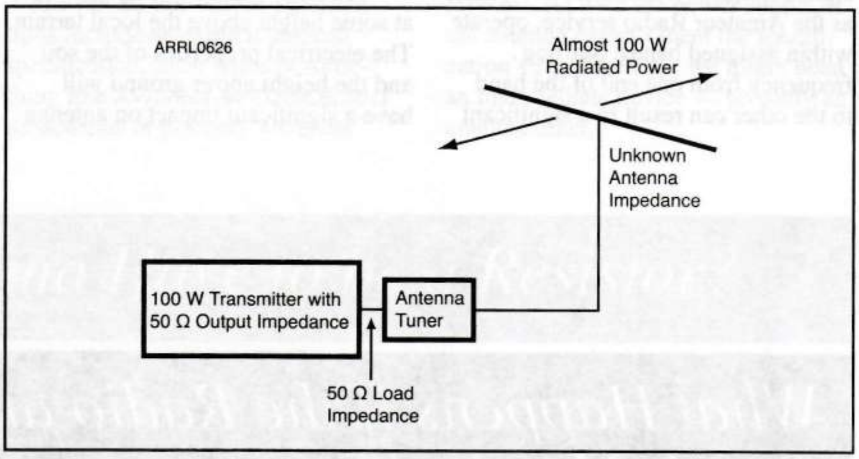
ARRL0626
Almost 100 W
Radiated Power
Unknown
Antenna
Impedance
100 W Transmitter with
50 Ω Output Impedance
Antenna
Tuner
50 Ω Load
Impedance

Figure 1-3 — Radio and antenna system with tuner located at the radio end of the transmission line.

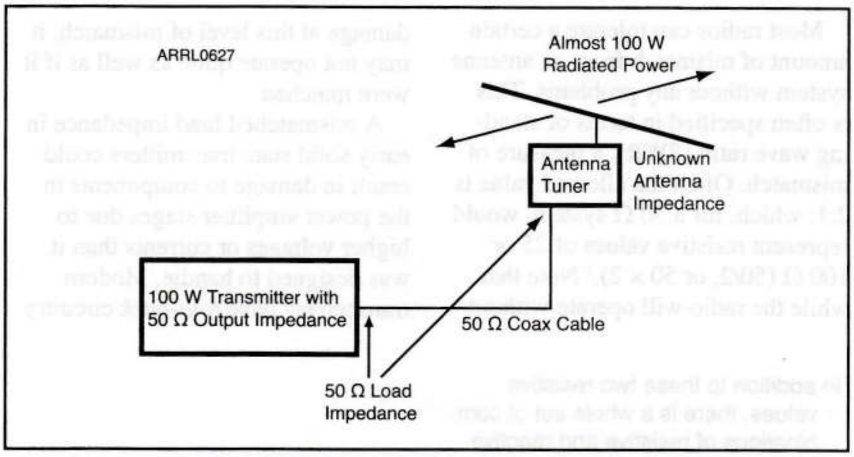
ARRL0627
Almost 100 W
Radiated Power
Antenna
Tuner
Unknown
Antenna
Impedance
100 W Transmitter with
50 Ω Output Impedance
50 Ω Coax Cable
50 Ω Load
Impedance

Figure 1-4 — Radio and antenna system with tuner located at the antenna end of the transmission line.

What's in a Name?

Over the years there has been some discussion of what to call this device I call an *antenna tuner*. There are those who would say that *antenna tuner* is a misnomer because it doesn't actually tune the antenna, but really tunes the transmission line in most applications.

Over the years many other terms have been used to describe this device. You will see references to an *antenna coupler*, *antenna tuning unit* (ATU) and a *transmatch* — all within various ARRL magazines and books. They all mean the same thing and refer to the same type of device.

The best name I've heard for this function is *antenna system matching device*. Since that doesn't resolve itself to an easy to pronounce acronym, throughout this book I will use the most common term *antenna tuner* with neither apology nor regret.

Review Questions

1-1 Under what conditions is there no requirement for an antenna tuner in a transmitting system?

1-2 How can you tell if an antenna tuner is needed?

1-3 What are the likely consequences of not having an antenna tuner, or having a misadjusted antenna tuner?

1-4 If an antenna tuner is properly adjusted on one frequency, what must be done if the operating frequency is changed?

Chapter 2

A Look at Typical Configurations

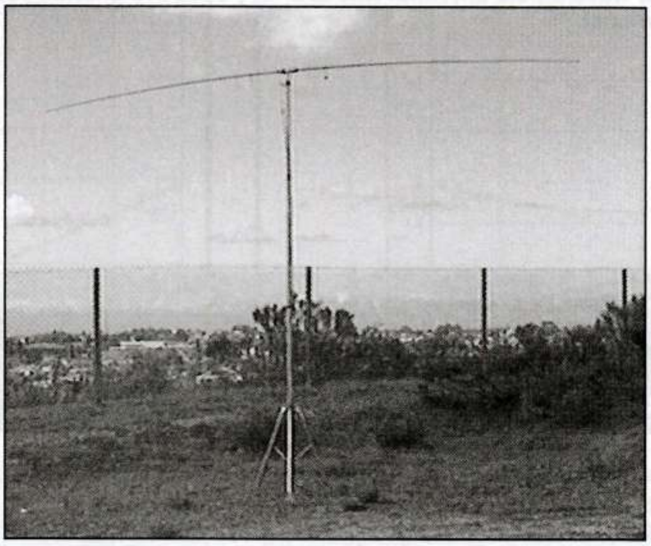

The classic half-wave dipole — an antenna mainstay at many amateur stations.

Contents

A Transmission Line — The Piece in the Middle........................2-2

Dipole for the 40 Meter Amateur Band..2-3

The Dipole for the 80 Meter Amateur Band..................................2-4

Fractional Bandwidth 2-5

Review Questions..2-6

A very popular antenna for a new radio amateur is called a *half-wave dipole*. This antenna is easy to construct from two pieces of wire with a total length of about 468/f, where f is the frequency in MHz and the result is in feet. The antenna is often fed with 50 Ω coaxial cable as shown in **Figure 2-1**. Generally, the antenna length is trimmed equally on each side of center until the closest match to 50 Ω is found. While it will function as shown, more predictable performance is often obtained by inserting a device between the inherently balanced antenna and unbalanced (with respect to ground) coaxial cable to provide an appropriate transition between them. This device is generally called a balun — for balanced to unbalanced transition. We will discuss this in detail later in the book.

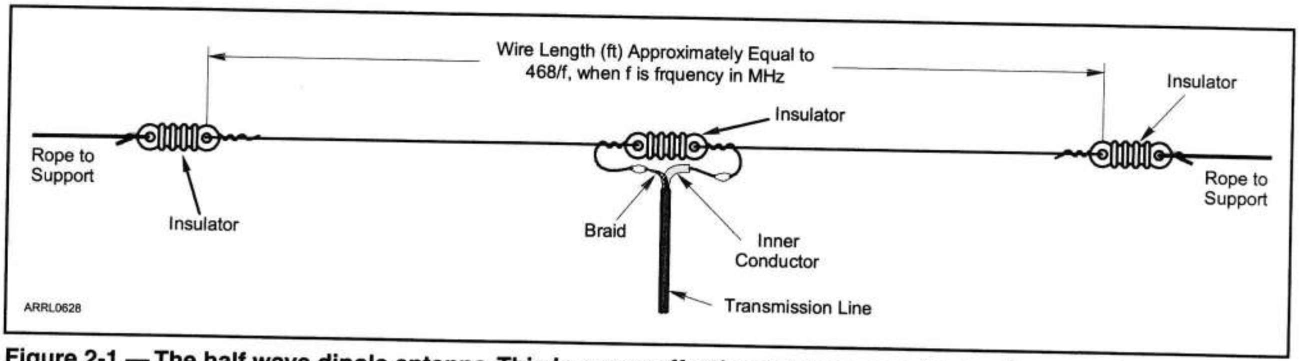

Figure 2-1 — The half wave dipole antenna. This is a very effective antenna, considering its ease of construction and low cost.

A Transmission Line — The Piece in the Middle

In most cases the antenna and radio are located in separate areas. Most often, the optimum location for the radio is somewhere that a person can comfortably operate a radio. For many types of radio systems, this means indoors. The antenna, on the other hand, is typically located outside so it can radiate with minimal obstructions and with at optimum height above the terrain.

It is possible to meet both requirements if an antenna is connected to the radio system through a *transmission line*. An ideal transmission line conveys energy from one end to the other in a totally transparent manner. A real transmission line, on the other hand does so with some limitations:

- Each type of transmission line has a characteristic impedance, Z_0. If the transmission line is terminated with an impedance equal to the Z_0, the same impedance will be seen at any length of line. Fortunately, there are many types of line that have a Z_0 of 50 Ω, so such a line connected to an antenna with a 50 Ω impedance will present a 50 Ω load to the radio no matter the length of the line. It is worth mentioning briefly at this point that if the line is terminated in an impedance other than its Z_0, the impedance at the far end in general will be neither the terminating impedance nor the Z_0 of the line.

These topics will be covered in much more detail later, but it's good to have an idea as we go through the basics.

- Real transmission lines have some associated losses. Thus not all the power put into the bottom by the transmitter comes out at the antenna. The attenuation is higher with increasing frequency, for a given type of line. The attenuation is also directly proportional to line length. In addition, the attenuation increases with standing wave ratio (SWR, a measure of mismatch) — another reason to have a matched system, if possible. We will discuss this phenomenon later.

- SWR is generally determined by measuring the ratio of reflected power to forward power. An SWR of 1:1 results in no reflected power. If the line has significant attenuation, both the power getting to the antenna, and any reflected power will be attenuated. Thus a measurement of SWR at the radio end will appear to be lower than the real SWR at the antenna end.

Dipole for the 40 Meter Amateur Band

If we were to make a dipole of the form of Figure 2-1 for the center of the US 40 meter band, we would use the formula 468/f, with the f equal to 7.15 MHz to result in a starting length of 65.45 feet, or 65 feet, 5.4 inches. As noted previously, this will vary with height, but let's say we get it all correct and have a height that provides close to a 50 Ω load at that frequency. If we then measure the standing wave ratio (SWR) across the band, 7.0 – 7.3 MHz, we would observe a set of SWR data very much like that plotted in **Figure 2-2.**[1,2]

Note that, as desired, we observe an almost perfectly matched condition (SWR of 1:1) at our design frequency of 7.15 MHz. Note also that as we change frequency across the band, our SWR rises to 2:1 at each band edge. Thus, even though our antenna is only closely matched at the band center, if the transmitter can operate at full power with a 2:1 SWR, this antenna will be usable across the band, if trimmed very carefully.

What About Height?

We have previously mentioned that changing height will also change the

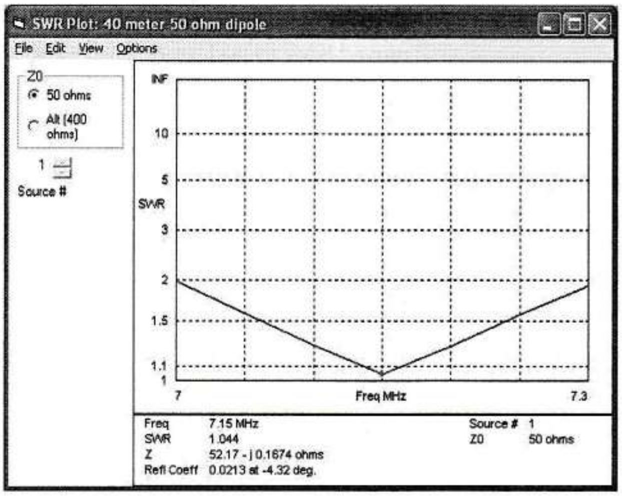
SWR Plot: 40 meter 50 ohm dipole
File Edit View Options
INF
10
5
SWR
3
2
1.5
1.1
1
7
Freq MHz
7.3

Z0
● 50 ohms
○ Alt (400 ohms)
1
Source #

Freq 7.15 MHz
SWR 1.044
Z 52.17 - j 0.1674 ohms
Refl Coeff 0.0213 at -4.32 deg.
Source # 1
Z0 50 ohms

Figure 2-2 — Plot of SWR of a 40 meter dipole with height and length adjusted for 50 Ω at 7.15 MHz.

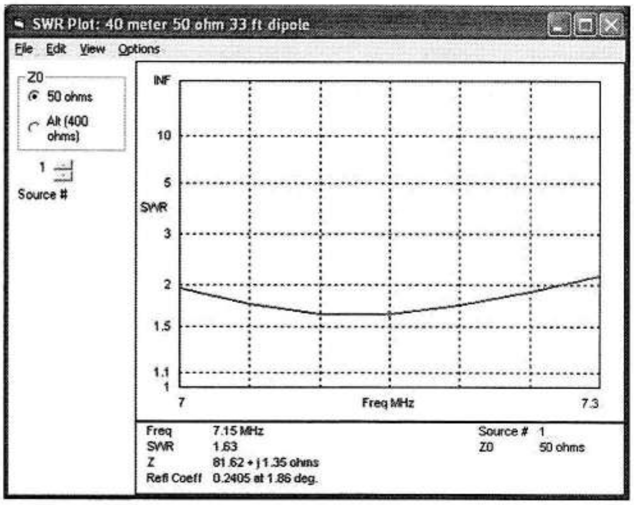
SWR Plot: 40 meter 50 ohm 33 ft dipole
File Edit View Options
INF
10
5
SWR
3
2
1.5
1.1
1
7
Freq MHz
7.3

Z0
● 50 ohms
○ Alt (400 ohms)
1
Source #

Freq 7.15 MHz
SWR 1.63
Z 81.62 + j 1.35 ohms
Refl Coeff 0.2405 at 1.86 deg.
Source # 1
Z0 50 ohms

Figure 2-3 — Plot of SWR of a 40 meter dipole with height set to λ/4, about 33 feet.

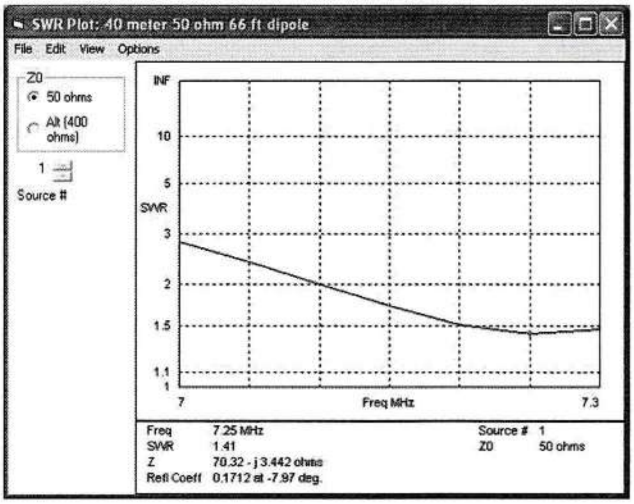
SWR Plot: 40 meter 50 ohm 66 ft dipole
File Edit View Options
INF
10
5
SWR
3
2
1.5
1.1
1
7
Freq MHz
7.3

Z0
● 50 ohms
○ Alt (400 ohms)
1
Source #

Freq 7.25 MHz
SWR 1.41
Z 70.32 - j 3.442 ohms
Refl Coeff 0.1712 at -7.97 deg.
Source # 1
Z0 50 ohms

Figure 2-4 — Plot of SWR of 40 meter dipole of Figure 2-3 raised to a height of λ/2, about 66 feet. Note the change in frequency of minimum SWR. It could be tuned back to mid-band by lengthening.

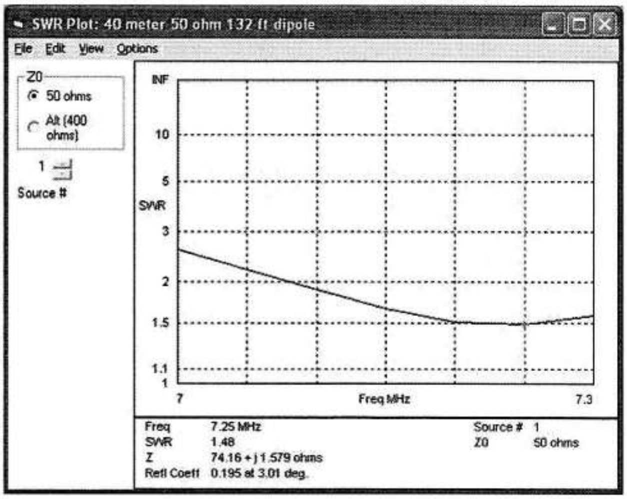
SWR Plot: 40 meter 50 ohm 132 ft dipole
File Edit View Options
INF
10
5
SWR
3
2
1.5
1.1
1
7
Freq MHz
7.3

Z0
● 50 ohms
○ Alt (400 ohms)
1
Source #

Freq 7.25 MHz
SWR 1.48
Z 74.16 + j 1.579 ohms
Refl Coeff 0.195 at 3.01 deg.
Source # 1
Z0 50 ohms

Figure 2-5 — Plot of same antenna raised to 1 λ, about 132 feet.

---

[1] Notes appear on page 6.

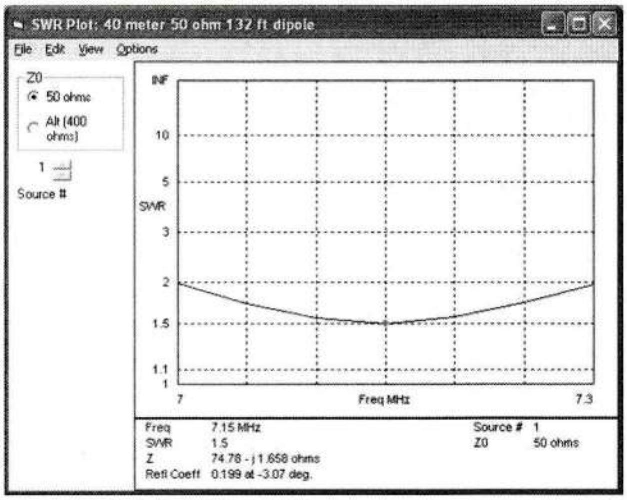

Z0
● 50 ohms
○ Alt (400 ohms)

Source # 1

Freq 7.15 MHz
SWR 1.5
Z 74.78 - j 1.658 ohms
Refl Coeff 0.199 at -3.07 deg.

Source # 1
Z0 50 ohms

Figure 2-6 — Plot of SWR of a 1 λ high 40 meter dipole trimmed for mid band.

antenna impedance. In fact it will also change the resonant frequency of the antenna. To obtain the plot in Figure 2-2, we lowered the antenna to a height of 12 feet. While such an antenna will operate, the combination of its direct signal and the signal reflected from the ground will result in most of the radiation going skyward. This makes it suitable for short and medium range communication. In many cases antennas are raised higher to increase low angle radiation to support communication over longer distances.

Impedance and Resonant Frequency at Greater Height

Figs 2-3 through **2-5** are a series of SWR plots of the same antenna as it is raised, first to 33 feet (λ/4), then 66 feet (λ/2) and finally 132 feet (1 λ). Observe the two phenomena happening as the antenna is raised — first, the impedance is changing and second the resonant frequency (frequency at which the impedance is resistive) is changing. Note that the impedance shown is that close to the resonant frequency.

To improve the SWR, we can adjust the length by making it long enough so that the resonant frequency is moved back to mid band. The result is shown in **Figure 2-6**. Note that the antenna now has an SWR range across the band of 2:1 or less so it might still work well across the band without a tuner. Note that to achieve this result, it must be very carefully trimmed.

A Dipole for the 80 Meter Amateur Band

A dipole for 80 meters (don't worry, this is the last example) has a very different result. The US 80 meter (sometimes called 75 and 80 meters) band is wider than 40 meters, covering 3.5 to 4 MHz. The result, again with the height lowered and length adjusted to match 50 Ω at mid band is shown in **Figure 2-7**. This is a very different result from our 40 meter dipole's impedance over the band. The SWR at the band edges is now above 7:1, and the bandwidth over which the SWR is 2:1 or less, the SWR bandwidth, is about 150 kHz — about half what it was on 40 meters.

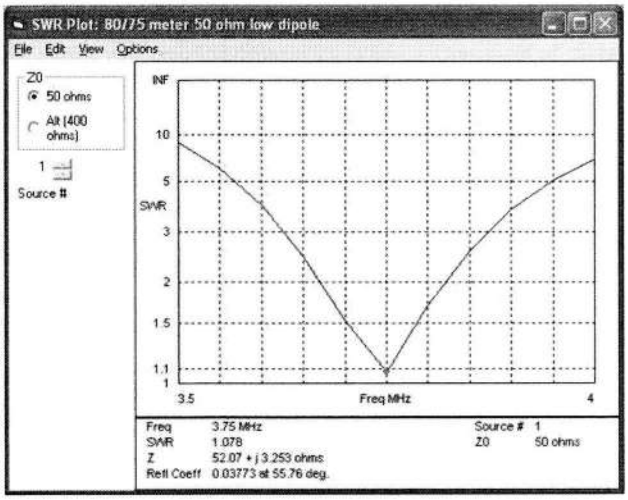

Z0
● 50 ohms
○ Alt (400 ohms)

Source # 1

Freq 3.75 MHz
SWR 1.078
Z 52.07 + j 3.253 ohms
Refl Coeff 0.03773 at 55.76 deg.

Source # 1
Z0 50 ohms

Figure 2-7 — Plot of SWR of a 80/75 meter dipole with height and length adjusted for 50 Ω at 3.75 MHz.

Fractional Bandwidth

The reduced bandwidth on 80 meters illustrates a few interesting principles. First, for a band this wide at this frequency, it would take at least three dipoles with slightly different lengths to cover the whole band if we didn't have a device to transform the impedance to a value we wanted — an *antenna tuner*. Second, we note that the key factor is not the bandwidth itself, but rather the fraction (or percentage) of the center frequency that we wish to be able to cover.

In the case of 40 meters, we were able to cover a bandwidth (7300 – 7000) of 300 kHz with an SWR of 2:1. This is the fraction 300/7150 = 0.042, or 4.2%. Note that on 80 meters we had about 150/3750 = 4%. We can use this as a predictor of whether a single thin dipole can cover each amateur band with the results shown in Table 2-1.

Table 2-1

Percentage Bandwidth of Each HF and VHF US Amateur Band.

Band (Meters)	*Upper Frequency (kHz)*	*Lower Frequency (kHz)*	*Bandwidth (kHz)*	*Bandwidth (%)*
160	2000	1800	200	10.5
80	4000	3500	500	13.3
60	5407	5330	77	1.4
40	7300	7000	300	4.2
30	10,100	10,150	50	0.5
20	14,000	14,350	350	2.5
17	18,068	18,168	100	0.6
15	21,000	21,450	450	2.1
12	24,890	24,990	100	0.4
10	28,000	29,700	1700	5.9
6	50,000	54,000	4000	7.7
2	144,000	148,000	4000	2.7
1.25	222,000	225,000	3000	1.3
70 cm	420,000	450,000	30,000	6.9

Clearly, the problem is most pronounced on 80 meters, perhaps why we chose that example. In addition, 160, 10 and 6 meters are wider than our threshold, as is the 70 cm band. Fortunately, the higher three of these bands have operations divided by mode, so that it is not unreasonable to have, for example, a horizontal 6 meter antenna centered at the low end for CW and SSB operation and a separate vertical antenna for FM designed for the high end. Many operators, however, like to work all portions of 160 and 80 meters, so that can be an issue.

Notes

[1] This plot was obtained using an antenna modeling program named *EZNEC*. We will call upon this program repeatedly throughout the book. Several versions of *EZNEC* antenna modeling software, including a free trial version, are available from developer Roy Lewallen, W7EL, at **www.eznec.com**.

[2] For an introduction to *EZNEC* modeling, see Appendix A of J. Hallas, W1ZR, *Basic Antennas — Understanding Practical Antennas and Design*. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9994. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/arrl-store**; **pubsales@arrl.org**. See chapters 2-6.

Review Questions

2-1 Under what conditions might a ½ wave dipole not be compatible with a radio system?

2-2 Why can a transmission line tend to mask a high SWR?

2-3 If a dipole should be able to cover a complete amateur band, but doesn't because the minimum SWR is higher than the middle of the band, what can you do to improve the situation?

2-4 If you have no antenna tuner, how many separate dipoles would be required to cover the entire 10 meter amateur band with an SWR of less than 2:1?

Chapter 3

So Just What Is an Antenna Tuner?

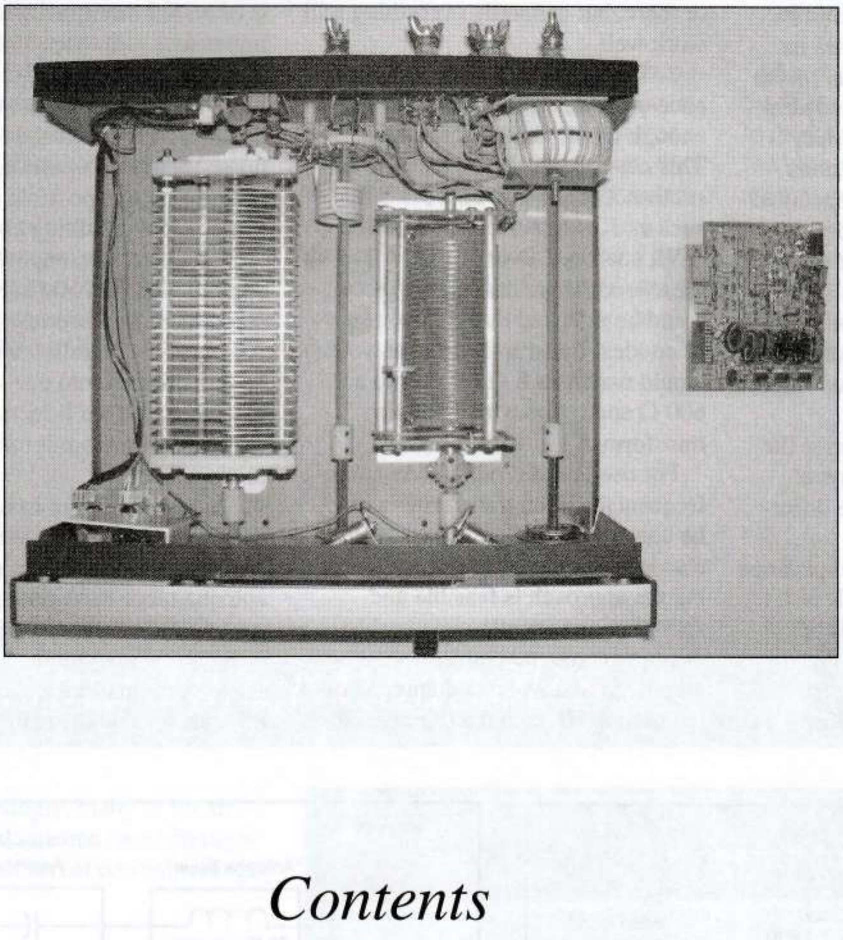

On the left, inside view of a manually tuned 1500 W L-network HF tuner (Ten-Tec 238); on the right, a computer controlled 100 W automatic tuner (Elecraft KAT3) designed to fit inside the compact Elecraft K-3 HF and 6 meter transceiver.

Contents

What's In an Antenna Tuner.. 3-2

Antenna Tuner Controls.. 3-3

Sidebar — A Quick Discussion of Standing Wave Ratio 3-4

Where are the Knobs on an Automatic Antenna Tuner? 3-8

What's the Best Tuner for You? 3-9

Review Questions... 3-10

As noted in Chapter 1, *the antenna tuner is a variable impedance transforming device that can transform the impedance of an antenna system to appear to the transmitter as a 50 Ω load, while causing almost all of the transmitter power to be radiated from the antenna, just as if everything were matched.* This sounds like just what we need to solve the problem of using an antenna system that has an SWR that the transmitter doesn't like to deal with.

While an antenna tuner may well be just what we need — they come in many flavors, shapes and sizes — not to mention specification ratings. Much of this book will deal with how to choose the best type of antenna tuner, as well as to determine the best place in the system to locate it.

What's In an Antenna Tuner?

In order to perform its transformation function, an antenna tuner must deal with any load that is within its specified limits (everything has its limits — no antenna tuner can match absolutely everything!). As noted in the sidebar, an SWR greater than 1:1 happens if the load of the antenna system is different from the specified output load impedance of the transmitter, 50 Ω for most current designs. Table B in the sidebar shows some representative values of resistance and capacitive or inductive series reactance that result in various values of SWR.

Note that we could categorize the load impedance into four general categories that could all have different solutions:

1. The load is a resistive impedance of 50 Ω. In this case the SWR is 1:1 and we don't need an antenna tuner at all. We could bundle with this all loads that work well with the transmitter without change — perhaps those with an SWR of less than 2:1. There may be a bit of extra loss here or there, but generally everything will work well.

2. The load is a resistive impedance other than 50 Ω , and not close enough to allow proper operation. This could be loads with a 2:1 SWR such as 25 Ω or 100 Ω, a 3:1 SWR such as 16.7 Ω or 150 Ω, or a 10:1 SWR such as 5 Ω or 500 Ω. All of these loads share the fact that they could be matched through the use of an ideal transformer, just as you would match an 8 Ω speaker to a 600 Ω audio line with an audio transformer.

For use in radio rather than audio frequencies, such transformers must be constructed differently to avoid also introducing reactive components, but the approach is feasible and sometimes encountered. One classic example is the matching arrangement employed in a wideband antenna often used at HF called a rhombic. It offers an almost resistive load of around 600 Ω over a 4:1 frequency range and is often fed through a wideband 12:1 impedance ratio transformer by a 50 Ω system. **Figure 3-1** shows the idea. While most folks would call it a transformer rather than an antenna tuner, there's no question that it is performing one possible class of the function we're talking about.

3. The load is composed of a combination of a 50 Ω resistive component and a series reactance. A load with a parallel reactance can be transformed into one with a series reactance, but we'll focus on the series version since it may be easier to visualize.

This makes for an easy to imagine antenna tuner. If the series reactance of the load is inductive ($+X_L$), all the antenna tuner need contain is a component with a capacitive reactance of the same magnitude, ($-X_C$). The series combination of the two is just 0 Ω reactive, so the resultant is the

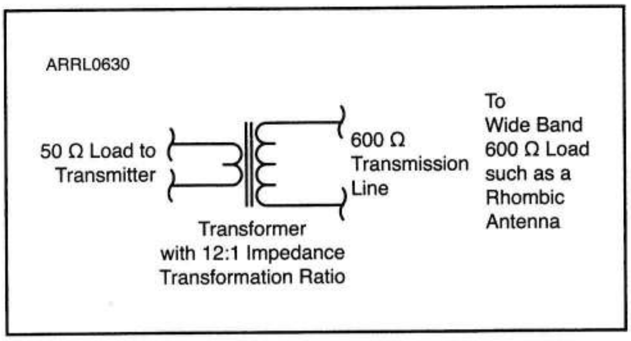
ARRL0630
50 Ω Load to Transmitter
600 Ω Transmission Line
To Wide Band 600 Ω Load such as a Rhombic Antenna
Transformer with 12:1 Impedance Transformation Ratio

Figure 3-1 — A wide-band impedance transformer can serve the function of an antenna tuner for a resistive load.

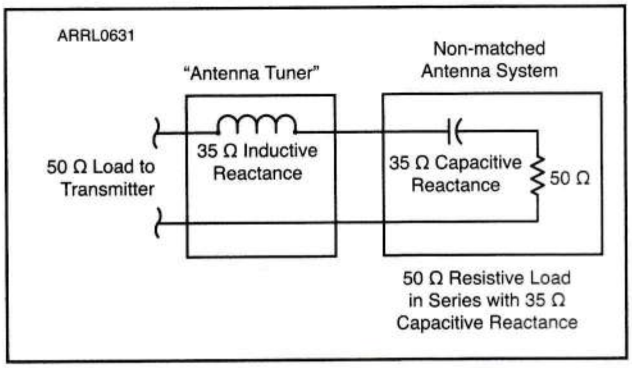
ARRL0631
"Antenna Tuner"
Non-matched Antenna System
50 Ω Load to Transmitter
35 Ω Inductive Reactance
35 Ω Capacitive Reactance
50 Ω
50 Ω Resistive Load in Series with 35 Ω Capacitive Reactance

Figure 3-2 — Antenna tuner for special case of reactance in series with resistive part equal to Z_0.

resistive component —50 Ω as shown in **Figure 3-2**.

Note that, unlike the wideband transformer, this will likely work only on a single frequency. If the antenna has an inductive reactance, it will tend to increase with frequency ($X_L = 2 \times \pi \times f \times L$), at least over a limited frequency range. In order to keep the total impedance resistive, we will need a variable capacitor, so we can adjust the capacitive reactance to match the changing inductive reactance of the load. Note that our "antenna tuner" now has a knob on the front, and is starting to look more like an antenna tuner.

4. The load is a combination of resistive and reactive components that have an SWR greater than we can deal with and that falls in none of the other three categories. This is the general case — the one all general purpose antenna tuners must deal with, at least over some range of SWR.

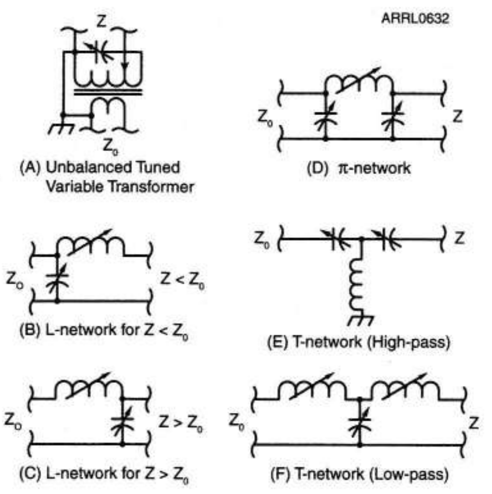
ARRL0632
Z
Z_0
(A) Unbalanced Tuned Variable Transformer
Z_0
Z
(D) π-network
Z_0
Z < Z_0
(B) L-network for Z < Z_0
Z_0
Z
(E) T-network (High-pass)
Z_0
Z > Z_0
(C) L-network for Z > Z_0
Z_0
Z
(F) T-network (Low-pass)

Figure 3-3 — Antenna tuner configurations that can be used to match the more general mismatched load.

There are a number of circuit configurations that can be employed that will do a combination of cancelling the reactance and transforming the resistive impedance. We have cataloged them in **Figure 3-3** so you will recognize the configurations as we discuss them. In many cases, if constructed of the same quality components, they will provide equivalent performance. As with all general purpose devices, there are certain advantages and disadvantages of each, and some may be more appropriate for some applications than for others.

Another categorization that can be made is whether the tuner is designed to feed an unbalanced load, typically fed with coaxial cable or a single wire against ground, or is intended to feed a balanced load such as a dipole fed with balanced open wire transmission line. Each of these configurations can be adapted to either task, however, we will defer that discussion until later.

Antenna Tuner Controls

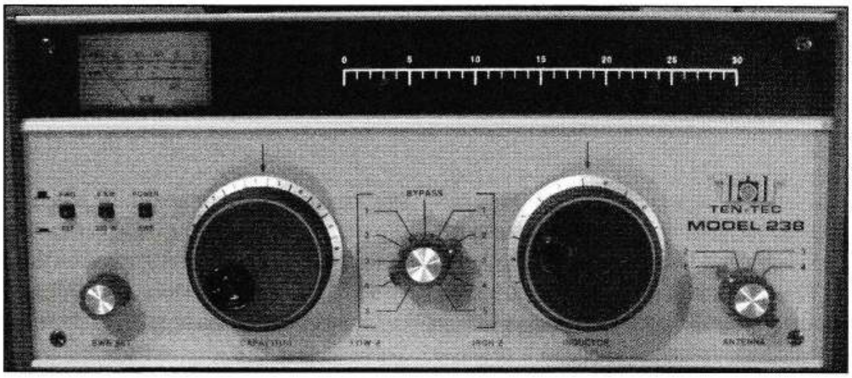

Figure 3-4 — Front panel view of an L-network tuner, an early version of the Ten-Tec 238.

Not surprisingly, many of the different manual antenna tuner configurations have different control arrangements.

L-Network Tuner Controls

The L-network has two primary adjustments, an inductor and a capacitor. In the Ten-Tec 238 tuner shown in **Figure 3-4**, and in most L-network tuners, each control is adjustable throughout its entire range on any band, providing maximum flexibility. In addition, there is a switch to move

A Quick Discussion of Standing Wave Ratio

We discussed standing wave ratio in Chapter 2 and noted that a matched antenna system has a standing wave ratio (SWR) of 1:1. That is the case in which there are no standing waves and all energy from the transmitter is delivered to the load with no reflections on the transmission line. This is a very straightforward situation that is easy to understand. Getting into the details of this topic can get lengthy and there are entire books devoted to the topic.[A] Fortunately, we won't need to get that deeply involved to understand antenna tuners and their applications. Still, we need to have some idea about SWR, what it means and what its consequences are to understand, when and if it is a problem and how to solve it.

The Matched Transmission Line

If we were to connect a battery to a long real transmission line with no load and monitor the current that would flow with a high speed oscilloscope we would notice an interesting effect. The combination of the line series inductance and shunt capacitance, and to a lesser effect the wire resistance, would result in an initial current that would flow, even though there is an open circuit at the far end.

The current will continue to flow until the signal propagates to the end of the line and returns. When it reaches the end of the line (traveling somewhat slower than the speed of light) a *reflected* wave of the opposite polarity will propagate back because there can be no current flow at the open circuit. When the reflected wave returns to the source end of the line the combination of the forward and reflected wave will result in zero current — just what we would expect for an open transmission line.

The ratio of the applied voltage to initial current is an impedance and we call this the characteristic impedance or Z_0 of the transmission line. This is the current that would flow into an infinite length of line no matter what is at the far end. Note that we didn't say anything about signal frequency here, this is strictly a matter of the way the line is constructed, especially the capacitance and inductance that is distributed along its length. For the usual low loss line, ignoring resistance, we can determine Z_0 as follows:

$$Z_0 = \sqrt{L/C}$$

where L and C are the inductance and capacitance per unit length, typically per foot or per meter. These values can often be found in many manufacturers' specification sheets.

The Matched Transmission Line

If, instead of having an infinite line, we put a resistor with a resistance equal to the line Z_0 on the far end of the line it will absorb the power as it arrives and there will be no reflection. The voltage and current relationship at the signal end of the line will appear as if it were the same resistor, just located at the source.

This is exactly what we wanted! If the source is a radio designed to drive a 50 Ω load, Z_0 is 50 Ω and our load is an antenna with an impedance of 50 Ω, we get just what we want — to our transmitter it appears as if the antenna is connected directly to the antenna. There is no reflected power, the SWR is 1:1 and all is well with the world.

The Mismatched Transmission Line

There are an infinite number of cases in which a transmission line is terminated, not with its Z_0, nor with an open or short, but with some other impedance. This results in a reflection of a signal that is smaller than the signal that was sent down the line. It is important to note that this signal does not represent lost power, because it is re-reflected from the transmitter back up to the line. It is true that the power delivered to the antenna will equal the forward power less the reflected power, however, because the power is re-reflected at the transmitter so the numbers go up together. This is illustrated in **Table A** for the case of perfect (lossless) transmission line, assuming the transmitter can put power into the SWR shown.

So What's the Big Deal

The key is the two reasonable assumptions noted above. Let's discuss each:

- Transmission lines are not lossless, and their losses increase with increasing SWR as we will discuss. Whether or not that's a problem depends on the line

Table A

Example of Net Transmitted Power with 100 W Transmitter versus SWR with Lossless Transmission Line

SWR	*Reflection %*	*Forward Power*	*Reflected Power*	*Antenna Power*
1:1	0	100	0	100
2:1	10	111	11	100
3:1	25	133	33	100

the variable capacitor between the two ends of the inductor to allow tuning of both high and low impedance loads.

This particular tuner uses a rotary inductor (see **Figure 3-5**) for adjustment of inductance. The coil turns with rotation of the shaft on the left end. As it turns, its point of connection to the small wheel on the top in the picture is moved up and down the coil. It requires about 30 turns of the knob to adjust from the minimum to the maximum amount of inductance. Some tuners use a switched inductance instead. The switch allows a quicker change of settings, but does not provide for values of inductance between the switched taps. This is not generally a problem, although there are some loads that can't be tuned to a 1:1 SWR at some frequencies — still close enough works. Note that there is no bandswitch, on this model, since any available adjustment type, frequency and SWR as we will discuss later.

- Often a more serious problem is that somewhere in the range of SWR in Table A, typically by an SWR of 2:1, transmitters will reduce power to avoid damage due to the higher current or voltage that results from the higher than specified SWR.

Note that the second problem is not a fundamental issue of SWR itself, but rather a design choice made by transmitter designers. Still, for transmitters that we may choose to use, it is a real problem that we have to deal with.

What Kind of Load Results in High SWR

Any load that is different from Z_0 of the transmission line will result in a SWR greater than 1:1. For the case of the frequently encountered 50 Ω coaxial cable, that means any load that is not 50 Ω resistive. This can mean a load that is resistive with a value of other than 50 Ω or a load that is resistive, but also has capacitive (–X) or inductive (+X) reactance, or any of an infinite number of combinations of the two. Some examples of loads with different SWRs are shown in **Table B**.

Note that it is very easy to determine the SWR for the resistive case — it's just Z_0/R or R/Z_0 depending on whether the load is greater or less than the Z_0.

Thus it is important to know not just the resistive part, but also the equivalent series inductive or capacitive reactance to determine the SWR.

Determining SWR from Impedance Data

While the case for resistive loads is simple, the case for loads with reactance or complex loads is well — more complex. Still, there are at least three ways that I know of to determine the SWR of a complex load, not counting measuring with an SWR meter, as follows:

- Use the software TLW (Transmission Line for Windows) that comes packaged with recent editions of The ARRL Antenna Book.[B] If you plug in the R and X values and the appropriate transmission line Z_0, it provides the SWR at each end of the line. You can select which end of the line has the measured Z (input or load). This is very handy since it will also calculate the line loss.
- This may be the most simple calculation to make graphically with a Smith Chart.[C] Recall that if the Z is just resistive (X=0), the SWR is either Z_0/Z or Z/Z_0, depending on whether the Z is lower or higher than the Z_0. Recall also that a circle on a Smith Chart represents constant SWR. Thus if you enter the Z on the Smith Chart and draw a circle centered on the chart center, it will show the transformed Z for any length of line. Either point at which the circle crosses the resistive axis can be used to calculate the SWR as described above.
- If you have neither computer nor Smith Chart, you can compute the SWR directly. This can be found, for example, in a book I had as a text many years ago, a classic by the late John Kraus, W8JK, *Antennas*. In the first edition it is in the appendix on page 507. The calculation is done in two steps:

1. Find the voltage reflection coefficient σ:
$\sigma = (Z - Z_0)/(Z + Z_0)$
Note that the Z is a complex number, so the calculation is a bit tedious.
2. Find the SWR = $(1 + |\sigma|) / (1 - |\sigma|)$
where $|\sigma|$ indicates the magnitude of the complex reflection coefficient, σ, found in step 1.

Table B

Impedance of Loads that Result in Different SWRs.

For Ideal 50 Ω Coaxial Cable

R (Ω)	*X (Ω)*	*SWR*
50	0	1:1
25	0	2:1
100	0	2:1
50	±35	2:1
30	±18	2:1
16.7	0	3:1
150	0	3:1
50	±58	3:1
30	±40	3:1
5	0	10:1
500	0	10:1
50	±142	10:1
250	±250	10:1

Notes

[A] W. Maxwell, W2DU, *Reflections*, check his Web site for information on availability of the latest edition at **w2du.com**.

[B] R. D. Straw, Editor, *The ARRL Antenna Book*, 21st Edition. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9876. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

[C] Paper copies of 50 Ω Smith Charts are available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 1341. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

is usable on any band — you just crank until you are within range. It may even be helpful for upper body development.

The capacitor can be a single, or multiple section type or a combination of a variable and fixed capacitors. In my tuner, the switch that moves the capacitor from one end of the inductor to the other has additional positions that switch in fixed capacitors in parallel to extend the tuning range. The fixed and variable combination allows somewhat finer tuning, but in multiple steps. A larger variable with a multi-turn vernier dial may be equivalent, and may be quicker to adjust.

An advantage of the L-network tuner is that there is only a single combination of L and C that will match a given load. As noted below, some tuners can offer more than one, but only one is optimum.

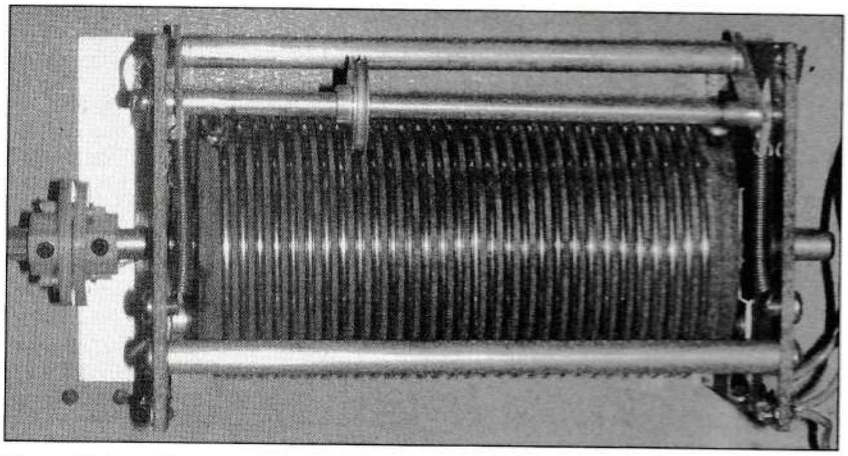

Figure 3-5 — Close up of a rotary inductor.

Pi- and T-Network Tuner Controls

Pi- and T-network tuners have three tuning controls — usually one inductor and two capacitors. Pi-network tuners that started out in the age of vacuum tube transmitters with pi-network output circuits often are built with similar components and are thus bandswitched as in the Drake MN-2000 tuner shown in **Figure 3-6**. In this tuner, the bandswitch sets the inductance to a preset coil tap for the middle of the expected range and the capacitors on both ends are used to provide the closest match. This configuration can be useful, but generally has less tuning range than those with adjustable or switched inductors in small increments.

Note that a pi-network tuner set to minimum capacitance on either of the capacitors essentially becomes an L-network tuner, with the high or low impedance determination dependent on which capacitor is set to around zero.

With either the pi- or T-network tuners, sometimes multiple very different settings can be found that will tune a particular load. In general, a setting with the minimum inductance is more efficient since the wire resistance of the inductor can dissipate power.

Tuned Transformer Tuner Controls

The tuned transformer is essentially the same as an output circuit from an earlier, before the pi-network, kind of transmitter. It typically has a balanced tuned circuit designed to resonate on a particular band (there weren't as many back then) using a fixed inductor and variable capacitor. The typically balanced transmission line was tapped symmetrically on the coil — close to the center for low impedance loads, close to the ends for high impedance loads. Small clips were generally used to provide load connection points. Very low impedance loads were connected in series with the center of the inductor.

Any reactance in the load would detune the tuned circuit that would then be readjusted by retuning the variable capacitor. The usual transmitter side connection was a link of a few turns that could be mechanically moved in and out of the center of the coil to provide adjustable coupling.

It was actually simpler to adjust than describe. Once the tap points were located for a particular band using the usual single antenna, the clips were fixed in place, and this became just one more coil to be changed while changing frequencies. Both transmitters and receivers of the pre WW2 era often had plug-in coils that required changing each time a band change was desired.

The 1950s era E. F. Johnson *Matchbox* antenna tuner (see **Figure 3-7**) was a radical design change to the tuner transformer type of tuner. This was the era in which transmitters and receivers were generally bandswitched and the Matchbox was an effort to make antenna tuning just as easy. What made the matchbox unique was a patented differential capacitor design that allowed a variable capacitance to be made of two capacitors — one that increased and one that decreased as it was rotated.

A pair of these were placed across the tuner resonant circuit allowing the transmission lines in the middle of each to be tapped across the resonant circuit using capacitive "taps" instead of the previous inductive taps. This allowed a continuously adjustable impedance setting to be made from the front panel. The bandswitching allowed bands to be changed without changing coils. This was very popular at the time, and they are still popular today. They do have some limitations, principally that they only work on the amateur bands of that day, not 30, 17 and 12 meters — although they may

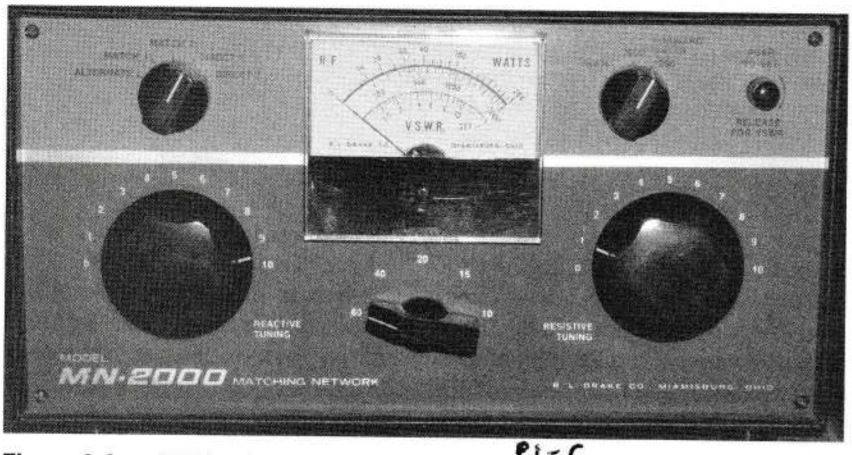

Figure 3-6 — 1970's vintage bandswitched ~~pi-network~~ PI-C tuner.

THE DRAKE MN-2000 IS NOT A PI-TYPE CIRCUIT BUT A PI-C CIRCUIT WITH THE INPUT CAPACITOR AND THE COIL BEING FIXED (BAND-SWITCHED) AND THE TWO OUTPUT CAPACITOR BEING VARIABLE. THIS IS VERY CLEVER DESIGN, THE Q IS

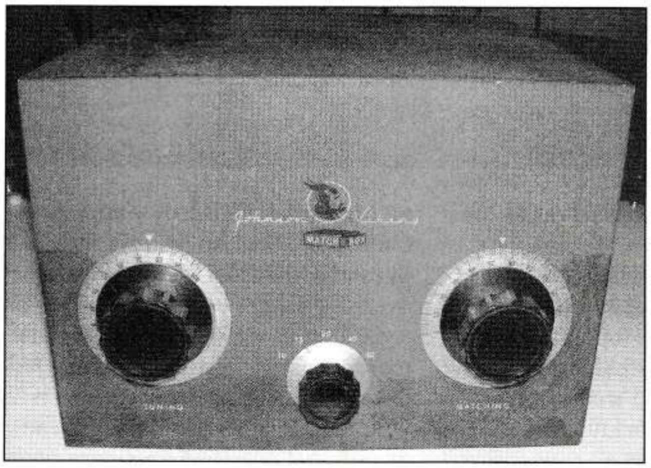

Figure 3-7 — 1950's vintage E. F. Johnson *Matchbox* antenna tuner.

drive some loads on the higher bands by setting to an adjacent band.

Antenna Tuner Metering

Many antenna tuners include metering, either directional power meters or an SWR bridge. These can be used to monitor progress while adjusting the antenna tuner — either for minimum reflected power, or for minimum SWR. Arguably, this is somewhat redundant if the connected radio transmitter also has SWR or reflected power measurement capability, although some meters in tuners may be more sensitive or easy to read.

While there is certainly no harm in having redundant metering, if your radio meter is easily visible from the antenna tuner location, and has sufficient sensitivity to allow proper tuning, there isn't much benefit in my opinion. The early transceivers, and especially pre-transceiver transmitters, did not generally include that function, perhaps explaining why most tuner manufacturers still provide them. For most users with metering in the transmitter, transmission line current meters on the output side of the tuner would actually be more useful. To my knowledge, alas, no tuner manufacturers now include these.[1,2]

Tuner Limitations

All tuners are composed of combinations of inductors and capacitors in various combinations. Since the impedance at a high SWR can vary widely, the tuner needs to have a wide range of values in order to be able to accommodate all possibilities at every frequency within its range. It is generally the case that the lower the frequency, the more inductance and capacitance will be required. Thus the maximum available inductance and capacitance, generally determines the low frequency limit of operation.

In the case of high power tuners with large components, it is not unusual for tuners that cover down to 160 meters (1.8 MHz) to be in larger enclosures simply because the size of the components is larger. In a similar way, it is the minimum capacitance and inductance that sets the high frequency limit. In this case, one must often look beyond the specified values of the components themselves and look to the stray capacitance of wires and component frames to the cabinet, as well as the inductance of connecting wires between components. Some tuners have trouble matching loads on 10 meters (28 to 29.7 MHz in the US) that would be easy to match at lower frequencies for this reason.

Increasing power with high SWR can result in very high currents and voltages within the tuner, requiring large components for high power tuners. The large components, in addition to making the box much bigger, also make the control of stray inductance and capacitance more of a challenge for the designer.

---

[1] Notes appear on page 9.

CONSTANT (18) OVER THE TUNING RANGE AND THERE IS ONLY ONE MATCH FOR A GIVEN IMPEDANCE IN CONTRARY TO THE T-TUNER.

Where are the Knobs on an Automatic Antenna Tuner?

Automatic antenna tuners don't generally have tuning knobs. Instead, they have a microcontroller that monitors SWR or reflected power and uses an algorithm to step through allowable values of L and C until a match is found. Any of the tuner topologies in Figure 3-3 could be used in an automatic tuner with an appropriate control mechanism.

Some automatic tuners use rotating variable capacitors and rotary inductors, just as in many manual tuners. In fact some are adaptations of manual tuners with servo motors and controllers added on as an additional physical and logical layer.

I do think it is safe to say, however, that the majority of auto tuners use a set of discrete components instead of variable ones. A range of inductance is provided by using relays to switch different sizes in series to reach the desired value. Similarly, a range of capacitance can be provided by having multiple capacitors switched in parallel. By having values available in a series such as 1, 2, 4, 8, 16, 32, 64, 128, 256 for example, any value from 1 to 511 (either pF or µH) can be obtained in steps of 1 unit. Thus any needed resolution and range can be easily provided.

While such switching, if done manually with toggle switches, might make us go over the top, it's duck soup for a microprocessor. The hard part may be defining a computer algorithm that can quickly find the right values without trying all combinations. Fortunately, that just needs to be done once and is provided "free" with the tuner. Most auto tuners can find a suitable match within a defined threshold, often 1.5:1, within a few seconds with no additional information about the antenna being tuned.

Many auto tuners have the capability to memorize the appropriate settings for a particular load once it's found. There are two approaches to this, both can be quite successful:

Memory stored by frequency. Perhaps most commonly found in tuners is the storage of tuning data based on frequency. Of course, this requires that the tuner measure the frequency before it can tune.[3] Most tuners have a number of frequency "bins" of a certain bandwidth, for example perhaps 25 kHz. If the frequency is within that bin, the last successful setting from a frequency within that bin will be used. If it doesn't meet the threshold, the tuner will start tuning over.

Obviously, the smaller the bin, the more successful the tune is likely to be, assuming the same antenna is used. On the other hand, this depends on previous operation within the bin, so it takes longer to accumulate data for all bins. Most tuners seem to have it about right, in terms of trade off.

Memory stored by past success. A different approach is to just keep track of the last number of successful tuner settings, independent of frequency. Let's say the tuner remembers the last 100 successful tuner settings. It just very quickly tries all those before it initiates its regular tuning algorithm. It is quite likely that recent conditions will be repeated. This approach is particularly advantageous for a single port tuner that is switched between multiple antennas making multiple successful tuner settings appropriate within a single frequency bin. Of course an advantage for the manufacturer is that the tuner doesn't need to measure or otherwise determine transmitting frequency.

What's the Best Tuner for You?

The best tuner for a particular application will depend largely depend on your antenna system and operational needs. Automatic antenna tuners inside transceivers are convenient and don't require extra space or wiring. Check the specs carefully, many do not cover a wide impedance range (many are spec'd at 3:1, although some may do better than spec). Interestingly, the internal tuners from Elecraft, FlexRadio and Ten-Tec do provide wide range (to 10:1 SWR) tuning, perhaps others will follow in this direction. The other disadvantage of internal tuners is that you will need a tuner with a higher power rating if you use, or consider using a linear amplifier. A high power tuner will work fine at low power, if the linear is in your plans.

Manual Tuners often offer antenna switching and control functions, not present in many auto tuners. Manual tuners also often provide continuous tuning to allow setting the controls for an exact 1:1 SWR, although that doesn't really provide a significant benefit.

Automatic tuners have a number of advantages and not many disadvantages. Some can be operated remotely at or near the antenna to reduce system losses. Either of the memory tuner types can reduce the tuning time for previously encountered antennas from a few seconds to a few tenths of seconds. Arguably either is significantly quicker to get into operation than a manual tuner, even with recorded dial settings.

The rest of the book will explore these issues in more detail.

Notes

[1] E. Nichols, KL7AJ, "Keeping Current with Antenna Performance," *QST*, Feb 2009, pp 34-36.

[2] P. Danzer, N1II, "A Simple Transformer to Measure Your Antenna Current," *QST*, Sep 2009, p 35.

[3] Internal, or brand specific tuners have the advantage that the frequency information can be made instantly available from the radio circuitry.

Review Questions

3-1 What are two reasons to provide a matched, or nearly matched, load to a transmitter?

3-2 What is the SWR of a 50 Ω cable with a 500 Ω resistive load? How about a 500 Ω transmission line with a 50 Ω resistive load?

3-3 If a 100 W output transmitter is connected to a lossless antenna system with a 3:1 SWR and it can deliver its full power, how much power reaches the antenna? How much power is reflected from the load?

Chapter 4

Tuning an Antenna Tuner

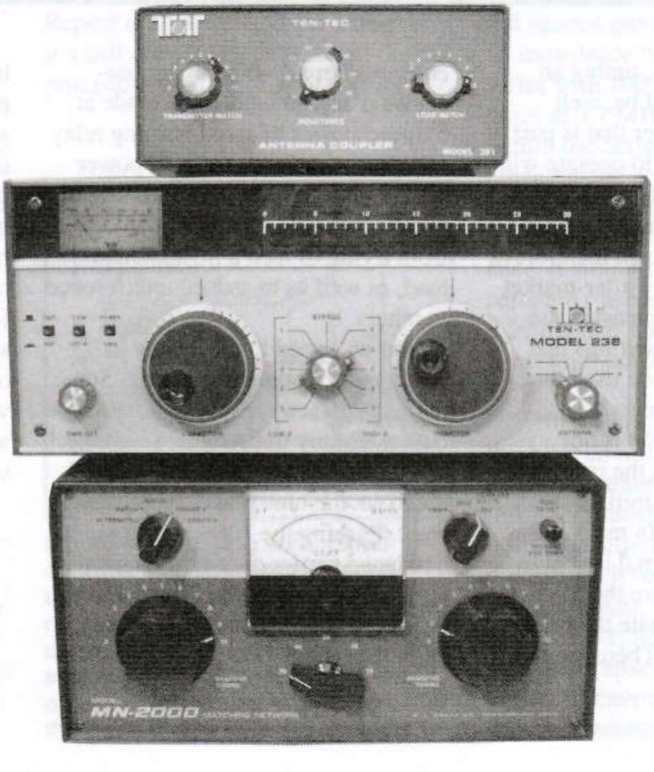
TEN-TEC
ANTENNA COUPLER
TEN-TEC
MODEL 238
MN-2000

Tuning controls of three manual tuners. On top, a 100 W Pi-network, in the middle a 1500 W L network tuner and on the bottom a bandswitched 1000 W Pi-network tuner.

Contents

Tuning an Automatic Tuner 4-2
Tuning a Manual Tuner 4-3
A Simple Transformer to Measure Antenna Current 4-6
Review Questions 4-7

By now we, hopefully, know what an antenna tuner is, and why it may be a good idea. The next question should be "how do we work it?" The answer to this depends a bit on the type of tuner and your configuration. In all cases, the most important rule is: *Don't interfere with other users of the band!* This is an issue because all types of tuners require the transmission of RF for at least the final adjustment.

This is because tuning is considered complete only if the SWR is close to 1:1, and it takes a signal to be able to measure SWR. Even if the signal is of a reduced amplitude, people have been known to make worldwide contacts with transmitters running less than 1 W, so don't assume a low powered TUNE signal won't go anywhere.

If you are worried about loosing track of the current frequency that has a particularly desirable station; just put it in memory, tune off the frequency, adjust the tuner and, when tuning is complete, recall the desired frequency from the memory. You should be good to go, and you won't ruin it for someone else.

Tuning an Automatic Tuner

As its name implies, tuning an automatic tuner should be, well *automatic*. If it's a tuner that is part of the radio, or designed to operate with the radio, there is likely a TUNE button that reduces power, sends a carrier on the desired frequency and then forces the tuner to tune. Many after-market automatic tuners first measure the frequency and try the settings that worked the last time you used that frequency, or one close to it.

If the tuner is one that is not designed to work with the radio, you generally will need to initiate the process in some way. In many cases, if a transmitted RF signal is sent to the tuner it will measure the SWR and automatically initiate the tuning process if it needs to. There are two concerns here — first, many tuners want adjustments to be made at reduced power to avoid burning relay contacts — second, reduced power should also be used during tuning to avoid potential transmitter damage from trying to feed a mismatched load, as well as to reduce interference to others.

You will need to find a way to easily reduce power for tuning. Some radios do exactly that with a TUNE button. On my transceiver, I have such a button and a menu selection allows me to set the tune power to any level I want. Pushing the TUNE button sends a reduced carrier out of the transmitter until I push it again. This button is also useful to start tuning a linear amplifier. If you don't have a TUNE button, you may need to manually reduce power and then hit a key to send a signal. Note that I said *key*, as in radiotelegraph mode. Hitting the push-to-talk button on a SSB voice transmitter will switch to transmit, but not put out a signal until you start talking — so that doesn't work well for this application. If your transmitter supports AM voice, it will usually put out a carrier of about 25% of full PEP when keyed, so that is a possibility as well, although that is usually somewhat more power than desired.[1]

[1] Transmitters such as the early R. L. Drake vacuum tube series use controlled carrier in AM mode and may not put out a carrier without speech.

Tuning a Manual Tuner

To tune a manual tuner, you can actually start with the transmitter turned off. First look at the instruction manual and see if it provides suggested settings for each band. While your actual settings will be somewhat different, since you will have a different impedance on each band, they should make a good starting point. Hopefully the controls on your tuner have calibration marks on each control, if not, imagine an analog clock face on each control.

Next, practice tuning your tuner into a dummy load. Even though the tuner usually isn't needed to match the 50 Ω dummy load to the transmitter, if the tuner is in line, the tuner will require just as careful adjustments to tune into that load. Try the tuner on each band and record the settings.

Adjusting a Manual Tuner

Most manual antenna tuners have three controls. If it's a Pi- or T-network, they will all be adjustable inductors or capacitors, usually one of the former and two of the latter. An L-network will just have one of each but will usually have a switch to move the variable capacitor between the input and output side of the inductor as shown in Figure 3-3 in Chapter 3. Because the three controls interact to some degree, finding the right settings can be arduous, can stress the transmitter and can also cause interference to others as you try to find the right settings, hence the recorded table. I provide the following suggestions.

- Many antenna tuner instruction manuals have tables of typical values for each band. Use these as a starting point.
- Practice tuning your antenna tuner into a dummy load, as discussed above.
- First tune as close as you can using the receiver as an indicator. If there are signals on the band, adjust for maximum S-meter indication. If there are no signals adjust for loudest band noise or meter indication.
- If you have an antenna analyzer, use it instead of the radio to find settings that offer a matched load to the radio port as shown in **Figure 4-1**.
- If you can't quite get to a 1:1 SWR, shift one of the controls off its minimum position and adjust the other one (or two) for minimum. If the result is better, move the first control a bit more in the same direction. Repeat until it starts to get worse. If it's still not satisfactory, repeat the procedure with a different control.

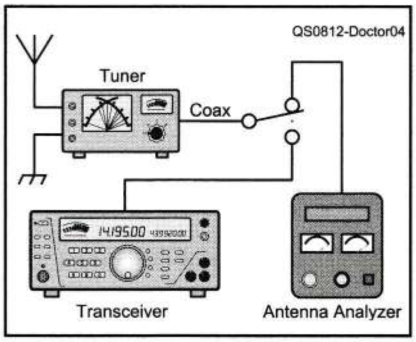

Figure 4-1 — Using an antenna analyzer to adjust your antenna tuner reduces both stress to your radio and interference to other users of the band. The coax switch shown can be used to switch the output of the antenna tuner to either the radio or the antenna analyzer.

- If there are multiple settings yielding a match, as there will likely be with T- and Pi-network tuners, select the one with minimum inductance — it will be more efficient because inductors tend to have more loss than air variable capacitors. The following example illustrates the point.

A fairly typical "random" length of wire likely to be matched with an antenna tuner would be 57.5 feet long fed against ground. This would have an impedance of 25.6 Ω resistive in series with 100 Ω capacitive reactance at 3.7 MHz — a 10:1 SWR, just within the design range of most wide range tuners. A low-pass L-network tuner will just have a single value of L and C that would mach this impedance. It would be, according to *TLW* software (discussed later in detail), 5.29 µH and 1000 pF. The loss in the L-network tuner would be 0.11 dB or 2.5%.

With a high-pass T-network tuner, there are an infinite number of combinations of tuner components that can result in a satisfactory match. Setting one value requires the other two components each be set to a particular value to obtain a match. As shown in **Table 4-1**, listed by selected value of output capacitor, not all combinations will have the same efficiency. It shows dramatically that by having a

Table 4-1

Examination of High-Pass T-Network Tuner Efficiency as a Function of Tuning: Load 25.6 Ω Resistive in Series with 100 Ω Capacitive at 3.7 MHz.

Capacitor Output (pF)	*Intput (pF)*	*Inductor (µH)*	*Loss (dB)*	*Loss (%)*
5	9.2	130.6	8.5	85.9
10	12.5	82.1	5.4	71.3
30	25.7	34.4	2.4	41.7
50	37.5	22.5	1.6	29.9
100	63.2	12.8	0.9	18.4
200	102.6	7.8	0.5	11.5
500	171.2	4.7	0.3	6.8
1000	223.2	3.7	0.23	5.2
2000	264	3.2	0.19	4.4
5000	297	2.6	0.17	3.9

Table 4-2

W1ZR Antenna Tuner Settings Table

Antenna Tuner Settings — 100 Ft Dipole (4)

					AMP SET	
FREQ	*C*	*IMPED*	*L*	*SWR*	*Load*	*Tune*
1.8	8.0	L5	4 - 9.9	2.5	0.0	8.0
3.55	5.6	H1	20 - 6.9	12.0	4.0	5.8
3.8	1.0	H1	18 - 1.9	9.8	8.8	4.7
5.3	4.3	H1	12 - 9.5	5.8		
7.0	1.2	H2	6 - 5.0	4.6	7.8	2.7
7.2	8.5	H1	8 - 1.0	5.6		
10.1-.15	8.8	H1	4 - 6.3	4.4		
14.0	2.0	H1	4 - 5.6	4.0	7.0	1.5
14.2	1.3	H1	4 - 6.4	3.7	7.0	1.4
18.068-.168	4.1	H1	2 - 7.7	2.9	7.0	2.0
21.0	3.0	L1	2 - 6.5	3.7	6.2	1.5
21.2	2.6	L1	2 - 6.2	4.3		
21.3	2.5	L1	2 - 6.3	4.5	6.0	1.3
24.890-.990	1.5	H1	1 - 1.8	1.8		
28.0	1.0	L1	3 - 5.4	1.6	5.0	2.0
28.5	1.0	L1	2 - 9.6	2.5	4.5	2.0
29.0	1.0	L1	2 - 6.0	3.6		

Antenna Tuner Settings — Tri-Band (2)

					AMP SET	
FREQ	*C*	*IMPED*	*L*	*SWR*	*Load*	*Tune*
10.1-.15	4.0	L1	9 - 1.3	5		
18.068-.168	3.3	L1	4 - 8.2	5	7	2.0

Antenna Tuner Settings — Ground Plane (3)

					AMP SET	
FREQ	*C*	*IMPED*	*L*	*SWR*	*Load*	
3.6	10.0	L1	9 - 1.7	2.3	4	5.8

high inductance value, the loss will be higher than if an inductor with a lower value were used. All combinations will provide a 1:1 match to the transmitter which will happily pump power into the system. While some values would likely not be realizable at 3.7 MHz, the reactances all would be easily employed at 30 MHz. A 1500 W transmitter with the 86% loss of the first entry would dissipate 1290 W in the tuner as heat. The tuner would likely melt before your eyes!

Note that Table 1 indicates that the highest efficiency is with the lowest inductance value — often the case. It also can happen that the high Q of a low inductance solution can result in large circulating currents resulting in additional loss. The best way to be sure is to measure the relative current going from the tuner to the antenna. The solution with the highest current will be the most efficient.

The actual value of the current isn't usually significant since the antenna impedance is unknown. For any given antenna on a given frequency, what is important is that you have the solution that provides the most current towards the antenna with the transmitter operating into its design load and within ratings. Unfortunately, I am not aware of any tuners now available that provide for the measurement of antenna current. RF ammeters are available on the surplus market and one is available from MFJ. Alternately, a fairly simple antenna current transformer project article in *QST* described how to make your own. It is reproduced in the following sidebar.

Once you have a match, reduce the inductance, retune the capacitors, and see if you can still obtain a match with an even higher current. Once you have the settings for a particular antenna, you won't need to measure the current again unless something changes in the system.

- Record your data! All manual antenna tuners I've seen have calibrated scales on their front panel for each control. As soon as you have a match, write down the settings you found for future reference and use them for a starting point next time. I find that having a set of dial readings for at least the phone and CW portions of the band is useful. **Table 4-2** is the actual table that I use at my station. Note that there are separate table segments for three different antennas. The fourth position is connected to a dummy load for testing and amplifier adjustment purposes.

The Controls in My Table

Since there's a bit more here than you might need, let me explain the columns I have included in my chart. The tuner I use is an early Ten-Tec 238 with front panel layout shown in the chapter title (center) figure. This is a 1.5 kW rated L-network tuner that provides four antenna connection ports, one of which (# 4) can be set up for either a balanced or unbalanced load using an internal balun on the output side.

- FREQ is the approximate frequency that I made the adjustment for.
- C is the approximate setting of the variable capacitor dial. It has 10 divisions, but I "eyeballed" how far between the marks the setting was to get the tenths position.
- IMPED is a switch that can set the configuration for high (H) or low (L) impedance and has five fixed capacitor settings for each, hence the numbers 1 through 5, in addition to the letter.
- L is the setting for the variable inductor. The inductor setting has two numbers because it is a multiple turn rotary inductor. The first number is the value of the TURNS scale, while the second is the knob setting from the skirt on the knob.[2]

---

[2] Current models of this tuner have an improved drive for the rotary inductor that provides a single reading for the L value.

- SWR is the SWR measured with the tuner bypassed. This doesn't play a part in the tuning, but is useful as a check to make sure the antenna hasn't changed or fallen down since it was installed. The tuner will just as happily tune the antenna if it breaks and falls down, but the signal won't go as far.
- The two AMP SET adjustments are for my linear amplifier tuning controls. They share the spreadsheet because, as with the antenna tuner, if I change bands or modes, they need adjustment. These columns save having an additional piece of paper on the operating desk. If you don't have a linear amplifier you don't need these.

Your tuner may have different control labels, and you may not want to include all my data, however, it should give you a starting point for your own record keeping. I maintain my records on a PC spreadsheet, although any convenient medium can be used. A small set of data might fit on a 3 × 5 (or perhaps 5 × 7) inch file card. The spreadsheet has the advantage that it can be easily updated as antennas are changed or added. In my experience, amateur stations are never quite done!

A Simple Transformer to Measure Your Antenna Current

SWR doesn't give you the whole story — you need an RF current meter.

Paul Danzer, N1II

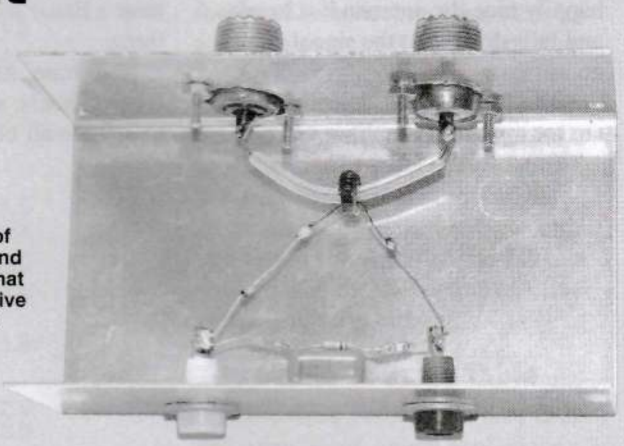

Figure 1 — View of the transformer and simple circuitry that make up the relative RF current meter.

In a recent *QST* article, Eric Nichols, KL7AJ, presented a good idea — monitoring the RF current into your antenna system to insure optimum tuning.[1] Actually, it was the second time it was mentioned to me — George Peters, K1EHW, suggested the same thing to me several months before. At the end of Eric's article, he proposed using a current transformer to do the monitoring.

Making a Current Transformer

This could be as simple as a turn of wire through a ferrite core and several turns of wire around the core to form a transformer.[2] The output of the transformer would be proportional to the current through the wire.

Making it Happen

This seemed simple enough, and the result of one hour's work is shown in Figure 1. The core used was a T37-6. T37 designates the size (0.37 inches OD), picked so the insulated center conductor of RG-58 (or RG-59) would fit comfortably through the core center. The –6 designation relates to the frequency application of the ferrite mix, in this case 2 to 50 MHz.

Searching the ARRL suppliers' data base, it appears that Alstar Magnetics offers this core; an alternate would be a Palomar F37 with mix 61. There is no criticality here — if you want to try it, strip a core from any old source — perhaps from a junked PC power supply or computer cable. It may not be the most efficient RF transformer ever built, but if it works it will do the job.

Wrap 20 turns of 24 gauge enameled wire as the transformer secondary. The secondary is connected to half wave rectifier consisting of a silicon diode (1N914), a 10 kΩ resistor as the load and a 0.1 μF capacitor as a filter (see Figure 2). A high-impedance voltmeter (the $10 variety) is connected to the two pin jacks to serve as an indicator.

The unit shown was tested with a 100 W transmitter on all bands from 80 through 10 meters. Performance across each band was relatively uniform, considering the probable variation of SWR and power though the feed line as the frequency was varied.

Hook up the rest of the circuit as shown in Figure 2 and connect a meter to the terminals.

What we Have

The object here was not to get an exact measure of the antenna current. What I wanted was a relative measure, so I could see if anything was going wrong, or use it as a way to adjust my antenna tuner for the maximum signal to the antenna. Commercial stations use a current meter, mounted at the connection of the feed line to the antenna, to monitor output. Since they know the antenna impedance and their meters are calibrated, they can determine precise power into the antenna.

If you enjoy low power (QRP, typically 5 W or less) or very low power (QRPP, less than 1 W) operation, more turns may be needed and can easily be added.[3] Similarly, if your meter does not have enough sensitivity, more turns may be called for. If after assembly the core is not firmly in place, held by the friction of the secondary on the primary wire, a drop of glue can be used to secure it all together.

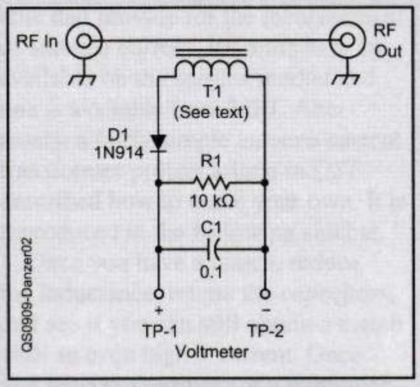

Figure 2 — Schematic diagram of the relative RF current meter. Nothing about the circuit is critical. See text for parts information.

Putting it to Use

The current meter consumes a miniscule fraction of the output power, so, can be left in the line, or removed when not in use. You may even find a meter case and a surplus meter at a hamfest that will work with it to give continuous indication without tying up your bench meter. I suggest writing the relative current indication for each band in your log. Later, if something seems amiss, it is then an easy job to compare your readings to the recorded ones to find out if the problem is in your antenna system.

Notes

[1] E. Nichols, KL7AJ, "Keeping Current with Antenna Performance," *QST*, Feb 2009, pp 34-36.

[2] Each pass through the center of a toroid counts as a full turn.

[3] Low power operators may obtain better results using a germanium diode, such as a 1N34, in place of the silicon diode due to its lower forward voltage drop.

ARRL Member Paul Danzer, N1II, was first licensed in 1953, and now holds an Amateur Extra class license. Paul has been operating 40 meter CW almost constantly since he first started. He uses his years of experience as an electronic engineer to design and build small, one-night ham radio projects. Currently he is a Professor of Computer Science at Housatonic Community College in Connecticut. He can be reached at **n1ii@arrl.net**.

From September 2009 QST © ARRL

Review Questions

4-1 Why is it important to listen to the operating frequency before you start the tuning process?

4-2 Why is it a good idea to manually initiate tuning with your auto-tuner before transmitting on a new frequency at full power?

4-3 If there are more than one set of tuner settings that provide a 1:1 SWR, which should you choose?

Chapter 5

The Internal Tuner — How Does it Help?

The Yaesu FT-2000 transceiver, one of many modern transceivers equipped with an automatic internal antenna tuner.

Contents

The Flavors of Internal Tuners........5-2

Tuner Impedance Range........5-2

Impedance Limits of Internal Tuners........5-3

That Coax Connector........5-3

Coax Loss — If Things Seem Too Good to Be True........5-3

What Else is Going to be in Your Station?........5-3

Review Questions........5-4

So far, our discussions about antenna tuners have made it appear that they are a separate device that is inserted between a transceiver and an antenna system. While that's often true, we should also consider the case of a transceiver that has an internal, almost always automatic, antenna tuner.

The idea of an automatic antenna tuner is that it can sense the SWR and adjust the tuner to minimize it. This is most commonly accomplished by a configuration consisting of a collection of inductors and capacitors that can be switched in or out of a matching network until the lowest SWR is obtained. A few automatic tuners actually rotate the shaft of a variable capacitor using an electric motor as the driving element.

The Flavors of Internal Tuners

Some transceivers offer an internal tuner as an extra cost option, in many cases one that can be added if needed. In other cases, the transceiver with tuner is essentially a separate model that has to be specified at the time of purchase. Other transceiver lines come only with or only without an internal tuner. Thus, if you're in the process of deciding on your next (or first) transceiver it's good to think carefully about your requirements before you make a decision. **Figures 5-1 – 5-3** illustrate radios with some of the choices.

Tuner Impedance Range

As noted previously, all antenna tuners have limits on the range of impedances that they can transform to the desired load — almost always 50 Ω. In many cases the range of values is dependent on the frequency, often narrower on 6 meters than on the HF bands. Most tuner designs are limited on the low frequency end by the maximum values of the inductors and capacitors, and on the high frequency end by their minimum values, often the minimum capacitance of the elements to the metal case.

Thus it is often the case that the range of impedance values that the tuner is able to match is greater in the middle of the range than on the ends — typically 80 or 160 meters on the low end and 10 or 6 meters on the high end. Most manufacturers provide a single specification such as "tunes SWR at least 10:1" or "matches from 5 to 500 Ω." Note that the latter specification is not as general as the first — does it mean it only matches

Figure 5-1 — The Elecraft K-3 is an entirely modular design, including the option of a wide-range internal antenna tuner.

Figure 5-2 — The ICOM IC-7600 is only available with a built in antenna tuner.

Figure 5-3 — The Kenwood TS-480 is available as either a 100 W version with internal tuner (TS-480 SAT) or 200 W version without (TS-480 HX).

resistive loads? Probably not, since it would be hard to design a tuner that just did that, nor would it be very useful. I tend to think that both ratings actually mean the same 10:1 SWR.

If they deliver on their promise, that would mean that they would actually tune the 10:1 range on the highest and lowest bands, and actually tune a wider range in the middle — and many seem to do just that.

Impedance Limits of Internal Tuners

While most external tuners have rating similar to the 10:1 discussed above, many transceivers have internal tuners rated at a 3:1, or 16.5 to 150 Ω tuning range. This seems to be particularly true of the radios provided by the large Japanese companies — ICOM, Kenwood and Yaesu. In contrast, the US made HF transceivers, Elecraft, FlexRadio and Ten-Tec, offer internal tuners specified to tune a 10:1 range.

The tuners with a 3:1 range are intended to deal with a matched antenna operated somewhat away from its design frequency. A look at the SWR plots in Chapter 2 provides the idea. While a 3:1 range will easily cover a 40 meter dipole across the whole band, it will only cover about 300 kHz of the 500 kHz width of 80 meters with a standard 80 meter wire dipole (see Figure 2-7).

The tuners with a tuning range of 10:1 or greater are what I call *wide range* tuners. These tuners can not only deal with an 80 meter dipole across the whole band, but can be used to match a whole spectrum of antenna types ranging from random, non resonant, wires to dipoles fed on multiple bands and exotic high gain antennas such as V beams or rhombics.

That Coax Connector

All current internal tuners terminate on the back of the radio in a coax connector, just as do radios without a tuner. This is a signal that they are intended to drive into coax fed *unbalanced* antenna systems. We will discuss the various benefits of balanced and unbalanced transmission lines a bit later, however, it is worth pointing out that many external tuners are set up to drive either. If your antenna system is balanced, you will need to make the transition from coax to balanced line outside the radio, tending to reduce the "all in one box" benefit of the internal tuner.

Coax Loss — If Things Seem Too Good to Be True...

It is tempting to try to use the antenna tuner to "force fit" an antenna to work on a band other than the one it's designed for. This can often appear to work, but appearances can be deceiving. Let's take one example.

Say you have a 20 meter dipole fed with 100 feet of RG-58 coaxial cable. The antenna works very well on 20 meters, and requires just a bit of trimming with the tuner for proper operation. So far so good.

All of a sudden sunspots reappear after years of absence and 10 meters is wide open with great signals world wide (sooner or later it's bound to happen). You have no antenna for 10 meters, so try using your 20 meter dipole. Wonder of wonders, it actually tunes with the internal tuner — a perfect match! You hear your neighbors working distant stations right and left, but no one hears you — what's the story?

This is an interesting case that provides a cautionary tale, as well as outlines one of the challenges of designing multiband antennas. On 20 meters, our dipole (depending on height) will have an SWR of around 1:1. The 100 feet of RG-58 coax will have a loss of about 37% (less than 2 dB), not noticeable in most cases, and all will be fine.[1]

Move to 10 meters and the 20 meter dipole will no longer provide a good match. In fact, my model indicates that its impedance will be 3836 Ω resistive in series with 790 Ω capacitive reactance — an SWR of 81:1! The high SWR increases the loss of the transmission line to about 15 dB. That means that of our 100 W, only 3.1 W will reach the antenna — that explains why we aren't being heard as well as our neighbors!

The perhaps more interesting aspect of this is that the power reflected by the mismatched antenna also is reduced by 15 dB on its way back to the tuner. Because of this, the tuner only sees an SWR at the bottom of the cable of about 3:1 and will happily adjust for a match. Thus the radio is happily transmitting, but 97% of its power output is heating up the transmission line.

What Else is Going to be in Your Station?

Another consideration as to whether an internal tuner will be your best choice is whether you have, or expect to have, a linear amplifier following your transceiver. For some types of operation, being able to select whether you transmit 100 or 1500 W can make a big difference.

If you have a linear amplifier and need a tuner, you will need an antenna tuner on the output side of your linear. A tuner in your transceiver can adjust for a mismatch in the input side of the amplifier (rarely needed), but can't help with tuning the antenna system. That has to be done at the output of the amplifier — and at the higher power level. If the amplifier is turned off, the radio's tuner can be used to match the antenna, but you will want to disable it when you switch to high power.

Thus, if you have or expect to have, a linear amplifier as a part of your station; you may want to consider a transceiver without a tuner and instead purchase a tuner with a power rating that can handle your linear. Note that this is mostly a monetary statement, since there is no harm in having the additional tuner in your transceiver. It can be turned off if you don't need it. I actually have one in my transceiver and another following the linear. I find the transceiver's internal tuner handy when I take my transceiver on vacation and am forced to use "antennas of opportunity."

---

[1] We will use decibels (dB) to express gain and loss ratios throughout the book. If you need a refresher on the topic, please see Appendix A.

Review Questions

5-1 What are some of the advantages of an internal antenna tuner in a transceiver?

5-2 What kind of antennas can be tuned with a tuner that can tune an SWR of 3:1?

5-3 Why isn't an antenna that can be tuned with an internal tuner always effective?

Chapter 6

An External Tuner at the Radio

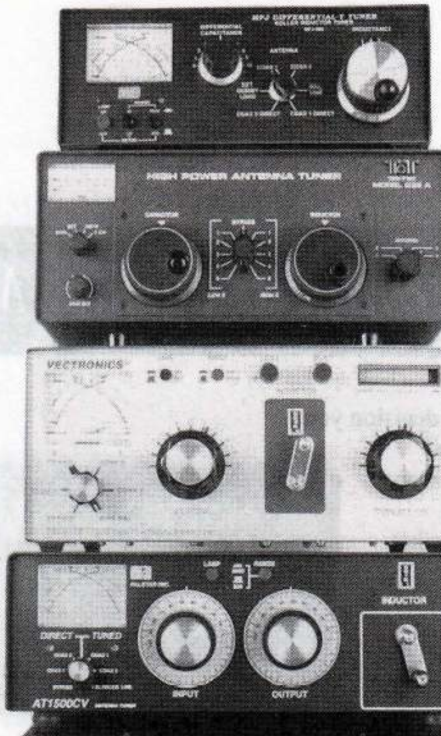
MFJ DIFFERENTIAL-T TUNER
HIGH POWER ANTENNA TUNER
VECTRONICS
AT1500CV
INPUT
OUTPUT
INDUCTOR
TRANSMITTER
ANTENNA

Five high power (1 kW or higher rating) manual tuners, one choice for an external tuner to place near the transceiver

Contents

Automatic versus Manual Tuners..6-2
Review Questions..6-4

In place of an antenna tuner that is built into the transceiver, another option is to have one next to the radio. Note that this is the functional equivalent of an internal tuner, and shares the potential problem of appearing to work well with an inappropriate antenna, as described in the previous chapter.

There are many available choices for such a device, including automatic tuners from some radio manufacturers that integrate directly with radio controls. There are others from after market manufacturers that offer additional features and, in some cases, are less expensive than internal tuners. While you do have to have an extra box on or near your radio, external tuners may offer some significant advantages:

- As discussed in Chapter 3, they can have a wider tuning range than many internal tuners.
- If you think you may want to add (or if you have) a high powered linear amplifier, you can select a tuner compatible with the higher power.
- Some external tuners offer both balanced and unbalanced outputs.
- Some external tuners offer multiple switched connections so different antennas can be quickly brought on line.
- If you have multiple radios, they can be switched to the tuner input so only one tuner is needed.

Automatic versus Manual Tuners

The first decision you will be faced with if you want an external tuner is whether you prefer an automatic or a manual one. As the name implies, an automatic antenna tuner finds the optimum, or near optimum, settings by itself. The manual tuner usually has three controls that require adjustment to obtain the best match.

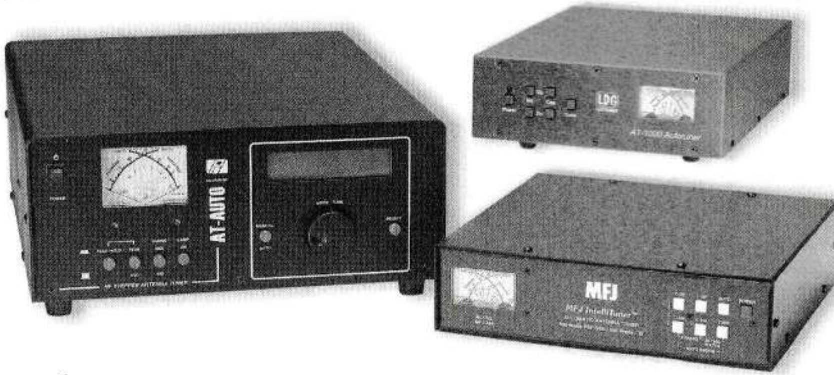

Figure 6-1 — High power (1 kW or greater) automatic antenna tuners from Palstar, LDG and MFJ).

In the past, automatic tuners were not available at high power levels, so those with high powered stations had no choice but to use manual tuners. That is no longer the case, as automatic tuners are available at all legal amateur power levels (See **Figure 6-1**). Over the years, the ARRL has tested many manual and automatic tuners as part of *QST* product reviews. I was surprised to find that to our level of measurement precision, there was not a big difference in efficiency between tuners in the two groups. Still, there are some benefits to manual tuners and many amateurs are happy with the results they receive.

Automatic Antenna Tuners

To operate an automatic, you just feed it some power, hit a TUNE button and listen and watch while it finds satisfactory settings. Many tuners can do this in tens of seconds the first time and often remember the settings so the next time it takes just a few seconds or less for that antenna at that frequency.

In my experience there's not a lot not to like about an automatic antenna tuner. I have a number of them for particular applications, although I still use a manual tuner for my main station tuner. The key is to select a tuner that will operate with your radio — some are radio independent, while some work with the same radio controls that would operate one from the manufacturer of the radio.

One issue with some automatic tuners is just what happens if they sense a mismatch as you operate. Most automatic tuners are rated to tune at reduced power to avoid excessive arcing at relay contacts. If you change frequency or even if the wind blows your antenna around and the tuner starts tuning, the results could be dramatic. Some allow disabling the auto tune function once you have it tuned, others automatically cut back on power. In my opinion, it's better to be able to take control yourself.

Manual Antenna Tuners

Manual tuners do often provide an additional level of control capability. Some have switch selectable outputs for multiple balanced and unbalanced antenna connections. One can be dedicated to a dummy load, if you don't (yet) have enough antenna connections to fill up the jacks. Manual tuners often provide SWR and power metering independent of that of the radio.

Most automatic tuners tune until they reach what they consider a reasonable match, perhaps 1.5:1. Hopefully it is a value that will allow your radio to operate at full power. The manual tuner let's you set it to whatever threshold you think is best — although if your transmitter can put out full power, it won't make much difference.

Review Questions

6-1 What are three potential benefits of using an external tuner at the radio rather than an internal tuner in the radio?

6-2 What are the benefits of an internal tuner?

6-3 What limitation is common to either internal or external tuners located at the radio?

Chapter 7

Transmission Lines

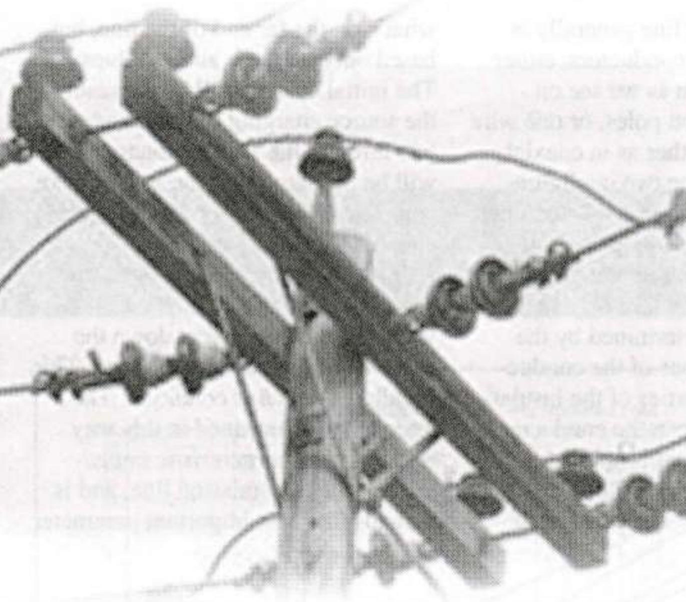

Transmission lines come in many forms, serving many applications.

Contents

Characteristic Impedance 7-2

Propagation Velocity 7-3

Attenuation 7-3

What Can We Do to Reduce Transmission Line Loss? 7-5

Sidebar: Determining the Input Impedance of an Unmatched Transmission Line 7-6

Review Questions 7-7

As we have mentioned previously, frequently the antenna and radio are not in exactly the same place. There are some notable exceptions, particularly in portable hand-held systems and various microwave communications and radar systems. In most other cases, optimum performance requires the transmitter and receiver to be at some distance from the antenna. It may also be a matter of combat survival, especially if your enemy is equipped with anti-radiation weaponry designed to home in on a signal. The component that makes the interconnection is called a *transmission line*.

Transmission lines are used in places besides radio systems — for example, power distribution lines are a kind of transmission line, as are telephone wires and cable TV connections. In addition to just transporting signals, transmission lines have some important properties that we will need to understand to allow us to make proper use of them. This section will briefly discuss the key parameters.

Characteristic Impedance

A transmission line generally is composed of two conductors, either parallel wires such as we see on power transmission poles, or one wire surrounding the other as in coaxial cable TV wire. The two configurations are shown in **Figure 7-1**. Either type has a certain inductance and capacitance per unit length and can be modeled as shown in **Figure 7-2**, with the values determined by the physical dimensions of the conductors and the properties of the insulating material between the conductors. If a voltage or signal is applied to such a network, there will be an initial current flow independent of what's on the far end of the line, but based only on the L and C values. The initial current will be the result of the source charging the shunt capacitors through the series inductors and will be the same as if the source were connected to a resistor whose value is equal to the square root of L/C.

If the far end of the line is terminated in a resistive load of the same value, all the power sent down the line will be delivered to the load. This is called a *matched* condition. The impedance determined in this way is called the *characteristic impedance* of the transmission line, and is perhaps the most important parameter associated a transmission line. Common coaxial transmission lines have characteristic impedances (referred to as Z_0) between 35 and 100 Ω, while balanced lines are found in the range of 70 to 600 Ω. What this means to us as radio people, is that if we have an antenna that has an impedance of 50 Ω and a radio transmitter designed to drive a 50 Ω load, we can connect the two with any length of the appropriate 50 Ω coaxial cable and the transmitter will think it is right next to the antenna. The antenna will receive most (see next section) of the transmitted power and all is well with the world!

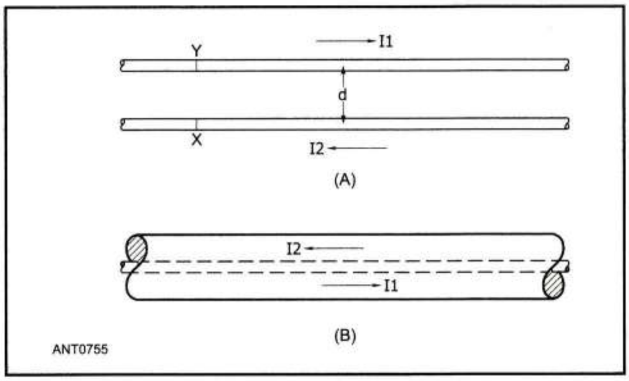

Figure 7-1 — Parallel wire and coaxial transmission lines.

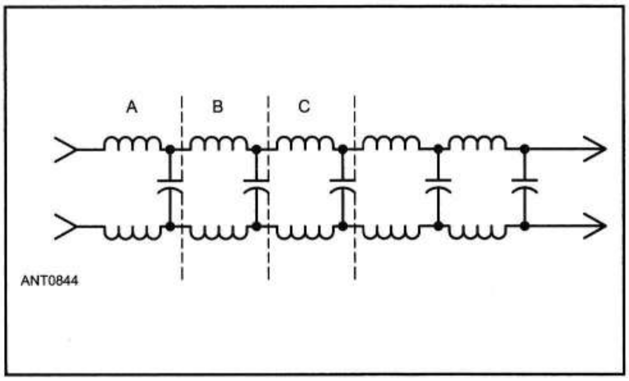

Figure 7-2 — Lumped constant equivalent of an ideal transmission line.

Propagation Velocity

Signals in air dielectric transmission lines propagate at almost the speed of light in air. Other dielectric materials between the conductors cause the signals in transmission lines to slow down just as we observe with light rays traveling through water. In many cases, this is not a matter of concern, since we often only care that the signals get out the other end, however, there are some exceptions.

The velocity can be shown to be reduced by a factor of one over the square root of the relative dielectric constant of the insulating material. Some cable specifications provide the relative velocity as a fraction of the speed of light. If not, and you know the material, most engineering handbooks include tables of properties of materials. For example, polyethylene is a common cable insulating material and has a relative dielectric constant of 2.26. The square root of 2.25 is 1.5, so the propagation velocity in polyethylene insulated coaxial cable is about $3/1.5 \times 10^8$ or 2×10^8 m/sec.

Some applications actually use coaxial cables to provide delayed signals in pulse and other applications. Having a way to accurately predict the delay just by knowing the cable characteristics and measuring the length of the cable can save a lot of lab time.

Attenuation

The ideal transmission line model shown in Figure 7-2 passes all input power to a matched load at the output. A real transmission line also has resistance associated with the wire conductors and some loss of signal due to the nature of the insulating material. As transmission lines are made larger, the resistance is reduced and as the dielectric material gets closer to low-loss air, the losses are reduced. The skin effect causes currents to travel nearer to the surface of the conductors at higher frequencies, and the effective loss thus increases as the frequency is increased.

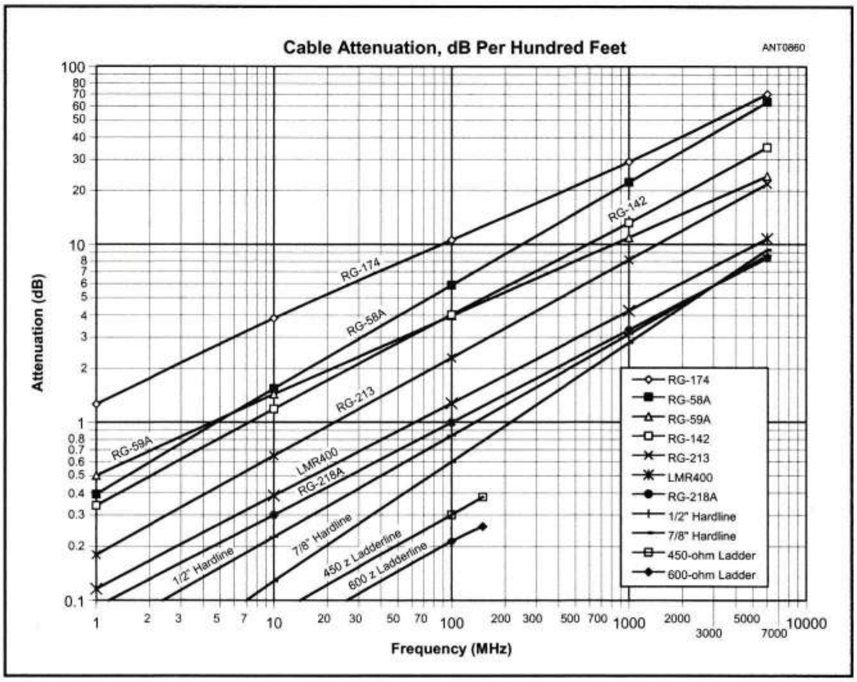

Figure 7-3 — Loss of some typical transmission lines in dB per 100 feet as a function of frequency. The RG-58 transmission lines are 50 Ω polyethelene insulated coaxial cable slightly less than ¼ inch in diameter. The RG-8 through RG-216 are 50 and 70 Ω polyethelene insulated transmission lines with a diameter somewhat less than ½ inch. The "hardline" types have a foam dielectric very near to air.

Figure 7-3 provides some real world examples of the losses as a function of frequency for the most common types of transmission line. Note that the loss increases linearly with length and the values are for a length of 100 feet. Note also that the losses shown are for transmission lines feeding into loads matched to their Z_0. As will be discussed shortly, losses can increase significantly if the line is not matched. The "open wire" line shown consists of two parallel wires with air dielectric and infrequent spacers, typically resulting in a Z_0 of around 600 Ω. While the losses of such a line are low, they only work well if spaced from metal objects and lossey material and not coiled up, while coaxial cables have higher loss, all the signal is within the outer conductor and they can be run in conduit, coiled up, placed next to other wires and are therefore much more convenient to work with.

Sometimes a long straight run of open wire line will be transformed to 50 Ω at the ends with coaxial cable used at the antenna and radio ends to take advantage of the benefits of both.

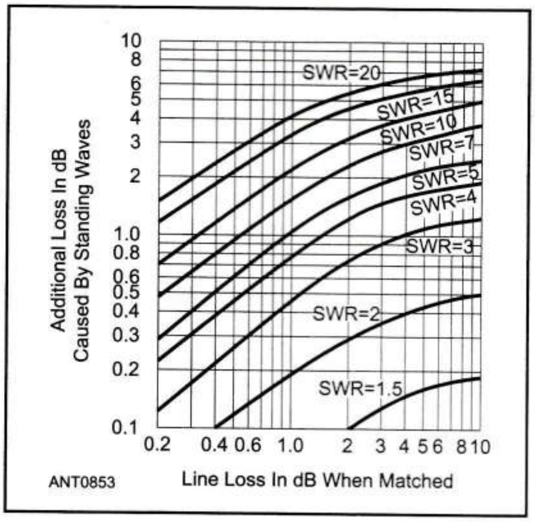

Figure 7-4 — Additional loss of a transmission line when mismatched. This loss needs to be added to the loss in Figure 7-3 for mismatched lines.

Lines with Unmatched Terminations

In our discussions so far, we have been talking about transmission lines feeding terminations matched to their characteristic impedance. In that case, the voltage-current relationship at the load will reflect the impedance of the load — not the characteristic impedance. Along the line the voltage and current will vary with distance providing a load to the transmitter end that is generally neither that of the far end Z_L, nor the Z_0 of the transmission line. The transmitter load can be calculated knowing the Z_L, the Z_0 and the electrical length of the line as discussed in the sidebar.

The ratio of maximum voltage on the line to minimum voltage on the line is called the *standing wave ratio* or SWR. A matched line has an SWR of 1:1, a 50 Ω line terminated with a 25 or 100 Ω load will have an SWR of 2:1. There are a whole family of complex impedances that will also have a 2:1 SWR, by the way. The computation is easier with resistive loads.

There are some interesting special cases with a mismatched line. The load impedance, resistive or complex, repeats every λ/2, for example. The impedance goes to the opposite extreme in odd multiples of a λ/4. For example our 25 Ω load would get transformed to 100 Ω in λ/4 or ¾ λ transmission line sections and vice versa. This effect can be used to our advantage if we wish to transform impedances at a specific frequency. In this case, the line is actually acting like a kind of antenna tuner.

A generally less desirable effect of mismatched lines is that the losses increase. This is easy to see, if voltages and currents are higher, we might expect losses to increase as well. **Figure 7-4** provides the additional loss for a mismatched line that needs to be added to the matched loss in Figure 7-3. As is evident, the combination of matched loss and high SWR results in dramatic increases in loss. This is why antenna designs that don't use matched transmission lines often use air-dielectric lines.

What Can We Do to Reduce Transmission Line Loss?

By now you should have a clear understanding of what can happen to a signal if applied to an antenna system through even a relatively short length of coaxial cable that has a high SWR. Depending on the matched loss, length and SWR, much of your signal can disappear before it gets to the antenna.

The insidious aspect of this is that while this is happening, the SWR at the radio with a collocated or internal tuner looks great — a perfect match and easy to get it. Losses are like that — they tend to result in a wide SWR bandwidth, as does a dummy load. Not a good thing if you want to communicate over the air. Fortunately, there are three fairly straightforward approaches to solving this problem:

- Change the antenna to one with either a wide enough bandwidth to cover all frequencies of interest, or one that covers multiple amateur bands with low SWR. This is not a topic for a book on antenna tuners, but is covered in many books on antennas.[1,2]
- Move the antenna tuner to, or close to, the antenna location. This is the subject of Chapter 8, or;
- Use a transmission line that has low enough matched loss so that even a high SWR doesn't result in high losses. This is the subject of Chapter 9.

Notes

[1] J. Hallas, W1ZR, *Basic Antennas — Understanding Practical Antennas and Designs,* Available from your ARRL dealer or the ARRL Book-store, ARRL order no. 9994. Tele-phone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/arrl-store/**; **pubsales@arrl.org**. See Chapter 11, "Wideband Dipole Arrays;" Chapter 12 "Multiband Dipole Arrays" or Chapter 21 "Log Periodic Dipole Arrays," for example.

[2] R. D. Straw, Editor, *The ARRL Antenna Book,* 21st Edition. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9876. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/arrl-store/**; **pubsales@arrl.org**.

Determining the Input Impedance of an Unmatched Transmission Line

The input impedance of a transmission line of any length with any terminating impedance can be determined in a number of different ways.

The most straightforward way is through direct calculation. Unfortunately, this is also perhaps the most time consuming and perhaps most error prone method, at least until you have it set up on a spreadsheet or other program that can handle he hyperbolic function.

My favorite way to determine the input impedance, standing wave ratio (SWR) as well as the line loss, is to use *TLW* (*Transmission Line for Windows*) software that comes with *The ARRL Antenna Book*.[3] The main screen is shown in Figure 7-A performing an analysis of an antenna with a complex load impedance fed through 100 feet of coaxial cable.

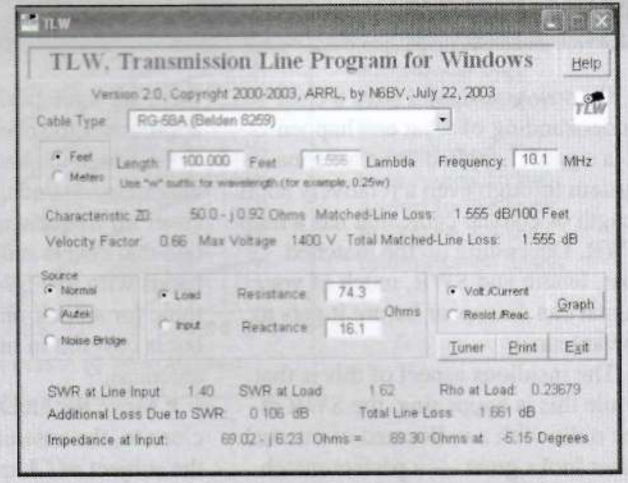
TLW
Help
TLW
Graph
Tuner
Print
Exit

TLW, Transmission Line Program for Windows

Version 2.0, Copyright 2000-2003, ARRL, by N6BV, July 22, 2003

Cable Type: RG-58A (Belden 8259)

● Feet ○ Meters Length: 100.000 Feet 1.555 Lambda Frequency: 10.1 MHz

Use "w" suffix for wavelength (for example, 0.25w)

Characteristic Z0: 50.0 - j 0.92 Ohms Matched-Line Loss: 1.555 dB/100 Feet

Velocity Factor: 0.66 Max Voltage: 1400 V Total Matched-Line Loss: 1.555 dB

Source: ● Normal ○ Autek ○ Noise Bridge

● Load ○ Input Resistance: 74.3 Reactance: 16.1 Ohms

● Volt./Current ○ Resist./React.

SWR at Line Input: 1.40 SWR at Load: 1.62 Rho at Load: 0.23679

Additional Loss Due to SWR: 0.106 dB Total Line Loss: 1.661 dB

Impedance at Input: 69.02 - j 6.23 Ohms = 69.30 Ohms at -5.15 Degrees

Figure 7-A — The main screen of *TLW* (*Transmission Line for Windows*) software is shown performing an analysis of the condition of the L/D of 10,000 case.

The antenna input impedance (74.3 + *j*16.1) is inserted in the LOAD box. The + *j*16.1 indicates an inductive reactance with a value of 16.1 Ω at the frequency of interest (10.1 MHz). A –*j* value would indicate a capacitive reactance, which would be entered with a minus sign. The output impedance through 100 feet of RG-58A 50 Ω coax is provided at the bottom in both rectangular (69.02 – *j*6.23) and polar coordinates (69.53 @ –5.15°), along with the SWR at line input (1.40:1) and output (1.62:1) as well as line loss (1.661 dB) — both for the matched case (1.555 dB) and the additional loss due to mismatch (0.106 dB). That's about everything I could think to ask, except which pile in the basement has the RG-58A!

A third way to evaluate the input impedance is through a graphical method. A Smith chart, see Figure 7-B, can be used to determine the input impedance of a transmission line. This was very commonly used before the personal computer became ubiquitous. In addition to the accuracy limitations due to the input and output resolution inherent in a chart, the Smith chart assumes that the line is lossless. This may result in significant errors depending on the amount of line loss.

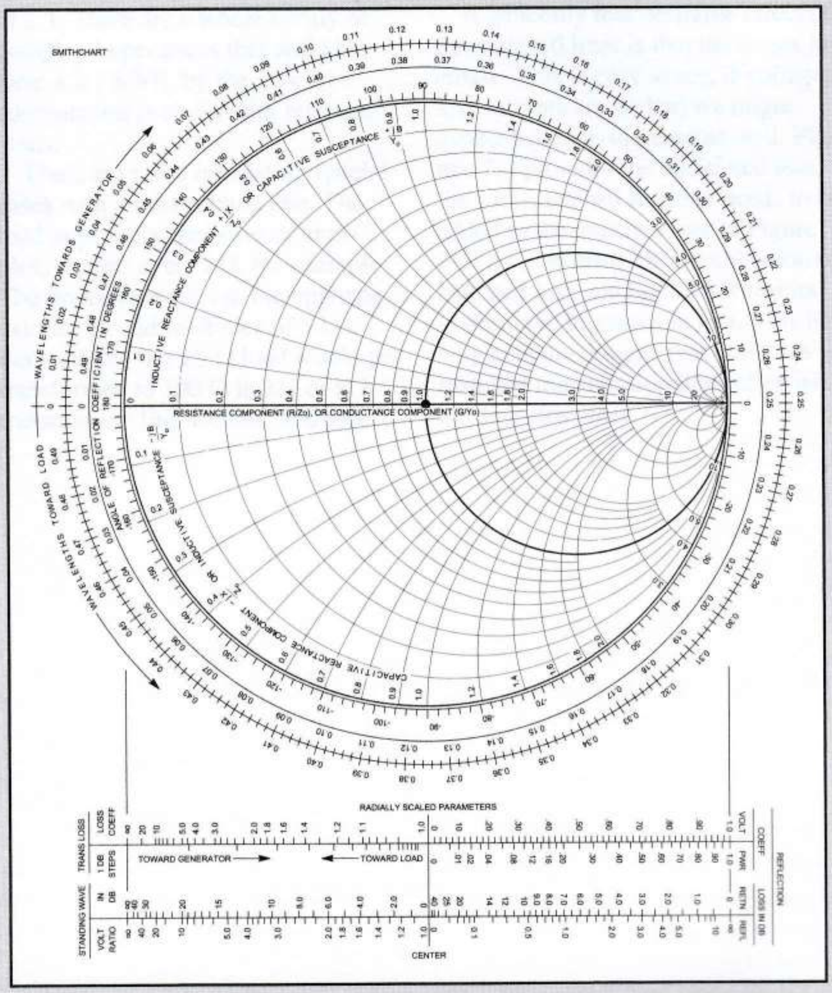
SMITHCHART
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
INDUCTIVE REACTANCE COMPONENT
CAPACITIVE REACTANCE COMPONENT
CAPACITIVE SUSCEPTANCE
INDUCTIVE SUSCEPTANCE
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
RADIALLY SCALED PARAMETERS
TOWARD GENERATOR
TOWARD LOAD
CENTER
STANDING WAVE
VOLT RATIO
IN DB
TRANS LOSS
1 DB STEPS
LOSS COEFF
REFLECTION COEFF
VOLT
PWR
RETN LOSS IN DB
REFL

Figure 7-B — A Smith chart used for the graphical determination of the input impedance of a lossless transmission line.

---

[3] *TLW* is supplied with *The ARRL Antenna Book*, 20th Edition, available from the ARRL Bookstore at **www.arrl.org/catalog/** order number 9043 — $39.95.

Review Questions

7-1 Describe three reasons it might be desirable to have a transmitter and an antenna in different locations.

7-2 If you want to make a $\lambda/4$ section of RG-213 for our 10 MHz system, how long would you make it?

7-3 A 1000 W transmitter at 15 MHz is feeding a matched load through 200 feet of RG-8 transmission line. How much power reaches the antenna? Repeat if the frequency is 150 MHz. Repeat both cases if the antenna has an SWR of 3:1.

Chapter 8

Moving the Tuner to the Back 40

LDG Electronics
1445 Parran Road
St. Leonard, MD 20685
Ph. 410-586-2177 Fax: 410-586-8475
ldg@ldgelectronics.com
www.ldgelectronics.com
LDG
Antenna
Transmitter
Remote Interface
RT-11 Automatic Tuner

Automatic antenna tuners designed for remote mounting. On top, one from SGC, beneath a model from LDG.

Contents

An Example of Using a Remote Tuner to Minimize Loss.... 8-2
Intermediate Steps 8-4
Remote Antenna Tuners — the Downside......................... 8-5
Review Questions... 8-6

Some automatic antenna tuners are designed to be mounted remotely, rather than being collocated with the radio. We can take advantage of this in order to minimize losses in the transmission line between the antenna system and the transceiver. With the tuner at the antenna end, the load at the antenna end of the transmission line provides very close to a 1:1 SWR. The line will then have only the matched loss between the radio and the tuner, not the additional loss due to mismatch.

An Example of Using a Remote Tuner to Minimize Loss

Figure 8-1 shows the configuration of a transceiver with a remote automatic antenna tuner at the antenna. Let's assume that the transmission line is RG-58 and that it is 100 feet between the radio and the tuner plus the tuner is connected directly to the feed point of the coax fed antenna.

We will again use the example from Chapter 4 of a 20 meter dipole tuned by the tuner to operate on 10 meters where its impedance will

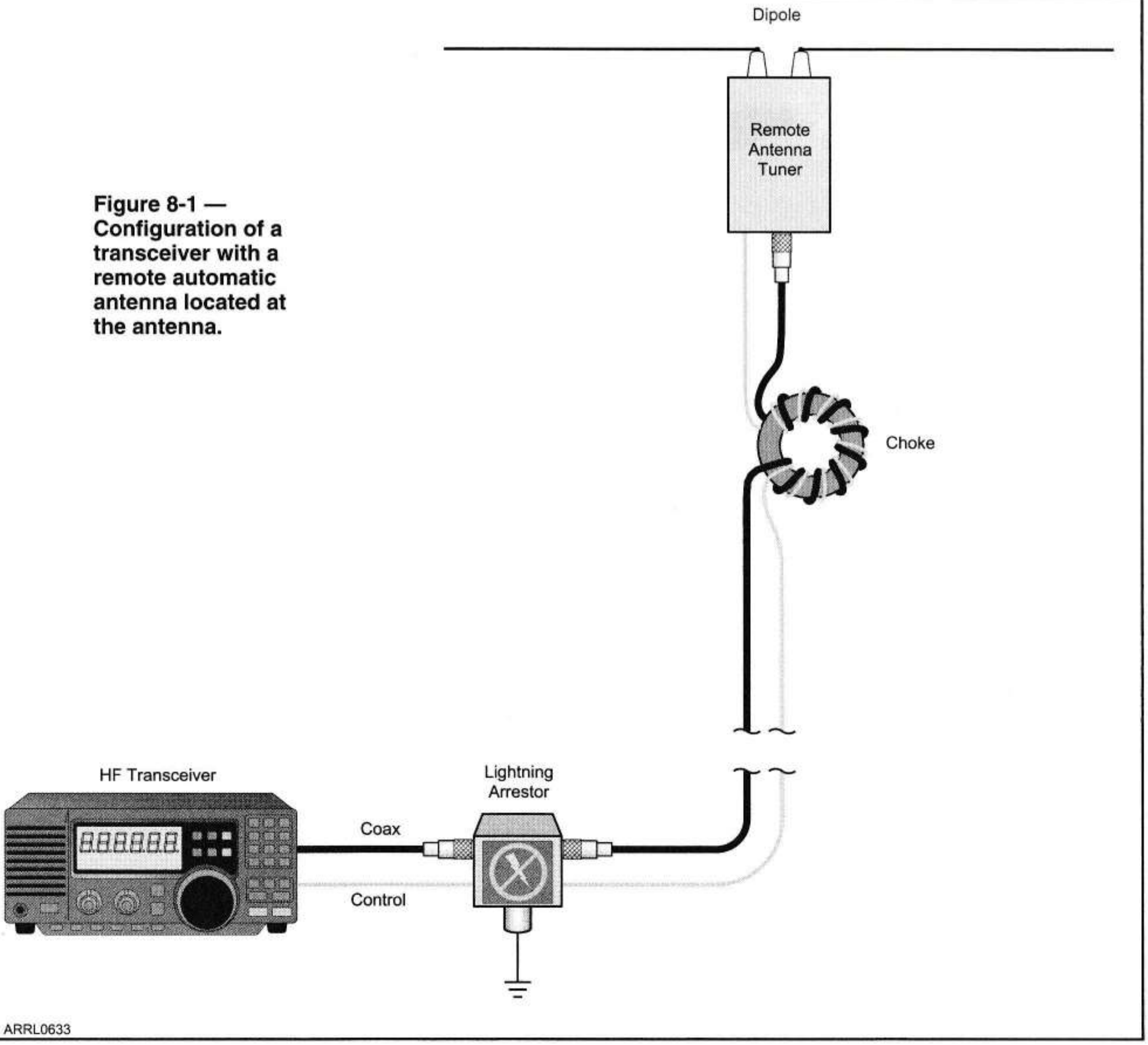

Figure 8-1 — Configuration of a transceiver with a remote automatic antenna located at the antenna.

Table 8-1

Comparison of System Losses for Remote and Collocated Tuners — RG-58

Loss Element	*Tuner at Radio*	*Remote Tuner Beneath Antenna*
Loss in 100' Cable (dB)	14.7	2.9
Percent Power Lost (%)	96.6	48.1
Power to Antenna at 100 W (W)	3.4	51.9

Table 8-2

Comparison of System Losses for Remote and Collocated Tuners — LMR-400

Loss Element	*Tuner at Radio*	*Remote Tuner Beneath Antenna*
Loss in 100' Cable (dB)	8.62	0.66
Percent Power Lost (%)	86.3	14.1
Power to Antenna at 100 W (W)	13.7	85.8

be 3836 Ω resistive in series with 790 Ω capacitive reactance — an SWR of 81:1. **Table 8-1** provides the comparison in loss for a 100 W output transceiver, in each case assuming the antenna tuner contributes no loss. While this is slightly optimistic, that loss is usually very small. The result is rather dramatic and illustrates why just hitting the TUNE button on your internal tuner may not provide satisfactory results even if the resulting radio SWR ends up at a perfect match.

Perhaps you will also not be satisfied with moving your tuner to the antenna and still losing almost half your power in the RG-58. The lesson here is that RG-58, while available and relatively inexpensive, is not the best choice for even a 100 foot run at 28 MHz. Of course, it's much worse as you move to VHF, and keep in mind that it effects both transmit and receive signals.

One of the lowest loss flexible coaxial transmission lines that fits the usual UHF coax connectors is Times Wire LMR-400. There are similar cables by other manufacturers. **Table 8-2** repeats the exercise shown in Table 1 with the lower loss cable. Note that while there is measurable loss even for the matched case, it is less than 1/10 of an S-unit, so hardly noticeable, and much less than the loss for the tuner located at the radio that will result in about a 1.5 S-unit reduction on the receiving station's meter.[1]

[1] Notes appear on page 5.

Intermediate Steps

Not all antenna systems are structured in a way to allow a tuner to be at a feed point. An example might be a dipole suspended between two supports with a transmission line hanging in the middle. In this case, it is tempting to consider a short coax run from the antenna itself to the tuner as shown in **Figure 8-2**. This arrangement may also be necessary if the tuner is not waterproof and requires shelter from the elements.

This is a feasible arrangement; however, while a substantial portion of the loss can be eliminated in this way, the loss in the section between the tuner and the antenna can still be surprisingly high. For example, for the RG-58 case of Table 8-1, a 30 foot run between the antenna and the tuner would have a loss of 4.4 dB, or 64% of the power, leaving only 36 W from your 100 W transmitter to be radiated — if it got to and through the tuner without other losses.

A better way to accomplish this in many cases will be to use a section of the low loss transmission line discussed in Chapter 9 between the antenna and the tuner.

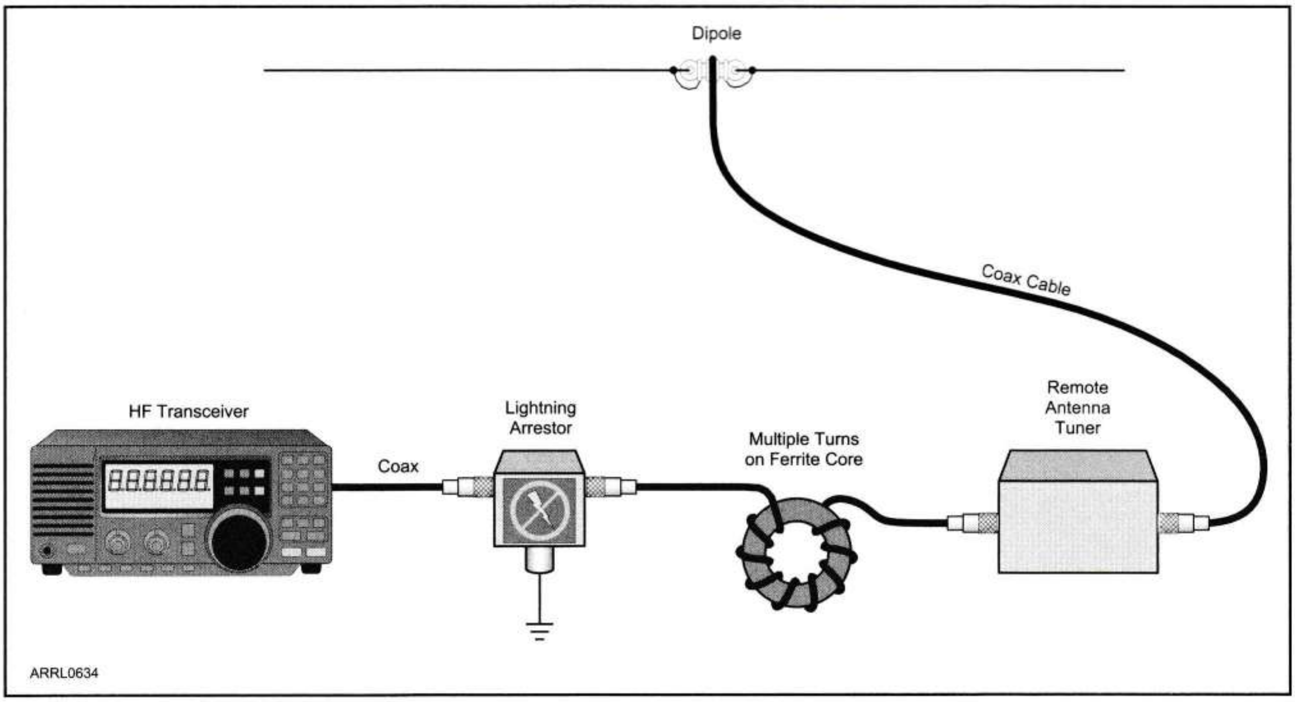
Dipole
Coax Cable
HF Transceiver
Coax
Lightning Arrestor
Multiple Turns on Ferrite Core
Remote Antenna Tuner
888888
ARRL0634

Figure 8-2 — Configuration of a transceiver with a remote automatic antenna located beneath the antenna.

Remote Antenna Tuners — the Downside

The use of a remote antenna tuner has a few disadvantages. I currently use three in various station locations and am quite happy with them in spite of the downsides:

- One remote tuner is located near my main basement station. I share antennas with my spouse (Nancy, W1NCY), who has an upstairs station. With an automatic tuner that can be patched to antennas from the W1ZR main station, she has access to any W1ZR antenna with low loss in the 100 foot coax run to her desktop setup.
- On our sailboat, I use an insulated backstay as an HF antenna. This non-resonant system has a high SWR on all bands and requires a tuner at the feed point to feed it against ground. I have a short lead from a corner under the lazerette beneath the antenna feed with another short lead to the engine block where a ground connection is available. Not only does it work well on all HF bands, but by keeping the radiating portion of the system in the corner of the boat, I minimize interference between the radio and marine navigation systems.
- I also have a remote tuner in the trunk of my car for my mobile HF system. Typical HF mobile antennas have very narrow SWR bandwidths. By using a remote antenna tuner at the antenna base, I can cover entire bands, or even tune to lower frequency bands than the antenna is designed for. While this sort of operation isn't as efficient as adjusting the antenna itself, it is much easier than making adjustments while driving at highway speeds!

There are a few aspects of remote tuner operation that might be disadvantageous for some. One is the fact that in most such installations the tuner becomes integrated into a particular antenna structure. If we have multiple antennas, each will need its own tuner — a potentially expensive proposition.[2] It also is the case that in addition to providing coaxial cable to the tuner, one must also provide operating power and sometimes control signals to the tuner. Some tuner manufacturers package the coax, power and control into extension cable assemblies that can be used for this purpose. Another possibility is to use a *bias-T* to insert the power onto the coax cable, sharing it between the two functions.[3]

Notes

[1] The Collins Radio established, now industry standard, S-meter calibration is that S-9 equals 50 µV at the antenna terminals and each S-unit represents a change of 6 dB. To say that not all amateur S-meters follow this standard would be an understatement.

[2] Remotly controlled antenna switches may be used to solve this problem, although commercial units cost as much as some remote 100 W antenna tuners.

[3] S. Ford, WB8IMY, "Short Takes — MFJ-4712 Two-Position Remote Antenna Switch," *QST*, Sep 2006, p 48.

Review Questions

8-1 Why is the quality (loss) of a transmission line more important in the mismatched than the matched case?

8-2 If using an internal tuner in a transceiver, why might the SWR *before* tuning be an important indicator of antenna system efficiency?

8-3 Why is the SWR *before* tuning insignificant if an antenna located tuner can match to the transmission line Z_0?

Chapter 9

Transmission Line Choices for Low Loss

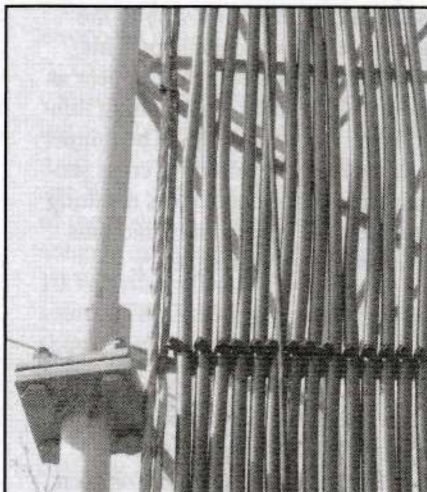

Coaxial cables going up one of the towers at W1AW, the ARRL Headquarters station

Contents

Hello Real World! 9-2

How Can We Determine Loss? 9-3

How is Transmission Line Made? 9-4

I Know What's Happening at the Shack —

Sidebar: What's Happening at the Other End of my Feed Line? 9-5

Review Questions 9-6

The ideal transmission line accepts the output power (as well as received signal) from its source and delivers it to its destination load without loss. If the characteristic impedance (Z_0) is the same as the source and load impedances, it delivers it with the same ratio of voltage to current. That means, for all practical purposes, the ideal transmission line is invisible to the system. Devices connected to both ends act as if they were collocated, except for the delay in signal arrival time, unimportant except in special circumstances.

Hello Real World!

Unfortunately, ideal transmission lines don't exist, although for many applications we can come remarkably close. The primary issue in most cases is loss. We just don't get as much power to the far end as we put in. There are three primary mechanisms that result in transmission line loss. They are conductor resistance, dielectric loss and leakage or radiation.

Conductor Resistance

Figure 7-1 shows the geometry of the two primary types of transmission line. At (A) is a balanced transmission line, while at (B) we have coaxial cable of the type discussed in the earlier chapters of the book. Both propagate signals down the line as currents with their associated electric and magnetic fields. In a properly terminated transmission line, the magnitude of I_1 will equal that of I_2 with the result that fields at some distance will cancel. In the case of coaxial cable they cancel at the shield, while for balanced line they exist for some distance around the line.

The action of the currents in an ideal line can be modeled as shown in Figure 7-2 in Chapter 7. The model can be analyzed with reasonable accuracy in comparison to real cable if there are at least 10 sections per wavelength, and if the inductors and capacitors are modeled as real elements, including the resistance of the wire and the dielectric loss of the capacitors.

For the case of dc, the wire loss just equals the dc resistance of the wire, a function of its diameter and material. As frequencies increase, *skin effect* results in the current moving to the outer edge of each conductor. This makes the wires electrically appear as tubes, with the tube walls getting thinner as the frequency rises. The thinner walls decrease the effective cross sectional area of the conductors resulting in an increase in resistance and thus increased loss with frequency.

Dielectric Loss

An ideal transmission line would consist of conductors with the insulation between them a perfect lossless vacuum. While some transmission lines approach this, with air (or even special gasses) as the primary dielectric, real lines always have some lossey material between the conductors.

A lossey dielectric has two significant effects on signals: First, it slows the signal to some fraction of the speed of light. This is often an important consideration in multielement driven antennas that rely on accurate phasing of signals, but does not enter into this discussion.

Second, dielectric loss results in yet another contribution to line loss, and this also increases with increasing frequency.

Radiation or Leakage Loss

While generally less significant a loss factor than the previous two, some signal is lost in a real line due to it leaving the transmission line before it reaches the intended load. In a sense this is more a misdirected signal than a lost signal, and, because the signals end up in undesired places, the interference to other systems may be more important than the slight reduction in signal delivered to the desired load. Still, for either reason it can represents a significant issue for real lines.

In the case of coaxial cable, a portion of the shield resistance (depending on the shield coverage) results in a current that appears on the outside of the shield. The outer shield current actually makes the outside of the shield act as an antenna with resulting radiation detectable along the length of the shield. For balanced transmission line, the fields cancel wherever the distance to both wires is the same. For a perfectly installed line, that means all the places on a plane centered between the conductors and perpendicular to the line between them.

If the line is not perfectly installed, there will be some unbalance between the capacitance of each wire to ground resulting in a current unbalance that results in radiation. A similar effect occurs if the load is not perfectly balanced. Even in the case of perfect balance, there is some radiation in directions in which the distance to the wires is different. If the spacing is a small fraction of a wavelength, this radiation is diminished within a few times the wire spacing.

How Can We Determine Loss?

Transmission line loss is a design parameter that is specified by each manufacturer for their cables. The line loss is generally specified over the usual operating range at a number of frequencies, often 1, 10, 100, 1000 and 10,000 MHz.

Matched Loss

Figure 7-3 shows the *matched* loss per 100 feet of representative transmission line types. As you make choices of transmission line, it is important to note that this data is representative — the variation between attenuation of different manufacturers, and even different part numbers from the same manufacturer, can be striking. Check the manufacturer's Web page for the data sheet of the transmission line you are considering to be sure you know what you are getting.

For lengths other than 100 feet, the loss scales linearly. For example if a line has a matched loss of 2 dB at 100 feet, the same line, at the same frequency, will have a loss of 1 dB if 50 feet long, and 5 dB if 250 feet long.

Additional Loss Due to Mismatch

By now it should come as no surprise to find that a mismatched line has more loss than a matched one. The additional loss due to mismatch is a function of both the mismatch and the loss if matched. Figure 7-4 shows the additional loss in dB that occurs as a result of a line not being matched.

Note that the SWR shown is the SWR as measured, or calculated, at the load not the SWR measured at the transmitter end of the cable. This is particularly important in the case of a lossey line since the loss will reduce both the power reaching the antenna and the power of the reflected wave that is used to determine the SWR. This can give very optimistic, and erroneous results.

To give an example of this effect, consider a 100 W transmitter driving a 100 feet of coax with a loss of 3 dB (50% loss). The antenna will see 50 W of power. Lets say 20% of the power is reflected due to the antenna mismatch. That will result in 10 W being reflected back toward the source. The 3 dB loss results in 5 W showing up as reflected power at the bottom of the cable. This is quite different than if the 100 W were applied to the antenna on lossless line — in that case, a reflected power of 20 W, not 5 W, would, show up at the SWR measurement device.

Table 9-1 summarizes what we have at the two locations. Note the rather distressing result. A very acceptable measurement of an SWR of 1.6:1 at the bottom of the coax is the result of an unpleasant SWR of 4:1 at the antenna. In this example, our 100 W of power results in only 40 W radiated from the antenna — yet all of our measurements make us think we're doing well. Unfortunately, this example is not unusual, especially at the upper end of HF into the VHF range. If it happens at higher frequencies, it is usually more evident since nothing much ends up going in or out of the system! The sidebar discusses ways that this can be calculated, and perhaps avoided, through the use of software.[1]

Table 9-1

Forward and Reflected Power and SWR as Seen at Each End of a Transmission Line with 3 dB Loss

Measurement	*Bottom of Cable*	*Top of Cable*
Forward Power (W)	100	50
Reflected Power (W)	5	10
Indicated Reflection Coefficient	0.224	0.447
Indicated SWR	1.6	4.0

---

[1] J. Hallas, W1ZR, "I Know What's Happening at the Shack — What's Happening at the Other End of my Feed Line?" *QST*, Feb 2007, p 63.

How is Transmission line Made?

You may wonder what is it about transmission lines that makes their loss performance so different. It all comes down to how and of what they are made. **Figure 9-1** shows typical construction techniques of parallel and coaxial transmission line types. Not shown are lines with mostly air as a dielectric, the lines with the least loss. Parallel line, so-called open-wire line, consists of two parallel bare wires with occasional spacers designed to keep them apart and at about the same distance. The traditional insulators were ceramic, however, recent lines have used plastic or PVC spacers. Mostly air-dielectric coax is sometimes encountered. Such cable that I've seen is formed from two concentric copper tubes with ceramic donut shaped spacers every foot or so.

The characteristic impedance of cable is a function of conductor diameter and spacing. Thus as line gets larger, each conductor gets larger for the same Z_0 and thus the resistance is reduced, resulting in lower loss with larger cable.

The balanced line (more next chapter) has progressively less attenuation as the Z_0 gets higher (the ratio of series resistance to load resistance goes down) and as the polyethylene dielectric is replaced by air, partially in the case of window line (C) and more completely in open wire line described above.

Coaxial cable loss is similar in that foam dielectric is part polyethylene and part air. The lines with double (E) and solid metal (F, G) shields have lower resistance and much less leakage.

Perhaps not surprisingly, all the factors that make line have lower attenuation, make it more expensive. Still, in most systems, for reasonable distances, transmission line cost is a small part of total system cost, and often provides a worthwhile investment.

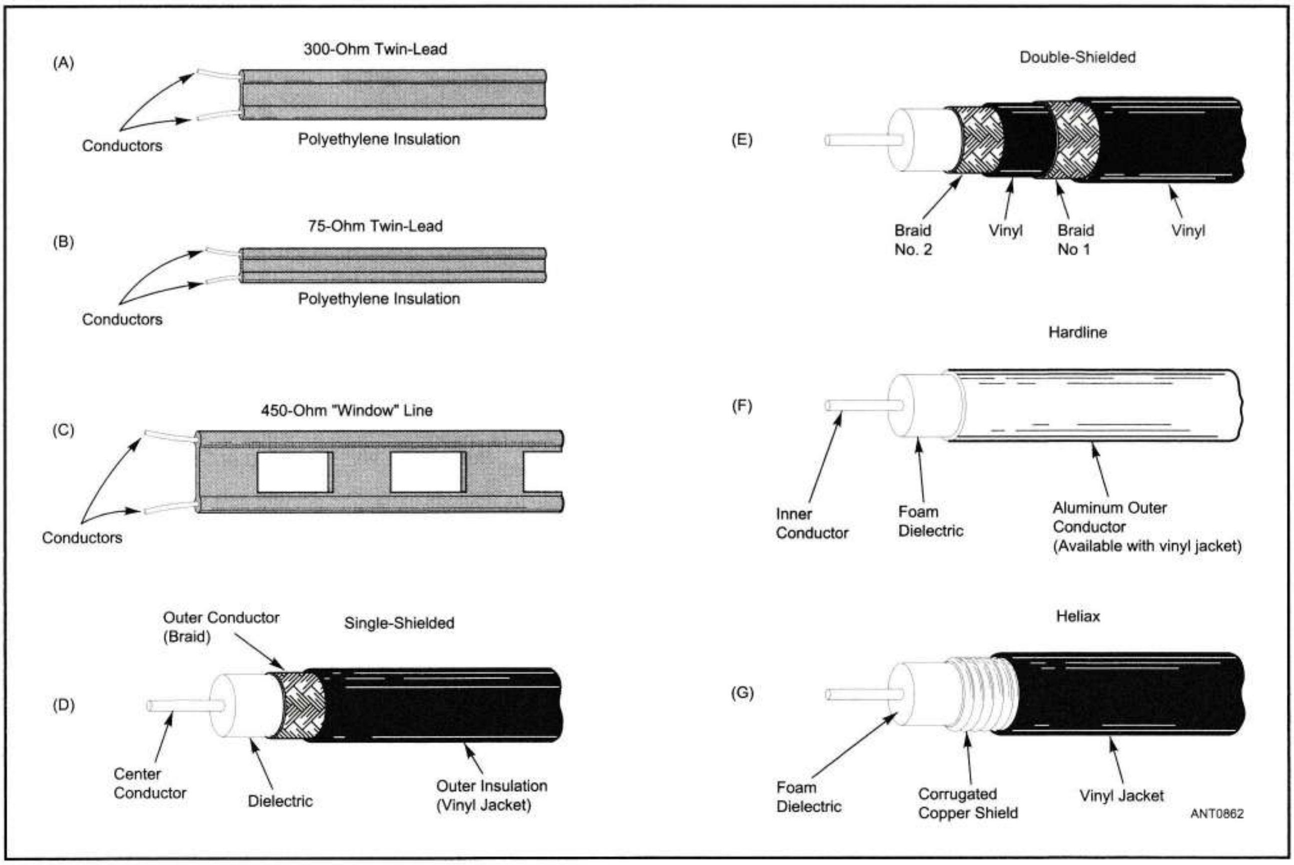
(A)
300-Ohm Twin-Lead
Conductors
Polyethylene Insulation
(B)
75-Ohm Twin-Lead
Conductors
Polyethylene Insulation
(C)
450-Ohm "Window" Line
Conductors
(D)
Outer Conductor (Braid)
Single-Shielded
Center Conductor
Dielectric
Outer Insulation (Vinyl Jacket)
(E)
Double-Shielded
Braid No. 2
Vinyl
Braid No 1
Vinyl
(F)
Hardline
Inner Conductor
Foam Dielectric
Aluminum Outer Conductor (Available with vinyl jacket)
(G)
Heliax
Foam Dielectric
Corrugated Copper Shield
Vinyl Jacket
ANT0862

Figure 9-1 — Construction techniques used in making flexible transmission lines.

I Know What's Happening at the Shack — What's Happening at the Other End of my Feed Line?

If you want to find out — here's the easy way using TLW.

Joel R. Hallas, W1ZR
Technical Editor, QST

I'm told that one of the more frequent questions received by *QST's* "Doctor" has to do with folks wanting to determine the impact of transmission line losses on the effectiveness of their antenna system. These questions are often along the lines of "I measure an SWR of 2.5:1 at the transmitter end of 135 feet of RG-8X coaxial cable. My transceiver's auto-tuner can tune it to 1.1, but how can I tell what my losses are?" or "How much difference will I have if I have a tuner at the antenna instead of using the built-in tuner?"

These are important questions that almost every amateur operator is faced with from time to time. An approximate answer can be obtained by using the graphs found in any recent edition of The ARRL Antenna Book showing the loss characteristics of many transmission line types, plus adding in the effect of an SWR greater than 1:1. The SWR at the antenna end can be determined from the bottom end SWR and the cable loss. Using these graphs requires a bit of interpolation or Kentucky windage, but can result in useful data.

But There's an Even Better Way!

Packaged with each of the last few editions of *The ARRL Antenna Book* is a CD containing the pages of the whole Antenna Book as well as some very useful software. The program that I use almost daily is one written by Antenna Book Editor R. Dean Straw, N6BV, called *TLW* for Transmission Line for Windows.

TLW provides a very easy to operate mechanism to determine everything I usually need to know about what's happening on a transmission line. When you open the program, you are presented with a screen as shown in Figure A. This has the values plugged in from the last time you used it, often saving a step. Let's take a quick tour of the inputs:

Cable Type — This allows you to select the cable you would like to analyze. A drop-down box provides for the selection of one of 32 of the most common types of coax and balanced lines. An additional entry is provided for User Defined Transmission Lines that can be specified by propagation velocity and attenuation.

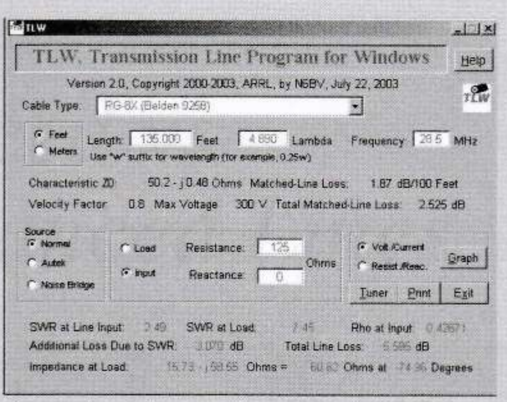

TLW, Transmission Line Program for Windows

Version 2.0, Copyright 2000-2003, ARRL, by N6BV, July 22, 2003

Cable Type: RG-8X (Belden 9258)

● Feet ○ Meters Length: 135.000 Feet 4.890 Lambda Frequency 28.5 MHz

Use "w" suffix for wavelength (for example, 0.25w)

Characteristic Z0: 50.2 - j 0.48 Ohms Matched-Line Loss: 1.87 dB/100 Feet

Velocity Factor 0.8 Max Voltage 300 V Total Matched-Line Loss: 2.525 dB

Source: ● Normal ○ Autek ○ Noise Bridge

○ Load ● Input Resistance: 125 Reactance: 0 Ohms

● Volt./Current ○ Resist./Reac. Graph

Tuner Print Exit

SWR at Line Input: 2.49 SWR at Load: 7.45 Rho at Input 0.42671

Additional Loss Due to SWR: 3.070 dB Total Line Loss: 5.595 dB

Impedance at Load: 15.73 - j 58.55 Ohms = 60.62 Ohms at -74.96 Degrees

Figure A — The opening screen of *TLW*, illustrating the process described in the article.

Length — In feet or meters, your choice.

Frequency — This is an important parameter when dealing with transmission line effects.

Source — This defines the form of the input impedance data. Generally, you can use NORMAL.

Impedance — The impedance can be specified as what you measure, resistive (real) and reactive (imaginary, minus means capacitive). This could come from your antenna analyzer at either end of the transmission line. Note, if you only know the SWR, not the actual impedance, all is not lost — see below.

Now for the Outputs

SWR — The SWR is provided at each end of the cable. This is an important difference that many people miss, important even with a moderate SWR at the transmitter end, as we'll see — the SWR at the antenna will be much higher due to the cable loss. With TLW, you instantly know the SWR at both ends, and the loss in the cable itself

Rho at Load — This is the reflection coefficient, the fraction of the power reflected back from the load.

Additional Loss Due to SWR — This is one of the answers we were after.

Total Loss — And this is the other, the total loss in the line, including that caused by the mismatch.

But Doctor, What if I can Only Measure the SWR — Not the Actual Impedance?

Often the only measurement data available is the SWR at the transmitter end of the cable. Because the losses are a function of the SWR, not the particular impedance, you can just put in an arbitrary impedance with that same SWR and click the INPUT button. An easy arbitrary impedance to use is just the SWR times the Z_0 of the cable, usually 50 Ω. For example, you could use a resistance of 125 Ω to represent an SWR of 2.5:1. This is what we've done in Figure A, using 135 feet of popular Belden RG-8X.

The results are interesting. Note that the 2.5:1 SWR as seen at the radio on 28.5 MHz results from a 7.45:1 SWR at the antenna — perhaps this is an eye-opener! Note that of the 5.6 dB loss, more than half, or 3.1 dB, is due to the mismatch.

Note that if we used something other the actual measured impedance, we can't make use of the impedance data that TLW provides. We can use the SWR and loss data, however, but that's probably what we wanted to find out.

We can now do some "what ifs." We can see how much loss we have on other bands by just changing the frequency. For example, on 80 meters, with the same 2.5:1 at the transmitter end, the SWR at the antenna is about 3:1 and the loss is slightly more than 1 dB. We could also plug in an impedance calculated at the antenna end and see what difference other cable types would make. For example, with the same 28.5 MHz SWR of 7.45 at the antenna and 135 feet of 1/2 inch Andrew Heliax, we will have a total loss of 1.5 dB at 28.5 MHz. Note that the SWR seen at the bottom will now be 5.5:1 and our radio's auto-tuner might not be able to match the new load.

But Wait There's More!

You can also click the GRAPH button and get a plot of either voltage and current or resistance and reactance along the cable. Note that these will only be useful if we have started with actual impedance, rather than SWR.

Pushing the TUNER button results in a page asking you to select some specifications for your tuner parts. TLW effectively designs a tuner of the type you asked for at the shack end of the cable. It also calculates the power lost in the tuner and gives a summary of the transmitted and lost power in watts, so you don't need to calculate it!

When you've finished, be sure to hit the EXIT button, don't just close the window. Otherwise TLW may not start properly the next time you want to use it.

Review Questions

9-1 What parameters make a difference in the matched loss of transmission line, mismatched loss?

9-2 Why is it important to know the SWR at the load, rather than at the transmitter to determine transmission line loss?

9-3 Does coaxial cable or parallel line transmission line tend to have the lowest matched loss?

Chapter 10

Balanced vs Unbalanced Lines

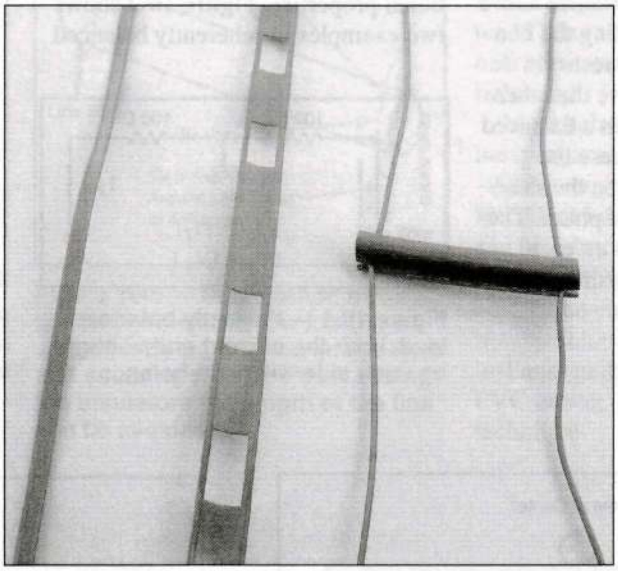

View of three popular types of balanced line. On far left, TV type 300 Ω twinlead, nominal 450 Ω "window" line and 600 Ω open-wire line.

Contents

What's the Story About Balanced Line? 10-2

The Benefits of Balanced Line... 10-3

The Downside of Balanced Line.. 10-4

What About Unbalanced Currents on Coax? 10-4

What Happens If I Feed my Dipole Directly with Coax? 10-5

Review Questions... 10-6

The careful observer might have noted that in Figure 7-3, in Chapter 7, the lowest loss line was not the fancy, expensive "hardline" coax, but rather the much less expensive balanced open-wire and window line. This is a hard combination to beat — low cost and low attenuation, which explains its popularity. The other side of the coin is that most current radio equipment is designed to operate with unbalanced coaxial line.

What's the Story About Balanced Line?

Before we can go too far, we need to discuss what we mean by balanced and unbalanced. By a balanced system, we mean one in which the two sides are at the same impedance above ground. For example feeding the non ground end of two 100 Ω resistors (see **Figure 10-1**) that have the other end grounded will result in a balanced 200 Ω system. In such a case the voltage on each side will be the same magnitude but 180° out of phase. The magnitude of the currents on each side will also be the same since the voltages and resistances are equal.

Note that the system would be balanced whether or not the ground connection were there. No current flows in the ground lead of the perfectly balanced system, so it could be removed without changing the operational properties. **Figure 10-2** shows two examples of inherently balanced

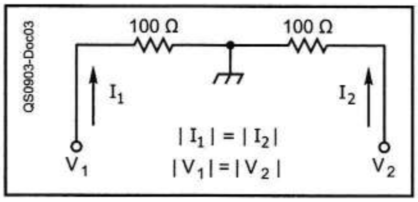
QS0903-Doc03
100 Ω
100 Ω
I_1
I_2
$|I_1| = |I_2|$
$|V_1| = |V_2|$
V_1
V_2

Figure 10-1 — Perfectly balanced load, both the current and voltage on each side will be in balance.

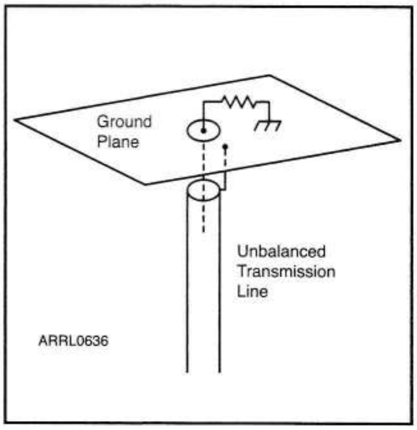
Ground
Plane
Unbalanced
Transmission
Line
ARRL0636

Figure 10-3 — Unbalanced resistive load. The signal is applied with reference to ground.

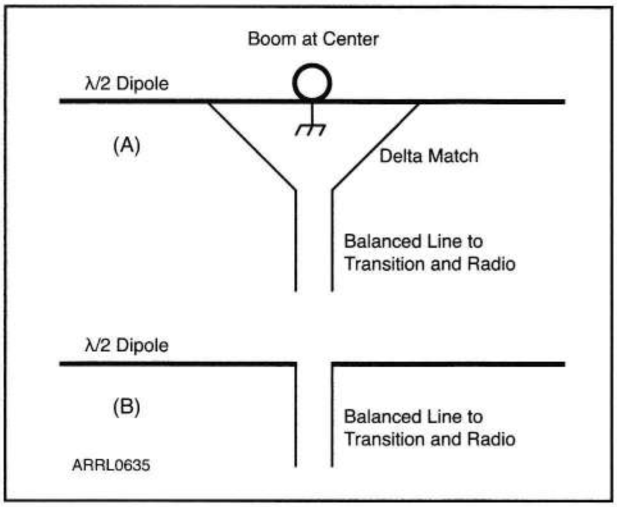
Boom at Center
λ/2 Dipole
(A)
Delta Match
Balanced Line to
Transition and Radio
λ/2 Dipole
(B)
Balanced Line to
Transition and Radio
ARRL0635

Figure 10-2 — Two examples of inherently balanced antennas, one (delta matched dipole) with a central ground (A) and one (split dipole) without (B).

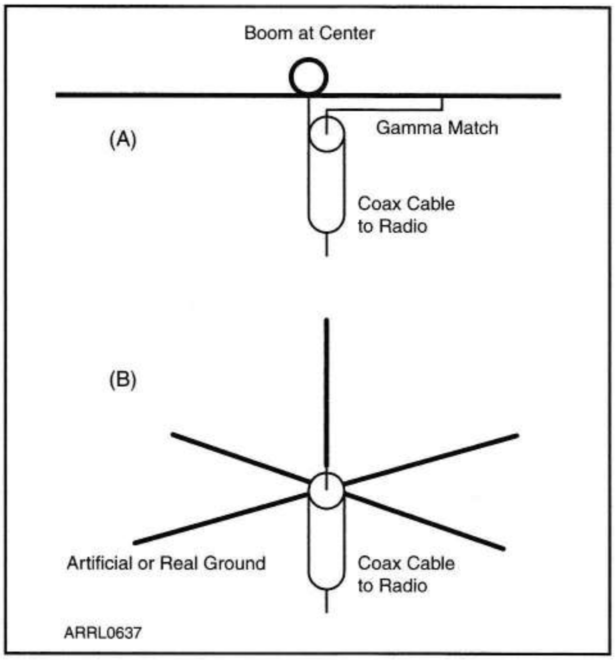
Boom at Center
(A)
Gamma Match
Coax Cable
to Radio
(B)
Artificial or Real Ground
Coax Cable
to Radio
ARRL0637

Figure 10-4 — Two examples of inherently unbalanced antennas, a gamma matched dipole at (A) and a vertical monopole at (B).

antennas, one with a central ground (A) and one without (B). The ground in the first is not actually necessary, but can be beneficial for lightning protection purposes. For both cases, we note that some kind of transition is required to shift to an unbalanced system for connection to the radio. The types of transition will be the subject of the next chapter.

An unbalanced system, on the other hand, is fed with respect to ground. That is, one side of the load is at ground potential. **Figure 10-3** is an example of an unbalanced resistive load, and **Figure 10-4** is shows two antennas with inherently unbalanced feed points.

The Benefits of Balanced Line

As noted earlier, balanced line has two primary benefits in comparison to coaxial cable. Most balanced lines, especially those with large portions of air dielectric, such as window line or ladder line, have a significantly lower matched loss than most coaxial cables. Even though their characteristic impedance is often higher than desired for many antenna systems, the total of matched and mismatched loss is generally much less than the loss of coax — particularly if the coax is not well matched to the impedance of the load.

Balanced line is often very conveniently connected to balanced antennas and is most often encountered in that application. It is also frequently found in very long runs from unbalanced antennas in which the losses (or cost) of coax would be prohibitive. In that case, it may be worth the effort to transition (Chapter 11) from unbalanced to balanced near the antenna and then back at the radio end.

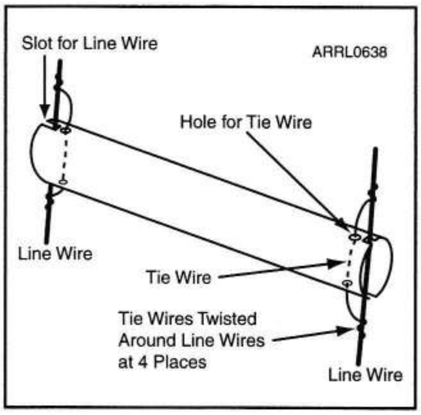

Figure 10-5 — One method of making homemade low loss open wire line. By putting the line in slots and securing with tie wires, sliding the insulators the length of the line can be avoided.

Most balanced line is relatively inexpensive when compared to coax — often an important consideration. Excellent homemade open wire line can be fabricated at low cost from two rolls of wire (stranded works best, in my experience, particularly if subject to flexing) and insulators made from inexpensive household PVC tubing. **Figure 10-5** shows one technique.

The Downsides of Balanced Line

As with most aspects of life, there are downsides of balanced line as well as benefits. While coax cable, with its fields contained within the shield, can be rolled, buried (if it has a direct burial rating) or installed within or near pipes without impacting performance, this is not true of balanced line. With balanced line, the fields conveying the signal down the line are significant not only between the wires but outside as well for a distance of a few times the wire spacing. This means that balanced line can't be placed on the ground, run in metal ducts, run through lossey material or rolled up, without causing additional loss.

While fields at a distance from balanced line are small, nearby fields can cause interference to other systems, particularly those interconnected by wiring. In addition, on receive, balanced line can pick up interfering signals from computers and other systems if run too close to them.

Another potential issue, particularly with the twinlead or window line variants of balanced line is a change in characteristics when wet. Rainwater can accumulate on the web material between the conductors resulting in a change to the dielectric properties of the line.[1]

There are a few things that can be done if this is a problem:

- Some have been known to carefully cut away some of the web to reduce the water collection surface. To avoid weakening the line, don't cut into the wire if you try this.
- In the "old days," some amateurs waxed their twinlead. This was reputed to cause the water to bead up and run off. Remaining water would be concentrated in droplets leaving most of the dielectric clear. I haven't seen any data on how effective this actually is, nor have I heard of any modern amateurs having the patience to actually do it!

In many cases, these potential problems can be largely avoided through careful planning and installation design. If not, it is sometimes efficient to make a transition to coax cable for some portion of the transmission line run.

---

[1] B. Allison, WB1GCM, J. Hallas, W1ZR," A Closer Look at Window Transmission Line," *QST*, Nov 2009, pp 66-67.

What About Unbalanced Currents on Coax?

In the above section, it probably sounded as if the fields in coax are completely within the coax, thus avoiding the downsides of parallel or balanced line. That is true for the ideal case of properly terminated coax, however, there are at least three ways in which coax can suffer the same limitations as balanced line in this regard. In all cases, this results in currents on the outside of the shield. I have listed them in order of likelihood and severity, based on my experience.

- The coax is not terminated in a properly unbalanced load. This happens if the coax shield is not connected to a ground reference at the antenna. The result is that the current to the shield splits between the inside and outside of the shield.
- The coax run is coupled to the antenna such that the net coupling is not balanced at the cable. For a center fed antenna, coax should be run perpendicular to the antenna element so the coupling from each side cancels. If not, even with proper termination, there will be currents induced on the outside of the shield. For some antennas, monopoles or off center fed antennas for example, it is almost impossible to avoid such coupling.
- The coax shield itself can be leaky. The shield of most coax is not perfect and some is far from it. The specification sheets often list "percent coverage" indicating how well the shield covers the center conductor. Cables with lower numbers will be leakier than those with higher coverage. The current tends to increase as it is coupled over longer distances.

In many cases, the effects of such currents are small, but be aware that they exist and check them out if strange problems appear. A quick test is to note the antenna tuning or SWR as you run your hand along the cable. With properly installed and isolated coax, the tuning shouldn't change.

What Happens If I Feed my Dipole Directly with Coax?

Many amateurs connect their coax directly to a balanced center fed dipole — center conductor to one side, shield to the other as shown in **Figure 10-6**. This is a topic of some controversy with as many folk swearing that it works fine, as there are those who say it causes problems.

The reason for the disagreement may be that there are actually people who end up in both camps because of the dimensions of their system. Recall that due to skin effect, the currents on the inside of the shield of a coax cable are within a small thickness from the inside wall. Thus the outside of the coax acts like a completely different conductor — it's as if the insulated coax were installed in a pipe that is connected to the shield at the antenna end.

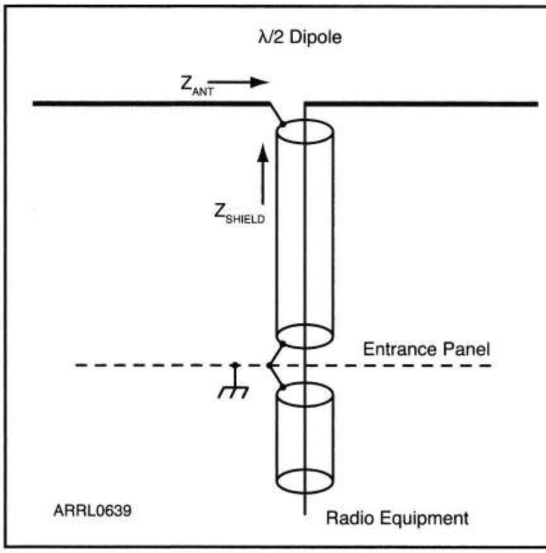

Figure 10-6 — Dipole fed directly with coax. The current on the left side divides between the left side antenna segment and the outside of the coax inversely proportional to the impedances as shown.

If the half wave dipole in Figure 10-6 had a center impedance of 50 Ω, typical for relatively low dipoles, the balanced antenna would act like a 25 Ω load to each connection of the coax, Z_{ANT} in Figure 10-6. The impedance of the outside of the shield at the antenna is a bit less obvious. This impedance depends on the impedance to ground at the bottom of the cable, usually at a ground level and grounded entrance panel, and the length of the cable between the antenna and the ground terminal.

Just as with an antenna wire or transmission line, this impedance varies depending on electrical length. If the ground impedance is low and the length is a multiple of a half wave, the impedance at the top will repeat and also be low. On the other hand, if the length is an odd number of quarter wavelengths, the impedance will reverse and be high. Lengths in between will result in intermediate values of impedance, both resistive and reactive.

If the impedance at the top of the shield is of the same order as the half dipole impedance, the current will divide between the two paths. This can have three effects:

- If the ground at the bottom is not of a very low impedance, RF currents can enter the radio room and cause a number of strange equipment problems such as feedback or transmitter lockup.
- The radiation from the transmission line will distort the antenna pattern, usually adding an omnidirectional vertical component. For a single element antenna, this may not be all bad, since it may tend to fill in the nulls that otherwise would be in some directions. For a directional array, however, such radiation will reduce both forward gain and front-to-back ratio — not good things.
- If the transmission line runs past sensitive electronic systems, fire or intrusion alarm wiring, computers or telephone systems, for example, the radiation may cause harmful interference in either direction.

Note that for a low impedance antenna, such as our dipole and most coax fed antennas, there are a small number of lengths that will result in such problems. With a multiband antenna the possibilities of problems increase directly with the number of bands covered. Thus it is not surprising that many don't think any precautions are necessary. Still, it can't hurt to avoid them, much of the subject of the next chapter.

Review Questions

10-1 Under what conditions is unbalanced transmission line most appropriate for antenna to radio connections?

10-2 Repeat question 10-1 for balanced transmission line.

10-3 Why do some people report problems with balanced antennas connected directly to coax while others observe no difficulties?

Chapter 11

So What's a Balun, an Unun, a Choke?

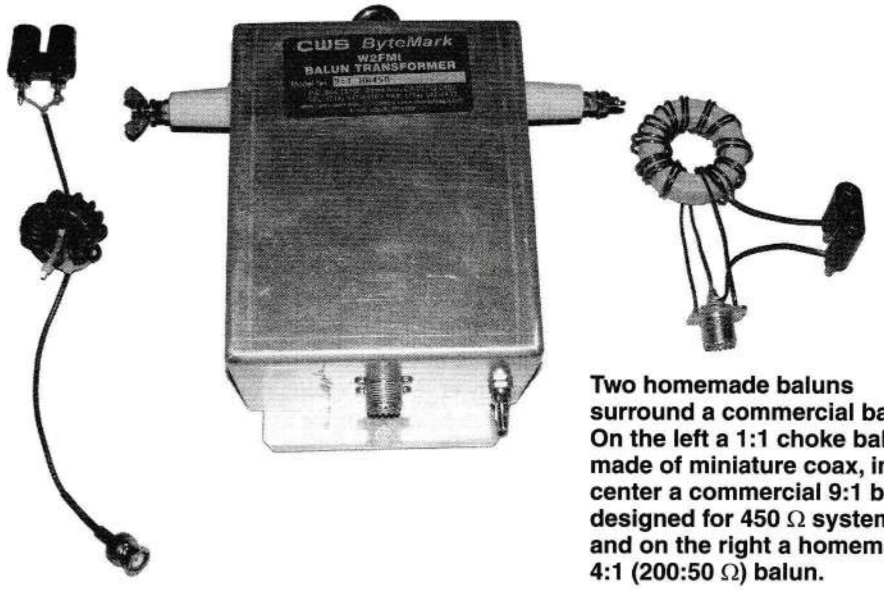

Two homemade baluns surround a commercial balun. On the left a 1:1 choke balun made of miniature coax, in the center a commercial 9:1 balun designed for 450 Ω systems and on the right a homemade 4:1 (200:50 Ω) balun.

Contents

When are Baluns Necessary? 11-2
What Do We Mean by Balanced to Unbalanced? 11-3
Balun Location 11-4
The Different Balun Configurations 11-5
Selecting Baluns for Use in Your System 11-7
Review Questions 11-8

Baluns are devices that provide a transition between unbalanced and balanced systems, impedance transformation or a combination of both. The name *balun* comes from *BAL*anced to *UN*balanced and each part is pronounced the same way as in the separate words.

Sometimes you will encounter an *unun.* The unun (*UN*balanced to *UN*balanced) provides an interconnection between two unbalanced systems, either to change the system impedance, to force all currents into the inside of a coaxial transmission line or usually both.

A *choke* refers to an inductance that offers a high impedance over a particular frequency range. A choke wound with coaxial cable has its high impedance only to currents on the outside of the coax shield, the inductances of the inner conductors effectively cancel. As we will discuss, such a choke can be used either as a balun or an unun.

When are Baluns Necessary?

We have discussed unbalanced line such as coax cable, balanced line such as twinlead, window or ladder line as well as balanced and unbalanced loads. We should point out that virtually all transmitters or transceivers built since the 1950s are designed to work into an unbalanced load — made evident by the coaxial connector used for the antenna connection.[1]

There are many antennas designed to operate with a coaxial cable connection. These antennas can be connected directly to the radio with coaxial cable if they present the desired load to the transmitter. In this case, no transition is required, the entire system is unbalanced. This system can work well and neither an antenna tuner, nor a balun, is required for proper operation. The one exception occurs if there is coupling from the antenna to the transmission line, resulting in undesired coax shield current in which case a common mode choke is recommended.

What's Wrong with this Picture?

The unbalanced antenna, unbalanced transmission line and unbalanced radio make a simple and straightforward arrangement that is easy to operate. Unfortunately, in many cases, the antenna only provides an appropriate match over a narrow frequency range — often narrower than an amateur band, for example. This has two effects:

- If the SWR at the transmitter reaches a higher value than the transmitter can operate into, it will usually reduce the power output to avoid damage to internal components.
- While the losses for matched coaxial cable can be made acceptable by selecting the proper cable for the frequency and length, the losses generally go up quickly with increasing SWR. As noted previously, this actually makes the SWR at the transmitter look better; however, significant power may be lost (turned into heat) in the cable.

Many amateurs choose instead to use a balanced antenna fed with lower loss balanced line in this case. As will be discussed later, this often makes what would be a single band antenna, if fed with coax, into a multiband antenna. We then are faced with a balanced load and an unbalanced transmitter output connector.

What Happens if We Plug the Balanced Line into our Coax Connector?

While not recommended. you can actually make the connection, if pressed. Arguably, it is not much worse than the previously discussed case of hooking unbalanced coax line to a dipole antenna. In this case, however, the shield side of the transmitter output is connected to ground and a fraction of your output power will flow in that direction radiating within the room instead of towards the desired direction. If the ground is not solid, you can end up with RF on all your equipment, causing various strange and sometimes uncomfortable effects.

---

[1] The shift to the use of coaxial cable happened because coaxial cable became available after WW2 coincidentally with the popularity of broadcast television. Transmitters with shielding, filtering and coaxial RF connections were less likely, if properly designed, to cause interference problems with TV sets.

What do We Mean by Balanced to Unbalanced?

The balun makes the transition from an unbalanced system, such as the transceiver, to a balanced system, such as a balanced transmission line. A balun can be used between a balanced dipole and a coaxial cable to avoid the problem of current being driven down the shield of the coax, or it can be used between the unbalanced transmitter and a balanced line to result in the current going towards the antenna as shown in **Figure 11-1**.

As noted previously, sometimes a balun is also used to transform impedances. For example, in the case of a 50 Ω antenna, a 50 Ω transmitter and the use of low loss 450 Ω transmission line, a balun with a 9:1 impedance transformation ratio allows a straightforward interconnection as shown in **Figures 11-2** and **11-3**.

Of course there is still no such thing as a free lunch. In order to use the low loss balanced line in this configuration, the added loss in each of the two baluns must be considered as part of the equation. Measurements I've taken of commercial HF 9:1 baluns have indicated that losses of about 0.5 dB each should be anticipated. It doesn't usually take too long a run of line before the balun loss is compensated for by the higher loss of the coax. This depends on both length and frequency.

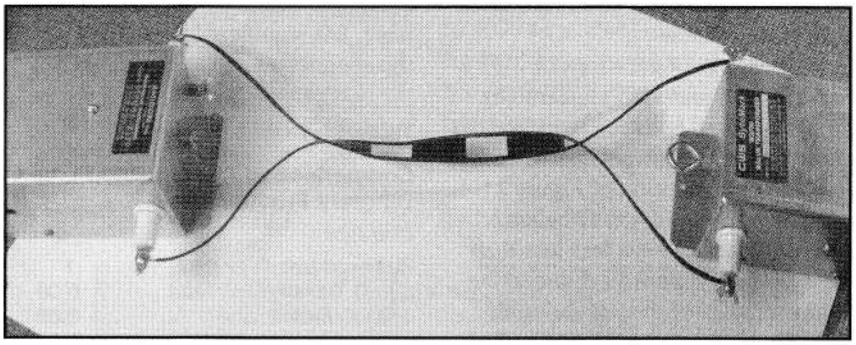

Figure 11-3 — Photo of a model of the system of Figure 11-2 using commercial 9:1 baluns. The twist in the window line is intentional. An occasional twist reduces wind effects and tends to maintain balance in the system if near objects.

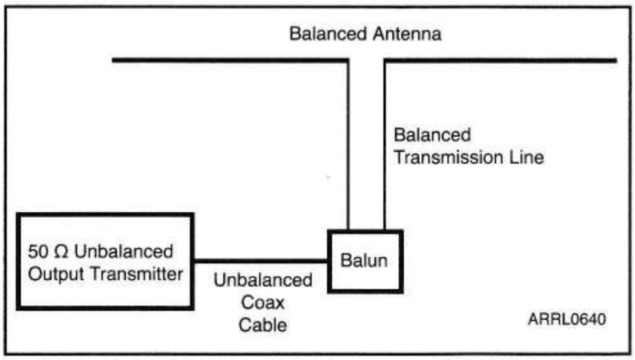
Balanced Antenna
Balanced Transmission Line
50 Ω Unbalanced Output Transmitter
Balun
Unbalanced Coax Cable
ARRL0640

Figure 11-1 — A balun inserted between the unbalanced output of a transmitter or transceiver and a balanced antenna system.

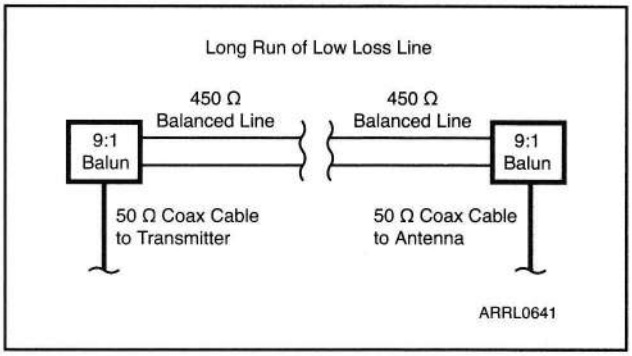
Long Run of Low Loss Line
450 Ω Balanced Line
450 Ω Balanced Line
9:1 Balun
9:1 Balun
50 Ω Coax Cable to Transmitter
50 Ω Coax Cable to Antenna
ARRL0641

Figure 11-2 — The use of baluns with a 9:1 impedance transformation ratio makes it easy to use low loss 450 Ω balanced transmission line for long runs.

Balun Location

A balun can be inserted anywhere in a system between a coaxial connection and the balanced load. As shown in **Figures 11-4** and **11-5**, the connection between an unbalanced radio and a balanced dipole can include a balun at either end of the transmission line run. Here we have used a dipole with a *delta match* adjusted to provide a 450 Ω balanced load at the antenna.

While the configurations of Figures 11-4 and 11-5 appear to be equivalent, and both can work well as shown, there are advantages to each:

- The configuration of Figure 11-4, using a long run of 450 Ω balanced line, will usually have less loss than the system in Figure 11-5 due to the characteristics of the transmission line. The difference will depend on the frequency, line length and type of coax. Some representative loss figures are shown in **Table 11-1**.
- Using coaxial cable for the run, as in Figure 11-5, has its own advantages. The principal one is that coax is less fussy about how it is placed than is balanced line. While coax can be run in metal conduit, coiled up and some types can be buried — this is not the case with balanced line. With properly installed coax, the fields are entirely within the coax, making it insensitive to its surroundings. Parallel conductor balanced line, on the other hand, has significant fields between the conductors and extending out to a few times the wire separation in all directions. In order to obtain the benefits of such line, it needs to be spaced that far from metal structures and lossey media.

Note also that it is not necessary to have the run entirely with one type of line. For example, a long aerial run of balanced line can be transitioned to coax near the radio with a short run of coax going through conduit or other balanced line hostile environments.

Table 11-1

Comparison of Transmission Line Losses of Different Lines at Different Frequencies

Line Type	*Matched Loss at Frequency (MHz, dB per 100 feet)*			
Balanced Line	*1*	*10*	*100*	*1000**
300 Ω Transmit Twinlead	0.09	0.3	1.1	3.9
450 Ω Window Line	0.02	0.08	0.3	1.1
600 Ω Open Wire Line	0.02	0.06	0.2	0.75
Coax Cable	*1*	*10*	*100*	*1000*
RG-58	0.37	1.4	5.3	20
RG-213	0.18	0.7	2.5	8.6
LMR-400	0.12	0.39	1.3	4.2

*Spacing may be too wide for efficient operation at this frequency.

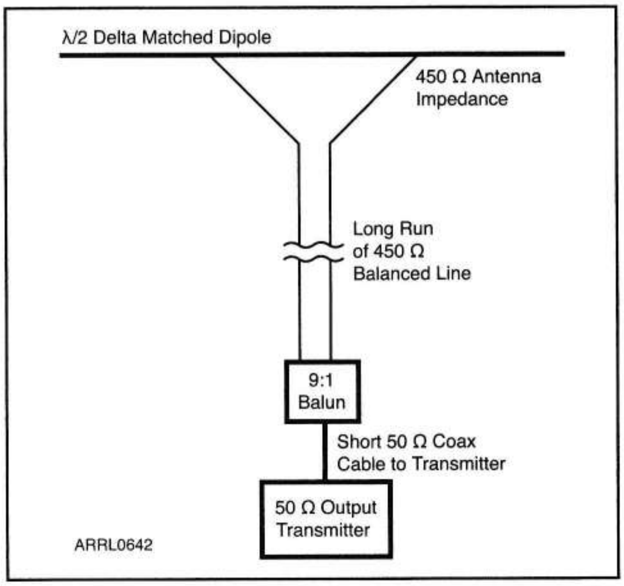
λ/2 Delta Matched Dipole
450 Ω Antenna Impedance
Long Run of 450 Ω Balanced Line
9:1 Balun
Short 50 Ω Coax Cable to Transmitter
50 Ω Output Transmitter
ARRL0642

Figure 11-4 — Delta matched 450 Ω dipole fed by low loss 450 Ω balanced line to a 9:1 balun at the transmitter.

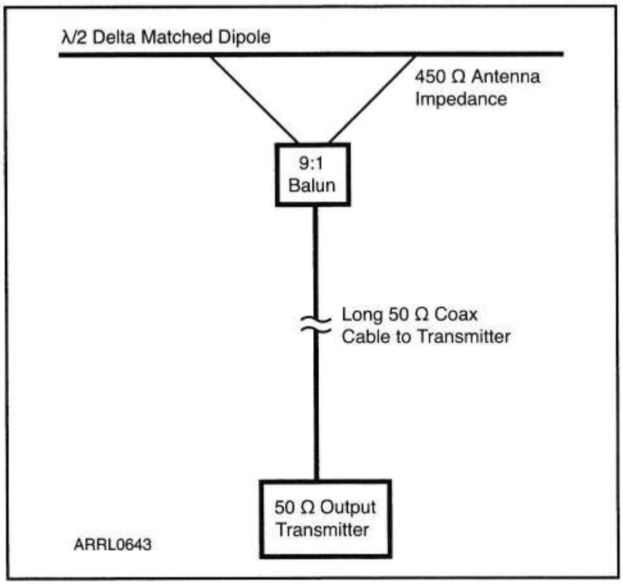
λ/2 Delta Matched Dipole
450 Ω Antenna Impedance
9:1 Balun
Long 50 Ω Coax Cable to Transmitter
50 Ω Output Transmitter
ARRL0643

Figure 11-5 — Dipole fed by a 9:1 balun and 50 Ω coaxial cable to the transmitter.

The Different Balun Configurations

There are a number of different balun configurations. They range from simple chokes that force (most of) the current that otherwise would be on the shield of coax to the intended load to more complex devices that combine balanced to unbalanced transitions and impedance transformations in a single device. The following is a summary of the more common types.

Choke Baluns

Choke baluns are found on the coax side of a balanced to unbalanced transition. As noted previously, if a balanced load, such as a dipole antenna, is fed directly by coax, the current on the shield side of the connection will split between the connected half dipole and the outside of the shield in a ratio inversely proportional to the two impedances. The idea of the choke balun is to increase the impedance of the outside of the shield by adding inductance and thus reducing the current that flows down the outside of the shield. If no current were to flow down the shield the current to the each side of the antenna would be the same as if the coax were a balanced line on the antenna side.

Air wound coax loops — The simplest form of choke is just a coil of coaxial cable as shown in **Figure 11-6**. Note that for the so called *differential mode* signal, the desired signal inside the coax, this coil just acts like a few feet of transmission line. The *common mode* signal, any undesired signal on the outside of the coax, sees the effect of the coil inductance.

This configuration has been evaluated at length with results reported in *The ARRL Antenna Book*.[2] An example is a six turns in single layer coil of coax with a diameter of 4.25 inches. This coil will have an impedance ranging from 514 Ω at 14 MHz to 1079 Ω at 29 MHz, very suitable for feeding a dipole or triband Yagi, for example. Such coils are sometimes scramble wound rather than being wound in a single layer. While

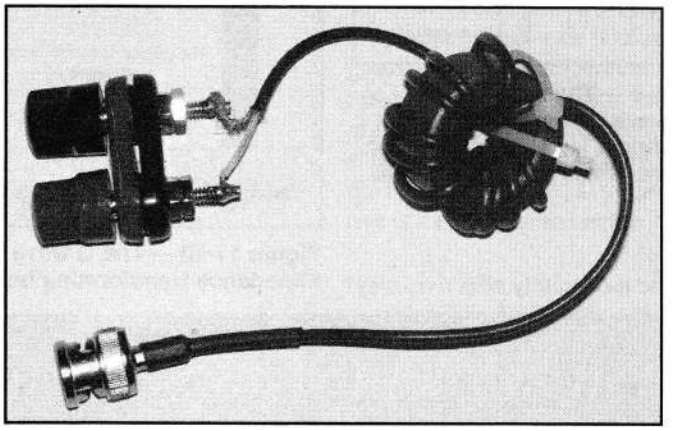

Figure 11-7 — Choke balun made by passing multiple turns of coax through a ferrite toroidal core.

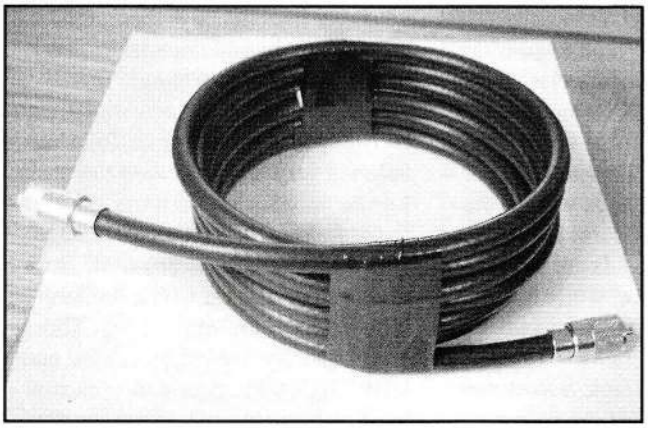

Figure 11-6 — Choke balun made from a coil of coax. Such a balun can be effective over about a 2:1 frequency range.

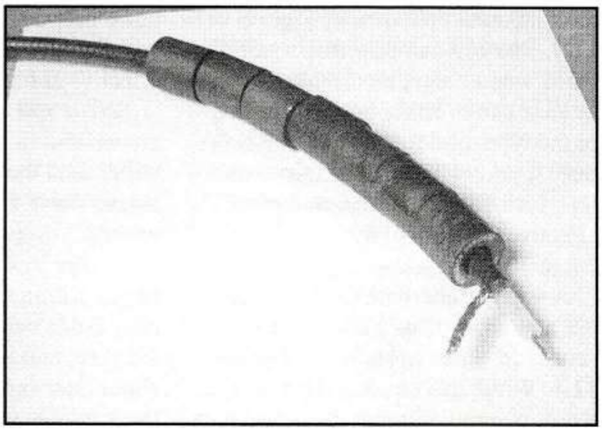

Figure 11-8 — Using multiple single turn "coils" in series — a good solution for large diameter or stiff coax.

---

[2] R. D. Straw, Editor, *The ARRL Antenna Book*, 21st Edition. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9876. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/arrl-store**; **pubsales@arrl.org**.

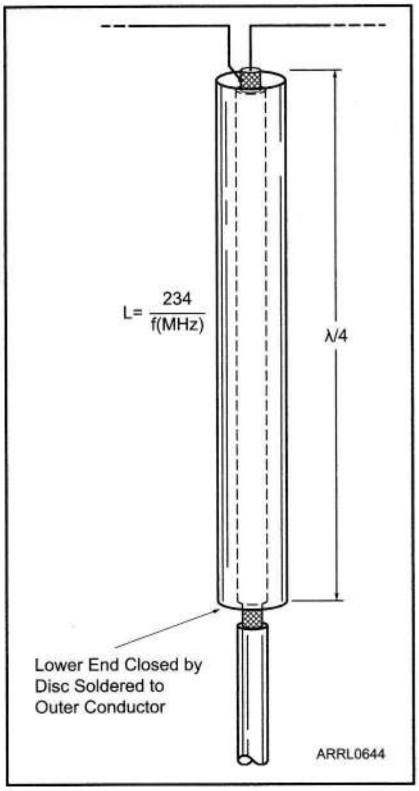

Figure 11-9 — The bazooka quarter wave choke balun made by providing a shorted line section around the shield of a coax cable.

they can be moderately effective, they usually suffer at higher frequencies due to excessive capacitance between turns. The only other effect of the coax coil is any attenuation due to the additional length of coax, in this case about 7 feet.

Ferrite toroid coils — The effectiveness of a choke coil can be improved by winding it on a ferrite toroidal core, as shown in **Figure 11-7**. The coil can be either a pair of wires wound together around the core or coax cable, either extra thin coax**,** or the type used for the transmission line. A coil of 12 turns of thin coax on a type FT240-61 core makes an effective choke over the range from 1.8 to 30 MHz.

A popular alternate method is to use a number of single turn ferrite "coils" in series as shown in **Figure 11-8**. While this can be effective, it is worth pointing out that the inductance of N such inductors in series equals N times the inductance of a single inductor. On the other hand, the inductance of multiple turns through a single toroid goes up by the square of the number of turns. Still if the coax is thick, or hard to bend tightly, this is a very workable solution. It has the additional advantage that the capacitance between "turns" is less than with a coil, so it will tend to be more effective at higher frequencies.

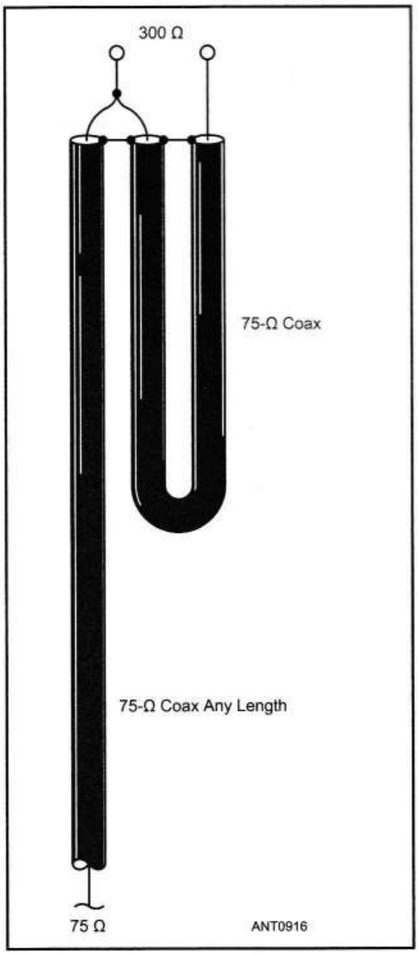

Figure 11-10 — The ½ wave 4:1 impedance transforming balun.

Ferrite rod coils — Some baluns are wound on sections of ferrite rod rather than the toroid, donut shaped, cores. These can work as well as the toroids, however, the toroids have the advantage that the fields are contained within the core, while the rods have fields between the ends outside the core, making them more fussy about their surroundings. Sometimes the available space will point towards one versus the other. All other things being equal, I prefer the self shielding toroids, if feasible.

Transmission line sections — A quarter wave long section of transmission line, shorted at the far end, has a high impedance at the near end. This property can be used as a choke at the particular frequency that the line section is that length (or odd multiples). This is called a bazooka, due to the use of large concentric structures as shown in **Figure 11-9**. This is a particularly popular arrangement at VHF and higher, since the dimensions are reasonable and wideband properties of the ferrite baluns and chokes are often not required.

Transformer Baluns

The half-wave coax loop — One of the early forms of transforming balun was also made of coax cable, this time an electrical ½ wave long. This cable was formed into a loop with one end on each side of the balanced load as shown in **Figure 11-10**. The unbalanced line was connected to one side of the balanced load.

The half wave loop provides a copy of the signal from the coax feedline delayed by 180°, and thus out of phase. Each of the two signals see an impedance of half of total 300 Ω load, or 150 Ω. Because the impedance on a halfwave (or multiple) of a transmission line repeats itself, the loop will bring the 150 Ω impedance back to the feedline junction. The parallel combination of the two 150 Ω signals at the feedline junction is thus ¼ of the total load or 75 Ω. Note that since a half wave length of transmission line repeats whatever impedance is there, this will transform any impedance to ¼ the value on the single frequency.

The half-wave loop transforming balun is very efficient, with the only loss or imbalance a result of the loss in the loop of coax. A disadvantage is that they only operate on (or close to) the frequency at which the loop is ½ wave (or multiples) long. These are frequently used to feed VHF and UHF Yagis that operate on a narrow band of frequencies making the narrow bandwidth irrelevant.

Ferrite transformers — Broadband

baluns can be made using the same kind of ferrite structures described previously as transformers. They frequently are made with multiple windings that can be connected in series or parallel to provide multiple transformation ratios. The 4:1 balun shown in **Figure 11-11** is an example of such a balun that functions similarly to the ½ wave coax loop balun.

Current and Voltage Baluns — Depending on the details of their design, a balun can be made to force either the current *or* the voltage on the balanced side to equal each other. Note that in a perfectly balanced load, in which each side has the same impedance to ground, if the voltages are equal so will be the currents and vice versa.

Unfortunately, in our real world, often an intended balanced load, isn't quite. For example, an otherwise balanced antenna may become unbalanced if one side is closer to the ground, or to a metal structure, than the other. If the balun is intended, as most are, to keep currents off the outside of coax shields, then by forcing the currents to be equal your objective is most likely to be met.

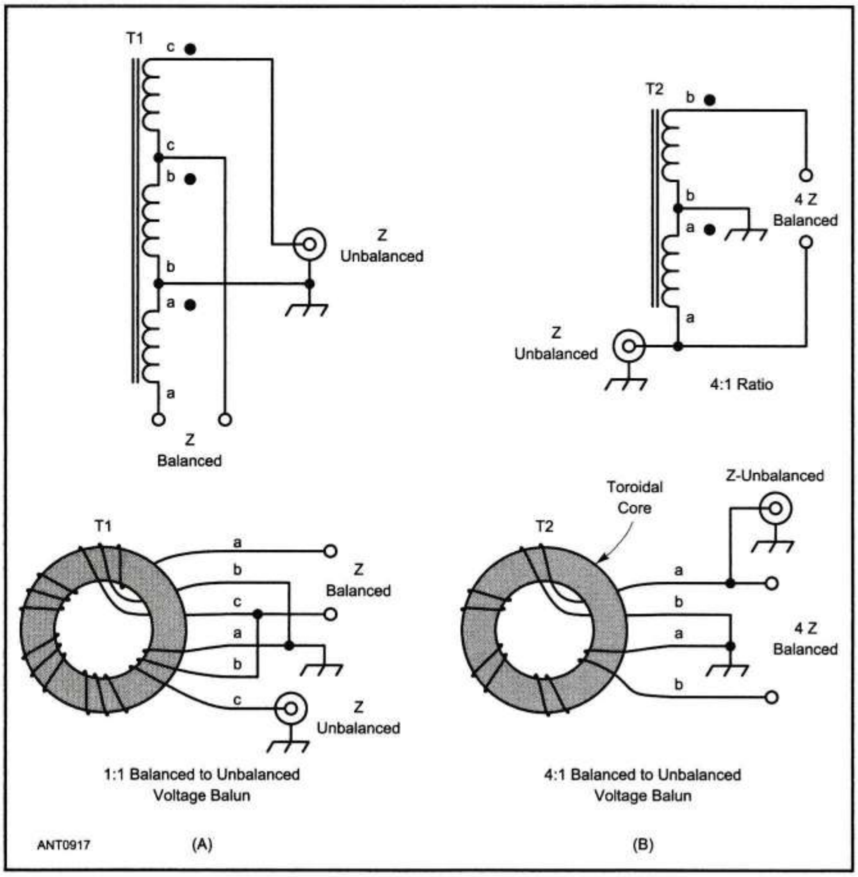

Figure 11-11 — Schematic and pictorial view of a broadband ferrite core 4:1 impedance transforming balun.

Selecting Baluns for Use in Your System

Within the constraints listed above, any of the general type of baluns described can be selected for use. Most baluns have a specified input and output impedance and power rating. These should be matched to your application. The transformation ratio does not apply for any impedance, rather for the design center of the balun. For example a 4:1 balun designed to transform a 200 Ω balanced load to 50 Ω coax, can not counted on to provide a 4:1 transformation to a 2000 Ω balanced load.

In practice, the transformation is likely to be close to the specified ratio for perhaps a 4:1 mismatch — in the previous case that would be from 50 to 800 Ω. Outside of that range, they may continue to function as a balun, however, the transformation ratio will be different and the losses will tend to increase.

A mismatch will also increase the stress on the balun. The current or the voltage will be higher by the square root of the SWR or mismatch ratio and thus the ratings should be increased appropriately. After use (with power off) a good check at a 100 W or higher power station is to put a finger on the balun to check the temperature. Any increase above ambient is cause for reconsideration.

Review Questions

11-1 Provide three reasons for using a balun.

11-2 What are the relative advantages of having a balun at a balanced antenna feed point versus having a long run of balanced line to a balun near the transmitter?

11-3 What are the consequences of operating a balun at an impedance far from its design point?

Chapter 12

Balanced Antenna Tuners

A classic 275 W E. F. Johnson "Matchbox" balanced antenna tuner from the 1950s. Still a viable tuner useful to 750 W PEP and available at hamfests and auction sites. A 1 kW model, useful above the US legal limit, was also offered.

Contents

What's a Balanced Tuner? 12-2

A Balanced Tuner versus an Unbalanced Tuner with a Balun 12-3

Another Possibility — The "Hot" Tuner 12-4

Review Questions 12-5

We've been talking about antenna tuners and balanced line in the same sentence, even though we haven't addressed the issue of the kind of tuners to use with a balanced load. Now it's time.

What's a Balanced Tuner?

While we haven't specifically addressed the question, the tuners we've been talking about so far have been really what could be called *unbalanced* tuners. You can usually spot an unbalanced tuner by its input and output connector arrangements. In some cases, they have a coax connector for unbalanced cable on both the radio and antenna system sides. Other unbalanced tuners have a coax connector for the radio side and a single terminal intended for a wire antenna and another for a ground connection.

Some unbalanced tuners have another pair of connectors that are designated for "balanced loads" but the transition occurs after the tuner function using an internal balun. While these can be useful for driving balanced loads, as we will discuss, they are really two distinct pieces, an unbalanced tuner followed by a balun.

Enter a Truly Balanced Tuner

The tuner configurations shown in Figure 3-3, in Chapter 3, are all inherently unbalanced since ground is on one side of each. Any of them can be converted to a balanced configuration by essentially providing a mirror image on the other side of ground. The configurations of Figure 3-3 are shown in balanced form in **Figure 12-1**. A quick look at the figure will indicate that we now have something we probably didn't want — a balanced to balanced tuner!

The one configuration that can be easily transformed to a balanced to unbalanced tuner is the tuned transformer in Figure 12-1 (A). This is because the transformer coupling to the radio (left) side of the transformer does not maintain the same ground reference. Thus one side of the primary (left) winding can be grounded resulting in an unbalanced connection. This is exactly what the 1950s era Johnson Matchbox did, perhaps explaining why it is still in demand today.

Unbalancing One Side

The other configurations in Figure 12-1 require a balun on the 50 Ω side to fit into coaxial fed systems. While this may seem about the same as putting the balun on the antenna side of the tuner (more in next section), the difference here is that we transform the widely variable balanced antenna system impedance on the right side to a 50 Ω balanced load on the left in Figure 12-1. We then can use a 1:1 50 Ω balun to transform to unbalanced for coax connectivity. Alternately, some tuners are designed to transform to a 200 Ω balanced load, to be followed by a 4:1 balun for connectivity to 50 Ω coax.

In either the 1:1 or 4:1 case, we have the advantage that the baluns are operating at their design impedance. They will offer minimum loss and will operate with their expected transformation ratios.

There's Always a Price

A possible downside of the balanced antenna tuner designs is that the nature of the beast requires significantly more — but not quite twice — as many components as the same unbalanced configuration. The usual antenna tuner components, for medium to high power tuners, are both large and pricy. This results in ramifications in terms of both cost and size. There is, after all, still no such thing as a free lunch.

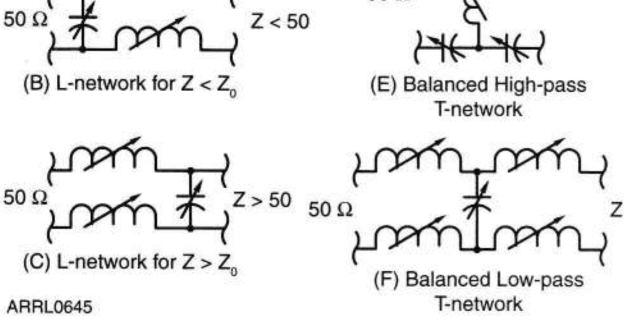
50 Ω
Z
(A) Balanced Tuned Variable Transformer
50 Ω
Z
(D) Balanced π-network
50 Ω
Z < 50
(B) L-network for Z < Z_0
50 Ω
Z
(E) Balanced High-pass T-network
50 Ω
Z > 50
(C) L-network for Z > Z_0
50 Ω
Z
(F) Balanced Low-pass T-network
ARRL0645

Figure 12-1 — Configurations of balanced antenna tuner versions of the unbalanced tuners in Figure 3-3, in Chapter 3.

A Balanced Tuner versus an Unbalanced Tuner with a Balun

As you look over the features of commercial antenna tuners, you will see many with "balanced outputs." In most cases, these are inherently unbalanced tuners, with a balun on the antenna side to transform to a balanced load. Many people, including your humble servant, have been happily using such tuners for years — so what's the problem?

Well, the simple answer is — there may not be a problem! If the impedance at the bottom of your balanced feed line (generally quite different from either the antenna impedance or the transmission line characteristic impedance, if the line is not matched to the antenna) can transform to an unbalanced impedance through the balun without difficulty — there is no problem — you're done!

On the other hand, if the impedance at the balun is very different from its design point, you may have a number of problems, most significantly loss in the balun. While balun loss is not great in its own right — toroid cores are not very good heat radiators. Any balun loss will transform itself into heat and may destroy the balun's ferrite core in the process.

The Good News Is

While there are exceptions, many unbalanced tuners, including those with a balun for balanced loads, offer additional features compared to many "balanced" tuners. My fairly typical unbalanced L-network tuner, for example, has three outputs for unbalanced coax terminations and a fourth that can either be coax or balanced, using an internal 4:1 balun on the output side. In addition to my main balanced multiband antenna, I have a few other antennas that are fed by coax. Thus the tuner provides a convenient point for them all to come together with easy switching between them. With a single switch on the tuner, I can instantly select any of the four antennas — either through the tuner, or bypassing the tuner, if they are matched systems. Most balanced tuners just provide a connection to a single balanced antenna system.

In terms of keeping common mode RF currents out of your station, it really doesn't matter which side of the tuner has the balun — its choking impedance has the same effect.

If it Doesn't Play

If an unbalanced tuner with a balun doesn't work well with your antenna system on a particular band, you may be able to recover. The fact that it doesn't tune usually means that the balanced line has a high SWR (if it were matched, to say 450 Ω, the tuner and balun combination could probably handle it without difficulty). Because of the nature of the mismatched line, changing the length will have a profound effect on the impedance as seen by the tuner. Try inserting about ⅛ wavelength of line and see what happens. Be sure to observe the usual cautions about balanced line — such as not coiling it up.

Another Possibility — The "Hot" Tuner

Impedance transformation through an unbalanced tuner will not "know" it is unbalanced unless one side of the tuner is grounded. Thus it is quite possible to use an inherently unbalanced tuner into a balanced load if it is isolated from ground. This can be accomplished as shown in **Figure 12-2** for a remote automatic tuner.

The configuration in **Figure 12-3** makes this quite convenient, however, it can also be accomplished using a manual tuner in the station. The difficulty with the tuner in the station is that it generally is designed so that the cabinet is at what it thinks is "ground" potential. If it is isolated from ground, the whole cabinet is at the potential of half the RF voltage feeding the antenna. This can be uncomfortable for the operator, unless she uses oven mitts to make adjustments. An unsuspecting cat can get quite a curl as it passes by during a transmission!

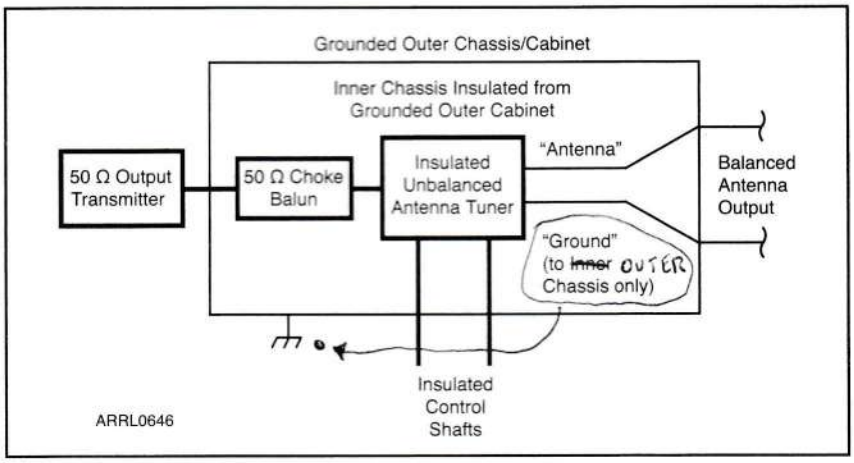
Grounded Outer Chassis/Cabinet
Inner Chassis Insulated from Grounded Outer Cabinet
50 Ω Output Transmitter
50 Ω Choke Balun
Insulated Unbalanced Antenna Tuner
"Antenna"
Balanced Antenna Output
"Ground" (to ~~Inner~~ OUTER Chassis only)
Insulated Control Shafts
ARRL0646

Figure 12-3 — Manually tuned version of an unbalanced tuner isolated from ground driving a balanced load. Note the insulated additional layer of cabinetry.

The hot chassis problem has been solved by at least one published design in which what would be the cabinet is insulated from and inside an additional cabinet with insulated shaft extensions on the controls as shown in **Figure 12-3**. I'm not aware that this has been offered as a commercial product, however, it is feasible — if not simple to implement.

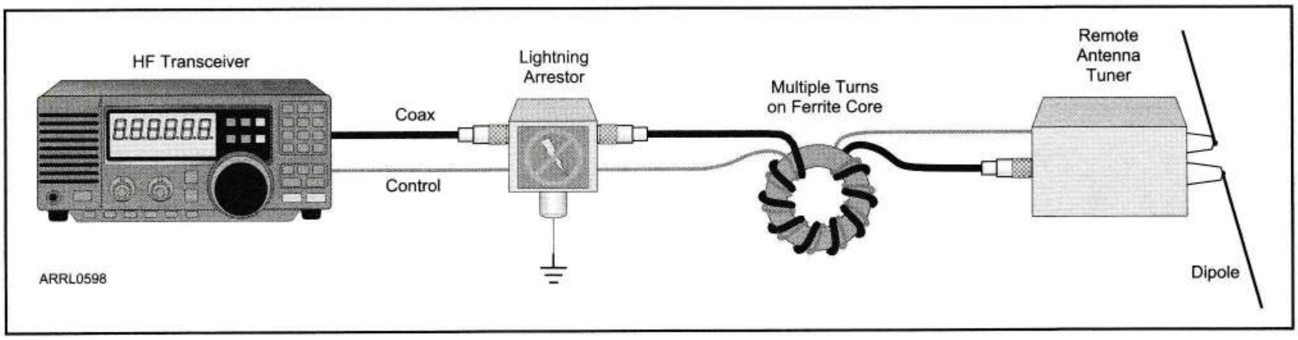
HF Transceiver
888888
Coax
Control
Lightning Arrestor
Multiple Turns on Ferrite Core
Remote Antenna Tuner
Dipole
ARRL0598

Figure 12-2 — Unbalanced tuner isolated from ground driving a balanced load. The common mode choke on the radio side maintains the balance on the antenna side. Note that it is necessary to provide a choke on all ~~balun~~ REMOTE TUNER connections, as shown, in order to elevate the whole tuner above RF ground.

Review Questions

12-1 Why might an inherently balanced tuner be a good idea?

12-2 What difference does it make whether a balun is on the input or output of an antenna tuner?

12-3 What kind of problems might you have using an unbalanced tuner with a balanced load?

Chapter 13

Antennas That Work Well with Antenna Tuners

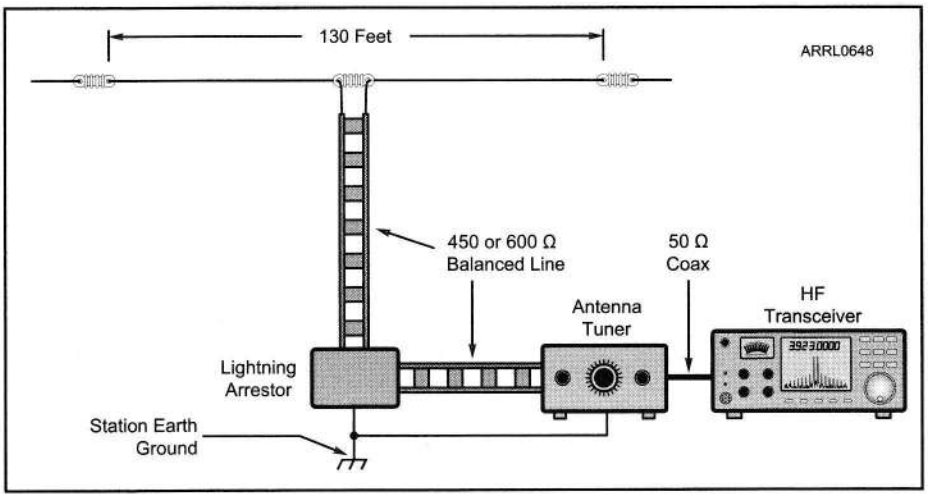
130 Feet
ARRL0648
450 or 600 Ω
Balanced Line
50 Ω
Coax
HF
Transceiver
Antenna
Tuner
3923.0000
Lightning
Arrestor
Station Earth
Ground

The center-fed Zepp is a very popular antenna with those having a wide-range antenna tuner.

Contents

The Coax Fed Dipole 13-2
Tuned Yagi Arrays 13-3
Single Wire Fed Antennas 13-4
The Vertical Monopole 13-5
Balanced Antenna Systems 13-7
Mobile Antenna Systems 13-8
Review Questions 13-10

There's no magic antenna that goes well with a tuner — in fact, virtually any antenna can benefit from a tuner under some circumstances. Here we will discuss how a tuner can increase the flexibility and usefulness of some popular antennas.

The Coax Fed Dipole

In Chapter 2, we discussed the nominally matched dipole and the limitations on its bandwidth within a given SWR. For example, to have a single low 80 meter dipole be useful across the entire band, we may need to deal with an SWR as high as a 10:1 as shown in Figure 2-7, in Chapter 2, although at many heights, the SWR will not rise beyond 5 or 6:1 at band edges. Most other examples shown in Chapter 2 have more reasonable SWR characteristics. Still, in many cases an SWR higher than 2:1 may be encountered, requiring an antenna tuner to permit proper transmitter operation at band edges.

A Dipole Operated on an Odd Harmonic

A special case of the dipole is one used on its third harmonic. A very popular amateur antenna is a 40 meter (7 to 7.3 MHz in the US) dipole also operated on 15 meters (21 to 21.450 MHz). This is often stated as if it were a straightforward arrangement, however, it is just a bit more difficult to achieve than would be implied by some. A tuner makes it much easier. The difficulties arise particularly due to two factors:

- The nominal impedance of a ¾ λ dipole in free space is about 110 Ω, rather than 72 Ω for a λ/2 dipole. While the actual impedance will vary above and below this number, depending on height above ground, this often gives us a higher SWR than 2:1 to start with.
- The third harmonic of 7.15 MHz, the center of the 40 meter band, is 21.45 MHz, the very top of the 15 meter band.

Figure 13-1 shows the modeled SWR of a wire 40 meter dipole cut for mid band at a height of 40 feet. **Figure 13-2** shows the SWR of the same antenna operated on 15 meters. Obviously, this isn't going to work out as well as some folks hypothesize.

Many wide range tuners could compensate for the SWR at the transmitter. As discussed in Chapter 7, however, the losses in a coax cable with a 10:1 SWR at 21 MHz may result in more loss than can generally be tolerated, although it certainly could be used in a pinch. Because the losses in mismatched coax are less at lower frequencies, the concept can be salvaged by optimizing the dipole as a 15 meter ¾ λ dipole with the results shown in **Figures 13-3** and **13-4**.

Note that the lengthened antenna has an SWR of less than 3:1 across all of 15 meters, significantly reducing the transmission line loss associated with Figure 13-2. While the SWR on 40 meters is no longer delightful, it is reasonable from a transmission line loss standpoint. The resulting antenna can work well with some nominal "3:1" SWR tuners that work beyond

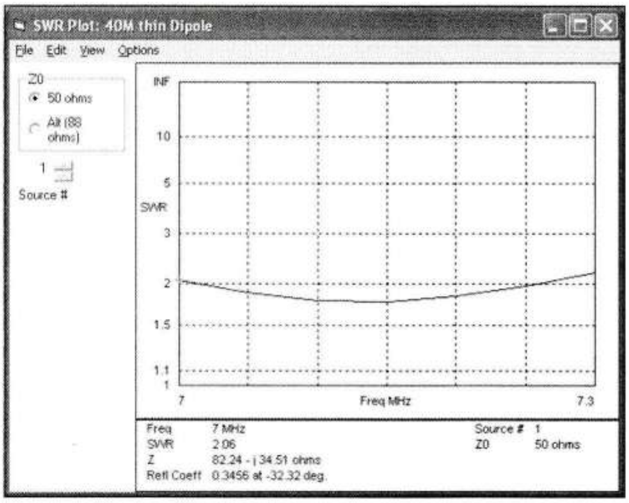

Z0: ● 50 ohms ○ Alt (88 ohms)

Source # 1

Freq 7 MHz
SWR 2.06
Z 82.24 - j 34.51 ohms
Refl Coeff 0.3456 at -32.32 deg.
Source # 1
Z0 50 ohms

Figure 13-1 — 50 Ω SWR of a 40 meter dipole at 40 feet cut for mid band.

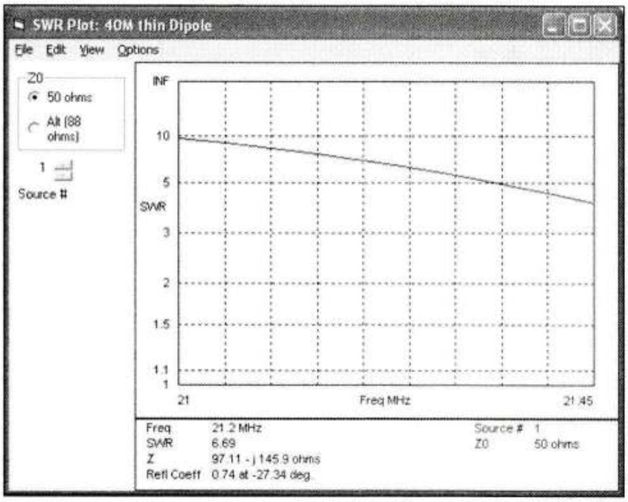

Z0: ● 50 ohms ○ Alt (88 ohms)

Source # 1

Freq 21.2 MHz
SWR 6.69
Z 97.11 - j 145.9 ohms
Refl Coeff 0.74 at -27.34 deg.
Source # 1
Z0 50 ohms

Figure 13-2 — SWR of the 40 meter dipole of Figure 13-1 across 15 meters.

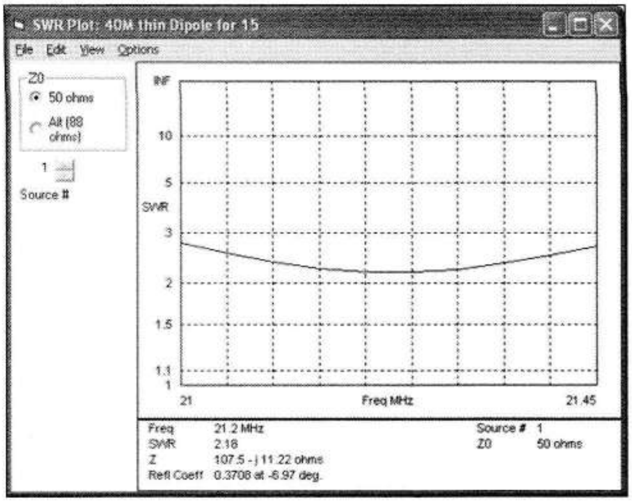
SWR Plot: 40M thin Dipole for 15
File Edit View Options
Z0
● 50 ohms
○ Alt (88 ohms)
1
Source #
INF
10
5
SWR
3
2
1.5
1.1
1
21
Freq MHz
21.45

```
Freq        21.2 MHz                         Source #  1
SWR         2.18                             Z0        50 ohms
Z           107.5 - j 11.22 ohms
Refl Coeff  0.3708 at -6.97 deg.
```

Figure 13-3 — SWR of the 40 meter dipole of Figure 13-1 lengthened to be optimized for 15 meters across 15 the 15 meter band.

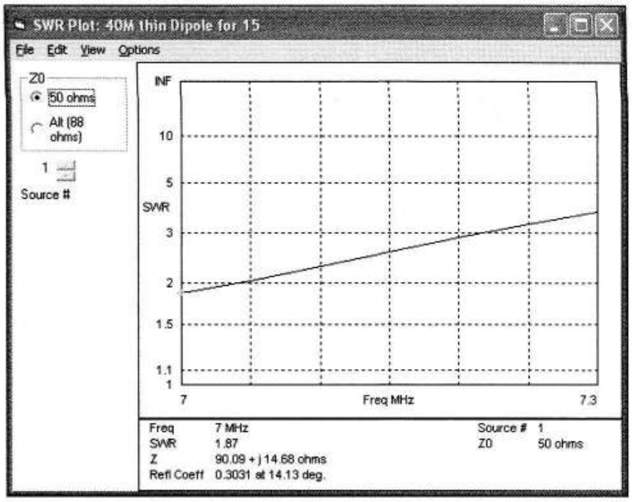
SWR Plot: 40M thin Dipole for 15
File Edit View Options
Z0
● 50 ohms
○ Alt (88 ohms)
1
Source #
INF
10
5
SWR
3
2
1.5
1.1
1
7
Freq MHz
7.3

```
Freq        7 MHz                            Source #  1
SWR         1.87                             Z0        50 ohms
Z           90.09 + j 14.68 ohms
Refl Coeff  0.3031 at 14.13 deg.
```

Figure 13-4 — SWR of the lengthened (15 meter optimized) 40 meter dipole of Figure 13-3 across 40 meters.

their specs, and is an easy job for wider range tuners.

Dipoles Can Benefit

The preceding examples illustrate ways in which a simple dipole can benefit from an antenna tuner. For any dipole, the tuner can compensate for imperfect length adjustment, often a convenience. The one caution is that even if the tuner can match the resulting impedance, the loss in the mismatched transmission line between the tuner and the antenna can be an issue, even if all seems well at the transmitter. Take another look at Chapter 7.

Tuned Yagi Arrays

We often think of the popular HF Yagi beam antenna as one that is fed directly with 50 Ω coax with low SWR and without any need for a tuner. This is certainly the sense one gets from the specification sheet, yet some highly tuned Yagis, particularly those operating on multiple bands, have trouble providing a close match from band edge to band edge, particularly if optimized for the CW or SSB segment. In most cases, even though the SWR may be marginal at the far edge, there is still beneficial gain and front to back ratio. This is exactly the place for a limited range internal tuner to be applied, since the SWR is often less than 3:1. My sense is this is what they had in mind when these 3:1 tuners were designed — compensating for the just out of tune coax fed dipole or Yagi.

Single Wire Fed Antennas

There are a number of antennas that are designed to be fed as a single wire, without any transmission line. Some are field expedients, however, others are well thought out designs that can offer excellent performance.

Single Wire Antennas

Single wire antennas go back to the very beginning of amateur radio — the days before anyone thought of a transmission line. The most commonly used transmitting antenna in the early days was a short vertical (we're talking below 200 meters, so everything was short!) with top capacitance loading of multiple horizontal wires. They were used on shipboard, shore and ham stations alike, and worked well enough to get the signals through.

These days, a "random length" single wire is often used as an easy to deploy temporary antenna or, for some, using very thin wire, as an unnoticeable stealthy antenna if neighbors or deed restrictions object to "antennas." They can be effective, especially if fed against a good ground system, and if they are at least $\lambda/4$ long at the lowest frequency. Such a $\lambda/4$ wire fed against ground even has a name! This is called a Marconi antenna, and yes. it dates back more than a century.

Random wire antennas will still work if shorter, but ground losses rapidly reduce efficiency as lengths get much shorter than $\lambda/4$.

The success of such single wiere antennas depends on all the usual factors including height, length, and ground conditions all played against desired frequency and geographical coverage. With a ground mounted radio and a higher wire antenna, there will always be some vertically and often horizontally polarized radiation as well. The combination can provide a useful mix of coverage, although the coverage in any direction will not be as good as that of an antenna designed for the job.

The key element in making such an antenna play is a wide range antenna tuner that can match the widely variable load of the single wire on multiple bands to the $50\ \Omega$ that modern radios require. So we have a place here for the antenna tuner, the question is "which place."

Antenna Tuner at the Shack — This can work and has been done, but in the most successful implementations, the antenna tuner is near the boundary between indoors and out. In some stations, that boundary is within or at the edge of the shack eliminating the distinction.

A portion of a radiating antenna inside a building is generally asking for some kind of trouble. First there is the RF safety issue as applied to any people inside the structure. Even if your transmit power level does not require a formal assessment, you are not released from the requirement to not expose people to excessive levels of RF. Second, your indoor receiving antenna will pick up electrical noise from other equipment in the building. Third, if the building contains any sensitive electronic systems, your signal has a good chance of causing trouble, such as setting off fire alarms, or getting in to a television receiver while a spouse is watching an important show. In any case, any radiation that occurs within the building structure is not likely to help you get your signal where you want it.

None of this is to suggest that indoor antennas can't be used, however all of the above issues need to be considered. Here we are talking about an antenna that is partly inside a building, and partly outside. The point is that having as much as possible outside is generally a better solution.

Antenna Tuner at the Antenna — The antenna tuner can also be located at the antenna, if it's remotely tuned or sometimes remotely switched. Some remote tuners are waterproof, while others will need a waterproof enclosure to make them last. With the tuner at or near the antenna feed point, matched coax can run from the tuner back to the station, as discussed previously. By running the coax from the station out to the antenna feed point, the issues of radiation within the building can be largely avoided.

This is the configuration used for

Table 13-1

Random Wire Antenna Lengths that Should be Avoided Due to Resonance and High Impedance.

Band (Meters)	0.5λ	1λ	1.5λ	2λ	2.5λ	3λ
160	234 to 260	468 to 520	702 to 780	936 to 1040	1170.0 to 1300.0	1404.0 to 1560.0
80	117.0 to 133.7	234.0 to 267.4	351.0 to 401.1	468.0 to 534.9	585.0 to 668.6	702.0 to 802.3
40	64.1 to 66.9	128.2 to 133.7	192.3 to 200.6	256.4 to 267.4	320.5 to 334.3	384.7 to 401.1
30	46.1 to 46.3	92.2 to 92.7	138.3 to 139.0	184.4 to 185.3	230.5 to 231.7	276.7 to 278.0
20	32.6 to 33.4	65.2 to 66.9	97.8 to 100.3	130.5 to 133.7	163.1 to 167.1	195.7 to 200.6
17	25.8 to 25.9	51.5 to 51.8	77.3 to 77.7	103.0 to 103.6	128.8 to 129.5	154.6 to 155.4
15	21.8 to 22.3	43.6 to 44.6	65.5 to 66.9	87.3 to 89.1	109.1 to 111.4	130.9 to 133.7
12	18.7 to 18.8	37.5 to 37.6	56.2 to 56.4	74.9 to 75.2	93.6 to 94.0	112.4 to 112.8
10	15.8 to 16.7	31.5 to 33.4	47.3 to 50.1	63.0 to 66.9	78.8 to 83.6	94.5 to 100.3
6	8.7 to 9.4	17.3 to 18.7	26.0 to 28.1	34.7 to 37.4	43.3 to 46.8	52.0 to 56.2

many marine HF antennas. On a sailboat, the typical HF antenna is an insulated backstay fed by a remote single wire tuner located near the antenna base and fed against the ship ground system, including salt water, if possible.

But Watch Out for Tricky Lengths! — The so called *random length* wire antenna is not really random since any antenna does have a specific length. Such an antenna will work better at some lengths than others, making it best if not entirely random in its installation. The problems most often show up if the length happens to be a multiple of λ/2. At this length, the impedance hits a maximum, often in the 1000 to 2000 Ω range, depending on the usual factors. This results in an SWR of 20 to 40:1, higher than even a wide range tuner promises to deal with.

The impedance gets somewhat lower with each multiple of a λ/2, but may still be a problem. A second concern is that the voltage on the end at the tuner is high. The voltage can be nearly 500 V_{RMS} with a 100 W transmitter. This can stress tuner capacitors as well as pose a safety hazard to unsuspecting people. Instead of the shock experienced by coming in contact with dc or 60 Hz ac, contact with RF results in a painful burn that is usually slow to heal.

While it sounds like it should be an easy task to select a length, the multiplicity of bands without harmonic relationship makes it difficult to find a length that won't be likely a problem on some band. A look at **Table 13-1**, indicating approximate λ/2 resonant lengths for each band, along with multiples, provides a list of lengths to avoid for the amateur bands. There are so few lengths available, that sometimes it makes sense to just try the length that fits the available spot and, if it has problems tuning on a band, add a few feet and try again.

The Vertical Monopole

A popular antenna with those who like to operate long distances on the lower HF and MF bands is a vertical monopole. This can perhaps be considered a special case of the random wire antenna, except that it is generally of specific electrical length — typically λ/4 or ⅝ λ, depending on frequency and available supports. While a λ/4 vertical monopole can usually be fed directly with 50 Ω coax against ground, the other lengths do require matching networks or an antenna tuner for operation with a transmission line. A low impedance ground is a requirement for efficient operation of such antennas.

A 43 foot tall vertical made from wire or tubing is a very popular antenna in some quarters. It is a ⅝ λ long monopole on 20 meters and thus provides the optimum low angle radiation from a single vertical radiator (see **Figure 13-5**), yet can provide excellent low angle performance on 40 meters where it is somewhat longer than λ/4, as well as on 80 meters where it is somewhat less than λ/4, but generally a manageable height. All the lengths share a complex feed point impedance that responds nicely to a remote tuner at the base, although at these frequencies, a short run of high quality coax may have sufficiently low in loss to be suitable for use with a tuner at the station.

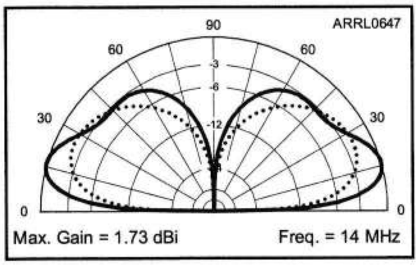

Figure 13-5 — Vertical radiation patterns of a 43 foot, ⅝ λ, vertical on 20 meters (solid) compared to a ¼ λ monopole. Note the increased radiation at low elevation angles.
THE DOTTED PATTERN IS FOR THE 1/4 λ MONOPOLE.

Inverted L

A popular variant of the vertical monopole, particularly on the lower frequency bands, is the inverted L. A λ/4 monopole for 160 meters, for example, needs to be around 130 feet long — a bit of a stretch for most as a vertical arrangement. The inverted L is obtained by taking the same length of wire and going as high as feasible, then bending the remainder horizontally. The antenna has the appearance of an upside down letter L — hence its name.

The lower (vertical) portion has the highest current and will thus provide low angle radiation very similar to that of a full sized monopole. At the λ/4 frequency, the inverted L can often be fed directly by 50 Ω coax with a reasonable SWR. An additional advantage of this arrangement, if fed with a wide range tuner at its base, is that it will provide low angle vertically polarized radiation at higher frequencies up to the frequency at which the vertical portion is just longer than ⅝ λ. Note the earlier concerns about feeding an antenna that is a multiple of λ/2, and if multiband operation is in the plan, make the inverted L longer than λ/4 on the lowest band.

The Bobtail Curtain

An interesting, inexpensive and surprisingly effective directional array called the Bobtail Curtain is composed of three λ/4 vertical elements fed in phase and driven from the single wire center element as shown

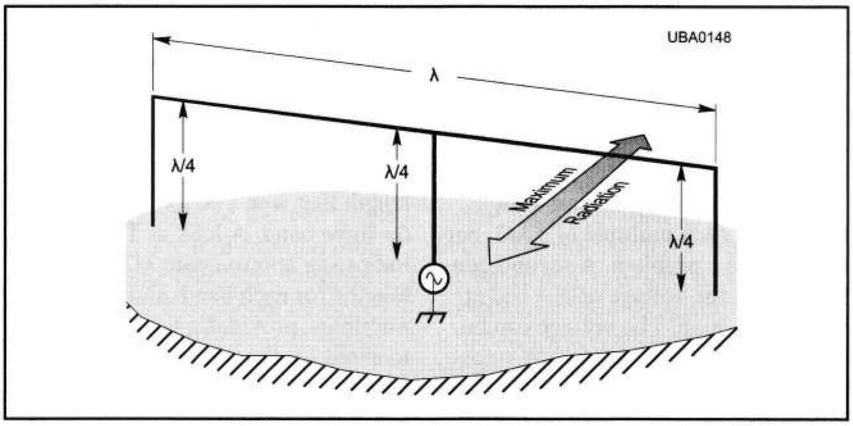

Figure 13-6 — The Bobtail Curtain three element vertical phased array fed by a single wire.

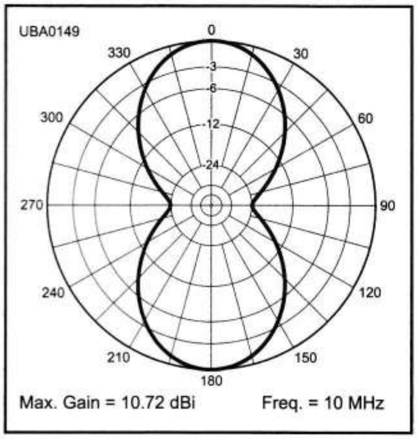

Figure 13-7 — Azimuth pattern of a Bobtail Curtain array at 10° elevation. Note the high gain over 60° on each side, broadside to the array.

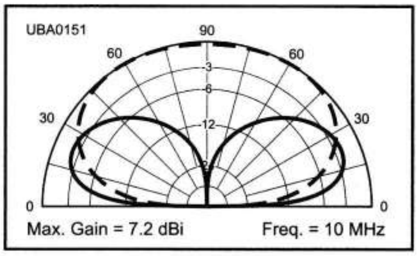

Figure 13-8 — Elevation pattern of a Bobtail Curtain array (solid) compared to a full wave center fed antenna on the same supports (dashed). Note the superior low angle coverage.

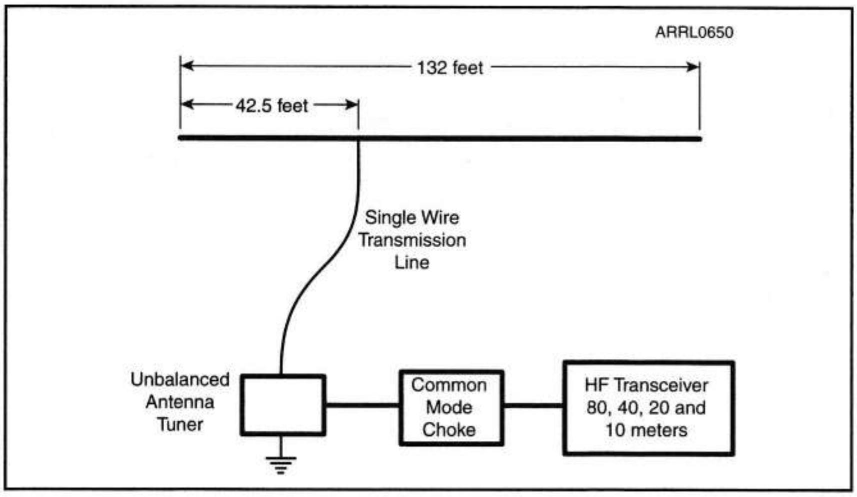

Figure 13-9 — Single wire fed Windom antenna configuration for 80, 40, 20 and 10 meters.

in **Figure 13-6**.[1] This antenna is composed of just four pieces of wire, yet provides significant gain in two fixed directions as shown in **Figure 13-7**. The low angle gain of this antenna favors long distance communication (see **Figure 13-8**) and is higher than that of horizontal antennas of the same height.

The Bobtail Curtain can be fed by a single wire tuner located at the source shown in Figure 13-6, although the impedance is quite high, as discussed in the previous section. Making the middle wire a few feet longer may make it usable with most wide range tuners. The antenna can be used on other bands as a kind of top loaded monopole, although the results will not be in the same league as those at its design frequency.

Single Wire Fed Windom

An unbroken half wave antenna off-center fed by a single wire was quite a popular multiband radiator in the 1930s, following publication of a note by Windom in *QST*.[2] The Windom antenna, as shown in **Figure 13-9,** with a fundamental frequency in the lower portion of 80 meters, was said to operate equally well on 40, 20 and 10 meters, thus covering all the amateur HF bands of the day.

There is some controversy over just how well this antenna works as a horizontal antenna; however, there is no question that it can work. Figure 13-9 shows a choke or current balun in the coax from the tuner to the radio, since there will likely be significant common mode current that should be kept out of the radio room.

[1] For more about the Bobtail Curtain array, see J. Hallas, W1ZR, *Basic Antennas — Understanding Practical Antennas and Designs,* Chapter 15. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9994. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop/**; **pubsales@arrl.org**.

[2] L. Windom, W8GZ, "Notes on Ethereal Adornments," *QST*, Sep 1929, pp 19-22, 84.

Balanced Antenna Systems

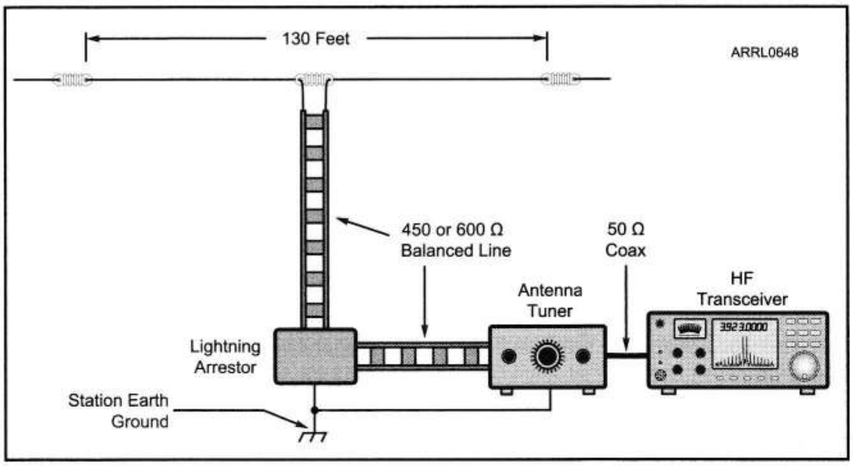

Figure 13-10 — Traditional center-fed Zepp antenna. With the length shown, the antenna will work well from 80 through 6 meters, and may even tune on 160 meters. The antenna length is not critical, however, it works best if close to λ/2 on the lowest operating frequency as described in the text.

For many amateurs, the antenna tuner really earns its keep by matching balanced systems for multiple bands. The classic system is one that is often called a center-fed Zepp, although there are some who suggest it really should have different names on different bands.

The Center-Fed Zepp

This antenna is actually a half-wave dipole if fed on its lowest band. The difference between it and the usual dipole is that it is fed with low-loss balanced transmission line, such as open wire line or window type line. Since it doesn't really need to be resonant to work well, and thus can be almost any length, a more accurate name might be the *not always resonant tuned feeder dipole*. Still more folk will know what you mean if you call it a center-fed Zepp (CFZ). The configuration is shown in **Figure 13-10**.

For the case in which it is resonant at its half wave frequency, the radiation pattern is the same as any half-wave dipole at the same height. The difference is that the SWR on the transmission line will be in the range of 4:1 to 10:1 depending on the height of the antenna and the transmission line type. While the loss with this kind of mismatch would be a problem with coax, it isn't with low loss line. Including the loss due to mismatch, the total loss is comparable or even less than most matched coax situations at the same frequency.

The antenna can easily be fed with a wide range antenna tuner, preferably a balanced antenna tuner, but usually also with an unbalanced tuner equipped with a balun. The antenna will work very well on higher bands — into the VHF range. There may be some lengths of transmission line that will result in an impedance that the tuner can deal with on all bands.

The antenna patterns become fairly complex at higher frequencies. For resonant antennas, there are generally 2 × N main lobes where N is the number of full wavelengths.

As mentioned, the antenna has no particular need to be resonant anywhere to work well. The impedance is easier to match if the antenna is at least a half wave long at its lowest frequency, but there is nothing magic about that length. I use a 100 foot long dipole, center fed with 450 Ω window line quite successfully on all bands from 80 through 6 meters, for example.

Another popular length is 86 feet. This provides the optimum gain on 20 meters in single lobes in each direction perpendicular to the wire. While the lobes are sharp compared to a dipole, 35.6° on either side compared to 83°, the gain at the peak is more than 3 dB higher than a dipole.

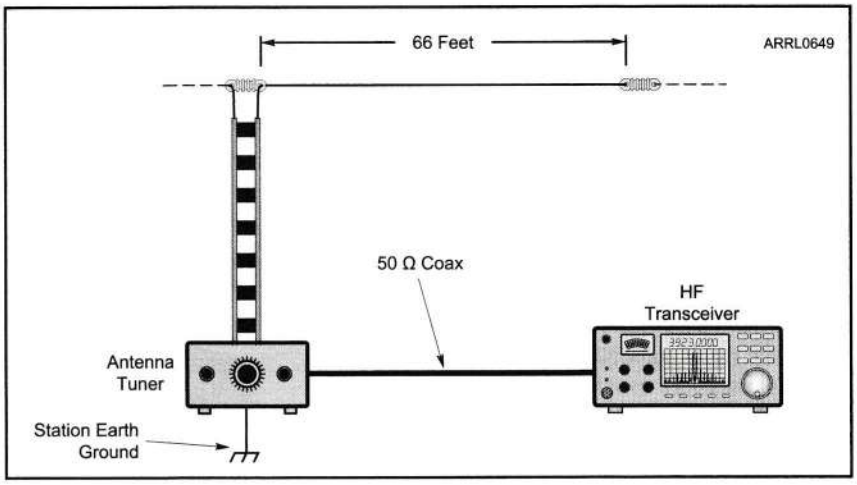

Figure 13-11 — End-fed Zepp antenna for 40, 20, 15 and 10 meters. This works as a horizontal antenna if the length is a multiple of λ/2 on each frequency used. On other frequencies the feed line will also radiate, usually with vertical polarization, and precautions must be taken to keep common mode currents from the radio equipment. Lightning arrestor not shown.

This length will usually be useable from 80 meters through the higher frequency bands, although some transmission line lengths may have problems with some wide range tuners on 80 meters.

If there are problems matching particular impedances, usually an extra 10 feet or so of the transmission line will change the impedance enough to make the match possible. Be sure to check the other bands to make sure you haven't just moved the problem around. Usually a length can be found that will work on all bands. Be sure to avoid coiling up the excess balanced line. Always keep it off the ground and away from metal — at least three or four times the distance between the conductors.

The End-Fed Zepp

A half wave antenna fed on the end with a λ/4 matching section of low loss balanced line is called an end-fed Zepp. This is often a convenient antenna if the station is close to a property boundary, making a center-fed antenna a difficult proposition. The matching section transforms the high impedance on the end, typically close to 2000 Ω, to a relatively low impedance that can be fed by a transmitter, or a narrow range tuner. This is the original Zeppelin antenna that was deployed beneath the aircraft's gondola. It is also popular as a VHF vertical called a "J-antenna."

Some have had better results by providing a λ/10 or so of wire as a counterpoise on the other side of the transmission line, particularly if using it on multiple bands. On bands at which the antenna is a multiple of λ/2, it provides mostly horizontal polarization with minimal common mode current on the transmission line. It can be made to work on bands at which it is not resonant, however, there will be a pronounced lack of balance on the transmission line.

While the λ/4 matching section is a key part of the original Zepp design, it is not necessary if a wide range antenna tuner is provided to perform the matching as shown in **Figure 13-11**.

The unbalanced line currents result in two effects.

- First, the line will radiate. This may not be a terrible thing, if the line is in the clear, but can cause problems if it is running near other equipment.
- Second, the line will bring common mode current into the station. This can usually be addressed through a common mode choke at the point at which a transition to coax is made. Make sure the coax run is short, since the mismatch will result in much higher losses in coax.

Mobile Antenna Systems

Antennas used in vehicles operate under the same principles as antennas in other applications, however, they often provide unique challenges. Antennas for MF and the lower portion of the HF region must be shortened considerably in order to be practical on most motor vehicles, large ships are a notable exception. In addition, unlike the usual antenna environment, these antennas are usually quite close to both the radio equipment as well as the vehicle body. With the exception of aircraft antennas, most mobile antennas are also quite close to the ground.

While none of the above differences by itself results in major problems, they all lead to certain limitations in

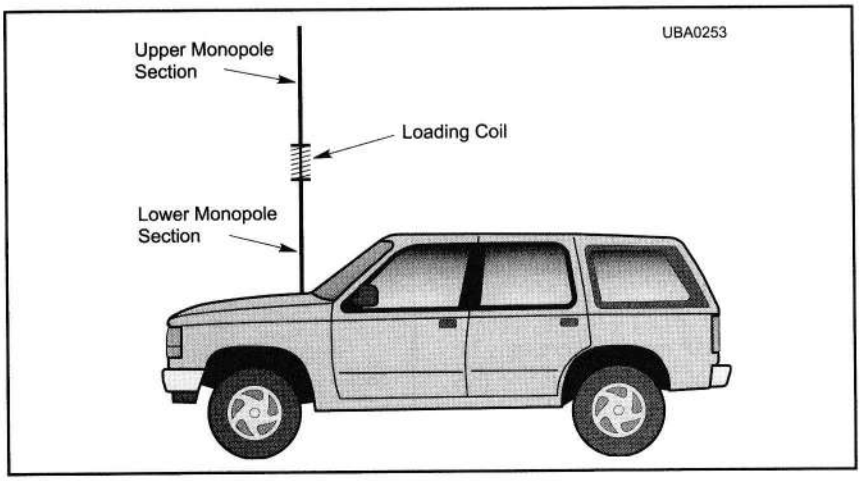

Figure 13-12 — Typical vehicle mounted monopole for the lower HF range. Because of practical limitations to do with garages, tunnels and trees, lengths tend to be 8 feet or less, from about ⅛ to 1⁄32 wavelength. The center loading shown improves efficiency.

antenna design and performance.

HF Mobile Antennas on Land Vehicles

The lower HF bands, especially 80, 40 and 20 meters, are popular with mobile operators. Depending on the sunspots and time of day, one or more can usually be counted on for reliable medium to long range communication. The usual antenna is a relatively short monopole antenna mounted on, and fed against, the vehicle body as shown in **Figure 13-12**.

As in the antenna in Figure 13-12, center loading is often used because it results in higher current in the lower section and thus less loss, although base loading, or even just using a solid whip with a tuner at the base are often encountered. A limitation of such an antenna is its SWR bandwidth, a typical case shown in **Figure 13-13** — much narrower than a full size antenna.

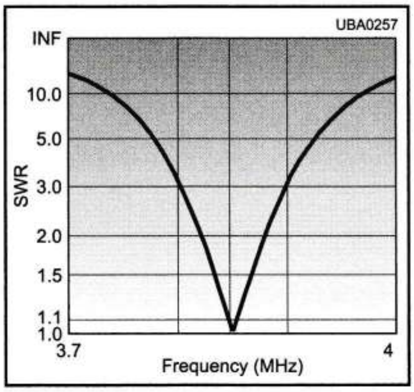

Figure 13-13 — Typical SWR plot of a mobile antenna of the type shown in Figure 13-12. The 2:1 SWR bandwidth is likely to range from 20 to 50 kHz, with antennas having less efficiency tending to have wider bandwidth.

This makes the shortened mobile antenna a good candidate for a wide range antenna tuner. The tuner can be used to allow full band coverage to be obtained. I have even used a tuner to allow a 40 meter mobile whip to be used on 80 meters, although it is not as efficient as the center loaded 80 meter arrangement.

HF Mobile Antennas on Other Vehicles

Those using HF on boats or aircraft do not generally share the same size restrictions as on land vehicles. Often either will use a convenient length of wire, for example the insulated backstay on a sailboat, or a wire from a short mast to a handy surface on an aircraft, to form an antenna similar to the random wire described earlier. This is almost always accompanied by a remote, usually automatic, antenna tuner very near one end or the other of the wire. This system can work very well if fed against the aircraft body or a metal hull. On my fiberglass sailboat, I feed it against the ship ground system.

Review Questions

13-1 Under what circumstances can an antenna tuner be of benefit in feeding a simple coax fed $\lambda/2$ dipole?

13-2 Why is it important to use low-loss transmission line such as window or ladder line if using a dipole on multiple bands?

13-3 What are the possible consequences of common mode currents on a transmission line?

Chapter 14

A Survey of Available Tuners

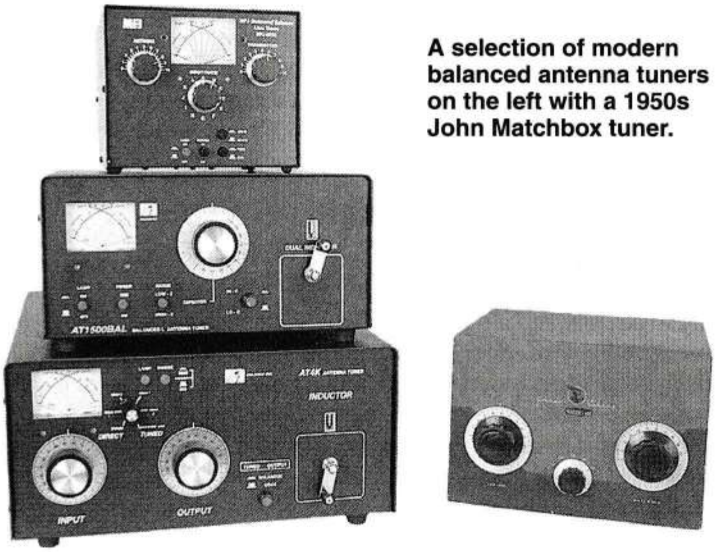

A selection of modern balanced antenna tuners on the left with a 1950s John Matchbox tuner.

Contents

Key Tuner Parameters 14-2
Product Review Testing 14-4
Selected QST antenn Tuner Product Review Articles
QST Reviews Five High Power Antenna Tuners 14-5
Automatic Antenna Tuners — A Sample of the Field 14-12
Medium to High Power Auto Antenna Tuners — The Evolution Continues 14-19
A New Generation of Balanced Antenna Tuners 14-26
Review Questions 14-36

Key Tuner Paramaters

Virtually every configuration of antenna tuner, capable of operation at every legal power level, is available commercially from one or more vendors. This can make selecting a commercial tuner a bit of a challenge, just as it is much tougher to select a meal from a restaurant with many good choices, compared to just a few.

Fortunately, the choices can be quickly reduced if one can make a few decisions on key tuner parameters. I will try to organize them in a way that will help make the selection easier. The following are the major decisions to be made, in this person's opinion:

What Frequencies Do You Want the Tuner to Cover?

When talking about antenna tuners, we almost always are talking about the HF range, although there are also some tuners for VHF bands. This means that we want to cover the amateur bands from 80 through 10 meters — that's a *given*. Most current amateur transceivers also cover 160 meters. Is that of interest? If so, because it often doubles the size of many tuner components, it will likely add significantly to the cost and size of a tuner, all other things being equal. The 160 meter band is gaining in popularity since it provides interesting operational conditions at the bottom of the sunspot cycle, as I write this. If you have room for effective 160 meter antennas, having a tuner that supports operation here may be a good choice.

On the other end of the spectrum is 6 meters. More and more current amateur "HF" transceiver cross into operation on our lowest VHF band. This band is also quite popular now since there are long distance modes that don't depend on sunspot activity. If you have, or expect to have, a dedicated 6 meter antenna, the chances are that it will be suitable to be fed directly with 50 Ω coax and a tuner will not be needed. On the other hand, if you will need to use your all band tuner fed antenna, such as the center-fed Zepp, on this band, you will want to include that capability in your tuner.

The other frequency decisions may be more subtle. Operation on 10 meters is sometimes a problem with larger tuners that have trouble ending up with sufficiently low minimum inductance and capacitance to provide the same impedance coverage on 10 meters as on the other bands — check the specifications to get an idea. In a similar vein, some tuners don't have quite enough maximum inductance or capacitance to provide the tuning capability to cover 80 (or 160) meters with the same range as on the other bands.

How Much Power Do You Expect to Run?

For any given tuner design, there is no factor that influences tuner cost and size more than power level. This should not be a surprise since higher power means higher voltage and current stress on the components within. The obvious answer might be to pick a tuner with a power rating equal to the most power that your transmitter can put out — but be careful, there are four issues hidden here:

- The power rating is based on the manufacturer's SWR specification (if they did a proper design job). If it says a tuner is good for an SWR of 10:1, typical for a wide range tuner, and you tune an antenna with a 20:1 SWR, you may be able to get a match, but either the voltage or the current in the tuner can be higher by a factor of $\sqrt{(20/10)}$, or 1.41 times. That may explain the arcing, or the burning smell, you observe on some frequencies! The answer may be to go somewhat higher in power rating than you expect to actually run.
- So you now have that shiny new 100 W transceiver — you love it to death and can't imagine that you will ever need more power. You buy a 100 W antenna tuner (or better a 200 W tuner, as discussed above) and happily make many contacts with your stealthy antenna. Now you decide to try for DXCC and notice that the DXpedition in Outer Slovenia can almost hear you, but can't quite copy your call letters. A bigger antenna is out of the question — the only solution is a high power amplifier. Oops, now you have an almost new 100 W tuner that you can no longer use.
- In addition to taking up more space, and more lawnmowing money, high power tuners often, because of the physical size of the components, have trouble providing the tuning range some antennas need on the higher bands. Some switch out portions of the circuit on higher bands to avoid the problem, but some don't. Keep your eye on that ball, if looking at higher power tuners.
- On the plus side, higher power tuners usually have inductors that are wound with heavier wire. This reduces losses, although it may not be noticeable in the test results.

Do You Want a Balanced or Unbalanced Tuner?

If you expect to be feeding single wire or coax-fed antennas, an unbalanced antenna tuner will be fine. If you want to feed an antenna such as a center fed Zepp with balanced line, you will need to have some provision to feed a balanced load. This can be handled in one of three ways:

- A fully balanced tuner is ideally suited to matching to balanced loads within its matching range. These are available in multiple power ranges and can work quite well. A disadvantage of these is that many do not include all the switching options of many unbalanced tuners — not a problem if you expect to use a single antenna.
- Many inherently unbalanced

tuners include an output connection through an internal balun. These work best with balanced loads within about 4:1 SWR of the balun design impedance. Most use 4:1 baluns designed for a 200 Ω load. That means they will work reasonably well with resistive loads from 50 to 800 Ω, and with complex loads of a similar SWR. They will also work outside that range, although their response can be a bit unpredictable. If they don't get hot, they are probably putting the power where you want it.

- The same kind of balun can be added outside of an unbalanced tuner if you decide to move from coax to a balanced feeder system. The same concerns as noted above apply. However there are some advantages such as that you can custom design the balun for the impedance that each antenna has, and you can have the balun some distance from the station equipment — but watch out for mismatched coax losses if the distance is more than a few feet.

Do Your Want a Remote or Local Tuner?

An antenna tuner that is located next to the radio in the station area offers a lot of plusses in terms of control and monitoring. Some antennas, however, really want the impedance matching performed at the antenna. A big consideration here is that if the tuner is located with the antenna, it usually can only operate with that single antenna, although remote switching is possible. If the antenna tuner is near the radio, switching between antennas systems is usually much easier — sometimes provided within the tuner, sometimes outside — but always within reach.

Do you Want a Manual or an Automatic Tuner?

A manual tuner requires adjusting two or three controls (and often switches) every time either frequency is changed, typically by around 50 kHz, depending on what band is selected or what type antenna is used. The adjustments can be tedious, although recording them makes returning to a proper tune under previously encountered conditions fairly easy.

An automatic tuner, on the other hand, just requires a push of the button to obtain a tune. Many have some sort of memory arrangement so they can quickly return to previous settings. The tuning usually takes a few seconds the first time, a fraction thereafter. While it is possible to imagine a remote manual tuner, to my knowledge, all remote tuners are of the automatic flavor. So if you want remote, you get automatic.

The idea of an automatic tuner sounds pretty good, and it is. There are just a few items to be aware of, as noted below:

- In general, an auto tuner will make a match within its specified SWR threshold, often 1.5:1, but check and compare with your radio's requirement. With your manual tuner, you can usually find a 1:1 setting. If the radio will put out its full power into the load, this is not worth worrying about.
- Watch out for tuners that initiate a tune on their own. Some can be just *too* automatic to suit me. You tune your radio 25 kHz off your previous frequency to capture a neat DX station, start to call, and hear your antenna tuner relays clacking. Chances are the up and down tuning will result in the station missing some part of your call sign.

Of more concern, if there is no circuitry to prevent it, many automatic tuners that can handle 100 W, for example, want tuning to occur at a lower power level to avoid burning relay contacts. If it starts up while you are operating, you can easily hit it at full power before you know what's happening. In the mobile environment, this can also happen as you pass that 18 wheeler with a giant piece of aluminum detuning your antenna (it can also serve as a reflector, if you get it in the right place!).

- I once thought that the small components in the usual compact automatic antenna tuner would result in more losses compared to the larger components, especially inductors, in the usual manual tuners. To my surprise, following the test results presented below, this does not seem to be the case, although it might be buried in the 5% minimum loss, below which we don't report due to measurement uncertainty. If it is a loss of that level (0.2 dB), in my opinion, it should not be part of the decision process.

Product Review Testing

The American Radio Relay League (ARRL), the US Amateur Radio organization publishes a monthly magazine, *QST*, for its members. In addition to its many technical and general interest features, each month's issue includes detailed reviews of products of interest to its readers. These Product Reviews generally include laboratory evaluation of equipment operational parameters using professional grade, independently calibrated test instruments. Over the years, we have had many such reviews of antenna tuners and I have reproduced the results from some here.

While many of the reviews include currently available equipment, some models covered are no longer available as new equipment. Even so, it may be of benefit in two ways:

- There is an active used equipment market in Amateur Radio equipment. This may provide a way to knowledgeably purchase an antenna tuner at a lower cost.
- In some cases, manufacturers have made relatively small changes to equipment, so data on previous models may be directly applicable.

ARRL Lab Antenna Tuner Testing Methodology

The following method was used in the ARRL Lab to measure the loss of the tuners presented below. Two test fixtures were built which would hold combinations of high power 50 Ω "non-inductive" carbon resistors (one fixture for parallel combinations and one for series), one shown in **Figure 14-1**. Even with the non-inductive resistors, some net fixture inductance was apparent at some frequencies, so a variable capacitor was used to compensate. The fixture accuracy was measured using the Lab's vector impedance meter.

Each fixture used an input connection (for the tuner) and an output connection. The output connection went to a 50 Ω input power attenuator, which took the place of one of the resistors in the load (for the series loads, it was always the one on the ground side of the network). The output of the power attenuator was connected to a high accuracy laboratory wattmeter, and the actual attenuation was measured for each frequency. The tuners were matched at low power, then 100 W of RF was applied at the input, with the output being measured by the Lab's wattmeter.

Each tuner was connected to the load fixture at each resistive impedance level for testing in the ARRL Lab. A 100 W RF source fed the input of the tuner. The resistive load was connected to the antenna output. The load was connected to a 50 Ω. power attenuator, which took the place of the final resistors. This is a close-up of the parallel load test fixture. The variable capacitor was used to compensate for fixture inductance encountered at some frequencies. While complex impedance loads could have been tested as well, the additional data points would have made testing impractical. We believe that the results at a particular SWR should be representative of other loads with the same SWR.

Note that the reviews were conducted over a period of a few years and the data taken, as well as the presentation, varied slightly from review to review. Please make sure that you are comparing equivalent data.

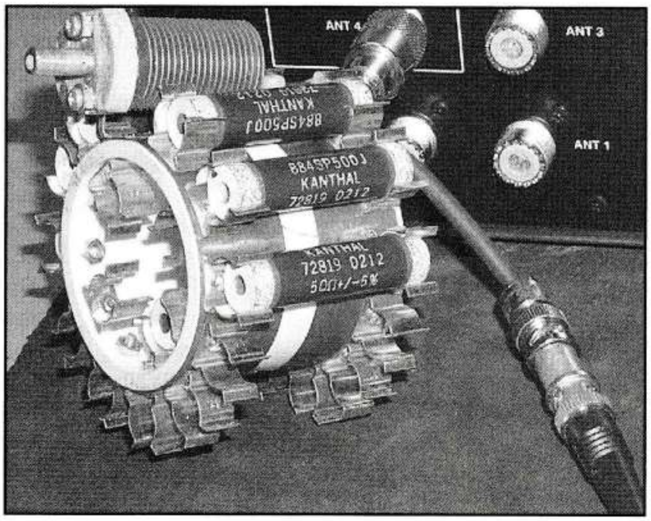

Figure 14-1 — Close up of test figure used to test antenna tuner efficiency into loads from 50 Ω down using paralleled 50 Ω noninductive loads.

QST Reviews Five High Power Antenna Tuners

Reviewed by Jim Parise, W1UK
ARRL Technical Advisor
QST *February, 2003*

One piece of gear that finds its way into most everyone's shack at one time or another is an HF antenna tuner. First of all, the name antenna tuner is something of a misnomer. It does not tune the antenna at all, but acts as an impedance transformer that provides your transmitting equipment with the proper load, usually 50 Ω.

Who needs an antenna tuner? Well, anyone who has the need to match an antenna with an impedance outside the range of their transmitter or amplifier's output circuits. Modern transceivers often have built-in tuners that are capable of matching SWR mismatches up to 3:1. Beyond that they need help, and if you use an amplifier you most surely will need one. Many hams find themselves with limited antenna choices and the desire to operate on frequencies other than what they were designed for, or use non-resonant multiband antennas that require a tuner. The five HF tuners we tested are all in the kilowatt class, meaning the manufacturer rates their power handling capability at 1 kW output or more.

With the wide range of frequencies in the HF spectrum and the huge diversity of antenna types in use, tuners are expected to perform under an incredible number of possible combinations. Some are more efficient at it than others. A measure of a tuner's ability to transform impedances efficiently is energy loss. Under extreme conditions a tuner can get quite hot or arc over at power levels well under the manufacturers rating. Heat in a tuner is a product of loss. RF energy being dissipated as heat is lost power that will not find its way to your antenna and onto the air.

Tuner losses generally get higher as the impedance of the load decreases. If a ham were running 1500 W into a tuner that was 50% efficient, 750 W would be dissipated in the tuner. Most of the loss in a tuner occurs in the coil, and no coil can withstand 750 W of power. A high-power tuner could probably be safely used at 50% efficiency and 100 W, but hams should be careful with high SWR and high power, or a tuner failure is a real possibility. Each tuner was tested into resistive impedances that ranged from 6.25 to 400 Ω, and their respective percentage of loss and 1.5 SWR bandwidth measured. The tuners were also used in everyday communications on all bands at power levels up to 1 kW, matching a G5RV fed with balanced ladder line and a 160 meter inverted L fed with coax.

The characteristicvs of the five uners tested are described in **Table 14-1**. Each tuner is described separately below.

Table 14-1
Comparison of Five Antenna Tuners

	Ameritron ATR-30	*MFJ MFJ-986*	*Palstar AT1500 CV*	*Ten-Tec 238A*	*Vectronics HFT-1500*
Circuit Type	T-network	T-network	T-network	L-network	T-network
SWR/wattmeter	Cross-needle	Cross-needle	Cross-needle	Single-needle	Cross-needle
Balun type	4:1 current	4:1 current	4:1 voltage	4:1voltage	4:1 voltage
Manufacturer's claimed PEP rating	3000 W	3000 W	1500 W	2000 W	2000 W
Manufacturer's claimed Matching range	35-500	35-500	20-1500	5-3000	Not specified
Physical dimensions (HWD) in inches	5.25×13×14.4	4.1×11×15.2	4.5×12.5×12	5.5×13×11	5.5×12.5×12

AMERITRON ATR-30

The ATR-30 (**Figures 14-2, 3, 4**) is Ameritron's legal limit T-network antenna tuner offering. It is housed in a plain, all black aluminum enclosure with a scratch resistant coated front panel. A look under the cover reveals two large variable capacitors and an air core edge wound silver plated roller inductor. Like most of the other tuners reviewed here, the capacitors are adjusted with vernier reduction drives, and although they tune smoothly, they require a lot of force to turn. The roller inductor is adjusted with a plastic lever type knob and is quite stiff to crank. The roller itself is a pinch roller, and the physical resistance in turning the crank may be offset by lessoned contact resistance.

The balun is constructed with three large cores and wound with wire covered with Teflon tubing. The cross-needle meter displays both average and peak power in switchable 300 W or 3 kW ranges. The wattmeter requires dc to function. It is illuminated by either a 12 V barrel connector on the back panel or a 9 V battery accessible through a trap door on the bottom of the unit. With 12 V, both the meter and inductor turns counter are illuminated. When a 9 V battery is used, the meter will function, but the panel lamps will not light.

Finding a match on 80 meters on the G5RV required quite a bit of time finding the right combination of capacitor tuning and inductance, and the stiff controls didn't make it much fun. Finding the sweet spot on the higher bands was much easier. There are three coax outputs on the ATR-30, including two that may be switched direct or through the T-network and a third direct only. Single or balanced feed lines connect to large ceramic binding posts with wing nuts. Test results are summarized in **Table 14-2**.

Manufacturer: Ameritron, 116 Willow Rd, Starkville, MS 39759; tel 662-323-8211; fax 662-323-6551; **www. ameritron.com**. 2003 price: $599.95.

Table 14-2
Ameritron ATR-30 Loss and Bandwidth Test Results

SWR	*Load (Ω)*		*160 m*	*80 m*	*40 m*	*20 m*	*10 m*
8:1	6.25	Power Loss %	20	12	<10	<10	<10
		1.5 SWR BW	1	3	>5	4	>5
4.1	12.5	Power Loss %	15	<10	<10	<10	<10
		1.5 SWR BW	3	4	>5	>5	>5
2:1	25	Power Loss %	10	<10	<10	<10	<10
		1.5 SWR BW	4	>5	>5	>5	>5
1:1	50	Power Loss %	<10	<10	<10	<10	<10
		1.5 SWR BW	>5	>5	>5	>5	>5
2:1	100	Power Loss %	<10	<10	<10	<10	<10
		1.5 SWR BW	>5	>5	>5	>5	>5
4:1	200	Power Loss %	<10	<10	<10	<10	<10
		1.5 SWR BW	>5	>5	>5	>5	5
8:1	400	Power Loss %	<10	<10	<10	<10	<10
		1.5 SWR BW	>5	>5	>5	>5	3

Notes
Power losses are expressed as a percentage. A 21% loss of power is 1 dB.
The 1.5-SWR Bandwidth (SWR BW) represents the bandwidth over which an SWR of 1.5:1 or less was maintained as a percentage of the measurement frequencies (1.8, 3.5, 7.2, 14.2 and 29.7 MHz).

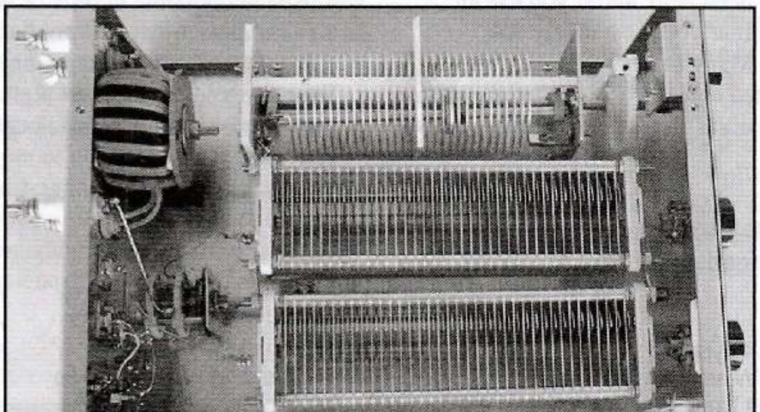

Figure 14-3 — Interior circuitry of the Ameritron ATR-30.

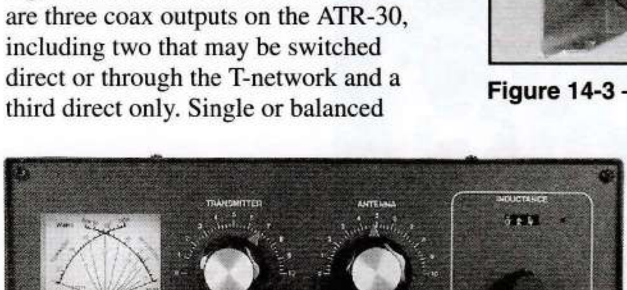

Figure 14-2 — Front panel of the Ameritron ATR-30.

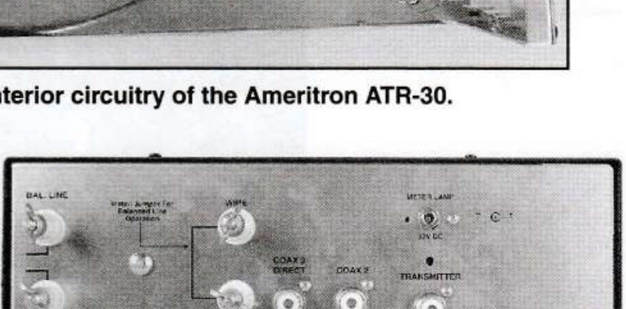

Figure 14-4 — Rear panel connections on the Ameritron ATR-30.

MFJ-986

The '986 is one of more than 20 antenna tuner products offered by the prolific MFJ (in 2003). It is rated for 1500 W PEP within the 35-500 Ω impedance range, positioning it alongside the '989C as one of their high power units. Unlike the traditional T-network with a pair of variable capacitors, the '986 (**Figures 14-5, 6, 7**)makes use of a single variable capacitor and a wire wound air core roller inductor with a three digit turns counter. The tuner uses what MFJ describes as a "differential" capacitor. It has two discrete sets of fixed plates and one variable set, with a single capacitor control, forming a T-network with the inductor.

The cabinet is narrow and extends nearly 18 inches deep. The cross-needle meter has a high (3000 W) and low (300 W) setting switched by a front panel button, and it can display average and peak power as well as SWR. Providing 12 V to the connector on the back panel brightly illuminates the meter. Both the capacitor and roller inductor are directly driven. The inductance knob has a finger depression, which are often found on a transceiver's tuning knob. It is somewhat awkward to turn due to binding. The tuner had a tendency to creep around the operating desk when the inductor is rapidly turned. A check inside revealed several loose screws holding the inductor to the chassis. Tightening these did improve the smoothness somewhat.

The '986 provided a match to both antennas quickly, but arced over on 80 meters with the inverted L at 900 W. While this tuner does provide ease of use, one should give careful consideration to the loss figures in the tables. During operation the meter developed an intermittent problem with both needles going off scale even with very low power applied and a low SWR. Movement of the SWR bridge circuit board on the inside back panel seemed to correct the problem.

Two coax outputs are provided that can be switched to bypass the tuning circuit, as well as a third output for a dummy load. Balanced and single-wire feed lines connect to large ceramic feed through posts with wing nuts. A 4:1 two-core current balun is provided at the output. Test results are summarized in **Table 14-3**.

Manufacturer: MFJ Enterprises, Inc, PO Box 494, Mississippi State, MS 39762; tel 800-647-1800; fax 662-323-6551; **www.mfjenterprises.com**. 2003 price: $329.95.

Table 14-3
MFJ-986 Loss and Bandwidth Test Results

SWR	*Load (Ω)*		*160 m*	*80 m*	*40 m*	*20 m*	*10 m*
8:1	6.25	Power Loss %	47	31	21	16	13
		1.5 SWR BW	1	1	2	4	>5
4.1	12.5	PowerLoss %	33	22	14	12	11
		1.5 SWR BW	1	1	4	5	>5
2:1	25	PowerLoss %	25	20	10	<10	10
		1.5 SWR BW	1	2	4	>5	>5
1:1	50	PowerLoss %	22	12	<10	<10	<10
		1.5 SWR BW	2	3	>5	>5	>5
2:1	100	PowerLoss %	15	10	<10	<10	19
		1.5 SWR BW	3	5	>5	>5	>5
4:1	200	PowerLoss %	11	<10	<10	<10	<10
		1.5 SWR BW	3	>5	>5	>5	>5
8:1	400	PowerLoss %	10	<10	<10	11	16
		1.5 SWR BW	3	>5	>5	>5	5

Notes

Power losses are expressed as a percentage. A 21% loss of power is 1 dB.

The 1.5-SWR Bandwidth (SWR BW) represents the bandwidth over which an SWR of 1.5:1 or less was maintained as a percentage of the measurement frequencies (1.8, 3.5, 7.2, 14.2 and 29.7 MHz).

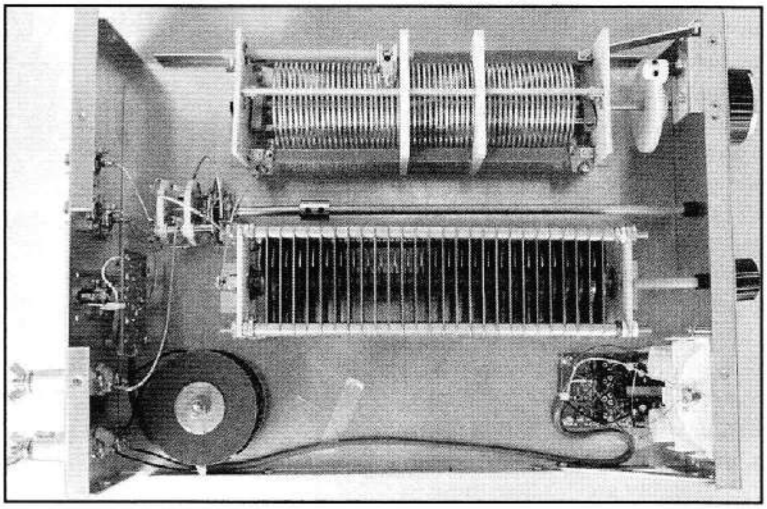

Figure 14-6 — Interior circuitry of the MFJ-986.

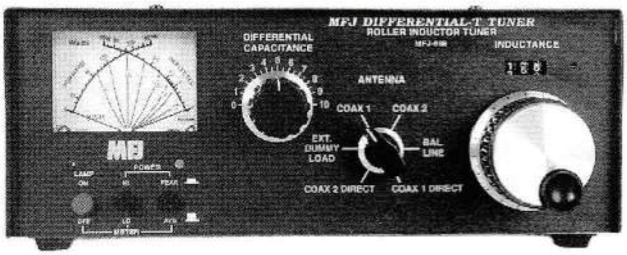
MFJ DIFFERENTIAL-T TUNER
ROLLER INDUCTOR TUNER
DIFFERENTIAL CAPACITANCE
0
1
2
3
4
5
6
7
8
9
10
INDUCTANCE
ANTENNA
COAX 1
COAX 2
EXT. DUMMY LOAD
BAL LINE
COAX 2 DIRECT
COAX 1 DIRECT
POWER
LAMP
ON
OFF

Figure 14-5 — Front panel of the MFJ-986.

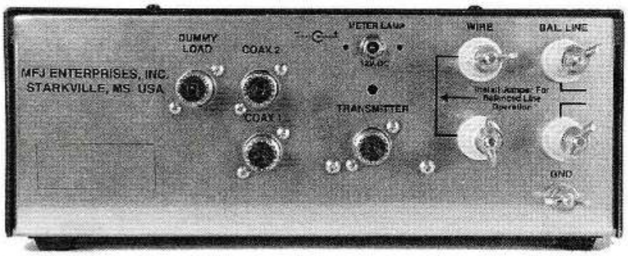
DUMMY LOAD
COAX 2
COAX 1
MFJ ENTERPRISES, INC.
STARKVILLE, MS USA
METER LAMP
12V DC
TRANSMITTER
WIRE
BAL LINE
Install Jumper For Balanced Line Operation
GND

Figure 14-7 — Rear panel connections on the MFJ-986.

PALSTAR AT1500CV

The Palstar AT1500CV (**Figures 14-8, 9, 10**) is a T-network tuner solidly constructed in a compact aluminum enclosure. Inside, the metering and SWR bridge circuitry are encased in their own aluminum boxes. The moderately sized variable capacitors are mounted side by side and feature large calibrated dials and silky smooth vernier tuning that allows for precise adjustment. Palstar indicates that newer units incorporate a roller bearing assembly which makes inductor adjustment smoother. The roller inductor is quite large. It is an air core silver plated edge wound coil with heavy-duty ceramic forms. Control of the inductor is via a lever handle and mechanical turns counter. It is not as smooth as some of the other tuners and tends to lurch as it is rotated during fine adjustments. The cross-needle meter displays SWR and average forward and reflected power in 300 and 3000 W ranges, with no option for peak power metering. It is illuminated with 12 V from an included wall adapter.

Achieving a 1:1 SWR on the G5RV was possible on all bands from 80 through 10 meters at its rated power maximum of 1000 W single tone. Similar results were noted on the inverted L. Although the tuning chart provided in the manual does not include settings for 30 meters, the tuner easily matched both antennas on that band. This tuner has a solid feel to it and doesn't creep around while making adjustments.

In laboratory tests, the AT1500CV had more difficulties on 160 meters than on other bands. A 1:1 SWR was only obtainable with the 25 and 50 Ω loads, the power losses at 160 meters were measurably greater than those on other bands at the same loads. The variable capacitors are 240 pF maximum, which explains the 160 meter performance. The tradeoff results in lighter weight.

Antenna connections on the back panel include three coax inputs. Two are available direct or through the tuning network and one is bypassed

Table 14-4
Palstar AT1500CV Loss and Bandwidth Test Results

SWR	*Load (Ω)*		*160 m*	*80 m*	*40 m*	*20 m*	*10 m*
8:1	6.25	Power Loss %	No Match	25	16	12	12
		1.5 SWR BW		1	2	>5	3
4.1	12.5	Power Loss %	No Match	16	13	<10	11
		1.5 SWR BW		2	4	>5	>5
2:1	25	Power Loss %	21	13	<10	<10	<10
		1.5 SWR BW	2	3	>5	>5	>5
1:1	50	Power Loss %	13	<10	<10	<10	<10
		1.5 SWR BW	3	>5	>5	>5	>5
2:1	100	Power Loss %	No Match	<10	<10	<10	13
		1.5 SWR BW		>5	>5	>5	>5
4:1	200	Power Loss %	No Match	<10	<10	<10	12
		1.5 SWR BW		>5	>5	>5	>5
8:1	400	Power Loss %	No Match	<10	<10	<10	<10
		1.5 SWR BW		>5	>5	>5	5

Notes

"No Match" means that a 1:1 SWR could not be obtained.
Power losses are expressed as a percentage. A 21% loss of power is 1 dB.
The 1.5 SWR Bandwidth (SWR BW) represents the bandwidth over which an SWR of 1.5:1 or less was maintained as a

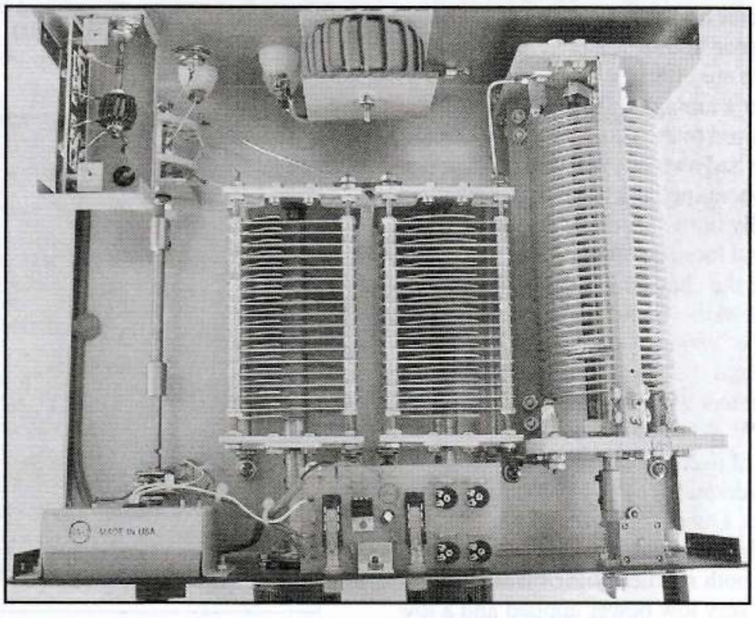

Figure 14-10 — Interior circuitry of the Palstar AT1500CV.

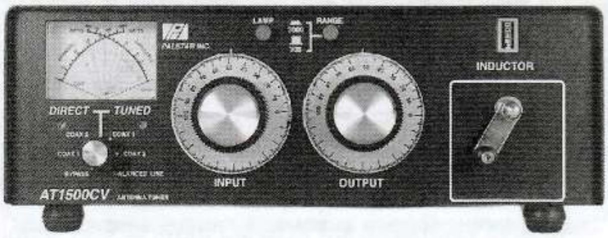

Figure 14-8 — Front panel of the Palstar AT1500CV.

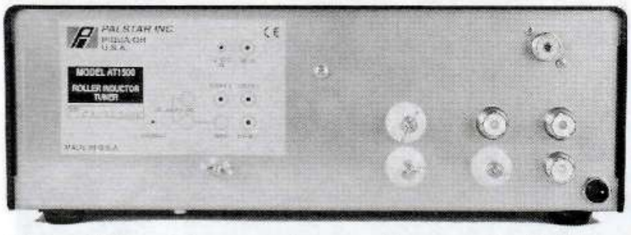

Figure 14-9 — Rear panel connections on the Palstar AT1500CV.

straight through. The bypass output can also be used with an optional 4:1 balun available for $39.95 [2003]. Balanced or single wire feed lines are attached to Delrin terminal posts with wing nuts.

Longer threaded material on the terminal posts would be a welcome improvement and make attachment of larger gauge wire much easier. In one isolated circumstance during lab testing, our engineer felt an RF bite on 10 meters through the metal portion of the inductor crank. This did not seem to be a pervasive problem, and Palstar indicates that the grounding on the crank has also recently been improved. Test results are summarized in Table 14-4.

Manufacturer: Palstar, Inc, 9676 N Looney Rd, PO Box 1136, Piqua, OH 45356; tel 937-773-6255; fax 937-773-8003; **www.palstarinc.com.** 2003 price: $429.95, optional 4:1 balun $39.95.

TEN-TEC 238A

The Ten-Tec 238A (**Figures 14-11**, **12** and **13**) is the only tuner in our roundup to utilize an L-network. Using an innovative switching arrangement, the '238A actually provides five different circuit configurations to maximize efficiency. A look at the low loss figures in the table confirms this. The well-constructed tuner makes use of a single variable capacitor and a smooth turning wire-wound roller inductor to match impedances up to 3000 Ω. Additional capacitance can be switched into the circuit by means of a front panel switch. Five settings each are available for high and low impedances, as well as a bypass choice. A ceramic feed through post on the back panel provides a connection to add an additional 1000 pF capacitor (included with the tuner) to the circuit for matching 160 meter antennas. The small dual use meter can be switched to display SWR or RF power in either a 2000 or 200 W range. The meter lamp is powered by 12 V on the back panel.

The '238A handily matched all bands on both test antennas at full legal limit power. On 10 meters using the G5RV, the capacitor and inductor controls were a bit sensitive to small adjustments. One minor complaint: Measuring forward power in the high power range on the meter caused the indicator to slap the right extreme position during CW keying.

In laboratory tests, the '238A did not match on 160 meters into a 6.25 Ω until 2700 pF of external capacitance was added. Ten-Tec indicates they will provide additional capacitors to purchasers at no extra charge. Also, the power losses on 10 meters were considerably greater than those on other bands. However, the '238A performed very admirably on most other bands, quite often exhibiting the least losses on given band and load combinations.

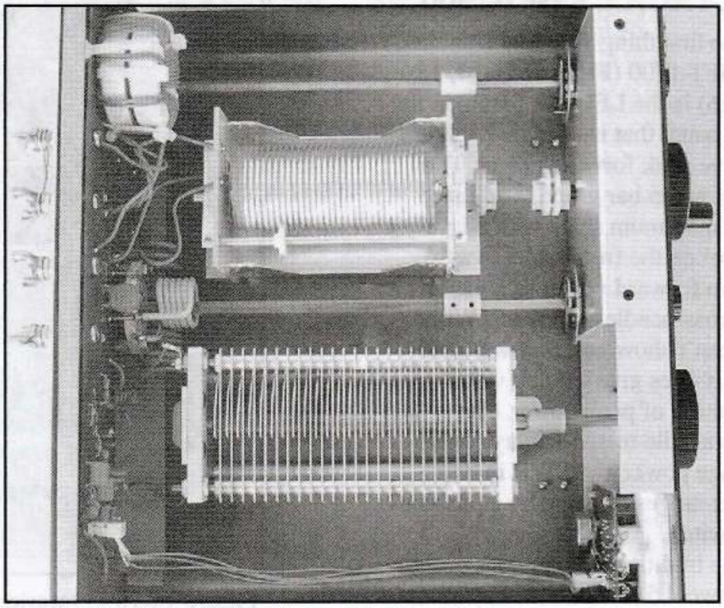

Figure 14-13 — Interior circuitry of the Ten-Tec 238A.

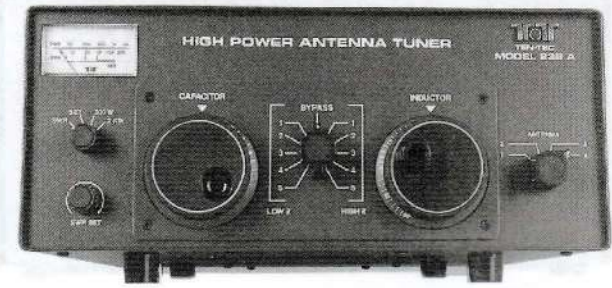

Figure 14-11 — Front panel of the Ten-Tec 238A.

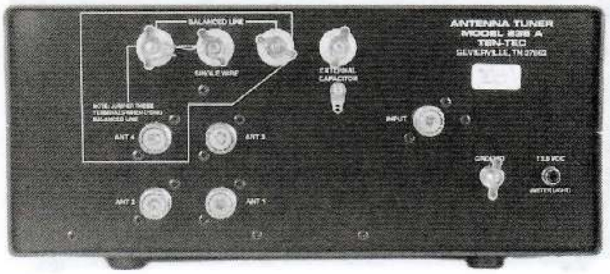

Figure 14-12 — Rear panel connections on the Ten-Tec 238A.

The four-position antenna switch on the front panel allows selection of four coax outputs or a balanced/random wire. The balanced output and one coax connector share position four. A two-core balun is provided for matching the balanced output. The '238A is the only one of the five tuners reviewed here without an air-core inductor. The inductor here has a linen phenolic core.

Since we purchased our review unit, Ten-Tec has produced a substantially identical antenna tuner, the '238B. The Tennessee manufacturer is now selling only the '238B model. Test results are summarized in **Table 14-5**.

Manufacturer: Ten-Tec, 1185 Dolly Parton Pky, Sevierville, TN 37862; tel 865-453-7172; fax 865-428-4483; **www.tentec.com**. 2003 price: $475.00 ('238B).

Table 14-5
Ten-Tec 238A Loss and Bandwidth Test Results

SWR	*Load (Ω)*		*160 m*	*80 m*	*40 m*	*20 m*	*10 m*
8:1	6.25	Power Loss %	No Match	<10	<10	<10	28
		1.5 SWR BW		>5	>5	>5	2
4.1	12.5	Power Loss %	<10	10	<10	<10	22
		1.5 SWR BW	>5	>5	>5	>5	2
2:1	25	Power Loss %	<10	10	<10	<10	17
		1.5 SWR BW	>5	>5	>5	>5	3
1:1	50	Power Loss %	<10	<10	<10	<10	<10
		1.5 SWR BW	>5	>5	>5	>5	>5
2:1	100	Power Loss %	<10	<10	<10	<10	20
		1.5 SWR BW	>5	>5	>5	>5	4
4:1	200	Power Loss %	<10	<10	<10	<10	20
		1.5 SWR BW	>5	>5	>5	>5	4
8:1	400	Power Loss %	<10	<10	<10	10	13
		1.5 SWR BW	>5	>5	>5	>5	1

Notes

"No Match" means that a 1:1 SWR could not be obtained, even with the included 1000 pF external capacitor attached. A match on 160 m at 6.25 was achieved with 2700 pF of external capacitance; the power loss was less than 10%, and the 1.5 SWR BW was greater than 5%.

Power losses are expressed as a percentage. A 21% loss of power is 1 dB. The 1.5-SWR Bandwidth (SWR BW) represents the bandwidth over which an SWR of 1.5:1 or less was maintained as a percentage of the measurement frequencies (1.8, 3.5, 7.2, 14.2 and 29.7 MHz).

VECTRONICS HFT-1500

The first thing you notice about the HFT-1500 (**Figures 14-14, 15** and **16**) is the LED bar graph on the front panel that is used to display relative peak forward power. The bright green bar graph is adjusted to its maximum scale with a level control on the front panel to a known forward power reading on the cross-needle wattmeter. While it doesn't show actual power readings, it does give a quick visual indication of peak power. The cross-needle meter displays only average power and SWR in two power ranges selected by a pushbutton switch.

The traditional T-network design uses two 4.5 kV variable capacitors adjusted with large comfortable

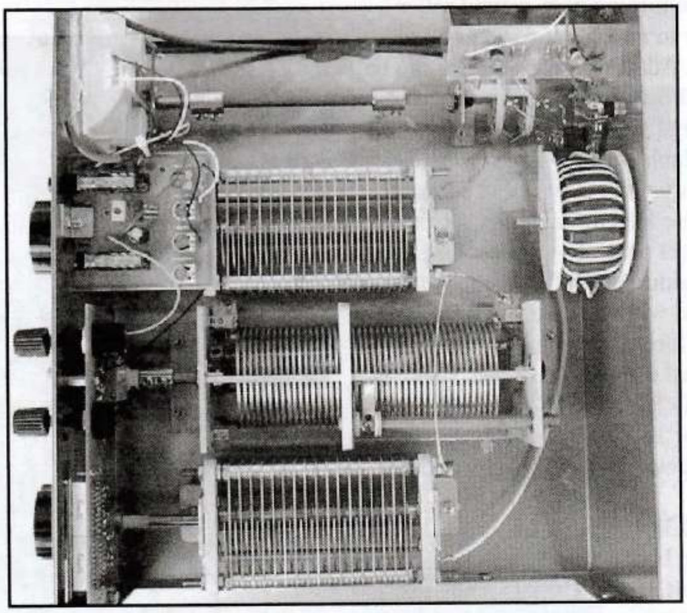

Figure 14-16 — Interior circuitry of the VECTRONICS HFT-1500.

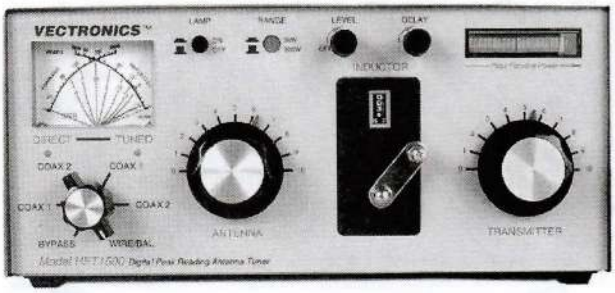
VECTRONICS
LAMP
RANGE
LEVEL
DELAY
INDUCTOR
DIRECT
TUNED
COAX 2
COAX 1
COAX 1
COAX 2
BYPASS
WIRE/BAL
ANTENNA
TRANSMITTER
Model HFT1500 Digital Peak Reading Antenna Tuner

Figure 14-14 — Front panel of the VECTRONICS HFT-1500.

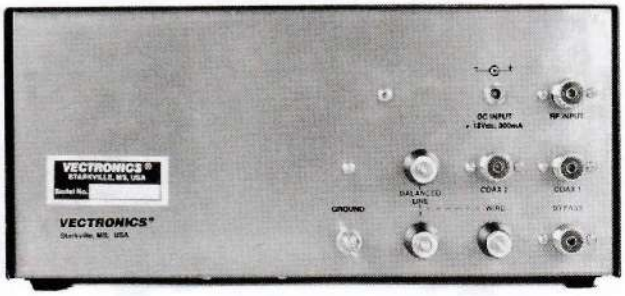
DC INPUT
+12Vdc 300mA
RF INPUT
VECTRONICS
STARKVILLE, MS, USA
Serial No.
VECTRONICS
Starkville, MS USA
BALANCED LINE
COAX 2
COAX 1
GROUND
WIRE
BYPASS

Figure 14-15 — Rear panel connections on the VECTRONICS HFT-1500.

Table 14-6
Vectronics HFT-1500 Loss and Bandwidth Test Results

SWR	*Load (Ω)*		*160 m*	*80 m*	*40 m*	*20 m*	*10 m*
8:1	6.25	Power Loss %	45	42	16	15	8
		1.5 SWR BW	1	<1	2	5	>5
4.1	12.5	Power Loss %	32	31	11	<10	<10
		1.5 SWR BW	>5	>5	3	>5	>5
2:1	25	Power Loss %	19	24	<10	<10	<10
		1.5 SWR BW	2	1	>5	>5	>5
1:1	50	Power Loss %	12	<10	<10	<10	<10
		1.5 SWR BW	3	>5	>5	>5	>5
2:1	100	Power Loss %	12	<10	<10	<10	<10
		1.5 SWR BW	3	>5	>5	>5	>5
4:1	200	Power Loss %	<10	<10	<10	<10	<10
		1.5 SWR BW	4	>5	>5	>5	>5
8:1	400	Power Loss %	<10	<10	<10	11	16
		1.5 SWR BW	4	>5	>5	>5	4

Notes

Power losses are expressed as a percentage. A 21% loss of power is 1 dB.

The 1.5-SWR Bandwidth (SWR BW) represents the bandwidth over which an SWR of 1.5:1 or less was maintained as a percentage of the measurement frequencies (1.8, 3.5, 7.2, 14.2 and 29.7 MHz).

knobs and very smooth vernier tuning. This tuner stays put on the desk while you utilize a lever handle and gear driven five digit mechanical turns counter to adjust the air wound roller inductor.

Tuning the 160-meter inverted L and the G5RV on all bands except 10 meters was easily accomplished, with no problems handling 1 kW. In the field, the best SWR that could be obtained on 10 was 2.33:1; this is possibly due to the additional "loading" when a human touches the all-metal inductor crank. The ARRL Lab adjusted the inductor with a wooden pencil in some cases to obtain a match.

The HFT-1500 provides two coax inputs and a third that completely bypasses the tuning network. The balanced and single wire inputs are Delrin terminal posts connected to a large single core 4:1 voltage balun. In some circumstances during lab testing, our engineer felt RF bites through the metal portion of the inductor crank. Later units feature a metal shaft bushing that provides a better ground. This was not experienced in field testing, but tuner adjustments were not made at high power. Test results are summarized in **Table 14-6**.

Manufacturer: Vectronics, 300 Industrial Park Rd, Starkville MS 39759; tel 662-323-5800; fax 662-323-6551; **www.vectronics.com**. 2003 price: $459.95.

Automatic Antenna Tuners — A Sample of the Field

Reviewed by Joel R. Hallas, W1ZR
QST *Technical Editor*
QST *May, 2004*

Automatic antenna tuners can provide a quick and easy method of matching multiband or nonresonant antennas quickly and effortlessly. Performance is on par with manual tuners and there is now a selection available providing a fit to most requirements. A few years back, we surveyed the antenna tuner market in these pages. All those tuners were manual ones.

A few automatic tuners were available at that time, but they were the exception rather than the rule. Since then, a number of manufacturers have focused their attention on automatic antenna tuners and it seems as if there's now one for every application.

For this review we selected representative tuners from a number of manufacturers (see **Figure 14-17**). There are two general configurations: those providing remote tuning, with the tuner intended to be near the antenna, and those intended to be collocated with the station equipment. For those manufacturers having models in each camp, we selected one of each.

As noted in the specifications, each tuner is different and they don't exactly line up head-to-head, so a direct comparison is not always appropriate. Still, for each we will provide the basic data you would need to decide if one of these is a good fit for your needs. If you like the looks of a particular manufacturer, but the tuner reviewed doesn't exactly meet your requirements, check their Web page or your dealer for other models at different power levels or with different features.

For this review we stuck to those at the 100 to 200 W level and tried for a representative self-contained and remote tuner from each manufacturer. For these tests, we looked at units designed to work with any radio. In addition to these, most HF radio manufacturers offer external auto tuners designed to operate exclusively with their radios. One manufacturer, LDG, also offers tuners with the capabilities to operate with specific radios' commands. Many HF radio manufacturers also offer internal auto-tuners as a part of their transceivers. We generally review these as a part of the radios. We elected to stick to general purpose models for this review.

Those selected for this test include the RT-11 and Z-100 from LDG, the MFJ-991 and from SGC the SG-237 and MAC-200.

Remote vs Collocated Autotuners

It's probably worth a paragraph or two to discuss the relative merits and appropriateness of the two configurations. If the antenna(s) to be matched terminate right at the station equipment, without a transmission line (or with a very short one), there is no particular advantage to having a remote tuner.

On the other hand, that setup often results in problems with "RF in the shack," or EMI within the house. I have one antenna setup of that form and whenever I use it, the RF resets the kitchen FM radio, causing it to lose track of time and memory. Most find it a better solution to have the antenna system stop somewhere outside the house and have a transmission line interconnect the radio and antenna.

With a remote tuner, the matching can be at or near the antenna, resulting in a well matched transmission line coming back to the radio. With the tuner at the radio, the matching takes place near the radio, and the SWR on the line out to the antenna will often be high. Note that in both cases the radio will be provided a good match and will happily put full power into the system. With the remote tuner, often more of the power will be radiated by the antenna and less will be dissipated as heat by the transmission line.

Some radios now come with automatic tuners as part of the radio. These can be functionally equivalent to the local tuner. The separate tuner does offer two advantages over the

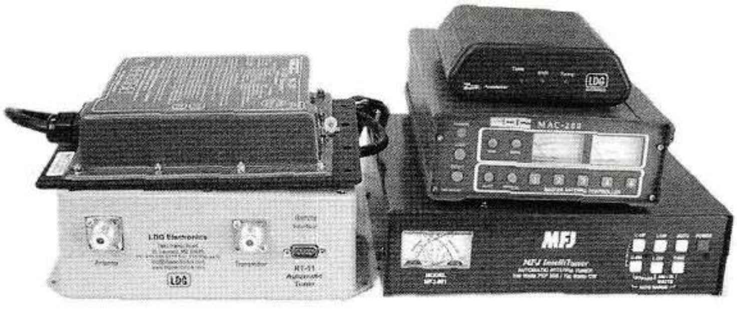

Figure 14-17 — Automatic tuners from LDG, MFJ and SGC — ready for test.

internal tuner — it can support multiple radios, and most can match a higher SWR than many of the internal tuners. The key factor in deciding between a local and a remote tuner then becomes how much of the transmitter power actually reaches the antenna system and how much will heat up the transmission line. This is a function of operating frequency, SWR and line length.

The place to start the analysis is with *The ARRL Antenna Book.*[1] Figure 23 in Chapter 24 provides the attenuation in 100 feet of *matched* transmission line for each of the common types. Table 14 in the same chapter provides the *additional* loss for the non-matched case as a function of SWR and matched loss. [The same data is in Chapter 7 of this book.] Using the two tables, you can determine the total loss of your transmission line.

If the loss is acceptable, you can tune at the transmitter end. If not, by tuning at the antenna end you can eliminate most of the additional loss due to mismatch.

An example might be helpful. A multiband antenna such as my G5RV might have a 5:1 SWR on parts of 10 meters. If we feed it with 150 feet of RG-58 cable we have a matched loss of about 3.5 dB and the 5:1 SWR adds about another 2 dB for a total of 5.5 dB loss or about 71% of our power heating up the line! By moving the tuner to the antenna end of the coax, we reduce that to 55%. Another option is to change the transmission line (my solution immediately after making that calculation!).

If we go with lower loss RG-8 instead, we have a matched loss of about 2 dB (1.3 dB per 100 ft) and the SWR adds about another 0.3 dB for a total of 2.3 dB, or a loss of about 41% of our power. With the tuner at the antenna, the total loss would be the 2 dB or 37%—not much of a difference. Note that I didn't consider the tuner losses here, under the assumption that they would be the same at either location. The numbers get much more grim as the SWR goes up, by the way.

Two notes of caution—first, if you measure the SWR at the shack end of the cable, it will appear better than it is, due to the cable loss; second, the matched loss from Figure 23 is the *best* matched loss for new and dry cable. The actual loss can only go up from there, especially if the cable came from some dark corner of your basement, or worse, from an unknown source at a flea market.

Note that all of these tuners are designed for unbalanced (coax) cable between the tuner and the radio and, with the exceptions noted, unbalanced operation on the antenna side (coax or single wire antennas) as well. In order to use these tuners with a balanced load, it is best to decouple the ground side of the tuner from the transmission line. A coax choke can be used, or a balun (balanced to unbalanced transformer) can be employed. If you are using a balun, it is best to keep the line impedance within a SWR of about 4:1 (with respect to the usual 200 Ω impedance of a 4:1 balun) to avoid heating, loss and possible balun damage. A subsequent planned review will look at the special class of unbalanced to balanced tuners that sidestep this issue in an elegant way and are once again appearing on the market.

A word or two about the data shown may be in order. Please note that we used the same precision resistive load set for each tuner. In most cases, this includes data outside the manufacturers' specification range. We measured at a single frequency in each band, and for 6 meters we just measured the case with a 1:1 load for all the tuners that claimed operation there.

---

[1]Notes appear on page 14-35.

LDG RT-11 and Z-100

A look at LDG's Web page, **www.ldgelectronics.com**, will show you their selection of tuners and accessories. In addition to the RT-11 remote control tuner and the Z-100 local tuner, tested (see **Figure 14-18**), they offer an autotuner rated at 1000 W. As this review was being prepared in 2004, they also announced a much more feature rich desktop unit to complement the very basic, but lower cost Z-100.

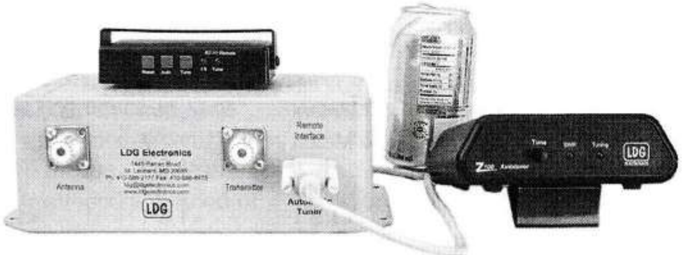

Figure 14-18 — LDG RT-11 (left) with control box and Z-100. A 12 ounce beverage container provides scale.

LDG RT-11

The RT-11 is a rugged, weatherproof tuner designed to be used inside or outdoors. It has coax connectors on input and output and is rated at 125 W from 1.8 to 30 and 50 W to 54 MHz. It is rated to match antennas with an SWR of 10:1 to 30 MHz and 3:1 from 50 to 54 MHz. This unit has an optional remote control designed to be used up to 100 feet from the tuner (15 foot cable supplied).

The remote control provides the ability to force a new tune, set for semi or full auto operation, keep the unit from tuning and allows bypassing of the tuner (straight through operation for matched antennas or out of band receive). In addition, LED indicators let you know the status and whether the SWR is below 1.5:1. One caution with the remote, the mounting bracket screws are not captive so you may want to use double-sided tape to secure it to a surface.

The RT-11 also has a jack to allow interfacing to radios designed to control proprietary tuners by emulating their commands. Special cables are offered to hook to various Alinco, ICOM, Kenwood and Yaesu radios for this purpose.

In my trials at home, I found that the RT-11 effortlessly tuned every antenna I had on all bands with a very low SWR, always well below 1.5:1. The tuning seemed to take about 5 seconds and hit the result the first time.

While automatic operation is possible without radio interface or remote control, I would have trouble imagining it and would suggest the remote for those without a compatible radio. For one thing, once tuned, it won't automatically restart the tuning process until the SWR reaches 3:1, a value that might cause many radios to "fold back" (reduce power) significantly.

The manual mentions that tuning should take place at the 25 W level, with full power applied only after tune is achieved, so that relay life will not be reduced. They also indicate that if a radio has the common power output fold back with high SWR it will automatically reduce power appropriately. I wouldn't count on that providing sufficient reduction, especially as the match is approached; I would instead manually reduce power while tuning.

LDG Z-100

The Z-100 is a self contained desk top unit with a 125 W rating (50 W on 6 meters). This unit is the first of a new series of tuners being released by LDG, with others providing additional features and controls. This is a basic unit corresponding to its low price, but none the less provides the needed tuning functions with straightforward "one button" operation. It includes memory of the settings for the last 200 matches made and tries those before it branches to its *unknown tune* algorithm. This is a clever idea that avoids the need to measure frequency. It is particularly useful for the case in which more than one antenna is used, since it really doesn't care — if it made the match before it will do it again, in a fraction of a second.

A feature of this unit is that it only draws power when actually tuning. Upon completion, the relays remain latched without power required until

Table 14-7
LDG RT-11 Remote Automatic Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input SWR range: to 10:1 (3:1 on 6 meters).	See below.
Output SWR range: not specified.	See below.
Frequency coverage: 1.8 to 54 MHz.	As specified.
Input power: 0.1 to 125 W, 50 W on 6 meters.	Tested at 50 W on HF, 20 W on 6 meters.
Power requirements: 11 to 14 V dc, 7 to 250 mA.	Idle 0 mA, max 140 mA at 13.8 V dc.

Measured power loss into resistive loads (%) / Input SWR at match:

SWR	*Load (Ω)*	*160 Meters*	*80 Meters*	*40 Meters*	*20 Meters*	*10 Meters*	*6 Meters*
16:1	3.125	No Match	No Match	27 <1.5:1	22 2.6:1	<10 <1.5:1	
8:1	6.25	No Match	17 <1.5:1	17 <1.5:1	19 <1.5:1	20 <1.5:1	
4:1	12.5	11 <1.5:1	11 <1.5:1	12 <1.5:1	13 <1.5:1	26 <2:1	
2:1	25	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	23 2.1:1	
1:1	50	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	21 2.1:1
2:1	100	<10 <2:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	17 <1.5:1	
4:1	200	<10 <2:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	15 <1.5:1	
8:1	400	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	28 <1.5:1	
16:1	800	<10 <1.5:1	12 <1.5:1	<10 <1.5:1	<10 <1.5:1	19 <1.5:1	

Table 14-8
LDG Z-100 Desktop Automatic Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input SWR range: to 10:1 (3:1 on 6 meters).	See below.
Output SWR range: not specified.	See below.
Frequency coverage: 1.8 to 54 MHz.	See text.
Input power: 0.1 to 125 W, 50 W on 6 meters.	Tested at 50 W on HF, 20 W on 6 meters.
Power requirements: 7 to 18 V dc, 250 mA.	Idle 0 mA, max 160 mA at 13.8 V dc.

Measured power loss into resistive loads (%) / Input SWR at match:

SWR	*Load (Ω)*	*160 Meters*	*80 Meters*	*40 Meters*	*20 Meters*	*10 Meters*	*6 Meters*
16:1	3.125	No Match	No Match	15 <1.5:1	10 <1.5:1	<10 <1.5:1	
8:1	6.25	No Match	13 2.6:1	15 <2:1	15 <1.5:1	20 <2:1	
4:1	12.5	No Match	<10 <1.5:1	<10 <1.5:1	<10 <2:1	<10 <1.5:1	
2:1	25	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	17 <1.5:1	
1:1	50	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	20 <2:1
2:1	100	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	
4:1	200	<10 2.1:1	<10 2.2:1	<10 <1.5:1	11 <1.5:1	<10 <1.5:1	
8:1	400	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	11 <1.5:1	<10 <2:1	
16:1	800	11 2.2:1	14 2.3:1	15 <1.5:1	33 <1.5:1	43 <1.5:1	

the next tune is requested. This is particularly appealing for portable low power operations in which every milliamp-hour is carried in.

The operation of this tuner is fundamentally different from the others in that it will only tune when the TUNE button is pushed. This has an advantage in that it will not start tuning inadvertently on high SWR, potentially applying full power during tune. Upon changing frequency, the operator can decide whether or not to initiate the tune cycle. By pushing the button for less than 0.5 seconds, the tuner goes to bypass. Between 0.5 and 3 seconds the tuner will attempt to find a match from memory (typically less than half a second), and only initiate a full tune if needed. If the button is pushed for more than 3 seconds a full tune cycle is initiated, ignoring any memorized settings.

This is easier to deal with than it sounds since the LEDs provide a confirmation of tuning mode. A green LED also indicates if the match is less than 1.5:1 or between 1.5 and 2:1. An early production unit exhibited some inconsistent tuning results on 6 meters. The manufacturer has developed a method of improved internal decoupling that solved this problem and the results reflect that change. LDG states that this will be included in future production units and that anyone having problems with 6 meter tuning on early units should contact them directly.

Manufacturer: LDG Electronics Inc, 1445 Parran Rd, PO Box 48, St Leonard, MD 20685; tel 877-890-3003; fax 410-586-8475; **www.ldgelectronics.com**. 2004 prices: RT-11 Tuner, $209; Remote head, $39; Z-100, $149; 12 V power cube for either tuner, $10.

MFJ-991

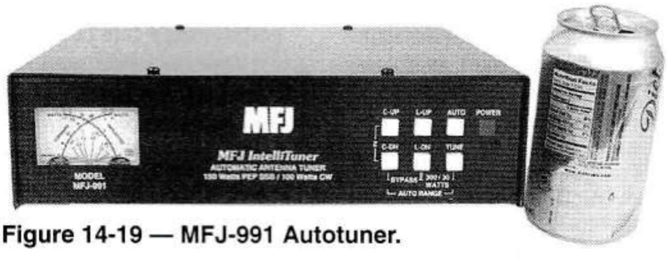

Figure 14-19 — MFJ-991 Autotuner.

MFJ has recently announced a series of autotuners. All are in the typical indoor/ colocated configuration, but with the available remote kit (ordered, but not received in time for the review) can be used as a remote tuner if kept out of the weather. They offer the MFJ-991 (tested, as shown in **Figure 14-19**) 150 W rated tuner, a tuner with similar features rated at 600 W and a 300 W autotuner with digital display, balun and two port antenna switch. As with the LDG, the MFJ tuners require their tuning be accomplished at reduced power.

The MFJ-991 front panel provides full control capability and forward and reflected power metering with "crossed-needle" SWR measurement display. Two features of the '991 are not found on other tested units and are worth mentioning. First, the '991 has the capability to allow manually adjusting the C and L values of the tuner while in receive. This is helpful if you are using your radio for receive only and wish to attempt to match in order to increase signal strength. The second feature is the ability to select the SWR threshold at either 1.5 or 2:1.

In our testing the tuner went to the best match it could find in either case. The difference was in how far you could change frequency before it would automatically retune when you transmitted. If your radio can tolerate a 2:1 SWR, this could provide a benefit. The '991 could tune my G5RV on all bands and remembered the last setting for each frequency (1000 memories), with a one-click reset. Tuning the first time on a frequency was typical of the cycle time of the other units. This unit sounded somewhat louder (it has 18 heavy duty relays) than the other test samples, although not an issue for the single click associated with a memorized tune.

The MFJ-991 can interoperate with ICOM and Alinco tuner control commands. Appropriate interface cables are available from MFJ.

In addition to the obvious controls and indicators, there are some subtle ones as well. There are a number of meter "signals" sent to the operator that will not be clear without a cruise through the manual. For example, if you forget to reduce power before tuning, the power meter swings to full scale and the tuner is bypassed. If you toggle between 1.5 and 2:1 SWR threshold, the meter briefly jumps up to indicate that SWR.

Manufacturer: To order, or for your nearest MFJ dealer, call MFJ Enterprises at 800-647-1800 or order at **www.mfjenterprises. com**, fax 662-323-6551; or write MFJ Enterprises, Inc, 300 Industrial Park Rd, Starkville, MS 39759. [2004] Price: $220.

Table 14-9
MFJ-991 Desktop/Remote Automatic Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input Impedance range: 6 to 3200 Ω.	See below.
Output SWR range: 1.5:1 or 2:1 selectable.	See below.
Frequency coverage: 1.8 to 30 MHz.	As specified.
Input power: 5 to 150 W.	Tested at 50 W.
Power requirements: 12 to 15 V dc, 1.0 A.	Idle 90-300 mA, max 500 mA at 13.8 V dc.

Measured power loss into resistive loads (%) / Input SWR at match:

SWR	*Load (Ω)*	*160 Meters*	*80 Meters*	*40 Meters*	*20 Meters*	*10 Meters*
16:1	3.125	No Match	27	20	15	<10
			<1.5:1	<1.5:1	<1.5:1	<2:1
8:1	6.25	12	14	15	24	19
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	2.6:1
4:1	12.5	<10	<10	<10	<10	12
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1
2:1	25	<10	<10	<10	<10	<10
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1
1:1	50	<10	<10	<10	<10	<10
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1
2:1	100	<10	<10	<10	<10	<10
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1
4:1	200	<10	<10	<10	<10	<10
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1
8:1	400	<10	<10	<10	11	11
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1
16:1	800	<10	16	<10	<10	23
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1

SGC SG-237 and MAC-200

SGC has been making automatic antenna tuners for many years, and offers a wide selection. Until recently, they concentrated on remote tuners, usually the weatherproof type, at power ratings of 100 to 500 W. More recently, they have introduced a number of variations of their SG-237, a board version to build into your own radio or antenna cabinet, an open enclosure model for remote but indoor use and one with controls and indicators intended for portable use. They have recently released a Multiple Antenna Controller, MAC-200. The units we tested were an SG-237 (one that has fed the backstay of my sailboat for some years) and a relatively new combination unit, the MAC-200 (as shown in **Figure 14-20**).

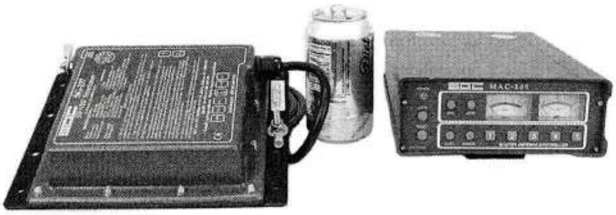

Figure 14-20 — SGC SG-237 (left) and MAC-200.

SGC SG-237 Autotuner

The SG-237 was the first in a series of SGC tuners that were much more compact than their previous (and continuing) line, the SG-230 (200 W, 1.6-30 MHz), SG-231 (100 W, 1-60 MHz) and SG-235 (500 W, 1.8-30 MHz). The SG-237 is rated at 100 W PEP, both for operation and for tuning (40 W continuous carrier power). The transmitter connection is via a 9 foot combined RF, control and power cable (25 and 50 foot extensions are available), while the antenna connection is via a single ceramic insulated terminal. While the tuner is designed to feed a single wire against ground, I have successfully employed mine into coax fed antennas as well as loops and single wires.

This is a rugged unit in a compact waterproof aluminum enclosure that can slide in almost anywhere. It needs no remote control box to work, and I've used mine that way for years. An optional remote control is offered (and we tested with it) to provide a positive indication of match and to allow a lockout or a forced reset. The lockout can be handy in a portable or mobile environment to avoid the tuner attempting to retune due to the temporary proximity of outside objects while in motion, for example.

The '237 provides a memory function that uses the measured transmitter frequency to result in the tuner settings last used to successfully tune to that frequency. This means that instead of taking 5 to 10 seconds to tune, the previously stored settings for a frequency are applied in milliseconds and used if the resulting SWR

Table 14-10
SGC SG-237 Remote Automatic Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input SWR range: corresponding to a minimum 7 foot antenna above 3.5 MHz, 23 feet above 1.8 MHz.	See below.
Output SWR range: 1.4:1 typical.	See below.
Frequency coverage: 1.8 to 60 MHz.	As specified.
Input power: 3 to 100 W.	Tested at 50 W.

Power requirements: 10.5 to 18 V dc, 300 mA. Idle and tuning 270 to 480 mA at 13.8 V dc.

Measured power loss into resistive loads (%) / Input SWR at match:

SWR	*Load (Ω)*	*160 Meters*	*80 Meters*	*40 Meters*	*20 Meters*	*10 Meters*	*6 Meters*
16:1	3.125	21 <1.5:1	21 <1.5:1	19 <1.5:1	No Match	<10 <1.5:1	
8:1	6.25	16 <1.5:1	18 2.1:1	20 <1.5:1	24 <1.5:1	23 <1.5:1	
4:1	12.5	11 <1.5:1	12 <1.5:1	15 <1.5:1	17 <1.5:1	23 <1.5:1	
2:1	25	<10 <1.5:1	11 <1.5:1	21 <2:1	18 <2:1	26 <1.5:1	
1:1	50	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	<10 <1.5:1	12 <1.5:1	29 <1.5:1
2:1	100	<10 <2:1	<10 <2:1	<10 <1.5:1	13 <1.5:1	14 <1.5:1	
4:1	200	No Match	<10 2.1:1	<10 <1.5:1	16 <1.5:1	10 <1.5:1	
8:1	400	No Match	<10 2.1:1	10 <1.5:1	16 <1.5:1	18 <1.5:1	
16:1	800	<10 <1.5:1	<10 <1.5:1	14 <1.5:1	30 <1.5:1	58 <2:1	

is less than 2:1. In some cases they won't be right (and a new tune will be initiated), if a different antenna is used, for example, but their concept is that this tuner is out where the antenna is and thus should see the same impedance for a given frequency each time that frequency is employed.

SGC MAC-200 Multiple Antenna Controller

This device combines a 200 W autotuner, a five port automatic or manual antenna switch, power and SWR meters and a balun in one compact package. This is a nice unit and quickly grew on me in operation. It has enough controls and indicators to keep me entertained even when the bands aren't open.

For some time, most manual tuners have had provisions for switching antennas as well as tuning them. This is very handy, since even those with just one antenna should have a dummy load to switch to for tune up. I have no numbers, but I would guess more hams have more than one HF antenna than have only one. I don't have fancy HF antennas, but I have a G5RV I use on 80 through 10, a Lazy-H, perpendicular to the G5RV for 30 through 10 meters and a 70 foot wire vertical for use when it makes sense (not often in my part of New England).

With the MAC-200, I can memorize which antenna to use on which band. It switches antennas on the first dot of a CW string (full power—no problem) and I'm there. As noted, at my station, on some bands the antenna is a function of the direction to the distant station, I can just touch the MANUAL button and I can select a different antenna by touching a button numbered 1 to 5.

As configured, antenna one is intended for balanced feed, antenna two has a terminal for a single wire feed, and three through five are terminated in UHF (SO-239) coax connector jacks.

The separate meters for power and SWR are handy. The SWR meter is especially convenient in that it computes and indicates SWR independent of power level without an adjustment. The manual warns that the meters are not of the precision type and I can attest to that. Their meter indicated 80 W while my Bird 43 showed 100. This is still very useful to let you know if all is well.

This is a handy unit, in a cabinet matching the cross section of the SGC-2020 transceiver (but somewhat longer), it can be made to fit in most shacks. In mine it can slide beneath the shelf that holds my equipment three inches above the desk, for example.

The one limitation some may have is it provides a "typical" match of 2:1. Well 2:1 is fine for some transmitters, while others may start to "fold back" above 1.5:1 and this may be a problem if yours is in that group. On the other hand, looking at the data it is clear that most of the time it was below 1.5:1 with our loads.

Our original unit remembered which antenna to use on each band, but didn't memorize tuner settings. A warranty replacement unit had a flawless memory function.

Manufacturer: SGC Inc, 13737 SE 26th St, Bellevue, WA 98005; tel 425-746-6310; fax 425-746-6384; **www.sgcworld.com**. 2004 price: SG-237, $360; Smartlock remote control, $59; MAC-200, $360.

Table 14-11
SGC MAC-200 Desktop Automatic Antenna Tuner and Controller

Manufacturer's Specifications	*Measured in ARRL Lab*
Input impedance range: 5-1000 (feed line) 0.2-5000 (long wire).	See below.
Output SWR range: 2:1 typical.	See below.
Frequency coverage: 1.8 to 60 MHz.	As specified.
Input power: 1.5 to 200 W.	Tested at 50 W.
Power requirements: 10 to 18.5 V dc, 230 mA. Idle 240 mA, max 490 mA at 13.8 V dc (meter light off)	Idle 280 mA, max 530 mA at 13.8 V dc (meter light on)

Measured power loss into resistive loads (%) / Input SWR at match:

SWR	*Load (Ω)*	*160 Meters*	*80 Meters*	*40 Meters*	*20 Meters*	*10 Meters*	*6 Meters*
	6 Meters						
16:1	3.125	No Match	39	24	<10	<10	
			<2:1	<2:1	<1.5:1	<1.5:1	
8:1	6.25	No Match	16	18	15	30	
			<1.5:1	<1.5:1	<1.5:1	<1.5:1	
4:1	12.5	<10	<10	<10	18	27	
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1	
2:1	25	<10	<10	<10	<10	25	
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	2:1	
1:1	50	<10	<10	<10	<10	12	27
		<2:1	<1.5:1	<1.5:1	<1.5:1	<2:1	<2:1
2:1	100	<10	<10	<10	<10	14	
		<1.5:1	<2:1	<2:1	<2:1	<1.5:1	
4:1	200	<10	<10	<10	<10	14	
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<2:1	
8:1	400	<10	<10	13	15	17	
		<1.5:1	<1.5:1	<1.5:1	<1.5:1	<1.5:1	
16:1	800	No Match	22	14	25	31	
			<1.5:1	<1.5:1	<1.5:1	<1.5:1	

Medium to High Power Auto Antenna Tuners — The Evolution Continues

Reviewed by Joel R. Hallas, W1ZR
QST *Technical Editor,*
QST *August 2006*

We last looked at a selection of automatic antenna tuners in May 2004 (see previous section). At that time, we picked four examples, two designed for remote operation and two desktop units. While different in many ways, they all were limited to 100 to 200 W. Higher powered automatic antenna tuners have been available, but they were either built into high powered amplifiers (ICOM or Yaesu, for example), were designed for remote operation as in the SGC SG-235, or have been discontinued, as in the case of the Ten-Tec 253.[2]

High powered automatic antenna tuners initially appeared in the 1980s — the Heath 2500, J.W. Miller AT2500, and Daiwa CNA-1001 — but they were not as sophisticated as the units reviewed here, requiring some manual adjustment or presetting of the inductor for best operation.[3,4,5]

Modern solid state transceivers and linear amplifiers with their "no tune" output stages have been well received, especially by those of us who grew up in the vacuum tube era.

The typical transmitter of that time required careful adjustment of the transmitter TUNE, LOAD and DRIVE controls, at a minimum, every time we changed frequency. Failure to do it right could result in destruction of the final amplifier tubes — sometimes a pricey proposition. The other side of the coin is that those transmitters could be adjusted to match antenna loads with a much higher SWR than the current crop of radio gear.

The modern no-tune feature generally means "no tune if the SWR is 1.5:1 or less," something that often doesn't fit with real world antennas and wide amateur bands. The answer has been an antenna system matching unit, often called an antenna tuner, even though many would argue that it doesn't really tune the antenna — rather it transforms the impedance at the bottom of an antenna system to the 50 Ω the radio equipment wants to operate into. Setting semantics aside, it seems that the controls we used to have to adjust haven't disappeared but have moved from the radio to another box — the antenna tuner.

Enter the Automatic Antenna Tuner

To extend the no-tune concept all the way to the antenna requires a no-tune, or *automatic*, antenna tuner. As noted, these have been generally available for some time at the power levels of modern transceivers, Each of these tuners can effectively add an automatic antenna matching system to your medium to high power HF station, at least up to their power limits. Honor the caveats and you will appreciate the freedom they provide.

For this review we have selected three desktop models that span the medium to high power range and highlight some other differences as well. While they all fill the same spot in a station block diagram, they are quite different in several respects and thus should not be directly compared on a 1:1 basis. Rather, determine if one or more will meet your needs, desires and budget.

All of the tuners reviewed in the 2004 report shared the same adjustment mechanism — fixed inductors and capacitors switched by relays. The LDG and MFJ tuners reviewed here share that architecture as well, but with beefier components designed to meet the higher power requirements of their ratings. The Palstar AT-Auto is of a completely different design, a T network with a differential variable capacitor, half on each side of a shunting rotary inductor. The Palstar makes its adjustments via stepper motors rather than relays. Interestingly, while standalone tuners have tended to be of the relay type, many internal tuners have used motor driven capacitors. Motor driven components were also used in the high powered automatic tuners from the 1980s mentioned previously, so this is a time-tested approach. We'll look at each tuner in alphabetical order. These tuners from LDG, MFJ and Palstar offer the convenience of automatic tuning at medium to high power levels.

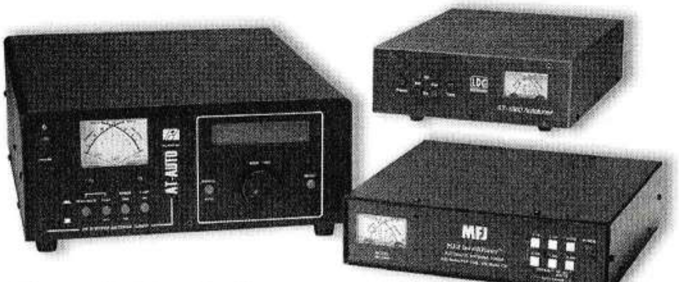

Figure 14-21 — Medium power tuners from Palstar, LDG and MFJ.

LDG AT-1000 Automatic Memory Antenna Tuner

The LDG AT-1000 sits squarely in the middle of the pack, both in terms of price and power rating. The 1000 W level fits nicely with a good selection of popular, not quite legal limit, linear amplifiers, including a couple announced at Dayton in 2006. All of the tuners reviewed here cover 160 through 10 meters, but the AT-1000 offers operation on 6 meters at the 100 W level as well.

The LDG, in common with the other units, has memories to allow a rapid return to the last previously matched settings. In the case of the AT-1000, there are 200 memory locations and the tuner will reset to them in less than half a second. If the antenna you are tuning does not have a memorized setting that works, the tuner can spend up to 8 seconds finding a match, but will remember for the next time.

The LDG approach to memory management is different from some. Some others tuners store settings based on frequency. This is a great approach if you will be tuning a single antenna for each band. In the LDG AT-1000, the last 200 successful matches are stored independent of frequency. A new match request quickly cycles through the 200 to see if a match is found before it begins its tune algorithm. This is a particularly beneficial approach if you switch the tuner output to different antennas on the same band, since it doesn't have to relearn parameters for a particular frequency every time you change antennas. It also avoids the need for frequency measuring circuitry within the tuner.

The AT-1000 has a fairly Spartan front panel that doesn't immediately show its capabilities. LDG has made

Table 14-12
LDG AT-1000

Manufacturer's Specifications
Maximum power, 1000 W SSB, 750 W CW,
500 W digital, 100 W 6 meters.
20 W while tuning (100 W with foldback),
75 W (3:1 SWR), 125 W (1:1 SWR).
Minimum power for tuning: 5 W.
Frequency range: 1.8 to 54 MHz.

Tuning time: 1 to 8 seconds.
0.5 second if memorized

Input impedance: 6 to 800 ..
Power required: 11 to 15 V, 1 A.
Dimensions (height, width, depth):
3 × 13 × 9 inches.
Weight: 6 pounds.

Measured power loss into resistive loads (%) / Input SWR at match. Tested at 50 W.

SWR	*Load (Ω)*	*160 m*	*80 m*	*40 m*	*20 m*	*10 m*
16:1	3.125	No match n/a	No match n/a	No match n/a	No match n/a	No match n/a
8:1	6.25	No match n/a	No match n/a	* **	20 1.6:1	22 **
4:1	12.5	12 2.3:1	* **	* **	* **	18 2.2:1
2:1	25	* **	* **	* **	* **	* **
1:1	50	* **	* **	* **	* **	* **
2:1	100	* **	* **	* **	* 2.1:1	* **
4:1	200	* **	* **	* **	* **	* **
8:1	400	16 2.1:1	* **	* **	22 2.1:1	17 **
16:1	800	No match n/a	18 **	17 **	55 **	No match n/a

*Power loss less than or equal to 10%.
**Matched SWR less than or equal to 1.5:1.

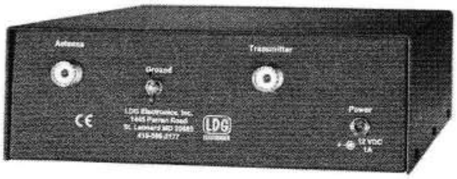

Figure 14-23 — Rear panel connections on the LDG AT-1000.

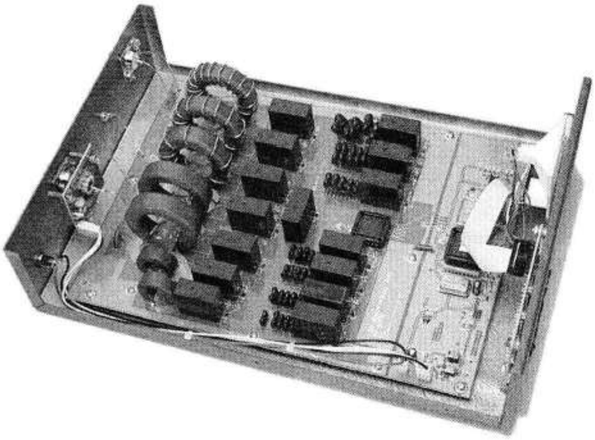

Figure 14-24 — Interior circuitry of the LDG AT-1000.

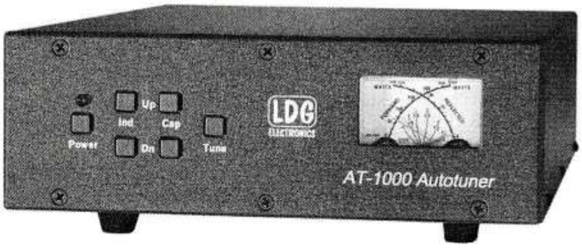

Figure 14-22 — Front panel of the LDG AT-1000.

maximum use of the cross needle wattmeter on the front panel, going further with information delivery than I would have imagined with such a straightforward display. The cross needle wattmeter has two scales, reading forward (1000 W) and reflected (180 W) power simultaneously. The point at which the meter pointers cross is a third scale directly reading in SWR. So far it's just like every other cross needle meter we've seen.

What sets the AT-1000 apart is that LDG has additional indicator functionality that appears as a set of semaphore signals between tuner and operator. Each needle is driven independently in this mode and each can vary in speed and position as it fills you in on the tuner's condition and status. For example, if during a tuning cycle, both arms quickly raise fully (in surrender?) it means too much power has been supplied to allow it to safely tune. It then switches to bypass mode and goes on break until you reduce power and restart the tuning cycle.

Another feature is the ability to change the inductance and capacitance parameters manually. This can be a benefit in at least two ways. First it allows manual fine trimming of the tuning in case you want a better SWR than the tuner finds automatically. Manual tuning can also be used to peak up the receive signal strength if you're listening to a station on a frequency that you can't transmit on. This could include shortwave broadcast frequencies or bands or segments for which you don't yet have license privileges.

How'd it Work? At my station the AT-1000 did everything I asked of it without problems. My antennas tuned satisfactorily on all bands. Memorized settings were recalled without difficulty in a fraction of a second. The 15 page multicolor manual does a good job of describing how to set up and use the tuner. A table of "meter bounce codes" is included, and you will likely want to keep a copy at hand, unless your memory is much better than mine. ARRL Lab test results are shown in **Table 14-12**.

Manufacturer: LDG Electronics, 1445 Parran Rd, St Leonard, MD 20685; tel 410-586-8475; **ldg@ldgelectronics.com**; **www. ldgelectronics.com**. 2006 price: $599.

MFJ-994 Intellituner Automatic Antenna Tuner

At first glance, the MFJ-994 appears to be a very similar tuner to the AT-1000. It has almost the same controls and indicators as the LDG tuner, although they are on opposite sides of the front panel. It makes use of a similar set of semaphore indications. The major differences are:

- The MFJ tuner has a lower maximum power rating (600 vs 1000 W PEP and 300 vs 750 W CW) and a correspondingly lower price.
- The '994 provides an interface to allow operation using the TUNE button on Alinco, ICOM, Kenwood or Yaesu radios equipped to control an antenna tuner. An accessory cable is required to activate this function.
- The '994 has more memory locations, all indexed by frequency.
- The '994 offers a choice of two power meter scales, 3000 or 300 W forward power with corresponding 600 or 60 W reflected power at full scale.

During the review process, we determined that the MFJ-994 has now been superseded by the MFJ-994B. The functionality and operation are similar, with the same tuning and switching elements. The main difference seems to be the memory management function, with the B

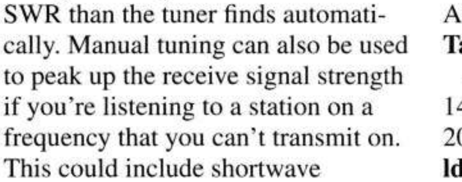

Figure 14-26 — Rear panel connections on the MFJ-994.

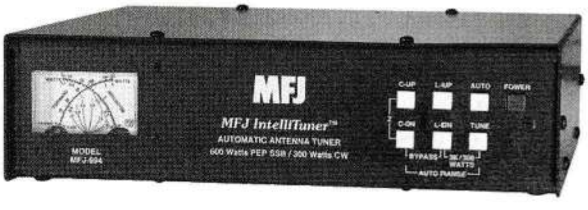

Figure 14-25 — Front panel of the MFJ-994.

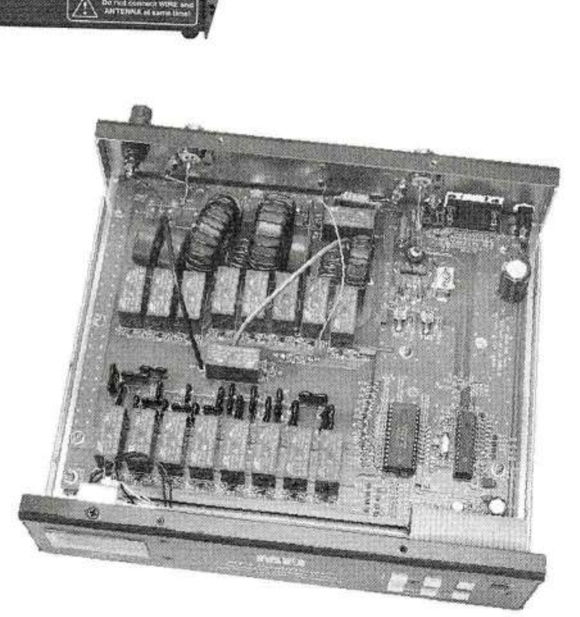

Figure 14-27 — Interior circuitry of the MFJ-994.

Table 14-13
MFJ-994

Manufacturer's Specifications

Maximum power, 600 W PEP SSB, 300 W CW, 20 W while tuning (100 W with foldback).
Minimum power for tuning: 10 W.
Frequency range: 1.8 to 30 MHz.
Input impedance: 12 to 800.
Power required: 12 to 15 V, 1 A.
Dimensions (height, width, depth) 2.8 × 10.1 × 9.2 inches.
Weight: 3.7 pounds.

Measured power loss into resistive loads (%) / Input SWR at match. Tested at 50 W.

SWR	*Load (Ω)*	*160 m*	*80 m*	*40 m*	*20 m*	*10 m*
16:1	3.125	No Match	No Match	10	12	*
		n/a	n/a	2.6:1	2.2:1	2.0:1
8:1	6.25	13	13	*	13	15
		2.4:1	1.8:1	**	**	2.2:1
4:1	12.5	*	*	*	*	15
		**	**	**	**	1.8:1
2:1	25	*	*	*	*	*
		**	**	**	**	**
1:1	50	*	*	*	*	*
		**	**	**	**	**
2:1	100	*	*	*	*	*
		**	**	**	**	**
4:1	200	*	*	*	*	*
		**	**	**	**	**
8:1	400	*	*	*	14	*
		**	**	**	**	1.8:1
16:1	800	*	17	16	11	*
		**	**	**	1.8:1	1.7:1

*Power loss less than or equal to 10%.
**Matched SWR less than or equal to 1.5:1.

model including a virtual memory function said to store up to 10,000 frequency data points versus 1000 in the earlier '994. The other difference noted in the specs is that the new one can tune with a minimum of 2 W, versus 10 W in the original.

The MFJ has some control flexibility hidden behind its few buttons. The "tuned" threshold can be set to either 1.5:1 or 2:1. With the 1.5:1 setting, the match is closer, but it will take a bit longer to find. While 2:1 should happen more quickly, some radios won't put out full power into a 2:1 load. It's nice to have the choice. An "auto tune" function is also provided that causes the tuner to activate if the SWR goes above your threshold by a user selectable 0.5, 1.0 or 1.5. This is handy as you change frequency, but be careful it doesn't drop into tuning mode while you are at full power. Pushing in the AUTO button unambiguously disables the auto tune function, entirely avoiding this concern.

In Summary The MFJ-994 did everything just as promised. It showed no problem handling the 600 W we could deliver, the only tuner of the group that we could test to the limit! It was too quick for me to stopwatch the tuning speed, but my sense was, as expected, that the frequency based memories were a bit quicker if all operations were using the same antenna. It also appeared that the MFJ relays were a bit quieter than the other relay operated tuner, but that could have been because there were generally fewer cycles needed before it was matched. ARRL Lab test results are shown in **Table 14-13**.

MFJ provided a comprehensive 23 page, well illustrated, instruction manual that was clear in its description of how the unit operated and what to expect of it.

Manufacturer: MFJ Enterprises, 300 Industrial Park Rd, Starkville, MS 39759; tel 662-323-5869, fax 662-323-6551; **www. mfjenterprises. com**. 2006 price: $359.95. Auto tuner radio interface cables: MFJ-5124A for Alinco, $19.95 MFJ-5124I for ICOM, $19.95 MFJ-5124K for Kenwood, $59.95; MFJ-5124Y for Yaesu, $59.95.

PALSTAR AT-AUTO

As noted previously, the AT-Auto is of a different architecture than the other tuners in this review. Palstar is noted for its line of high quality manual antenna tuners, so it should not be a surprise that Palstar decided to build its first auto tuner using the same type of in-house constructed parts that have been successful in their other offerings.

This is a large and complex device with features that I haven't seen before in any tuner. The first feature you might notice is the cross needle wattmeter that takes up the left side of the front panel. This is the same display and circuitry built into their PM2000A standalone wattmeter. It provides both average and peak forward and reflected power readings at full scale power levels of 3000/600 (forward/reflected) and 300/60 W. This is an upgrade to the Palstar WM150 wattmeter reviewed in July 2002 with the addition of a PEAK HOLD setting with a 2 second hold time.

The right side of the panel includes a two-line liquid crystal display (LCD), a knob for manual tuning (more later), MANUAL/AUTO and SELECT buttons. It is on this side of the panel that interesting things happen. As with most auto tuners, you apply power and it starts to match your antenna to 50 Ω. What's different with the AT-Auto's tuning sequence is that it first measures frequency, tells you the operating frequency and then politely asks you to turn off your transmitter while it finds a close match. It has a default set of L and C values for each frequency and runs its quiet motors until they are in range. It then asks for another dose of RF to refine the tune. Every time it does this, like a good smart machine, it memorizes its settings and uses them the next time you use that frequency. This approach minimizes interference to others.

A well thought out and, to my knowledge, unique feature of this tuner is its use of a radio interface. This, and many other auto tuners, make use of a radio's tuner interface to initiate the tuning sequence and cause a reduced-power tune signal to be sent from the radio. The AT-Auto has an additional connection from the radio's CAT (computer interface) port. The CAT port provides frequency data from the radio to the tuner. As you tune the receiver, the tuner indicates the frequency and follows along with either its default or memorized values. This means that, if you've been on that frequency before, the tuner will be tuned before you hit the key or PTT! If you haven't been there before, it will use its default values and still be pretty close.

This is particularly handy if you are listening to a frequency at which you aren't licensed to transmit, for example WWV or shortwave broadcast frequencies. The cable for ICOM radios is provided. You do have to think a bit while you use it (yes we have to be even smarter than our smart devices!). The CAT frequency data reflects the frequency of whichever VFO is active. If you are operating split frequency, it will tune back and forth between the two VFO frequencies as you switch between them. This is not likely much of an issue if you are operating in the typical "up five" (VFOs just 5 kHz apart) environment, but could be a disaster if you are operating cross-band split for some reason. The solution — while you have the VFO on your transmit frequency and the tuner properly tuned, tap the MANUAL/AUTO button so that it will be in manual mode and the automation will take a break.

The CAT function, as we wrote this, is available only for radios that support ICOM's CI-V protocol. Palstar expects to have cables and software upgrades to allow interoperation with Kenwood and Yaesu radios soon. Speaking of software, Palstar kindly provides a 9-pin serial cable designed to allow you to install

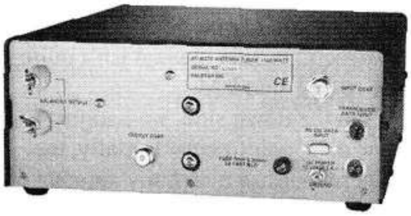

Figure 14-29 — Rear panel connections on the Palstar AT-Auto.

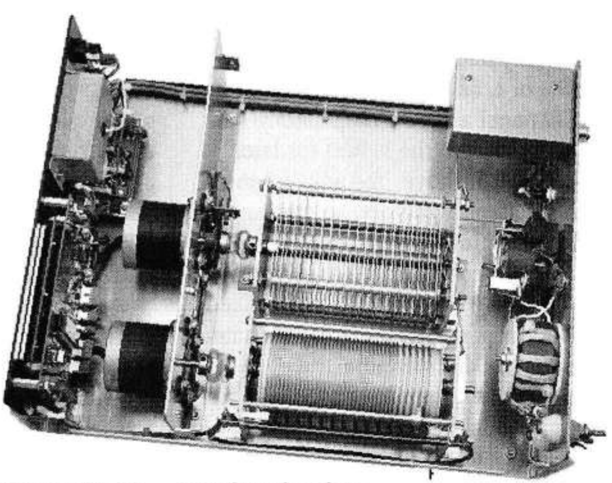

Figure 14-30 — Interior circuitry of the Palstar AT-Auto.

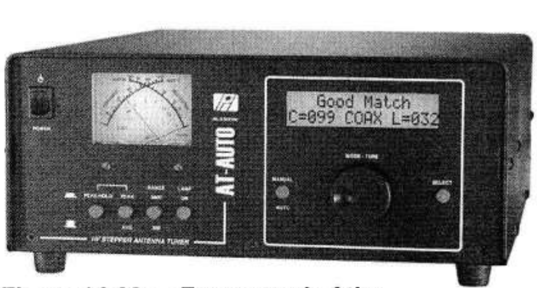

Figure 14-28 — Front panel of the Palstar AT-Auto.

upgraded software when it is available. As we write this, there is no information on how to do this in either the manual or on the Palstar Web site. On the other hand, there is no new software either, so presumably when there's something new to download, they will tell us how to do it. We did successfully download upgraded software for their ZM-30 Antenna Analyzer during that review, so they seem to be able to deal with the process. We understand from Palstar that revised firmware to improve 10 meter operation has been developed and will be available shortly.

Another nice feature of the AT-Auto is that it has provisions for switch selection (via the front panel SELECT button) of either a balanced (through a 4:1 balun at the output) or an unbalanced (coax connected) antenna. The switch doesn't just insert the balun, but provides two completely independent antenna connections, so both can be always available. The memory associates the settings with each antenna port separately, so if you switch between antennas while on the same band, it thoughtfully remembers the settings for each.

New production tuners will also include BYPASS selection. A factory upgrade is said to be available for early production units. Want to do it Yourself? Poke the MANUAL/AUTO button and you are into manual tuning mode. The single knob below the LCD display is used to change the inductance or capacitance. It comes up ready to tune the C; push the knob in briefly and it changes the L. The value of L and C are displayed on the LCD panel as you change them, so you can record the values for later reference. It moves the adjustments via the stepper motors, so minimum effort is required — no heavy cranking needed.

So how'd it work? We really liked the concept of this tuner. By its nature it can provide finer adjustment than the relay types we have seen, although their resolution could be made arbitrarily fine with additional relays and smaller increments of C and L, and higher cost. We also liked the quietness of the motors compared to the clacking relays. The downside can be tuning time. As noted above, with a CATconnected radio and memorized settings it rapidly follows you as you tune. A change in bands can take longer. We measured the following times to retune between bands:

From (Meters)	*To*	*Tune Time (Seconds)*
80	40	11
40	30	4
30	20	3
20	15	2
15	10	2
160	10	45

The times shown are between memorized settings. To tune from a default setting to a finely tuned condition takes about 2 additional seconds if there isn't already a memorized setting for the frequency. It seems especially quick with the CAT function, since it feels like it's all tuned and waiting for you to catch up!

The AT-Auto comes with a well illustrated 20 page manual. The manual assumes you know why you wanted an antenna tuner and where you are going to connect it. This wasn't much of a problem since it has such an informative display, although, if you're expecting "open left flap, remove tuner from box..." type instructions, this manual isn't quite up to that standard.

On the down side, we had just a few disappointments. Initially, the tuner wouldn't match my antenna on 12 and 10 meters. On 12 meters, my measured SWR was 6:1 at the tuner. While it couldn't find a perfect match, and announced MATCH NOT FOUND! on the LCD panel, it actually got to within a 2.4:1 SWR, which my radio found acceptable. On 10 meters my antenna showed an SWR of 3:1 and the best the tuner could do was 2.78:1. In both of these cases, the LCD panel noted that the inductance was at L=000.

This suggested to me that it was the particular value of complex impedance rather than the SWR that was causing the problem. I put in an additional 6 feet of 50 Ω coax to move "around the Smith Chart" and the tuner was then able to find a perfect match in both cases. This is a good trick to keep in mind if faced with this kind of problem with any tuner. The indicator was very helpful in pointing me in the right direction.

A less significant problem was that the frequency metering system required more power to determine the frequency on the higher bands. On 160 through 15 meters, the 10 W that the ICOM transceiver put out in TUNE mode was plenty to allow the frequency to be read and the tune function to be initiated. On 12 and 10 meters, at 10 W, I received a FREQUENCY ERROR CHECK CARRIER LEVEL message on the LCD. It took more power to allow the frequency to be read. This was no problem with another transceiver that allows me to crank it up as far as I want when the TUNE button is pressed, but the ICOM had a fixed level. A quick fix is to switch to AM or FM mode to put out more power. Just hit the bar on your PTT microphone or use a straight key in CW mode.

Upon further investigation, I found that it worked fine on all bands with 10 W to a matched load. At some impedances, apparently there was not enough signal for the measuring circuits. Another workaround would be to manually set it within range and then let the AT-Auto fine tune to perfection. A handy feature of this tuner is that it can be tuned at 1500 W with no relay contacts to burn. This is good since it sometimes notices a bit of reflected power and trims the adjustment after you turn on the amplifier. ARRL Lab test results are shown in **Table 14-14**.

Manufacturer: Palstar Inc, 9676 N Looney Rd, Piqua, OH 45356; tel 800-773-7931, fax 937-773-8003; **info@palstar.com**; **www.palstar.com**. 2006 price: $1195.

In Summary

Each of these tuners did what they said they would over most of their frequency range and each could be of benefit to operators using higher power than that handled by other automatic antenna tuners. Some care needs to be taken with their operation, beyond that found in the lower

Table 14-14
Palstar AT-Auto

Manufacturer's Specifications

Maximum power: 1500 W single tone.
Frequency range: 1.8 to 30 MHz.
Input impedance: 160 to 20 meters, 1200 ±j1200; 15 to 10 meters 1000 ±*j*1000.
Tuning time: 1-30 seconds.
Dimensions (height, width, depth): 5.5 × 12.5 × 16 inches.
Weight: 25 pounds.

Measured power loss into resistive loads (%) / Input SWR at match. Tested at 50 W.

SWR	*Load (Ω)*	*160 m*	*80 m*	*40 m*	*20 m*	*10 m*
16:1	3.125	54 **	33 **	19 **	* **	No Match n/a
8:1	6.25	42 **	24 **	15 **	12 **	No Match n/a
4:1	12.5	30 **	17 **	14 **	* **	No Match n/a
2:1	25	24 **	12 **	* **	* **	26 **
1:1	50	16 **	* **	* **	* **	* **
2:1	100	13 **	* **	* **	* **	20 **
4:1	200	11 **	* **	* **	12 **	43 **
8:1	400	12 **	* **	11 **	23 **	42 **
16:1	800	16 1.6:1	25 **	24 **	55 **	49 **

*Power loss less than or equal to 10%.
**Matched SWR less than or equal to 1.5:1.

powered units, because of the power involved and particularly because of the difference between "tuning power" and "operating power." Each unit is specified to be tuned using reduced power from the exciter at somewhere in the 10 to 100 W range.

Each tuner finds a match at that power level and then it should be safe to switch on the linear amplifier — but is it? With the MFJ and Palstar tuners, there were conditions under which it would initiate the tuning cycle again at full power, even after being within specification at the lower tuning power. The reason is likely a combination of a larger sample of RF to make reflected power measurement and nonlinearity of the detectors, not surprising for analog circuitry. Different tuners will react differently under this circumstance. The Palstar actually doesn't mind tuning at 1500 W and will calmly trim up without a whimper. The MFJ tuner will note that it is tuning above the specified tuning power spec and switch to bypass — whoops, you had the linear pre-tuned with a dummy load so that it would be ready to go, and now it sees a 9:1 SWR, or whatever your antenna has, going directly to the amplifier. This could cause a catastrophic reaction in your amplifier! The LDG tuner is different in this regard. It only initiates a tune cycle if the TUNE button is pressed.

So what to do? There are two potential solutions. One, manually trim up the tuning for an even lower SWR using the L and C up and down buttons. If the reflected power is fully nulled, it should avoid this problem. The other approach is to switch the tuner to manual mode after it finds its initial match. That will prevent it from automatically retuning while high power is applied. Or do both. Arguably this changes tuner operation from "auto" to "semi-auto," but I recommend it for any high power relay switched tuner. The consequences of not doing so could be extreme to either the amplifier or the relay contacts or both.

A New Generation of Balanced Antenna Tuners

Reviewed by Joel R. Hallas, W1ZR
QST *Technical Editor*
QST *September 2004*

A new breed of antenna tuner available in different flavors from multiple manufacturers addresses concerns about using baluns with high SWR to feed balanced antenna systems. Paul Danzer's article in the April 2004 issue of *QST* brought to paper a concern many of us have had for some time — the use of baluns at the output of antenna tuners.[6] As noted in his article, if the balanced load is near the balun's design impedance (typically 200 Ω for the usual 4:1 balun) all is well.

Unfortunately, the typical random sized center fed antenna with random length ladder line feed has an impedance at the feed point that varies dramatically with frequency. The result can be heating and loss (and occasional damage) at the balun. These effects were well documented in a series of *QST* articles by Frank Witt and later in a performance evaluation of unbalanced tuners with both balanced and unbalanced loads.[7,8]

So How Do We Fix the Problem?

As Danzer noted, the classic solution has been the use of an inherently balanced tuner. The commercially manufactured E.F. Johnson Matchbox tuners of the '50s worked reasonably well in their day, and over the bands that they covered (we didn't have the 60, 30, 17 and 12 meter bands back then). Now at least three manufacturers have begun offering balanced antenna tuners of a different configuration. For this review we selected the MFJ-974H and the Palstar AT1500BAL and AT4K. Interestingly, each of the three uses a different architecture, and each is different from the design of the old Matchbox!

In addition to those units, SGC has announced a low power self-contained auto-tuner, the SG-211, which shares the design concept of the AT4K and will be the subject of an upcoming Short Takes column. We thought it would be good to evaluate the performance of the medium and high-power units and provide a comparison to the old Johnson tuner.

What's in the New Tuners?

Glad you asked! As noted, each of the tuners uses a different design configuration. Each can be directly compared to some of the common unbalanced configurations. Note that the power ratings and price of the Palstar tuners put them in a different category from the MFJ units and thus, direct comparisons may not be appropriate. Both the MFJ and the Palstar AT1500BAL are fully balanced tuners — the MFJ a dual T section design with shunt L, and the Palstar a dual L section with shunt C. The Palstar AT4K tuner takes a completely different approach. It uses the insulated unbalanced scheme suggested in Paul's article and described in detail in *The ARRL Antenna Book*.[9]

The relationship of the designs to their more commonly encountered unbalanced configurations are shown in simplified schematics in **Figures 14-32** through **14-33**. The other differences between the units are in their ratings. The MFJ tuner covers 160 through 6 meters, while the Palstar tuners top out at 10 meters. Both Palstar tuners are rated at 1500 W or greater (the AT4K has a reduced rating below a 25 Ω load) while the MFJ tuner is rated at 300 W PEP, 150 W CW.

The ratings of the Johnson Matchboxes were established in a day when amateur power levels were specified based on average dc power *input*, rather than the current PEP RF *output*

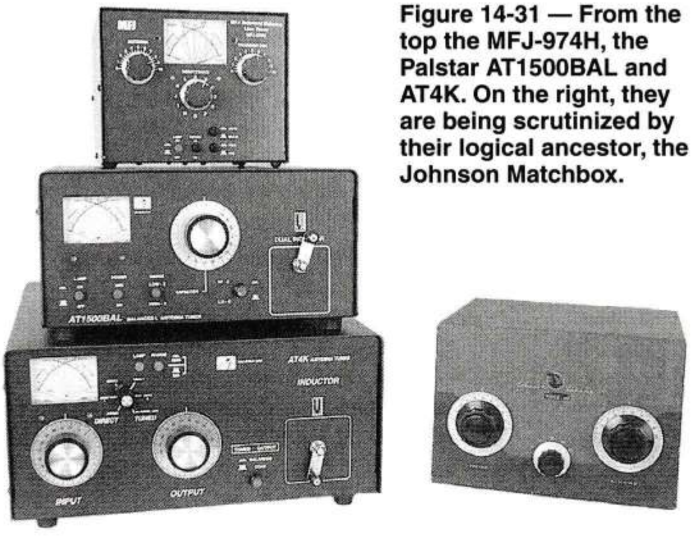

Figure 14-31 — From the top the MFJ-974H, the Palstar AT1500BAL and AT4K. On the right, they are being scrutinized by their logical ancestor, the Johnson Matchbox.

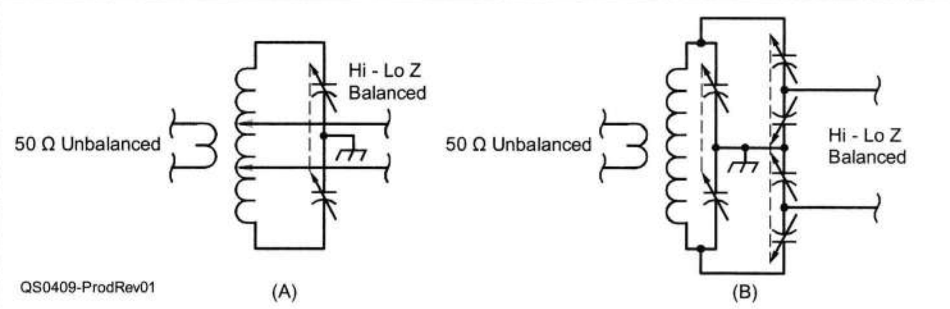
Hi - Lo Z
Balanced
50 Ω Unbalanced
50 Ω Unbalanced
Hi - Lo Z
Balanced
QS0409-ProdRev01
(A)
(B)

Figure 14-32 — At A, the traditional balanced antenna tuner. At B, the Johnson Matchbox antenna tuner. Simplified view without bandswitching.

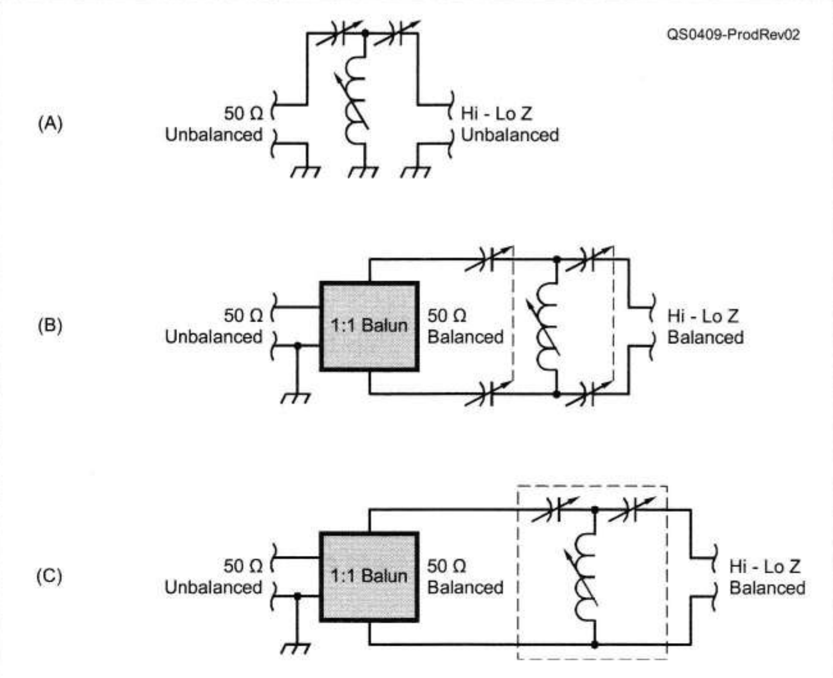
QS0409-ProdRev02
(A)
50 Ω
Unbalanced
Hi - Lo Z
Unbalanced
(B)
50 Ω
Unbalanced
1:1 Balun
50 Ω
Balanced
Hi - Lo Z
Balanced
(C)
50 Ω
Unbalanced
1:1 Balun
50 Ω
Balanced
Hi - Lo Z
Balanced

Figure 14-33 — At A, an unbalanced T-network tuner. At B, a balanced T-network tuner. At C, an unbalanced T-network tuner for balanced loads.

power. A "275 W" Johnson matchbox was thus rated to work with transmitters running 275 W dc input, or about 200 W average output. This was in the day of plate modulated AM service, so that rating further translates to 800 W PEP under today's rules. Similarly, a "kW Matchbox" would likely be rated today as a 3 kW PEP tuner.

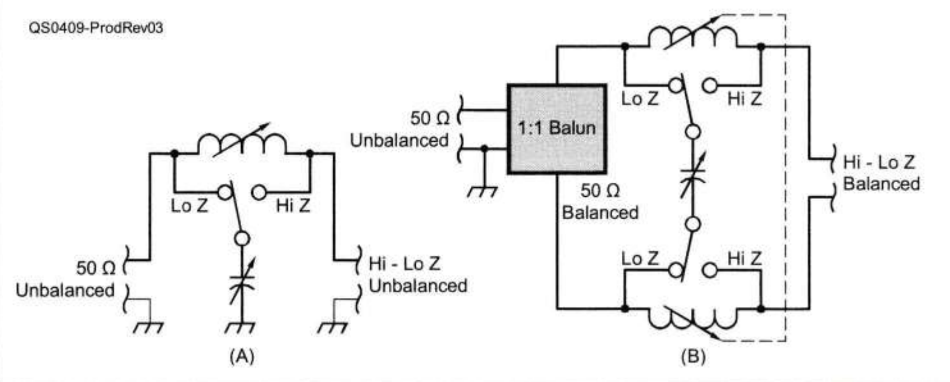
QS0409-ProdRev03
Lo Z
Hi Z
50 Ω
Unbalanced
Hi - Lo Z
Unbalanced
(A)
50 Ω
Unbalanced
1:1 Balun
50 Ω
Balanced
Lo Z
Hi Z
Lo Z
Hi Z
Hi - Lo Z
Balanced
(B)

Figure 14-34 — At A, an unbalanced L-network tuner with switched capacitor. At B, a balanced L-network tuner with switched capacitor.

MFJ-974(H) Balanced Antenna Tuner

MFJ provides two versions of their '974, with ('974H) and without ('974) the capability to cover 160 meters. Both are rated at 300 W PEP, and 150 W CW. At only 7½ inches wide, it takes up a small amount of desk space. Interestingly, it has about the same shape as the old 275 W Johnson Matchbox, but is 1 to 2 inches smaller in each dimension. As noted above, this unit uses a T section design as in **Figure 14-33B**, so it is not necessary to switch capacitors from one side to the other as the impedance changes from high to low.

The inductor is varied by selecting among 12 steps. The H model has an extra inductance for 160 meters that changes the tuning range as needed. A cross-needle two-range (30 and 300 W full scale) power/SWR meter can be set to read peak or average power. Both the power level and null matched the meter on my transceiver quite closely. The capacitors are direct (1:1) drive, with plastic gearing between the ganged units. I found them easy to turn and to get a null. I used the old trick of tuning for maximum receive signal, first with the L switch, then with the capacitors.

MFJ provides a page of nominal settings by band that you may find helpful. When you switch to adjusting the tuner while transmitting, be sure to note their caution about not changing the inductor switch with power applied. You may burn the contacts otherwise. I also make it a practice to always tune at reduced power to avoid hitting my finals with a heavy mismatch at full power. When you have it set, turn up the power and confirm that all is well.

One caution that applies to all T section tuners is to note that multiple settings can result in low SWR. Unfortunately, some will have higher losses than others. The rule of thumb is to use the setting with the smallest inductance (the L switch position on

Table 14-15
MFJ-974H Balanced Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input load range: 12 to 2000 Ω.	See below.
Output SWR range: Not specified.	See below.
Frequency coverage: 1.8 to 54 MHz.	See below.
Input power: 150 W average, 300 W PEP.	Not tested.
Size: 6 × 7.5 × 9 inches (HWD).	

SWR	*Load (Ω)*		*160 Meters*	*80 Meters*	*40 Meters*	*20 Meters*	*10 Meters*	*6 Meters**
		Power Loss %	55	41	32	20	16	
8:1	6.25	SWR BW %	0.9	0.6	1.0	2.0	2.0	Not Tested
		Imbalance**	0	0	0	0	0	
		Power Loss %	40	33	23	<10	10	
4:1	12.5	SWR BW %	0.6	0.9	1.4	2.8	3.4	Not Tested
		Imbalance	0	0	0	0	0	
		Power Loss %	25	22	<10	<10	<10	
2:1	25	SWR BW %	0.8	1.1	3.2	4.9	2.4	Not Tested
		Imbalance	0	0	0	0	0	
		Power Loss %	22	15	32	20	<10	<10
1:1	50	SWR BW %	1.1	1.4	2.1	2.1	12	10
		Imbalance	0	0	0	0	0	0
		Power Loss %	18	<10	<10	<10	<10	
2:1	100	SWR BW %	1.1	1.7	4.1	7.0	25	Not Tested
		Imbalance	0	0	0	0	0	
		Power Loss %	17	<10	<10	<10	<10	
4:1	200	SWR BW %	1.1	2.8	2.5	9.1	10	Not Tested
		Imbalance	0	0	0	0	0	
		Power Loss %	12	<10	<10	<10	<10	
8:1	400	SWR BW %	1.5	3.4	7.5	7.0	3.7	Not Tested
		Imbalance	0	0	0	0	0	
		Power Loss %	10	<10	<10	<10	<10	
16:1	800	SWR BW %	1.9	4.3	8.3	7.7	3.7	Not Tested
		Imbalance	0	0	0	0	0	
		Power Loss %	<10	<10	<10	10	11	
32:1	1600	SWR BW %	2.1	2.6	5.6	4.4	3.0	Not Tested
		Imbalance	0	0	0	0	0	

Notes
Power losses are expressed as a percentage. A 10% power loss represents less than half (0.46) a dB.
The SWR bandwidth is the percentage of the measurement frequency that can be changed with the SWR staying under 1.5:1.
*The ARRL test fixture is only usable at 50 . on 6 meters.
**As defined in F. Witt, "Evaluation of Antenna Tuners and Baluns—An Update," *QEX*, Sep-Oct 2003, pp 3-14.

this tuner) and highest capacitance that will provide a match. One limitation of the switched rather than roller inductor is that you can't get values in between. I didn't find that much of a problem, as I was always able to get to 1.5:1 or better SWR with my antennas.

As you make adjustments, I suggest you log the settings for each frequency in a table such as the one supplied in the back of the manual. You may want to make photocopies or set up a spreadsheet so you will be able to repeat the process for your next antenna. If you find that you can't get a match at some frequency, try adding 6 feet of ladder line at a time. The different length will move the impedance around and may solve the problem, but you will have to retune on the other bands and hope to find a length that will tune properly on all bands.

Balanced loads are connected via standard (¾ inch) spaced multipurpose binding posts. These accept wire ends, or dual banana plugs, a handy connector for balanced feed lines at this power level. The '974H also provides a capability to match unbalanced loads and provides a coax connector for that purpose. It is necessary to provide a jumper between the lower balanced load binding post and the ground terminal to complete the unbalanced hookup. Note that the output is not switched, so an antenna should be connected to either the coax connector or the balanced jacks, but not both.

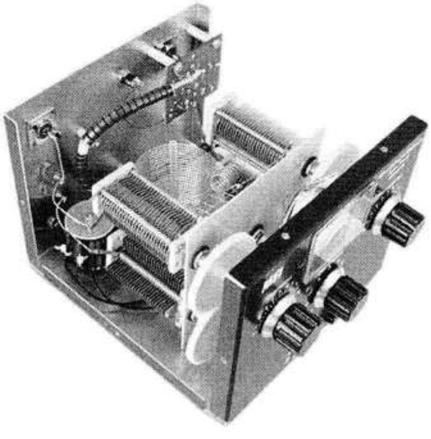

Figure 14-35 — MFJ-974H, inside view.

Note that while the '974 is specified to match a wide range of impedances on 6 meters, neither the Lab nor W1ZR could test at any impedance other than 50 Ω. I tried the unbalanced connection arrangement with my coax-fed G5RV. It worked fine until I tried it on 160 meters, not one of the bands the antenna is supposed to operate on. I was able to quickly find a match at low power. When I turned up the power past 70 W, arcing was evident inside the unit. I opened it up and was surprised to find that rather than the expected arcing tuner element, I found the arcing source was at the metering circuit board. Measured Lab data is shown in **Table 14-15**.

The unit went back to MFJ for repair under their warranty. Upon return, it worked fine at 100 W (my maximum power with the transceiver) and I was told by MFJ that a signal wire had been dressed too close to the meter board and arced. It was repositioned and MFJ installed the intended tie-wrap to hold it away from the meter board. The tuner was retested and had exhibited no problems.

Manufacturer: MFJ Enterprises Inc, 300 Industrial Park Rd, Starkville, MS 39759; tel 800-647-1800; fax 662-323-6551; **www.mfjenterprises.com**. 2004 price: MFJ-974, $179.95; MFJ-974H, $199.95.

Palstar AT4K Antenna Tuner

Unlike the other tuners in this review, the AT4K might best be described as an unbalanced tuner designed to feed unbalanced or balanced loads. This sounds contradictory, but really isn't, as described in footnotes 1 and 6. If you just think of this tuner as a "four-terminal black box" (see **Figure 14-33C**) connecting a balanced load to a balanced source, followed by a 1:1 balun going to the transceiver, you may get the picture.

An unbalanced T section, isolated from ground within the box, performs the required impedance transformation to match the impedance at the bottom of the balanced transmission line to 50 Ω to operate properly through the internal balun. This tuner also provides a heavy duty relay to connect the common points of the input, output and inductor to the chassis to convert it to a typical T section tuner for unbalanced loads.

An additional feature, not found in any of the other tuners in this review, is antenna switching capability. The AT4K has a single pair of balanced output terminals, but also has three coax outputs. Two can be tuned, or switched to straight through, while the third provides straight though operation only. This is a great spot to connect a dummy load, or a well matched Yagi.

The inside view gives a sense of the heavy duty construction of this massive tuner. The rotary inductor is fabricated from silver plated strip stock, rather than the usual wire, and the rolling contact is designed to grip a significant portion or the coil, not just the edge. The inductor is driven by a smooth turn-count dial reading out to 1⁄10 of a turn and resetable in

Table 14-16
Palstar AT4K Balanced/Unbalanced Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input load range: 8 to 2000 Ω.	See below.
Output SWR range: Not specified.	See below.
Frequency coverage: 1.8 to 30 MHz.	See below.
Input power: 2500 W single tone.	Not tested.
Size: 5 × 15 × 16 inches (HWD).	

SWR	*Load (Ω)*		*160 Meters*	*80 Meters*	*40 Meters*	*20 Meters*	*10 Meters*	*6 Meters**
8:1	6.25	Power Loss %	24	14	10	10	34	
		SWR BW %	1.1	2.6	6.4	4.9	0.5	
		Imbalance	0.6	0.7	0	0.6	5.2	
4:1	12.5	Power Loss %	12	<10	<10	<10	18	
		SWR BW %	1.9	4.0	17	8.5	0.7	
		Imbalance	0	0.4	0	0.4	1.7	
2:1	25	Power Loss %	<10	<10	<10	<10	<10	
		SWR BW %	3.0	7.4	39	21	1.5	
		Imbalance	0	0.2	0	0.5	0	
1:1	50	Power Loss %	<10	<10	<10	<10	<10	
		SWR BW %	5.5	16	69	42	3.0	
		Imbalance	0.3	0	0	0.4	0.3	
2:1	100	Power Loss %	<10	<10	<10	<10	<10	
		SWR BW %	6.1	26	49	44	10	
		Imbalance	0.4	0	0	0	0	
4:1	200	Power Loss %	<10	<10	<10	<10	37[1]	
		SWR BW %	6.9	22	24	32	0.8	
		Imbalance	0.5	0	0	0	3.4	
8:1	400	Power Loss %	<10	<10	<10	<10	Note 1	
		SWR BW %	7.8	15	14	15		
		Imbalance	1.5	0.3	0	0		
16:1	800	Power Loss %	<10	<10	<10	<10	Note 1	
		SWR BW %	7.2	10	10	8.5		
		Imbalance	2.6	0.5	0	0		
32:1	1600	Power Loss %	<10	<10	<10			
		SWR BW %	4.4	6.9	5.6	No Match	No Match	
		Imbalance	4.9	0	0			

Notes

[1] According to the owner's manual for the AT4K, the matching range is limited on 10 meters. In testing, it was found that matches (SWR <1.5:1) could be obtained on 10 meters with some higher impedance loads by using a large amount of inductance. However, it is very likely that the tuner was close to self-resonance under these conditions, and this type of operation should be avoided due to the high losses in the tuner.

Power losses are expressed as a percentage. A 10% power loss represents less than half (0.46) a dB.

The SWR bandwidth is the percentage of the measurement frequency that can be changed with the SWR staying under 1.5:1.

between marks. As with its sister units, the capacitors are driven by 5:1 reduction drives with 0 to 100 logging scales.

One design challenge with a tuner built of these large parts is to keep the minimum capacitance low enough so the unit will meet specifications at 10 meters. As shown in the data, Palstar has done a reasonably good job with this, but the limitations should be noted. I found that this unit could match all my antennas, balanced or unbalanced, except on 17 and 10 meters where not all impedances are within range. Again, a change in transmission line length may move your impedance to a value within the tunable range.

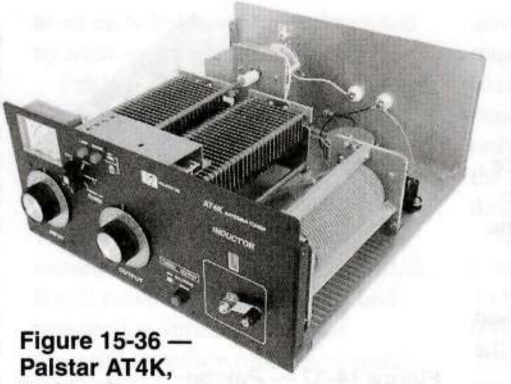
Figure 15-36 — Palstar AT4K, inside view.

The metering was consistent with my other wattmeters, both in position of reflected null and forward power. I was able to easily return to a previous setting using the resolution of the dial scales. Measured Lab data is shown in **Table 14-16**.

Palstar AT1500BAL Balanced Antenna Tuner

The Palstar AT1500BAL is of the configuration shown in **Figure 14-34B**. It is a legal limit device and a look inside makes me believe that it can handle serious power. The rotary inductors, two edge-wound silver plated ball-bearing monsters, driven by a toothed fiberglass belt and turns counting dial take up a lot of the interior space. The rest is largely occupied by a 450 pF 4.5 kV variable driven by a smooth 5:1 vernier dial.

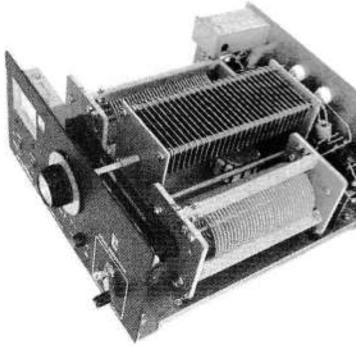

Figure 14-37 — Palstar AT1500BAL, inside view.

The inductors and capacitor are made in-house and work very well. I was particularly impressed with the resetability of the controls. The turns counter reads out to tenths of a turn, and can be easily reset to ½ a tenth out of the 32 turns end-to-end. The vernier drive has a dial with 100 divisions and no perceivable backlash. Once you make up a table of tuning values by frequency, retuning should be a snap. I found the tuner easy to use at W1ZR, matching all reasonable antenna configurations I tried. I am only able to test at the 500 W level and, as expected, encountered no difficulties.

The cross-needle meter (300 and 3000 W full scale ranges for forward average power) agreed closely with those on my equipment. One note, on many tuners, power is only required for meter lighting. On this unit, relays are used — both to switch capacitors from one end to the other and to switch in additional capacitance. The tuner will work only in the "high impedance/ no extra capacitance" mode if you neglect to plug in the

Table 14-17
Palstar AT1500BAL Balanced Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input load range: 160 to 20 meters, 2500 ± j2500 Ω.	See below.
20 meters 1000 ± j1000 Ω; 10 meters, 500 ± j500 Ω.	See below.
Output SWR range: Not specified.	See below.
Frequency coverage: 1.8 to 30 MHz.	See below.
Input power: 1500 W PEP.	Not tested.
Size: 6.5 × 12.5 × 15 inches (HWD).	

SWR	*Load (Ω)*		*160 Meters*	*80 Meters*	*40 Meters*	*20 Meters*	*10 Meters*	*6 Meters**
		Power Loss %	<10	<10	<10			
8:1	6.25	SWR BW %	No Match	No Match	9.9	6.6	5.4	
		Imbalance	0.3	0	0			
		Power Loss %	<10	<10	<10			
4:1	12.5	SWR BW %	No Match	No Match	1.3	1.3	10	
		Imbalance			0	0	0	
		Power Loss %	<10	<10	<10	17	<10	
2:1	25	SWR BW %	1.5	46	42	19	13	
		Imbalance	0	0	0	0	0	
		Power Loss %	<10	<10	<10	<10	<10	
1:1	50	SWR BW %	>100	>100	>100	>100	20	
		Imbalance	0.3	0	0	0	0	
		Power Loss %	<10	<10	<10	<10	<10	
2:1	100	SWR BW %	>100	54	58	56	61	
		Imbalance	0	0	0	0	0	
		Power Loss %	<10	<10	<10	<10	<10	
4:1	200	SWR BW %	>100	29	26	30	3.7	
		Imbalance	0.2	0	0	0	0	
		Power Loss %	<10	<10	<10	<10	13	
8:1	400	SWR BW %	17	23	18	20	1.7	
		Imbalance	0	0	0	0	0	
		Power Loss %	<10	<10	<10	<10	18	
16:1	800	SWR BW %	11	11	13	13	4.0	
		Imbalance	0	0	0	0	0	
		Power Loss %	<10	<10	<10	<10		
32:1	1600	SWR BW %	5.6	8.0	7.8	5	No Match	
		Imbalance	0	0	0	0		

Notes
Power losses are expressed as a percentage. A 10% power loss represents less than half (0.46) a dB.
The SWR bandwidth is the percentage of the measurement frequency that can be changed with the SWR staying under 1.5:1.

supplied wall-wart or connect an external 12 V dc supply. I found I needed to exercise both relays to tune my antenna on all bands.

Just after we purchased the AT1500BAL tuner for evaluation, Palstar added the AT1KBAL tuner to their line. In spite of their product numbers, they both are specified to handle the US legal limit. There is a key difference, however. The AT1KBAL tuner has the capacitors on the output side only. This results in a tuner that is designed to match from around 100 Ω up and is similar to an early balanced tuner described by Measures.[10,11]

Palstar has informed us that later this year they will replace the AT-1500BAL and the newer (and not tested) AT1KBAL with a new balanced tuner, the model BT1500A. This tuner will share the basic architecture and design of the AT1500BAL. It will make use of a new switched two-section variable capacitor designed to reduce minimum capacitance. This is intended to improve the tuning range, particularly on 10 meters. It will also have the inductors driven in tandem, removing the requirement for the drive belt and said to make for smoother tuning. Palstar expects to offer the BT1500A as their only balanced tuner once material for the others is exhausted. Measured Lab data is shown in **Table 14-17**.

Manufacturer: Palstar Inc, 9676 N Looney Rd, PO Box 1136, Piqua, OH 45356; tel 937-773-6255; fax 937-773-8003; **www.palstar.com**. 2004 prices: AT4K, $895; AT1500BAL, $695.95; AT1KBAL, $595.

E. F. Johnson Matchbox Antenna Tuners

Before WW II, the E. F. Johnson Company of Waseca, Minnesota, manufactured high quality variable capacitors, inductors and ceramic parts for electronics. By 1950 they were advertising their innovative Viking I transmitter in *The ARRL Radio Amateur's Handbook* and the next year their Matchbox antenna tuner for $48.95. A kW version was advertised in 1957. These tuners were based on the classic prewar balanced tuned circuit tuner (**Figure 14-32A**), but with an added twist. While the classic tuner of the day used plug-in coils to change bands, and manually attached tap points on the coil to change impedances, the Matchbox was bandswitching and had a unique differential capacitor arrangement to adjust the transformation ratio from the front panel, as shown in **Figure 14-32B**.

Models were available with and without SWR metering and the units included a TR relay useful for the typical separate transmitter and receiver setups of the day. By the '60s, the typical amateur antenna system had become one designed for resonant matched operation using coaxial cable without antenna tuners. Radios moved from AM to SSB, and Johnson changed their focus from amateur equipment to the commercial two-way FM radio market.

Their tuners are still a popular item at flea markets and on electronic auction sites. While not exactly a part of this review, Matchbox data reported by Witt in his earlier article is presented for comparison. As noted in the tables, the Matchbox tuners work well, but over a typically smaller impedance range than their modern counterparts. They also do not have band switch positions for any of the newer bands, although there is enough tuning range so they can typically cover 17 and 12, but not 30 or 60 meters. Within these limits, however, they can still be put to good use. Until the units discussed in this review became available, the Matchboxes were the only commercial tuner

Table 14-18
E. F. Johnson 275 W Matchbox, Balanced Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input load range: 25 to 1250 Ω.	See below.
Output SWR range: not specified.	See below.
Frequency coverage: 80, 40, 20, 15 and 10 meters	See below.
Input power: 275 W dc input.	Not tested.
Size: 7 × 10 × 10.5 inches (HWD).	

SWR	*Load (Ω)*		*80 Meters*	*40 Meters*	*20 Meters*	*15 Meters*	*10 Meters*
		Power Loss %					
4:1	12.5	SWR BW %	No Match	No Match	No Match	No Match	No Match
		Imbalance					
		Power Loss %				<10	<10
2:1	25	SWR BW %	No Match	No Match	No Match	4	4
		Imbalance				0	0
		Power Loss %	<10		<10	<10	<10
1:1	50	SWR BW %	2	No Match	>5	>5	3
		Imbalance	0	0	0	0	
		Power Loss %	<10	<10	12	11	<10
2:1	100	SWR BW %	2	4	>5	>5	3
		Imbalance	0	0	0	0	0
		Power Loss %	11	<10	<10	<10	11
4:1	200	SWR BW %	2	4	>5	>5	3
		Imbalance	0	0	0	0	0
		Power Loss %	12	<10	<10	<10	11
8:1	400	SWR BW %	2	4	>5	>5	3
		Imbalance	0	0	0	0	0
		Power Loss %	10	<10	<10	<10	11
16:1	800	SWR BW %	2	4	>5	>5	3
		Imbalance	0	0	0	0	0
		Power Loss %	10	<10			
32:1	1600	SWR BW %	2	4	No Match	No Match	No Match
		Imbalance	0	0			
		Power Loss %	10				
64:1	3200	SWR BW %	2	No Match			
		Imbalance	0				

Notes
Power losses are expressed as a percentage. A 10% power loss represents less than half (0.46) a dB.
The SWR bandwidth is the percentage of the measurement frequency that can be changed with the SWR staying under 1.5:1.

in wide use specifically designed to work with balanced antenna systems. Measured data from Frank Witt is shown in **Tables 14-18** and **14-19**.[12]

Notes

[1] R. D. Straw, Editor, *The ARRL Antenna Book,* 21st Edition. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9876. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

[2] The SG-235 is a 500 W version of the SG-230 with similar specifications, except needing a longer antenna. The SG-230 was reviewed in *QST* in November 1993.

[3] P. Pagel, "Product Review: Heath SA-2500 Antenna Tuner," *QST,* Mar 1985, pp 39-41.

[4] L. Aurick, "Product Review: J.W. Miller Automatic Antenna Tuner Auto-Trak Model AT2500," *QST,* Jul 1981, pp 42-43.

[5] S. Gerli, "Product Review: Daiwa CNA-1001 Automatic Antenna Tuner," *QST,* Nov 1981, pp 41-42.

[6] P. Danzer, N1II, "Open Wire Feed Line—A Second Look," *QST,* Apr 2004, pp 34-36.

[7] F. Witt, AI1H, "How to Evaluate Your Antenna Tuner—Parts 1 and 2," *QST,* Apr 1995, pp 30-34 and May 1995, pp 33-37.

[8] R. Lindquist, N1RL, "Product Review—QST Compares: Four High-Power Antenna Tuners," *QST,* Mar 1997, pp 73-77.

[9] R. D. Straw, Editor, *The ARRL Antenna Book,* 21st Edition. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9876. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

[10] Most center-fed antennas longer than 0.5 Ω are likely to meet this criterion. Very short antennas, or some low impedance driven arrays (8JK, for example) may have problems with this configuration, depending on feed line length.

[11] R. Measures, AG6K, "A Balanced Balanced Antenna Tuner," *QST,* Feb 1990, pp 28-32.

[12] See Note 7.

Table 14-19

E. F. Johnson kW Matchbox, Balanced Antenna Tuner

Manufacturer's Specifications	*Measured in ARRL Lab*
Input load range: 50 to 2000 Ω.	See below.
Output SWR range: not specified.	See below.
Frequency coverage: 80, 40, 20, 15 and 10 meters	See below.
Input power: 1000 W dc input.	Not tested.
Size: 12.5 × 17.25 × 11 inches (HWD).	

SWR	*Load (Ω)*		*80 Meters*	*40 Meters*	*20 Meters*	*15 Meters*	*10 Meters*
4:1	12.5	Power Loss %				15	<10
		SWR BW %	No Match	No Match	No Match	1	3
		Imbalance				0	0
2:1	25	Power Loss %			<10	13	<10
		SWR BW %	No Match	No Match	2	2	4
		Imbalance			0	0	0
1:1	50	Power Loss %		<10	<10	<10	<10
		SWR BW %	No Match	2	2	2	4
		Imbalance		0	0	0	0
2:1	100	Power Loss %	<10	<10	<10	<10	<10
		SWR BW %	1	2	2	2	4
		Imbalance	0	0	0	0	0
4:1	200	Power Loss %	<10	<10	11	<10	<10
		SWR BW %	1	2	2	2	4
		Imbalance	0	0	0	0	0
8:1	400	Power Loss %	<10	<10	<10	14	11
		SWR BW %	1	2	2	2	4
		Imbalance	0	0	0	0	0
16:1	800	Power Loss %	10	<10	<10	11	11
		SWR BW %	1	3	2	2	4
		Imbalance	0	0	0	0	0
32:1	1600	Power Loss %	10	<10	<10		
		SWR BW %	1	2	2	No Match	No Match
		Imbalance	0	0	0		
64:1	3200	Power Loss %	10				
		SWR BW %	2	No Match	No Match	No Match	No Match
		Imbalance	0				

Notes

Power losses are expressed as a percentage. A 10% power loss represents less than half (0.46) a dB.
The SWR bandwidth is the percentage of the measurement frequency that can be changed with the SWR staying under 1.5:1.

Review Questions

14-1 What are some of the key issues to keep in mind if selecting a commercial tuner?

14-2 What is the major advantage of an automatic antenna tuner?

14-3 Why might it be important to use a remotely tuned antenna tuner near the antenna feed point?

Chapter 15

Building Your Own Tuners

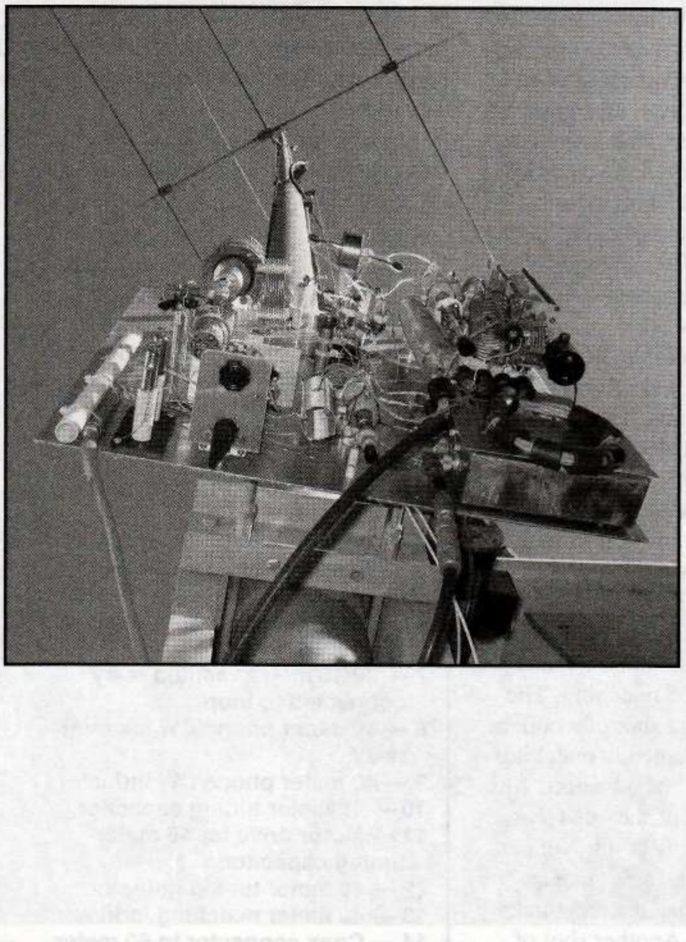

Contents

Why Build a Tuner...15-2

Special Purpose Tuners...15-2

Home Built General Purpose Tuners 15-7

Review Questions...15-12

Why Build a Tuner

A quick look at the schematic diagram of a basic manual antenna tuner (see Chapter 3), not counting metering and switching functions, reveals perhaps the simplest of amateur equipment schematics. A tuner has very few parts, and very few connections. That should make it easy to duplicate — and it does. Unfortunately each of the parts can be quite expensive and perhaps hard to find. On the other hand, if you have — or can easily obtain — the needed parts, there is no reason not to try building one or more tuners.

Special Purpose Tuners

Most of the tuners we have discussed heretofore have been of the general purpose variety. That is, they were designed to match varying loads on multiple frequencies from a single box. In any given amateur station, that may not be what is needed. You may have a need to match your particular antenna system to your particular radio on a particular frequency. That describes a "tuner" that needs no knobs or controls — just a fixed matching network.

The fixed network is more common than we may think. Every Yagi antenna with a gamma, T or hairpin match at the driven element has essentially done exactly that. Each is a matching network located at the antenna to provide a match from the antenna to the transmission line for a low loss run to the radio.

In a similar way, we can provide a network for each of our antennas that provides needed matching. The chapter title figure, a shot of a collection of specialized antenna matching systems at, W6TC, the advanced Amateur Radio station of the late George Badger, illustrates the point. None of those networks looks quite like any of the usual commercial antenna tuners of the last chapter! Another shot of this collection of networks, annotated with descriptions of each tuning element is shown in **Figure 15-1**.

Designing Your Special Purpose Tuner

While the usual circuit design and analysis tools of an electronic engineer can be used to design an impedance matching network, it can be a much simpler proposition for the amateur with a few software and hardware tools. The steps are straightforward:

•*Determine the (complex) impedance that the antenna system will present to the radio or transmission line.*

Determining the impedance can be done in a number of ways. The best approach is probably to measure the impedance at the location that will have the tuner by using an antenna analyzer. Such analyzers include tun-

1 — RG-213 coax cable from shack to input of vacuum relay box behind board (37) to connect to SteppIR Yagi antenna or 30, 40, 60, 80 and 160 meter loop tuners.
2 — RG-213 coax cable to SteppIR antenna.
3 — Coax connector to 30 and 40 meter balun (38).
4 — 30 meter tuning inductor.
5 — 30 meter vacuum variable capacitor.
6 — 30/40 meter vacuum relay connected to loop.
7 — 30/40 meter vacuum relay connected to loop.
8 — 40 meter phone/CW vacuum relay.
9 — 40 meter phone/CW inductor.
10 — 40 meter tuning capacitor.
11 — Motor drive for 40 meter tuning capacitor.
12 — 40 meter tuning inductor.
13 — 40 meter matching inductor.
14 — Coax connector to 60 meter balun.
15 — 60 meter balun.
16 — 60 meter tuning inductor.
17 — 60 meter tuning vacuum variable capacitor.
18 — 60 meter vacuum relays.
19 — 80 meter coax connector to 80 meter balun.
20 — 80 meter balun.
21 — Motor drive for 80 meter vacuum variable capacitor.
22 — Limit switches for 80 meter vacuum variable capacitor.
23 — 80 meter vacuum variable capacitor.
24, 25 — 80 meter tuning and matching inductors.
26 — 80/160 meter vacuum relay.
27 — 160 meter open end vacuum relay.
28, 29 — Connection to loop on tower.
30 — 160 meter receive vacuum relay.
31-33 — 160 meter receive RC network.
34 — Spark gap protector for 160 meter receive RC network.
35, 36 — 25-pair control cable to shack.
37 — Vacuum relay switch box to connect to SteppIR Yagi or loop.
38 — 30 and 40 meter balun.
39, 40 — Feed-through to 80 and 160 meter radial system.
41 — Remote motor control.
42 — Control shaft for 30 meter vacuum variable capacitor (5).
43 — Motor control relay.
44 — 160 meter balun.
45 — Manual control for 60 meter vacuum variable capacitor.
46 — Connector to 160 meter balun.

PHOTO AND ANNOTATION BARRY SHACKLEFORD, W6YE

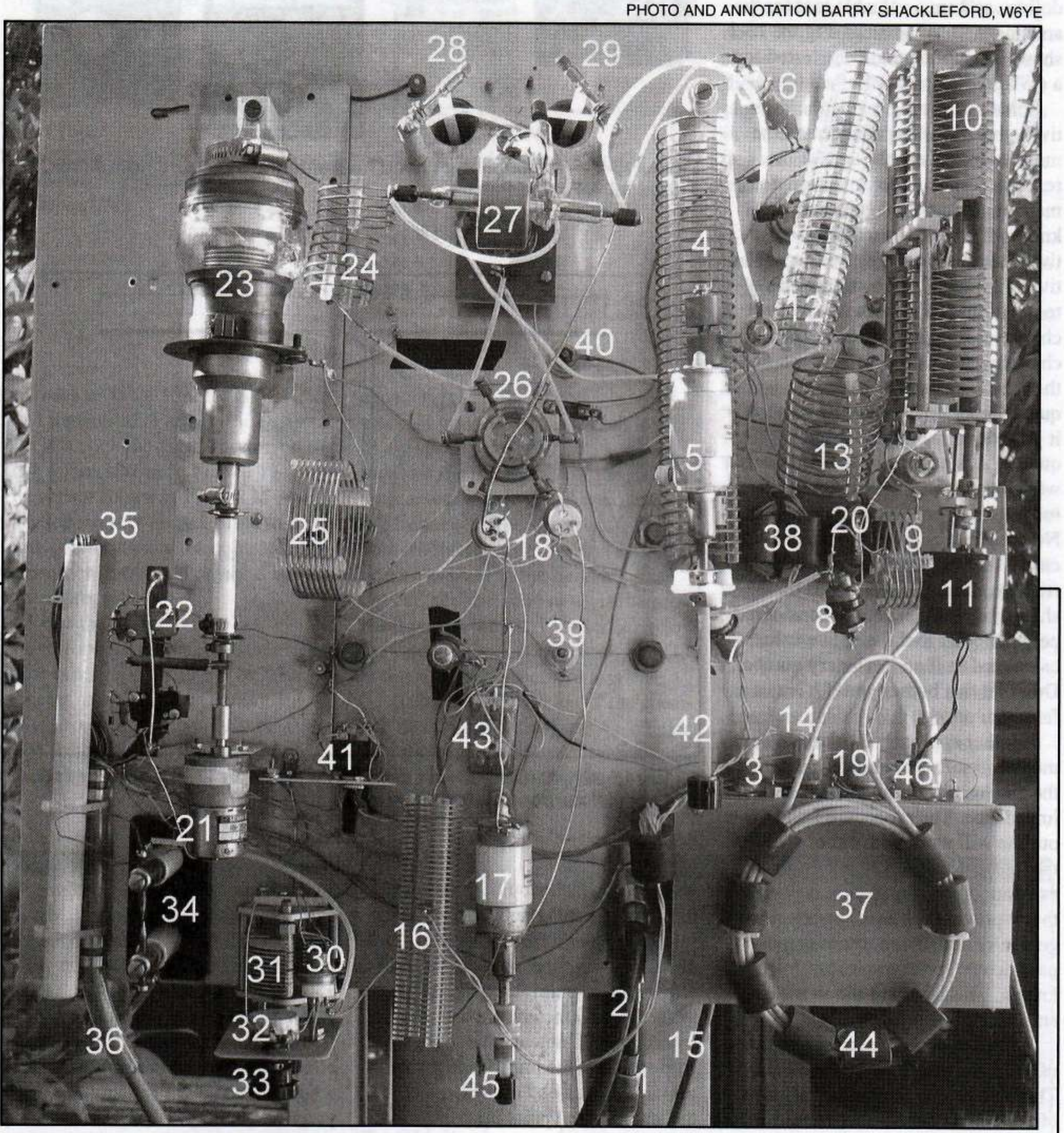

Figure 15-1 — Example of perhaps the extreme in homemade antenna tuning arrangements. This collection of remotely controlled special purpose antenna tuners, all on an outside panel (cover removed for photo), was made to switch between and tune the multiple antenna systems of the late George Badger, W6TC, the designer and builder. The annotations refer to the elements in the list below.

The tuner panel includes separate remote controlled matching networks and connections for each of the HF bands from 30 to 160 meters. It is mounted at the base of the tower. At the top of the tuner panel are two connectors that connect to the two lower ends of the 30, 40, 60, 80 and 160 meter loop that extends from the left connector up to the 5 foot fiberglass boom extension on the left, up and over the 8 foot fiberglass mast extension at the top to the boom extension on the right, then down to the right connector.

able RF generators and displays that indicate actual complex (resistive and reactive) impedance, not just SWR. These are available from a number of manufacturers for a few hundred dollars, and are a valuable tool for the antenna inclined amateur. **Figure 15-2** shows a sampling that were tested for a *QST* product review.[1]

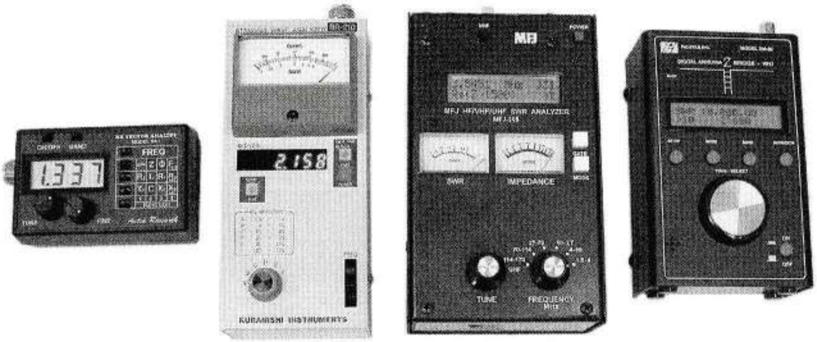

Figure 15-2 — A sampling of hand held antenna analyzers from Autek, Kuranishi, MFJ and Palstar.

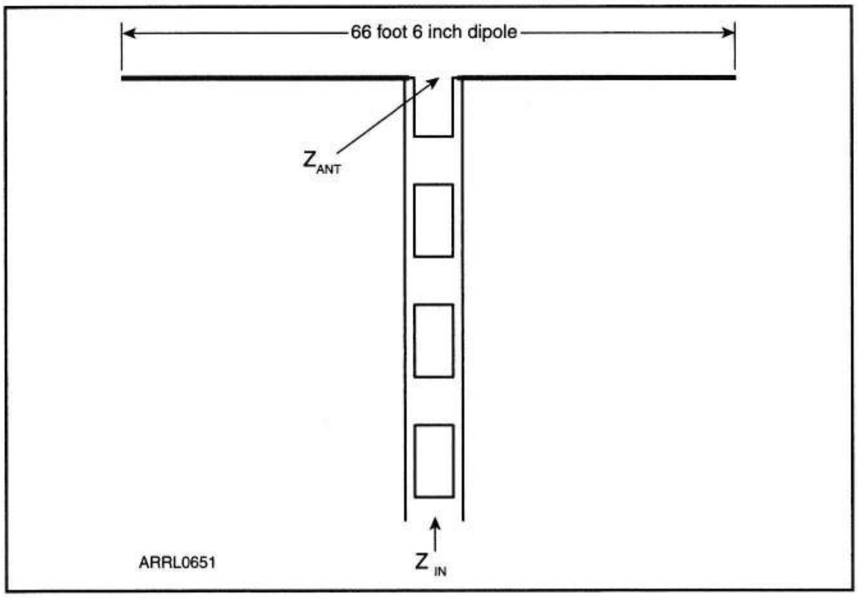

Figure 15-3 — 40 meter dipole used as the example for the determination of a special purpose matching network.

Some analyzers provide the resistive value of the impedance and the magnitude, but not the sign of the reactive part. In order to design a matching network, it is important to know whether the reactance is positive (inductive) or negative (capacitive). Fortunately, it is usually easy to determine this. Just make a slight change in frequency and note the change in the value of reactance. If the reactance goes up as the frequency goes up it is inductive (+). If it goes down with increasing frequency, it is capacitive (–). This only works if the reactance does not go to zero between the two data points. Note that many programs use a lower case italicized letter *j* as part of the reactive value. This indicates that it should be mathematically considered perpendicular to the resistive value, or calculated as if an imaginary number. Don't let this bother you if it is not a familiar concept.

Another approach is to use antenna modeling software to determine the antenna, or the antenna system, impedance. The procedure I will outline will work with either. While modeling will result in a reasonable estimate of the impedance, it is rare that it provides an exact result, since generally not all conditions (ground parameters, for example) are fully known. This is what I did in the example that follows.

A third approach is to rely on manufacturers' or other published data. This also has its limitations in terms of the effect of the actual conditions at your location.

Figure 15-3 provides the basis for a short example of how this process can easily be conducted. The antenna under consideration is a dipole cut for the low end of the 40 meter amateur band. Instead of feeding it directly with coax, which would likely not require a matching network, we have selected to use 70 feet of nominal 450 Ω window line, so it can be fed efficiently on multiple bands. In order to use it on 40 meters, we determine (using *EZNEC* antenna modeling software) that the impedance at the center and each end of the band at the antenna feed point is as shown in **Table 15-1**.[2]

We then use the transmission line analysis program *TLW*, furnished with recent editions of *The ARRL Antenna Book*, to determine the impedance at the bottom of the mismatched transmission line as shown in **Table 15-2**.[3] Note that we could have gone directly to Table 15-2 using many antenna analysis programs that are equipped with transmission line models. My version of *EZNEC* included an early transmission line capability that did not take losses into account, so I used *TLW* instead. Since we need *TLW* for the next step, it wasn't any extra effort. One hint about using *TLW* — when finished don't just close the program window — click on the EXIT button to avoid problems with it starting properly next time.

- *Design a network that will transform the impedance to the desired impedance of the radio or transmission line.*

There are many techniques that could be employed to design a matching network to match any of the im-

[1] Notes appear on page 11.

Table 15-1

Impedance Predicted at the Antenna (Z_{ANT}) as Shown in Figure 15-1.

Frequency(MHz)	*Resistive Part (Ω)*	*Reactive Part (Ω)*	*50 Ω SWR*
7.0	82.3	–34.0	2.0
7.15	88.0	–0.37	1.8
7.3	94.1	+33.1	2.2

Table 15-2

Impedance Predicted at Bottom of the 70 Foot Transmission Line (Z_{IN}) as Shown in Figure 15-1.

Frequency(MHz)	*Resistive Part (Ω)*	*Reactive Part (Ω)*	*450 Ω SWR*
7.0	87.7	+82.9	4.9
7.15	103.3	+148.6	4.6
7.3	236.5	+229.7	4.3

pedances in Table 15-2 to our desired 50 Ω transmitter. Perhaps the easiest is to just click the TUNER button on the *TLW* screen shown in **Figure 15-4**. The screen shown in **Figure 15-5** will then appear, providing a place to input some common parameters including transmitter power and physical component properties.

While the design power is a straightforward parameter, some of the other terms may not be. The inductor and capacitor Q are related to the losses in the components. Typically, the biggest loss in an HF tuner, especially one with an air dielectric capacitor(s), is due to coil wire resistance. If you are buying a coil, it will likely specify the Q. If not, you can measure the impedance of the coil on your antenna analyzer at the operating frequency. The Q is just X_L/R, both available on an antenna analyzer screen.[4] Values in the hundreds are typical, and the good news is the program results are not very sensitive to the value of Q. A 100% change in Q, in this example, results in less than a 0.5% change in the value of the specified inductor and capacitor — smaller than the usual component tolerance, thus we specify a Q of 200 and the resulting design should be fine for most real inductors.

The OUTPUT STRAY CAPACITANCE value represents the capacitance of the wiring and components on the output side to the chassis. All wiring and components in the usual metal chassis or cabinet have capacitance to the metal walls. This is most important as the frequency goes up. Values in the tens of picofarads are common depending largely on component size and placement.

The other choice on this screen is the TUNER NETWORK TYPE. There really isn't too much to say about this — pick the one you like best, or better yet — try them all. I am somewhat partial to the low-pass L-network because it only has two components and provides some attenuation of harmonic signals. It also has only one solution for any load — more later. Pick any one to start, you will likely want to look at them all to find which one has components that are available or easiest to obtain.

If you pick any configuration besides the L-network, it will ask you for a value to use for the capacitor on the output side (towards antenna). It will try to design a tuner for any value you pick. If it can't make a tuner work with your value it will suggest making it larger or smaller. Make note of the power loss and try different values of output capacitor

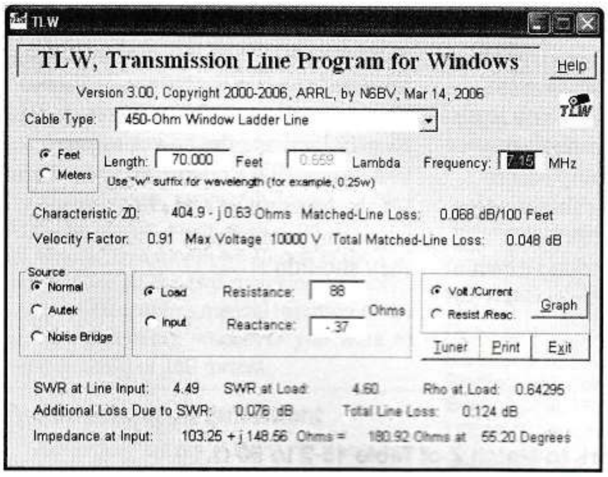
TLW

TLW, Transmission Line Program for Windows

Version 3.00, Copyright 2000-2006, ARRL, by N6BV, Mar 14, 2006

Cable Type: 450-Ohm Window Ladder Line

● Feet ○ Meters

Length: 70.000 Feet 0.559 Lambda Frequency: 7.15 MHz

Use "w" suffix for wavelength (for example, 0.25w)

Characteristic Z0: 404.9 - j 0.63 Ohms Matched-Line Loss: 0.068 dB/100 Feet

Velocity Factor: 0.91 Max Voltage 10000 V Total Matched-Line Loss: 0.048 dB

Source: ● Normal ○ Autek ○ Noise Bridge

● Load ○ Input

Resistance: 88 Ohms

Reactance: -37 Ohms

● Volt./Current ○ Resist./Reac.

SWR at Line Input: 4.49 SWR at Load: 4.60 Rho at Load: 0.64295

Additional Loss Due to SWR: 0.076 dB Total Line Loss: 0.124 dB

Impedance at Input: 103.25 + j 148.56 Ohms = 180.92 Ohms at 55.20 Degrees

Figure 15-4 — Input screen of *TLW* transmission line analysis software used to determine the impedance at the bottom of the mismatched transmission line of Figure 15-1.

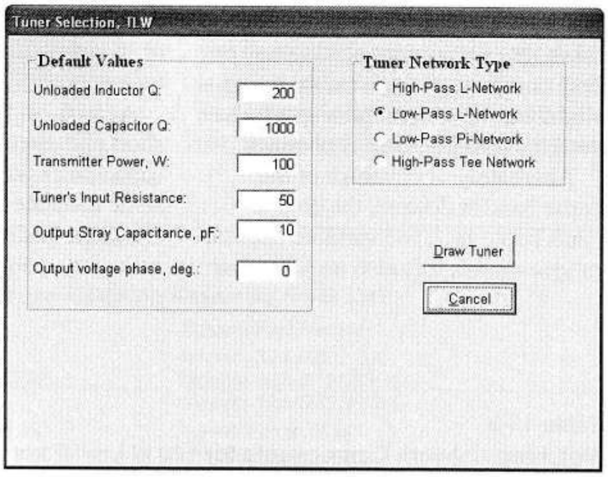
Tuner Selection, TLW

Default Values

Unloaded Inductor Q: 200

Unloaded Capacitor Q: 1000

Transmitter Power, W: 100

Tuner's Input Resistance: 50

Output Stray Capacitance, pF: 10

Output voltage phase, deg.: 0

Tuner Network Type

○ High-Pass L-Network

● Low-Pass L-Network

○ Low-Pass Pi-Network

○ High-Pass Tee Network

Figure 15-5 — Antenna tuner Input screen of *TLW* software. Note the tuner configuration selected, LOW-PASS L-NETWORK, the power level, 100 W, that determines component ratings as well as the other parameters as described in the text.

until you achieve a design that has an acceptable amount of loss and reasonably sized components. This illustrates the fact that the three element tuners — Pi- and T-network — have more than one set of adjustments that will provide a satisfactory match to the radio. Unfortunately, while they will all look good to the radio, some will have more loss than others.

Push the DRAW TUNER button. The screen shown in **Figure 15-6** should appear providing the design details for the low-pass L-network tuner I selected. In addition to the component values needed to provide the match, the table above the schematic provides information on component stress levels and other key data. The primary selection criterion can be found by looking at the resulting tuner designs (next paragraph) in each configuration and seeing which ones have component values that are reasonable, and even better, that you have on hand. A look at the resulting efficiency, expressed as POWER INTO LOAD, may also help you select a preferred topology.

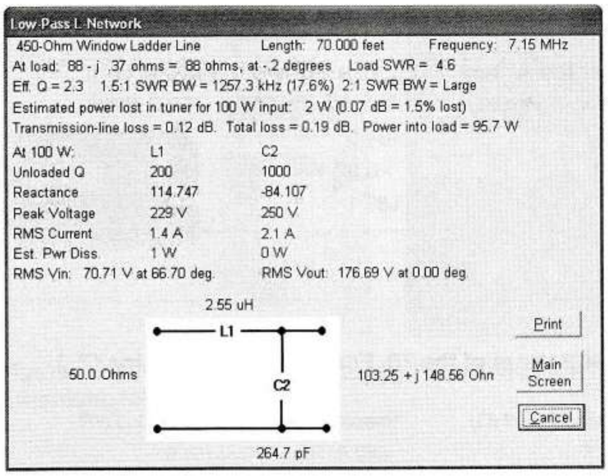

```
450-Ohm Window Ladder Line          Length:  70.000 feet        Frequency:  7.15 MHz
At load:  88 - j  .37 ohms =  88 ohms, at -.2 degrees    Load SWR =  4.6
Eff. Q = 2.3    1.5:1 SWR BW = 1257.3 kHz (17.6%)  2:1 SWR BW = Large
Estimated power lost in tuner for 100 W input:   2 W (0.07 dB = 1.5% lost)
Transmission-line loss = 0.12 dB.  Total loss = 0.19 dB.  Power into load = 95.7 W
At 100 W:           L1               C2
Unloaded Q          200              1000
Reactance           114.747          -84.107
Peak Voltage        229 V            250 V
RMS Current         1.4 A            2.1 A
Est. Pwr Diss.      1 W              0 W
RMS Vin:   70.71 V at 66.70 deg.     RMS Vout:  176.69 V at 0.00 deg.
```

Figure 15-6 — Antenna tuner output (design) screen of *TLW* software. Note the tuner schematic with parts values shown. The data and table above provide additional important information.

Table 15-3 shows the resulting design for the middle and both edges of the 40 meter band. This design can be used as a starting point for a single frequency, single load, matching network. I say starting point because even with careful analysis there are a few variables that are not known exactly, including actual component values — typical tolerances are ±10% — for example. This is often taken into account by making one or both components variable, or trying slightly different component values until the desired result is obtained.

Alternately, if coverage of the entire band is desired, the components could be made variable over the ranges shown. It is also possible that the values for the middle of the band will provide a satisfactory match over at least a portion of a band without requiring change. This can be determined by trial and error, or simulated in an antenna analysis program by using modeled lumped "loads" at the bottom of the transmission line.

- *Build and test the network.*

As noted, the hardest part is often finding the needed components, especially those needed for high power networks. Hamfest flea markets are often a good source for such items, although they often come with a level of uncertainty regarding ancestry and component value. Again, the trusty antenna analyzer can be used to measure the reactance. Fixed or tapped inductors can be easily wound using plastic rod or tubing as a form. *The ARRL Handbook* provides a source of formulae for winding the desired inductance.

As with all RF projects, keep leads short and direct. To minimize stray capacitance, keep the components some distance from the cabinet edges. Use high quality connectors and make sure you have provided direct connectivity for all needed interconnections.

The best way to test the network is to start with the antenna analyzer. With the network connected to the proper place on the transmission line, the impedance should read close to 50 Ω at the design frequency. Now take *swept frequency* type data by measuring the impedance and SWR every 25 or 50 kHz. By plotting these on a graph, you should have a good idea of what the network will do and whether or not it will work across the range. If it isn't quite right, make a small change in one value and repeat the sweep. You should see an indication of the extent and direction of the change. Since there are only two or three components, you should quickly converge on your result.

Next hook it to your radio tuned to an unused frequency and note the transceiver's SWR reading. It should be close to that of the analyzer. Slowly increase the power to just below the level you designed to. Turn off the transmitter and check to see if any components have become hot — they shouldn't!

Table 15-3
Matching Network Components for 100 W Low-Pass L-Network to Match Z of Table 15-2 to 50 Ω.

Frequency(MHz)	*Inductance (μH)*	*Inductor Current (A_P)*	*Capacitance (pF)*	*Capacitor Voltage (V_P)*	*Efficiency (%)*
7.0	1.72	1.4	339.4	181	96.1
7.15	2.55	1.4	264.7	250	95.7
7.3	3.28	1.4	205.2	317	95.4

160 and 80 Meter Matching Network for Your 43 Foot Vertical — *Part 1*

The popular 43 foot vertical works best with matching at the antenna base — here's a way to do it for our two MF bands.

Phil Salas, AD5X

I've recently made the move to a 43 foot vertical for much of my lower frequency operations. This length antenna offers higher radiation resistance than shorter loaded monopoles. Increased radiation resistance improves efficiency by reducing the effects of ground losses, especially when you have an electrically short antenna — a characteristic of even a 43 foot antenna on 160 and 80 meters. If fed with a 1:4 unbalanced to unbalanced transformer (unun), a 43 foot antenna has a reasonable compromise SWR on 60 through 10 meters, which means that cable and unun losses are pretty much negligible on these bands.

This antenna is really not a good performer on 160 meters, and to a lesser extent on 80 meters, unless you provide matching right at the antenna base. This is due to the high capacitive reactance and still relatively low radiation resistance of a 43 foot antenna on 160 and 80 meters. This makes the mismatch so bad that it is almost impossible to efficiently match from your shack. If you can match the antenna system from your shack, you will throw away a lot of power in your coax and unun due to the very bad mismatch at the antenna.

I thus started experimenting with matching networks and wound up with two external impedance matching devices designed to significantly reduce SWR related coax losses and unun mismatch losses, and to help the inside tuner match on 160 and 80 meters. This month we'll look at the simpler version of the two matching networks. This version requires manual insertion of the matching network whenever you want to operate on 80 or 160 meters.

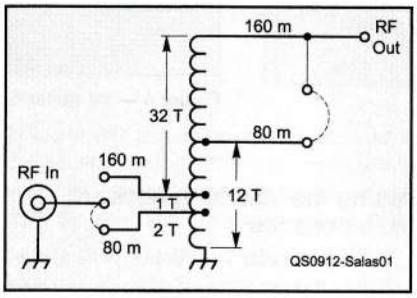

Figure 1 — Schematic diagram of the two band matching unit. 80 meter connections shown with dashed lines.

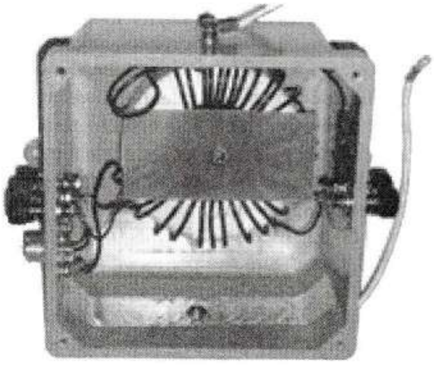

Figure 2 — T400A-2 160 and 80 meter matching unit.

The Matching Requirement

According to my AIM4170C antenna analyzer, my 43 foot vertical antenna has a capacitive reactance of about 580 Ω on 160 meters. This will vary based on the particular construction of your 43 foot vertical, its proximity to other objects, and other factors. The reactance will almost certainly be in the 550 to 650 Ω range. This amount of capacitive reactance needs approximately 50 μH of inductance in order to resonate the antenna. On 80 meters, approximately 9 μH is needed to resonate the antenna. A 50 μH high Q inductor is going to be large. For this first solution, I elected to go with a toroidal inductor in order to keep the matching unit as compact as possible.

Toroid Inductor Matching Solution

This compact design will handle the full legal limit on 80 and 160 meters for low duty cycle SSB and CW modes. The inductor consists of 35 turns of #14 AWG solid copper insulated house wiring wound on a T400A-2 toroid core. The antenna feed point is tapped two turns from the ground end for 80 meters, and three turns from the ground end for 160 meters. You should start with 38 turns total on your toroid, but then remove turns as necessary to get the network to resonate where you want it in the 160 meter band (more on this later).

I mounted the toroid assembly in a 6 × 6 × 4 inch NEMA enclosure using a 2½ inch long #10 machine screw and associated hardware along with a 2 × 4 inch piece of unplated fiberglass PC board material. Before you mount the toroid, prepare

Table 1
160 Meter Toroid Impedance Matching Assembly Parts List

Description	*Source/Part Number**
Banana plug (4 required)	Mouser 174-R802-EX
Binding post, black (4 required)	Mouser 164-R126B-EX
Binding post, red	Mouser 164-R126R-EX
NEMA Enclosure, 6 × 6 × 4 inch	Lowes/Home Depot
Glass cloth tape, 3M #27	ACE Hardware
SO-239 connector	Mouser 601-25-7350
Toroid, T400A-2 powdered iron	Amidon T400A-2

*Amidon parts are available from **www.amidoncorp.com** and Mouser parts from **www.mouser.com**.

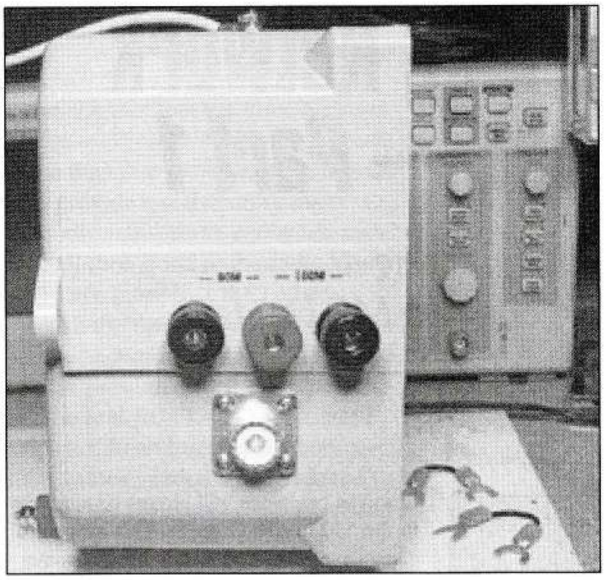

Figure 3 — 160 and 80 meter input tap points.

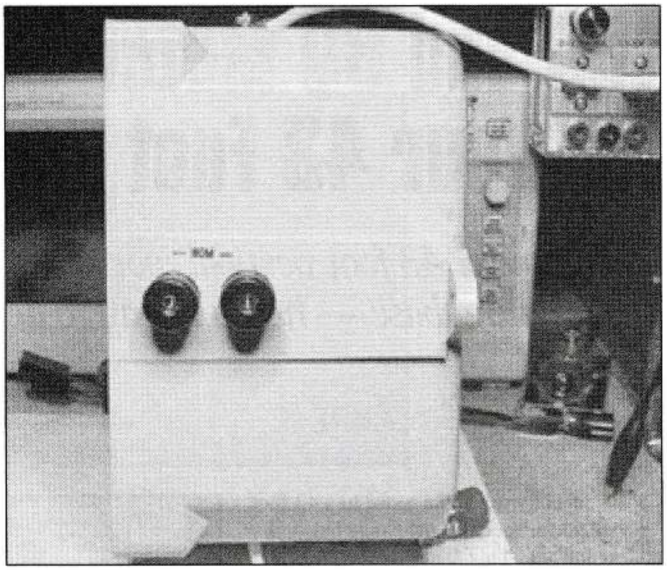

Figure 4 — 80 meter coil shorting jacks.

it by scraping the insulation off the outside second, third and 11th through 13th wire turns. Because of the high voltages possible at legal limit power levels, especially on 160 meters, wrap the toroid with two layers of 3M #27 glass cloth electrical tape for added insulation between the #14 AWG wires and the toroid core.

Figure 1 is the schematic of the matching assembly, Figure 2 shows the internal details of the assembly, and Table 1 lists the parts necessary. To select between 160 and 80 meter operation, I used external jumpers across binding posts as shown in Figures 3 and 4. Stainless steel #8 hardware (screws, washers, lockwashers and nuts) are used for the matching unit ground and RF output terminals. Internal to the matching unit, I used a 2 inch wide strip of aluminum duct repair tape as a good low impedance ground between the UHF connector and the ground screw on the bottom of the case. Finally, I used #14 AWG stranded insulated wire for all internal connections.

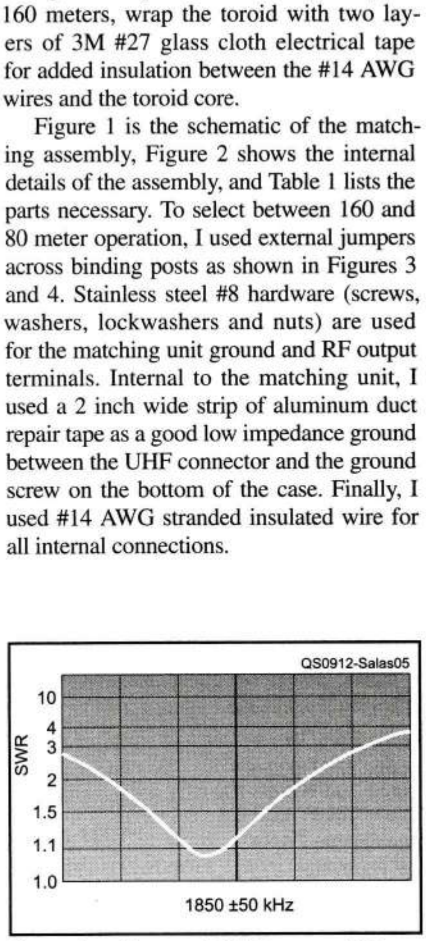

Figure 5 — Measured SWR across 160 meters indicating a 2:1 SWR bandwidth of approximately 50 kHz.

Tuning the Matching Network to Resonance

Your particular installation will almost certainly require you to change the resonant frequency of the matching network. This is because there will be some variations of the antenna impedance based on your particular antenna physical construction, proximity to other objects and final length, as well as your desired operating frequency range. The design is such that the overall inductance is too large for 160 meters, so the network should resonate at or below the lower band edge. Therefore, you will need to remove one or more of the upper inductor turns in order to resonate the network for the desired frequency on 160 meters.

To do this, first solder wires from the second and third turn tap points on the coil to the two outer binding posts by the SO-239 connector. The input tap points tend to be fairly noncritical and will probably be the same for all installations. Now solder a short wire from the SO-239 center pin to the middle binding post. Next, externally jumper the middle binding post to the 160 meter binding post (third turn).

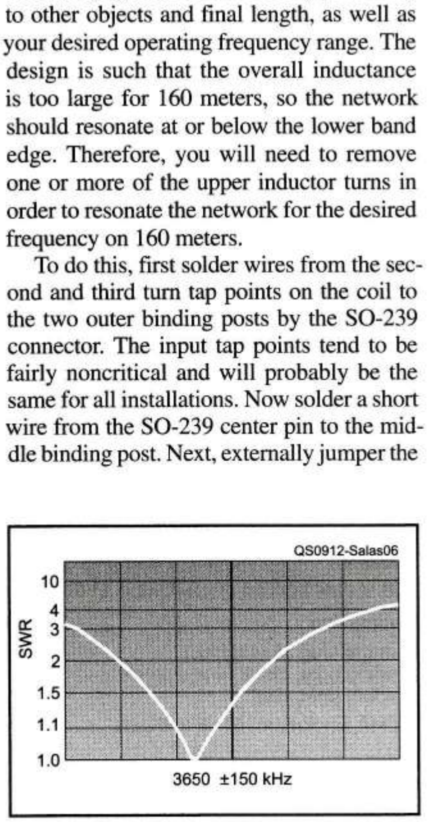

Figure 6 — SWR across 80 meters indicating a measured 2:1 SWR bandwidth of about 150 kHz.

Connect the matching assembly to the base of your 43 foot vertical and see where the minimum SWR point is on 160 meters using your antenna analyzer. If the resonant frequency is too low, remove a turn of wire and see where the minimum SWR point is again. You'll see about a 50 kHz upward move in frequency per turn of wire removed. When you have your desired resonant point on 160 meters, it is time to move to 80 meters. Externally jumper the input tap middle binding post to the 80 meter binding post (second turn), and use a clip lead to short from the top of the coil to turn number 12 and see where your minimum SWR frequency occurs. Move the tap point up or down until your resonance point (lowest SWR) is where you want it. Solder a wire from this tap point to one of the binding posts. Solder another wire from the top of the coil to another binding post. Now you will be able to externally jumper these binding posts to select either 160 or 80 meters.

My final test results for 160 and 80 meters are shown in Figures 5 and 6 as measured with my RigExperts AA-200 antenna analyzer connected directly to the matching network input at the base of the antenna. I'm a CW operator, so I favor resonance in the lower part of these bands, but you can adjust for your favorite portion of each band. The 2:1 SWR bandwidth on 160 meters is about 50 kHz, and about 150 kHz on 80 meters. Even a 3:1 SWR on these bands results in negligible SWR related cable losses for any reasonable length cable and is easily matched with my MFJ-998 in-shack tuner or most transceivers' internal tuners.

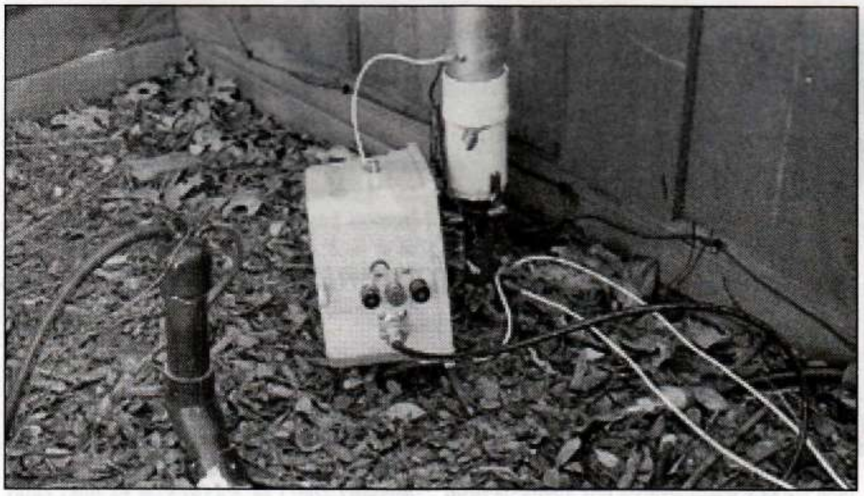

Figure 7 — Matching unit at the base of the author's 43 foot antenna. In this view, it is strapped for 80 meters.

Operation

Using the matching unit is simple. Just disconnect your normal unun when you want to operate on 160 or 80 meters and connect this matching unit to the base of the antenna. Select either 160 or 80 meters with the external straps. You can connect both the unun and this matching unit to the antenna at the same time, and just leave off the ground wire from the unit that is not used. The matching unit connected to the base of my 43 foot vertical is shown in Figure 7.

Conclusion

The matching network discussed in this article will permit very effective operation of your 43 foot vertical on the 160 and 80 meter bands. In Part 2, we'll look at a remotely switchable 160, 80 and 60 through 10 meter base matching unit. It is more complex, but it is also more convenient. Until then, see you on top band!

Amateur Extra class operator and ARRL Life Member Phil Salas, AD5X, was first licensed as WN3BCQ in 1964. Ham radio became the reason he subsequently pursued a career in electrical engineering. Phil earned BSEE and MSEE degrees from Virginia Tech and SMU, respectively, and worked in new product development for the next 33 years. Phil is now retired and spends his days split between ham radio related projects and enjoying time with his wife, Debbie, N5UPT. You can reach Phil at 1517 Creekside Dr, Richardson, TX 75081 or at **ad5x@arrl.net**.

160 and 80 Meter Matching Network for Your 43 Foot Vertical — *Part 2*

This dual band matching section design provides for remote band changing.

Phil Salas, AD5X

In Part 1, I described a simple 160 and 80 meter base matching assembly for a 43 foot antenna. While that unit is very effective and inexpensive to build, it is also a little inconvenient in that you must connect it when it is needed, and you must also manually enable 160 or 80 meter operation using straps. This month, we'll look at a more versatile matching assembly that is completely remote-controllable for operation on all bands. Again, this base matching network will significantly eliminate SWR-related coax losses and unun mismatch losses, and assist inside tuners in providing a match on 160 and 80 meters.

The All Band Matching Solution

As discussed in Part 1,[1] a 43 foot vertical antenna requires approximately 50 μH of inductance to resonate the antenna. On 80 meters, approximately 9 μH is needed for resonance. For the matching solution shown in Figure 1, I used an MFJ 404-0669 air wound coil and two Array Solutions RF-30 relays for up to full legal limit matching on all bands from 160 to 10 meters. A 2.1 × 5.5 mm dc power jack located below the SO-239 connector provides for relay control voltage inputs of 0 V, +12 or –12 V dc. The control input MOVs and bypass capacitors are mounted on a 6 terminal strip inside the matching assembly.

The matching unit operates as follows: With no control voltage applied, the inductor is disconnected and the unun is connected across the antenna feed point thereby preserving the original compromise SWR on 60-10 meters.

When +12 V is applied, the inductor is connected across the antenna feed point and inductor turns are shorted to resonate the antenna on 80 meters.

When the voltage is reversed, the short across the coil is removed and the matching section resonates the antenna on 160 meters. In both the 160 and 80 meter cases, the unun secondary taps into the inductor at the 200 Ω point, providing a proper match to the unun on these bands. This keeps the unun secondary voltage reasonable, and the feed line and unun losses very low as well.

I built the matching unit into an 8 × 8 × 4 inch NEMA weatherproof box available from home supply stores (see Figure 2). This box is too small to fit the entire inductor length needed, so I split the inductor into two pieces as can be seen in Figure 1. The long inductor section consists of 61 turns, and the short inductor section consists of 12 turns. While I preferred the more compact size of this box, you may wish to use the larger 12 × 12 × 4 inch NEMA box that can contain the full uncut inductor length needed.

All internal wiring uses insulated #14 AWG stranded wire. I attached the wires to the coil tap points using the MFJ coil clips called out in the parts list. You can solder the wires directly to the coil, but this is difficult due to the #12 AWG size of the coil wire and the spacing of the turns. The coil is mounted with #8 stainless steel screws, washers, lock washers and nuts also used for the ground and antenna feed terminals. Figure 1 shows the final matching unit with all components mounted. Note the terminal strip with the MOVs and bypass capacitors.

Since the antenna is unbalanced, a current balun or voltage unun is typically used. A voltage unun should be wired as shown in Figure 2.

Relay Connections

Some comments are necessary regarding the relay connections. See the sidebar on the *QST* binaries Web site for a discussion of how high the peak voltage across the full coil can be on 160 meters. As you can see, I connected relay contacts in series to increase the overall breakdown voltage. Also, the tap points on the inductor provide additional voltage above ground, which helps with the overall breakdown voltage.

The main concern is with the contact-to-coil 5.3 kV peak breakdown rating. When I examined the relays, I observed that the outer SPDT contacts are connected via insulated internal wires that are well separated from the coil by 0.1 to 0.2 inches. The problem is with the center SPDT relay common wire, which is in contact with the coil. While the coil and common wire are both insulated, there is no air-gap separation between them. This common wire-to-coil contact is obviously what

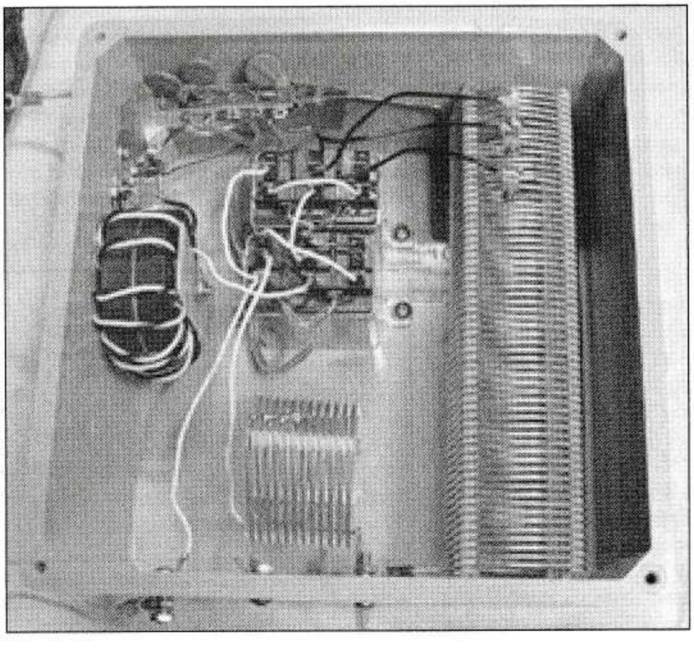

Figure 1 — All parts mounted in the NEMA enclosure.

---

[1] Notes appear on page 35.

determines the breakdown voltage rating. The way I got around this problem was to use the center SPDT relay contacts for the lowest potential interfaces as indicated in the schematic (Figure 2).

For the switched control voltage input, I used a 12 V, 1 A wall wart to keep the +12 V control voltages separate and isolated from the regular station voltage. This is to eliminate any possibility of shorting the main power supply when the control voltage polarity is flipped.

I mounted the switch in a small plastic box that I attached to my transceiver support shelf. The unlabeled center-off position is for 60 through 10 meter operation. The separate switch you see in the photo is for controlling other outdoor accessories, as I have two voltage feeds going out to my antenna location.

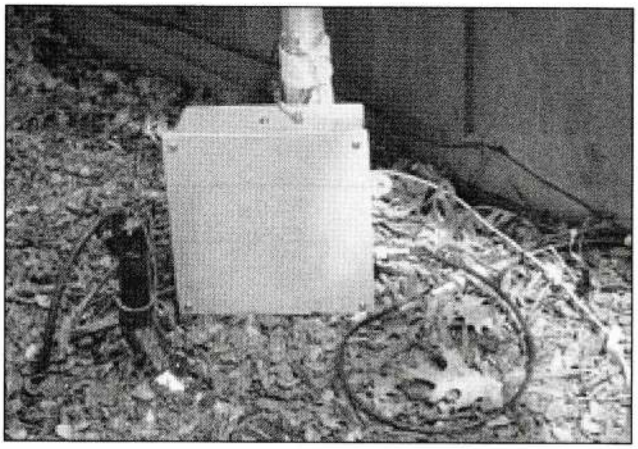

Figure 3 — The completed all band remotely controlled switching and impedance matching network mounted at the antenna base.

Matching Network Resonance

As with the previous matching network, you will almost certainly need to adjust the resonant frequency of the matching network due to variations in particular antenna physical construction, proximity to other objects and final length, as well as your desired operating frequency range. The overall starting inductance is too large for 160 meters, so the network will resonate at or below the lower band edge. Therefore, you can simply short one or more of the upper (short coil) inductor turns in order to tune the network higher to your desired 160 meter frequency.

To do this, first leave off the wire leads that attach between the relay contacts and the tap points on the coil. Build up short jumpers using the test-clips and micro-clips (perfect for the coil turn taps) called out in the parts list, and attach these between the relay contacts and coil, using the suggested tap points shown on the schematic.

Connect the matching assembly to the base of your 43 foot vertical (see Figure 3). Then enable 160 meter operation by applying –12 V dc, and jumper turns on the 12 turn inductor with a short clip lead until you find your desired 160 meter resonance point. Next, move the 160 meter relay tap point until you get minimum SWR. Now permanently short the turns on the short coil by soldering a piece of #16 AWG buss wire across these turns.

As you can see from my photos, I needed to short six turns on this coil. Next enable 80 meter operation (apply +12 V to the relays assembly) and select the coil shorting point for your desired resonant frequency and the tap point for best SWR. Finally remove the test clips leads, attach the coil clips, and solder wires between the coil clips and relay.

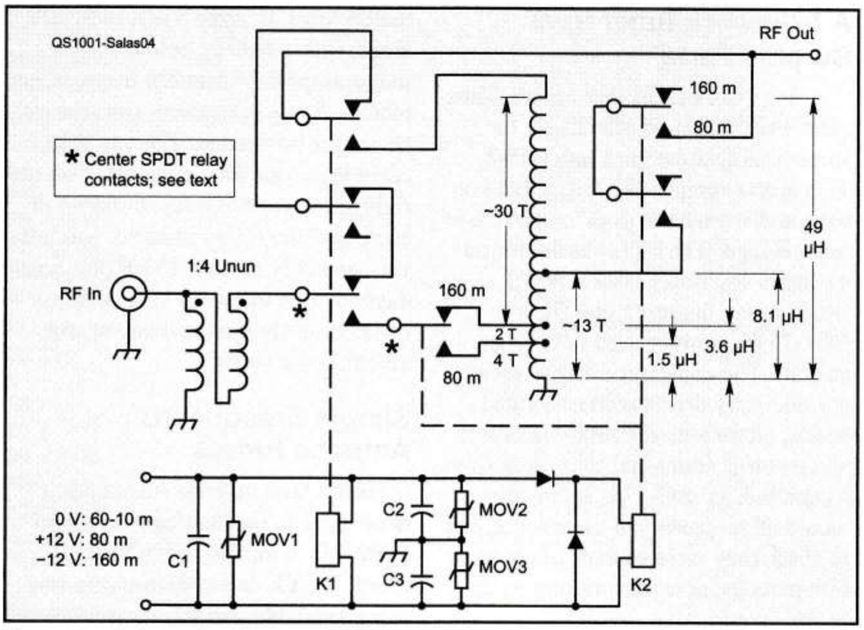

Figure 2 — Schematic of the remotely controlled matching network. The parts list is on the QST binaries version.[2]

The 2:1 SWR bandwidth on 160 meters is about 50 kHz, and about 150 kHz on 80 meters. Even a 4:1 SWR on these bands results in negligible SWR related cable and unun loss, and is easily matched with my MFJ-998, an in-shack tuner.

Operation

Operation of this matching unit couldn't be simpler. When no control voltage is applied, the antenna functions as it always has on 60 through 10 meters. For 80 meter operation, apply +12 V dc, and for 160 meter operation apply –12 V dc. The matching unit connected to the base of my 43 foot vertical is shown in Figure 10.

Conclusion

The matching network discussed in this article will permit very effective operation of your 43 foot vertical on all bands from 160 through 10 meters. Although on 10, 12 and 15 meters the elevation angle of the peak is higher than optimum for DX, it will still allow contacts. Feel free to experiment a little. You might prefer to use the toroid from Part 1 in this design instead of the air-core inductor. If you are not running more than about 500 W, the less expensive RF-10 relays are all you will need. Have fun, and I'll see you on the low bands.

Notes

[1] P. Salas, AD5X, "160 Meter Matching Network for Your 43 Foot Vertical — *Part 1*," *QST*, Dec 2009, pp 30-32.

[2] **www.arrl.org/files/qst-binaries**.

See Part 1 for Phil's bio. You can reach Phil at 1517 Creekside Dr, Richardson, TX 75081 or at **ad5x@arrl.net**.

Home Built General Purpose Tuners

There are a number of approaches to the design and construction of general purpose antenna tuners. Perhaps the easiest is to replicate existing successful designs, although finding the exact parts called for is sometimes a problem. Fortunately, the exact values are not very critical. It is also possible, if somewhat tedious, to go through the *TLW* based procedure described above at the extremes of frequency and mismatch to determine the range of values needed and then construct a tuner with variable and/or switched elements, often in combination, that will cover the needed range.

I have done this for the most popular configurations, see **Figure 15-7**, the high-pass T-network, Pi-network and low pass L-network with the results shown in **Tables 15-4** through **15-6**. I have provided the values for matching loads with a 10:1 SWR, the usual spec for a wide range tuner. In addition to the resistive loads at 5 and 500 Ω, I have made runs for moderate reactive loads of 25 ± *j*25 Ω and extreme reactive loads of 250 ± *j*250 Ω, all with approximately a 10:1 SWR.

As noted in Chapter 4, the Pi- and T-network tuners have multiple solutions for any load, while the L-network has a single solution. It is worth stressing this in case you use one, whether a commercial unit or a home-built one. Look again at Table 4-1, in Chapter 4, to get the idea.

Such "HF" tuners sometimes include the MF 160 meter amateur band as well as the 80 through 10 meter HF bands. I have provided data for 160, 80 and above the top of 10 meters so you can decide which way you want to go. Note that the addition of 160 meters makes a big difference in the component values required, one reason it often is not included. In addition, in some cases the larger inductor and capacitor(s) may have too high a minimum value to allow proper operation on 10 meters.

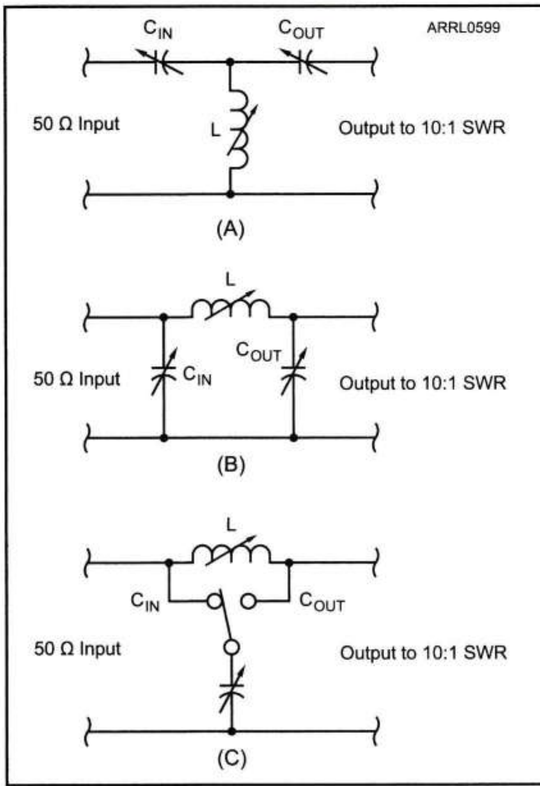

Figure 15-7 — Schematic diagrams of a high-pass T-network (A), Pi-Network (B) and a low-pass L-network (C) used as design examples in the text and Tables 15-4 to 15-6 on the following pages.

A T-Network Tuner from Surplus Parts

A T-network tuner that I made some years back and used effectively for some time is shown in **Figure 15-8**. This is an example of a "use what you have and see what it does" approach to tuner design. The unit is built around a chassis and panel from a WW2 surplus plug in tuning unit from a BC-375 transmitter used in bomber aircraft. The capacitor controls nice 4:1 planetary drives with calibrated scales, along with the rotary indicator and its turns count dial also came from a cannibalized BC-375. Today they would all be preserved as antiques, but in 1965, they were sources of cheap or free parts for new transmitting equipment.

Figure 15-9 shows the inside of the tuner, including a home built "Monimatch Mark II" type SWR indicator, along with a battery holder for the meter amplifier.[5] Modern transceivers include SWR indication, so these may no longer be needed. The variable capacitors, one above the other on the right side, were of a few hundred pF each, and the rotary inductor (on left) was probably around 15 μH. In those days, no one knew for sure — but it successfully matched any of my antennas for years.

Simple Breadboard Antenna Tuners

Home-built antenna tuners don't need to be as mechanically complex as my old T-match. Barry Shackleford, W6YE, has constructed a few using available variable capacitors and inductors fabricated in his workshop. While these won't handle full power,

Table 15-4

Component Requirements for High-Pass (Shunt L) T-Network Antenna Tuners at 10:1 SWR

Frequency/Z (Ω)	*Capacitor*		*Inductor (µH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
1.8 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	1136	3000	2.1	180	710	96
500	548	500	13.9	323	1250	98
25 + *j*100	343	300	10.3	790	3070	92
25 – *j*100	170	300	20	1040	4030	86
250 + *j*250	308	200	10.5	380	1470	98
250 – *j*250	337	300	16.9	525	2030	96

Frequency/Z (Ω)	*Capacitor*		*Inductor (µH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
3.5 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	563	1500	1.1	190	720	96
500	265	200	7.3	343	1330	98
25 + *j*100	275	200	3.5	613	2373	95
25 – *j*100	104	200	8.6	880	3403	88
250 + *j*250	333	100	5.6	381	1475	98
250 – *j*250	136	100	10.8	670	2600	94

Frequency/Z (Ω)	*Capacitor*		*Inductor (µH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
30 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	79	200	0.12	160	640	96
500	29	50	0.77	370	1470	97
25 + *j*100	91	30	0.24	400	1560	98
25 – *j*100	24	100	0.46	440	1710	93
250 + *j*250	36	100	0.9	300	1150	98
250 – *j*250	29	100	0.6	360	1410	97

Table 15-5

Component Requirements for Low-Pass (Series L) L-Network Antenna Tuners at 10:1 SWR

Frequency/Z (Ω)	*Capacitor*		*Inductor (µH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
1.8 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	5254	n/a	1.34	100	390	98
500	n/a	536	13.5	310	1210	98
25 + *j*100	n/a	1408	12	290	1120	98
25 – *j*100	1760	n/a	11	100	390	97
250 + *j*250	n/a	713	13	310	1210	98
250 – *j*250	n/a	359	13	310	1210	98

Frequency/Z (Ω)	*Capacitor*		*Inductor (µH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
3.5 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	2700	n/a	0.69	100	400	98
500	n/a	275	6.8	310	1200	98
25 + *j*100	n/a	720	6.2	290	1120	98
25 – *j*100	926	n/a	5.6	100	390	97
250 + *j*250	n/a	367	6.8	310	1210	98
250 – *j*250	n/a	184	6.8	310	1210	98

Frequency/Z (Ω)	*Capacitor*		*Inductor (µH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
30 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	315	n/a	0.08	100	390	98
500	n/a	32	0.79	310	1210	98
25 + *j*100	n/a	85	0.72	290	1120	98
25 – *j*100	140	n/a	0.58	100	390	97
250 + *j*250	n/a	43	0.79	310	1210	98
250 – *j*250	n/a	22	0.79	310	1210	98

Table 15-6

Component Requirements for Low-Pass Pi-Network Antenna Tuners at 10:1 SWR

Frequency/Z (Ω)	*Capacitor*		*Inductor (μH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
1.8 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	5256	500	1.4	100	390	98
500	2602	1000	9.6	310	1200	96
25 + *j*100	966	1500	12.5	280	1110	97
25 – *j*100	3410	500	7.5	280	1100	96
250 + *j*250	1931	1000	11.3	310	1210	97
250 – *j*250	1284	500	12.9	310	1210	97

Frequency/Z (Ω)	*Capacitor*		*Inductor (μH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
3.5 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	2706	500	0.7	100	390	98
500	1287	500	5.1	310	1200	96
25 + *j*100	643	800	6.2	280	1110	97
25 – *j*100	1886	300	3.7	280	1430	95
250 + *j*250	934	500	6.0	310	1200	97
250 – *j*250	859	300	6.2	310	1200	97

Frequency/Z (Ω)	*Capacitor*		*Inductor (μH)*	*Capacitor Voltage (V_P)*		*Efficiency (%)*
30 MHz	*Input (pF)*	*Output (pF)*		*100 W*	*1500 W*	
5	321	200	0.08	100	390	98
500	118	50	0.7	310	1200	97
25 + *j*100	103	100	0.7	290	1100	97
25 – *j*100	205	30	0.5	285	1100	96
250 + *j*250	71	50	0.8	310	1200	97
250 – *j*250	77	30	0.8	310	1200	97

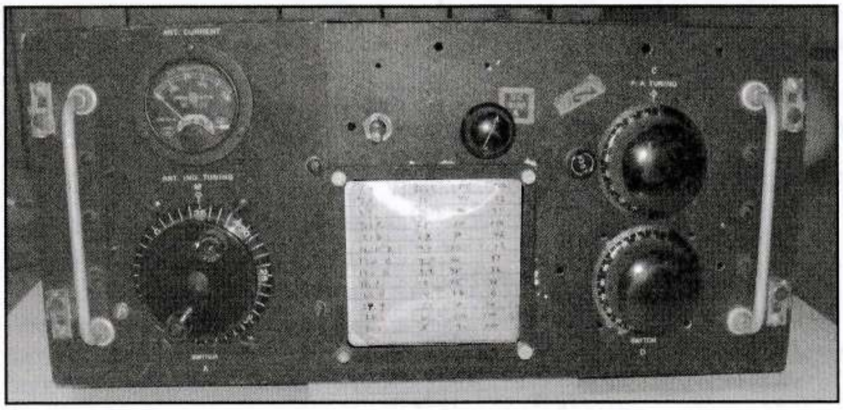

Figure 15-8 — Front view of an early homemade antenna tuner. The unit is completely built of surplus parts. Note the handy tuning chart that was part of the BC-375 transmitter tuning drawer.

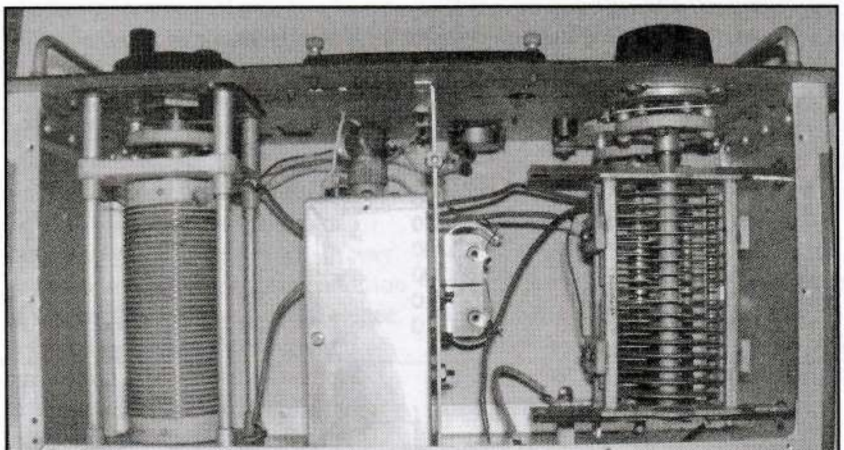

Figure 15-9 — Inside view of the home brew T-network tuner. The SWR meter (in 2 × 2 × 5 inch box in left partition) and associated circuitry would likely not be needed to use with current transceivers.

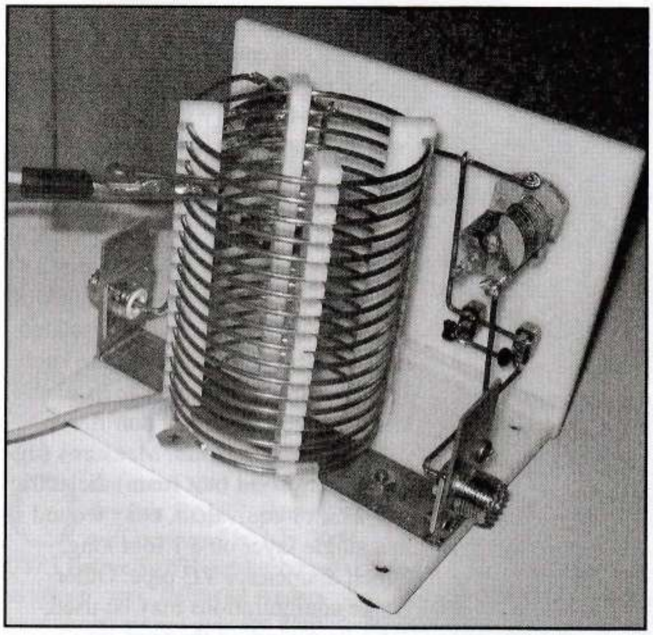

Figure 15-10 — Homebrew breadboard T-network tuner made by Barry Shackleford, W6YE. This tuner uses smaller variable capacitors with a voltage rating appropriate for 100 W use. The inductor was homemade, as described in the text. Note the inductance can be changed by moving (with transmitter off) the alligator clip to a different turn.

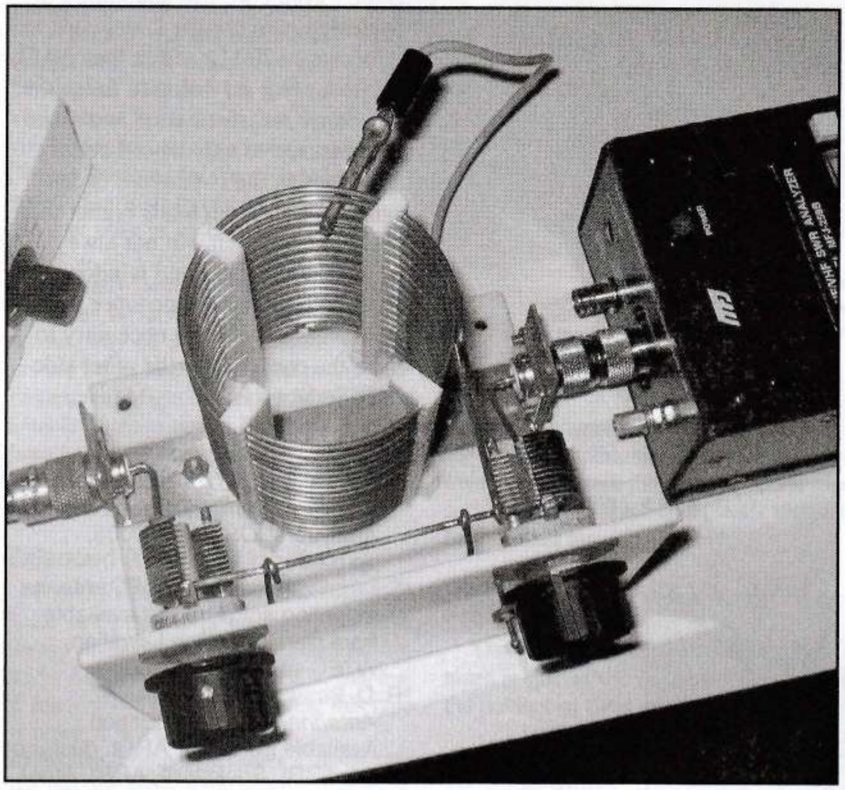

Figure 15-11 — Another view of Shackleford's T-network tuner.

they should be fine at the 100 W level. **Figures 15-10** and **15-11** show a T-network tuner that Barry made. Note the simplicity of construction. The only complication may be the homemade inductors. Barry used acrylic stock, cut and notched with his table saw, using a jig to set the spacing. You could use any technique that allowed the turns to be spaced so that his alligator clip tuning method can be employed. Barry stretched the wire before winding to work harden it and then secured it in his form with hot glue.

Barry also made the L-network tuner shown in **Figure 15-12**. This uses two capacitors in parallel, one an old broadcast three-gang unit, making up the single capacitor for the L-network. The smaller capacitor makes it easy to make fine adjustments, once you're close with the larger unit.

Making Tuners for Balanced Loads

Many tuners; home made as well as commercial units, provide for balanced loads by including a balun between an unbalanced tuner and the balanced load. This is an appropriate approach if the load is within perhaps 4:1 of the typical 200 to 50 Ω balun often used for the purpose. If beyond that range, the balun does not perform very well and losses and even damage may result.

A different approach was presented by Richard Measures, AG6K, in a *QST* article some years back.[6] Measures presented a design that first transformed the balanced antenna system load to 50 Ω balanced and then made the transition from balanced to unbalanced using a 1:1 choke balun operating at its design impedance.

Measures used a balanced L-network to perform the transition to 50 Ω and followed it with a coax choke. The schematic is shown in **Figure 15-13**, with a photo of his breadboard version, less balun, in **Figure 15-14**. He went on to show an elegant design suitable for remote control using stepper motors with remote position indicators, available on the ARRL members Web site. However, for many, a duplication of his breadboard design may be more feasible.

Measures found that inductors in the range of 15 to 20 μH maximum inductance were suitable, with a 5 A current rating for 1.5 kW use. Inductors may be found at hamfests, or are available from MFJ (**www.mfjenter-**

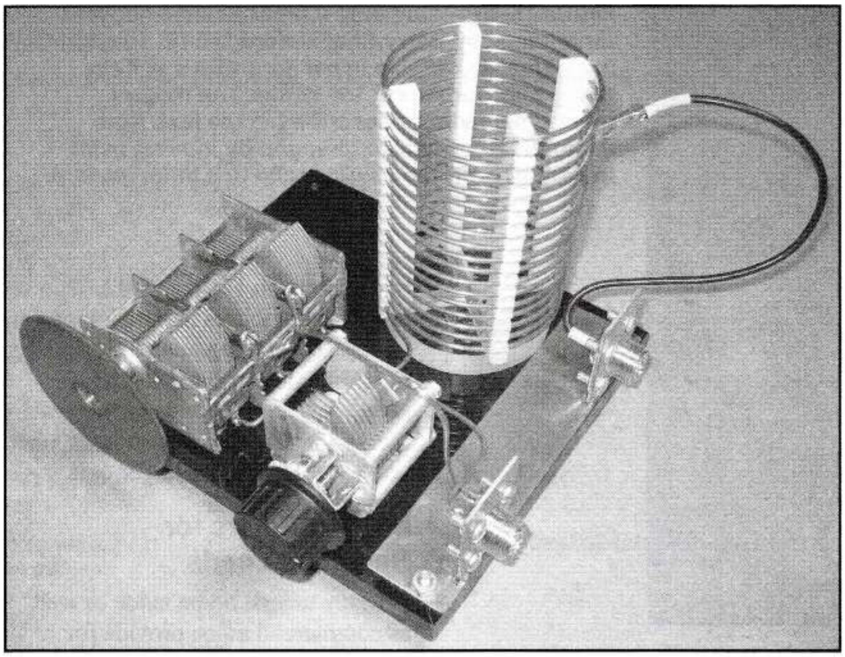

Figure 15-12 — An L-network tuner made by Barry Shackleford, W6YE. The two capacitors are in parallel making a single equivalent capacitance with the smaller one used for fine adjustment.

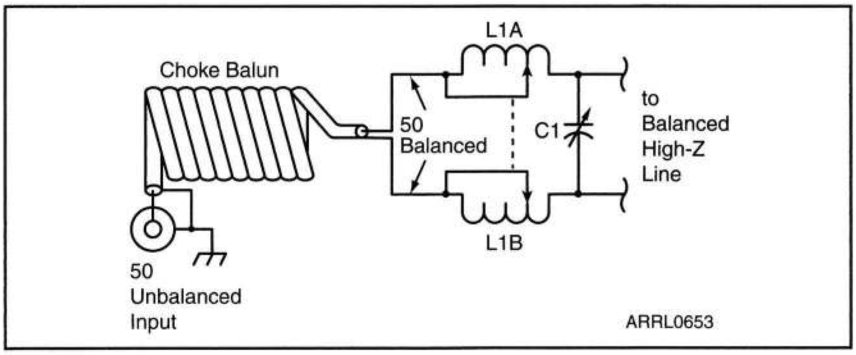

Figure 15-13 — The balanced L-network tuner designed by Richard Measures, AG6K.

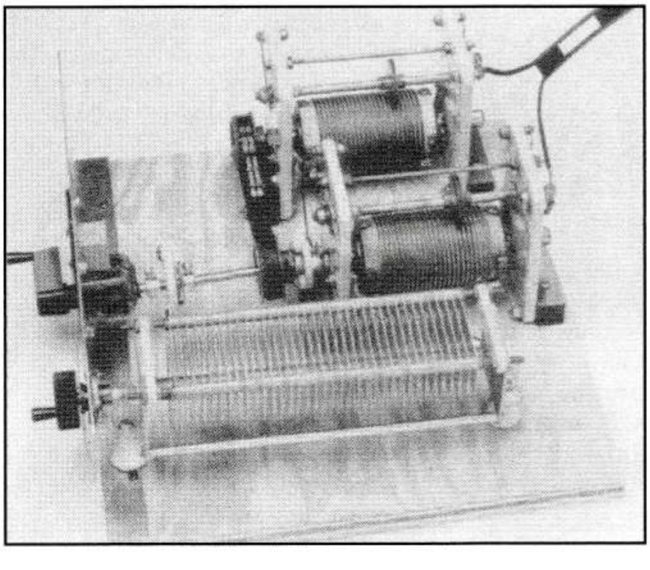

Figure 15-14 — Photo of Measures' balanced L-network tuner. The choke-balun is not shown.

prises.com). Suitable sprockets and drive belts to couple the inductors together are available from McMaster-Carr (**www.mcmaster.com**), or they may be coupled end-to-end with some loss of symmetry.

Capacitors should be in the 200 to 300 pF range, with a 1.5 kV rating if used at the 1.5 kW level. Note that both sides of the capacitor are hot with RF, so to avoid burns use an insulated shaft coupling on the way to the knob. Trust me, the set screw in an insulated knob will get you if you don't!

For the choke balun, Measures suggests using solid (not foam) dielectric RG-213, or equivalent, coax wound in a single layer on a 1 foot long, 5 inch diameter, PVC pipe. Other choke configurations may be used, including winding the coax on a ferrite toroid core. An adaptation of the design is to use the L-network to transform to 200 Ω, rather than 50 Ω, and then use a 4:1 balun to finish the impedance transformation and make the transition to unbalanced coax.

As shown, the L-network is designed to transform loads higher then 50 Ω. This is often, but not always the case with ladder and window line systems. If a lower impedance load is encountered, it will be necessary to move the capacitor to the other side of the inductors.

Notes

[1] J. Hallas, W1ZR, "Product Review — A Look at Some High-End Antenna Analyzers" *QST*, May 2005, pp 65-69.

[2] Several versions of *EZNEC* antenna modeling software are available from developer Roy Lewallen, W7EL, at **www.eznec.com**.

[3] R. D. Straw, Editor, *The ARRL Antenna Book,* 21st Edition. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9876. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/arrl-store**; **pubsales@arrl.org**.

[4] Note that the value of R will be higher than the dc resistance because of skin effect.

[5] L. McCoy, W1ICP (SK), "The Monimatch," *QST*, Feb 1957, pp 38-40.

[6] Richard Measures, AG6K, "A *Balanced* Balanced Antenna Tuner," *QST*, Feb 1990, pp 28-32.

Hairpin Tuners for Matching Balanced Antenna Systems

Balanced Transmatch designs for 28 to 450 MHz.

John Stanley, K4ERO

Once more, the advantages of ladder line for HF antennas were well presented in July 2008 *QST*.[1] Before WWII surplus brought us cheap coax, balanced feeders were almost always used for VHF as well. In the January 1942 *QST* reproduction that was included with The 2007 ARRL Handbook, we find that both home-brew and commercial VHF gear all used balanced lines.[2] For a given cost, open wire ladder line, window line or even TV twin lead can give you a lower loss installation than trying to buy large diameter coax in an effort to keep the losses to an acceptable level. This is dramatically demonstrated by comparing losses in various line types.[3] So, we wonder, why do so few present-day operators use ladder line or twin lead on the VHF frequencies? Might one reason be the lack of suitable *antenna tuners* (transmatches) for those bands? If balanced tuners were available, would ladder line be as popular at VHF as it has become at HF?

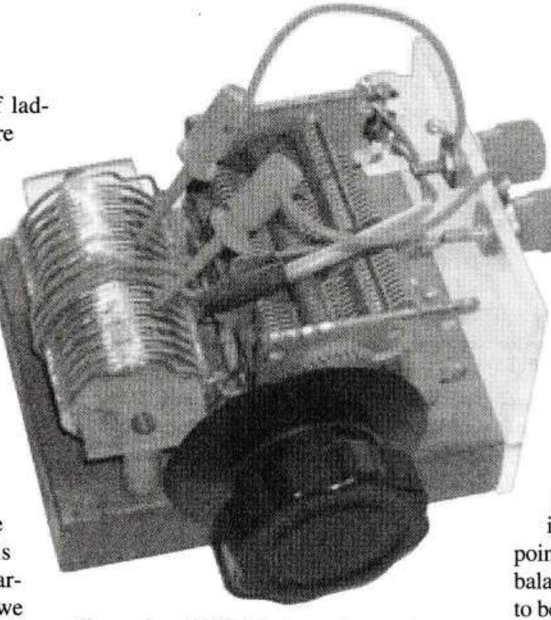

Figure 1 — K4ERO balanced tuner for 60 through 15 meters.

Balanced Tuners for HF Use

The ARRL Handbook, *The ARRL Antenna Book* and other ham publications have always included designs for balanced tuners.[4,5] Adam Nathanson, N4EKV, has one of many good Web sites showing this type of tuner at **www.n4ekv.com/tuners.asp**. I have used tuners like this for years with good results. Figure 1 shows the one I use at up to 100 W on the HF bands.

As noted in Volume 6 of *The ARRL Antenna Compendium*, I lean toward tuners with a fixed link and tapped coil.[6] The match is found by tuning the capacitor and adjusting the output side to connect to a variable number of turns, keeping the taps equidistant from the coil ends. Other versions use the taps for coarse tune and a variable capacitor in series with the link for finer tuning.

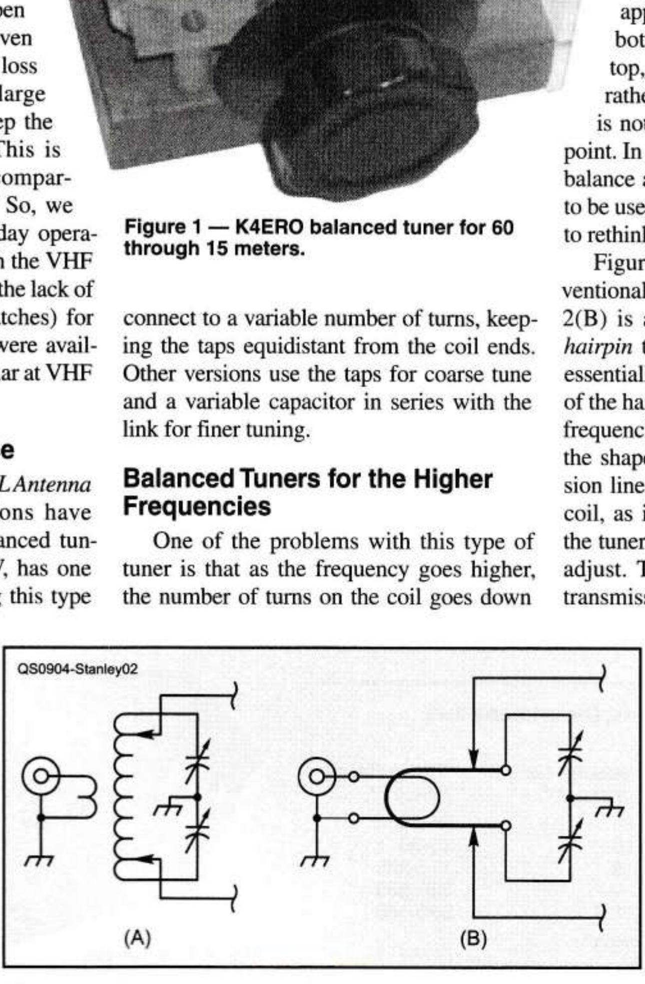

Figure 2 — Balanced tuner configurations. At (A) conventional tapped coil based tuner, at (B) the hairpin equivalent.

Balanced Tuners for the Higher Frequencies

One of the problems with this type of tuner is that as the frequency goes higher, the number of turns on the coil goes down rather quickly. By the time you get to 28 MHz, there may be only four turns on the coil. This means that the adjustment *steps* available are very limited. You can tap to either two or four turns, or if you are willing to unbalance the tap positions, or can access the opposite side of the coil, you could use one or three turns. In either case, the operation is compromised. Another approach is to connect one tap to the bottom of the coil, instead of to the top, effectively using fractions of a turn rather than whole turns, but this often is not feasible from a mechanical standpoint. In addition the symmetry and thus the balance are affected. If this type of tuner is to be used on 10 meters and higher, we need to rethink the design.

Figure 2(A) is a schematic of the conventional tapped link coupled tuner. Figure 2(B) is a representation of what I call a *hairpin* tuner. The electrical properties are essentially identical, but the physical layout of the hairpin type is optimum for the higher frequencies. By making the main inductor in the shape of a hairpin, or shorted transmission line, instead of a single layer solenoid coil, as is used in the conventional design, the tuner becomes much easier to build and adjust. The use of a short short-circuited transmission line section as an inductor is nothing new. It has been used for VHF/UHF circuits for many years.

Building Hairpin Tuners

For some time I have been using a hairpin inductor in a balanced tuner for 6 and 10 meters, and recently I built one for each of the 144, 220 and 432 MHz bands. The approach is the same on each frequency. A section of transmission line was used instead of the coil typically used on lower frequencies.

[1]Notes appear on page 36.

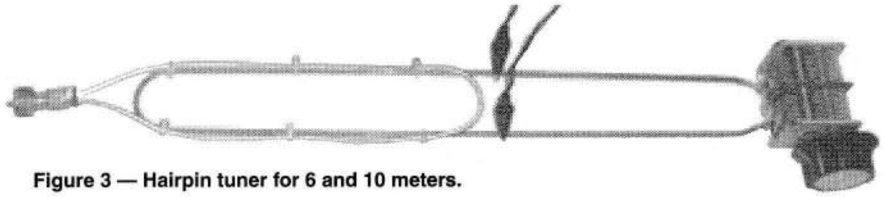

Figure 3 — Hairpin tuner for 6 and 10 meters.

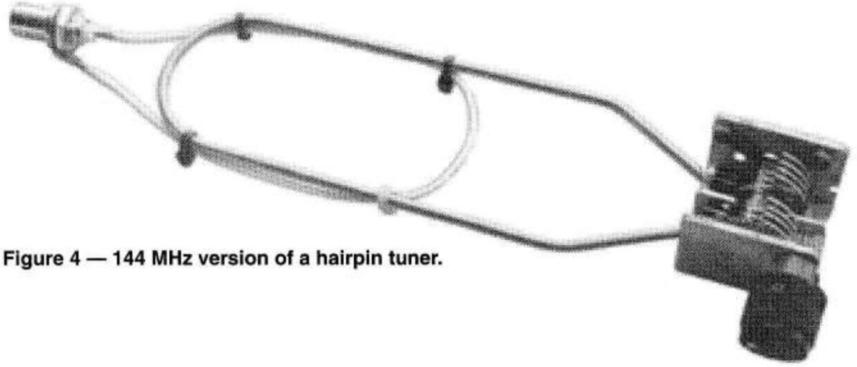

Figure 4 — 144 MHz version of a hairpin tuner.

The capacitor was a split stator or butterfly design. The input link is a single turn inductor that overlaps a portion of the main transmission line inductor and the output taps were taken at whatever point of the hairpin that gives the best match.

The advantage of this layout is that moving the taps to any point on the hairpin is the same as tapping on different numbers of turns on a coil, except that with the hairpin, it is very convenient to make the adjustment in as fine a step as may be desired. In all of my designs, the coupling loop is held to the main hairpin by cable ties. This allows some adjustment of the coupling loop, but holds the loop sufficiently snug so that it is not likely to move around accidentally. Use less loop coupling for higher Q with more selectivity and more loop coupling for lowest loss.

Making them Play

The frequency tuning range percentage will depend on the capacitor used. For widest range, select one with a high minimum to maximum capacitance ratio. All of the capacitors I have tried have provided adequate range to cover the desired amateur band. The range can be extended to cover a second band by putting fixed capacitors in parallel with the variable tuning capacitor. For example, the 6 meter version, shown in Figure 3, works for 10 meters with the addition of a parallel 40 pF fixed ceramic capacitor, while the 220 MHz version, shown in Figure 5, works fine on 2 meters with the addition of a 12 pF ceramic. As it is, the 2 meter version just makes it to 222 MHz, so two bands are possible without switching caps. A single tuner could also work on 10 and 6 meters without switching by careful component selection.

Capacitor Options

The most difficult component to find will likely be the split stator capacitor. There are ways of designing your own capacitor and it is made easier because of the relatively small capacitance required at these frequencies. For all of the designs shown here, I took capacitors from my well stocked junk box. You less well equipped folks will have to search at a hamfest or check out the basement of one of the local old timers. Commercial capacitors are available, but the price may shock you. You could also choose to use a single section capacitor instead of the split stator. In that case, the capacitor shaft and frame will be "hot" and must be floated above ground. You will have to tune it via a long insulated shaft. And, of course, the balance will be somewhat compromised. The method does work and a suitable single section capacitor may be easier to find, however . Alternatively you could use a pair of identical capacitors to ground if you adjusted them each to the same setting or worked out a common shaft arrangement. Fussy, but it could work. Target dimensions and component values for the bands in this range are shown in Table 1.

Inductor Choices

The length of the hairpin will depend on the value of your capacitor. The values below represent tuners I have built and should give you a good starting point. Match the hairpin width to the spacing of the capacitor terminals, or bend the ends of the hairpin in or out at the capacitor end in order to make the connections. Spacing does affect the inductance value so keep it close to what you see in the photos.

My inductors are ⅛ inch diameter brass rod, but they could be soft copper tubing or wire in sizes from 12 gauge up to ¼ inch. Brass welding rods from the hardware store could also be used. A smaller diameter means the hairpin should be shortened a bit as the inductance per inch will be higher. The links should be insulated wire, either enameled or PVC coated or, best of all, Teflon insulated. The 6 meter tuner in Figure 3 uses the shield of Teflon coax as the link. My links do not make electrical contact anywhere with the main hairpin. The coax shield and center of the hairpin could both

Table 1

Hairpin Tuner Component Values, Dimensions and Frequency Range

Band	*Capacitor Value (pF/Section)**	*Inductor Length (Inches)***	*Tuning Range (MHz)****
10 Meters	95-170	15	26-34
6 Meters	15-90	15	36-90
2 Meters	3-20	6	110-225
222 MHz	3-10	4	220-330
432 MHz	3-4	1.5	390-440

*Effective capacity is ½ of the value/section.
**All inductors made of ⅛ inch brass rod.
***Range may by reduced or shifted by reactive loads.

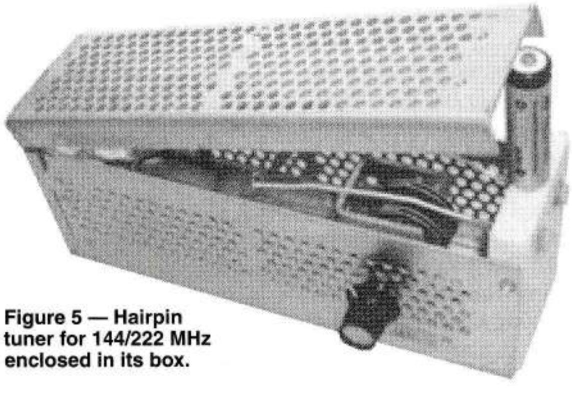

Figure 5 — Hairpin tuner for 144/222 MHz enclosed in its box.

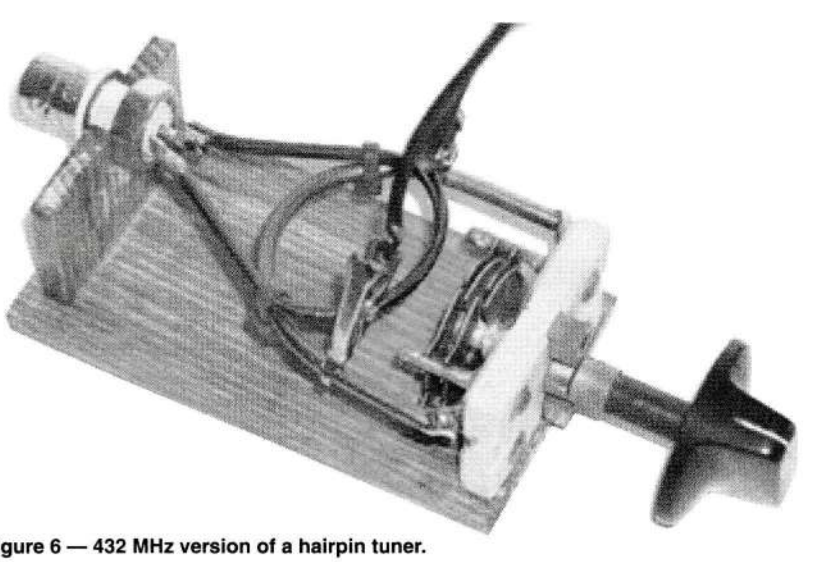

Figure 6 — 432 MHz version of a hairpin tuner.

be grounded to a chassis, if desired.

Tuning Up

Tuning consists of setting the taps to an intermediate position and adjusting the capacitor for minimum SWR. If SWR is not as low as desired, move the taps a bit either towards or away from the capacitor and readjust the capacitor. You should be able to find tap settings that allow the SWR to be reduced to 1:1.

Safety Considerations

Note that neither of the tuners described above are shown installed within a chassis of any kind. This is to show the construction more clearly. Also, I must admit that during tests, I got away with this because I used low power and am very careful not to touch the *hot* parts of the circuit. I also have a non-metallic operating desk. The open construction is useful during experiments.

I definitely do not recommend this approach for general use. You will want to put your tuner in a shielded box of some kind, probably with a hinged lid to allow you to move the taps as shown in Figure 5. You should *never* adjust the tap clips with power applied. The tuning shaft should be brought out through the enclosure where an insulated knob should be installed for tuning even though with a split stator capacitor the shaft should be at ground potential. For high power, the use of a suitable enclosure is essential, especially if the tuner is to be installed close to the operating position. This is to prevent RF burns from contact as well as exposure to excess RF levels.

Remember that your body is more susceptible to pick up from VHF fields than is the case at HF. If you put the tuner well away from the operating position in a place that is protected from access by family members or pets, you may be able to use a somewhat more open construction as is sometimes done with conventional home brew tuners. Radiation from an unenclosed tuner of any type can be enough to cause interference with nearby electronics devices and could cause fires if anything flammable comes in contact with the hot parts of the circuit.

These tuners have been tested with 100 W on 10 and 6 meters, 50 W on 2 meters and 20 W on 70 cm, the maximum output of my rig. For higher power, the designs can be scaled, remembering that the bigger it gets, the lower the frequency for the same geometry. Thus, a design similar to that used here for 432 MHz, but three times larger, would probably work fine with a full kW, but on 144 MHz. Since my 20 W, 432 MHz design uses an inductor that is about as short as is practical (see Figure 6), getting up to 1 kW at 432 MHz, might prove difficult with this design. At the least, a different type of capacitor would be required.

I hope that these simple to build and adjust tuners will start a trend towards greater use of balanced feeders on the higher frequencies just as ladder line has become the favorite for many on the lower bands.

Notes

[1] J. Hallas, W1ZR, "Getting on the Air—Your Second HF Antenna," *QST*, Jul 2008, pp 69-70.

[2] B. Goodman, W1JPE, "Receivers for 112-Mc. Emergency Work," *QST*, Jan 1942, pp 18-25, 74-75.

[3] *The ARRL Handbook for Radio Communications*, 2009 Edition, Figure 21.4. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 0261 (Hardcover 0292). Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

[4] See Note 3, p 21.13.

[5] R. D. Straw, Editor, *The ARRL Antenna Book*, 21st Edition, p 25-3. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9876. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

[6] J. Stanley, K4ERO, "The Filtuner," *ARRL Antenna Compendium, Volume 6*. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 7431. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

John Stanley, K4ERO, holds an Amateur Extra class license and has been licensed for over 50 years. He has worked as a broadcast engineer most of his life, mainly with religious shortwave stations in many parts of the world. He graduated from MIT in 1962 with a BSEE degree. John is an ARRL Technical Adviser. He and his wife, Ruth, WB4LUA, live in Rising Fawn, Georgia. You can reach John at 524 White Pine Ln, Rising Fawn, GA 30738 or at **jnrstanley@alum.MIT.edu**.

Review Questions

15-1 What are some major challenges involved with building an antenna tuner?

15-2 Under what conditions is a special purpose tuner a viable option?

15-3 Why might you decide not to include 160 meters in your tuner design?

APPENDIX A

Making Sense of Decibels

Joel R. Hallas, W1ZR

Much of the technical data found in *QST* and other ARRL publications is expressed in decibels, generally shown as *dB*. Contrary to the belief of some, this is not done to mystify beginners, but rather is intended to make comparisons easier and avoid the use of very large or small numbers.

So What's this Decibels Business All About?

Decibels are just a way of expressing ratios, often power ratios. If we are looking at the gain of an amplifier stage, the pattern of an antenna or the loss of a transmission line we are generally interested in the ratio of the power out to the power in, or the ratio of the power in front of a beam antenna to that coming from the back. These are some of the places that we will find the results expressed in dB.

Decibels are a logarithmic function. Logarithms are a handy mathematical tool based on exponents. An important feature of logarithms is that multiplication can be performed by adding the logarithmic quantities instead of multiplying them. Similarly, divisions can be accomplished by subtracting in the same manner. This becomes a benefit if you are dealing with multiple stages of amplification and attenuation — as we often are in radios — snatching a minuscule signal from the ether — amplifying and processing it so we can hear it out the loudspeaker. Instead of having to multiply and divide at each stage to keep track of the progress of our signal processing — often with signal levels with many zeros to the right of the decimal point — we can just tally all the dB and have total gain of the system.

So How do We Compute the Decibels?

The *deci* in decibels refers to a factor of ⅒, as in *deciliters* for ⅒ of a liter, while the *bel* relates to the idea of a logarithmic ratio, originally used to define sound power.

- To convert a power ratio into decibels, just:

1. Find the base 10 logarithm of the power ratio.
2. Multiply by 10.

For example, if we have an amplifier with a power gain of 275, we find the logarithm of 275 (see below, if you don't do logs in your head) to be 2.44. We multiply by 10 and the result is that a power gain of 275 can be represented as 24.4 dB.

- To convert decibels to a power ratio, we do the opposite:

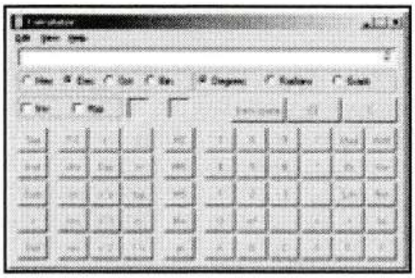

Figure 1 — The *Windows Scientific Calculator* ready to calculate the effect of a 2 dB cable loss.

1. Divide by 10.
2. Find the base 10 antilog of the result.

Note that the base 10 antilog of a number is just 10 raised to the power of the number. This is also something you probably don't do in your head, so let's see how to easily perform the computations.

Understanding a few characteristics of logs will help avoid problems interpreting results. Note that a gain of 0 dB means that there is no change to the signal — not that the signal has vanished! The other important fact is that a power ratio of less than one (a loss rather than a gain) is represented as a negative number in decibels.

Enter the *Windows Scientific Calculator*

In the *very* old days, engineers and technicians used tables to make accurate logarithmic calculations, and mechanical slide rules if three significant digits was sufficient precision.[1] Starting around 1970, scientific calculators became available. Initially they were expensive typewriter-sized devices that were typically shared within an engineering department. Within a few years pocket-sized units were available for less than $200, and now everyone could make calculations to a precision of nine significant digits, whether warranted by the data accuracy or not. Tables and slide rules were relegated to the pages of history with spark transmitters.

The dawn of the reasonably priced personal computer seemed to push the fancy scientific calculator out of sight only about 10 years later. Unfortunately, for many functions a calculator may be a better choice, and decibel calculations may be one. The calculations may be easily performed on a PC by putting the equations in cells of an Excel spreadsheet, but not everyone can do that without a lot of thought.

If you have a suitable scientific calculator, that should easily make the calculations. Not all have an ANTILOG button, but if not, they will likely have a button that says X^Y, which can be used as above. If you don't have a handheld calculator, you may not know that there is a very capable one included as an "accessory" within the Microsoft *Windows* operating system! Just click START then ALL PROGRAMS then ACCESSORIES. You should find an icon for the *Windows Calculator*. You could open it, but why not drag it onto your desktop first.

On first opening, you may find a four-function grocery store type calculator. Have no fear, just click on VIEW then SCIENTIFIC to get the one you want. It should look about like Figure 1.

Give it a Test Drive

Let's say you have a mismatched coax cable with a loss of 2 dB. You may want to know how many of the 100 W from your transmitter actually reach your antenna. Remember a 2 dB loss is a "gain" of –2 dB! We'll go through the inverse dB calculation of column 2. Using your *Windows Calculator*, either hit 2 on your keyboard, or:

- Click on the 2 on the calculator "keypad," then:
- Click on the +/– key; the display should show – 2, as in Figure 1.
- Click on the / key to select the "divide by" operation. Enter the digits 1 and 0 for the number 10 and hit ENTER or =. Make note of the result (–0.2) or store it in memory (M+).
- Enter the digits 1 and 0 for the number 10.
- Click on X^Y to raise 10 to a power.
- Enter your earlier result, –.2 and hit ENTER.
- The display should show 0.63095734448019324943436013662234 or about 0.63. That is the fraction of your power left after a 2 dB loss. That means your antenna sees 63 W and 37 W is heating your transmission line.

While you are using the *Windows Scientific Calculator* look over its other features. It provides painless conversion between decimal and binary, octal or hex numbers, for example.

Joel R. Hallas, W1ZR, is QST *Technical Editor. He can be reached at* **w1zr@arrl.org**.

---

[1] See, for example, *Standard Mathematical Tables*, CRC Press, any edition. In addition to tables of logarithms and trigonometric functions, it includes many handy formulas from geometry, trigonometry and calculus.

From April 2007 QST © ARRL

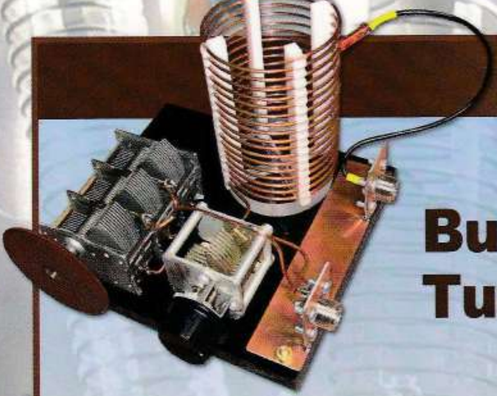

Buy or Build the Best Tuner for Your Station.

Antenna tuners are devices often misunderstood in the Amateur Radio world. While not every station requires an antenna tuner to transmit radio signals, often an incompatibility between the transmitter and the antenna system results in poor performance. An antenna tuner between them is the way to obtain efficient operation.

The ARRL Guide to Antenna Tuners discusses the details of the different configurations and requirements of antenna tuners. It explores the design, construction and applications of the different types available. You'll learn if one is necessary in your station, what type of tuner is needed and where to install it for maximum improvement. This book will give you a better understanding of your antenna system and the way it can be enhanced through the selection and use of the appropriate antenna tuner.

Contents:

- Why Might I Need an Antenna Tuner
- A Look at a Typical Configuration
- So Just What is an Antenna Tuner?
- Tuning an Antenna Tuner
- The Internal Tuner—How Does it Help?
- An External Tuner at the Radio
- Transmission Lines and Loss
- Moving the Tuner to the Back 40
- Transmission Line Choices for Low Loss
- Balanced versus Unbalanced Lines
- So What's a Balun, an Unun, a Choke?
- Balanced Antenna Tuners
- Antennas that Work Well with Tuners
- A Survey of Available Tuners
- Making Your Own Tuner

Joel Hallas has been a radar and telecommunications systems engineer, engineering manager, telecommunications executive and college teacher, and is now Technical Editor of ***QST***, the monthly journal of ARRL, the national association for Amateur Radio. He has earned BS and MS degrees in electrical engineering. Joel is a radio amateur, and holds the license call sign W1ZR.

***ARRL** — devoted entirely to Amateur Radio*

ARRL has books, software, online courses and other resources for licensing, operating and education. Look for these other popular ARRL titles:

- ***The ARRL Antenna Book***
- ***Basic Antennas: Understanding Practical Antennas and Design***
- ***The ARRL Antenna Designer's Notebook***

Published by:
ARRL *The national association for* **AMATEUR RADIO™**
225 Main Street
Newington, CT 06111-1494 USA
www.arrl.org

ISBN 978-0-87259-098-4
52295
9 780872 590984
USA $22.95 ARRL Order No. 0984

PART II—
BUILDING AND USING BALUNS AND UNUNS

Chapter 7

The 1:1 Balun

Sec 7.1 Introduction

In this chapter I will introduce the most popular broadband Balun in amateur radio use—the 1:1 Balun. This topic been discussed in the amateur radio literature since the publication of Turrin's 1964 article.[8] Although Turrin's Balun is really a version of Ruthroff's, which was introduced in his classic 1959 article,[9] the real beginning of the broadband 1:1 Balun dates back to Guanella's classic paper of 1944.[3] Guanella's objective was to design a broadband 16:1 Balun to match the balanced output impedance of 960 ohms of a push-pull, 100-watt vacuum-tube amplifier to the unbalanced load of a matched 60-ohm coaxial cable. Use of his approach for 1:1, 4:1, and 9:1 Baluns has produced the designs of choice. They are presently called current or choke Baluns.

This chapter begins with an introduction to the technology of transmission line transformers. Other topics discussed include: 1) when to use a Balun; 2) highlights of significant articles in the professional and amateur radio literature; 3) high-power, medium-power, and low-power designs; and 4) isolation transformers. The latter is presented here for the first time. The chapter closes with a brief summary of the significant points included within.

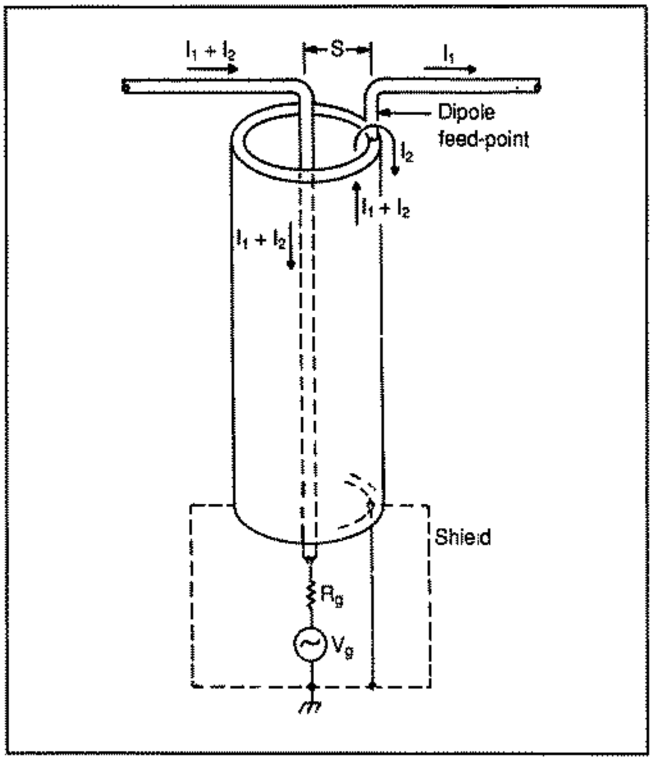

Figure 7-1. *An illustration of the various currents at the feedpoint of a dipole.* I_1 *is the dipole current and* I_2*, the inverted L (imbalance) current.*

Sec 7.2 When to Use a Balun

Baluns have taken on a more significant role in the past few decades with the advent of solid-state transceivers and Class B linear amplifiers with unbalanced outputs. That is, the voltage on the center conductor of their output chassis connectors varies (plus and minus) with respect to ground. In many cases, coaxial cables are used as the transmission lines from these unbalanced outputs to antennas like dipoles, inverted Vs, and Yagi beams that favor a balanced feed. In essence, they prefer a source of power whose terminals are balanced (voltages being equal and opposite) with respect to actual ground or to the virtual ground that bisects the center of the antenna. The question that is asked most frequently is whether a 1:1 Balun is really needed.

To illustrate the problem involved and to give a basis for my suggestions, I refer you to **Figure 7-1**. Here we have, at the feedpoint of the dipole, two equal and opposite transmission line currents with two components each—I_1 and I_2. Also shown is the spacing, s, between the center conductor and the outside braid. Theoretically, a balanced antenna with a balanced feed would have a ground (zero potential) plane bisecting this spacing. However, because a coax-feed is unbalanced and the outer braid is also connected to ground at some point, an imbalance exists at the feedpoint giving rise to two antenna modes. One lies with I_1, providing a dipole mode; the other lies with I_2, providing an inverted L mode.

If the spacing, s, is increased, the imbalance at the feedpoint becomes greater—giving rise to more current on the outer braid and a larger imbalance of currents on the antenna's arms. Several steps can be taken to eliminate or minimize the undesirable inverted L mode (that is, eliminate or minimize I_2). The obvious choice is to use a well-designed Balun that not only provides a balanced feed, but also minimizes (by its choking reactance) I_2—if the coaxial cable does not lie in the ground plane that bisects the center of the dipole. The other step is to ground the coaxial cable at a quarter-wave (or odd-multiple thereof) from the feedpoint. This discourages the inverted L mode because any radiating element will want to see a high impedance at these lengths instead of the low impedance of a ground connection.

I conducted experiments with Baluns on a 20-meter half-wave dipole at a height of 0.17 wavelengths, which gave a resonant impedance of 50 ohms. VSWR curves were compared under various conditions. When the coaxial cable was in the ground plane of the antenna (that is, perpendicular to the axis of the antenna), the VSWR curves were identical with or without a well-designed Balun—no matter where the outer braid was grounded. A significant difference was noted only when the coaxial cable was out of the ground plane. When the cable dropped down at a 45-degree angle under the dipole, a large change in the VSWR took place. This meant that the inverted L mode was appreciable.

It should also be mentioned that the direction of I_2, the imbalance current, can depend upon the side on which the coaxial cable is out of the ground plane of the dipole. For example, if the cable comes down under the right side in **Figure 7-1** (that is, the angle between the horizontal arm and the coax is less than 90 degrees on the right side and more than 90 degrees on the left side), then the direction of I_2 can be reversed by the imbalance in the induced currents on the outside of the braid. By the same token, by having the coaxial cable coming down on the other side, the value of I_2 is only increased in magnitude.

However, feeding a Yagi beam without a well-designed 1:1 Balun is a different matter. Because most Yagi designs use shunt-feeding (usually by hairpin matching networks) in order to raise the input impedance close to 50 ohms, the effective spacing (s) is greatly increased. Furthermore, the center of the driven element is actually grounded. Thus, connecting the outer braid (which is grounded at some point) to one of the input terminals, creates a large imbalance and a real need for a Balun. An interesting solution, which would eliminate the matching network, is to use a step-down Balun designed to match 50-ohm cable directly to the lower balanced-impedance of the driven element.[2]

In summary, if you concur with the theoretical model of **Figure 7-1**, my experiments performed on 20 meters, and the reports from radio amateurs using dipoles and inverted Vs without Baluns, then it appears that 1:1 Baluns are really needed for: a) Yagi beam antennas where severe pattern distortion can take place without one, and b) dipoles and inverted Vs that have the coaxial cable feed lines out of the ground plane that bisects the antennas, or that are unbalanced by their proximity to manmade or natural structures. In general, the need for a Balun is not so critical with dipoles and inverted Vs (especially on 40, 80, and 160 meters) because the diameter of the coaxial cable connector at the feedpoint is much smaller than the wavelength.

If my model, which assumes that a part of the problem when feeding balanced antennas with coaxial cable is related to the size of the spacing, s (shown in **Figure 7-1**), then the possibility exists for using Ununs for matching into balanced antennas with impedances other than 50 ohms and with small values of s. For example, half-wave dipoles at a height of about a half-wave, quads, and center-fed 3/2-wave dipoles—which all have impedances close to 100 ohms—could very well be matched to 50-ohm cable by a 2:1 Unun. As I will show, they are considerably easier to construct than 2:1 Baluns. Furthermore, Genaille[10] has recently shown considerable success using Ununs in this kind of application.

To wind up this section, I would like to comment on an article published by Eggers,[11] WA9NEW, concerning the use of a Balun with a half-wave dipole. While at North Carolina State University, he conducted an experimental investigation of pattern distortion without a Balun at 1.6 GHz in an RF anechoic chamber (which simulates "free space"). Briefly, his results showed that, with a Balun (bazooka type), the antenna radiation pattern compared very favorably with the classic "figure-eight." Without the Balun, the radiation pattern was severely distorted.

Even though the author expressed difficulty in obtaining accurate measurements at this very high frequency, I have a question regarding the validity of performing the experiment in the first place. From the photograph in the article, it appears that conventional coaxial cable and connectors were used in the experiment. If we assume an effective diameter of 0.375 inches for these components, then scaling up to 3.5 MHz (457.14 fold) results in a coaxial cable with a diameter of 14.28 feet! I am quite sure that the large spacing, s, of 7.14 feet would bring about a noticeable imbalance resulting in appreciable pattern distortion even at 3.5 MHz.

Sec 7.3 Highlights of Significant Articles on 1:1 Baluns

Although there have been many articles on 1:1 Baluns published in the professional and amateur literature, I have selected for review a few that I believe have had the most impact on 1:1 Baluns for amateur radio use. As you will see, even though I consider some of the amateur articles significant, their impact upon the use and understanding of these devices has not always been positive. In fact, in some cases, the opposite has been true.

In the professional literature

There are actually only two significant articles in the professional literature that provide the fundamental principles upon which the theory and design of this class of transformers are based. It can be said that succeeding investigators simply extended the works of the authors of these two articles.

The first presentation on broadband matching transformers using transmission lines was given by Guanella in 1944.[3] He coiled transmission lines forming a choke such that only transmission line currents were allowed to flow, no matter where a ground was connected to the load. His single, coiled transmission line resulted in a 1:1 Balun. It is shown schematically in **Figure 7-2A**. Prior to this, RF Baluns were achieved by the use of quarter and half-wave transmission lines, and as a result, had narrow bandwidths. Guanella then demonstrated broadband Baluns with impedance transformations of 1:n^2 where *n* is the number of transmission lines he connected in a series-parallel arrangement.

Several important points should be made regarding Guanella's 1:1 Balun shown in **Figure 7-2A**. With sufficient choking reactance, the output is isolated from the input and only flux-canceling transmission line currents are allowed to flow. With matched or very short transmission lines, the grounding of terminal 5 (actually or virtually like the center of a dipole) results in terminal 4 becoming $+V_1/2$ and terminal 2 becoming $-V_1/2$—creating a balanced output. This type of Balun has lately been called a "current" or "choke" Balun. A significant feature of this model is that a potential gradient of $-V_1/2$ exists along the length of the transmission line. This gradient, which exists on both conductors, accounts for practically all of the loss in these transformers because the loss mechanism is voltage dependent (a dielectric-type loss). All transmission line transformers have some sort of voltage gradient along their transmission lines and are, thus, subject to the same type of losses. Furthermore, the theory and loss mechanism is the same whether the transmission lines are coax or twin-lead, or coiled around cores, or threaded through ferrite beads. Additionally, it was shown[2] that higher-impedance Baluns or Baluns subjected to higher VSWRs have more loss because the voltage gradients are also larger.

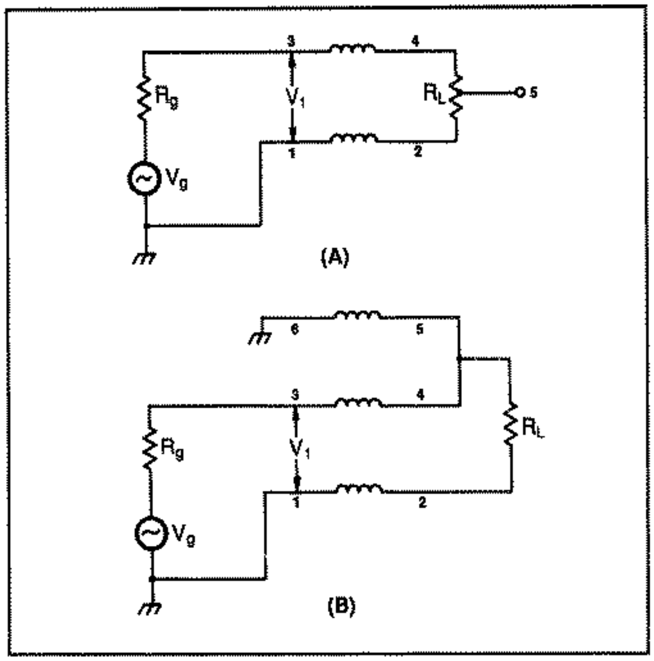

Figure 7-2. *Two versions of the 1:1 Balun: (A) the Guanella Balun and the basic building block; (B) the Ruthroff Balun as originally drawn.*

The second significant article on broadband transmission line transformers was published by Ruthroff.[9] His 1:1 Balun, which is shown as originally drawn in

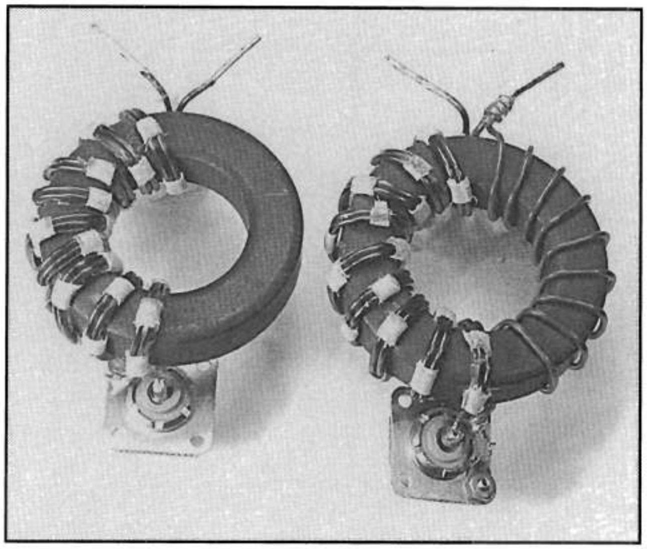

Photo 7-A. *The two basic forms of the 1:1 Balun that first appeared in the professional literature. The two-conductor Guanella Balun is on the left and the three-conductor Ruthroff Balun is on the right.*

Figure 7-2B, used an extra winding to complete (as he said) the path for the magnetizing current. Even though his schematic drawing appeared to look like a trifilar winding, his pictorial in the article clearly showed that the third winding (5-6) was on a separate part of the toroid. With an equal number of turns, it forms a voltage divider with winding (3-4) placing terminal 4 at $+V_1/2$ and terminal 2 at $-V_1/2$. In his classic paper, Ruthroff also presented his form of the 4:1 Balun (which is also different from Guanella's), a 4:1 Unun, and various hybrids. **Photo 7-A** shows the two basic forms of the 1:1 Balun that first appeared in the professional literature. The two-conductor Guanella 1:1 Balun is on the left and the three-conductor Ruthroff Balun is on the right. As was mentioned before, the Guanella Balun has also been called a "current" or "choke" Balun.

Before moving on to the significant articles in the amateur radio literature, some mention should be made of the differences between the two basic forms shown in **Photo 7-A**. Guanella's 1:1 Balun came to be known as the *basic building block* for this whole class of broadband transformers. This term was coined by Ruthroff as he showed its 1:1 Balun capability when the load was grounded at its center (terminal 5), and as a phase-inverter when the load was grounded at the top (terminal 4). By connecting terminal 2 to terminal 3 and connecting the bottom of the load to ground, Ruthroff demonstrated his very popular 4:1 Unun. I call this type of arrangement the bootstrap connection. By grounding terminal 2, there is no potential drop along the transmission line and, therefore, no need for magnetic cores or beads. This arrangement, which turns out to be an important function for extending the high frequency performance of this class of transformers, I call the "phase-delay" connection.

Thus, with the flexibility shown by Guanella's basic building block, a 1:1 Balun is now realized that not only presents a balanced power source to a balanced antenna system, but can also prevent an imbalance current (an inverted L antenna current) by its choking reactance when the load is unbalanced or mismatched or when the feedline is not perpendicular to the axis of the antenna.

Interestingly enough, except at the very low end of the frequency response of the Ruthroff 1:1 Balun where autotransformer action can take place, his Balun takes on the characteristics of the Guanella Balun. The reactance of the third winding becomes great enough to make it literally transparent. This is not the nature of the trifilar-wound (voltage) Balun, which is sensitive to unbalanced and mismatched loads over its entire passband because it is actually two tightly coupled transmission lines. This distinction was not recognized by most of those who published in the amateur radio literature.

In the amateur radio literature

R. Turrin, W2IMU—1964

The first presentation in the amateur radio literature on 1:1 Baluns using ferrite cores was by Turrin in 1964.[8] Turrin, who was a colleague of Ruthroff's at Bell Labs, took his small-signal design (which used No. 37 or 38 wire on toroids with ODs of 0.25 inches or less) and adapted it to high-power use. This was done by using thicker wire, larger cores, and (very importantly for high efficiency[2]) low permeability ferrite. Ruthroff used lossy manganese-zinc ferrites with permeabilities of about 3000 because efficiency was not a major consideration.

Figure 7-3 is a pictorial and a schematic of Turrin's design. As you can see, the third wire (winding 3-4) is placed between the two current-carrying wires (windings 1-2 and 4-5). **Photo 7-B** shows (on the left) his actual design using a ferrite core; a popular design (on the right) using a powdered-iron core is available in kit form from Amidon Associates, Inc. Both Baluns use 10 trifilar turns of a single-coated wire like Formex™ or Formvar™ on a toroid. Turrin's design

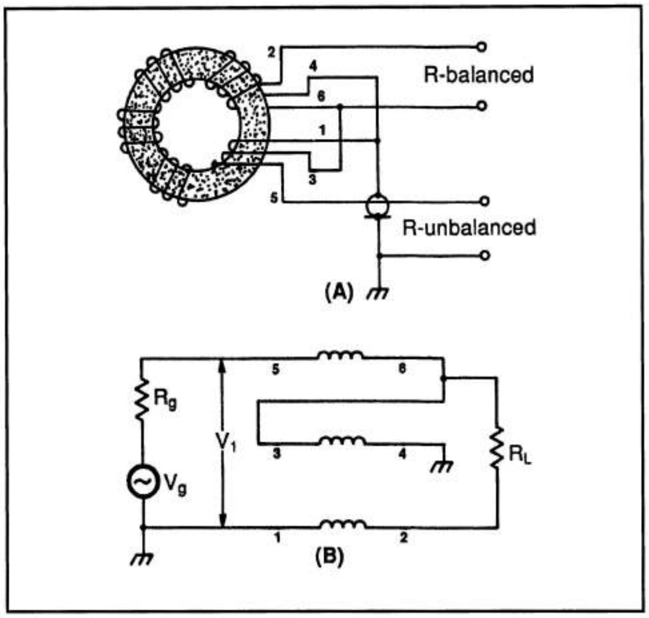

Figure 7-3. *(A) A pictorial of Turrin's 1:1 Balun, and (B) a schematic of his Balun.*

uses a ferrite toroid with an OD of 2.4 inches and a permeability of 40. The Amidon Associates' Balun uses a powdered-iron toroid with a 2-inch OD and a permeability of only 10. Both Baluns are specified to handle 1000 watts of power from 1.8 to 30 MHz.

Figure 7-4 shows the response curves for these two Baluns when terminated with 50-ohm loads. The response curve for a popular 1:1 rod-type Balun that uses the same schematic and wire is also shown. It has 8 trifilar turns, tightly wound on a rod of 0.5-inch diameter, 2.5 inches long, and with a permeability of 125. The rod-type Balun is shown in **Photo 7-C**.

Several important features should be brought out regarding the results shown in **Figure 7-4**. They are:

1. All Baluns had insufficient choking reactance and, hence, poor low-frequency responses. The powdered-iron version was especially poor. They all showed a drop in the input impedance and an inductive component at 2 MHz. This meant flux in the cores and an undesirable condition—especially for ferrite, which is a nonlinear material. Ferrite cores could not only suffer damage, but they could also generate spurious frequencies under these conditions. In fact, the same condition could occur at 4 MHz with a VSWR of 2:1! Therefore, I don't recommend any of these Baluns for use on 160 or 80 meters.

2. The major problem at the low-frequency end is the role of the third winding (3-4) in **Figure 7-3B**. It has been claimed[12] that the third winding improves the low-frequency response (over the two-conductor Guanella 1:1 Balun) because it enables autotransformer action at the low end. However, recent measurements I have made on two-conductor Guanella Baluns and three-conductor Ruthroff (or Turrin) Baluns, with loads grounded at their centers, show insignificant differences. This type of load approximates the actual condition when feeding a balanced antenna system. The negative feature of the third winding (3-4) is that, at the low-frequency end, there can be insufficient reactance to prevent harmful flux in the core because of a direct shunting path to ground. With the two-conductor 1:1 Balun, the only flux-inducing current is that of the imbalance current (the inverted L mode), which is usually far smaller.

3. Another important feature of the curves shown in **Figure 7-4** is the effect of the characteristic impedances of the coiled transmission lines. For example, a bifilar winding (wires tight together) on a toroid with spacing between adjacent bifilar turns exhibits a char-

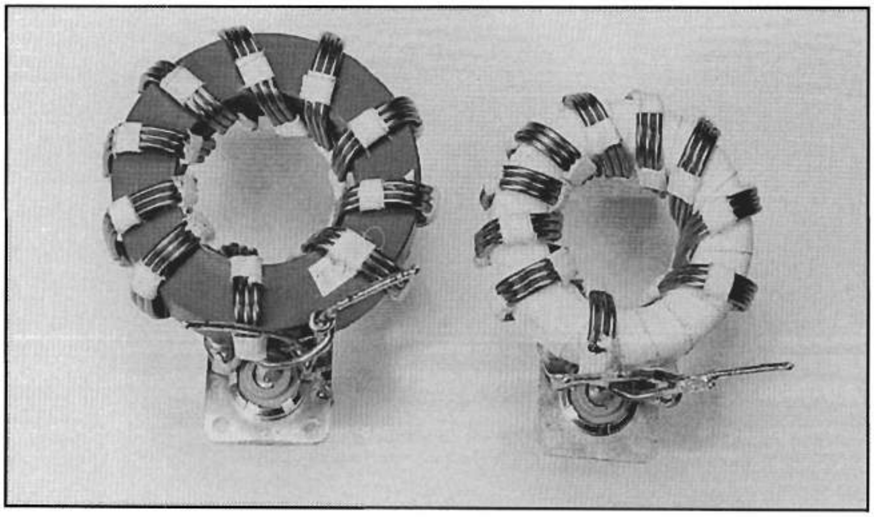

Photo 7-B. *Two versions of Turrin's design: On the left, the 1:1 Balun that has appeared in the amateur radio literature; on the right, a 1:1 Balun that has been readily available in kit form from Amidon Associates, Inc.*

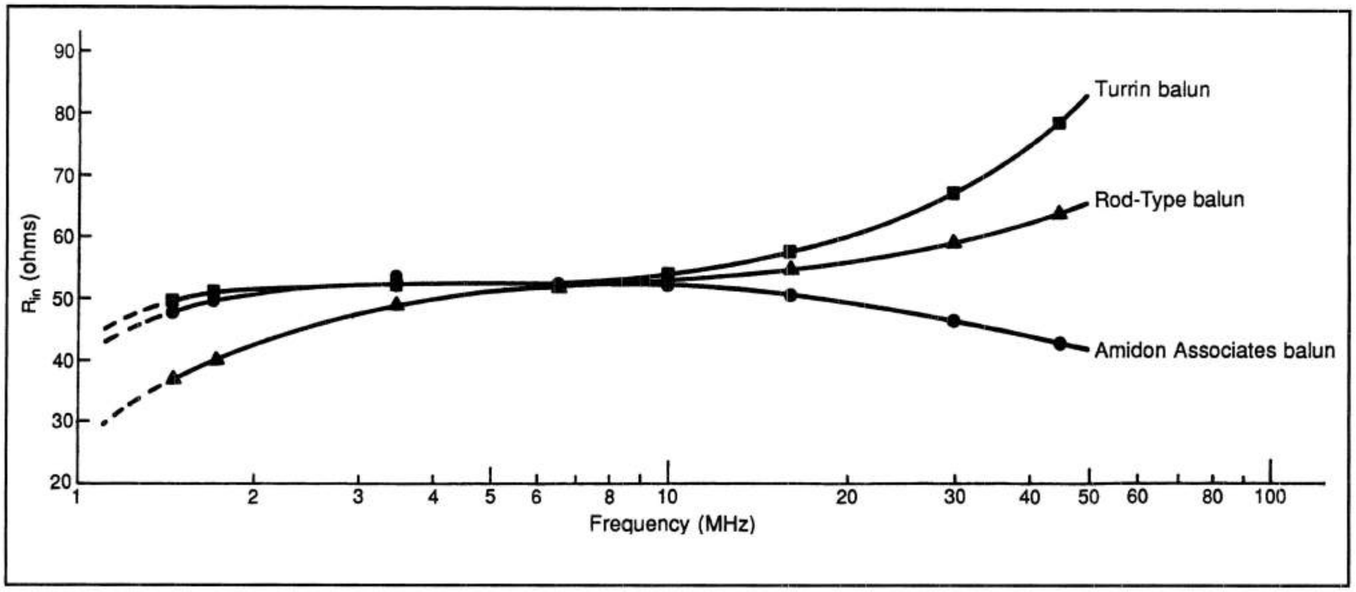

Figure 7-4. *The input impedance versus frequency, when terminated with 50 ohms, for the Turrin, typical rod-type, and Amidon Associates 1:1 Baluns.*

acteristic impedance of about 45 ohms. When wound on a rod with no space between adjacent bifilar turns, the characteristic impedance drops to about 25 ohms. With the third winding (3-4) between the other two as shown in **Figure 7-3B**, the characteristic impedance is raised to approximately 70 ohms in the toroidal case and to about 47 ohms in the rod case. If the toroidal Baluns were terminated in 70 ohms and the rod Balun terminated in 47 ohms, the high-frequency responses would be practically flat to at least 30 MHz. The difference in high-frequency response between the two toroidal Baluns (with 50-ohm loads) is due to the differences in the lengths of their transmission lines. The transmission line on the powdered-iron core is appreciably less because the OD, as well as the cross-sectional area, is smaller.

4. The trifilar-wound form of the 1:1 Balun also has an additional undesirable property. Its high frequency response is sensitive to unbalanced and mismatched loads. This is because the third wire now forms two tightly coupled transmission lines. It is unlike the Ruthroff version shown on the right in **Photo 7-A**. In his second article,[12] Turrin pointed out this important distinction.

J. Reisert, W1JR—1978

The next significant article on 1:1 Baluns was published by Reisert in 1978.[7] Reisert proposed winding some of the smaller (but still high-powered) coaxial cables around a 2.4-inch OD ferrite toroid with a permeability of 125. The windings also included a crossover, which is shown in **Figure 7-5** and **Photo 7-D**. In addition, he recommended various numbers of turns, depending upon the low-frequency requirement. For example, he suggested 12 turns to cover 3.5 MHz, 10 turns for 7 MHz, 6 for 14 MHz, and 4 for 21 and 28 MHz. Because the characteristic impedance of the coaxial cable is the same as the coax feedline, the Balun only introduces a foot or two of extra length to the feedline. This is true in the HF and VHF bands. The coaxial cables recommended in the article were RG-141/U, RG-142/U, and RG-303/U.

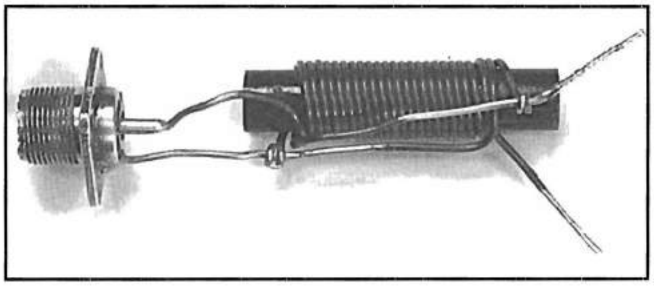

Photo 7-C. *A typical rod-type Balun.*

From the articles that followed in the amateur radio literature, it became apparent that few recognized all of the important features of Reisert's Balun. They are listed below.

1. An efficient, low-loss ferrite was used.

2. The Baluns had sufficient choking reactances for the various low-frequency requirements.

3. The characteristic impedance of the coiled transmission line was the same as that of the feedline, thus eliminating the extra transformer action of a length of transmission line with a different characteristic impedance.

4. The Balun is a form of Guanella's two-conductor 1:1 Balun, which is not prone to core flux and, hence, saturation and the generation of spurious frequencies. It is also not susceptible to mismatched and unbalanced loads as are the Turrin and "voltage" Baluns.

After constructing several of Reisert's Baluns and comparing them with other Guanella designs, I found that the crossover winding had virtually no effect up to 100 MHz (the limit of my equipment). In regards to Reisert's VSWR comparison with a rod-type Balun when feeding a triband Yagi beam on 20 meters, I found that his Balun had a lower VSWR (practically 1:1) at the best match point. The rod-type Balun had a best VSWR of about 1.3:1, but at a slightly higher frequency. He attributed the higher (and somewhat flatter) VSWR curve of the rod-type Balun to its greater ohmic loss. Because the rod-type Baluns I have investigated used the same low-loss ferrite that Reisert's did, I suspect that the differences in the VSWR curves were mainly due to the mismatch loss introduced by the rod-type Balun.

G. Badger, W6TC—1980

Badger published an in-depth, two-part series[13,14] in *Ham Radio* magazine in 1980 on air-core Baluns and Ununs. I am sure it was instrumental in advancing the technology of this class of wideband transformers. An article by Orr[15] also shows that there are many other radio amateurs who see the advantages of air-core transformers.

What are the claims for air-core Baluns over their ferrite-core counterparts? First and foremost, proponents of air-core Baluns claim they don't suffer the consequences of saturation that leads to spurious frequencies, heating, and ultimate damage. Secondly, it's said that they are not subject to arcing from the windings to the core.

What are the claims for the ferrite-core Baluns over their air-core counterparts? Simply put, they have wider bandwidths and are more compact.

After reading Badger's two-part series, I found that my curiosity was piqued by his experimental data on harmonic distortion due to saturation in a ferrite-core 1:1 Balun. Although many have expressed concerns regarding saturation in ferrite-core Baluns, Badger's data could very well provide the only results available. He used the two-tone test method, which combined two RF sources of 2.001 and 2.003 MHz, amplified it to 2 kW PEP and then fed it through a commercial 1:1 rod-type ferrite Balun. The data showed considerable distortion in the 3rd order and 9th order distortion products. In other words, appreciable nonlinearity took place at this high power level.

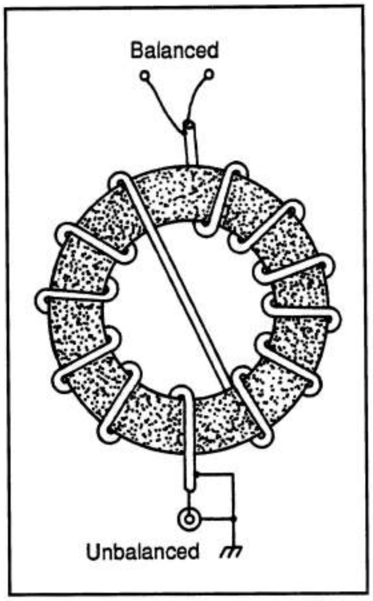

Figure 7-5. *A pictorial of the crossover used in Reisert's 1:1 Balun.*

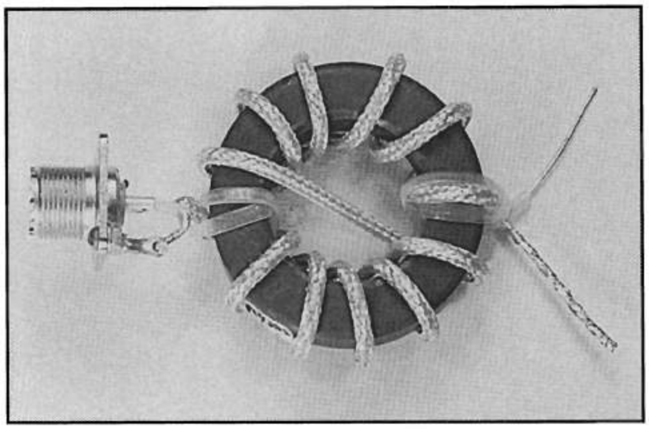

Photo 7-D. *A Reisert, W1JR, 1:1 Balun.*

Several questions come to mind regarding these measurements. What was the low-frequency response of the commercial 1:1 rod-type Balun Badger used? From my measurements on a rod-type Balun (**Figure 7-4**), I found a drop in the input impedance and an inductive component at 2 MHz. This indicates flux in the core and a problem when using this Balun at 2 MHz. Because many rod-type 1:1 Baluns have been used over the years, it would have been instructive if he had also made these measurements at 4 and 7

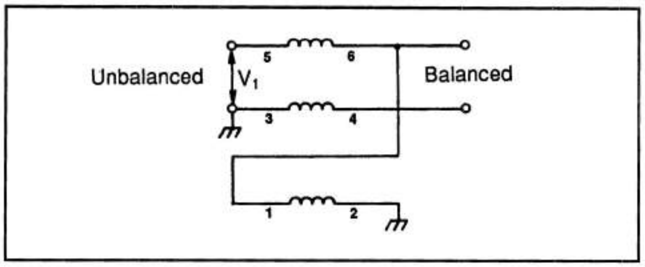

Figure 7-6. *Schematic of Badger's 1:1 Balun with a compensating winding (1-2). Winding (3-4) is the outer-braid of the coax and winding (5-6) is the inner conductor.*

MHz. They would have given the readers a safe lower-frequency limit for these Baluns.

I wondered why Badger didn't make similar measurements on Reisert's 1:1 Balun, which he included in his articles. As noted earlier, I consider Reisert's 1:1 Balun a very good design! I am sure that no distortion products would have been found at 2 MHz with this Balun. The end result is that Badger chose a poor ferrite-core design for making his comparisons. This helped to contribute to an undeserved reputation for the ferrite-core Balun.

Badger also suggested placing an insulated wire in parallel with the coax winding on Reisert's 1:1 Balun. He called this a compensating winding, which provided a superior balanced output. The third winding (1-2) is shown **Figure 7-6**. Winding (5-6) is the inner conductor and winding (3-4), the outer braid. Later experiments by myself and others have shown that a well-designed two conductor (Guanella) 1:1 Balun has a completely satisfactory balanced output for antenna applications. Furthermore, it does not suffer from an unbalanced and/or mismatched load and core saturation. Incidentally, Badger's schematic of **Figure 7-6** now adds up to four different versions of the 1:1 Balun. They are the two-conductor version of Guanella's and the three, three-conductor versions of Ruthroff's, Turrin's, and (now) Badger's.

Badger and Orr also mentioned the Collins Balun in their articles. This Balun is comprised of a dummy length of coax wound as a continuation of the original coiled coax winding. Interestingly, it is connected as a Ruthroff 1:1 Balun (**Figure 7-2B**), which also uses a third winding. Because there is appreciable coupling between the two coiled windings, the Collins Balun should also be susceptible to mismatched and/or unbalanced loads. Badger claimed it was, by far, the best 1:1 Balun he had ever used. Again, it would have been very informative if he had compared it with the Reisert Balun (without the compensating third wire).

M.W. Maxwell, W2DU—1983

One of the more significant articles on 1:1 Baluns was published by Maxwell[4] in 1983. Maxwell introduced, what he called, the "choke" Balun. It was formed by placing high-permeability ferrite beads over about one foot of small (but high-powered) coaxial cables similar to the ones used in the Reisert Balun. **Photo 7-E** shows the W2DU "choke" Balun removed from its plastic enclosure.

Maxwell compared his Balun with (what he termed) a "transformer-type" Balun by measuring the input impedances versus frequency when the outputs were terminated in 50 ohms. The "transformer-type" Balun didn't yield a true 1:1 impedance transfer ratio, which he claimed was due to losses, leakage reactance, and less than optimum coupling. Because Maxwell gave no description of the "transformer-type" Balun, I assumed it was the popular rod-type Balun shown in **Photo 7-C**. As you can see in **Figure 7-4**, this Balun has a poor low-frequency response. Furthermore, it is really optimized for a load of 47 ohms and not 50 ohms.

What Maxwell failed to realize was that his Balun was a form of Guanella's two-conductor type. That is, it is both a choke (a lumped element) and a transmission line (a distributed element). Additionally, Guanella's theory applies whether the transmission lines are coiled (about a core) or beaded, twin-lead or coaxial cable. From Ruthroff's classic paper,[9] which extended Guanella's work,[3] we became aware of the

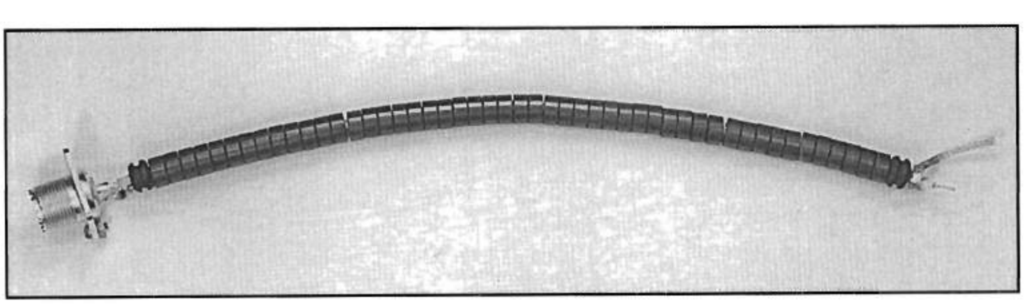

Photo 7-E. *The Maxwell, W2DU, "choke" 1:1 Balun.*

voltage drops along the lengths of the transmission lines. From very accurate insertion loss measurements,[2] we learned that the losses were mainly in the magnetic medium—and that they were related to the voltage levels and the permeabilities. Maxwell didn't take into account these latter findings. He used lossy high-permeability beads (2500) and assumed that the main loss was in the transmission line. He claimed that the CW power-handling capability of his Balun was 3.5 kW at 50 MHz and 9 kW at 10 MHz—the same as the coaxial cable itself. I seriously question these power ratings. Ironically, it is very likely that Maxwell's Balun had more real loss than the so-called "transformer-type" Balun!

R.W. Lewallen, W7EL—1985

There is very little doubt that Lewallen's interesting article[6] in 1985 contributed significantly to the better understanding and design of 1:1 and 4:1 Baluns. In it, he coined the (now very popular) terms "voltage" and "current" Baluns. The "voltage" Balun, a three-conductor type, has output ports which have voltages that are balanced to ground. It is brought about (see **Figure 7-6**) by the voltage-divider action of windings (5-6) and (1-2). Because we have two tightly coupled transmission lines in the passband with the same potential gradients, terminal 6 is at $+V_1/2$ and terminal 4 at $-V_1/2$ where V_1 is the input voltage. The "current" Balun, on the other hand, is a two-conductor Balun that produces equal and opposite currents on the output ports for any form of load impedance.

Lewallen conducted a series of experiments on 10 meters to compare the performances of "voltage" and "current" Baluns under balanced and unbalanced conditions. In the unbalanced (nonsymetrical) condition, the dipole was lengthened by five inches on one side and shortened by five inches on the other side. He then obtained a figure of merit for both Baluns (as well as for the case without a Balun) defined as the ratio of the average magnitude of the currents at the feedpoint over the magnitude of the imbalance (the inverted L) current. The magnitudes of the currents were obtained by current-probe toroids. Measurements were made at the antenna feedpoint and at a half-wave (physically) from it.

The "current" Balun consisted of 15 turns of very small RG-178/U coax on an FT82-61 core (a ferrite toroid with an OD of 0.825 inches and a permeability of 125). The "voltage" Balun had 10 turns of RG-178/U coax with a No. 26 wire in parallel (closely coupled) on the same toroid. The schematic is shown in **Figure 7-6**.

Lewallen concluded (and I agree) that his experiments clearly showed that the "current" Balun gave superior performance at every measured point in each experiment. However, the "voltage" Balun still improved the balance over the no-Balun case. He also concluded that other experiments should be performed in order to better compare the two forms of the Balun. One is to ascertain the difference when the feedline is placed nonsymetrical with respect to the antenna (to induce an imbalance current into the feedline). Others include determining the optimum point in the feedline to place the Balun, and testing the various kinds of core and beaded Baluns.

Although Lewallen's article pretty much speaks to Badger's proposal of adding a third wire to Reisert's Balun for better balance (i.e., avoid it), there are some comments and questions I have regarding his experiments and findings. They are:

1. Why didn't Lewallen use Reisert's Balun as the "current" Balun and Badger's suggested third-wire design as the "voltage" Balun? These would have been more realistic designs for comparisons. Instead, he used very small structures (which will only find use in QRP operations) and, as such, have higher frequency capabilities. Also, because the "voltage" Balun only had 10 turns (and hence a shorter transmission line and a poorer low-frequency response), it was favored in the comparisons on 10 meters. Had Lewallen used transmission lines of equal lengths on Reisert's cores, the differences between the two Baluns would have been even more dramatic.

2. It would also have been very useful if Lewallen had made comparisons between a "current" Balun that could handle the full legal limit of amateur radio power (again like the Reisert Balun) and "voltage" Baluns using Turrin's schematic (**Figure 7-3**) with rod and toroidal cores which have been readily available for nearly three decades.

3. Additionally, comparisons should not only be limited to 10 meters. Because 1:1 "voltage" Baluns are configurations of coupled transmission lines with various characteristic impedances, their performances with mismatched and unbalanced loads are more sensitive to the higher frequencies than their "current" Balun counterparts. Therefore, making similar measurements on 20 meters would also provide more useful information.

4. Even though Ruthroff's classic 1959 paper[9] has been the industry standard over the years, his 1:1

Balun design has been practically nonexistent in the amateur literature. Turrin mentioned its advantage over his first design in his second article.[12] But Turrin's first design has prevailed in our amateur literature. Because Ruthroff's design has the third conductor on a separate part of the toroid, it has the balanced output mentioned by Badger,[13] but still retains the flexibility of the Guanella[3] Balun. In other words, as the frequency is increased, the choking action of the third wire makes it practically transparent. This enables it to handle any form of load impedance. It would have been informative if Lewallen had pointed this out and also noted that Ruthroff's 1:1 Balun, although looking like a "voltage" Balun, is really a "current" Balun.

5. Lewallen and the others who have published in the amateur radio literature have failed to reference the first presentation on what are now known as "current" or "choke" Baluns, made by Guanella[3] in 1944. Even though Guanella used coiled transmission lines without a magnetic core, his theory on how these devices work is still applicable today.

J.S. Belrose, VE2CV—1991

The last article on 1:1 Baluns I considered worth mentioning was written by Belrose[5] in 1991. In it, he described the W2DU Balun by Maxwell and how his technique of threading coaxial cable through ferrite beads could be easily applied to 4:1 and 9:1 Baluns.

What immediately grabbed my attention in this article was the *deck head* (the editor's comments), which included highly complimentary remarks regarding the beaded-coax Balun. In essence, it said, "In this breakthrough article, W2DU's peerless 1:1 current-Balun design serves as the basis for excellent ferrite-bead-choke current Baluns capable of 4:1 and 9:1 impedance transformation."

However, if one reads the article carefully, it becomes apparent that this is not what Belrose said. His words were, "The *current Balun* of the *type* developed by Walt Maxwell, W2DU—a Balun consisting of ferrite beads slipped over a length of coaxial cable—is the best so far devised (italics mine)." He did not say that W2DU's Balun was "peerless." In fact, in the article he said just the opposite. He pointed out that the W2DU Balun's main disadvantage is that the beads are lossy at HF and heating becomes a concern when the transmitting power exceeds 125 watts! For high power (that is 1-kW CW), Belrose recommended Roehm's[16] designs, which use lower permeability (850) beads nearest the Balun's balanced output (where most of the heating takes place).

However, I do question two of the advantages he claims for the W2DU Balun. They are:

1. Its excellent power-loss and impedance-versus-frequency characteristics are much superior to those of a bifilar current Balun wound on a ferrite toroid.
2. It has excellent power-handling capability, and can function quite satisfactorily when working into highly reactive loads. This is so because the magnetic flux produced by currents flowing on this Balun's wires cannot saturate its ferrite beads.

Belrose obtained evidence for advantage number 1 by comparing the input impedance and power loss versus frequency of the W2DU Balun with a commercial Balun when they were terminated in 50 ohms. The commercial Balun was a bifilar wound toroidal type used in a differential-T tuner. What Belrose failed to realize was that the commercial Balun had heavily insulated wires, resulting in a characteristic impedance greater than 100 ohms. Thus, he was actually comparing a 50-ohm transmission line with a longer line that had a characteristic impedance in excess of 100 ohms! As expected, his input impedance versus frequency curve for the commercial Balun was even more severe than that of the Turrin Balun shown in **Figure 7-4**.

Advantage number 2 is based upon the premise that the magnetic flux produced by currents on the W2DU Balun's wires cannot saturate the ferrite beads, while the windings of a bifilar wound toroidal current Balun can. This is an incorrect assumption because the magnetic flux of a two-conductor type Balun, like the beaded-coax or the bifilar-wound toroidal Balun, is generated by the imbalance (inverted L) current and hence is *much* lower than the transmission line currents. This is especially true with sufficient choking reactances. The perception that the toroidal type Balun still transmits the energy to the output circuit by flux linkages, could very well lead to this mistaken impression.

For high power beaded-coax Baluns, Belrose referred to designs by Roehm,[16] who increased the power capability of this type of Balun by using lower permeability beads near the balanced output. He also increased the length considerably. For operation from 80 meters to 10 meters, Roehm used 28 inches of beaded coax. For 160 meters to 10 meters, he used 36 inches of beaded coax. Belrose's suggestion of connecting beaded coaxes in parallel on the low-imped-

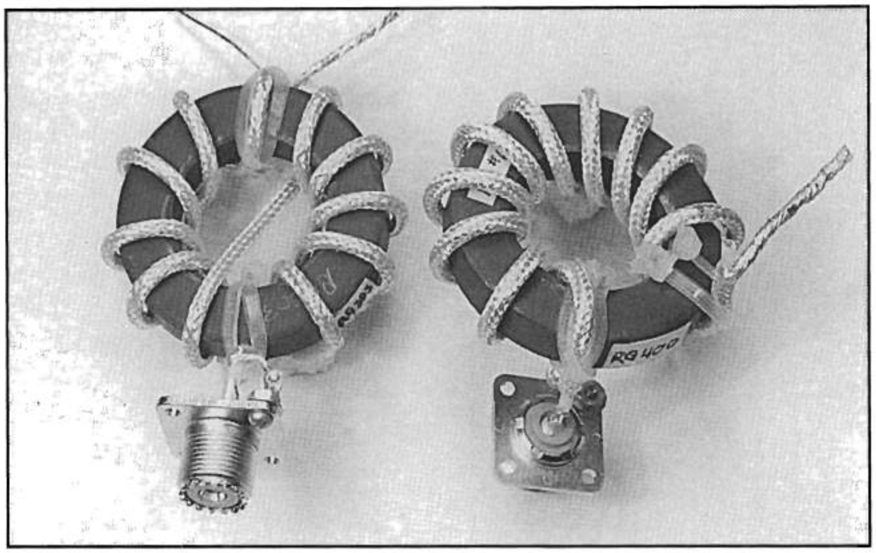

Photo 7-F. *Two versions of Reisert's 1:1 Balun. The Balun on the left uses the crossover shown in Figure 7-5. The Balun on the right is continuously wound. Both have the same electrical performance in the HF band.*

ance side and in series on the high impedance side to obtain a broadband 4:1 transformation ratio would require transmission lines with characteristic impedances of 100 ohms. This means, for a high power 4:1 Balun using beaded transmission lines, about 56 inches of beaded line would be required for the 80-meter to 10-meter operation and 72 inches for the 160-meter to 10-meter coverage. For a 9:1 Balun, it would be necessary to increase these lengths by 50 percent!

It remains to be seen what Belrose would have said or done if he had compared the W2DU Balun of Maxwell's with the W1JR Balun of Reisert's. He certainly couldn't claim the advantages listed in his article for the W2DU Balun. Would he still have claimed that the type of Balun developed by Maxwell is the best so far devised? Given the evidence, I doubt it.

Sec 7.4 High-, Medium-, and Low-power Designs

In this section I'll present my latest 1:1 Balun designs. Except for one Balun that appeared in the June 1993 issue of *CQ*, the others appeared for the first time in the magazine's April 1994 issue. Because I have favored Reisert's design throughout this chapter, the first Baluns described here are my versions of his technique of coiling small (but high power) coaxial cable around a low-permeability ferrite toroid. For my wire versions, I could have used all sorts of adjectives to describe them like Guanella, two-conductor, choke, and current. However, in the process of writing this section, I thought Belrose's adjectives were the most direct. Using his words, I call my wire versions of the 1:1 Balun simply—*bifilar toroidal Baluns*.

Photo 7-F shows two versions of Reisert's Balun. The one on the left uses the crossover shown in **Figure 7-5**. Because no difference in performance at HF was noticed without the crossover, a continuous-wound version is also shown on the right. The main advantage in the HF band with the crossover winding is purely mechanical. Having the input and output connections on opposite sides of the toroid is not only more convenient, but it also offers a much stronger method of mounting.

For operation from 1.8 to 30 MHz, 10 turns of small coax like RG-303/U, RG-142B/U, or RG-400/U are wound on a 2.4-inch OD ferrite toroid with a permeability of 250. If the use is limited from 3.5 to 30 MHz, then a permeability of 125 is recommended because it would yield a slightly higher efficiency at the high end. If one wants the highest possible efficiency and limits the operation from 14 to 30 MHz, then a permeability of 40 is recommended. With loads grounded at their centers, these conditions were found to give ample margins (handle a VSWR of 3:1 without any appreciable flux) at their low frequency ends.

For ease of winding, I found TY-RAP™ Cable Ties *very* useful. Two were used at each end. Removing the covering on the outer braid also helps. Because about 24 inches of cable is wound on the toroid, I recommend you start with at least 32 inches. Of the three cables noted above, I found RG-303/U cable the easiest to wind and connect. Although it only has a single-thickness braid (the others have double-thickness braids), its power rating is still the same—9 kW at 10 MHz and 3.5 kW at 50 MHz.

The next high power design is shown in **Photo 7-G** mounted in a 4 inch long by 3 inch wide by 2.25 inch

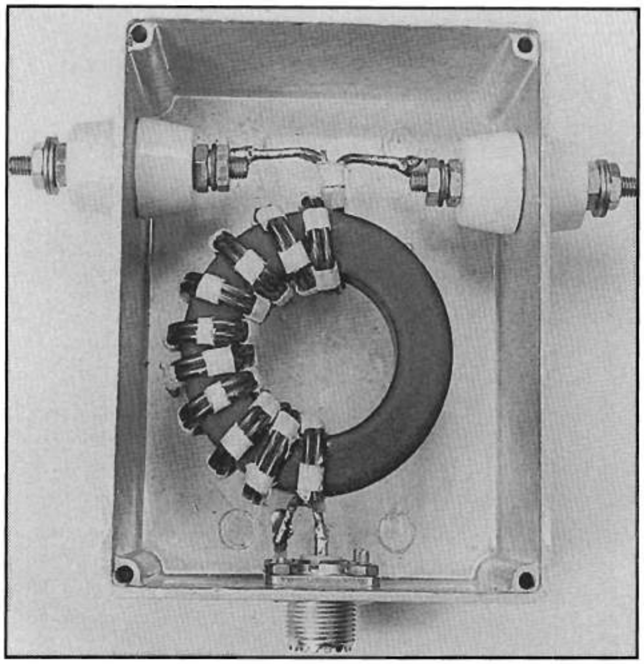

Photo 7-G. *My high-power design of a bifilar toroidal (Guanella/current) 1:1 Balun mounted in a 4 inch long by 3 inch wide by 2.25 inch high Bud aluminum box.*

high Bud CU 234 aluminum box. It has 10 bifilar turns of No. 12 H Thermaleze wire on a 2.4-inch OD ferrite toroid. As with the Reisert versions, a permeability of 250 is recommended for 1.8 to 30 MHz, 125 for 3.5 to 30 MHz, and 40 for 14 to 30 MHz. One wire is also covered with two layers of Scotch No. 92 polyimide tape in order to raise the characteristic impedance to 50 ohms. With this added insulation, the voltage breakdown of this twin-lead transmission line compares very favorably with RG-8/U cable (4000 volts). In order to preserve the spacing, the wires are also clamped together about every 1/2 inch with strips of Scotch No. 27 glass tape 3/16 inches wide and a little over 1 inch long.

Two "economy" versions of the high power bifilar toroidal Balun are shown in **Photo 7-H**. The one on the left shows the windings crowded on one-half of the toroid. The one on the right provides the same positions of the input and output connections by using the crossover. Their performances are identical. Both Baluns have 10 bifilar turns of No. 14 H Thermaleze wire on a 2.4-inch OD ferrite toroid. The choices of permeability, which trade off bandwidth for efficiency, are the same as those used in the two previous high power designs. The word "economy" here refers to economy in labor—this Balun is actually very easy to construct.

This "economy" Balun, which also handles the full legal limit of amateur radio power, has a small tradeoff in high frequency response. Because no extra insulation is used, the characteristic impedance of two tightly clamped No. 14 H Thermaleze wires is 45 ohms. With one layer of Scotch No. 92 tape, it increases to 50 ohms. But for most of the HF band, the difference in performance between Baluns using transmission lines of 45 and 50 ohms should be negligible. Even without the extra insulation, the voltage breakdown should compare very favorably with the smaller, high-power coaxes used in the Reisert versions (1900 volts).

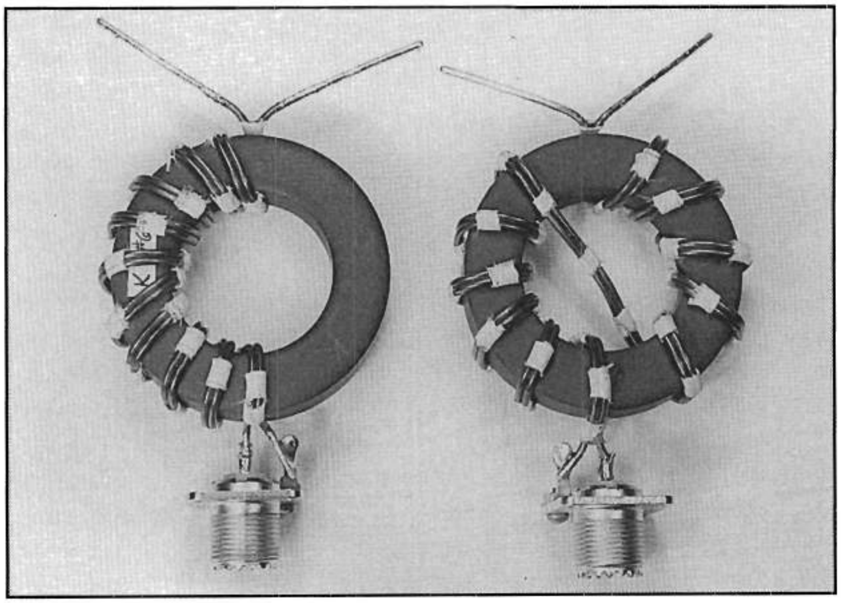

Photo 7-H. *Two "economy" versions of the high-power bifilar toroidal (Guanella/current) 1:1 Balun. The one on the right uses Reisert's crossover technique.*

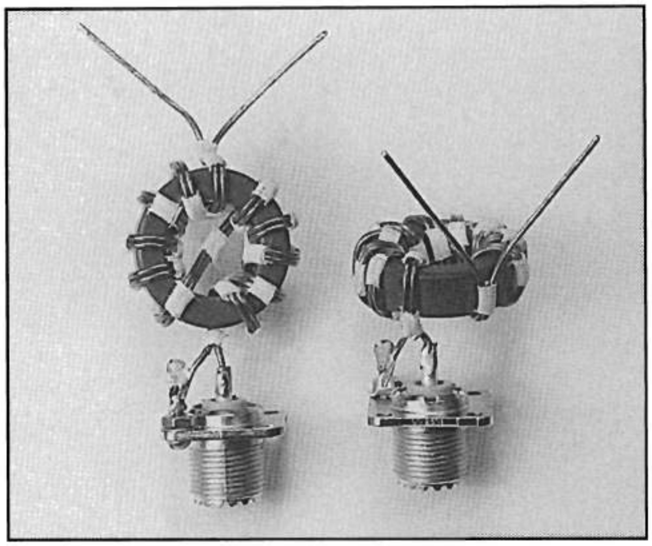

Photo 7-I. *Two low-power versions of the bifilar toroidal (Guanella/current) 1:1 Balun capable of handling the output of practically any HF transceiver. The Balun on the left has the crossover.*

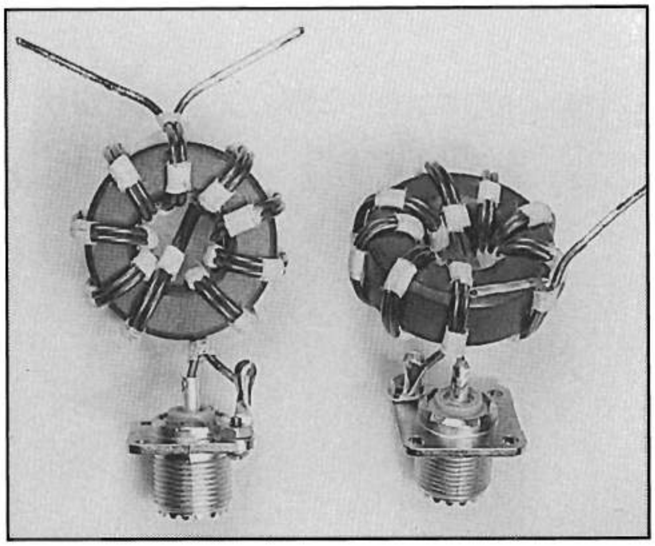

Photo 7-J. *Two medium-power versions of the bifilar toroidal (Guanella/current) 1:1 Balun capable of handling the full legal limit of amateur radio power when the VSWR is less than 2:1. The Balun on the left has the crossover.*

Photo 7-I shows two low-power versions of a bifilar toroidal Balun capable of handling the output of practically any HF transceiver. One has a crossover winding and the other, a continuous winding. They both have 10 bifilar turns of No. 16 H Thermaleze wire on a 1.25-inch OD ferrite toroid with a permeability of 250. Since efficiency is not a major problem in low-power use, I found no reason to suggest the other two versions, which use lower permeabilities. It is also interesting to note that two tightly clamped No. 16 H Thermaleze wires have a characteristic impedance close to 50 ohms. Therefore, this small Balun (particularly with its short leads) has a very good high frequency response.

Photo 7-J shows two medium-power versions of a bifilar toroidal Balun capable of handling the full legal limit of amateur radio power under controlled conditions—when the VSWR is less than 2:1. Being smaller than its larger (2.4-inch OD) counterpart, its heat-sinking capability and, hence, power rating is less. As before, one Balun uses a crossover while the other doesn't. Each has 8 bifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid. The ferrite permeabilities and expected bandwidths are the same as the other high-power Baluns. Because the average magnetic path length in the core is about two-thirds that of the 2.4-inch core, only 8 bifilar turns are required in order to produce a similar low-frequency capability. Even though the characteristic impedances of their bifilar windings are 45 ohms, their responses on 10 meters should be somewhat better than the "economy" models because the lengths of their transmission lines are shorter (18 compared to 24 inches).

And now a few words on what sort of efficiency one can expect in trading-off low frequency response by using lower permeability ferrite cores. From earlier studies,[2] it was found that the efficiency (with sufficient choking, so only transmission line currents flow) is related to the permeability, the voltage drop along the length of the transmission line, and the frequency. The higher the permeability and/or voltage drop, the greater the loss. Additionally, the higher the permeability, the greater the loss with frequency. It was also found that a permeability of less than 300 was necessary in order to obtain the very high efficiencies of which these devices are capable.

Here are some efficiencies that might be expected from ferrites under matched conditions, based on the results of the studies:

1. With 250 material, an efficiency near 99 percent at 1.8 MHz and 97 percent at 30 MHz.
2. With 125 material, an efficiency near 99 percent at 3.5 MHz and 98 percent at 30 MHz.
3. With 40 material, an efficiency of 99 percent at 14 MHz and at 30 MHz.

When a Balun is exposed to a high impedance resulting in a VSWR of 2:1, the voltage, and hence loss, increases by about 40 percent. With a VSWR of

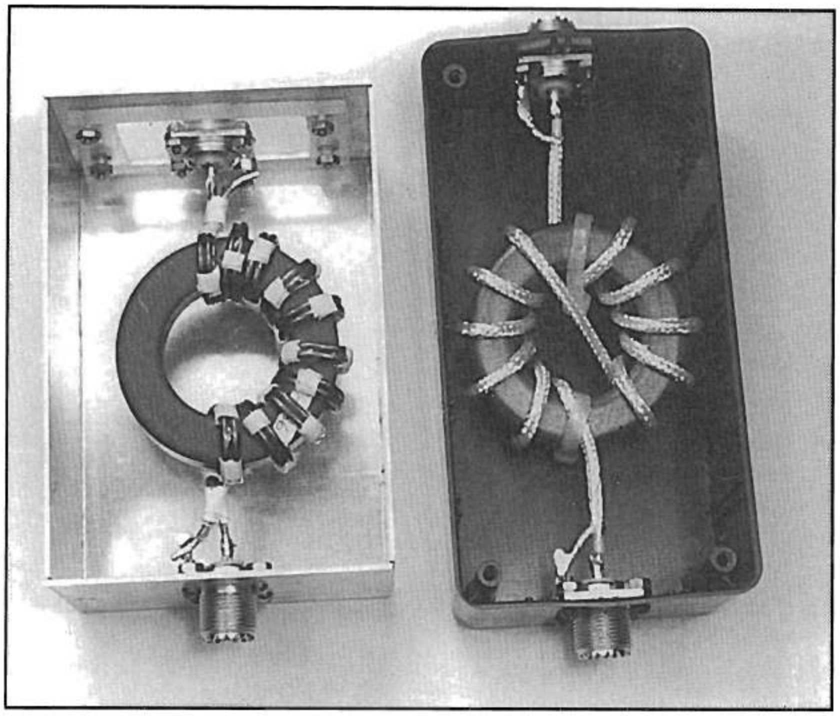

Photo 7-K. *Two examples of mounting the high-power 1:1 Guanella Balun in enclosures that allow isolation transformer operation.*

4:1, the loss doubles. With a VSWR of 10:1, the loss is more than three-fold. Since limited data was obtained in this study,[2] these increases in losses with increases in VSWR could very well be even greater.

Sec 7.5 Isolation Transformers

As was shown in the preceding sections, the 1:1 Guanella (current/choke) Balun is in reality a two-conductor RF choke, or an isolation transformer. However, practical Balun designs don't usually lend themselves to isolation use for mechanical reasons. They cannot be conveniently inserted into a coaxial cable system. Therefore, the objective is to mount the "Guanella Baluns" in enclosures that will allow them to perform as isolation transformers.

Photo 7-K shows two examples of enclosures that should accomplish the mission. The example on the left shows one of the high-power units mounted in a 5 inch long by 3.5 inch wide by 2.25 inch high aluminum enclosure. The output connector (on the top) is insulated by a 0.125-inch-thick piece of plastic. The one on the right shows another high-power unit mounted in a 6 inch long by 3 inch wide by 2 inch high plastic enclosure, which is available from RadioShack. The descriptions of the high-power units are given in the preceding section.

These isolation transformers can be used as Baluns when inserted in a coaxial cable one-half wavelength (physically) from a half-wave dipole[6] or between a coaxial cable and a balanced L-C tuner.[16] They should never be inserted in a coaxial cable system that presents a high impedance and, consequently, a high voltage. This could be harmful to the unit.

Note that the high-power unit on the left in **Photo 7-K** can also be used as a filament choke in a Class B linear amplifier. Its low-frequency response is much better than any rod-type, assuring 160-meter operation. Obviously, it would not be mounted in the aluminum enclosure shown in the photo. In addition, the two layers of Scotch No. 92 tape would not be required because the characteristic impedance of the winding is unimportant and the voltages involved are quite low.

By the way, the designs discussed above can be used as phase-inverters. Simply connect the hot lead on one side to the ground on the other side.

Sec 7.6 Summary

In investigating the 1:1 Balun, I was quite surprised to see the ferrite- and powdered-iron-core designs that have been available in the literature and off-the-shelf since 1964. They not only had poor low- and high-frequency responses, but they were also susceptible to flux in the cores at their low-frequency ends. Furthermore, since they only used single-coated wires, they were also prone to voltage breakdown. No doubt, these designs are responsible for the poor reputation that the Balun has had for many years.

It wasn't until 1978, when Reisert published his article, that a Balun which had all of the attributes of a good design became available. Namely:

a) It is efficient because it uses a low-permeability core.

b) It has sufficient choking reactance to meet its low-frequency requirement.

c) It is not prone to flux in the core (and, hence, saturation) because it has no third winding.

d) It has a 50-ohm characteristic impedance and maintains a 1:1 transformation ratio with a 50-ohm load.

e) It has a good voltage breakdown capability (1900 volts).

f) It can handle a mismatched and/or unbalanced load.

However, succeeding investigators failed to see the advantages of Reisert's design and proposed their own. Surprisingly, they belonged to two distinct groups. One group favored "air-core" Baluns and the other, "choke" (beaded-coax) Baluns.

The main argument given by the "air-core" followers was that their Balun would never experience problems with saturation, while the "ferrite-core" Balun would. However, the Reisert Balun is a current/choke type Balun that could only have flux in the core due to the imbalance (inverted L) current, which is much smaller than the transmission line currents. In fact, with any degree of choking reactance by the coiled transmission line, the imbalance current is essentially negligible. Therefore, saturation is not a concern with a Balun like Reisert's. In all fairness, it should be noted that it is a different story with the 4:1 current/choke and voltage Baluns. All three of these types of Baluns have a "magnetizing inductance" in their low-frequency models and hence a possibility of saturation with a poor design.

The advocates of the "choke" 1:1 Balun claim that their beaded-coax Balun can't saturate while the bifilar (current) toroidal Balun can. This is untrue; they are basically the same kind of structure, and neither has a third conductor which could allow a flux-causing current at the very low-frequency end. But of all of the attributes listed above for the Reisert Balun, the first one has the "choke" Balun at a disadvantage in the HF band. Because its transmission line is not coiled about a toroid, it does not have the multiplication factor of N^2 (due to mutual coupling), where N is the number of turns, while the toroidal Balun does. Therefore, higher-permeability beads are required in order to obtain sufficient choking reactance. This results in lower efficiency.

Finally, it should be pointed out that the 1:1 Balun using the small, but high-power, coaxial cable is capable of 5-kW operation in the HF band. Because of current crowding, the bifilar toroidal Balun, even with No. 12 wire, has shown excessive wire heating at this level. Therefore, No. 10 or No. 8 wire, with added layers of Scotch No. 92 tape in order to obtain a 50-ohm characteristic impedance, is recommended.

Chapter 8

The 4:1 Balun

Sec 8.1 Introduction

Chapter 7 in this book discussed the most popular of all Baluns—the 1:1 Balun designed to match 50 ohms unbalanced to 50 ohms balanced. It not only gave a review of the history, theory, and design of these broadband transformers, but also my viewpoint on published articles advocating later designs using coaxial cable wound around a toroid, threaded through ferrite beads, or just plain coiled in air (an air-core Balun). As was noted, I am in considerable disagreement with many of the claims advanced for these later 1:1 Baluns.

Chapter 8 deals with the next most popular Balun —the 4:1 Balun designed to match 50 ohms unbalanced to 200 ohms balanced. Baluns matching 50 ohms unbalanced to 12.5 ohms balanced are also included. It begins with a little information on the history and design of these Baluns, followed by high- and low-power designs and comparisons with other Baluns that have been on the market or in the amateur literature. Guanella's approach to the design of a 4:1 Balun is particularly noteworthy. It not only yields an excellent Balun design, but also an Unun (unbalanced-to-unbalanced transformer) with practically the same performance. His design could very well be called a *Balun/Unun*. As in the first chapter, this one also closes with a brief summary of the significant points.

Sec 8.2 A Little History and Design Information

There are really only two classic papers that have established the principles upon which the transmission line transformer (the Balun being a subset thereof) is based. The first one was written by Guanella in 1944. Guanella proposed the idea of coiling a transmission line to isolate the input from the output, resulting in the (now popular) *current* or *choke* Balun.[3] The second was by Ruthroff in 1959, whose analysis of these transmission line transformers is the present industry standard.[9] Ruthroff also introduced the Unun and the hybrid transformer.

Interestingly enough, both Guanella and Ruthroff had different approaches to their 1:1 and 4:1 Balun designs. Guanella used a two-conductor 1:1 Balun design, while Ruthroff used a three-conductor design. Ruthroff's third conductor (which was said to increase the low-frequency response over the two-conductor Balun[12]) lay on a separate part of a toroidal core. Investigators who followed failed to recognize this fact. Their comparisons were made with a three-conductor Balun that had the third wire in parallel with the other two, which then formed two coupled transmission lines. This gave rise to the term *voltage* Balun—an inferior design.[6]

However, the differences between Guanella's and Ruthroff's approaches to 4:1 Baluns were even more striking. Guanella connected coiled transmission lines in a parallel-series arrangement, so in-phase voltages were summed at the high impedance side. His Balun has been called a *current* Balun.[6] Ruthroff, on the other hand, obtained a 4:1 transformation ratio by summing a direct voltage with a delayed voltage that traversed a single transmission line in a phase-inverter connection (see **Chapter 7**). His Balun has been called a *voltage* Balun. The distinction between the operation of these two Baluns was also overlooked by practically everyone who followed.

This section reviews the two different approaches taken by Guanella and Ruthroff in obtaining 4:1 Baluns. Of particular importance are the descriptions of the potential drops along the lengths of the transmission lines when the loads are grounded at various points. These voltage drops are not only relevant to the ohmic losses in the Baluns, but also to their electrical performances. These descriptions were quite

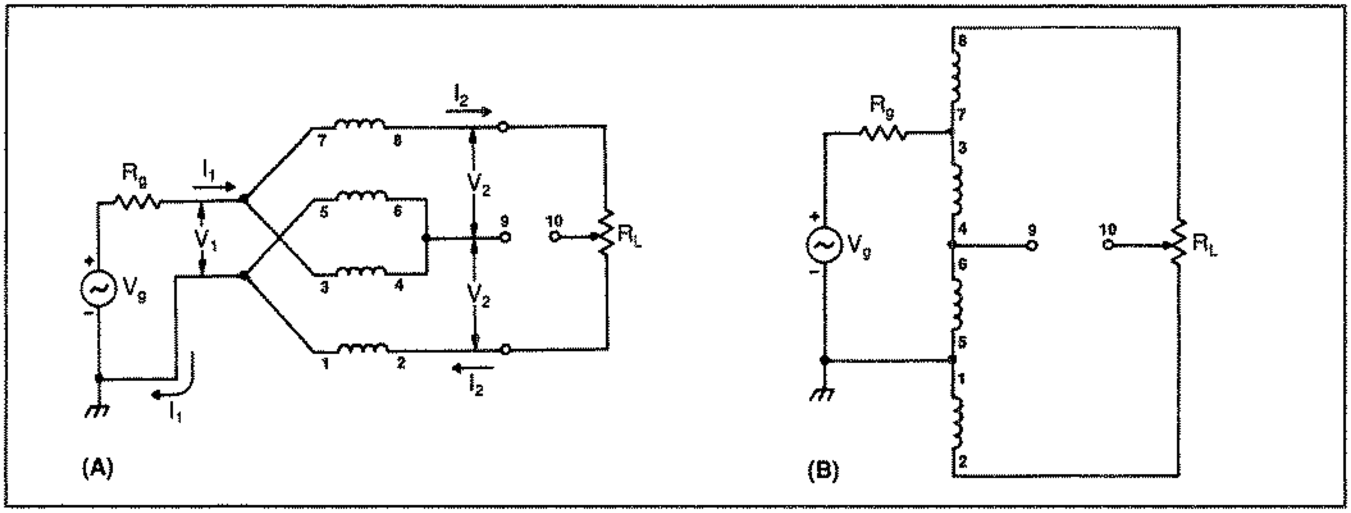

Figure 8-1. *Electrical models of the Guanella 4:1 Balun: (A) high-frequency, (B) low-frequency.*

possibly presented for the first time in the second edition of my book *Transmission Line Transformers.*[2]

Sec 8.2.1 Guanella's 4:1 Balun

Figures 8-1A and **B** show the high- and low-frequency models of Guanella's method of connecting transmission lines in parallel-series to obtain a 4:1 Balun. The high-frequency model (**Figure 8-1A**) assumes that the choking reactances of the coiled (or beaded) transmission lines are sufficient to isolate the input from the output, so only transmission line currents are allowed to flow. This occurs when the reactance of windings 3-4 and 5-6 (which are in series) is much greater than R_g (at least by a factor of ten.[2] If two cores are used, the reactance is the sum of the reactances of windings 3-4 and 5-6. If a single core is used, the reactance is twice as large because of the mutual coupling between the windings. The other advantage of Guanella's method (besides only using one core) is that shorter transmission lines can be used, resulting in better high-frequency performance.

As with all transmission line transformers, the objective is to have the transmission lines see loads equal to their characteristic impedances, resulting in "flat lines." This yields the highest frequency response. Because each transmission line in **Figure 8-1A** sees one half of the load, R_L, the optimum value of the characteristic impedance is $R_L/2$. Consequently, the input impedance, V_1/I_1, is simply the impedance of two identical transmission lines connected in parallel. It follows that the impedance transformation ratio is the load, R_L, divided by the input impedance.

Because the Guanella 4:1 Balun sums voltages of equal delays from identical transmission lines, his Balun is only limited in high-frequency performance by the deviation of the characteristic impedance of the transmission lines from the optimum values and the parasitics not absorbed into the characteristic impedance of the lines. I (and practically everyone else) had overlooked the simple and important statement, "a frequency independent transformation," which appeared in Guanella's 1944 article[3]—a fact that is evidenced by the scarcity of his designs in the literature. Another interesting aspect of the Guanella 4:1 Balun is the analysis of his Balun when the load is floating or grounded at different points. This leads to the determination of the voltage gradients that exist along the transmission lines and the various functions of which his 4:1 design is capable. Assuming a matched load or very short transmission lines resulting in $V_2 = V_1$, they are as follows.

Floating Load

With terminal 10 (which is at the center of R_L) floating, the potential gradient along the top transmission line in **Figure 8-1A** (windings 5-6 and 7-8) is $-1/2V_1$; along the bottom transmission line (winding 1-2 and 3-4) it's $-3/2V_1$. The voltage to ground on terminal 9, V_{90}, is $-1/2V_1$. Because the bottom transmission line (in **Figure 8-1A**) has a voltage drop along its length three times greater than the top transmission line, it results in three times more loss because losses in transmission line transformers are voltage dependent (dielectric-type losses).[2]

Even though a single-core Guanella 4:1 Balun maintains the voltages (stated above) when feeding a folded dipole (of about 200 ohms) which has a virtual-ground potential at terminal 10, it still feeds equal currents to each side of the antenna because of the series-connection at its output. Since the output voltages are not balanced to ground, a reactive component is probably introduced into the input impedance. Additionally, the choking reactance of the windings also prevents antenna currents from flowing on the outside of the coaxial cable feedline.

Load Grounded at Center

When two cores are used and terminal 10 (the center of R_L) is grounded, the voltage gradient along the top transmission line in **Figure 8-1A** is zero, and along the bottom transmission line it is $-V_1$. The voltage to ground on terminal 9 (V_{90}) is also zero. In fact, the core for the top transmission line isn't needed. It merely acts as a mechanical support for the top transmission line, which now only operates as a delay line. Also, all of the loss now occurs in the core of the bottom transmission line where a longitudinal potential gradient exists. Furthermore, the low-frequency response, as seen from **Figure 8-1B**, is now determined by the reactances of windings 1-2 and 3-4. This means that the low-frequency response with a floating load is better by a factor of two over the case where the load is grounded at its center.

But the single-core case is a different matter. Since the potential at terminal 9 (V_{90}) wants to be at $-1/2V_1$, connecting a ground directly to the center of R_L causes an imbalance that renders the single-core Balun unusable. If the ground were placed at a point 25 percent below terminal 8 (50 ohms from terminal 8 with a 200-ohm load), no difference would be noted from a floating load. This condition also exists when two cores are used.

Load Grounded at the Bottom

It is when the load is grounded at the bottom (at terminal 2), that we have what is probably the most interesting case. The 4:1 Balun (with two cores) is now converted into a very broadband Unun (unbalanced-to-unbalanced transformer). Because the bottom transmission line in **Figure 8-1A** has no potential drop along its length, it only acts as a delay line. The voltage to ground at terminal 9 (V_{90}) is $+V_1$, and the voltage gradient along the top transmission line is $+V_1$. This results in a voltage of $2V_1$ across the load. The low-frequency response is now determined by the reactances of windings 5-6 and 7-8. This is just the opposite of the Balun case when the center of the load was grounded. A single-core 4:1 Guanella Balun can also be converted to an Unun by putting a 1:1 Balun (for isolation) in series with the 4:1 Balun.[2]

There is a reason for claiming a very broadband response for a Guanella Unun (converted from his Balun)—the two in-phase voltages are now summed at the high-impedance side. The only other competition for a 4:1 Unun design is that of Ruthroff's,[9] where a direct voltage is summed with a delayed voltage that traversed a single transmission line (and, hence, had a built-in, high-frequency cut-off). In fact, very little information can be found in the literature on a *Guanella 4:1 Unun.*

Sec 8.2.2 Ruthroff's 4:1 Balun

Figures 8-2A and **B** show the high- and low-frequency models of Ruthroff's approach for a 4:1 Balun. The high-frequency model (**Figure 8-2A**) assumes that the

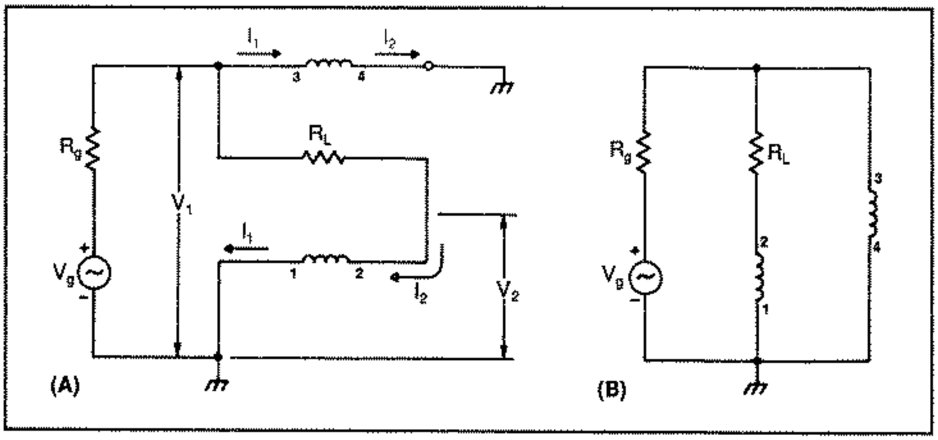

Figure 8-2. *Electrical models of the Ruthroff 4:1 Balun: (A) high-frequency, (B) low-frequency.*

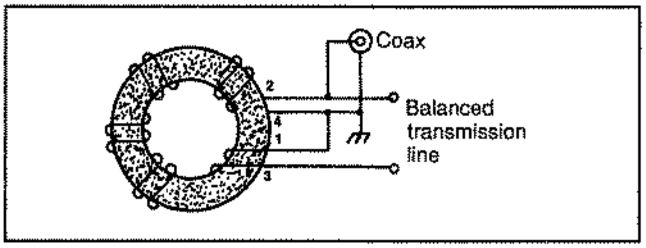

Figure 8-3. *A pictorial of the connections for a 4:1 Ruthroff (voltage) Balun.*

choking reactance of the coiled (or beaded) transmission line is sufficient to isolate the input from the output, so only transmission line currents are allowed to flow. This occurs when the reactance of winding 3-4 (or 1-2, because they are the same) is much greater than R_g (at least by a factor of ten[2]).

As **Figure 8-2A** indicates, the transmission line is connected in a *phase-inverter* function (see **Chapter 7**). That is, a $-V_1$ voltage gradient now exists along the length of the transmission line. Therefore, the voltage across R_L now becomes $V_1 + V_2$. Although Ruthroff analyzed his 4:1 Unun in his classic paper,[9] his results also apply to his Balun with a floating load, because both devices sum a direct voltage with a delayed voltage. In essence, he used loop equations on the input and output and transmission line equations to eliminate one set of variables (I_2 and V_2). Ruthroff also used a maxima technique (setting a derivative to zero) to solve for the optimum characteristic impedance of the transmission line. As in the Guanella case, he found the optimum value to be $1/2R_L$.

An inspection of **Figure 8-2A** shows that the left side of R_L (terminal 3) has a direct voltage (V_1) to ground, and the right side (terminal 2) a delayed voltage ($-V_2$) to ground, which traveled the length of the transmission line. Also note, that if the line is electrically one-half wavelength long, the output is *zero*. As a result, Ruthroff's design (which has a built-in cutoff) is sensitive to the transmission line length.

I have recently unearthed another interesting aspect of Ruthroff's design.[2] If the center of the load is grounded, the high-frequency performance is vastly improved. The built-in high-frequency cut-off is eliminated and the Balun appears to take on the character of a Guanella Balun that sums voltages of equal delays. A closer inspection reveals that the input impedance now consists of two impedances in parallel: one consisting of $R_L/2$, and the other of a "flat line" terminated by $R_L/2$. As a result, the currents are not in phase!

Surprisingly, when matching into a folded dipole with an input impedance of 200 ohms, Ruthroff's Balun exhibits a high frequency response that is much greater than expected (see later sections). Because the folded dipole has a *virtual* ground at its center, the Balun could very well be summing voltages of equal phases.

Sec 8.2.3 Amateur Radio History and Design

Looking back at old issues of amateur radio handbooks (I don't have a complete set), I found that the first presentation on broadband 4:1 Baluns appeared in the 1955 edition of *The ARRL Handbook.* The section was called "Coil Baluns." The schematic diagram was that of Guanella's, shown in **Figure 8-1A**. What surprised me was that this section appeared to use many of the important words contained in Guanella's article.[3] It mentioned that the choking action of the coiled transmission lines should be great enough to isolate the input from the output at the lowest frequency of interest. It also included the requirement on the characteristic impedance of the coiled transmission lines; namely, that the characteristic impedance should be equal to $R_L/2$, where R_L is the load.

However, the section also included two other statements which are not correct in light of today's design practices. One recommended that the length of the winding in each coil be equal to about a quarter wavelength. The other stated that the principal application is in going from a 300-ohm balanced load to a 75-ohm coaxial line. With magnetic cores, the lengths of the windings are now considerably shorter than a quarter wavelength, and the applications include a host of different impedance levels.

Recent issues of the handbooks now include the 4:1 broadband *coil Balun* (along with the same write-up that appeared in the 1955 issue), and one with windings on ferrite cores. They are now called 4:1 air-core current Baluns and "just plain" 4:1 current Baluns (ferrite cores being assumed). Unfortunately, what is lacking in the description of the 4:1 current Balun is information on the importance of the characteristic impedance of the windings and the value of the permeability of the ferrite cores. The literature states that 8 to 10 turns (of No. 14 Formvar-coated, close-spaced—I guess) on a toroidal core or 10 to 15 turns on a rod are typical values for the HF range. Ferrites with permeabilities from 850 to 2500 are also suggested. Nothing is mentioned regarding the dimensions of the cores.

In essence, there's very little information available today in our handbooks that would help one understand and construct the "popular" *current* Balun. Even the choices of recommended ferrites to be used are found wanting. Accurate loss measurements^2 have shown ferrites with permeabilities of 850 to 2500 to be lossy in Balun and Unun applications. Only when the permeabilities of ferrites are 300 or less will Baluns and Ununs exhibit the very high efficiencies of which they are capable.

Even though the 4:1 *voltage* Balun has actually had a shorter history than the *current* Balun, considerably more construction detail (including an actual photograph) has been available in the amateur radio handbooks. As far as I can tell, the first presentation took place between 1965 and 1968. In looking through succeeding issues (including the 1993 issue), I find the write-up hasn't changed much (if any) over the years. The 4:1 Balun from the handbooks and a commercial rod-type are described and compared with my designs in the succeeding sections.

Sec 8.3 4:1 Ruthroff Balun Designs

As mentioned above, the schematic of Ruthroff's Balun is shown in **Figure 8-2A**. A pictorial representation is shown in **Figure 8-3**. **Photo 8-A** (on the left) shows my construction of a design close to the one shown in the handbook's photographs. It has 10 close-spaced, bifilar turns of No. 14 Formvar-coated wire on a 2.4-inch OD ferrite toroid with a permeability of 40. **Figure 8-4** shows a plot of the input impedance versus frequency when the 200-ohm load is center-tapped-to-ground (which is close to the actual case when matching into balanced antenna systems). As you can see, when compared to a design that has the proper characteristic impedance of the winding and sufficient choking, the response is very poor. Although this Balun has been rated at 1000 watts of RF power from 1.8 through 60 MHz, I would suggest it not be used below 6 MHz for fear of excessive flux in the core (especially when the magnitude of the load is greater than 200 ohms). Also, above 14 MHz, the transformation becomes considerably greater than 4:1.

My design, on the right in **Photo 8-A**, has 14 bifilar turns of No. 14 tinned copper wire on a 2.4-inch OD ferrite toroid with a permeability of 125 or 250. The wires are threaded through No. 13 Teflon tubing with a wall thickness of 20 mils. As you can see by its excellent high frequency response in **Figure 8-4**, the characteristic impedance of the bifilar winding must be very close to the ideal value of 100 ohms. **Photo 8-B** shows two different views of my design mounted in a 4 inch long by 3 inch wide by 2.25 inch high Bud CU 234 aluminum box. The Balun, which is placed equidistant between the top, bottom, and sides of the enclosure, is securely mounted by soldering its leads to the two feedthrough insulators and the SO-239 chassis connector.

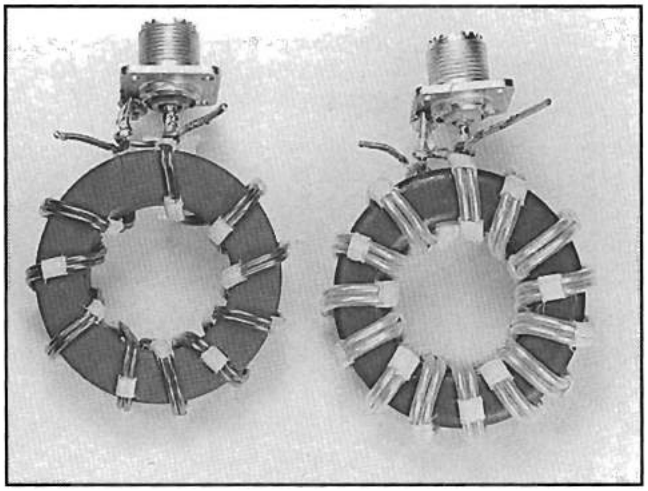

Photo 8-A. *Two designs of the 4:1 Ruthroff (voltage) Balun. The one on the left is taken from the amateur radio handbook. The one on the right is my improved version.*

It should be mentioned that if the Balun is to be used mainly on the lower portion of the HF band (including 160 meters), then the 250 permeability ferrite is recommended. Even though the difference in low frequency response between permeabilities of 125 and 250 doesn't show up in **Figure 8-4**, the 250 permeability would provide an extra safety margin (from flux in the core) at the low frequency end. The trade-off lies in giving up a little in efficiency (about 1 percent) for an increase in the safety margin (a factor of 2) at the low end.

Incidentally, the handbooks also state that the Balun can be used between a balanced 300-ohm point and a 75-ohm unbalanced line. Because I suspected this statement as well, I again measured the input impedances versus frequency of both Baluns when terminated in a 300-ohm center-tapped-to-ground load. **Figure 8-5** shows the deterioration that takes place, especially at the high end. Even a Balun that is well designed for a 50:200-ohm impedance level is not recommended

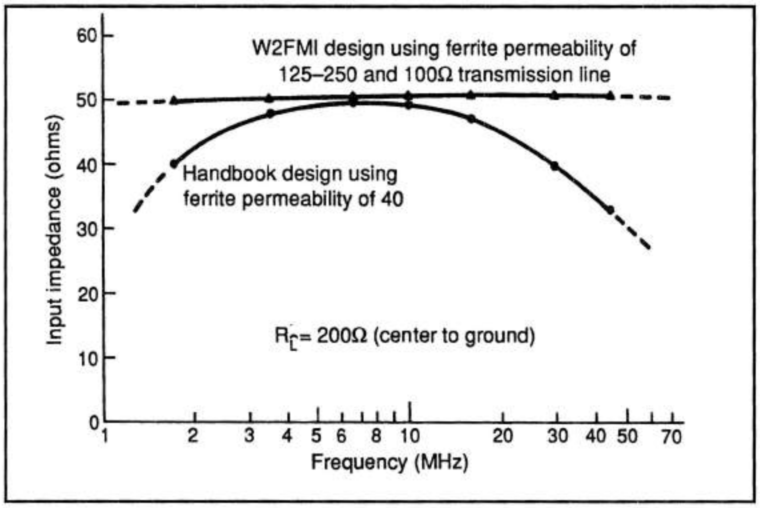

Figure 8-4. *The input impedance versus frequency for a 4:1 Ruthroff (voltage) Balun design from the amateur radio handbook and one optimized for the 50:200-ohm level. The load is grounded at its center.*

Figure 8-5. *The input impedance versus frequency for the two Ruthroff Baluns of Figure 8-4, but with a 300-ohm load. Note the deterioration of the W2FMI design, which was optimized for the 50:200-ohm level.* ↓

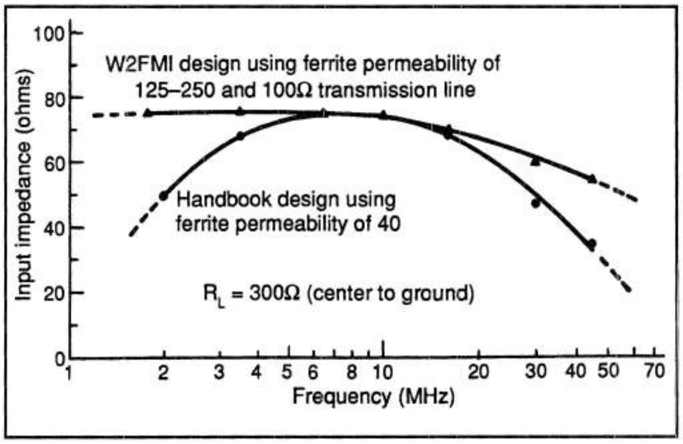

for the 75:300-ohm level. Because the length of the transmission line becomes significant beyond 10 MHz, standing waves then change the impedance ratio due to the mismatch with the Balun's transmission line. My design also shows more safety margin at the low end. I'm surprised that these simple measurements weren't made many years ago.

However, the Balun shown in the handbook does have one interesting feature. It uses a very low permeability ferrite (40), which has been shown by very accurate insertion loss measurements^2 to yield efficiencies in Baluns (and Ununs) of *99 percent* at the

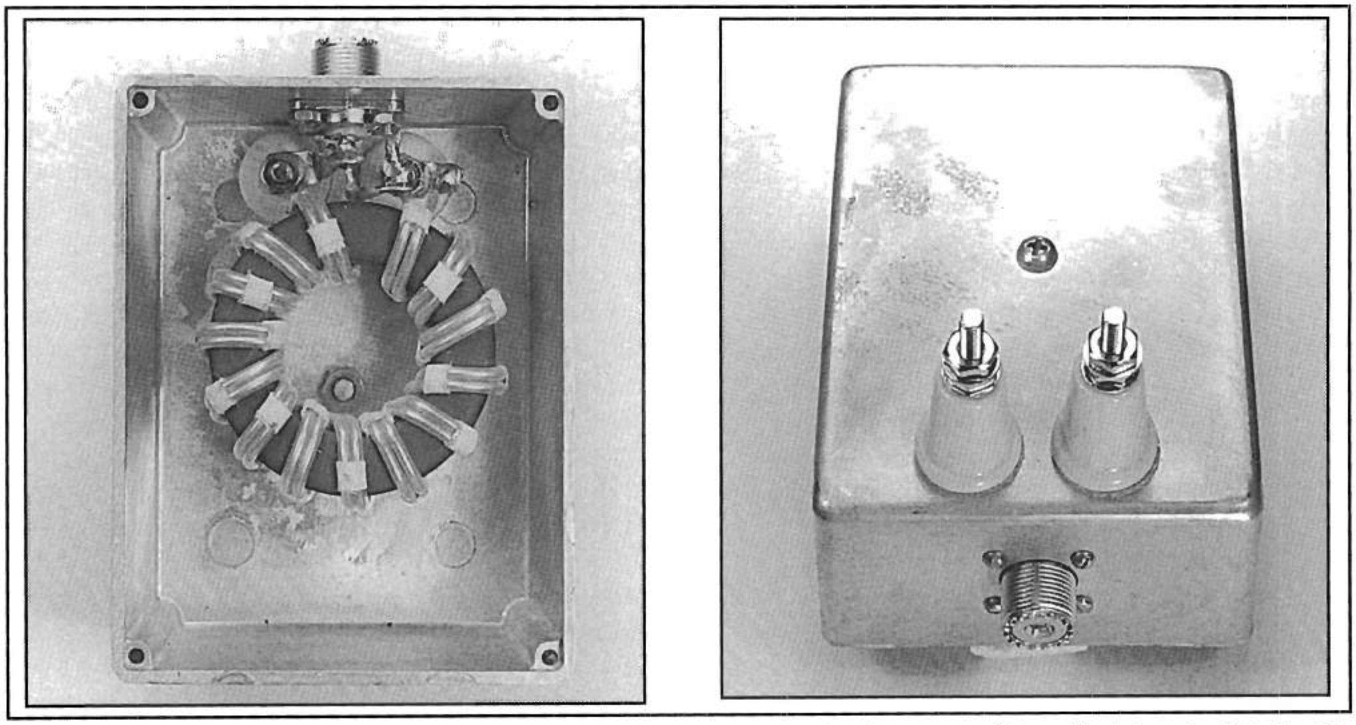

Photo 8-B. *Two different views of the optimized version of the 4:1 Ruthroff Balun mounted in a 4 inch long by 3 inch wide by 2.25 inch high Bud CU 234 aluminum enclosure.*

50:200-ohm impedance level! This is even a percent or two better than the ferrite with a permeability of 125. Because this ferrite permeability is so low, the major problem lies in obtaining sufficient choking reactance at the lowest frequency of interest, so that only transmission line currents are allowed to flow.

The design chosen (in order to exploit this very high efficiency) is shown **Photo 8-C**. It uses 14 bifilar turns of the same wire, as with my previous Balun shown on the right in **Photo 8A**, on two 2.4-inch OD cores (bound together with No. 27 glass tape) with permeabilities of 40. The unmounted view shows how the two cores are bound together by glass tape. The other views attempt to give an example for mounting the Balun. The Balun is supported by two acrylic end pieces which are, in turn, held fast to the enclosure by a long bolt. The Balun is placed equidistant between the top, bottom, and sides of a 5 inch long by 3.5 inch wide by 2.25 inch high aluminum enclosure. A few washers at the point where the bolt comes through the enclosure help to position the Balun between the top and bottom.

When matching 50 ohms (unbalanced) to 200 ohms (balanced), the response of this Balun is practically the same as mine, shown in **Figure 8-4** using a single core. From 1.7 to 30 MHz, it can certainly handle the full legal limit of amateur radio power with an efficiency close to 99 percent. However, if the operation of this Balun is restricted to the HF band only (that is, from 3 to 30 MHz), then it could be rated conservatively at 10 kW of peak power and 5 kW of average power. It would be an ideal Balun for a log-periodic beam antenna.

Finally, **Photo 8-D** shows three different views of a low power 4:1 Ruthroff (voltage) Balun designed to handle the output power of any HF transceiver easily. It has 10 bifilar turns of No. 18 hook-up wire on a 1.5-inch OD ferrite toroid with a permeability of 250. The enclosure is a 2.75 inch long by 2.125 inch wide by 1.625 inch high CU 3000-A minibox.

Section 8.4 4:1 Guanella Balun Designs

Photo 8-E shows two high-power Guanella 4:1 Baluns designed to match 50-ohm coaxial cable to loads of 200 ohms. They both use No.14 H Thermaleze wire with a covering of Teflon tubing giving characteristic impedances very close to 100 ohms (the objective). Their responses are flat from 1.5 MHz

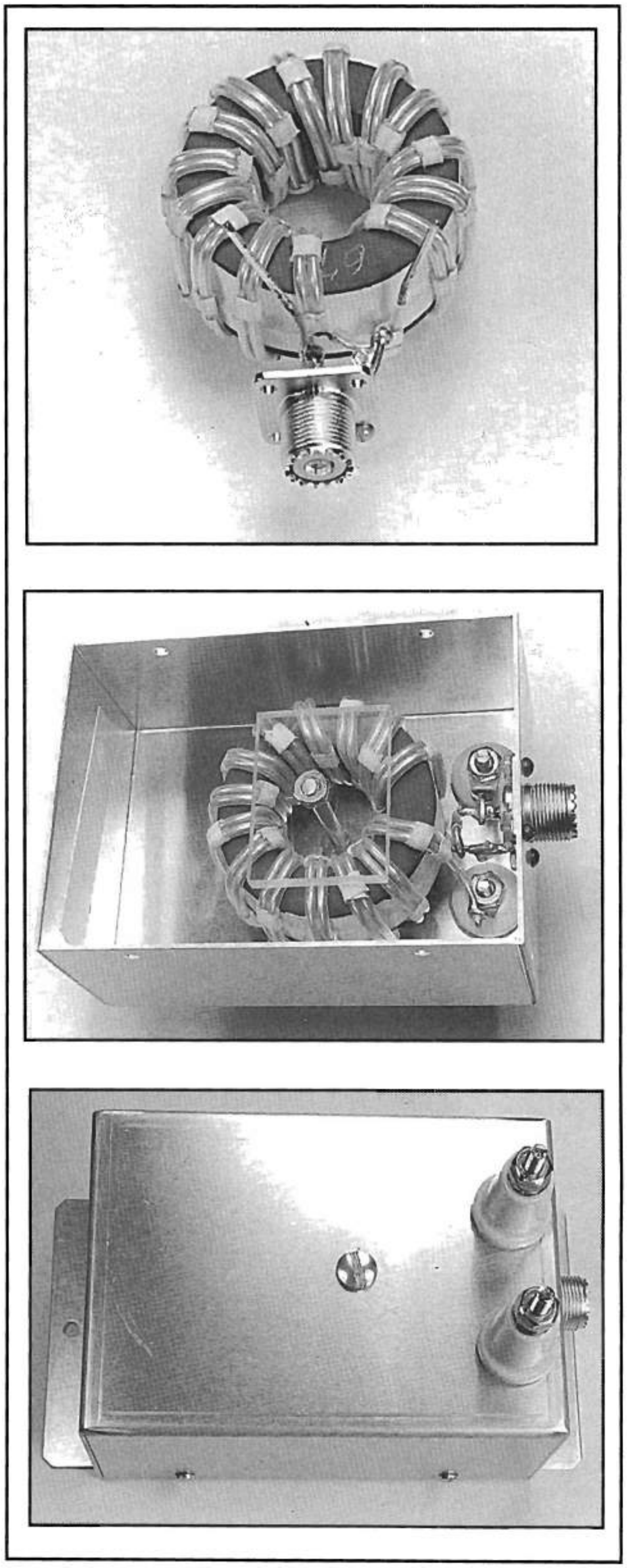

Photo 8-C. *Three views of the very high-power 4:1 Ruthroff Balun using two low-permeability (40) ferrite cores. Dimensions of the aluminum enclosure are: length, 5 inches; width, 3.5 inches; height, 2.25 inches.*

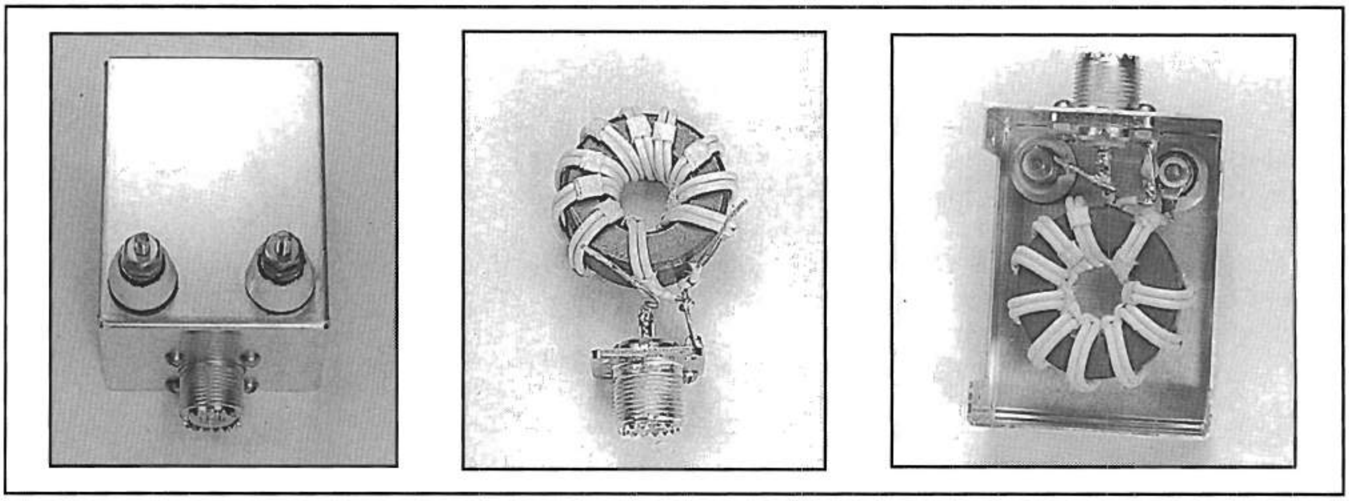

Photo 8-D. *Three views of the low-power 4:1 Ruthroff Balun designed to handle the output of any HF transceiver. The aluminum enclosure is a 2.75 inches high by 2.125 inches wide by 1.625 inches high CU 3000-A minibox.*

to well beyond 30 MHz. Both can easily handle the full legal limit of amateur radio power.

The single-core version (on the left) has 8 bifilar turns on each of its two transmission lines. The dual-core version (on the right) has 16 turns on each core. The wires are clamped together with strips of Scotch No. 27 glass tape placed about every 3/4 inch. The cores are 2.4-inch OD ferrite toroids with a permeability of 250. The connectors are on the low-impedance sides. For ease of connection, the dual-core version has one winding clockwise and the other, counterclockwise. Also, in the dual-core case, the spacing between the two cores (which isn't critical) can be as small as 1/4 inch.

These transformers can also be wound with ordinary No. 14 (solid) house wire. The several samples I tried yielded characteristic impedances close to 100 ohms (and, thus, were acceptable). The major difference lies in the voltage-breakdown capability. Units wound with Teflon-sleeved No. 14 H Thermaleze wire have been reported to withstand 10,000 volts without breakdown! Obviously, this is beyond the capability of ordinary house-wire.

Photo 8-F shows a Guanella (current) 4:1 Balun mounted in a 5 inch long by 4 inch wide by 3 inch high CU 3005-A minibox. It has 14 bifilar turns of No. 14 H Thermaleze wire on each of the two 2.4-inch OD ferrite toroids with permeabilities of 250. Each wire is covered with Teflon tubing, resulting in a characteristic impedance close to 100 ohms (the optimum). The windings also employ a crossover after the seventh turn, as shown in **Figure 8-6**. For ease of con-

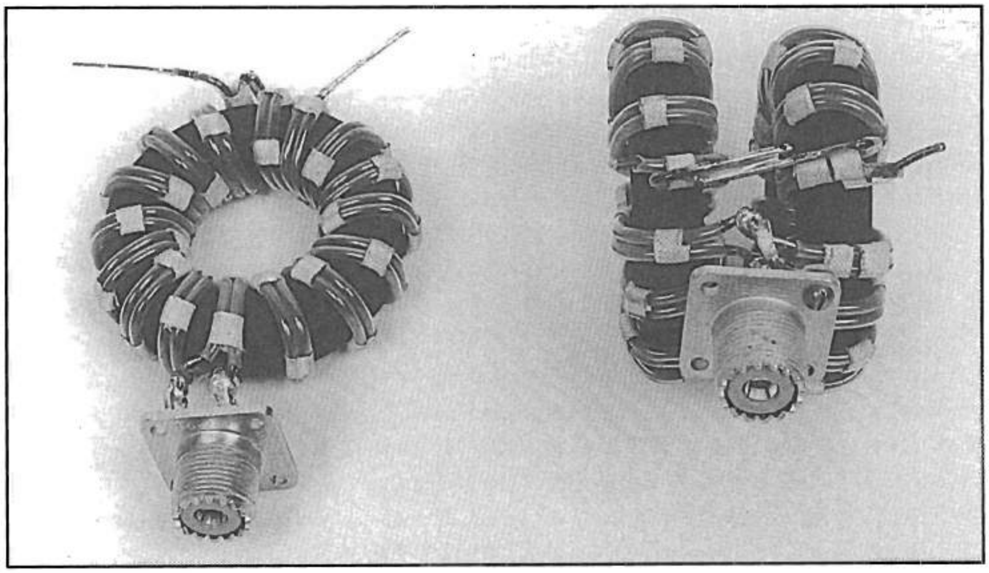

Photo 8-E. *Two high-power versions of the Guanella 4:1 Balun. The Balun on the left uses a single core while the one on the right uses two cores. The connectors are on the low-impedance sides.*

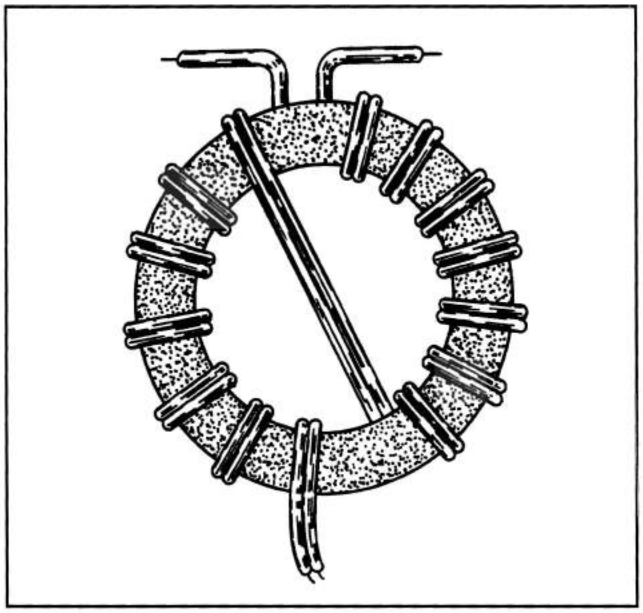

Figure 8-6. *Construction of a crossover placing input and output connections on opposite sides of the toroid.*

nection, one toroid is wound clockwise and the other, counterclockwise. The spacing between the toroids can be between 1/4 and 1/2 inch.

When matching 50-ohm cable to a balanced load of 200 ohms, the transformation ratio is constant (within 2 percent) from 1.5 to 45 MHz. This Balun can also handle the legal limit of amateur radio power. It would probably perform satisfactorily if wound with ordinary No. 14 house wire (solid), or with Teflon-covered No. 14 tinned wire. However, the design in **Photo 8-F** has withstood peak pulses of 10,000 volts! Considering this fact, it might be worthwhile to take the extra effort and use Teflon-covered No. 14 H Thermaleze wire.

A single-core version using two coiled transmission lines on a single core looks interesting and should be investigated. It results in balanced currents and unbalanced voltages. I would use two coiled transmission lines with 7 bifilar turns of the same wire on the same core as above.

Photo 8-G shows two low-power Guanella 4:1 Baluns designed to match 50-ohm coaxial cable to loads of 200 ohms. They both use No. 20 hook-up wire (solid) giving a characteristic impedance very close to the objective of 100 ohms. Their responses are flat from 1.5 MHz to well beyond 50 MHz. They are conservatively rated at 150 watts of continuous power and 300 watts of peak power. They have been exposed to 500 watts of continuous power (in a matched condition) for a considerable length of time with virtually no rise in temperature.

The single-core version (on the left) has 7 bifilar turns on each of its two transmission lines, while the dual-core version (on the right) has 14 turns on each core. The wires are clamped together about every 1/2 inch with strips of Scotch No. 27 glass tape. The cores are 1.25-inch OD ferrite toroids with a permeability of 250. The connectors are on the low-impedance sides. As above, the dual-core version has one winding clockwise and the other, counterclockwise.

Photo 8-H is a step-down version of the Guanella 4:1 Balun and uses two ferrite rod-cores 3/8 inch in diameter and 3.5 inches in length. Core permeabilities are 125. It uses the schematic of **Figure 8-1A,** but with the generator (which is grounded) on the right side and the load (ungrounded) on the left side. This 4:1 Balun is designed to match 50-ohm coaxial cable (on the right side) to a balanced load of 12.5 ohms. Each rod has 13.5 bifilar turns of No. 14 H Thermaleze wire. Again, for ease of connection, one rod is wound clockwise and the other, counterclockwise. The response is flat from 1.5 MHz to well over 30 MHz. This Balun is fully capable of handling the legal limit of amateur radio power. The connector is on the high-impedance (50 ohms) side. Beaded versions of Guanella's step-down 4:1 Balun also look very promising for operations on the VHF and UHF bands. The technique requires minimizing the parasitics in the interconnections. Some examples will be shown later.

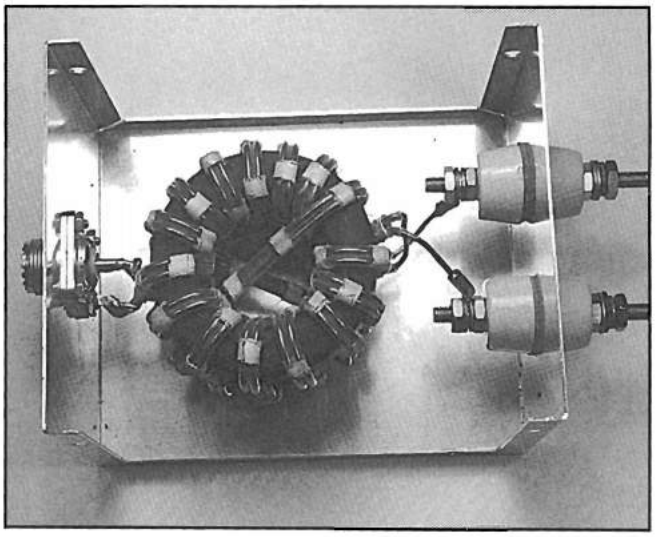

Photo 8-F. *A 4:1 Guanella Balun using the crossover connection of Figure 8-6 and mounted in a 5 inch long by 4 inch wide by 3 inch high CU 3005-A minibox.*

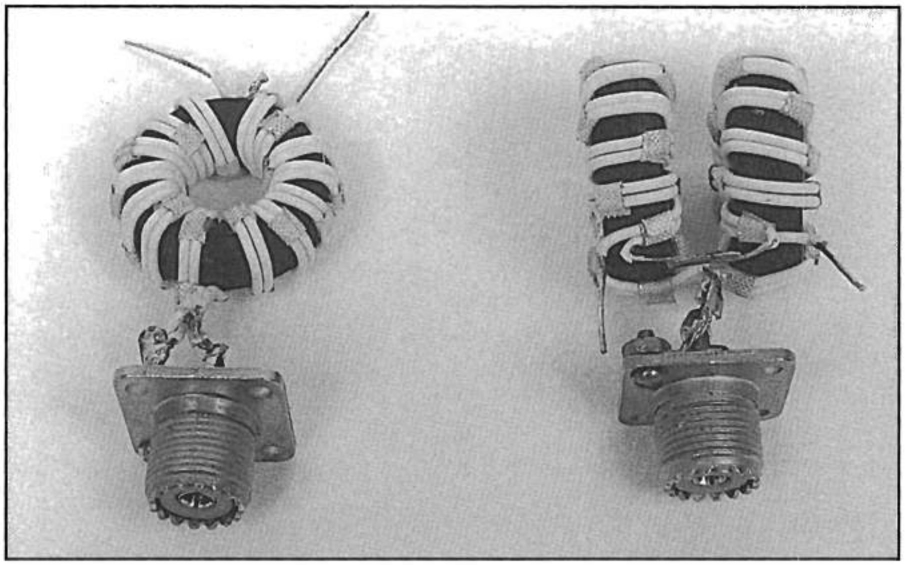

Photo 8-G. *Two low-power versions of the Guanella 4:1 Balun. The Balun on the left uses a single core while the one on the right uses two cores. The connectors are on the low-impedance sides.*

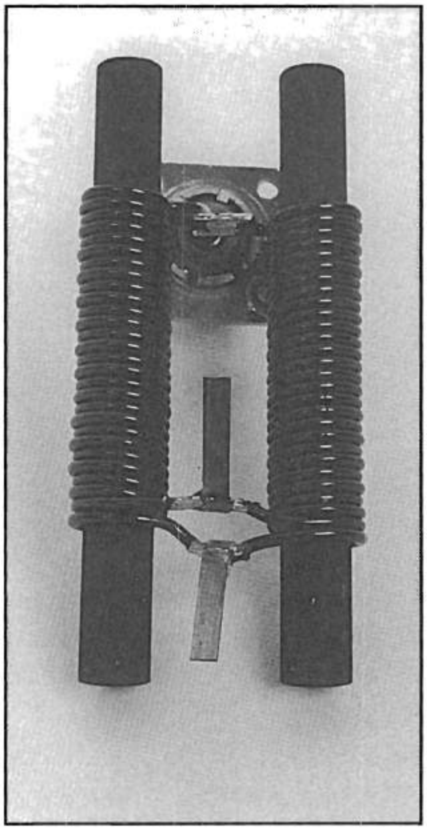

Photo 8-H. *A dual rod-core 4:1 Guanella step-down Balun designed to match 50-ohm cable to a balanced load of 12.5 ohms. The connector is on the 50-ohm unbalanced side.* ⟶

It should be mentioned again that the three dual-core Baluns above also make excellent broadband Ununs. They only sacrifice a little in low-frequency response. However, because of their conservative designs, they can still handle the 160-meter band.

Sec 8.5 Comparisons with Other Baluns

After completing the study on 4:1 Baluns, I thought it would be interesting to characterize other Baluns that are commercially available, or that have been recently described in the amateur radio literature. My findings are as follows.

The 4:1 Rod-type Ruthroff Balun

Photo 8-I shows a photograph of the typical rod-type 4:1 Balun, which was practically the only one available for the past three to four decades. The Balun in the photograph is the **HI-Q Balun.** It is the Ruthroff design (now called a *voltage* Balun[6]) with 10 bifilar turns of No. 14 wire on a 1/2-inch diameter ferrite rod 2 inches in length. In terminating this Balun with 200 ohms, the useful range was found to be from 7 to 15 MHz. Below 7 MHz, the input impedance showed a considerable inductive component—indicating autotransformer action and flux in the core (which could be harmful). Above 15 MHz, the transformation ratio increased and became complex. The optimum impedance level was found when matching 100 ohms to 25 ohms (indicating a characteristic impedance of the windings of only 50 ohms). The useful frequency range at this impedance level increased to 3.5 MHz to 30 MHz.

When matching 50-ohm coaxial cable to a 20-meter folded dipole at a height of 0.17 wavelengths (resulting in a resonant input impedance of 200 ohms), the VSWR curve was indistinguishable from that of the best Guanella 4:1 (current) Baluns.[6] This Balun also presented no difficulty in handling the full power limit. However, on 10 meters, the difference due to a very low characteristic impedance of the coiled transmission line became evident. **Figure 8-7** shows the poor VSWR curve of the rod-type Balun when compared to other Guanella and Ruthroff Baluns with characteristic impedances close to the optimum value of 100 ohms. The rod-type Balun with all of its inadequacies, which include voltage-breakdown, is certainly *not* recommended.

The high-power Ruthroff Balun is close to McCoy's design.[17] It uses 11 bifilar turns of No. 14 H Thermaleze wire on a stack of three T200-2 cores. The wires

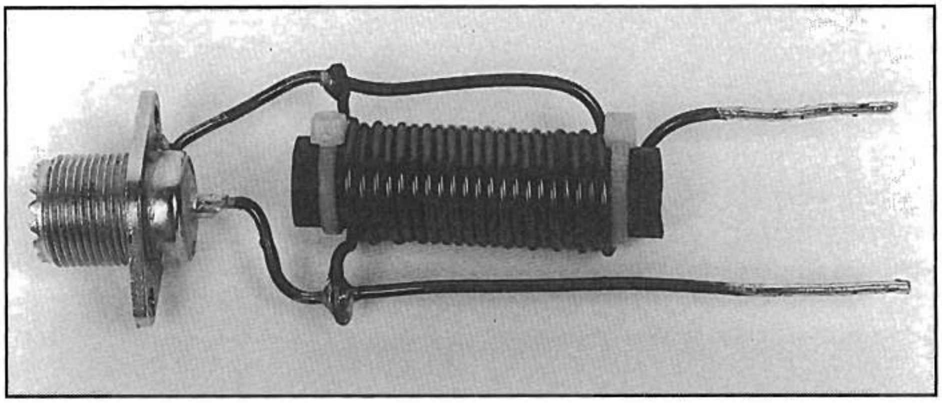

Photo 8-I. *A typical 4:1 rod-type Ruthroff (voltage) Balun (HI-Q).*

are also covered with 20-mil wall Teflon tubing yielding a characteristic impedance close to 100 ohms. I differ with McCoy's design because I feel the characteristic impedance of his transmission line could be closer to 50 ohms. His Balun and others that use powdered-iron cores will be described further in the next chapter on Baluns for antenna tuners.

The low-power units in **Figure 8-7** used 14 bifilar turns of No. 18 hook-up wire on 1.25-inch OD ferrite toroids with a permeability of 250. The transmission lines of the low-power Baluns were just under 20 inches, while those of the high-power Balun were 50 inches in length.

As can be seen in **Figure 8-7**, there are **very** small differences between the VSWR curves of the two low-power Baluns and the high-power Ruthroff (voltage) Balun. The differences could very well be attributed to the small variations in the characteristic impedances of the windings. Very likely, the most important information gleaned from **Figure 8-7** is that, when feeding a balanced dipole with a virtual ground-plain bisecting it, the Ruthroff Balun takes on the character of a two-core Guanella Balun. In other words, the Ruthroff Balun loses its built-in high-frequency cut-off!

4:1 Current Baluns

I also characterized several so-called *current* Baluns^6 that recently appeared on the market. These are my findings:

a) They are the dual-core (toroids) version of the Guanella Balun, which sums voltages of equal delays.

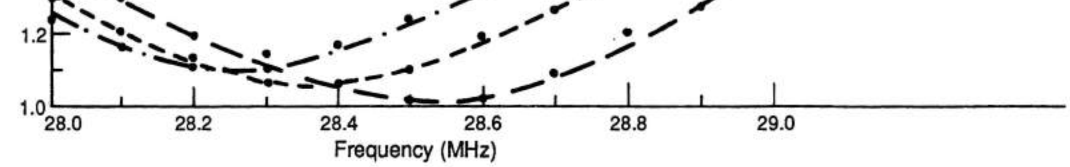

Figure 8-7. *Plots of VSWR curves on 10 meters for four different 4:1 Baluns. The high-power Ruthroff Balun is actually McCoy's design with a 100-ohm characteristic impedance winding (and is described in the next chapter). The Hi-Q Balun is shown in Photo 8-I. The comparisons show the importance of having the optimum value of the characteristic impedance of the windings and that the high-power Balun with its much longer transmission line indicates a Guanella-type operation.*

b) The electrical performances of these Baluns are vastly superior to the rod-type Balun described earlier.

c) These Baluns should meet their electrical and power-rating specifications.

d) My only criticism is that they could have more of a safety margin at the low-frequency end, where excessive core flux (due to higher than expected impedances) could take place. More inductance in the windings is recommended.

The Beaded-coax 4:1 Balun

A design in an amateur radio journal[5] advocated using beaded coaxial cable (of 100 ohms) in a 4:1 Guanella design. Various claims were advanced for this approach. I constructed one of these Baluns using No.14 wire with Teflon sleeving, resulting in the required 100-ohm characteristic impedance. Here are my findings:

a) The Balun had excellent margins at both the high- and low-frequency ends. The performance of this Balun verified the analysis (expressed earlier) with the high- and low-frequency models and the subsequent voltage gradients. In fact, the high-frequency performance exceeded the capability of my simple test equipment.[2]

b) The major disadvantage is in efficiency. Because high-permeability (2500) beads are required in order to obtain the required choking reactance in the HF band, this Balun had considerably more loss than coiled-type Baluns using low-permeability (less than 300) ferrite toroids.[2] A soak-test[2] (transformers connected back-to-back and about 500 watts applied into a dummy load) with the dual-core low-power unit in **Photo 8-G** showed that the smaller Balun ran considerably cooler! The beaded transmission line technique is recommended mainly for Baluns (and Ununs) operating at low impedance levels or on the higher-frequency bands.

Sec 8.6 Summary

Unlike the 1:1 Balun, the 4:1 Balun matching 50 ohms unbalanced to 200 ohms balanced has had no real standard for comparison in the amateur radio literature. As **Chapter 7** showed, Reisert's Balun in his 1978 article[7] had all of the attributes of a good 1:1 design. Therefore, he set a legitimate standard for others to follow or even attempt to exceed. Also shown were some of my variations in his design for increased efficiency and ease of construction.

However, **Chapter 8** has illustrated that the designs in the amateur literature (particularly the handbooks) are found lacking in bandwidth, or efficiency, or both. Even the 4:1 Baluns on the commercial market can be improved. This is *especially* true of the rod-type Balun that has been available for decades!

In the process of investigating the 4:1 Balun for my series of articles in *CQ* and *Communications Quarterly*, I have arrived at some designs that could provide the beginnings of standards for this device. They are included in this chapter and are:

1. For *balanced* applications like matching 50-ohm cable to the 200-ohm balanced input impedance of folded-dipole or log-periodic antennas, I recommend the single-core, Ruthroff design of **Photo 8-A**. It is capable of handling the full legal limit of amateur radio power. For a higher-power capability and a little less bandwidth, I recommend the Ruthroff design shown in **Photo 8-C**. These designs are presently called *voltage* Baluns.[6]

2. For unbalanced applications like the OCFD (off-center-fed dipole), or a dipole that could be unbalanced by surrounding structures or by construction errors, I recommend the two-core Guanella design of **Photo 8-E** or **Photo 8-F**. It is a much more flexible unit that can operate successfully as a Balun when the load is grounded at its center (**Figure 8-1A**), or as an Unun when the load is grounded at the bottom. Although not shown, it can even be grounded at the top, yielding a 4:1 phase-inverter.

3. For low-power 4:1 Baluns, there really have been no designs in the literature for comparisons. Therefore, by default, the designs in this chapter are suggested as standards. Applications of the single-core Ruthroff Balun and the two-core Guanella Balun are the same as their higher-power counterparts.

I am sure there are some who don't agree with the recommendations proposed above. The Guanella (current) Balun appears to be the main Balun of choice. The question is: Why use a two-core Guanella (current) Balun when a single-core Ruthroff (voltage) Balun will do? They both have the same power ratings! For those who disagree with my views, designs, or recommendations, I encourage them to (as the classic TV commercial used to say) *put it in writing*. Then we will all benefit from the new information.

Chapter 9

Baluns for Antenna Tuners

Sec 9.1 Introduction

The 4:1 Balun, matching 50 ohms (unbalanced) to 200 ohms (balanced), has found its most popular use in antenna tuners. Because the Balun rarely sees a resistive load of 200 ohms in this application, the primary objective is to take the balanced impedance (with respect to ground) of the input to an open-wire (or twin-lead) feedline and transform it into an unbalanced impedance which has one side grounded and can be transformed into 50 ohms by an L-C matching network. This was well described in two *CQ* articles by McCoy.[17,18]

As **Chapter 8** has shown, there are two different forms of the 4:1 Balun. One uses two transmission lines wound on separate cores (or threaded through ferrite beads in some cases) and connected in parallel at the 50-ohm side and in series at the 200-ohm side. This design was first presented by Guanella in 1944,[3] and is presently called a *current* Balun.[6] The other design, using a single transmission line wound around a core and connected in a phase-inverter configuration, was introduced by Ruthroff in 1959.[9] This design, which has recently been called a *voltage* Balun,[6] is now perceived as the inferior design.

This chapter presents another view of the 4:1 Balun. It not only includes the optimum design considerations for antenna tuner use, but also the design parameters for multiband antenna systems using center-fed dipoles with open-wire or twin-lead feedlines. Also included are my views on the G5RV antenna, the tuner using a 1:1 Balun before the L-C matching network, and the special cases of Ruthroff's 4:1 design matching into a load that is actually or virtually grounded at its center.

Sec 9.2 The Two Forms of Antenna Tuners

Three of the more common items in amateur radio jargon are VSWR, antenna tuners, and multiband antennas (especially the G5RV). These have appeared upon the scene because of the ease at which bands can now be changed and the narrow limits in the range of matching impedances with modern rigs.

The concept of using a wire antenna on many different bands isn't new. Designs have been around for more than six decades. In fact, satisfactory circuits have also been available which couple transmitters to balanced lines that present loads different than the transmitter output impedance. These were known as series and parallel-tuned circuits.[19] The transforming of a balanced impedance to an unbalanced impedance was accomplished by the isolation provided by magnetic coupling. Energy was transmitted from one circuit to the other by either having two coils in close proximity or by "link" coupling. However, these methods of coupling have fallen by the wayside, together with rock-bound rigs, plug-in coils, separate

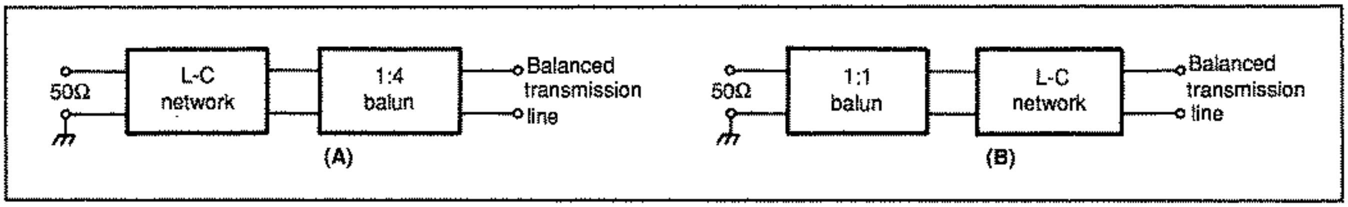

Figure 9-1. *The two basic forms of the Transmatch (antenna tuner): (A) the more popular design using an unbalanced L-C network and a 1:4 Balun; (B) a 1:1 Balun and a balanced L-C network.*

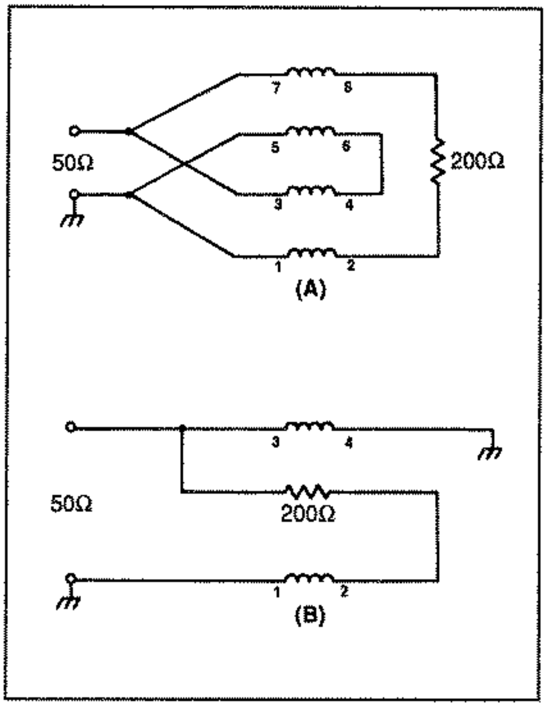

Figure 9-2. *The two basic forms of the 1:4 Balun: (A) the Guanella (current) Balun; (B) the Ruthroff (voltage) Balun.*

Figure 9-3. *The frequency response of a 4:1 Ruthroff (voltage) Balun with the load floating and with its center grounded. In the grounded case, the high-frequency response is similar to that of a 4:1 Guanella (current) Balun.*

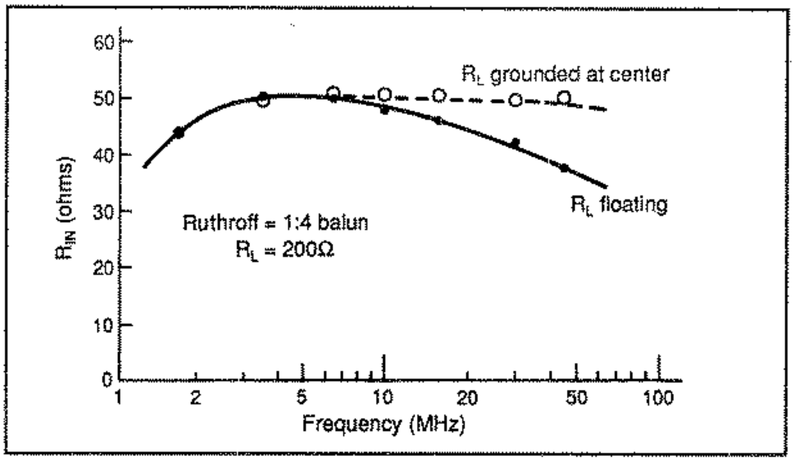

receivers and transmitters, and (sad to say) the exciting flashing of mercury-vapor rectifiers.

Today, the *transmatch* is most often used to convert the reactive/resistive load presented by an antenna system to a nonreactive, grounded 50-ohm load. The transmatch is also commonly known as an antenna tuner. The isolation role, that of converting a balanced impedance to an unbalanced one, is now provided by the Balun transformer.

There are two basic forms of the transmatch, and they are shown in **Figure 9-1**. **Figure 9-1A**, which shows a 4:1 Balun between the L-C network and the balanced transmission line, has been the most popular. In some designs, a 1:1 Balun has been used. In either case, this form of antenna tuner places the burden on the Balun, not on the L-C network. Depending upon the dimensions of the antenna and open-wire (or twin-lead) transmission line, the Balun can see very high impedances that may be harmful. In turn, the L-C networks are simple because of their unbalanced nature. Among the most popular networks are L, pi, T, Ultimate, and SPC types.[20–23] An added advantage to this approach is that the Balun can be placed outside the operating area and connected to the L-C network by a coaxial cable.[17,18]

On the other hand, **Figure 9-1B** takes the complexity out of the Balun and places it on the L-C network. With a balanced network, the 1:1 Balun should see a lower voltage drop along the length of its transmission line (and, hence, less loss[2]) because its load is always close to 50 ohms. Additionally, the choking requirements of a 1:1 Balun are considerably less than that of the 4:1 Balun in **Figure 9-1A**.

Roehm has addressed the problems related to this form of transmatch in an article.[16] In fact, he suggests a design using an unbalanced T network and a 1:1 beaded-coax Balun. Although a balanced L-C network is inherently more complex and costly, it would be interesting to see its comparison with Roehm's unbalanced design. Additionally, a comparison with a 1:1 Balun using 50-ohm twin-lead or coaxial cable wound around a ferrite toroid with a permeability of less than 300 would also be useful. Because the 1:1 beaded-coax (choke-type) Balun requires ferrite beads with permeabilities considerably greater than 300, it has more loss.[2]

In any event, the basic form of the transmatch using the design in **Figure 9-1B** looks promising and merits further investigation.

Sec 9.3 Another View of the 4:1 Balun

This section presents my views and the results of my work on a 4:1 Balun designed for use in the very popular "antenna tuner" shown in **Figure 9-1A**. Because the Balun may be exposed to harmful high voltage conditions in this application, the efficiency and

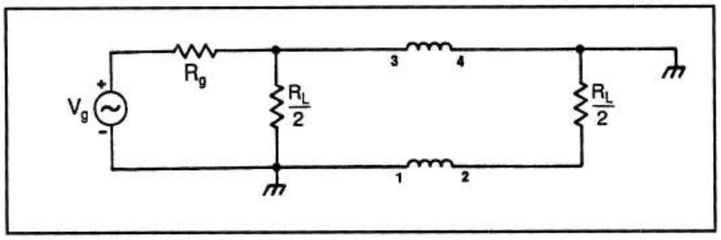

Figure 9-4. *Suggested model of the 4:1 Ruthroff (voltage) Balun when the load, R_L, is grounded at its center.*

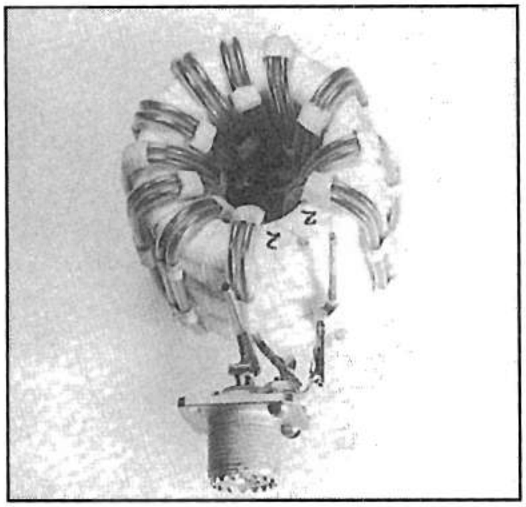

Photo 9-A. *The high-power Ruthroff 4:1 Balun used in the comparison with other Baluns (see Figure 9-7, and Figure 8-7 of Chapter 8). Except for some difference in the characteristic impedance of the bifilar winding, it is essentially the McCoy 4:1 Balun.*

ruggedness of the core materials are important considerations. Experiments have shown that losses in Baluns are related to the impedance level[2] (and, hence, voltage level), and the permeability of the materials. Therefore, the losses are of a dielectric-type and not of a current-type, as in conventional transformers. Moreover, it is well known that powdered-iron is a more rugged and linear material than ferrite. A very important question to ask is which form of the 4:1 Balun should be used in this application—Guanella's or Ruthroff's? My conclusions may surprise many readers.

As was mentioned at the beginning of this chapter and above, there are two basic forms of the 4:1 Balun. They are shown in **Figure 9-2A** and **B**. **Figure 9-2A** is Guanella's approach. It uses two coiled transmission lines (on separate cores) connected in parallel on the 50-ohm side and in series on the 200-ohm side. This has recently been called a *current* Balun.[6] In order to have "flat" transmission lines and obtain the highest frequency response, the characteristic impedance of the coiled transmission lines should be equal to the loads they see—namely, $1/2R_L$ and, in this case, 100 ohms.

As Guanella said in his classic paper,[3] this Balun is literally "frequency independent." At the low frequency end, the reactance of the coiled transmission line should be much greater than 100 ohms (in this case) in order to assure that the energy is transmitted from input to output by an efficient transmission line mode. Beaded transmission lines aren't recommended for use on the HF band at these impedance levels because of excessive dielectric loss.

Figure 9-2B shows Ruthroff's approach, which uses a single transmission line connected in the phase-inverter configuration (see **Chapter 7**). By grounding terminal 4, a voltage drop of $-V_1$ appears across the length of the transmission line. As a result, terminal 3 is at $+V_1$ and terminal 2 is at $-V_1$—a 4:1 transformation ratio.

It is when the load is actually grounded at its center that we see a very interesting feature of this approach. A simple impedance measurement will show that the high frequency response is *vastly* improved and that the Ruthroff (voltage) Balun appears to take on the character of a Guanella (current) Balun! **Figure 9-3** illustrates the measurements of the input impedance of a Ruthroff 4:1 Balun with the load floating, and when it is grounded at its center. These measurements (with a simple resistive bridge) were made on my design using a powdered-iron core, which will be described later.

A model for the Ruthroff 4:1 Balun, when the load is grounded at its center, is provided in **Figure 9-4**. If the characteristic impedance of the transmission line in the 4:1 Balun is 100 ohms (the optimum value if the load is 200 ohms), then the generator sees $R_L/2$ in parallel with a $R_L/2$ from a "flat" line. As a result, the generator sees its match of 50 ohms—even though the currents in the loads are not in phase!

However, when the Balun is connected to a center-fed, folded dipole, or log-periodic beam antenna with 200-ohm input impedances, the virtual ground-plane bisecting the antennas presents an interesting case. Because there is no metallic connection to the center of the load, the currents in both halves of the antennas are in phase. This is unlike the situation in **Figure 9-4**, where the current in the load on the right is delayed compared to the current in the load on the left. As is evident in **Figure 8-7** of **Chapter 8**, the VSWR curve

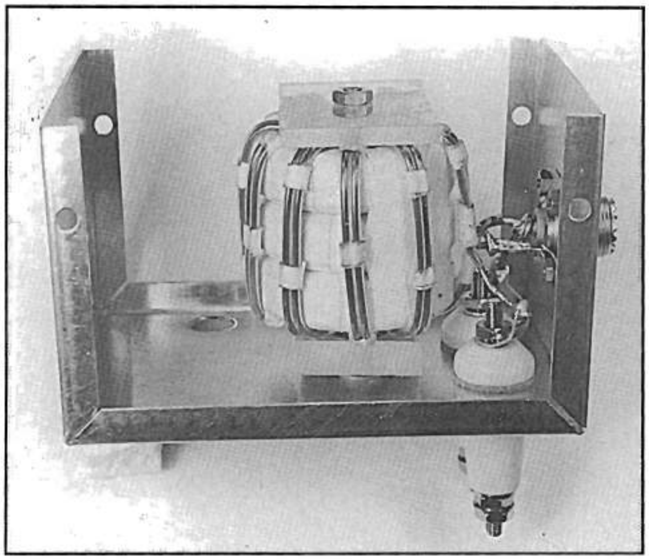

Photo 9-B. *The Balun of Photo 9-A mounted in a 5 inch long by 4 inch wide by 3 inch high minibox.*

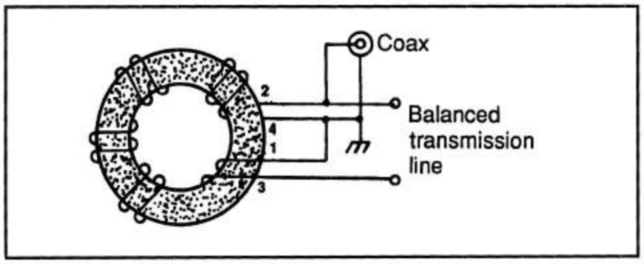

Figure 9-5. *A pictorial of the connections for a 4:1 Ruthroff (voltage) Balun.*

for the high-power Ruthroff Balun has practically the same shape as those of the two lower-power units—even though its transmission line is more than two and one half times longer.

Because the high-frequency response of a Ruthroff 4:1 Balun with a floating load is highly dependent upon the length of the transmission line, **Figure 8-7** suggests that, in the virtual ground case, the Balun acts as a Guanella Balun which sums voltages of equal phases. It also suggests that the effective electrical length of the transmission line is one-half of its actual length. I am quite sure that this model of Ruthroff's Balun was not proposed by him or by others that followed. However, it should be remembered that this condition only exists in the *balanced* case. With an unbalanced load, the Balun should introduce a reactive component that will limit the high frequency response. In the unbalanced case, the Guanella Balun with two cores is the Balun of choice.

Sec 9.4 Some "Hardy" 4:1 Designs

Photos 9-A and **9-B** show two views of a Ruthroff 4:1 Balun, much like McCoy's design, which has been used in his highly popular transmatch.[22] **Figure 9-5** shows a pictorial representation of the connections. It has 11 bifilar turns of No. 14 H Thermaleze wire on three (stacked) T200-2 cores. The cores are powdered-iron material with a permeability of 10. The OD

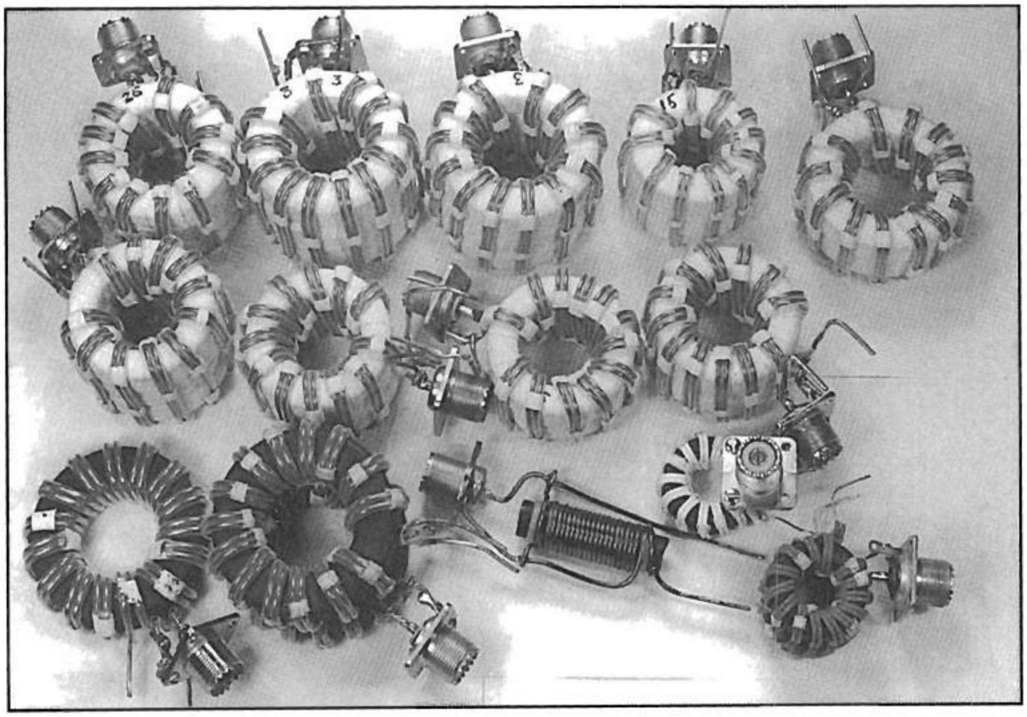

Photo 9-C. *The various Baluns used in the study of 4:1 Baluns for antenna tuners.*

is 2 inches. The wires are also covered with a 15-mil wall Teflon tubing, yielding a characteristic impedance close to the optimum value of 100 ohms.

I differ with McCoy's design here, because the characteristic impedance of his Balun could be closer to 50 ohms. **Photo 9-B** shows the Balun mounted in a 5 inch long by 4 inch wide by 3 inch high minibox. Because McCoy didn't use a thickly insulated wire, he wound a layer of Scotch No. 27 glass tape on each toroid before stacking. This was followed by another layer in the stacking process. With Teflon sleeving over the wire, the extra insulation provided by the glass tape could be dispensed with.

In order to improve the low-frequency response of McCoy's Balun, I made a study of higher permeability powdered-iron cores. **Photo 9-C** shows the various Baluns used in the study. The object of this study was to determine the best core material for a 4:1 Balun to be used in antenna tuners where they can be exposed to high impedances (and, hence, hostile environments). I knew, as the result of very accurate insertion loss measurements[2] that loss with ferrite materials was related to the voltage drop along the length of the transmission line and to the value of the permeability. Permeabilities of 40 (No. 67 ferrite) exhibited the lowest loss. The results taken on a single powdered-iron material—No. 2 material with a permeability of 10—also showed the very same low loss. Because powdered-iron material has been known to be more rugged and linear than ferrite material, this suggested that other powdered-irons with greater permeabilities should also be investigated.

I investigated four other powdered-irons with permeabilities of 20, 25, 35, and 75. Their designations were Nos. 1, 15, 3, and 26, respectively. Comparisons were made on input impedances (with the outputs terminated in 200 ohms) and temperature rises (when handling 500 watts of power). The power test showed, convincingly, that No. 26 material was not to be used because it showed a definite rise in temperature while the other three didn't. However, all four materials showed a definite lower input impedance than the No. 2 material, which has a permeability of 10. As expected, the higher the permeability, the larger the difference with No. 2 material. Although an input impedance measurement does provide some indication of loss because it appears as a shunting path to ground, a very accurate insertion loss measurement would provide a more precise indication of the trade-off that can be made in efficiency for low frequency response.

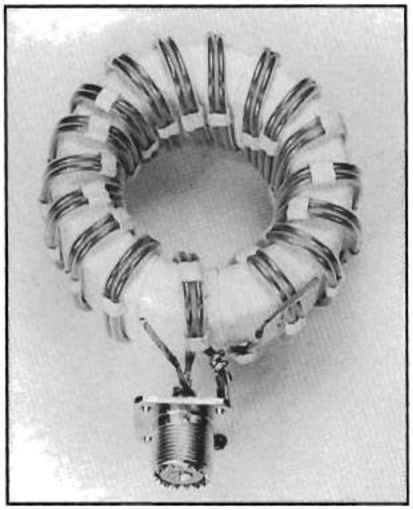

Photo 9-D. *An improved 4:1 Ruthroff design for antenna tuners.*

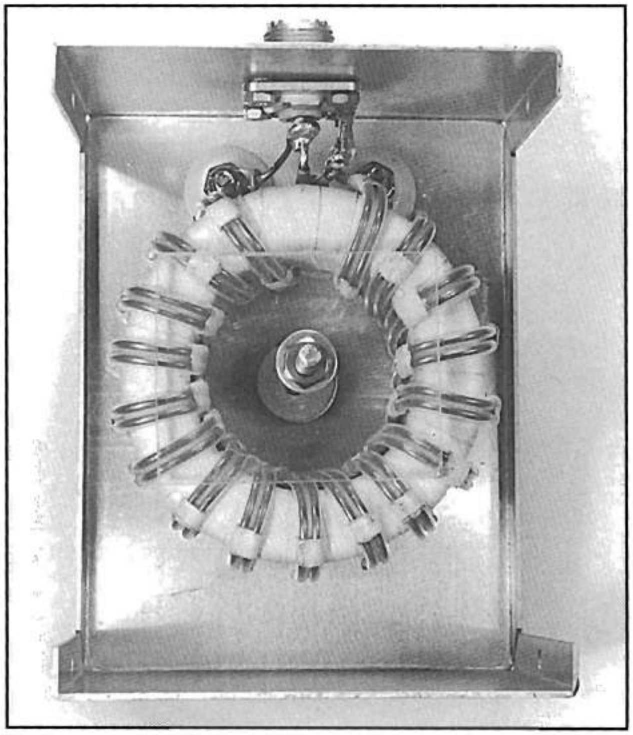

Photo 9-E. *The improved 4:1 Ruthroff Balun mounted in a 5 inch long by 4 inch wide by 3 inch high minibox.*

Because my simple loss measurements indicated that the higher permeability powdered-irons had more loss than the No. 2 material, I decided to design a 4:1 Ruthroff Balun using this material—but with a larger core and more turns than the McCoy[17,18] Balun. Although McCoy's design has enjoyed considerable success over the years, I felt that a larger inductive reactance was desirable in order to assure better performance on the lower frequency bands (particularly 160 meters).

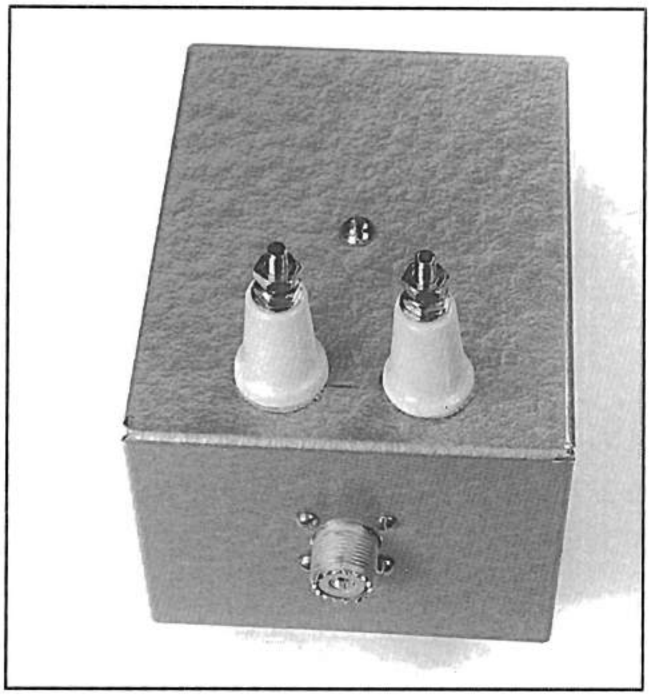

Photo 9-F. *The top view of the 4:1 Balun of Photo 9-E.*

The specific design is shown in **Photo 9-D**. It has 17 bifilar turns of No. 14 H Thermaleze wire on a T300A-2 powdered-iron core, which has an OD of 3 inches and a permeability of 10. With this number of turns and a larger cross section than the three T200-2 cores, the low frequency response improved by a factor of about two over the McCoy Balun, which has 10 to 12 turns on a stack of three T200-2 cores. Furthermore, the wires are also covered with 15-mil wall Teflon tubing, resulting in a characteristic impedance of 100 ohms (the objective). This well-insulated transmission line has been reported to handle 10,000 volts without breakdown. **Figure 9-3** illustrates the performance of this Balun (under a matched condition) when the load is floating and when it is center-tapped-to-ground. **Photo 9-E** shows the Balun mounted in a minibox 5 inches long by 4 inches wide by 3 inches high. **Photo 9-F** shows the top view of the mounted Balun.

Because the Balun with the larger core and more turns showed an improvement by a factor of two in the low-frequency response over the McCoy Balun, I constructed an even larger design. This is shown in **Photo 9-G**. It has 21 turns of the same wire on a T400A-2 core with an OD of 4 inches and a permeability of 10. Even though **Photo 9-G** shows the toroid wrapped with Scotch No. 27 glass tape, as mentioned earlier, this extra insulation isn't required because the wires are covered with Teflon sleeving.

Figure 9-6 shows the comparisons in the input impedances versus frequency of these three "hardy" Baluns when they are terminated in 200-ohm loads grounded at their centers. As can be seen, the low-frequency response of the Balun with the most turns and largest core is the best. Even though the Balun with the 3-inch core has a poorer low-frequency response, it is an improvement over the McCoy Balun and should find considerable use in transmatches. **Figure 9-6** also shows that the high frequency responses of these Baluns with the loads grounded at their centers are remarkably similar. This is especially interesting, as the length of the transmission line on the larger

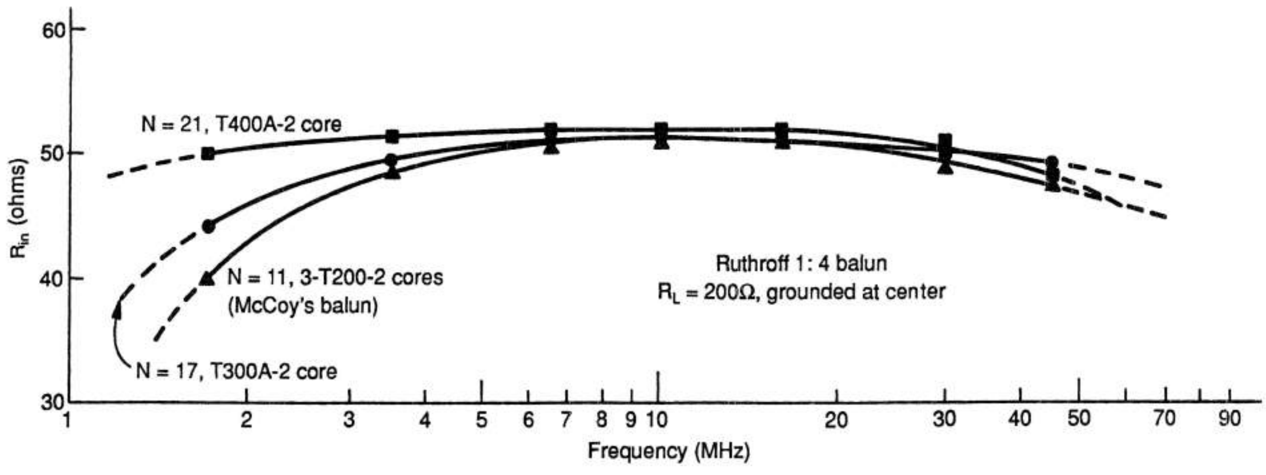

Figure 9-6. *A comparison of the performance of the McCoy 4:1 Balun with Baluns having larger cores and more turns.*

Balun (with the T400A-2 core) is more than twice as long as the other two (115 inches compared to 50 and 55 inches). If it is correct that my model of a Ruthroff (voltage) Balun feeding a balanced antenna or transmission line takes on the character of a Guanella (current) Balun because of the virtual ground, then this large Balun could have many applications.

Sec 9.5 Multiband Dipoles

Antenna tuners have been known to work well for some radio amateurs and not for others. This is due to the differences in the dimensions of their antenna systems. With high impedances seen by the 4:1 Baluns in the antenna tuners, the Baluns not only fail to provide a good balanced-to-unbalanced conversion, but they can also be damaged by excessive heating. The high current, low impedance condition seen by the Balun isn't a problem. Therefore, the object in multiband antenna design is to provide the most favorable impedances for the Baluns, especially on the three lowest frequency bands—40, 80, and 160 meters. Usually on the higher frequency bands, the impedances seen by the Baluns are not as high and the Balun's choking reactances are greater (assuring balanced-to-unbalanced conversion). This section discusses three cases of multiband center-fed dipole designs. They are 1) the "worst case" design, 2) a smaller design—the G5RV, and 3) a larger design. Others may have better designs for multiband operation, but it's clear that the "worst case" design ought to be avoided. **Figure 9-7** shows the symbols for the dimensions of the center-fed dipoles.

Sec 9.5.1 The "Worst Case" Design

Apparently, an 80-meter dipole with a quarter-wave open-wire (or twin-lead) feedline is a logical design for a multiband dipole. In **Figure 9-7** this would mean that $L_1 = L_2 = 59$ to 67 feet, depending upon the favorite operating frequency. If one were to use 450-ohm twin-lead with "open windows," L_2 would be diminished by 10 percent; with 300-ohm TV twin-lead, it would be diminished by about 20 percent. This would make a good antenna system on 160 meters. Because $L_1 + L_2$ is close to a quarter-wave, the current at the input to the feedline is at its highest value and the impedance at its lowest—a very favorable condition for the Balun. In fact, a 1:1 Balun at the feedpoint would probably do a good job.

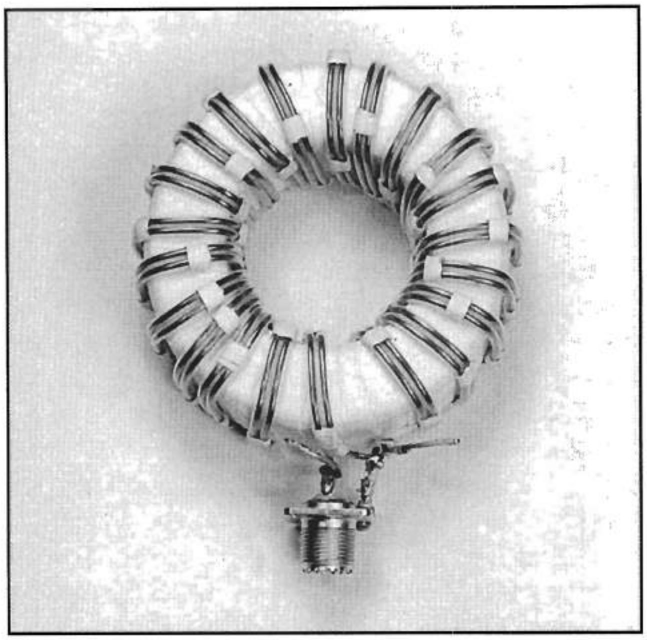

Photo 9-G. *A large 4:1 Ruthroff design using a 4-inch OD powdered-iron core and 21 bifilar turns.*

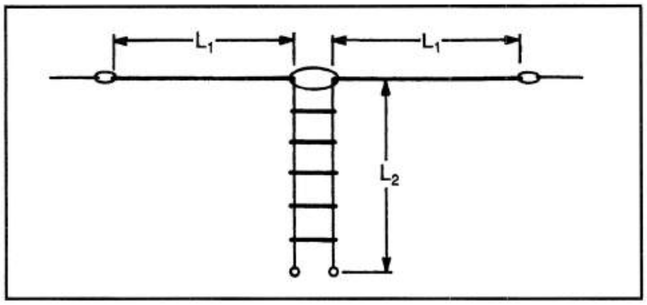

Figure 9-7. *The symbols used for the dimensions of a center-fed dipole with open-wire feeders (or twin-lead with appropriate consideration of velocity factors).*

However, what does the input to the feedline look like on 80 meters? If 450-ohm feedline is used, the 4:1 Balun sees a quarter-wave 450-ohm feedline terminated in approximately 50 ohms. Transmission line theory tells us that the Balun would see 4050 ohms—an impossible condition for most Baluns. The situation becomes worse on 40 meters, where we may have a center-fed, full-wave dipole with a half-wave transmission line (a 1:1 matching transformer). In this case, the Balun could see an impedance approaching 10,000 ohms—a more than impossible condition! Although high impedances would also be seen on 10, 15, and 20 meters, the conditions are not quite as severe because the Balun's choking reactances are usually greater and the impedances lower. In essence, the "worst case" design for a 4:1 Balun is the "best case" design (except for 160 meters), if one still uses inductive coupling to a parallel-tuned circuit.[19]

Sec 9.5.2 A Smaller Design— The G5RV

Varney,[24] G5RV, designed a multiband center-fed antenna system capable of operation on all HF bands from 3.5 to 30 MHz. In contrast to multiband antennas designed as half-wave dipoles on 80 meters (the "worst case" design), the full-size G5RV antenna was designed as a three-half-wave antenna on 14,150 MHz with a 1:1 transmission line matching transformer. It was possible to accomplish this using the dimensions of 51 feet for L_1 and 34 feet for L_2. For 450-ohm twin-lead with "windows," L_2 would be 31 feet; for 300-ohm TV ribbon, L_2 would be 28 feet. Consequently, the input impedance at the base of the matching transmission line was about 100 ohms on 14,150 MHz, and a manageable impedance for 50 or 80-ohm coaxial cable.

However, with a total length for $L_1 + L_2$ of 85 feet, the impedances at the input to the transmission line on 40, 80, and 160 meters are also manageable. Even though they have a reactive component on these bands, they aren't so high that a well-designed 4:1 Balun in an antenna tuner can't handle them easily. Varney also showed that the highest impedances occurred on the 18-, 21-, and 28-MHz bands. However, this doesn't present a problem with the design shown in **Photo 9-D** because it uses an efficient core material, and it also has the highest reactances of its windings at these frequencies.

Varney[24] also wrote, at considerable length, on the unsuitability of a Balun being used to connect the base of the 34-foot open feeders to a coaxial cable feedline. He stated that if a Balun is connected to a reactive load with a VSWR of more than 2:1, its internal losses would increase. Varney also mentioned heating of the wires and saturation of the core. Evidently Varney was not familiar with McCoy's design, which uses a powdered-iron core (with a permeability of 10) that can withstand VSWRs considerably greater than 2:1—without showing any temperature rise. Furthermore, the wire doesn't heat up; however, the core itself does via dielectric heating. Additionally, with sufficient choking reactance, Baluns can handle (equally) the resistive and reactive components of an impedance.

Finally, after observing the voltage and current distributions on all of the bands, it appears that a 2:1 (100:50-ohm) Balun might be an interesting one to try on the G5RV antenna. It could be that many of the bands would not require the added matching of an antenna tuner. If some of the bands require an antenna tuner in order to be used, then I would suggest using a "hardened" Balun. That is what I call McCoy's approach, which uses efficient and hardy powdered-iron cores. A 2:1 Balun, comprised of a 1:2 Unun in series with a 1:1 Guanella (current) Balun, could be easily designed and built (see **Chapter 10**).

Sec 9.5.3 A Larger Design

Even though the G5RV antenna can be made to operate on 160 meters with a suitable antenna tuner, an antenna system larger than the "worst case" design can provide better operation on the 40-, 80-, and 160-meter bands. As you might expect, an antenna system about twice as large as the G5RV offers these advantages. Suggested dimensions are L_1 = 80 feet and L_2 = 100 feet. If the feedline is 450-ohm twin-line with "windows," then L_2 = 90 feet. If it is 300-ohm TV ribbon, then L_2 = 82 feet. As with the G5RV, it's the total length of $L_1 + L_2$ that presents favorable or unfavorable impedances to the 4:1 Balun in the antenna tuner. Therefore, L_1 and L_2 could both be 90 feet, as well. Only very small differences in performance would be noticed between these two systems, particularly on the lower-frequency bands. Obviously, other combinations totaling 180 feet are also possible. In the G5RV case, it's 85 feet.

Sec 9.6 Summary

After reading this chapter, one might think that I have set this technology back a few years by advocating voltage Baluns and powdered-iron cores. I have even questioned the professional literature. However, my conclusions were based upon three experimental results. These were: 1) measurements with my resistive bridge on input impedances of 4:1 Ruthroff (voltage) Baluns with loads floating and center-tapped-to-ground, 2) VSWR measurements on folded dipoles with various 4:1 Baluns (large and small and, therefore, with many different lengths of transmission lines), and 3) McCoy's success with his 4:1 Balun. Also, as this chapter points out, it helps to have the dimensions of a multiband, center-fed dipole, and feeders favor the operation of a 4:1 Balun in antenna tuners. And, *yes*, there is a "worst case" antenna design!

As in many investigations, supplying answers to some questions can lead to others that appear to be important. Specifically, for powdered-irons, how would permeabilities in the 20 to 35 range perform? Simple impedance measurements showing lower values on input impedances, indicate that there is more loss than with a permeability of 10; but accurate insertion loss measurements are needed in order to tell the complete story—the trade-off in efficiency for low-frequency response.

Finally, low permeability ferrite-like No. 67 material with a permeability of 40 looks interesting for use in Baluns for antenna tuners. Accurate insertion loss measurements[2] have also shown the very same high efficiency that was exhibited by powdered-iron having a permeability of 10. With a sufficient number of turns on an appropriate size core, a Balun made of this material could be practical. Even though the amateur radio literature still refers to the problem of core saturation, there has been only one recorded case. This was on 2 MHz with a rod-type 1:1 Balun where insufficient choking reactance exists.[13,14]

Chapter 10

1.5:1 and 2:1 Baluns

Sec 10.1 Introduction

There are many applications for broadband Baluns with impedance transformation ratios close to 1.5:1 and 2:1. Two applications involve matching 50-ohm cable to balanced loads of 75 or 100 ohms, which are the input impedances of a half-wave dipole at heights of 0.22 or 0.34 wavelengths above ground. Another, is the matching of 50-ohm cable to the 100-ohm input impedance of a quad antenna. An interesting, and somewhat unexpected, application is the matching of 50-ohm cable directly to the input impedance of the driven element of a Yagi beam antenna of 33 or 25 ohms. This would eliminate the common hairpin matching network presently used to raise their input impedances to 50 ohms.

There are many versions of these two Baluns. They include: 1) high- and low-power designs, 2) designs matching 50-ohm cable to higher or lower impedances, 3) series- or parallel-type designs, 4) single- or dual-core designs, 5) dual-ratio designs, and 6) HF and VHF designs. The series-type Baluns use an Unun (unbalanced-to-unbalanced transformer) in series with a Guanella (current) Balun. More details on these Ununs are provided in later chapters. In this chapter, you'll read about many high-power designs capable of handling the full legal limit of amateur radio power. They are optimized for sufficient margins in choking reactance at their low frequency ends, and in efficiency throughout their passbands. Two of the 2:1 Baluns are specifically designed for 2-meter operation.

Sec 10.2 1.5:1 Baluns

In this section, I'll present two series-type 1.5:1 Baluns (actually 1.56:1, which should be close enough). They both use 1.56:1 Ununs in series with Guanella 1:1 Baluns. **Figures 10-1A** and **B** show their schematic diagrams. **Figure 10-1B** has an extra input (to a tap), which provides another ratio of 1.33:1.

The left-hand side of **Photo 10-A** shows a design using **Figure 10-1A** mounted in a CU 3006 minibox 5.25 inches long by 3 inches wide by 2.25 inches high (RadioShack has carried a similar enclosure). The 1:1.56 Unun has 4 quintufilar turns on a 1.5-inch OD

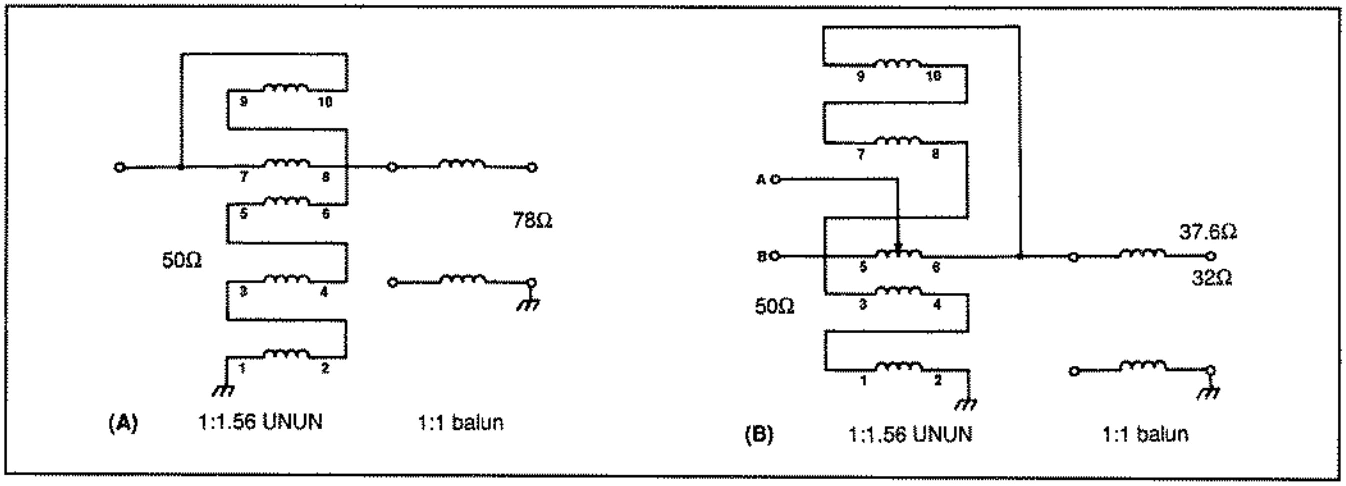

Figure 10-1. *Schematic diagrams of two 1.56:1 Baluns: (A) step-up, 50:78 ohms; (B) step-down, 50:37.6 ohms—connection A, 50:32 ohms—connection B.*

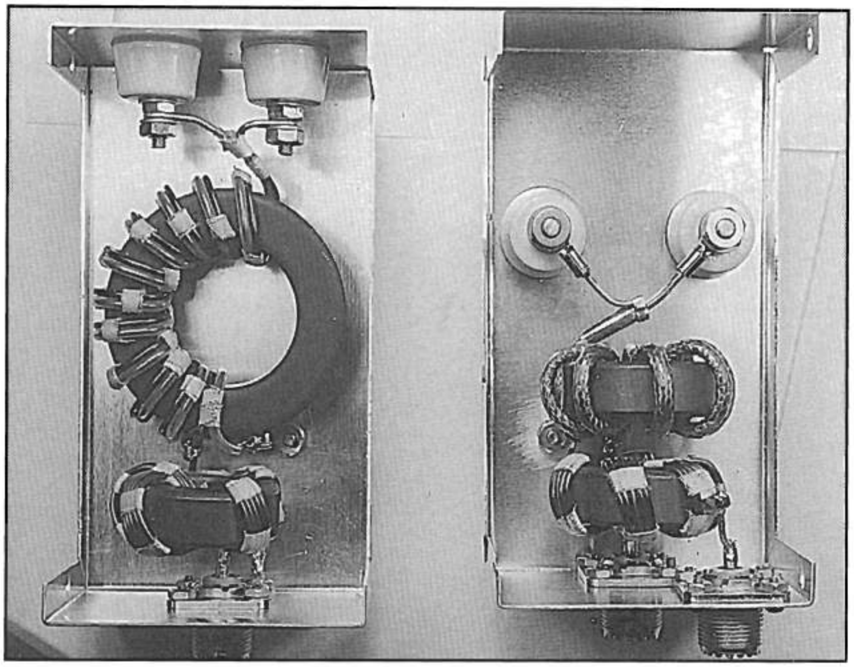

Photo 10-A. *Baluns using the schematic diagrams of Figure 10-1. Balun on the left matches 50-ohm cable to a balanced load of 78 ohms. Balun on the right matches 50-ohm cable to balanced loads of 37.6 or 32 ohms.*

ferrite toroid with a permeability of 250. Winding 7-8 is No. 14 H Thermaleze wire and the other four are No. 16 H Thermaleze wire.

The 1:1 Guanella Balun has 11 bifilar turns of No. 14 H Thermaleze wire on a 2.4-inch OD ferrite toroid with a permeability of 250. One wire is covered with Teflon tubing, resulting in a characteristic impedance very close to 78 ohms (the optimum value).

When matching 50-ohm cable to a balanced load of 78 ohms, the impedance transformation ratio is literally flat (within a percent or two) from 1.5 MHz to 40 MHz! You might be interested to know that (separately) the 1:1 (75:75-ohm) Balun would make an excellent isolation transformer for 75-ohm hardline, and the 1.56:1 (78:50-ohm) Unun an excellent match between 75-ohm hardline and 50-ohm cable.

The right-hand side of **Photo 10-A** shows a design using **Figure 10-1B** mounted in a similar enclosure. The 1.56:1 Unun has 5 quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire and is tapped at one turn from terminal 5. The other four wires are No. 16 H Thermaleze.

The 1:1 Guanella Balun has 7 turns of homemade coaxial cable on a 1.5-inch OD ferrite toroid with a permeability of 250. The inner conductor is No. 14 H Thermaleze wire and is covered with Teflon tubing. The outer braid, which is from a small coaxial cable or from 1/8-inch tubular braid, is also tightly wrapped with Scotch No. 92 tape to preserve the low characteristic impedance.

In matching 50-ohm cable to a balanced load of 37.6 ohms (connection A), or to a balanced load of 32 ohms (connection B), the response is essentially flat (within a percent or two) from 1.5 to 30 MHz.

Sec 10.3 2:1 Baluns

The 2:1 Balun lends itself to more choices in design than the 1.56:1 Balun. This is especially true because the parallel-type design, which provides a 2.25:1 Balun with the widest possible bandwidth, can easily be employed. The 1.56:1 Balun is at a disadvantage here. This section presents many Baluns using both series and parallel-type designs.

Sec 10.3.1 Series-type Baluns

Figure 10-2 shows circuit diagrams for two versions of the series-type Balun. **Photo 10-B** shows a design using **Figure 10-2A** mounted in a CU 3005-A minibox 5 inches long by 4 inches wide by 3 inches high. The 1:2 Unun has 7 trifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. The output tap is located 6 turns from terminal 5. Winding 5-6 is No. 14 H Thermaleze wire and the other two are No. 16 H Thermaleze wire.

The 1:1 Guanella Balun has 14 bifilar turns of No. 14 H Thermaleze wire on a 2.4-inch OD ferrite toroid with a permeability of 250. Both wires are covered with Teflon tubing, which results in a characteristic impedance of 100 ohms (the optimum value). A crossover, placing 7 turns on one side of the toroid

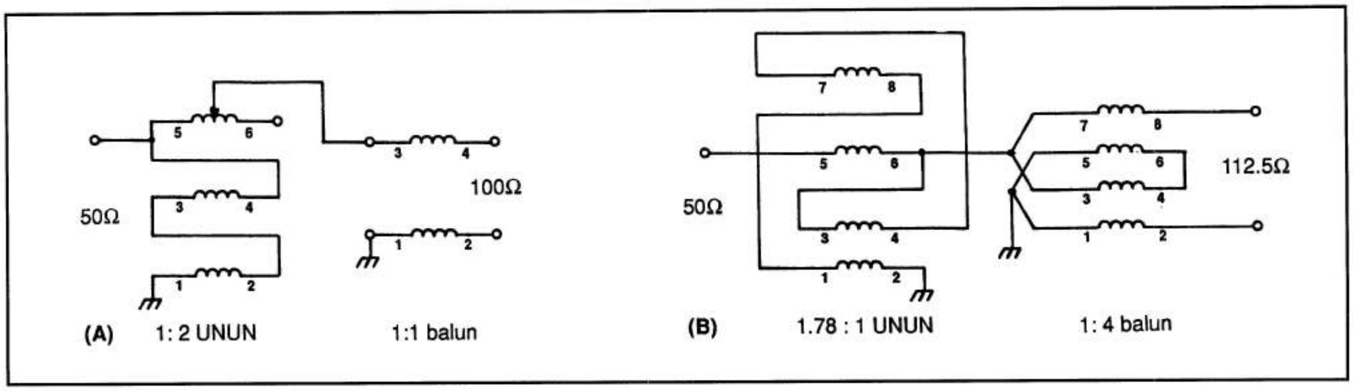

Figure 10-2. *Schematic diagrams of two versions of the series-type Balun: (A) 1:2 (50:100 ohms) Balun; (B) 1:2.25 (50:112.5 ohms) Balun.*

and 7 turns on the other, is used so the output and input are on opposite sides of the toroid. **Figure 10-3** is a drawing of the crossover. Although this technique has no electrical advantage at HF, the mechanical advantage is obvious.

When matching 50-ohm cable to a balanced load of 100 ohms, the response is literally flat (within 2 to 3 percent) from 1.5 to 30 MHz. By connecting the output of the Unun to terminal 6 instead of to the tap, the Balun would match 50-ohm cable to a balanced load of 112.5 ohms with about the same response.

Photo 10-C shows two slightly different versions of series-type 2.25:1 Baluns using the circuit of **Figure 10-2B**. Both have the same 1.78:1 step-down Unun, which has 5 quadrifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire, and the other three are No. 16 H Thermaleze wire. Each version also has 8 bifilar turns of No. 14 H Thermaleze wire on both of the 1.5-inch OD ferrite toroids, with a permeability of 250.

The differences are: 1) the Balun on the left in **Photo 10-C** has one layer of Scotch No. 92 tape on one of the wires in each bifilar winding and a crossover after the fourth turn, and 2) the Balun on the right has two layers of Scotch No. 92 tape on one of the wires on one toroid and no extra insulation on the wires of the other toroid. Therefore, one of the windings in the 1:4 Guanella Balun has a characteristic impedance a little less than 50 ohms and the other a little greater than 50 ohms, resulting in a canceling effect. Furthermore, the crossover isn't used in this design. The Balun on the left is mounted in a CU 3006 minibox 5.25 inches long by 3 inches wide by 2.25 inches high. The Balun on the right is mounted in a CU 3015-A minibox 4 inches long by 2 inches wide by 2.75 inches high.

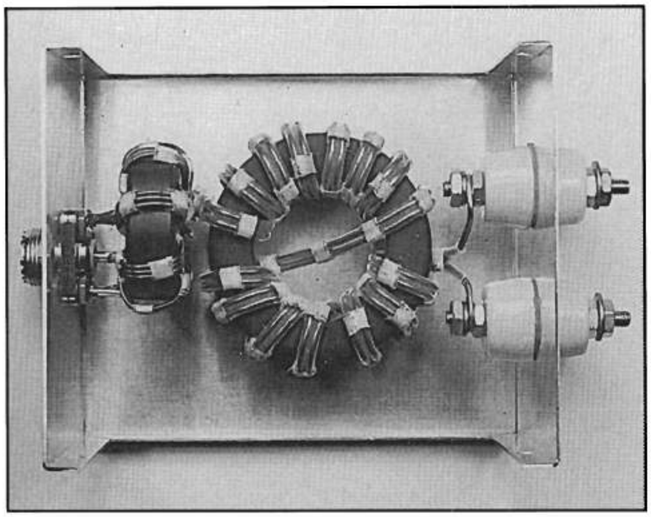

Photo 10-B. *A series-type Balun using the schematic diagram of Figure 10-2A designed to match 50-ohm cable to a balanced load of 100 ohms.*

The performance of these two Baluns is essentially the same. When matching 50-ohm cable to balanced loads of 112.5 ohms, the responses are essentially flat (within 2 to 3 percent) from 1.5 to 30 MHz.

From preliminary measurements on series-type 2:1 Baluns, the Balun in **Photo 10-B** is the one I'd recommend for matching 50-ohm cable to balanced loads of 100 ohms, while the Baluns in **Photo 10-C** would be best for matching to balanced loads of 112.5 ohms. Also, by replacing the 1.78:1 Unun in **Figure 10-2B** with a 2.25:1 Unun, and not adding any extra insulation to the windings of the 1:4 Balun, it's possible to obtain an excellent Balun matching 50-ohm cable to a balanced load of 89 ohms.

Figure 10-4 shows the schematic diagram of a series-type Balun designed to match 50-ohm cable to balanced loads of 25 or 22.22 ohms. **Photo 10-D**

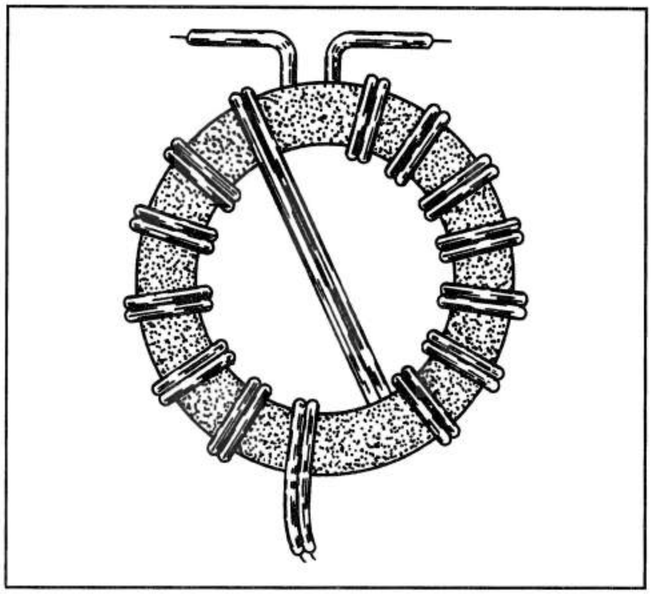

Figure 10-3. *Construction of a 1:1 Guanella Balun with a crossover placing the input and output terminals on opposite sides of the toroid.*

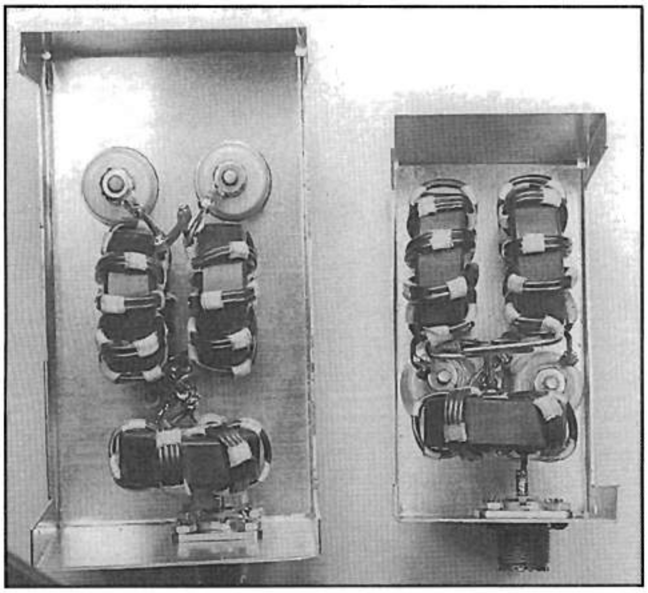

Photo 10-C. *Two series-type Baluns using the schematic diagram of Figure 10-2B designed to match 50-ohm cable to balanced loads of 112.5 ohms.*

shows two versions of this dual-ratio Balun. The Balun on the left, for its Unun, has 6 trifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 3-4 is tapped one turn from terminal 3, yielding the 2:1 ratio. The 1:1 Guanella Balun has 6 turns of homemade coaxial cable on a similar toroid. The inner conductor is No. 12 H Thermaleze wire and is covered with Teflon tubing. The outer braid, from a small coax or from 1/8-inch tubular braid, is tightly wrapped with Scotch No. 92 tape to preserve its low characteristic impedance.

In matching 50-ohm cable to balanced loads of 25 ohms (connection A) or to 22.22 ohms (connection B), the response is essentially flat (within a percent or two) from 1.5 to 30 MHz.

The Balun on the right in **Photo 10-D** has similar windings on a single 2.4-inch OD ferrite toroid with a permeability of 250. Its performance is quite comparable to the Balun on the left.

Sec 10.3.2 Parallel-type Baluns

As you saw in the previous section, the series-type Baluns presented here are combinations of Ununs with ratios of 1.33:1, 1.78:1, 2:1, and 2.25:1 in series with Guanella 1:1 or 4:1 Baluns. The Ununs, which are really an extension of Ruthroff's[9] bootstrap technique for obtaining a 4:1 Unun, sum direct voltages with delayed voltages that traverse a single transmission line. Therefore, the Unun eventually limits the high-frequency response of the series-type Balun.

On the other hand, the parallel-type Balun is an extension of Guanella's technique of summing voltages of equal delays. Instead of simply connecting transmission lines in parallel-series, the parallel-type Balun connects Guanella Baluns in parallel-series. As I noted in **Reference 25**, two 4:1 Guanella Baluns can be connected in parallel-series, yielding very broadband ratios of 6.25:1. This section shows how a 1:1 Guanella Balun can be connected with a 4:1 Guanella Balun in parallel-series, yielding a very broadband ratio of 2.25:1.

Figure 10-5 is the schematic diagram of the high-frequency model of a 2.25:1 Unun which is used for analysis purposes. Because the current through the load is $3/2I_1$, the transformation ratio is $(3/2)^2$, or 2.25:1. Therefore, if the impedance seen on the left side is 50 ohms, a matched impedance on the right side is 22.22 ohms. Because two thirds of the 50 ohms appears across the input of the Guanella 1:1 Balun, its optimum characteristic impedance is 33.33 ohms. Similarly, this is also the value of the optimum characteristic impedance for the windings of the 4:1 Balun. Because the 1:1 Balun wants to see 33.33 ohms on its output (a matched condition) and the 4:1 Balun wants to see 66.66 ohms, placing these two values in parallel results in the confirming value of 22.22 ohms.

If the 50-ohm generator is placed on the right side in **Figure 10-5**, the circuit becomes a step-up Unun matching 50 ohms to 112.5 ohms (on the left). If the ground is removed on the left side, the transformer becomes a Balun. A similar analysis as above, shows that the optimum characteristic impedance of the three bifilar windings now becomes 75 ohms.

Photo 10-E shows a parallel-type 2.25:1 Balun designed to match 50-ohm cable to a balanced load of 112.5 ohms. It has 9 bifilar turns of No. 14 H Thermaleze wire on each of the three toroids that have a 1.5-inch OD and a permeability of 250. Also, one of the wires on each toroid is covered with Teflon tubing, resulting in a characteristic impedance of 75 ohms (the optimum value). When operating as a Balun, the response is essentially flat from 7 MHz to over 45 MHz. As an Unun, the flat response is broadened from 1.5 MHz to over 45 MHz.

Because the coiled wire, parallel-type Balun didn't provide any real advantage over the series-type Balun (in fact, the low-frequency response was poorer), I investigated the beaded transmission line version for possible use in the VHF band. **Figure 10-6** shows a schematic diagram of one using coaxial cable. Obviously, twin lead could be substituted for the coaxial cable. **Photo 10-F** shows both versions.

The top design in **Photo 10-F** has 4 inches of 3/8-inch OD ferrite beads with a permeability of 125 on each of the three 75-ohm transmission lines. It is designed to match 50-ohm cable on the right to a balanced load of 112.5 ohms on the left (with the ground on terminal 1 removed). The transmission lines consist of two No. 14 H Thermaleze wires separated by the Teflon tubing covering one of them. When matching 50-ohm cable to a balanced load of 112.5 ohms, the response is essentially flat from 30 MHz to over 100 MHz (the limit of my bridge).

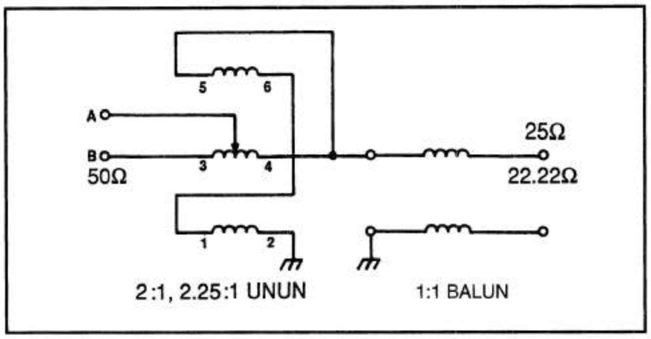

Figure 10-4. *Schematic diagram of a dual-ratio series-type Balun. Connection A matches 50-ohm cable to a balanced load of 25 ohms. Connection B matches 50-ohm cable to a balanced load of 22.22 ohms.*

The bottom design in **Photo 10-F** also has 4 inches of 3/8-inch OD ferrite beads with a permeability of 250. However, they are now threaded by homemade coaxial cable with a characteristic impedance of 33.33 ohms. It is designed to match 50-ohm cable on the left to a balanced load of 22.22 ohms (with the ground on terminal 2 removed). The inner conductor of the coax is No. 14 H Thermaleze wire and is covered with

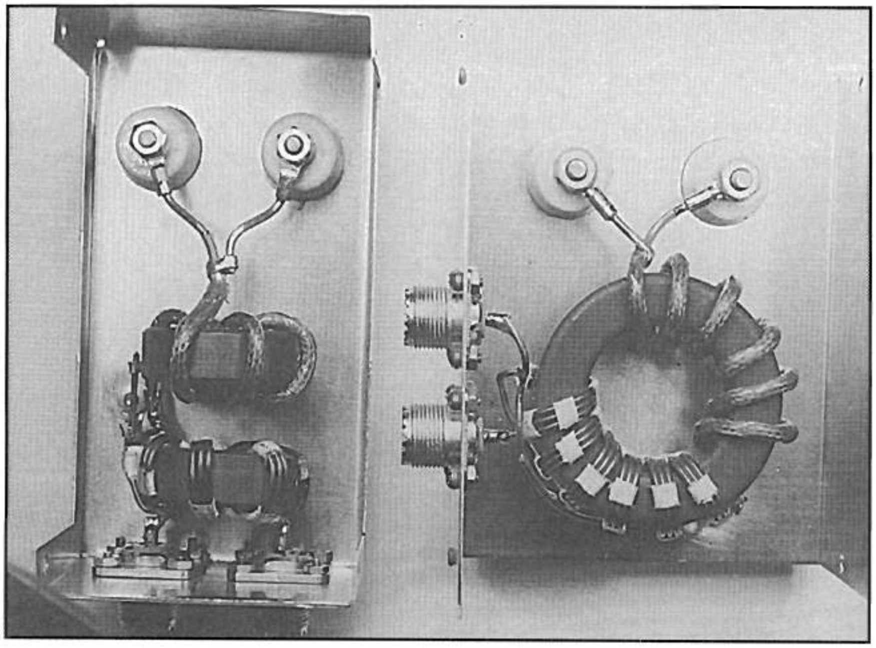

Photo 10-D. *Two series-type Baluns using the schematic diagram of Figure 10-4 designed to match 50-ohm cable to balanced loads of 25 ohms or 22.22 ohms.*

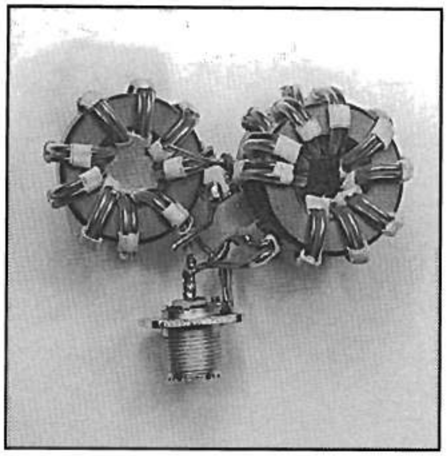

Photo 10-E. *Parallel-type 2.25:1 Balun matching 50-ohm cable to a balanced load of 112.5 ohms.*

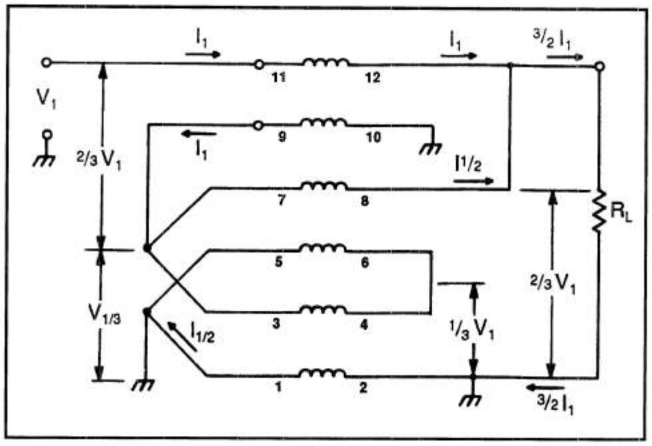

Figure 10-5. *High-frequency model of the parallel-type 2.25:1 transformer. Connections shown are for Unun operation.*

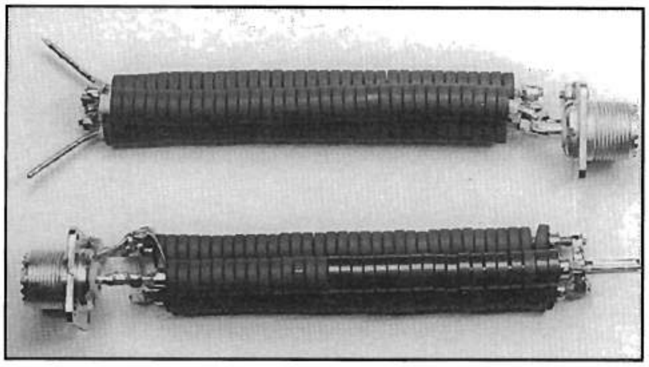

Photo 10-F. *Beaded transmission line versions of the 2.25:1 parallel-type Balun for operation in the VHF band. The top Balun matches 50-ohm cable to a balanced load of 112.5 ohms. The bottom Balun matches it to a balanced load of 22.22 ohms.*

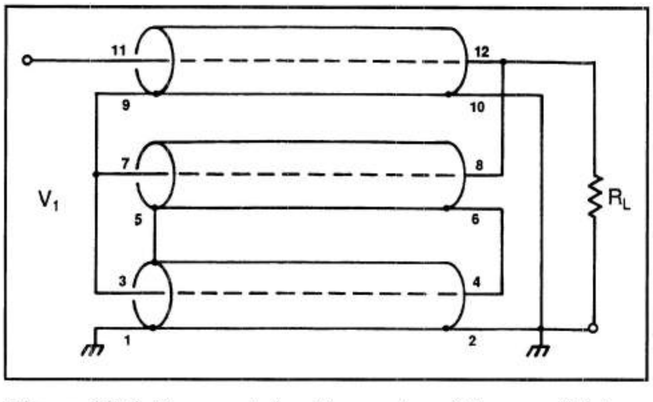

Figure 10-6. *The coaxial cable version of the parallel-type 2.25:1 transformer of Figure 10-5.*

Teflon tubing. The braid is from small coaxial cable or from 1/8-inch tubular braid. The braid is also tightly wrapped with Scotch No. 92 tape in order to preserve the low characteristic impedance. When matching 50-ohm cable to a balanced load of 22.22 ohms, the response is essentially flat from 14 MHz to over 100 MHz (again, the limit of my bridge).

Sec 10.4 Closing Comments

In closing, I'd like to make a couple of comments regarding parallel-type Baluns.

First, if you are interested in a 1.78:1 ratio, replace the 1:4 Balun in **Figures 10-5** and **10-6** with a 1:9 Guanella Balun (three transmission lines connected in parallel-series). This would yield an output current of $4/3I_1$ and a ratio of $(4/3)^2$, or 1.78:1. If you replace the 1:4 Balun with a 1:16 Guanella Balun (four transmission lines connected in parallel-series), the output current will be $5/4I_1$, with a ratio of $(5/4)^2$, or 1.56:1.

Second, because the parallel-type Balun is really an extension of Guanella's technique of summing voltages of equal delays,[26,27] the high-frequency response is mainly limited by the parasitics in the interconnections. Therefore, beaded transmission lines offer the best opportunity for successful operation on the VHF band. It is also recommended that the ferrite beads have permeabilities of 300 or less[2] in order to achieve the very high efficiencies of which these transformers are capable.

Chapter 11

6:1, 9:1, and 12:1 Baluns

Sec 11.1 Introduction

This chapter is a combination of two articles I wrote for *Communications Quarterly*. One, on the 6:1 and 9:1 Baluns, appeared in the Winter 1993 issue and the other, on the 12:1 Balun, ran in the Summer 1993 issue. I combined these two articles here because the three Baluns discussed have much in common. Commonalities include: 1) They are the most difficult to construct and are, therefore, the most expensive; 2) two of them, the 6:1 and 12:1 Baluns, use Ununs in series with Guanella Baluns; 3) two of them use the 9:1 Guanella Balun—the 12:1 Balun with a series 1.33:1 Unun; 4) they are generally associated with two-wire transmission lines with characteristic impedances of 300, 450, and 600 ohms, respectively; 5) when matching 50-ohm cable to higher impedances, they have more loss than the Baluns in the preceding chapters; and 6) the optimum application of these Baluns requires a more critical understanding of the trade-offs in low-frequency response for efficiency.

Like the 2:1 Balun in **Chapter 10**, the 6:1 (actually 6.25:1) Balun also comes in two forms: the series-type, which offers better low-frequency response in the HF band, and the parallel-type, which has a vastly greater high-frequency capability. The parallel-type 6:1 Balun, together with the 9:1 Guanella Balun (which is also a parallel-type), offer the potential for designs capable of efficient and broadband operation on the VHF and UHF bands.

Sec 11.2 6:1 and 9:1 Baluns

Many radio amateurs associate the use of the 6:1 and 9:1 Baluns with 300-ohm twin lead feeding folded dipoles, 450-ohm "ladder" line feeding single or multi-band antenna systems. However, what is neglected (in some cases) is the effect of the height of these antennas above earth and the length of the transmission lines feeding them.

Broadband 6:1 and 9:1 Baluns are considerably more difficult to construct than the more common 1:1 and 4:1 Baluns. This is especially true when matching 50-ohm cable to impedances of 300 and 450 ohms. Furthermore, there are some important trade-offs in low-frequency response for efficiency.

From what I could gather "on the air" or talking to radio clubs, I have determined what I believe are probably two of the most common misconceptions regarding the use of these Baluns:

1. 6:1 Baluns. In free-space, the folded dipole with 300-ohm twin-lead has a resonant impedance close to 300 ohms. The dipole also has this value at a height of about 0.225 wavelength above ground. However, it's only 200 ohms at a height of about 0.17 wavelength and 400 ohms (the maximum) and at 0.35 wavelength. In many cases, the 4:1 Balun would actually do a better job of matching.

2. 9:1 Baluns. Some are unaware of the relationship between the impedance at the input of a transmission line, the characteristic impedance of the line, and the impedance at the end of the line. Just because a transmission line has a characteristic impedance of 450 ohms doesn't necessarily mean that a 9:1 Balun will perform a satisfactory match to 50-ohm cable. Far from it. For example, if the line is terminated by a half-wave dipole with an impedance of 50 ohms, the 9:1 Balun would see 50 ohms when the line is a half-wave long and 4050 ohms when it's a quarter-wave long! Broadband Baluns cannot be designed to handle impedances as high as 4050 ohms. It's very likely that a well-designed 50:450-ohm Balun would experience (particularly on 80 and 160 meters) harmful flux in the core *and* excessive heating because of the large voltage drop along the length of its transmission lines. This problem of presenting very high (and harmful) impedances to Baluns is quite prevalent with multi-band antenna systems.

Clearly, there are many applications for 6:1 and 9:1 Baluns. They not only include matching 50-ohm cable

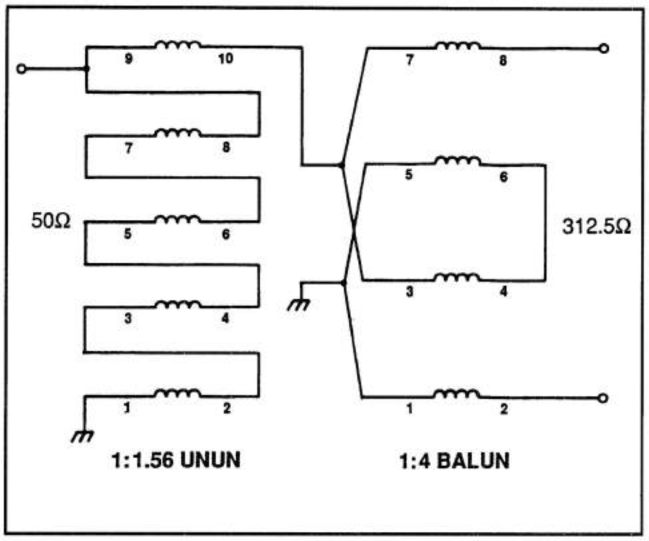

Figure 11-1. *Schematic diagram of the series-type Balun with a 1:6.25 ratio designed to match 50 to 312.5 ohms.*

to balanced loads of 300 and 450 ohms, but also to balanced loads of 8 and 5.6 ohms, as well. Furthermore, many of the designs in this chapter will perform almost as well in Unun (unbalanced-to-unbalanced) applications. The trade-off (which is usually very small) is in low-frequency response. Additionally, these Baluns could be used to exploit the low-loss properties of 300- and 450-ohm twin-lead where very long transmission lines are used. This is especially important at 14 MHz and above.

In the pages that follow, I'll present a variety of Baluns matching 50-ohm cable to 300 ohms (actually to 312.5 ohms, a 6.25:1 ratio) and to 450 ohms, as well as 50-ohm cable to 8 and 5.6 ohms. Also included are two different versions of 6.25:1 Baluns. One is a series-type using a 1.56:1 Unun in series with a 4:1 Guanella Balun and the other a series-parallel arrangement using two 4:1 Guanella Baluns. Because the series-parallel Balun adds voltages of equal delays, you'll find its high-frequency capability is much greater.

The 9:1 Balun is a conventional Guanella Balun with three transmission lines connected in series at the high-impedance side and in parallel at the low-impedance side. Therefore, it also sums voltages of equal delays. Some of the comparisons and analyses of these 6.25:1 and 9:1 Baluns were probably published for the first time in my Winter 1993 *Communications Quarterly* article.

Sec 11.2.1 6.25:1 Series-Type Baluns

Figure 11-1 shows the schematic diagram of a series-type Balun designed to match 50-ohm cable to a balanced load of 312.5 ohms. It consists of a 1:1.56 Unun in series with a 1:4 Guanella Balun. The overall ratio of 1:6.25 should satisfy most of the 1:6 requirements. **Photo 11-A** shows three examples. All three Baluns use the same step-up Unun that has four quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 9-10 is No. 14 H Thermaleze wire and the other four are No. 16 H Thermaleze wire. Because this Unun sums only one delayed voltage with four equal direct voltages, it has an excellent high-frequency response.[2]

The Balun on the left in **Photo 11-A** has eight bifilar turns of No. 18 hook-up wire on each transmission line of its 1:4 Balun. The wires are further spaced with No. 18 Teflon tubing providing a characteristic impedance close to 150 ohms (the optimum value). The ferrite toroid has a 2.4-inch OD and a permeability of 250. When matching 50-ohm cable to a floating

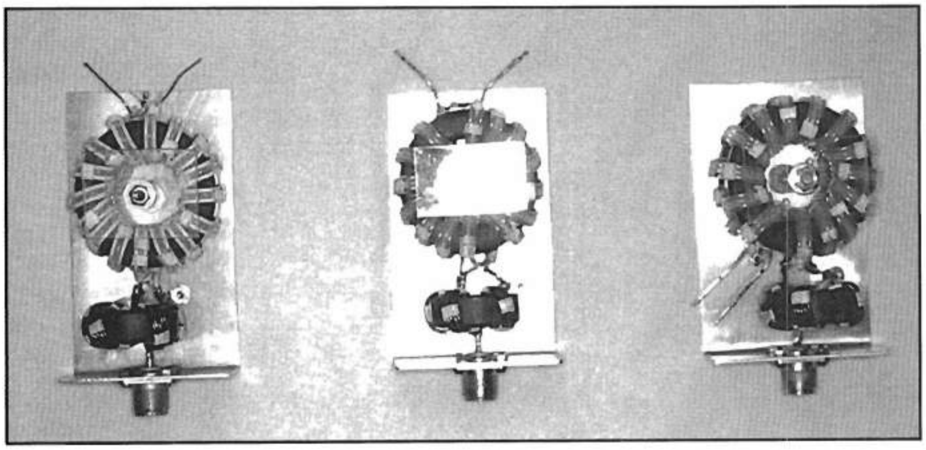

Photo 11-A. *Three examples of series-type 1:6.25 Baluns. The Balun on the right, with a double-core 1:4 Balun, has both a balanced voltage and current output. The other two only have balanced-current outputs.*

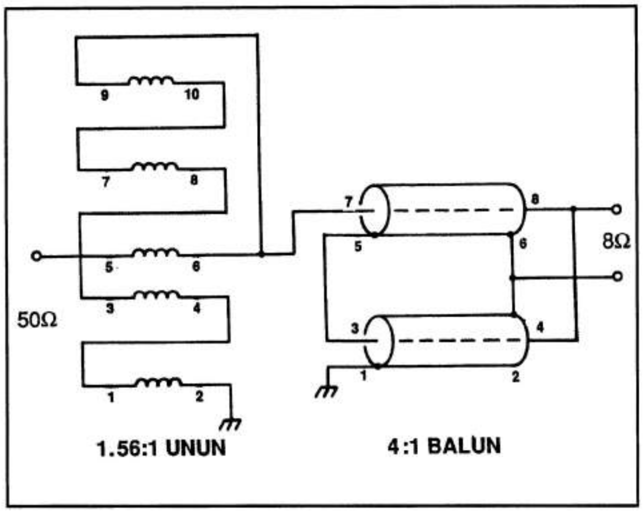

Figure 11-2. *Schematic diagram of the series-type Balun with a 6.25:1 ratio designed to match 50 to 8 ohms.*

load of 312.5 ohms, the response is essentially flat from 1.7 to 30 MHz. Under matched conditions, 500 watts of continuous power and 1 kW of peak power is a conservative power rating. Because the 1:4 Balun in this series-type 1:6.25 Balun uses only one core instead of two, this transformer should never be used when the load is grounded at its center. Also, it is not recommended for balanced antennas. This series-type Balun presents balanced currents, but does *not* present balanced voltages.

The Balun in the center of **Photo 11-A** has seven bifilar turns of No. 16 SF Formvar wire on each transmission line on its 1:4 Balun. The wires are covered with Telfon sleeving and further separated by No. 16 Teflon tubing. Like the Balun on the left, the characteristic impedance is also close to the objective of 150 ohms. The toroid also has a 2.4-inch OD and a permeability of 250. When matching 50-ohm cable to a floating load of 312.5 ohms, the response is essentially flat from 3.5 to 30 MHz. Over this frequency range, this Balun can easily handle the full legal limit of amateur radio power. Because this Balun also presents balanced currents and not balanced voltages, it should not be used when the loads are balanced to ground or grounded at their centers.

The Balun on the right in **Photo 11-A** has 14 bifilar turns of No. 16 SF Formvar wire on each of the two toroids of the 1:4 Balun. The wires are also covered with Teflon sleeving and further separated by No. 16 Teflon tubing. For ease of connection, one core is wound clockwise and the other counterclockwise. The two cores are spaced 1/4 inch apart with acrylic sections. When matching 50-ohm cable to a 312.5 load that is either floating, balanced to ground, grounded at its center, or grounded at its bottom (a broadband Unun), the response is essentially flat from 1.7 to 30 MHz. Under this matched condition, it can easily handle the full legal limit of amateur radio power. Furthermore, this is a true Balun because it presents equal currents and equal voltages. If one were to measure the voltages-to-ground, when matching into a balanced load, one would find them to be equal and opposite. The other two 1:6.25 Baluns using single-core 1:4 Baluns, would have equal currents but not equal voltages (see **Chapter 8**). Because they are easier to construct, it would be interesting to compare them with a true Balun.

Figure 11-2 shows the schematic diagram of a series-type Balun designed to match 50-ohm cable to a balanced load of 8 ohms (perhaps a short-boom Yagi). It consists of a 1.56:1 step-down Unun in series with a Guanella 4:1 step-down Balun. The overall ratio is 6.25:1. **Photo 11-B** shows two examples. Both Baluns use the same step-down Unun, which has four quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire and the other four are No. 16 H Thermaleze wire. The interleaving of the wires is such that the performance is optimized for matching 50 to 32 ohms.

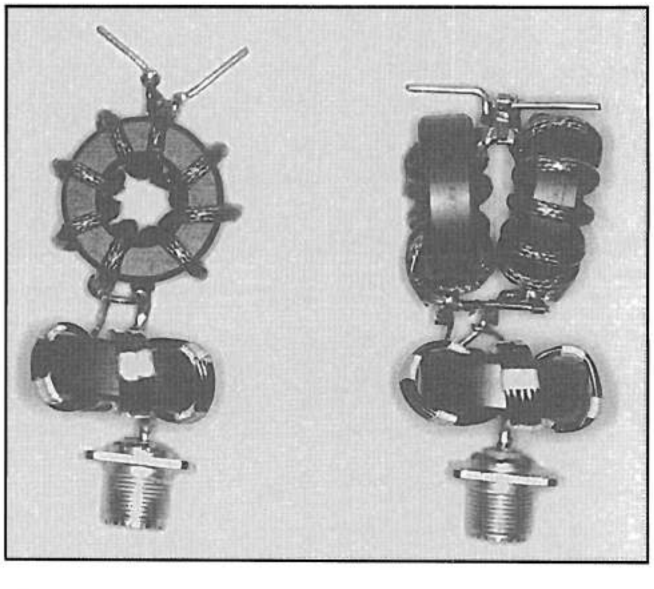

Photo 11-B. *Two examples of the series-type 6.25:1 Balun optimized at the 50:8-ohm level. The Balun on the left is designed to match into a floating 8-ohm load. The Balun on the right is designed to match into an 8-ohm floating, center-tapped-to-ground or grounded load (Unun).*

The Balun on the left in **Photo 11-B** has four turns of low-impedance coaxial cable on each transmission line on the single-core 4:1 Balun. The inner conductor is No. 14 H Thermaleze wire, and it has two layers of Scotch No. 92 polyimide tape. The outer braid is from a small coax (or 1/8-inch tubular braid) and is tightly wrapped with Scotch No. 92 tape to achieve the 17-ohm characteristic impedance (the optimum value). The ferrite toroid has a 1.5-inch OD and a permeability of 250. When matching 50-ohm cable to a floating load of 8 ohms, the response is flat from 1 to 40 MHz. In a matched condition, this Balun can easily handle the full legal limit of amateur radio power.

The Balun on the right in **Photo 11-B** has six turns of the same coaxial cable on each of the two cores of the Guanella step-down Balun. The cores also have a 1.5-inch OD and a permeability of 250. The performance of this Balun is practically the same as the above with the single-core 4:1 Balun. The important differences are that this 6.25:1 Balun also performs equally well whether the load is center-tapped-to-ground, balanced-to-ground, or grounded at the bottom (a broadband Unun). This is the one recommended for feeding a short-spaced Yagi beam antenna.

Sec 11.2.2 6.25:1 Parallel-Type Baluns

The 6.25:1 series-type Baluns described in the preceding section consisted of a 1.56:1 Unun, which is an extension of Ruthroff's bootstrap approach for Ununs,[9] in series with a Guanella 4:1 Balun.[2] The upper-frequency limit for this combination is really set by the Unun, which sums a delayed voltage with four direct voltages. The parallel-type 6.25:1 Baluns described in this section are really extensions of Guanella's approach, which sums voltages of equal delays. Therefore, the upper-frequency limit is mainly dependent upon the parasitics in the interconnections.

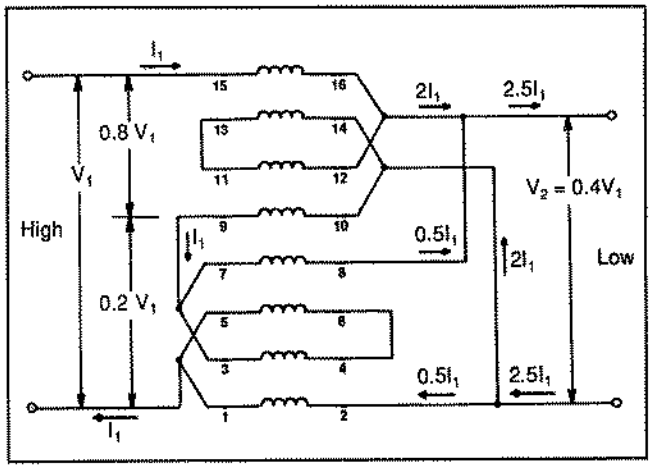

Figure 11-3. *Schematic diagram of the parallel-type Balun (and Unun) with a 6.25:1 ratio. The currents and voltages are shown for analysis purposes (see text).*

The 6.25:1 parallel-type Balun uses two 4:1 Guanella Baluns connected in parallel on the low-impedance side and in series on the high-impedance side. As you will see, one of the Baluns is reversed, giving the desired ratio of 6.25:1. Other combinations can produce different fractional-ratios (other than $1{:}n^2$ where n is 1, 2, 3, . . .), like 2.25:1 and 1.78:1. Because very little practical design information is available regarding this family of very broadband Baluns,[26,27] this section also includes my high-frequency analysis of the 6.25:1 parallel-type Balun. It should also be pointed out that very little sacrifice in performance occurs whether the load is grounded at its center or at the bottom (as an Unun).

Figure 11-3 shows the coiled-wire version of the 6.25:1 parallel-type Balun. For analysis purposes, the voltages and currents are also shown. As can be seen, the top 4:1 Balun is connected as a step-down Balun, while the bottom 4:1 Balun is connected as a step-up Balun. The Baluns are in series on the high-impedance side (on the left) and in parallel on the low-impedance side (on the right). As **Figure 11-3** illustrates, the lower 1:4 Balun adds a current of $0.5I_1$ to the load, resulting in a total current of $2.5I_1$. Thus, the impedance transformation ratio is 2.5^2, or 6.25:1.

For maximum high-frequency response, each transmission line should see a load equal to its characteristic impedance. In other words, they should be "flat" lines. If the high side on the left is 50 ohms, then 40 ohms appears on the input of the top Balun and 10 ohms on the input of the bottom Balun. Consequently, the optimum characteristic impedance for *all* transmission lines is 20 ohms. On the low-impedance side on the right, the top Balun wants to see 10 ohms, while the bottom Balun wants to see 40 ohms. Because 10 ohms in parallel with 40 ohms equals 8 ohms, each Balun conveniently sees its ideal load and a broadband ratio of 6.25:1 is obtained.

If the Balun is required to match 50-ohm cable (on the right side) to a balanced load of 312.5 ohms (on the left side), the same analysis shows that the optimum characteristic impedance of all the transmission lines is 125 ohms.

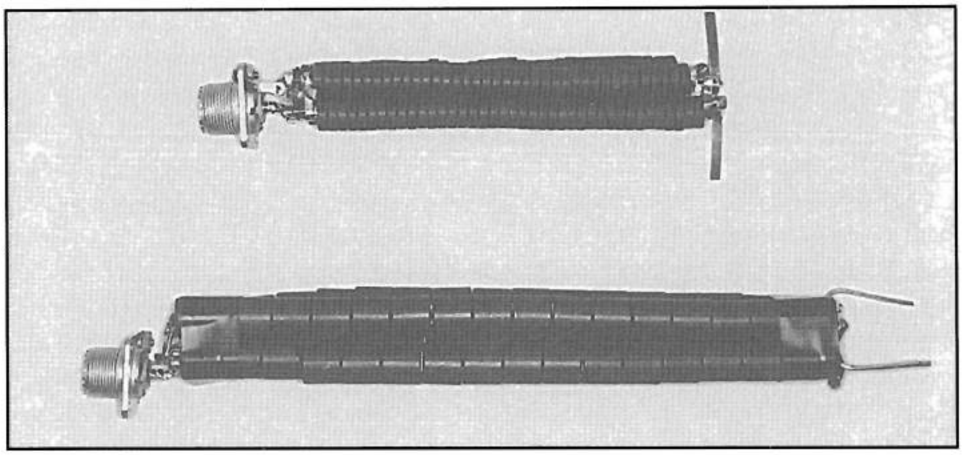

Photo 11-C. *Two beaded-line versions of the parallel-type 6.25:1 Balun (and Unun). The top transformer is designed to match 50-ohm cable to 8 ohms. The bottom transformer is designed to match 50-ohm cable to 312.5 ohms.*

Because the parallel-type Balun (or Unun) sums voltages of equal delays and, therefore, has no built-in high-frequency cut-off, it has a real advantage over the series-type Balun on the VHF bands and above. Furthermore, beaded transmission lines with low-permeability ferrite beads (125 and less) can be used, resulting in high efficiencies. On the HF band, where coiled windings are generally used, the series-type Balun is preferred because of its simplicity.

Photo 11-C shows two beaded-line 6.25:1 transformers. The top Balun, designed to match 50-ohm cable (on the left) to 8 ohms (on the right), uses low-impedance coaxial cable lines. The schematic diagram is shown in **Figure 11-4**. It has 5 inches of 0.375-inch OD beads (permeability 125) on four coaxial cables with characteristic impedances of 20 ohms. The inner-conductors of No. 12 H Thermaleze wire have two layers of Scotch No. 92 polyimide tape. The outer braids, from small coaxial cable or 1/8-inch tubular braid, are also wrapped tightly with the same tape in order to preserve the 20-ohm characteristic impedance. When matching 50 ohms to 8 ohms, the response is essentially flat from 10 MHz to beyond 100 MHz (the limit of my simple bridge). Under this matched condition, this Balun can easily handle the full legal limit of amateur radio power. Furthermore, it has practically the same performance when operating as an Unun (both terminals 1 and 2 grounded). In the Unun application, the bottom transmission line has no voltage along it and, therefore, requires no beads.

The bottom Balun in **Photo 11-C**, which is designed to match 50-ohm cable to a balanced load of 312.5 ohms, has 8 inches of 0.5-inch OD beads on 125-ohm twin-lead transmission lines. The ferrite beads also have a permeability of 125. The wires are No. 14 H Thermaleze wire and are covered with Teflon sleeving. They are further separated by No. 18 Teflon tubing. When matching 50-ohm cable (on the right side) to 312.5 ohms (on the left side), the response is essentially flat from 20 MHz to over 100 MHz. Under this matched condition, this Balun can also easily handle the full legal limit of amateur radio power. Additionally, this transformer performs practically as well when used as an Unun.

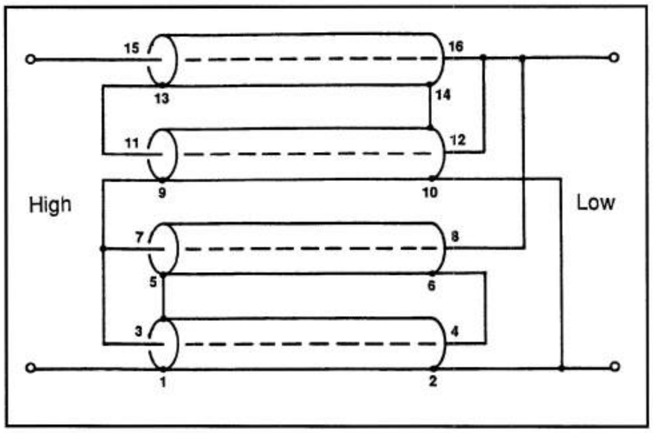

Figure 11-4. *Schematic diagram of the coaxial-cable version of the parallel-type 6.25:1 Balun (and Unun).*

Sec 11.3 9:1 Baluns

The broadband 9:1 Balun, matching 50-ohm cable to a balanced load of 450 ohms, is one of the most difficult ones to construct because high-impedance transmission lines (150 ohms) are required for maximum high-frequency response, and greater reactances are needed in order to isolate the input from the output. So one can appreciate the task at hand, this section also provides a brief review of the theory of these devices.[2]

Figure 11-5 shows the high- and low-frequency models of the Guanella 9:1 Balun that connect three transmission lines in series at the high-impedance side and in parallel at the low-impedance side. Because Guanella's Baluns (which can be easily converted to Ununs) sum voltages of equal delays, they offer the highest frequency capability.

The high-frequency model (**Figure 11-5A**) assumes there is sufficient choking reactance in the coiled (or beaded) transmission lines to isolate the input from the output and only allow transmission line currents to flow. Under this condition, the analysis is rather straightforward as it only involves transmission line theory. Simply stated—the maximum high-frequency response occurs when each transmission line is terminated in a load equal to its characteristic impedance, Z_0. Thus, the transmission lines in the 9:1 Balun have no standing waves. Because each transmission line sees one third of the load, the optimum value of Z_0 is $R_L/3$. Except for parasitics in the interconnections and self-resonances in coiled windings, Guanella's approach is literally "frequency independent."

On the other hand, the low-frequency analysis of the Guanella 9:1 Balun is most important because it reveals the major difficulty in designing them for low-loss, wideband operation. **Figure 11-5B** is the model for determining the low-frequency response. It assumes that no energy is transmitted to the load by a transmission line mode. Although the terminology and analysis is the same as that used for conventional autotransformers, the similarity ends when there is sufficient choking reactance to only allow for the efficient transmission line mode.

As with conventional transformers, one can analyze the low-frequency response of the 9:1 Balun from either the low- or high-impedance side. By putting the generator on the low-impedance side in **Figure 11-5B**, I've chosen to analyze it from that side. With the output open-circuited, the generator sees four coiled (or beaded) lines connected in series-parallel. The net result is that the generator sees the reactance of only *one* coiled (or beaded) line. To prevent a shunting current to ground (and/or autotransformer operation), the reactance the generator sees should be much greater than R_g (at least by a factor of 10 at the lowest frequency of interest). The inductance of the coiled or beaded line that prevents the unwanted currents is still known as the magnetizing inductance, L_M.

What's important to note here is that the low-frequency model of the Guanella 4:1 Balun does not have the series-parallel combination of coiled or beaded lines.[2] Only two lines, which are in series, exist in its model. Therefore, for a two-core Guanella 4:1 Balun having the same number of turns (and same cores) as a 9:1 Guanella Balun, its low-frequency response is better by a factor of two!

Another advantage that goes to the Guanella 4:1 Balun when matching 50 to 200 ohms, is in the number of turns that can be wound on the same cores. Since 4:1 Baluns require characteristic impedances of 100 ohms (instead of 150), the width of the transmission lines is considerably less, thus allowing for more turns. Also, as will be shown later, the efficiency of the 4:1 Balun is greater because the potential drops

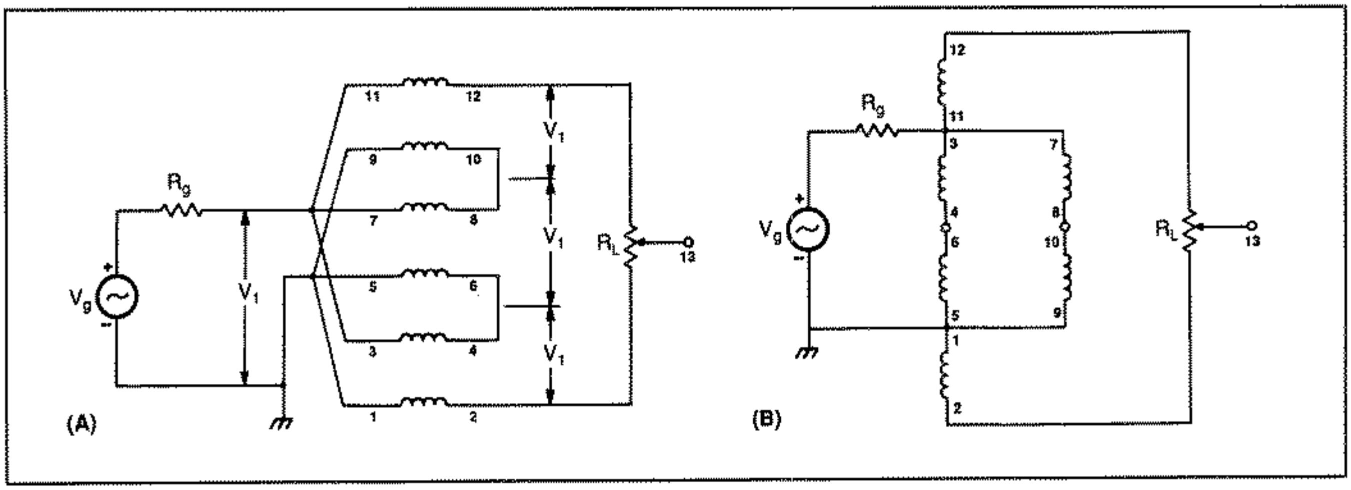

Figure 11-5. *Models of Guanella's 1:9 Balun. The high-frequency model, (A), assumes that $Z_o = R_L/3$, and therefore V_2, the output of each transmission line, equals V_1. The low-frequency model, (B), assumes no energy is transmitted to the load, R_L, by a transmission line mode.*

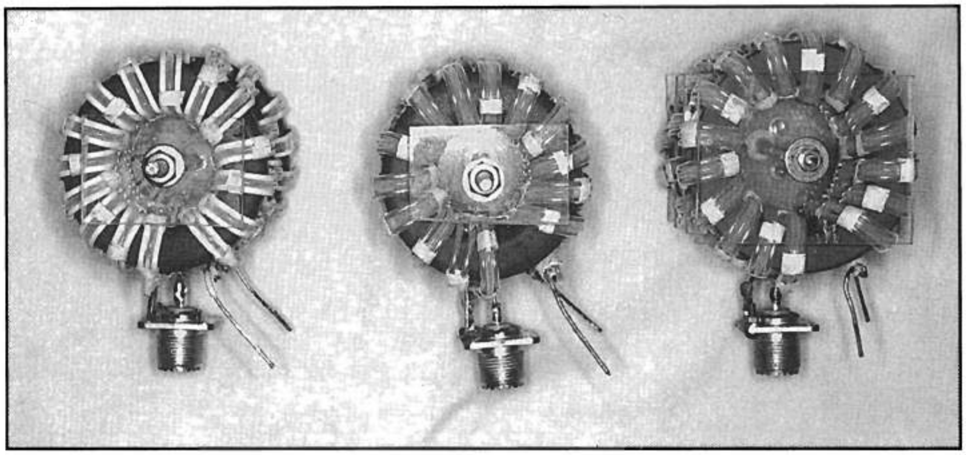

Photo 11-D. *Three broadband Guanella 9:1 Baluns designed to match 50-ohm cable to 450 ohms. The transformer on the left, using No. 18 hook-up wire, can handle 500 watts from 1.7 to 45 MHz. The transformer in the center, using No. 16 wire, can handle 1 kW from 3.5 to 45 MHz. The transformer on the right (with larger cores), using No. 16 wire, can handle 1 kW from 1.7 to 45 MHz.*

along the transmission lines are lower (less dielectric loss). Finally, as you can see from **Figure 11-5B**, by also grounding terminal 2 (Unun operation), windings 1-2 and 3-4 are both shorted—degrading the low-frequency response because L_M is reduced by one third.

Another interesting analysis with Baluns and Ununs concerns the potential gradients (voltage drops) along the transmission lines. Because the loss with these transformers, when transferring the energy via a transmission line mode, is a dielectric-type (voltage dependent), the higher the gradient, the greater the loss. The interesting cases occur when the load is: a) floating, b) grounded at the center, and c) grounded at the bottom (an Unun).

Floating load. With terminal 13 in **Figure 11-5A** ungrounded, the top transmission line has a gradient of $+V_1$ and the bottom transmission line has a gradient of $-V_1$. The center transmission line has a gradient of zero. Therefore, the center transmission line only acts as a delay line and doesn't require a magnetic core or beads. As a result, the top and bottom cores (or beads) account for the dielectric loss.

Load grounded at the center. With terminal 13 grounded at the center of the load, the top transmission line has a gradient of $+V_1$, the bottom transmission line has a gradient of $-V_1$, and the center transmission line has a gradient of $-V_1/2$. This configuration results in about 25 percent more loss because of the extra gradient along the center transmission line. Incidentally, this condition exists when matching into balanced systems like 450-ohm transmission lines or antennas because they have virtual grounds at the center of the loads they present.

Load grounded at the bottom. With terminal 13 grounded at the bottom of the load (an Unun), the top and center transmission lines have gradients of $+V_1$. The bottom transmission line has no gradient and, therefore, no loss. It only acts as a delay line and thus requires no magnetic core or beads.

Sec 11.3.1 Some Practical 9:1 Balun Designs

Photo 11-D shows three versions of the broadband Guanella 9:1 Balun designed to match 50-ohm cable to 450-ohm loads. The transformer on the left has 15 bifilar turns of No. 18 hook-up wire on each of the three ferrite toroids with a 2.4-inch OD and permeability of 250. The wires are further separated by No. 16 Teflon tubing, resulting in a characteristic impedance close to 150 ohms (the optimum value). The cores in this Balun, as well as the other two that follow, are spaced 1/4 inch apart by sections of acrylic. In matching 50 to 450 ohms, the response is essentially flat from 1.7 to 45 MHz. In this matched condition, this transformer can easily handle 500 watts of continuous power and 1 kW of peak power.

The transformer in the center of **Photo 11-D** has 14 bifilar turns of No. 16 SF Formvar wire on each of the three ferrite toroids with a 2.4-inch OD and permeability of 250. The wires are covered with Teflon sleeving and further separated by No. 16 Teflon tubing. The characteristic impedance is also close to the optimum value of 150 ohms. In matching 50 to 450 ohms, the response is essentially flat from 3.5 to 45 MHz. In this matched condition, this transformer can easily handle 1 kW of continuous power and 2 kW of peak power. **Photo 11-E** shows this Balun mounted in a minibox 6 inches long by 5 inches wide by 4 inches high.

The transformer on the right in **Photo 11-D** is designed to handle 1 kW of continuous power and 2 kW

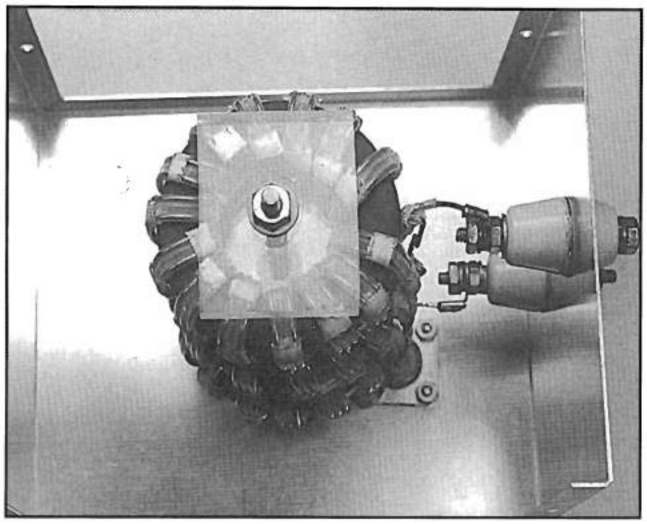

Photo 11-E. *The Guanella 9:1 Balun shown mounted in a large minibox.*

of peak power from 1.7 to 45 MHz. Because this transformer uses larger cores with slightly higher permeabilities (2.68-inch OD and 290 permeability), which are not as popular as those used in the other two Baluns, it is much more expensive to construct. It has 16 bifilar turns of No. 16 SF Formvar wire on each toroid. The wires are covered with No. 16 Teflon sleeving and further separated with No. 16 Teflon tubing. In matching 50 to 450 ohms, the response is flat from 1.7 to 45 MHz.

Finally, **Figure 11-6** shows the schematic diagram of a 9:1 Balun (or Unun) using coaxial cables wound around ferrite cores or threaded through ferrite beads. This form of the transformer is especially useful when matching 50-ohm cable to 5.6 ohms because the choking reactance of the magnetizing inductance, L_M, only need be much greater than 5.6 ohms. **Photo 11-F** shows two different designs.

The transformer on the top has 9 1/2 turns of low-impedance coaxial cable on each rod. The rods are 1/2 inch in diameter, 2 1/2 inches long, and have a permeability of 125. The low-frequency response of this Balun is quite insensitive to the length and permeability of the rods.[2] The inner conductors of the coaxial cables are No. 12 H Thermaleze wire with two layers of Scotch No. 92 tape. The outer braids are made from small coax (or 1/8-inch tubular braid). They are further wrapped with Scotch No. 92 tape to preserve the low-impedance of 17 ohms. In matching 50-ohm cable to 5.6 ohms, the response is essentially flat from 1.7 to 30 MHz. The optimum impedance level was found when matching 40 to 4.45 ohms. The addition of another layer of Scotch No. 92 tape would optimize this transformer at the 50:56-ohm level, and the high-frequency response would exceed 100 MHz. This transformer is very efficient and should handle the full legal limit of amateur radio power easily.

The bottom transformer in **Photo 11-F** is designed to match 50-ohm cable to a load of 5.6 ohms in the VHF band. It uses 3 1/2 inches of beads on three low-impedance coaxial cables. The ferrite beads have an OD of 3/8 inch and a permeability of 125. The inner conductors of the coaxes are No. 12 H Thermaleze wire with one layer of Scotch No. 92 tape. The outer conductors are from small coaxes or 1/8-inch tubular braid, and are also tightly wrapped with Scotch No. 92 tape to preserve the low characteristic impedance. In matching 50-ohm cable to 5.6 ohms, the response is essentially flat from 7 MHz to over 100 MHz (the limit of my test equipment). This 9:1 Balun (which can be used as an Unun) can handle the full legal limit of amateur radio power under matched conditions, because of the low-permeability beads and the low-voltage gradients along the lengths of its transmission lines.

Sec 11.4 Concluding Remarks on 6:1 and 9:1 Baluns

One of the most important properties of broadband Baluns and Ununs (which all use ferrites) is their capability of having extremely high efficiencies. Knowing the loss mechanism in these transformers and the trade-off in low-frequency response for efficiency allows one to optimize their applications. In the paragraphs that follow, I'll discuss the losses and

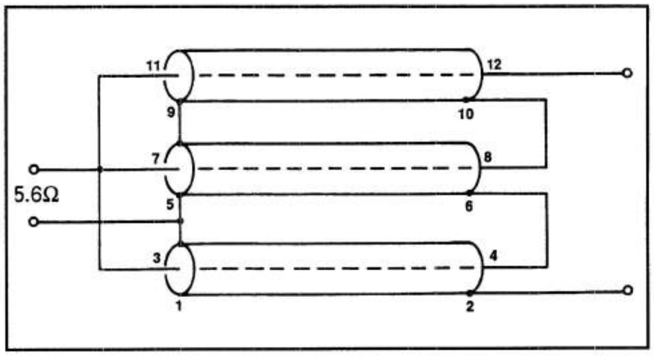

Figure 11-6. *Schematic diagram of a coaxial cable (beaded or coiled) 9:1 Guanella Balun (or Unun). This design is especially useful in matching 50-ohm cable to 5.6 ohms over a very wide bandwidth.*

trade-offs involved with the transformers presented in the preceding sections. The approach used here should be applicable to all forms of transmission line transformers. This section ends with a review of two articles that contained 9:1 Baluns. As you'll see, I have some rather different views on the claims made in these articles.

Accurate measurements on many broadband Ununs have found the losses to be related to the permeability and the impedance level.[2] Permeabilities greater than 300 resulted in excessive losses. Because these losses are unlike the conventional transformer whose losses are current-dependent, it can only be assumed that their losses are voltage dependent; in other words a dielectric-type loss. Therefore, higher-impedance transformers have higher voltage gradients along their transmission line and, thus, have greater losses. Additionally, it was found that the higher the permeability, the greater the loss with frequency. Taking into account the accurate measurements and the factors noted above, I offer these loss values for the transformers in the preceding sections:

1.56:1 Ununs. The 1.56:1 Ununs (either step-up or step-down) used in series with Guanella 4:1 Baluns to form 6.25:1 Baluns have the lowest potential gradients along their transmission lines. Voltage drops of only about 0.2 V_1, where V_1 is the input voltage, exist along their transmission lines. Accurate measurements have shown losses, in a matched condition, of only 0.04 dB. If the cores, which have a permeability of 250, were replaced with cores having a permeability of 125, the losses could be as low as 0.02 dB over much of the passband. The low-frequency response would still be acceptable at 1.7 MHz. This Unun is a natural for matching into 75-ohm hard line when long transmission lines are required.

6.25:1 and 9:1 Low-impedance Baluns. Baluns matching 50-ohm cable to 8 or 5.6 ohms, also have very low voltage drops along their transmission lines. Generally, they are about twice that of the 1.56:1 Unun. Therefore, the losses with these Baluns should be on the order of 0.1 dB in their passbands.

6.25:1 High-impedance Baluns. The losses in the series-type Baluns are mainly in the 1:4 Guanella Baluns, which have potential gradients of about 1.25 V_1, where V_1 is the input voltage. From previous measurements at this impedance level, the suggestion is that the losses (with ferrites of 250 permeability) should be about 0.1 dB at 7 MHz and 0.2 dB at 30 MHz. By using toroids with permeabilities of 125, the losses could be 0.07 dB and 0.15 dB, respectively. However, with a permeability of 40, the losses could be as low as 0.05 dB within the passband. However, one must consider the sacrifice in low-frequency response incurred when using these lower-permeability ferrites. With a permeability of 125, it's poorer by a factor of 2. With a permeability of 40, it's poorer by a factor of 8! The 6.25:1 parallel-type Balun in this article uses ferrite beads with a permeability of 125 and, therefore, should have losses similar to its series-type counterpart.

9:1 High-impedance Baluns. As was shown in the preceding section, the potential gradient along two of the transmission lines is V_1, where V_1 is the input voltage. The third transmission line, with a balanced load (or as an Unun), has no potential gradient and, consequently, no loss in its core. Because the loss with the series-type Balun mainly exists in one core, the loss with the 1:9 Balun should be a little less than twice as great. With ferrite cores of 250 permeability, the suggested losses are 0.2 dB at 7 MHz and 0.4 dB at 30 MHz. With cores of 125 permeability, the losses are about 0.14 and 0.28 dB, respectively. Again, by using cores with permeabilities of 40, the losses are practically negligible—approximately 0.1 dB within its passband.

As in the case of the 1:6.25 Baluns above, similar trade-offs occur in the low-frequency response. That is, if 125 permeability cores are used, the low-frequency response is poorer by a factor of 2; with 40 permeability cores, it's poorer by a factor of 8. The major difference here is that the low-frequency performance of the

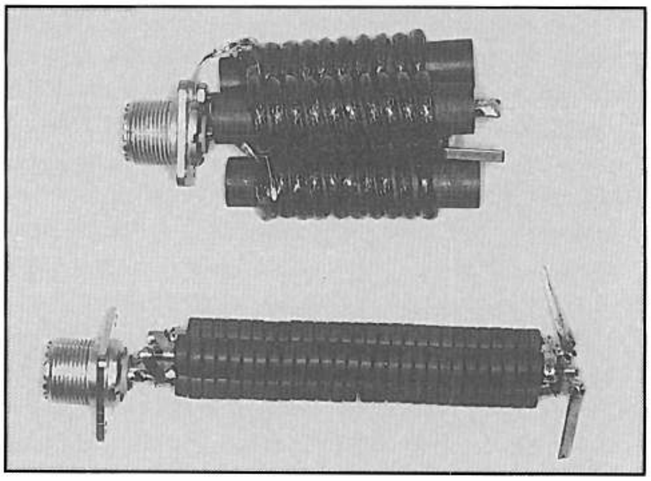

Photo 11-F. *Two versions of the coaxial-cable 9:1 Guanella Balun (or Unun) designed to match 50-ohm cable to 5.6 ohms*

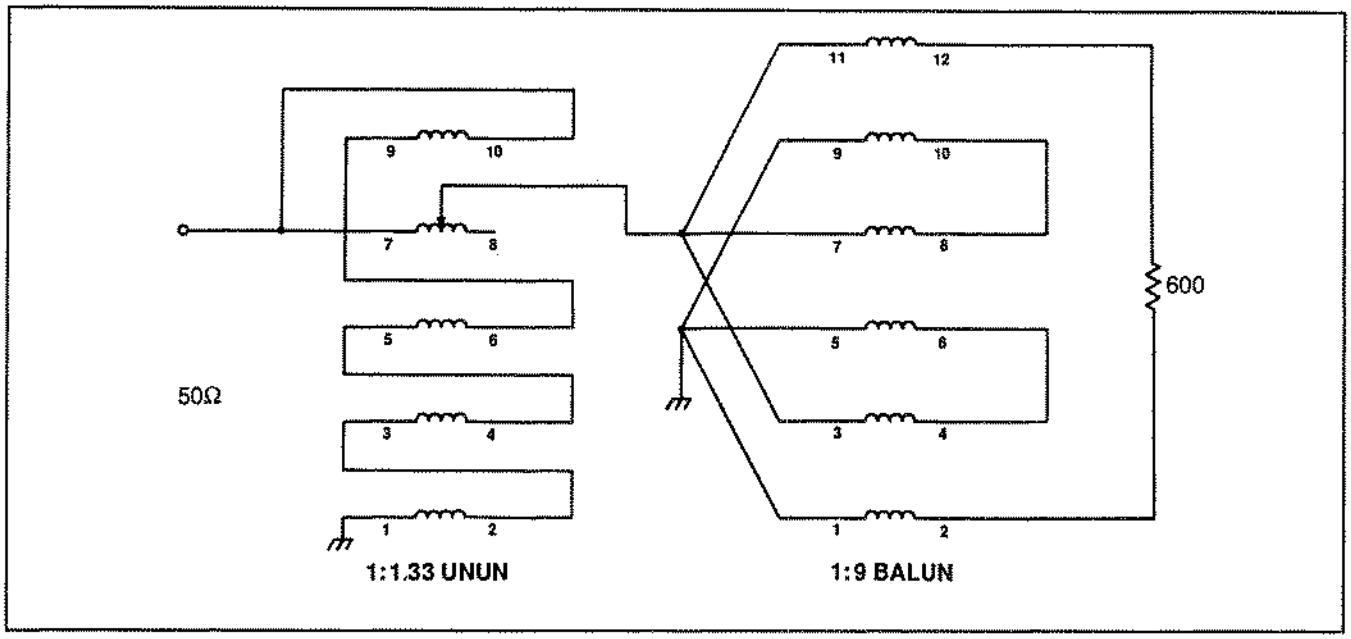

Figure 11-7. *Schematic diagram of the series-type 12:1 Balun using a 1:1.33 Unun in series with a 1:9 Guanella Balun.*

9:1 Balun, as seen by its low-frequency model, isn't as good as the 4:1 Balun that controls the low-frequency response of the series-type 1:6.25 Balun. Additionally, it should be pointed out that all of the suggested losses for the transformers in this chapter are for matched conditions—that is with VSWRs of 1:1. If the VSWR is 2:1 due to a load twice as large as the objective, the input voltage to the Balun increases by about 40 percent. Therefore, the losses should increase by close to the same percentage.

In closing this section, I would like to report my findings on two recent articles in amateur radio journals that also described 9:1 Baluns matching 50-ohm cable to 450 ohms. One[5] advocated using three 150-ohm coaxial cables threaded through high-permeability ferrite beads—a 9:1 Guanella Balun. Because of the low voltage-breakdown capability of the coaxial cable and the high loss found by accurate measurements on Ununs using these high-permeability ferrites, the design was suspect. I built a copy of the design and found it to be, as expected, unable to handle any appreciable power. The second article[28] advocated using 14 trifilar turns of "magnet wire" on a 2-inch OD powdered-iron core (permeability of 10). This Balun was also constructed and tested. Again, as was expected, when matching 50-ohm cable to a floating load of 450 ohms, the 9:1 Balun barely reached a true 9:1 ratio at 7 MHz. Above 7 MHz, the ratio became greater than 9:1 and also introduced a reactive component. Below 7 MHz, there was insufficient choking reactance to prevent flux in the core. My three objections to this design are: 1) a trifilar design has a poor high-frequency response because it sums a direct voltage with a delayed voltage that traverses a single transmission line and a delayed voltage that traverses two transmission lines, 2) the characteristic impedances of the transmission lines are only 50 ohms (the objective is 150 ohms), and 3) the low-frequency response is poor because of the low-permeability powdered-iron core. I do not recommend either of the designs in these two articles.

Sec 11.5 12:1 Baluns

Over the years, the broadband, 12:1 Balun has been of special interest to users of rhombic and V antennas. With the aid of this Balun, certain advantages over multi-element arrays can be fully exploited. Rhombics and Vs are easier to construct, both electrically and mechanically, and there are no particularly critical dimensions or adjustments. Furthermore, they give satisfactory gain and directivity over a 2-to-1 frequency range. These antennas have also been found to be more effective in reception. Because their designs can present input impedances of 600 ohms, and very long lengths of highly efficient 600-ohm open-wire line can be used between the shack and the antenna, an efficient and broadband 12:1 Balun is a natural for

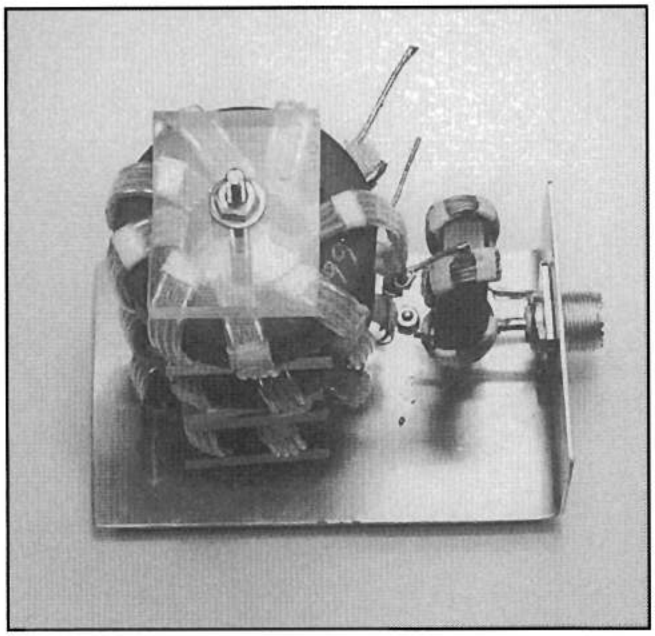

Photo 11-G. *Top view of the high-power 12:1 Balun.*

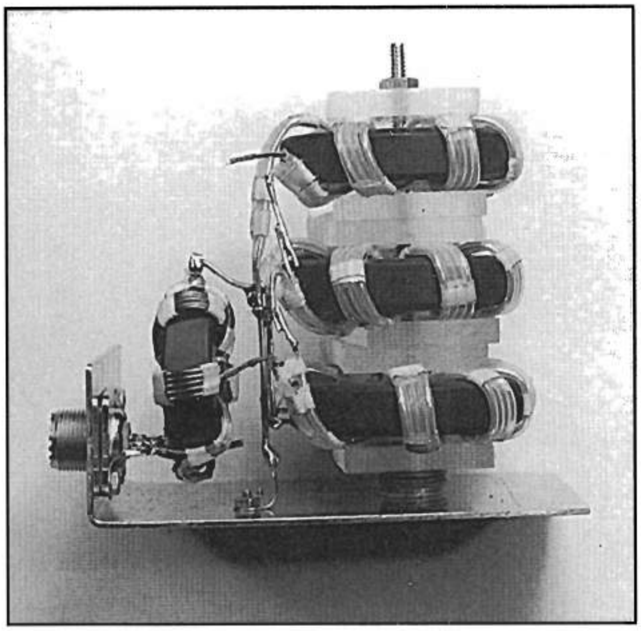

Photo 11-H. *Side view of the high-power 12:1 Balun.*

this application. However, to my knowledge, satisfactory Baluns have not been available for this use.

I will present two versions of a series-type Balun designed to match 50-ohm cable to a balanced load of 600 ohms. One is a high-power unit designed to handle the full legal limit of amateur radio power over a bandwidth of 7 to 30 MHz. The other is a medium-power unit capable of handling approximately one-half the legal limit of amateur radio power from 3.5 to 30 MHz. Both Baluns use a 1:1.33 Unun in series with a 1:9 Guanella Balun.

As you will see, these Baluns are not especially easy to design and construct. The major difficulties lie in trying to obtain sufficient choking reactances in the coiled windings to meet the low-frequency requirements, and large enough characteristic impedances of the windings to meet the high-frequency requirements. Because a coiled winding with a characteristic impedance of 200 ohms (the objective) is practically impossible to obtain with any reasonable wire size and number of turns, I used the compensating technique first described in my book.[2] Because the characteristic impedances of the 9:1 Guanella Balun are somewhat less than 200 ohms, a compensating effect (and hence higher frequency response) can be obtained by having a higher (than the normal objective) characteristic impedance of the windings in the 1:1.33 Unun. Earlier work (also described in my book) presented a 12:1 Balun using a 1:3 Unun in series with a 1:4 Ruthroff Balun. The Baluns presented in this section using a 1:1.33 Unun in series with a Guanella 1:9 Balun, are much improved designs.

Sec 11.5.1 A High-power 12:1 Balun

Figure 11-7 shows the schematic diagram of the series-type 12:1 Balun used in both the high- and medium-power versions. **Photo 11-G** shows a top view of the high-power Balun. **Photo 11-H** shows a side view.

The 1:1.33 Unun has 5 quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 7-8 is No. 14 H Thermaleze wire and the other four are No. 16 H Thermaleze wire. Winding 7-8 is also tapped at 3 turns from terminal 7.

The 1:9 Guanella Balun has 8 bifilar turns of tinned No. 16 wire on each of the three toroids. Each wire is covered with Teflon tubing and further separated by two Teflon tubings. The characteristic impedance of the windings is about 190 ohms (the objective is 200 ohms). The ferrite toroids have an OD of 2.4 inches and a permeability of 250. The spacing between the toroids is 1/2 inch.

In matching 50-ohm cable to a balanced load of 600 ohms, the response is literally flat (within a percent or two) from 7 to 30 MHz. Within this bandwidth, it is capable of handling the full legal limit of amateur

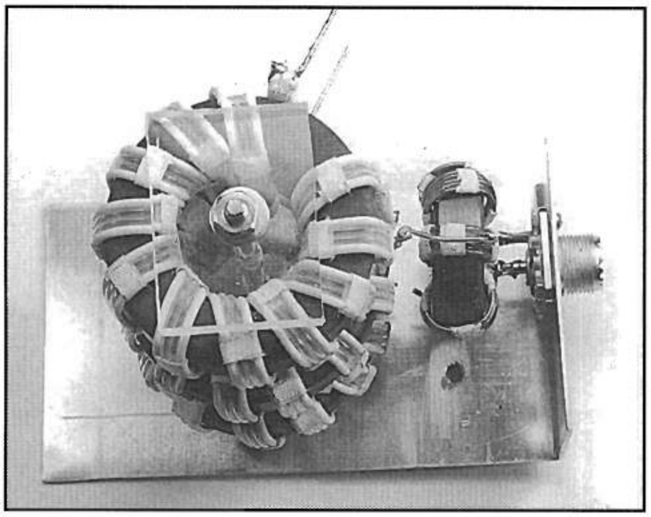

Photo 11-I. *Top view of the medium-power 12:1 Balun.*

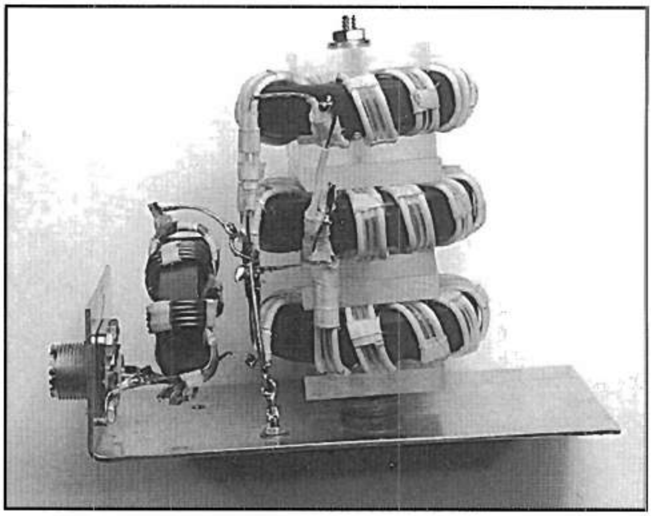

Photo 11-J. *Side view of the medium-power 12:1 Balun.*

radio power. In a matched condition, the expected insertion loss is about 0.25 dB.

Sec 11.5.2 A Medium-power 12:1 Balun

Photo 11-I shows the top view of the medium-power 12:1 series-type Balun. **Photo 11-J** shows the side view. This Balun also has the same 1:1.33 Unun as described earlier.

The 1:9 Guanella Balun has 11 bifilar turns of No. 18 hook-up wire on each toroid. The wires are further separated by two No. 18 Teflon tubings. The characteristic impedance of the windings is about 170 ohms. The toroids have an OD of 2.4 inches and a permeability of 250. The spacing between the toroids is also 1/2 inch.

In matching 50-ohm cable to a balanced load of 600 ohms, the response varies less than 5 percent from 3.5 to 30 MHz. Within this bandwidth, the Balun can handle about one-half the legal limit of amateur radio power. As with the high-power version, the expected insertion loss is also 0.25 dB.

Sec 11.6 Concluding Remarks on 12:1 Baluns

Many of the concluding remarks from the discussion on 6:1 and 9:1 Baluns (**Sec 11.4**) also apply to 12:1 Baluns; therefore, I won't repeat them here. But the following four remarks are specific to 12:1 Baluns and warrant mentioning:

First, high-impedance transmission line transformers like the 12:1 Balun are particularly sensitive to metallic enclosures. If a minibox were to be used, I would suggest the one shown in **Photo 11-E**, which is 6 inches long by 5 inches wide by 4 inches high. Smaller metallic enclosures would reduce the characteristic impedances of the windings and affect the high-frequency response. Even the spacing between cores had to be increased from 1/4 inch (for a 50:450-ohm Balun) to 1/2 inch. The subchassis shown in the photographs were used because they provided the necessary electrical and mechanical support.

Second, the 12:1 Baluns described in this section also make excellent Ununs, albeit with some compromise in the low-frequency response. I would suggest using the high-power unit only between 14 and 30 MHz and the medium-power unit only between 7 and 30 MHz.

Third, for the readers interested in VHF operation, I would suggest the parallel-type approach described in the earlier section on 6:1 Baluns. In this case, a 9:1 Guanella Balun is connected in series-parallel with a 1:4 Guanella Balun. This would produce a broadband ratio of 12.25:1. By using 170-ohm twin-lead (about 10 inches long) threaded through ferrite beads with a permeability of 125, it appears that it is possible to match 50-ohm cable to a balanced load of 612.5 ohms throughout the VHF band.

Fourth, by using torids with a permeability of 125 in the 1:9 Guanella Baluns (of the 12:1 Baluns), the insertion loss would be reduced by around one half (0.12 dB), with a trade-off in low-frequency response. The high-power unit would now cover about 10 to 30 MHz, and the medium-power Balun would cover about 7 to 30 MHz.

The 4:1 Unun

Sec 12.1 Introduction

From an analysis standpoint, the 4:1 Unun can be said to have received the most attention in the literature. It began with Ruthroff's introduction and complete analysis of this device in his classic paper published in 1959.[9] Ruthroff's paper then became the industry standard for this class of devices known as *transmission line transformers*. These are devices that transmit the energy from the input to the output by an efficient transmission line mode, and not by flux linkages (as in conventional transformers).

However 15 years earlier, Guanella had introduced, in his classic 1944 paper,[3] the first broadband Baluns by combining coiled transmission lines in a series-parallel arrangement, yielding ratios of $1{:}n^2$ where n = 1, 2, 3,... and so on. It has also been shown that Guanella's technique also lends itself to Ununs as well.[2] In fact, in this chapter, you will see that his technique of summing voltages of equal delays promises to yield high-power designs capable of operating on the VHF and UHF bands.

The 4:1 Unun also exemplifies (more than any other transformer) the many choices that can be made in its design. These include: 1) Ruthroff's or Guanella's designs, 2) wire or coaxial cable transmission lines, 3) coiled or beaded lines, 4) rods or toroids, 5) low-power or high-power designs, 6) HF, VHF, or UHF designs, and 7) the trade-offs in efficiency for low-frequency response or for high VSWR. The 4:1 Unun is the most prevalent of all the Ununs. It finds extensive use in solid-state circuits and in many antenna applications involving the matching of ground-fed antennas—where impedances of 12 to 13 ohms must be matched to 50-ohm coaxial cable. This chapter provides information on many 4:1 Unun designs.

Sec 12.2 The Ruthroff 4:1 Unun

Figure 12-1 illustrates two versions of Ruthroff's approach to obtaining a 4:1 unbalanced-to-unbalanced transformer (Unun). As can be seen, one uses a coiled wire transmission line, while the other uses a coiled coaxial cable. Depending upon the frequency, beaded transmission lines may also be used.

Ruthroff's design uses a single transmission line connected in, what I call, the bootstrap configuration. That is, terminal 2 is connected to terminal 3, lifting the transmission line (at the high-impedance side) by the voltage V_1. If the reactance of the coiled winding or beaded line is much greater than R_G, then only flux-canceling transmission line currents are allowed to flow. It is also apparent that the output voltage is the sum of a direct voltage, V_1, and a delayed voltage,

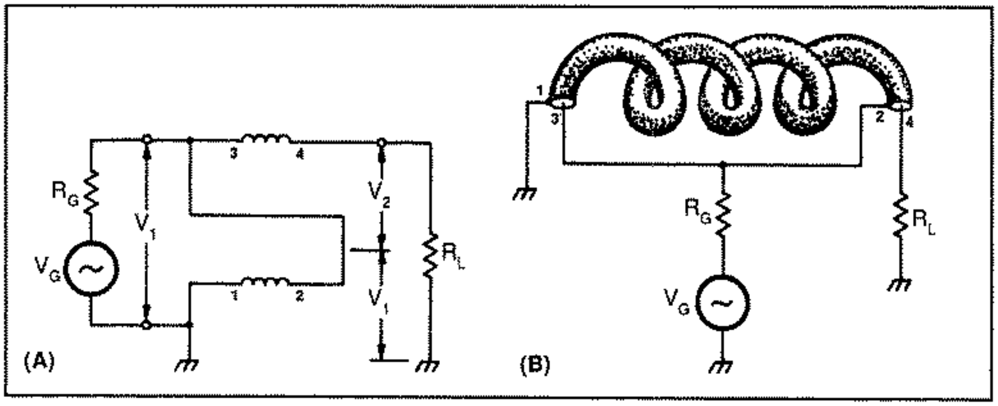

Figure 12-1. *The Ruthroff 4:1 Unun (R_L=$4R_G$): (A) coiled bifilar winding; (B) coiled coaxial cable.*

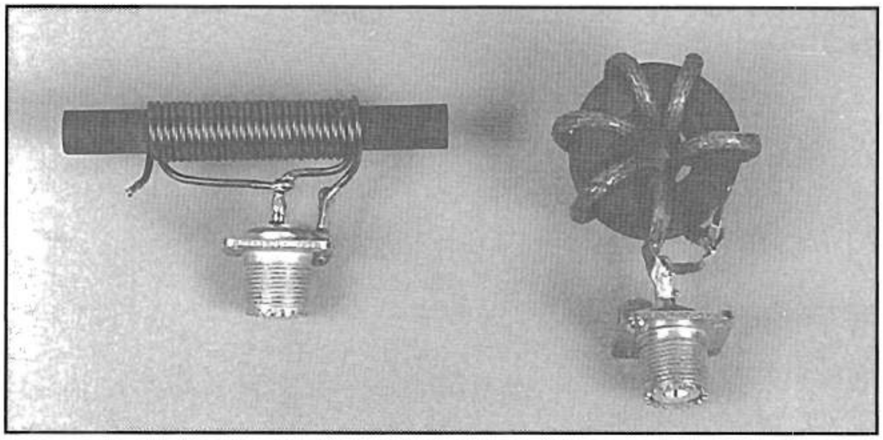

Photo 12-A. *Two versions of the Ruthroff 4:1 (50:12.5-ohm) Unun: coiled wire rod (on the left); coiled coaxial cable toroid (on the right).*

V_2, which traverses a single transmission line. This delay in V_2 eventually limits the high-frequency response. For example, if the electrical length of the line is a 1/2 wave, the output is zero. Ruthroff also found that the optimum value of the characteristic impedance of the transmission line (for maximum high-frequency response) is $R_L/2$.

Therefore, the electrical length and characteristic impedance of the transmission line play major roles in Ruthroff's design. Because his work was mainly concerned with small-signal applications, Ruthroff was able to obtain broad bands of a few tens of kilohertz to over a thousand megahertz. This was possible because he used a few turns (5 to 10) of fine wire (Nos. 37 and 38) on high-permeability toroids as small as 0.08 inches in OD. As a result, the phase-delay with these very short transmission lines was very small. However, large-signal (power) applications present an entirely different picture. For operation in the HF band (including 160 meters), transmission lines vary between one to three feet in length (depending upon impedance level). Consequently, phase-delay can play a major role, as will be seen in the following examples.

Sec. 12.2.1 50:12.5-ohm Ununs

Photo 12-A shows two examples of efficient and broadband 4:1 Ununs matching 50 to 12.5 ohms. The rod version (on the left) has 14 bifilar turns of No. 14 H Thermaleze wire on a low-permeability (125) ferrite rod 0.375 inches in diameter and 3.5 inches long. The connections are shown in **Figure 12-1A**. The cable connector is on the low-impedance side. The response is flat from 1.5 to 30 MHz. In a matched condition, this Unun can easily handle the full legal limit of amateur radio power. Because a tightly wound rod Unun yields a characteristic impedance very close to 25 ohms (the optimum value), this is quite likely the easiest one to construct that covers the above bandwidth.

The toroidal version (on the right in **Photo 12-A**) has 6 turns of homemade, low-impedance coaxial cable on a 1.5-inch OD ferrite toroid with a permeability of 250. The connections are shown in **Figure**

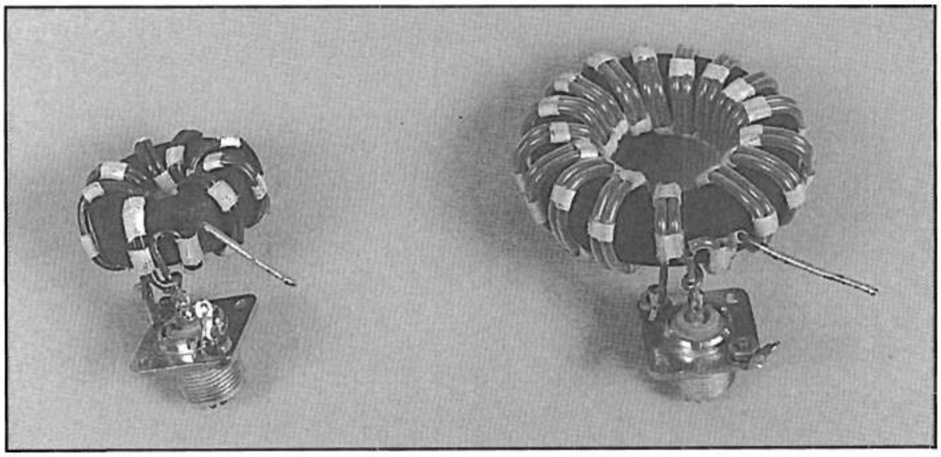

Photo 12-B. *Two higher-impedance Ruthroff 4:1 Ununs: 100:25-ohm (on the left); 200:50-ohm (on the right).*

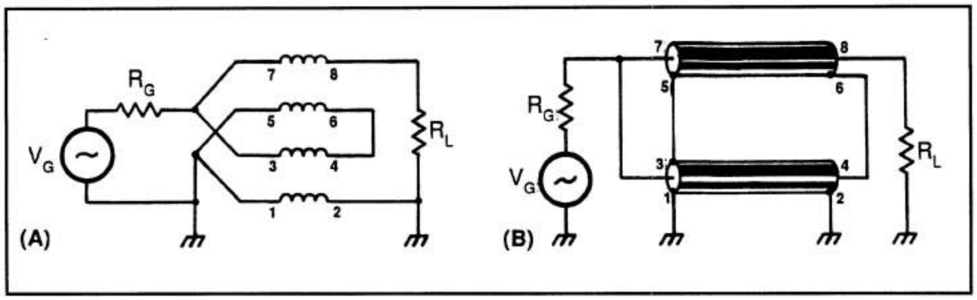

Figure 12-2. *The Guanella 4:1 Unun ($R_L=4R_G$): (A) coiled bifilar windings; (B) coiled or beaded coaxial cables.*

12-1B. The cable connector is on the low-impedance side. The inner conductor is No. 14 H Thermaleze wire and is covered with Teflon tubing. The outer braid is from a small coaxial cable (or from 1/8-inch tubular braid) tightly wrapped with Scotch No. 92 tape in order to obtain the desired characteristic impedance. In matching 50 to 12.5 ohms, the response is flat from 1.5 to 50 MHz. Because the current is evenly distributed on the inner conductor, this small Unun has an exceptionally high power capability—at least 5 kW of continuous power and 10 kW of peak power (in a matched condition).

Sec 12.2.2 100:25-ohm Unun

In some combiner applications, an Unun matching 100 to 25 ohms is required. The smaller toroidal version, pictured on the left in **Photo 12-B**, shows a Ruthroff design that can satisfy many of these requirements. It has 8 bifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid with a permeability of 250. One wire is also covered with a single layer of Scotch No. 92 tape, providing a characteristic impedance close to the desired value of 50 ohms. In matching 100 to 25 ohms, the response is essentially flat from 1.5 to 30 MHz. This Unun can easily handle the full legal limit of amateur radio power.

Sec 12.2.3 200:50-ohm Unun

When dealing with this type of Balun, the Ruthroff approach cannot yield the broadband response of the lower-impedance designs shown above. Because more turns are required in order to obtain the necessary choking reactance, and a 100-ohm characteristic impedance that requires more spacing between the wires is used, the cores must be considerably larger. This results in longer transmission lines. Consequently, the high-frequency response is now limited by the greater phase delay of this high-impedance Unun.

The larger transformer, shown on the right in **Photo 12-B** is my optimized version of a Ruthroff 200:50-ohm Unun. It has 16 bifilar turns of No. 14 H Thermaleze wire on a low-permeability (250) 2.4-inch OD ferrite toroid. Each wire is covered with Teflon tubing, resulting in a characteristic impedance of 97 ohms. Because of the long transmission line (36 inches), the impedance transformation ratio (in matching 200 ohms to 50 ohms) varies from 4 to 4.44 from 1.5 to 30 MHz. A conservative power rating (under a matched condition) is 2 kW of continuous power and 4 kW of peak power. Because this higher-impedance Unun has a larger voltage drop along the length of its windings, its loss (a dielectric-type[2]) is a little greater than the lower-impedance Ununs described earlier. In a matched condition, the efficiency is about 97 percent, while the others experience efficiencies of 98 to 99 percent.

Sec 12.3 The Guanella 4:1 Unun

Even though Guanella's investigation[3] was directed toward developing a broadband Balun to match the balanced output of a 100-watt, push-pull, vacuum-tube amplifier to the unbalanced load of a coaxial

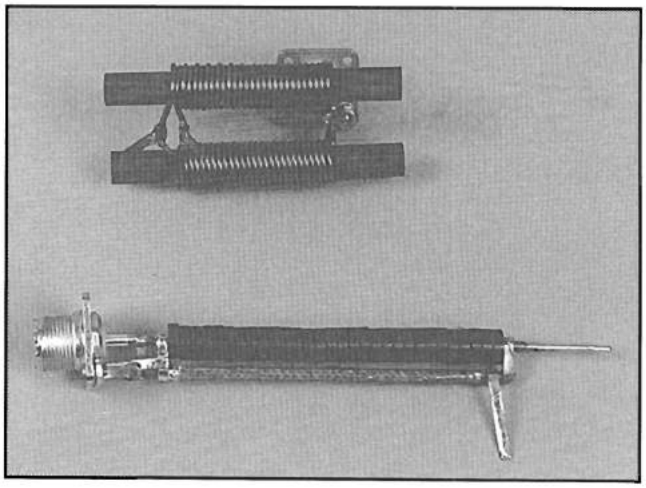

Photo 12-C. *Two Guanella 4:1 (50:12.5-ohm) Ununs: rod version (on the top), 1.5 to 50 MHz; beaded version (on the bottom), 10 MHz to over 100 MHz.*

cable, his technique of connecting transmission lines in a parallel-series arrangement has only recently been recognized as the design for the widest possible bandwidth in an unbalanced-to-unbalanced application.[2] Some have labeled his approach the "equal-delay network".[26] The major difference in Guanella's approach (from Ruthroff's) is that by summing the equal-delay voltages of coiled (or beaded) transmission lines, he minimizes the dependence of the high-frequency response on the lengths of the transmission lines. As was mentioned before, Ruthroff's method of summing a direct voltage with a delayed voltage that traversed a single transmission line has a limited application, especially with high-power, high-impedance Ununs (like 200:50 and 300:75 ohms).

Furthermore, Guanella's approach is also important in designing high- and low-impedance Baluns and Ununs with impedance transformation ratios other than 4:1. Connecting three transmission lines in parallel-series results in a 9:1 ratio, four in a 16:1. Also by connecting a fractional-ratio Unun in series with his Baluns, or by using various combinations of parallel-series transmission lines,[26,27] Ununs and Baluns are now available with a continuum of ratios from 1.36:1 to 16:1. Moreover, these ratios now make it possible to match 50-ohm cable to impedances as low as 3.125 ohms and as high as 800 ohms. A major factor in the success of these designs rests in the understanding of the low-frequency models of these various transformers.[2] This section looks at the 4:1 Unun using Guanella's approach. As in the Ruthroff case, the optimum value of the characteristic impedances of the transmission lines for a Guanella 4:1 transformer is also $R_L/2$.

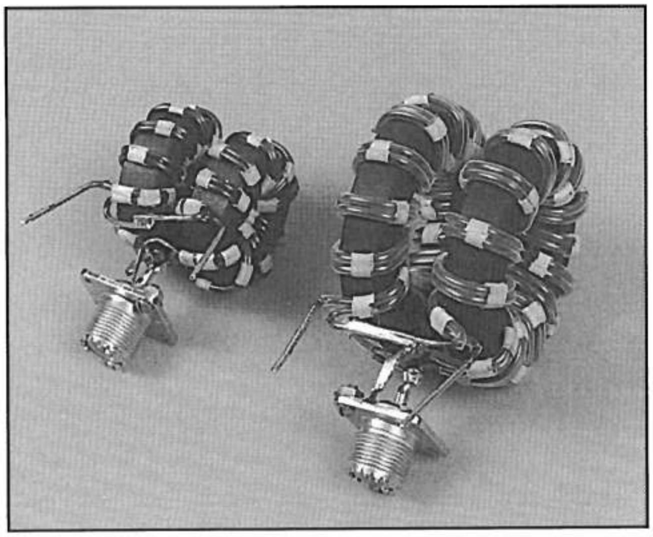

Photo 12-D. *Two higher-impedance Guanella 4:1 Ununs: 100:25-ohm (on the left); 200:50-ohm (on the right).*

Sec 12.3.1 50:12.5-ohm Ununs

Figure 12-2 shows the schematic diagrams of the coiled-wire and coaxial cable (coiled or beaded) versions of 4:1 Ununs using Guanella's technique of connecting transmission lines in parallel-series arrangements. As can be seen in **Figure 12-2**, the lower transmission lines are grounded at both ends and, therefore, have no potential drop along their lengths. Thus, the coiling or beading has no effect. The core only acts as a mechanical support and the beads can be removed. In essence, the bottom transmission line plays the important role of a delay line. In addition, the low-frequency response of this form of Unun is solely determined by the reactance of the top coiled or beaded transmission line.

The top Unun in **Photo 12-C** shows a rod version of Guanella's 4:1 Unun. There are 13.5 bifilar turns of No. 14 H Thermaleze wire on low-permeability (125) ferrite rods 0.375 inches in diameter and 3.5 inches long. For ease of connection, one winding is clockwise and the other is counterclockwise. The cable connector is on the high-impedance side. In matching 50 to 12.5 ohms, the response is flat from 1.5 to over 50 MHz! This Unun, in a matched condition, is capable of handling the full legal limit of amateur radio power. Furthermore, with the 50-ohm generator on the right (in **Figure 12-2A**) and a 12.5-ohm balanced load on the left (perhaps a Yagi beam), this transformer makes an excellent step-down Balun.

The bottom transformer in **Photo 12-C** shows a beaded-coax version of a 50:12.5-ohm step-down Unun designed for 2-meter operation. It has 3.5 inches of beaded coax on the top transmission line (**Figure 12-2B**) and no beads on the bottom transmission line. (Actually, the bottom rod in **Figure 12-2A** can also be removed with no change in performance.) The beads are low-permeability (125) ferrite. The inner conductor of the coaxial cable is No. 12 H Thermaleze wire with about 3.5 layers of Scotch No. 92 tape (two 0.5-inch tapes wound edgewise like a window shade), providing a characteristic impedance close to the optimum value. The outer braid is from a small coaxial cable (or from 1/8-inch tubular braid). This homemade coax is further wrapped tightly with Scotch No. 92 tape in order to preserve its low characteristic impedance. The cable connector is on the low-impedance side. The response of this Unun is essentially flat from 10 to 100 MHz (the limit of my bridge). It can also (easily) handle the full legal limit of amateur radio power.

Sec 12.3.2 100:25-ohm Unun

The Unun on the left in **Photo 12-D** is a Guanella version that matches 100 to 25 ohms. There are 8 bifilar turns of No. 14 H Thermaleze wire on each 1.5-inch OD low-permeability (250) toroid. One toroid is wound clockwise and the other is wound counterclockwise. One of the wires (on each toroid) is covered with one layer of Scotch No. 92 tape. The cable connector is on the low-impedance side. The response is flat from 1.5 MHz to well over 30 MHz. This Unun can also handle the full legal limit of amateur radio power.

It is interesting to note that when used as a Balun (the ground removed from terminal 2), and placed in series (on the left side) with a 1.78:1 Unun (see **Chapter 13**), this compound arrangement provides an excellent Balun for matching 50-ohm coaxial cable directly to quad antennas having impedances of 100 to 110 ohms.

Sec 12.3.3 200:50-ohm Unun

The transformer on the right in **Photo 12-D** is an excellent Unun (or Balun with terminal 2 removed from ground) for matching 50 to 200 ohms. It has 14 bifilar turns of No. 14 H Thermaleze wire on each low-permeability (250) toroid with a 2.4-inch OD. Each wire is covered with Teflon tubing, providing a characteristic impedance of 98 ohms (which is quite good because the optimum value is 100). Again, for ease of connection, one winding is clockwise and the other is counterclockwise. When operating as an Unun or a Balun and matching 50 to 200 ohms, the response is essentially flat from 1.5 to 30 MHz. A conservative power rating (in a matched condition) is 5 kW of continuous power and 10 kW of peak power. This transformer has been reported to handle peak pulses of 10,000 volts!

Summary

Since its introduction by Ruthroff in 1959,[9] the 4:1 Unun has been the most popular transmission line transformer matching unbalanced impedances to unbalanced impedances. As I mentioned at the beginning of this chapter, there are many choices to consider when designing these broadband and efficient transformers. One of the most important choices involves whether to use the Ruthroff or Guanella approach. In fact, the Guanella design should probably be designated a *Balun/Unun*. Recently, it has become the design of choice in the higher frequency bands. From the designs presented in this chapter, I offer the following recommendations:

1. For Ununs in the HF band with impedance levels of 100:25 ohms and lower, the Ruthroff approach is recommended because of its simplicity.
2. For high impedance levels in the HF band (like 200:50 and 300:75 ohms), the Guanella approach is recommended.
3. For low-impedance operation on the VHF band, the beaded-coax Guanella approach is recommended.
4. For high-impedance operation on the VHF band, the coiled-wire Guanella approach appears to be the preferred choice, and should be investigated first. Obviously, the number of turns should be reduced from the examples shown in this chapter because the reactance of the winding is proportional to the frequency.
5. For high-power use on the HF band, the Ruthroff Unun with low-impedance coaxial cable on a toroid (on the right on **Photo 12-A**) is recommended. It is easy to construct and can very likely handle more than 5 kW of continuous power.
6. Also, at high-impedance levels, one might consider using lower permeability ferrites for higher efficiencies. Look at permeabilities of 125 and 40.

Chapter 13

1.33:1, 1.5:1, and 2:1 Ununs

Sec 13.1 Introduction

Little practical design information has been available on Ununs with impedance transformation ratios of less than 4:1 (these are called *fractional-ratio* Ununs). However, many important applications can be found for efficient and broadband Ununs with ratios like 1.33:1, 1.5:1, and 2:1. Some examples include the matching of 50-ohm coaxial cable to: a) vertical antennas, inverted Ls, and ground-fed slopers (all over good ground systems), b) 75-ohm hardline cable, c) a junction of two 50-ohm coaxial cables, d) shunt-fed towers performing as vertical antennas, and e) the output of a transceiver or class B linear amplifier when an unfavorable VSWR condition exists.

These three Ununs also play an important role in making other useful Baluns possible. Examples given in earlier chapters include: a) connecting a 1.5:1 Unun (50:75 ohms) in series with a 1:1 Balun (75:75 ohms) results in a broadband 1.5:1 Balun (50:75 ohms); b) connecting a 2:1 Unun (50:100 ohms) in series with a 1:1 Balun (100:100 ohms) results in a broadband 2:1 Balun (50:100 ohms); c) connecting a 1.5:1 Unun (50:75 ohms) in series with a 4:1 Balun (75:300 ohms) results in a broadband 6:1 Balun (50:300 ohms), and d) connecting a 1.33:1 Unun (50:66.7 ohms) in series with a 9:1 Balun (66.7:600 ohms) results in a broadband 12:1 Balun (50:600 ohms).

It has been shown[2] that a continuum of ratios can now be obtained with Ununs matching 50-ohm cable to impedances as low as 3.125 ohms and as high as 800 ohms. In addition, by using higher-order windings (trifilar, quadfilar, etc.), Ununs can be constructed with two broadband ratios like 1.5:1 and 3:1, or 2:1 and 4:1. Furthermore, by tapping some of the windings of these higher-order Ununs, multimatch transformers can be constructed with many broadband ratios. As a result of this class of fractional-ratio Ununs, a continuum of Ununs and Baluns is now available to match 50 ohms unbalanced to unbalanced or balanced impedances as low as 3.125 ohms and as high as 800 ohms.

My first attempt to obtain ratios less than 4:1 was made by tapping one of the wires in a Ruthroff 4:1 bifilar Unun. My experiment met with only moderate success.[2] An adequate low-frequency response with a 1.33:1 ratio was difficult to obtain. Also, the 2:1 ratio had considerably greater loss than higher or lower ratios. Recently, I found that higher-order windings (trifilar, quadrifilar, etc.), some with taps, provide much wider bandwidths and higher efficiencies. This

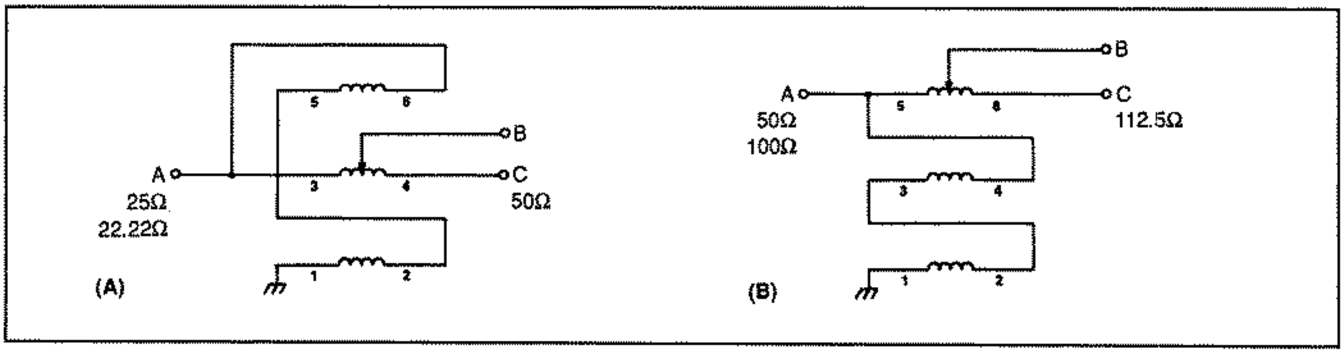

Figure 13-1. *Schematic diagrams: (A) matching 50 to 25 ohms (B-A) and 50 to 22.22 ohms (C-A); (B) matching 100 to 50 ohms (B-A) and 112.5 to 50 ohms (C-A).*

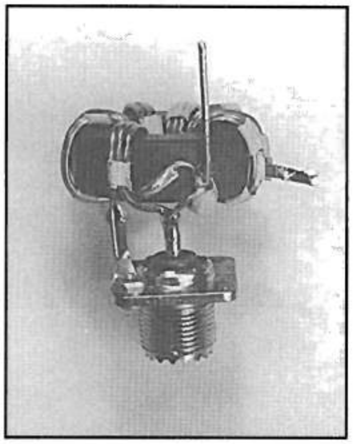

Photo 13-A. *Bottom view of the 2:1 Unun designed to match 50 ohms to 25 ohms or 22.2 ohms (Figure 13-1A). The connector is on the low-impedance side.*

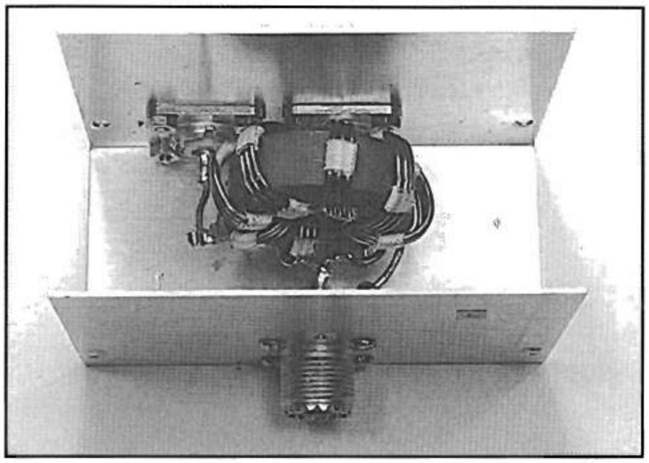

Photo 13-B. *The 2:1 Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox.*

chapter describes fractional-ratio Ununs using these higher-order windings.

The next section discusses the practical aspects of the 2:1 Unun. This Unun is not only one of the more useful transformers, but it also serves as a good introduction to the trifilar and quadrifilar designs. What follows is an introduction to the most difficult fractional-ratio Unun—the quintufilar design, which results in very broadband 1.33:1 and 1.5:1 Ununs. A more complete discussion appears in **Part I** of this book.

This chapter closes with construction tips. As you will see, these Ununs can be difficult to construct.

Sec 13.2 2:1 Ununs

Figure 13-1A shows the schematic diagram of an Unun designed to match 50-ohm cable to an unbalanced load of 25 ohms (2:1 ratio with connections A-B) or 22.22 ohms (2.25:1 ratio with connections A-C). It has 6 trifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 3-4 is tapped at 5 turns from terminal 3. **Photo 13-A** is a photograph showing the various connections. The connector is on the low-impedance side. **Photo 13-B** shows the transformer mounted in a CU-3015A (4 inches long by 2 inches wide by 2.75 inches high) minibox. In matching 50 ohms to either 25 or 22.22 ohms, the transformation ratio is constant from 1 to 30 MHz.

Because the transmission lines are very short, this Unun does quite well as a step-up transformer. That is, when matching 50 ohms (on the left side) to 100 ohms (connections A-B) or 112.5 ohms (connections A-C) on the right side, the transformation ratio is constant from 1 to 15 MHz. Because of the extremely high efficiency of this transformer (98 to 99 percent under matched conditions), this small version can easily handle the full legal limit of amateur radio power.

Figure 13-1B shows the schematic diagram of an Unun designed to match 50-ohm cable to an unbalanced load of 100 ohms (2:1 ratio with connections A-B) or 112.5-ohms (2.25:1 ratio with connections A-C). It has 7 trifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. The top winding 5-6 is No. 14 Thermaleze wire and is tapped at 6 turns from terminal 5. The other two windings are No. 16 H Thermaleze wire. **Photo 13-C** shows the various connections. The connector is on the low-impedance side. In matching 50-ohm cable to 100 ohms (A-B) or 112.5 ohms (A-C), the transformation ratio is constant from 1 to 30 MHz.

Again, because the transmission lines are very short, this Unun does quite well as a step-down transformer. In matching 50-ohm cable (on the right side) to 25 ohms (A-B) or 22.22 ohms (A-C), the transformation ratios are constant from 1 to 15 MHz. As above, this transformer can easily handle the full legal limit of amateur radio power.

Although the quadrifilar Unun shown in the schematic diagram in **Figure 13-2** and in **Photo 13-D** has an impedance transformation ratio of 1.78:1, it should also satisfy many of the 2:1 requirements. This Unun, which is designed to match 50-ohm cable to an unbalanced load of 28 ohms, not only has a very broadband response (1 MHz to over 50 MHz), but also offers other possible wideband ratios that will be covered in succeeding chapters.

Specifically, the Unun has 5 quadrifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire and the other three are No. 16 H Thermaleze wire. Like the

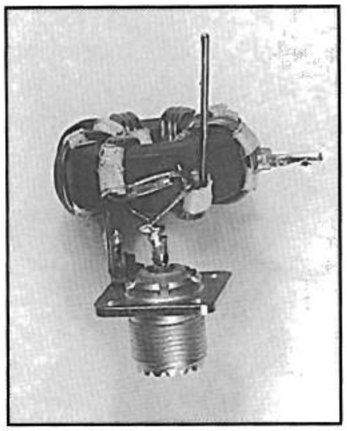

Photo 13-C. *Bottom view of the 2:1 Unun designed to match 50 ohms to 100 ohms or 112.5 ohms (Figure 13-1B). The connector is on the low-impedance side.*

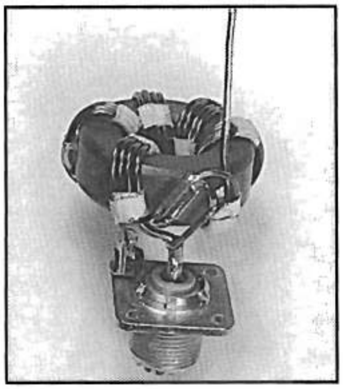

Photo 13-D. *Bottom view of the 1.78:1 Unun designed to match 50 to 28 ohms. The connector is on the low-impedance side.*

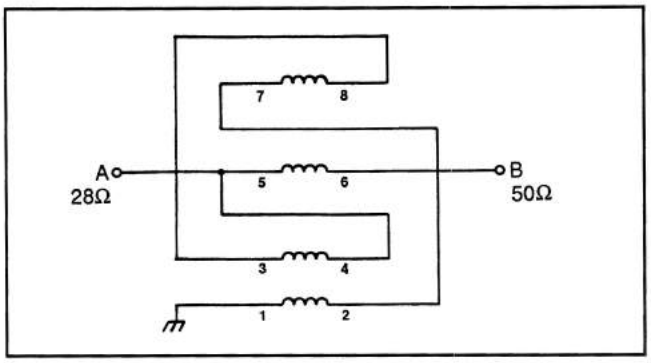

Figure 13-2. *Schematic diagram of the quadrifilar Unun designed to match 50 to 28 ohms (1.78:1 ratio).*

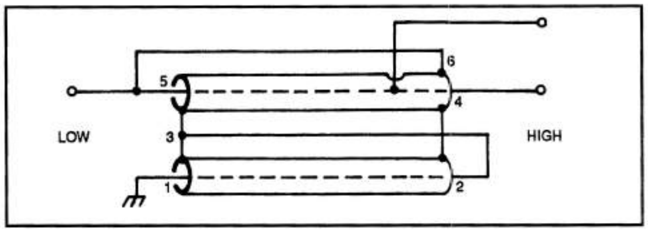

Figure 13-3. *Schematic diagram of a tapped-trifilar Unun that uses two sections of coaxial cable, yielding impedance ratios of 2:1 and 2.25:1.*

two 2:1 Ununs described above, this one also easily handles the full legal limit of amateur radio power.

As with most coiled Ununs that have little spacing between adjacent turns, current-crowding (between adjacent turns) can eventually limit the power-handling capability of these devices. It's possible to improve the ability for handling higher currents by using thicker wires, or by using coaxial cables where current-crowding is nonexistent. **Figure 13-3** shows the schematic diagram of a tapped-trifilar transformer that uses two sections of coaxial cable yielding impedance ratios of 2:1 and 2.25:1. **Photo 13-E** shows two trifilar toroidal transformers using low-impedance coaxial cables with their outer braids connected in parallel and acting as the third conductor. These transformers are conservatively rated at 5 kW of continuous power.

The smaller transformer in **Photo 13-E** has 7 trifilar turns of low-impedance coax on a 2-inch OD toroid with a permeability of 290. The No. 14 H Thermaleze wire inner conductors have four layers of Scotch No. 92 tape. The outer braids are made from small coaxial cables (or 1/8-inch tubular braid), and are also wrapped with Scotch No. 92 tape in order to preserve the low characteristic impedances. The inner conductor of the top coax in **Figure 13-3** is tapped at 6 turns from terminal 5. When matching 50 ohms to 22.22 ohms or 25 ohms, the impedance ratio is constant from 1 to over 50 MHz.

The larger transformer in **Photo 13-E** also has 7 trifilar turns, but on a 2.4-inch OD toroid with a permeability of 125. The inner conductors of No. 14 H Thermaleze wire now have a 15-mil wall of Teflon sleeving, yielding the low-impedance coaxial cable. The outer braids are the same. Because the ferrite permeability is lower and the lengths of the transmission lines are longer than those of the smaller unit, this transformer's bandwidth is not quite as good. When matching 50 to 22.22 ohms (this particular Unun doesn't have a tapped winding), the impedance ratio is constant from 1.7 to 30 MHz.

Although not shown, these very high-powered 2:1 Ununs can also be easily designed to match 50-ohm cable to unbalanced loads of 100 and 112.5 ohms. This is done by using small but high-powered coaxes like RG-303/U, RG-141/U, or RG-142/U.

Sec 13.3 1.5:1 Ununs

Figure 13-4 shows three basic forms of a quintufilar 1.56:1 Unun that should satisfy most of the 1.5:1 requirements. As can be seen, the only difference in

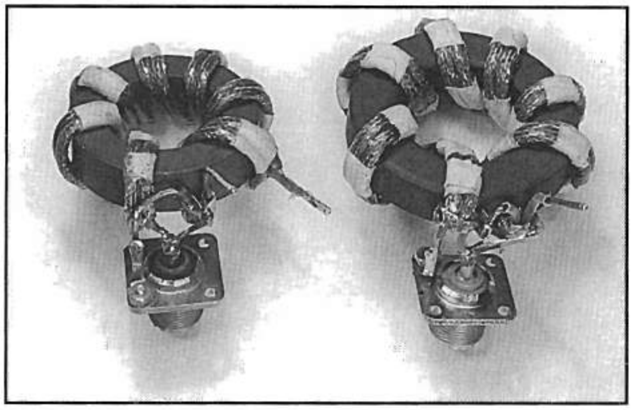

Photo 13-E. *Two toroidal transformers using coaxial cable and designed to match 50 to 22.22 ohms (2.25:1). The smaller transformer is also tapped, yielding a match of 50 to 25 ohms (2:1).*

the schematic diagrams is in the interleaving of the windings. This is done to optimize the performance of these Ununs at the various impedance levels. Schematic A is optimized for matching 50 to 75 ohms. Schematic B is optimized for matching 32 to 50 ohms. Schematic C, while optimized for matching 40 to 62 ohms, still yields quite broadband ratios at both 50:75 and 32:50-ohm levels. It should be a useful, general-purpose Unun.

Photo 13-F is a photograph of the bottom views (showing the connections) of the three different designs. They appear in the same order as the schematics of **Figure 13-4**; i.e., a) the Unun on the left is designed to match 50 to 75 ohms, b) the Unun in the center is design to match 32 to 50 ohms, and c) the Unun on the right is designed to work quite well at both impedance levels. The SO-239 connectors are all on the low impedance side of the Ununs.

All three transformers have four quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Their differences are:

1. **50:75 ohms** (on the left on **Figure 13-4** and **Photo 13-F**).

Winding 9-10 is No. 14 H Thermaleze wire. The other four windings are No. 16 H Thermaleze wire. When matching 50 to 75 ohms (actually to 78 ohms), the transformation ratio is constant from 1 to over 30 MHz. In matching 50 ohms (on the right side in **Figure 13-3A**) to 32 ohms, it is still constant from 1 to 15 MHz.

2. **32:50 ohms** (in the center in **Figure 13-4** and **Photo 13-F**).

Winding 5-6 is No. 14 H Thermaleze wire. The other four windings are No. 16 H Thermaleze wire. When matching 32 to 50 ohms, the transformation ratio is constant from 1 to over 30 MHz. In matching 75 ohms (on the right side in **Figure 13-3B**) to 50 ohms, it is still constant from 1 to 15 MHz.

3. **50:75 ohms; 32:50 ohms** (on the right in **Figure 13-4** and **Photo 13-F**).

Winding 7-8 is No. 14 H Thermaleze wire. The other four windings are No. 16 H Thermaleze wire. In matching 32 to 50 ohms, the transformation ratio is constant from 1 to 30 MHz. In matching 75 ohms (on the right side in **Figure 13-3C**) to 50 ohms, it is still constant to 21 MHz. This is quite a good general-purpose design.

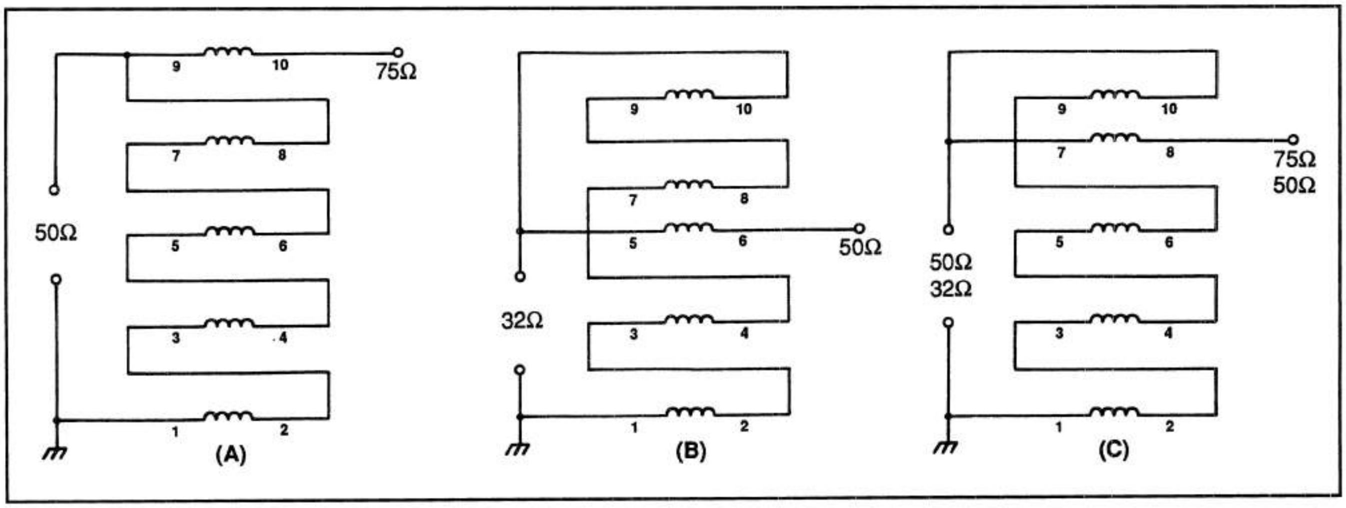

Figure 13-4. *Three basic forms of a quintufilar 1.56:1 Unun: (A) optimized to match 50 to 75 ohms, (B) optimized to match 32 to 50 ohms, and (C) optimized to match 40 to 62 ohms, resulting in a good general-purpose design.*

Even though a small toroid (with only a 1.5-inch OD) is used, these Ununs are still very sturdy transformers. Because their efficiencies are so high (98 to 99 percent), they can easily handle the full legal limit of amateur radio power.[2] Furthermore, the windings carrying the majority of the current (80 percent) are all No. 14 wire. Only when well-designed Ununs are subjected to very high VSWRs will excessive heating occur. Ununs (and Baluns) should never be exposed to these severe conditions.

Sec 13.4 A 1.33:1 Unun

The circuit shown in **Figure 13-5** evolved after many attempts were made at obtaining a broadband match of 50 to 66.7 ohms (1.33:1). **Photo 13-G** shows the bottom view of an actual design. The SO-239 connector is on the low impedance side. **Photo 13-H** shows the Unun mounted in a CU-3015A minibox.

Specifically, this Unun has five quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire and is tapped at three turns from terminal 5 (**Figure 13-4**). It is also covered with one layer of Scotch No. 92 polyimide tape, optimizing the performance at the 50:66.7-ohm level. The other four windings are No. 16 H Thermaleze wire.

In matching 50 to 66.7 ohms (A-B), the transformation ratio is practically constant from 1 to 30 MHz. The ratio only decreases by 3 percent across the band. In matching 50 to 32 ohms (C-A), the transformation ratio is constant from 1 to 30 MHz. In matching 75 to 50 ohms (C-A), the ratio is constant from 1 to 15 MHz. In matching 50 to 37.6 ohms (B-A), the ratio is constant from 1 to 15 MHz. As you can see, the Unun has some useful broadband multimatches.

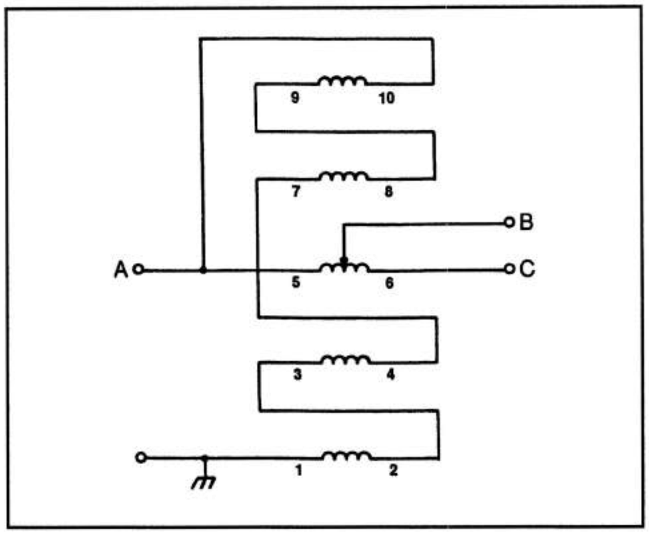

Figure 13-5. *Schematic diagram of a quintufilar Unun specifically designed to yield a broadband 1.33:1 ratio (66.7:50 ohms; connection B-A). Connection C-A also yields a broadband 1.56:1 ratio (50:32 ohms).*

As in the cases of the other three Ununs, this tapped-Unun also easily handles the full legal limit of amateur radio power. Like the 2:1 Unun, for higher power capabilities, thicker wires or a three-coax quintufilar design can be used.[2]

Sec 13.5 Construction Tips

Most of my Unun designs use the bootstrap connection that sums direct voltages (on the high-impedance side) with a delayed voltage, which traverses a single transmission line.[2] Therefore, in order to achieve the very wideband responses, small toroids (which allow the shortest transmission lines) are used. The small 1.5-inch OD toroids offer this advantage. Furthermore,

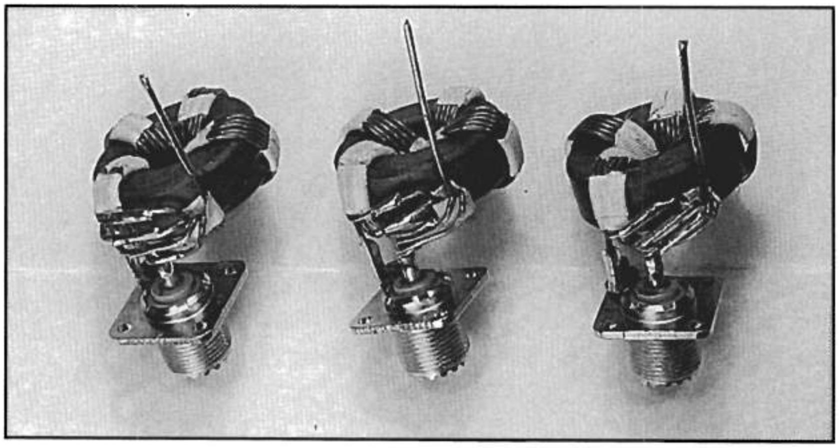

Photo 13-F. *Photograph of the three different Unun designs shown in Figure 13-4: A) on the left, 50:75 ohms, B) in the center 32:50 ohms, and C) on the right, a general purpose design matching both impedance levels quite well.*

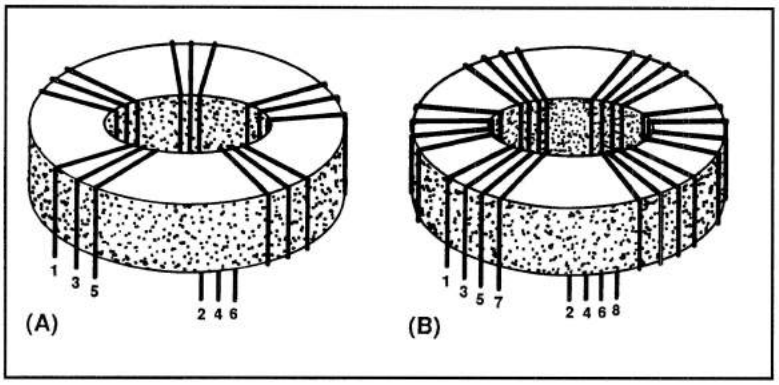

Figure 13-6. *Pictorials of higher-order windings: (A) trifilar and (B) quadrifilar.*

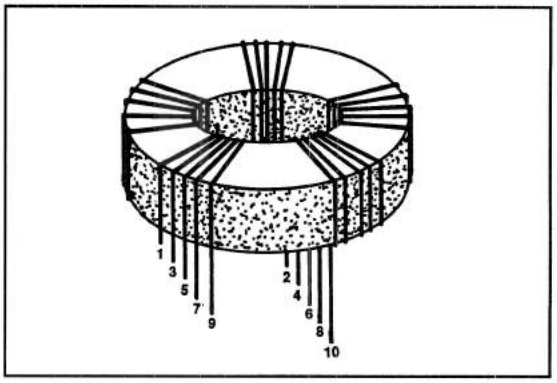

Figure 13-7. *Pictorial of a quintufilar winding.*

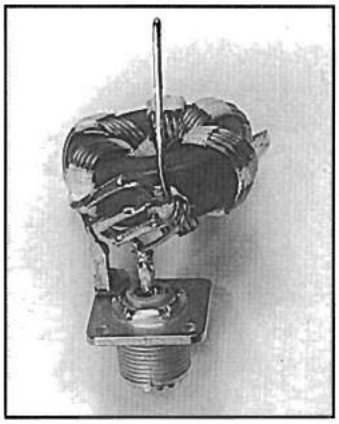

Photo 13-G. *Photograph of the 1.33:1 Unun (66.7:50 ohms). This transformer also has a broadband 1.56:1 ratio (50:32 ohms).*

quintufilar (and higher-order windings) Ununs are also eventually limited in their high-frequency responses by self-resonances. Shorter winding lengths keep these self-resonances well out of the HF band. For the Ununs in this chapter, they occur between 45 and 65 MHz, with **Figure 13-4A** having the higher value.

There is also a mechanical advantage in using the smaller toroids. You'll find that the popular CU-3015A minibox makes an excellent enclosure for the 1.5-inch OD toroid. Furthermore, because ferrite is a ceramic and, therefore, unaffected by moisture, no special precautions need be taken for out-of-doors use. Potting the transformer in plastic is unnecessary. One must only keep the Unun out of a pool of water.

Because well-designed transformers have virtually no flux in the core, their power ratings are mainly determined by the ability of the transmission lines to handle the voltages and currents. Furthermore, it can be shown that the losses in these transformers are related to the voltage gradients along the transmission lines.[2] Thus, they are dielectric-type ferrite losses. This means that the efficiency can be severely degraded with very high VSWRs since higher voltage gradients occur under these conditions.

Several suggestions can be made regarding the construction of Ununs using these higher-order windings (trifilar, quadrifilar, etc.). They are:

1) Make a ribbon out of the wires and wind them all at the same time. This keeps the wires as close as possible, resulting in the maintenance of the optimum characteristic impedance of the transmission lines. I found that strips made with 1.25- by 0.375- to 0.55-inch glass tape (Scotch No. 27), clamped about every 1/2 inch, hold the wires in place very well. The starting lengths of the wires should be about 5 inches longer than one would calculate knowing the number of turns and the length around each turn.

2) Because work-hardening of the copper wire takes place in coiling it around a toroid, a pair of pliers and

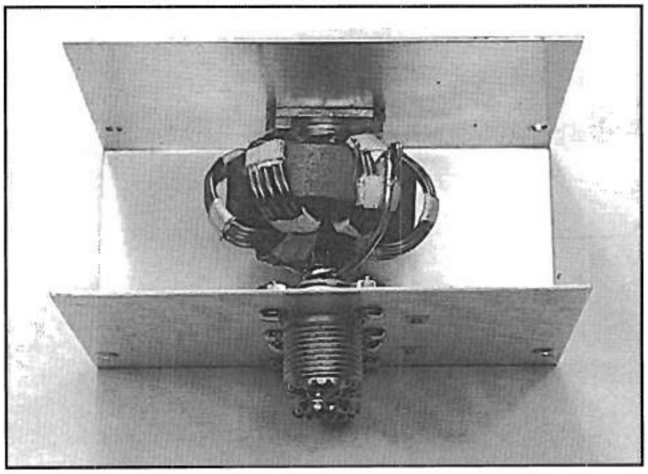

Photo 13-H. *The 1.33:1 Unun mounted in a CU-3015A minibox.*

a strong thumb (and arms) are indispensable tools. It takes considerable effort to wind these transformers. Also, because these designs have adequate margins at their low-frequency ends, some space between the windings and the toroids can be tolerated.

3) It is helpful to recognize the various patterns that appear at the ends of the windings. **Figure 13-6** shows a drawing of the trifilar and quadrifilar patterns and **Figure 13-7** shows the quintufilar pattern. Note that terminal 1 and terminal 6 or terminal 8 or 10 are the outside terminals of the patterns. Also, note that terminal 1 is always grounded in the schematic diagrams.

4) Tapping windings can be one of the more difficult tasks in constructing these transformers. Winding a tapped transformer is also more difficult. I found that the edge of a small, fine file does the best job in removing the insulation. About 1/8 to 1/4 inch is removed around the wire. It also helps to remove some of the copper. Then a flat 1/8-inch copper strip or No. 14 wire flattened on one end is soldered to the bare wire. The soldered connection is then rendered smooth using the edge of the file. Finally, two pieces of Scotch No. 92 polyimide tape are placed on the joint to insulate it from the neighboring turns. I also found that a tap placed one turn from the end of the winding is best made approximately 4 inches from the end of the wire (when the wire is straight).

Finally, a comment should be made about low-power Ununs. Practically all of the transformers in this chapter can be easily designed for low-power use. Designs capable of handing the full output of HF transceivers can be readily constructed. Cores with an OD of 1.25 inches are recommended. The same number of turns, but with one size smaller wire, is also recommended. Because smaller cores and thinner wires are used, these lower-power units are not only easier to construct, but they also have wider bandwidths due to the shorter lengths of the windings.

Chapter 14

Dual-Ratio Ununs

Sec 14.1 Introduction

Earlier chapters in this book on Ununs have basically covered single-ratio transformers. Broadband ratios of 1.33:1, 1.5:1, 2:1, and 4:1 were the design objectives. The 1.33:1 ratio was obtained by tapping a 1.5:1 (actually 1.56:1) quintufilar-wound Unun. The 2:1 ratio was obtained by tapping a 2.25:1 trifilar-wound Unun. Although these transformers can be considered to have two broadband ratios (1.33:1 and 1.56:1 or 2:1 and 2.25:1), their two ratios were not different enough for many practical applications. This is especially true of antennas where the input impedance varies with frequency.

My earlier work[2] and an article by Genaille[10] have shown that a host of ratios (less than 4:1) can be obtained by tapping the bifilar winding of a Ruthroff 4:1 Unun.[9] However, the bandwidths obtained using this technique are quite limited with each ratio and are highly dependent upon the impedance level. This is particularly true when a rod core is used because it requires more turns (resulting in longer transmission lines) in order to obtain the necessary choking reactance that isolates the input from the output. Furthermore, ratios around 2:1 exhibit more loss because autotransformer action also enters into the matching process.[2]

By using quadrifilar and quintufilar windings on small 1.5-inch OD cores, and connecting them in such a way that the characteristic impedances of the windings are near optimum, two very different and broadband ratios matching 50 ohms to lower impedances are obtained. Furthermore, because the transmission lines in these transformers are so very short (8 to 9 inches in length), these transformers do quite well in matching 50 ohms to higher impedances (as step-up transformers).

This chapter presents two Ununs which have two broadband ratios that differ by a factor of two! One has a 1.5:1 and a 3:1 ratio (actually 1.56:1 and 2.78:1). The other has a 2:1 (actually 1.78:1) and a 4:1 ratio. Also, this chapter introduces the novel tech-

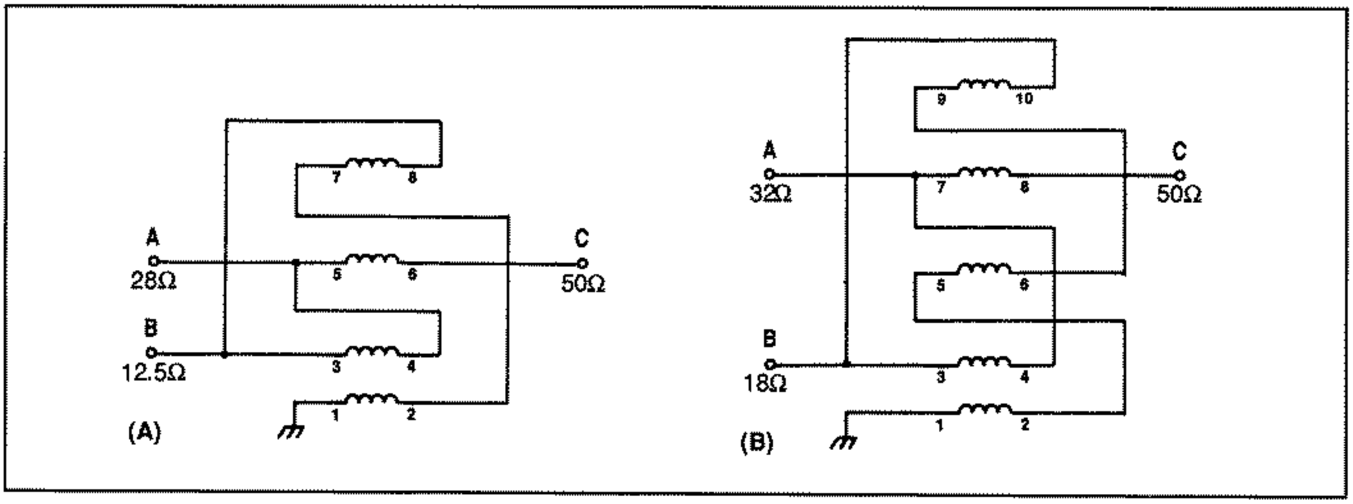

Figure 14-1. *Schematic diagrams of dual-ratio Ununs: (A) 1.78:1 connection C-A, 4:1 connection C-B; (B) 1.56:1 connection C-A, 2.78:1 connection C-B.*

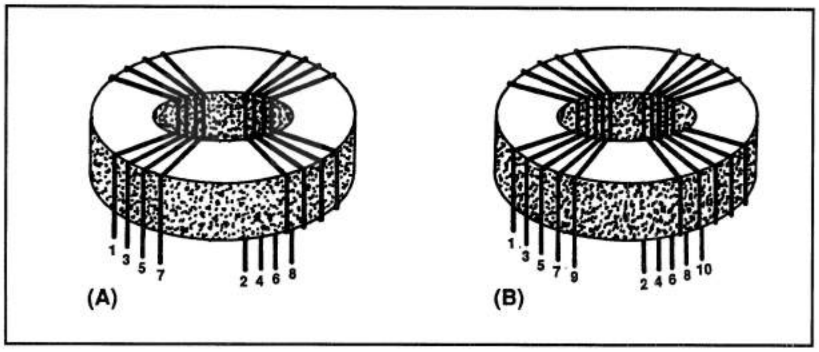

Figure 14-2. *Pictorials of higher-order windings: (A) quadrifilar, (B) quintufilar.*

nique of connecting these two transformers in parallel on their 50-ohm sides, resulting in four very broadband ratios.

Sec 14.2 2:1 and 4:1 Ratios

Figure 14-1A shows the schematic diagram of the quadrifilar Unun yielding ratios of 2:1 (actually 1.78:1) and 4:1. **Figure 14-2A** provides a pictorial of its windings. On the left in **Photo 14-A**, you see the bottom view of an Unun. The cable connector is on the 28-ohm (1.78:1) side. This Unun has four quadrifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD toroid with a permeability of 250.

In matching 50 to 28 ohms (connection C-A), the response is flat to within 1 percent from 1 to 30 MHz. From 1 to 50 MHz, it's flat to within 2 percent. In matching 50 to 12.5 ohms (connection C-B), the response is flat to within 3 percent from 1 to 30 MHz.

Because very short transmission lines are used, this transformer performs quite well as a step-up transformer. In matching 50 to 200 ohms (connection B-C), the response is flat to within 3 percent from 1.5 to 10 MHz. In matching 50 to 89 ohms (connection A-C), the response is flat to within 5 percent from 1.5 to 30 MHz!

Sec 14.2.1 Construction Tips

Start with about 14 inches of straightened wire. Form the wires into a ribbon with clamps of Scotch No. 27 glass tape every 1/2 inch. I found that strips 3/16 inch wide and about 1.5 inches long do a good job. The clamps should be long enough to go around the wires twice. After winding, connect terminals 2 and 7. Then connect terminals 3 and 8. Finally, connect terminals 4 and 5. Because work-hardening takes place quickly, you will find that a pair of pliers and a strong thumb (and arms) are necessary tools. You will also find that winding these exceptionally performing transformers is not easy. As in all endeavors, practice really pays off.

Sec 14.3 1.5:1 and 3:1 Ratios

Figure 14-1B shows the schematic diagram of the quintufilar Unun yielding ratios of 1.5:1 and 3:1

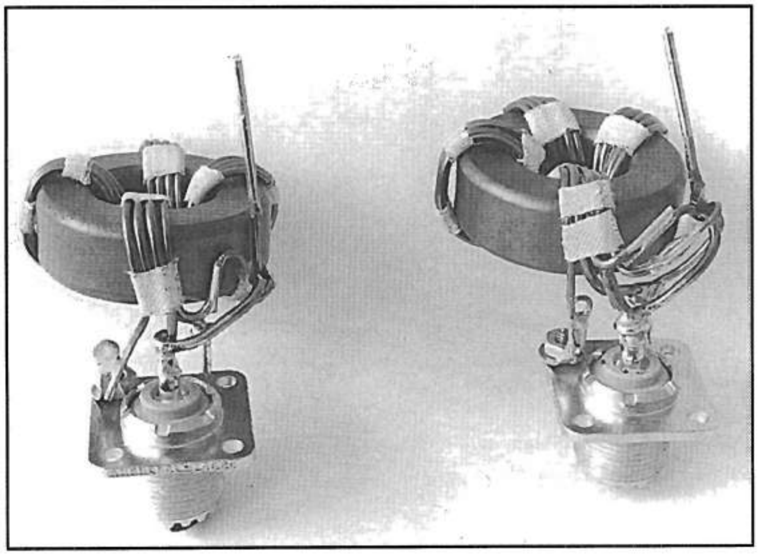

Photo 14-A. *On the left, a quadrifilar Unun with ratios of 1.78:1 and 4:1; on the right, a quintufilar Unun with ratios of 1.56:1 and 2.78:1.*

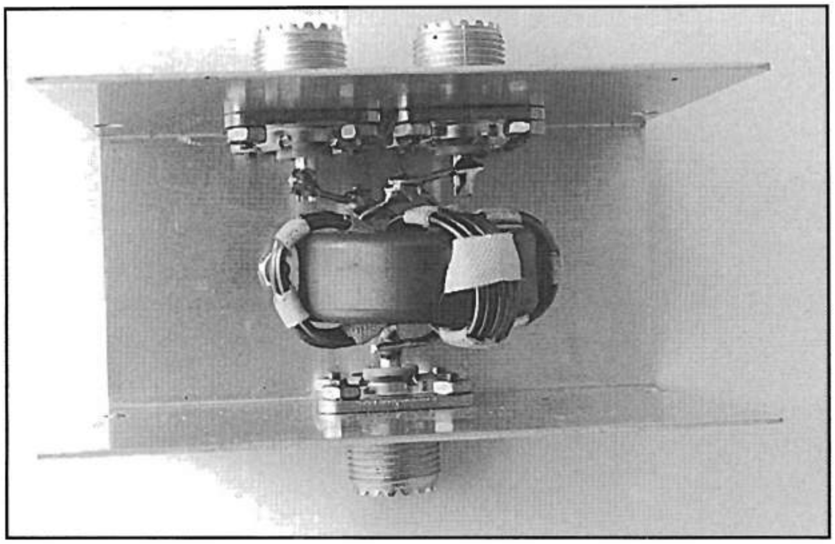

Photo 14-B. *The dual-ratio quintufilar Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox.*

(actually 1.56:1 and 2.78:1). **Figure 14-2B** is a pictorial of its windings. On the right side in **Photo 14-A** you see a bottom view of an Unun. The cable connector is on the 32-ohm (1.56:1) side. This Unun has four quintufilar turns on a 1.5-inch OD toroid with a permeability of 250. Windings 3-4 and 7-8 are No. 14 H Thermaleze wire. The other three are No. 16 H Thermaleze wire. Winding 7-8 also has two layers of Scotch No. 92 polyimide tape, which optimizes the 1.56:1 ratio.

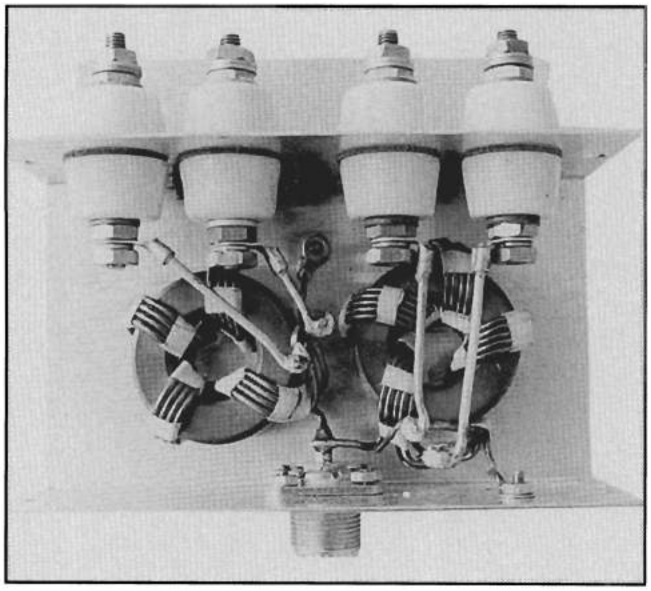

Photo 14-C. *The two dual-ratio Ununs connected in parallel on their 50-ohm sides providing four broadband ratios close to 1.5:1, 2:1, 3:1, and 4:1. The quadrifilar unit is on the left and the quintufilar unit is on the right. The enclosure is a 5 inch long by 3 inch wide by 2 inch high minibox.*

In matching 50 to 32 ohms (connection C-A), the response is essentially flat (less than 1 percent variation) from 1 MHz to over 40 MHz. Without the two layers of Scotch No. 92 tape, the response varies by 4 percent from 1 to 30 MHz. When used as a step-up transformer matching 50 to 78 ohms (connection A-C), and with the two layers of Scotch No. 92 tape on winding 7-8, the response is flat to within 5 percent from 1 to 15 MHz. Without the extra insulation on winding 7-8, the response is flat to within 5 percent from 1 to 7.5 MHz.

In matching 50 to 18 ohms (connection C-B), the variation in response is less than 3 percent from 1 to 40 MHz. The response is the same whether winding 7-8 is covered with the extra insulation or not. As a step-up transformer matching 50 to 139 ohms (connection B-C), the response is flat to within 3 percent from 1 to 10 MHz (with or without the extra insulation on winding 7-8). **Photo 14-B** shows this Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox. The two cable connectors on the low impedance side could be replaced with feedthrough insulators for antenna use.

Sec 14.3.1 Construction Tips

Prepare the ribbon as was described for the quadrifilar Unun. If you choose to use the two extra layers of Scotch No. 92 tape on winding 7-8, make sure this winding is on the outside position of the ribbon (refer to **Figure 14-2B**). I found the best order in which to connect the wires is as follows: first, connect terminal 2 to 5; second, connect terminal 6 to 9; third, connect terminal 3 to 10; and, finally, connect terminal 4 to 7.

As was mentioned before, because a small toroid is used in order to achieve the best response (due to shorter transmission lines), this transformer also requires considerable strength and patience in the winding process.

Sec 14.4 Parallel Transformers

One of the most pleasant surprises I received with these efficient and broadband transformers was to find that they can be connected in parallel on their 50-ohm sides and still possess the same performance levels. Because the loading effect of one transformer on the other is minimal (like a short length of transmission line), the transformer that is properly terminated takes the power, while the other one is transparent. This lets you obtain four very wideband ratios with the two dual-ratio Ununs described in this chapter. Obviously, this technique eliminates one transmission line. Furthermore, Baluns can also be connected this way for feeding, with a single coaxial cable, beams and dipoles with different resonant impedances.

Photo 14-C shows two transformers connected in parallel on their 50-ohm sides using the two dual-ratio Ununs described in this chapter. As was mentioned, it now yields four wideband ratios very close to 1.5:1, 2:1, 3:1, and 4:1. I have used this matching network to feed a host of ground-fed antennas (over a good ground system). In one case, I had a 10-, 15-, and 20-meter trap vertical, slopers for 40 and 160 meters, a 12-meter vertical, and an inverted L for 80 meters all matched to a single coaxial cable at the same time. It was a simple matter of connecting each antenna to the output terminal that presented the best match (lowest VSWR). This technique is actually an extension of connecting dipoles for different bands, in parallel. The antenna that presents the correct impedance takes the power, and the others are essentially transparent. In many cases, I found that only one transformer with two broadband ratios performed adequately.

Chapter 15

Multimatch Ununs

Sec 15.1 Introduction

Broadband multimatch Ununs capable of high-power applications have been the goal of many designers over the years. Some have resorted to using conventional autotransformers with tapped windings to obtain the many impedance transformation ratios. However, these attempts met with little success because of the device's limited bandwidths and efficiencies. Others (including myself)[2] have tried tapping a bifilar Ruthroff Unun.[9] Although these designs yielded the high efficiencies of transmission line transformers, they had limited bandwidths. Furthermore, their best bandwidths (for the various ratios) occurred at odd impedance levels. In other words, they didn't meet the objective of broadband operation with one of the input or output ports at 50 ohms.

Chapter 14 presented two Ununs which had two broadband ratios that differed by a factor of two. One had a 1.5:1 and a 3:1 ratio (actually 1.56:1 and 2.78:1). The other had a 2:1 (actually 1.78:1) and a 4:1 ratio. This chapter describes two multimatch designs that are capable of many more broadband ratios. For the most part, both are capable of broadband operation from 1.7 to 30 MHz.

One Unun has the following five ratios (which are close to): 1.5:1, 2:1, 4:1, 6:1, and 9:1. Because the two lower ratios work well in either direction (that is stepping up or down from 50 ohms), this design can match 50-ohm cable to impedances as high as 100 ohms (actually 112.5 ohms) and as low as 5.6 ohms over the frequency range. As a result, it has seven usable applications. Furthermore, because this is a transmission line transformer that cancels out the flux in the core, losses (in a matched condition) of only 0.04 to 0.08 dB can be expected.

The novelty in this design lies in the use of a trifilar winding (with one winding tapped) on a very small ferrite toroid, resulting in the shortest possible lengths of transmission lines. The windings are also connected in such a manner as to optimize their characteristic impedances from an overall standpoint.

I have used the adjective *ultimate* to describe the second Unun design. Although it might be risky business, I assume that this design will meet one of the most common definitions for this adjective—namely, *beyond which it is impossible to go.* For many of us, the classic use of this adjective was made by Lew McCoy in describing his popular transmatch.[22] Although there have been some improvements to Lew's design, his use of this definite (and strong) adjective can be said to have withstood the test of time. I hope my use meets with similar success.

While the tapped-trifilar design provides five broadband ratios and seven practical applications, the *ultimate* design presented in this chapter goes well

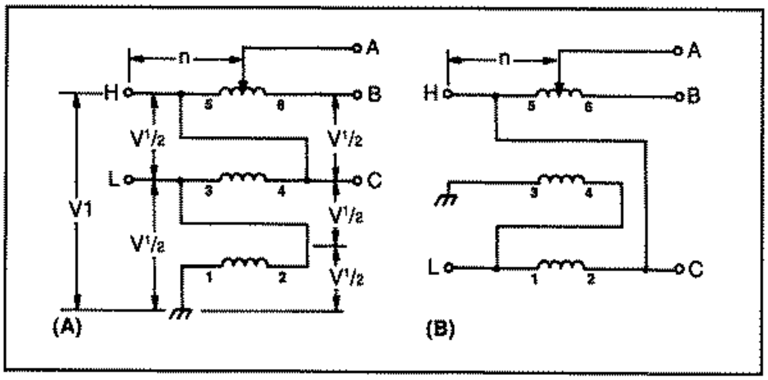

Figure 15-1. *Circuit diagrams for the 5-ratio Unun: (A) diagram for analysis; (B) transposed windings for best overall performance.*

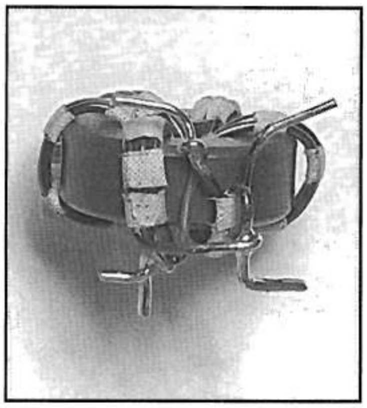

Photo 15-A. *Bottom view of the 5-ratio Unun of Figure 15-1B. The upper-left lead is terminal C. The upper-right lead is terminal B. The lower-left lead is terminal H. The lead pointing straight down is grounded (terminal 3). The lower-right lead is terminal L.*

beyond this number. It uses a tapped-quadrifilar design that yields the following 10 broadband ratios: 1.36:1, 1.56:1, 1.78:1, 2.25:1, 3.06:1, 4:1, 6.25:1, 9:1, 12.25:1, and 16:1. Because the four lower ratios also work quite well in either direction, this design offers *fourteen* applications in matching 50-ohm cable to impedances as high as 112.5 ohms and as low as 3.125 ohms. It also has the advantage of using a small low-loss toroidal core. Additionally, the windings are also interleaved in a pattern that optimizes their characteristic impedance.

However, this achievement comes at a price—*difficulty*. The 5-ratio Unun, which uses a trifilar winding, is considerably easier to wind. In addition, the quadrifilar 10-ratio Unun has two of its windings tapped, while the 5-ratio Unun has only one (see **Chapter 13** on tapping windings). If you have had little experience in winding Ununs or Baluns, attempt simplified versions of these two multimatch Ununs first. These versions eliminate the tapping of the windings. For the trifilar Unun, the remaining ratios would be: 2.25:1, 4:1, and 9:1. For the quadrifilar Unun, they would be: 1.78:1, 2.25:1, 4:1, 9:1, and 16:1.

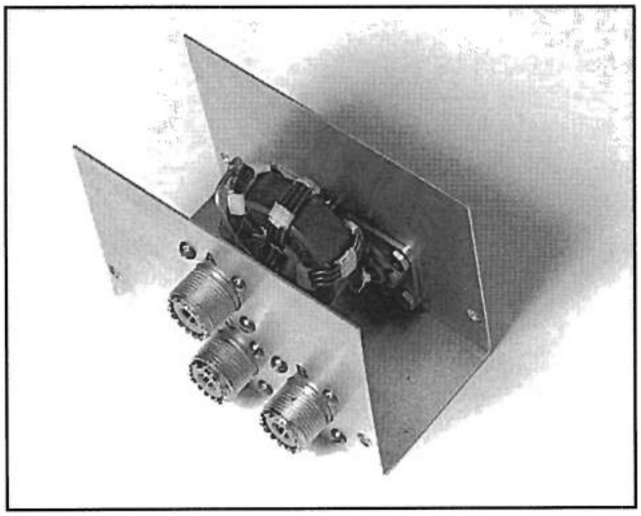

Photo 15-B. *The high-power unit mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox.*

For those interested in the design considerations of these broadband multimatch transformers, a brief review is presented in each section. These sections are followed by others describing high-power designs capable of handling the full legal limit of amateur radio power. Finally, the remaining sections present low-power designs capable of handling the output of any HF transceiver. Because transmission line transformers can be made so efficient in matching 50 to 100 ohms or less, their small sizes will surprise many readers. Therefore, the combination of using small ferrite toroids with the maximum allowable permeability (less than 300) for high efficiency,[2] and with sufficient turns to meet the low-frequency objective, results in the excellent performance exhibited by the designs in this chapter.

Sec 15.2 The 5-Ratio Unun

Let's first look at **Figure 15-1A** because it is the easiest form of the trifilar-wound Unun to explain. For example, if the input voltage to ground, V_1, is connected to terminal H, the output terminal B, has a voltage to ground of $3/2V_1$. This results in a transformation ratio, g, of $(3/2)^2$ or 2.25:1. This should satisfy most 2:1 requirements. If the output is at terminal A to ground, then the output voltage is:

$$\begin{aligned} V_o &= V_1 + V_1(n/2N) \\ &= V_1(1 + n/2N) \qquad \text{(Eq 15-1)} \end{aligned}$$

where:

N = the total number of turns on the winding

n = the number of turns from terminal 5.

The transformation ratio, g, then becomes:

$$\begin{aligned} g &= (V_o/V_1)^2 \\ &= (1 + n/2N)^2 \qquad \text{(Eq 15-2)} \end{aligned}$$

If the input voltage to ground, V_1, is connected to terminal L, then terminal C has twice the voltage of V_1—resulting in a 4:1 ratio. Terminal B has three times the voltage resulting in a 9:1 ratio. With terminal A, the output voltage is:

$$V_o = 2V_1 + V_1(n/N)$$
$$= V_1(2 + n/N) \quad \text{(Eq 15-3)}$$

The transformation ratio, g, then becomes:

$$g = (2 + n/N)^2 \quad \text{(Eq 15-4)}$$

Sec 15.2.1 A High-power 5-Ratio Unun

After several attempts at rearranging the windings of **Figure 15-1A** for the best overall performance (optimizing the effective characteristic impedances of the windings), **Figure 15-1B** evolved. **Photo 15-A** shows the bottom view of an Unun, using the circuit of **Figure 15-1B**, capable of handling the full legal limit of amateur radio power. **Photo 15-B** shows the unit mounted in a CU-3015A minibox. It has five trifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is tapped at two turns (n = 2) from terminal 5.

If the 9:1 ratio matching 50 to 5.6 ohms (connection B-L) is to be used at full power, then winding 3-4 should be No. 12 H Thermaleze wire. If not, then all windings can be No. 14 H Thermaleze wire.

A listing of the expected performance across the band from 1.7 MHz to 30 MHz, with the various ratios, is as follows:

9:1 (B-L); 50:5.6 ohms
Ratio is within 1 percent!
5.75:1 (A-L); 50:8.7 ohms
Ratio decreases by 5 percent.
4:1 (C-L); 50:12.5 ohms
Ratio increases by 15 percent (the greatest deviation of all the ratios).
2.25:1
a) (B-H); 50:22.22 ohms
Ratio decreases by 4 percent.
b) (H-B); 50:112.5 ohms
Ratio increases by 8 percent.
1.44:1
a) (A-H); 50:35 ohms
Ratio decreases by 10 percent.
b) (H-A); 50:72 ohms
Ratio increases by 2 percent.

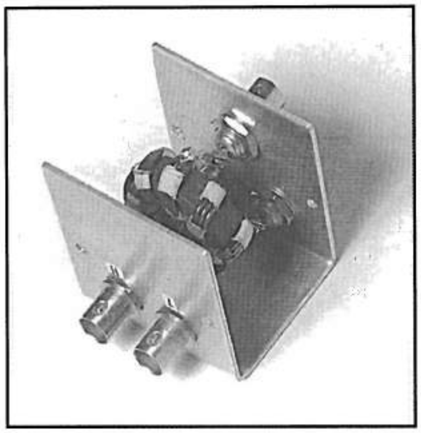

Photo 15-C. *The low-power unit mounted in a homemade 2 inch long by 1.5 inch wide by 2.25 inch high minibox.*

Several comments should be made regarding the expected results shown above. First of all, the greatest deviation from a flat response at any ratio occurs when matching 50 to 12.5 ohms (connection C-L; a 4:1 ratio). If an accurate insertion loss measurement was made at this ratio and impedance level, the result would show an insignificant difference across the band. Secondly, the major part of the deviations for all ratios occurs beyond 15 MHz (the effect of standing waves). Finally, the higher ratios should *never* be used to match 50 ohms to 450 ohms, 288 ohms, and 200 ohms, respectively. The characteristic impedances and choking reactances *do not* allow for broadband operation under these conditions.

Sec 15.2.2 A Low-power 5-Ratio Unun

Photo 15-C shows a low-power unit mounted in a homemade 2 inch long by 1.5 inch wide by 2.25 inch

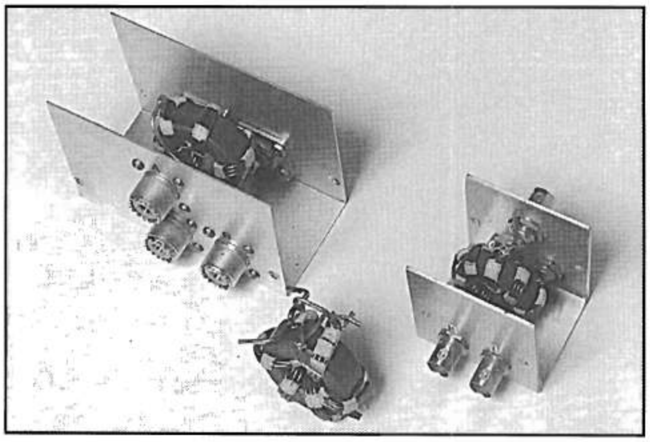

Photo 15-D. *The three 5-ratio Ununs together. From left to right, the high-power unit mounted and unmounted, the low-power unit.*

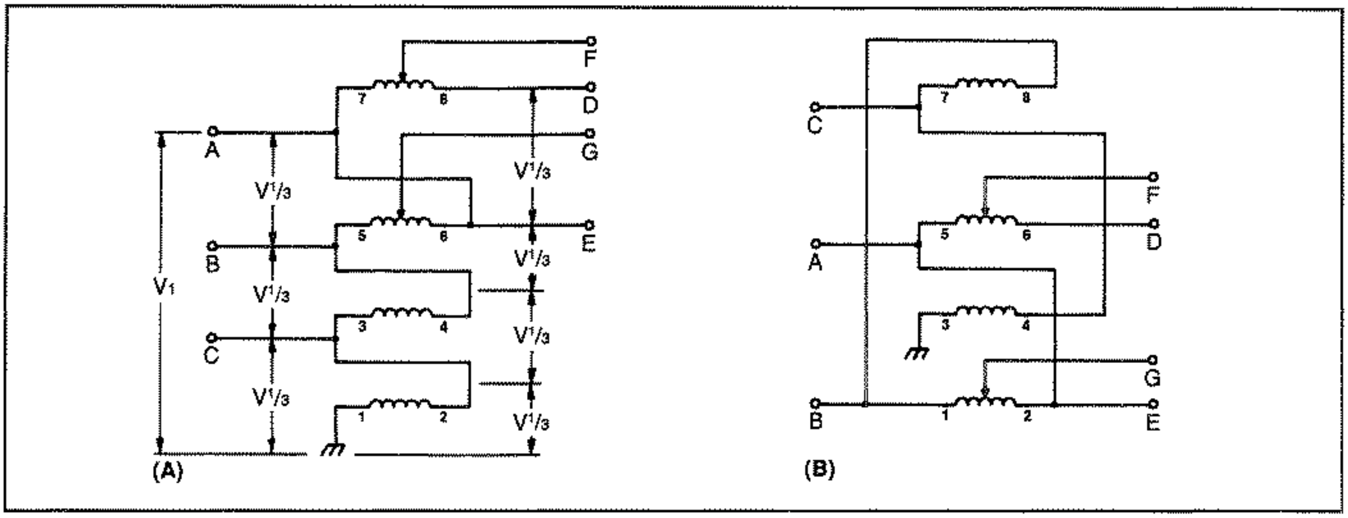

Figure 15-2. *Circuit diagrams for the 10-ratio Unun: (A) diagram for analysis; (B) transposed windings for best overall performance.*

high minibox. It has six trifilar turns of No. 16 H Thermaleze wire on a 1.25-inch OD ferrite toroid with a permeability of 250. The tap on winding 5-6 is located three turns from terminal 5, yielding ratios of 6.25:1 and 1.56:1 instead of the 5.75:1 and 1.44:1 ratios of the high-power unit. In actual use, these differences should be negligible.

Because this Unun has shorter transmission lines than its high-power counterpart, the deviations of the ratios across the band are even smaller. It is also interesting to note that, if No. 14 H Thermaleze wire was used in winding 3-4, this very small Unun could very well be rated at 500 watts of continuous power!

Photo 15-D shows all three 5-ratio Ununs together.

Sec 15.3 The 10-Ratio Unun— The Ultimate Multimatch

Figure 15-2A is presented here because it is the easiest form of a quadrifilar-wound Unun to explain. With the input voltage, V_1, connected to the various terminals on the left (the low-impedance side), and with very short transmission lines compared to the wavelength, we have the following transformation ratios:

1. V_1 connected to terminal A

a) At terminal D the output voltage V_o is $4/3V_1$. Therefore, the transformation ratio, g, with connection A-D is:

$$g = (4/3)^2 = 1{:}1.78 \qquad \text{(Eq 15-5)}$$

b) At terminal F the output voltage is:

$$V_o = V_1(1 + n/3N) \qquad \text{(Eq 15-6)}$$

where:
N = total number of turns
n = number of turns from terminal 7.

The transformation ratio with connection A-F then becomes:

$$g = (V_o/V_1)^2 = (1 + n/3N)^2 \qquad \text{(Eq 15-7)}$$

2. V_1 connected to terminal B

a) At terminal E the output voltage is $3/2V_1$. Thus, the transformation ratio with connection B-E is:

$$g = (3/2)^2 = 1{:}2.25 \qquad \text{(Eq 15-8)}$$

b) At terminal G the output voltage is:

$$V_o = V_1(1 + n/2N) \qquad \text{(Eq 15-9)}$$

where n = number of turns from terminal 5.

The transformation ratio with connection B-G becomes:

$$g = (V_o/V_1)^2 = (1 + n/2N)^2 \qquad \text{(Eq 15-10)}$$

c) At terminal D the output voltage is $2V_1$. The transformation ratio with connection B-D is:

$$g = (2)^2 = 1{:}4 \qquad \text{(Eq 15-11)}$$

d) At terminal F the output voltage is:

$$V_o = V_1(3/2 + n/2N) \qquad \text{(Eq 15-12)}$$

where n = number of turns from terminal 7.

The transformation ratio with connection B-F, then, is:

$$g = (V_o/V_1)^2 = (3/2 + n/2N) \qquad \text{(Eq 15-13)}$$

3. V_1 connected to terminal C

a) At terminal E, the output voltage is $3V_1$. The transformation ratio with connection C-E becomes:

$$g = (3)^2 = 1{:}9 \qquad \text{(Eq 15-14)}$$

b) At terminal G, the output voltage is:

$$V_o = V_1(2 + n/N) \qquad \text{(Eq 15-15)}$$

where n = number of turns from terminal 5.

The transformation ratio with connection C-G is:

$$g = (V_o/V_1)^2 = (2 + n/N)^2 \qquad \text{(Eq 15-16)}$$

c) At terminal D, the output voltage is $4V_1$. The transformation with connection C-D becomes:

$$g = (4)^2 = 1{:}16 \qquad \text{(Eq 15-17)}$$

d) At terminal F the output voltage is:

$$V_o = V_1(3 + n/N) \qquad \text{(Eq 15-18)}$$

where n = number of turns from terminal 7.

The transformation ratio with connection C-F is:

$$g = (V_o/V_1)^2 = (3 + n/N)^2 \qquad \text{(Eq 15-19)}$$

Sec 15.3.1 A High-power 10-Ratio Unun

Figure 15-2B evolved after several attempts at rearranging the windings of **Figure 15-2A** for best overall performance (optimizing the effective characteristic impedances of the windings). **Photo 15-E** shows the bottom view of an unmounted Unun using the circuit of **Figure 15-2B**. The top-left lead is terminal E. The top-right lead is terminal D. The bottom-left lead is terminal B. The center lead (connected to the SO-239 connector) is terminal A.

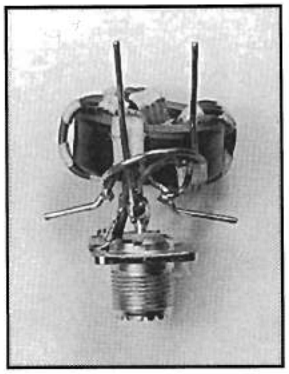

Photo 15-E. *Bottom view of the 10-ratio Unun. The connector is on terminal A.*

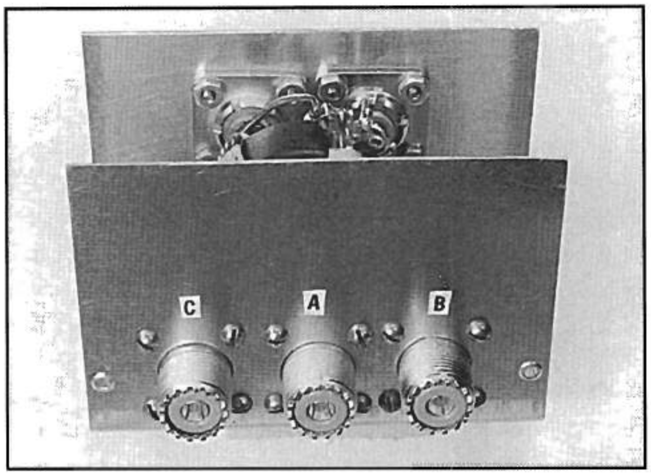

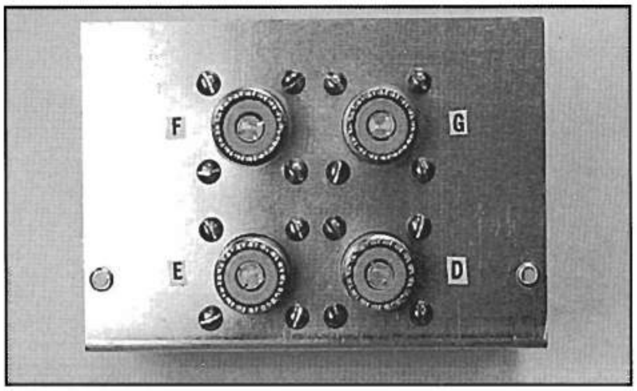

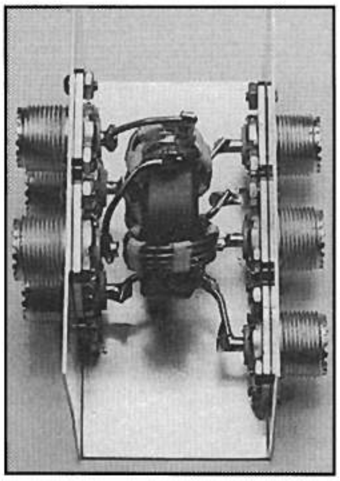

Photo 15-F. *Three different views of the 10-ratio Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high CU-3015A minibox.*

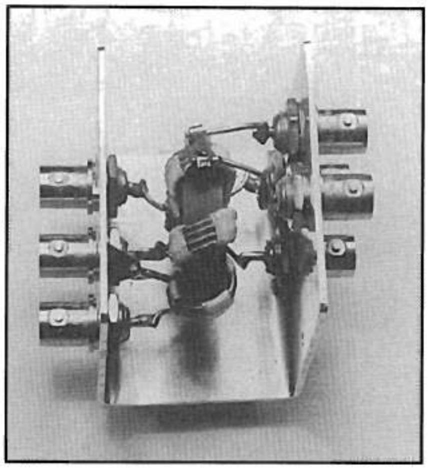

Photo 15-G. *The low-power unit mounted in a homemade 2.25 inch long by 1.5 inch wide by 2.25 inch high minibox.*

The bottom-right lead is terminal C. Below these three bottom leads is a ground connection (terminal 3 in **Figure 15-2B** to the SO-239 connector. **Photo 15-F** shows three different views of this high-power unit mounted in a 4 inch long by 2 inch wide by 2.75 inch high CU-3015A minibox.

This 10-ratio Unun has four quadrifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is tapped at 2 turns from terminal 5 and winding 7-8 is tapped at 2 turns from terminal 7.

If the 9:1 ratio (connection C-E), the 12.25:1 ratio (connection C-F), and the 16:1 ratio (connection C-D), are to be used at the full legal limit of amateur radio power, then I suggest that winding 3-4 be replaced with No. 12 H Thermaleze wire. If not, then these three ratios should be used at lower power levels (of 500 watts continuous and 1 kW peak). It should also be mentioned that using No. 12 wire for winding 3-4 adds a greater degree of *difficulty* to the construction process.

A listing of the expected performance across the band from 1.7 to 30 MHz, with the various ratios, is as follows:

16:1 (D-C); 50:3.125 ohms

Ratio is constant up to 21 MHz. It then decreases by 15 percent.

12.25:1 (F-C); 50:4.08 ohms

Ratio is constant.

9:1 (E-C); 50:5.56 ohms

Ratio increases by 5 percent.

6.25:1 (G-C); 50:8 ohms

Ratio is constant.

4:1 (D-B); 50:12.5 ohms

Ratio decreases by 5 percent.

3.06:1 (F-B); 50:16.3 ohms

Ratio decreases by 10 percent.

2.25:1

a) (E-B); 50:22.22 ohms

Ratio increases by 4 percent.

b) (B-E); 50:112.5 ohms

Ratio increases by 50 percent (the greatest deviation across the band of any of the ratios).

1.78:1

a) (D-A); 50:28.1 ohms.

Ratio is constant.

b) (A-D); 50:89 ohms.

Ratio increases by 15 percent.

1.56:1

a) (G-B); 50:32 ohms

Ratio increases by 10 percent.

b) (B-G); 50:78 ohms

Ratio increases by 40 percent.

1.36:1

a) (F-A); 50:36.8 ohms

Ratio decreases by 9 percent.

b) (A-F); 50:68 ohms

Ratio increases by 1.5 percent.

Sec 15.3.2 A Low-power 10-Ratio Unun

Photo 15-G shows a low-power unit mounted in a homemade 2.25 inch long by 1.5 inch wide by 2.25 inch high minibox. It has five quadrifilar turns of No. 16 H Thermaleze wire on a 1.25-inch OD ferrite toroid with a permeability of 250. The tap on winding 5-6 (**Figure 15-2B**) is at three turns from terminal 5 and on winding 1-2; it is three turns from terminal 1. Because the number of turns is different from the high-power unit, so are the ratios that use the taps. In this case, they are a little larger. Specifically, the tapped ratios are now: 1:12.96, 1:6.76, 1:3.24, 1:1.69, and 1:1.44. If the taps were at two turns from terminals 5 and 1, the ratios would be a little less than those of the high-power unit. You can play with the equations in the first section of this chapter and arrive at many different ratios.

Because this Unun has shorter transmission lines than its high-power counterpart, the deviations of the ratios across the HF band are generally smaller. Also, if winding 3-4 (in **Figure 15-2B**) were replaced with No. 14 H Thermaleze wire, this low-power unit could very well be rated at 500 watts of continuous power for all ratios!

Chapter 16

Ununs for Beverage Antennas

Section 16.1 Introduction

The Beverage antenna[19] is well known by 160-meter enthusiasts for enhanced signal-to-noise ratios when there are high levels of interference and atmospheric noise. If erected properly, Beverages also have excellent directivity. However, they are quite inefficient and, therefore, not generally suitable as transmitting antennas. Important considerations with Beverages are the terminating resistor (for the more common single-wire version) and the input matching Unun (unbalanced-to-unbalanced transformer). The terminating resistor and the impedance ratio of the Unun are determined by the characteristic impedance of the antenna acting as a long transmission line with one good conductor and one poor conductor (the earth). This line is generally between 400 and 600 ohms, and theoretically given by:

$$Z_0 = 138 \times \log(4h/d) \qquad \text{(Eq 16-1)}$$

where:
Z_0 = characteristic impedance of the Beverage
h = height of the wire above ground
d = diameter of the wire.

This chapter presents low- and high-power versions of multimatch Ununs designed to match 50-ohm cable to unbalanced loads from 450 to 800 ohms. The low-power unit, which is capable of handling continuous power levels up to 100 watts, is specifically designed for the Beverage antenna when it is performing as a receiving antenna. The high-power unit, which is capable of handling 1 kW of continuous power, can be used with the Beverage or any other traveling wave antenna when used as a transmitting antenna. Also presented are high-power designs capable of flat response, including the entire AM broadcast band. These multimatch Ununs could be of interest to designers of high-power amplifiers for the broadcast band. A little theory on how these devices are designed is also provided.

Sec 16.2 A Little Theory

Transmission line transformers[2] (the Unun being a subset thereof) are known for having greater bandwidths and efficiencies than their counterparts, the conventional transformers. Design considerations for the two types of transformers are also vastly different. Transmission line transformers use chokes and transmission lines, while conventional transformers use flux linkages.

High-impedance Ununs (and Baluns), which match 50 ohms unbalanced to impedances as high as 800 ohms, lie at about the edge of this technology's capability. The reasons are: 1) the windings require more turns because higher reactances are needed for isolating the input from the output, and 2) they require higher characteristic impedances in the transmission lines because the loads they see are greater. Therefore, when winding one of these devices, you'll *just* run out of space on your toroidal cores when trying to satisfy the low frequency and high frequency objectives. Incidentally, beaded transmission lines are not recommended at these impedance levels because of their excessive losses.

There are two methods for obtaining broadband operation at these high impedance levels. One uses Guanella's 9:1 and 16:1 Baluns, which are converted to Unun operation.[2] The other uses higher-order windings (quadrifilar in this case) on a single core, which is an extension of Ruthroff's bootstrap approach.[2] The Guanella approach, which uses coiled transmission line connected in series at the high-impedance side and in parallel at the low-impedance side, results in very broad bandwidths—but with difficulty in meeting low-frequency objectives. Low-frequency models[2] show that, with ratios above 4:1,

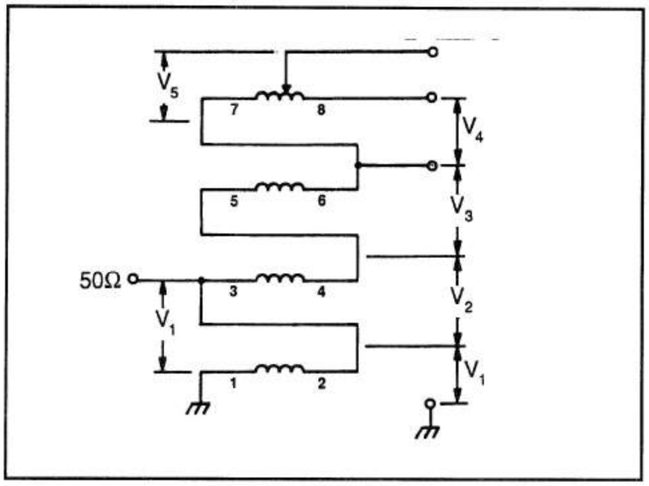

Figure 16-1. *Schematic diagram of the quadrifilar design, using Ruthroff's approach for high-impedance, low-frequency Ununs like the Beverage matching transformer.*

some of the coiled windings are connected in parallel, resulting in reduced reactances. However, with the Ruthroff approach, all of the inductances (at the low-frequency end) end up mutually aiding each other. However, Ruthroff's approach suffers at the high-frequency end because a direct voltage is summed with three voltages that traverse various lengths of transmission lines. As a result, Guanella's approach of summing voltages of equal delays is preferred for the higher frequency bands, and Ruthroff's approach is most often used for the lower frequency bands. This chapter presents designs using Ruthroff's approach.

Figure 16-1 shows the schematic diagram of a quadrifilar-wound Unun. If the lengths of the transmission lines are very short compared to the wavelength (therefore, phase delay and standing waves are negligible), then:

$$V_1 = V_2 = V_3 = V_4 \qquad \text{(Eq 16-2)}$$

at terminal 6,

$$V_o = V_1 + V_2 + V_3 = 3V_1 \qquad \text{(Eq 16-3)}$$

and the impedance ratio becomes:

$$g = (V_o/V_1)^2 = 9 \qquad \text{(Eq 16-4)}$$

At terminal 8, it becomes:

$$g = 16 \qquad \text{(Eq 16-5)}$$

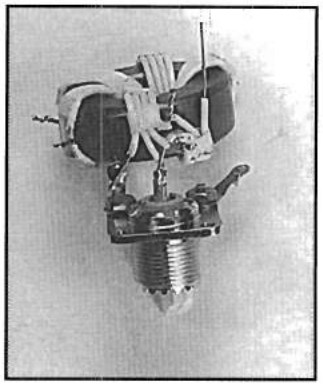

Photo 16-A. *The bottom view of the low-power Beverage antenna Unun.*

The voltage at the tap in winding 7-8 is:

$$\begin{aligned} V_o &= 3V_1 + V_5 \\ &= 3V_1 + n/NV_1 = V_1(3 + n/N) \\ &= V_1(3 + n/N) \qquad \text{(Eq 16-6)} \end{aligned}$$

where:
N = total number of turns
n = number of turns from terminal 7

The impedance ratio, using the tapped winding, becomes:

$$g = (V_o/V_1)^2 = (3 + n/N)^2 \qquad \text{(Eq 16-7)}$$

When the lengths of the transmission lines are significant, then important phase delays can occur and reduce the high frequency response. As you can see in **Figure 16-1**, V_2 travels one transmission line, V_3 travels two transmission lines, and V_4 travels three transmission lines. Additionally, the high frequency response is further diminished if the characteristic impedances of the transmission lines are not at their optimum values (which is hard to do at these impedances levels). Even with these major flaws, the Ruthroff approach is better for Beverage antenna use because this antenna's greatest advantages are on the lower frequency bands (80 and 160 meters).

Sec 16.3 A Low-power Design

Photo 16-A shows the bottom view of a 5-turn quadrifilar-wound Unun designed to handle 100 watts of continuous power with constant ratios from 9:1 to 16:1 in the 40- and 80-meter bands. It uses the Ruthroff approach of **Figure 16-1** and is shown here to give the reader a method for making the various interconnections. For operation on the 80- and 160-meter bands, I would use 6 quadrifilar turns on a 1.5-

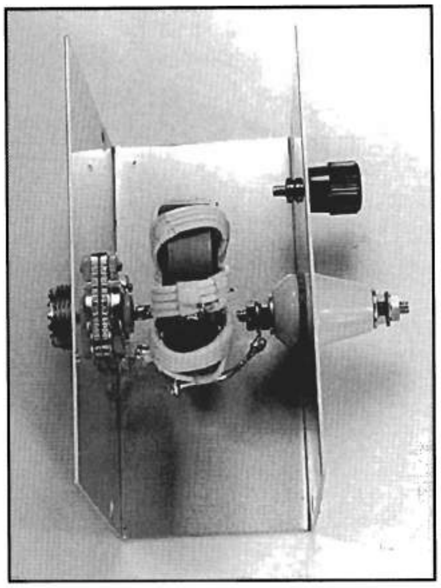

Photo 16-B. *The low-power Beverage antenna Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox.*

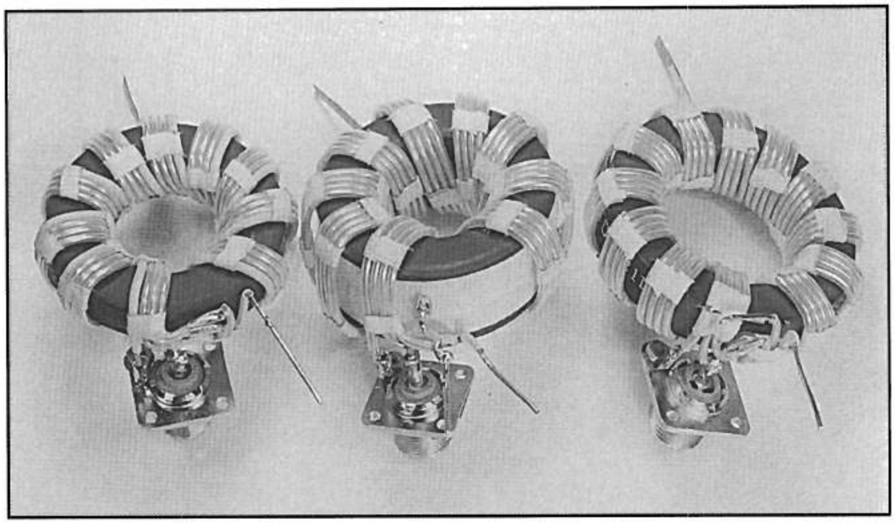

Photo 16-C. *Three high-power, low-frequency Ununs using a quadrifilar design with Ruthroff's approach. The one on the left is designed to cover the 80- and 160-meter bands. The other two are designed to cover the 160-meter and AM broadcast bands.*

inch OD ferrite toroid with a permeability of 250. The bottom winding is No. 20 hook-up wire and the other three are No. 22 hook-up wire. Winding 7-8 is tapped at 2 turns from terminal 7, yielding a 11.11:1 ratio, and at 3 turns from terminal 7, yielding a 12.25:1 ratio. Therefore, with outputs also at terminals 6 and 8, this Unun matches 50-ohm cable to loads of 450, 555.6, 612.5, and 800 ohms.

Photo 16-B shows the unit mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox. The output (the feedthrough insulator) is connected to one of the taps. A grounded binding post is also shown.

Sec 16.4 High-power Designs

Photo 16-C shows three high-power designs. The one on the left is specifically designed to cover the frequencies generally used with traveling wave structures like the Beverage antenna. This design has 10 quadrifilar turns on a 2.4-inch OD ferrite toroid with a permeability of 250. Winding 1-2 is No. 14 tinned copper wire, and the other three are No. 16 tinned copper wire. The wires are also covered with Teflon sleeving. Winding 7-8 is tapped at 5 turns from terminal 7, yielding a ratio of 12.25:1. When matching 50-ohm cable to loads of 450 ohms (terminal 6), 612.5 ohms (the tap), and 800 ohms (terminal 8), the variation in response is less than 5 percent from 1.5 to 4 MHz. At 6.5 MHz, the variation (which is an increase in the impedance ratio) increases to about 20 percent. **Photo 16-D** attempts to provide a better view of the connections.

The other two high-power Ununs in **Photo 16-C** are specifically designed to cover the broadcast and 160-meter bands. The one in the center has 9 quadrifilar turns (of the same wires as above) on a stack of two 2.4-inch OD ferrite toroids with permeabilities of 250. The tap on winding 7-8 is now at 4 turns from terminal 7, yielding a ratio of 11.86:1. When matching 50-ohm cable to 450 ohms (terminal 6), 593 ohms (the tap), or 800 ohms (terminal 8), the response is literally

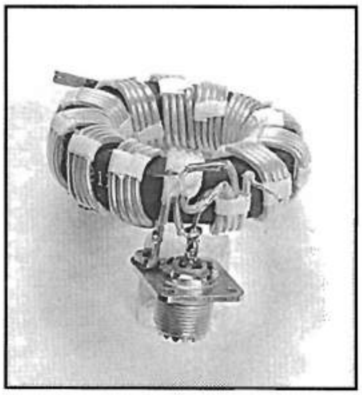

Photo 16-D. *The bottom view of the high-power Beverage antenna Unun.*

flat from 0.5 to 2 MHz. At 4 MHz, the ratios increase by about 6 percent. At 7 MHz, they increase by about 20 percent.

The Unun on the right in **Photo 16-C** illustrates another way of obtaining the same performance as above. In this case, the design has 12 quadrifilar turns (of the same wires as above) on a 2.68-inch OD ferrite with a permeability of 290. The tap on winding 7-8 is at 6 turns from terminal 7, yielding a 12.25:1 ratio (instead of 11.86:1 as above). Although the performance of this design is practically the same as the one above (using the two 2.4-inch OD cores), it is a much more expensive design because the 2.68-inch OD core is not nearly as popular. However, if a broadband, high-power and high-impedance Unun (or Balun) is required to cover 1.5 to 30 MHz, then these expensive 2.68-inch OD ferrite cores are very likely the *only* alternative!

Chapter 17

Concluding Remarks

Sec 17.1 Introduction

As I've often said, there is little information that deals with aspects of design and applications of practical hardware. Most companies are reluctant to publish their results for fear of giving away their hard-earned secrets. Textbooks only contain a few paragraphs on the subject. Therefore, very few people fully understand this technology, and, as a result, it is still far from reaching its full potential.

This book not only contains my designs, which have appeared in series written for *CQ* and *Communications Quarterly*, but also my views on other articles that have appeared in the amateur radio literature.

In the process of converting and combining the articles into appropriate chapters, one section in my article in the *CQ*, March 1993 issue entitled "Dual-Ratio Ununs," stood out as having broader applications. It is entitled "Reflections on Power Ratings" and is presented here. This information is followed by a section on misconceptions and one on the "state of the art."

After reading this book, some might think I was overly critical and didn't agree with any of the designs or explanations (or both) presented in the amateur radio literature. This is quite true. In taking this stand, I was hoping to provoke, in return, critical comment on my work. In this way, we can help our amateur friends—and perhaps even our professional friends—by advancing the understanding and application of these very useful transformers.

Sec 17.2 Reflections on Power Ratings

Power rating is one of the most controversial and least understood specifications for the transmission line transformer. As of today, no professional group has yet set the standard for this specification (as well as any other) for this popular class of matching transformers. In fact, manufacturers of ferrites, which are mainly the materials used with these devices, only specify them for their uses as conventional transformers and inductors, or microwave devices.

It is well known that power ratings for practically all conventional devices are based upon catastrophic failures (usually exceeding a voltage or current limit), and failures over a relatively short period of time due to an excessive rise in temperature.

With transmission line transformers, there are really two catastrophic-type failures that can occur. One is voltage breakdown. If the device is misterminated with a high impedance (especially an open circuit), a breakdown of the insulation can occur. This is particularly true of the 1:1 Balun (50:50 ohms) that is terminated with the very high impedance of a full-wave dipole or inverted V. Using heavily coated wires (but still maintaining a characteristic impedance close to 50 ohms), or small but high-power coaxial cable, can help under these conditions.

The second catastrophic failure occurs at the low-frequency end of the transformer's passband, when the energy is not completely transmitted to the output circuit by a transmission line mode. This takes place when the reactance of the coiled or beaded transmission line is not sufficient to prevent conventional transformer currents or shunting currents to ground. Under these conditions, harmful flux can take place in the core or beads. Nonlinearities can also occur if the flux becomes appreciable. The objective in design at the low-frequency end is to have a margin of safety such that, with a termination of about three times (hence VSWR of 3:1) that of a matched condition,[2] no flux will appear in the core or beads.

The failure due to an excessive rise in temperature is the least understood of the two because it involves the failure mechanism in transmission line transformers when only transmission line currents are allowed to flow. Unlike the conventional transformer whose losses are current dependent (wire, eddy current, and hystereis losses), the transmission line transformer's losses are voltage dependent (a dielectric-type). That is, the greater the voltage drop along the length of the transmission lines, the greater the loss. Furthermore, it can be shown that only low-permeability ferrites (less than 300) yield the extremely high efficiencies of which these transformers are capable.[2]

Because all transmission line transformers have voltage drops along their transmission lines, we must look at their high-frequency models to determine the magnitude of these drops and, hence, the temperature rise that can be expected. Here are some examples:

If V_1 is the voltage that appears on the 50-ohm side of the transformer, then:

1) For a quintufilar-wound Unun, the longitudinal voltage-drop is $V_1/5$ in matching to lower impedances (like 32 or 18 ohms) and $V_1/4$ in matching to higher impedances (like 78 or 139 ohms).

2) For a quadrifilar-wound Unun, the longitudinal voltage-drop is $V_1/4$ in matching to lower impedances (like 12.5 or 28 ohms) and $V_1/3$ in matching to higher impedances (like 89 or 200 ohms).

3) For a trifilar-wound Unun, the longitudinal voltage-drop is $V_1/3$ in matching to lower impedances (like 22.22 or 25 ohms) and $V_1/2$ in matching to higher impedances (like 100 or 112.5 ohms).

4) For a bifilar-wound Unun, the longitudinal voltage-drop is $V_1/2$ in matching to a lower impedance of 12.5 ohms and V_1 in matching to a higher impedance of 200 ohms.

Because the quintufilar-wound Unun has the lowest voltage drop, it is expected to have the highest efficiency. Furthermore, it can be seen that the highest efficiencies occur in matching to lower impedances. Very accurate measurements[2] have shown that 4:1 Ununs, using ferrite cores with permeabilities of 125, have exhibited losses of only 0.02 to 0.04 dB in matching 50 to 12.5 ohms from 1 MHz to over 30 MHz. Even though the Ununs in this book have mostly used permeabilities of 250, and should have slightly greater losses, many use higher-order windings (trifilar, quadrifilar, and quintufilar) and, hence, have lower longitudinal voltage drops. Therefore, they should have losses of only 0.02 to 0.04 dB, as well.

When matching at the 1 kW level, the figures above mean that only 5 to 10 watts would be dissipated in the Unun. As a heatsink, these small transformers should be able to handle this loss easily. In fact, they should be able to handle several times this level of continuous power. Also, because they use heavily coated wires, their peak power ratings should be greater by more than a factor of two!

Another important power rating consideration is to determine what happens when the transformers are misterminated. Because the losses being considered now are dielectric-types and, hence, voltage-dependent, the harmful terminations are greater than that for which the transformers were designed. For example, if the termination is three times greater (a VSWR of 3:1), the voltages along the transmission lines would increase by a factor of 1.73. This means the losses would practically double. The Ununs described in this series, when matching to impedances lower than 50 ohms, should easily handle this mismatch. Obviously, mismatches in the range of 10:1 would result in much lower efficiencies and should be avoided.

The analysis of the losses in Baluns follows the same pattern. The voltages are as follows: under matched conditions for a 1:1 Balun (50:50 ohms), $V_1/2$; for a 4:1 Balun (50:200 ohms), V_1; for a 9:1 Balun (50:450 ohms), $1.5V_1$; and for higher-impedance Baluns it could be $2V_1$. The higher voltage drops, together with high VSWRs, means that high-impedance Baluns (and Ununs) have more loss and require larger structures to dissipate the heat. It should also be pointed out that there is a tradeoff in efficiency for low-frequency response with Baluns (and Ununs) when matching 50 ohms to higher impedances like 200 ohms, 300 ohms, 450 ohms, and higher. This is done by using permeabilities of 125 and lower.

Finally, I thought it might be useful to give some general guidelines as to what efficiencies you might expect with Baluns and Ununs when using ferrite cores or beads with a permeability of 250. Here are some expected efficiencies when matching 50 ohms to various loads under matched conditions:

Loads	**Efficiency**
50 ohms or less	98 to 99.5 percent
50 to 100 ohms	97 to 98 percent
100 to 200 ohms	96 to 97 percent
200 ohms and above	93 to 96 percent

As I mentioned earlier, these efficiencies would be reduced by a percent or two with a VSWR of 3:1, which increases the loss by a factor of about two. Also, the efficiencies can be increased by a percent or two with high-impedance loads (greater than 100 ohms) by resorting to lower permeability ferrites that trade off efficiency for low-frequency response. In closing, I would like to say that high-permeability manganese-zinc ferrites should be avoided because of their much higher losses. Furthermore, their losses are highly frequency dependent, while low-permeability nickel-zinc ferrites are not.

Sec 17.3 Misconceptions

From recent discussions on the air and phone calls concerning Baluns, I think the most *expensive* misconception regarding Baluns is the assumption that a 9:1 (450:50 ohm) Balun would match 50-ohm cable (or the output of a linear or transceiver) to 450-ohm twin lead, without considering the effect of its termination. In truth, the 9:1 Balun would *only* see 450 ohms if the line were terminated in 450 ohms. In reality, if the line were terminated in a 50-ohm dipole, the Balun would see 50 ohms when the line is a half-wave long and 4050 ohms when is a quarter-wave long. The 9:1 Balun is clearly useless in this application.

By far, most misconceptions regarding Baluns are due to the many radio amateurs who perceive these devices as conventional transformers that transmit the energy from input to output by flux linkages and not as transmission line transformers, which transmit energy by an efficient transmission line mode. This is clearly shown by the writers who have compared their "new" coaxial cable (coiled about a toroid or threaded through ferrite beads) Baluns with Baluns using wire transmission lines coiled about a ferrite rod or toroid. They claim their Baluns are better because the others: 1) were limited by leakage inductance, 2) did not exhibit true 1:1 impedance transformations, 3) were prone to core saturation, 4) added a reactive component to the input impedance, 5) were susceptible to unbalanced and mismatched loads, and more importantly, 6) had *more* loss.

If the writers had accepted the correct model for these devices (given to us by Guanella and Ruthroff), which shows that they are really chokes (lumped elements) and configurations of transmission lines (distributed elements), then there are several parameters they should have considered in their comparisons. They are: 1) the characteristic impedances and lengths of the transmission lines (the high-frequency capability), 2) what form of the 1:1 Balun or 4:1 Balun is used by the other Balun, 3) the low-frequency capabilities (safety margins), 4) power capabilities, and finally 5) efficiencies.

Now, had the writers used the proper parameters in their comparisons, they would have found that mismatch loss was mistaken for real (ohmic) loss; high-frequency response was limited by standing waves, and not leakage inductance or shunting capacitance; the beaded-coax Balun had *more* loss than a well-designed Balun using wire or coax transmission lines coiled about a toroid; and that their comparisons were made with either the trifilar (voltage) 1:1 Balun or the Ruthroff 4:1 Balun, which are inferior designs.

In fact, the perception that the transmission line transformer is actually a conventional transformer is so prevalent, that a new name for this class of devices should be considered—*broadband transmission line matching networks*. This name (without the word *transformer*) would help in dispelling inaccurate perceptions and in standardizing the schematic diagrams. It would place the coiled or beaded transmission lines (in the high-frequency models) horizontally, and eliminate the phasing or polarity dots.

Sec 17.4 The State of the Art

Until very recently, the radio amateur had only two types of Baluns available in the literature and on the market. They were the so-called 1:1 and 4:1 "voltage" Baluns. As was shown in **Chapter 7**, the comparisons by others with new 1:1 designs using coaxial cable (called "current" Baluns) were made with an inferior trifilar-wound Balun, instead of Ruthoff's design that appeared in his 1959 paper and became the industry's standard. Ruthroff's third conductor on his 1:1 Balun was on a separate part of the toroid, thus giving it practically the same characteristics as the Guanella ("current") Balun. These articles on newer designs not only gave a new language to our Baluns, but also presented questionable statements regarding their performances. It would be interesting if the authors of these articles compared their Baluns with well-designed Ruthroff or Guanella Baluns using 50-ohm bifilar windings or coaxial cables on low-loss ferrite toroids (less than 300 permeability). I am quite sure their claims would be greatly diminished.

As was noted in **Chapter 8**, the 4:1 voltage Balun appeared in the amateur radio journals about 25 years ago (the same time as the "inferior" 1:1 voltage Balun). Considerable design information appeared in the handbooks of the time regarding the construction and performance of this Balun. Furthermore, this information also stayed the same over these many years. As was shown in **Chapter 8**, the design was found lacking. However, with some rather simple changes, like doubling the cross-sectional area of the core, increasing the number of turns from 10 to 14, and using extra insulation on the wires to increase the characteristic impedance of the coiled transmission line from about 50 to 100 ohms (the objective), a much better design emerged. In fact, for balanced antenna systems, this new design might well be described as "peerless."

A 4:1 Guanella (current) Balun has now appeared in our handbooks. This more flexible Balun uses two transmission lines wound on separate cores and connected in series at one end and in parallel on the other. Literally, no design information is given on its construction. What is offered are recommendations for the permeability of the ferrite cores. Values from 850 to 2500 are proposed. However, use of these high permeabilities would result in lossy Baluns.

I also found it interesting, in my work on these devices, that the classic papers of Guanella[3] and Ruthroff[9] are still the cornerstones of this technology known as *transmission line transformers*. To be sure, some of us have extended the work of these two by using better measuring equipment, creating more complicated configurations, and finding new applications. However, it is apparent from the articles published in the amateur radio journals and discussions on the air and at club meetings that most radio amateurs still perceive these devices as conventional transformers. They don't look at these devices as Guanella and Ruthroff did—as chokes and transmission lines. As a result, there has been a lack of good design information in our literature.

There are many new and useful designs possible with this technology, as discussed in **Part I** of this book. They include: higher power levels, applications on the VHF and UHF bands and above, and new Baluns and Ununs with ratios other than 1:n^2 where n = 1, 2, 3, . . . , etc. This book presents some designs* and suggestions for higher-power and higher-frequency applications.

I see two reasons for the lack of emergence in this technology. They are:

1) This subject is not adequately covered in any college textbook, and it generally has not been of interest to academics who rightfully view their role as basic research and not applications. As a result, there are few graduates with any skill in the design of transmission line transformers—in contrast to the areas of transmission line, waveguide, and antenna theory.

2) The professional societies don't receive enough application papers. Although much of the research and development work performed in industry is highly innovative, important to the advancement of the technology, and certainly publishable in scientific journals, corporations are often reluctant to allow publication for fear of "aiding" their competition. It has been stated[29] that in the past few decades, the submission of application papers to the technical journals of the IEEE has declined. In fact, a survey by one of the technical societies showed that 85 percent of the submissions now come from universities, not industry!

In order to assist technologies, like *transmission line transformers*, which are far from reaching their potential applications, IEEE has instituted a program called *Emerging Practices in Technology* (EPT). The object of the EPT program is to facilitate the development of new standards by disseminating and making available various EPT papers to the broadest possible audience worldwide. The papers on practices in various areas of technology are peer reviewed by relevant IEEE Technical Committees, and have the potential for standardization in the future. The papers (mine is **Reference 30**) are published by the IEEE Standards Press.

FOOTNOTE

*Kits and finished units available from Amidon Associates. Inc., 240 Briggs Ave., Costa Mesa, CA 92626.

Part II

References

1. Jerry Sevick, W2FMI, *Transmission Line Transformers*, 1st edition, Amateur Radio Relay League, 1987.

2. Jerry Sevick, W2FMI, *Transmission Line Transformers*, 4th edition, Noble Publishing Corporation, 2001.

3. G. Guanella, "Novel Matching Systems for High Frequencies," *Brown-Boveri Review*, Volume 31, September 1944, pages 327–329.

4. Walt Maxwell, W2DU, "Some Aspects of the Balun Problem," *QST*, March 1983, pages 38–40.

5. John Belrose, VE2CU, "Transforming the Balun," *QST*, June 1991, pages 30–33.

6. Roy Lewallen, W7EL, "Baluns: What They Do and How They Do It," *The ARRL Antenna Compendium*, Volume 1, Amateur Radio Relay League, 1985, pages 12–15.

7. Joe Reisert, W1JR, "Simple and Efficient Broadband Balun," *Ham Radio*, September 1978, pages 12–15.

8. Richard H. Turrin, W2IMU, "Broad-Band Balun Transformers," *QST*, August 1964, pages 33–35.

9. C. L. Ruthroff, "Some Broad-Band Transformers," *Proceedings of the IRE*, Volume 47, August 1959, pages 1337–1342.

10. Richard A. Genaille, W4UW, "How to Build a Multi-Tap Unun," *CQ* May 1992, pages 28–32.

11. Bruce Eggers, WA9NEW, "An Analysis of the Balun," *QST*, April 1980, pages 19–21.

12. Richard H. Turrin, W2IMU, "Application of Broad-Band Balun Transformers," *QST*, April 1969, pages 42, 43.

13. George Badger, W6TC, "A New Class of Coaxial-Line Transformers, Part 1," *Ham Radio*, February 1980, pages 12–18.

14. George Badger, W6TC, "A New Class Of Coaxial-Line Transformers, Part 2," *Ham Radio*, March 1980, pages 18–29, 70.

15. Bill Orr, W6SAI, "Radio Fundamentals: The Coax Balun," *CQ*, November 1993, pages 60–65.

16. Albert Roehm, W2OBJ, "Some Additional Aspects of the Balun Problem," *The ARRL Antenna Compendium*, Volume 2, Amateur Radio Relay League, pages 172–174.

17. Lew McCoy, W1ICP, "Let's Talk About Wire," Part 1, *CQ*, January 1993, pages 42–45.

18. Lew McCoy, W1ICP, "Let's Talk About Wire," Part 2, *CQ*, February 1993, pages 32–42.

19. Gerald Hall, K1TD, Editor, *The ARRL Antenna Book*, 16th edition, Amateur Radio Relay League, 1991, Chapter 25.

20. George Grammer, W1DF, "Simplified Design of Impedance-Matching Networks," in three parts, *QST*, March, April, and May 1957.

21. D.K. Belcher, "RF Matching Techniques, Design and Example," *QST*, October 1972, page 24.

22. Lew McCoy, W1ICP, "The Ultimate Transmatch," *QST*, July 1970, pages 24–27, 58.

23. Warren Bruene, W5OLY, "Introducing the Series-Parallel Network," *QST*, June 1986, page 21.

24. Louis Varney, G5RV, "The G5RV Muliband Antenna...Up-to-Date," *The Antenna Compendium*, Volume 1, Amateur Radio Relay League, pages 86–90.

25. Jerry Sevick, W2FMI, "6:1 and 9:1 Baluns," *Communications Quarterly*, Winter 1993, pages 43–51.

26. D. Meyer, "Equal-Delay Networks Match Impedances Over Wide Bandwidths," *MICROWAVES & RF*, April 1990, pages 179–188.

27. S.E. London and S.V. Tomeshevich, "Line Transformers with Fractional Transformation Factor," *Telecommunications and Radio Engineering*, Volume 28/29, April 1974.

28. Carl Markle, K8IHQ, "Wideband RF Baluns," *73 Amateur Radio Today*, September, 1992, pages 20–23.

29. S. Maas, D. Hornbuckle, D. Masse, "Applications Papers for the *MTT Transactions*," Summer 1992, IEEE MTT-S Newsletter, pages 3, 5.

30. Jerry Sevick, W2FMI, "Design and Realization of Broadband Transmission Line Matching Transformer," *Emerging Practices in Technology*, IEEE Standards Press, 1993.

Index

Amidon Associates, Inc...18, 30–32
    kits...30–31, 88
Antenna modes
    dipole mode...28
    inverted L mode...28, 31
Antenna tuners, Baluns for...55–63
Autotransformers...6, 9, 19, 21, 30–31, 52, 76, 97, 101

Badger, George...33–36, 89, 114
    two-tone test method...33
Baluns
    air-core...33, 41, 43, 46
    and dipoles...27–29, 35, 40, 45–46, 52–55, 57, 61–62, 65, 71, 100–111, 113
    and inverted Vs...27–28, 111
    and log-periodics...49, 54, 57,
    and multiband dipoles...61
    and quads...28, 65, 87
    and rhombics...80
    and Yagis...27–28, 33, 65, 73–74, 86
    attributes of a good design...41
    Balun/Unun...43, 87
    beaded-coax...36, 41, 54, 56, 86–87, 113
    bifilar toroidal...37–39, 40
    broadband...27, 29, 65, 71, 74, 78, 83, 85, 115
    coil...46
    Collins...34
    commercial...33, 36, 47, 52–54
    choke...4, 19, 27, 29–30, 34, 36–37, 40–41, 43, 56
    current...4, 27, 29–30, 35–36, 38–41, 43, 46–47, 50, 52–55–57, 61, 62, 65, 113–114
        bifilar current type...36
    dual-core...50–54, 65
    dual-ratio...65, 68–69
    economy...38–39
    ferrite core...30–31, 33–34, 39, 41, 46, 49, 78–79, 110
    for antenna tuners...55–63
    for 2-meter operation...65
    fractional-ratio...74
    "hardened"...62
    HF designs...65
    high-impedance...15, 79–80
    high-power...27, 37–40, 49–53, 57–58, 65, 81–82
    HI-Q...52–53
    low-power...27, 37, 39, 43, 50–54, 65
    medium-power...27, 37, 39, 81
    misconceptions about...113
    parallel-type...65, 66, 68–71, 74–75, 79, 82
    RF...29, 115
    rod-type...4, 31–34, 40, 47, 52–54, 63
    series-type...16, 65–69, 72–75, 79–82
    single-core...45, 50–51, 54, 73–74
    step-down...28, 51–52, 65, 73–74, 86
    step-up...65, 74
    three-conductor...30–31, 34–35, 43
    toroidal...32, 36–39, 41
    transformer-type...34–35
    trifilar-wound...30, 32, 113
    two-conductor...30–31, 33–37, 53
    VHF designs...65
    voltage...33, 35–36, 43, 47, 52, 54–55, 62, 113
    when to use...27–29
    "worst case" design...61–63
    1:1...2, 4, 15–16, 27–41, 43, 45, 54–56, 61–66, 68, 71, 89, 111–113
        articles on...29
        bifilar toroidal...37–39, 41
        in amateur radio literature...30–37
        trifilar-wound...30, 32, 113
        using ferrite cores...30, 112
    1:4...5–7, 11–12, 16, 21–22, 55–56, 67, 70, 72–74
    1.5:1...65–70, 89
    1.56:1...65–66
    1:6...72
    1:6.25...16, 72–73, 79–80
    1:9...15–16, 70–71, 76, 78–82
    2:1...28, 62, 65–71, 89
    4:1...5, 7, 27, 30, 35–37, 41, 43–62
        history...43–44
        beaded coax...54
        current...53
    6:1...71–82, 89
    6.25:1...71–75, 79
        high-impedance...74, 79
        low-impedance...74, 79

  - 9:1...27, 36–37, 71–82, 89
    - high-impedance...79
    - low-impedance...79
    - practical designs...77–78
  - 12:1...71–82, 89
  - 16:1...27, 107
  - 50:450-ohm...71, 82
- Balun/Unun...43, 87
- Beach umbrella antenna...Preface
- Belrose, John...36–37, 115
- Beverage antennas...107–110
- Bootstrap...5, 7, 10–11, 13, 17, 21–22, 30, 68, 74, 83, 93, 107
- Broadband matching transformers...29
- Broadband transmission line matching networks...113

- Characterization...17–20
- Cores
  - air...33, 41, 43, 46
  - ferrite...30–31, 39, 46, 49, 78–79, 110, 112, 114
  - geometry of...17
  - low permeability...41
  - magnetic...17, 30, 36, 46, 77
  - materials...18
  - powdered-iron...30, 32, 40, 53, 57–62, 80
  - rod...3–4, 17–18, 51–52, 97
  - toroidal...3, 11, 17, 35, 43, 46, 102, 107
- Crossover...32–33, 37–39, 50–51, 66–68,

- Delay line...4–5, 20, 22, 45, 77, 86
- Dipole antennas...27–29, 35, 40, 45–46, 52–55, 57, 61–62, 65, 71, 100, 111, 113

- Eggers, Bruce...28, 114

- Ferrite
  - Balun...33
    - beads...4, 7, 29, 34, 36, 43, 55–56, 69–70, 75, 78–80, 82, 113
  - high-permeability...34–35, 54, 80
  - core...30–31, 39, 46, 49, 78–79, 110, 112, 114
  - efficiencies...39
  - low-loss...32–33, 113
  - low permeability...30, 37, 48, 54, 63, 75, 86, 112
  - manganese-zinc...4, 30, 113
  - nickel-zinc...18, 113
  - rod...11, 20, 52, 84, 86
  - toroids...13, 17, 31–32, 35–39, 47, 49–51, 53–54, 56, 66–68, 72, 74, 77, 81, 84–85, 90, 92–93, 101–104, 106, 109, 113
- Flux canceling...21, 29, 83
- Floating load...44–46, 58, 72–74, 77, 80
- Fractional-ratio Ununs...86, 89–90

- Genaille, Richard...28, 97, 115
- Guanella, G.
  - basic building block...3–5, 10–11, 16, 21–22, 29–30
  - classic paper of 1944...3–5, 23, 27, 43, 57, 83, 114–115
  - current/choke Balun...4, 27, 30, 36, 38–40, 43, 50, 54, 56
  - dual-core Balun...52–53
  - equal-delay approach...11, 13, 22, 86
  - single-core Balun...44–45, 49–52
  - two conductor...30–31, 33–34, 37, 40, 43
  - 1:1 Balun...4, 29–31, 33–34, 38–40, 43, 65–66, 68
  - 1:4 Balun...5–7, 11, 16, 21, 56, 67, 72, 79, 82
  - 1:4 TLT...5–7, 22
  - 1:4 Unun...5, 7, 20–22
  - 1:9 Balun...15–16, 70, 76, 80–82
  - 4:1 Balun...5, 7, 27, 43–45, 49–54, 56–57,68, 72–74, 76, 79, 114
  - 4:1 Unun...7, 45, 83, 85–86
  - 9:1 Balun...27, 71–72, 75–82, 107
  - 16:1 Balun...27, 86, 107
- G5RV antenna...55, 61–62, 115

- Hairpin matching networks...28, 65
- HI-Q Balun... 52–53

- Imbalance current...28, 30–31, 35, 41
- Inverted V antennas...27–28, 111
- Isolation transformers...4, 27, 40, 66

- L-C matching network...55
  - L...56
  - pi...56
  - T...56
  - Ultimate...56
  - SPC...56
- Lewallen, Roy...35–36, 115
  - 1:1 Balun...35
  - 4:1 Balun...35
  - current Balun...35
  - voltage Balun...35
- Link coupling...55
- Loads
  - balanced...46, 51–52, 65–70, 72–75, 79, 81–82, 86
  - center-tapped-to-ground...47, 60, 62, 73–74
  - floating...44–46, 56–58, 60, 62, 72–74, 77, 80
  - grounded at bottom...4, 45, 54, 73–74, 77
  - grounded at center...3–4, 30–31, 37, 45–46, 54–55–57, 60, 73–74, 77
  - mismatched...30, 32–35, 41, 113

  reactive...36, 62
  unbalanced...27, 30, 33–35, 41, 58, 85, 90–91, 107
Log-periodic antennas...49, 54, 57

Magnetizing inductance...17–18
Materials, core...18
Maxwell, Walt...34–37, 115
  choke Balun...34
  transformer-type Balun...34–35
  1983 article...34
McCoy, Lew...52–53, 55, 57–60, 62, 101, 115
Multiband antennas...55, 61–62
  dipoles...61
Multimatch transformers...89, 102

Orr, Bill...33, 115

Phase-delay...30, 37, 39, 73, 85, 87, 94, 108
Phase-inverter...4–6, 21, 30, 40, 43, 46, 54–55, 57
Pi network...56
Polarity dots...113
Power ratings...35, 54, 94, 111–112
  loads and efficiencies...112–113

Quad antennas...28, 65, 87

Reisert, Joe...32–35, 37–38, 41, 54, 115
  crossover...32–33, 37–39
Rhombic antennas...80
Roehm, Albert...36, 56, 115
Ruthroff, C.L.
  bifilar Unun...89
  bootstrap...5, 7, 10–11, 13, 15, 17, 21–22, 30, 68, 74, 83, 93, 107
  broadband transmission line transformers...22, 29
  classic paper of 1959...23, 27, 30, 34–35, 43, 46, 83, 113–115
  high-frequency model...5, 45
  hybrid transformer...43
  maxima technique...46
  phase-inverter configuration...21, 55, 57
  three-conductor Balun...30–31, 34, 43
  trifilar-wound Balun...15, 30, 32, 113
  voltage Balun...43, 52, 54–55, 62, 113
  1959 article...1, 4, 9, 10, 17, 29, 42, 57, 61, 87, 88
  1:1 Balun...3–4, 29–30, 34, 36, 43, 113
  1:4 Balun...5–7, 22, 56, 81
  1:4 TLTs...6
  1:4 Unun...5–6, 9, 11, 16, 21
  4:1 Balun...30, 43, 45–50, 53, 55–59, 61–62, 113
    rod-type...52–53
  4:1 bifilar Unun...89, 97
  4:1 Unun...30, 45–46, 68, 83–84, 87, 97
  100:25-ohm Unun... 84–85
  200:50-ohm Unun...84–85

Series and parallel-tuned circuits...55
Single-core Balun...45, 50–51, 54, 73–74
Step-down transformers...90
Step-up transformers...90, 97–99

Terman, Frederick E...22–23
Toroids...17, 20–21, 30, 51, 69, 73, 79, 81–82
  cores...3, 11, 35, 46, 107
  current-probe...35
  ferrite...13, 17, 31–32, 35–39, 47, 49–51, 53–54, 56, 66–68, 72, 74, 77, 81, 84–85, 90, 92–93, 101–104, 106, 109, 113
  powdered-iron...17, 31
  4C4...9, 18
Transformation ratios...37, 41, 43–44, 51–52, 57, 65–66, 68, 74, 85–86, 89–90, 92–93, 101–105
Transformers
  air-core...33
  Balun...56, 115
  bifilar, tapped...9
  broadband...30, 43, 100
    multimatch...89, 102
  conventional...21, 57, 76, 79, 83, 107, 111, 113–114
  equal-delay...12
  flux linkages...21, 83, 107, 113
  hybrid...43
  isolation...4, 27, 40, 66
  matching...29, 61–62, 108, 111
  multimatch...63, 76
  parallel...100
  step-down...90
  step-up... 90, 97–99
  toroidal...17, 91
  transmission line...3, 5, 21, 23, 27, 29, 43–44, 79, 82–83, 101–102, 107, 111–115
  unbalanced-to-unbalanced...43, 45, 65, 83, 107
  wideband..33
Transmatch...55–56, 58, 60, 101, 115
Transmission lines
  balanced...56
  beaded...17, 34, 37, 54, 57, 69–70, 75–76, 83, 86, 70, 107, 111, 113
  coax...29, 34, 83, 113
  coiled...5, 17, 29, 31, 33–34, 36, 41, 43–44, 46, 51–52, 57, 76, 83, 86, 107, 111, 114

  coupled...30, 32, 35, 43
  currents...27, 29, 36, 39, 41, 44, 46, 49, 76, 83, 111
  flat line...44, 46, 47, 57, 69, 74
  half-wave...29, 61
  high-impedance...75, 82
  ladder line...71
  matched...29, 62
  McCoy...53
  open-wire...5, 56
  quarter-wave...29
  theory...61, 76
  twin-lead...29, 34, 38, 56, 75
  very short...29, 44, 84, 90, 97–98, 104, 108
Transmission line transformers (TLTs)
  book by W2FMI...12, 44, 81, 115
  broadband...10, 13, 22, 29–30, 43, 100, 102, 113, 115
  catastrophic failures...111
  coiled wire...12
  equal-delay...12
  isolation transformers...4, 27, 40, 66
  low-ratio...9, 13
  power ratings...111–113
  trifilar-wound...15
  technology of...27, 113–114
  with 1:1 ratios...3–4
  with 1:4 ratios...5–7
  with ratios greater than 1:4...15–16
  with ratios less than 1:4...9–13
  1:1...5
  1:4...5–7, 21–22
  1:9...21
  1:16...21
Turrin, Richard
  articles...27, 30–32, 36, 115
  Baluns...27, 31–34, 36

Ultimate multimatch...104–105
Ununs
  air-core...33
  bifilar...89, 101, 112
  broadband ratios...89, 92, 97, 99, 101
  broadband...45, 52, 73–74, 79, 89
  coiled...91
  dual-ratio...97–100
    1.5:1 and 3:1 ratios...98–99
    2:1 and 4:1 ratios...98
  for Beverage antennas...107–110
    low-power design...108–109
    high-power designs...109–110
  for 2-meter operation...86
  fractional ratio...86, 89–90
  high-impedance...85–86, 107, 110
  high-power...83, 109–110
  input-matching...107
  low-power...83, 95
  multimatch...101–107
  parallel transformer...100
  parallel-type...74–75
  quadrifilar...67, 90–91, 94, 98–99, 102, 104, 106, 108–109, 112
    tapped...102
  quintufilar...12, 16, 65–66 72–73, 81, 90–94, 96–99, 112
  step-down...67, 73, 86
  step-up...69, 72, 79, 90
  trifilar...10, 66, 68, 89-90, 94, 97, 102
    tapped...91, 101–102, 112
  1:1.33...16, 80–82
  1:1.5...22
  1:2...22, 62, 66
  1:3...81
  1.33:1...71, 89, 93–94
  1:4...5–7, 9, 11, 16, 20–22
  1.5:1...89–93
  1.56:1...65–66, 72, 74, 79, 91–92
    quintufilar...66, 91-93, 97–98, 102
  2:1...28, 89–95
  4:1...7, 30, 45-46, 68, 83–87, 97, 112
  5-ratio...102–103
    high-power version...103
    low-power version...103–104
  9:1...78–79
  10-ratio...102, 104–105
    high-power version...105–106
    low-power version...106
    quadrifilar...102
  50:12.5-ohm...84–85
  100:25-ohm...85
  200:50-ohm...85

V antennas...80
Varney, Louis...62, 115
Vertical antennas...89
  shunt-fed towers performing as...89

Wideband transformers...33
Wire antennas...55

Yagi antennas...27–28, 33, 65, 73–74, 86

$4.95

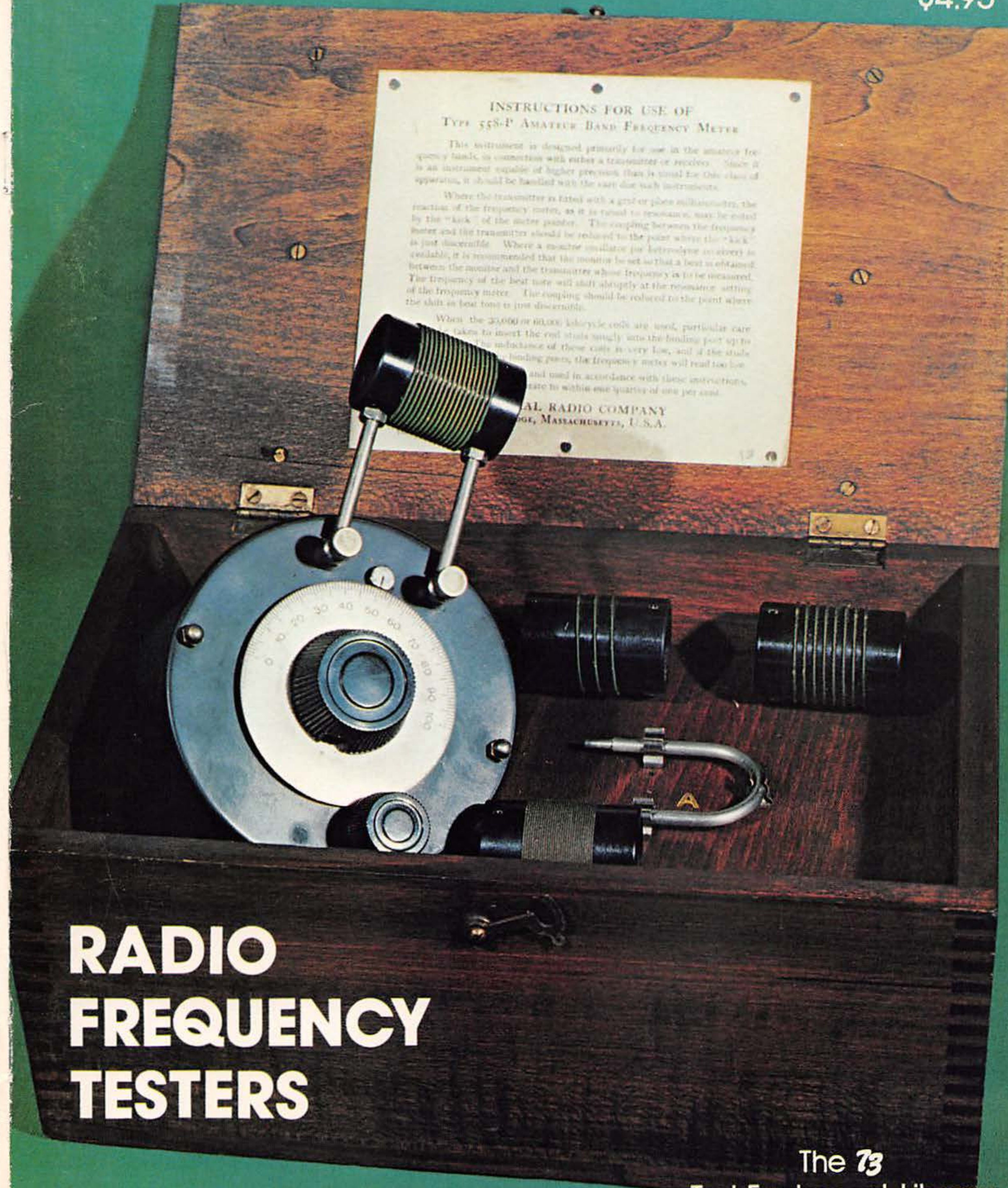
INSTRUCTIONS FOR USE OF
TYPE 558-P AMATEUR BAND FREQUENCY METER
AL RADIO COMPANY
GE, MASSACHUSETTS, U.S.A.

RADIO FREQUENCY TESTERS

77 Pieces of Test Equipment You Can Build!

The 73
Test Equipment Library
volume III

Table of Contents

The 73 Test Equipment Library

Volume III: Radio Frequency Testers

Chapter I: What is Your SWR, OM?

SWR Power Meter for 80 Thru 10m Bands W6VAT 5
An Etched Circuit SWR Bridge WA5SWD 6
Checking Your SWR Indicator W5EHC 8
Home Brew Bridge Calibration W4OAB 8
Mini SWR Bridge W2EEY 8
VSWR Supreme W4BRS 9
A Simple VHF SWR Meter W2EEY/K3EZ 10
Directional Coupler and VSWR Bridge for VHF and UHF WA6SXC 13
A Slotted Line for 1250 MHz WA8CHD 14
Some Directional Wattmeters and a Novel VSWR Meter G3PDM 16
SWR Bridges Using Zero-Center Meters W2EEY/K3EZ 18

Chapter II: How to Measure RF Impedance

The Antennascope – An Effective Tool VE3CEA 21
How to Build an Antennascope WA1CCH 22
The Mark III RF Impedance Bridge W6CRT 22
Measure RF Impedance with Your SWR Bridge W2AOO 26
Use Your GDO and Z Meter W0BMW 27

Chapter III: Measuring RF Power Output

Measuring RF Output WA6CPP 30
Light Bulbs as RF Indicators WB2GQY 30
Build Yourself a Light Wattmeter Henry 32
Low Power Light Wattmeter K1CLL 33
Yet Another Light Wattmeter K5CXN 34
A Low Cost RF Wattmeter WA3AJR 35
RF Power Measurement Using Hot Carrier Diodes W6AJF 36
VHF Dummy Load Wattmeter K9STH 38

Chapter IV: Measure Your Field Strength and Frequency

The RF Sniffer K5JKX 39
A Useful Accessory for the Ham Shack W2WHX 40
Two Thru Twenty FSM W7OE 40
An Amplified, Calibrated Signal Strength Meter VE3CES 41
Frequency Measuring Equipment at Microwave Frequencies WA9VFG 42
Frequency Meter-1 to 10 GHz Amateur Microwave K1CLL 43
Lambda Lines W2DXH 46

Chapter V: RF Signal Generators

Zero-Beating with a Frequency Meter WB4TBX 48
Hand Calibrate the BC-221 Frequency Meter Henry 48
High Accuracy VHF Frequency Measurements W5WGF 49
Transistorized LM Meter W5SOT 50
An I-F Spotter W5WGF 52
430-470 kHz Sweep Frequency Generator WA5SWD 53

An Armstrong Sweeper W2EMF 54
A New Broom K6JHJ 54
Low Cost Signal Source VE3GFW 55
Receiver Tweaker K6UAW 56
Citizen Band Alignment Aid WA5SWD 56
FET Chirper K6QKL 57
UHF Signal Source K1CLL 58
Crystal Controlled Signal Source with Infinite Attenuator for 144, 432 and 1296 MHz K1CLL 59
Low Cost 220 MHz Signal Generator K1CLL 60
UHF Signal Generator K6MIO 62
Lab Type IF/RF Sweep Generator Using ICs K4DHC 63
100 kHz Thin-Line Pulse Generator W2DXH 70

Chapter VI: Crystal Calibrators

All Band Band-Edge Marker WA2HRZ 74
The Multical K9VXL 74
100 kHz Marker Generator W7CJB 75
All Band Frequency Marker W1KNI 75
Poor Man's Universal Frequency Generator W2EEY 76
Calibrate that Calibrator W2KPE 78
A Simple Frequency Deviation Meter K6BJ 78

Chapter VII: GDOs

The Indicating Oscillator KH6AF 80
The Little Gate Dipper W5ETT 81
The Great Dipper WA0AYP 83
The Great Dipper-Modification WA0AYP 85
Etched Circuit UHF Dipmeter W1JJL 86
UHF Grid Dipper WA6BSO 87
Wide Range VHF-UHF Dipper K1CLL 91
Extra Services from Your Grid Dip Oscillator WA4UZM 94
GDO to Find C G3OGR 94
GDO Coil Extension WA0ABI 94

Chapter VIII: Noise Generators

Simple Noise Generator K9ONT 96
A Crystal Diode Noise Generator W5TEV 96
Temperature Limited Diode Noise Generator Thomas 97
Calibrated Noise Generator for 432 W6GXN 98
Another Way to Measure Noise Figure K5JKX 99

Chapter IX: Attenuators and Dummy Loads

Low Power Attenuators for the Amateur Bands WB6AIG/WA6RDZ 102
S-Unit Attenuator WA5SWD 103
The Ever-Useful T-Pad K5JKX 103
Combination Dummy Load/Attenuator Network W2EEY 104
The Minican W6JTT 106
The Tiny Terminator K5JKX 106

Chapter I

What is Your SWR, OM?

SWR POWER METER FOR 80 THRU 10m BANDS

James Lee W6VAT

With the declining sunspots, usage of the lower frequency bands is definitely increasing. The resulting increased QRM is helpful to no one and only efficient equipment and operating procedures will result in a maximum number of QSO's. The SWR/PWR meter described here won't make you a better operator, but it can help you be sure that you are delivering the most rf to your antenna from your rig. Fig. 1 shows a front view of the meter.

The basic circuit is a directional coupler switched to sample either forward or reverse voltage and a voltmeter to read this voltage.

This type of coupler has an output proportional to length, power, and frequency. The longer it is, the more output it gives. Since it puts a small impedance "bump" in the line, the length of the coupler should be limited to not over 1/20 wavelength at the highest frequency, or it may begin to contribute noticeably to the SWR itself. For a given power, if the frequency of the rf flowing through the coupler is reduced, the maximum coupler output is reduced. This means that you can get full scale readings on 10 M with a lot less power than on 80 M. The meter described here gives half scale deflection on 40 M, at maximum sensitivity, with about 350 watts of forward power. If your rig is a KW this meter will fill the requirements for SWR/PWR measurements nicely. It can be left in the line at all times to monitor SWR or Power delivered to the antenna or other load. If desired, a Barker and Williamson type 551A coaxial switch can be used to insert the meter in the line for test purposes and then switch it out during operating periods.

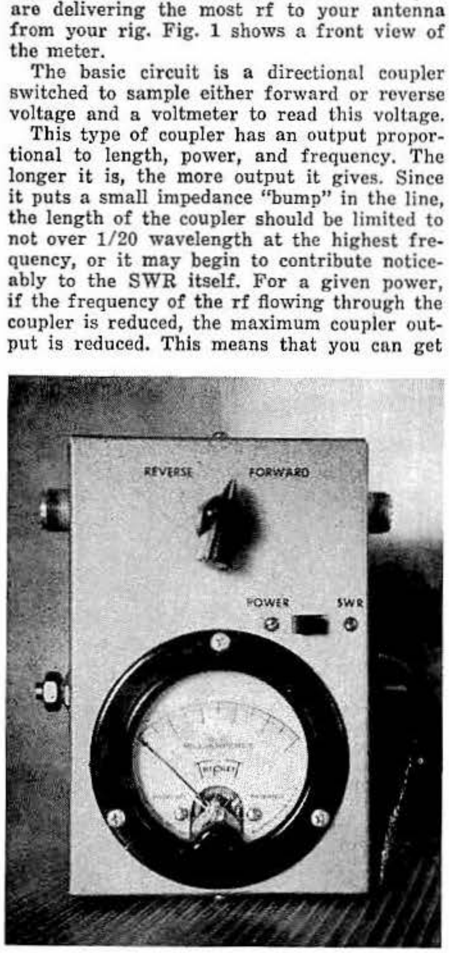

Construction

The unit is built in a gray hammertone LMB type 141 box. The dimensions of this box are 3" x 4" x 6". Fig. 2 shows the parts placement and should answer any questions concerning layout.

The coax directional coupler is made from a 14" length of RG-8/U. The outer covering is slit lengthwise with a knife and peeled off. Take care here not to cut into the woven braid. The woven braid is then bunched toward the center to loosen it. Next, a length of #22 enameled wire is passed through the braid at about 2¼" from one end and run under the braid next to the inner insulation. It is brought out at the other end, again 2¼" from the end of the shield braid. When this is done, smooth the braid back to its original position carefully to avoid scratching the enamel on the #22 wire. The #22 enameled wire should lay as straight as possible under the braid. It should have no slack nor should it twist around the inner insulation to any great degree. The ends of the shield braid are trimmed back far enough to be soldered to Amphenol 83-1H hoods. The inner insulation is trimmed off so as to expose about 3/16" of the inner conductor. The inner conductors are then soldered to Amphenol 83-1R female type chassis mount coax connectors. The coupler may now be set aside and the rest of the meter constructed.

Potentiometer R_1 is mounted on 1/16" x 1" x 4" piece of bakelite and positioned approximately in the center of the box to minimize capacity to ground. The metal rear cover is removed from the pot for the same reason. The pot ground lead is to the center of the coax and is wire in last. The diode D_1 should be protected from damage by heat when soldering it in the circuit. Long nose pliers gripping the leads near the body of the diode are satisfactory. The directional coupler is the last item soldered in place. As can be seen in the photos—the coax connectors are mounted from the inside of the box. When the #22 enameled leads are soldered to S_1 they should be the same length, and again, take care not to scratch the enamel or a short circuit may occur and you'll have to do it all over again.

Calibration

There are any number of ways to calibrate the SWR/PWR meter, but the way most hams will use is their own rig and a suitable dummy load. Though very limited in power, a 2 watt 50 ohm resistor such as made by Ohmite, IRC and others, mounted inside an Amphenol 83-1SP male coax plug makes a very good dummy load. Although not completely non-inductive, this dummy load is far superior to such real unknowns as light bulbs, electric iron heating elements, etc. This particular load is 50 ohms shunted by 6 mmfd over the range of 3-30 mc. It is a good dummy load—though a low power one.

With the back cover off, attach the dummy load to J_2, the load jack. Set R_1, the Sensitivity Control, to maximum and S_1 to FORWARD. Next apply power by hooking your rig (or

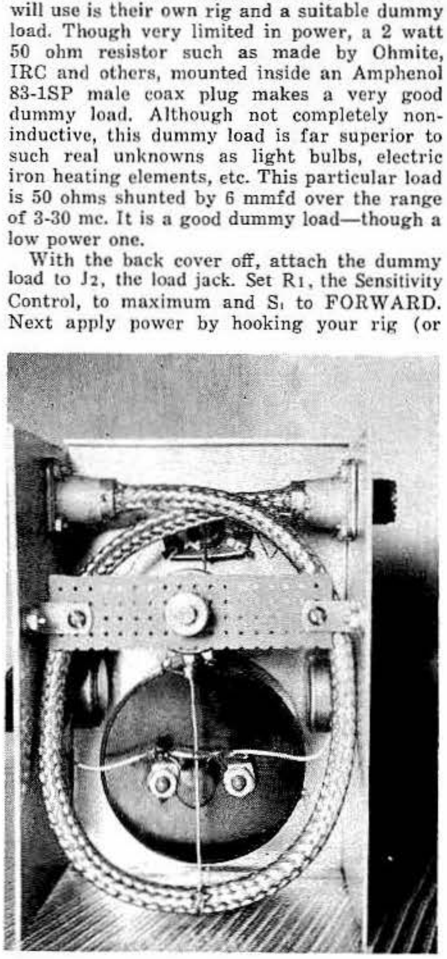

other rf source) to J_1, the transmitter jack. Make sure that the rf applied is 28-30 mc, or, the highest frequency you operate in the 3-30 mc region. This meter is sensitive to frequency. Stray capacities and other unbalances will have their greatest effect at the highest frequencies. In any event, calibrate the unit on the highest frequency your rig will tune in the 3-30 mc region. With power applied, set R_2 so the meter reading is at least half scale and switch S_1 to REVERSE. This will result in a lower meter reading. R_1 should then be adjusted for a minimum reading on the meter. Using the suggested load, you will not get a complete null, but the null should not be much more than 50 ua for half scale deflection in the FORWARD position on S_1. Don't put too much power into the dummy load since excessive dissipation can ruin it and change its characteristics greatly. If you have a higher power dummy load whose characteristics you know accurately, by all means use it, but remember a light bulb is not a good load. Once nulled, lock the nut on R_1 taking care not to disturb the setting. Replace the back cover and using the dummy load, recheck the null to make certain that it has not shifted.

Using The Meter

To make SWR measurements you need only insert it in the line and set S_1 to FORWARD and S_2 to SWR. Adjust the sensitivity control for at least a half scale reading. Then switch S_1 to REVERSE and read the value. SWR is then calculated by the following:

$$SWR = \frac{I_{fwd} + I_{rev}}{I_{fwd} - I_{rev}}$$

For example:

let $I_{fwd} = 500$ ua and $I_{rev} = 50$ ua, then

$$SWR = \frac{500 + 50}{500 - 50} = \frac{550}{450} = 1.22:1$$

The meter is useful for tune-up purposes where exact SWR is not needed. Just keep the FORWARD reading at a constant value and tune for minimum REVERSE readings. The exact SWR can be calculated when you have found the lowest REVERSE position.

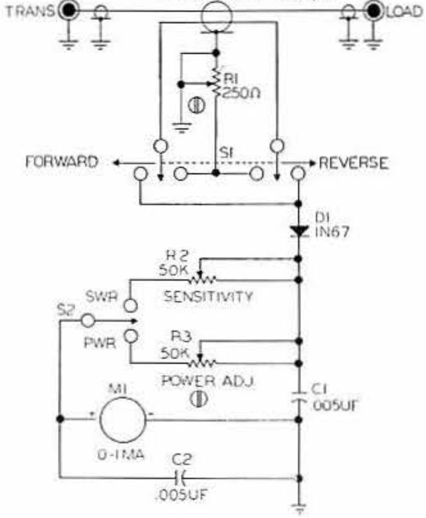

The POWER position may appear to be essentially the same as the SWR position and it is. When measuring power into a load of known fixed value you only need to know the voltage across (or the current through) the load. The POWER position on S_2 is used with S_1 set to FORWARD. R_3 is merely set to a scale reading that is convenient for all bands if only a relative reading is used. If a good 50 ohm high power dummy load is available, you may make accurate calibrations by using a VTVM plus high frequency detector probe and measuring the actual voltage across the load. Just put a T-connector on the load jack J_2. Put the dummy load on one arm of the "T" and read the voltage at the other arm. 200 volts across a pure resistive load of 50 ohms is equal to 800 watts of power. If your rig delivers a key down 800 watts to a load then you could set R_3 at 0.8 ma on the scale, etc. The scale will not be precisely linear, particularly at low powers, but if enough resistance is used at R_3 the effects should be minimum. Since the coupler voltage is frequency sensitive one setting of R_3 will not hold for all bands. If desired, R_3 could be replaced with a switch and a number of selected resistors (one for each band). This would keep the scale factor constant between bands. This could be done for R_2 as well, but complicates an otherwise simple device

Operation

Once calibrated, the meter is very simple to use. Just hook it in the line at some convenient point and apply power. With S_1 set to FORWARD and S_2 set to SWR, adjust R_2 for at least a half scale reading. Switch S_1 to REVERSE and you will then be able to continuously monitor your reverse power. For power measurements, set S_1 to FORWARD and S_2 to PWR. Apply power and the meter will be monitoring your forward power continuously. R_3 is a screw driver adjusted pot. It can be set, and locked at a point which allows operation on all the bands you operate. The scale reading on each band will be different, but once set, these readings can be jotted down in your log book and any change quickly noted. This meter is also an excellent device for the antenna "tinkerer," or if you have coax coupling between stages in your rig it can be used to provide proper power transfer. Best of all, it is not expensive nor difficult to construct. Try one and you won't be without one again.

AN ETCHED CIRCUIT SWR BRIDGE

Ed Lawrence WA5SWD

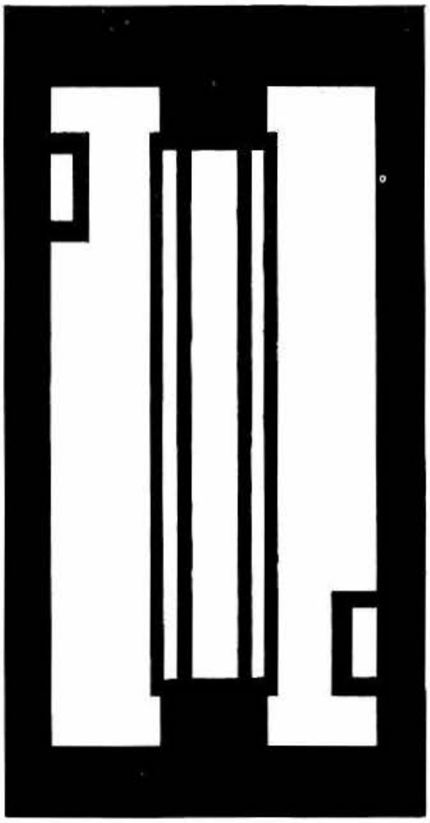

Several of these boards have been etched and units assembled, and all have been satisfactory. I have used mine, with a 200 microamp meter, from a kw on the low bands up to a "Twoer," with no problems with sensitivity.

Since Bud Miniboxes are commonly available, I scaled this board to fit into the Bud

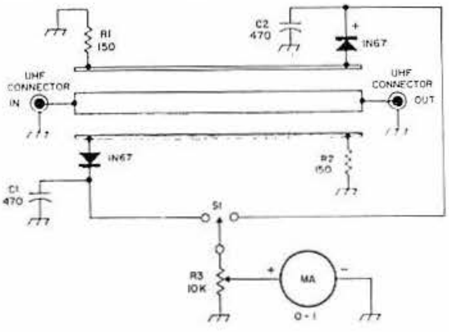

Schematic diagram of SWR bridge.

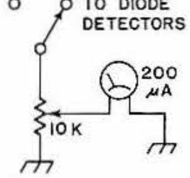

Fig. 1. Metering circuit for the etched circuit SWR bridge.

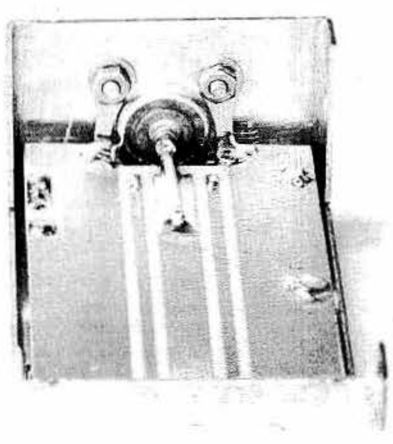

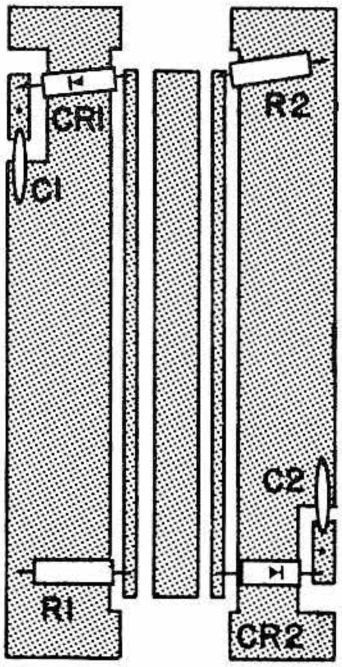

Fig. 2. Layout for the printed circuit board.

Minibox CU 2102A or CU 3002A. If you want to include the metering in the same package, the CU 2103A or CU 3003A should be used to give more room.

If you use the CU 2102A, center the coax connectors on the ends. If the CU 2103A, mount the connectors .8 inches from the open end. I used solder lugs bent at right angles to mount the P. C. Board to the chassis. The photograph shows the mounting much better than 10,000 words.

The meter sensitivity control circuit shown has a wider control range than the one shown in the *ARRL Handbook*, since the pot shunts the meter at low settings. This action could be accentuated by putting a fixed resistance in series with the meter movement, at the expense of sensitivity.

Although the virtues of operating transmission lines with low standing wave ratios (SWR) have been discussed many times in the past, evidently the economics of maintaining low SWR's are not readily apparent, particularly if the frequency of operation is low and the transmission line short. This has been reflected in various pseudo-technical QSO's where many have been led to the utter disregard for standing wave ratios. Most members of the amateur fraternity exist on limited budgets at best and when a significant portion of that precious transmitted power is eaten up by transmission line losses and misinterpreted standing wave ratios for naught, something should be done.

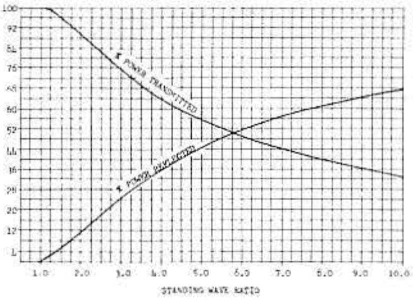

SWR vs % Power Transmitted and Power Reflected

Figure 1

A look at the graph in Fig. 1 will show you the percentage of power reflected for various standing wave ratios. For instance, if you are presently tolerating an SWR of about 5.8:1 (not uncommon in many ham shacks), 50% of the power which reaches the antenna is actually reflected back down the transmission line, heating up the final tank and causing TVI. Nor is only the transmitted signal effected, a high SWR will similarly degrade the received signal. This is particularly important in the reception of the extremely low level signals often encountered in DX and VHF operating. Stereophonic buffs should take heed too. A recent report by the IEEE (Institute of Electronic and Electrical Engineers) Professional Group on Broadcasting noted that a high SWR on receiver antenna inputs causes a reduction of stereo quality.

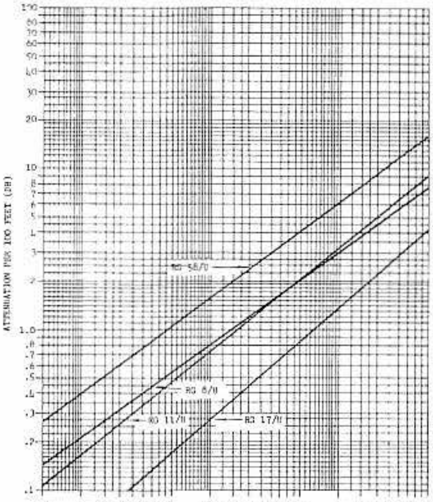

Figure 2

Attenuation vs frequency.

Many of the amateur stations on the air today make use of RG8A/U coaxial cable. Its excellence is proven out by its extensive use by the military, but a look at the loss graph (Fig. 2) for this cable indicates that it is not completely lossless! Even at 4 mc it has approximately 0.3 db loss per 100 feet, and on six meters there is a loss of 1.4 db for the same length. A look at Fig. 3 indicates that on 75 meters only 94% of the transmitted power is delivered to the antenna if 100 feet of RG8A/U is in use. At 50 mc the loss has sky rocketed to 26% for the same length of line. However, there is one big hooker for these conditions to exist: the SWR must be 1:1. For any other value of SWR there will be further line losses as shown in Fig. 4 because standing waves have the property of multiplying attenuation. This graph indicates that if a transmission line is operating at an SWR of 3.7:1, the line loss will be multiplied by a factor of two. For the previously mentioned situation on 50 mc, an additional 24% loss could be expected with an RG8A/U line operating at an SWR of 3.7:1.

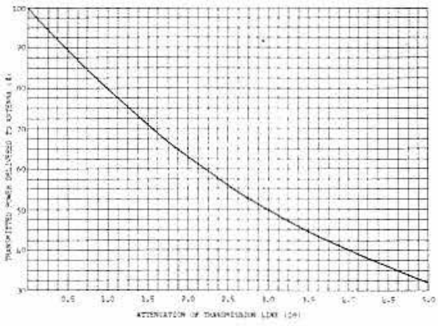

Figure 3

Attenuation vs Power transmitted.

It should be obvious by now that the use of an SWR bridge in the line at all times is very advantageous in the maintenance of a low SWR at the operating frequency. However, contrary to popular belief, the SWR bridge does not tell all. Since there is loss or attenuation in any length of transmission line, the reflected wave will be attenuated in the same matter as the transmitted or incident signal. Because the standing wave ratio is the ratio of the incident wave to the reflected wave, attenuation of the reflected wave will give erroneous SWR measurements when the SWR bridge is conveniently located at the transmitter. In this location the bridge will see the full power of the transmitter, but only a portion of the reflected signal. In some cases where the length of the transmission line is excessively long, the reflected wave will be attenuated to such a degree that the SWR will appear to be very close to 1:1, while in reality it will be a good deal higher. This fact is graphically represented in Fig. 5.

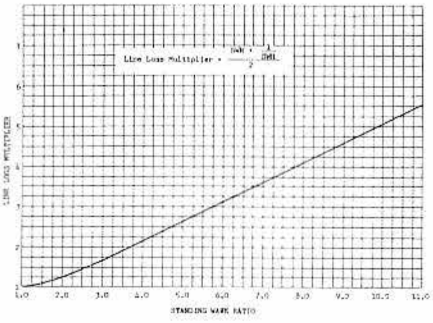

Figure 4

Line loss multiplier vs SWR.

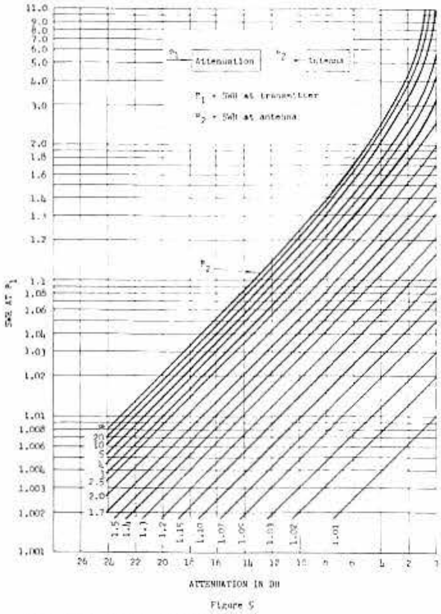

Figure 5

SWR vs attenuation.

For example, 143 feet of RG8A/U at 50 mc would result in approximately 2 db attenuation. If an SWR bridge inserted in the line at the transmitter indicated an SWR of 2:1, this graph shows that an SWR of 3:1 exists at the antenna. A look at Figs. 3 and 4 will indicate that a 3.3 db loss (2 db times 1.65 multiplier) occurs, amounting to 47% loss of transmitted power in transmission line losses. Of the remaining 53% power arriving at the antenna, 24% will be reflected back down the line. A little simple arithmetic will show that of the total power transmitted, only 30% will be radiated! This simple mathematical fact should make the merits of low standing wave ratios immediately obvious if we wish to get the most out of our equipment. By keeping transmission lines short and by insuring that the SWR is as close to 1:1 as practicable, line losses will be minimized, maximum power will be delivered to the antenna and more successful and reliable radio communications will result.

CHECKING YOUR SWR INDICATOR

Carl Drumeller W5EHC

Many articles have been published on how to build and even some on how to calibrate a VSWR indicator. The calibration instructions usually tell you to terminate the indicator's output with a purely-resistive 52-ohm load and then to adjust the device so that a maximum forward and minimum reflected meter deflections are obtained. Sometimes they'll go further and tell you to reverse the device and recheck for opposite indications.

This is all very well. It assures you that the VSWR indicator will be telling you the truth when it says "All's well!" while looking into an utterly-flat transmission line. It doesn't tell you a thing about what the indicator will have to say when it gets tangled up with a line that has a wildly-mismatched termination.

As most transmission lines, in actual practice, are terminated in loads which are not only mismatched in the matter of resistance but also in the inclusion of a considerable magnitude of reactance, it would be well to explore the indications you'll get under realistic circumstances. After all, these are the situations under which you'd want to take corrective steps. Accurate indications of undesired conditions, therefore, are imperative if intelligent remedial actions are to be taken.

Fortunately, some quite enlightening tests are made easily. All you'll need are some lengths of coax transmission line (the same as you're using in your antenna feedline) equipped with male fittings at each end and a few female-to-female junctions. Select the frequency at which you want to make the test. Usually it's wise to make the test on the highest frequency band you plan to use. With this in mind, make up three one-eighth wavelength sections of transmission line and mount the male fittings on the ends of each section.

If your antenna presents an unmatched load to your transmission line, you may elect to skip over this paragraph and go directly to the next one. If it does not (Ah, you dreamer!), you'll need another piece of transmission line. It should be fairly long, perhaps a half wavelength. Put a male fitting at one end and attach a termination, which is deliberately made to be a sad mismatch, at the other end. Don't just mismatch it by using too high or too low a value of resistance. Throw in some reactance, too! You might use a resistor with an inductor in series. Or, you might try a capacitor in series with the resistor. Or, you could use either an inductor or a capacitor in parallel with a resistor. In fact, it would be best to experiment with all four!

Now that you have a transmission line available that you know is mismatched, you're ready to start the test. The first check (the "control", you might call it) is made with everything normal. That is, you'll have the transmitter feeding directly into the VSWR indicator's transfer box and the transmission line (either the one to your antenna or the substitute line to the mismatched load) attached to the output of the transfer box. Note the VSWR indicated. Also note your transmitter; insure that it's tuned to resonance and is adjusted to a power you can maintain throughout the first portion of the tests. Jot down these indications. Now, insert an eighth-wave section between the transfer box and the transmission line and without making *any* other changes or adjustments, note the VSWR. Repeat these steps, adding an additional eighth-wave section each time until you've used all three. Did you detect any change in VSWR? If there was even the slightest change, your VSWR indicator is not trustworthy!

Now for two more checks. Try varying the transmitter power output. Does this have any effect upon the indicated VSWR? If it does, your VSWR indicator is not trustworthy! Then try varying the transmitter output tuning, deliberately throwing the stage out of resonance. Does this have any effect upon the indicated VSWR? If it does, your VSWR indicator is not trustworthy!

Few VSWR indicators under the $150 class will pass these basic tests. If yours doesn't, don't be perturbed. You have an instrument that still has a useful field of application. You can use it as a comparative indicator. For instance, if you're adjusting the gamma match at an antenna, it'll serve quite well; in this application, you're holding all of the significant variables constant, with the exception of one (the gamma match), the effect of which you want to observe. Your tests will have shown you the parameters you'll have to hold constant for any other than simple comparisons. In all probability, you will have found that measurements taken with different (electrical) lengths of transmission line are invalid. Also, it's probable that, owing to the non-congruity of diode curves, measurements will have to be taken at precisely the same level of rf power if accurate comparisons are to be made.

If you'll keep its very real limitations in mind, you'll find that even an inexpensive VSWR indicator has excellent potentials for useful measurements. But don't ask it to perform at levels that even its expensive siblings can't attain!

HOME BREW BRIDGE CALIBRATION

Clifford Honess W4OAB

This table is useful for all homebrew SWR bridges and gives the % reflected power and the % full scale reading in the reflected mode, when the meter is set at full scale in the forward mode.

SWR	% Pwr. Refl.	Refl. Rdg. in % of Full Scale
1.1	.2%	4.8%
1.2	.8	9.1
1.3	1.7	13.1
1.4	2.8	16.7
1.5	4.0	20.0
1.6	5.3	23.1
1.7	6.7	25.9
1.8	8.2	28.6
1.9	9.7	31.1
2.0	11.0	33.3
2.1	12.6	35.5
2.2	14.0	37.5
2.3	15.5	39.4
2.4	17.0	41.2
2.5	18.4	42.9
2.6	19.7	44.4
2.7	21.1	45.9
2.8	22.5	47.4
2.9	23.8	48.7
3.0	25.0	50.0
4.0	36.0	60.0
5.7	49.0	70.0
9.0	64.0	80.0
19.0	81.0	90.0

MINI SWR BRIDGE

John Schultz W2EEY

Most SWR meters today are of the coupled variety which can be left in a transmission line while a transmitter is operated at full power. However, for a number of prolonged tune-up operations, involving antenna matching systems, for instance, such couplers have several disadvantages.

On 160 and 80 meters, especially, a reasonable amount of power is necessary to produce full deflection—up to 100 watts with some configurations. With a very low-powered transmitter, making adjustments at this power level certainly may damage the output tube or tank-circuit components with a high SWR. This will not be the case with higher-powered circuits but, in any case, a signal strong enough to cause needless QRM will be radiated.

Another disadvantage of the coupled SWR meter, if it is home constructed, is that it must be carefully calibrated since its response is very dependent upon the mechanical configuration of the coupling circuit. This is unlike the bridge-type SWR meter (described in this article) where a standard SWR curve may be used with a good degree of accuracy.

The above factors, plus the fact that I didn't need an SWR meter continuously in the transmission line, led me to construct the little resistance type SWR bridge shown in the photograph. It is just about as simple and inexpensive a unit as can possible be built.

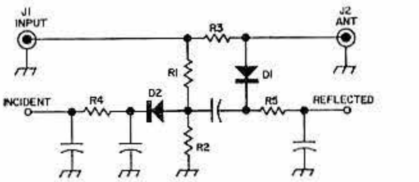

Fig. 1. SWR Bridge Circuit. See text for values of R_1, R_2 and R_3. R_4 and R_5 may be any matched value from 10 k to 47 k. D_1 and D_2 are IN34, IN54, or similar types. All capacitors are disc ceramic, .005 MF, 100V.

Construction

Two SO-239 coax chassis connectors are joined back to back by two 1¼ inch threaded hex spacers. The two four-lug terminal strips are mounted at the ends of one of the spacers. The wiring of diodes D_1 and D_2 as shown in **Fig. 1**, should be such that the incidental voltage-measuring point appears on the terminal strip mounted on the "input" SO-239 connector, in order to avoid confusion in measurement. Short leads, of course, should be used but hardly anything else is possible with only 1¼ inches between connectors.

Some attention must be paid to the components used if accurate readings are to be obtained. Resistor R_3 must closely match the impedance of the coaxial line used (52 or 75 ohms). For 52-ohm lines, a suitable resistor (within ½ to 1 ohm) can usually be found from a group of standard 10% tolerance, 47-ohm resistors; and for 75-ohm lines, from a group of 68-ohm resistors. Resistors R_1 and R_2 can have any value from about 30 to 100 ohms, but it is important that they are as closely matched as possible. One trick which may be used to affect very small resistance changes is to file "V" notches in a composition resistor to raise its resistance. Two-watt units are suggested for these resistors because of their longer-term stability and endurance in case too much input power is applied.

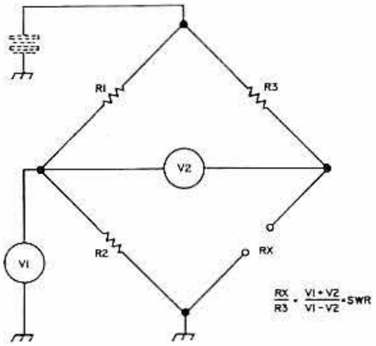

Fig. 2. Simplified diagram of SWR Bridge. V_1 represents incident voltage and V_2 the reflected voltage.

Resistors R_4 and R_5 serve as linearizing resistors so that almost any meter with a basic movement of 1 mA or less can be used as an indicator. The lower dc voltage ranges on almost any VOM will work fine. These resistors as well as diodes D_1 and D_2 should be checked to see that they match reasonably well (the resistors within a few precent and the diodes within a few percent for their forward and reverse resistance readings).

Calibration

There are really no adjustments that can be made to the bridge, and calibration really consists of checking the balance. **Fig. 2** is the dc circuit of the bridge (a simple Wheatstone bridge with resistance arms). If the balance of the bridge is good, V_2 should be the same when points R_X are opened or shorted so long as V_1 is held constant. This can be checked on the actual bridge by applying an input at the highest frequency of interest (6 or 10 meters), shorting J_2, and checking that incident and reflecting voltages are the same. The same is done with J_2 open. If the voltages are not equal, the difference can be taken as an indication of how accurate the SWR readings will be. If the difference is too great, R_1 or R_2 will have to be changed for a better match or the mounting of the components changed to reduce stray couplings.

A further check is to connect a known 52 or 75 ohm resistor across J_2. The reflected voltage should, of course, read zero.

Operation

As noted in **Fig. 2** the actual SWR is a simple function of the incident and reflected voltage readings. **Fig. 3** presents this function in graphical form. The incident voltage is simply adjusted for some convenient value, say 10 volts reading on the dc scale of a VOM (possible with most SSB transmitters by adjusting the carrier balance control with no audio input). The reflected voltage is then read as a percentage of the incident voltage and the SWR found from **Fig. 3.** The input power required to operate the bridge is essentially independent of frequency, being about 1-2 watts maximum.

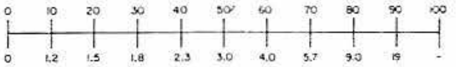

Fig. 3. SWR values for selected reflected voltage readings taken as % of incident voltage reading.

It should be remembered that such a bridge can measure only the resistive portion of an impedance. When using it to adjust a circuit, if a SWR minimum null but not a zero reading for reflected voltage can be obtained, it indicates some reactive component must still be present.

VSWR SUPREME

E. L. Klein W4BRS

One of the most valuable tools used by the amateur is the Voltage Standing Wave Ratio Meter. It ranks with the grid-dipper and the plate current meter as an indispensable instrument around the ham shack. We know that the VSWR meter is very handy in indicating relative power output when tuning a transmitter, particularly when the plate dip is not too discernable. It is most useful, however, in proving that the last available watt has reached the antenna where it can do some good.

Why a good match?

Although a good copper connection is made all the way to the antenna, an efficient transfer of power may not be achieved because of a mismatch between the characteristic impedances of the various portions of the transmission system. The interesting thing is that the "match" is different for each frequency because the antenna is essentially a single frequency device.

We can appreciate the importance of a proper match between the transmitter and the antenna when we are told, for example, that a VSWR of 3 to 1 causes a power loss of nearly 3 dB for 200 feet of RG-8/U coaxial line at 30 MHz. The table below provides

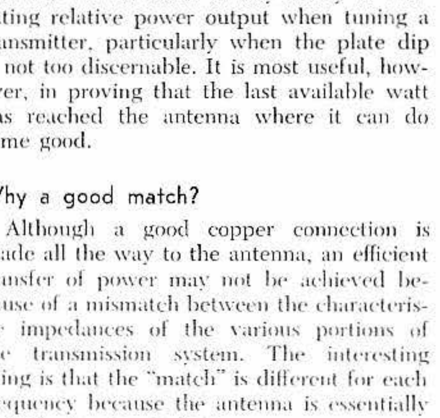

Fig. 1 Schematic of basic VSWR Meter and associated switching circuits.

Table I

VSWR	Power Loss (dB)	Transmitter Power Needed to Provide 1 kW at the Antenna
1.5:1	2.1	1600 watts
2:1	2.3	1700
3:1	2.8	1900
4:1	3.3	2000
5:1	3.7	2300
7:1	4.5	2800
10:1	5.3	3400

Additional power needed to compensate for a poor impedance match between transmitter and antenna. Figures are based upon 200 feet of RG-8/U cable at 30 MHz.

the real reason why we should be concerned with the impedance match. Notice how much the transmitter power would have to be increased to make up for a poor match between the transmitter and the antenna. Incidently, this match involves each and every part of the total transmission system including connectors, antenna relay, low-pass filter, balun, etc., as well as the transmission line itself and that particularly critical point at which it is connected to the antenna.

A new approach

Most VSWR meters today are an external accessory to the transmitter. But this practice is not good. Coaxial connectors are expensive and cause unwarranted mismatch and power loss. Meter faces usually end up behind the transmitter or in some other inaccessible location. When switching from forward to reverse, the little accessory box scoots across the table leaving scratches and a distraught operator.

The transmitter plate current meter is no longer a plug-in accessory. Why should the VSWR meter be? (Believe it or not, plate meters used to be plugged in with phone jacks.) Using the simple design described here, the home constructor as well as the commercial manufacturer can now build the VSWR meter into the transmitter in the smallest possible space and at only pennies of cost.

The circuit

Nothing is new about the circuit. It has been adequately described in the past in magazine articles and handbooks. However, for the convenience of the reader, the VSWR meter circuit is reproduced in **Fig. 1** for handy reference. Terminating resistors R1 and R2 should be 33 ohms for a 50-ohm transmission line when the physical configuration, as shown here, is used. One-half watt or smaller size resistors may be used. Diodes CR1 and CR2 are any matched pair of silicon diodes or germanium. The types which are enclosed in glass cases are the easiest to use because of their small size. The ohmmeter can be used to select and match the diodes of the ten-cent surplus variety found in advertisements in ham magazines such as *73*. Bypass capacitors C1 and C2 are 1500 pF Centralab type FT-1500.

Physical components

Parts used in this VSWR meter are illustrated in **Fig. 2.** The brass tubing is about 5 to 7 inches long and of 5/16 inch outside diameter. This size tubing fits snugly around the inner polyethelyne insulation from RG-8/U coaxial cable. About 10 inches of coax is stripped of its outer jacket and braid. The inner insulation is trimmed to extend 1/8 inch past each end of the brass tubing. Two large solder lugs are selected to fit over the 1/4 inch threaded shank of the bypass capacitors. These lugs should be of the long variety so they may be shaped and soldered to the brass tubing as shown in **Fig. 3.** Two 8-inch pieces of #22 enameled copper wire are also required.

Assembly

After soldering the lugs to the brass tubing about 5/8 inch in from each end, the bypass capacitors are assembled to the lugs. Place several fiber washers under the ring nuts prior to tightening them down on the threaded shank of the capacitors. This permits careful soldering of the capacitors to the lugs without danger of also soldering the nuts in place.

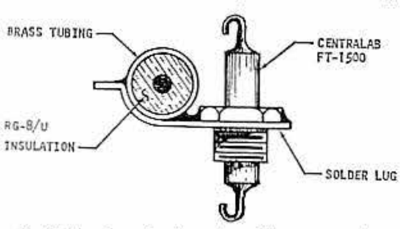

Fig. 3 Solder lug is shaped to fit one-quarter way around the brass tubing.

Two small grooves are now cut 180 degrees apart for the total length of the polyethelyne insulation. A small wood carving gouge or carefully manipulated razor blade can be used for this purpose. These grooves provide a space for the enameled copper wire which is held in place when assembling as shown in **Fig. 4.** Prior to this operation, the wire should be stretched and work-hardened by jerking it between two pairs of pliers. Be sure that the plane described by the two wires lies at right angles to the chassis on which the unit is mounted. This permits all resistors and diodes to have equal lead lengths.

When the inner assembly has been tugged and shoved into place within the brass tubing, the #22 wire ends are trimmed, stripped and soldered to their respective resistors and diodes. Much care should be exercised at this point to prevent melting the insulation or damaging the near zero-length component leads.

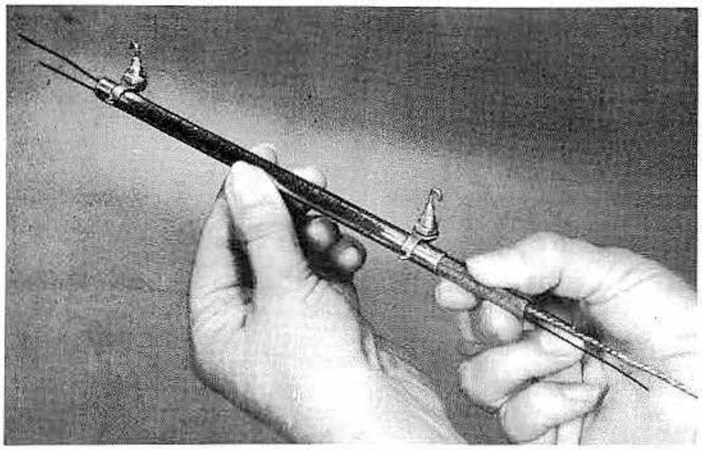

Fig. 4 Assembling the inner components into the brass tubing.

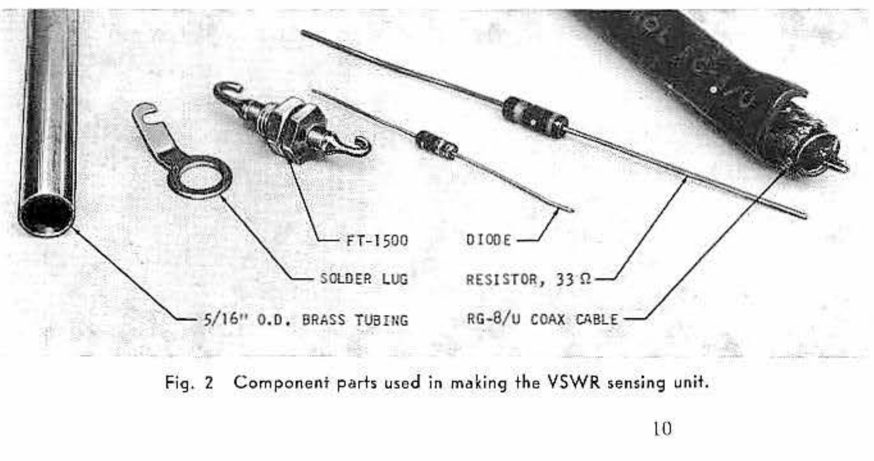

Fig. 2 Component parts used in making the VSWR sensing unit.

Application

A completed sensing unit for the VSWR meter is shown mounted on a typical chassis in **Fig. 5.** It will be noted that no conventional box or housing is used because the total outside of the unit is at ground rf and dc potential, save for the component connections at each end. By mounting the bypass capacitor in the chassis, the low-voltage rectified current fed to the meter switch is

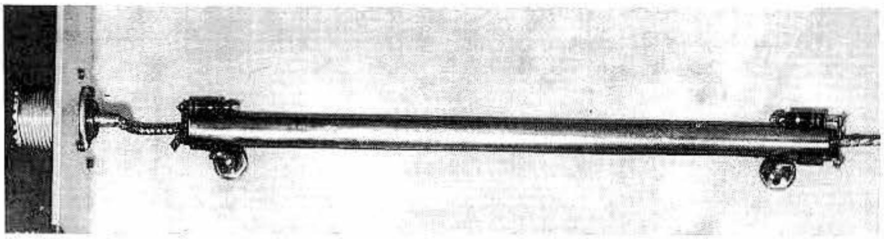

Fig. 5 Finished VSWR sensing unit mounted on a typical chassis.

isolated from high-power rf on the other side of the chassis. It can readily be seen that the finished sensing unit occupies no more space than would be used by a coaxial lead running from an antenna relay to the antenna connector on the chassis.

A further refinement is shown in **Fig. 6.** Complete isolation of the high-power rf is provided by the coaxial hood. Impedance discontinuity is also minimized by use of the hood, which was designed for this purpose and is readily available.

Length of sensor

The dimensions given for the length of the sensor element, including its outer tubing and inner conductor, are not critical. They are, however, directly related to the power of the transmitter with which the VSWR meter is used. For example, with a one kilowatt high-frequency CW transmitter and a 0-1 milliammeter as the indicating meter, the length of the sensor can be as short as 2-3 inches. A sensor which is constructed approximately 7 inches long, as illustrated in this article, will work fine with the same meter on a 25-200 watt high frequency transmitter. If meters with higher current ratings are used, a longer sensor is required, and, conversely, a more sensitive meter would provide adequate full-scale deflection with a shorter sensor element. Obviously, it is impracticable to vary the length of the sensor element in order to vary the sensitivity of the VSWR meter as a whole. It is for this reason that the adjusting resistor is provided in series with the meter. For VHF use, the sensor can be shorter.

All that has been said above can be depicted graphically. **Fig. 7** shows the generalized relationship between the sensor length and transmitter power with which it is used. The two curves represent different meter sensitivities. A 0-1 milliammeter offers a good compromise. With a 5-7 inch long sensor and the proper series resistor, all powers normally encountered in amateur work can be handled. However, if space requirements so dictate, a shorter sensor unit can be employed with some small sacrifice in accuracy.

Terminating resistor

Small variations in mechanical construction and lead dress will have an effect on the value of the terminating resistors, R_1 and R_2. Also, a carbon resistor does not display the same reactance at high frequencies as its measured resistance at dc. The value of the 33-ohm resistor was therefore determined empirically.

To verify the proper value of the terminating resistors, the test set-up shown in **Fig. 8** is used. A radio-frequency source of approximately 10 to 20 watts is required. A transmitter exciter stage operating on the 10-meter band is preferred for this purpose. Ten meters, or even fifteen meters, will provide better accuracy than one of the lower frequency bands. A dummy load is also required. This load must be capable of dissipating the power of the radio-frequency source used in making the test. Three or four 2-watt carbon resistors of the proper value in parallel to provide 52 ohms will suffice if the power is not left on continuously.

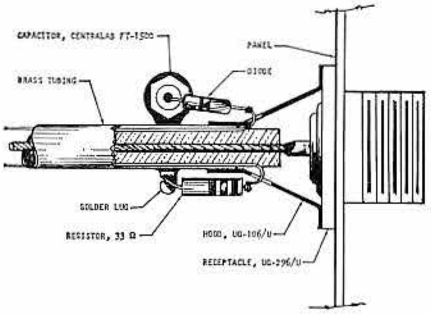

Fig. 6 Recommended chassis connection for output of VSWR meter sensing unit.

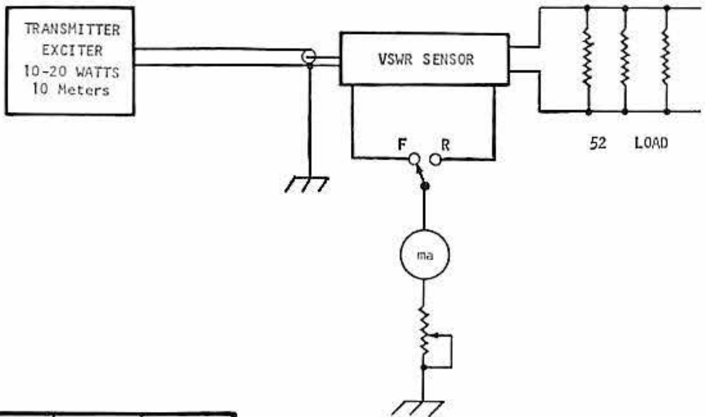

Fig. 8 Test set-up for verifying the proper value of the terminating resistors which are a part of the sensor unit. The value of the dummy load should match the characteristic impedance of the sensor unit and have a total wattage rating nearly equal to the source power.

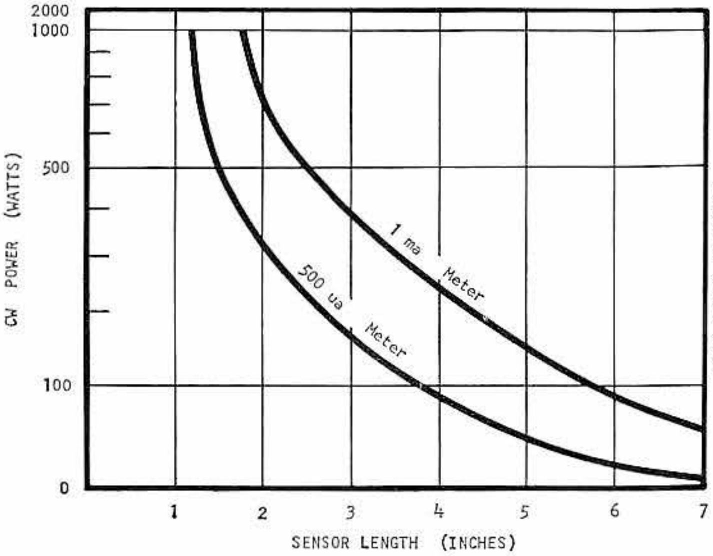

Fig. 7 The approximate relationship between sensor length and transmitter power is shown for two commonly used meter movements. Other meter values may be used as discussed in the text.

In making the test, the selector switch is first placed in the "forward" position. With power applied, immediately adjust the sensitivity control so that the meter reads full scale. Upon switching to the "reverse" position, the meter should read near zero and be at or below the 1:1 calibration point on the meter scale. Several resistors may be substituted until the proper value is found. The important thing to remember is that both of the resistors should be simultaneously substituted and that they must be as near identical as possible as measured on a reasonably good ohmeter. Lead lengths should also be as short as possible and of identical length.

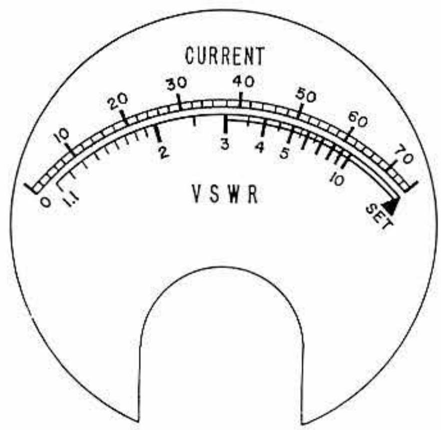

Fig. 9 Full-size photomaster of the VSWR meter dial. A multi-purpose meter was used in the author's transmitter so that the grid and plate currents could also be read on the upper scale.

The dial scale

Using the standard formula for calculating VSWR, it is possible to calibrate the meter face as follows:

$$\text{VSWR} = \frac{\text{forward} + \text{reverse}}{\text{forward} - \text{reverse}}$$

Fig. 9 is a full-scale illustration of a meter face used with the *VSWR Supreme.* This scale fits the Triplett Model 327, as well as a number of other meters of the same size category. A word of caution—don't assume that the scale calibration, or linearity will be the same for all makes of meters. The individual meter movement selected should be checked by using the above formula and marking off radials representing 4-5 different VSWR values. With the scale from your meter at the center of an oversized radial(s) drawing, it is possible to verify the angular placement of each VSWR calibration point.

The *VSWR Supreme* is truly a novel approach to an old standby. Using the construction methods outlined in this article, it is possible to fabricate the sensor unit so that it occupies the smallest possible space. This sensor can now be built into a transmitter and take up no more room than the coaxial lead which it replaces.

A SIMPLE VHF SWR METER

John Schultz W2EEY/K3EZ

I recently wanted to do some work on a 2 meter antenna and since no other instrument was handy, started to use an SWR meter manufactured for use on the high frequency bands. After some erratic results, it was decided to check the SWR meter accuracy on 2 meters with some carbon resistors to simulate different SWR's. The results readily confirmed that the SWR meter was useless at VHF unless one didn't care whether a SWR was really 1:2 or 1:5. Rather than purchase another SWR meter, it was decided to construct one that would render reasonable results, within 10% accuracy or so, on the VHF bands, particularly 144 and 220 MHz.

There is nothing basically new in the circuitry of the SWR meter to be described. What is different about it is that it utilizes a particularly simple and inexpensive method of construction that yields reasonable results. It can be constructed as a completely self-contained SWR meter or only the pickup element can be constructed and used with an external meter. The circuitry as shown here for the meter utilizes two meter movements so one can read forward and reflected power simultaneously and avoid the annoyance of having a forward-reflected switch arrangement for a single meter.

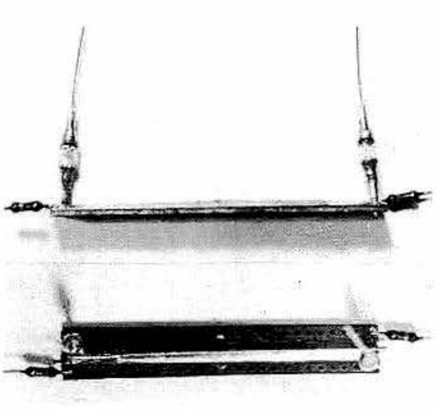

The heart of the SWR meter is a carefully constructed pickup element. Details of construction are discussed in the text but the photo shows how closely the diodes and terminating resistors must be soldered to the pickup element.

Pickup Element

The "heart" of any SWR meter of the reflectometer type is the pickup element. Many elaborate forms for such elements have been devised which involve complicated mechanical construction. Such complicated construction does become necessary if very accurate results are desired and if the meter is to maintain linearity over a very wide frequency range. However, over a smaller frequency range and with some minor sacrifice in accuracy, the construction of a pickup element can be greatly simplified. Basically, the pickup element should not cause any discontinuity in the transmission line section in which it is inserted but yet be long enough so enough voltage can be picked up in both the forward and reflected directions to make the meter usable with even low power transmitters.

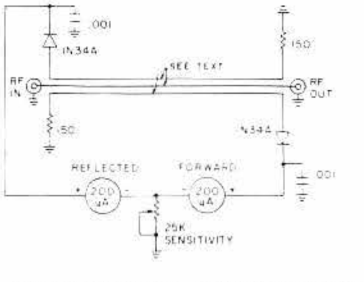

Fig. 1. SWR meter circuit. Two identical meters should be used (current range and internal resistance).

The pickup element I devised is shown in the photo. It is a 2-7/8" long piece of alternate grid pre-punched perf-board stock. The board is about 7/16" wide and within this width contained 4 separate copper strips spaced about 1/16" or less apart. The center two strips are soldered together to act as the inner conductor continuation of a coaxial line section. To solder the two inner strips together tack solder a bare piece of hookup wire between the two strips and then cover the entire two strips with solder.

Without the use of pickup wire, it will be nearly impossible to develop a smooth solder flow between the strip. Each outer strip acts as a pickup element for the SWR meter circuit shown in Fig. 1. The terminating resistor and diode are soldered to each end of the outer strip as shown in the photo and with minimum excess lead length to the strip. The use of a heat sink on the diode is necessary to prevent damage during soldering.

Mounting The Pickup Element

The pickup element made from the board stock is mounted between two approximately spaced SO-239 chassis connectors. The center strip of the board is soldered at each end to the center post of the SO-239 connector. The terminating resistor at each end is grounded as directly as possible to a ground lug held in place by one screw of the SO-239 mounting hardware. These details are shown in Fig. 2. It is important that the terminating resistor be grounded in this manner with minimum lead length. The enclosure in which the pickup element is contained should just be wide enough to accommodate the SO 239 connectors so that when the enclosure is secured together, the pickup element is boxed in by a metal surface on each side except directly above it. Many chassis or enclosure types are suitable for this purpose and the overall size of the enclosure will depend, of course, on the meter used and sensitivity control placement. These details are not covered here because they can be made as desired. They will not affect the basic accuracy of the meter as long as the pickup element is properly mounted and enclosed. The bypassing of the pickup rectifier diodes must also be done with as short leads as possible on the bypass capacitor. As shown in Fig. 2, a two lug terminal strip (one lug grounded) mounted on the side wall of the enclosure immediately at the cathode end of the diode will perform this task very well. The length of the IN34A diodes is such that the bypass capacitor cannot be connected to the same ground lug used for the terminating resistor. The leads on the capacitor would be too long and it will be ineffective.

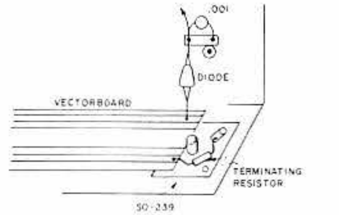

Fig. 2. Center strip of vector board is soldered to center post of SO-239. Other components are mounted at each end of board as shown (only one end shown here).

Operation and Results

The dual meter circuit of Fig. 1 reads forward and reflected power simultaneously. The sensitivity potentiometer is set to read full scale on the forward meter and the SWR read directly from the reflected meter. The latter meter can be calibrated for various SWR's by the use of small carbon resistors (100Ω to simulate a 1:2 SWR in a 50Ω line, etc.). Usually, it is only necessary to calibrate the reflected meter for SWR's of 1.5, 2 and 3 at the frequency of interest. Calibrated in this manner, the accuracy of measurement will be roughly 10% and is certainly good enough for most general antenna work. A particularly nice meter display can be made if one can find a two meter movement in one enclosure. I purchased a surplus stereo VU meter which had dual 200 μA movements and used it in the SWR meter.

The sensitivity of the SWR meter is such that transceivers of the 1-2 watt output class can easily be used with it on the VHF bands. The basic meter, of course, can also be used on the lower frequency bands as well and it will retain good accuracy. The only problem on the lower frequency bands is that the pickup strips are so short that more transmitter power has to be used to activate the meter than is usually convenient to use during antenna experiments. No exact tests were made but probably 70–100 watts would be needed to use the meter on as low a band as 80 meters. The meter was used and checked, however, on 40 meters. The accuracy of the meter remained very good and full scale deflection of the forward level meter required a power level of 60 watts. Being an in-line type meter, it can be left permanently in line when used on any band with a minimum of loss.

DIRECTIONAL COUPLER AND VSWR BRIDGE FOR VHF AND UHF

Bob Kolb WA6SXC

The aerospace industry has fostered the development of many new components and materials. Hams, being the kind of people they are, are quick to see practical applications for these materials that never occur to design engineers. I have often heard the criticism that it is impractical to publish articles or design ham gear with these new or expensive materials, because most OM's don't have access to them. Yet I've often been dismayed when I learn of an application for a piece of surplus equipment after it is no longer available. For this reason I feel that we should publish any application that is practical regardless of how immediately it can be used. Sooner or later, the material will show up on the surplus market and then we'll know what use can be made of it.

Reliable test equipment for the VHF-UHF bands is difficult to come by on a low budget. The literature is full of "relative" measuring devices but few pieces of homebrew gear are engineered for repeatable performance. Several directional couplers have been built according to the description in this article and each has measured to within ±.5 dB of the designed value. This is due in part to the mechanical rigidity and close tolerance of RG-141 "coaxitube." No special tools are required outside of a cheap pair of vernier calipers. The tools used to make the original coupler were an Xacto knife, file, soldering gun, vernier calipers and a vise. Don't let the calipers scare you. If you're not after a closely calibrated device they may be omitted.

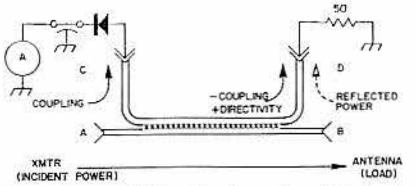

Fig. 1. A typical directional coupler. This device is the heart of a VSWR bridge, and can also be used for many other applications.

The design goal was a directional coupler with about 30 dB directivity in the passband with a low insertion loss. Each milliwatt measured at the coupling arm equals one watt through the main line. Such a device is the heart of a good quality VSWR Bridge. The measured values were 30.3 dB coupling and 38 dB directivity at 432 MHz. Data presented in the graph was taken using HP608C and 614A signal generators and a General Microwave R. F. Power Meter. The measured insertion loss was 0.2 dB.

Resolution of the smallest possible VSWR is limited by directivity. Few of the handbook VSWR bridges or the low cost type attractive to the CB trade achieve as much as 20 dB directivity. Thus the minimum discernable VSWR is approximately 1.7:1. With 38 dB directivity, 1.02:1 VSWR's can be accurately measured.

Directivity may be defined as the isolation of arm D from arm A, over and above the coupling as shown in the **Fig. 1.** Coupling is achieved by removing part of the jacket between adjacent coax conductors. If the input is at arm A, incident power can be sampled 30 dB down at arm C but appears −68 dB at arm D. Reflected power entering arm B is sampled −30 dB at arm D while at arm C it is −68 dB. It stands to reason if the directivity is low, one cannot tell with certainty if he is measuring incident or reflected power. Port D may be used as the dc return for a detector at port C and vice versa.

This device will have its fundamental passband where the length exposed between the two lines is $\lambda/4\sqrt{\epsilon_r}$. It will also have a passboard at $(2n - 1)\ \lambda/4\sqrt{\epsilon_r}$ or at three, five, seven, etc., times the frequency for which it is a quarter wave. Hence a coupler designed at 432 MHz is usable at 1296 MHz.

This coupler has also been used to measure relative power and modulation at 2 meters where its coupling factor for incident power is approximately −40 dB but the directivity is poor, hence arm D must be terminated in 50 ohms. It's a real aid for tune up and will give a good indication of increased power with AM modulation right in the r.f. line. RG141 will handle 500 watts of rf up to 2000 MHz.

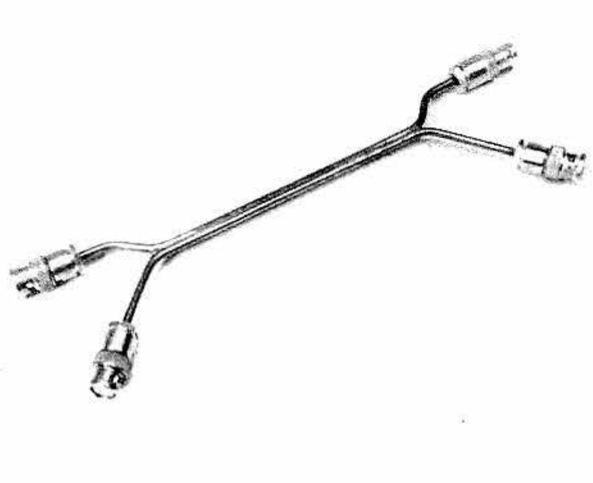

This UHF directional coupler is very simple to make, yet offers excellent performance on the 70 and 23 cm bands.

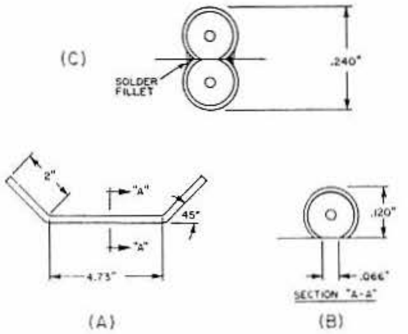

Fig. 2. Details of the construction of the UHF directional coupler.

The formula for determining coupling length is

$$\frac{c}{4 f_o \sqrt{\epsilon_r}} = \frac{\lambda c}{4\sqrt{\epsilon_r}} \text{ or } \frac{300 \times 10^{8} \text{ cm}}{4 \times f_o \times \sqrt{2.1} \times 2.54 \ {}^{\text{cm}}/_{\text{in}}}$$

$$= \frac{\lambda \text{ c inches}}{4\sqrt{\epsilon_r}}$$

$\sqrt{\epsilon_r}$ for Teflon $= \sqrt{2.1} = 1.449$

From these calculations $\frac{\lambda \text{coupling}}{4}$ at 432 MHz is 4.73 inches. With an Xacto knife cut two pieces of line 8.73 inches long and carefully bend them so that they form the shape shown in **Fig. 2A.**

Clamp the bent coax into the vise and file away the copper jacket taking care that the filed surface is smooth and flat. A belt or stationery disc sander works well too. A cross section of the filed piece should look like **Fig. 2B.** Next fit the two pieces together so that a cross section would look like a figure 8 and secure in a vise. Heat with a soldering gun only. Do not use a torch. Avoid excessive heating. Flow solder between the two lines as shown in **Fig. 2C.** The "arms" can now be bent into any convenient configuration provided enough allowance is made at the ends for connector assembly. A good rule to follow is a minimum bend radius of half an inch although a quarter inch radius is permissible. The arms should be approximately two inches long.

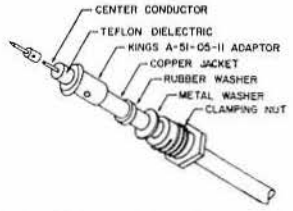

Fig. 3. Use of Kings A-51-05-11 adapter for using GR-141 with standard BNC connectors.

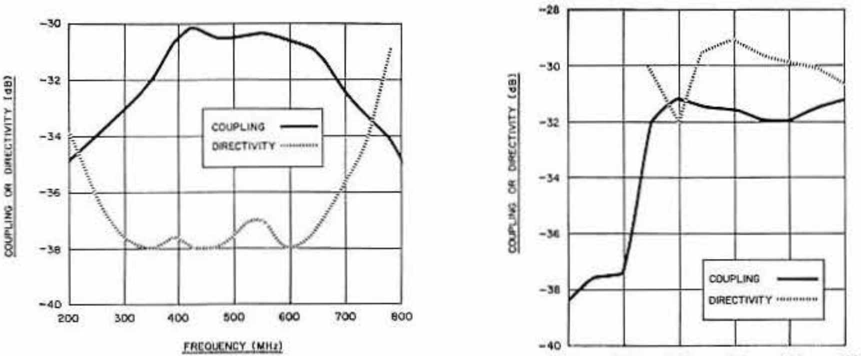

Fig. 4. Coupling and directivity for a directional coupler similar to the one discussed in the text. This device used a coupling wavelength of 4.635 inches rather than the 4.73 inches specified in the text. The only effect of the longer wavelength is to center the curves on 432MHz instead of about 500 MHz.

RG 141 has the same cross section as RG 58/U *without* the vinyl jacket therefore any connector that will accept RG 58/U can be used on RG 141 provided a sleeve is made up to make a snug fit in the clamping nut. A special adaptor is made by Kings for this purpose and sells for 45 cents. The connector assembly is shown in **Fig. 3.** Three RG 88E/U and one RG 89C/U connectors were used on the coupler shown in the photo.

Fig. 4 gives the measured directivity and coupling for this type of directional coupler at both 70 and 23 cm. You can see that performance is quite satisfactory.

Fig. 5 lists a number of applications for a directional coupler. The detectors in the measuring instruments should be suitable for use at 500 or 1300 MHz.

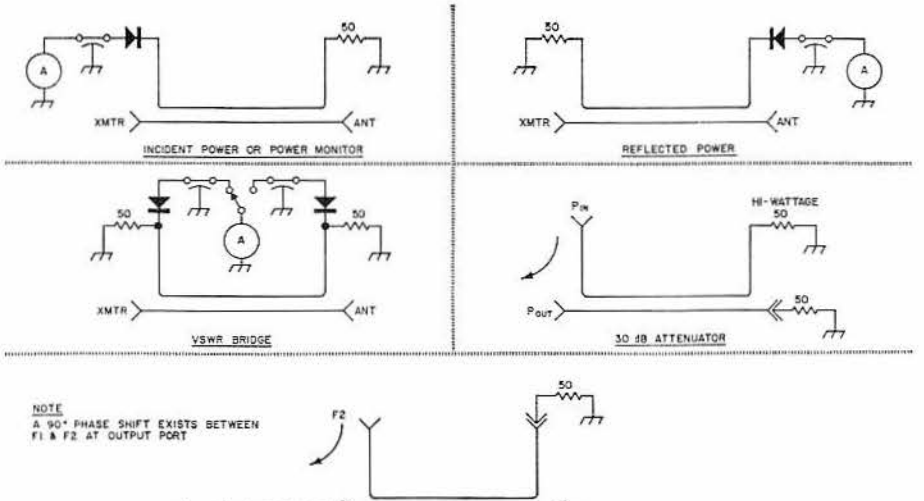

Fig. 5. Applications of the directional coupler described in the text. Unlike most pieces of ham-made test equipment, this one is good at 450 and even 1300 MHz.

A SLOTTED LINE FOR 1250 MHz

Silas Smith WA8CHD

The SWR bridge is a very useful—and sometimes badly neglected—tool. Especially on 1250 mc with only two or three watts of power, you should measure your SWR and do something about it if necessary. I recently joined the 23 cm boys, though I have only worked 1234 mc so far. My first attempt to communicate with W8VKQ on the other end failed but after two evenings of diligent work with a meagre amount of test equipment we made contact. It would have been a lot easier with test equipment good in the 1250 mc range since most hams are not equipped to measure frequency and adjust their rigs at these frequencies.

After that experience, I made the slotted line indicator described in this article. The cost is next to nothing and it's easy to build, but it does a good job. The unit is built around a one inch thin wall copper tube 10¼" long (**Fig. 1**). This tube has a ⅛" slot cut lengthwise for 7½", about ¾ wavelength at 1250 mc. This is long enough to get a fair sampling of the standing wave or null points.

Three other pieces of 1" copper pipe are needed. Two are 5/16" long and one is 1⅜" long. The two 5/16" pieces each have ⅛" cut out and are reshaped to fit inside the 1" tube. These two pieces were each soldered to 1/16" plates as in the detail in **Fig. 1.** This process makes two cup-like structures which should

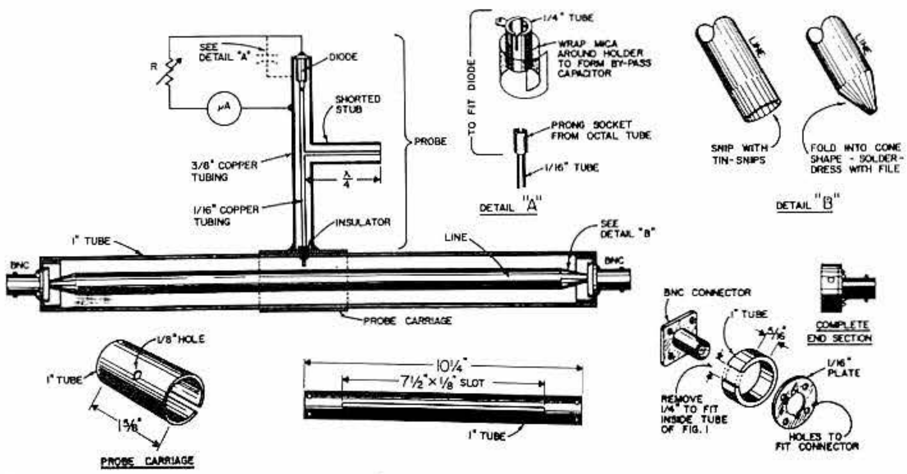
SEE DETAIL "A"
DIODE
R
μA
SHORTED STUB
PROBE
3/8" COPPER TUBING
$\frac{\lambda}{4}$
1/16" COPPER TUBING
INSULATOR
1" TUBE
BNC
LINE
SEE DETAIL "B"
BNC
PROBE CARRIAGE
TO FIT DIODE
1/4" TUBE
WRAP MICA AROUND HOLDER TO FORM BY-PASS CAPACITOR
PRONG SOCKET FROM OCTAL TUBE
1/16" TUBE
DETAIL "A"
LINE
LINE
SNIP WITH TIN-SNIPS
FOLD INTO CONE SHAPE - SOLDER - DRESS WITH FILE
DETAIL "B"
BNC CONNECTOR
1" TUBE
COMPLETE END SECTION
1/16" PLATE
REMOVE 1/4" TO FIT INSIDE TUBE OF FIG. 1
HOLES TO FIT CONNECTOR
1/8" HOLE
1" TUBE
1⅜"
PROBE CARRIAGE
10¼"
7½" x ⅛" SLOT
1" TUBE

Fig. 1. The slotted coaxial line and probe for 1215 to 1300 mc described in this article.

fit within the end of the 1" pipe. Drill nine holes in each cup: four holes for attaching the cup to the slotted line, four to fit the BNC connector mounting plate and one ⅜" hole to pass the main body of the connector. File two notches for the connector ears. You can fasten the cups to the slotted tube with small sheet metal screws, threaded holes or nuts soldered to the back of the holes.

The 1" x 1⅜" piece of pipe is for the probe carriage. It is cut lengthwise on one side, slipped over the 1" tube, centered, and a ⅛" hole drilled through it over the slot in the 1" tube.

The probe is built from ⅜" and ⅛" brass tubing from your hobby shop. At the probe end is an insulator from a coax fitting and on the other end is a jack made from an octal tube socket. A small piece of ¼" brass tubing is used to make the diode socket (see detail A). A piece of 22 gauge wire is inserted in the probe end of the 1/16" tubing and soldered. A 1N21 diode works very well.

Use **Fig. 2** to choose the proper size for the line in the center. Each end of this line should be tapered and soldered to the coax fitting as in detail B. The center conductor should be 7/16" for 50 ohms or 5/16" for 73 ohms.

The meter used should have a range that permits a full scale reading and should be calibrated with some scale.

Uses: Measuring frequency

Frequency can be measured with the slotted line in two ways: the distance between null points at one half wavelength ($\lambda/2$) or the distance between peaks. See **Fig. 3.** Measuring between the nulls is preferable since the peaks are very broad. If your line is flat, the peaks and nulls may be very small, so you may have to induce a mismatch in the line. Likewise, you may have to induce a mismatch to get the two nulls to fall within the 7½" of the line. The easiest way to induce a mismatch is a short in the line.

Measure the distance between the nulls carefully with a centimeter rule. Twice this distance divided into 30,000 will give you your frequency in megacycles. It would be a good idea to check this against a standard or at least avoid operating too close to the edge of the band.

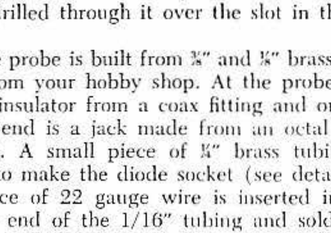

Fig. 2. Ratio of outer tube to inner line for various impedances.

Measuring VSWR

Measuring voltage standing wave ratio (VSWR) with the slotted line is easy, too. We are looking for a ratio of E_{min}, the minimum voltage on the line, and E_{max}. The ratio:

$$\frac{E_{max}}{E_{min}} = \text{VSWR}$$

However, there is a slight complication of vectors in E_{max} and E_{min}, so that to find them, you must first determine the highest voltage read on the meter, E_1 and the lowest voltage read on the meter, E_2 with the same setting of the shunt resistor, as the probe carriage is slid along the line. Now, $E_{max} = E_1 + E_2$ and $E_{min} = E_1 - E_2$ so that

$$\text{VSWR} = \frac{E_1 + E_2}{E_1 - E_2}$$

An even simpler method is to set the meter to full scale (100) at the highest reading, and reading the VSWR directly from **Fig. 5.**

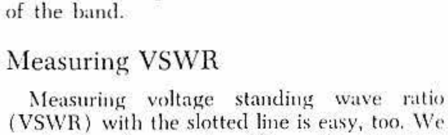
GENERATOR
SLOTTED LINE
LOAD
$\frac{\lambda}{2}$
STANDING WAVE
$\frac{\lambda}{2}$
GENERATOR
SLOTTED LINE
LOAD
FORWARD WAVE
REFLECTED WAVE

Fig. 3, Left. Measuring frequency. Fig. 4, Right. Measuring VSWR.

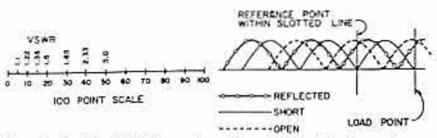
VSWR
1.1
1.22
1.35
1.5
1.85
2.33
3.0
0
10
20
30
40
50
60
70
80
90
100
100 POINT SCALE
REFERENCE POINT WITHIN SLOTTED LINE
REFLECTED
SHORT
OPEN
LOAD POINT

Fig. 5, Left. VSWR scale. Fig. 6, Right. Impedance measurements.

Impedance measurements

It's a little harder to measure impedance, but even it's not bad if you do it step by step. First, consider the high impedance on your line when your antenna is open or shorted. Perhaps you don't think of it as a change in impedance. This is what we are looking for. Remember that there is a definite relationship between impedance and frequency. **Fig. 6** shows the voltage relationship between a short and open. On a short set the probe for a null point, usually the first null from load end of the slotted line. This will be our reference point. You will note that when the load point is open there is a shift of 90° or a quarter wavelength either side of the reference point. This adds up to one half wavelength. Any impedance between infinity and zero ohms will lie somewhere along the half wavelength of the line. It will be noted that an open is equidistant from the reference point toward the generator and toward the load, however the line has moved toward the load (capacitive reactance). To measure this reactance, we replace the load that we measure with a short. We set our probe to the first minimum reading from the load end of the slotted line. Mark this spot with a scribe. This is our reference point. Now we remove the short and add our load under measurement. It will be noted that the voltage has moved upward. Move the probe toward the load. If the voltage goes down, you are heading in the right direction. If the voltage starts to go up change direction toward the generator. Move the probe as above to the new null point and note the direction that the probe has been moved.

Here's a step-by-step example. First, you'll need the frequency and wavelength, VSWR, direction of probe, distance between reference point and new null, and impedance of the slotted line.

Let's say as we measured the frequency, the distance between the null points was 12 cm. This is half the wavelength, so the wavelength is 24 cm. The frequency is 30000/24 or 1250 mc.

Let's say we read 80 as a maximum and 20 as a minimum on a point scale in measuring the VSWR:

$E_{max} = E_1 + E_2 = 80 + 20 = 100$
$E_{min} = E_1 - E_2 = 60$

$$\text{VSWR} = \frac{100}{60} = 1.66$$

Or set E_1 to 100 by adjustment of R, then E_2 would read 25 on the meter, a VSWR of 1.66 by **Fig. 5.**

As in **Fig. 6,** the probe was moved toward the load.

Let's assume the distance from the reference point to the new null was 2 cm. What part of a wavelength is 2 cm? Wavelength is 24 cm, so 2/24 = .083 wavelengths.

This is a movement of .083 wavelengths toward the load.

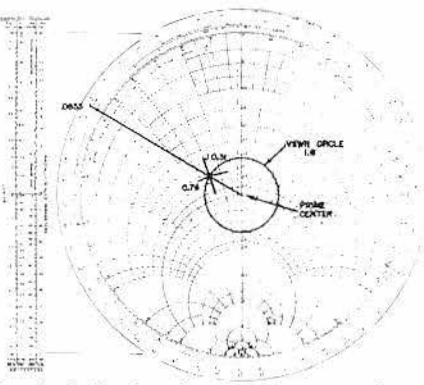

Fig. 7. Smith chart for determining impedance.

The slotted line used has an impedance of 50 ohms.

We are now ready to plot this information on the Smith chart (**Fig. 7**, better get out your magnifying glass).

Draw the VSWR circle at the prime center with a radius of 1.6.

Draw a line from .083 wavelength toward the load from the prime center.

At the intersection of the circle and line, follow the constant resistance circle to .74 and the capacitive reactance circle to .31. Our load impedance is then:

$Z_L = Z_o\ (.74 - .31j)$
$Z_L = 50\ (.74 - .31j)$
$Z_L = 37 - 15.5j$

Well, there are some of the things your slotted coax line can do besides telling you that you're on the air. A commercial version would cost $750. I guarantee that you'll get your dollar's worth out of this one as it costs less than $1 to build. I hope that it helps you get on 1250 mc.

SOME DIRECTIONAL WATTMETERS AND A NOVEL VSWR METER

P. G. Martin G3PDM

Most conventional reflectometers cannot be used for accurate power measurement because their sensitivities are frequency dependent. This is due to the use of combinations of reactance and resistance in the sampling circuits which detect the transmission line current and voltage.

This basic problem can be solved by the use of conventional lumped components instead of the distributed parameters of a transmission line. The line voltage can be sampled by two resistors or two capacitors used as a voltage divider, rather than one resistor and some distributed capacitance. The line current can be monitored by a properly designed current transformer instead of an inductance and resistance. High frequency current transformers consist of primary and secondary windings on a ferrite or iron dust toroidal core, with a low value of load resistance across the secondary winding.

All SWR bridges and directional wattmeters need to generate two dc voltages proportional to the forward and reflected voltages or currents of the transmission line. To achieve this one has either the current detector or the voltage detector providing two antiphase signals so that addition and subtraction can be performed.

A Frequency-Independent Directional Wattmeter

M. B. Allenson G3TGD, has designed a wattmeter using the above principles, where the low resistance in the current transformer secondary circuit is split into two equal parts. The center connection is taken to the voltage sampling point so that sum and difference voltages are available at the ends of the transformer secondary winding, see Fig. 1.

With two meters, this circuit can be used as a versatile calibrated directional wattmeter over the freqnecy range 100 kHz to 70 MHz, with an accuracy of about 10 per cent. Precise calculations of SWR and transmitter efficiency can be made.

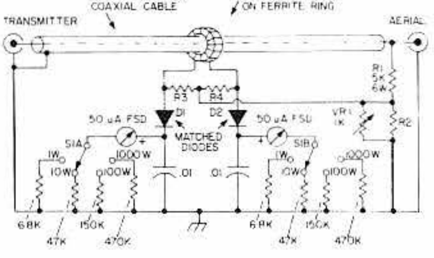

Fig. 1. Circuit of the basic frequency-dependent directional wattmeter due to G3TGD. The two meters indicate forward and reflected powers.

Maximum sensitivity with a 50 μA meter is less than five milliwatts, but with the multiplier resistors specified in Fig. 1, full scale deflection corresponds to power of 1, 10, 100 and 1000 watts. Calibration is non-linear, because the instrument samples voltage, and power is proportional to voltage squared.

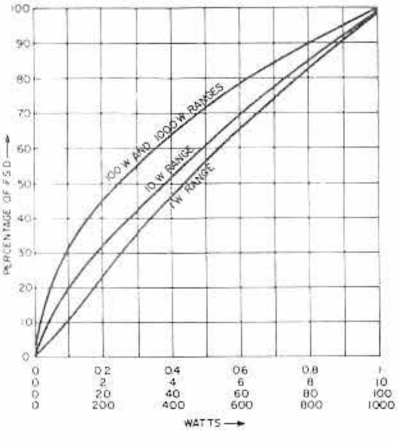

Fig. 2. Calibration curves for the instrument described in Fig. 1.

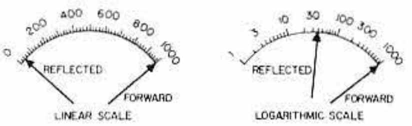

Fig. 3. Linear and logarithmic scales. The inherent advantages of the logarithmic form are immediately obvious.

Unfortunately, two transmission line impedances are in common use in coaxial systems: 50Ω and 75Ω. As it is not possible to design instruments whose sensitivities are independent of line impedance, some component values must depend on the impedance in use. For simplicity, only one of the voltage driver resistors need be changed, but instrument calibration will be different. By changing the current transformer resistors as well as one of the voltage divider resistors, the calibration is the same for both line impedances. This technique has been adopted here, and the calibration curves in Fig. 2, are correct for 50 or 75Ω lines provided the resistor values in Table I are used.

The Logarithmic Wattmeter

The basic instrument can be simplified by including a logarithmic network so that the power range switch is redundant and a single meter scale can be used for powers from, say, one watt to 1000 watts. A logarithmic scale has the 1, 10, 100 and 1000 watt points equally spaced (see Fig. 3).

The advantage of a logarithmic instrument is that one can measure very low reflected powers and very high forward powers simultaneously with the same percentage accuracy, without having to switch meter ranges.

It is simple to add a reasonably accurate *wide-range* logarithmic network to the meter in Fig. 1 (see Fig. 5). The basis of its

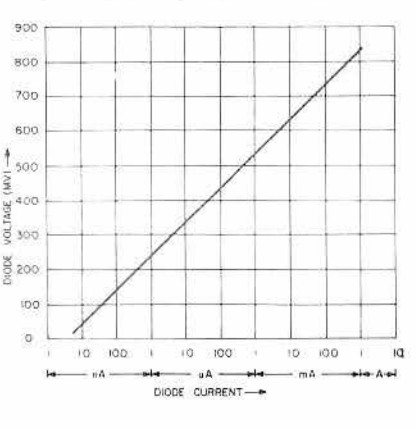

Fig. 4. Smoothed experimental plot of the current /voltage characteristic of a 1N4002 silicon junction diode, showing its logarithmic properties.

operation is that the voltage dropped across a forward-biased p-n junction diode is proportional to the logarithm of the current passing through it (see Fig. 4). To reduce the potential dynamic range of the circuit, a relatively insensitive meter is used, and a small resistance is added in series with the logarithmic diode to restore a logarithmic form to the scale (see Fig. 6).

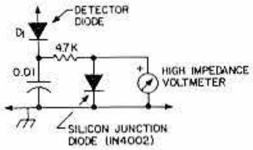

Fig. 5. Basic wide-range logarithmic converter.

An experimental logarithmic wattmeter is shown in Fig. 7. Figure 8 gives the calibration scale for 50 or 75Ω lines, provided the correct resistors are used (Table I).

A Direct-Reading SWR Meter

A particularly useful device would be an instrument giving a direct measurement of the standing wave ratio on a transmission line, independent of the absolute power levels or the frequency in use. Such an instrument, with its single meter, would be ideal for incorporation into transmitters and transceivers (especially with the physically small sampling circuits associated with it).

The swr can be expressed in terms of the forward and reflected voltages according to:

$$\text{swr} = \frac{E_f + E_r}{E_f - E_r} \qquad (1)$$

We wish to generate this function electronically, so that outputs of the two detectors can be used to generate a meter current proportional to SWR. This would be rather tedious, though not impossible.

Conveniently, manipulation of equation (1) shows that:

$$\frac{E_f}{E_r} = \frac{SWR + 1}{SWR - 1} \qquad (2)$$

which although not proportional to SWR, is a mathematical function of it only. Electronic division of E_f by E_r is easily done by taking logarithms and subtracting. That is:

$$\log \frac{E_f}{E_r} = \log E_f - \log E_r$$

Table I

	Ω	Ω
Line impedance	50	75
R3 and R4	27	33
R2	220	180

Values for R2, R3 and R4 to be used in 50 and 75Ω transmission lines.

In Fig. 9, the two silicon diode voltages are proportional to the logarithms of their currents, which in turn are proportional to the forward and reflected voltages. The two diode voltages can be subtracted directly by connecting a meter between them, rather than from each one to chassis.

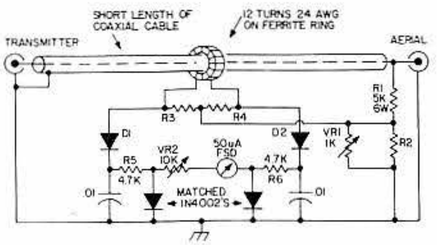

Fig. 6. Circuit of the logarithmic directional wattmeter. D3 and D4 are matched (see text).

The meter cannot be calibrated linearly in SWR, because of equation (2), and because the circuit does not take anti-logarithms after subtracting the logarithms. The outcome of this is beneficial: the SWR meter is increasingly sensitive as the standing wave ratio approaches 1:1. This is where one wants most sensitivity: to make the final adjustments to aerial arrays, to measure the variations in SWR over a band, and so on. Fig. 10, shows a calibration curve for SWR meters. Naturally the meter sensitivity cannot be completely independent of the power level in use. Accuracy falls when the reflected power is less than about half a watt (this corresponds to an SWR of 1.05:1 when the forward power is 1 kW).

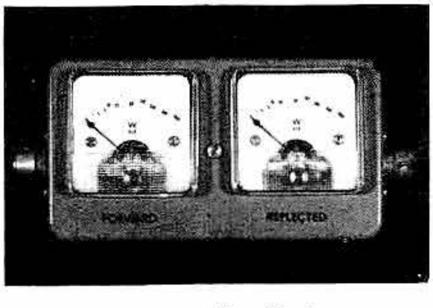

Fig. 7. An experimental logarithmic wattmeter.

A differential amplifier could be added to the circuit of Fig. 9, to enable a less sensitive meter to be used.

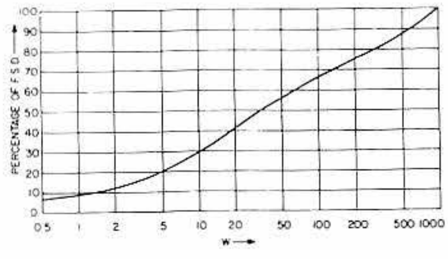

Fig. 8. Calibration curve for logarithmic wattmeters.

Construction of the Instruments

Layout of the sampling circuits is fairly critical, see Fig. 11. The input and output sockets should be set a few inches apart, and connected together with a short length of coaxial cable. The coax braid must be grounded at one end only, so that it acts as an electrostatic screen between the primary and secondary windings of the toroidal transformer. Twelve turns of 24 AWG enamelled wire, equally spaced around the circumference of the ring, form the secondary winding. The primary is formed by simply threading the ring onto the coax.

A suitable ferrite ring is the Mullard FX1596, made in England, although other types are suitable. The FX1596 has an outside diameter of half an inch, and is designed for wideband rf applications between 5 and 20 MHz. The main requirement is that the ferrite material should maintain a high permeability over the frequency range in use.

Other components in the sampling circuits should have the shortest possible leads. R1, R2 and R must be non-inductive solid carbon types: for high power levels (about 100 watts) R1 should consist of two or three 2 watt carbon resistors in parallel. VR1 should be a miniature skeleton potentiometer to keep stray reactance to a minimum, although it may be dispensed with by trying various fixed resistors for R2 until the reflected indication under matched conditions is zero.

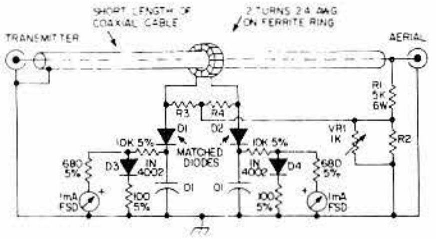

Fig. 9. Complete circuit for a power-independent, frequency-independent direct-reading SWR meter.

The detector diodes need to be matched for similar voltage drop, using the circuit in Fig. 12. Point contact germanium types with a PIV rating of 80 volts or so are recommended.

Logarithmic diodes should be modern medium-current silicon junction types, such as conventional rectifier diodes. The 1N4002 is specially recommended for its good logarithmic properties. Log diodes should also be matched with the circuit in Fig. 12.

The 0.01 μF decoupling capacitors should be a disc ceramic type.

In designing a toroidal transformer different to that specified, several factors should be borne in mind. As the number of secondary turns increases, the self-capacitance rises and causes the response to fall at high

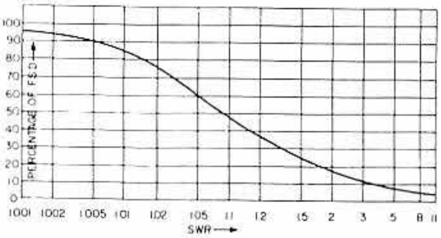

Fig. 10. Calibration curve for SWR meters of the type described in Fig. 9.

frequencies. Failure of this nature causes the reflected power indication to rise; in other words the directivity of the instrument falls. If the 27Ω resistors are raised appreciably in value, the instruments will eventually become frequency sensitive.

The ratio of the voltage sampling resistors (R1 and R2) in the HF designs is determined by the sensitivity of the current sensing circuit, and the two sampling voltages must be equal in magnitude under matched conditions. VR1 provides fine adjustment of the ratio. Absolute values of the resistors can be varied considerably, bearing in mind that as their values increase the stray capacitance across them may need to be compensated for.

Useful Equations

Let the line current be I amps, the line voltage be V volts, and the characteristic impedance of the transmission line be Z_o. Then $V = IZ_o$.

If the current transformer has a ratio of 1:n, and each of the resistors in its secondary circuit has a value of RΩ, then the rf voltage across each of them is given by:

$$V_{(i)} = \frac{IR}{n} \qquad (3)$$

The voltage detector output is obviously

$$V_{(v)} = \frac{VR_2}{R_1 + R_2} = \frac{R_2}{R_1 + R_2}\,IZ_o$$

Which is, to a good approximation,

$$V_{(v)} = \frac{R_2}{R_1}\,IZ_o \qquad (4)$$

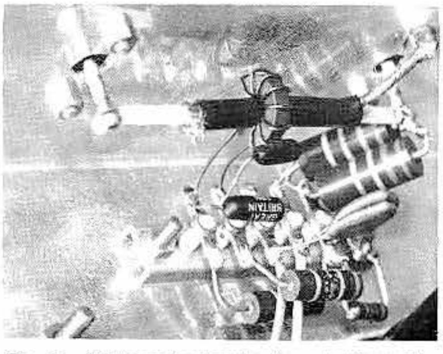

Fig. 11. Photograph showing layout of sampling circuits used in an experimental swr meter.

The main design equation for all the HF instruments is therefore:

$$R_2 = \frac{R.R_1}{n.Z_o}$$

where the value for R2 includes the effect of VR1, if fitted.

The dissipation of some of the components specified is quite high. For those planning different circuits, the following equations express the dissipation of R1 and the current transformer resistors R:

$$W_{(R1)} = \frac{Z_o.W}{R_1} \quad \text{watts,}$$

where W is the transmitter output power.

$$W_{(R)} = \frac{W.r}{n2.Z_o} \quad \text{watts.}$$

In the instruments described, $W_{(R1)}$ is about 5 watts, and $W_{(R)}$ 2 watts for a transmitter power of 500 watts.

Calibration

If any of the instruments are built exactly as described, and used in systems of the correct impedance, the calibration given in Figs. 2, 8 and 10 will be sufficiently accurate for most purposes. For those designing their own circuits, the following procedure is recommended.

Test equipment needed includes a high power rf source (a transmitter) and an rf voltmeter. The instruments can be calibrated with less accuracy without the rf voltmeter. The wattmeters are calibrated by feeding power through the meter into an appropriate dummy load (50 or 75Ω). VR1 is set for minimum reflected power indication, and the power scale is marked according to the rf voltage appearing across the load. If an rf voltmeter is not available, a peak-reading type can be made with a diode, capacitor and dc voltmeter. As the detector output is equal to the peak rf voltage applied to it, equation (4) leads to:

$$V_{(det)} = 2.8\ V\ \frac{R_2}{R_1} = 2.8\sqrt{WZ_o\,\frac{R_2}{R_1}}$$

It would be difficult for most amateurs to obtain sufficient high power carbon resistors to calibrate an SWR meter by means of deliberate mismatching. An indirect method is therefore recommended.

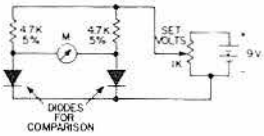

Fig. 12. Hookup circuit for matching detector diodes for equal forward voltage drop, and silicon junction diodes for similar logarithmic properties. The meter should be as sensitive as possible (say 50 µA fsd), and should not deflect appreciably as the voltage is varied between zero and nine volts.

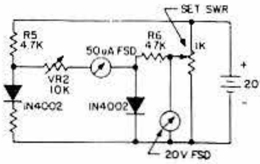

Fig. 13. Circuit used to calibrate SWR meters(see text).

Disconnect R5 and R6, Fig. 9, from the detectors, and connect them instead as shown in Fig. 13. One voltage is fixed at about 20 volts, and the other is varied between zero and 20 volts. The ratio of these voltages corresponds to a definite SWR which can be determined from equation (1). Before carrying out this procedure, however, VR2 should be set for full scale deflection of the meter under matched conditions at the highest power level to be used.

Conclusions

All of the instruments described in this article have been tested under actual operating conditions, on all amateur bands between 1.8 MHz and 30 MHz. Power levels used varied from 100 to 1200 watts. With the components specified, the instruments will sustain power levels well above the kilowatt level for periods of tens of seconds.

It is hoped that by introducing frequency independent directional wattmeters, one will be able to make useful comparisons of absolute power levels and accurate assessments of standing wave situations. The logarithmic scales are an added convenience, and the direct-reading SWR meter offers a saving in meters.

SWR BRIDGES USING ZERO-CENTER METERS

John Schultz W2EEY/K3EZ

The basic coaxial cable swr meter, which samples both the forward and reflected voltages on a transmission line, is an essential instrument in setting up an amateur station. Although a low transmission line swr in itself doesn't guarantee that an antenna system will work, knowing what the swr is remains an essential bit of information in evaluating the status of the transmitter/antenna interface.

The virtues of the usual coaxial line swr meter – economy and simple circuitry – are offset by the awkwardness involved in using the instrument: that is, switching back and forth between "forward" and "reflected" switch positions and adjusting a sensitivity control for sull-scale deflection in the forward position. This awkwardness of operation is partly relieved by dual-instrument swr meters so the forward/reflected switch is no longer necessary – but the cost of the instrument increases. The awkwardness of operation can be completely relieved by special dual movement meters with crossing pointers and/or special circuitry, but then the cost really soars.

Another approach is the usage of some different circuitry to detect the swr which can then activate a zero-center meter so one can peak up the antenna coupler, etc., for minimum swr on a transmission line with the same ease as one tunes an FM entertainment receiver for a zero-center reading on the discriminator meter. In the end analysis, practically no line is set up for any specific swr, but simply for the lowest swr that can be achieved. So, why not make the process as simple as possible? Circuitry to do this has actually been available for many years and was especially developed for use with the automatic servo-tuning of military transmitters. The zero output when tuning conditions were correct, and the plus or minus output when they were not, allowed, after sufficient power amplification, the activation of motors to tune transmitter PA task circuits, coupling circuits, etc. Now, with the economical availability of toroids and imported zero-center meters, zero-center swr type circuits can be applied to amateur usage and meet the criteria of being easy to construct and economical. They offer a real challenge to the conventional type of swr meter, especially when one is using a multiband antenna and tuner arrangement, or just leaves an swr meter in a transmission line to verify that the line swr is holding constant. In the former case, one usually knows the approximate transmitter/tuner settings on each band, and the swr meter is used to peak up the adjustments. Such peaking-up, as mentioned before in the case of an FM receiver, is certainly facilitated by just having to look at a zero-center meter. In the latter case, a zero-center meter will more readily reveal slight changes in transmission line swr from an established norm. Perhaps the only place where a conventional swr meter is still to be preferred is when one is dealing with experimental antennas where extreme swrs exist and a conventional swr meter can help to more quickly obtain "ballpark" antenna coupler or loading coil settings. The zero-center type circuits operate on a slightly different basis than the conventional swr meter circuits, and just a brief review of transmission line operation will be helpful to understand the different basis for the functioning of the two types.

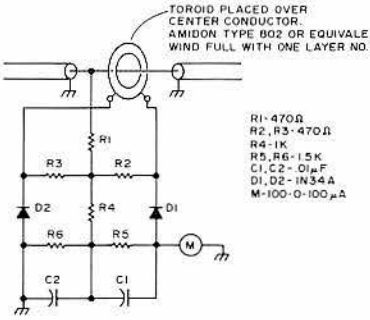

Figure 1.

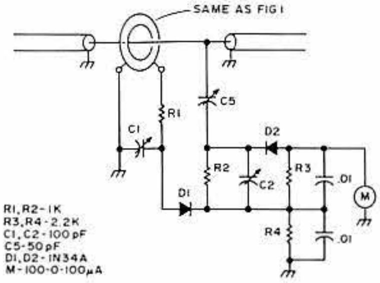

Figure 2.

If a transmission line is attempting to deliver power to a load, both voltage and current values exist along the line. If the antenna load impedance is resistive and equals that of the line impedance, all the power feed into the line reaches the antenna less any power loss in the line itself because of the line's inherent loss. If the load is not equal to the line impedance or reactive, part of the power the line tries to deliver is rejected by the load and reflected back along the line. This sets up standing waves of voltage and current which are not in phase with the original voltage and current waves the transmitter is trying to "pump" into the line.

Conventional swr meters sample a portion of the voltage or current wave traveling towards the load and a sample of that traveling back from the load. The relative amplitude of the two samples is the swr and, of course, the desired condition is that there be no reflected wave from the load. If there is no reflected power from the load, the line is often referred to as being "flat." This term is a bit misleading. Although the reflected power may be zero under matched conditions, the power flowing to the load via the line will still set up normal sine wave distributions of voltage and current along the line.

Another way to sense that the load is matched to the line is to determine first of all that the phase difference between the voltage and current on the line is zero and that the ratio of the in-phase voltage and current is correct for the impedance of the line being used. If there is no reactance present in the load, there will be no phase difference between voltage and current, as in any ac circuit feeding a resistive load. If the load is not only non-reactive but also of the correct resistance, the ratio of voltage and current on the line will have a definite relationship correct for a given line impedance. The zero-center meter circuits are designed to monitor both of these line conditions.

Practical Circuits

Two separate circuits are used to monitor the line conditions just described. One could switch a separate zero-center meter between the two line monitor circuits but with imported zero-center meters available for as low as $1.50, it is hardly worthwhile to do so. (A good source of the 100-0-100uA meters needed is Edlie Electronics, 2700 Hempstead Turnpike, Levittown NY 11756. Their meters with order number DA798 or DA792 cost only $1.50 but they have a $7.50 mail order minimum.) Also, for many installations, it is not necessary to use both circuits. For instance, in a home installation feeding a set of various fixed antennas, it would probably be of primary interest to have the impedance monitor circuit (voltage/current ratio). Once the antennas have been set up and operate satisfactorily, one would primarily just be interested to know that some impedance condition on the line has not changed. In a mobile installation, where some reactance cancelling L or C needs to be adjusted with frequency changes, it would be of prime interest to have the phase monitor circuit once the installation has been initially set up for a correct impedance match, and the reactance cancelling component only needs adjustment during actual operation.

Phase Monitor

Fig. 1 shows the circuit of the phase monitor. It is a basic discriminator circuit. The voltage on the transmission line to ground is sampled via R1. The current on the line induces a voltage in the windings on the toroid placed in the line which, of course, is 90° out of phase with the voltage sample. The voltage proportional to the line voltage is divided across R1 and R4 (C1 and C2 are rf bypass capacitors). The voltage proportional to line current (across the toroid winding) is divided between R2 and R3. Referenced to the center point of R2 and R3, these voltages are 180° out of phase with respect to each other and either one is 90° out of phase with respect to the voltage across R4. The voltages across R2 and R4 are rectified by D1 and, because of the diode polarity, produce a voltage with the polarity indicated across R5. Similarly, R3 and R4 and D2 produce a voltage of opposite polarity across R6. When the transmission line current and voltage are in phase, the voltages across R5 and R6 are equal and opposite, and the resultant voltage output which drives the meter is zero. For an out-of-phase condition, some output voltage exists whose magnitude indicates the amount of the phase difference and whose polarity would indicate whether the line voltage leads or lags the line current.

One could actually calibrate the output meter for different magnitudes and directions of phase difference, but this would be very tedious and not of real use for just checking the adjustment of a line for reactance cancellation. The only calibration that has to be checked is to excite the circuit into a resistive dummy load and to be sure that the zero-center feature works. If it does not, it indicates that some unbalance has been introduced during construction or some of the components (R2/R3 and R4/R5) are not closely enough matched. If the meter excursions off zero center become too violent with a high power transmitter, some additional resistance can be introduced in series with the meter to limit the maximum current flow.

Impedance Monitor

Fig. 2 shows the circuit of the impedance monitor, which is also a form of discriminator circuit. Samples of the transmission line voltage and current when they achieve the desired ratio produce an output voltage from the circuit which is zero. As in the previous circuit, a toroid placed around the center conductor of the transmission line develops by transformer action a voltage proportional to the line current. A sample of the line voltage to ground is taken via the variable capacitor C5. Unlike the previous circuit, both the sampled voltages are rectified, producing voltages across load resistors R3 and R4, without paying any attention to the phase differences between the sample voltages. It is only the amplitude of the sample voltages which determines the amplitude of the voltage across R4 (for the line voltage) or across R3 (for the line current). Because of the diode polarities, the voltages across R3 and R4 are of opposite polarity. They will cancel when equal, producing zero output voltage, and since C5 is variable, this condition can be set to occur at any desired line voltage to current ratio.

The difficulty with this circuit is that the sample voltages generated by the toroid transformers and the C5 coupling are frequency dependent. The rest of the components in the circuit (R1/C1 and R2/C2) are there to help eliminate the frequency sensitive output of the coupling elements over a reasonable range.

C5 is set with C1 and C2 at mid-range to provide a zero voltage output from the circuit on 20m when the line is terminated in a resistive load equal to the line impedance. Going to 10m, C2 is adjusted to restore balance and on 80m, C1 is used to restore balance. The adjustment has to be gone through several times before a good balance is achieved over the entire HF range. It would be possible to calibrate the meter with different value load resistors for different swrs. For high value loads, the meter would move off center in the opposite direction for low value loads. But, such calibration would only hold true for non-reactive loads.

Construction

Many conventional swr bridge circuits have been described over the last year or two in amateur literature where a toroid transformer, rather than a parallel wire to the center conductor of a coaxial line, has been used as a pickup element. The construction of such units applies equally well to the circuits shown here. It is only the circuitry following the pickup element that is really different, plus the added R1 pickup in Fig. 1, or the C5 pickup in Fig. 2. Note that it doesn't matter on which side of the toroid these pickups are placed. However, as in any rf circuit, all leads must be kept as direct and as short as possible. The circuits shown are for the HF range, but by changing the toroid and values of the bypass capacitors, they can be extended, by experienced experimenters, into the VHF range.

Chapter II

How to Measure RF Impedance

THE ANTENNASCOPE – AN EFFECTIVE TOOL

W. R. Carruthers VE3CEA

There are two types of antennas, commercial and amateur. A commercial antenna is generally designed for one frequency, has many acres of ground around it, no obstructions and miles of heavy copper cable buried underground to provide an "effective" ground. These antennas work as designed – very well. The amateur antenna, on the other hand, is just that – an amateur design and construction.

This antenna is subject to all ills, roof tops, buildings, trees, TV masts, house electric wiring, telephone wires and what not. It's a wonder they work at all! But they can be made to work and thousands of amateurs make them work. They make them work by pruning or lengthening the feeder cable and by using an antenna coupler. These are always empirical steps, the "let's cut and try and see what happens" method. How much better it would be, and a time saver too, if we tested our antenna systems electrically and *knew* what was happening and then could take intelligent action to put the whole antenna system into resonance.

This fact is well known – an antenna can only accept power and radiate properly when it is operating at its resonant frequency. This is no problem for the commercial people who operate at one frequency. The amateur, however, wants to "roam the band" and may wish to operate over frequencies hundreds of thousands of cycles wide, even megacycles wide. How can he do this with a fixed antenna system? The answer is, he can't! But he can construct an antenna system for a certain frequency and take the penalty of reduced radiation when he moves far away from it. However this actually works very well, because each amateur has his own particular part of a band in which he likes to operate – and his friends tend to stay there too. On this particular spot, the amateur works diligently to "put out a good signal."

The question arises – how can we make sure our antenna system is radiating well at the particulaf frequency we wish to use? One answer is to use electrical test equipment to show us what is happening on the whole antenna system, which includes the antenna and the feed line.

One of the most useful devices for this purpose is the rf bridge, generally called the Antennascope. It is a simple device, inexpensive to construct and very effective in results. It is usually powered by a grid dip oscillator. Such bridges should be used at the junction of the feed line and the antenna and will show the resonant frequency of the antenna itself and the radiation resistance at the feed point.

Making such measurement up in the air is a difficult thing for the average amateur and impossible for those whose antennas are supported at the ends. If we are willing, however, to accept a small degradation in results, we can use the *rf* bridge at the station end if we have a half wave, or multiple of a half wave, feed cable. At every half wave point on a feeder cable the voltage and current vectors are in phase, which simply means that the electrical condition seen at the end of the cable is repeated every half wavelength in the cable. We can use the *rf* bridge then, at the station end of the feed line, if we are willing to agree that the results will not be 100% but reasonably close to it. The results will be affected by all the various factors that affect amateur antenna resonance and these effects may give us some peculiar results, but they can be overcome and the final results may be quite valuable to us.

Let me give you an example to illustrate what I'm talking about and to show you how effective the use of the *rf* bridge can be: –

A friend of mine constructed a 40 meter inverted V antenna, held at the feed point 40' up on his beam tower, 66' legs down to supports which held the ends about 8' off the ground. Feed line was 100' of Twin Amphenol cable, velocity factor .68. The antenna was difficult to feed, swr was high, radiation was poor. He asked me to have a look (electrical) at it. I took my grid dip meter, *rf* bridge and vtvm.

The first thing done was to check the feed line length. 1/2 wave length at 7.1 MHz was 492 x .68/7.1 or 47.1 feet. Two 1/2 wave lengths (to get into the station) would be 94.2 feet.

The first conclusion was that the feed line was 5.8 feet too long.

Next Test No. 1 was made using the *rf* bridge with results as shown in **Fig. 1**, the results being shown in table form and also plotted in graphical form.

It was obvious from this graph that the antenna system was resonating outside the

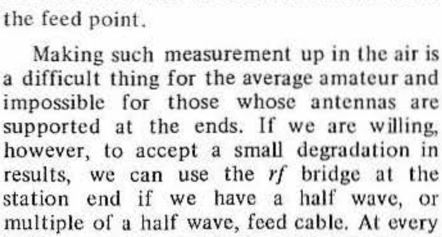

F MC	R OHMS
7.0	55
7.1	80
7.2	110
7.3	115

Fig. 1. 100' Feedline Test No. 1.

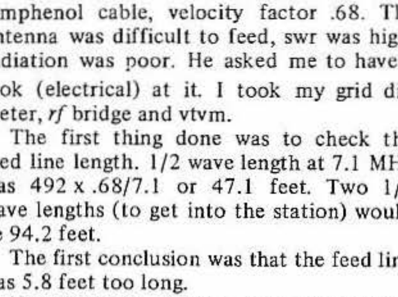

F MC	R OHMS
7.0	48
7.05	51
7.1	69
7.2	80
7.3	115

Fig. 2. 94.2' Feedline Test No. 2.

band as shown by the dotted lines. This test was repeated and the results were taken down to 6.4 MHz. They showed the system to be resonant at 6.6 MHz.

Test No. 2 was made next using the feed line cut to 94.2 feet. **Fig. 2** shows the results.

It was obvious the resonant point of the system was rising.

Test No. 3 was made next, cutting the feed line to 91.2 feet long. **Fig.** 3 shows the results.

F MC	R OHMS
7.0	43
7.1	46
7.2	76
7.3	81

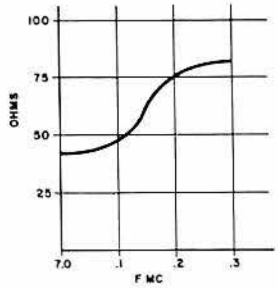

Fig. 3. 91.2' Feedline Test No. 3.

The resonant point was rising, but not far enough yet.

Test No. 4 was made using the feed line cut to 88.2 feet long. **Fig. 4** shows the results. It was obvious that we were very close to the resonant frequency of 7.1 MHz which my friend wished to use.

F MC	R OHMS
7.0	45
7.05	47
7.1	51
7.2	60
7.3	75

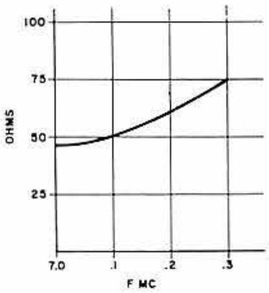

Fig. 4. 88.2' Feedline Test No. 4.

Test No. 5 was with 85.2 feet in the feed line. **Fig.** 5 shows the results.

F MC	R OHMS
7.0	78
7.1	61
7.2	60
7.3	66

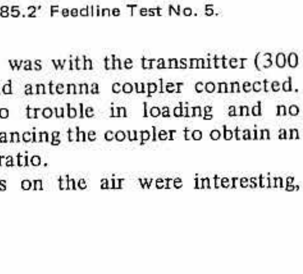

Fig. 5. 85.2' Feedline Test No. 5.

Test No. 6 was with the transmitter (300 watts CW) and antenna coupler connected. There was no trouble in loading and no trouble in balancing the coupler to obtain an swr of 1 to 1 ratio.

The results on the air were interesting, 5/9+ reports to the Eastern half of the U. S. A., 5/8 reports to Germany etc. *Conclusion:* The results shown above are not precise, nor can they be expected to be precise. There are too many unknown factors entering the electrical picture, such as those which required a shortening of the feed line, in this example, to somewhat less than a half wave length. But the bridge showed us the overall picture and suggested what was required to be done. The on-the-air results show that it was giving us a good picture and a result that was very satisfactory for my friend's needs.

Why not construct an *rf* bridge and check you own antenna system? I suggest it will pay off and be very informative to you, showing you what your antenna system looks like electrically and what to do to bring your whole system to the resonant frequency you wish to obtain. . . . VE3CEA

HOW TO BUILD AN ANTENNASCOPE

Paul Franson WA1CCH

A type of simple bridge used for measuring antenna impedance is called the Antennascope, shown in Fig. 1. This bridge is designed for low power operation – a grid dip meter usually gives plenty of power. It should be built very compactly with short leads. The potentiometer should be of high quality; an Allen-Bradley Type J is fine. The bridge can be calibrated with regular composition resistors. Simply connect the resistors in turn to the antenna terminal and adjust the pot until the meter reading dips to zero. Then mark the value of the resistor by the pot pointer. In use, the meter reading will not null completely except for resistive loads, so it will not read zero for reactive antennas. Nevertheless, the minimum reading will occur at the approximate impedance reading. Remember that all antenna bridges should be used between the antenna and the transmission line.

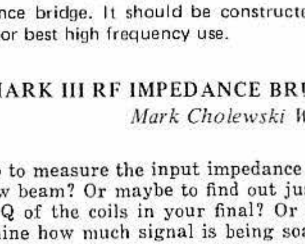

Fig. 1. The Antennascope is a simple antenna impedance bridge. It should be constructed compactly for best high frequency use.

THE MARK III RF IMPEDANCE BRIDGE

Mark Cholewski W6CRT

NEED to measure the input impedance of that new beam? Or maybe to find out just what is the Q of the coils in your final? Or even to determine how much signal is being soaked up by your coax? If you ever want to do these, or any similar jobs, then the Mark III RF Impedance Bridge is the thing for you.

You can build it for a total cost of about $30 (exclusive of sheet metal) and, if you follow instructions closely, it will be accurate to closer than 10 percent throughout its operating range. Unlike the more common resistive-bridge and reflectometer methods of measuring impedance, the Mark III operates equally well at resonance or far away. It will measure both resistance and reactance present in resistors, capacitors, inductors, antennas, and transmission lines at any frequency between 2 and 30 mc.

Before we start into the actual construction of the Mark III, one thing must be emphasized. Accuracy can be assured only if the components, circuit, and parts layout are absolutely duplicated. The original instrument's calibration was obtained through tedious laboratory techniques. If you make any changes, the calibration curves will no longer apply. However, if instructions are followed to the letter you need have no worries about accuracy. A test model, built by W6BJU following these instructions, checked out to 2 percent accuracy at 2 mc and 10 percent at 30 mc.

Construction of the Mark III divides into three major sections: Preliminary metalwork, actual wiring, and calibration. Each will be described separately. Ready? Let's go!

Preliminary Metalwork

1. Cut, drill, and bend to shape from soft aluminum shields S1, S2, and S3 as shown in Figs. C1, C2, and C3.
2. Cut, drill, and tap plexiglas insulators I1, I2, and I3 from bulk rod stock as shown in Figs. C4 and C5. When tapping plexiglas, use water as lubricant.
3. Cut, drill, bend, and solder tubular shields S1A, S2A, and S3A as shown in Fig. C6. Copper or brass may be used; aluminum should be avoided because of solding difficulties.
4. Assemble shielded resistor assembly R2/S4 as shown in Fig. C7. The copper tubing must be drilled out to clear the body of R2. When soldering, hold the assembly in a vise to protect R2 from excessive heat.
5. Cut, drill, bend to shape, and solder box shields S1B, S2B, and S3B as shown in Fig. C8.
6. Drill S5 (a 3x4x6 LMB unpainted chassis box) as shown in Fig. C9.
7. Cut, drill, and bend to shape shield partition S5A as shown in Fig. C10.

Shielded Transformer

While classified under the "preliminary metalwork" section for reasons which will become obvious, construction of the shielded transformer is the most critical part of the entire project. Before proceeding, read and re-read steps 8 through 24 and be sure that you understand them fully. Take special care when soldering—three transformers were built for the original instrument before a non-shorted one was achieved.

8. Cut to length, drill, and tap transformer mounting insulator I4 from 5/16-inch plexiglas rod.

9. Cut and drill two bobbin-end washers as shown in Fig. C12 and solder them to a length of copper tube as also shown. Then cut halfway through the bobbin with a hacksaw.

10. Pull the shielding from an 18-inch length of RG58/U. Save 8 inches for step 19 and use the rest in step 11.

11. Solder one end of the 10-inch shielding into the U-shaped slot on the bobbin end, using an aluminum rod or small drill to keep the inside of the shielding open. It must pass

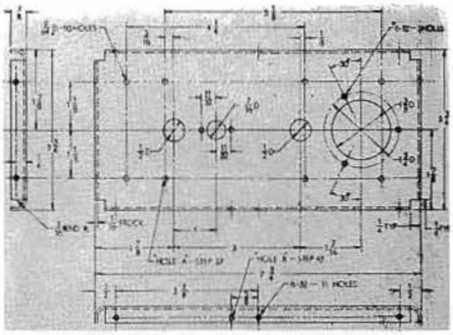

C1—Shield S1 (1 REQ), Mat'l Aluminum

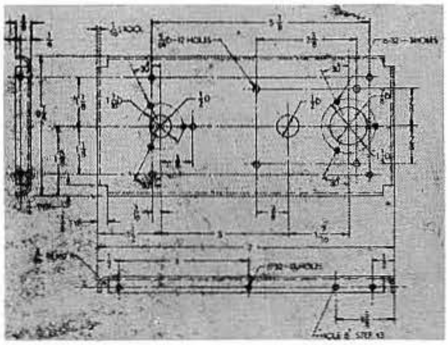

C2—Shield S2 (1 REQ), Mat'l Aluminum

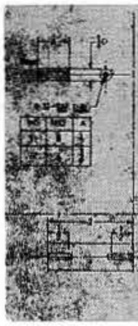

C4 — Insulators, Mat'l Plexglas rod
C5 — Insulator I3 (8 REQ), Mat'l Plexiglas rod

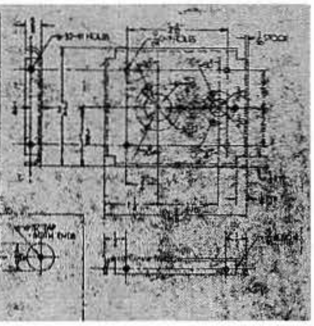

C3—Shield S3 (1 REQ), Mat'l Aluminum

No. 26 plastic-covered wire. Clean off all metal burrs and solder splatters.

12. Wind 48 turns of No. 26 plastic-covered hookup wire on the bobbin in three lays of 16 turns each. Solder the start of the winding at point A (see Fig. C12) and wind in the direction shown by the arrow.

13. The last turn will end at the shielding attached in step 11. Feed the free end of the wire into the shielding, draw the turns tight, and secure the winding with plastic tape.

14. Cut a piece of brass or copper shim stock as shown in Fig. C13 to a length which will wrap around the bobbin but will not allow the ends of the shim stock to touch each other.

15. Tin the shim stock along the edges.

16. Place the wound bobbin in a vise, wrap the shim stock around it, lining up the free ends of the shim with the slot cut in step 9, and solder the shim to the bobbin ends. Caution. *Do not overheat the winding; the plastic covering melts easily and a short is almost impossible to detect.*

17. Connect an SO-239 coax connector and UG-177/U hood to the shielded primary lead as shown in Fig. C15.

18. Wind one turn of ¼-inch diameter half-hard copper tubing around a 1-inch diameter form. Saw the tubing as shown by "phantom lines" in Fig. C14. Drill as shown in Fig. C14 and clean off all burrs.

19. Locate the 8-inch piece of shielding left over from step 10.

20. Using same technique as in step 11, solder one end of the shielding into the ⅛-inch hole. Clean off all solder splatter and burrs.

21. Solder one end of another length of No. 26 plastic-covered hookup wire to point A (see Fig. C14) and wind three turns inside the tubing in the direction shown.

22. Feed the free end of the hookup wire through the shielding. Bend the tubing into final shape as shown in Fig. C14. Pull up the three turns snugly, making sure that the plastic coating is undamaged.

23. Connect a male phono plug to the free end of the shielded wire as shown in Fig. C15. Length of the wire is critical.

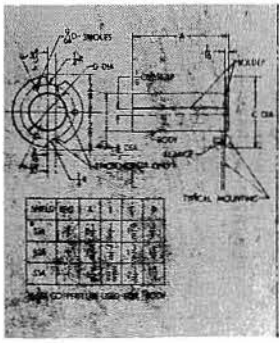

C6—Tube shield, Mat'l .010 Brass

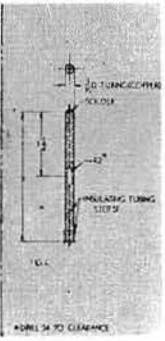

C7 — Shield Assembly S4

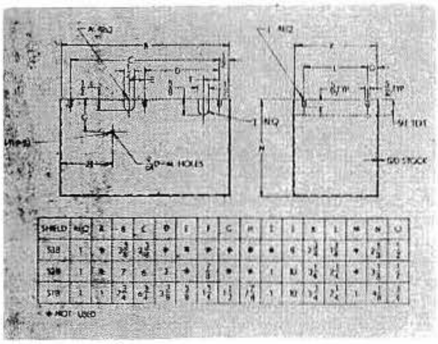

C8—Shield can, Mat'l Aluminum

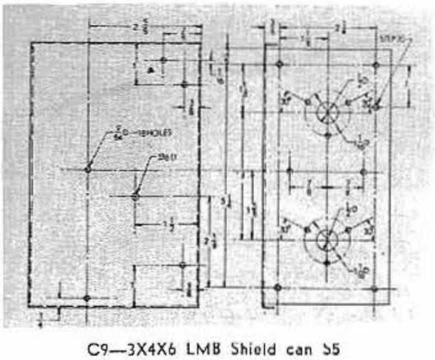

C9—3X4X6 LMB Shield can S5

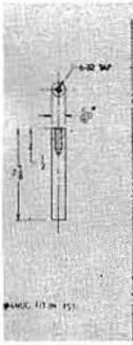

C11 — Insulator I4, Mat'l Plexglas

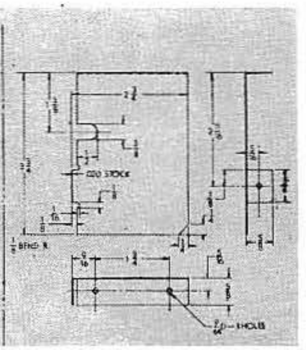

C10—Shield Partition S5A, Mat'l Aluminum

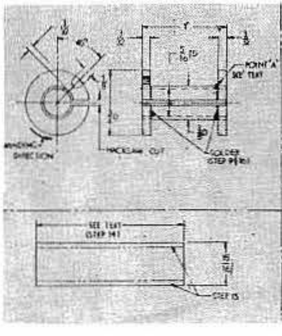

C13—Shield PS1A, Mat'l .005 Brass or Copper
C12 — Primary bobbin PS1, Mat'l Brass or Copper

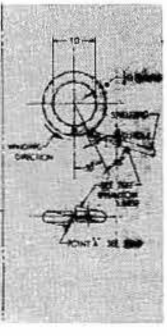

C14—Shield PS2, Mat'l copper

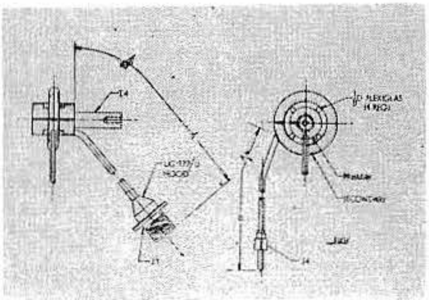

C15—Transformer assembly

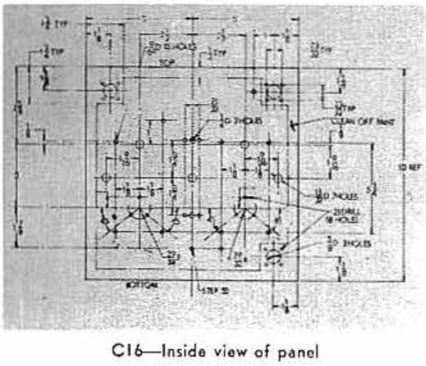

C16—Inside view of panel

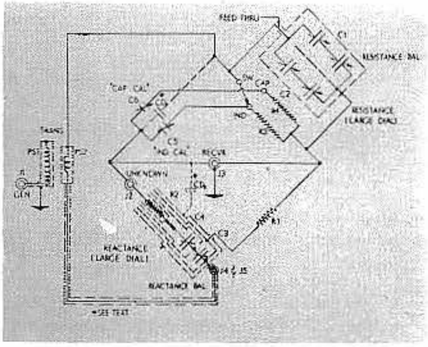

Fig. 1—HF. RF. Bridge Mark III Schematic

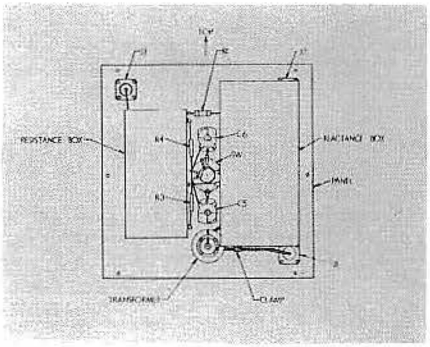

Fig. 2—Inside view

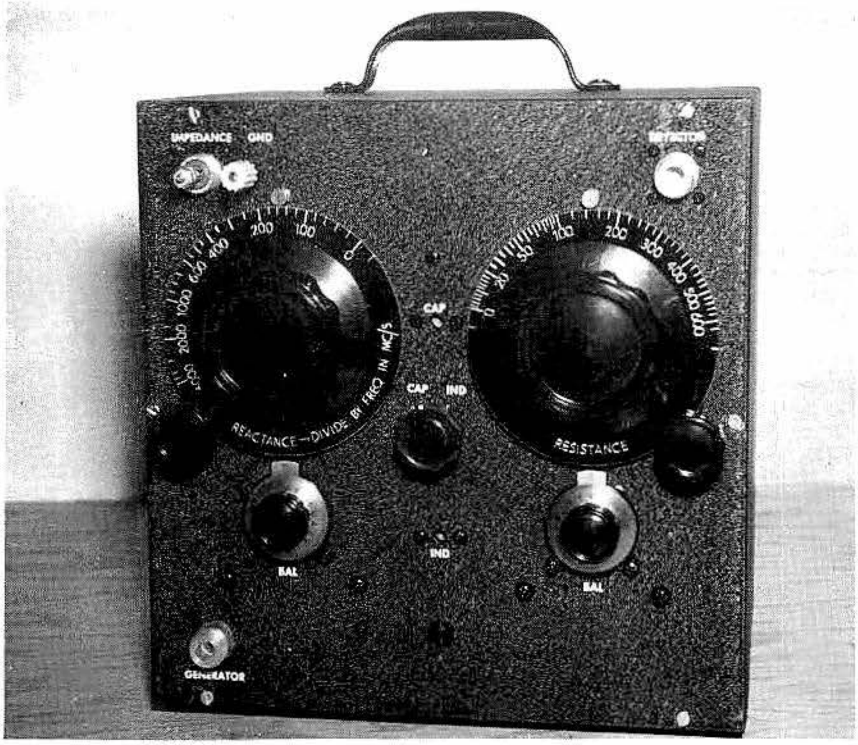

24. Using four pieces of ⅛-inch diameter plexiglas rod as spacers, assemble the primary and secondary shielded windings as shown in Fig. C15. Attach insulator I4 to the transformer by cementing it into the bobbin hole with Duco. Cement both windings to spacers with Duco and allow to dry overnight. This completes the transformer.

25. Remove top and bottom from the 8x10x10 utility box. Remove all paint from flanges; clean to bare metal to provide adequate rf shielding on reassembly.

26. Remove paint from inside of bottom plate for ½ inch in from each edge.

27. Remove paint from inside of top plate as shown in Fig. C16 by "phantom lines."

28. Drill the top plate as shown in Fig. C16.

Note that drawing shows INSIDE surface of plate.

29. Cut shafts of all four variable capacitors to ⅜-inch length. Remove all trimmer capacitors.

30. Tap the threeholes on the face of each capacitor, using a 6-32 tap. Take care not to damage the first stator plate; a bottom tap may be necessary. Attach three type I2 insulators to C1 and C2, using ¼-inch-long 6-32 set screws as shown in Fig. 3. This completes preliminary metalwork.

Actual Wiring

31. Attach C1 and C2 to shield box S5 using six ¼-inch-long 6-32 machine screws. Connect stator lugs of C1 to those of C2 with No. 18 tinned wire, as shown in Fig. 3.

32. Mount two soldering lugs as shown in Fig. 3 and connect remaining stator lugs to them, using No. 18 tinned wire.

33. Press the Erie CF-408 feed-thru into the 0.136-inch diameter hole in S5. Solder a short No. 18 tinned lead from the inside terminal of this insulator to the wire installed in step 31, as shown in Fig. 3.

34. Attach shield partition S5A, using three ¼-inch-long 6-32 machine screws. One of the screws installed in step 32 must be temporarily loosened and removed.

35. Attach four type I3 insulators to shield box S5 as shown in Figs. 3 and C9.

36. Attach three more soldering lugs to S5 as shown in Fig. 3, using ¼-inch-long 6-32 machine screws.

37. Attach four type I3 insulators to shield platform S1, using the 4¼-inch-spaced holes in S1 and 6-32 screws. Attach S1A to S1, using 6-32 screws from the inside of S1. Attach the female phono socket to S1 in the 7/16-inch diameter hole. Attach four type I1 insulators, using ¼-inch-long 6-32 screws. Do not tighten the screw in the hole marked "Hole A" in Fig. C1; this screw will hold a cable clamp later. See Fig. 4 for details of insulator placement.

38. Attach four type I1 insulators to shield platform S2, using the 2⅜-inch-spaced holes in S2 and ¼-inch-long 6-32 screws. Attach S2A to S2. Attach capacitor C3 with 1-¼-inch-long 6-32 screws, using two nuts on each screw as shown in Fig. 4. Mount a soldering lug under one nut as shown. Align the capacitor by adjustment of the mounting screws and nuts.

39. Solder a No. 18 tinned wire to the female phono socket and pass the wire through the corresponding hole in S2. Attach S2 to S1 with ¼-inch-long 6-32 screws going into the type I1 insulators attached to S1 in step 37.

40. Attach C4 to shield platform S3 using ⅝-inch-long 6-32 screws with dual nuts (same as in step 38). Attach S3A to S3.

41. Connect two of the stator lugs of C4 with No. 18 tinned wire as shown in Fig. 4. Slide shield assembly S4 into S3A. Center assembly S4 in S3A, using a piece of ¼-inch-long insulating tubing. Make certain that opposite ends of S4 and S3A are even as shown in Fig. 4, and cement tubing in place with Duco. Solder the shorted end of resistor R2 (which is in S4) to the wire connecting stator lugs of C4.

42. Attach shield platform S3 to S2, using ¼-inch-long 6-32 screws going into the type I1 insulators installed on S2 in step 38.

43. Place a ¼-inch-long 6-32 screw in the flange of S3 as shown in Fig. 4, with a soldering lug. Connect this lug to the stator lug of C3 with No. 18 tinned wire. Using ¼-inch-long 6-32 screws, mount a soldering lug in Hole B (see Fig. C1) of S1 and another in S2 as shown in Fig. C2. These lugs are mounted in a direction opposite to that of the platform flanges.

44. Attach the "unknown" ground lug to the panel next to the "unknown" coax connector hole, as shown in Fig. C16. Attach the "receiver" coax connector to the panel, from the inside. Attach the "IND-CAP" switch to the center of the panel, using an extra nut to position the switch as far as possible from the panel. Orient the switch as shown in Fig. 5. Place a soldering lug under each nut. Mount C5 and C6, using ⅜-inch spacers between the capacitor frames and the panel. Connect the

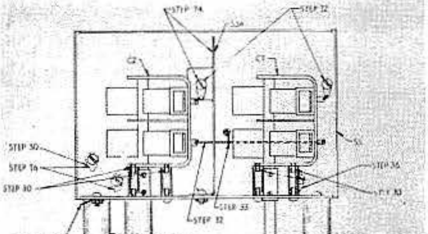

Fig. 3—Resistance box—Inboard view

rotors of C5 and C6 to the lugs of the switch with No. 18 tinned wire as shown in Fig. 5.

45. Solder a 2½-inch length of No. 18 tinned wire to the "COMMON" terminal of the switch. Connect the stators of C5 and C6 to the remaining switch terminals, as well as resistors R3 and R4. Complete connections are shown in Fig. 5.

46. Attach the "unknown" coax connector to the panel, placing a soldering lug under one mounting screw. Connect the "unknown" ground lug to this soldering lug to provide a good bond.

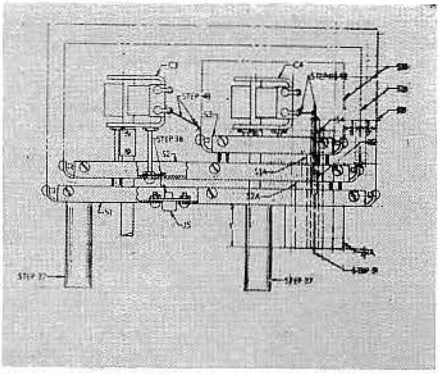

Fig. 4—Reactance Assembly—Inboard view

47. Mount panel bearings for capacitors C2 and C4. Mount the special panel bearings furnished with the Johnson Vernier Dial assemblies in place. Attach the two large dials to dummy shafts and mount the dial indicator in the position you prefer. Remove the large dials and dummy shafts after placing the dial indicators.

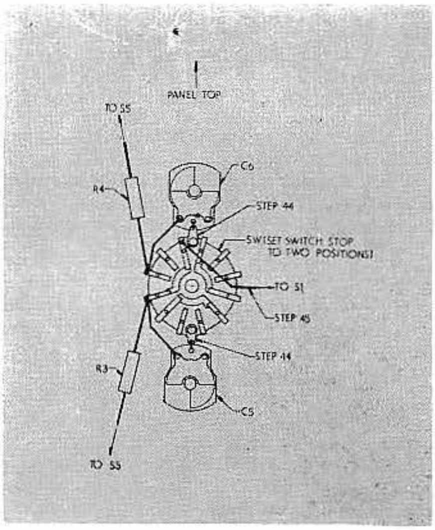

Fig. 5—Switch Assembly

48. Mount S5 at the left side of the panel (as shown in Fig. C16) and mount the assembly of S1, S2, and S3 at the right side. See Fig. 2.

49. Attach shaft couplers to the four capacitors. Cut the plexiglas shafts to length and mount them in place. Attach the large vernier dials. Set the dial of C2 so that it reads "0" at minimum capacity. Set the dial at C4 so that it reads "100" at maximum capacity. Mount the two Calrad dials on the panel, setting them so that they both read "0" at maximum capacity. Attach the knob to the "IND-CAP" switch.

50. Connect resistor R1 (270-ohm deposited-carbon) from the soldering lug on S2 to the lug in line on S5. Connect R3 (220 ohms) and R4 (100 ohms) to the soldering lugs on S5 which were installed in step 36. R3 will be the resistor nearest the bottom of the panel.

51. Connect a No. 18 tinned-wire lead from the "detector" coax connector to the lug on S5. Connect a No. 12 (note different wire size) lead from the "unknown" connector to the free end of R2. *Make sure that R2 is not shorted to any shield.*

52. Attach the transformer, completed in step 24, to the bottom of the panel as shown in Fig. 2. Attach the coax connector connected to the transformer to the panel in the "signal generator" hole. In the hole between the transformer and the connector, mount a cable clamp to hold the shielded primary lead. Secure the shielded secondary lead with another clamp held by the loose screw installed in step 37.

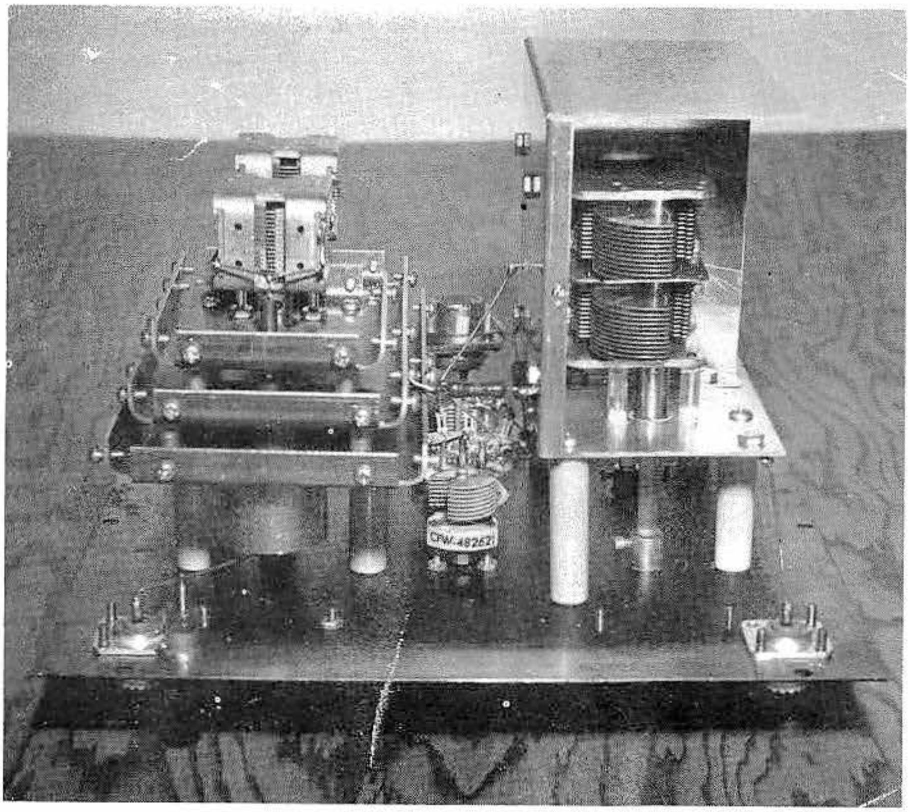

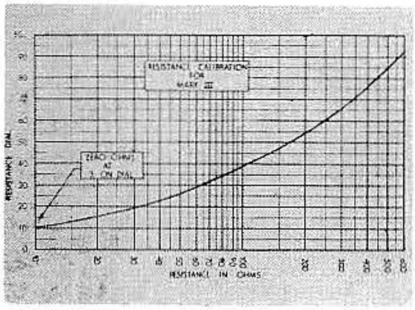

Fig. 6

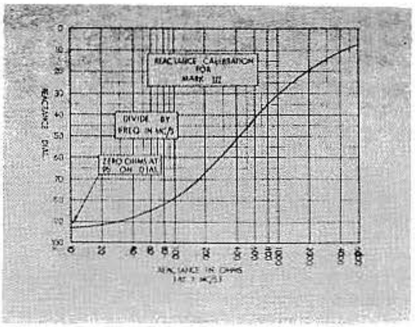

Fig. 7

53. Mount a soldering lug to S1B with a ¼-inch-long 6-32 screw, placing the lug in the direction of the box opening.

54. Place ¼-inch-long 6-32 screws in all remaining tapped holes in S1, S2, and S3. Starting with S3, place all shield boxes in place and secure screws. Connect a No. 18 tinned lead from the soldering lug installed in step 53 to the Erie feed-thru on S5. Set C5 and C6 to mid-capacity.

55. Attach four rubber feet to the bottom plate of the case and four more to the bottom side of the utility cabinet. Restore the front plate in place and secure with the sheet-metal screws provided. This completes construction of the bridge. After calibration, it will be ready for use.

Calibration and Use

Since the Mark III is a null-type instrument (adapted from the Schering bridge circuit) it can only be used with a signal generator and a detector. Both must be shielded; however, a Heath SG-8 will do nicely as the signal generator and any decent communications receiver will serve as the detector. For best results, it should be calibrated with the signal generator and receiver with which it will be used.

56. Connect the bridge to the signal generator and the receiver, using coax cable from the panel connectors to each.

57. Set both the signal generator and the receiver to 2 mc.

58. Short the "unknown" terminal and ground terminal using a banana plug in the connector or a coaxial short made by soldering the pin of a PL-259 connector to the shell through a copper disc.

59. Set the "IND-CAP" switch to "CAP."

60. Set the "Reactance" dial to 15.

61. Set the "Resistance" dial to 5.

62. Tune the signal generator, only, for maximum signal in the receiver.

63. Using the two small "balance" dials, reduce the signal in the receiver to the lowest level possible. This is the "null" or "balance" condition. The dials will interact, and multiple adjustment will be necessary. However, all receiver and signal generator adjustments must be left alone during this step.

64. Replace the short across the "unknown" connector with the 620-ohm deposited-carbon "test" resistor.

65. Using only the large "resistance" and "reactance" dials, null out the signal once more. Record the final reading of the "resistance" dial. If this reading is 95, you are extremely lucky and the first half of the calibration is complete. If not, proceed with step 66.

66. If the "resistance" dial reading is less than 95, increase the capacity of C6 slightly and repeat steps 60 through 65. If the reading is larger than 95, decrease the setting of C6 and repeat steps 60 through 65. Continue this process until the reading comes out at 95.

67. When capacitive calibration is complete, set the "IND-CAP" switch to "IND," and the "reactance" dial to 30.

68. Repeat steps 60 through 65; if the final reading of the "resistance" dial is other than 95, proceed with step 69.

69. If the reading is less than 95, increase the capacity of C5 slightly and repeat steps 60 through 65. If the reading is larger than 95, decrease the setting of C5 and repeat. Continue until the reading is 95. This completes calibration, and the curves shown here can be used for readings.

Using the Mark III

To use the Mark III, set the instrument up as described in step 56 of the calibration procedure. Set the signal generator to the desired frequency and short out the "unknown" terminal of the bridge as described in step 58.

If you're measuring an inductive impedance, set the "IND-CAP" switch to "IND," the reactance dial to 95, and the resistance dial to 5. Null the signal with the small balance dials.

Now remove the short and connect the unknown. Rebalance the bridge using the large dials, note the readings, and convert the readings to ohms by use of the calibration curves.

These initial values must be corrected. The reactance value is corrected by dividing the value read from the curve by the frequency (in mc) at which the reading was taken. The resistance value is corrected by obtaining the correction factor due to frequency from Fig. 8 and multiplying the reading by this factor.

If you're measuring a capacitive impedance, set the switch to "CAP." Then estimate the reactance of the unknown and set the large reactance dial to a value just larger. Set the resistance dial to 5 and short out the "unknown" connector. Null the signal with the short and again with the unknown, the same as for an inductive impedance.

Correct the resistance reading in the same manner as for an inductive impedance. However, the reactance reading is corrected differently: Subtract the value taken from the calibration curve from the value originally set on the reactance dial. Now divide this remainder by the frequency (in mc) at which the reading was taken. The result is the true value of capacitive reactance.

A few minutes' practice will make operation of the Mark III far more simple than the detailed directions would indicate; in practice, you can make a reading in less time than it takes to read these paragraphs.

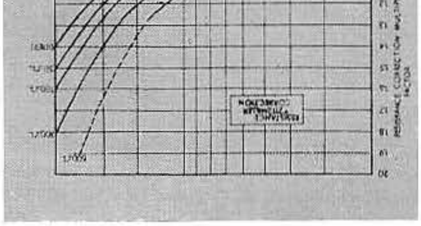

Fig. 8

Parts Substitutions and Design Changes

Since few hams are content to build a "Chinese copy" of someone else's design, a few words on the effect of changes are necessary.

Naturally, the Mark III doesn't represent the only possible—or even necessarily the best—way in which such a bridge can be built. Any part, or all, may be changed. However, any such change will invalidate the calibration curves, and is not recommended unless the builder has access to laboratory equipment.

Even then, before making any substitutions, these three design articles should be read and fully understood:

"A Radio-Frequency Bridge for Impedance Measurements From 4000 KC/S to 60 MC/S," D. B. Sinclair, Proceedings of the I.R.E., November, 1940, pages 497-502.

"A High Frequency Model of The Precision Condenser," D. B. Sinclair, General Radio Experimenter, October-November, 1938, pages 1-7.

"The Effect of Stray Capacitances to Ground in Substitution Measurements," M. Reed, Wireless Engineering, May, 1936, page 284 ff.

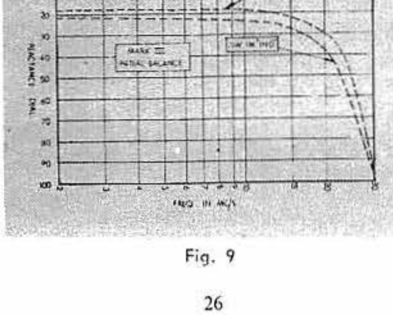

Fig. 9

Parts List

Quan.	Item	Description	Cat. No.
1 ea.	C1, C2	Miller, 10-365 mmf/sect.	2112
1 ea.	C3, C4	Miller, 10-365 mmf	2111
1 ea.	C5, C6	50 mmf air padder (good quality)	APC-50
1 ea.	J1, J2, J3	SO-239 Panel coax connector	
1	J4	Phono plug	
1	J5	Phono socket	
1	R1	Aerovox 1% 1/2W 270 ohm "Carbofilm"	Type-CP
1	R2	330 ohm 5% 1/2W Carbon Resistor	
1	R3	220 ohm 5% 1W Carbon Resistor	
1	R4	100 ohm 5% 1W Carbon Resistor	
1	Test	Aerovox 1% 1/2W 620 ohm "Carbofilm"	Type-CP
1	SW	Centralab 3p; 1 Sect; 2-3 Position (reset stop to two positions only)	2506
*2	Dials	E. F. Johnson Vernier (0-100 CW)	116-285-1
2	Brgs	E. F. Johnson Panel Bearings	115-255
4	Cplrs	Metal ¼ shaft couplers	
3	Rod	¼ inch diameter plastic rod x 12"	
1	Rod	⅛ inch diameter plastic rod x 6"	
2	Rod	⅜ inch diameter plastic rod x 12"	
2	Dials	"Calrad" 1-13/32 inch Vernier Dial	VD-36
1	BP	Binding Post (ground lug)	
1	Case	Bud carrying case 10"x10"x8"	CC-1100
1	Box	LMB box (plain) 3" x 4" x 6"	
1	Hood	UG 177/U coaxial cable hood	
1	Insul.	Erie feed-thru insulator	CF-408
1	Short	PL-259 coaxial male connector	
2	Clamps	Walsco clamp for ⅛ to 3/16 cable	7502-N

Of the above list the critical parts are:
C1, C2, C3, C4, J2, R1, Test resistor, Case, LMB box, UG 177/U, Erie feed-thru.

---

* This author hand engraved these dials per Figs. 6 and 7.

MEASURE RF IMPEDANCE WITH YOUR SWR BRIDGE

Warren McDowell W2AOO

So far, the great majority of antenna impedance bridges that we have found in construction articles are devices that function only with a low power rf source. The run-of-the-mill impedance bridge is designed to operate with a grid dipper as the source of rf excitation. Operation with tube type dippers is generally intended as the transistor dippers produce an rf level that is too low for excitation of this bridge type.

The conventional antenna bridge cannot be left in the transmission line continually as excessive rf energy would soon destroy the device. This means that each time measurement of antenna impedance is desired, the transmission line must be opened and the bridge inserted and grid dip excitation applied. Grid dippers are not necessarily the most accurate rf source for a specific frequency in an amateur band . . . therefore the station receiver must monitor the dipper output for any bridge accuracy. A low power bridge will not often present the true operating impedance of the antenna . . . especially antennas with parasitic elements. A bridge that operates under full transmitter power will present a much more accurate picture of your antenna system at a specific frequency.

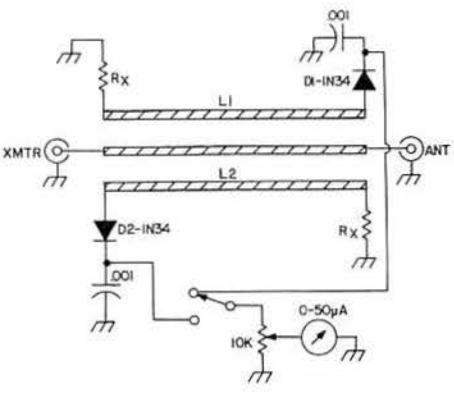

Fig. 1. Conventional swr bridge.

Inspiration for the "In Line" full power bridge came from information concerning the standard swr bridge. Just about every amateur has in his possession some sort of swr bridge and the great majority are of the type illustrated in Fig. 1. This bridge consists of a section of transmission line near which are placed two inductors. These inductors are actually two bridges along with their associated diodes and resistors. One of the bridges reads forward power and the other reflected power. The resistors (Rx) at the end of the inductors L1 and L2 are critical for accurate bridge null (balance) and therefore must be the proper value for the specific transmission line used. For the average swr bridge the value for Rx is 100Ω for 75Ω line and 150Ω for 50Ω transmission line. Considering that resistor Rx is critical for the impedance of the line in use, varying the value of Rx and devising a system of calibration for Rx would enable us to determine the impedance of a line when a null is achieved on the bridge meter.

The "reflected" inductor which is L1 in Fig. 1 is the portion of the bridge circuit we are interested in for impedance measurements. The value of Rx and the transmission line must balance the bridge for a null to be realized. Any variation from the above parameters will mean changing the value of Rx so that the bridge again balances at a new impedance value.

By experimenting with various values of resistance at Rx, it was determined that a 1000Ω potentiometer represents a fair value. The 1000Ω potentiometer is inserted in place of Rx on inductor L1 (see Fig. 2). This is the inductor with the diode pickup located toward the load or antenna end of the swr bridge.

Make sure that all leads to the 1000Ω potentiometer are short and that the metal case (shell) of the potentiometer is well grounded. Excessive lead length or inductance will create inaccuracy of the device. The position of the potentiometer will be determined by the physical layout of your particular swr bridge. It must be set at a point where the shaft can be extended through the front panel of your swr bridge. Allowance must also be made for a dial or other indicating device which can be calibrated in ohms (impedance) on the front panel. It might even be desirable to mount your entire present bridge in another larger case so that all functions can be accommodated.

Calibration of this in-line bridge was the major problem. An ordinary grid dip meter will not provide sufficient excitation for readings. With full power applied, especially a kilowatt, it becomes difficult to find resistive dummy loads of various values to calibrate the bridge. Even with 100W of rf, proper resistive load values are not common.

The solution to the calibration problem came to us in the form of an (ouch!) CB transmitter. A CB transmitter is fortunate if it is able to put out 3W of rf and at the same time is well within the frequency range of an swr bridge. The most important fact is that a CB transmitter will provide adequate excitation for calibration of the bridge with ordinary 5W 5% carbon (garden variety) resistors. For calibration, a good assortment of these resistors is necessary. Use values such as 5, 27, 47, 75, 100, 150, 220 and 470Ω. Intermediate values can be then interpolated on your scale. The calibration procedure is simple . . . first borrow your neighbor's CB, then attach the 5W resistors across the antenna coax connector of the bridge and excitation of the CB transmitter is applied to the remaining connector on the bridge. The bridge sensitivity should be set for a middle scale reading of the meter and the 1000Ω potentiometer is varied until you reach a null on the meter. Mark the value of the calibration resistors on the potentiometer scale (dial). Do this for all of the available resistors and your bridge will be in fair calibration.

At this point we should mention that this system does not measure reactive components in the antenna system. If your antenna is reactive, either inductive or capacitive, the meter will present a shallow, poorly defined null at the operating frequency. A sharp, well defined null will indicate a purely resistive impedance.

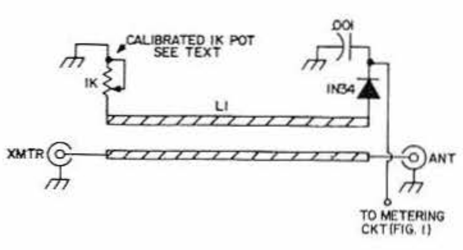

Fig. 2. The modified bridge leads to 1K pot should be as short as possible and shell (case) of pot grounded.

When using the bridge in its former function as an swr bridge, set the resistance dial to the value of your transmission line. When measuring impedance, vary the dial for maximum dip on the meter and read the resistance (impedance) directly.

As a final point, it is wise to insert the swr/impedance bridge at a half-wave or an even multiple of a half-wavelength from your antenna. At half-wave points from the antenna, the antenna impedance is repeated. This will enable your measurements to be much more accurate. When determing half-wavelength points, take into consideration the velocity factor of your particular coax.

USE YOUR GDO AND Z METER

Denys Fredrickson W0BMW

The GDO is one of the most versatile pieces of test equipment available. Yet there are many hams who don't know how or when to use one. The writer will try to describe and explain some of its various functions.

The GDO is basically a variable high frequency oscillator with a frequency range of approximately 550 kc to 250 Mc. It may also be used as a diode detector or wave meter. The GDO gets its name from the fact that a meter measures the grid current and when the oscillator circuit is coupled to a resonant circuit a reduction in grid current is obtained. This is called the grid dip. However, when it is used as a wavemeter and coupled to an rf source, an increase in current is obtained at resonance.

The GDO and impedance meter can be used to accomplish the following:

1. Determine the resonant frequency of tuned circuits, including antennas.
2. Determine the impedance of circuits, receiver inputs and antennas.
3. Determine the length of half-wavelength or quarter-wavelength transmission or tuning stub lines.
4. Determine the "Q" of a circuit or component with the aid of a VTVM.
5. Determine the resonant frequency of individual coils, capacitors or crystals that are within the range of the GDO.
6. Determine the rf frequency of energized circuits.
7. Monitor a radiated rf signal with the aid of headphones.
8. Neutralize rf stages.
9. Locate parasitic oscillations.
10. To align receivers and television sets.
11. Determine where BCI and TVI is entering the radio or television receivers.
12. Determine unknown inductance.
13. Determine unknown capacitance.

Now if you will step into the lab we will try to demonstrate how these instruments can be put through their paces. Let's begin with the simple functions and then gradually creep up to those which are more complex so they don't scare us before we get started.

An oscillator-detector

Simply plug in a pair of headphones (if GDO has facilities for them) and "zero-beat" with the radiating signal. This then will be the frequency of the radiating signal.

Crystal frequencies

Connect a one turn loop of wire across the crystal and couple the GDO close enough to get a dip of the meter when resonance is obtained. It is always wise to check lower frequencies to be sure it is the fundamental frequency that is being indicated.

Frequency determination

Generally the GDO has a switch which is used to remove the plate voltage from the tube. The tube will then serve the function of a diode and the meter as a diode load. When a peak deflection of the meter is obtained this will indicate the frequency of the radiating signal.

Resonance of an RF choke

When an rf choke is used as a parallel or shunt fed circuit, it must be free of self resonance over the operating frequency range of that circuit or it may burn up. The popular pi tank circuit is an example. Place a short circuit across the choke and then determine its self resonant frequency by coupling the GDO close enough to indicate a dip on the meter when the resonant frequency is obtained.

Neutralization

Apply plate power to the exciter stages and filament power only to the stage being neutralized. Use GDO as a wavemeter and couple close to the tank coil in the stage being neutralized. Vary the frequency of the GDO until maximum reading is obtained and then adjust the neutralization for minimum GDO meter reading. The circuit being neutralized may have to be retuned and the above procedure repeated with a closer coupling of the GDO to the Tank coil.

IF alignment

Tune the GDO to the desired frequency and couple it close to the if coil to be aligned. Adjust the if coil until a dip is observed on the meter. The if coil will then be tuned to the desired frequency.

Inductance and capacitance checking

To determine the value of an unknown capacitor, connect it across a known inductance and use the GDO to find the resonant frequency of the circuit. With these known values a reactance chart will give the value of the capacitor. Some GDO's supply a chart which corresponds to the coils supplied with the GDO as the known inductances. To determine the value of an unknown inductance, connect a known capacitance across the coil and use the GDO to find the resonant frequency. Again, the reactance chart may be used or the following formula (which may be used for either inductance or capacitance) for resonant circuits:

$$L = \frac{1}{39.48\ (f^2)\ C} \quad \text{or} \quad C = \frac{1}{39.48\ (f^2)\ L}$$

Where f = cycles per second
L = inductance in henries
C = capacitance in farads

The inductance of an air core coil can be estimated by the following formula:

$$L = \frac{(rN)^2}{9r \neq 10w}$$

Where L = inductance in microhenries
N = number of turns
r = radius of coil in inches
w = length of coil in inches

Q measurements

Connect a condenser across the coil so the tank circuit resonates at the desired frequency. Connect a VTVM across the tuned circuit and tune the GDO until maximum reading is obtained on the VTVM. The GDO coupling may be changed until a convenient value is obtained on the VTVM and then it must not be moved during the remainder of the test. Note the resonant frequency f_c then detune the GDO to a lower frequency until the VTVM reads 70.7 percent of its original or peak value and call this frequency f_1. Now detune the GDO to a higher frequency until the VTVM again reads 70.7 percent of its original or peak value and call this frequency f_2. The Q is then calculated by using the following formula:

$$Q = \frac{f_c}{f_3}$$

Where f_c = is the center of resonant frequency
f_3 = the difference between f_1 and f_2

Parasitic oscillations

By using a pair of headphones with the GDO, the parasitic oscillation frequency may be determined. Turn the power off of the stage being checked and then use GDO to find the circuit which resonates at the parasitic frequency by moving the GDO slowly around the wiring. When a "dip" is observed, moisten the finger and touch an ungrounded point of the circuit. If a change in the dip is observed, it indicates that it is the portion of the circuit that would be a likely suspect.

BCI and TVI locator

Most of the BCI and TVI problems can only be resolved at the receiver, either by installation of filters, resistors or condensers or a combination of all three. The problem is—where is the rf entering the receiver? Use the GDO tuned to the frequency which produces the greatest amount of interference. Probe around with the GDO until the most sensitive spot is located, which is indicated by watching or listening to the receiver interference. After the point of entry is determined then the appropriate corrective action can be accomplished.

Antenna measurements

Space does not permit to discuss all types of antennas and adjustments so only a few will be mentioned to give some idea on the use of the instruments. At this point it should be mentioned that inductive type coupling should be used between the GDO and antenna when checking near the current maximum point and capacitive coupling when checking near the voltage-maximum point.

The beam antenna has gained tremendous popularity in recent years plus many headaches for those striving to obtain the maximum effectiveness. Most of the headaches can virtually be eliminated by using the GDO and Impedance meter (Z-meter). Let's take a look at a 3 element yagi and see what has to be done to obtain a good adjustment. The element lengths must be physically adjusted or electrically loaded to obtain resonance at the desired frequencies and the feed point impedance must match the impedance of the transmission line. These two points are not the only considerations for beam adjustment but they are the most important factors. The GDO can be inductively coupled to each element and the elements adjusted until each one is resonant at the desired frequencies. It is best to make the measurements while the antenna is in operating position. This is very difficult to do in many cases but let's assume you can. After the elements have been adjusted, the feed point must be adjusted to match the line. The Z-meter and GDO will be used to accomplish this adjustment. The Z-meter is basically a resistance type bridge with a calibrated potentiometer as one of the bridge arms. Connect the Z-meter directly to the antenna feed point. Couple the GDO to the Z-meter inductively thru a couple loops of wire connected to the other terminals of the Z-meter. Tune the GDO to the resonant frequency of the beam and adjust the Z-meter to the dip or null. If the impedance indicated by the Z-meter is not the same as the transmission line then readjust the matching network and redip the Z-meter until the impedances are equal.

Now—if you can't adjust the antenna in the operational position you still can determine the resonant frequency and the impedance by standing on the good old Terra-Firma. The procedure is a little more involved but effective. First—we must have a means of electrically connecting the instruments to the antenna. This is best accomplished by a transmission line a half-wave or a multiple of a half-wave in length. Determine the height of the antenna above ground and calculate how many half-wave lengths of line will be required by using the following formula for a half-wave length of line:

$$L = \frac{(492)\ (K)}{f}$$

Where L = feet
f = megacycles
K = propagation constant
(RG/8 is .66)

This is an approximate length so be sure and cut it extra long because now we will find the exact physical length. Why an exact physical length? A halfwave length of transmission line will reflect the resistance placed across the output at the input end of the line, i.e. if a 50 ohm non-reactive resistor is placed across one end of a half wave or multiple length thereof, the GDO and Z-meter will indicate 50 ohms at the other end of the line. Cut the line somewhat longer than calculated above, short one end and connect the Z-meter to the other end of the line. NOTE: Keep twin lead off the ground and away from metal objects. Set the Z-meter to zero impedance and couple the GDO inductively to the Z-meter. Adjust GDO frequency until the fundamental frequency causes the Z-meter to dip or indicate a null. The frequency indicated should be lower than the desired frequency. Simply cut a few inches of cable off, short the end again and readjust the GDO. Repeat this procedure until the desired frequency (which should be the same as the resonant frequency of the antenna) is obtained. You will then have an electrically halfwave length of line or a multiple thereof.

Coax or twin-lead may be used for the halfwave length line when checking the impedance of the antenna. Connect the line to the antenna, hoist the antenna up to its operating position and adjust both the Z-meter and GDO for the null indication. If the antenna is not resonant at the desired frequency, the driven element should be readjusted a measured amount and then note the frequency change. This will give you an idea how much the resonant frequency changes with a corresponding element change. Now adjust the matching network to the desired impedance. This will be accomplished when the Z-meter dips at the desired impedance with the GDO set at the resonant frequency of the Antenna.

What would you do if your 100 foot coax cable developed a short someplace along the line? Replacing the whole line would be too expensive. Simply connect the Z-meter to one end of the line, adjust the Z-meter for zero impedance and then adjust the GDO for lowest frequency which will produce a null on the Z-meter. Use this frequency in the formula given for a halfwave length line and carefully calculate the length which will be the distance from the input end to the short.

A quarter wave length tuning stub can also be determined by using the procedures just outlined for the halfwave length line except a quarter wave line reflects a short at the input when the output end is electrically open. Now that we have mentioned the quarter wave length line, some may be wondering just what useful purpose does it serve. The quarter wave tuning stub (as it is sometimes called) may be used for antenna matching, TVI elimination or matching two units which have different impedances. The quarter wave matching stub can be used as a matching device on antennas which is explained in most antenna handbooks. It may also be used to eliminate an interfering frequency from entering the TV. This is accomplished by connecting a quarter wave stub to the TV antenna terminals which is a quarter wave in length at the interfering frequency.

Another use for the quarter wavelength matching stub is to permit maximum signal transfer between the source and a load which have different impedances. If the signal source impedance was 100 ohms and the load impedance was 52 ohms, a 72 ohm quarter wavelength of line would give a good impedance match. Hold it just a minute, how in the world did we come up with that 72 ohm business? Simple—another formula will give us this information.

$$Z_o = \sqrt{Z_s Z_a}$$

where Z_o = Impedance of quarterwave matching stub
Z_s = Impedance of the source
Z_a = Impedance of the load

Very little has been said concerning the various methods of GDO coupling. Actually—only two types of coupling are used; inductive and capacitive. Capacitive type coupling may be used on shielded coax cable, the ends of antenna elements and generally where the voltage maximum exists. To obtain the greatest accuracy, the GDO should be loosely coupled. Parallel coupling to inductors can be used to obtain maximum coupling.

Chapter III

Measuring RF Power Output

MEASURING RF OUTPUT

Paul Schuett WA6CPP

So your new Bandjammer 5000-Q is rated at 752W peak power. Big deal. How much of this is getting out where it counts?

It's easy to find out by inserting an rf ammeter in series with the line. A more exact reading would be to have the rf ammeter at the antenna input terminals, but that might be impractical when it comes to reading the meter (although you could put a diode there and a remote-reading meter in the shack).

Recently I found an rf ammeter at one of the mail-order surplus houses for $2.95. I installed this in a little cabinet, put two coax connectors on the back (in and out) and now can read rf current in the line, into the dummy antenna, or wherever it is going.

Remember the formula $P=I^2 R$? Square the reading on the meter, multiply by the impedance of the line you're using, and you have the power past that point. At the antenna, you could determine the antenna resistance and take the current reading at that point.

For those who have difficulty with mental computations and can't find a pencil, the following chart fives the computed power levels present in matched 50 and 70 Ohm coaxial lines for various levels of rf current:

RF Current Amperes	Power Output in Watts	
	50 Ohm Line	70 Ohm Line
0.5	12.5	17.5
1.0	50	70
2.0	200	280
3.0	450	630
4.0	800	1120
5.0	1250	1750

When tuning up the rig, place the ammeter on the antenna side of any tuning or matching devices and tune for maximum current.

Inside wiring is extremely simple. Note ground wire installed to insure continuity of the shield circuit.

The completed instrument showing the coax connectors on the rear.

It's amazing to learn how much (or how little) current these rigs produce. For instance, my Swan 250-C puts 1.35A into a 50Ω line. My friend's SB101 from Heathkit put out 0.7A until we worked on the antenna – then it put out about 1.25A. My Heathkit SB401 puts anywhere from 2 to 0.6A into a 50Ω Cantenna, depending on what band it's on.

Commercial stations determine their power by the antenna current. Knowing the antenna resistance at the operating frequency, they multiply that by the current squared. If the antenna resistance is 62½Ω, 4A rf would be 1 kW into the antenna. 126.48A would be 1 MW.

Remember the ammeter does insert a little reactance in the line (I never leave it in all the time), and the calibration changes with frequency, although you can tell what side of the ballpark you are on. An rf ammeter in the shack makes a nice piece of test equipment.

LIGHT BULBS AS RF INDICATORS

John Houser WB2GQY

The major appeal to the amateur – as well as some commercial applications – of light bulbs as rf power indicators is low cost. To this must be added the universal availability of bulbs and screwbase sockets for pennies.

As low cost is of primary interest to well over 50% of those interested in any project, and as I have always had an insatiable desire to find out the why's and wherefore's of standard light bulbs as rf power indicators, I decided it might be the opportune time to do a resarch project and determine once and for all just which bulbs might be suitable and which might not be, and also to determine whether light bulbs would make *good* rf power indicators, or poor, and to find out what precautions might have to be taken if one decided he was going to take this low-cost path of determining his transmitter output power rather than go for a more expensive power output meter.

Also, power output meters in the higher wattage ranges become quite expensive compared to the $2 to $5 which might be expended in a light bulb indicator. In general, porcelain screw bases are available for from 12¢ to 25¢ each, and bulbs from 15¢ to 65¢ each, and not more than four of each are necessary for up to 3 KW power indication.

Table 1 lists most of the common types of electric light bulbs readily available. One look at this table immediately reveals why such light bulbs might not be such good rf power indicators as some folks may have thought they were in the past. It also reveals that some very special precautions have to be taken in using them, or the user may find he has overloaded his transmitter and burned up a few components which might be expensive to replace.

Table 1

Variation in Resistance, Cold to Hot State, Common Variety of Electric Light Bulbs.

Bulb Rating Watts At 115V	Cold Filament Resistance	Hot Filament Resistance	Ratio Cold to Hot Filament Res. (Approx.)
7.5	166	1750	1 to 10
25	40	529	1 to 13
40	27	331	1 to 12
60	20	219	1 to 11
100	9	132	1 to 15
150	6	83	1 to 14
200	4.5	65	1 to 14
250	3.5	53	1 to 15
500	2	26	1 to 13
750	1+	17.7	1 to 15

Table 2

Possible Configurations for Various Power Outputs at Various Impedance Terminations

A. – Nominal 70 to 73Ω Impedance Loads:		
175W Load:	3–60W bulbs in parallel	(73Ω)
	OR	
	7–25W bulbs in parallel	(70Ω)
3,000W Load:	4–750W bulbs in series	(71Ω)
B.) Nominal 50 to 55Ω Impedance Loads:		
250W Load:	4–60W bulbs in parallel	(54.9Ω)
500W Load:	2–150W bulbs in series, both paralleled by 1–150W bulb	(55Ω)
1,000W Load:	2–500W bulbs in series	(53.6Ω)
2,250W load:	3–750W bulbs in series	(53.1Ω)

The extremely high ratio of cold to hot filament resistance in all types of these bulbs immediately struck me as being the most undesirable factor in using them.

It is very easy to see, for instance, that if one wished to use a 250W bulb for indication on a 250W transmitter, and he computed the resistance at 250W to be 53Ω, (which it is, but only when *hot*), he would assume he had just about a perfectly matched indicator to plug in in place of his 52.5Ω feed line.

However, from this table, it is apparent that this 53Ω resistance is attained *only* at full brilliance and wattage, and the actual cold resistance is only 3.5Ω. In other words, if the bulb were connected to the antenna terminals of the transmitter, and the transmitter keyed full power, the transmitter would be looking into *not* 53Ω, but 3.5Ω, which is a lot of difference, and an extremely low value for any pi network to match.

For a few seconds, until the filament attained full brilliance, the transmitter would be subjected to a terrific overload, due to this impedance mismatch.

Therefore the first precaution which might be emphasized in using light bulbs would be *not* to key the transmitter at full power with a cold bulb, but to gradually bring the power from some lower value to full power as the bulb attains full brightness (and hot, matching resistance).

Not until I got into this project did I realize the very high ratio of resistance of these filaments from the cold to hot state; I don't suppose very many people do. It also brings to mind how the house electric meter must jump every time a bulb is snapped on in the house. This is not an ad for those light dimmers being sold at all the electrical stores, but it sure brings to mind that power bills could be cut appreciably through their use, i.e., bringing the bulb gradually to full brilliance instead of just snapping on a switch.

Getting back to the bulbs, Table 2 gives in various configurations series, and/or parallel combinations which would be most likely to give the amateur a load for a particular transmitter power output, in nominal impedances near 52 and 72Ω. If the configuration mentions 200W, this does not mean that it would be suitable for indicating the output of a 100W output transmitter, because at half brightness, the resistance offered by the bulb is not identical to that at full brightness.

While a differance of an ohm or two would not be serious, nor would a difference of as much as five, or even ten watts, at high power levels, at low power levels less than 100W, for instance, such differences would be seen to become increasingly serious from the matched impedance standpoint. The configurations given match quite a variety of standard line impedances and a wide range of power outputs. Matches can be obtained for RG-8, 11, 17, 13, 58 and 59 type cable.[1]

One may not realize without measurement that the lead length of the filament support wires alone inside the 25–150W bulbs is very close to 18 cm. Even though they are coiled on a 2 to 1 ratio, the filament is inductive in every sense of the word. At higher frequencies, the filament support wires would appear inductive, and to these factors must be added the parallel capacity of the screwbase shell and the central base contact wafer. Even though such capacity is small, it would become significant at most amateur frequencies above the 30 MHz range. Though the 22 cm total wire path would perhaps indicate a bulb could be used up to 300 MHz, such is not at all the case.

It is easy to see that the sometimes suggested trick of using a capacitor in series with a light bulb as a load should be approached with caution, for it would be very easy indeed to run into a series resonant circuit which might result in damage to the transmitter to which such circuit were connected.

In the course of my preparation of this article, I discussed the ramifications with a number of interested hams. Some of them suggested I extend the research to include the use of the smaller types of indicator (pilot) bulbs as loads for testing out transmitters with power outputs in the 1W to 20W range, not only just for amateur applications, but also with a view to using them as loads in testing FM transmitters.

When one considers that there are well over 100 types of these small bulbs, rated from .001W to 2W, and if all of these were to be considered individually, it could take a vast amount of time – and eventually one would end up with perhaps only five or so of these bulbs that would be at all suitable, so such research was not included in this article. However it did open up a field in which there may be a demand for information and may be the subject of a subsequent article.

Frequency Ranges

The use of standard screw-base ceramic or steatite porcelain light bulb sockets is entirely feasible for all of the configurations shown and will handle all amateur bands, 160 through 10. Naturally the leads from socket to socket should be as short as possible in either the series and/or parallel configurations. I found these leads can be kept to approximately 2 cm for such interconnections. Likewise, the coax termination lead should be kept to 2 cm or less.

If extra precautions as to lead lengths are observed, and the bases of the bulbs removed to enable connections directly to the stem wires, it would appear reasonable to suspect that these bulbs might be used for 6, 5, and perhaps 2 meter bands, but it is also quite evident the 2 meter band would be the practical limit.

One should be able to conjecture that light bulbs as power rf indicators are not quite the equal of well-designed power output meters which maintain their rated impedances over a very wide power output range – bulbs do not – but then, they *are* cheap in comparison.

Visual comparison of brightness is completely satisfactory for comparison purposes.

For instance, a 500W bulb connected to the 115V mains should show the same brilliance as one of the 500W bulbs as used in the 1 KW load.

Actually a transmitter supposedly putting out 2,000W PEP is putting out something less than 1,000W with average voice modulation; it would be more of the order of 500–750W average power. Remember that the light bulb is only going to show average power output, not peak, and as ham transmitters are limited to 1,000W dc input to the final amplifier, one cannot expect much more than 500–750W output (average) unless the efficiency of the final amplifier stage approaches 85% which is very unusual, although I am hearing lately that certain high-power transistors are in development which will deliver such high efficiency figures; a bit above that which heretofore has been obtainable with tubes. You should be hearing a lot more about these super-efficiency transistors in the near future; and I expect them to be appearing in certain ham transmitters within a year or so.

Naturally a CW transmitter with the final operated Class C may deliver as much as 850W with 1,000W dc input, while a DSB transmitter on phone could not be expected to deliver more than 650W with Class A or B modulation.

The research and conclusions I reached on this project brought to mind the old subject of using light bulbs in series with primaries of transformers to reduce the secondary output voltages, which is a trick which has been used for years by hams and others. The information contained herein indicates they are not only quite suitable for such usage, but in fact make quite ideal voltage regulators of a sort.

In fact, the question immediately arises as to why bulbs would not make rather ideal voltage regulators for high voltage supplies if used as a variable-resistance dc regulator in the dc leg. This again opens up a field which might bear intense investigation.

[1] Solid Dielectric RF Transmission Lines, W8LUQ, *Radio News* Oct. 1946.
Line Matching: Table of Power and Voltage Loss in DB, *Radio News* Feb. 1947.

BUILD YOURSELF A LIGHT WATTMETER

Carl Henry

Every week brings something new in the hectic field (or pasture) of electronics. Attempting to solve old problems with new components is an interesting pastime for electronics enthusiasts, but they must be careful not to put their foot in the wrong thing. One of the new components is the cadmium sulphide/selenide photocell. A semiconductor sensitive to light is not an entirely new concept, since selenium cells have been around for some time, but the degree of sensitivity makes the cadmium cell stand out.

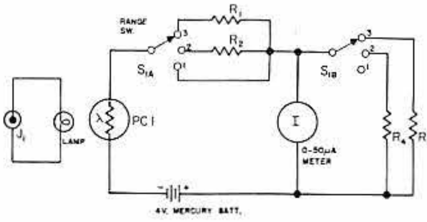

Fig. 1. Circuit of Light Wattmeter.

In measurements especially, there are many possible applications for cadmium photocells. One measurement in particular is usually difficult for the amateur, and this is power measurement. There is a way of using cadmium photocells to measure power, and I call this circuit a "light wattmeter". Operation is just as the name implies, that is, the power is used to generate light which is measured by the photocell.

You have probably realized by now that we are going to use a light bulb as a load. Now this is frequently done in amateur circles, but no one will go out on a limb as to its accuracy. Except me. We know that the ordinary lamp filament has a positive temperature coefficient. Fig. 5 illustrates the variation of a typical lamp filament resistance with input power. By keeping this in mind, fairly good accuracy can be had. Of course, if you use a different lamp, the curve will still apply, but the resistance will be different.

You can run your own graph, however, by using the setup shown in Fig. 2. Either dc or ac can be used, and since the lamp has little inductive or capacitive effect, the readings taken will be good to better than 100 mc. After you have a graph on the lamp you are using, you can effect any kind of impedance match you wish.

Using a lamp as a load simplifies the problem of power measurements because loads for this service are hard to come by. Power resistors are too inductive, and when they approach 1000 watts, they become downright expensive. However, even a 1000 watt lamp is not too expensive.

Fig. 1 shows the circuit that I used as the basis of this article. A standard 150 watt lamp was used here as a load. The photocell is mounted about five inches from the bulb. A wooden box houses the wattmeter, completely sealed internally against extraneous light. Figs. 3 and 4 illustrate the variation of resistance of the photocell with variation of power applied to the lamp. Even very small amounts of power are measureable, if the cell

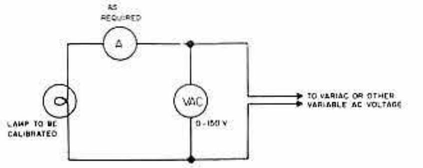

Fig. 2. Method of calibrating lamp and determining resistance.

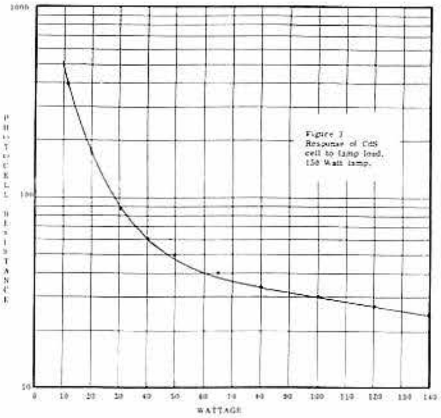

Fig. 3. Response of CdS cell to 150 watt lamp.

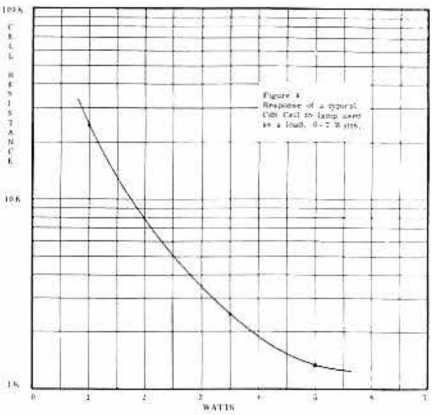

Fig. 4. Response of cell to 7 watt lamp.

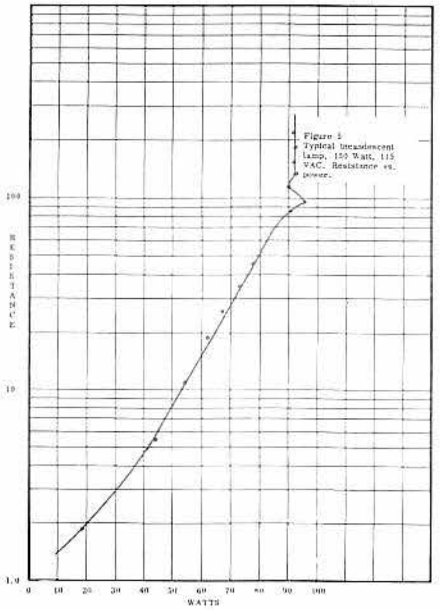

Fig. 5. Resistance versus power input, 150 watt lamp.

is shielded from external light.

The wires from the input connector to the lamp are kept as short as possible by removing the lamp base and soldering the lamp wires directly to the coaxial connector. Switch S1 selects three ranges, which can be set by the builder to anything he desires. In my case I used three ranges which cover from 0.5 watt full scale to 250 watts full scale. With some photocells a 1.35 volt mercury battery can be used instead of the 4 volt battery shown. Also, I used a 50 microamp meter, because of convenience (mine), but even a 10 ma meter will work. Don't exceed the rated power dissipation of the photocell, and remember this may derate with increasing ambient temperature. One thousand watts in a box can be a lot of ambient temperature.

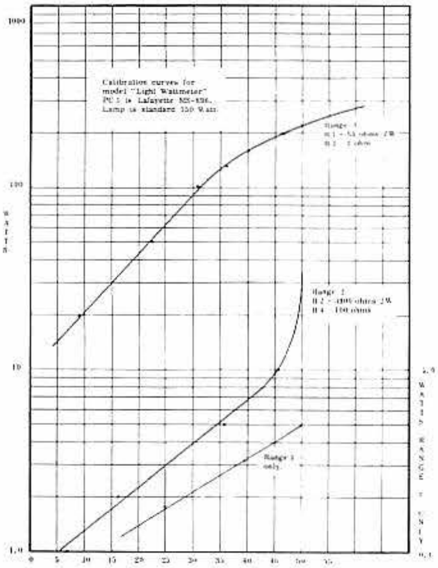

Fig. 6. Calibration curve for Light Wattmeter.

Getting back to the accuracy, if I were to tell you what the meter would read with such and such a lamp, and a given power input, your calibration would be off ten to twenty per cent when the construction was finished. The best way of getting a good calibration is something you've heard before; "if you want a job done right around here you've got to do it yourself".

Now obviously you can't calibrate the wattmeter at 14 mc, because this is what you are trying to measure to start with. Luckily the light from the lamp filament is primarily a power function, and it doesn't matter whether this is dc power, or 100 mc power.

So, to calibrate the wattmeter, apply ac or dc as you wish. Measure the input voltage and current, and read the meter. Make a graph of input power vs. meter reading, and there you have it. Fig. 6 is the graph that I use with the circuit in Fig. 1. Keep the impedance variation in mind, and select a lamp with the power impedance so that standing waves won't eat up lots of your power. As a dummy load and approximate power indictor this is not too critical, of course. But if you wish accurate power measurements, your impedance should be approximately matched.

LOW POWER LIGHT WATTMETER

Bill Hoisington K1CLL

This article describes a very useful gadget for determining the rf power output of solid-state VHF–UHF transmitters in the difficult range to measure, from about 10 mW up to 5 watts. It does not read watts directly; but by a simple comparison of calibrated pilot light brilliance, it will tell you how many watts you are putting out, to within less than 5%. It allows you to check power increases and estimate your efficiency quite close.

Principle Involved

We'll start right in with this part because, while this unit is not by any means a "trick," it does not read rf directly. You first light a pilot light as a good dummy load, matching it into the rf tank circuit of your transmitter by the normal means, also noted here.

You then switch on a second bulb of the same type by means of a battery, controlling the light output with a $1.30 wirewound potentiometer in series, as shown in Fig. 1. This pot must be previously calibrated in milliwatts, as by the method of "volts times milliamperes equals milliwatts." You then match the brilliance of the bulb lit up with rf or its dull glow at some 18 to 25 milliwatts if you're just getting your transmitter going, and read the watts on the wattmeter dial. It's astonishing how well it works, how repeatable it is, and how you wouldn't be without it once you build and calibrate it.

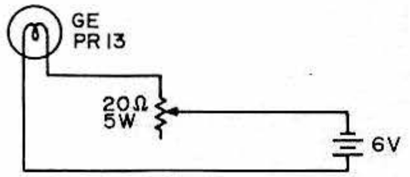

Fig. 1. The circuit supplies the brilliance "standard" for comparison. When the "standard" lamp is mounted adjacent to the dummy load, the pot permits variation of the standard to match the load. If the resistance is panel-marked in watts, a good power indication is achieved.

Brilliance Standard

Figure 2 tells almost the whole story at a glance. You can, of course, put as much calibration on the dial as you have time for. It is quite important to orient the bulb filaments in the same relation to your eyes for best matching. There isn't much in back of the panel except one 6V battery which can be obtained in any hardware store.

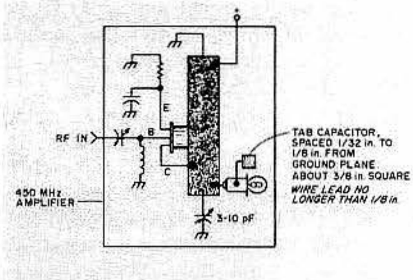
RF IN
E
B
C
TAB CAPACITOR, SPACED 1/32 in. TO 1/8 in. FROM GROUND PLANE. ABOUT 3/8 in. SQUARE
WIRE LEAD NO LONGER THAN 1/8 in.
450 MHz AMPLIFIER
3-10 pF

Fig. 3. Matched pilot light load for the UHF version.

RF Matching

Not that it is particularly critical, but be sure and note the need for a large range of series capacitors for the rf pilot lights as you go up in frequency. This can be seen clearly in Figs. 2 and 3. The block diagram, Fig. 2, shows a 6 meter setup. As you go up in frequency the series capacitance drops. A good matched load on 432 MHz can be obtained as shown in Fig. 3. I sometimes remove the tin base from the bulbs, but this is not an absolute neces-

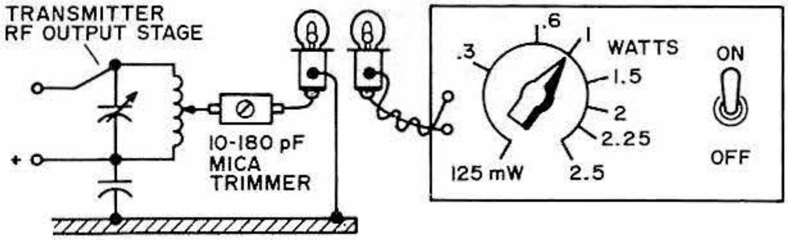

Fig. 2. A series capacitance loads the rf indicator for comparison. The capacitance value will decrease inversely with frequency increases.

sity. It is important to vary the amount of coupling, and thus the series capacity, by spacing the tab capacitor closer or further away from the ground plane, as detailed in Fig. 3. You also can use as many bulbs as you can solder onto a tuned rf inductor, even though they don't all light up with the same brilliance. You can match them all up, but you don't have to. Just check the wattage, or milliwattage, of each one and add them up for the total.

The number 48 or 49 bulb, listed at 2V and 60 mA, is rated at 120 mW, and glows dim at about 12 to 15 mW; so it can be used for low-power receiver oscillators, etc. With two other bulbs found in hardware stores, connected and matched to the rf inductor, such as the PR13 (5V at 500 mA), you can read correctly up to 5W. From there on up you're on your own, although a good variable 115V dc supply can be made up to work around 50 to 100W. I generally use a variety of 115V bulbs of different wattage, light them up with rf, and use their rated wattage.

YET ANOTHER LIGHT WATTMETER

John Meisner K5CXN

Many VHF operators would like a cheap, accurate instrument to measure rf power. The same desire is frequently expressed by operators of the HF bands. Here is a very simple wattmeter which when used with a 50Ω transmission line or load has all of the following desirable characteristics:

1. Easily calibrated to good accuracy (±5%) with your multimeter and a variable dc source;
2. Perfectly flat from dc to 450 MHz;
3. Insertion VSWR less than 1.05
4. Power readings in the 2 to 50 watt range (higher power can be measured with slight design changes);

The operating principle of this wattmeter is stark simplicity. A pilot lamp across the rf line senses a small portion of power in the line and glows brighter with increasing power. An appropriately located photovoltaic cell connected to a microammeter measures the light output which is proportional to the power flowing in the line. Of course a good deal of nonlinearity is involved in the various elements – both the lamp's resistance and its spectrum output change with heating; output of the photovoltaic cell varies considerably with both the amount and frequency of the light shining upon it. Some of these factors tend to cancel out however, because the photovoltaic cell produces some current with only infra-red lamp output at low power levels before the lamp even produces a visible glow, and the cell tends to saturate, increasing its output quite slowly at more intense illumination levels.

Probably the biggest single requirement of any wattmeter is that it must be capable of being inserted into a transmission line without disturbing the operating conditions in the line (low insertion VSWR). It might be argued that hanging a lightbulb across a transmission line will seriously affect the line impedance. Ordinarily, this is true, but the undesirable changes can be minimized and indeed approach an insignificant level if the resistance of the lamp filament is very large when compared to the line impedance. In general, a factor of 15 or more times the line impedance is sufficiently large to produce negligible effects. In the case of the suggested 10V, .014 Ampere pilot lamp, the mismatch produced in the line gives a VSWR of 1.05:1 at the power level of 1 watt. This mismatch decreases rapidly with increasing power. It falls to well below 1.01:1 at 50 watts. Another point of interest with regard to this particular choice of lamp is that it is a long-life type with a life expectancy of 10,000 hours. This implies two advantages: (a) The lamp will operate at well over its rated voltage without burnout (50.0V at 50 watts), and (b) the interior of the lamp envelope will resist darkening which would negate the wattmeter calibration. In addition, the filament structure of this type of lamp is a single-strand straight tungsten wire. The coiled type of filament structure introduces undesirable inductance into the circuit which can distort wattmeter readings in the UHF range.

The wattmeter shown is constructed in two boxes for the sake of convenience. The sensor can be located in the transmission line at any point and the meter can be placed beside the transmitter. For test work, the whole unit could easily be put into a single box. Coaxial fittings and cable were used for interconnection, but since dc only flows in this circuit, any type of wiring would be satisfactory.

The photovoltaic cell used in this wattmeter is a unit obtained from the local Allied/Radio Shack store. (Catalog No. 276-115.) Output of the cell is rated .5V at .6 mA in sunlight.

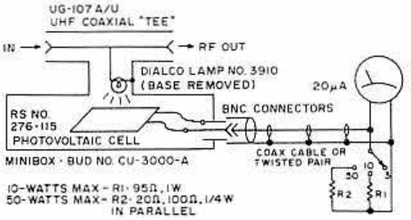

Fig. 1. Schematic diagram.

Since the cell generates considerably more than 20 mA under moderate illumination, a switching arrangement is incorporated into the wattmeter to shunt it into progressively higher current ranges with increasing power. Alternatively, a 50 or 100 milliammeter could be used with less switching at a sacrifice of sensitivity in the 1–3 watt range.

Calibration of the wattmeter is a simple process. By solution of the formula $P = V^2/R$ for voltage, the following tabulation is made for a 50Ω line impedance:

VOLTAGE (rms or dc)	POWER (Watts into 50Ω)
7.07	1
10.00	2
12.24	3
14.14	4
15.71	5
17.32	6
18.71	7
20.00	8
21.21	9
22.36	10
31.62	20
38.71	30
44.71	40
50.00	50

Using the tabulation, fasten a metered variable dc supply into the wattmeter according to the following diagram. Now simply note the reading on your wattmeter for each of the selected voltages in the table and tabulate this reading with the corresponding power in watts in a table of your own. It may even be possible to remove the front of the meter case and mark new calibrations directly on the dial. This was not possible with some hermetically sealed meters. It is best to disconnect the wattmeter from the antenna feedline for this calibration. If the meter is left with a transmission line attached and the antenna happens to be fed through a balun device with near zero resistance, the power supply, and perhaps the balun, will suffer.

It must be pointed out again that this power meter is intended for use either with 50Ω coaxial line systems with low VSWR or with 50Ω dummy loads. A coaxial line that is not "flat" (unity VSWR) or a dummy antenna that does not look like" 50Ω, which is the case with

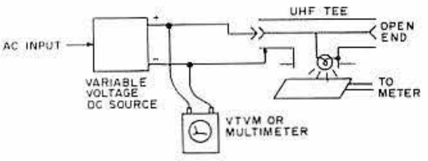

Fig. 2. Calibration diagram.

most HF loads used at VHF or UHF, may cause distorted power readings. If you have doubt about your transmitter's power output when using this meter, you should check it while using a known 50Ω load rated at the transmitter's output frequency.

I mentioned before that this power meter was easily adaptable to higher power readings. To upgrade the meter, the only change required is to insert the correct higher voltage pilot lamp having low current, long-life specifications. Suggestions for some of these are the following Dialco Lamps:

Part No.	Max Power Level (Watts)
24CS	100
48CS	200
60PSB	340
120PSB	800

A LOW COST RF WATTMETER

Mark Leavey WA3AJR

You say you just finished building that 2 meter rig and want to find out how much power you're running but can't afford to buy a wattmeter? You've been calling "CQ 80 QRP" all day with your quarter-watt wonder and nobody is answering? Then get yourself up, go down into the workshop, and build yourself a neato-keeno handy-dandy wattmeter.

If we are going to build a wattmeter, let's consider what we want. Accuracy and ease of calibration, as well as simplicity in construction are prime requisites. The meter described here is as accurate as components allow, and it's easy to build. The calibration is logarithmic, which means that a simple graph is possible, and easier than changing the meter scale.

For the mathematicians, I will present the formulas upon which this device is based, and ways of modifying it; for those of you who avoid math whenever you can, look at the graphs and skip these few paragraphs.

To spare undue complexity, assume 51Ω line – other values can be dealt with later. Perhaps the easiest parameter to measure, and one that is proportional to power, is rf voltage. A voltmeter can be made most easily with a series resistor and a 0–1 mA meter. Now let's plunge into the actual calculations.

Assume W is the full-scale meter reading in watts, Z is the line impedance, E is the voltage measured, I is the full-scale meter reading in amps of the basic meter, and R is the value of the series resistor in ohms. We know that the voltage (IR) is equal to the square root of "impedance times power," or 14.270V. Now, since the voltage and the current (.001A because the full-scale movement is 1 mA) are known, simple division yields 14,270. The resistor value, then, is 14.270 kΩ. The upper portion of Fig. 1 is a graph that will enable the nonmathematician to choose the value of the resistor for full-scale readings up to 4 kW, with 51Ω line and a 0–1 mA meter.

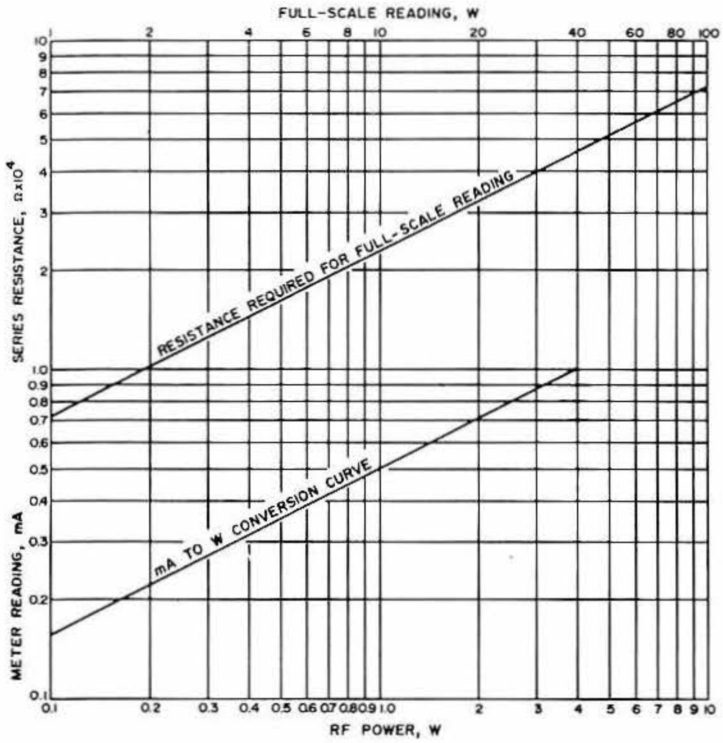

Fig. 1. Logarithmic plots for determining power. The upper curve gives resistance values for determining what the full-scale meter deflection will be (remember to multiply the series resistance value shown on the chart by 12 kΩ). The lower curve will allow you to determine your precise power out if you use a 0–1 mA meter.

Why 4 kW with an amateur power limit of 1 kW? A look at the bottom half of Fig. 1 will explain. Although this is the calibration of the prototype, for 4W full scale, it will double for 40, 400, or 4000W. A half-scale reading, 0.5 mA, corresponds to 1W (1 kW, etc.). This spreads out the range below 1 kW for ease of reading and measuring.

Now get out that soldering copper and gas pliers, and build it. As the schematic (Fig. 2) shows, the circuit is a basic rectifying type rf voltmeter. The prototype was built in a small can of the plug-in-module variety that was scrounged from the junkbox. About the only critical part is the series resistor. The capacitors in the prototype were mica, but ceramic disks would work as well. The diode can be a 1N34A, 1N270, 1N52, 1N38A, or just about anything else. Use the old ham's rule of thumb: "When in doubt, try it out!"

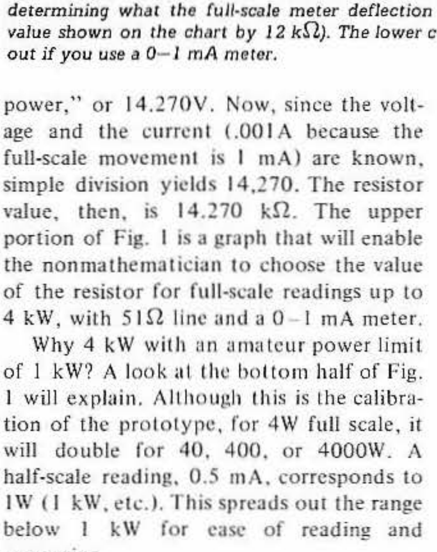

Front view of unit.

Two sockets might prove more convenient rather than one with a coaxial tee as shown. Conventional minibox construction or building into a new or existing rig will be more than adequate. Point-to-point wiring is used to permit compactness and reduce lead length.

"Fine," you say, "but I don't have a huge mound of test equipment. How do I calibrate it?" That is the beauty of it – you don't! If the series resistor is accurate, the meter will be self-calibrating to a log scale. Remember, you know R and Z, and the full-scale W. Now assume a half-scale

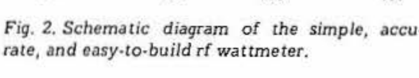

Fig. 2. Schematic diagram of the simple, accurate, and easy-to-build rf wattmeter.

Connection to transmitter with dummy load (see text).

reading, I = 0.0005, and calculate W for half-scale. Plot these two points at 1.0 and 0.5 mA on Fig. 1, and connect by a straight line, which you may extend the length of the graph.

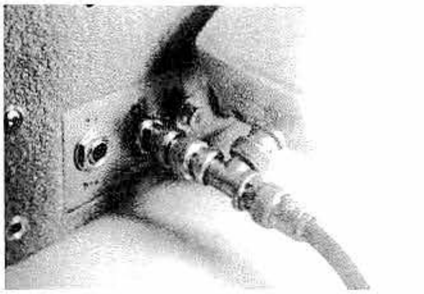

Connection to transmitter with antenna connected.

Install the meter through a coaxial tee at your antenna connector, or through some other predetermined means, and terminate with a dummy load. The one seen in the picture is three 150Ω resistors in parallel, dipped in epoxy, shielded with a copper braid, and installed on a BNC plug. Apply power and read the meter. That's it! The meter can be used with an antenna if your swr is below about 1.2:1.

So what did I promise? A low-cost, rf wattmeter that is inexpensive enough for the Novice, practical and useful enough for the General, and "Extra" accurate. Go raid the junkbox, and add a worthwhile piece of gear to your shack.

RF POWER MEASUREMENT USING HOT CARRIER DIODES

Frank Jones W6AJF

Two rf wattmeters are shown here, one with a range of 25 mW to 10W and the other covering the range of 5 to 300W. Both are useful from low radio frequencies on up through 450 MHz.

The low-power version (Fig. 1) makes use of a 20W Sierra dummy antenna built into the meter case, though the metering circuit only goes up to 10W. If the maximum is to be 20W, the reference meter reading could be about 45 μA instead of 30. The minimum power reading would be doubled. In this wattmeter, the power range potentiometer is calibrated and only a reference line on the meter is used when making rf measurements. The dummy 50Ω antenna resistor is rated up to 1000 MHz so is excellent from 450 MHz down.

The range potentiometer had an audio (nonlinear) taper. By connecting the "high" resistance end to the diode, the watt range scale is spread out quite well in the 0.1 – 10W range. The hot carrier diode, an HP 2900, has a 10 PIV rating, which means that the rms rf voltage across it should be less than 3V for safe operation. At 10W of rf power, the rms voltage would be a little over 22V, which means a voltage divider is needed to keep the applied diode voltage down to about 2V. An HP 2800 diode with a 75 PIV rating would be more desirable, especially if the meter was to be calibrated for 20W maximum. This diode is about $1 and has a little higher capacitance, which would require a different shunt capacitance across parts of the resistor divider to make the device work with the same power range calibration.

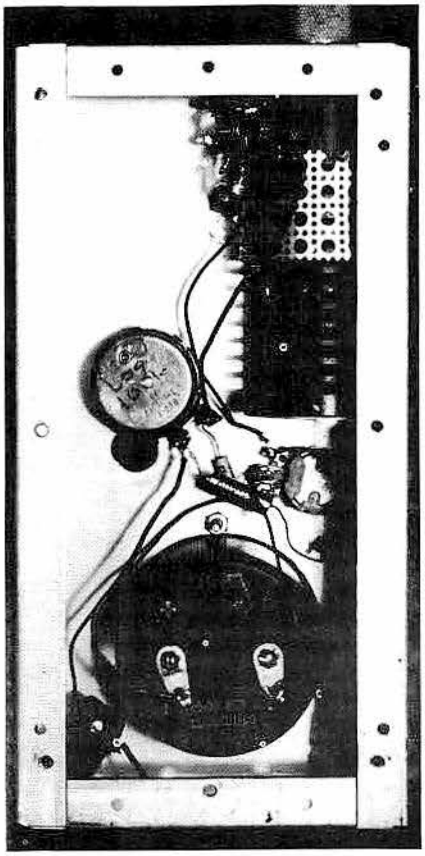

Bottom view of 10W unit with the rf dummy antenna clamped in one corner.

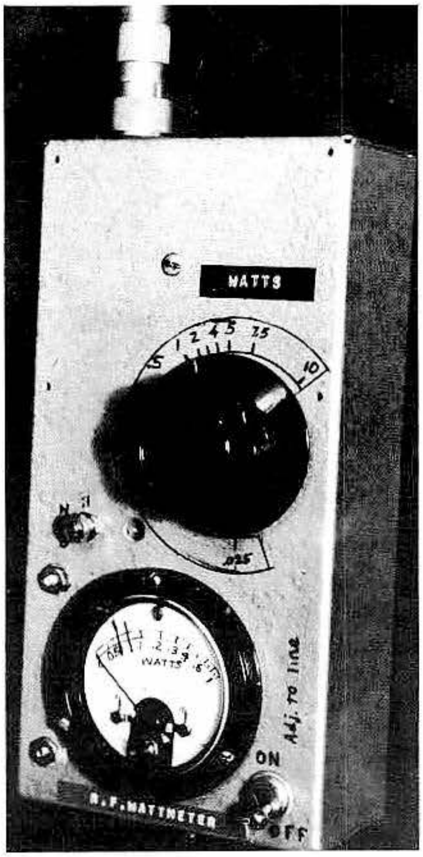

Top view of low powered rf wattmeter covering .025 to 10W. Built into a 8x4x2 chassis with wire screen bottom plate for ventilation.

The divider should use ½W resistors of the carbon or metal film type, since these units are part of the rf circuit. It is better to use three ½W 300Ω resistors in the string rather than a single 900Ω 2W resistor: this is because the rf resistance characteristic is usually better in ¼ or ½W types in certain ranges of resistance. Every resis-

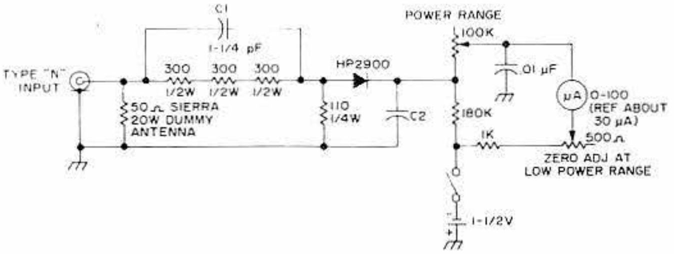

Fig. 1. .025 to 10W RF wattmeter.

tor has some inductance and shunt capacitance which becomes part of the voltage divider. The diode shunt capacitance is in parallel with that of the 110Ω ¼W resistor in Fig. 1. However, nearly any combination of resistor sizes can be equalized within 10 to 20% over the desired frequency range. This divider is across the 50Ω dummy antenna, so should not shunt the value down to less than 49 or 48Ω. This divider has to dissipate a little rf power also. Its total resistance should be at least 20 times as high as the dummy antenna load resistor.

The values shown in Fig. 1 are just about the minimum that should be used. Too high values makes it more difficult to extend the frequency range to the upper end, though it can be done, as was discovered in the higher-powered wattmeter of Fig. 2.

All diodes are poor rectifiers at applied rf voltages below their forward bias values of 300–700 mV (peak). By using a forward dc bias voltage to make the diode conduct at least 5 or 10 mA, the detection sensitivity is increased as much as 5 or 10 times. This requires a small battery, a couple of fixed-value resistors, and an adjustable pot to balance this current out of the meter when measuring rf powers below 100 mV. If the power range is limited to a minimum of ¼ or ½W, no bias circuit is needed in this 10W instrument. The range scale in either case has to be hand calibrated.

A low-powered radio transmitter or exciter can be used as a 10W power source when calibrating the power range pot scale. The transmitter can use stage detuning to reduce power outputs down to the lower values needed. Many swr meters have watts of power calibration and one of these can be put in the coax line to the rf wattmeter for calibration service. A more accurate calibration can be made by comparing the power readings against some reliable commercial rf wattmeter within its frequency range and calibration charts. This scheme is usually necessary for checking the calibration at VHF or UHF. Another method is to use an accurate rf voltmeter across the dummy antenna connection to ground and read the power values in watts = E^2/R. For example, 5V (rms) squared is 25; and divided by 50Ω is equal to 500 mW.

The Sierra 50Ω dummy antenna has no connection available at the high end of the resistor, which terminates in a type N fitting. The metering circuit has to connect to this point as close as possible by getting into the inner conductor of a coax fitting, or by drilling a 3/8 or ½ in. hole through the shell of the dummy antenna close to the rf fitting end. This can be done and the first 300Ω resistor in the voltage divider soldered to the inner connection to the large 50Ω resistor. A long 1/8 in. diameter soldering iron tip is needed. The divider resistors, diode, and four .001 μF stud-mounted bypass capacitors were all mounted around this large hole in tapped 6-32 holes for the four capacitors. Larger values of bypass capacitors can be shunted across these 0.001 μF values to ground to extend the frequency range down to low rf or even af values. For example, a .02 μF capacitor shunt would allow operation to 2 MHz. A miniature 50 or 100 μF electrolytic shunt would function at audio frequencies down to 300 Hz. The diode must have a low-impedance path to ground over the desired frequency range to function as a peak rectifier and get as much dc output voltage as possible for the meter circuit. The microammeter in series with a variable range resistor is simply a dc voltmeter. The diode rectifier converts rf voltage to dc, so the diode should be equally efficient over the whole rf range.

The 5–300W unit was built to use with a large dummy antenna rated up to 500 MHz, which is a massive unit external to the box shown in the photographs. Quite a bit of rebuilding went into this device to make one calibration of the range potentiometer fit all frequencies from 450 to 2 MHz. The input and output coax fittings had to be finally mounted so the inner conductor tips could be soldered together and the resistor divider connected to this point. The latter consisted of two 4300Ω 2W carbon resistors and a 68Ω 1W resistor in series to a copper sheet inside of the aluminum box.

The watt range variable resistor was a 500 kΩ linear potentiometer which was limited to a lower value by shunting it from the moving arm to the diode connection end with a 220 kΩ resistor. This gave a maximum power reading of 300W when the reference line was drawn on the meter face at 12 μA. The import, low priced, 0–30 μA meter had a large meter scale. A smaller 0–50 μA meter would have been usable, since the meter is used only as a reference. The range pot knob is adjusted when rf power is applied to run the meter reading up to the line drawn on the meter scale face.

ANT.
INPUT
25
100
75
50
WATTS
DC MICROAMPERES

Top view of 5 to 300W metering circuit for use with external high powered dummy antenna.

The circuit shown in Fig. 2 was equalized to within about 15% error over the range of 2 to 450 MHz by shunting a 5 pF capacitor across the 68Ω resistor in the rf divider.

Calibration of this device was made at 144 MHz using a transmitter having up to 400W available carrier output. The meterizing unit was connected to a large Bird rf wattmeter at the external fittings of the latter. Several thermocouples had to be used to cover the wide range of power for the calibration. This required reading a chart curve for each Bird wattmeter reading and using correction factors for frequency in order to obtain the actual watts of rf power. Now, the large unit is used without the thermocouples, charts, rf choke, etc. simply as a dummy antenna. The new metering circuit connects directly into the antenna fitting, with a few feet of 50Ω coax over to the transmitters being tested.

This power measuring device can be used in any 50Ω coaxial line to monitor the actual power going toward the antenna. The swr in the line should be low, or near unity, in order for the calibration to be reasonably accurate.

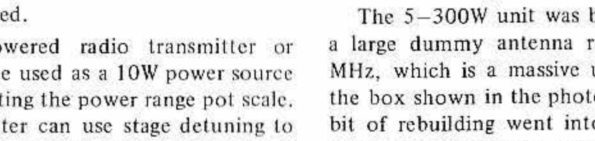

Fig. 2. 5 to 300W RF wattmeter metering circuit. External 300 or 400W during antenna load.

VHF DUMMY LOAD WATTMETER

Glen Zook K9STH

The unit

The unit described herein is similar to some 60 watt units which may be found around many commercial two-way radio shops. This dummy load has provision for connection to an external relative output meter. This external output meter may become an accurate wattmeter if the following criteria are met:

1. Frequency bandwidth of ±10% of calibration frequency.
2. RF output kept within power dissipation of dummy load.
3. Accurate initial calibration.

These criteria may be easily met in amateur vhf operation if only one band is considered for each set of calibration data. Since most vhf amateurs operate on 50 mhz, 144 mhz or 432 mhz, the ± 10% frequency limitations may be easily met. This limitation gives a 10 mhz bandwidth at 50 mhz, 29 mhz bandwidth at 144 mhz, and 86 mhz at 432 mhz. The limitation to the power ratings of the dummy load is only common sense, for if a resistive network is overloaded, the impedance may be drastically increased, caused by damage to the load resistors. The calibration limitation may be overcome if a standard, previously calibrated unit, or commercial unit is used.

The unit consists basically of 16 220 Ohm resistors in a series parallel arrangement. The metering circuit consists of a germanium diode pickup with necessary rf filtering. The meter movement is generally a vom, but any 50 ua meter movement should suffice. Exact physical layout is not extremely critical, but it is suggested that the layout be made similar to the unit shown in the accompanying photographs. This unit is acceptable for 60 Watt output transmitters without modification. The power capability may be increased to about 200 Watts if the resistive network is suspended in 1 quart of oil. If this is done, care must be taken to keep the metering circuit out of the oil. The lead from the diode to the resistive network must, of course, be partly submerged, but keep the diode itself out of the oil. I do not personally use this arrangement, but I know of two units which have been in use at a large Southeastern two-way radio shop for several years. The same shop has incorporated a range switch with several meter shunts for various maximum scale power readings. This feature is especially useful to the amateur vhf FM operator who may be working with equipment of from ½ to 250 Watt outputs. The schematic appears as Fig. 1, and the basic circuit for various meter shunts as Fig. 2.

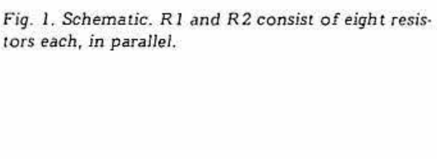

Fig. 1. Schematic. R1 and R2 consist of eight resistors each, in parallel.

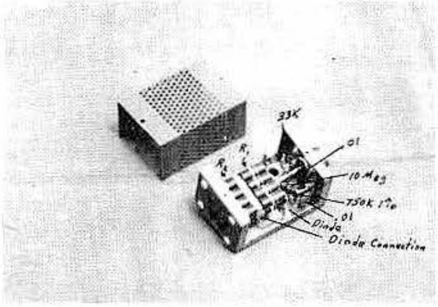

Parts layout and interior view of Dummy Load – Wattmeter

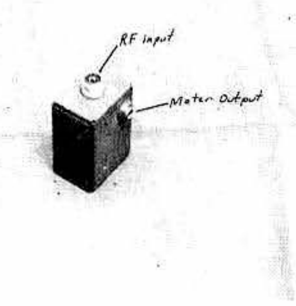

External view.

Calibration

Calibration is best accomplished by using a Bird "Thruline" or similar commercial vhf inline wattmeter. Second choice is a Bird "Termaline" or similar dummy load-wattmeter. In both cases, a graph should be created by plotting meter divisions on the horizontal axis, and power on the vertical axis. The meter shunt should be placed at minimum resistance and increased to give maximum reading at the desired power level (this holds primarily true for units using the metering circuit of Fig. 2) or, if a vom or vtvm is being used the range switch should be placed on a high voltage setting and reduced a setting at a time until the desired reading is obtained. The transmitter should be adjusted for various power levels on the standard wattmeter and the voltage or current reading on the new meter recorded on the graph. In the case where the standard meter is of the dummy load-wattmeter type, it will be necessary to switch the coax from one unit to the other. Do not retune the transmitter, for each unit will present almost the same load to the transmitter (50 ohms). Take the reading and record as with an inline type of meter. The points on the graph should now be connected with a smooth curve (use of a draftsman's range switch is used, it will be necessary to calibrate for each switch position. Also, if multi-band use is expected, the graphs must be made for each band. Use of the wattmeter now requires only the connection to the transmitter, setting of range switch to the proper level, and reading the graph.

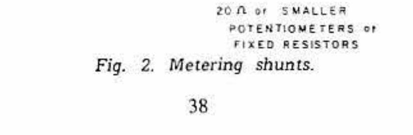

Fig. 2. Metering shunts.

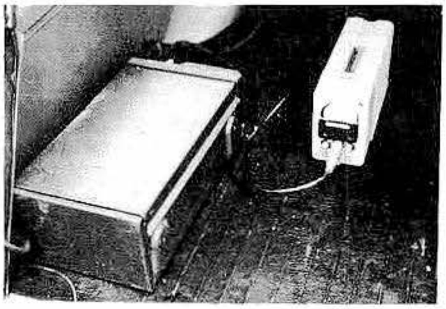

Using the Dummy Load – Wattmeter to check output of FM unit.

Uses

The uses of this dummy load-wattmeter are as varied as the amateur mind can devise. One very important use is determining the losses of 50 ohm coax. Measure the output of the transmitter at the transmitter. Then measure the output at the end of the length of coax. The losses in the line become apparent. The loss in db may be calculated by the standard power ratio formula, 10 $\log_{10}$ Power out of coax/Power into coax.

Another use is the determination of efficiency of final amplifier stages. This efficiency may be calculated by Power out (measured by dummy load-wattmeter)/Power in (measured by plate current/plate voltage meter) x 100%. A third use is determining once and for all which amateur really has the most output. This list may be expanded by the builder to suit his own tastes.

Conclusion

This dummy load-wattmeter is not a Bird "Termaline" nor should it be regarded as a substitute for any other laboratory equipment. However, with a little care in calibration, (assuming a 5% accuracy standard is used for initial calibration) the accuracy should be within 10%, and this, my friend, is not bad for a wattmeter costing less than $10.

Chapter IV

Measure Your Field Strength and Frequency

THE RF SNIFFER

Jim Kyle K5JKX

Every now and then there's a need to know if any rf is present in a circuit. Frequency isn't so important—the question is simply. "Is there rf here?"

Your grid-dipper can frequently answer this, if used in the wavemeter mode, but occasionally it's not sensitive enough—particularly if you're working with a receiver oscillator where power is measured in microwatts.

Here's an rf Sniffer which will indicate the slightest trace of rf in a circuit. In addition to checking receiver oscillators, it's a perfect gadget to ensure perfect neutralization of a transmitter final.

Connect the components as shown in the schematic. Use long-nosed pliers as a heat sink between the diode and the solder joint when wiring, to prevent diode damage. Note that the pickup loop of 14 gauge wire is insulated with a strip of spaghetti.

Mount capacitor and diode on back of meter with shortest possible leads. Attach pickup loop directly to negative meter terminal; it's stiff enough to do without other mechanical support.

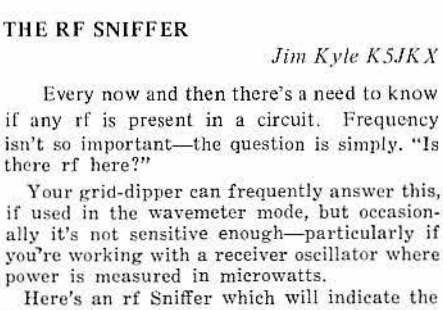

Amplifier Neutralization—Couple the Sniffer to the antenna terminal with a temporary two-turn link around the pickup loop. Remove plate and screen voltage from the final amplifier. Apply drive. Adjust neutralization for minimum indication on the Sniffer—but don't expect to be able to get it down to zero.

Oscillator Checking—Place the pickup loop near the oscillator coil. If the oscillator's working, the Sniffer will indicate rf. Touching either the grid or plate lead (use an *insulated* tool for this test, not your fingers) should reduce the Sniffer's indication.

Receiver Troubleshooting—Check the oscillator as described above. If it's okay, next check the mixer plate coil by placing the Sniffer pickup loop near it. If you get an indication here, move to the first if stage and place the pickup loop near the plate pin of the tube socket. Proceed through the receiver until you lose the indication. The trouble is somewhere between the last indication and the point at which it disappeared.

Field Strength Meter—Couple a short antenna to the pickup loop by two turns of wire around the loop. Field strength will be indicated in a comparative manner by the meter. It cannot be calibrated, but proves useful in tuning mobile or beam antennas, etc.

SWR Measurement—(Parallel lines only). Move the Sniffer long the line. Mark maximum reading and minimum reading over a half-wavelength. Divide minimum into maximum. The quotient is, roughly, your VSWR. This method is by no means exact, but will indicate whether the line is under or over a 2:1 SWR.

UHF Frequency Measurement — Set up Lecher wires. Couple the rf Sniffer lightly to the tank circuit instead of using a flashlight bulb. Use Lecher wires in normal fashion, reading Sniffer indications for maximum and minimum. This is much more exact than the normal methods.

Improvised Grid-Dipper—If you have a signal generator available, it can be used with the rf Sniffer to serve as a "grid-dip" meter to locate resonance for any tank circuit. Couple both the generator and the Sniffer lightly to the unknown tank. Vary generator frequency. A sharp rise in Sniffer indication indicates the resonance point.

THE PINK TICKET REJECTOR

A. D. Taylor GW8PG

Hams sometimes get into trouble with the FCC because they have made a mistake in tuning a frequency multiplier or PA stage. For example, if the multiplier stage following a 3.5 MHz oscillator is accidentally tuned to 10.5 MHz instead of 7 MHz, the PA can be loaded on 10.5 MHz; and if this is done even a short transmission is likely to produce an unwanted "QSL" from the FCC! Mistakes like this can be prevented by using the simple, easily built absorption wavemeter shown in Fig. 1. Only one coil is required in this meter, the low frequency range being obtained by switching a padding capacitor in parallel with the variable tuning capacitor. With switch S1 open the tuning range is approximately 6.8 to 30 MHz, and with S1 closed it is approximately 3.5 to 6.8 MHz.

Construction of the wavemeter is simple. It can be put together in a metal utility box, on a wooden panel and baseboard, or even in a cracker can of the right size. The coil can be wound on any type of ¾ in. diameter form. If nothing else is handy, a short length of dowel could be used. The coil winding consists of 7 turns of 20-gage wire, close wound. Both the tuning capacitor (C1) and the padding capacitor (C2) should be 500 pF components, but the values of the other components are not critical.

Any meter having a full-scale deflection between 100 µA and 1 mA can be used. If the wavemeter is built in a metal box, L1 must be mounted outside the box. A slow-motion drive is not needed. If C1 is fitted with a pointer type knob and a cardboard scale is cemented onto the front panel, the calibration points can be written on the scale in ink.

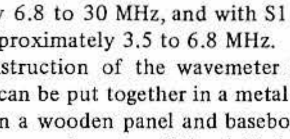

Fig. 1. L1 – 7 turns 20 gage enameled copper wire, close-wound on a ¾ in. form; C1 – 500 pF variable capacitor; C2 – 500 pF fixed capacitor; C3 – Fixed capacitor, any value between 1000 pF and 0.01 µF; D1 – Silicon or germanium; M1 – Moving coil-meter (100 mA–1 mA).

When it came to calibrating the wavemeter I thought of a simple method that I have not seen described before. I soldered about 2 in. of wire onto one end of L1, and about 10 ft of wire onto the other end. I then connected the short lead to the antenna terminal of my receiver and strung the 10 ft of wire up as a temporary antenna. I then tuned the receiver to a steady signal at each frequency at which I wanted a calibration point, and adjusted C1 until the strength of the received signal suddenly dropped sharply, indicating that the wavemeter was tuned to the frequency of the signal. I was then able to mark this frequency on the cardboard tuning scale. The drop in signal strengh was very sharp so calibration was easy. Once enough calibration points had been obtained the temporary wires were unsoldered and the wavemeter was ready for use.

The only frequency which may require a little adjustment is 30 MHz. If the wavemeter will not tune as high as this, push the top turn of coil L1 about 1/8 in. away from the other turns. This should reduce the inductance sufficiently; a spot of cement will hold the turn in its new position.

To use the wavemeter for checking a transmitter, bring coil L1 close to the tank coil of the stage being checked, apply power to the stage and rotate C1 until a maximum reading is obtained on the meter. The output frequency of the stage can then be read off from the wavemeter tuning scale.

The wavemeter can also be used as a radiation meter for tuning up single-wire antennas. If it is tuned to the transmitter output frequency and placed near to the antenna wire, maximum reading on meter M1 will indicate maximum output power from the transmitter.

Many readers will have realized that the calibration method I suggest uses the principle of the rejector circuit. That is why I have called the little gadget "the pink ticket rejector"!

A USEFUL ACCESSORY FOR THE HAM SHACK

Rex Morris W2WHX

BECAUSE we are dealing with something we cannot see, namely electrons and electromagnetic radiation (we only see their effects), we must acquire the ability to use test equipment, in order to understand and find our way in this invisible realm.

The piece of test equipment about to be described is one of the more useful instruments that the amateur should have in the shack. While this instrument serves primarily as a field strength meter, it will also serve as a phone monitor, neutralization indicator and a sensitive wavemeter.

Referring to the circuit diagram, meter M1 is a sensitive instrument which indicates the pressure of rf when the device is used as a field strength meter, wavemeter or neutralization indicator.

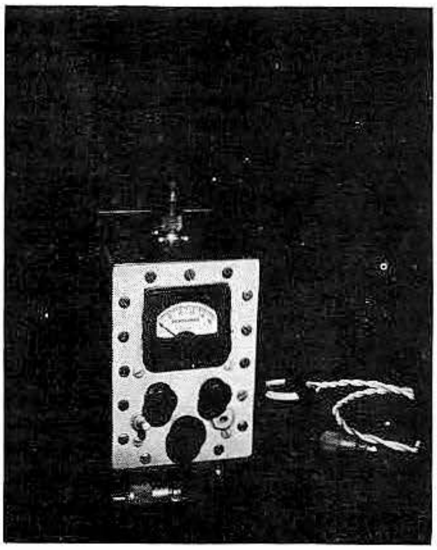

To use this instrument as a wavemeter, a pickup loop is substituted for the short whip antenna at the top. Switch S1, the band switch on the right of the panel, is placed on the proper position. Coil L1 tunes the vhf range of approximately 90 mc to 170 mc (2 meters). L2 tunes 28 mc to 100 mc (10 meter and six meter bands). L3 tunes to 7 mc to 30 mc (40, 20, 15 and 10 meter bands). L4 tunes from 2 mc to 7 mc (80 and 40 meter bands). With switch S1 on the proper tap, condenser C2, on left of panel, is used to peak the reading on meter M1. Using a calibrated dial, frequency may be read directly. Toggle switch S2 is a high-low range switch for meter M1, providing a means of keeping the meter on scale and protecting it against burn-out. Potentiometer R1 (center of panel) is a vernier shunt control, also for keeping the meter on scale.

For use as a phone monitor, rf should be fed into the input jack with a link or pick-up wire. Once again the LC circuit is resonated to the frequency we desire. Earphones inserted in jack J2 will open the meter circuit and allow you to monitor the signal. Switch S2 is placed in the Hi position.

Field strength readings can be taken by using a short pickup antenna. Again the LC circuit should be tuned to resonance. Meter M1 will give an indication of field strength.

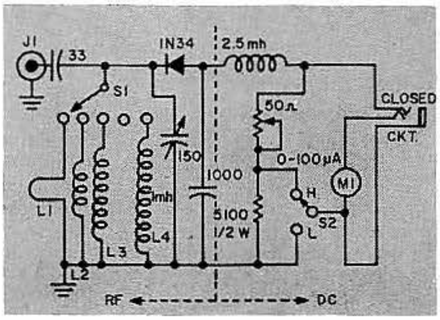

With switch S2 in the Hi position the meter is very sensitive and potentiometer R1 is a variable shunt providing much range of scale adjustment. For use as a remote reading field strength meter an external microammeter (with up to 200 feet of wire) may be plugged into J2. For making transmitter adjustments this is a very desirable feature.

Now notice that one position on the bandswitch S1 is vacant, this vacant position provides a very broad band—low sensitivity position for those extremely high rf fields where even R1 and S2 cannot provide enough attenuation.

Neutralization measurements are made by coupling the instrument through a pick-up link to the tank coil involved, with S2 in the Lo position. When the instrument is tuned to resonance it becomes a very sensitive rf indicator. It is so sensitive that it will readily be seen that complete neutralization exists in theory only.

As in all simple gadgets there are a few simple construction techniques which make the difference between gadget and instrument. In this case the important thing to keep in mind is that from the antenna to the crystal is the rf portion, and from the crystal to the jack J2 is the dc portion (with audio superimposed). With this in mind, construction is such that the two parts are separated, thereby giving some measure of protection from rf energy to the very sensitive microammeter M1. Also note the extensive use of sheet metal screws on the aluminum case. The only rf we want to enter the case is the rf we are attempting to measure via the antenna jack. Note again, we have here an instrument for detecting electromagnetic radiation, from approximately 170 mc to 2 mc. We now have an instrument for visualizing what cannot be seen.

The usefulness of this instrument is limited only to one's ability to apply it and interpret the results it gives. These only come with experience, trial and error and determined application.

Coil Data

L1—One turn hairpin loop.
L2—5 turns of #18 enamel wire space wound 1" dia.
L3—24 turns #22 cloth covered wire, close wound, ⅝" dia.

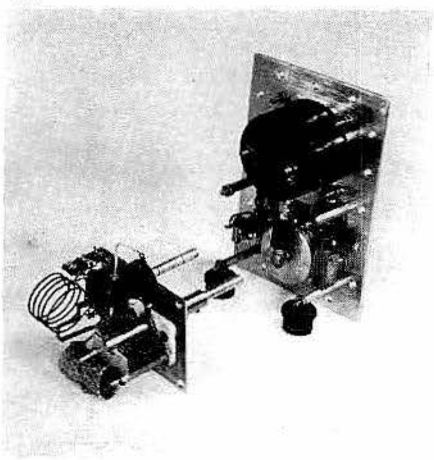

TWO THRU TWENTY FSM

Howard Pyle W7OE

I used a small LMB aluminum meter cabinet with a hole for a 2" meter; any equivalent cabinet can of course be used. With a small enclosure such as this (4"x4") the meter occupies the face of the cabinet and all controls are on the rear. Using a somewhat larger housing both the meter and the control knobs can be placed on a front panel if you prefer.

Rather than wind up the three coils I found that the J. W. Miller Company RF chokes with minor modifications were admirably suited for a compact job. One each of their catalog numbers 4606, 4588 and 4580 were required to cover the three frequency spreads from 2 through 20 meters. The #4606 coil should have five turns carefully removed, to cover the 10-20 meter bands; remove three turns from #4588 for six meters and two turns from the #4580 coil to handle 2 meters. These can all be resonated with a 50 pf midget variable tuning capacitor; a Hammarlund HF-50 or the equivalent is excellent. Any of the small multi-point selector switches can be used for band switching. A JBT lever switch type SS-14-ILS will handle the job or, if you prefer a rotary switch you can use a Mallory type 3215J as I did, leaving two positions unused. (I might want to add a couple of coils and try 40 and 80 later!).

MIN
6
10
15
MA
20
BAND
PHONES
ON
OFF

Resistors R-1 and R-2 shown in the schematic, need be only ½ watt. A single flashlight cell will serve for the battery although I chose a 1.4 volt mercury transistor battery to conserve space. Other items shown on the schematic are obvious and all parts are readily available from most electronic parts distributors as well as from the electronic mail order houses.

Build this little F/S meter and know what your VHF outputs are doing. If you want to check your modulation quality, a phone jack may be added as shown, making this little gadget really versatile.

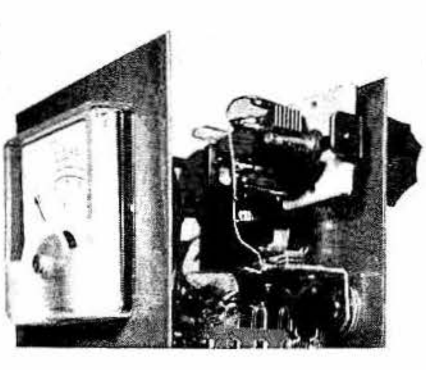

Side view of the FSM.

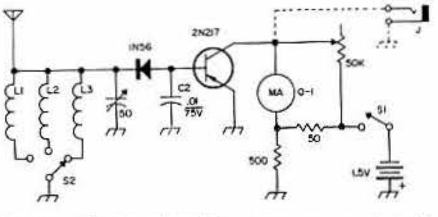

Fig. 1. The bandswitching, transistorized field strength meter for two through twenty meters.

AN AMPLIFIED, CALIBRATED SIGNAL STRENGTH METER

J. L. Iliffe VE3CES

Recently, I had the problem of tuning a four element quad. As you may or may not be aware, these beasties are supposed to be tuned from the rear for minimum signal. A quick check showed none of my friends had a signal strength meter, so I prepared to degrade myself and buy one. A look at a few prices convinced me to build.

Since I wanted some other information on the quad, like front to back ratio and the effect of more or fewer elements, I decided to add a calibrated attenuator and enough gain to make a fairly wide input range. It also had to be cheap!

The result is shown in Fig. 1.

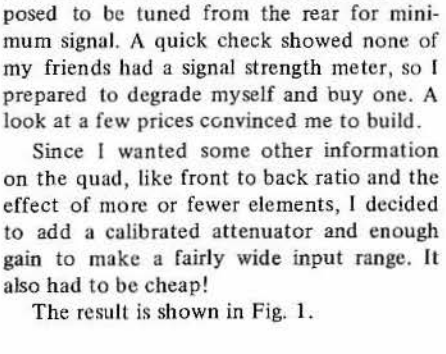

Fig. 1. Diagram of the field strength meter.

To use the normal rf attenuator method of switched T-pads requires complicated shields and quite a few resistors. Also the attenuator has to be terminated in its characteristic impedance to read correctly. To bypass this problem I first detect the rf, then attenuate the dc. This has the added advantage that the circuit is no longer frequency sensitive.

The incoming rf is tuned by C1-L2. C1 can be any small variable. I used both sections of a dual 15 pF because my local surplus store has them for 60¢. For VHF use only 1 section.

L2 is wound on a plastic pill bottle about 1" in diameter and tapering to 7/8". To cover 13–24 MHz, I used 11 turns spaced over about an inch. L1 is 2 turns over top of L2. I tried bandswitching with another pill bottle fastened on the other side of the shield from L2. The idea was to bandswitch another frequency range but I find it more convenient to wind on coils as needed. Use the grid dipper to get you in the ballpark. I have used this meter as high as 72 MHz without trouble. D1 can be any diode. I used a 1N34 because I could then specify it and know it would work, but I tried a computer type which also worked. If you prefer the meter to peak rather than dip, reverse D1.

Rectified rf from D1 is put on the top of R1, the calibrated attenuator. R2 in series with R1 gives the 0 dB point at its junction. For a 30 dB range, R2 is 47K if R1 is 1 MΩ This doesn't quite fill the range but is close. Changing the value of R2 will change the range but 30 dB is considerably more than the F/B ratio of most beams.

Q1 is a 2N5033 FET. The high input impedance of Q1 allows us to set the calibration of R1 directly by dc voltage measurements on the VTVM since it does not draw any base (gate) current. It is a p-channel device. If you use an n-channel type you will have to change the entire biasing of the circuit and also reverse D1.

Q2 and Q3 form a differential amplifier to drive the meter. Q2 is necessary to avoid loading Q1 and I had quite a bit of trouble balancing the meter against battery voltage changes until I added Q3. It will now operate from 8.5–9.2 volts with no trouble. The 10K pot in Q3's base centers the meter. The meter I used is a 250 μA tuning meter with no markings on it except a red/white/blue bar. This is all you need since we calibrate on R1, not the meter.

Operation

The meter is quite sensitive and with a two foot antenna I could get a reading several hundred feet behind my quad at 60 watts input. First tune the input (which is quite sharp) with the attenuator set at zero. This is the least sensitive position. Now set the meter for a convenient reading near the center scale with the incoming signal still on using R3. Adjust your antenna. When you feed it power again the reading will not be quite on scale on the meter but turning up R1 will allow you to put the meter back to the original position. Do not touch R3. The reading on R1 is now the increased gain in dB needed to bring the signal back to its original strength. In other words, the decrease in signal strength.

Note that during measurements you need a received signal to use R1. With no signal the meter will be off scale.

I have also used the meter to align oscillators and doublers in my two meter receiver. A probe can be made for this from two turns of wire on the end of a piece of coax. The high gain available allows the pickup loop to be quite far away which reduces detuning. Adjustments show up well on the meter.

FREQUENCY MEASURING EQUIPMENT AT MICROWAVE FREQUENCIES

Silas Smith WA9VFG

This article is not intended to give the theory, but rather a practical solution to the building and use of wavemeters at microwave frequencies.

In microwave work, frequency is one of the most important measurements. It must be understood the wave length in the devices described here is not the exact frequency wave length. A well-constructed wavemeter that has been calibrated can be very precise. They can be within 1.5 MHz at 10 GHz or less than half of a MHz at 1250 MHz. Temperature has some effect on the frequency. Most commercial wavemeters are constructed of Invar, a metal that changes very little with temperature. Some parts are of bi-metal construction to compensate for temperature. For the average experimenter, brass and copper will have to suffice. Although silver plating is desirable, it isn't an absolute necessity. Frequency at microwave frequencies can be measured by three methods: wavemeters, slotted lines, and frequency comparisons. All of these methods are used commercially. The frequency comparison is usually used in the laboratory to calibrate the wavemeter and the slotted line. As a general rule, any method of frequency measurement used at lower frequencies can also be used at the microwave frequencies, but are not always practical. The resonate cavity as a wavemeter is used in microwave measurements.

There are three types of cavity wavemeters: the transmission type, Fig. 1A and 1B, the reaction type, Fig. 1C, and the assorption or absorption type, Fig. 1D. All are resonate cavities. The way in which the wavemeter is used determines the type.

All wavemeters are adjusted for maximum readings except the assorption type. The assorption is adjusted for a dip in power output. The most popular wavemeter used by the beginner is the open circuited transmission line type, Fig. 2. This type of wavemeter is the equivalent of lecher wires. (Open circuit refers to the standing wave within the cavity, not the physical construction except as it pertains to the frequency wave length.) The practical physical dimensions are not many. The inner circumference of the main tube should be less than one wave length at the highest frequency to be measured. The rod should be small compared to the tube. If inductive coupling is used, the inductive coupling should be close to the shorted end. For probe coupling, the probe should be close to the middle. The open circuited transmission line is generally used in two ways. This type can be used "in line' (Fig. 1B) as it has very little loss when it is resonated. However, it should be removed from the line before transmitting, as it will act as a narrow band filter. The half wave length is the measurement between the two successive points at which the generator will load to maximum, as the rod is inserted or withdrawn.

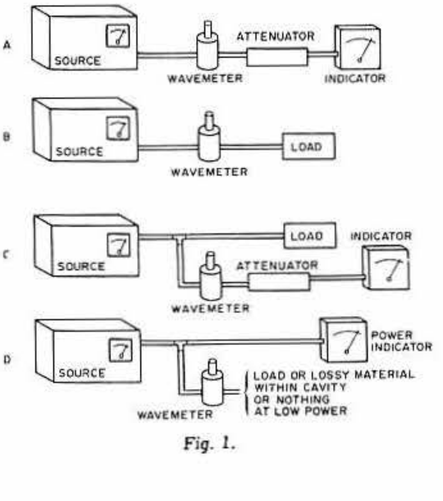

Fig. 1.

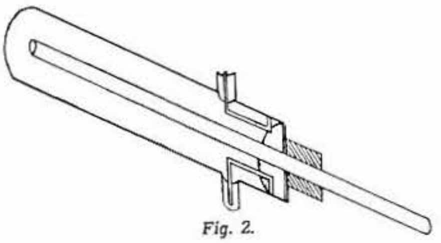

Fig. 2.

Another method in the use of the open circuited transmission line calls for the use of an additional circuit, as in Fig. 3. The circuit is a simple crystal diode detector connected to a microampmeter. The diode and condenser are usually built into the connector, as the leads should be kept as short as possible. The half wave length measurement is made on the rod between two successive maximum readings on the meter, as the rod is inserted or withdrawn. See Fig. 1A and 1B for the setup.

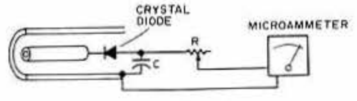

Fig. 3.

The quarter wave coaxial cavity is actually a shorted coaxial line one quarter wave length long (Fig. 4, 5, 6). As illustrated in these figures this type makes a very good cavity to use as a standard. To calibrate, a chart is made of the micrometer settings at different frequency wave lengths from a calibrated source generator. Don't tell them so, but the Public Relations Department of the Telephone Company may help you here, if they have any microwave technicians close by. The quarter wave coaxial cavity can be either physically open or closed. If closed, the closed end should extend at least a quarter of an inch beyond the center conductor at its lowest frequency. The closing of the end will lower the resonate frequency. Probe (capacitance) coupling as used for coupling in Fig. 5 will shorten the center conductor, and loop (inductive) as used in Fig. 4 will lengthen the conductor. In Fig. 4 we change the length of the center conductor to change its one quarter wave length. In Figs. 5 and 6 the center conductor remains the same, and we change the resonate frequency by capacitance. This method makes it necessary to construct the center conductor very short as compared to the full quarter wave length as in Fig. 4. These devices are not longer, so caution must be used when calibrating. The closed wavemeter as indicated in Fig. 7 is a shorted coax line at each end. The wavemeter uses a shorting plunger which is movable along part of its length. If used as the quarter wave coaxial cavity, the center conductor must be longer than a quarter wave length.

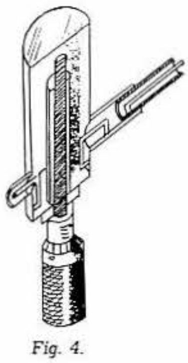

Fig. 4.

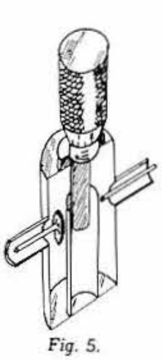

Fig. 5.

Up to this point we have covered most of the wavemeters that could be used from around 144 MHz up to approximately 3000 MHz. 1000 MHz to 3000 MHz are usually called the lower microwave frequencies. If the inner circumference of the outer tube is kept less than one wavelength, these wavemeters will operate in the desired TM mode.

There are four ways to couple energy into a wavemeter, loop (Fig. 4), probe (Fig. 5), direct (Fig. 6) and slit (Fig. 8). The most commonly used is the loop, as it has very little effect upon the electric field. The usual methods for changing loop coupling is to change the size and orientation of the loop. Loop coupling is usually placed in the high current area of the wavemeter. Capacitative coupling is changed by the size of probe and the distance from the center conductor. Capacitative coupling is usually placed at the high voltage portion of the wavemeter. As in Fig. 5, a small probe – say ½ in. piece of No. 22 wire, for example – may require an external voltage amplifier. The smaller the probe, the less effect on the resonate fre-

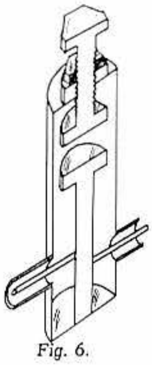

Fig. 6.

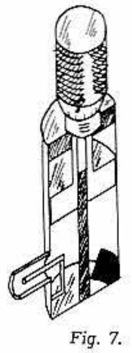

Fig. 7.

quency of the wavemeter. In direct coupling as in Fig. 6 the primary concern is impedance. To increase the impedance, move the coupling up the line away from the shorted end. To decrease the impedance, move the coupling down the line toward the shorted end. Slit coupling (Fig. 8) can be a small hole or a series of small holes or a slit. Its purpose is to allow a certain amount of leakage. In all forms of coupling, it is desirable to use loose coupling, as the wavemeter will have less effect on the system, and the Q of the circuit will be higher.

Above 3000 MHz, usually only the tunable wavemeter is used. The cavity is one quarter wave length long. The cavity can be coupled in three ways: loop, probe and slit. Because of the high frequencies, the slit is usually used, and the meter is most often kept as an assorption meter. In the assorption method, the wavemeter should be detuned when not in use. Some of the cavity wavemeters have a little lossy material added to absorb some of the energy, as in Fig. 7. Lossy material can be made from graphite impregnated cloth in epoxy.

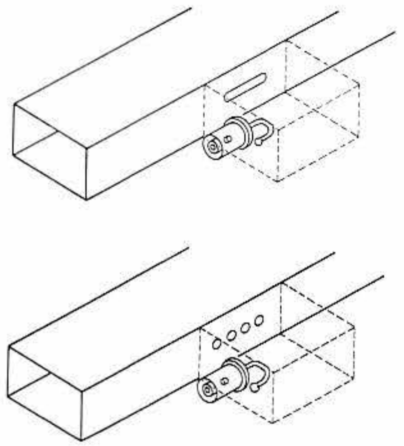

Fig. 8.

There is one other type of wavemeter that can be briefly mentioned; it is the reference wavemeter. The reference wavemeter is of any design as described, but would be constructed more like Fig. 6. It can be locked when adjusted to a selected frequency and used as a reference standard. The micrometer assemblies can be made from any micrometer with additional parts welded on. I constructed one using an oversized tube over the main cavity, and dimpled it at various places around its circumference until it fit smoothly over the cavity, and I used a piece of 3/8 threaded brass pipe as the main adjusting screw. I am sure you can come up with a good one without any backlash. This wavemeter spread the 1250 MHz band out to over 100 inches by rough measurements. I haven't calibrated it, so I can't say for sure just how far. There are three nice veeder root counters in the APX 6 which would make excellent wavemeters plus sliding contact material. One could even use the entire cavity.

The last method of microwave measurement that we will look at is the slotted line (Fig. 9). The slotted line is a section of coax line along which is cut a slot. A probe, which is a simple crystal detector with a one quarter wave length shorted stub for a dc return path, is moved along near the center conductor of the slotted section. In this case we are looking for two successive minimum readings along the line. The distance between these readings is one half wave length. A slotted line should also be calibrated. If calibrated at one spot near the intended frequency to be measured, a chart will not have to be made – just a K factor obtained. The distance between two successive readings times the K factor should equal the frequency half wave length. There are a few accessories that can be either built in or used externally. One is a coax attenuator shown in Fig. 4. It will slide in and out of the other half. It too can be calibrated if one wishes. A 50Ω resistor can be used for an impedance match for 50Ω lines if inductive coupling is used such as in Fig. 10 and when the loop is small. The line stretcher (Fig. 11) is useful with the slotted line. It merely consists of two coax sections, one sliding into the other.

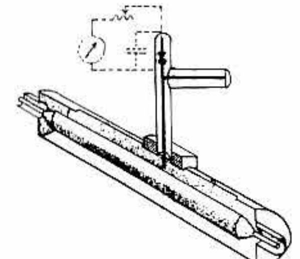

Fig. 9

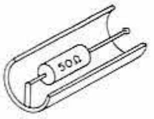

Fig. 10

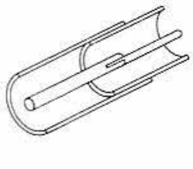

Fig. 11.

FREQUENCY METER – 1 to 10 GHz AMATEUR MICROWAVE

Bill Hoisington K1CLL

A quarter wave coaxial cavity is used up to about 5 ghz, and from there to over 10 ghz the three quarter mode is used. A complete explanation of these types of operation is given.

The same type of unit can be used as a very good tuned mixer from 1 to 10 ghz.

The Coaxial Cavity

The basic circuit of the coaxial cavity is shown in **Fig. 1**. A cylindrical outer cavity wall encloses a round rod some 4 inches long which is the center conductor–this center conductor is grounded at one end.

The Shape of the Cavity

The exterior shape of the cavity is shown in Fig.2. and is seen to be rectangular in cross section, with two thin walls and two thick side walls. Believe me, this configuration was not arrived at in one day! De-

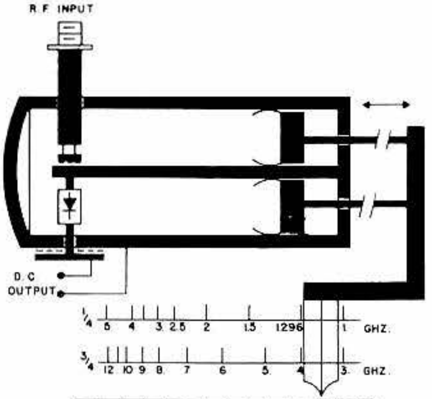

Fig. 1. Basic coaxial cavity.

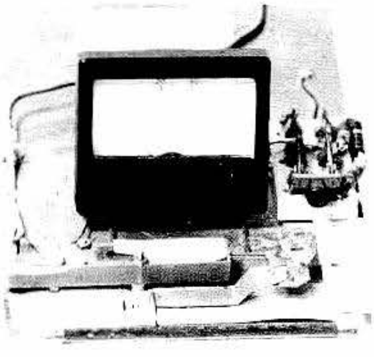

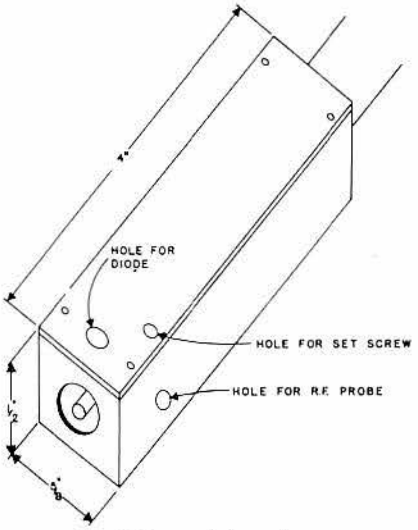

Fig. 2. Shape of the cavity.

signing tuners for X Band, I gaily started in with sections of thin-wall round pipe, the way I'd always done on uhf. The first thing you run up against is, how do you make the diode bypass capacitor? Machine out a curved saddle piece to fit exactly over the outer wall? Possible, but too expensive. And then how do you introduce the rf probe coupling into the cavity? Add on a "saddle" with a hole in it? These considerations and others, such as mounting (more saddles?) led to the abandonment of the pipe as a shape for microwave cavities; but not until a lot of time had been spent on the above mentioned items.

Diode holder and capacity

Looking at **Fig. 3.** you will see the first answer arrived at; but only after weeks and weeks of making different types and shapes. The center conductor is slightly flattened and drilled out to fit the diode prong. An 8/32 copper machine screw is drilled out to fit the other prong, then slotted with a fine jeweller's saw, and then compressed slightly to an inside diameter a shade less than the OD of the diode prong. In this way the copper screw will hold the diode as you insert it into the cavity. Believe me, that helps!

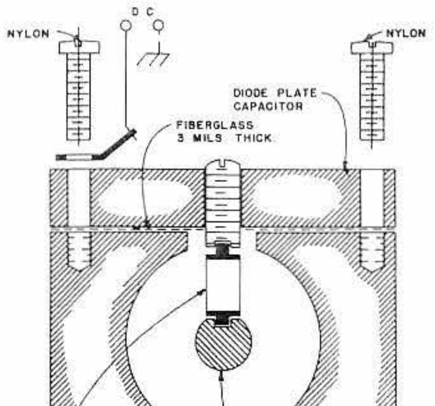

Fig. 3. Diode holder and capacitor.

The second answer is also evident from **Fig. 3.** as the diode bypass capacity can now be made efficient at X-Band. As mentioned before, you cannot "buy" a capacitor "good for X-Band. You can make it though, as shown in **Fig. 3.**, if the cavity body has been designed correctly for it. One of the thin wall sides of the cavity is drilled out (or machined out) just wide enough to clear the diode and it's holder, which is the 8/32 copper screw. The copper capacitor plate, which is thick enough to take at least a half dozen 2/56 threads, is drilled and tapped for the 8/32 screw, and clearance drilled in the corners for the 2/56 mounting screws. A soldering lug for the dc connection is used under one of these, and a three mil (three thousandth of an inch) thick sheet of fiberglass cut out to fit, larger than the plate. This helps to keep metal particles from lodging inside the tiny crack that might be there if the fiberglass sheet did not extend out beyond the plate all the way around. You can begin to see some of the detail needed at X-Band.

Further reasons for the rectangular cross-section now show up in **Fig. 4.**, which details the rf probe connections. This item was also very troublesome in first models using pipe walls, where "more saddles" was the only solution. All "saddles" are eliminated by the rectangular shape. Small semi-rigid cable is used for the connector. I have some short lengths with X-Band antennas connected to them for use as "In-Space" pick-ups, feeding directly into the wavemeter cavity. There is at times an advantage in this type of "energy collection" (antennas) which will be taken up later.

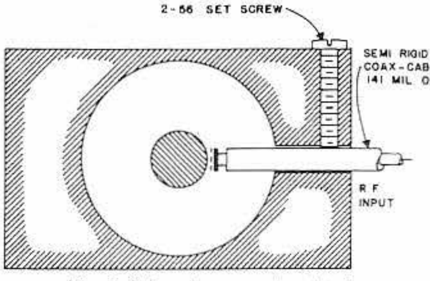

Fig. 4. RF probe connector detail.

Fig. 5. shows detail of the treatment of the cavity end of the rf cable, or probe. The outer conductor is cut away for about one quarter inch in length and removed. About a sixteenth or so of the Teflon is left, which is then removed from the center conductor. A thin copper washer (which I generally cut out of sheet copper since the hole to solder the center conductor is quite small) is then soldered to the center conductor, making the "capacity probe", as shown in **Fig. 5.**

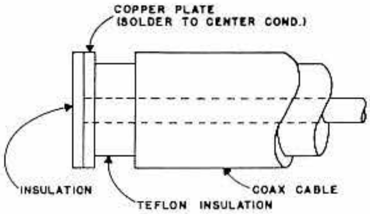

Fig. 5. RF probe detail.

Mylar tape or other good insulation is fastened to the side of this washer facing the center conductor. With this insulation in place you can push the probe all the way in, while testing, and still not have a dead short. Different thicknesses of fiberglass sheet can also be cemented on, to make up more permanent types of fixed capacitors, of different values.

For some uses, particularly in this one as a wavemeter, loose coupling is desired, but it must be securely locked with the set screw, otherwise your dial calibration and frequency reading will suffer.

Plunger fingers

Here is the most difficult item. It is hoped to have stock pieces made up for this work that you can purchase at reasonable cost. The fingers should be made of tempered beryllium-copper, which is not easy to work with.

Fig. 6. shows some details of the plunger and fingers. I assume, having been told so by "well-informed sources" (mechanical engineers) that these units should be made in a machine shop by competent machinists. Maybe so, as the ones I have made here in the shack by hand tend to lose their tension if not handled carefully.

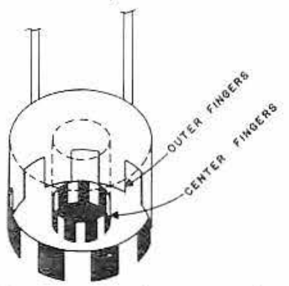

Fig. 6. Plunger details: A)End view, B)Outer fingers, C)Center fingers.

Fig. 7. shows the desired fit for these fingers. The plunger body should be an easily slide-fit *inside* the ¼ inch cavity, and the center hole in the plunger after the fingers should also be an easy fit *over* the center conductor.

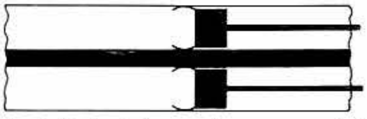

Fig. 7. Desired shape and curvature of the plunger fingers.

Two steel push rods lead back from the plunger through small holes in the back end of the cavity (see **Fig. 1.**); these terminate in the brass block which is furnished with a pointer for the frequency scale. Maximum extension of the plunger should be up against the end piece, as a positive reference point for the dial, in case of trouble after calibration. This point should be indicated on the scale as "minimum frequency" in order to reset the pointer if it should ever become displaced after calibration.

The diode

At present, the diode used is an X-Band "pill package," with a prong at each end as shown in **Fig. 3.** These are point-contact diodes, like the famous 1N23 ceramic cartridge types of World War II fame, only a lot smaller. Referring again to **Fig. 3.** always make sure that the ceramic part of the diode is, as nearly as possible, in the open space between the inner and outer conductors. This space is where the rf is! It is also important to make sure that there is as much metal surface continuity as possible along the cavity wall, across the fiberglass sheet X-Band capacitor insulation onto the diode capacitor plate, and from there over to the diode holder and onto the metal end of the diode.

The rf is at a maximum between the inner and, outer conductors, which is an air space of a sixteenth of an inch. and that is where the diode should be.

The diode rf bypass capacitor, formed by the diode plate and the flat top of the cavity body, need only have a capacity which is relatively small; anything over about 20 pF is sufficient. What it *must* have is the proper *lack* of inductance! The details of how this act has been covered in previous paragraphs, and if you follow those details you will find little or no rf on the *outside* of the diode capacity plate or the dc lead from it.

X-Band is not just short waves. It is really short; like a quarter wave at X-Band equals 9/32 of an inch as you can plainly see, if you get one (or more) of those little plastic millimeter rulers in a stationery store for 5 or 10¢. Be sure and get some, by the way, if you're going to do anything above two meters.

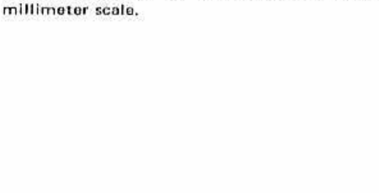

Fig. 8. Full waves for S, C and X-Bands on millimeter scale.

Fig. 8. shows the millimeter scale, with s, C, and X-Band plainly showing.

A handy wavelength-frequency chart is included here for your convenience, which is useful from the khz range way up *above* X-Band. See **Fig. 9.** Get to know the easy reciprocals, like 1 centimeter equals 30,000 mhz, 3 centimeters equals X-Band, 10 centimeters equals S-Band (3,000 mhz), 1,000 mhz equals 30 centimeters, etc. Very useful!

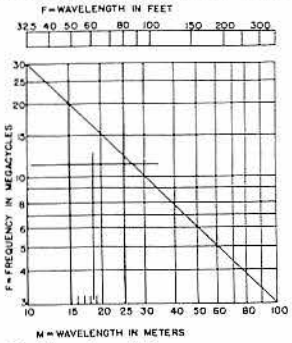

Fig. 9. Wave-length/frequency converter. Use of multiplying factors such as those at the bottom of the graph will cover any portion of the electromagnetic-wave spectrum.

The 3/4 mode and harmonics

Don't worry about that word "mode." Generally when something odd takes place in a cavity or waveguide, certain types of engineers tend to fall back on obscurantism (I seem to have fallen for that $64 word. It just means covering up). They say, "It jumped mode", or, "Spurious showed up."

Here's the straight dope. **Fig. 10** shows the quarter wave "mode" of operation. Starting at 1 ghz you will find *one* point of maximum dc output. If the oscillator under measurement is "running hard" with lots of 2nd and 3rd harmonic energy content, these will be found at 2 and 3 thousand megahertz, and possibly higher ones, which should drop steadily in power as you go up. The diode itself may cause some of these if hit too hard with the rf input.

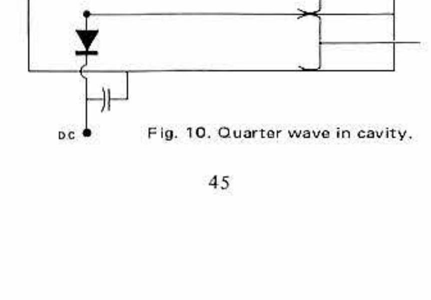

Fig. 10. Quarter wave in cavity.

Fig.11. shows the 3/4 wave mode, which is a very "natural" type of operation. Don't forget that in an instrument of this kind you are looking for standing waves and you want them to be of the greatest amplitude possible (within reason). So, if you tune the cavity by the plunger so that it measures some three quarter waves on it (allowing for length-loading of the diode on the first quarter), you will find two peaks on the meter due to the situation shown in **Fig.11.** The higher the Q, and the lower the losses along the line, the more quarter waves can be found. For the 4 inch cavity shown, three quarter waves at S-Band are the longest that will fit.

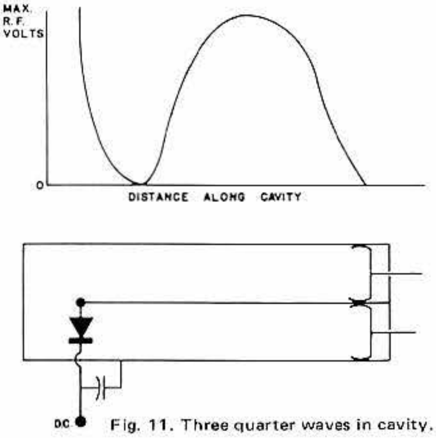

Fig. 11. Three quarter waves in cavity.

A check on this operation is easy. Using the millimeter scale on the "dial", take several readings between maximums, for example, 22, 37, 51, and 67, add the spacings together, which comes to 45 millimeters, divide by three (the number of samples), and you will find an average of 15 millimeters for the waves which are standing on the center conductor (or "along the cavity", if you prefer) and there you are, 15 millimeters for the half wave, 3 centimeters for the full wave. Which is X-Band at 10,000 mhz or 10 ghz.

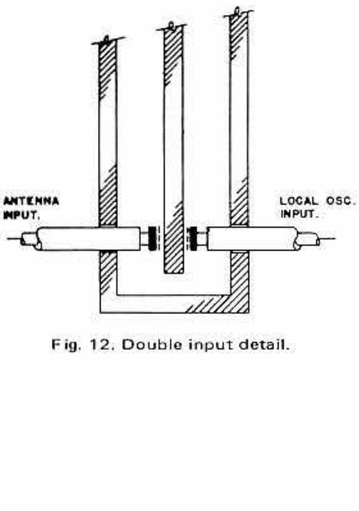

Fig. 12. Double input detail.

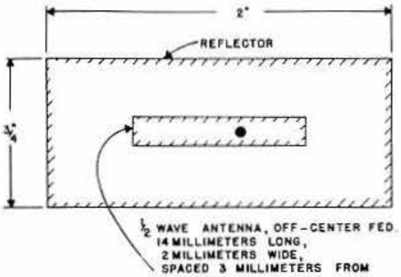

Fig. 13. Test antenna–X-Band "two-element".

If you find numbers which are not well known, you can find the frequency on the chart, at least close enough to put you in one of the microwave amateur bands, such as 5,650 or 10,500 mhz.

Use as a Microwave Mixer

This same type of cavity can be used from 1 to 10 ghz as a mixer for the front end of a superhet receiver covering those frequencies.

This application will only be touched on briefly here as the whole receiver is detailed in another article in *73 Magazine*.

Fig. 12. shows how to do it, so you can plan on this use, and make more than one, if you wish to.

Looking at Fig. 12., you can see how useful it is to have *two* thick sides on the cavity, one for rf input and one for the local oscillator input.

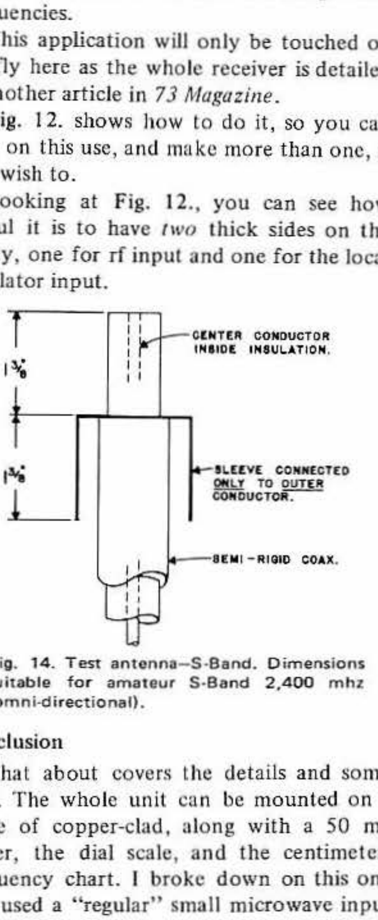
CENTER CONDUCTOR INSIDE INSULATION.
1¾
1¾
SLEEVE CONNECTED ONLY TO OUTER CONDUCTOR.
SEMI-RIGID COAX.

Fig. 14. Test antenna–S-Band. Dimensions suitable for amateur S-Band 2,400 mhz (omni-directional).

Conclusion

That about covers the details and some uses. The whole unit can be mounted on a piece of copper-clad, along with a 50 ma meter, the dial scale, and the centimeter-frequency chart. I broke down on this one and used a "regular" small microwave input connector for the rf. (Instead of an "RCA Phono Jack".) For connections to other units, such as oscillators and multipliers, small flexible cable may be used.

Fig. 13 shows a test antenna for X band, not the best in the world but good enough for a starter. With a lens in front it really picks up signals. Fig. 14 shows an S band antenna for the 2,400 MHz amateur frequencies.

LAMBDA LINES

James Ashe W2DXH

Over the past few years, new kinds of commercial gear and military surplus have made it quite easy to generate VHF frequencies. But the problem of measuring the frequencies has not become any easier for amateurs short on calibrated instruments. Of course this means the newcomer to VHF! What can he do to find what ballpark a circuit is radiating in, if there are no accurate devices available? When this problem came up recently, a simple solution appeared quite by accident. It was so simple, in fact, that its simplicity must be the feature that has kept it out of the ham publications. Another first for 73!

The traditional solution to the rough frequency measurement problem is to make up a Lecher Wire system. There is some question about the value of one of these in the modern ham shack. Narrow-band crystal-controlled techniques guarantee frequency and stability once the multipliers are tuned properly. In the old modulated oscillator days things were not that stable . . . so why go to all that carpentry and construction work for what fairly well promises to be a use-it-once gadget? Particularly when a little reflection (pun intended!) may bring out a cheaper, faster and better arrangement?

Ham and commercial builders of VHF gear have been using tuned stubs for years to match impedances, tune out frequencies, tune in others, etc. Yet it seems to have occurred to very few workers indeed that it might be possible to cut stubs to length accurately enough to serve as frequency standards. Apparently this can be done, with an accuracy of about 5%! This compares very favorably indeed with the performance of lower-frequency grid dip meters and some signal generators. It's pretty good for a pencil and yardstick operation: the only other items required are some understanding of how it works and a piece of 300 ohm twin lead. Belden #8235 recommended. You can calibrate that new GDO for 432 at a cost of just a few cents!

Theory

Many kinds of things show a property of tuning sharply to a certain frequency. This property is called resonance. We hear it when a struck piece of metal rings, and see it in the pendulum of an old grandfather clock. The grid dip meter shows a drop in grid current of an oscillator when a nearby resonant circuit steals energy from the oscillator. And it is the nearby resonant circuit that is the subject of this article.

The basic circuit is the quarter-wave stub. A little browsing around in the handbooks and earlier issues of 73 and other ham magazines will tell you lots about quarter wave stubs. The important practical points are that the stub resonates at certain frequencies, and that at these frequencies it can be dipped at its shorted end in the same way as any other resonant circuit.

But the term 'quarter-wave' has to be taken with a grain of salt. The tuned stub will be shorter than a free-space quarter wave, because the dielectric has a slowing-down effect on the rate at which the RF bounces end-to-end along it. Suppose you laid out a mile or so of twin-lead and transmitted a signal, at the same time sending off a reference signal by space wave. The reference signal would arrive at the other end first, in about 5.35 microseconds. The twin-lead signal would arrive a full microsecond later, about a 20% delay. Since this applies even to short lengths of twin-lead, the delay must be taken into account for accurate measurements. Also, there is a considerable difference in velocity factors between different brands and qualities of twin-lead.

Crystals are often used in overtone oscillators for generating stable VHF frequencies. The various modes of oscillation are pictured in the handbooks. Tuned lines will also show overtone resonances, and in the case of large uncertainty, it might just happen that reasonable errors could lead to a consistent but very wrong result. A halfwave line will resonate at a frequency f, and also at 2f, 3f, and so on. Note both odd and even multiples! All other resonant lines have a similar overtone resonance property. The problem is slightly aggravated by the convenience of using relatively long lines at the higher frequencies because they are easier to handle. The solution is to cut a pair of lines whose collections of resonant frequencies have only one resonance in common. The recommended lengths are a halfwave and a three-quarter wave line.

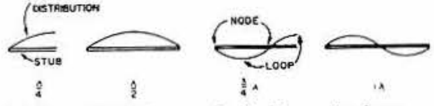

Fig. 1. Four kinds of tuned stubs, taking their fundamental resonances as the frequency producing the illustrated voltage distribution. This shows that none of them has a unique resonant frequency.

Fig. 1 shows four basic tuned lines. Just which resonance is an overtone and which is not depends somewhat on the application. The simplest way out of this problem in semantics is to say that the three-quarter wave line really doesn't have that resonance at f/3, ignore the quarter and fullwave lines, and stick to the remaining two for test work.

At 432 MHz a wave in free space is about 27.3 inches long. Suppose we are using Belden #8235 twin lead, which Belden says has a velocity factor or propagation constant of 0.77. The twinlead wavelength than is 27.3 times 0.77 or 21 inches. The halfwave stub must be 10.5 inches long, shorted on both ends; and the three-quarter wave stub 15.75 inches long, shorted at one end. These are convenient lengths, not too long to use on the workbench, nor so short that percentage accuracy in cutting becomes a big question.

Using the stubs

The commercially available grid dip meters are not noted for accuracy. It's commonly estimated that the scale calibration can be

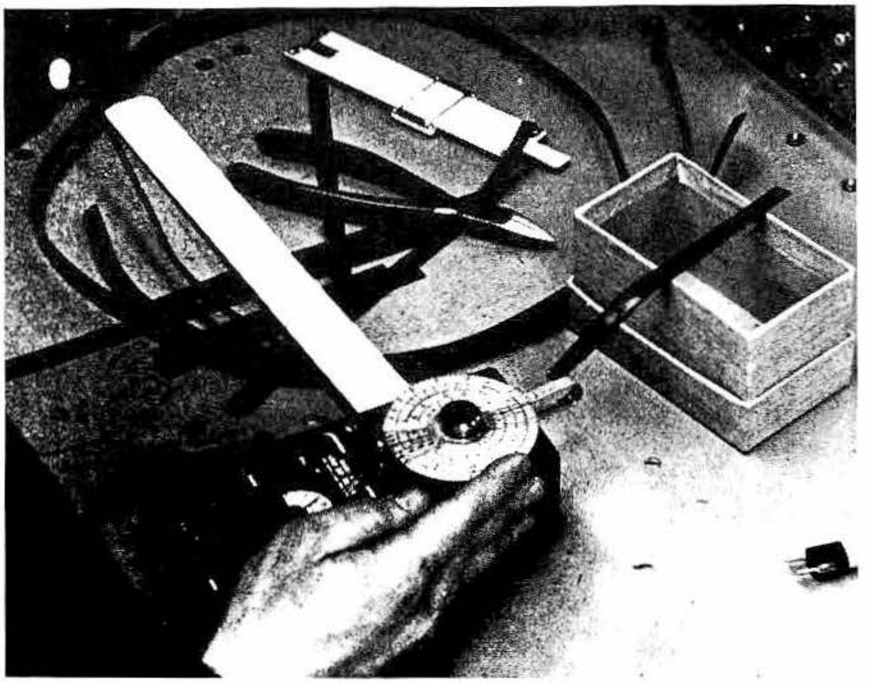

trusted to within about 20%. With some care, calibration points taken from twin-lead resonators appear to be good to about 5%. The first precaution is accurate construction. Cut the strips slightly long, short one end of each, and cut the three-quarter wave line to length. Then go more carefully at the other end of the halfwave line, which must be shorted at both ends. It should not be too hard to get the correct lengths within one sixteenth inch.

When making frequency checks, the lines must be held off the workbench an inch or two. Use small boxes or pieces of cardboard. Probably the better part of a foot distance is in order if the workbench is of metal or has a copper surface. At two meters, a perceptible change in calibration can be detected if the line is laid out on a wood surface! It's very good practice to make up lines for two meters or lower, and practice dipping them. Some refinement of technique will certainly be required before a halfwave and a three-quarter wave line can be made to dip at the same point on a standard dip meter. Once the trick is mastered, it can be carried up to the higher frequencies.

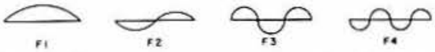

Fig. 2. Some overtone resonances of a halfwave stub (shorted at both ends). F1, F2, F3 and F4 should be 1F, 2F, 3F, 4F.

The lines are dipped in the same way as any coil. Because they have a very high Q, there will be a tendency for the dip oscillator to pull, or to seem to give different readings when tuning down to frequency and tuning up to frequency. The remedy is less coupling: move the dip meter a little further away from the line and try again. Dip the stub at its shorted end!

But what was that trick for calibrating a dip oscillator, mentioned earlier? Can't make up a pair of lines for each frequency. No need to! That's simply the reliable way for finding the right ballpark. When you're there, you can set the lines aside, make up another three-quarter wave resonator cut for the lowest frequency, and after marking that point on the scale, trim the stub up to the next calibration frequency. Throw the remainder away when done calibrating.

The half-wave stub is also useful as a tuned coupler. Suppose you want to tune an oscillator to a particular frequency but have nothing to indicate at that frequency. Loosely couple the RF into one end of the half-wave stub, and take it out the other end with a hairpin loop, through a diode to a 50 µA meter. You will only get a reading at the resonant frequency of the half-wave stub. Simple!

Chapter V

RF Signal Generators

ZERO-BEATING WITH A FREQUENCY METER

Jim Harrison WB4TBX

After putting an LM-18 in service, I found it a nuisance to plug in headphones every time it was necessary to adjust the corrector control to produce Zero Beat at one of the crystal check points. I hooked up a phone plug to the plate winding of an audio output transformer, and then connected a small loudspeaker to the secondary winding. I leave the phone plug inserted in the LM phone jack at all times. After using this setup for a while, it occurred to me that the last few cycles either side of zero beat might not be audible, so I tried a parallel hook-up off the primary side of the transformer. The other end of the cable was connected to the Vertical input of an oscilloscope. Now, when adjusting the corrector control for zero beat, you will see low amplitude sine waves even after you can no longer hear the beat note. You tune out the sine waves until you have a perfectly straight reference line on the scope. The Heterodyne oscillator is then corrected to calibration much more accurately than trusting to the ear alone. The principle is not new, yet the idea may not have occurred to some people. Any old scope will do, as the frequencies are in the audio range.

Another bit of frustration occurred when trying to adjust the corrector knob. I would pass the zero beat point time and time again, due to the stiffness of the control. Remember, these units are built for ship-board use, and loose controls cannot be tolerated. I dug up a small vernier tuning drive, similar to the Jackson Planetary-Vernier drive, (about 5 to 1 ratio). I mounted this drive on a right-angle aluminum bracket and attached it to the LM using the two screws on the upper right side of the unit. Bore holes slightly oversize to allow for accurate alignment of the control shaft so that it will not bind. Cut out a portion of the bracket so the "High-Low" knob can be moved. The precise Zero beat adjustment can now be made very easily with the reduction drive.

HAND CALIBRATE THE BC-221 FREQUENCY METER

Carl Henry

At present many amateurs have an opportunity that may not repeat itself. I refer to the many army and navy surplus frequency meters now available, through MARS and surplus sales. Many of these meters have the original calibration chart missing, and are considered virtually useless. This is demonstrated graphically by prices, the frequency meters with charts costing $100, the meters without charts selling for one-third this amount. Meters without charts are worth more than this would seem to indicate. It is no great job to calibrate one of these meters, but it is time consuming. A little care is required, but you do not need to be a super technician or electronic engineer to do the job. Neither is elaborate equipment required.

Essential to the calibration are several items not commonly available around the shack, but easy to build. Figure 1 illustrates the first requirement. This is a 10 kc multivibrator, which will operate from a 100 kc xtal calibrator, and give 10 kc markers up to 20 mc. I am assuming here that you have or can beg, borrow, or steal a 100 kc calibrator, this being a common item. In fact, some receivers have them built in. Using this 10 kc marker source, the high band of the frequency meter can be calibrated. Interpolation to 1 kc points is then possible, and if care is used an accuracy of better than 0.005% will result.

Fig. 2 shows a simple beat detector that can be used with earphones to get an exact zero beat. It took about ten minutes to wire up a 1629 (army surplus magic eye tube) temporarily for this purpose, and it is certainly worth while in the interest of increased accuracy.

With your equipment assembled, connect the frequency meter to its power supply. Allow at least one hour for proper warm-up so that all equipment will stabilize, including the 100 kc calibrator and 10 kc multivibrator. Set the CALIBRATE control on the frequency meter to center, and be especially careful not to move it until the calibration is complete. After the warm-up period the 100 kc calibrator should be zeroed against WWV, at as high a frequency as possible. The 10 kc multivibrator must be checked to see that it gives nine beats between each 100 kc beat. This can be done with your communications receiver, and it will be easiest to do it on as low a frequency as possible, say 600 to 700 kc. Now type several sheets of paper, listing the high range of your frequency meter by 10 kc points. If you are in doubt as to the range, check it with your receiver. To do this set the dial to the low end of its range, turn on the internal xtal, and zero in on the strongest beat at or near the low end. Turn off the internal xtal oscillator and find the signal from the frequency meter with your receiver. Note the frequency. Tune your receiver until you reach another harmonic from the frequency meter. The difference between the first and second readings is the frequency of the meter at the low end

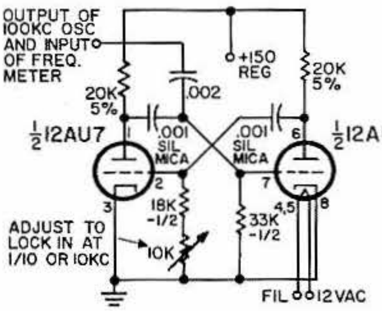

Figure 1. 10 kc multivibrator for use with 100 kc crystal oscillator, providing 10 kc beats with frequency meter.

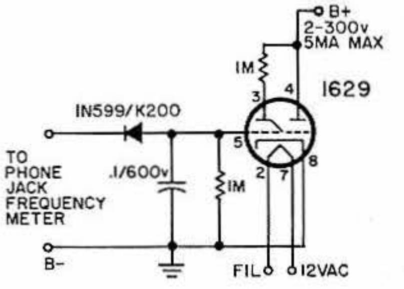

Figure 2. A simple zero-beat indicator for use in conjunction with headphones.

of its dial. Now do the same thing at the high end, and you have the primary coverage of your frequency meter. Most have a range of 125 to 250 kc on the low band and 2 to 4 mc on the high band.

When you write up the frequency range by 10 kc points, leave a space between notations, since you will be adding information here later. Now, using the 100 kc xtal, check the high band of the frequency meter at 100 kc points, listing the dial readings. Now go back and check the dial at 10 kc points, listing all the dial readings. The checks should agree with the 100 kc checks at every tenth point. This is a handy method of cross checking yourself. You might note that on equipment of this type, always approach the final reading from below. If you pass the zero beat point, don't jockey the dial back and forth for zero beat. Go back below the beat by 5 kc or so and approach again from below. This procedure will help to eliminate error from dial backlash.

After your 10 kc points are all listed, interpolate to 1 kc points. This is the hardest part of the job, and is very time consuming. A big pot of coffee and a patient and loving XYL will be a big help here. The patient XYL can be replaced with an adding machine if you have one available, but you'll still need the coffee.

The difference between each 10 kc point must be listed. Note this in the space you left on your sheets. Each 1 kc will be 10% of this, so add 10% to the 10 kc listing for 11 kc, 10% more for 12 kc, 10% more for 13 kc, and so on. When you reach the next listed 10 kc point, the calculated listing and the measured listing must agree, another cross-check. This will take 2000 individual additions.

After the high band is finished, you may wish to calibrate the low band in the same fashion. Referring to Figure 1, change the grid-plate capacitors to 0.01 mfd. This will put the multivibrator at approximately 1 kc intervals. If you have trouble syncing the multivibrator, you may have to build another to operate at 10 kc, and sync the 1 kc from this. No trouble should be had with a strong output xtal calibrator, however. Calibrate the multivibrator as before, except with 1 kc intervals instead of 10 kc.

A calibration book can be prepared when you finish. It is a good idea to file all your original calculations and papers, should the book ever be destroyed. The frequency meter is now as good as any with original calibration book, at a good saving of money.

HIGH ACCURACY VHF FREQUENCY MEASUREMENTS

Howard Burgess W5WGF

Accurate frequency measurements can be a problem for the amateur with a limited budget. However with some home construction and careful operation the average ham can make VHF frequency measurements to an accuracy better than .00015% at two meters. This is the equivalent of measuring the distance from New York to Los Angeles with an error of only 25 feet. Many commercial units cannot equal this figure. The same method can also be used for HF and UHF measurements.

There are many ways to measure frequency but few of them are satisfactory for use at the very high frequencies. The well known hetrodyne frequency meter becomes unstable when its oscillator is operated at VHF. It can no longer be held or read to any degree of accuracy. The oscillator can be operated at a low frequency and one of the harmonics used at VHF, but any error in the oscillator will be multiplied by the number of the harmonic used. A frequency meter that can be held to within 200 hertz at 4 MHz will be off by 7.4 kHz at the 148 MHz harmonic.

A second method of much greater accuracy uses low frequency crystals which are referenced to a known standard such as WWV. The harmonics of these oscillators will be quite accurate and useful far into the UHF region. However this system has its limitations. Even when used with multivibrators and harmonic amplifiers it produces only spot frequencies.

Although neither of these two methods is satisfactory when used alone, they can be combined to make an accurate and versatile system. If you haven't guessed it by now, the system works like this. A crystal oscillator operates on 5 MHz. This oscillator can be kept to zero beat with WWV with very little effort. With a simple harmonic amplifier following it, strong markers are available every 5 MHz far into the UHF region. To fill in between the 5 MHz points, and get full tuneable coverage, all that is required is to add the output of a stable low frequency VFO to the proper marker. Example: To measure 146.25 MHz just add 1.25 MHz from a calibrated tuneable oscillator to the 145 MHz harmonic of the crystal. The same results can be had by using the 150 MHz marker and substracting 3.75 MHz.

The tuneable low frequency oscillator of this hetrodyne system can be any stable, calibrated, oscillator that will give the desired frequencies. A good signal generator can be used but better yet is the old faithful BC-221 frequency meter. The crystal oscillator that supplies the 5 MHz markers should be designed for high stability. However, even simple crystal-controlled units can be kept zero beat with WWV for periods long enough to make most measurements.

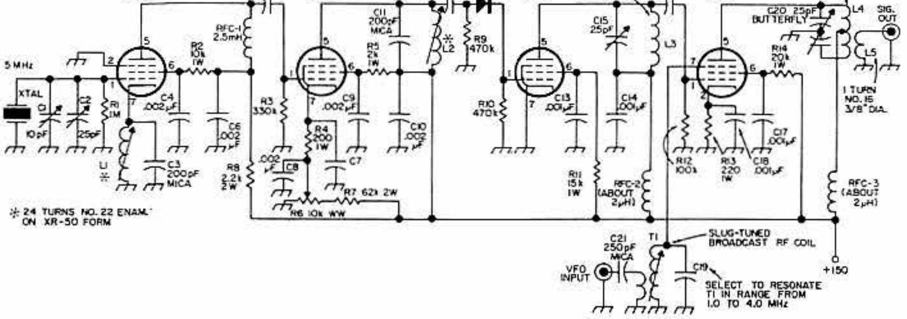
OSC.
6AUG
5 MHz
XTAL
C1
10pF
C2
25pF
R1
1M
L1
C3
200pF
MICA
C4
.002µF
R2
10k
1W
C5
RFC-1
2.5mH
C6
.002
µF
R3
330k
R8
2.2k
2W
AMP.
6AG5
.002
µF
C8
R4
200
1W
C7
R6 10k WW
R7 62k 2W
C9
.002µF
R5
2k
1W
C11
200pF
MICA
C10
.002
µF
C12
200pF
MICA
L2
R9
470k
CR1
1N34
ETC.
MULT.
6AK5
R10
470k
C13
.001µF
R11
15k
1W
2 TURNS
NO.16
1/2" DIA.
C15
25pF
C14
.001µF
RFC-2
(ABOUT
2µH)
L3
C16
20pF
MICA
MIXER
6AS6
R12
100k
R13
220
1W
C18
.001µF
C17
.001µF
R14
20k
1W
C20 25pF
BUTTERFLY
3 TURNS NO.16
3/8" DIA.
L4
SIG.
OUT
L5
1 TURN
NO. 16
3/8" DIA.
RFC-3
(ABOUT
2µH)
+150
C21
250pF
MICA
VFO
INPUT
T1
SLUG-TUNED
BROADCAST RF COIL
C19
SELECT TO RESONATE
T1 IN RANGE FROM
1.0 TO 4.0 MHz
✳ 24 TURNS NO. 22 ENAM.
ON XR-50 FORM

Fig. 2. Schematic of the oscillator, multipliers and mixer for using a BC-221 on the VHF ham bands.

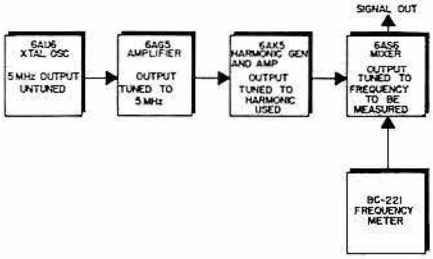
6AU6
XTAL OSC
5 MHz OUTPUT
UNTUNED
6AG5
AMPLIFIER
OUTPUT
TUNED TO
5 MHz
6AK5
HARMONIC GEN
AND AMP
OUTPUT
TUNED TO
HARMONIC
USED
6AS6
MIXER
OUTPUT
TUNED TO
FREQUENCY
TO BE
MEASURED
SIGNAL OUT
BC-221
FREQUENCY
METER

Fig. 1. Block diagram of the VHF frequency meter.

Earlier we quoted a figure of .00015% or better for the accuracy of this system. Perhaps we should show how this is possible. The crystal oscillator can be held to near zero beat with WWV but due to propagation errors in the signal of WWV, we can never be sure that our crystal is closer than 2 parts in 10 million. This would be 2 hertz of error at 10 MHz or an uncertainty of 29 Hz in the 145 MHz marker. The BC-221 is normally considered to be a .05% instrument. This would be an error of about 1.75 kHz at 3.5 MHz. However with care in calibration, and reading it is not difficult to reduce this value to 200 hertz or less. In a hetrodyne system the error of the VFO is not multiplied at VHF but is just added to the error of the crystal marker used.

The total error at 2 meters is 29 Hz contributed by the crystal and 200 Hz by the VFO for a total of 229 Hz. This is a little more than 1.5 hertz per million hertz for a tuneable system. Of course these values are approximate and with careful operation they can be reduced by 50% or more.

In the 146.25 MHz example used earlier, the VFO was required to furnish less than 1% of the total output. To put it another way, the only wobble is in the smallest cog and its contribution is so small it can't shake up the machinery too much.

The circuit shown in **Figs. 1 and 2** has been used for monitoring MARS, CAP, and several other services. The crystal oscillator is quite stable but can be tuned enough to zero with WWV. Tuning is done with C1. One stage of harmonic amplification is sufficient to give strong signals well above 150 MHz. The plate circuit of this amplifier stage is tuned to the harmonic to be used. This feeds one input grid of the mixer. The other grid of the mixer is driven by the output of the BC-221 frequency meter. The tuned circuit shown in this grid resonates broadly in the 2-4 MHz range of the BC-221. This helps to keep the higher harmonics of the BC-221 out of the mixer.

The output of the mixer is resonated to the desired operating frequency. This will be either the sum or difference of the two input signals. The level of the output signal can be controlled by R6.

Operation of this system is simple. The "cook book" would read as follows:

1. Couple the output to the antenna of the VHF receiver.
2. Determine the crystal harmonic and VFO frequency that will give the frequency of the signal to be checked.
3. Tune the VHF receiver to the signal to be checked.
4. Tune the BC-221 until the output of the frequency monitor zero beats the received signal.
5. If required, peak the tuned circuits in the monitor for maximum output and adjust R6 as needed.
6. The frequency of the received signal will be the crystal harmonic plus (or minus) the reading of the BC-221.

Many details cannot be covered in one story due to lack of space. The operator will have to determine the most effeceive method of coupling to this particular receiver. He will also have to explore the many combinations of frequencies which can be used. These and many other questions cannot be included at this time. However those who require such a system as this will probably be capable of filling these details.

One word of caution is in order. With two oscillators that are rich in harmonics, there can be many unwanted "birdies." These present no problem after the operator has gained experience but the new user should be very cautious. Many times an unwanted beat can be eliminated at a critical spot by changing the two frequencies that are being mixed (shift from sum to difference).

Perhaps we should emphasize that this system is a "trade off" where the amateur can trade his skill and patience for highly accurate measurements with simple equipment.

A TRANSISTORIZED LM METER

Charles Landahl W5SOT

The last word may never be written about the BC-221 and LM frequency meters. The LM is particularly attractive because it is in the smaller package. With transistors replacing tubes, it has features most everyone wants – it is rugged, portable, and accurate, to name a few. I will describe a conversion of an LM-15 frequency meter in which field-effect transistors replace tubes; the power supply becomes a standard 9V transistor radio battery and the current drain is less than 3 mA when all functions are energized. In addition, I offer calibration information which will be of interest to anyone having a BC-221 or LM without the official calibration book. I bought an LM-15 for a temptingly low price (Fair Radio Sales Co., Lima, Ohio, $14.95). The set is sold in the "as is" condition with tubes and crystal but without calibration book. It is a good idea, but not necessary, to start with a set which is working before making the change to FETs. Resistance measurements will show if the circuits are complete. Important values are marked on the schematic of Fig. 1.

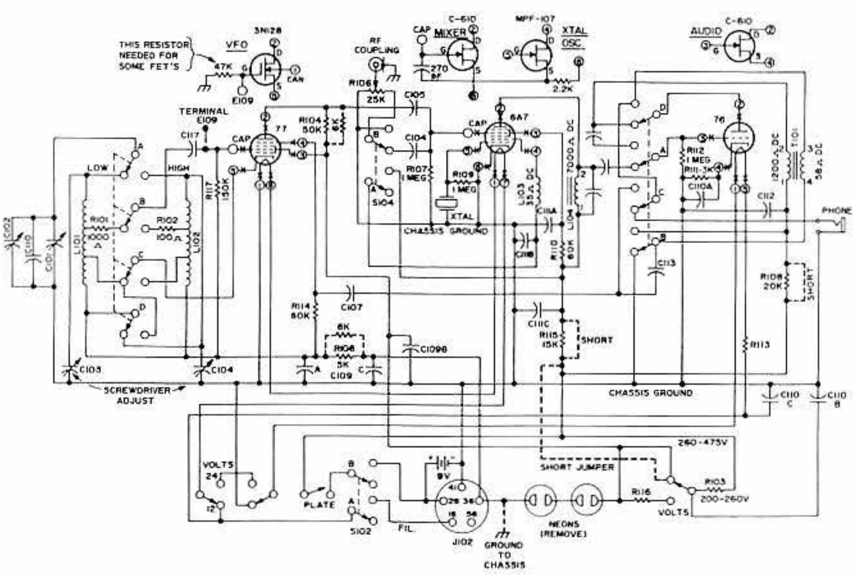
THIS RESISTOR NEEDED FOR SOME FET'S
VFO
3N128
47K
G
D
S
CAN
E109
TERMINAL E109
C117
CAP
77
R104 80K
6K
LOW
HIGH
A
B
C
D
R101 1000 Ω
R102 100 Ω
L101
L102
R117 130K
C102
C110
C101
C103
C104
SCREWDRIVER ADJUST
RF COUPLING
R106 25K
C105
S104
R107 1 MEG
C-610
CAP
MIXER
270 PF
MPF-107
XTAL OSC
2.2K
6A7
R109 1 MEG
XTAL
CHASSIS GROUND
L103 35 Ω DC
L104 7000 Ω DC
C118A
C118B
R110 60K
C-610
AUDIO
76
R112 1 MEG
R111-3K
C110A
C112
1200 Ω DC
T101
58 Ω DC
PHONES
C113
R108 20K
SHORT
R114 80K
C107
8K
R108 5K
C109
C109B
C111C
R115 15K
SHORT
R113
CHASSIS GROUND
C110 C
C110 B
260-475V
SHORT JUMPER
VOLTS 24
12
PLATE
S102
FIL.
9V
J102
GROUND TO CHASSIS
NEONS (REMOVE)
R116
R103
200-260V
VOLTS

Fig. 1. Modified schematic of frequency meter.

Smash Tubes

The most difficult part of my conversion was getting up the courage to smash the tubes! I wanted the bases for mounting transistors. Place the tubes one at a time in a paper sack, hold the top closed and with a metal object, strike the glass through the paper. The flying glass is caught and collected for disposal. Scrape and clean the mastic from the inside of the tube bases. Should you choose to mount a transistor socket in the wall of each tube socket, you can use the original wires; otherwise, unsolder the old wires and replace the needed ones with about 2 in. of sturdy new tinned wires. The appropriate tube base pin connections are shown in Fig. 2. Actually there is no preferred mounting scheme. Use whatever appeals to you.

Check for clearance between socket and walls. My conversion used transistor sockets mounted on metal plates which were bolted to the wall of the salvaged tube bases. This allowed FET substitution to determine which ones would work best in the several circuits of the LM. All FETs used are N-channel.

Modification

With cover removed and the LM in the upright position, front panel toward you, on the left side wall, look through two oblong machined slots and see mounted on a phenolic board a 50 kΩ plate resistor. Parallel it with about 6 kΩ. Turn the LM upside down, panel toward you. On the underside, two resistors must be shorted and a jumper wire made up and connected. Short R115, which is a 15 kΩ wirewound resistor, quite visible on a phenolic board at the left of the 1000 kHz crystal can. Run an insulated wire from a terminal of this shorted resistor to the 260–470V tap contact of the link switch. This wire can be about 6 in. long and conveniently passes through a wall slot behind the crystal socket. The link switch and its terminals are on a phenolic board in the compartment aft of the crystal socket. The jumper wire will cross near the grid resistor, R109, of the crystal oscillator. While there, change the 100 kΩ (R109) to 1 MΩ. Next, unfasten the screws holding the phenolic board located to the left of the power plug. Tip up the board and short across R108. This is a 20 kΩ composition resistor which is in the plate voltage line to the audio amplifier. Also, at the power plug, locate pin 36. Short it to chassis ground. On most sets pin 36 is the ground return for the vfo cathode. The circuit was closed through external connections in a power supply. You have completed the surprisingly few changes needed to make the LM work on FETs and a 9V battery.

The VFO

With FET source connected to pin 5, drain to pin 2, can to pin 1 (if needed), plug the FET into the vfo socket. Connect a solid wire between terminal E109 and the gate of the FET. (Terminal E109 held the grid cap wire for the vfo tube.) Connect a 9V battery to the power plug pins. PLUS to 26 and MINUS to terminal 41. If you have a milliammeter in the battery lead, it should read about 1.5 mA when you turn on the FIL and PLATE switches. Provided you were fortunate in the choice of FET, you should hear a clear CW signal in your receiver. Set your receiver to 2 MHz or 4 MHz. You may need to connect a wire from the rf coupling post on the front of the LM to your receiver antenna. Rotate the LM dial between 0300 and 0600 on the readout. Your vfo will be on the low end of 125–250 kHz or 2–4 MHz depending on the position of the low or high band switch. The XTAL and MOD switches should be off. The FET selected for the vfo may require a 47 kΩ resistor between gate and chassis ground. I found this to be true for the RCA 3N128, 3N142,and *one* of the two 40559A FETs. On the other hand, one RCA 40559A and one of several 2N3085 silicon N-channel FET from Poly-Paks worked beautifully without adding 47 kΩ to the gate.

Apparently junction and insulated-gate field-effect transistors have slightly different characteristics which show up in this peculiar vfo circuit. My own choice is the 3N128 with the additional resistor on the gate. You may find it necessary to "tune" the source, drain, and gate resistances in order to have your vfo working well with a particular FET. I used potentiometers across the various elements to arrive at the recommended values. My vfo works reliably from 10V down to 6V and the maximum drain current is 1.5 mA.

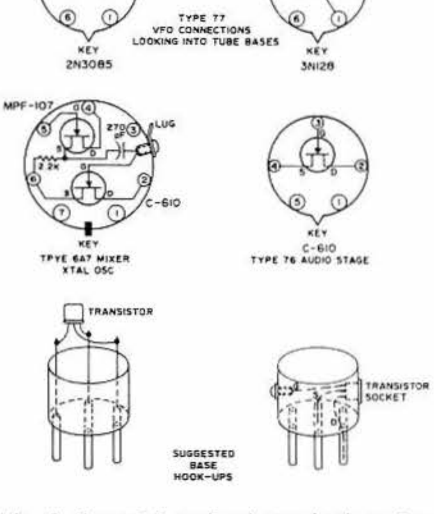

Fig. 2. Suggested semiconductor hookups for tube sockets.

Caution

Other oscillator hook-ups may occur to you, and they will work – but the tuning range and linearity of the vfo will suffer! Linear tuning is most important, so stick with that shown.

The AFO and AF Amplifier

The audio oscillator and amplifier wasn't as fussy as the vfo. I used a Radio Shack 276-664 FET – it is said to replace a C-610 or 2N3088. I found that the Poly-Paks "hobby" FET 92CU588 will work equally well. In making up the socket, gate goes to tube pin 3, source to cathode pin 4, and drain to plate pin 2. That is all there is to this one. Plug in the FET. When you next turn on the 9V power, the milliammeter will barely increase a few hundred microamps as you switch on the modulation control. At this moment a rather pleasant 500 Hz tone will appear on the vfo frequency no matter which harmonic you have tuned in on receiver. Your modulator is finished. The audio amplifier is too, for that matter. You just won't hear anything in the headphones until you complete the crystal oscillator and the mixer circuits.

Crystal Oscillator

The reference oscillator is not much trouble. You have already changed the gate resistor from 100 kΩ to 1 MΩ. Actually this change may not be necessary because some crystals are more active and will oscillate well with the original resistor. Mine went into oscillation better with the higher value. The FET you select for this circuit can be one of several. Mine is a Motorola MPF-107. I found the Radio Shack 276-112 and the Poly-Paks 2N3085 also work, but draw more current. Whichever you choose, the gate connects to base grid pin 5, source to cathode pin 6 *through* a 2.2 kΩ resistor. Drain hooks to plate pin 4.

Now, when 9V is turned on, MOD off, XTAL on, you should hear the crystal oscillator signal every 1000 kHz on your receiver. The milliammeter should increase about 1.5 mA or less when XTAL is turned on. If you don't hear the crystal frequency, bring the receiver antenna wire close to the crystal FET. We still haven't made the connection which adds the crystal-oscillator signal to the rf coupling post on the front of the LM. Assuming you have all circuits in working order up to this point, we move to the mixer.

Mixer

There is no single FET substitute for a pentagrid converter tube. The dual-gate MOSFET comes closest; however, use of one would have defeated my goal of simplest conversion. Therefore, four N-channel FETs are needed to do the work of three tubes, but what a saving in power supply! The mixer concerns itself only with beat frequencies occurring between the reference oscillator, vfo, or an external signal – all audio work. Thus, a hobby FET was selected. I used the Poly-Paks N-channel FET. A Radio Shack C-610 replacement will also work. Connect source to chassis ground through pin 6 of the tube socket; the drain connects to plate pin 2, gate to mixer grid cap wire through the lug in tube socket wall.

Finally, connect a capacitor (200–300 pF) from the top of the 2.2 kΩ crystal oscillator source resistor to the gate of the mixer FET (grid cap wire). You are in business.

With 600Ω phones plugged into the LM, you should hear all the necessary beat frequencies occurring between the vfo and the crystal oscillator as you tune the vfo through its range. XTAL must be on and the MOD switch off. Otherwise, the audio amplifier becomes the modulator and you hear nothing in the LM phones.

What Next?

With the beat notes loud and clear you are ready to calibrate. This is quite the most fun part of the work because the linear tuning rate of the LM is almost unbelievable. The slow rate is due to the series combination of the A section of C109, C101, tuning L101 or L102. The amount of the matter is that one revolution of the 100-division circular dial produces about 3 kHz change on 125–250 kHz range, and about 50 kHz per revolution on the 2–4 MHz range. The actual calibration of my unit was 2.89 kHz and 45.17 kHz per revolution.

The linearity can be checked by how little you need to vary the "corrector" for each zero-beat checkpoint. Each LM or BC-221 will be slightly different. Now, when you consider that the vernier allows you to split one division into tenths, then it is clear that you can set a frequency to better than 0.5 kHz over the range of the frequency meter. May I repeat:

1 dial revolution of
100 div = 45.17 kHz
1 division = 0.4517 kHz
1/10 div = 45 Hz

Therefore, all you need is a checkpoint at which to zero the vfo and start counting revolutions, divisions and tenths of divisions to accurately set any frequency within the two ranges of the vfo. I found it useful to construct graphs on K&E 358 11L graph paper. The grid is 10 X 10 (per 0.5 in.). The paper has 20 units vertical and 30 horizontal. This allows graphing 100 division and leaves room for 10 vernier divisions on the right hand end of the paper. Use the crystal checkpoints listed in Table I to locate *your* dial settings. Once graphed, a frequency can be selected directly from the chart, or, depending on the accuracy desired, interpolated between checkpoints.

Table I. Crystal Checkpoints

KHz	VFO	XTAL	Approximate Dial Settings
Low Band			
125	8	1	0320
150	20	3	1192
166.667	6	1	1750
200	5	1	2935
222.222	9	2	3700
250	4	1	4647
High Band			
2000	1	2	0396
2250	4	9	0945
2500	2	5	1493
2750	4	11	1050
3000	1	3	2606
3250	4	13	3150
3500	2	7	3711
3750	4	15	4262
4000	1	4	4812

It is obvious from this discussion that the low band of the LM is fabulous. You can squeeze down to about 3 Hz by use of the vernier scale. By the way, hidden behind two cover plates just beneath the corrector knob, are "high" and "low" padder capacitors. These were used when vfo tubes were replaced to bring calibration book values into usefulness. The padders should be set near the middle of their range.

Make it Handy!

Fasten a handle to the case, strap on a 9V battery, go forth and have fun with your rejuvenated frequency meter. I use mine for its intended purposes as well as a band-edge marker and keying monitor.

AN I-F SPOTTER

Howard Burgess W5WGF

If you have ever tried to find the *if* frequencies of unfamiliar and inoperative pieces of surplus gear with no schematic, it is a waste of time to tell you how rough it can be. Even a single *if* transformer from the junk box can be a problem if it has no part number or identification.

Of course in some cases a grid dip meter can be used to find the operating frequency. However, few grid dippers cover the important *if* frequencies below 2 mc. To complicate things, if a dipper is used on a shielded transformer above 2 mc, the resonant frequency of the transformer may be shifted if the shield is removed.

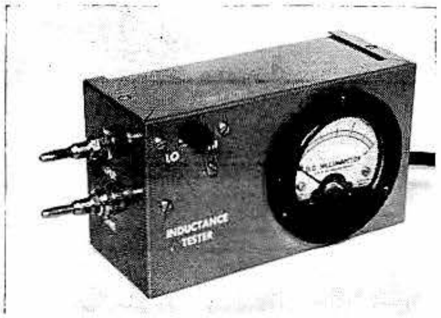

The if spotter. The switch for changing the coupling capacitor is at the upper left marked "HI" and "LO." The posts on the end are X1 and X2.

If these problems sound familiar to you, we would like to suggest a little gadget that can help solve them. With just two resistors, two capacitors, and a tube, don't expect it to give a digital read-out to all your questions, but it can put you in the ball park.

The principle of operation is as simple as the construction. The tuned circuit in question is merely made to oscillate at its resonate frequency. The frequency can then be determined by tuning in its radiated signal on the ham receiver. To set the unknown coil into oscillation requires the use of a simple "two terminal" oscillator. Such an oscillator is shown in the schematic of Fig. 1. When any tuned circuit is connected to the two points marked X, the circuit will oscillate at its resonant frequency.

In this oscillator the twin triode is a tube such as the 12AT7. The section V2 furnishes the necessary feedback and eliminates the need for extra coils or feedback connections.

The construction is simple. The unit could have been built in a larger case with its own power supply and would have become a nice piece of bench equipment. However, due to the few parts required and the small amount of plate power used (3 mils at 90 volts) it was built as an overgrown probe. The power is robbed from another piece of test equipment or the receiver.

As a probe it can be used on the work bench to test individual coils and transformers or it can be held in contact with the various transformers in a receiver.

There is only one point of caution that should be observed in construction. The lead from the grid contact of VI to the XI post should be kept as short and direct as possible with the least capacity to ground. This lead becames part of the oscillating tuned circuit and limits the upper frequency to which the unit will operate.

The coupling capacitor from the grid of V1 to the plate of V2 furnishes the feedback required to maintain oscillation. To reduce the loading on the tuned circuit, this capacitor should be held to the smallest value that will sustain oscillation. Because of the wide range over which this instrument operates, a switch is provided to change values. With the capacitors shown, operation is possible from about 60 mc down to well below 50 kc. The larger value is used at the lower frequencies and is switched in only when required. With coils of medium Q the switching point is around 3 to 5 mc.

The connectors X1 and X2 can be almost any kind of posts. The ones shown on the unit here are banana plugs. These can be used as test points, or alligator clips can be slid over them for use in clipping to coil leads.

The meter shown is a three mil meter and is used to read the total plate current of both halves of the tube. This will indicate when the circuit is oscillating. When the tester is not oscillating the meter will indicate a current of about 1 mil (with 90 plate volts). Under oscillating conditions the meter will rise to as much as 3 mils with a high Q coil. The actual amount of current is not so important as the fact that the upward shift indicates that the coil is not open and is oscillating. A 5 or 10 mil meter will serve the purpose just as well.

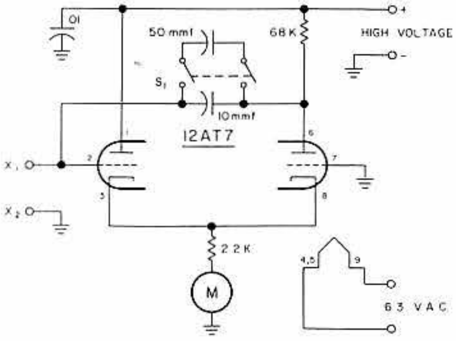

FIGURE 1

Circuit diagram of the if spotter

.01 Ceramic capacitor
50 pf Mica capacitor
10 pf Ceramic
68 K ½ watt carbon resistor
22 K ½ watt carbon resistor
Milliammeter in range of 3 to 10 ma
Slide action Double Pole-Single Throw switch (S1)
2 Banana plugs (x1 and x2)
12AT7 tube
Minibox (size depends on type of meter used)
9 pin miniature tube socket

As the pictures show, this unit was built in a small 5" x 3" x 2" box. The unit will work just as well if it is built on a small piece of peg board with a couple of leads run out to the multimeter. This is for the man in a hurry.

When the unit is finished, apply power with the "X" points open (no coil across them). Because of the open grid of V1, the meter will drift about. After the tube has had time to warm up, short the X posts with a heavy piece of bus or copper. The meter will now come to rest somewhere around 1 mil. This is the "no oscillation" current and should be kept in mind as a reference point for future use. Now remove the short and connect almost any kind of an LC circuit across the posts. The meter reading will now rise from the "no oscillation" value indicating that the coil is oscillating. Do not use the large coupling capacitor unless the circuit refuses to oscillate with the smaller value.

To check a single *if* transformer, all that is necessary is to hook one of the *tuned* coils

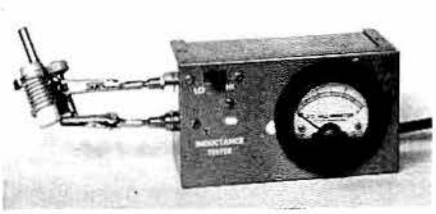

Checking a ten meter tank circuit

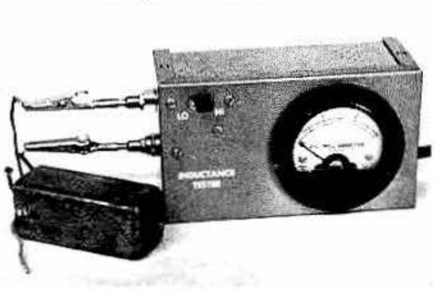

Checking an if transformer

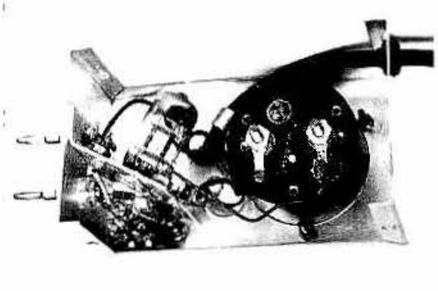

Inside view showing extreme simplicity of the "Spotter." The tube is mounted at an angle to keep the grid lead to the "unknown" as short as possible.

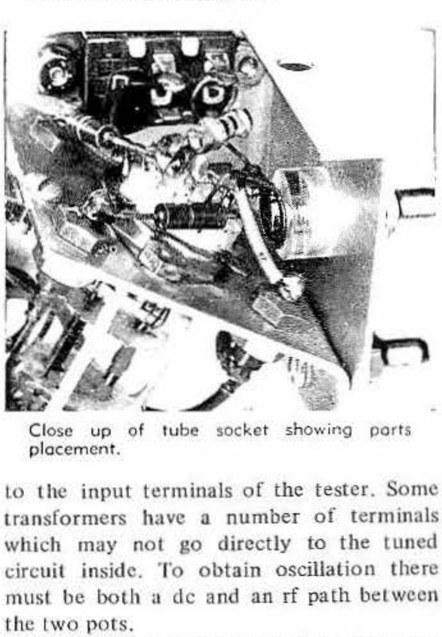

Close up of tube socket showing parts placement.

to the input terminals of the tester. Some transformers have a number of terminals which may not go directly to the tuned circuit inside. To obtain oscillation there must be both a dc and an rf path between the two pots.

If the transformer is one whose frequency falls in the range below the broadcast band it is quite convenient to have one of the surplus receivers that covers the range down to 200 kc. However the check can still be made with a regular broadcast receiver. All that is necessary is to find the harmonics of the tester as they fall in the broadcast band. They will be separated by a value equal to the frequency of the coil being tested. As an example, if a signal is spotted at 900 kc and the next one higher is found at 135 kc, it is a pretty good bet that the transformer is operating on 450 kc (1350 − 900 = 450).

To find the operating frequency of an *if* stage it is not necessary to have the amplifier in operating condition or the tubes hot. Just connect the two contacts across the primary or secondary of the transformer in question and watch for signs of oscillation on the meter. Some transformers have a portion of the bias system inside of the can. This can usually be overcome by connecting the tester from grid to ground of the tube in the stage being tested.

In addition to checking transformers it can also be useful in testing the range over which a transmitter tank will tune. Just make sure that the high voltage is turned off and connect the probe across the tank to be tested. Now you can tune the tank and follow its entire usable range with the receiver. If the tank being checked happens to be the final, the meter on the probe will indicate when the antenna is brought into resonance.

This little tester was built to do just one thing—sort out some old *ifs* in the junk box. After we tried it we found that it would do a lot of useful chores around the ham shack. With proper care and feeding, it can probably learn to do tricks that we haven't even thought of.

430-470 kHz SWEEP FREQUENCY GENERATOR

Edward Lawrence WA5SWD

Here is a simple sweep frequency generator for aligning the most common *if* strips. The unit has only one transistor because the sweep voltage is taken from the oscilloscope it is used with. Most general purpose oscilloscopes have a sawtooth output jack. By using this sawtooth, the frequency is locked to the position of the trace.

The oscillator is made to deviate in frequency in step with the voltage applied to the base bias circuit, either the sweep voltage or the dc voltage applied at the control point. If no sweep or control voltage is applied, the oscillator runs at the center frequency, and may be used as a conventional signal generator.

The 2.5 mh choke and the two 150 pf capacitors form a very broad tank circuit, so it is easy to FM the oscillator without a great change in amplitude. As to how the change in base voltage causes the frequency shift, I am not quite sure. All I know is that it is quite linear and is a positive shift (**Fig. 3**). That is, an increase in base voltage causes an increase in frequency. Also, a change in collector supply voltage will shift the frequency, so be sure to use a stable supply.

Since an rfc is used as part of the tank, you may have to compensate for a variation in center frequency by changing the 150 pf capacitors to some other value to get 455 khz as your center value.

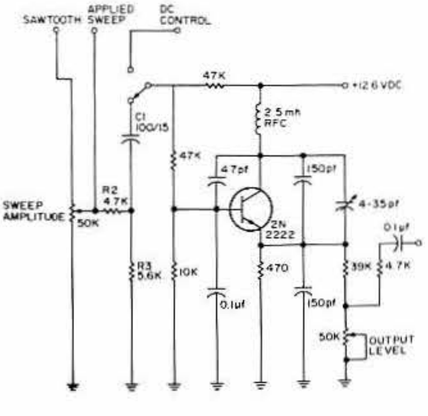

Fig.1. Sweep generator circuit diagram.

If you do, then run a plot of dc control voltage vs frequency and adjust the voltage divider R2 and R3 so that 1 volt P-P at the sweep pot wiper gives plus and minus 5 khz (10 khz total). This is not as hard as it sounds. Just listen for the harmonics on a broadcast band radio. For instance, the second harmonic of 430 khz is 860 khz, and the second harmonic of 470 khz is 940 khz, both handy on the BC radio. Note the dc voltages at the control point required to obtain these frequencies. Take the difference and divide by four. The answer is the P-P voltage required to shift the oscillator plus and minus 5 khz. Then apply a low, known ac voltage to the wiper of the sweep pot with the switch in the *sweep applied* position. Adjust the value of the divider resistors to get the proper fraction. In my case it was 0.43.

It is advisable to run the sweep rate as slowly as possible, in order to display the response curve as accurately as possible. The sharper the skirts, the more slowly the generator must pass through the bandpass. With this generator, the amount of frequency deviation is controlled by the P-P amplitude at the sweep pot arm. If you are looking at the response of a regular *if,* you would set the sweep amplitude high to see the entire response curve. As far as an accurate display is concerned, this is fine, since the slope of the skirts is shallow. But if you were looking at the response of a sharp filter, you could not tolerate such a wide sweep, because the fast rate-of-change would tend to skew the display. To correct this, reduce the sweep amplitude to reduce the frequency deviation down to the edges of the skirts of the response curve. This reduces the rate of change and minimizes the skewing of the display. Also, it is better to display the *if* before detection, if possible, to prevent the detector time constant from possibly distorting the display.

If you have the sweep generator set up as described, then it is easy to set the sweep for a known deviation, and then procede to read the 3 or 6 db points from the face of the scope. Be sure to disable the AGC for this test.

For those who aren't familiar with the setup for obtaining the response curve display, refer to **Fig. 2.** and the following outline.

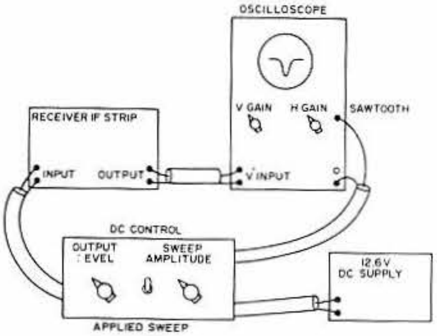

Fig.2. Interconnections of sweep generator.

First, hook up the equipment as in Fig. 1 Set *horz gain* for desired sweep width. Adjust *sweep amplitude* pot on generator for desired frequency range. Set *output level* to mid-range. Adjust the *vert gain* for the desired pattern height.

If we wish to change the total frequency deviation and the horizontal display width at the same time, use the scope's *horz gain* control. This presumes that the amplitude of the sawtooth output is also varied by the *horz gain* control.

If we wish to change the frequency spread and not change the width of the scope display, adjust the *sweep amplitude* control. This allows you to take a better look at the sidelobes or any ripples in the passband, depending on how the controls are adjusted.

You don't have to use a sawtooth to sweep with if it isn't handy.

Sixty hz can also be used, but it will probably skew the passband you are trying to display, so I don't recommend it. One note here: when the sweep generator is first turned on, or the setting of the sweep amplitude is changed, the frequency will drift for a few seconds. This is due to the charge on C1 changing to a new level. C1 is large to couple the low frequency sweep with as little distortion as possible.

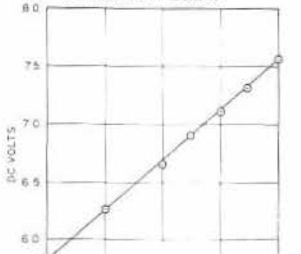

Fig.3. Sweep frequency vs dc voltage.

DC CONTROL VOLTS	FREQUENCY
5.86	430
6.27	440
6.69	450
6.90	455
7.12	460
7.33	465
7.55	470

AN ARMSTRONG SWEEPER

Al Donkin W2EMF

Sometime ago while experimenting with SSB crystal filters at 480 kc, it occurred to me that a manually swept Armstrong oscillator with a pot coupled to the tuning capacitor for 'scope sweep voltage would save point by point frequency measurements with a BC221 while adjusting the filter for minimum in-band ripple. A circuit was hurriedly developed, and the results were even better than expected. Simple though it may be, I'm sure anyone who is working with crystal filters will appreciate its usefulness. The effects of tuning and "diddling" were immediately visible in the in-band ripple and the skirts down to about 20 db could be observed with a few twists of the knob.

With this encouraging experience, a sturdier model was constructed. The circuit was assembled on a small piece of Vector-board, with the tuning capacitor and batteries mounted in a 5 × 7 × 3" aluminum chassis. An aluminum bracket was bent to hold the "sweep" pot and mounted behind the tuning capacitor, spaced to accomodate the mechanical shaft coupling. The Vectorboard was also mounted on the bracket.

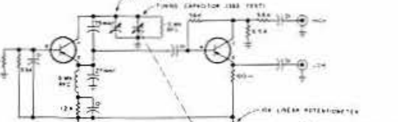

The oscillator circuit is simple, using a 1 mh rfc for the tank inductor. The range of the circuit as shown is about 400 kc to 500 kc, at center frequency. The tuning capacitor is a Hammerlund MC-50-M with two of the rotor plates removed, resulting in a range of approximately 7-25 mmfd. An untuned amplifier provides isolation of the oscillator from the reactance of the load. I used 2N404 transistors but most PNP *if* transistors will work.

In addition to sweeping crystal filters, I have found this little box handy as a general purpose low frequency oscillator and sweeper for receiver *if*'s. The persistency of most oscilloscopes produces a much better display than might be expected, and it certainly beats the point by point method of plotting filter responses.

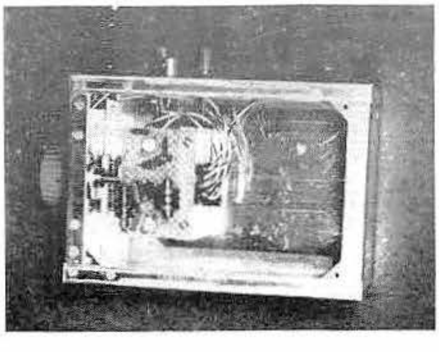

A NEW BROOM

Dick Gridley K6JHJ

This unit is a sweep generator of simple design, with many uses. Besides being useful for filter alignment, as shown in the diagram, it is useful for optimizing receiver and transmitter performance.

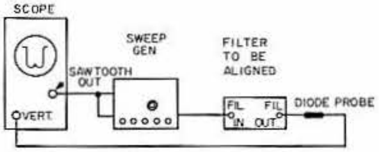

The heart of the rig is an inexpensive item called the Vari-Cap by International Rectifier Corp. Two are discussed here, the 6.8SC20 and the 100SC2. The 6.8SC20, according to the manufacturer, will vary its capacitance from 50 mmfd at .1v to 2.5 mmfd at -200v. The 100SC2 will vary from 600 mmfd at .1v to 100 mmfd at -10v. These units each cost less than $3.00. The construction is not critical in any way, nor is layout a problem—just use good ham practices. Any power supply delivering 150 v dc at 60 ma regulated will do. The oscillator tank must be grounded and I did have some trouble until I removed the unused winding from the 455 *if* transformer used for the oscillator tank.

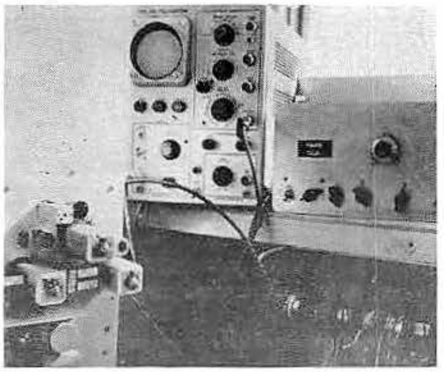

Sawtooth energy is applied to R1 which regulates the amount to be amplified by V1. R2 linear control merely puts a portion of regulated voltage on the Vari-Cap and thus sets the capacity of Vari-Cap close to its center, so the sawtooth pulse creates a linear frequency swing. This circuit gives about 6 kc sweep. I have used TV horizontal coils for L1 and increased C1 and C2 to .003 for 85 kc and 50 kc operation. However, to get enough swing at low frequency, the Vari-Cap was changed to a 100 SC2. For 200 kc to 10 mc the 6.8SC20 gives plenty of sweep.

A note here about scopes. Most of the average do not go below 15 cycles and the sweep rate should be about 10 cycles. Some do not have a sawtooth output jack. The sawtooth output can be picked off the sync circuit. I found the best point is on the feed side of the sync pot through a .25 mfd capacitor. Some scopes have a couple of jack pins to add capacity externally, some don't. You will find little trouble adding about .25 mfd across the slowest speed.

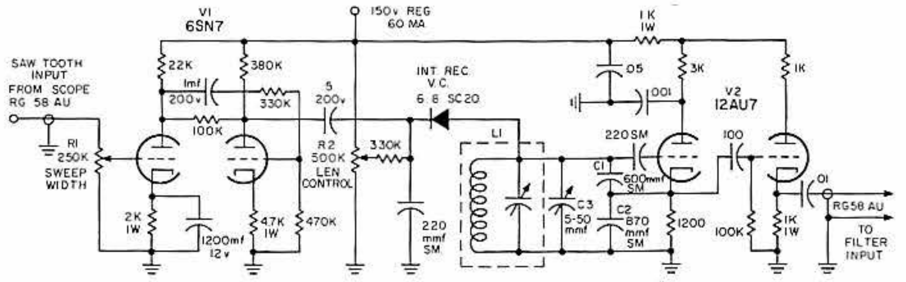

L1 STANDARD I F TRANSFORMER WITH ONE WINDING REMOVED
SM- INDICATES SILVER MICA.

The use of the rig is simple. Plug the lead from R1 into the scope's sawtooth output. Couple the output from the cathode follower section of V2 to the filter input through a .001 mfd capacitor. The vertical input of the scope is connected to the filter output through an rf probe. Adjust R1 and R2 about ½ open. With C3 about ½ meshed, adjust the trimmer in the *if* can to put you in the ball park, and then use C3 for fine adjustments of frequency. Additional minor adjustments of R1 and R2 may be necessary. Of course, when used to align a receiver, the diode probe is unnecessary if the input to the scope is picked off of the diode detector. It goes without saying that it is necessary to plot an index for the scope bezel so that a ready reference is available for band pass width and depth of skirt.

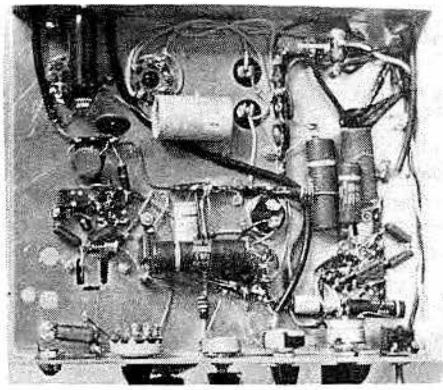

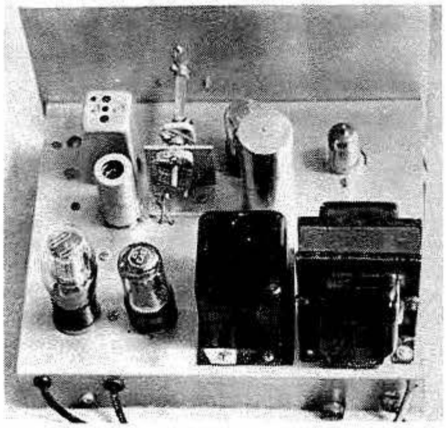

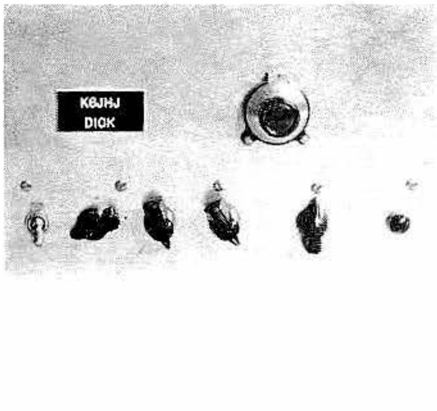

Some one now says "This is just like setting ye old TV—*if* band pass." That it is. The only difference here is the frequency and the sweep generator cost was junk and $5.00.

The unused switch, shown in the photo, was installed to select one of three frequencies. However, this rig has been used on about six. I don't think I ever will decide which three to use. The unidentified tube is a 6U8 used as a 1 kc audio oscillator with cathode follower output.

LOW COST SIGNAL SOURCE

Howard White VE3GFW

For many years the Measurements Corporation Model 80 signalgenerator has been the industry standard for tuning rigs. Many hams have wanted to have a laboratory-grade signal source of this type but the $800 price tag is prohibitive. This article describes a signal source that has many of the same features as the 80 except that it only costs 1% as much. That's why I call this the Model 0.8.

You wonder at some of the features of the marvelous little device? To briefly list them, it has:

- Variable output from about 80 nV to 50 mV of rf power.
- Frequency range from 1.8 to 450 MHz so you can cover 160 through ¾ meters.
- Crystal-control frequency stability.
- Fairly clean output signal.
- A 51Ω antenna load.
- Safety feature to prevent the destruction of the device in case a transmitter is accidentally loaded into it.

The Circuit

The circuit is the combined brainchild of many ham engineers and technicians in our local 2m FM club – the Toronto FM Communications Association. More than 300 of these units have been built in the past two years, so the circuit has been exceptionally well tested in the field.

The unit consists of a single 2N5130 NPN transistor oscillator configuration. Other transistors, such as the 2N706, can be used in this circuit with slightly different biasing arrangements; however, the 2N5130 appeared to give the cleanest output signal.

Output level can be varied from less than 80 nV (the best we could measure) to 50 mV. It is controlled by a 1 kΩ pot in the transistor collector circuit. Power is obtained from a 1.5V penlight cell bypassed for rf by a 0.1 μF capacitor.

My present "0.8" signal source has been used intermittently for more than a year without any degradation of output. However, remember to shut the unit off when you're finished with it; the battery doesn't last forever.

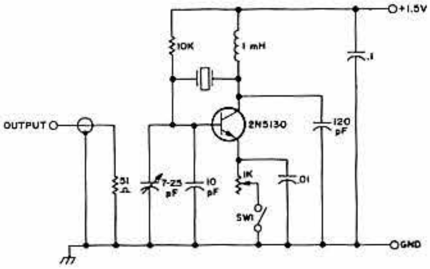

Fig. 1. Schematic diagram of the "Point Eight" Signal Source, the "poor man's Model 80."

The unit is built on a PC board. The parts layout is almost exactly the same as that of the schematic shown in Fig. 1. The electronic components are placed on the copper side of the board. The on-off switch-level control, frequency adjustment, output jack, and crystal socket are on the other side (Fig. 2). The unit can be packaged in a minibox.

Almost any crystal in the 1.8–12 MHz will oscillate in this unit. The output is rich in harmonics to 450 MHz, so you can tune up any receiver from 160 to ¾ meters.

The output impedance of the signal source is 51Ω, a value which simulates a perfect antenna load to the receiver. This is the manner in which this is accomplished: A 51Ω, ½W resistor is connected across the terminals of an RCA phono jack. No physical connection is made to the circuit except through the common

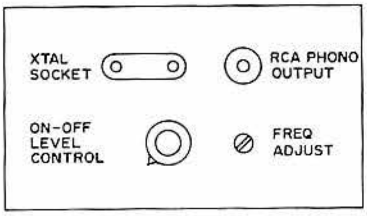

Fig. 2. Minibox layout of the signal source panel.

ground of the PC board. The maximum output level depends on the lead lengths of the resistor. (The unit has plenty of output when the lengths are about ½ in. long.) This design has an added safety feature, too; if by accident you load a transmitter into the unit, all that is destroyed is the ½W resistor.

Operation and Uses

These units were originally designed to tune up 2m FM receivers. The procedure is quite simple. You connect the signal source to the antenna input on your receiver. Plug in a transmit crystal, and adjust the frequency control for a zero reading on the discriminator. Adjust the output level to the desired signal strength (below first limiter saturation) and tune up the receiver.

There are a myriad of other uses for the signal source. Using a 3.5 MHz crystal, you have a band edge marker. With the transmit crystals on a Twoer or any other transmitter you have instant frequency spotting without modifying the circuit of the transmitter. Of course, the signal source can be used to tune up any receiver, peak tuned circuits, be an rf source for an antenna noise bridge, and so on.

RECEIVER TWEAKER

Malcolm Oakes K6UAW

I've found that the singlemost piece of needed test equipment by the amateur on FM is a receiver alignment generator. Most of us, however, do not have access to a signal generator. (Come on now, you wouldn't really call that *TV thing* a signal generator would you?)

The unit to be described is a very functional device that will allow you to scrape every ounce of sensitivity from your receiver. The generator can be built so small you can carry it around in your shirt pocket. Compactness, combined with its battery-powered portability make it ideal for servicing mobile receivers. Stability? It's crystal controlled and is as good as the rock you plug in (you obtain the rock from your transmitter). Output reactance? Nearly zero degrees, allows proper tuning of rf stage. The attenuator shown does not have a great deal of dynamic range due to distributed capacity of the potentiometer. However, I said before, the unit was "functional" and it is just that. The signal can be attenuated into the noise and brought up to approximately a 30 uv level; overly sufficient for a normal alignment.

Construction

The circuit layout is not particularly critical, and if laid out in a manner similar to the schematic, no problems should be encountered. The attenuator section should be completely shielded away from the oscillator and multiplier stages, so rf leakage will not be a problem. A crystal socket should be provided so the unit is versatile for any frequency. However, if you plan on using it on only one channel and can spare a crystal, build it with the crystal inside. Two glass piston screwdriver adjusted trimmers must be provided (to "rubber the crystal" and tune the output to resonance) as front panel controls.

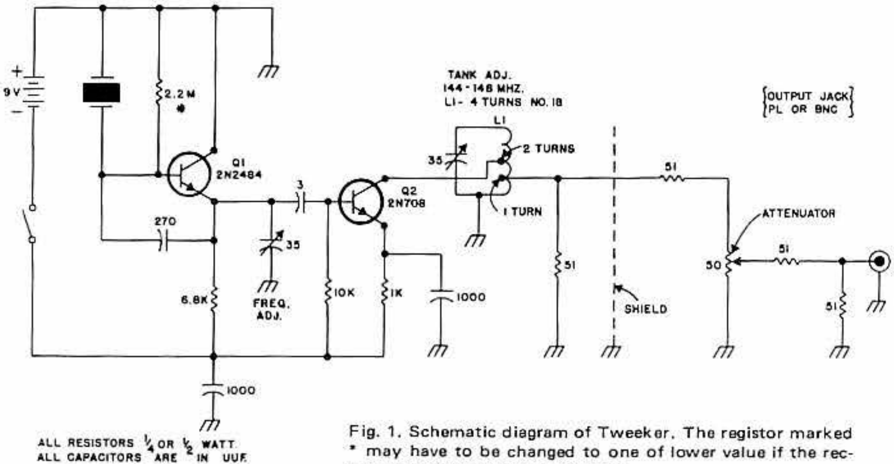

Fig. 1. Schematic diagram of Tweeker. The registor marked * may have to be changed to one of lower value if the recommended transistor is not used.

Circuit

Transistor Q1 in the first stage is a crystal oscillator which is very loosely coupled to Q2 the multiplier. This stage is biased into "class C" so as to multiply the 6 mhz crystal 24 times, up to the two-meter band. (3 mhz crystals will also work at a multiplication of 48.) Another important function of this stage is to attenuate the oscillator output (about 6-7 volts peak-to-peak) to a level usable at the two-meter frequency for alignment – 30 to 40 uv. The potentiometer adjusted attenuator takes the signal down to a level as desired by the operator.

I developed this unit for use on two meters. Several have been built and all work fine. But, for those of you who need a six-meter generator, this same circuit should work fine with the only modification needed being the final tank frequency. (A few more turns on the coil and a slightly larger trimmer capacitor.)

Adjustment and operation

Connect the output of the generator to your rig. Plug in a crystal and tune for zero (with frequency adjustment) on the discriminator (center frequency). Now, looking at the first limiter voltage, peak the final tank. Attenuate the generator as necessary to avoid saturation of the limiter. Repeat as necessary until a definite peak is reached.

If the generator will not go down into the noise with the attenuator control, the final tank may be detuned as necessary to provide the desired range on the attenuator.

CITIZENS BAND ALIGNMENT AID

Edward Lawrence WA5SWD

It is obvious that the front end of a receiver should be tuned to respond equally well across the band, and thus have consistent sensitivity at any frequency in the band, if maximum usefulness is to be obtained. The usual method of insuring reasonable results is to tune at the middle, and check the ends. This is rather spotty, at best, as it would be better to check all channels. This is involved since you must set both frequency and amplitude 23 times in a single pass.

Naturally, a good sweep generator could be used, and the rf-mixer tracking set with it instead, but this was not available to me at the time. The circuit in Fig. 1 was originally a super-regenerative receiver taken from the GE Transistor Manual. These, as is well known, tend to spray a lot of rf back out the antenna. With this in mind, I listened with a communications receiver and noted strong modulation products about 100 kHz above and below the "carrier." Deducing this to be caused by the quench frequency, I brilliantly decided to drop the quench frequency into the audio range, by increasing the capacitance across the base bias resistor. Lo and behold! Suddenly the entire Citizens Band was alive and jumping with a mass of "line noise," which at first appeared to be like

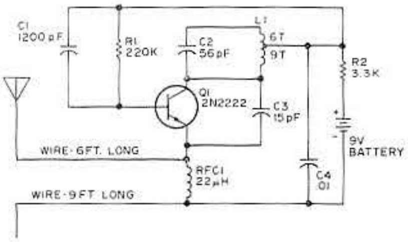

Fig. 1. Schematic of the alignment aid. L1 is 15 turns No. 22 tapped at 9 turns from the collector end. No switch is shown, so remove battery when not in use.

73 LIBRARY

NEW REPEATER ATLAS

Hundreds of new listings . . . listed by both location and frequency . . . dual listing, invaluable for your car . . . find those repeaters as you travel. This is the ONLY complete list of repeaters being published. Almost 3000 repeaters listed in this issue . . . repeaters from all over the entire world. only $1.95

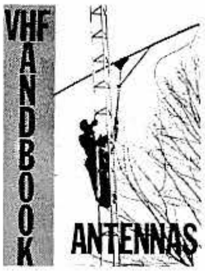

VHF ANTENNA HANDBOOK

The NEW VHF Antenna Handbook details the theory, design and construction of hundreds of different VHF and UHF antennas . . .

A practical book written for the average amateur who takes joy in building, not full of complex formulas for the design engineer. Packed with fabulous antenna projects you can build. $4.95

73 VERTICAL, BEAM AND TRIANGLE ANTENNAS by Edward M. Noll W3FQJ

Describes 73 different antennas for amateurs. Each design is the result of the author's own experiments: each has actually been built and air-tested. Includes appendices covering the construction of noise bridges and antenna line tuners, as well as methods for measuring resonant frequency, velocity factor, and standing-wave ratios. 160 pages. $5.50

73 DIPOLE AND LONG-WIRE ANTENNAS by Edward M. Noll W3FQJ

This is the first collection of virtually every type of wire antenna used by amateurs. Includes dimensions, configurations, and detailed construction data for 73 different antenna types. Appendices describe the construction of noise bridges, line tuners, and data on measuring resonant frequency, velocity factor, and swr. 160 pages. $5.50

2M FM HANDBOOK

Contains almost every conceivable circuit that might be needed for use with a repeater. All circuits explained in detail. All aspects covered, from the operator to the antenna. $5.95

Practical Test Instruments You Can Build

37 simple test instruments you can make — covers VOMs, VTVMs, semiconductor testing units, dip meters, wattmeters, and just about anything else you might need around the test lab and ham shack. $4.95

IC OP-AMP COOKBOOK

by Walter G. Jung. Covers not only the basic theory of the IC op amp is great detail, but also includes over 250 practical circuit applications, liberally illustrated. 592 pages, 5½ x 8½, soft bound. $12.95

RF AND DIGITAL TEST EQUIPMENT YOU CAN BUILD

RF burst, function, square wave generators, variable length pulse generators — 100 kHz marker, i-f and rf sweep generators, audio osc, af/rf signal injector, 146 MHz synthesizer, digital readouts for counters, several counters, prescaler, microwavemeter, etc. 252 pages. $5.95.

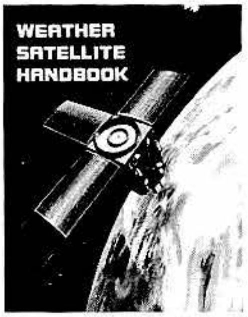

WEATHER SATELLITE HANDBOOK

Simple equipment and methods for getting good pictures from the weather satellite. Antennas, receivers, monitors, facsimile you can build, tracking, automatic control (you don't even have to be home). Dr. Taggart WB8DQT $4.95.

SSTV HANDBOOK

This excellent book tells all about it, from its history and basics to the present state of the art techniques. Contains chapters on circuits, monitors, cameras, color SSTV, test equipment and much more.

Hardbound $7 Softbound $5

FASCINATING WORLD OF RADIO COMMUNICATION

Interesting stories of the history of radio pioneering and discovery. Also includes the fundamentals of broadcast band DXing. A must for every radio amateur. $3.95.

TEST EQUIPMENT LIBRARY

A multi-volume anthology that you won't find anywhere for three times the price! The editors of *73* present a complete guide to testing that is both basic and complete and aimed at everyone from beginners to amateur Extras. Everything from how to test diodes to frequency counters. Plus a cumulative index . . . all in *73*'s all new TEST EQUIPMENT LIBRARY Vol. I — Vol. III.

Vol. I — Component Testers — $4.95
Vol. II — Audio Frequency Testers — $4.95
Vol. III — Radio Frequency Testers — $4.95

1001 PRACTICAL ELECTRONIC CIRCUITS

Tab's new 1001 circuits is available for only $9.95 ppd. The next time you want a circuit for just about anything, eat your heart out that you didn't send for this book the first time you read about it. You'd better order the book right away, before they run out. $9.95

VHF PROJECTS FOR AMATEUR AND EXPERIMENTER

A must for the VHF op. Opening chapters on operating practices and getting started in VHF, both AM and FM, followed by 58 chapters on building useful test equipment, modifying existing and surplus gear. $4.95

73 MAGAZINE • PETERBOROUGH NH 03458

Hewlett Packard 6384A power supply

SPECIFICATIONS:

DC Output: 4 to 5.5 Vdc @ 8 Amps
Load Regulation: 2 mV no load to full load
Line Regulation: 2 mV for line changes between 104 and 127 Vac at any output voltage and current within specifications
Ripple & Noise: Less than 5 mV peak to peak
Crowbar, Factory Set at 6.25 V, Automatic cutback limit circuit and overvoltage crowbar protection.

HEWLETT PACKARD 6384A Power Supply $99.95

Power design 1210 regulated power supply

Output: 0 to 12 V @ 0 to 10 Amps
Regulation: 0.01% or better.
Ripple and Noise: Less than 500 uV rms. Voltage Controlled, Metered.

POWER SUPPLY 1210 - Regulated Power Supply - $149.00

a full range, 5-function 3½ digit multimeter

RANGES
DC VOLTS: 0 to ±1.999/19.99/199.9/1500 Volts
AC VOLTS: 0 to 1.999/19.99/199.9/1000 Volts
DC CURRENT: 0 to ±1.999/19.99/199.9/1999 mA
AC CURRENT: 0 to 1.999/19.99/199.9/1999 mA
RESISTANCE: 0 to 199.9 Ω through 19.99 Megohms (6 ranges!)

ACCURACIES
DC VOLTS: ±0.2% F.S. ±0.2% of reading
(1000V range: ±0.5% F.S. ±0.5% of reading).
AC VOLTS: ±0.7% F.S. ±1% of reading
(1000V range: ±0.7% F.S. ±2% of reading).
DC CURRENT: ±0.5% F.S. ±0.8% of reading
(1000 mA range: ±0.5% F.S. ±1.8% of reading).
AC CURRENT: ±0.5% F.S. ±1.3% of reading
(1000 mA range: ±0.5% F.S. ±2.3% of reading).
RESISTANCE: ±0.5% F.S. ±0.8% of reading
(10 megohm range: ±0.5% F.S. ±1.8% of reading).

DATA PRECISION MODEL 134 – $169.00

INSTRUMENTS & COMPONENTS

AST/SERVO SYSTEMS, INC.
20 Republic Road, North Billerica MA 01862
(617) 667-8541

NEW 73 BOOKS

TEST EQUIPMENT LIBRARY
Vol. III, Radio Frequency Testers

Radio frequency waves are the common denominator of Amateur Radio so here is a book for all hams. No matter what your specific interest, such items as SWR, antenna impedance, line impedance, RF output and field strength should interest you. This book not only gives detailed instructions on testing these items but includes sections on signal generators, crystal calibrators, grid dip oscillators, noise generators, dummy loads and much more. It's a must for all up-to-date shacks.
Test Equipment Library
Vol. III, Radio Frequency Testers $4.95

TEST EQUIPMENT LIBRARY
VOL. II, Audio Frequency Testers

HOW'S YOUR SPEECH POWER? YOUR SHIFT? YOUR SYNC?

You can find out easily with a little time and a junk box full of parts. It's all right there in the new Volume II of the 73 Test Equipment Library . . . Audio Frequency Testers . . . jam packed full of all kinds of audio frequency test equipment. So if you're into SSB, RTTY, SSTV, etc., this book is a must for you . . . good book for hi-fi addicts and experimenters too!
Test Equipment Library
Vol. II, Audio Frequency Testers $4.95

BUILD YOUR OWN TEST EQUIPMENT!

SAVE A BUNDLE!

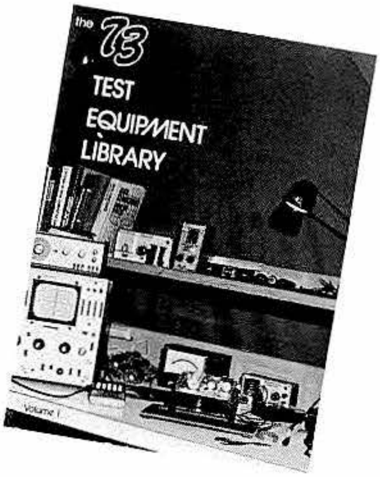

TEST EQUIPMENT LIBRARY
VOL. I, Component Testers

Just out is Volume I of the 73 Test Equipment Library . . . how to build transistor testers (eight of 'em), diode testers (3), IC testers (3), voltmeters and VTVMs (9), ohmmeters (8 different kinds), inductance (3), capacity (9), Q measurement, crystal checking (6), temperature (2), aural meters for the blind (3) and all sorts of miscellaneous data on meters . . . using them, making them more versatile, making standards, things like that. Invaluable book, ridiculously low priced.
Test Equipment Library, Vol. I, Component Testers $4.95

Send me:
☐ Vol. I . . . . .Component Testers @ $4.95
☐ Vol. II . . . . . Audio Frequency Testers @ $4.95
☐ Vol. III . . . . . Radio Frequency Testers @ $4.95

☐ Cash enclosed ☐ Check enclosed

☐ BankAmericard ☐ Master Charge ☐ American Express

Card # ____________________ Interbank # __________ Expiration date ________
Name ____________________ Signature ____________________
Address ____
City ____________________ State __________ Zip __________

73 magazine • Peterborough NH 03458

Meshna

New England's #1

Surplus Dealer For Over 20 Years.

Send for our latest catalog.

Meshna PO Box 62 E. Lynn, Mass. 01904

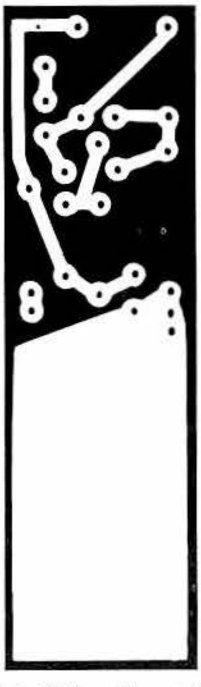

Fig. 2. Full size PC board layout (foil side).

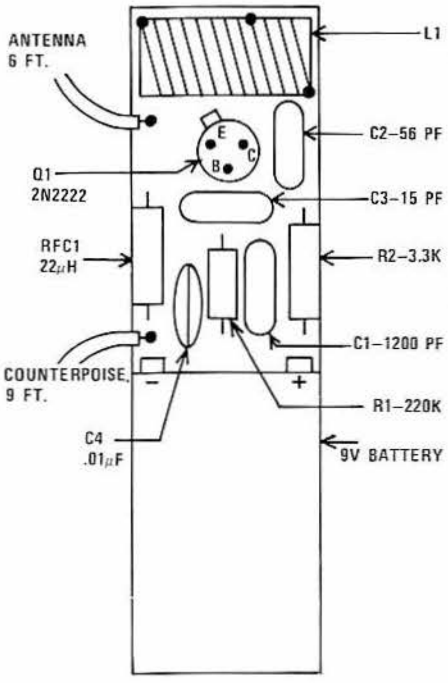

Fig. 3. Component side parts placement.

natural noise. Then I turned on the BFO, which showed it to be many, many, many closely spaced "carriers." I presume the lower quench frequency develops a much larger bias swing on the base of the transistor, FM'ing it across the entire band. This stuff is concentrated in the region of 27 MHz, with very little spill-over into 10 or 15 meters, although the level across the design range is rather constant. The nearly constant rf level is what I needed to simplify tuneup of the front ends of the units I was working on.

Figure 2 is the layout of the P.C. board I made to facilitate reproduction of this useful circuit for a few interested friends and Fig. 3 is the component layout.

I used 6 ft of wire as an antenna connected to the emitter side of the rfc, with 9 ft of wire on the battery side of the rfc as a counterpoise. Since this made it a bit large, I carefully hung it vertically in a handy tree nearby. Of course, since this put a strong signal on every channel into every CB receiver within 300 yards, I limited the use of this device to avoid unnecessary interference to the deserving users of the band. However, one fellow accidentally launched his into a tree too high to recover with an over-ambitious heave, and it ran for about a month in the dead of winter in a nameless northern city, helping every CBer for miles around align their sets! It only draws about half a milliamp from a 9 volt battery.

Since I am sure many others would like to align their 11 meter receivers for optimum results, I am presenting this circuit as a very low cost and worthwhile aid.

FET CHIRPER

Chuck Hines K6QKL

The Chirper is an automatically keyed, crystal controlled, signal source which may be used to optimize the signal-to-noise ratio of a receiving converter. Homebrew or commercial, converters are a common thing around an amateur station. And, most of the VHF Tribe have read thru a jungle of esoterica dealing with low noise front ends, the velvet beauty of FET's on Two, noise generators and eternal truth, and how to copy 20 db below the noise by the selective use of liquid helium. With a kind of relentless evolution converters have been getting better and better, noise figures become lower, and the prices of suitable front end devices are dropping by the hour. But when it comes to aligning these converters the scene is one of wretchedness. A black art at best, the job is taken up with an enduring combination of blunt instrument and myth. The latter have a certain charm. Are you convinced your converter is in top notch condition because you can "hear noise" when you attach the antenna – or better yet, when you place a 50 ohm resistor across the input? Try putting a complete short across that same input. Shorts aren't much good as noise sources. You'll find the short gives about the same change in noise level as the 50 ohm termination. What has changed is the impedance the front end "sees". The same is partially true of the noise from the antenna. Neither is indicative of the performance of the converter. Peaking the system up for maximum on either a weak signal or on noise gets you nowhere. The diode noise generator which every VHF book of substance describes is a good and useful tool when used properly. The assumption is that everyone already knows full well how to use it and does so. Few in fact do.

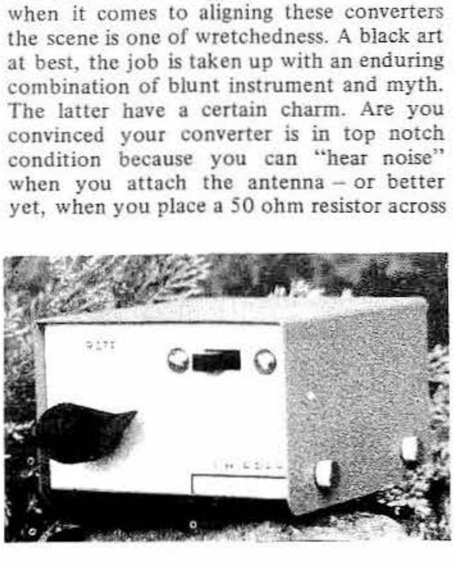

I'm sure you've read of it before in many places, but a little redundancy is in order. The noise with which you are concerned is the noise generated *internally* by the first tube, transistor, or other active device the signal encounters upon its arrival at your converter. By fiddling with the external reactances, adjusting the voltage and current and otherwise manipulating the things soldered to the device, one may minimize the internally generated noise. At the same time the reason the front end exists is to amplify the signal. One usually desires as much amplification possible, short of smoke and oscillation. Minimum noise and maximum amplification is the game. Though the two are not quite mutually exclusive a certain amount of compromise takes place. Thus, the signal to noise ratio. When aligning a converter's first stage every adjustment effects *both* signal and noise. Given a constant signal source coupled into the converter thru an appropriate impedance, the job is finished when the front end has been adjusted for the greatest *difference* between signal and noise of which it is capable.

The Chirper is designed to help you do all this by letting you see what effect each adjustment has on both signal and noise. The TIS34 oscillates at a frequency controlled by the crystal. With the constants shown, that can be anywhere between 8.2 and 36 MHz. The variable capacitor must be adjusted for resonance. It isn't particularily critical but its setting peaks the *rf* output at either the fundamental or some harmonic. For 6 meters an 8.35 MHz crystal is used. A 9.0 MHz rock will pin the meter when the

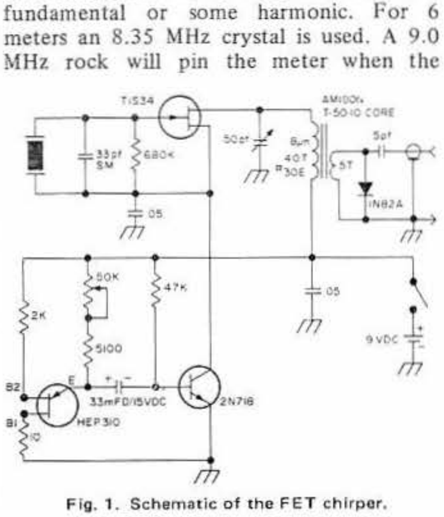

Fig. 1. Schematic of the FET chirper.

Chirper is connected into a 2 meter converter. The Amidon toroid is wound with No. 30 enameled, 40 turns for the primary and 5 turns for the secondary. After it is wound, spread the turns to fill the toroid and paint it with Q dope. The diode and 5 pF coupling capacitor are connected with the shortest possible leads, the diode being grounded at the *rf* connector. The harmonic output is excellent and quite useable at 1200 Mhz.

The oscillator is turned on and off by a multivibrator combination of unijunction and NPN transistor adapted from the *G. E. Transistor Manual.* The rate at which the multivibrator cycles is determined by the large value capacitor, in this case 33 mF. The polarity of the capacitor is critical. Observe it. To increase the cycling rate, decrease the capacitance; and, to decrease the rate, put in a larger value. Mine cycles a little under once per second. A value somewhere between 30 and 40 mF should suit your needs. You are better off scrounging some odd value from a defunct computer board because of the tolerance problem. If it says 33.2 mF, its probably pretty close to that value. Otherwise you're dealing with tolerances of plus 100% and minus 50% or something equally grotesque. The 50K pot determines the portion of the cycle during which the oscillator is On and is mislabeled rate on the Chirper shown. The HEP-310 is generally available and inexpensive. Other unijunctions were not tried. On the other hand almost any NPN of reasonable quality will work in place of the 2N718. A number of 2N388 and 2N3478's were tried and behaved well. It's a good place to use those transistors you've replaced with FET's. Use something with a Beta of 50 or better for best results. The 5100 ohm resistor in series with the pot is for current limiting. It's deletion will increase battery drain with no increase in Chirper performance. Normal current from the 9 volt battery is around 5 mA.

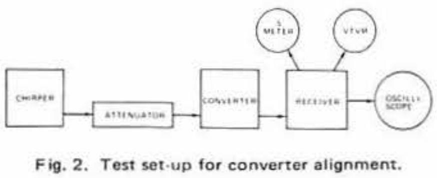

Fig. 2. Test set-up for converter alignment.

Construction is non-critical and pretty much a matter of taste. Mine is built on a piece of vector-board and mounted in a Suzurando box, model M-1N. It measures 3¼ X 2 X 4" and there is still room inside for additions. It sells for 330 yen, about 92c. A slide switch is used to turn the power on and off. Paint one well of the slide switch with red paint – Testor's Pla, a model plastic paint, is good – and the switch will indicate its position. Red for On and black for Off. It saves batteries. Check your work and the polarity of the large capacitor. Re-check the connections to all the semi-conductors. With four different kinds of devices things can become confused. Set the pot to the middle

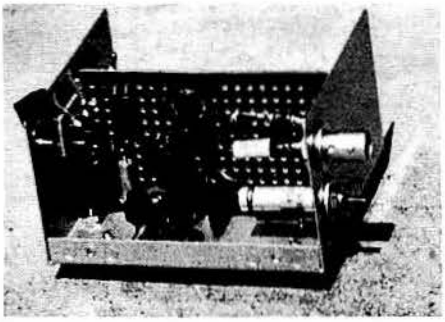

of its range. Insert a crystal in the socket. Connect the Chirper to your converter, turn the switch on and adjust the variable capacitor for the highest reading on your S-meter. The oscillator will turn on and off. Varying the pot will extend or diminish the amount of time the oscillator is on. Whatever you do, *don't connect the chirper to an external antenna.* The harmonic content is high and even at this power level is sufficient to cause severe interference to television receivers within a two block radius.

For converter alignment, the test set-up is illustrated in Fig. 2. The Chirper is fed to the converter thru an attenuator for two reasons. First, the power output of the Chirper is too high on six and two. You don't want to align with a forty over nine signal. Something around S-5 to S-7 is desired. Second, the attenuator maintains a 50 ohm termination for the converter. A converter cannot be aligned with a floating input impedance. Fixed and variable attenuators of excellent quality are available thru surplus and homebrew data is available. See *73,* January 67, p. 40 for one that will do the job. Turn the receiver avc off. The read-out options are diverse. The best is probably a scope connected to the *if.* A vtvm can be used, connected to the audio output. And, the S-meter can be used with the avc on *fast.* This will vary with the receiver and it's particular time constants. What needs to be avoided is avc pumping that interferes with your readings.

Turn the Chirper on, adjust the attenuator for a convenient signal level. When the oscillator is on, you're reading signal. When the oscillator is off, you're reading noise (on the scope, vtvm, S-meter, etc.). As you make adjustments on your converter, observe the effect on both signal and noise. Adjust for the greatest difference between the two. Turn the Chirper off and re-check the converter neutralization. If necessary, re-neutralize the converter and go thru the whole thing again. Talking about it makes it seem somewhat complex. It really isn't and the whole business won't take long once you've done it. It will become quickly apparent that highest signal level and lowest noise level do not coincide. You can vary the bias, voltage, etc., and observe the effects of each on the signal-to-noise relationship. You can, in short, optimize your converter's performance.

A number of things can be done with the Chirper. There is room in the box to build another oscillator section connected to the transistor collector, operating in parallel and simultaneous with the first oscillator. By appropriate choice of crystal and attenuation, both signals can be introduced into the converter in order to adjust the mixer for minimum cross modulation.

Or, instead of using an oscillator at all, you can use the switching section of the Chirper to key a noise generator on and off. This has a certain attraction where an integrating network is used prior to a vtvm. In this case noise is used as a signal.

In spite of it's name, the Chirper is remarkably stable. Chirp becomes apparent from two meters or so, but is no problem. Build one and take the myths out of your converter.

UHF SIGNAL SOURCE

Bill Hoisington K1CLL

The UHF experimenter learns early that the UHF ham bands don't always furnish a signal when you need them. When I build receivers and converters for 432 and 1296, I find that I need a small signal generator for alignment and band spotting. Not much power is needed for this work, so transistors are the ideal choice for generating the signal. This signal source is in three parts, a 432 source, a tripler to 1296 and a modulator.

The 432 Driver

I started with a low-cost 27 mc crystal that I found in my junk box. (Never mind how it got there!) I used my usual phase-reversing crystal circuit followed by a bunch of doublers. Lower-priced transistors were used in the early stages and the UHF ones saved for later. You really need good ones on 1296! Each doubler is biased from rectification of rf from

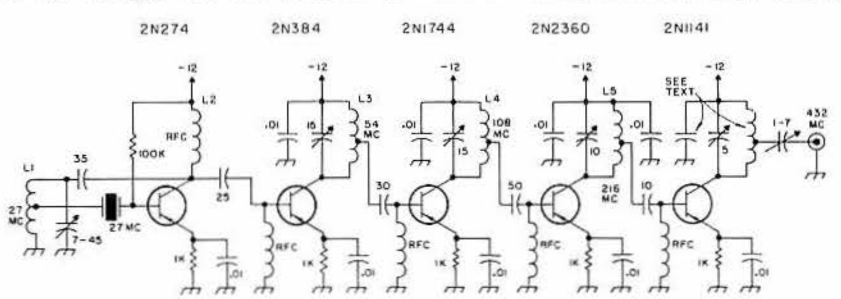

Fig. 1. Schematic diagram of the 432 mc signal source. The transistors aren't critical in most cases and other UHF and VHF ones will work fine.

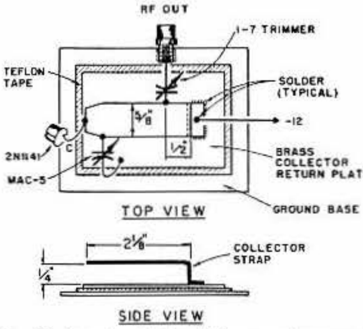
RF OUT
I-7 TRIMMER
TEFLON TAPE
SOLDER (TYPICAL)
3/8
-12
1/2
2N1141
C
MAC-5
BRASS COLLECTOR RETURN PLATE
TOP VIEW
GROUND BASE
2 1/8
COLLECTOR STRAP
1/4
SIDE VIEW

Fig. 2. Details of the 432 mc collector circuit.

the previous stage. If you need more output than this circuit gives you, use less than a 1 k resistor in the emitter, but watch out for high collector current. **Fig. 1** shows the schematic of the 432 mc generator and **Fig. 2** gives details of the 432 mc collector circuit.

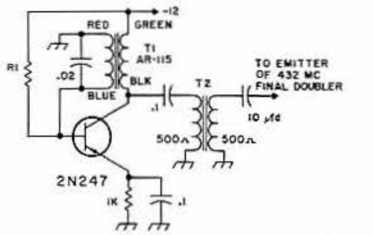
-12
RED
GREEN
T1 AR-115
R1
.02
BLK
BLUE
T2
TO EMITTER OF 432 MC FINAL DOUBLER
.1
10 μfd
500Ω
500Ω
2N247
IK
.1

Fig. 3. The tone oscillator for the 432 mc signal source.

The tone modulator

This is a crude modulator (**Fig. 3**), but it works. The modulation transformer is not absolutely necessary, but seems to improve results. You can apply the modulation almost anywhere for this application, but modulating across the 1 k emitter resistor gave the cleanest sound with the doubler used.

Now to 1296

I built this tripler to 1296 just for the fun of it. But it worked quite well. I normally don't hold with triplers at this frequency, but it's an easy way to get 1296 mc energy from the 432 mc driver. The transistor I used was a Motorola 2N1141. It's several years old and there are better ones that are far cheaper now. But it does work on 1296. I couldn't get it to work with grounded emitter, but grounded base is fine. **Fig. 4** gives the schematic but Fig. 5 gives the details, which are vital. The input on 432 tunes very nicely. But I had to reduce the emitter resistor in the doubler to 432 to get enough drive. It ended up at 200 ohms. The collector circuit is short, but tunes smoothly. The 1296 mc output registers 100 a in the 1296 cavity in the May 73.

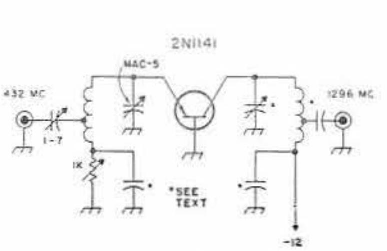
2N1141
MAC-5
432 MC
I-7
IK
*SEE TEXT
1296 MC
-12

Fig. 4. Schematic of the 1296 mc tripler.

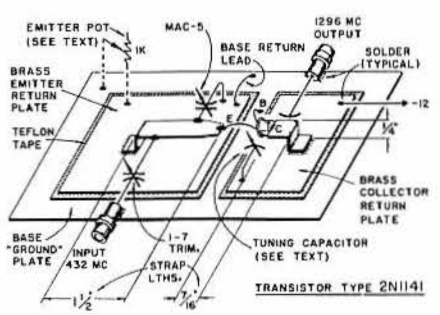
EMITTER POT (SEE TEXT)
IK
MAC-5
1296 MC OUTPUT
BASE RETURN LEAD
SOLDER (TYPICAL)
BRASS EMITTER RETURN PLATE
-12
TEFLON TAPE
E
B
C
1/4
BRASS COLLECTOR RETURN PLATE
BASE "GROUND" PLATE
INPUT 432 MC
I-7 TRIM.
STRAP LTHS.
TUNING CAPACITOR (SEE TEXT)
TRANSISTOR TYPE 2N1141
1 1/2
7/16

Fig. 5. Pictorial layout of the 1296 mc tripler.

You might try a small amount of modulation on the 1296 mc tripler. Also a waveguide attenuator. Be seeing you on 23 cm.

CRYSTAL CONTROLLED SIGNAL SOURCE WITH INFINITE ATTENUATOR FOR 144, 432 AND 1296 MHz

Bill Hoisington K1CLL

One of the most useful test-equipment gadgets the homebrewer can build is a signal generator. The one described here is of commercial quality and it can be completely contained inside a waveguide. Positioning, by sliding along the waveguide, provides a variable-strength stable signal of one millivolt, one microvolt, one nanovolt, or less, dropping down gradually to a true zero. It does this in a perfectly smooth fashion without steps or jumps so that every fraction of a decibel in lower noise figure shows immediately on the slide dial. What's more, the slide can be calibrated so that FM'ers can use the device for directly measuring receiver sensitivity in tenths of a microvolt.

In building a 6 meter receiver recently for maximum absolute sensitivity I naturally had to check especially on the first-stage rf transistor and circuit for minimum noise figure. (For this type of work you *must* have a signal generator capable of being attenuated out of sight with any receiver you can buy for any money.) The usual generators on the market under $100 do not do this. And many of the very expensive generators get so leaky that they have to be used 200 ft from the receiver. At any rate, the generator described here can be made up quickly and at low cost, and it *is* stable, reliable, and infinitely variable.

Waveguide

The only possible difficulty might be in obtaining the piece of waveguide needed. The piece I used is 4¼ in. wide by 2 1/8 in. high, and is 24 in. long. If you have a choice, get a piece a little longer. You could make up this item out of brass or copper if you had to, because in this case it is not used to *carry* energy but to *attenuate* it, so the worse you make it the better!

The waveguide *must not* have any holes in it and should be reasonably smooth inside; otherwise your dial would not read smoothly in attenuation. You *could* use copper or aluminum drain pipe, although I have not tried them yet. Working directly on the rf, this attenuator is good for any kind of modulation, including SSB, FM, pulse, or what have you.

Construction

Figure 1 shows the basic idea. When the signal generator plate is close to the

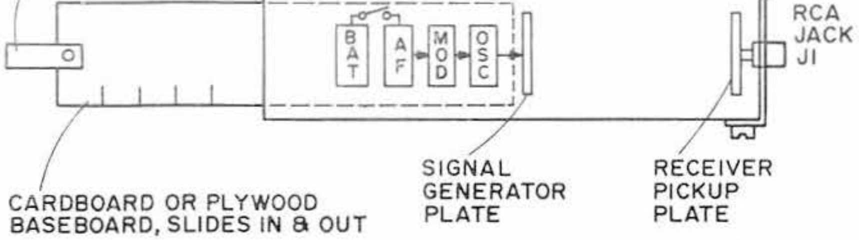
GENERATOR
DOWEL
WAVEGUIDE CLOSED AT REC. END, OPEN AT THE OTHER END
COPPER END CAP MUST BE TIGHT!
RCA JACK J1
BAT
AF
MOD
OSC
SIGNAL GENERATOR PLATE
RECEIVER PICKUP PLATE
CARDBOARD OR PLYWOOD BASEBOARD, SLIDES IN & OUT

Fig. 1. Sketch shows plan-view layout of unit inside waveguide attenuator. The oscillator unit is mounted on a flat wood or cardboard strip that can be calibrated to give accurate indications of output signal.

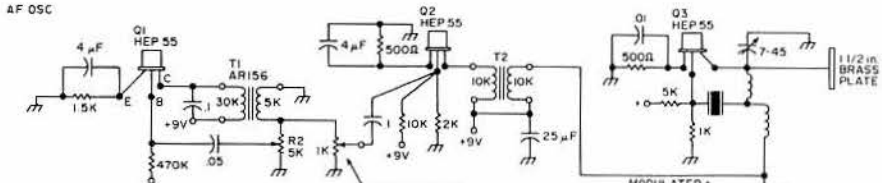

Fig. 2. Schematic of generator. This entire assembly about 3 x 4 in. including small 9V battery and switch must be entirely inside the waveguide. No wire or metal of any kind can be brought outside.

receiver pickup plate, you can get about 100 mV of signal into the receiver, and it is handy for checking diode receivers. When the two plates are about 8 in. apart, the signal is just detectable on a good receiver. Additional spacing between plates amounts to "waveguide beyond cutoff." I do not believe that there is any receiver in the world that can pick up the signal much beyond the 8¾ in. point.

Pretty soon in your receiver "peaking" work you get to that signal that may be but a tenth of a microvolt or so, and you begin dreaming about cryogenic front ends, masers, and such. As mentioned, every fraction of a decibel lower in noise figure, every improvement in sensitivity comes out rigorously and relentlessly on that slide dial. You can easily check which of your low-noise transistors is really low, whether that MOSFET will do a better or worse job for you, and in which circuit.

As you go up in frequency you may have to make smaller and smaller oscillators in order to fit in smaller waveguides to get the cutoff effect. (That will not be a problem if you read 73; the May issue described a "postage-stamp-sized" rf generator that is an ideal candidate for the signal source.)

Circuit

A crystal oscillator, an af oscillator, and a simple class A modulator do an excellent job to start with. Figure 2 shows the present unit as used on 6 meters. It must be stressed again that no wire or other piece of metal may be allowed to reach the outside from this assembly. I'm making up another for 2 meters soon (still my favorite band) and will try one on 450 a little later.

Audio

A controlled-feedback transformer-coupled af oscillator does a good job in furnishing a sine wave. A Motorola HEP55 is used for the oscillator, with feedback to the base from the collector through transformer T1, controlled by resistor R2. Audio output is taken off the 5 kΩ winding of T1, is fed through R4 the modulation control, and then to the base of af modulator Q2. Transistor Q2 is set up for low-power class A operation because not much modulation is needed for the signal generator. Transformer T2 is an old 5W unit from "tube-type portable" days. The secondary of T2 feeds a modulated +9V signal to Q3, the crystal-controlled 50 MHz oscillator.

This rf oscillator is one of my negative-feedback jobs with phase reversal in the crystal. A 1½ in. square plate is tied onto the collector, radiating energy to the receiver pickup plate facing it inside the waveguide. This energy is rapidly attenuated as you move the plates apart, and should be impossible to detect after some 9 or 10 in. of separation.

Once again, do *not* bring any wires or any other metal or conductor out from the oscillator assembly. If you want an outside controlled switch or other control, bring it out as a wooden dowel handle.

That's about it. Tune everything up outside the waveguide on the bench; when you're satisfied, plug your best 6 meter receiver into J1, push the oscillator plank along the waveguide (or rather I should say pull it along) away from J1. You'll get a surprise! Hope this helps you with your low-noise receiver work. It did a lot for me.

LOW COST 220 MHz SIGNAL GENERATOR

Bill Hoisington K1CLL

This article describes the design and construction of an easy to build, inexpensive, crystal controlled signal generator for the 220 MHz band, including a very low cost attenuator that goes from a quarter volt down through 1/20th of a microvolt and on to a real zero (of rf power). It is very useful for receiver front-end tuneup, low noise tests, and as a portable field generator for overall antenna tests through the receiver. For signal identification purposes, af and FM modulation are included.

If you really want to fight for a low-noise front end, this piece of equipment will be of great assistance to you, because the attenuation really is infinite and without any difficult bypassing or shielding.

Design of the attenuator.

Infinite attenuation is achieved here by the use of a 50¢ piece of aluminum tubing, as shown in Fig. 1. You cannot drive 220 MHz signals more than a few inches down inside of a piece of aluminum tubing. By putting *everything* – battery, on-off switch, circuit and all – on the movable generator strip and sliding it in and out of the tubing, you avoid all touchy, difficult and expensive bypassing, costly attentuator pots, shielding, etc., and provide a simplified means of varying the attenuation with stable, smooth, easy repeatability. Calibration is of the slide rule variety and also simple as far as writing down the microvolts on the scale is concerned.

This principle is older than radio tubes; in fact Sir Oliver Lodge used it in his 1890 microwave work.

Attenuator Construction Details

Figure 1 tells most of the story, with details in Figs. 2 through 9. An adequate rf seal can be made at the pickup end of the aluminum tubing, standard TV masting, 1¼ O.D., by 2 or 4 tabs in one end as in Fig. 3 and bending them back as shown, then cutting off the excess tubing. Install the pickup, plate, output jack, and end plate as shown in Fig. 1. I used time-saving external mounts for fastening it down to the wood baseboard as shown in Figs. 2 through 5. Drill a ¼ in. hole for a screwdriver as in Fig. 2, and use angles for the pickup end. Figure 2 also shows the scale in use for attenuation settings, and Fig. 3 shows pointer details.

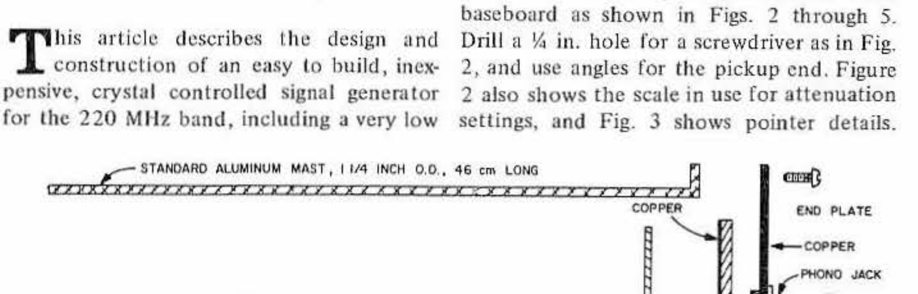
STANDARD ALUMINUM MAST, 1 1/4 INCH O.D., 46 cm LONG
COPPER
END PLATE
COPPER
PHONO JACK
SWITCH
BAT.
AF
OSC.
QUINT
OUTPUT
WOOD DOWEL
COPPER CLAD, 2.3 cm WIDE, 21.5 cm LONG
2 cm ROUND
INSULATE

Fig. 1. Sideview, 220 signal generator and infinite attentuator.

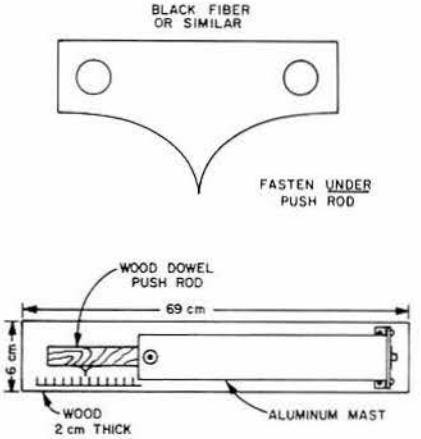
BLACK FIBER OR SIMILAR
FASTEN UNDER PUSH ROD
WOOD DOWEL PUSH ROD
69 cm
6 cm
WOOD 2 cm THICK
ALUMINUM MAST

Fig. 2. Top view of the 220 signal generator and details of the calibration pointer.

Figures 6 and 7 show pictorials of the layout, top view and side view.

The Generator

Nothing too fussy here, but attention to details will assure reliable af and rf oscillation at low power and low battery drain and good frequency multiplication. Figure 8 shows the schematic with the details of the two oscillators, the crystal in the 44 MHz range, and the quintupler. The af uses a standard circuit which, however, has one item to watch. Contrary to a transformer coupled circuit, which is seldom mentioned, this twin-T job has a nasty trick of not starting every time. However a small cap from collector to ground cures this and makes it 100% reliable in that respect. The emitter being grounded, I suppose this establishes the correct in-phase relation with the collector, in which both of these elements should be in phase. You can put a small trim pot af gain control between the modulator and the oscillator if you wish, watching out for dc voltages of course. As shown here, there is plenty of modulation for signal identification, both AM and FM.

Referring to Fig. 8, at the left is the af oscillator. It is not down symmetrically, but you can note the two 22K resistors and the two .02 frequency settling caps, along with the .05 and the 2.2K terminating the lines. All of these set the frequency, and to change the frequency you should vary all of them in at least their approximate present ratios. It is around 500 cycles as shown. Do not forget the "starting cap" from collector to ground.

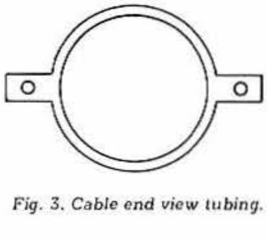

Fig. 3. Cable end view tubing.

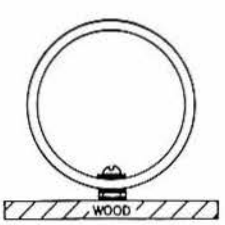

Fig. 4. Open end view.

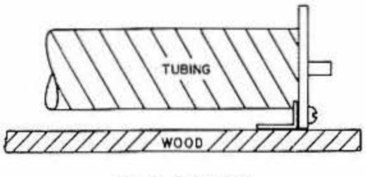
TUBING
WOOD

Fig. 5. Side view.

This audio is fed to the base of the multiplier where it provides some AM and some FM modulation for signal identification purposes. When working with receiver oscillators, and in particular with high-ratio multipliers, this is very important. It may also be locked into any old scope sync for noise figure and sensitivity comparisons. The scope sync gives a reference point where the signal to noise ratio will always be the same, without resorting to guess work. The crystal oscillator is my old tried and true crystal phase-reversing job, which uses negative feedback from the collector coil, which, after going through the crystal, reverses phase and becomes positive, thus assuring oscillation but *only* on the crystal frequency. A HEP 75 (similar to the famous 3866) is used for the quintupler. A lot more output is noted with this powerful but smooth operating old faithful, still good to 450 MHz.

The output goes to the antenna plate on the forward end of the baseboard plank. When this plate is moved all the way in so that it is only ¼ in. from the cable pickup plate in the end of the aluminum tube, a full scale reading may be obtained on a 50 microamp meter on the output of a tuned diode detector (see Fig. 9). My first model here has marks on the scale (see Fig. 2) for 1 rf stage; 2 rf stages; mixer (feeding into a good i-f strip); mixer plus 1 rf, etc. At the 1 rf plus mixer, you begin to hear all the repeaters within 100 miles or so. With 2 rf plus mixer (followed by a sensitive i-f of course) you are really getting sensitivity. This is where you put a scope in line in order to have an electronic comparison point for signal coming out of the noise purposes, and then can really get into the low-noise bit, if you have a hermit location. Just happens I have one here. When I hear the noise of a car, it's a visitor!

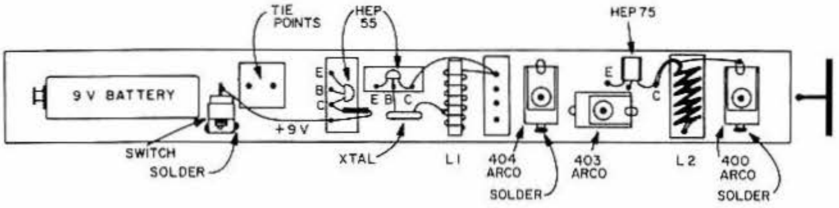
TIE POINTS
HEP 55
HEP 75
9 V BATTERY
E
B
C
E B C
+9V
E
C
SWITCH
SOLDER
XTAL
L1
404 ARCO
SOLDER
403 ARCO
L2
400 ARCO
SOLDER

Fig. 6A. Layout, top view.

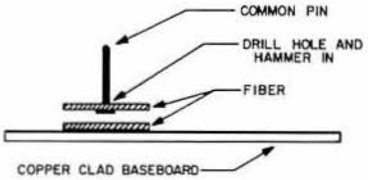
COMMON PIN
DRILL HOLE AND HAMMER IN
FIBER
COPPER CLAD BASEBOARD

Fig. 6B. Tie-point construction.

Output

You will see for yourself as soon as you start testing that the attenuator is smooth-working and stepless, and that true infinite attenuation is at hand. An rf input state (pre-amp) with a noise figure a fraction better than another shows right away on the scale. For example, adjustment of the fixed bias voltages on the two gates of a 3N200 or 3N201 shows right away on the scale as the

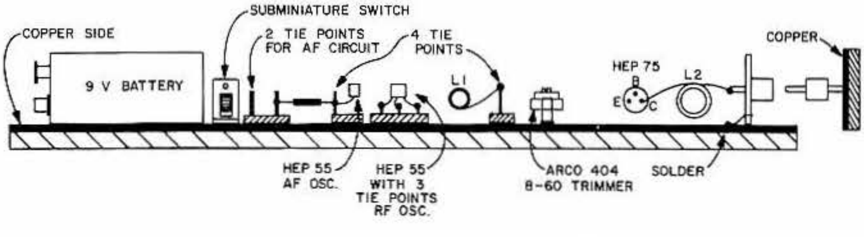
SUBMINIATURE SWITCH
COPPER SIDE
2 TIE POINTS FOR AF CIRCUIT
4 TIE POINTS
COPPER
HEP 75
L2
L1
B
E
C
9 V BATTERY
HEP 55 AF OSC.
HEP 55 WITH 3 TIE POINTS RF OSC.
ARCO 404 8-60 TRIMMER
SOLDER

Fig. 7. Side view of the signal generator, 220 MHz.

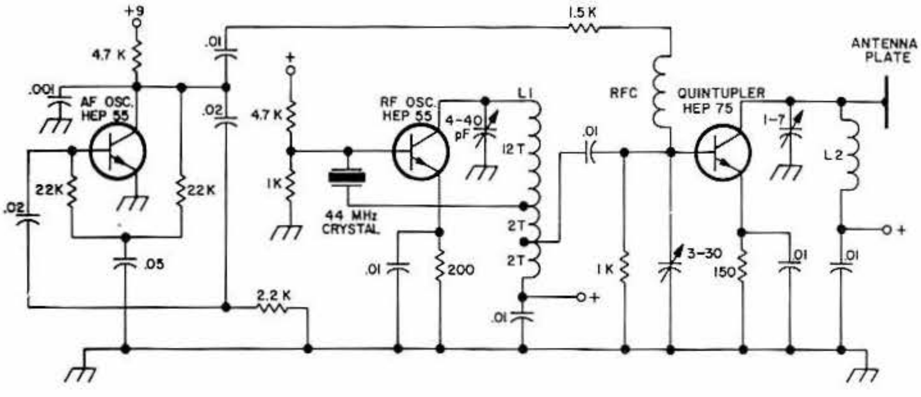
+9
4.7 K
.001
AF OSC.
HEP 55
22K
.02
22K
.05
.01
.02
4.7 K
+
1K
44 MHz
CRYSTAL
2.2 K
1.5 K
RF OSC.
HEP 55
4-40
pF
L1
12 T
2T
2T
.01
200
.01
+
.01
RFC
1K
3-30
QUINTUPLER
HEP 75
1-7
150
.01
L2
.01
+
ANTENNA
PLATE

Fig. 8. Schematic. L1 = 16 turns No. 26 output tap at 2 turns, crystal feedback tap at 4 turns, from cold end. Wound on phenloic form .6 cm O.D. L2 = 6 turns No. 18 bare, air wound, .6 cm O.D., 2.5 cm long between tie points. RFC = about 40 turns No. 40, on phenolic form .3 cm O.D., 1 cm long (not critical).

push-rod is moved in and out and the signal is locked onto the scope. This work you can do right on the bench and at low cost.

There is quite a bit of mechanical work in the unit, depending on just how much "finish" you want it to have. You can also bring a dowel rod for "on and off" use. Do not, under any circumstances, bring out a *conductor*. You *can* do this, but only with an extreme amount of filtering, which is not part of this article. Be sure to set up the baseboard, antenna plate and battery first, and get them working mechanically. With a drawn-out shape like this I generally start with a longer piece of copper-clad than needed, build from one end, and then cut off what is left over. Understandably, once you have made the first one you can always see, after it is done, many ways of improving it. However, someone has to make the first one, and that's generally my job.

Antenna and Field Tests

Out of the tubing, and with a small antenna connected via a one turn link around the quadrupler coil and then to ground, returning L2 for maximum output, this little rig puts out a lot of signal on 220. Especially if you reduce the oscillator emitter resistor! Up to several volts of rf can be obtained in a tuned diode receiver if you push things along, which is around 5 to 10 mW. If you place this generator out in a field several hundred yards or more away, you can then line up your antenna on the car or house, check antenna cables, antenna input alignment, and match or mismatch for lowest noise figure, etc.

Front end alignment should first be done with a relatively broadband i-f strip on 10.7 MHz. Be sure nothing metallic on the generator strip protrudes enough to touch the inner wall of the tubing, or "scratch" will occur in the high gain receiver. A piece of thin fiberglass or other insulating sheet wrapped around the whole generator movable plank is a good precaution.

Once again I include a 220 MHz tuned diode detector, which is an absolute must for frequency multiplication, especially quintupling and such, where the other unwanted harmonics are as little as 20% away from the desired frequency. Figure 9 shows this piece of test equipment in pictorial schematic form. Remember, shape is of considerable importance as you go from VHF into UHF. It is quite easy to make the square trough line out of an old piece of copper clad. Or even a new piece! And this particular one described and shown in Fig. 9 goes very well to over 450 MHz and thus is very nice for the next band also. If you make it just as I've shown it, it will do a good job for you.

Calibration will present some difficulties, so line up some other lads around who are already on these bands and get your calibration that way.

So good luck, friends, more coming – lot's more! Keep reading.

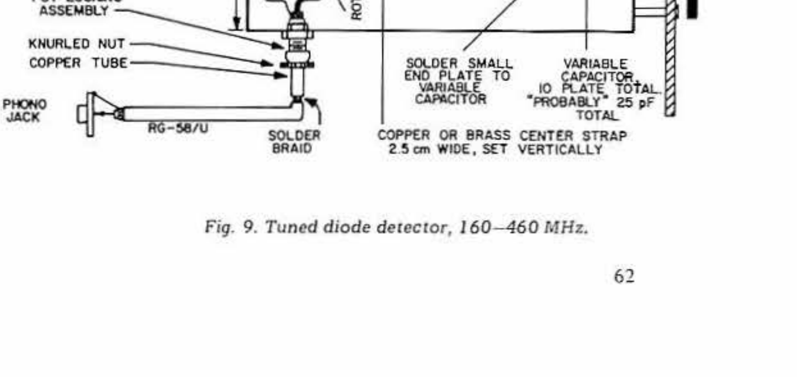
AF
DC
OUTPUT
"PHONO" JACK
ADAPTER
(EXTERNAL)
1.4 cm SPACERS
COPPER-CLAD
COPPER INSIDE
TWO SMALL
LAFAYETTE
.001 CAPS
IN295
DIODE
10 X 10 cm
PANEL
POINTER
KNOB
10.2 cm
5 cm
7.6 cm
INSIDE TEFLON
OF RG-58/U
2.2 pF
3 cm
ROTATES
POT LOCKING
ASSEMBLY
KNURLED NUT
COPPER TUBE
PHONO
JACK
RG-58/U
SOLDER
BRAID
SOLDER SMALL
END PLATE TO
VARIABLE
CAPACITOR
VARIABLE
CAPACITOR,
10 PLATE TOTAL
"PROBABLY" 25 pF
TOTAL
COPPER OR BRASS CENTER STRAP
2.5 cm WIDE, SET VERTICALLY

Fig. 9. Tuned diode detector, 160–460 MHz.

UHF SIGNAL GENERATOR

Jim Kennedy K6MIO

Several requirements were the basis for the design of this signal generator: operation on both 432 mc and 1296 mc, high stability e.g., crystal control, variable rf amplitude, provision for insertion of various types of modulation, and lowest possible cost consistent with satisfactory operation.

I had a 27.005 mc overtone crystal left over from a brief period of disillusionment about CB and, since 27 times 6 is 432 which times 3 is 1296, this seemed a likely place to start.

Fig. 1 shows the final result. The overtone oscillator is voltage regulated and is left operating at all times when the generator is on. The crystal is mounted underneath the chassis to protect it from rapid temperature changes from drafts, etc., further enhancing the stability.

The use of diode multipliers at 432 mc and 1296 mc greatly simplifies developing UHF rf.

Direct coupling the modulation into the cathode of the last vacuum tube multiplier provides a modulation input that will accomodate almost any signal from audio to video, or even pulse.

A four position mode switch on the front panel allows the choice of carrier, no carrier, carrier with 60 cycle modulation and finally carrier with external modulation.

Construction

Building the unit presents no special problems except getting it all under the chassis. The 5 x 9½ x 2½ chassis doesn't leave much useful

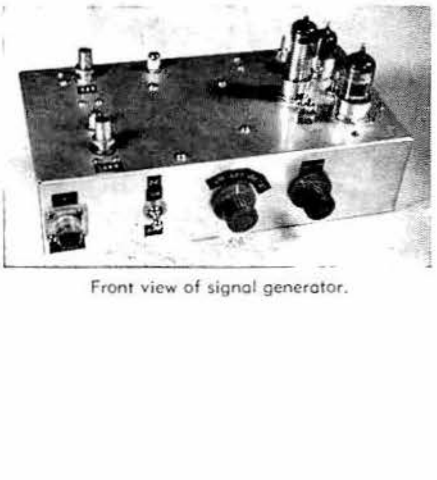

Front view of signal generator.

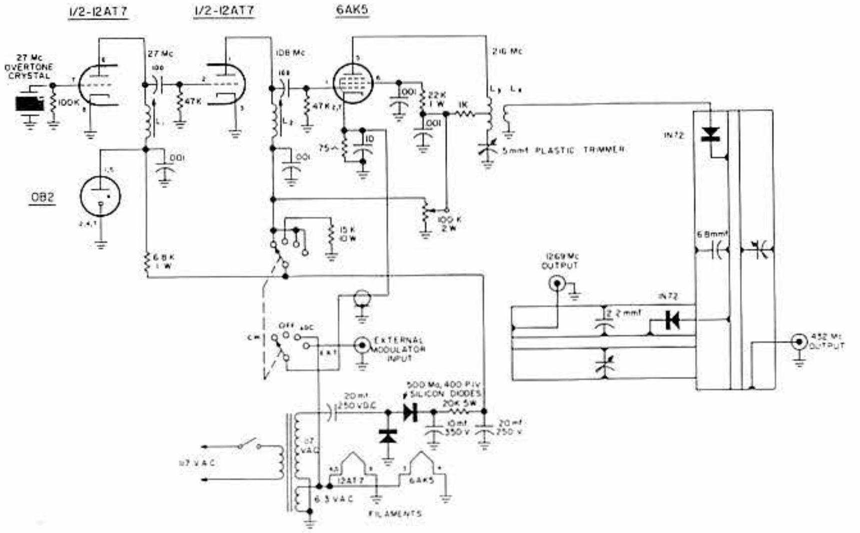

FIGURE 1

L1—10 turns # 28, 3/8" slug form.
L2—3 turns #28, 3/8" slug form.
L3—10 turns #18, 1/2" dia.
L4—3 turns #18, 1/4" dia.

room to spare, but it does make for a very compact device.

The lengths of the diode multiplier tuned lines were determined primarily by the space available and bear only the slightest relation to the wave lengths involved; hence, it was necessary to pad both lines with additional capacitances in order to resonate them properly.

Both lines were folded from 1/32 inch brass sheet stock. The inner conductors are made from ¼ inch brass tubing. The 432 line is 1" x 1" x 7". The input diode is tapped onto the inner conductor 1¼ inches from the other end. The signal output loop is about 1¼ inches long. The output diode is tapped 2 inches from the loop end of the line.

The tuning capacitor is made by soldering a 10-32 nut over a hole in the middle of the line shell. This provides the threads for the 10-32 screw which actually serves as the capacitor.

Round ½ inch disks are soldered to the bottom of the center of the inner conductor and to the end of the turning screw to provide additional range.

The 1296 line is similar except it is only 3⅜ inches long. The multiplier diode is tapped on the inner conductor ⅝ inches from one end. The signal output loop is about ¾ inches long and is located at the other end of the line. The tuning capacitor is identical to the one used in the 432 line.

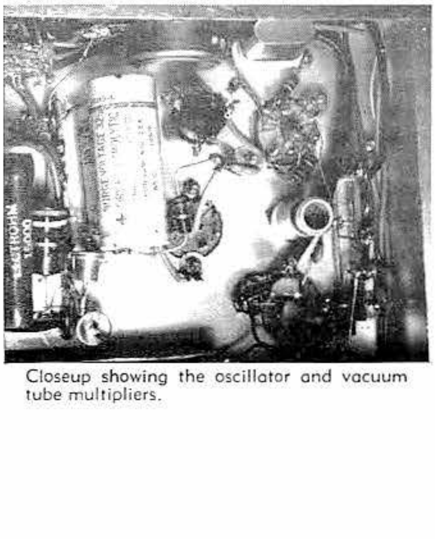

Closeup showing the oscillator and vacuum tube multipliers.

The 15K 10 watt resistor in the carrier OFF position limits the no load voltage from exceeding the ratings on the filter capacitors.

Operation

The tune up of the tube stages can be easily accomplished with a grid dipper. The UHF stages should only need to be peaked with a converter and receiver.

Shields on all tubes and a bottom cover on the chassis are important requirements if stray leakage at the signal frequency or lower harmonics is to be eliminated as a possible source of measurement error. These were removed for the photographs but are always used in practice.

For best results it is a good idea to employ a 50 ohm (or 75 ohm as the case may be) pad between the signal generator and the equipment under test. This will insure that the generator will appear as a resistive load as much as possible. This is necessary if the generator is to duplicate the conditions presented to the receiver by a good antenna.

This pad can take many forms. The two easiest ways to make one would be to use sufficient lengths of some lossy coax. At 432 mc higher, 20 or 30 feet of RG-58 (or RG-59) seems to be quite adequate for most applications. The other approach would involve building a conventional resistor pad into a male and female BNC connector soldered back to back. Fig. 2 shows such a pad.

Aside from receiver testing, this device has several other uses to which it can be put. It is also useful as a signal source for antenna tests and it may be used as frequency standard.

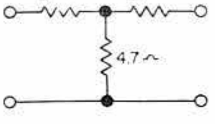

FIGURE 2

Fig. 2. 20 db 52 ohm pad. Resistors mounted in BNC connector.

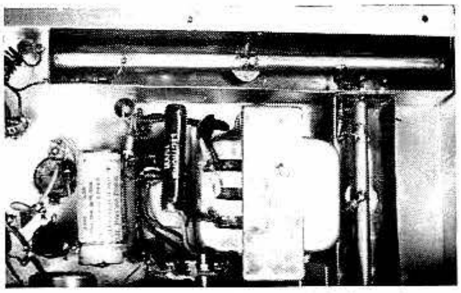

Closeup view showing details of diode multipliers lines.

This last use can be of great value at UHF because of the persistent problem of crystal tolerance versus frequency multiplication. A converter can easily be 40 or 50 kc or more off of the indicated frequency at 1296 mc. This can be quite a problem if you are hunting for a moonbounce signal through a 5 cps filter.

Though not shown on the diagram or in the pictures, it was found that a 50 mmfd variable capacitor connected across the crystal was sufficient to reduce 27.005 mc to the 27.000 . . . mc required to put the generator on 432.00 . . . and 1296.00 . . . mc. It will probably be necessary to retune the oscillator plate circuit for best operation; some loss in output will be encountered but this should not be a serious problem. However, if such operation is contemplated, it might be of some advantage to obtain a 27.002 mc crystal and employ a smaller variable capacitor.

In order to make the frequency standard really something of a standard, a 1 mc calibration oscillator should be employed. Some harmonic of the oscillator (5 mc, 10 mc, etc.) should be zero beated against WWV. The signal generator should then be zero beated against the 27 mc harmonic of the 1 mc calibrator—presto!—a 432-1296 mc frequency standard.

Though there are many refinements which can be made, the unit described has done a yeoman service in all the varied tasks asked of it.

LAB TYPE IF/RF SWEEP GENERATOR USING ICS

Ray Megirian K4DHC

Just about every kit-type sweep generator on the market starts life out at around 3 or 4 MHz and ends up at 200 MHz or so. This always made analysis or display of 456 kHz i-f amplifier response characteristics a tedious step-by-step plotting procedure or required time out to rig up a temporary gimmick to do the job. No more. The sweep generator to be described here will do the job with ease and precision. Even if you do not intend to duplicate the instrument, you may find parts of it useful for other applications.

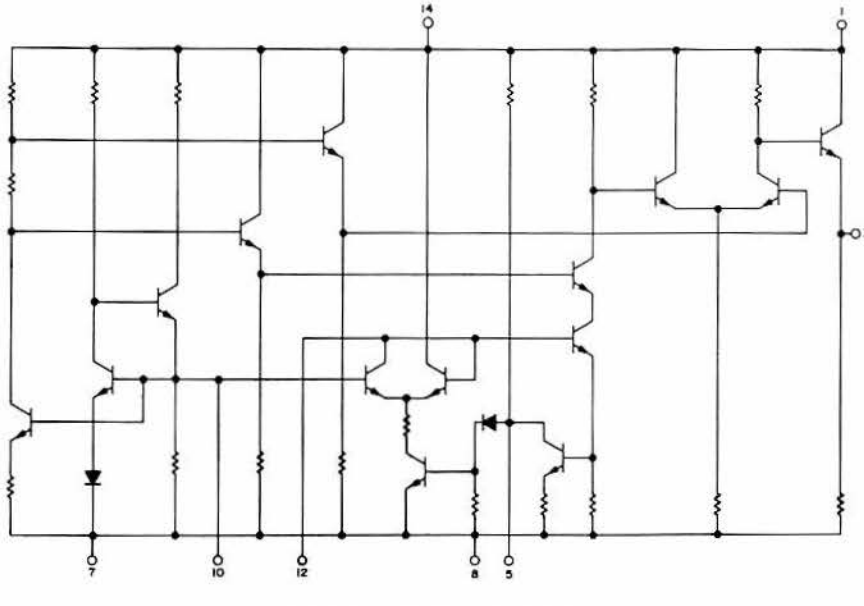

Fig. 1. Schematic of the MC1648 IC.

The generator covers a frequency range of from 400 kHz to 30 MHz in five bands. It can be operated in the CW mode as well as swept, thus allowing it to be used as an ordinary signal generator. Maximum output is 350 mV p-p across 50Ω. When sweeping, the return trace may be blanked or not, as desired. Two calibrated dials are provided for setting the start and stop frequencies and the maximum sweep width would be the entire band in use. The frequencies covered by the five bands are: 400 to 900 kHz, 850 kHz to 2.3 MHz, 2 to 6 MHz, 5 to 15 MHz and 10 to 30 MHz. Sweep time is variable between 20 ms and 6 seconds per sweep. A step attenuator in conjection with a vernier control provide a maximum attenuation of the output of 120 dB. An input is provided for a post-injection marker system built into the unit. A synchronous ramp with gain control is provided for driving the oscilloscope horizontal sweep. Blanking pulses are also brought out to a connector in case they are needed for synchronizing external equipment.

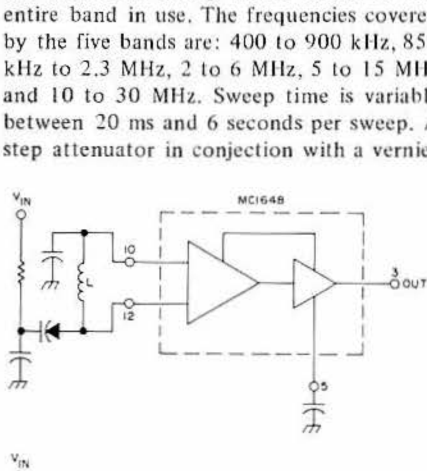

Fig. 2. Tuning the MC1648 with a single diode (A) and using back-to-back diodes (B).

A New IC

The heart of the generator is a new Motorola IC, the MC1648 emitter-coupled oscillator. It was intended for use in phase-locked loop systems operating in excess of 150 MHz, but may be used in many other applications such as this one. The device provides output of high spectral purity and incorporates an internal agc system which simplifies design of the sweep generator by eliminating the need for external leveling. A buffer amplifier and emitter-follower output are also incorporated on the chip, eliminating the need for external amplifiers. Figure 1 is a schematic of the MC1648 IC. Figure 2 illustrates two methods for tuning the oscillator. The device is packaged in a 14-pin DIP and requires a 5V dc supply. Since it was intended to be used with Motorola MECL III logic, either polarity is permissible. A positive supply is used here, connected to pins 1 and 14 with 7 and 8 grounded.

All the information covered by the data sheet for the MC1648 dealt with operation from 10 MHz up to about 180 MHz. Since I was interested in going as low as 400 kHz, I had to do some experimenting with tank circuits. My best results were obtained with the use of cup cores for the two lowest bands. I also found that ordinary molded iron core rf chokes of the miniature variety did an excellent job the rest of the way. Use of these tiny components made possible a very compact 5-band assembly.

The VCO

In order to facilitate shielding and simplify construction, electronic band switching is employed. Figure 3 is the schematic for the rf portion of the generator. The 2N4391 J-FET has a low 'on' resistance and works very well as a switch. One of these transistors is placed in series with each tank circuit

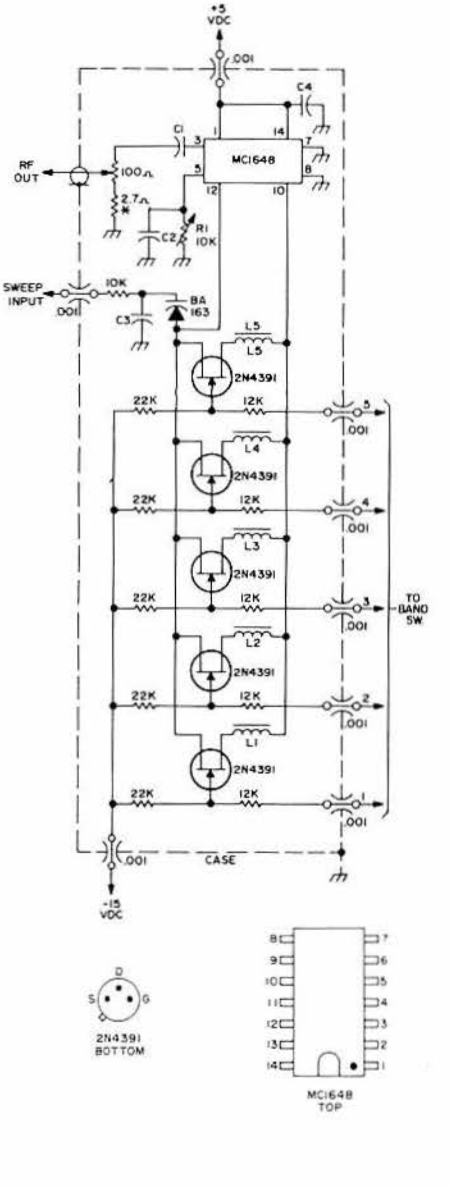

ALL RESISTORS-1/4W, 5%
L1-28 TURNS NO. 30 AWG ENAMEL.
L2-12 TURNS NO. 28 AWG ENAMEL.
(L1 & L2 MOUNTED IN CUP CORE-See Text)
L3-18 μH MINIATURE MOLDED RFC.
L4-4.7 μH MINIATURE MOLDED RFC.
L5-1.1 μH MINIATURE MOLDED RFC
C1, C2-TYPE 150D TANTALUM-2.2 μF/20V.
R1-BECKMAN 89PR10K 15-T TRIMMER.

*-SELECTED PART-See Text.

Fig. 3. Schematic of the rf assembly using dc band switching.

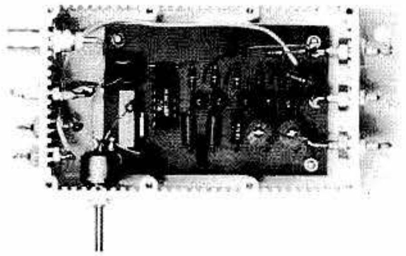

Completed rf assembly ready for mounting in the cabinet. A plastic version of the 2N4391 was used with a considerable savings in parts cost.

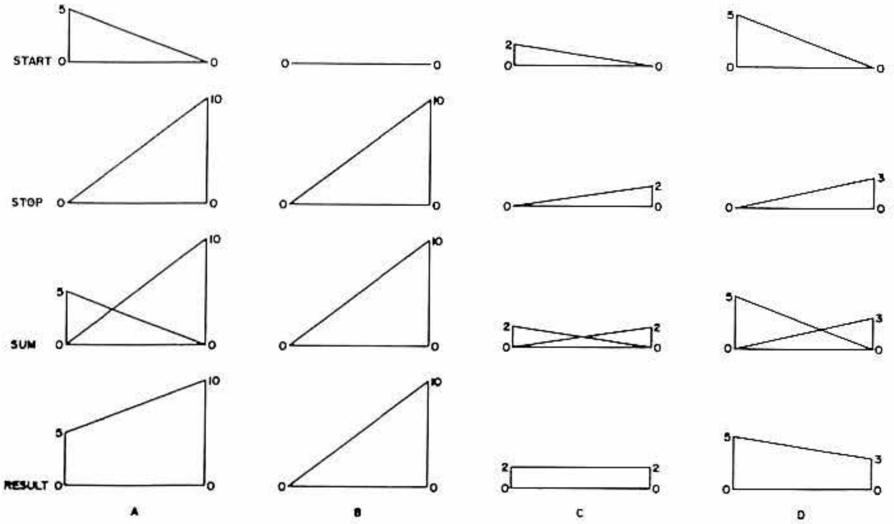

Fig. 5. Pictorial demonstration of ramp generation.

and can be turned on by application of a positive voltage at the gate. This allows the use of a strictly dc-operated remote switching arrangement. The entire assembly is built on a 2 x 3 in. pc board and mounted in a Pomona Model 3306 enclosure with a Model 3328 bottom mounting plate.

A BA163 tuning diode by ITT is used to sweep the oscillator. Although it is intended for bc band use, it performs admirably at these higher frequencies and its high capacitance ratio allows wide sweep excursions. If you are planning to use this oscillator for VHF applications, a more suitable diode would be in order. According to the data sheet on the MC1648, typical maximum output frequency is 225 MHz.

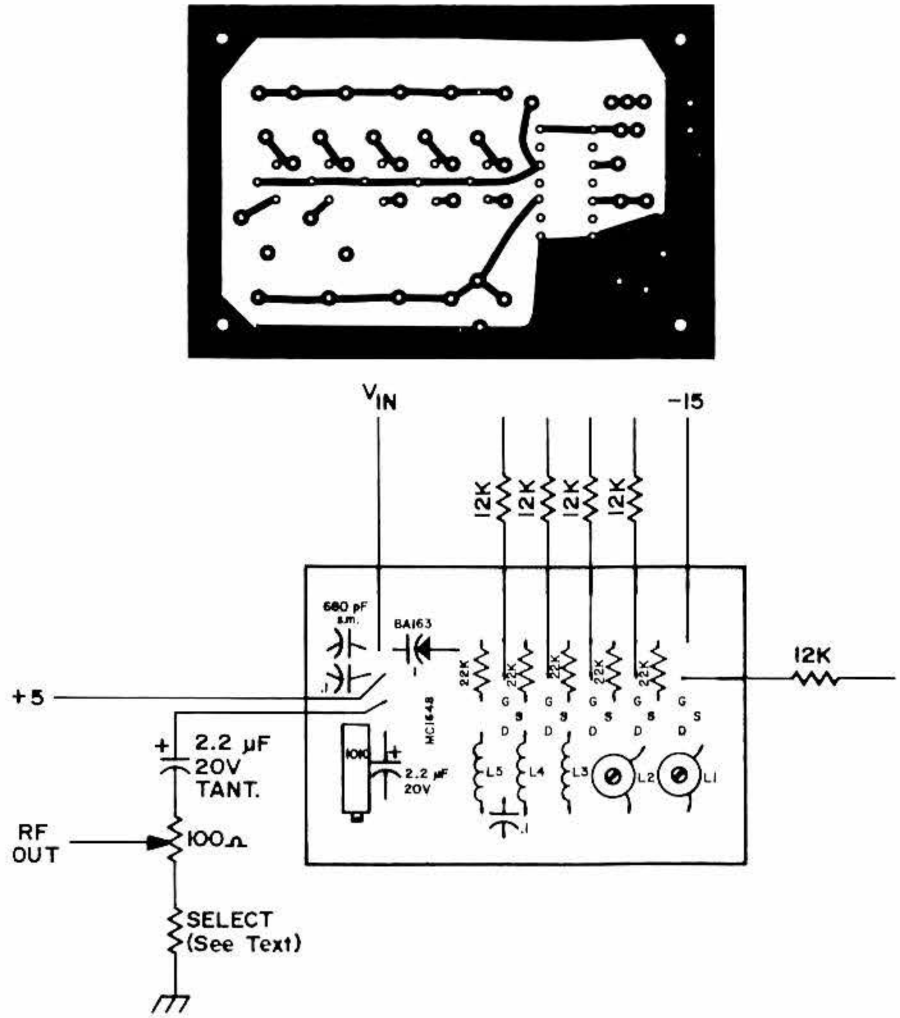

Fig. 4. (A) Foil side of rf pc board. (B) Location of parts on component side.

The vernier attenautor is a 100Ω RV6 style potentiometer mounted inside the enclosure. A coupling and extension shaft are used to bring the control out to the front panel. The value of the resistor at the bottom end of the pot will have to be selected so that a range of 20 dB is provided with full swing of the control. Since end resistance varies from one control to another, an exact value will have to be arrived at experimentally.

The output from the VCO is a square wave. If sine wave output is desired, the agc characteristic may be modified by introducing resistance between pin 5 and ground. The small trimmer incorporated on the board is used for this purpose. It was my experience, however, that this may introduce instability with some MC1648s. Although the trimmer was left in, the one in my unit is turned to maximum resistance where it has no effect and left there. It would do no harm to leave it out altogether since no other changes would be needed.

If any readers are contemplating exact duplication of this generator, you will find the cup cores I used are no longer available. These were Ferrox-cube part number 332P133B4-3C and are obsolete. I had a large number of these on hand left over from another project and saw no sense in buying more. I am sure some of the presently available cup cores will make excellent substitutes. Those used here are about 3/8 in. in diameter and are ungapped. You may have to experiment with the number of turns in the coil to obtain the coverage you require. One handy trick you can pull with a cup core is to rotate the slots in each half so that fractional turns are produced when the coil wires are brought out separately through the

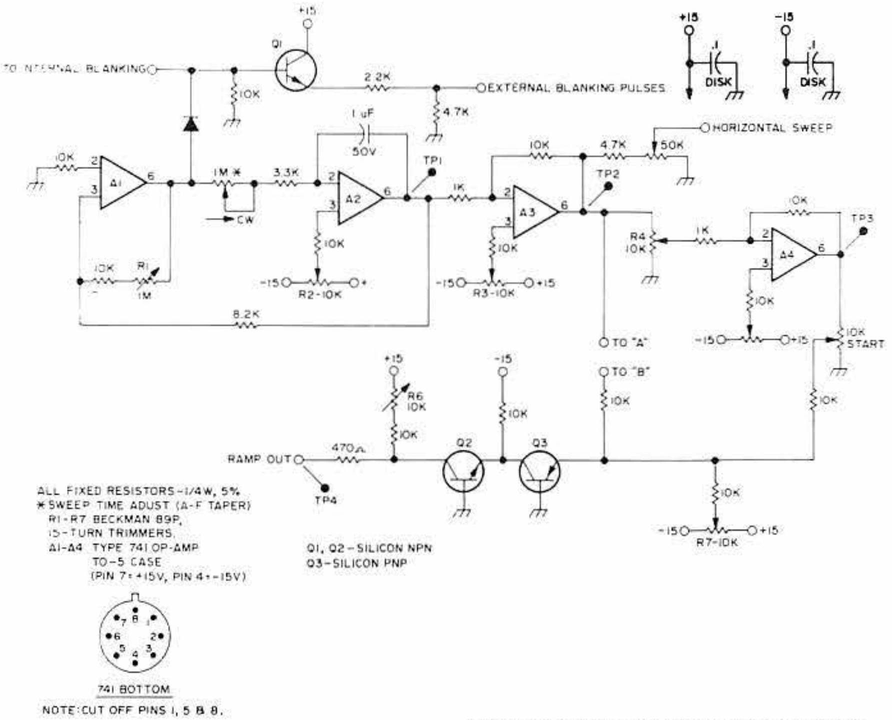

Fig. 6. Schematic of the ramp generator circuit.

two spaced holes. Exerting pressure on the core by means of the mounting screw will also shift the frequency and is almost like having a slug to tune. Increasing pressure seems to raise the frequency. The rf chokes used for the higher bands may also vary slightly from piece to piece and several may have to be tried to get the desired coverage. There isn't much we can do to alter the frequency where the chokes are used. Of course, slug-tuned coils may be used if the pc layout is modified. I do not recommend their use on the two lowest bands, however, since my results were rather unsatisfactory with this sort of tank circuit.

One final point would be in order before leaving the VCO. Note that the 10 KΩ resistor going to the tuning diode, the 2.2 μF rf output capacitor and each of the 12 KΩ resistors going to the transistor gates are not mounted on the board. One end of each has a hole provided on the board and then the component itself is used to make the connection to its final destination. This saves board space and eliminates separate wires. The small resistor at the bottom of the vernier attenuator is similarly mounted between the board and the bottom lug on the control. These points are more clearly seen in the photo. Also note that the cup cores are fastened directly to the board by means of 1-72 screws into threaded holes.

The Ramp Generator

In contrast with kit-type sweep generators that use the 60 Hz line for sweeping the oscillator, laboratory instruments have internal circuitry designed for this purpose. In our case, as with most sweep generators, a voltage ramp is used to control the oscillator during the sweep period. The circuitry has been arranged so that we can adjust the starting point (dc level) of the ramp independently of the stopping point. This simply means that we can set the frequency at which the sweep starts as well as the frequency at which it stops. We can also control, over a fairly wide range, the time it takes for a complete sweep. In this case, the sweep time is continuously adjustable between 20 ms and 6 seconds per sweep.

To get a better idea of how the ramp generator produces the results we've described above, let's look at Fig. 4. Note that a negative-going linear ramp (start) is combined with a positive-going linear ramp (stop) to produce the resultant ramp being applied to the varactor diode. If either input ramp is zero, the resultant will be the same as the single ramp since adding zero to anything will not change its value. Naturally, if both are zero, the output will be zero and the oscillator output frequency will be constant during the sweep period. The same is true whenever both ramps are of equal amplitude, except that the output will be a steady voltage other than zero. Also note that the resultant ramp could have a negative slope if the 'stop' input is lower than the 'start' input, and the oscillator would obviously sweep down in frequency rather than up.

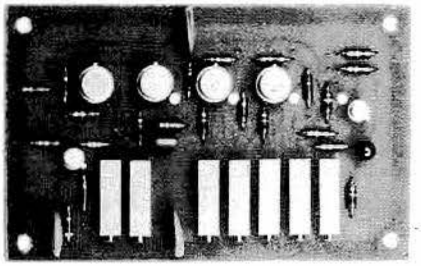

The ramp and blanking generator pc assembly. The leads going to external points will go in the holes visible on the board.

The schematic for the ramp generator is shown in Fig. 6. Operational amplifiers A1 and A2 form a triangle wave generator with A1 acting as a threshold detector and A2 as an integrator. One half of the triangle (negative-going) represents the sweep period while the other half represents the retrace time. Since a square wave is generated at the output of A1, we have a convenient source of blanking voltage built right in. This signal is positive during the sweep period and is connected to the band switch when Swept 1 (blanked retrace) operation is selected. During retrace the square wave drops to zero and actually disconnects the tank circuit until the next sweep starts. An emitter-follower, Q3, buffers the square wave output for external use.

The output of inverting amplifier A3 is a positive-going ramp during the sweep period and is used for both the 'stop' signal and horizontal sweep for the oscilloscope. Since an additional inversion takes place in the output of A4, the negative-going ramp at this point is used as the 'start' signal. These two ramps are applied across front panel controls which have 6:1 reduction drives and are fitted with dials calibrated in frequency for the five operating bands. The outputs from these two pots are fed to summing amplifier Q1, where the output will be the resultant ramp we discussed above. Because the output Q1 is in the negative region (PNP transistor), a second common-base (NPN) amplifier is used to shift the output back to where it will always be positive.

The ramp and blanking generator is constructed on a pc board 2.6 x 4.1 in. in size. For convenience, all the trimmer resistors were mounted along one edge. As seen in the photo, this board was mounted by means of two small brackets in a vertical position with the trimmers facing up. To simplify pattern layout, unused pins 1, 5 and 8 on each of the four 741 opamps were clipped off right at the case. Color-coded wires were connected to all necessary points on the board before it was mounted, with the leads made long enough to reach their destinations. For those interested in building this unit, a detailed procedure for setting up the ramp generator will be given later.

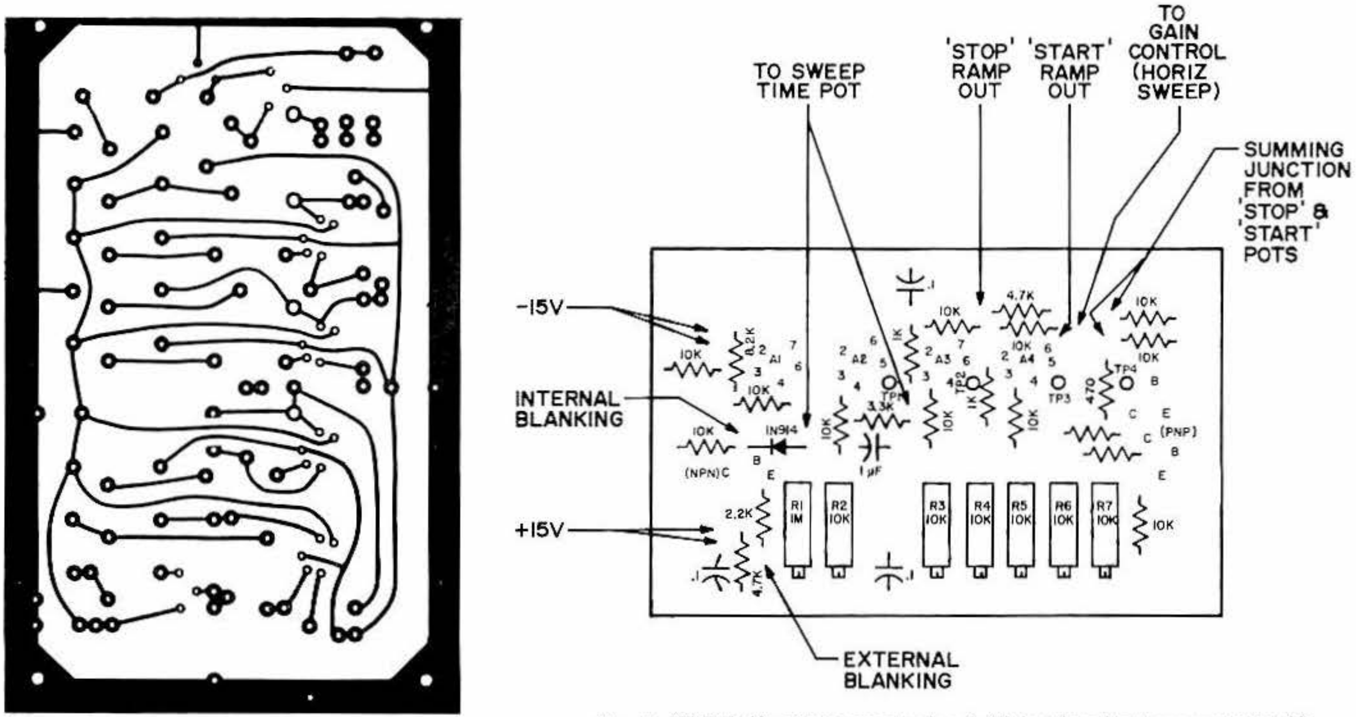
TO SWEEP TIME POT
'STOP' RAMP OUT
'START' RAMP OUT
TO GAIN CONTROL (HORIZ SWEEP)
SUMMING JUNCTION FROM 'STOP' & 'START' POTS
-15V
INTERNAL BLANKING
+15V
EXTERNAL BLANKING
A1
A2
A3
A4
TP1
TP2
TP3
TP4
IN914
(NPN)C
(PNP)
R1 1M
R2 10K
R3 10K
R4 10K
R5 10K
R6 10K
R7 10K
2.2K
4.7K
8.2K
3.3K
1K
470
10K
.1
1 µF

Fig. 7 (A) Foil side of ramp generator board. (B) location of parts on component side.

The Mixer/Amplifier

In order to provide a means for displaying frequency markers, a mixer and amplifier circuit was built into the sweep generator. The schematic for this circuit is shown in Fig. 6. A marker generator was not included inside the sweeper because other rf generators were available for use as markers and the added expense seemed unnecessary. Because the mixer board was to be installed inside a shielded box, dc gain control was resorted to as with the band switching discussed earlier. A 2N4391 is used as a variable resistor to control the signal input to the amplifier stage. This in turn controls the size of the marker being displayed on the curve. A 2N3819 FET is used in a simple mixer circuit which worked the best of several tried. Incidentally, the PNP and NPN transistors used here and on the ramp generator board are unknown surplus silicon units.

The pc board for the mixer is 1.5 x 2.9 in. and is mounted in a Bud CU-124 die-cast aluminum box. Note that here again some of the components are mounted between the board and their external connectors. These include both the 10 pF and 100 pF input capacitors to the 2N3819 and the 10 KΩ resistor going to the gate of the 2N4391.

The Power Supply

Figure 7 is a schematic for the power supply. Voltage requirements for the sweep generator are +15V, −15V and +5V dc. The current drawn by any of the above circuits runs no more than 30 or 40 mA maximum,

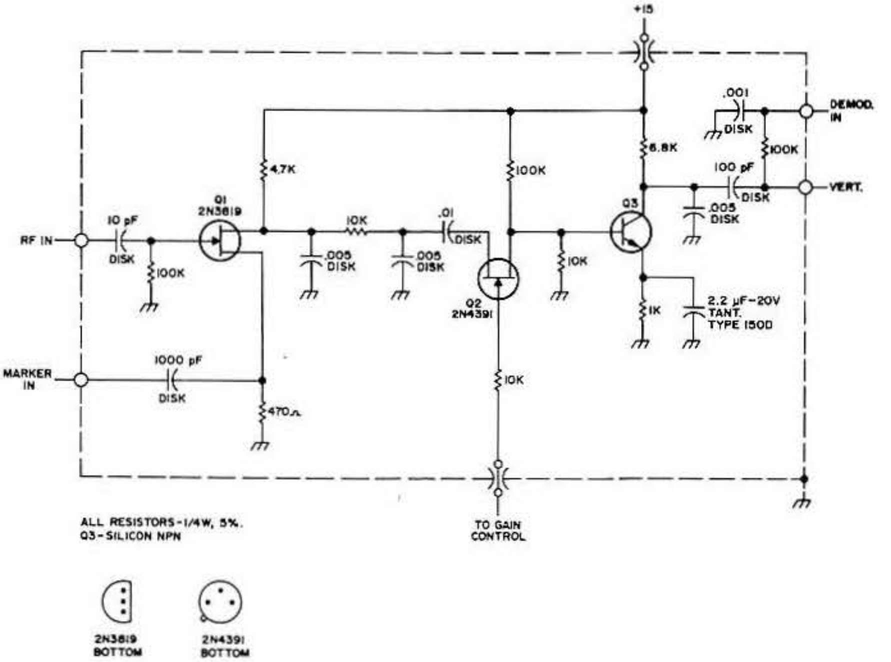
+15
RF IN
MARKER IN
10 pF DISK
100K
1000 pF DISK
470Ω
4.7K
Q1 2N3819
10K
.005 DISK
.01 DISK
.005 DISK
100K
Q2 2N4391
10K
10K
6.8K
Q3
1K
.001 DISK
100K
100 pF DISK
.005 DISK
2.2 µF-20V TANT. TYPE 150D
DEMOD. IN
VERT.
TO GAIN CONTROL
ALL RESISTORS-1/4W, 5%.
Q3-SILICON NPN
2N3819 BOTTOM
2N4391 BOTTOM

Fig. 8. Post-injection marker circuit with dc gain control of marker size.

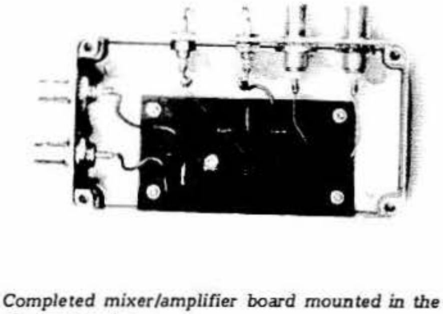

Completed mixer/amplifier board mounted in the die-cast housing.

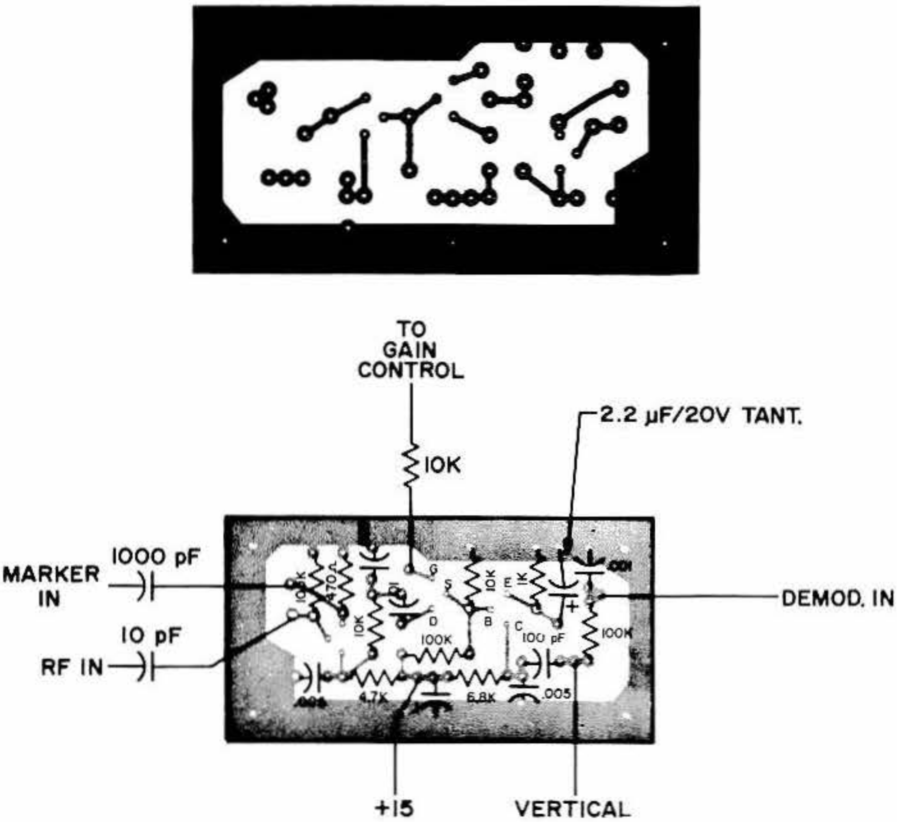
TO
GAIN
CONTROL
10K
2.2 µF/20V TANT.
MARKER IN
1000 pF
RF IN
10 pF
DEMOD. IN
470Ω
100K
6.8K
4.7K
100 pF
.001
.005
+15
VERTICAL

Fig. .9. (A) Foil side of mixer/amplifier pc board. (B) Location of parts on component side.

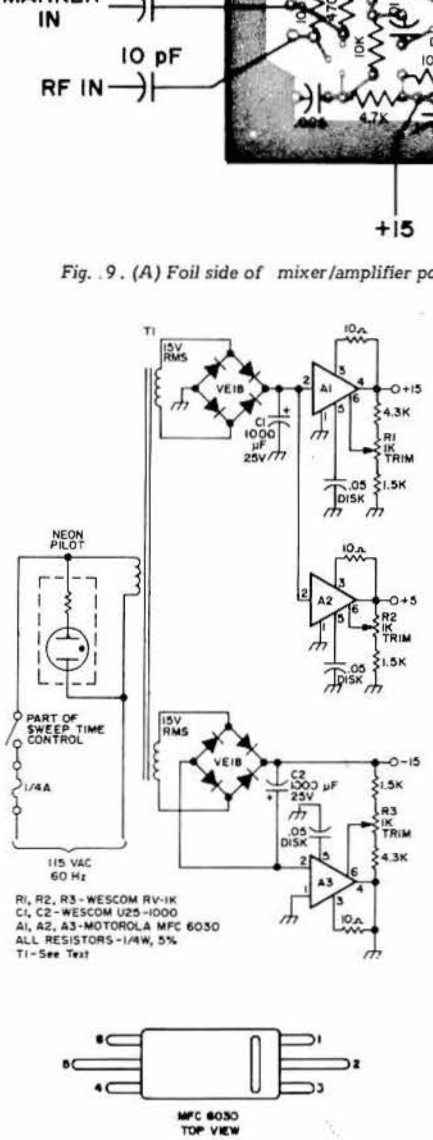
T1
15V RMS
VE18
C1 1000 µF 25V
A1
+15
4.3K
R1 1K TRIM
.05 DISK
1.5K
10Ω
NEON PILOT
A2
+5
R2 1K TRIM
PART OF SWEEP TIME CONTROL
1/4A
115 VAC 60 Hz
C2 1000 µF 25V
-15
R3 1K TRIM
A3
R1, R2, R3-WESCOM RV-1K
C1, C2-WESCOM U25-1000
A1, A2, A3-MOTOROLA MFC 6030
ALL RESISTORS-1/4W, 5%
T1-See Text
MFC 6030
TOP VIEW

Fig. 10. (A) Foil side of power supply pc board. (B) location of parts on component side.

making the use of small IC voltage regulators ideal. Motorola MFC6030 (plastic) regulators are used in each supply and are overload protected against accidental short circuit. Varo type VE18 molded bridge rectifiers, together with 1000 μF filter capacitors, supply the dc input to the regulators. The +15V and +5V regulators are both fed from the same source. The power transformer was a surplus unit with a 30V CT secondary. The center tap was uncovered and the two leads separated so as to provide two independent 15V windings. A small trimmer is provided in each supply for voltage adjustment.

The pc board for the power supply is 2.6 x 4.1 in. All components but the transformer are mounted on the board. The transformer is mounted directly on the chassis. Relative placement of the various components making up the complete generator can be seen quite clearly in the photo.

Control Circuits

Figure 8 is the wiring diagram for the control circuitry. Note that for CW operation of the generator, a dc voltage is applied to the varactor by way of the Start pot. In order for the dial calibration to be valid in either mode, the dc level applied to the control must be exactly the same as the peak amplitude of the ramp during swept operation. As will be explained in the set-up procedure, the ramp will vary from a starting point of approximately +1.4V and peak at +12V. Resistor R1 in series with the +15V line is selected for a +12V level at the high end of the pot. R2 is selected for a level of +1.4V at the low end. The reason for the offset of 1.4V at the low end is because the pin to which the tuning diode is connected at the MC1648, sits at this level. The source of this bias is the drop developed across two forward-biased junctions within the IC and not from any external source. Once these dc levels have been set for CW with the fixed resistors, the ramp can be made to match by means of the trimmers in the ramp generator circuit.

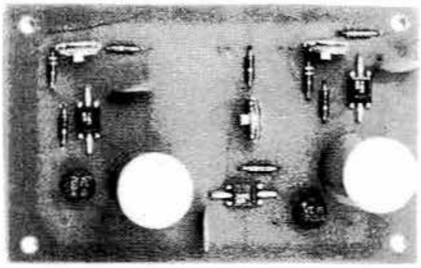

The power supply pc assembly.

The rest of the control circuitry is quite straightforward. In position 2 (Swept 1) of the mode switch, positive pulses from the ramp generator are fed to the band switch during the sweep period. When retrace occurs, the pulse drops to zero and the oscillator shuts off until the next sweep starts. In the CW and Swept 2 positions of the mode switch, a steady +15V is applied to the band switch and the oscillator runs continuously.

The final function performed by the mode switch is to route the dc for CW or the ramp for swept operation to the tuning diode. The two fixed resistors are mounted point-to-point behind the panel since all points are readily accessible within every short distances. The fixed resistors associated with the marker gain control are also mounted the same way behind the panel. The switches are ordinary rotary types and the potentiometers are all ordinary carbon controls. The Start and Stop controls as well as all gain controls have linear tapers. The Sweep Time pot is an audio taper type with built-in switch for ac power.

Setting Up the Ramp Generator

For best results, a calibrated dc scope is required to properly adjust the ramp generator. Test points have been provided on the pc board and short pieces of bare wire connected to each of these points make excellent tie points for the scope probe. The

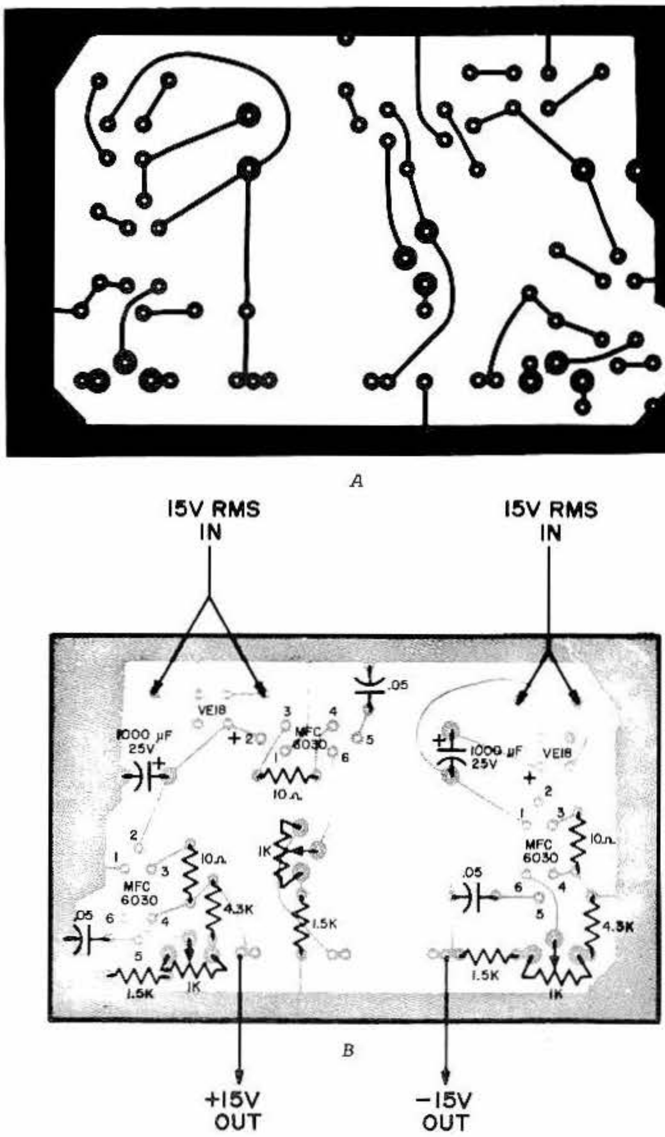
A
15V RMS IN
15V RMS IN
VEI8
1000 µF 25V
+
2
3
4
MFC 6030
5
6
.05
10Ω
IK
1.5K
4.3K
1
IK
B
+15V OUT
−15V OUT

Fig. 11. (A) Foil side of power supply PC board. (B) Location of parts on component side.

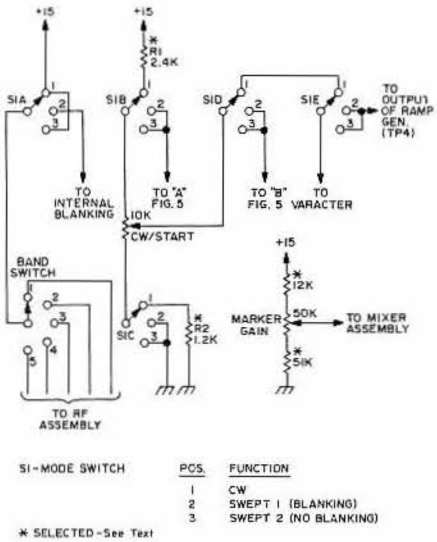
+15
+15
*
R1
2.4K
S1A
S1B
S1D
S1E
1
2
3
TO OUTPUT OF RAMP GEN. (TP4)
TO INTERNAL BLANKING
TO "A" FIG. 5
TO "B" FIG. 5
TO VARACTER
10K
CW/START
BAND SWITCH
4
5
S1C
R2
1.2K
+15
12K
MARKER GAIN
50K
TO MIXER ASSEMBLY
51K
TO RF ASSEMBLY
S1-MODE SWITCH
POS.
FUNCTION
1 CW
2 SWEPT 1 (BLANKING)
3 SWEPT 2 (NO BLANKING)
* SELECTED - See Text

Fig.12 Control circuit wiring schematic.

circuit may be aligned either before or after installation. The Start, Stop and Sweep Time pots can be connected at the ends of their respective leads if the unit is checked outside the cabinet.

Set all seven trimmers to mid-range. Connect the scope to TP1 and adjust the horizontal for a full sweep of 20 ms. Set the Sweep Time control to minimum resistance and apply power. Some sort of triangular wave should be displayed. Adjust R1 for a 3V p-p amplitude of the waveform. Turn R2 cw until the negative-going portion of the triangle is 20 ms long. Since R1 and R2 interact, you will have to stop occasionally and reset R1 for proper amplitude. Once the ramp is set at 20 ms with the Sweep Time pot at minimum, the slow speed end will automatically be about 5 or 6 seconds with a 1 MΩ pot.

Transfer the scope probe to TP2 where a positive-going ramp should be seen. Set the starting point of the ramp to zero volts by means of R3. Amplitude of this ramp must be 10V. If it is not, go back and adjust R1 slightly until it is. If necessary, reset R2 for 20 ms trace length. As soon as a 10V, 20 ms, zero-based, positive-going ramp has been achieved at TP2, go on to TP3.

At TP3 there should be a negative-going ramp. Once again we require a zero base line or starting point. Adjust R5 to accomplish this. Amplitude should once again be 10V and is controlled by R4. Once you have a 10V, zero-based, negative-going ramp, move on to TP4.

Connect the probe to TP4 and set the Stop control to full cw. The Start pot should be at minimum setting. A positive-going ramp should be present at TP4. By means of trimmer R7, move the ramp up or down until the starting point is approximately +1.4V. Ordinarily the ramp will not go below zero but will flatten out as the positioning control is adjusted. This is due to clipping in the output amplifier. If the ramp generator is going to be used with another type of VCO, the ramp can be set to exactly zero if desired. In our case we can't as explained previously.

Because a substantial amount of capacitance change occurs in the tuning diode up to 12V, the final ramp amplitude is set to 12V by means of R6. Because R6 stretches the signal in a negative direction, R7 will have to be used to reset the base line at 1.4V. Other than for checking tracking between the CW and swept modes, this completes the alignment procedure for the ramp generator.

When the generator is finished, you can check the tracking by observing actual signals at the input to the VCO. Connect the scope probe to the feedthrough capacitor on the VCO assembly that goes to the tuning diode and set the Start control to maximum. Switch rapidly between the CW and Swept 1 positions of the mode switch. The CW input will be a straight line sitting at +12V while the swept input will be a ramp peaking at +12V. Next, turn the Start control to minimum and again compare levels. If the Stop control is also at minimum, both signals will be straight lines at +1.4V. If there is any substantial difference between the two modes, use R6 and R7 on the generator board to make the ramp agree with the dc levels. That completes all phases of the ramp set-up procedure.

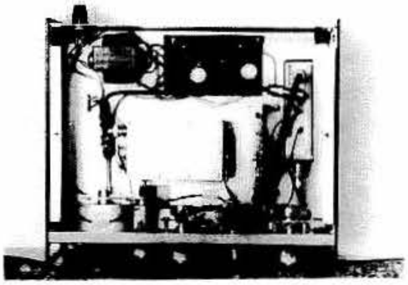

This bird's eye view of the interior shows the location of all major assemblies.

Construction

The cabinet used here is one manufactured by Sorensen Electronics in their Mod-U-Line series. These are the most reasonably priced instrument enclosures I've come across so far and I've used them for several projects. This one is a Model MCH-5129 with a CP-129 chassis plate. Dimensions are: 5¼ in. high by 12 in. wide by 9 in. deep.

The 5-step attenuator was picked up surplus from Fertik's Electronics for about \$10. It is well made and designed for 50Ω systems operating up to 1 GHz. It has an integral female BNC output connector on the front face along with four threaded mounting holes for ease of installation. The input connector is a BNC male at the end of a short piece of coax.

All rf assemblies are interconnected inside the cabinet by means of coax cables. This is clearly evident in the photo.

The Start and Stop pots were mounted on brackets behind the front panel so that Jackson Brothers type 4511/DAF reduction drives could be installed for easier tuning. The two circular dials are slightly under 2 in. in diameter and were cut from sheet plastic. While operating in the CW mode, one of these was calibrated in pencil to provide a pattern for the finished product. A master was laid out using Rubylith® film and rub-on numbers. A negative of this was then made using 3M reversing film. From the negative a finished set of dials was printed on aluminum material of the presensitized variety. These were cut out and stuck to the plastic by means of their own pressure-sensitive adhesive backing. The nameplate was made from the same material. The two index pointers are clear plastic with hairlines scored on the inside surface. They are mounted on spacers directly over each dial. All remaining labeling was done with tub-on lettering.

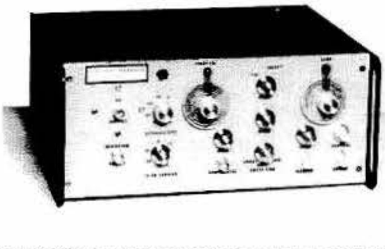

The finished sweep generator makes a professional appearance.

Vendor Addresses:

Fertik's Electronics, 9th & Tioga Sts., Philadelphia PA 19140.

Sorensen Electronics Co., Inc., 418 Queens Lane, San Jose CA 95112.

Wescom, P.O. Box 2436, El Cajon CA 92021.

100 KHZ THIN-LINE PULSE GENERATOR

James Ashe W2DXH

Ordinary 100-kHz frequency standards are usually audible up to a few tens of megahertz. A good one might be usable at 50 MHz. The circuit described here uses a dual NAND gate to generate a 100 kHz signal whose harmonics are usable to 432 MHz or higher. And it can be built without benefit of special instruments and knowledge.

The thin line pulse

One rather surprising result of higher mathematics is that all repetitive signals are composed of harmonically related sine and cosine waves. For example, the familiar square wave is composed of a fundamental frequency, which sets its basic repetition rate, and of odd harmonics only of its fundamental, which contribute to its square corners. If the harmonics' amplitude or phase relationship is upset, the square wave is distorted. This feature makes the square wave very useful for amplifier testing, but its harmonic content is not very good for frequency standard applications.

Now suppose that we start adding up signals of F, 2F, 3F, and so on, phased in so that they all reinforce each other once per cycle. Let's say thay are all the same amplitude. What would we get? See **Fig. 1A.**

The five equal amplitude sine waves peak simultaneously at the beginning of the fundamental's cycle. Everywhere else, until near the end of the cycle, they are more or less out of phase. Trying to see what will happen, we try adding the first two frequencies. **Fig. 1B**, the result, might suggest something to a mathematician.

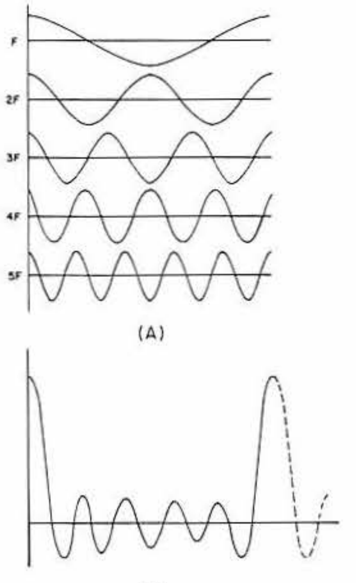

Fig. 1. Five sine waves (A) and the waveform as a result of point-by-point addition (B).

As the number of frequencies is increased, their amplitudes tend to average to zero everywhere except at the beginning of the cycle. Here, they all add up to a short, sharp pulse. It follows that a short, repetitive, one-sided pulse should contain odd and even multiples of the fundamental frequency.

An ideal thin line pulse has infinite frequency content.* No real signal could meet

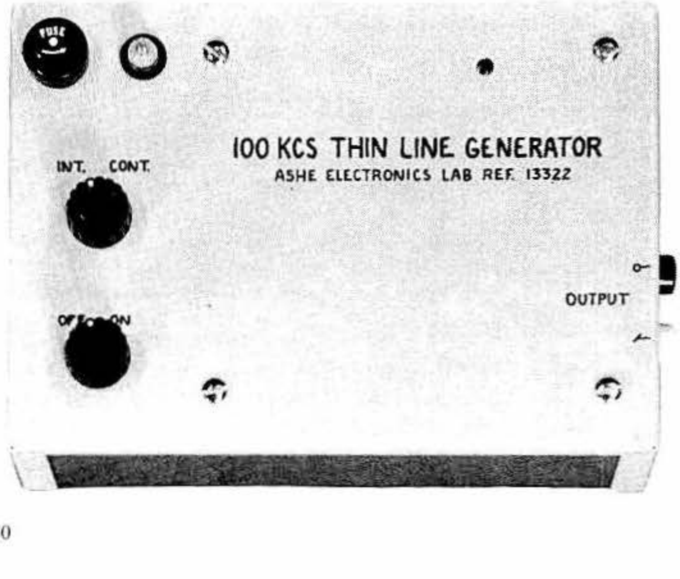

External view of the 100 Hz thin-line generator.

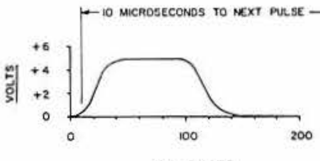

Fig. 2. Real circuit output as seen by a Tektronix 545A oscilloscope. A faster scope shows shorter rise-time and sharper corners.

this spec, but a fast digital IC can produce a very workable approximation. Fig. 2 shows a Tektronix 545A view of the generator output and tests with other scopes indicate the real pulse has better rise time and sharper corners than shown here.

If this pulse is viewed on a low-performance service variety scope, its appearance will be greatly changed. There will be an apparent loss in amplitude, since the pulse occurs and terminates before the slow circuitry can properly respond. The apparent duration is increased, also because of the slower viewing circuitry. And the fast pulse may excite circuit resonances, so that the thin line pulse appears as a damped oscillation. But these problems do not interfere with constructing the generator, because the very simple NAND gate circuitry contains no critical elements or adjustments.

How it works

There are four circuit sections, shown in Fig. 3. A 100-kHz crystal-stabilized oscillator sets the basic frequency, and a dual NAND gate circuit converts the oscillator output to a thin line pulse. A 1-Hz astable generates the output marking signal. A 6 volt dc power source is provided by a voltage doubler zener-regulated supply.

Multivibrator oscillators are not ordinarily very stable frequency sources. But if the oscillator is designed to run slightly below required frequency, and an appropriate crystal is connected between transistor base terminals, oscillations are stabilized at the crystal frequency.

The crystal does not change the multivibrator's style of operation. It synchronizes the astable to its own frequency, by triggering the OFF transistor into conduction shortly before normal RC turn-on. The output is a squarish wave with good fall time, but a long rise time as shown in **Fig. 4A.**

In passing through the first NAND gate the pulse is squared up and becomes slightly unsymmetrical. See **Fig. 4B.** A differentiating network, C7 and R11, converts the square wave into the pulses shown in **Fig. 4C.** These pulses, applied to the second NAND gate, reappear as the thin line pulses shown in **Fig. 4D.**

Since one CW signal sounds just like another and there may be several in the vicinity of a check point, a marker feature is required. This is provided by the 1-Hz astable, which paralyzes the second NAND gate part of the time. Its base bias resistors are unequal, giving a distinctive duty cycle to the output signal. A switch disables the astable if a continuous signal is required. **Fig. 5** shows the output when the second astable is operating: the output is locked in the up condition during half of each 1-Hz astable cycle.

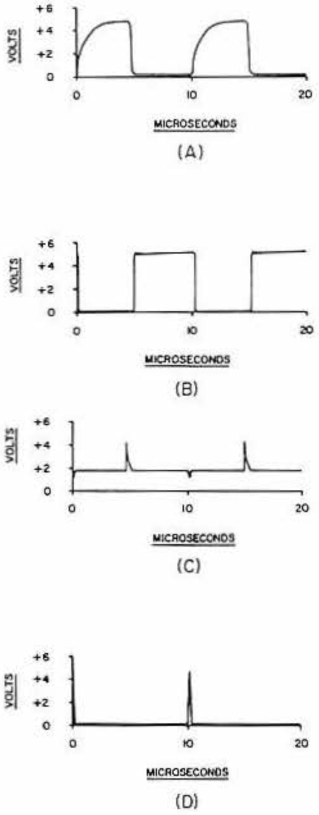

(A)

(B)

(C)

(D)

Fig. 4. Signals at four critical points in the generator, as displayed on a Tektronix 545A oscilloscope. They are shown in time coincidence.

Sometimes an astable oscillator will refuse to start oscillating when it is turned on. It does not start because both transistors are in saturation. This reduces loop gain so that available noise cannot be amplified around the loop. It would never start without some strong, outside interference.

A pair of diodes, D1 and D2, provide a reliable remedy. The diodes are arranged so that base bias must come from whichever collector is at the higher voltage. If both transistors are in saturation, their collectors are at perhaps 1 volt, which cannot provide enough base current to keep the transistors in saturation. This contradictory situation does not arise in the real circuit, which starts reliably.

Additional diodes, D5 through D8, appear in the base circuit of the 1-Hz astable. These are protective diodes. The collector swing at turnoff of about 5 volts is conveyed powerfully to the opposite base through the large coupling capacitors C5 and C6. The reverse B-E breakdown voltage of these transistors is not known, so the diodes are provided to prevent the turnoff voltage exceeding 2 volts or so.

DC power for the Generator circuitry comes from a voltage doubler supply based on a low-current filament transformer. Its design is conventional, but a large capacitor, C12, is provided across its output to mini-

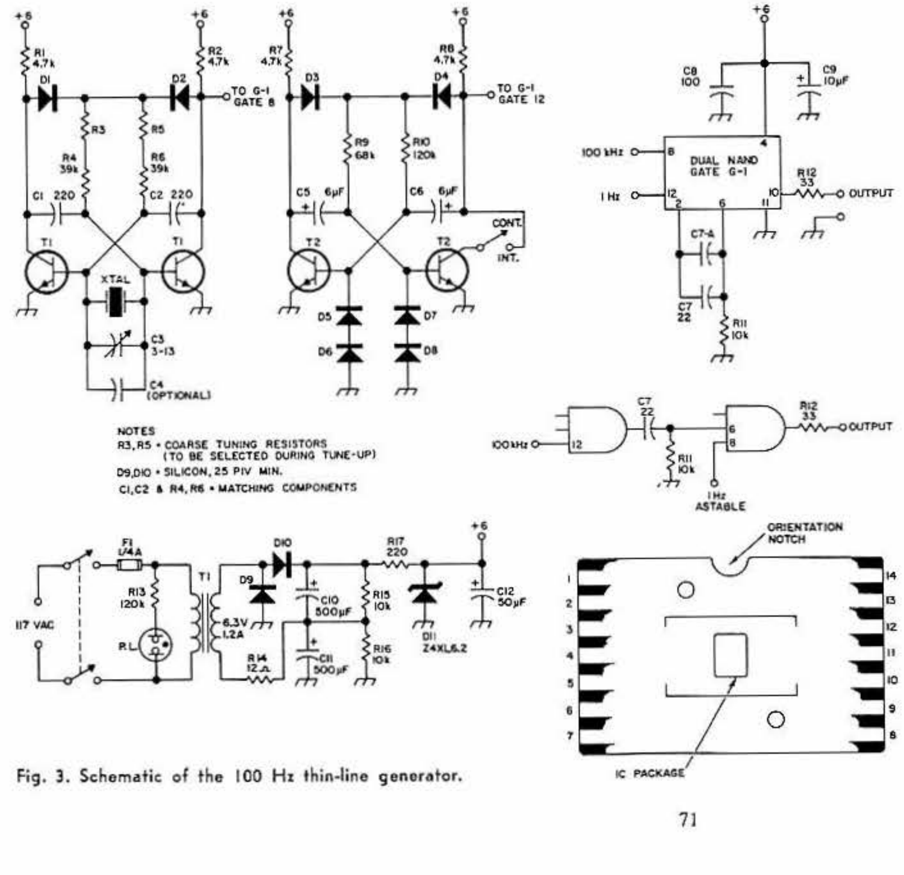

NOTES
R3, R5 • COARSE TUNING RESISTORS (TO BE SELECTED DURING TUNE-UP)
D9, D10 • SILICON, 25 PIV MIN.
C1, C2 & R4, R6 • MATCHING COMPONENTS

Fig. 3. Schematic of the 100 Hz thin-line generator.

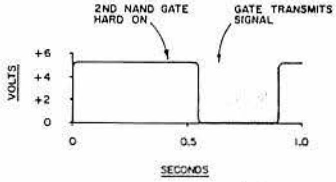

Fig. 5. The second NAND gate locks in its up position part of the time to produce an intermittent output.

mize noise on the supply line. The supply could be replaced with some batteries, shunted by a 50 μF or larger capacitor to absorb transients. The original breadboard ran very well, powered by four flashlight batteries.

Construction

The generator is built in a Premier #PMC 1008 3x5x7 inch heavy aluminum box. Its top cover was refinished in light green enamel, and four ⅜ inch grommets in the bottom piece serve as protective feet.

Inside the box, the 6.3-volt transformer and cheater cord connector are mounted on the left-hand wall. A pilot lamp, fuse, and two switches are mounted on the horizontal panel, at the extreme left. This leaves just enough open space for the two circuit boards which occupy most of the box. Two banana jack output connectors are placed on the right-hand side, just below the panel.

The circuit boards are cut to 4½ x 5 inches, from Vector 3/32 inch pattern A stock and mounted parallel to the panel. The upper board is spaced an inch from the panel, and carries both astable oscillator circuits. The other board is mounted one half inch below, and carries the digital IC and the power supply circuitry. Assembled, the two boards make a sandwich with wiring sides together.

Both boards are mounted on the same four centers. These are through the second hole diagonally inward from each corner. The 1 inch 6-32 internally threaded spacers are modified by adding a short length of 6-32 threaded shaft to one end of each, simplifying assembly.

Component assembly on the boards is largely a matter of plugging in Vector T9.4 lugs. The finished product looks much better if some thought is given to facing the lugs in one of two directions. Mounting and transistor holes should be drilled and reamed to size before installing lugs.

The general arrangement puts all wiring on one side of the board and practically all components on the other side. This approach seems a little inflexible but is straightforward and looks good.

Possible board orientation problems may be overcome by working out a handling and wiring procedure that doesn't require constant reference to actual components. A good approach assumes that the board is only turned over an imaginary hinge at its bottom edge, so that top down when one side is up becomes bottom up when the other side is down. This preserves left-right relationships. Another useful convention is that all supply wiring goes to left-hand end of components.

Wiring is carried out one network (plus supply lines; ground lines; interstage lines, etc.) at a time, with prearranged color coding. Bare wire goes for short runs and where there is no chance of a short. Solder each lug when convenient. #22 solid wire fits the T9.4 lugs well, but flexible stranded wire is used for the four lines from one board to the other.

Transistors precede other components into the board, because they are convenient position markers. They are placed in their mounting holes in the board from the component side, and their leads brought to the T9.4 lugs. Then the other components are mounted on the boards. Diode and electrolytic capacitor mounting polarity should be double checked. The T9.4 lugs may need a little bending before they will take a good grip on the components, but no component soldering is done until everything is installed.

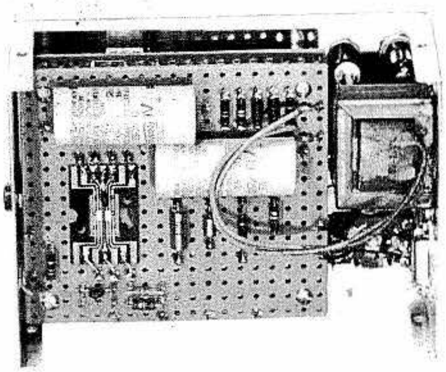

Inside the assembled thin-line generator showing the component side of the power supply and IC board.

Trimmer capacitor C3 is mounted on its tabs just under the top panel. Then a small screwdriver access hole is drilled over it in the panel, before painting, for vernier frequency adjustment after final assembly.

Certain components are matched before installation. An ohmmeter and a capacitor checker will do a satisfactory job of selecting C1 and C2, and R4 and R6, for equal values. These components are chosen alike for best symmetry of the 100-kHz oscillator operation. It might be good planning to leave these components unsoldered until tuning is completed, but everything else can be soldered to the board at this point. Note that the R3 and R5 sites do not get resistors until later.

Two optional capacitor sites are included. These are for C4, an additional and probably unnecessary padder across the crystal; and C7A, which can be added to increase the width of the thin line pulse.

Apparently, the digital IC comes in a specially designed package for testing before use. To mount the IC, solder a ⅜ inch piece of #22 wire in each of the T9.4 lugs carrying supply and signal voltages to the IC. Place the IC between the two rows of lugs, bend the wires against the proper terminals, and solder. No other mounting is required.

The original breadboard showed a lot of transient noise in its supply circuit. This originated from the IC, which was trying to get big chunks of current to manufacture pulses. Since the IC cannot deliver frequencies not available from the supply lines, very careful bypassing is indicated.

High-frequency bypassing consists of C9, a .01μF disc ceramic capacitor across the IC supply terminals on the wiring side of the board, and C10, a 100 picofarad capacitor soldered directly between supply terminals on the IC. The capacitor leads are provided with spaghetti insulation and placed for minimum open space between the capacitor leads and the IC's supply leads.

Testing before final assembly is very easy, because the odd appearing board layouts go together giving a structure that opens out like a book. The hinge is the four leads between boards. Leave transformer leads long, so that the circuit may be tested well free of its cabinet.

The upper half of the Premier box is prepared by a powerful cleaner which removes its original paint. After thorough removal of the cleaner, the metal is roughened with wet sandpaper, rinsed in vinegar solution and then clear water, leaving a very good surface that does not require priming for excellent paint adhesion. Watch out for greasy fingerprints.

Rustoleum #868 Green applied from a convenient spray can gives a fine finish. Follow instructions on the can. After drying, the fresh, clean enamel will take waterproof India ink, applied with a Leroy drafting pen. When the ink is thoroughly dry, a final coat of Rustoleum #717 Clear finishes the job. The enamel is soft at first, but hardens into a coat durable in normal lab use.

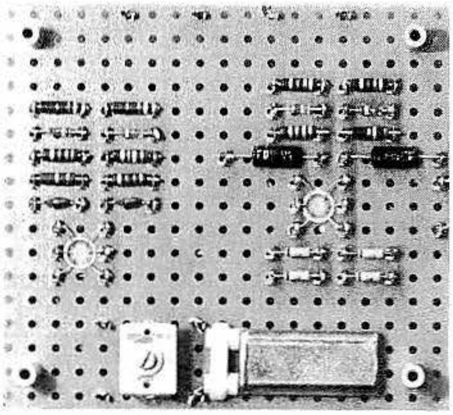

View of the component side of the astable oscillators board.

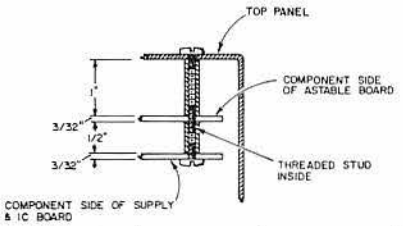
TOP PANEL
COMPONENT SIDE OF ASTABLE BOARD
1"
3/32"
1/2"
3/32"
THREADED STUD INSIDE
COMPONENT SIDE OF SUPPLY & IC BOARD

Fig. 6. Mounting dimensions and spacer assembly diagram.

Table of special parts

Crystal: 100 kHz parallel resonant 32 pF. shunt capacitance normally designed quartz crystal.

The following parts were obtained from Solid State Sales, P. O. Box 74, Somerville, Mass. 02143.

- T1 & T2: 2N2060 type dual NPN transistor
- D1, D2, D3, D4: fast point-contact Germanium diodes coded 1N59
- D5, D6, D7, D8: fast point-contact Silicon diodes marked S284GM
- G1: surplus digital integrated circuit Solid State Sales type G1. (comes with data sheet)

Tuning up

The generator should be zeroed to frequency before installation in its case. This is a two-step process. First, the 100-kHz astable base resistances are adjusted by choosing resistors for R3 and R5 to bring the oscillator frequency within trimmer range of 100 kHz, perhaps a few hundred cycles high at 15 MHz. Then the trimming capacitor brings the frequency to accurate coincidence with WWV.

To roughly zero the generator, set the trimmer capacitor, C3, at minimum capacitance. Identify WWV on a short-wave receiver, and tune around a bit to familiarize yourself with what's happening in the vicinity. It would be nice if things are fairly quiet.

Then put 4.7k resistors into the astable board at the R3 and R5 sites, turn on the generator, and look around for the signal. Depending upon actual values of C1 and C2, the signal may be on either side of WWV but is likely to be on the high side. If so, try again with resistors one size larger, which will lower the frequency. You should shortly find resistors that bring the frequency near enough to WWV for final zeroing with the capacitor. Verify tuning range on both sides of WWV.

Correct values for R3 and R5 may be approximated very quickly if a good triggered scope is available. Try selecting resistors for a period of 11.4 microseconds with the crystal removed.

Using the thin line generator

A breadboard test showed that (as might have been expected) there should be some way to distinguish generator signals from other CW signals. The continuous/intermittent feature provides the marking, and once the correct signal is located the generator can be switched to "continuous" for accurate work.

At low frequencies, the generator output and behavior resembles a conventional 100-kHz standard. The signal simply is not as strong. A greater difference appears at higher frequencies: the original model yields an audible beat note at 80-MHz from a diode mixer through an inexpensive audio amplifier. And another test shows a usable signal at 432 MHz: the 4,320th harmonic.

Some connection to the receiver or other detector is required. This is a natural consequence of a circuit design that puts the signal where it belongs, rather than spraying it all over the lab. A few picofarads coupling capacitance is sufficient at all frequencies.

Perhaps this circuit can be used for purposes other than a frequency standard. Its moderate amplitude but wideband output should be ideal for detecting changes in receiver sensitivity over a broad tuning range. In fact, with a little decoupling of the input leads and provision of a coax output connector the generator should do well as a stable, reliable small-signal source. A piece of adjustable waveguide-below-cutoff would make an excellent attenuator for work not requiring exact measurements. Another thought that occurs is possible further development by provision of some arrangement for detecting which harmonic is actually being heard.

Chapter VI

Crystal Calibrators

ALL BAND BAND-EDGE MARKER

Charles Berner WA2HRZ

This calibrator is complete with its own ac supply, eliminating the need for taking power from the receiver. If ac outlets are at a premium at your shack, the on-off switch and the line plug can be eliminated and the ac line connected across the receivers ac input so that the calibrator comes on whenever the receiver is turned on. A high density selenium rectifier was used and this contributes greatly to the compactness of the unit. The whole thing runs very cool and even after 24 hours operation, still isn't hot.

Wiring isn't too critical, but keep all leads short and direct. Using an octal socket for a xtal holder lets you use the unused pins as tiepoints and is recommended.

A Petersen type Z-2 xtal was used because it has a tolerance of .002%. This is one reason why no provision is made for "zeroing" the calibrator against a standard. The other is that there isn't any such standard as WWV for use. However, if this provision is desired, C1, the 56 mmfd capacitor, connected to pin 1 of the 6BH6 thru the 18K resistor, can be made variable. Also, if band edge markers are wanted for 50 mc and up, a 5 mc xtal may be substituted when this function is wanted and C1 made variable for "zeroing" against WWV.

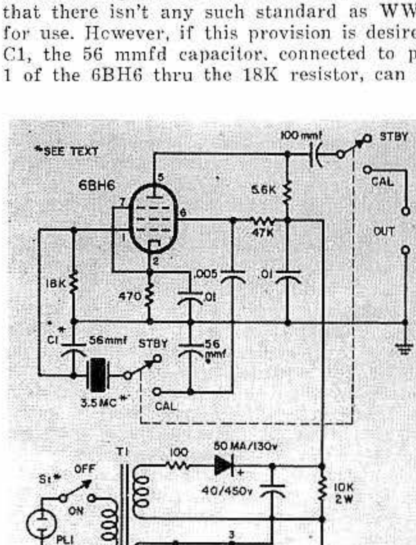

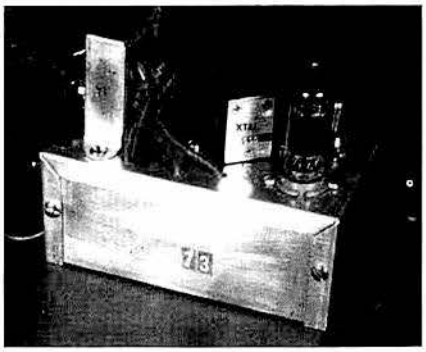

This gadget can also be used as a xtal activity checker by simply replacing the marker xtal with the xtal to be tested. Tuning your receiver should get you a strong signal at the xtal's fundamental if it is OK.

Table 1:

Band	Harmonic 3.5 mc	100 kc	Marker Frequency
80 meters	1	35	3.5 mc
40 "	2	70	7.0 mc
20 "	4	140	14.0 mc
15 "	6	210	21.0 mc
10 "	8	280	28.0 mc
6 "	15	525	52.5 mc
2 "	42	1470	147.0 mc
1¼ "	63	2205	220.5 mc
1¼ "	64	2240	224.0 mc

THE MULTICAL

E. R. Davisson K9VXL

What is the "Multical"? As the name implies, "multi" would suggest several uses, and "cal" might infer a calibrator of some sort.

Well, that's right, but there is slightly more significance to the name. "Multi" is also a short form term used to describe flip-flop circuits known as multivibrators.

By combining the basic characteristics of a free-running multivibrator (astable) with crystal control, you have a simple, stable, virtually-insensitive-to-temperature-changes, crystal calibrator for that receiver you have been wondering about.

The circuit uses no inductors and depends upon the crystal for the proper feedback for oscillations. Temperature stability is partially due to the absence of capacitors.

Transistor stage Q_2 operates with unity gain, whereas transistor Q_1 operates at considerably more gain. Both stages are operating as feedback amplifiers. The harmonic generator diode D_1 is a 1N128. Any general purpose diode may be used.

By using the multivibrator circuit, the waveform obtained is comparatively rich in harmonics and could be used without any further refinements. However, to insure useful harmonics through 30 MHz starting from a 100 kHz crystal, a harmonic generator consisting of R_6 and D_1 shown in **Fig. 1** was added. The capacitors C_1 and C_2 are used strictly for coupling and have no effect on frequency stability.

Crystals from 100 kHz up to 1 MHz may be used in the Multical with no changes. The circuit will oscillate from voltages as low as 2 volts and can be operated safely from voltages as high as 20 volts. This wide range of voltage operation allows the source to be obtained from virtually any place.

Output from the calibrator may be fed directly into the receiver's input, or may be coupled to a short whip antenna. With a whip antenna, close coupling to the receiver's input

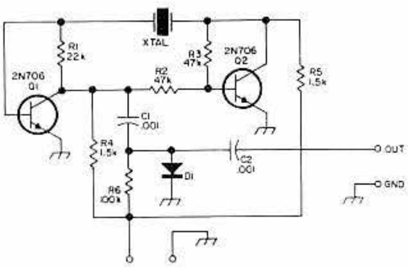

Fig. 1. Schematic of the Multical.

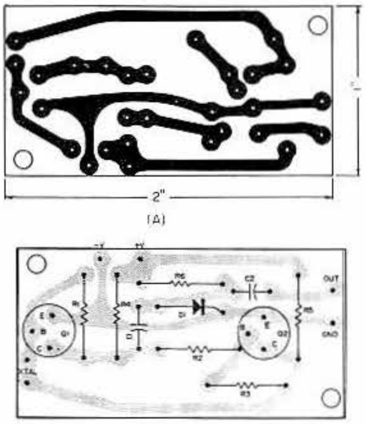

Fig. 2. Suggested printed circuit board layout for the Multical. A gives the copper side, B the component side. A board for the Multical is available for $1 from the Harris Company, 56 E. Main Street, Torrington, Conn.

may be required at higher frequencies. (Especially at the lower voltage levels.)

For the more ambitious builders, **Fig. 2** shows the printed circuit board layout for the Multical. Due to its small physical size (1" x 2"), room can probably be found even in the most compact of receivers. **Fig. 2A** shows the foil side, and **2B** shows the parts placement.

So the next time you wonder about the accuracy of your receiver calibration, give this simple circuit a try and you'll know for sure.

100 KHZ MARKER GENERATOR

W. W. Davey W7CJB

This useful piece of equipment generates usable harmonics from 100 kHz to 225 MHz. It is completely self contained and portable which makes it convenient not only to use in the ham shack, but also in the mobile unit or at a field day location. Its use lies mainly in accurately spotting band edges and 100 kHz calibration points throughout the ham bands.

Most modern day home receivers are equipped with calibrators, but these calibrators are of little use when needed to spot frequencies on VHF and UHF converters or portable equipment.

The generator is constructed in a 2¼ x 4¾ x 1½ inch handi-box. The parts are mounted on a vector board, and the entire unit is powered by one #216 nine volt battery or its equivalent.

Hints on construction First obtain some vector board. The piece I used was cut from the board supplied in a "GE experimenters aid hobbyist kit." The board must be cut to size before construction and will measure 3½ x 2 inches. This will allow room for the 9 volt battery in the end of the handi-box. Make sure the newly cut vector board will fit inside the handi-box before you start mounting parts. It might save a lot of trimming at a later date.

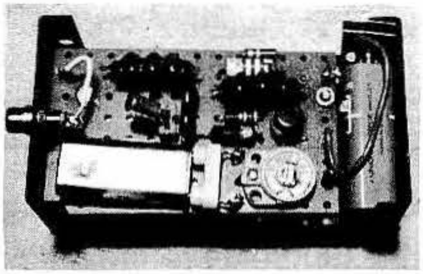

The parts layout is not critical. Components may be arranged as shown in the photo or in any other arrangement suitable to the components you may be using. I used sockets for the transistors, as I wanted to be able to experiment to see which transistors from my junk box would give the most output in the VHF and UHF bands. I ended up using the 2N404 for the oscillator and a 2N384 in the multiplier stage. I also found that Japanese 2SA83 transistors which had been removed from the *if* stages of a junked transistor radio, would work equally well in both sockets.

All components are mounted on the top of the Vector board with the exception of C4 and C5. For the most part, wiring can be completed with existing leads on components. The push-in terminals furnished with the GE experimenters kit were used for the battery connections, antenna connection and for mounting the crystal socket. The circuit board can be mounted to the handi-box with three 11/16 inch bushings. This leaves room for a slide switch to be mounted on the cover of the handi-box. Two of these bushings were purposely placed at the end of the board to form a sort of socket to hold the 9 Volt battery. The antenna output connector which is mounted on the handi-box is a switchcraft #3501FP phono jack. A small hole may be drilled in the bottom of the handi-box through which a screwdriver may be inserted for adjusting C1. For extreme accuracy C1 is adjusted to zero beat with WWV.

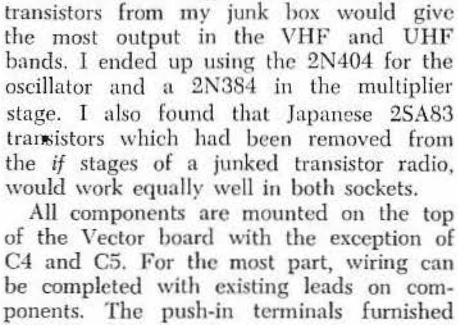

A 36 inch piece of insulated wire soldered into a phono plug may be inserted into the phono jack and used as a test antenna. The intensity of the markers may be varied by moving this test antenna near your receiver antenna lead-in. If you are using a coax lead-in you can couple by drilling a small hole in your coax relay so that the test antenna can be inserted near the relay armature. You will find that most SWR meters provide an easy method of coupling to the center conductor of the coax. As a last resort you can always couple to the receiver or converter antenna coil.

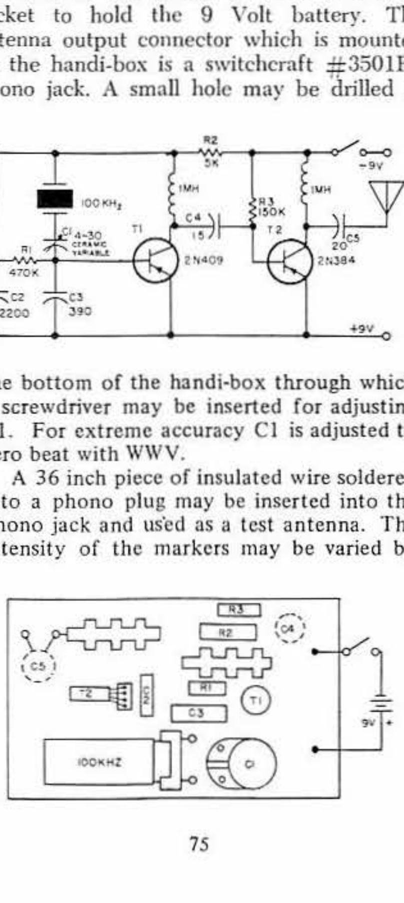

I have made very good use of this gadget to spot frequencies in the 144 and 220 MHz bands. It was well worth the time and effort it took to build it.

ALL BAND FREQUENCY MARKER

Kenneth W. Robbins W1KNI

Crystal controlled marker generators are useful adjuncts in any frequency determining situation requiring high accuracy, such as locating band edges, sub-bands and calibrating receivers. If you've been entertaining thoughts about construction of one, a version is described here which uses the new C/MOS integrated circuits powered by a 9 volt transistor radio battery. And instead of the usual rotary harmonic selector switch, a multi-pin IC connector strip and three test plugs serve as a miniature patch panel to enable various divisions of the reference crystal, with a maximum countdown of 256. "Rocks" from 100 kHz to 4 MHz oscillate readily in this circuit. In this model an FT241 xtal set to 400,000 Hz has been chosen for control and has usable receiver calibration divisions down to 2.5 kHz. The harmonic spectrum extends to at least 160 MHz, the tuning limit of a transistor superregen used in testing. When used in densely occupied HF bands, an AM beeper can be switched on as an identification aid.

Divide by	Output, kHz
1	400
2	200
4	100
8	50
10	40
16	25
20	20
40	10
80	5
100	4
160	2.5

Table 1.

Referring to Fig. 1, one third of a hex inverter makes up a crystal controlled oscillator and buffer, another third is a slow rate pulser and the two remaining units function in the dividing section. These are all standard circuits described in RCA's COS/MOS Data Book #SSD-203. An emitter follower minimizes loading on the IC outputs, speeds up rise time to increase harmonic content, and provides a low impedance output. The AM

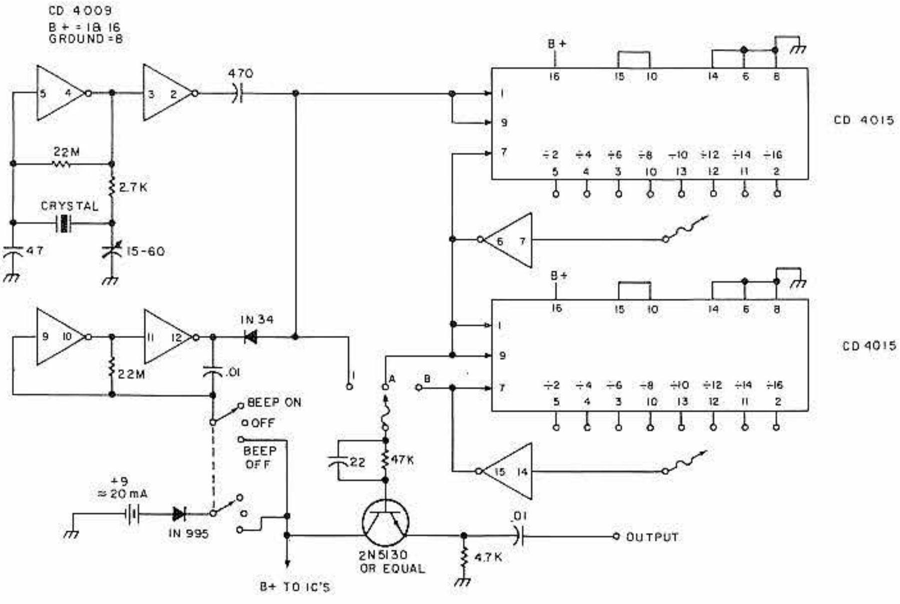

Fig. 1. Schematic.

beeper is a simple clamp that gates rf on or off to following stages.

Photo 1 shows all components mounted on Vector P pattern perf-board that fits inside a Bud minibox. Sleeving 3/8" (10mm) long is slipped over the wire trap terminals of the contact strip to space it up from the board. A DPDT center-off miniature toggle switch acts as one board to panel spacer. Diagonally across from it, a 4-40 threaded rod conducts emitter follower output up through the front panel via a 1/2" (13mm) insulating spacer and plastic shoulder washers. Two regular 4-40 screws and spacers complete the four corner mounting. This spacing allows the contact strip to project partly through a panel cutout so that it is mechanically secure without fastening.

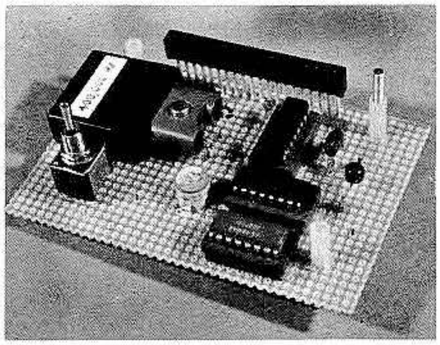

Photo 1.

Photo 2 (completed assembly) shows a stick-on label with patching connection call-outs for various division ratios. If only one crystal is employed, labeling could indicate most used frequencies instead. A typical frequency vs division listing for this model is shown in Table 1. You can easily make up a complete table of all possible ratios, remembering that each CD4015 shift register divides by even numbers ONLY, starting at 2 and ending at 16.

Uses to which a marker generator may be put have been described before: i-f alignment, BFO, scope linearity, etc. A type that divides down to the audio range like this one is especially useful in checking superhets. A very broad and flat spectrum of overlapping signals is generated and an audio tone will be heard no matter where the set is tuned. If its tracking and sensitivity are top-notch, the S meter will hold steady over the tuning range. Tracking adjustment amounts to tweaking for maximum meter reading or loudest audio tone. Then patch for 100 KHz markers and check calibration. It's a lot faster and easier than using a conventional signal generator.

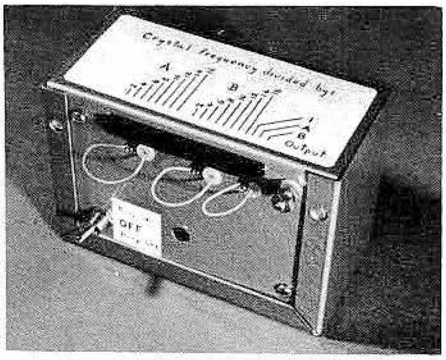

Photo 2.

1	CD4009	1	15/60pF trimmer
2	CD4015	1	.01 uF
1	1N34	1	xtal; see text
1	1N995	1	xtal socket
1	2N5130 or equiv.	1	Alco #MST205P switch
1	2.7k	1	Bud #CU-2115HG minibox
1	4.7k	1	Vector #44P29-062 perfboard
1	47k	1	#216 battery
2	22M	1	Battery connector
1	22pF	3	16 pin IC sockets
1	47pF	3	Augat patch pins or equiv.
1	100pF	1	20 pin contact strip SAE #Series 7000
1	470pF		

Table 2. Parts List.

POOR MAN'S UNIVERSAL FREQUENCY GENERATOR

John Schultz W2EEY

As precise frequency control and measurement becomes more and more a part of the amateur radio game, the need develops for test instruments that deliver a wide range of both rf and af signals of high accuracy. It would be ideal if everyone could have a frequency counter and a synthesizer type rf and af generator but that is hardly the case. Most amateurs must utilize their basic station gear along with selected accessory items to test out and adjust equipment. This article describes a very useful accessory item that for a modest cost goes a long way toward having some of the expensive test equipment just mentioned. The item to be described is somewhat like a grid-dip meter in that it is basically a simple type of oscillator but as one gets to know and use it, new uses for it are found and its versatility constantly expands.

Circuit Description

Figure 1 shows the circuit diagram of the test generator. Basically, it consists of a string of SN7490 decade counters which are used to divide down a selected input signal by a factor of 10 or 2. The input signal can come from a 1 MHz master oscillator, a special crystal oscillator for externally used crystals or from any external sine-wave source. The special crystal oscillator which uses a SN7400 will operate with almost any basic or overtone crystal in the hf range. It can be used for crystals in the low frequency and lower VHF range also by a simple modification. One gate of the SN7400 crystal oscillator is used to drive a LED which will indicate that the crystal is oscillating so it serves as a crystal activity indicator as well. When an external sine-wave source is used, it is first coupled through a SN74121 multivibrator. This stage squares off the sine wave so it can better drive the subsequent frequency divider chain.

The frequency divider chain is fixed, although one could easily switch the individual SN7490 units to divide by different ratios when desired. This should be obvious by noting the wiring of the divide by 2 SN7490 with that of the divide by 10 units. However, the variety of frequencies which can be generated then with different input sources becomes confusing and more than would normally be needed.

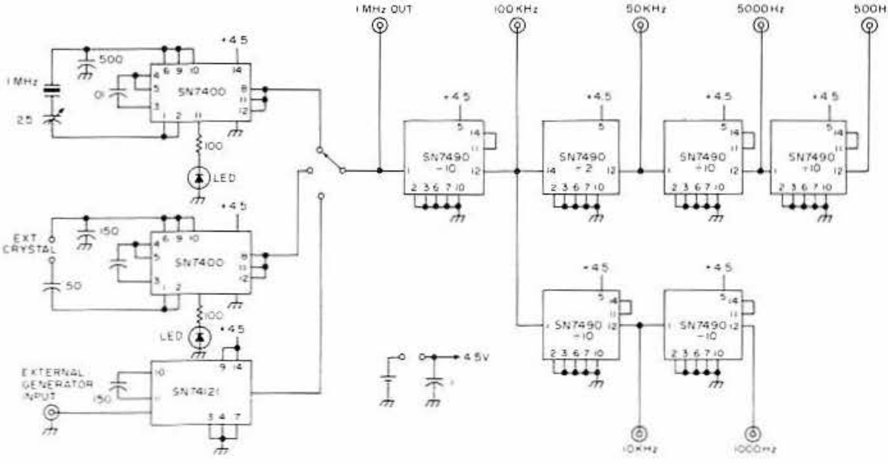

Fig. 1. Diagram of universal frequency generator. Output frequencies shown are for using 1 MHz oscillator.

The fixed divider chain follows the sequence: divide by 10, divide by 2, divide by 10, divide by 10. A separate branch after the first divide by 10 unit goes through two other divide by 10 stages. In the case of the divider chain being driven by the 1 MHz master oscillator, this results in the following output frequencies being simultaneously present: 1 MHz (basic oscillator output), 100 kHz, 50 kHz, 10 kHz, 5000 Hz, 1000 Hz and 500 Hz. With any other frequency input source you can easily calculate what frequency outputs the divider chain will bring in both the rf and af regions. Many surplus crystals will produce interesting frequencies of high stability in the af region that can be used for test purposes.

When using the special crystal oscillator, the LED will glow to indicate that oscillation is taking place. As shown with a 150 pF capacitor from one side of the crystal oscillator circuit to ground, the oscillator will work satisfactorily with hf crystals. Its range of oscillation can be extended to lf as well as high frequency overtone crystals by changing this capacitor. The value of capacitor required in picofarads is 500 divided by the frequency of the crystal in MHz. This value need, however, to be only approximate unless you require an absolutely square wave output from the unit.

When using the multivibrator input about a 1½ to 2V peak input, either sine-wave or approximate square wave is required.

Construction

The whole unit can be constructed on a piece of perforated board about 3 x 2 in. and made completely portable if powered by a 4½V battery (Burgess No.532) or just three D cells in series. This arrangement does not provide the absolutely best stability for the 1 MHz master oscillator but unless you intend to use the unit for marker frequency generation in the VHF range, it is a perfectly satisfactory arrangement. Alternatively, one could power the ICs from any standard 5.5V regulated supply used for IC digital circuitry.

I constructed my unit for battery powered operation and enclosed the unit in a small aluminum mini-box. The output of each divider was brought to a pin jack on the front panel of the unit.

One simple way to wire the relatively small number of ICs involved is to purchase perforated board which has hole spacing to fit standard DIP and preferably with a copper pad still left around each hole. The ICs are then placed on the board and the appropriate pins which either go to ground or to the 4.5V line bent in different directions. The ground line is run along one side of the IC and the 4.5V line along the other side and bare wire used to connect the appropriate pins to either line. Figure 2, illustrates the wiring for one of the divide by 10 ICs. When one starts this process on the board, it will be surprising how fast the wiring is completed. Individual insulated wire jumpers are used to make the input/output connections between ICs. The wiring is not critical and using a receiver to hear the markers, or an audio amplifier for the lower frequency outputs, one should be able to determine quickly if the circuit is working. The frequency of the 1 MHz master oscillator may be brought exactly on frequency using the 25 pF trimmer in the circuit and checking against WWV with a harmonic of the oscillator or by using a counter.

Applications

As I mentioned before, the applications that you can find for the generator really begin to unfold only after you have had it around the shack for awhile. Some of the applications would be:

1. A frequency marker generator for receiver calibration. The markers are usable up into the VHF range.

2. To extend the range of present rf or af signal generators into lower frequency ranges than they presently cover.

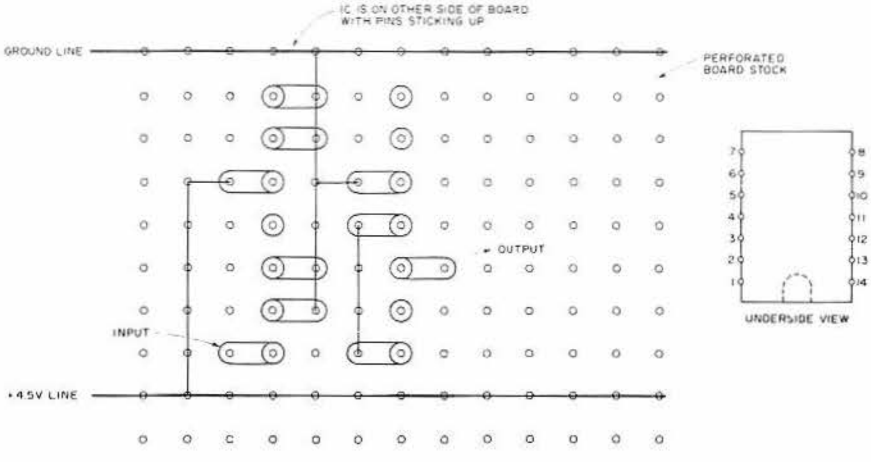

Fig. 2. Perforated board wiring of ICs. One SN 7490 divide by 10 unit is shown wired.

3. To perform stability checks on high frequency variable oscillators. The divider chain will always perform precisely and you can monitor the change in frequency of a higher frequency oscillator with a stable low frequency receiver.

4. A frequency generator to generate precise rf or af square wave signals at any frequency desired by choosing the proper crystal.

5. A crystal activity checker.

6. By taking two or more of the simultaneous outputs together via mixing diodes and a series tuned circuit resonant at the desired frequency, you can also mix the divider outputs to generate a variety of intermediate frequency outputs.

CALIBRATE THAT CALIBRATOR

Mitchel Katz W2KPE

Most modern receivers and transceivers in use today rely upon a 100 kHz crystal oscillator to calibrate the tuning dial. While some of the calibrators are built in, others come as outboard accessories. In any event the operation of each is the same.

By this time we all probably know what "zero beating" is. The 100 kHz oscillator in order to serve as a calibrator must be "zeroed" to some standard frequency such as the WWV carrier frequency on 5, 10, 15, 20, etc. MHz. With a CW or AM receiver, we can very easily tune through the zero beat point. On SSB receivers because one of the sidebands is missing we can only hear the one side as we approach zero. The other side of zero is greatly attenuated and may possibly not be heard at all. To further complicate matters for us in trying to calibrate the 100 kHz oscillator, as we can only hear down to about 20 Hz, we can't zero in any closer than this. Leaving the receiver at this point, we next turn on the 100 kHz calibrate oscillator. After a suitable warm up period, we turn the tune control of the oscillator and again adjust for a zero beat condition against the WWV frequency. With this method of calibration we have several possible sources of error. First in zeroing WWV with the receiver beat frequency oscillator and then zero beating the 100 kHz calibrator against the bfo. Each of these adjustments is limited to the lower limit of our hearing range, as well as the fact that we are obtaining the zero beat at a relatively low i-f frequency.

A more accurate method of calibrating the 100 kHz oscillator will now be discussed. After the receiver and calibrator oscillator have been warmed up for about 30 minutes, tune in WWV on a frequency that produces a fairly good, steady signal. Adjust the tuning for maximum reading on the S-meter. Having tuned in WWV, turn *off* the beat frequency oscillator. Now turn on the 100 kHz oscillator that is to be calibrated. If a harmonic of this oscillator is fairly close to the WWV frequency, a beat note will be heard. At this time adjust the calibrator "crystal tune" control and the S-meter will start pulsing from a maximum to a minimum value. The closer you get to dead center the slower the pulsing action will become. It is fairly easy to come down to 1 pulse per second with this method. If your receiver doesn't have a meter, you can also hear this pulsation very clearly. In any event you would always tune for the slowest pulse rate.

Note that with this method we have adjusted the calibrator frequency harmonic directly to the WWV carrier rather than to a low i-f. We have eliminated one zero beating step, and this, together with the fact that we are obtaining the zero beat at a much higher frequency, will provide greater accuracy.

Having described the method, here are a few points of general interest:

1. Before attempting any calibration let the equipment heat up for at least a half hour to stabilize.

2. After tuning in WWV, wait until the 400 Hz modulating tone goes off before adjusting the calibrator. If not, you may find later that you zero beat the 400 Hz instead of the carrier frequency!

3. The levels produced in the receiver by WWV and the calibrator oscillator should be about equal to produce a good beat between the two frequencies.

4. Use the highest WWV frequency that will produce a good, stable signal in the receiver. Certainly a 1 pulse per second beat at 20 MHz will provide greater calibration accuracy than 1 pps beat at 5 MHz or better yet than 455 kHz! The accuracy will be considerably greater and it is no more difficult to come by.

A SIMPLE FREQUENCY DEVIATION METER

John Reinartz K6BJ

A frequency-deviation meter allows one to read the deviation plus or minus that a received signal is off frequency. Depending on the meter range desired and used, a deviation of ten or less cycles can be read either high or low. Such a device is especially useful when used on MARS nets or when frequency checks are desired of any incoming signal and the answer must be in cycles low or high of a desired frequency. Those grinding their own crystals or desiring to compare crystals will find this device especially useful.

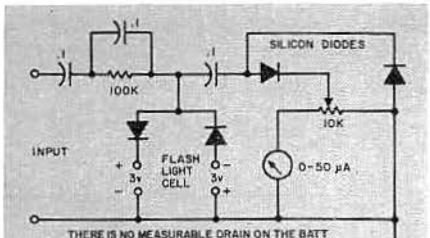
.1
.1
.1
100K
SILICON DIODES
10K
INPUT
+
3v
-
FLASH
LIGHT
CELL
-
3v
+
0-50 μA
THERE IS NO MEASURABLE DRAIN ON THE BATT
WHEN METER IS NOT IN USE.

Fig. 1

Two fundamental circuits were investigated, one using diodes only and the other using transistors only. These are shown in Fig. 1 and 2 respectively. A 0-50 microampere meter should be used for the diode type and a 0-1 milliampere meter will serve nicely for the transistor type, although a 0-50 or 0-100 microampere meter will also serve nicely in the transistor type frequency-deviation meter system. Silicon diodes were used for the diode type and 2N123 for the transistor type. An input voltage of 25 is needed for the diode type and 7 volts or less for the transistor type depending on the meter sensitivity, being 2 volts when a 50 microampere meter is used.

Whatever scale reading is desired, be it 250 cycles low or high or 500 cycles low or high, the meter cover is removed and new figures are added below the meter scale with a zero in the center of the scale and maximum readings at each end of the scale as appropriate. Pencil markings will do. The 250 cycle can be read to 10 cycles per division and the 500 cycle scale can be read to 20 cycles per scale division. Each can be read to half these values or 5 and 10 cycles respectively.

In use, you set your frequency meter, LM or 221 either 250 cycles or 500 cycles lower than the frequency to be checked. If the frequency to be checked is right on, the frequency-deviation meter will read zero at the center of the scale on the meter; if the frequency is low, the meter will read low and if higher, the meter will read higher. The answer in cycles will be the value indicated by your new markings. In use the frequency-deviation meter is connected across the high impedance output of your receiver in the case of the diode type and across the low impedance output in the case of the transistor type.

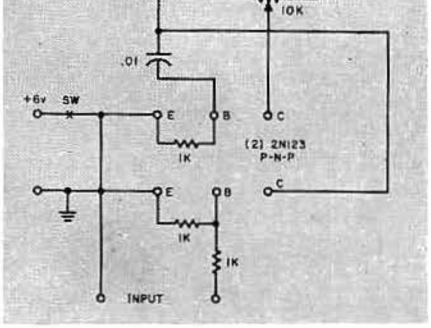

Fig. 2

In those cases where a definite frequency will be under observation, it will be found advantageous to grind or obtain a crystal that is adjustable to 250 cycles low or 500 cycles low, as appropriate, and to use it in the transistorized oscillator shown in Fig. 3. Any crystal holder that has an adjustable air gap will do. Some of the TCS surplus crystal holders have a three point adjustable top plate and are about the best obtainable. Since your best and probably only method of adjusting the crystal is by the use of your LM or BC221, be sure that your frequency standard is accurate. It is best to use the low frequency position and with the 1 mc crystal switch on, tune the meter to that portion of the desired frequency less the mc part. For instance, to set the LM for a reading of 2,732,000 cycles, set the LM on the low frequency for 732,000 cycles or 732 kc. The 1 mc crystal will furnish the mc part of the reading. In my case I set the LM to 731.5 kc and adjust the crystal to that frequency in the adjustable TCS holder. The transistorized oscillator holds the frequency to such a close tolerance it has not been necessary to make adjustments in weeks. A hand held push switch connected into the positive battery lead allows the oscillator to be turned on as needed to check the frequency of a MARS station on 2732 kc to an accuracy of plus or minus 10 cycles of that frequency. The frequency-deviation meter is of course checked against the 440 or 600 cycle tone of WWV, no other check is necessary since the scale is linear.

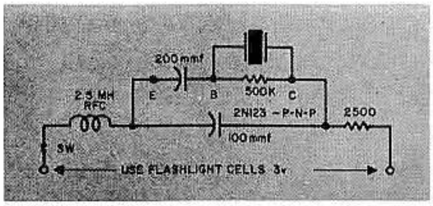

Fig. 3

Chapter VII

GDOs

THE INDICATING OSCILLATOR

Ken Brown KH6AF

Relatives of this little gadget have been around a long time! Even the transistorized versions, which usually leave something to be desired. With an FET, however, we are back in business as with tubes, but with many advantages.

The range of this oscillator is fantastic. That is, without any circuit tricks or special handling. The low frequency end was carried down to the 1 MHz in order to cover the lowest ham band. The high end takes care of 250 MHz easily. This can be extended with a little more effort.

Use of a field-effect transistor (FET) allows operation more nearly like the tubes with which we are, perhaps, more familiar. However, we are not tied to the power lines, which alone makes it worthwhile.

A 2N4221 FET was used in this indicating oscillator. Very likely other FETs will work also. The MOSFETs also should be as good, if not better. The new ones with built-in diode protection would be much easier to handle.

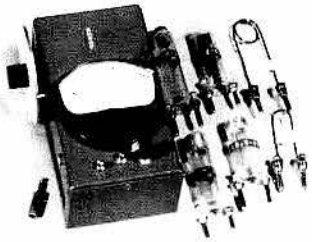

Standard banana plugs spaced ¾ in. make ideal bases for the coils. The lower frequency coils are wound on polystyrene forms.

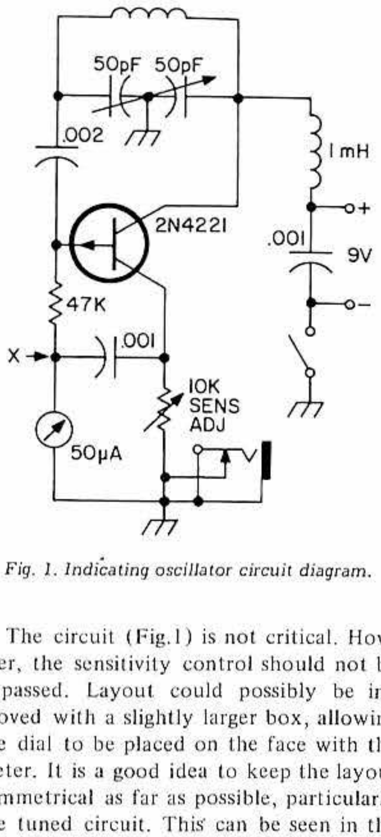

Fig. 1. Indicating oscillator circuit diagram.

The circuit (Fig.1) is not critical. However, the sensitivity control should not be bypassed. Layout could possibly be improved with a slightly larger box, allowing the dial to be placed on the face with the meter. It is a good idea to keep the layout symmetrical as far as possible, particularly the tuned circuit. This can be seen in the photo of the inside view. The box used was an LMB 532 EL with the cover reversed to allow for coil mounting insulator (polystryrene or other good rf insulating material). The meter should be a 50 μA movement; otherwise, a meter amplifier such as the one shown schematically in Fig. 2 will be necessary. This is no problem, as there is room for this amplifier on the circuit board.

A thumbwheel from a BC-375 tuning unit could be used very nicely as a dial. Three sides have been left clear for ease in placement of unit when in use. Plug-in coils makes for easy bank change and application to the job at hand. Standard-spaced banana pins allow for use with other accessories. Use 5/8 in. polystryrene tubing and stud-type banana pins.

A dual banana plug can be used during construction for setting the spacing accurately. The coils shown have the pins wired in place for stronger mechanical assembly. Drill a hole for about 24 AWG copper wire on each side of the pin studs which will lie along each side of the poly tubing. Use a number 59 or 60 drill. One wire is enough on each pin. Form a hairpin with about an inch of wire, push it through the drilled holes from the outside. Now twist tightly with longnose pliers, cut it off short, but not so short as to allow the wire to untwist! Then apply several coats of liquid "poly" cement. Be sure to move the coils frequently during the hardening period to make sure the liquid "poly" flows evenly over the stud and forms a slight fillet with the tube. Epoxy doesn't seem to work well with polystyrene. Neither Allied nor Newark list liquid polystyrene any longer, but your neighborhood hobby shop should be well stocked.

The lowest frequency coil (number 1, 0.95–2.2 MHz) was made from a Miller

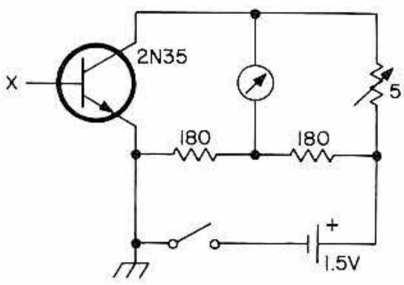

Fig. 2. This meter amplifier will increase the level of the signal so that a less-sensitive meter than 50 µA may be used.

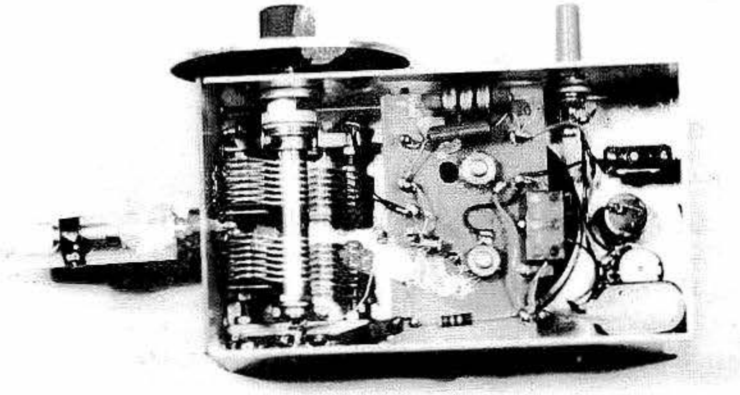

Interior view of oscillator with coil attached shows construction of the author's version.

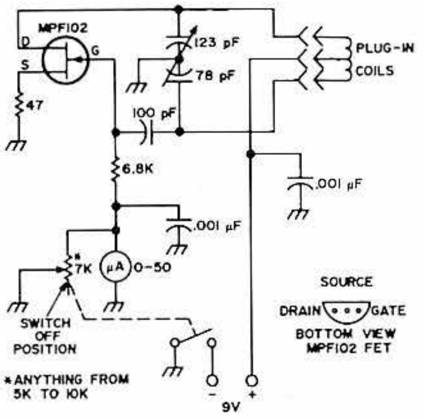

Fig. 1. Schematic diagram of the little gate dipper.

951 ferrite 0.5 mH choke with 6 or 8 turns removed – just enough to slip inside the poly tubing. Some reaming may be necessary. The number 2 coil (2.2–5.4 MHz) was made from a Miller ferrite antenna unit with the slug permanently installed in the top end and all lugs and mounting hardware removed. This coil was also mounted inside the poly tubing. The number 3 coil (5.4–13.5 MHz) consists of 32 turns of 28 AWG enameled wire, close-wound on the outside of the poly tubing. All coils are 2½ in. long with windings as near the end as practical. The number 4 coil (27–50 MHz) consists of 10 turns of 24 AWG enameled wire, close-wound. The number 5 coil (45–100 MHz) consists of 2 turns of 14 AWG enameled wire, self-supporting. The number 6 coil (60–270 MHz) is one hairpin loop 3/8 x 1 in.

A jack is provided for headphone use. The screwdriver-adjusted miniature pot is for zero adjustment of the meter amplifier. It can be seen in the photograph.

THE LITTLE GATE DIPPER

John Aggers W5ETT

Possibly you now own a grid dipper, but is it small, easy to handle, and cordless, making it completely portable? If not, you will want to build this gate dip meter. The cost is extremely low – only about $7. All parts are readily obtainable and construction is simple. The plug-in coil forms, using battery plugs and polystyrene tubing, are easy to make.

The Circuit

An MPF 102 FET is used in a modified Colpitts circuit. Except for the #1 coil, where a choke is used, the B+ is fed to the centertap of the coil. This is necessary to obtain a fairly constant gate current as the oscillator is tuned to its end frequencies. Drain current varies from 4 to 1 mA proceeding from 225 to 1.7 MHz. At the same time the gate current varies from 20 to well over 50 µA.

From this, it is apparent that the stronger the oscillations the smaller the drain current and the larger the gate current. In gate dip operation, as power is drawn from the oscillator the drain current will increase and the gate current will decrease or dip.

Limited wavemeter operation, obtained by switching off the B+, is accompanied by a slight shift in calibration. When the circuit picks up rf, the FET suddenly goes into oscillation using the rf as its battery. Thus, the amount of rf picked up must be large enough or there will be no oscillation and no meter indication. However, despite these deficiencies, it is still considered a useful mode of operation and for that reason has been included. It is only necessary to wire the sensitivity control so that the resistance is maximum when the switch is in the off position.

Construction

A natural finish aluminum minibox (4 x 2-1/8 x 1-5/8 in.) is used for the meter case. The variable capacitor came from an old transistor radio and measured 1 3/8 x 1/2 x 1 in. The shaft was already squared and tapped for a small screw. Since those listed in the catalogs have a plain or flat shaft, you will have to use a collar with setscrew, or drill and tap the shaft. The trimmer capacitors are not used and should be removed.

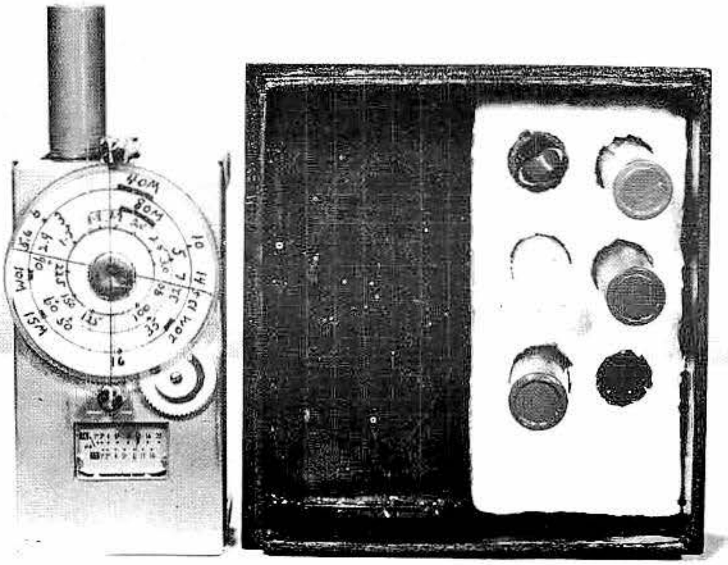

The little gate dipper with spare coils.

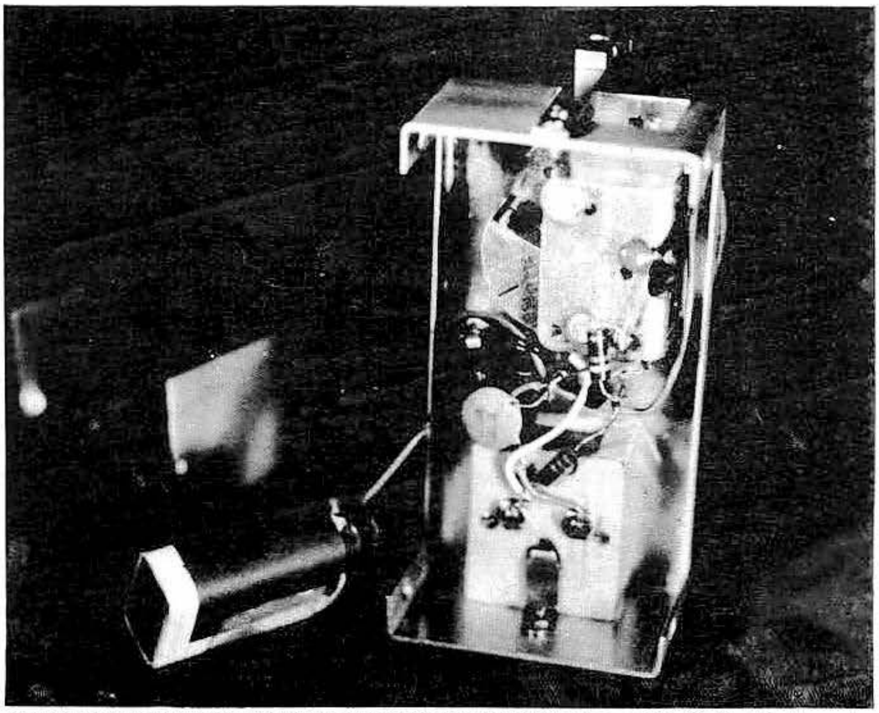

Meter is held against the front panel by a small bracket. The FET is the small black object in the center.

To make the coil socket you will need three pin receptacles from an octal socket, two pieces of 1/8 in. Plexiglas approximately 7/8 x 3/4 in., and one battery plug for a pattern. The pins of the battery plug form a triangle. I shall refer to the holes at the base as the outside holes. Drill holes in one piece of plastic to match the pins of the battery plug. Match the two pieces of plastic, clamp in a vise, and drill the two outside holes in the second piece. Bend the lug part of each socket pin to a right angle. Slip one over each outside pin of the battery plug. Using this as a jig, solder the lug portions to the stators of the variable capacitor. Remove the plug, and the pieces of plastic should fit down over the variable capacitor. The lug part of the center socket pin is brought out between the two layers of plastic.

File a small notch in the bottom piece to accommodate the lug. Before cementing the two pieces together and to the frame, make each hole slightly larger than the diameter of the socket pins. This will allow for expansion when the plug is inserted.

The dial is made of 2-1/4 in. diameter 1/8 in. Plexiglas. To give the dial a rough edge, for good thumb traction, I heated an old gear wheel and rigged up an arrangement to rotate the dial against it. The gear should have rather coarse teeth and rotate with the dial, or you will create flat spots.

The variable capacitor can now be mounted in the case. Position it so that the top and sides of the dial will be just about even with the edges of the case.

The dial marker is mounted on square aluminum posts. The top post (2 in. long) has 1½ in. of its length filed down to a 1/8 in. thickness to reduce its bulky appearance. To make the hairline, scribe a line in a ½ in. wide piece of plastic and fill in with a ballpoint pen.

The sensitivity control I used was already prepared for the knob shown. If you don't have one like it, use a dime-size pot and a setscrew knob. Any resistance from 5 to 10 kΩ will be fine.

Keystone light meters are available from Olson Electronics in a package of five ($3.99) or Transistors Unlimited Co. (75¢ each). Some modification of the meter is necessary. Remove the light cell and series resistor. Drill two holes, spaced ½ in. apart in the back of the case to pass 4-40 machine screws for easy soldering, make sure the heads and nuts are clean and free of any nickel plating. The screws should be filed even with nuts in order to make room for the battery. Solder the leads from the meter movement to the terminals, but be quick because the plastic case tends to melt in a hurry.

Wiring is just a here-to-there proposition, requiring no terminal boards or terminal lugs. The FET is soldered in place supported by its own leads. With reasonable care you should not damage it. A battery holder was found unnecessary; however, it is a good idea to wrap a layer of tape or stiff fiber paper around the battery to prevent the metal case from shorting out the meter terminals.

Coil Construction

Figure 2 and the photo give the necessary dimensions and show the parts needed to make the coil forms. The battery plugs are listed in the catalogs to fit #482 and

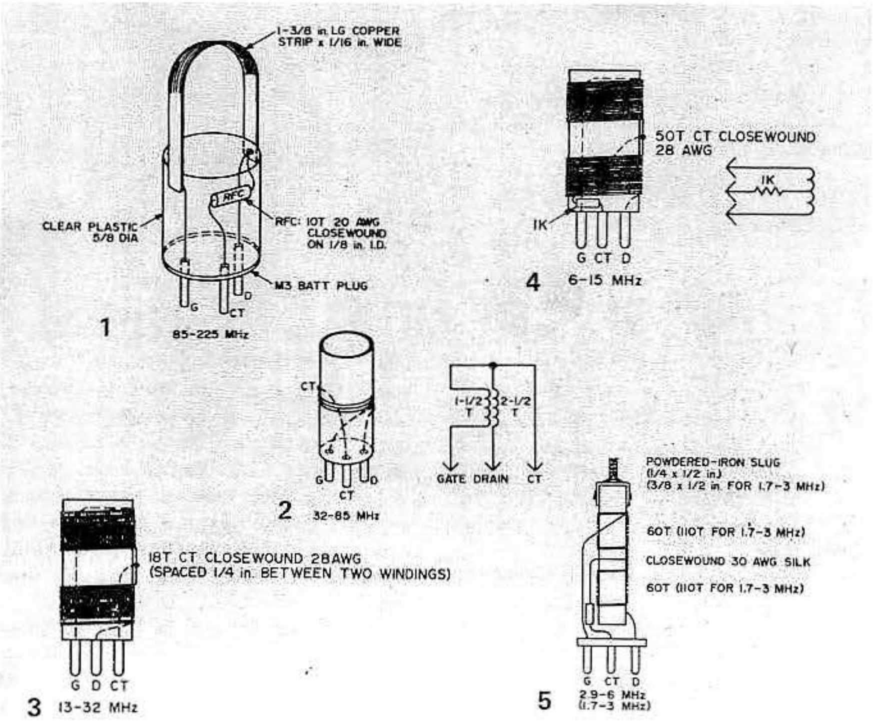

1-3/8 in. LG COPPER STRIP x 1/16 in. WIDE
CLEAR PLASTIC 5/8 DIA
RFC
RFC: 10T 20 AWG CLOSEWOUND ON 1/8 in. I.D.
M3 BATT PLUG
G
CT
D
1
85-225 MHz
CT
G
CT
D
2
32-85 MHz
1-1/2 T
2-1/2 T
GATE
DRAIN
CT
18T CT CLOSEWOUND 28AWG (SPACED 1/4 in. BETWEEN TWO WINDINGS)
G D CT
3
13-32 MHz
50T CT CLOSEWOUND 28 AWG
1K
1K
G CT D
4
6-15 MHz
POWDERED-IRON SLUG (1/4 x 1/2 in.) (3/8 x 1/2 in. FOR 1.7-3 MHz)
60T (110T FOR 1.7-3 MHz)
CLOSEWOUND 30 AWG SILK
60T (110T FOR 1.7-3 MHz)
G CT D
5
2.9-6 MHz (1.7-3 MHz)

Fig. 2. Coil configurations for various frequencies of resonance.

M3 batteries. The center pin should be filed slightly shorter to make the plug seat evenly in the socket. While you are at it, file the nickel plating from the ends of all plug prongs. This will make for easier soldering.

Complete coil information is given in Fig. 2. However a little explanation may be in order. The irregular method of winding the #2 coil is necessary to reach 85 MHz and still maintain oscillation. With 4 turns close-wound the highest was too low. With the 4 turns spaced, oscillation ceased at the highest frequency. The 30 AWG silk wire was taken from a TV flyback transformer. The resistors in the centertap of the last three coils improve the meter's sensitivity slightly. They are mounted right next to the coil winding. With a slight groove filed inside the insulating sleeve, it should slip over the resistor.

The #6 coil is layer-wound as space permits and scrambled wound the rest of the necessary turns. The top winding of all coils should end near the very edge of the coil form. This will make for easier coupling to a tuned circuit. After the coils are checked out the insulating sleeves may be glued to the plug base.

Allow the glue to dry for several days before plugging the open ends of each coil with a small cardboard disk. The coils are painted with colored lacquers. Colored paper between the coil and the insulating sleeve will probably work just as well.

Calibration

For calibration purposes, you will need another indicating oscillator or dip meter. Operate it in the diode or wavemeter mode and loosely coupled to the gate dipper. I calibrated only 5 points on each scale plus any ham bands which appeared. Remember the dipper is not a frequency meter but something to get you in the ballpark.

Conclusion

The little gate dipper was checked against a well known commercial tube equivalent and, as near as I could tell, they were just about even. The battery should last for a long time because the current drain is extremely low.

THE GREAT DIPPER

John E. Boyd WA0AYP

Test equipment is essential in the hamshack, as those of us have found when we attempted to get that new piece of homebrew perking for the first time. One of the most useful pieces of test equipment is the grid dip oscillator or simply, the GDO; besides being relatively inexpensive, it is particularly versatile. Need an indicating absorption wave meter? The GDO will do that. How about a modulated signal source? It handles that too. If you are interested in discovering its whole variety of uses, why not purchase one of several books on the subject.

Every item used in this GDO was selected with an eye toward the average home builder. There are no parts which must be specially purchased from the West Indies Export Company or similar outfit. Nearly all the parts, except for the meter, miniature pot, and mode switch, were obtained from the junkbox, or rather, from several junkboxes. If you insist on buying all new parts, total cost of the project will be about $20.

Circuit description

The grid-dip oscillator, in this case more properly termed an emitter-dip oscillator, gets its name from the fact that emitter current in transistor Q_1 decreases when the tuned circuit C_1-L_1 is in resonance with a nearby circuit. This decrease is easily seen by the dip of the meter indicating pointer.

When switched to the diode position, B+ is removed from the oscillator and the incoming rf is rectified by diode D_x; the voltage developed across the 2k resistor is amplified by the meter amplifier and monitored by the 0-1 mA meter. In switching to the signal position, B+ is removed from the meter amplifier but applied to the modulator, and a 1 kHz tone is available from one of the output jacks. In the modulated oscillator position, B+ is reapplied to the oscillator, and the oscillator is modulated by the 1 kHz tone.

Like a patch-work quilt, this GDO was built using circuits from already published articles or books and modified where necessary. The whole circuit is composed of three separate entities–oscillator, meter amplifier, and audio tone generator. The circuit is not particularly critical, but lead lengths and dress in the oscillator must follow good VHF practice, if stable VHF oscillation is to be maintained.

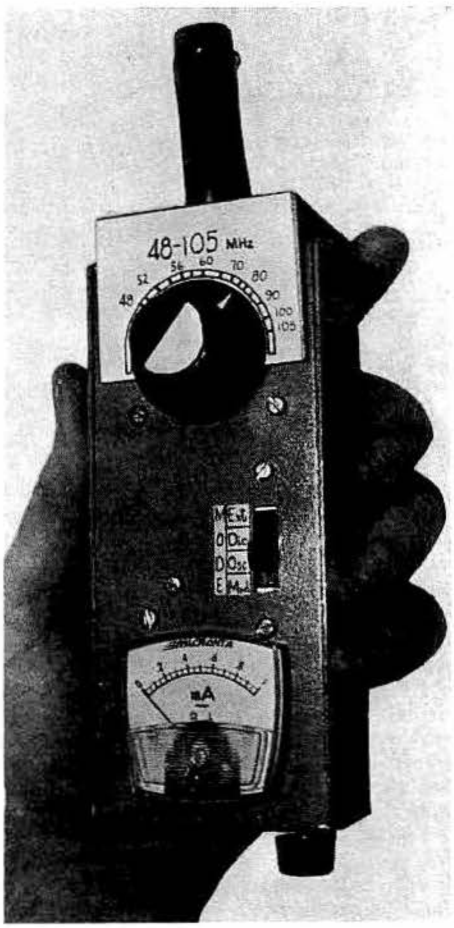

My operating time is spent on the various bands from 28 mHz to 432 mHz. Quite naturally, when I discovered that the GDO would not oscillate satisfactorily over the entire range from 2 mHz to 200 mHz, I juggled values so that it would oscillate well at 216 mHz (for tuning frequency doublers to 432 mHz); then I tried to get as low in frequency as possible. Oscillation was vigorous to about 20 mHz. Coils and scales were then made to cover the respective ranges. It you don't do any homebrewing on the VHF bands perhaps you will find it necessary to

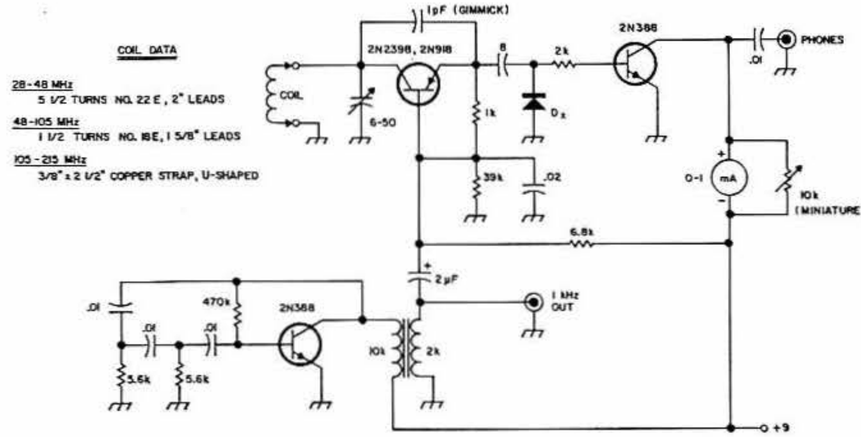

Fig. 1. Circuit diagram of the great dipper. Note that although the 2N2398 is a PNP transistor, the 2N918 is NPN, and if used as the oscillator transistor, problems would arise with voltage polarity. The diode D_x may be almost anything that you have available. The 1 pF gimmick capacitor consists of 1½" of twisted wire.

Internal construction of the great dipper. The modulator and oscillator boards are to the right—the oscillator transistor is mounted right next to the coil jacks.

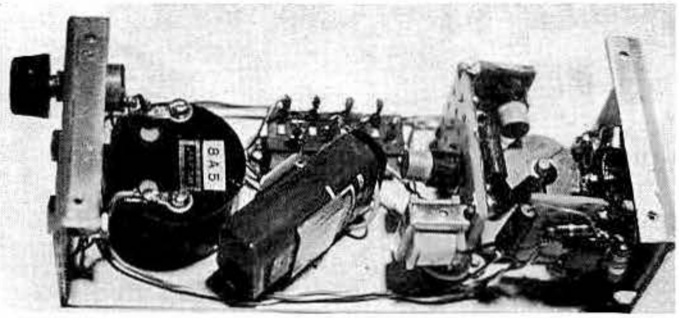

change the value of the emitter-collector feedback capacitor, and to juggle the emitter and base resistor values in order to sustain oscillation at your desired frequencies.

Construction

Vector board was used, mainly because I wanted to experiment with component values; however, a printed circuit would be just as good, especially for something such as a club project. It can be seen from the photograph that the meter amplifier and audio oscillator are built on separate boards. This is due to the fact that I built several different amplifiers; the layout would look neater if they were on the same board. Positioning of the rf oscillator and capacitor C_1 as shown in the photograph is recommended, but the placement of other parts is not critical.

Fiberglass board is used as an insulator for mounting the banana jacks and plugs. It cuts and drills easily and appears to work fine. Three banana jacks were used, the third jack being used merely to provide mechanical rigidity. It could also be used, if necessary, to shunt additional capacitance across the emitter and collector on the lower frequencies.

Because a shear and a brake were available, I constructed my own chassis, consisting of two U-shaped pieces of 1/16" aluminum. Using the GDO is a breeze, for it fits the hand very comfortably; if placed on the workbench, it doesn't roll off each time it is bumped. The completed case (1¾" H x 2¼" W x 6¾" L) is exceptionally rigid and imparts a reassuringly solid feel when handled. Commercially available miniboxes could be used if you don't have facilities available for rolling your own.

In building the rf oscillator, keep all the rf leads as short as possible; especially the short lead from C_1 to Q_1 and from circuit ground to chassis ground. It was found that false dips could be completely eliminated if a copper strap ¼" wide was added from the capacitor ground lug directly to chassis ground. Apparently the ground on the variable capacitor C_1 is not quite good enough at frequencies above 100 mHz. Various transistors were tried in the oscillator; the PNP type 2N2398 was found to be a good performer, as was the NPN type 2N918. However, the use of the NPN type could lead to problems with battery polarity. Capacitor C_2 is a 1½" length of twisted wire positioned near the collector lead of Q_1. This slightly modified oscillator circuit is from a book describing, among other things, a transistorized GDO which you may wish to use as a reference.

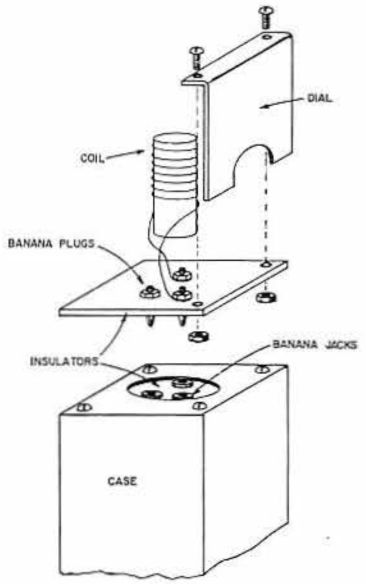

Fig. 3. Construction of the plug-in coil assemblies. The coil forms were made from the plastic containers which hold Polaroid print coater.

To keep cost low, a 0-1 mA meter was used in conjunction with a simple meter amplifier. If you happen to have a 0-50 μA meter lying in the junkbox, that would work equally well, and the circuit could be simplified accordingly. Several circuits were built for the meter amplifier; the one chosen was a compromise between cost and performance. A germanium transistor was used because it requires less voltage to turn it on. Leakage is low, the pointer of the meter resting just off zero when no coil is in place. Further information can be obtained from January 1966 *73 Magazine*, which was the source for this circuit.

A transistorized audio tone generator is coupled by a 2 μF capacitor to the base of the oscillator transistor for modulation. This modulated oscillator allows the GDO to be used as a versatile signal source. An output jack is included on the panel to allow the 1 kHz tone to be used without turning on the oscillator.

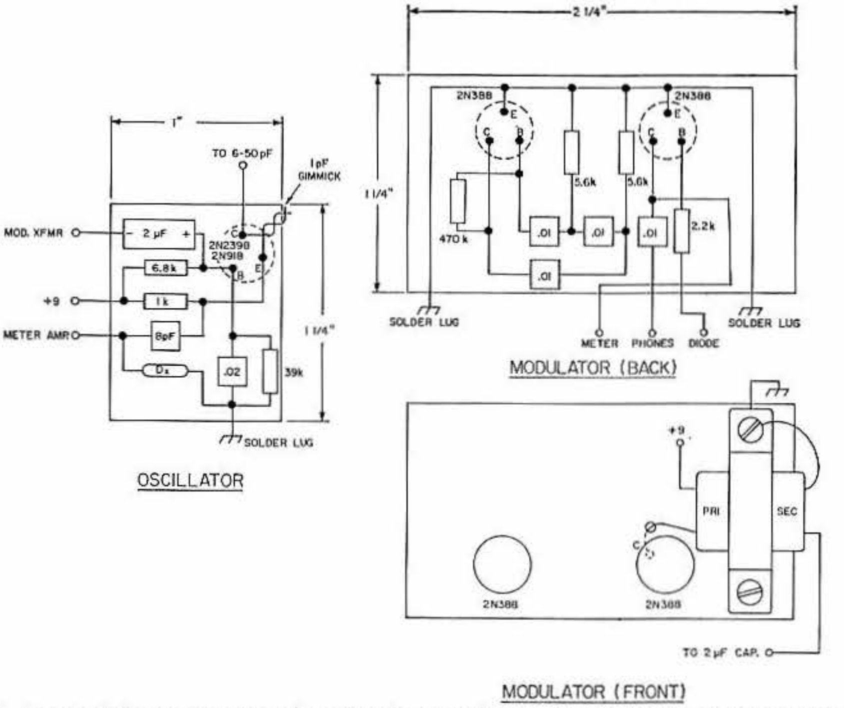

Fig. 2. Layout of the two circuit boards used in the great dipper. Although two boards were used in this case, the circuit could be easily adapted to one board, and even to printed circuitry.

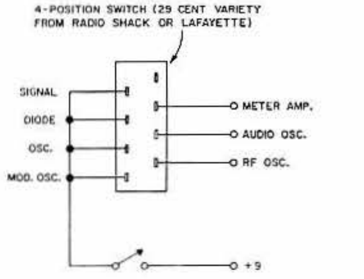
4-POSITION SWITCH (29 CENT VARIETY FROM RADIO SHACK OR LAFAYETTE)
SIGNAL
DIODE
OSC.
MOD. OSC.
METER AMP.
AUDIO OSC.
RF OSC.
+9

Fig. 4. Wiring the four-position slide switch for the great dipper.

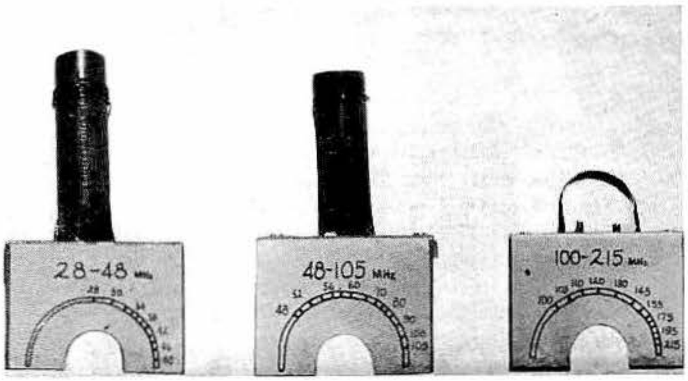

The three plug-in coils for the Great Dipper. Three ranges cover from 28 to 216 MHz. The second harmonic of 216 MHz may be used for tuning up 432 MHz converters and such.

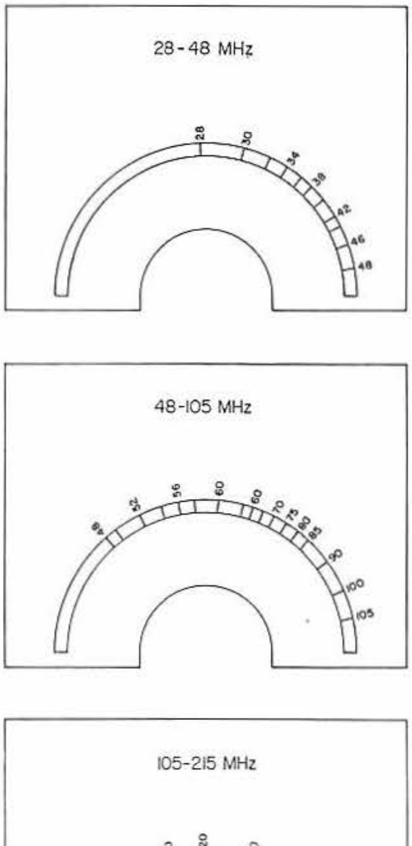

Fig. 5. Full-scale dials for the great dipper. If the construction shown in the photographs is followed closely, the calibration of these dials should be within several percent.

There are a couple of components which not everyone will want to duplicate. One, the sub-miniature 10k pot with SPST switch, was chosen because a very limited amount of space was available; if the unit is built on a larger chassis, the more commonly available *Midgetrol* could be used. The other, a four position switch used to select the desired mode, is a 29c variety available from Lafayette or Radio Shack. It has a peculiar switching arrangement and if you duplicate this project, several hours of experimenting could be eliminated by following the pictorial diagram included in this article. A disadvantage of this particular switch is that the meter does not indicate in the modulated oscillator position.

Using individual scales on each plug-in coil assembly greatly enhances scale legibility, reducing the chance of reading error and speeding frequency identification. This scheme; however, is not original. It was described in a 1957 issue of *Short-wave Magazine*, and is currently being used on a commercial GDO. It requires little additional effort to build the coil assembly in this manner and is, to me, well worth that extra effort. For want of anything else, the coil forms were made from the plastic tubes which contain the film coater supplied with each roll of Polaroid film.

Lastly, ease of tuning is accomplished largely through the use of a 1" skirted knob. Small knobs are simply too difficult to use comfortably.

Calibration and operation

It is best to calibrate this GDO by listening for the oscillator, modulated by the 1 kHz tone, on a general coverage receiver. An alternate method is to use another GDO, placing one in oscillate and the other in diode, tuning for either peak or dip. The scales which were used on this GDO will serve if parts and layout are followed closely.

To use this unit as a dipper, place the mode switch in oscillate, and place the dipper coil next to the coil under test. The turns of both coils should be parallel, and not at right angles to each other. To keep from pulling the oscillator frequency, keep the two coils separated as much as possible, while still maintaining a meter dip. This assures that dial accuracy will be kept high. If a coil is inaccessable, twist a pair of wires together, forming a two turn coil on each end; slip this coupling link over the two coils. Keep in mind that a coil, when it is in a circuit, may not dip at the same frequency as when it is out of this same circuit.

GREATER DIPPER – MODIFICATION

John E. Boyd WA0AYP

Introduction

To the uninitiated, a grid dip oscillator is neither glamorous nor exotic, even if it does happen to be transistorized. But to those who have ever tried to put a coil on frequency, one is worth its weight in microcircuits.

The original model of the *Great Dipper* proved to be a versatile, if unexciting, grid dip oscillator. Intended to be used mostly at VHF, the dipper was limited to frequencies above 28 MHz. To those who seldom operate above 10 meters, the lack of frequency coverage below 28 MHz was a definite handicap. Expanding the frequency coverage downward extends the dipper's usefulness considerably.

Component changes made were not extensive, nor were the physical dimensions of the unit changed. The schematic diagram of the modified circuit is shown in **Fig. 1**, and the pictorial diagram of **Fig. 2** is included to supplement the photographs printed in the original article.

Modification

First, the oscillator circuit board should be rotated 90 degrees from its original position so that transistor substitution or replacement can be made easily. A microwave diode might be substituted for the original glass computer-type diode; this results in better performance at VHF, but its effect cannot be accurately predicted. A 1N21 or 1N23 diode could be used in place of the D4900 called for in the schematic. Best oscillator performance over a wide frequency range results from using a 5 pF capacitor as the collector feedback coupler, but this was a compromise value. If operation is likely to be confined to either the HF or VHF segments of the radio spectrum, some experimentation with this capacitor should result in improved performance over that particular range. For experimentation, use a 3-12 pF or 4-30 pF trimmer.

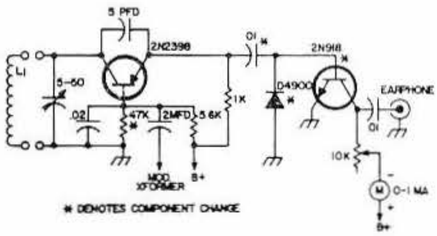

*Fig. 1. Schematic of modified circuit, with component changes indicated by *. L1: 13-23 MHz, 14½ turns No. 20; 22-44 MHz, 5¼ turns No. 24, 45-90 MHz, 1½ turns No. 24; 90-195 MHz, 2½ x 3/8 in.*

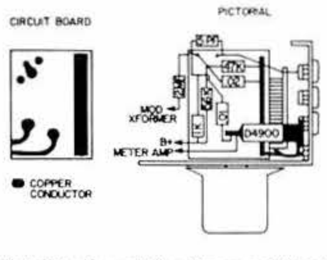

Fig. 2. Pictorial view of the dipper and its circuit board.

To perk up meter amplifier performance, the 2N388 meter amplifier is replaced by a 2N918, a relatively high gain, low leakage transistor.

Although the change was not made in this unit, a 6-100 pF (MAPC-100) variable capacitor could be substituted for the existing 5-50 pF unit. As can be seen from the full-size scales, frequency coverage is narrow on the 13-23 MHz plug-in coil. Those who frequently construct HF gear might benefit from the substitution.

Operation

Operation of the dipper is the same as before, except the frequency range is greater. Another coil must be wound for the 10-25 MHz range, but existing coils were not modified. Calibration is now different from before; full-size scales are included with this article for those who do not have a general-coverage receiver or another grid dip oscillator on hand for calibration.

Oscillation is still vigorous at 13 MHz, which indicates lower frequencies may be reached with a suitable coil plugged in. Similarly, oscillation is still noticeable at 195 MHz, so operation at higher frequencies may be possible.

Finishing Touches

Appearance of the dipper was improved after enlisting the help of a draftsman. The dial and meter scales were professionally inked. **Figure 3** provides full-size copies of these scales.

To improve the legibility of the labels and scales, they were inked on yellow cards rather than on conventional white paper. It's a small point, but for those who use test equipment frequently, the black-on-yellow technique results in less eye strain.

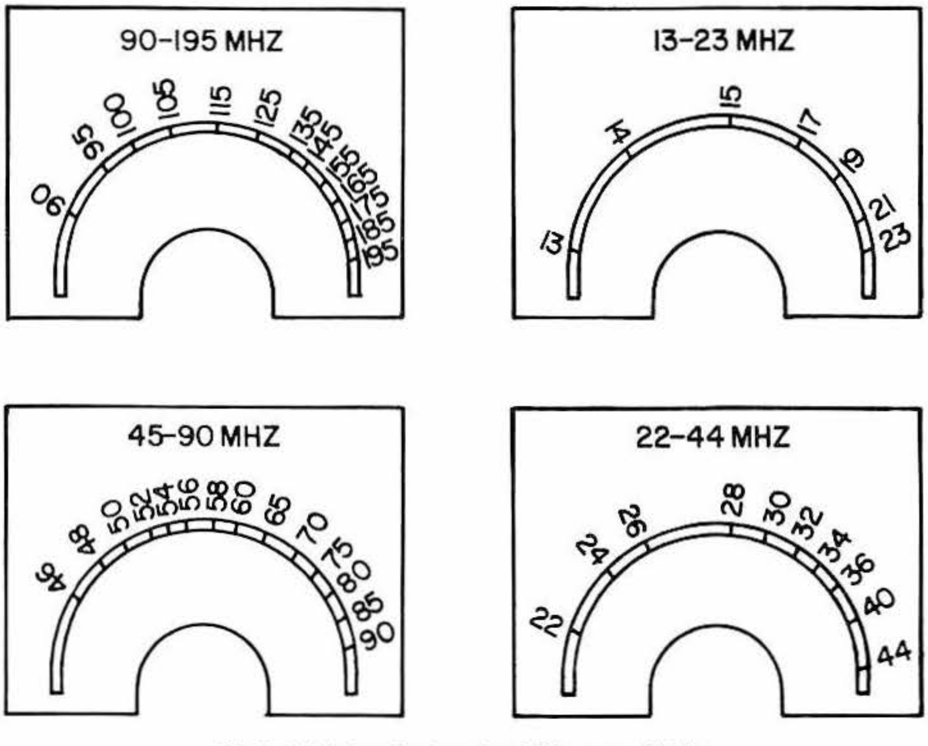

Fig. 3. Dial designs for the various MHz ranges, full size.

ETCHED CIRCUIT UHF DIPMETER

Bob Corbett W1JJL

I made three models of the dipper. They cover 130-175 MHz, 175-250 MHz, and 250-480 MHz. The first two models use the same size inductor, with the 130-175 MHz model using a larger capacitor for tuning with a ceramic trimmer across it. This trimmer is not shown in the schematic; its value is 7-45 pF and it should be adjusted to cover the proper range.

The 250-480 MHz dipper uses a smaller inductor than the others. It also has a copper jumper (shown in the layout) that the lower frequency dippers don't have.

Each dipper is complete (including the battery) except for the meter. The meters were omitted to save space and money, but can be included if you wish to use a slightly larger case.

Each dipper uses one RCA 2N3478 NPN silicon transistor. These transistors cost only $1.90 apiece, but it's likely that other transistors that are even cheaper could be used. The 2N3478 has odd basing—the only reference is the short case-shield lead—so don't shorten any leads until you're sure that you can keep track of the connections.

The copper side of the board for the dipper is shown in Fig. 2 with the component side in Fig. 3. Use glass or Teflon based board. Paper or bakelite board probably wouldn't be satisfactory. Trim the board to the proper shape with a nibbler. The inductor should be coated with coil dope to keep from shorting it when you use it.

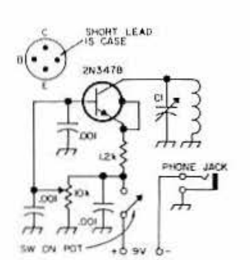

Fig. 1. The etched circuit dipper is very simple. The circuit is almost identical to the one described by WA1CCH in the December 1965 73, but the construction is quite different. C1 is Johnson 160-104 (9 pF) for the two higher frequency dippers, and 160-107 (14 pF) for the 130-180 MHz model. There is a trimmer across C1 in the 130-180 MHz model; see text.

To mount the boards, you'll have to cut a thin slot near the edge of the Minibox used as a case. One way to do it is to drill a number of holes of the proper size in a row, then use a file to finish the slot. You'll have to bend that side of the Minibox out to get the board in. It's held in place by an extra set of nuts on the shafts of the potentiometer and the tuning capacitor. Be sure to trim the leads projecting from the copper side of the board so that they won't touch the metal of the case. The battery is held in place with a simple clip made from scrap metal.

The dipper is very easy to use. But before we get to that, let's check it out and calibrate it.

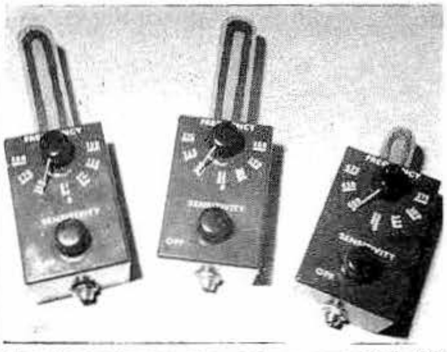

The three dippers shown here cover 130-480 MHz.

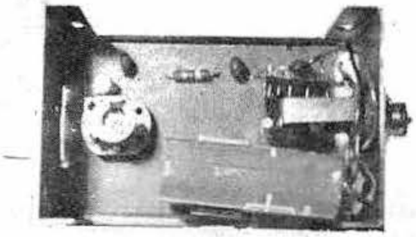

Here's the inside of one of the dippers.

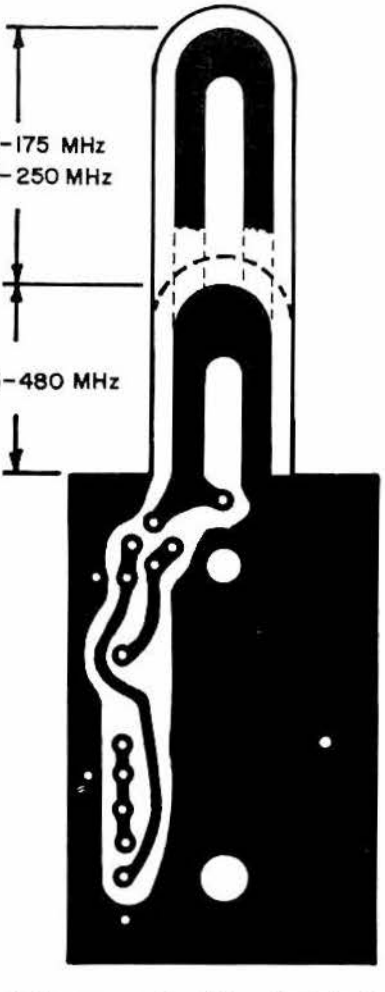

Fig. 2. The copper side of the etched circuit board used in the dipper. This layout is full size. Use board suitable for these frequencies: fiber glass or Teflon.

Plug a 2 to 5 mA meter into the meter jack. You can use a more sensitive meter if you shunt it with a resistor that gives the proper scale. Put the resistor across the meter jack terminals in the dipper if you use the meter for other things.

Turn on the dipper by twisting the potentiometer knob clockwise until it clicks. The meter should show very low current. As you turn the pot, the current should suddenly jump to about 1 mA. That means that the transistor is oscillating. If you touch the coil, the meter reading should drop and the dipper may stop oscillating completely. Now tune the capacitor through its range. There should be a little variation in current, but not too much.

Now you're ready to calibrate the dipper. The easiest method is a sensitive wave meter that covers the range, but it's quite easy to do the job with a TV set. A TV set covers 176 to 216 MHz (channels 7-13) for the low calibration. Then the second harmonics of the dipper tuning 235 to 445 MHz can be received on a UHF TV set (470-890 MHz). If you have a two meter receiver, that gives you another maker at 146 MHz. You can put on the panel markings with Ami-Tron or Datak rub-on lettering.

The dipper should be complete now, and ready for use. Bring the dipper near a resonant circuit in the dipper's range and tune the frequency control. You should get a prominent dip in current when both circuits are tuned to the same frequency. The amount of dip depends on the setting of the pot in the dipper, the distance from the tuned circuit, the Q of the circuit, and the type of coupling. In many cases it's easiest to leave the dipper stationary and tune the other circuit.

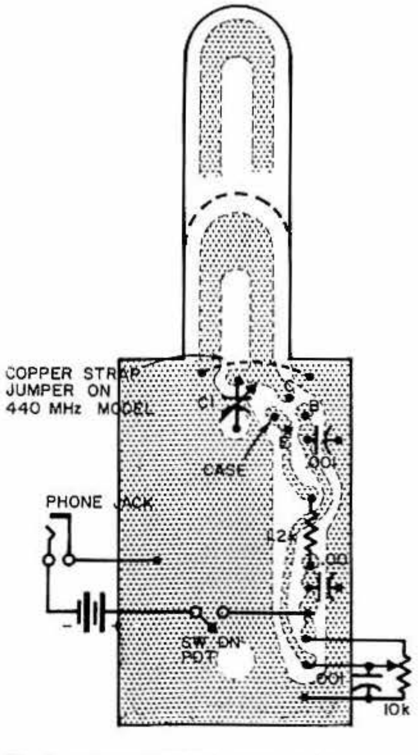

Fig. 3. Component side of the dippers. There is a 7-45 pF ceramic trimmer across C1 in the lowest frequency model. See the text.

The dipper can also be used for monitoring AM transmitters by plugging a set of headphones in the meter jack and adjusting the tuning and pot. You can also use the dipper for determining the frequency of another oscillator. Simply tune the oscillating dipper with headphones plugged in until you hear a slight click. You probably won't be able to get a zero beat at these frequencies.

Be careful that you don't use the dipper around an energized transmitter of more than a few watts output or the dipper may be damaged.

These dippers are simple, inexpensive and non-critical to build. After you've build them, you'll wonder how you ever tried to build UHF equipment without a good dipper.

UHF GRID DIPPER

J. Fisk WA6BSO

When building or testing equipment for the 420 mc band, amateurs invariably run into the problem of, "Where am I?". It becomes a little difficult to tell whether you are actually in the band or somewhere nearby. The uninitiated will counter that you should be able to figure close enough, after all the band is 30 mc wide; but even the experienced old timer will confirm that this just isn't so.

The 420-450 mc amateur band falls between the VHF and UHF television assignments and there is very little to use for a frequency reference point. It is nearly impossible to tell exactly where you are without resorting to expensive commercial gear or lecher wires. Nor is the problem confined to the ¾ meter band; when building crystal controlled converters for 432 or 1296, it's nice to know that the last tripler is really tripling and not doubling or quadrupling. Even with a 45 or 50 mc crystal, it's quite easy to tune up on the wrong harmonic. Remember that a 40 mc change at 400 mc is analogous to a 400 kc change on 75 meters.

There are several commercial instruments that fill this requirement nicely, but the cost of the least expensive of these would buy a pretty respectable all band receiver. Occasionally suitable test equipment appears on the surplus market, but again, the price is prohibitive. The simple UHF grid-dipper described in this article was designed specifically to economically fill this need. It covers the frequencies

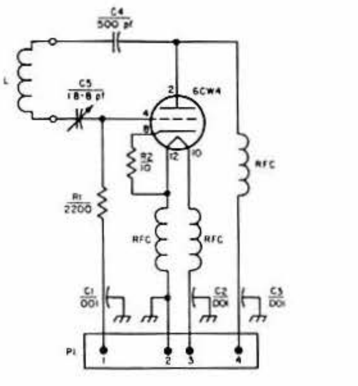

C1,C2,C3	STANDOFF BUTTON BYPASS CAPACITORS
C4	BUTTON CAPACITOR (SEE TEXT)
C5	E.F JOHNSON TYPE 160-104 VARIABLE
L	PLUG-IN COIL
P1	AMPHENOL 91-MC4M MALE CABLE PLUG
RFC's	8-1/2" No 2? ENAMELED WIRE WOUND ON 1/2 W, 100K COMPOSITION RESISTORS

Fig. 1. Schematic of the rf head of the UHF grid dipper.

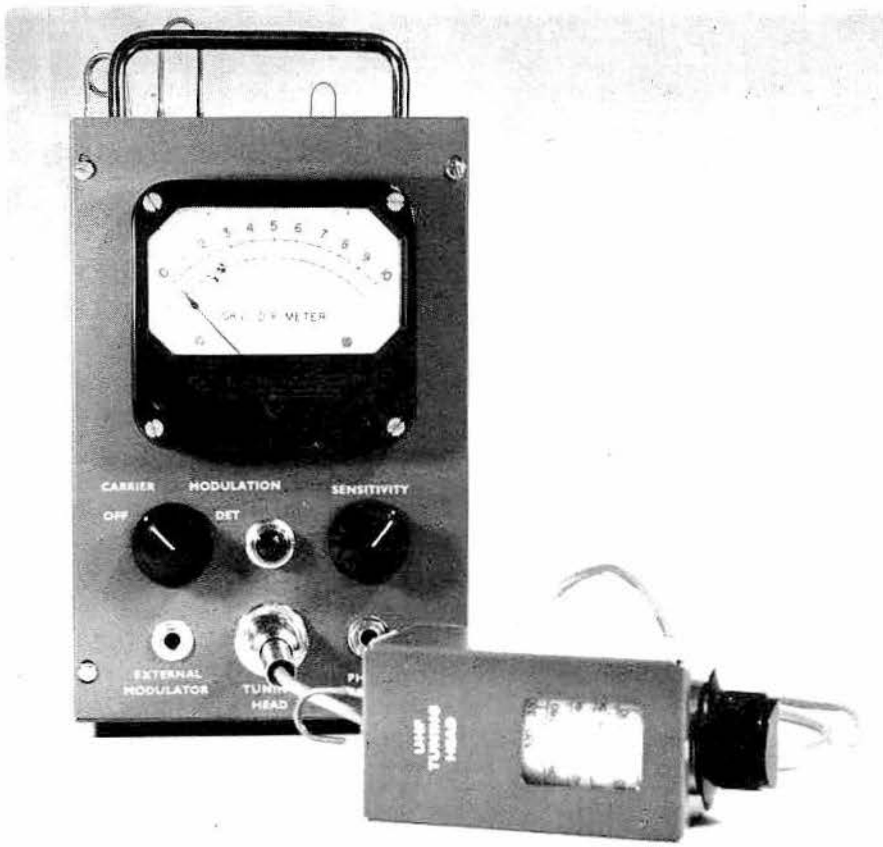

GRID DIP METER
CARRIER
OFF
MODULATION
DET
SENSITIVITY
EXTERNAL MODULATOR
TUNING HEAD
UHF TUNING HEAD

from 300 to 680 mc at a total cost that is comparable to a low-frequency kit-type grid-dipper. By using junk-box parts and smart horsetradesmanship this cost may be substantially reduced. I should hasten to point out that substitution of parts in the tuning head should be avoided if accurate dial calibration is desired.

This particular grid-dipper has been duplicated several times and in each case calibration has been within several percent of the original unit; at least as accurate as the familiar kit-type grid-dipper. This is helpful to the ham who does not have access to commercial equipment for calibration purposes. By using the layout and parts described in this article, the precalibrated dial illustrated in **Fig. 8** may be used with a minimum of error.

Although an accessory power supply, modulator and indicator are included in this design, an enterprising ham could use the power supply and meter presently available in any lower frequency tube type grid-dipper. Although using a modified low-frequency dipper is simple and inexpensive, the power/indicator box specifically designed for the UHF tuning head has several features that are not available in inexpensive kit-type units. These include internal 1,000 cycle modulation, provision for external modulation, and voltage regulation which provides the necessary stability at the ultra high frequencies.

The rf tuning head

The heart of this instrument lies in the tuning head itself. There is nothing particularly new or different about the circuit, but at these frequencies stray inductance and capacitance in circuit layout and construction will seriously affect the end result. All the lead lengths must be as short as humanly possible and physical circuit layout must follow standard UHF practice. One of the big problems in building tunable oscillators at UHF is to obtain an oscillator that will tune from one end of the range to the other with no "holes," frequency jumping or instability. The series tuned 6CW4 nuvistor oscillator shown in **Fig. 1** fills these requirements.

Insofar as possible, all the tuning head wiring is done on a point-to-point basis with the components mounted directly to the 6CW4 tube socket or button capacitor lugs as illustrated in **Fig. 3.** To obtain the desired accuracy with the precalibrated dial, this diagram should be followed as closely as possible.

To keep stray circuit inductance to a minimum, all wiring between the coil and tube socket is done with one-quarter inch wide strips of thin copper sheet. In addition, the rotor of the variable capacitor is connected to the 6CW4 grid (pin 4) with a short strip of the same material. The only tricky part of this wiring is the installation of the series capacitor C4. This capacitor is a 500 picofarad button type mounted as shown in **Fig. 5.** The "S" shaped bracket ("A") is made from a one-quarter inch wide copper strip, one and one half inches long and soldered to the mounting stud of the capacitor. Connection to the 6CW4 plate (pin 2) is made with the soldering lug on top of the button.

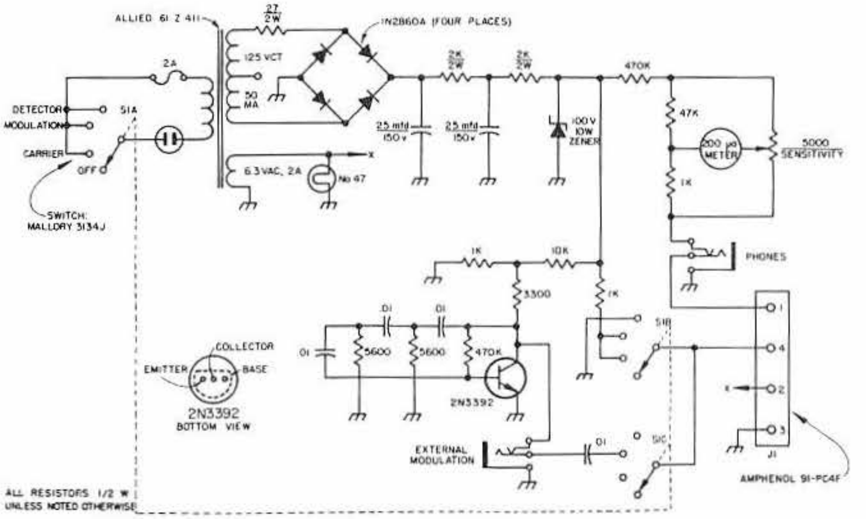

ALLIED 61 Z 411
27 2W
1N2860A (FOUR PLACES)
2A
125 VCT
50 MA
DETECTOR
MODULATION
CARRIER
OFF
S1A
SWITCH: MALLORY 3134J
6.3VAC, 2A
No 47
X
25 mfd 150 v
25 mfd 150 v
2K 2W
2K 2W
100 V 10W ZENER
470K
47K
200 μa METER
5000 SENSITIVITY
1K
PHONES
1K
10K
3300
1K
.01
.01
.01
5600
5600
470K
2N3392
S1B
COLLECTOR
EMITTER
BASE
2N3392 BOTTOM VIEW
EXTERNAL MODULATION
.01
S1C
1
4
2
3
J1
AMPHENOL 91-PC4F
ALL RESISTORS 1/2 W UNLESS NOTED OTHERWISE

Fig. 2. Power supply and modulator for the UHF grid dipper.

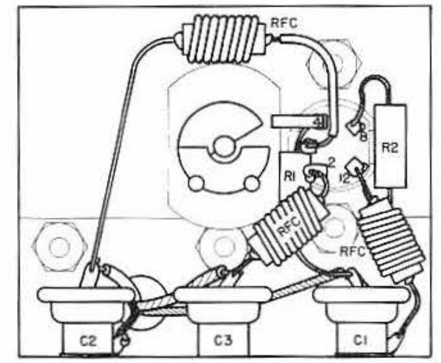

RFC
R2
R1
RFC
RFC
C2
C3
C1

Fig. 3. Critical wiring around the 6CW4 socket.

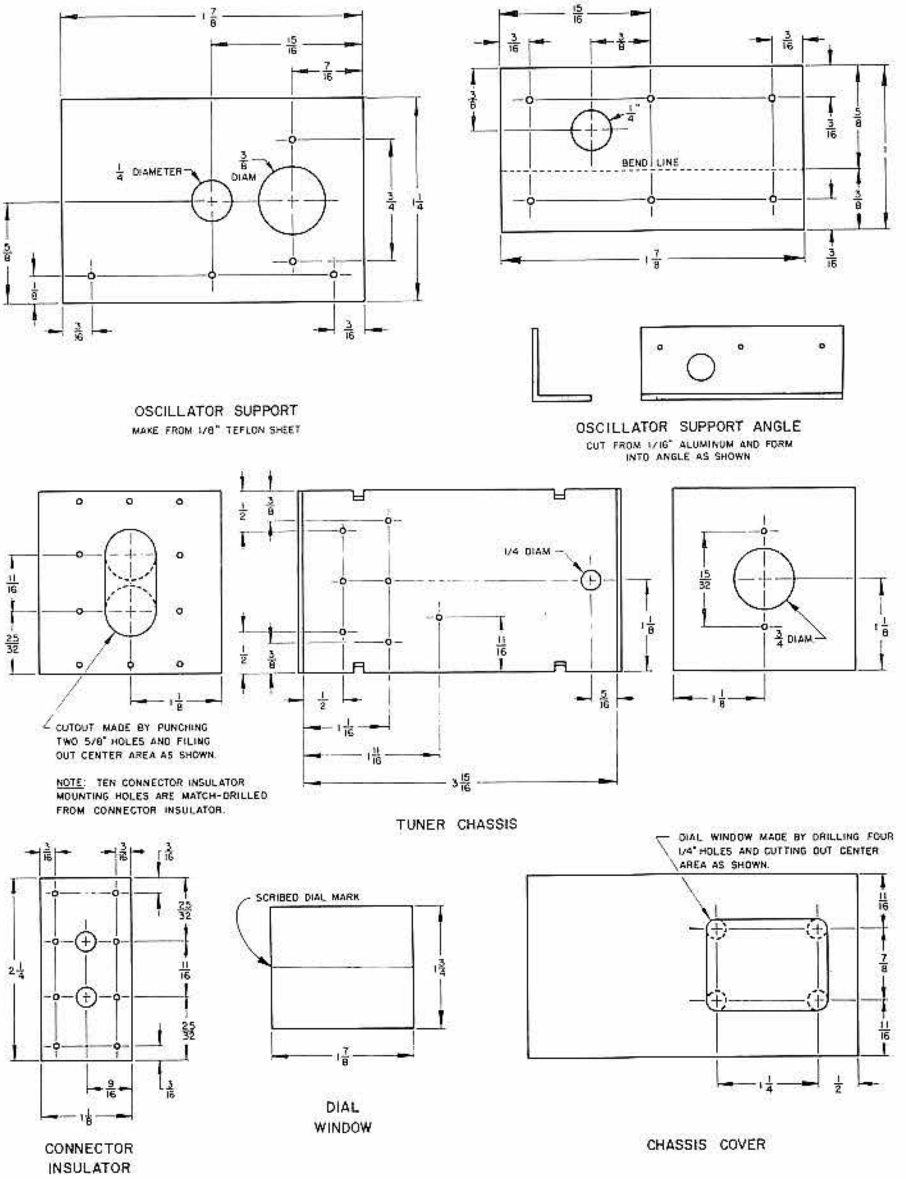
OSCILLATOR SUPPORT
MAKE FROM 1/8" TEFLON SHEET
1/4 DIAMETER
3/8 DIAM
OSCILLATOR SUPPORT ANGLE
CUT FROM 1/16" ALUMINUM AND FORM INTO ANGLE AS SHOWN
BEND LINE
CUTOUT MADE BY PUNCHING TWO 5/8" HOLES AND FILING OUT CENTER AREA AS SHOWN.
NOTE: TEN CONNECTOR INSULATOR MOUNTING HOLES ARE MATCH-DRILLED FROM CONNECTOR INSULATOR.
1/4 DIAM
TUNER CHASSIS
3/4 DIAM
DIAL WINDOW MADE BY DRILLING FOUR 1/4" HOLES AND CUTTING OUT CENTER AREA AS SHOWN.
SCRIBED DIAL MARK
DIAL WINDOW
CONNECTOR INSULATOR
CHASSIS COVER

Fig. 4. Layout of the tuner chassis, chassis cover, dial, connector insulator, oscillator support and oscillator support angle.

Connection to the stator of the variable capacitor is accomplished with another short strip of thin copper as shown in **Fig. 5** ("B"). This piece of copper is bent so it touches both stator mounting pins when the unit is assembled; then it is soldered in place.

The oscillator and tuning mechanism are housed in a standard 2¼ × 2¼ × 4 inch chassis box laid out as shown in **Fig. 4**. Although the author's unit is based on an LMB type 107 chassis box, other manufacturers have similarly sized boxes which are equally suitable. The layout of the enclosure is straight-forward and no difficulty should be found in duplicating it. The dial and coil socket cutouts are made by drilling or punching round holes and then cutting out the area between them as shown in the drawing. This is easily done with an Adel "nibbling" tool. All of the small screw holes are drilled with a standard ⅛ inch drill to pass 4-40 screws; 5/32 inch holes will be required if 6-32 screws are used.

The dial window is cut from a sheet of 1/16 inch clear plastic to the dimensions shown in **Fig. 4**. A hairline is scribed in the center of this window; if a more distinct line is desired, this scribe mark may be filled in with black india ink. After the tuner enclosure is completed, this dial window is epoxied in place over the dial cutout.

The coil socket is made by installing two banana jacks (E. F. Johnson type 108-740) on 11/16 inch centers in the Teflon connector insulator illustrated in **Fig. 4** (if Teflon is not available, Polystyrene may be used). This "socket" is installed over the large oblong hole cut in the end of the chassis. The screw holes used for mounting the connector insulator are match-drilled to the holes in the insulator itself. In this way it is properly mated to the enclosure. Although nylon attaching screws were used in the original model, they are not necessary and regular metallic screws will not alter any of the oscillator's characteristics.

To reduce stray capacity to a minimum, the oscillator circuitry is mounted on an insulating sheet. Teflon was used for this purpose in the original unit, but epoxy board with the copper peeled off would be perfectly suitable. Sheet polystyrene is not too desirable in this location because of its susceptibility to heat. The envelope of the 6CW4 gets very hot and the ambient temperature within the confines of the small chassis is quite high.

The general layout of the Teflon oscillator support and associated support angle are also shown in **Fig. 4**. The angle is cut out from a piece of 1/16 inch aluminum sheet and bent in a vise to form the angle. When these two pieces are mated together, the metal angle will probably interfere with the lower nuvistor socket mounting screw. It must be drilled out using the hole in the oscillator support as a guide. This additional hole is not shown because its exact location will vary from unit to unit and depends upon the accuracy with which the parts are laid out.

The dial mechanism is not complicated, but it is hard to ascertain from the photographs exactly how it is put together. The exploded drawing of **Fig. 6** should help in this respect. The vernier mechanism is an Eddystone 10:1 planetary drive that provides both smooth

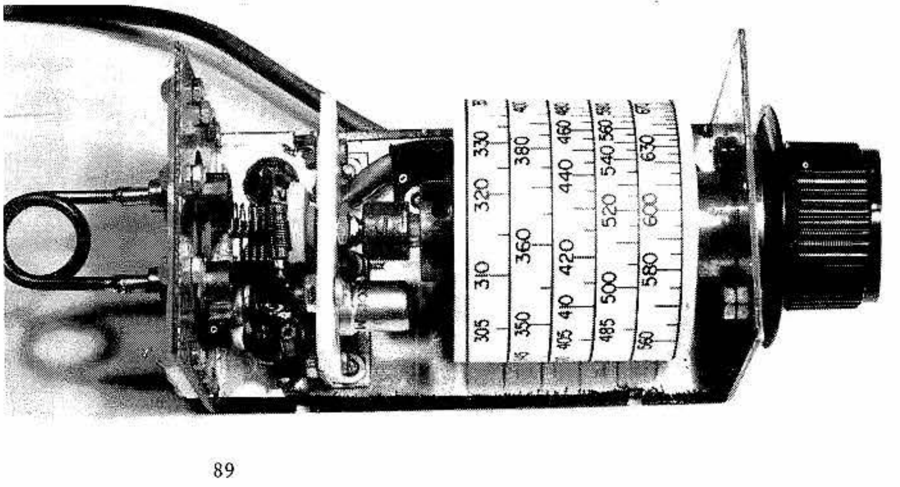
305
310
320
330
350
360
380
405
410
420
440
460
485
500
520
540
560
580
600
630

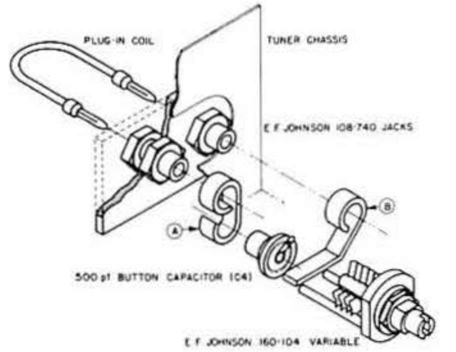
PLUG-IN COIL
TUNER CHASSIS
E F JOHNSON 108-740 JACKS
B
A
500 pf BUTTON CAPACITOR (C4)
E F JOHNSON 160-104 VARIABLE

BOTH THE BUTTON CAPACITOR AND VARIABLE CAPACITOR CONNECTING STRAPS (A AND B RESPECTIVELY) ARE MADE FROM 3/16" WIDE THIN COPPER STRIPS AND FORMED INTO SHAPE AS SHOWN. NOTE THAT THE STATOR SUPPORTING PINS ON THE VARIABLE CAPACITOR HAVE LENGTHENED SOMEWHAT FOR PURPOSES OF THIS ILLUSTRATION.

Fig. 5. Exploded view of the connections between the 6CW4 and the coil socket.

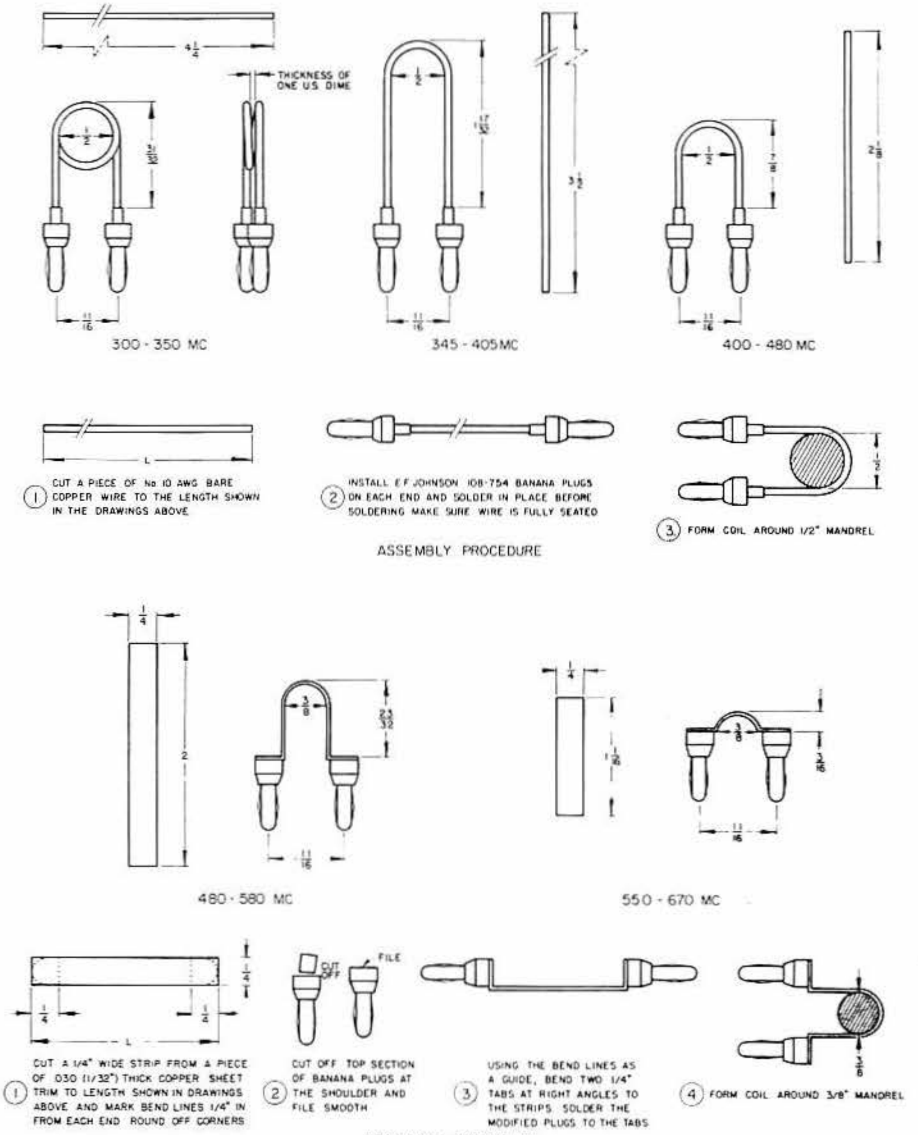

① CUT A PIECE OF No 10 AWG BARE COPPER WIRE TO THE LENGTH SHOWN IN THE DRAWINGS ABOVE

② INSTALL E F JOHNSON 108-754 BANANA PLUGS ON EACH END AND SOLDER IN PLACE BEFORE SOLDERING MAKE SURE WIRE IS FULLY SEATED

③ FORM COIL AROUND 1/2" MANDREL

ASSEMBLY PROCEDURE

① CUT A 1/4" WIDE STRIP FROM A PIECE OF .030 (1/32") THICK COPPER SHEET TRIM TO LENGTH SHOWN IN DRAWINGS ABOVE AND MARK BEND LINES 1/4" IN FROM EACH END ROUND OFF CORNERS

② CUT OFF TOP SECTION OF BANANA PLUGS AT THE SHOULDER AND FILE SMOOTH

③ USING THE BEND LINES AS A GUIDE, BEND TWO 1/4" TABS AT RIGHT ANGLES TO THE STRIPS. SOLDER THE MODIFIED PLUGS TO THE TABS

④ FORM COIL AROUND 3/8" MANDREL

ASSEMBLY PROCEDURE

Fig. 7. The coils for the UHF grid dipper.

action and repeatability in a small package. Although this unit is manufactured in England, it is available from many of the larger electronics parts houses in this country.

Substitution of similar drives should be perfectly satisfactory as long as they don't extend more than one inch beyond the front panel of the tuner chasssis.

The vernier drive is connected to the variable capacitor through a ¼ inch polystyrene shaft 1-¼ inches long and the usual "¼ to ¼" shaft couplers. Polystyrene or some other insulator must be used here because the rotor of the capacitor must be isolated from ground. Because of the space limitations inside the enclosure, the coupler at the variable capacitor end of this shaft is only one-half of a standard coupler. A standard "¼ to ¼" coupler is sawed in two and one-half is epoxied to the end of the polystyrene shaft. Save the other half; it will be used for the dial drum hub.

The 8 pf variable capacitor was designed for screwdriver adjustment and its 3/16 inch shaft must be made compatible with the standard coupler. This is accomplished with a bushing made from sheet copper. A short piece of ¼ inch wide, 1/32 inch thick copper strap is formed around the capacitor shaft and takes up the slack between the shaft and the coupler.

The drum dial in the original unit was made from the metal top of a Johnson's Shoe Shine Kit (49¢ at the local grocer's), but any similar closed cylinder 1-15/16 inches in diameter and about 1-11/16 inches long should be suitable; other diameters will void the accuracy of the precalibrated dial. The "skirt" or bottom rim is cut off the metal can at the circumferential notch and a ¼ inch hole is drilled in the center of the top. The remaining half of the ¼ inch shaft coupler that was left over from the polystyrene shaft is then epoxied in place over the hole to provide a dial drum hub. Another hole is drilled in the side of the can 11/16 inches from the top; this provides access to the shaft coupler on the rear end of the vernier drive.

The precalibrated paper dial may now be cemented in place. It's a good idea to cement a piece of white paper the same size as the dial between the dial and the drum, otherwise the label on the can will show through the paper dial. Rubber cement is recommended at this point to prevent excessive wrinkling and distortion of the dial.

When all of the dial parts are completed, they are put together as shown in **Fig. 6**. It's a little crowded in the small box, but all of the parts *will* fit. However, in order to get all of the dial machinery into the box in the right order, a correct assembly sequence must be followed. First the polystyrene shaft and coupler are inserted into the dial drum from the rear. Next insert the Eddystone drive assembly through the hole in the front panel and mate it with the end of the polystyrene shaft. Place the bushing over the capacitor shaft, attach the polystyrene shaft and tighten the coupler; also tighten the coupler at the back end of the Eddystone drive thru the access hole provided in the dial drum. Install the vernier drive mounting screws. Now completely mesh the capacitor plates, center the low edge of the dial in the window and tighten up the dial drum hub. Disassembly must be accomplished in reverse order.

The tuning head is attached to the indicator/power unit through a four-conductor cable three feet long. This cable is attached to the box with a plastic cable clamp mounted in a hole provided immediately adjacent to the oscillator support angle and exits through a rubber grommeted hole at the front end of the enclosure. There is not enough room at the rubber grommet to use a cable clamp, so the cable is epoxied to the box at this point. Before installing the cable, however, check for sufficient clearance between it and the dial drum. It will probably be necessary to route the cable along the corner of the chassis to gain enough clearance.

Power indicator modulator

The power/indicator unit is housed in a standard 9 x 6 x 5 inch utility box (Bud type AU-1040 or equivalent). The layout of this

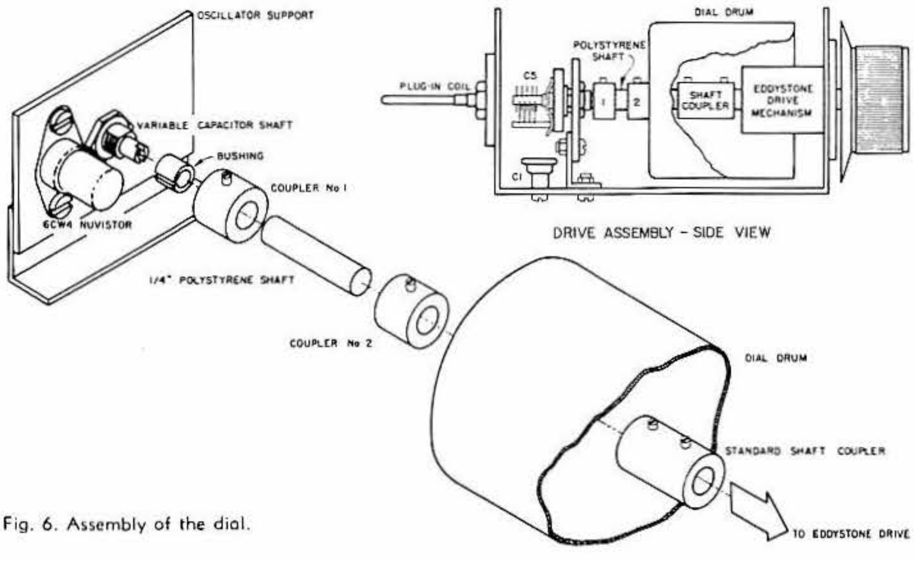

Fig. 6. Assembly of the dial.

circuitry is not at all critical, and just about anything that suits the builder may be used. The only particular caution that must be observed is with the transistorized 1 kc phase shift oscillator. This unit is built on a piece of perforated epoxy board (Vector 32AA18) 1-3/16 inches wide and 1-11/16 inches long. To preclude any 60 cycle pickup, this board is situated on the opposite side of the chassis from the power transformer.

In the author's case all the power/indicator components were mounted on a 4 x 5 x 1 inch aluminum chassis. This chassis was then mounted to the front panel of the utility box with the phone jack and power plug mounting nuts. Two large diameter holes (1-⅜ inch) are punched in the rear panel to provide access to the fuse holder and to pass the AC power plug. A chassis handle (Bud H-9168) on the top and rubber feet on the bottom just about complete the unit. There is one other addition however; five pairs of 5/32 inch holes, drilled on 11/16 inch centers along the rear edge of the top of the box provide convenient storage for the five frequency determining coils. A coat of spray paint and some Datak "Letraset" dry transfer labels are the finishing touches.

Calibration and operation

Without access to existing 420 mc equipment with known frequency characteristics, exact calibration in this band is impossible. However, this grid-dipper may be checked on the other ranges with the aid of an all-band television receiver. If the circuit layout and construction techniques described are closely followed, good correlation can be obtained and reasonable accuracy insured. The bands that are likely to be the furthest off are the lowest and the highest. In the lowest the coil spacing is quite critical and in the highest a slight change in length will move the frequency several megacycles. An accuracy of 2% at 600 mc is plus or minus 12 mc; close attention to the specified dimensions should provide accuracy better than this.

Operation of this grid-dipper is exactly the same as any lower frequency unit. It may be used in determining circuit resonance, detecting parasitic oscillations or as a signal generator. Because of its extended range, it has been found to be very useful in determining the series resonance point of rf chokes and ceramic bypass capacitors.

In the detect mode, this unit will indicate rf voltages as low as 50,000 microvolts. More sensitive operation may be obtained if it is used as an oscillating detector. In this case headphones are used and an audio beat note will be heard when the grid-dipper is tuned to the oscillator being checked. In the upper frequency ranges, it is usually difficult to obtain an actual beat note, and only a "tick" will be discernible when you tune by the frequency of the unknown energy.

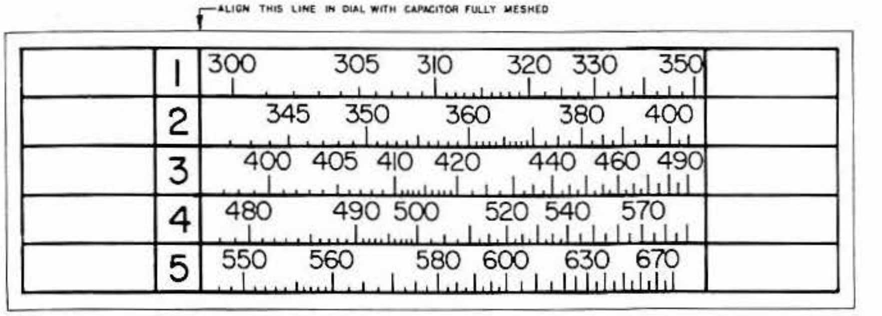

Fig. 8. Full size dial scale.

When the oscillator is tuned between 605 and 650 mc, there is sufficient second harmonic energy to provide a strong reference signal in the 1215 mc amateur band. This is particularly useful in the initial tune-up of converters for this band. Use of the internal 1000 cycle modulation aids in distinguishing the grid-dipper signal from other rf sources that are present throughout the spectrum.

WIDE RANGE VHF-UHF DIPPER

Bill Hoisington K1CLL

Most dippers for amateurs that I have seen so far, not counting the $400 ones, stop around 200 MHz just as you are about to enter the fascinating UHF region. We do have the 432 and 1296 bands, so let's become more familiar with them.

After all these years of "grid-dipping" we find ourselves without a grid, so it just becomes a "dipper". To retain the prestige of a hyphenated name we can call it a "dipper-generator". Most grid-dippers have been used as generators, but this one has built-in modulation, variable input-output coupling, controlled Q, and several other interesting features. Best of all, it goes all the way up to 1296 MHz.

When this little unit is completed it may be used as a dipper for determining the resonant frequency of VHF and UHF circuits, as an indicating frequency meter with an adjustable Q-multiplier, a field strength meter and modulation monitor, a sensitive regenerative receiver, or a CW and MCW signal generator. You can also use it as a harmonic monitor or as a frequency transfer unit from one transmitter to another.

Several circuits must be considered when building a wide band instrument such as this. For example, you should change circuits around 100 MHz and again at 600 MHz, give or take a few hundred. Below 100 MHz coils are good; from there to 600 MHz you can use ¼ wave resonators, and after that the ½ wave job becomes rapidly the best method, up to 1300 MHz.

Plug-in rf heads

I have made no attempt to cover the complete range from 130 to 1300 MHz with one oscillator. By using plug-in tuners you may vary the components to suit the frequency. On 50 MHz for example, you may use a low cost transistor, a coil, and a 25 or 50 pF capacitor. From 100 to 600 MHz you use a better transistor, a ¼ wave strap, and a 10 or 15 pF capacitor. In the microwave region up to 1296 MHz you use the best transistor you've got, ½ wave lines,

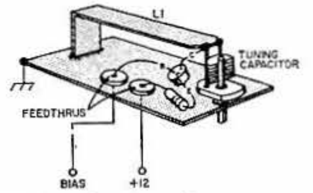

Fig. 1. Basic VHF/UHF oscillator circuit.

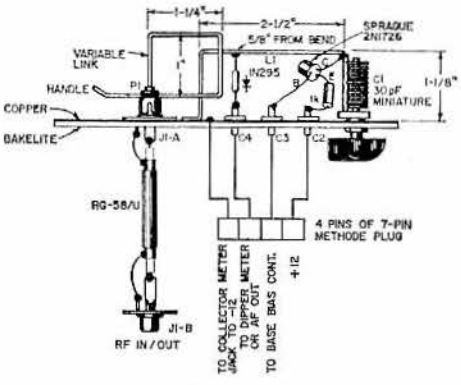
1-1/4"
2-1/2"
SPRAGUE 2N1726
5/8" FROM BEND
VARIABLE LINK
L1
1N295
1"
HANDLE
P1
C
E
B
1k
C1 30 pF MINIATURE
1-1/8"
COPPER
BAKELITE
J1-A
C4
C3
C2
RG-58/U
4 PINS OF 7-PIN METHODE PLUG
TO COLLECTOR METER JACK TO −12
TO DIPPER METER OR AF OUT
TO BASE BIAS CONT.
+12
J1-B
RF IN/OUT

Fig. 2. 130 to 300 MHz tuning head.

and a small butterfly capacitor of 3 to 5 pF.

If you break the circuit at the right point, it simplifies things—then the two halves may be connected through a miniature 7-pin socket and plug as shown in Fig. 2. All four leads are reasonably dead to rf. You can leave out some of the audio if you like, but it's very handy to have a modulated signal. If you're running triple or quadruple conversion, it's nice to know by it's modulation which is the signal and which might be a *birdie*. As far as dials are concerned—it makes calibration and reading a lot easier to have only one band or range per dial.

130 to 300 MHz oscillator

Fig. 1 shows the basic ¼ wave circuit; Fig. 2 the complete rf unit with control, af output and modulation.

The circuit itself is very simplified, as seen in Fig. 1; there being only one inductance, L1, and no choke coils. This should make for a flat tuning oscillator without power dips as it is tuned over a 2 to 1 range in frequency, and it does just that. With a 2N1726 in the circuit there is a smooth power output curve from about one volt rf at 130 MHz down to ½ volt at 300 MHz.

The rf coupling jack J1 couples the rf energy both in and out. This is because L1 acts as either a detector resonator or an oscillator resonator, as required. Actually this rf jack can be used as shown in Fig. 2. P1 is a variable link to L1 and is plugged into J1; J1A has a few inches of cable between the white ABS plastic front panel and the copper clad bakelite sub-panel. Because the phono plug is rotatable, a nice variation in rf coupling can be obtained. The coax cable and J1B get the rf out to the front panel for easy use with antennas, probes, cables, etc.

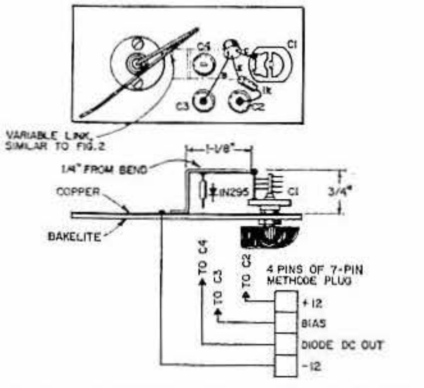
C4
C1
E
B
C3
C2
1k
VARIABLE LINK, SIMILAR TO FIG.2
1-1/8"
1/4" FROM BEND
COPPER
1N295
C1
3/4"
BAKELITE
TO C4
TO C3
TO C2
4 PINS OF 7-PIN METHODE PLUG
+12
BIAS
DIODE DC OUT
−12

Fig. 3. 300 to 600 MHz oscillator with variable link.

The emitter goes to a 1K resistor then through a coaxial bypass capacitor which gets the dc in and out and leaves the rf behind. These feed-through type bypasses are very necessary—do *not* skimp on this item.

300 to 600 MHz unit

Fig. 3 shows that this unit is essentially the same as the last, except for dimensions. I used a 2N1141 here although many others will work too. It tunes smoothly from 300 to 600 MHz; use the variable link feature as in Fig. 2.

900 to 1100 MHz

For this frequency range we need a little different approach. From Fig. 4 we can see that we now have two ½ wave lines on which low-voltage points can be found to attach the base and collector resistors. Most of the ¼ wave portion of the lines on the transistor end are actually inside the case. The places where the base resistor and the 500 ohm collector resistor are attached to the ½ wave lines can be found, or checked, by watching the rf meter and touching the lines with a pencil. At the proper point no change occurs in the rf output; sometimes it even increases.

The diode circuit of Fig. 6 is not ideal but it works. I have several of these around the shack and they work very well for detecting 1296 Mhz energy. Even ordinary hook-up wire will support the assembly of Fig. 6 about ¼ inch below the rf lines; you will soon find the best spot with the unit oscillating. The rf input jack and associated loop L3 are fastened so that L3 is in place over L1 and L2, and it's coupling can be varied in a semi-fixed fashion.

At this point we should mention that as a "dipper" the circuit is still working fine; also as a signal generator. It also serves as an rf detector but as the frequency gets up into the microwave ranges it is not quite as good as the tuned rf detectors featured in another article in 73 Magazine. Ideally, you should use the dipper on microwaves as a modulated generator and couple it into the unknown circuit; then a probe attached to *another* tuned detector should be coupled into the unknown circuit. There are quite a few variations using the dipper as an oscillator that you will find useful if you use a little ingenuity.

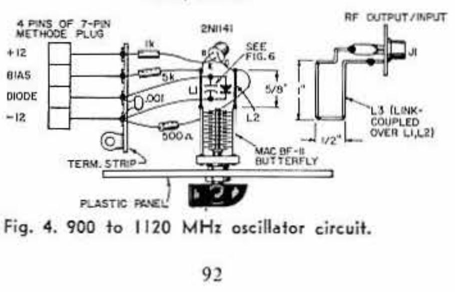
NOTE
L1 & L2 ARE THIN BRASS STRAP, 3/8" x 5/8"
4 PINS OF 7-PIN METHODE PLUG
+12
BIAS
DIODE
−12
1k
5k
.001
500Ω
2N1141
B
E
SEE FIG. 6
L1
L2
5/8"
1"
RF OUTPUT/INPUT
J1
L3 (LINK-COUPLED OVER L1,L2)
1/2"
MAC BF-11 BUTTERFLY
TERM. STRIP
PLASTIC PANEL

Fig. 4. 900 to 1120 MHz oscillator circuit.

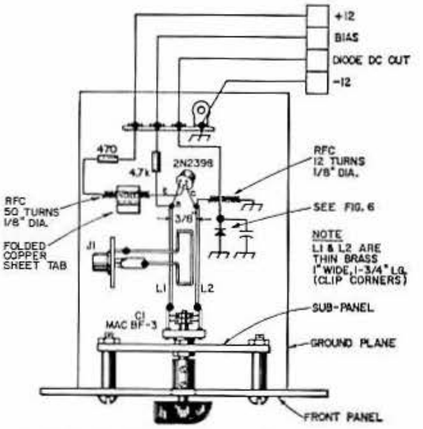
+12
BIAS
DIODE DC OUT
−12
470
4.7k
2N2398
E
B
C
RFC 12 TURNS 1/8" DIA.
RFC 50 TURNS 1/8" DIA.
3/8"
SEE FIG. 6
FOLDED COPPER SHEET TAB
J1
NOTE
L1 & L2 ARE THIN BRASS 1" WIDE, 1-3/4" LG. (CLIP CORNERS)
L1
L2
C1 MAC BF-3
SUB-PANEL
GROUND PLANE
FRONT PANEL

Fig. 5. 1200 to 1300 MHz oscillator and layout.

In the microwave detector line, my experience indicates that the plunger tuned coax cavity line is the best, the tuned trough line next, and the circuit of Fig. 5 next best. As a dipper, generator and regenerative receiver it is still good at 1296 MHz. Just to check, I plugged an antenna into J1, put an audio amplifier across the diode and copied a small transistor oscillator across the room. The base bias control works as a very smooth regeneration control. Smooth regeneration, as we will see later, is very important for maximum sensitivity when looking for harmonics and weak signals.

1200 to 1300 MHz unit

Fig. 5 shows the 1296 unit; I have used this circuit for many months as a dipper, variable-frequency generator, modulated-oscillator source, and as a regenerative receiver for 1296. In this circuit I used a negative dc grounded collector return. Don't short the base plate to the modulator base. Note that one end of the diode is tied to the base plate; this lead is brought out as the minus 12 volt lead. You can also use it ungrounded as in Fig. 4—you can use a 5th lead in the 7-pin plug and keep the diode isolated from the minus 12 volts. Suit yourself, just remember that *all* units have to use the same leads, as they all plug into the single modulator rf unit.

I just plugged a little 12 element Yagi antenna into this dipper and it works nice and smooth as a regenerative receiver. Please note, this is only for test purposes around the shack. You can hear with it, but not *that* good!

I had to put a choke in the cathode lead on this one, and tune it (the choke) with a piece of copper foil. A choke was needed in the collector lead too; after all this is the L-band microwave region.

The rf input jack J1 is mounted on a bakelite upright. Be careful of vertical metal pieces attached to the base plate; they only need to be two inches long or so to become Marconi antennas on 1296! Bring the base resistor and the collector choke away from the lines in a perpendicular fashion—it helps.

The total length of the diode and it's two leads, from ground to the tiny .001 capacitor C2 is about 1⅜ inches; it is spaced about 3/16 from the ground plane. The bottom edges of the lines are about ½ inch from the ground plane.

The transistor presently in the unit is a selected 2N2398; about half of the dozen or so I have here go to 1300 MHz, a couple go to 1400, and the rest to 1100 or 1200.

Don't be alarmed that L1 and L2 are longer than those of Fig. 5, the smaller butterfly capacitor does that. You can make a choice as to capacity, length of brass, and desired rf range. You can use a 5 or 10 pF capacitor for C1, shorter lines, and tune over 1300 MHz. In fact, I have reached 1600 MHz with this circuit!

Modulation and control

Fig. 7 shows the circuitry for bias control, modulation and audio. Don't let it scare you. It's just the same old deal of doing what has to be done for control purposes, and from then on just turning the knobs to get what you want.

I have found that a very good plug and jack can be made by using an ordinary 7 pin socket and a Methode 7 pin bakelite plug. Unfortunately, I have never found a miniature tube with a bakelite base; they are always made of glass, so you will probably have to buy the 7 pin plug.

The minus 12 volt lead goes to the collector meter jack and then to the minus 12 volt of the power supply. This puts the rf panel ground at minus 12 volts and the audio af panel at plus 12. Of course, you don't have to ground the plus 12 on the audio panel, I just have that habit.

The base return goes through a 5K resistor and then to the bias control potentiometer. The diode dc/af output goes to af jack J2 and the meter M1 through the meter sensitivity control. This potentiometer is selected to suit the meter; I have used a 10 K unit with a 500 microampere meter. Note that part of this resistor should be used when af output is desired, otherwise the meter shorts out the af.

Audio modulation

You might think that just about everyone knows how to build an audio oscillator. Mine did oscillate, but the tone! And the wave-shapes, hoo-boy! I used an af transformer with the collector on one side and the base on the other. So, once again to the handbooks and once again practically zilch. I did get the idea for a phase shift oscillator out of one of them, even if the circuit didn't work at first. After considerable experimentation I can recommend the circuit shown in Fig. 7. It works! In addition, the modulation may be adjusted to exactly 1000 MHz. This is very useful as many microwave test amplifiers have built-in narrow band af audio filters centered on 1 KHz. There is also a modulation gain control. This helps if a nice tone is desired.

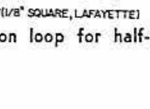

Fig. 6. RF detection loop for half-wave oscillators.

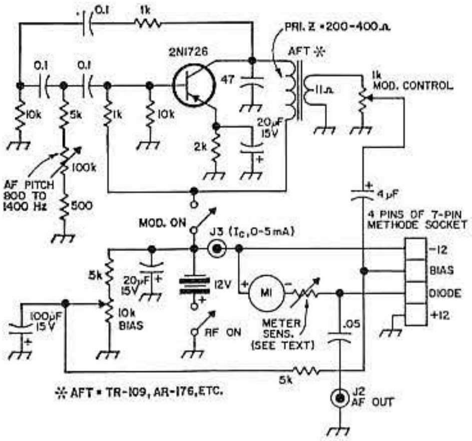

Fig. 7. Power control, modulation and audio circuitry. This circuit is used with all the rf heads.

Almost any small transistor output transformer will do the job for the af transformer, but don't go over 400 ohms impedance in the collector winding. Note the 1K resistor between the collector and the phase shift network; this reduces feedback to the base and may have to be increased or decreased depending on the gain of your transistor at 1000 Hertz.

I usually run the rf current (collector dc) between one and two mils with the 2N1726 transistors; other transistors may take more. A number 48 bulb in the collector lead may save you a $3 transistor. With the 2N1141 the rf output keeps climbing up to 4 or 5 mA collector current. You will readily find the best place to operate. When the collector mils keep climbing and the rf output starts to drop off, back off!

Operation

Dipper

As a general rule "dipping" is easier in the VHF region and gets more difficult as you go into the microwave region. On HF you couple one coil end on to the other; on VHF you bring it near, and on UHF you have to work to get the necessary coupling or use a probe.

With nothing near the dipper, swing the dial through the range to make sure it has no dips of it's own. The VHF/UHF units without chokes described here do not generally have such dips. Unwanted dips can be caused by chokes, resonant feed wire lengths, metal supports, and rf links and cables among other things.

When you do get a dip after coupling to the unknown circuit be sure and change the resonance of that circuit for a final check while watching the dipper meter. If the test circuit is a tuned circuit vary the tuning and see if the dipper will follow it—it *should*. If all else fails, use the dipper as a generator. Since it is very difficult to get the far end of a cable matched exactly over much of a frequency range, *expect* to find external dips in the dipper when using a cable.

Indicating frequency meter

Always keep the transistor plugged in and the base bias at zero so the diode is doing the work. When you advance the bias, collector capacity will cause the dipper frequency to change a little—more with some transistors than others.

For finding a weak signal you can use regeneration by turning up the base bias, but watch out for slight frequency changes. This regeneration can be very handy for finding weak oscillators or hard-to-find rf energy. Use the rf input loop with care; the least coupling is the best. Remember that some cables and terminations will detune L1.

Field strength meter and modulation monitor

The first part is obvious; use a small antenna or probe, get some signal in, and go ahead. Do not use any base bias to start with. If you are working with a very weak signal you might have to push the bias up for regeneration.

The modulation monitor is simply our often-described system of diode detector, transistor amplifier, and padded ear-phones. You can actually hear what your own transmitter sounds like to others. I use it on every new rig and after every circuit change. You don't need much of an antenna or probe when listening to your own rig; don't overload the diode when checking modulation. In fact, use light rf coupling and plenty of audio gain to hear yourself as others hear you.

Regenerative Receiver

Plug an audio amplifier into J2, Fig. 7, advance the bias control, and tune. I have heard several UHF TV stations from Massachusetts up here in Peterborough, so it is really sensitive. One nice feature of this circuit is that the regeneration turns into oscillation very smoothly. Stability is good too. You can heterodyne a crystal controlled two meter signal and copy CW with it. Not bad for a 144 MHz blooper!

To transfer signals from one transmitter (A) to another (B), just tune in A, then shut it off; listen for B and tune it in. That's all. Harmonic monitoring is easy: just tune over the suspected range in the regenerative condition. It is particularly good because only *one* frequency is present in the receiver. This is *not* the case when using a super-het receiver for monitoring harmonics.

Signal generator

One of the big features of this circuit is the presence of an rf meter right in the proper place circuit-wise. The modulation also helps, especially when running triple or quadruple conversion in a receiver. The modulation control is very convenient, at full on its spreads the signal across 20 or 30 kHz on a selective receiver. For checking a difficult to get at circuit, use a cable and probe, either capacitive or inductive, to get the signal into the unknown circuit.

I often use one of these units for antenna and receiver tests. I just plug a little two element beam into the rf jack and set it out away from the shack; often one or two hundred yards away. There is nothing like tuning up pre-amps with your antenna system connected. For antenna tests it is used in reverse.

EXTRA SERVICES FROM YOUR GRID-DIP OSCILLATOR

W. B. Cameron WA4UZM

One of the handiest instruments around the average ham shack is the grid dip oscillator, but many hams do not get all of the advantages this little instrument can offer. Most people know that it is useful for finding the resonant frequency of an LC circuit, or for generating a signal someplace in the range, although it is not quite a substitute for a well calibrated signal generator in this respect.

Many people do not realize that the grid dip oscillator makes an excellent crystal calibrator as well. The typical grid dip oscillator uses a two-terminal coil and a two-section variable capacitor, connected in the form of a Colpitts oscillator circuit, with the variable capacitor providing the voltage division for proper feedback. This same circuit, with a quartz crystal replacing the coil, becomes a very satisfactory crystal oscillator. In this case, the variable capacitor becomes a trimming capacitor capable of putting the crystal on the proper frequency with considerable accuracy and also allowing a fairly wide latitude higher or lower for checking bandpass.

The meter reading gives an index of crystal activity. What is required is an adapter to match the pins of the crystal holders to the coil socket of the grid dip oscillator. I have made up such an adapter for my particular unit which consists of a 1½ inch by 4½ inch piece of masonite which carries two banana plugs to insert into the oscillator socket, and several different sizes of crystal holders, all connected in parallel, to accommodate any crystals I happen to have around the shack. I carry in my tool kit various assorted crystals. These give me spot frequencies for band edges and other reference points on all of the amateur bands.

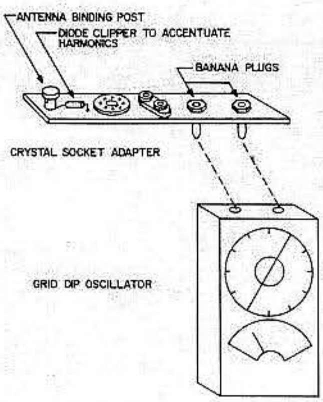

Crystal socket adapter for grid dip oscillator extends the unit's capability by enabling it to read crystal activity.

In checking the higher frequency bands, I find it useful to add a diode clipper and have one permanently wired on the same adapter board. This is a 1N69 which simply goes from one side of the crystal to a terminal point at which I can attach an antenna or a lead to the input jack of the receiver. By simply attaching a short length of wire approximating a quarter wavelength at this frequency, I can provide a strong enough signal from the shack to enable me to check out my direction finding loop in the automobile fifty feet away.

In tuning up my 2 meter FM gear, I plug in a 3061.25 crystal (which is the transmitter oscillator crystal for my Link 2m FM unit) and then couple from the 1N69 to the input of the receiver. This provides a signal stronger than most of the local signals on the band, and adequate for initial tune-up. For finer adjustment of the i-f strip, I clip the lead from the 1N69 to the chassis close to the antenna input jack and the leakage provides a signal of the order of .2 μV, which is useful for final adjustment of the critical tuned circuits, the squelch control, and others. For best results, I leave the oscillator and the receiver on for a half hour or so to warm up, and then with the variable capacitor in the grid dip oscillator, I zero-beat the incoming signal of a station generally considered to be on frequency. Once this is done, I have a signal source on the bench with which to check the receiver for passband balance as indicated on the microammeter connected to the discriminator output.

GDO TO FIND C

F. C. Rayer G3OGR

Most amateurs have a grid dip oscillator lying around, and an easy accessory can allow it to be used to read the values of unknown capacitances up to about 1000 pF. This is useful for unmarked surplus, those with obliterated markings, or to find the swing of small variables. Or we may check the best value found in some circuit position with a preset or variable, then measure this and substitute a near value fixed capacitor.

Figure 1 is the circuit. C1 and C2 are on the lid of an insulated box, carrying also spring terminals for CX. The capacitors have good knobs with pointers. Coil L can be half a dozen turns of stout wire, self-supporting, or anything which comes within a convenient range of the GDO (say 2.5–10 MHz) with both variables fully closed.

To calibrate, close C1 and C2 fully. Tune the GDO for the usual dip. Note the frequency on the box for future use. Take a few 1% capacitors, such as 100 pF, 200 pF, and so on, up to a total of about 1000 pF. Clip one to CX. Open C2 to restore the dip on the GDO. Mark the capacitor value on C2 dial. Series and parallel capacitors give more values. For example, 100 pF plus 200 pF in parallel gives 300 pF, while 500 plus 200 gives 700 pF, and so on.

Restore C2 to its fully closed mark. Repeat to calibrate C1, this time using capacitors such as 5 pF, etc., up to about 50 pF. When you see how the scales mark, estimate intermediate markings, to fill in.

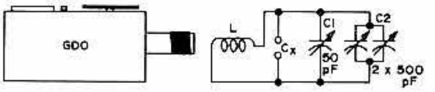

Fig. 1. GDO to find C.

Once calibration is finished, it is easy to find a capacitance value from about 2.5 to 1000 pF. Close C1 and C2. Put the GDO near L and tune the GDO for dip. Clip the unknown capacitor across CX. Open C1 or C2, as appropriate, to restore the dip. Read off the value from the scale. That's all there is to it!

GDO COIL EXTENSION

Bill Turner WAØABI

A grid dipper is a great piece of test equipment and belongs wherever rf equipment is to be designed, constructed, or serviced. All writings on the subject seem to be prefaced with this comment, and I hesitate to break tradition. Now that the formalities are over, we will proceed with the subject matter.

Probably the one most inconvenient feature of all grid dip meters is the physical size of the instrument. This holds true for the two unit commercial models as well as the more common self-contained types. It is not so much that they are large but rather that the head will not always fit into the nooks and crannies which seem to be designed into most equipment. This situation will more than likely become more severe as integrated circuits come into more general use. At lower frequencies the coil assembly is usually long enough to reach to within shouting distance of the desired circuit, but as the frequency is increased the dipper coil shrinks to perhaps an inch protruding from the case. Try getting this even near a circuit buried 2 or 3 in. deep in a chassis and you will fully appreciate the magnitude of the problem.

There is a way of relieving the difficulty which requires neither a great deal of time nor expense. It is only necessary to put to use some basic theory which each of us was required to know in order to get a ham license. Required are perhaps a 3 ft length of small coax (RG 174/U), 6 in. of hookup wire and a ballpoint pen casing. The innards of the pen are removed, and the business end reamed slightly to accept the ends of a ½ in. diameter one turn link formed of hookup wire. Inside the barrel of the pen the ends are soldered to the coax, insulated, and cemented in place. The remaining wire is formed into a 2 turn link which fits snugly on the coil form of your grid dipper. This link is soldered to the opposite end of the coax and secured. Shrinkable tubing is perfect to cover the joint.

The next time you face this perplexing problem hold your head high and fear not. Merely slip the link over the coil form, place your probe so as to allow coupling into the desired circuit and proceed as usual. The only difference will be a more or less fixed load on the oscillator which doesn't affect its ability to indicate resonance. As always, best results are obtained with minimum coupling.

Chapter VIII

Noise Generators

SIMPLE NOISE GENERATOR

George Rubis K9ONT

As anyone knows that has done any work at all on receivers, whether it is a conversion or simply substituting a "hotter" tube in the front end, we get to the point where we begin to wonder if the adaptation was worth while or have we been fooling ourselves.

A noise generator using one of the noise diodes (IN21 or IN23) can give an indication if any improvement has been made.

The circuit is straight forward, but with one addition that others that I have seen do not have. The voltage is regulated by a Zener diode.

The reason is obvious to anyone who has worked with the simpler type of noise generator. The results are not always consistent from measurement to measurement and from day to day. The voltage and current vary with the setting of the variable resistance and due to the normal aging of the battery.

The Zener Diode eliminates this by maintaining a constant line voltage. In our particular instrument it is six volts. Of course we must use a battery in excess of six volts. Nine volts is a good value. I have found that used transistor radio batteries still have enough life in most instances to last for many tests.

One of the main requisites of a noise generator is that it must be shielded throughout. Therefore we must give some thought as to the placement of the various components.

A Mini-box 2¼ x 2¼ x 4 is an ideal size. As for a connector I used the SO-239 coaxial. I find that this connector allows more flexibility than any other. If a direct connection to the receiver is desired merely attach it through the double connector type DKF-2 made by Dow Key. On the other hand if it is desired to have the controls of the noise generator close at hand merely connect a length of Coax of eighteen inches or so. I haven't been able to discover that it has affected any measurements to any degree.

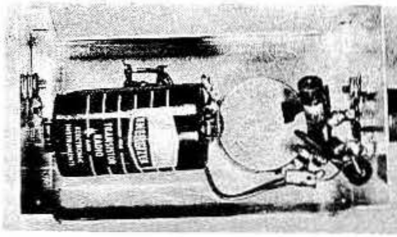

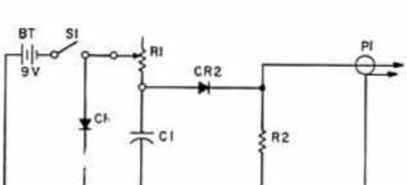

To avoid excess wear and tear on the zener diode and the battery a 200 ohm resistor should be inserted between the 9 v battery and S1.

In the construction of this noise generator just remember a few basic rules. Keep all connections as short as possible. The noise diode and bypass condenser and resistor (50 or 75 ohms as the case may be) as close as possible to the output plug. Remember to use pliers to absorb the heat when soldering the leads of the diodes.

To mount the silicon diode, which has one large end and one small, we must improvise to a certain extent. For the small end a lug from one of the old tube sockets will do. For the large end use a small fuse clip.

Don't be too fussy about the variable resistor. For most purposes any value from 10M up to 50M can be used.

The battery you choose will determine the manner of mounting.

No need to give detailed instructions as to the use of this noise generator. There are ample instructions to be found in various magazines as well as handbooks.

. . . K9ONT

Parts List

2—¼ x 2¼ x 4 Minibox
Bt—9 volt battery
Cr1—6 Volt Zener Diode
Cr2—IN 21 or IN 23 Silicon Diode
R1—10M-50M Variable
R2—51 ohm or 75 ohm (according to your line)
C1—.001 to .005 disk ceramic
S1—S.P.ST. this may be on your variable resistance
P1—So-239

A CRYSTAL DIODE NOISE GENERATOR

Karl Tipple W5TEV

Most amateurs who do very much operating at the higher frequencies where receiver sensitivity rather than QRM often determines success eventually come to the conclusion that their present receiving equipment is not as sensitive as would be desirable and that a change is in order. And once that change has been made, whether it takes the form of a new converter or modification of existing rf stages, one nearly always wonders whether the new really *is* better than the old. Comparisons of signal strength with new and old equipment are erratic and unreliable at best and do not necessarily give the desired information since the absolute output level or "S" meter reading is not the problem. What is actually of concern is whether the signal to noise ratio of the receiving system has been improved. Or in other words, does a given rf signal produce an audio output signal from the new receiving equipment that is louder in comparison to the background noise than did the same input signal with the original receiving equipment. If it does, the noise figure of the new receiving equipment is smaller (better) and it should be possible to copy a weaker signal than before.

Unfortunately few of us own or have access to a signal generator of the proper frequency range, with an accurately calibrated attenuator and shielded output, for making the type of measurements mentioned above. For the majority of amateurs something less expensive and more readily available is needed. A crystal noise generator, which can be constructed for a few dollars and which will enable anyone to make before and after comparisons of noise figure, is such a device.

Briefly, the theory behind the use of the noise generator follows. The sensitivity of a receiver is highly dependent on the ability of the rf stages to amplify a weak signal as much as possible while adding a minimum amount of noise to the signal. Noise of this type will

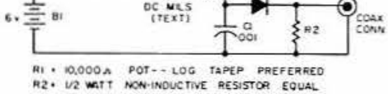

RI • 10,000Ω POT-- LOG TAPEP PREFERRED
R2 • 1/2 WATT NON-INDUCTIVE RESISTOR EQUAL
TO LINE IMPEDANCE

FIG. 1

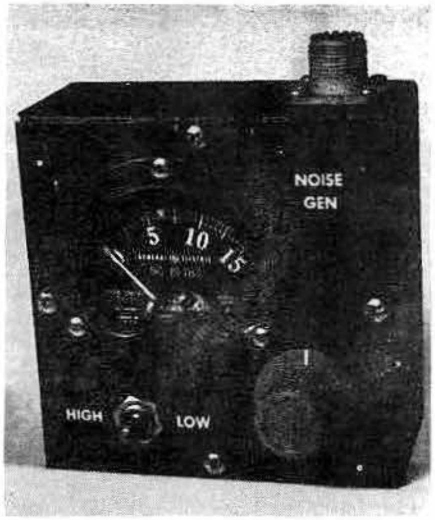

be heard at the receiver output as a familiar hiss or "rushing" sound. Now, if an external source of broadband or white noise (random noise independent of frequency) is connected to the receiver antenna terminals, an increase in noise or hiss will be noted at the receiver output. And if the noise generator output is adjusted to provide exactly twice as much receiver output noise as existed before the application of the noise generator, then the generator must be providing a noise signal equal to that generated in the receiver. Thus, if the receiver can be modified so as to reduce the noise generator signal required to double the receiver noise output, the noise figure of the receiver will have been improved.

A noise generator of the type described in this article can not be used for measurement of absolute noise figure without additional calibrating equipment. It can, however, be used for *comparison* of noise figures. It enables one to determine whether a receiver really has been improved and it can be used to determine which of a number of possible modifications will yield the best results in terms of minimum noise figure (maximum sensitivity). Use of the device will be discussed in more detail later.

A basic noise generator circuit is shown in Fig. 1. The noise is produced by CR1, a silicon diode of the 1N21 or 1N23 variety. These diodes have been widely available on the surplus market for several years. They are also readily available new at small cost; however as will be explained later, it may be desirable to use a surplus diode rather than a new one.

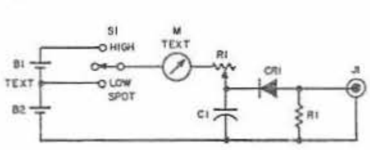

FIG. 2

R1 = 10,000 ohm pot—log taper preferred
R2 = ½ watt non inductive resistor equal to line impedance
Other parts same as Fig. 1

The current supplied to the diode from the battery and therefore the noise generated by the diode is determined by the setting of R1. The best value for R1 is necessarily a compromise since in its maximum resistance position it must limit the current to a low value but it must also not change resistance too rapidly at the low resistance end of the range or the normal operating range of the instrument will be limited to only a few degrees rotation of the R1 shaft. A 10,000 ohm log taper pot should prove satisfactory.

The value of R2 should be about the same as the transmission line impedance. A good value might be 56 or 68 ohms, although if it is desired to use the noise generator with equipment which normally operates from higher impedance lines, a larger value should be selected for R2. In this case one might compromise with a value of 100 ohms. C1 serves as bypass capacitor and may be either mica or disc ceramic.

A more elaborate arrangement was used in the construction of the unit shown in Photo. 3. A schematic of this generator appears in Fig. 2. Here a tapped battery (2 pencells) was used to minimize the range to be covered by R1. It should also be noted that the diode is shown reverse biased in Fig. 2. Many of the surplus diodes available have very poor reverse resistances, which will allow reverse bias operation, and when operated under these conditions will generate considerably more noise for the same bias current than when operated in the more conventional configuration shown in Fig. 1. The diode must have a reverse resistance of only a few thousand ohms or less if it is to work successfully in the reverse bias connection. Therefore unless the reader can obtain some surplus diodes and select one with a low reverse resistance, it will probably be necessary to operate the diode forward biased and to use the higher battery voltage of Fig. 1.

The meter range depends upon the battery voltage and the resistance of the diode in whichever connection it is used. The meter used in Fig. 2 was a O-4 ma surplus meter. However, if the diode is operated in the forward biased connection a meter of somewhat higher current rating will probably be needed. Therefore, it is suggested that a low range meter be obtained and if more than full scale current is required for the diode to develop sufficient noise, a shunt resistance can be placed across the meter terminals to extend the meter range. An inexpensive meter can be used since all that is needed is a relative indication of current.

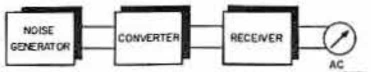

FIG. 3

For those not familiar with the IN21-1N23 type diodes it might be mentioned that they have no wire leads. The anode connection is a brass shell on one end of the unit and the cathode connection is a brass pin at the other end. Although external connecting wires might be soldered to the brass ends, there is danger that the diode will be over heated and destroyed by such techniques. It is safer to make connections by clamps made from small plate or grid caps or fahnstock clips.

It should be evident from the above discussion that there is nothing critical about the construction of a noise generator of this kind and that a number of different circuit variations are possible. Even if the device were constructed entirely of new components the cost should be less than $8 and with a little ingenuity the cost can be reduced substantially through the use of surplus parts.

Perhaps use of the noise generator can best be illustrated by considering an example of noise generator application. Suppose that you want to use either one of two available 6 meter converters with your low band receiver and that you wish to determine which converter will provide the best weak signal reception. A test set-up is shown in Fig. 3.

The ac voltmeter is used to measure the audio output of the receiver and may be either a multimeter or a VTVM. It can be connected across the speaker terminals or across the headphone line. Higher voltages can generally be obtained from the headphone output and this connection will probably have to be used if a multimeter without a very low ac range is used.

The measurements can be started by measuring the ac noise output voltage from the receiver with the noise generator turned "off". Then the noise generator should be turned "on" and the generator level adjusted until the output meter reading has doubled. At this point a note should be made of the noise generator setting. The above procedure can now be repeated with the second converter connected to the receiver. Whichever receiver-converter combination requires the smaller noise generator current setting to double the noise output is the combination that will provide the best weak signal reception.

The noise passed by the receiver is a function of the bandwidth of the receiver. Also the reading given by a particular output meter is dependent on its bandwidth. Therefore the bandwidth of the receiver should not be changed during measurements and the same output meter should be used thoughout a set of measurements.

The technique described above can also be used to determine which of a number of rf amplifier tubes will give the best performance. The noise generator can be used for alignment of rf stages.

For the man who constructs his own converters or receivers the device is a tremendous boon since such adjustments as determining the optimum antenna to grid coil coupling can be made simply by adjusting for maximum increase in noise when the generator is switched "on". (This technique gives the same results as adjusting for minimum noise generator setting to double noise output and it is a bit easier to use when adjustments must be made.) If the input circuit uses a single tapped coil instead of two coils, the best tap position can be determined.

The uses mentioned above are only a few of the possible applications of this noise generator, and for the amateur who constructs his own receiving equipment the device is indispensable.

TEMPERATURE LIMITED DIODE NOISE GENERATOR

F. L. Thomas

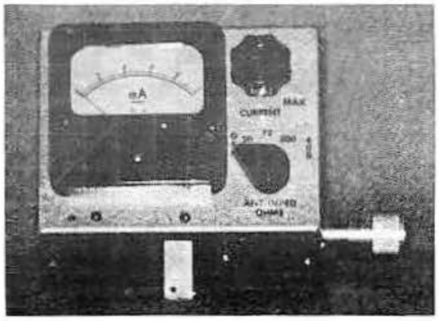

Recently I was confronted with the necessity for a good noise generator. All of the noise generator designs available in any of the amateur publications at hand were of the crystal diode type. The disadvantage of this type of instrument lies in the fact that the current through the diode has no simple relation to the noise output. Unless expensive calibration equipment is on hand this type of generator is useful only on a comparative basis. Consequently it was decided that a temperature limited diode noise generator would be built.

The noise output of a temperature limited diode noise generator is simply related to the current flowing through the diode.[1] The noise figure of a receiver may be calculated directly from the magnitude of the current by the following equation:[2]

$$\text{Noise figure in db.} = 10 \log (20\ I\ R) \quad (1)$$

Where I=current through temperature limited diode required to make the noise output power of the receiver double the value it was with no current through the diode.

R=antenna impedance

The actual noise generator is quite simple, consisting of half a 6AL5 with a milliammeter in the cathode leg and the appropriate resistance for the antenna circuit connected to the plate by means of a selector switch. The heater current is controlled by a 20 ohm potentiometer in series with the heater. The unit is battery operated for convenience. High current capacity, small sized mercury cells are used throughout. The whole unit is contained in a 3" x 4" x 5" Minibox.

Construction

The resistors are mounted on the switch, making the leads as short as possible. The tube socket is mounted on a small bracket screwed onto the side of the chassis, placed so that the distance from the switch is a minimum. The meter calibration potentiometer and the heater supply battery holders are mounted on another bracket placed over the meter as shown in the photograph.

Operation

To use the generator it is connected to the receiver antenna terminals, the AVC is turned off, and the audio output is measured with the generator off. The generator is then turned on and the current through the diode is increased until the power output is double what it was before. At this point the voltage output will only be 1.41 times the original. The current necessary to give this noise increase is noted and the noise figure calculated according to equation (1), or estimated from the figures given in Table I.

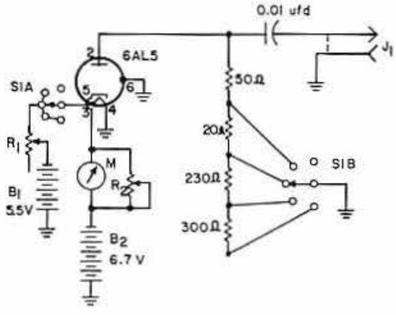

B1—four Mallory RM12 or RM12R in series (lifetime of this circuit should be greater than 10 operating hours)
B2—Mallory TR-135R
M—0-1 ma (Lafayette TM-60 or equivalent)
R1—20 ohm potentiometer
R2—20 ohm Potentiometer
J1—coax connector
S1—DP 5 pos miniature rotary switch

Noise Figure in db.	Current, ma., for antenna impedance of:			
	50Ω	72Ω	300Ω	600Ω
2	1.40	1.15	-	-
4	2.25	1.80	0.42	-
6	3.55	2.85	-	-
8	5.65	4.50	1.05	0.53
10	8.85	7.15	-	-
12	14.2	11.3	2.65	1.32
16	-	-	6.65	3.30
20	-	-	16.6	8.42

The unit described is useful for noise figures up to 13 db at 72 ohms. For higher noise figures a higher current capacity diode and a higher heater voltage and current supply is required, necessitating an ac operated power supply. With a suitable power supply a 6X4 may be satisfactorily operated up to a noise figure of about 22 db. In both cases the maximum noise is obtained by passing more than the allowable average current through the diode. Consequently readings at these extremes should be made and the current lowered again in only a few seconds, or damage to the diode will result.

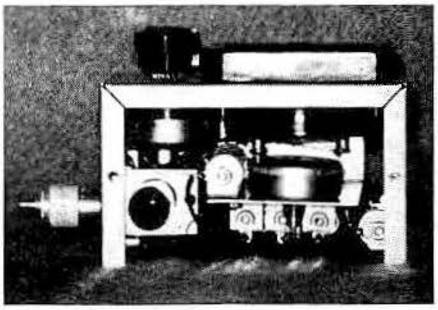

Table I shows the current readings for various noise figures with different antenna impedances.

The generator will operate satisfactorily up to at least 50 mc, and probably considerably higher.

Editor's Note: By substituting flashlight batteries for the mercury cells and a 20 mA meter instead of the 1 mA meter with a shunt across it, the cost of construction can be greatly reduced.

REFERENCES

1. Radiotron Designers Handbook, fourth edition, page 1307, reproduced and distributed by RCA Victor Division, Radio Corporation of America, Harrison, N. J., 1953.
2. Radio Engineering Handbook, fifth edition, page 19-10, Edited by Keith Henney, Published by McGraw-Hill Book Company, 1959.
3. Terman, F. E., Electronic and Radio Engineering, fourth edition, Published by McGraw-Hill Book Company, 1955.

CALIBRATED NOISE GENERATOR FOR 432

Hank Olson W6GXN

As pointed out in previous articles—the temperature-limited noise diode is the best tool for measuring the sensitivity of one's receiver. The measurement made with this device is "noise-figure" or "noise-factor" and this is the magic number by which you can compare your VHF or UHF receiver with Joe's down the street, or Sam's in Massachusetts. These measurements will be comparable between any two receivers because bandwidth, type of detector, and other miscellaneous features (different for each receiver) do not affect the measurement, if it's carefully made (that is, if one's measurement of 3 db power increase, when the diode is turned on, is true).

Past amateur articles on noise-diodes and their use in noise-factor measurement have only shown how to construct units for up to about 200 mc. The noise generators described previously, for amateur use, have used the 5722, the diode wired 24G (3C24), or the diode wired 801A, as their tungsten-filament diodes. The 15E and 01A have also been used occasionally as diode-wired triodes, too. Commercial UHF noise generators use a variety of tungsten-filament tubes, all of which are a bit expensive for most hams. If you can lay your hands on one of these tubes, used or otherwise surplus, by all means employ it in your noise generator. Table I is a list of such gems.

Philco:	L1262A
RCA:	R6212
Bendix:	6144
Marconi:	CV2171

Table I

The tube I used is a triode, a Western Electric 708A, originally designed for grounded-grid UHF amplifier service. The grid and filament are used as the diode elements, ignoring the plate altogether; the plate, if used, could only increase transit time and shunt capacity. The tube is used, as it was intended to be used, with the metal shell (grid) grounded. The 50 Ω load is connected in the filament circuit; the filament power is fed in by means of a concentric inductor, which also tunes out the stray capacity of the tube.

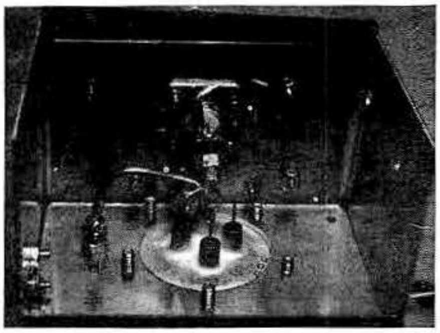

RF Circuitry.

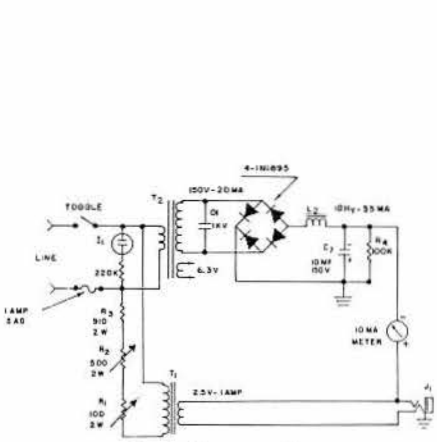

Fig. 1. Power supply.

C1, a small "tweaker," adds in a tiny additional capacity to make adjustment to 432 mc easier; it makes tuning to anywhere in the 420 to 450 mc band possible. L1 is constructed of a 1⅞" length of ⅛" copper tubing and has a piece of No. 20 teflon insulated wire inside it. The use of teflon insulated wire is only necessary because teflon will withstand the heat of soldering.

The effective circuit, then, is as in Figure 2.

At 432 mc, if $Cgk + C_1 = 3$ mmfd, then L must be 0.04 μh to be parallel resonant. The reactance of either $Cgk + C_1$ or L_1 is about 100 Ω, so the system has a Q of ½, and hence will be rather broad in its noise output spectrum—just as we want it to be.

The WE 708A tube was recently available from a Los Angeles surplus emporium at the price of 39 cents each or ten for a dollar. We bought a buck's worth, figuring some would be NG, but all were perfect and saturated well. One was lost in initial test, when we applied too much filament voltage; it was subsequently hack-sawed open to find out the details of its construction and to confirm connections. The details learned are presented in Figure 3 along with its saturation curve.

The filament is a single, fine, straight, tungsten wire through the grid helix. The grid helix is perhaps 1/16" diameter and is welded every turn to the shell. All this adds up to: good cylindrical diode configuration, close cathode-grid spacing to cut down transit time, and low grid to case inductance. In short, we have a nearly ideal noise diode for a dime a piece.

Construction details: The rf section of the generator is constructed on an aluminum plate bent into an L, the WE 708A protrudes through a 1⅜" round hole to expose its filament pins next to where the UG58A/U connector is mounted on the other side of the L. The WE 708A is held in place by five 8-32 binding-head screws that are tapped into the plate. The UG58A/U has four 200Ω, ½ w resistors soldered to it each at 90° to its neighbors to form a less inductive load, approximating a resistive sheet. These resistors are soldered to the UG58/U before it is mounted to the aluminum sheet to make soldering easier (less heat required). Then L1 is formed and soldered from the WE 708A filament pin to C3. The center filament lead of L1 is fed through and soldered to the other filament pin and to C4. C2 is then soldered in; be sure this is the type called for or a similar low inductance stand-off ceramic. The rest is straight forward. Diode and power supply were each built in an LMB 141 box chassis. The details of wiring the diode circuitry are shown in Fig. 4.

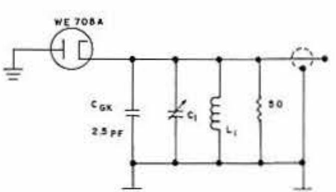

Fig. 2. Basic noise generator circuit.

A word about R1 (the "fine" adjust) is worthwhile. Make *sure* this one is a 2w type A.B. (ohmite) molded carbon pot, if not both R1 and R2. This will make smooth diode-plate current adjustment easy; a wire wound pot will cause the plate current to vary in steps because of the effect of the pot's sliding contact sequencially contacting each wire (the same applies if you use a Variac).

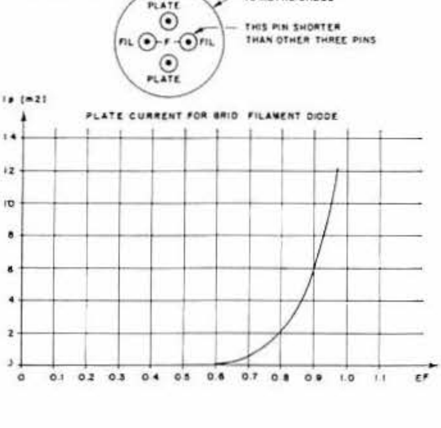

FIG 3

Fig. 3. Connections of 708A and graph of relationship between "plate" current and filament voltage.

To align on 432 mc, C2 is temporarily removed and a UHF grid dip meter coupled to L1, loosely. C1 is adjusted for a dip at 432 mc. Then C2 is reinstalled, and we should be ready for receiver checks.

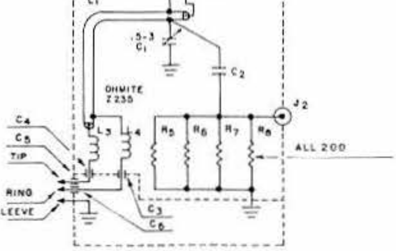

Fig. 4. Circuit of diode noise generator.

The above noise generator was compared with a Hewlett Packard 343A noise diode using my own 432 mc converter as the "to be measured" device. The results showed less than 0.5 db difference.

The author wishes to thank Gene Howell, W4RLU, for his photography of the unit.

ANOTHER WAY TO MEASURE NOISE FIGURE

Jim Kyle K5JKX

Many times it's been said, but it can always bear repeating. Noise figure is probably the least-well-known measurement in any amateur station.

At least a part of the difficulty with noise figure lies in the means usually employed to measure it. While it's not too difficult to figure out the power input to your final, or even the power *output* (often a surprisingly different figure), measurement of receiver noise figure tends to be a complicated and somewhat inaccurate process at best. It requires special equipment, and even then may be no more accurate than plus-or-minus 100 percent.

The classic means of measuring noise figure is to use a noise generator and crank in additional noise until receiver output is doubled. This means, of course, that the noise generator output is then exactly equal to the original noise, and if the noise-generator output is accurately known then the original noise is also known. All this has been gone into in detail in another article.

However, a noise generator with accurate calibration isn't so easy to come by, and an inaccurate noise generator doesn't do much good for measurement purposes (although it's fine for tune-up).

There *is* another way to do it, which is actually much more in line with amateur practice. This other way also requires some test equipment, but it might be more easy to come by.

Before we get into the details of the "other way" to measure noise figure, let's take another look at the reason for using noise-figure measurement as a yardstick for receiver sensitivity in the first place.

To start, we're really interested in the answer to the question "How weak a signal can I hear?" In the 3-30 mc range, the question can be answered directly—how many microvolts must the receiver have to give readable output?

As the vhf region is entered, though, the fractional microvolts become alarmingly small.

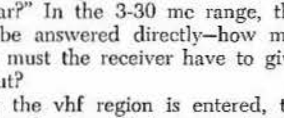

DB Loss	R1, R2	R3
3	8.2	150
6	16	68
9	24	39
20	39	10

All resistors composition, ½ watt, 5%

Fig. 1. Construction of T-pads

Somebody figured out that most of the problem lay in the receiver's own internal noise, and came up with the idea of a "perfect" receiver which would have no noise at all. This is a noise figure of 0 db. Now by comparing existing receivers to this perfect ideal, and comparing the internal noise in db, we had a way of discussing receiver sensitivity.

Since we're now talking about noise, which is equally present at all frequencies throughout the spectrum, we can see that the amount of noise present in a receiver's output is at least partially determined by how much of the spectrum we are looking at. A broad receiver has more noise output than a narrow one, all other things being equal. If you don't believe it, fiddle with the selectivity switch on your own rig and listen to the change of noise output.

This dependence of noise on bandwidth is another reason for using noise figure as a comparison. The *actual* amount of noise is cancelled out in the comparison, leaving only the relative amounts of noise in the "perfect" receiver and the receiver under test to be measured.

When all this became established, nobody was paying much attention to receiver bandwidth and it was felt that a true determination of the effective noise bandwidth of a receiver was much more complicated than the comparison measurement. However, in these days of SSB and special filters, that's not so true any more.

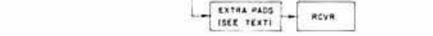

Fig. 2. Test set up.

As you may have guessed by now, the "other method" of determining noise figure depends on a microvolt measurement and knowledge of the receiver's effective noise bandwidth. The only reason for converting the results back to noise figure is to allow comparison with measurements made in the more conventional manner.

With typical ham measurement techniques, the results won't be of National-Bureau-of-Standards accuracy. However, if you're reasonably careful, results using this method will be at least comparable in accuracy to those made with a homebrew noise generator. Ready? Let's go:

You'll need two items of test equipment (only one if you're really lucky). These are an rf signal generator covering the desired frequency range on fundamental output, and an rf VTVM reasonably accurate at the desired frequency. If you have access to a "microvolter" or similar laboratory signal generator, you won't need the VTVM.

In addition, you'll need a whole handful of 50 ohm T-pads; these can easily be put together in a hurry by following the schematic in Fig. 1. You'll probably need about 8 20-db pads, as well as one each in 3-db, 6-db, and 12-db values.

Turn on both the receiver and the signal generator and let them warm up. For protection against any leakage from signal-generator to receiver through the power lines, it's best to supply them from separate circuits and to use a power-line filter such as that used to eliminate rf interference between the power line and the unit.

Connect a string of six 20-db pads to the signal-generator output as shown in Fig. 2, and adjust output of the signal generator to 0.1 volt. If you have a microvolter or equivalent, use only one 20-db pad and set generator output to 1 microvolt.

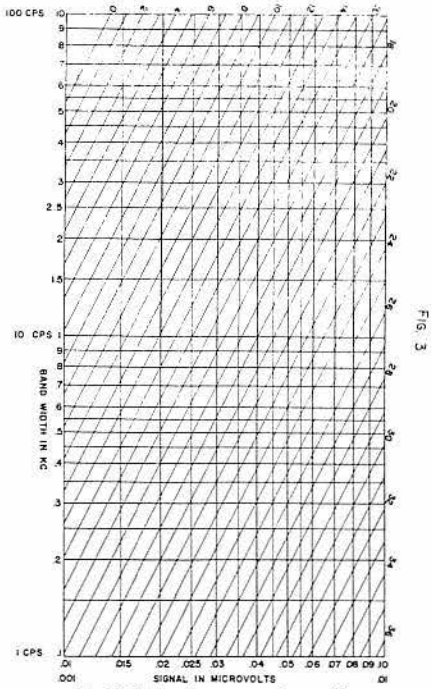

Fig. 3. Noise figure vs. microvolts.

In either event, the output of the final T-pad will be a 0.1 microvolt CW signal. This should be more than adequate for any reasonably-sensitive receiver to allow spotting of the signal.

Switch the receiver's avc off and the bfo on, and place the selectivity switch in any position for which the selectivity is accurately known. The selectivity marked on the front panel will not be the effective noise bandwidth, but you can use it as a starting point to guesstimate the noise bandwidth. If your receiver uses a mechanical filter or other device with approximately the same skirt selectivity, effective noise bandwidth will be about 1½ times the bandwidth marked on the front panel. If it is one of the older types with reasonably broad skirts, noise bandwidth will be about 3 times the marked value. Both these correction factors are approximate, of course; if you have any means of measuring effective noise bandwidth, use it instead.

For a start, use a fairly broad selectivity position; this requires more signal and makes things a bit easier.

Now tune in the signal from the generator, leaving gain controls at maximum but tuning for maximum signal strength just as if it were the new state you need. The 0.1-microvolt signal should be easy to find.

Next step is to reduce the generator output by hooking in additional T-pads until you locate the point of "minimum discernible signal." The 3, 6, and 12 db pads may be hooked up in series in any combination to give you from 3 to 24 db additional attenuation in 3-db steps. Using another 20-db pad will give you from 20 to 44 db more attenuation, and the signal is sure to become too weak to copy before you reach 44 db below one-tenth of a microvolt!

The point of MDS is approximately equal to a 0-db signal-to-noise ratio for most of us, and is considerably easier to determine than would be a true output S/N ratio. When you find this point, record the db below 0.1 microvolt and the selectivity (in kilocycles) used.

Now switch to a different bandwidth on the receiver and repeat the test. Record its results also. For maximum accuracy, repeat each of the tests 10 to 12 times and average the result.

The signal level in microvolts corresponding to db below 0.1 microvolt is given in Table I. Locate it there and move to Fig. 3, the graph of signal versus bandwidth by noise figure.

Enter the graph from the side with effective noise bandwidth, and move across until you intersect the line corresponding to signal level in microvolts. The diagonal lines are noise figure; if one passes through the intersection point, read noise figure in db from it. If not, interpolate between the lines.

In reading Fig. 3, use the 10 kc-100 cps scale with the .01-.1 microvolt scale, and the 100 cps-1 cps scale with the .001-.01 microvolt scale. If your bandwidth-signal level combination falls off the graph to the left, use the *lower* signal-level scale with the *higher* bandwidth scale and subtract 20 db from the resulting noise figure.

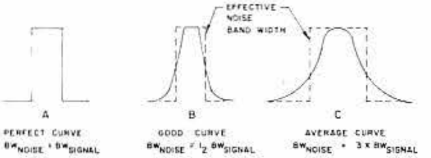
EFFECTIVE NOISE BAND WIDTH
A
PERFECT CURVE
$BW_{NOISE} = BW_{SIGNAL}$
B
GOOD CURVE
$BW_{NOISE} \approx \frac{1}{2} BW_{SIGNAL}$
C
AVERAGE CURVE
$BW_{NOISE} = 3 \times BW_{SIGNAL}$

Fig. 4. Relation of noise and signal bandwidths.

In the happy event that all your errors (and our approximations) cancel out, you'll find the noise figure to be the same at both the narrow and the broad bandwidth positions. However, it's more likely that you'll measure different noise figures at different positions of the selectivity control. It's safest to take the highest noise figure measured as being closest to correct, but you can average them if you prefer. Either way, you will probably be within 1 db of the real figure—and this is as accurate as most noise-generator techniques can be, also.

That completes the measurement, but before we wind this up let's take a more detailed look at the idea of "effective noise bandwidth" which is such a key part of this measurement technique.

Most of us are familiar with the idea of a "perfect" curve for receiver selectivity such as that shown at A in Fig. 4. Here the receiver has equal response over the desired band, and response drops to zero at the band edge. Such a curve is said to have a shape factor of 1, and is of course impossible to achieve in practice.

Now back to noise; it's spread out equally over the spectrum. A noise bandwidth of 1000 cycles per second contains 10 times as much noise as one of 100 cps. Thus "noise bandwidth" inherently has a shape factor of 1.

Since such a shape factor is impossible to achieve, it follows that "noise bandwidth" and actual receiver bandwidth must differ. If receiver bandwidth is measured at the —60 db points, the noise bandwidth will always be smaller than this receiver bandwidth. If receiver bandwidth is measured at the points where response drops 1 db below peak, the noise bandwidth will always be greater.

The mathematical expression for noise bandwidth is an integral equation involving differential gain, which is a cumbersome thing to solve. In general, the noise bandwidth of a receiver is said to be approximately equal to the bandwidth between points which are 3 db down from peak response.

In practice, if the shape factor (6 to 60 db) of the receiver is 2, the effective noise bandwidth will be approximately 1.3 times the 6-db bandwidth. If shape factor is between 2 and 10, noise bandwidth will be approximately equal to the square root of the shape factor (6 to 60 db) times the 6-db bandwidth. Few receivers have shape factors greater than 10.

db BELOW 0.1 MICROVOLT	MICROVOLTS
3	.07
6	.05
9	.035
12	.025
15	.018
18	.013
20	.01
21	.009
23	.007
24	.0063
26	.005
29	.0035
32	.0025
35	.0018
38	.0013
41	.0009

Table I.

The approximations quoted earlier (1.5 times marked bandwidth for SSB-selectivity receivers, 3 times marked bandwidth for others) are based on these relations. If you're really interested in calibrating your receiver's noise bandwidth for using this measurement technique, however, you might take a converter and have it measured for noise figure by the generator technique, then run this technique backwards to determine the effective noise bandwidth of your receiver in each position of the selectivity control.

The technique described here, incidentally, assumes that no audio filters are used following the detector. If they are, all results are off, since the effective noise bandwidth will have been changed in an unpredictable manner by the audio filters.

However, you can remove the audio filters from the hookup for measurement purposes, determine noise figure, then return the audio filters to the circuit and run the measurement backward to find out your effective noise bandwidth with filters present. Don't be surprised if it comes out in the region from 1 to 10 cycles per second; a good audio filter can work wonders with weak-signal reception.

For additional details on this technique of measuring noise figure, you can consult Reference Data for Radio Engineers, 4th edition, published by IT&T and available from Radio Bookshop, or any good radar text.

Chapter IX

Attenuators and Dummy Loads

LOW POWER ATTENUATORS FOR THE AMATEUR BANDS

George Daughters WB6AIG
Will Alexander WA6RDZ

In the evaluation of rf amplifiers, filters and many other devices, a variable attenuator is indispensable. This article describes attenuators built and tested by the authors. These attenuators are flat from dc to over 50 MHz and usable to over 450 MHz. They use low cost parts, are very simple to build, and are more accurate than ordinarily required in amateur applications.

The basic attenuator section is the symmetrical pi shown in **Fig. 1**. Resistance values are given by the relations:

$$R_1 = R_s\,(\sqrt{K} + 1)/(\sqrt{K} - 1)$$
$$R_2 = R_s\,(K - 1)/(2\sqrt{K})$$

where R_s is the characteristic impedance of the pad (equal to the source and load impedance) and K is the attenuation factor, P_{in}/P_{out}.

Resistor values for the most commonly used impedance (50 ohms) are shown in **Table 1.**

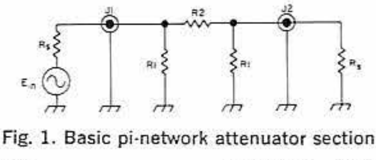

Fig. 1. Basic pi-network attenuator section.

Nominal atten- uation in dB	R1 in ohms	R2 in ohms	Calculated atten- uation in dB	Measured atten- uation in dB
1	910	6.2	1.1	1.1- 1.2
2	430	12	2.1	2.1- 2.3
3	300	18	3.0	3.2
6	150	39	6.2	6.3- 6.6
10	91	68	10.2	9.6-10.1
20	62	240	19.6	19.5-19.7

Annenuation measured at 50 MHz and lower.

Table I. Resistor values for 50-Ω attenuators.

Notice that the use of standard value 5% half watt composition resistors allows accuracy within 1 dB of the calculated value of attenuation and within 1 dB of the desired nominal value.

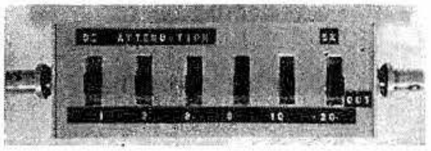

Amateur low-power attenuator made from inexpensive slide switches, 5% resistors and fiber glass, copper-clad board.

The attenuators are built in small channel boxes made of copper-clad etched circuit board material. Aluminum channel boxes commercially available would probably work equally well. Small, inexpensive DPDT slide switches (H. H. Smith No. 518 or equivalent) are soldered directly to the copper board and the resistors are soldered to the switch terminals (which have been cut short) with the shortest leads possible.

Two wiring variations have been tried, one with the series resistors (R_2) mounted between the switch wiper contacts (type B), and one with all resistors connected to the attenuator in terminals (type A). See **Fig. 2** and the photo of the interior of the attenuator. It was found that the latter arrangement, type A,

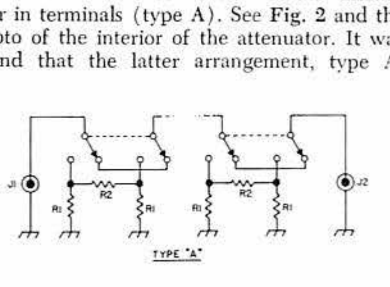

Fig. 2. Two types of attenuator construction. Type A has lower insertion loss than type B at high frequencies, so A is recommended. Resistors R1 and R2 are 5% composition, ½ watt. The switches are H. H. Smith 518 or equivalent. The connectors can be of any type to suit.

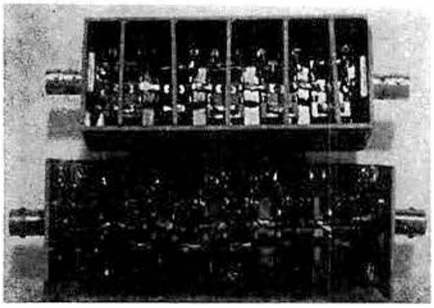

Details of attenuator construction. The top style, with complete shielding, is recommended.

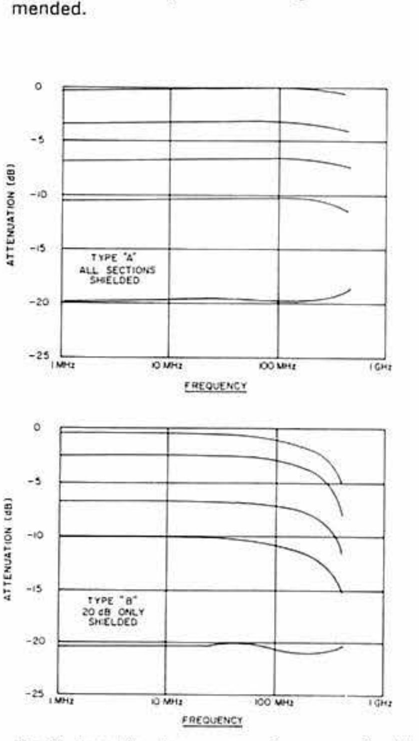

Fig. 3. Insertion loss versus frequency for the two types of attenuator construction: all sections shielded (A) and only the 20 db section shielded (B).

provided less insertion loss than the former at high frequencies, so this type of construction is recommended. Also note that dividing shields are desirable between input and output elements of a single section. These shields prevent capacitive feedthrough at the high frequencies, and are desirable on the high attenuation sections (anything over 10 dB) even at low frequencies. On the low attenuation sections, very little difference is evident below VHF. See **Fig. 3** for the attenuation of the attenuators up to over 450 MHz. Building attenuators with greater than 20 dB attenuation per section by this method is not recommended for high frequency use.

S-UNIT ATTENUATOR

Ed Lawrence WA5SWD

Since the topic of "S meters" is a popular one among radio amateurs, a lot of time is spent describing these devices, usually along the lines of how generous or "Scotch" the meters are at the QTH of the parties in the QSO. After a few such QSO's, I decided to build an attenuator, calibrated in "S" units. My aim was to attain an accuracy of 1 db or better, using 5% ½w resistors and simple construction so it would be easy to duplicate.

As a sidelight, I started out by calculating both "tee" and "pi" pads, and used "pi" because all values of resistance are close to standard values, but (especially for high attenuation pads) the values for "tee" pads can get quite small; and expensive.

I figured the values required from the tables in the Allied's "Electronics Data Handbook", page 8, 5th edition. (Allied Radio, 75c, full of good info.)

Since "S" units are supposed to be 6 db, I figured data for steps of 1,2,4 and 8 times that amount, or 6,12,24 and 48 db. With these steps, any number from 0 to 15 "S" units of attenuation could be selected. However, 8 "S" units proved to be too much for one step, as shown by the lowered attenuation at 30 MHz, due to the inherent shunt capacitance of the resistor used in the series leg, plus the stray capacitance of the switch. So I removed the 8 "S" unit step and installed another 4 "S" unit step. This allows selected steps of attentuation from 0 to 11 "S" units.

Here are the values I calculated, and the actual values used, based on 51 ohms. The steps are switched in series, as required for the desired attenuation.

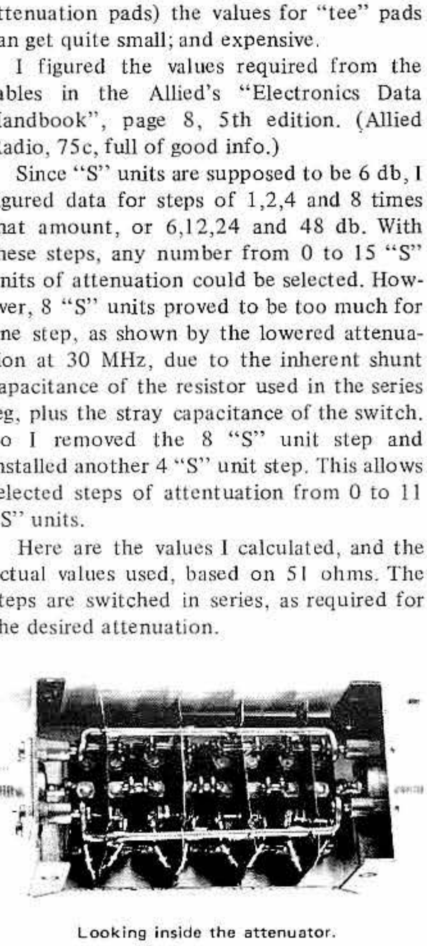

Looking inside the attenuator.

Resistance Values for 51 Ohm Attenuator:

		R1		R2	
"S" units	DB	ideal	actual	ideal	actual
1	6	154	150	38	39
2	12	85	82	96	100
4	24	58	56	405	390
8	48	51.5	51	6400	6800

After the attenuator was completed, the attenuation was measured at 3 kHz and at 30 MHz. With the test equipment available it was possible to measure more accurately at 30 MHz than at 3 kHz. Below is the data from the tests.

Atten Step "S" units	DB	Predicted atten DB	Measured at 3 kHz	Measured at 30 MHz
1	6	6.2	6.0	6.02
2	12	12.3	12.2	12.16
4	24	23.3	24.2	24.05
8	48	48.5	47.5	39.11

Now if we want to make an educated guess as to how far up we can expect good results, say 1 db error out of 24 db, then we can use the measured error in the 48 db step to calculate the capacitance across the series leg, and from that calculate the frequency where the 1 db error will occur. Go through the math if that is how you get your kicks, or take my word for it. It comes out to about 2 pF. And this will cause a reduction of 1 db at about 220 MHz. And since the resistor is of a lower value for the smaller steps, they should hold their values to even higher frequencies, but I expect other factors would get into the act along the line somewhere. I will state that still works well at 2 meters.

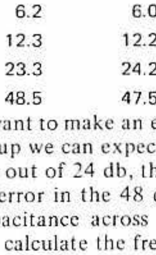

Fig. 1. Diagram for one step in the attenuator.

If you want to get fancy, you can always figure the values for 1, 2 and 3 db steps and have from 0 to 72 db attenuation in 1 db steps.

Referring to the photos; you can see I built my attenuator in a Bud Minibox CU-2102-A, 4" X 2-1/8" X 1-5/8". Four steps is the maximum in this size box, unless different switches are used. Mine are Cutler-Hammer 7592K6. The shielding was made from transformer strap, but could be any soft copper available. Try a Hobby Shop and get the thin sheet that is used for embossing if all else fails.

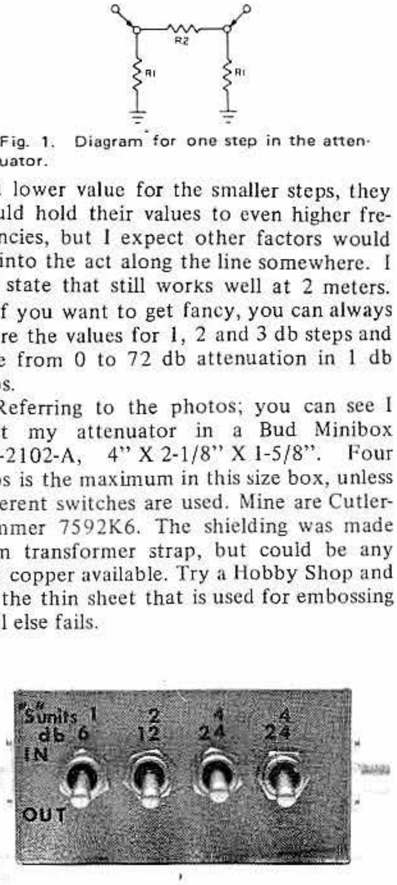

Front panel showing the switches for the various steps of attenuation.

Here are some of the uses an attenuator of this type is suited for:

Checking receiver "S" meter calibration.

Attenuating signals to aid in peaking receivers and converters.

Calibrating diode voltmeters for *rf* measurements.

Checking antenna gain. Or gain of that outboard *rf* stage.

P. S. My "S" meter lies, just as I thought!

THE EVER-USEFUL T-PAD

Jim Kyle K5JKX

Most of us have transmitters; let's hope that an equal number of us have receivers also. Antennas and microphones are usual station accessories, with a few determined diehards here and there clinging to the trusty old key. But how many of us have much in the way of test equipment?

Now and then somebody pops up with a VOM, and occasionally you can even find an operator who uses a scope. But the kind of special-purpose test gear you find in a well-equipped laboratory is almost always absent in the ham shack.

Which is more or less as it should be, since we're hams, not laboratory technicians. But with the present trends toward VHF, at least some specialized test gear is necessary. Otherwise, the regular station equipment can't be tuned for maximum performance.

One of the simplest such items is a noise generator, for getting the VHF receiver in perfect tune. This gadget has been described many times before, so we won't repeat it again —but we do have something which transforms the usual noise-generator lashup from a so-so item to an instrument capable of laboratory accuracy.

Before going into detail, let's look at the normal method of using a noise generator: you connect the generator to the antenna input, tie the converter to the receiver, turn the avc off, connect a voltmeter to the detector load resistor in the receiver (or put an ac voltmeter across the speaker leads), and measure the voltage produced by just noise. Then you turn on the generator and adjust it for a 3 db (1.4 times the voltage) increase in output; the object is to achieve the 3 db increase with the minimum amount of current flowing in the noise generator.

However, this technique of using the noise generator is pretty sloppy, since it assumes that the receiver's detector is absolutely linear for small signals—and this assumption is almost always incorrect.

A far better technique is to hook things up just as before, except now you place three T-pads in the line between converter and receiver. The T-pads on each end of the string serve merely to clamp the line impedance at 50 ohms, but the one in the center is built for precisely 3 db loss.

Now take the 3 db pad out and take your reading as before with the noise generator off; you don't have to turn off avc or hook up a voltmeter. The receiver's S-meter can be used instead, since we're not measuring anything with the meter itself. It merely serves as an indicator so we can come back to the same point.

Then replace the 3 db pad in the line, turn on the generator, and crank it up until you get the same meter reading as before. Since you now have 3 db of added loss between converter and receiver you must have increased the noise power output of the converter by that same 3 db, and you could care less about the linearity of the detector!

This whole method is far from new, but previous descriptions of it have left something to be desired in the way of telling how to build the T-pad. The gadget is so simple that it must have seemed obvious to previous writers—but it does have its tricky points too.

For instance, since a T-pad consists merely of 3 resistors, it is pretty easy to just wire them up by their leads. But they are hanging in the receiver antenna lead, and present-day receivers are rather sensitive. If you happen to find a 20-meter signal, it's going to foul up your measurements!

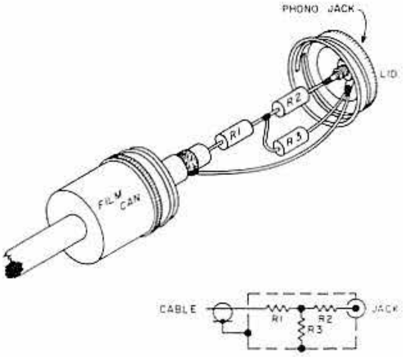

T-Pad construction details

One of the quickest ways to sidestep this problem is to shield the pad against all outside influences—but how do you shield anything so tiny?

The answer here is to use discarded 35-mm film cans, which all photographers who shoot 35-mm cameras have in abundance. The Kodak kind seem to work best. This type has a threaded cap, with a flat spot in its center just right for drilling a ¼ inch hole to take a single-hole-mounting phono jack. At the other end, a ⅜ inch hole can be punched and lined with a rubber grommet for coax cable to enter.

Next step is to place the jack in the hole in the lid, with its solder lug on the inside (be sure to clean the paint so that a good electrical contact will result). The T-pad can be assembled as shown in the drawing with shortest possible leads, and supported by its lead connecting to the jack. Thread a short length of RG-58 through the grommetted hole, and connect its center conductor to the other lead of the pad; the shield and shunt lead of the pad connect to the solder lug of the jack.

All that's left is to wire-brush the threads on the film can for good contact, and screw the lid down tight. Presto, a shielded T-pad. A phono connector should be put on the free end of the cable.

You can make up a whole bunch of these in various loss values, and get virtually any amount of loss you want by stringing them together. And this has a whole lot more use than just using with a noise generator . . .

For instance, when you want to test an antenna, arrange for a steady signal, strong enough to register on your S-meter with the back of the antenna pointed at it. Then swing the antenna in small steps, and bring the S-meter back to the same reading by inserting additional loss between converter and receiver with the T-pads. The difference between this and S-meter indications may amaze you.

Or if you are called upon to measure the difference in signal strength between two stations, the same approach can be used. Note the S-meter reading of the weaker, then knock the stronger one back to the same reading by putting T-pads in the converter-to-receiver line. Read the db off the pads and add them up.

You can even use this in place of an S-meter if you really want to know the signal-to-noise ratio of an incoming signal with accuracy; take a reading on noise, then knock the signal back to the same point. Total up the db, and there's your answer.

Though resistance values in the pad must be precise for absolute accuracy, the pad has an inherent tolerance of small errors and you should have better than 2 percent accuracy if you use 5-percent resistors. Specifically, a 10-percent error in the resistance value of any one arm produces less than ¼ db error in the pad loss, and less than 4 percent error in impedance. Using 5-percent resistors would, of course, cut these error limits in half.

If you want to follow the approach of using a string of these pads for all purposes in the shack, it's best to make them up on a "binary" approach since this gives you the maximum number of db values with the minimum number of parts. A basic assortment might be two 1 db pads, and one each of 2 db, 4 db, 8 db, 16 db, and 32 db. Using them in series in various combinations, you can get any whole number of db from 0 to 64, which pretty well covers the range of values you may ever need. For an example, to get 50 db you would use the 32 db pad, the 16, and the 2. For 60 db, you would use the 32, the 16, the 8, and the 4.

If the ability to increase loss in 1 db steps seems a bit exotic to you in view of the fact that 3 db is only half an S-unit, then you can use a binary progression in 3 db steps; this takes one 3 db pad, one 6 db, one 12 db, and one 24 db. The range is from 0 to 45 db, 3 db at a time.

For clamping a line's impedance, it's a good idea to use at least a 3 db pad and a 6 db unit might be even better. The lower-loss pads pads may not have the ability to swamp out impedance variations on their other sides.

About all that's left to make this complete is a chart of resistance values for various loss figures. Here it is; all are for use with a 50 ohm line; to use at any other impedance, multiply these values by the ratio of the new impedance to 50 ohms:

LOSS AMOUNT	R1, R2	R3
1 db	2.7	430
2 db	5.6	220
3 db	8.2	150
4 db	11	100
6 db	16	68
8 db	22	47
12 db	30	27
16 db	36	16
24 db	43	6.2
32 db	47	2.2

COMBINATION DUMMY LOAD/ ATTENUATOR NETWORK

John Schultz W2EEY

There are many instances when it is desired to use an existing transmitter as an exciter unit for a high-power linear amplifier. Many such linear amplifiers require a drive level that is only a fraction of the transmitter's output. To some degree, the transmitter can be detuned in order to reduce its output level, but this procedure is rarely possible when several orders of magnitude reduction in the power level are necessary. In such a case one can either internally modify the transmitter for a lower output level or use an alternator network between the transmitter and linear amplifier. In the latter case, the transmitter can be operated at its normal power input level and with its tuning controls at their normal settings.

The unit described in this article functions as both an *rf* attenuator and as a dummy load. The latter capability allows a transmitter to be properly tuned alone for correct operation before it is used to drive a linear amplifier. An optional wattmeter circuit is included which when calibrated allows direct reading, in watts, of the full transmitter output or of the drive level supplied to the linear amplifier.

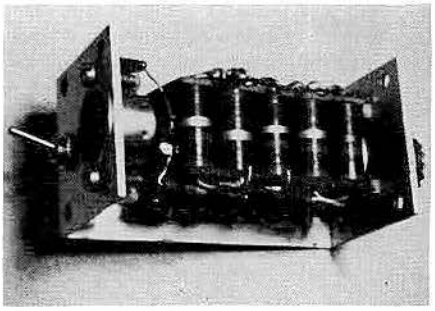

A simple method of construction is employed. Based mainly on "sandwiching" the resistors used between two pieces of vectorboard. Details are given in the text. SO-239 is used as coax input connector. The circuit function switch is located below the output connector.

The unit described was built for use with a nominal 100 output transmitter used primarily for SSB service. The construction used, however, can be extended to other power levels for transmitters operating on 80-10 meters. Also, using the information

Power Reduction	Resistor Factors a	Resistor Factors b
1/4 (6db)	.3	1.3
1/5 (7db)	.4	1.1
1/10 (10db)	.5	.7
1/20 (13db)	.6	.5

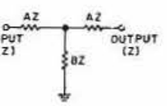

Fig. 1. Approximate resistor factors for "T" network attenuators over the ranges normally desired for exciter power output reduction.

supplied, the same type of attenuator/dummy load can be designed for other than 50 ohm transmission line systems. The attenuator was not designed as a precision network in order to allow the use of inexpensive resistors. However, the attenuation characteristics are quite satisfactory for the intended usage.

Besides its application as a power reducer when driving a linear amplifier, the unit can be used with a transmitter whenever a quick, known level of power output reduction is needed for operating purposes, approximate gain measurements, etc.

Circuit

Fig. 1 shows the circuit values for a generalized T network attenuator that can be used in any impedance unbalanced transmission line. The scaling factors are only shown for those power reduction levels most likely to be needed when driving a linear amplifier with a 75-200 watt transmitter, in order to avoid unnecessary detail. Factors for intermediate power reduction values can be found by interpolation to a satisfactory degree or one can consult an electronics handbook. The basis of the attenuator/dummy load network is to find the combination of resistor arms that will provide the desired attenuation and still be able to be connected together to form a dummy load of the correct value. Fig. 2 shows one possible combination. Each resistor bank has a value of about 20 ohms (5 resistors of 100 ohms each in parallel). In one position of the DPDT switch, the resistor banks are formed into a "T" network attenuator. In the other switch position, all three banks are placed in series as a dummy load connected across the input only. The resistance values which result are not exactly those shown in Fig. 1 for a 10db alternator. However, they are close enough to be effective and some tailoring of the individual legs is possible since each of the resistor bank values vary by a few ohms due to the tolerance of the resistors used. An optional voltmeter circuit is also shown in **Fig. 2** connected to one pole of the DPDT switch. It can be used as a relative power output indicator or if calibrated, as described later, actually measure the power output of the transmitter and of the attenuator.

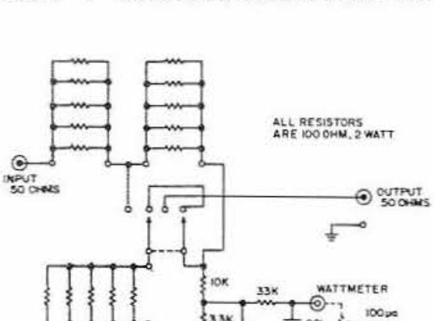

Fig. 2. Circuit of one possible dummy load/attenuator network providing about 10 db power reduction. Optional wattmeter circuit is also included.

Many variations of the basic idea are possible. **Fig. 3** shows the use of four banks of 100 ohm resistors. All four are used to form an attenuator that comes reasonably close to the values required for 7db attenuation in a 50 ohm system. Only three are used in series for the dummy load function. In this case only a simple SPST switch is necessary to disconnect the output. The same *rf* voltmeter circuit as used in **Fig. 2** may be added if desired. The switch, in fact, could be eliminated entirely if one were willing to disconnect the output termination in order to use the dummy load feature.

Whatever combination of resistance banks are used in order to achieve a desired attenuation value and the correct dummy load resistance, care must be taken that each resistance bank has sufficient power dissipation capability. The dissipation in each leg of the "T" network varies according to the attenuation level and can be calculated by Ohms Law. In general, a continuous power rating for a resistor bank equal to about one-third of the SSB peak power rating seems to suffice, including for quick tune-up on CW. For keyed CW service, the power rating should be increased to at least one-half the key-down power level.

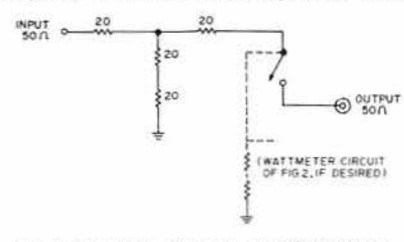

Fig. 3. Another dummy load/attenuator configuration possible with the 100 ohm resistor banks. It provides about a 1/5 power reduction (7db) when used as an attenuator.

Construction

The approach of using a relatively large number of 2 watt composition resistors is far less expensive than using specific value *rf* non-inductive resistors of 10-30 watts power rating. In quantities of more than 10, IRC type RC-2, 2 watt, 10% tolerance resistors cost about 9 cents each. So, one can achieve a 40 watt unit for less than $2 resistor cost. Banks composed of these resistors work well up to 30 mc as long as the interconnecting leads are kept short.

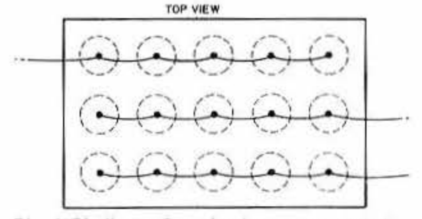

Fig. 4. Similar resistor banks are connected together on the underside of the assembly.

The photograph shows the construction used by the author for the circuit of **Fig. 2.** Similar construction can be used for larger size units as well. As shown in the photograph, the 15 resistors in rows of 5 each are sandwiched between two 1-1/8" x 1-7/8" pieces of vectorboard. None of the resistors physically touches. The wiring is done using the resistor leads. This construction is somewhat compact to expect full, continuous power dissipation from the unit but suffices for intermittent use. The frame measures 2¾" x 1½" x 1½". A cover is not absolutely necessary since the minor radiation that takes place is not important in this application. If a cover is used, it certainly should be of a perforated type to allow maximum air flow. A SD-239 connector is used at one end of the frame for the input. A dual connector is used at the other end, but normally one would use two RCA type phono jacks–one for the output and one for a meter circuit. The switch is located immediately below the output connector–a miniature Alco MST type.

Calibration

If it is desired to calibrate the voltmeter circuit as a wattmeter, it is necessary to use a probe and VTVM. Using the unit as a dummy load, the *rf* voltage is measured at the input and the power calculated. The 50K ohm potentiometer is used to set the meter at full scale for the highest power level used. The *rf* voltage is measured and the power level calculated in order to calibrate the meter for lesser power levels leaving the potentiometer at its "set" value. The same procedure is followed to calibrate the meter for the output power level by measuring the output *rf* voltage when the unit is used as a "T" attenuator and connected to a regular dummy load. The calibration should be made on the lowest frequency band used and rechecked on the highest frequency band used. If the readings differ significantly on the highest frequency band from those established, it may be necessary to add a few mmf capacitance across the diode in the voltmeter circuit in order to compensate for the slightest reactance present in the circuit.

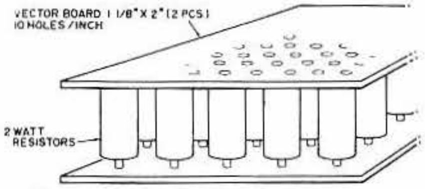

Fig. 5. Sketch showing details of resistor "sandwich" assembly.

Operation

When used between a transmitter and the 50 ohm input of a linear amplifier, the unit is first used as a dummy load for tune-up of the transmitter. The unit is then switched (with the transmitter unkeyed) to its attenuator position. In most cases, no returning of the transmitter should be necessary unless the input of the linear amplifier is particularly reactive.

Summary

The unit described is not intended as a precision attenuator or power measuring device. However, it will perform very well for its intended applications and costs far less than more sophisticated units performing the same functions.

THE MINICAN

Sam Kelly W6JTT

The Heathkit "Cantenna" has proven to be a major breakthrough in the field of dummy loads for ham use. Unfortunately, it isn't the most convenient thing to use on a small work bench with low power rigs! Borrowing their idea, I built this load for use with transmitters in the 5 to 15 watt range. The parts are few: a Campbell soup tin can, four one watt resistors, a UG-254-A connector, a short piece of 5/16 inch brass tubing and transformer oil.

Fig. 1 is a sketch of the assembled unit. The 50 ohm resistance was made up of three 15 ohm and one 5 ohm one watt carbon resistors.

First sand a can lid from a larger size can until it is free of paint. Drill a ½ inch hole through the center of the lid, and a ¼ inch hole on the perimiter. Mount the coaxial connector through the center hole. Solder a 1 inch length of 5/16 in. brass tubing over the ¼ inch hole. Solder the resistors as shown. Center the lid on the can and solder the lid to the can. Use a file to remove all rough edges. Mask the connector with masking tape and paint the can to prevent rusting.

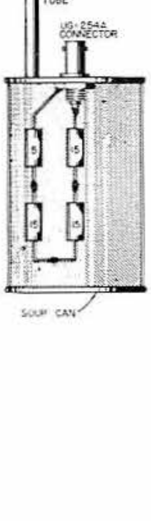

Fig. 1. Construction of the Minicon. Main parts are a soup can, coax connector and resistors.

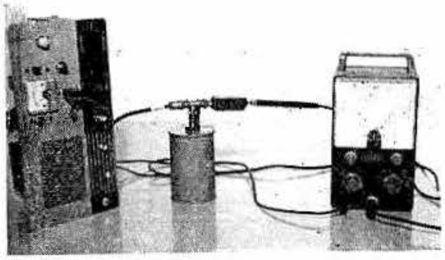

The Minicon in use with the companion detector.

Fill the can with transformer oil. A good grade of automatic transmission fluid was used in one load with no degredation in performance. However, it probably is not advisable as the fluid has a relatively low ignition temperature and might create a fire hazard.

The load was tested by running it for five hours with an input of 15 watts of 50 MHz rf. The can became warm, but the resistors showed no signs of overheating.

A maximum VSWR of 1.5:1 was obtained at 234 MHz. The measurement was made at this frequency because an automatic test set was available.

A companion rf detector unit shown in **Fig. 2** was built in a two inch section of ¾ inch square extruded brass stock. A Dage 394-1 BNC connector is mounted on one end for connecting to the RF circuit, while the DC output to the VTVM is a tip jack.

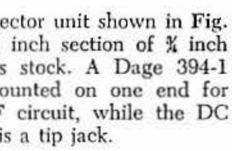

Fig. 2. Rf detector for use with the Minicon.

THE TINY TERMINATOR

Jim Kyle K5JKX

HAVE you ever felt the need for a dummy load which would also indicate accurately the actual rf output power of a transmitter?

You can buy such a beastie, you know. The Bird *Termaline*, standard of the two-way communication industry, shouldn't set you back much more than a C-note . . .

But you can also build one at considerable less outlay, and that's what this article is all about.

As the photographs show, there's almost nothing to the "Tiny Terminator"—nothing, that is, but a 50-ohm load for any transmitter, which will absorb 20 watts for days on end and will handle 40 watts for brief periods (a minor modification can double these ratings), will *not* radiate rf into the ether after the fashion of the "standard" light bulb load, will allow accurate measurement of output power up past 225 mc, and can be built in less than two hours for less than $5.

You can see by the schematic that there's nothing to the device electrically; the tricks are all mechanical. Electrically, the Tiny Terminator consists of 20 (that's what we said) 1000-ohm 1-watt resistors connected in parallel, with a crystal diode running from the hot end of the composite resistor to the meter jack and a 27 mmfd capacitor bypassing the meter.

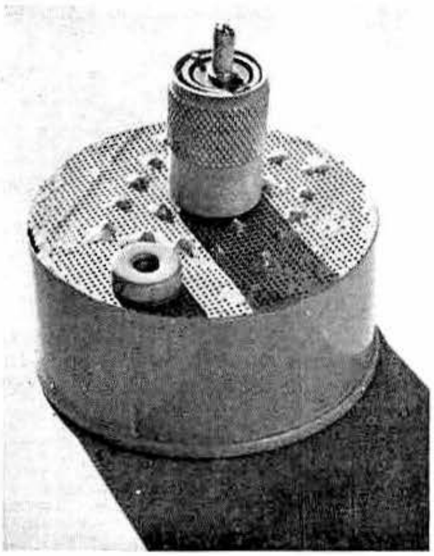

Mechanically, the big problem with a unit such as this is the problem of keeping the load resistive regardless of frequency. Resistor leads have inductance (about 25 millimicrohenries per inch) so we get rid of the leads as completely as possible. This is done by punching holes in a flat plate, passing the lead through the hole until the resistor is flush with the plate, and soldering rapidly with a hot iron (an Ungar 47-watt 1100-degree tip was used in building the model shown) so that the resistor body won't cook before the leads and plate are joined.

The flat plate has capacitance. This is avoided by keeping the plate shielded from outside influences, and separated as far as possible from its shield.

The shield, incidentally, is the bottom of 1½ inches of a Canada Dry cola, soldered all around the edge to the front resistor plate. This confines all rf inside the terminator.

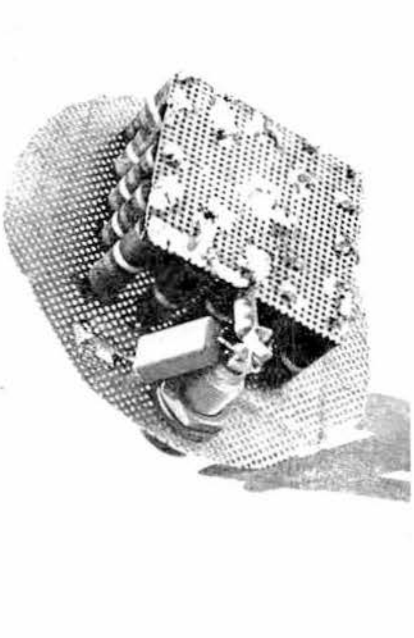

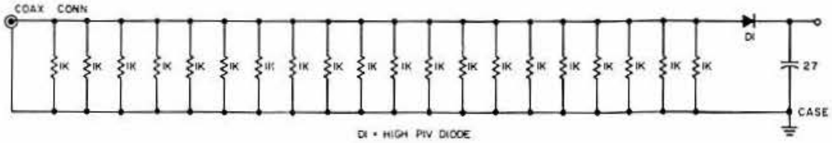

Addition of the crystal diode (a 1N34 was used, but a higher-voltage unit is recommended if you ever expect to measure power higher than 8 watts) and the bypass capacitor provide the power-measuring feature. These two components, in conjunction with an external voltmeter of at least 1000 ohms per volt sensitivity, provide you a peak-reading ac voltmeter. The voltage indicated on the meter will be equal to the peak value of rf voltage present across the load. Squaring the voltage and dividing by 50 (the resistance) will give you the *peak* power. Most power ratings are in rms values rather than peak; multiplying the indicated voltage by 0.707 before squaring will give you the rms power output.

The unit shown in the photos used perforated brass stock for the resistor plates; this happened only because a length of the brass was on hand in the junkbox at the time. A cut-out tin can lid will work equally well and will be much less expensive.

Not visible in the photos is the means of connecting the hot end of the resistor plate. A hole was punched in the middle of the hot plate and a length of No. 14 bare wire was soldered in, then filed off flush on the rear side. The bare wire was threaded through the coax connector and soldered to the center contact after checking for possible shorts.

The coax connector itself is held in place by a solder joint all around the rim of the cable-end aperture (see photo). This joint, if well made, provides plenty of strength.

Earlier, we mentioned that a minor modification would double the power rating. That modification is this: instead of 1-watt resistors, use the 2-watt variety. If you really want to go high-power, use 50 2700-ohm 2-watt resistors and have a terminator which will absorb 100 watts continuously and 200 watts ICAS. However, one that big will cost more—and you'll probably have to use a coffee can for the shield instead of the cola can used here.

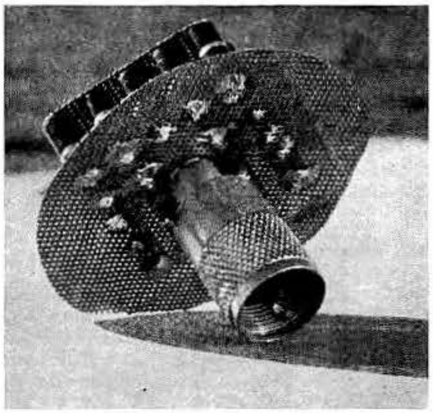

1

Basic Principles of Radiofrequency Identification

CONTENTS

1.1 Basics of RFID 1
1.2 Passive versus Active RFID Systems 5
1.3 Functional Classification of RFID Transponders 6
1.4 Applications and Frequency Selection 7

Although the widespread use of radiofrequency identification (RFID) is recent, the technology itself is several years old. RFID was first developed during World War II in order to distinguish between friend and enemy aircraft. Several years later, thanks to the technological advances in microelectronics, wireless communications, and computer networks, RFID has come of age. It has now become a technology that is mature enough to be mass-marketed at competitive cost and become a critical player in the global market. This chapter introduces the reader to the different parts that make up a RFID system and provides an overview of the different concepts that will be described in depth in the upcoming chapters.

1.1 Basics of RFID

The basic function of an RFID system is to automatically retrieve the information that has been previously inserted into specific integrated circuits, as seen in Figure 1.1. RFID systems were first developed in automatic identification laboratories as a natural evolution of the different technologies they use. Different automatic identification systems use different methods to transmit the identifying information. For example, bar code technology uses light as the transmission media, while RFID systems use radio waves. In this context, the two main elements of RFID systems are the devices used to carry this information and the equipment used to automatically capture or retrieve the information.

The devices that store and carry the information are called *transponders* or *tags* and are shown in Figure 1.2. Although the industry commonly refers to them as *tags* mainly due to their physical shape and the fact that they are

© 2011 by Taylor and Francis Group, LLC

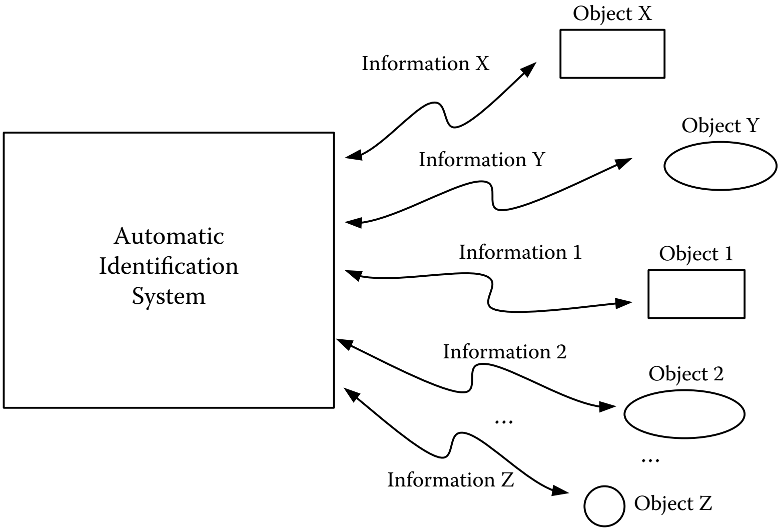
Object X
Information X
Object Y
Information Y
Automatic Identification System
Information 1
Object 1
Information 2
Object 2
...
...
Information Z
Object Z

FIGURE 1.1
Information transfer in an automatic identification system.

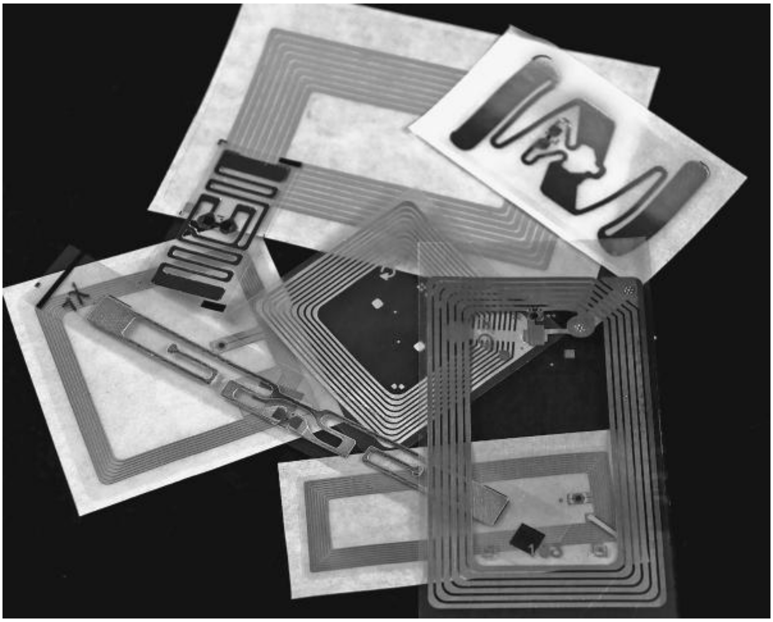

FIGURE 1.2
Several types of transponders used in RFID systems.

© 2011 by Taylor and Francis Group, LLC

mostly used to tag pallets or cases of goods, the name *transponder* reflects better the function of these devices.

The device that is used to capture and transfer information is commonly called a *reader*, because in earlier RFID systems they were only able to read the information sent by the transponders. However, with the development of new functions and applications of RFID systems, the name *interrogator* reflects better the function of this subsystem. Therefore, this book will use the names *transponder* and *interrogator* when referring to these elements.

Transponders must have the circuitry needed to harvest power from the electromagnetic fields generated by the interrogator, the necessary memory elements, as well as the different control circuits. The simplest transponders contain only read-only memory (ROM), while more sophisticated transponders also include random access memory (RAM) and nonvolatile programmable read-only memory (PROM) or electrically erasable programmable read-only memory (EEPROM). ROM usually contains the identification string for the transponder and instructions for its operating system. RAM is normally used for temporary data storage when the transponder communicates with the interrogator. PROM or EEPROM is normally used to incorporate additional functionality depending on the application. The memory capacity of transponders ranges from a single bit to several kilobits. Single-bit transponders are typically used in retail electronic surveillance in which there are only two possible states: *article paid* and *article not paid*. Memory sizes of up to 128 bits are enough to store the individual transponder identification number with several check bits. Memories up to 512 bits are normally user programmable, in which in addition to the individual identification number the memory can hold additional information required by its application. Higher capacity memories can be seen as carriers for the transport of data files. They are also used in applications in which there are several sensors attached to the transponder.

Interrogators have vastly different complexity levels depending on the type of transponders they support as well as their specific purpose. In any case, they all must provide the basic functionality to communicate with the transponders, first by energizing them and second by establishing a communications link. The complexity of the communications link can also vary considerably depending on the desired reliability. The reliability of the communications link between transponder and interrogator can be enhanced by adding parity checks, error detection, and error detection and correction. However, the use of these schemes will result in a lower transmission rate. Figure 1.3 shows a picture of a basic interrogator in which it is possible to see the coiled antenna and some of the electronic circuits.

RFID systems contain more elements in addition to the transponder and interrogator. First, the communication between transponder and interrogator is not possible—or becomes extremely deficient—without the appropriate radiant elements that will transfer the information between these two subsystems in the form of electromagnetic energy. Both transponder and

© 2011 by Taylor and Francis Group, LLC

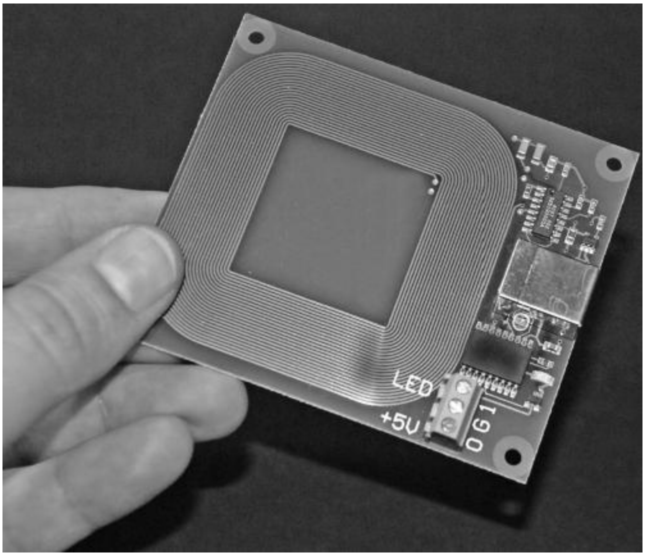

FIGURE 1.3
Basic RFID interrogator operating in the low-frequency (LF) range.

interrogator need to use the appropriate antennas to transfer information. As will be discussed in Chapter 2, the antenna dimensions in commercially available transponders and interrogators are much shorter than their ideal dimensions. This results in seriously limiting the transfer of energy between interrogators and transponders.

In addition to the transponders, the interrogators, and their antennas, the RFID system requires a *host computer* connected to the interrogator. This host computer provides a certain level of intelligence and acts like the interface between the RFID system and the ultimate application. Therefore, the interrogator must have the means to at least perform basic communication functions with the interrogator. A large number of today's applications require more than one interrogator running on a network rather than the single, stand-alone interrogator that was typical of earlier systems. Finally, an *edge server* is normally used between the host computer and the network in which the application is running.

In addition to these elements (shown in Figure 1.4), most of them hardware based, a typical RFID system contains several software elements. The *firmware* is the software that runs inside the interrogator in order to provide it with its basic functions. Some interrogators may also run one or more software *applications* that enhance the versatility of the interrogator. The name *middleware* encompasses the software applications that run in the background of the system, typically after the host computer. Finally, most applications require the use of one or several *databases* used to link the stored message

© 2011 by Taylor and Francis Group, LLC

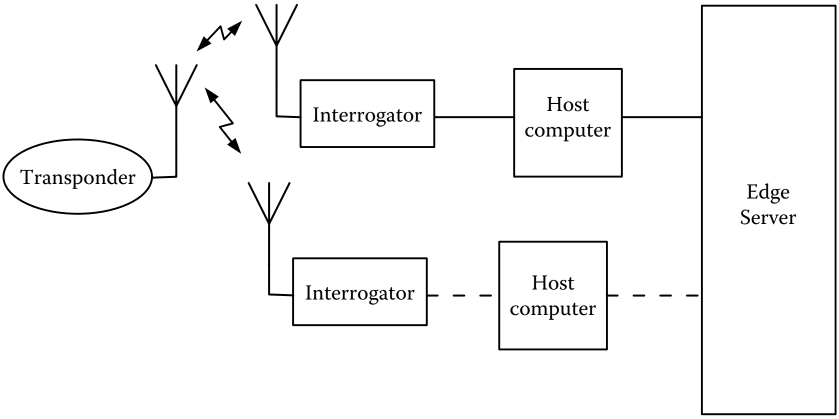

FIGURE 1.4
Basic structure of an RFID system.

sent by the transponder with more detailed information that can identify the transponder and provide additional information.

1.2 Passive versus Active RFID Systems

Passive RFID systems are those systems that use passive transponders. Passive transponders do not have an internal power source. They harvest the energy needed by their internal circuits from the electromagnetic field generated by the interrogator. For this reason, they have a short range, limited to a few feet and often, more realistically, to a few inches. Because they don't have an internal source of energy, the user does not need to worry about the status of the battery. Furthermore, their manufacturing and production costs are very low. The RFID transponders and systems described in this book are all passive.

Active RFID systems, on the other hand, use active transponders. Active transponders have an internal power source, typically a battery that allows broadcasting the signal to the interrogator. Because of not being limited to the power harvested by the antenna, they have an extended read range, typically several hundred feet. However, the inclusion of the power source increases their cost in two ways: the cost of the battery itself as well as the maintenance costs required to check the status of the internal power source and replace it when it has reached an unacceptably low level. The cost of an active transponder is approximately 100 times higher than the cost of a passive transponder. Active transponders typically operate in the ultra-high-frequency (UHF) and microwave ranges. A new type of active RFID transponder is

© 2011 by Taylor and Francis Group, LLC

linked to the use of RFID-enabled cell phones that can emulate an active tag. In this case, the problem of battery maintenance is solved as the user keeps the cell phone charged as part of its regular maintenance.

A third type of transponder is called *semipassive* or *battery-assisted transponder*. These transponders also include a battery, but contrary to active transponders, the battery is not used to generate the power to transmit the signal to the interrogator. Instead, the battery is used to support secondary functions like the data logging from different types of sensors. These transponders also harvest the energy from the electromagnetic field generated by the interrogator to power its internal circuits other than the sensing and data-logging parts.

1.3 Functional Classification of RFID Transponders

A functional classification of RFID transponders is based on their electronic product code (EPC) class. EPC is the application of one specific type of RFID technology within the consumer packaged goods industry. Within this classification, RFID transponders are divided into different classes and generations:

Generation 1, Class 0: Passive tags with read-only functionality. These are also called *write one, read many* (WORM) transponders. These transponders are programmed at the factory with their unique identification number. The user is not able to change it or include additional information.

Generation 1, Class 0+: These are also WORM transponders. They differ from Generation 1, Class 0 transponders in that it is the user who programs them. After they have been programmed by the user, no further programming or changing of data is allowed.

Generation 1, Class 1: These transponders were similar to Generation 1, Class 0 or 0+ transponders, but could be read by interrogators from other companies. Gen 1, Class 1 transponders have evolved into the different transponders from Generation 2.

Whereas Generation 1 transponders employ proprietary data structures and can be read only by interrogators manufactured by the same vendor, Generation 2 transponders are vendor neutral in their specifications. This means that as they have developed (following agreed upon standards), they can be read with interrogators from multiple vendors. These transponders provide additional features, for example the elimination of duplicate reads within the read range. They are also more reliable than Generation 1 transponders and support faster read rates. The user of Generation 2 transponders

© 2011 by Taylor and Francis Group, LLC

can access specific parts of their memory as these are rewritable transponders.

Generation 2, Class 1: These transponders are WORM transponders. They are also programmed at the factory, but they can be read with equipment from different vendors, support the higher read rates, and have more noise immunity than the Generation 1 transponders.

Generation 2, Class 2: These transponders are rewritable transponders. They can be written several times by the user using equipment different from the vendor's equipment.

Generation 2, Class 3: These transponders are the semipassive or battery-assisted transponders.

Generation 2, Class 4: This refers to active transponders.

Generation 2, Class 5: These transponders are essentially interrogators. Transponders that follow Generation 2, Class 5 must be able to power other transponders.

1.4 Applications and Frequency Selection

RFID systems come in different flavors from the point of view of the frequency they use. Selecting the most adequate frequency is a function of two variables: the technological developments of systems at the different operating frequencies—directly related to the cost of systems—as well as the properties of electromagnetic waves at those different frequencies.

The initial users of RFID systems found systems operating in the low-frequency (LF) range followed a few years later by systems operating in the high-frequency (HF) range. Therefore, their selection of systems was limited to these two broad frequency ranges even if those were not the most adequate for their specific applications. Recent years have seen an increase in the number of available RFID systems operating in the UHF range as well as a reduction of their economic cost. It is possible to expect a similar trend in upcoming years for RFID systems operating in the microwave range to become more available and more affordable. This will help the final users to make decisions on what type of RFID systems to deploy based on technical reasons rather than based on the availability and cost of these systems.

Table 1.1 presents a broad view of the frequency spectrum used by RFID systems as well as the most common frequencies for each band. It is necessary to note that Table 1.1 is not exhaustive because it is possible to find systems, within each frequency band, operating at other frequencies.

The LF band has been used by RFID systems for several years. This fact, together with the less stringent requirements of the electronics operating in this frequency range, has made these systems very cost-effective. Their

© 2011 by Taylor and Francis Group, LLC

TABLE 1.1

Commonly Used Frequency Band for RFID Systems

Frequency Band	Frequency Range	Typical Frequencies Used in RFID Systems
Low Frequency (LF)	100 kHz – 500 kHz	125 kHz 134.2 kHz
High Frequency (HF)	10 MHz – 15 MHz	13.56 MHz
Ultra High Frequency (UHF)	400 MHz – 950 MHz	866 MHz Europe 915 MHz United States
Microwaves (μW)	2.4 GHz – 6.8 GHz	2.45 GHz 3.0 GHz

major drawback is that due to the low frequency of the carrier, these systems can only communicate at low frequency rates. The main advantage of LF RFID systems is that electromagnetic waves operating in the LF range are the least affected by the presence of metals due to the penetration depth of these frequencies. This makes them ideal to be used in applications in which the transponders are surrounded by large metallic masses. Furthermore, LF waves can travel through water with minimal attenuation compared to waves of higher frequencies. This property makes them the frequency range of choice for the identification of animals due to their high water content. Most of these applications are based on implanting these transponders in livestock and pets. Figure 1.5 shows a picture of two different transponders that operate in the LF range in which it is possible to see the multiple loops used by these transponders.

RFID systems operating in the HF range can support higher read rates than LF systems. Although signals in the HF range are more affected by metal and have more attenuation when traveling through water, the transponders used for HF systems can be manufactured at a lower cost than the transponders operating in the LF range. This is due to the fact that the antennas for HF transponders can be made smaller, as shown in the two transponders seen in Figure 1.6. The need for less conductive material to construct the antennas results in a lower cost. Finally, the global uniformity across the world in the frequency of 13.56 MHz makes HF systems able to operate in any country.

While HF RFID systems operate at the single frequency of 13.56 MHz, UHF RFID systems can operate at different frequencies in the UHF band. Moreover, different countries have established different frequency ranges for their UHF RFID operation, making the compatibility between systems more difficult. For example, in North America the assigned band is from 902 MHz to 928 MHz, while in Europe it is from 860 MHz to 868 MHz and in Japan is from 950 MHz to 956 MHz. Other countries and regions have other different allocated frequency ranges. Despite this challenge and due to the developments in microelectronics that allowed the decrease in cost for these

© 2011 by Taylor and Francis Group, LLC

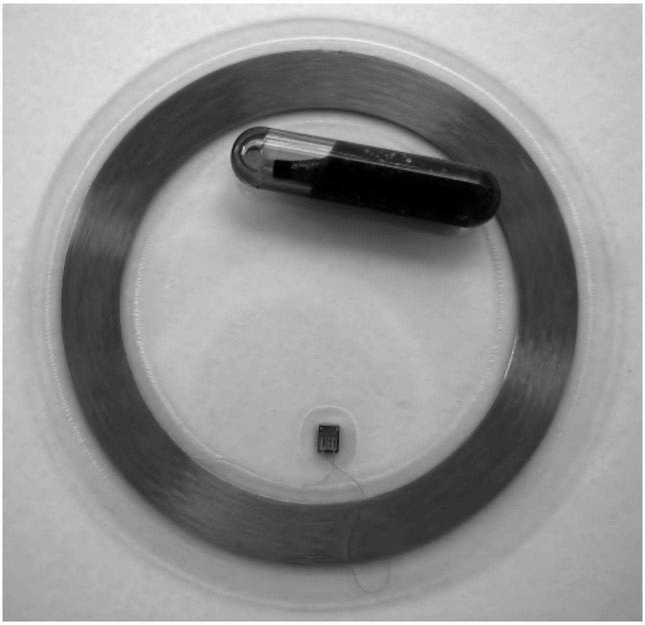

FIGURE 1.5
Low-frequency (LF) RFID transponders require antennas of several hundred feet of coiled wire.

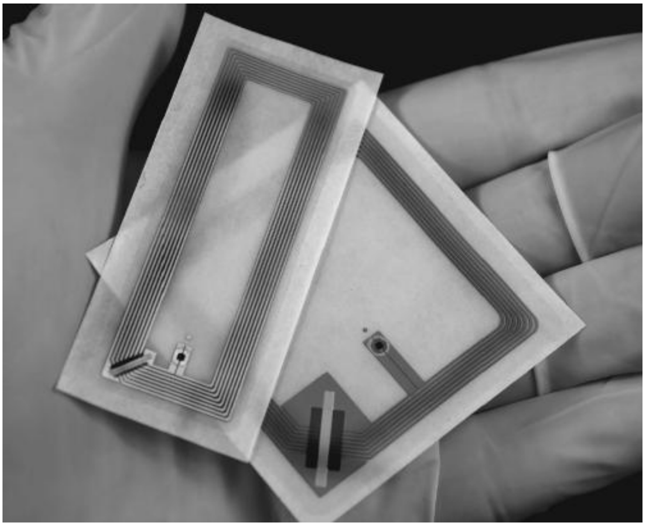

FIGURE 1.6
High-frequency (HF) RFID transponders can use shorter, smaller antennas.

© 2011 by Taylor and Francis Group, LLC

systems, UHF systems have established themselves as a viable alternative to the existing LF and HF systems. One of the main advantages of UHF RFID systems is that their higher carrier frequency allows for a much faster read rate of the transponder information, thus allowing the transmission of higher amounts of data. The communication between transponder and interrogator used in UHF systems is by electrical field instead of the magnetic coupling used by LF and HF systems. The antennas required by UHF systems can also be smaller than the antennas used for LF and HF systems, resulting in higher efficiency. These last two factors contribute to extended read ranges for UHF RFID systems. However, UHF systems also present several drawbacks compared to HF and LF systems. The antennas for transponders operating in the UHF range have very different physical dimensions depending on the parameters that the manufacturer wants to optimize in a given transponder. Figures 1.7 and 1.8 show examples of several UHF transponders in which it is possible to notice these differences.

UHF systems do not work well around metals as these reflect their electromagnetic waves. In addition, water absorbs UHF waves, making them unusable for animal implantation. Furthermore, the extended read range of UHF transponders may become a drawback for those applications that want to keep the read of transponders confined to a certain area.

Although 2.45 GHz is within the UHF range (UHF extends from 300 MHz to 3 GHz), the RFID systems operating in the frequency range extending from 2.45 GHz to 5.8 GHz are commonly called *microwave RFID systems*. In this frequency range, the propagation is limited to line of sight. Microwave signals are also attenuated by water and reflected by metallic objects. The cost of interrogators and transponders used in the microwave range is higher

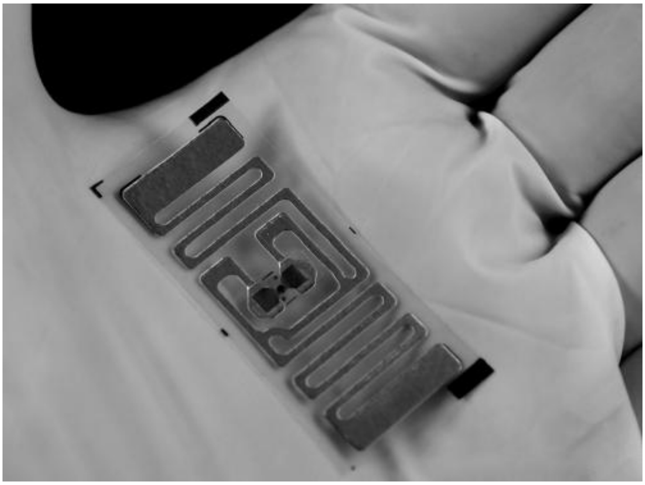

FIGURE 1.7
Ultra-high-frequency (UHF) RFID transponder.

© 2011 by Taylor and Francis Group, LLC

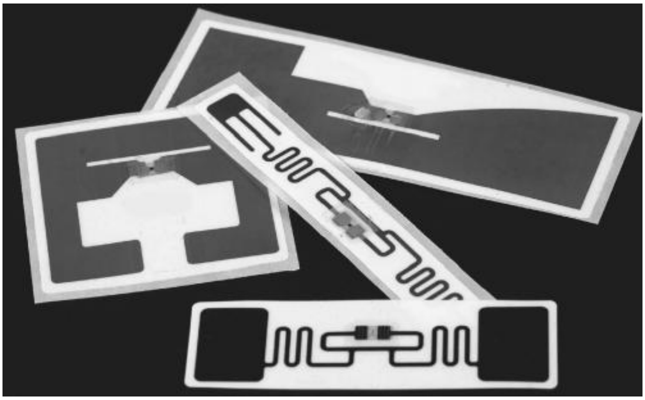

FIGURE 1.8
Several types of antennas used by ultra-high-frequency (UHF) RFID transponders.

than for any other type of system. On the positive side, it is possible to achieve very high read rates, and with the help from spread spectrum techniques, it is possible to achieve very high reliability rates and noise immunity. The higher end of the range, from 5.5 GHz to 6.8 GHz, is still currently under research and development.

© 2011 by Taylor and Francis Group, LLC

2

Antennas for RFID Transponders

CONTENTS

2.1 Review of Basic Antenna Theory for RFID Transponders 14
  2.1.1 Antennas for RFID Transponders Operating in the LH and HF Regions 14
  2.1.2 Inductance of Coil Antennas Operating in the LF and HF Regions 17
  2.1.3 Antennas for RFID Transponders Operating in the UHF Region 24
2.2 Antenna Tuning for RFID Transponders 29
2.3 Antenna Matching for RFID Transponders 33
2.4 Antennas in Commercial RFID Transponders 37
2.5 The Connection between the Chip and the Antenna in RFID Transponders 43
2.6 Additional Factors That Affect the Performance of Antennas in RFID Transponders 47

The purpose of the antenna in the transponder of an RFID system is multiple: first, it has to collect power from the electromagnetic field generated by the interrogator. In addition, the antenna must transfer the collected power to its load that is the chip in the transponder in order to turn it on. Finally, the antenna must radiate the data signals generated by the chip back to the interrogator. The antenna should be optimized to minimize energy losses during this process by choosing the type and dimensions of the antenna suited for a specific application and matching its impedance to the impedance of the load. Because these requirements may lead to contradictory solutions, the designer of the system must evaluate them carefully when considering the possible solutions.

This chapter explores the design and performance considerations for antennas in transponders. It starts by presenting a basic review of antenna theory with special emphasis on the differences used for systems in the low-frequency (LF) and high-frequency (HF) ranges compared to the antennas used for ultra-high-frequency (UHF) systems. The chapter continues by describing the types of antennas found in commercial transponders. This is followed by considering the requirements for antenna tuning and

© 2011 by Taylor and Francis Group, LLC

antenna matching. The following section describes the different methods for attaching the antenna to the chip on the transponder. The chapter finishes with a review of other antenna factors that affect their performance.

2.1 Review of Basic Antenna Theory for RFID Transponders

2.1.1 Antennas for RFID Transponders Operating in the LH and HF Regions

Electrical current flowing through a conductor generates electromagnetic fields. From the point of view of the concepts described in this chapter, it is possible to distinguish two field regions. The first region is called the *far-field region*. In this region, the generated fields are radiated fields, meaning that the energy propagates through the space with an energy density proportional to the inverse of the distance. The second area is called the *near-field region*. In this area, radiated fields are not prevalent. Instead, attenuating fields, in which the strength of the field decreases with $1/r^3$, are dominant. Furthermore, the power in this region is reactive.

Given these two regions, it seems obvious that antennas for RFID transponders should be designed to operate in the far-field region. However, the boundaries between far-field and near-field regions depend on the relationship between the physical dimensions of the antenna and the wavelength of the propagating signal. In particular, the dimensions of the antenna should be comparable to the wavelength of the signal in order to achieve optimal performance. The wavelength of signals operating in the LF region is around 2.4 km (1.5 miles), while the wavelength for signals operating in the HF region is 24 m (78 feet). Therefore, at these frequencies it would not be practical to build antennas with dimensions similar to the wavelength of their signals. Any antennas that can be built in a practical manner for RFID transponders operating in the LF or HF ranges will be electrically small and therefore highly inefficient. In contrast, RFID signals operating in the UHF region have a wavelength of approximately 30 cm. Although this dimension is still very large for building practical antennas, the following sections will explore how they can be modified to obtain antennas for UHF transponders that will operate in the far-field region.

RFID transponders operating in the LF or HF frequencies cannot use dipole antennas because of the mismatch in dimensions. An alternative solution to this problem is to use a small-loop antenna instead of a dipole. While the near fields radiated by a dipole are mainly electrical fields, the near field from a loop antenna is a magnetic field. A small-loop antenna is a closed loop with a maximum dimension that is less than about a tenth of the wavelength of the signal. It can be shown that the small-loop antenna is the dual equivalent of an ideal dipole, thus making it suitable for antennas incorporated in transponders operating at LF or HF frequencies.

© 2011 by Taylor and Francis Group, LLC

Let's consider a conductor of infinite length carrying a magnitude of current of I amps. The magnetic field (B_Φ) measured at a distance of r meters from the conductor can be found using Ampere's law as

$$B_\phi = \frac{\mu_0 I}{2\,\pi\, r} \;(Weber\,/\,m^2) \tag{2.1}$$

where μ_0 is the permeability of free space ($4\,\pi\,10^{-7}$ H/m).

However, a conductor of infinite length is not realistic. Its practical implementation is based on building a loop antenna by bending the original, finite wire, which carries a current of I amps in a circle with a radius of a meters. In practice, the wire is bent in such a way that produces a total of N turns as this allows using a longer wire with a relatively small diameter. In this situation, the value of the magnetic field in the z-coordinate direction (B_z) for a point located at a distance of r meters from the plane of the coil and located along the axis of the coil, as shown in Figure 2.1, can be found as

$$B_z = \frac{\mu_0 I N a^2}{2\,(a^2 + r^2)^{3/2}} \;(Weber\,/\,m^2) \tag{2.2}$$

where a is the radius of the loop in meters.

Assuming that the distance at which the field is measured, r, is much larger than the radius of the coil ($r^2 >> a^2$), equation (2.2) can be simplified as follows:

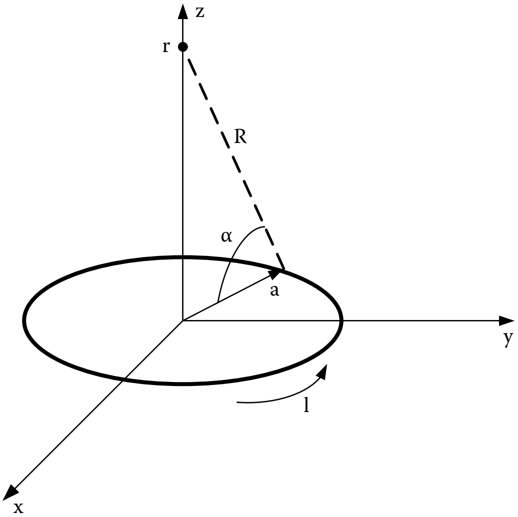

FIGURE 2.1
Magnetic field produced by loop antenna.

© 2011 by Taylor and Francis Group, LLC

$$B_z = \frac{\mu_0 I N a^2}{2 r^3} \quad (Weber / m^2) \tag{2.3}$$

As shown by equation (2.3), the magnetic field created by the small-loop antenna is dependent on $1/r^3$. This causes a strong decrease of field intensity as one moves away from the location of the coil antenna, and will seriously limit the read range of transponders using this type of antenna. However, even with this limitation, the small-coil antenna is more desirable than an electrically small dipole.

The antenna in the RFID transponder has to be both a receiving antenna and a transmitting antenna. It has to collect enough energy from the electromagnetic field generated by the interrogator to be able to power its integrated circuit, and it also has to transmit the data signal generated by the chip back to the interrogator. The *reciprocity theorem* indicates that the receiving and transmitting patterns for a given antenna are identical, and therefore the same antenna can be used for either task.

The voltage induced in a coil consisting of N loops immersed in a time-variant magnetic field can be calculated using Faraday's law as

$$V = -N \frac{d\Psi}{dt} \tag{2.4}$$

with Ψ being the magnetic flux.

The magnetic flux, in turn, can be calculated as

$$\Psi = \int B \bullet dS \tag{2.5}$$

where

B = the magnitude of the time-variant magnetic field

S = the surface area of the coil

$\cdot$ = the inner product operator

Equation (2.5) indicates that the magnetic flux, and consequently the voltage induced in the coil, is dependent on the relative orientation between the coil and the magnetic field. This has some important implications at the time of using interrogators and transponders. Assuming that the magnetic field B was generated by the interrogator using another loop antenna, the maximum received voltage will occur when both coils are placed parallel to each other. Because the voltage generated in the coil that will be used to power the chip in the transponder is proportional to the intensity of the magnetic field (B), this voltage will also be dependent on ($1/r^3$). This is important because

© 2011 by Taylor and Francis Group, LLC

it limits the maximum distance between the interrogator and transponder coils that will produce the threshold voltage necessary to power the chip on the transponder.

Although the previous equations describe the full underlying relationships between the variables involved in this process, it is more desirable from a practical point of view to study how these equations evolve in specific cases. Assuming that the coil is specifically tuned to the frequency of the time-variant magnetic field, the induced voltage in the coil is as follows:

$$V_0 = 2\pi f N S Q B_0 \cos\beta \tag{2.6}$$

where

V_0 = the induced voltage in the coil

f = the frequency of the time-variant magnetic field

B_o = the intensity of the field

N = the number of loops in the coil

Q = the quality factor of the coil antenna

S = the area of the coil

β = the angle between the plane of the coil and the direction of the field

Keeping in mind that the magnetic field generated by the coil in the interrogator is perpendicular to its plane, and that equation (2.6) will be maximized when $\beta = 0$, the induced voltage will be maximized when the interrogator and transponder coils are placed parallel to each other. Once again, the limiting factor in equation (2.6) comes from the intensity of the magnetic field (B_0) decreasing with the cube of the distance between coils ($1/r^3$).

2.1.2 Inductance of Coil Antennas Operating in the LF and HF Regions

Because, as shown in equation (2.6), the induced voltage is proportional to the quality factor (Q factor) of the coil antenna, it is desirable to increase the Q factor as much as possible. This can be achieved by carefully choosing the value of the inductance of the antenna. This section discusses the values of inductance found in typical antennas operating in the LF and HF regions.

Typical RFID systems use inductance values for antennas in transponders of few mH, while antennas for the interrogators have typical values between 10 and 100 times lower. It is necessary to keep in mind that the calculation of inductance values presented here can only be considered as approximated values. These calculations do not take into consideration the finite value of the conductivity of the wire that results in Ohmic losses, and, more importantly, they do not take into account the distributed stray capacitances that appear in the coil. A more detailed and comprehensive analysis and calculations of inductance values in antennas are beyond the scope of this book, but

© 2011 by Taylor and Francis Group, LLC

they can be found in any book on electromagnetism. If the designer requires an accurate value of inductance, the best option is to measure it with a complex impedance meter. This instrument is also able to measure the Q factor of the antenna, providing good characterization of the antenna.

The inductance of the coil shown in Figure 2.2, made with a single layer of wire and with n turns and a core of air, is approximately:

$$L = \frac{d^2 n^2}{18d + 40l} (\mu H) \tag{2.7}$$

where

d = the diameter of the coil in inches

n = the number of turns

l = the length of the coil in inches

Equation (2.7) shows the square relationship between inductance and number of turns; if the number of turns is doubled, the resulting inductance is quadrupled.

Example 2.1:

Calculate the number of turns required to achieve an inductance of 10 μH by winding a piece of wire over an air core with a diameter of 1 inch. The length of the resulting coil must be equal to 1.25 inches.

Solution:

In this situation, l = 1.25 inches, and d = 1.0 inches. Therefore, we can write

$$10 = \frac{(1.0)^2 n^2}{18 \cdot 1.0 + 40 \cdot 1.25}$$

Solving for n gives $n = 26.07$ turns. Given the limitations of the approximation for equation (2.7) plus the stray capacitance issues mentioned earlier, a good starting point would be to choose 26 turns. In this case, the value of inductance that we would measure in practice is likely to differ from 10 μH. The designer would have to adjust the number of turns after measuring the actual value of inductance. A recommended procedure is to start by choosing a larger number of turns and therefore a larger number of inductance, and proceed by removing turns until the coil achieved the required inductance.

Although this type of structure to create the antenna for an RFID system is not the most commonly used, it is nevertheless employed in some applications, especially in those that required a hermetically sealed transponder. Figure 2.3 shows an RFID transponder operating in the LF frequency in which it is possible to see the coil antenna and its windings.

© 2011 by Taylor and Francis Group, LLC

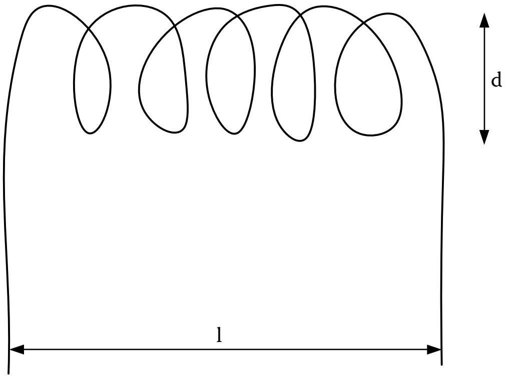

FIGURE 2.2
Single-layer, air-core coil.

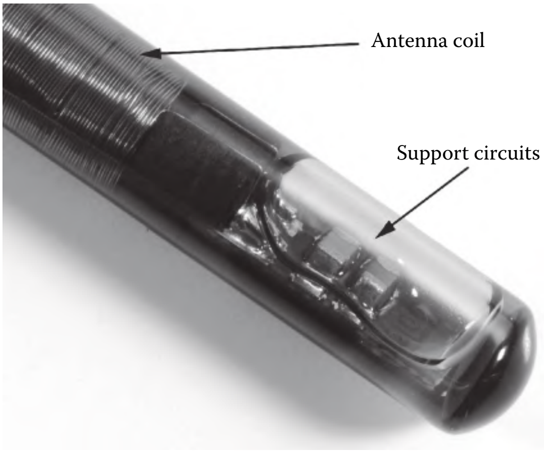

FIGURE 2.3
Detail of coil antenna in a commercial radiofrequency identification (RFID) transponder for the low-frequency (LF) range.

The practical problem associated with the use of this type of antenna is the three-dimensional space that it requires. Such antennas are not practical for applications that require a flat transponder, such as a credit card, and they are limited to applications that can support the tubular shape, for example in animal implantation. For this reason, the majority of transponders in the LF and HF frequency regions use antennas formed on a single plane, using several layers of conductor to achieve the desired inductance. Figure 2.4 shows this configuration for a multilayer circular coil with *n* turns.

© 2011 by Taylor and Francis Group, LLC

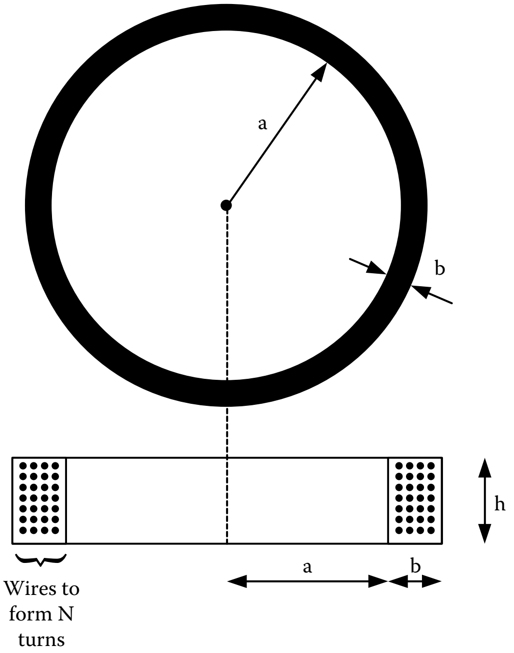

FIGURE 2.4
Multilayer, circular-coil, *N*-turn antenna.

The inductance of this coil is

$$L = \frac{0.31a^2 n^2}{6a + 9h + 10b} \quad (\mu H) \tag{2.8}$$

where

a = the radius of the coil in cm
n = the number of turns
b = the winding thickness of the coil in cm
h = the winding height in cm

This configuration allows building coils with the required amount of inductance with minimal thickness, therefore minimizing the transponder's footprint. A variation of this antenna that results in an even thinner antenna is the spiral wound coil with a single layer shown in Figure 2.5.

The inductance of a single-layer spiral antenna is

$$L = \frac{0.3937\, a^2 n^2}{8a + 11b} \quad (\mu H) \tag{2.9}$$

© 2011 by Taylor and Francis Group, LLC

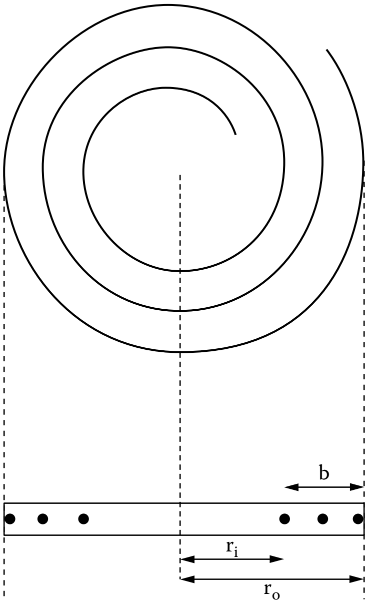

FIGURE 2.5
Single-layer, spiral-coil antenna.

where

a = the average radius of the coil in cm $(a = \frac{r_i + r_o}{2})$

b = ro – ri in cm

n = the number of turns in the spiral

This structure results in a thinner antenna, although the number of turns in the antenna is limited to the number of turns that is possible to pack in a single layer. This limits the inductance of the antenna as this value depends on the square of the number of turns. Other forms of planar coil antennas are the multilayer square-loop coil antenna with n turns and the multilayer rectangular-loop coil antenna, also with n turns, shown in Figure 2.6.

The inductance for the planar multilayer square loop coil antenna with n turns, shown in Figure 2.6(a), is

$$L = 0.008\, a\, n^2 \left[2.3 \log\left(\frac{a}{h+b} \right) + 0.2235\, \frac{b+h}{a} + 0.726 \right] (\mu H) \tag{2.10}$$

where

n = the number of turns

a = the distance from the center of the coil to the side in cm

h = the thickness of the winding in cm

b = the width of the winding in cm

© 2011 by Taylor and Francis Group, LLC

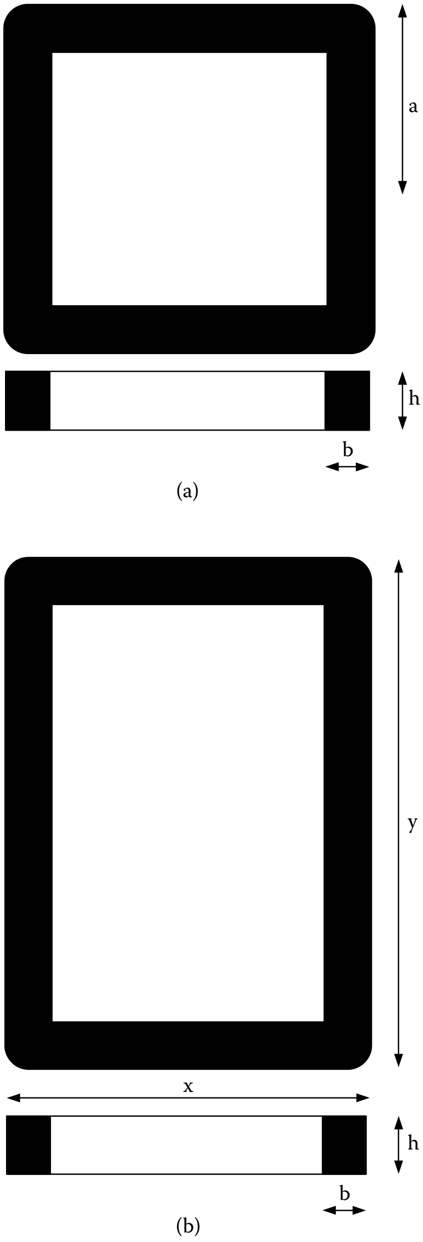

FIGURE 2.6
(a) Multilayer, square-coil antenna. (b) Multilayer, rectangular-coil antenna.

© 2011 by Taylor and Francis Group, LLC

The inductance for the planar multilayer rectangular-loop antenna with n turns, shown in Figure 2.6(b), is

$$L = \frac{0.0276\, n^2 (x + y + 2h)^2}{1.908\ (x + y + 2h) + 9b + 10h} \quad (\mu H) \tag{2.11}$$

where

x and y = the dimensions of the coil in cm
h = the thickness of the coil in cm
b = the width of the winding in cm
n = the number of turns

It is necessary to state once again that these equations, used to calculate the inductance of various coils are only approximations and need to be taken with some care. In particular, equation (2.11) is valid only when the dimensions x and y of the coil are quite different. For example, if the two dimensions of the coil in equation 2.11 (x and y) were identical, thus resulting in a square coil, the inductance values calculated by equations (2.10) and (2.11) will be different. These equations are provided as a starting point for the designer of antennas for RFID transponders.

Some antennas in transponders are configured as planar rectangular spiral structures, as shown in Figure 2.7. Figure 2.8 shows an example of a commercial transponder using this type of antenna.

There is not a full and closed equation to calculate the inductance of this type of antenna. However, its inductance can be calculated as the sum of the self-inductances for all the straight segments minus the sum of the mutual inductances existing in the arrangement. The mutual inductance is

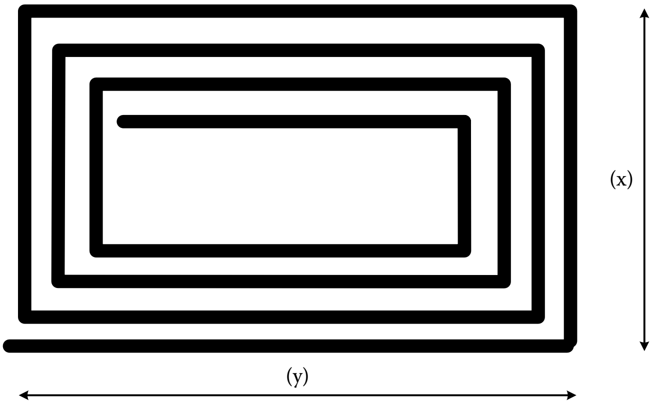

FIGURE 2.7
Single-layer, rectangular planar spiral antenna.

© 2011 by Taylor and Francis Group, LLC

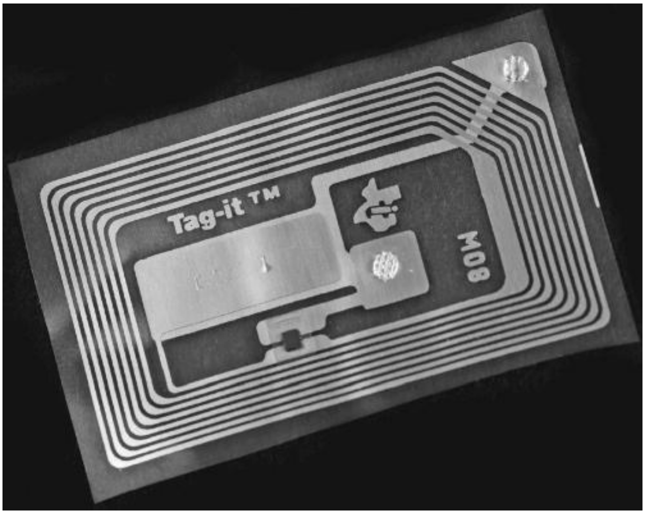

FIGURE 2.8
Commercial transponder using a planar, rectangular-spiral antenna.

the inductance that results from the magnetic fields produced by adjacent conductors. This mutual inductance will be positive when the direction of the current in the conductors is the same, and it will be negative when it is opposite.

2.1.3 Antennas for RFID Transponders Operating in the UHF Region

When the transponder is operating in the UHF region, the wavelength of the electromagnetic radiation is now comparable to the dimensions of the antenna. Because the antenna is no longer electrically small, a dipole can now be used. A *short dipole* is defined as a dipole whose length does not exceed one-fiftieth of the wavelength of the current signal. A short dipole, as depicted in Figure 2.9, produces an electric field that in near-field conditions has both radial and zenithal components, as given by equation (2.12):

$$E_r = \frac{2I_0 l}{4\pi} k^2 \eta_0 \, e^{-jkr} \left(\frac{1}{(kr)^2} - \frac{j}{(kr)^3} \right) \cos\theta$$
$$E_\theta = \frac{I_0 l}{4\pi} k^2 \eta_0 \, e^{-jkr} \left(\frac{j}{kr} - \frac{1}{(kr)^2} - \frac{j}{(kr)^3} \right) \sin\theta \tag{2.12}$$

© 2011 by Taylor and Francis Group, LLC

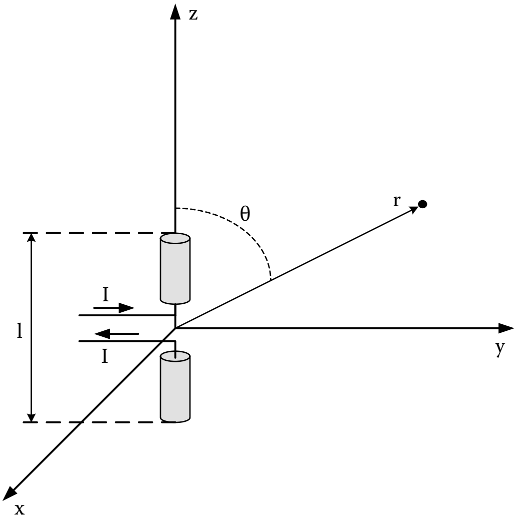

FIGURE 2.9
Electric field produced by a short dipole.

where

k = the wavenumber $\left(k = \frac{2\pi}{\lambda}\right)$

η_0 = the intrinsic impedance of free space $\left(\eta_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} \approx 120\pi\right)$

r = the distance for the center of the dipole at which the field is measured

θ = the zenith angle at which the field is measured

Assuming that the far-field approximation is valid ($r \gg \lambda$), the radial component of the electrical field can be neglected, and therefore equation (2.12) becomes

$$E_\theta = \frac{jI_0 l}{4\pi} k \eta_0 \frac{e^{-jkr}}{r} \sin\theta \tag{2.13}$$

The short dipole, however, does not radiate effectively because it has a very small radiation resistance, normally below 1 Ω, making it unsuitable for a good impedance match. Therefore, other dipole alternatives are preferred when possible, such as the half-wave dipole. The *half-wave dipole,* depicted in Figure 2.10, receives its name because its length equals half of the wavelength of the current signal traveling through the antenna.

A half-wave dipole also produces an electric field only in the zenith component that can be described by equation (2.14):

© 2011 by Taylor and Francis Group, LLC

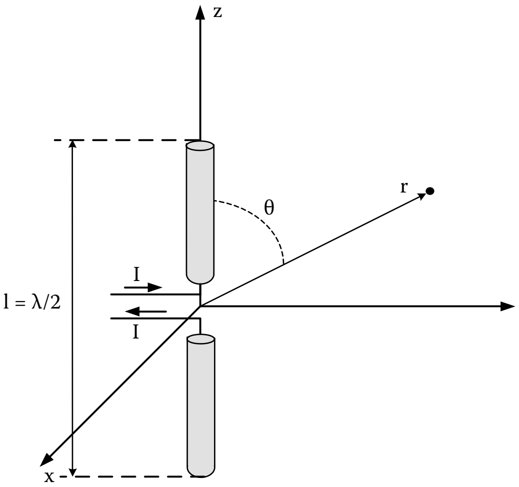

FIGURE 2.10
Electric field produced by a half-wave dipole.

$$E_\theta = \frac{j\,60 I_0}{r}\left(\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}\right)e^{-jkr} \tag{2.14}$$

One of the main advantages of the half-wave dipole is that its radiation resistance, around 73 Ω, is much larger than the radiation resistance of the short dipole. This makes the half-wave dipole better suited to match the impedance presented by its load, the integrated circuit in the transponder.

The wavelength for the typical UHF RFID frequency of 915 MHz is approximately 33 cm. Therefore, the length of a half-wave dipole is 16.5 cm. This value is valid only for free space; for an antenna mounted on an inlay, the actual value of its half wavelength decreases as the dielectric constant of the substrate increases. Although this distance is achievable, it presents some drawbacks. First of all, 16.5 cm may still be too large for some applications that require or would benefit from smaller transponders. More importantly, the power transfer characteristics for this antenna are very small and therefore require substantial matching approaches. Antenna matching is described in further detail in Section 2.2.

The space required by the antenna can be shortened while keeping the length of the conductors to the required 16.5 cm, for example by bending or even folding the wires that make up the dipole, as shown in Figure 2.11. This process can be repeated several times, bending the wires of the antenna in different directions as shown in Figure 2.12, which results in the structure known as the *meandered dipole*.

© 2011 by Taylor and Francis Group, LLC

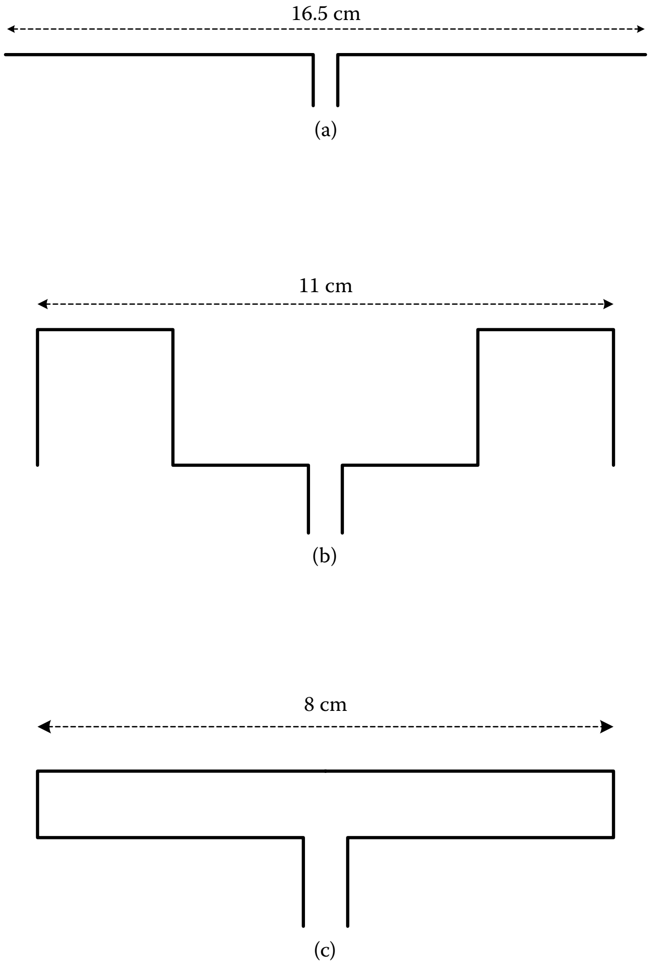

FIGURE 2.11
Strategies for reducing the size of ultra-high-frequency (UHF) antennas. (a) Original half-wave dipole. (b) Bent dipole. (c) Folded dipole.

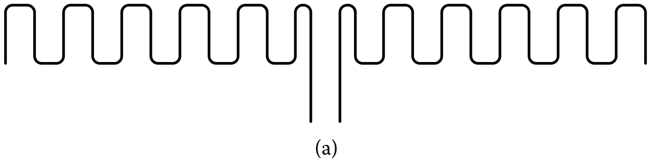

FIGURE 2.12
Examples of meandered dipole antennas. (a) Meandered dipole in a single direction. (b) Meandered dipole in two directions.

© 2011 by Taylor and Francis Group, LLC

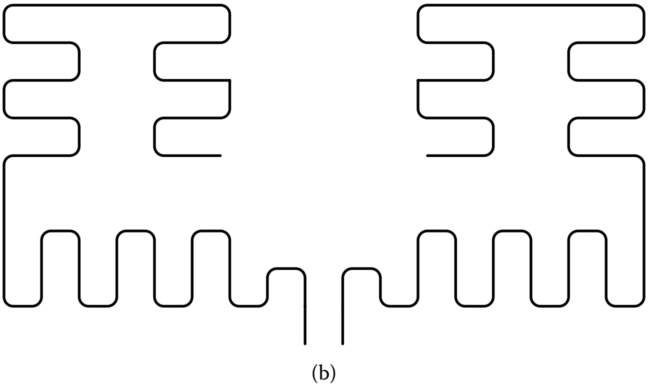

FIGURE 2.12 (Continued)
Examples of meandered dipole antennas. (a) Meandered dipole in a single direction. (b) Meandered dipole in two directions.

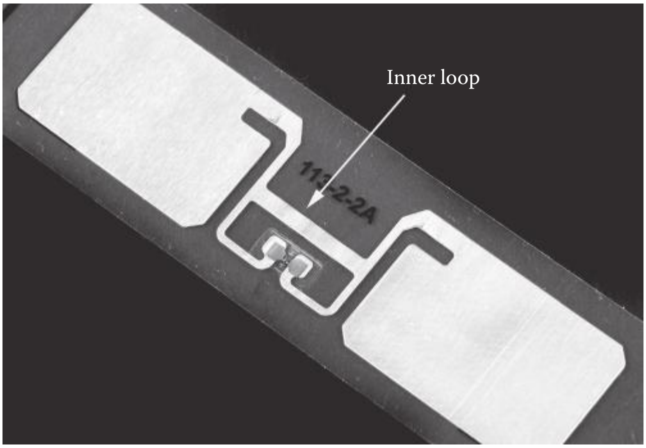

FIGURE 2.13
Detail of the inner loop used in some UHF antennas.

However, bending the dipole degrades the electrical characteristics of the antenna: the current lines flowing in opposite directions through adjacent paths have a canceling effect that in turn increases the radiation resistance of the antenna. Because this effect depends on the distance between the paths, the canceling effect becomes stronger when the antenna has multiple bends closely packed. In practice, the total length of a bent or meandered dipole is longer than the straight half-wave wavelength, although the total size of the antenna will be smaller.

Some antennas operating at the UHF range also incorporate a low DC resistance path closer to the center of the antenna known as the *inner loop*. This inner loop, shown in Figure 2.13, presents a series of advantages that makes its use common in this type of antenna.

© 2011 by Taylor and Francis Group, LLC

First, the inner loop presents a low DC resistance that helps reduce the potential of damage to the integrated circuit due to high-voltage electrostatic discharges that may occur in the antenna or its vicinity. Second, the inductance of the inner loop can be used in conjunction with the input capacitance of the integrated circuit to form a resonance circuit that will couple with the magnetic near field, producing enough voltage to power the integrated circuit. Near-field communications for UHF RFID systems, although not very common yet, are growing rapidly. Third, the inner loop can be used like a transformer to help match the impedance of the integrated circuit to the antenna. Finally, the impedance of the antenna can be lowered or increased by changing the width of the trace in the inner loop.

2.2 Antenna Tuning for RFID Transponders

Antenna tuning for RFID transponders operating at LF or HF frequencies attempts to make the antenna resonant at the frequency of transmission, thus maximizing the current or the voltage in the system. Because the antenna has a marked inductive component, resonance is achieved by placing a capacitor of the adequate value in series or in parallel with the antenna.

When the capacitor is placed in series with the antenna, the resulting series resonant circuit has an impedance minimum at the resonance frequency and therefore maximizes the current being delivered to its load. This type of configuration is typical for interrogators and will be studied in Chapter 4. When the capacitor is placed in parallel with the antenna, the result is a parallel circuit that exhibits an impedance maximum at the resonant frequency and therefore maximizes the voltage across its terminals. This is the typical configuration for transponders. The tuning capacitor in RFID transponders can be created over the same substrate as shown in Figure 2.14.

The resonant circuit can therefore be modeled as a parallel circuit, similar to the one shown in Figure 2.15, in which L represents the inductance of the antenna, C represents the capacitance of the capacitor that must be added, R represents the resistance of the load that is the integrated circuit that must be powered, and the value r represents the resistance of the coil. Because normally $r << R$, the value r will be ignored in the calculations needed to find the value of required capacitance. However, the resistance r of the inductor contributes to degrading the overall performance of the antenna. These effects will be studied in the last section of this chapter.

The complex impedance presented by the circuit shown in Figure 2.15, neglecting the Ohmic losses r, is

© 2011 by Taylor and Francis Group, LLC

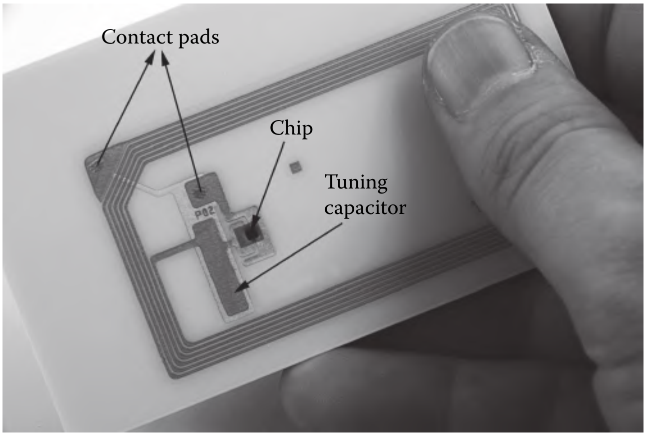

FIGURE 2.14
Tuning capacitor in the antenna of a high-frequency (HF) transponder.

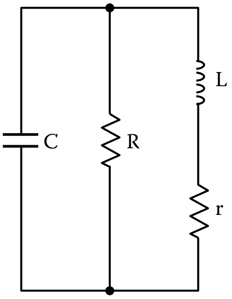

FIGURE 2.15
Parallel resonant model used to analyze antenna resonance. L: antenna inductance. C: external tuning capacitor. R: load (chip to energize). r: Ohmic losses in the antenna.

$$Z(j\omega) = \frac{\dfrac{j\omega}{C}}{\left(\dfrac{1}{LC} - \omega^2\right) + j\dfrac{\omega}{RC}} \tag{2.15}$$

where $\omega = 2\pi f$.

The resonant frequency for the parallel circuit is the frequency that creates the maximum for $Z(j\omega)$. This frequency is

$$\omega_o = \sqrt{\frac{1}{LC}} \quad or$$

$$f_o = \frac{1}{2\pi\sqrt{LC}} \tag{2.16}$$

© 2011 by Taylor and Francis Group, LLC

From equation (2.15), it can be seen that at the resonant frequency, the impedance of the parallel circuit is real and equal to the value of the resistance of the load. This is the maximum value for the impedance.

Example 2.2:

Consider the value of inductance for the antenna in Example 2.1 (10 μH). Calculate the value of parallel capacitance required to make the antenna resonant at the typical RFID frequencies of (a) 125 kHz for LF systems and (b) 13.56 MHz for HF systems.

Solution:

Solving for C in equation (2.16) yields

$$C = \frac{1}{L(2\pi f)^2}$$

Therefore, the values of capacitance required for resonance are as follows:
$C_{125\ \text{kHz}} = 162$ nF
and
$C_{13.56\ \text{MHz}} = 13.7$ pF

The same calculations for resonant frequency can be applied to the devices that communicate data by tuning and detuning the antenna. This can be accomplished by using, for example, two capacitors or two inductors in the parallel RCL circuit. In order to detune the antenna, the controlling device can, for example, shorten one of the capacitors or inductors, resulting in a new resonant frequency. This, in turn, results in a minimal transfer of energy back to the interrogator.

The devices MCRF355 and MCRF360 from Microchip Technologies operate following this principle. Figure 2.16 shows one of the configurations used. Here, the antenna will be tuned when both inductors are active and will be detuned when the device shortens the inductor L_2.

Using equation (2.13), the resonant frequency when the antenna is tuned is $f_{\text{tuned}} = 1/2\pi\sqrt{L_T\,C}$, with L_T being the total inductance between point A and ground. When the device detunes the antenna by shortening inductor L_2, the new resonant frequency becomes $f_{\text{detuned}} = 1/2\pi\sqrt{L_1\,C}$. Because $L_T > L_1$ ($L_T = L_1 + L_2 + 2k\sqrt{L_1 L_2}$, with k being the coupling coefficient between the coils), $f_{\text{tuned}} < f_{\text{detuned}}$.

The difference between the tuned and detuned frequencies must be chosen to maximize the Modulation Index and the read range. If it is necessary to choose the tuned and detuned frequencies such as $f_{\text{tuned}} < f_{\text{detuned}}$, this can be achieved by using the configuration shown in Figure 2.17.

The operation is similar as shown before. Here, $f_{\text{tuned}} = 1/2\pi\sqrt{L\,C_T}$, with $C_T = C_1C_2/\,C_1 + C_2$. The device detunes the circuit by shortening the capacitor

© 2011 by Taylor and Francis Group, LLC

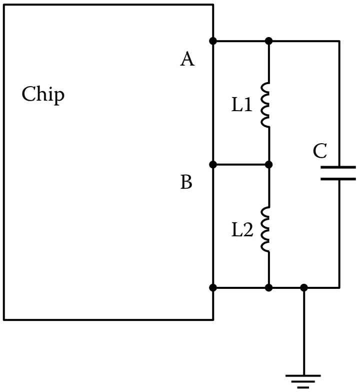

FIGURE 2.16
Tuned and detuned circuit using two inductors and one capacitor. This configuration results in $f_{\text{tuned}} < f_{\text{detuned}}$.

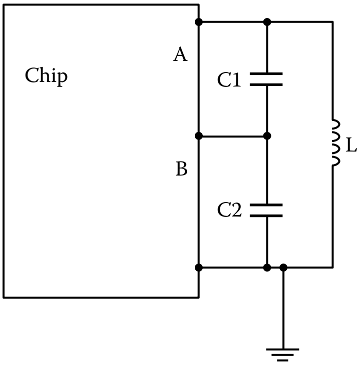

FIGURE 2.17
Tuned and detuned circuit using two capacitors and one inductor. This configuration results in $f_{\text{tuned}} > f_{\text{detuned}}$.

C_2, resulting in the following detuned frequency: $f_{\text{detuned}} = 1/2\pi\sqrt{LC_1}$. Because $C_T < C_1$, this results in $f_{\text{tuned}} < f_{\text{detuned}}$.

Equation (2.15) can also be used to calculate the bandwidth of the system defined as its –3 dB cutoff. This is

$$B = \frac{1}{2\pi RC} \quad (Hz) \tag{2.17}$$

An important parameter in a resonant circuit is its quality factor (Q), which compares the energy stored in the circuit with the energy that it dissipates:

© 2011 by Taylor and Francis Group, LLC

$$Q = \frac{f_o}{B} \tag{2.18}$$

Equation (2.17) can be rewritten taking into account the values of components in the resonant circuit by combining equations (2.16) and (2.17):

$$Q = R\sqrt{\frac{C}{L}} \tag{2.19}$$

Equation (2.6), found earlier, was used to calculate the voltage induced in the coil. Substituting equation (2.19) into equation (2.6) yields

$$V_0 = 2\pi f N S Q B_0 \cos\beta \ = \ 2\pi f N S R\sqrt{\frac{C}{L}}\, B_0 \cos\beta \tag{2.20}$$

This equation indicates that the voltage induced in the antenna transponder is inversely proportional to the square root of its inductance, and proportional to both the number of turns in the coil and its surface.

2.3 Antenna Matching for RFID Transponders

The purpose of matching an antenna to its load is to ensure that the antenna transfers the maximum amount of power to its load. In an RFID transponder, the load of the antenna is the integrated circuit that will be powered by the energy supplied by the antenna. Antenna matching is generally based on altering the complex impedance presented by the antenna at its operating frequency, by modifying its physical dimensions, inserting a reactive component, or a combination of both.

Consider a simple circuit, as shown in Figure 2.18. The power transferred by the voltage source V_s to the load resistance R_L is

$$P_L = \frac{R_L}{(R_L + R_s)^2}\, V_s^2 \tag{2.21}$$

The goal of power matching is to find a relationship between R_s and R_L that will maximize the power transferred to the load. In mathematical terms, this

© 2011 by Taylor and Francis Group, LLC

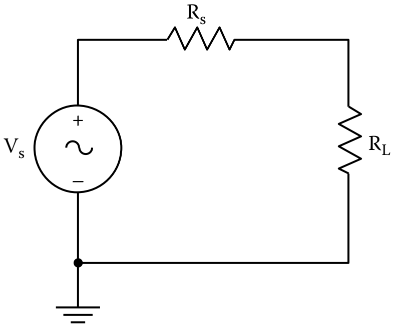

FIGURE 2.18
Simplified model for power transfer calculations.

is equivalent to finding the relationship between R_L and the rest of the circuit elements that will make the first derivative of equation (2.21) zero. That is,

$$\frac{dP_L}{dR_L} = \frac{(R_L + R_s)^2 - 2R_L\ (R_L + R_s)}{(R_L + R_s)^4} V_s^2 = 0 \quad \Rightarrow \quad R_L = R_s \tag{2.22}$$

In a transponder, the voltage source models the voltage induced in the antenna, and the load models the integrated circuit that must be powered. These are not purely resistive elements such as those depicted in Figure 2.18, but they have resistance and reactance. Therefore, a more accurate model for this situation is shown in Figure 2.19.

Equation (2.23) shows the power transferred to the load that is the real part of the impedance of the integrated circuit:

$$P_L = \frac{R_{ic}}{2\left[\left(R_{ant} + R_{ic}\right)^2 + \left(X_{ant} + X_{ic}\right)^2\right]} V_{ant}^2 \tag{2.23}$$

There are now two conditions that must be met in order to transfer the maximum amount of power between the source (voltage generated by the antenna) and the load (integrated circuit to be powered). These are

$$\begin{cases} R_{ic} = R_{ant} \\ X_{ic} = -X_{ant} \end{cases} \tag{2.24}$$

Equation (2.24) can now be understood by saying that the amount of power transferred from the antenna to the integrated circuit will be maximum when their impedances are conjugate. This is an advantageous condition for RFID systems as the impedance of the antenna has an inductive behavior and the

© 2011 by Taylor and Francis Group, LLC

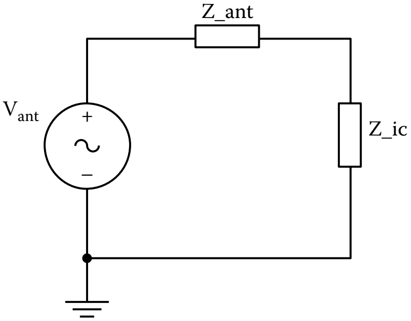

FIGURE 2.19
Model for power transfer calculations in an RFID transponder.

input impedance of the integrated circuit presents a capacitive behavior. The model shown in Figure 2.19 can be further refined as the model shown in Figure 2.20.

It is necessary to point out that the series impedance model used in Figure 2.20 is not the most common way to represent the input impedance of an integrated circuit or any other electronic device. Generally, input impedance is expressed using a parallel model with the input resistance and the input capacitance. However, to preserve the parallelism used in Figures 2.18 and 2.19, it is more convenient to express it using the series model. Given the values of parallel resistance and capacitance, it is possible to find their equivalent series values using equation (2.25):

$$\begin{cases} R_s = \dfrac{R_p}{1+(\omega R_p C_p)^2} \\ C_s = \dfrac{1+(\omega R_p C_p)^2}{\omega^2 R_p^2 C_p} \end{cases} \tag{2.25}$$

Example 2.3:

The specifications for the chip MCRF200 from Microchip used in RFID transponders indicate a parallel capacitance of 2 pF. Assuming a parallel resistance of 3 kΩ, find the equivalent series model at the frequency of 915 MHz.

Solution:

Equation (2.25) gives the following values of the equivalent series model:

$R_s = 2.6\ \Omega$ and $C_s = 1.96$ pF. In general, the values of series and parallel capacitance will be approximately the same. However, the values of the series and parallel resistances are very different for each model.

© 2011 by Taylor and Francis Group, LLC

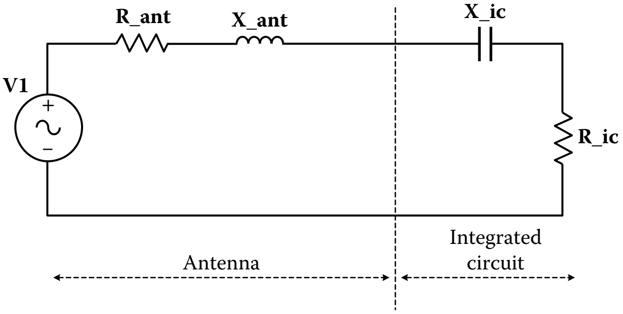

FIGURE 2.20
Series model for the transponder antenna and chip to be energized.

It is also interesting to calculate the gain or attenuation between the voltage delivered to the integrated circuit and the voltage generated by the antenna in the optimal case of maximum power transfer. Using the model shown in Figure 2.16, the voltage across the integrated circuit produced by the antenna can be calculated as follows:

$$V_{ic} \approx j \frac{1}{2\pi\omega \, R_{ic} \, C_{ic}} \, V_{ant} \tag{2.26}$$

Example 2.4:

Using the values given in Example 2.3, calculate the voltage gain or attenuation at the input of the integrated circuit at the frequency of 915 MHz.

Solution:

Using equation (2.26), $V_{\text{ic}} / V_{\text{ant}}$ is equal to 34. This means that the voltage at the input of the integrated circuit in the transponder at the resonance frequency will be 34 times larger than the voltage generated by the antenna. In practice, this number will be lower because equation (2.26) has not considered factors such as the effect of the Q factor and other losses in the system. Typical voltage gains in practice are around 10 to 15. In any case, this represents an important benefit for the designer of the RFID system.

It is necessary to consider that, in practice, there will be some mismatch between the impedance of the integrated circuit and the impedance of the antenna as these will not be perfect conjugates. This mismatch will produce losses in the power being delivered to the integrated circuit, resulting in decreasing the sensitivity and read range for the transponder. This factor becomes more important for UHF antennas designed for global operations that have a frequency bandwidth of 860 MHz to 960 MHz rather than those

© 2011 by Taylor and Francis Group, LLC

designed for operation at a single frequency. In this case, the usual design approach is to place the matched frequency at the geometrical mean of the two frequency ends, resulting in 908 MHz. This will result, however, in mismatch losses at frequencies close to 860 MHz and frequencies close to 960 MHz.

2.4 Antennas in Commercial RFID Transponders

Coil antennas are used in the different type of encapsulated glass transponders, similar to Figure 2.21, operating in the LF region manufactured by Texas Instruments. Glass transponders are hermetically sealed, making them waterproof and therefore suitable to be immersed in water and other fluids.

These transponders are marketed toward access control, vehicle identification, container tracking, asset management, animal identification, and waste management applications, as described by the literature provided by their manufacturer. Texas Instruments has different versions of these transponders depending on the size of the memory in the chip and its ability to be rewritten by the user. The transponders are also available in different sizes, in particular lengths of 12 mm, 23 mm, and 32 mm. Figure 2.22 shows the different sizes of these glass transponders that operate at the central frequency of 134.2 kHz, corresponding to the LF range. Other operational characteristics of these glass transponders are listed in Table 2.1.

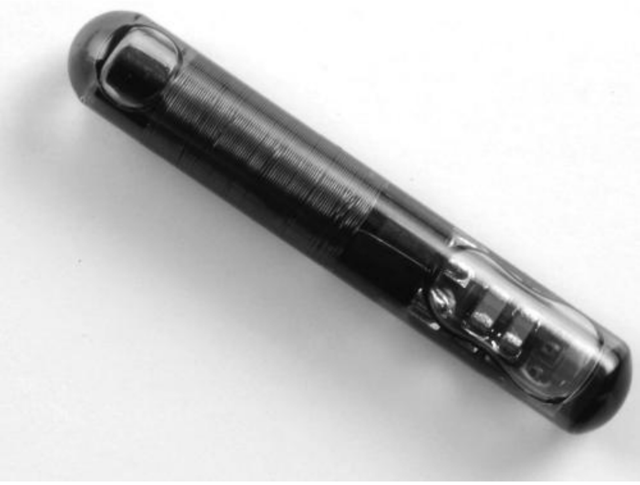

FIGURE 2.21
Glass transponder used for RFID applications in the LF range.

© 2011 by Taylor and Francis Group, LLC

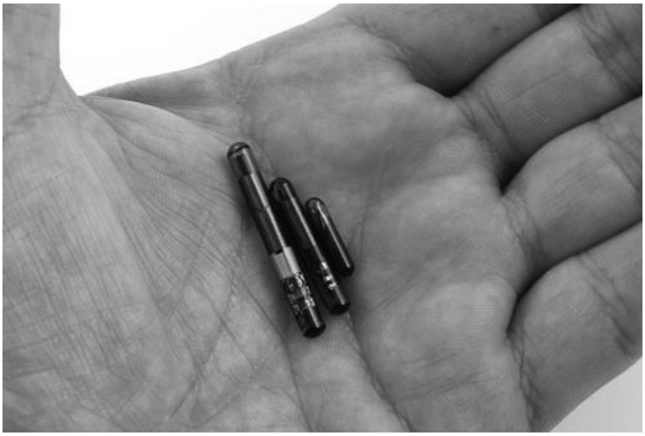

FIGURE 2.22
Three commercial glass transponders of different sizes used for application in the LF range.

TABLE 2.1
Operational Characteristics of Glass Transponders

Main Operational Parameters Glass Transponders Texas Instruments – LF Region	
Modulation	FSK: 134.2 kHz and 123.2 kHz
Transmission Principle	Half Duplex
Read Range	Less than 100 cm
Maximum Operating Temperature Range	– 25°C to + 85°C
Read Time	Less than 70 ms
Case material	Glass
Protection	Hermetic seal
Weight	0.8 grams
Memory on board	4 different options depending on specific part number

Other transponders operating in the LF region use planar coil antennas similar to those shown in Figure 2.23. The dimensions of the larger transponder are 25 mm for the outer ring and 20 mm for the inner ring, while the dimensions for the smaller antenna are 14 mm for the outer ring and 9 mm for the inner ring.

These antennas consist of a length of conductor wire coiled multiple times in order to create the desired inductance. The manufacturer specifies their use for the same general type of applications as the glass transponders, with the advantage of being much thinner.

Antennas for transponders operating in the HF frequency (13.56 MHz) are normally planar antennas, having a structure similar to the planar antennas used in the LF region. However, because HF antennas require a lower

© 2011 by Taylor and Francis Group, LLC

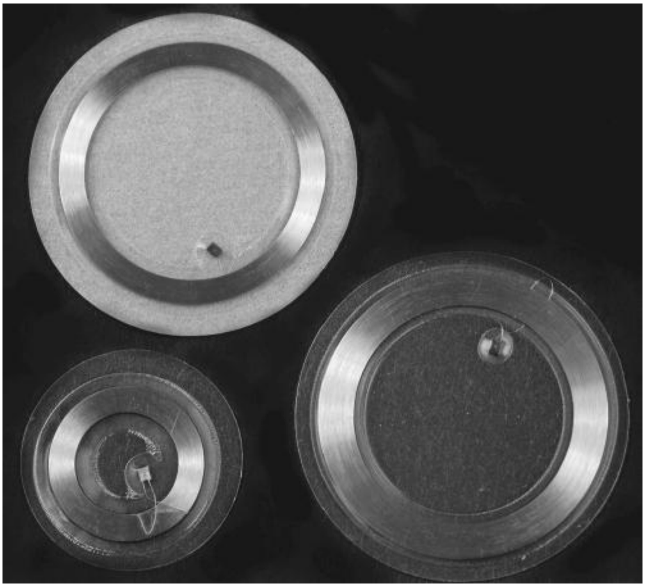

FIGURE 2.23
Commercial planar coil antennas.

inductance, the number of turns for these antennas is much lower, normally less than 10 turns. Planar antennas for RFID transponders operating in the HF region are also manufactured in different sizes and thicknesses depending on the intended application. Figure 2.8 showed a picture of an HF transponder manufactured by Texas Instruments under the name of Tag-it™. This type of transponder is marketed toward product authentication, library applications, supply chain management, asset management, and ticketing or stored value applications. The antennas are built over a polyethylene-terephthalate (PET) substrate giving a maximum thickness of 0.085 mm for the antenna area and 0.355 mm for the chip area. Figure 2.24 shows three antennas from this same family. The larger rectangular antenna has maximum dimensions of 75 mm × 45 mm and consists of seven turns; the small rectangular antenna has a maximum dimension of 39 mm × 22 mm and has six turns; the square antenna has maximum dimensions of 45 mm × 45 mm and contains nine turns.

The number and type of different antennas for transponders operating in the UHF region are higher as each commercial product has been optimized with an application in mind in terms of the different antenna parameters or the type of substrate better suited for a specific purpose. Figures 2.25, 2.26, and 2.27 show three examples of the diverse UHF transponders manufactured by Alien Technology.

Figure 2.25 shows the Alien ALN-9640 Squiggle® Inlay from Alien Technology. This is a general-purpose transponder used for packages that

© 2011 by Taylor and Francis Group, LLC

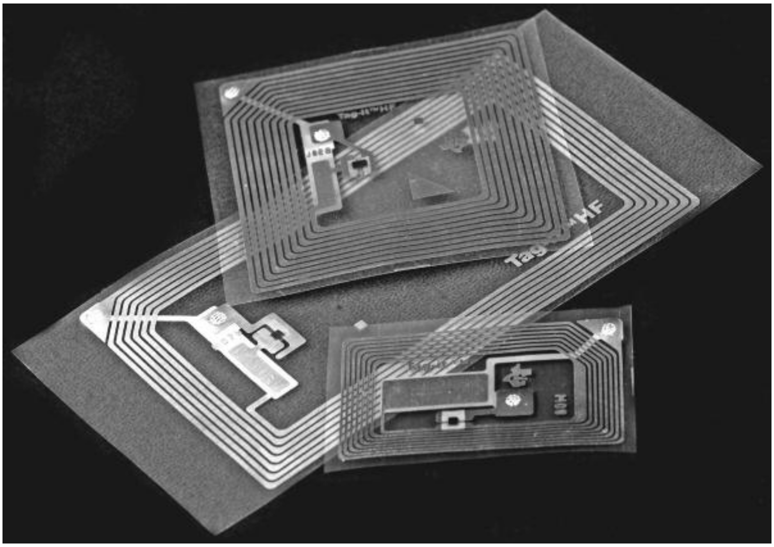

FIGURE 2.24
Diverse planar rectangular antennas used in commercial HF transponders.

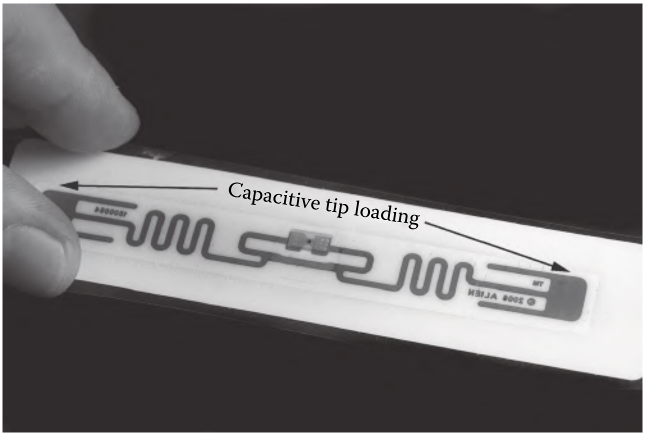

FIGURE 2.25
Commercial UHF transponder based on a meandered dipole and using capacitive tip loading.

include metal or water. The largest length dimension of this antenna is 98 mm, and the largest width of the overall antenna is 8 mm. This antenna exhibits a radiation pattern similar to a half-wave dipole. This antenna is based on a half-wave dipole that has been made shorter by bending the wires and adding shunt and series inductors to create the adequate resonant frequency for the integrated circuit that will be powered by the antenna. The large conductive areas located at each end of the antenna serve as capacitors

© 2011 by Taylor and Francis Group, LLC

used to increase the antenna capacitance. This technique is known as *capacitive tip loading*. Since the magnitude of capacitive reactance decreases as the capacitance increases, a tip-loaded dipole exhibits a more inductive behavior than a dipole of the same length without the additional capacitance at its ends.

Figure 2.26 shows the Alien ALN-9534 2x2 Inlay from Alien Technology. This antenna exhibits a more omnidirectional radiation pattern and is marketed toward item-level tagging applications such as apparel or baggage. The maximum dimensions of this antenna are 46 mm × 40 mm. This type of antenna, in which the conductors are thicker than the previously studied wire antennas, presents the advantage of having higher capacitance and lower inductance than the wire antennas, therefore making it easier to match to the characteristics of its integrated circuits. Furthermore, these antennas made with thicker conductors exhibit a wider bandwidth than the wire antennas.

Figure 2.27 shows the Alien ALN-9554 M Inlay from Alien Technology. This antenna exhibits a better omnidirectional radiation pattern and is marketed toward tagging plastic totes, pallets, and other reusable assets. The maximum dimensions of this antenna are 98 mm × 38 mm. This type of antenna presents the higher bandwidth for the same reasons described earlier. When using conductive inks to create the antenna, the additional ink required to create them results in a cost increase for this type of antenna.

Figure 2.28 shows an example of a meandered dipole antenna with an inner loop for use in the UHF range manufactured by RSIID technologies.

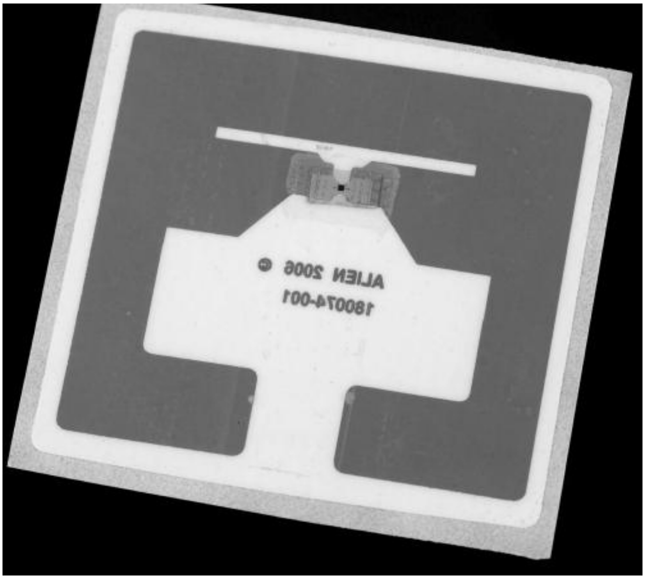

FIGURE 2.26
Commercial UHF transponder using a more omnidirectional antenna.

© 2011 by Taylor and Francis Group, LLC

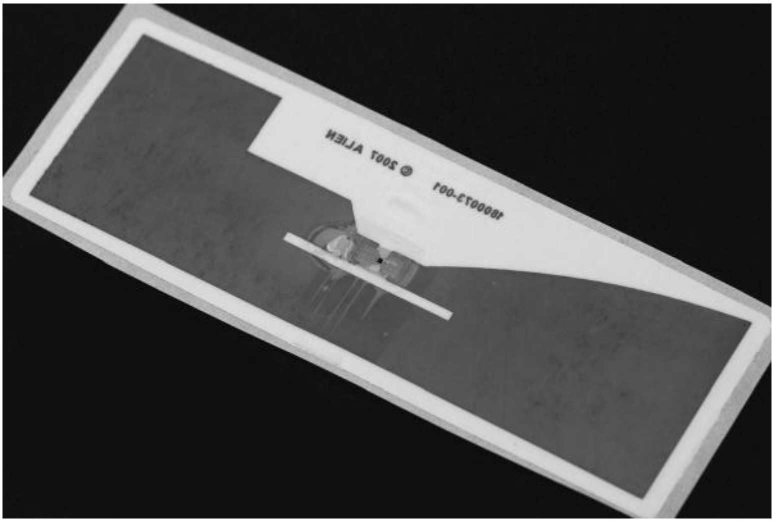

FIGURE 2.27
Commercial UHF transponder having a higher bandwidth response.

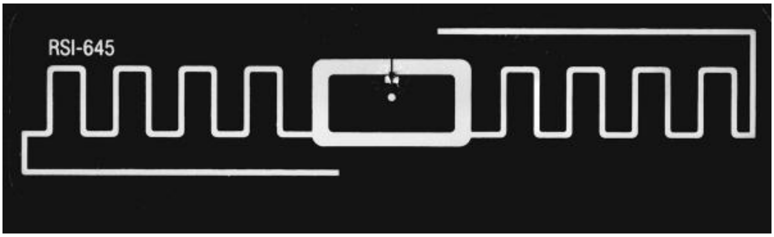

FIGURE 2.28
Meandered dipole antenna for UHF applications.

The maximum dimensions of this antenna are 67 mm × 13.5 mm, thus making it suitable for applications with limited available space. Figure 2.28 is just one example of the different configurations available for meandered dipoles.

Figure 2.29 shows a dual dipole transponder. These transponders have two dipole antennas to make them more insensitive to the orientation of the transponder in reference to the interrogator. The potential drawback of this approach is the additional space required by the transponder that may not be suitable for all products. Furthermore, the front end of the integrated circuit that will be powered by the antenna increases in complexity as it needs to respond to two different antennas. However, for longer range applications in which the position of the object can change, this type of antenna presents a clear advantage.

© 2011 by Taylor and Francis Group, LLC

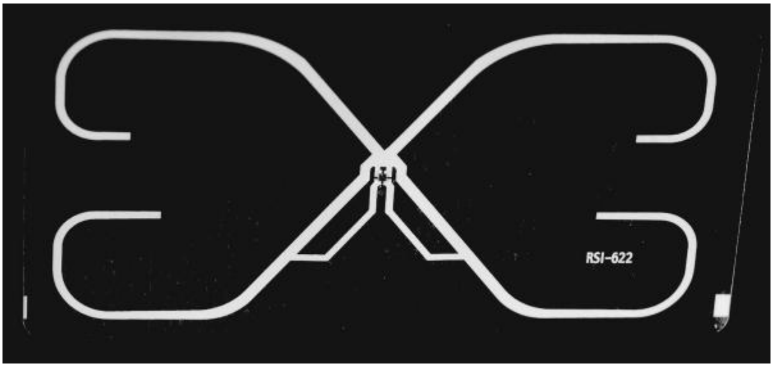

FIGURE 2.29
Dual dipole antenna for UHF transponders.

While the majority of transponders operating in the UHF range are powered by far-field electromagnetic radiation, in some cases the relatively large area that they require may be larger than the available surface in the transponder. The solution in this case is to use UHF antennas that operate in the near-field region, accepting the drawbacks associated with this mode of operation. These antennas operate based on the same inductive coupling principles used by LF and HF antennas. However, it is possible to add shunt and series inductors or capacitive tip loads to create what are known as *hybrid antennas*. These hybrid antennas, examples of which are shown in Figures 2.30 and 2.31, are able to operate in the near and far fields even though the performance in the far-field region is degraded compared to the performance of a typical dipole antenna.

2.5 The Connection between the Chip and the Antenna in RFID Transponders

A critical step in the manufacture of RFID transponders is the connection between the antenna and the integrated circuit in the transponder. There are main methods of doing this connection: chip-on-board and direct die attachment.

Chip-on-board (COB) is an assembly technology for semiconductors in which the electronic chip is directly mounted on and connected to its final board instead of undergoing the traditional packaging process for integrated circuits. This process reduces space requirements and cost, and also increases the performance of the system due to the decrease in connection lengths and increase in reliability. This approach is normally used for wire-wound antennas. During the COB process, the capacitor used to create resonance is also packaged in the device. The device is then covered in epoxy in such a

© 2011 by Taylor and Francis Group, LLC

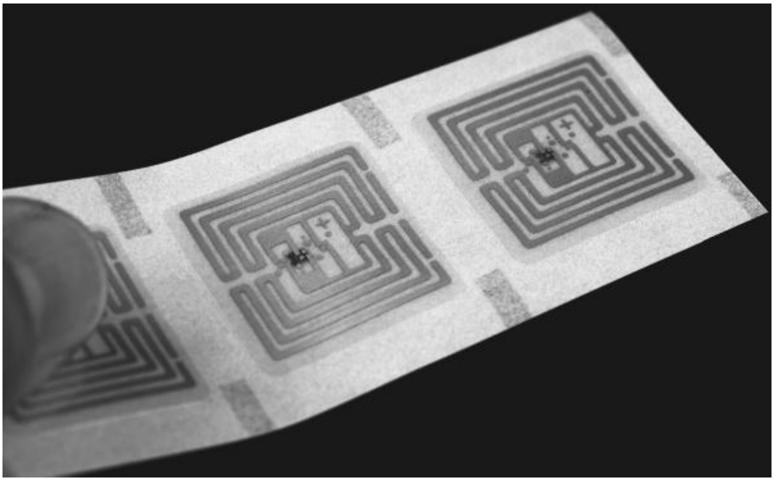

FIGURE 2.30
Hybrid UHF antenna able to communicate in near field and far field.

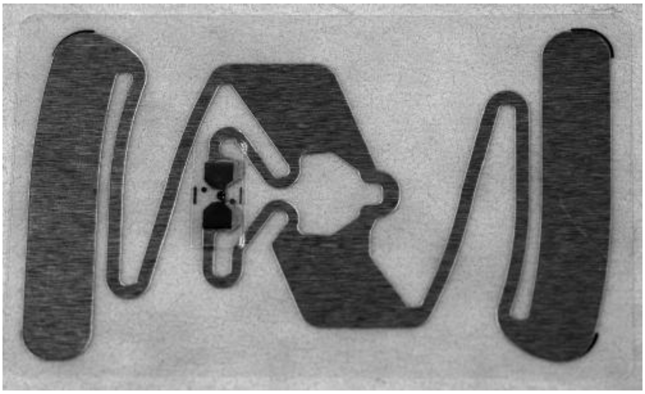

FIGURE 2.31
Hybrid UHF antenna with tip loading.

way that only the two cables to connect to the antenna emerge from the package, as shown in Figure 2.32. The majority of COB packages are used in cards that need to meet the requirements for standard thickness of these cards at 0.76 mm. With this requirement, the typical thickness of COB packages is about 0.41 mm. Although the COB package is designed to protect the internal silicon device during the card lamination process, it is necessary to be careful to prevent mechanical cracks on the device as a result of the mechanical pressures and hot temperatures.

Direct die attachment can be achieved using two techniques: wire bonding or flip-chip. Flip-chip is a technology for interconnecting semiconductor

© 2011 by Taylor and Francis Group, LLC

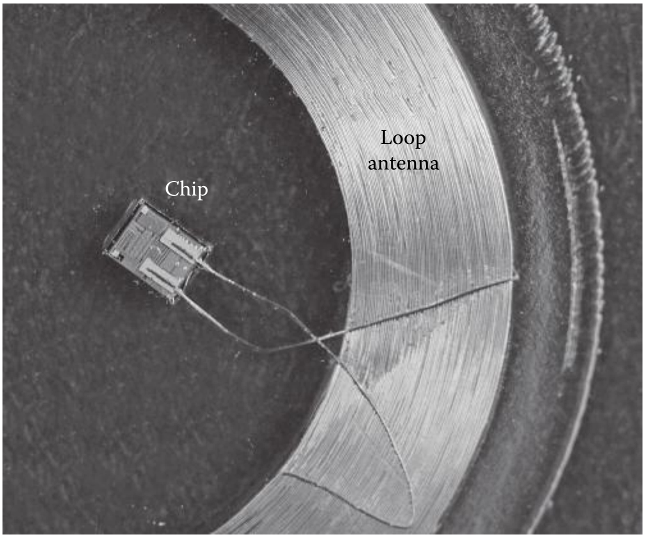

FIGURE 2.32
Connection of the antenna wires to the chip in the transponder.

devices by creating conductive bumps in the pads of the chip. In order to connect the chip, this has to be flipped so the bond pads are in contact with the matching pads in the circuit. In the wire-bonding technology, the chip is mounted upright and wires are used to connect the chip to the circuit. In the case of RFID transponders, the bumps or wires are connected to the antenna, thus being an ideal method to use for printed, etched, or stamped antennas. In this case, the resonance capacitor can also be etched on the substrate of the transponder, and therefore this technology does not require binding it with the chip as COB does. Figure 2.33 shows an example of this type of connection.

The choice between printing, etching, or stamping the antenna is a trade-off between cost and performance. The Q factor of the antenna, which as discussed earlier has a strong effect on the read range for the transponder, is inversely proportional to the resistance of the antenna. With this in mind, it is desirable to use an etched antenna instead of a printed antenna made of conductive material. The etching process is also less expensive than the printing process. For large antennas, however, etching and stamping waste too much unwanted material. In this case, a printed antenna becomes a more desirable solution. Figure 2.34 shows the general structure of the chip-to-antenna connection, independently of the specific technique being used. The antenna is connected between one of the terminals in the die and its ground connection (GND).

This connection creates a problem due to the crossing of the connection to the antenna with the ground ring around the die, creating a parasitic

© 2011 by Taylor and Francis Group, LLC

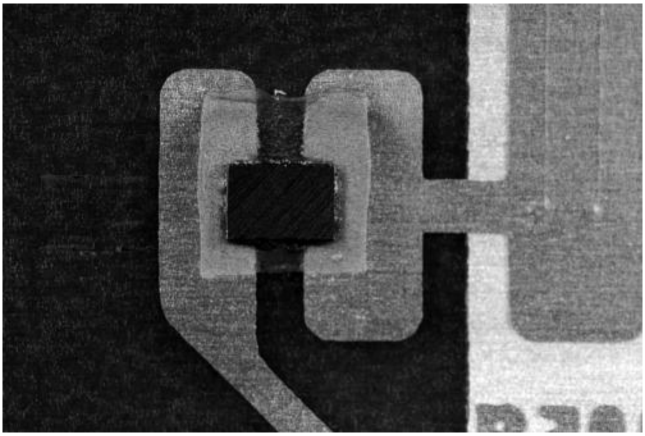

FIGURE 2.33
Direct connection between the chip and the antenna in the transponder.

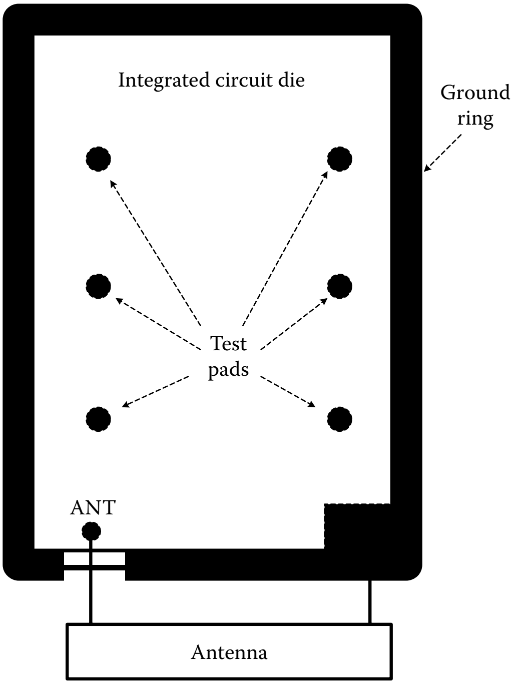
Integrated circuit die
Ground ring
Test pads
ANT
Antenna

FIGURE 2.34
Strategy for reducing the stray capacitance when connecting the chip to the antenna in an RFID transponder.

© 2011 by Taylor and Francis Group, LLC

capacitance. Depending on the technology being used to make this connection, the parasitic capacitance can have a value of up to 20 fF, influencing the resonant frequency for the antenna. To limit this effect, the ground ring around the die is narrowed in the vicinity of the pad for the antenna connection, thus reducing the parasitic capacitance.

2.6 Additional Factors That Affect the Performance of Antennas in RFID Transponders

As discussed in the preceding sections, the quality factor of an antenna is influenced by its Ohmic losses. For the coil antennas used for transponders operating in the LF region, these Ohmic losses come from the wire used to build the coil. The resistance in DC of a conductor of length *l*, with a uniform cross-section *S*, is

$$R_{DC} = \frac{l}{\sigma S} \tag{2.27}$$

where σ is the conductivity of the material.

Because of the need to manufacture the antenna as small as possible, designers opt for choosing the narrowest possible diameter, resulting in an increase in its resistance as shown by equation (2.27). More important, however, for the design of antennas is their losses in AC. The density of AC current through a wire is not constant through its surface, but the current tends to flow through the outer section of the wire, thus decreasing its effective area. This is known as the *skin effect*. Similarly, the *skin depth* for a conductor is defined as the value of its depth at which the current density falls to 37% of the current density along the surface. As shown in equation (2.28), for a given material, its skin depth is dependent on the frequency of the AC current:

$$\delta = \frac{1}{\sqrt{\pi \mu \sigma f}} \tag{2.28}$$

where

δ = the skin depth

μ = the permeability of the material

σ = the conductivity of the material

f = the frequency of the AC current through the material

© 2011 by Taylor and Francis Group, LLC

Example 2.5:

Calculate the skin depth for copper wire at (a) 125 kHz and (b) at 13.56 MHz.

Solution:

Copper has a relative permeability of 1, so its total permeability is $\mu = 4\ \pi\ 10^{-7}$ H/m. Its conductivity is equal to $\sigma = 5.8\ 10^{7}$ mho/m. Therefore, equation (2.28) can be rewritten for copper as

$$\delta = \frac{0.066}{\sqrt{f}} \quad (m)$$

A. The skin depth of copper at 125 kHz is equal to 0.186 mm. This means that the current density of a point located 0.186 mm away from the surface of the conductor is only 37% of the current density on the surface.
B. The skin depth of copper at 13.56 MHz is equal to 0.018 mm. At this frequency, the skin depth is about one-tenth of the skin depth at 125 kHz. Therefore, at 135 MHz the majority of the current is flowing through the surface of the conductor.

The total resistance of a wire to AC current (R_{ac}) is a combination of its DC resistance (R_{DC}) and its skin depth (δ):

$$R_{ac} = R_{DC}\frac{a}{2\delta} \tag{2.29}$$

where a is the radius of the wire.

For etched coil antennas, commonly found in planar transponders, the AC resistance is given by equation (2.30):

$$R_{ac} = \frac{1}{(w+t)}\sqrt{\frac{\pi\mu f}{\sigma}} \tag{2.30}$$

where

w = the width

t = the thickness of the conductor on the substrate

Etching is a wasteful process as metal is removed from the substrate in order to create the antenna. Although some of the materials may be recycled, etching also requires large amounts of chemicals and energy to use in the process. To respond to this concern, newer processes based on conductive inks or copper deposits have been developed, especially for antennas operating in the UHF frequency range. Because of the higher conductivity of copper compared to silver-based conductive inks, antennas made with deposited copper can be substantially thinner than those made with conductive inks. In general, both materials can be effectively used to create antennas.

© 2011 by Taylor and Francis Group, LLC

3

Transponders

CONTENTS

3.1 The Analog Front End 50
  3.1.1 Radiofrequency Stage 50
  3.1.2 Power Management 54
3.2 Signals in the Transponder 57
  3.2.1 Signal Encoding 57
  3.2.2 Modulation and Demodulation 59
3.3 The Logic System 66
  3.3.1 Data Memory 66
  3.3.2 Cyclic Redundancy Check for Error Detection 75
3.4 Transponder Programming 77
3.5 Summary of Electrical Specifications for RFID Transponders 79
3.6 Mechanical Considerations and Transponder Assembly 81

The two main components of a transponder are its internal circuits and its antenna. The antenna, studied in Chapter 2, is used to collect energy from the electromagnetic fields in which the transponder is embedded as well as to transmit the information back to the interrogator. The integrated circuit, also known as the *chip* or *device*, is the key component in the transponder, and for this reason this chapter uses the names *transponder*, *chip*, and *device* interchangeably. The integrated circuits have the ability to store information to be transmitted to the interrogator, execute a series of commands, and, in some cases, store new information sent by a remote station.

This chapter describes the fundamental blocks that make up the transponder. Because there are a large number of transponders in the market, each one with its own specific differences, this chapter focuses on the functional blocks common to them using commercial transponders to illustrate these concepts. Therefore, the devices used in the examples should not be seen as an exhaustive list of transponders but just as examples used to illustrate specific concepts and ideas. This chapter starts by describing the analog front end commonly encountered in these devices. The analog front end contains the radiofrequency section that is necessary to match the electrical properties of the device to the electrical properties of the antenna as well as the circuits used for harvesting and managing the power necessary to turn on the device. This section is followed by the description of the main encoding and modulation methods commonly used by transponders. This is followed

© 2011 by Taylor and Francis Group, LLC

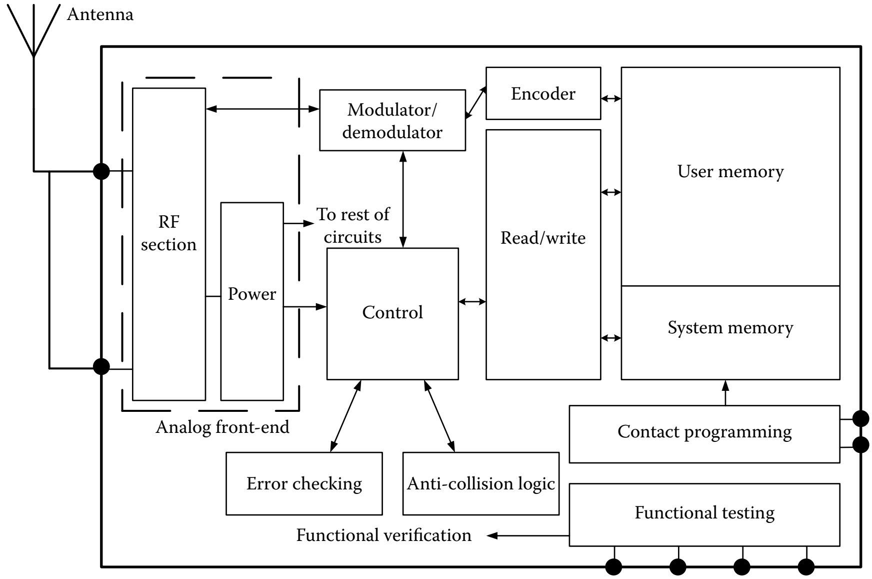
Antenna
RF section
Power
To rest of circuits
Modulator/ demodulator
Encoder
Read/write
User memory
System memory
Control
Analog front-end
Error checking
Anti-collision logic
Contact programming
Functional testing
Functional verification

FIGURE 3.1
Block diagram of a generic RFID transponder highlighting its main functional components.

by the description of the different memory structures to store data and configuration parameters as well as the codes used to ensure the integrity of the digital data. The next section briefly describes the methods for programming the transponder by either a contact programmer or a contactless programmer. The last two sections in this chapter present a summary of the electrical and mechanical specifications found in commercial transponders as well as a description of their most common assembly methods.

Figure 3.1 depicts the block diagram of a basic and generic RFID transponder showing its main functional blocks. It is important to note that not all the blocks shown in the figure will be always present in any transponder depending on its use, intended application, and cost.

3.1 The Analog Front End

3.1.1 Radiofrequency Stage

When the transponder is immersed in an electromagnetic field of the appropriate frequency, a radiofrequency voltage appears across the antenna terminals. The task of the front-end stage is to rectify that radiofrequency voltage and convert it into a continuous voltage (DC) with a value high enough to power the rest of the circuits inside the device. Different transponders have

© 2011 by Taylor and Francis Group, LLC

different requirements regarding the minimum voltage required to become operational depending on their internal structure and fabrication technology.

The first step in the energy conversion process between the electromagnetic field and the DC voltage is the resonant circuit tuned to the frequency of the field. The different possible configurations for the external resonant circuit described in Section 2.2 are now studied, taking into account the structures used in commercial transponders. Figure 3.2 shows the possible alternatives for the external resonant circuit using two commercial devices: the MCRF355 and the MCRF450, both from Microchip Technology and both operating at 13.56 MHz.

Some devices incorporate one or two capacitors inside their integrated circuit, eliminating the need for using an external capacitor and therefore reducing the size required for the overall transponder inlay. Figure 3.3 shows the resonant circuit using commercial devices that incorporate one or two capacitors in their internal circuitry.

The devices MCRF360, MCRF451, and MCRF455 from Microchip Technology have a single internal capacitor with values of 100 pF, 95 pF, and 50 pF, respectively. The device MCRF422, also from Microchip Technology, has two internal capacitors of 50.6 pF and 65.4 pF. All these devices operate at 13.56 MHz.

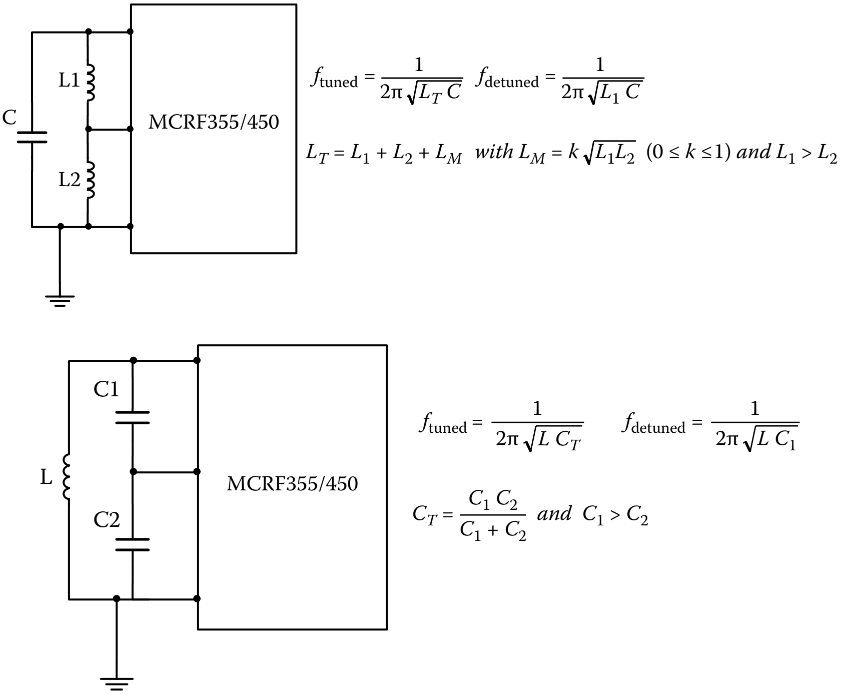

$$f_{\text{tuned}} = \frac{1}{2\pi\sqrt{L_T C}} \quad f_{\text{detuned}} = \frac{1}{2\pi\sqrt{L_1 C}}$$

$$L_T = L_1 + L_2 + L_M \ \textit{with}\ L_M = k\sqrt{L_1 L_2}\ \ (0 \le k \le 1)\ \textit{and}\ L_1 > L_2$$

$$f_{\text{tuned}} = \frac{1}{2\pi\sqrt{L\,C_T}} \qquad f_{\text{detuned}} = \frac{1}{2\pi\sqrt{L\,C_1}}$$

$$C_T = \frac{C_1 C_2}{C_1 + C_2}\ \textit{and}\ C_1 > C_2$$

FIGURE 3.2
Alternatives for the external resonant circuit found in commercial transponders. These transponders require a capacitor in the inlay to create the resonant circuit. The inductor in the circuits models the antenna.

© 2011 by Taylor and Francis Group, LLC

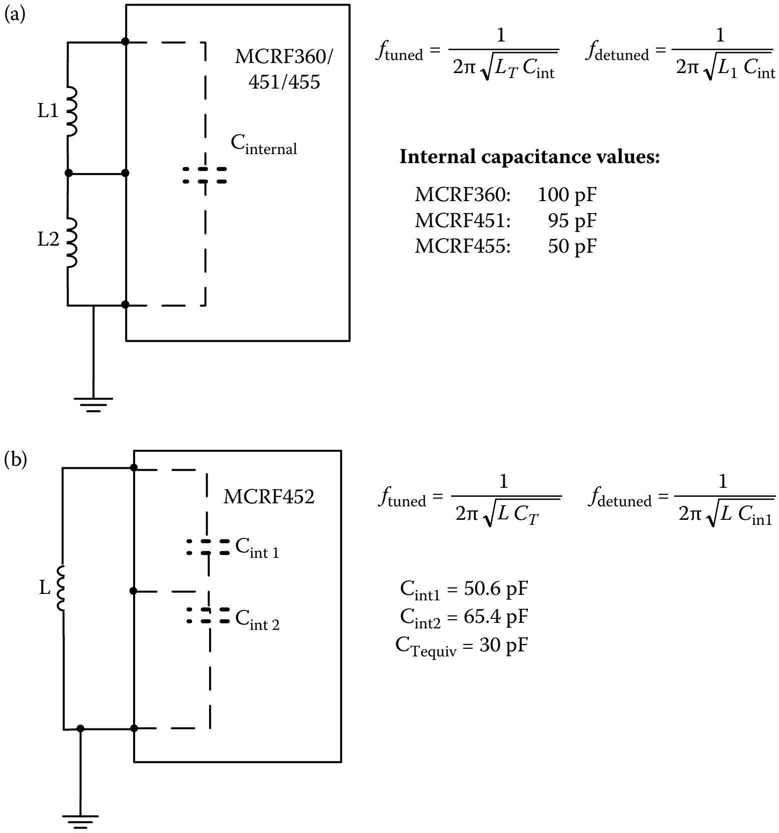

$$f_{\text{tuned}} = \frac{1}{2\pi\sqrt{L_T\, C_{\text{int}}}} \qquad f_{\text{detuned}} = \frac{1}{2\pi\sqrt{L_1\, C_{\text{int}}}}$$

Internal capacitance values:

MCRF360:	100 pF
MCRF451:	95 pF
MCRF455:	50 pF

$$f_{\text{tuned}} = \frac{1}{2\pi\sqrt{L\, C_T}} \qquad f_{\text{detuned}} = \frac{1}{2\pi\sqrt{L\, C_{\text{in1}}}}$$

$C_{\text{int1}} = 50.6$ pF
$C_{\text{int2}} = 65.4$ pF
$C_{\text{Tequiv}} = 30$ pF

FIGURE 3.3
Alternatives for the external resonant circuit for devices that incorporate the resonance capacitor in the die of the integrated circuit. (a) Structure with a single internal capacitor. (b) Structure with two internal capacitors.

The path between one of the antenna connection nodes and ground, shown in Figures 3.2 and 3.3, is shortened according to a control signal. The path is shortened using a transistor, called a *modulation transistor,* that exhibits a low ON resistance, normally below 5 Ω and a very high OFF resistance of several MΩ. In this situation, the transponder exhibits two different resonant frequencies.

The resulting resonance frequency when the modulation transistor is OFF is chosen to be equal to the frequency of the electromagnetic field emitted by the interrogator. As shown in Figure 3.4, because of the resonance condition, the energy of the field is transmitted through the front end to the device, resulting in the radiofrequency voltage at the input of the device being maximal. This situation is called *uncloaking*. When the modulation transistor is turned ON, it shortens one of the inductors or capacitors, resulting in a resonant frequency for the circuit different from the frequency of the electromagnetic field. When it shortens one of the two inductors connected

© 2011 by Taylor and Francis Group, LLC

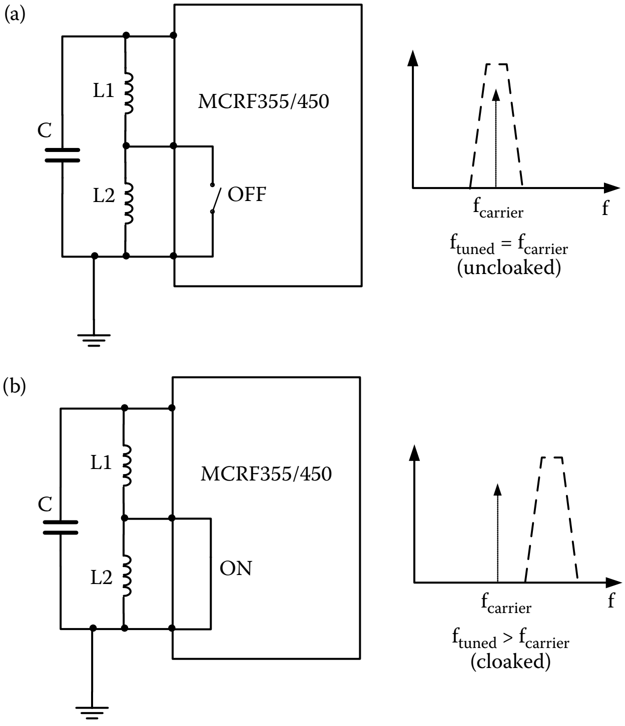

FIGURE 3.4
Cloaking and uncloaking modes produced by the control transistor in the transponder and the different resonant frequencies generated. (a) Uncloaked. (b) Cloaked.

in series, the resulting resonant frequency is higher than the frequency of the field; when it shortens one of the two capacitors connected in series, the resulting resonant frequency is lower. In any case, the frequency of the circuit and the frequency of the field are different, and therefore most of the energy of the electromagnetic field is rejected by the filter, resulting in the voltage generated at the input of the device being minimal, ideally zero. This situation is called *cloaking*. The same cloaking–uncloaking approach is used by the transponder to transmit data to the interrogator. Using this approach, the digital signal controls the modulation transistor.

Example 3.1:

Calculate the values of the external inductors needed to use in a transponder based on the MCRF451 device.

Solution:

The MCRF451 has an internal capacitance of 95 pF and operates at a frequency of 13.56 MHz. Using the equations shown in Figure 3.3, the total inductance

© 2011 by Taylor and Francis Group, LLC

necessary is L_T = 1.45 μH. With this inductance, the circuit resonates at 13.56 MHz, and therefore the energy from the electromagnetic field is passed to the device. The L_T = 1.45 μH of external inductance will be split between L_1 and L_2.

The value of L_1 can be calculated using the equations from Figure 3.3 now for the case of detuned frequency. The detuned frequency should be between 3 MHz and 6 MHz away from the tuned frequency. For the structure used by the MCRF451 device, the detuned frequency will be higher than the tuned frequency. Using for example a frequency shift of 4 MHz, the detuned frequency becomes 17.56 MHz, resulting in a value of L_1 equal to 864 nH. Assuming initially a value for the coupling coefficient between L_1 and L_2 equal to zero ($k = 0$), the required value for L_2 is 586 nH. It can be observed that $L_1 > L_2$ as it is required. In a more realistic situation, the value of the coupling coefficient will be between 0 and 1, resulting in a lower value for L_2.

3.1.2 Power Management

In order to be operative, the transponder needs to convert the radiofrequency voltage detected by the antenna into a DC voltage. The voltage required to bias the internal components in the transponder is higher than the voltage detected by the antenna. Therefore, the transponder requires the use of voltage multipliers to reach the values necessary by the biasing voltage.

A voltage multiplier is a circuit that converts a lower AC voltage into a higher DC voltage. Figure 3.5 shows the basic structure of a simple voltage multiplier called a *voltage doubler*.

Assuming initially ideal diodes for simplicity, diode D1 is forward biased during the phase of negative input voltage, while diode D2 is reverse biased and therefore capacitor C2 is disconnected from the rest of the circuit. In this stage, capacitor C1 charges to a DC voltage ideally equal to the amplitude of the AC signal. Afterward, during the phase of positive input voltage, the biasing of the diodes reverses: D2 is forward biased, and D1 becomes reverse biased. This causes capacitor C2 to be directly connected to the input of the AC signal. Taking into consideration the voltage at which capacitor C1 was originally charged, the DC voltage at capacitor C2 will be equal to twice the amplitude of

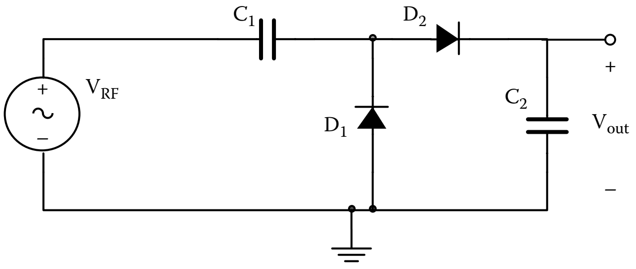

FIGURE 3.5
Structure of a single-stage basic voltage multiplier. V_{RF} models the voltage generated by the electromagnetic field in which the transponder is immersed.

© 2011 by Taylor and Francis Group, LLC

the input voltage. Considering now the minimum ON voltage for real diodes, the voltage at the output of the structure shown in Figure 3.5 is

$$V_{out} = 2(V_p - V_{ON}) \tag{3.1}$$

where

V_{out} = the output DC voltage,

Vp = the peak voltage detected at the antenna terminals

V_{ON} = the minimum direct voltage across the diode to turn it on

Figure 3.6 shows the intermediate signals observed at the different nodes in the voltage doubler structure. The voltage supply models the radiofrequency signal detected at the terminals of the antenna in the transponder. The graph of the output voltages shows how the output capacitor is being charged until it reaches the steady-state level predicted by equation (3.1). Once the transient disappears, the output voltage stabilizes and will remain at that level as long as the transponder is immersed in that electromagnetic field.

In practical RFID systems, because the RF voltage detected at the antenna terminals is relatively small, the voltage doubler structure shown in Figure 3.5 does not produce a high enough voltage to power the integrated circuit in the transponder. This situation can be resolved by connecting additional voltage doublers. The resulting structure is known as a *Dickson charge pump*, resulting in an output DC voltage equal to

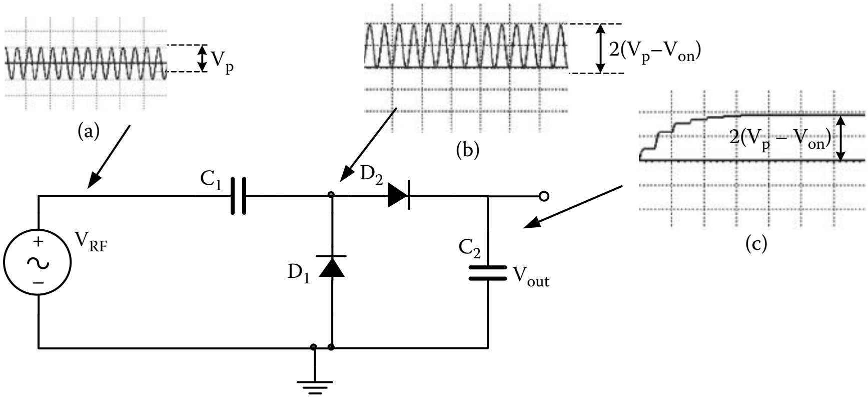

FIGURE 3.6
Signals observed at diverse points in the single-stage voltage multiplier. (a) Input signal. (b) Intermediate signal consisting of the input signal plus a DC component equal to its peak value. (c) Output signal. After the transient disappears, the output signal is a DC signal with a value equal to the double of the peak amplitude of the input signal.

© 2011 by Taylor and Francis Group, LLC

$$V_{out} = 2N(V_p - V_{ON}) \tag{3.2}$$

where N is the number of basic voltage doubler cells.

Figure 3.7 shows a three-stage Dickson charge pump. Its output voltage is equal to six times the peak voltage for the input minus the turn-on voltage for the diodes. Figure 3.8 shows the intermediate signals observed at several nodes in this structure. The sensitivity of the different graphs in volts per division has been kept the same to better observe how at each intermediate node, the AC signal has an additional offset equal to its peak value. The final DC output voltage is predicted by equation (3.2).

There is, however, a practical limit to the number of stages that is reasonable to consider for use in a practical application. As the number of stages increases, the number of required diodes increases by a factor of 2, and therefore the power dissipated by the charge pump also increases in the same ratio, decreasing the overall efficiency of the circuit. The efficiency of the charge pump is given by equation (3.3):

$$\eta_{pump} = \frac{P_{load}}{P_{load} + P_{consumed}} = \frac{V_{out}}{V_{out} + 2N\,V_{ON}} \tag{3.3}$$

where N is the number of stages in the charge pump.

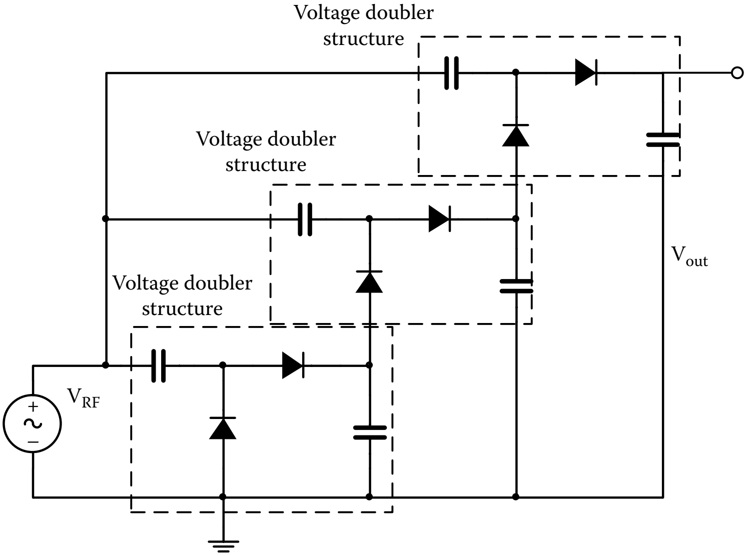
Voltage doubler structure
Voltage doubler structure
Voltage doubler structure
V_{out}
V_{RF}
+
−

FIGURE 3.7
Dickson charge pump structure used to increase the DC voltage extracted from the radiofrequency signal generated by the electromagnetic field in which the transponder is immersed.

© 2011 by Taylor and Francis Group, LLC

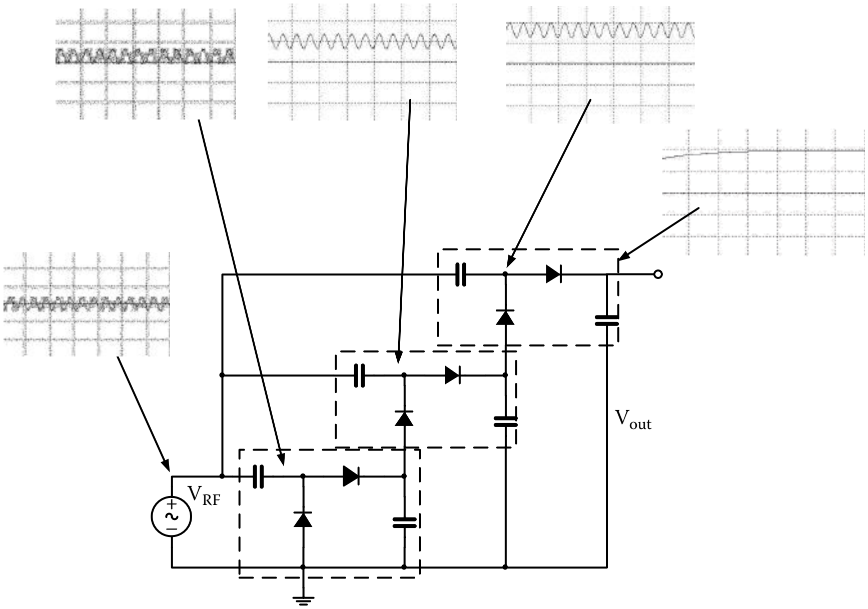

FIGURE 3.8
Signals observed at different points in the Dickson charge pump. After each stage, the overall signal increases its DC component until it only contains a DC component at the output of the last stage.

Equation (3.3) shows that the efficiency of the charge pump will increase by using diodes with a lower turn-on voltage. To this extent, several types of devices have been used and several more types are being currently investigated. The most common approach is to use Schottky diodes due to their electrical characteristics that make them extremely attractive for this purpose. However, using Schottky diodes increases the complexity of manufacturing the transponder, thus resulting in a higher cost. For this reason, the use of ultra-low-power complementary metal oxide semiconductor (CMOS) diodes appears to be an attractive alternative as they can be manufactured with the rest of the transponder and therefore they will not increase the cost of manufacturing the transponder.

3.2 Signals in the Transponder

3.2.1 Signal Encoding

The digital signals in the transponder are encoded using one or more of these common encoding schemes that are represented in Figure 3.9.

© 2011 by Taylor and Francis Group, LLC

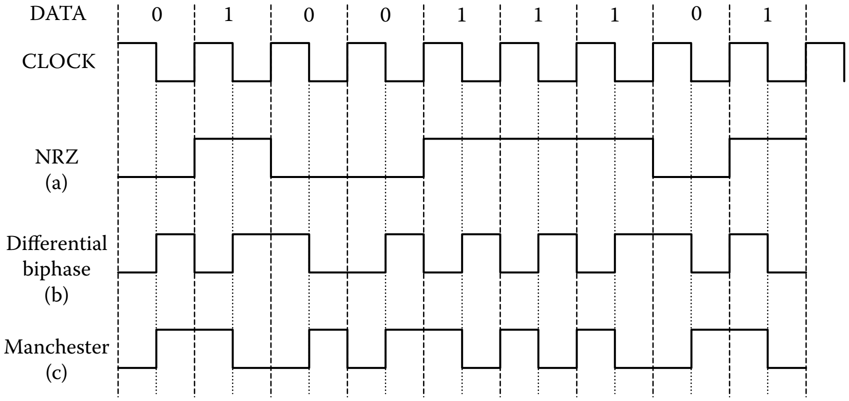

FIGURE 3.9
Common signal-encoding techniques. (a) Nonreturn-to-zero (NRZ): *1* represented by a high logic level; *0* represented by a low logic level. (b) Differential biphase: *1* represented by a change in level at the start of the clock period; *0* represented by no change in level at the start of the clock period. There is always a change of level at the middle of the clock period. (c) Manchester: *1* represented by a high-to-low transition at the middle of the clock period; *0* represented by a low-to-high transition at the middle of the clock period.

These encoding techniques are as follows:

NRZ (nonreturn to zero): There is no data encoding done. The 1′s and 0′s are clocked from the data array directly to the output transistor that will close or open the switch in the analog front end, as was shown in Figure 3.4. Although extremely simple, this encoding technique presents problems if the data to transmit have long strings of 1′s or 0′s because these may be misunderstood as the presence of a DC voltage in the line.

Differential biphase: Multiple encoding approaches using this method. This method embeds clocking information with the data and will help to synchronize the interrogator to the bit stream. In this approach, a level change occurs at the middle of every bit clock period. *1* is represented by a change in level at the start of the clock. *0* is represented by no change in level at the start of the clock.

Manchester code (biphase level, biphase_L, or split phase): This is a variation of the previous encoding method. There is not always a transition at the clock edge, but there is always a transition in the middle of the clock cycle, thus allowing one to extract the clock signal from the data signal. In Manchester code, a *1* is represented by a high to low level change in the middle of the clock. A *0* is represented by a low to high level change at the start of the clock. This type of encoding is used in the MCRF355/360 and the MCRF45X family of transponders manufactured by Microchip Technology.

© 2011 by Taylor and Francis Group, LLC

Example 3.2:

Assume that we are using a biphase-L format (Manchester) for encoding the digital data at a frequency of 70 kHz. The data bit *0* is sent by first detuning (cloaking) the device during 7 μs and then tuning (uncloaking) the device for an additional 7 μs. The data bit *1* is sent by uncloaking the device and then cloaking it for the same time periods.

3.2.2 Modulation and Demodulation

Modulation refers to changing one or more parameters in the carrier radiofrequency signal as a method to transmit the encoded signal. In practice, only one of the parameters (amplitude, frequency, or phase) of the radiofrequency is usually changed when transmitting simple digital data.

When the amplitude of the carrier changes depending on the encoded signal, the resulting modulated signal is called *amplitude shift keying*, or ASK. In this case, a high voltage in the envelope of the radiofrequency signal signifies a logic level *1,* while a low voltage in the envelope signifies a logic level *0,* following the appropriate signal-encoding method being used. ASK modulation is also known as *direct modulation* in RFID systems. ASK modulation offers the possibility of high data rates due to the simplicity of the process. The spectrum of the ASK signal is relatively narrow as it contains only energy at the frequency of the carrier signal and at each of the two sidebands. One sideband is located at the frequency of the carrier minus the frequency of the modulating signal, and the other sideband is located at the frequency of the carrier plus the frequency of the modulating signal. The bandwidth efficiency of a binary-modulated ASK signal is 1 bit/second/Hz. The detection of the ASK signal can be done using a coherent detector or a noncoherent detector. Noncoherent detection increases the simplicity of the overall system, although it reduces the ability to differentiate between the desired signal and noise. The immunity to noise can be increased by using a coherent detector, although the immunity to noise is overall lower than using an FSK or PSK modulation process. The bit error probability for noncoherent ASK modulation, assuming an additive white Gaussian noise limited channel, is

$$P\big|_{\substack{ASK\\NONCOH}} = 0.5e^{\frac{-E_b}{4N_0}} + 0.5erfc\left(\sqrt{E_b/2N_0}\right) \tag{3.4}$$

where

P = the bit error probability

E_b/N_0 = the average bit power

erfc(x) = the complementary error function for the argument *x*

© 2011 by Taylor and Francis Group, LLC

Similarly, the bit error probability for coherent ASK modulation is

$$P\big|_{\substack{ASK\\COH}} = 0.5 erfc\left(\sqrt{E_b/2N_0}\right) \tag{3.5}$$

Figure 3.10 shows the bit error probability values obtained using equations (3.4) and (3.5) for different values of energy transmitted. From this graph it is possible to observe the decrease in bit error rates for coherent ASK modulation, which is especially significant when increasing the energy per bit transmitted.

When the encoding signal changes the amplitude of the carrier being transmitted, the resulting modulated signal is called *frequency shift keying*, or FSK. When the modulating signal is binary, FSK results in using two different frequencies for transmitting the digital data. The most common approach to FSK in RFID systems is called Fc/8/10: this means that a *0* is transmitted as an amplitude modulated clock cycle with a period corresponding to the carrier frequency divided by 8, and a *1* uses a period that corresponds to the carrier frequency divided by 10. Figure 3.11 shows the time representation for this FSK modulation.

The logic level *0* is transmitted by sending a total of eight cycles of the RF signal; the first four cycles have a higher amplitude than the last four cycles. The logic level *1* is transmitted by sending a total of ten RF cycles,

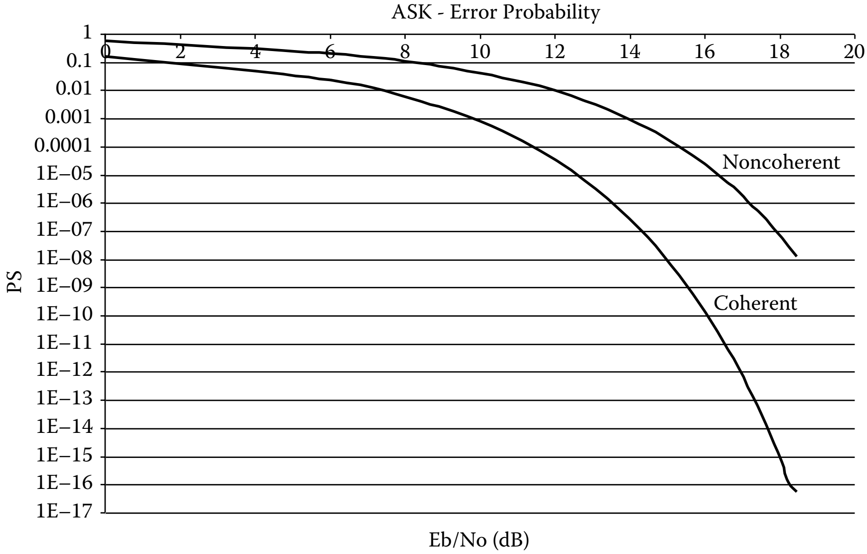

FIGURE 3.10
Error probability (Ps) for coherent and noncoherent ASK modulation as a function of the energy received (Eb/No).

© 2011 by Taylor and Francis Group, LLC

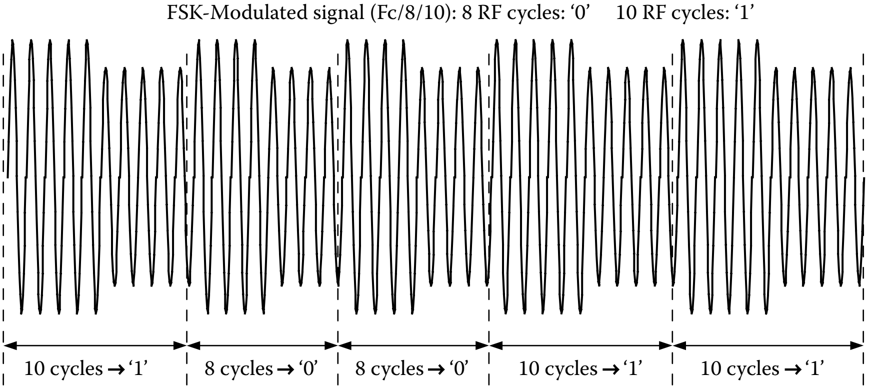

FIGURE 3.11
Structure of the FSK-modulated signal. The symbol *1* is represented by sending 10 RF cycles, while the symbol *0* is represented by sending 8 RF cycles. The message sent in the figure is *10011*.

with the first five cycles having a higher amplitude than the last five cycles. The change in amplitude between the first and second halves of the cycles allows one to count the number of cycles between transitions and therefore to differentiate between logic levels. In other words, the *1* and *0* are differentiated by looking at the number of periods between transitions.

Other FSK approaches used by other transponders are Fc/10/8, Fc/5/8, and Fc/8/5. Table 3.1 lists the relation between the encoding signal and the data transmitted.

The bandwidth required by the FSK signal is dependent on the existence or not of phase changes when changing frequencies. Obviously, in those FSK modulation approaches that exhibit abrupt phase changes, there will be more spectral components at higher frequencies, thus increasing the bandwidth. The bandwidth efficiency of a binary-modulated FSK signal depends on how the modulation process has been performed, but it approaches 2 bits/second/Hz. Similarly to ASK modulation, FSK signals can be recovered using noncoherent detectors or coherent detectors, with a coherent detector providing stronger noise immunity. The calculation of the bit error rate for FSK signals depends on the separation between the frequencies used to transmit the FSK signal. The optimal frequency selection occurs when these frequencies are a multiple of the frequency of the encoding signal. These are called *orthogonal frequencies.*

The bit error probability for noncoherent FSK modulation, assuming an additive white Gaussian noise limited channel, is

$$P\big|_{\substack{FSK \\ NONCOH}} = 0.5e^{\frac{-E_b}{2N_0}} \tag{3.6}$$

© 2011 by Taylor and Francis Group, LLC

TABLE 3.1

Different FSK Modulation Approaches

	Logic Level "0"	Logic Level "1"
Fc/8/10	8 RF cycles	10 RF cycles
Fc/10/8	10 RF cycles	8 RF cycles
Fc/5/8	5 RF cycles	8 RF cycles
Fc/8/5	8 RF cycles	5 RF cycles

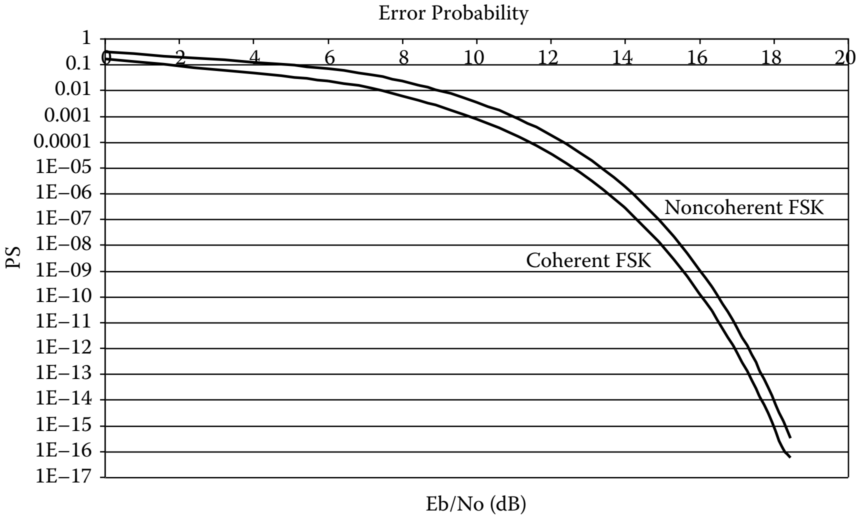

FIGURE 3.12
Error probability (Ps) for noncoherent FSK and coherent FSK as a function of the energy received (Eb/No).

The bit error probability for coherent FSK modulation is

$$P\big|_{\substack{FSK\\COH}} = 0.5erfc\left(\sqrt{E_b / 2N_0}\right) \tag{3.7}$$

Comparing equations (3.5) and (3.7) shows that if the two frequencies that make the FSK signal are chosen to be orthogonal, then the bit error rates for coherent ASK and coherent FSK signals are essentially the same. Figure 3.12 shows the probability of bit error as predicted by equations (3.3) and (3.4) for different values of energy per bit. This figure shows that the difference in performance between noncoherent detection and coherent detection for FSK signals is lower than for ASK signals. This has resulted in the widespread use of noncoherent FSK modulation due to its simplicity.

© 2011 by Taylor and Francis Group, LLC

Example 3.3:

Calculate the bit error rate for coherent and noncoherent ASK and FSK demodulation for the following signal-to-noise ratios (E_b/N_o): (a) 5 dB, (b) 10 dB, and (c) 15 dB.

Solution:

The bit error rates for each type of demodulation are given by equations (3.4) to (3.7). The first step is to express the signal-to-noise ratio in linear units. This results in the following values:

A. $E_b/N_o = 5$ dB ➔ $E_b/N_o = 3.16$ (linear units)
B. $E_b/N_o = 10$ dB ➔ $E_b/N_o = 10.0$ (linear units)
C. $E_b/N_o = 15$ dB ➔ $E_b/N_o = 31.6$ (linear units)

Bit error rates for noncoherent ASK demodulation are predicted by equation (3.4), resulting in the following:

With $E_b/N_o = 3.16$ (5 dB): $P_s = 0.264$ (approximately 1 error every 4 bits transmitted)
With $E_b/N_o = 10.0$ (10 dB): $P_s = 0.042$ (approximately 1 error every 24 bits transmitted)
With $E_b/N_o = 31.6$ (15 dB): $P_s = 0.185\ 10^{-3}$ (approximately 1 error every 5400 bits transmitted)

Bit error rates for coherent ASK demodulation are predicted by equation (3.5), resulting in the following:

With $E_b/N_o = 3.16$ (5 dB): $P_s = 0.037$ (approximately 1 error every 30 bits transmitted)
With $E_b/N_o = 10.0$ (10 dB): $P_s = 0.78\ 10^{-3}$ (approximately 1 error every 1300 bits transmitted)
With $E_b/N_o = 31.6$ (15 dB): $P_s = 9.47\ 10^{-9}$ (approximately 1 error every 100 million bits transmitted)

Bit error rates for noncoherent FSK demodulation are predicted by equation (3.6), resulting in the following:

With $E_b/N_o = 3.16$ (5 dB): $P_s = 0.103$ (approximately 1 error every 10 bits transmitted)
With $E_b/N_o = 10.0$ (10 dB): $P_s = 3.37\ 10^{-3}$ (approximately 1 error every 300 bits transmitted)
With $E_b/N_o = 31.6$ (15 dB): $P_s = 6.87\ 10^{-8}$ (approximately 1 error every 14.5 million bits transmitted)

Bit error rates for coherent FSK demodulation are predicted by equation (3.7), resulting in the following:

With $E_b/N_o = 3.16$ (5 dB): $P_s = 0.037$ (approximately 1 error every 30 bits transmitted)
With $E_b/N_o = 10.0$ (10 dB): $P_s = 0.78\ 10^{-3}$ (approximately 1 error every 1300 bits transmitted)
With $E_b/N_o = 31.6$ (15 dB): $P_s = 9.47\ 10^{-9}$ (approximately 1 error every 100 million bits transmitted)

© 2011 by Taylor and Francis Group, LLC

> These results show that the performance of coherent demodulation has a much better performance than noncoherent demodulation. Even for relatively noisy signals, coherent ASK shows a much lower bit error rate than when the demodulation is performed noncoherently. Therefore, it seems obvious to choose a coherent demodulation scheme when using ASK. While coherent FSK demodulation also shows a lower bit error rate than noncoherent FSK demodulation, the difference is not as strong as in the ASK case. For this reason, a noncoherent FSK demodulation may be preferred as it will reduce the complexity of the system without excessively degrading its performance.

When the encoding signal changes the phase of the carrier, the resulting modulated signal is called *phase shift keying* (PSK), as shown in Figure 3.13. PSK is basically the same approach as FSK with the difference being that only one frequency is used. The information is encoded here in the change of phase between clock cycles. PSK also allows for different approaches on this encoding. Because the receiver is only looking for changes of phase, it is possible to transmit information in a smaller number of radiofrequency cycles, resulting in using faster data transfer rates compared to FSK.

Some devices like the MCRF200 and the MCRF250, both from Microchip Technology, support two PSK modulation modes called PSK_1 and PSK_2. In PSK_1, the phase of the RF signal changes when the encoding data change. This means that, for example, the phase of the RF signal will change when the encoding signal changes from *1* to *0* or from *0* to *1*. In

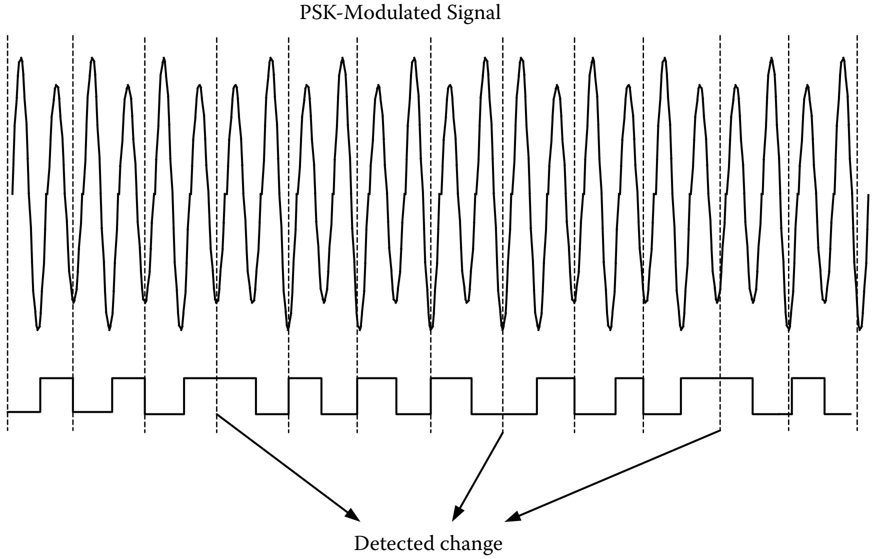

FIGURE 3.13
A generic PSK-modulated signal. The figure shows that the PSK signal requires less RF cycles to transmit the digital data.

© 2011 by Taylor and Francis Group, LLC

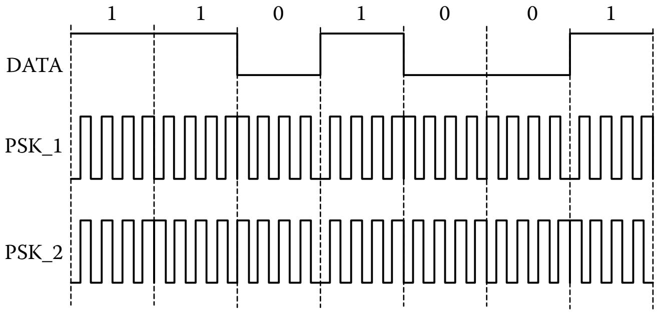

FIGURE 3.14
Two variations of PSK modulation. In PSK_1, the phase of the carrier changes every time there is a change in the symbol transmitted. In PSK_2, the phase of the carrier changes every time the symbol transmitted is *1*.

PSK_2, the phase of the RF signal changes every time the encoding data have the logic level *1*. For example, the phase will change when the encoding data change from *0* to *1* or from *1* to *1*, but it will not change when the encoding data move from *1* to *0* or from *0* to *0*. Figure 3.14 depicts these two PSK modulation approaches.

A third PSK modulation mode available in some other devices—for example, the T5557 from Atmel®—is based on a phase change on the rising edge of the encoding signal.

The spectral occupancy of the PSK signal is identical to the ASK signal assuming that there are no phase changes abruptly occurring at the symbol boundaries. However, in contrast to ASK and FSK, the detection of PSK signals requires a coherent demodulator. The bit error rate for PSK signals is

$$P\big|_{\substack{PSK \\ COH}} = 0.5erfc\left(\sqrt{E_b / N_0}\right) \tag{3.8}$$

When comparing the different modulation methods and assuming equal average energy per symbol transmitted, the different PSK modulation approaches present the strongest immunity against noise. Coherent FSK modulation with orthogonal symbols or coherent ASK modulation is the next best approach followed by the noncoherent FSK modulation. Noncoherent ASK modulation exhibits the poorest performance. For example, a bit error rate of 10^{-6} requires a value of Eb/No slightly above 10 dB for a PSK signal, but it will require a value of Eb/No of about 18 dB for a noncoherent ASK signal. Similarly, an energy Eb/No of 10 dB will result in a bit error rate of 1.5 10^{-5} in using PSK, but the bit error rate will increase up to 1.5 10^{-2} when using noncoherent ASK. This represents moving from 1.5 errors every 100,000 bits transmitted (PSK) to 1.5 errors every 100 bits transmitted (noncoherent ASK).

© 2011 by Taylor and Francis Group, LLC

3.3 The Logic System

3.3.1 Data Memory

The memory of RFID transponders is divided into data memory and configuration memory. Data memory stores the information that will be transmitted back to the interrogator. Configuration memory stores data regarding the configuration parameters for the transponder. Memory varies in size, rewriting capabilities, structure, and characteristics similar to those found in memory used in other products. The size of the memory ranges from 1 bit to 64 kbits or even larger, with the most common values being around a few hundred bits. One of the most important differences between transponders is whether the data stored in their memory can be rewritten by the user. This gives rise to the distinction between *read-only* transponders and *read/write* transponders. Because memory types, sizes, and structures can greatly differ between different transponders, the following examples are not intended to be viewed as a comprehensive guide, but to highlight some of the most commonly used approaches to memory in transponders.

The device MCRF200 from Microchip Technology contains a memory of 140 bits; 128 bits are data bits, and the remaining 12 bits are used for the configuration register. This is a one-time programmable (OTP) device that operates as a read-only device once it is programmed. The data memory can be programmed by the user using a contactless programmer or can be directly programmed by the factory at the time of production. The configuration register can only be programmed at the time of production. The manufacturer specifies memory data retention better than 200 years.

Figure 3.15 shows the block diagram of this device. The EEPROM memory is addressed by the column and row decoders at the clock rate. The output from the memory array is a bit stream that is directly fed into the modulation control and modulation circuit in order to be transmitted back to the interrogator. Figure 3.16 shows the description of the different bits for the configuration register in this memory.

CB12 is the bit that allows programming the device. Once CB12 is set to 1, the device cannot be programmed or erased. Because the MCRF200 does not support anticollision, bit CB11 is always 0. However, for the device MCRF250 that is similar to the MCRF200 with the only exception that MCRF250 supports anticollision, its bit CB11 is always set to 1. CB10 is used to establish the transmission rate when PSK modulation has been selected. Bits CB9 and CB8 establish the modulation for the encoded data using the modulation types shown in Section 3.2.2. Bits CB7 and CB6 determine how the data are encoded, as described in Section 3.2.1. Bit CB5 is always set to 0. Bits CB4, CB3, and CB2 are used to set the baud rate at which data are transferred. The default timing (CB4 = CB3 = CB2 = 0) is MOD128, meaning that the transmission rate is set at 128 RF cycles for bit.

© 2011 by Taylor and Francis Group, LLC

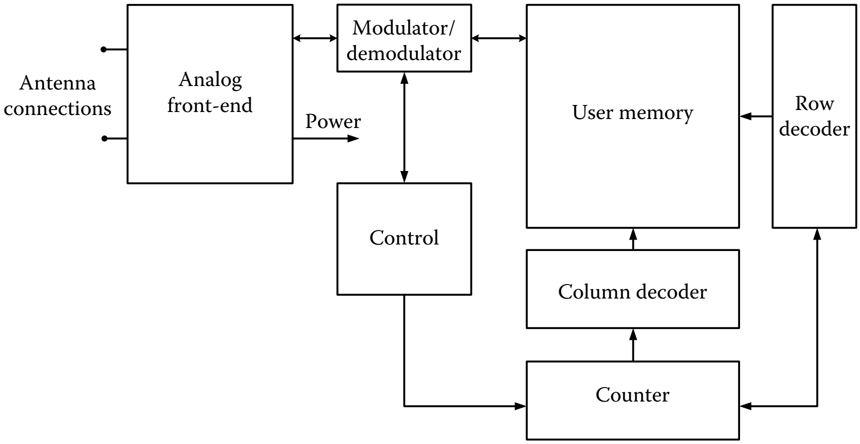

FIGURE 3.15
Structure of the circuitry for memory reading, writing, and control.

Because the MCRF250/251 device operates in the LF range, assuming a carrier frequency of 128 kHz, this setting will result in a transmission rate of 1 kHz. Conversely, the fastest transmission rate is MOD16, which will result in a transmission rate of 8 kHz. It is also important to keep in mind that MOD128 is the data rate used when programming the device. Finally, CB1 is used to specify the size of the data memory of 128 bits or 96 bits depending upon the needs of the user.

Example 3.4:

The configuration register for a MCRF 200 device is set to $08D. Describe the settings for the device.

Solution:

The binary values for the configuration register are shown as follows:

CB12	CB11	CB10	CB9	CB8	CB7	CB6	CB5	CB4	CB3	CB2	CB1
0	0	0	0	1	0	0	0	1	1	0	1

CB12 = 0. The device has not been programmed yet.
CB11 = 0. MCRF200 does not support anticollision; therefore, CB11 is set to 0.
CB10 = 0. Rate of PSK modulation is equal to half of the carrier frequency.
CB9 = 0 and CB8 = 1. The modulation chosen is PSK_1. This means that the phase of the modulation signal changes at the change of the encoded data.
CB7 = CB6 = 0. Data encoded using nonreturn to zero level (NRZ_L).
CB5 = 0. No special meaning because CB5 is always set to 0.
CB4 = 1; CB3 = 1; CB2 = 0. Baud rate is MOD32, resulting in a 4 kHz baud rate.
CB1 = 0. Memory size is 128 bits.

© 2011 by Taylor and Francis Group, LLC

Example 3.5:

State the configuration register for a transponder based on the MCRF250 device with the following parameters:

Blank (not programmed device)
Modulation: PSK_1.
PSK rate: Frequency half of the carrier frequency
Data encoding: NRZ_L
Baud rate: MOD 32
Memory size: 128 bits

Solution:

The MCRF250 is a device similar to the MCRF200, with the only difference being that it supports anticollision. Therefore, the configuration bit CB11 will be set to 1. The rest of the configuration bits have the same meaning as the MCRF200. The values of the configuration bits in this case are as follows:

CB12	CB11	CB10	CB9	CB8	CB7	CB6	CB5	CB4	CB3	CB2	CB1
0	1	0	0	1	0	0	0	1	1	0	1

Therefore, the hexadecimal value stored in the configuration register is $48D.

The devices MCRF355 and MCRF360, also manufactured by Microchip Technology, operate at the HF frequency of 13.56 MHz. They both have a total of 154 bits of memory that must be programmed using a contact programmer. This makes them act as read-only devices when they are used in the field but gives them the flexibility to be reprogrammed by the user when necessary. The data retention specified by the manufacturer is also better than 200 years.

The family of devices MCRF450/451/452/455 also operates in the HF frequency range at 13.56 MHz. As was discussed in Chapter 2, the different devices in this family use internal or external capacitors to modulate the radiofrequency signal. In discussing their internal memory structure, the concepts described for the device MCRF450 are also valid for the rest of the devices in this family. All of these devices have their memory organized in two areas: main memory and stored cyclic redundancy check (SCRC) memory. SCRC memory is used in the operation of the cyclic redundancy code necessary for the operation of the anticollision algorithm, as all of these devices support anticollision. The explanation of the CRC procedure is shown in Section 3.3.2.

The main memory of the MCRF450 and similar devices contains a total of 1024 bits and is structured in 32 blocks, each one with 32 bits. The SCRC memory is organized in 32 blocks, each one with 16 bits. Figure 3.17 shows the structure of the user memory in these devices.

The first three blocks (B0–B2) of the main memory are used to set up the operation of the device in a manner similar to a configuration register, while the remaining 29 blocks (B3–B31) are for user data. All the blocks of the main

© 2011 by Taylor and Francis Group, LLC

	BYTE 3								BYTE 2								BYTE 1								BYTE 0							
	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Block 0	FR	TF	TFT		DF		MT		TM			Available touser																				
Block 1									32-bit unique identification code for transponder																							
Block 2							Bits for write-protect mode. Each bit write-protects one block																									
Block 3									Data Fields for Fast Read Response (LS Block)																							
Block 4									Data Fields for Fast Read Response																							
Block 5									Data Fields for Fast Read Response (MS Block)																							
Block 6	DATA																															
Block 7	DATA																															
...	...																															
Block 31	DATA																															

FIGURE 3.17
Structure of the user memory for the MCRF 450 family of transponders.

memory with the exception of Block 1, which contains the ID or serial number of the transponder, can be individually rewritten block by block by a contactless programmer. Block 1 is programmed at the factory and is protected against being rewritten. The main memory is read or written in blocks of 32 bits, with the exception of bits 30 and 31 in Block 0 that can be selected individually. Blocks 3 to 5 constitute the fast read field (FRF) that is used by the interrogator in the transmission of commands to the transponder.

Block 0 is split into 21 bits of general memory available to the user and 11 bits containing operational parameters for the transponder. These operational bits, shown in Figure 3.17, are as follows:

FR (Bit 31). When FR = 0, the transponder will respond to the *fast read bypass* (FRB) command but will not respond to the *fast read request* (FRR) command from the interrogator. When FR = 1, the previous situation is reversed. These interrogator commands are further explained and described in Chapter 6.

TF (Bit 30). TF = 0 sets the transponder in *interrogator talks first* (ITF) mode. This means that it will wait for an FRR command. TF = 1 sets the transponder in *tag talks first* (TTF) mode if FR = 1. In this case, the transponder will send a *fast read response* without waiting for an FRR.

TFT (Bits 29–28). These two bits set the value for the parameter TCMAX when the transponder is in TTF mode. TCMAX is the number of

© 2011 by Taylor and Francis Group, LLC

TABLE 3.2

Configurations for TFT Bits

Bit 29	Bit 28	TCMAX
0	0	1
0	1	2
1	0	4
1	1	Continuous

TABLE 3.3

Configuration for DF Bits

Bit 27	Bit 26	TCMAX
0	0	32 Bits
0	1	48 Bits
1	0	64 Bits
1	1	96 Bits

fast read responses that the transponder can send after an FRR command, as shown in Table 3.2. For example, TCMAX = 2 means that the transponder can send its response twice for acknowledgment. If TCMAX is set to continuous mode, the transponder will send its response approximately every 80 ms until it receives the correct response from the interrogator. This is the default mode.

DF (Bits 27–26). These two bits set the length of the data transmitted by the transponder according to Table 3.3. The default data length is 32 bits.

MT (Bits 25–24). These two bits are hardwired and cannot be changed by the user. At the current state of production, these two bits are both set to 0. The remaining values are reserved for future applications.

TM (Bits 23–22–21). These three bits set up the total memory size. These three bits are also hardwired. Currently, there are only two possibilities allowed: 000 and 001. The code *000* describes a transponder with 512 bits of memory, while the code *001* describes a transponder with a memory size of 1 kbit.

Block 1 contains a unique 32-bit identification number for the transponder. This identification word is serialized by the manufacturer.

Block 2 contains the bits that set each block in write-protect mode. Each bit corresponds to its 32-bit block. For example, bit 5 will set Block 5 as writable or write-protected. When the write-protect bit is set to 1, the block is writable; when the bit is set to 0, the block is write-protected. It is important to note that once a block is write-protected, it cannot be changed back to being

© 2011 by Taylor and Francis Group, LLC

writable. The transponders are shipped to the user with all the write-protect bits to 1 with the exception of bit 1. This means that Block 1, which contains the unique ID for the transponder, cannot be changed by the user. Bits 31 and 30 in Block 0 (bits FR and TF) are not write-protectable, thus allowing the user to change how the transponder is operating. Finally, it is also important to note that Block 2 can also be set to be write-protected.

Blocks 3 to 5 contain data bits for the fast response. The size of the fast response was set by bits 27 and 26 (bits DF) in Block 0.

The SCRC memory section is organized into 32 blocks, each one with 16 bits as shown in Figure 3.18. The blocks contain the CRC code for the corresponding memory block sent back to the interrogator in order to check for anticollision.

One of the family of transponders operating at 13.56 MHz manufactured by Atmel® has an internal user memory ranging from 1 kbit to 64 kbits with the distinct feature that this memory is encrypted, thus providing an additional layer of protection. The *CryptoRF EEPROM* family offers user memory of 1 kbit, 2 kbits, 4 kbits, 8 kbits, 16 kbits, 32 kbits, and 64 kbits. In addition to the user memory, all these devices have a configuration memory of 2 kbits. This configuration memory is used to store eight sets of passwords for reading and writing, four crypto key sets, security access registers for each user zone and password, and key registers for each zone. The transponders in this family also support anticollision. Figure 3.19 shows the block diagram for these types of devices. As shown in the figure, while the analog front end is similar to that of any transponder, it differs in the existence of authentication and password verification prior to the device sending data.

Bit #	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Stored CRC Block 0																
Stored CRC Block 1																
...	...															
Stored CRC Block 31																

FIGURE 3.18
Structure of the stored CRC memory used in the MCRF 450 family of transponders.

© 2011 by Taylor and Francis Group, LLC

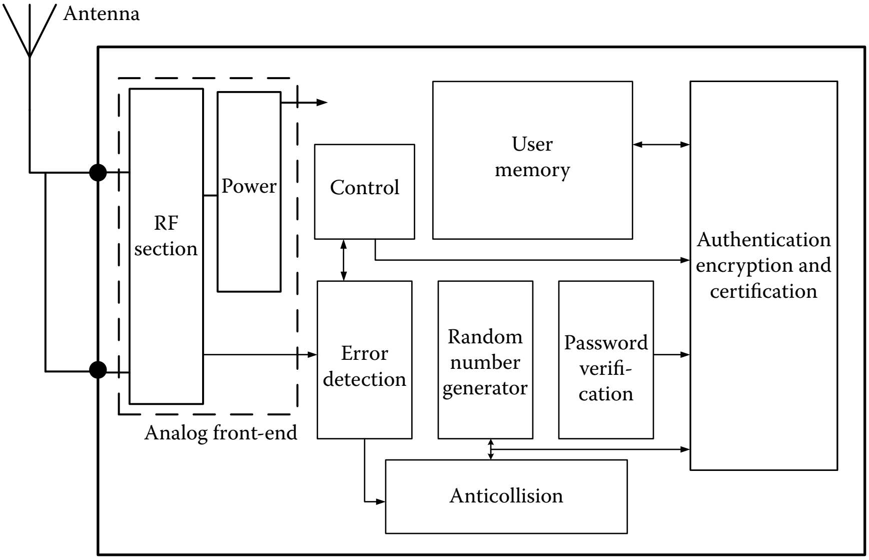
Antenna
Power
Control
User memory
RF section
Authentication encryption and certification
Error detection
Random number generator
Password verifi-cation
Analog front-end
Anticollision

FIGURE 3.19
Transponder with encrypted memory showing the additional components needed for this process.

In the transponders with user memory size of 64 kbits, 32 kbits, or 16 kbits, the EEPROM user memory is divided into 16 user zones. Figure 3.20 shows the memory map for the 64 kbit memory transponder. As seen in the figure, each user zone is configured for storing 512 bytes (4096 bits), giving a total user memory of 64 kbits. For the 32 kbit transponder, its memory is also divided into 16 user zones, each one of 256 bytes (2048 bits); for the 16 kbit memory, the size of each user zone is 128 bytes (1024 bits).

The transponder from the same family with a memory of 8 kbits has a similar structure, although the memory is divided into eight different user zones. Each zone can store 128 bytes (1024 bits). The memory for the transponders with 4 kbits, 2 kbits, and 1 kbit of user memory are all divided into four user zones. For the 4 kbit memory, each user zone stores 128 bytes (1024 bits); for the 2 kbit memory, each user zone stores 64 bytes (512 bits); and for the 1 kbit memory, each user zones stores 32 bytes (256 bits).

In all cases, the access to the user zones is only possible after having some security requirements are met. These are defined by the user during the setup of the device. The initial state for the transponders in this family is to have all the security features disabled. The user must set up the level of protection by configuring the security requirements for the transponder. These security requirements are stored in the configuration memory.

All the transponders in the CryptoRF family have a configuration memory of 2048 bits. This memory stores passwords, keys, codes, and security

© 2011 by Taylor and Francis Group, LLC

	Address	Bit 7	Bit 6	Bit 5	Bit 4	Bit 3	Bit 2	Bit 1	Bit 0
User Zone 0	0								
	1								
	...	Total of 512 bytes (4096 bits) in Zone 0							
	512								
User Zone 1	0								
	1								
	...	Total of 512 bytes (4096 bits) in Zone 1							
	512								
User Zones 2 to 14	...								
User Zone 15	0								
	1								
	...	Total of 512 bytes (4096 bits) in Zone 15							
	512								

FIGURE 3.20
Memory map divided into 16 user zones, each one containing 4096 bits for a total of 65,536 bits (64 kbits).

definitions for each one of the user zones. The access rights to the configuration memory are defined in the control logic and cannot be altered by the user. The transponder can be programmed so each zone uses a different set of passwords. This is especially useful in the case of different users accessing the transponder. In addition, the transponders include three fuses that must be blown during the setup process. The fuses lock some of the portions of the configuration memory.

Atmel® also manufactures families of transponders that operate in the LF range, as well as in the UHF range. For example, the transponder ATA5558 can operate between 100 kHz and 200 kHz, while the ATA5590 operates in the UHF range from 860 MHz to 960 MHz. Both devices contain an internal EEPROM divided into a user memory of 1024 bits and a system memory of 320 bits. The user memory is organized in pages of 128 bits, and each page in blocks of 32 bits. Each one of the blocks must be programmed separately. Pages are protected against overwriting by the use of a lock bit that, as shown in Figure 3.21, is the most significant bit of each block.

The system memory in the ATA5590 transponder contains a page with the identification information for the transponder (Tag_ID page), a page with system-level information available to the user, and two blocks with system-level information for the manufacturer of the transponder.

© 2011 by Taylor and Francis Group, LLC

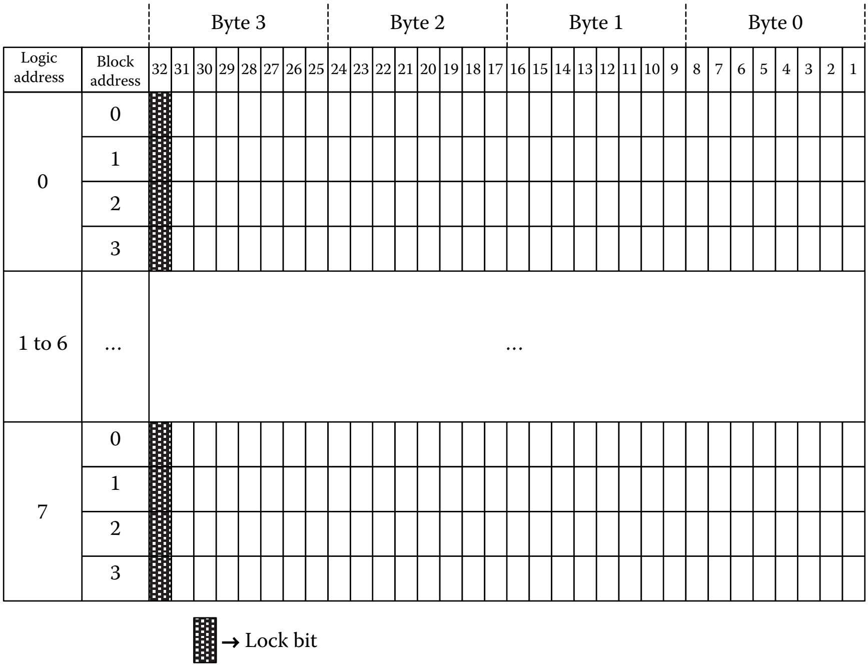

FIGURE 3.21
Memory map with lock bit. The lock bit is located at the MSB in each address.

3.3.2 Cyclic Redundancy Check for Error Detection

The cyclic redundancy check is the most used method for ensuring the integrity of the digital signals in some RFID transponders. In particular, CRC is used as a method of error detection in the digital stream transmitted and received by the transponders that support CRC. It is important to note that although the CRC algorithms detect the existence of errors, they do not make corrections.

The CRC method is based on appending additional information to the message being transmitted. This additional information is mathematically related to the message and therefore is redundant. At the receiving end, the receiver checks this additional information to verify that it agrees with the message being transmitted. This allows the receiver to determine with a certain degree of probability the existence of an error in the transmission. However, this method implies that not all errors are subject to detection. Simply stated, the CRC method is implemented by dividing the entire numeric binary value of the message by a constant. This constant is called a *generator polynomial*. The reminder in the division is then appended to the message. The polynomial that is used for the division is chosen from a family of polynomials with certain mathematic properties whose study is beyond the scope of this book. The transponders from Atmel and Microchip both use a polynomial known as the CRC-CCITT-16, although their specific

© 2011 by Taylor and Francis Group, LLC

implementation is somewhat different. This polynomial is also used in other digital transmissions such as CDMA and Bluetooth, among others. The CRC-CCITT-16 polynomial is

$$CRC - CCITT - 16: x^{16} + x^{12} + x^{5} + x^{0}$$

Reading the polynomial from most significant bit (MSB) to least significant bit (LSB) (normal representation) results in a value of 1021 hexadecimals. We must take into consideration that CRC polynomials have always had their MSB equal to 1, and for this reason it is not considered in the calculation of their values. Reading the polynomial from LSB to MSB (reverse representation) results in the value of 8048 hexadecimals. Microchip Technology uses reverse representation, while Atmel® uses normal representation.

CRC can be hardware-implemented using shift registers and exclusive-OR (XOR) gates in the feedback process shown in Figure 3.22. The encoder shown in Figure 3.22 consists of 16 shift registers and the exclusive-OR gates. The initial value loaded into the shift registers is $FFF in the Microchip Technology approach and $000 for the Atmel approach. The encoder performs the exclusive-OR function and shifts the registers until the last bit of the data stream is entered. At that point, the CRC value of the data set is equal to the values in the shift registers. When the transponder transmits data, the calculated CRC value is attached to the data.

The receiver performs an identical process and verifies that the CRC code generated by the receiver is equal to the CRC code appended to the message. The advantage of the CRC method is that the current state of one of the shift registers is a result of considerable past history, as it has to go through a lengthy process. Therefore, it is unlikely that a burst of errors will produce a CRC code equal to the one generated in the absence of errors. In fact, the CRC-CCIT-16 code is able to detect all error bursts of 16 bits or less as well as 99.9% of error bursts of more than 16 bits.

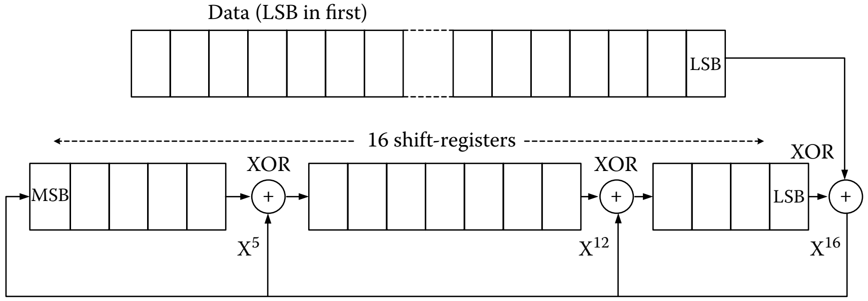

FIGURE 3.22
Structure of a generic CRC encoder using the CRC-CCITT-16 polynomial.

© 2011 by Taylor and Francis Group, LLC

Different manufacturers may also follow different approaches as to whether they append the CRC value with its LSB or MSB first. The specification documents for the transponders describe how the CRC value is attached to the data. The previous section described the structure of memory in some specific transponders. For example, the user memory for the MCRF450 family of transponders from Microchip Technology is divided into 32 blocks. When the interrogator tries to write data into any of these blocks, before doing any processing, the transponders check the CRC. If the CRC data are correct, the transponder proceeds with storing the data and the CRC in memory. Then, the transponder immediately transmits back the data and stored CRC to the interrogator for verification. If the CRC received by the transponder is incorrect, the transponder ignores the received message and waits for the next command with a valid CRC.

3.4 Transponder Programming

The programming of the transponders can be done by the manufacturer at the time of production or by the end user. *Programming the transponders* means to store a unique identification number in its memory, as well as to set up the configuration parameters for the communication between the transponder and the interrogator and additional parameters such as password protection. Transponders can be programmed using contact or noncontact methods.

The MCRF200/250 family of transponders manufactured by Microchip Technology is an example of a contactless programmable device. Because the MCRF200/250 family is made up of one-time programmable devices, the device can only be programmed when it is blank from the manufacturing process. Programming this device requires a specific sequence of radiofrequency signals applied following this sequence:

Initial, power-up RF signal (125 kHz for the MCRF200/250 family) lasting between 80 µs and 180 µs.

Absence of RF field lasting between 50 µs and 100 µs.

A continuous FSK signal that serves as a *verify* signal, lasting 131 ms. This signal is required for the transponder to be energized in order to output all the data contained in its 128 bits of memory. Because the device is blank at the time of programming, the output data are all *1*.

Programming data: After completing the verify mode, the device enters the programming mode. The device is programmed starting with bit 1 and finishing with bit 128. The symbol *1* is programmed by sending a low-power RF signal with an amplitude similar to the one used for the initial power-up. The symbol *0* is programmed by sending a high-power RF signal with an amplitude approximately equal to 2.2 times the amplitude used for the symbol *1*.

© 2011 by Taylor and Francis Group, LLC

Device response: After the 128 bits have been programmed, the device responds by transmitting the programmed data.

The MCRF355/360 family of transponders, also from Microchip Technology, has 154 bits of memory that can be programmed by using a contact programmer. Also, the memory in this family of devices can be reprogrammed. Figure 3.23 shows the pin diagram for this device.

Note that Pin 3 (Ant. A) and Pin 6 (Ant. B) are used to connect the external capacitors and inductors, as was shown in Figures 2.16 and 2.17, while Pin 5 (Vss) serves a ground. In order to program these devices, Pins 3, 6, and 5 are connected to ground, Pin 8 (Vdd) is connected to the positive voltage source, Pin 2 (CLK) is connected to the clock, and Pin 1 (Vprg) is connected to the signal that will be programmed in the transponder.

Some specific codes are used to carry out specific functions in the memory of the device. For example, because this device is reprogrammable, when the user wants to write new data into memory, the existing data must be erased first. This is done by sending the code 0111010100 through the VPRG pin, as shown in Figure 3.24.

Similarly, to start the programming sequence, the user must transmit the code 0111010010, followed by the data bits to store in memory. In total, there will be a total of 166 bits sent to the device: the first 12 bits are the command to program, and the remaining 154 bits are the data to be stored in memory. Finally, the data stored in memory can be read by transmitting the code 0111010110 to the device. After this code has been transmitted, the device will output the data stored in its memory.

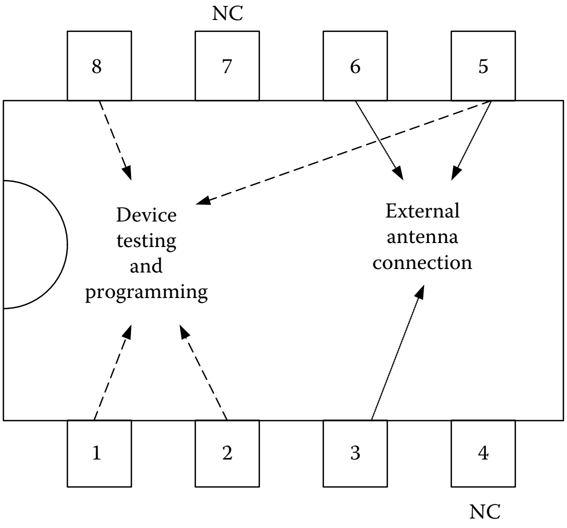

Pin	Function
1	VPRG
2	CLK
3	Antenna A
4	NC
5	V_{ss}
6	Antenna B
7	NC
8	V_{dd}

FIGURE 3.23
Pins for the MCRF355/360 transponders.

© 2011 by Taylor and Francis Group, LLC

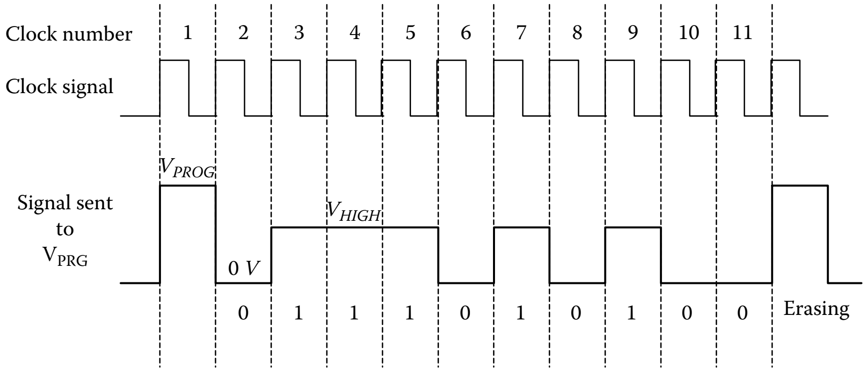

FIGURE 3.24
Code sent to erase existing data in memory using a contact programmer.

3.5 Summary of Electrical Specifications for RFID Transponders

The manner in which manufacturers specify the performance of their transponders varies greatly, not only among different manufacturers but also for devices from the same manufacturer. Tables 3.4 to 3.7 attempt to show the most important performance specifications from different manufacturers of RFID transponders. Table 3.4 does not pretend to be an exhaustive list of available devices, but just aims to show specific examples of device performance. The values of the parameters found in Tables 3.4 to 3.7 should be seen as somewhat typical performance values.

Table 3.4 shows the performance characteristics of three RFID transponders manufactured by Microchip Technology. The device MCRF200 is listed as a *125 kHz microID® Passive RFID Device*; the MCRF250 is exactly similar to the MCRF200 but incorporates anticollision; the MCRF355/360 family of devices is listed as *13.56 MHz Passive RFID Devices with Anti-Collision Feature*; and the devices in the MCRF450 family are listed as *13.56 MHz Read/Write Passive RFID Devices.*

Table 3.5 shows the electrical performance characteristics for four transponders manufactured by Atmel®. The ATA5567 device is listed as a *Multifunctional 330-Bit Read/Write RF Identification IC*; the ATA5570 is listed as a *Multifunctional 330-Bit Read/Write RF Sensor Identification IC*; the ATA 5558 is listed as a *1 kbit R/W IDID® with Deterministic Anticollision*; and the T5557 is listed as a *Multifunctional 330-Bit Read/Write RF Identification IC.*

Table 3.6 shows the performance parameters of two additional transponders from Atmel®. The device TK5530 is listed as a *read-only transponder*, while the TK5551 is listed as a *standards read/write ID transponder with anticollision.* The difference between these transponders and those evaluated earlier is

© 2011 by Taylor and Francis Group, LLC

TABLE 3.4

Summary of Electrical Parameters for Several RFID Transponders Manufactured by Microchip Technology

Parameter	MCRF 200/250	MCRF 350/360	MCRF 450/451/452/455
Operating frequency	100 kHz to 400 kHz	13.56 MHz	2 MHz to 35 MHz
Dynamic coil current	50 µA		
Operating current	5 µA	7 µA	20 µA
Maximum current through antenna pads	50 mA	40 mA	40 mA
Current leakage		10 nA	
Turn-on voltage	10 Vpp between antenna pads 2 V_{DC}		
Reading voltage		2.4 V	2.8 V
Testing voltage		4.0 V	
Programming voltage		V_{IH}: 2.8 V V_{IL}: 1.2 V	
Coil voltage while reading		4 Vpp	4 Vpp
Coil clamp voltage		32 Vpp	
Programming time	2 sec (for all 128 bit array)		5 ms (for a 32-bit block)
Sleep time		50 ms to 200 ms	
Modulation resistance		< 4 Ω	< 5 Ω
Internal capacitance	2 pF	100 pF	95 pF (MCRF451) 30 pF (MCRF 452) 50 pF (MCRF 455)
Data retention	200 years	200 years	200 years
Maximum storage temperature	-65°C to +150°C	-65°C to +150°C	-65°C to +150°C
Maximum ambient temperature	-40°C to +125°C		-40°C to +125°C
Maximum dissipated power			500 mW

[a] Blank spaces in the table indicate that the manufacturer did not provide information for that specific parameter.

that the ones shown below incorporate the coil for the antenna within the transponder. Moreover, the operating characteristics are given as strength of magnetic field instead of voltage or current.

Table 3.7 summarizes the most important characteristics of several transponders manufactured by Texas Instruments under their Tag-it™ commercial name. The manufactured has also chosen to specify their reading and writing thresholds as field intensity.

© 2011 by Taylor and Francis Group, LLC

TABLE 3.5

Summary of Electrical Parameters for Several RFID Transponders Manufactured by Atmel®

Parameter	ATA5567	ATA5570	ATA5558	T5557
Operating frequency	100 kHz – 150 kHz	100 kHz – 150 kHz	100 kHz – 250 kHz	100 kHz – 150 kHz
Supply current read mode	4 µA	4 µA	7 µA	4 µA
Supply current programming	40 µA		40 µA	40 µA
Maximum DC current	20 mA	20 mA	20 mA	20 mA
Maximum AC current	20 mA	20 mA	20 mA	20 mA
Coil voltage normal mode	3.6 V		4 V	4 V
Coil voltage Read / Write	> 6 V	> 10 V	> 6 V	> 6 V
Coil voltage Programming	> 8 V		> 6 V	> 8 v
Clamp voltage	17 V – 23 V	17 V – 23 V	7 V – 16 V	17 V – 23 V
Internal capacitance	78 pF	340 pF	78 pF	78 pF
Start-up time	< 1 msec	< 3 msec	< 3 msec	< 3 msec
Programming cycles	100,000 cycles	100,000 cycles	100,000 cycles	100,000 cycles
Data Retention	50 years	50 years	50 years	50 years
Maximum storage temperature	-40°C to +150°C	-40°C to +150°C	-40°C to +150°C	-40°C to +150°C
Maximum operating temperature	-40°C to +85°C	-25°C to +105°C	-40°C to +85°C	-40°C to +85°C
Maximum dissipated power	100 mW	100 mW	100 mW	100 mW

3.6 Mechanical Considerations and Transponder Assembly

The integrated circuits or chips in which the transponders are based are available in different packages. While commercial transponders use the chip-on-board (COB) technique described in Chapter 2, this package is not adequate for the development of prototypes. For this reason, most of the transponders are also available in the more common plastic dual-in-line packages (PDIPs) or small outline plastic packages (SOICs).

For example, some of the transponders described in this chapter, such as the MCRF200/250, the MCRF355/360, and the family of MCRF450 transponders, all manufactured by Microchip Technology, are available in these three packages (COB, SOIC, and PDIP). In their SOICs or PDIPs, these are available in eight pins; four of the pins are for testing purposes, two of the pins are for

© 2011 by Taylor and Francis Group, LLC

TABLE 3.6

Summary of Electrical Parameters for Several RFID Transponders Manufactured by Atmel®

Parameter	TK5530	TK5551
Resonance frequency	121.4 kHz – 129.2 kHz	120 kHz – 130 kHz
Inductance	3.95 mH	3.8 mH
Quality factor	13	13
Max field strength before modulation	2 A/m	4 A/m
Field strength operation	30 A/m (-40°C) – 17 A/m (+85°C)	30 A/m (-40°C) – 17 A/m (+85°C)
Field strength programming mode		50 A/m
Modulation Range	4.0 V (20 A/m) – 8.0 V (100 A/m)	4.0 V (20 A/m) – 8.0 V (100 A/m)
Programming time		16 ms per block
Data retention		10 years
Programming cycles		100,000 cycles
Absolute maximum field strength	1000 A/m	1000 A/m
Maximum storage temperature	-40°C to +125°C	-40°C to +125°C
Maximum operating temperature	-40°C to +85°C	-40°C to +85°C

connecting the transponder antenna, and the two remaining pins are not-connect pins. The devices ATA5570 and ATA5567/5558 from Atmel® are also available in COB packages and PDIPs, also using eight pins. The ATA5570 uses four not-connect pins, two pins for connecting the antenna, one pin for ground, and one pin for connecting the output of an external sensor. Because the ATA5567 and ATA5558 do not use an external sensor, their pinout is much simpler using six not-connect pins and two pins for connecting the external antenna. The dimensions of the COB version for all these transponders are 8 mm × 5 mm with a thickness of 0.4 mm. This extremely reduced size makes it ideal for incorporating the transponder and its antenna into the inlay. Transponders, especially in their COB form, are extremely sensitive to issues such as electrostatic discharge and their exposure to ultraviolet light that may result in erasing the contents in their memory cells. However, the use of X-rays for die inspection does not harm the device or erase their memory cell contents.

The process for assembling the whole transponder varies depending on the frequency range at which it will operate as well as the specifications for the final product. For transponders operating in the LF region, the assembly process starts by preparing the die and the capacitor and testing the functionality of the combined die-capacitor element. This is followed with preparing the antenna coil used in this frequency range and creating the inlay formed by the coil and the COB. After the inlay has successfully passed the test, the finished transponder is processed depending on the

© 2011 by Taylor and Francis Group, LLC

desired final product, such as card lamination and plastic molding. The last step in this process is a final functional test.

Transponders operating in the HF region can be assembled using three different methods. The first method, which results in the transponder with the COB, is based on the same assembly process as described earlier for LF transponders. A second method results in creating the transponder with direct wire bonding to the antenna that can be etched, printed, or stamped. This assembly method starts by preparing the die; preparing the antenna on the substrate etched, printed, or stamped; and then making the inlay by direct wire bonding to the antenna. After the inlay test, the finished transponder is created by using card lamination, plastic molding, or another method. The last step is a final functional test. The third method creates a transponder with a flip-chip process using a bumped die. The process starts by die bumping for flip-chip assembly, preparing the antenna on the substrate by etching, printing, or stamping it. The inlay is then created using the flip-chip process to attach the antenna to the die. This is also followed by the inlay test, and the finishing of the transponder by lamination or plastic molding. As always, the last step is a final functional test.

© 2011 by Taylor and Francis Group, LLC

4

Antennas for Interrogators

CONTENTS

4.1 Antennas for HF and LF Interrogators 86
  4.1.1 Design of Antennas for HF and LF Interrogators 86
  4.1.2 Antenna Matching 96
  4.1.3 Multiple Antenna Systems 103
4.2 Auxiliary Elements for LF and HF Interrogator Antennas 104
  4.2.1 Antenna-Tuning Indicators 104
  4.2.2 Antenna Multiplexers 104
  4.2.3 Preamplifiers 105
4.3 Antennas for UHF Interrogators 106
  4.3.1 Design of Antennas for UHF Interrogators 106
  4.3.2 Transmission Lines and Connectors for UHF Interrogators 111
4.4 Commercial Antennas for RFID Interrogators 114
  4.4.1 Antennas for RFID Interrogators 114
  4.4.2 Tuning Boards for Interrogator Antennas 116

This chapter describes the principles of design for interrogator antennas. The basic theory that describes the behavior of the antennas used in interrogators is the same that was used to describe the behavior of antennas used in transponders. However, in the case of antennas for interrogators, the size of the antenna is much less critical: because the interrogator has less size restrictions, the antenna can have larger dimensions. This results in an increased antenna performance. Furthermore, antennas for interrogators can be created with thicker conductors that can substation higher currents, resulting in stronger electromagnetic fields. The chapter starts by describing the basic principles used in the design of antennas for interrogators operating in the high-frequency (HF) and low-frequency (LF) bands. It continues by describing auxiliary elements such as tuning indicators, multiplexers, and preamplifiers used with those antennas in order to improve the overall performance of the system. The next section describes the basic principles used in the design of antennas in interrogators operating in the ultra-high-frequency (UHF) band. The last section in this chapter shows different examples of commercially available antennas for interrogators.

© 2011 by Taylor and Francis Group, LLC

4.1 Antennas for HF and LF Interrogators

4.1.1 Design of Antennas for HF and LF Interrogators

Antennas for interrogators operating in the HF and LF ranges are designed for parameters such as maximum read distance for transponders, transponder read rate, as well as regulations such as the maximum field allowed in some specific conditions. The majority of antennas used in HF and LF interrogators are variations of the coil antenna that generates a magnetic field. These antennas are typically called *portal antennas* because due to their physical dimensions and geometry, they can be placed in portals. A small number of antennas for HF and LF interrogators are based on a small dipole that generates an electric field. These antennas are normally called *stick antennas*.

In general, the distance at which a transponder can be read increases with the size of the antenna. However, this size cannot increase arbitrarily. This is due to the practical limitations in using an antenna of a large size and also to the limitations due to regulatory issues. Examples of these factors are the decrease in the signal-to-noise ratio as the size of the antenna increases; the potential for exceeding the regulatory limits of field strength; the need for shielding nearby electronic equipment to avoid unwanted interaction with the generated field; the increase in the area of null zones in which the generated field is zero; and the increase in the inductance of the antenna to a point that makes matching the antenna to the reader extremely difficult, if not impossible.

Interrogators working in the HF range expect an antenna tuned at the frequency of 13.56 MHz, with an input impedance of 50 Ω and a loaded quality factor (Q factor) of less than 20. Similarly, interrogators operating in the LF range expect an antenna tuned to the appropriate LF frequency, an input impedance of 50 Ω, and a loaded Q factor of less than 20. In both cases, for best performance, the match of the interrogator with the antenna should have a *voltage standing wave ratio* (VSWR) lower than 1.2.

The VSWR is a measure of how much of the energy sent through a transmission line is reflected back to the transmitter instead of being transferred to the antenna in order to be converted to the electromagnetic field. Standing waves occur as a result of the existence of incident and reflected waves. The ratio of reflected voltage to incident voltage is called the *reflection coefficient* and is denoted by the symbol Γ. Mathematically, the reflection coefficient is

$$\Gamma = \frac{E_r}{E_i} \tag{4.1}$$

where

Γ = the reflection coefficient

E_r = the reflected voltage

E_i = the incident voltage

© 2011 by Taylor and Francis Group, LLC

The reflection coefficient can also be expressed in terms of the impedance of the transmission line and the impedance of the load as

$$\Gamma = \frac{Z_L - Z_o}{Z_L + Z_o} \tag{4.2}$$

where Z_L = the impedance of the load and Z_o = the characteristic impedance of the transmission line.

Equation (4.2) shows that when the transmission line is terminated with a load with an impedance equal to its characteristic impedance, the reflection coefficient is zero. This means that all the energy present in the line has been transferred to the load. When the line is terminated with either an open circuit or a short circuit, the absolute value of the reflection coefficient is equal to 1. This indicates that none of the energy present in the line has been transferred to the load, but is being reflected back to the transmitter. For any other terminations, including complex impedance values, the value of the reflection coefficient is, in absolute value, between 0 and 1. This indicates that a portion of the energy present in the line has been transferred to the load and some other portion of the energy remains in the line.

The VSWR is defined as the ratio of the maximum voltage in the line ($E_{\max}$) compared to the minimum voltage in the line ($E_{\min}$):

$$VSWR = \frac{E_{\max}}{E_{\min}} \tag{4.3}$$

The VSWR can also be defined as a function of the reflection coefficient (Γ):

$$VSWR = \frac{1+|\Gamma|}{1-|\Gamma|} \tag{4.4}$$

Equation (4.4) shows that in the case of maximum reflection due to the transmission line terminated in an open or a short circuit, which results in $|\Gamma| = 1$, the value of the VSWR is infinity. Conversely, when the line is terminated with its characteristic impedance that results in $|\Gamma| = 0$, the value of VSWR is equal to zero.

Example 4.1:

The designer of a RFID system specifies a match between an interrogator and the antenna, resulting in a VSWR better than 1.2 using a 50 Ω coaxial cable as the transmission line. Calculate the maximum reflection coefficient and the allowed range of antenna impedance, assuming it is perfectly resistive.

© 2011 by Taylor and Francis Group, LLC

Solution:

The relationship between VSWR and the reflection coefficient is given by equation (4.4). Algebraic manipulation of this equation yields $|\Gamma| = 0.1$. This is the maximum value allowed for the reflection coefficient.

The relationship between the reflection coefficient and the impedance of the load is given by equation (4.2). When solving for the value of Z_L using this equation, it is necessary to consider the two signs associated with the previous result of $|\Gamma| = 0.1$ as the impedance of the load can exceed or be below the impedance of the line.

$\Gamma = 0.1$ results in $Z_L = 61\ \Omega$, while $\Gamma = -0.1$ results in $Z_L = 41\ \Omega$. Therefore, an antenna impedance ranging from 41 Ω to 61 Ω will result in a VSWR better than 1.2.

Note than the range of allowed load impedances is not symmetric referred to the impedance of the line (Z_o). There is more tolerance for values above Z_o than for values below Z_o.

A transmission line that is not perfectly matched (i.e., a VSWR value higher than 1) reduces the amount of power from the interrogator that reaches the antenna. More importantly, however, is the fact that a fraction of the power that remains in the transmission line can break down the dielectric of the line or increase its temperature due to the power dissipated. Therefore, it is critical to achieve the best possible matching.

Antenna analyzers are a very versatile and useful tool at the time of designing antennas and matching them to their transmission lines. Figure 4.1 shows an image of a low-cost antenna analyzer. The configuration of this analyzer is done through software, and the results are displayed and stored in a personal computer.

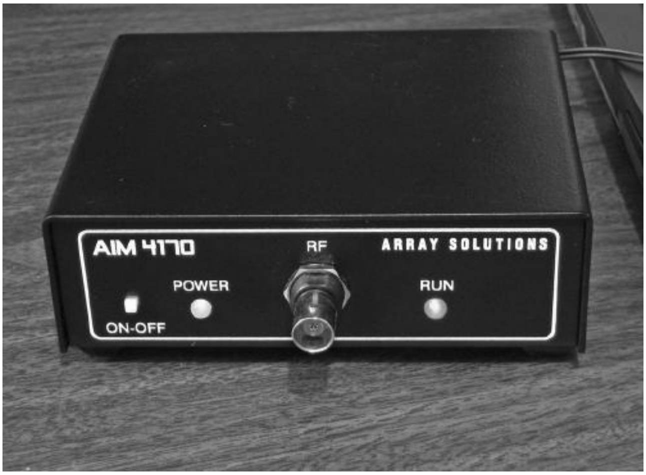

FIGURE 4.1
Antenna analyzer from Array Solutions. The control of the unit and the display of results are done by software.

© 2011 by Taylor and Francis Group, LLC

This unit shown in Figure 4.1 allows the measurement of the complex impedance of a transmission line or an antenna, the measurement of reflection coefficient, insertion losses, and other parameters of interest. The front panel only contains an ON/OFF switch, two light-emitting diodes (LEDs) that indicate whether the unit is energized and whether the unit is running a test, and a single Bayonet Neill-Concelman (BNC) connector. The back panel contains the power supply connection and a DB9 connector for a serial cable from the control computer to configure the instrument and transmit the results. The user can select the range of frequencies for the sweep-starting frequency, ending frequency, and step-or central frequency and bandwidth. This unit can also act as a fixed radio frequency source by generating a user-selected single frequency signal. The output values are shown graphically on the screen as conventional graphs (parameter versus frequency) or Smith charts. The data can also be stored as a text file for further processing. The graph can display one or more of the following parameters: impedance magnitude, impedance phase, reflection coefficient, return loss, stationary wave ratio, series load circuit, or parallel load circuit. Figure 4.2 shows an example of the output showing the magnitude and phase of a transmission line between 25 MHz and 40 MHz.

Figure 4.2 depicts the magnitude and phase of the impedance measured by the antenna analyzer for a transmission line with an electrical length

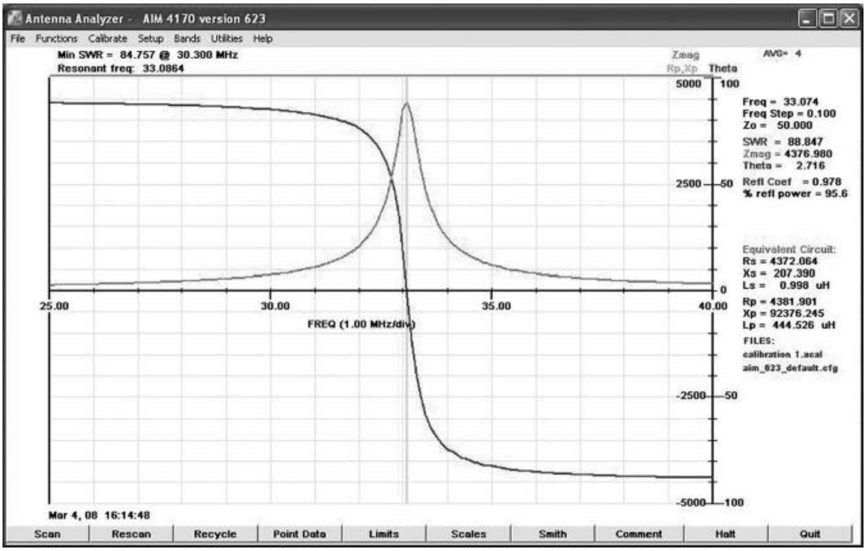
Antenna Analyzer - AIM 4170 version 623
File
Functions
Calibrate
Setup
Bands
Utilities
Help
Zmag
Rp,Xp
Theta
AVG= 4
5000
100
2500
50
0
-2500
-50
-5000
-100
25.00
30.00
35.00
40.00
FREQ (1.00 MHz/div)
Scan
Rescan
Recycle
Point Data
Limits
Scales
Smith
Comment
Halt
Quit

```
Min SWR =  84.757 @  30.300 MHz
Resonant freq:  33.0864

Freq =  33.074
Freq Step = 0.100
Zo =   50.000

SWR  =  88.847
Zmag = 4376.980
Theta =    2.716

Refl Coef   = 0.978
% refl power = 95.6

Equivalent Circuit:
Rs = 4372.064
Xs =  207.390
Ls =    0.998  uH

Rp = 4381.901
Xp = 92376.245
Lp =  444.526  uH

FILES:
calibration 1.acal
aim_623_default.cfg

Mar 4, 08  16:14:48
```

FIGURE 4.2
Frequency behavior of a transmission line terminated with a short circuit for frequencies around a quarter wavelength. This display shows the magnitude and the phase of the impedance at the near end.

© 2011 by Taylor and Francis Group, LLC

equal to one-quarter of the wavelength, terminated in a short circuit. Due to the impedance transformations at this specific electrical length, the theoretical value of the transformed impedance seen at the near end of the line is infinity. In practice, the losses of the line make the impedance have a large, yet finite, value. The output also displays other values described earlier in this section, such as the reflection coefficient and the VSWR, among others. Similar results could also be obtained using other commercially available antenna analyzers.

When designing the antenna, it is also necessary to take into consideration the write range of transponders that is about 70% of their read range. This is due to the higher amount of power needed to rewrite a transponder compared with just reading it. The orientation of the antenna with respect to the transponder is also a factor in determining the maximum read range in a given application. As shown in Figure 4.3, the field generated by the coil antenna does not have the same strength in all the spatial directions, but instead it has regions of maximum values alternating with regions of minimal strength.

An additional parameter that may limit the maximum read range of a system is the presence of metal objects in the vicinity of the antenna. Metal objects close to the antenna have the effect of detuning it, and therefore it loses its efficiency at the desired frequency. As the size of the antenna increases, the minimum required separation distance between the antenna and the metal objects also increases. For antennas operating in the HF frequency range, the minimum distance between antenna and metal has to be greater than 10 cm.

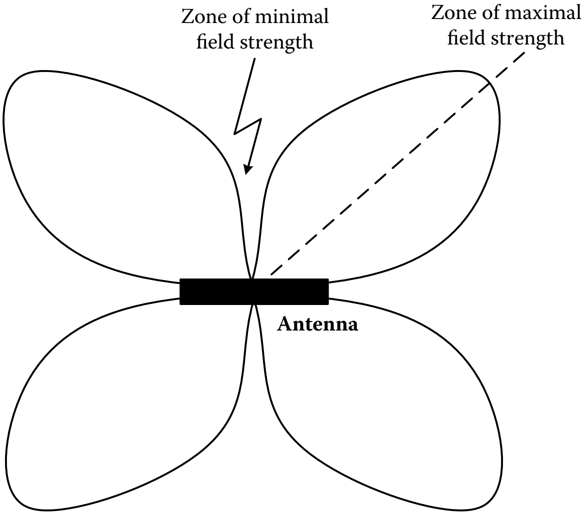
Zone of minimal field strength
Zone of maximal field strength
Antenna

FIGURE 4.3
The radiation diagram for the loop antenna in the interrogator shows areas of maximum and minimum field strength.

© 2011 by Taylor and Francis Group, LLC

Even at 30 cm of separation, there is a considerable decrease in the reading range. When the distance between antenna and metal object reaches 50 cm, there is no measurable difference. Similarly, the presence of other antennas in the vicinity also affects the overall performance of the system.

Antennas can be built using any conductive material. However, they are normally built using copper or aluminum. The type of material used for the antenna and the dimensions of the tube or strip lines depend on the desired mechanical parameters of the antenna as well as the maximum inductance that must be matched to the interrogator. In addition to the conductors, the matching process requires the use of capacitors and resistors. These must be rated for the high voltages that may arise in resonant antennas as well as for the power that they must dissipate. Typical power values for these resistors range from 2 W to 50 W.

The dimensions of the antenna have an effect on the read distance between interrogators and transponders. The strength of the magnetic field for an electrically small-loop antenna is

$$B_z = \frac{\mu_0 I N a^2}{2\,(a^2 + r^2)^{3/2}} \quad (Weber / m^2) \tag{4.5}$$

with a being the radius of the loop and r being the radial distance from the center of the loop. To study the effects of the antenna radius, the rest of the parameters in equation (4.5) can be considered constant. Assuming that all antennas have the same amount of current and also the same number of turns, the only parameter that changes is the radius of the loop. Figure 4.4 depicts the values of magnetic fields for coils with three different radii (0.2 m, 0.4 m, and 0.8 m). Because the graph is only intended to highlight the dependence of the field with the distance from the center of the coil, it uses arbitrary units for the strength of the field depicted in the vertical axis.

It is interesting to note that, although larger coils produce stronger fields, this only happens after a certain distance threshold has been reached. In particular, for read distances closer to the antenna, the stronger fields are produced by the coils with the smallest radius. Antennas with smaller radii produce stronger fields in the vicinity of the antenna, but these fields decrease faster for longer distances. Therefore, the design of the antenna must consider the range of distances between transponders and the interrogator when selecting the size of the loop.

The resonant frequency of an antenna (ω_o or f_o) is

$$\begin{gathered} \omega_o = \sqrt{\frac{1}{LC}} \quad or \\ f_o = \frac{1}{2\pi\sqrt{LC}} \end{gathered} \tag{4.6}$$

© 2011 by Taylor and Francis Group, LLC

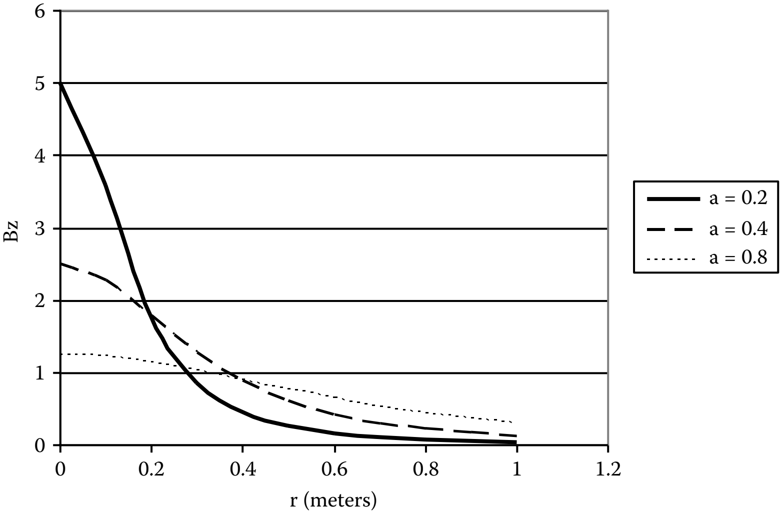

FIGURE 4.4
Strength of the magnetic field as a function of the distance from the coil for loop antennas of different diameters.

An antenna will reach its resonant frequency when its inductive impedance is equal to its capacitive impedance. The loop antenna has an inherently inductive behavior. Therefore, it needs an external capacitor to provide the capacitive impedance required for resonance. The value of this external capacitor can be calculated using equation (4.6). It is important to note that the capacitance and inductance of the antenna are inversely related. A problem may arise if the inductance of the antenna increases to a point that may make it difficult to use a capacitor of suitable value. For example, a loop antenna with an inductance of 5 µH operating in the HF range requires a matching capacitor of 27 pF. For these small capacitance values, the parasitic capacitances that have not been taken into account when modeling the antenna may contribute significantly to the overall capacitance and therefore modify its resonant frequency. It is then important for antennas to be designed with lower inductive values. This may require one to use, for example, low-resistance conductive tube instead of wire at the expense of increasing its physical dimensions.

The inductance of the antenna made with copper tube, shown in Figure 4.5, can be estimated as

$$L(\mu H) = 0.008\ a\left[\ln\left(0.707\,\frac{a}{d}\right) + 0.379\right] \tag{4.7}$$

where

L = the estimated value of inductance

d = the diameter of the tube in cm

© 2011 by Taylor and Francis Group, LLC

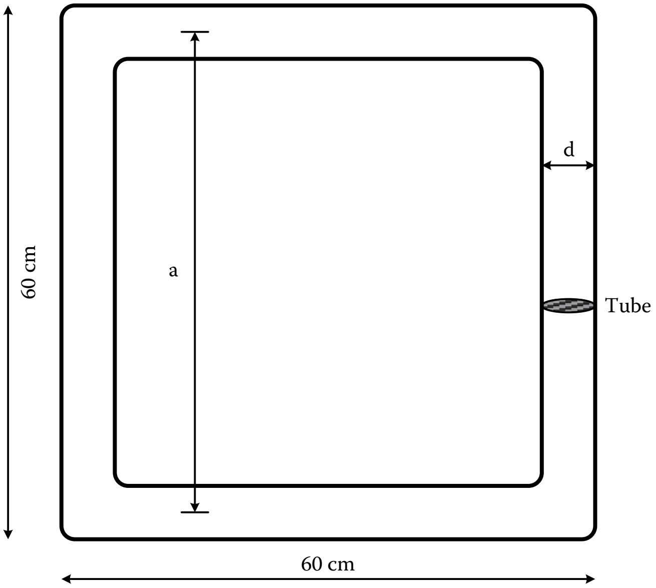

FIGURE 4.5
Basic structure of a loop antenna made with copper tube.

a = the length of the side of the antenna from center to center (as shown in Figure 4.5)

Example 4.2:

An interrogator operating at HF uses a loop antenna made of copper tube with a diameter of 2 cm, with a rectangular shape of 60 cm × 60 cm. Estimate the value of the inductance for the antenna. Estimate also the value of the capacitor required to tune the antenna for resonance.

Solution:

Using equation (4.7) with d = 2 and a = 59 yields L = 1.6 μH. This value of inductance can now be used in equation (4.6) to calculate the estimated value of capacitance. The estimated value of capacitance is 86 pF.

The previous equations provide an estimated value of antenna inductance. After the antenna has been built, its inductance must be measured in order to choose the correct value for the tuning capacitor. This can be done with an impedance analyzer that allows the selection of the measurement frequency. However, impedance analyzers are expensive and may not be available in most cases. In this case, a low-cost LCR (inductance [L], capacitance [C], and resistance [R]) meter that operates at the frequency of 1 kHz can give a reasonable approximation of the antenna inductance. The antenna analyzer shown in Figure 4.1 is another low-cost alternative for measuring

© 2011 by Taylor and Francis Group, LLC

the inductance of the antenna and the resonant frequency of the antenna-capacitor system.

The last parameter that needs to be calculated and measured when designing the antenna is its Q factor. The Q factor is a measure the frequency selectivity of an antenna. The Q factor is related to the range of frequencies above and below its resonant frequency that the antenna will be able to transmit or receive with losses below 3 dB. The Q factor can be estimated as

$$Q = \frac{f_r}{BW_{-3dB}} \tag{4.8}$$

where Q is the quality factor of the antenna, f_r is its resonant frequency, and $BW_{-3\mathrm{dB}}$ is the bandwidth of the antenna at 3dB of attenuation.

A high Q factor indicates an antenna that is very narrow in frequency, that is, highly selective. This may be good in order to reject interfering signals that otherwise would be detected by the antenna. However, the RFID signal has a certain bandwidth. If the antenna is too selective, it may distort the components of the transmitted signal that fall outside the selectivity of the antenna, as shown in Figure 4.6. A good experimental compromise is to choose a Q factor of around 20.

If the Q factor of the system made of the antenna and capacitor is higher than 20, it needs to be brought down by using a dampening resistor in parallel with the antenna in order to widen its bandwidth. This technique is shown in Figure 4.7.

The introduction of the dampening resistor produces Ohmic losses in the system, lowering the Q factor. The Q factor is related to the parallel resistance as

$$Q = \frac{R_{par}}{2\pi f L} \tag{4.9}$$

An additional problem arises when using equation (4.9) in order to select the value of the parallel dampening resistor: the value R_{par} in the equation is the combination of the external dampening resistor (R_{damp}) in parallel with the resistance of the antenna at the frequency of interest (R_{ant}). The measurement of the resistance of the antenna at the operating frequency is difficult and requires the use of expensive instrumentation. However, the Q factor for an antenna is relatively easy to measure experimentally with an antenna analyzer or using a radiofrequency signal generator and a spectrum analyzer. Therefore, the resistance of the antenna R_{ant} at the operating frequency can be calculated indirectly by measuring the Q factor of the antenna without an external dampening resistor. Once this value has been found, it is now possible to calculate the required value of the total parallel resistance R_{par} to achieve the desired value of Q. Because R_{par} is the parallel of R_{ant} and the

© 2011 by Taylor and Francis Group, LLC

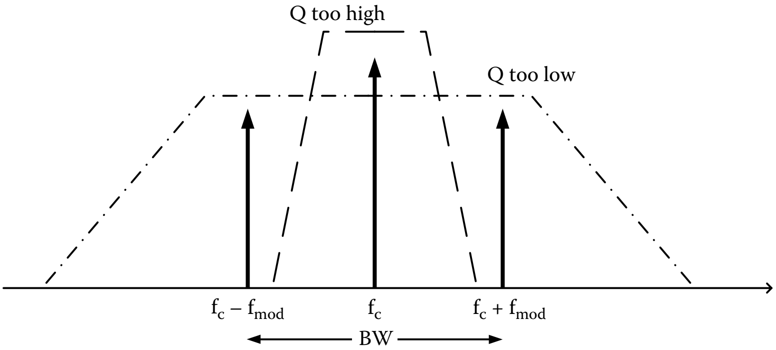

FIGURE 4.6
Q factor of an antenna and its relationship with the antenna performance. (a) A very low quality factor (Q factor) does not provide adequate selectivity. (b) A very high Q factor may reject components of the transmitted signal.

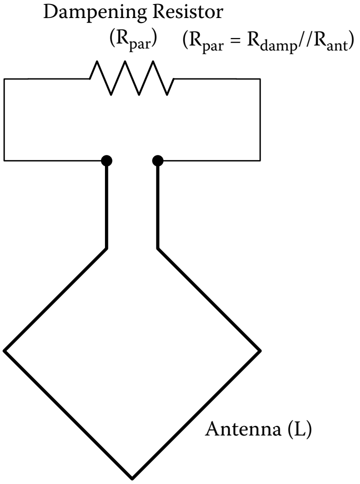

FIGURE 4.7
Dampening resistor connected in parallel with the antenna. This is used to lower the Q factor to the desired value.

external dampening resistor R_{damp}, it is now possible to estimate the value of this external resistor. Example 4.3 illustrates this procedure.

Example 4.3:

Consider the antenna for the RFID system from Example 4.2. Estimate the value of the external dampening resistor necessary for the system to achieve a Q factor of 20. The initial measurement of the Q factor for the antenna-capacitor system resulted in a value of 55.

© 2011 by Taylor and Francis Group, LLC

Solution:

Using equation (4.9) with a Q factor of 55, an inductance value for the antenna of 1.6 μH, and a frequency of 13.56 MHz results in a resistance equal to 7.5 kΩ. Because the Q factor of 55 was measured without the use of external resistors, the value of resistance obtained is the resistance of the antenna: $R_{ant} = 7.5$ kΩ.

Now, we must calculate the value of the total parallel resistance needed to reduce the Q factor to the desired value of Q. We use again equation (4.9) with the same operational parameters as before, with the exception that the new value for Q results in a resistance equal to $R_{par} = 2.7$ kΩ.

Considering that the total value of the resistance found above is equal to the resistance of the antenna in parallel with the dampening resistor ($R_{par} = R_{ant} // R_{damp}$), it is now possible to calculate the value of the external dampening resistor. This results in $R_{damp} = 4.2$ kΩ. Because this is not a standard resistor value, the designer should use the next standard value, which is 4.3 kΩ.

Therefore, the addition of 4.3 kΩ results in lowering the Q factor from its initial value of 55 to the desired value of 20. Because the value of inductance used for this calculation could be an estimated value, the design process involves measuring the Q factor with the external dampening resistor of 4.3 kΩ and adjusting its value if necessary to reach a Q factor of 20. Finally, it is also necessary to keep in mind that this external dampening resistor must be sized to dissipate the power transmitted by the interrogator.

4.1.2 Antenna Matching

Matching the impedance of the antenna to the impedance of the transmission line (normally a coaxial cable of 50 Ω) ensures that all the power from the interrogator is transferred to the antenna, and there is no power reflected in the line. In practice, a VSWR of 1.2 or lower is acceptable as the majority of the power is transmitted to the antenna. From the several approaches to match an antenna to its transmission line, the most widely used matching methods for interrogator antennas in RFID systems are gamma matching, T-matching, transformer matching, and capacitance matching.

Gamma matching is a technique based on connecting the two conductors of the coaxial cable that make up the transmission line to the two points in the antenna that exhibit a resistive impedance equal to 50 Ω. The shield of the coaxial cable is normally connected to the point in the antenna that is opposite to the gap used to connect the resonance capacitor and the dampening resistor. The central connector of the coaxial cable is then connected, by means of an additional conductor, to the appropriate point in the antenna as shown in Figure 4.8.

In practice, the point at which the central conductor of the coaxial cable is connected (point A) must be determined by experimentation. The designer of the system connects an antenna analyzer at the far end of the coaxial cable and moves point A until the antenna analyzer reads a VSWR value equal to 1. This indicates that the impedance of the antenna between point A and the connection of the central conductor is equal to 50 Ω. The value of the capacitor used to tune the antenna to the resonance frequency was

© 2011 by Taylor and Francis Group, LLC

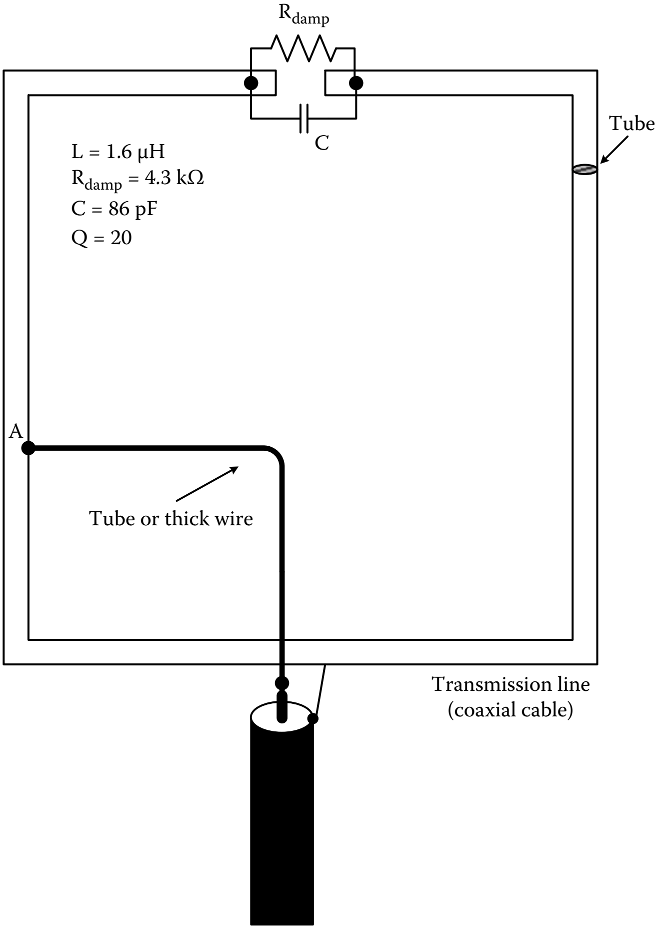

FIGURE 4.8
Connections for gamma matching.

determined using an approximation for the value of the antenna inductance. Therefore, a variable capacitor is initially used in order to tune the antenna to its resonant frequency. Once the correct value of capacitance has been determined, it is then possible to substitute the bulkier variable capacitor with a fixed capacitor.

T-matching is similar to gamma matching with the difference being that in T-matching, both the internal and external conductors of the coaxial cable are tapped to two points in the antenna, as shown in Figure 4.9. It is based on the fact that two points of the antenna, equidistant from the center, exhibit a purely resistive impedance.

The two tapping points are located symmetrically with respect to the gap used for connecting the tuning capacitor and the dampening resistor. The

© 2011 by Taylor and Francis Group, LLC

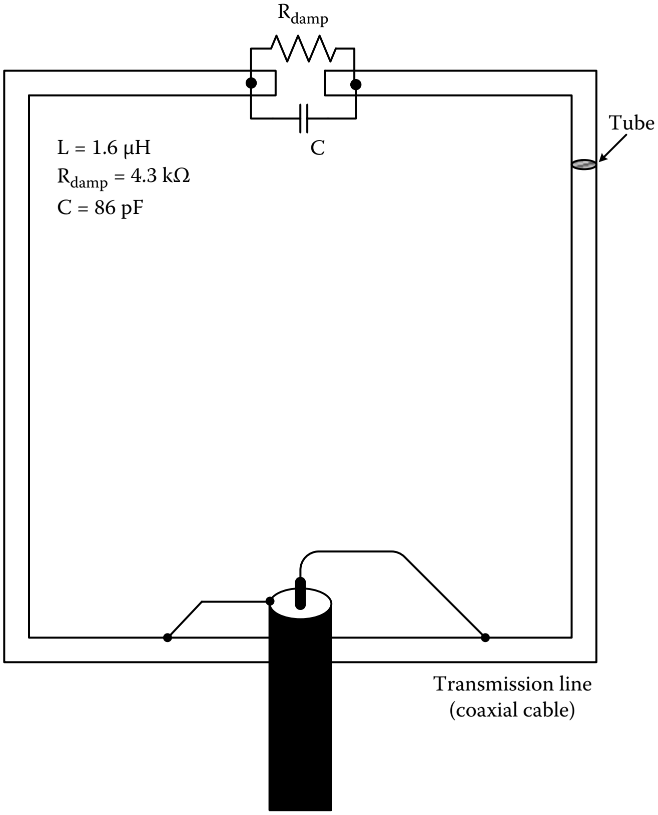

FIGURE 4.9
T-matching.

two tapping points are found by using an antenna analyzer until the VSWR value is equal to 1. Higher values of the Q factor results in the two matching points being closer together.

Transformer matching has the advantage that it electrically isolates the antenna from the interrogator, thus breaking the Ohmic continuity between these two parts of the system. This is useful when the RFID system is affected by noise as it breaks any ground loop that may exist. Figure 4.10 shows the electric diagram of transformer matching. It is important to note that the matching occurs at the gap in the antenna that is used to connect the resonance capacitor and the damping resistor instead of their opposite point, as it happens with gamma and T-matching.

Transformer matching requires the design of two elements: the matching transformer and a balum (balanced–unbalanced) transformer. The design of the matching transformer requires one to size the number of turns of the primary (*N*) that is connected to the transmission line and the secondary (*M*)

© 2011 by Taylor and Francis Group, LLC

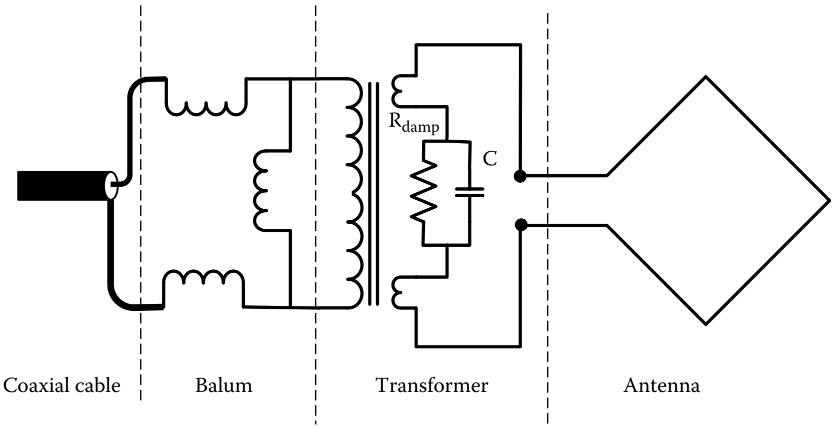

FIGURE 4.10
Transformer matching. Notice how the matching point is the same point at which the capacitor and resistor are connected.

that is connected to the antenna. The ratio of turns can be found using the following:

$$\left(\frac{N}{M}\right)^2 = \frac{Z_o R_{par}}{(2\pi f L)^2} \tag{4.10}$$

where (N/M) is the primary to secondary ratio of turns, Z_o is the impedance of the transmission line, R_{par} is the parallel resistance of the antenna and the dampening resistor, and L is the inductance of the antenna.

Example 4.4:

Using the values from the previous examples in this chapter, calculate the number of turns for the primary and secondary of the matching transformer. Assume a characteristic impedance of 50 Ω.

Solution:

Using equation (4.10) with R_{par} = 2.7 kΩ, L = 1.6 μH, and f = 13.56 MHz, the ratio of turns is equal to 2.7. Using, for example, M = 3 results in N = 8 windings.

The balum transformer is used to connect unbalanced transmission lines such as a coaxial cable to balanced antennas such as the loop antenna. Although they can be connected directly, as with the gamma matching, the common mode currents that appear in the antenna and line due to the unbalance may cause excessive noise and disturb the circuits. In this case, the connection of a balum transformer between line and transformer

© 2011 by Taylor and Francis Group, LLC

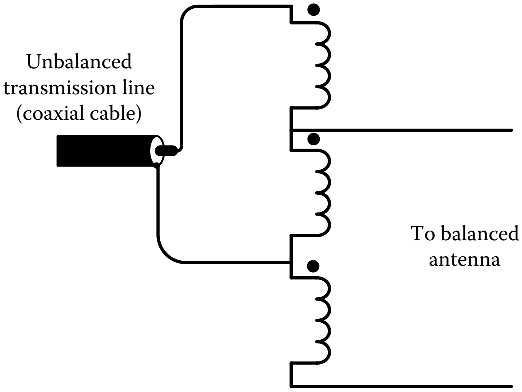

FIGURE 4.11
A balum is used to connect unbalanced transmission lines to balanced antennas.

solves these problems. The structure of a balum transformer is shown in Figure 4.11.

Both the matching transformer and the balum transformer are normally made with a toroid ferrite core. The balum transformer can be tested by connecting a 50 Ω load to the balanced side and measuring the VSWR through a coaxial cable connected to the unbalanced side. The VSWR should be close to 1.0.

Capacitance matching attempts to match the antenna and its transmission line by adding to external capacitors between the resonance capacitor and the dampening resistor, as shown in Figure 4.12.

Because of the stray capacitances inherent to the system, it is very difficult to predict the adequate values for the two matching capacitors, which requires extensive experimentation. Furthermore, for larger antennas, the values of the two matching capacitors tend to be very small, making the matching process very sensitive to stray capacitances. In Figure 4.12, the series combination of C_1 and C_2 must be equal to the capacitance value required to tune the antenna at the correct frequency.

The value of the matching capacitor C_{m1} can be estimated as

$$C_2 = C_{tuning}\sqrt{\frac{R_{par}}{R_{out}}} \tag{4.11}$$

where C is the total capacitance required for tuning the antenna, R_{par} is the equivalent value of the AC resistance of the antenna in parallel with the damping resistor, and R_{out} is the desired matching impedance, typically 50 Ω. The value of C_1 can be calculated through the series capacitance of C_1 and C_2.

© 2011 by Taylor and Francis Group, LLC

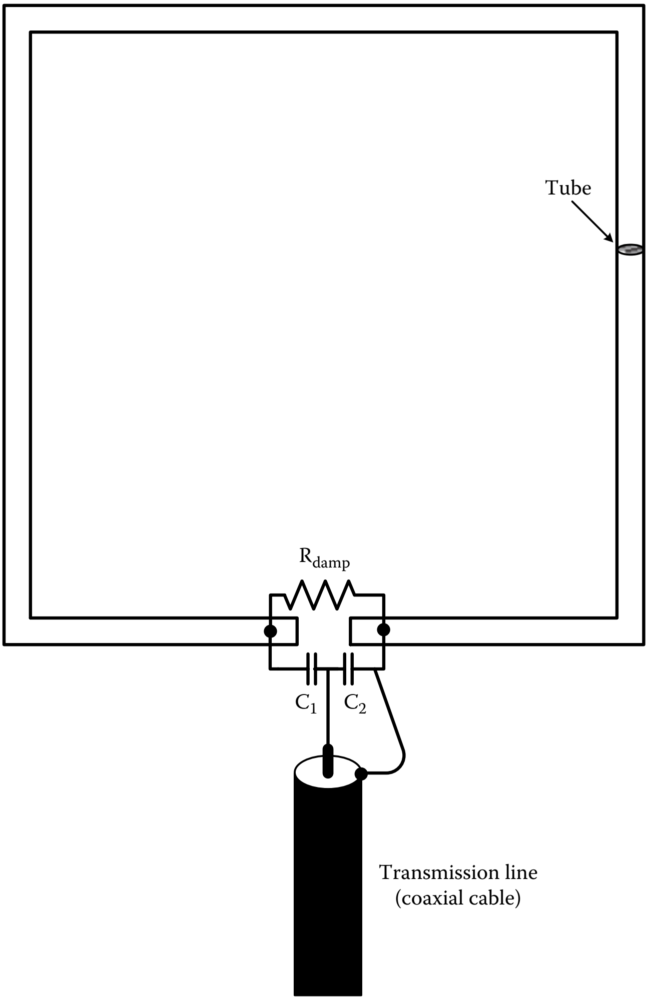

FIGURE 4.12
Capacitance matching.

Example 4.5:

The antenna used in Example 4.4 must be matched to a 50 Ω transmission line using capacitance matching. Calculate the values of the two matching capacitors required.

Solution:

The required tuning capacitor has a value of 86 pF, and the parallel resistance has a value of 2.7 kΩ. The required value for R_{out} is 50 Ω. Using equation (4.11) yields a value of C_2 = 632 pF. The equivalent series capacitance is 86 pF, allowing one to calculate the value of C_1 equal to 99 pF.

© 2011 by Taylor and Francis Group, LLC

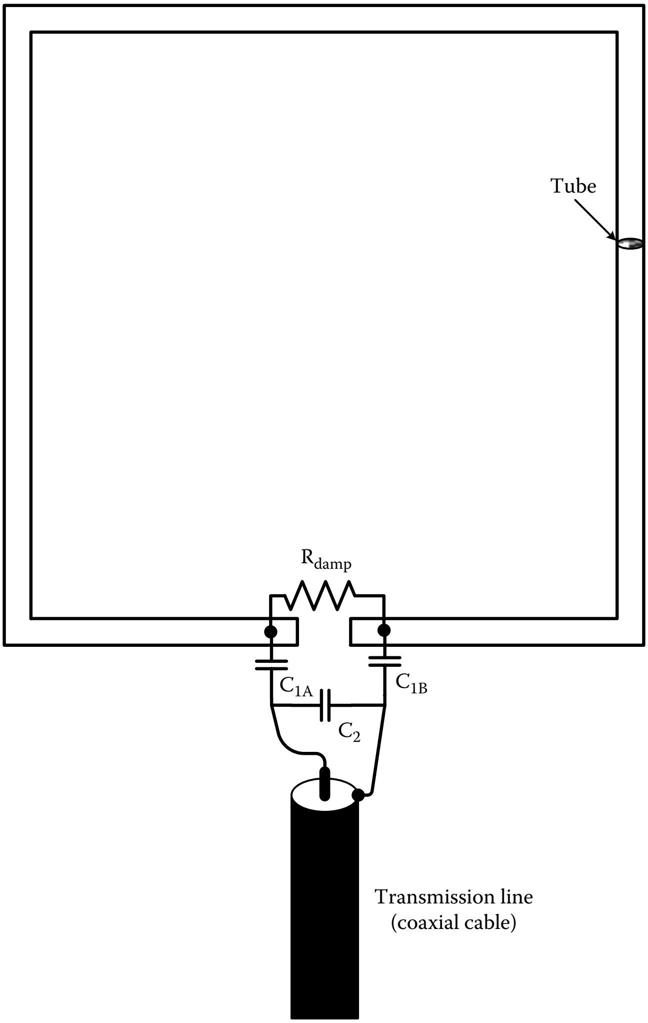

FIGURE 4.13
Balanced capacitance matching.

Some antennas require a matched capacitance so the capacitance seen by both conductors of the coaxial cable is the same. This can be achieved by using the structure shown in Figure 4.13. In this structure, the value of C_1 is doubled and added to each side of C_2.

Example 4.6:

Using the results of Example 4.5, calculate the capacitor values required for balanced matched capacitance.

© 2011 by Taylor and Francis Group, LLC

Solution:

From Example 4.5: $C_1 = 99$ pF and $C_2 = 632$ pF. To achieve balanced capacitance, $C_{1A} = C_{1B} = 198$ pF. C_2 remains at the same value of 632 pF.

4.1.3 Multiple Antenna Systems

It is not uncommon for several antennas to be located in close proximity. This may be due to different interrogators being close to each other or for multiple antennas to be connected to a single interrogator. When designed correctly, the mutual coupling between the antennas can enhance the performance of the overall system. However, a careless positioning of the multiple antennas may result in their mutual coupling, severely degrading their performance.

When several single-antenna interrogators coexist in close vicinity to each other, it is necessary to minimize their interaction with each other. The coupling between these antennas depends on the distance between them as well as their relative angles. Although the radiation pattern of a single antenna is well known and easy to predict, this radiation pattern becomes distorted due to the coupling. The best approach for locating the optimal placement of two or more antennas is by using an antenna analyzer in order to find the minimum of radiation by measuring the voltage generated in the nearby antenna.

A technique sometimes used to enhance the radiation pattern of the antenna in a single interrogator is to place another matched but unconnected antenna, opposite to the antenna connected to the interrogator. This connected antenna is a *driven antenna*, while the unconnected antenna is a *reflective antenna*. The combination of both antennas directs the energy from the driven antenna to a narrower area. Once again, an antenna analyzer should be used in order to determine the best arrangement between the two antennas.

Instead of leaving the second antenna unconnected, making it a reflective antenna, it is also possible to drive it with the signal from the interrogator by means of a splitter. In this case, the output of the interrogator is connected to the input of the splitter, and each output from the splitter is connected to a matched antenna. Splitters also offer the possibility to have the two output signals in-phase or out-of-phase. Splitters introduce losses, but when the two signals are connected in-phase, the reading distance between interrogator and transponders increases more than the reading distance with just one antenna without the losses from the splitter. The increase in the reading distance is at the cost of reducing the radiation pattern in the directions perpendicular to the arrangement of antennas. Therefore, this technique is only useful when the transponders are located only in known areas. When the signals fed to the two antennas are out-of-phase, the radiation pattern changes significantly, resulting in a pattern opposed to when the two signals are in phase. This gives the designer of the system the ability to configure the radiation pattern depending on the needs of the system.

Other techniques based on systems with multiple antennas are directed toward specific applications. In one of them, two identical antennas located

© 2011 by Taylor and Francis Group, LLC

in a crossed arrangement are connected to the output of a splitter but use different lengths of transmission line. In particular, if the different between lengths is a quarter of a wavelength, the two fields are shifted in 90°. This produces a field that changes the direction of its maximum, giving the appearance of a rotating field. Some other commercial antennas have an additional input that allows connecting an out-of-phase signal from a complementary antenna in order to maximize field strength in a single direction.

4.2 Auxiliary Elements for LF and HF Interrogator Antennas

4.2.1 Antenna-Tuning Indicators

Tuning indicators are used to facilitate the process of antenna tuning without the need to use antenna analyzers or other instrumentation. They present the advantage of being a self-contained system that has been optimized for specific interrogators and antennas. On the other hand, because of their specific nature, they can only be used with a limited number of antennas.

For example, Texas Instruments manufactures the RI-ACC-ATI2 antenna-tuning indicator used to help tune their LF antennas. The tuning indicator is connected to the interrogator and indicates to the user if the inductance or capacitance in the interrogator needs to be increased or decreased until the optimal value has been achieved. Some interrogators have a tuning capacitor, while some others have a tuning inductance that can be adjusted for optimal performance.

4.2.2 Antenna Multiplexers

Antenna multiplexers are used to connect multiple antennas to a single interrogator. This can be done to take advantage of the multiple coupling between antennas or when it is necessary to read data from multiple interrogators located at distant points not covered by a single antenna. The majority of multiplexers in the market work with two, four, or eight antennas. The multiplexer is connected between the radiofrequency output of the interrogator and the antennas, as shown in Figure 4.14.

It is not necessary for all the antennas connected to the multiplexer to have the same inductance. This presents the advantage of being able to choose the antenna most adequate for a specific application without having to physically connect and disconnect them from the interrogator. However, because most multiplexers do not include a tuning feature, each antenna must be tuned to the appropriate resonance frequency with their own individual tuning method described in Sections 2.2 and 2.3.

Antenna multiplexers are normally created by the manufacturers of interrogators, although there are some third-party companies that also

© 2011 by Taylor and Francis Group, LLC

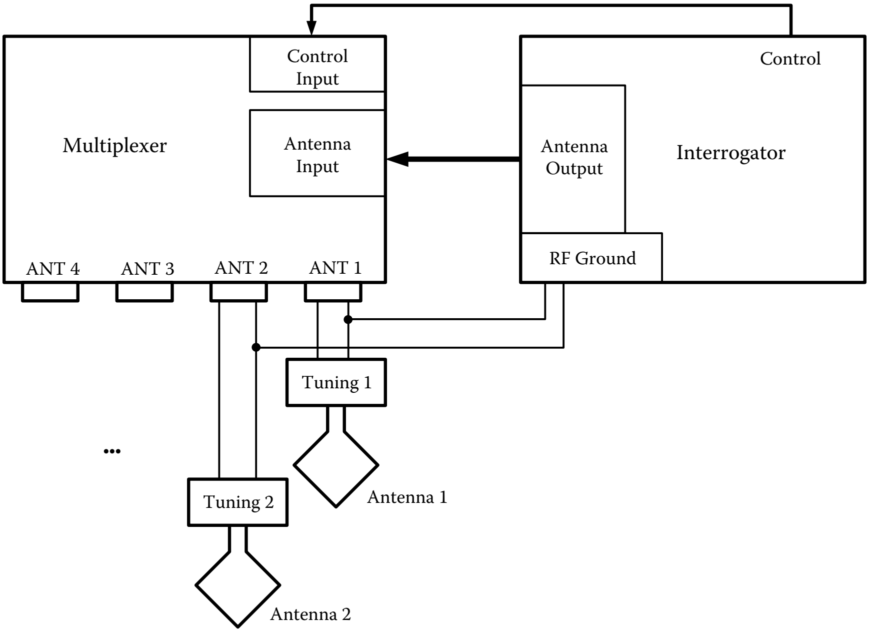

FIGURE 4.14
Multiplexer for connecting multiple antennas to a single interrogator.

manufacture them. For example, Skyetek manufactures multiplexers for the HF frequency range that can control four or eight antennas, having insertion losses lower than 0.7 dB and isolation between adjacent channels better than 45 dB.

4.2.3 Preamplifiers

Increasing the reading distance between interrogators and transponders requires a double approach. First, the electromagnetic field transmitted by the interrogator has to increase in order to generate enough power to turn on the distant transponders. This is relatively easy to do and is mainly limited by standards and regulations on maximum power and maximum allowed field strength. Second, reading distant transponders requires the interrogator to detect signals closer to the noise level. The increase in the detection sensitivity can be achieved with the use of preamplifiers, as shown in Figure 4.15.

The factor that limits the read range in RFID is ultimately the inherent noise that accompanies the signal received by the interrogator. A correctly designed preamplifier is able to increase the level of the signal without degrading the signal-to-noise ratio. The design of the main preamplifier is normally based on a low pass filter in order to reject the frequency generated by the field and

© 2011 by Taylor and Francis Group, LLC

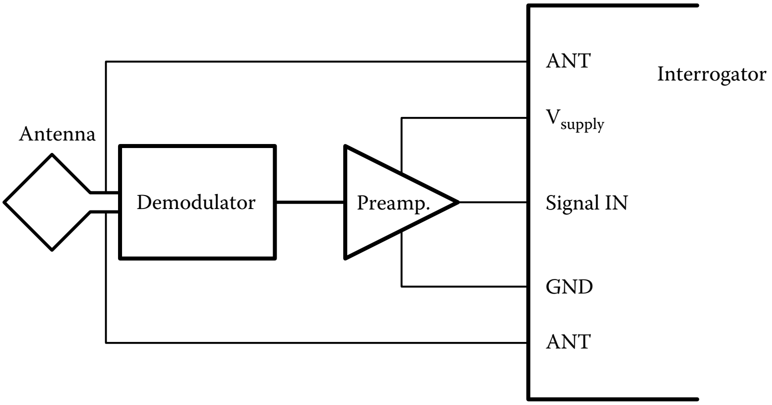

FIGURE 4.15
Preamplifier for extending reading distance.

allow the transmission of data. This stage is followed by an AC amplifier that increases the signal levels to the point that can be easily understood by the interrogator. The design of the preamplifier must consider the trade-offs between increasing the reading distance and the settling time required by the system. This means that the interrogator will be optimized for reading distance or for settling time, but cannot be optimized for both parameters.

Even when using preamplifiers, it is necessary to minimize the other potential sources of noise in the system. These can be the use of bad connectors between the interrogator and the antenna, improper installation of the transmission line between interrogator and antenna, an antenna that is not tuned correctly, the existence of interfering signals from nearby electronic equipment, and noise emanating from the power supply lines, among others. The designer of the RFID system should ensure that all these sources are minimized or at least known, and their values measured.

4.3 Antennas for UHF Interrogators

4.3.1 Design of Antennas for UHF Interrogators

Given the wavelength of signals in the UHF range, it is possible to design antennas for interrogators that operate in the far-field region as interrogators do not have the same space limitations of transponders. The resulting radiation pattern of two dipoles located at a short distance from each other can be easily modified by carefully arranging the distance between them as well as the phase shift of the signals that they radiate. A further modification of the radiation pattern comes from placing a large metal surface at a

© 2011 by Taylor and Francis Group, LLC

specific distance from the two dipoles. The metal surface acts like a shield that effectively removes half of the radiation pattern. Furthermore, if the distance between the antennas and the metal shield is one-fourth of the signal wavelength, the phase shift due to the distance between antenna and shield, combined with the phase shift created by the reflection in the metal shield, places the reflected signals in phase with the signals originated by the antenna. This, in turn, has the effect of doubling the power transmitted in the area away from the metal shield, as shown in Figure 4.16. It is necessary to remember that the increase of power in that specific direction comes at the expense of having no power in the area behind the metal shield, also known as a *passive reflector*. This structure is commonly used in *patch* or *flat* antennas.

An additional advantage of working in the far-field region is the possibility to take advantage of polarization. Polarized signals have their electric fields oscillating in a known and constant direction in space. This way, the signal recovered by the receiving antenna has a marked dependence on the geometrical relationship between the direction of the electric field and the direction of the antenna. Therefore, the energy recovered by the antenna in the interrogator ranges from a maximal value to zero depending on the relative orientation between the antennas in the interrogator and transponder. Using polarized signals allows discriminating between transponders. When the discrimination between transponders is

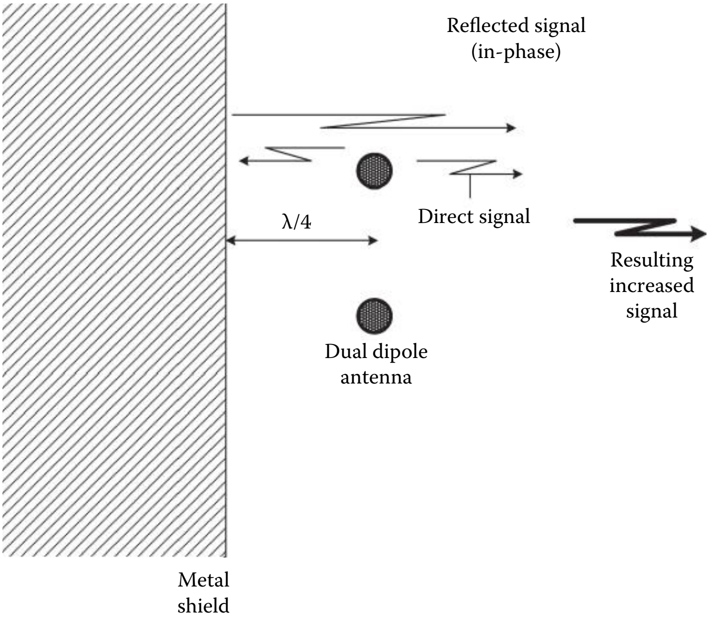

FIGURE 4.16
Metal shield increases effective radiated power in one specific direction.

© 2011 by Taylor and Francis Group, LLC

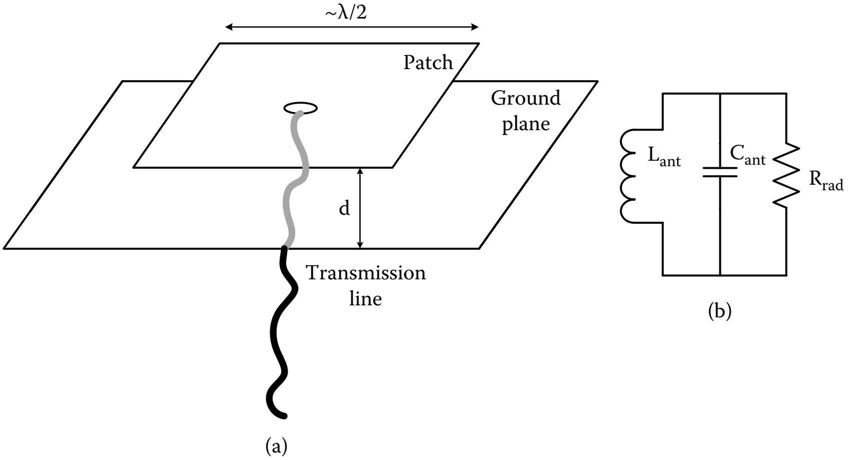

FIGURE 4.17
(a) Basic structure of a patch antenna. (b) Parallel equivalent circuit for a patch antenna.

not required, the use of dual dipole antennas in the transponders ensures that the interrogator will receive an adequate amount of energy.

The design of an antenna for the UHF RFID band represents an increased engineering challenge compared to designing it for the HF or LF bands because of the higher bandwidth needed for signals in the UHF band. It also becomes more difficult to maximize the performance of the antenna over the wider frequency range allocated for RFID systems in the UHF band. For this reason, while most of the HF and LF antennas follow the same basic design, engineers have developed several approaches to the design of UHF antennas for interrogators based on different performance parameters. This is especially critical for handheld applications in which the limited size of the antenna degrades its performance.

The patch antenna shown in Figure 4.17(a) is one of the most commonly used antennas for UHF interrogators. In its basic form, it consists of a metal patch of approximately one-half of a wavelength long and mounted over a ground plane. Figure 4.17(b) shows its equivalent electrical model.

Typical values for the equivalent circuit shown in Figure 4.17(b) are around 150 Ω for the radiation resistance of the antenna, between 9 and 10 pF for its capacitance, and from 2 to 4 nH for its inductance. The matching of the transmission line with a 50 Ω characteristic impedance to the relative high radiation resistance of the antenna is done by changing the location at which the transmission line feeds into the metal patch away from its center. The bandwidth of the patch antenna is a function of its dimensions and the separation between the metal patch and the ground plane. The type of dielectric used between the planes is also important, as dielectrics other than air will increase its equivalent capacitance.

© 2011 by Taylor and Francis Group, LLC

The patch antenna, by its very own design of using a ground reference as part of the antenna, is an unbalanced antenna. Because unbalanced antennas do not operate well when they are close to dielectric materials, and the human body can be essentially considered as a dielectric material at UHF frequencies, patch antennas are not the best choice for handheld interrogators. Furthermore, the size and weight limitations required by mobile and handheld interrogators further limit the use of patch antennas in these applications. A better solution for handheld and mobile interrogators are balanced antennas, that is, antennas with symmetric and opposed currents as shown in Figure 4.18.

When the antenna is fed using an unbalanced transmission line, such as a coaxial cable, it is necessary to use a balum transformer to match the unbalanced line to the balanced antenna. The basic theory behind this balum is the same as was described for HF and LF antennas. Although the structure shown in Figure 4.18 is technically feasible, an additional problem may arise due to the size required by half of a wavelength, especially for handheld interrogators. At the frequency of 915 MHz, the length of half of a wavelength is equal to 16 cm. While this is a very reasonable distance for a stationary antenna, it may be excessively large for applications involving handheld readers. In this case, it is possible to bend the dipoles in a similar way as is done for the antennas in transponders, as shown in Figure 4.19.

The reduction in size for the bent dipoles comes at the expense of degrading the electrical specifications of the antenna. In particular, its radiation resistance decreases and therefore the bandwidth of the bent dipole structure also decreases compared to the bandwidth of the regular dipole.

A different type of antenna used in UHF interrogators (either stationary or mobile) is based on the Yagi-Uda antenna. This structure, shown in Figure 4.20, consists of a driven element and two or more parasitic elements: one reflector and one or more directors. The parasitic elements direct the majority of the

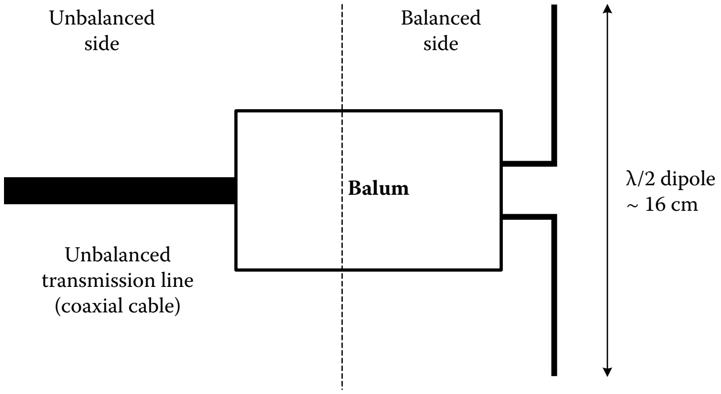

FIGURE 4.18
Balanced half-wavelength antenna.

© 2011 by Taylor and Francis Group, LLC

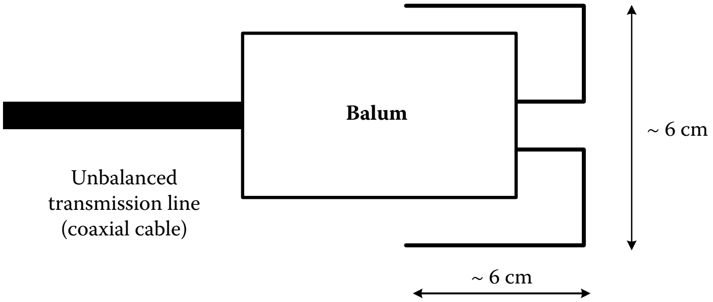
Balum
Unbalanced transmission line (coaxial cable)
~ 6 cm
~ 6 cm

FIGURE 4.19
Bent dipole for a UHF interrogator antenna.

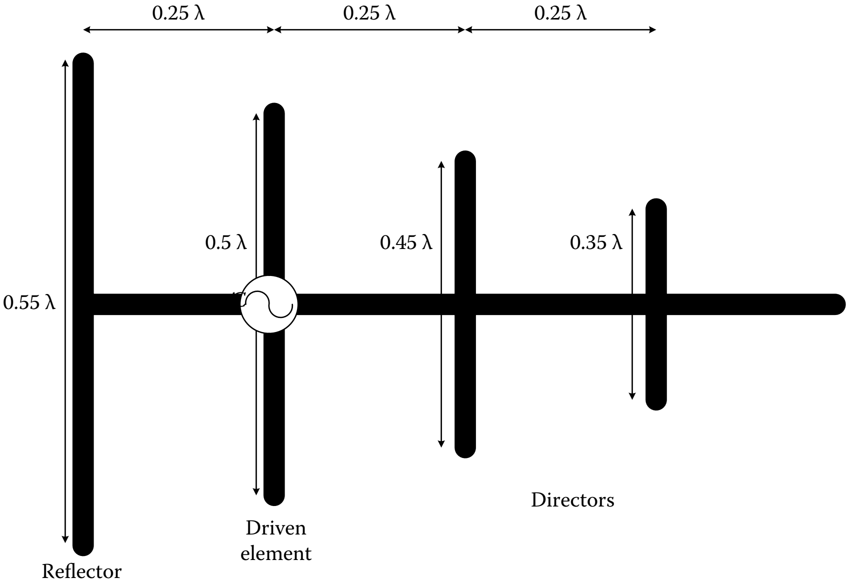
0.25 λ
0.25 λ
0.25 λ
0.55 λ
0.5 λ
0.45 λ
0.35 λ
Reflector
Driven element
Directors

FIGURE 4.20
Yagi-Uda antenna.

energy radiated by the antenna toward the director. Therefore, this structure helps to minimize the interaction of the radiated field with a person who may be holding an antenna for handheld interrogators. Moreover, because this is a balanced antenna, this interaction between the energy radiated by the antenna and that radiated by the human body is further minimized.

In its original form, however, Yagi-Uda antennas can still be considered too large to use with handheld interrogators. In this case, it is possible to bend the dipole and longer elements, as described above. The existence of the parasitic elements will provide additional gain compared to the single dipole.

© 2011 by Taylor and Francis Group, LLC

4.3.2 Transmission Lines and Connectors for UHF Interrogators

Contrary to LF and HF systems, the coaxial cables and connectors typically used in UHF RFID systems have the potential to extremely degrade the signal they transmit. The main reason for this problem is the skin depth. *Skin depth* describes the effect that the current flowing through a conductor flows through a portion of its inner surface. Equation (4.12) shows that the skin depth (δ) is inversely proportional to the frequency of the signal:

$$\delta = \sqrt{\frac{1}{\pi \mu \sigma f}} \tag{4.12}$$

where δ is the skin depth, μ the magnetic permeability of the conductor, σ its conductivity, and f the frequency of the current flowing through the conductor.

Therefore, an increase in the signal frequency results in a smaller portion of the conductor being used to transmit the signal. For a typical conductor made of copper, the skin depth for an LF signal is 0.2 mm, the skin depth for an HF signal is 55 μm, and for a UHF signal it is 2 μm. Consider a copper conductor with a radius of 0.2 mm. At LF frequencies and below, the current will flow through all the conductor area. When the frequency increases to 13.56 MHz, only the first 55 μm are actually available as a path for the current. The effective area has been reduced to a half of the area available for an LF signal. Because the conductivity of the conductor depends on its area, the conductivity of this conductor at HF is about half of its conductivity at LF. The situation changes more dramatically at UHF frequencies, as the skin depth of 2 μm means that current can flow only through the outer 1 percent of the conductor. With this, the conductivity of the copper wire for UHF signals has been reduced in a factor of 62 compared to the conductivity at LF. This strong decrease in conductivity introduces considerable losses that may render the system unusable. Figure 4.21 further illustrates this situation.

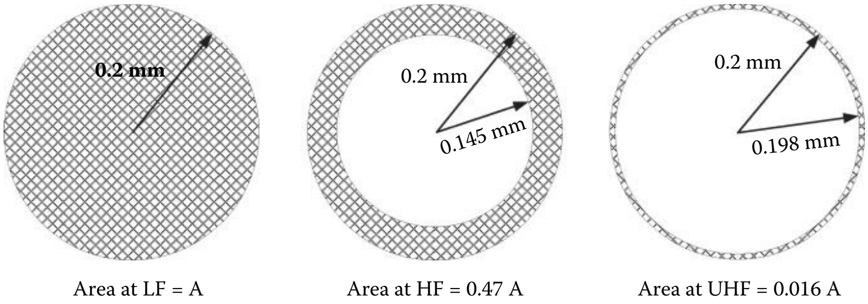

FIGURE 4.21
Skin effect at (a) low frequency (LF), (b) high frequency (HF), and (c) ultra-high frequency (UHF). Current flows through the shaded areas.

© 2011 by Taylor and Francis Group, LLC

With these considerations, the layout of the transmission line used to connect the UHF interrogator to its antenna becomes critical. The most effective approach to improving the performance of the system is to minimize the length of transmission line between the output of the interrogator and its antenna. It can also be possible to use transmission lines with larger diameter wires; however, the slight increase in effective area does not compensate for the additional cost of these lines or their increased rigidity. It is also important to choose a coaxial cable that has a continuous outer conductor instead of a multistrand ribbon for its shield.

The use of adequate connectors is also critical for RFID systems working in the UHF band. The BNC connectors that are commonly used in the LF and HF bands could be used without significant performance degradation in UHF applications. However, other types of connectors, specifically designed for higher frequencies, produce more reliable results. TNC connectors were conceived as a threaded version of the BNC connectors, being suitable for use up to 10 GHz. Type-N connectors are commonly used in high-frequency applications as they are suitable to carry signals up to 10 GHz and even 18 GHz with their recent enhancements. They are also mechanically robust and well suited for outdoor use. Figure 4.22 shows a male type-N connector, while Figure 4.23 shows a female type-N connector.

Because type-N connectors are available in 50 Ω and 75 Ω versions and the connectors are typically not labeled, it is very easy to confuse them, originating an impedance mismatch. Furthermore, the dimensions of their inner conductors are slightly different, so mismatching the connectors may

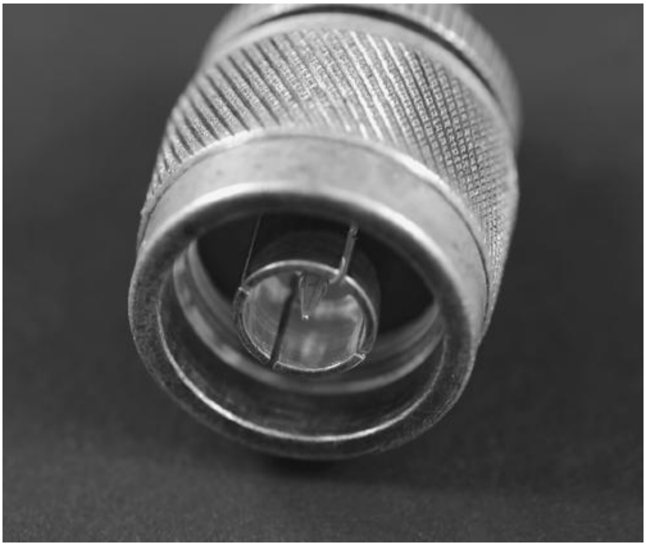

FIGURE 4.22
Male type-N connector.

© 2011 by Taylor and Francis Group, LLC

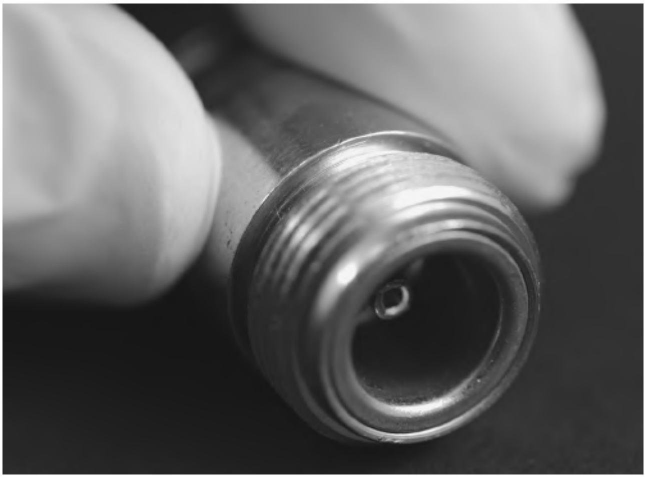

FIGURE 4.23
Female type N-connector.

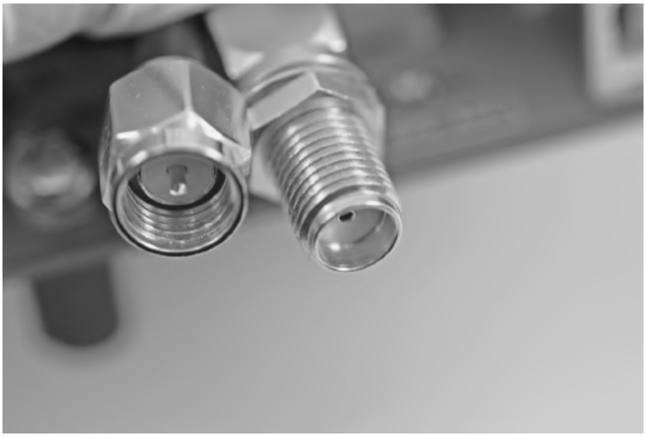

FIGURE 4.24
Male and female subminiature version A (SMA) connectors.

also damage them physically. Subminiature version A (SMA) connectors, shown in Figure 4.24, are perhaps the most widely used microwave connector and are becoming more used in RFID systems. SMA connectors can transfer signals up to 18 GHz and have a 50 Ω impedance. The main problem using this type of connector is that they have not been designed to be connected and disconnected constantly, like for example a BNC connector. Even with the best of care, they cannot exceed a few hundred connections and disconnections.

© 2011 by Taylor and Francis Group, LLC

4.4 Commercial Antennas for RFID Interrogators

4.4.1 Antennas for RFID Interrogators

Antennas for interrogators are available from the same manufacturers that make transponders and interrogators, as well as third-party vendors that build interrogator antennas for specific applications. Some manufacturers offer systems in which the antenna and the interrogator are integrated together, as is the case with handheld RFID systems.

Texas Instruments manufactures two types of antennas for the LF band under the generic name of Series 2000. These antennas are divided into *gate antennas* and *stick antennas*. Table 4.1 shows the specifications of these antennas.

As shown on Table 4.1, gate antennas mainly differ in their size. Designers of RFID systems choose the size of the antennas based on the required application, depending on the area that they have to cover. The stick antennas S01C and S02C are identical, with the only difference being the length of their coaxial cable: 1 meter for S01C and 3 meters for S02C. Choosing between gate antennas and stick antennas is also application dependent. In general, gate antennas offer a larger read zone and more read distance, while stick antennas offer a more focused zone. Therefore, if for example it is necessary to distinguish between different transponders, stick antennas will be more useful due to their more defined radiation field.

Some manufacturers have designed HF antennas that are shaped in a particular way for a specific application. For example, Unified Transponder Concepts, Inc., produces several antennas specific for laundry systems. One of them is shaped like a garment portal to be used with standard garment racks for clothes with an embedded RFID transponder. Another example from the same manufacturer is an antenna in the shape of a chute, used for the detection and control of deposited items tagged with the appropriate transponders. In addition, they also manufacture the more traditional panel and handheld antennas. The power levels accepted by these antennas ranges from 1 W to 4 W, and they all exhibit a resistance of 50 Ω at 13.56 MHz.

TABLE 4.1

Specifications for Some Texas Instruments Series 2000 Antennas Operating in the LF Band (134.2 kHz)

	Model	Inductance	Dimensions	Weight
Gate Antennas	G01E	27 μH	0.7 m x 0.2 m	745 grams
	G02E	27 μH	0.2 m x 0.2 m	425 grams
	G04E	26 μH	1.1 m x 0.52 m	2500 grams
Stick Antennas	S01C	27 μH	0.14 m x 2mm (Ø)	134 grams
	S02C	27 μH	0.14 m x 2 mm(Ø)	185 grams
	P02A	116 μH	0.13 m x 21.3 mm (Ø)	105 grams

© 2011 by Taylor and Francis Group, LLC

TABLE 4.2

Specifications for UHF Antennas Manufactured by NetHome

	Dimensions (mm)	Gain	Polarization	Connector
NT-900-SPP	90 x 90 x 25	0 dBi	Right / Left hand	SMA
NT-900-MPP	120 x 120 x 25	0 dBi	Circular	SMA
NT-900-CP	200 x 200 x 25	6 dBi	Right / Left had	SMA
NT-900-FP	360 x 260 x 25	6 dBi	Circular	TNC

TABLE 4.3

Specifications for MTI Wireless Edge Antennas Operating in the 400 MHz Range

Model	Frequency (MHz)	Dimensions (mm)	Gain	Polarization	Connector
MT182018	430 – 440	305 x 305 x 25	3.5 dBi	Circular LH	N-type Female
MT182016	430 – 437	370 x 370 x 30	9 dBi	Circular LH	N-type Female
MT181033	430 – 440	305 x 305 x 25	3 dBi	Linear	N-type Female
MT182011	405 – 450	370 x 370 x 40	8 dBi	Linear	N-type Female

TABLE 4.4

Specifications for MTI Wireless Edge Antennas Operating in the 865 MHz to 870 MHz Range

Model	Dimensions (mm)	Gain	Polarization	Connector
MT242032	190 x 190 x 30	7 dBi	Circular LH or RH	N-type Female
MT242040	190 x 190 x 30	7.5 dBi	Circular LH or RH	N-type Female
MT242033	540 x 470 x 220	8.5 dBi	Circular LH or RH	N-type Female
MT242034	685 x 626 x 220	8.5 dBi	Circular LH or RH	N-type Female
MT242027	260 x 260 x 30	8.5 dBi	Circular LH or RH	N-type Female
MT242014	305 x 305 x 25	8.5 dBi	Circular LH or RH	N-type Female
MT242017	371 x 371 x 40	10 dBi	Circular LH or RH	N-type Female
MT243017	650 x 320 x 40	11 dBi	Circular LH or RH	N-type Female

There is a larger number of independent manufacturers for antennas operating in the UHF range. For example, the Korean manufacturer NetHom offers flat, patch antennas that operate in the 902 MHz to 928 MHz range and have an input impedance of 50 Ω, as shown in Table 4.2.

Another manufacturer, MTI Wireless Edge Ltd., offers a vast array of standalone antennas for a wide number of frequencies and applications. Table 4.3 shows specifications for their line of antennas in 400 MHz UHF RFID range. All the antennas shown in Table 4.3 have a maximum input power of 6 W and an input impedance of 50 Ω.

The same manufacturer also offers antennas that operate in the 865 MHz to 870 MHz frequency range, have an input impedance of 50 Ω, and have a maximum allowable power of 6 W. Table 4.4 shows the rest of the specifications for these antennas.

© 2011 by Taylor and Francis Group, LLC

TABLE 4.5

Specifications for MTI Wireless Edge Antennas Operating in the 2.4 GHz Range

Model	Dimensions (mm)	Gain	Polarization	Connector
MT343024	190 x 190 x 30	13 dBi	Circular LH or RH	N-type Female
MT344034	305 x 305 x 15	15.5 dBi	Circular LH or RH	N-type Female
MT345014	371 x 371 x 40	19 dBi	Circular LH or RH	N-type Female

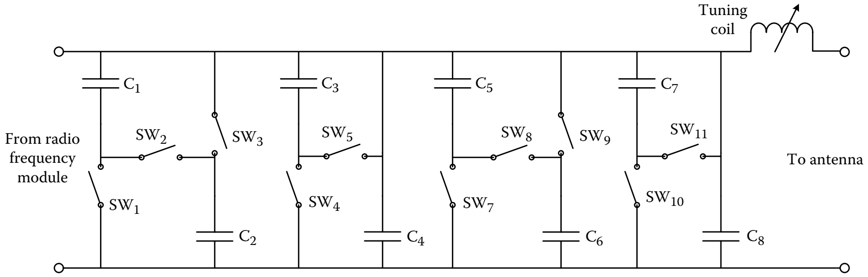

FIGURE 4.25
Schematic for the tuning board in the Series 2000 Reader from Texas Instruments.

It is also possible to purchase antennas similar to those shown in Table 4.4 dual polarized, that is, being both right-hand (RH) and left-hand (LH) circularly polarized. In this case, the antennas have two N-type female connectors, one for each polarization. Other models for antennas in the same frequency range of 865–870 MHz are available for linear polarization. In addition, this manufacturer also offers antennas for the 902–928 MHz and 950–956 MHz frequency ranges with dimensions and gains similar to those shown in Table 4.4 for either single or dual polarization.

Table 4.5 shows the specifications for a sample of antennas from MTI Wireless Edge Ltd. operating in the 2.4 GHz to 2.5 GHz frequency band and having an input impedance of 50 Ω.

Once again, antennas with similar dimensions and gains for the 2.4 GHz range are also available for single- and dual-linear polarizations.

4.4.2 Tuning Boards for Interrogator Antennas

Tuning boards are the circuits that are used to provide the high resonance needed to increase the strength of the field generated by the antennas in the system. Tuning boards can be found as stand-alone systems or as part of the interrogators. This subsection describes the Antenna Tuning Board that is part of the 2000 Series Reader System manufactured by Texas Instruments as an example of this type of circuits. As shown in Figure 4.25, the tuning board consists of a series of capacitors that are connected in parallel with the inductance of the antenna by means of several switches

© 2011 by Taylor and Francis Group, LLC

TABLE 4.6

Resulting Capacitance for Different Configurations of Switches in Tuning Board

Switch	Capacitance (nF)	Connection
SW1	33	C_2 ON
SW2	16.5	C_1, C_2 in series
SW3	33	C_1 ON
SW4	33	C_4 ON
SW5	16.5	C_3, C_4 in series
SW7	15	C_6 ON
SW8	7.5	C_5, C_6 in series
SW9	15	C_5 ON
SW10	10	C_8 ON
SW11	5	C_7, C_8 in series

(SW1–SW11) in the same board, thus allowing one to match a wide range of antenna inductances.

Table 4.6 shows the capacitive effect of each one of the switches in the tuning module. It can be seen that the lowest capacitance occurs when only SW11 is closed, resulting in a total capacitance of 5 nF. The highest capacitance that can be achieved with the board is equal to 139 nF, which occurs when SW1, SW3, SW4, SW7, SW9, and SW10 are closed.

The initial, coarse tuning adjustment is done with the capacitors, while the fine tuning due to the steps in the overall capacitance is done with the tuning coil located at the end of the tuning board. This process is repeated while monitoring the strength of the field generated by the antenna until it reaches a maximum.

© 2011 by Taylor and Francis Group, LLC

5

Interrogators

CONTENTS

5.1 Transmitter and Receiver.. 120
5.2 Modulator and Demodulator 126
  5.2.1 Amplitude-Shift Keying (ASK) Modulation............................... 126
  5.2.2 Frequency-Shift Keying (FSK) Modulation................................. 127
  5.2.3 Phase-Shift Keying (PSK) Modulation 129
5.3 Long-Range Considerations 133
5.4 Interrogator Synchronization... 134
  5.4.1 No Synchronization at the Interrogator Level............................ 134
  5.4.2 Wireless Synchronization 135
  5.4.3 Wired Synchronization 137
  5.4.4 Combined Wired–Wireless Synchronization 137
  5.4.5 Master–Slave Synchronization 137
  5.4.6 Carrier Phase Synchronization... 139
5.5 Structure of Commercially Available Interrogators 140
  5.5.1 Interrogators for LF RFID Systems.. 140
    5.5.1.1 Radiofrequency Modules.. 141
    5.5.1.2 Control Module 144
  5.5.2 Interrogators for HF RFID Systems... 149
  5.5.3 Single-Chip Interrogators 150

Interrogators have two basic functions: to generate and transmit the radiofrequency signal used to energize the transponders and to receive and decode the backscattered signal generated by the transponders. In addition, some interrogators may also transmit specific commands to the transponders. The interrogator also handles the bidirectional communication with a host computer used to process the information from the transponders and to issue commands to the interrogator. Figure 5.1 shows a basic block diagram of a generic interrogator with its basic constituent blocks.

If the transponder is able to accept commands from the interrogator, the radiofrequency signal generated by the interrogator has to be further modulated with the specific code; otherwise, the interrogator will only transmit the radiofrequency carrier. In either case, the radiofrequency signal is then amplified to specific power levels and passed through the tuning circuits before reaching the antenna. The antenna is also used to detect the backscatter signal generated by the transponder, which, after being filtered, is

© 2011 by Taylor and Francis Group, LLC

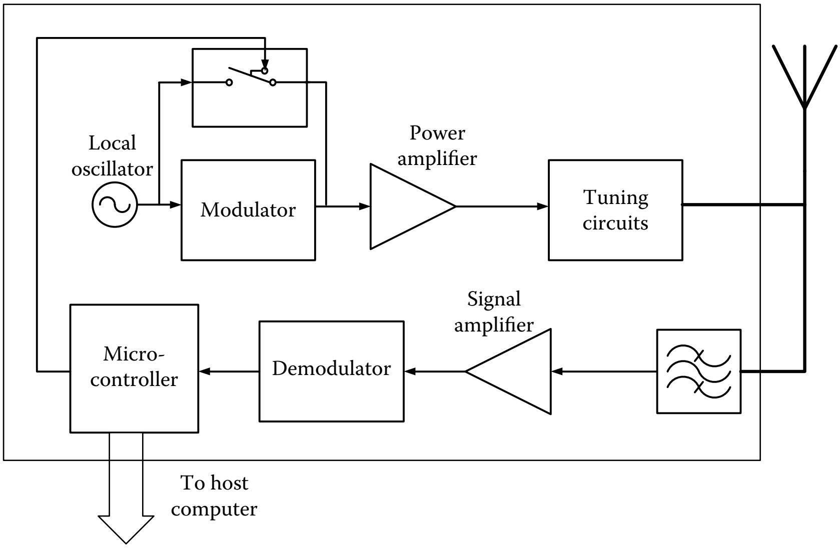

FIGURE 5.1
Basic block diagram of a generic interrogator.

then amplified and demodulated if necessary. The resulting signal is then sent to a microcontroller that will transmit it to the host computer using the appropriate communications protocol. The microcontroller also manages the timing for the interrogator to act as a transmitter or a receiver.

This chapter starts describing the transmitter and receiver sections in an interrogator. It continues by studying the most commonly used digital modulations in RFID systems. Afterwards, the chapter discusses issues related to extending the read range between transponder and interrogator as well as the requirements for interrogator synchronization when two or more interrogators coexist within the same general area. The last section in this chapter provides a description of the structure in different commercial interrogators.

5.1 Transmitter and Receiver

The basic structure of the transmitter section in the RFID interrogator contains an oscillator operating at the frequency of interest, a voltage or power amplifier, and the tuning circuits to match the impedance of the transmitter to the characteristics of the antenna. In addition to providing the radiofrequency signal to energize the transponders, some interrogators also transmit commands to the transponder. They can also be able to write new data into

© 2011 by Taylor and Francis Group, LLC

the transponder's memory. Transmitters in a RFID system must generate an accurate radiofrequency signal while minimizing the spurious radiation that they produce. *Spurious radiation*, common to all transmitters, is the amount of energy that falls outside the desired transmission band. In addition, they have to be as efficient as possible in order to reduce the amount of energy dissipated as heat. Excessive heat can severely reduce the life span of electronic components and must be reduced to acceptable levels.

For read-only transponders, the interrogator continuously transmits the radiofrequency signal and receives the response from the transponders. This mode of operation is known as *tag talks first*. When read/write transponders are being used, the interrogator must send commands for reading or writing them. For example, in the devices used by Microchip Technology, the interrogator uses amplitude modulation to transmit commands to the transponders. The communication starts by the interrogator sending specially timed gap pulses called *fast read request* (FRR) and *fast read bypass* (FRB). These gaps that are periods of time without electromagnetic field, are 175 μs wide, having a 100% modulation depth. The FRR or FRB pulses contain five of these gaps within 1.575 ms, as shown in Figure 5.2.

For RFID systems using magnetic coupling, one of the easiest approaches to drive the resonant circuit connected to the antenna is by using a D-drive circuit. Although either a full-bridge or a half-bridge will work correctly, most of the circuits use the half-bridge structure as it is less costly and easier to implement. Figure 5.3 shows a basic schematic for the half-bridge drive.

With a digital signal driving the circuit, the voltage at the node V_S is a square signal with voltages between V_{SS} and V_{DD}. Depending on the resonant quality factor (Q factor), the voltage across the antenna can reach a few hundred volts peak to peak. This is important to note at the time of selecting the electronic components used. An additional advantage of this structure is that although the driving digital signal has a very rich harmonic component,

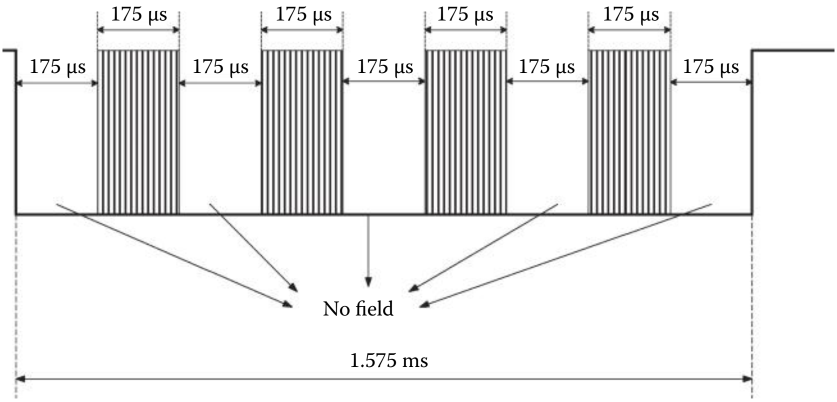

FIGURE 5.2
Fast read request (FRR) and fast read bypass (FRB) pulses.

© 2011 by Taylor and Francis Group, LLC

the signal across the antenna consists mostly of the fundamental component due to the filtering effect and the high Q factor value at the antenna. There are several devices commercially available that integrate the inverter and the two metal oxide semiconductors (MOSs) or metal oxide semiconductor field effect transistors (MOSFETs) in a single package. In addition to being able to operate at the high voltages generated, the capacitors used must have very low tolerance values due to the requirements for an elevated Q factor.

The typical rise and fall times for the signals using the circuit shown in Figure 5.3 range between 120 µs and 200 µs. This limits the data transmission rate that can be achieved. A modification of the circuit to increase transmission rates by reducing rise and fall times needs to accelerate the turn-on or turn-off response of the driver. This can be achieved by using the modified driver circuit shown in Figure 5.4.

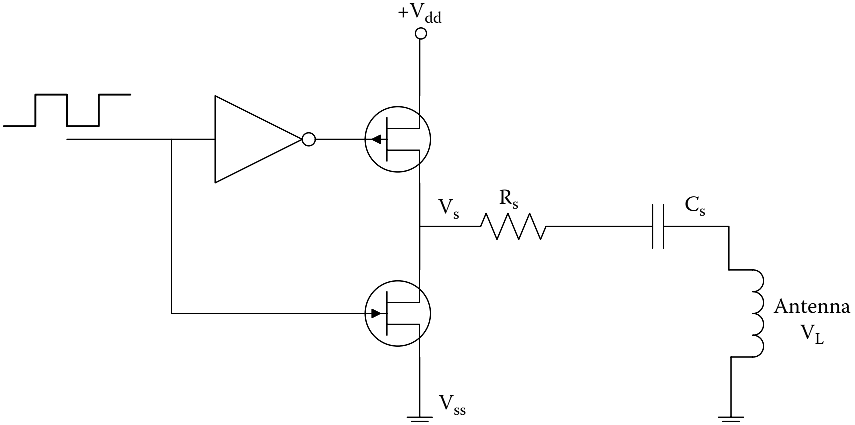

FIGURE 5.3
Half-bridge structure used to drive the resonant circuit.

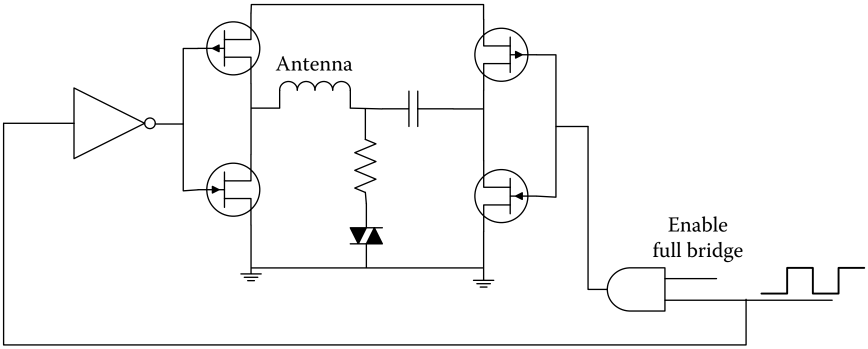

FIGURE 5.4
Modified bridge for increased transmission rates.

© 2011 by Taylor and Francis Group, LLC

The idea behind this circuit is to decrease the start-up time by starting the driver in full-bridge mode until the voltage has reached the desired value, at which point it can be switched back to half-bridge mode. Because the driver is started by driving the two half-bridges 180° out of phase, it effectively operates like doubling the input source voltage. With this structure, the circuit can respond up to three times faster. It is also necessary to keep in mind that the voltage across the capacitor will also increase and therefore the capacitor must be rated for that higher voltage. In order to reduce the electromagnetic interference generated by the circuit, the Triac is typically fired at zero.

The basic structure of the receiver in the RFID interrogator also contains a detector, typically an envelope detector, a signal amplifier, filters, and the appropriate pulse-shaping circuits. These circuits convert the distorted received signal at the output of the filters into pulses. These pulses are, in turn, connected to the input of a microcontroller that will send the appropriate data to the middleware or host computers for further processing.

Selectivity and sensitivity are key parameters for any radiofrequency receiver. *Selectivity* is defined as the extent to which a receiver can differentiate between the desired signal and unwanted signals present in its bandwidth. *Sensitivity* is defined as the minimum input signal that produces the minimal required output to be interpreted as information. Other important characteristics for receivers are their dynamic range, cross-modulation, and intermodulation distortion. *Dynamic range* refers to the range of signal levels simultaneously applied to the input for which the receiver will operate correctly. *Cross-modulation* and *intermodulation distortion* originate as a result of the nonlinear effects in the receiver that cause it to create new, undesired components.

For RFID systems with magnetic coupling, the transponder communicates with the interrogator by changing the value of the magnetic field. The transponder is magnetically coupled to the front end of the receiver in the interrogator. In this structure, changes in the magnetic field become voltage changes at the antenna in the interrogator. Figure 5.5 shows a basic block diagram following this approach. The peak detector extracts the envelope of the signal being transmitted. The peak detector is followed by a filter to remove the DC component of the envelope. A further low-pass filter removes the radiofrequency component, leaving only the envelope of the modulated signal. The output of the comparator is a digital signal with the information from the transmitted signal.

The peak detector must operate with the relative high voltages that may appear at the input of the tuning circuit. Moreover, it should not become a load to the tuning circuit either. Other requirements for the peak detector are to reduce the ripple in the carrier as much as possible, maintain the modulation of the signal, and possess low setting times, among others. It is not possible to optimize all these parameters simultaneously, thus requiring the designer of the system to find an acceptable compromise depending on its overall requirements.

© 2011 by Taylor and Francis Group, LLC

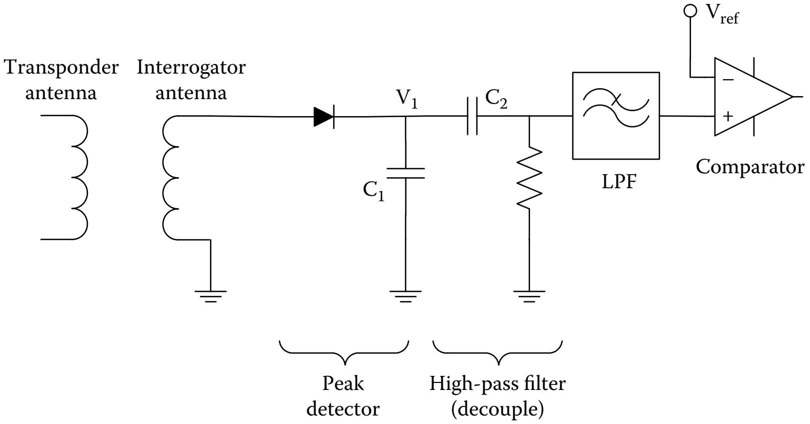

FIGURE 5.5
Basic structure for the detector in the RFID interrogator.

Example 5.1:

Discuss possible solutions for reducing ripple in the carrier.

Solution:

The ripple in the carrier can be reduced by increasing the value of the capacitor used in the peak detector. However, increasing the value of the capacitor will degrade the settling time for the detector, and will increase its loading effects. A possible solution would be to use an active peak detector. However, given the range of voltages expected at the input of the detector, the cost of the detector would significantly increase.

The voltage at node V_1 increases quickly when the input voltage increases due to the low impedance of the diode in forward mode. When the input voltage decreases, the voltage at node V_2 decreases extremely slowly due to the very high impedance presented by the diode in reverse conduction. This situation could be solved to some extent by placing a resistor in parallel with capacitor C_1 to help discharge it. However, this resistor has to be large enough for the peak detector to avoid loading the input circuits. A value of 10 MΩ is typically used.

The goal for the decoupling stage is to reject the high DC voltages present in the signal without excessive loading the front end. This stage should also have a fast dynamic response and stabilize quickly after detecting energy. The time constant of this stage is equal to the product of the resistance by capacitance C_2. The component values can be calculated by considering the delay time in establishing the communication after detecting the presence of radiofrequency energy as well as the minimum allowable voltage after the settling time. The cutoff frequency of the resulting high-pass filter has a strong effect in distorting and attenuating the digital signal that contains the information. With these conflicting requirements for the time constant, the

© 2011 by Taylor and Francis Group, LLC

simple RC circuit is not appropriate for removing the DC component of the received signal. Instead, it is necessary to use a circuit with either a nonlinear or a controllable time constant, as shown in Figure 5.6.

The diodes start conducting when the voltage across the resistor is above $(V_{REF} + V_{ON})$ or below $-(V_{REF} + V_{ON})$. This results in lowering the effective resistance seen by the circuit, and therefore it lowers its time constant. An additional benefit from using the diodes is that they protect the low-pass filter against positive or negative voltage spikes. The additional resistor R_2 is controlled by the switch. The switch is closed during transient periods and opened when data are being received by the circuit.

The low-pass filter stage is best implemented using an operational amplifier configured as an inverting amplifier, as shown in Figure 5.7. This structure not only allows controlling the gain in the pass band, but also minimizes its output impedance.

The cutoff frequency depends on the values of R_f and C_f. This frequency should be chosen at least two decades above the cutoff frequency from the DC blocker. It is also necessary to limit the gain of the filter to avoid the saturation of the operational amplifier. For some systems, the simple filter shown in Figure 5.7 may not provide enough attenuation at the desired frequencies. If this is the case, it can be substituted by a more complex filter. The three most common active filter topologies are Chebyshev, Butterworth, and Bessel filters. Chebyshev filters have the steepest transition between the pass and stop band, although they have ripple in the pass band. Butterworth filters have the flattest pass band, but their transition is not as steep as Chebyshev filters. Bessel filters have linear phase response with a smooth transition between

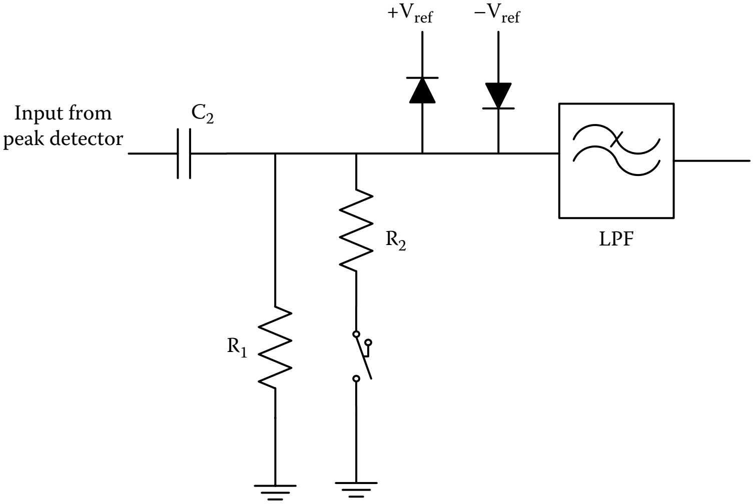

FIGURE 5.6
Circuit for implementing a nonlinear time constant.

© 2011 by Taylor and Francis Group, LLC

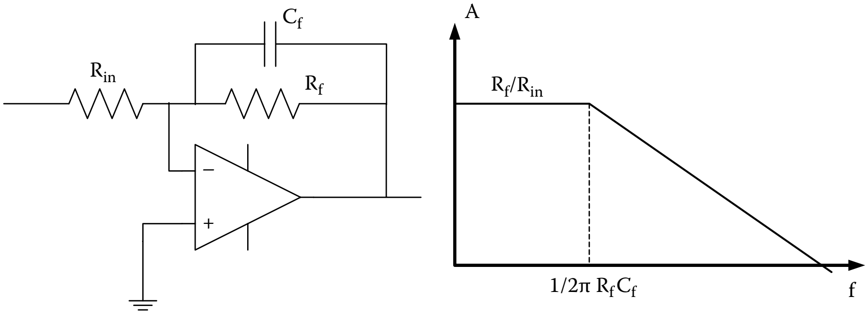

FIGURE 5.7
Low-pass filter stage.

pass band and stop band. When selecting one of these filter structures, it is necessary to consider not only their frequency response but also their transient response. For example, although Chebyshev filters have the best frequency response due to the fast transition between pass and stop bands, this results in an underdamped transient response with overshoot and ringing. Therefore, it has a very long settling time. Although Butterworth filters have a better transient response, they still exhibit some overshoot. The Bessel filter, although it has the worst frequency response, exhibits the best and fastest transient response and therefore is a very attractive candidate for this type of application.

The final stage is a comparator with some degree of hysteresis. The level of hysteresis is chosen to provide the required immunity against noise. The output of the comparator should be a digital signal that can be directly connected to the input of a microcontroller for further processing and transmission.

5.2 Modulator and Demodulator

5.2.1 Amplitude-Shift Keying (ASK) Modulation

Amplitude-shift keying (ASK) is the most basic form of modulation for digital signals. In ASK, data are represented as variations in the amplitude of the radiofrequency carrier. In binary ASK, there are only two symbols to be transmitted: the symbol *1* is normally represented by the full amplitude of the carrier, while the symbol *0* is represented by the minimal amplitude of the carrier. In this case, these two different symbols are represented by a single bit. In general, for M-ary ASK symbols, n bits are used to represent M symbols as

$$M = 2^n \tag{5.1}$$

© 2011 by Taylor and Francis Group, LLC

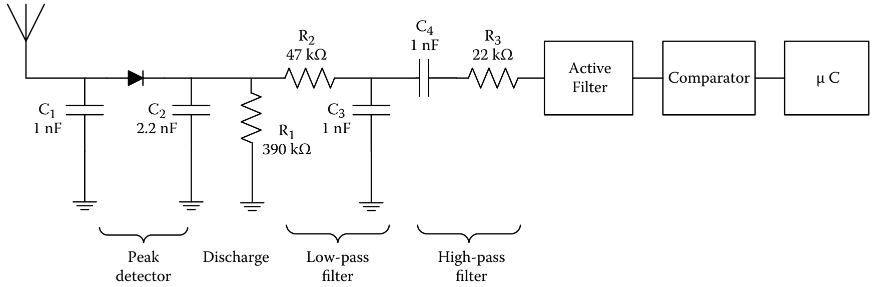

FIGURE 5.8
Detector in the receiver section of a commercial interrogator.

A *M*-ary ASK system requires distinguishing between *M* different amplitude levels for the radiofrequency carrier. While *M*-ary systems increase the transmission rate because each symbol encodes *n* different bits, the amplitude levels of the different symbols become closer to each other, making the system more susceptible to transmission errors. Therefore, there is a tradeoff between the transmission rate and acceptable transmission errors.

The ASK signal is demodulated by detecting the envelope of the carrier, normally using a half-wave rectifier. This detects the peak amplitude of the signal generated by the transponder and feeds it into an RC circuit that charges and discharges accordingly. The time constant of the RC circuit must be chosen with these two considerations: (1) it must be small enough so the voltage across the capacitor diminishes fast enough in order to keep with the changes in the envelope of the signal, and (2) it must be high enough to avoid excessive ripple in the detected signal. After the envelope has been detected, the resulting signal is passed through a low-pass filter and a signal-shaping circuit and finally fed into a microcontroller. Figure 5.8 shows the schematic of the detector in the receiver section of a commercially available interrogator working in the LF range at 125 kHz.

The capacitor C$_1$ provides the resonance for the inductance of the antenna. An additional advantage of this structure is that because C$_1$ is grounded, the carrier finds a path to ground. The diode and C$_2$ form the half-wave rectifier for which the resistor R$_1$ provides a discharging path. The time constant of this circuit is approximately 850 μs. The signal is further low-pass and high-pass filtered before entering the comparator for processing.

5.2.2 Frequency-Shift Keying (FSK) Modulation

Frequency-shift keying (FSK) is a form of modulation in which the digital signal to be transmitted changes the frequency of the carrier within a predetermined set of available frequencies. The most basic case, in which only two instantaneous frequencies can be transmitted, is called *binary FSK* (2FSK). For more complex systems, and in order to increase the bit transmission rate, the number

© 2011 by Taylor and Francis Group, LLC

of possible instantaneous carrier frequencies is higher, normally a power of 2. This situation is called *M*-ary FSK, where *M* represents the total number of different possible instantaneous frequencies, as shown in equation (5.1).

Binary FSK transmits two different frequencies f_0 and f_1 to represent a logic *0* or a logic *1* following an established protocol. The transmission of 2FSK can be easily accomplished by having two oscillators operating at frequencies f_0 and f_1, followed by a switch that selects one or the other based on the digital input. Although this is a simple procedure, it results in discontinuities of the carrier frequency at the time of switching, thus resulting in a greater prominence of high-frequency components that implies the need for a much wider bandwidth transmission. 2FSK can also be generated using the VCO approach as is done for analog signals. Because the digital modulating signal will never have a zero rise or settling time, this digital modulation is in fact an analog modulation with a very short transition time. In this case, the carrier has a continuous and smooth transition, therefore guaranteeing phase continuity and requiring a narrower spectrum for transmission. This method is known as continuous-frequency shift keying (CPSK) and can also be readily applied for multilevel modulation systems.

FSK signals can be detected by *noncoherent detection* or *coherent detection*. Noncoherent detection is based on simultaneously injecting the 2FSK signal into two band-pass filters, each one tuned to the transmitted frequencies, f_0 or f_1, as shown in Figure 5.9. In this case, the output of each filter is connected to an envelope detector and their outputs are combined to reconstruct the baseband modulating signal.

An alternative method for noncoherent FSK detection is based on injecting the FSK signal to the input of a phase-locked loop (PLL). The output of the internal *voltage-controlled oscillator* (VCO) will switch between frequencies following the signal at the output of the PLL, thus locking on to the incoming FSK signal so that the output changes between the logic levels of *0* and *1*, depending on the frequency of the incoming signal.

Figure 5.10 shows the basis for coherent detection of FSK signals. This approach requires the generation of two carriers with the same exact frequencies and phases as those used by the transmitter. In this case, only one

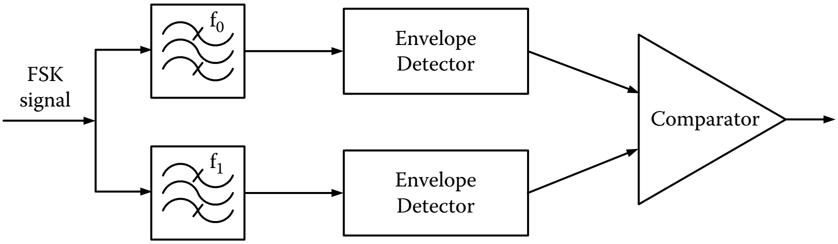

FIGURE 5.9
Noncoherent frequency-shift keying (FSK) demodulator.

© 2011 by Taylor and Francis Group, LLC

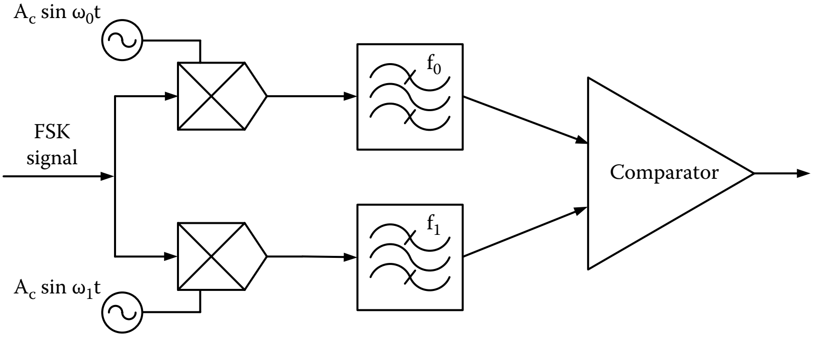

FIGURE 5.10
Coherent FSK demodulator.

of the filter outputs contains a signal at any given time. Therefore, after the addition of the two paths, the transmitting signal is recovered.

The coherent detection of FSK-modulated signals requires a more complex receiver than the noncoherent detection of the same signal, although it decreases the rate of errors for the transmission. Consequently, the designer of the system must consider the trade-offs between bit error rates and system complexity.

5.2.3 Phase-Shift Keying (PSK) Modulation

Phase-shift keying (PSK) modulation is one of the most effective methods to transmit digital signals. In PSK, the information is contained in the instantaneous phase of the modulated carrier. Usually, this phase is imposed and measured with respect to a known and selected phase of the carrier. For 2PSK, in which the modulating signal has two levels, the phase states are 0° and 180°. In M-ary modulation systems, the phase associated to each transmitted symbol is distributed equidistantly from each other around the 360° circle. When $M = 4$ ($n = 2$ bits), the resulting modulation is called *quadriphase phase-shift keying* (QPSK). QPSK, in which the phases are separated by 90°, is a common modulation used in communication systems.

Figure 5.11 shows a simplified version of one of the methods to generate a 2PSK signal. After the radiofrequency carrier is shifter by 180°, the resulting signal as well as the original signal are fed at the input of a 2-to-1 selector. This selector is driven by the digital signal. The output of the selector is either signal (a) or signal (b), depending on the value of the digital signal that controls the output of the selector.

PSK signals always require coherent demodulation, and therefore the PSK detector needs to know perfectly the phase of the unmodulated carrier. This requirement can be problematic as any phase error of the locally generated carrier reduces the effective voltage at the output of the detector in cos(Φ), thus degrading the signal-to-noise ratio. A solution to this problem is by the

© 2011 by Taylor and Francis Group, LLC

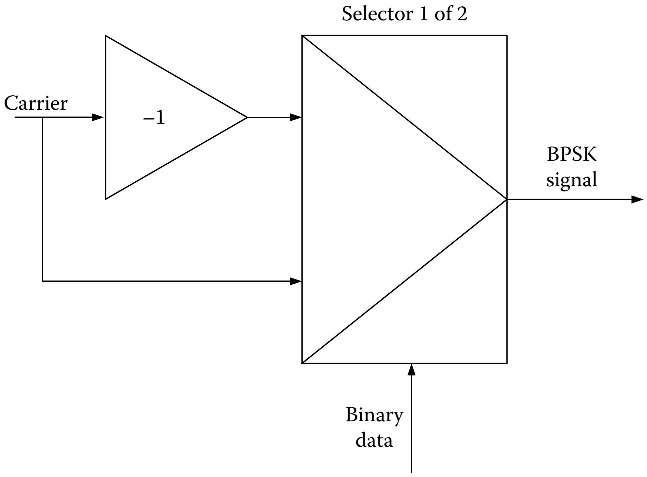

FIGURE 5.11
Binary phase-shift keying (2PSK) generation.

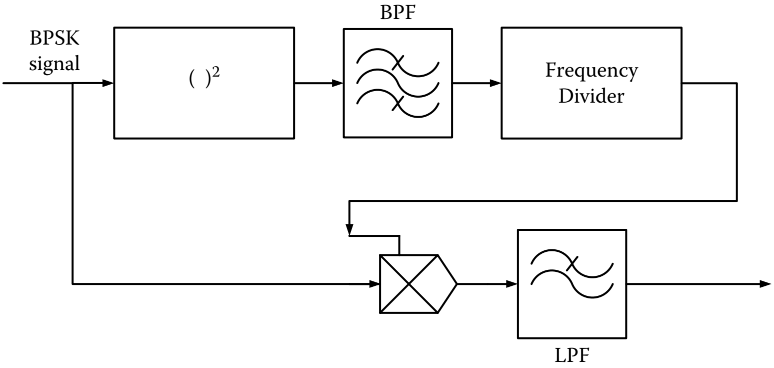

FIGURE 5.12
2PSK demodulation.

demodulator itself deriving a phase coherent reference to use in the demodulation process, as shown in Figure 5.12.

The output of the squaring circuit contains information at DC as well as at a double of the radiofrequency carrier. The band-pass filter eliminates the DC component. The frequency divider divides the frequency of the resulting signal by two and shifts its phase. The resulting signal is then multiplied by the original 2PSK signal that, after being filtered, will result in a DC signal with amplitude either positive or negative, thus shifted in 180°.

Figure 5.13 shows the generation process for QPSK signals. The digital-modulating signal $v_m(t)$ is connected to a 2-bit serial to parallel converter, so the resulting signals $v_1(t)$ and $v_2(t)$ change at half the rate of $v_m(t)$. This has the

© 2011 by Taylor and Francis Group, LLC

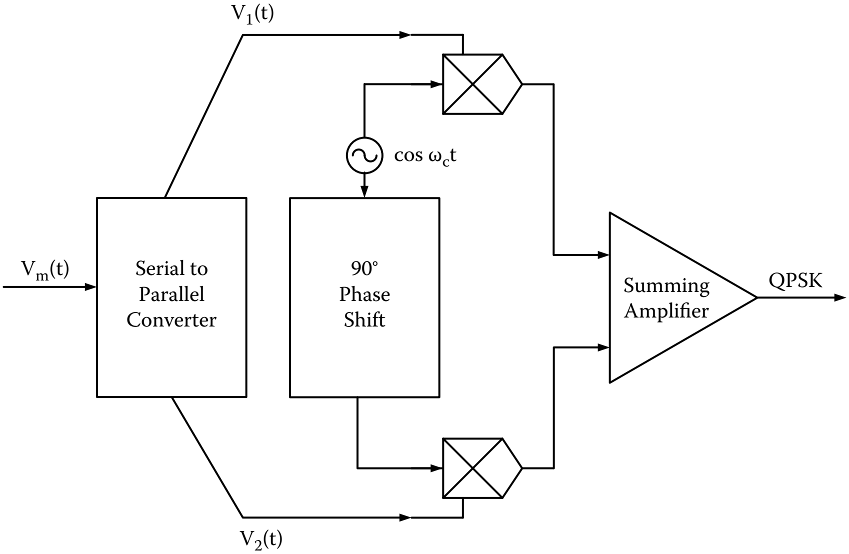
$V_1(t)$
cos $\omega_c t$
$V_m(t)$
Serial to Parallel Converter
90° Phase Shift
Summing Amplifier
QPSK
$V_2(t)$

FIGURE 5.13
Quadriphase phase-shift keying (QPSK) generation.

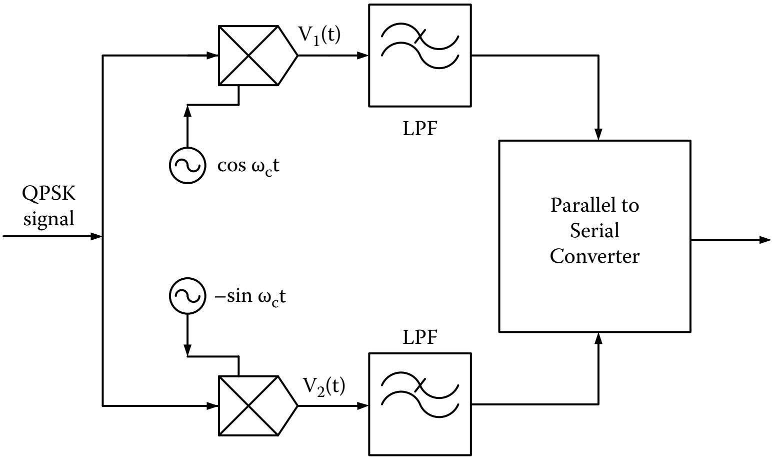
$V_1(t)$
LPF
cos $\omega_c t$
QPSK signal
Parallel to Serial Converter
−sin $\omega_c t$
LPF
$V_2(t)$

FIGURE 5.14
QPSK demodulation.

advantage of reducing the required transmission bandwidth as is the case of *M*-ary modulations. The output QPSK signal contains two components at the same frequency but shifted in 90°, and therefore there is no interference between them as they are in quadrature. That is, all the spectral components of the first term are 90° out of phase with those of the second term, making them independent from each other.

© 2011 by Taylor and Francis Group, LLC

The detection of the QPSK signal, as shown in Figure 5.14, follows an inverse process. The intermediate output signals $V_1(t)$ and $V_2(t)$ contain energy at DC as well as a frequency double the radiofrequency signal. This last component is eliminated by the low-pass filter, resulting in two separate streams of digital data. There are combined into a single data stream by the parallel-to-digital converter, thus recovering the original digital-modulating signal. QPSK modulation allows for sending information at twice the speed of 2PSK using the same bandwidth without compromising the performance of the system from the point of view of error rates.

Figure 5.15 shows a basic diagram of a 2PSK demodulator used in a commercial RFID interrogator operating in the LF range at 125 kHz. Capacitor C_1 provides the capacitance for resonance with the inductance of the antenna. The two diodes are used for detecting the envelope of the signal; each diode is on during half of the cycle. The 1 MΩ resistor is used to discharge the voltage store in capacitor C_2 used for the envelope detection. The envelope signal is fed into a series of active filters and pulse-shaping circuits not shown in Figure 5.15. The output of the last pulse-shaping circuit is a square wave at a frequency that is half of the radiofrequency carrier signal. This signal exhibits a 180° phase shifts according to the digital data sent by the transponders. This signal is now used as the clock for a D-type flip-flop. The input to the flip-flop is a signal at the same frequency derived from the transmitting signal. As the phase of the detected signal V_{out} changes, the output of the flip-flop also changes, thus recovering the digital information sent by the transponder.

One of the major problems with the design shown previously is that the phase of the returned signal can be, in certain conditions, unpredictable. For example, if the transitions of the incoming signal and the reference signal occur at the same time, the output of the flip-flop becomes unpredictable. This situation can be avoided with additional circuitry that intentionally shifts the phase between the signals to 90°, making it ideal for PSK detection.

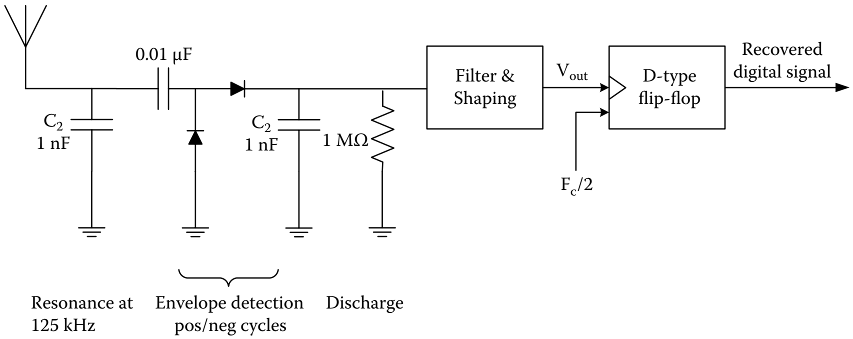

FIGURE 5.15
2PSK demodulator in a commercial interrogator.

© 2011 by Taylor and Francis Group, LLC

5.3 Long-Range Considerations

When evaluating the read range of transponders, it is necessary to understand that long-range operation may not always be a desirable condition. For example, in an environment where several interrogators communicate with their own transponders, it is necessary to ensure that a specific interrogator reads data from only the transponders that it has been assigned. Data from a distant transponder would then be treated like an interfering signal.

If long range is truly needed, it can be achieved by one or more of the following design considerations:

Increasing the transmitter output power. This can be achieved by designing the power amplifiers in the transmitter section in the interrogator with higher gain. Alternatively, when it is not possible to modify the interior of the interrogator, it is then possible to use an external preamplifier. In any case, it is necessary to comply with the limits set by the appropriate regulatory agencies on limits of power for different frequency range. Also, some regulatory agencies may also limit the ability of a user to modify a system that has already been approved and certified by the regulatory agency. This comes from the fact that the modification of the gain in the amplifiers can have effects on the frequency stability, spectral purity, and other characteristics of the transmitted signal that are specified by the regulatory agencies.

Modifying the size of the antenna. As described in Chapter 4, the size of the antenna and its quality factor (Q factor) affect the characteristics of the signal being transmitted and therefore the range at which transponders receive enough energy to activate their internal circuits. A correctly sized antenna becomes critical for long-range applications.

Optimizing the sensitivity of the receiver. Merely increasing the power in the transmitter section may not have a significant effect in the read range of the system if the signals generated by the transponder cannot be detected by the receiver. Therefore, interrogators to be used for long-range applications must have front-end amplifiers with high enough sensitivity that will allow them to detect the signals from the transponders at the longest distance from the interrogator that has been specified. In addition to the optimal design of the receiver, the sensitivity of the interrogator can be increased by reducing the amount of noise and interference added to the desired signal. This can be achieved by appropriate shielding of the front-end components, by placing the interrogator away from electrically noisy sources and using pass-band filters.

© 2011 by Taylor and Francis Group, LLC

5.4 Interrogator Synchronization

Synchronization is used to prevent interference between the interrogators in applications that have multiple interrogators operating in the same area. Synchronization is based on preventing interference by coordinating their transmission and reception windows. Synchronization is only required when the different interrogators are located physically close to each other. Therefore, when working with multiple interrogators, it is critical to determine if synchronization is required. However, the distance between interrogators alone is not sufficient to determine the need for synchronization. The electrical path between interrogators is affected by the presence of metallic structures such as buildings, the existence of conveyors, the layout of nearby power or data cables, and even the existence of reinforcing metal bars in concrete floors. Assuming only radiated signals, Texas Instruments has published some guidelines for the maximum distances that require synchronization for interrogators using some of their antennas as shown in Table 5.1. It is important to note that Table 5.1 assumes the two interrogators are using the same antenna type.

Table 5.1 shows that antennas with larger dimensions require more separation. Because this table assumes only radiated signals, in more realistic situations, this distance can be highly reduced due to the additional paths available to the different radiofrequency signals. This table can also be used as a starting point when working with interrogators using different antennas.

The next step, after determining that synchronization between interrogators is needed, is to decide on the method used to synchronize them. Although the following are the synchronization methods available to some of the Texas Instruments interrogators, the same methods are used by interrogators from other manufacturers.

5.4.1 No Synchronization at the Interrogator Level

This method can only be used with a single reader or in the case of synchronization by software. It represents the fastest method for receiving

TABLE 5.1

Separation between Interrogators After Which Synchronization Is No Longer Necessary Assuming Only Radiated Signals

Antenna Type	Antenna Dimensions	Minimum Separation Distance
S02-C (stick antenna)	Ø = 21 mm; l =140 mm	18 meters
G02-E (gate antenna)	200 mm x 200 mm	32 meters
G01-E (gate antenna)	715 mm x 170 mm	55 meters
G04-E (gate antenna)	1018 mm x 518 mm	105 meters

© 2011 by Taylor and Francis Group, LLC

information from the transponders. If all the interrogators are connected to the same network, the coordination can be carried out by the host computer issuing commands to read transponders simultaneously and buffering the results, or issuing different commands to different interrogators in turn. In any case, the coordination is done by the host computer rather than by the interrogators.

5.4.2 Wireless Synchronization

In wireless synchronization, when a specific interrogator detects increased energy at the frequency of the carrier, it assumes that this energy has been generated by another interrogator. This makes it stop itself before starting its own cycle. Obviously, this method only works when the background noise is low for the type of antenna being used. Larger antennas, which are more sensitive, present a clear disadvantage in this situation. When the environmental conditions change and there are additional reflections of signals in the immediate electrical environment, this approach is not valid either.

The receiver in each individual interrogator is adjusted in order to determine the level of background noise in the absence of signals from other interrogators. This determines the level after which the interrogator assumes that another interrogator is transmitting and therefore waits for a predetermined amount of time before starting its own cycle. This situation is illustrated in two situations depicted in Figure 5.16.

Figure 5.16(a) shows a time diagram involving wireless synchronization when the interrogator has not detected the presence of other signals. The interrogator starts by transmitting the radiofrequency signal for 50 ms that is used to energize the transponder. After the transponder is energized and recognizes the end of the radiofrequency energy burst, it starts transmitting its data for 20 ms. Once the transponder has finished sending its data, the interrogator waits for another 20 ms before starting the whole process again. This additional period is build as a precaution to avoid corrupting the data from other transponders to other interrogators. Because the interrogator does not detect any signals during this last 20 ms period, the whole cycle takes 90 ms.

In Figure 5.16(b), however, after the transponder has completed sending its information to the interrogator, the interrogator detects the presence of radiofrequency energy, and therefore responds by waiting 70 ms before starting its new cycle. This additional period gives enough time for the second interrogator to finish its own cycle. After the 140 ms have elapsed, the first interrogator starts its own cycle regardless of the presence of radiofrequency energy. This is done in order to prevent the interrogators always giving priority to other interrogators and therefore not being able to read their own transponders.

Wireless synchronization has the advantages that it does not need wires to interconnect the interrogators and all the interrogators are autonomous

© 2011 by Taylor and Francis Group, LLC

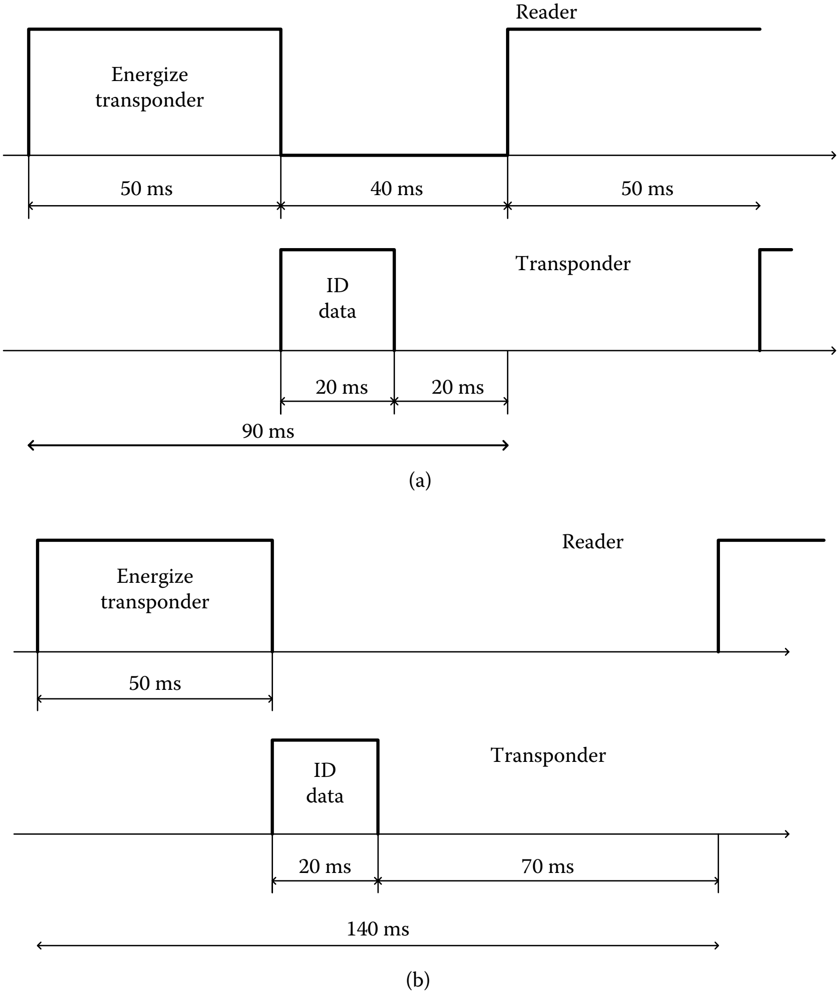
Reader
Energize transponder
50 ms
40 ms
50 ms
Transponder
ID data
20 ms
20 ms
90 ms
(a)
Reader
Energize transponder
50 ms
Transponder
ID data
20 ms
70 ms
140 ms
(b)

FIGURE 5.16
Wireless synchronization: (a) no other interrogators detected, and (b) delayed cycle due to other interrogators detected.

as there is not a master unit. It is very useful when handheld units have to coexist with fixed interrogators for these reasons. Wireless synchronization, however, is not adequate when large antennas are used in noisy environments because these large antennas are too sensitive to the background noise levels. Also, it cannot be used when the environmental conditions change, for example when vehicles enter or leave the read area as it changes the background energy. Finally, it cannot be used when other interrogators write data to the transponders.

© 2011 by Taylor and Francis Group, LLC

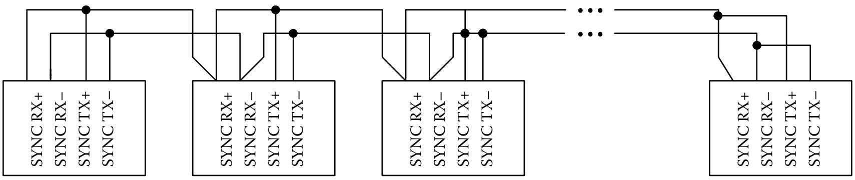

FIGURE 5.17
Structure of the control modules in interrogators using wired synchronization.

5.4.3 Wired Synchronization

The approach in wired synchronization is the same as for wireless synchronization with the only difference being that it uses wires to detect the presence of other interrogators instead of detecting radiofrequency energy. Therefore, the cycle for a single interrogator lasts 90 ms or 140 ms as described for wireless synchronization. This presents the advantage of being able to operate regardless of the environmental noise level, using a single twisted pair cable. This approach does not need a master unit. The main problem with wired synchronization is that if the power for a single interrogator fails, the bus fails and therefore the whole system fails. It is also not suitable for read/write transponders. Figure 5.17 shows the standard connection for the control modules of the interrogators using wired synchronization using the following nomenclature:

SYNC RX+: Noninverted data input
SYNC RX–: Inverted data input
SYNC TX+: Noninverted data output
SYNC TX–: Inverted data output
GND: Signal ground

5.4.4 Combined Wired–Wireless Synchronization

Using this approach, interrogators synchronized using a hard wire connection are also detecting the amount of radiofrequency energy to decide if another interrogator is already operational. This method resolves the individual disadvantages of each single approach.

5.4.5 Master–Slave Synchronization

In this approach, which is probably the most commonly used method of synchronization, one of the interrogators is configured as the master while the rest of the interrogators are configured as its slaves. Master–slave synchronization can be carried out using three variants:

© 2011 by Taylor and Francis Group, LLC

Master–slave without acknowledgment. This method is the fastest approach for reading transponders as it assumes that all the interrogators are part of the same synchronization bus. This requires the interrogators to have a synchronization bus that connects all the readers, using for example a single twisted pair as shown in Figure 5.18.

Once the master has finished a complete read cycle, it transmits a synchronization pulse in the bus. Any slave that has finished its own cycle is waiting for this pulse in order to start a new cycle, as shown in Figure 5.19.

This approach presents the advantage of needing only a single twisted pair. It also has the fastest read rate and can be used when transponders are rewritten by interrogators. It is necessary, however, to keep in mind that if the master fails, the rest of the units stop working. Also, if a slave reads a transponder and the master does not, the slave may miss the next pulse as it is busy processing the information from its transponder.

Master–slave with acknowledgment. In this approach, the master has to wait until all the slaves have completed their cycles before starting a new master cycle. This requires using four wires: two for the pulse from the master and two for the information from the slaves to the master, as shown in Figure 5.20.

All the interrogators are running free while buffering the data via a point of data connection. A slave will not respond to a command from the host computer unless it has received the pulse from the master. This method presents the advantage that all the interrogators wait for the slowest to be complete and all the units are able to write to the transponders. This happens at the price of using a twin twisted pair instead of a single.

Master–slave triggered synchronization. This is similar to the previous approach, although in this case, the pulse is generated by an external device instead of an interrogator in the chain. All the interrogators

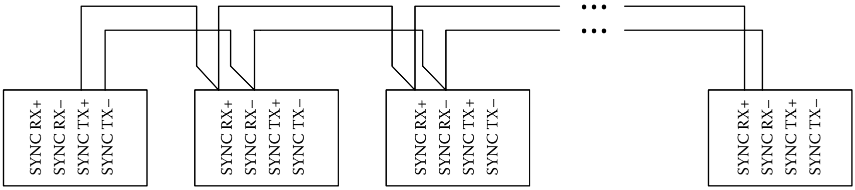

FIGURE 5.18
Structure of the control modules in interrogators using master–slave synchronization without acknowledgment.

© 2011 by Taylor and Francis Group, LLC

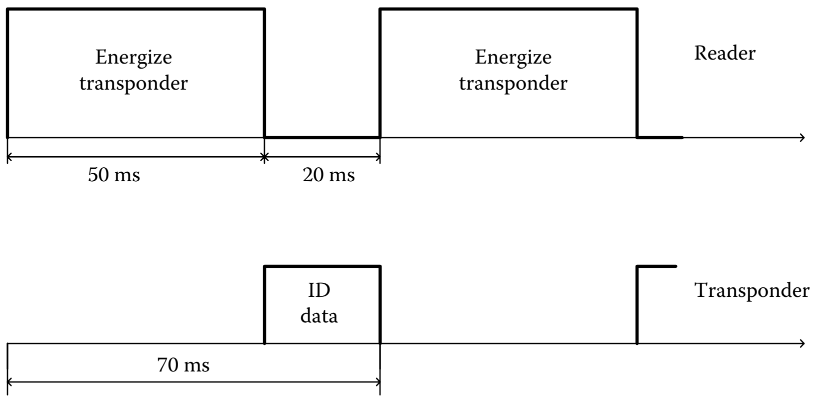
Energize
transponder
Energize
transponder
Reader
50 ms
20 ms
ID
data
Transponder
70 ms

FIGURE 5.19
Master–slave synchronization without acknowledgment.

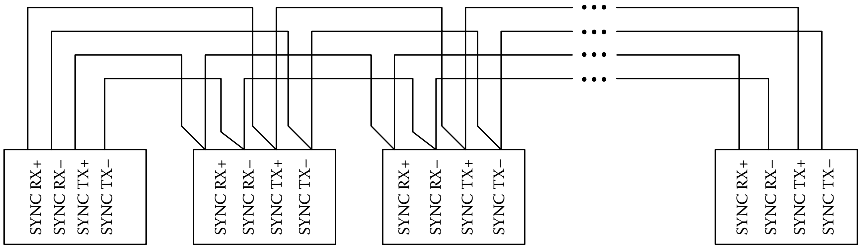
SYNC RX+
SYNC RX–
SYNC TX+
SYNC TX–
SYNC RX+
SYNC RX–
SYNC TX+
SYNC TX–
SYNC RX+
SYNC RX–
SYNC TX+
SYNC TX–
SYNC RX+
SYNC RX–
SYNC TX+
SYNC TX–

FIGURE 5.20
Structure of the control modules in interrogators using master–slave synchronization with acknowledgment.

are configured as slaves with the external master issuing the synchronization pulses at suitable intervals. This is required for interrogators that must read and write to multipage transponders in which different interrogators require different window times. In this case, the master has total control over the coordination of the devices in the chain. However, the master must be a more complex system and requires additional intelligence.

5.4.6 Carrier Phase Synchronization

This method is used when several antennas are located close together with the possibility that the field generated by each one of them may interact with the others. Because of the different transmission phases at each interrogator, the interaction may result in destructive interference. To avoid this effect, all the interrogators must be operated from the same carrier signal.

© 2011 by Taylor and Francis Group, LLC

5.5 Structure of Commercially Available Interrogators

Interrogators for RFID systems take different approaches depending on the specific needs of their intended application: from large and complex systems to single-chip interrogators. The following pages describe these different approaches through examples from commercial interrogators. It is necessary to view them just as specific examples as the number of available systems is higher, each one of them using a different approach. However, these examples provide valuable information about the structure of interrogators.

5.5.1 Interrogators for LF RFID Systems

The Series Reader 2000 from Texas Instruments is an example of a complex interrogator for an RFID system operating in the low-frequency range. Its main component is the S251B reader. This interrogator has a communications interface as RS232, RS422, or RS485; eight general-purpose input/output lines; two open collector outputs; one synchronization bus; one carrier phase synchronization bus; as well as a power connector and an antenna connector. Table 5.2 shows the electric specifications of the eight general-purpose input/output (I/O) lines in this unit.

This interrogator supports two different communication protocols: the ASCII protocol or Texas Instruments' own TIRIS protocol. The ASCII protocol that is used to send commands to the interrogator can only be used with the RS232 or RS422 interface. The TIRIS bus protocol is a binary protocol between a host computer and one or more interrogators.

The interrogator has two main modules: the *control module* and the *radiofrequency module*. The control module contains all the electronics needed for the communication between the interrogator and the host computer in order to control the radiofrequency module. It also has circuits used to optimize the performance of the antenna by correctly tuning the module. The radiofrequency module handles all the analog functions needed to generate the

TABLE 5.2

Electrical Specifications of S251B Reader (Texas Instruments)

Parameter		Minimum	Maximum
General Purpose I/O Voltage			
	Low Level	---	0.9 V
	High Level	3.15 V	5.25 V
General Purpose I/O Current			
	Low Level	----	25 mA
	High Level	----	16 mA
General Purpose I/O 1 to 4 Total output current		----	10 mA
General Purpose I/O 5 to 8 Total output current		----	10 mA

© 2011 by Taylor and Francis Group, LLC

signal that energizes the transponders, send commands to the transponders, and detect their responses. These two modules will be studied in deeper detail in Sections 5.5.1.1 and 5.5.1.2.

The synchronization bus operates in differential mode with a maximum data rate of 10 Mbits/s and allows for a maximum cable length of 1200 m. It can handle up to 32 different drives or receivers on the line. The outputs of the bus are protected against short circuits with a current limit of 150 mA to ground and 250 mA to the power supply. In addition, this unit contains three banks of dual in-line package (DIP) switches that allow for configuration changes. One of the banks is for setting up the control module, another bank for the settings of the communications protocol, and the third one for synchronization settings.

5.5.1.1 Radiofrequency Modules

The functions of the radiofrequency modules in the RFID system are to transmit the radiofrequency signal to the transponder via the antenna, to receive and demodulate the response of the transponder, and to write information to the transponder. Figure 5.21 shows a simplified block diagram of a generic radiofrequency unit in an interrogator.

For the modules that are part of the Series 2000 Reader manufactured by Texas Instruments, it is possible to distinguish the following submodules:

Power supply. This submodule provides the necessary voltage to power the rest of the electronics in the radiofrequency module. The majority

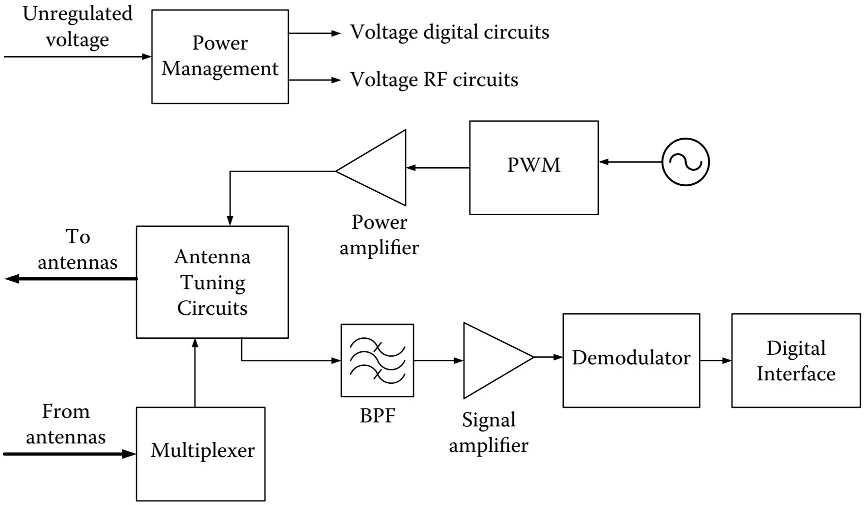
Unregulated voltage
Power Management
Voltage digital circuits
Voltage RF circuits
PWM
Power amplifier
To antennas
Antenna Tuning Circuits
BPF
Signal amplifier
Demodulator
Digital Interface
From antennas
Multiplexer

FIGURE 5.21
Simplified block diagram for the radiofrequency unit.

© 2011 by Taylor and Francis Group, LLC

of modules contain two voltage regulators, one to power the logic circuits and the other to power the rest of the module.

The manufacturer stresses not using a switching power supply, as at their typical frequency of operation their harmonics will create strong interfering signals.

Transmitter. The radiofrequency signal is generated by a crystal oscillator operating at the RFID frequency of 134.2 kHz. This signal is then fed into the pulse width modulator for which the user can select the width ratio between 3% and 50%. It is necessary to consider that decreasing this ratio also decreases the energy of the transmitted signal, thus providing a method to control the power transmitted by the interrogator.

Antenna circuit. The purpose of this submodule, shown in Figure 5.22, is to match the impedance of the transmitter to the impedance of the antenna in order to maximize the radiated power as well as its Q factor.

The circuit consists of the inductance of the antenna in series with an adjustable inductance for fine tuning as well as the capacitive part formed by the capacitors C_1, C_2, C_3, C_4, and C_{couple}. The switch is used to connect or disconnect C_4 for coarse tuning. It is necessary to connect several capacitors in series because of the high voltages that can be developed across the antenna terminals during resonance.

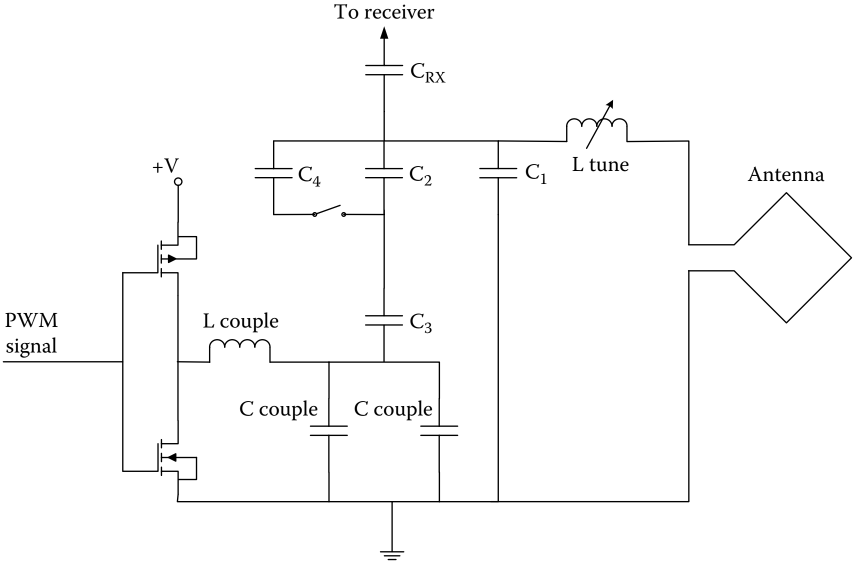

FIGURE 5.22
Antenna-tuning circuit.

© 2011 by Taylor and Francis Group, LLC

The capacitive resonator is driven by the *pulse-width modulation* (PWM) signal from the transmitter by two MOSFET power transistors through the inductor L_{couple}.

Receiver. The receiver in this module demodulates the FSK signals generated by the transponders, with a low symbol specified by a frequency of 134.2 kHz and a high symbol by a frequency of 123.2 kHz. The module can operate with a single antenna used for transmission and reception or with specific transmission and reception antennas. The path for the received signal depending on the type of antenna used is selected by the multiplexer at the front end.

Table 5.3 describes the different pins and its connections for the standard radiofrequency module in the Series 2000 Reader from Texas Instruments. Modules from other manufacturers have similar connections.

The most important electrical characteristics for these pins and connectors are shown in Table 5.4.

As shown in Table 5.4, the radiofrequency module incorporates a signal strength detector whose output is the RXSS- pin. This allows adjusting the threshold level for the detection of radiofrequency energy required by wireless synchronization. The adjustment is done by means of a potentiometer on the board. Different types of antennas require different threshold levels even if they operate in the same environment.

TABLE 5.3

Pins and Connectors in Standard Radiofrequency Module

Signal Name	Connection	Description
GND	IN	Ground logic circuits
TXCT-	IN	Transmitter control input for activation
VSL	IN	Supply voltage
RXDT	OUT	Receiver data signal output
RXSA	IN / OUT	Adjustment for receiver threshold level
RXCK	OUT	Clock output receiver
GNDP	IN	Ground power stage
RXA0	IN	Select signal for multiplexer
GNDP	IN	Ground power stage
RDTP	OUT	Test pin for receiver
VSP	IN	Supply voltage power stage
RSTP	OUT	Test pin for receiver
VSP	IN	Supply voltage power stage
RXSS-	OUT	Signal strength output in receiver
VD	IN / OUT	Internally regulated voltage
RCTP	OUT	Test pin for receiver

© 2011 by Taylor and Francis Group, LLC

TABLE 5.4

Summary of Electrical Characteristics Radiofrequency Module

Parameter	Min	Typ	Max	Unit
Internal regulated voltage	4.75	5.00	5.25	V
Supply current receive mode		9.0		mA
Supply current transmit mode		11.0		mA
Low level input voltage at TXCT- and RXA0	0		0.8	V
High level input voltage at TXCT- and RXA0	2.4		5.0	V
Low level output voltage at RXDT and RXCK	0		0.8	V
High level output voltage at RXDT and RXCK	4.0			V
Low level output voltage at RXSS-			0.8	V
Cable length for connection to control module		0.5	2.0	m
Number of slave modules that can be driven		1	5	
Inductance of antenna tuning coil	1.3	3.0	4.7	µH
Total antenna resonator capacitor	46	47	48	nF

This radiofrequency module, as well as most of the modules available in the market, can be used with the transmit–receive antenna or with special receive-only antennas. The type of antenna to be used is selected by a multiplexer in the receiver section.

5.5.1.2 Control Module

The control modules provide the functionality needed for the radiofrequency module to communicate with the rest of the devices. They control the Transmit and Receive commands from the radiofrequency modules according to the commands sent by the host computer. They also process the information from the transponders received by the radiofrequency modules, decode them, check their validity, and convert the signals to the protocol used for communication with the host computer. The control modules used in conjunction with the radiofrequency modules described in the previous section are available using the RS232C, RS422, or RS422/485 interfaces depending on the needs of the user. These control modules also handle the different synchronization methods described in Section 5.4 that need to be followed when more than a single interrogator is being used.

During the *charge time*, the control module activates the transmitter in the radiofrequency module. As long as this is active, the antenna resonates at the oscillator frequency and generates a radiofrequency field around its vicinity. This field charges the capacitor incorporated in the transponders energizing them. The typical length of the charge time is 50 ms. During this time, the receiver is disabled and cannot read any data. The charge time is followed by the *read time*, as shown in Figure 5.23. This time typically lasts 20 ms, during which the transponder sends the information back to the interrogator.

© 2011 by Taylor and Francis Group, LLC

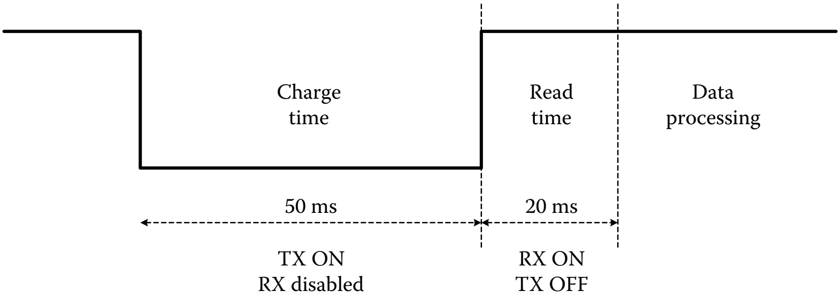

FIGURE 5.23
Charge and read times.

All the data formats consist of 128 bits. After being received by the radio-frequency module, they are checked by the control module. The different types of transponders supported by these interrogators are differentiated from each other by using different Start, Stop, and End bytes. The data structure for these transponders is shown in Figure 5.24. It is important to note that the first bit being transmitted is the least significant bit (LSB) and the last bit is the most significant bit (MSB).

The data structure for read-only transponders is shown in Figure 5.24(a). The data format starts with 16 pre-bits that are all set at zero, followed by the start byte ($7E). The next 64 bits are the identification data bits programmed into the transponder at the time of manufacturing; they cannot be changed by the user. The next 16 bits comprise the data protection bits (BCC). The following byte is the stop byte, which has the same structure as the start byte ($7E). Afterward, the transponder sends 16 end bits, all of them set to zero. During the last bit, the transponder discharges its capacitor, therefore deenergizing its internal circuits.

Figure 5.24(b) shows the data structure for read/write transponders. It also starts with the 16 pre-bits set at zero followed by the start byte that in this case is set at $FE. This tells the interrogator that the data come from a read/write transponder. The next 80 bits are the data bits that can be programmed by the user. The manufacturer recommends splitting these bits into 64 bits of data identification transmitted first and 16 bits for data protection bits (block check character, or BCC). The stop byte that is transmitted next has the same value as the start byte ($FE). The final 16 bits reflect the 16 LSBs of the read data that must be checked for validity by the control module. Once again, the last of these 16 bits makes the transponder discharge its capacitor and deenergize its internal circuits.

The data structure for the multipage read/write transponder is shown in Figure 5.24(c). The data start again with the 16 pre-bits set at zero, followed by the start byte with a value of $7E. This start byte is the same as for read-only transponders, meaning that the distinction between these two structures will be possible after the reception of the read address and read address frame

© 2011 by Taylor and Francis Group, LLC

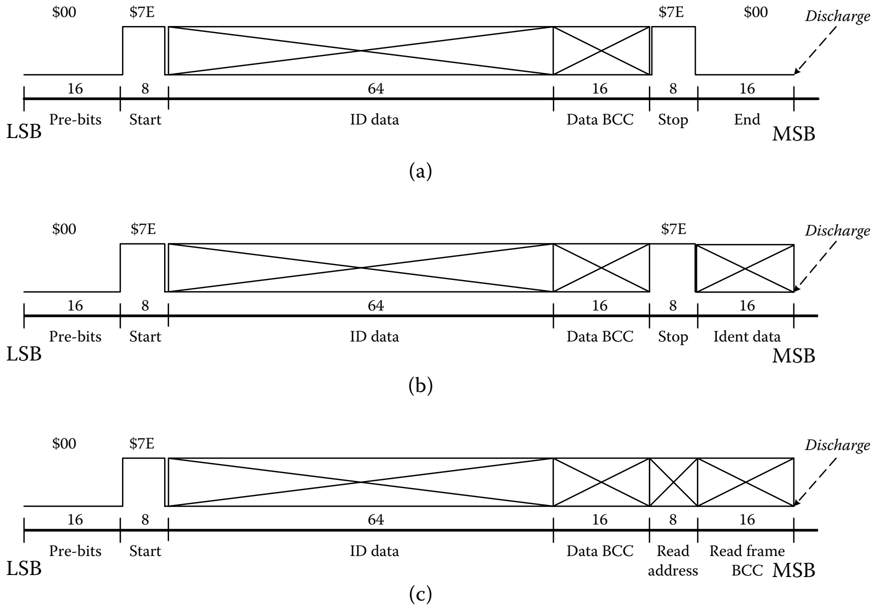
$00
$7E
$7E
$00
Discharge
16
8
64
16
8
16
LSB
Pre-bits
Start
ID data
Data BCC
Stop
End
MSB
(a)
$00
$7E
$7E
Discharge
16
8
64
16
8
16
LSB
Pre-bits
Start
ID data
Data BCC
Stop
Ident data
MSB
(b)
$00
$7E
Discharge
16
8
64
16
8
16
LSB
Pre-bits
Start
ID data
Data BCC
Read address
Read frame BCC
MSB
(c)

FIGURE 5.24
Data structure for several transponders: (a) read-only transponders, (b) read/write transponders, and (c) multipage read/write transponders.

BCC. The start byte is followed by 80 data bits that can be programmed by the user. The manufacturer also recommends splitting them into the first 64 data identification bits and 16 bits for protection data (BCC). The next 8 bits make up the read address byte; this contains a status field (2 bits), which is transmitted first, and a page field (6 bits). The 2 bits that make up the status field inform the control module about the function that the transponder has executed and the type of page that has been received. The final 16 bits make up the read frame BCC. One bit after having sent the first 128 bits, the transponder discharges the capacitor and deenergizes its internal circuits.

The function of the 16 pre-bits is to give enough time to the radiofrequency module to recover from the overload that it may have experienced during the 50 ms in which the transmitter has generated the radiofrequency signal. This is especially important because the transponders may start transmitting data immediately after having been energized. This period is known as the *waiting time*, lasting less than 1 ms.

After this time has elapsed, it is necessary to detect the first low-to-high transition for the start byte. After successful detection, the radiofrequency module synchronizes its clock and its data. The control module must also ensure that it receives all the high bits in the start byte. Otherwise, if a low bit is received, the process tops and the control module goes back to waiting for the detection of the low-to-high transitions. If the start byte has been successfully detected,

© 2011 by Taylor and Francis Group, LLC

the control module receives the next 80 bits and starts the generation of the cyclic redundancy check (CRC) to ensure the validity of the data. The last step in the decoding process is to determine whether the transponder that generated the data stream was a read-only transponder or a multipage read/write transponder as they both use the same start byte. This is done by checking the content of the CRC generator at the end of the process. If this value is equal to $0000, this indicates that a multipage read/write transponder originated the data. Otherwise, it indicates that it came from a read-only transponder. Figure 5.25 shows an overview of this sequence for the control module.

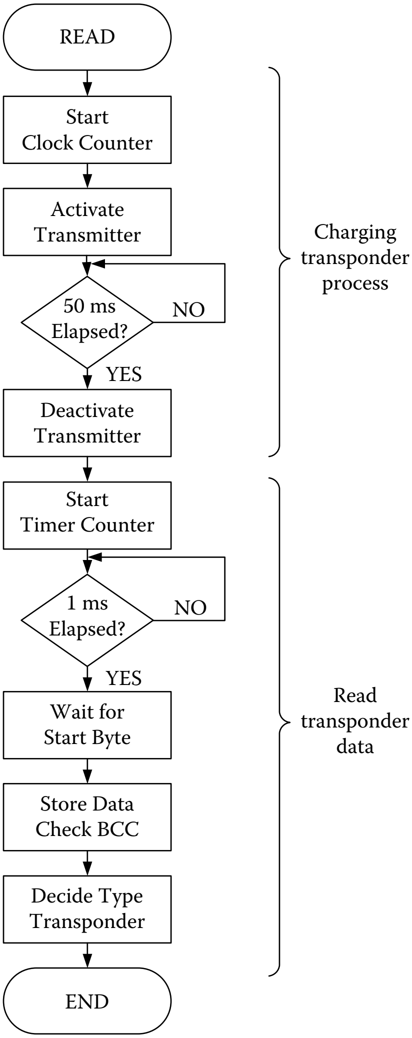
READ
Start Clock Counter
Activate Transmitter
50 ms Elapsed?
NO
YES
Deactivate Transmitter
Charging transponder process
Start Timer Counter
1 ms Elapsed?
NO
YES
Wait for Start Byte
Store Data Check BCC
Decide Type Transponder
Read transponder data
END

FIGURE 5.25
Control module sequence.

© 2011 by Taylor and Francis Group, LLC

The control module is also used to control the flow for writing to the transponder. Writing to the transponder is used to transfer commands, addresses, and data to the transponder. Before writing to the transponder, this must be energized. Therefore, writing to the transponder occurs after the transmitter has sent the 50 ms radiofrequency pulse to energize the circuits in the transponder. The actual process of writing is done by switching the transmitter off and on. To write a high bit, the transmitter is inactive during 1 ms followed by being active during 1 ms. To write a low bit, the transmitter is inactive during 0.3 ms followed by being active during 1.7 ms. The total time for writing a single bit is 2.0 ms.

To program the write data into the transponder's memory, the transmitter must be active for at least 15 ms immediately after the previously described write function has been executed. This is known as the *programming time*. Having the transmitter on for this time enables it to energize the circuits in the transponder. The write data format is shown in Figure 5.26.

All the data bits are transmitted to the transponder with the LSB first. The 8-bit write keyword ($BB) and the 8-bit write password ($EB) are sent to the transponder to initiate the transmission of the 80-bit data. The manufacturer recommends splitting the 80-bit data into 64-bit identification data transmitted first followed by 16 protection data bits (BCC). The write data format is terminated by a 16-bit write frame that must be set to $0300. After write frame, the transmitter will be active for 15 ms to allow the transponder to program the received data into its memory. The actual programming will only take place if the write keyword and write password bytes are received, the write data have the correct number of bits, and the radiofrequency field generated by the transmitter is high enough to produce the voltage required to program the transponder's memory. This voltage exceeds the voltage required to read the memory. Once after the programming time has elapsed, the transponder responds by sending the data that it has stored into its memory. This data follow the 128-bit format described previously in this section. If the read data received are structurally valid, they must be compared against the write data that were sent to the transponder to ensure that the correct data were written into the transponder. If the structure or the contents of the read data are not valid, the process must be repeated.

The last main function for the control module is to handle the synchronization tasks between different interrogators. For this to occur, a synchronization

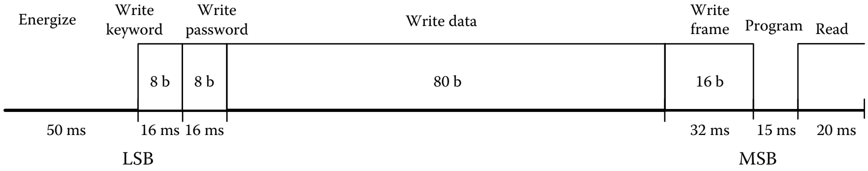

FIGURE 5.26
Format for writing data.

© 2011 by Taylor and Francis Group, LLC

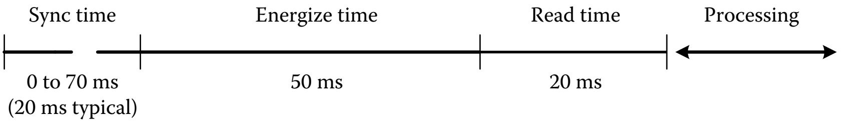

FIGURE 5.27
Synchronization sequence.

time must exist before the transmitter in the interrogator generates the radiofrequency signal to energize the transponders, as shown in Figure 5.27.

The synchronization allows the different interrogators in the overall system to detect each other and delay the activation of their own transmitters until other interrogators have finished their read functions. For this reason, the synchronization times vary depending on the number of interrogators in the application as well as the relative order of one interrogator in reference to the others in its vicinity.

5.5.2 Interrogators for HF RFID Systems

The Series Reader 6000 from Texas Instruments is an example of an RFID interrogator that operates in the HF range. The structure of the system is similar to the 2000 Series Reader described in the previous section. The Midrange Reader Module in this system provides the communication between transponder and interrogator. Output power levels can be adjusted between 100 mW and 1000 mW by means of a potentiometer placed on the radiofrequency module of the interrogator. The communication from the interrogator to the transponder is by ASK modulation, normally at 20%. The communication from the transponder to the interrogator is done by FSK modulation.

The TRF7960/1 also from Texas Instruments comprises multistandard fully integrated circuits containing the analog front end and the data-framing system. Their internal receiver enables them to modulate ASK and PSK, and it also includes automatic gain control and selectable bandwidth. In addition, the receiver includes a framing system allowing one to perform CRC checks, parity checks, or both. The output data are then sent to a microcontroller. The output of the transmitter only needs to be connected to the same impedance-matching circuit used for reception. Figure 5.28 shows the basic structure of the RFID system using this integrated circuit.

The receiver has two inputs that must be connected to an external filter. The external filter also converts the PSK modulation from the transponder into an amplitude-modulated signal. The two receiver inputs are multiplexed to two receiver channels: a main receiver and an auxiliary receiver. The main receiver contains a radiofrequency detection stage, a gain, a filter with automatic gain control, a digitizing stage, and an indicator for the strength of the received signal. The auxiliary receiver has a similar structure, although it is mainly used to measure the strength of the received signal.

© 2011 by Taylor and Francis Group, LLC

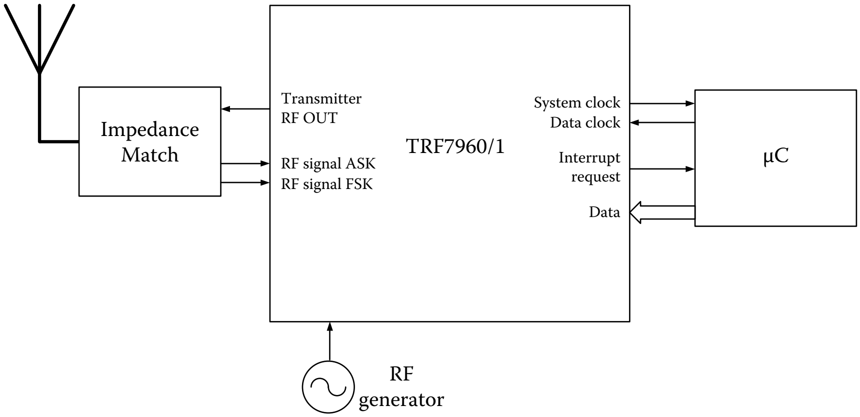

FIGURE 5.28
Structure of an interrogator using the TRF7960/1 circuits.

This is done by measuring the voltage of the demodulated signal and latching at its peak value.

The digital part of the receiver contains the bit decoders for the various protocols that supports the framing logic. The bit decoders convert the decoded signal into a bit stream. It has been specifically designed for maximum error tolerance, thus allowing the correct decoding of signals that have been partially corrupted due to noise or interference. The framing section packages the received bits into bytes, removing the control bytes sent by the transponder. This results in raw data that are sent to the microcontroller.

The transmitter contains the radiofrequency oscillator at 13.56 MHz, the processing circuits, and the radiofrequency power stage. The power levels available are 20 dBm (100 mW) or 23 dBm (200 mW) when the circuit is powered with 5V. When it is powered with 3 V, the available power levels are 15 dBm (33 mW) or 18 dBm (70 mW). The digital portion of the transmitter is very similar to that of the receiver as these circuits frame the command signals originated by the microcontroller.

5.5.3 Single-Chip Interrogators

In contrast to the complex systems described earlier, some interrogators have minimal components, normally an integrated circuit containing the transmitter and receiver and circuitry to communicate with a microcontroller or a microprocessor.

The core of the Series 6000 Reader from Texas Instruments is the RI-R6C-001 Integrated Circuit manufactured by the same company. This chip, whose block diagram is depicted in Figure 5.29, is a transceiver that provides the transmit and receive functions necessary for communication with transceivers operating at that frequency range.

© 2011 by Taylor and Francis Group, LLC

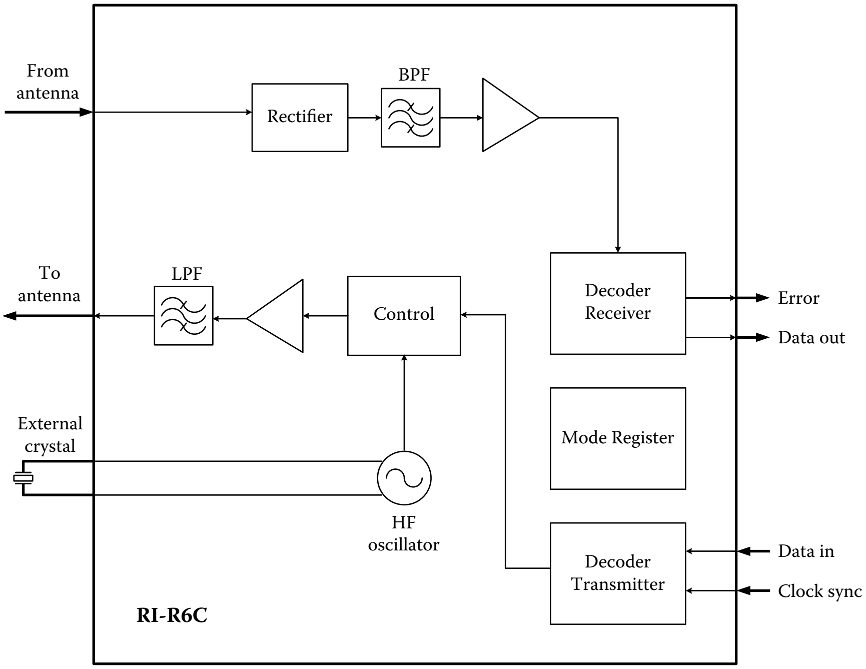
From antenna
Rectifier
BPF
To antenna
LPF
Control
Decoder Receiver
Error
Data out
Mode Register
External crystal
HF oscillator
Decoder Transmitter
Data in
Clock sync
RI-R6C

FIGURE 5.29
Simplified structure of the RI-R6C integrated circuit.

This transceiver integrated circuit communicates with its remote controller, normally a microprocessor, by a simple three-wire serial link. The communication between the transceiver and the transponders is generated by the microprocessor that sends the commands directly to the transceiver. The transceiver, in turn, applies the necessary carrier modulation and generates the radiofrequency signal.

The U2270B integrated circuit manufactured by Atmel® is another example of a single chip that serves as an interrogator. As shown in Figure 5.30, the U2270B incorporates an on-board power supply, an oscillator, tuning circuits, and other signal-processing circuits to produce an output that can be read by a microcontroller.

This specific chip has been designed to operate in the low-frequency range. It can be configured to work with different ranges of voltages for power supplies (from 5 V to 16 V). The low-pass filter in the receiver is a fourth-order Butterworth designed to remove the remaining energy of the carrier signal and other high frequencies after the demodulation process. The cutoff frequency can be selected by the user depending on the type of communication between the chip and the transponder that is being used. The possible choices are a Manchester code or biphase modulation typically at 5 kbaud.

© 2011 by Taylor and Francis Group, LLC

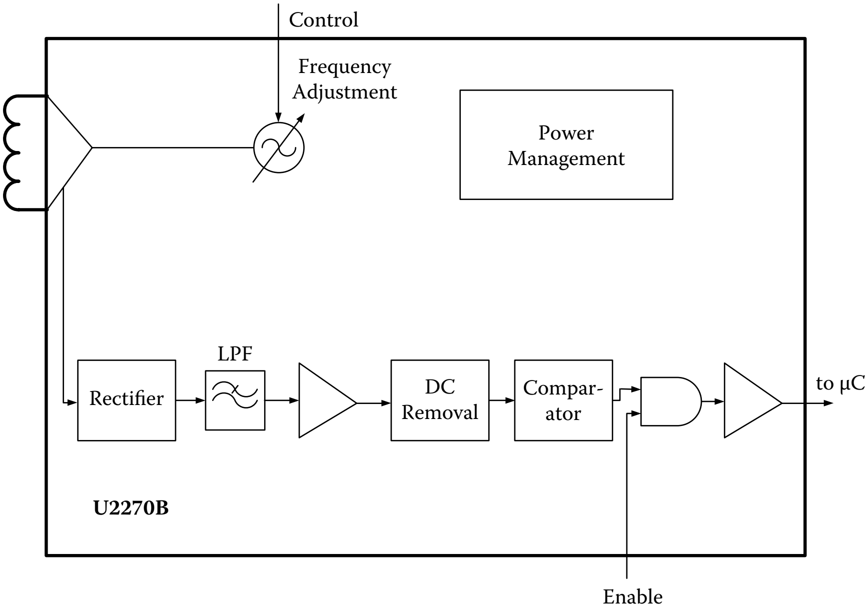
Control
Frequency
Adjustment
Power
Management
LPF
Rectifier
DC
Removal
Compar-
ator
to μC
U2270B
Enable

FIGURE 5.30
Simplified structure of the U2270B integrated circuit.

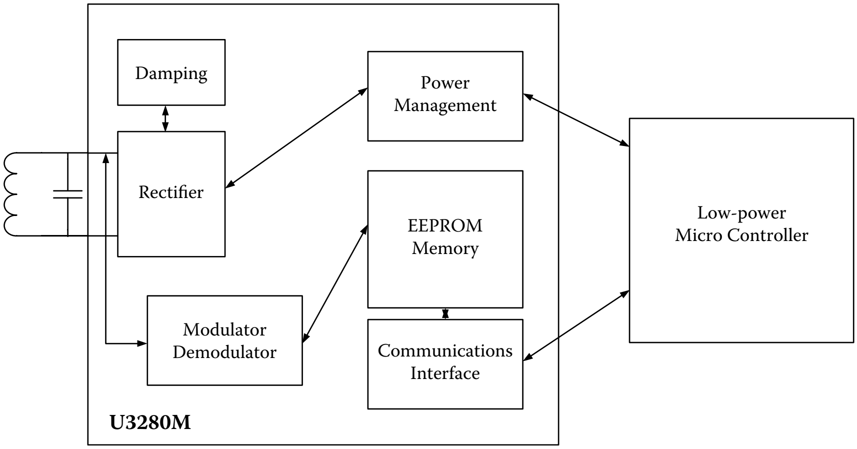
Damping
Rectifier
Power
Management
EEPROM
Memory
Modulator
Demodulator
Communications
Interface
Low-power
Micro Controller
U3280M

FIGURE 5.31
Simplified structure of the U3280M integrated circuit.

© 2011 by Taylor and Francis Group, LLC

One of the main applications of this chip is as an electronic immobilizer for the automotive industry in order to reduce auto theft. The interrogator and the associated microcontroller are located inside the vehicle, while the transponder with its unique identifier code is carried by the lawful owner or user in the vehicle. Unless the interrogator receives a valid code from a transponder, the microcontroller keeps the electronic ignition of the vehicle disabled. An additional benefit of this approach is that because the power to the transponder is supplied by the interrogator through the electromagnetic field it generates, the user of the vehicle does not need to worry about carrying fresh batteries like in traditional wireless fobs.

Atmel® also manufactures another chip, the U3280M shown in Figure 5.31, which, although marketed as a transponder interface for microcontrollers, functions as an interrogator. The chip can supply the energy required to power an external microcontroller and can also accept an external battery. The distribution of power is controlled by a power management circuit that either operates in automatic mode or is controlled by the microcontroller. The communication of this device with its microcontroller is carried out following a serial format. Modulation is carried out by changing the magnetic field generated by the antenna by varying its load. For example, a high logic level results in an increase of current through the antenna while the voltage becomes damped. The low logic level is generated by a decrease of current through the antenna that results in an increase in its voltage.

© 2011 by Taylor and Francis Group, LLC

6

Interrogator Communication and Control

CONTENTS

6.1 TIRIS Bus Protocol: Texas Instruments 155
6.2 ASCII Protocol for TIRIS Interrogators: Texas Instruments 159
6.3 Series 2000 Micro-Reader System: Texas Instruments 160
6.4 High-Frequency Interrogators: Texas Instruments........ 166
  6.4.1 TIRIS Protocol for Series 6000 Reader System Interrogators... 166
  6.4.2 Host Protocol for Tag-it™ Series 6000 Reader System Interrogators 166

Interrogators communicate with the host computers in order to transmit commands to the transponders and collect their information. This chapter explores some of the protocols used for linking interrogators with the host computers using as examples the different protocols employed by some interrogators manufactured by Texas Instruments. These examples show the basic communication processes between interrogators and their host computers as a stepping-stone to understand the control of interrogators from other manufacturers. The chapter starts by describing the TIRIS Bus Protocol, developed by Texas Instruments. It continues by studying how the ASCII protocol is used in some commercial interrogators. The next section explains the protocol used in some LF interrogators manufactured by Texas Instruments. The chapter ends by describing the communication protocols used in interrogators operating in the HF range.

6.1 TIRIS Bus Protocol: Texas Instruments

The TIRIS Bus Protocol is a proprietary protocol developed by Texas Instruments for communications between its interrogators or between interrogators and a host. This is a half-duplex byte-count protocol. It allows efficient communication between a single master and multiple slaves. The protocol allows communication with a single slave unit on a RS232 line and up to 31 slaves on a RS422 or RS485 line or using parallel links.

All the data transfers using the TIRIS protocol are initiated by the host; the slave units only send data upon request from the host unit, therefore

© 2011 by Taylor and Francis Group, LLC

eliminating the possibility of data collision on the bus. Furthermore, the protocol uses positive acknowledgment with the slave units indicating the reception of a message from the master. All the messages have the same format independently of the direction they travel. Table 6.1 shows the structure of an *N*-byte message using this protocol.

Command-message-codes describe the action that the master requires from the slave. Its most significant bit (MSB; Bit 7) is the *queued-response flag*, while the following seven bits contain the command code. When the queued-response flag is set to 1, it indicates to the destination unit that it should acknowledge and execute the command but its response should be queued until requested by the master. This is normally used with commands that have a delayed response in order to avoid suspension of the line activity. When the flag is set to zero, the response is immediate following the execution of the command.

TABLE 6.1

TIRIS Bus Protocol Format

Byte	Content
0	Start Mark[a]
1	Destination Address[b]
2	Source Address[c]
3	Message Code[d]
4	Data length
5	Data field (1)
6	Data field (2)
…	…
N + 4	Data field (N)
N + 5	MSB for CRC-field (1)
N + 6	CRC – field
N + 7	End Mark

[a] The *start mark* signifies the beginning of a message and is represented by $01. This corresponds to the ASCII character Start-of-Heading.

[b] The second byte, *destination address,* indicates the address of the unit to which the message is directed. The binary value of this byte corresponds to the ID of the destination address. The ID for the different units ranges from $00 to $FE, although the maximum number of units in a specific application is limited by the physical limitations of the link. Broadcast messages are indicated by using a destination address of $FF.

[c] The *source address* is the third byte in the message format. It indicates the ID of the unit sending the message.

[d] The fourth byte is the *message code* byte. This byte is used to define the content of the message. When the message is from the master to the slave, this byte is a *command-message-code*. When the message is from slave to master, this byte is a *response-message-code*.

© 2011 by Taylor and Francis Group, LLC

The response-message-code describes the acknowledgment that the slave unit has received the command. The four MSBs in the code are flags, while the four remaining least significant bits (LSBs) are the response code. The different flags are as follows:

The MSB (Bit 7) is the *error-flag*. When it is set to 1, it indicates that the slave has detected an error in the transmission or in the contents of the command. The error code is indicated in the last four bits.

Bit 6 is the *busy-flag*. When it is set to 1, it indicates that the slave is temporarily unable to accept commands.

Bit 5 is the *data-available-flag*. When it is set to 1, it indicates to the master that there is at least one message in the output queue in the slave. In order to read this message, the master must send the appropriate command.

Bit 4 is the *broadcast-received-flag*. When it is set to 1, it indicates that a broadcast message, issued previously, has been received correctly. The response to the broadcast command is placed in the queue. The master can read the response by issuing the appropriate commands at which point the flag is reset to 0.

Bits 5-8 in the Response-Message-Code byte indicate the type of response. Their meaning varies depending on the state of the Error-Flag bit as shown in Table 6.2.

Bit 5 in the TIRIS protocol is the *data length* byte. This byte shows the number of bytes that follow as data field. If the command or response does not require any data, this value will be zero.

The following N bytes in the TIRIS protocol make up the *data field,* with N being the value specified in the previous byte (Data Length byte). The content of this field depends on the message being sent.

The two bytes after the Data-Field bytes are *CRC-field* bytes. These bytes contain the Cyclic Redundancy Check values to ensure the integrity of the message. The TIRIS Bus protocol supports two methods of generating the CRC field. The first method is the Reverse CRC-CCITT algorithm that was described in Section 3.3.2. The second method supported by this protocol is called Longitudinal Redundancy Check (LRC). This is based on XORing each character with the previous value and storing the result in the LSB byte fin the CRC-Field. The MSB byte of the CRC-Field is generated by storing the 1′s complement of the LSB byte. Using this method, if the MSB and LSB are ANDed together, the result should be zero. Consequently, any value different from zero indicates an error in the received message. This method is not as secure as the Reverse CRC method but it is faster.

The final byte is the *end mark* byte that indicates the end of the message. This is represented by $04 that corresponds to the ASC II character *end of transmission*.

© 2011 by Taylor and Francis Group, LLC

TABLE 6.2

Response Codes

Bit	Error-Flag = 0 Process Response		Error-Flag = 1 Error-Response	
0	Command Complete	A non-queued command received and executed correctly	Transmission Error	Slave received a message with its ID but found error on data
1	Accepted	A queued response command received correctly. Response will be placed on the queue	Command Invalid	Message received correctly but command code is not valid for this slave
2	Queue Empty	No message in the queue	Error – Data field length	Data field length wrong. Command not executed
3	Nothing to Resend	No information to resend available	Error - Parameter	Invalid parameter. Command not executed

Note: The following *N* bytes make up the *data field,* with *N* being the value specified in the previous byte (data length byte). The content of this field depends on the message being sent.

Using a two-wire, half-duplex circuit is not possible to execute flow control. Therefore, all the units in the bus must be able to receive and buffer the maximum length of the message at the operating speed for the line. When a unit is unable to handle correctly the incoming data, it cannot send an error message until the original message has ended. At that point, it can transmit the appropriate error message.

The data flow for the commands that do not require queuing from the slaves is always controlled by the master. The first steps for the master unit are to enable the transmitter, send the command message, and disable the transmitter. The slave unit, which is always monitoring the line, receives the incoming message and checks for the correct destination address. If the destination address is incorrect, it ignores the message and returns to monitoring the line. Because in this case the slave does not give any response to the master, the master will resend the command.

If the destination address is correct, the slave unit checks for errors in the message. If it detects errors, it sends an error response to the master, disables its transmitter, and returns to monitoring the line. Upon reception of the error response, the master unit processes the error code and resends the original command. If no errors are detected, the slave unit processes the command, sends the appropriate response message, disables the transmitter, and returns to monitoring the line. Upon reception of the response message, the master processes it and proceeds to transmitting the next command.

Broadcast messages transmitted by the master are not acknowledged in the normal way. After the master has sent the broadcast command, the slave checks for errors in the received message. If it detects errors, it ignores the message, sets the error flag, and returns to monitoring the line. If it does not detect errors, it sets the broadcast-received flag, processes the command, and places the response in its queue. The master accesses the special status bit in

© 2011 by Taylor and Francis Group, LLC

the response-code field in the slave's response to evaluate the correct reception of the previous message.

Example 6.1:

Describe the operations taken by a master unit that wants to initiate an operation on all slaves simultaneously.

Solution:

The master unit sends a command with a destination ID of $FF. The master unit then waits the expected operation time and proceeds to poll each individual slave. The master is looking for the *broadcast-received-flag* that verifies the correct reception of the command. The absence of this flag in a slave unit triggers the master to broadcast the command again.

Queued messages use a slightly different format from the format used by broadcast or immediate commands. The *N*th data field byte contains a sequence number, given by the master, that is used to have a reference when the queued messages are requested from the slave units. The slave unit indicates that a command requiring a queued response has been received and is being processed by sending a response message to the master with the accepted flag set. This response is not transmitted directly by the slave unit, but it is placed on its queue for the master unit to access it when ready.

When interrogators are asked to queue a response, these responses are placed in a 30-position circular buffer. The buffer has two pointers: store-next-message and read-next-message. Because of the circular structure of the buffer, it is not necessary to delete responses as they are overwritten when the pointer is incremented.

6.2 ASCII Protocol for TIRIS Interrogators: Texas Instruments

Some of the interrogators from Texas Instruments accept commands using a simple ASCII protocol. This ASCII protocol can only be used with a RS232 or RS422 communications line. Once the interrogator reinitializes, the first character to be transmitted is $02 (Start of Text, or STX), followed by $0D (Carriage Return) and $0A (Line Feed) through the serial interface. Any characters that may have been transmitted before the STX character are treated as random characters generated by the start-up process and therefore discarded.

Using this protocol, the interrogator can operate on four different modes: execute, normal, line, and gate. In execute mode, the interrogator triggers a single command, while in the other modes the read function is continuously triggered. The interrogators are also able to distinguish between three types of transponders: read-only, read/write, and multipage. The serial communications protocol from the interrogator to the host computer is

© 2011 by Taylor and Francis Group, LLC

introduced as a single character that describes the type of interrogator that originated the message: R (read-only), W (read/write), or M (multipage).

After receiving data from a transponder, the interrogator performs a data check. The result of this data check is one of the following: *identification received correctly*, *identification invalid*, or *no identification detected*. After receiving a correct identification, the interrogator proceeds to send the information through the bus. This is done by placing the identification string in a buffer that is accessed, read, and cleared by the host computer. The TIRIS protocol can distinguish between two groups of transponder identification: animal code numbers and industrial (nonanimal) code numbers.

6.3 Series 2000 Micro-Reader System: Texas Instruments

The Series 2000 RFID interrogators operate in the LF range. Table 6.3 describes the structure of an *N*-byte message sent from the host computer to the interrogator in these systems.

The following set of bytes are the *data field* bytes. These bytes only exist if required by the command field.

TABLE 6.3

Series 2000 Host-to-Interrogator Message Format

Byte	Content
0	Start Mark[a]
1	Message length[b]
2	Command Field[c] (1)
3	Command Field (2) - not always present
4	Data field (1)
5	Data field (2)
…	…
N + 3	Data field (N)
N + 4	BCC

Note: If the second command field is not used, the length of the message is $N + 3$ bytes. In addition, messages are limited to a total of 41 bytes.

[a] The first byte is the *start mark* that indicates the start of a message.

[b] The second byte is the *message length* byte that indicates the length, in bytes, of the command and data fields that follow.

[c] The third and fourth bytes (when the second command field byte exists) are the *command field* bytes. These bytes define the mode of operation for the interrogator and determine the operation that the interrogator must carry out on the transponder. Depending on the type of command, the information in the data fields may or may not be transmitted to the transponder. The meaning of the different bits for the command field (1) is described in Table 6.4.

© 2011 by Taylor and Francis Group, LLC

TABLE 6.4

Command Field (I) Bit Description

Bit	Description	Value	Operation
0 and 1	Operation Mode	00	Perform single command
		01	Read in continuous Normal Mode
		10	Read in continuous Line Mode
		11	Send interrogator software version
2	FBCC Calculation	0 or 1	1 = interrogator calculates FBCC
3	Power burst I	0 or 1	1 = duration determined in data field
4	Power burst pause	0 or 1	1 = duration determined in data field
5	Power burst II	0 or 1	1 = duration determined in data field
6	Data	0 or 1	1 = determined in data field
7	Command Expansion Field	0 or 1	1 = Command Field (2) is used

TABLE 6.5

Command Field (2) Bit Description

Bit	Description	Value	Operation
0	Special write timing	0 or 1	1 = Must be determined in data field.
1	Wireless synchronization	0 or 1	1 = Wireless synchronization being used.
2	Database consistency check (DBCC) calculation	0 or 1	1 = Interrogator calculates DBCC.
3 to 7	Reserved	0 or 1	Reserved.

The following set of bytes are the *data field* bytes. These bytes only exist if required by the command field.

The last byte in the message is the *block check character* (BCC) byte. This is a one-byte value that results from XORing each one of the bytes in the preceding message with the exception of the start mark byte.

As shown in Table 6.4, the existence of command field (2) is set by bit 7 of command field (1). Table 6.5 shows the meaning of the bits in command field (2) when it is present in the message.

Example 6.2:

Calculate the BCC for a message containing the following hex bytes: 04 23 38.

Solution:

BCC is calculated by XORing the bytes of the message. Start by first XORing 04 and 23:

04	0000 0100
23	0010 0011
XOR	0010 0111

The resulting value (0010 0111) is now XORed with 38:

Previous XOR	0010 0111
38	0011 1000
XOR	0001 1111

Therefore, the resulting BCC is $1F.

© 2011 by Taylor and Francis Group, LLC

The following three examples illustrate the structure of messages between the host computer and the interrogator for a variety of different commands and applications.

Example 6.3:

The following is the structure of a message transmitted from the host computer to the interrogator in commanding the interrogator to read a read-only transponder:

Byte	Code	Content	Description
0	$01	Start Mark	Start Mark
1	$02	Length	Two bytes follow (excludes BCC)
2	$08	Command Field (1)	Perform single command – Power burst
3	$32	Data Field (1)	Power burst for 50 ms to energize transponder
4	$38	BCC	BCC over previous bytes excluding Start Mark

Example 6.4:

The following is the structure of a message transmitted from the host computer to the interrogator in commanding the interrogator to write the following data in a read/write transponder.

00 0A	FE	32	66	00	00	01
MSB						LSB

Byte	Code	Content	Description
0	$01	Start Mark	Start Mark
1	$11	Length	17 bytes follow (excludes BCC)
2	$E8	Command Field (1)	Perform single command, no FBCC calculations, send power burst to energize transponder for reading and writing. Data command on following field
3	$06	Command Field (2)	Use wireless synchronization
4	$32	Data Field (1)	Power burst to energize transponder for reading (50 ms)
5	$0F	Data Field (2)	Power burst to energize transponder for writing (15 ms)
6	$0C	Data Field (3)	12 data bytes follow
7	$BB	Data Field (4)	Write Keyword
8	$EB	Data Field (5)	Write Password
9	$01	Data Field (6)	LSB of data to write
10	$00	Data Field (7)	Data to write
11	$00	Data Field (8)	Data to write
12	$66	Data Field (9)	Data to write
13	$32	Data Field (10)	Data to write
14	$FE	Data Field (11)	Data to write

© 2011 by Taylor and Francis Group, LLC

Byte	Code	Content	Description
15	$0A	Data Field (12)	Data to write
16	$00	Data Field (13)	MSB of data to write
17	$00	Data Field (14)	Write frame
18	$03	Data Field (15)	Write frame
19	$9C	BCC	BCC over previous bytes excluding Start Mark

Example 6.5:

The following is the structure of a message transmitted from the host computer to the interrogator in commanding the interrogator to read page 2 of a multipage transponder.

Byte	Code	Content	Description
0	$01	Start Mark	Start Mark
1	$04	Length	Four bytes follow (excludes BCC)
2	$48	Command Field (1)	Perform single command – Power burst with data
3	$32	Data Field (1)	Power burst for 50 ms to energize transponder
4	$01	Data Field (2)	One data field follows
5	$08	Data Field (3)	Write address specifying general read of page 2
6	$77	BCC	BCC over previous bytes excluding Start Mark

The structure of the messages sent from the interrogator to the host computer is described in Table 6.6.

The bytes following BCC are the *data* field bytes.

TABLE 6.6

Series 2000 Interrogator-to-Host Message Format

Byte	Content
0	Start Mark[a]
1	Message length[b]
2	Status[c]
3	Data field (1) LSB byte
4	Data field (2)
…	…
N + 2	Data field (N) MSB byte
N + 3	BCC

[a] The first byte is the *start mark* that indicates the start of a message.

[b] The second byte is the *message length* byte that indicates the length, in bytes, of the command and data fields that follow.

[c] The third byte is the *status byte*. This provides feedback about the preceding read or programming operation according to Table 6.7.

© 2011 by Taylor and Francis Group, LLC

TABLE 6.7

Description of Bits in the Status Byte

Bit	Value	Operation
0 and 1	00	Read Only transponder
	01	Read/Write transponder
	10	Multipage transponder
	11	Other transponder
2	0 or 1	1 = Start byte was successfully detected
3	0 or 1	1 = DBCC was successfully performed
4	0 or 1	1 = DBCC was successfully performed
5	0 or 1	1 = Software version for interrogator follows
6–7	Reserved	

TABLE 6.8

Description of Data Field Bytes

Type of Response	Number of Bytes in Data Field	Description
Read Only	8	Identification data. LSB first
Read/Write	8	Identification data. LSB first
Multipage	9	Identification data. LSB first and Read Address
Other	14	Complete transponder protocol without pre-bits assuming a successful start byte was detected
No Read	0	No data fields, no start byte. Status is $03
Software version	1	Software version. For example $12 means software version 1.2

The last byte in the message is the *BCC* byte. This is a one-byte value that results from XORing each one of the bytes in the preceding message with the exception of the start mark byte.

The length and meaning of the data field bytes are shown in Table 6.8.

Example 6.6:

The following is the structure of a message transmitted from the interrogator to the host computer after a successful read of a read-only transponder.

Byte	Code	Content	Description
0	$01	Start Mark	Start Mark
1	$09	Length	Nine bytes follow (excludes BCC)
2	$0C	Status	Valid RO, Start byte detected, DBCC is correct
3	$6A	Data Field (1)	Identification data (LSB)
4	$58	Data Field (2)	Identification data

© 2011 by Taylor and Francis Group, LLC

Byte	Code	Content	Description
5	$4C	Data Field (3)	Identification data
6	$00	Data Field (4)	Identification data
7	$00	Data Field (5)	Identification data
8	$00	Data Field (6)	Identification data
9	$00	Data Field (7)	Identification data
10	$00	Data Field (8)	Identification data (MSB)
11	$7B	BCC	BCC over previous bytes excluding Start Mark

Example 6.7:

The following is the structure of a message transmitted from the interrogator to the host computer after an unsuccessful read attempt.

Byte	Code	Content	Description
0	$01	Start Mark	Start Mark
1	$01	Length	One byte follows (excludes BCC)
2	$03	Status	Other, no Start byte detected, DBCC is not correct
3	$02	BCC	BCC over previous bytes excluding Start Mark

Example 6.8:

The following is the structure of a message transmitted from the interrogator to the host computer after successfully programming page 2 of a multipage transponder.

Byte	Code	Content	Description
0	$01	Start Mark	Start Mark
1	$0A	Length	Ten bytes follow (excludes BCC)
2	$1E	Status	Valid Multipaage, Start byte detected, DBCC is correct
3	$47	Data Field (1)	Identification data (LSB)
4	$C6	Data Field (2)	Identification data
5	$2D	Data Field (3)	Identification data
6	$00	Data Field (4)	Identification data
7	$00	Data Field (5)	Identification data
8	$00	Data Field (6)	Identification data
9	$00	Data Field (7)	Identification data
10	$00	Data Field (8)	Identification data (MSB)
11	$09	Data Field (9)	Read address specifying successful programming of page 2
12	$7B	BCC	BCC over previous bytes excluding Start Mark

© 2011 by Taylor and Francis Group, LLC

6.4 High-Frequency Interrogators: Texas Instruments

6.4.1 TIRIS Protocol for Series 6000 Reader System Interrogators

The Series 6000 Reader System interrogators are RFID interrogators operating in the HF band, manufactured by Texas Instruments. These interrogators use a RS232 line to communicate with the host computer. The data packet from the host to the interrogator is known as the *request*, while the data from the interrogator to the host are known as the *response*. All communication sequences are initiated by the host that is the primary station. Using this protocol, the host waits for the response before continuing, thus assuring that the host does not have to handle large sequences of open requests. This is a binary, byte-count-oriented protocol in which the data length is passed as a parameter in the message.

Table 6.9 shows the structure of the messages used by these interrogators for the request message (host to interrogator).

The contents of byte 5 (command flags) are used to control the actions of the interrogator. The meanings of the different bits in this byte are shown in Table 6.10.

The BCC, used to ensure the integrity of the message, is a 16-bit code calculated on all the bytes of the packet including the start mark. The BCC contains two parts: its least significant byte is a longitudinal redundancy check (LCR) calculated by performing a cumulative XOR operation on all the bytes of the packet. The most significant byte in the BCC is the one's complement of the LCR.

The structure of the message sent from the interrogator to the host (response) is shown in Table 6.11. Most of these bytes are similar to the bytes in the request message.

The data field in response messages is limited to a maximum of 23 bytes. The contents of byte 5 (command flags) are used to report on the actions of the interrogator. The meanings of the different bits in this byte are shown in Table 6.12.

The different error codes that the interrogator transmits to the host computer when bit 4 of the command flag byte is set to 1 are described in Table 6.13.

The contents of byte 5 (command flags) are used to control the actions of the interrogator. The meanings of the different bits in this byte are the same as for the request message shown in Table 6.10.

6.4.2 Host Protocol for Tag-it™ Series 6000 Reader System Interrogators

This protocol defines the communications between the Tag-it™ Series 6000 Reader System Interrogators and the host computer. It handles the communication between the host and the interrogator, and carries out the requests

© 2011 by Taylor and Francis Group, LLC

TABLE 6.9

Series 6000 Host-to-Interrogator Message Format

Byte	Content
0	Start of Frame[a]
1	Message length[b] - LSB
2	Message length - MSB
3	Node address[c] - LSB
4	Node address - MSB
5	Command flags[d]
6	Command[e]
7	Data[f]
8	Data
…	Data
N+6	Data
N+7	BCC[g]
N+8	BCC

[a] The first byte is the *start mark* that indicates the start of a message. This byte is equal to $01.

[b] The second and third bytes are the *message length* bytes that indicate the length of the whole packet, including the start mark.

[c] Bytes 3 and 4 constitute the *node address*. Its value is $0000.

[d] Byte 5 is the *command flags* byte. It controls the actions taken by the interrogator. The value of this byte is dependent on the command being transmitted.

[e] Byte 6 is the *command* byte. It specifies the actions that must be taken by the interrogator. The value of this byte is dependent on the command being transmitted.

[f] The following *N* bytes are the *data* bytes. These contain the parameters and the data for the command.

[g] The last two bytes are the *BCC* bytes used to ensure the integrity of the preceding message. BCC is performed on the previous bytes, including the start mark.

TABLE 6.10

Bit Functionality in Command Flag Byte (Request)

Bit	Description
0	Reserved for future use
1	Reserved for future use
2	Reserved for future use
3	Reserved for future use
4	1 = Command to be performed only on transponders with an address that matches the data section of the message
5	Reserved for future use
6	Reserved for future use
7	Reserved for future use

© 2011 by Taylor and Francis Group, LLC

TABLE 6.11

Series 6000 Interrogator-to-Host Message Format

Byte	Content
0	Start of Frame
1	Message length - LSB
2	Message length - MSB
3	Node address - LSB
4	Node address - MSB
5	Command flags
6	Command
7	Data
8	Data
…	Data
N+6	Data
N+7	BCC
N+8	BCC

TABLE 6.12

Bit Functionality in Command Flag Byte (Response)

Bit	Description
0	Reserved for future use
1	Reserved for future use
2	Reserved for future use
3	Reserved for future use
4	Error flag. 1 = command was unsuccessful. Data section of response packet contains the code of the error
5	Reserved for future use
6	Reserved for future use
7	Reserved for future use

TABLE 6.13

Error Codes

Code	Description
$01	Transponder not found
$02	Command not supported
$03	Packet BCC invalid
$04	Packet flags invalid for command
$05	General write failure
$06	Write failure because of locked blocks
$07	Transponder does not support command
$0F	Undefined error

© 2011 by Taylor and Francis Group, LLC

for commands as well as their responses. Additionally, the protocol manages some aspects of the functionality of the reader and enhances the functionality of basic transponders by requesting the execution of compounded commands. This protocol is intended for point-to-point, half-duplex communications in which the host is the primary station. In some situations, however, the host may receive responses without sending a request first. This is a binary, byte-count-oriented protocol in which the data length is passed as a parameter in the message. Similar to the TIRIS protocol, the message from the host to the interrogator is called a *request*, while the message from the interrogator to the host is called a *response*.

The communication between host and interrogator is performed using a frame structure. The frame structure is essentially the same for requests and responses with the exception of the flag as they have different meanings. Requests use control flags, while responses use status flags. In this protocol, the length of the message data is limited to 4091 bytes. Table 6.14 describes the structure of the request and response frames.

In *pass-through-service* mode, the interrogator passes the information that it has received from the host to the transponder immediately without doing any other action. When the interrogator receives a response from the transponder, it also passes it directly to the host without further action. If the interrogator does not receive a response from the transponder in a specified period of time, it sends a message to the host with the error flag set and the

TABLE 6.14

Structure of Request and Response Frames

	Description	
Byte	**Request Frame**	**Response Frame**
1	Start Code[a]	Start Code
2	Data Length[b]	Data Length
3	Data Length	Data Length
4	Service Code[c]	Service Code
5	Control Flag	Status Flag
6	Message Data	Message Data
…	Message Data	Message Data
N-1	Message Data	Message Data
N+6	BCC	BCC
N+7	BCC	BCC

[a] The first byte contains the *start code*. This is the start frame delimiter using the value $D5.

[b] The next block is the two *data length* bytes that specify the number of data bytes that follow that block. The minimum length is 4 bytes (the service code, control flag, and two BCC bytes), while the maximum length is limited to 4095 bytes.

[c] The *service code* indicates the type of command that is being sent, as shown in Table 6.15.

© 2011 by Taylor and Francis Group, LLC

TABLE 6.15

Service Codes

Code	Description
$01	Pass-through Service
$02	Single Service
$03	Compound Service
$04	Ancillary Service

appropriate error code. Using pass-through service, the host directly controls the transponder.

In *single-service* mode, after receiving a request from the host, the interrogator analyzes and verifies it before formatting and sending a single command to the transponder. Also, when the interrogator receives a response from the transponder, it analyzes it, verifies it, formats it, and sends a unique response to the host.

The *compound-service* mode is used by the host to instruct the interrogator to perform specific actions on transponders using the transponder protocol. The difference between these actions and the actions from a single-service request is that in compound service, the interrogator performs complex operations involving multiple commands before returning a response to the host computer. After receiving a compound-service request, the interrogator also analyzes and verifies it before sending the appropriate commands to the transponder. The interrogator also analyzes and verifies all the responses received from the transponder before sending a single response to the host. This response, however, may consist of more than one frame depending on its size. Compound service increases the performance of the system because, for example, it takes less time to send a single command to read four blocks of memory and transmit the information to the host, rather than issuing four different reading commands while transmitting the intermediate information to the host after each one has been performed.

Ancillary-service mode is used to instruct the interrogator to perform certain actions that do not involve communication with the transponder. Examples of such actions are interrogator reset, read interrogator version, and perform interrogator diagnostics. They are also used to upload new software versions in the interrogator or to set up new communication parameters.

Table 6.16 lists the available service codes with the exception of the pass-through-service codes as they are defined by the transponder protocol. The different transponder protocols are described in Chapter 7. The different commands are described in further detail in the next sections of this chapter. Table 6.16 can also be used to understand the type of commands available for each service code.

For control flags, four of the bits in byte 5 are used to specify instructions to the interrogator or they carry information about the message, as shown in Table 6.17.

© 2011 by Taylor and Francis Group, LLC

TABLE 6.16

Commands Available for Different Service Codes

		Service Code		
Command	**Command Code**	**Single**	**Compound**	**Ancillary**
Tag Version	$03	$02		
Read Block	$12	$02		
Read Multiblock	$02	$02	$03	
Read Block SID	$FE		$03	
Read Multiblock SID	$FD		$03	
Write Block	$05	$02		
Write and Lock Block	$07	$02		
Write Multiblock	$06	$02	$03	
Lock Block	$08	$02		
Lock Multiblock	$09	$02	$03	
Repeat Last Request	$01			$04
Send Last Request	$02			$04
Stop Continuous	$04			$04
Start Synch	$05			$04
Reader Reset	$10			$04
Reader Version	$11			$04
Reader Diagnostic	$12			$04
Read Reader Setup	$13			$04
Start Flash Loader	$16			$04
Factory Lock Block	$3D	$02		
Write SID Code	$3E	$02		
Factory Programming Off	$3F	$02		

The next byte, byte 5, is the *control flags* byte. This byte has a different meaning depending on the direction of the message. For Request Frames (host-to-interrogator), byte 5 is called the *control flag byte*. For response frames (interrogator-to-host), byte 5 is called the *status flag byte*.

For status flags, four of the bits in byte 5 are used to provide information about the result of the request, as shown in Table 6.18.

Bit 0 (exception flag) indicates whether the request was successfully routed to the interrogator. When bit 0 is reset to 0, it indicates that no exception occurred and the request was completed satisfactorily. When bit 0 is set to 1, it indicates that an exception occurred. In this case, the error code that describes the exception can be found in the first byte of the message. These errors are shown in Table 6.19.

Bit 1 (more flag) informs the interrogator whether more data follow (bit 1 = 1). Bit 2 (emulation flag) informs the interrogator whether the transponder completed the request using a compound command to emulate a complex command (bit 2 = 1). Bit 3 (auto-repeat flag) informs the interrogator where

© 2011 by Taylor and Francis Group, LLC

TABLE 6.17

Bits for Control Flags

Code	Description
0 (LSB)	Reserved
1[a]	More Flag
2[b]	Emulation Flag
3[c]	Auto Repeat Flag
4[d]	BCC Flag
5 to 7	Reserved

[a] When bit 1 (more flag) is set to 1, it indicates to the interrogator that it should expect more frames until the host sends a frame in which bit 1 is reset to 0. When bit 1 is reset to 0, it indicates to the interrogator that there are no more data associated with the request either in a single message or at the end of a series of messages.

[b] Bit 2 (emulation flag) indicates the interrogator if it is allowed to emulate a complex request by sending multiple single requests. This results in substituting a complex command with a compound command. When bit 2 is set to 1, this means that if the command cannot be performed using the current transponder version, the interrogator will attempt to use a compound command. When bit 2 is reset to 0, the substitution is not allowed.

[c] Bit 3 (auto-repeat flag) specifies whether or not the interrogator should automatically repeat the execution of the request. When Bit 3 is set to 1, auto-repeat is on. In this case, the interrogator will execute the request, send the response it has received from the transponder back to the interrogator, wait for a specified amount of time, and send the request again. This process is repeated until it receives another instruction with bit 3 reset to 0.

[d] Bit 4 is the BCC flag. It specifies the type of block code check used to ensure the integrity of the message. When bit 4 is reset to 0, the message uses CRC-CCIT, while when bit 4 is set to 1, the message uses LRC. Although LRC is simpler to implement and requires less processing time than CRC, it does not offer as much protection as CRC.

TABLE 6.18

Bits for Status Flags

Code	Description
0 (LSB)	Exception Flag
1	More Flag
2	Emulation Flag
3	Auto Repeat Flag
4	BCC Flag
5 to 7	Reserved

© 2011 by Taylor and Francis Group, LLC

TABLE 6.19

Error Messages Associated to Exception Flag

Code	Description
$00	Reserved
$01	Request data corrupted and not executed
$02	Application not supported
$03	Data format error and request aborted
$04	Continuous mode not available for this request
$05	Reserved
$06	Reserved
$07	Reserved
$08	Reserved
$09	Reserved
$0A	Reserved
$0B	Reserved
$0C	Reserved
$0D	Reserved
$0E	Reserved
$0F	Undefined error and request aborted

TABLE 6.20

Structure of Single Service Messages

Length	Description
5 bytes	Frame
1 byte	Command
1 byte	Format
1 byte	Synch
≤ 4088 bytes	Message Data
2 bytes	Frame

the transponder performed the request once (bit 3 = 0) or continuously (bit 3 = 1). Bit 4 (BCC flag) specifies the method used to generate the BCC byte: CRC (bit 4 = 0) or LRC (bit 4 = 1).

The next byte or set of bytes are the *message data* bytes that carry the actual message. The actual format of the message depends on the type of command that was specified in the service code byte. In pass-through-service messages, the interrogator is transparent to the request from the host to the transponder. The communication from the host to the transponder uses the transponder protocol. Because this is a bit-oriented protocol, it may be necessary to add the necessary bits to the message to ensure that the whole frame falls within a byte-oriented protocol. These are called *pad bits*. The structure for single-service messages is shown in Table 6.20.

© 2011 by Taylor and Francis Group, LLC

TABLE 6.21

Format Byte for a Request Message

Bit	Name	Description
0 (LSB)	Reserved	
1	Code Extension	0 = No extension code 1 = Code extension used
2	Addressing	0 = Not addressed 1 = Addressed. Address contained in first 4 bytes of data field
3	Format	0 = Fixed format 1 = Variable format
4	Transmitter control	0= Turn transmitter off after executing command 1 = Leave transmitter on
5	Reserved	
6	Reserved	
7 (MSB)	Reserved	

TABLE 6.22

Format Byte for a Response Message

Bit	Name	Description
0 (LSB)	Error	0 = No error 1 = Error occurred. Status code is first byte of data field after any address information
1	Code Extension	0 = No extension code 1 = Code extension used
2	Addressing	0 = Not addressed 1 = Addressed. Address contained in first 4 bytes of data field
3	Format	0 = Fixed format 1 = Variable format
4	Reserved	
5	Reserved	
6	Reserved	
7 (MSB)	Reserved	

The structure for single-service messages is shown in Table 6.20.

The format byte contains information about the format and status of the message. The contents are different depending on whether the message is a request or a response. Table 6.21 shows the description of the format byte for a request message, while Table 6.22 shows the description of the format byte for a response message.

Table 6.23 shows the meaning of the synchronization codes for both request and response messages.

The structure for compound messages is shown in Table 6.24.

Tables 6.25 and 6.26 show the structure of the format byte for request messages and response messages, respectively.

© 2011 by Taylor and Francis Group, LLC

TABLE 6.23

Synch Byte for a Request or for a Response Message

Bit	Name	Description
0–2	Synchronization type	000 = No synchronization 001 = Master Slave synchronization (Master) 010 = Master Slave synchronization (Slave) 011 = Cascaded synchronization
3	Action on "No Transponder"	0 = All responses sent to host 1 = "No transponder" responses not sent to host
4	Alternate Operation during continuous mode	0 = All valid data responses sent to host 1 = Identical responses sent to host only once
5	Reserved	
6	Reserved	
7	Reserved	

TABLE 6.24

Structure of Compound Messages

Length	Description
5 bytes	Frame
1 byte	Command
1 byte	Format
1 byte	Synch
≤ 4088 bytes	Message Data
2 bytes	Frame

TABLE 6.25

Format Byte for a Request Compound Message

Bit	Name	Description
0 (LSB)	Reserved	
1	Reserved	
2	Addressing	0 = Not addressed 1 = Addressed. Address contained in first 4 bytes of data field
3	Format	0 = Fixed format 1 = Variable format
4	Transmitter control	0= Turn transmitter off after executing command 1 = Leave transmitter on
5	Reserved	
6	Reserved	
7 (MSB)	Reserved	

© 2011 by Taylor and Francis Group, LLC

TABLE 6.26

Format Byte for a Response Compound Message

Bit	Name	Description
0 (LSB)	Error	0 = No error 1 = Error occurred. Status code is first byte of data field after any address information
1	Reserved	
2	Addressing	0 = Not addressed 1 = Addressed. Address contained in first 4 bytes of data field
3	Format	0 = Fixed format 1 = Variable format
4	Reserved	
5	Reserved	
6	Reserved	
7 (MSB)	Reserved	

TABLE 6.27

Structure of Ancillary Messages

Length	Description
5 bytes	Frame
1 byte	Command
≤ 4090 bytes	Message Data
2 bytes	Frame

Note: Because this type of command does not involve communication with the transponder, it does not require the format or sync bytes present in the other types of messages. The last two bytes contain the BCC code of either CRC or LRC. This is performed over all the preceding data with the exception of the start-of-frame byte.

The description of the synchronization bit is the same as for single messages that was shown in Table 6.23. The structure of ancillary messages is shown in Table 6.27.

The method of flow control demands for each request to receive a response. The flags in the response message indicate the presence of an error or other exceptions. After a certain time without receiving a response, the request can time out. The control is done by the host that acts as the primary station. Data lengths are variable for each message, although the maximum data length in a single message is 4091 bytes.

© 2011 by Taylor and Francis Group, LLC

7

The Air Communication Link

CONTENTS

7.1 Communication Protocols ... 178
  7.1.1 Communication Protocols for Systems Operating in the LF and HF Ranges ... 178
  7.1.2 Communication Protocols for Systems Operating in the UHF Range ... 181
7.2 The Communications Link ... 184
  7.2.1 Elements of the Communications Link ... 184
  7.2.2 Forward Link ... 186
  7.2.3 Return Link ... 189
7.3 Transponder Operating Modes ... 191
  7.3.1 Transponders Operating in the LF and HF Bands ... 191
  7.3.2 Transponders Operating in the UHF Bands ... 193
7.4 Arbitration for Transponders ... 196
  7.4.1 Principles of Arbitration ... 196
  7.4.2 Principles of Anticollision ... 197
  7.4.3 Deterministic Anticollision Procedures ... 199
  7.4.4 Aloha-Based Anticollision Procedures ... 200

This chapter describes the communication methods used to transfer information from the interrogator to the transponder and from the transponder to the interrogator. The communication from interrogator to transponder is usually known as the *forward communications link*. This consists mostly of commands sent from the host computer to the transponder via the interrogator. It also contains the data to be written for those transponders whose memories can be rewritten. The communication from transponder to interrogator is known as the *return link* or the *backscattered communication*. This consists mostly of the transmission of the data stored in the transponder's memory.

While there are several standards that specify the critical elements of the communications protocol, the implementation of these standards especially varies not only among different manufacturers but also among different products from the same manufacturers. Therefore, this chapter does not pretend to be a comprehensive collection of all the air communication methods employed by RFID systems. Instead, this chapter explores the most common

© 2011 by Taylor and Francis Group, LLC

approaches to the communication between these elements of the RFID system using specific systems as general examples. This should give the readers the basic information while preparing them to search for more specific details if they so require.

This chapter starts by describing the different air communication protocols followed by the study of the elements that make up the forward and the return communications link. It continues by talking about the different modes of operation available to transponders. The last section in this chapter describes the different arbitration methods and the anticollision procedures that allow effective communication when there is more than a single transponder in the radiofrequency field.

7.1 Communication Protocols

7.1.1 Communication Protocols for Systems Operating in the LF and HF Ranges

The T5557 is a transponder with a 300-bit memory manufactured by Atmel® that operates in the low-frequency range and is similar to other transponders from different families of manufacturers. The communication between transponder and interrogators starts with the chip in the transponder being initialized after it has reached its threshold voltage. Afterward, the chip will be ready or will experience an additional delay of 67 ms depending on the status of one of its configuration bits. In read mode, the data from the memory are transmitted serially by modulating the load across the antenna terminals, starting with bit 1 from block 1. This process continues until the last bit from the specified number of blocks has been transmitted. The number of blocks to be read is defined by one of the bits in the transmitted command.

The transmission from the transponder to the interrogator uses amplitude-shift keying (ASK) modulation by switching a load connected to the antenna terminals on and off. Figure 7.1 shows the waveform across the load for different types of coding.

The transponder has the ability to insert a special pattern in order to synchronize the interrogator before transmitting the first block. This pattern consists of four bit periods in which the modulation is turned off and on, as shown in Figure 7.2.

In order to write data to the transponders, the interrogator uses on-off keying. It interrupts on and off the electromagnetic field generated as shown in Figure 7.3. The time between the gaps is used to encode the data to be transmitted: 24 periods of the RF field for a *0* and 54 periods of the RF field for a *1*. The process finishes after the transponder has detected that there is no gap

© 2011 by Taylor and Francis Group, LLC

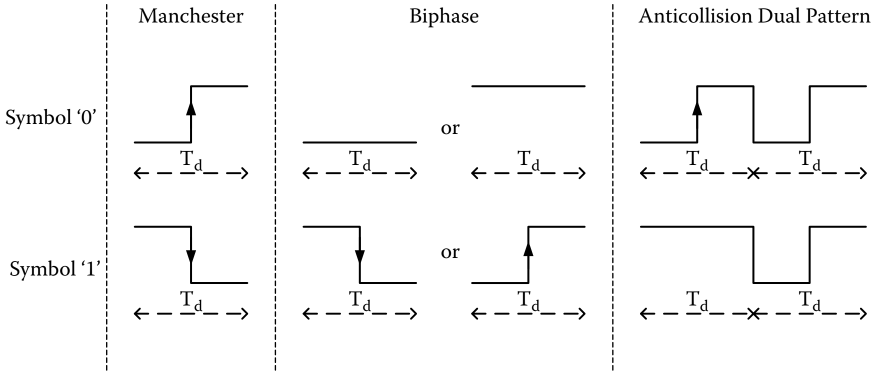
Manchester
Biphase
Anticollision Dual Pattern
Symbol '0'
Symbol '1'
or
or
T_d

FIGURE 7.1
Waveforms across load for different types of coding.

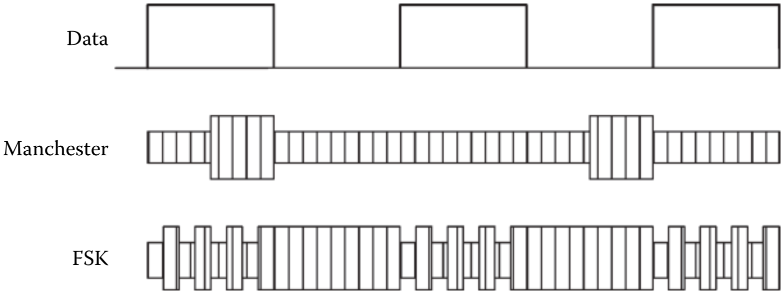
Data
Manchester
FSK

FIGURE 7.2
Synchronization patterns for Manchester and frequency-shift keying (FSK) modulations.

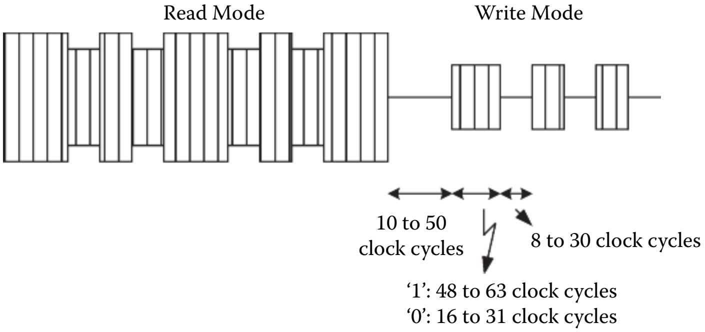
Read Mode
Write Mode
10 to 50 clock cycles
8 to 30 clock cycles
'1': 48 to 63 clock cycles
'0': 16 to 31 clock cycles

FIGURE 7.3
Gap sequences used in data transmission.

© 2011 by Taylor and Francis Group, LLC

for more than 64 periods. The initial gap triggers the communication from the interrogator to the transponder. This initial gap can be accepted at any time after the system has been initialized.

After the initial gap, the transponder expects to receive a valid command sequence. A standard write command needs the operation code, the lock bit, the 32 bits that make up the data, and the 3-bit address. Writing to a protected transponder requires the additional 32-bit password that is placed between the operation code and the data address bits. The password bits are compared bit by bit with the contents of the password stored in memory. If the bits do not match, the transponder returns to read mode. The complete writing process is shown in Figure 7.4.

After all the information has been correctly received, the transponder can be programmed. There is, however, a delay of one clock period between the end of the writing sequence and the start of the programming. The typical programming time is 5.6 ms, including the reading for data verification. Figure 7.5 displays the voltage measured across the antenna after programming a memory block.

For the Tag-it™ family of transponders manufactured by Texas Instruments, the communication between interrogator and transponder is done by amplitude modulation of the carrier. This allows the interrogator to generate enough energy to power the chips on board the transponders while simultaneously sending data. The encoding for the logic symbols *1* and *0* is done

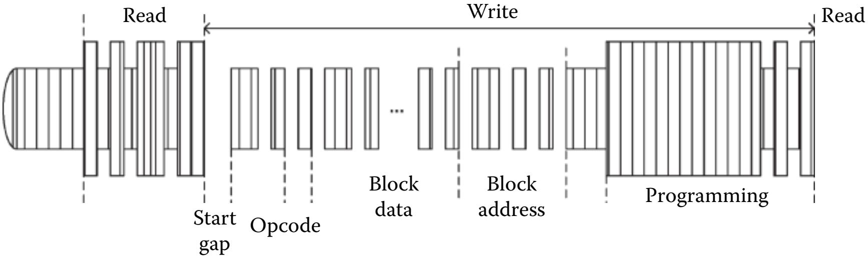

FIGURE 7.4
Writing sequence.

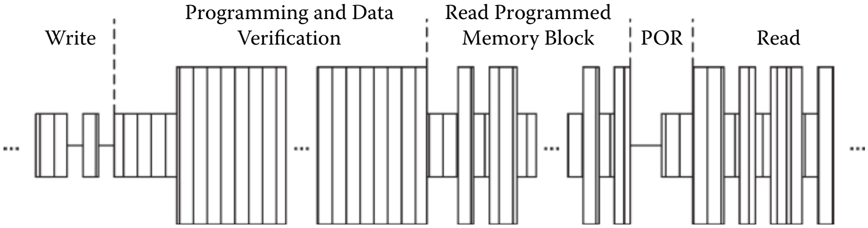
Write
Programming and Data Verification
Read Programmed Memory Block
POR
Read
...
...
...
...

FIGURE 7.5
Voltage across antenna terminals after programming a block.

© 2011 by Taylor and Francis Group, LLC

by pulse width modulation: *1* is encoded as a power burst lasting 150 µs, while *0* is encoded by power bursts lasting 100 µs. Data are framed between a start-of-frame (SOF) signal and end-of-frame (EOF) signal. SOF consists of a 50 µs burst followed by a standard duration pulse, terminating with a 150 µs pulse. The EOF signal consists of a 50 µs burst followed by a standard duration pulse, terminating with a 100 µs pulse.

The communication between the transponder and the interrogator uses the modulation of the antenna as it tunes and detunes the signal to generate the backscattered information. The signals being transmitted are encoded using a Manchester code, taking advantage of the clock frequency being embedded in the transmitted signal. With the load modulation, a Manchester high level is represented by a frequency of 423.75 kHz, while a Manchester low level is represented by a frequency of 484.29 kHz.

7.1.2 Communication Protocols for Systems Operating in the UHF Range

Transponders operating in the UHF range manufactured by Atmel® are based on the *interrogator talks first* (ITF) approach to establish communication with the interrogator. The communication always starts with a forward message, that is, a message from the interrogator to the transponder. After receiving this initial message, the transponder responds by backscattering a message to the interrogator. This is known as the *return link*. This backscatter communication is done using a binary phase-shift keying (2PSK) modulation based on changing the imaginary part of the input impedance in the transponder as a function of the message between sent.

The communication is based on a system of frames, as shown in Figure 7.6. The first frame is the header frame that is used to adjust the timing. The second frame is the data frame that is used to transport the data. The length of this frame depends on the command sent and the addressing mechanism being used. The last frame is the end-of-transmission (EOT) frame that marks the end of the transmission. Once the transmission has finished, the interrogator enters into listening mode after having received another header frame.

While the interrogator is transmitting data, the transponder remains in listening mode. During the transmission of the interrogator header frame, it is possible for the transponder to operate in full duplex. It transmits data back to the interrogator in a loop until the EOF frame is sent.

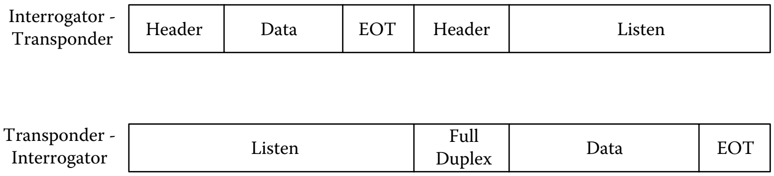

FIGURE 7.6
Frame communications during default transmissions.

© 2011 by Taylor and Francis Group, LLC

The communication between transponder and interrogator must be reliable even in the noisy environments in which these systems normally operate. To improve the reliability of the communication link, the interrogator transmits periodic clock ticks to the transponders, ensuring reliability because it helps to block other radiofrequency sources. In addition, the link between transponder and interrogator operates synchronously, and because the interrogator controls the speed and modulation ticks, the resulting transmission has a higher signal-to-noise ratio than an asynchronous communication.

The general structure of the frames is different between the forward or return link as well as the type of commands that it carries. All the frames start with a header frame. The data frame in the forward link depends on the type of command carried, as shown in Figure 7.7, which shows frames for different types of commands. Data frames contain the command and its cyclic redundancy check (CRC), data-setting parameters, address information, and data for programming. For example, short commands consist of 6 bits and its 2-bit CRC, while long commands consist at least of the command fame, the CRC, and the parameter field. In the transmission of a long command, these frames are followed by a CRC frame of 16 bits. The transmission ends with an end-of-transmission frame that is made of two EOF symbols.

Figure 7.8 shows the general structure for the return link frames. The communication link starts with a header frame that is used to obtain reference symbols, adjust modulation, and set the timing for the EOF characters. This is followed by a data frame, a CRC frame 16 bits long, and an

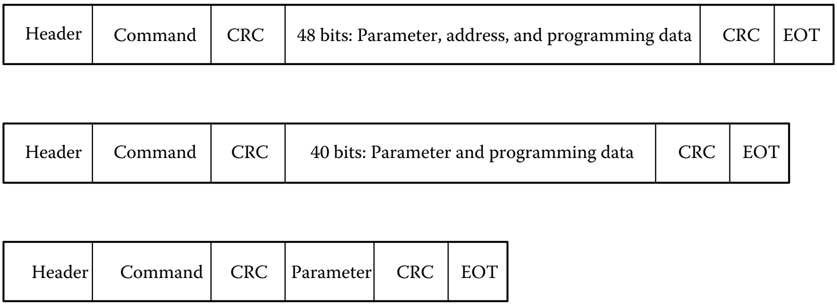
Header
Command
CRC
48 bits: Parameter, address, and programming data
CRC
EOT
Header
Command
CRC
40 bits: Parameter and programming data
CRC
EOT
Header
Command
CRC
Parameter
CRC
EOT

FIGURE 7.7
Examples of data frames in a forward link.

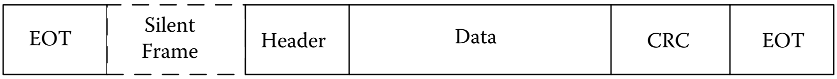
EOT
Silent Frame
Header
Data
CRC
EOT

FIGURE 7.8
General structure for return link frames.

© 2011 by Taylor and Francis Group, LLC

end-of-transmission frame. The data frame can have a length of 8 or 16 bits depending on the type of data carried. The silent frame shown in Figure 7.8 is only used during programming. During the silent frame period, the interrogator only transmits a continuous wave. The length of this period depends on the strength of the electromagnetic field.

The general structure of the communications flow for these transponders is shown in Figure 7.9. The transponder is continuously waiting for a notch indicating the presence of the external field. After it has been received, it will check for a valid header frame that indicates the beginning of a transmission. After the header frame has been received correctly, the transponder performs several additional verifications while transmitting the different parts of the message, including checking for the anticollision requests. The process ends after receiving the EOF data.

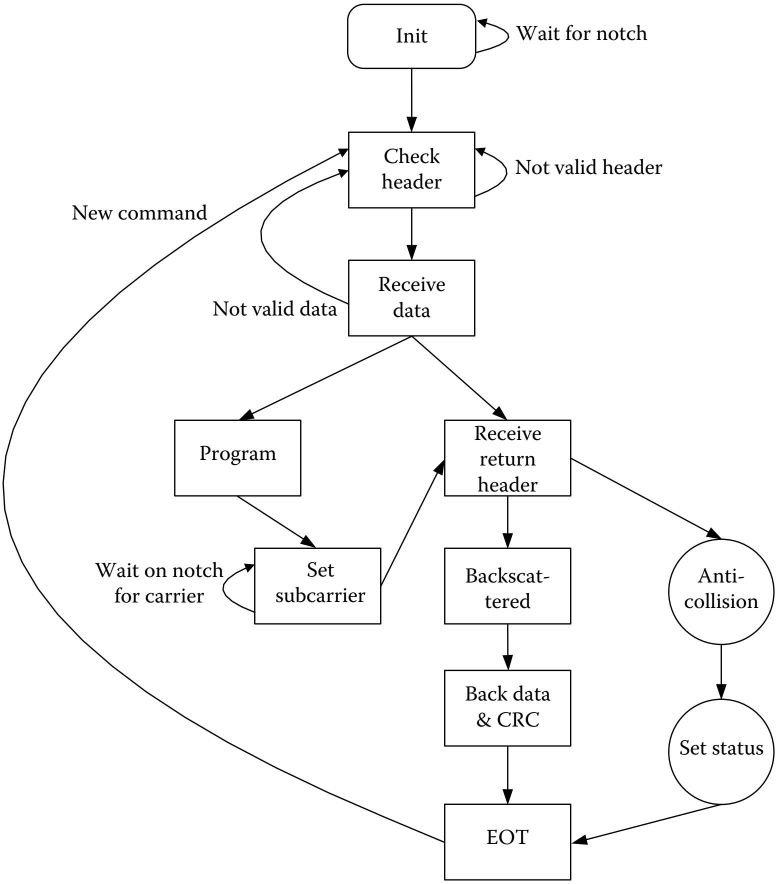
Init
Wait for notch
Check header
Not valid header
New command
Receive data
Not valid data
Program
Receive return header
Wait on notch for carrier
Set subcarrier
Backscat-tered
Anti-collision
Back data & CRC
Set status
EOT

FIGURE 7.9
General structure of communications flow.

© 2011 by Taylor and Francis Group, LLC

7.2 The Communications Link

7.2.1 Elements of the Communications Link

Because the transponder is powered by the electromagnetic field generated by the interrogator, it must measure the strength of this field continuously. If the strength of the field decreases in such a way that causes the voltage after the rectifier to be lower than a certain threshold, the circuits in the transponder activate an internal reset. For voltages higher than this threshold, the oscillator and the rest of the circuits in the transponder circuit are enabled. Figure 7.10 shows the time events between the different signals involved after the transponder is turned on for some transponders manufactured by Atmel®.

The duration of the power on reset (POR) period depends on how fast the signal ramps up. This in turn depends on the strength of the electromagnetic field. After the POR has executed, there is an additional waiting period called the *band gap* that is needed for stabilization of the rest of the internal circuits before the signal to enable the oscillator switches its state. Finally, there is a delay of 64 clock cycles (approximately 160 μs) before the oscillator in the transponder achieves operational status.

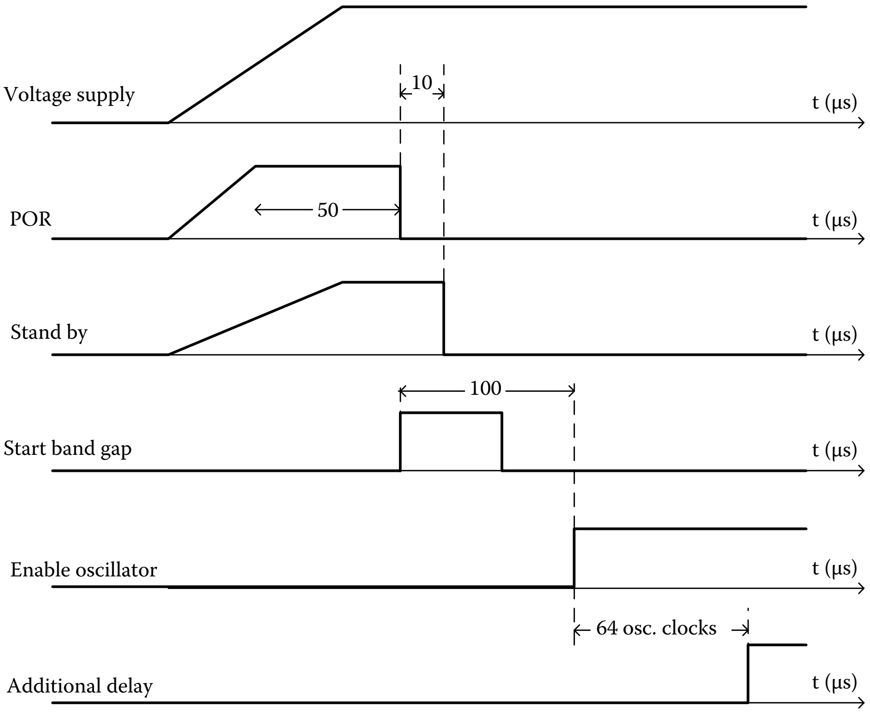

FIGURE 7.10
Chronogram of signals during power up.

© 2011 by Taylor and Francis Group, LLC

When the electromagnetic field decreases, the voltage after the internal rectifier also decreases. Once it reaches a certain threshold level, the oscillator stops. This results in lowering the current required by the circuit that in turn results in the voltage supply decreasing with a lower slope as shown in Figure 7.11. Finally, if the voltage decreases below another threshold, the circuit executes a reset. The difference between the two voltage levels allows for taking into account variations in the strength of the field without having to execute a reset.

The communication between transponder and interrogator normally happens in a noisy environment. Therefore, the designers of the system must ensure bit error rates as low as possible. This is achieved by using a combination of several techniques including generating notches in the system clock, making the return link synchronous to notch signals generated by the interrogator, using different coding techniques, and using CRC to protect the integrity of the data. CRCs for commands are 2 bits long, while CRCs for data are 16 bits long. Figure 7.12 shows the concept of notch timing, which results from generating notches at periodic intervals in the signal generated. For a typical value of signal level between 450 mV and 600 mV, T_1 lasts 4 µs and T_{notch} lasts 2 µs.

Other techniques include using *pulse interval encoding* to define the symbols *1* and *0*, whose timing and timing differences are controlled by the interrogator as an additional element to increase the robustness of the link. The fact that the return link is synchronous to the interrogator allows it to close the communication as soon as it detects a transmission error.

The communication from the interrogator to the transponder is known as the *forward link*, while the communication from the transponder back to the interrogator is known as the *return link*. The forward link carries the different type of commands to the transponder, while the return link carries the data from the transponder to the interrogator. These data will only be available for those commands that require the return of data. Because the

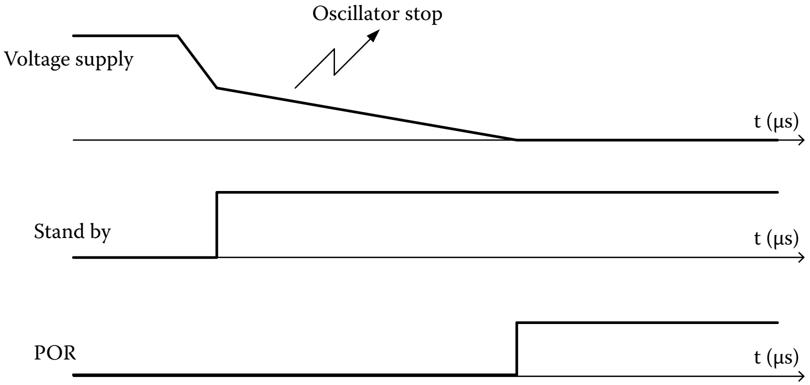

FIGURE 7.11
Chronogram of signals during power down.

© 2011 by Taylor and Francis Group, LLC

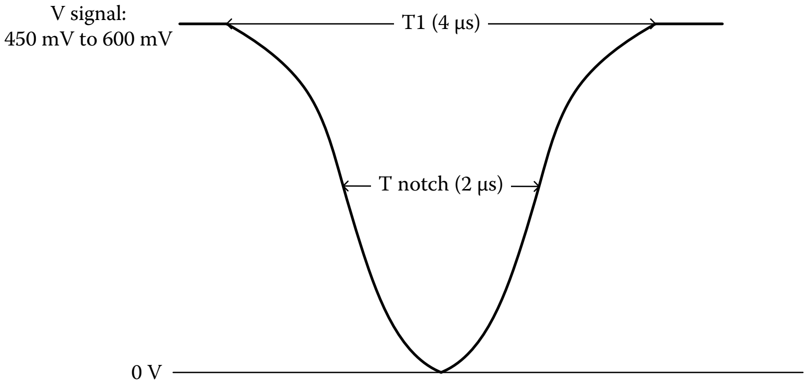

FIGURE 7.12
Notch timing chronogram.

handover mechanism from the forward to the return link happens synchronously, there is no need to wait for a cycle between them. Once the forward link sends its last frame that consists of an end-of-transmission message, the transponder is ready to transmit its return link header.

Texas Instruments RFID systems operating at low frequency use a frequency of 134.2 kHz to encode a low bit and a frequency of 123.2 kHz to encode a high bit. Because each bit uses 16 radiofrequency cycles to transmit, the duration of a low bit is different from the duration of a high bit. The low bit needs 119.9 µs to transmit, while a high bit needs 129.2 µs. The interrogator sends information to the transponder to transfer commands, addresses, or data. The writing process is always initiated by turning the radiofrequency transmitter on for a period of time of up to 50 ms. Afterward, the interrogator starts sending the bits that make up the command and associated data. Write bits last 2 ms. The differentiation between high and low bits is based on the duty cycle for each one of these bits. As shown in Figure 7.13, a high bit has a duty cycle approximately equal to 50%, while the low bit has a duty cycle approximately equal to 20%. The program phase consists of a continuous radiofrequency field for the duration of this period.

7.2.2 Forward Link

The forward header in most Atmel® transponders has four symbols, as shown in Figure 7.14. This header is accepted by the transponder if the length of the symbols *check1* and *EOF** is larger than the length of the symbol *0**.

Figure 7.15 shows that the information being transmitted is encoded in the timing between pulses. After the frame has been accepted and the

© 2011 by Taylor and Francis Group, LLC

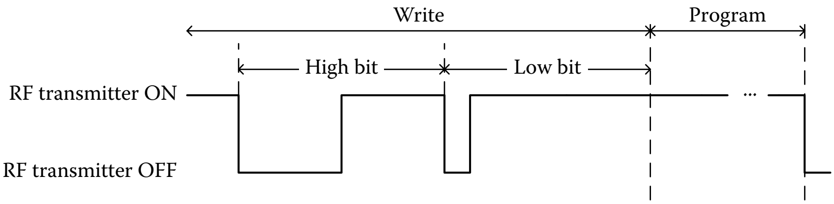
Write
Program
High bit
Low bit
RF transmitter ON
RF transmitter OFF
...

FIGURE 7.13
Write and program signals.

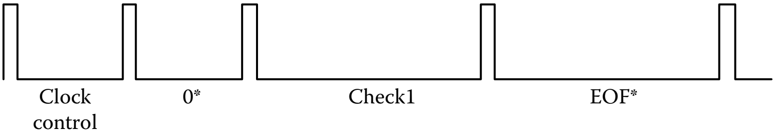
Clock control
0*
Check1
EOF*

FIGURE 7.14
The four symbols in the forward header.

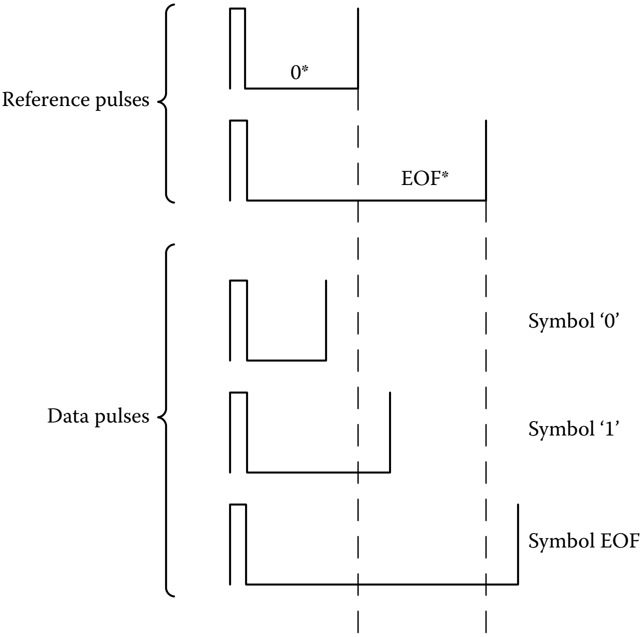
Reference pulses
0*
EOF*
Data pulses
Symbol '0'
Symbol '1'
Symbol EOF

FIGURE 7.15
Pulse timing used to encode information.

© 2011 by Taylor and Francis Group, LLC

interrogator transmits data, the transponder measures the time between two pulses or notch signals. This is used by the transponder to decide whether the received symbol is a *0* or a *1* by comparing its length with the length of the symbols *0** and *EOF**, as shown in Figure 7.15.

This figure shows that a symbol will be interpreted as a *0* if its timing is shorter than the time that was sent in the Header for the symbol *0**. A symbol will be interpreted as a *1* if the time used by that symbol is longer than for the symbol *0** but shorter than the time used in the header for the symbol *EOF**. Finally, EOF will be interpreted for a symbol that lasts longer than the *EOF** symbol. This approach increases the reliability of the link as the transponder does not have to measure absolute but relative time intervals.

The communication from the interrogator to the transponder for the Series 2000 Texas Instruments RFID systems is based on transmitting bursts of energy of specific duration at specific intervals. For example, the commands used to read read-only or read/write transponders consist of an initial power burst lasting 50 ms followed by a period of 20 ms in which the radiofrequency field is turned off. In fact, all these transponders use the 50 ms power burst as the initial identifier for an upcoming command in addition to providing the energy to power the transponder. The following are some examples of formats for the different commands supported by these transponders. Figure 7.16 shows the protocol to transmit the command used to program a rewritable transponder.

The command starts with the 50 ms power burst followed by the specific command keyword and the password if necessary. These are followed by the 80 data bits used by this family of transponders and the 16 bits that make up the write frame. The last information being sent is the PBII, after which the transponder will read the new data stored in its memory to be verified by the interrogator. Figures 7.17 and 7.18 show the communication protocols for the commands *lock page* (of multipage transponders) and *selective read page* (of multipage transponders), respectively.

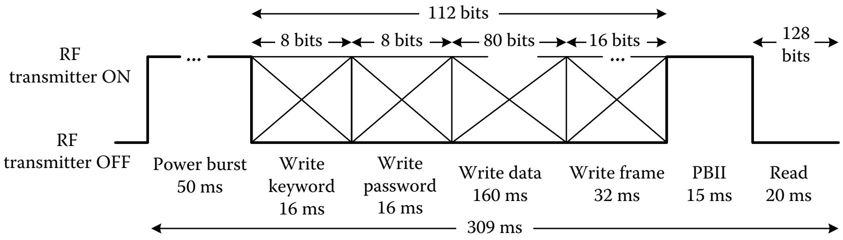

FIGURE 7.16
Protocol used to transmit a command to program a rewritable transponder.

© 2011 by Taylor and Francis Group, LLC

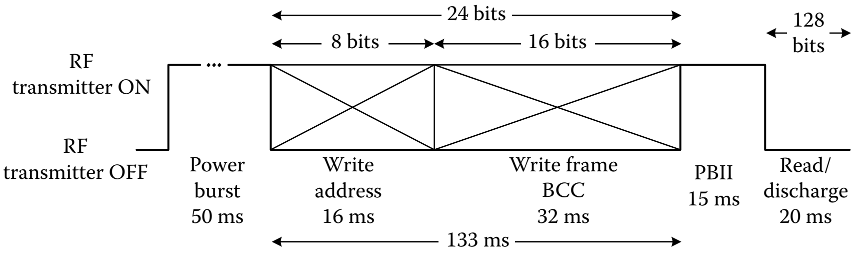
24 bits
8 bits
16 bits
128 bits
RF transmitter ON
RF transmitter OFF
...
Power burst 50 ms
Write address 16 ms
Write frame BCC 32 ms
PBII 15 ms
Read/ discharge 20 ms
133 ms

FIGURE 7.17
Protocol for the *lock page* command.

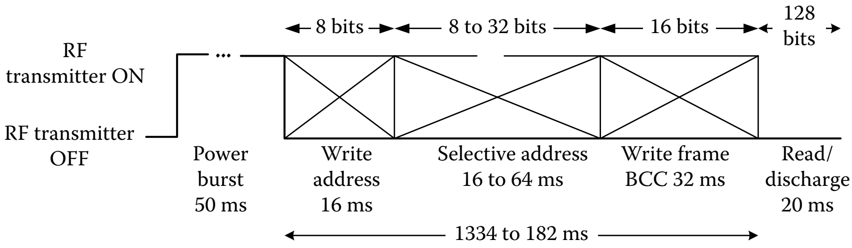
8 bits
8 to 32 bits
16 bits
128 bits
RF transmitter ON
RF transmitter OFF
...
Power burst 50 ms
Write address 16 ms
Selective address 16 to 64 ms
Write frame BCC 32 ms
Read/ discharge 20 ms
1334 to 182 ms

FIGURE 7.18
Protocol for the *selective read page* command.

7.2.3 Return Link

The return link in the majority of Atmel® transponders uses synchronous communication and angular modulation (binary PSK) in order to increase the signal-to-noise ratio and reduce the bit error rates because the return signals have extremely low power values. The synchronicity occurs as the transponder transmits its signal between two notches from the signal generated by the interrogator.

The return link starts with the return link header that is used to provide the timing information for the signal similar to the forward link header. Figure 7.19 shows the structure of the return link header for different types of coding styles used for the communication. In any case, the return link header always contains four symbols. The function and structure of each one of these symbols are shown in Table 7.1.

The return link header waits for a notch in the radiofrequency field generated by the interrogator before starting the generation of the reference timing. The data in the return link can be encoded using any of these coding techniques: bit stream (raw data), notch-locked non-return-to-zero-inverted (NRZI) encoding, soft-locked NRZI, FM0, and 3phase1. The EOT frame contains two EOF symbols.

© 2011 by Taylor and Francis Group, LLC

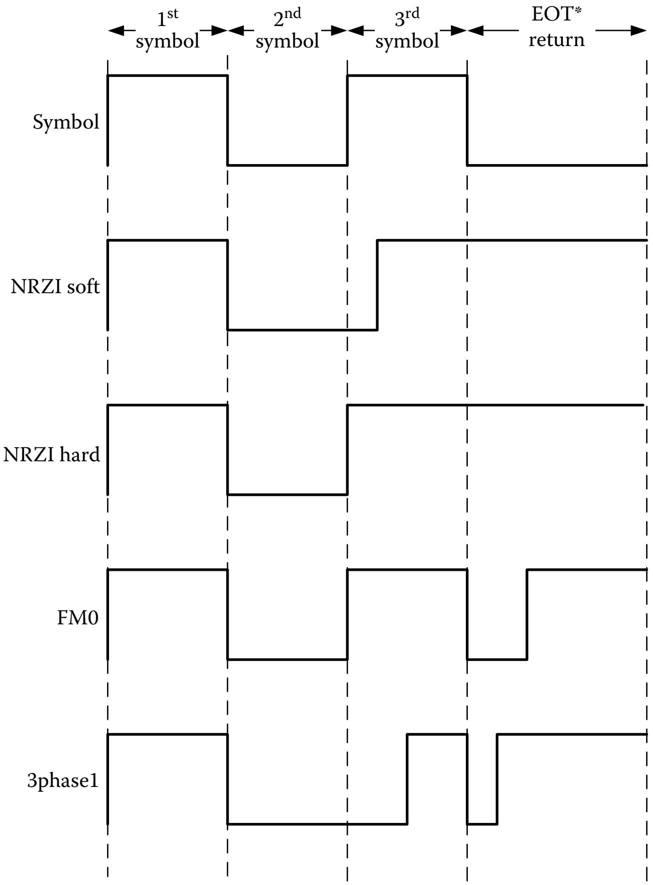

FIGURE 7.19
Structure of the return link header for different coding styles.

TABLE 7.1
Function of the Symbols in the Return Link Header

Symbol	Function	Additional Comments
1st	Main timing and power management	Used by interrogator to affect transmission rate of return link
2nd	Timing Reference	Used for anticollision control
3rd	Same timing as symbol 4	When using NIRZ, modulation is on
4th	Timing reference for EOF detection	

Figure 7.20 shows the transmission protocol used by the transponder to send information back to the interrogator as its response to a Read Data command for Texas Instruments transponders using their TIRIS protocol. All parts of this message are transmitted, with the least significant bit (LSB) being the first bit transmitted.

© 2011 by Taylor and Francis Group, LLC

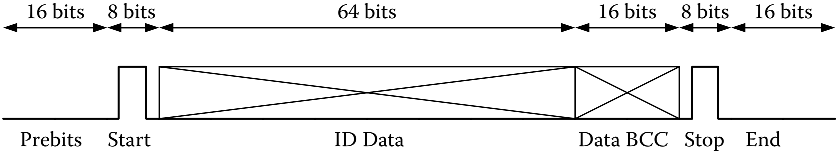

FIGURE 7.20
Transponder data sent back to the interrogator.

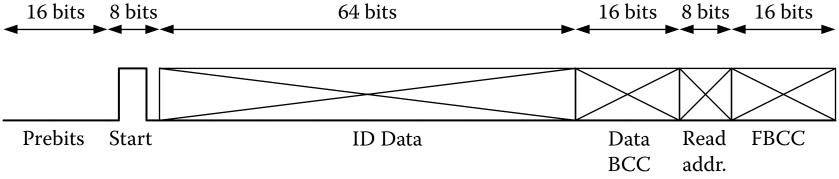

FIGURE 7.21
Return link for a multipage transponder.

After transmitting the pre-bits and eight bits that specify the start sequence, the transponder sends the 80 bits stored in its memory. The 64 first bits are the data bits, while the following 16 bits are the block check code (BCC) used to ensure the integrity of the data. These are followed by the stop and end bits. In order to be ready for a new activation, the transponder finishes the process by discharging its internal capacitor at the end of the transmission. The same protocol is used for rewritable transponders. Figure 7.21 shows the return link when accessing a multipage transponder.

In this case, following the BCC data, the transponder transmits the read address. This consists of a 2-bit status field that provides information about the function that that transponder has executed and a 6-bit page field that provides information about the page that was affected by the execution. The data format ends with reading the frame block check character (FBCC).

7.3 Transponder Operating Modes

7.3.1 Transponders Operating in the LF and HF Bands

After the transponders are initialized, they can enter into two different states, *interrogator talks first* (ITF) or *public mode* or *electronic surveillance*, depending on the status of their operating mode flags. In ITF mode, the data stream received through modulation from the interrogator

© 2011 by Taylor and Francis Group, LLC

undergoes different verifications. If they all result in a valid status, the transponder is synchronized with the data stream. At that point, the transponder is ready for communication. This communication always starts with a forward message, which is a message from the interrogator to the transponder. This communication method is used in multitransponder applications due to its anticollision capabilities. In this mode of operation, the transponder can be in any of the following states: *power down*, *ready*, *selected*, or *quiet*.

When the transponder is not able to harvest enough energy from the surrounding electromagnetic field to energize its internal chip, it enters power down state. The transponder can also enter the ready state after having received a reset command. The transponder enters the selected mode after having received an explicit command with its own serial number from the interrogator. It is possible to set several transponders into the ready state by sending a specific command containing a matching partial identification number. It is necessary to set a transponder into the selected mode before being able to access it. The quit state is entered after a selected state after the interrogator selects another transponder or receives a specific command to enter this state.

Public mode communications start with the SOF pattern, which is followed by continuously streaming the data from the user memory. As shown in Figure 7.22, the first bit being sent is bit 31 from data block 0. When the transponder reaches its final data block, it starts transmitting data block 0 again. This process terminates when the electromagnetic field that energizes the transponder is turned off or when the transponder receives the appropriate command from the interrogator.

Electronic surveillance mode is a particular operating mode in which the transponders are attached to specific retail articles. By programming all bits in block 0 of the user memory to *0*, the transponder is set to an *unpaid state*. After the article has been purchased, the cash register sets the bits of block 0 to *1*, thus entering the *paid state*. Additional security can be achieved by protecting the bits of block 0 with a write password. If a transponder in the *unpaid state* enters the electromagnetic field generated by the interrogator located at the doors of the store, it will modulate the field. Consequently, the presence of a transponder marked in the *unpaid state* will be detected by the

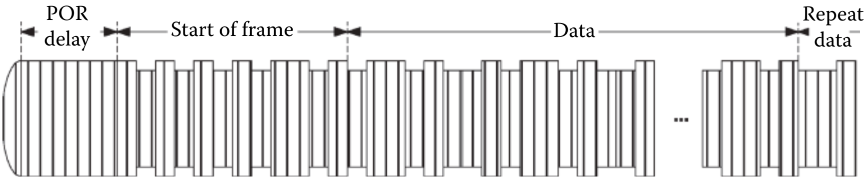

FIGURE 7.22
Initial sequence for the public mode.

© 2011 by Taylor and Francis Group, LLC

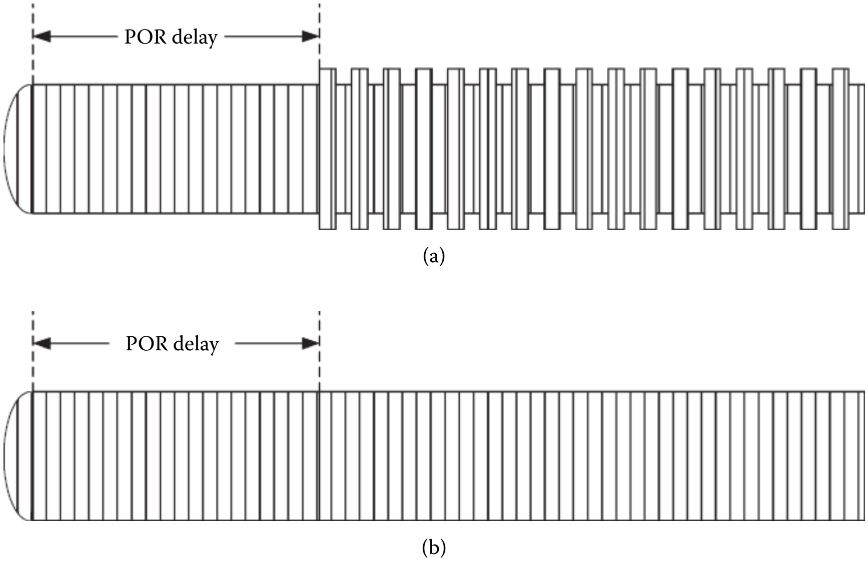

FIGURE 7.23
Signals from the transponder in electronic surveillance mode: (a) article unpaid, and (b) article paid.

interrogator, triggering the appropriate alarms. Conversely, a transponder marked as being in the *paid state* will not alter the electromagnetic field and will be transparent to the interrogator. Figure 7.23 shows the signals for these two cases.

Figure 7.24 shows the relationships between the different states for the transponders depending on their operational mode.

7.3.2 Transponders Operating in the UHF Bands

The possible states for the transponders manufactured by Atmel® operating in the UHF band are RF Field OFF, RF Field ON, Start, Mute, Observe, Ready, Active, and Isolated. These states and the paths between them are shown in Figure 7.25 and discussed here.

- The initial state is the *RF Field Off* state. Here, the transponder is not able to operate because there is no external radiofrequency field for which it can draw power. The transponder is able to store some configuration information for about 8 seconds after the field has ceased. After this time, the contents of these registers become undefined.
- When the transponder enters a region with an existing radiofrequency field whose magnitude allows the transponder to power

© 2011 by Taylor and Francis Group, LLC

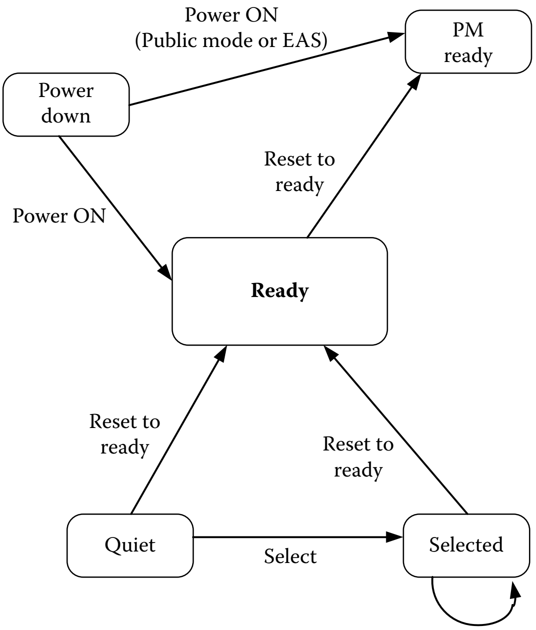
Power ON
(Public mode or EAS)
PM ready
Power down
Reset to ready
Power ON
Ready
Reset to ready
Reset to ready
Quiet
Select
Selected

FIGURE 7.24
Different transponder states.

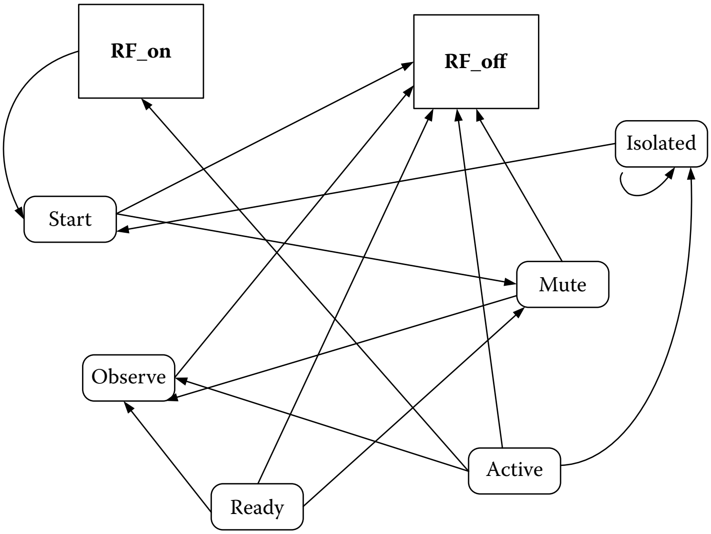
RF_on
RF_off
Isolated
Start
Mute
Observe
Active
Ready

FIGURE 7.25
Operational states for UHF transponders.

© 2011 by Taylor and Francis Group, LLC

its internal circuitry, this generates a power on reset. This is the *RF Field On* state. After the POR has occurred, the internal oscillator in the transponder is switched on. It is important to note that for this process to occur, the frequency of the external radiofrequency field must match the resonant frequency of the transponder's antenna and associated circuitry. Once this process has finished, the transponder moves into the *Start* state.

- In the *Start* state, the transponder refreshes the contents of the status registers and reads their configuration from its memory, enabling or disabling its trigger function. After this process, the transponder is able to detect the modulation ticks that signify information being sent from the interrogator. From this state, the transponder can move to the *Mute* state or to the *Observe* state.
- In the *Mute* state, the transponder cannot respond to operations such as programming, arbitration, or backscattering data. The transponder leaves the *Mute* state after some of its internal configuration registers have been modified.
- The *Observe* state allows the transponder to become synchronous with the interrogator for performing operations such as programming. To become synchronous the transponder checks that the forward header received from the interrogator is valid, that the command received is within the list of known commands, and that its CRC is also valid. If all these checks are passed successfully, the transponder moves to the *Ready* state.
- During the *Ready* state, the transponder receives the parameters associated with the type of commands sent by the interrogator. If any of the parameters are false, the transponder enters the *Mute* state. Otherwise, the transponder enters the *Active* state.
- In the *Active* state, the transponder is able to execute the command sent by the interrogator after receiving and checking the rest of the forward frame. If there are any errors in the information sent after the header frame, the transponder moves into an *Isolated* state or *Mute* state depending on the type of errors received. If the information sent in the forward frame is correct, the transponder executes the command. Once the transponder has finished the execution it moves back into the *Ready* state. Figure 7.26 shows with more detail the processes that occur during the *Active* state.
- The first step in this process is to verify the length of the received stream. If this is correct, the transponder checks for the validity of the position of the end-of-transmission frame. Depending on the result of this check, the transponder knows if the anticollision procedure has been activated in order to select that specific transponder. If this is the case, the transponder will enter the appropriate

© 2011 by Taylor and Francis Group, LLC

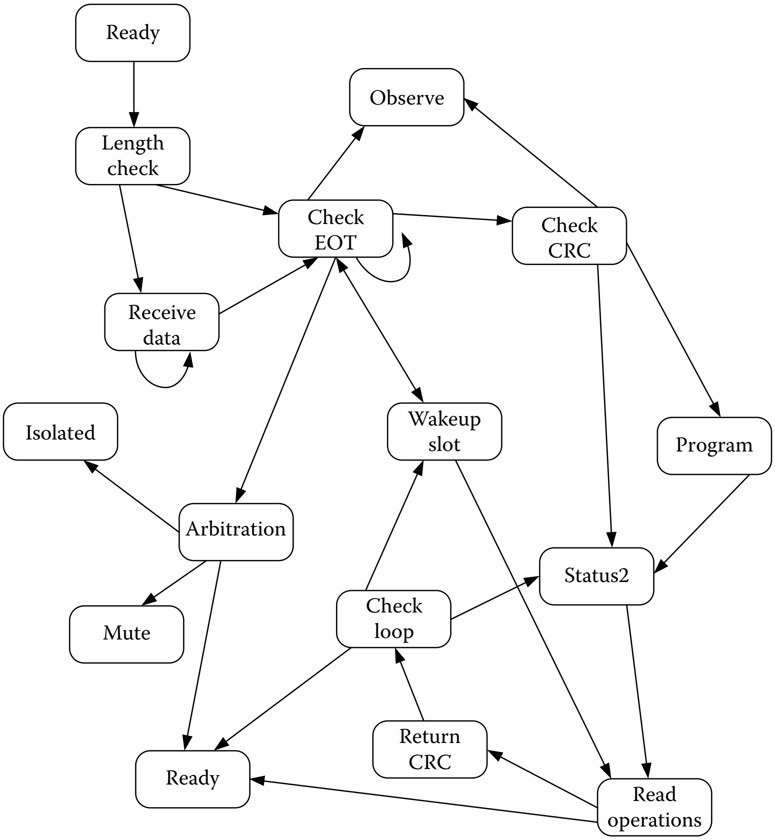
Ready
Observe
Length check
Check EOT
Check CRC
Receive data
Isolated
Wakeup slot
Program
Arbitration
Status2
Mute
Check loop
Return CRC
Ready
Read operations

FIGURE 7.26
Processes during the active state of transponders.

anticollision and arbitration procedures. The transponder can also enter into receiving wakeup commands or other types of commands sent by the interrogator.

- In the *Isolated* state, the transponder is isolated and waits for either a POR or reset command to leave this state.

7.4 Arbitration for Transponders

7.4.1 Principles of Arbitration

When there is more than one transponder immersed in the electromagnetic field generated by an interrogator, it is necessary to develop and establish a

© 2011 by Taylor and Francis Group, LLC

procedure to select a single transponder or a specific group of transponders for carrying out specific functions. Alternatively, it can also be necessary to isolate a specific transponder or a group of transponders so they do not disturb communications. In some applications it may also be interesting to know whether there are transponders in the field or not. All these functions require the use of arbitration procedures. The two main types of arbitration procedures are *deterministic arbitration procedures* and time-slot based procedures, also known as *aloha procedures*.

Deterministic procedures are based on a *tree-walking function* in which the interrogator acting as the control node uses the individual IDs of transponders to select one or a group of them. The interrogator starts by polling all the transponders in the field. If there is more than one, their responses will collide and the interrogator will not receive the message. Therefore, in the next step, the interrogator polls only a specific subgroup of transponders based on their IDs. If the responses also collide, the interrogator reduces the population by polling only a specific subgroup from the previous subgroup. This process is repeated until the interrogator receives a response from a single transponder.

Aloha procedures are controlled by slot commands. After the transponders are turned on, they calculate random values for as long as they are powered. Once the transponder has received a slot command, it creates a random number from its generator, which points to a slot number in the chip. After receiving a start command, the transponders transmit their information to the reader during the first slop. If the random number matches the first slot number, it calculates a new one during this slot. When the random number matches the slot number controlled by the interrogator, that transponder becomes active. This allows the transponder to communicate with the interrogator. If the numbers do not match, the transponder enters a mute state and therefore does not respond to any commands. When the transponder receives a new slot command, it calculates if it is supposed to be active or mute.

7.4.2 Principles of Anticollision

The aim of anticollision procedures is to detect and identify the transponders that are present within the range of the electromagnetic field generated by the interrogator. While the implementation of specific anticollision procedures may vary between manufacturer and models, the principles of anticollision are common to all of them. For transponders manufactured by Atmel®, the communication with a single transponder or a group of transponders is controlled by the interrogator. The arbitration is initiated by the interrogator issuing a command for the transponders to transmit their identification numbers. The command can be tailored to the transponders that the interrogator expects to be in its area of influence. If the interrogator does not know anything about the transponders that are expected to be, the command is general and the detection encompasses all transponders. Otherwise,

© 2011 by Taylor and Francis Group, LLC

the command can use parameters such as a partial transponder identification number that will cause only transponders with partially matching IDs to respond to the command.

After the interrogator receives responses from the transponders, it scans their ID number from the most significant to the least significant bits. Each time slot corresponds to a specific bit position in the ID number. Because all transponders respond simultaneously with their modulated signal, their responses will be superimposed on one another. Therefore, because of the modulation being used, a damped signal overwrites a nondamped signal. With this, a logical *1* prevails over a logical *0*. Consequently, the transponders with these ID numbers are removed from the process. When the interrogator detects a logical *1*, it responds by broadcasting a gap in the radiofrequency field. The detection of a logical *0* does not originate any response from the interrogator. The gap in the field that is created as a response to the logical *0* is monitored by the transponders in the field and deduces, by comparing their ID number to the gap that was just broadcasted, if that specific transponder should continue in the current anticollision loop or eliminate itself from it. The transponders that are eliminated are muted and return to the ready state, in which they will remain until the interrogator initiates a new anticollision loop.

The transponders that have not been eliminated remain in the current loop as long as the final bit of the identification number has not been reached yet, and the process repeats itself. By the time the last bit has been interrogated, there will be a single transponder active as all the other transponders will have been eliminated through the previous steps. At this point, the interrogator is able to identify the associated ID for that specific transponder, and the transponder enters into the Selected state. Now, the transponder can establish direct communication with the interrogator and can be read or written. The transponder remains in the Selected state until the interrogator starts a new anticollision loop.

The MCRF450 family of transponders manufactured by Microchip Technology also supports anticollision. Their algorithm is based on time division multiplexing the transponder responses in which each transponder is only allowed to communicate with the interrogator during its time slot. This algorithm contains four control loops: Detection, Processing, Sleeping, and Reactivation.

The control of these loops is done by five commands as well as the transponder internal flag bits. The commands that control the anticollision loops are as follows:

- *Fast read request (FRR).* With an internal flag set, the transponder will respond only to an FRR command. An FFR command consists of five timed gap pulses, as shown in Figure 7.27, in which the period of the gap contains specific information. There are a total of nine possible pulse positions allowed in this protocol.

© 2011 by Taylor and Francis Group, LLC

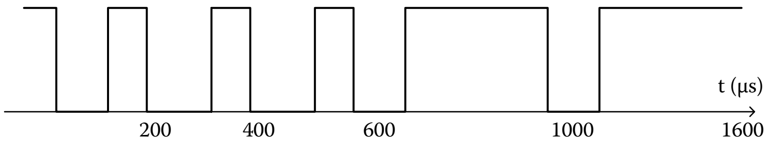

FIGURE 7.27
Structure of a fast read request (FRR).

- *Fast read bypass (FRB).* This command is used in the reactivation loop. Upon reception of this command, the transponder responds with a block of data. The structure of this command is similar to the FRR command, with the difference being in the period of the gaps.
- *Matching code 1 (MC1).* This command is used with the transponder and does not require any further processing. The command consists of time calibration pulses and pulse position modulation (PPM) signals. Upon the reception of this command, a transponder in the detection loops enters the sleeping loop.
- *Matching code 2 (MC2).* When a transponder in the detection loop receives this command, it enters the processing loop. The structure of this command is similar to that of MC1.
- *End process (EP).* This command causes a transponder in the processing loop to exit it and enter the sleeping loop.

Transponders can enter the detection loop by waiting for an FFR command or by transmitting a fast read (FR) response without waiting for a command. The first operation mode is called *interrogator talks first* (ITF), while the second operation mode is called *tag talks first* (TTF). The read and write operations are carried out during the processing loop. Only transponders with a special flag set in this loop respond to write or read operations. The sleeping loop is used to keep all processed devices in a silent condition in which they do not respond to external commands. The reactivation loop is used as an intermediate step by the interrogator to set transponders into different loops depending on the command that is transmitted to transponders in this loop.

7.4.3 Deterministic Anticollision Procedures

Deterministic anticollision procedures are based on the transponder backscattering the information to the interrogator only if the result of a compare condition is true. The compare condition is set up by the interrogator while broadcasting a specific command. Deterministic anticollision procedures are also called tree-based procedures. The interrogator sends a series of queries, each one with a prefix that is part of the transponder's ID code. Only those transponders with matching IDs respond to the interrogator's query. If there is more than one transponder with matching partial IDs, these transponders respond at the same time

© 2011 by Taylor and Francis Group, LLC

QRP SWL HOMEBUILDER

RF — Test and Measurement Homepage

Topics 1998 - 2005

- SWL Page
- RF Preamps
- QRPHB Software
- Popcorn DC Receiver Mainframe
- 40 M Popcorn Superhet Receiver
- Broadband Transformer Topics
- Diplexer Topics
- Ugly Construction
- Wee Willy 75 M DSB Transceiver
- W7ZOI JFET Biasing Tutorial
- EMRFD Review
- Tapped Cap Impedance Transformation
- More Active Antenna Experiments
- Fun with LEDs
- SWL Receiving Antenna Experiments

Topics 2006 - 2009

- Experiments with JFET Biasing
- TRF Receiver for WWV
- MF and HF SWL Receiving Antenna
- MF and HF Receive Antenna Splitter
- MF TRF Receiver
- Junk Box Low Pass NDB Filter
- More NDB Ideas and Circuits
- Low Pass Audio Filter Experiments
- TRF WWV Receiver for 5 MHz
- Complementary-Symmetry Amplifier Biasing Basics
- Low Power Audio Amp Experiments
- Two Bravo Receiver Experiments
- ICOM IC-7200 Review
- Electronic Hobbyist Circuits
- Output of a Diode Ring Mixer in an Oscilloscope

Topics 2010 - 2011

- Low Noise Crystal Oscillators
- Crystal Parameter Checker Supplement
- Hobbyist Circuits 2010
- RF Workbench Page 1
- RF Workbench Page 2
- AF Transistor Input Z Experiments
- RF Workbench Page 3
- Misc RF Experiments 2011
- Hobby and Fun 2011
- QRP Modules 2011
- Receiver Band-pass Filters
- RF Workbench Page 4
- VFO-2011
- VHF Butler Oscillator

Topics 2012 - 2014

- VHF to the Max !
- HF Ragbag
- RF Workbench Page 5
- VHF-FM
- Sundry Experiments 2012-13
- VHF - Veronica
- RF Workbench Page 6
- HF Embarcadero
- Caitlyn 310 : UHF Beginnings

Make it — then measure it!

RF — Test and Measurement

Frequently Asked Questions

1. How to you make your schematics?

I hand draw my schematics using the Paint Program that ships with all Microsoft Windows(tm) operating systems.

I strive to make schematics as clear and as small in file size as possible.The Windows Clipboard is used extensively to copy and paste the desired components from previous schematics to new schematics — very few new components are ever drawn; rather they are recycled from schematic to schematic. My system allows me to paste color 3-D labels and small photographic bitmaps into schematics.

The raw 24-bit bitmap drawings are compressed to 16-bit png files. Prior to May 17, 2010, bitmaps were compressed as 8-bit gif files. An example drafted circuit using the Windows 7 Paint program follows.

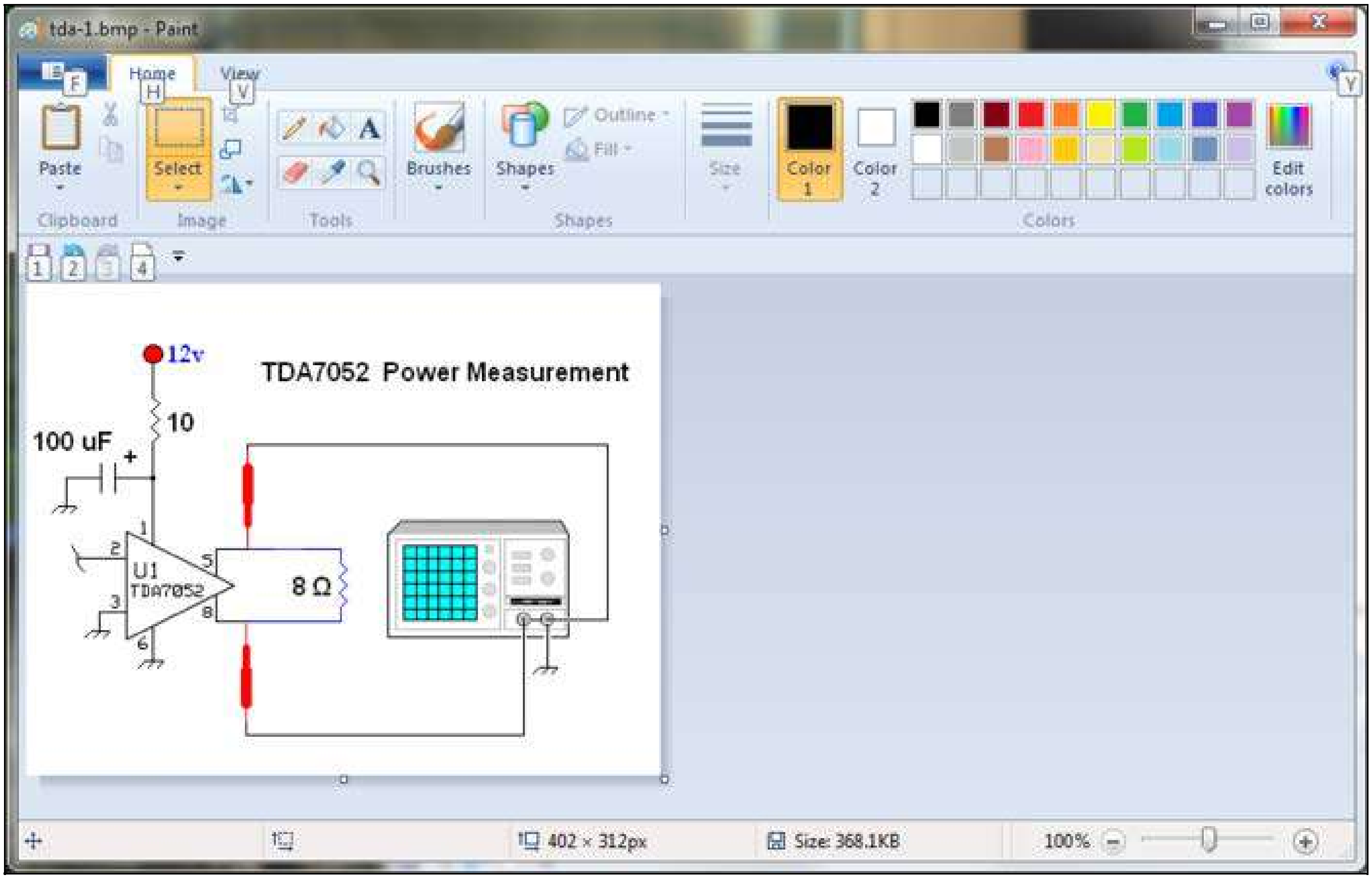
tda-1.bmp - Paint
Home
View
Paste
Select
Brushes
Shapes
Outline
Fill
Size
Color 1
Color 2
Edit colors
Clipboard
Image
Tools
Shapes
Colors
12v
TDA7052 Power Measurement
10
100 uF
+
1
2
3
5
6
8
U1
TDA7052
8 Ω
402 × 312px
Size: 368.1KB
100%

At least 25 people have sent or recommended software for making schematics. Thank you for this kind gesture, but I prefer my current method and after 15 years — make schematics quickly.

2. How come you don't supply parts lists? Other people do.

The answer is simple; lack of time. It takes considerable time and effort to put up a new web page and also to maintain a large web site. I save time by leaving the parts list up to the builder. In addition, this site is about experimentation and using what parts you have on hand.

3. Why didn't you answer my email?

I answer all legitimate emails as soon as possible. Our POP3 server gets an average of 2016 spam emails per month (December 2013 data), however, our software removes 99.27% of these and I never see them. Occasionally, legitimate emails are filtered in error and I apologize. I've received as many as 83 legitimate emails in a single day so can get behind. Please keep emailing — when readers stop emailing, I'll know the site has fallen totally obsolete [it may already be obsolete] and delete it.

Our mail server software logs and analyzes the network information of all spammers and may automatically filter and/or block their addresses or even their entire ISPs at the router level. Analysis indicates that 90% of our SPAM comes from just 3 countries and if you happen to live in one of these countries, the filtering will be especially sensitive.

This sounds dogmatic and unfriendly, however, until you've set up a domain and must handle ++ spam emails, endure and then develop router control software and other strategies to handle DDoS attacks and so forth, your completely naive about the 'dark side' of running a homebrew web server. Massive amounts of bandwidth might be otherwise wasted by allowing unwanted server use unless we actively counter these activities to keep the site running well for legitimate hobbyists. Not to mention all the wasted time. Further:

The 3 host web servers, the power, the server software, the security apps, the internet bandwidth etc.are owned or purchased by my family, and as you know, nothing is free. Despite many offers by companies to place ads on my pages, I've kept the site advertisment free and running pops.net costs us a few thousand dollars each year. We ask you to please respect our site for the sake of the experimenters who visit.

Pops.net server rack in our warehouse.

While I appreciate that some people might want to email invite me onto their social networks, I do not have time to participate. All email traffic from or involving social networks see this page for a list is deleted automatically by our POP3 server control software.

I **never** buy or sell parts via email, nor exchange hyperlinks. Never. I do give free parts to those in need though.

All email with the .info domain is blocked.

The number of people selling kits has jumped up by ~4 dB in the past 5 years. Increasingly, builders who need help with kits were emailing me for support. I rarely build kits and my knowledge regarding kit building is nearly 0. Please contact your kit seller for help.

You may wish to enquire with the kit seller about their online support polices and promptness prior to purchase. Additionally you might try the "support" email address provided and see if and how promptly they reply. Most of the popular kit sellers (AADE, Kits and Parts, etc.) provide excellent support to their customers. Like anything else online; buyer beware.

4. How come you didn't link to my web site - I linked to yours?

A big thanks to the folks who link to this web site! The QRP/SWL HomeBuilder site focus is content, not web links. Making a lot of links means spending time testing for and tracking down dead links - the so called "link rot". Time spent on the web site is time away from the electronics work bench. In addition, it is not logistically possible to reciprocate in kind, as hundreds of web sites and blogs have linked this site.

5. I see the word "popcorn" used a lot on this site- what's this all about?

Popcorn connotes the essential theme of the web site; simple, frugal, without fuss and over use of technical jargon, or complex math and engineering techniques. The QRP/SWL HomeBuilder web site is referred to as the popcorn site by many. The site targets hobbyists. The emphasis is fun. The hope is that it will attract new people to electronic design, measurement and experimentation. Hopefully, this site stimulates interest in QRP homebrew electronics.

Soon after I began building electronic circuits, my teachers and the popular electronic-related media of the day pushed me towards etched, printed circuit boards. I complied and this killed my passion for electronics. For me, habitually stuffing circuit boards lacks creativity and freedom.

Later, I discovered people were building guitar and bass amps using point to point wiring techniques with terminal strips and partial circuit boards. I became interested in building and repairing guitar amps and this passion continues today. In 1992, the discovery of 2 QST articles changed everything *for me (complete reference provided):* The Ugly Weekender: parts 1 and 2 by Roger Hayward, KA7EXM and Wes Hayward, W7ZOI; published in QST for August 1981 and June 1992. This was my first exposure to Ugly Construction and it was immediately adopted as the defacto standard bread boarding method in my electronics work shop. In fairness, etched circuit boards are a great tool, but not essential for the experimenter.
After working with Ugly Construction over time, considerable progress was made in understanding RF circuits and one output was the launch of this web site in 1998.

Currently, little has changed, I continue to prefer scratch-homebrew rather than kit-homebrew electronics. My interest in Short Wave radio and analog electronics has grown considerably. For me, electronic circuits hold a certain mystique which arouses my curiosity to learn, enjoy and share. As a lay person, this web site has facilitated meeting some awesome people through email from all continents and it has been a privilege to learn from them, my mentors, book and web authors and often enough; from my mistakes.

5. What do you mean by a 5K1 or 3K3 resistor value?

For E24 or 5% tolerance resistors 5K1 = 5.1K, 3K3 = 3.3K and so on. For E96 or 1% resistors 31.6K is written as 31.6K. All resistors are 1/4 watt unless otherwise specified.

6. How do you measure audio amp output power?

Please see Figure 4 on this web page. Any amp when cranked, outputs much greater power than when it is providing a clean sine wave. The quoted power for any audio power amp on this web site is the maximum average power it will give before the pure sine wave becomes distorted.

7. I noticed a new web page appears and then it is edited for 1-2 weeks. When is the web page completed?

When a new web page is added, it takes a week or so to find and change some of the grammar and spelling errors. Sometimes new ideas or feedback will cause me to further edit a web page at any point in time. This whole web site is a work in progress. The last date any given web page was edited is posted on the bottom of the web page.

8. Do you buy or sell stuff?

No and no. I receive numerous emails from people asking me to sell them stuff. I do not sell anything - no parts, books, coffee cups, ball caps, tee-shirts, ad space — nothing. I do not buy parts in commercial-quantity volumes and have no need to make contracts for obtaining any electronic components. Every week, Asian companies email to ask about buying their parts — please note, my answer is *always* the same: no thank you.

9. Questions and concerns about printing and printability

Each year, a few readers email to complain how poorly the web pages print. This is true and I apologize.

Some people prefer pdf files for easy printing. I have resisted going to pdf format for 3 main reasons:

1. The web site audience is international and many are using web translators. PDF files are 8-bit graphic image files and do not translate.
2. More and more readers are using mobile computer devices and pdf files are a pain for them.
3. We should all print less often to save resources

As an experimenter, I dislike crammed, small-size schematics and feel they should be drawn for *maximum* clarity. Therefore, my schematics tend to have a lot of white space and color contrast. I try to make them no wider than 700 pixels, although sometimes it's impossible to do this. Big schematics are *not* printer friendly. The only practical solution is to click on and open them in a separate browser window for easier sizing and printing.

I also feature big photos which burn up a lot of printer paper. Project photos are important to me; they provide a more intimate glimpse into the bench work and promote the real purpose of the site — building stuff.

A potential printing solution for Microsoft Explorer 8 users; Click

10. I have noticed in your CMOS logic photographs, you don't always ground unused input gates. Isn't this bad?

Proper CMOS logic practice mandates the grounding of all unused input gates. In prototypes and experiments, I don't always do this as I generally want to re-use the IC in other experiments. This is a cost saving measure. When you build a lot of stuff, it can get expensive and recycling parts makes sense. In keeper circuits or critical prototypes, unused input gates are directly soldered to the copper clad board. This also anchors the IC very well.

11. What is the proper URL of the home page?

http://www.qrp.pops.net

The following pre-2006 URL was decommissioned August 6, 2010: http://www.qrp.pops.net/default.htm

12. What are the QRPHB Design Centers and Professor Ivanenko character about?

This web site is about design and not just providing circuits to copy; I'm hopeful that the QRPHB Design Center concept initiated in 2011 will invigorate the site. Design Centers are the presentation of simple, but useful algorithms for amateur builders to advance their skills on the bench.

Professor Vasily Ivanenko ( ), a *fictitious* retired Russian physics professor wants to share his knowledge and give back to society. He signifies each Design Center. Professor Ivanenko was drawn for me by Rod Adams in 1996 using the Paint program that ships with Windows (the same app I make my schematics with). Rod did all of the other original bitmap art for this website including the coil guy and junk box pictures. This character was inspired by one of my favorite photographers: Irving Penn — this photograph, which is all over the web.

A new character; Dr. Natasha Petrovna appeared in late Summer 2012.

The professors are just a good bit of fun — add intrigue, characters on whom to focus and a means of identifying Design Centers. Electronics with just math and physics bores us all. Adding splash, color, clear photographs and characters such as the coil guy or the Professors boosts the site's appeal and provides a creative outlet for me.

13. Why did you kill your blog?

Time mostly. *My* blog wasted *yours* and my time. I carefully analyzed my personal yield from blogs in 2012 to 2013 — for the most part, blogs just entertain + share trivia, or rehash someone elses idea(s), or 'innocently' attract you in hope to sell stuff — and sometimes, just fulfill the author's need for attention. I don't seek, nor have time for entertainment or spectacle within my RF hobby and I certainly don't wish to waste *your* time. Each to his own, I suppose.

My analysis showed that unfortunately, blogs rarely boosted my understanding of electronic design or measurement practices. My ardent focus is to learn + improve and then pay some of this knowledge forward on a web site. Of course, everything in context — many exceptional people blog. For example, Dave AA7EE, or Jason, NT7S.

Most of the RF design and measurement people I follow keep old fashion web sites and provide generous email support. Design and measurement web sites, plus reputable and/or peer-reviewed industry and hobbist books, journals and multimedia work best for me.

Further, great elmers don't just publicly hang-out on blogs, or web server groups, or publish Utube videos — some just check their emails and when asked — give wonderful support without fanfare. Hats off to these humble folks. Thank you!

14. What oscilloscope should I buy?

Yikes — a tough question I get nearly every month. Please do your research. My best answer is buy the best 'scope you can afford. Are you a casual experimenter, or sit in your lab a lot? I prefer DSO's , however, made due with a old boat- anchor CRT for my first 10 years. The Rigol 1052/1152 seem popular entry-level choices due to their cost versus performance ratio. On the other hand, view this video to see how much Rigol DSO technology has changed.

I owned and sold my 1052 to a builder in Michigan — a worthy choice like many other 'scopes. IMHO, the FFT and math functions on the 1052 and 1152 suffer due to low memory depth and clock jitter.

Even if you only work at HF, a bandwidth 5X higher than your main frequencies of interest works better for showing harmonics. Again, I advise people to simply buy the best 'scope they can afford since it will form the heart of your test bench.

15. I'm a beginner — what toroids should I buy?

Opinions will vary, but here's what I recommend. Buy any quantity you wish, but sometimes minimum quantities apply and shipping to some countries costs dearly, so I tend to order enough parts to last me for awhile.

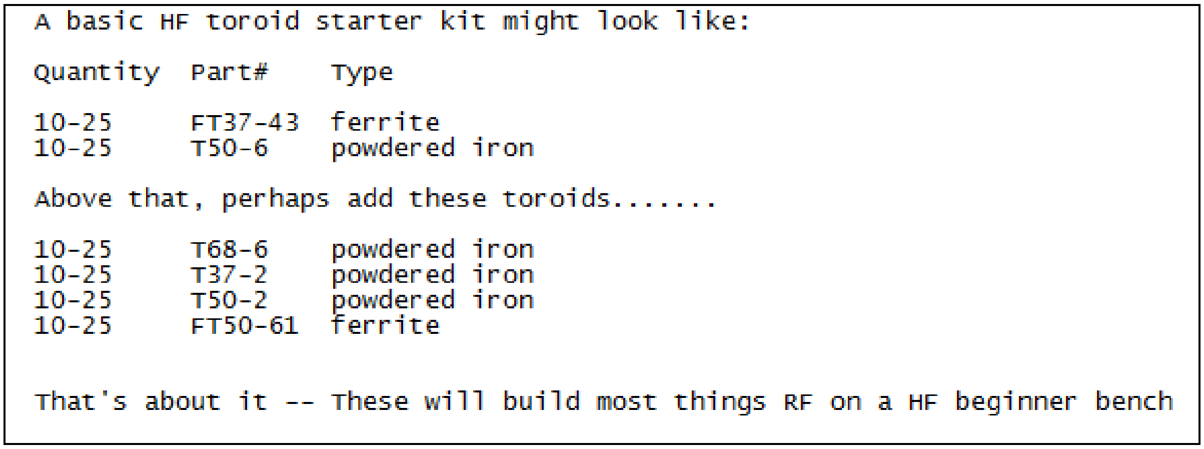

```
A basic HF toroid starter kit might look like:

Quantity  Part#    Type

10-25     FT37-43  ferrite
10-25     T50-6    powdered iron

Above that, perhaps add these toroids.......

10-25     T68-6    powdered iron
10-25     T37-2    powdered iron
10-25     T50-2    powdered iron
10-25     FT50-61  ferrite


That's about it -- These will build most things RF on a HF beginner bench
```

RF — Test and Measurement

QRP — Log: Updates To Permanent Content

15 dB ATTENUATOR
50 OHM
0403

Oct 5, 2014 --- The end of QRPHB? Expect the Qrp site to go down each day for 1-4 hours. As I wrote before, we're losing an average of 10% of the outgoing packets and this peaks as high as 40% when many people are accessing the site. See an example graph below.

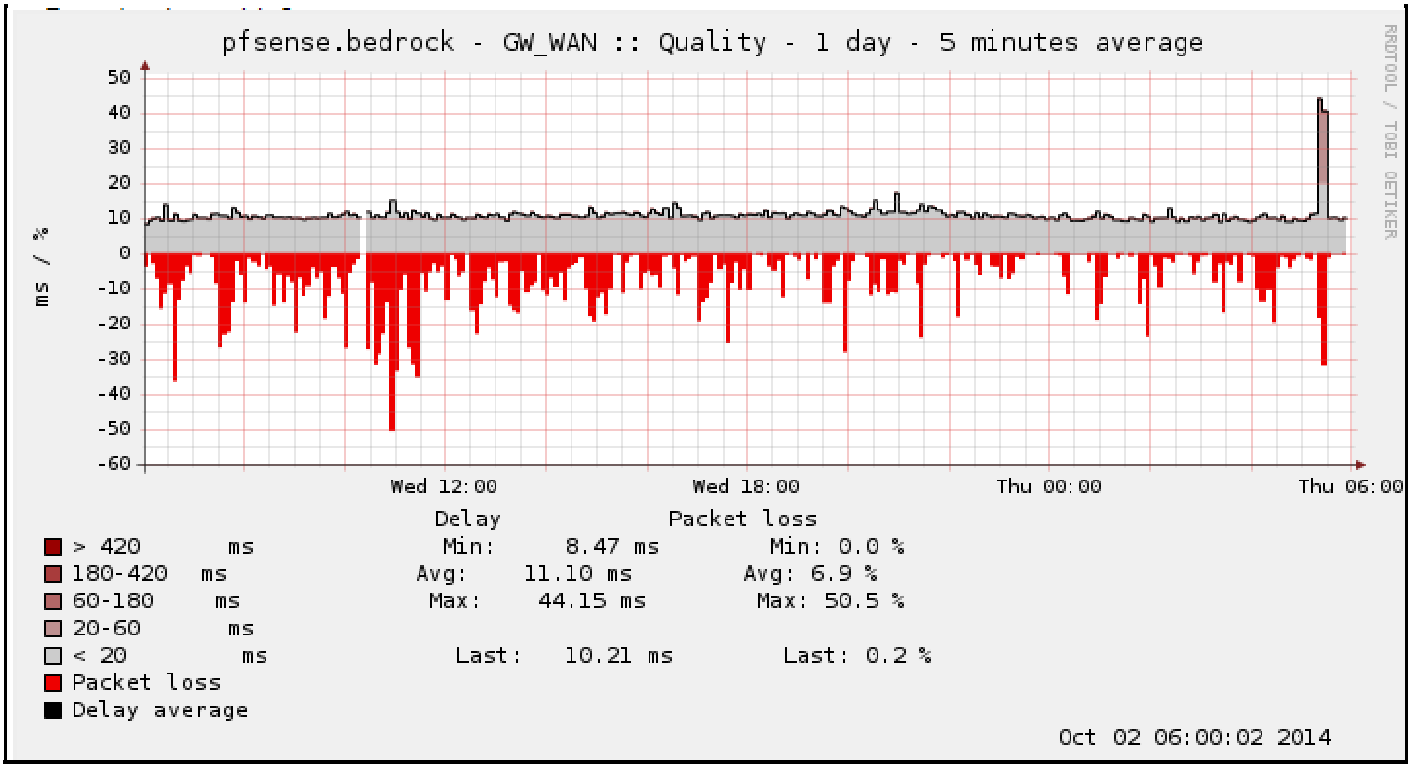

On Oct 2, 1 of our ISP's techs came and tested our WAN -- "it's better than most" he said. Hopefully they can do something to boost the performance on their side. Wave bought out all the local competition and we're stuck with them [this is happening everywhere]. Click for 1 of the many links writing about the emerging internet cable company monopoly.

You'd think that in 2014 we could properly host a web site like we've done since 1998? To boot, Stuart is paying for a high bandwidth business account and getting less than 1% of the promised bandwidth. We chatted today and will wait patiently for Wave to improve our outflow - but only for so long.

If Wave can't or won't boost performance, I am leaning towards taking this web site down as opposed to moving it -- it's been a good run. QRPHB creates a lot of work and I could get way more done without the web site hassles. So, 2014 might be it for me. If so, thanks for coming here all these years and best of luck with your experiments!

--- Sep. 19 --- Wave Broadband, our Internet provider suffered major problems recently and I shut the site down for 4 days rather than have it limp along pathetically. Well the Wave saga *continues*: the site loads slowly and pictures are missing etc.

Our email server remains working even when the main site is down. You can always email me if the site is not working for you -- or if you have a comment or question.

Vistor volume rose this Summer so we added a better router that provides bit by bit data performance collection and new features to help throughput. At minimum user bandwidth, a unique visitor hits the site every ~0.7 minutes. I'll start formal bench experiments on October 3 --- to kick off the sites' 16th season.

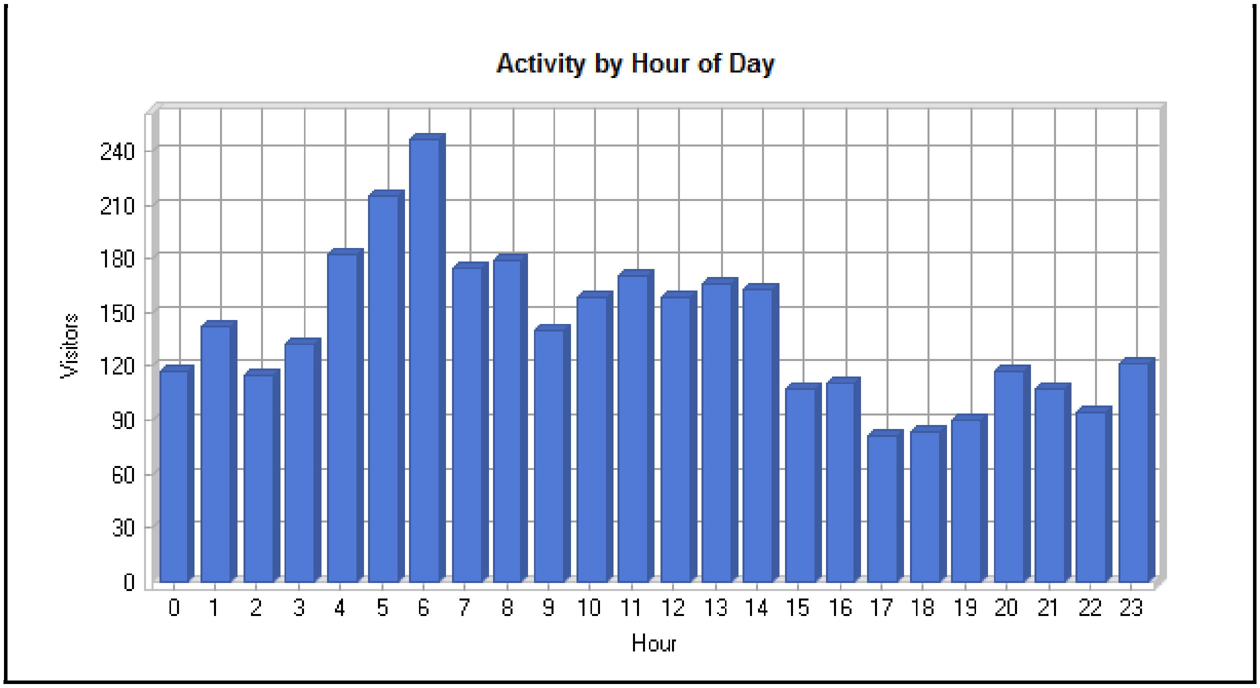

August 31, 2014 — Kit to upgrade my HP frequency counter to 3 GHz added as Section 5 on Caitlyn 310 — UHF Beginnings.

August 18, 2014 — QRP-POS Data on the Sundry Web Page. Look at the end of Section 8: Popcorn AF Amps For Receivers — Reprise. This new, all-discrete AF stage will go into the Funster Line receiver and ranks as 1 of my best in terms of power and headroom.

August 12, 2014 — Funster Line: a QRP 40M band CW trans-receiver added to HF Embarcadero web page as Menu item 4. Click here for Funster. Only the transmitter is presented for now.

June 18, 2014 — I tweaked the page now called About... on the top level menu. Also, on this web page, I added a new essay for 2014 just under the essay for 2010. I'm off the bench until Fall, but hope to add a little content on rainy days or such. I'm about 2 years behind in presenting some of my experiments.

May 22 and 26, 2014 — New Supplemental Web Page for VHF-FM launched. This page and another supplement will house some new receivers over time. The new supplement is linked on the original VHF-FM web page in section number 4.

May 5, 2014 — Section 7: NE612 Mixer Diddy added to VHF Veronica

April 14, 2014 — Completion of Return Loss Bridge Experiments [added Bridge #4] on Caitlyn 310.

April 7, 2014 — Return loss and VCO experiments added to Caitlyn 310.

March 7, 2014 — Section 3 added to VHF-FM. A DC Converter for VCOs. **March 25** QRP-POSDATA for March 2014 Poor Hams Scalar Network Analyzer (PHSNA) added to Section 1 of Sundry.

February 14, 2014 — Caitlyn 310 — New repository web page for my venture into UHF. Click. Surprisingly, the site averages ~ 3000 unique visitors every 24 hours. Click for the Feb 25 tally

January 15, 2014 — The FAQ was often missed and therefore moved to the top level menu + editted/augmented.

January 4, 2014 — A follow-on version of the K3NHI QEX power meter added to RF Workbench 5 as section 6.

December 15, 2013 — I added Section 3 to HF Embarcadero. VXO and VCXO Notes. I significantly updated the VFO-2011 web page on Dec 17, 2013.

November 7, 2013 — Section 6 added to VHF — Véronique. 1 photo added to the end of the Ugly Construction page. Severe Fall weather conditions took down the server for ~1 hour today.

November 1, 2013 — I added Section 2 to HF Embarcadero. A vestigal set of notes regarding my attempts to update the Popcorn Superhet receiver.

October 31, 2013 — Section 5 added to VHF — Véronique.

October 27, 2013 — VHF-FM re-formatted and Section 2 added. I imagine this page will disinterest many.

October 18, 2013 — 3 rarely accessed web pages removed from drop down menu, but not deleted from the server. New QRP-Posdata added to HF-Ragbag (near the end). Section 3 edited and Section 4 added to VHF — Véronique.

October 11, 2013 — Build Season 15 begins [I bench experiment Oct to May]. I added HF Embarcadero to hold all my HF and perhaps AF experiments this season. I'll also add content to VHF FM, VHF — Véronique and RF Workbench 6 over the next 12 months. Thanks.

September 9, 2013 — Blog deleted.

August 8, 2013 — Section 2 of RF Workbench 6 added. Measuring PA collector V and I to calculate efficiency. On August 16 — I added a new essay on the Ugly Construction page called Is Ugly Construction Less Reproducible than Manhattan?

June 25, 2013 — I started RF Workbench 6. It will take 1 year to complete

April 14, 2013 — I added Section 3: 50 Ω MMIC Bench Amplifer to the VHF 2013 Veronica web page.

March 31, 2013 — I started a new content page called Pin Outs

March 25, 2013 — Section 9: an essay about the 1981 Progressive Receiver, plus the final section — 10: Miscellaneous Pictures and Figures added to Sundry 2012-13

March 18, 2013 — Section 8. Popcorn AF Amplifier for Receivers — Reprise added to Sundry 2012-13

March 10, 2013 — I added a new VHF content page for 2013: VHF-2013 - Veronica

February 16, 2012 — I added Section **7** -- A Journey Above HF -- to the Sundry Experiments 2012-13 web page.

February 14, 2012 — I added Section **6** -- Non-Mechanical Iambic Paddle -- to the Sundry 2012-13 web page.

February 1, 2013 — A seperate QRP—POSDATA added to RF Workbench 4. Now QRP— POSDATA 1, 2 and 3.

December 15, 2012 — I added Section 4 to Sundry Experiments 2012 - 2013 . A PLL circuit from EMRFD.

January 5, 2013 — I added Section 5 to Sundry Experiments 2012 - 2013 . A simple AF feedback amp. An essay concerning L-C meters was also added on RF Workbench 5. Section **5.**

December 22, 2012 — I deleted the web page ***Tuning VFOs with a PN Junction*** since some of the experiments were poor quality and performed back in 1998 when I was more ignorant than now. I've learned much since then and my new VCOs from the past 1-2 years reflect this knowledge. The ***Selected QRP Reading list*** and ***Cascode 7 Receiver*** web pages were also wiped.

December 15, 2012 — I added Section 4 to Sundry Experiments 2012 - 2013 . A PLL circuit from EMRFD.

December 3, 2012 — I added Section 3 to Sundry Experiments 2012 - 2013 . Interview with Jason from Etherkit

November 13, 2012 — QRP — Posdata added to the VHF to the Max web page. Section 5: Z-Comm VCO.

November 1, 2012 — Sundry Experiments 2012 - 2013 web page added.
QRP — PosData added to Section 2 of **Power Meter Calibrators** on RF Workbench 5 and updated again on Nov 22, 2012.

October 16, 2012 — VHF FM web page added. Already, it has spawned a first supplemental web page

Sept 24, 2012 — QRP — Posdata added to the end of the HF Ragbag web page. I added a bypass and decouple network for HF to lower VHF.

August 17, 2012 — QRP — Posdata #2 added to bottom of the Receiver Band-pass Filters web page.

August 6, 2012 — Section 5: **Some Experiments with RF Bypass Capacitors** added to the HF Ragbag web page. Also a new QRP — Posdata added to the bottom of the Crystal Parameter Checker web page.

August 1, 2012 — QRP — Posdata added to the Receiver Band-pass Filters web page.

July 12, 2012 — I added a corrected schematic on the Wee Willy page: Wayne. M0WAY — 14 MHz PA under August 25, 2011. Also, added a new essay on **Microphonics in DC Receivers**. See Section 4 on the HF Ragbag Page

June 23, 2012 — RF Workbench 5 added. Click here.

May 31, 2012 — Galina discovered that I neglected to publish the proper version of the Hobby and Fun 2011 page and corrected my error. A fine-tuneable Wien Bridge Oscillator idea from Ken Kuhn now appears at the page bottom.

April 21, 2012 — I heavily edited RF Workbench 1 and 2.

April 13, 2012 — Web site purge. I removed RF Filters, VFO 1998, QRP Workshop Ideas, Miscellaneous Schematics and Photos, Base-biased VFO, Funster Transceiver, Miscellaneous Circuits and Ideas 2005 and Crystal Oscillator Offsets. **Reason:** substandard.

April 6, 2012 — 50 MHz Receiver Pre-amp and Filter added to VHF to the Max web page .Section 4.

March 26, 2012 — HF Ragbag web page added to top-level menu. Non-VHF experiment repository for 2012.

March 19, 2012 — VHF to the Max web page added to top-level menu. 50 MHz VCO experiments added to this page in Section 3.

March 14, 2012 — EMRFD review edited.

February 29, 2012 — Minor edits to the Audio Transistor Input Impedance Experiments web page. Also, I updated the calculation of the common emiiter amplifier base input resistance using the better formula: **Rin = (B+1)*(re + RE')** [while ignoring REB]. This is my favorite web page on the site.

February 26, 2012 — QRP— Posdata added to the bottom of the 2nd NDB web page

February 17, 2012 — I introduced a new miscellaneous VHF page: VHF to the Max — I'll slowly add stuff over 2012. Major editing done to the Broadband Transformer web page.

February 4, 2012 — RF Workbench 1 and 2 significantly edited. 2 new photos added.

January 28, 2012 — QRP — Posdata added to Crystal Parameter Checker web page and to QRP Modules 2011 under 7 MHz VCO Experiments on this web page. I re-wrote the temperature compensation section of the VFO-2011 web page and added 3 photographs.

January 3, 2012 — Web site change: I update web essays with an "epilogue section". In 2012 and on, they will be called QRP — Posdata (Spanish for post-script or epilogue). Posdata #2 added to the RF Workbench 4 web page.

December 17, 2011 — The Butler Did It ! - First VHF Experiments 2011 web page added. A 50 MHz frequency doubler added Dec 26, 2011.

Nov 12, 2011 — Our server went down for 16 hours. Both AC power and cable Internet to the warehouse failed after a rain storm and wind gusts knocked down some trees that severed the hydro and cable wires. Expect more weather-related down time as Winter approaches.

Oct 15, 2011 — VFO-2011 added.

Oct 2, 2011 — RF Workbench Page 4 added. On **Oct 17, 2011** I added an epilogue.

Sept 20, 2011 — Double Stacked Toroid VFO 2008 web page pulled off. It was substandard and some of the material will re-emerge on a VFO 2011 page this Winter.

Sept 19, 2011 — I updated the SWL essay since it was 6 years old and much has changed with respect to Internet radio. Over time I have received ++ emails expressing different views. I am more a SWL than a Ham and offer just 1 opinion and live by a "each to his own" mantra. Context is everything - this is a radio electronic experimenters site that recognizes SWL'ers are important members of the radio community. Click. Wee Willy web page updated again!

Sept 12, 2011 — Design Center concerning popcorn receiver band-pass filters added. Click.

Aug 26, 2011 — Minor update at the end of Wee Willy DSB transceiver. ***New Junk Box Blog format***. Change is good.

July 30, 2011 — New web page QRP Modules listed on the main menu. Currently under construction.

March 31, 2011 — 2 photos added to the Ugly Construction web page

March 19, 2011 — New content; Hobby and Fun 2011 . I'll slowly add more stuff over the year.

Feb 12, 2011 — New content. Miscellaneous RF Experiments 2011

Dec 29, 2010 — New content. RF Workbench Page 3

Dec 12, 2010 — Final additions to the the 2010 Hobbyist Page added — these concern matching FETs, BJTS and diodes.

Nov 10, 2010 — Some editing and 2 photos added to RF Workbench page 2.

Oct 9, 2010 — This October marks the 12th season of experiments for the site. I have 4 partially completed web pages on the go — pure craziness. I decided to finish them 1 at a time and then add them sequentially. Today, a new web page was added and is 1 of 2 supplements to a future main QRP audio page: Audio Transistor Input Impedance Experiments. The first new "permanent" content since March 2010.

Oct 7, 2010 —The Junk box page lay out was simplified: Accommodating the various modern Web devices plus screen resolutions proved difficult with the old html code. Although less impressive, the new format updates quickly and looks the same on every computer.

Aug 15, 2010 — The cable supplying the Internet connection failed. The site went down for 25 hours. Expect more shut downs over the next couple of days as we resolve any remaining problems.

August 6, 2010 — I decommissioned the historic home page URL http://www.qrp.pops.net/default.htm. This page was a html hard-coded parallel version of the the *correct* home page URL http://www.qrp.pops.net. It was just too much work to continue to update the old (pre-2006) home page in addition to the proper home page.

June 28, 2010 — Web subtitle change : **Amateur Radio Electronic Design** to shorten the name and reflect the site's main purpose.

Effective May 17, 2010 — Schematics now use the 16-bit png format. I have abandoned the 8-bit gif format

May and June 2010: The pops.net net control crew rest up for the next big wind storm. The servers lost power for many hours on May 4 and June 12 due to bad weather. Severe storms arising from the Pacific Ocean threaten our AC power lines each Spring.

March 2010

Wide Range L-C Oscillator added to Hobbyist 2010 page. Link **March 7**

JavaScript Applet **K** added to the QRP Tools page **March 11**

RF Workbench Page 2 added. Link **March 15**

February 2010

JavaScript Applet **J** added to the QRP Tools page **Feb 6**

Editing of the Low Noise Crystal Oscillator web page **Feb 22**

RF Workbench Page 1 added. Link **Feb 18**

January 2010

Experiment #6 added to the Hobbyist page **Jan 2**

New content Low Noise Crystal Oscillator **Jan 10**

2 photos added to the bottom of the Ugly Construction web page. **Jan 12**

1 photo added to the Broadband Transformers page **Jan 12**

New content Hobbyist Page 2010 **Jan 13**

New content: Supplement to JavaScript Applet G which is located on the QRP Tools page **Jan 11**

Additions to the Low Noise Crystal Osc page: the DSO versus CRO essay, plus 5 MHz crystal oscillator added **Jan 31**

JavaScript Applets H and I added to the QRP Tools web page **Jan 31**

December 2009

It has been brought to my attention that the email replies I am sending are not compatible with some of the latest email software such as Thunderbird etc. The reason was that I was using a homebrew email program written over 10 years ago. It is now obsolete. A new email platform is now in place along with a completely **new email address**. Consult the email web page for more information. **Dec 30**

Experiment #5 added to the Hobbyist page FAQ updated and edited. **Dec 30**

Experiment #4 added to the Hobbyist page **Dec 26**

Experiment #3 added to the Hobbyist page **Dec 23**

What does the output of a diode ring mixer look like in your oscilloscope? This has become a FAQ. The question is answered in a contribution by Wes, W7ZOI. Big thanks to Wes for this content. **Dec 22**

I am very pleased to present the **Mike, KL7R Memorial Receiver Experiments**. Click here **Dec 19**

New bulleted list format added to this page to improve readability **Dec 19**

A minor addition was added to the bottom of the Ugly Construction web page. Flux pen photo and text. **Dec 19**

Minor updates to the VFO 2008 "Stacked Toroids" web page under **Epilogue - December 19, 2009**

Update to JavaScript applet **Item E: Calculate Cut off Frequency for an RC Low Pass Filter**. Now has a capacitor range from 0.1 nF to 1500 nF. QRP tools page. **Dec 19**

Editing plus a photo added to the broadband transformer page. This page was improved to support an upcoming project. **Dec 17**

Drafting errors on this schematic corrected (fuse position+ negative rail LED polarity). From this web page. Thanks to Paul, K0EET and Tom for the good eyes and their emails. **Dec 16**

VFO Experiments 2009 updated again; 3 images added. Supplemental web page added and updated **Dec 13**

New JavaScript applet added to QRP Tools page; **Item F: dBm calculator. Dec 12**

Nov 22-27, 2009

1 Hertz Precision Time Base added to the Hobbyist Page.
VFO Experiments 2009 added
The top level menu item "Java Tools" was renamed **QRP Tools.** This menu provides a link to my Webmaster's page and some basic JavaScript applets. Some new material will appear on the Junk Box page during this time.

Nov 20, 2009

Finally; Fall-Winter experiments begin. A small update on the VFO 2008 web page under A 3.5 MHz VFO for Diode Ring Mixers was added.
The RF Preamp page was updated with some content that was first presented on the Junk Box web page. I have received a lot of email about these amps and decided to permanently add them to the site.
The W7ZOI file linked on this page is now in pdf format.

Nov 1, 2009

The Ugly Construction web page was augmented and re-written. 2 new photos were added. This term actually came from Wes and Roger Hayward.
A new Electronic Hobbyist page was added to the drop down 2009 menu.

Oct 25, 2009

Some small updates to the Junk Box page were made. I don't ever think I have been so excited about upcoming Fall and Winter experiments as there are a number of cool, new ideas in my notebook. Extra work and travelling have kept me off the bench, but this will cease in mid November. After that, it's back to the work bench. The Fall-Winter experiments will include some HAM, SWL, general electronics and tube guitar amp experiments. Thanks for your feedback and ideas!

Oct 19, 2009

RSS feed. Click on the **orange RSS icon** above to establish a feed. I will only show 1 item; the latest addition of major new content to the site. Additions to the Junk Box page will occasionally be counted as "new content" and will be included on the feed.

Can anyone guess what brand of beer is in the tool box ? Yes.... its

Oct 12, 2009

Fall weather has come to Western North America! Thanks to Cor, PA3COR for debugging the JavaScript code on this page. Apps number D and E now work in Firefox. Rediscover the fun and learning of scratch homebrew electronics!

RF — Test and Measurement

About My Web Site

OUT
GEN2
GEN1

Welcome Friends!

Introduction

Welcome to the QRP and SWL Homebuilder web site. I write about my experiments with relatively simple and primitive electronic circuits.

Avoiding excessive algebra and obscure parts, I emphasize and show fundamental bench practices.

Through real experiments I examine topics to challenge and intrigue amateur designers — providing examples and describing ways to plan, problem solve, breadboard and measure your circuits.

As amateur experimenters we ought to advance in our hobby; not just perform cookbook electronics. Designing and improving your circuits requires considerable knowledge and effort. Fortunately, others selflessly share their ideas to teach us.

In time, you may recognize your electronics workbench as your greatest teacher. Bench experiments involve us thinking about and measuring our circuits so we know what's happening instead of relying too much on folklore, guessing and copying others. Designing and/or simulating circuits with software can enhance your learning but does not obviate the need to spend time in the trenches with meters, wires and solder.

People often learn skills by modeling others. We need sound examples of how other builders work and think to inform our own designs — inspired, creative and active learning driven by experience and reflection.

At some level, our bench experiences are stories of growth and realization sparked by going and doing. For example, why did the designer choose a particular resistor value? You try different resistor values while measuring the results and increase your knowledge. Collecting schematics, kit building or just thinking do not provide as intimate a learning experience as soldering your ideas on a bare copper board. Talking, tweeting or day dreaming about design is not the same as *doing it*.

Russian novelist Fyodor Dostoyevsky describes the contrast between real life and passivism; "love in action is a harsh and dreadful thing compared to love in dreams", The Brothers Karamazov, Братья Карамазовы. Dare to dream, but better yet, dare to innovate — to design and build *your own* circuits. You may start by just modifying a favorite circuit or scaling a stage to another frequency. We need more innovators and less imitators to grow and sustain our great hobby.

Electronic design produces more than a completed circuit. On the bench, even joy is experiential – a moment of discovery (or several discrete moments) yields more pleasure than stuffing a circuit board or operating a piece of gear. Creativity trumps process every time!

I hope this site demonstrates my passion for building basic, "popcorn" circuits and sharing ideas. Please remember I am just a lay person experimenter and not an electrical engineer.

Regards, Todd, VE7BPO

Essay for 2010

Building or buying test equipment and acquiring a good reference library are important to your experiments. Spurred by the realization that sound bench measurement practices are at the heart of good design, test equipment receives greater focus in 2010 and on.

A reference library is vital to our electronic experiments; good examples lead to better experiments. Poor circuits are everywhere and some builders can't tell a good design from a bad one. Minimalism and simplicity aren't excuses for sloppy design when your goal is to learn. Collecting and sharing well designed circuits helps us avoid wasting time and experiencing frustration. Circuits with attributes like well defined input or output impedances, low noise or harmonic distortion are desirable to fuel experimentation. Look for better quality circuit examples in 2010 and on.

The Internet is changing how we read and write. The prevalence of small portable web devices such as iPhones, ever increasing numbers of web sites and blogs, and the use of search engines create fierce competition among sites. Modern sites attract your attention with varied visual, aural and textural media and unfortunately, hype and pseudo-journalism. Narrative writing is more skimmed than read. Brief is in — bullets, subtitles, lists and graphics replace long lines of narrative prose which no one seems to have time for anymore.

Have you noticed the changes on this website? New content still contains lots of narrative writing, but assumes an active voice, with emphasis on brevity, clarity and speaking directly to you, the reader. Sharing mostly obsolete, analog 1970's-style circuits, QRP/SWL HomeBuilder attracts a tiny, niche audience. I believe the success of this website depends on providing good and diverse content — not Tags, RS feeds, adopting net-speak, or self promotion. You be the judge.

Essay for 2014

The Internet of Everything?

Bucking the trend, my contribution to amateur RF homebrew remains informational and not social. Why?

Social media information represents a Pandora’s Box of good science and opinion, mediocre thought, or trash potentially created and/or disseminated by anyone who’s connected. We accept that much of our social media content doesn’t come from the best or brightest — some people are just plain interesting, or express themselves vigorously, seem like-minded, or touch our hearts.

Some builders, like me, seek objectivity and not just “likes” and “follows” based on sentiment and spectacle. While a few radio builders may prefer to join hands and sing Kumbaya, or pat themselves and others on the back simply because their breadboard actually works, a trifling of us care more about how and why our circuits work. We *like* measures and measurement tools and *follow* science, experiments and the works of those who shine brightly.

What’s wrong with plain information, unfettered discovery, experimental rigor, objectivity and rational, kind thought?

It’s not that these characteristics don’t attribute social web clients — they do, but the negative impacts of social media worry me a little. A brief list of concerns: loss of privacy, the threat of wasting time while really just isolating ourselves from our real friends + family. The numbing exposure to the Internet of Ads and Spam. Still too, bubbling up like purulent sores come the charlatans, the misinformation peddlers, the opinion spammers, and those who anonymously leave stinging sarcasm, or outright hatred [hostile online comments that attack people, or divert a healthy flow of ideas].

Running a low-tech web site with nearly 0 commercials suits me better.

SEO — Search Engine Optimization

I’ve read that Google analyses your web site content, the number and quality of the sites that link to your pages, their search engine clicks and so forth. In part, Google seems to rank a site based on how relevant and authoritative they believe it is. Some people specifically employ SEO techniques to gather in more traffic.
To my surprise, each year, tens of thousands come to this site via search engines like Google. I don’t think my material seems too relevant or authoritative. I’ve made no effort at SEO, so I conclude that you, my readers have more to do with the site’s success than anything I’ve ever done.

Thank you.

Hope Invigorates

Invigorated by the excellent work either emailed to me by experimenters such as Michel F6FEO, or Dick, N4HAY; or posted on blogs or Community sites like Yahoo, I feel hopeful about the future of our hobby in 2014. The PHSNA Yahoo group leaders, the recent work of Jason, NT7S, Steve VE7SL and many others show that *amateur design* experiments still have a pulse + respirations.
The aforementioned get my vote for their MOF like behavior: a strong blend of creativity, tradition and quality.

Looking Ahead — Future Site Content

Most of my new receiver work involves quadrature and in-phase mixers fed with (2) local oscillators; 1 output shifted 90 degrees from the other — essentially, EMRFD Chapter 9. Even my Funster [a personal, lowbrow trans-receiver I drag onto hill and dale] now contains phasing receiver circuitry to reduce the opposite sideband by 20 dB along with further low-pass filtering. I hope to add some Funster content to HF Embarcadero in Winter 2014. Like many of you, the 1 resource I lack the most is time.
While I'm thrilled with the notion of a receiver appliance that contains just an antenna, LNA, ADC and some sort of "wonderfall" display or speaker, I'm still smitten by analog design with hardware. Still, the I-Q mixer will offer a nice transition into SDR should I ever wish to spend my free time writing C# and not melting solder.

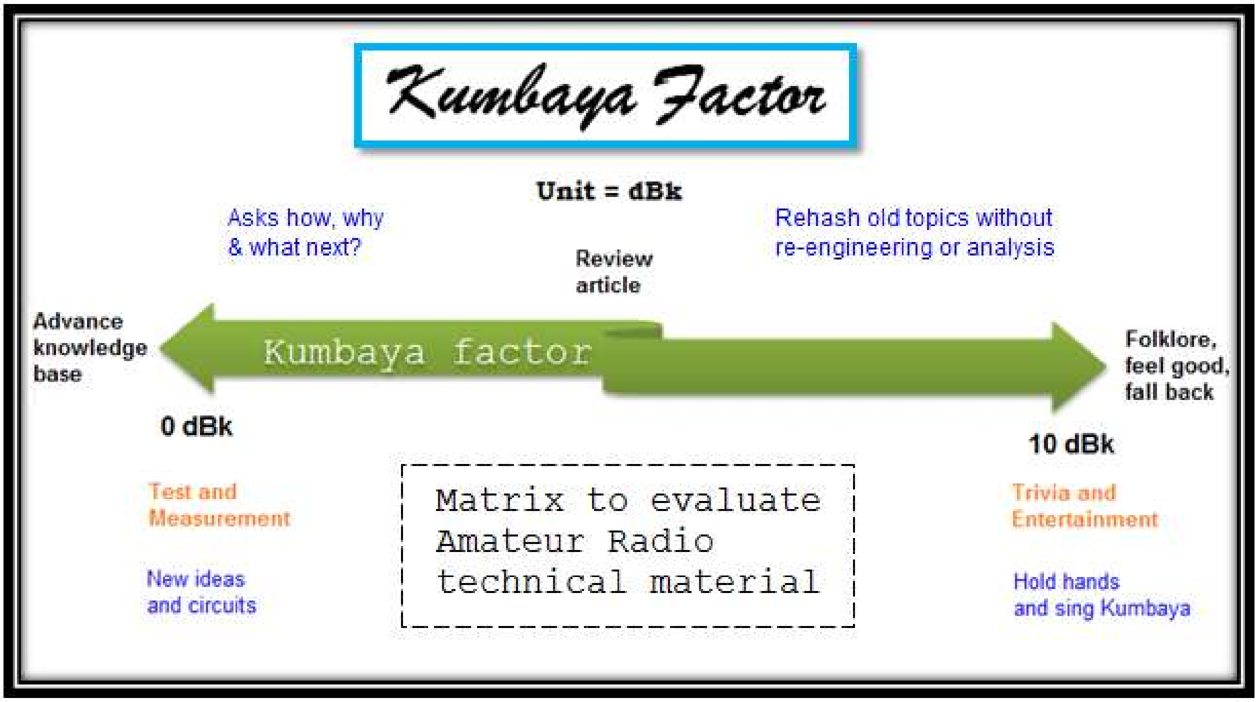
Kumbaya Factor
Unit = dBk
Asks how, why & what next?
Review article
Rehash old topics without re-engineering or analysis
Advance knowledge base
Kumbaya factor
Folklore, feel good, fall back
0 dBk
10 dBk
Test and Measurement
New ideas and circuits
Matrix to evaluate Amateur Radio technical material
Trivia and Entertainment
Hold hands and sing Kumbaya

Best to you!

Miscellaneous

My special thanks to Wes Hayward, W7ZOI for his generous support and elmering over many years.

EMRFD is the main reference of my site

All permanant content circuits were built and tested. Schematics are drafted as carefully as possible. Please accept that bench and/or drafting errors may occur. No liability arising from the application, use, or misuse of these projects that results in direct or indirect damage or loss is assumed.

Full price is paid for all parts used and no monies are or were received for promoting any products or companies on this web site. Any ads, hyperlinks or mention of commercial products or companies is out of courtesy only.

"Until you build and measure it, you don't know what you don't know";
Rick Campbell, KK7B; VHF Open Sources — Design of Low Power High-Stability Low Phase Noise Single Frequency VHF Sources with High Spectral Purity; 2008

Information Regarding the Compression of Schematics

I see many electronics web authors compressing black and white schematics as jpg files. This results in distortion of the schematic. Schematics are best compressed using the 8-bit, lossless LZW algorithm which means converting the file to a png, gif, or even pdf format. The files sizes will typically be smaller than .jpg compression, have no distortion and can be edited easily.

- RAC is the National Amateur Radio Society of Canada
- For my web page concerning support of the **Radio Amateurs of Canada**, please click here

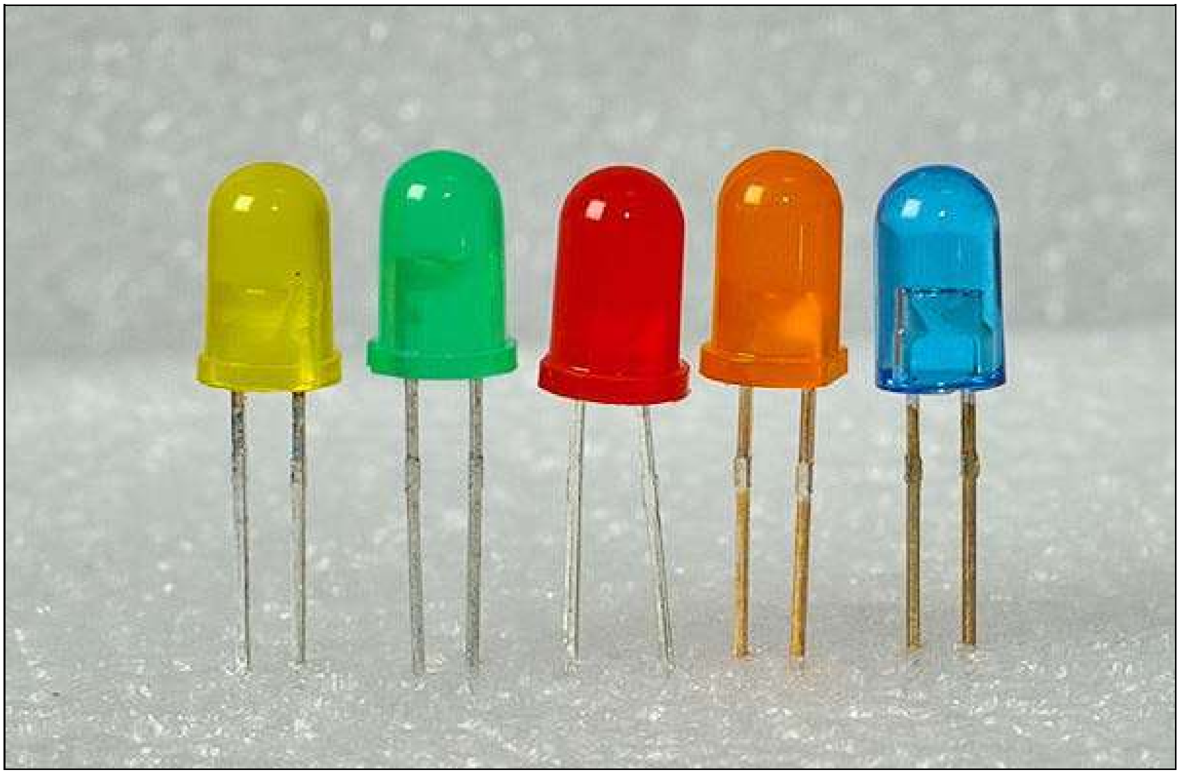

The hand drawn image bitmaps on this web site (logo etc.) are by Rod Adams. All website photographs were taken by VE7BPO except as indicated.

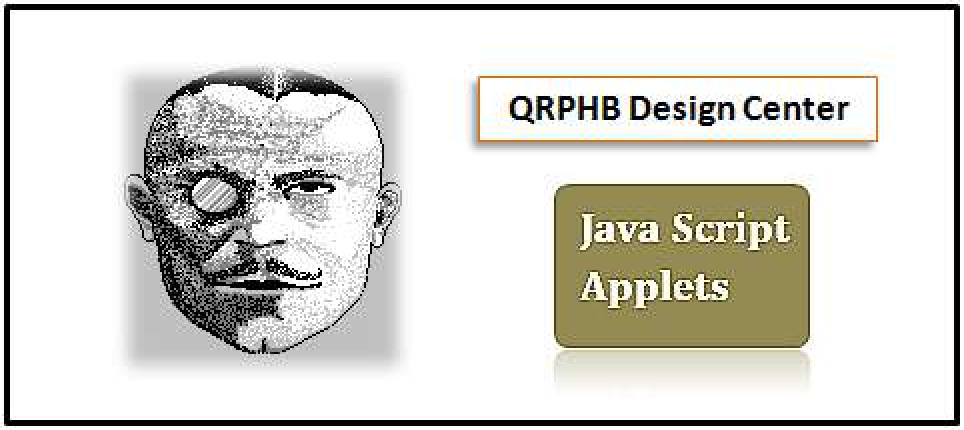
QRPHB Design Center
Java Script Applets

1. Click for the old QRP HomeBuilder Graphics page. Click for my Pin Outs page. Click for my Homepage

---

2. Some simple tools written in JavaScript for the QRP/SWL HomeBuilder:
`Minimal input error checking`

A. Calculate DC Voltage Divider Bias

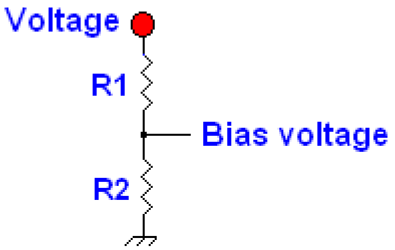
Voltage
R1
Bias voltage
R2

Enter Voltage: Enter R1: Enter R2: Bias voltage =

B. Calculate Inductive Reactance

Enter Inductance in uH: Enter Freq in MHz: XL (ohms) =

C. Calculate DC Current for a Current Mirror

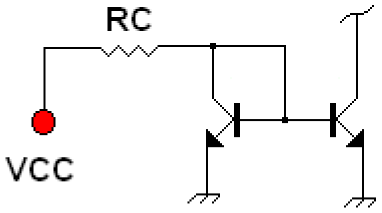

Enter VCC: Enter RC in ohms Current (mA) =

D. Calculate # of Turns To Obtain a Desired Inductance on a Ferrite Torroid

Enter Inductance in millihenries: Select Core: FT37-43 Turns =

The *AL for this ferrite core is* =

E. Calculate Cut off Frequency for an RC Low Pass Filter

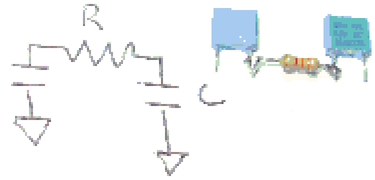

Enter resistor value in ohms: Select capacitor values in uF: 0.01 3 dB down frequency (Hertz) =

F. Calculate Power in dBm and mW from Peak to Peak Voltage

Enter measured peak-to-peak voltage into a 50 ohm load: dBm, mW

Application Note: This web site follows the EMRFD standard for dBm power measurement. dBm = the power delivered into a 50 Ohm resistive load which is temporarily substituted at that point in the signal chain.

G. Calculate Lm and Cm For a Crystal using the G3UUR Method

Enter frequency in **MHz** written on crystal (series resonant frequency):

Enter measured frequency in **MHz** with switch open: Enter measured frequency in **MHz** with switch thrown:

Enter crystal capacitance in **pF**: Enter open switch circuit capacitance in **pF**:

Cm = femto Farads , **Lm** = Henries

Supplemental web page for this applet: Crystal filter measurement and adjustment Link

---

H. Calculate Decibel Power Gain or Loss from 2 Peak-peak Voltages

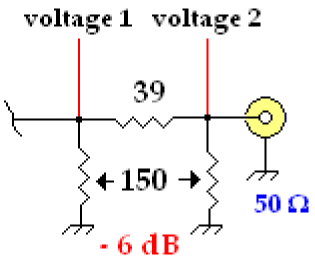

Enter voltage 1: Enter voltage 2 dB gain or loss =

I. Calculate Decibel Power Gain or Loss from Input and Output power

Enter input power in watts: Enter output power in watts dB gain or loss =

J. Calculate dBm and mW from RMS Voltage (50 ohms)

Enter RMS voltage dBm mW

K. Calculate Return Loss and VSWR (50 Ohms Detector)

Enter the detector signal in pk-pk volts when the unknown port is terminated in an open circuit: Enter the detector signal in pk-pk volts when the unknown port is terminated in the unknown impedance:

Return loss = dB, VSWR=1:

Measurements per Figure 7.41 EMRFD. Schematic here

L. Calculate Power from the DC Output of an AD8307 Meter

1. Linear calibration steps:

Enter measured DVM voltage at -10 dBm: Enter measured DVM voltage at -20 dBm:

2. Calculate power in dBm from DVM voltage:

Enter measured DVM voltage: Power =

M. L-C-C Tee Network

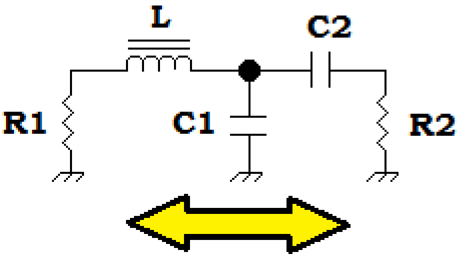

Enter **frequency** in MHz:

Enter **R1** in Ω: Enter **R2** in Ω: *R1 must be < R2, but the network is bi-directional*

Enter Q: *Perhaps start with 2-5*

C1 = pF, C2 = pF, L = uH

N. Parallel Resistor Values (2-4 resistors)

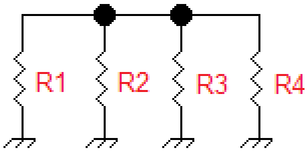

Enter R1: Enter R2 Enter R3: Enter R4:

R = ohms

O. LCR -- a Reactance Calculator In beta -- do not use !

Mode: L C to Reactance Inductance: uH Enter frequency in MHz

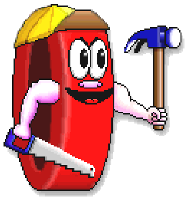

This page last updated: September 21, 2013

SWL HOMEBUILDER

RF — Test and Measurement

Short Wave Listening коротковолновое радио

Introduction to Short and Medium Wave Radio Listening

Short wave radio listening was a childhood passion and I enjoy being an SWL just as much today and log at least 800 hours of SWL per year. There seems to be many web pages devoted to construction of radio equipment for the amateur radio experimenter but relatively few for the shortwave radio devotee. I decided to expand this web site to include projects for the SWL Homebuilder in 2005.

My favorite bands are 49 meters (5.9 - 6.2 MHz) at night-time and 19 meters (15.1-15.8 MHz) during the daylight hours. I also listen to medium wave DX around 1400 - 1600 KHz.

Why Listen to Analog Short Wave Radio?

Is analog short wave radio dead? I think not.

No doubt, short wave radio has passed its prime and is slowly dying, however, it's still fun and/or relevant to some.

World band radio: Almost 1/2 of the world's population lives on $2.00 *USD* or less per day. The Internet (the main alternative to shortwave radio) poses a luxury to many poor people living in lower-income countries — experienced travelers or those who support people in developing countries will understand this statement. In some countries now, ranking in the middle class just means you have a full-time job. In addition, oppressive governmental regimes may limit foreign media and Internet access: LW, MW and SW radio can break through obstacles such as natural or man-made disasters, borders, poverty and censorship.

For SWL hobbyists, analog shortwave radio entertains, informs and best of all, provides opportunities to analyze propagation and experiment with real radio topics including static, solar flares, QRN, antennas, grounding, baluns, coax, and wire. SWLing poses an adventure — it's unpredictable, challenging and increasingly difficult as stations decrease and QRN increases. I've built many antennas and even some noise cancelling circuits just to pull in a few Dx stations. The sport of SWL lies in making DX contacts: a theme shared with Ham radio.

What About Internet Radio?

Radio by definition is the transmission and reception of electromagnetic waves of **radio frequency**; but perhaps blue-tooth or Wi-Fi reception from a hot spot qualifies as *radio* in the modern era? Just as peanuts aren't nuts, Internet server or webcasted radio is not RF broadcasted radio. I think Internet radio is great, but fundamentally a very different medium from that enjoyed by SWL fans

Internet radio involves a radio player decoding a stream of compressed bits fed from a Internet radio station or virtual receiver. In some cases, the material originates from a real radio station that also broadcasts an AM or FM signal. For example, you can tune FM station Rooskie Radio "Русское Радио" in much of Slavik Europe or play them on a computer device anywhere you can get an Internet connection.

For lovers of foreign content, listening to Internet radio makes sense; providing convenience, a good signal when bandwidth is high and 24 hour per day listening on 1 IP address. Internet radio offers a much cheaper way for content providers to beam their news and music services around the globe — we've seen numerous large broadcast radio services such as the BBC World Service reduce or drop analog SW and add Internet radio, satellite and digital SW transmissions for their customers.

The exciting growth of independent and niche Internet radio stations increases personal freedom of choice and provides opportunities for unique interest providers and consumers to find each. Media streaming companies and manufacturers of Internet radio *players* and their worldwide distributors benefit too.

This technology is a far cry from tuning the SW bands with a homebrew or commercial radio frequency *receiver* and a length of wire slung in a tree. Perhaps, the greatest advantages of Internet radio are that you don't have to get up early, or stay up late to pull in some rare Dx, nor do you need any radio skills or special equipment — perfect for the majority of listeners. But we're SWL radio hobbyists: people who listen for both content and because we love radio propagation and gear.

There is nothing wrong with Internet radio, or any of the modern data streaming techniques however, SWL aficionados driven by skill, the thrill of Dx and love of their experimental hobby share a special bond that Internet radio doesn't give them.

ICOM
IC-R75
COMMUNICATIONS RECEIVER
10.000.00
PHONES
POWER
AF-Ø-RF/SQL
TWIN PBT
SSB
CW/RTTY
FIL
AM
FM
TS
V/M
MW
CLR
SEL
SCAN
ENT
CLOCK
ATT
NR
AGC

Assembling a station The most important component in your radio shack is your antenna. Don't hesitate to safely experiment with the many antenna designs available on the world wide web. Your sure to find a commercial unit or home brew antenna design that suits your real estate and budget.

Your next task is to find a receiver. It is difficult to recommend any one receiver because there are so many excellent commercial receivers to choose from. If you are thinking about purchasing a used receiver, you might consider checking eBay to find a receiver or to learn the going price for used gear. The ultimate SWL experience in my opinion is to build and operate a receiver on at least 1 band.

Favorite SWL and SWL-related Web Sites

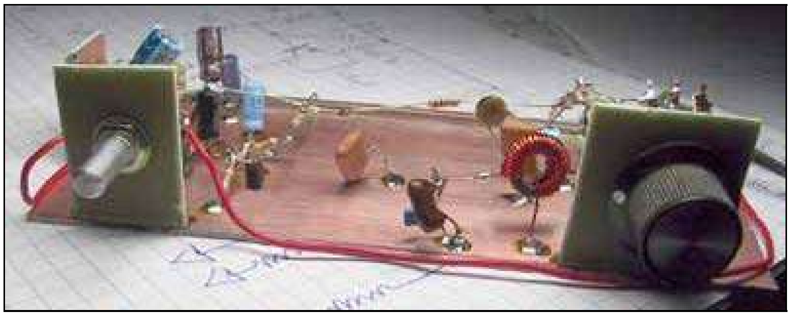

Wikipedia-Shortwave Bands A good description of the bands and their general propagation.

Canada's SWL-DXer website Hard core Canadian web site dedicated to SWL. Thanks gentlemen!

http://www.bobsamerica.com/swl

http://www.dxing.info/

Doug's Shortwave Radio Page

AA6V's SWL Links

Method for soldering a PL-259 to RG-213 or RG-8

Digital Modes For SWL Fans

PSK31

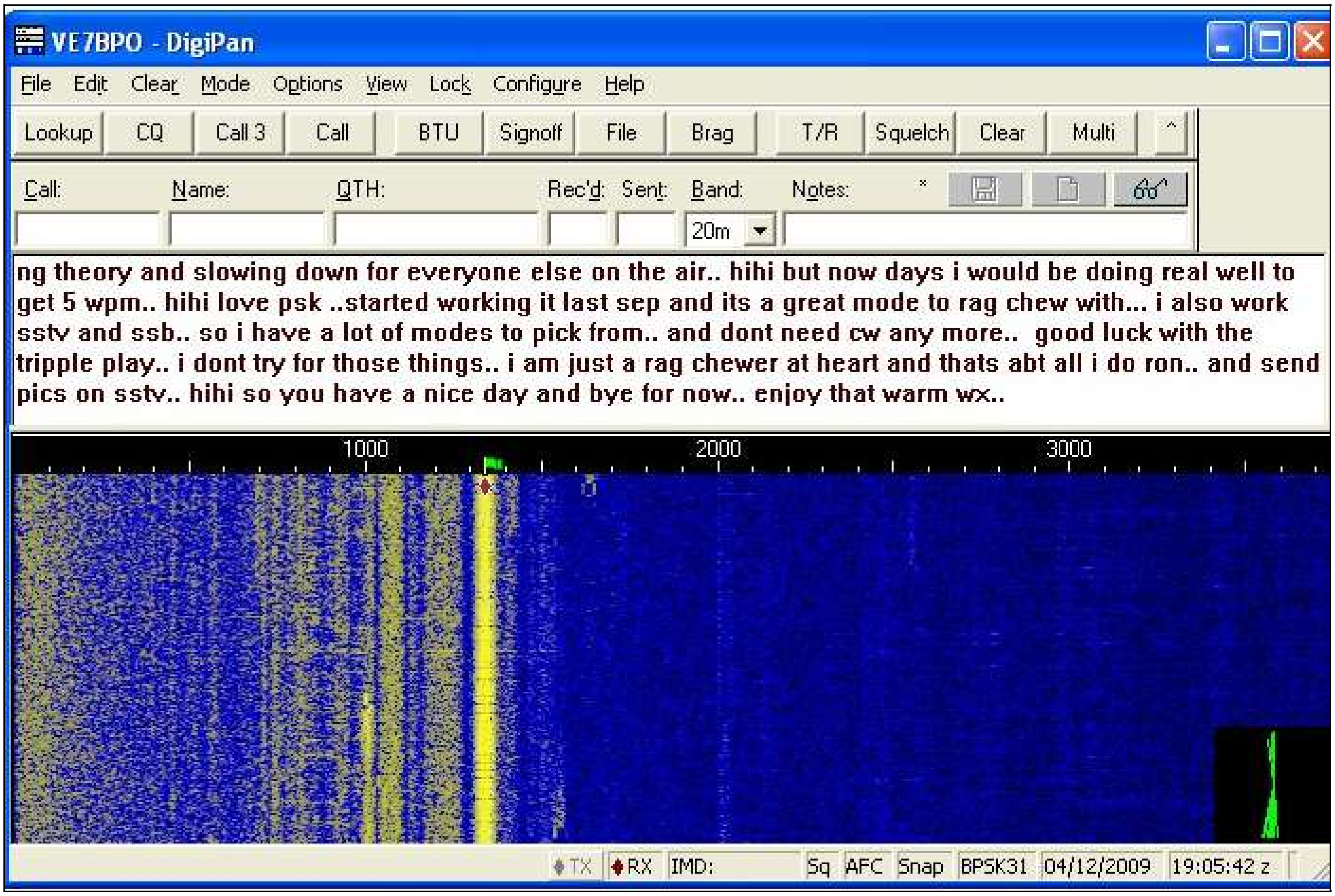

VE7BPO - DigiPan
File Edit Clear Mode Options View Lock Configure Help
Lookup CQ Call 3 Call BTU Signoff File Brag T/R Squelch Clear Multi
Call: Name: QTH: Rec'd: Sent: Band: Notes:
1000 2000 3000
TX RX IMD: Sq AFC Snap

Band: 20m

ng theory and slowing down for everyone else on the air.. hihi but now days i would be doing real well to get 5 wpm.. hihi love psk ..started working it last sep and its a great mode to rag chew with... i also work sstv and ssb.. so i have a lot of modes to pick from.. and dont need cw any more.. good luck with the tripple play.. i dont try for those things.. i am just a rag chewer at heart and thats abt all i do ron.. and send pics on sstv.. hihi so you have a nice day and bye for now.. enjoy that warm wx..

BPSK31 04/12/2009 19:05:42 z

There are a number of good sites about this relatively new HAM mode. All that you minimally need is a receiver dialed in at 14.070 USB (or another

PSK31 frequency), a microphone hooked to your computer sound card and some free software. The software (DigiPan 2.0) is available at http://www.digipan.net/ .
I use a USB interfaced microphone and place it about 2 cm from my receiver speaker. If HAMs are operating; you should hear some warbles and see some waterfalls on your screen. Click on one of the waterfalls to begin receiving the text. I knew nothing about this mode, but was up and running in 15 minutes.

Additional Short Wave and Medium Wave Receiver Photos

OPTIMUS
FM
MW
LW
SW
MHz
kHz

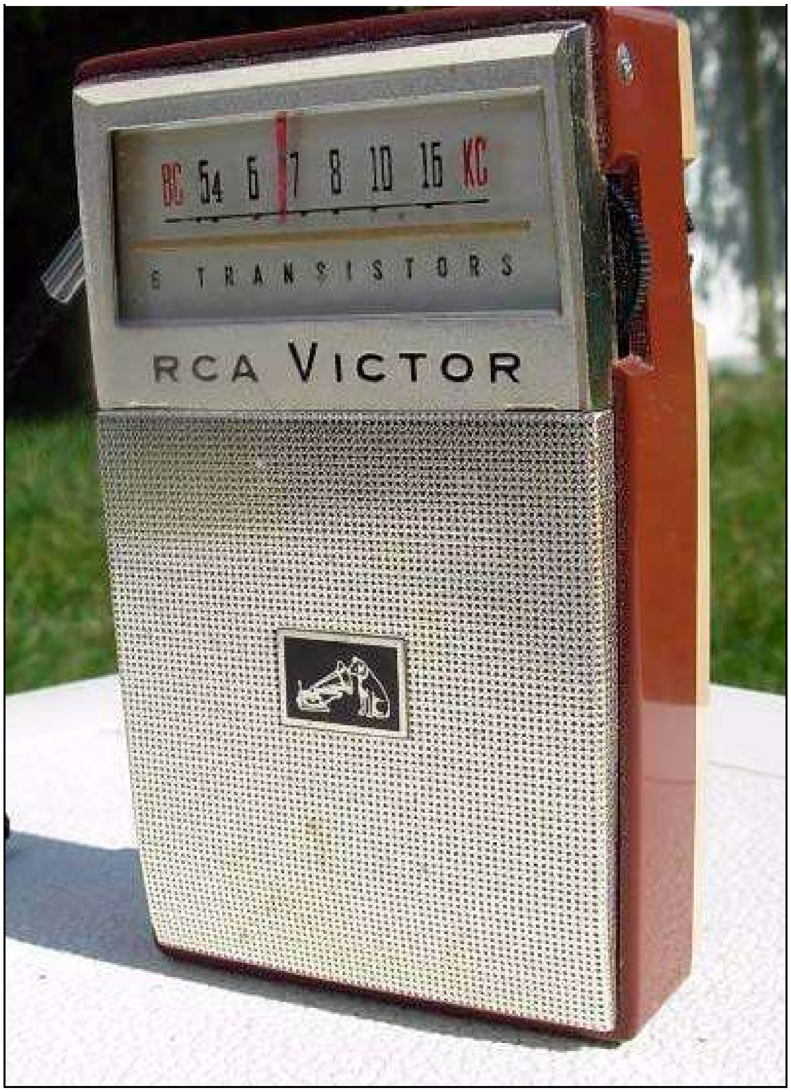
BC 54 6 7 8 10 16 KC
6 TRANSISTORS
RCA VICTOR

RCA VICTOR
SISTOR
DIO
MODEL GP-526
RCA VICTOR
RCA VICTOR

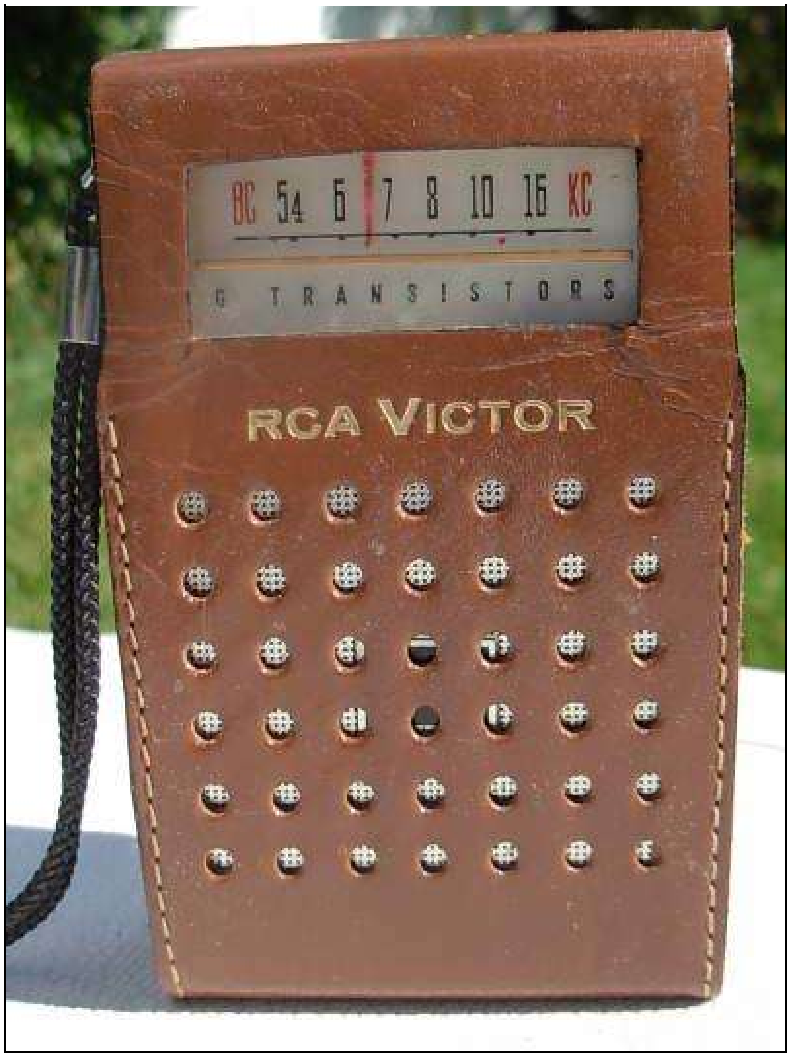
BC 54 6 7 8 10 16 KC
TRANSISTORS
RCA VICTOR

RF Preamps

W1FB 6M RF Preamp

Discussion:

Here is a schematic sent to me by W1FB many years ago. It is very similar to a 6M two-stage preamp that he published in QST in the mid eighties. Doug really favored the grounded gate FET for narrow band preamps. His published work is replete with examples of them on just about every band. I built that amp and remember getting about 10 dB gain, which is all that I wanted for the 6M direct conversion receiver using a diode ring detector that I was building. The great feature of the amp is that it combines a band pass filter and preamp in one. I lost the original schematic that Doug sent me but was delighted to see that I made a bitmapped drawing of it on a floppy disk that was recently re-discovered when we were moving an old desk. The shield shown in the schematic was a small piece of grounded ,double sided PC board in which, I made a small chamfered hole in to pass the lead going to the T2 tap. The shield, along with very short component leads will help minimize parasitic oscillations. The T2 tap is 3 turns down from the end of the T2 main winding that connects to the variable capacitor. Doug specified T37-10 cores for the inductors, but I substituted T37-6 cores and used the same number of windings as specified for the T37-10 core inductors. It worked fine.

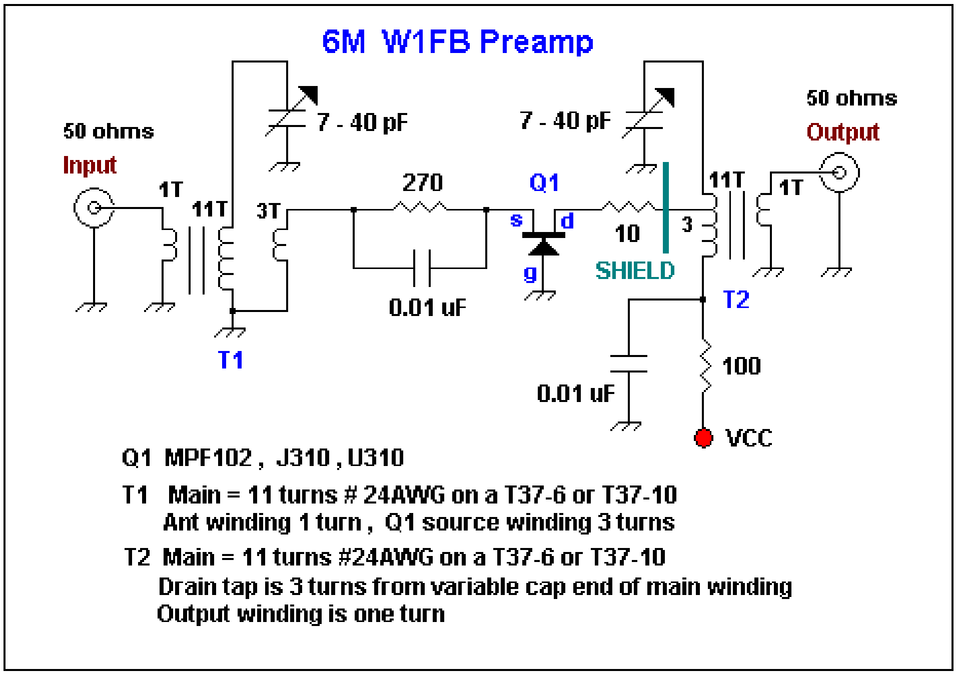

Q1 MPF102 , J310 , U310

T1 Main = 11 turns # 24AWG on a T37-6 or T37-10
Ant winding 1 turn , Q1 source winding 3 turns

T2 Main = 11 turns #24AWG on a T37-6 or T37-10
Drain tap is 3 turns from variable cap end of main winding
Output winding is one turn

VE7GC Popcorn RF Preamp

Discussion:

Here is an easy RF preamp by Dick Pattinson, VE7GC. It uses a single tuned circuit at the front end and can connect directly to a mixer or product detector in a simple receiver project. Note how Dick provided adjustable RF gain control for this circuit in his Wee Willy project on this website. If you can not find Tak Lee green 10.7 MHz IF coils, probably any other brand of 10.7 MHz slug tuned IF transformer would work. The Mouser catalog number is 421F123 . If your 10.7 MHz IF coil has a built in capacitor at the base , remove it. A fixed inductor may also be wound using a powdered iron torroid core and then all or a portion of the C1 capacity would be made variable. The input impedance is 50 ohms and the output impedance is low due to the Q2 emitter follower stage.

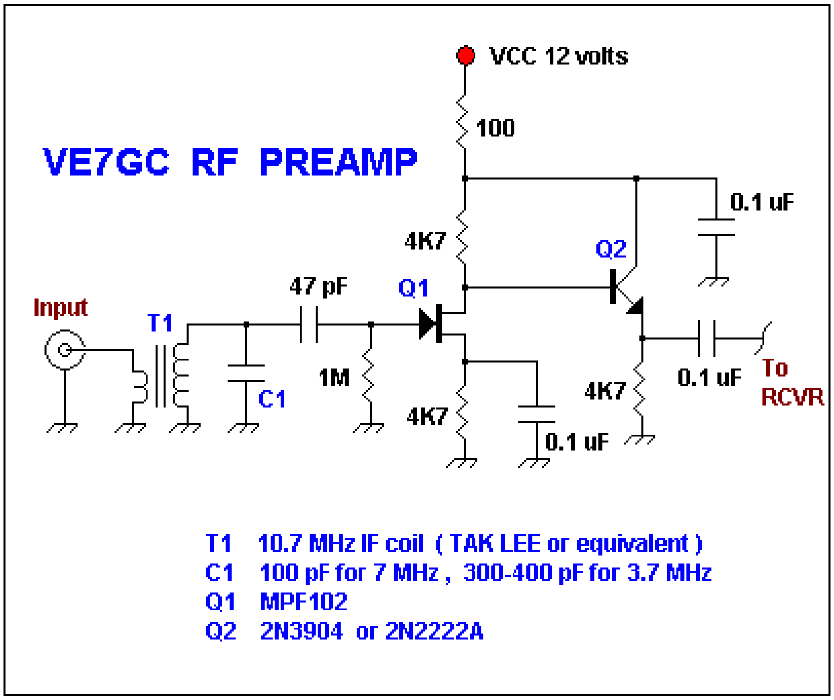
VE7GC RF PREAMP
VCC 12 volts
100
0.1 uF
4K7
Q2
47 pF
Q1
Input
T1
1M
C1
4K7
0.1 uF
4K7
0.1 uF
To RCVR
T1 10.7 MHz IF coil (TAK LEE or equivalent)
C1 100 pF for 7 MHz , 300-400 pF for 3.7 MHz
Q1 MPF102
Q2 2N3904 or 2N2222A

A Low Noise, High Dynamic Range Broadband RF Amp

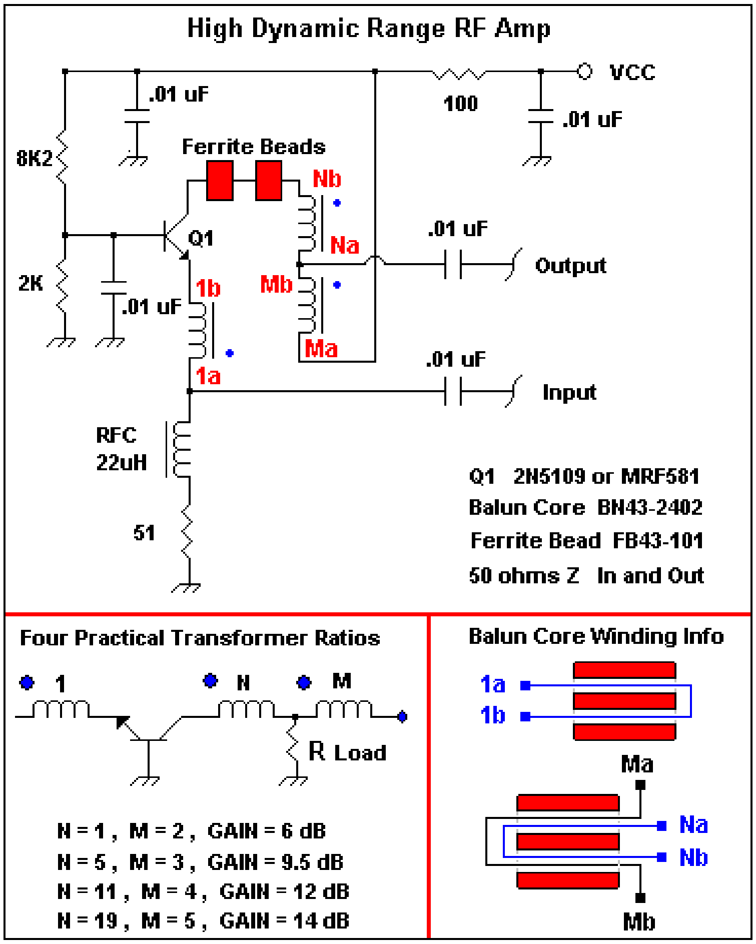
High Dynamic Range RF Amp
.01 uF
100
VCC
.01 uF
8K2
Ferrite Beads
Nb
Q1
Na
.01 uF
Output
2K
1b
Mb
.01 uF
Ma
1a
.01 uF
Input
RFC 22uH
51
Q1 2N5109 or MRF581
Balun Core BN43-2402
Ferrite Bead FB43-101
50 ohms Z In and Out
Four Practical Transformer Ratios
1
N
M
R Load
N = 1 , M = 2 , GAIN = 6 dB
N = 5 , M = 3 , GAIN = 9.5 dB
N = 11 , M = 4 , GAIN = 12 dB
N = 19 , M = 5 , GAIN = 14 dB
Balun Core Winding Info
1a
1b
Ma
Na
Nb
Mb

Discussion:

This schematic is a version of a circuit developed and patented by David Norton and Allen Podell in June 1974. This variation was described by Joe Reisert, W1JR in the now defunct Ham Radio Magazine. The Norton design uses transformer coupling to achieve "noiseless negative feedback" and is really outstanding.
A great article utilizing and augmenting on this technique receivers is by Jacob Makhinson, N6NWP in QST magazine for Feb 1993 with "A High Dynamic Range MF/HF Receiver Front End". Makhinson arranged 2 in push-pull to obtain excellent results. Obtain a back-issue of QST for closer study. **Note** that the fore mentioned Feb QST article has the coil phasing wrong and the correct phasing can be seen at this web site from QST for July 1996. There is also information about Norton feedback RF amplifiers in EMFRD.

If you are building a contest-grade receiver and need a good RF preamp and/or post mixer amplifier, the Norton type is quite suitable. An amp built using a 2N5109 can have a noise figure in the 2.5 - 3dB range. I have also built them with 2N3866, MRF517, MRF581 and a 2N5179 although the last transistor would be a somewhat poorer choice. This schematic with a 2N5109 is good from 1.8 to 150 MHz with a 1.2:1 VSWR or less according to Joe Reisert. I have even put one in a friends CB radio and he was delighted.

Winding and Construction Hints

Making the Norton amps requires some planning to keep all component leads as short as possible. The transistor leads and any connecting components should be trimmed as short as practical to promote stability. Sketch the component layout on a piece of paper and modify it until you are satisfied you have designed a good layout. The ferrite beads on the transistor collector aid in stability and should be used to preserve the noise figure by squashing any oscillations should they develop. The 22 uH choke can be the little epoxy coated units that are color coded and look somewhat like resistors. Do not use a choke less than 22 uH.

Before winding, the builder must first decide how much gain is needed from the amp. For an RF preamp, the stage should have gain equal to or greater than the passive stages after it. Also there will be losses in the transformer, so the theoretical gain of the Norton amp maybe 1 dB off and will need to be factored in. For the purposes of discussion, a 9.5 dB amp is desired , so N = 5 and M = 3. The first step is to mark one side of the core with a dab of liquid paper, paint or a small piece of tape. This will allow you to keep track of the transformer later. To mark, hold the core so that both channels are parallel to the floor, one on top of the other. Apply your dab of paint to the top of the core and use the marked top to denote the A windings. 1a, Ma and Na will all start from the top channel in the balun core. Using 32 AWG wire for all three windings, start with winding 1 and wind the single turn from point 1a to 1b. Cut off the leads so they are shorter than 5 centimeters (2 inches). Next, wind Ma to Mb three complete turns through the binocular core and trim the leads if needed. Tie a small knot in the wire at both ends. This will clearly mark this M winding. Both windings should look like the diagram under the schematic. 1a to 1b are on the left of the balun core and winding Ma to Mb are on the right side of the core. Mb has a distinguishing knot at the tip of both wire ends. Ma starts from the top of the core which you have marked with a dab of paint or something. Finally, wind Na to Nb five complete turns through the core in the same direction as the previous winding M. Strip wires Na and Mb (Mb has the knot), twist together and solder. Scrape the enamel off the leads **very** gently with a sharp hobbyist knife.

Insert the transformer in your circuit and cut the leads to their proper length and then solder away. It maybe preferable to pre-strip the leads on winding 1 as it is hard to strip the enamel off a fine wire that has only one turn and it may accidentally pull out of the core. If it does, just re-insert it into the balun core on the correct side. Once you have soldered Na and Mb you can always identify the windings later because you have marked the top of the balun core which denotes the A windings. Try and make your windings gently tight as if there is too much slack you may have difficulty getting the last few windings thru the core channels. A 14 dB gain amp maybe **impossible** to wind with 32 AWG wire, it may best to use 34 AWG for that amplifier. I have never built one for greater than 12 dB. The transformers are a bit tedious to wind, however persevere and the results will be well worth it. For HF, you can substitute 0.1 uF caps for the 0.01 caps shown if you like.

Toroidal Inductor Norton Amp Experiments

Discussion:

The amp shown in the schematic to the right uses a ferrite torroid for the transformer and has ~10 dB gain. Winding 1 turn of wire over the cold end as shown in the schematic is tricky. Try to keep this link as short as possible. A ferrite bead or a 22-51 ohm resistor on the transistor collector is desirable. You can try increasing the turns (1:21:5 etc) to experimentally obtain more gain from this amp. The torroid version is a valid option for builders who do not have balun (binocular) core ferrites in their junk box. Toroidal inductors are certainly easier to wind then binocular core versions.

In 2007, I built several Norton "noiseless feedback" RF amps using FT50-43 and FT37-43 ferrite toroids. These are outstanding and I recommend using them in projects. The input and output Z is 50 ohms. The overall BJT topology is reminiscent of a common base amp. I have some basic information concerning this amp on this web page . They are straight forward to build. The biggest problem is the phasing of the single turn link. Get it wrong and your amp can turn into an oscillator.

Figure 1

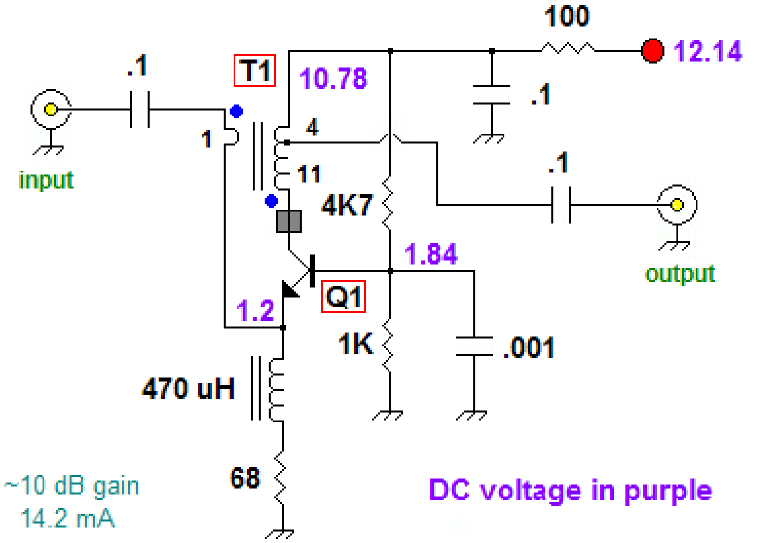

Q1 = 2N5109 or NTE 123 etc
T1 = Secondary: 15 turns with a tap a 4
turns from the cold end.
Primary: 1 link over cold end
T1: use FT50-43 or FT37-43 ferrite

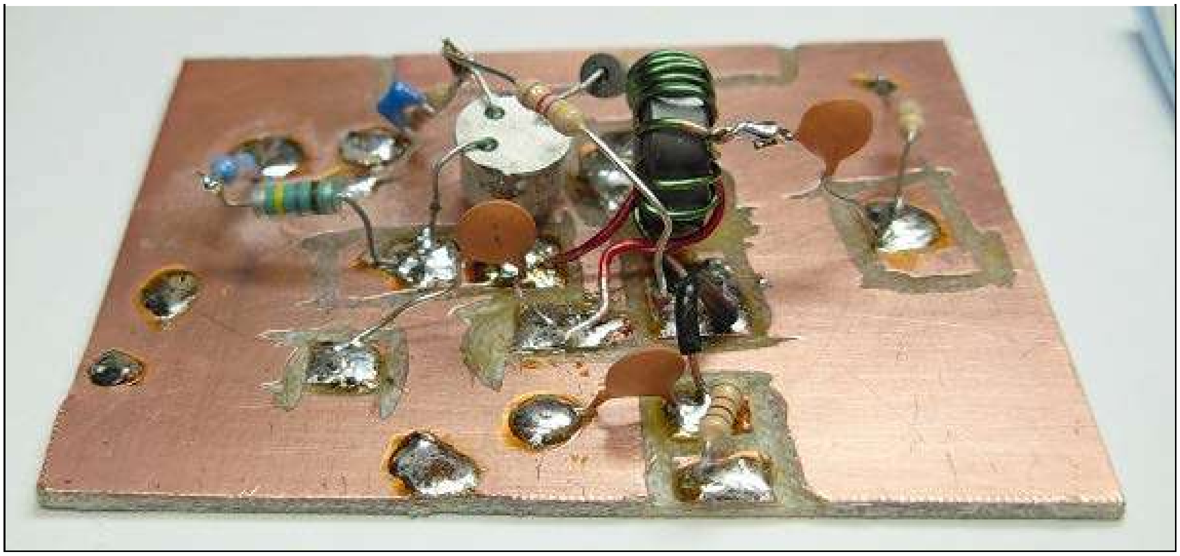

Shown above. The breadboard of Figure 1.

Figure 2

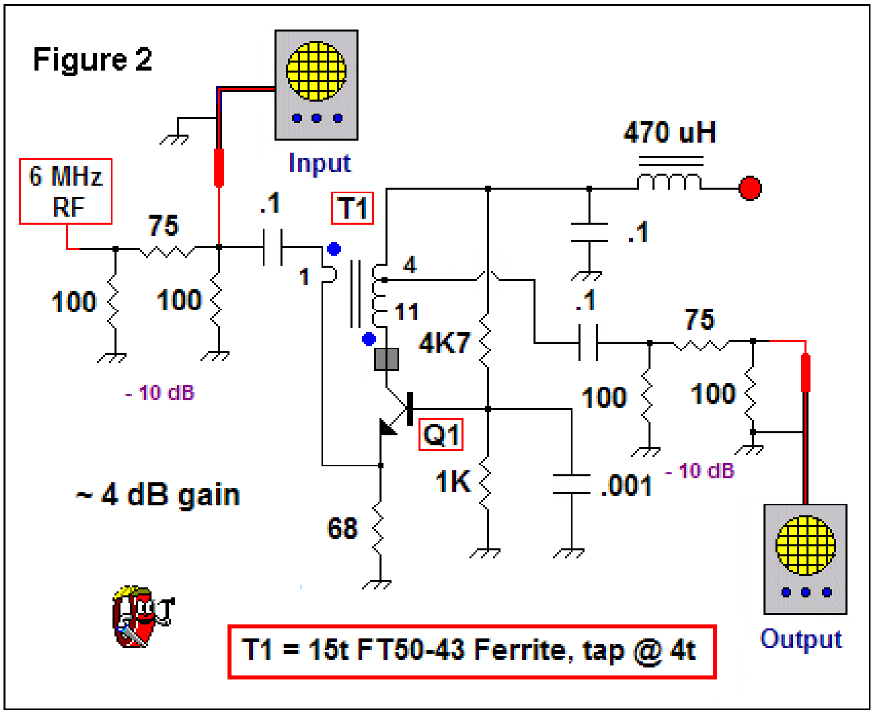
Input
6 MHz RF
75
.1
T1
470 uH
.1
100
100
1
4
11
4K7
.1
75
- 10 dB
100
100
Q1
~ 4 dB gain
1K
.001
- 10 dB
68
Output
T1 = 15t FT50-43 Ferrite, tap @ 4t

Shown above is the Figure 1 amp above (labeled Figure 2) with a 50 ohm -10 dB pad on the input and output, so gain is low. I used these pads to evaluate the amp in a number of experiments. I never got around to writing up these experiments on the web site and likely never will. I wish I had more time as my notebooks are full of unpublished experiments that would be great content for this web site.

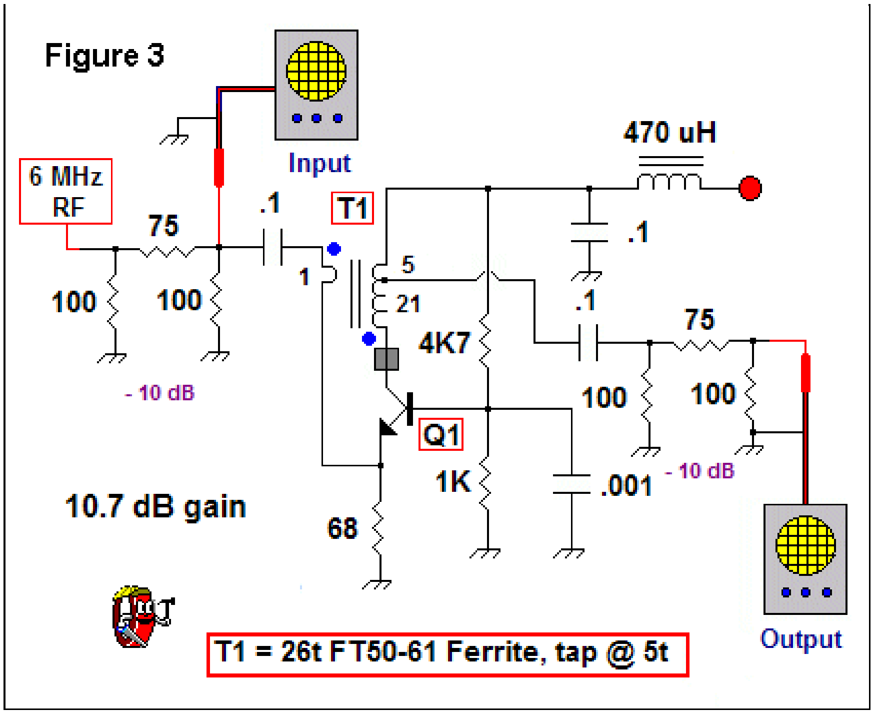
Figure 3
Input
6 MHz RF
75
.1
T1
470 uH
.1
1
5
21
100
100
4K7
.1
75
- 10 dB
100
100
Q1
1K
.001
- 10 dB
10.7 dB gain
68
Output
T1 = 26t FT50-61 Ferrite, tap @ 5t

The amp above (labeled Figure 3) is a hot one; 10.7 dB gain even with 10 dB of attenuation. You can leave off the input pad and decrease the output pad to -6dB if you want or require a wide band, low noise RF amp with lots of guts. Most builders use binocular ferrite cores for the inductor, but torroids work fine for many applications.

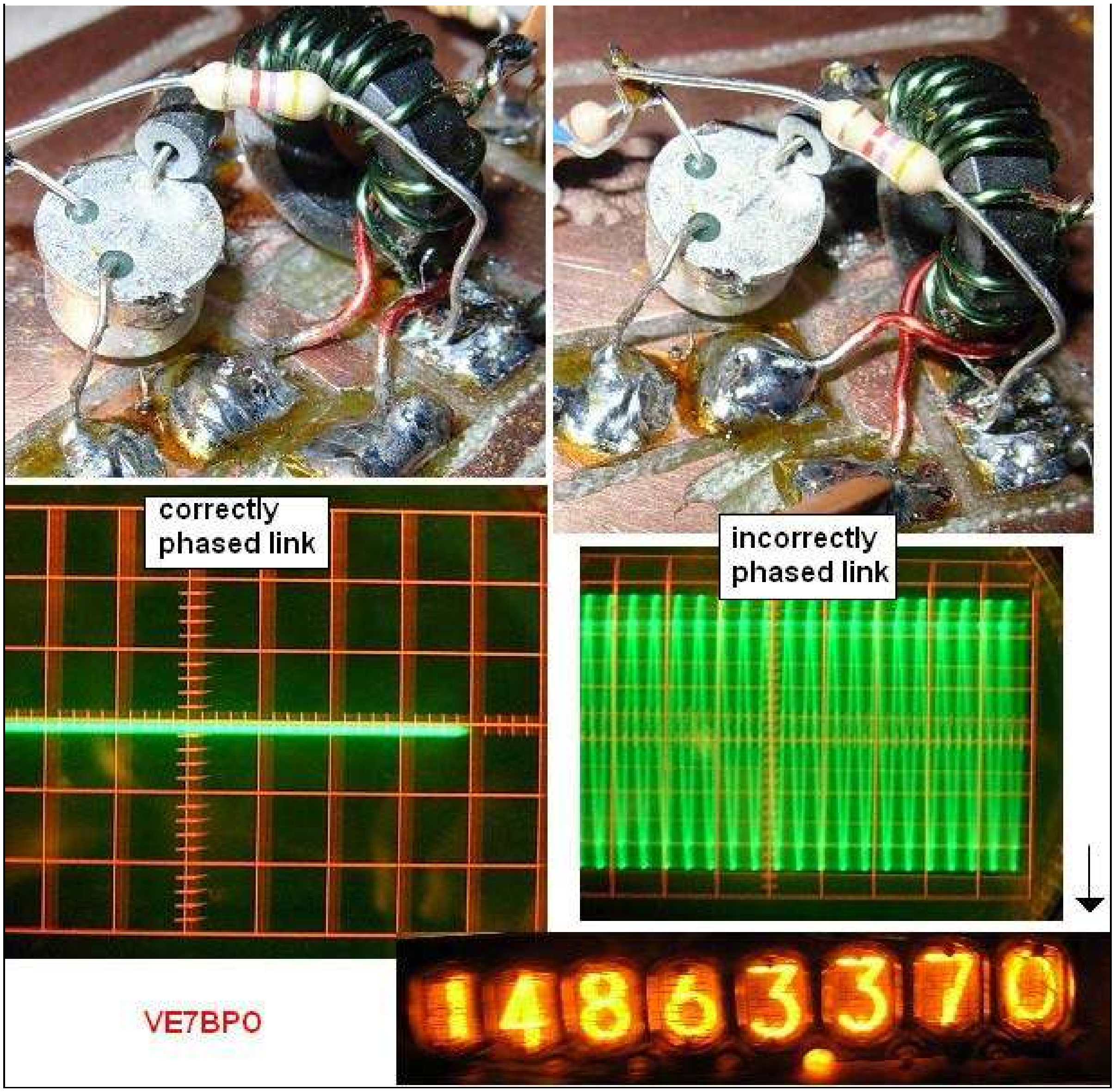
correctly phased link
incorrectly phased link
VE7BPO
14863370

Shown above is a photograph of 1 of the experiments from 2007. The one turn link from the Norton amplifiers just above is shown phased correctly and then phased incorrectly. Note the oscillation in the "badly wired" amp at 14.86 MHz. I routinely check all of my noiseless RF amps using the oscilloscope. Occasionally, I will put a shunt coil and cap (from input to ground) on the input to "exaggerate" any oscillations. This has proven to be a useful technique for testing if the phasing of the one turn link was done correctly.

RF preamp for the 40 Meter band with 3 tuned filters

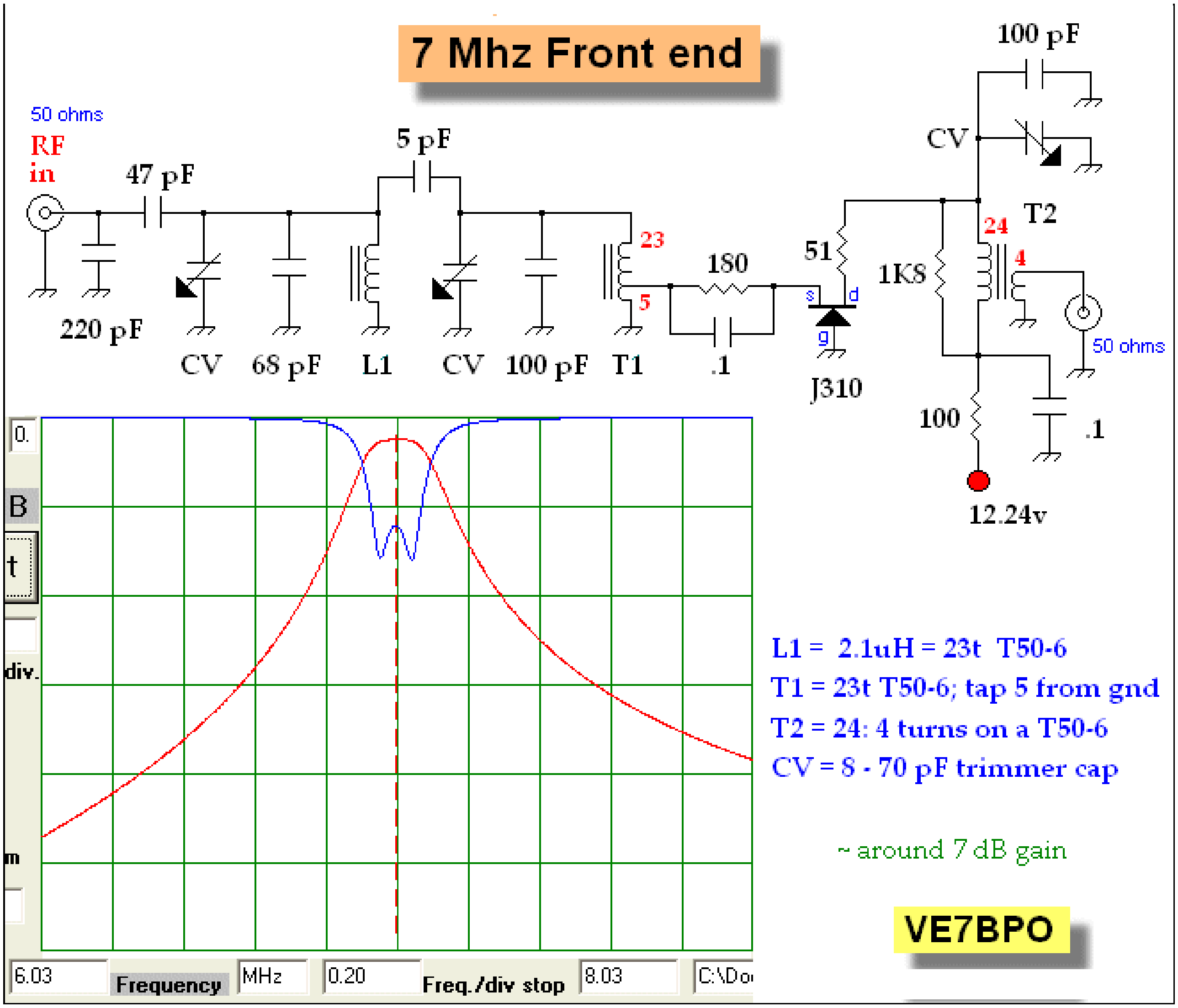
7 Mhz Front end
50 ohms
RF in
47 pF
5 pF
220 pF
CV
68 pF
L1
CV
100 pF
T1
23
5
180
.1
51
s
d
g
J310
1K8
CV
100 pF
T2
24
4
50 ohms
100
.1
12.24v
0.
B
t
div.
m
6.03
Frequency
MHz
0.20
Freq./div stop
8.03
C:\Do
L1 = 2.1uH = 23t T50-6
T1 = 23t T50-6; tap 5 from gnd
T2 = 24: 4 turns on a T50-6
CV = 8 - 70 pF trimmer cap
~ around 7 dB gain
VE7BPO

An experimenter's 40 Meter band front end for CW. This has a double-tuned filter and a low gain, lower noise RF amp. Great circuit for isolation of a product detector or mixer in a popcorn receiver.

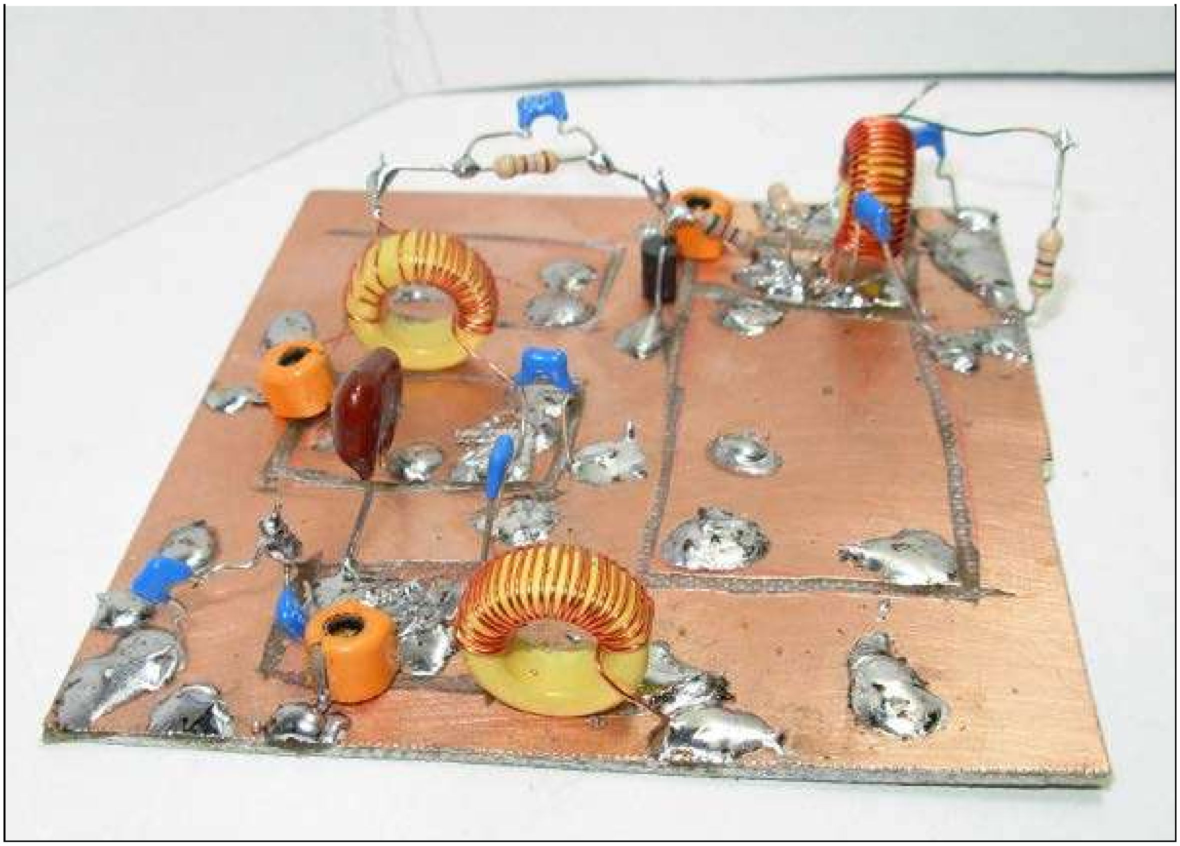

A photograph of the above 40 meter band front end, double-tuned filter plus tuned common gate RF amplifier. Input and output Z is 50 ohms.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

QRP HomeBuilder Software

Disclaimer:

THE QRP HOMEBUILDER SOFTWARE OFFERED HEREIN ("THE SOFTWARE") DOES NOT COME WITH ANY WARRANTY, EXPRESS OR IMPLIED. IF YOU MAKE USE OF THE SOFTWARE, PLEASE BE AWARE THAT YOU DO SO AT YOUR OWN RISK. NEITHER THE AUTHORS OF THE SOFTWARE AT JENNA DESIGN NOR ANY OTHER PARTY WILL ACCEPT RESPONSIBILITY FOR ANY OCURRING OR UNFORESEEN CONSEQUENCES OR DAMAGES THAT ARISE AS A RESULT OF THE USE OR MISUSE OF THE SOFTWARE.

Technical Info and Distribution

The QRP HomeBuilder applications are written in C++ for speed and compactness. Apps will specified as GUI (graphical user interface) or 32-bit console based (DOS look). Anyone may display or distribute these applications via website or diskette providing that they do not charge for the program(s). I know longer have a C++ compiler and no future work on these applications is anticipated.

QRP - RELATED APPLICATIONS FOR DOWNLOADING

CoilBuilder_99

CoilBuilder_99 is a powdered iron inductor winding application. Enter desired inductance, select core size and mix and press the Calculate button to determine the correct number of windings for your inductor. Data is also given showing, core color, permeability, frequency range, AL value and maximum number of turns versus wire guage for the chosen core size. Encompasses 12 different core sizes and 8 different mixes of powdered iron. Calculated results can be stored on a disk file or printed out.

Style: GUI, File size: 90K, zipped, 44K.

Bug Fixes: Some missing AL values for # 7 material added April 24/99. K6WHP's superior version is linked below.

Current Version is: 4 / 24 / 1999

Download the CB99.zip file

PI Filter Designer

PI Filter Designer is a simple 3 element 50 ohm input and output impedance pi filter designing application. This program allows the user to design simple lowpass filters by selecting from a variety of standard capacitor values either empirically or to suit what you have on hand. The filter 3 dB cutoff frequency and required L1 inductance are automatically calculated and displayed. In addition, the user may select an additional capacitor value to put in parallel with both caps C1 and C2. In this app XL = XC = 50 ohms impedance. No other impedances can be calculated with this program.

Style: GUI, File size: 47K, zipped, 22K.

Current Version is: 1 / 14 / 1999

Download the pifilter.zip file

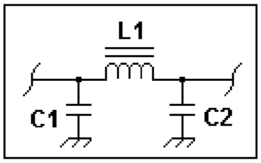

CapCoder

CapCoder gives the capacitance in microfarads, nanofarads and picofarads and tolerance of any capacitor code entered into its input section. Example : 104J. This app uses numeric spin-buttons and a combo box so that no typing is required for data entry.

Style: GUI, File size: 48K, zipped, 22K.

Bug Fixes : A nanofarad conversion error was corrected July 2, 1999.

Current Version is: 07 / 02 / 1999
Download the capcoder.zip file

Resistor Coder

Resistor Coder gives the resistance in ohms of any resistor color code entered into its input section. Four or five band resistors can be accommodated by this program. This app uses drop-down combo boxes so that no typing is required for data entry. Results may be saved to a disk file or directly printed.

Style: GUI, File size: 58K, zipped, 27K.

Current Version is: 1 / 16 / 1999
Download the resistor coder.zip file

Ferrite

Ferrite is used to calculate the number of turns required on toroidal ferrite cores to achieve the desired millihenry-value inductance. 15 different ferrite toroids are included in this application. This program will calculate the winding data for an inductance range of 0.001 to 27 millihenries.

Style: Console, File size: 64K, zipped, 31K.

Bug Fixes: Thanks to PA3CKR for the bug report; fixed Jan 19/99.

Current Version is: 1 / 19 / 1999
Download the ferrite.zip file

Universal Diplexer

Universal Diplexer calculates the inductance and capacitance values for a Bridge-Tee diplexer based upon a chosen superhet receiver intermediate frequency. The diplexer is the Joe Reisert, W1JR popularized design discussed under Diplexer Topics on this web site. The user inputs an IF and presses the Calculate button to have the capacitor and inductor values given in pF and uH respectively. The diplexer schematic is included in the application. Note that the this is for the Q = 1 version of the Bridge-Tee Diplexer.

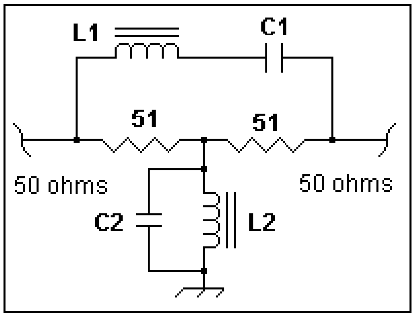

Style: GUI, File size: 49K, zipped, 22K.

Current Version is: 1 / 19 / 1999
Download the diplexer.zip file

HF Dipole

A very basic program for calculating the length of each leg of a 1/2 wave wire dipole antenna. Program good for 1 - 500 MHz, although intended for MF - HF useage. This app does nothing more than the standard 468/freq (MHz) type calculations. It was written for DOS many years ago and ported to Windows. The output shows the 1/2 wavelength and 1/4 wavelength design wire length in feet and meters. This app is probably of no help to experienced antenna designers.

Style: GUI, File size: 46K, zipped, 22K.

Update : Minor improvements made Feb 9, 1999

Current Version is: 2 / 9 / 1999
Download the hf_dipole.zip file

Resonator

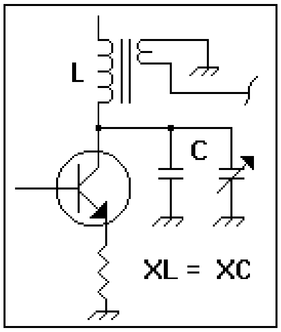

This application calculates the inductor and capacitor values for the tank circuit of a simple bipolar transistor RF amp. The basic schematic is shown above. Enter the center frequency plus the inductive/capacitive reactance you desire and press the Calculate button to calculate the necessary inductance and capacitance for L and C respectively.

Style: GUI, File size: 50K, zipped, 21K.

Current Version is: 1 / 23 / 1999

Download the resonator.zip file

NPN DC-BIAS

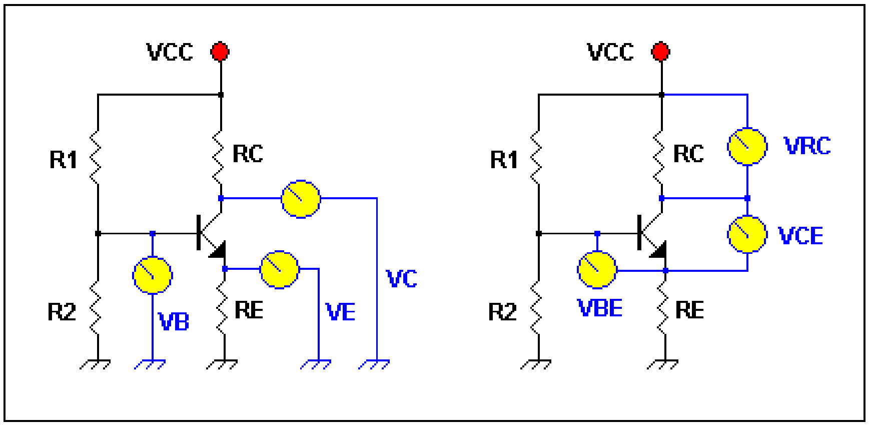

This application calculates the various voltages and currents of a simple voltage divider bias NPN bipolar transistor amp. The following is calculated: IB, IC, IE, VE, VB, VC, VCE and detection of Saturation or Cutoff. The user can alter the VCC, VBE, transistor beta and any of four resistor values R1, R2, RC and RE by picking the transistor value from a standard-value resistor table or manually entering the value. The schematic illustrates some of the voltage measuring points on the transistor schematic. *This app is in final BETA.*

Style: GUI, File size: 73K, zipped, 32K.

Current Version is: 16 / 04 / 1999

Download the nbias.zip file

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Popcorn Direct Conversion Main Frame

Discussion:

Note: if you click on a schematic, a larger version will appear in a new web browser window.

Shown to the right is the schematic to a low cost popcorn direct conversion receiver main frame. To complete the receiver, a front end band pass filter and a VFO with an output power of 7 dBm is required. This is indeed a frugal project using 4 cheap transistors, an RC low pass filter and an LM386N for output power to a pair of low-impedance headphones. The builder also has a choice of 5 diplexers and an optional mute circuit. This receiver is easily built using Ugly Construction and can be built in 3-4 hours with a bit of luck.

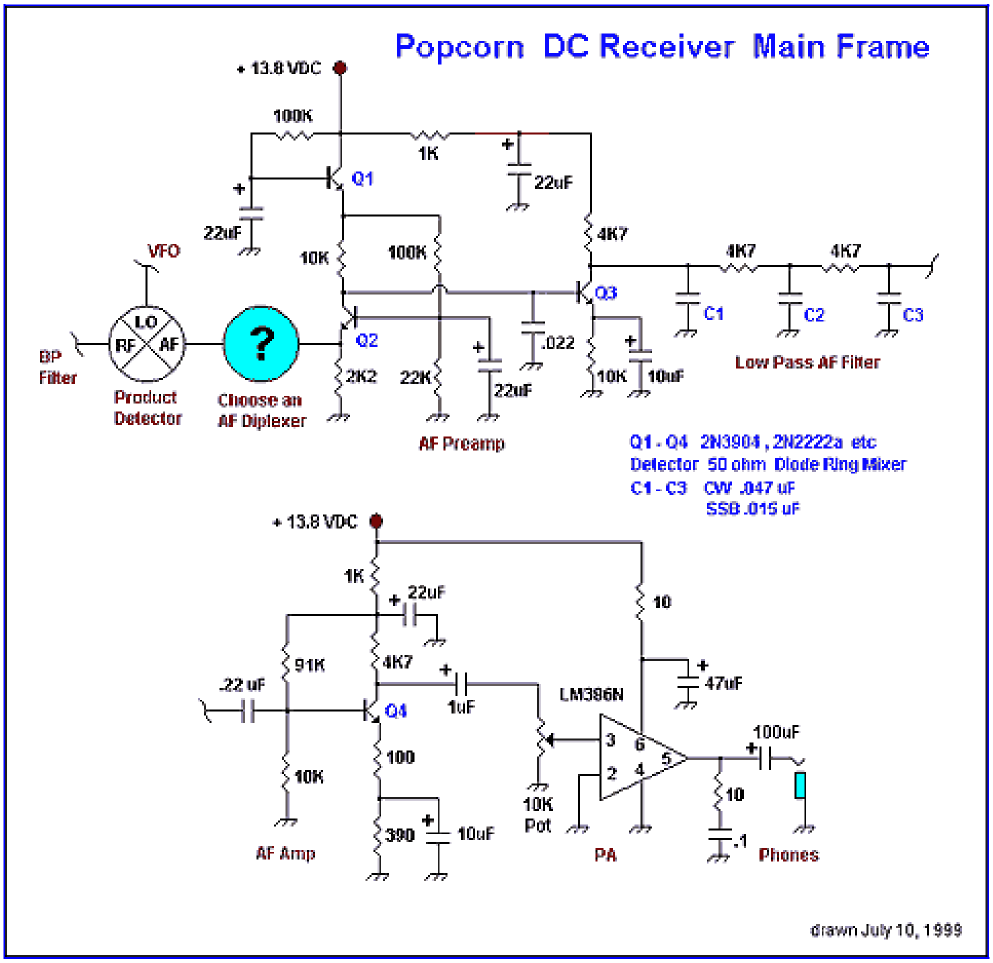
Popcorn DC Receiver Main Frame
+ 13.8 VDC
100K
1K
22uF
Q1
22uF
10K
100K
4K7
4K7
4K7
VFO
LO
RF
AF
?
BP Filter
Product Detector
Choose an AF Diplexer
Q2
Q3
.022
C1
C2
C3
2K2
22K
22uF
10K
10uF
Low Pass AF Filter
AF Preamp
Q1 - Q4 2N3904 , 2N2222a etc
Detector 50 ohm Diode Ring Mixer
C1 - C3 CW .047 uF
SSB .015 uF
+ 13.8 VDC
1K
22uF
10
91K
4K7
1uF
47uF
.22 uF
Q4
LM386N
3
6
2
4
5
100uF
10K
100
10
390
10uF
10K Pot
.1
AF Amp
PA
Phones
drawn July 10, 1999

Product Detector and Diplexers

The 50 ohm diode ring product detector can be commercial units such as the Mini-Circuits SBL1 or TUF-1 or homebrewed 50 ohm impedance units. Five simple "diplexers" are shown in the lower "Adjuncts" schematic for you to choose from. The one you choose will depend on available parts, cost and your requirements in a popcorn receiver such as this. These diplexers are mostly of the low pass filter variety and provide a ~50 ohm termination to the diode ring mixer and some matching to preserve the product detector dynamic range. I realize that except for (A) and (D) these audio frequency filters are not truly diplexers and will not provide DC to daylight matching. The intent of this web site is not high performance-high cost

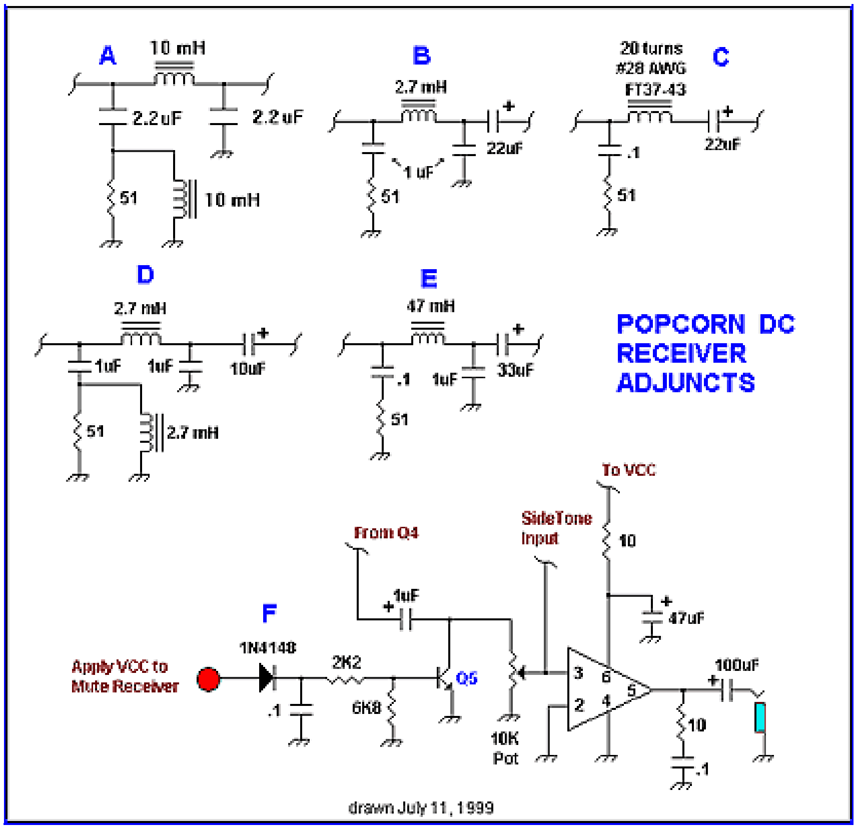
A
10 mH
2.2uF
2.2uF
51
10 mH
B
2.7 mH
22uF
1 uF
51
20 turns
#28 AWG
FT37-43
C
.1
22uF
51
D
2.7 mH
1uF
1uF
10uF
51
2.7 mH
E
47 mH
.1
1uF
33uF
51
POPCORN DC
RECEIVER
ADJUNCTS
To VCC
SideTone
Input
10
From Q4
1uF
47uF
F
1N4148
2K2
Apply VCC to
Mute Receiver
Q5
.1
6K8
10K
Pot
3
2
6
4
5
100uF
10
.1
drawn July 11, 1999

design and please do not confuse it as such. Note that electrolytic capacitors that bypass to ground such as the 1 uF caps must be non-polarized or bipolar for best results.

The (A) diplexer is by W7ZOI and is described on the Diplexer Web Page on this site.

The (B) and (D) diplexers are my designs and the (D) diplexer is the (B) diplexer with out the high pass component.

The (B) diplexer shown has a 3000 hertz 2 pole high pass/2 pole low pass design. This 2nd order filter provides reasonable overall matching Capacitors are standard-value, non-polar electrolytic types.

The (C) diplexer is a very basic, but very practical choice for this receiver.

The (E) diplexer is one that I used in one of my first DC receivers and the 47 millihenry inductor is a standard value unit sold by Mouser Electronics and others.

Another diplexer choice for this receiver might be the unit described by Rick Campbell, KK7B in his Binaural I-Q receiver project published in the March 1999 issue of QST.

Update May 15, 2009

There was confusion regarding the 2.7 to 47 mH inductors mentioned on this web page. I originally wound just the 2.7 mH inductor on a ferrite, but not the others. This is not a great idea as losses are high. For millihenries-value inductors, commercial parts should be purchased. A good brand to consider might be Epcos. Sorry for causing confusion.

10 milliHenry inductor
* Not wound at home

AF Preamps

The AF preamp section follows that of the Ugly Weekender Receiver designed by Wes Hayward, W7ZOI. I tried many other configurations and came to the conclusion that these two simple but elegant stages give a winning combination of low noise, good gain, low parts count, low hum and good AM broadcast band rejection. The Q1 transistor decouples the receiver preamp very well and no hum was detected in the headphones providing a well filtered DC power supply was used. The Q2 grounded base amp provides a low impedance termination of the product detector and diplexer stages. Q2 and Q3 are direct coupled and provide lots of gain to drive the succeeding low pass filter without it adding a huge abundance of noise to the signal. The bypass capacitor (0.022 uF) is essential to bypass any broadcast AM detected in the Q1 stage to ground. Other values of capacitors maybe tried, but do not omit this critical part.

Low Pass Filter

I cannot handle listening to a DC receiver on a crowded band without some low pass filtering. The high pitch heterodynes effect my concentration and give me a headache. Nevertheless, it is neat to temporarily listen to an unfiltered DC receiver; to hear the pure and wonderful signals possible by beating RF directly into audio. I prefer low pass to band pass filters at audio and have used many combinations of active filters using discrete components and op-amps, as well as passive designs using AF inductors to build wave filters. This receiver uses none of these devices, however they could be easily substituted for the filter shown. Connected to Q3 is a simple, cheap RC low pass filter based upon the design criteria given on the Discrete Component RC Audio Filters web page on this web site. The cutoff values you calculate will be ballpark and values of 0.047 uF for CW and 0.015 uF for SSB were chosen, but other values could just as easily been used and please do not hesitate to experiment with the caps and/or the resistors to suit the parts you have on hand. For the capacitors in the low pass filters, avoid using ceramic disk type caps if you want the best possible performance. Polyester, polypropylene, polystyrene or polyester film type are all suitable, however, ceramic caps will work if you are really going junk box/low cost.

I attempted to make a wave file to demonstrate the low pass filter. I came right off the headphone jack into the input of the of my 16-bit PC sound card via a step up audio transformer and the results were a little disappointing. Sixty-cycle hum and distortion of loud stations were added by the sound card. The sound file is big (636 KB) and is a digital recording of me tuning through a 30 meter pile-up using the lowest sample rate and frequency possible on my computer. The low sample rate/frequency also degraded the sound somewhat as well, but I decided to put it on the page, warts and all. The DX station was a VK2 and sure did cause a lot of excitement on 30 meters that night around sunset on the left coast. Actually the wave file demonstrates how good the receiver sensitivity and AM radio immunity is. In addition, the low receiver background noise is also very apparent underneath the constant 60 cycle hum. The 60 cycle hum and clipping of loud CW signals is not heard in the headphones and is a soundcard manifestation. Perhaps the best method would be to come of Q4 and go right into the sound card with a smaller line-in signal voltage. Download the popdc wave file

---

AF Driver and Final Amp

Connected to the input and output of the Q4 stage are small value capacitors to provide some high pass filtering for the receiver amplifier chain. Some emitter degeneration is used on Q4 to provide a better termination of the preceding RC low pass filter. The receiver amplifier chain has a lot of gain and when the 10K pot is turned to minimal resistance (cranked) , the LM386N can be driven into distortion. You may want to limit the maximum gain with a series resistor connected to the 10K pot after building and testing this receiver.

The final AF amp is the perennial LM386N, a low cost, easy to use AF amp. Turn it upside down and solder pins 2 and 4 right to your copper ground plane to anchor this part. It can easily be configured to drive a small speaker.

An optional mute circuit is shown in the "Adjuncts" schematic and is labeled (F). This circuit is a simple transistor switch which grounds the output from Q4 and mutes the receiver audio. This circuit switches rapidly and there are no annoying pops or clicks to be heard in the headphones when it is switched. Apply the VCC to the diode as shown to mute the receiver during transmit if the receiver is used in conjunction with a transmitter. Q5 in the mute circuit can be a 2N3904 or 2N2222a or substitute.

In addition, a suggested side tone input to the LM386N is shown. I have started to use simple one section RC filters on the output of my side tone oscillators to smooth the waveform into a more pleasing audio tone.

Conclusion:

This popcorn receiver can be made very inexpensively and has good sensitivity and a reasonable noise level and selectivity. I tested this main frame on 30 and 40 meters and really enjoyed it. This receiver main frame could be combined with an inexpensive VFO using tuning diodes to keep cost down and the popcorn factor up. Although it does not use tuning diodes, a 40 Meter band VFO schematic has been placed on the VFO page.

Here is a YouTube Link using the receiver with a different front end filter and VFO. This is not my radio or video.

A blog post from Peter AK6L --- it's good to see builders moving beyond kits.

---

40 Meter Popcorn Superhet Receiver

Discussion:

Note: if you click on a schematic, a larger version will appear in a new web browser window.

To the right is the schematic for a no-frills, relatively low-cost CW superhet receiver with a 4.00 MHz Intermediate frequency. There is no AGC or RF gain control, however this receiver has good large signal handling capability. This receiver uses just 6 bipolar transistors and an op amp for reasonable volume into headphones. Much of the ideas/design of the various stages must be credited to Wes Hayward as I borrowed heavily from his previous work and through ideas obtained by discussion. If one were to homebrew the diode ring mixers, indeed this would be a very low cost receiver giving reasonable performance which outperforms any NE602 based superhet receivers that I have built or listened to. Below the main schematic is a diplexer diagram that allows the builder to choose from one of two RF and AF diplexers used to terminate the diode ring mixers.

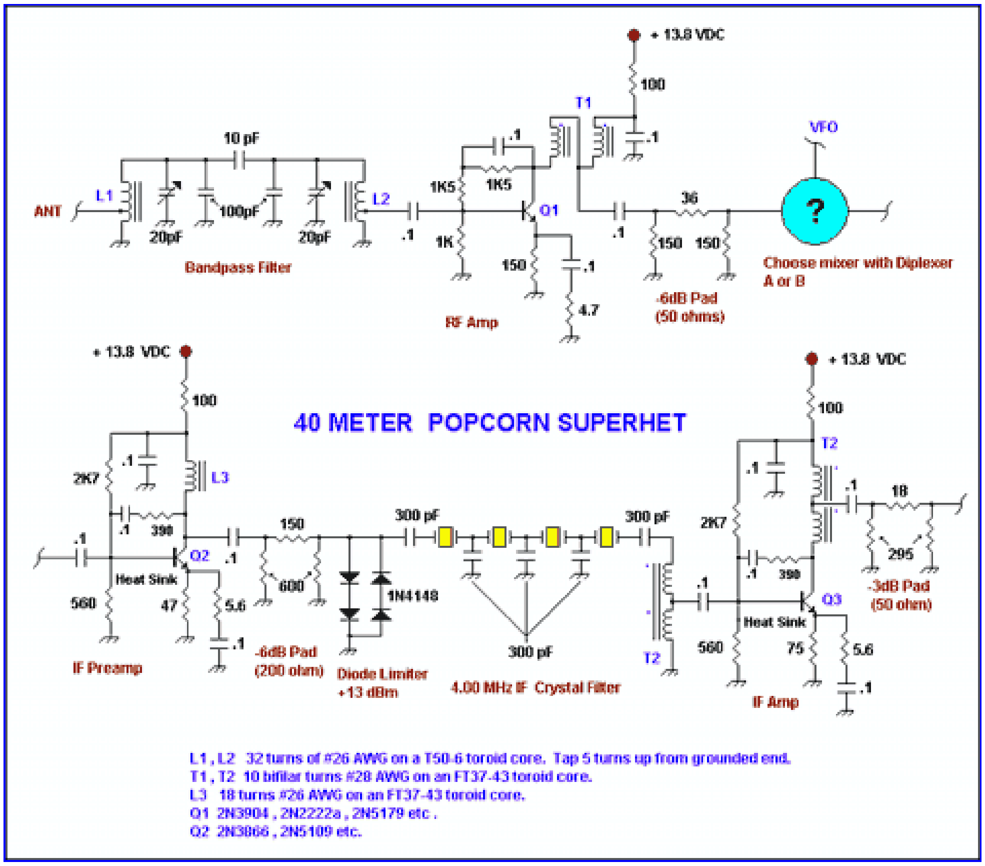
40 METER POPCORN SUPERHET
ANT
L1
L2
10 pF
100pF
20pF
20pF
Bandpass Filter
.1
1K5
1K5
1K
Q1
150
.1
4.7
RF Amp
+ 13.8 VDC
100
T1
.1
.1
36
150
150
-6dB Pad
(50 ohms)
VFO
?
Choose mixer with Diplexer
A or B
+ 13.8 VDC
100
2K7
.1
L3
.1
390
.1
Q2
Heat Sink
560
47
5.6
.1
IF Preamp
150
600
-6dB Pad
(200 ohm)
1N4148
Diode Limiter
+13 dBm
300 pF
300 pF
300 pF
4.00 MHz IF Crystal Filter
T2
.1
560
2K7
+ 13.8 VDC
100
T2
.1
.1
18
.1
390
295
Q3
Heat Sink
75
5.6
.1
-3dB Pad
(50 ohm)
IF Amp
L1 , L2 32 turns of #26 AWG on a T50-6 toroid core. Tap 5 turns up from grounded end.
T1 , T2 10 bifilar turns #28 AWG on an FT37-43 toroid core.
L3 18 turns #26 AWG on an FT37-43 toroid core.
Q1 2N3904 , 2N2222a , 2N5179 etc .
Q2 2N3866 , 2N5109 etc.

Band pass Filter and RF Preamp

From the 50 ohm receiver antenna jack, first off is a double-tuned band pass filter which was designed by Rick Campbell, KK7B and works very well. The trimmer caps can be the 5 - 20 pF units sold by Digi-key and Mouser. The fixed-value caps in my prototype were inexpensive monolithic ceramic capacitors purchased from Digi-key. Rick used an NP0 ceramic for the 10 pF coupling cap plus silver-mica type for the 100 pF caps in his original design. For possible lower insertion loss, the probable best/cheapest way would be to use all NPO ceramics for the fixed value caps in this filter.

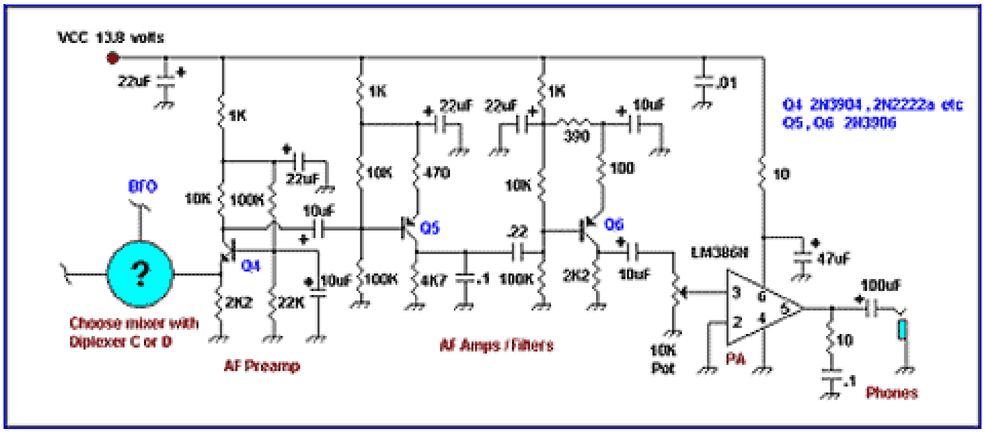
VCC 13.8 volts
22uF
1K
BFO
?
Choose mixer with
Diplexer C or D
10K
100K
Q4
2K2
22K
22uF
10uF
10uF
AF Preamp
1K
10K
470
Q5
100K
4K7
.1
22uF
22uF
1K
390
10K
.22
100K
Q6
100
2K2
10uF
10uF
AF Amps / Filters
.01
Q4 2N3904 , 2N2222a etc
Q5 , Q6 2N3906
10
LM386N
47uF
3
6
2
4
5
PA
10K
Pot
100uF
10
.1
Phones

The RF amp is my favorite popcorn RF amp ; a 50 ohm feedback amp. A grounded-gate JFET amp was tried in its place and was also found to be quite suitable and does not require the -6dB pi attenuator that follows the feedback amp as shown in the schematic. The feedback amp's 50 ohm input impedance properly terminates the band pass filter. The -6 dB attenuator pad following the amp to helps provide a 50 ohm input impedance for the mixer and to reduce stage gain which aids in preserving the signal to noise ratio of the receiver. If a builder wants a little more sensitivity, the pi attenuation pad could be reduced to -3 dB however this may effect the receiver dynamic range. The transformer T1 is one of 2 broadband transmission line transformers in this receiver. It transforms the 200 ohm collector impedance to 50 ohms for the succeeding stage.

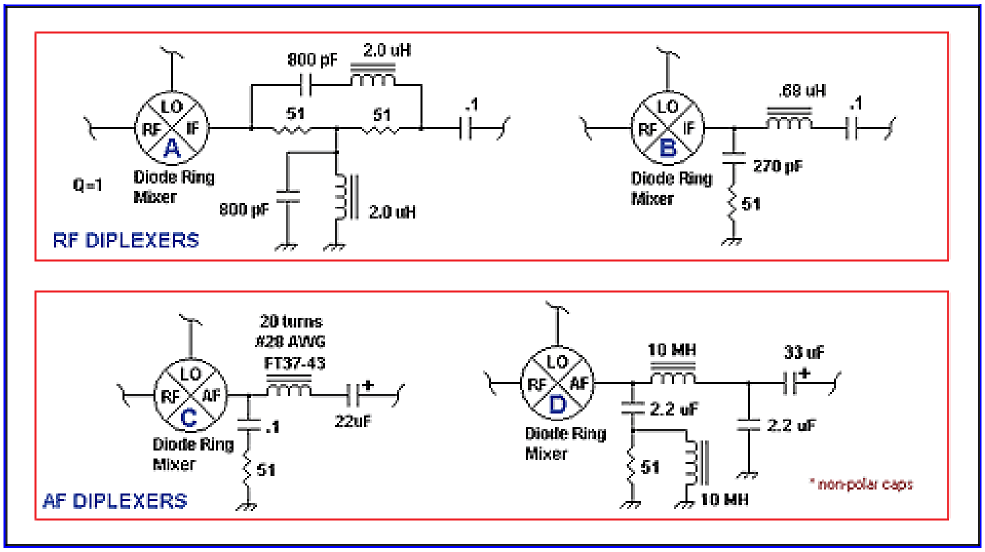

Mixer and Diplexer

A 50 ohm diode ring mixer (7dBm) such as the Mini Circuits SBL-1 or TUF-1 or homebrew are all suitable. Following the mixer is an RF diplexer of your choice. The more complex Brifge-Tee (Q = 1)diplexer (A) is an excellent design, however maybe overkill in a popcorn superhet such as this. For the (A) diplexer, to get the necessary 800 pF for the capacitors, simply parallel a 470 with a 330 pF or a 120 pF with a 680 pf capacitor. The inductors at 2.0 uH are wound on powdered-iron torroids. You can use # 26 AWG wire and it requires 22 turns on a T37-2 core or 20 turns on a T50-2 core. In addition, you can use a #6 material torroid to wind the inductors. This diplexer is described elsewhere on this web site.

The simpler (B) diplexer uses a ~3 times the IF frequency that I have seen this basic design in many textbooks and articles and provides reasonable matching with a 50 ohm inductive and capacitive reactance. The cutoff frequency chosen was 11.78 MHz as this allows the use of a standard value capacitor (270pF). To wind the 0.68 uH inductor use 13 turns on a T37-2 torroid or 12 turns on a T50-2 powdered iron torroid core. You can easily use 24 - 26 AWG wire for the inductor.

IF Preamp , Crystal Filter and IF Amplifier

Except for the inductors, the IF preamp and IF amp are identical and both warrant a small clip-on heat sink as they draw reasonable current. The standing current maybe increased or reduced by changing the 47 and 75 ohm resistors connected to the Q2 and Q3 emitter respectively. Factors such as available power supply current versus dynamic range requirements may come into play. One may want to stand more current in the IF preamp and less current in the IF amp. For example, the 75 ohm resistor on the Q3 emitter could be increased considerably and/or the 5.6 ohm degeneration resistor could be increased as well if less stage current draw is wanted. The 2N3866 transistor is usually a cheaper way to go for these amps than the 2N5109, but the choice is up to you as you may have something available in your junk box. The 200 ohm -6dB pad following the IF preamp should not be omitted as it helps prevents the stage from seeing reactance's created ahead by the crystal filter. The four diodes form a 13dB limiter to protect the crystal filter should a catastrophically large signal be present in the receiver's front end. They maybe omitted. A -3dB 50 ohm resistive pad terminates the IF amp and helps establish a 50 ohm input impedance for the product detector ahead. **Click here** for more on the IF preamp.

This receiver has a narrow IF Cohn Crystal filter. Bandwidth is ~ 405 hertz, which unfortunately makes tuning quite sharp however this filter is very nice for crowded band conditions. The IF filter crystals should be closely matched in frequency to prevent unwanted ripple in the pass band. Generally, you have to buy 10 and then if you have a frequency counter, use the receiver BFO stage to test your crystals for matching. Pick the closest 4 crystals and use them in your filter. It does not matter if the crystals have series or 20 pF load capacitance, but it does matter that they are matched in frequency within 40 hertz of one another or better for this receiver. For my prototype receiver, I purchased ten 20 pF load capacitance 4 MHz crystals and luckily found 4 that matched each other within 9 hertz! For those builders who do not have a frequency counter, some QRP parts retailers sell matched sets of crystals. It is important to note that the BFO should be set on the high side of the IF frequency as simple crystal ladder filters have a steeper upper passband than lower pass band.

The crystal filter is terminated by the 4:1 transmission line transformer and then 50 ohm impedance of the IF amplifier. The -3dB pad following the IF feedback amplifier helps to terminate the crystal filter by helping ensure a 50 ohm IF amp input impedance and should not be omitted. Place a 75 and a 220 ohm resistor in series to get the required 295 ohm resistance on each leg of the pi attenuator.

Many may balk at just one stage of IF amplification, but since there is no AGC and this is a CW receiver, it works well. A feedback amp is once again used to provide correct input and output impedances for stages connected to the IF amp. Following the IF amp is another attenuator set for -3dB and then a 50 ohm diode ring mixer.

Product Detector, AF Diplexer and Audio Amplifiers

The mixer/detector can be SBL-1 or TUF-1 types or homebrew if you want to reduce costs further as the mixers are the single most expensive components in this receiver.

Again a choice of diplexers is required. The (C) AF diplexer is very simplistic but very practical if you are trying to keep costs low. The (D) diplexer is designed by W7ZOI and is from the Diplexer Web Page on this site.

Following the diplexer, a grounded base audio amp provides a 50 ohm termination to the product detector. AF gain and some AF filtering are provided by Q5 and Q6 which together attenuate frequencies less than 72 hertz and greater than 638 hertz. This amplifier pair are described on the discrete AF filters web page on this web site. Keep your leads short on all the AF transistors.

The final AF amp is the perennial LM386N, a low cost, easy to use AF amp. Turn it upside down and solder pins 2 and 4 right to your copper ground plane to anchor this part. There are a number of low-noise alternatives to the LM386 available which are generally more expensive but would be quite suitable. Discrete component AF amps can also be used, but a popcorn part such as the LM386N maybe cheaper and easier.

VFO and BFO

VFO schematic

BFO schematic

Alternate version of this receiver

Construction Ideas

When constructing any project, build in small modules and test each one separately. For instance, the AF amp should be built first and then tested by injecting a very low-level audio frequency tone into that stage and listening for output in your headphones. Every QRP workbench should have a simple AF tone oscillator from a schematic similar to the ones used for keying side tones in CW transmitters. The encased oscillator should have to a 100K or so potentiometer connected to the output to vary the output signal amplitude. Generally use maximum resistance on the 100K pot to start with and reduce this resistance slowly as the in-test amplified oscillator output could be very loud!. After testing the AF amp, build the 3rd AF preamp stage including the 10K panel mounted pot so you can vary the gain going into the AF amp. Now inject the AF oscillator output into the input on the pot and vary the 10K pot to ensure that the stage you built is working. It should be a lot louder now and should go up and down in volume with the 10K pot. Finally build the remaining preamp stage and once again test the circuit with your AF oscillator. The output into the phones should be painfully loud now when cranked up! The next stage to build would be the BFO. If you do not have a scope, peak the tuned circuit by watching the S meter on a radio receiver located nearby. Ensure that you put a load on the output winding of the BFO such as 47 ohm resistor to ground. A small piece of wire can be used as an antenna if the BFO signal is too weak to activate the S meter on your receiver. Once peaked, you can now use the BFO to match your IF crystals. To use the BFO to match your crystals, use a small wire to bypass or disconnect the 60 pF variable capacitor that is used to connect the crystal to ground. In other words, the bottom lead of the crystal is connected to ground with a short piece of wire. This makes testing your crystals a little more scientific as the variable capacitor cannot influence the crystal frequency during testing. You can also use the BFO in conjunction with a scope or voltage probe to test the various RF amps in the receiver. I do this all the time with my scope. Proceed with this build a stage, test a stage method and you should be rewarded with a functional end product.

KK4RF's version of the 40 Meter Band Popcorn Superhet

Marty, KK4RF emailed me and described his version of the popcorn superhet from this web page and contributed some great info and photos of his receiver. Of note is Marty's use of Radio Shack IC boards for mounting the components for each stage other than the VFO circuit. This is yet another variation from ugly construction that I have also used which works extremely well. Marty built the VFO using pure ugly construction and I was glad to hear that he is enjoying good frequency stability even with the lid off the VFO enclosure.

He built the receiver into an old Heathkit HW-12 single-bander case from the 1960's and it is a very attractive receiver to say the least. He found an old National Velvet Vernier Drive at the Virginia Beach Hamfest this year and used it to tune the VFO. Don't you love Hamfests! For the BFO he used an BFO circuit with a 4 mHz crystal from a different receiver project (a project that never quite worked.) He built a small power supply and located it along with the BFO under the chassis.

I like Marty's generous use of ground plane and neat stage layout. He reports good selectivity with his IF filter and apparently built four superhets that did not work before building this popcorn version. This is more a testimony to Marty's perservance to home building than to this receiver design in my opinion. I won't tell you how many rig failures I have personally incurred, as it would take a long time! Many thanks to Marty, KK4RF for the feedback and great pictures.

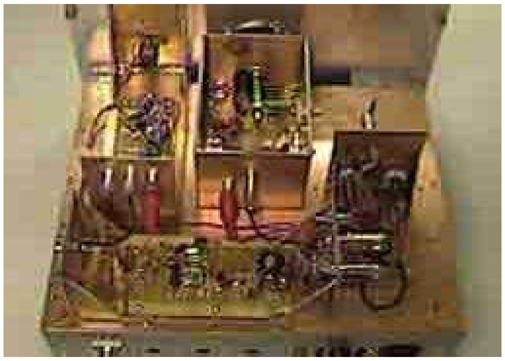
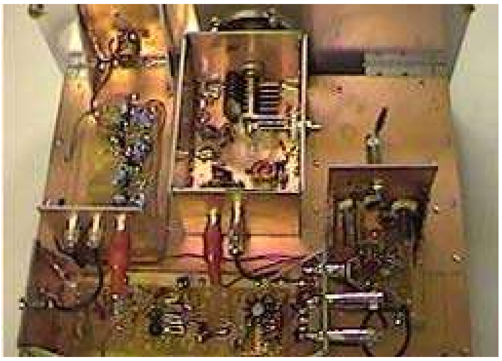

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Broadband Transformers plus Diode Ring Mixers

Discussion

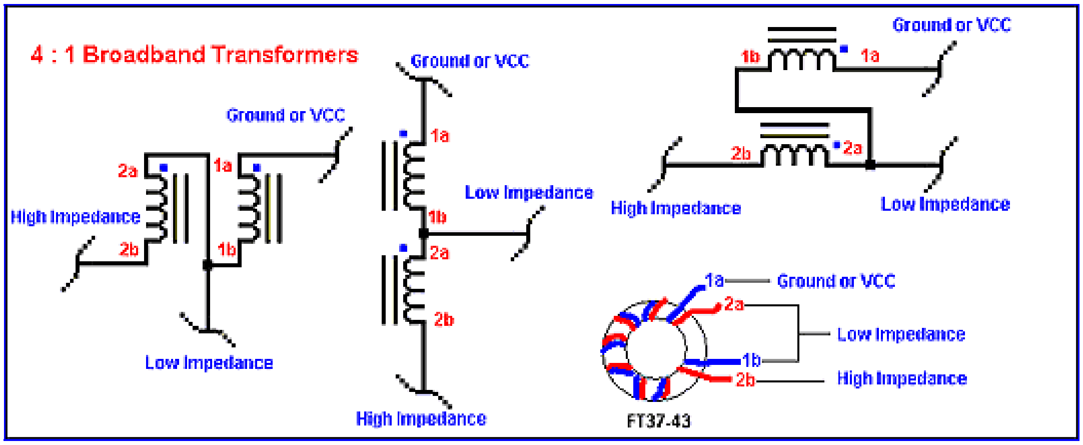

There are 2 basic types of broadband transformers used in most QRP work — conventional and transmission line style. Both types may be wound on ferrite toroids, pot cores or rods, however, I only discuss the toroidal transformers we employ to give a 4:1 impedance transformation.

I use these transmission line transformers on many projects on the QRP / SWL HomeBuilder web site. For MF and HF uses, a ferrite core permeability of 850-900 is generally required and the FT37-43 ferrite core proves suitable. Shown above are 3 equivalent schematics of the 4:1 transmission line transformer. You'll probably find that the center drawing easiest to conceptualize, however, with closer examination, all 3 schematics are the same and transform signals from unbalanced 50 Ω impedance up to 200 Ω unbalanced impedance or visa-versa.

The high impedance is 200 Ω and the low impedance is 50 Ω in all cases. It is important to know that these transformers are symmetrical and the points labeled Ground or VCC can be switched with the point labeled High Impedance. *Click on the schematic to enlarge it.*

Winding the 4:1 Transformers

Wire Twisting

Transmission line transformers are wound with bifilar (2 wires — generally twisted together). Winding these transformers is very easy. All you need are two ~18 cm (7 inch) pieces of #28 AWG enamel coated wire and an FT37-43 ferrite toroidal core.

A shop vise, a ruler, plus a brace and bit hand drill may aid your construction — I bought my brace and bit drill at a garage sale for 2 dollars. You need to twist the 2 pieces of wire together to get ~3-4 twists per cm (8-10 twists per inch) in the wire. To do this, loosely twist the wires at one end and place these twisted ends in a bench vise. Next, place the free wire ends together in your brace and bit drill chuck (no drill bit) and tighten up the chuck so that the wires are held securely.

Try to keep wire length and tension equal. Start turning your drill to twist the wires together and every once and a while measure how many twists per cm with a ruler. When you get to 3-4 twists per cm (8-10 twists per inch) you're done and then may trim the excess leads with a wire cutter in preparation for final winding and soldering.

At VHF, I often use just 3 or 4 total turns on an FT23-43 toroid with a piece of wire just a few cm long. I place the 2 wires in the vise and twist them using a pair of pliers held parallel to the wire.

Transmission line transformers will also work if the wires are untwisted. 3-4 twists per cm only serves as a non-criticial guide. Never wind your simple 4:1 transformers with bare wire.

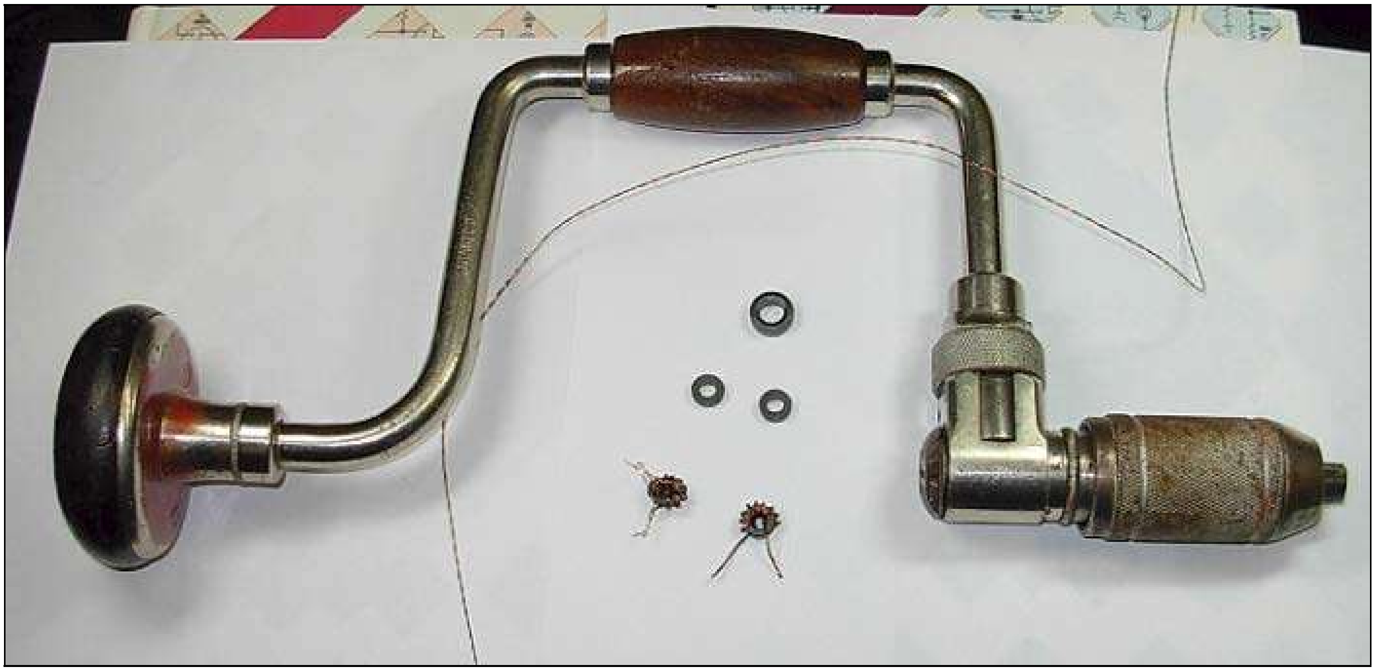

A brace and bit hand drill plus a vise provides a good way to twist your wires.

Final Wiring and Soldering

Leaving a 2.5 cm (1 inch) lead, wind ten complete loops through the toroidal core leaving a small gap between the start and finish leads.

Untwist the leads a little so that you have 4 separate wires. One set of these wires wires will be called winding #1 and the other winding #2. You need to identify them and further break them into 1a, 1b and 2a and 2b. Generally I regard the the top two windings as (A) and the the bottom two wires (B), however, use whatever system works best for you. Strip off the enamel at the tips of all four leads and then get your ohmmeter or better yet, a beeping continuity tester.

Start on one of the top (A) wires by connecting the ohmmeter or continuity beeper to it and then touch one of the bottom wires and then the other bottom wire. Whatever bottom wire (B) shows continuity with your top wire should be marked along with the source top (A) wire with paint, liquid paper, tape, or whatever you like.

I prefer to wind 2 different colored wires if possible.

Designate the marked wire pair winding number 1. You may also want to test for a short circuit — there should be no connection between wire set 1 and wire set 2 at all. So now you have 2 wires sets, winding set 1 is marked and winding set 2 is unmarked. The top two wires are arbitrarily labeled A and the bottom two wires are labeled B . Refer to the schematic above for clarification. Connect 1b to 2a and twist them together and then solder. Your done!

It's really easy to make these things don't you think?.

A trio of bifilar transformers wound on FT37-43 ferrite toroids. 2 colors of wire reduces errors and speeds up construction. Consider making up 5 at a time, so you have them on hand and do not have to interrupt your experiments.

Homebuilding Diode Ring Mixers

Discussion:

Easy to make, homebuilt diode ring mixers give a low-cost alternative to commercial diode ring mixers. A double-balanced diode ring mixer has 2 unbalanced to balanced transformers and a diode ring. The impedances at the three ports is 50 Ω. The transformers are wound with #28 AWG enamel coated wire on a FT37-43 ferrite toroidal core using a trifilar (3 wire) technique.

The wire twisting and winding technique is done as described above for the bifilar transformers. The connections 2b and 3a are twisted together and soldered. Again you will have to develop a technique to help you distinguish the wires from 1 another. *Click on the schematic to enlarge*

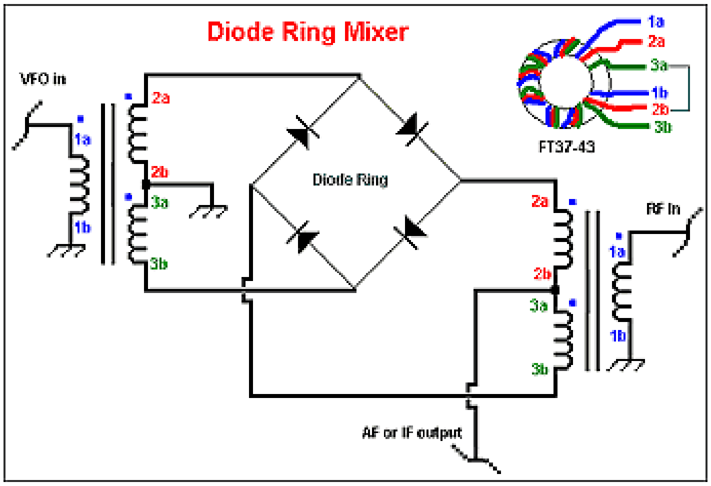

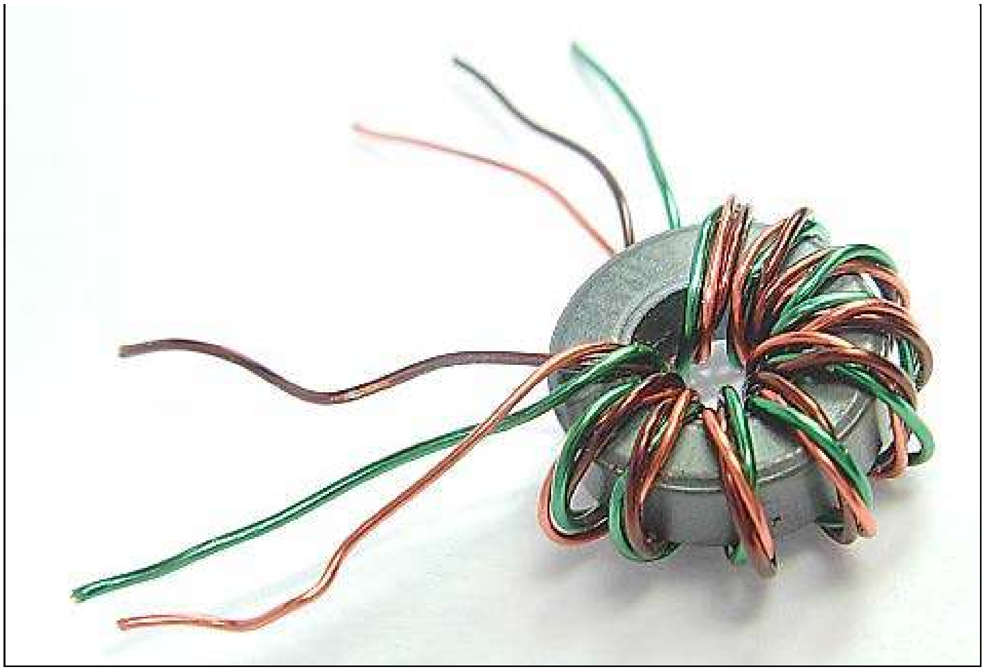

A trifilar transformer wound using 3 colors of wire on an FT37-43 toroid

#32 BELD. 1/4 LB.
20
30
28

A sample of the enamel coated copper wire collection I wind inductors and transformers with. In stock are wire gauges from 18 to 34. Like everything else, this collection started small and grew over time. Be vigilant for bargains and when you find a good price, purchase a whole bunch as it does not go bad. The Belden wire (orange spool ends) is over 40 years old and the enamel insulation remains perfect.

Diode Matching for Mixers

Discussion

For optimal results Schottky or Hot-Carrier diodes should be used. However, common diodes such as the 1N914, 1N4148 or 1N4454 are all quite suitable and are much cheaper. The four ring diodes should be matched to help mixer balance and thus carrier suppression. At MF and HF the most critical matching required is the forward voltage drop across the diode and this is easily performed with a sensitive voltmeter.

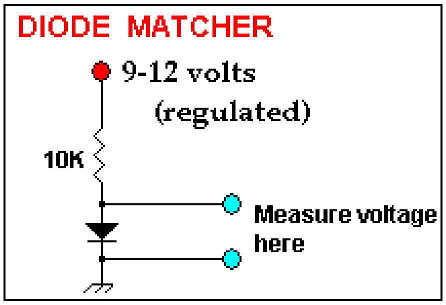

Set your voltmeter on the 2 volt scale to give you 3 decimal places for matching the voltage drops. Try and find 4 diodes close to one another. In addition, best results maybe obtained if all the diodes are the same type (i.e. all 1N4148) and if they are all from the same manufacturer. Look above for easy schematic to match your diodes with a voltmeter. Give the diode under test at least 20 seconds to warm up and stabilize before taking your voltage measurement.

RF — Test and Measurement

Diplexers Topics

Introduction

My original web page on diplexers was rather incomplete and received some criticism from electronic engineers albeit the focus of this web site is "popcorn" designs. Wes Hayward, W7ZOI provided me some excellent schematics, analysis and simulations for diplexers which terminate doubly balanced mixers and these are presented below. After this section, the W1JR Bridge-Tee RF Diplexer from the original QRP HomeBuilder diplexer web page is presented along with new commentary and simulations by W7ZOI.

The final section presents a practical diplexer for terminating a product detector. All graphical images labeled as Figures 1-24 are copyright and property of W7ZOI and may not be presented elsewhere. Updated September 23, 2000.

W7ZOI Diplexer Notes

The usual amplifier is a two-port circuit. That is, it has an input port consisting of two terminals and an output consisting of two more. One terminal (ground) can be shared between the ports. Many filters are also two-port networks, including most of the ladders networks we use so often. Many other networks have three or even more ports. A common example is a mixer, which has three ports. Another example of a three port network is a diplexer. This linear network is usually designed around two port filters where one end of two different filters are paralleled to form an input port. This is illustrated as Figure 1. The purpose of a diplexer is usually to force a frequency constant impedance to occur at the input port, even though we usually only use one of the two output ports for signals. The simplest form of diplexer uses a pair of 1 element filters, a low pass and a high pass. This is shown in Figure 2.

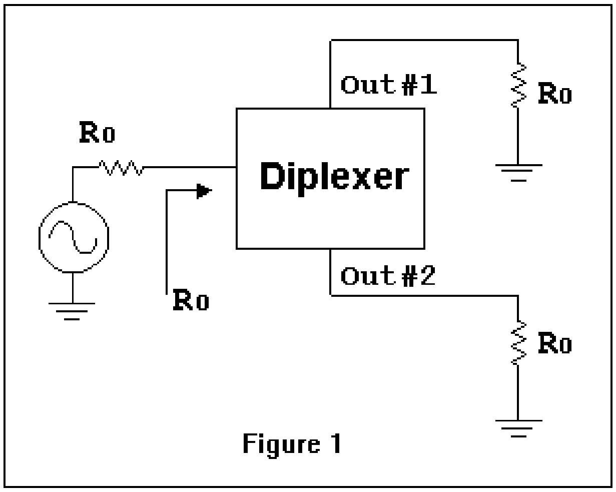

Figure 1

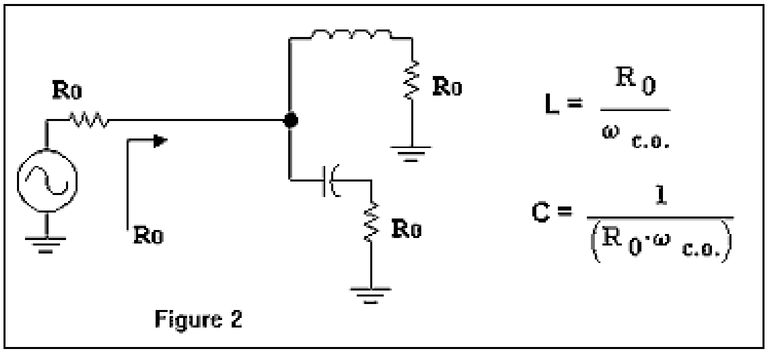

Figure 2

Where the equations give the L and C that provide a perfect match. The angular frequency is called the cross-over. A familiar example is the cross over used in audio systems. The network that splits signals is a diplexer. Here is an example where both outputs are used. Another form of diplexer is the band pass/band-stop combination. This is shown in Figure 3:

Let's now consider further some examples, some that work and some that don't work as well. First, let's look at an audio diplexer that follows a product detector in a DC receiver. The load of interest is the first audio stage, which has a 50 Ohm input resistance. The diplexer offered is Figure 4. Note that this is not the combination of filters. It just looks like a low pass with an extra resistor. The response of this circuit is shown in Figure 5. The transmitted signal never gets up to the desired 1 volt in the low pass passband while the impedance match, represented by reflection coefficient, never gets down to the desired zero. The response is just that of a lossy low pass filter.

In 2012; Click on many of the diplexer images to see the original sized version

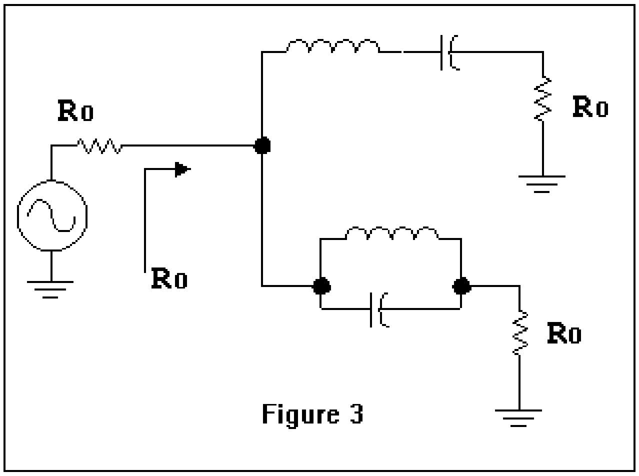

Figure 3

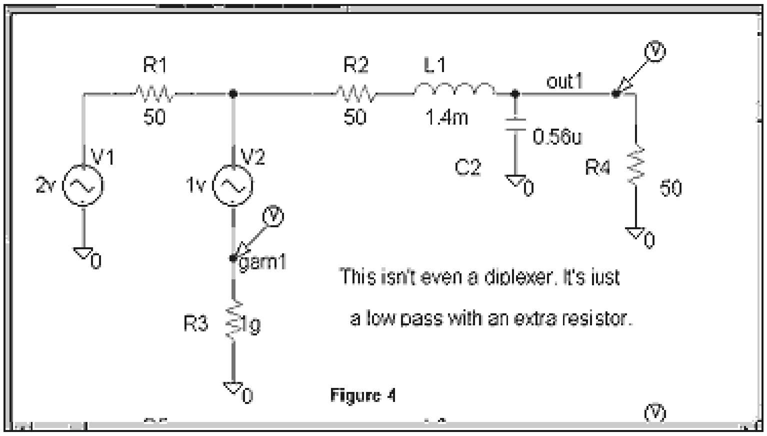
R1
50
R2
50
L1
1.4m
out1
0.56u
C2
R4
50
V1
2v
V2
1v
gam1
R3
1g
0

This isn't even a diplexer. It's just a low pass with an extra resistor.

Figure 4

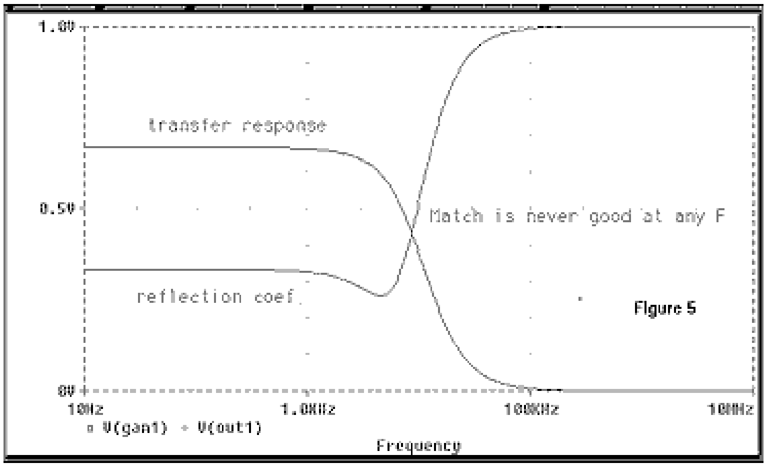

Match is never good at any F

Figure 5

The normal filter circuit without the extra resistor is Figure 6. The corresponding output response is shown in Figure 7. Note that the transmitted signal is now up at 1 while the reflection is down to zero, both within the passband. Transmission goes to zero while reflection is 1 in the stopband.

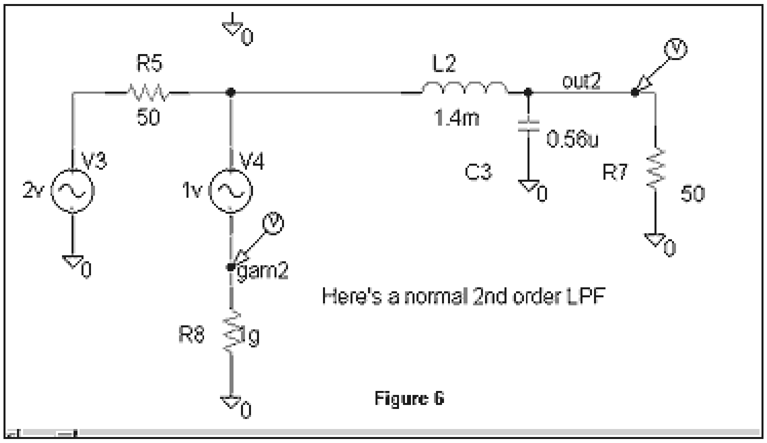
R5
50
L2
1.4m
out2
0.56u
C3
R7
50
V3
2v
V4
1v
gam2
R8
1g
0

Here's a normal 2nd order LPF

Figure 6

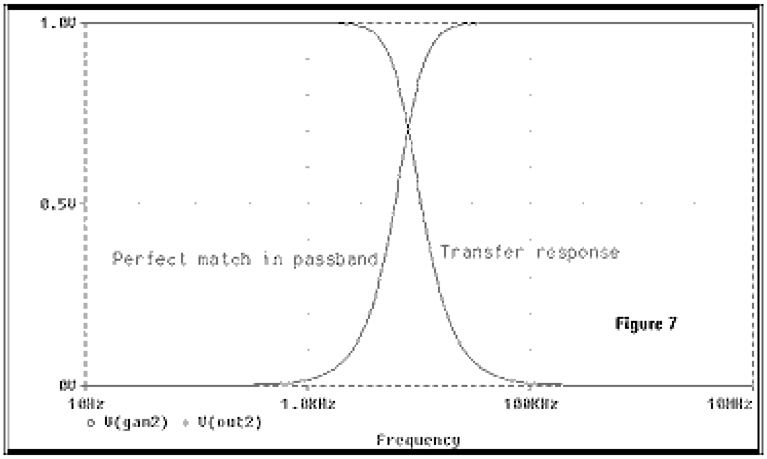

Figure 7

Now let's use the low pass and put a high pass with it to try to form a diplexer. This is shown in Figure 8 where we now have just guessed at component values. The response, shown in Figure 9, has high pass and low pass outputs that we might expect. The match is good at the frequency extremes, but is only so-so in the transition band.

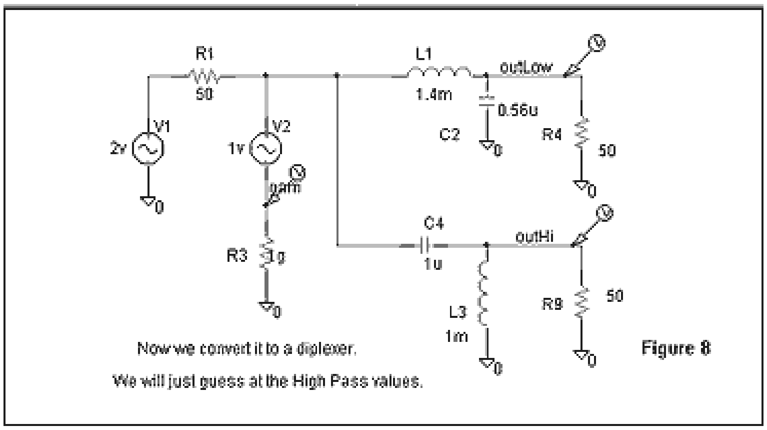
R1
50
V1
2v
V2
1v
gam
R3
1g
L1
1.4m
outLow
C2
0.56u
R4
50
C4
1u
outHi
L3
1m
R9
50
0

Now we convert it to a diplexer.

We will just guess at the High Pass values.

Figure 8

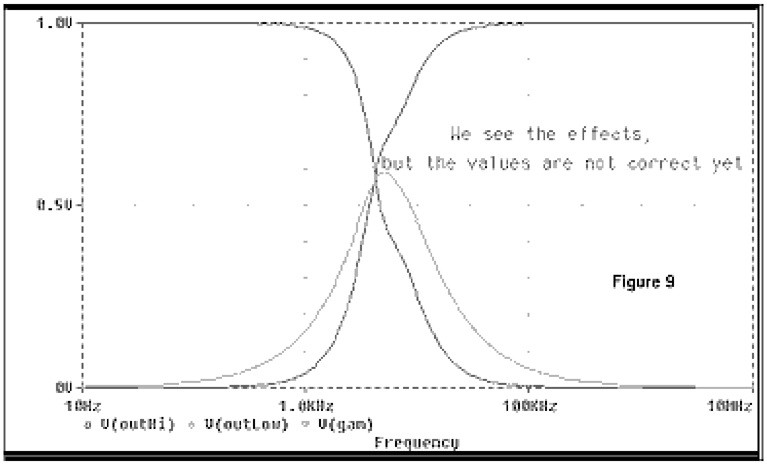

We see the effects,
but the values are not correct yet

Figure 9

Let's now look at a carefully designed pair of two element filters. The circuit is Figure 10 an is a final example. The corresponding response is Figure 11. It is hard to see, for the response merges in with the baseline. However, the reflection is zero and it is zero everywhere. This filter was designed for a 1 kHz crossover, so it can be scaled to other frequencies with ease.

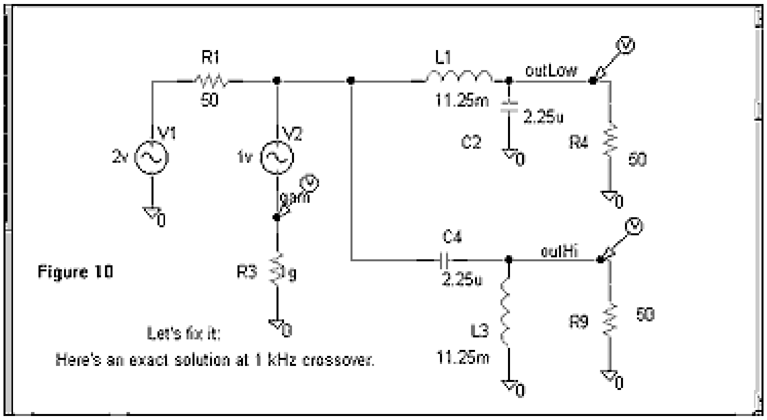
R1
50
V1
2v
V2
1v
gam
R3
1g
L1
11.25m
outLow
C2
2.25u
R4
50
C4
2.25u
outHi
L3
11.25m
R9
50
0

Figure 10

Let's fix it:
Here's an exact solution at 1 kHz crossover.

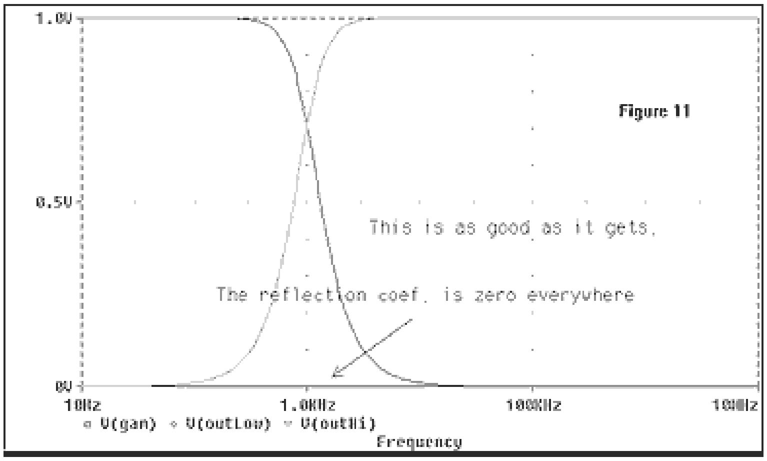

Figure 11

This is as good as it gets.

The reflection coef. is zero everywhere

Figure 12 is another final audio example. This circuit is very similar to the one used in the past by Roy Lewallen, W7EL, although the inductor was smaller at 100 uH in his Optimized rig. The response of this diplexer is shown in Figure 13. This is not perfect, but it is probably quite a good performer in typical receiver situations.

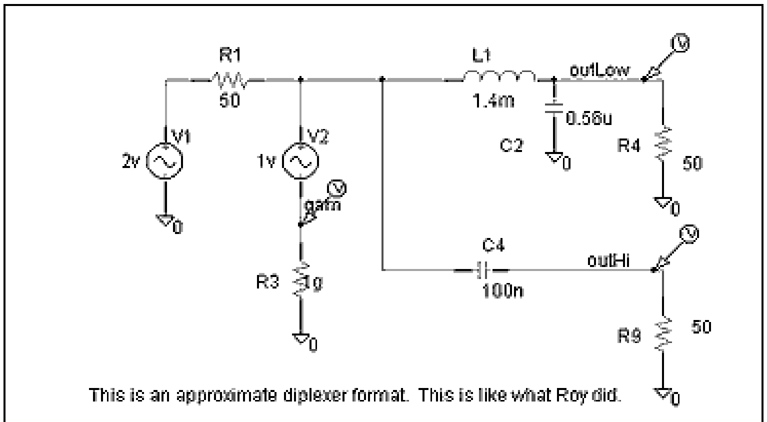
R1
50
V1
2v
V2
1v
gam
R3
1g
L1
1.4m
outLow
C2
0.56u
R4
50
C4
100n
outHi
R9
50
0

This is an approximate diplexer format. This is like what Roy did.

Figure 12

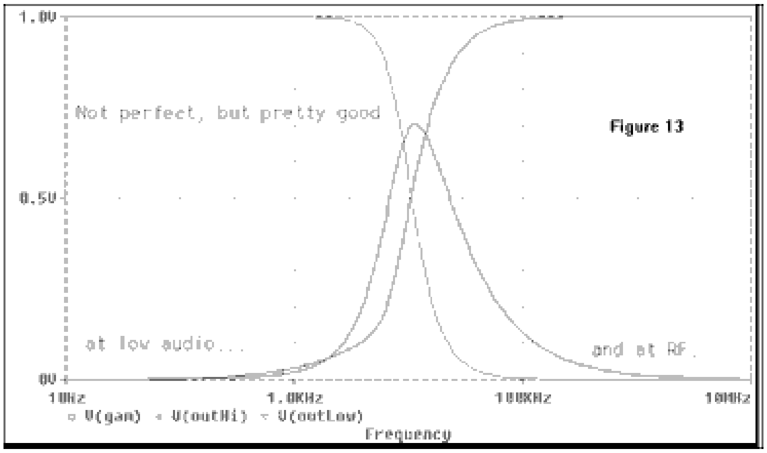

Not perfect, but pretty good

Figure 13

at low audio...

and at RF.

Finally, here's a higher frequency example. 5th order low pass and high pass filters are combined. The filters have a cross over at about 150 MHz. Note that there is a slight reflection in the transition band. This is probably just the result of our having rounded some values in the design process. Figures 14 and 15. An outstanding reference on this is Nic Hamilton, G4TXG, "Improving Direct Conversion Receiver Design," Radio

Communications, April 1991.

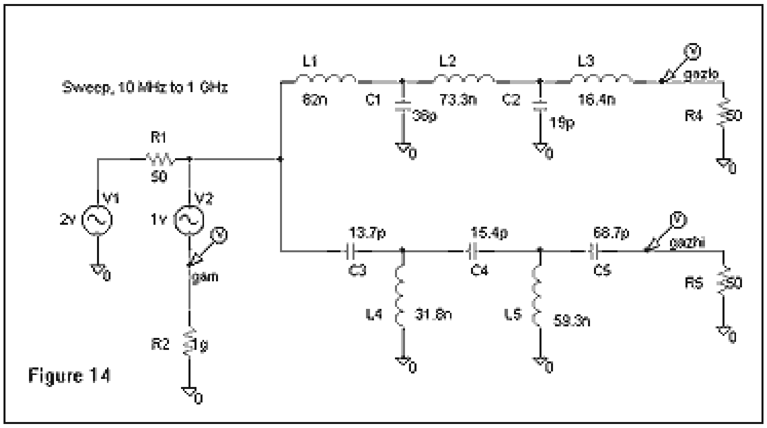

Figure 14

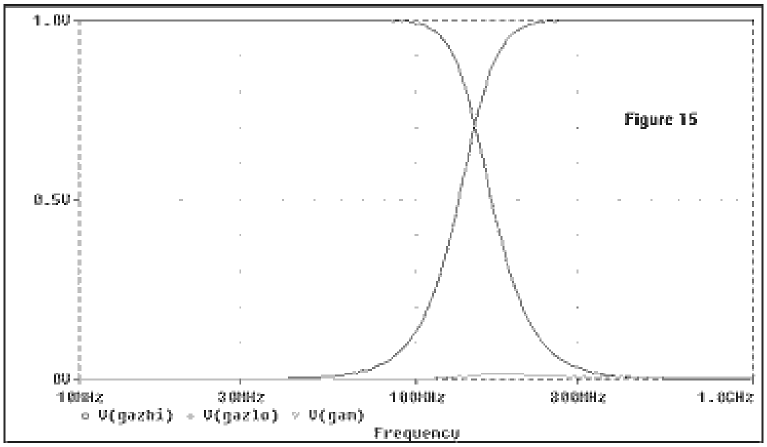

Figure 15

Bridge-Tee RF Diplexer

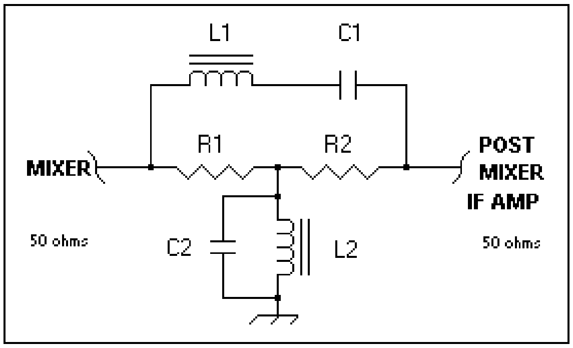

This is an excellent bandstop/bandpass diplexer popularized by Joe Reisert W1JR. This easy to build diplexer has a low parts count and is easily built using Ugly Construction. Resistors R1 and R2 present a 50 ohm impedance to the mixer output and a 50 ohm impedance to the input of the post mixer amplifier. The IF frequency is passed through the diplexer while out of passband RF is given a low impedance path to ground. The capacitance for C1 is generally built up by substituting the nearest standard value capacitor or by placing 2 or more capacitors in parallel with each other to achieve the desired value. The same procedure is then repeated for the C2 capacitance. For more strenuous purposes, a portion of C1 and C2 or the inductors L1 and L2 can be variable and adjusted on the bench. The inductors can easily wound on powdered-iron toroid cores. I have used T50-2 or T50-6 type toroids with good results. The Q of the inductors is 1.

It is possible to design a more generalized form of this diplexer with a higher loaded Q in the resonators. The diplexer shown and used in the program has a Q of 1. This was used by W1JR in his VHF/UHF World Column in the now defunct HAM Radio Magazine for March and November 1984. It was also more recently used by Jacob Makhinson, N6NWP in his A High-Dynamic Range MF/HF Receiver Front End in QST for February 1993. The actual formulae for this diplexer is far more complex than the simplified formula shown below or used in the program, but both provide a very good approximation for the Q = 1 version as used by W1JR and N6NWP. If you wanted Q=10, the series tuned circuit would use L that is 10 times as high with C to resonate. The parallel tuned circuit would then use C that was 10 times higher with L to resonate.

A supplemental web page with some hard-core mathematics for this diplexer can be found on the Diplexer Supplemental Page.

Simplified Formulae (Q = 1):
R1 and R2 are always 51 ohm resistors.
Inductors L1 and L2 -> 50 / (6.283 * frequency in Megahertz)
Capacitors C1 and C2 -> 1 / (6.283 * 50 * frequency in Hertz)

Example 1: For a 9 Mhz IF , L1 and L2 = 0.88 microhenrys and C1 and C2 = 350 picofarads
Example 2: For a 4.92 MHz IF , L1 and L2 = 1.62 microhenries and C1 and C2 = 647 picofarads

I wrote a simple program to do the math for the Q = 1 version. Download the Bridge-Tee RF Diplexer Diplexer Program

---

Comments and analysis by W7ZOI

This is a double ended version of the first order bandpass/bandstop design presented earlier. But it's a good one, within the constraints of what it can do. The first is the simulation schematic for the diplexer, which is better termed a Bridge-Tee Diplexer. (There are bridge Tee filters and attenuators too.) That figure is entitled Figure 16.

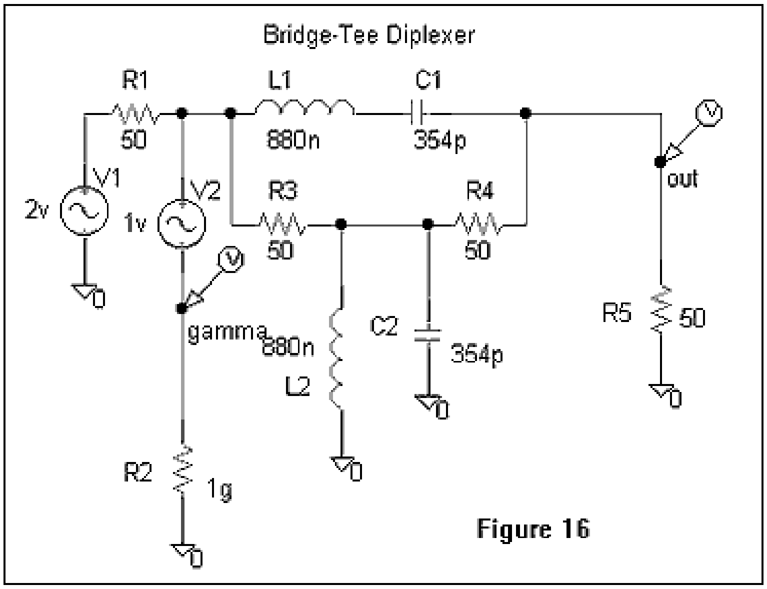
Bridge-Tee Diplexer
R1
50
L1
880n
C1
354p
V1
2v
V2
1v
R3
50
R4
50
out
gamma
880n
L2
C2
354p
R5
50
R2
1g
0

Figure 16

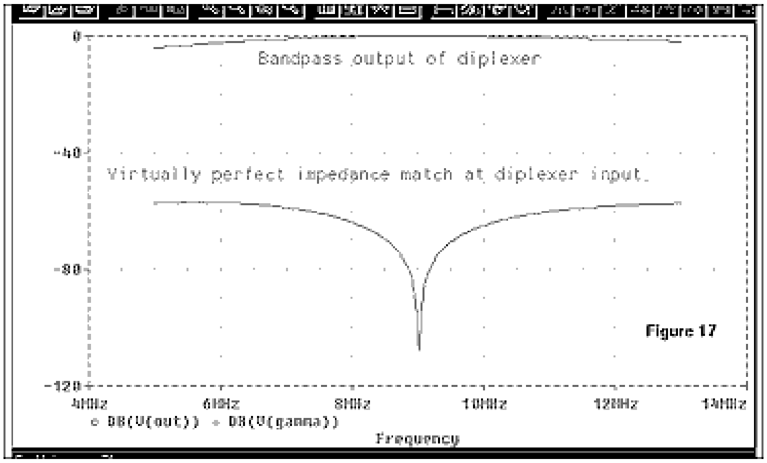

Bandpass output of diplexer

Virtually perfect impedance match at diplexer input.

Figure 17

The response for this circuit is shown in Figure 17. This is extremely good. The through response is very flat owing to the low Q of the series tuned circuit. But even better is the match. It is very good. Indeed, it would have been perfect except for slight roundoff errors that occurred as we designed the networks.

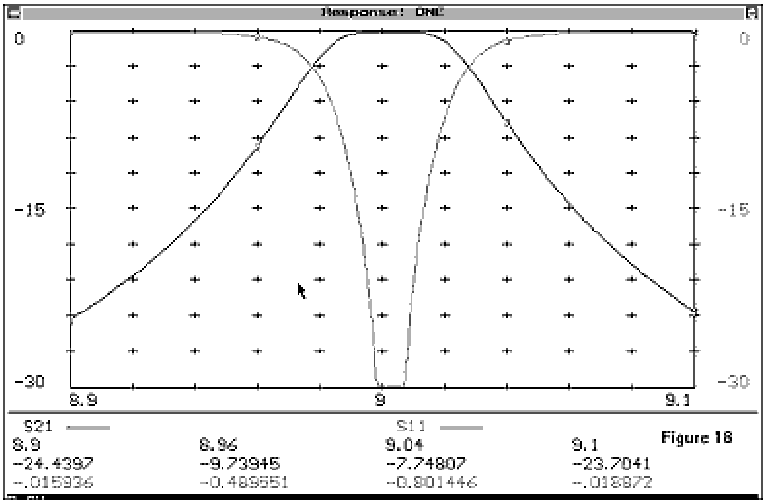

```
S21 ——                     S11 ——
8.9          8.96          9.04          9.1
-24.4397     -9.73945      -7.74807      -23.7041
-.015936     -0.489551     -0.801446     -.018872
```

Figure 18

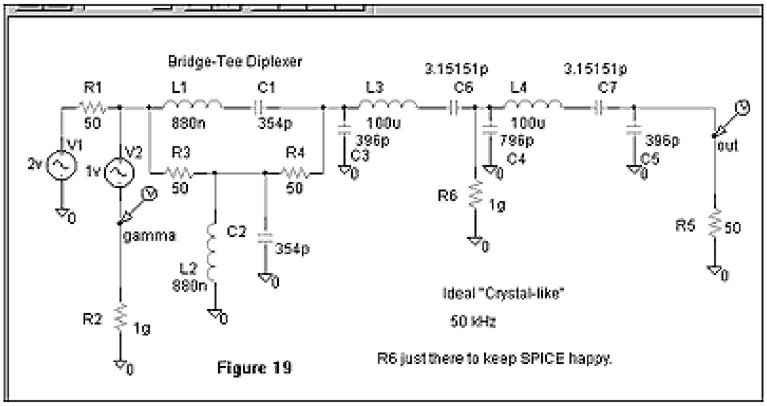
Bridge-Tee Diplexer
R1
50
L1
880n
C1
354p
L3
100u
3.15151p
C6
L4
100u
3.15151p
C7
V1
2v
V2
1v
R3
50
R4
50
396p
C3
796p
C4
396p
C5
out
gamma
C2
354p
L2
880n
R6
1g
R5
50
R2
1g
0
Ideal "Crystal-like"
50 kHz

Figure 19

R6 just there to keep SPICE happy.

This kind of thing works fine if you really have a perfect match following the diplexer. But what if you don't. There are some places where they do not do the job that some folks think they will do. For example, a diplexer WILL NOT cause the impedance to be flat if it is followed by a filter. The diplexer must still be properly terminated at both output ports. In Todd's usual applications, he is worried about providing a good mixer termination for a product detector. The audio amp that he uses will usually have a common base first stage and that will present a good wideband load to the diplexer, so he is okay. But other folks have placed a diplexer after a switching mode mixer that then drives a narrow filter. The diplexer then does little good. To illustrate this situation, I designed a "crystal like" two pole LC bandpass filter with a 50 kHz bandwidth. This represents the general case where we try to put a diplexer between a mixer and a filter. The filter response by itself is shown in Figure 18. The schematic for the diplexer and following filter is in Figure 19.

The response for the combination is in Figure 20. Here we see a passband response that is fine; it's just the repeat of the filter response we already saw. However, the input impedance looking into the diplexer, the impedance that would be seen by a mixer, is terrible. The return loss is 0 dB at all frequencies except where we get within the passband of the filter.

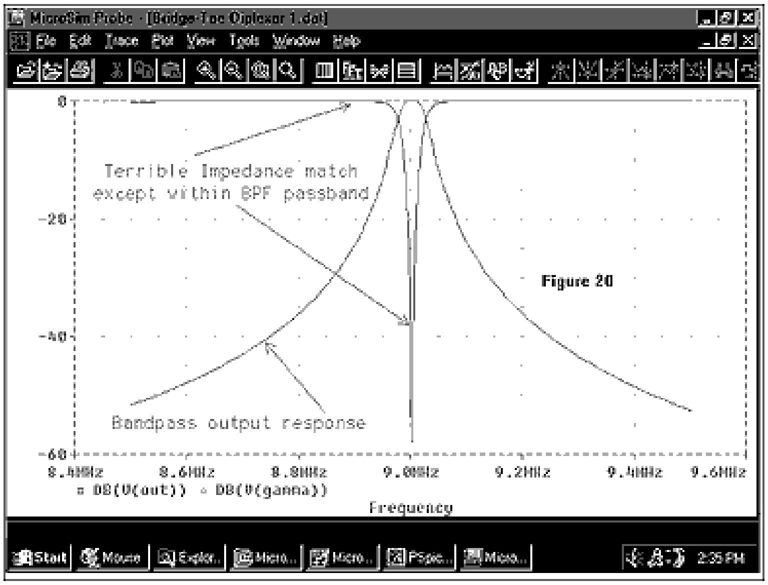
MicroSim Probe - [Bridge-Tee Diplexer 1.dat]
File Edit Trace Plot View Tools Window Help
0
-20
-40
-60
Terrible Impedance match
except within BPF passband
Figure 20
Bandpass output response
8.4MHz
8.6MHz
8.8MHz
9.0MHz
9.2MHz
9.4MHz
9.6MHz
DB(V(out)) DB(V(gamma))
Frequency
Start
Mouse
Explor...
Micro...
Micro...
PSpic...
Micro...
2:35 PM

Building a diplexer to follow a product detector is not a cheap endeavor. Audio inductors and capacitors such as metalized polyester film types are not common in many builder's junk boxes. I really like the design shown in Figure 10 and wished to use it because it uses just 2 inductors and capacitors which is in keeping with the popcorn nature of this website. The main difficulty is that the inductors and capacitors are not standard value types and series connecting components to achieve the desired values would add to both the cost and size of the finished product.

Obviously, it will not likely match from DC to daylight. That is not the intention of this simple design or this web site in general. I asked Wes to place just 2 standard value capacitors and inductors in the Figure 10 diplexer design and see what happens.

Here was his response to my request:
OK, here are some "practical values." Note that things don't really change that much. We start with 11.x mH and 2.25 uF. Change the inductor to 10 mH and get Figure 21. Then change the cap to 2.2 uF and see almost no change in Figure 22.

Practical Diplexer for Popcorn Receivers

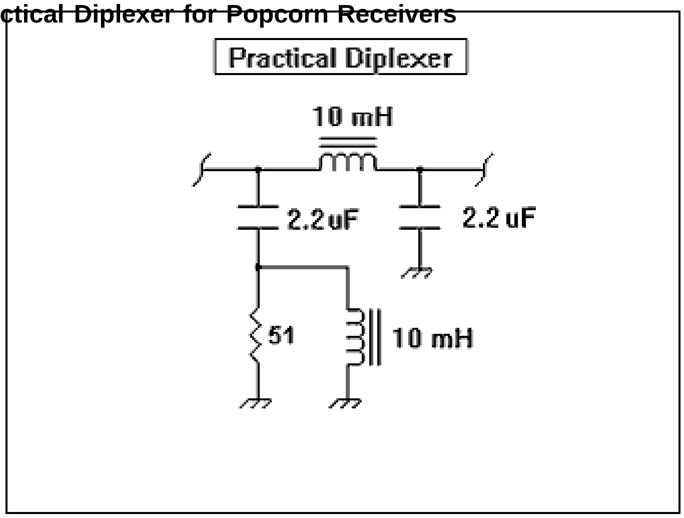
Practical Diplexer
10 mH
2.2uF
2.2 uF
51
10 mH

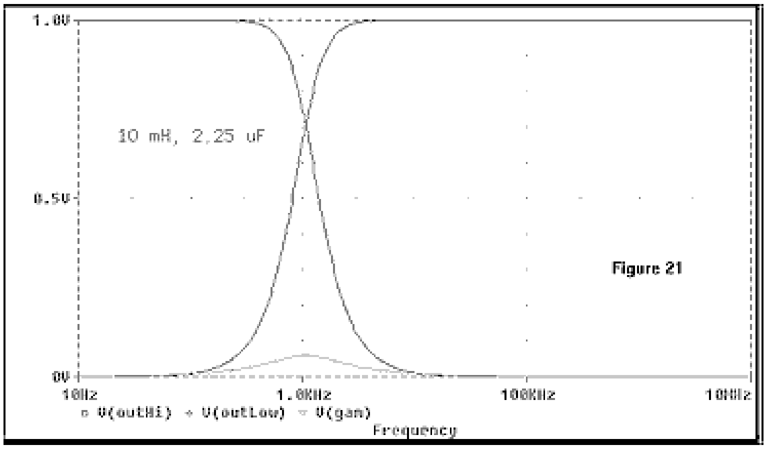

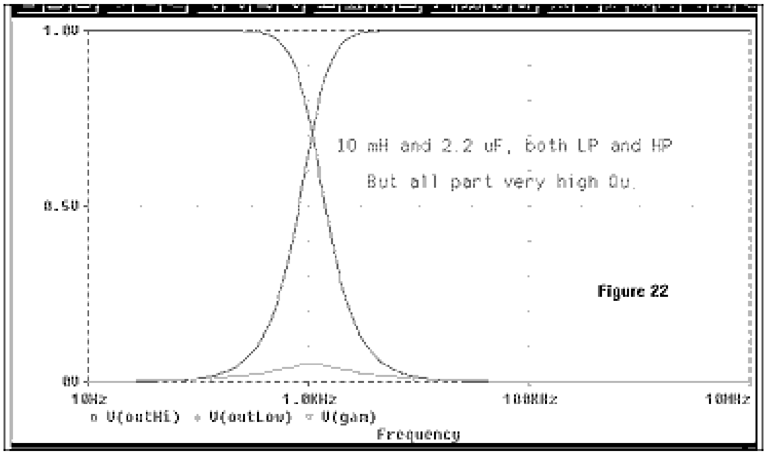

But now move into the world of even greater reality and acknowledge that many of the inductors we use at audio are very low Q. Change Qu of L to 10 at 1 kHz, so I put 6.3 Ohms in series with each L to get Figure 23. And do the same thing, but with a dB scale, for Figure 24. Note that we can see the difficulties, but things are still pretty good. We see some loss (about 1 dB) in the low pass path and less than perfect match. But the match is still very good. 20 dB is about 1.1:1 vswr, much better than 99.9% of the hams can really measure. (A 10 dB match is about 2:1.) Hope this is what you were after....Wes

It was and I will use this "practical" diplexer in my next popcorn DC receiver project. Note that the practical diplexer input and output impedance is 50 ohms and the 2.2 uF caps should not be polarized capacitors such as regular electrolytic types which have a positive and negative polarity.

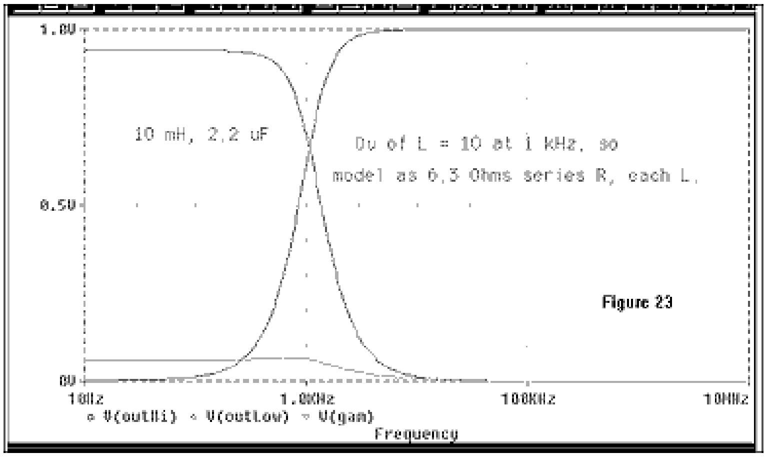

10 mH, 2.2 uF

Qu of L = 10 at 1 kHz, so model as 6.3 Ohms series R, each L.

Figure 23

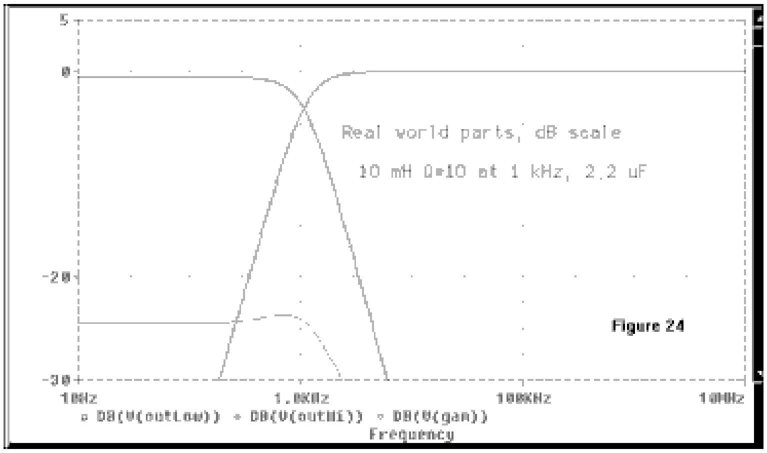

Real world parts, dB scale

10 mH Q=10 at 1 kHz, 2.2 uF

Figure 24

Many thanks to Wes Hayward, W7ZOI for his work on this page.
A version of this web page in Russian Cyrillic

Ugly Construction

Discussion

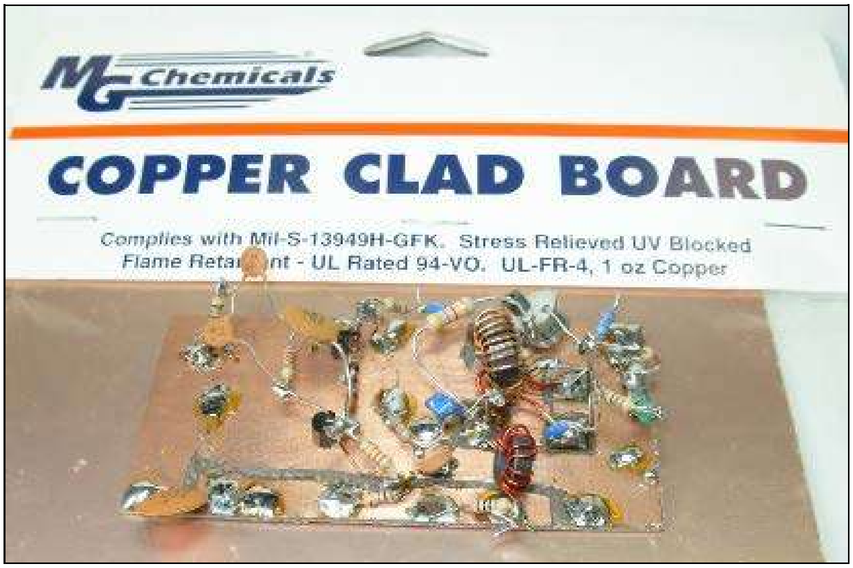
MG Chemicals
COPPER CLAD BOARD
Complies with Mil-S-13949H-GFK. Stress Relieved UV Blocked
Flame Retardant - UL Rated 94-VO. UL-FR-4, 1 oz Copper

We enjoy many ways to build electronic circuits. For example, you might breadboard on a perforated circuit board, an etched PC board, a sheet of copper clad board, or even a piece of copper wire. In the hay day of tube electronics, builders used terminal strips and point-to-point wiring within the project chassis. I mostly breadboard using Ugly Construction.

Ugly Construction, "dead bug", or "ground-plane construction" involves building circuits on top of a double or single-sided copper clad board (copper side up for single-sided board). The copper ground-plane provides a low impedance ground and mechanically supports the parts soldered to it. Component leads requiring grounding are soldered *directly* to the copper surface, while the ungrounded leads of these parts anchor any ungrounded parts connected to them. Isolated sections called stand-offs hold other ungrounded or remotely located parts.

Example stand-offs include high value resistors (10 Megohm or greater), terminal strips, or small copper islands glued onto or cut into the copper ground-plane. Parts such as transistors, IC's or commercial diode ring mixers are generally flipped upside down and anchored by their grounded lead(s). Metal encased parts such as crystals can be grounded by a short wire or directly soldered upside down to the copper board.

DC voltage wires, or decoupling resistors may be supported by soldering 1 lead of a bypass capacitor to the ground plane while the other lead holds the DC voltage carrying part up off the copper board a short distance. I mostly use grounded caps for stand-offs and not 10M resistors.

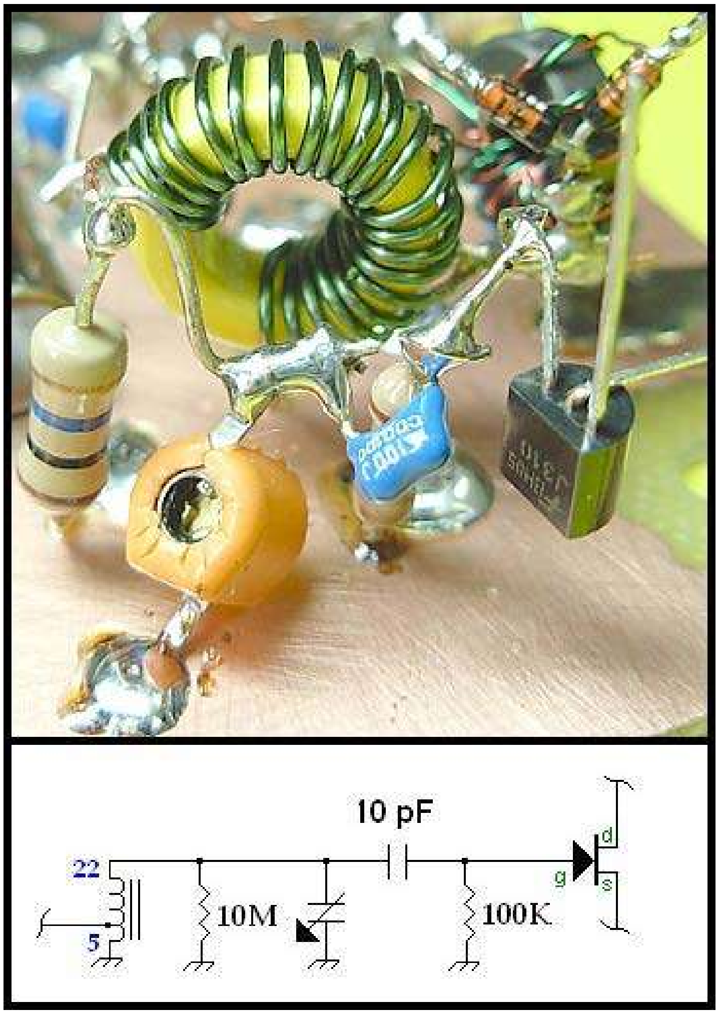

Shown above — a 10 Megohm, half-watt, stand-off resistor anchors the "hot" inductor terminal plus supports the ungrounded trimmer capacitor terminal. The coil's 24 gauge wire provides additional mechanical stability. The signal loss from adding the 10M resistor was about 0.1 volts peak to peak in 1 experiment.

Ugly Construction allows the experimenter total control over the design of a project and in my opinion, its greatest strength is *speed*. Ugly Construction yields rapid and flexible bread boarding — very appealing for prodigious home builders.

The Origin of the Term "Ugly Construction"

Roger Hayward, KA7EXM and Wes Hayward, W7ZOI coined the term "Ugly Construction" while writing the "Ugly Weekender" published in the August 1981 issue of QST. I asked Wes about this in 2009. The term was a takeoff from the 1958 book entitled The Ugly American by William Lederer and Eugene Burdick.

A big part of the learning of the QST article was Ugly Construction! The term and the bread boarding technique emphasized the fact that there is no correlation between the "prettiness" of a construction project and the way it works. According to Wes, the goal had a couple of corollaries. "First, people like myself who do NOT have the knack for doing pretty projects can still build successful radios. Second, is that we all need to look at our projects after the fact to discover what it is that really makes them work well. In the case of the Ugly Weekender, the thing that makes it fly is that there is a wonderful ground plane with that PC board material".

Indeed, this transceiver functions very well; especially after you temperature compensate the VFO. Wes also built versions for the 30 and 80 meter bands. I have versions on 15, 40 and 80 meters. The transmitter portion is a true QRP classic; both as a Ham radio transmitter and because it promoted "ugly" ground-plane or dead bug bread boarding techniques to the scratch homebrew community.

Classic Ugly Construction

This term emerged in Spring 2010 and describes the archetype popularized by Roger and Wes Hayward. All ungrounded leads not anchored to other parts are attached to the copper ground-plane via high ohm resistors — *no* glued pads or islands cut in the copper board.

In-situ comparisons of a 10 megohm resistor versus islands cut into the ground plane and glued-on Manhattan-style pads demonstrated that the resistor had the lowest capacitance; around 1 pF versus 4 pF or greater for the pads or islands. Click for a high resolution transmitter chain built with 100% Classic Ugly Construction. Click for a crystal oscillator.

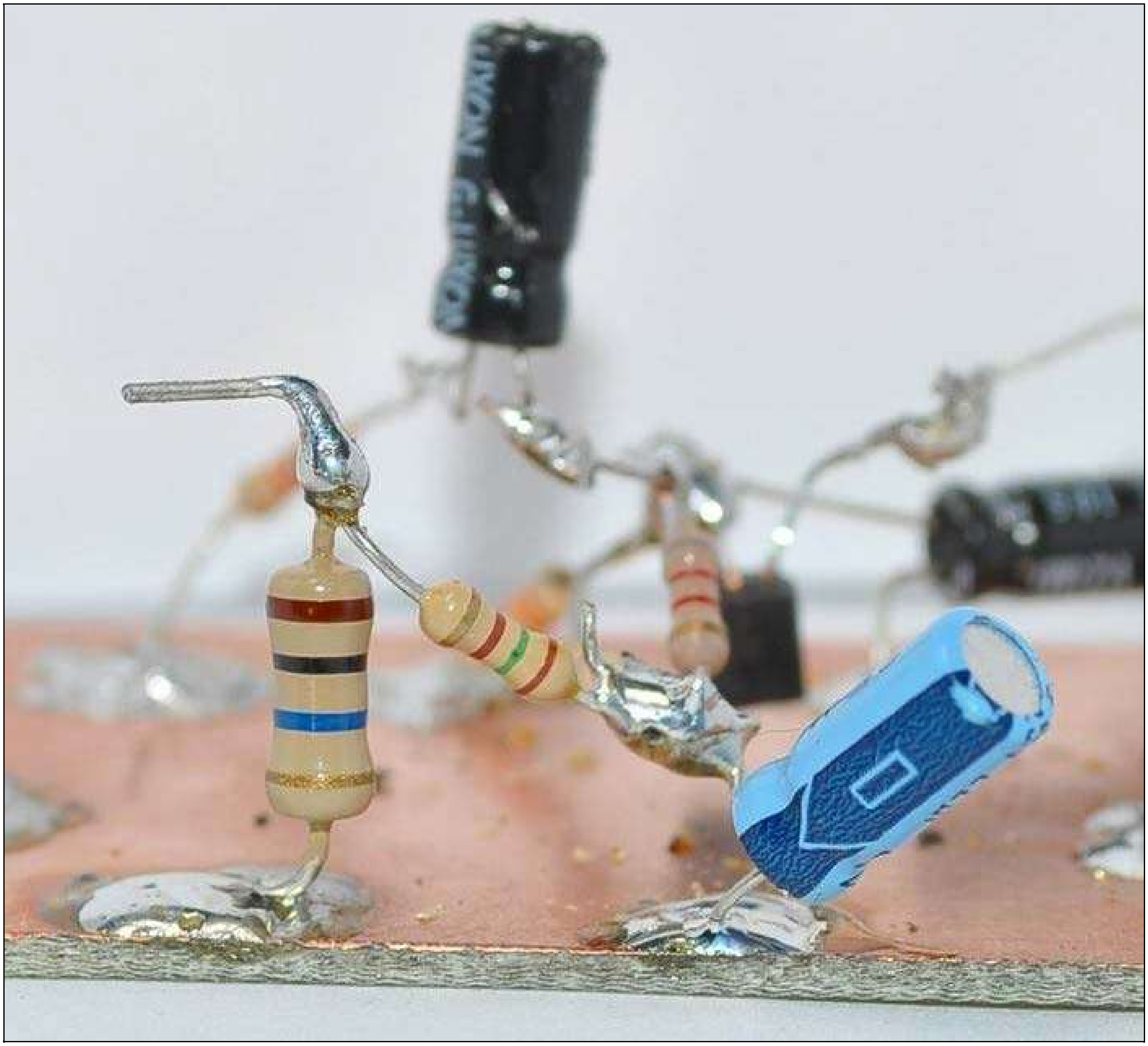

Above — *Classic* Ugly Construction using a high ohm stand-off resistor. The top of the 10 megohm resistor is the VCC connection point. It feeds a 150 ohm / 47 uF decoupling network connected to a transistor collector resistor. Bypass capacitors also serve as stand-offs and I prefer thick lead (100 volt or greater) caps for stand-off duty.

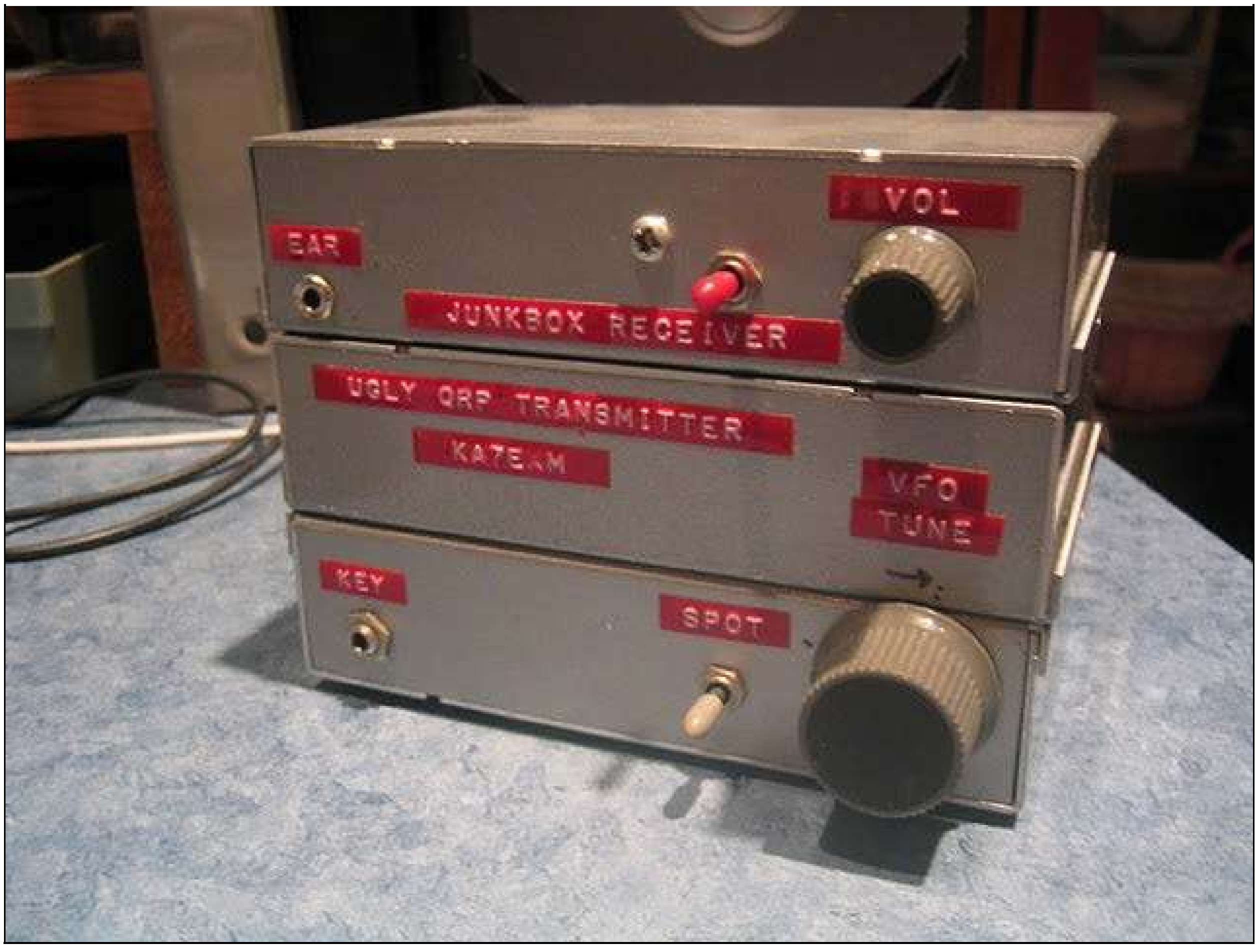
VOL
EAR
JUNKBOX RECEIVER
UGLY QRP TRANSMITTER
KA7EXM
VFO
TUNE
KEY
SPOT

Above — The original Ugly Weekender. Photograph used by permission of Roger, KA7EXM.

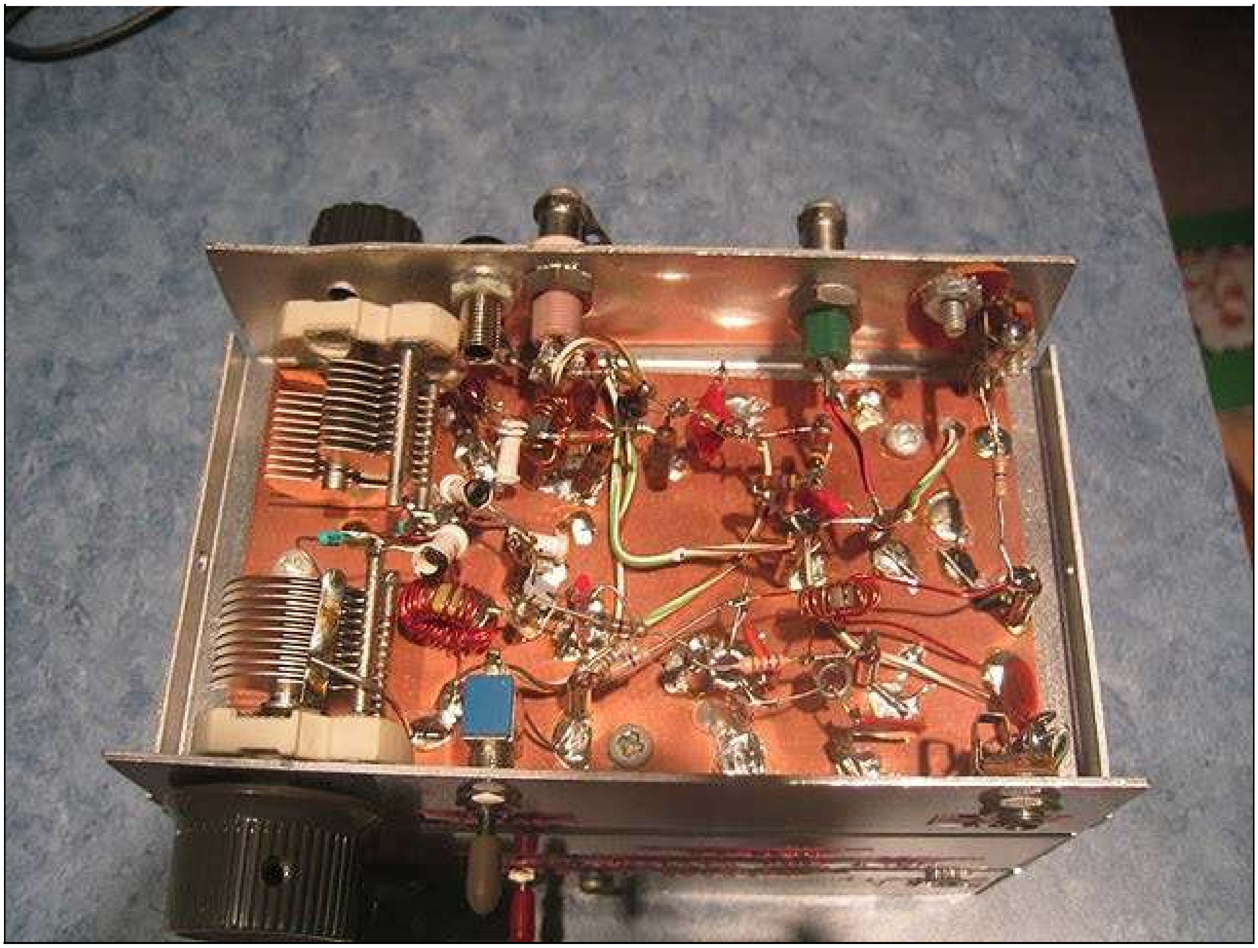

Above — The original Ugly Weekender. Now this is *Classic* Ugly Construction. Please refrain from building this transceiver and adding modern notions such as a PIC microcontroller keyer — that's just wrong!

Photograph used by permission of Roger, KA7EXM.

Ugly Construction Variants

The most popular Ugly Construction variant is called Manhattan style. Manhattan or "paddy board" construction uses small square or round pads cut or stamped from PC board that are glued copper side up onto a large copper clad board also placed copper side up. The small pads or "islands" serve to anchor ungrounded components. Components soldered to the pads such as transistors or ICs are generally not positioned upside down like in Classic Ugly Construction. Many Manhattan style builders use IC sockets as well. These hobbyists sometimes build beautiful looking layouts — Manhattan is a wonderful bread boarding technique. Google for more information. The best Manhattan construction and copper board chassis bashing I've seen comes from Dave, AA7EE. Click for a blog describing his version of the WBR regenerative receiver. *Linked with the permission of Dave, AA7EE.*

Another interesting variant is used by Dick Pattinson, VE7GC. The circuit board is placed copper side up and holes are drilled and countersunk so that the holes are isolated from the ground plane. Ungrounded components are connected underneath the main board through the countersunk holes. There are many such variations. Each Ugly Construction variant has advantages and disadvantages.

On this website, stand-offs are created by cutting a few lines into the copper board with a small, motorized hobbyist cutting tool; with high megohm value resistors, and occasionally by a small Manhattan style pad or 2. Manhattan pads are great for supporting components needing solid anchoring such as a trimmer capacitor or potentiometer.

Classic Ugly Construction dominates circuits breadboarded after May 2010.

The motor tool may also be used to grind off the copper underneath where VFO toroidal inductors will lie, so that the inductor Q is not effected by the being glued onto a copper surface. In audio projects, I may also grind off the copper around the copper board mounting bolts so that they are isolated from the chassis and do not provide multiple grounds and create the potential for ground loops.

Is Ugly Construction Less Reproducible than Manhattan?

I've received well over 1000 emails about Ugly Construction since launching the site in 1998. Some feel that circuits made with Manhattan pads are more reproducible than Classic Ugly Construction builds. This might be true, but to my knowledge nobody has performed a comparison trial.

The important question is why would this happen? I've read/heard opinions that the stray L and C from the long component leads associated with Ugly Construction might wreck circuit reproducibility, but respectfully disagree from DC to HF. In microwave breadboards, we fabricate lumped element inductors and capacitors (*i.e. precisely dimensioned Manhattan pads*) right into the PC board — Manhattan pads glued all around a breadboard may potentially exhibit much more stray L and C than a few component leads in an Ugly build. Also, wise Ugly builders keep their lead lengths short where it counts: for example, RF bypass + ground and at the input/output of a BJT or FET that offers gain into UHF.

I conjecture human error probably inflicts more problems for Ugly Construction builders — Manhattan building, with its slower pace might trigger less mistakes by newbies. Still, too, Manhattan builders tend to make prettier, squared and aligned circuits and it's easier to spot trouble — plus they look nicer in photographs and some builders carefully document and photograph their builds for others to admire and strictly copy. I've see Manhattan build photos where every resistor tolerance band pointed in the same direction — wow! I think it might be difficult to put such a 'work of art' into an RF-tight metal box for much-sought isolation.

Further, in Ugly Construction — upside down parts might wreak havoc on the "spatially challenged" builder. Who knows? I'm comforted knowing that kit sellers who provide a screened printed circuit board with explicit instructions, still must provide major email support to mitigate build errors. To err in an ugly fashion is human?

Whatever variant of construction you choose, it's sure to be a winner!

Further Discussion

Wire

Non-stranded (solid core) copper wire such as the 22 AWG 3-color package sold by Radio Shack seems a good choice for hook up wire. With non-stranded wire, you do not have to worry about little stay wire hairs causing shorts and it's easier to wrap around components leads. I use red for wires that carry positive voltage, green for grounding and black for wires that carry AC signals short distances. In addition, RG-174 or shielded wire is used to carry AC signals for distances greater than 10 cm, and for connecting stages requiring 50 ohm input or output impedances such as diode ring mixers or low-pass filters.

2781306
NEXXTECH
Hook Up Wire
22 Gauge Solid

Your Health

Please consider the following safety comments:

For regular soldering, ensure ventilation of your room — flux fumes can be harmful. Open your shop window and/or use a small fan to improve fresh air intake;

Whenever possible, perform high wattage soldering outdoors;

When grinding paths on copper clad boards, wear a small particulate respirator, gloves, plus ear and eye protection and most importantly; **do it outside** for yours and your family's health;

Wash your hands after soldering and handling freshly cut, fiberglass dust laden copper clad boards.

Soldering Irons

For soldering copper clad boards together, AC grounds on tube guitar amp chassis and performing antenna work, I currently use a Weller SP 80L (80 watt) soldering iron. It is heavy and unwieldy, so you have to be very careful when its plugged in. These high wattage soldering irons produce lots of smoke.

Weller
SP 80L

Shown above is an 80 watt "heat torch". My main soldering irons are typically in the 30-35 watt range. Consider keeping at least 1 back up soldering iron, as you never know when a soldering iron is going to burn up. My current 35 watt iron is shown below. These Weller irons have a built in lamp which lights when they are plugged in; a very nice feature. I also keep a small stock of new soldering iron tips.

Copper Clad Board

Some builders ask about sources for copper clad board. I personally use boards made by MG Chemicals as they have dealers in my city and are reasonably priced and good quality. Try the search words **copper clad board** plus **your country name** in your favorite web search engine. A few links follow, but as I have only used boards sold locally, I can't comment about the online companies.

MG Chemicals Worldwide distributer index

Electronic Goldmine Online store

Circuit Specialists Online store

Miscellaneous

Shown below is a schematic and the Ugly Constructed version of it.

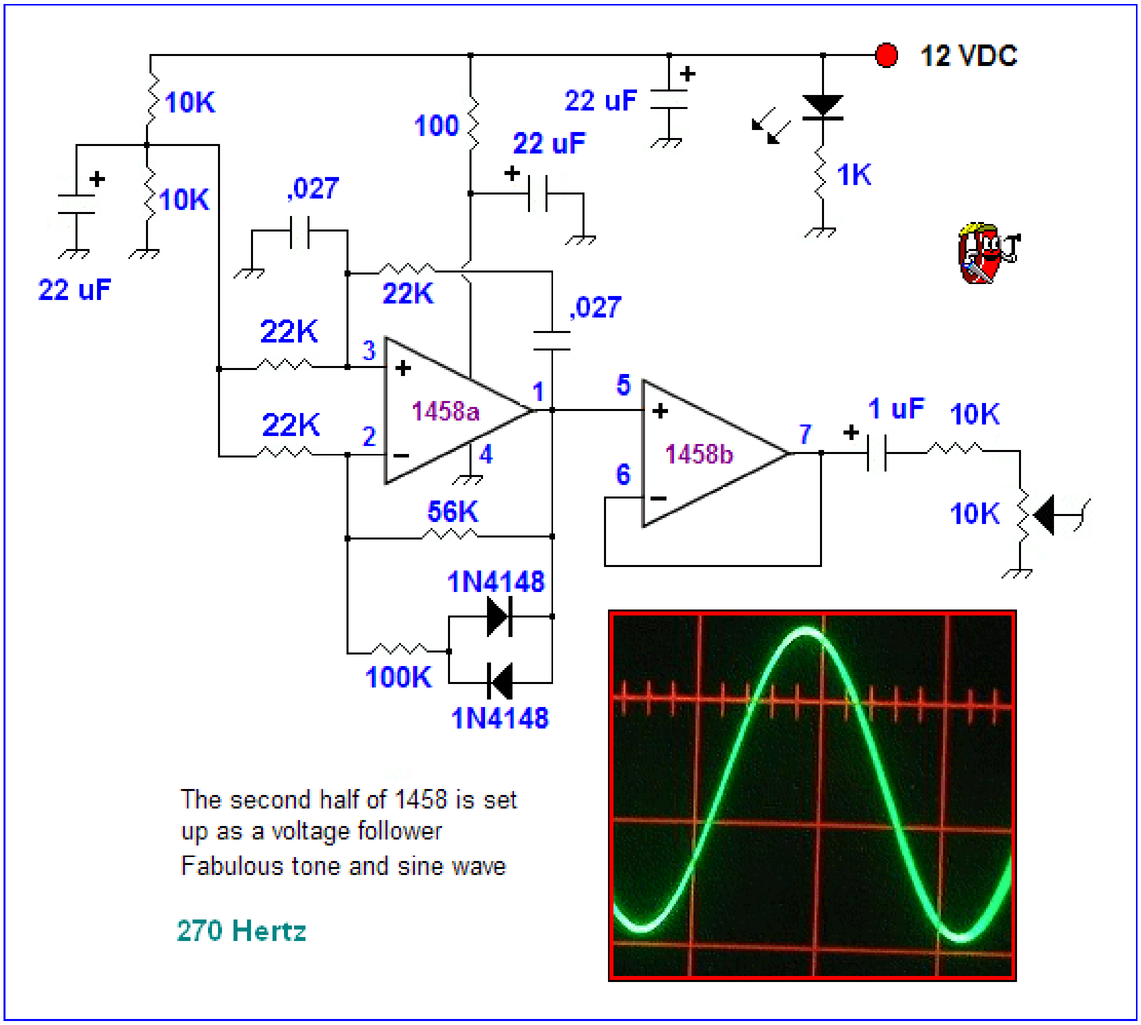
12 VDC
10K
100
22 uF
22 uF
1K
10K
,027
22 uF
22K
,027
22K
3
+
1458a
1
5
+
1458b
7
1 uF
10K
22K
2
−
4
6
−
10K
56K
1N4148
100K
1N4148
The second half of 1458 is set
up as a voltage follower
Fabulous tone and sine wave
270 Hertz

Above. The schematic of an adapted sine wave audio frequency oscillator taken with permission from EMRFD , Figure 12.4. EMRFD is the main reference for this web site. The original schematic author was Wes, W7ZOI

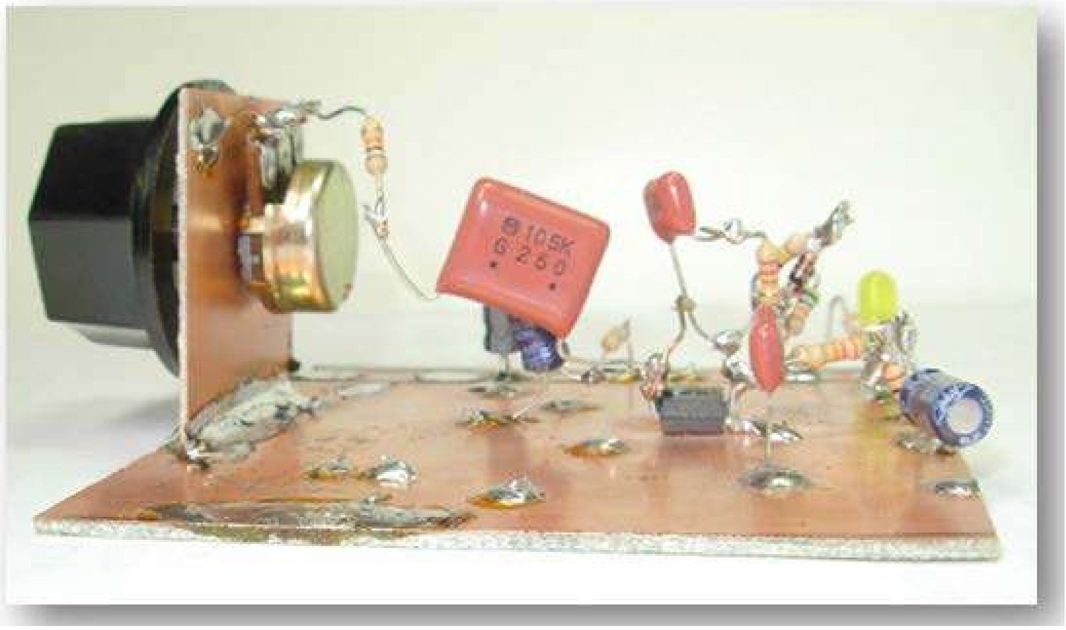

Above. I built this circuit from start to scope in about 25 minutes. This was a scrap, pre-used board with a positive voltage path and a potentiometer holder already on it When miniaturization is not your goal, construction is much easier and faster. I re-use parts and boards to save money. You may remove entire stages from 1 board and solder them onto another.

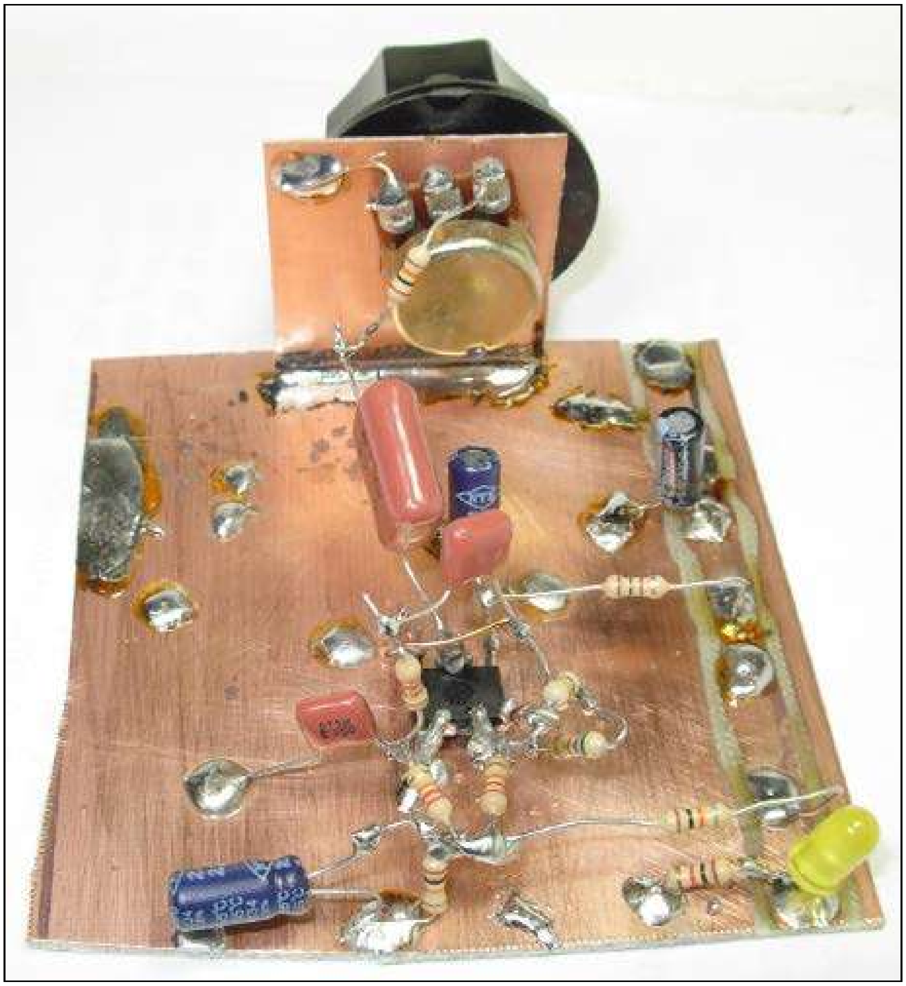

Above. Note how the 10K output potentiometer holder is soldered to the main copper clad board. The grounded 10K resistor is used to anchor the 22K resistors connected to IC pins 2 and 3 and can be seen in the foreground.

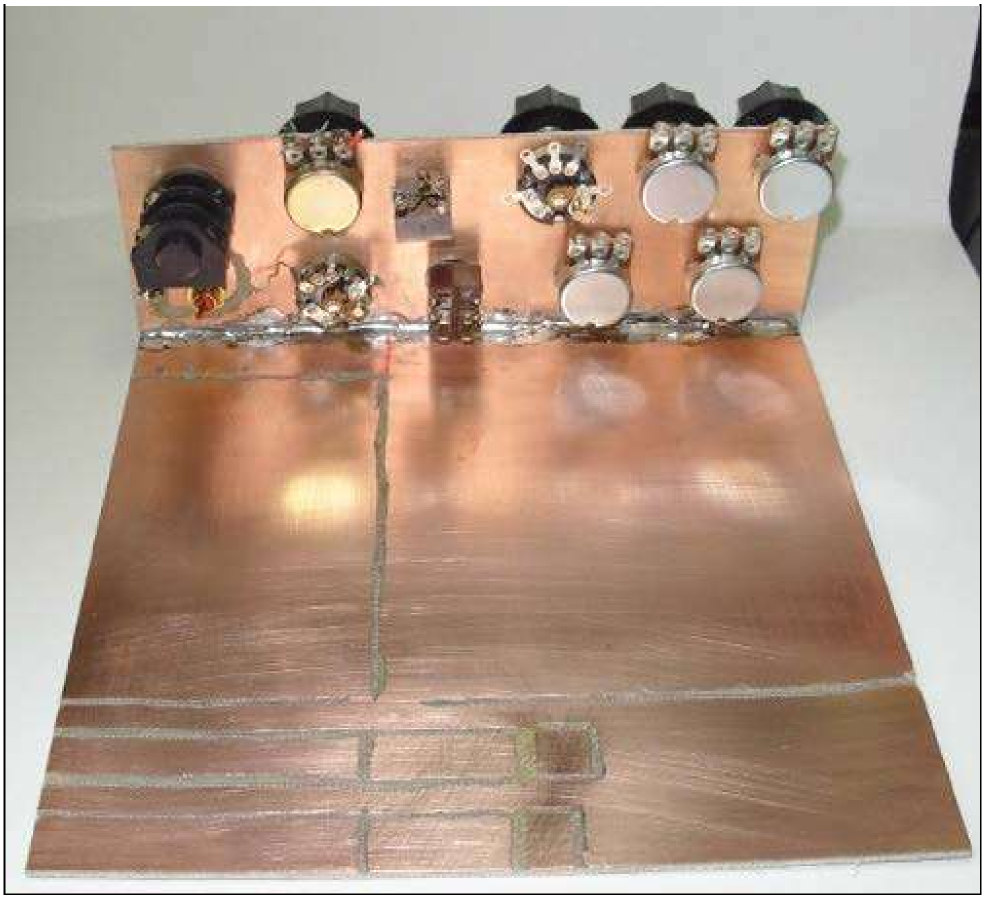

Shown above is another project. Entire control panels can be built from copper clad board for prototype circuits. In this board are numerous cut paths, 7 potentiometers, 3 jacks and a switch. Do not build an LC VFO over double sided copper clad board; lest it become "a capacitor" and affect your frequency stability.

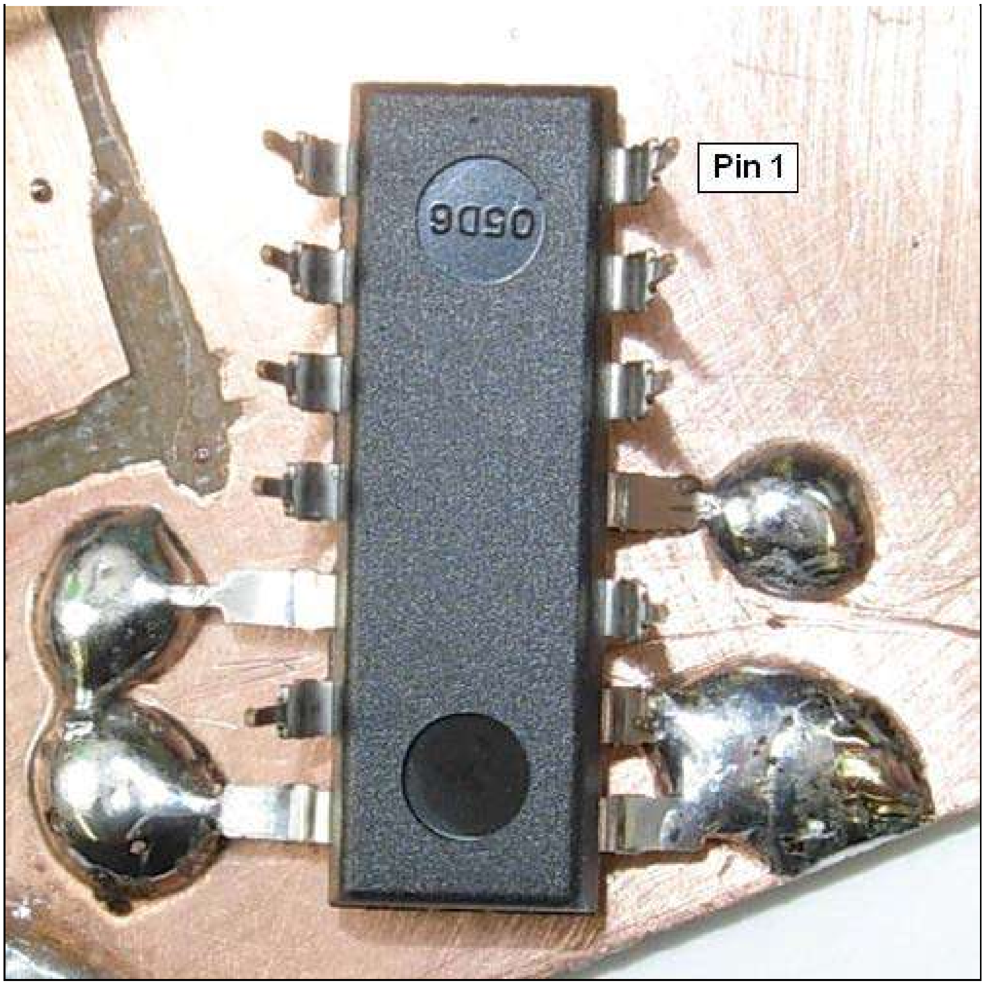

Shown above is a CD4013B D Flip-Flop soldered "dead bug" style. Pins 4, 6,7, 8 and 10 are grounded to the copper surface; well anchoring this part. Using proper static precautions, I have never had a CMOS device failure using Ugly Construction and also save the price of an IC socket.

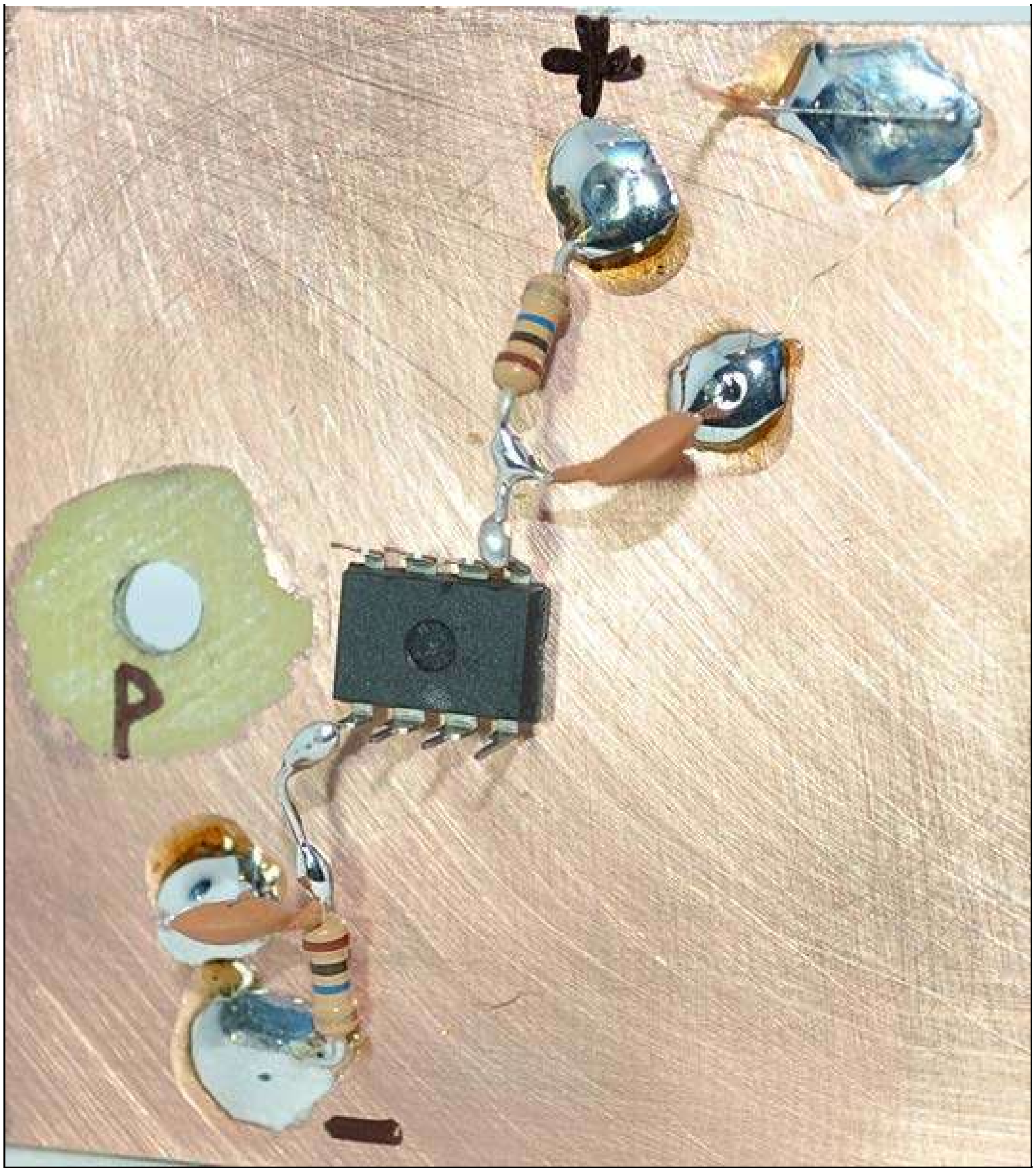

Above — a method to anchor op-amps using a split (negative and positive) power supply. Pins 4 and 8 are soldered to the copper board via a 10 megohm resistor. A 0.1 uF ceramic bypass capacitor is also connected to these pins. The resistor leads were left a little long to allow easy connection of the power supply wires. I write each pin's polarity on the board to avoid wiring mistakes.

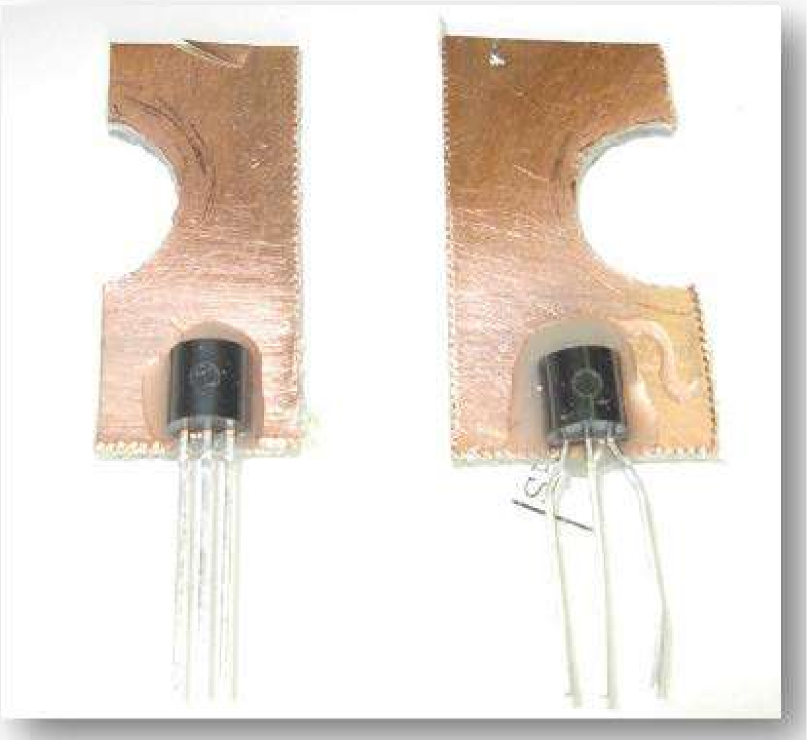

Shown above is another use for copper clad board; heat sinks. In this case, 2 scraps are epoxy glued to 2N3904 and 2N3906 transistors.

MG Chemicals®
ROSIN FLUX PEN
Cat. No. 835-P Net. Fill 10 ml
Non-Corrosive RA Type Flux. FLAMMABLE / INFLAMMABLE

A flux pen like this one from MG chemicals is a handy item for the QRP workshop. They are great for applying liquid flux to allow easy and precise soldering of SMT components. Also sometimes when adding components to ground in already built up circuit boards, it can be difficult to get your soldering iron down at a low angle for proper heat transfer. Some flux can help solder a part to the copper ground plane in these situations.

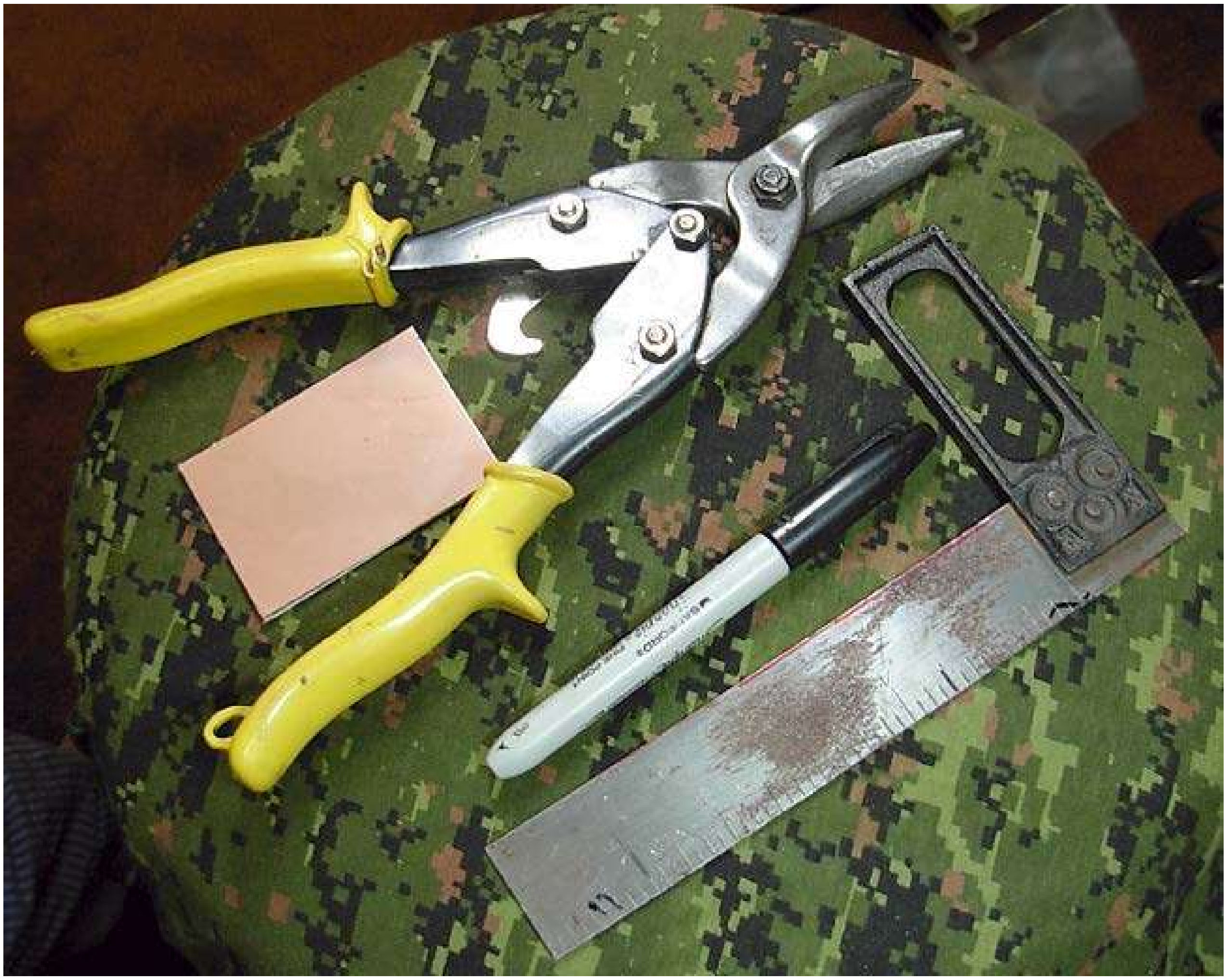

Shown above are the basic tools used to cut copper clad board. A felt pen marker, small square and a set of straight aviation shears. If you press one end of the copper board into the side of your bench and keep pressure on the handle of the aviation shears with your thigh, it is possible to make long, straight cuts. The board will flex and move out of the way as you cut. Your leg provides the force to advance the aviation shears.

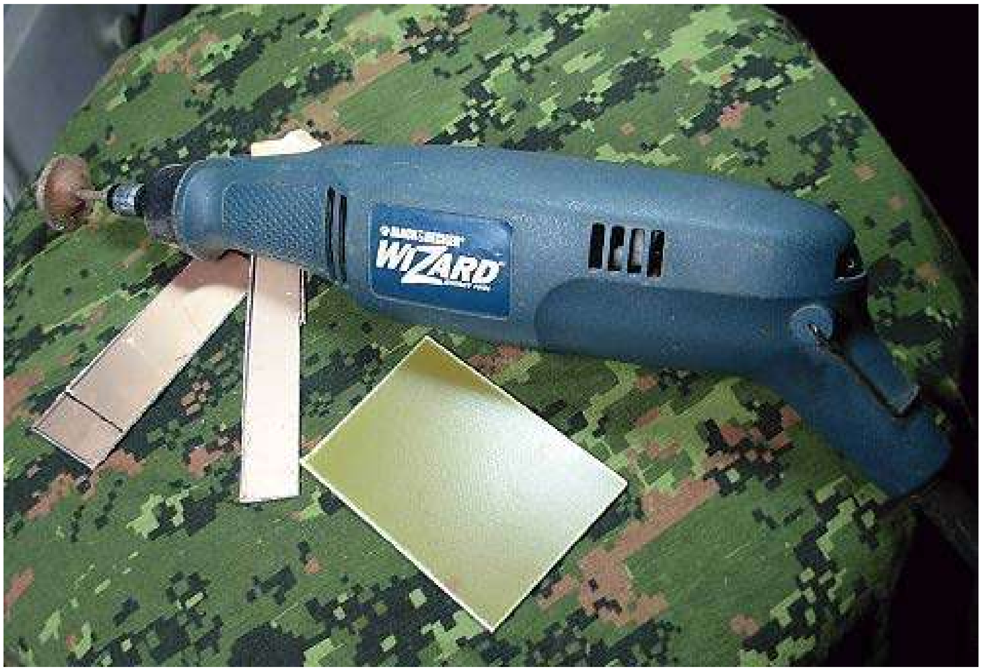

The motorized grinding/cutting tool used to carve out small pathways in copper clad board.

NET CONTENTS
6 ROLLS
BULL DOG
CAUTION:
GLOVES RECOMMENDED
WHEN USING THIS PRODUCT
SUPERFINE
STEEL WOOL
MANUFACTURED BY
THAMESVILLE METAL PRODUCTS LIMITED
THAMESVILLE - ONTARIO

Steel wool used to clean up copper clad board before construction. A box like this will last for years.

OLFA SILVER

For 15 years, I've used this 9 mm cutter to scrape the enamel insulation off the magnet wire wound on toroidal inductors.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

VE7GC Wee Willy 75 Meter DSB Transceiver Project

Introduction

I meet a lot of interesting amateur radio enthusiasts online and on the air. Among them is Dick Pattinson, VE7GC who was first licensed in 1934. When Dick isn't sailing he builds and operates his homebrew QRP gear to hams around the west coast of Canada and the U.S.

Presented is a double sideband transceiver that Dick calls "Wee Willy". This rig has a low parts count and is easily built using non-etched PC board techniques. Dick even makes his own radio case using copper clad PC board for the front and back panels and cardboard for the rest of the case. The set itself is in a case 1 1/4 by 2 1/4 by 5 1/2 inches. There is a separate container which holds a 6 volt rechargeable battery and speaker. The speaker/battery case is about 2 1/2 inches cubed and is not shown.

The circuits are built in three sections on the circuit board, namely TX, RX, and VFO. An electret condenser microphone is on front panel along with T/R switch, volume and frequency adjust. The back panel has the antenna jack, power input cord and the speaker jack. The battery pack has a 100 uF capacitor to across the power leads for additional filtering and is shown in the VFO schematic.

The text that follows is clipboard pastings from email that Dick sent me with some additional comments and expansions by me. Dick's project exemplifies practicality and innovation and with that is a major contribution for the QRPHB site.

Construction Methods

Dick's electronic construction method is quite fascinating and represents yet another derivation of ugly construction. The electronic wiring is done on single sided PC board, copper side up. Small holes are drilled through the PC board material to allow component leads to pass through them. Then the holes carrying active leads are chamfered (countersunk) with a larger drill bit which is not run all the way through the PC board. This leaves an ground-insulated side to the hole and prevents a component lead short circuit. The copper being topside allows both convenient and short component grounding. The Wee Willy parts layout is extremely neat and compact. Dick, presumably through practice has great skill with this technique and I plan to try it in the future.

The project case is constructed from 1/16th inch cardboard which is cut and bent to fit the electronic PC board. Once cut, the outer surface and edges at the front are covered with tissue paper or Kleenex (tm) type tissues soaked in white glue. The applied tissue paper and glue is allowed to dry and then additional coats are added to build up a body. Alternately, the cardboard case can be coated with lots of glue and the covering material imbedded in the glue. The air bubbles are pressed out and extra glue is added where necessary. When enough material has been added to cover up and strengthen the case joints and the glue is perfectly dry, the case is painted with Rust Coat Enamel available at hardware stores. The end result is a glossy, durable finish which looks very sharp.

Transmitter

This transmitter uses an electret condenser microphone (Dick used an Archer 270-90). The mic is built right into the front panel of the chassis and this of course guarantees short mic leads. A 741 op amp is used as a speech amplifier which in turn drives the balanced modulator a Signetics NE602 doubly balanced mixer. The input and output impedance of the NE602 mixer is around 1500 ohms.

To adjust the transmitter, set the bias control on the VN10 stage to ground and tweak L1 to resonance using an RF probe or scope on the VN10 input. The input signal must be audio, spoken into the front panel microphone to get the DSB. Once L1 is tuned, connect a 50 ohm load to the antenna with some sort of RF indicator (such as a RF power meter) and advance the bias control to give a watt or so output. Then speaking into the microphone should result in a DSB signal suitable for communicating on QRP! No audio input should result in no RF output. The supplied voltage should be kept at 6 volts, remembering that NE602's cannot stand voltage greater than 9 volts. With suitable voltage control such as a 6.8 volt zener diode on these chips, one could use higher input voltage with a corresponding RF output.

There is another way of setting the bias on the VN10. After aligning L1, with a ammeter in the six volt supply line, advance the bias control until the input current increases about 10 mA (with no modulation). If you do not have an FT37-77 ferrite core, substitute 10 bifilar turns on a FT37-43 ferrite core for the T1 transformer

.

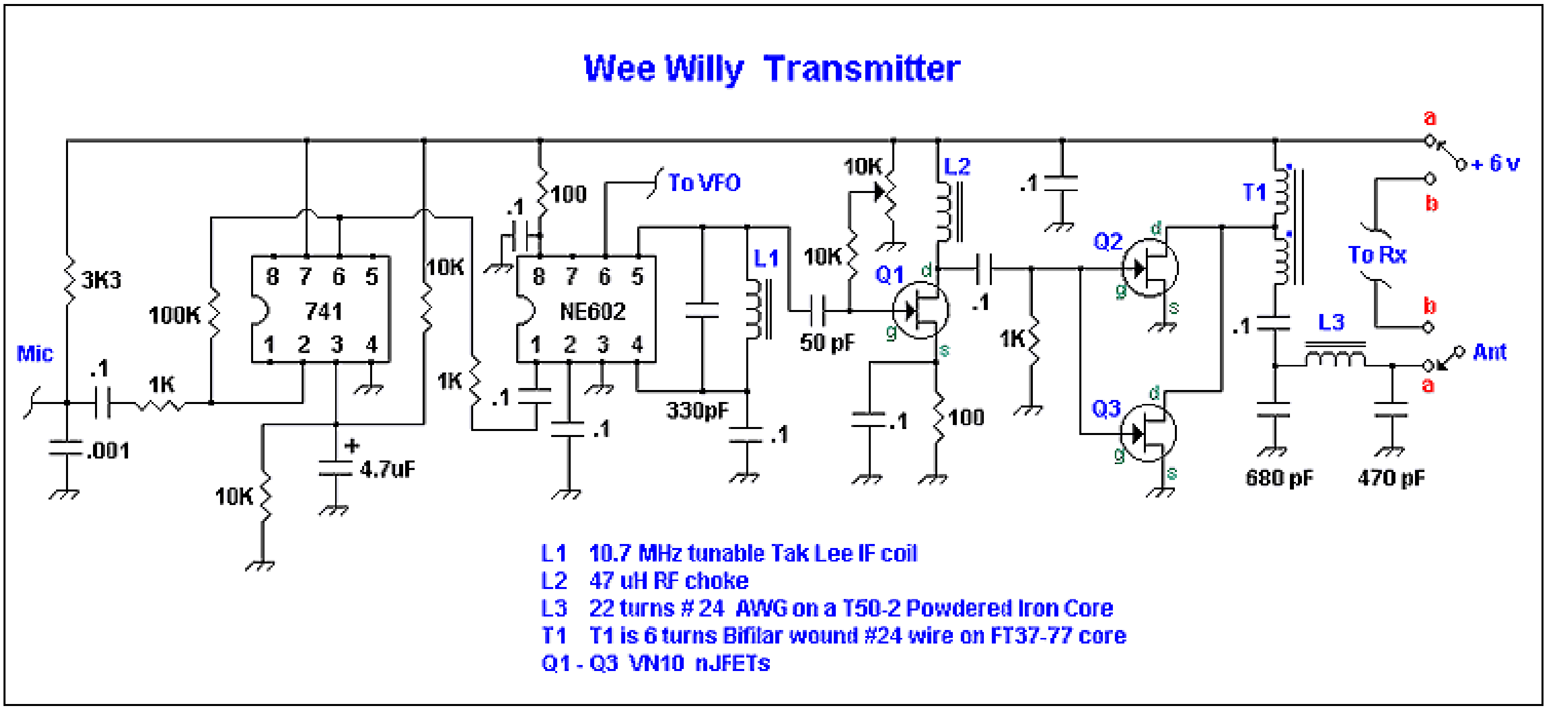
Wee Willy Transmitter

L1 10.7 MHz tunable Tak Lee IF coil
L2 47 uH RF choke
L3 22 turns # 24 AWG on a T50-2 Powdered Iron Core
T1 T1 is 6 turns Bifilar wound #24 wire on FT37-77 core
Q1 - Q3 VN10 nJFETs

Receiver

The receiver is a direct conversion type with a manual RF gain control in the form of a 5K potentiometer. Listening to a weak signal on the desired frequency the RF stage and the mixer core (T1 and T2) adjustments are made until you hear the loudest possible signal, keeping the input test signal as low as possible. When the receiver is connected to a doublet antenna there is no lack of incoming signal, which can be controlled by the front panel RF gain control.

The antenna and 6 volt supply is switched manually from TX to RX mode and back by a front panel mounted switch. If you can not find Tak Lee green 10.7 MHz IF coils, probably any other brand of 10.7 MHz slug tuned IF transformer would work. The Mouser catalog number 421F123 would work well and in another 80 meter project I used it with a 470 pF capacitor instead of the 330 pF cap shown. I would start with Dick's 330 pf cap and if it will not tune to resonance sharply, slightly increase the cap value up to see if a bit more capacity is required to resonate it on the desired 75 Meter frequency. Note that the secondary coil on the L1 transformer in the transmitter schematic is unused. If your 10.7 MHz IF coil has a built in capacitor at the base, remove it.

During receive, the standby drain current at 6.0 volts was 24 mA and on loud signals it rose to 100 mA. If this is too much, probably the easiest thing to do would be to put in a series resistor from positive to the LM386 to limit the drain current. To get output on the speaker it is a matter of how loud you want it for the drain you draw. If earphone only reception is okay, then the drain could be reduced considerably.

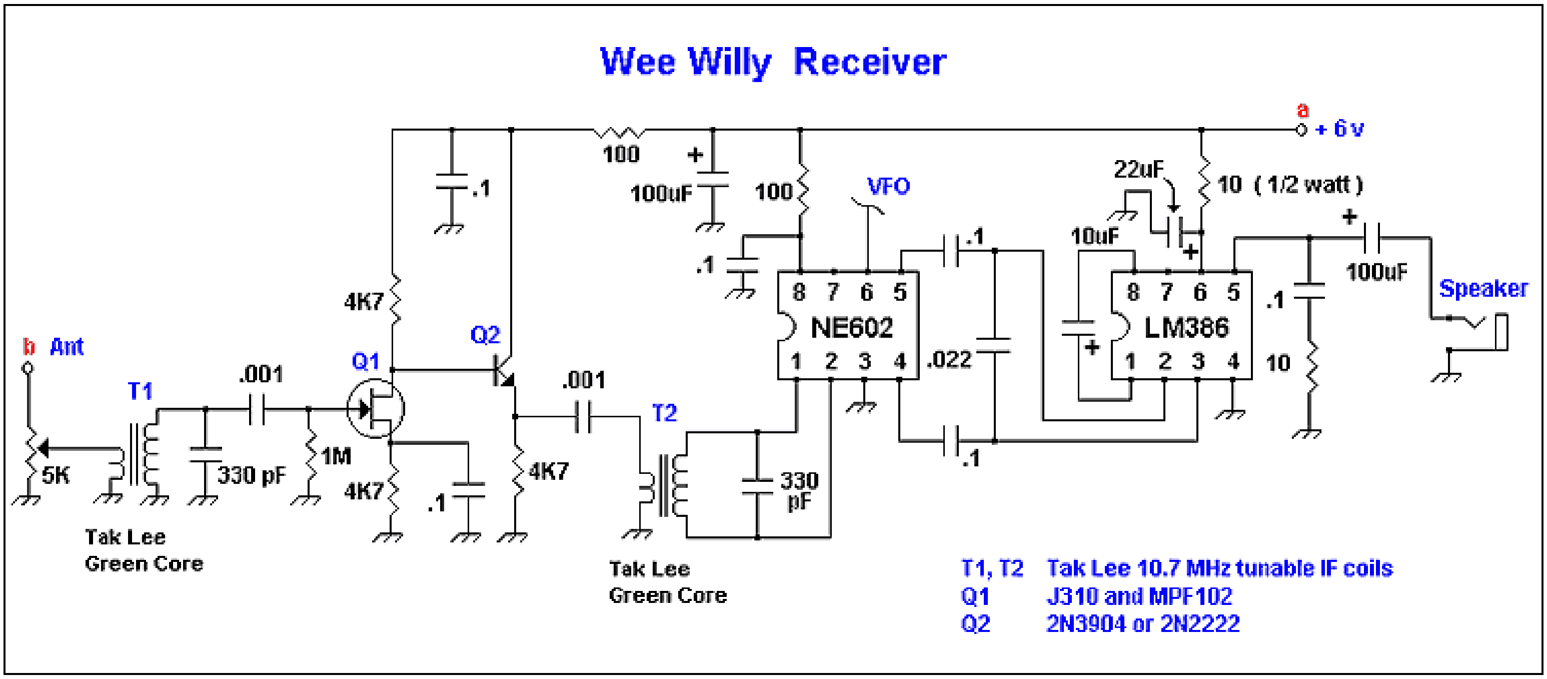
Wee Willy Receiver
a
+ 6 v
100
+
100uF
.1
100
VFO
.1
.1
22uF
10 (1/2 watt)
10uF
+
+
100uF
Speaker
8 7 6 5
NE602
1 2 3 4
.022
8 7 6 5
LM386
1 2 3 4
.1
10
.1
b Ant
4K7
Q2
Q1
T1
.001
5K
330 pF
1M
4K7
.1
4K7
.001
T2
330 pF
Tak Lee
Green Core
Tak Lee
Green Core
T1, T2 Tak Lee 10.7 MHz tunable IF coils
Q1 J310 and MPF102
Q2 2N3904 or 2N2222

VFO

Dick's diagram indicate that this VFO was based upon a design presented in SPRAT for summer 1995. The main inductor L1 is wound with #32 AWG wire on a 1/4 inch slug-tuned coil former. This coil would have an XL somewhere between 250 - 310 ohms, so if you cannot find a coil former as described , you could easily wind one on a powdered iron toroid and make a portion of the C1 capacity variable for adjustment. A suggested alternate inductor is 53 turns of #26 AWG on a T68-6 core powdered iron core.

Dick suggests checking an old television to find suitable coil formers such as the one he used. It would probably be best to distribute the 120 pF C1 capacity among 3-4 capacitors to enhance stability. These caps should be NP0 ceramic for best results with frequency stability.

Dick's oscillator uses the slug tuned core to put VFO frequency close in frequency to where you want to operate and the variable resistor tuner on the front panel allows adjustment around the incoming signal to get the correct pitch. The desired band-edge is easily set by adjusting the slug while listening to the VFO frequency as audio on another receiver that has a frequency readout or directly with a frequency counter.

The L2 150 uH RF choke can be a simple epoxy unit which resembles a resistor. The D1 variable capacitance diode is a BB104 which has ~ 35 pF capacitance on each side. These are available at Dan's Small Parts and Kits whose URL is in the Links section of the site info web page. Experimentation with other tuning diodes could produce a practical alternative to the specified D1 part.

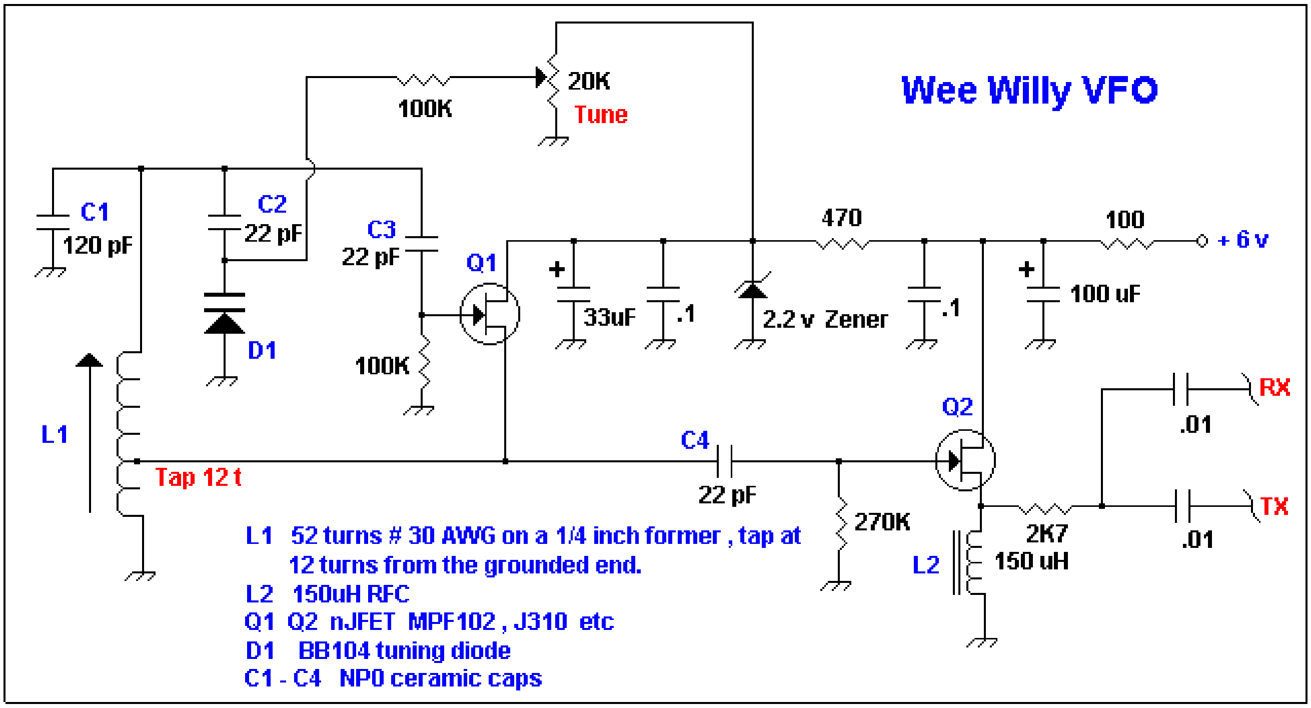
Wee Willy VFO
100K
20K
Tune
C1
120 pF
C2
22 pF
C3
22 pF
D1
Q1
100K
+
33uF
.1
2.2 v Zener
470
.1
100
+ 6 v
+
100 uF
L1
Tap 12 t
C4
22 pF
270K
Q2
RX
.01
TX
.01
2K7
150 uH
L2
L1 52 turns # 30 AWG on a 1/4 inch former , tap at 12 turns from the grounded end.
L2 150uH RFC
Q1 Q2 nJFET MPF102 , J310 etc
D1 BB104 tuning diode
C1 - C4 NP0 ceramic caps

Operating Wee Willy

Dick sent me Willy in the mail and the first available moment I fired the little rig up and spanned the VFO which tuned from ~ 3721 to 3738 KHz. I then proceeded to tune 3729, the frequency of BCEN, our SSB provincial public service net and on my first break was able to check in with one call sign repeat to the net control station. The band conditions were noisy and most signals were S8 or lower, however Wee Willy's 1.5 watts P.E.P. were able to check me in with my folded Marconi antenna. I later changed the L1 slug and worked some stations higher up the band. My audio reports were favorible and no one knew I was running DSB. This is a fun radio and it looks cute to boot!

When transmitting, I had to be careful to keep my hand away from the VFO compartment to prevent pulling the VFO frequency with the capacitance change from my hand. The VFO has reasonable long term frequency stability and copying CW stations with the receiver was possible without frequent tuning readjustments.

The following are some digital photos of Wee Willy taken by VE7ZAC.

Move your mouse over the images for a larger version.

From left to right the electret mic, TX/RX switch, RF gain control and VFO tuning control. The number 375 is Dick's project identification number. He has given me many schematics of his projects and each has a unique number and case color. For the non-Canadians, the large coin on top of Willy is our "Loonie" a 1 dollar coin. The large 2 color coin in front of the rig is our "Toonie" a 2 dollar coin.

The hole on the left of this rear panel shot is for adjusting the L1 slug for the VFO. From wherever on the band the VFO is set a front panel tuning range of about 17 KHz was possible. Also shown are the speaker jack , BNC antenna connection and DC power leads.

Wee Willy with the cover off.

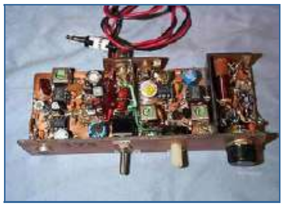

These top view shots clearly shows the 3 distinct compartments. From left to right the TX , RX and VFO sections. The VFO has a PC board seperator for shielding. Dick's construction method is well illustrated with this photo. The IC's are in sockets. The 3 VN10s have a small tab on the top. There are no heat sinks on the VN10 finals and they do require any for a 6 volt supply voltage. The rear panel jacks and VFO inductor are all mounted in PC board material.

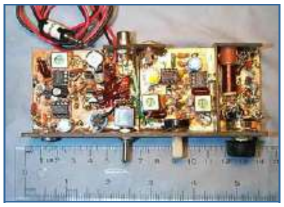

Here is a bottom shot showing the connecting wiring through the chamfered PC board holes. Many thanks to Dick Pattinson, VE7GC for allowing me to present one of his projects.

Wee Wee Willy by NM8T

NM8T Builds Wee Wee Willy

Steve White, NM8T emailed me a few weeks ago that he was building Wee Willy and was not getting the expected output power. I forwarded his email to Dick and they problem solved the issue. It turns out that the V10KM's were the culprit.

Here are their final two emails:

Steve: I think you have found a solution to your problem. I did not know that VN10's are different. I bought mine in 1994 from a radio parts supplier and they are marked VN10KM F324A1. Have fun with your new rig and let us know how you are doing with it. 73, Dick

Dick, I have completed my little rig and have had a blast with it so far. My first CQ yielded me 3 contacts in Pittsburg, Pa. My friend across town gave me a 20 over S9 report. The three hams in Pa. were astounded that I had only 1.5 watt PEP. I had to use 12 volts and used small 5 volt regulators to feed the NE602's. This works very well. I even used a little homebrew tuner with it and got my swr flat. I am going to take the little rig to Dayton Hamvention next year. I go every year and take a little project each year to show everyone. I hope it is ok to call my little rig Wee Wee Willy, since it is a little smaller than your version. I want to thank you for your help in getting this rig completed and also thank Todd for his help in getting me in contact with you. Your little rig is very neat and right up my alley for a QRP rig. I look forward to working you someday on QRP. I'm 47yrs. old on the 24th of this month. I have a whole room full of QRP rigs such as HW-7, HW-8, HW-9, Argonaut 509, 515, Powermite PM-3a. Have built a Cascade, LCK, Spider, MXM, Keyers, Tuners, and all kind of other projects. I love QRP and hope to build many more soon. The Wee Willy is one of my favorite rigs. Here are some pictures for you and Todd. Thanks and Hpe CU Agn. 73's NM8T (No more 8 Tracks) Steve White - Fayetteville, WV

75m DSB
Tune - Wee Wee WILLY - Gain

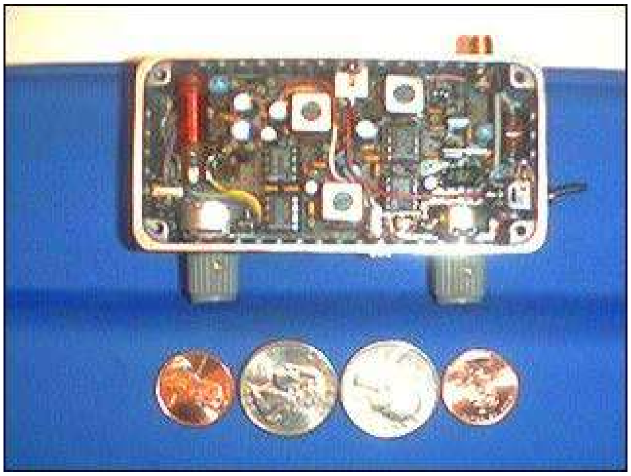

My special thanks to Steve White and Dick Pattinson.

Countless builders have constructed the Wee Willy on several bands. Here are 2 more examples. Paul, KE7HR Link1 and Link 2. The finals even made it into the the BITX17 transceiver. Allan Yates published some good work on the final amp.

I've also tweaked Dick's design, but decided to leave his little rig and this web page alone out of respect to him. It's always easier to improve rather than actually *design* something. I'll never better the experiences I had operating and documenting his transceiver.

Update February 15, 2011

I lost touch with Dick and decided to call him on Feb 10, 2011. He's 95 years old and no longer building. Sadly, his wife Christina died 2 years ago. She was a beautiful person and a gifted painter. Dick remains astonished that Wee Willy - Project 375 QRP Radiophone has been built by at least 50 people (that I know about) world wide, the subject of blogs and the basis/inspiration for numerous radio projects. I met Dick on 80 meters and my wife and I once visited his residence and spent 2 glorious, days on Saltspring Island with he and his wife.

Dick sent me the radio and documentation on Aug 13, 1998. He was a hardcore builder and a professional electronics technician who, for a time, flew his float plane to service equipment in the Gulf Islands of British Columbia. Dick keeps many binders full of his carefully numbered and well documented projects. He gave me a few sheets for my reference library as a gift. Dick's work inspired me early on and I cherish the time I had Wee Willy in my radio shack - blown away by DSB, QRP and such a novel little radio.

Dick asked me to send his greetings to all QRP Homebuilders everywhere.

More Updates

Aug 25, 2011

Hi. Been viewing your web site for a while, lot of great information for builders. I thought i would let you know that I built the "Wee Willy", it works great, so good that I decided to try building one for 20m. I built the 20m version with a 14.318 crystal instead of a VFO ideal for my local net, also changed the PA circuit using IRF510 giving me 2.5W output, the transistor can run 5W, but it does get rather hot, I must say that the quality of the transmitted audio is amazing, everyone comments how good it is and some do not believe it is home-brew!!! I have put a few CQ calls out with a simple dipole and made contacts into Belgium, Netherlands, Finland and Saudi Arabia. Many thanks again for putting the Wee Willy on your website, it has opened my eyes to how well you get get out with QRP power.

Regards Wayne. M0WAY. Click for Wayne's fabulous Wee Willy bread board.

Sept 19, 2011

Wayne, M0WAY wrote again,

I built another Wee Willy and put it into a box, but with a few extra bits:

A new VFO was added that runs 14.190 – 14.350 MHz (an 11 MHz crystal oscillator + a 3.2 MHz VFO + a band-pass filter) I got the power up to 4W, however the audio was a little over powered from the extra RF, so I had to reduce the audio gain and added an antenna switching circuit, after I burnt out my receive circuit with the 4W output.

Here's Wayne's PA (from the NE602 on ** updated July 12, 2012 **) Click for a photo of his latest incarnation. Thanks again!

Oct 19, 2011

Mark, WA4JAT built a "0 cost" Wee Willy from 100% junk box parts including a CB radio transistor for the final. Click and click. Mark tunes the VFO using an air variable capacitor mounted in the former Heath Kit signal tracer project chassis. The cable running out of the front is the microphone coax with an electret mic at the end. Employing large size VFO, Rx and Tx boards facilitate easy future modifications. Awesome work Mark!

RF — Test and Measurement

JFET BIASING TUTORIAL BY W7ZOI

This tutorial is copyright © 2000-2001 by Wes Hayward, W7ZOI.

JFET Biasing and an Amplifier Design W7ZOI, 30Dec00.

Junction Field Effect Transistors (JFETs) are generally no more difficult to bias than a bipolar transistor. In many ways they behave essentially like a classic triode vacuum tube. The current flowing between FET drain and source is controlled by the gate-to-source voltage, V_{GS}. (This is the voltage you would measure with a voltmeter connected between the gate and the source.)

The Fig. 1 schematic defines the parameters we use to describe the DC behavior of a JFET. The 100 Ohm resistor is added to preserve UHF stability during measurements. The gate voltage, V, is then varied while measuring drain current. The results are then plotted, as shown in Fig. 1. The FET is characterized by the simple quadratic equation below the schematic with the two FET related parameters of I_{DSS} and V_P. I_{DSS} is the drain current when the gate is short circuited to the source. The pinchoff voltage, V_P, is V_{GS} when the drain current is essentially zero. The example in Fig. 1 has I_{DSS}=20 mA and V_P=-4 volts. Source and drain current are identical in the JFET.

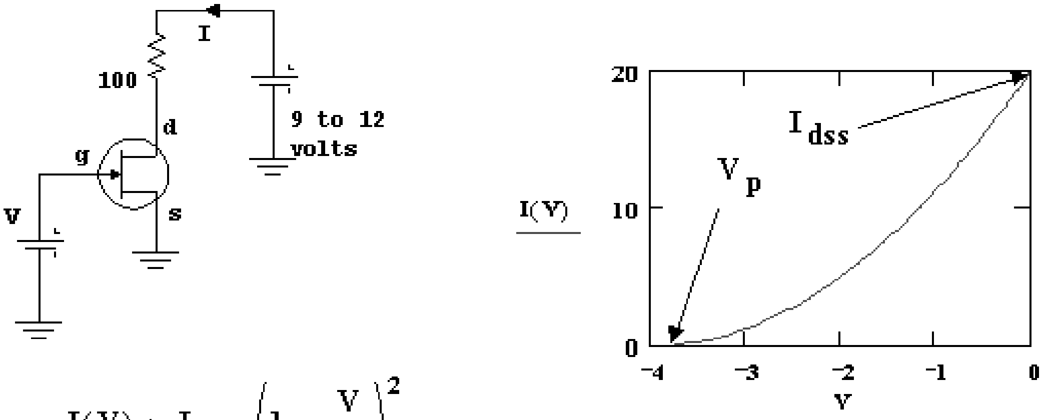

$$I(V) := I_{dss} \cdot \left(1 - \frac{V}{V_p}\right)^2$$

$I_{dss} := 20$ mA $V_p := -4$ volts

Figure 1

Basic behavior of an N-Channel depletion mode JFET.
The numbers shown basically illustrate the ideas.

The experimenter wishing to use a JFET in a circuit will most often use the method called "self biasing" where a resistor is placed in the source lead. The gate is grounded. The FET current also flows in the resistor. The current generates a voltage drop that elevates the source voltage above that of the gate, which is the same as biasing the gate negative with respect to the source. This establishes a drain current less than I_{DSS}.

Drain voltage is not critical so long as it is a few volts greater than the source.

The resistor value for a desired bias is calculated with Equation 1 if I_{DSS} and V_P are known. But these parameters are rarely well specified. For example, the popular J310 is characterized only by ranges: I_{DSS} is 24 to 60 mA and V_P is -2 to -6.5 volts.

$$R_s := \frac{V_p}{I_D} \cdot \left(\sqrt{\frac{I_D}{I_{dss}}} - 1\right) \qquad \text{Eq.1}$$

I_{DSS} and V_P are easily measured for a specific FET with the circuit shown in Fig. 2. The circuit is built with a variety of source resistor values and the source voltage is then measured for each with a high impedance voltmeter. The pinchoff voltage will then be the negative of the source voltage when R_S is a large value such as 100K. Grounding the source through a current meter generates a direct reading of I_{DSS}. This brief measurement will provide enough data to allow biasing.

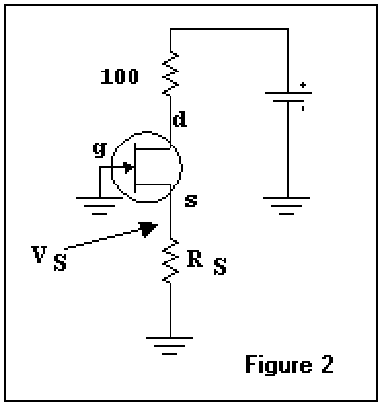

Figure 2

Circuit used to determine FET DC parameters.

This experiment was performed with a junk box 2N5454, producing the parameters I_{DSS}=11 mA and V_P=-2.8 volts. The current equation from Fig. 1 is plotted in Fig. 3 with these parameters. Additional points were measured producing points close to the curve.

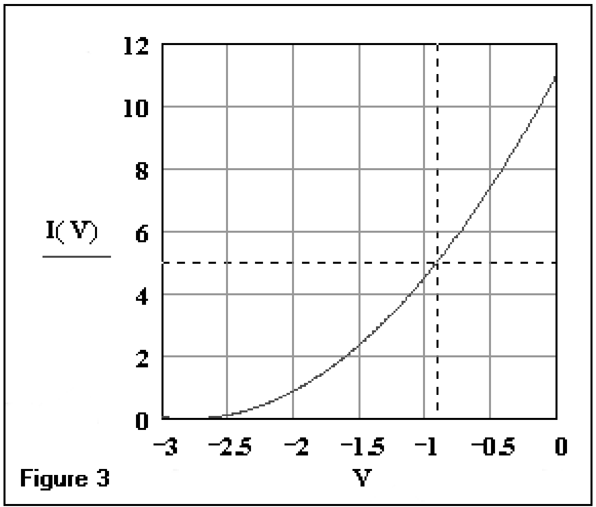

Figure 3

Data and smooth curve for a 2N5454 we measured.
This FET has a pinch off of -2.8 volts and a IDSS of 11 mA.

This exercise SHOULD be done with virtually every FET that the experimenter uses. It is not enough to obtain data for a given FET type, although if a group of parts are available from one manufacturing batch, moderate uniformity might be expected.

We mentioned the J310 earlier with an I_{DSS} spread from 24 to 60 mA. That 2.5:1 span range is one of the better of commonly available parts. The historically popular MPF-102 has a 10:1 spread in I_{DSS} (2 to 20 mA) with V_P extending from -0.5 to -7.5 volts! Clearly, this part should be used only with extreme care. It's a good RF part, but one MPF-102 may not resemble the next.

Assume the FET we summarized in Fig. 3 was to be used in an amplifier and that we wished to bias the FET for 5 mA. The 5 mA point can be picked on the graph and matches V_{GS} of approximately -0.9 volt. 0.9 volt drop from 5 mA matches a resistor of 180 Ohms. Application of Eq. 1 produces 182 Ohms. Crosshairs indicating the 5 mA point are included in Fig. 3.

Once a FET has been biased, it can be used in a small-signal amplifier. The significant gain parameter is the transconductance, given by Eq. 2. Transconductance is the drain signal current flowing for each volt of signal applied to the gate. Note that g_M depends upon the bias current. The g_M value given in data sheets corresponds to operation at I_{dss}.

$$g_m = 2 \cdot \frac{I_{dss}}{V_p} \cdot \left(1 - \frac{V_{GS}}{V_p}\right) \qquad \text{Eq.2}$$

$$I_{dss} := .011 \quad V_p := -2.8 \quad V_{GS} := -0.9$$

$$g_m := 2 \cdot \frac{I_{dss}}{V_p} \cdot \left(1 - \frac{V_{GS}}{V_p}\right) \qquad g_m = -5.33163 \cdot 10^{-3}$$

Figure 4

Data used to produce transconductance for the FET used in our sample amplifier.

Evaluation of g_M from Eq. 2 for our 5 mA bias produces the data in Fig. 4.

Consider the amplifier shown in Fig. 5. We will use this as an output buffer for an oscillator. Assume that we have tapped a signal from the oscillator tank where a voltage of 1 volt peak-to-peak is available. When this is applied to the FET gate, a current of 5.33 mA pk-pk will flow. The transformer uses a 15 turn drain winding and a 3 turn output link on a FT37-43 ferrite toroid. This turns ratio transforms the 50 Ohm load to appear as 1250 Ohms in the drain line. A current of 5.3 mA pk-pk flowing here produces a voltage of 6.25 V pk-pk. The no-signal drain voltage will be about 11 volts, so positive and negative 3 volt excursions will not cause voltage limiting. The current will vary from the 5 mA DC value down to nearly 3 mA (still well above zero) and up to nearly 8 mA, still below I_{dss}. The signal at the 50 Ohm load is reduced by the transformer to 1.33 V pk-pk. This is about +6 dBm, enough power to drive a diode ring mixer. This simple circuit is a low noise, efficient, and very useful amplifier.

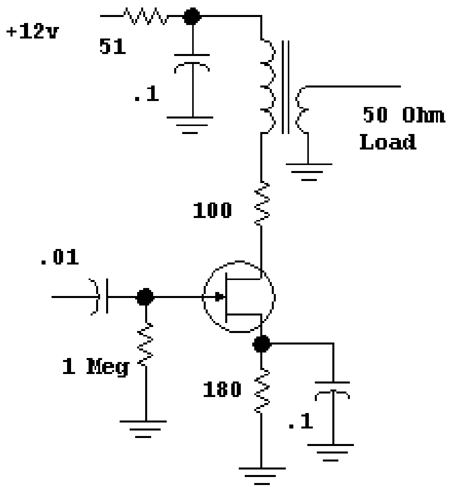

Figure 5

Common source amplifier biased for 5 mA drain current.

Figure 6 shows a FET amplifier that uses "long tail" biasing. A negative power supply is now used, forcing the biasing resistor to behave like a constant current source. This method can accommodate a wide range of FETs of a given type if the bias is set up for $I_d<I_{dss}$-min.

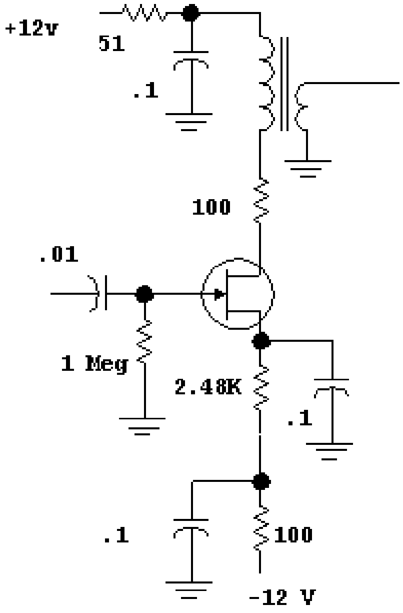
+12v
51
.1
100
.01
1 Meg
2.48K
.1
.1
100
-12 V

Figure 6

Amplifier with "long tail" biasing. This amplifier is biased for 5 mA and is identical in performance to that of Figure 5, but does not require the careful device characterization.

RF — Test and Measurement

EMRFD Review

Experimental Methods in RF Design

$49.95

EXPERIMENTAL METHODS in RF DESIGN

Wes Hayward, W7ZOI
Rick Campbell, KK7B
Bob Larkin, W7PUA

Published by ARRL

CD-ROM INCLUDED

First published by the ARRL in 2003, EMRFD serves as the main QRP/SWL Homebuilder site reference.

Written by Wes Hayward, W7ZOI, Rick Campbell, KK7B and Bob Larkin, W7PUA, EMRFD is the follow-on to the 1977 ARRL publication, Solid State Design For The Radio Amateur (SSD).

A treasure trove of narrative and tabled information, schematics, references and photographs adorn this lengthy (512 page) book that comes with a data CD filled with software and key reference papers. With the included CD, EMRFD is about twice the size of the original SSD.

Wes and crew emphasize fundamental radio equipment design and bench testing rather than providing the usual catalog of circuits to just casually copy and not carefully examine.

Since we amateur builders own varied skills, abilities and test equipment, some sections may intimidate readers, while others may inspire and drive your bench practices to a higher level of competence and joy. The heart of RF design lies in measurement and reason: EMRFD emphasizes this and in doing so alienates some readers.

For some, homebrew radio electronics and kit building are synonymous — plenty of kits are sold to builders who chose to build someone else's design, rather than capture their own ideas on a scrap of copper board, or at least, to modify their equipment to suit their needs. EMRFD may appeal to builders who enjoy learning about RF *design*, or want more innovative and creative bench experiences — 1 stage at a time.

Although stressing that *build and measure* = a proven way to improve in this hobby, Wes and team share other pearls including bread boarding techniques, parts lay out, hot parts and pitfalls to avoid.

The first chapter is simply called Getting Started. On page 8, Wes shares his first simple receiver design — this page starts your breadboarding in haste! EMRFD features information and designs for all levels of experimenters, although basic electronics knowledge and some experience are needed to get the most out of its content.

Replete with sidebar examples about measuring or calculating data for common circuits, EMRFD also offers general purpose stages including a universal monoband superhet receiver front end or AF chain instead of less versatile, single- application circuits. The design information feels vibrant and flexible and some of the material is original, or presented in a way that adds to the existing amateur radio knowledge base. For example, new front-end mute circuits, or the cascode JFET mixer and RF amplifier.

Wes also shares some new ideas for RF and IF amplifiers ranging from simple to state-of-the-art. This book has something for everyone — I prefer to describe EMRFD's influence on our hobby in simple action verbs such as: improve, innovate, inspire, explore and transform.

Software:

The Microsoft Windows programs run on everything up to and including Windows 7. While a few programs are new, others represent updates of the historic W7ZOI Ladpac applications. Upon mastery of the Ladpac program suite, you'll generate the needed stages to boost *your* designs.

For example you can design complex double or triple tuned band-pass filters for your projects without any math — the software gives you the series capacitor values so you need not worry about putting links on your inductors to match input or output impedances.

EMRFD = the **must have** reference book for your homebrew workshop.

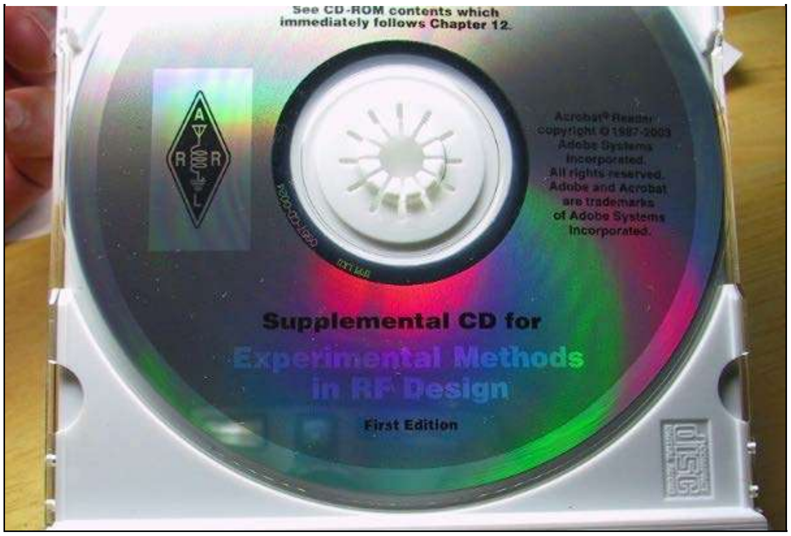
See CD-ROM contents which immediately follows Chapter 12.
ARRL
Acrobat® Reader copyright © 1987-2003 Adobe Systems Incorporated. All rights reserved. Adobe and Acrobat are trademarks of Adobe Systems Incorporated.
Supplemental CD for
Experimental Methods in RF Design
First Edition

ARRL EMRFD web link

Link to the W7ZOI errata page for EMRFD

Tapped Capacitor Impedance Transformation in LC Bandpass Filters

copyright © Wes Hayward, W7ZOI, April 30, 2003.

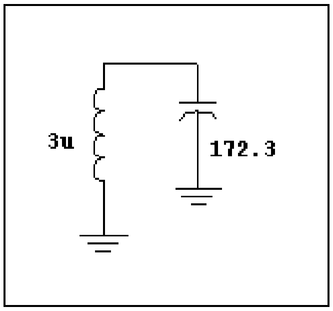

We often use a pair of capacitors to match impedances at the termination ends of LC bandpass filters. The circuit consists of a shunt capacitor at the termination followed by a series capacitor connecting to the high Z end of a parallel tuned circuit. Some readers have asked about how the capacitors are picked. Although there is considerable flexibility in some of the choices, it is not empirical as some have guessed. A simplified double tuned circuit design sequence is presented in the sidebar on page 3.14 of EMRFD and this is the beginning of the analysis used in the programs DTC.exe and TTC.exe contained on the EMRFD CD. I'll not go into too much detail here, for it's in EMRFD, beginning on page 3.8, and in chapter 3 of IRFD. Here is what happens in the programs:

Bandpass filter design begins with an almost arbitrary choice of inductor. We pick 3 uH for a 7 MHz double tuned bandpass filter that we will use to illustrate the ideas. This L resonates at 7 MHz with 172.3 pF.

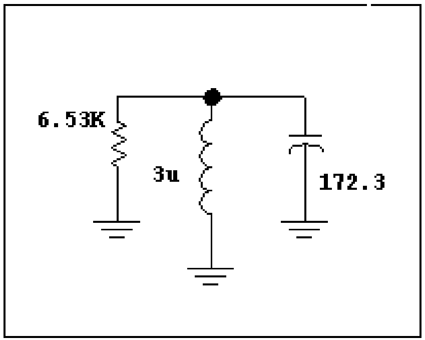

A bandwidth and center frequency are picked for the filter. This establishes a filter Q. The Q of an end section is then determined by the desired filter shape (Butterworth, etc)Let's say we want to do a 7 MHz center frequency Butterworth filter with BW=0.2 MHz. Filter Q is then 35=7/0.2. QE will then be 35x(root(2))= 49.5. The QE value in the sidebar (p3.14) includes the effect of finite inductor Q. Assume a lossless inductor for this example.

So what does this mean? It means that the end tuned circuit, when not coupled to the rest of the filter, needs to be set up to have a Q of 49.5. This is experimentally significant. (ref: QST, Dec, 1991). The reactance at 7 MHz of our 3 uH inductor is X=131.9 Ohms, so we need to load each end with a resistance R=QX=6.53K-Ohm.

What this means is that we would realize our double tuned bandpass filter with center frequency of 7 MHz and bandwidth of 200 kHz if we terminated a simple double tuned circuit in 6.5K at each end. The filter is shown below where we have used additional equations (EMRFD p 3.14) to calculate the coupling and tuning capacitors.

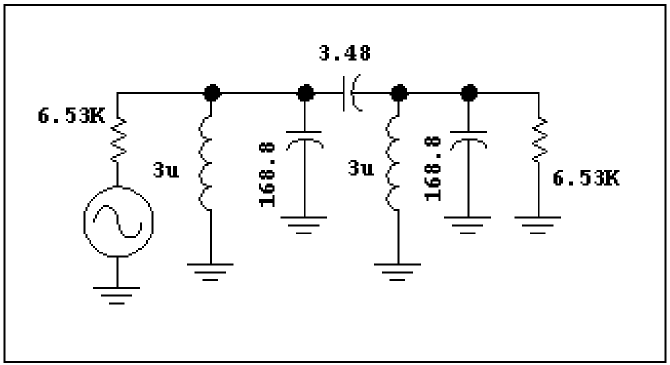

This is a useable filter design, for it will generate the desired shape and bandwidth. But, it is not very practical; it does not fit in our low impedance world.

The filter can be redesigned. One classic, but usually impractical solution is to scale the filter to lower impedance levels. For example, if we dropped L from 3 uH to 23 nH, we could directly load our filter with 50 Ohms and get the required end section Q. But this is not at all practical. First, it's difficult to build inductors with L this small and still have reasonable Q at 7 MHz. Second, the parasitic inductance of the rather large (high C) capacitors that we would attach to this inductor would begin to compare to 23 nH. A better re-design would use transformation circuits, schemes that will let us use a 50 Ohm termination (or whatever we need) and make it function as a 6.5K resistance when seen by the inductor.

One such scheme is a transformer. This could be realized with ferrite cores or with links inductors wound on the existing 3 microHenry parts. But link coupling with design precision is a challenge of it's own. Conventional two and three element transforming networks (L, pi, Tee) are also suitable.

The simplest transforming circuit uses a series capacitor. Let’s do some analysis to see how this works:

We have arbitrarily picked a 10 pF capacitor to illustrate the idea. At 7 MHz, the reactance of a 10 pF capacitor is 2274 Ohms. Hence, the complex impedance of the 50 Ohm resistance and the series capacitor is Z=50-j2274. The impedance transformation behavior of this circuit is studied by transforming the series impedance to a parallel admittance. Recall that Y=1/Z. So Y=1/(50-j2274). The result is Y=(G+jB)=9.67E-6 + j4.4E-4. Of special importance is the resistive real part of the admittance.

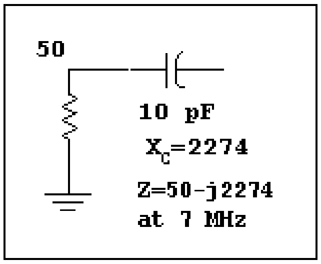

The 10 pF series capacitor in our example makes the 50 Ohm resistor look like a 103 K-Ohm resistor in parallel with a capacitor that is nearly 10 pF. The general case causes a R0 value resistor to look like a value of Rp with a series capacitor with reactance given by

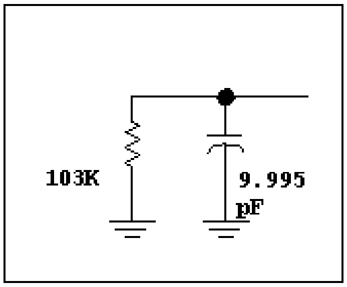

This equation is #3.1-5 from IRFD where it is derived.

$$X_C = \sqrt{R_p \cdot R_0 - R_0^2} \qquad \text{Eq. AA}$$

Our original example filter needed a parallel resistance, Rp, of 6.5K-Ohms, which is produced by a series capacitor with a reactance of 568 Ohms. This is a capacitor of 40 pF. Another version of our double tuned bandpass filter is then

Notice that the tuning capacitors, the elements across the inductor, have dropped as C is added at the ends.

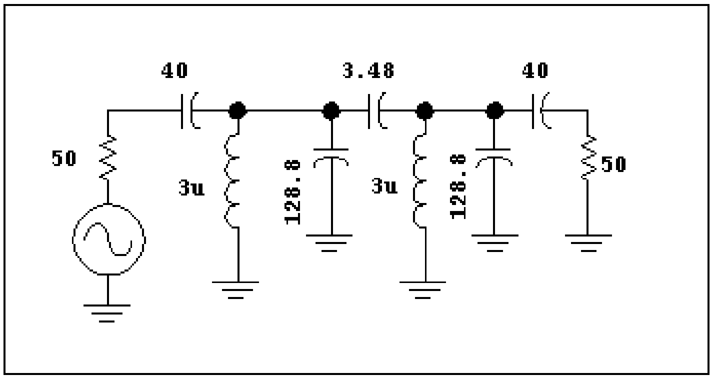

A capacitor in series with a resistance produces higher equivalent resistance. Consider a parallel combination. Here we pick a value of 200 pF as the parallel C and calculate the admittance. This is then converted to an impedance with Z=1/Y and the individual components are evaluated. This result is shown to the right.

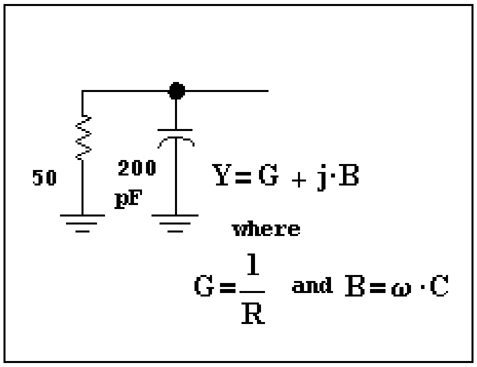

The parallel capacitor transforms the 50 Ohms to behave like a lower value.
In this case, we obtain about 42 Ohms in series with a 1236 pF capacitor.

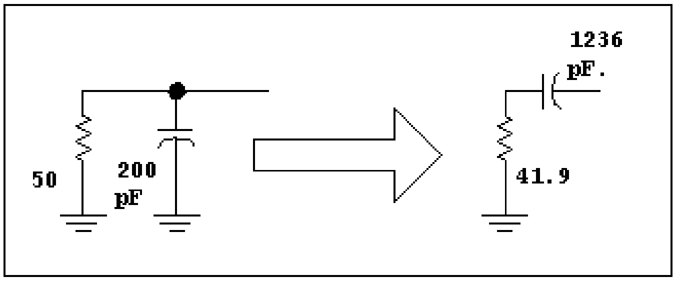

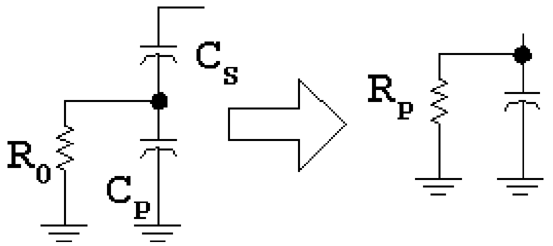

$$C_p = \sqrt{\frac{R_p \cdot C_s^2}{R_0} - \frac{1}{R_0^2 \cdot \omega^2}} - C_s$$

Eq. BB

A series capacitor transforms a termination to “look like” a higher resistance while a parallel capacitor “transforms to” a lower R. Clearly, the combination of the two can generate about any result we need, realizing that they will also produce reactance that must be absorbed into the existing tuned circuits. The mathematics (now symbolic and not just number manipulation) is messy, but not difficult. A resistive termination R0 (50 Ohms or whatever) is paralleled by a capacitor Cp. The admittance is calculated and is converted to an impedance. The impedance of a series capacitor, Cs, is then added. The result is converted back to an admittance. The resistive real part is extracted and inverted to yield Rp. This expression is then solved to yield a design equation:

This is the equation used in the programs. A related expression provides the equivalent capacitance, needed to calculate the tuning capacitor for each resonator.

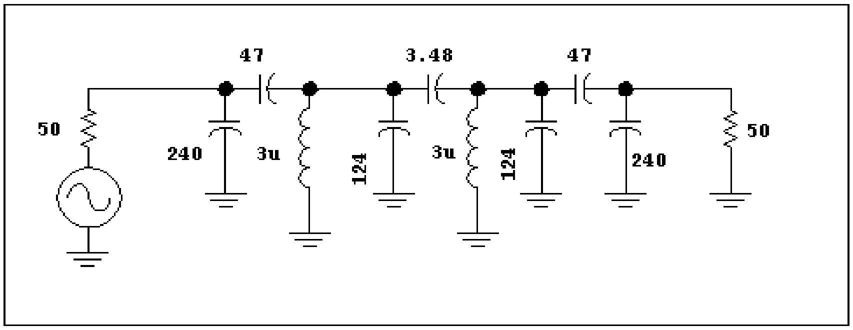

If we use the program with the center frequency and bandwidth presented earlier, we find that the minimum allowed series capacitance is 40 pF. If we then insert a value of 47 pF, we see that parallel 240 pF capacitors are needed to properly load the resonators. This variation is shown to the right. Circuits using the capacitor tap are practical, for they allow existing junk box parts to be used. The topology has little other utility, offering virtually the same response as a filter using only series capacitors for loading.

Useful link shared by Tim, KE7VYD on Feb 4, 2014. Click

QRP SWL HOMEBUILDER

Amateur and Short Wave Radio Electronics Experimenter's Web Site

More Active Antenna Experiments

Introduction

Many builders emailed me requesting a simple, broadband VPA (voltage probe antenna) design with more power gain than the common gate versions I have presented elsewhere on this web site. Connecting a whip antenna to a cascode JFET stage described by W7ZOI in Experimental Methods in RF Design is 1 method I considered.

I built the version shown in Figure 1 almost 2 years ago. This VPA. although more powerful, overloaded the front end of my test receiver with multiple RF signals. Clearly some tuning on the input was needed.

Experimental Broadband Cascode JFET VPA

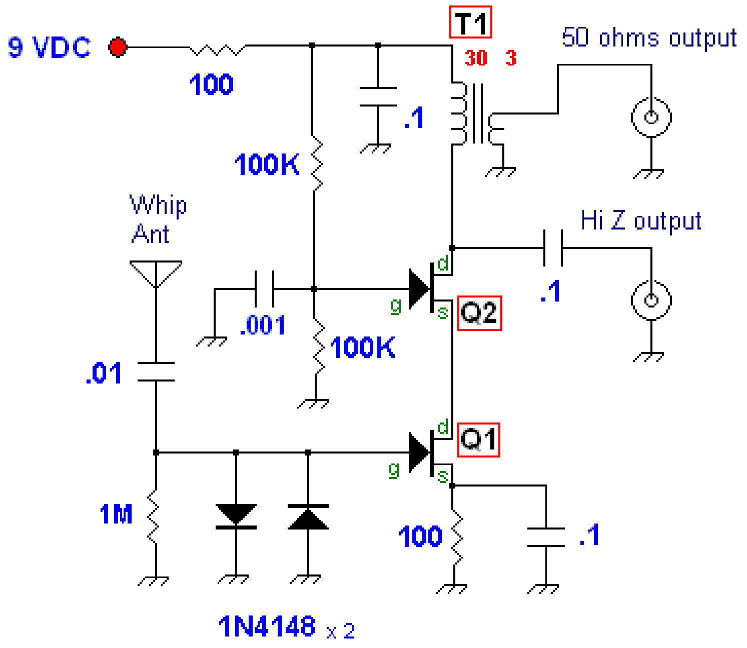

Q1 , Q2 MPF102
T1 30t primary, 3t secondary on a FT50-43 core

Figure 1

Feb 2005 by VE7BPO

The Tuned Whip

Previous experimentation confirmed that it is easy to tune a short whip antenna by connecting it to the hot end of an L C (inductor and capacitor) tank circuit. The high impedance whip antenna was "matched" to a JFET RF amplifier by placing a high value (1 megohm or greater) on the JFET gate to ground. Although this method is practical, I desired a network to transform the output impedance of the tuned whip tank tank circuit to a known impedance. I do not possess the knowledge or mathematical skill to design such a network and asked Wes Hayward if he might consider doing this for me. My desired parameters for the network were 10.0 MHz, a 50 ohm output impedance and a 4 foot (122 cm) whip. Please refer to Wes' calculations and schematic in Figure 2 below. This math is difficult, however, a practical design for experimentation is provided.

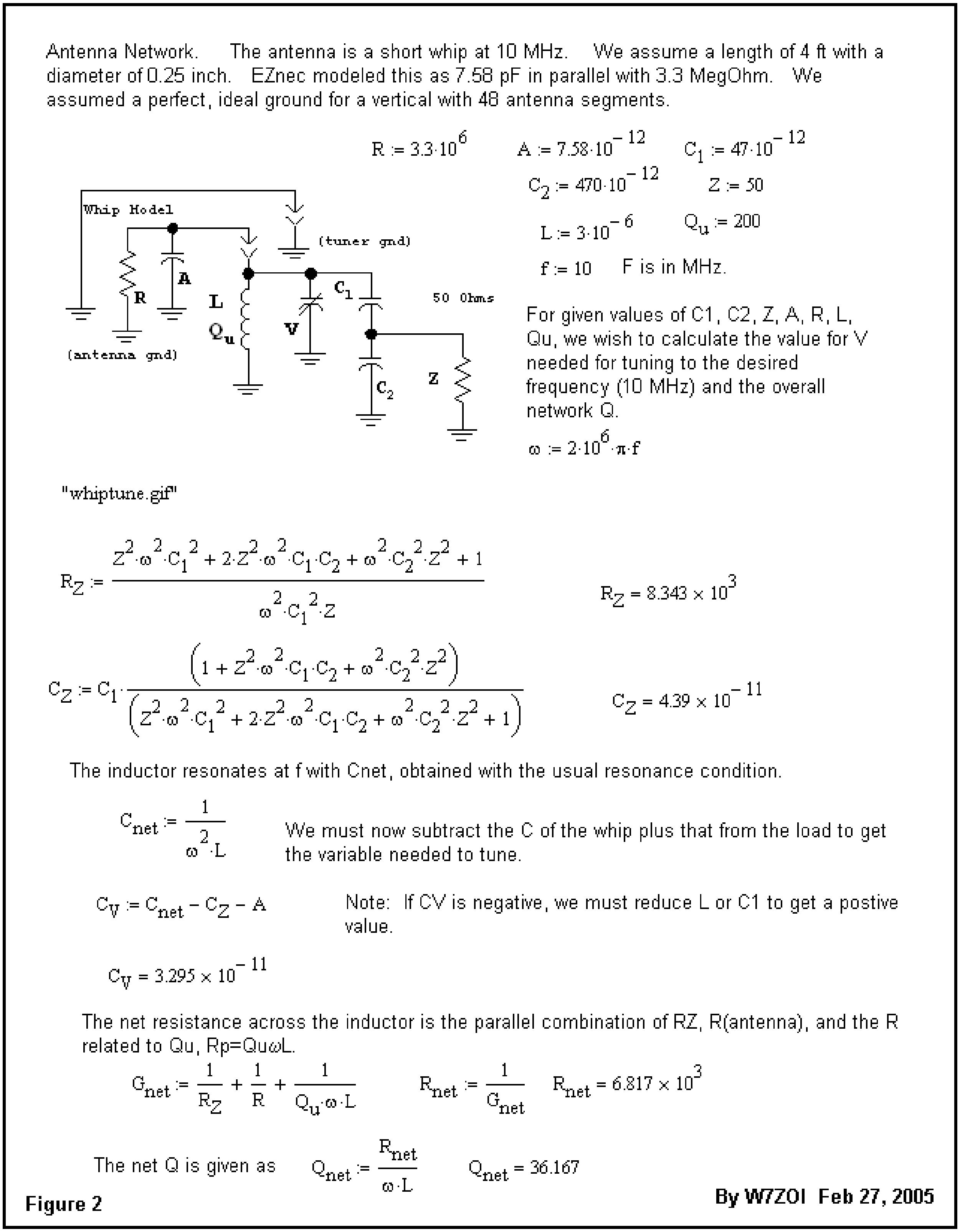

Antenna Network. The antenna is a short whip at 10 MHz. We assume a length of 4 ft with a diameter of 0.25 inch. EZnec modeled this as 7.58 pF in parallel with 3.3 MegOhm. We assumed a perfect, ideal ground for a vertical with 48 antenna segments.

$R := 3.3 \cdot 10^{6}$ $A := 7.58 \cdot 10^{-12}$ $C_1 := 47 \cdot 10^{-12}$

$C_2 := 470 \cdot 10^{-12}$ $Z := 50$

$L := 3 \cdot 10^{-6}$ $Q_u := 200$

$f := 10$ F is in MHz.

For given values of C1, C2, Z, A, R, L, Qu, we wish to calculate the value for V needed for tuning to the desired frequency (10 MHz) and the overall network Q.

$\omega := 2 \cdot 10^{6} \cdot \pi \cdot f$

"whiptune.gif"

$$R_Z := \frac{Z^2 \cdot \omega^2 \cdot C_1^{\,2} + 2 \cdot Z^2 \cdot \omega^2 \cdot C_1 \cdot C_2 + \omega^2 \cdot C_2^{\,2} \cdot Z^2 + 1}{\omega^2 \cdot C_1^{\,2} \cdot Z} \qquad R_Z = 8.343 \times 10^{3}$$

$$C_Z := C_1 \cdot \frac{\left(1 + Z^2 \cdot \omega^2 \cdot C_1 \cdot C_2 + \omega^2 \cdot C_2^{\,2} \cdot Z^2\right)}{\left(Z^2 \cdot \omega^2 \cdot C_1^{\,2} + 2 \cdot Z^2 \cdot \omega^2 \cdot C_1 \cdot C_2 + \omega^2 \cdot C_2^{\,2} \cdot Z^2 + 1\right)} \qquad C_Z = 4.39 \times 10^{-11}$$

The inductor resonates at f with Cnet, obtained with the usual resonance condition.

$$C_{net} := \frac{1}{\omega^2 \cdot L}$$

We must now subtract the C of the whip plus that from the load to get the variable needed to tune.

$C_V := C_{net} - C_Z - A$

Note: If CV is negative, we must reduce L or C1 to get a postive value.

$C_V = 3.295 \times 10^{-11}$

The net resistance across the inductor is the parallel combination of RZ, R(antenna), and the R related to Qu, Rp=QuωL.

$$G_{net} := \frac{1}{R_Z} + \frac{1}{R} + \frac{1}{Q_u \cdot \omega \cdot L} \qquad R_{net} := \frac{1}{G_{net}} \qquad R_{net} = 6.817 \times 10^{3}$$

The net Q is given as $Q_{net} := \dfrac{R_{net}}{\omega \cdot L}$ $Q_{net} = 36.167$

Figure 2

By W7ZOI Feb 27, 2005

Common Gate Amplifier Version

I built the circuit shown in Figure 3 and Figure 4 and tested it on a medium grade SWL receiver (Realistic DX 300), rather than an expensive Amateur Radio receiver. I required this active antenna for experimenting with a 10 MHz WWV superheterodyne receiver I am designing. For practical analysis, VE7TW and I did listening tests with a commercially made 4 foot telescopic whip antenna that is fitted to a standard PL-259 connector (Figure 5) and his deluxe multi-band commercial SWL antenna up a 25 foot tower.

The Figure 3 VPA was very quiet
and pulled in WWV much better

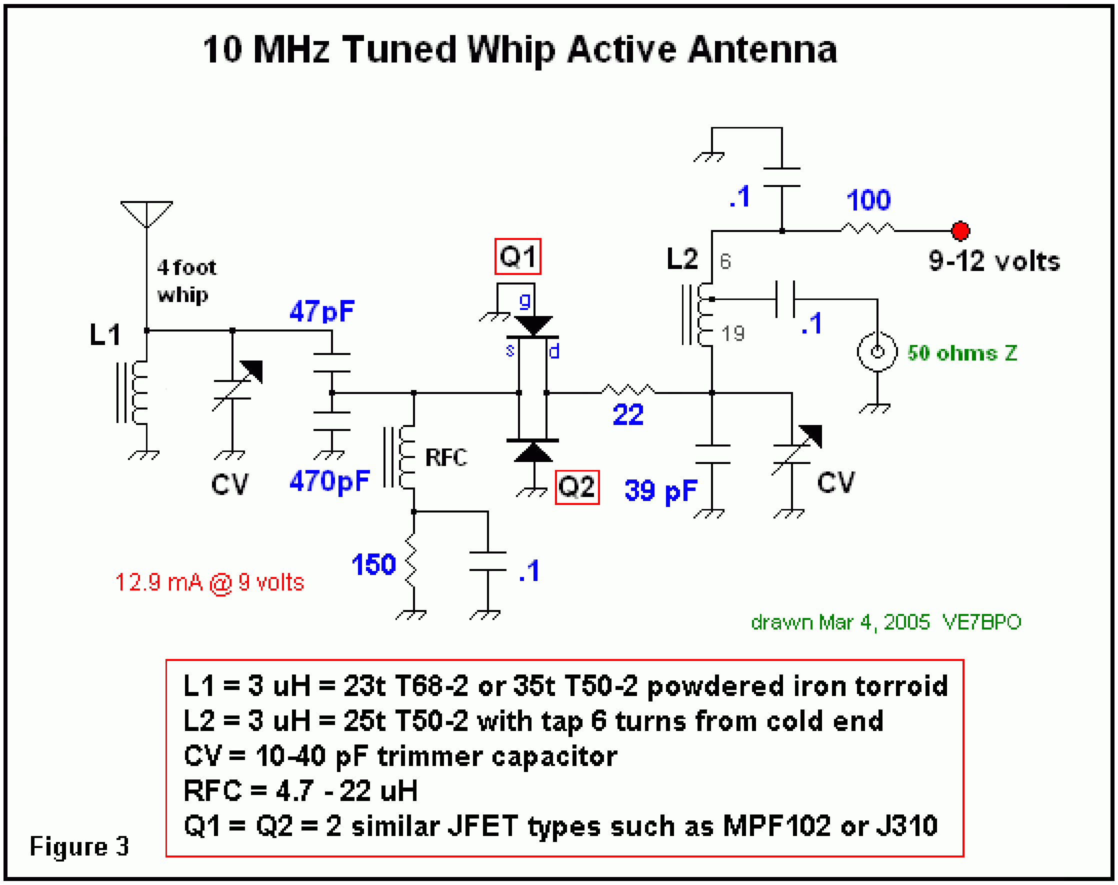
10 MHz Tuned Whip Active Antenna

L1 = 3 uH = 23t T68-2 or 35t T50-2 powdered iron torroid
L2 = 3 uH = 25t T50-2 with tap 6 turns from cold end
CV = 10-40 pF trimmer capacitor
RFC = 4.7 - 22 uH
Q1 = Q2 = 2 similar JFET types such as MPF102 or J310

Figure 3

than the plain 4 foot whip of Figure 5, however, received station signal strength was quite weak when compared to the outside antenna. Our conclusion was that considering the significant losses of the 4 foot whip antenna it was connected to, the common gate RF amp does not likely have enough voltage gain to please most builders. This amp did present a low impedance to the whip network and no spurious oscillation were measured on the bench. Do not omit the 22 ohm or similar value resistor in the drain of the FET. It is used to to push the UHF parasitic oscillation tendency into the ground. Such oscillations will trash your receiver mixer intermodulation performance.

The 150 ohm source resistor can be increased in value and/or 1 of the JFETs removed if you wish to reduce the current draw on a 9 volt battery. It might be better to substitute 1 better JFET such as the J310 rather than use the "popcorn" MPF102 as shown. This VPA may be practical for a receiver that has an existing broadband RF preamplifier. The tap on L2 was found experimentally and the output impedance is probably higher than 50 ohms, but is likely a reasonable low impedance match to most receiver front ends. A broadband transformer for L2 might also be a good choice.

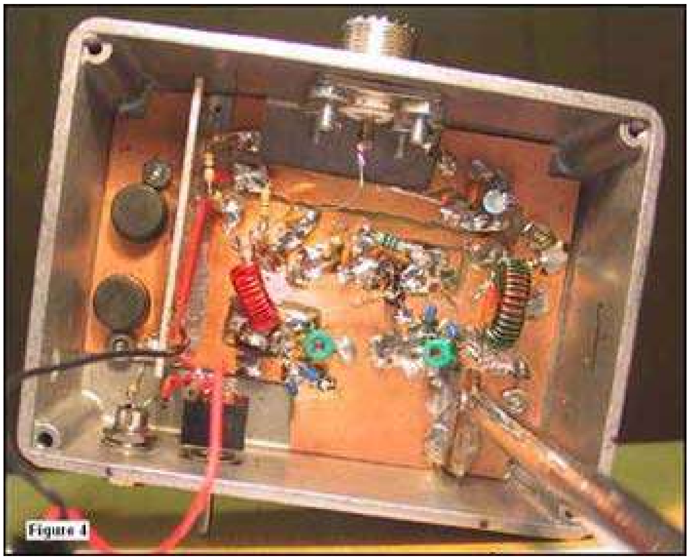
Figure 4

Figure 5

Figure 4 above: For the whip network, it is critical that you use a inductor that has an unloaded Q of 200 or above. Practically speaking, this means you cannot use a fixed value inductor such as an epoxy coated or molded RF choke. Use a powdered iron torroid instead. In my test VPA designs, for L1, I used a T68-2 core wound with (the green) 22 gauge enamel coated wire to get as high an unloaded Q as possible. If you use the T50-2 core, use 24 gauge wire if possible. Higher Q = lower losses.

Figure 5 above: This is a 4 foot whip antenna factory connected to a PL-259 that came with the Realistic DX300. It presents a very high impedance to the test receiver front end and probably wasn't a good choice to compare the VPA designs to.

Cascode JFET Amplifier Version

It was decided to use a cascode JFET amplifier to obtain more power gain. The whip network was changed to try to match the 10K input of the JFET amplifier shown in Figure 6. The whip network capacitor values (150 and 33 pF) were calculated to the best of my ability. This amplifier was tested in the same manner as the Figure 3 design. It worked very well. The WWV signal that morning was not very strong and could not even be heard with the plain 4 foot whip. The signal strength of the tuned whip was just below that of the outside antenna. The outside antenna was much quieter however and had less fading. The tuned whip antenna was quite noisy in comparison. The high gain RF amp brought up the strength of the environmental noise sources in the house. The RF gain of the receiver was reduced to compensate for the added noise.

Another problem was noted with this and other tuned drain versions of the cascode JFET amplifier; instability.

Recently, I connected a tuned drain version of the Figure 1 VPA to a receiver that contained a tuned input stage and was able to measure oscillations in the VPA with my scope. The FET drain tank and the tuned input amp seemed to be interacting.

A "swamping" resistor was placed across the VPA drain tank circuit. I had to use a resistor value of less than 1200 ohms to eliminate this instability. This greatly reduced the gain and selectivity advantage of a tuned output and I realized that output tuning may be impractical for many reasons. Some SWL builders use regenerative receivers and such a problem would be disastrous. I sent the Figure 6 schematic to Wes Hayward and he suggested using a broad or wide band amplifier as shown in Figure 8.

Instability can also occur in broadband output versions and a swamping resistor is still necessary but is used mostly to force an output impedance so that a transformer can be designed.

All of the cascode JFET amplifiers shown have fixed bias on Q2. Variable gain is possible by changing the bias voltage on Q2 with a voltage divider and/or modifying the amplifier circuit to give a greater range of bias controlled voltage gain. Please refer to EMRFD page 6.17 for information regarding this. A switchable resistor attenuator might also be practical for some builders.

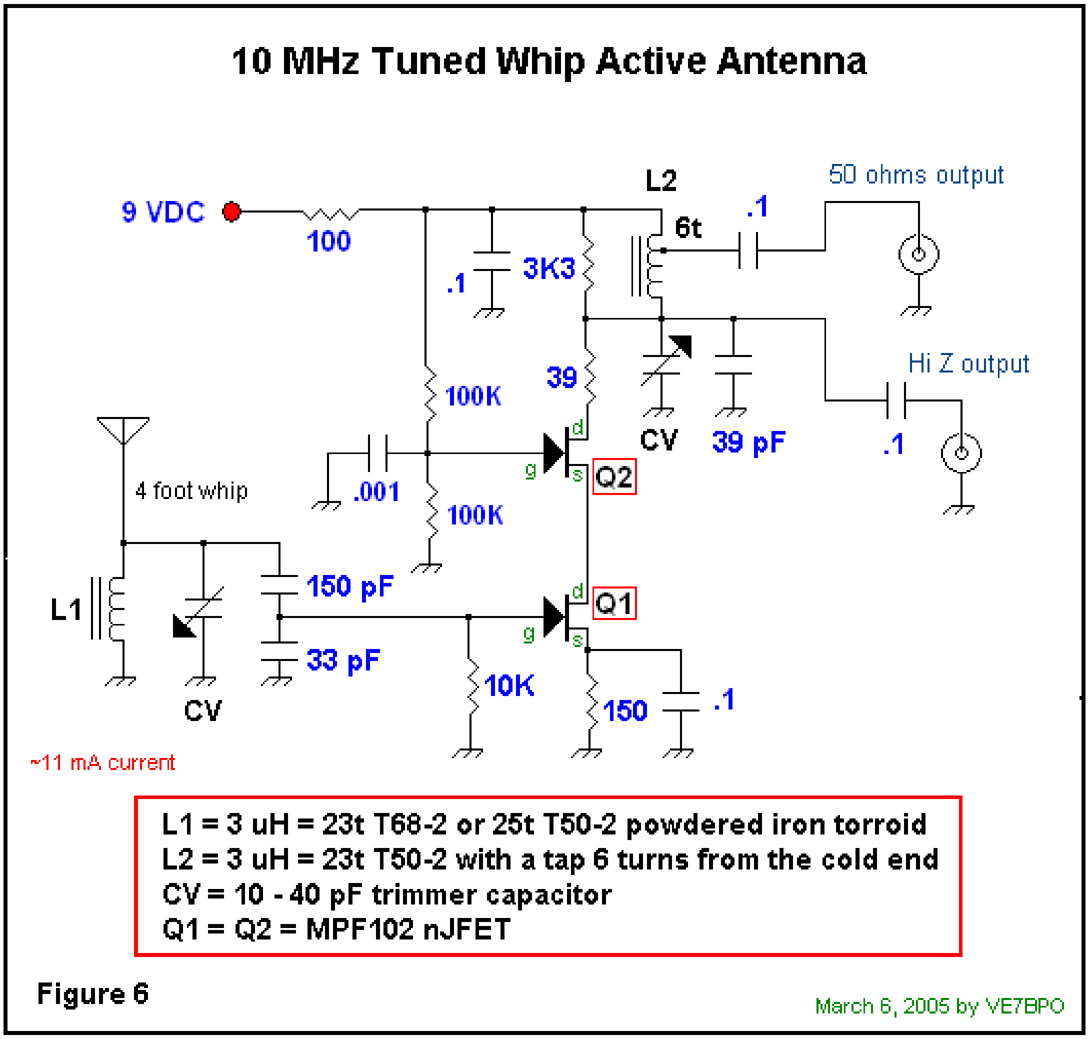
10 MHz Tuned Whip Active Antenna
9 VDC
100
.1
3K3
L2
6t
.1
50 ohms output
39
100K
CV
39 pF
Hi Z output
.1
.001
4 foot whip
d
g
s
Q2
100K
L1
150 pF
33 pF
d
g
s
Q1
CV
10K
150
.1
~11 mA current
L1 = 3 uH = 23t T68-2 or 25t T50-2 powdered iron torroid
L2 = 3 uH = 23t T50-2 with a tap 6 turns from the cold end
CV = 10 - 40 pF trimmer capacitor
Q1 = Q2 = MPF102 nJFET
March 6, 2005 by VE7BPO

Figure 6

Figure 6 above: The tuned whip network is connected to a cascode JFET amplifier. A dual gate MOSFET would also be a great choice. I have many on hand, but chose the cascode JFET topology because many builders no longer have access to these devices or prefer not to use the more available surface mount types. They are also more expensive.

Figure 7 to the right: Detail of the tapped L2 inductor should you decide to experiment with a tuned drain version or need one for another project. Wind your coil and leave an extra long loop for your wire tap. Cut the tapped loop at the midpoint and use a small piece of folded ~150 grit sandpaper to remove the enamel from each of the 2 wires. Twist the now bare wires together and lightly solder them. Cut the end wires to the required length and use the sandpaper to remove the enamel. A method to strip enamel off wire is a frequently asked question for me and sandpaper works

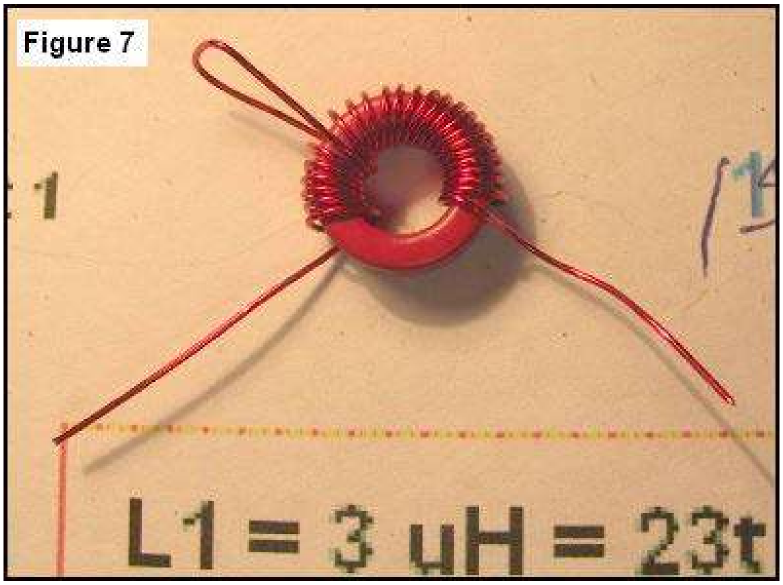
Figure 7

well.

Broadband Output Version

Figure 8: The broadband version W7ZOI suggested to try building. I modified the output transformer in the Figure 6 project and tested it with Tom, VE7TW. We really liked it. By adjusting the network trimmer capacitor, I was also able to tune the 30 meter Amateur radio band as well. For 30 meter band use, I peaked the tank circuit at 10.125 MHz by listening to receiver noise with a home brew direct conversion receiver and was suitably impressed.

This is the active antenna design I wll use for my future projects where strong voltage gain is required. If your receiver has a higher impedance such as 500 ohms, you might try using a couple more links on the output transformer secondary winding.

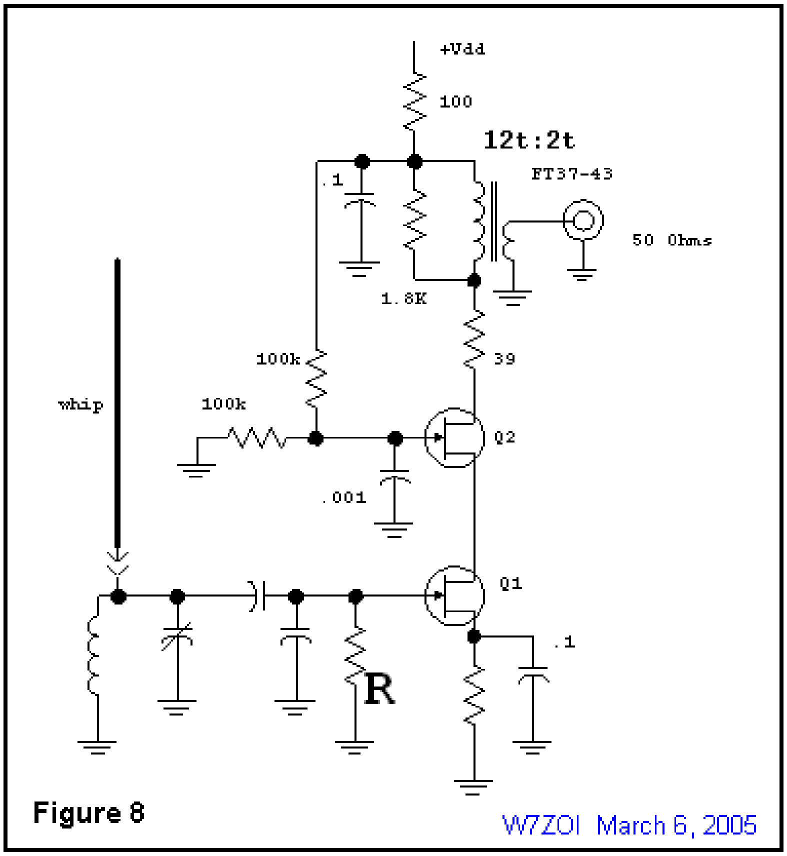

Figure 8
W7ZOI March 6, 2005

Tuning a Whip To Other Frequencies

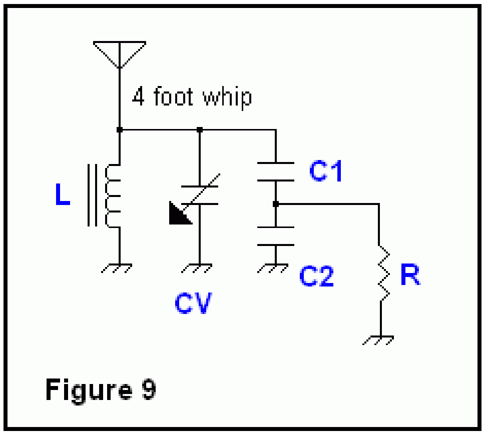

Figure 9

The ability to calculate the network values for different frequencies may prove difficult for those who lack software and/or math skills. To that end, a table follows which has some radio frequency bands and some suggested starting values for the Figure 9 parameters. Please note these are calculated and are suggested starting points based upon my limited understanding of radio electronics. Experimentation is the best method to find what component values will work for you.

Emails regarding the component values used in actual experiments is greatly welcomed.

The **R** value is the input impedance of your RF amplifier. In the case of the cascode JFET amp, it is the Q1 gate resistor.

Note that the actual circuit CV value is typically much lower than the suggested (calculated) CV value from the chart. CV is used to **resonate** the tank. CV is dependent on several factors including the capacitance of the whip antenna, your RF amplifier input capacitance , your circuit layout, component lead lengths and variations in the powdered iron core and C1 and C2 capacitors values. Expect that your whip antenna will exhibit between 8 and 15 pF of capacitance. You need to subtract this from the suggested (calculated) CV value from the Figure 10 table.

Wes, W7ZOI told me that the whip antenna capacitance will remain constant as you change frequency providing you are below 1/4 wavelength for a given frequency. Here is a good web site applet to calculate wire or whip 1/4 or 1/2 wave lengths per frequency: http://www.csgnetwork.com/antennagenericfreqlencalc.html.

The actual circuit CV might include a trimmer capacitor plus a parallel fixed value capacitor.

How to find the correct trimmer capacitor for any tuned circuit you wish to resonate

I suggest you chose a circuit CV value by placing a variable trimmer capacitor in your circuit that when set to minimum will be below half or more than the calculated CV value. Then peak the whip tank circuit using a test oscillator and scope or RF probe or by just using receiver noise. Now temporarily add a 5-10 pF capacitor in parallel with trimmer capacitor. Just barely solder it in place or even just hold it in place without touching the leads. If the output increases, you were under the correct circuit CV value. Add more capacitance and check again. Repeat until you are satisfied with your chosen circuit CV value.

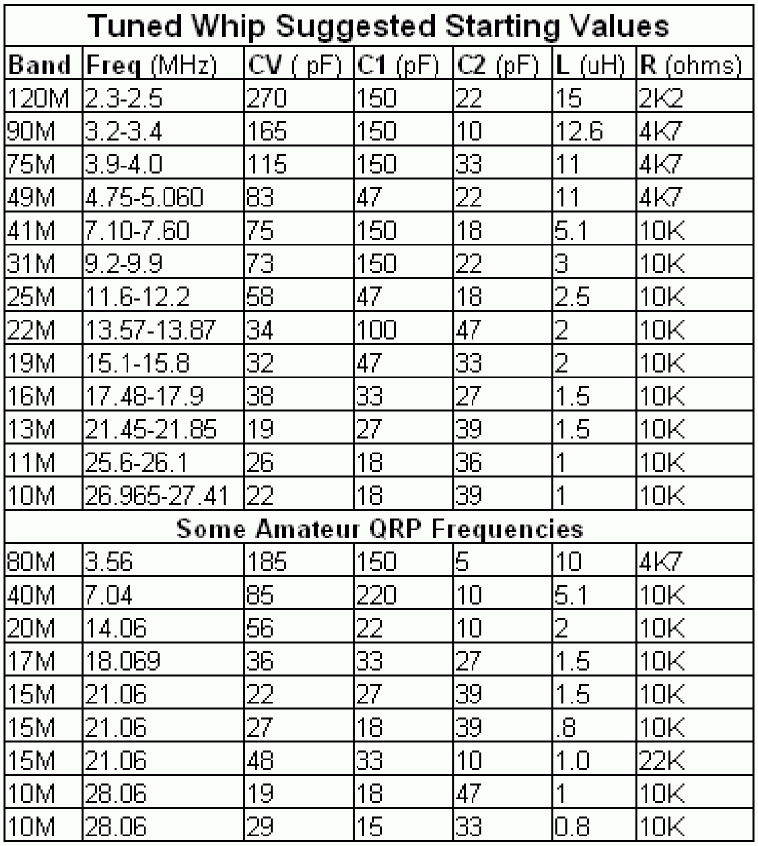

Tuned Whip Suggested Starting Values

Band	Freq (MHz)	CV (pF)	C1 (pF)	C2 (pF)	L (uH)	R (ohms)
120M	2.3-2.5	270	150	22	15	2K2
90M	3.2-3.4	165	150	10	12.6	4K7
75M	3.9-4.0	115	150	33	11	4K7
49M	4.75-5.060	83	47	22	11	4K7
41M	7.10-7.60	75	150	18	5.1	10K
31M	9.2-9.9	73	150	22	3	10K
25M	11.6-12.2	58	47	18	2.5	10K
22M	13.57-13.87	34	100	47	2	10K
19M	15.1-15.8	32	47	33	2	10K
16M	17.48-17.9	38	33	27	1.5	10K
13M	21.45-21.85	19	27	39	1.5	10K
11M	25.6-26.1	26	18	36	1	10K
10M	26.965-27.41	22	18	39	1	10K
Some Amateur QRP Frequencies						
80M	3.56	185	150	5	10	4K7
40M	7.04	85	220	10	5.1	10K
20M	14.06	56	22	10	2	10K
17M	18.069	36	33	27	1.5	10K
15M	21.06	22	27	39	1.5	10K
15M	21.06	27	18	39	.8	10K
15M	21.06	48	33	10	1.0	22K
10M	28.06	19	18	47	1	10K
10M	28.06	29	15	33	0.8	10K

Figure 10

If after adding the initial 5-10 pF capacitor, the output decreases, try peaking the tank again to see if you can restore the signal strength you had before you added the temporary capacitor. If after peaking, the signal strength is down, you now have too much capacitance and can remove the temporary capacitor.

You just might also have too much capacitance. You might try a smaller variable cap or reduce the value of any fixed capacitors in parallel with your trimmer to make sure your minimum capacitance is not too high to properly resonate the input tank circuit.

The point is you need to be able to tell if you have too little or too much capacitance for CV and by going under and over you can tell if you are truly resonating the tank when you adjust the trimmer capacitor. Experimentation will tell you.

Another option is to put in a front panel adjusted variable capacitor. Front panel switchable inductors might also allow other bands to be tuned with 1 tuned whip network.

Moving your body as you adjust the trimmer capacitor can change the tuning, so please keep this in mind.

Figure 10 above. Picking an inductor value for the whip network can be tricky and sometimes trial and error is required. This table may be used to find starting values for the Figure 9 network. Below 41 meters, I suggest trying a lowered RF amp input impedance as shown to allow practical component values. Most of these calculations have not been tested.

I think an indoor active antenna for 74 meters or below might just be a noise generator.

I chose a frequency mid band for any given SWL band on the chart. The bandwidth of the input network is wide enough so this should be suitable to cover a good portion of the band.

Building An Active Antenna

To build this active antenna, chose the input tank network values from the Figure 10 chart or from your own calculations and then use them in the Figure 6 circuit. The Q1 source resistor can practically be from 100 to 390 ohms depending on how long you need your 9 volt battery to last. Increasing this resistor value will reduce the amplifier power gain. Try different values and see for yourself!

Some Practical Examples: 40 and 41 Meter Band

An active antenna that would provide coverage of the entire 40 meter Amateur Radio and 41 meter Shortwave band was designed. A varactor diode was used as the tuning element. The tuning voltage to the varactor was controlled by a 10K potentiometer which also had an integral switch. The finished VPA is shown in Figures 11 and 12.

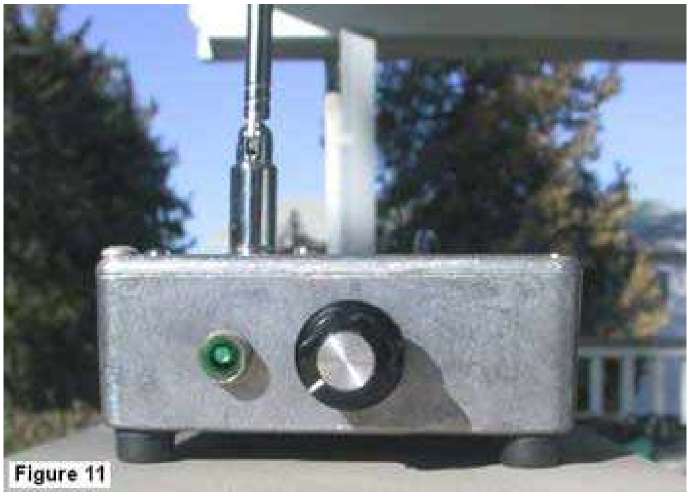
Figure 11

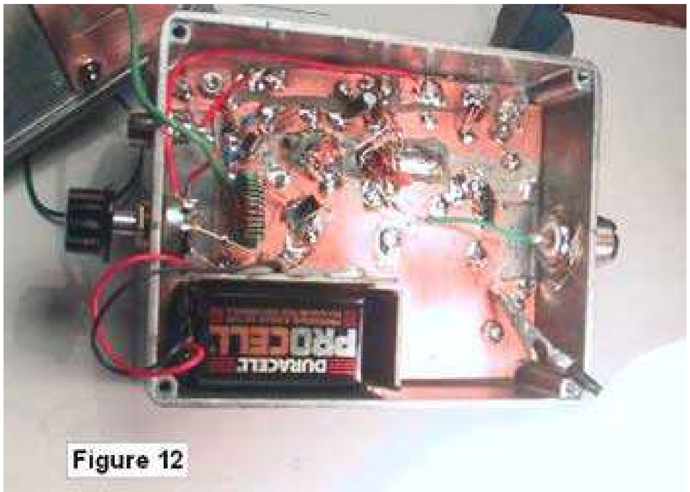
Figure 12

Figure 13 below: A hotter JFET, the J310 was used in this VPA. In the test receiver, I was able to peak a signal from ~ 6.90 to 7.60 MHz. Tuning is very sharp but peaking is easily performed by turning the potentiometer gently back and forth while listening to receiver noise or a station. It might have been better to use a smaller value zener diode as when the 9 volt battery fades below 7.5 VDC, the zener diode will not conduct and the voltage regulation will fail. Having said that, this "hotter" VPA is intended for use with an external power supply as current draw at 9 and 12 volts is 19 mA and 28.9 mA respectively.

I tried using this VPA as the antenna for the Cascode 7 receiver shown elsewhere on this web site. When the VPA was peaked at the receiver tuning frequency, loud oscillations occurred. The receiver and the VPA were about 1 meter apart. I had to turn on the -10 dB attenuator and detune the VPA for the oscillations to stop. Moving my hand near the whip antenna varied the oscillations. The high gain, tuned circuits of the Cascode 7 receiver are not a good choice for an active antenna.

Future receivers projects will have a integral VPA and clearly the front end of these receivers will have to be designed carefully. A low cost Grundig receiver was also overloaded with this VPA.

This VPA worked well with other receivers which did not have a tuned, high gain preamp.

Figure 14: From the chart, the MV209 exhibits about 44 pF (guessing) when 0 volts are

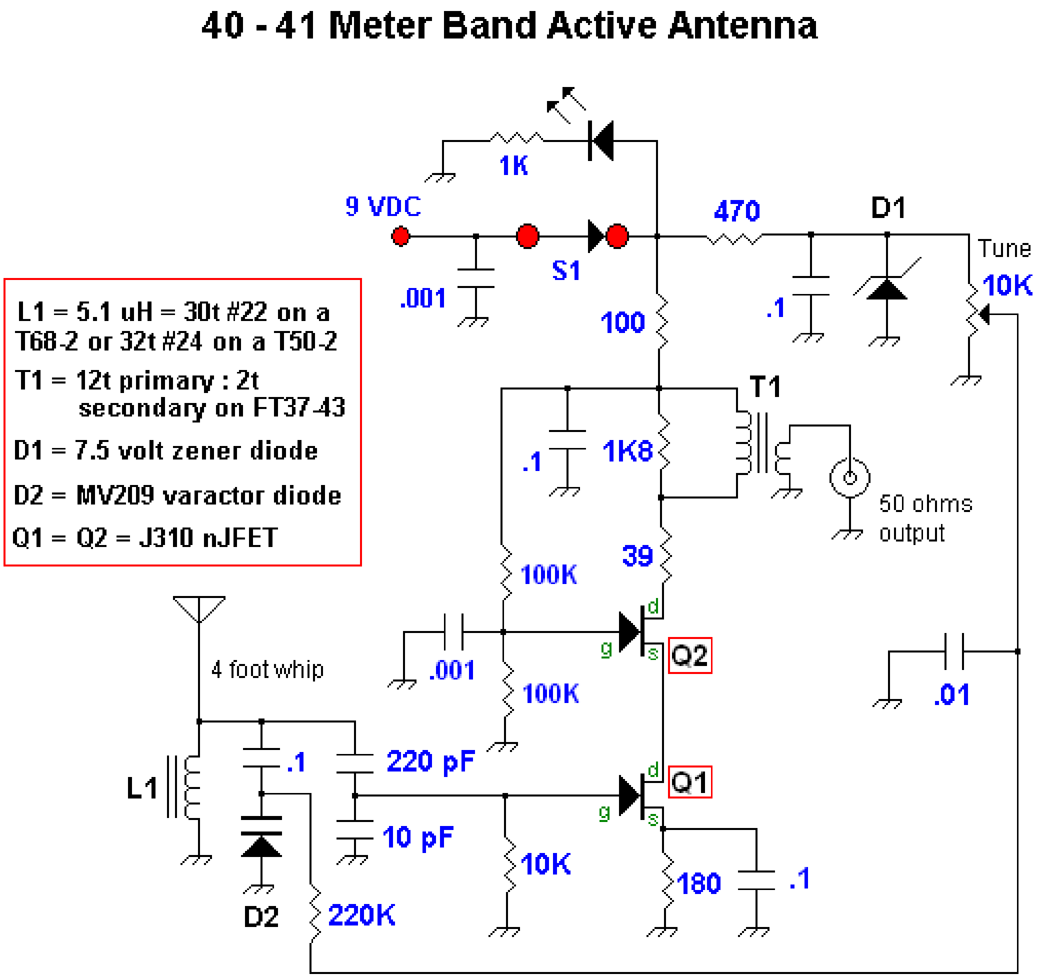
40 - 41 Meter Band Active Antenna
1K
9 VDC
S1
470
D1
Tune
10K
.001
100
.1
L1 = 5.1 uH = 30t #22 on a T68-2 or 32t #24 on a T50-2
T1 = 12t primary : 2t secondary on FT37-43
D1 = 7.5 volt zener diode
D2 = MV209 varactor diode
Q1 = Q2 = J310 nJFET
T1
.1
1K8
50 ohms output
100K
39
d
g
s
Q2
4 foot whip
.001
100K
.01
.1
220 pF
d
Q1
L1
g
s
10 pF
10K
180
.1
D2
220K

Figure 13

Mar 12, 2005 VE7BPO

applied to it.

The VPA was built and tested before the tuning diode components were added. A 7.039 MHz crystal oscillator with a piece of wire for an antenna was used as an RF source. The T2 secondary was connected to ground via a 47 ohm load resistor.

A 47 pF fixed capacitor was lightly soldered in parallel with L1 and the voltage was measured with an oscilloscope. A 10 pF capacitor was then carefully held across L1 and the voltage increased by 0.25 volts. A 22 pF capacitor was then tried and the voltage decreased much below that of just the 47 pF capacitor. The nearest standard value I had on hand above 47 pF was 56 pF. The 47 pF capacitor was removed and replaced with the 56 pF one. I tried holding a 5 pF capacitor in parallel with the 56 pF and the measured voltage decreased. I had experimentally determined that to resonate L1 at 7.039 MHz I needed between 47 and 56 pF for the CV value. This range should be close enough to resonate the tank at 7.0 MHz as well.

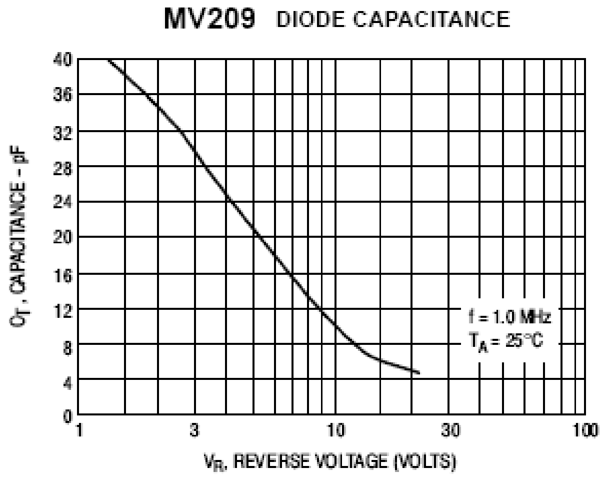

Figure 14

I then chose a varactor diode. The MV209 would be perfect for my project based upon the Figure 14 chart. I anticipated that I might have to place a small trimmer capacitor in parallel with the varactor to resonate the tank at the my lowest design frequency which was 7.0 MHz. As it turned out, in addition to the varactor capacitance, the voltage control circuit added additional capacitance and I actually needed 0.30 volts (measured between the 10K pot and the 220K resistor) to resonate the whip at 7.039 MHz. This was perfect; I did not need a trimmer capacitor! At 0 volts to the varactor diode, my whip resonated ~ 6.90 MHz.

---

5 MHz WWV Cascode Bipolar Amplifier

I wanted to build a non-FET version as shown in Figure 15. The tuned whip tank was originally resonated with a 5 - 40 pF trimmer capacitor. I unsoldered this trimmer capacitor and measured it with a meter; it was 27 pF. A 27 pF fixed value capacitor was soldered in and tested. The circuit was resonant at 4.98 MHz. This was close enough for me and also the 3 high Q fixed value capacitors provided a very narrow 6 dB bandwidth along with the inductor. The output impedance value of the tuned whip is around 200 ohms to match the Q1 bipolar amp input impedance.

Listening tests indicated that this circuit probably had too much gain at 5 MHz. It might be favorable to lower the Q1 emitter current to 7 mA or so by raising the Q1 emitter resistor or decreasing the Q1 bias voltage. Also, a series feedback, degenerative resistor on the Q1 emitter might be considered. This active antenna was comparable to the outdoor reference aerial for signal strength, however, predictably was much noisier.

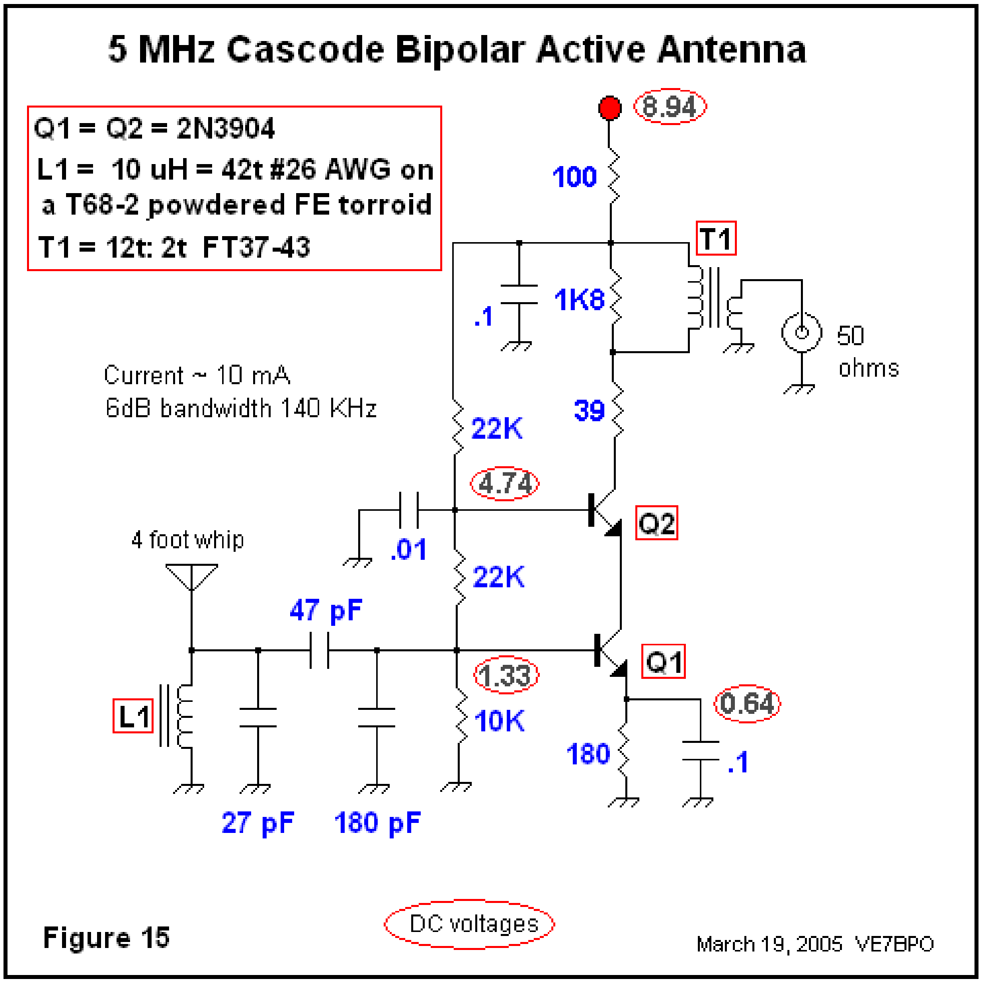

Figure 15

Final Comments

I found that using lower Q trimmer and fixed value capacitors undesirably increased the tuned whip bandwidth presumably by lowering the resonant circuit Q. The inductor unloaded Q was the dominant factor however. The worst case scenario was a tuned whip built with junk-quality parts which had a -6 dB bandwidth of ~390 KHz.

I also learned that you should expect high gain amplifiers to oscillate and specifically design to reduce or suppress this tendency.

The 2005 Active Antenna experiments were fun and provided many learning opportunities. An active antenna is a perfect weekend project. There is no substitute to learning by building and testing electronic circuits with your own hands.

My sincere thanks to all of the friends who helped me with these experiments.

Experiments by Other Builders

What follows are some VPA experiments by others that were sent to me by email. I seek your feedback and photographs to help improve this web site and also to gain motivation to add more new content.

Joe, K9LY
Hi, Todd :

Attached are some photos of a voltage probe amplifier that I built using ideas from your excellent website. I use a TenTec 1254 receiver in the car and listen to some shortwave broadcasts during the daily commute to and from work. The antenna is a 4-foot whip that screws into the trunk-lip mount shown. The amplifier is held to the bottom of the trunk lid by a magnet and has survived for several weeks without falling off.

The amplifier is tuned by a varicap diode and covers approx. 9-14 MHz. The tuning voltage comes from a potentiometer that I added to the front panel of the receiver.

I decided to use the 2N3904 cascode amplifier because I liked the idea of using the most common transistor possible. The LC-tuned input is nice because the antenna whip is held at chassis ground potential, which should help prevent damage to the amp caused by static buildup.

The TenTec 1254 Receiver. http://radio.tentec.com/kits/Receiver

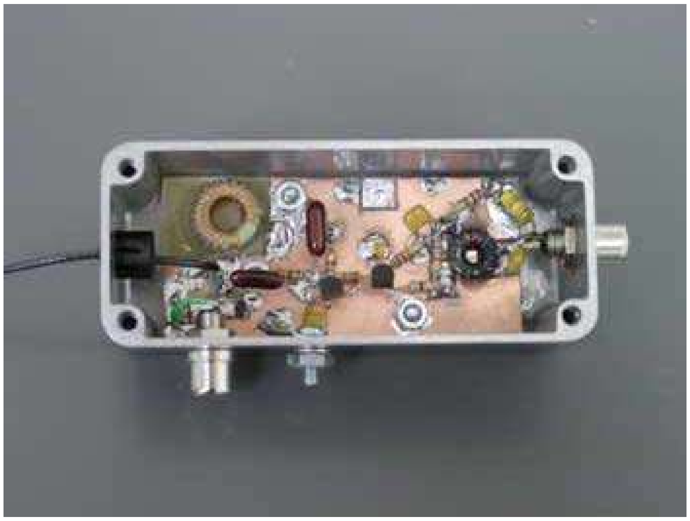

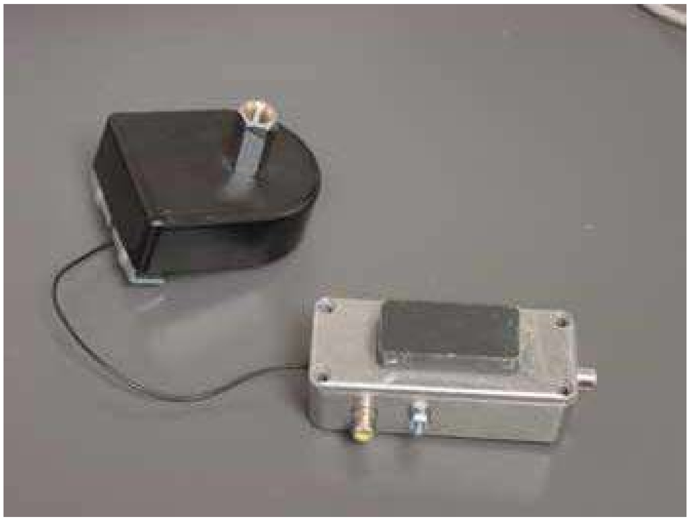

Above left: Inside Joe's trunk lid SWL Active antenna. Great ugly construction in a sturdy Hammond chassis.
Above right: Joe's remotely tuned SWL active (or voltage probe) antenna amp and whip holder. Thanks Joe!

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Fun with LEDs

Introduction

This summer I built several LED projects including sequential LED chasers (sequentially left to right) and also "Nite-Rider" style which go (sequentially left-right-left-right-etc). Many more LED schematics can be found on the World Wide Web via a Google search. LED projects are great fun for both HAM's and SWL's alike. They are also a lot of fun for children to experiment with. Currently, I am experimenting with PIC microcontrollers to perform LED "tricks".

I also built several very bright LED flashlights which run on a single 1.5 volt battery. For ultra-bright LED flashlight schematics, check out Dick Cappel's excellent and very informative web site. He has a number of LED driver circuits and other great schematics and theory. To wind the inductor for these LED flashlights, I had good success using an FT-37-43 ferrite torroid core. I used at least 40 turns of wire which is generally center tapped.

LED Chasers

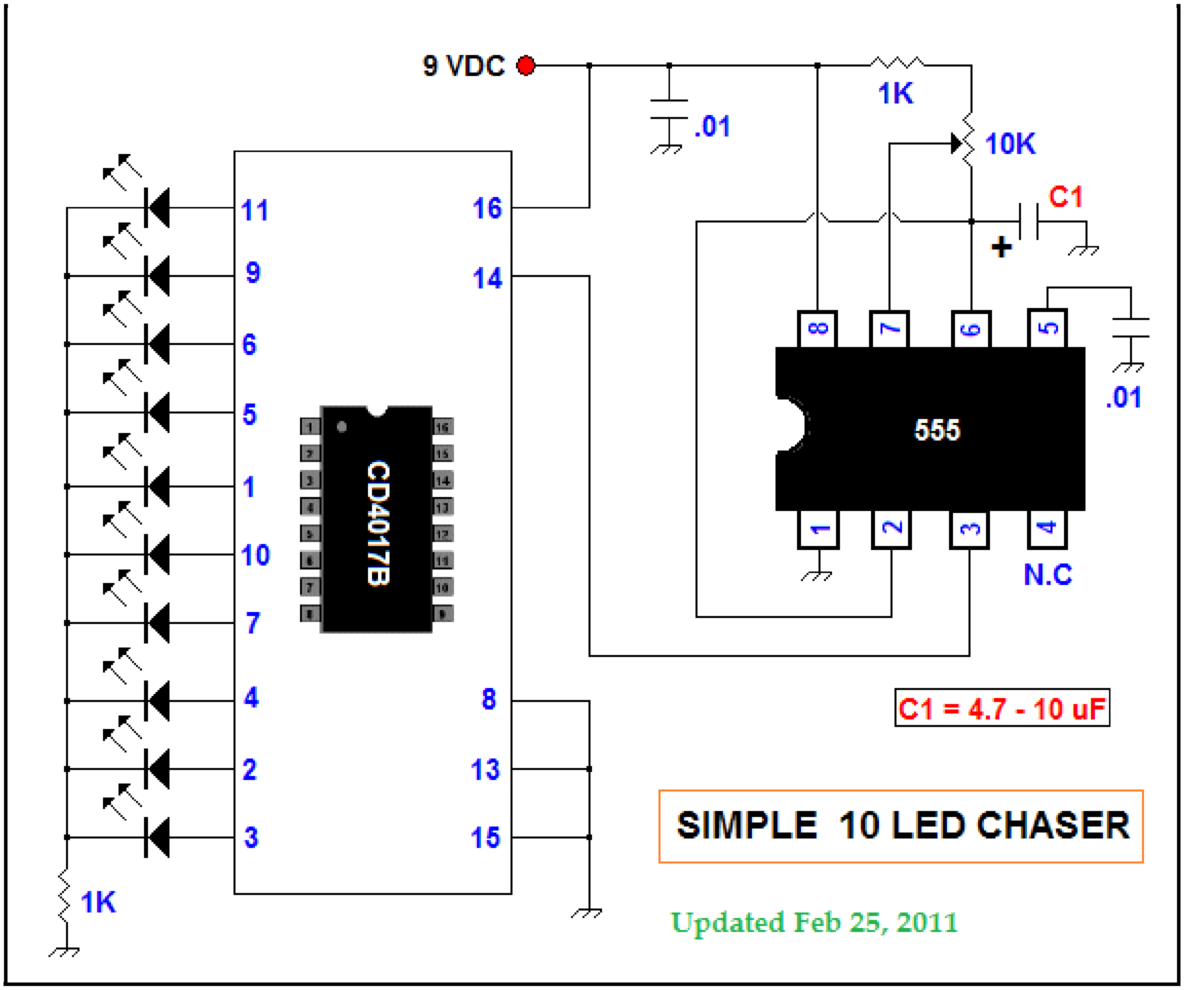

Above. This is the schematic for a ***very*** basic 10 LED chaser I built. I prefer my "chasers" to run slower than most and chose a 10 uF capacitor for C1. The 10K pot can reduce the flash speed from not moving to whatever minimum time constant is possible with the C1 value you choose. Don't bother with ultra-bright LEDs for these "chaser" projects as cheaper, lower millicandela (mcd) LEDS work fine. I favor blue and green LEDS. The 4017 decade counter is a fabulous part and can be driven to flash a row of LEDs with a 555 timer chip or a discrete BJT multivibrator.

Update Feb 25, 2011: Many "experts" have emailed, or flamed me on web forums to say this circuit can't possibly work. I really hope these unhappy men cultivate enough humility to study and understand the 555 and more importantly; to reap some of the happiness and joy that comes from being positive and helpful to others. Since 2005, greater than 300 builders have emailed to say this simple circuit works and they want to learn more about electronics.

My intent was to have the least number of parts to flash some LEDs. Some new builders become overwhelmed when the parts count is high — I once shared this fear and relate. You'll see a number of different bias circuits for the 555. Many builders run the reset pin; **Pin 4** high (connected to the 9 volt battery) and as a rule, this is a good thing to do, but it's not necessary for the circuit to work. Pin 7 is an open collector output to ground — its primary purpose is to discharge the capacitor.

It's important for the DC voltage in the pot wiper to not become too close to the + 9 volt rail or VCC (This happens when when the pot is rotated so that maximal DC voltage appears on pin 7), as pin 7 would draw excessive current. In my original schematic I left out a series resistor from VCC to the pot to eliminate this problem. After some thought, I added a 1K resistor on Feb 25, 2011, although this limits the rate somewhat. I run the pot on my circuit about mid-range and it hasn't been re-adjusted (or turned off) since 2005.

Also, the rotated pot wiper shouldn't get within a couple of hundred ohms of the capacitor as that too would cause excessive current spikes into pin 7. Generally, I prefer not to have much current on the wiper of a pot or, at least, try to keep the current small. Often, you can use a pot to set the desired timer speed and then remove, measure and substitute 2 standard value fixed resistors. A better way is to use math and calculate the resistor values, but this involves math and some people want nothing to do with equations.

This circuit is meant to provide a minimalist working circuit, but doesn't provide a great example of 555 design. Happily, for those wanting to learn more, countless great 555 tutorials may be found on the web. One of my 555 favorite sites is that managed by fellow Canadian Rob Paisley. Increasingly, I am exchanging electronics-related emails with model railroad enthusiasts across the globe and many of them know of Rob's wonderful web site.

Matthew Ritchie built and posted a nice version of the LED Chaser on YouTube. A reader sent in this breadboard photo.

By far, the coolest device incorporating the LED chaser lies within a sculpture called Cyanic by Seattle-area artist Allet. Click for his web site. Cyanic may be found on the New Sculpture Build section on his web site. You have to start the Quick Time video manually with a mouse click. I love Allett's work and his lastest light sculpture exemplifies how the Internet can unite creative people with a positive attitude.

A 10 LED sequential flasher in a blue Hammond chassis. The schematic is shown above.

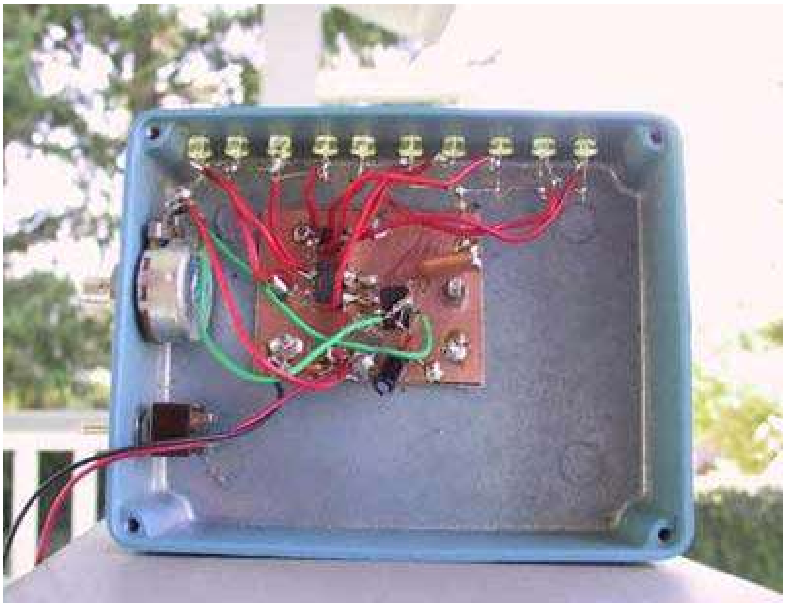

An RC oscillator designed for a 3 volt LED chaser. It oscillates quite slowly so the LED chaser it triggers will not be overly distracting. Some RC oscillator design details are discussed later. This oscillator triggered a 4017 decade counter instead of the 555 timer chip shown in the "Simple 10 LED Chaser" schematic. There are many links describing the theory of the 2 transistor astable multivibrator on the World Wide Web. I also have some information on this web page.

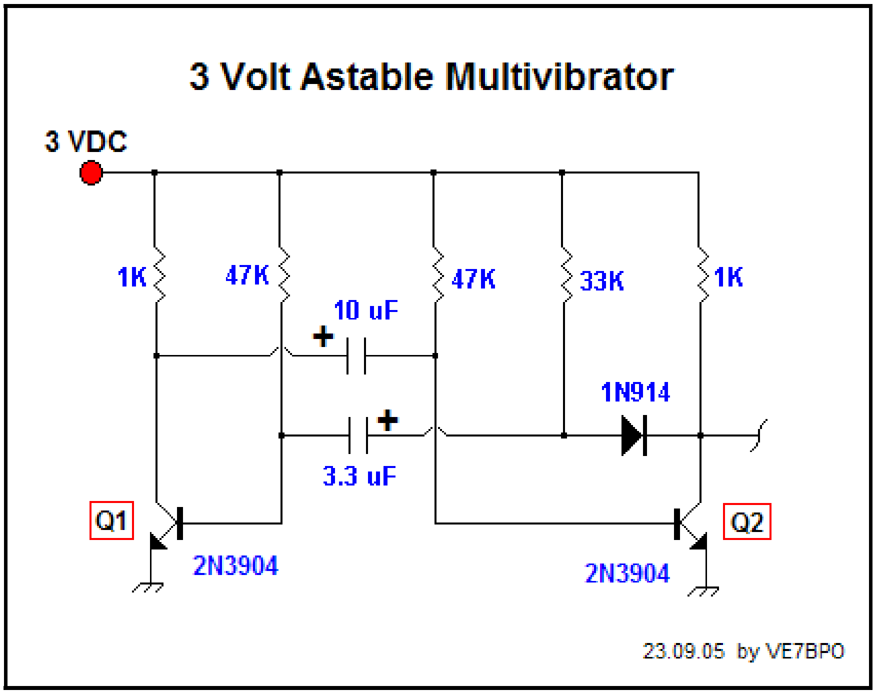

Above . This is a tiny 3 volt chaser which uses an LED bar instead of discrete LEDs. It draws 3.8 mA peak current on pulses. It uses the optimized BJT astable multivibrator shown directly above which fires at ~120 cycles per minute (slowly). The 3 volt battery pack is hidden behind it and should last several months. Soldering the LED bar was not an easy task. The plastic Hammond case measures 2.46 by 1.38 inches (6.25 by 3.5 cm).

A schematic to allow the 4017 decade counter to sequentially flash 6 LEDS left-right-left-right-etc. Connect your favorite square wave oscillator to pin 14. I built 4 of these using various oscillators and LED colors. You might consider using lower DC voltages and if so, may adjust the 1K current limiting resistor by using ohm's law. The 10 small signal diodes may be any appropriate type including the 1N914 or 1N4148. None of my 4 projects exceeded 6 mA peak current draw, so battery life is excellent. I increased the 1K resistor to 1K5 in my 4th project as I found the LED's that I used too bright.

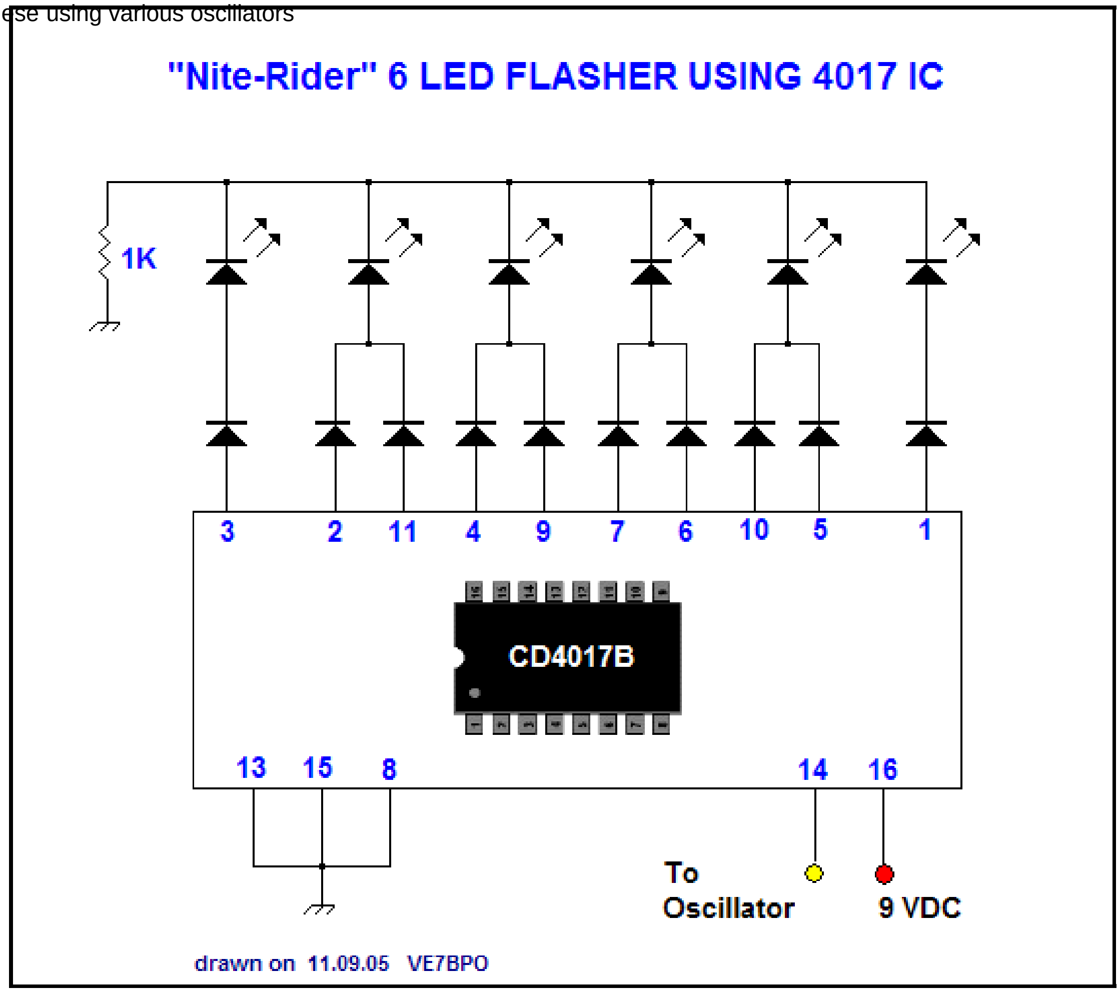
"Nite-Rider" 6 LED FLASHER USING 4017 IC
1K
3
2
11
4
9
7
6
10
5
1
CD4017B
13
15
8
14
16
To Oscillator
9 VDC
drawn on 11.09.05 VE7BPO

The prototype "nite-rider" project with messy wiring. The holes for the LEDs were bored with a hand drill and it shows! The discrete transistor multivibrator can be seen behind the 4017 IC.

One of the four "nite-rider"

project chassis I built. After completion, this one was given to the son of VE7KPB. When drilling in a plastic chassis, I learned it is best to use a drill press set to a lower speed.

Sequentially Off LED Pulser

This circuit uses a series of transistors with an RC pair to pulse a string of LEDs.

This the favorite LED experiment I performed this summer. This flasher circuit is different in that it turns **off** alternate LEDs for about 1 second in sequence. When you connect this circuit to the 9 volt battery, all of the transistors are usually placed in saturation and therefore all the LEDs are **on**. Closing the switch on the base terminal of Q1 for a moment initiates the correct pulse sequence. The pulse initiates in Q1 which turns off the LED connected to the Q1 collector for about 1

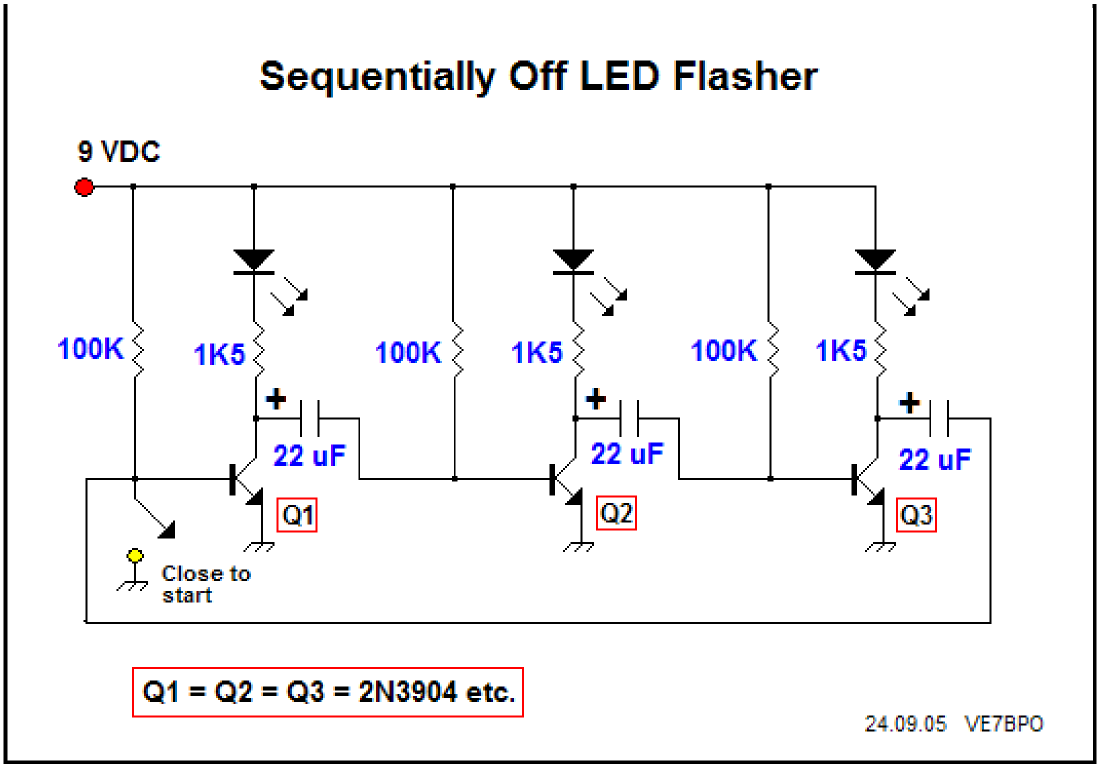
Sequentially Off LED Flasher
9 VDC
100K
1K5
100K
1K5
100K
1K5
22 uF
22 uF
22 uF
Q1
Q2
Q3
Close to start
Q1 = Q2 = Q3 = 2N3904 etc.
24.09.05 VE7BPO

second. When Q1 turns back on (goes into saturation), Q2 turns off. When Q2 turns back on then Q3 turns off and so on. The circuit is a closed loop and many more stages may be added.

You can experiment with different base resistor and coupling capacitor values to vary the speed of the LED string or to create a sense of randomness by varying each transistor's RC stage separately. This is a fun circuit!

Youtube link (not mine).

The prototype 3 transistor version. I just used a piece of wire to ground the Q1 base terminal and establish the correct pulse sequence after powering it up. For the LEDs, transistors, resistors and capacitors you can use whatever appropriate parts that you happen to have on hand. Current draw is less than 10 mA with a fresh 9 volt battery. Decrease the 1K5 current limiting resistor to 1K or so if you want brighter LEDs at the expense of more current draw. Do not operate this circuit above 9 volts unless you connect diodes from the transistor emitters to ground to prevent emitter-base breakdown.

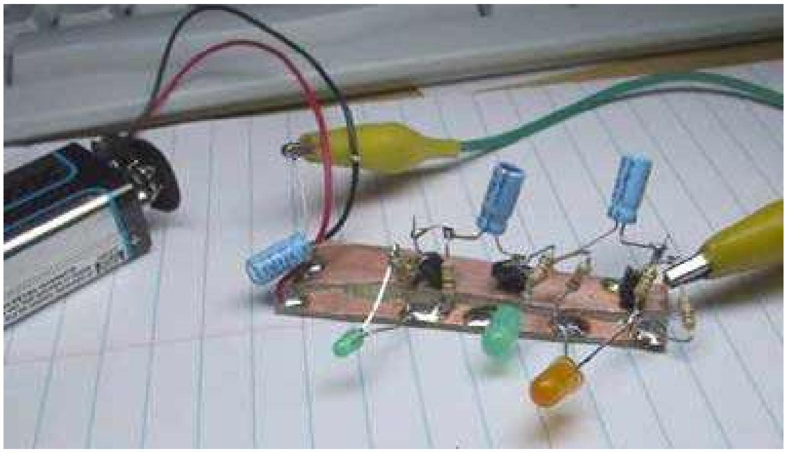

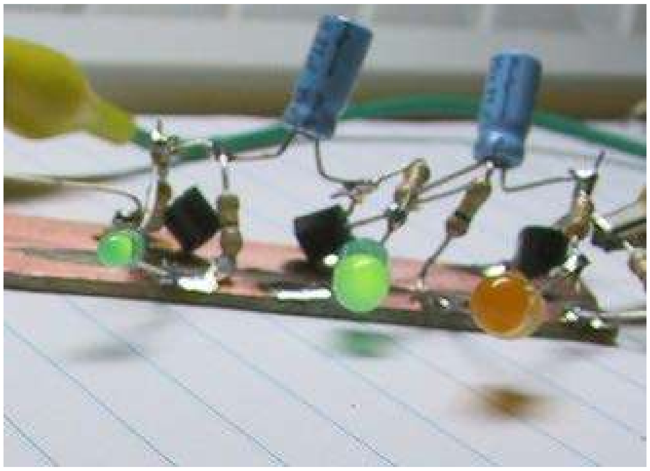

LED 1 and 2 are on and LED 3 is off at this moment in time.

Above and below photographs. This low current version has 9 LEDs connected in a chain and is powered by 3 volts. The 10th LED (extreme right hand side) is a flashing LED which is directly connected to the 3 volt supply and also uses a 1K current limiting resistor. Total peak current draw is only ~ 7 mA, yet it is still bright enough to see at night-time. The power supply is 2 D-cell batteries connected in series and then to the circuit by soldering wires directly onto the batteries with a 100 watt soldering iron.

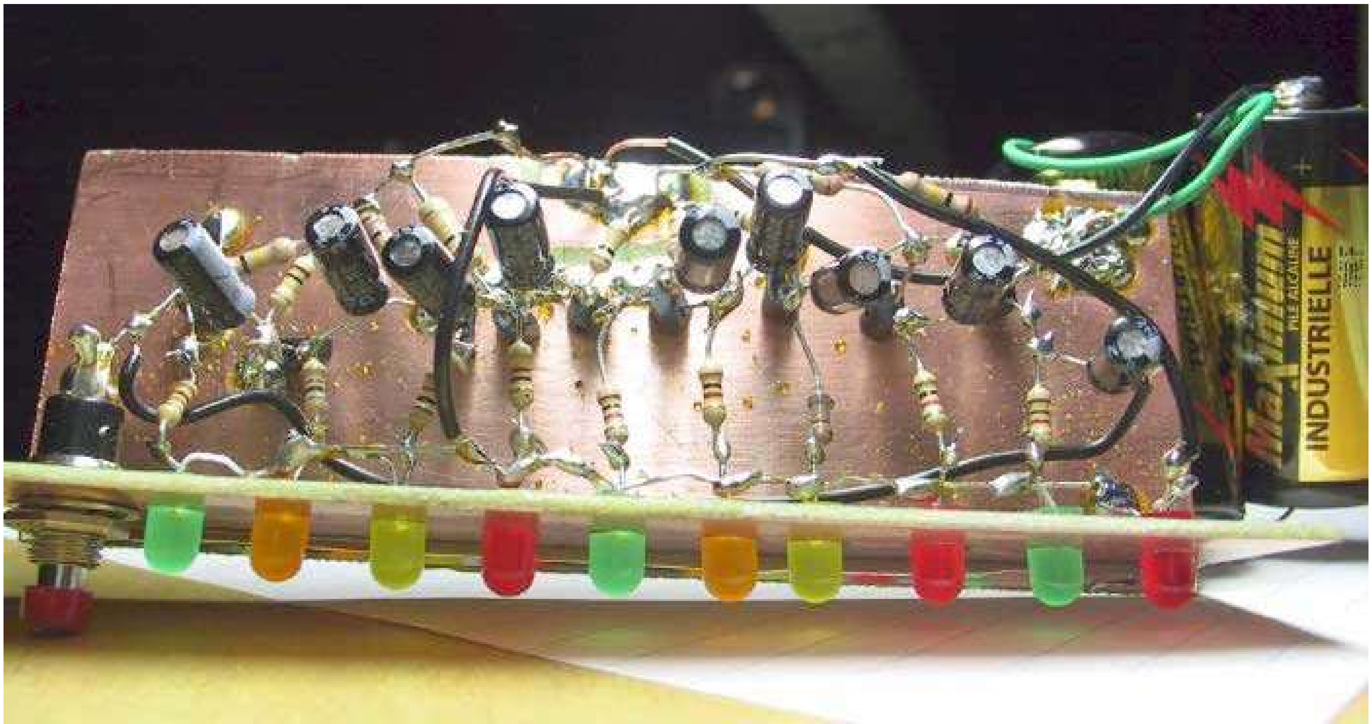

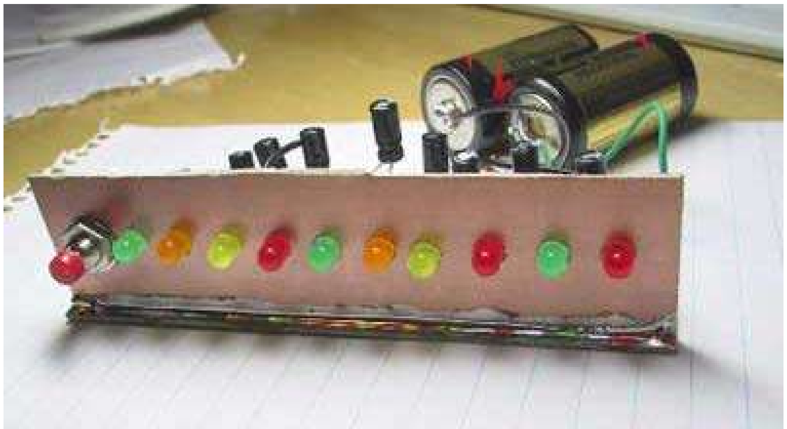

NoNot counting the 10th flashing LED, 5 of the 9 LEDs are on at any given moment. A sequential flash effect is noted (the state of each LED flip-flops and shifts over 1 position each flash). If you build this project with an even number of LEDs, the sequential effect is not seen. Half of the LEDS (spaced every other LED) are on and the other half are off at any moment. The same LEDs are lit or unlit each pulse. Thus the effect is more like a typical multivibrator LED flasher. This variable, even versus odd number of stages property makes the circuit quite versatile.

Conclusion

I hope that you have some fun experimenting with these and other circuits.

Amateur Radio Electronics Designer's Web Site

SWL Receiving Antenna Experiments

Introduction

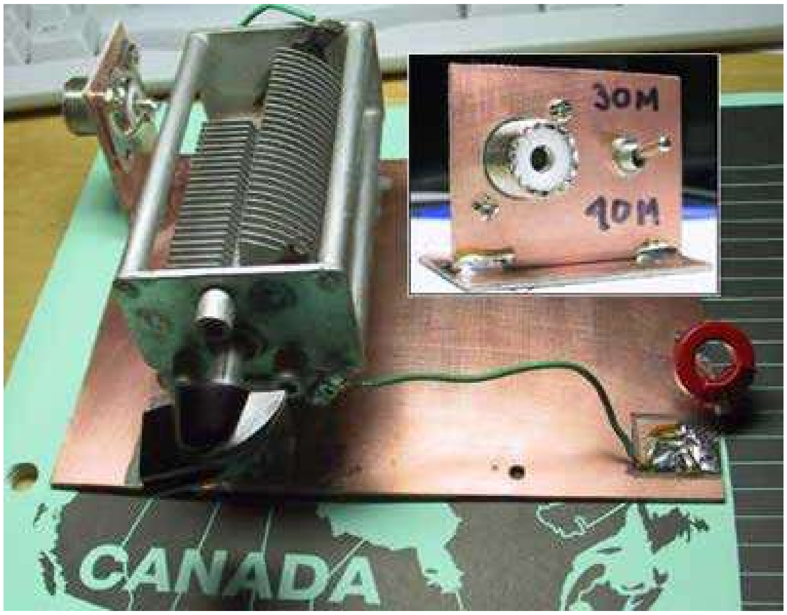

I have a lot to learn about SWL antennas. What follows are some brief experiments I performed in late October 2005. I have been experimenting with a half wavelength end-fed wire for use as a portable 40M band HAM transceiver (receive and transmit) antenna. This wire antenna is 67 feet long. End fed wires are very popular with those who pack a small portable transceiver when backpacking and camping. No feed line is required and the far end of the wire can be strung up using objects such as nearby trees or collapsible, portable poles. An elaborate ground system is not required. The return for the RF energy to ground might be grounding rod(s), short or long radial(s), or even just capacitively coupling to the local environment (including the operator!). Simple tuners are easily built to transform the high (thousands of ohms) wire impedance to the 50 ohms or so required by the transmitter.

I wanted to know if I could use this antenna as a tunable receive antenna for the the 30 and 31 Meter bands in addition to a tunable transmit/receiver antenna for my HAM radio work on 40 Meters. **What I verified is that tuning a multiband receive only antenna is not very practical.** When you tune a receive antenna you increase received noise and desired signals proportionally and therefore do not improve the signal to noise ratio in a meaningful way. Sometimes until you perform some experimentation, you don't really believe even good advice.

To the right: A computer simulation of a 40 Meter band end-fed Wire performed by W7ZOI on W7EL's EZnec program. The simulation was for 10.1 and 7.0 MHz with a 22 and 12 gauge wire. One 33 ft radial was used from the base of the 23.7 ft piece up 0.3 feet from the ground in this simulation. Z is impedance. Z and j are complex numbers used to represent the multi-dimensional quantities of the AC analysis of this antenna. In actual fact, j is an imaginary number. I suggest you might just ignore j unless your are well informed about impedance arithmetic.

Antenna Matching

A tuner can help match the impedance of the wire antenna and feed line (if used) to the input impedance of the receiver at a given frequency. This will result in more received signal and noise voltage to the receiver's input. HAM radio enthusiasts use antenna tuners to transform the impedance between the radio and the antenna tuner to 50 ohms to allow maximal output power from their transmitter. Non- amateur radio operators, can not use transmitters to

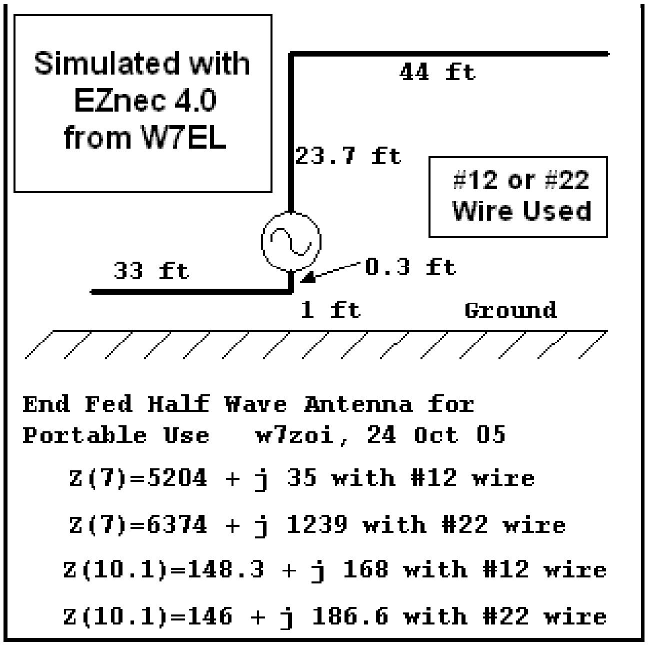
Simulated with EZnec 4.0 from W7EL
44 ft
23.7 ft
#12 or #22 Wire Used
33 ft
0.3 ft
1 ft
Ground
End Fed Half Wave Antenna for Portable Use w7zoi, 24 Oct 05
Z(7)=5204 + j 35 with #12 wire
Z(7)=6374 + j 1239 with #22 wire
Z(10.1)=148.3 + j 168 with #12 wire
Z(10.1)=146 + j 186.6 with #22 wire

match their antennas. This leaves either using receiver noise, S-Meter or an antenna analyzer such as the MFJ259. I just used my ears and S-meter. All of the tuners presented work as transmitter tuners as well. Any network used for transmitter work must be able to handle the output power of the transmitter final amplifier.

The Wire Antenna Experiments

I tried 3 different antenna tuners to see if I could tune an 18 gauge wire on 40-41 and 30-31 meters. My wire went from my computer room in the basement out a hole in the wall and sloped at ~ 50 degrees up to a rope tied to a tree in my backyard. The tip of the antenna is about 50 feet (15 meters) high. A 10 gauge insulated ground wire also passes from the computer room outside under the back lawn. It is "earthed" to two, 2 meter copper grounding rods hammered into the ground. No direct connection to the house ground system and the outside antenna grounding system should be made as this may result in increased receiver noise.

Antenna Tuner 1

The schematic on the left below, is very popular with HAMS who use it to to tune monoband end-fed wires. It is very simple and works reasonably well. Although the capacitor was able to resonate the 35 turns inductor, the T1 turns ratio was wrong and reducing it to 28:3 was required to get the maximum receiver signal in my experiments on the 40 and 41 meter bands. The alternative was to shorten or lengthen the wire antenna which is not very practical as it meant repeatedly climbing a tree.

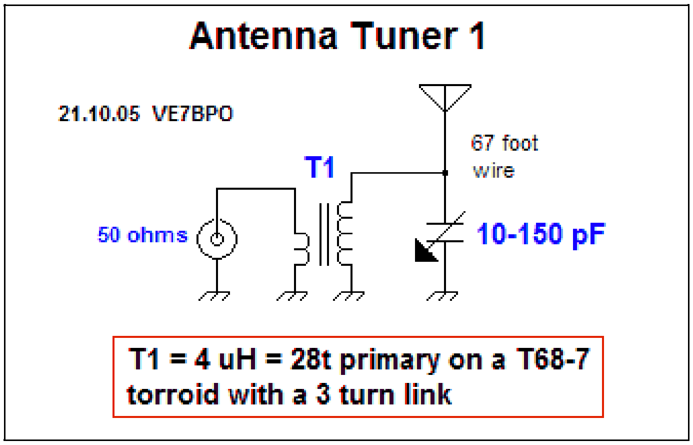
Antenna Tuner 1
21.10.05 VE7BPO
T1
67 foot wire
50 ohms
10-150 pF
T1 = 4 uH = 28t primary on a T68-7 torroid with a 3 turn link

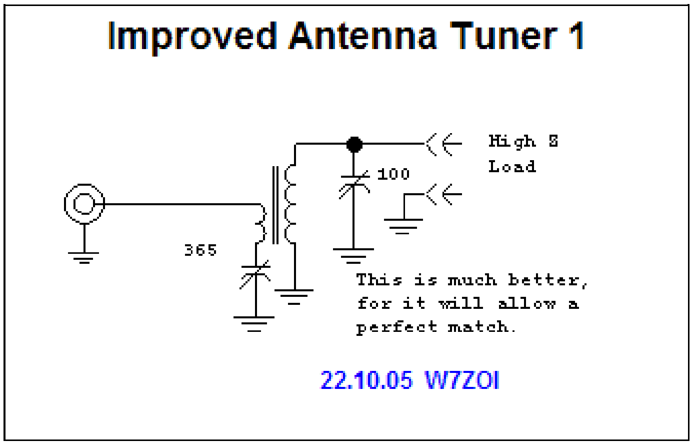
Improved Antenna Tuner 1
High Z Load
100
365
This is much better, for it will allow a perfect match.
22.10.05 W7ZOI

Above right: An improved version of the Antenna Tuner 1 schematic from W7ZOI. This tuner has 2 user "tweaking" adjustments much like most modern commercial antenna tuners (which typically also have a band changing adjustment).
I did not build this version.

In addition, this tuner would not tune the 67 foot wire on the 30 or 31 meters band. This was no surprise. When I turned the variable capacitor, I noticed some change in received signal, but not much. The signal strength was very poor and I could not hear much of anything.

Below. Two built up views of the Antenna Tuner 1 schematic. I soldered the antenna wire to the circuit in the isolated area connecting the inductor and capacitor. The outside antenna ground wire was soldered to the large copper ground plane.

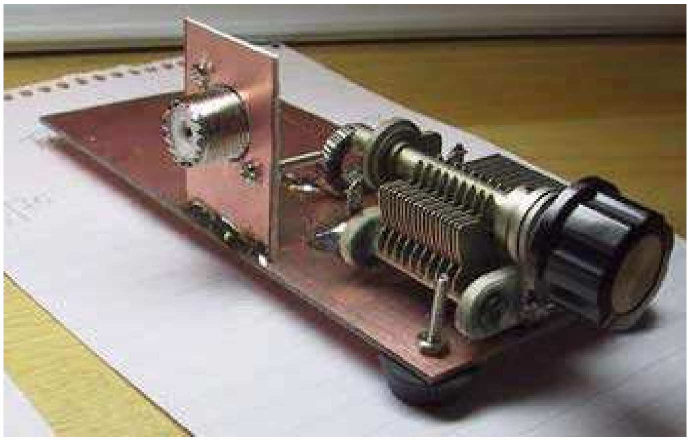

Antenna Tuner 2

This was the next tuner I built. Antenna tuner 2 tuned very sharply on the 40 and 41 Meter bands. It is designed to match a high impedance antenna, so it could not match the medium impedance (~150 ohms) wire to the 30 and 31 meter bands very well at all. Since ~ 150 ohms is fairly close to my receivers 50 ohm input impedance, I just connected the wire antenna directly to my receiver. The received noise and signals were then much stronger than those with the Antenna Tuner 2 network in the circuit on 30-31 meters.

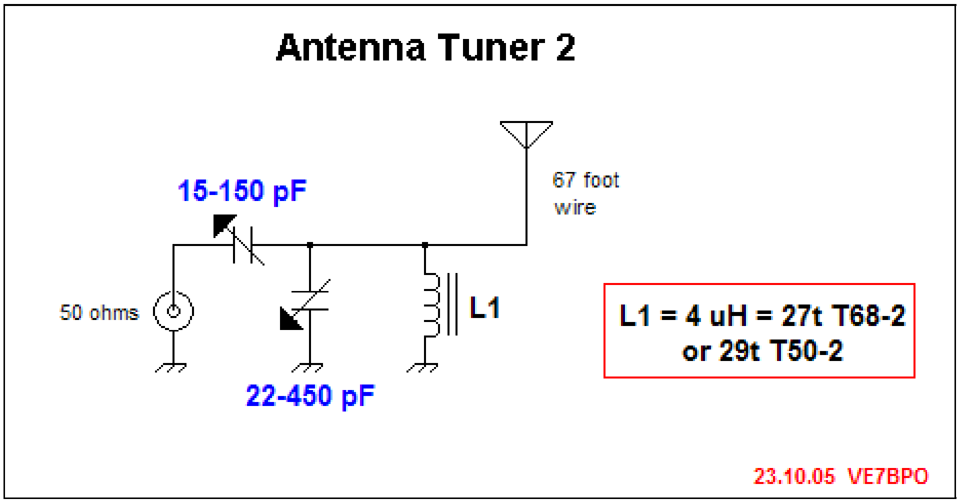

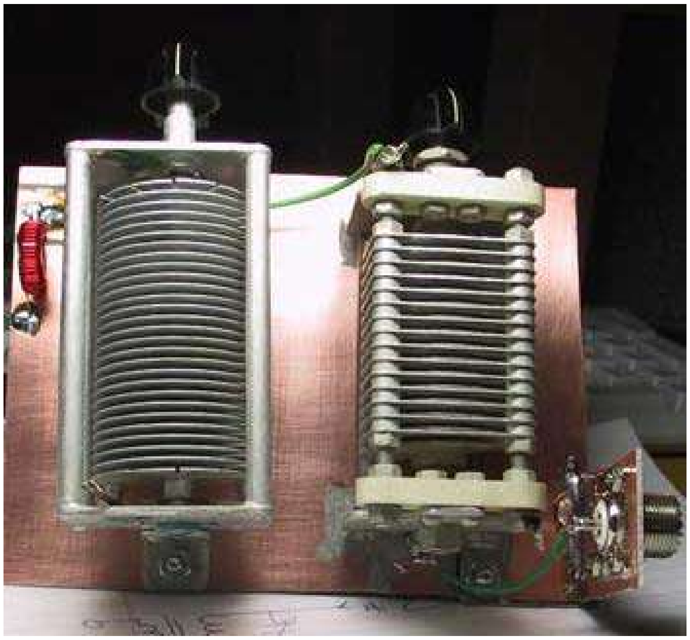

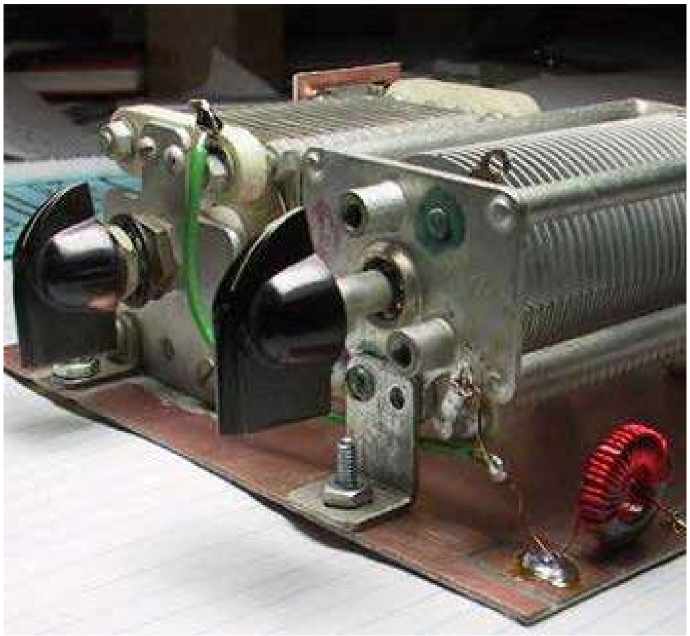

Above: Two constructed views of the Antenna Tuner 2 schematic. The variable capacitors were bought at a HAM festival in 1992 for 2 dollars each.

Antenna Tuner 3

The next antenna tuner topology I tried was the familiar L network. Circuit **A** is configured in a shunt L (inductor to ground) and series C (capacitor) and is a high pass L network. Circuit **B** is configured in a series L and shunt C and is a low pass L network.

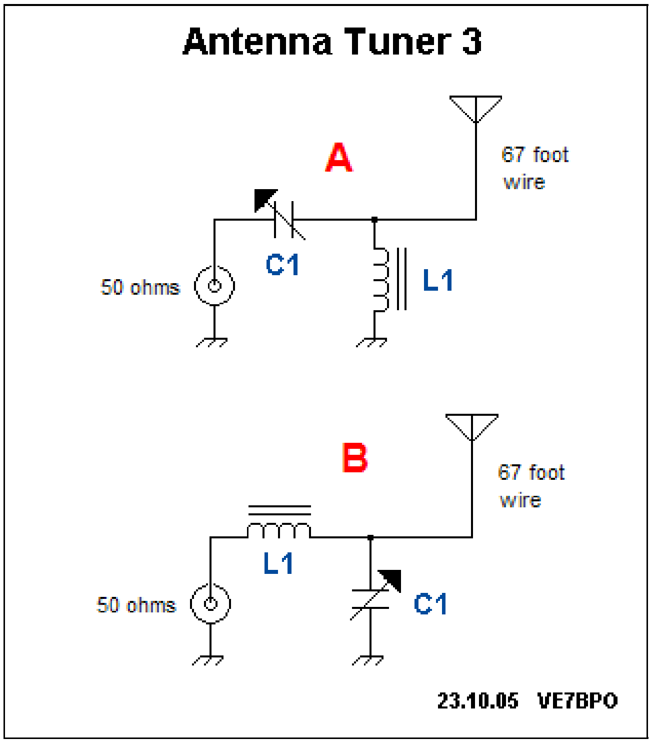
Antenna Tuner 3
A
67 foot wire
C1
50 ohms
L1
B
67 foot wire
L1
50 ohms
C1
23.10.05 VE7BPO

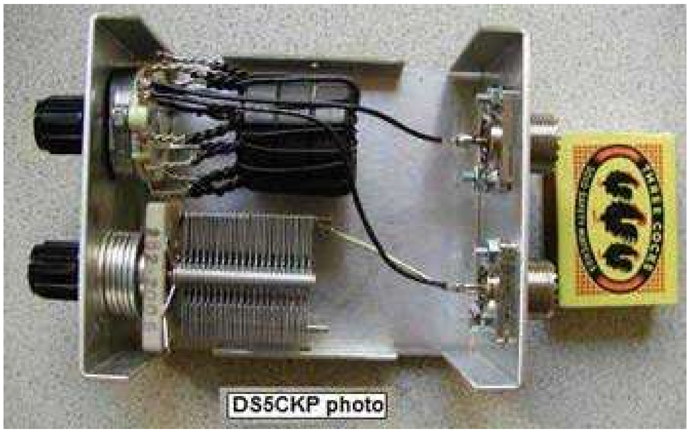
DS5CKP photo

L networks (especially the low pass form) are very popular as random wire tuners. MFJ sells an excellent version as the model MFJ-16010. A photograph of this tuner is shown below right. The above left photograph was taken by **DS5CKP** who also sells a random wire antenna tuner at: http://user.chollian.net/~cyberline/ckptuner.htm

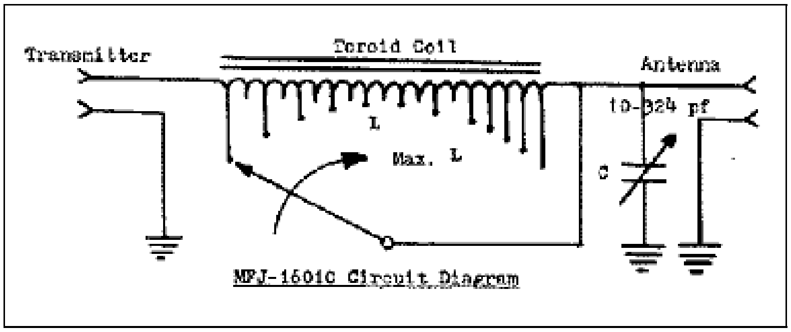
Transmitter
Toroid Coil
Antenna
10-324 pf
L
Max. L
C
MFJ-16010 Circuit Diagram

INDUCTANCE
CAPACITANCE
MFJ ANTENNA TUNER
Model MFJ-16010

Above. MFJ web publishes their manual including the schematic for the MFJ-16010 random wire antenna tuner. The inductor is actually wound on 3 stacked (probably ferrite) cores which are tapped. The taps are connected to a front panel mounted 12 position switch. This allows coverage from 2-30 MHz. As a simple experiment, I tried stacking 2 and then 3 FT-50-61 ferrite cores and was able to get a wind range of inductances from the 6 taps I made on my test inductor.

Numerous examples of the L network antenna tuner can be found on the web and in print including the 2006 A.R.R.L. handbook. I decided to try the high pass L network topology to experiment with.

I built part **A** of the Antenna Tuner 3 schematic with shunt L and series C. On the 40-41 meter bands I tried 3 different inductors; 4 uH, 2.1 uH and 1.3 uH. I tuned the network to get the greatest receive noise and S meter reading and measured the variable capacitor value. At 7.30 MHz, the capacitance values were 86 pF, 141 pF and 182 pF respectively. Although non-critical, I settled on a 2.6 uH inductor (23 turns on a T50-2 powdered iron torroid) so I could use a junk-box 10-150 pF variable capacitor to resonate it. Later I decided just a trimmer capacitor might do. You need around 125 pF at 7.30 MHz to tune the network with a 2.6 uH inductor (to give you a ballpark C value to start with).

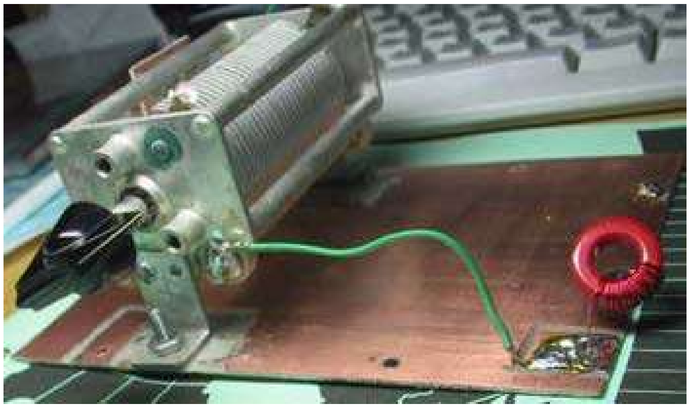

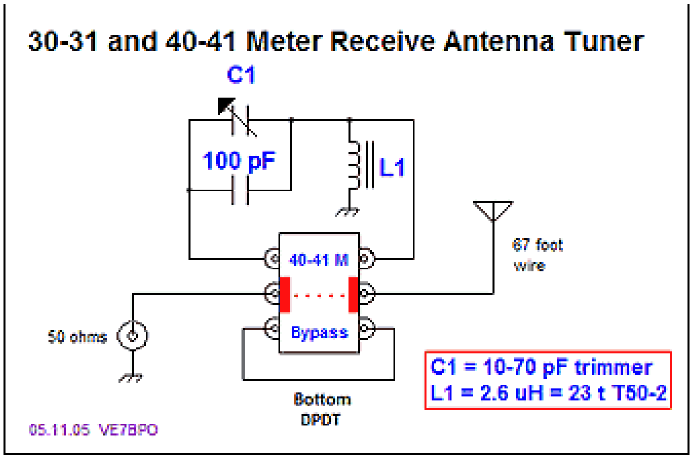
30-31 and 40-41 Meter Receive Antenna Tuner
C1
100 pF
L1
67 foot wire
40-41 M
Bypass
50 ohms
Bottom DPDT
C1 = 10-70 pF trimmer
L1 = 2.6 uH = 23 t T50-2
05.11.05 VE7BPO

Schematic to the right: A very small receiver tuner that allowed the L network to be switched in and out of the antenna path was constructed. The L network is tuned on 40-41 meters via a small trimmer cap on top of the double pole, double throw switch. A small plastic alignment screw driver is used for signal peaking. An air variable capacitor would be much easier to tune as you move up and down these bands.

Below left: Front view. Technically, the switch label should indicate **40M** and **bypass** as the bypassed antenna could be used on any band.

Below right: Rear view The trimmer cap can be seen on top of the switch.

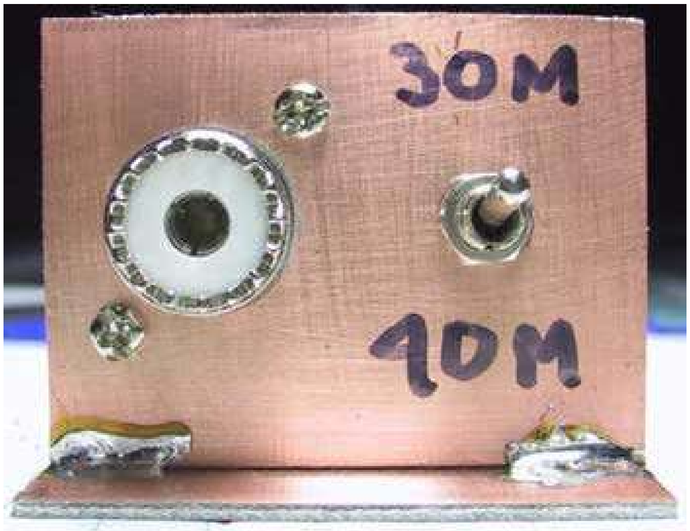

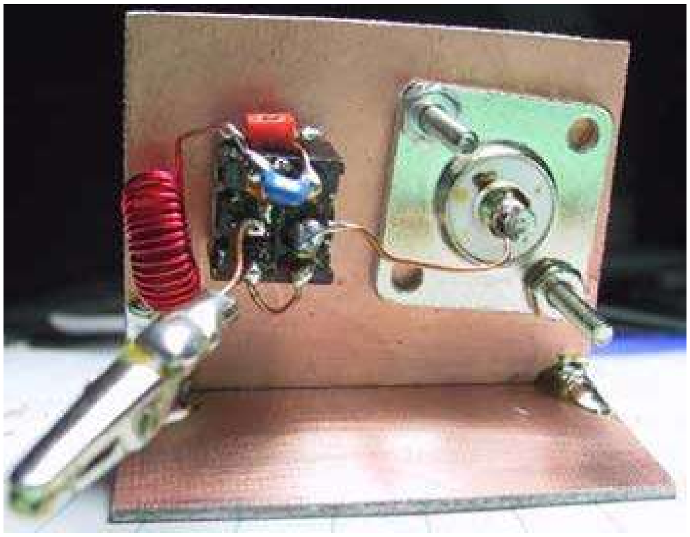

Below left: Side view, the antenna wire is held via an alligator clip. This allows me to unclip and ground the antenna when it is not in use. The antenna ground wire is soldered to the copper ground plane.

Below right: A T6-8 core wound with with 24 gauge wire. I used larger powered iron torroids to wind my inductors, however, they can also be wound on ferrite cores or be air-wound. The core size, wire gauge and type of inductor used can affect the selectivity and insertion loss of the tuner network, however for practical receiving purposes it is not a major concern.

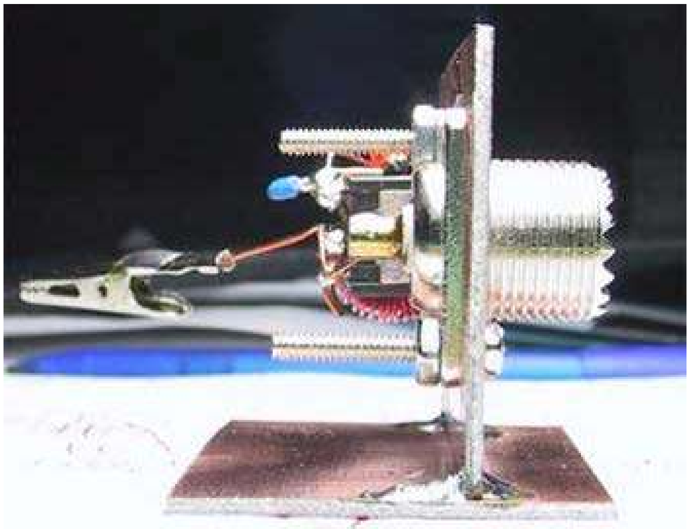

Conclusion

The final experiment allow me to easily switch between a matched and direct 67 foot wire antenna on 40 and 41 meters. I performed several listening tests and generally agree with those who say receive antenna tuners offer little to no improvement in signal readability. The only advantages I can think of for matching a receive antenna to a reasonably quiet and sensitive receiver are:

1. There may be some improvement in the receiver front end filter function as these filters are designed to have a specific input impedance.
2. Certain balanced mixers may function better with the correct impedance on their RF port.
3. The tuner itself (depending on design) may marginally improve the front end selectivity of the receiver it is connected to.

I do like the noise roar and louder signals with the matched antenna on 40 and 41 meters, although this is totally subjective. I think the reason for this is that as a HAM radio operator who always matches the antenna for any band I am on, I am used to louder signals and noise levels. I also spent most of my first 10 years of HAM radio operation on 80 and 160 meters which are relatively noisy bands and have been conditioned so that noise is "normal".

On 30-31 meters the bypassed antenna worked quite well and the L network can be switched in as an attenuator. Perhaps you might build up a tuner and try for yourself!

Although they are simple, low cost and easy to put up, it is likely unwise to use an end-fed wire without coaxial feed line as a receiving antenna. The ground wire is part of the antenna system and easily picks up household generated noise which will present to your receivers input. The time honored and easiest methods to reduce receiver noise are to get your antenna away from the house and other noise sources, use buried coaxial cable feed line to the house and directly earth ground the shield of the feed line with stakes. Antenna/ feed line "link coupling" by a transformer may also reduce noise (especially if the antenna system is balanced) and this topic begs further study.

Suggested Links

There are some fabulous web sites on the topic of SWL antennas and reducing receiver noise. I suggest these 5. Try a Google search for more.

http://www.dxing.info/equipment/

http://www.hard-core-dx.com/nordicdx/antenna/feed/feed1.html

http://www.aa5tb.com/efha.html

http://www.nyx.net/~dgrunber/

As usual, I learned a great deal from the process of experimenting. I look forward to spring when some more antenna experiments can be performed.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Experiments with JFET Biasing

The most common way of biasing a Junction Field Effect Transistor (JFET) is with a source resistor. This method, shown in Fig 1 below, has the advantage of offering negative feedback to stabilize the bias conditions. This is the same thing that happens when a bipolar transistor uses an emitter resistor. Self bias can be used as a method to evaluate a JFET to determine the critical parameters that describe it: Idss and Vp. These are discussed in Chapter 2 of **Experimental Methods in RF Design** and many other places.

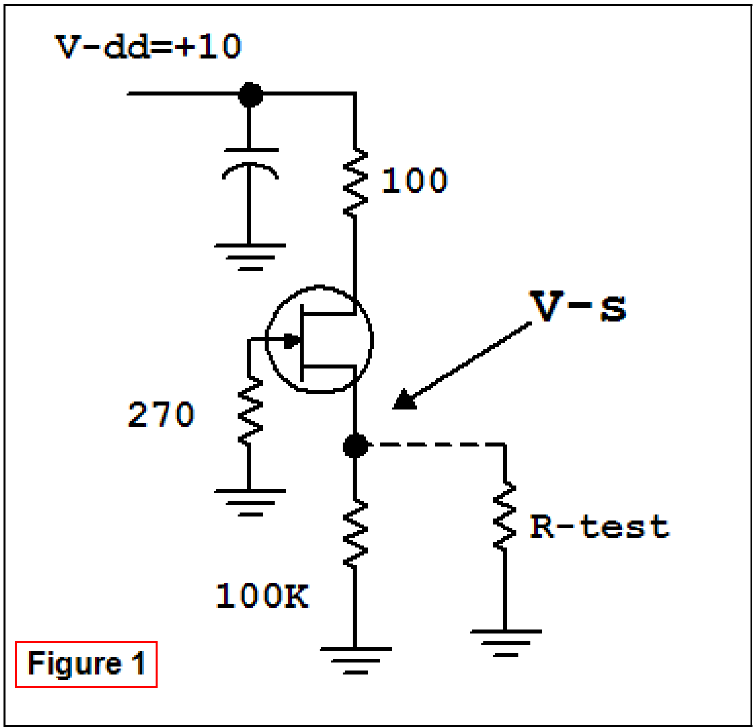

Figure 1

The method used in our experiment is to set up the FET of interest in a test fixture with a power supply, bypass capacitor, resistors in the drain and gate to suppress parasitic oscillations, and a handful of extra resistors, R-test, that can be paralleled with an existing 100K source resistor. A digital volt meter (DVM) is the basis for the measurements. We begin by using the DVM to measure the resistance of our test resistors, for the values will be used in calculations. The DVM is then attached to the FET source to measure the DC voltage. The first value we measure is with no attached R-test. The measured value will be very close to the FET pinchoff voltage.

The measurements we will perform infer drain current as a function of gate to source voltage. The physics of the FET support the model that there is no gate current so long as the gate is not forward biased with regard to the source. Hence, the drain current equals that in the source. We will measure the source current by measuring the voltage drop across the source resistor. The gate is at ground potential, for there is no gate current, so the gate to source voltage is just the negative of the source to ground voltage.

The resistors that I pulled from my stock for some measurements were marked as 22, 39, 68, 100, 150, 300, 510, 680, 1K, 2K, 3.3K, 6.8K, and 10K Ohm. The measured values are shown in attached figures. A systematic pattern was noticed with all of the measured resistances under the marked value, suggesting an error in the calibration of my DVM, a Fluke Model 73. All resistors were 2% carbon film 0.25 Watt. However, when I measured a 499 Ohm, 1% metal film resistor, it came up exactly at 500 Ohms. The differences between the measured values and those marked on the part were small enough that I neglected the details and used measured values for calculations.

The first FET I examined was a 2N5454, a common JFET that I had in my junk box. The source voltage was 3.26 with nothing but the 100K for source bias. I started my measurements with the largest resistor, 10K. The voltage dropped to 2.90 and was stable. I merely held the resistor in place rather than soldering it. The resistor was kept in place long enough to get a stable reading that I could record in my lab notebook. All results were of the same character until I got to the 300 Ohm resistor. At that point I started to notice a slight heating effect. The source voltage was 1.523, but slowly dropped to 1.518 volts. This behavior continued through the lower value resistors. The 22 Ohms produced 264 mV on the source that then dropped to 256 mV.

Later I examined a J310 JFET. This is a much larger area part than the 2N5454 with an Idss that is about three times larger. With the 22 Ohms in the test fixture, V-source went to 701 mV, but settled at 652 mV. The drain current was then 32 mA. With a 10 volt power supply, there was nearly 200 mW dissipated in the FET. This is within ratings, but high enough to produce heating. Operation at higher voltages and at Idss would further tax the part. One must take care when doing these measurements to be sure that the source voltage is observed quickly.

Attached are the MathCad documents that I used to examine the data. A spread sheet such as Excel could be used, but I prefer the graphics of MathCad. The second page for the 2N5454 shows a graph for the observed data as well as a calculated one. The two FET parameters for the 2N5454 were varied to obtain a good correspondence between the two. The part had Idss=15 mA with Vp=-3.5 volts. This is similar to the popular MPF-102, but close to the high Idss extreme for that part.

The data presented for the J310 is more abbreviated with only two points shown. I picked the 22 Ohm and 1K source resistors. This still produced data that is very close to that obtained with many more data points.

My initial analysis suggested that we could characterize the FET by measuring the source voltage with 100K in place to approximately determine Vp, and to then short circuit the source through the mA scale on the DVM to obtain Idss. This is a reasonable start. However, the pinchoff will usually be a few percent more negative (for an N-channel depletion mode part). The long leads in the source also make me feel uncomfortable with regard to parasitic oscillations.

After the DC measurements were done, I thought it wise to look at the potential for oscillation. The J310 was in the test circuit at this time. The TO-92 J310s are parts that are well known for their propensity for oscillation, so I guessed that it would not be difficult to coax this one into such a mode. But this was not what I found. I eliminated the 100 Ohm drain resistor, but moved the FET close to the 0.1 uF bypass. This bypass is not a very good one for VHF and upward. No oscillation was seen. I then eliminated the gate resistor, replacing it with the normal gate lead. Still no oscillation. I eventually added a gate inductor and a parallel tuned circuit in the source. The source bias resistor had a RF choke in series with it. I finally saw a robust VHF oscillation, but nothing else up through 1.5 GHz.

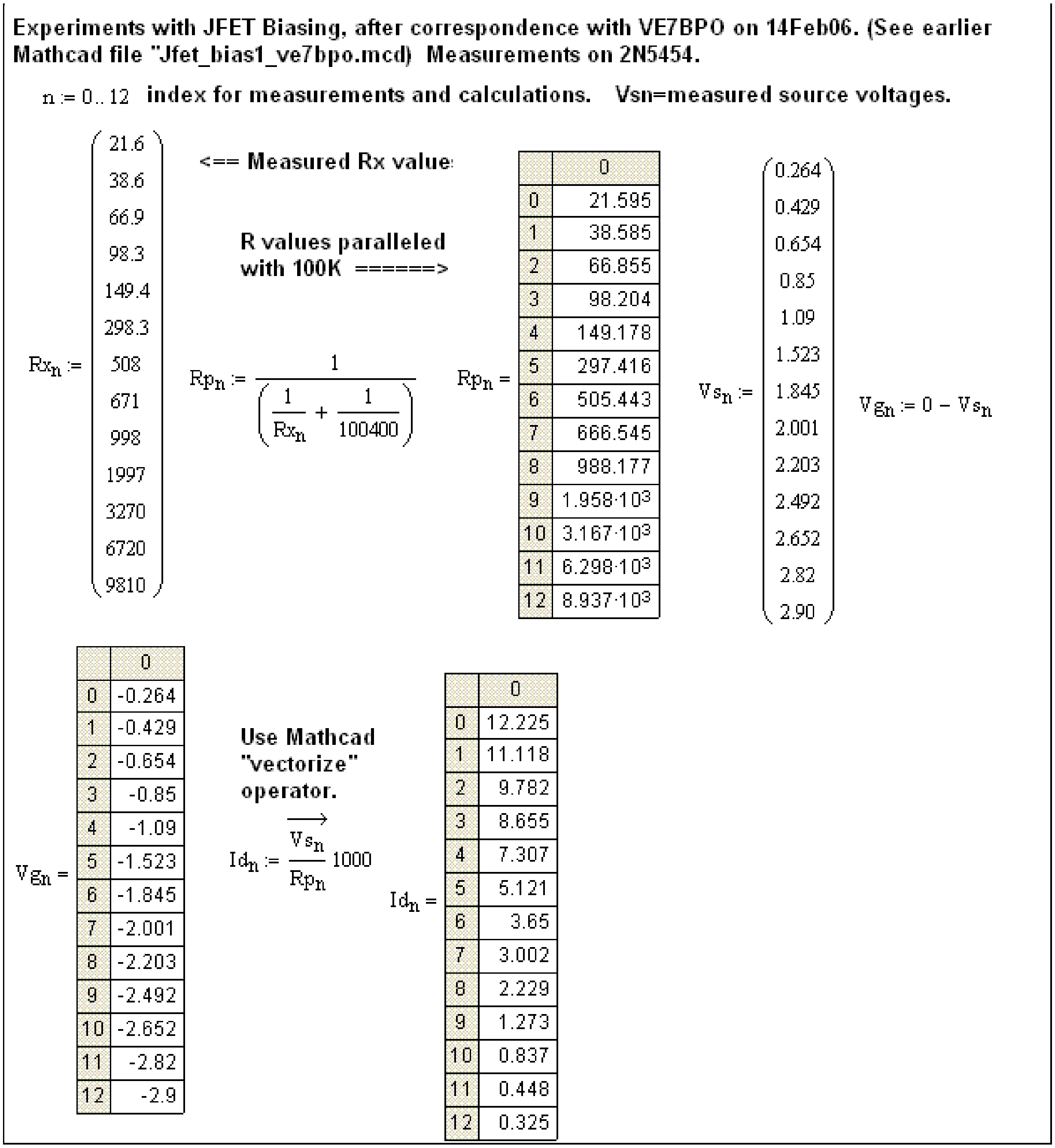

Experiments with JFET Biasing, after correspondence with VE7BPO on 14Feb06. (See earlier Mathcad file "Jfet_bias1_ve7bpo.mcd) Measurements on 2N5454.

$n := 0..12$ **index for measurements and calculations. Vsn=measured source voltages.**

$$Rx_n := \begin{pmatrix} 21.6 \\ 38.6 \\ 66.9 \\ 98.3 \\ 149.4 \\ 298.3 \\ 508 \\ 671 \\ 998 \\ 1997 \\ 3270 \\ 6720 \\ 9810 \end{pmatrix}$$

<== Measured Rx value

R values paralleled with 100K ======>

$$Rp_n := \frac{1}{\left(\frac{1}{Rx_n} + \frac{1}{100400}\right)}$$

$Rp_n =$

	0
0	21.595
1	38.585
2	66.855
3	98.204
4	149.178
5	297.416
6	505.443
7	666.545
8	988.177
9	$1.958 \cdot 10^3$
10	$3.167 \cdot 10^3$
11	$6.298 \cdot 10^3$
12	$8.937 \cdot 10^3$

$$Vs_n := \begin{pmatrix} 0.264 \\ 0.429 \\ 0.654 \\ 0.85 \\ 1.09 \\ 1.523 \\ 1.845 \\ 2.001 \\ 2.203 \\ 2.492 \\ 2.652 \\ 2.82 \\ 2.90 \end{pmatrix}$$

$Vg_n := 0 - Vs_n$

$Vg_n =$

	0
0	-0.264
1	-0.429
2	-0.654
3	-0.85
4	-1.09
5	-1.523
6	-1.845
7	-2.001
8	-2.203
9	-2.492
10	-2.652
11	-2.82
12	-2.9

Use Mathcad "vectorize" operator.

$$Id_n := \overrightarrow{\frac{Vs_n}{Rp_n}} 1000$$

$Id_n =$

	0
0	12.225
1	11.118
2	9.782
3	8.655
4	7.307
5	5.121
6	3.65
7	3.002
8	2.229
9	1.273
10	0.837
11	0.448
12	0.325

Let pinchoff be $u_p := -3.5$ **and** $I_{dss} := 15$

$$u := -3, -2.9 .. 0$$

$$I(u) := I_{dss} \cdot \left(1 - \frac{u}{u_p}\right)^2$$

Try various values for pinchoff and Idss to get a reasonable match.

Blue is measured data while red shows calculated behavior.

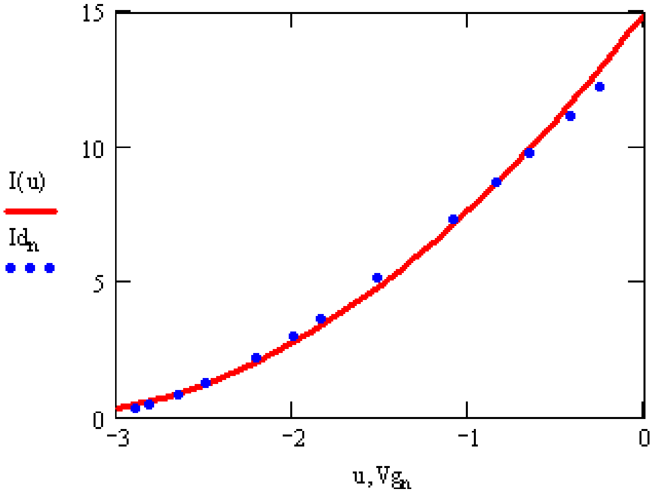

Experimental Note: During the measurements, I noticed a definite thermal change in Idss when dissipation was high. This occurred at higher currents, with low source bias resistors. The Vg values I used in my data were the first ones recorded with the DVM. The most extreme case was with the 22 Ohm source R. The source V started at 264 mV and went down to 256 mV. The effect was more dramatic with a larger part, a J310. That part, when operated with a 22 Ohm R, started with 701 mV and dropped to 652 mV.

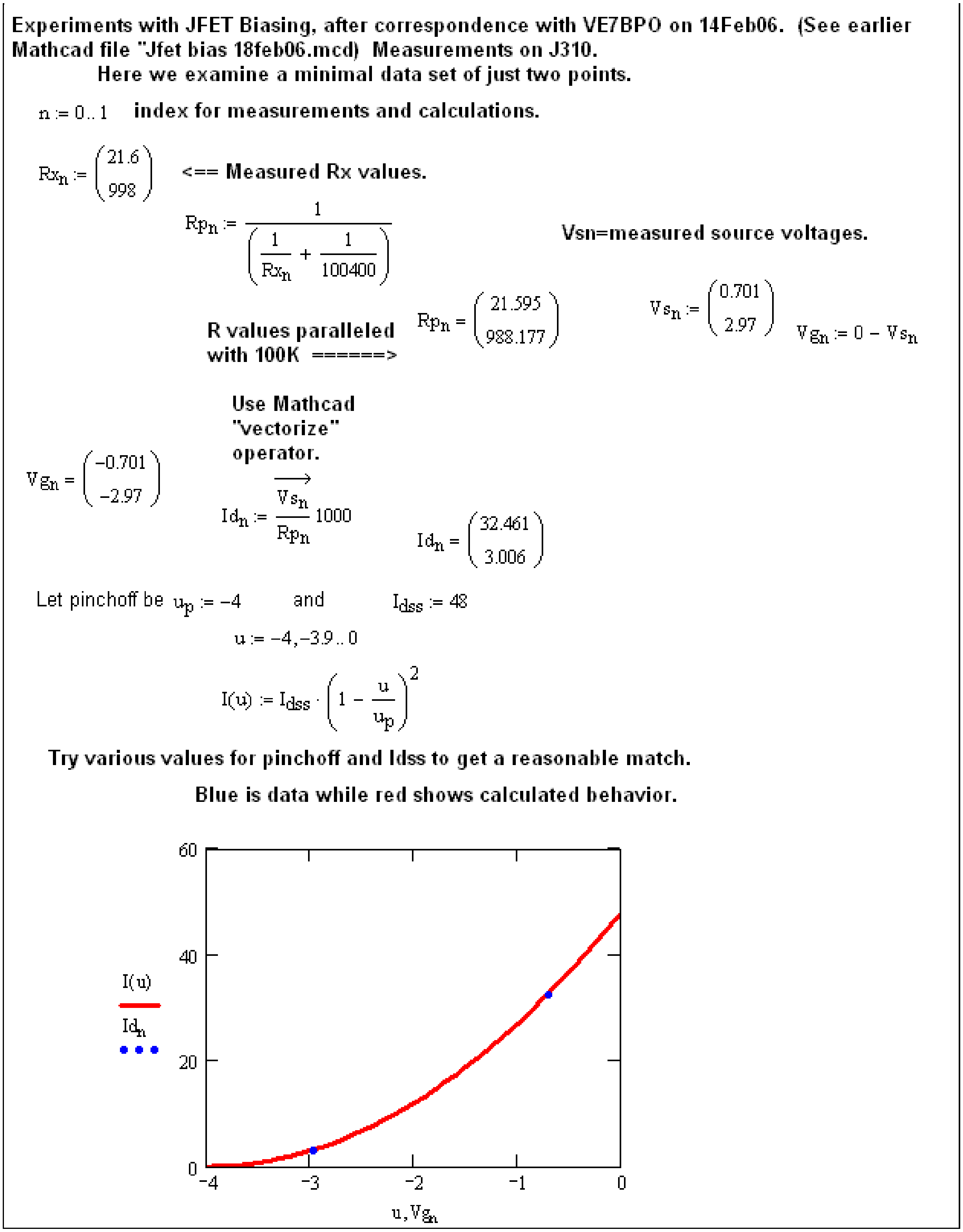

Experiments with JFET Biasing, after correspondence with VE7BPO on 14Feb06. (See earlier Mathcad file "Jfet bias 18feb06.mcd) Measurements on J310.

Here we examine a minimal data set of just two points.

$n := 0 .. 1$ **index for measurements and calculations.**

$Rx_n := \begin{pmatrix} 21.6 \\ 998 \end{pmatrix}$ **<== Measured Rx values.**

$$Rp_n := \frac{1}{\left(\frac{1}{Rx_n} + \frac{1}{100400}\right)}$$

Vsn=measured source voltages.

R values paralleled with 100K ======> $Rp_n = \begin{pmatrix} 21.595 \\ 988.177 \end{pmatrix}$

$Vs_n := \begin{pmatrix} 0.701 \\ 2.97 \end{pmatrix}$ $Vg_n := 0 - Vs_n$

$Vg_n = \begin{pmatrix} -0.701 \\ -2.97 \end{pmatrix}$

Use Mathcad "vectorize" operator.

$Id_n := \overrightarrow{\frac{Vs_n}{Rp_n} 1000}$ $Id_n = \begin{pmatrix} 32.461 \\ 3.006 \end{pmatrix}$

Let pinchoff be $u_p := -4$ and $I_{dss} := 48$

$u := -4, -3.9 .. 0$

$$I(u) := I_{dss} \cdot \left(1 - \frac{u}{u_p}\right)^2$$

Try various values for pinchoff and Idss to get a reasonable match.

Blue is data while red shows calculated behavior.

18 Feb, 2006 by W7ZOI

Many thanks to Wes, W7ZOI for this contribution.

10 MHz WWV Receiver Experiments

New feature: Click on a schematic to load it into a separate browser page for printing

Introduction

For nearly a year, I have been trying to develop a tuned radio frequency (TRF) 10 MHz, WWV, AM receiver. My initial RF stages were common emitter or common source stages with tuned input and output. Despite careful layout, parasitic oscillations plagued these designs and they were discarded. Later, I discovered that only tuning the input of RF stages reduced this tendency towards instability and still provided reasonable selectivity. Different detectors were also tried and evaluated.

A simple receiver that sounds great and is fun to build and experiment with follows. My special thanks to Wes, W7ZOI for performing many of the simulations and providing suggestions which kept me going.

WWV Audio Files from Sept 26, 2006

WWV File 1

WWV File 2

WWV File 3

The audio was digitally recorded using an electret condenser microphone held 3 cm away from the receiver speaker. The files were compressed using the WMA format.

Supplemental Web Page added June 29, 2007

Receiver Front End

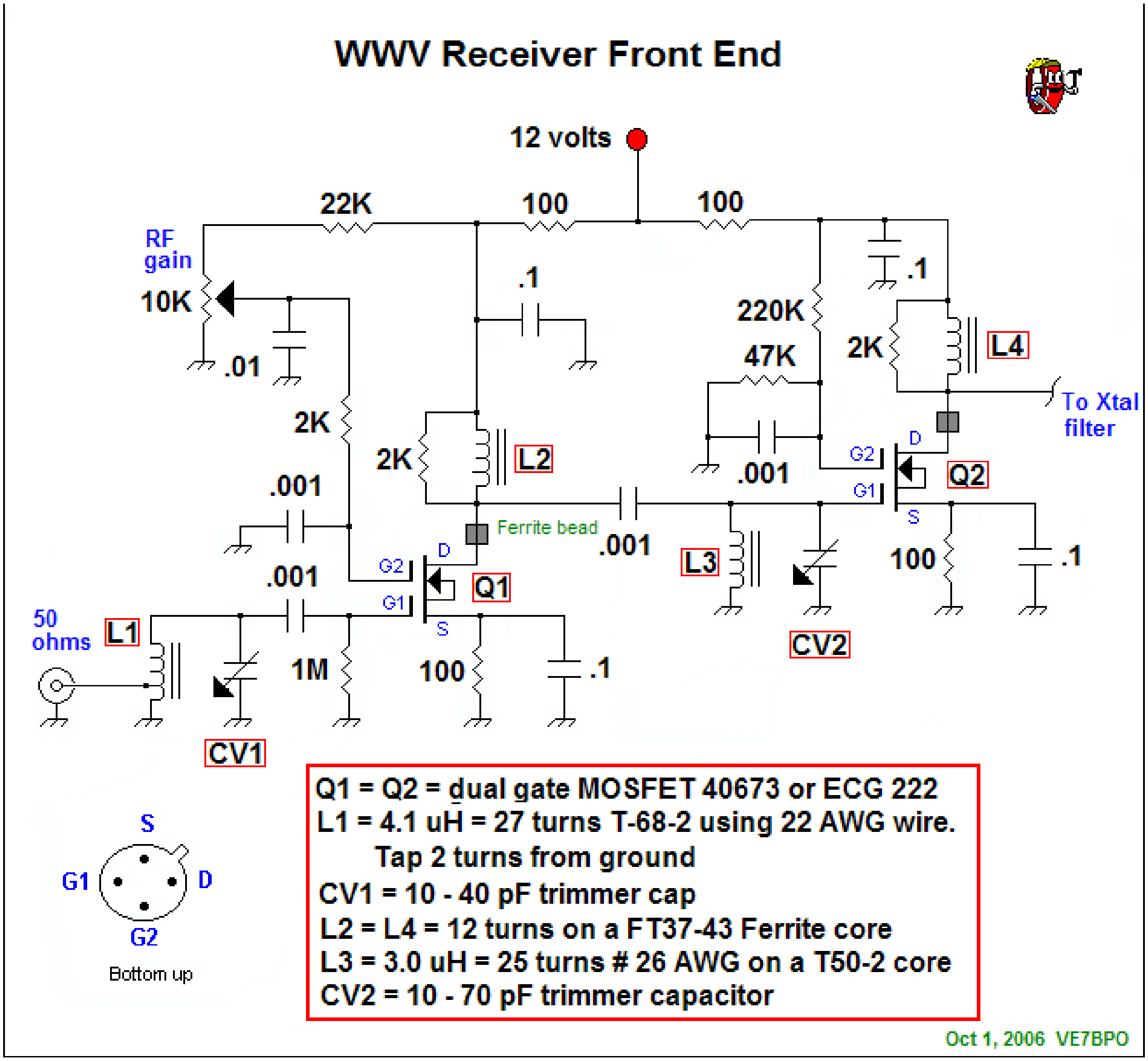

WWV Receiver Front End

Q1 = Q2 = dual gate MOSFET 40673 or ECG 222
L1 = 4.1 uH = 27 turns T-68-2 using 22 AWG wire.
Tap 2 turns from ground
CV1 = 10 - 40 pF trimmer cap
L2 = L4 = 12 turns on a FT37-43 Ferrite core
L3 = 3.0 uH = 25 turns # 26 AWG on a T50-2 core
CV2 = 10 - 70 pF trimmer capacitor

Oct 1, 2006 VE7BPO

Above schematic. The receiver front end has just 1 single-pole filter. For even greater selectivity (but greater insertion loss), consider moving the L1 tap to 1 turn from ground. My receiver was connected to a 80 meter dipole via an antenna tuner. The antenna tuner provided additional selectivity. No local broadcast band (BCB) signals were heard when the chassis lid was tightened on. You may require additional RF high pass or band pass filtering in your location.

The RF gain control is very basic and only the first 1/3 of the 10K pot is used to go from minimal to maximal gain. Modifications to allow more precise variation in RF gain for dual gate MOSFETs or cascode JFETS are shown in EMRFD. The method shown works fine. For the most part, I keep it set to minimal gain. Using higher gain than necessary, increases receiver noise and may overdrive the detector.

TRF receivers require careful layout. A piece of wire greater than 2-3 cm between the stages may be enough to plague your receiver with local BCB interference depending on your layout and chassis integrity. For interstage connections that had greater than a 2 cm gap, shielded 50 ohm cable was used to prevent BCB interference.

Dual gate MOSFETs provide adequate gain and low noise. you might consider cascoding 2 JFETS for each RF stage if you cannot obtain them. titleernatively, bipolar feedback amplifiers may be used and examples are provided later on this web page.

XTal Filter and RF Amplifier

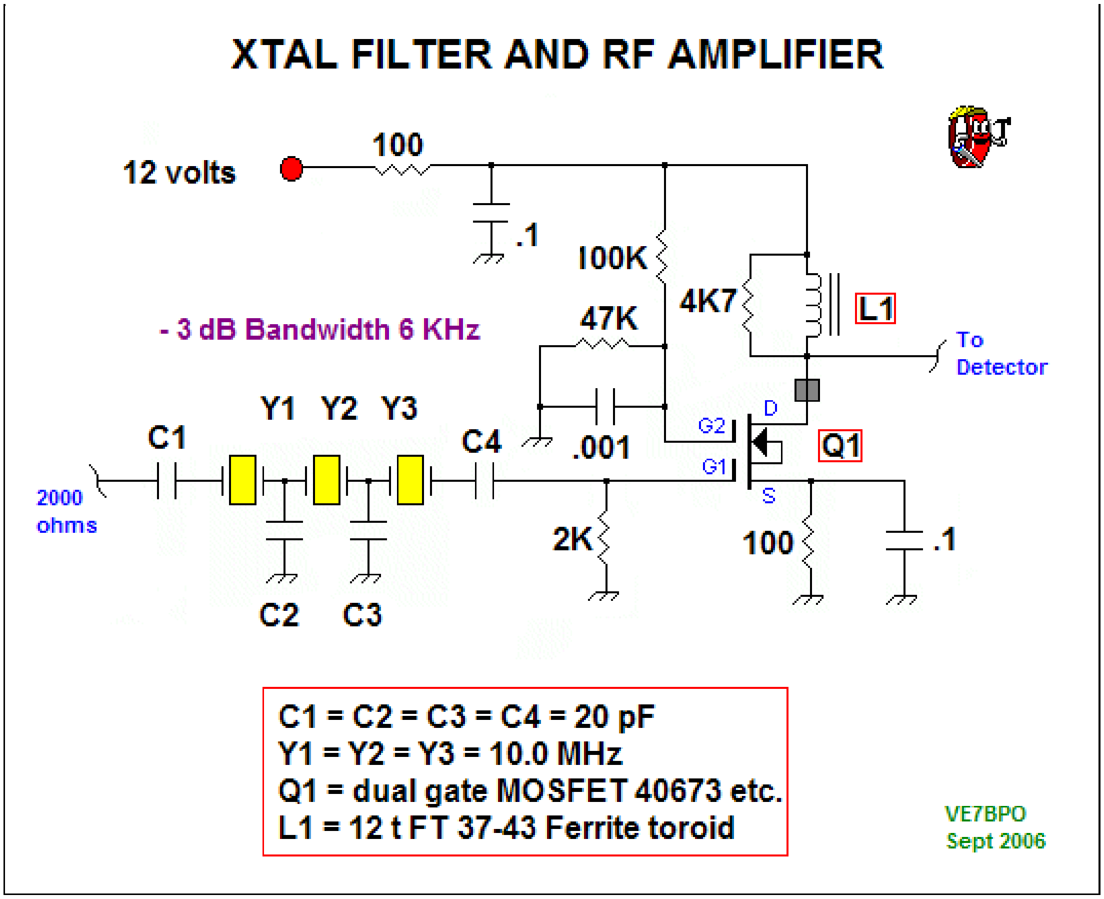
XTAL FILTER AND RF AMPLIFIER
12 volts
100
.1
100K
47K
4K7
L1
To Detector
- 3 dB Bandwidth 6 KHz
Y1 Y2 Y3
C1
C4
.001
G2
D
G1
Q1
S
2000 ohms
2K
100
.1
C2 C3
VE7BPO
Sept 2006

C1 = C2 = C3 = C4 = 20 pF
Y1 = Y2 = Y3 = 10.0 MHz
Q1 = dual gate MOSFET 40673 etc.
L1 = 12 t FT 37-43 Ferrite toroid

Schematic: The Q2 output impedance is 2000 ohms to match the input impedance of the Cohn crystal filter. This filter was designed by Wes, W7ZOI. Matched, computer grade, 10 MHz crystals were used. Choose 10 MHz crystals that are marked for a 20 pF or 32 pF load capacitance if possible. Using a 10 MHz crystal oscillator, find 3 that are closest to one another in frequency.

You may substitute 2K2 resistors instead of the specified 2K with a slight penalty in pass band shape.

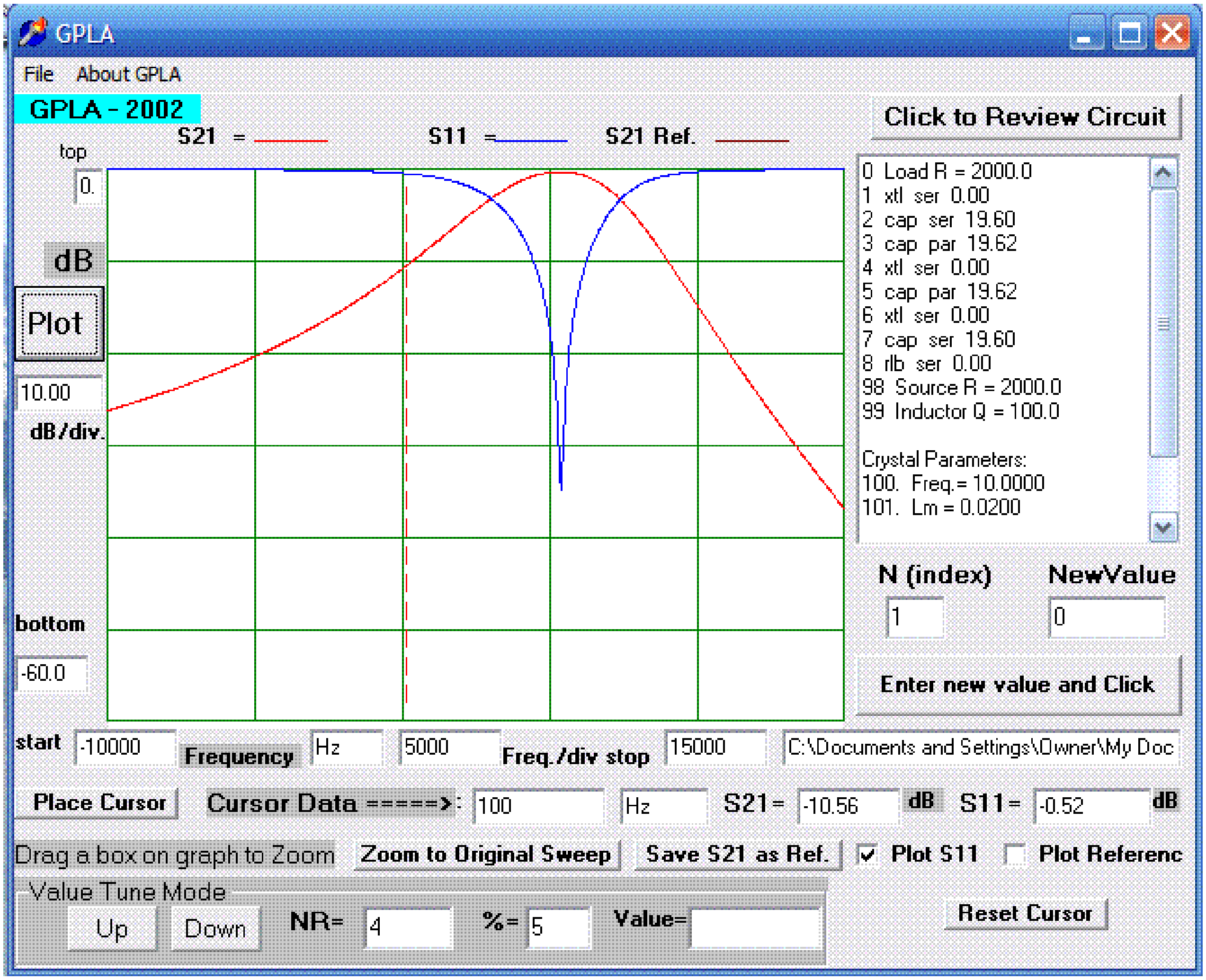
GPLA
File
About GPLA
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
dB
Plot
Enter new value and Click
Place Cursor
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Value Tune Mode
Up
Down
Reset Cursor

top 0.

10.00 dB/div.

bottom -60.0

start -10000 Frequency Hz 5000 Freq./div stop 15000 C:\Documents and Settings\Owner\My Doc

Cursor Data =====>: 100 Hz S21= -10.56 dB S11= -0.52 dB

☑ Plot S11 ☐ Plot Referenc

NR= 4 %= 5 Value=

0 Load R = 2000.0
1 xtl ser 0.00
2 cap ser 19.60
3 cap par 19.62
4 xtl ser 0.00
5 cap par 19.62
6 xtl ser 0.00
7 cap ser 19.60
8 rlb ser 0.00
98 Source R = 2000.0
99 Inductor Q = 100.0

Crystal Parameters:
100. Freq.= 10.0000
101. Lm = 0.0200

N (index) 1
NewValue 0

Above graphic. A simulation of the receiver crystal filter using GPLA, a program written by Wes, W7ZOI that comes with EMRFD. EMRFD is the major reference for this web site and I recommend that you add this book and companion software to your home library. The pass band is not symmetrical. It is mistuned for the lower pass band frequencies and would serve better as an upper sideband filter. Nevertheless, it works reasonably well and is simple to build and tolerant to component variation and match. At certain times, a very strong shortwave station at 9.985 MHz can be heard along with WWV. This usually occurs in the early evening when the WWV signal is not that strong at my location. For most of the day and night, whether WWV is present or not, very little interference has been detected.

Bypassing the crystal filter is an interesting experiment. As many as 5 stations were heard simultaneously and these varied as time passed. I heard Radio Vatican, Radio Habana and many other broadcasts during 1 evening. At one point I heard a station at 9.75 MHZ, WWV and a strong CW carrier at 10 .110 MHz!

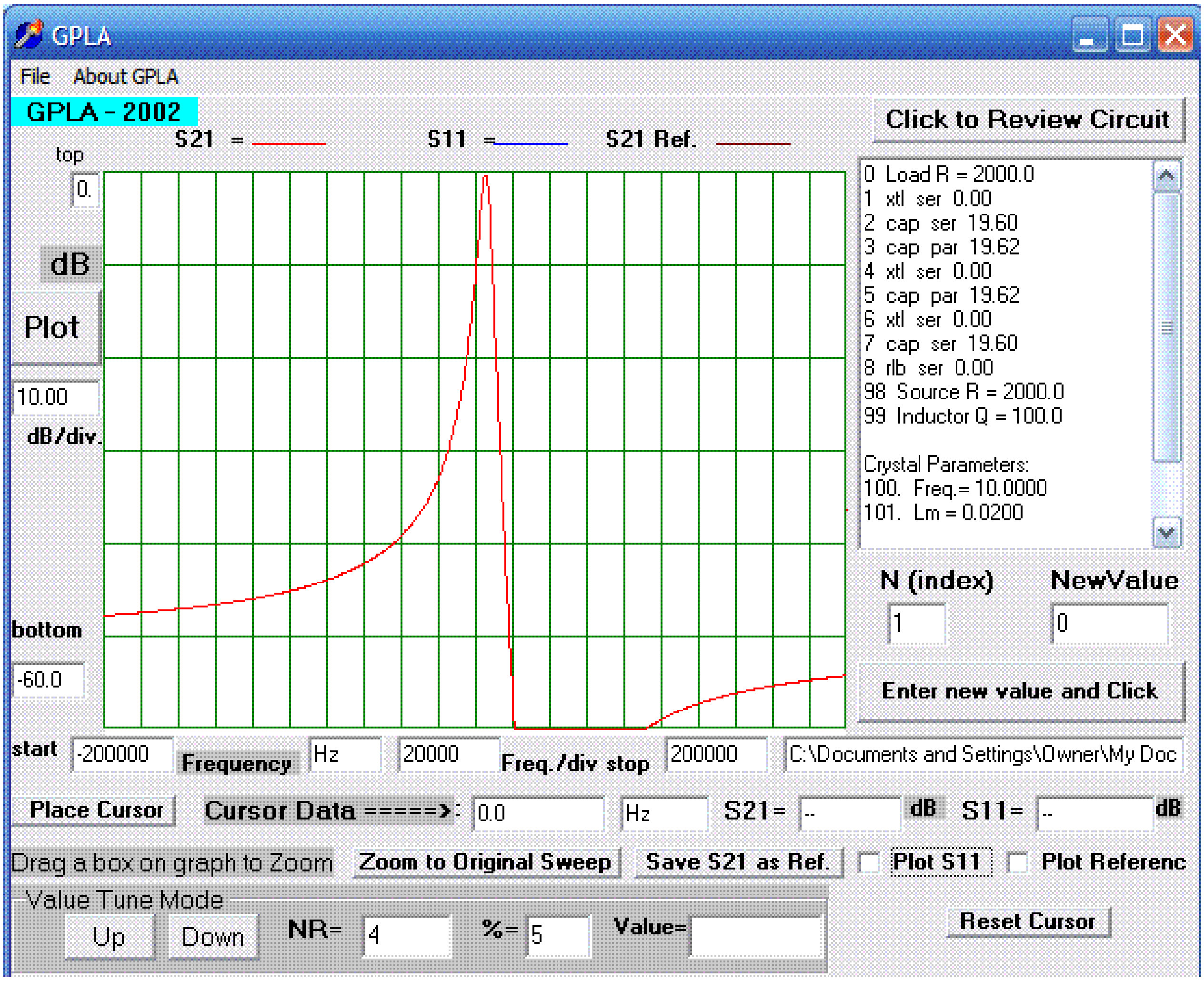
GPLA
File
About GPLA
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
dB
Plot
dB/div.
Frequency
Freq./div
Place Cursor
Cursor Data =====>:
dB
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
☐ Plot S11
☐ Plot Referenc
Value Tune Mode
Up
Down
Reset Cursor
Enter new value and Click

top: 0.

10.00 dB/div.

bottom: -60.0

0 Load R = 2000.0
1 xtl ser 0.00
2 cap ser 19.60
3 cap par 19.62
4 xtl ser 0.00
5 cap par 19.62
6 xtl ser 0.00
7 cap ser 19.60
8 rlb ser 0.00
98 Source R = 2000.0
99 Inductor Q = 100.0

Crystal Parameters:
100. Freq.= 10.0000
101. Lm = 0.0200

N (index): 1
NewValue: 0

start: -200000 Hz
Freq./div: 20000
stop: 200000
C:\Documents and Settings\Owner\My Doc

Cursor Data =====>: 0.0 Hz S21= -- dB S11= -- dB

NR= 4 %= 5 Value=

Above graphic. This is a sweep that goes from 200 kHz below to 200 kHz above 10 MHz to show the stop band response of this filter. This filter has a pretty decent response considering the low cost and effort involved.

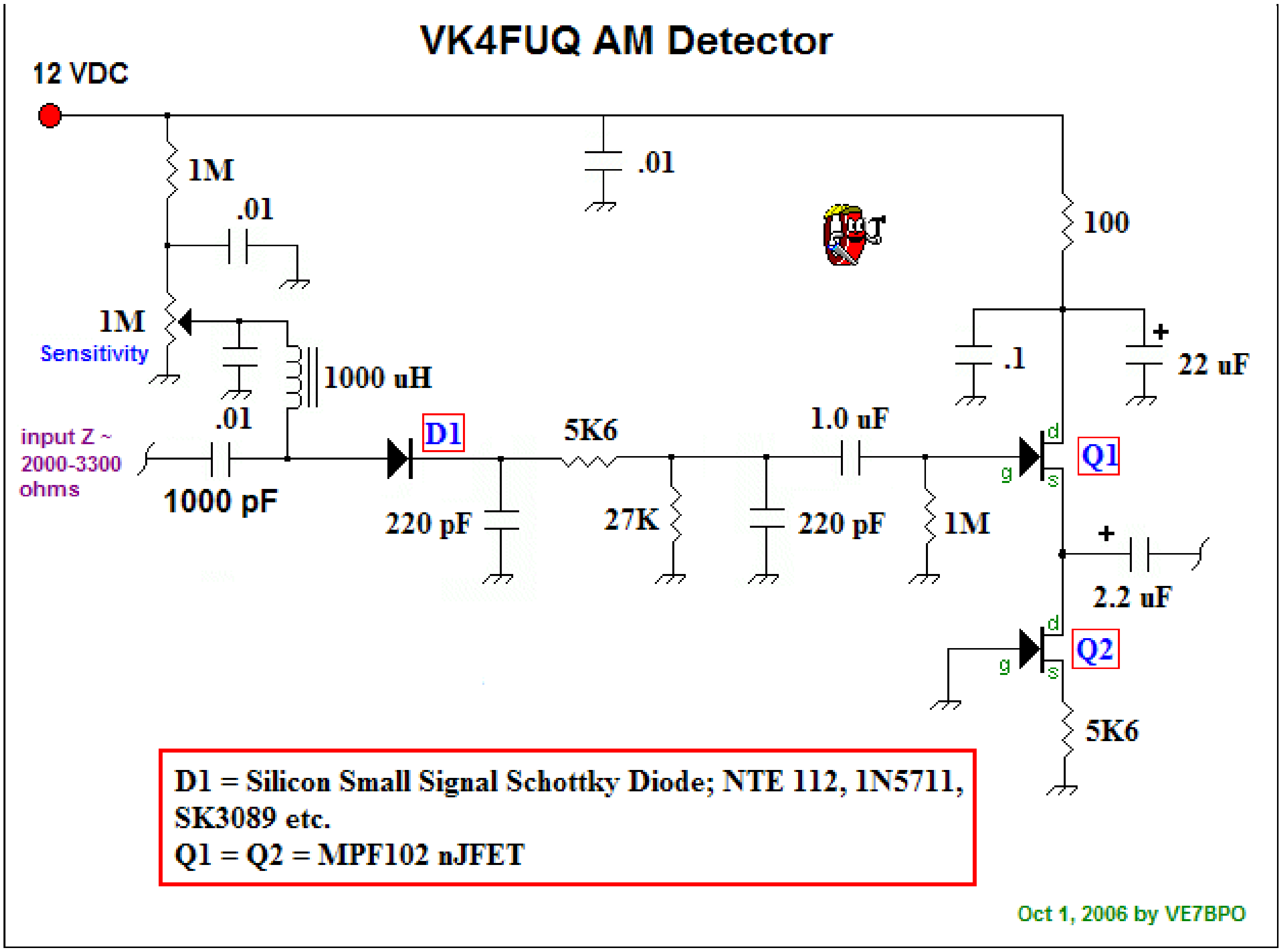
VK4FUQ AM Detector
12 VDC
1M
.01
.01
100
1M
Sensitivity
1000 uH
.1
+
22 uF
input Z ~
2000-3300
ohms
.01
D1
5K6
1.0 uF
Q1
d
g
s
1000 pF
220 pF
27K
220 pF
1M
+
2.2 uF
d
g
s
Q2
5K6
D1 = Silicon Small Signal Schottky Diode; NTE 112, 1N5711, SK3089 etc.
Q1 = Q2 = MPF102 nJFET
Oct 1, 2006 by VE7BPO

Above schematic. This detector is fabulous. It was designed by Felix Scerri, VK4FUQ. He has a web page explaining his high fidelity detectors at the Elliot Sound Products (ESP) site: http://sound.westhost.com/articles/am-radio.htm

The ESP web site is a personal favorite. Rod Elliot has one of the best do-it-yourself electronic web sites available. The main URL for his site is http://sound.westhost.com/index2.html

My sincere thanks to Rod and Felix for permission to present Felix's detector on this web page.

His improved AM detector has 3 positive advantages; it has high bandwidth, low distortion and incredible (and variable) sensitivity. I cannot get over how nice this detector sounds compared to others I have built and analyzed during weak and strong signal testing. The variable bias control allows the listener to adjust the bias to maintain detected audio fidelity even when the RF signal is weak.

This detector uses a UHF mixer diode often found in older television sets. Increasing the diode bias from O volts towards maximum causes three things to happen:

1. Increased sensitivity.
2. Increased audio high frequency response.
3. Slight increase in receiver noise.

When the WWV RF signal is weak, turning the bias off may result in the detected WWV signal disappearing. Increasing the bias will bring WWV back in. I generally run the bias control pot about 1/2 way and of course, higher as WWV fades out. I like the fidelity that the bias adds even when the WWV signal is strong. Note how the WWV audio quality continues to be high in fidelity as WWV fades out in this sound file.

Felix called for a 1 mH radio frequency choke. The largest I had in stock was a 1000 uH choke. I had to decouple it as shown to prevent oscillations from occurring in my receiver. For the 1 uF and 2.2 uF capacitors, I used polyester film types which sounded better than electrolytic capacitors.

Oct 13, 2006: Note. The 1000 pf input cap to the detector was omitted in error in the original schematic which is now correct.

To the right: The detector board. On the left is an op amp preamp stage that was later disconnected as it was not needed. Note the copper is removed where the chassis mounting nut contacts the copper board. Both audio boards were isolated from chassis ground and star grounded to a single point. The speaker negative terminal was also directly connected to this point. There is no hum.

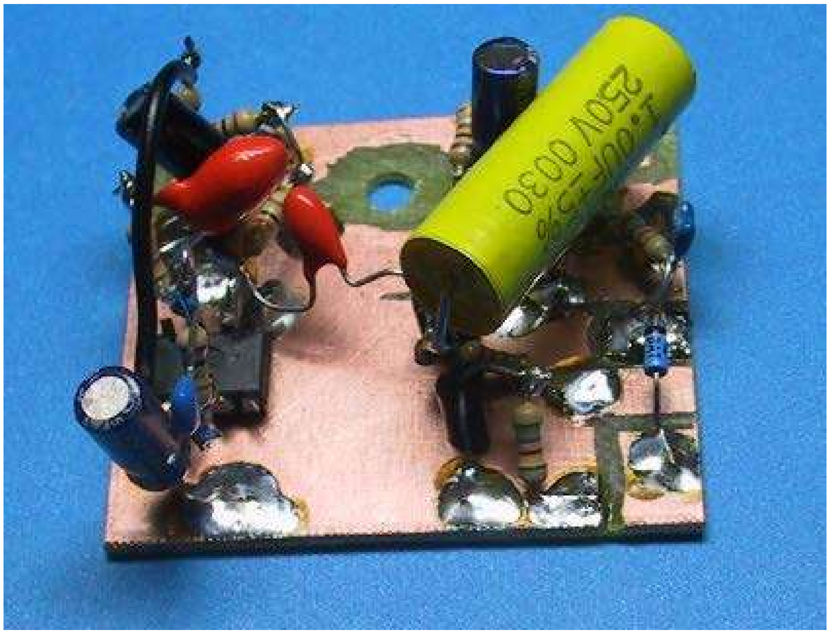

Audio Amplifier

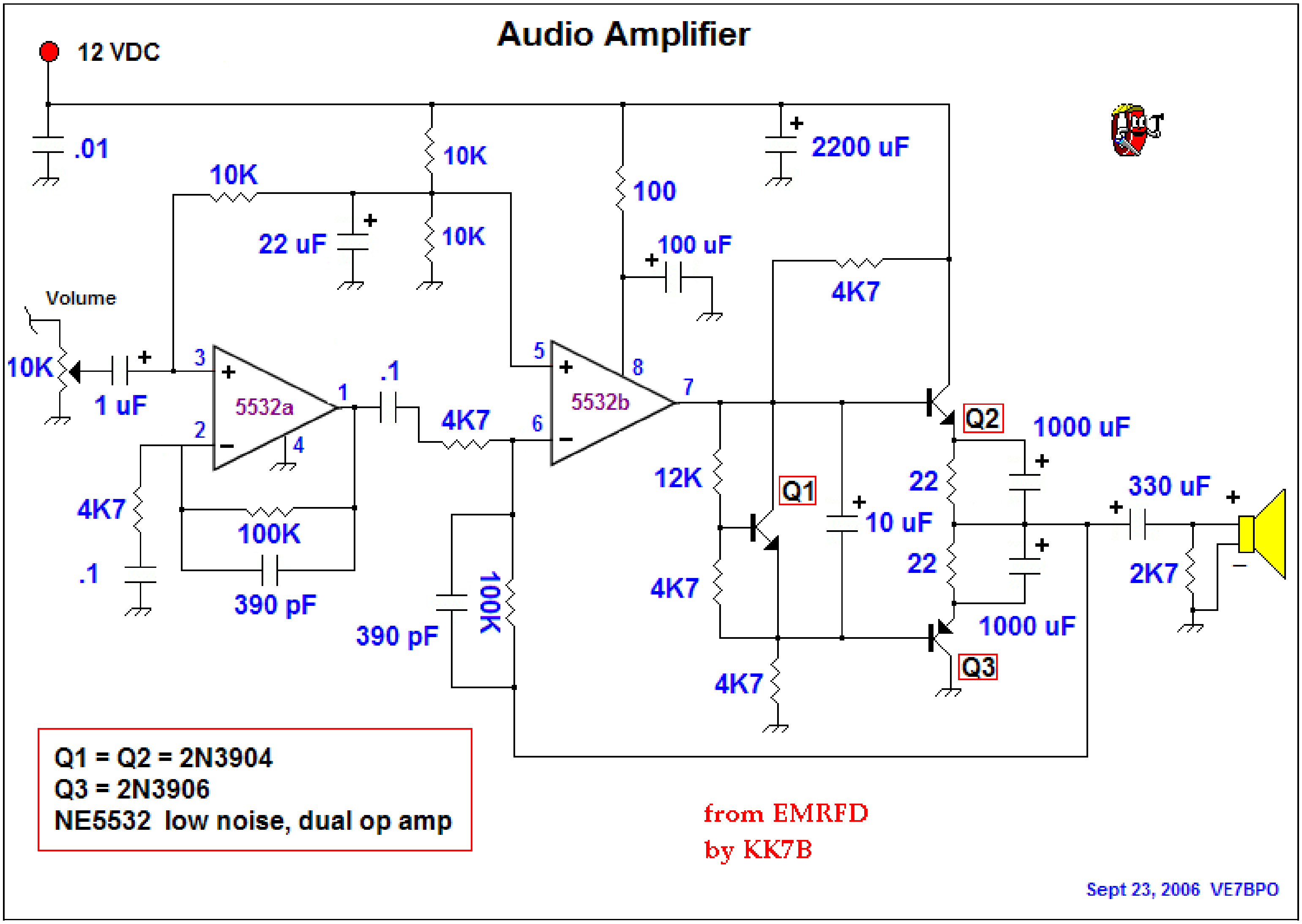
Audio Amplifier
12 VDC
.01
10K
10K
22 uF
10K
100
2200 uF
100 uF
4K7
Volume
10K
1 uF
3
2
5532a
1
4
.1
4K7
5
6
5532b
8
7
4K7
100K
390 pF
.1
390 pF
100K
12K
4K7
4K7
Q1
10 uF
22
22
Q2
Q3
1000 uF
1000 uF
330 uF
2K7
Q1 = Q2 = 2N3904
Q3 = 2N3906
NE5532 low noise, dual op amp
from EMRFD
by KK7B
Sept 23, 2006 VE7BPO

Above schematic. Rick, KK7B designed this low noise audio amplifier. It is from EMRFD. This superb AF amp greatly compliments the VK4FUQ detector. This is the best speaker audio amp under 1 watt I have ever used. Distortion is very low as long as it is not over-driven. I increased some capacitor values compared to the original schematic. Please refer to EMRFD for details on this stage. The chassis of this receiver greatly increases the low frequency response. On the 1 second pulses of WWV, the receiver "knocks" like a metronome. This does not occur when the chassis lid is off.

WWV web site: http://tf.nist.gov/stations/wwv.html All the often subtle pulses and tones transmitted at various times during the hour can be heard with this receiver.

To the right: A bread board of the AF amp. This is the audio amp I shall use in future projects which contain a speaker. Kudos to KK7B.

Below 3 images: Different views of the TRF receiver. On the front from left to right are the bias "sensitivity" control, volume control with integral power switch and blue LED "power on" indicator.

On the rear from left to right are the 12 VDC input jack, an unused switch (was an -10 dB attenuator at 1 point), the RF gain control and a coaxial SO239 connection.

Further Experiments

What follows are some of the ideas and circuits tried over the past year.

Schematic to the right: A 10 MHz, double tuned RF band pass filter that may be used ahead of the receiver. Insertion loss is ~ 3 dB and this filter uses a 5 pF coupling capacitor which are not too difficult to find. Filters with bandwidths of 150 - 180 KHz were also tested.

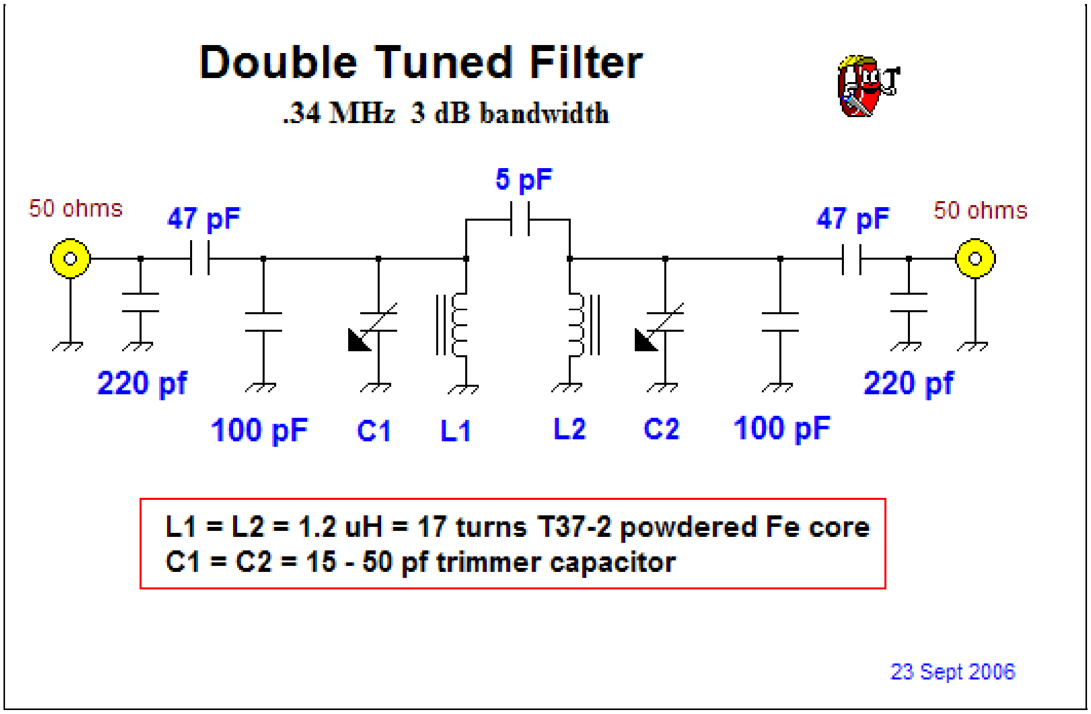
Double Tuned Filter
.34 MHz 3 dB bandwidth
50 ohms
47 pF
5 pF
47 pF
50 ohms
220 pf
100 pF
C1
L1
L2
C2
100 pF
220 pf
L1 = L2 = 1.2 uH = 17 turns T37-2 powdered Fe core
C1 = C2 = 15 - 50 pf trimmer capacitor
23 Sept 2006

To the right: A GPLA simulation of the popcorn DTC shown above. titlehough a little mistuned, it is reasonable for a filter that uses common junk box values and has low insertion loss.

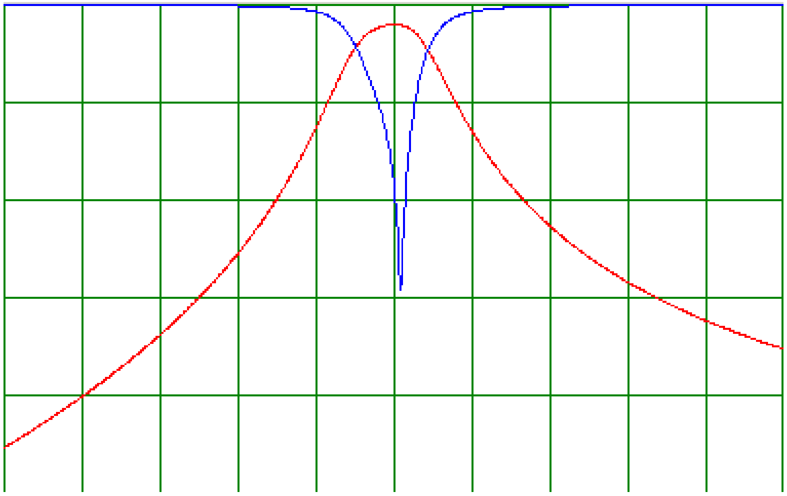

L-Match AM Detector

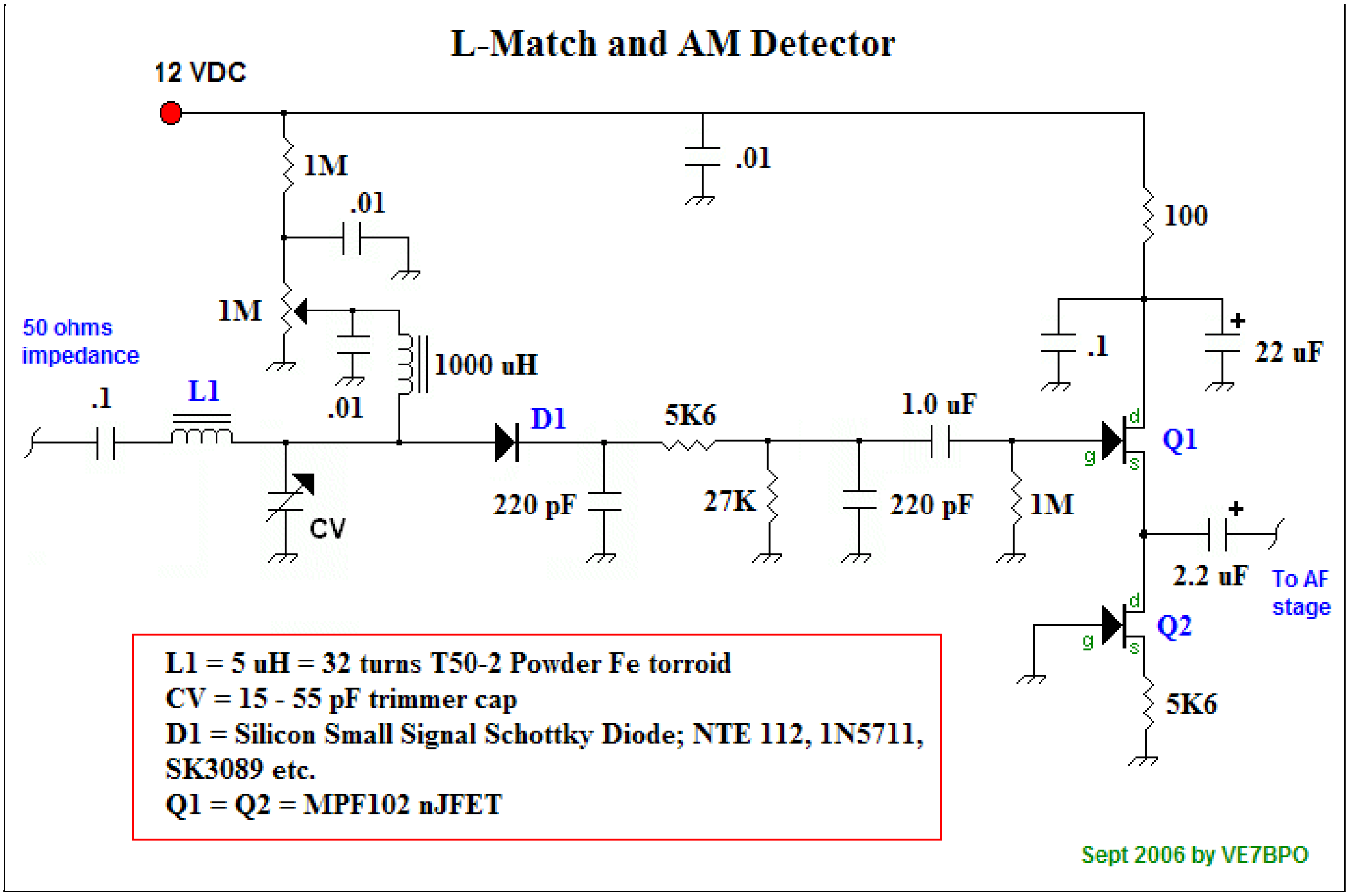
L-Match and AM Detector
12 VDC
1M
.01
.01
100
1M
50 ohms impedance
1000 uH
.1
22 uF
+
.1
L1
.01
D1
5K6
1.0 uF
Q1
g
d
s
CV
220 pF
27K
220 pF
1M
+
2.2 uF
To AF stage
Q2
g
d
s
5K6
L1 = 5 uH = 32 turns T50-2 Powder Fe torroid
CV = 15 - 55 pF trimmer cap
D1 = Silicon Small Signal Schottky Diode; NTE 112, 1N5711, SK3089 etc.
Q1 = Q2 = MPF102 nJFET
Sept 2006 by VE7BPO

Above schematic. The VK4FUQ detector can also be used to follow a 50 ohm output impedance stage by using an L-match as shown. The L match tunes very sharply. I peaked the L-match with the bias at 0 volts.

The input impedance of the detector is related to the DC current flowing in the diode. This is established by the adjustable bias current or "sensitivity control". The input resistance will be 26/I, where I is the current in mA.

For example, if the current in the diode is 10 microamps (0.01 mA) the input Z is 2600 Ohms. I have found that any input Z value from 2000 to 5100 ohms worked well with this detector.

Image to the right: A photograph of the L-match connected to a - 6dB 50 ohm pad which terminated the 50 ohm feedback amplifier that drove the L-match.

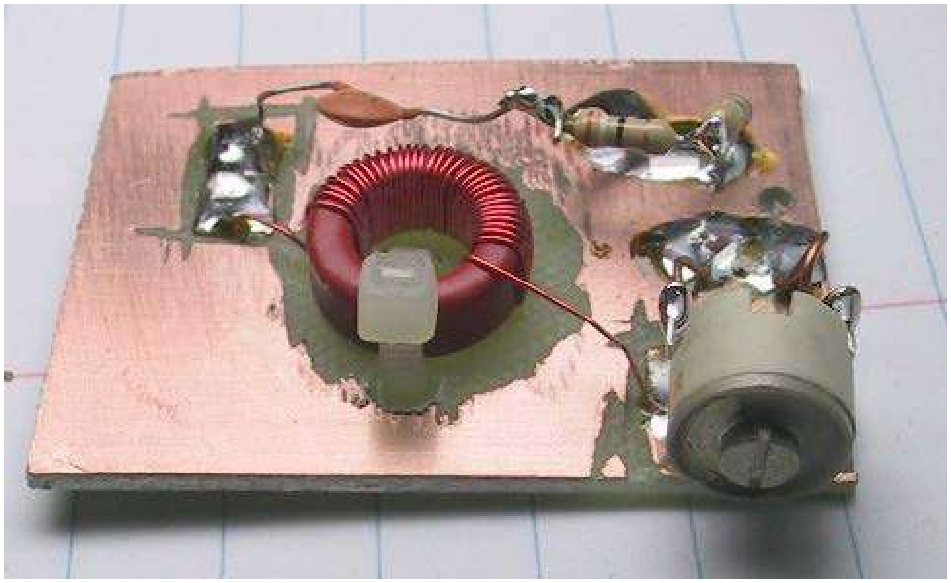

Feedback amplifiers

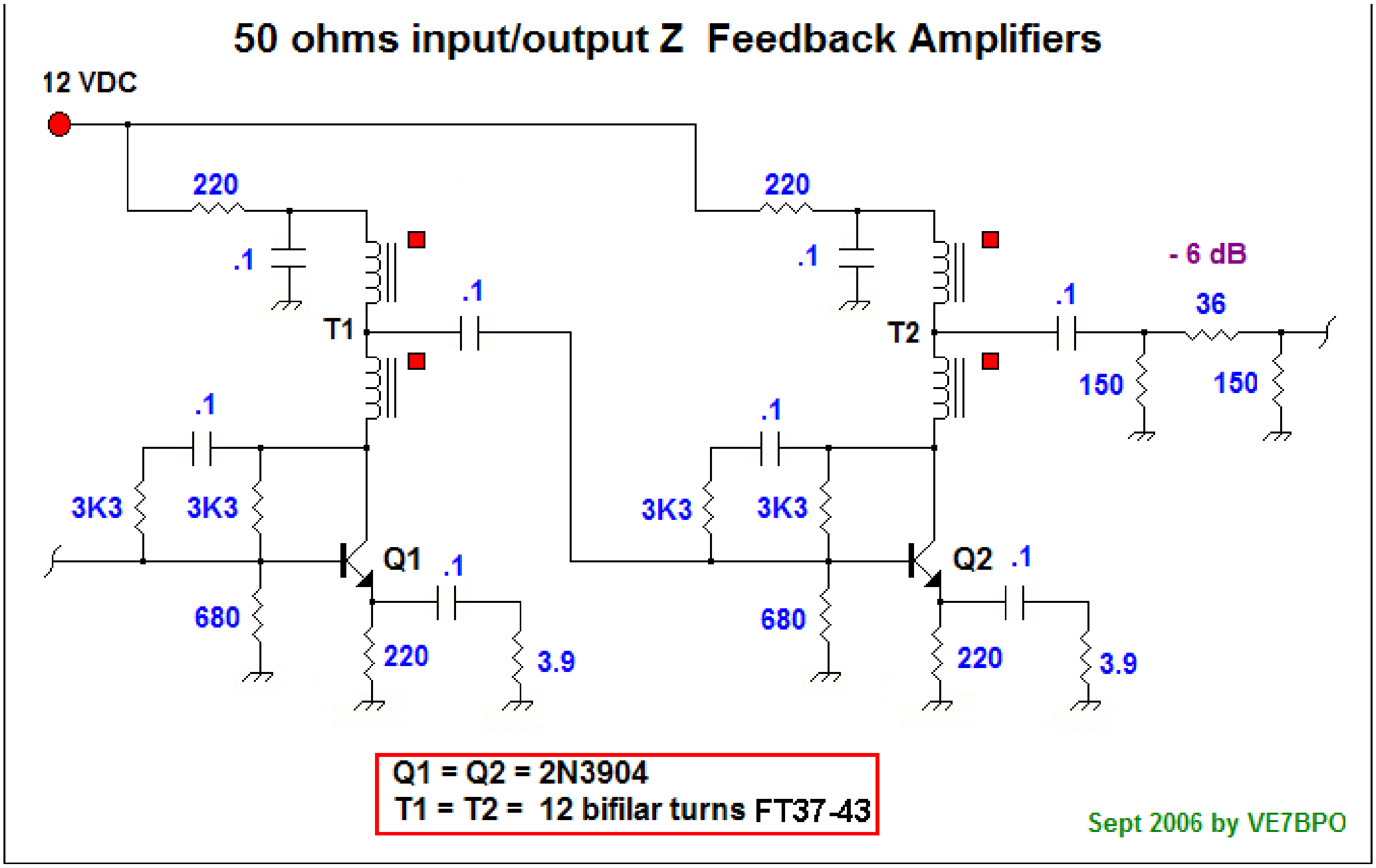
50 ohms input/output Z Feedback Amplifiers
12 VDC
220
.1
T1
.1
.1
3K3
3K3
Q1
.1
680
220
3.9
220
.1
T2
.1
- 6 dB
36
150
150
.1
3K3
3K3
Q2
.1
680
220
3.9
Q1 = Q2 = 2N3904
T1 = T2 = 12 bifilar turns FT37-43
Sept 2006 by VE7BPO

Above schematic. Feedback amplifiers may be used as RF amplifiers for a TRF receiver. This stage followed the crystal filter in one version of my TRF receiver. Stability was excellent. This feedback amplifier was designed by Wes, W7ZOI. It has ~ 20 dB gain and draws a little over 5 mA current.

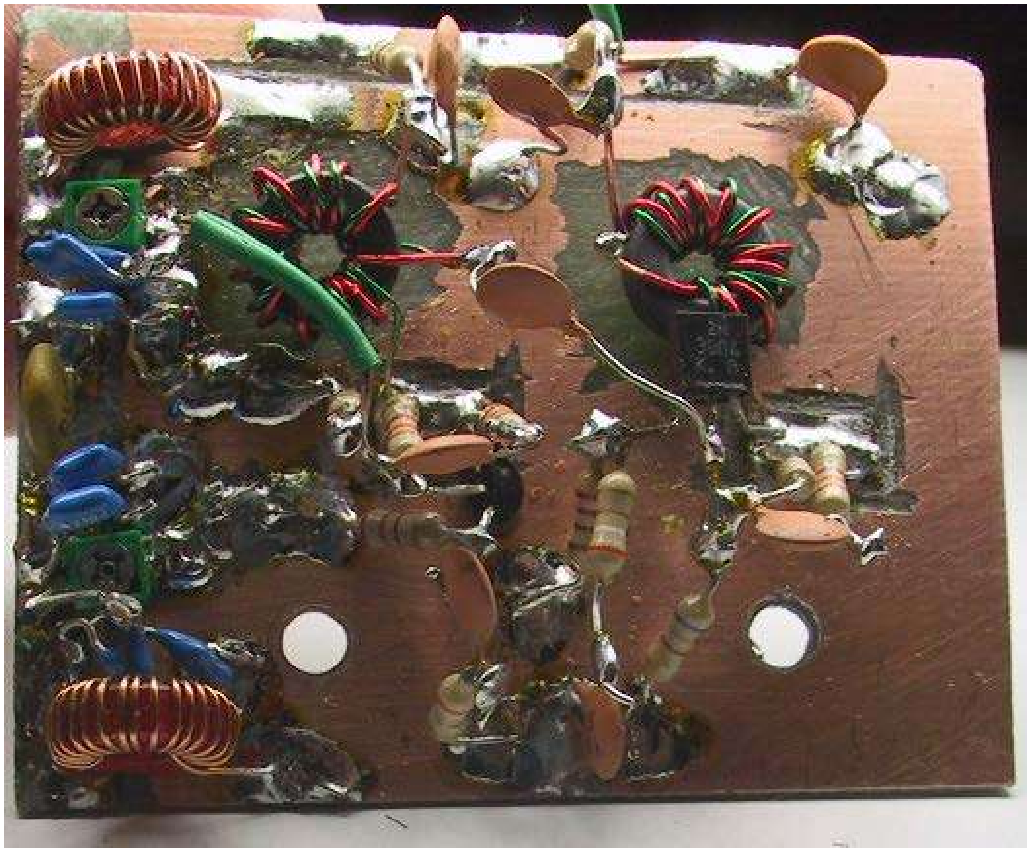

Above graphic. 2 feedback amplifiers are shown on this breadboard. The double tuned filter (DTC) shown earlier is also built on this board. In this version, the crystal filter was omitted and replaced with the DTC.

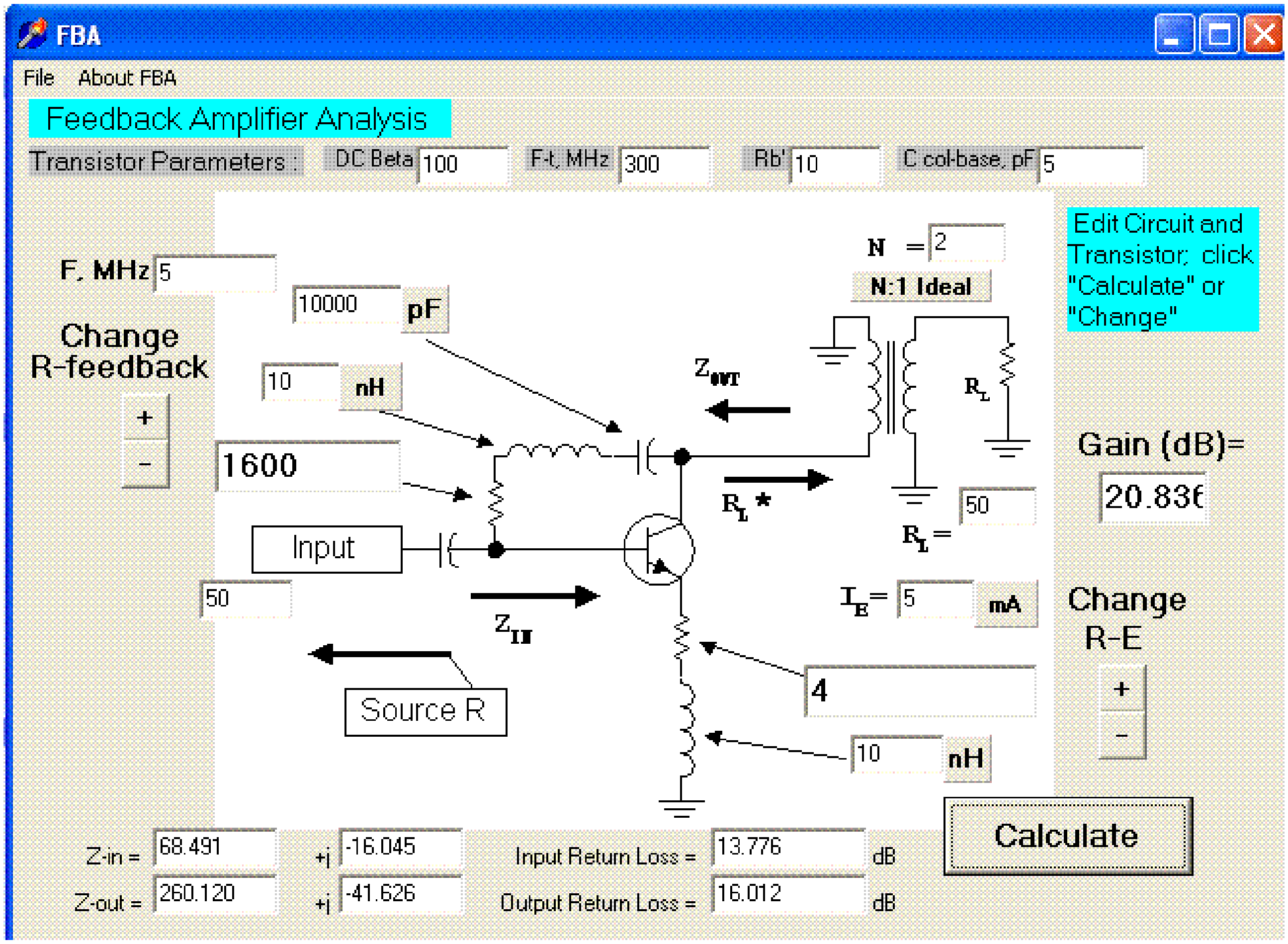
FBA
File
About FBA
Feedback Amplifier Analysis
Edit Circuit and Transistor; click "Calculate" or "Change"
Change R-feedback
+
-
Change R-E
+
-
Input
Source R
Z_{OUT}
Z_{IN}
R_L *
R_L
N:1 Ideal
Calculate

Transistor Parameters : DC Beta 100 F-t, MHz 300 Rb' 10 C col-base, pF 5

F, MHz 5

10000 pF

10 nH

1600

50

N = 2

R_L = 50

I_E = 5 mA

4

10 nH

Gain (dB)= 20.83

Z-in = 68.491 +j -16.045 Input Return Loss = 13.776 dB

Z-out = 260.120 +j -41.626 Output Return Loss = 16.012 dB

Above graphic. Using software that ships with EMRFD, W7ZOI designed the feedback amplifier used in this version of the receiver.

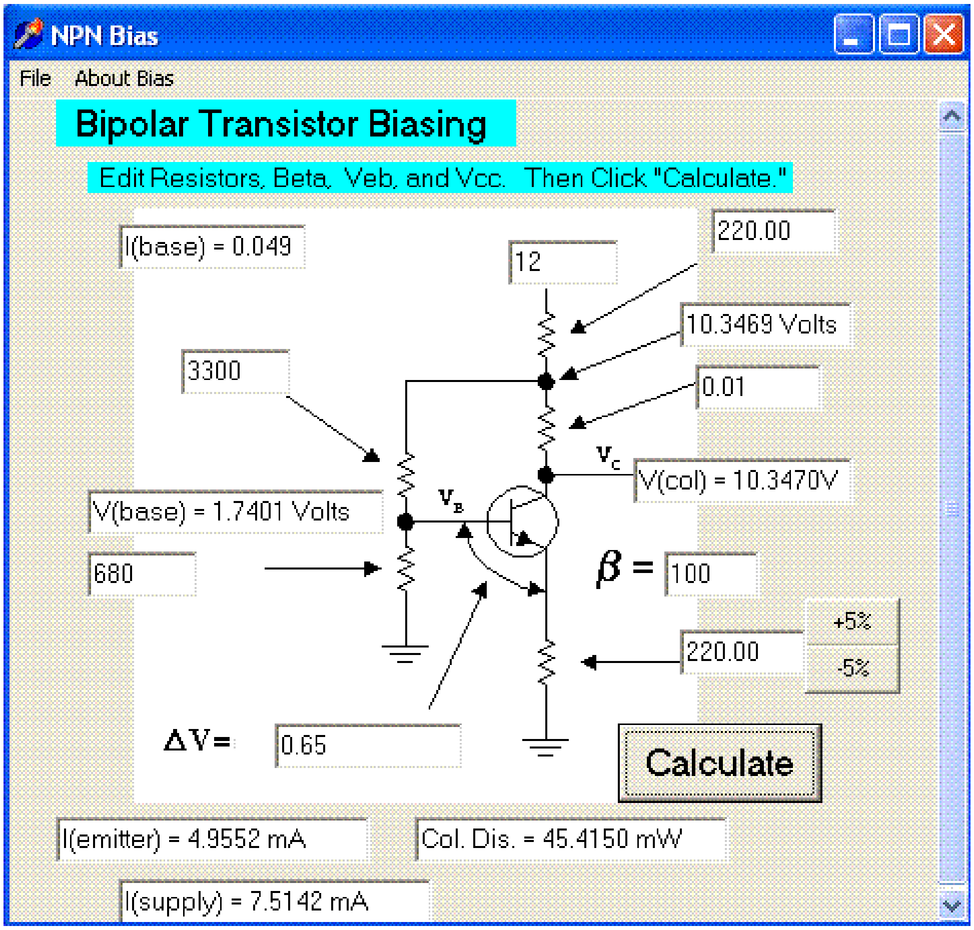
NPN Bias
File
About Bias
V_C
V_B
$\beta =$
$\Delta V=$
+5%
-5%
Calculate

Bipolar Transistor Biasing

Edit Resistors, Beta, Veb, and Vcc. Then Click "Calculate."

I(base) = 0.049

12

220.00

10.3469 Volts

3300

0.01

V(col) = 10.3470V

V(base) = 1.7401 Volts

680

100

220.00

0.65

I(emitter) = 4.9552 mA

Col. Dis. = 45.4150 mW

I(supply) = 7.5142 mA

Above graphic. The feedback amplifier bias resistor values were also calculated using software written by W7ZOI and included with EMRFD.

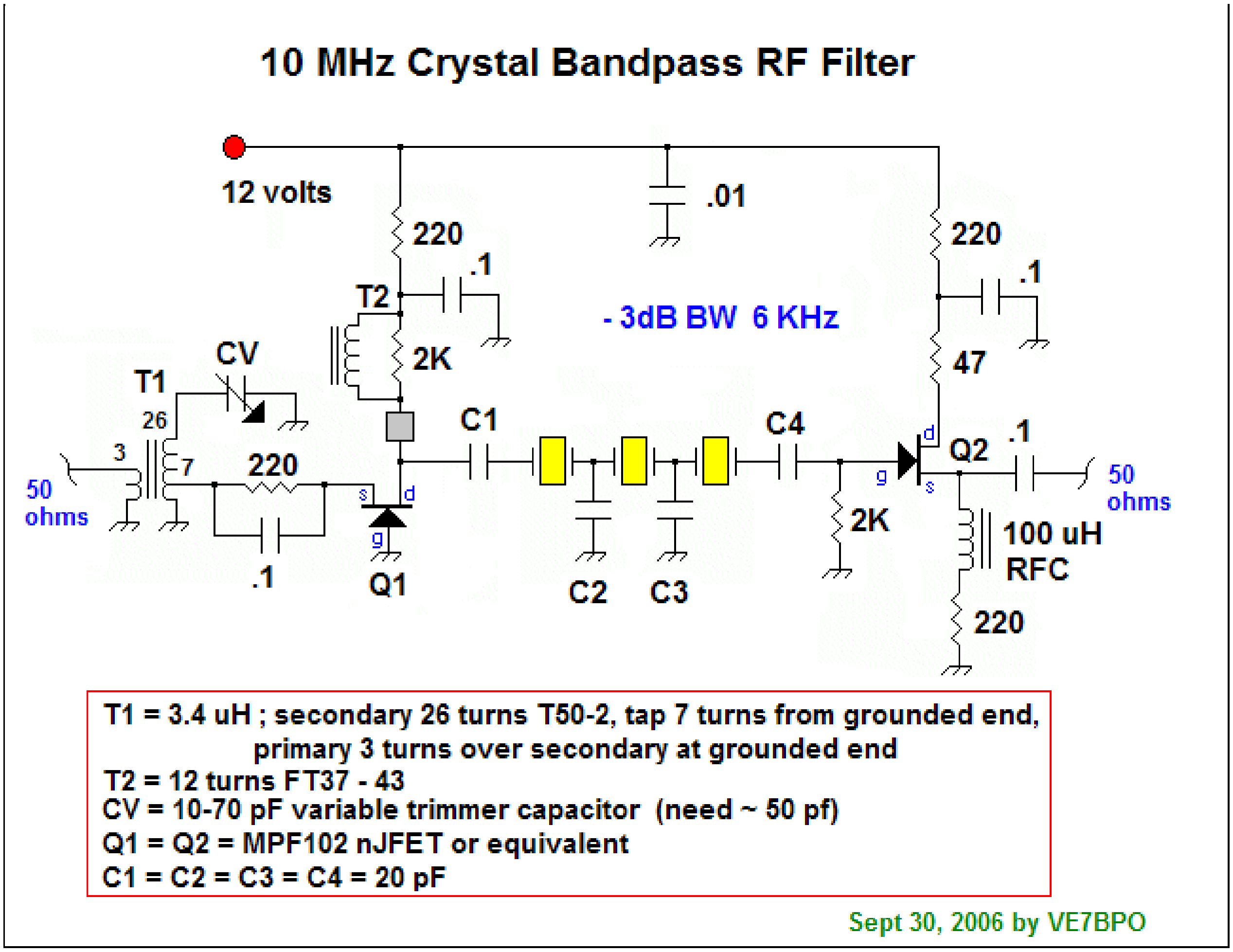
10 MHz Crystal Bandpass RF Filter
12 volts
220
.01
220
.1
T2
.1
- 3dB BW 6 KHz
2K
47
CV
T1
26
C1
C4
3
7
220
d
Q2
.1
50 ohms
s
d
g
s
50 ohms
g
2K
100 uH RFC
.1
Q1
C2
C3
220
T1 = 3.4 uH ; secondary 26 turns T50-2, tap 7 turns from grounded end, primary 3 turns over secondary at grounded end
T2 = 12 turns FT37 - 43
CV = 10-70 pF variable trimmer capacitor (need ~ 50 pf)
Q1 = Q2 = MPF102 nJFET or equivalent
C1 = C2 = C3 = C4 = 20 pF
Sept 30, 2006 by VE7BPO

Above schematic. This circuit has the crystal filter matched to 50 ohms input and output stages using JFETs. The JFETs also serve to provide a little more gain. Careful layout is required to reduce BCB interference for all stages in a TRF receiver.

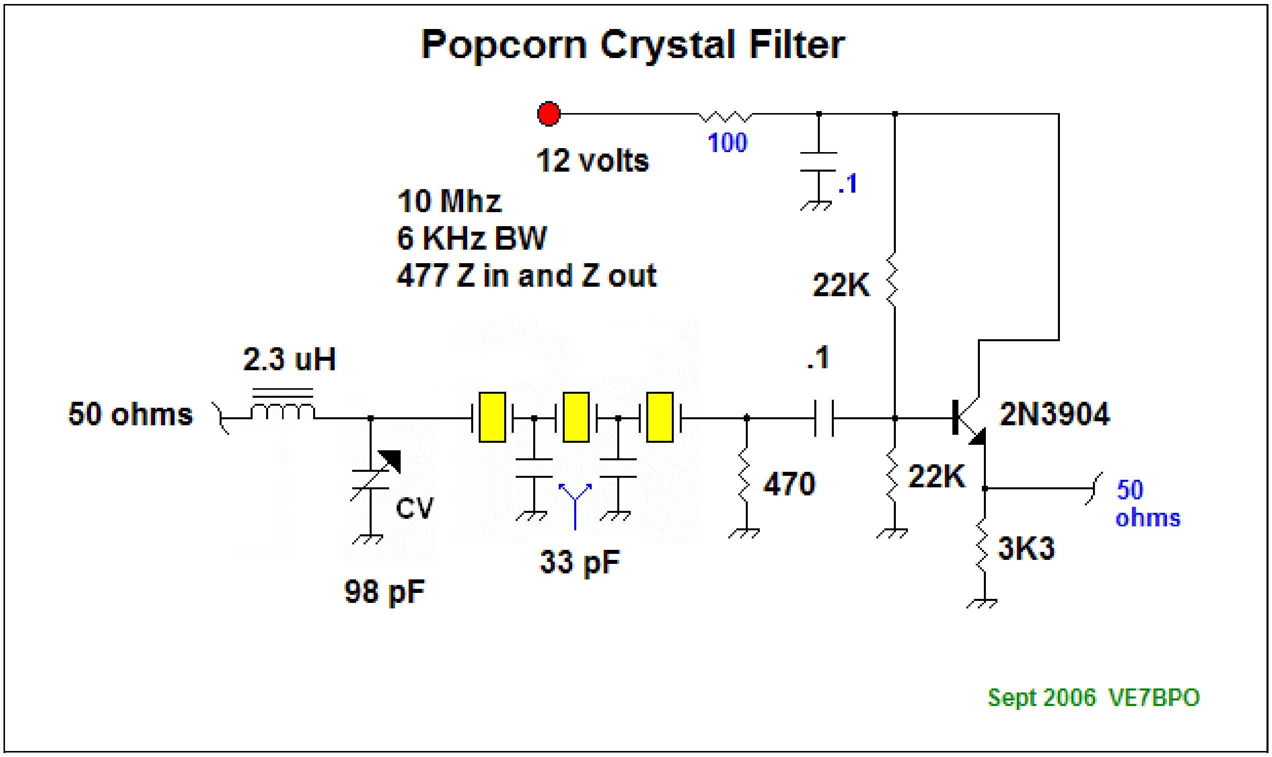
Popcorn Crystal Filter
12 volts
100
.1
10 Mhz
6 KHz BW
477 Z in and Z out
22K
2.3 uH
.1
50 ohms
2N3904
470
22K
50 ohms
CV
3K3
33 pF
98 pF
Sept 2006 VE7BPO

Above schematic. This is the original crystal filter that I designed. The input and output impedance is 477 ohms. The input was matched to the preceding 50 ohm stage using an L-network. A emitter follower is used to match the output to the 50 ohm stage which followed. Later, the emitter follower was replaced with the source follower (with a 470 ohm gate resistor) that is shown in the schematic directly above. The source follower had greater immunity to BCB interference and provided a better termination for this filter. This popcorn filter worked well, titlehough occasionally there was another station in addition to WWV, in the pass band. This also happened with the filter used in the final version of this receiver.

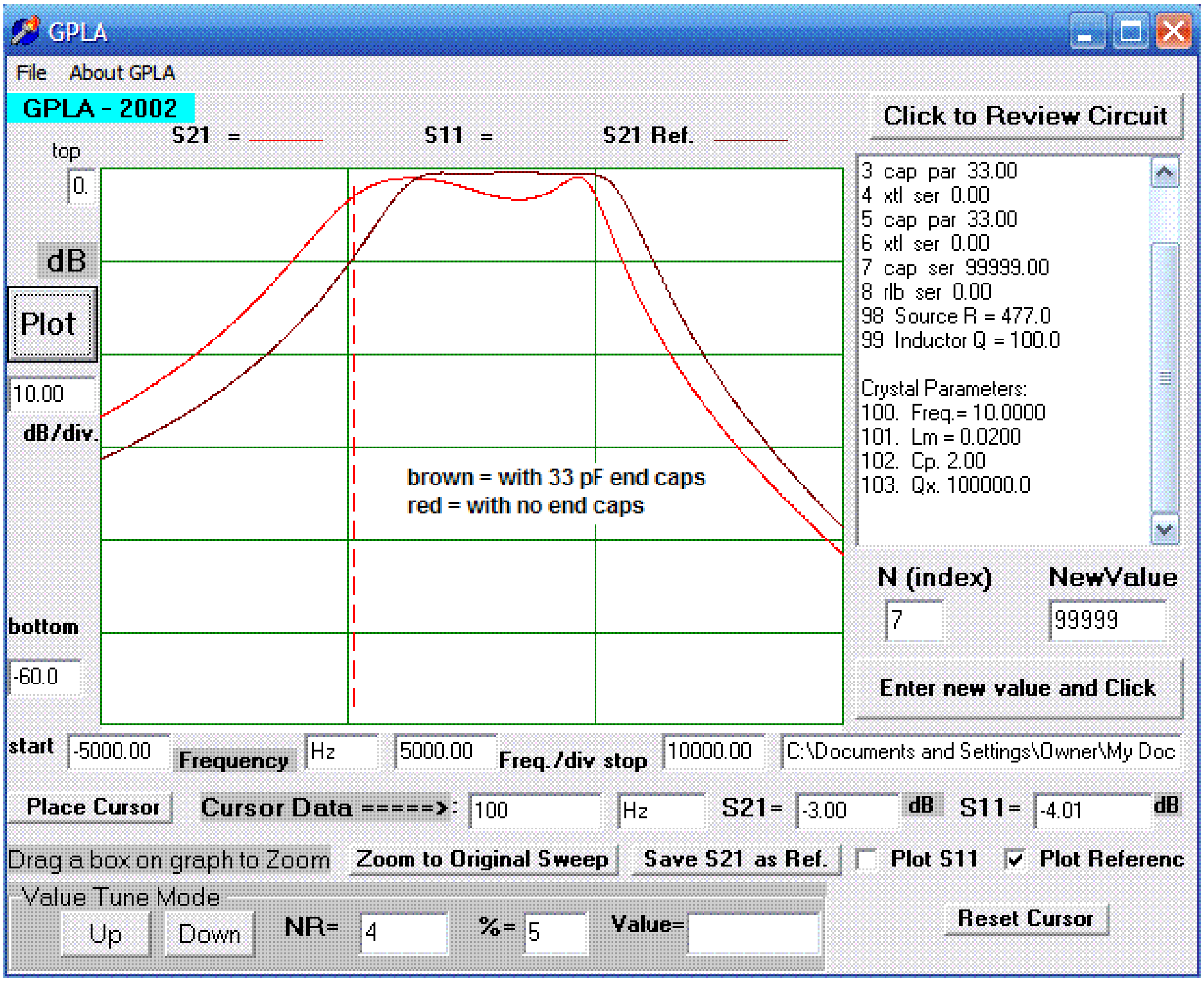
GPLA
File
About GPLA
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
dB
Plot
Frequency
Hz
Freq./div
Place Cursor
Cursor Data =====>
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Enter new value and Click
Value Tune Mode
Up
Down
Reset Cursor

brown = with 33 pF end caps
red = with no end caps

```
3  cap  par  33.00
4  xtl  ser  0.00
5  cap  par  33.00
6  xtl  ser  0.00
7  cap  ser  99999.00
8  rlb  ser  0.00
98  Source R = 477.0
99  Inductor Q = 100.0

Crystal Parameters:
100.  Freq.= 10.0000
101.  Lm = 0.0200
102.  Cp. 2.00
103.  Qx. 100000.0
```

top: 0.
10.00 dB/div.
bottom: -60.0
start: -5000.00
Freq./div: 5000.00
stop: 10000.00
C:\Documents and Settings\Owner\My Doc
Cursor Data =====>: 100 Hz
S21= -3.00 dB
S11= -4.01 dB
N (index): 7
NewValue: 99999
☐ Plot S11
☑ Plot Referenc
NR= 4
%= 5
Value=

Above graphic. Here are GPLA simulations of the popcorn filter with and with out 33 pF series end capacitors which serve to tune the filter. The brown tracing illustrates that it is better to include a series 33 pF cap at each end of the "popcorn" crystal filter. I did not use this filter because the dual gate MOSFET RF amps used in the final version, have better gain driving or following the 2000 ohm filter designed by W7ZOI.

GPLA is a "must-have" program. You can "tune" filters with different or asymmetric input/output impedances.

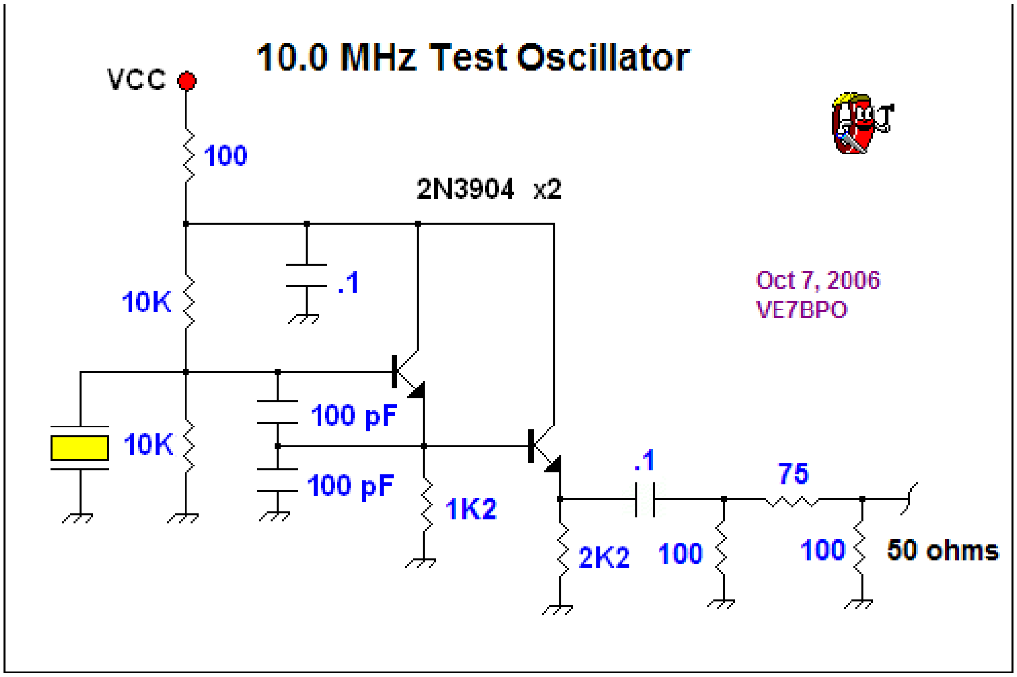
10.0 MHz Test Oscillator
VCC
100
2N3904 x2
.1
10K
Oct 7, 2006
VE7BPO
100 pF
10K
100 pF
1K2
.1
75
2K2
100
100
50 ohms

Above schematic. This 10.0 MHz crystal oscillator has a - 10 dB, 50 ohm pad on the output and was used to match the crystals, test the RF amps and align the filters used in these experiments.

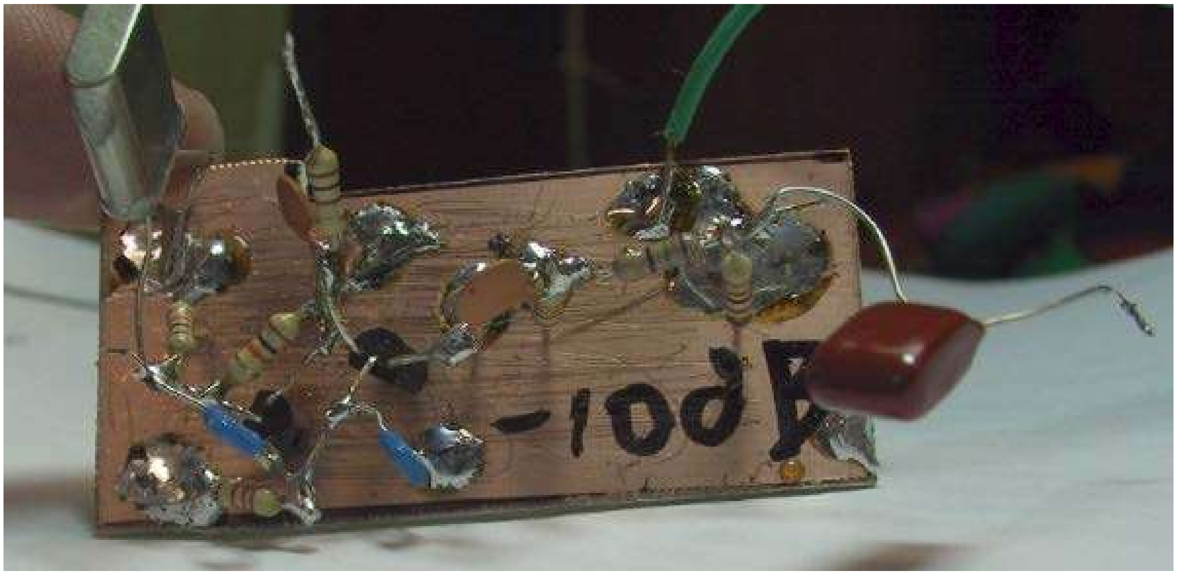

Above graphic. A breadboard of the test oscillator shown above.

Above graphic. Some of the bread boards developed during experimentation. My final receiver layout (and potentiometer positioning) is not optimal, however this is a prototype and I had no idea what the finished version would look like.

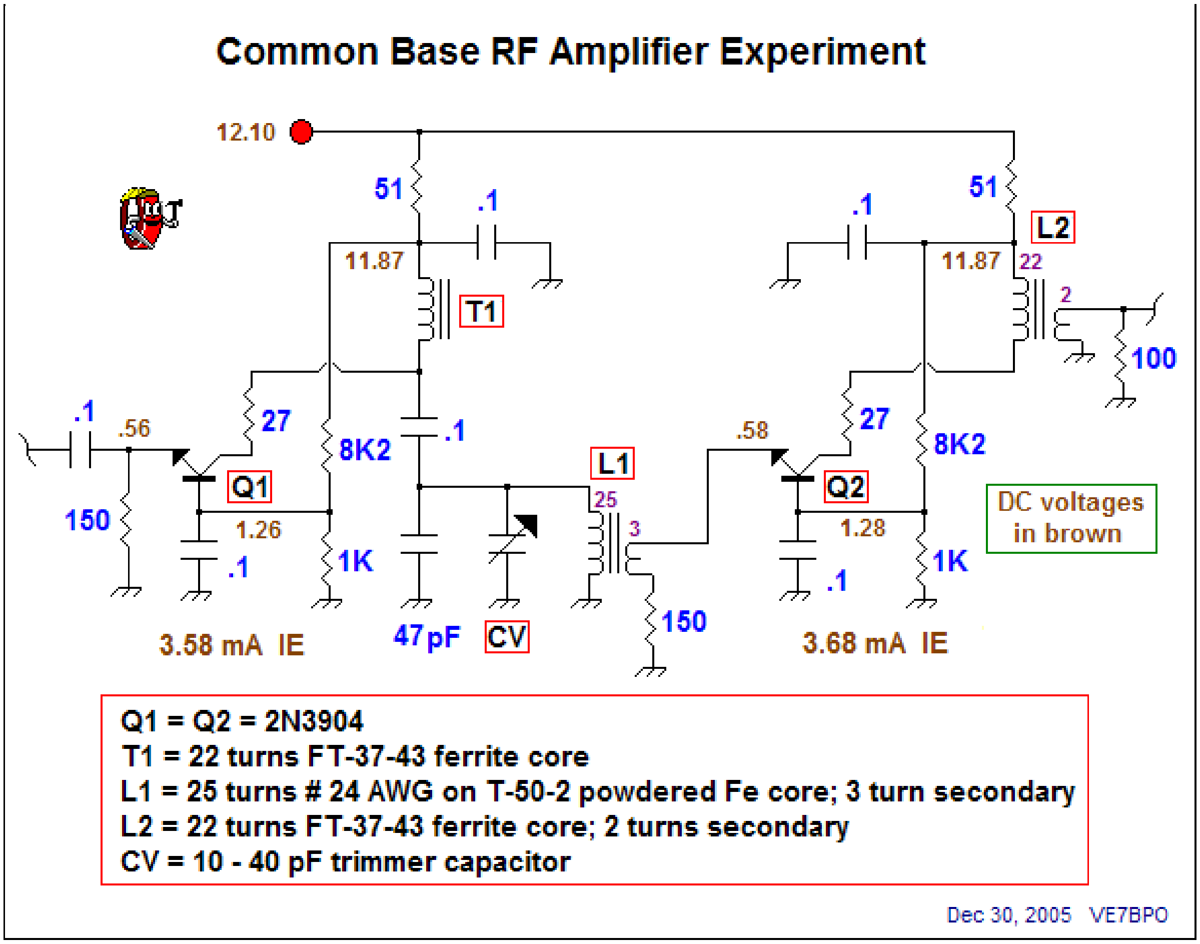

12.10
51
.1
11.87
T1
27
8K2
.1
.56
150
Q1
1.26
1K
47 pF
CV
L1
25
3
150
.58
Q2
1.28
.1
1K
3.58 mA IE
3.68 mA IE
51
.1
11.87
L2
22
2
100
27
8K2
DC voltages in brown

Common Base RF Amplifier Experiment

Q1 = Q2 = 2N3904
T1 = 22 turns FT-37-43 ferrite core
L1 = 25 turns # 24 AWG on T-50-2 powdered Fe core; 3 turn secondary
L2 = 22 turns FT-37-43 ferrite core; 2 turns secondary
CV = 10 - 40 pF trimmer capacitor

Dec 30, 2005 VE7BPO

Above schematic. This is one RF amp that was built for this receiver. The turns ratios on L1 is too drastic to afford much gain.

Above graphic. I have been told many times that my breadboards are very ugly looking. This breadboard of the schematic directly above, shows that occasionally, I can build a nice looking circuit!

Conclusion

The highlights of these experiments were VK4FUQ's detector and KK7B's AF amplifier.

When constructing such a receiver, *build backwards*. Install the speaker and then build and test the AF amp. Test it by touching your finger to the input and listening for noise or BCB radio. Turn the 10K pot and verify that the noise increases or decreases appropriately. Perhaps test it using an AF oscillator.

If it works, you get immediate positive feedback and motivation to continue. If it does not work, you only have 1 stage to trouble shoot.

Next, build the detector. To test it, connect a piece of wire about 25 cm long between the RFC and the anode of the diode. You should then hear local BCB radio. Slowly turn the bias potentiometer from 0 to fully on. Notice how increasing the bias may bring in 2 or more stations compared to when it was at 0. Also notice how it changes the tone and sensitivity of the detector. Try shortening the "test antenna" and observing how sensitive this detector is with the bias increased.

If all went well, you now have an AM radio!

Next add in the Q3 RF amp and again test it using a short piece of wire. Then continue on until you arrive at the antenna connection for your receiver.

Best regards, VE7BPO

MF and HF Receiving Antenna

Introduction

My first shortwave antenna was a simple end-fed wire which started at my bedroom window and extended out horizontally to a tree which was 25 feet away from our house. The antenna feed line was a short piece of wire that connected to the near end of the antenna and entered the house through a small hole I made in my wooden window sill. This feed line was directly connected to my receiver's high impedance antenna input. My station ground was long piece of wire that was connected to a copper pipe located in the bathroom next door. While this antenna brought in "the world" to my bedroom, it was extremely noisy. Directly connecting your antenna feed line and house ground system to your receiver are not good RFI reduction practices. This web page will explore some experiments in trying to minimize the Radio Frequency Interference (RFI) arising from my local environment.

Indoor RFI sources are usually plentiful. Electrical appliances such as washing machines, televisions, DVD players, computers and electrical wiring may all emit RFI which your antenna, or directly connected house ground system may pick up and feed to your receiver. Certain indoor devices may be really strong RFI sources and will have to be eliminated or decoupled. Outside of your house are also potential sources of RFI. These may include such things as power transformers, electric fence and garage door openers. RFI location and reduction is out of scope for this web page, however a good place to learn more is the ARRL RFI book (out-of-print: search on Amazon). To find RFI sources in your home and neighborhood, try using a battery powered AM radio. At my QTH, I located a noisy VCR inside the house my Grundig S350. We rarely use this VCR and now just leave it unplugged until we actually need to operate it. I tuned the receiver to an empty frequency and found this VCR by trial and error. **Please note this web page is concerned with feeding a shortwave listening antenna and does not describe providing protection against lightning.** For web sites which covers lightening plus RF ground please refer to this offering from W8JI or eHam.net.
Protect your home and family from lightening !!

Outdoor MF and HF Antenna

The schematic to the left summarizes the outdoor VE7BPO MF and HF receiving antenna system for summer 2007. Although modest for a big city lot, this antenna seems to pull in the DX and is relatively free of RFI. This antenna was just a case of "putting as much wire in the sky as possible" and the dimensions are indicated for interest sake only. The 27 meter long horizontal section is supported between 2 trees at a height of about 14 meters high. The weight of the

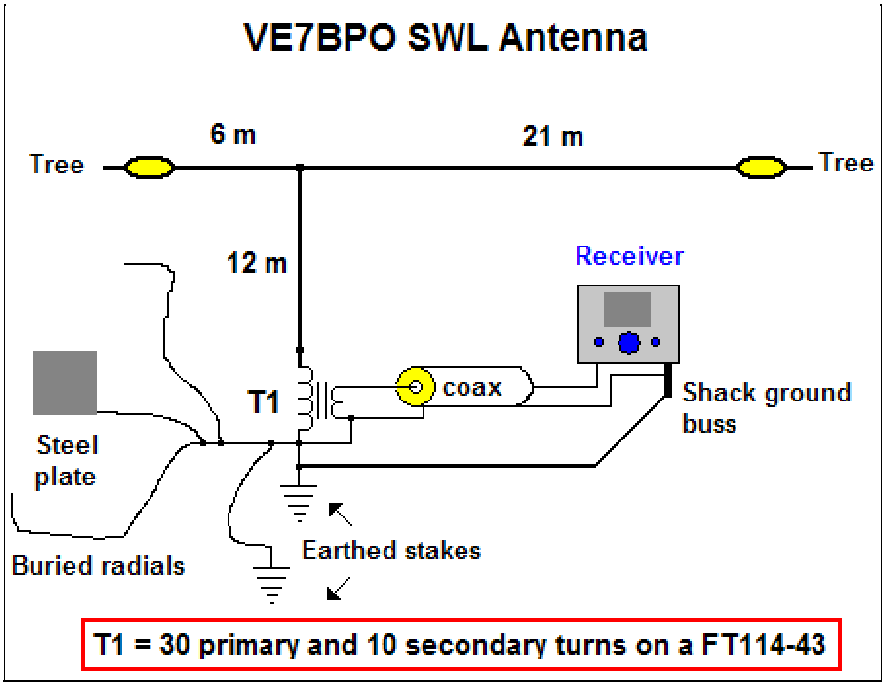

vertical element wire plus slack in the horizontal wire droop it to about 13 meters high in the center. The vertical section is soldered to the horizontal wire 6 meters from the nearest anchoring tree and runs straight down to the antenna feed point which is about 1 meter off the ground. The feed point is a piece of copper-clad PC board (with isolated sections created with a hobbyist motor tool) and is bolted to a long copper pipe which serves as the first station earth-grounding stake. A transformer (T1) configured as a UNUN **(unbalanced-to-unbalanced)** is used to interface the antenna with 50 ohm coax that runs through the house and into the radio shack. Some rudimentary experiments with the UNUN and the earth-grounding system were undertaken.

The methods I used to potentially lower unwanted RFI to my antenna system are as follows:

1. The receiver and power supply are independently connected to a single, central ground point (ground buss) in the radio shack.
2. 6-10 gauge wire is used for my ground system (not including the radials which are bare 12 gauge wire).
3. The ground wire connecting to my first earth stake to the station ground buss is just outside the shack window and is short as possible to provide a low impedance and low inductance path for MF and HF frequencies.
4. There is a second ground stake located 1 meter from the primary ground stake (I will add 2-4 more in time).
5. I have a large piece of steel buried underneath the soil tied in to my system as well as 3 bare copper radials. The radials are 3 - 7 meters in length.
6. New RG58/U coax was used as the feed line.
7. All wire splices in the grounding system are soldered and taped up. I used conductive grease (to prevent oxidation at the wire-stake interface) on any clamps connected to ground stakes. My ground stakes are ~ 2 meters long.
8. The earth grounding area soil is moist and peat-laden and is watered regularly.
9. I plan to maintain this ground system every 2 years.

4:1 UNUN

My antenna is almost an end-fed wire with both a vertical and horizontal section. I do not have the gear to measure the impedance versus frequency in the MF and HF bands. I do know that on some bands it may present an impedance of several thousand ohms and a transformer can smooth out the variation in impedance versus frequency so my receiver sees a relatively low impedance on most bands. The transformer also serves to help reduce RFI from my antenna system by eliminating unwanted common mode currents flowing on the outside of the coax braid. Grounding the antenna via the UNUN will also prevent static electricity from building up on the antenna. My first UNUN had a 4:1 impedance ratio. It is shown to the right. The antenna connects to point **A**. The ground stake connects to point **B.** Point **C** connects to the inner wire of the coax and point **D** is connected to the braid and also the grounding stake. I used 24 AWG wire and A FT114-43 ferrite core. You can clearly see there are 20 primary windings and 10 secondary windings loosely coupled. I chose the FT114-43 core because I had it on hand and the 24 gauge wire provides good mechanical support for the coil. I could have used an FT50-43 ferrite torroid as well with a smaller wire gauge. You can also use a bifilar transmission line type transformer. I was very happy with this UNUN and however it did not have as much signal strength as I expected on the 160 meter amateur band and below.

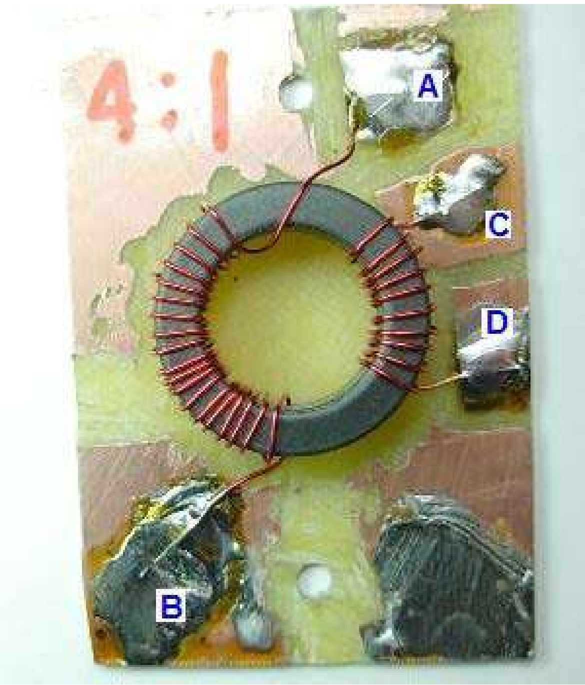

9:1 UNUN

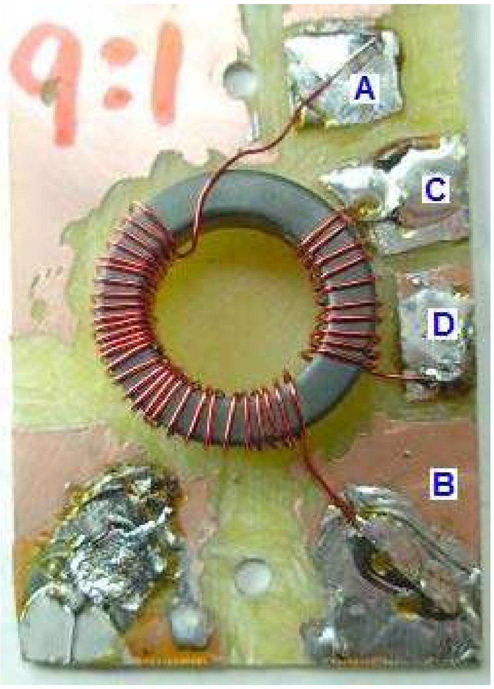

Next I tried a 9:1 impedance ratio UNUN. This an extremely popular impedance transformation ratio for end-fed or random wire SWL antennas. I wound 30 primary and 10 secondary turns on a FT114-43 ferrite torroid. The connection points are identical to those described in the above 4:1 UNUN. Remember that the impedance transformation ratio is the square of the actual turns ratio on your transformer; thus my 3:1 turns ratio is a 9:1 impedance ratio. Electrical engineers commonly use a rule when winding broadband transformers such as these. The inductive reactance (XL) of the smaller winding must be at least 4 times the load impedance at the lowest frequency that the transformer "looks" into. So for 50 ohm coax, the XL should be at least 200 ohms at 500 KHz which is the lowest frequency I intend to receive. The formula for XL is **XL = 6.28 X F X L**. Frequency (F) is in Hertz and L is the inductance in Henries.

At 500 KHz my inductor has an XL of 189 ohms which is almost perfect. I should have used 11 turns which is an XL of 229 ohms and strictly observes the 4X rule. Therefore my UNUN ideally should have used 33:11 turns on the FT114-43 torroid. If you use a FT50-43 torroid, use the same 33:11 turns ratio; this will provide 198 ohms XL at 500 KHz. For practical purposes, my 30:10 UNUN should work fine as I rarely tune frequencies less than 1000 KHz. I found this UNUN to have strong signals all the way down to MF and decided to use a 9:1 impedance ratio for my antenna system. Many experimenters and a few commercial UNUNs recommend the 9:1 impedance ratio for multiband end-fed or random wire antennas. Eventually I will encase it in a water and UV proof enclosure.

Conclusion

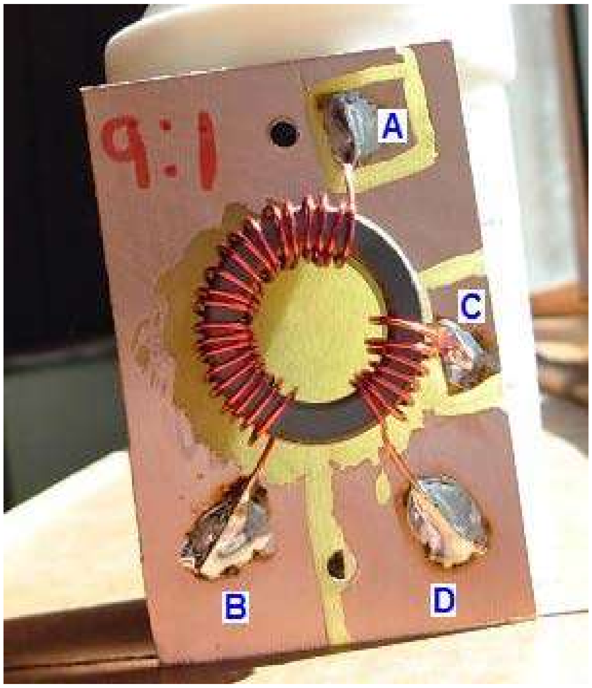

My experiments while constructing a reasonable quality MF and HF receiving antenna confirmed that using a UNUN, coax and a good RF ground system can reduce common mode RFI in my receivers.
I also tried temporarily connecting my ground system to a copper water pipe located in my shack while listening to WWV at 5 MHz and immediately the noise level rose 2 S-units on my receiver.
This pipe was clearly not grounded in my house where there is a mixture of plastic and copper water pipes. Additionally, my antenna wire and feed point is away from the house in a quiet area according to listening tests using a Grundig S350.

It is relatively easy to construct a UNUN on your bench using a ferrite torroid. Many builders have emailed me to say they do not feel comfortable winding torroids. Torroids are easy to use and by winding a couple and experiencing some success, your confidence working with them will surely improve. If you live in a part of the world where you can not easily obtain a suitable ferrite core, just email me and I may send you an FT50-43. You can also choose a ferrite with a different core permeability. Some builders use number 75 material. I used the FT114-43 because I get all my torroids from W8DIZ and just use what he has in stock for my projects. If you really do not want to construct a UNUN, commercial products are available on the web on sites such as http://www.arraysolutions.com/Products/baluns.htm.

I wish you good luck with your own antenna experiments and please be safe!

Some SWL Antenna Related Links

L.B. Cebik, W4RNL was a respected antenna expert. There is great information on his web pages

Build a Shortwave Antenna. A good overview of home brew multi-band antennas by N4UJW

Amateur and Short Wave Radio Electronics Experimenter's Web Site

MF and HF Receive Antenna Splitter

Introduction

As a radio experimenter, I have numerous MF and HF receivers to listen to but usually only 1 main outdoor antenna. Typically, this means that only my main radio receiver is connected to the outdoor antenna and my other receivers must use small indoor antennas with or without RF preamplifiers. I wanted to to permanently connect my main radio shack receiver and the receiver in the room directly above the shack to my main MF and HF antenna at the same time. The solution was to build a simple antenna splitter which allows the 2 radio receivers to connect to the single coaxial antenna feed line while preserving the correct impedance at all connection points. This project is based upon the splitter presented in EMRFD labeled Figure 3.81.

Each receiver and the antenna feed line have a 50 ohm characteristic impedance. This in-phase splitter is passive and has a loss of just over 3 dB. It is designed to operate from 500 KHz up to 30 MHz.

Please do not transmit through this device.

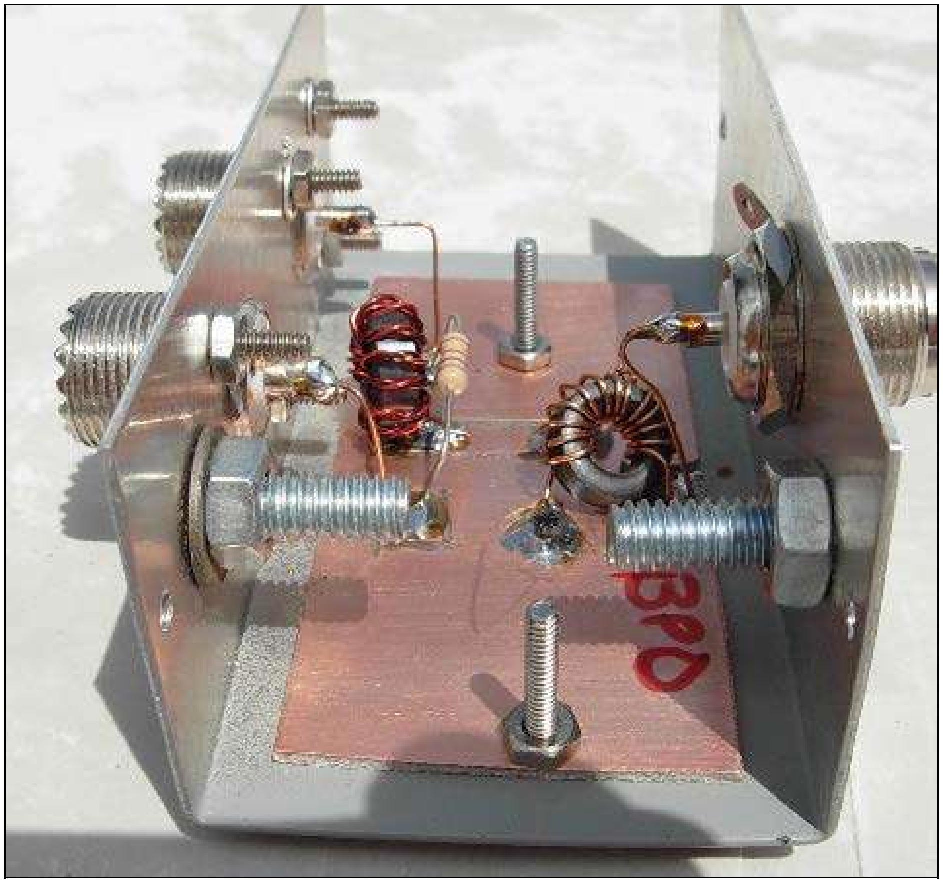

Project Schematic

The schematic to the left illustrates the entire splitter network from the antenna input to the input of the 2 receivers. T1 and T2 are broadband ferrite transformers with enough inductive reactance to tune down to the bottom of the broadcast band. If you only require a splitter for HF, then wind T1 with 10 total turns and a tap at 7 turns from the grounded end and T2 with just 10 bifilar windings. I used FT50-43 cores to allow the use of

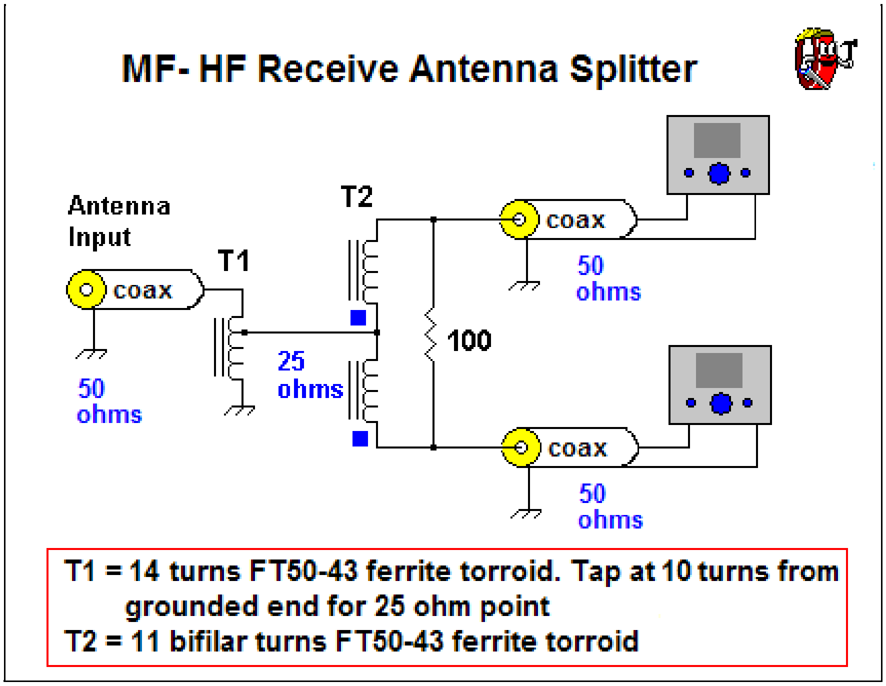

T1 = 14 turns FT50-43 ferrite torroid. Tap at 10 turns from grounded end for 25 ohm point
T2 = 11 bifilar turns FT50-43 ferrite torroid

thicker gauge wire which provides reasonable securement of the coils without external anchors, and because bigger inductors are easier to photograph. The FT37-43 ferrite core would also be a good choice, especially if miniaturization is a design goal.

Circuit Building Details

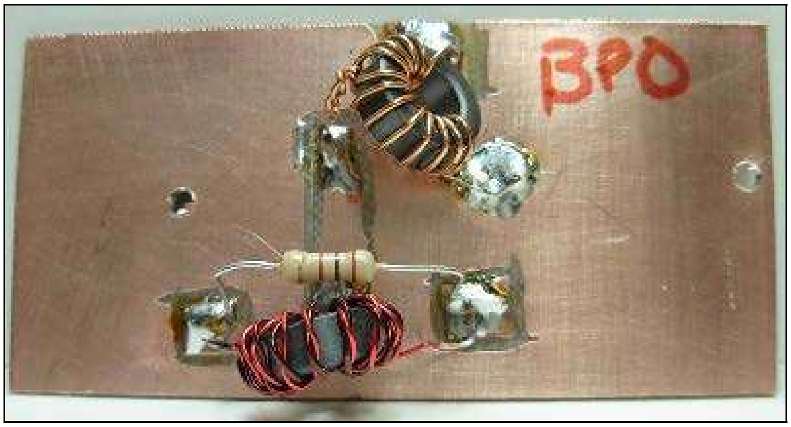

The antenna splitter breadboard is shown to the right. I used 3 colors of 22 gauge enamel covered wire to make my inductors. T2 is the actual splitter network coil and is the lower transformer in the photograph. The 100 ohm resistor serves to isolate the 2 ports connected to the receivers and absorbs impedance mismatches which may present when one terminal is not properly terminated. Note that the characteristic impedance at the input of this 3 port network is only 25 ohms. You can choose to ignore this or use an additional network such as a broadband transformer or an L- match between your antenna coax and the splitter to match this 25 ohm impedance.

I chose to use T1 which is an autotransformer with a tap at approximately the 25 ohm point. The splitter network worked well during my tests. Having only 1 receiver versus 2 receivers connected made no difference to the signal strength due to the excellent output port isolation.

Note that Wes, W7ZOI uses this 3 port network network several times in EMRFD. One example is the Lichen transceiver while another is the 6M superhet receiver presented in Chapter 6. Consult EMRFD for further discussion of this and other multiple port networks.

T2 is a bifilar transformer. The 2 wires were twisted together by securing one end of the 2 wires in a vise and the other end of these 2 wires in the chuck of a brace and bit (manual) hand drill. I twisted the hand drill until I had 8 twists per inch on the 2 wires . I used 2 color wire for ease of construction, however, it is almost as easy to tell the windings from one another by using an ohm meter or audible continuity tester.

Chassis

To the right is the completed project showing the SO-239 connectors which are wired to the antenna splitter output ports. A chassis from an old project was recycled for this new project. The large bolts seen in some of the of the photos were used to fill in holes which had been drilled for the old project. This was done to provide improved RF shielding. The bolts also increased the weight of the chassis and help keep it from tipping over. Although it does not look as attractive as if I had used a brand new project chassis, considerable cost savings were realized. These little Hammond project boxes are getting very expensive. Also the splitter is kept an the back of my main radio desk where it is out of sight anyway.

Update August 10, 2008 - Contribution by Dave, G4AON

This original network was designed for use in the MF to HF spectrum. Limiting this network is the input matching transformer T1 which negatively effects the T2 output port isolation; especially at 41 meters and higher. Testing by Dave, G4AON confirmed this. Dave designed, built and tested a trifilar wound, UNUN input matching transformer which provides a much flatter response for T2 port isolation from 0.1 to 52 MHz.

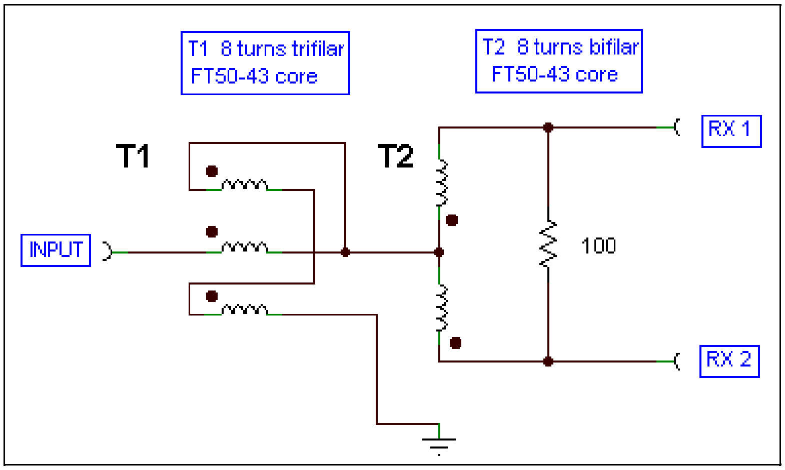

To the left you can see the G4AON input circuit for T1. In keeping with a design optimized for higher frequencies, less total turns are used on the transformers. His trifilar wound input transformer version is going to generate an impedance of (16/24)^2 x 50 = 22 Ohms at mid-band. My variation will generate (10/14)^2 * 50 = 25.5 ohms at mid-band. Using his version of T1 as opposed to my simple auto-transformer, Dave was able to provide better isolation of the output ports than the original design across a wider range of frequencies.

In the two popular, commercially sold RF splitters we have examined, the company did not even bother to match the input to the T2 transformer and some builders have written me to say they just omitted T1 and for their typical SWL listening this worked out fine for them. Increasing isolation across a wider frequency band and also matching the T2 input are issues that you the builder will have to consider. Certainly the lossy and often non-predictable #43 ferrite material is a factor which might affect your transformer performance. While a trifilar transformer is a little more difficult build for a novice as compared to an auto-transformer, this improved design might work very well at your QTH. Testing like Dave did is certainly the way to go and I greatly appreciate his contribution.

To the right are Dave's excellent bench measurements. He used a Marconi 2018 signal generator, a Racal 9301 RF millivoltmeter and a Bird load on the other port. Kudos to Dave for performing this experiment and contributing to the receive antenna splitter knowledge base.

Dave's web site.

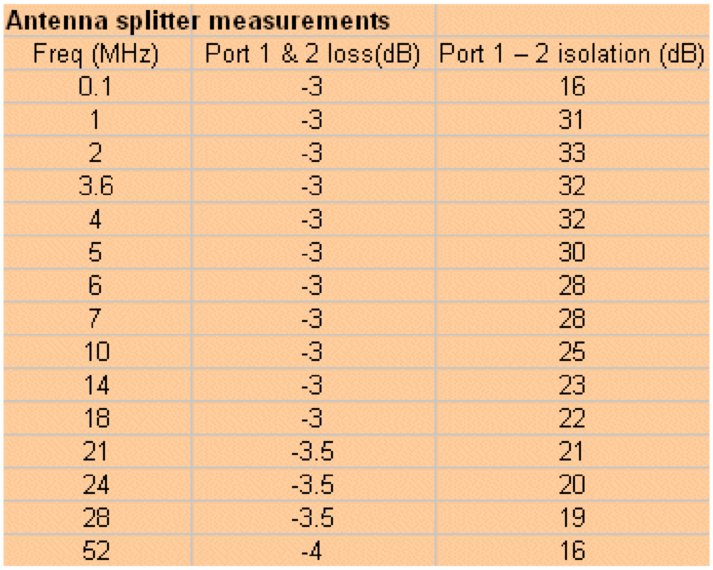

Antenna splitter measurements		
Freq (MHz)	Port 1 & 2 loss(dB)	Port 1 – 2 isolation (dB)
0.1	-3	16
1	-3	31
2	-3	33
3.6	-3	32
4	-3	32
5	-3	30
6	-3	28
7	-3	28
10	-3	25
14	-3	23
18	-3	22
21	-3.5	21
24	-3.5	20
28	-3.5	19
52	-4	16

Conclusion

It is really awesome to be able to connect 2 receivers to the same outdoor antenna. The 3 port network and cabling to the additional receiver does not seem to increase receive noise levels from RFI in the house. Most likely this is due to the fact I am using shielded coax, a shielded project box and have a good RF ground system.

This is a simple project you can build in one evening.
I hope you receive some good DX! 73 es CUL, VE7BPO

Here is a link from F6AOJ

Additional Photos

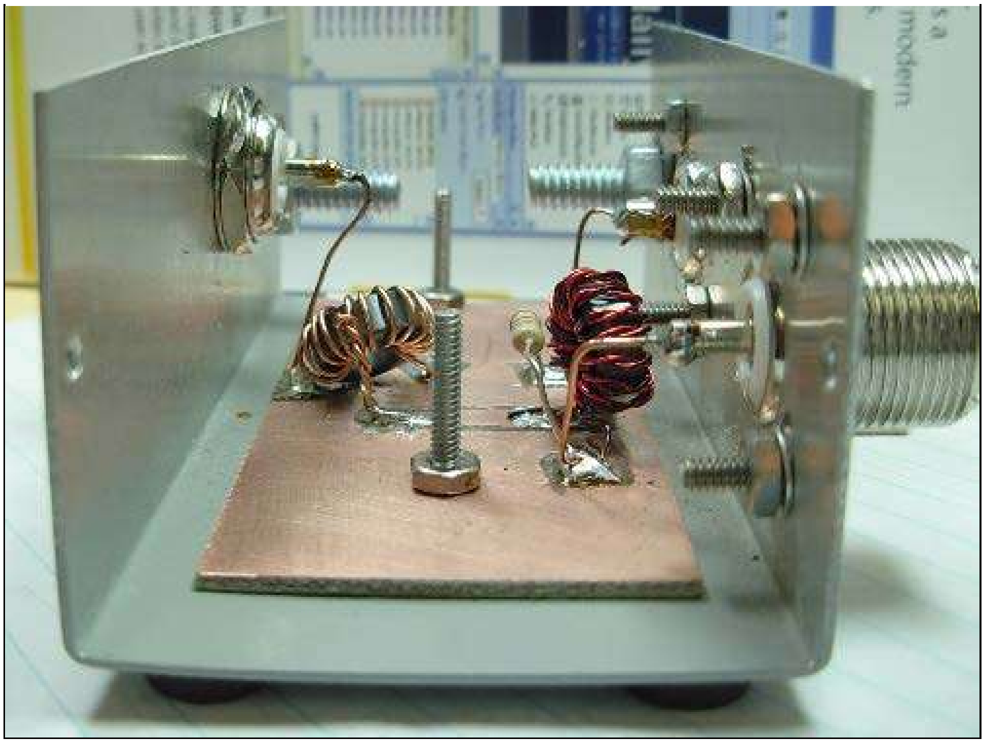

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Medium Frequency TRF Receiver

Introduction

This series of experiments was initiated in 2006, stalled, and was finally completed 16 months later with the inspiration provided by work regarding zero power receivers web-published by Wes, W7ZOI in late summer 2007.

Described is a complete receiver, built and presented backwards from the audio stage to the antenna. The design goals were to build a Tuned Radio Frequency broadcast band receiver with one RF amplifier, a high performance detector and a simple, headphone-level audio stage.

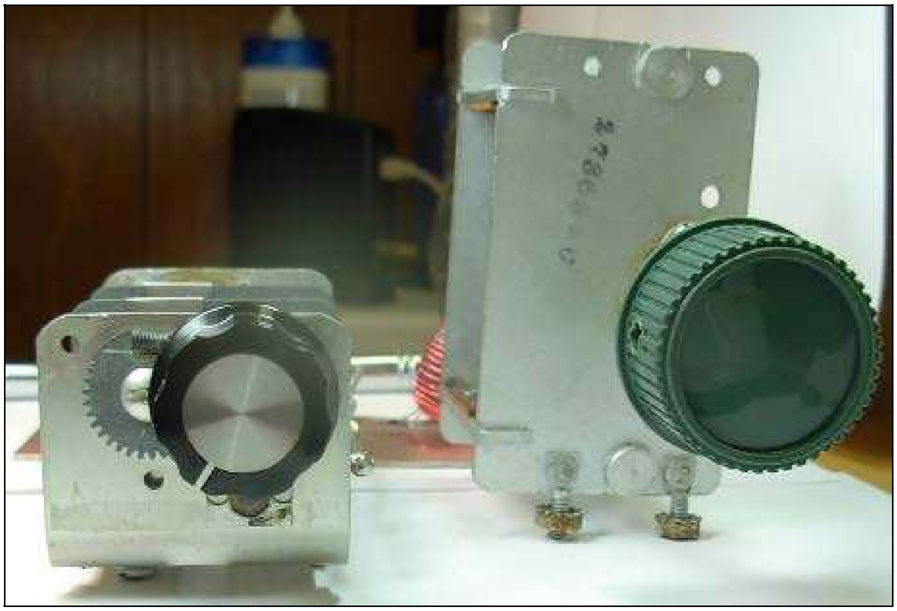

Receiver Block Diagram

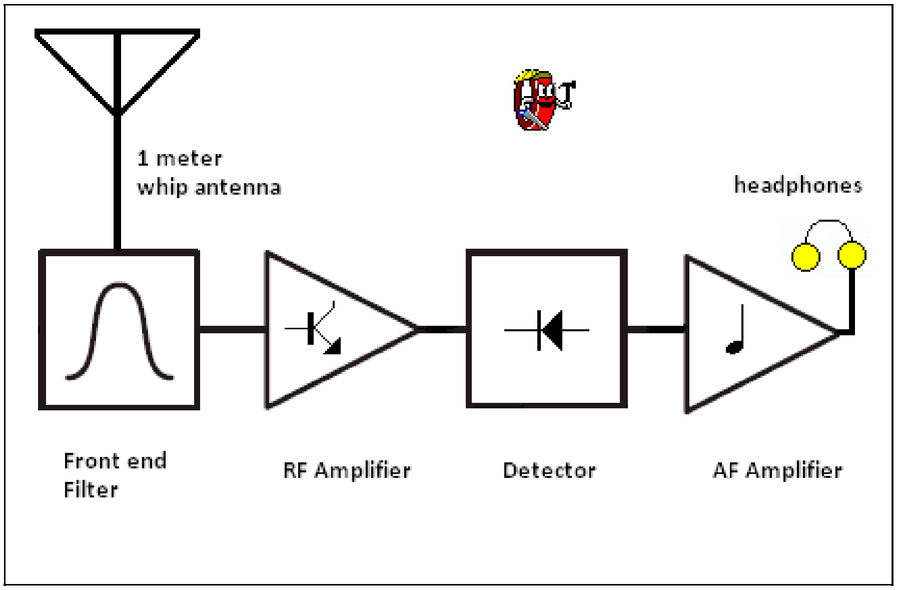

The receiver block diagram is shown to the left. The antenna is a ~ 1 meter long whip purchased from Radio Shack in the USA. A single cascode bipolar junction transistor amplifier boosts the RF voltage and drives an envelope detector which is terminated by a JFET source follower. The source follower connects to a 10K volume potentiometer which controls the AF signal voltage into a headphone-level audio amplifier. Like most of the projects on this site, the intention is to present some circuits and ideas for experimentation. This receiver is designed for local broadcast band AM radio reception, however, the various circuits could be used in or titleered for DX receivers as well.

A Supplemental Page can be found here

Audio Stage

The AF amplifier is a superb design by Rick, KK7B and is featured in many projects in EMRFD. This audio amplifier uses one 5532 op amp and has low noise and high gain. The 220 pF feedback capacitors can be increased to boost the low frequency response. I have built 6 or 7 versions of this stage and have used feedback capacitors up to 560 pF for this purpose. In the audio path, polyester film capacitors were used to try to improve the audio quality. Additionally, the value of the 15 uF capacitor connected to pin 2 is flexible. The quiescent current draw of this stage at 12.2 VDC is 12. 3 mA. Some builders may have to increase the 100 uF filter capacitor on the main 12.2 volt line to overcome motorboat oscillation. None occurred in my breadboard version. I suggest using this audio stage instead of the LM386 or discrete component final audio amplifiers in all projects which call for a headphone-level audio power amplifier on this web site.

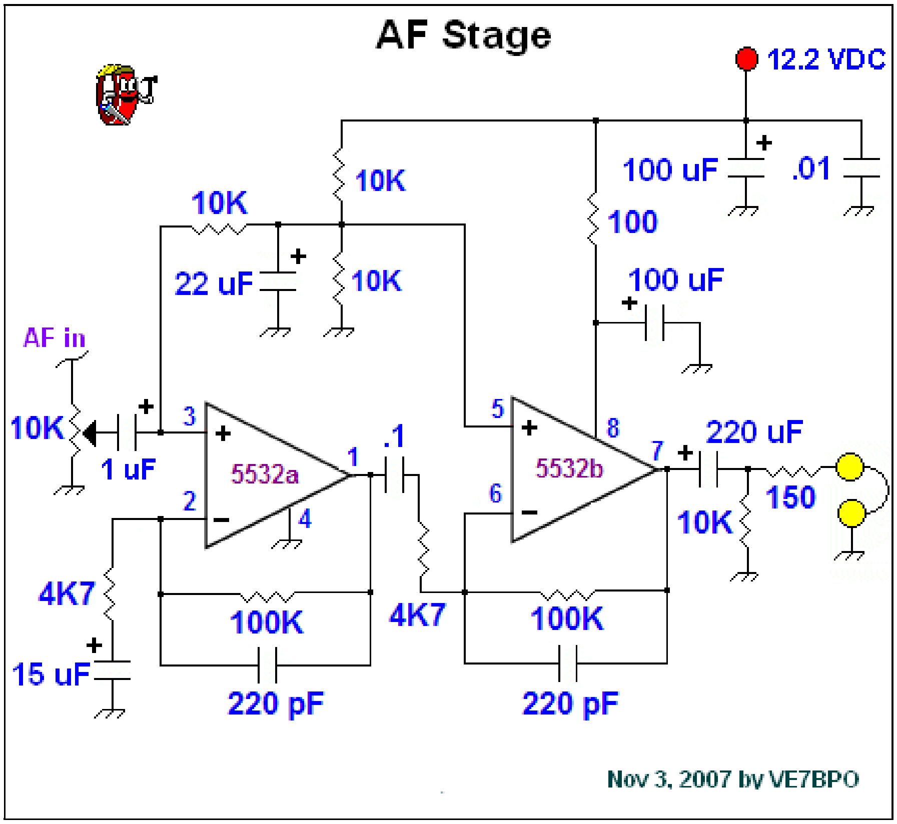
AF Stage
12.2 VDC
10K
10K
22 uF
10K
100 uF
.01
100
100 uF
AF in
10K
1 uF
3
2
5532a
1
4
.1
5
6
8
7
5532b
220 uF
150
10K
4K7
15 uF
100K
220 pF
4K7
100K
220 pF
Nov 3, 2007 by VE7BPO

Close up of the KK7B audio amplifier breadboard from the 10K potentiometer to headphone jack

---

Cascode BJT RF Amplifier and High Performance Detector

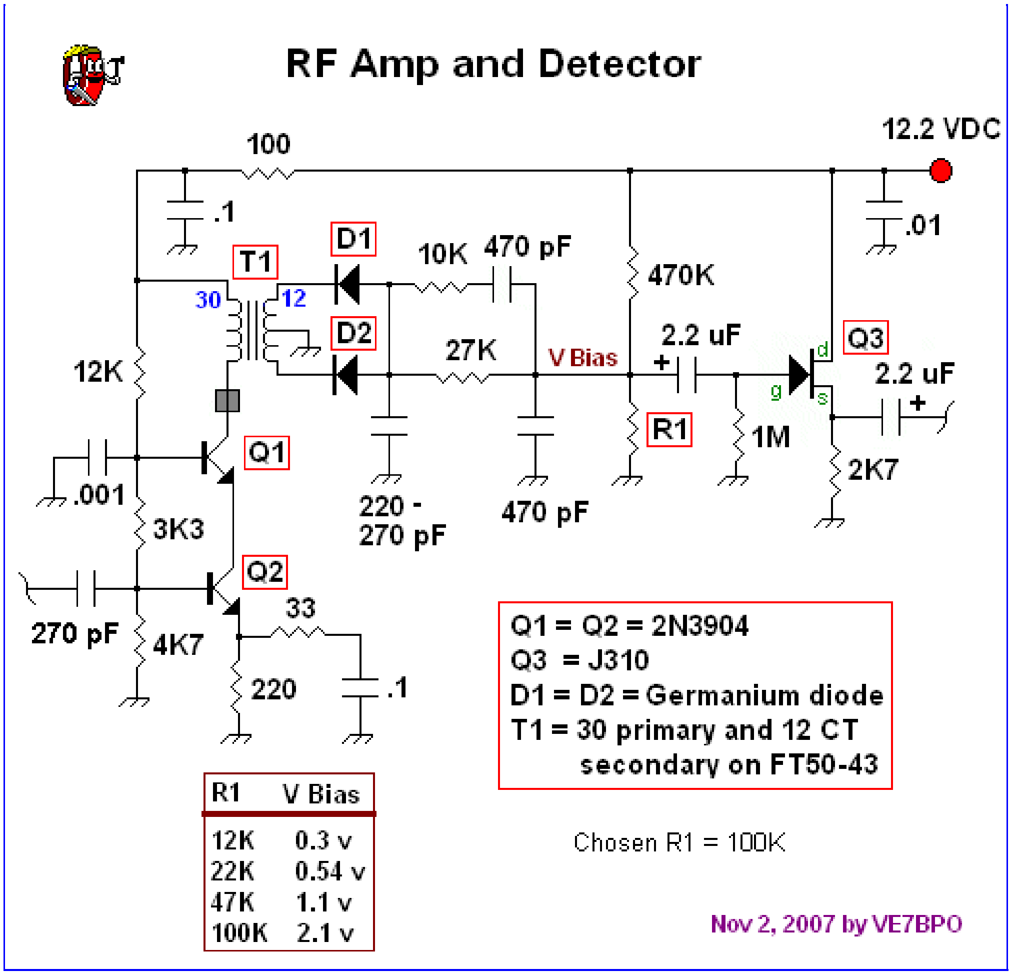
RF Amp and Detector
100
12.2 VDC
.1
.01
T1
D1
10K
470 pF
470K
30
12
D2
27K
V Bias
2.2 uF
+
Q3
d
g
s
2.2 uF
+
12K
R1
1M
2K7
Q1
.001
3K3
220 -
270 pF
470 pF
Q2
33
270 pF
4K7
220
.1

Q1 = Q2 = 2N3904
Q3 = J310
D1 = D2 = Germanium diode
T1 = 30 primary and 12 CT
secondary on FT50-43

R1	V Bias
12K	0.3 v
22K	0.54 v
47K	1.1 v
100K	2.1 v

Chosen R1 = 100K

Nov 2, 2007 by VE7BPO

Above is the combined RF amplifier, detector and JFET source follower schematic.

Cascode BJT RF Amplifier

To the left is a simplified RF amp diagram taken from the schematic above indicating the measured DC voltages for reference purposes.

T1, the output transformer was wound on an FT-50-43 ferrite toroid. An FT37-43 would also be suitable. Number 28 gauge enamel coated magnet wire was used for the 30 turn primary and 26 gauge wire was wound over top to make the secondary 12 turn, center-tapped winding. The 26 gauge wire was used for the secondary winding because it provided good anchoring of the transformer by the center tapped ground connection.

You may consider substituting a 22 to 100 ohm resistor for the ferrite bead on Q1. It suppresses VHF oscillations.

Detector

To the right is a photograph of the detector from the Q1 transformer through to the JFET source follower. Schottky/hot carrier diodes or germanium diodes such as the 1N34A with a low forward voltage drop are strongly recommended. I have found there to be significant variation in

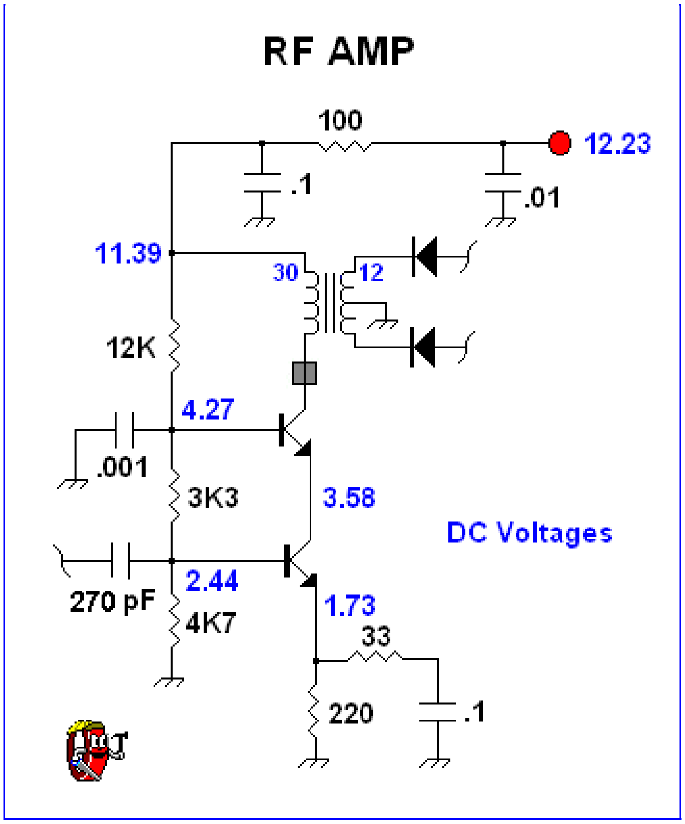

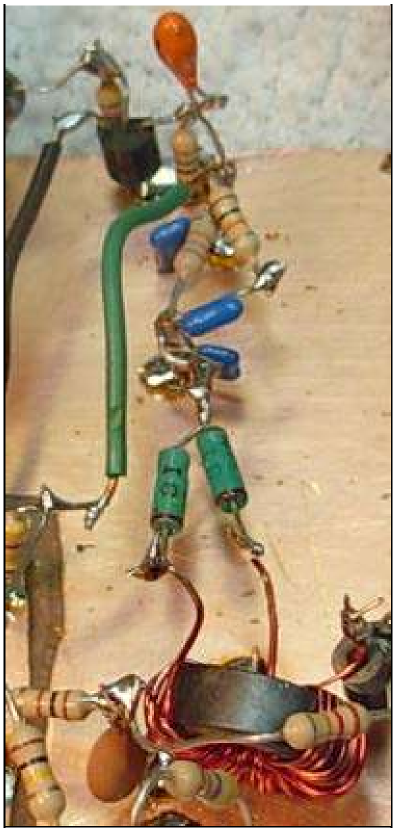

sensitivity between different types of these diodes. The 2 germanium diodes I used were matched as described on this web page. A number of detectors were built and tested for this receiver, however, the design shown had the best audio quality when compared to the others. The virtues of this detector include low noise, high bandwidth, high sensitivity and low distortion. although a little complex, this is a detector worthy of consideration in your AM receiver projects. The center-tapped Q2 transformer secondary and the 2 diodes provide full wave detection. This serves to reduce distortion somewhat and cancel even-order harmonics in the carrier signal. You may eliminate one of the diodes and convert the Q2 transformer secondary to a conventional, single link.

A 470K ohm resistor and R1 form a voltage divider that sets the detector bias voltage (V Bias). Some measured R1 values and corresponding bias voltages are shown in the schematic. I chose an R1 value of 100K for my final version. You may have to increase or decrease the R1 value to suit your local detector sensitivity requirements. You could also substitute a bias potentiometer for front panel adjustment of the receiver sensitivity. In this detector, changing the R1 value also changed the detector frequency response. I built a separate voltage divider with roughly the same V bias consisting of a 68K and a 15K resistor and swapped it for the 470K and 100K pair. Interestingly, the 470 K and 100K pair had better low frequency response and slightly higher sensitivity than the 68K and 15K voltage divider.
Diode detectors are best driven with a high impedance source and followed by a high impedance load. Q3, a simple JFET source follower provides a high impedance load. You might want to substitute a "popcorn" MPF102 for the high Idss J310 indicated in the schematic as a J310 is not really required here. If you substitute a MPF102, please increase the source resistor from 2K7 to 4K7 ohms.

Front-end Band Pass Filter and Antenna

In late summer 2007, Wes, W7ZOI conducted experiments with zero power receivers (crystal sets and such). He wound some inductors using ferrites with an unloaded Q of over 270 at MF! Please check out Wes' web site. His work with high Q ferrite inductors illustrates the importance of quantitative measurement and also provided the following revelation; we really do not have to resort to large, air core, Litz wire coils to build high-performance inductors at MF! The early prototype front end for this project was built using FT50-61 ferrite cores, however after Wes emailed me his work on zero power

receivers, I had to get some FT-114-61 ferrites for the front end of this receiver. The next day, I emailed Mark Laurain from Amidon Associates Inc and ordered some FT-114-61 ferrite toroids. The arrival of these ferrites prompted me to finish this project and put it up on the web.

The schematic on the right is the final band pass filter used for the front end. I initially tried using just L3 for the front end, but I was unable to just tune a single station. In my city, there are 2 powerful AM radio stations at 630 and 1150 KHz. With a single inductor, I could peak one of the stations, but the other could be heard in the background. Thus, the double-tuned band pass filter presented was designed and built. Now only one station can be detected with this circuit and tuning is sharp. Most builders would use a dual-ganged variable tuning capacitor, however, I elected to use 2 separate variable capacitors. Considerable flexibility with this circuit is possible. You will have to experiment to best determine your local sensitivity versus selectivity needs and to suit the variable capacitors you have available. Large AM receiver capacitors are getting hard to find. I obtained the 2 variable capacitors shown in the photographs below from 2 old receivers found in a second hand store. One of the receivers was a Marconi tube radio that was in poor condition. I paid $5.00 for both radios and harvested the 2 beautiful variable capacitors as well as some other parts such as knobs, switches and terminal strips. Never pass up on an old, derelict radio as a potential variable capacitor source!

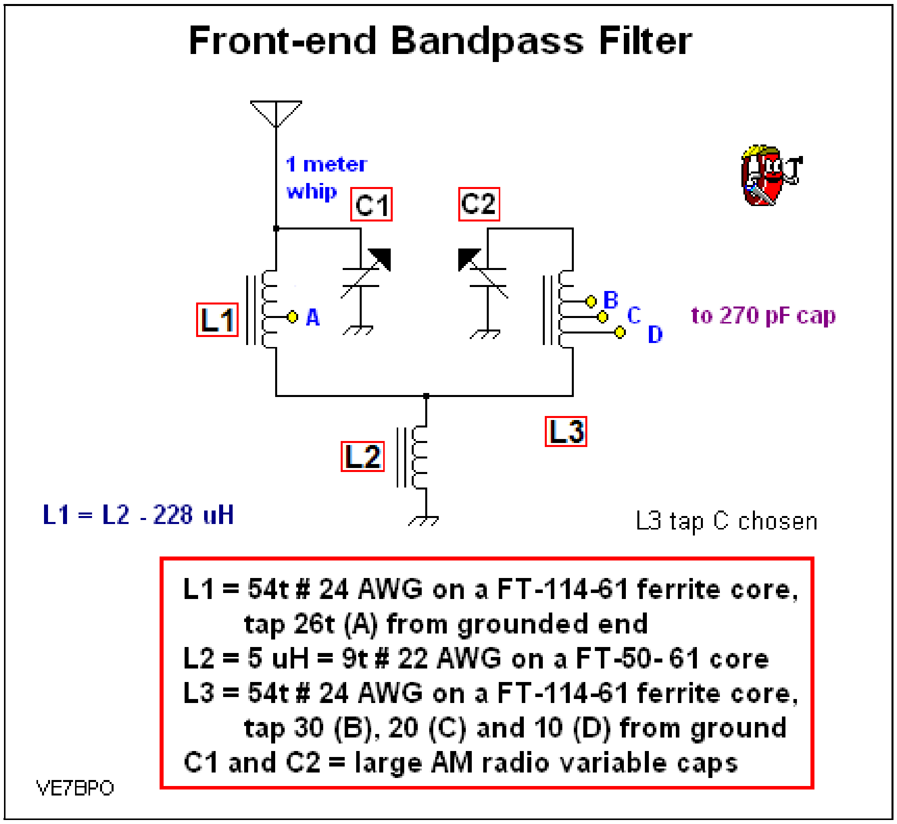
Front-end Bandpass Filter
1 meter whip
C1
C2
L1
A
B
C
D
to 270 pF cap
L2
L3
L1 = L2 - 228 uH
L3 tap C chosen
L1 = 54t # 24 AWG on a FT-114-61 ferrite core, tap 26t (A) from grounded end
L2 = 5 uH = 9t # 22 AWG on a FT-50-61 core
L3 = 54t # 24 AWG on a FT-114-61 ferrite core, tap 30 (B), 20 (C) and 10 (D) from ground
C1 and C2 = large AM radio variable caps
VE7BPO

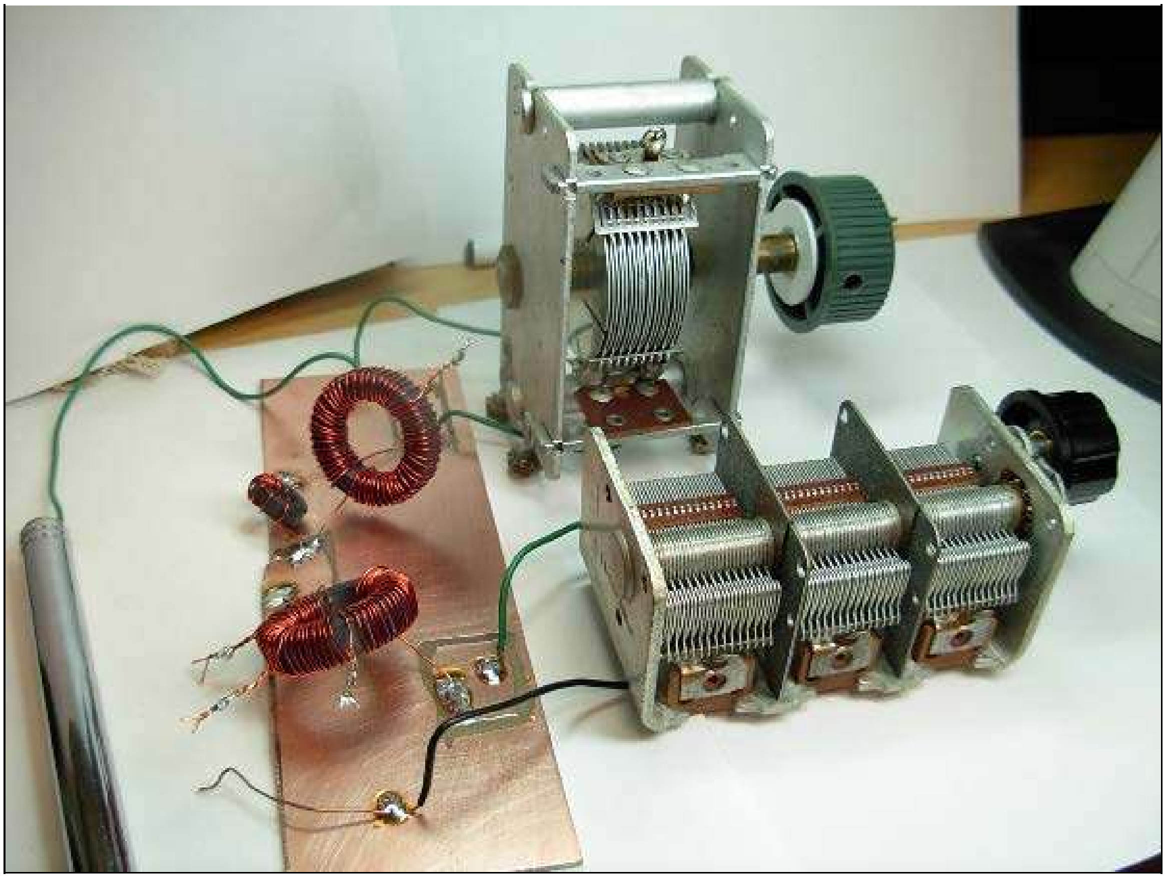

Shown above are 2 photographs of the band pass filter breadboards. The 2 variable capacitors had a variation of ~ 24 to 500 pF. There is considerable interplay between the 2 capacitors. For my QTH, it was better to peak C2 first and afterward to peak C1. Consider that L1 and L2 have a hot end and a grounded end. The antenna is connected to the the hot end of L1. Predictably, when substituting the L1 center tap as the antenna connection, the selectivity of the L1-C1 tank is increased and the sensitivity or received signal strength is reduced. This also occurs when testing the various tap points on L3 to feed the RF amplifier-detector stages. In the final circuit, I settled on Point C, 20 turns from ground. Using Point D, reduces sensitivity and increased selectivity. The opposite is true when using Point B. You the builder, have to determine which L1 and L3 connection points to use based on your own experiments and local factors. You may also change the receiver sensitivity by making changes such as increasing or decreasing the 270 pF coupling capacitor value, the emitter degeneration on Q2, or the detector bias.

Band pass Filter Analysis

It is impractical to sweep a BCB band filter using variable capacitors, so some analysis using GPLA, a program that ships with EMRFD was used to plot and better understand the double-tuned band pass filter response. A worst case inductor unloaded Q of 200 was used, but I imagine that the actual Q of L1 and L3 is much higher. For the source impedance, 100K ohm was used conjecturing that a short whip antenna at 1150 KHz would have a very high input impedance and not load down the L1 inductor. In reality, it is likely the antenna input impedance might be closer to 1 Megohm, however, I am using the worst case scenario. If the filter performs better than simulated - all is great! Higher source and load impedances and higher unloaded Q inductors would decrease the bandwidth of this filter which is desirable.
Note that I am concerned that L2 at 5 uH may may overcouple the 2 tank circuits. I did not see a double humped response on GPLA analysis, however, experimentation with L2 may be in order for the more astute homebuilder. You might consider lowering the L2 value to 3 or 4 uH and performing some testing. The load impedance for L3 was rather arbitrarily chosen. Considering that various taps on L3 may be used, the XC of the 270 pF coupling capacitor and the input impedance of the RF amplifier, I just chose 47K as the L3 load impedance. Below are 2 screen captures of GPLA plots. The top graph is the double tuned band pass filter and below it is the single tuned band pass filter consisting just of L3 and C3.
These graphs lead to 2 main conclusions:

1. The final band pass filter design appears to be reasonably sound.
2. We can understand why I could not tune in a single radio station with just L3 and C3 as the band pass filter; the filter skirts are not very steep and the second unwanted station was also amplified and detected.

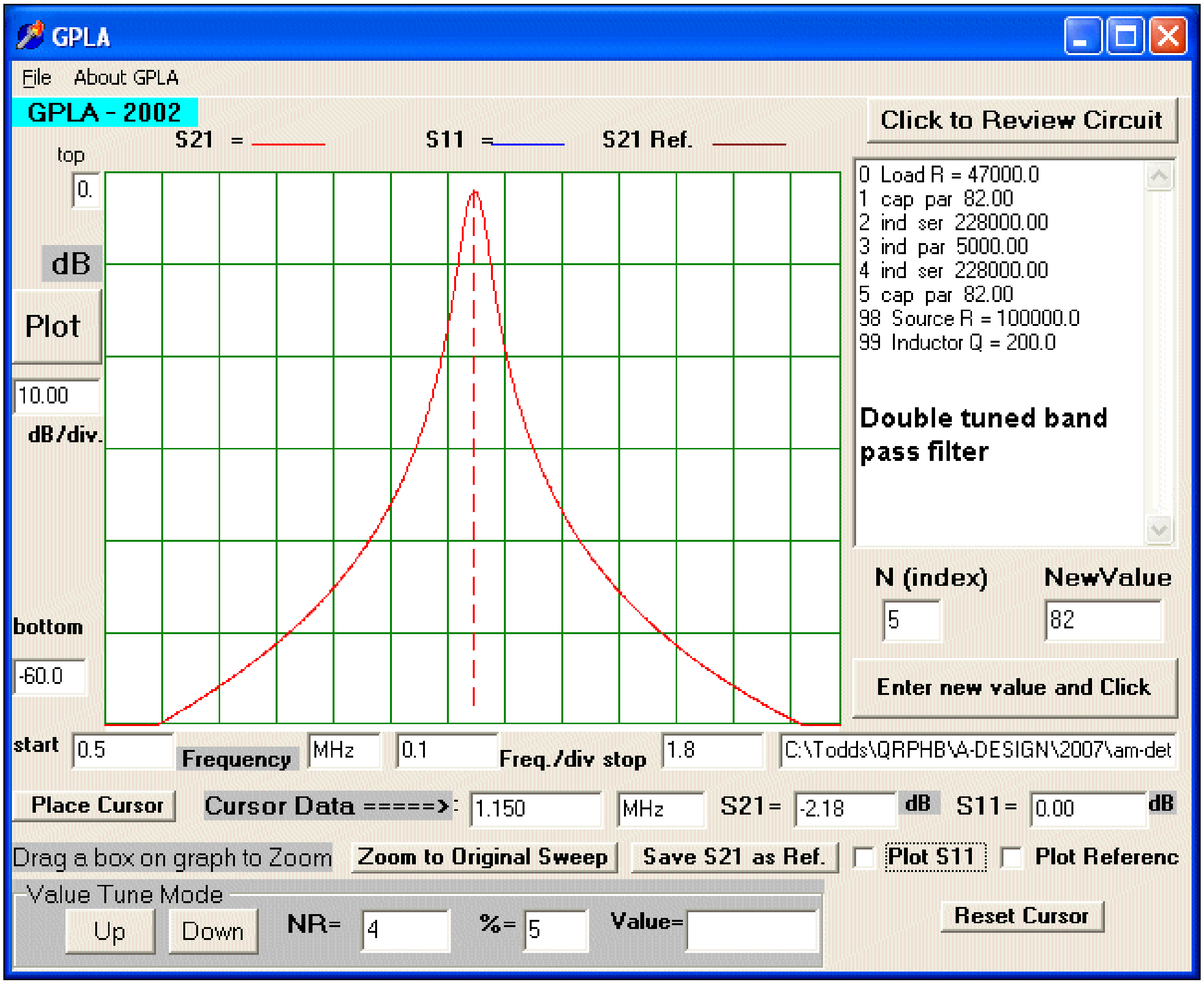

```
0  Load R = 47000.0
1  cap  par  82.00
2  ind  ser  228000.00
3  ind  par  5000.00
4  ind  ser  228000.00
5  cap  par  82.00
98  Source R = 100000.0
99  Inductor Q = 200.0
```

Double tuned band pass filter

top: 0.

dB

10.00 dB/div.

bottom: -60.0

start: 0.5 Frequency MHz

Freq./div: 0.1

stop: 1.8

C:\Todds\QRPHB\A-DESIGN\2007\am-det

N (index): 5

NewValue: 82

Cursor Data =====>: 1.150 MHz

S21= -2.18 dB

S11= 0.00 dB

☐ Plot S11

☐ Plot Referenc

NR= 4

%= 5

Value=

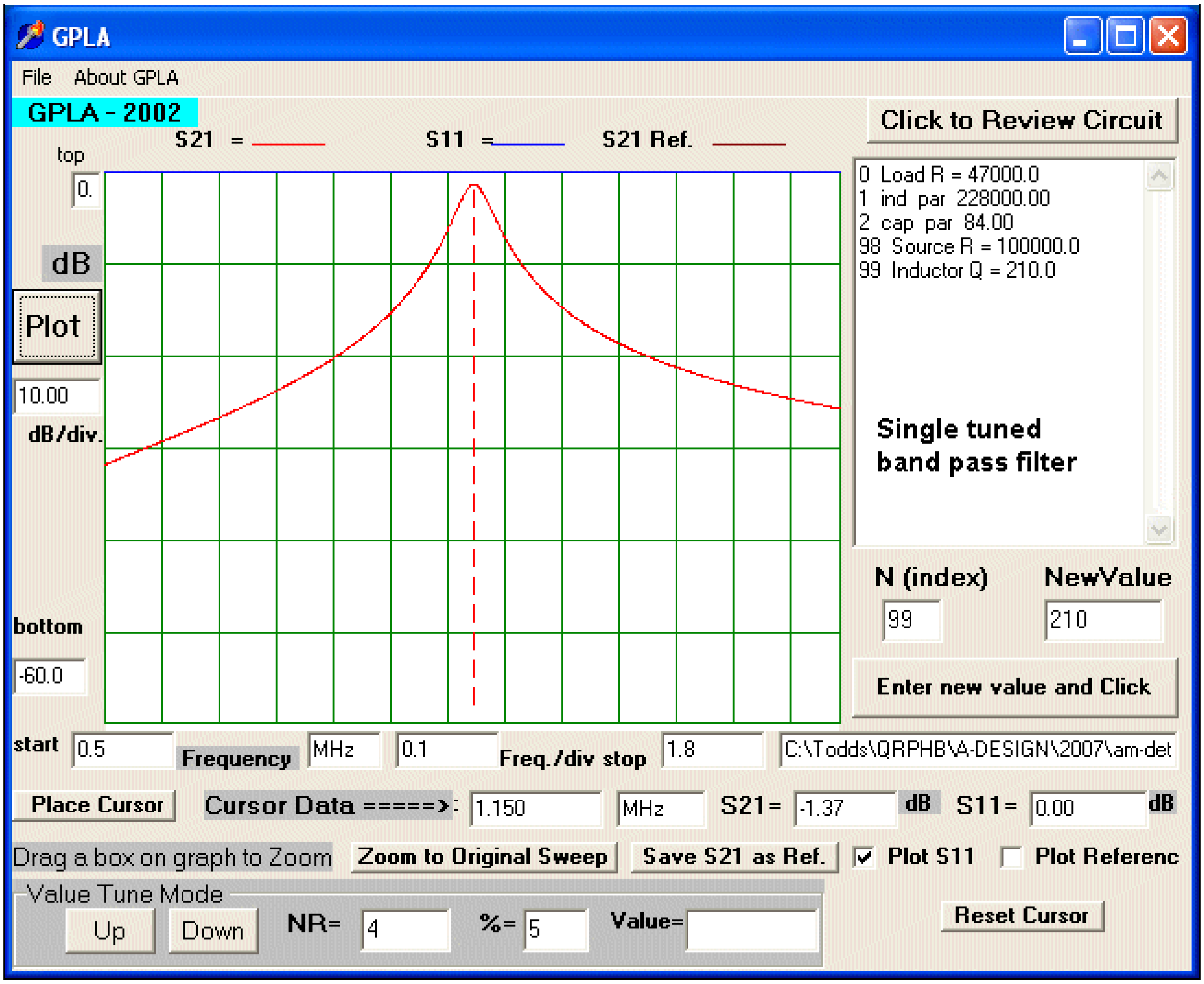
GPLA
File
About GPLA
GPLA - 2002
S21 =
S11 =
S21 Ref.
dB
Plot
Click to Review Circuit
Enter new value and Click
Place Cursor
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Value Tune Mode
Up
Down
Reset Cursor

top 0.

10.00 dB/div.

bottom -60.0

0 Load R = 47000.0
1 ind par 228000.00
2 cap par 84.00
98 Source R = 100000.0
99 Inductor Q = 210.0

Single tuned band pass filter

N (index) 99
NewValue 210

start 0.5 Frequency MHz 0.1 Freq./div stop 1.8 C:\Todds\QRPHB\A-DESIGN\2007\am-det

Cursor Data =====> 1.150 MHz S21= -1.37 dB S11= 0.00 dB

☑ Plot S11 ☐ Plot Referenc

NR= 4 %= 5 Value=

Varactor Tuned Front-end Filter

On November 11, 2007, I decided to investigate whether or not variable capacitance or varactor diodes could effectively replace the air variable capacitors in the band pass filter. In my parts cabinet were several MVAM -109 which is an obsolete but still readily available part. Another varactor, especially designed for tuning AM receivers is the 1SV149. This varactor is manufactured from Toshiba and is also appropriate.

While not comparable to the Q of 300 or greater of a good quality air variable capacitor, varactors are smaller, cheaper and can be easily ganged together so that only 1 potentiometer is required to tune the front-end filter. To the right is a photo of the varactor tuned front-end filter breadboard.

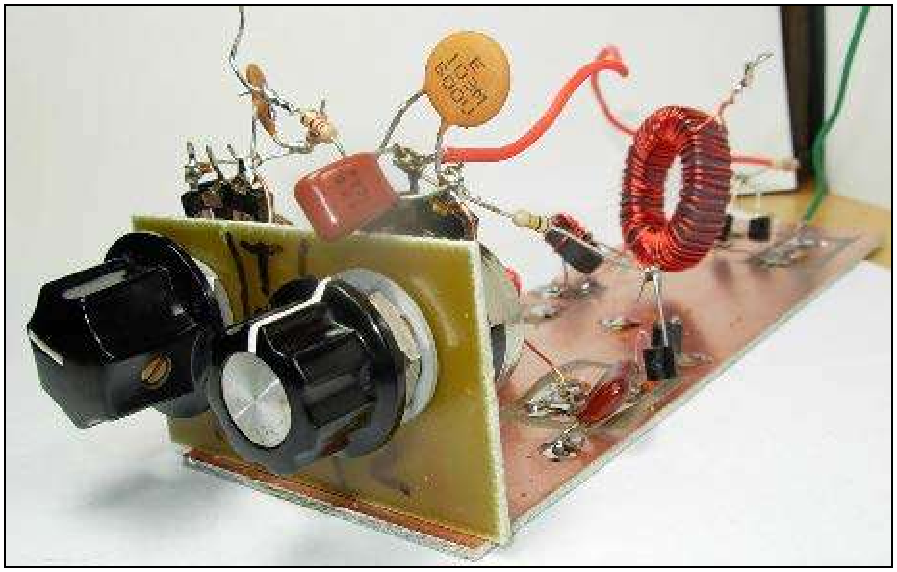

To the left is the schematic of the varactor tuned front-end band pass filter. The air variable capacitors were unsoldered from the original filter breadboard and a small board drilled and fitted with two 250K potentiometers was soldered to it.

Varactor Controlled Bandpass Filter

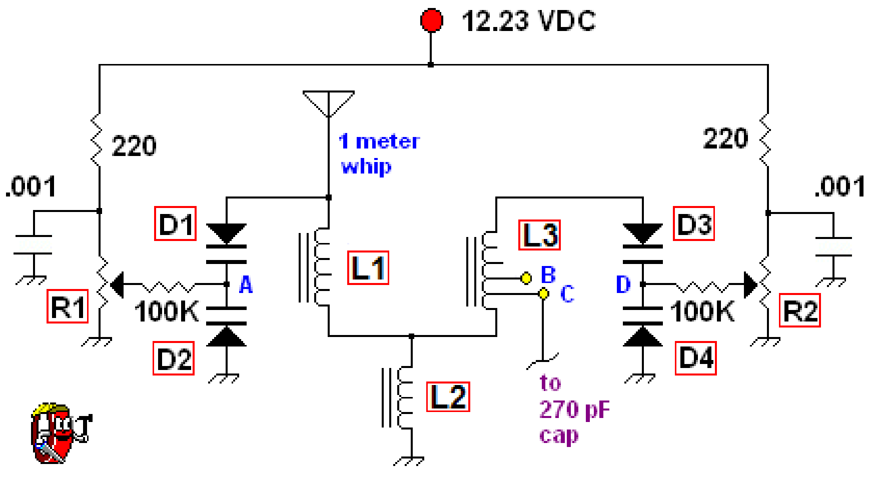

Tuning voltages	630 KHz	A = 0.76v	D = 1.0v
	1150 KHz	A = 4.42v	D = 4.77v

L1 = 228 uH = 54t FT-114-61 ferrite torroid
L2 = 5 uH = 9t FT-50-61 ferrite torroid
L3 = same as L1 with tap 15t (B) and 10t (C) from ground
R1 = R2 = 250K potentiometer
D1 = D2 = D3 = D4 = MVAM-109 varactor diode

L3 was also modified to have taps at 10 and 15 turns from ground. I conjectured that since the varactor diodes have less Q than their air variable cousins, it would be wise to tap down on L3 to try and increase the selectivity of the L3 tank circuit. In the end, I used the tap at 10 turns from ground for my receiver as signal strength was still acceptably strong. You may choose to use the tap at 15 or some other point to suit your local selectivity/sensitivity requirements.

I was able to tune in single stations as I previously did with the air variable capacitors. Tuning is "touchy". Ten-turn pots would be a better choice, however, are not very frugal for such a project. You get used to tuning with conventional potentiometers after a few minutes or so. I measured the reverse voltages required to tune the 2 main local AM radio stations and they are tabled in the schematic. The L1 tank requires slightly more capacitance to resonate than the L3 tank. Thus it takes a little less applied reverse voltage to the varactor pair resonating the L1 tank compared to the varactor pair resonating the L3 tank .A side view photograph of the varactor breadboard is shown directly below. The component leads have been kept long so that I can recycle parts from experimental project to project as possible. This helps contain costs. Shorter lead lengths and proper lay out should be pursued in any final projects you build.

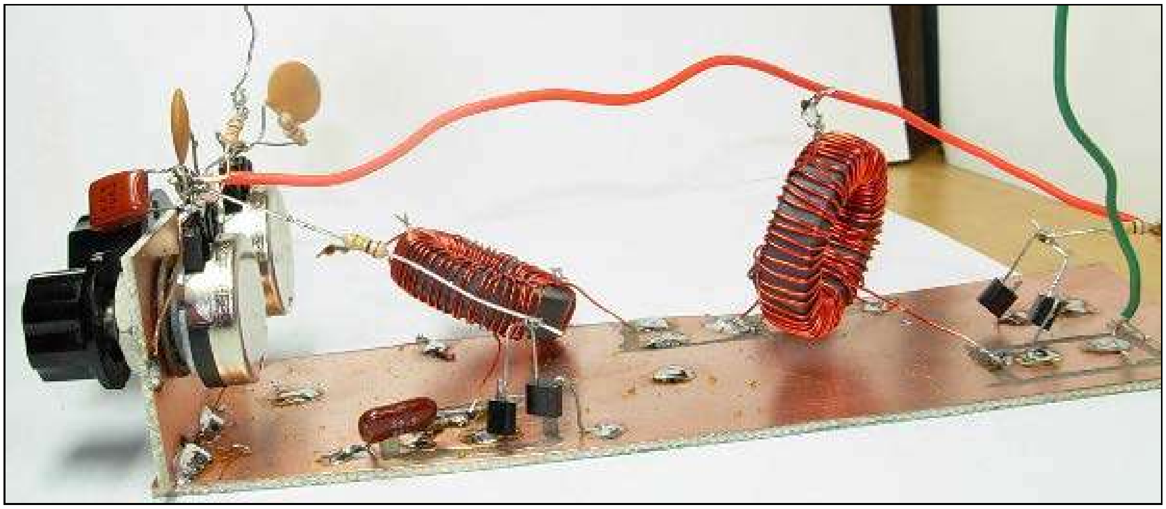

Single Varactor Tuned Front-end Filter

Tuning with a single potentiometer ganged to both varactors is easy to do after learning from the experiment above. All that is required is a method to compensate for the differences in capacitance between the the 2 LC tanks. I placed a high-Q (Q=300) variable trimmer capacitor in parallel with L1. By listening to the receive signal strength and tuning in one radio station using the potentiometer, I was able to peak CV for the strongest signal. I did this for both 650 and 1150 KHz and actually unsoldered CV and measured its value with a capacitance meter. The CV value was ~ 6 pF for both frequencies. I decided to replace CV with a fixed 5 pF silver mica capacitor and left it there in my final filter version. Your results will probably be different. I suggest just leaving CV

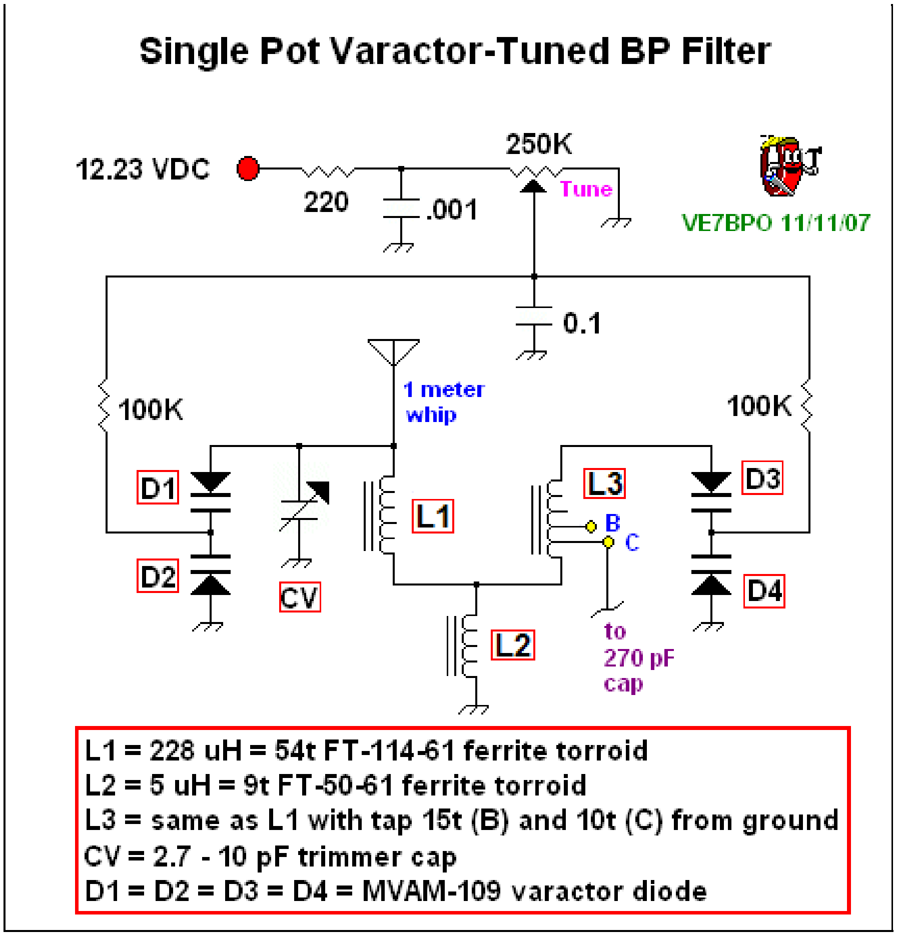

L1 = 228 uH = 54t FT-114-61 ferrite torroid
L2 = 5 uH = 9t FT-50-61 ferrite torroid
L3 = same as L1 with tap 15t (B) and 10t (C) from ground
CV = 2.7 - 10 pF trimmer cap
D1 = D2 = D3 = D4 = MVAM-109 varactor diode

and using this trimmer cap to peak the signal once you have tuned a desired radio station with the main tuning potentiometer. An alternative to using CV is to vary L1. You could try compressing the number of L1 windings to allow tracking of the 2 LC tank circuits.

For the varactors, I used back-to-back VVC diodes as opposed to just a single varactor to resonate each tank. This was done in an effort preserve the highest varactor Q possible. The RF voltage of the AM RF signal may be high enough to forward bias a single varactor during a portion of the AC signal and degrade Q. This does not happen when back-to-back diodes are used. Almost all high-grade FM tuner schematics I have seen use back-to-back varactor diodes in their various ganged, tunable band pass filters. The major drawback of back-to-back diodes is your tuning range is reduced because you now have 2 capacitors in series. Experimentation may be required to achieve the BCB band-spread that you desire. You can add another pair of varactors in parallel or add some parallel fixed capacitance or even change the L1 and L3 inductance values for example.

This receiver tunes nicely and sounds fabulous. Last evening I was able to tune in 5 different AM stations, however, other than the local 2 radio stations, the others were quite faint. This is not bad considering this receiver has only 1 RF amp and a 1 meter long antenna. This band pass filter could be adapted as a pre-selector for AM radio reception. To match 50 ohms, lower L1 and L3 tap points could be chosen or a few links of wire may be wound around the inductors.

In the photograph below, you can see the 5 pF capacitor soldered in parallel with the MVAM-109 pair associated with L1. The antenna also connects to the ungrounded end of the 5 pF capacitor. Below in the last photograph; since only one potentiometer is used for tuning, a large knob was screwed on to the pot control shaft to make tuning a little easier. The solder-laden 220 ohm resistor is the connection point for the regulated 12.2 VDC. The 220 ohm resistor on the left has been cut from the 12.2 VDC connection point so 0 voltage goes to the left potentiometer.

Final Thoughts

I emailed Wes, W7ZOI and asked him why it is better to inductively couple a tuned circuit which use air variable tuning capacitors. Wes wrote his answer in the form of a complete web article entitled Coupling Methods in the Double Tuned Circuit. Big thanks Wes! From his summary, when the inductors used to resonate each tuned circuit are constant, and inductive coupling is used, the coupling of the resonators will remain constant as the variable capacitors are tuned across the band. Please download and study his web article for it not only discusses coupling in the double tuned circuit, but provides some insight into using his LadBuild and GPLA software from EMRFD.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Junk Box NDB Low Pass Filter

Introduction

With winter approaching, many HAM and SWL hobbyists find intrigue in tuning in NDB or **N**on-**D**irectional **B**eacons. Although the tunable NDB band depends on your location, in Canada it may be found in a band ranging from about 190 to 535 KHz. Canadian beacons either have just a carrier (no offset)or are tuned using the USB with about a 400 Hertz offset, however, different offset frequencies and certainly LSB are used when receiving DX from other countries.

Less than 10 Km away from my QTH is a 10 KW AM radio station at 1150 KHz. On my test receiver, the S-meter reads off the scale (> 60 dB over S-9) when tuned to this frequency. This local radio station causes spurious, second-order intermodulation products (direct mixing) that all but wipes out some weaker NDB stations that I am trying to tune in. Certainly, having a 500 Hertz crystal IF filter is useful, but attenuating this local QRM is also desirable and is the topic of this web page. Many general coverage receivers offer limited or in some cases no filtering of the NDB band, however an outboard filter is an easy project to build in one afternoon.

Update Oct 11, 2010: Here is a link to a version of this project built by Robert, K5TD

Project Schematic

To the left is the project schematic. It seems odd to build a low pass filter to reduce BCB interference (as usually a high pass filter is required for this purpose at HF) however for NDB, an aggressive low pass filter is required. For simplicity sake, a 7 element Chebychev low pass filter was chosen. Since it is easy to wind reasonably high-Q inductors for 10 uH and greater inductance using number 61 material on a ferrite torroid, the FT50-61 core was chosen for all of the inductors. Number 22 gauge wire was used for the coils to keep the unloaded Q as high as possible. The FT37-61 ferrite is also suitable, but will have less Q and require smaller gauge wire. Use 19 turns instead of 17 for the 20.2 uH and 21 turns instead of 19 for the 24.1 uH coil. Do not use number 43 material ferrite cores.

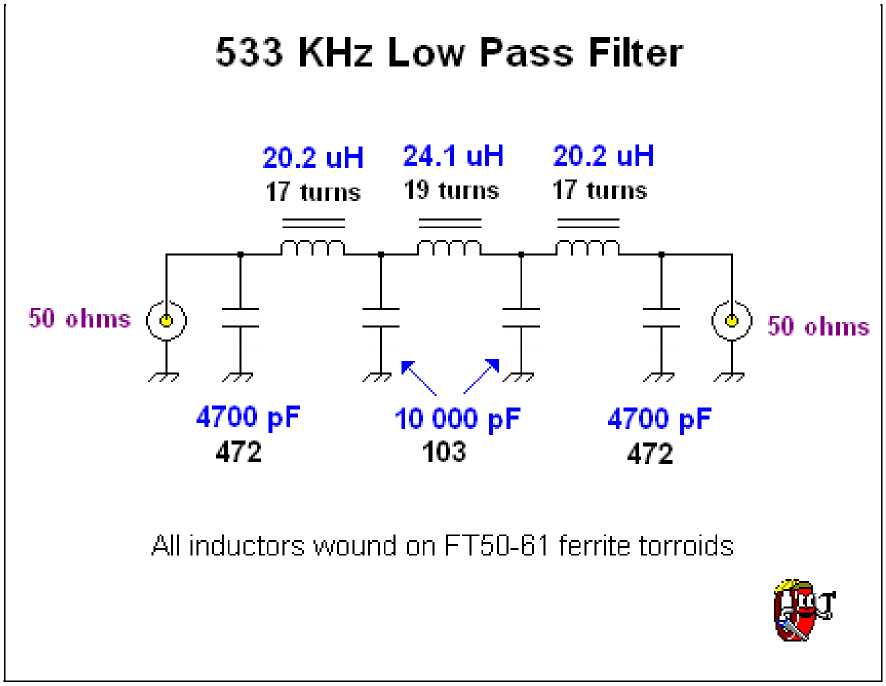
533 KHz Low Pass Filter
20.2 uH
17 turns
24.1 uH
19 turns
20.2 uH
17 turns
50 ohms
50 ohms
4700 pF
472
10 000 pF
103
4700 pF
472
All inductors wound on FT50-61 ferrite torroids

Components

I do not stock RF capacitors greater than 2200 pF, so junk box ceramic capacitors were used to build this filter. In fact, this design specifically uses more common, standard value capacitors to reduce cost and to not have to order in parts. Certainly, the astute builder could use higher quality capacitors or even large powdered iron torroids instead of the ferrite cores for inductors if higher performance is desired. Try to use high Q capacitors if you can find or are purchasing them. Poly or silver mica caps would be great choices. You can substitute a 5000 pF capacitor for the 4700 pF called for in the schematic.
To the right is a photograph of the components I used in the project breadboard.

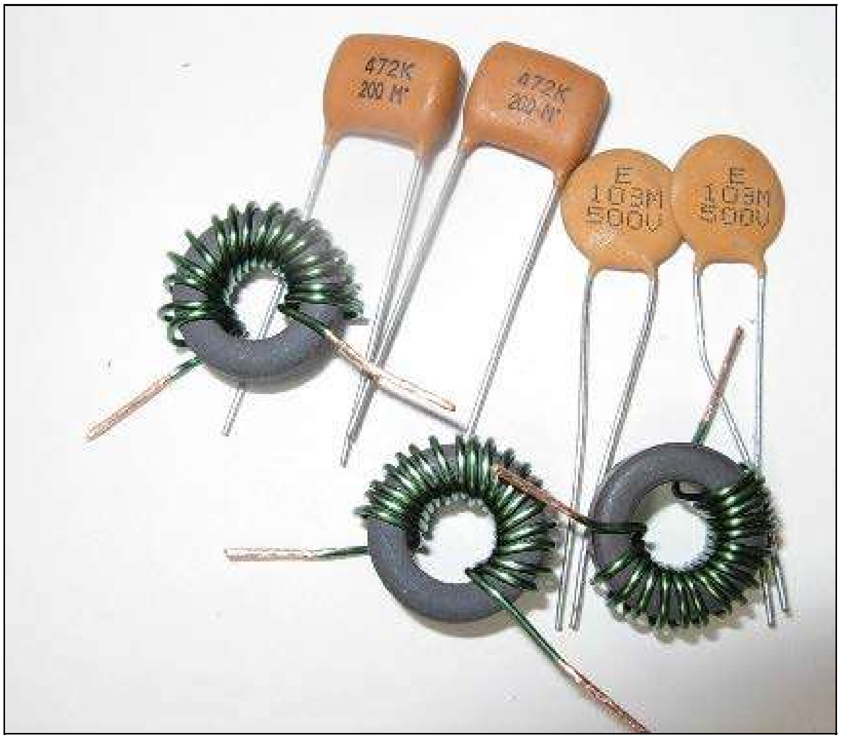
472K
200 M*
472K
200 M*
E
103M
500V
E
103M
500V

Breadboard

To the left is the completed project. Ugly construction as always, was used. The inductors were spaced at least 2.5 cm (1 inch) apart at right angles to try to minimize unwanted coupling.

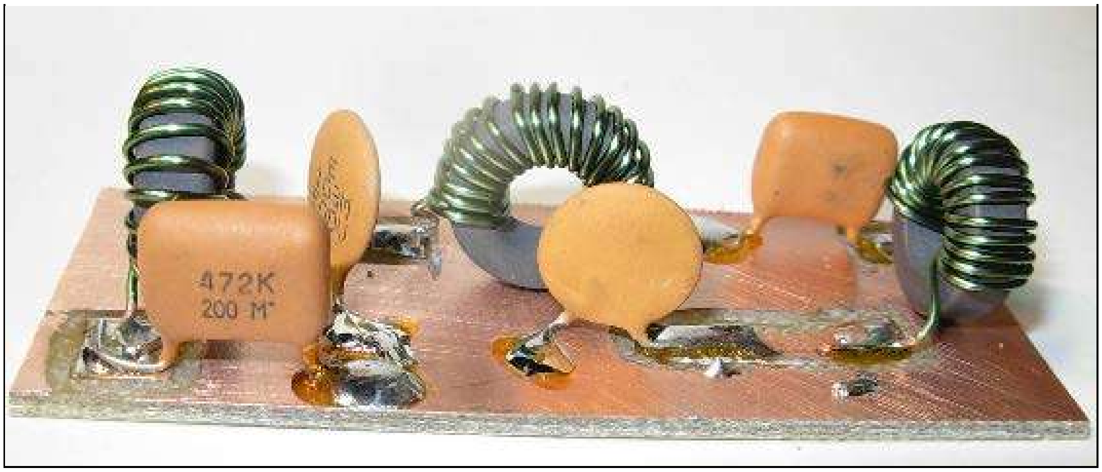

GPLA Simulation

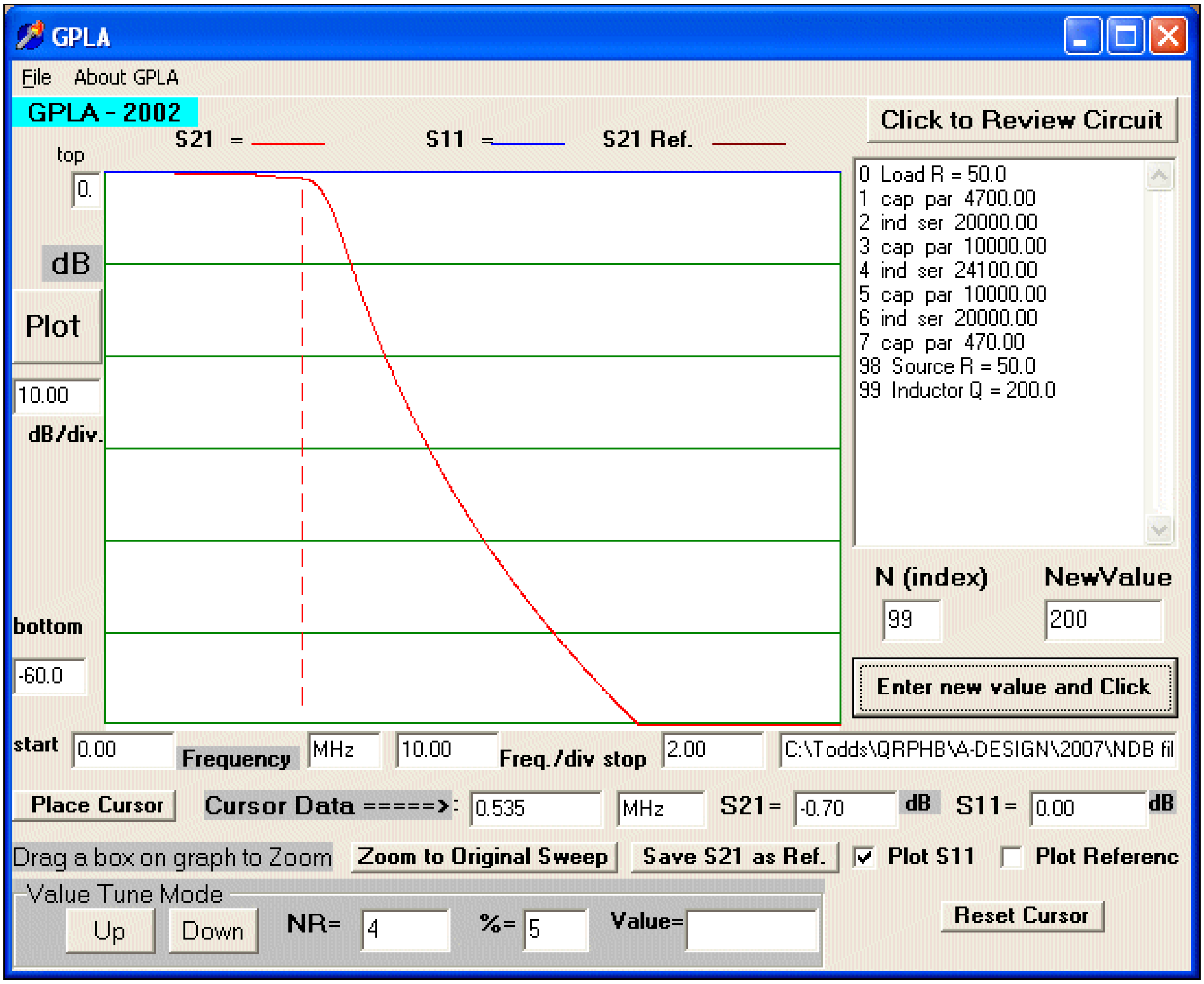

GPLA
File
About GPLA
GPLA - 2002
S21 =
S11 =
S21 Ref.
dB
Plot
Click to Review Circuit
Enter new value and Click
Place Cursor
Cursor Data =====>:
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Value Tune Mode
Up
Down
Reset Cursor

top 0.

10.00 dB/div.

bottom -60.0

```
0  Load R = 50.0
1  cap  par  4700.00
2  ind  ser  20000.00
3  cap  par  10000.00
4  ind  ser  24100.00
5  cap  par  10000.00
6  ind  ser  20000.00
7  cap  par  470.00
98  Source R = 50.0
99  Inductor Q = 200.0
```

N (index) 99 NewValue 200

start 0.00 Frequency MHz 10.00 Freq./div stop 2.00 C:\Todds\QRPHB\A-DESIGN\2007\NDB fil

Cursor Data =====>: 0.535 MHz S21= -0.70 dB S11= 0.00 dB

☑ Plot S11 ☐ Plot Referenc

NR= 4 %= 5 Value=

Above is the plot of the filter during simulation with GPLA. The simulation calculated an attenuation of ~46dB at 1150 KHz. At 1000 KHz the signal was 40 dB down, at 800 KHz it was ~24 dB down and at 630 KHz, the attenuation was only ~5.7 dB! Clearly this filter is not suitable if the offending BCB interference is from a station significantly less than 0.8 MHz. For my situation, this filter is acceptable. A 5 element Chebychev filter was also designed and plotted but was discarded as there was only 32 dB attenuation at 1150 KHz.
Since I wanted to tune as high as 535 KHz, the 533 KHz cutoff frequency was chosen. Additional work to help those with strong BCB interference at the lower BCB will be attempted in the future and presented on another web page.

Receiver Testing and Comments

USB
RX
DSP
0.389.00
ANT 1
F.AGC
S1 3 5 7 9 20 40 60dB
MEMO 22

Click on the picture to the left to hear the beacon YWB at 389 KHz with a 500 crystal hertz IF filter engaged on a borrowed Icom R-75 receiver. W7ZOI did some measurements on his R75 receiver S-meter using a signal generator and step attenuator. From S9 on up to 60 over, the steps were very accurate. However, below S9, correlation was poor. The built in attenuator is -20 dB when engaged. On my test receiver, I did some A/B testing with the filter in or out. For 1150 KHz (without the low pass filter) I had to engage the receiver's attenuator as without it, the S-meter reads off scale. With the attenuator engaged, the S-meter reads 50 dB over S-9 when tuned to 1150 KHz.

With the filter connected between the receiver and the feed line, (and the attenuator engaged) the S-meter read S-9.This is a drop of about 50 dB at 1150 KHz which means that this filter pretty much works as designed. I love the Icom R75 receiver; it is good value with it's many features in a compact package. Further testing was undertaken on other frequencies.

When listening to **WWV**; At 2.5 MHz, without the filter, the S-meter read S-9. With the filter inline, I could not hear WWV or see any S-meter reading. At 5 MHz without the filter, the S-meter read was at 20 dB over S-9. With the filter inline, I could still hear WWV very faintly, but the S-meter did not register. There was little noticeable attenuation at less than 700 KHz when using this filter.

Additional Photos and information

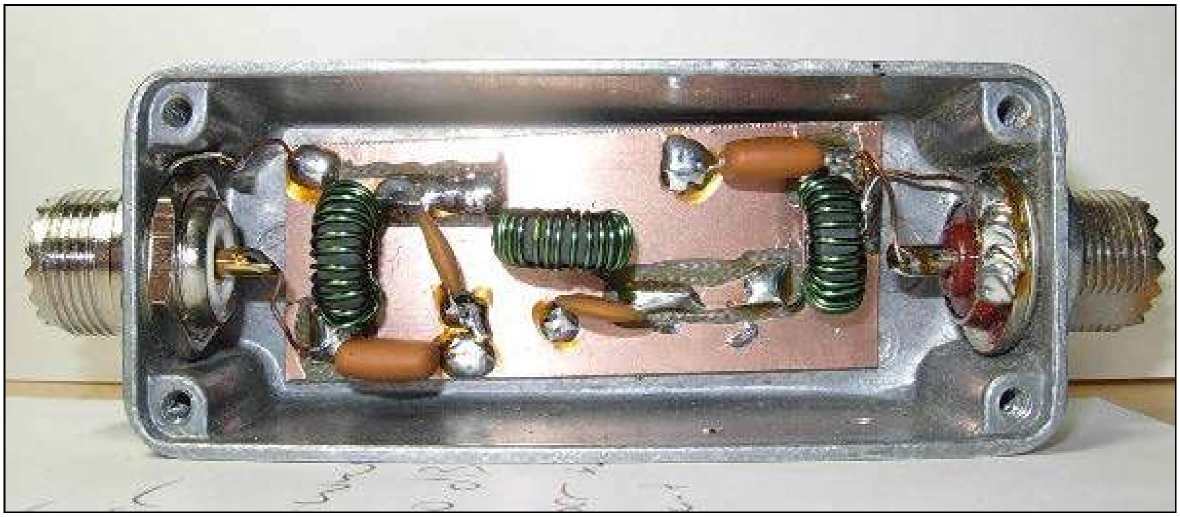

Shown above is the completed project in a Hammond die-cast case with SO-239 connectors at each end.

Shown above a photo of the Skookum beacon SX. It is on 389 KHz. This NDB is located in Skookumchuck BC, Canada

Links

My friend and fellow NDB enthusiast, Ken, VE7KPB has a posting on his web site showing some of the beacons he has logged from his QTH. Consider trying some of these frequencies from your own QTH to get used to finding beacons. Note you must temporarily allow pop ups to see his excellent log.

We recommend this non-directional beacon search and log utility program called WWSU from VE3GOP It must be registered and is a wonderful low cost tool.
I was near beacon L in Balti, Moldova (Балти, Молдова) in 2006. Below is a snippet from the VE3GOP program showing beacon L and also some nearby beacons. (Я изучаю русский язык). Приветствую Вас дорогие друзья!

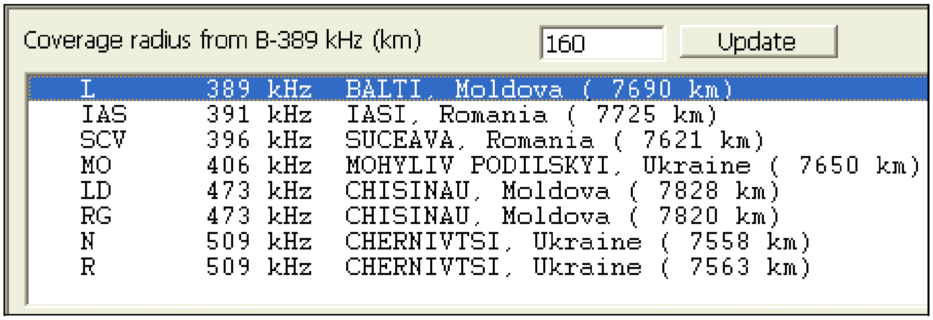

Coverage radius from B-389 kHz (km) 160

```
L          389 kHz  BALTI, Moldova ( 7690 km)
IAS        391 kHz  IASI, Romania ( 7725 km)
SCV        396 kHz  SUCEAVA, Romania ( 7621 km)
MO         406 kHz  MOHYLIV PODILSKYI, Ukraine ( 7650 km)
LD         473 kHz  CHISINAU, Moldova ( 7828 km)
RG         473 kHz  CHISINAU, Moldova ( 7820 km)
N          509 kHz  CHERNIVTSI, Ukraine ( 7558 km)
R          509 kHz  CHERNIVTSI, Ukraine ( 7563 km)
```

Martin Francis has an excellent NDB web site including the free program called NDB WEBLOG for a number of platforms

Some beacons may be located using this NavAid web site

Clint, KA7OEI has an informative web site regarding NDB listening including using digital computer processing to dig out weak signals. This is also a great overview site for newcomers to NDB.

Conclusion

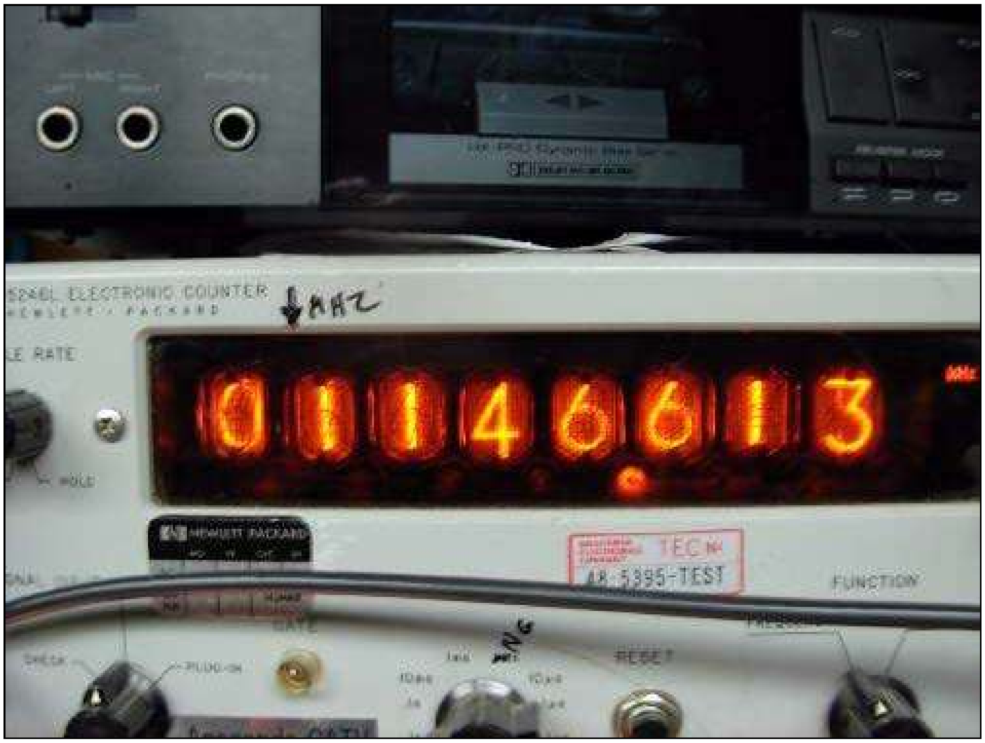

To the right is the outcome when I connect my frequency counter directly to my antenna coax cable feed line. 1150 KHz is my nemesis frequency! Happily it can be tamed with a little filtering to allow NDB listening and logging.

To the right is the outcome with my SWL antenna coaxial feed line connected directly to the scope. The scope was on the 0.5 volts per cm scale, so the peak to peak voltage is 0.2 volts. In just about any high-gain audio amplifier I build, if I touch my finger to the input, I can hear AM 1150 loud and clear - no wonder!

Good luck with your own NDB adventures!

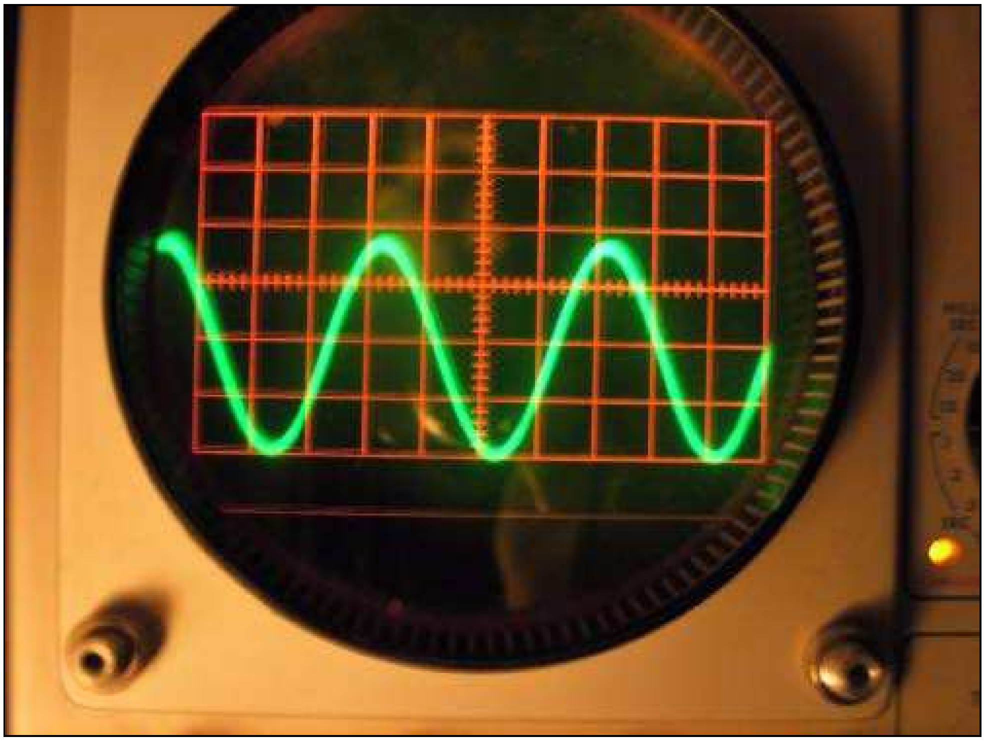

Amateur and Short Wave Radio Electronics Experimenter's Web Site

More NDB Information and Circuits

Introduction

Latest Update: December 3, 2012

This web page holds a collection of NDB-related ideas, experiences and hopefully will include some feedback from fellow NDB enthusiasts.

I devoted a new notebook to this topic and hopefully with inspiration from band listening and communicating/learning from others, I will fill it over time. New content will be added to the bottom of the existing material as QRP-Postadata

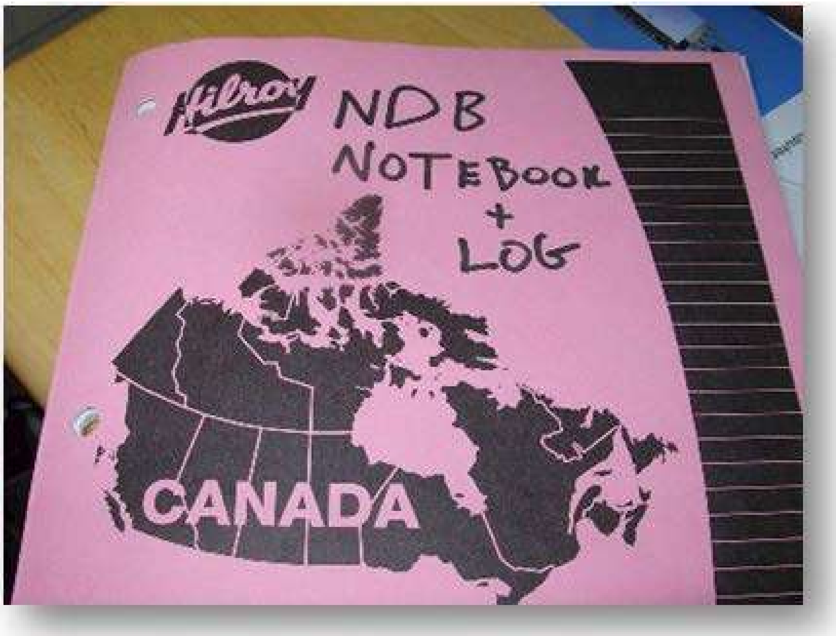

Improved NDB Chebyshev Low Pass Filter

A popcorn or "junk box" low pass filter was designed and presented [on this web page](). After discussion with VE7TW and testing a Realistic DX-300 and other receivers, it became apparent that even more attenuation of a strong local BCB station at 1150 KHz was desirable. In addition, there are other moderately strong AM radio stations from 630 to 800 KHz (especially at night time) which maybe causing mixer intermodulation distortion products. A fault of the junk box low pass filter is poor attenuation below 800 KHz and a better design was a prudent goal. Building on the learning obtained from the junk box filter experiments, an improved 7 element Chebyshev low pass filter was designed and is presented directly below. The 3 dB cut off of this filter is calculated to be 526 KHz. This is the filter that I now use for my home radio station. At my

nemesis frequency of 1150 KHz (where a powerful local radio station broadcasts), the attenuation is calculated to be 68 dB. It takes careful layout and a conductive chassis to realize this level of attenuation, but the effort is worth it. In very strong AM BCB locations, you might consider placing 2 such filters in series between your antenna and receiver if required.

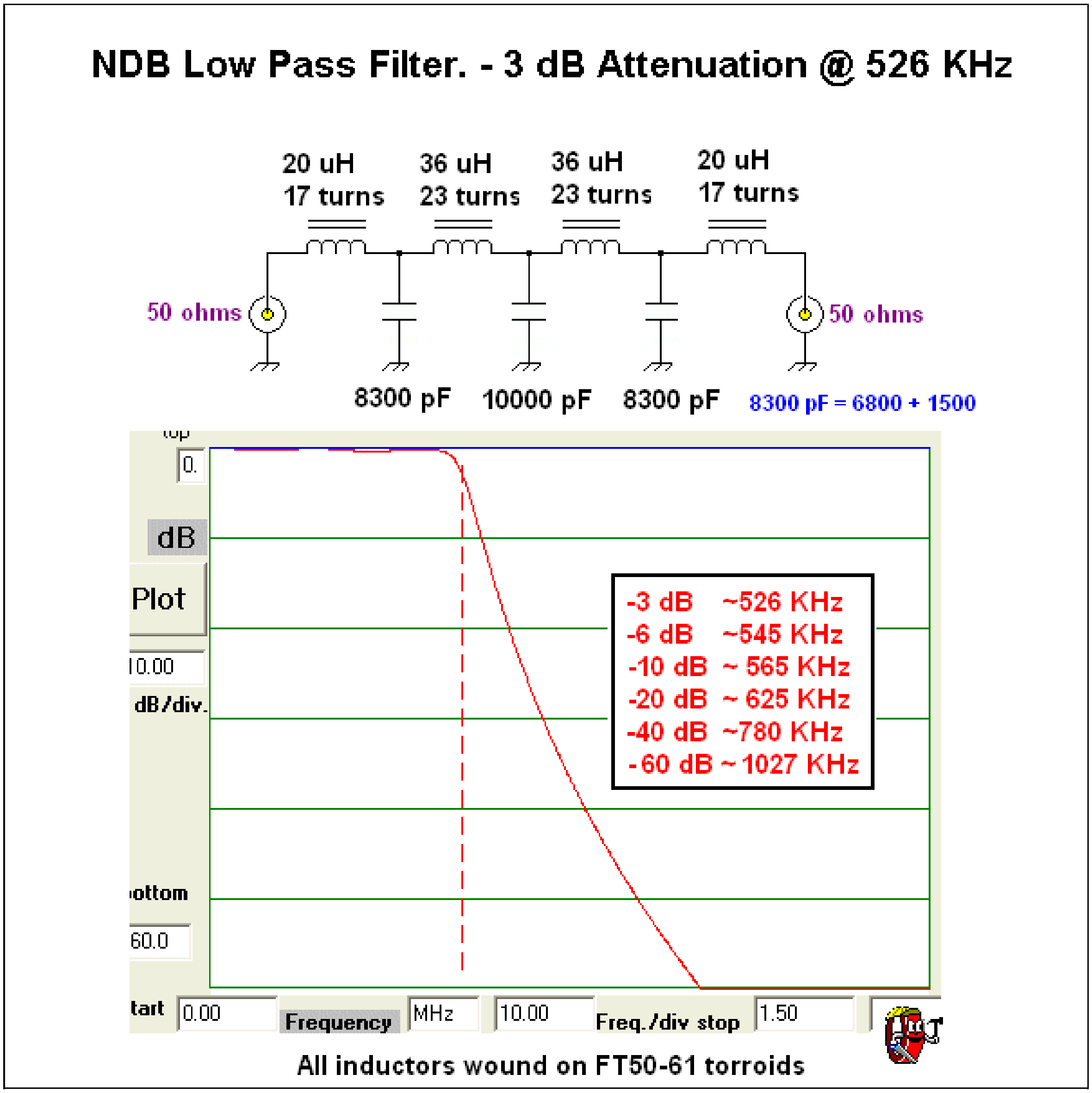

NDB Low Pass Filter. - 3 dB Attenuation @ 526 KHz

20 uH
17 turns
36 uH
23 turns
36 uH
23 turns
20 uH
17 turns
50 ohms
50 ohms
8300 pF
10000 pF
8300 pF
8300 pF = 6800 + 1500
top
0.
dB
Plot
10.00
dB/div.
ottom
60.0
tart
0.00
Frequency
MHz
10.00
Freq./div stop
1.50
-3 dB ~526 KHz
-6 dB ~545 KHz
-10 dB ~ 565 KHz
-20 dB ~ 625 KHz
-40 dB ~780 KHz
-60 dB ~ 1027 KHz

All inductors wound on FT50-61 torroids

The schematic and simulation of the improved NDB low pass filter is shown above.

Non-directional Beacon Identification

It is interesting to visit nearby beacons. In the photograph to the right is XC which broadcasts at 242 KHz.

I have learned that it is very important to confirm the NDB stations your hear via a database or list. What you hear on the air should match the database/list for **both** call sign and frequency, else suspect that you may have copied it incorrectly. **RNA**, the definitive signal list for North and Central America plus Hawaii may be found here.

Three Questions

Steve Ratzlaff, AA7U is an experienced NDB DXer and has been listening to beacons since the mid-1980's.
I asked him the following 3 questions:

1. LF beacons do little more than send their station identification in Morse code, are mostly low power and generally might be perceived by some people as boring and low tech. Yet, on the World Wide Web, one finds numerous web sites, software, projects and commercial equipment all passionately dedicated to NDB listening. What's all this fuss about listening to beacons?

Steve: It's a hobby that requires quite a bit of skill and technical accomplishment to get the most from the equipment. Most folks have AC noise to deal with, which can be particularly bad at LF. Finding an antenna that works at LF and that can be used at your own location can be a major task; finding a radio that has decent LF sensitivity, or an LF converter to use with an existing radio--all these must be detitle with just to begin hearing anything at LF. I find it to be quite a challenge. If it were easy to receive LF beacons then I probably would have lost interest years ago! It's true that in recent years several software programs have become available that allow finding beacons somewhat easier--one simply looks for them on the computer screen and decodes the dots and dashes of the beacon being received. This is quite popular among beginners and veterans alike. But the traditional method of aurally listening for the morse code idents of beacons is probably used more often, though many are combining both aural and software techniques now.

2. Let's say I live in a small city lot or even an apartment. I have modest equipment and/or not a huge amount of cash to spend on gear for NDB listening. From the antenna through to the headphones, what are some basic recommendations you might give to a newcomer wanting to get started in NDB listening?

Steve: The radio must have decent sensitivity at LF, or else an LF converter must be used. Due to high local AC noise, any type of LF antenna used indoors will be a poor substitute to one that can be placed outdoors. A few portable radios cover the LF NDB frequency range that will work for hearing local beacons, though the radio may need to be used outside to get away from AC noise. The discontinued Sony 2010 was the standard for portable radios for reasonable LF performance. Newer radios like the Degen DE1103 have been found to work reasonably well at LF and can be bought for well under $100 by mail order from eBay sellers; or the more expensive Kaito 1103 version, which has a warranty, can be obtained from several distributors like Universal Shortwave. The much more expensive semi-portable Eton E1 works well at LF, but is more in the price range of a tabletop radio. The Icom R75 is currently the best bargain in a tabletop radio that has very good LF sensitivity as well as 1 Hz tuning, which is an asset if a narrow external audio filter is used. I'm not too optimistic about what someone living in an apartment or high rise building might do to successfully receive LF beacons indoors. Often the AC noise level is too high to be able to use an indoor antenna. But some have been able to use loop antennas indoors for the stronger signals. An example of a top of the line commercial loop would be the Wellbrook ALA1530 or LFL1010. Unlike at shortwave frequencies, where simply tossing a wire out the window to a nearby tree or other support, or even running the wire around the room inside, will usually work fairly well, at LF a wire less than several hundred feet generally doesn't perform very well. It can be argued that an active whip antenna makes a very good LF antenna, and doesn't take up much room, but it must be used outdoors. And if there are strong AMBCB signals, then the active antenna, either loop or whip, must have very good overload resistance otherwise it can generate distortion of its own from the strong BCB signals.

3. What kind of distances are considered DX for NDB?

Steve: NDB DX is pretty much a relative thing. One just starting out might be thrilled to hear a beacon from the next town, or from the other side of his own state or province. As one improves his listening setup and gains experience, then usually DX goals also expand to try to hear beacons farther and farther away. NDB DXing generally is not a competitive hobby, unlike amateur radio with its various competitive "contests". Each person's listening setup, local noise level, etc. is usually very different from someone else's, even someone in the same town or general area. One person might live in the suburbs and have a lower noise level than his friend who lives right in town and has a much higher noise level. One might have room to put an antenna in a quiet spot; the other might be limited to much less. People who live near an ocean generally have a much better chance at hearing something exotic offshore than folks living far inland. Folks living in the central or eastern part of North America have many more beacons available to be heard than folks in western North America. But there are always a few beacons that are much stronger than most, and can be heard from long distances of 1000 miles or more, pretty much anywhere in North America at night. One example would be 206 GLS in Galveston, Texas, which runs around 2000 watts, has a large antenna, and is generally readily heard anywhere in North America at night--that beacon might be 1500 miles or more away, and might be considered real DX. However another 25 watt beacon from the same general area in Texas might be hard to hear only several hundred miles from that beacon. So "DX" is pretty much a relative term. Ndblist, an international email list devoted to beacons, is open to anyone with an interest in beacons--members post their loggings there. What might be a local beacon to someone might be DX to someone in a different part of the country. All levels of experience are welcome.

Thanks Steve.

NDB High Pass Filter

A high pass filter using standard value capacitors was designed using GPLA. although, such a filter would not help AC line noise and RFI generated in the house, I suspected my antenna was picking up local noise from below the NDB band. This filter was mounted inside a die-cast Hammond box with a SO-239 at each end. I used 22 gauge enamel covered wire for the inductors. A photo of the filter is shown to the right.
For the 0.01 uF caps, I used junk box ceramic capacitors with a 20% tolerance, however, I did measure a bunch and found 2 within 5% tolerance for my filter bread board.

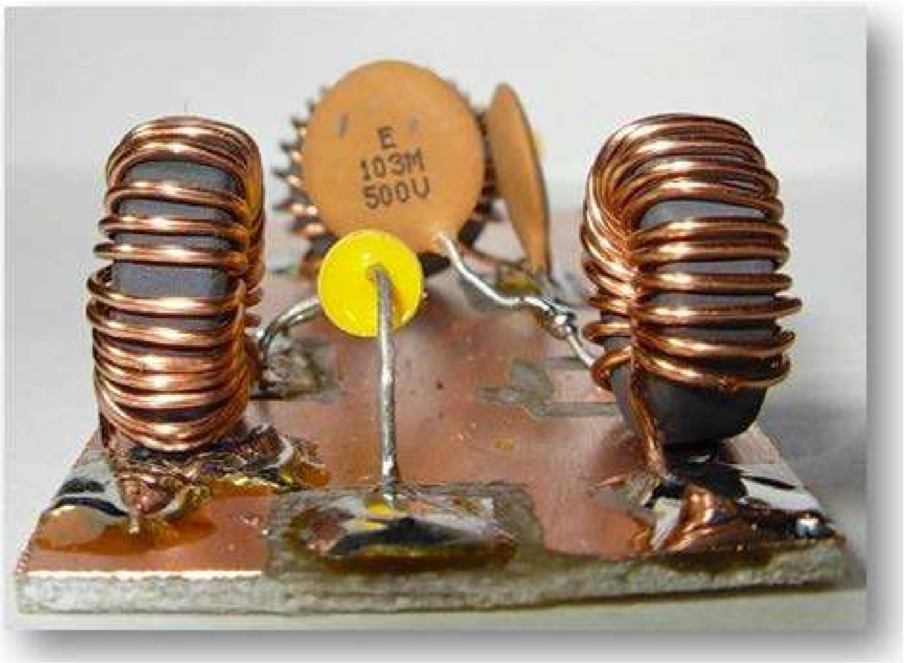

To the left is the filter schematic. This is an N = 7 Chebyshev high pass filter with a 3 dB cut off of 157 KHz. This cutoff frequency allowed the use of common, standard value capacitors and also even turns numbers to reach the desired inductance for the inductors when wound with FT50-61 ferrite cores.

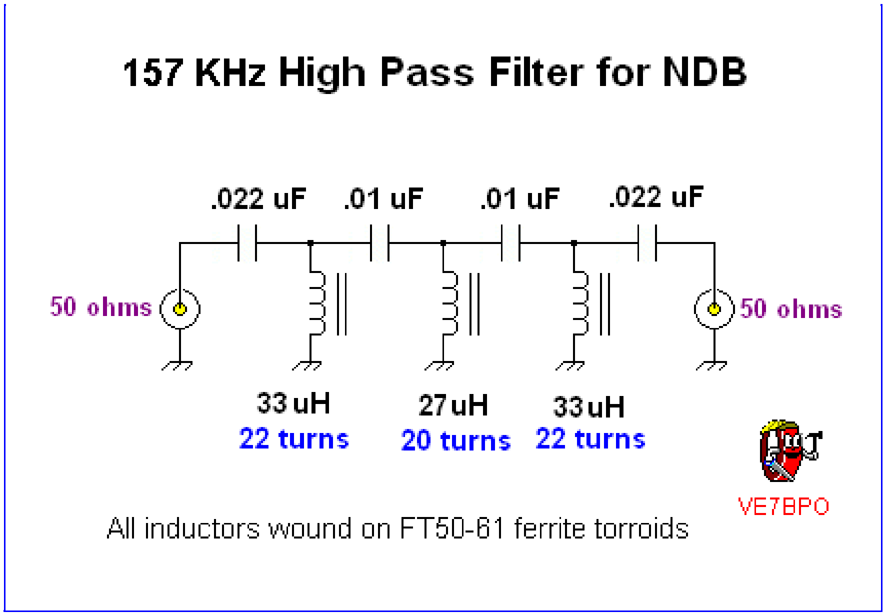
157 KHz High Pass Filter for NDB
.022 uF
.01 uF
.01 uF
.022 uF
50 ohms
50 ohms
33 uH
22 turns
27uH
20 turns
33 uH
22 turns
VE7BPO
All inductors wound on FT50-61 ferrite torroids

Use 5% tolerance, high Q caps such as polystyrene or NP0 ceramic and **not** junk box bypass-grade ceramic capacitors as possible. I used trashy ceramic caps for the 0.01 uF parts due to lack of better parts at the time of building and testing.

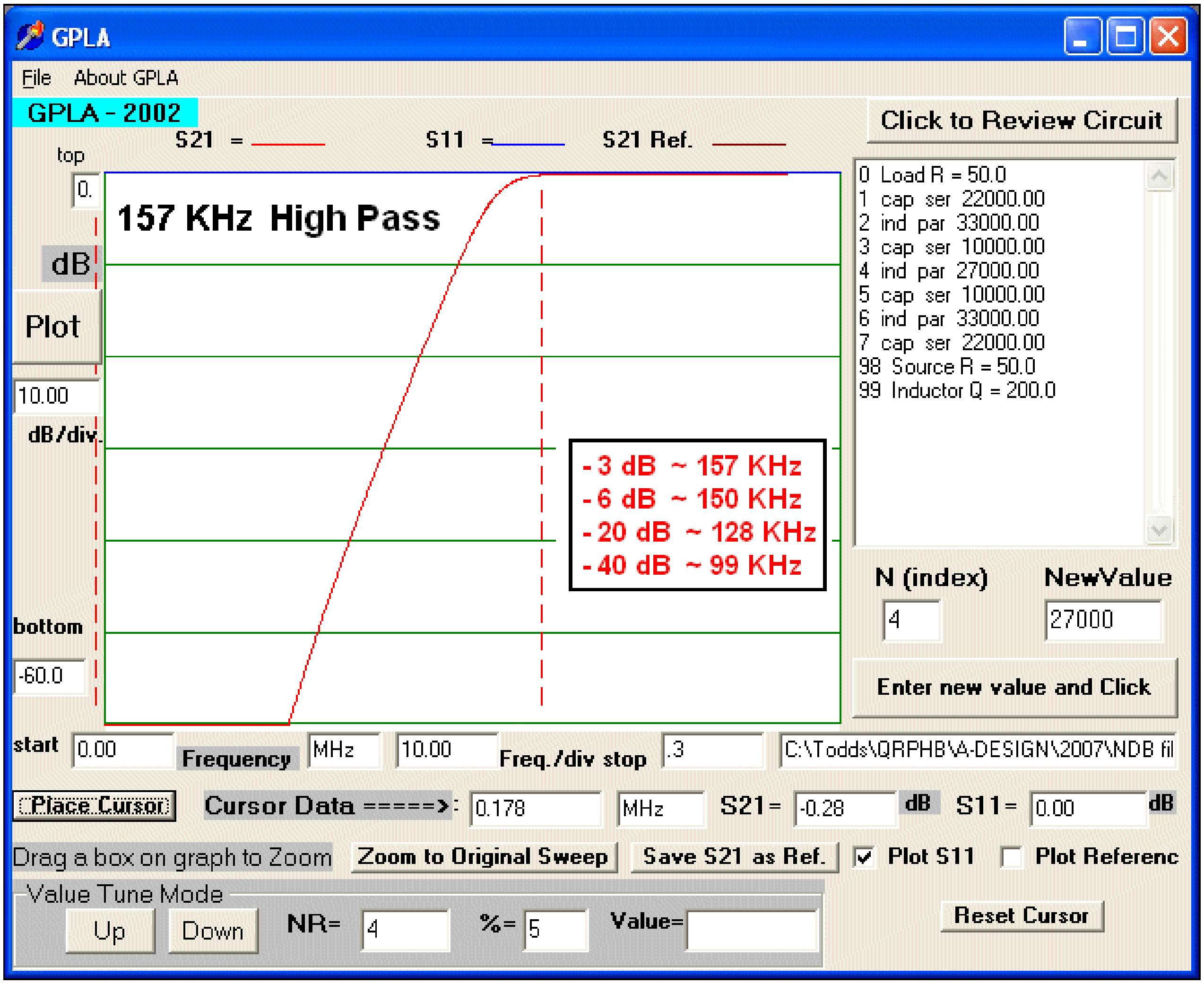
GPLA
File
About GPLA
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
top
dB
Plot
dB/div
bottom
157 KHz High Pass
start
Frequency
Freq./div stop
Place Cursor
Cursor Data =====>
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Value Tune Mode
Up
Down
N (index)
NewValue
Enter new value and Click
Reset Cursor

```
0  Load R = 50.0
1  cap  ser  22000.00
2  ind  par  33000.00
3  cap  ser  10000.00
4  ind  par  27000.00
5  cap  ser  10000.00
6  ind  par  33000.00
7  cap  ser  22000.00
98  Source R = 50.0
99  Inductor Q = 200.0
```

- 3 dB ~ 157 KHz
- 6 dB ~ 150 KHz
- 20 dB ~ 128 KHz
- 40 dB ~ 99 KHz

top: 0.
dB/div: 10.00
bottom: -60.0
start: 0.00 MHz
Freq./div: 10.00
stop: .3
C:\Todds\QRPHB\A-DESIGN\2007\NDB fil
Cursor Data: 0.178 MHz
S21= -0.28 dB
S11= 0.00 dB
☑ Plot S11
☐ Plot Referenc
N (index): 4
NewValue: 27000
NR= 4
%= 5
Value=

Above is the filter GPLA simulation. In particular, I have harsh noise from about 110 KHz on down. At 78 KHz, where this filter has a calculated attenuation of ~ 56 dB, I made an audio file of the band noise. This is in AM mode with the filter out for a few seconds and then in line. With the high pass filter in line, there is pronounced attenuation of the noise and my local 10 KW BCB station at 1150 KHz suddenly appears. Prior to this it was hidden by the harsh noise. At frequencies less than ~200 KHz (without my low pass filter) I can hear this BCB station intermittently as I tune around. I suspect that the R75 filtering down at 200 KHz and down is insufficient to stop this monster station.

At my QTH, using a high pass filter reduces some of the noise on the NDB band. At my location, a high pass plus a low pass filter in cascade between my antenna and my receiver results in less QRN and easier weak signal copying.

Long Wave Broadcast Radio Filter

I learned about LW Broadcast radio from Steve Ratzlaff. In particular, радио россий "Rah-deo RaSEE" (make sure you roll the R!) can occasionally be heard on the west coast and broadcasts at night-time using 500-1000 KW power. The frequencies he recommended to try were 153, 180, 189, and perhaps 171, 234 and 279 KHz. I have terrible problems with a local BCB radio station at 1150 KHZ that causes intermodulation distortion and/or blow-by detection at and below 200 KHz in addition to a terrible noise source at 78-120 KHz. Therefore, I built another cascade low pass/high pass filter and placed it in the same chassis as my regular NDB low pass/high pass combination filter for use when tuning LWBC and perhaps for when listening to frequencies less than 200 KHz.

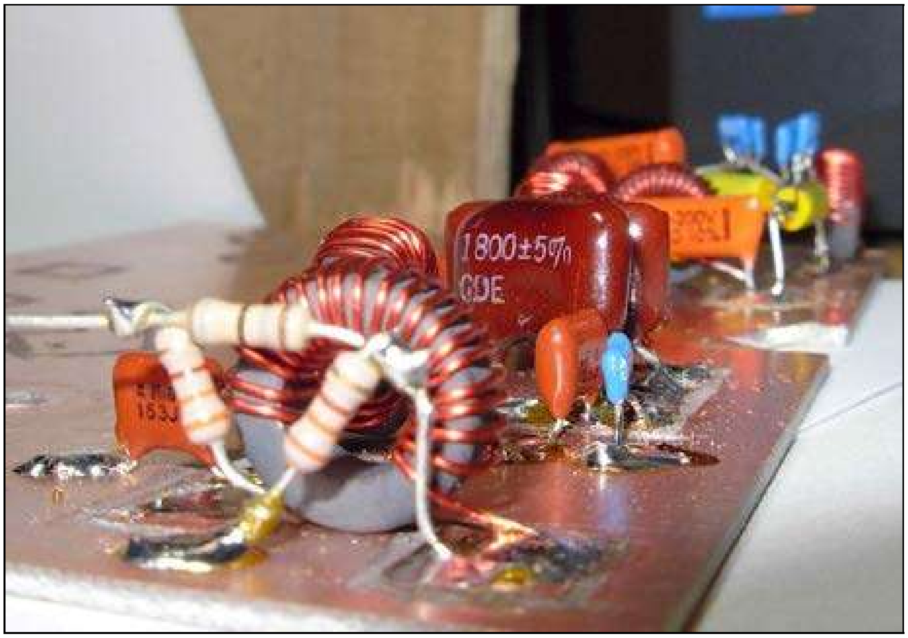
1800±5%
CDE
153J

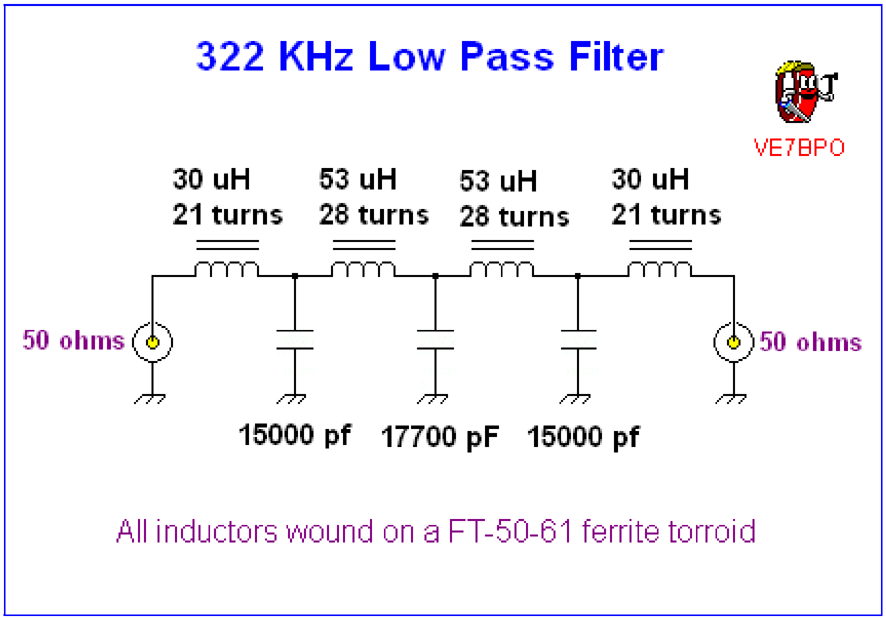
322 KHz Low Pass Filter
VE7BPO
30 uH
21 turns
53 uH
28 turns
53 uH
28 turns
30 uH
21 turns
50 ohms
50 ohms
15000 pf
17700 pF
15000 pf
All inductors wound on a FT-50-61 ferrite torroid

Above is. the schematic of the 322 KHz low pass filter. In the photograph above, you can see a 50 ohm pad at the input that was used only during testing. This filter offers a calculated attenuation of ~ 98 dB to my 1150 KHz interfering station. In reality it is not possible to achieve this level of attenuation, however, there is no detectable 1150 KHz signal interference with the filter in line which makes me happy. Click here to listen to the dramatic difference with regard to interference this filter makes at my QTH with my receiver tuned to 199 KHz. The receiver is set for wide band AM detection; first without the LWBC filter and then with the filter switched in. When the filter is switched in, the BCB interference disappears and a Canadian NDB (UAB @ 200 KHz can faintly be heard along with our cat meowing in the background. It is not possible to listen to LF without aggressive low pass/high pass filtering at my QTH.

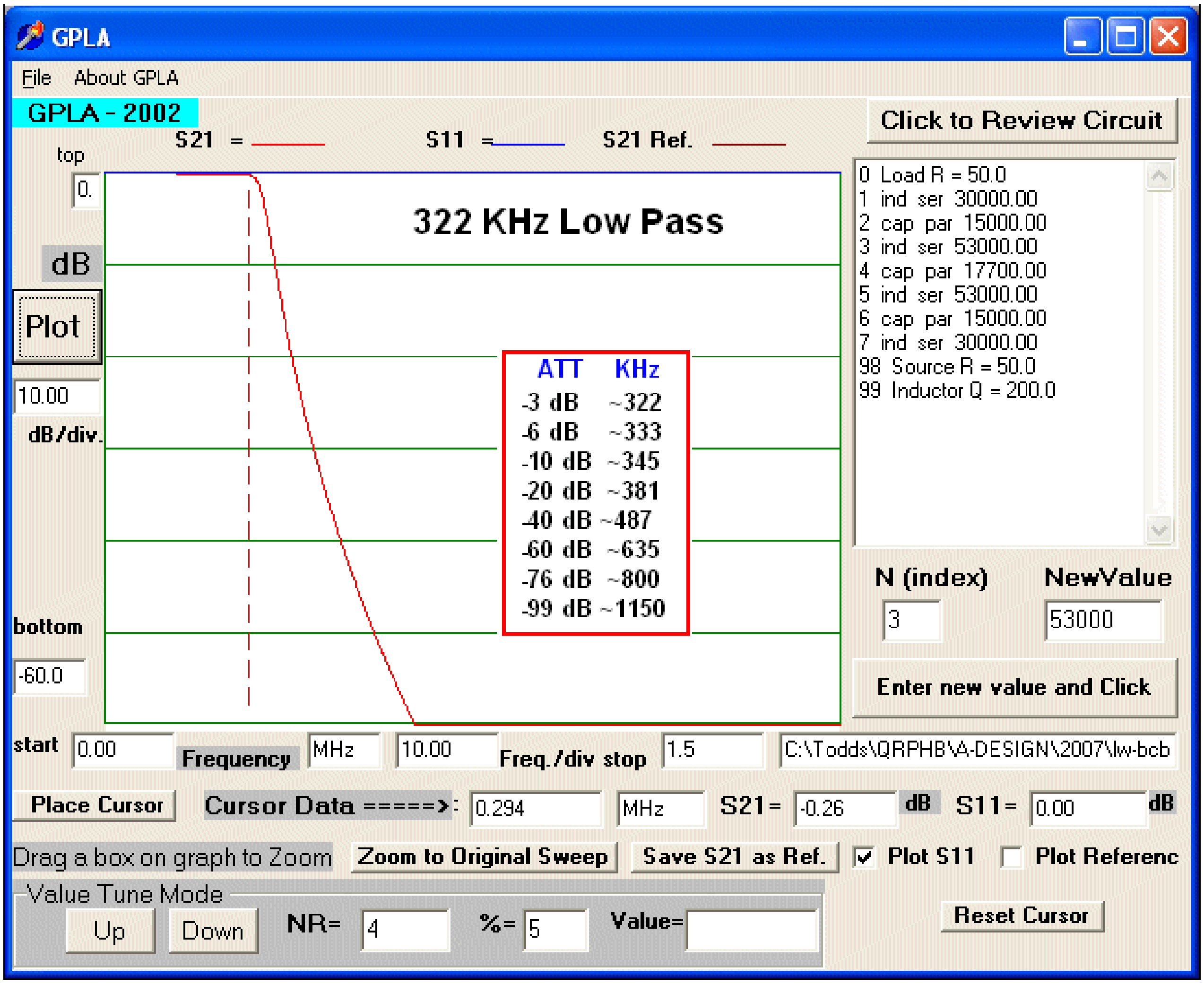
GPLA
File
About GPLA
GPLA - 2002
Click to Review Circuit
S21 =
S11 =
S21 Ref.
top
dB
Plot
dB/div.
bottom
Frequency
Freq./div stop
Place Cursor
Cursor Data =====>:
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Value Tune Mode
Up
Down
Reset Cursor
Enter new value and Click

322 KHz Low Pass

ATT	KHz
-3 dB	~322
-6 dB	~333
-10 dB	~345
-20 dB	~381
-40 dB	~487
-60 dB	~635
-76 dB	~800
-99 dB	~1150

```
0  Load R = 50.0
1  ind  ser  30000.00
2  cap  par  15000.00
3  ind  ser  53000.00
4  cap  par  17700.00
5  ind  ser  53000.00
6  cap  par  15000.00
7  ind  ser  30000.00
98  Source R = 50.0
99  Inductor Q = 200.0
```

top: 0.

10.00 dB/div.

bottom: -60.0

start: 0.00 MHz

Freq./div: 10.00

stop: 1.5

C:\Todds\QRPHB\A-DESIGN\2007\lw-bcb

N (index): 3

NewValue: 53000

Cursor Data =====>: 0.294 MHz S21= -0.26 dB S11= 0.00 dB

☑ Plot S11 ☐ Plot Referenc

NR= 4 %= 5 Value=

Above is the GPLA simulation of the LWBC low pass filter.

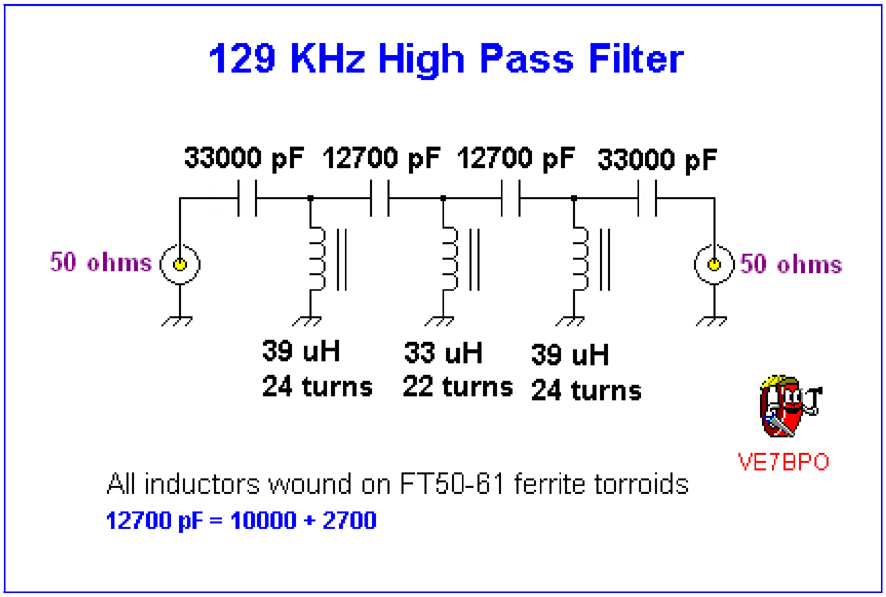
129 KHz High Pass Filter
33000 pF
12700 pF
12700 pF
33000 pF
50 ohms
50 ohms
39 uH
24 turns
33 uH
22 turns
39 uH
24 turns
VE7BPO

All inductors wound on FT50-61 ferrite torroids

12700 pF = 10000 + 2700

Above is the schematic of the 129 KHz LWBC high pass filter. In either of the 2 filters, capacitor values can be obtained using 1 or 2 standard value capacitors in parallel. The cutoff frequencies of both filters were chosen to allow using practical component values.

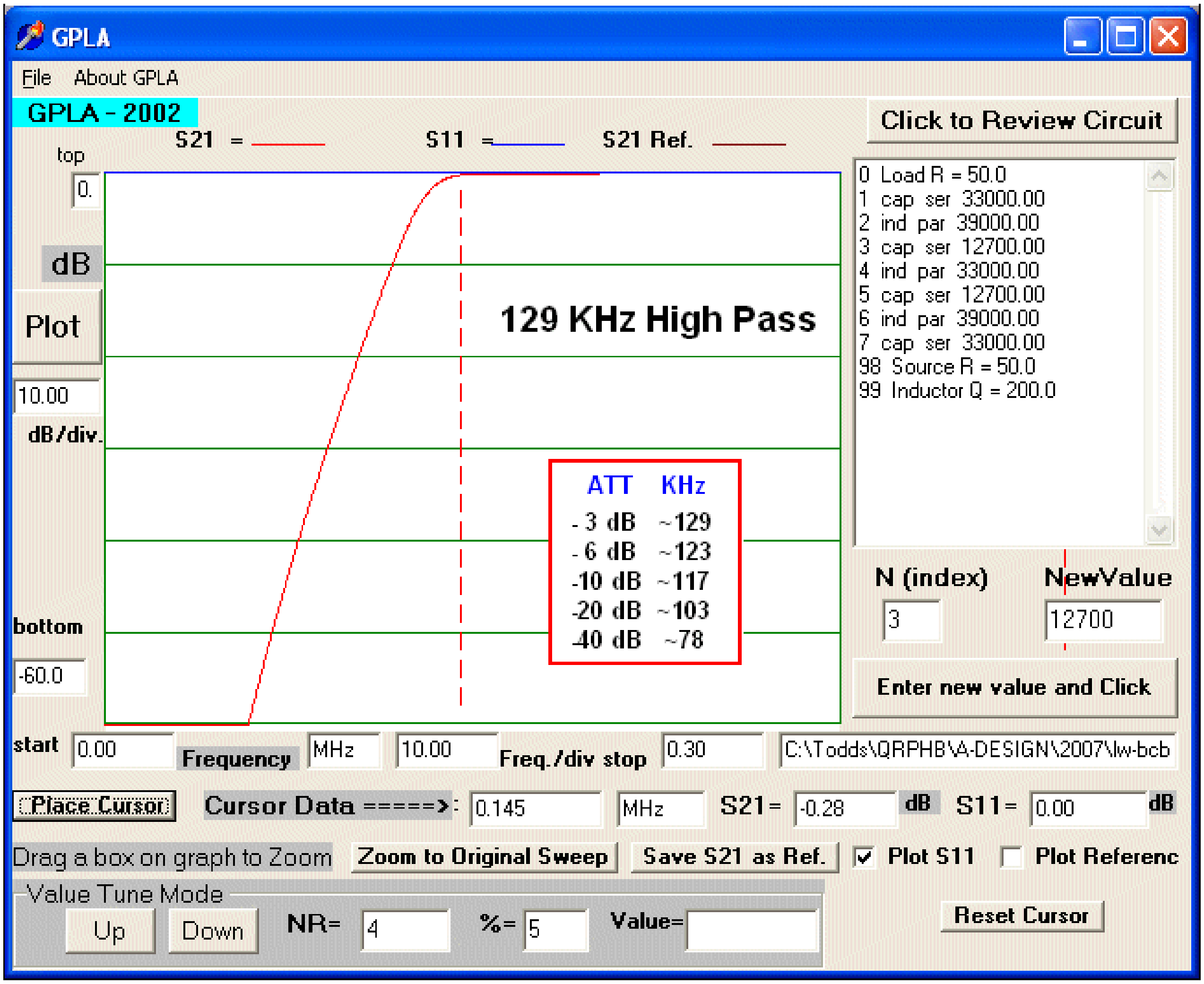
GPLA
File
About GPLA
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
dB
Plot
Enter new value and Click
Place Cursor
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Value Tune Mode
Up
Down
Reset Cursor

129 KHz High Pass

ATT	KHz
- 3 dB	~129
- 6 dB	~123
-10 dB	~117
-20 dB	~103
-40 dB	~78

0 Load R = 50.0
1 cap ser 33000.00
2 ind par 39000.00
3 cap ser 12700.00
4 ind par 33000.00
5 cap ser 12700.00
6 ind par 39000.00
7 cap ser 33000.00
98 Source R = 50.0
99 Inductor Q = 200.0

top 0.
10.00 dB/div.
bottom -60.0
N (index) 3
NewValue 12700
start 0.00 Frequency MHz 10.00 Freq./div stop 0.30
C:\Todds\QRPHB\A-DESIGN\2007\lw-bcb
Cursor Data =====>: 0.145 MHz S21= -0.28 dB S11= 0.00 dB
☑ Plot S11 ☐ Plot Referenc
NR= 4 %= 5 Value=

Above is the GPLA simulation of the LWBC high pass filter. The high pass filter might not be needed at your QTH. My LWBC filter has the low pass filter before the high pass filter. I.e. they are in series or cascaded.

Dual NDB and LWBC Filters

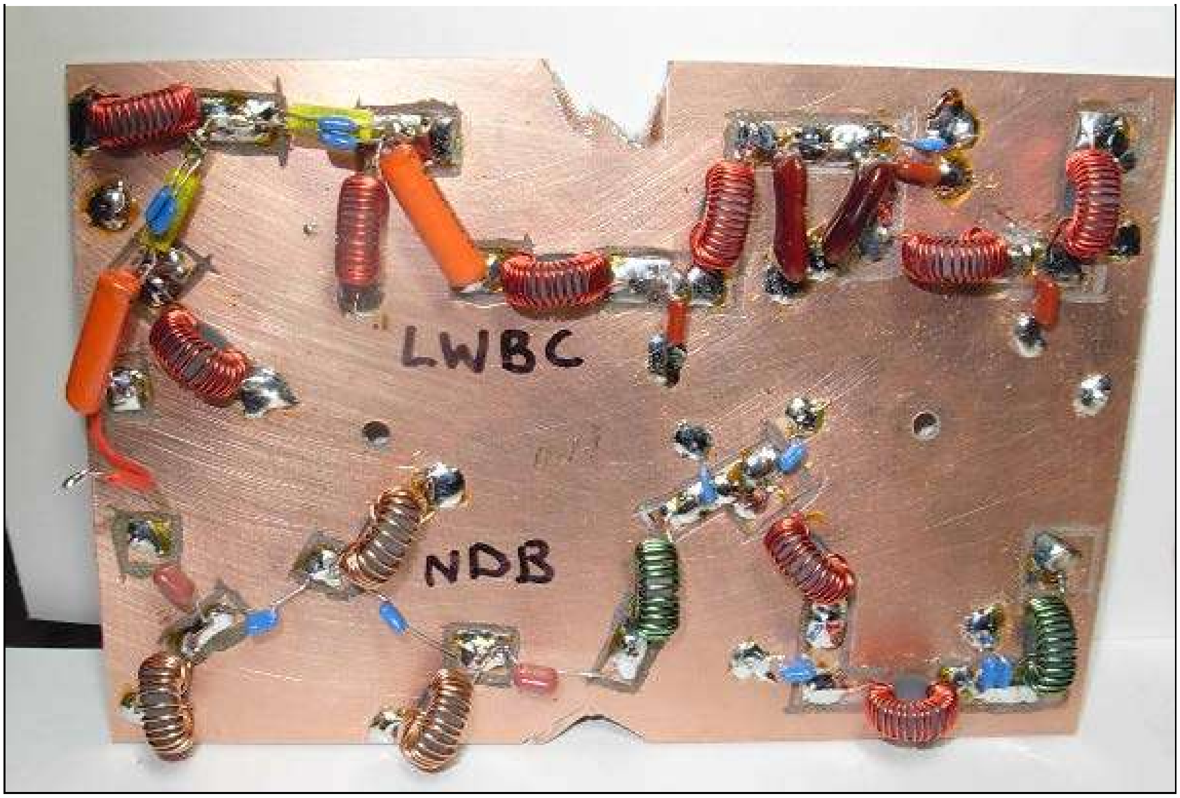

For use in my radio shack, I built LWBC and NDB filters inside 1 chassis with separate inputs and outputs. Some photos of this project are shown directly above and below. The NDB filter is the 526 KHz low pass filter in series with the 157 KHz high pass filter. The LWBC filter is the 322 KHz low pass filter in series with the 129 KHz high pass filter. High Q caps were used and the inductors were wound with either 22 or 24 gauge wire to obtain a relatively high unloaded Q. The large Hammond project case allowed reasonable spacing of the inductors and a nice long input to output layout.

Beacon XJ @326 KHz

Above is NDB XJ in Fort St. John, BC. Photo by VE7KPB in August 2008.

QRP — Posdata: NDB Low-pass Filter with Trap

An email from Rick, NU7Z spawned this 2012 addition.

Depending on their design, typical NDB low-pass filters provide less than 20 dB attenuation at 620 - 630 KHz, and if you hear a strong station on this frequency — good luck!

Rick sought a filter with a trap at ~ 620 KHz — after mulling around, we encountered design problems with a trap frequency so close to the low-pass cut-off frequency and later asked Wes, W7ZOI if he might help design our filter.

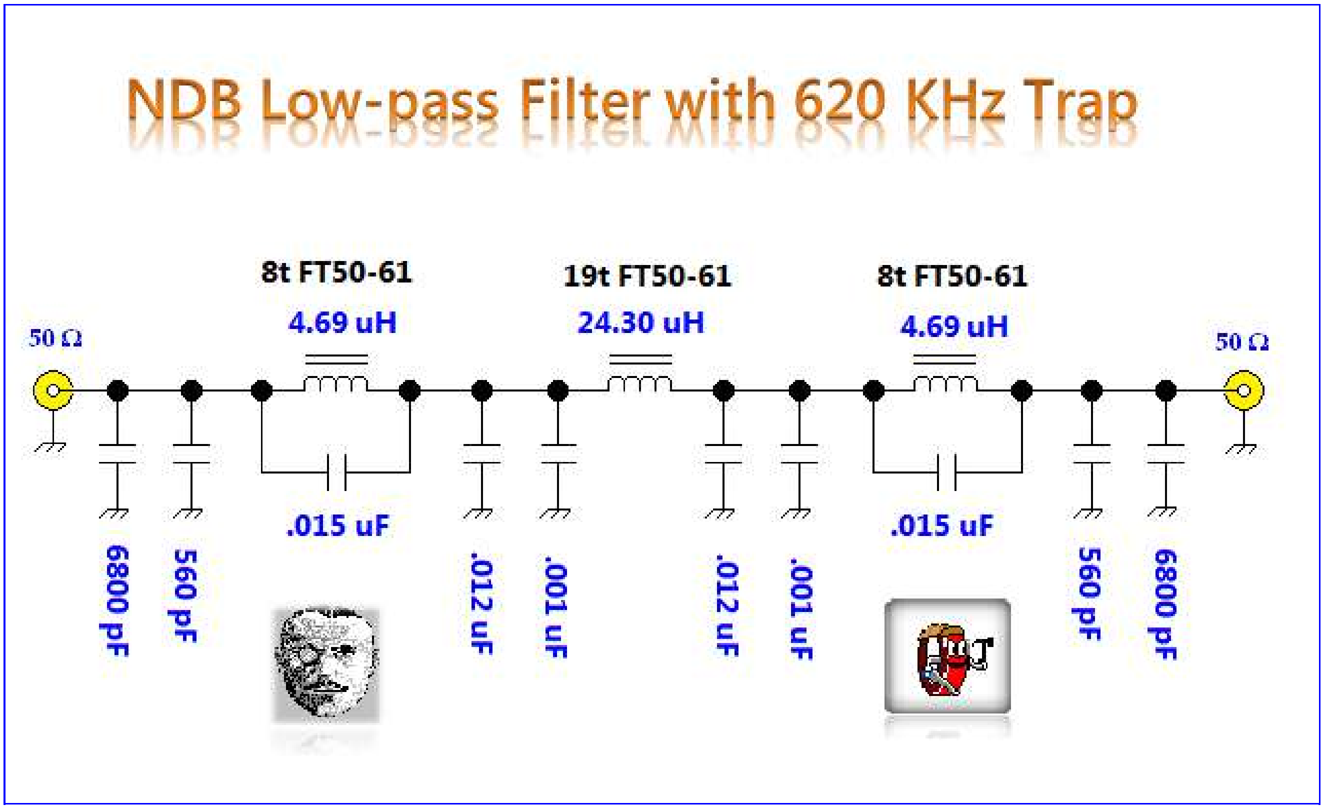
NDB Low-pass Filter with 620 KHz Trap
50 Ω
8t FT50-61
4.69 uH
19t FT50-61
24.30 uH
8t FT50-61
4.69 uH
50 Ω
6800 pF
560 pF
.015 uF
.012 uF
.001 uF
.012 uF
.001 uF
.015 uF
560 pF
6800 pF

Above — My version of the Wes, W7ZOI designed NDB low-pass filter with a trap. **Red** filter below

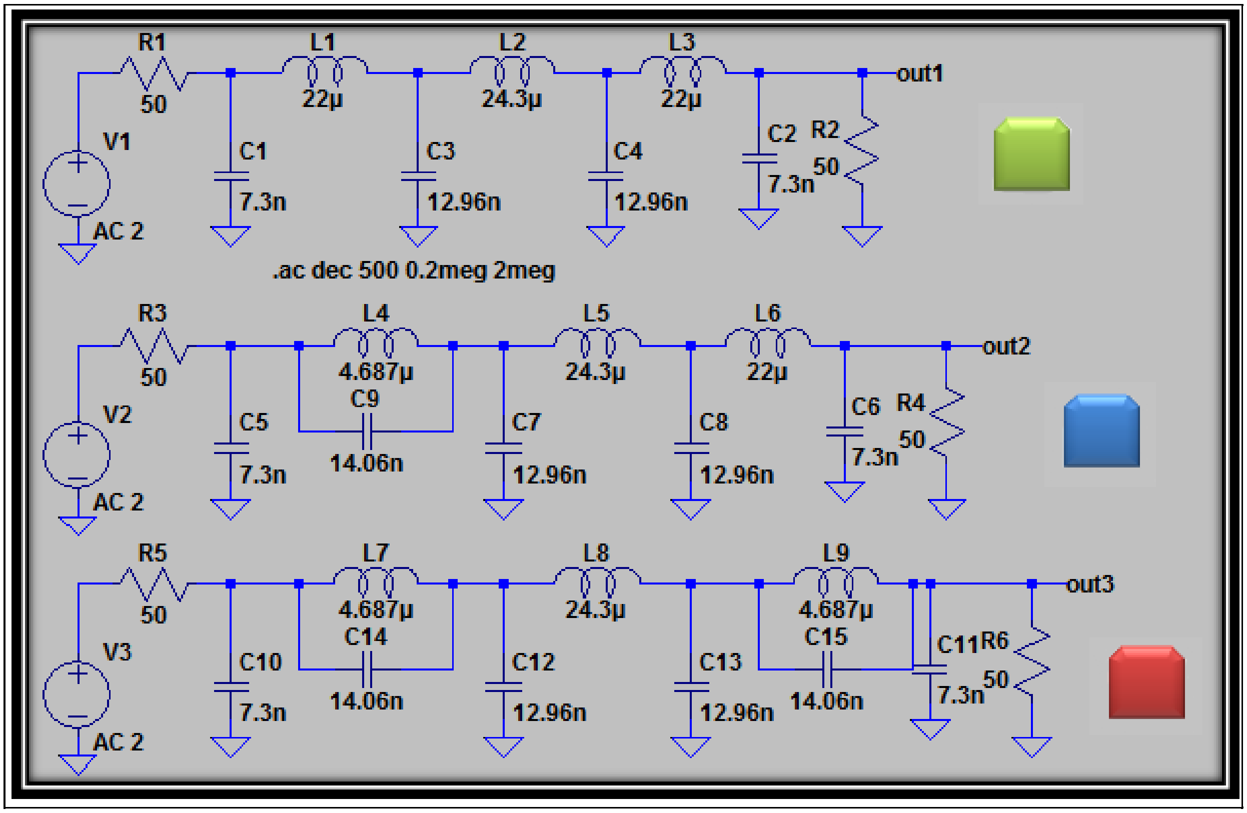
R1
50
L1
22µ
L2
24.3µ
L3
22µ
out1
V1
AC 2
C1
7.3n
C3
12.96n
C4
12.96n
C2
7.3n
R2
50
R3
50
L4
4.687µ
C9
14.06n
L5
24.3µ
L6
22µ
out2
V2
AC 2
C5
7.3n
C7
12.96n
C8
12.96n
C6
7.3n
R4
50
R5
50
L7
4.687µ
C14
14.06n
L8
24.3µ
L9
4.687µ
C15
14.06n
out3
V3
AC 2
C10
7.3n
C12
12.96n
C13
12.96n
C11
7.3n
R6
50
🟩
🟦
🟥

```
.ac dec 500 0.2meg 2meg
```

Above — A 7th order, 0.1 dB ripple **Chebyshev low-pass filter** with a 550 KHz cut-off filter evolved to include **1 trap**, and then **2 traps** at 620 KHz.

We learned that in simple situations, you may modify the elements of a low-pass filter so that the usual inductor is replaced by a parallel trap. See Wes' work in EMRFD Chapter 3; in particular, Figure 3.10.

Wes wrote he's employed this technique successfully before — for example, to add harmonic suppression to a simple output network for a QRP transmitter, although he hadn't added traps to higher order filters like the 1 we wanted. Click for a file containing the math contributed by Wes, W7ZOI.

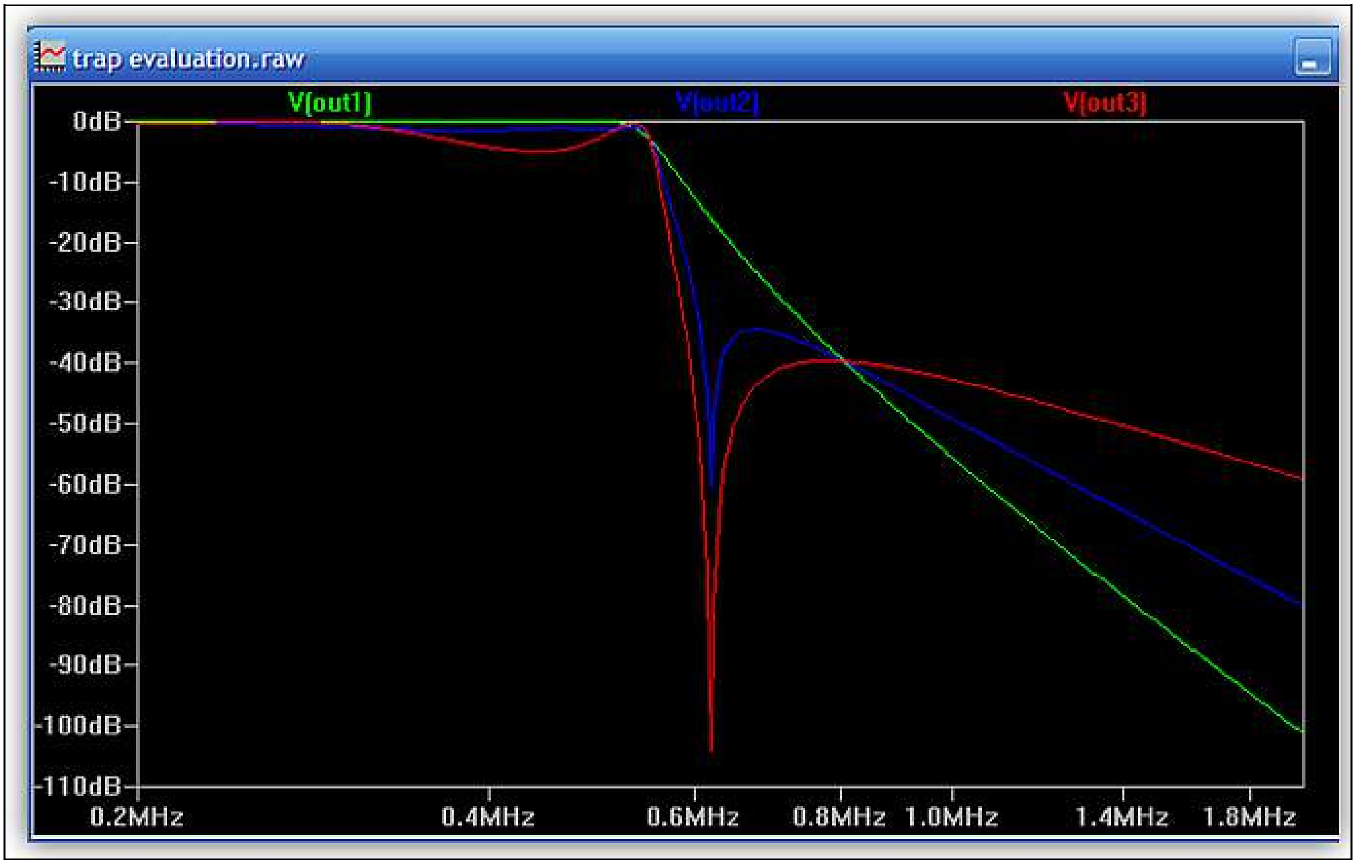

Above — The SPICE analysis of the 3 color-coded filters above. This design excludes the impact of finite L and C and the unloaded Q that could significantly affect function since the trap frequency is close to the low-pass cutoff frequency. These factors usually worsen the insertion loss near cutoff, but since we're using this filter in a noisy RF environment, filter misperformance should be tolerable.

In the future, Wes recommended designing an elliptical low-pass filter with software such as that distributed by AADE.

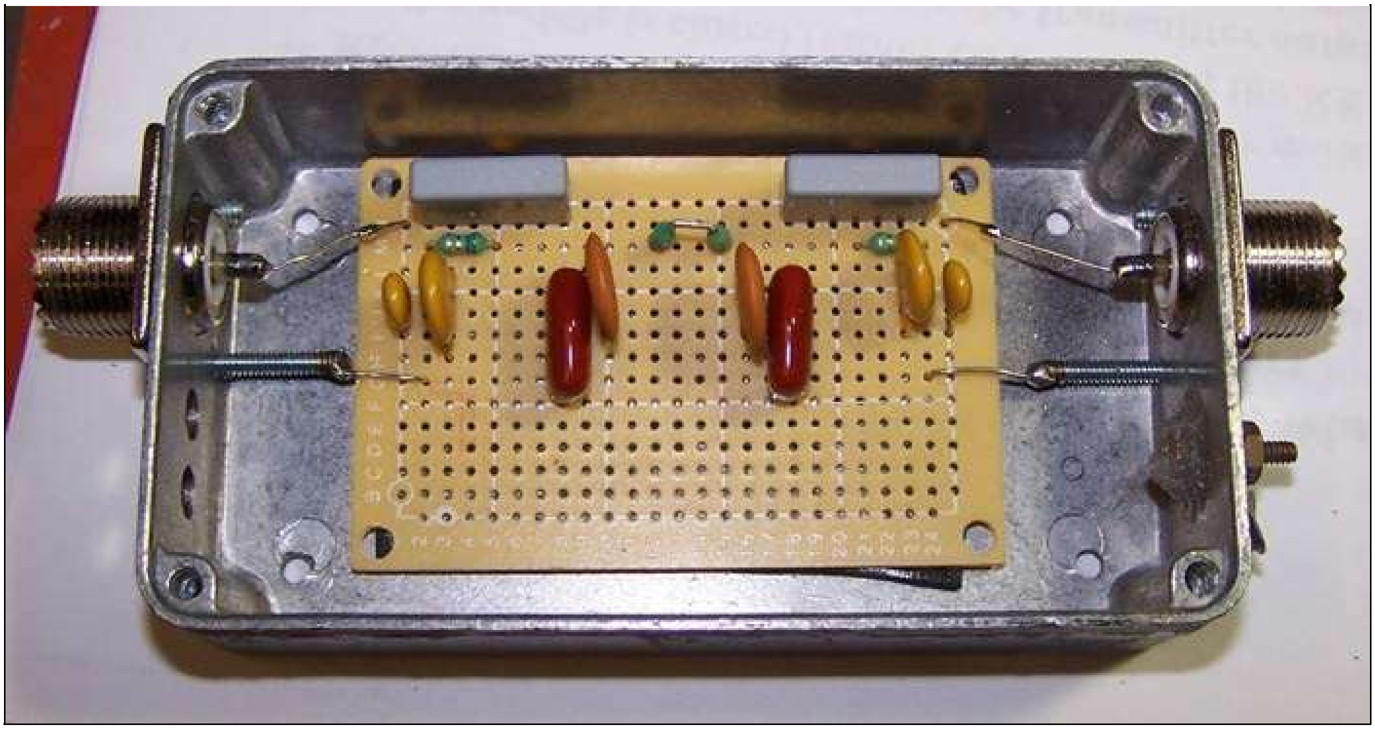

Above — A version of the filter built by Rick, NU7Z using epoxy-coated inductors for the L's. The insertion loss with these inductors = ~ 5 dB, although he runs a 40 dB receive preamp and can accommodate such losses.

Despite employing a loop receiving antenna, he could not listen around 500 KHz due to a loud, local broadcast station at 630 KHz. Inserting this filter reduced this 630 KHz signal from 40 dB over S-9 down to S-1 on his receiver S-meter.

Fantastic! Big thanks (большое спасибо) to Wes, W7ZOI.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

440 Hertz Peaked Low Pass Audio Filter Experiments

Introduction

At audio frequencies, low pass filtering can go a long way to improving CW and beacon reception. It has a been a long time since I built one and therefore decided to experiment with some designs using the 5532 op amp. EMRFD has a great section on RC active audio filters starting at Chapter 3.5 and this is where I began. After some experimentation, I remembered a peaked low pass filter designed and published by Wes, W7ZOI in the 1970's. This filter became very popular in Russia after publication in a 1971 Russian Amateur Radio Journal. I asked Wes if he might design another low pass filter peaked for 440 Hertz, which is my favorite CW beat note frequency. This filter is intended for use as an out board headphone jack device for the Icom R75 or other receiver.

AF LOW PASS
PHONES
INPUT

Base Project Schematic

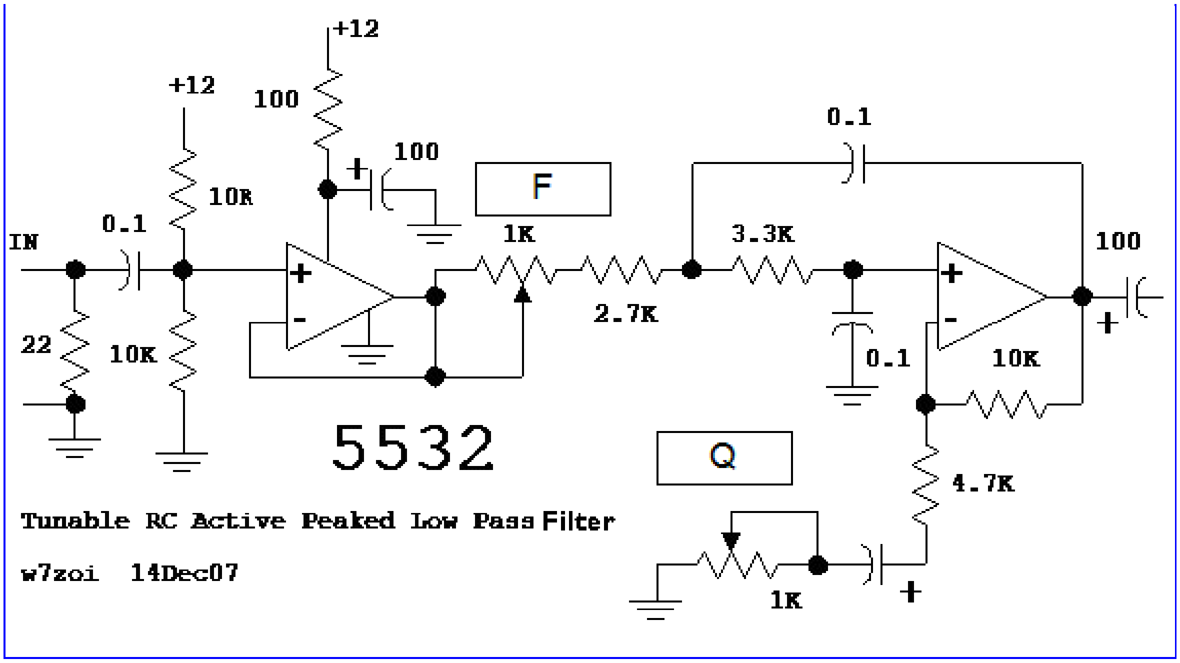
+12
+12
100
10K
100
0.1
IN
0.1
22
10K
F
1K
2.7K
3.3K
0.1
100
10K
5532
Q
4.7K
1K

Tunable RC Active Peaked Low Pass Filter

w7zoi 14Dec07

Above is the filter designed by Wes, W7ZOI. It has two 1K pots you can tweak; a subtle frequency control and a Q adjustment. It is theoretically possible to adjust the Q too strongly as to cause oscillation, although this did not happen with my bread board.

Schematic and Circuit Building Details

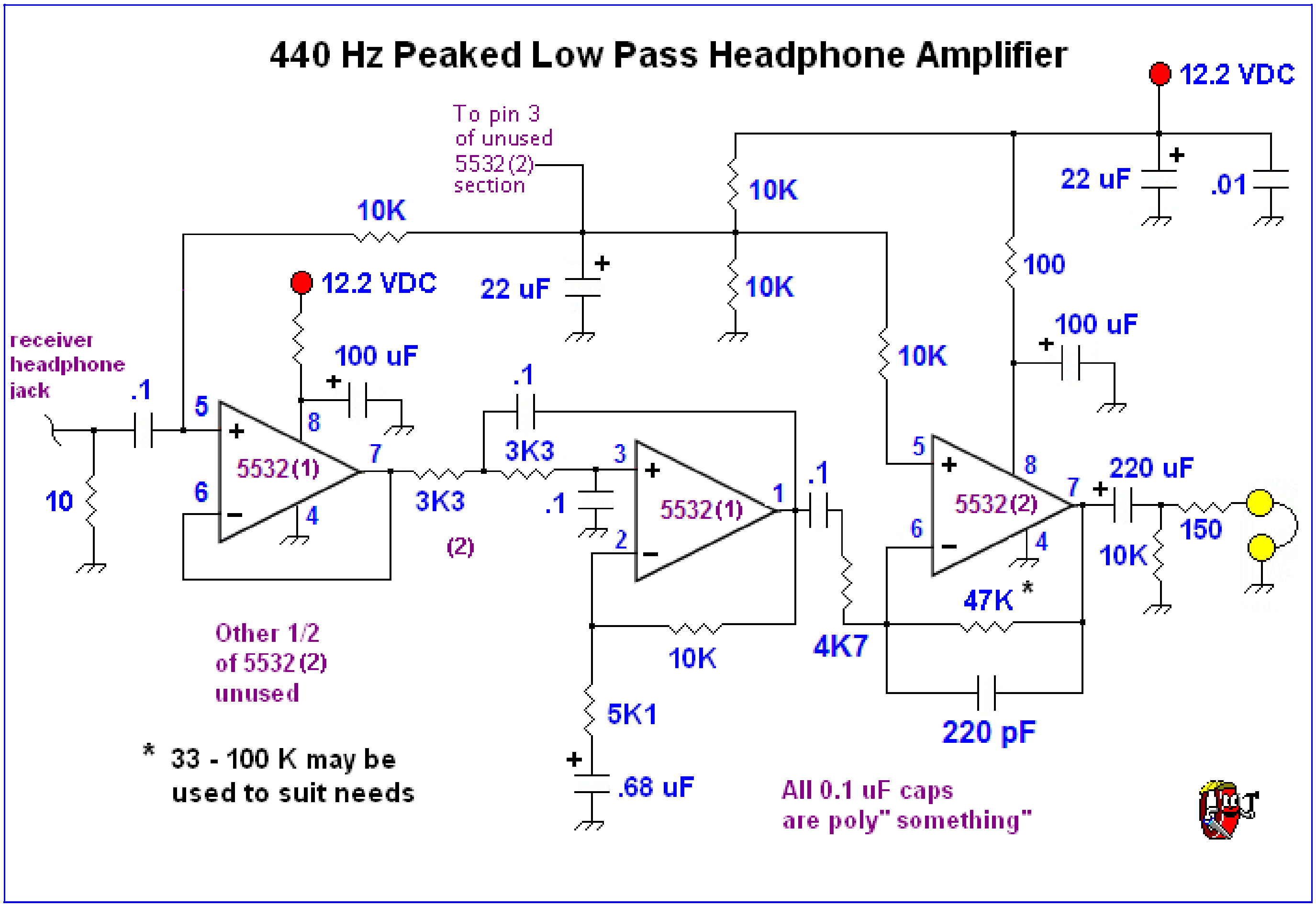
12.2 VDC
10K
10K
22 uF
.01
22 uF
10K
100
12.2 VDC
100 uF
receiver
headphone
jack
.1
5
8
100 uF
.1
10K
10
6
+
5532(1)
7
3K3
3K3
3
5532(1)
.1
1
.1
2
(2)
5
8
5532(2)
7
220 uF
6
4
150
10K
47K *
10K
4K7
5K1
220 pF
.68 uF

440 Hz Peaked Low Pass Headphone Amplifier

To pin 3 of unused 5532(2) section

Other 1/2 of 5532(2) unused

* 33 - 100 K may be used to suit needs

All 0.1 uF caps are poly" something"

Above is the final filter design. My prototype filter had 2 variable pots and I disconnected and measured them after discovering my favorite setting and then constructed the version shown above using fixed-value resistors instead of the potentiometers. If you like to tweak knobs, you might leave 1 or both of the pots in. The 0.68 uF cap can be raised as high as 4.7 uF if you have poly caps this large in capacitance value. although, it seems wasteful, I did not use one half of the second 5532 op amp. Feel free to add another pole of low pass filtering or something else to utilize this stage if you like. Pin 3 of the unused, second op amp 1/2 is connected to the 10K/10K voltage divider bias with a wire as indicated in the text.

This filter sounds the best when the R75 volume control is minimally turned on as I suspect some of the wide band noise heard is from the ICOM AF chain. More importantly, If the R75 audio gain control is turned too high, the filter will be overdriven and sound distorted This is especially true when using the 2.4 KHz wide SSB filter on the R75. The 10 ohm filter input resistor attenuates the receiver output and makes it more difficult to overdrive the audio filter. The 500 Hz filter at the 9 MHz Receiver IF has quite a bit of of loss and with this filter switched in, it is difficult to overdrive the audio filter. For best results, an audio filter should be placed just after the first AF preamp stage, however, using the headphone jack is the only option available for adding AF filtering in most commercial receivers I have used.
The second op amp stage is used to increase the headphone volume and the 47K feedback resistor can be adjusted to suit your needs. It is really important to experiment with the component values which will match your receiver and the IF filters and antenna you have. For example the input shunt resistor may be increased from 10 to 18-22 ohms if you always use a narrow IF filter during CW and beacon listening or received signals are low in volume. This is an experimenter's circuit, not a finished project. Overall, this circuit has low output volume and is really gentle on the ears in terms of noise and amplitude.

Project Breadboard and Samples

Shown to the right is a side photo of the experimental project. The big yellow Mallory polyester caps were used as I did not have any other desirable AF filtering caps in my parts stock. The day after, I built this filter, my parts order (including a big selection of polyester film audio caps) from Digi-Key arrived, however this is Murphy's law. These Mallorys are good quality capacitors- just a little large!

Low pass filters can really help reduce noise during reception. Two example audio files follow. These were heard in the NDB band using a 2.4 KHz SSB IF filter.
290 YYF
312 UNT
The occasional scratchy noises are me moving as I held the headphones around the microphone.

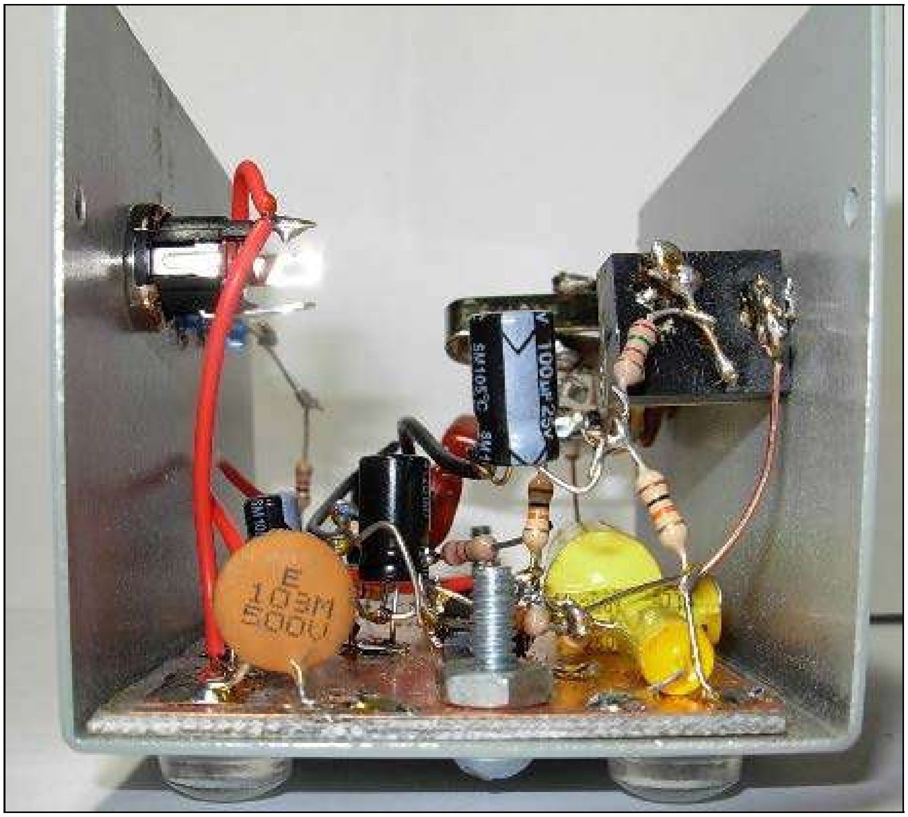

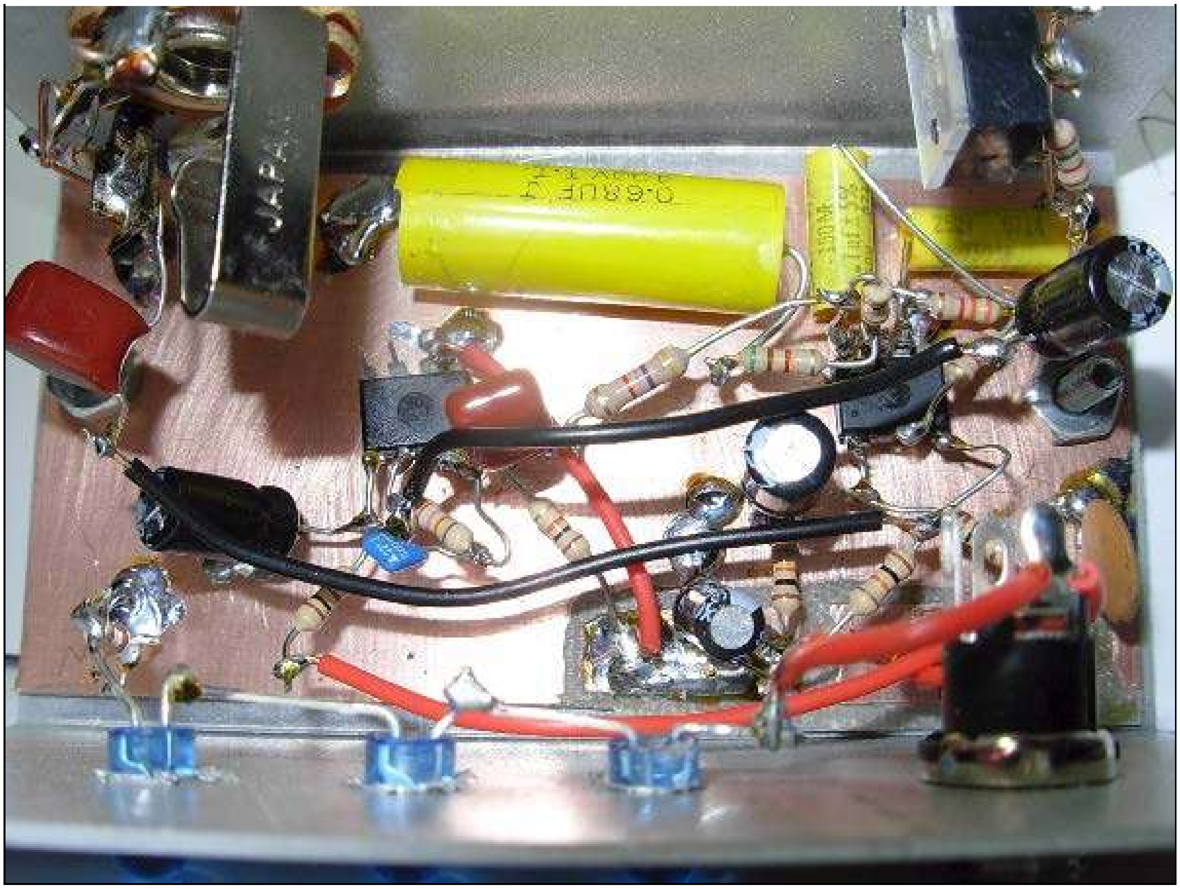

Above is the filter photographed from above. The 3 blue LEDs are used to light the lower row of buttons on the R75 as it is difficult to see them with a low level of light in the radio shack. No hum is heard with this filter. When you build AF amps or filters with the 5532, after soldering pin 4 to the copper ground plane, start out by connecting the components associated with pins 6 and 7 and then 1 and 2. I suggest this as placing components between adjacent pins is often the most difficult part of building when using ugly construction with op amps.

If by accident you make an unintentional solder bridge between 2 adjacent pins on an op amp, simply heat the 2 bridged pins up and gently drive a small screw driver between the pins. This should remove your unwanted solder bridge. The 5532 op amp is quiet and relatively inexpensive. In EMRFD, there are countless examples of how one can use them in a variety of applications.

Additional Information and Photographs

Above is a photo of the audio filter in action on the NDB band. The blue LED reflection can be seen on the receiver. For serious NDB and CW pile up work, narrow band pass audio filters are generally required. This simple audio filter experiment might be useful as a spring board for your own AF filter experiments and to learn the filter requirements of your own particular receiver.

Above is the filter photographed from the rear. The DC power cord has a built in RFI filter. For homebuilt projects, DC power cords can be obtain by cutting the power cord off old unwanted or broken "wall wart" power supplies. This provides you with a nice cord with a built-in plug. I collect old AC "wall wart" transformers for this purpose. This filter is powered by the main 12 volt DC supply on my radio bench.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Cascode Hybrid-Based WWV Receiver for 5 MHz

Introduction

This was my favorite project of 2007. When I web published the original TRF WWV receiver for 10 MHz in 2006, there were many complaints that I used hard-to-find dual-gate MOSFETs and also that the AF stage lacked the popcorn factor that this web site has become strongly associated with. In this experimental project, these 2 concerns are addressed.

The cascode JFET and BJT amplifier stage used in this receiver is based upon the amplifer described in the Hybrid Cascode IF Amplifier article which was published in QST for December 2007 and designed by W7ZOI and WA7MLH. This amplifier topology has many advantages including high gain + low noise, that it can function well at DC voltages less than 12 VDC and that the noise figure does not degrade when the BJT bias (and stage gain) is lowered during AGC action. Please read the QST article and also refer to the W7ZOI web site for more details on the IF amplifier and the cascode hybrid topology.

RF GAIN
Volume

Receiver Block Diagram

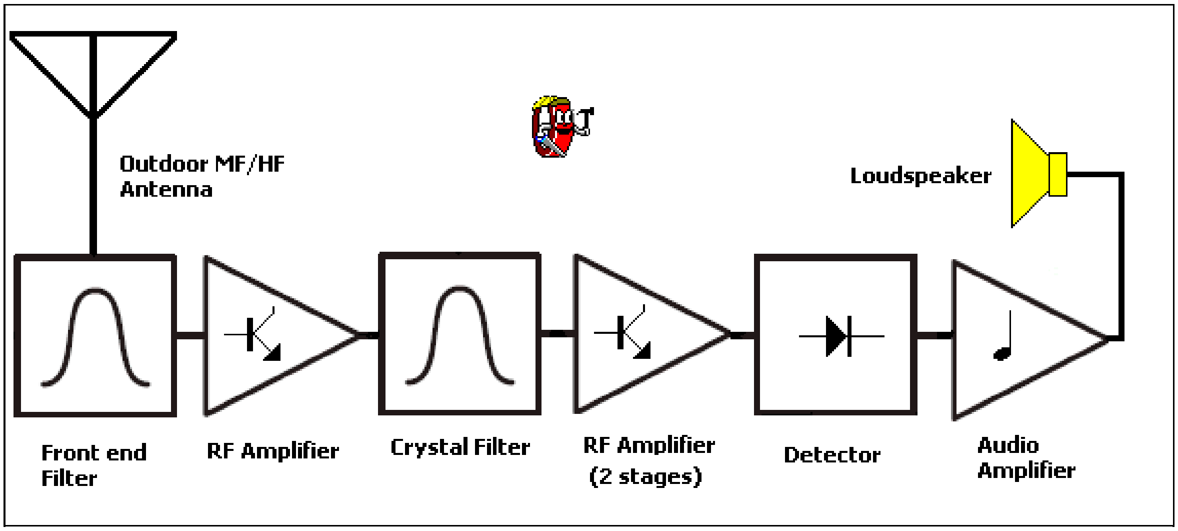
Outdoor MF/HF Antenna
Loudspeaker
Front end Filter
RF Amplifier
Crystal Filter
RF Amplifier (2 stages)
Detector
Audio Amplifier

The receiver block diagram is shown above. To hear digitally recorded examples of this receiver, click here , here or here. No attempt was made to make these files sound better than they really are- there is signal fading, room noise etc. A electret microphone was placed near the loud speaker to record these audio samples. *Note I am now compressing audio files in the mp3 format to allow listening by those who use Linux as their operating system.* **A supplemental web page** to this main web page is linked here

---

Receiver Front End: Band Pass Filter and First RF Amplifier

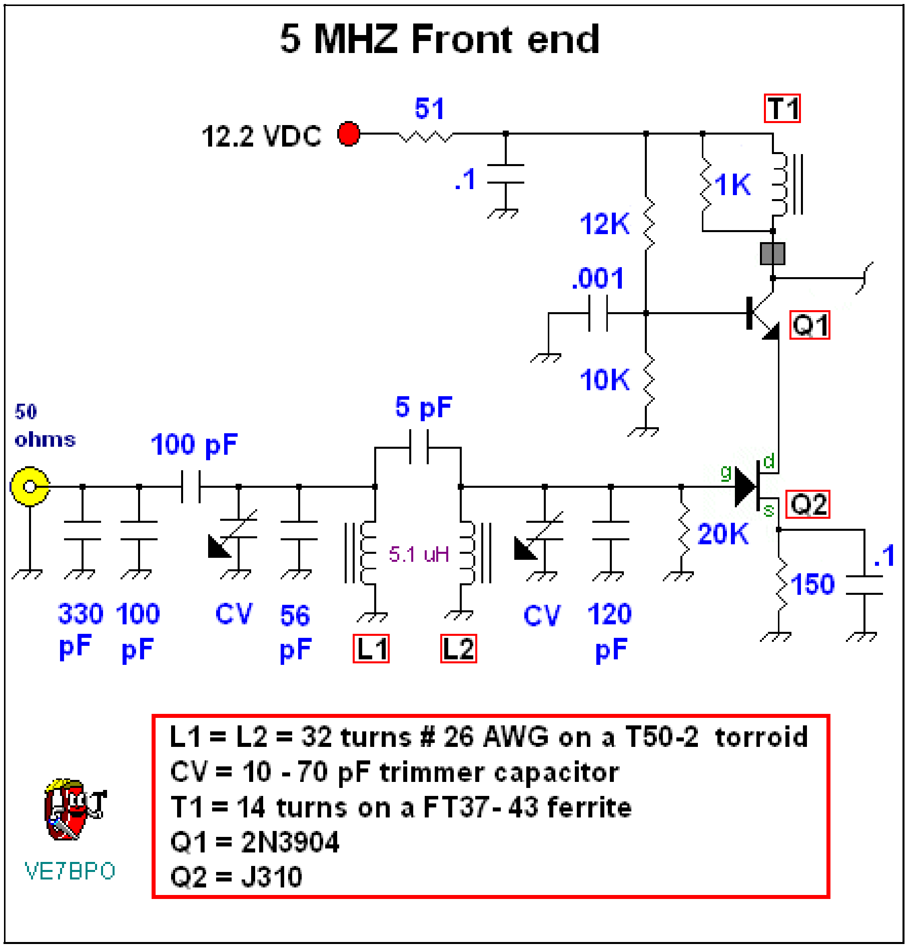
5 MHZ Front end
12.2 VDC
51
T1
.1
1K
12K
.001
Q1
10K
5 pF
50 ohms
100 pF
g
d
Q2
s
20K
.1
150
5.1 uH
330 pF
100 pF
CV
56 pF
L1
L2
CV
120 pF
L1 = L2 = 32 turns # 26 AWG on a T50-2 torroid
CV = 10 - 70 pF trimmer capacitor
T1 = 14 turns on a FT37-43 ferrite
Q1 = 2N3904
Q2 = J310
VE7BPO

This receiver is meant to interface with a standard 50 ohm feed line. Testing was performed using my MF/HF antenna.
I built 3 separate bread boards of this receiver and tried varying the number of RF amps, using different detectors (as well as different detector followers) and eventually built and tested this basic receiver design for 5, 6 and 10 MHz. With respect to using the cascode hybrid amp (and probably any other amplifier type) **in a TRF receiver**, I learned 3 things:

1. Do not operate the RF amps at maximum gain. I built some very powerful amps with a Q2 source resistor of 47 ohms and over 6 volts bias on Q1. While powerful, this amp broke into oscillation and also consumed much current (nearly 20 ma).

2. Keep the RF stages at least 2-3 cm apart to reduce the chance of parasitic oscillations.

3. Keep the input band pass filter at least 2 cm from the Q1/Q2 amp or you might encounter some unwanted oscillations.

For the front end band pass filter, a reasonably narrow bandwidth was desired. When sweeping early filter designs using a signal generator and oscilloscope, a double humped response was noted. These filter designs used a 10 pF coupling capacitor. The coupling capacitor was then decreased to 5 pF. To obtain the required 5 pF, two 10 pF capacitors were placed in series as shown in the photograph directly to the right.

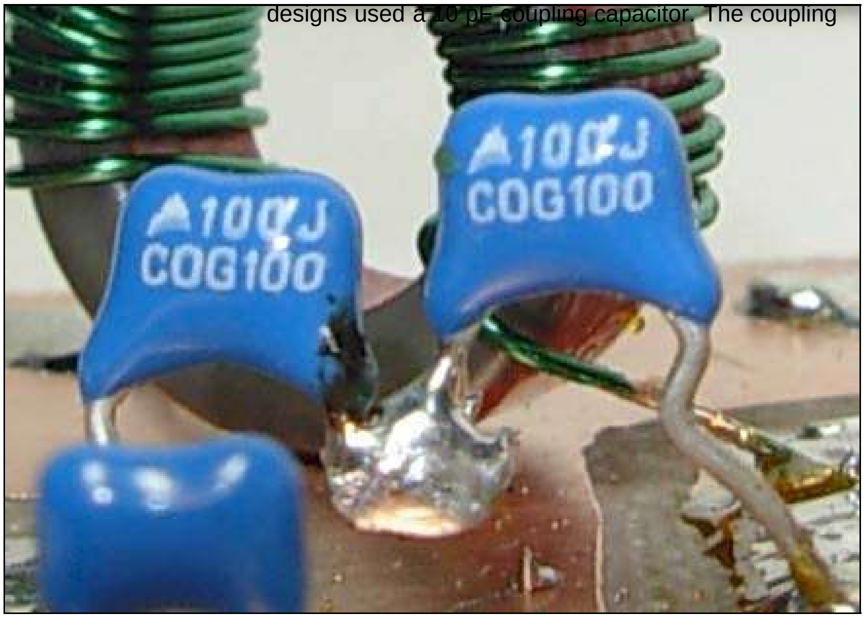
A100J
COG100
A100J
COG100

I struggled with this filter design because one end is terminated in the gate resistance of Q2 of the hybrid cascode amplifier and was not the standard 50 ohm impedance termination. My early filter designs suffered severe insertion loss or poor selectivity. I asked Wes, W7ZOI, for some instruction on solving my filter problems. I learned that this filter topology is referred to as a singly terminated, double tuned band pass filter. Wes designed the front end band pass filter for the 5 MHz receiver for us all to learn from and for this I am very grateful to him.

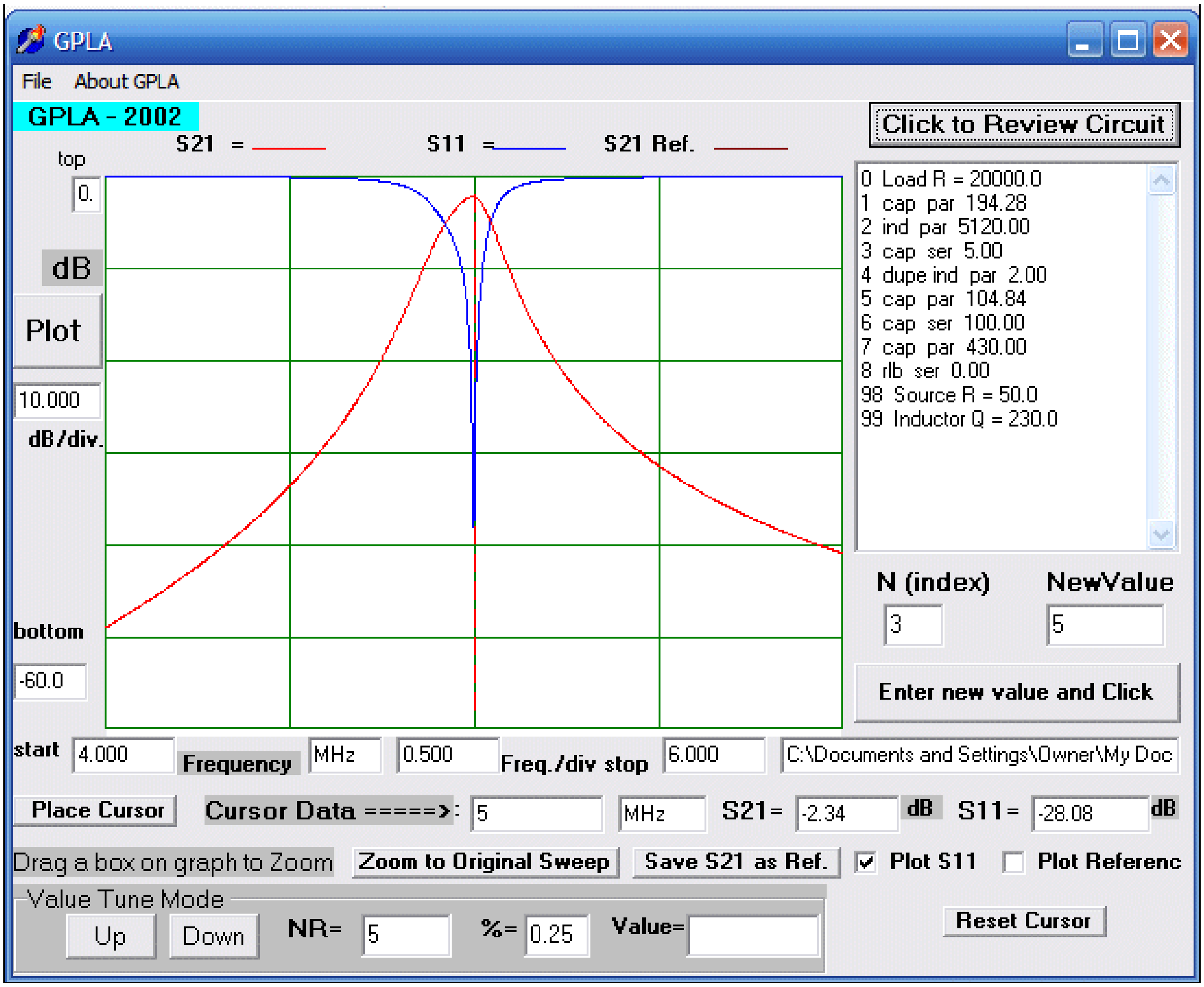
GPLA
File
About GPLA
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
top
dB
Plot
dB/div.
bottom
N (index)
NewValue
Enter new value and Click
Place Cursor
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Plot Referenc
Value Tune Mode
Up
Down
Reset Cursor

0 Load R = 20000.0
1 cap par 194.28
2 ind par 5120.00
3 cap ser 5.00
4 dupe ind par 2.00
5 cap par 104.84
6 cap ser 100.00
7 cap par 430.00
8 rlb ser 0.00
98 Source R = 50.0
99 Inductor Q = 230.0

top: 0.

10.000 dB/div.

bottom: -60.0

N (index): 3

NewValue: 5

start 4.000 Frequency MHz 0.500 Freq./div stop 6.000 C:\Documents and Settings\Owner\My Doc

Cursor Data =====>: 5 MHz S21= -2.34 dB S11= -28.08 dB

☑ Plot S11 ☐ Plot Referenc

NR= 5 %= 0.25 Value=

Above. A GPLA simulation of the singly terminated, double tuned filter designed by W7ZOI. A double tuned circuit is mandatory ahead of the WWV receiver as local BCB and other RF energy will be amplified by the first RF amp and may distort the WWV signal in the crystal filter or even might blow-by the crystal filter and be detected and heard in the speaker.

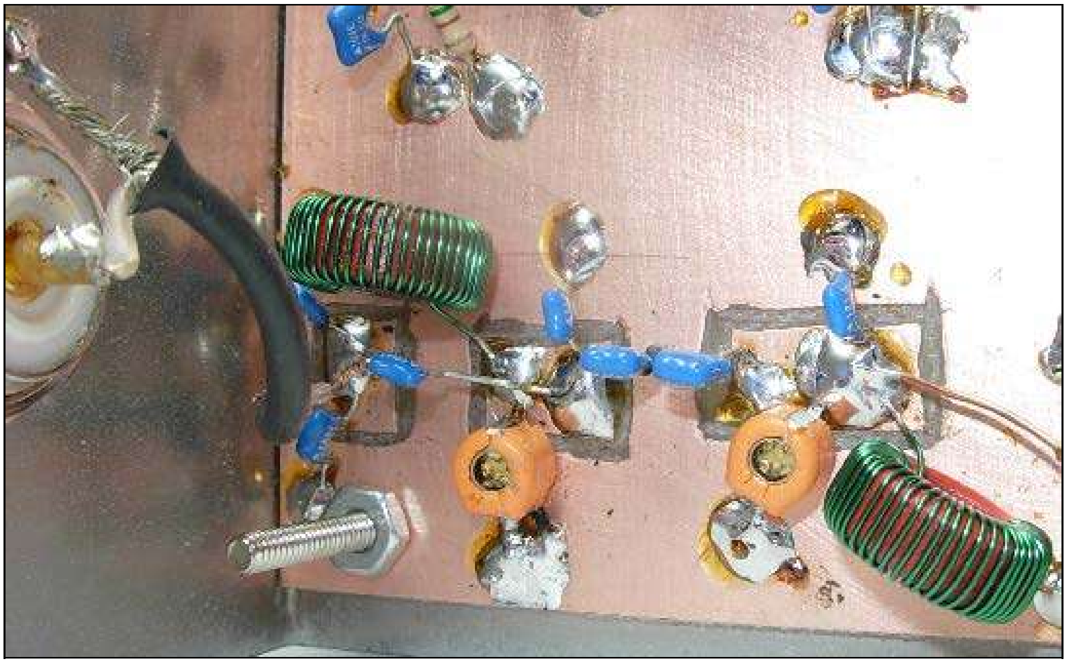

Directly above is a close up photograph of the input filter bread board. Filter tuning was done by ear (and screwdriver!) Simply tune the trimmer capacitors for the loudest audible WWV pulses in the speaker and you are set. If you can't locate a 20K gate resistor for Q2, a 22K resistor will work okay.

Crystal Filter and Second RF Amplifier Stage

In the schematic to the right is the crystal filter and second RF amplifier. The input impedance of the crystal filter is established by the 1K shunt resistor across the output transformer on Q1. The output impedance of the crystal filter is set by the 1K gate resistor of Q4. A filter input/output Z of 1000 ohms gave the best overall shape and bandwidth during my testing.

Developing this filter was difficult. My first batch of junk box crystals had a low motional inductance and with the filter I built I could hear stations ~400 KHz below and/or above the filter center frequency in addition to WWV. After giving up in frustration for nearly 2 months, a batch of 10 crystals were ordered from Digi-Key. These were microprocessor crystals; ones with 18 pF load capacitance in a HC49/U holder. The new filter was tweaked and tested and now provides single signal reception of WWV. Your own results may vary according to your crystal parameters. The Digi-Key part number is provided for reference purposes only.

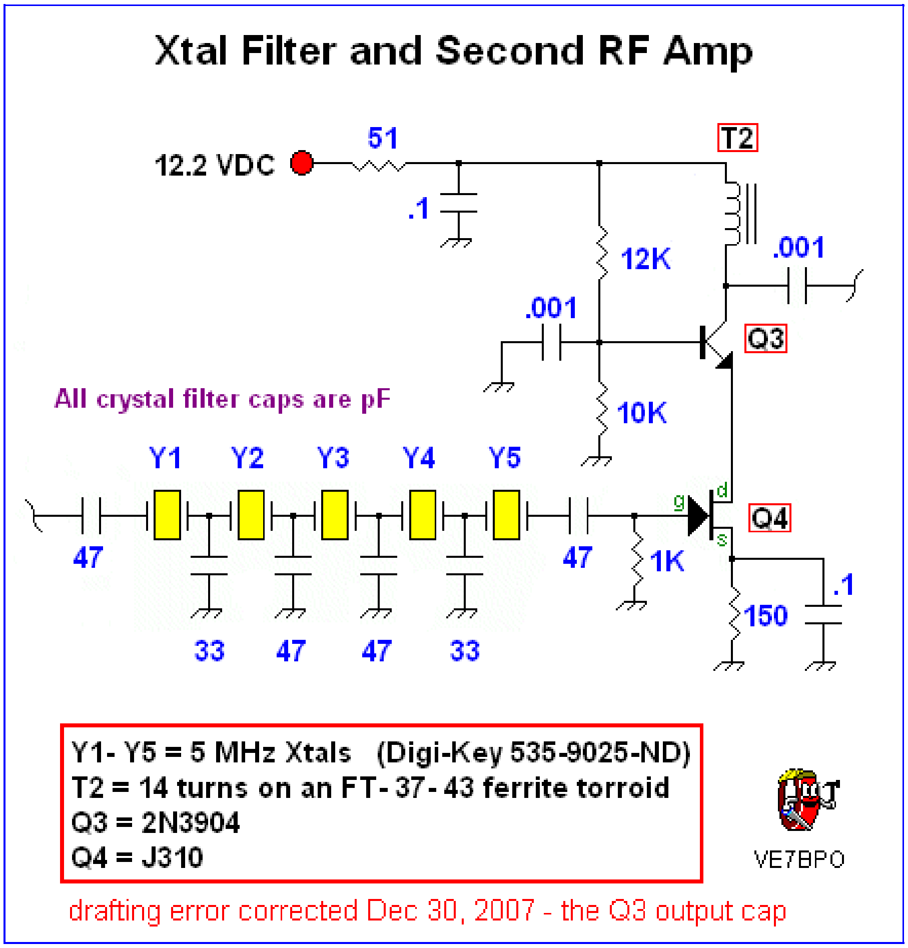
Xtal Filter and Second RF Amp
12.2 VDC
51
.1
T2
12K
.001
.001
Q3
All crystal filter caps are pF
10K
Y1
Y2
Y3
Y4
Y5
g
d
Q4
s
47
47
1K
.1
150
33
47
47
33
Y1- Y5 = 5 MHz Xtals (Digi-Key 535-9025-ND)
T2 = 14 turns on an FT-37-43 ferrite torroid
Q3 = 2N3904
Q4 = J310
VE7BPO
drafting error corrected Dec 30, 2007 - the Q3 output cap

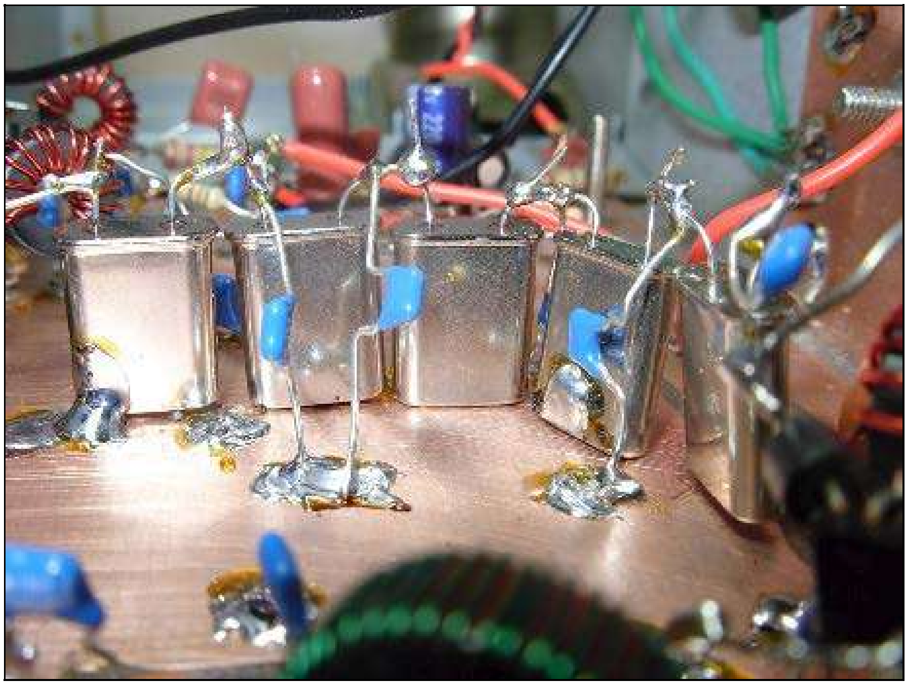

To the left is a close up photograph of the 5 MHz crystal filter. The crystals were turned upside down and the outer cases were directly soldered to the copper ground plane as you can easily see in the crystal to the left of the others. The rest of the crystals as well as one of the 47 pF tuning capacitors were soldered on the other side and solder points are hidden from view. The crystals were positioned to keep the output of Q1 away from the input of Q4. ***Stage lay out is very important in TRF receivers.*** I found stage layout to be far more important than keeping lead lengths short from my experimentation.

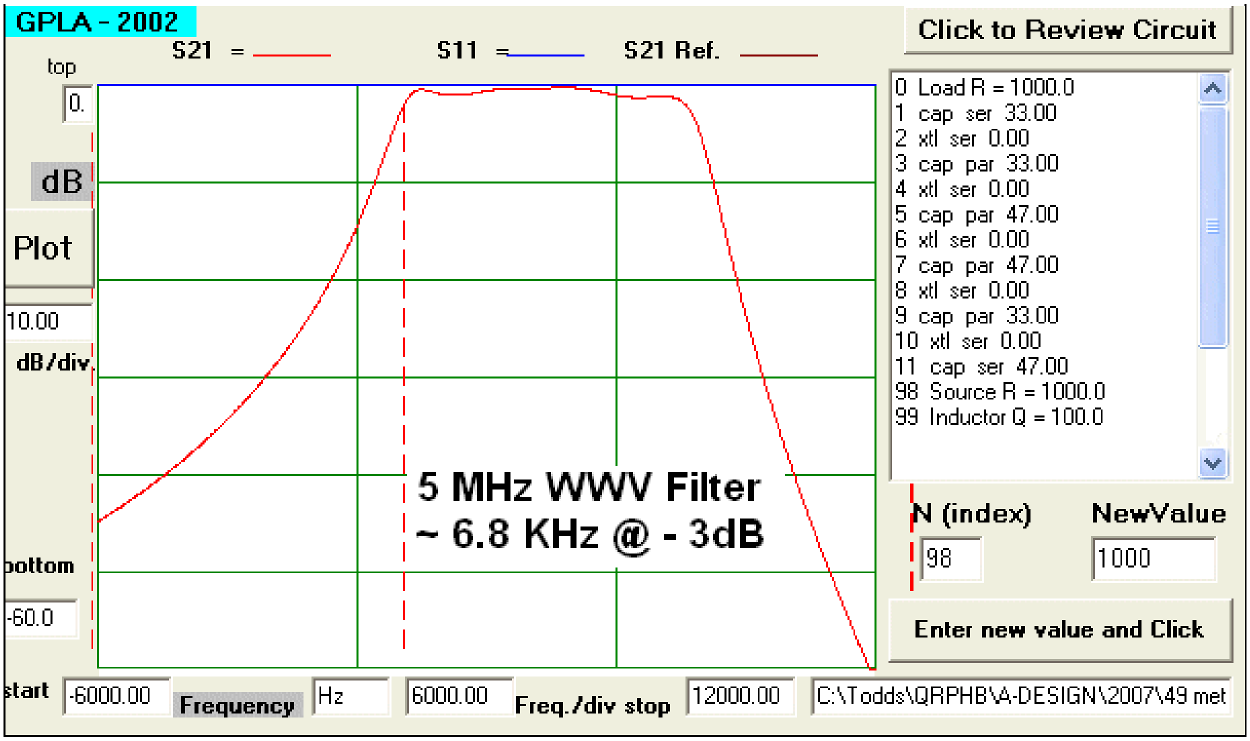

```
0 Load R = 1000.0
1 cap ser 33.00
2 xtl ser 0.00
3 cap par 33.00
4 xtl ser 0.00
5 cap par 47.00
6 xtl ser 0.00
7 cap par 47.00
8 xtl ser 0.00
9 cap par 33.00
10 xtl ser 0.00
11 cap ser 47.00
98 Source R = 1000.0
99 Inductor Q = 100.0
```

top: 0.
10.00 dB/div
bottom: -60.0
start: -6000.00
Frequency: Hz
Freq./div: 6000.00
stop: 12000.00
C:\Todds\QRPHB\A-DESIGN\2007\49 met
N (index): 98
NewValue: 1000

Directly above is the GPLA simulation of my crystal filter. The 5 MHz point is not centered exactly in the middle of the pass band, but a reasonable AM filter was built nonetheless. Crystal parameters, especially motional inductance and capacitance can make or break your filter. Motional inductance and capacitance describe the L and C values that make up the crystal's electrical LC model. Very large inductive and capacitive reactance values at the specified operating frequency give the crystal its extraordinarily high "quality factor" or "Q". For example, If the motional L is too low, your filter may not work as expected; providing single signal reception of WWV. The Lm was 0.02 and the Cp was 5 in the crystals which I used for my filter. In general, low Q crystals will give poor results. Oppositely, crystals with very high Q may give a lower then expected bandwidth and this may reduce AM receive fidelity. Experimentation is necessary.

Third RF Amplifier Stage and Detector

To the left is schematic of the final RF stage and the envelope detector. This RF stage has variable gain by means of a front-panel mounted 10K potentiometer which is used to vary the bias on Q5. The input Z of this stage is 100 K and is set by the Q6 gate resistor. The output of Q5 is AC coupled to a detector designed by Wes, W7ZOI. I performed considerable experimentation with basic diode detectors as well as detector source followers; some of which I sent to Wes for his consideration. He designed and emailed me back this simple, good sounding detector design which uses the gate voltage of Q7 to bias the germanium diode. Other types of diodes such as as hot carrier diodes will likely not have the output voltage of the Germanium type. Germanium diodes, when biased, had more noise and high frequency response in addition to higher output when compared to others I tried during my experiments. Diode detector guru, Felix, VK4FUQ advised me of an excellent diode he is now using called the BAT46. The audio samples of a local AM radio station using this diode and his other hi-fi lab equipment that he sent me are beyond fantastic.

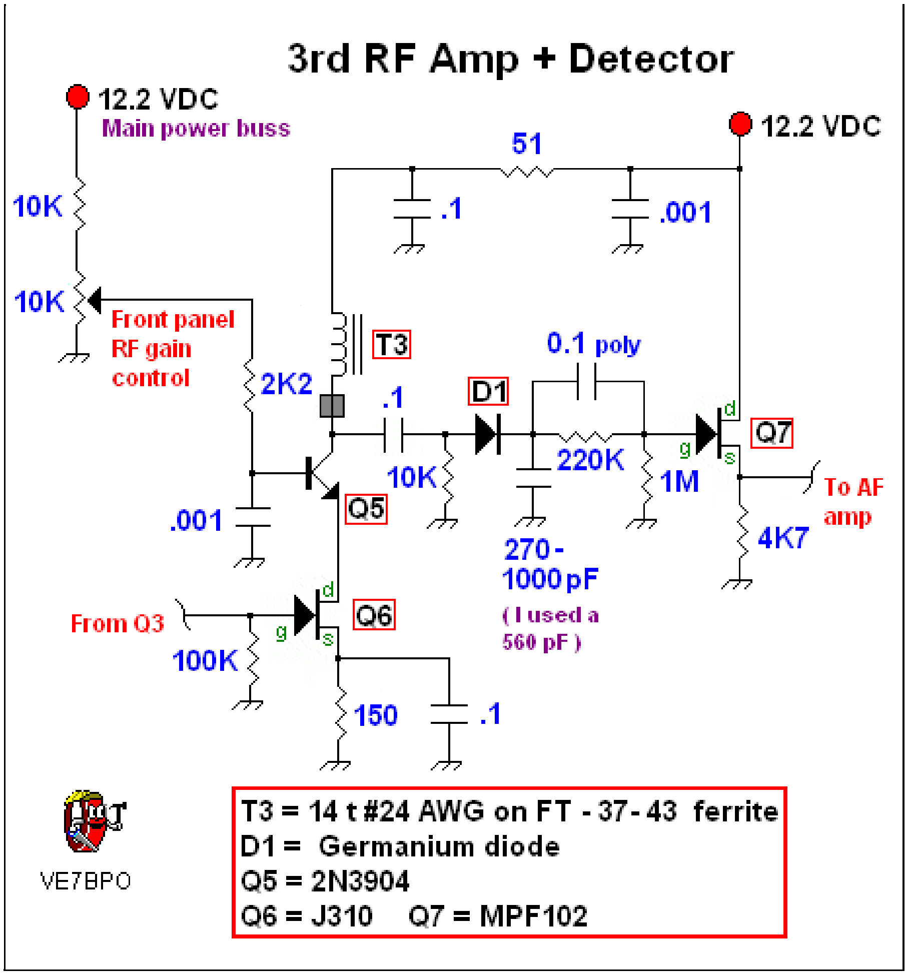
3rd RF Amp + Detector
12.2 VDC
Main power buss
51
12.2 VDC
10K
.1
.001
10K
Front panel
RF gain
control
T3
0.1 poly
2K2
.1
D1
d
Q7
g
s
220K
10K
1M
To AF
amp
.001
Q5
270 -
1000 pF
4K7
(I used a
560 pF)
From Q3
d
Q6
g
s
100K
150
.1
T3 = 14 t #24 AWG on FT - 37 - 43 ferrite
D1 = Germanium diode
Q5 = 2N3904
Q6 = J310 Q7 = MPF102
VE7BPO

The photograph on the right is a close up of the enveloped detector designed by W7ZOI. The germanium diode was purchased from The Source in Canada (Radio Shack in the USA). The blue, partially hidden shunt capacitor is a multi-layer ceramic 560 pF cap. The other capacitors are metalized, polyester film types. Ensure correct diode polarity.

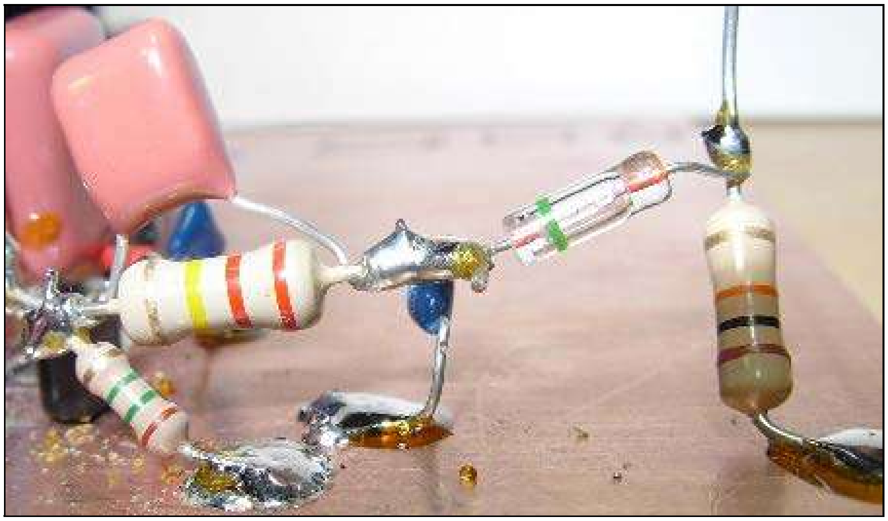

Audio Stage

To the left is the schematic of the audio stage. The very "popcorn" LM386 AF chip is used to please the audience who complained about my AF stages not having enough popcorn factor. A 4K7 resistor was inserted between pins 1 and 8 to reduce the gain somewhat. Thus, the LM386 is still being operated in the high gain mode but won't hurt your ears with loud noise and distortion. The 470 pF cap on pin 3 may be changed or eliminated. It is a simple low pass filter.

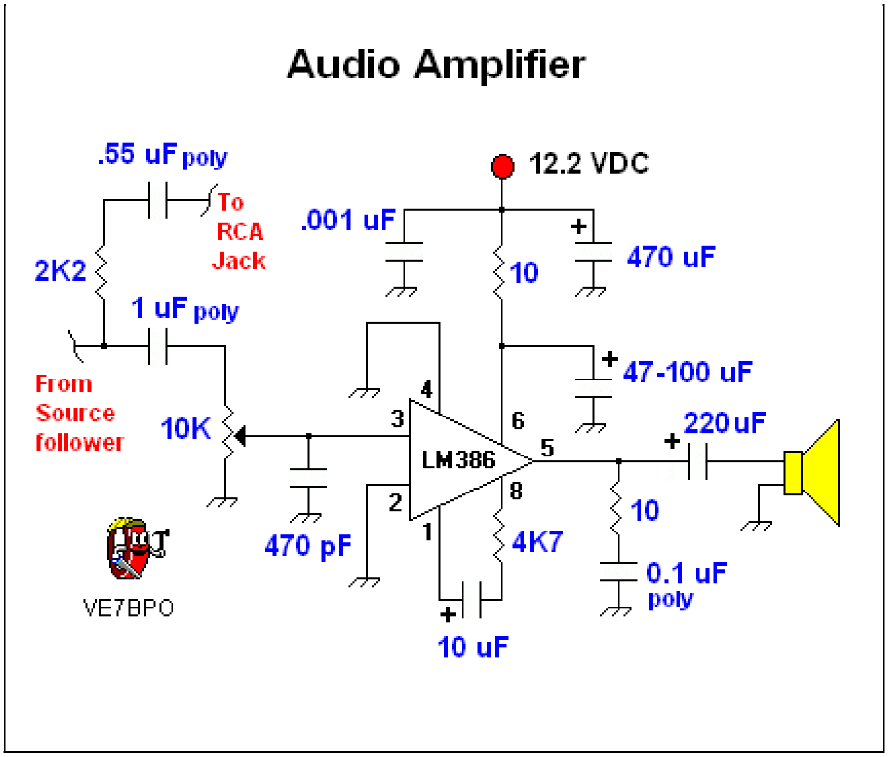

A secondary, audio output connects to a front panel mounted RCA phono jack. This allows me to use my lab grade (KK7B AF amp) and turn the audio off on the normal receiver AF amp.

The photograph to the right is a close up of the LM386-based audio stage. This is where I started. After drilling the chassis, wiring the speaker, installing the chassis potentiometers, making the main power buss and LED indicator, the AF stage was built on the main board. The main board was then temporarily soldered in and tested.

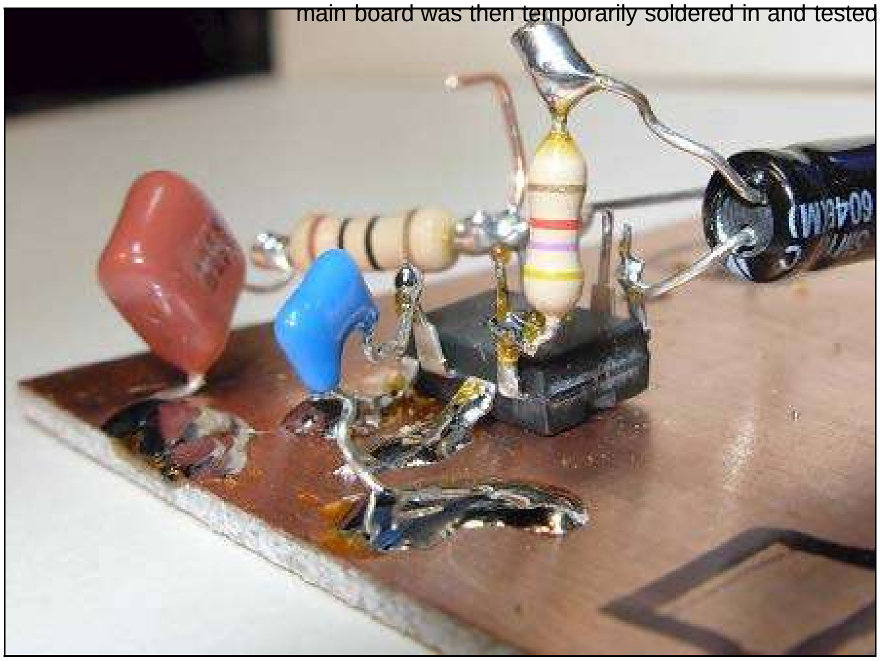

(Some connections were made via alligator clips such as the speaker wires). When the AF amp worked as expected, the main board was removed from the receiver chassis and the net stage was built. Up next were the detector and source follower. After bread boarding these, again the main board was laid in the chassis, wired, tested and then removed when all was functioning well. To test the detector I touched my finger to the input and heard local BCB radio. Following this, RF Amp #3 was added to the main board and again the main board was temporarily wired up and tested by touching the input of Q6 with my finger and observing that a local broadcast radio station increased/decreased in amplitude when the RF gain control was turned up and down. DC voltages were also measured and considered from project start to finish.

Actually, all you need to do is connect a band pass filter such as this to Q6 and the components after and you will have a nice TRF BCB AM radio. Each successive stage was built and tested, so when the receiver was finished, I already knew that it worked. I cannot emphasize enough how important it is to build your receiver backwards and test each stage as you go. There is strong temptation to start at the antenna connection and work until you get to the speaker, but please consider doing the opposite.

The bare copper wire in the photograph is the positive connection point for the speaker wire. It was trimmed somewhat during final assembly to reduce the possibility of it shorting.

Miscellaneous Photographs

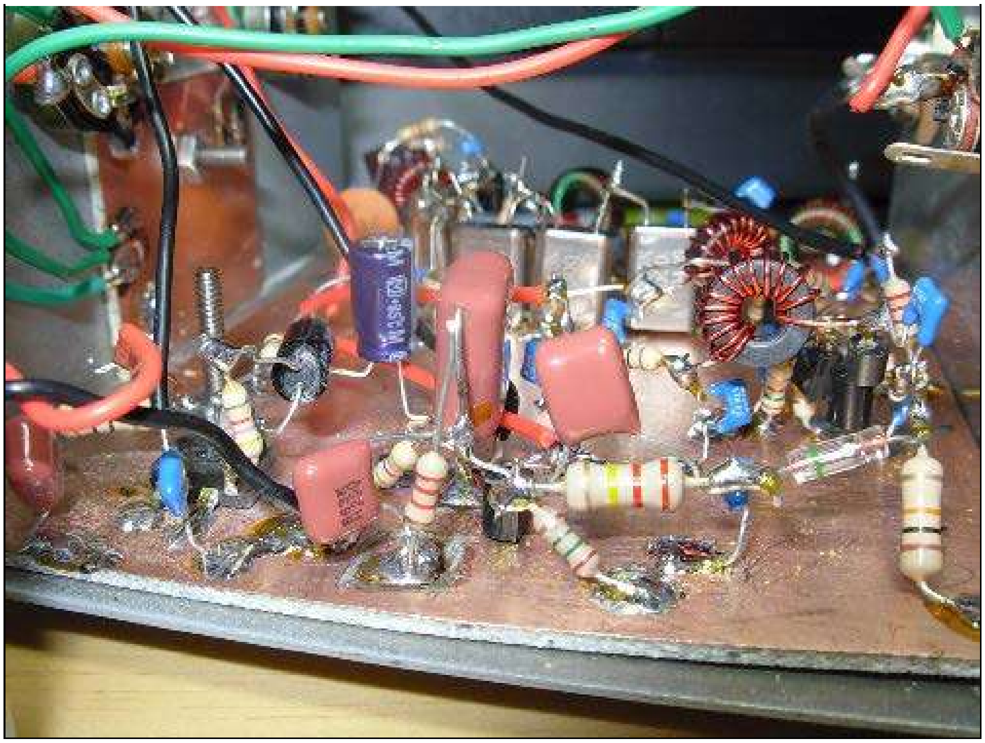

The photograph above shows some of the detail of the receiver main board from the right hand side which contains the detector, source follower and audio amp stages from right to left.

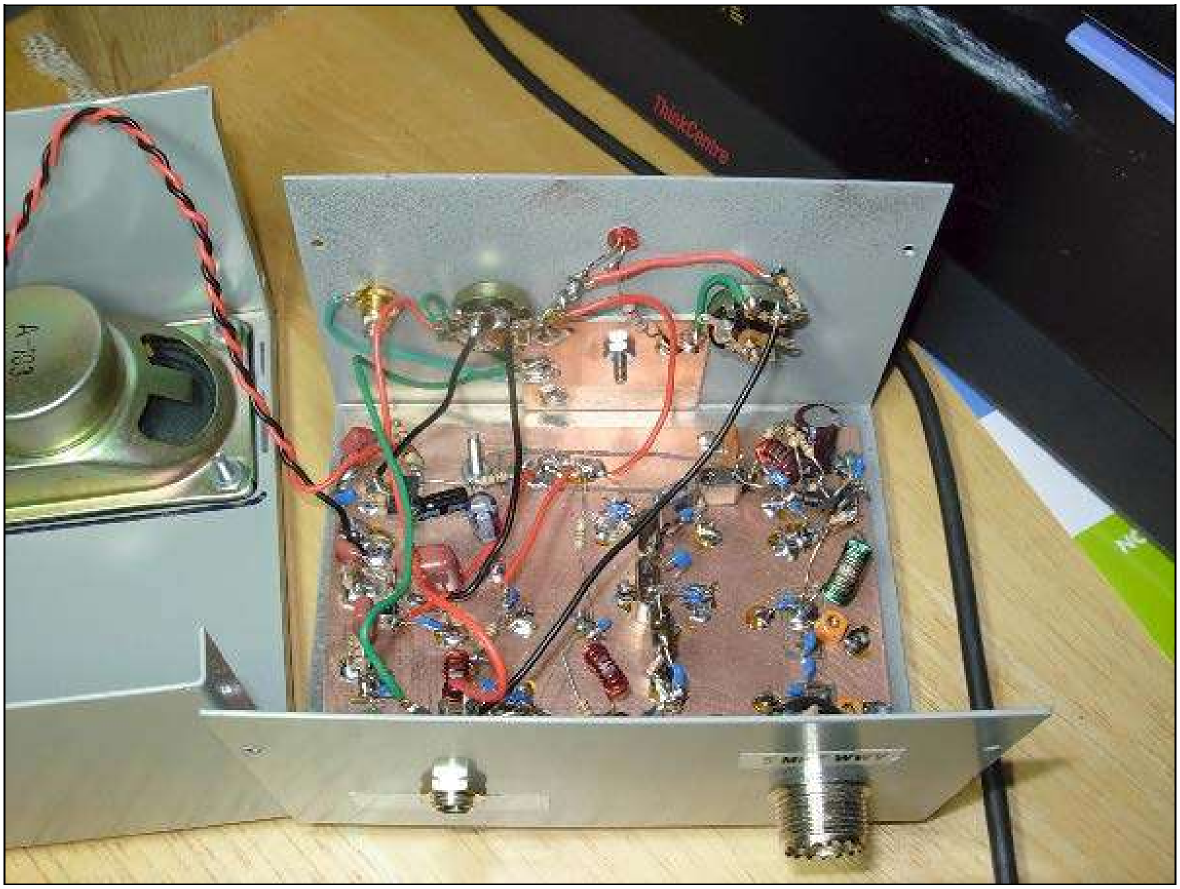

The photograph above shows a top view of the main chassis and also the chassis cover with the speaker bolted on and wired up.

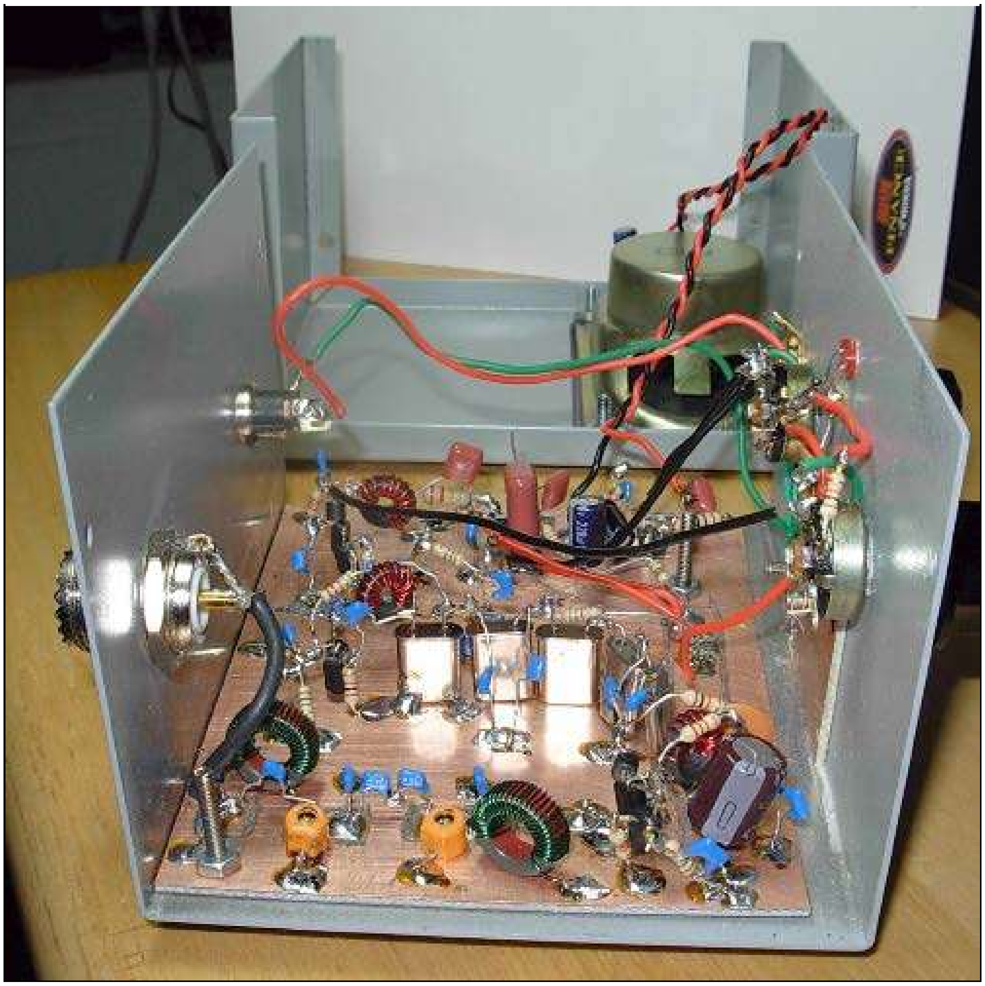

This wider angle photograph shows the main board from the left side. From left to right in the nearground are the SO-239 antenna connector, LC band pass filter and first RF amplifier.

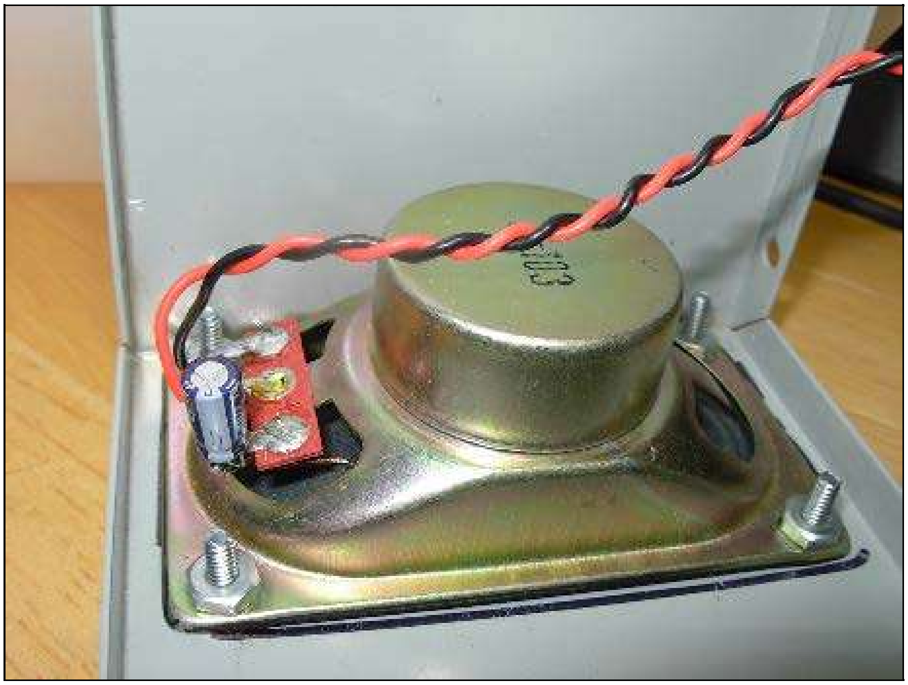

The photograph above shows the speaker attached to the Hammond chassis top. Holes were drilled in the chassis lid with a drill press to allow the sound to pass through.

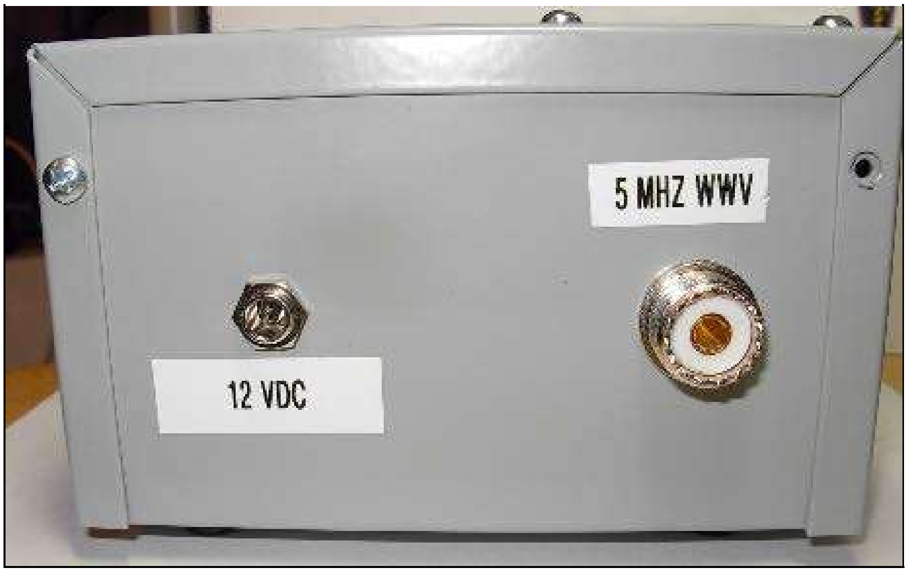
5 MHZ WWV
12 VDC

The photograph above shows the reverse view of the receiver chassis.

RF — Test and Measurement

Complementary-Symmetry Amplifier Biasing Basics

Introduction

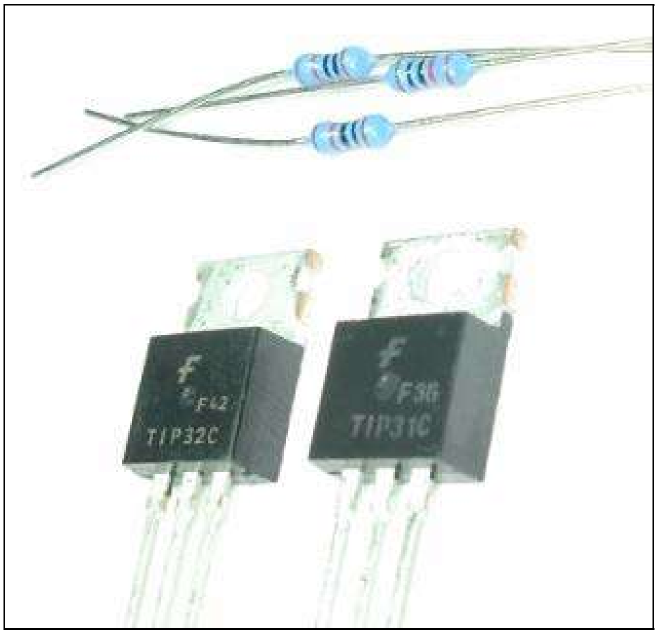

This page provides information concerning the biasing of Class-AB, complementary-symmetry audio amplifiers. These schematics should be considered theoretical, as design considerations such as thermal stability, negative feedback and component power ratings are minimized or excluded for sake of clarity.

The basic 2 transistor complementary-symmetry amplifier may be used as a simple low power AF amp or as a building block for a high powered stage such as a 50 watt guitar amplifier. It is important to understand how to properly bias your AF power amps to reduce distortion and to promote easy troubleshooting when problems arise. This web page describes the hows and whys of biasing in a progressive manner with minimal math.

Discussion

Figure 1

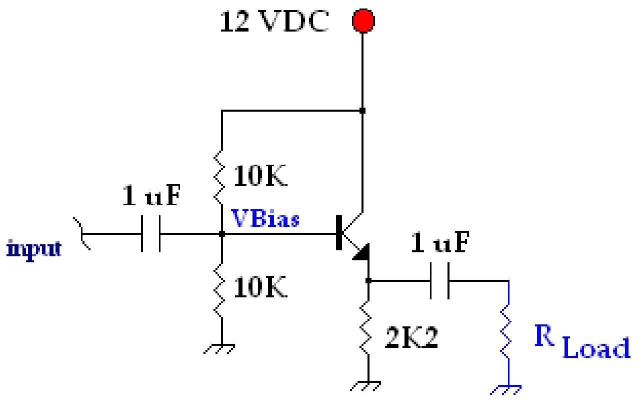

A review of the common collector amplifier (which is more commonly called the emitter follower) is a good place to start. We may refer to the complementary-symmetry transistor pair as complementary emitter followers since they are an NPN and PNP emitter follower connected in series. An emitter follower amp is shown in Figure 1. Its properties include:

✓ input on the base - output on the emitter
✓ high input impedance and low output impedance
✓ a voltage gain of 1
✓ good current and power gain

In an appropriate configuration, these qualities are perfect for driving a low impedance load such as an 8 ohm speaker with large output currents that are not provided by our typical transistor or op amp voltage amplifier stages. In many cases, we bias the emitter follower with a voltage divider network comprised of 2 identical value resistors. In Figure 1, the voltage divider consists of a series pair of 10K resistors and thus VBias = 6 volts. These 10K bias resistors will be used throughout this web page as the circuits evolve.

In Figure 2 is a pair of complementary emitter followers which have their bases biased with our now familiar series connected 10K bias network for a VBias of 6 volts. When the power is turned on, output capacitor C2 charges through the NPN transistor until it reaches about 6 volts *(theoretical value used to keep things simple)*. When the voltage at point V Emitter reaches 6 volts, the NPN transistor goes into cutoff because VBias voltage now equals the V emitter voltage. Recall that the NPN transistor base must be positive with respect to the emitter for current to flow. Both the NPN and the PNP transistor are in cutoff. This is the amplifier's quiescent state *(assuming no signal is applied to the input via C1)* and is called Class B bias.

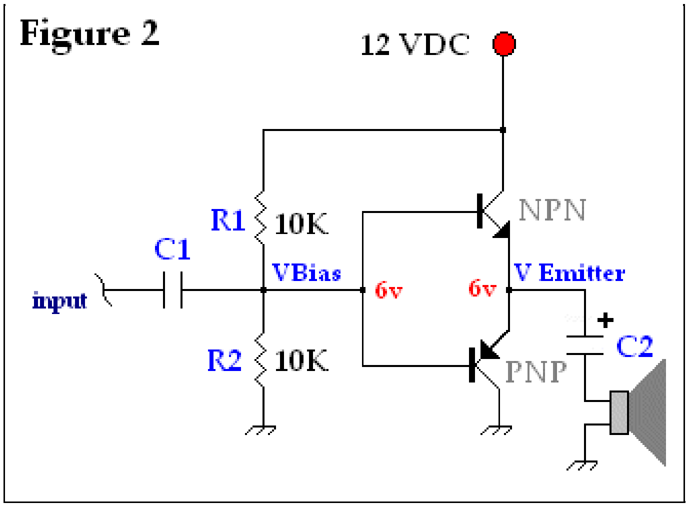

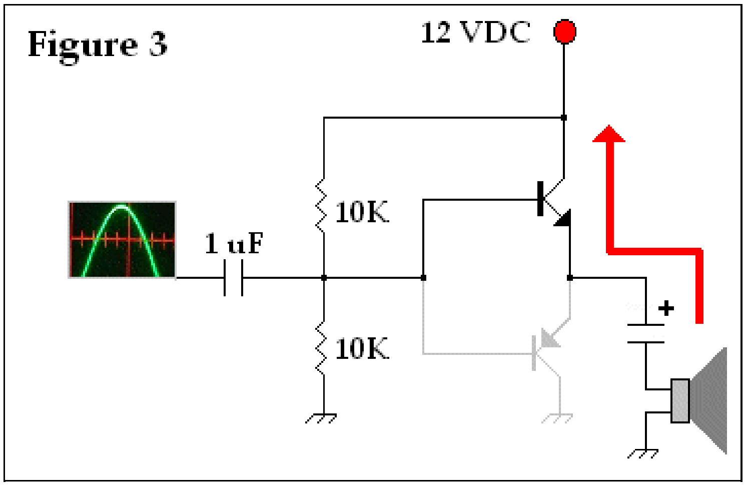

In Figure 3, a positive going signal is applied to the input capacitor. The NPN transistor becomes forward biased and turns ON. Current flows **through the NPN transistor** and charges capacitor C2 to a higher potential. The PNP transistor stays in cut off. The NPN transistor is an emitter follower connected to the speaker.

In Figure 4 a negative going signal the (negative half-cycle) is applied to the input. Q3 turns ON and discharges the output capacitor through the speaker as shown in **red.** The PNP transistor is an emitter follower connected to the speaker.

Thus the NPN and the PNP transistor conduct on alternate half cycles which causes AC current to flow through the speaker. The complementary emitter followers are said to be in push-pull operation.

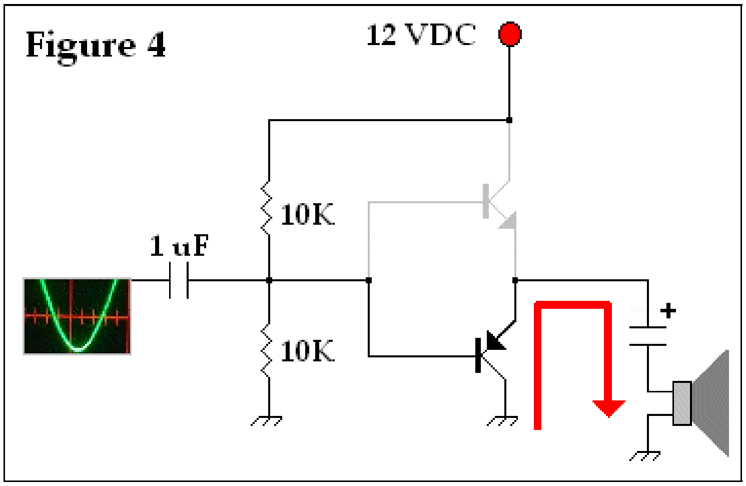

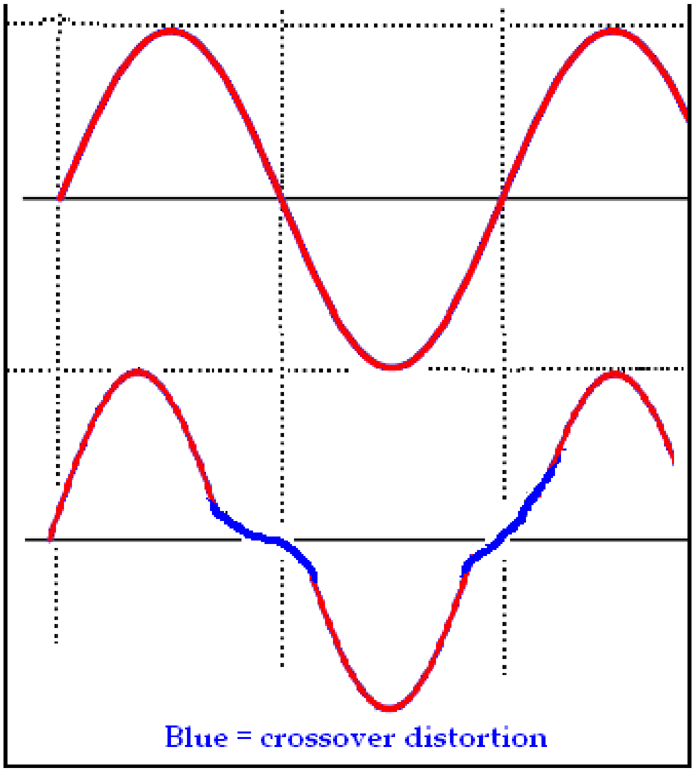

The circuit of Figure 2 has a significant problem; output signal distortion. Silicon transistors such as the 2N3904 and 2N3906 will not conduct until their bases are forward biased by somewhere around 0.7 volts. For the NPN transistor, this means that it will not conduct until the input signal has gone positive by about 0.7 volts. Oppositely, the PNP transistor will remain in cut off until the input signal goes negative by approximately 0.7 volts. As a result, there is a dead zone during the point in time when one transistor cuts OFF and the other turns ON. Shown above is a normal sinusoidal AC waveform in red and another with the distorted waveform of Figure 2 in red and blue. This distortion is called crossover distortion because it occurs at the zero crossing point of the AC waveform. This introduces odd-order harmonics into the output signal. Such is the drawback of the Class B amplifier.

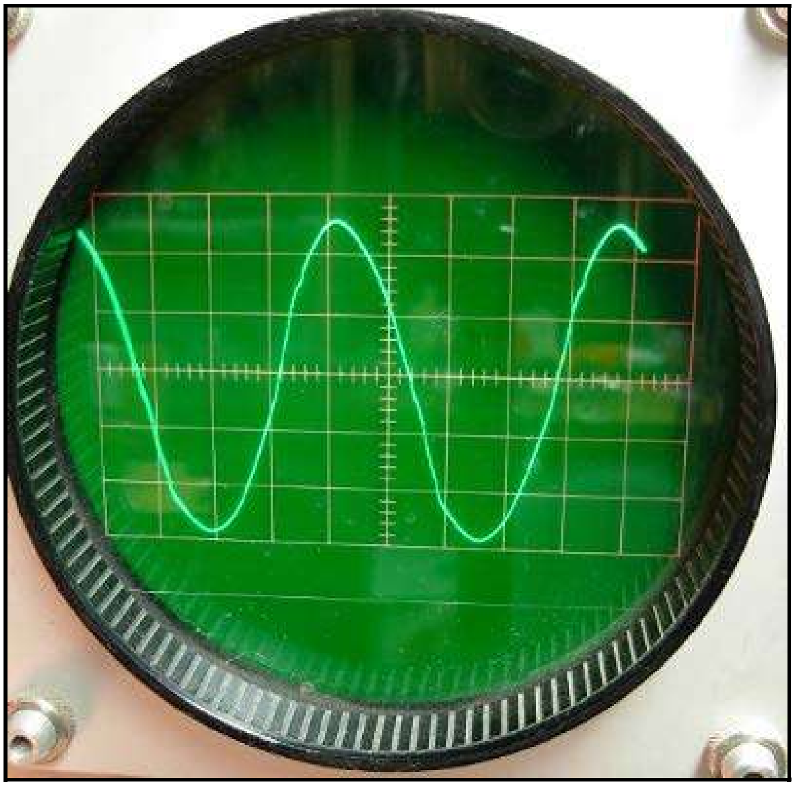

The above photograph shows crossover distortion in an under-biased power amp.

Figure 5

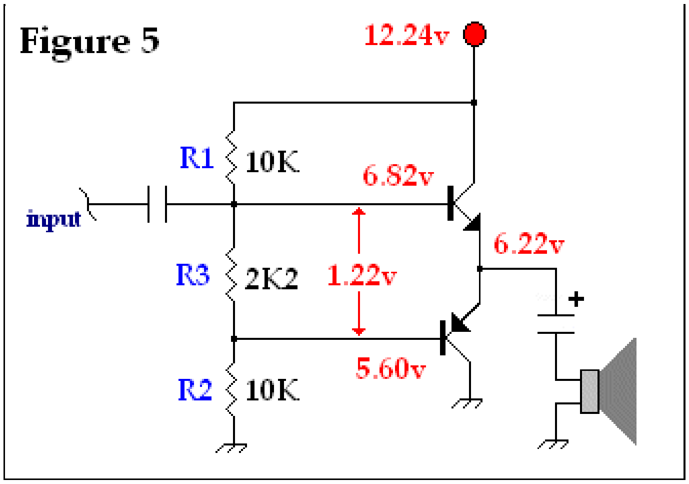

Figure 5 shows the principle technique used to reduce crossover distortion; both transistors are (slightly) forward biased almost to conduction in their quiescent state. As a result, any amplitude of positive or negative going signal will bias the appropriate transistor into conduction. An easy way to achieve this biasing is by adding 1 resistor to our 10K voltage divider network. In Figure 5 is a circuit I built, measured and listened to. R3, a 2K2 ohm resistor was placed in between R1 and R2, our usual 10K bias resistors. As a result , both transistors are forward biased.

That is: the base of the PNP transistor is negative with respect to its emitter and the the base of the NPN transistor is positive with respect to its emitter. As a rule of thumb, you need to drop at least 1 volt across R3. I chose a 2K2 resistor and it worked fine in my particular amp. The Figure 5 biasing topology is rarely used as it puts a series resistance on the PNP input among other problems; however, it exemplifies the basic principles of biasing our complimentary pair. With the forward bias on the transistor pair we now are in Class AB. The output capacitor serves to block the quiescent DC current from flowing through the speaker.

Figure 6

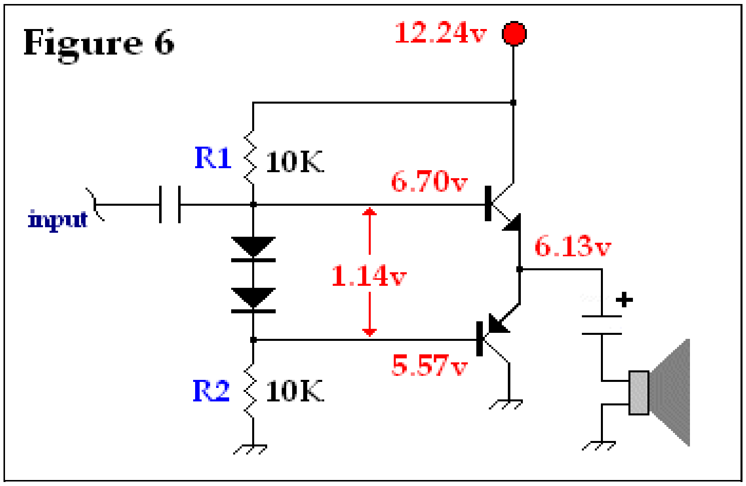

Figure 6 illustrates an improved biasing method over that of Figure 5 by using a pair of silicon diodes. You see this circuit used a lot by hobbyists. The voltage divider consists of 2 resistors and the 2 diodes. The 2 series connected diodes are connected in parallel to the NPN and PNP transistor base-emitter junctions which serves to keep the transistors turned on slightly. The net effect of the diode pair is the same as R3 in Figure 5. The voltage drop per diode was measured at 0.57 volts. The AC resistance of these 2 forward biased diodes is non-significant. There is major problem with the diode/resistor voltage divider; no way to adjust the diodes forward voltage drop. If each diode's forward threshold voltage is unequal to the base-emitter junction voltage of each transistor, either not enough forward bias is applied, or the 2 transistors may be turned on too much reducing efficiency and possibly cause excessive heating. Additionally, the pair of diodes lack the ability to provide temperature compensation when the transistors get hot.

Figure 7

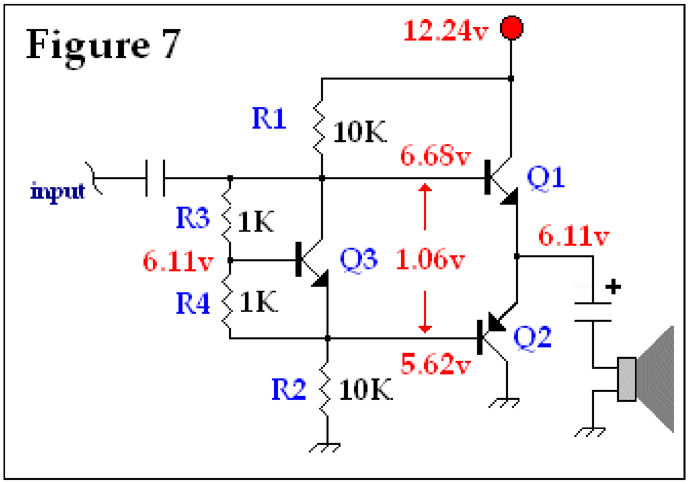

Figure 7 shows the best way to bias our complimentary pair. Our familiar 10K-10K voltage divider is kept, but a transistor Q3 with its own biasing resistors R3 and R4 are added. You might think of R3 and R4 as a voltage divider within a voltage divider. Q3 is referred to as an amplified diode or DC level shifter. It often receives local thermal feedback from the power follower output transistors. This usually involves mounting Q3 on the same heat sink as the finals. If the output transistors heat up, so does Q3 and this results in a a smaller voltage drop across Q3 which translates into less forward bias to Q1 and Q2. Within limits, Q3 with its own base-emitter junction provides variable forward bias for the output transistors.

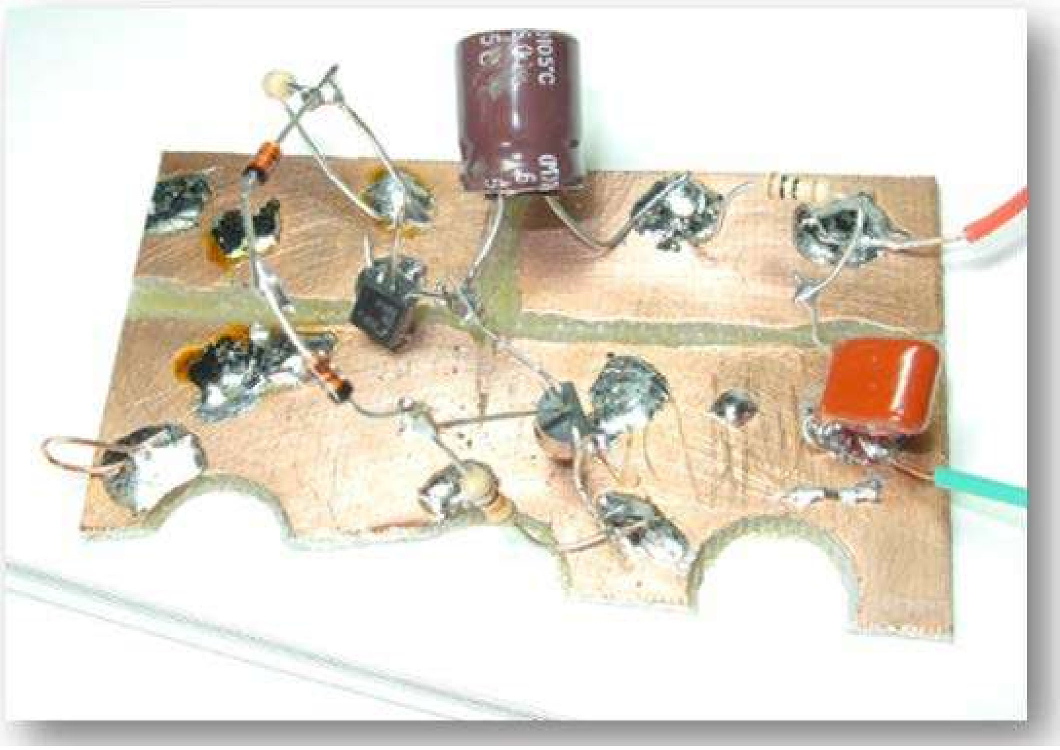

Shown above is the breadboard of the Figure 6 circuit built on scrap of copper clad board. Transistors were 2N3904 and 2N3906 types, diodes were 1N4148. The capacitor and resistor to the right were a low pass filter (10 ohm and 0.1 uF) to stabilize the output. The unseen speaker was connected to the red and green wires.

Figure 8

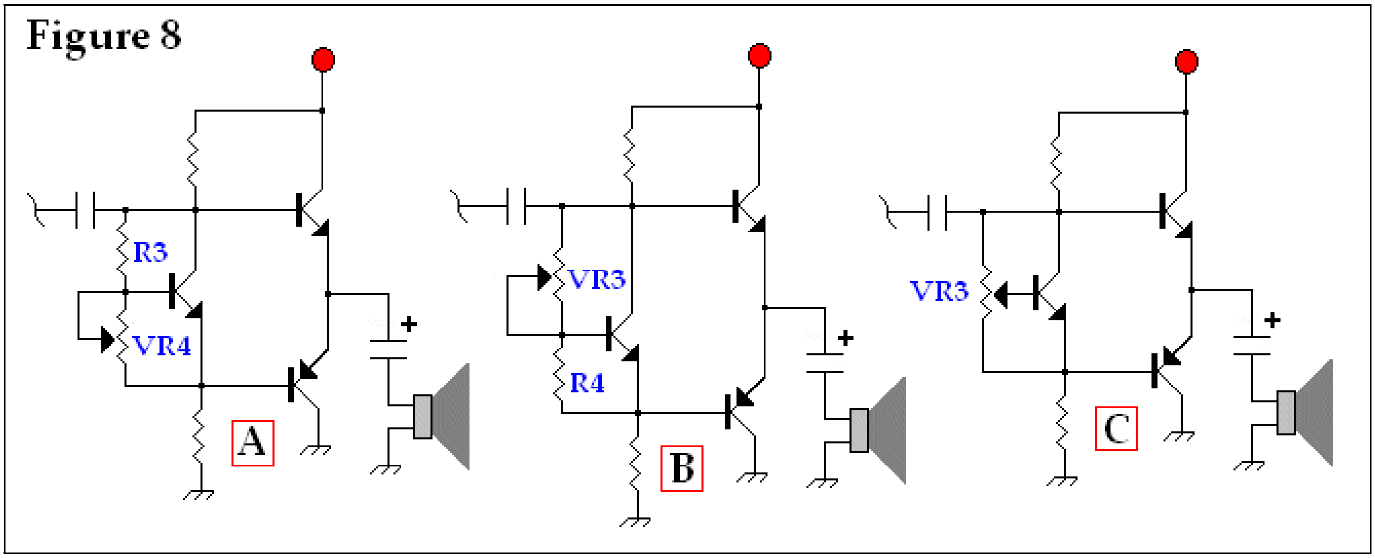

In practice, either R3 or R4 is often replaced with a trimmer potentiometer or a trimmer potentiometer is used instead of R3 and R4 and sometimes R3 and R4 are not of equal value. Shown above in Figure 8 are 3 variable bias topologies for Q3 that I have used. In some cases you will notice that the builder places a fixed value resistor or even a diode in series with the potentiometer in circuits like A or B. Using a potentiometer allows precise adjustment of the quiescent bias current and the ability to dial in the lowest crossover distortion possible. You can set the bias current using any combination of an oscilloscope and signal generator, a voltmeter, an ammeter or possibly try do it by ear when listening for and removing crossover distortion at low volume levels. The procedure I have read to adjust the bias by listening is as follows: Allow some low level signal through the amp so you can just hear it in the speaker. Turn the potentiometer from 1 extreme to another until crossover distortion is heard. Move the pot in the opposite direction until the crossover distortion disappears.
From my limited experience; in some amplifiers under 2 watts or so, you may not hear much of an audible change in crossover distortion when adjusting the bias control potentiometer, so the listening method is not useful in certain cases. It is worth mentioning, that crossover distortion sounds awful and you can usually hear it in amplifiers that are under biased.

Many builders just have a multimeter. In this case, measure the voltage drop across Q3 (the amplified diode) and ensure that is a least 1.1 volts and then slowly adjust the bias up or down from that point. Ultimately, you may have to just make the final bias setting by deciding what voltage drop across Q3 and/or what complementary pair quiescent current you want to establish. It is really not that difficult. *Whatever method you use,*

always re-check the Q3 voltage drop and amplifier bias current with no input signal to inform yourself of what is happening.
I am uncertain of the best method to measure the amplifier quiescent current, however I normally measure it using an ammeter connected in series with the emitter of the NPN transistor of the complementary pair.

Shown to the right and below are more images of what crossover distortion can look like in an under-biased power follower stage. The *right* output waveform also contains a little harmonic distortion, however that's a separate issue. I listened to this amp when connected to a speaker and music audio source; the audio had a noticeable "grungy" distorted sound. As mentioned, crossover distortion sounds terrible. The bias to the power followers (a complimentary pair of two 2N3906 and two 2N3904 transistors set up as Darlington emitter followers) was increased and the crossover distortion disappeared. A post bias adjustment audio listening test confirmed that the crossover distortion was gone.

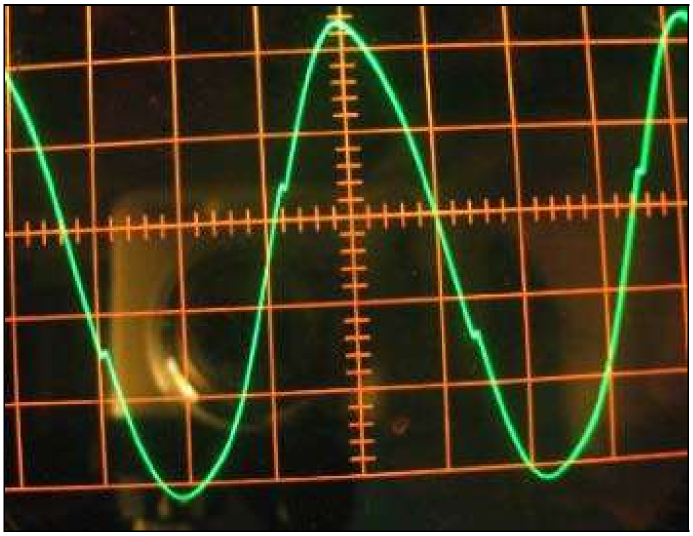

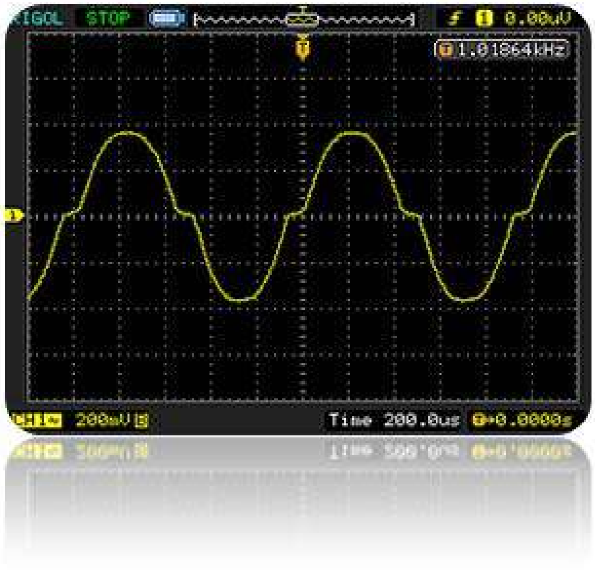

[Click](#) for a Russian language mp3 audio file. Vladimir (Volodya), a fellow builder in Ukraine, wanted an A - B comparison of Class AB versus Class B (cross-over distortion). In this audio file, I tweak a potentiometer biasing a pair of power followers to give contrast between the 2 amplifier classes. Under biased AF amps sound terrible in any language! The audio source was a cassette player. The speaker output was recorded, digitized and stuffed into the mp3 file.

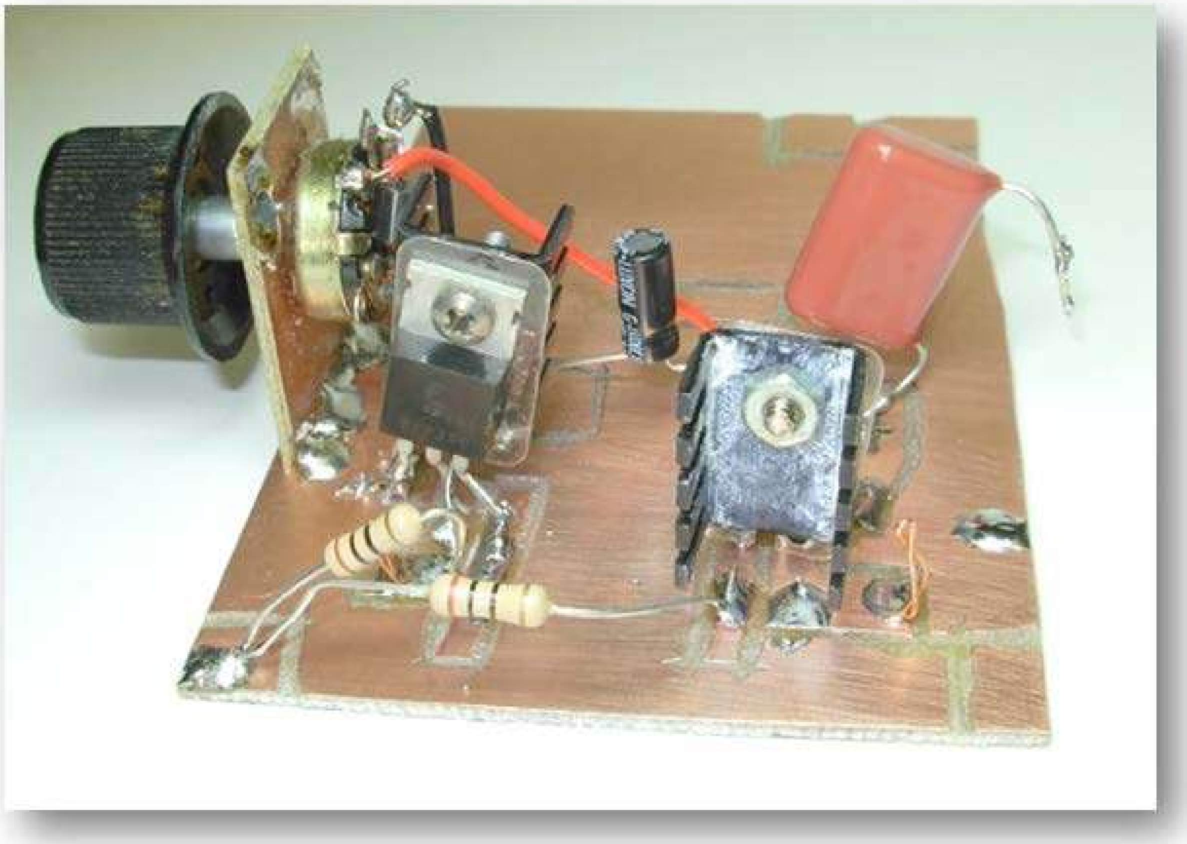

Shown above is a bread board of a complete amplifier utilizing a 10K pot to vary the bias on the amplified diode *(see Figure 8 c)*. With a 12.22 volt power supply, turning the potentiometer from one extreme to the other varied the current draw of the amplifier from about 0 to 95 mA. The average quiescent current draw of a properly biased **single complimentary pair** was somewhere between 5 and 10 mA in my bread boards.

Figure 9

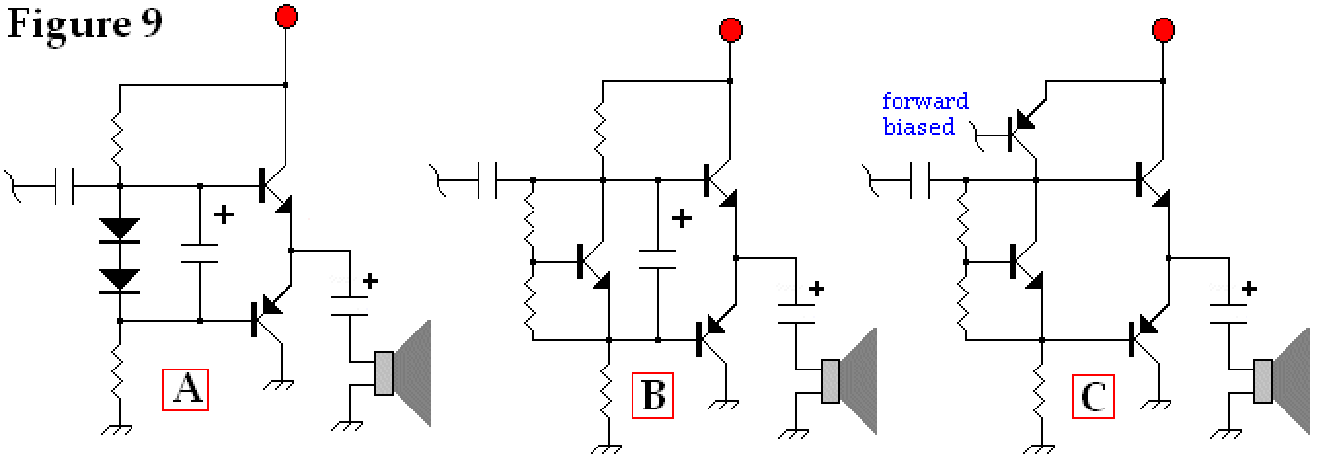

You may see a capacitor inserted between the output transistor bases as shown in Figure 9A and 9B. I have seen capacitor values from 4.7 uF to 100 uF used and the value is not critical, however, from my experiments, I have learned it is mandatory. This capacitor serves to keep the bias voltage constant as the AC signal swings up and down. Some engineers refer to the amplified diode an **NPN shifter bias amplifier** or a **level shifter**. Its function is to charge up the capacitor between the bases of the power follower NPN and PNP pair to a voltage difference that establishes the quiescent current.

In 9C, R1 has been replaced with a PNP transistor which is usually forward biased by another transistor. You may observe any number of variations of the basic biasing circuit presented in Figures 6 and 7, including 3 or more small signal diodes, 2 amplified diodes, current sources, feedback loops and more. although the techniques vary, the authors are still just biasing the complementary emitter followers to achieve low crossover distortion, stability and/or thermal tracking.

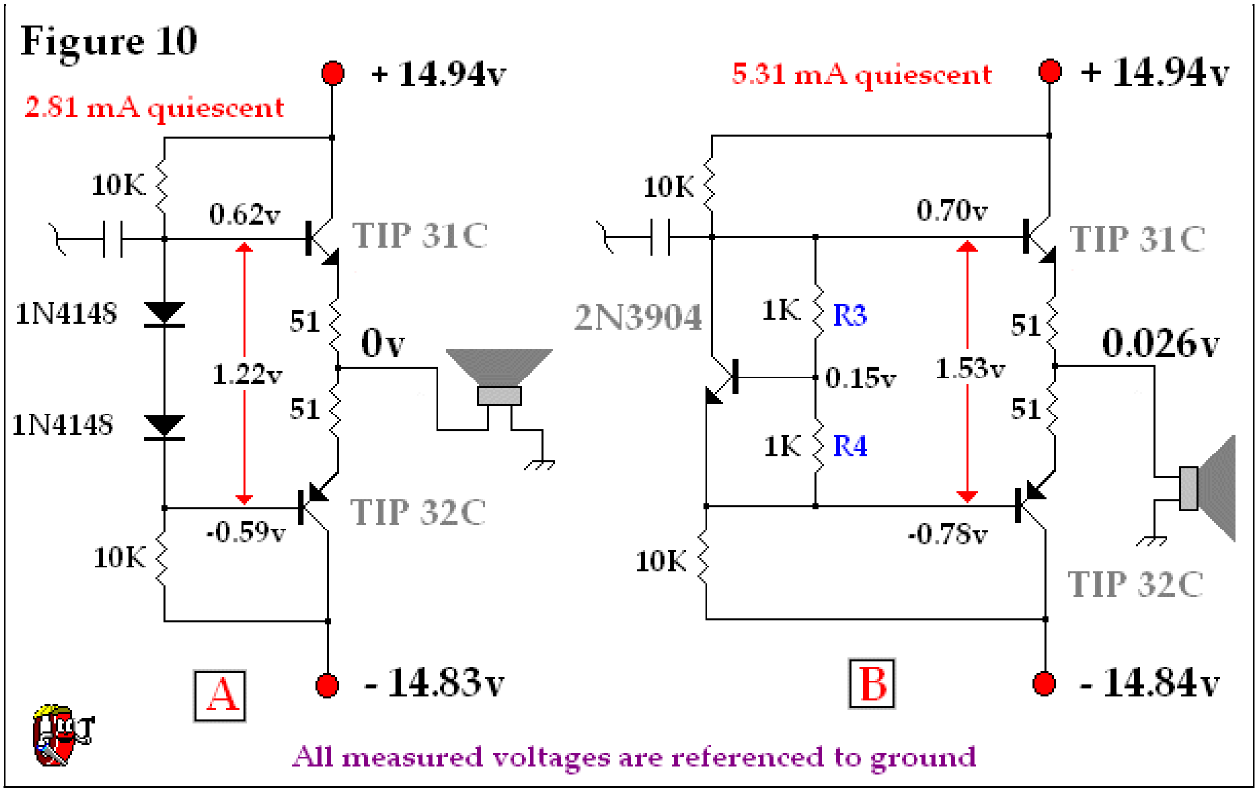
Figure 10
2.81 mA quiescent
+ 14.94v
10K
0.62v
TIP 31C
1N4148
51
0v
1.22v
51
1N4148
TIP 32C
-0.59v
10K
- 14.83v
A
5.31 mA quiescent
+ 14.94v
10K
0.70v
TIP 31C
2N3904
1K
R3
51
0.026v
0.15v
1.53v
51
1K
R4
-0.78v
10K
TIP 32C
B
- 14.84v
All measured voltages are referenced to ground

Shown above in Figure 10 are 2 amplifiers using a split power supply. The split power supply offers increased headroom due to a greater AC voltage swing as well as increases the available RMS output power without using super high AC power transformer secondary voltages. In addition, the split supply works well with op amps and if desired, enables you to reduce the number of coupling capacitors by allowing direct coupling of the preamp and speaker to the power amplifier. Coupling capacitors alter frequency response and perhaps may present phase shift issues. In some cases, we as builders use coupling capacitors to provide effects such as high pass filtering, however in Hi-Fi amps, enhanced low frequency response is usually desired; which necessitates the use of high value coupling caps in single power supply amplifiers. In split supply amps, the choice of using a coupling capacitor or not is available to you. In the Figure 10 a and b circuits above, the speaker is directly coupled to the complementary emitter followers output. Note that the voltage at this point is 0 or nearly 0 volts. For any given power supply voltage you chose (split or not), please ensure the amplifier components can handle the current and subsequent heat when a signal voltage is applied. This topic is out of scope. Build and measure...build and measure...

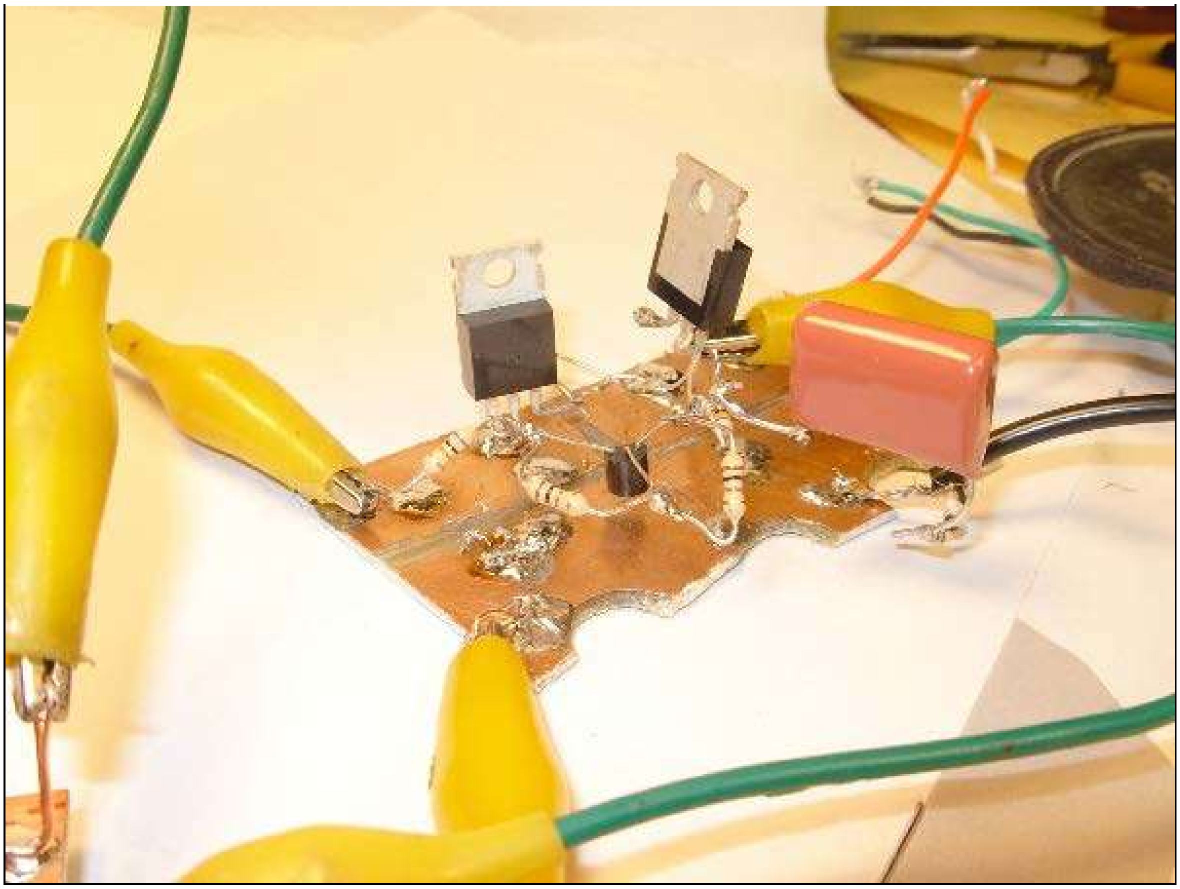

Shown above is a breadboard of the Figure 10b circuit. Additional experiments using even higher voltages were also performed, hence the moderate power TIP transistors were utilized. I burnt up four 2N3904/6 transistors performing many experiments with biasing over 3 nights. Some of the outputs of these experiments will be presented in future projects.

Figure 11

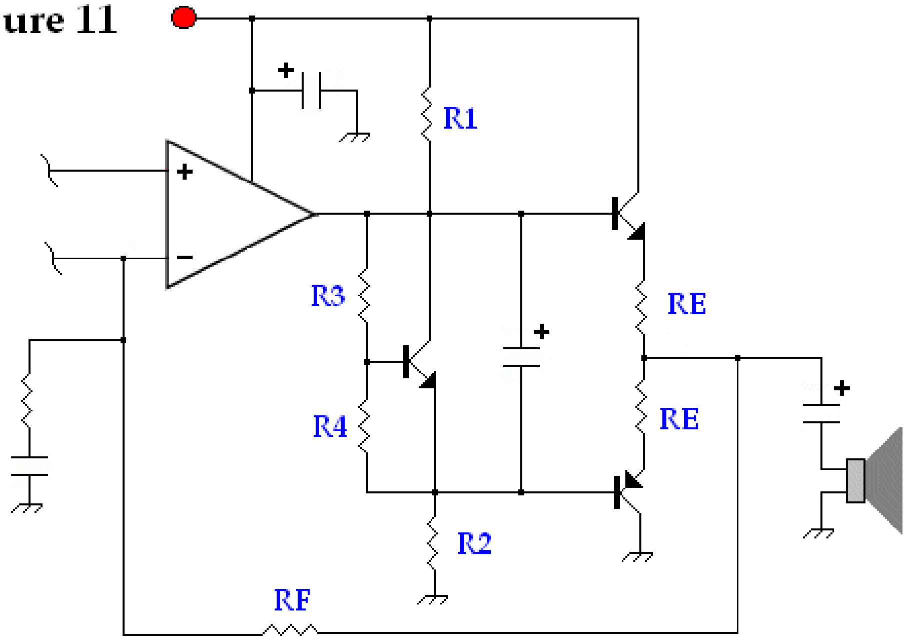

Shown above in Figure 11 is a complimentary emitter follower pair directly coupled to an op amp. The amplified diode and its biasing network is inside the op amp feedback loop. There are examples of this circuit in EMRFD and also on this web site. In single supply powered op amps, it is possible to omit R1. An example of this may be found in Figure 12.30 in EMRFD. Using a low noise op amp such as the NE5532 to drive your power followers can give outstanding results.

Figure 12

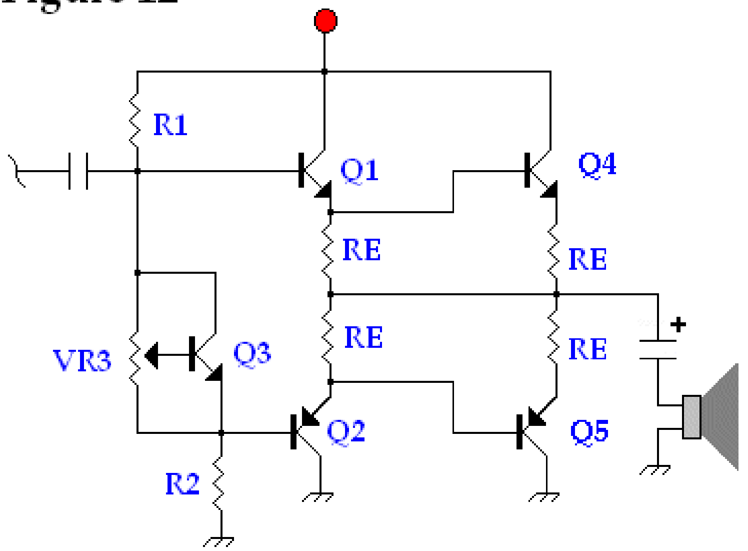

Shown above in Figure 12 is another theoretical power amp which illustrates the building block aspects of the simple stage we have been discussing. Q4 and Q5 are cascaded with Q1 and Q2 to build up the current (Darlington emitter followers). Such an amp could have several watts of output power depending on the supply voltage. The emitter resistors on Q4 and Q5 are often 0.47 to 1 ohm power resistors.

References

EMRFD Although Rick Campbell and Bob Larsen contributed chapters and circuits, the principal author is Wes Hayward. It amazes me that any human being could know so much about electronics and is so willing to share his knowledge. Respect.

Henderson, John. Electronic Devices Published in 1991 by Prentice Hall

Oleksy, Jerome E. Practical Solid-state Circuit Design Published 1974 by Howard W. Sams and Co.,Inc

Slone, Randy G. Understanding Electricity and Electronics Published 1996 by TAB Books-McGraw Hill

QRP SWL HOMEBUILDER

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Low Power Audio Amp Experiments

Introduction

This web page contains some experiments on simple, low power, speaker output audio amplifiers. Presented are ideas, some measurements and examples of audio amplifiers which will likely sound better than the IC audio chips commonly seen in many receiver projects. This web page is a follow-on to this one and is a completely new area of experimentation for me. Audio amps were built using both split and positive power supplies. In all cases the complimentary power followers were driven by an op-amp. I tried building some power amps using discrete transistor differential amplifier stages with current sources as the driver, but the noise performance and simplicity of the NE5532 or NE5534 op-amp was superior.

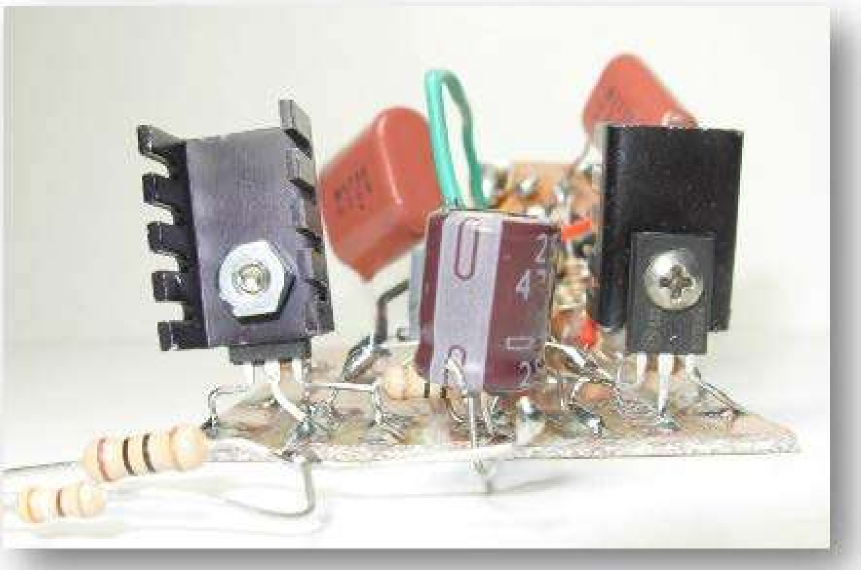

Split or Bi-polar Power Supply Audio Amplifiers

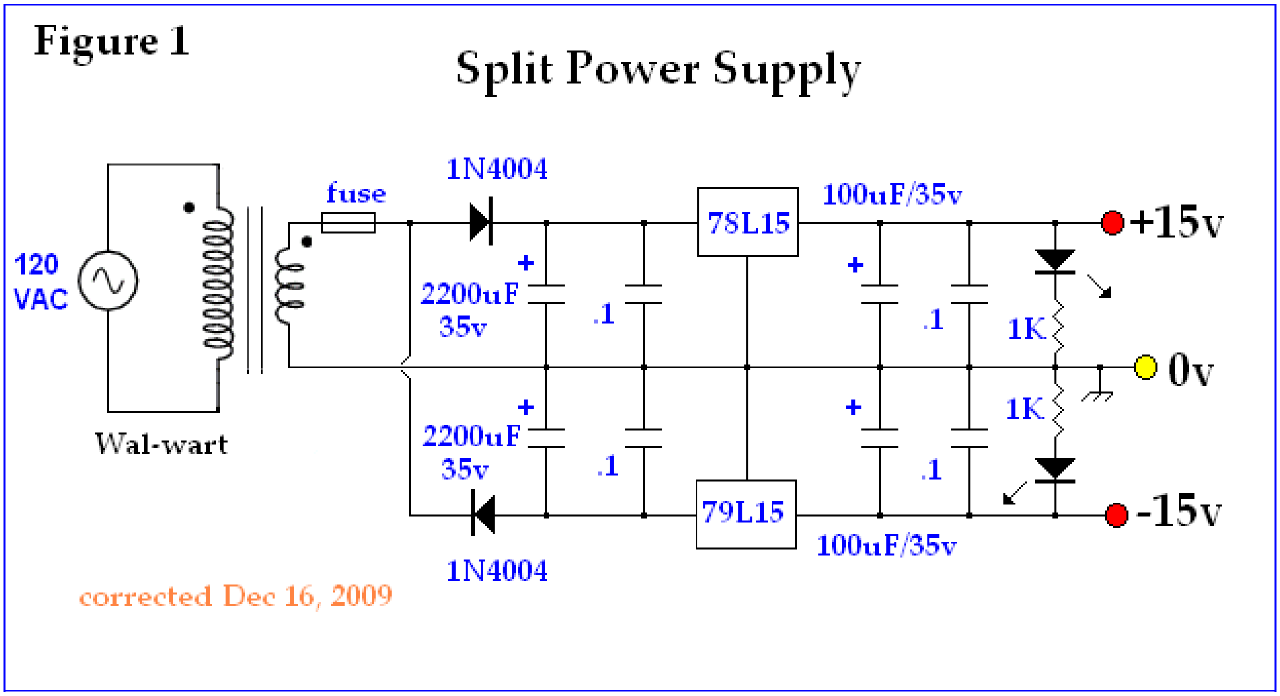

Figure 1

Split Power Supply

corrected Dec 16, 2009

In order to build up some split power supply amplifiers, a basic power supply was constructed and the schematic is shown in Figure 1 above. I found it was essential to regulate the voltage or hum would appear on the output. Choose a standard value fuse that is rated somewhere just above the maximum current you measure. I used 2 different AC output power transformers which were in the 18-24 volt, 375 mA to 1 amp range. The LEDS are strongly suggested. They inform you when there is power applied and their relative brightness will also often fall when higher current is being drawn on one side or the other. This alerted me to an accidental solder bridge to ground on more than 1 occasion.

The split power supply is shown in the photograph to the left. The retro Bakelite fuse holder is from an old tube audio amplifier. If you are wondering why the copper clad board is so large, the power supply is part of a future project. Some builders would use even greater value filter capacitors than those shown. Heat sinks on the voltage regulators are required for supplying DC to higher power amplifiers.

Rod Elliot Headphone Amp

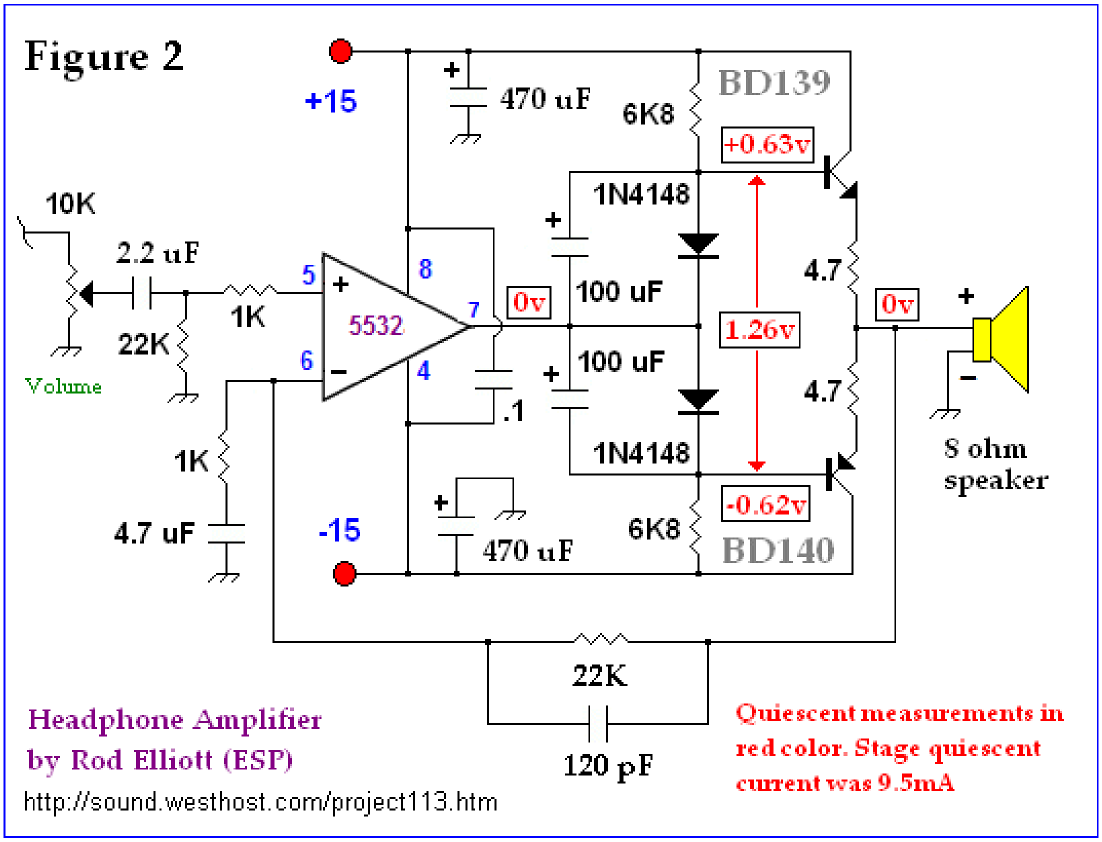
Figure 2
+15
470 uF
6K8
BD139
+0.63v
10K
1N4148
2.2 uF
5
8
1K
7
0v
100 uF
4.7
0v
22K
5532
1.26v
6
4
Volume
100 uF
4.7
.1
1K
1N4148
8 ohm
speaker
-0.62v
4.7 uF
-15
470 uF
6K8
BD140
22K
Headphone Amplifier
by Rod Elliott (ESP)
http://sound.westhost.com/project113.htm
120 pF
Quiescent measurements in
red color. Stage quiescent
current was 9.5mA

The first amp I built is shown in the Figure 2 schematic above. This amplifier was designed by Rod Elliot and is used with his permission. Rod's ESP web site is a virtual treasury of audio design information. If you are into understanding audio design, visiting his web site is strongly recommended. Rod sells printed circuit boards for all of his circuits if you prefer this building method . Note I have made some minor modifications to some part values. although primarily designed as a low distortion, high power, headphone amplifier, it drives an 8 ohm speaker very well. I was able to drive this amplifier as high as 0.68 watts average power with a pure sine wave output during analysis. Power measurement is discussed in the next section. Note that on this web page, I quote the entire stage quiescent current. Since the op-amp and the

2 (or more) power followers are a "package", it is a lot easier to just measure the current at the power supply lead(s) of the stage than unsolder and lift up a transistor lead. In this case, with no input single, the stage current was about 9.5 mA. I did check and about half the current is going to the op-amp with the other half to the transistor pair. This is a wonderful sounding amplifier and Rod has an entire web page devoted to it, so I will not comment further.

There are a variety of suitable transistor pairs for audio power amplifiers depending on the power output you are choosing. I stock just a few; BD139-140, TIP 41C-42C, NTE 128-129. The higher beta 2N3904-3906 or 2N4401-4403 pairs worked well in the low power, single power supply amplifiers shown on this web page. I also performed some higher output power experiments which required the TIP and BD transistors and these are not shown.

225K
H 250
475K
H 250

Above photograph. A breadboard of the Rod Elliot headphone amplifier. This early version had a temporary output capacitor. When first testing a new circuit that has a direct speaker output, it might be a good idea to temporarily use an output capacitor until you measure your voltages and current and feel your transistor temperatures. This will save your speaker if you made a big mistake and/or blow up the transistors when you first power it up.

Amplified Diode Biased Audio Amp

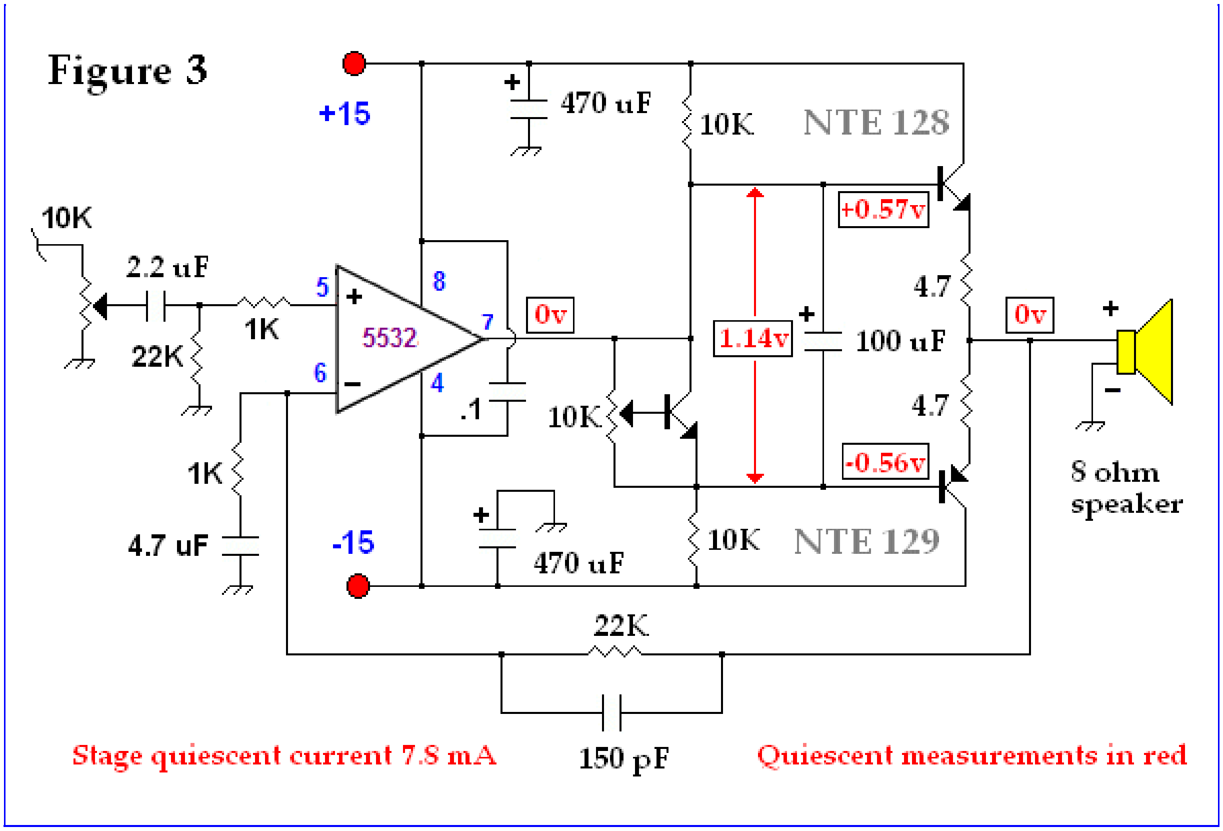
Figure 3
+15
470 uF
10K
NTE 128
10K
2.2 uF
+0.57v
5
8
4.7
1K
+
5532
7
0v
1.14v
100 uF
0v
+
22K
6
−
4
.1
10K
4.7
−
-0.56v
1K
8 ohm
speaker
4.7 uF
-15
470 uF
10K
NTE 129
22K
Stage quiescent current 7.8 mA
150 pF
Quiescent measurements in red

Above in Figure 3 is a split power supply audio amp using an "amplified diode" to control the bias. The bias transistor was wedged between one of the output transistors and a piece of copper clad board to allow thermal tracking. The 10K bias control resistor was a trimmer type suspended over the copper clad board in most of my bread boards. Usually, you just need to set and forget about this resistor after initial set up. I adjusted the bias by watching in my oscilloscope with a low level, 1 KHz sine wave connected to the input. I measured the various voltages and stage current at quiescent and have indicated these values in red for learning purposes. The bias current range was 7.2 to 154 mA when turning the 10K trimmer pot from 1 extreme to the other. The maximal clean output average power of this amp was 0.78 watts. I used press on heat sinks for the NTE128-129 pair and they ran quite warm to touch. These TO39 type packaged transistors are somewhat difficult to heat sink compared to the TIP/BD transistor packages where you can just bolt on a heat sink of any size that is required. Please remember that the metal tab on the TIP and BD transistors is connected to the collector terminal.

In the above photograph is my first bread board of the Figure 3 amplifier. This particular version had TIP transistors, a 4 ohm speaker and an 470 uF output capacitor. Note the full size 10K bias control potentiometer on the left hand side. This was purely a experimenter's bread board, but it sounded amazing when listening to music through it.

Harmonic Distortion and Measuring Output Power

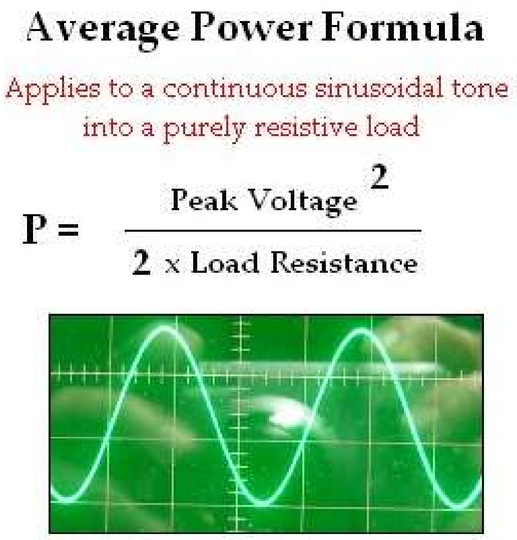

Average Power Formula

Applies to a continuous sinusoidal tone into a purely resistive load

$$P = \frac{\text{Peak Voltage}^2}{2 \times \text{Load Resistance}}$$

Figure 4

In Figure 4 is the formula used to calculate the average power of the circuits on this web page. For example if you measure 6 volts peak to peak on the oscilloscope, (3 volts peak voltage) and your resistive load is 8 ohms, the average power is 560 milliwatts. At any point in an AC waveform there is power and it may be reported using a variety of ways. Was it clean? distorted? a peak value? an RMS value? - often it is unclear.

To be clear, I measured the peak voltage on a pure, undistorted sine wave into an 8 ohm resistor. Stated power values are the mean sine wave power calculated with the formula shown. See this somewhat controversial link for details. You may not agree with my methodology, however, it allows you to compare the circuits on this web page. If you really must know the peak power, multiply the stated average sine wave power by 2. I will leave the power measurement and calculation debate up to scholars; as a lay-person, I need something simple.

The bench voltage measurement was as follows: The amplifier was connected to a 1 KHz pure sine wave generator and the 10K volume control pot was advanced just until any sign of distortion of the amplifier output sine wave appeared. Voltage measurement was taken at the point just before distortion occurred.

It is difficult to photograph a sine wave without a tripod. Motion, the angle, light reflection and jpeg graphic compression all wreck the perfect sine wave. In the Figure 4 graphic above is a typical 1 KHz output waveform from my power amplifiers (squeaky clean) at the amplifiers maximum average power level.

To the right is a photograph of my AF signal generator. This is an old, tube device but the output sine wave is beautiful. I did not perform spectrum analysis with a computer audio sound card program and will leave this up to audiophiles. These audio amps sound great; especially when compared to the IC audio power amps that many of us tend to use in our receivers.

Heathkit
AUDIO GENERATOR
POWER
FREQUENCY
OUTPUT

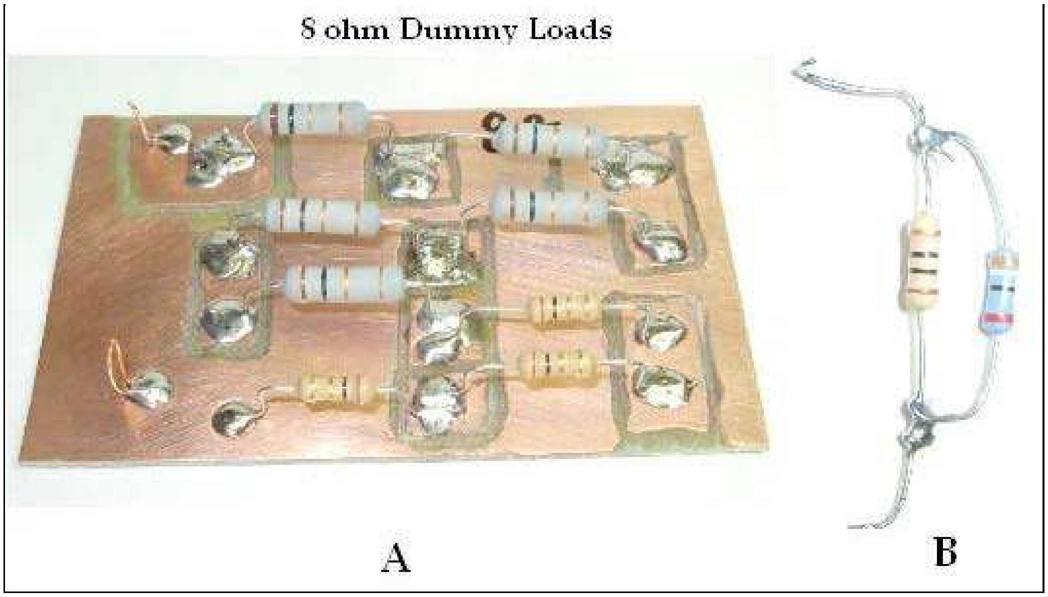

In the above photograph is **A**, an 8 x 1 ohm resistor load and **B**, an 8 ohm load made from parallel 1/2 watt 10 and 39 ohm resistors. In dummy load **A**, I used 5 two watt metal film resistors plus 3 half watt resistors. In the future, I will obtain 3 more 2 watt resistors and replace the 1/2 watt resistors for a 16 watt rating. For a quick resistive load, **B** is the way to go for most of the circuits on this web page.. You can make a 4 ohm load from parallel 4.7 and 27 ohm resistors. In truth, a single resistor or any combination of resistors adding up to the desired load R value will work.

Figure 5

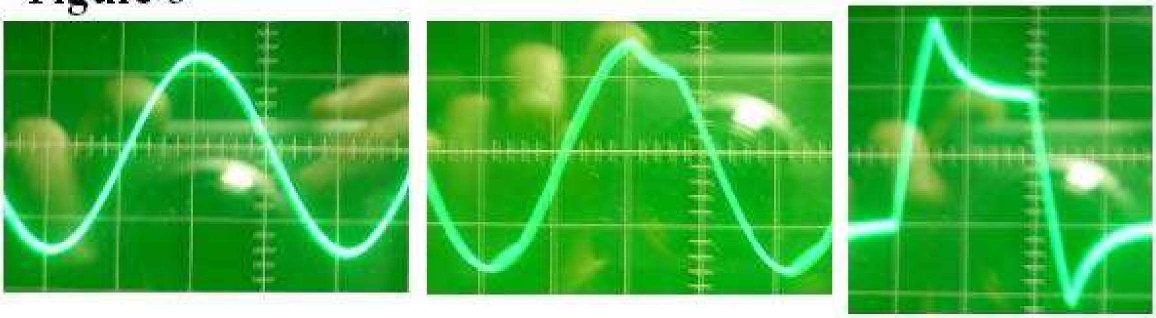

In Figure 5 above are some scope waveforms ranging from mildly distorted to full-on dirty.

Power Amplifier Concerns

Although the amplifiers on this page are 0.15-0.8 watts or so, they can consume relatively large current compared to the usual voltage amplifier circuits we build. Some potentially helpful tips to help keep away ground loops, oscillations and thermal run away are suggested as follows:

✓ Connect your negative speaker terminal directly to the AF power amp (do not use a common ground for the negative speaker terminal).
✓ Use big power supply line bypass capacitors (no 10 uF caps here)
✓ Keep your audio amplifier copper clad board separate from the rest of your circuit boards and star ground it to your main power supply ground point.
✓ Use heat sinks on your final transistors and voltage regulator(s) when you go for bigger power
✓ Watch your layout - keep the output away from the input etc.
✓ Watch your emitter resistor power ratings in "higher wattage" amplifiers. Burning resistors stink.

Single Power Supply Audio Power Amps

Since most 12 volt power supplies are actually closer to 14 volts; these experiments were performed with a typical radio bench DC power supply at 13.69 volts.. Figure 6, 7a and 7b represent evolving experiments aimed at obtaining greater output power.

Figure 6

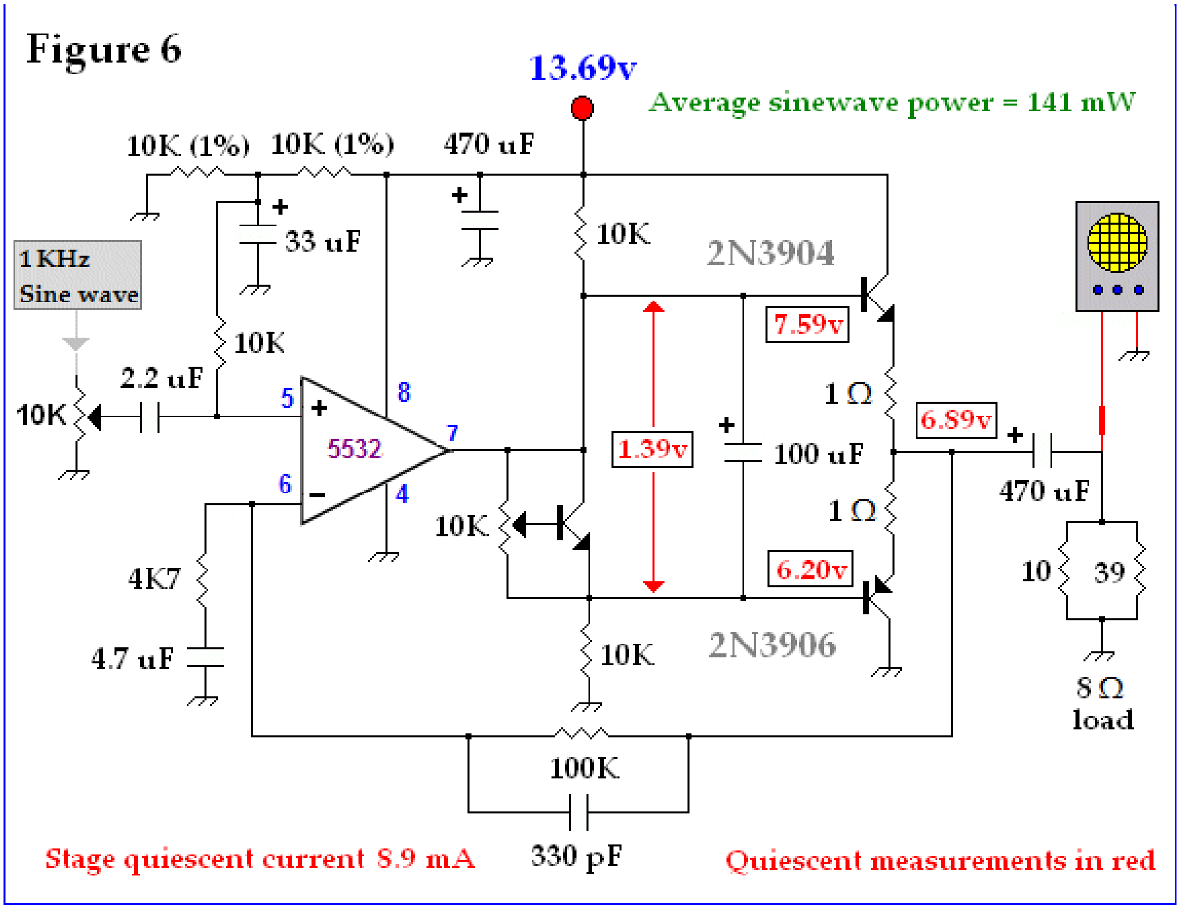
13.69v
Average sinewave power = 141 mW
10K (1%)
10K (1%)
470 uF
+
33 uF
10K
2N3904
1 KHz
Sine wave
10K
2.2 uF
10K
5
+
8
5532
7
6
–
4
7.59v
1 Ω
6.89v
+
1.39v
+
100 uF
470 uF
1 Ω
10K
4K7
6.20v
10
39
4.7 uF
10K
2N3906
8 Ω
load
100K
330 pF
Stage quiescent current 8.9 mA
Quiescent measurements in red

Shown above in Figure 6 is the fundamental design using one op-amp and 2 power followers. It is shown in an AC output power measurement configuration. The bias current range was about 5 to 100 mA when turning the 10K trimmer bias control pot from one extreme to the other. Maximal sine wave average output power was only 141 mW. Nevertheless, it might be loud enough for some receiver applications.
I connected this amplifier to a VCC of 15 volts. The maximal sine wave average output power was then 220 mW. In all of the single supply audio amps presented , increasing the VCC increased the maximal power output. Driving these amplifiers beyond a pure sine wave output power resulted in predictable harmonic distortion plus the re-emergence of crossover distortion in the output. This was an incredible learning; how could there be crossover distortion re-emerging in a amp that was properly biased to begin with? Increasing the bias current to the maximum level did not remove this crossover distortion. After emailing this question to Rick, KK7B and Wes, W7ZOI, and reading their replies, my best guess was that at some power level, the 5532 op-amp can not provide enough current to properly drive the complementary symmetry pair. The AC current in the output transistors may be limited by the base drive of the op-amp and they were no longer forward biased at the crossover point.

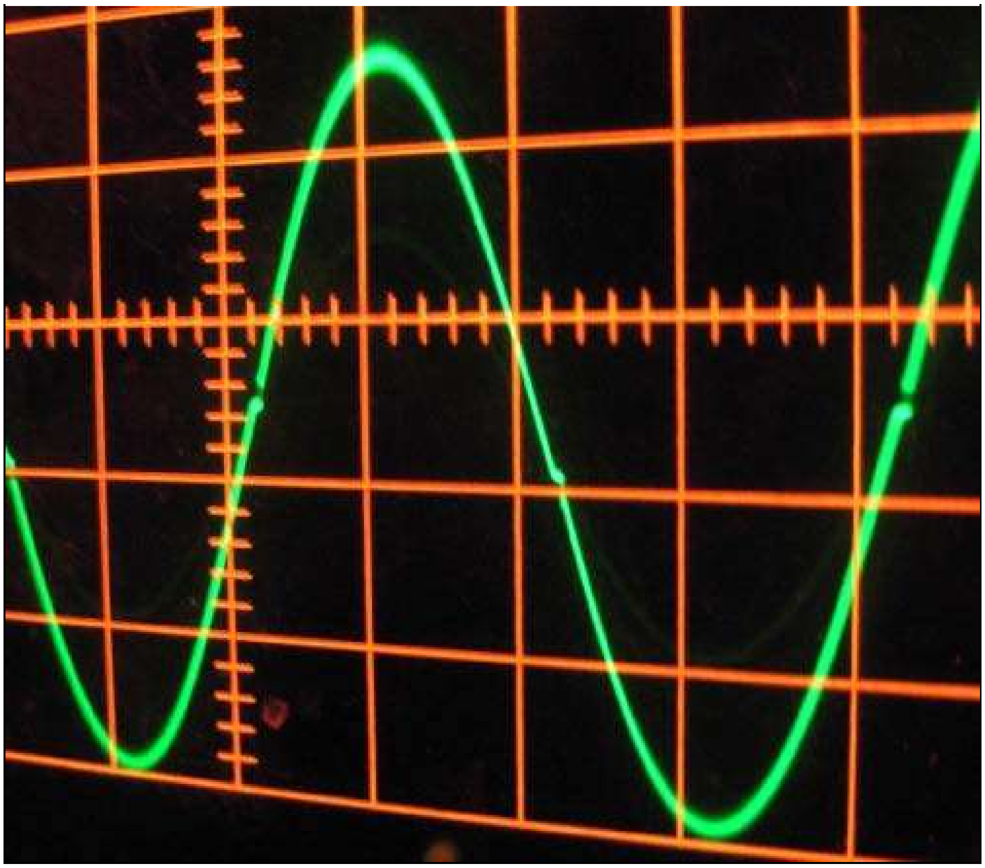

Above photograph. This is the Figure 6 amp driven past the point where the sine wave is pure. Note the crossover distortion blips on the sine wave. The base drive current for the power follower pair all comes from the op-amp. At this point there is likely not enough base drive to keep the base emitter junctions forward biased.

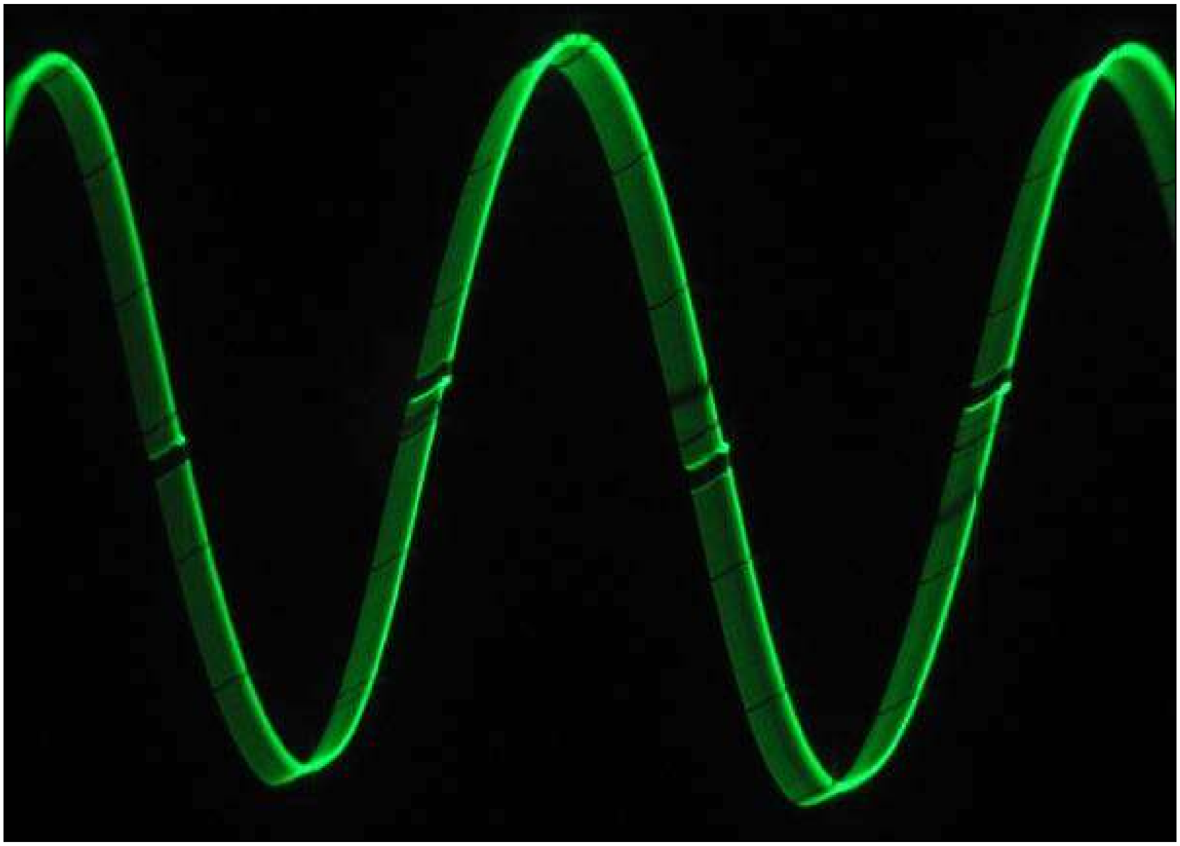

In the above photograph I blacked out the room and photographed the same scope waveform as above while shaking the camera from side to side. This adds some horizontal spreading of the signal and provided more information about what was happening as compared to a single,

clear oscilloscope trace.
I should mention that this crossover distortion blip occurred in all of the Figure 2 to 6 amps when they were driven past the point where a pure sine wave was seen. It is clear that maximal available power from a simple audio amplifier like this **(one NE5532 op-amp plus 2 power followers)** is constrained and thus its application is limited. Greater output power is possible using a split supply per Figures 2 and 3, however, a typical radio project has a 12 volt, single power supply. These basic amplifiers with a single power supply, may be very appropriate for projects such as a compact radio receivers or a code practice oscillator project, but not for applications where you require louder audio.

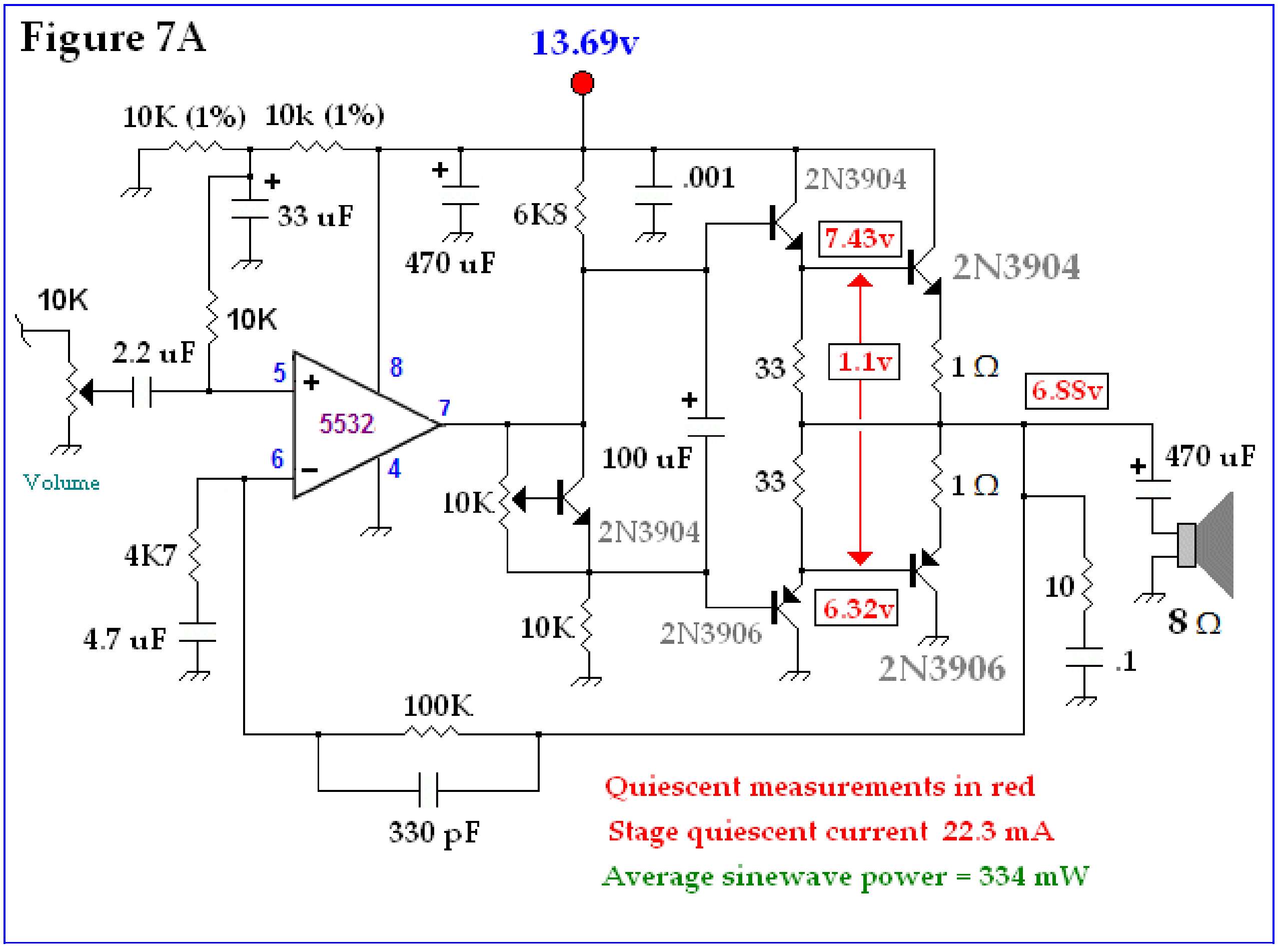

Quiescent measurements in red
Stage quiescent current 22.3 mA
Average sinewave power = 334 mW

Shown above in Figure 7a is an easy method to get more output power from the Figure 6 amplifier; add another set of complimentary pair current amplifiers. I found 33-39 ohms to be a good emitter resistor value during my experiments. Many hi-fi amp builders will use greater emitter resistor values, however, a design goal was to get more output power from our 12 volt supply. Series emitter resistors are used to improve linearity and operating-point stability. I kept the final power follower pair emitter resistor values at 1 ohm to get maximal output power. An output 10 ohm + 0.1 uF low pass filter was used to help prevent oscillations in view of the low emitter resistor values on the finals.

Biasing

The top 10K bias resistor was lowered to 6K8 to facilitate "more linear" setting of the output transistor bias with the 10K trimmer potentiometer. It did not help much. Setting the bias is very delicate procedure and you must turn the screw driver very slowly. In my bread board, the optimal stage bias current was 22.3 mA but anything around 20 mA should be fine. If you do not have an oscilloscope, after ensuring that there is no input signal, connect an ammeter in series with the positive power supply lead. Turn the bias potentiometer with with the screwdriver until you get close to 22 mA. If you only have a voltmeter, the rule of thumb of 1.1 seems to work... Measure the voltage across the final 2N3904-2N3906 bases and ensure the difference is at least 1.1 volts while adjusting the 10K trimmer pot. Personally I do my biasing with an oscilloscope at at least 2 different frequencies on the signal generator, however the for mentioned methods will work okay. This is a popcorn stage and a popcorn web site after all!

For the lowest potential noise, consider using metal film type resistors in your audio amps and "polysomething" capacitors wherever AC signal is coupled to another component or ground, excluding the output capacitor.

Figure 7b

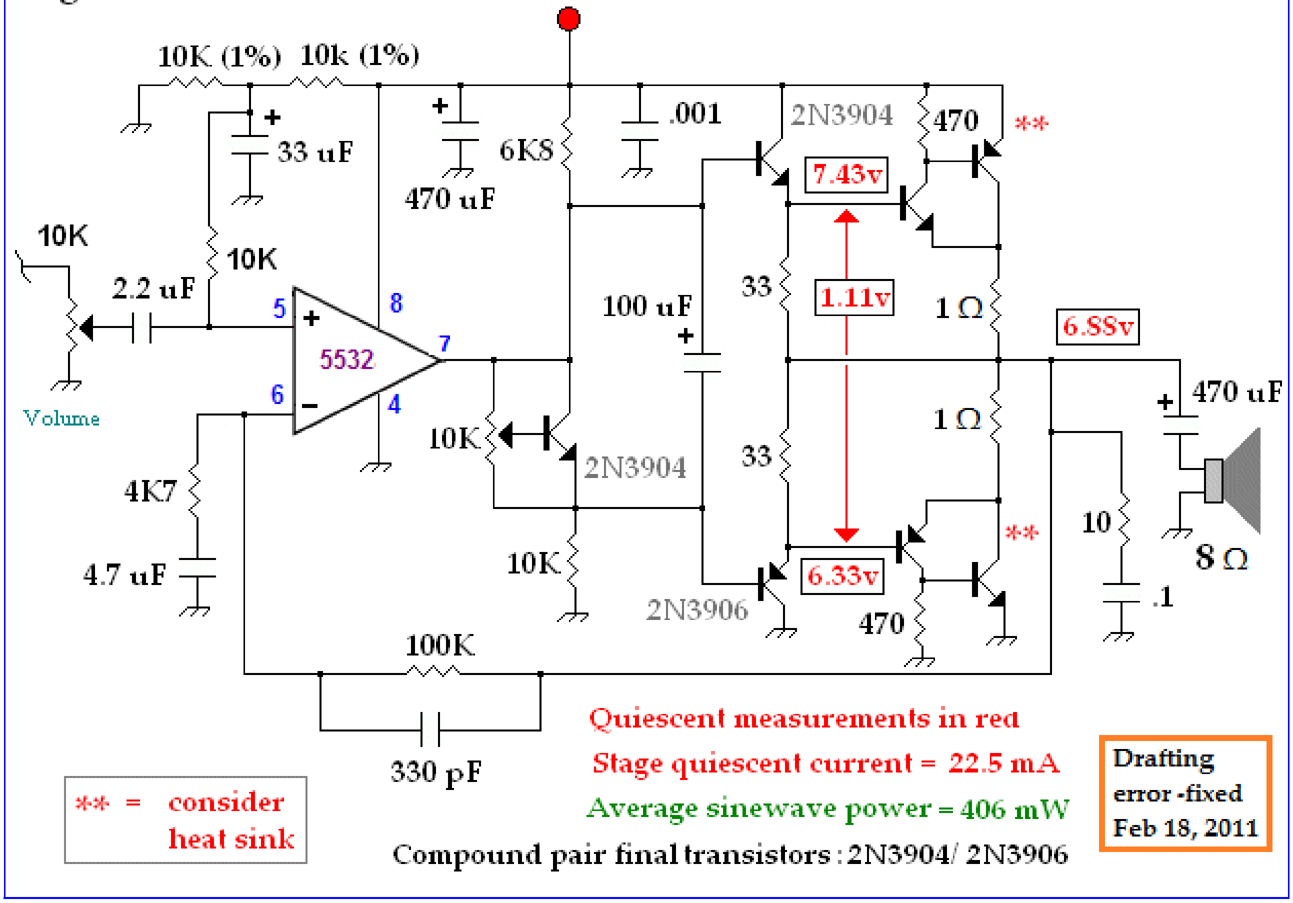
13.69v
10K (1%)
10k (1%)
+
33 uF
+
470 uF
6K8
.001
2N3904
470
**
7.43v
10K
10K
2.2 uF
5
+
8
5532
7
6
-
4
1.11v
100 uF
+
33
1 Ω
6.88v
Volume
10K
2N3904
33
1 Ω
+
470 uF
4K7
10
8 Ω
**
4.7 uF
10K
6.33v
2N3906
470
.1
100K
330 pF
** = consider heat sink
Drafting error -fixed Feb 18, 2011

Quiescent measurements in red

Stage quiescent current = 22.5 mA

Average sinewave power = 406 mW

Compound pair final transistors : 2N3904/ 2N3906

Further experiments to increase output power were frustrating. Finally a compound or Sziklai pair was trialed and increased the average power to over 400 mW as shown in Figure 7b . I used a small piece of copper clad board on the finals for a heat sink, although they really didn't get that warm. Ideally, the amplified diode should also be glued onto one of the heat sink boards for thermal tracking. This amplifier is now in a chassis as a bench reference audio amp for receiver testing.

In the above photograph is the Figure 7b prototype. I am using a new miniature potentiometer for my experiments that I bought from Digi-Key. The base has 2 leads which can be soldered right on the copper clad board for easy anchoring and removal after testing.

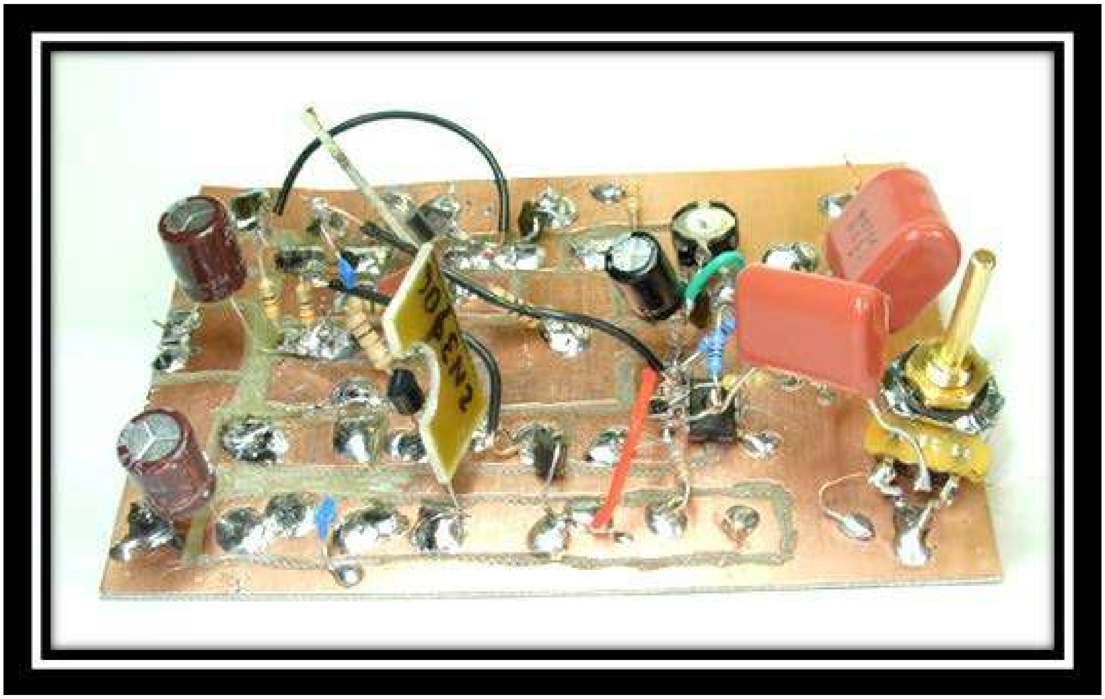

Another view of the bread board on which the Figure 6, 7a and 7b experiments were conducted.

KK7B Headroom Boosting Emitter Capacitors

We first learned about using large value emitter caps in audio amp complimentary pairs from EMRFD. Experimentation revealed that these capacitors do 2 things: **1.** Can increase the amplifier sine wave headroom and **2.**

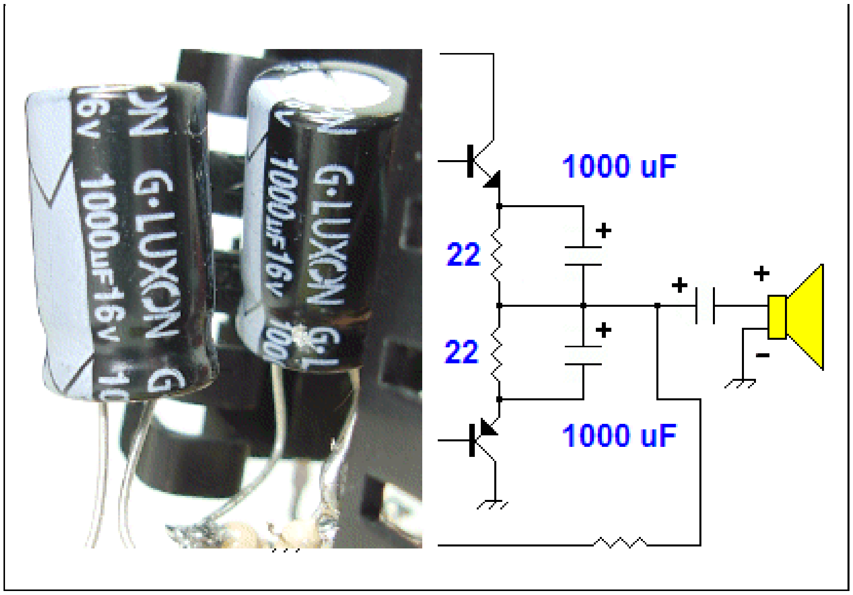

Add some low pass filtering. I learned from Rick, KK7B, that he designed his EMRFD amp to achieve low output power, low distortion and lower DC current drain. He desired a clean output audio amplifier for his R2 series of receivers without needing a lot of quiescent current or heat sinks on the 2N3904-2N3906 pair. The caps were added to make the amplifier think it had much lower emitter resistors at AC than the 22 ohm resistors he used in the EMRFD projects. When Rick made measurements and simulations of the amplifier, it was very stable, had low distortion and provided a very nice clean sound at all signal levels, from very weak signals in a few milliwatts of noise, to music driving the speaker. The result is outstanding and Rick's design was the catalyst for my own interest in audio amplifier experimentation.

I performed experiments with these capacitors and found that they increased my amplifier power and head room in some cases, and that the boost is indirectly proportional to the emitter resistor value. With 1 ohm output transistor emitter resistors the boost is generally not that significant. With 4.7 ohm or greater emitter resistors, they can make a big difference and you might consider trialing them for more power and headroom as appropriate. They can also add a nice, warm sound to your audio amp. Refer to EMRFD for numerous examples of this technique.

Popcorn Audio Amplifier

What follows is a popcorn or "poor man's" audio power amp using the 2N3904-2N3906 pair. To meet true popcorn criteria, all of the capacitors used in my breadboard were electrolytic and you can substitute different values from your own junk box. It would be better to use "polysomething" capacitors for the NE5532 pin 5 and 6 signal capacitors if you have them. I normally use a 1 uF to 4.7 uF poly-type capacitor in series with the 4K7 resistor on pin 6. The 270 pF feedback capacitor could be omitted or substituted with a higher or lower value to suit whatever high frequency roll off you desire.

The transistor glue-on heat sinks seen in the 1 bread board photograph are completely unnecessary. This BJT pair were used in other higher power experiments as well. The 22 uF capacitor between the transistor bases is essential from my experimentation. Without this capacitor, the amplifier headroom decreases and crossover distortion occurs. You can use the other half of the 5532 for a preamplifier or use a NE5534 instead. A 741 op-amp would be a horrible substitution. The NE5532 performance is breathtaking considering its low cost. 181 mW is surprisingly loud. All resistors are quarter watt rated. What a fun little amplifier! The schematic is Figure 8.

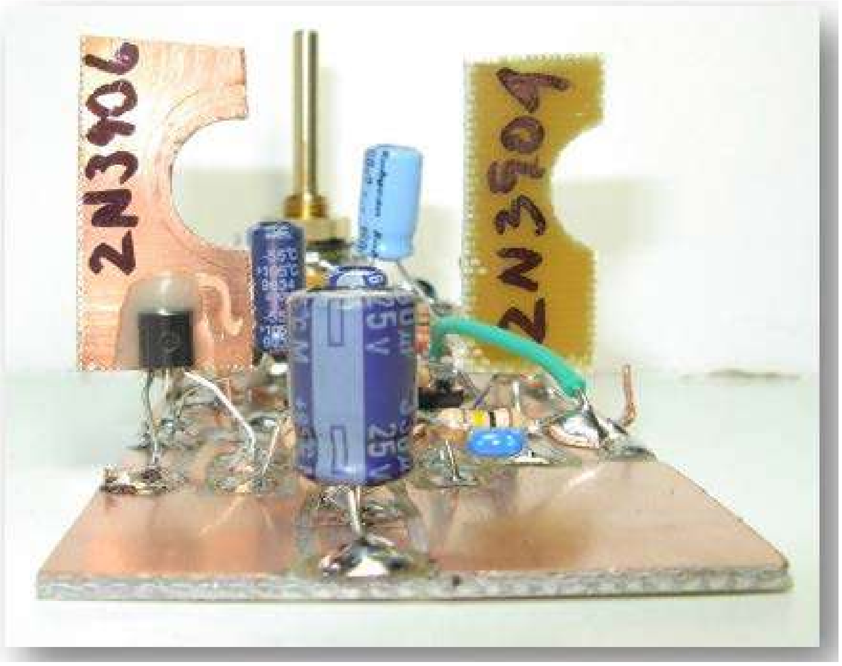

Figure 8

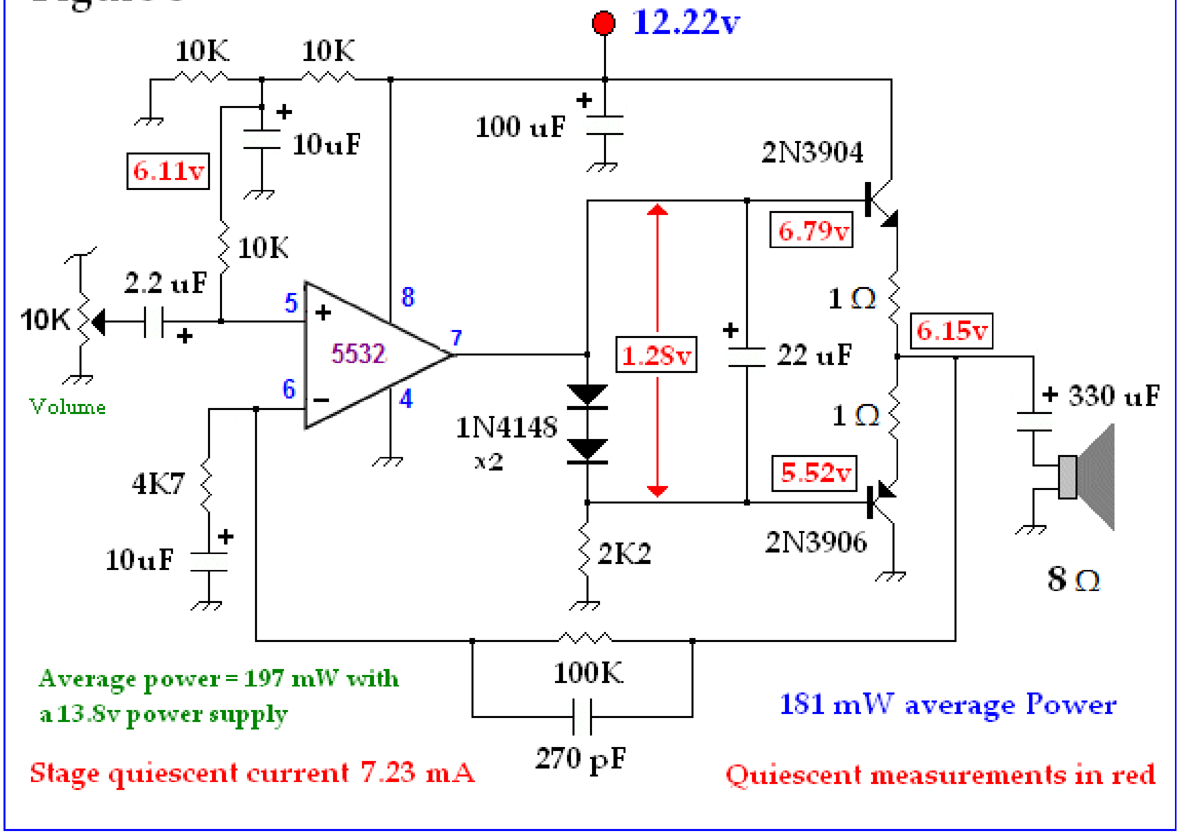
12.22v
10K
10K
+
10uF
6.11v
100 uF
+
2N3904
6.79v
10K
2.2 uF
10K
+
5
+
8
1 Ω
6.15v
5532
7
1.28v
+
22 uF
6
−
4
1 Ω
+ 330 uF
Volume
1N4148
x2
5.52v
4K7
+
10uF
2K2
2N3906
8 Ω
100K
270 pF

Average power = 197 mW with a 13.8v power supply

181 mW average Power

Stage quiescent current 7.23 mA

Quiescent measurements in red

missing
10K pot

The bread board of the Figure 8 popcorn amplifier using transistors without heat sinks.

Additional Outputs

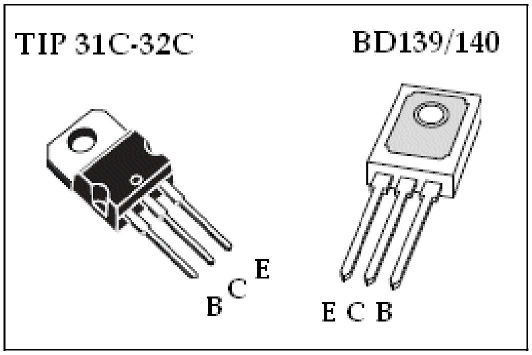
TIP 31C-32C
BD139/140
B C E
E C B

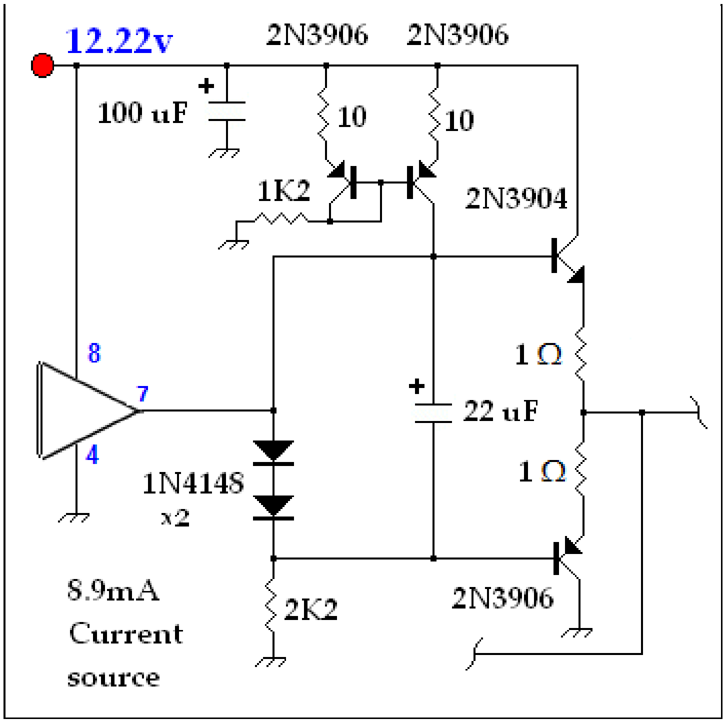
12.22v
2N3906
2N3906
100 uF
10
10
1K2
2N3904
8
7
4
1 Ω
22 uF
1 Ω
1N4148
x2
8.9mA
Current
source
2K2
2N3906

I tried putting a current source on the Figure 8 bias and it made no difference to the amplifier characteristics according to my simple oscilloscope, listening and DC analysis.

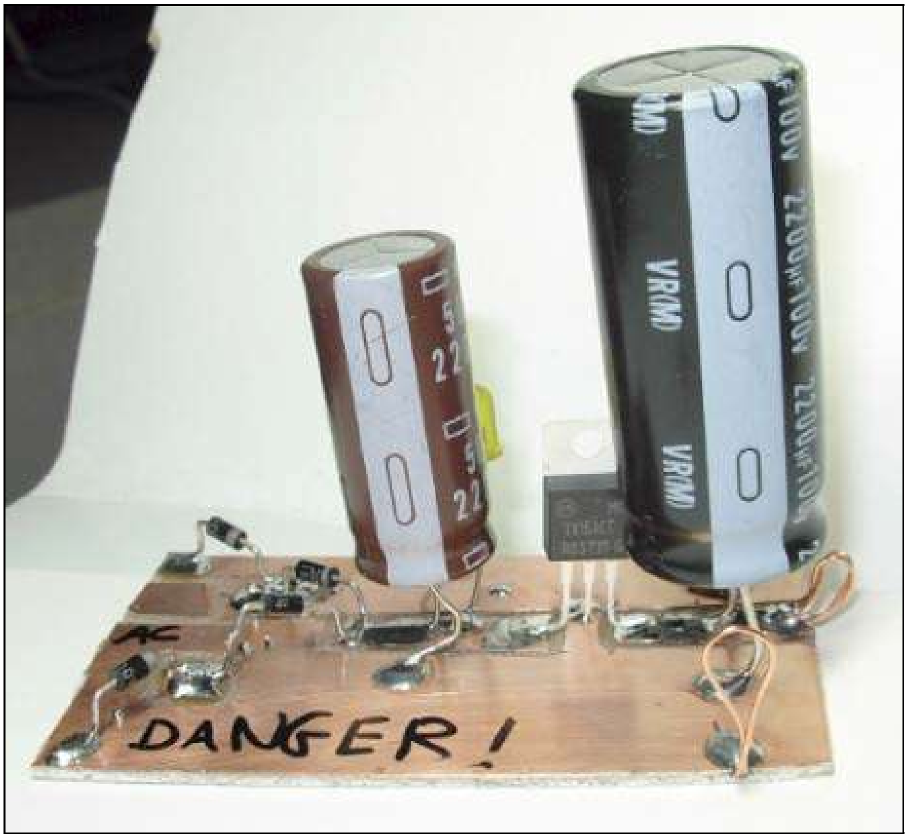
DANGER !
AC
VRM
2200uF100V

One of the full wave rectifier, voltage regulator and filter bread boards used in these experiments. I went as high as 24 VCC on some single supply amps I tested and was getting over 5 watts average output power

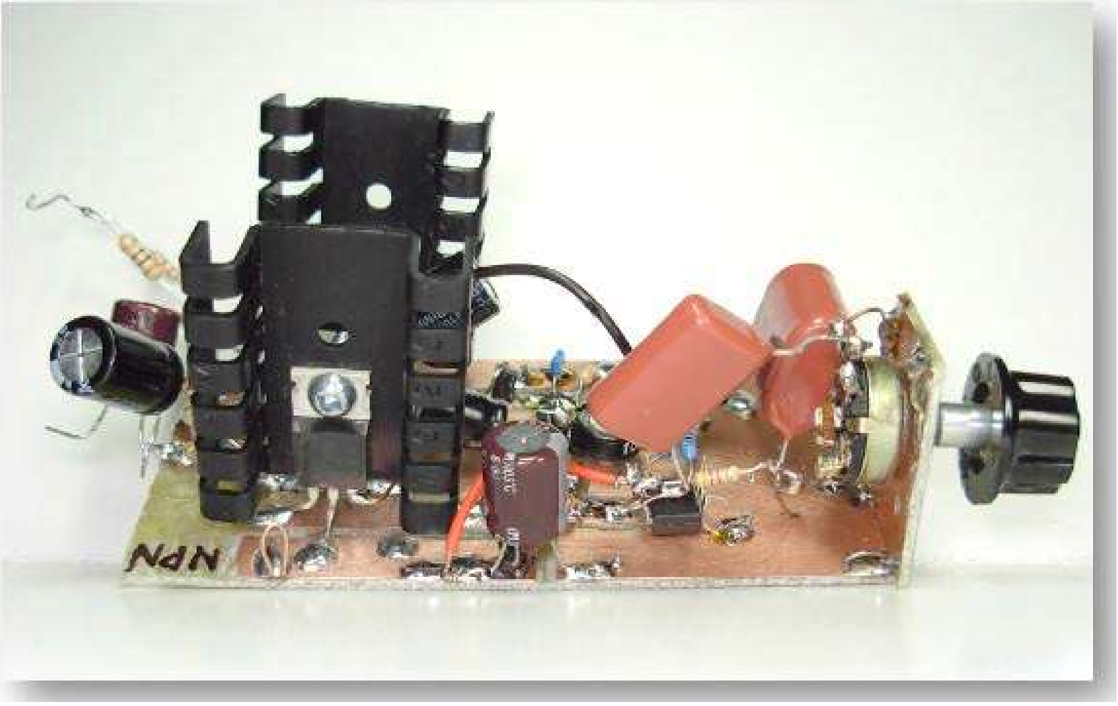
NPN

Two Bravo Receiver Experiments

Introduction

An experimental direct conversion receiver is presented. This 1990's style receiver was built to re-familiarize with DC receivers and try out a few new ideas. Design-on-the-bench bread boarding was used exclusively and was a pleasant way to both learn and pass time. Feedback has been received stating that that certain stages of previous receiver experiments were either too basic or too complex and thus a particular receiver was not built. This web site is as much a cookbook as anything. Kludge together whatever receiver stages you want; no project is meant to be set in stone. This receiver has a high popcorn factor with MPF102 and 2N3904s as the main semiconductors.

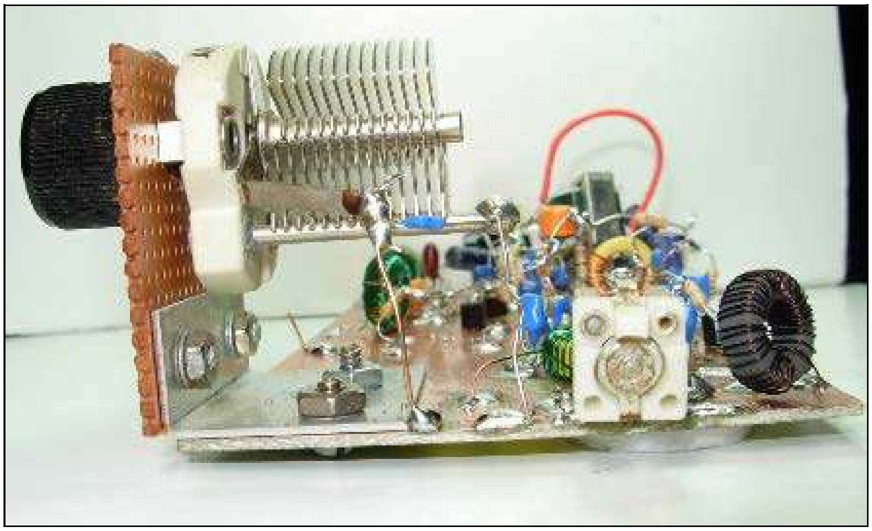

Variable Frequency Oscillator

Figure 1

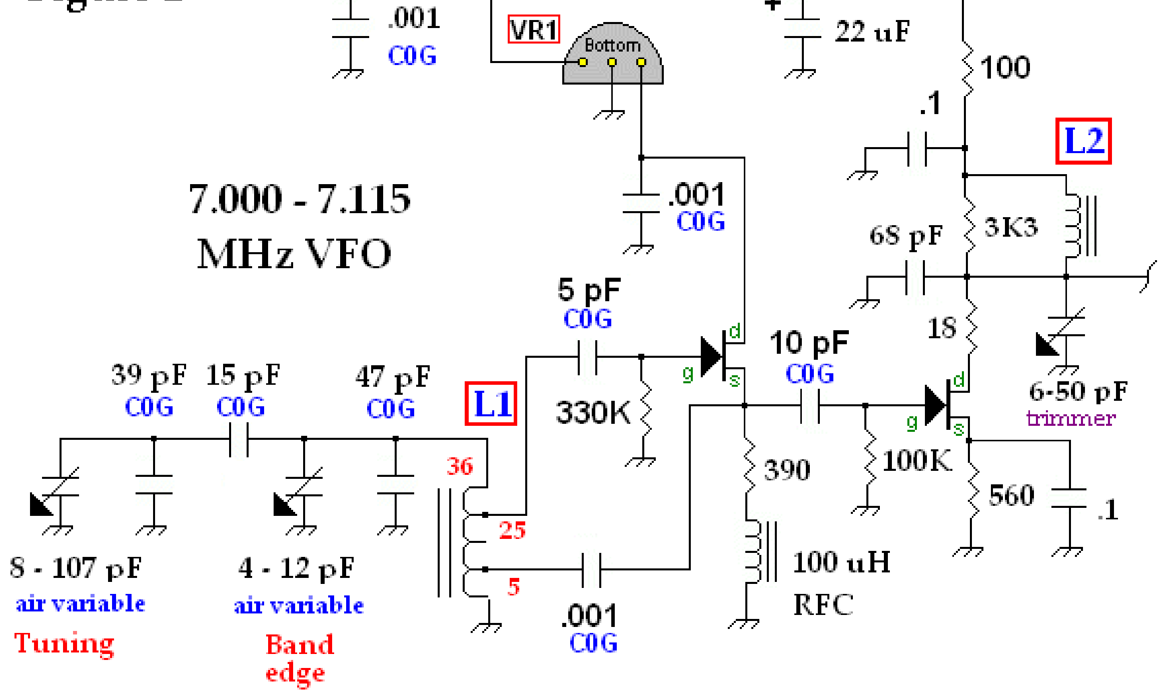
7.000 - 7.115
MHz VFO
12.24v
.001
C0G
VR1
Bottom
+
22 uF
100
.1
L2
.001
C0G
68 pF
3K3
5 pF
C0G
d
g
s
10 pF
C0G
18
39 pF
C0G
15 pF
C0G
47 pF
C0G
L1
330K
6-50 pF
trimmer
36
390
100K
560
.1
25
5
8 - 107 pF
air variable
Tuning
4 - 12 pF
air variable
Band
edge
.001
C0G
100 uH
RFC

L1 = 6.1 uH = 36t on a T68-6 torroid; tap at 5 and 25 turns (used 26 AWG)
L2 = 3.6 uH = 30t on a T50-6 torroid or 35t on a T37-6
N-channel JFETs : MPF102
100 uH RFC used was wound; 38t on a FT50-61 ferrite torroid

The first stage built was the VFO shown above in Figure 1. The oscillator portion is based upon Figure 4.15 from EMRFD. The VFO resonator tank is isolated from the JFET by tapping down as shown. This is an outstanding VFO topology. See this web page for a few more details and a coil tap calculator. I favor high L to C ratios in my RF tanks, although this does not affect the VFO function. The tapped inductor in this oscillator allows you to use a high RF voltage (low C + high L) while still keeping the FET gate AC voltage at a reasonable level. The buffer amp was designed for high output power and supplies nearly 5 volts peak to peak to the product detector local oscillator port. You can vary the output voltage by increasing or decreasing the 15 pF coupling capacitor for use in other projects.
To peak the L2 tank trimmer capacitor, use a scope, RF voltmeter, or temporarily connect a 10K (or greater value) resistor load to ground via a 10 - 47 pF output capacitor and adjust this capacitor while listening with a nearby CW receiver. (Use a short piece of wire as an antenna.) Additionally, you could also peak this trimmer cap while listening to a CW signal with the completed 2 Bravo receiver. It takes around 100 pF to resonate the L2 tank at 7.040 MHz in case you are wondering.

Since air variable capacitors were used for tuning and to set the band edge, Q is high and frequency stability is excellent. My 1 hour frequency drift was 50 Hertz uncovered. The high RF energy in the tank circuit results in low noise. The L1 taps also allow the use of a 5 pF gate coupling capacitor rather than the hard to locate 3.3 pF cap used in many example VFO schematics. With different buffer/amps as required, this is now my number 1 VFO topology and it is nothing short of stellar. Note that the 100 uH RFC can be wound with 15 turns on an FT37-43 ferrite torroid, or replaced with a fixed value choke.

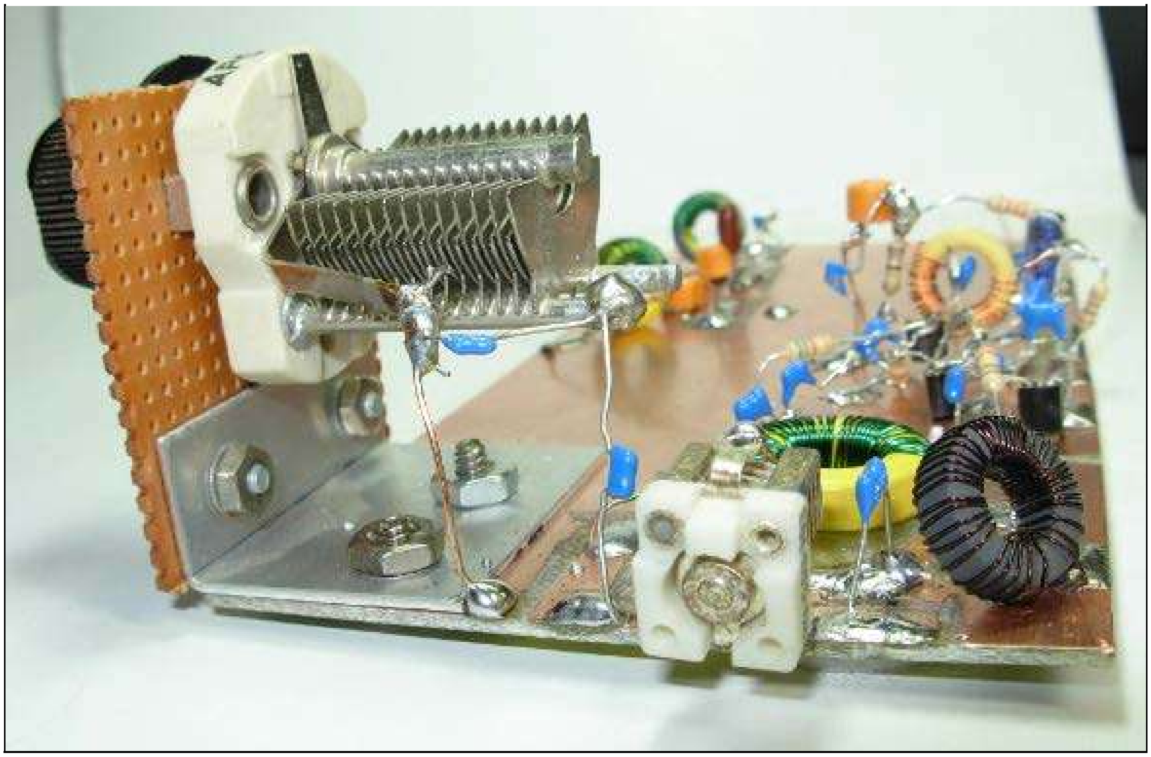

In the above photograph is the VFO bread board. I used 26 gauge wire for the inductor and took my time to make sure the wire was laying flat on the T68-6 torroidal core. You can pull the wire tighter if you wash your hands before winding.

Band pass Filter and Product Detector

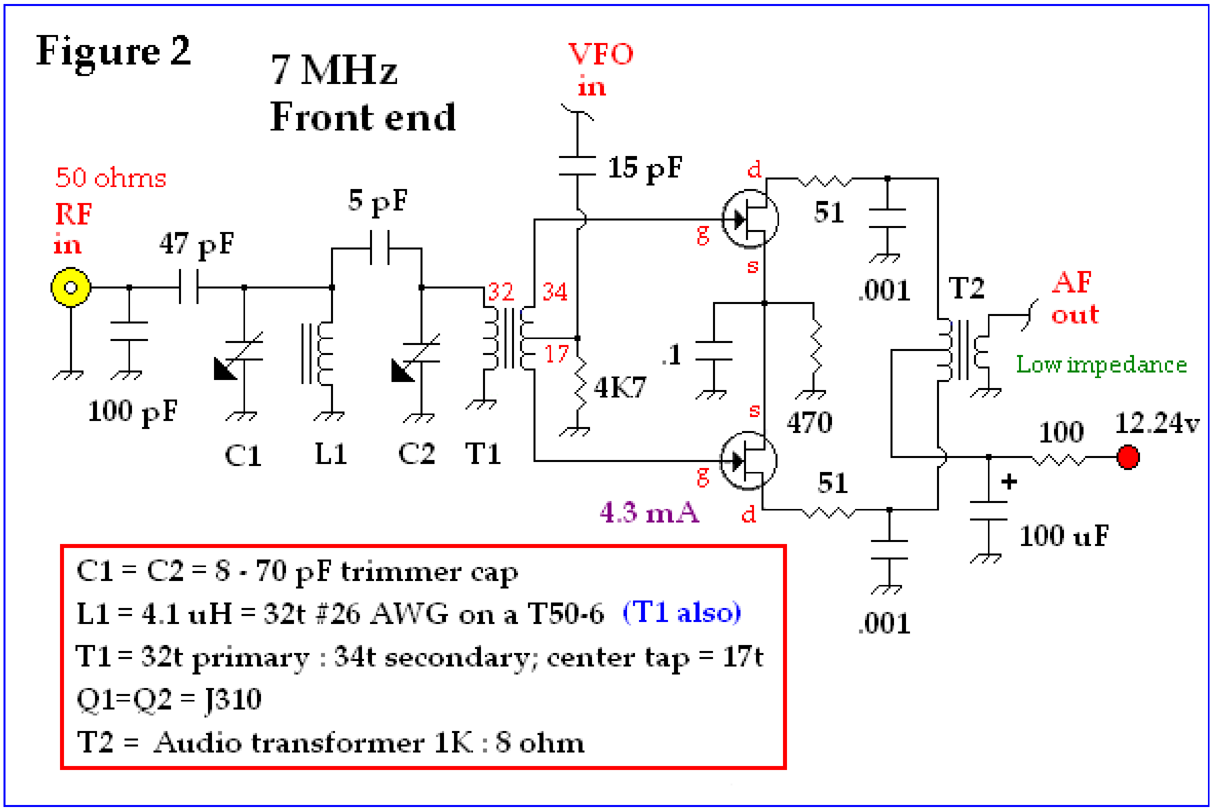
Figure 2
7 MHz
Front end
VFO
in
15 pF
50 ohms
RF
in
47 pF
5 pF
100 pF
C1
L1
C2
T1
32
34
17
4K7
d
g
s
51
.001
.1
470
s
g
d
51
.001
4.3 mA
T2
AF
out
Low impedance
100
12.24v
+
100 uF
C1 = C2 = 8 - 70 pF trimmer cap
L1 = 4.1 uH = 32t #26 AWG on a T50-6 (T1 also)
T1 = 32t primary : 34t secondary; center tap = 17t
Q1=Q2 = J310
T2 = Audio transformer 1K : 8 ohm

Please refer to Figure 2. The second stage constructed was the double tuned band pass filter. You will need about 50-54 pF to resonate L1 and T1 at 7.040 MHz. The C1 and C2 values chosen are thus perfect for tuning the 40 Meter CW band. I peaked my particular front end filter at a center frequency of 7.025 MHz using a 50 ohm output impedance RF generator and then did some fine tuning with an antenna connected after the receiver was constructed. You may also just tune C1 and C2 for maximum signal strength when listening to band noise and QSOs. Filter bandwidth is sufficient to cover the whole CW sub-band. No AM broadcast band radio was heard during several nights of testing.

The product detector is single balanced for improved port isolation and BCB rejection. Lay out your circuit to try to achieve symmetry. The schematic calls for J310s. I built the first prototype with MPF102 that were matched for Idss. To find two with the same Idss, I had to measure 16 transistors! This is too painful, and I recommend just using a pair of J310s. The words "matched" and "MPF102" should not be used in the same sentence! Ideally, your J310s should be matched, however, the process should not take as long as for MPF102 JFETS. The choice is yours to make. T1 is a little tricky to wind, however, your best effort should be good enough.
Some builders will be unhappy with using a audio transformer (T2), however, they are still in catalogs and online stores, or can be harvested from an old transistor radio. CB radio modulation transformers are also a possible source. A higher impedance audio transformer, will likely give even more conversion gain. Without the 51 ohm drain resistors, oscillations occurred in my bread board.

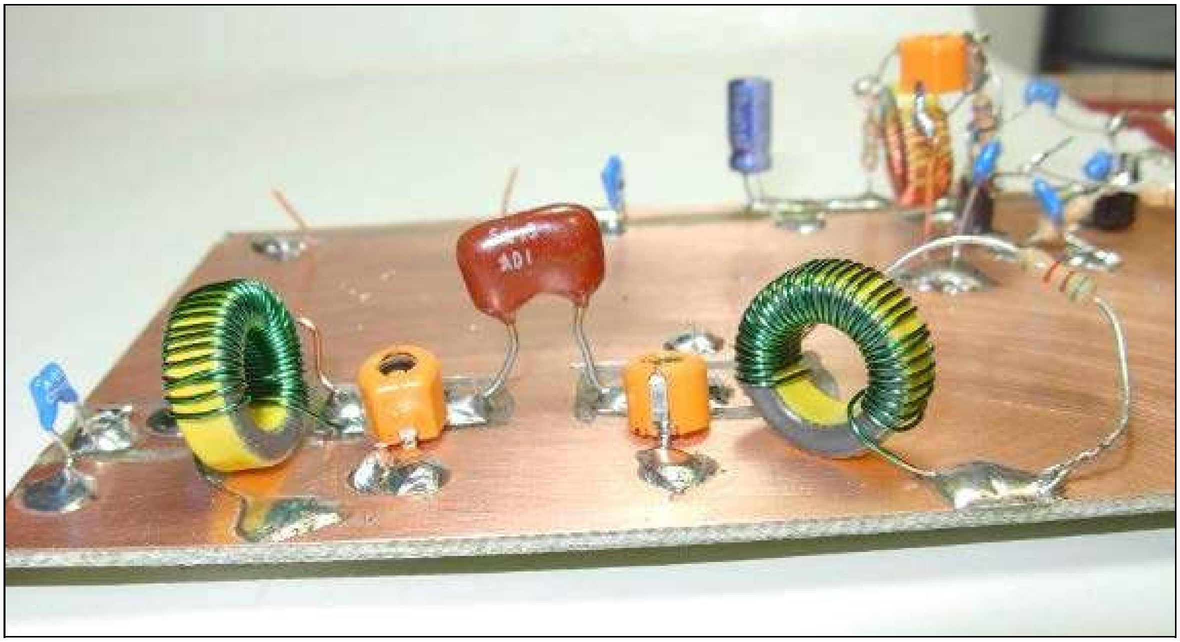

Above. A temporary 5K1 (5.1K) resistor was soldered across the second tank for testing when the front end filter was designed on my work bench.

Audio Pre-amplifier

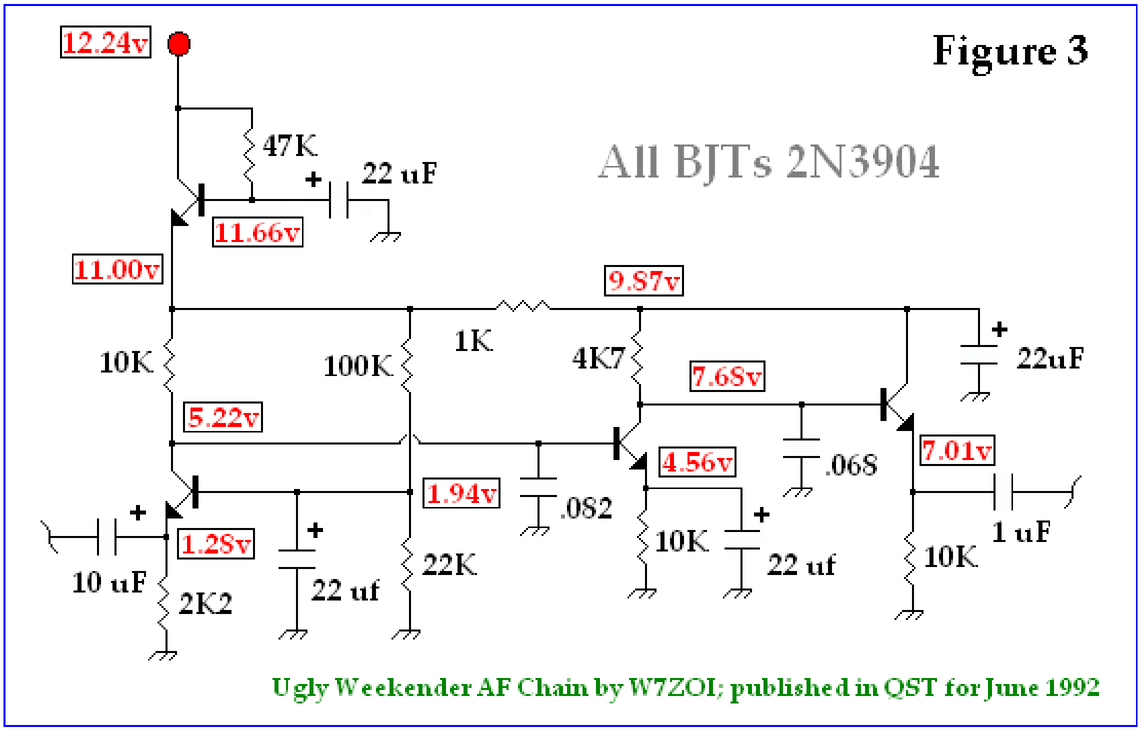
Figure 3
All BJTs 2N3904
12.24v
47K
22 uF
11.66v
11.00v
9.87v
10K
100K
1K
4K7
22uF
7.68v
5.22v
4.56v
.068
7.01v
1.94v
.082
1.28v
10 uF
2K2
22 uf
22K
10K
22 uf
10K
1 uF
Ugly Weekender AF Chain by W7ZOI; published in QST for June 1992

To match the low impedance winding of the audio transformer, a common base amp topology was chosen. I decided to use a favorite circuit; the

audio chain from the first amateur band receiver that I ever built -The Ugly Weekender. The final common emitter feedback amplifier from the original schematic was omitted as the 3 stages above provided enough voltage gain. The Figure 3 amplifier is worth studying. It is difficult to DC couple audio amplifier stages and not end up with your second and/or third stage in saturation. This example of good design by W7ZOI illustrates how to do it. The third stage, a common emitter amp is a level shifter and drops the DC voltage back down, although this stage is AC coupled to the volume potentiometer. For lower noise, you could AC couple the first common base amp to a 5532 op-amp, although, this would reduce the popcorn factor a bit. Do not expect ear blasting voltage gain from this humble circuit. It provides reasonable drive to the power amp stage.
The 0.82 and 0.68 uF capacitors shunt any detected RF energy to ground and also provide some low pass filtering. The original schematic called for 0.1 uF capacitors and which values work the best is yours to decide. Polyester film type capacitors were used in the bread board.

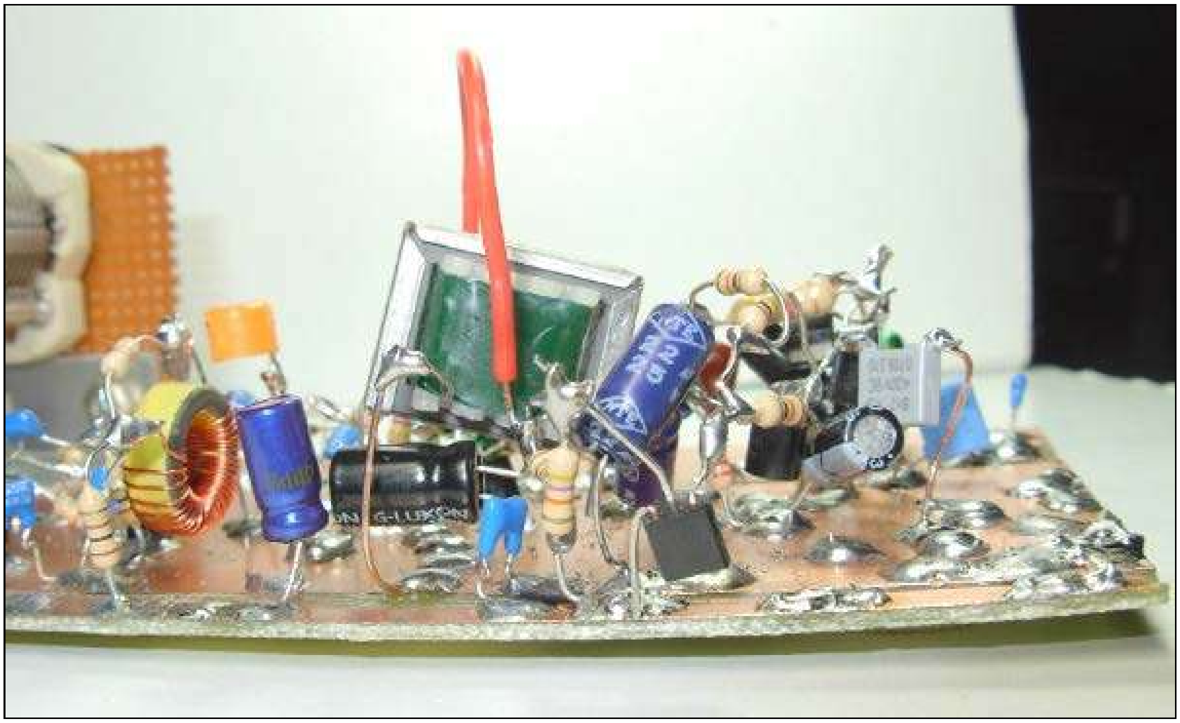

Photographed above is the Figure 3 bread board. You can also see the VFO buffer/amp and the audio transformer. I tried several AF preamplifiers, but preferred Figure 3 to all others.

Audio Power Amplifier

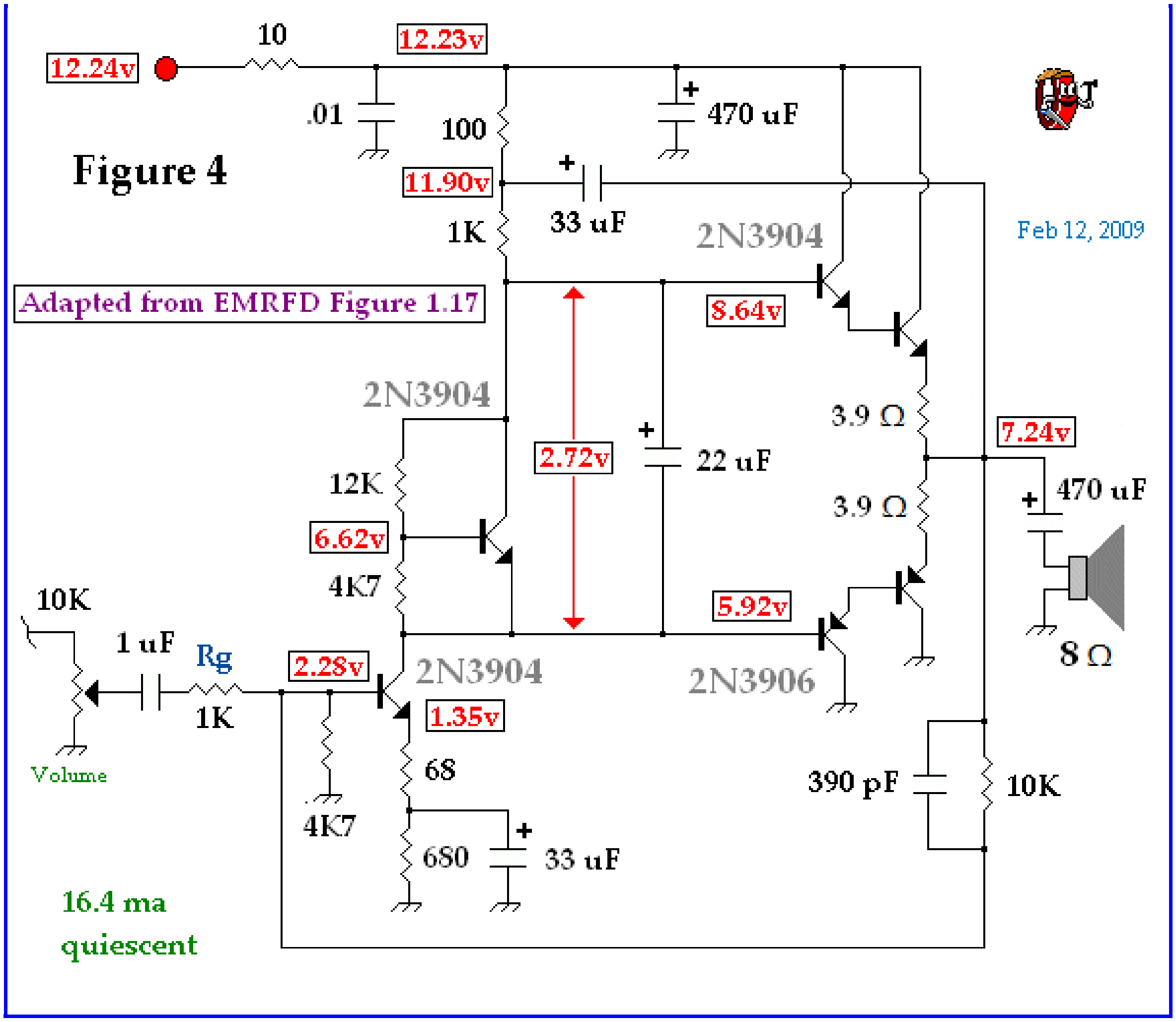
12.24v
10
12.23v
.01
100
470 uF
11.90v
33 uF
1K
2N3904
Feb 12, 2009
Adapted from EMRFD Figure 1.17
8.64v
2N3904
2.72v
22 uF
3.9 Ω
7.24v
12K
6.62v
3.9 Ω
470 uF
4K7
5.92v
10K
1 uF
Rg
2.28v
2N3904
2N3906
1.35v
8 Ω
1K
Volume
68
390 pF
10K
4K7
680
33 uF
16.4 ma
quiescent

Figure 4

Figure 4 provided 3 nights of experimentation. The base circuit for this amplifier was Figure 1.17 from EMRFD. To increase power gain and reduce harmonic distortion, Darlington configured emitter follower pairs were employed. This worked, except the power followers were under biased and had serious crossover distortion. To remedy this, the amplified diode (level shifter) bias was increased until a sine wave was seen on the oscilloscope. This was achieved by replacing the 10K resistor (R6 in the original schematic) with a 4K7 ohm resistor. The next task was to try to increase the voltage gain. Rg in EMRFD Figure 1.17 called for a 3K3 resistor. Rg was dropped to 1K; this worked. The degenerative feedback on the common emitter amplifier of EMRFD Fig 1.17 was also dropped somewhat. Next some bootstrapping feedback was AC coupled to the collector of the main common emitter amplifier. Each of the 2 collector resistor values was changed around and the outcome was recorded. Ultimately the 100 ohm plus 1K ohm resistor series pair was chosen and provided a boost of 0.85 volts peak-to-peak clean voltage gain to the output waveform. The power follower emitter degeneration resistors were also decreased from 22 to 3.9 ohms. The result is a low distortion power amp with about 150 mW of clean average power output. This receiver is not super loud, but it is reasonably loud and the audio is bell clear. If you use this amplifier stage in other projects that have a higher gain pre-amplifier, I recommend keeping Rg at 3K3 ohms as this amplifier will likely exhibit lower distortion characteristics.

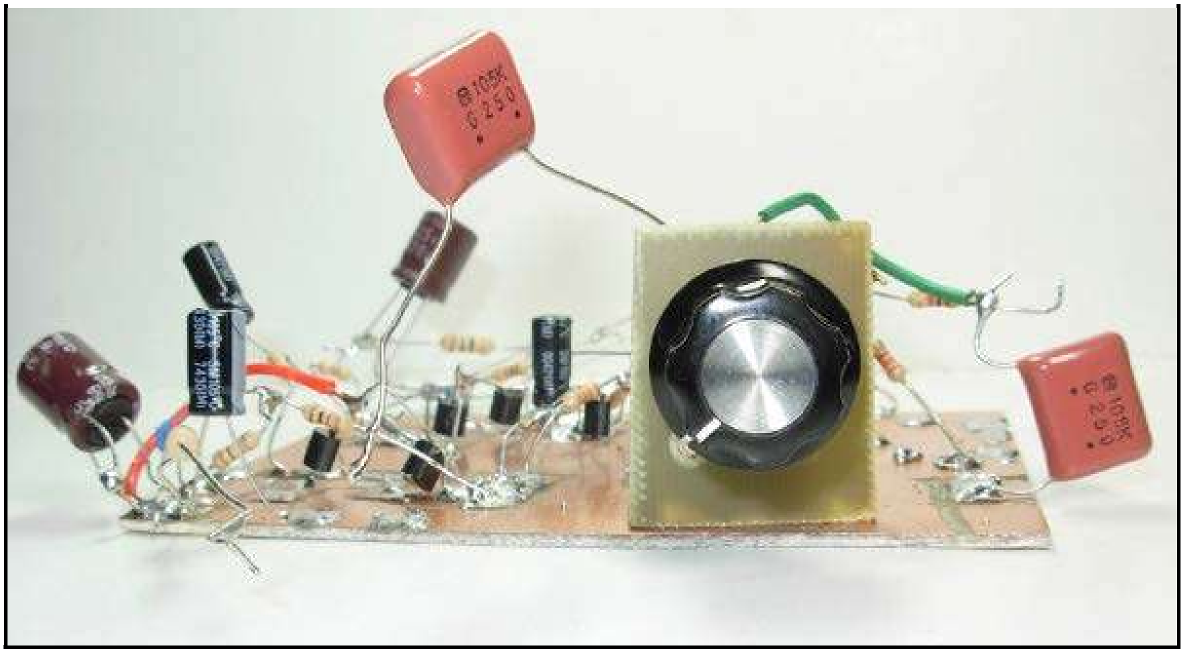

Shown above is the Figure 4 bread board. A 10 ohm resistor was used to decouple this stage. Without the resistor, audio oscillations at around 850 Hz manifested when the volume was greater than about half way up. The voltage drop across the 10 ohm resistor is trivial. You may have to increase this resistor value if you experience instability. Expect all amplifiers to oscillate and decouple them accordingly. The 390 pF feedback capacitor is required. The Figure 4 amplifier exhibits greater gain as frequency increases and in a direct conversion filter with no low pass filtering, this would be very harsh indeed. Feel free to experiment with the value of this feedback capacitor. Kudos to W7ZOI for the EMRFD Figure 1.17 schematic which serves as a great specimen to inform and challenge us experimenters. The original common emitter amplifier (Q1 in EMRFD Figure 1 .17) bias current is perfect and although I tried increasing and decreasing it, I returned to the originally specified bias resistor values.

Additional Outputs

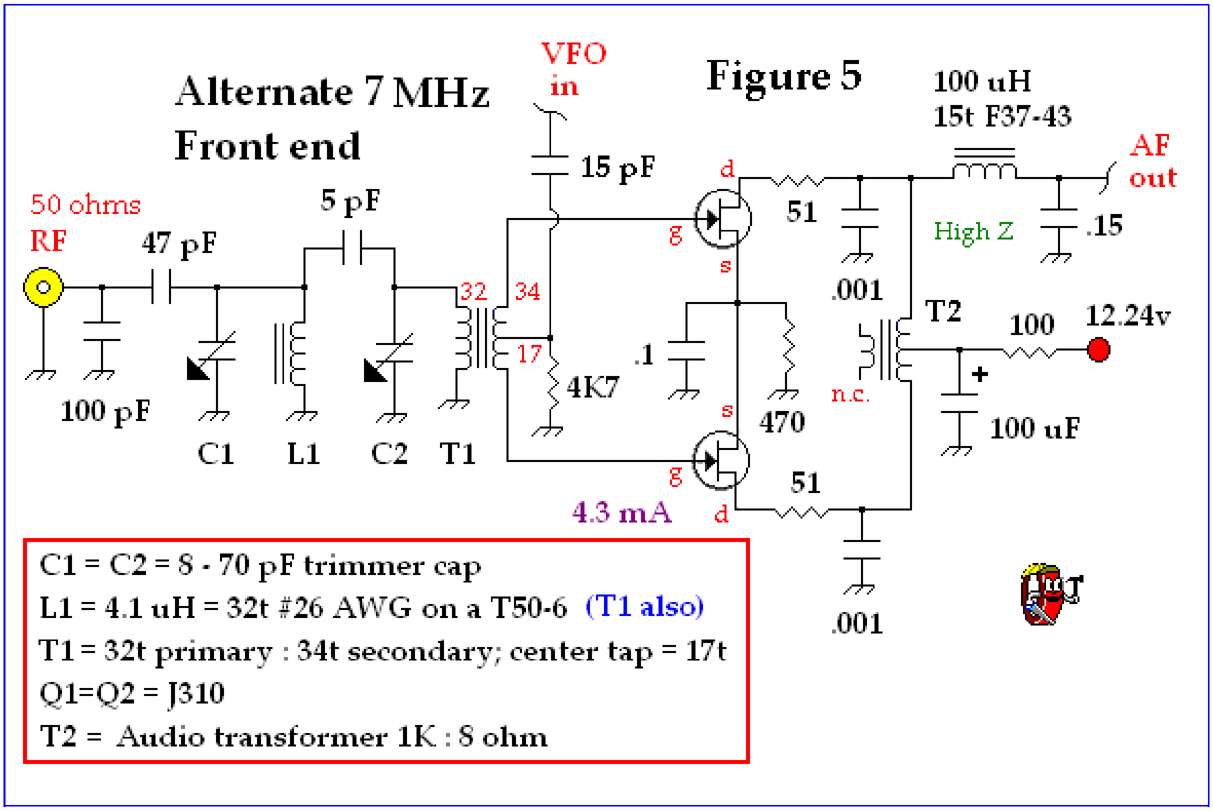
Alternate 7 MHz
Front end
VFO
in
Figure 5
100 uH
15t F37-43
AF
out
15 pF
d
51
50 ohms
RF
47 pF
5 pF
g
s
High Z
.15
.001
32
34
17
T2
100
12.24v
.1
4K7
n.c.
+
100 pF
C1
L1
C2
T1
s
470
100 uF
g
51
4.3 mA
d
.001

C1 = C2 = 8 - 70 pF trimmer cap
L1 = 4.1 uH = 32t #26 AWG on a T50-6 (T1 also)
T1 = 32t primary : 34t secondary; center tap = 17t
Q1=Q2 = J310
T2 = Audio transformer 1K : 8 ohm

Figure 5 depicts the product detector with a high impedance output. From my experiments at least, it was better to use the low Z coil for improved product detector balance and audio voltage gain.

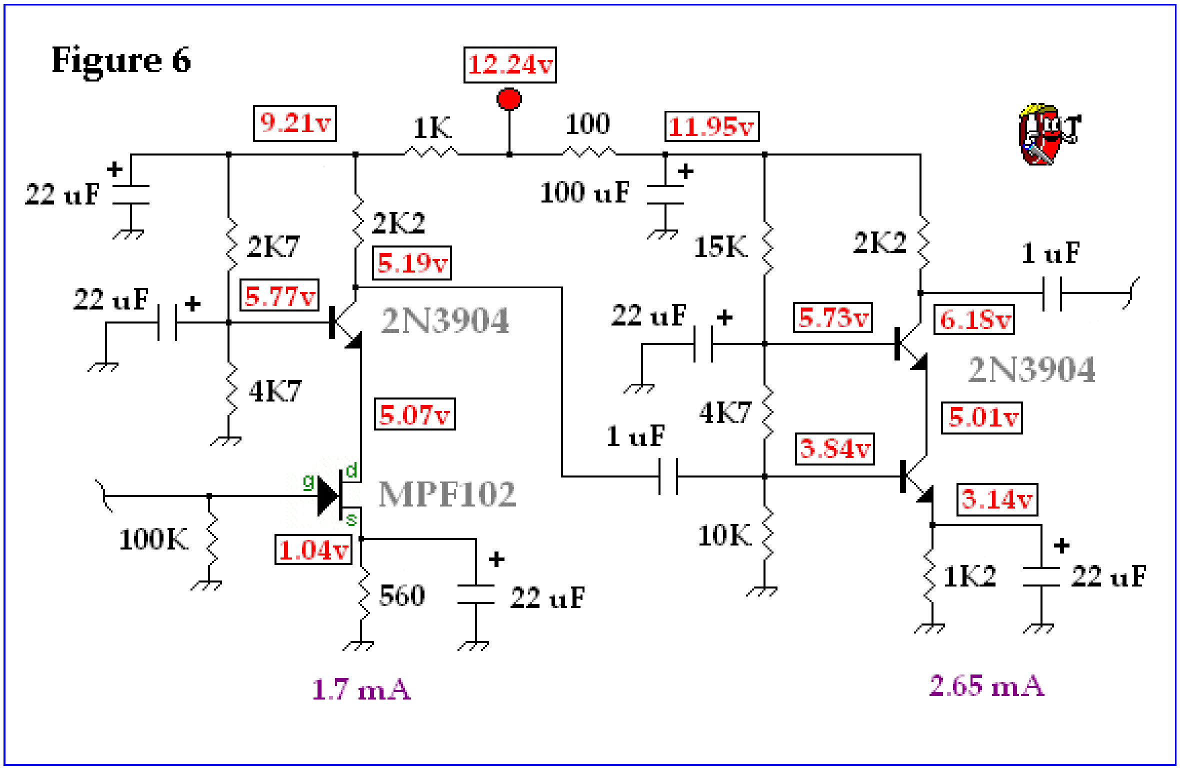
Figure 6
12.24v
9.21v
1K
100
11.95v
22 uF
2K7
2K2
5.19v
100 uF
15K
2K2
1 uF
22 uF
5.77v
2N3904
22 uF
5.73v
6.18v
2N3904
4K7
5.07v
4K7
5.01v
1 uF
3.84v
g
d
MPF102
s
3.14v
100K
1.04v
10K
1K2
22 uF
560
22 uF
1.7 mA
2.65 mA

Figure 6: A high impedance input audio stage. The first stage is a hybrid cascode. The second stage is common emitter, common base cascode. Care was taken with transistor biasing to try to optimize distortion characteristics. Certainly, I am a total novice with such amplifiers and more time on the bench and also with computer simulation is required to better understand these amplifiers. At any rate, the schematics with DC voltages are posted for others to study and hopefully improve.

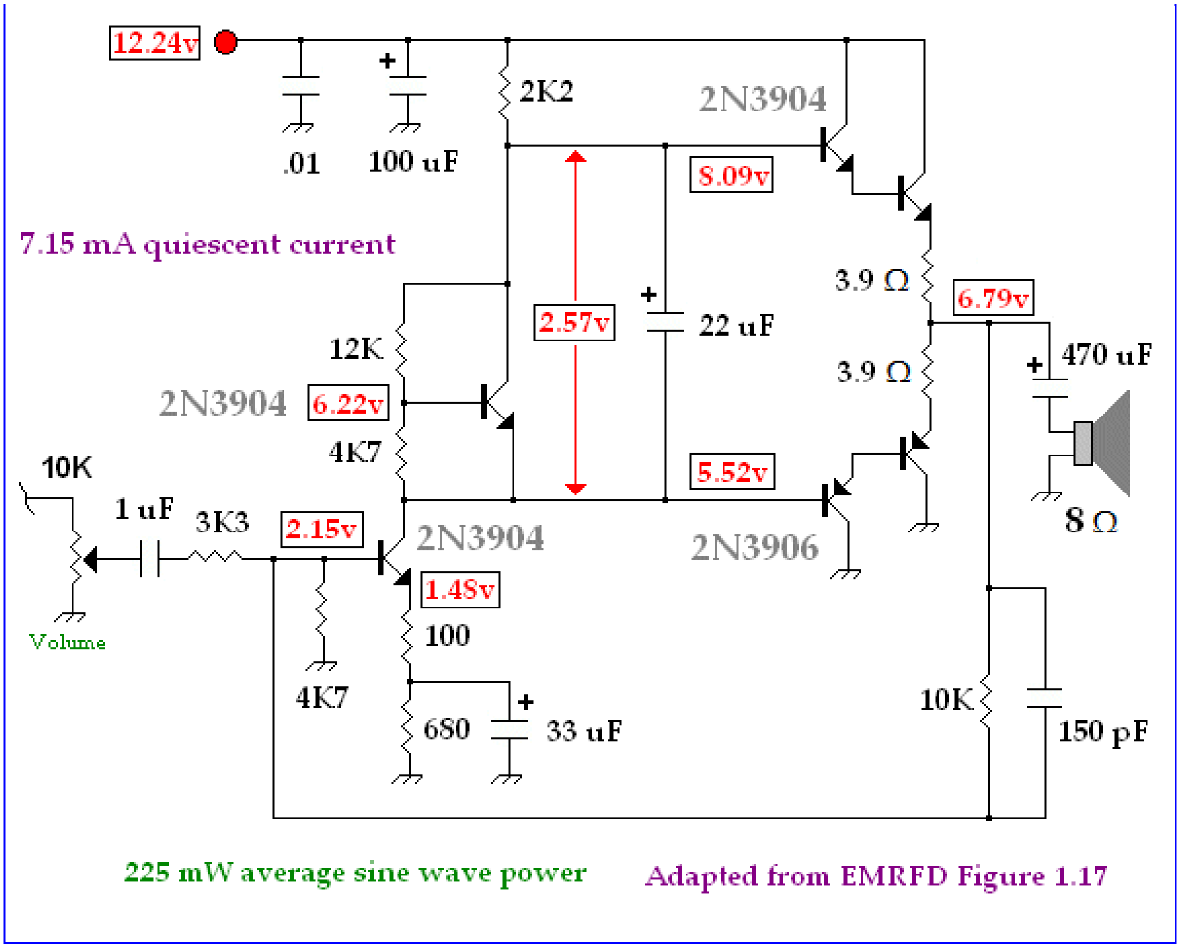
12.24v
.01
100 uF
2K2
2N3904
8.09v
2.57v
22 uF
3.9 Ω
6.79v
470 uF
3.9 Ω
12K
2N3904
6.22v
4K7
10K
1 uF
3K3
2.15v
2N3904
5.52v
2N3906
8 Ω
1.48v
100
Volume
4K7
680
33 uF
10K
150 pF

7.15 mA quiescent current

225 mW average sine wave power

Adapted from EMRFD Figure 1.17

I have 3 adapted versions of the EMRFD Figure 1.17 amplifier in my note book. This is my favorite and has the greatest clean maximum average power output of all of the 3 versions. This power amplifier somewhat lacks sufficient voltage gain for the 2 Bravo receiver (with its relatively low gain audio pre-amplifier) and thus Figure 4 was chosen as the more suitable power amplifier. You could lower the 3K3 resistor to increase the voltage gain. This circuit begs experimentation.

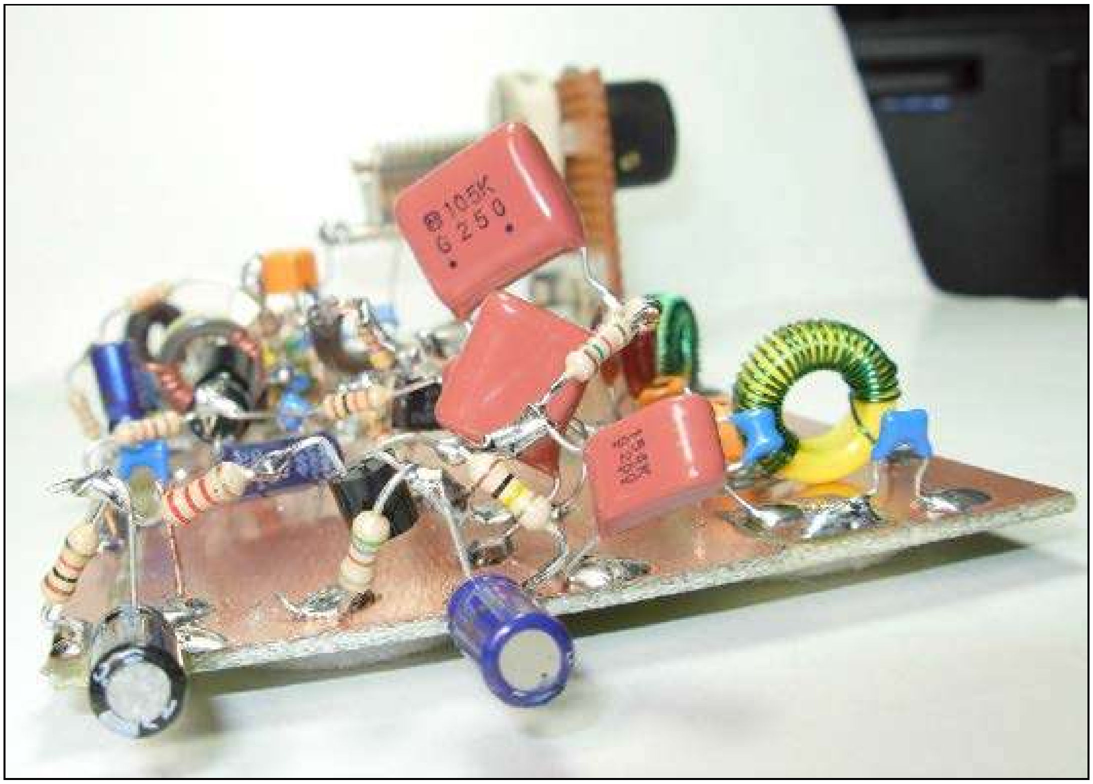

A couple of low pass RC audio filters were tried but later abandoned. One filter with its 1 uF AC coupling capacitors is shown in the photo above. It is quite an experience to hear an unfiltered direct conversion receiver. I love the purity. This is okay for an experimental or casual receiver, but not for a contest rig. Low pass filtering is definitely required in that context.

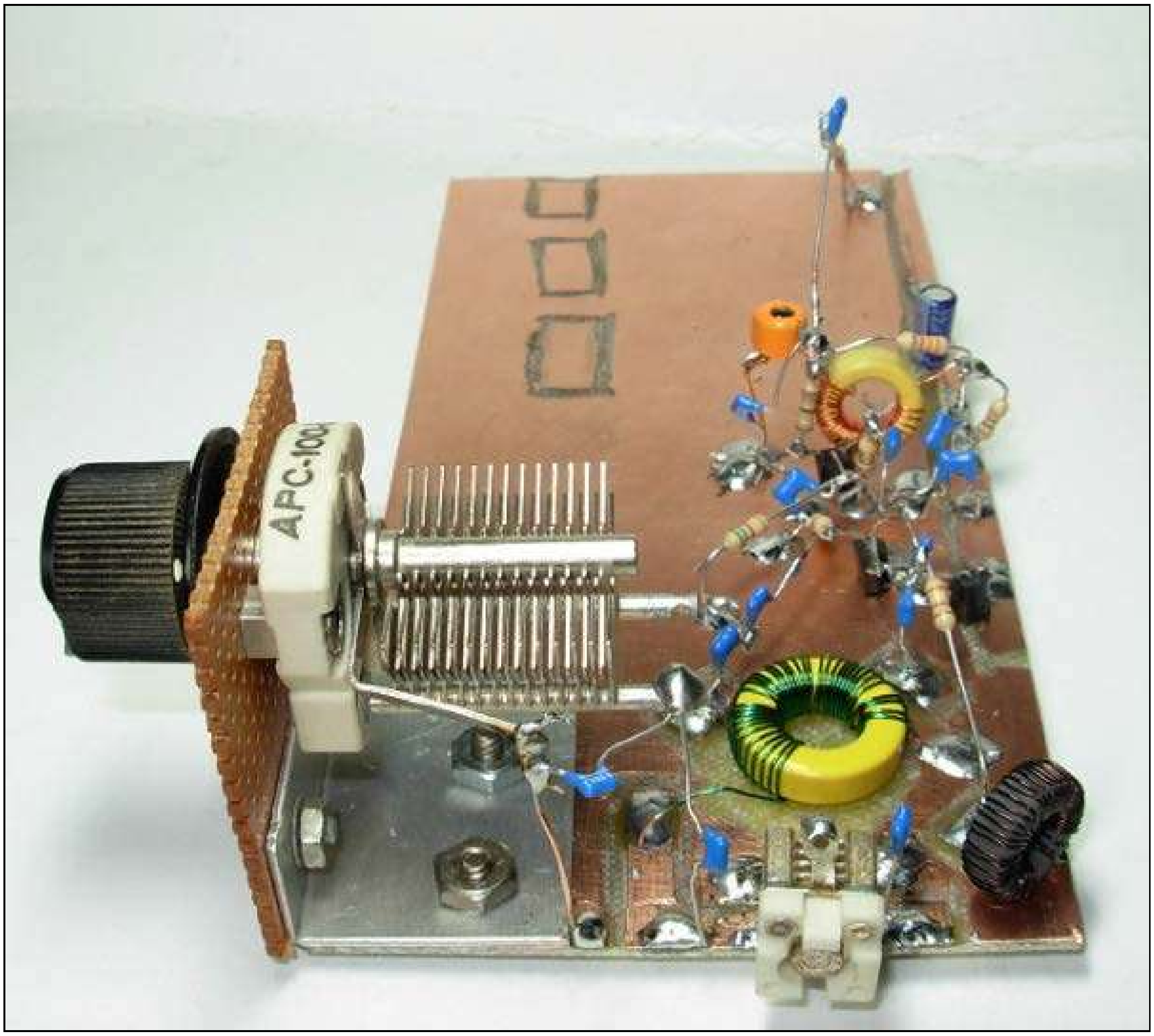

A different angle photo of the VFO bread board. My "build most of the project on 1 copper clad board" construction technique is not really suitable for "a keeper" receiver. VFOs should ideally be in a shielded box. Proper construction techniques and grounding ideas for DC receivers can be found in EMRFD, so they are not covered on this web page.

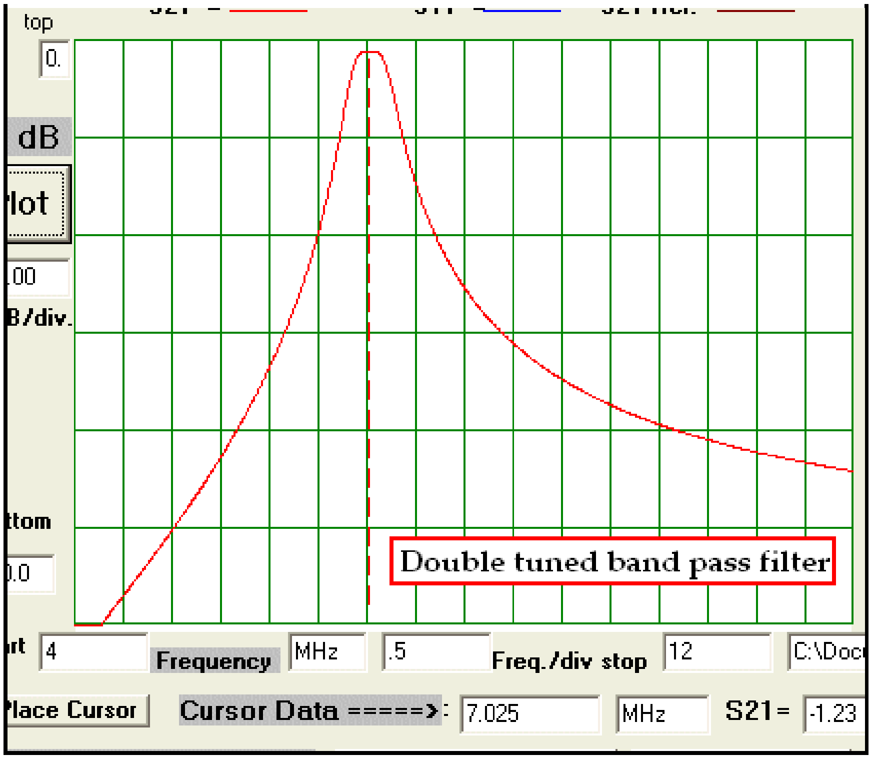
S21 =
S11 =
S21 ref.
dB
lot
Double tuned band pass filter
Place Cursor

top 0.

.00 B/div.

ttom .0

rt 4 Frequency MHz .5 Freq./div stop 12 C:\Doc

Cursor Data =====>: 7.025 MHz S21= -1.23

A GPLA simulation of the front end band pass filter centered at 7.025 MHz.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

ICOM IC-7200 Listening Tests and Observations

Introduction

NB
NR
ANF
METER
V/M 1 1.8
A/B 2 3.5
SPLIT 3 7
MW 4 10
M-CL 5 14
AGC 6 18
COMP 7 21
SCAN 8 24
VOX 9 28
RIT GENE
MNF 0 50
F-INP ENT
BAND
P.AMP
ATT
M-CH/RIT
SET
M-CH
MNF
RIT

In Spring 2009, I evaluated and photographed the ICOM IC-7200 transceiver. For this web page, only its receiver was evaluated for HAM and SWL purposes by using my eyes and ears. For technical evaluation, please click on these assessments written by Adam, VA7OJ/AB4OJ or Peter, G3SJX . Additionally, eHam.net has a review web page to consult and the ARRL publication, QST for June 2009 has a review. With my modest camera and lighting equipment, it was difficult to well photograph this very dark colored transceiver indoors.

I like ICOM radio equipment and am therefore biased in my review. Please consider trying this and any other radio out before you purchase it. Our needs, expectations and budgets tend to be uniquely different. Like others, I would rather own the new ICOM IC-7600, Yaesu FT2000D or Elecraft K3, however, my budget does not allow this.

It appears that the target audience for this transceiver is as follows: portable or field/emergency communication usage and/or it is oriented towards entry level HAMs or perhaps those wishing a modern DSP-filtered back up rig. My review is from a SWL/HAM perspective.

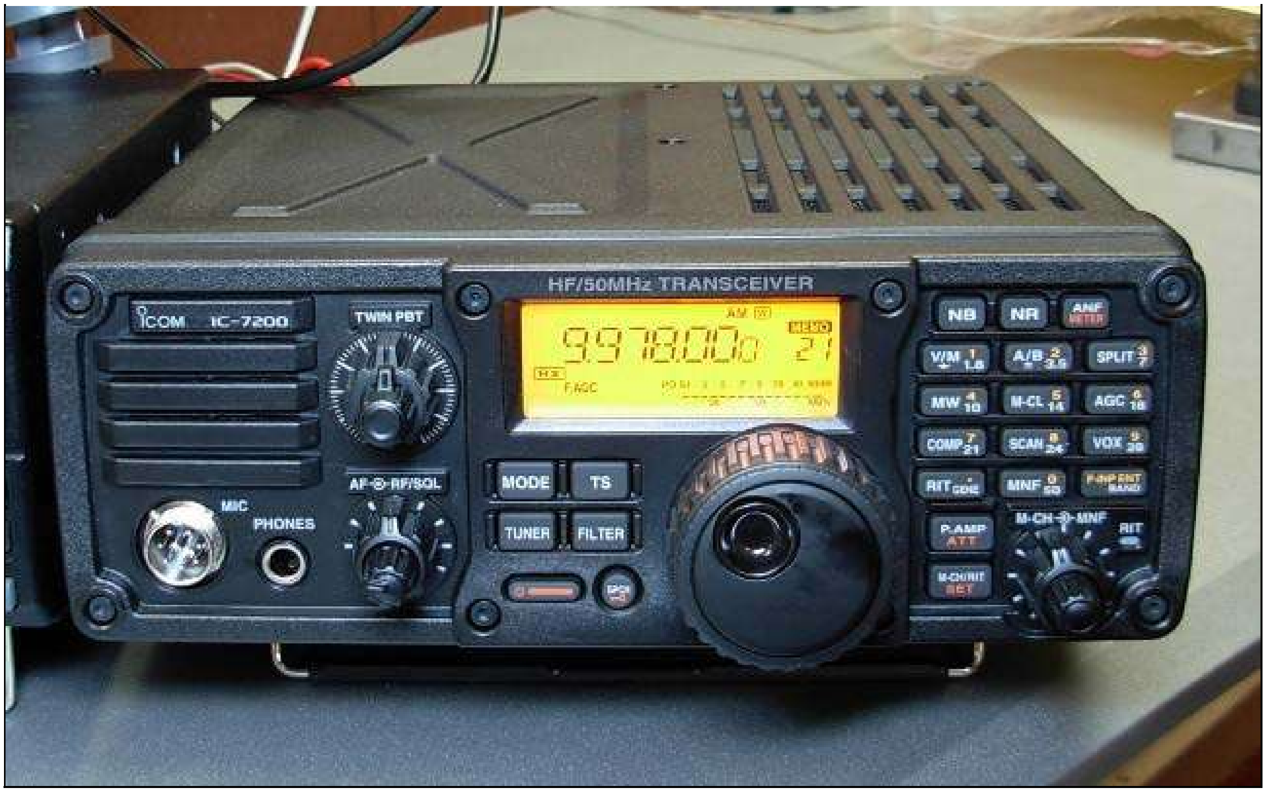

General Thoughts

Although modern and stylish, this radio is very easy to use. The owners manual is well written with clear examples of how to perform the various setting changes. After reading the manual and and trying the radio buttons and knobs out, I pretty much mastered receive operation on my first night. The LCD display is small, however is reasonably adequate considering that this whole radio is compact. I believe some operators will have difficulty with this diminutive display. The S-meter is a little difficult to see. You can push a front panel button which announces the S-Meter reading, set frequency and mode in English or Japanese language if you can't quite see the display at an odd viewing angle.

The VFO knob has adjustable resolution and works well, however, it lacks that silky smooth/weighted feeling of many other radios including the R75 receiver. The ICOM engineers had there work cut out for them; include all the modern interference fighting features; place them in a small chassis; make it easy to use and come in on budget. They did it! This radio is full of useful interference management tools. One example is the (DSP) manual notch filter; it works superbly. There is no FM mode. I imagine by dropping FM mode capability , the designers were able to use the specified 6 KHz roofing filter at the second IF which has the potential to improve dynamic range at close-in signal spacing for some modes. When I started in Amateur Radio many years ago, roofing filters were never mentioned, but have become a huge marketing lever and seemingly a topic of much confusion.

DSP IF Filtering

The DSP IF filtering works very well. Does it function better than more traditional crystal IF filtering? Yes and no. It is a question of compromise for me. The greatest DSP attribute is that there are no expensive crystal filters to purchase. Additionally, you can customize the desired IF filter bandwidth (wide, medium and narrow) for each mode and also set a "hard" versus "soft" filter shape. The soft shape equates less ringing and potentially less listener fatigue than the hard setting. I have never liked listening to CW signals through stiff, 6-8 pole 250-500 Hz crystal filters on any receiver, so for me, adjustable DSP IF filtering is preferable. DSP IF filtering is not perfect as authors like Rob Sherwood, NC0B have presented, however, ICOM have a pretty good DSP platform out now and I am pretty sure their latest technology from the IC-7600/PRO 3 on down have also been used in this receiver. After all, you can always try to find an old Drake and order some crystal filters for it, or spend significantly more cash on a high-end transceiver if you need better performance.

Shown to the the left is a few front panel controls and the front firing speaker. This speaker is reasonably nice sounding. There seems to be less harmonic distortion and audio gain than the R75 receiver. The radio front panel controls are well thought out and seemed intuitive to use after reading the manual and/or just trying them out. The various hardware components such as potentiometers, rotary switches, microphone jack, PL-259 jack etc seem not to be of the highest quality. When you compare such components to that used in their older designs, there is

ICOM IC-7200
HF/50MHz
TWIN PBT
AF RF/SQL
MIC
PHONES
MODE
TS
TUNER
FILTER
SPCH

evidence of modern cost containment. I have seen many reviews describing this transceiver as rugged. Certainly the diecast frame is solid, and I hope the forementioned hardware is as well. There is only one antenna connection. I believe that any radio offering the 6 meter band should offer two SO239 antenna connections, however, this likely would have crowded the back panel layout and increased cost for the RF in/out switching.

The VFO knob (shown to the right) looks and feels a little cheap. Functionally, however, the knob spins very well and precise tuning is easy to perform. One of my first tasks was to input various the frequencies I use for both amateur and short wave listening into the (201 maximum) memory slots. This was very easy to do and each memory channel also stores the mode and filter setting.

HF/50MHz TRANSCEIVER
MODE
TS
TUNER
FILTER
SPCH
V/M
MW
COMP
RIT

AM Reception (R75 versus IC-7200)

The IC-7200 is a good short wave receiver. With a high quality external speaker connected, pleasant and warm sounding AM audio may be heard; however it is not Hi-Fi. That is; the IC-7200 is not an audiophile AM receiver. There is no synchronous AM detection and no 15 KHz wide IF filter for example. For AM reception, the user may choose from an 8000 to 800 Hz IF bandwidth in the 3 switchable filter settings. During testing, I set my wide filter setting to 8000 Hz, my medium setting to 6000 Hz and my narrow setting to 3000 Hz for AM reception. Of course these bandwidths can be further adjusted at any time. Local MW broadcast radio sounded great with an 8000 Hz IF bandwidth during testing. The various AM reception appropriate interference controls such as pass band tuning (PBT), automatic notch filter (ANF) and digital noise reduction are all configurable or adjustable and I found the ANF and digital noise reduction helpful when listening on the crowded 49 meter band during a rain storm.
I performed A/B comparisons with the IC-7200 and the R75 simultaneously hooked to the same antenna. For AM reception, they are comparable with the R75 having marginally better sounding (more Hi-Fi) audio. I had the R75 IF filter bandwidth set at 15 KHz and the IC-7200 IF bandwidth set at 8 KHz (both at their maximum IF bandwidth).

CW/SSB Reception (R75 versus IC-7200)

In CW mode, both were set to have a 500 Hertz IF bandwidth.(The R75 had the FL-100 model 500 Hertz crystal filter in its 9 MHz IF slot). The IC-7200 was quieter, clearer and less overloaded by adjacent stations during pile ups. There were occasionally weak signals I was able to copy on the IC-7200 that I could not even hear on the R75 due to noise. The IC-7200 is a joy to use on CW; really fantastic.
On SSB, I found both receivers fairly comparable, but the IC-7200 was better for pulling weak signals out of the noise as the noise floor was a little quieter and the audio a little more crisp. My wife also agreed with me in her "independent tests" of the CW and SSB reception. I am certain to catch flak because my subjective comments are based upon listening tests and not measurements. It is interesting to note that in almost every aspect of our lives, preferences are made using 1 or more of our 5 senses. From listening, observing and also reading the reviews of others, it seems ICOM has a hit with the IC-7200.

Mouse over the images below left to view a full size photograph

The IC-7200 in the sun.

Rear view of the transceiver.

Top angle view. There are ventilation holes at the top right hand side. You can just see the ventilation screen just next to the right top panel screw.

Reverse angle top view. The military look is attractive.

The IC-7200 and R75 side by side.

Another photograph of the 2 radios from the front angle for comparison. The LCD display size difference is quite obvious in this shot.

The receivers at a right angle. The IC-7200 is a glad update, although, I would likely not upgrade from the R75 to the IC-7200 just for receiving purposes. If I was considering choosing the IC-718 (transceiver version of the R75), the IC-7200 is worth the extra money. It is also an excellent transceiver consideration for an entry-level rig, for back up purposes. or for the budget-minded operator.

Conclusion

The IC-7200 is a serious choice for amateurs seeking a good, modern HF plus 6 Meter transceiver. It is more than adequate as an SWL receiver if you are an amateur radio operator plus SW listener like myself. Please try one out for yourself if you are considering this transceiver.

Amateur Radio Electronic Design

Electronic Hobbyist Circuits

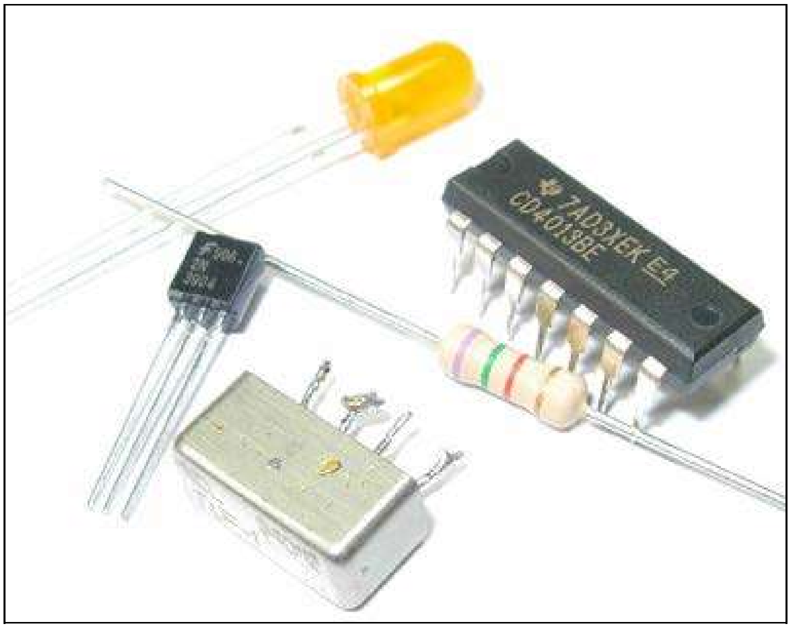

This page will house a collection of brief hobbyist experiments.

1. Pseudo-Random Number Generator

This circuit describes a simple, 6-bit random number pseudo-generator used to study binary counters and in particular, shift registers. Some very basic background information about binary counters and shift registers is provided. In reality there are dozens of different shift register topologies available and it can get quite complex. If you wish to find a good logic tutorial website, I strongly recommend Ken Bigelow's site as it has interactive diagrams. Flip-flops are also covered well on wikipedia and many other web sites.

Binary Counter: The circuit most often used as a counter is called a binary flip-flop. The basic flip-flop can be viewed as a toggle switch having either an ON or OFF position. This is the binary state 1 (HIGH) or 0 (LOW). Like the toggle switch, the binary flip-flop has 2 binary states 1 or 0. A binary flip flop counter counts in a sequence such as 0, 1, 0 ,1 etc. A straight binary counter can be built by using 1 or more flip-flops connected in a manner that the binary number stored in these flip-flops will represent the total number of trigger pulses received at the counter input.

Ring Counter: A ring counter has 2 or more flip-flops cascaded so that the output from one flip-flop becomes the input of the next flip-flop. The flip-flops are connected so that all of their outputs are at the binary state 0 except for one flip-flop. By pulsing the input of the ring counter, it will sequentially change the binary state of the succeeding flip-flop from binary 0 to binary 1. The flip-flop that contains the binary 1 indicates the count of this binary counter. The maximum number of pulses that can be counted by N flip-flops is N pulses.

Shift Register: A serial entry shift register is similar to a ring counter, except that the output flip-flop is not connected to the input flip-flop. Like the ring counter, the flip-flops are cascaded so that the output from one flip-flop becomes the input of the next flip-flop. All the set trigger and reset trigger inputs are tied together to form what is called a shift bus. Clock pulses are applied to the shift bus to cause the stored binary information to shift from left to right; one bit position per each received clock pulse. In Figure 2, this serial input/output + parallel output register has its 5th and 6th bits exclusive ORed to the serial input to form a pseudo-random sequencer, which is called a pseudo-random number generator by some.

The CMOS logic ICs used were one 4070 XOR (Exclusive OR) and three CD4013B D flip-flops. Junk box LEDs were used to observe the binary state of the clock and each of the 6 bits of the shift register.

Figure 1

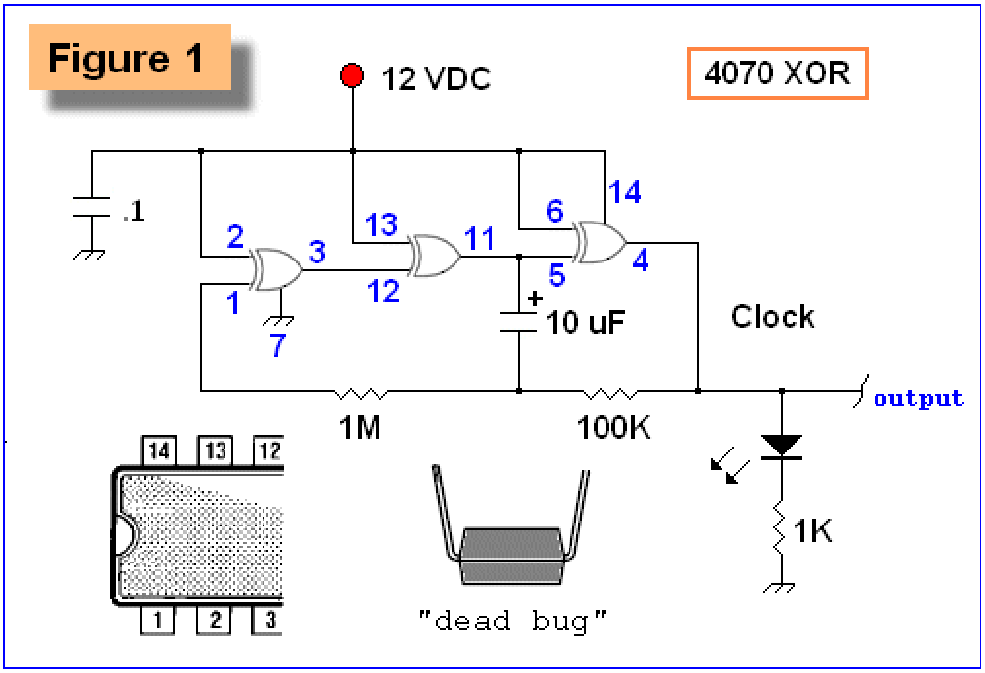
12 VDC
4070 XOR
.1
2
3
1
7
13
12
11
6
14
5
4
+
10 uF
Clock
output
1M
100K
1K
14
13
12
1
2
3
"dead bug"

Since only one XOR gate is needed for the shift register, the remaining gates were configured to make the clock. These gates are essentially wired up as inverters to form an astable multivibrator with a frequency of about 0.45 Hertz or 27 pulses per minute. Shown above in Figure 1 is the clock schematic and the pin 1 marking for all of the digital ICs on this web page. The output LED is not mandatory, but will instantly tell you whether or not your clock is working. I built this whole circuit using Ugly Construction with the ICs flipped upside down in a "dead bug" fashion. You can increase the clock speed by decreasing the 100K resistor or the capacitor values. F Hertz = 1/ (2.2 * R * C) with R in ohms and C in farads. The slow clock speed was chosen to better observe the digital output of the shift register.

Figure 2

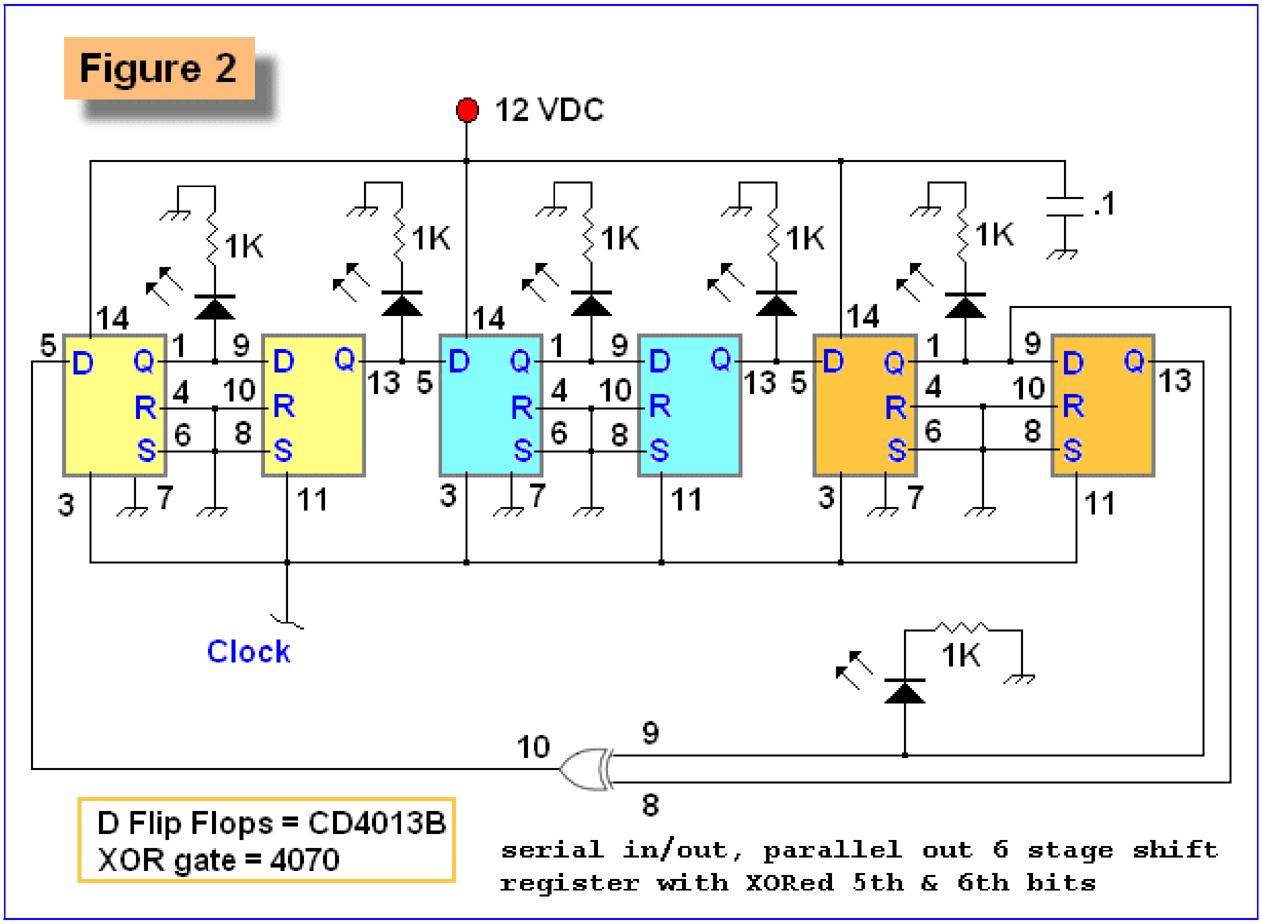
12 VDC
1K
1K
1K
1K
1K
.1
14
14
14
5
D
Q
1
9
D
Q
13
5
D
Q
1
9
D
Q
13
5
D
Q
1
9
D
Q
13
R
4
10
R
R
4
10
R
R
4
10
R
S
6
8
S
S
6
8
S
S
6
8
S
3
7
11
3
7
11
3
7
11
Clock
1K
10
9
8
D Flip Flops = CD4013B
XOR gate = 4070
serial in/out, parallel out 6 stage shift register with XORed 5th & 6th bits

In Figure 2 is the shift register. Each 4013 was wired up as 2 cascaded flip-flops and connected to the clock. Power was applied and then a test lead was used to bring pin 5 of the first flip-flop HIGH (connected to 12 volts for 1-2 seconds) . Both flip-flop state monitor LEDS turned ON in sequence with subsequent clock pulses. Afterwards, pin 5 was set LOW (shorted to ground with a test lead for a couple of seconds) and each LED turned OFF in sequence with subsequent clock pulses. The remaining two 4013s were wired up and tested the the same way and then finally the last XOR gate was wired up.

To avoid error, frequent pin counting and a systematic approach is recommended. For example, for each 4013, I soldered the ground pins, wired the pin 14 VDD, connected the clock to pins 3 and 11, then wired up the pin 1 and pin 13 LEDs. Systematic construction techniques are something that you the experimenter can develop and perfect over time. This approach saves time and grief. On some projects, when you have a lot of pins wired up, tracing and repairing an early mistake can be difficult.

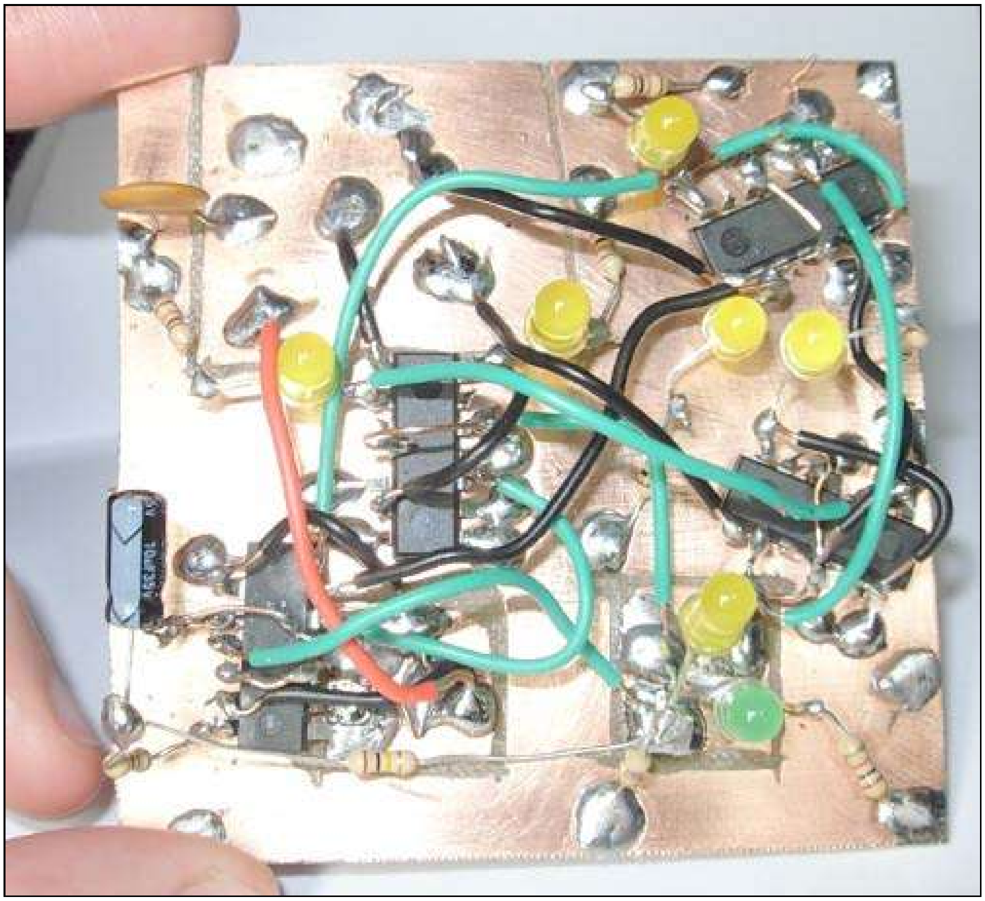

Shown above is a bread board of the entire pseudo-random number generator. I just built in on a scrap of board and did not lay it out so the LEDs were in a row, as I am not going to keep this project. The clock state monitor is the green colored LED. There are 63 possible states or combinations of the 6 bits (111000, 100110, 100101, 000101, 000001 etc.) State 000000 is disallowed and will hang up the shift register. If your clock LED is flashing and no shift register LEDs are lit, then "reset" by momentarily setting pin 5 of the first flip-flop HIGH (momentarily apply 12 volts). Long live the reset switch!

Pseudo-random numbers are now mostly generated by computer microprocessors controlled by software and have applications in cryptography, electronic music, security and many other applications. This "hardware" pseudo-random number generator experiment was really cool and if you want to randomly flash some LEDs, this could be the project to use!

If you are new to digital electronics; (like me) Welcome! Starting small with projects like this one will hopefully lead to increased confidence and problem solving skills for even bigger projects. You can also build the shift registers with J-K flip-flops, but it is more difficult and 4013s or other series D flip-flops are cheap as Борщ (borscht).

Figure 3

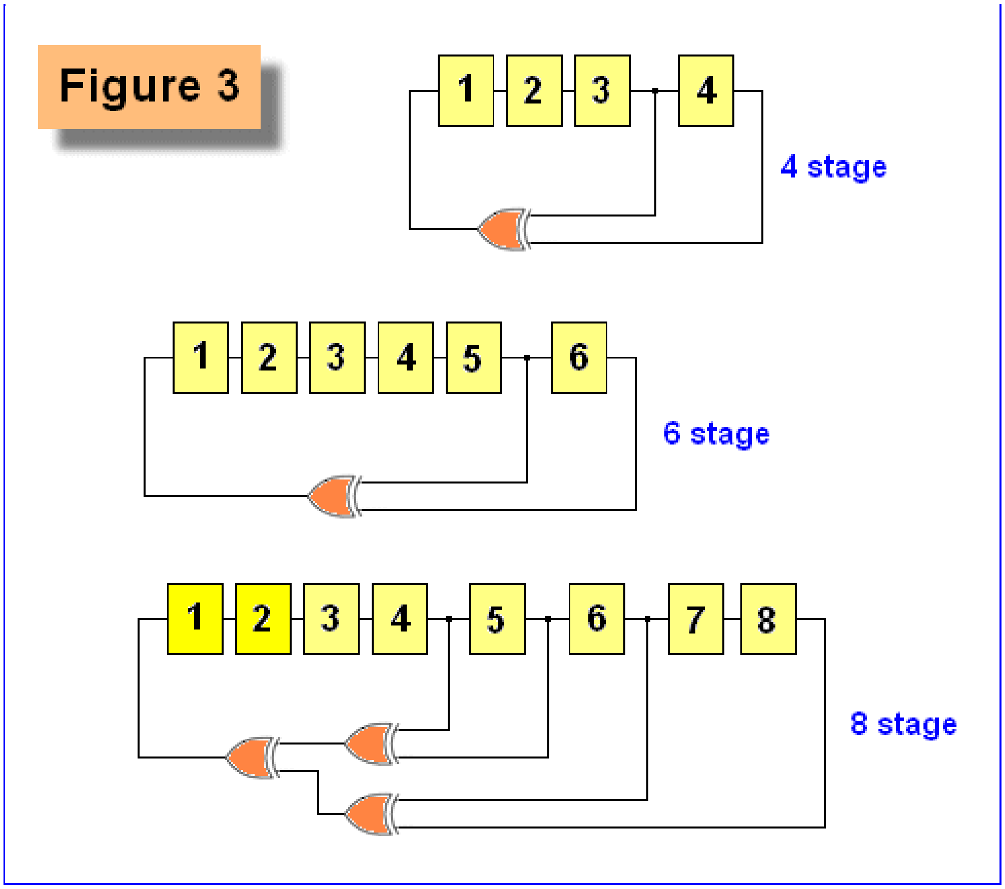

1
2
3
4
4 stage
1
2
3
4
5
6
6 stage
1
2
3
4
5
6
7
8
8 stage

Shown above in Figure 3 is how to hook up the XOR gate(s) for 4, 6 and 8 stage pseudo-random number generators. The 6 stage shift-register is of course, Figure 2 above and is presented for reference purposes. The 8 stage version = 1 byte.

2. One Hertz Precision Time Base

Digital clocks are very interesting. In the past 6 months, 10 -15 RC clocks have been constructed and tested. RC oscillators in the KHz to Hertz range are surprisingly frequency stable. For many projects, a plain RC clock is adequate, however, like in radio design, a crystal controlled time reference is sometimes required. Two examples of projects requiring precision clocks are time of day clocks and frequency counters. Presented is a 1 Hertz clock built from two 4000 series CMOS Logic ICs. Here is a great 4000 series tutorial with pin outs and more.

Figure 4

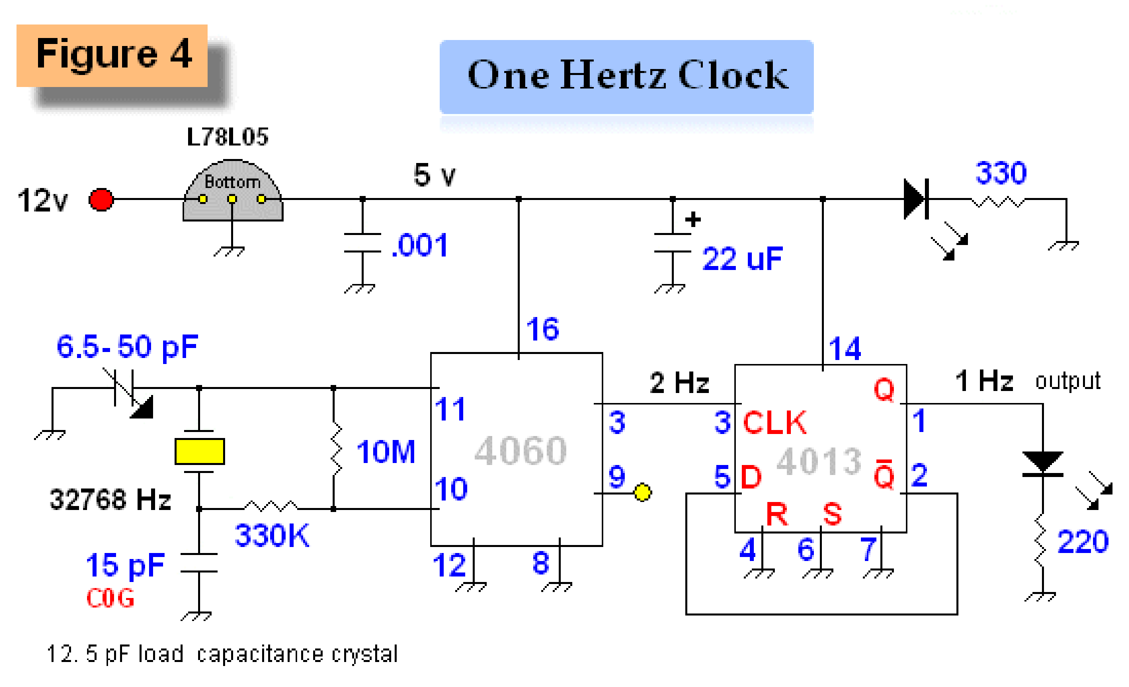

One Hertz Clock
L78L05
Bottom
12v
5 v
.001
22 uF
+
330
6.5- 50 pF
32768 Hz
10M
330K
15 pF
C0G
16
11
4060
10
12
8
3
9
2 Hz
14
3
CLK
5
D
4013
R
S
4
6
7
Q
1
Q̄
2
1 Hz output
220
12. 5 pF load capacitance crystal

Shown above in Figure 4 is the complete schematic with an output LED for testing. In the past, the MM5369 17 Stage Oscillator/Divider was popular for hobbyist precision time bases, however, it has gone obsolete. The 4060 ripple counter is a good "modern" replacement, although a different crystal is required. A 32768 Hertz crystal was used and is divided 16384 times to provide a 2 Hz output. The 4060 then drives a 4013 D Flip-flop configured as a divide by 2 to provide a 1 Hz output frequency. Key parts references may be found on the Webmaster's page.

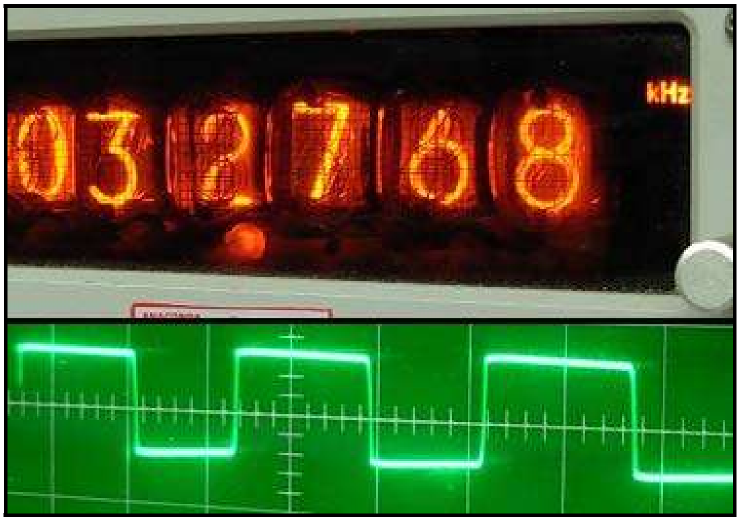

Shown above is the frequency and output waveform when a frequency counter and oscilloscope are (respectively) connected to pin 9 of the 4060 in Figure 4. The 6.5-50 pF trimmer pot is used for calibration. Originally, I used a trimmer cap instead of the fixed 15 pF capacitor shown between the 330K resistor and the crystal. After adjusting this trimmer capacitor for the best looking waveform, I removed the trimmer cap and measured it at 13 pF. I substituted the nearest standard value I had in my parts collection; 15 pF. It was interesting to measure the 4013 output frequency at 1 Hz.

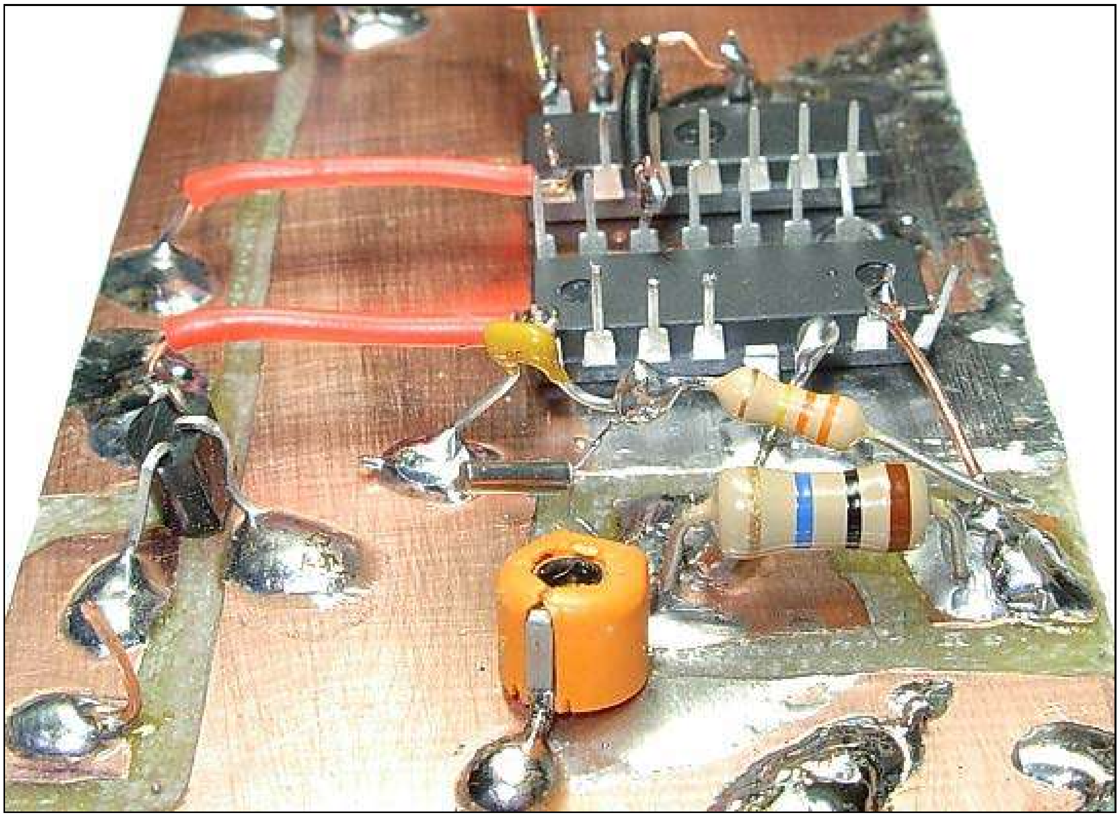

A close up photo of the 4060 oscillator/divider breadboard. The 10M resistor used was a 1/2 watt rated R as I have dozens of these in my parts collection. You can see the tiny cylindrical crystal just above and left of the orange Murata trimmer capacitor. It is oriented horizontally. This is a useful time base for the QRP workshop.

3. 10000 and 5000 Hz Multivibrator Clock

It is fun to occasionally build circuits using discrete semiconductors rather than with ICs. A 5000 Hz digital clock was needed for an experiment. It was decided to use multivibrators for the basic oscillator and a divide by 2.

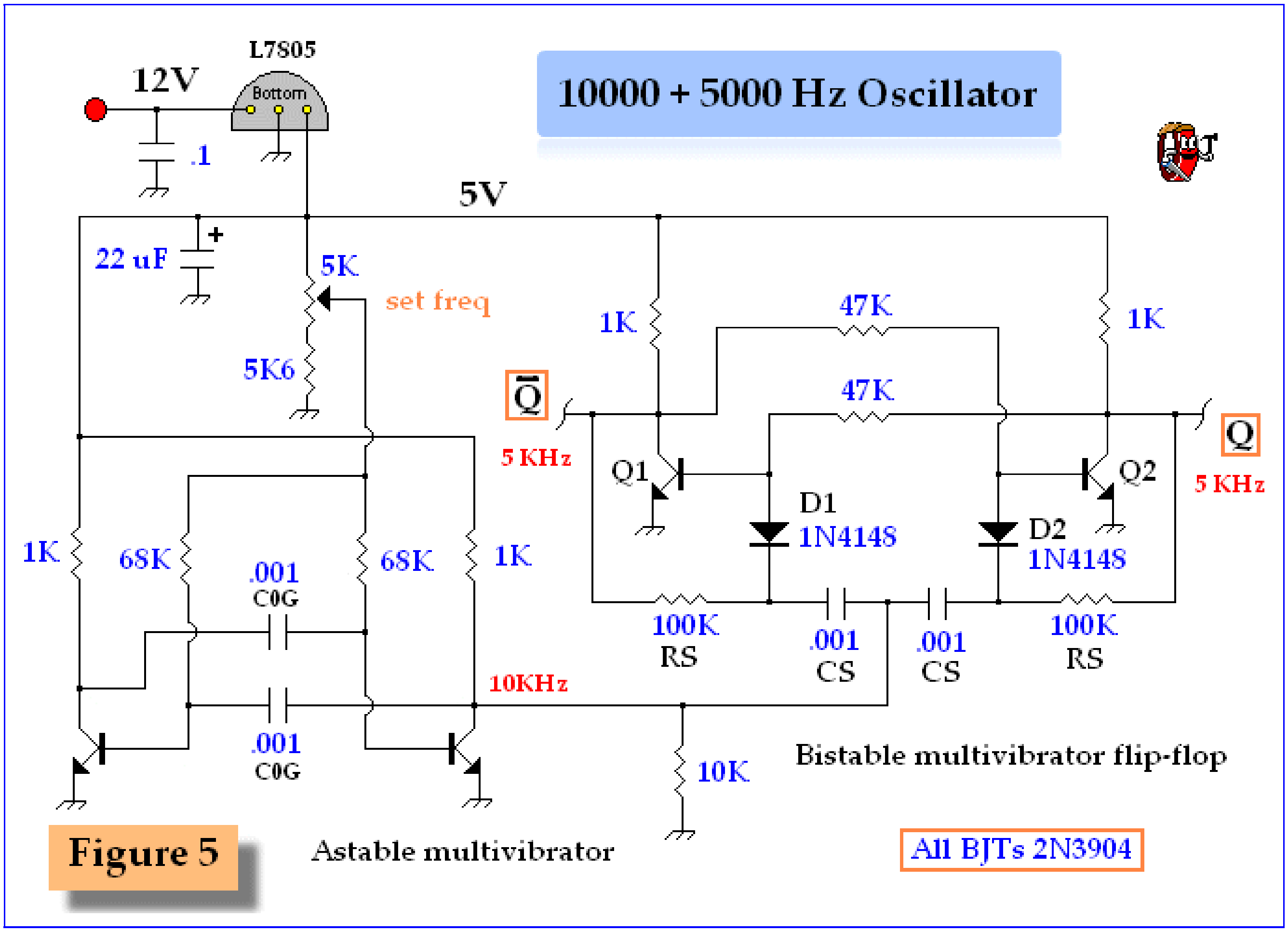

Figure 5

Figure 5 is the entire circuit. The tuning range of the astable multivibrator was about 7060-10650 Hz. The 5K pot was slowly adjusted until 10000 Hz was measured in a frequency counter. Following testing of the astable multivibrator, the flip flop was built and examined. Astable multivibrator function has been discussed previously on this web site.

Please refer to the bistable multivibrator. It is a one input circuit set up for toggle or flip-flop operation. Negative edge pulses applied between the two 0.001 capacitors will cause the binary state of Q1 and Q2 to change to the opposite state. The multivibrator circuit is made up of Q1, Q2 and the 47K and 1K base and collector resistors respectively. The other components D1, D2, the RS resistors and CS capacitors comprise a steering circuit to generate the proper response to the negative edge pulses. When a negative input pulse arrives, it is guided to the base terminal of the ON transistor, but prevented from reaching the base terminal of the OFF transistor.

In order to study this circuit at DC, I temporarily exchanged the 0.001 timing capacitors in the astable multivibrator with some 22 uF electrolytic caps to slow it down. Referring back to the bistable multivibrator, let us assume that Q1 is OFF and Q2 is ON. The collector voltage of Q1 is high (cut off). The collector voltage of Q2 is low (saturation). The Q1 collector is connected to the cathode of D1 by the 100K RS resistor. The cathode of D1 is reverse biased by the high Q1 collector voltage and also because its anode is held close to 0 volts by the 47K resistor connected to the collector terminal of Q2. It would take a very strong negative input pulse to forward bias D1 enough to reach the Q1 base terminal. The Q2 collector voltage is nearly 0 volts and therefore the D2 cathode has little to no reverse bias voltage via its RS. Thus, any small amplitude negative input pulse will cause D2 to become forward biased, reach the base of Q2 and drive Q2 OFF. Once Q2 switches off, in turn Q1 is toggled ON and its collector voltage goes low. The large reverse bias on D1 disappears. However, Q2 is now OFF and D2 will now be strongly reverse biased which will steer the next negative input pulse to the base of Q1. This is the basis of the circuit's negative edge flip-flop operation.

In another experiment, I changed the .001 C0G capacitors of the astable multivibrator to 470 pF. This gave a usable range of 22968 to 14832 Hertz (11484-7416 Hz at the Q1 and Q2 output) . Looking at the output of the flip-flop in the oscilloscope; at the higher frequency range, the flip-flop could not keep up and failed to divide by 2. I found experimentally that the time constant of each of the CS and RS components seemed to be the problem. When the CS capacitors were also decreased to 470 pF, the flip-flop worked properly.
As you increase the flip-flop operation frequency, speed up bypass capacitors might also be required across the 47K base resistors of Q1 and Q2 . A suggested starting value to try is 220 pF. Some builders also bypass the resistors in the RS steering circuit at higher frequencies, however, this is getting a little crazy. It is really important to look at the output waveform in the oscilloscope to ensure reasonable performance.

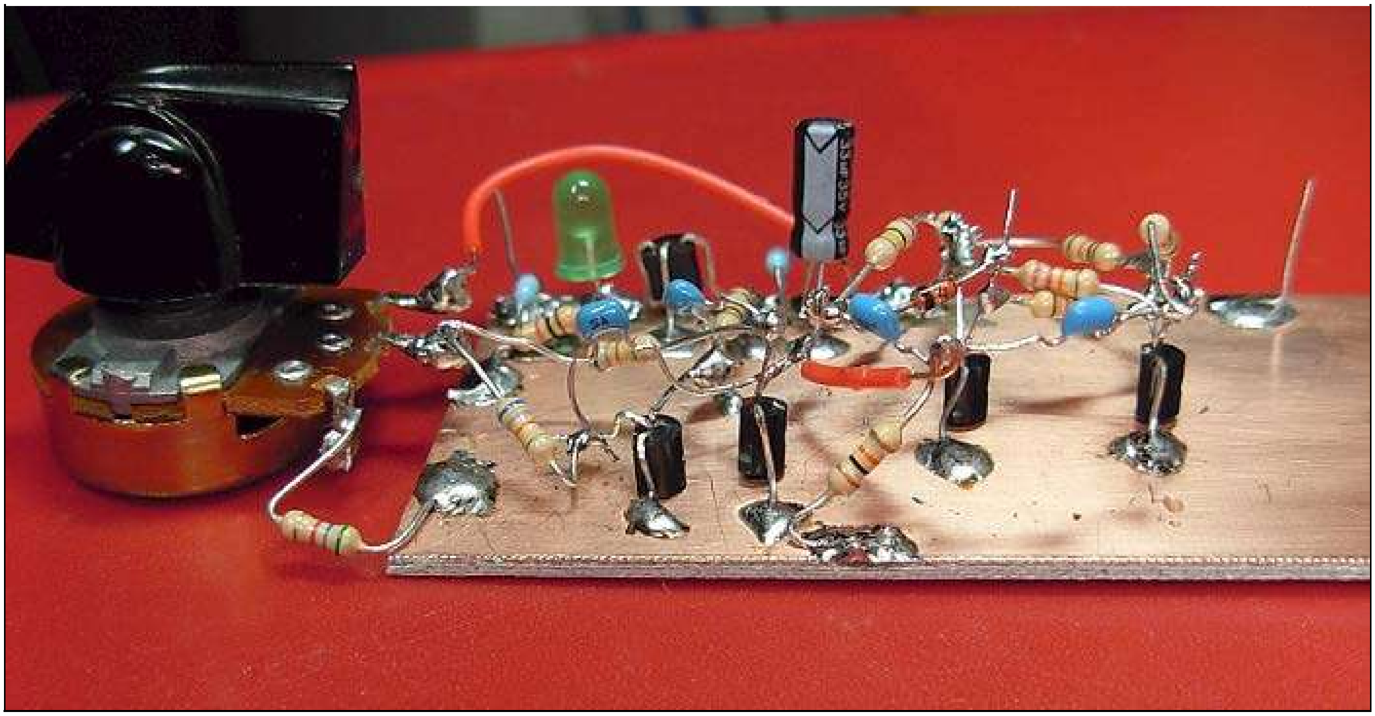

Shown above is the Figure 5 breadboard prototype.

Figure 6

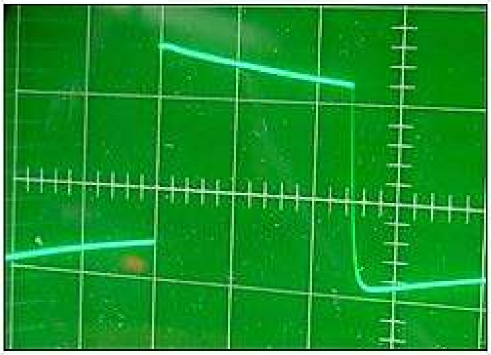

5 KHz output waveform of Q2

4. One KHz Digital and Analog Oscillator

A 1 KHz oscillator with 5 volt digital outputs 180 degrees apart and an analog output was sought. The frequency had to be near to, but not exactly 1000 Hertz. A major question to answer was how much low pass filtering is needed to remove the odd harmonics from digital circuits?

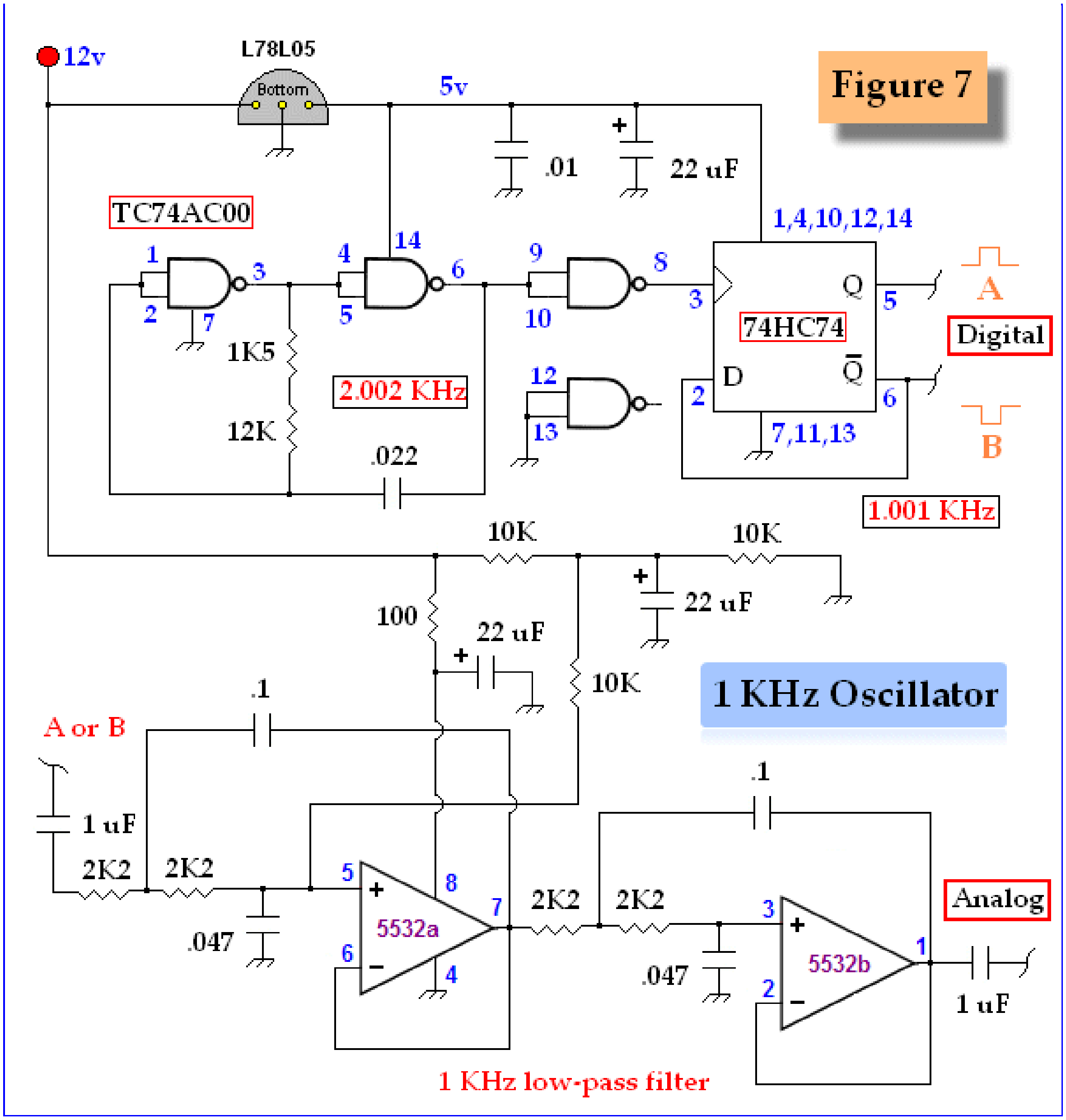
12v
L78L05
Bottom
5v
.01
+
22 uF
Figure 7
TC74AC00
1
3
4
14
6
9
8
2
7
5
10
3
Q
5
74HC74
A
Digital
1K5
2.002 KHz
12
D
Q̄
2
6
12K
.022
13
7,11,13
B
1,4,10,12,14
1.001 KHz
10K
10K
+
22 uF
100
22 uF
+
10K
1 KHz Oscillator
.1
A or B
1 uF
2K2
2K2
.047
5
+
8
7
5532a
6
−
4
.1
2K2
2K2
3
+
5532b
1
2
−
.047
Analog
1 uF
1 KHz low-pass filter

Figure 7 shows the complete schematic. NAND gates from a 74AC00 were wired as inverters and with the 13.5 K resistance and a 0.022 uF polyester capacitor, the frequency was 2002 Hertz. To improve the digital waveform and get the desired 2 outputs, a D flip-flop was used. The output frequency was 1001 Hertz. The digital part was completed!

For the analog filtering, active low-pass filters were tried, and in total 4 poles with a 1 KHz cuff off worked reasonably well. The filter uses the 5532 op-amp with common vales capacitors and resistors. Poly"something" caps were utilized.

Figure 8

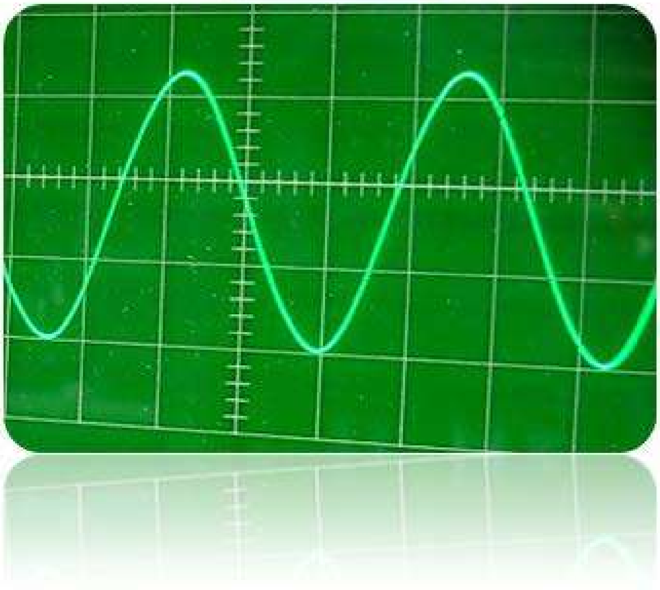

In Figure 8 is the output waveform of the low pass filter stages. A pretty nice sine wave was achieved and this oscillator could see duty for testing audio amplifiers. The scope was photographed at an angle to avoid the camera reflection and this distorts the sine waves a little.

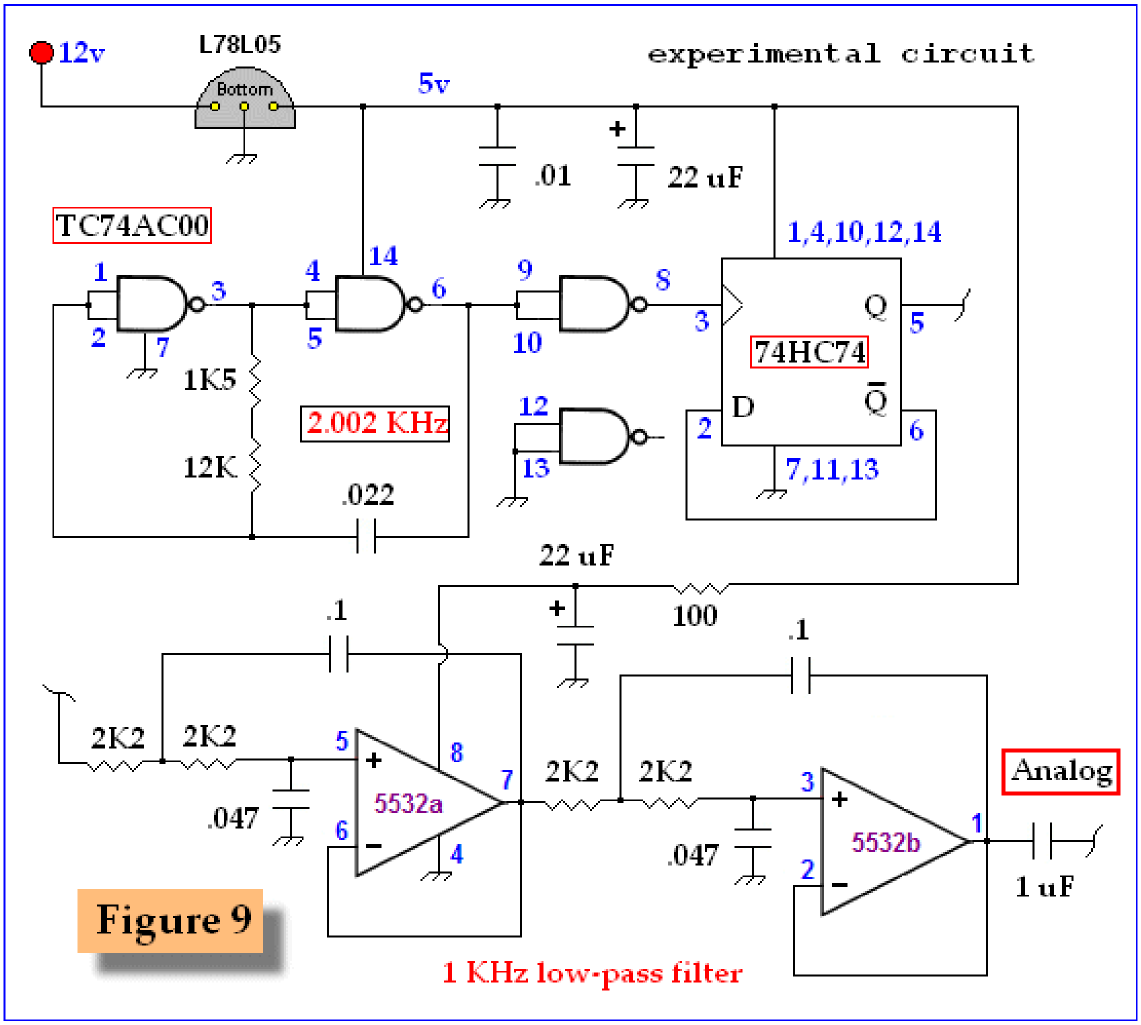
12v
L78L05
Bottom
5v
experimental circuit
.01
+
22 uF
TC74AC00
1,4,10,12,14
1
2
3
7
4
5
14
6
9
10
8
3
Q
5
74HC74
D
Q̄
2
6
7,11,13
1K5
12K
2.002 KHz
.022
12
13
22 uF
+
100
.1
.1
2K2
2K2
5
8
+
5532a
7
6
−
4
.047
2K2
2K2
3
+
5532b
1
2
−
.047
Analog
1 uF
1 KHz low-pass filter

Figure 9

Figure 9 depicts an experiment with the op-amp biasing. If the op-amp is run at 5 volts VCC, the bias requirement is 1/2 VCC or 2.5 volts. The DC voltage at the output of the D flip-flop was 2.55 volts. The 2K2 resistor was connected directly to this output and this eliminated the VCC/2 resistor bias network and a coupling capacitor.

Figure 10

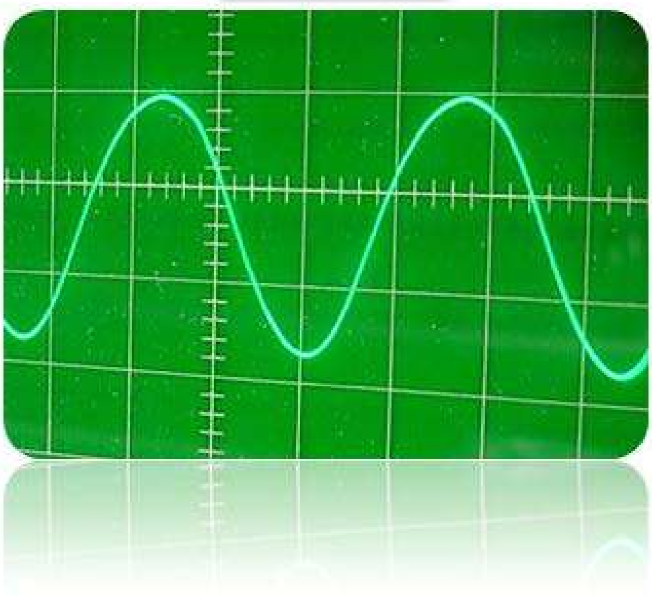

Figure 10 shows the output waveform of Figure 9. The AC waveform has harmonic distortion and thus the Figure 9 circuit will not be kept nor utilized.

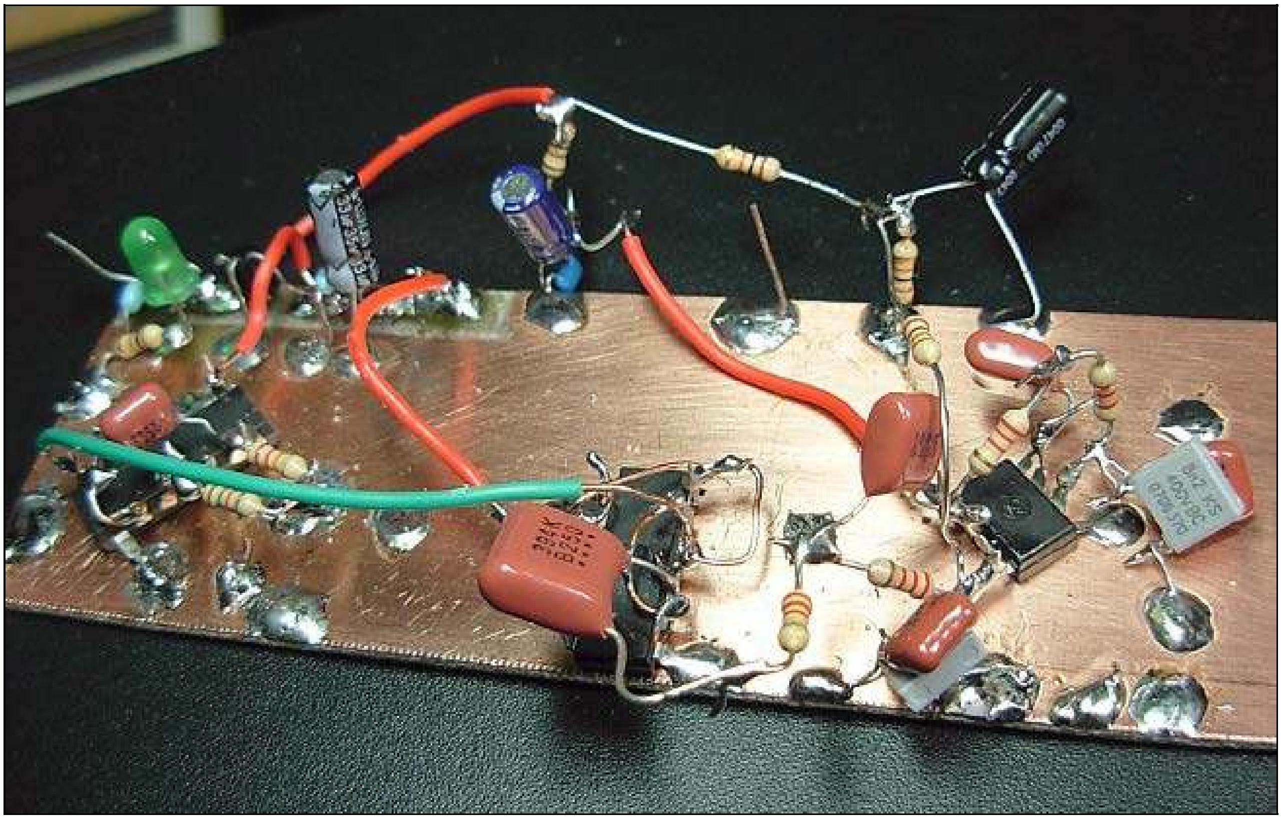

The Figure 7 breadboard. A 0.39 coupling capacitor (not 1 uF) was used between the D flip-flop and 5532a in this particular version. Unfortunately, no 0.047 uF caps were available for the low-pass filters and therefore a 0.039 plus a 0.0082 were placed in parallel for each of the .047 uF caps.

5. One KHz Low Distortion Signal Generator

Although I own a variable frequency wein bridge oscillator, it has been been set to 1 KHz for 2-3 years and is large and temperamental. It was decided to make a low distortion sine wave oscillator for just this one frequency. The circuit will be placed in a box along with another signal generator.

There are a number of ways to build signal generators using op amps. Countless example circuits may be found on the World Wide Web and some of them are really fantastic. Chose whichever method works best for you. Some might find my circuit to be overkill, but to each his own.

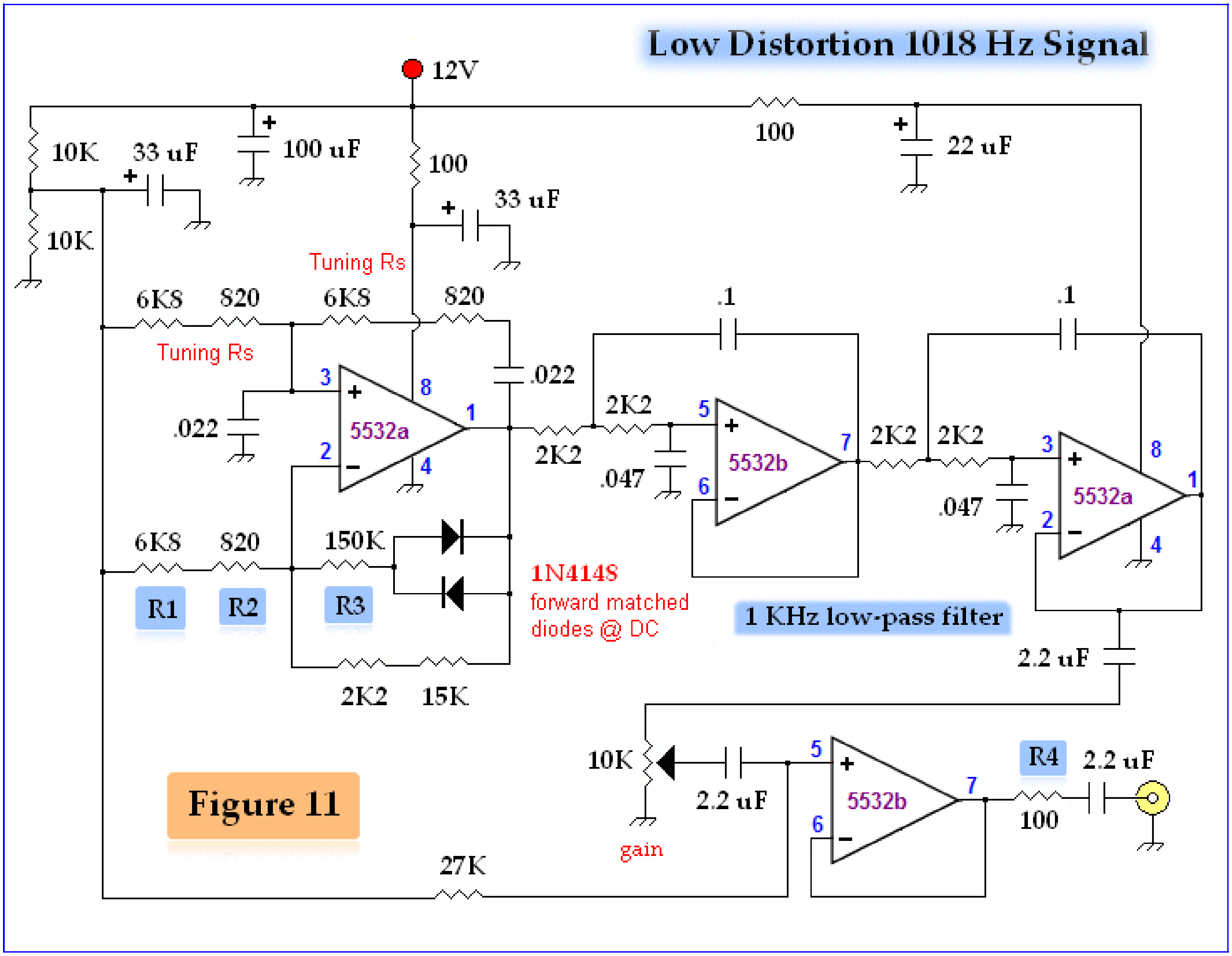
Low Distortion 1018 Hz Signal
12V
10K
33 uF
100 uF
100
100
22 uF
33 uF
10K
Tuning Rs
6K8
820
6K8
820
Tuning Rs
.022
.022
5532a
3
8
1
2
4
.1
.1
2K2
2K2
.047
5532b
5
6
7
2K2
2K2
.047
5532a
3
8
1
2
4
6K8
820
150K
R1
R2
R3
1N4148
forward matched
diodes @ DC
1 KHz low-pass filter
2.2 uF
2K2
15K
10K
2.2 uF
gain
5532b
5
6
7
R4
2.2 uF
100
27K
Figure 11

Figure 11 shows the entire circuit. The Wein bridge oscillator is from EMRFD and was designed by Wes, W7ZOI.

A 0.22 uF capacitor was chosen for the tuned circuits. Using resistors from my parts collection, 1018 Hz was the closest I could get to 1 KHz. This is the 6K8 + 820 ohms resistors labeled "tuning Rs". Other values were tried. For example, a single 6K8 gave 1142 Hz and a single 8K2 gave 939 Hz.

When the circuit was first built, I used a 10K on the VCC/2 bias point to pin 2 and a 22K feedback R from pin 1 back to pin 2. The sine wave had mild distortion. By experimentation, it was learned that the resistance from the VCC/2 bias point to pin 2 significantly affected the waveform purity. The 2K2+15K plus the R1 + R2 resistance values shown were determined by using a potentiometers rather than fixed resistors. Care was taken to adjust the feedback resistance from pin 1 back to pin 2 to keep away any overdrive distortion. I do not understand this, but even changing the 820 ohm R2 to 570 ohms, altered the sine wave purity.

The best looking sine wave came when the resistance from the VCC/2 bias point to pin 2 was the same as the tuning resistance; 6K8 plus 820 ohms. Later, the pin 1 to pin 2 feedback resistance was chosen for an unclipped; waveform with a reasonable output voltage using a potentiometer. The potentiometer was removed and measured at 17.1K, thus the 15K + 2K2 were soldered in. It was also discovered that by increasing R3 from 56K or 100K to 150K slightly improved the waveform.

The Figure 7 low-pass filter was connected to the main oscillator as shown. The final op-amp stage was used as a buffer between the low-pass filter and the gain control. R4 is used to set whatever output impedance you choose. Practically speaking, it could be any value between 47 and 620 ohms. Many AF oscillators have an output impedance of 600 ohms and 620 is the nearest E24 standard value. For my project, a 100 ohm

R4 was chosen. Output peak to peak voltage is 0.0 to 4.84v continuously.

Figure 12

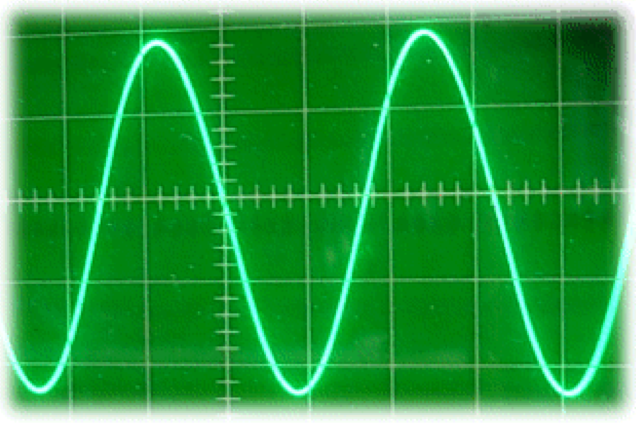

Here is the raw output of the basic Wien bridge oscillator. It is hard to photograph well, but it is stellar to say the least.

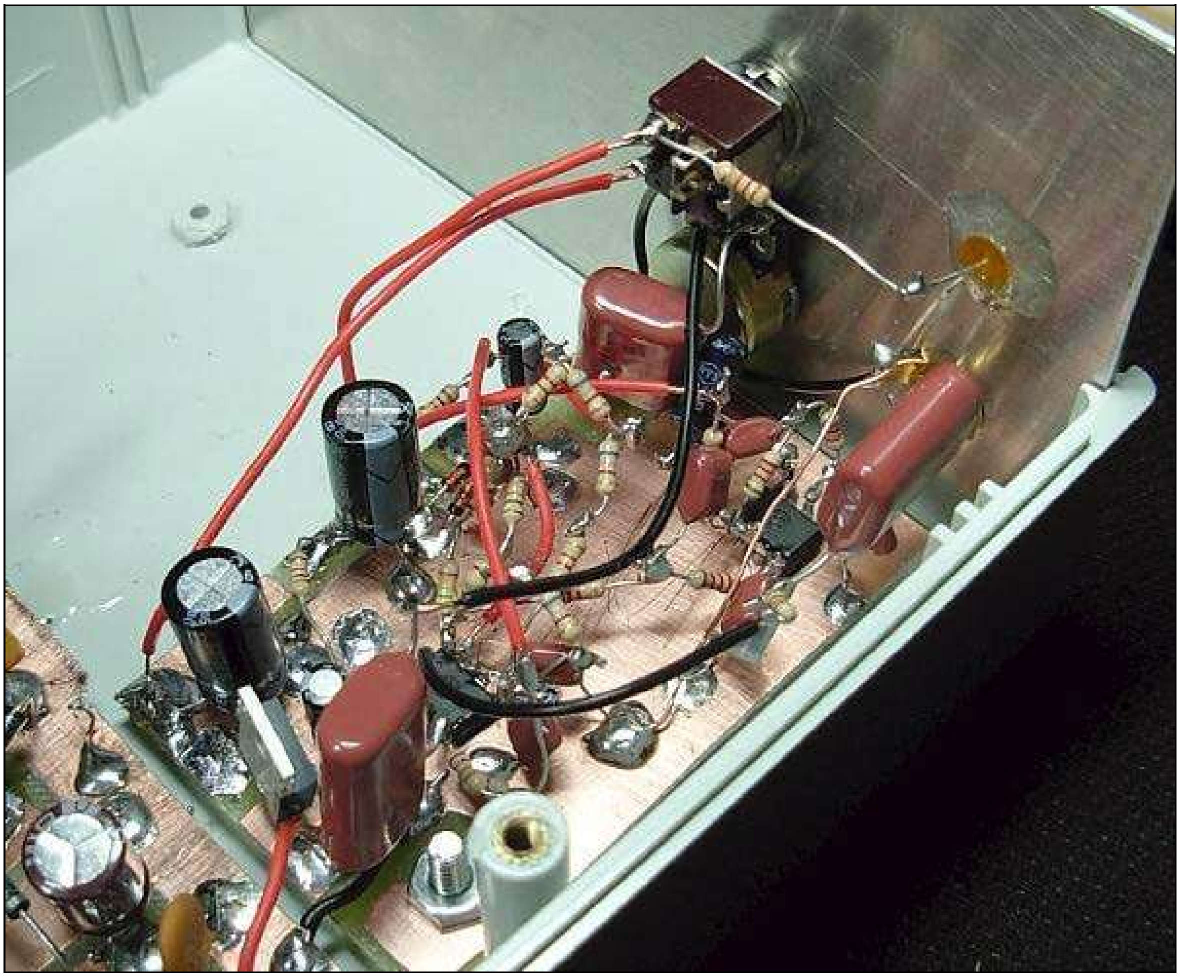

The Figure 11 breadboard mounted in a plastic Hammond chassis. The voltage regulator seen to the left of the bottom polyester cap is a 7812. This project has its own regulated power supply. [Other view](#) of Fig 11. Two 10 Megohm standoff resistors were used to help support all the resistors soldered to pins 1, 2 and 3 of the main oscillator.

1018 HERTZ
OFF
GAIN

A front view photo of the AF oscillator. A separate 7 plus 14 MHz oscillator circuit and controls will be placed on the right hand side of this box. The orange power ON indicator LED was epoxy glued into the chassis hole. Putting circuits in cabinets is one of the most expensive aspects of homebrew construction. One must be ever vigilant for bargain chassis boxes and hardware to keep costs down. Techniques such as gluing in the LED rather than purchasing a separate holder and recycling knobs and switches are also practiced for cost containment.

Figure 13

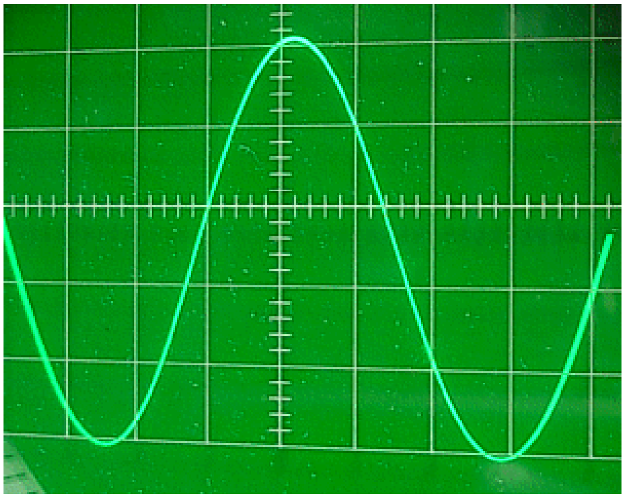

The output of Figure 11 is shown in Figure 13. This is the best sine wave seen ever on my scope. I looked at it closely and there is no change in line thickness or symmetry anywhere. It inverts with no change on the scope. Testing audio amplifiers will now be much more fun.

6. LM386 Power Experiments

The LM386 is an IC audio amp that has been used in thousands of hobbyist projects over the past 2 decades. By adding a capacitor +/- a resistor between pins 1 and 8, this device's internal gain can be changed from x20 to up to x200.

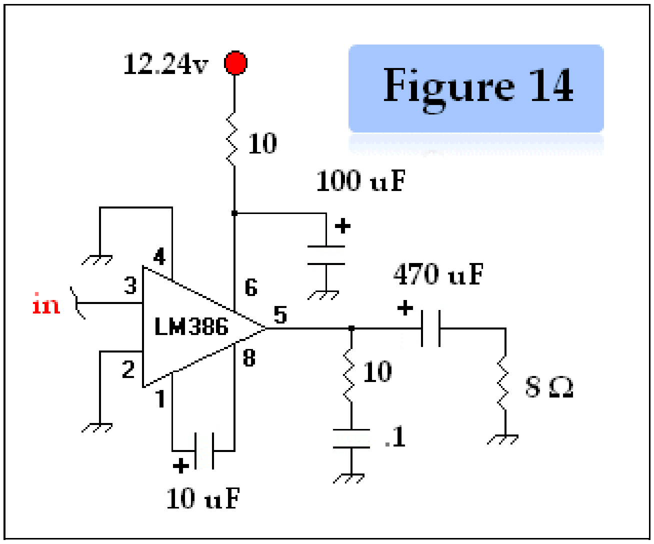
12.24v
Figure 14
10
100 uF
+
4
3
in
6
5
LM386
8
2
1
470 uF
+
10
8 Ω
.1
+
10 uF

Test circuit schematic in Figure 14.

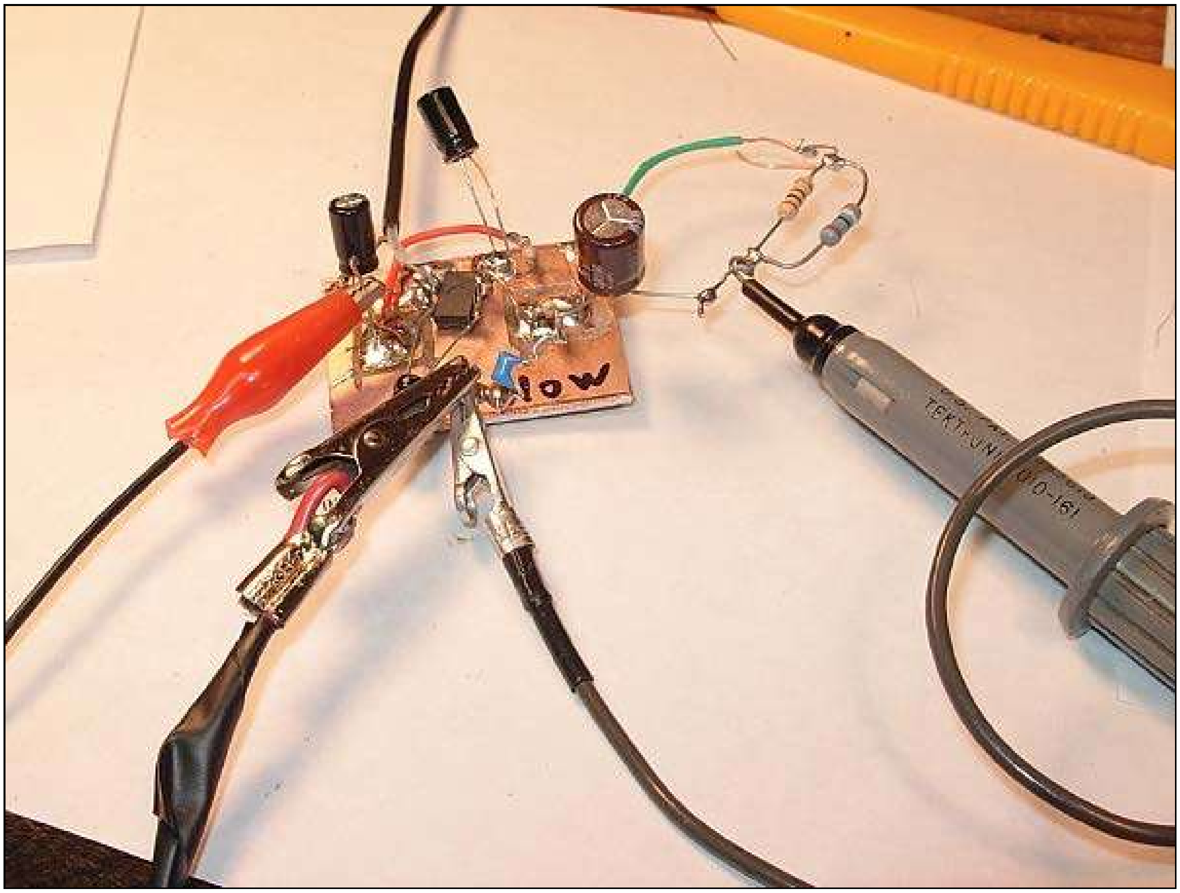
low
TEKTRONIX 010-161

The experiment breadboard is shown above. A very standard configuration. The amplifier drove an 8 ohm, 1 watt resistive load.

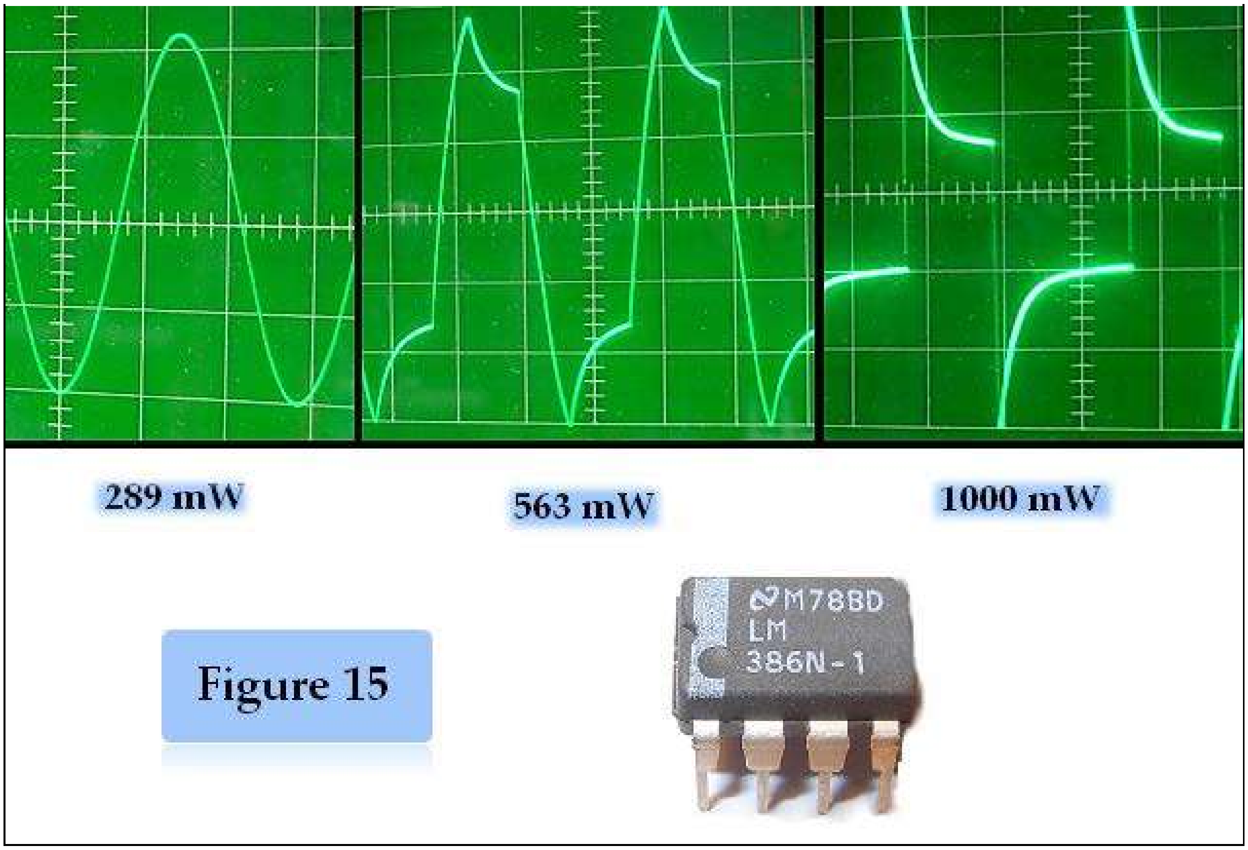

Figure 15

Over the years, I have noticed some kit sellers and project authors claiming that their LM386 based AF stages gave 1 or occasionally even 2 watts of output power into an 8 ohm speaker. This was confirmed on the bench. This device will output 1 watt into 8 ohms at 1018 Hertz with little problem. However, this is clearly 1 watt of square wave distortion.

The quiescent current of the LM386 was around 7 mA. The signal generator gain was increased until the first signs of distortion appeared. The gain was then backed off a little so a pure sine wave was observed in the oscilloscope. The current was ~ 155 mA and the measured power was 289 mW. Please refer to Figure 15 for the 289 mW sine wave. This was the clean signal power of the LM386 on my bench. The output waveforms at 563 mW and 1 watt are also shown. Extreme harmonic distortion occurred above 300 mW. This device will draw 240 mA or more when driven and clipping hard. It is not my intention to malign the LM386. It is a useful part, albeit a little dated. Its AF gain capability versus size is something to behold. Many builders have moved to the TDA7052 audio amplifier IC, or like myself, build their own low noise audio power amps.

ON
12.2 VDC
1018 HERTZ
OFF
GAIN

The 12.24 volt DC supply and the 1018 Hz AF audio oscillator used in these experiments.

RF — Test and Measurement

Time Domain Output from a Diode Ring Mixer

22 Dec 2009, w7zoi

Some folks wonder about the output that they should see on their oscilloscope when looking at the output from a diode ring mixer. There is no set, pat answer. The output can change dramatically as levels, frequencies, and even terminations are changed. This complication is illustrated here with a few screen shots, taken with a Rigol DS1052E 50 MHz bandwidth digital storage oscilloscope. The experiments started with the following pile of modules. Your collection will probably differ.

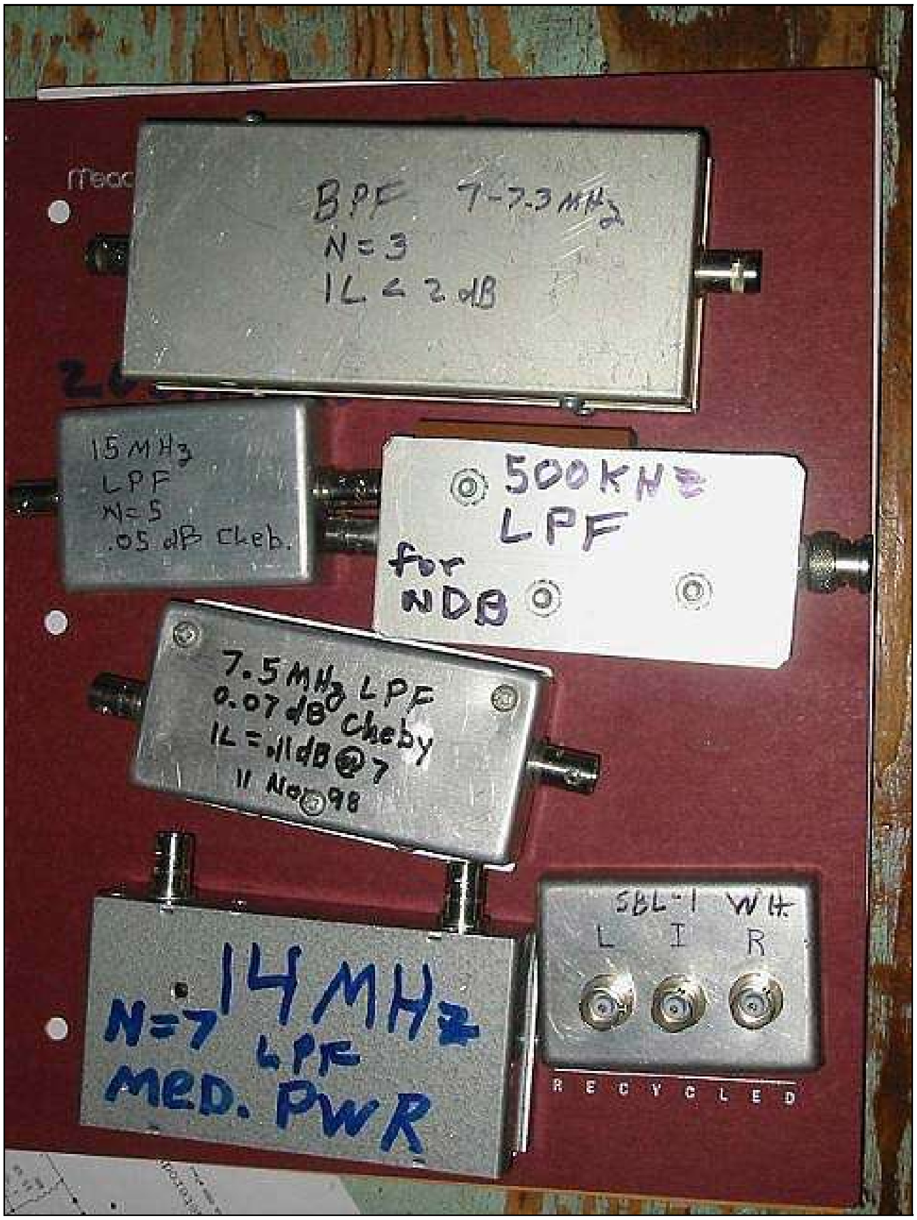
BPF 7-7.3 MHz
N=3
IL < 2 dB
15 MHz
LPF
N=5
.05 dB Cheb.
500 KHz
LPF
for
NDB
7.5 MHz LPF
0.07 dB Cheby
IL = .11 dB @ 7
11 Nov 98
SBL-1 WH
L I R
N=7 14 MHz
LPF
MED. PWR
RECYCLED

Figure 1. Some available experimental modules.

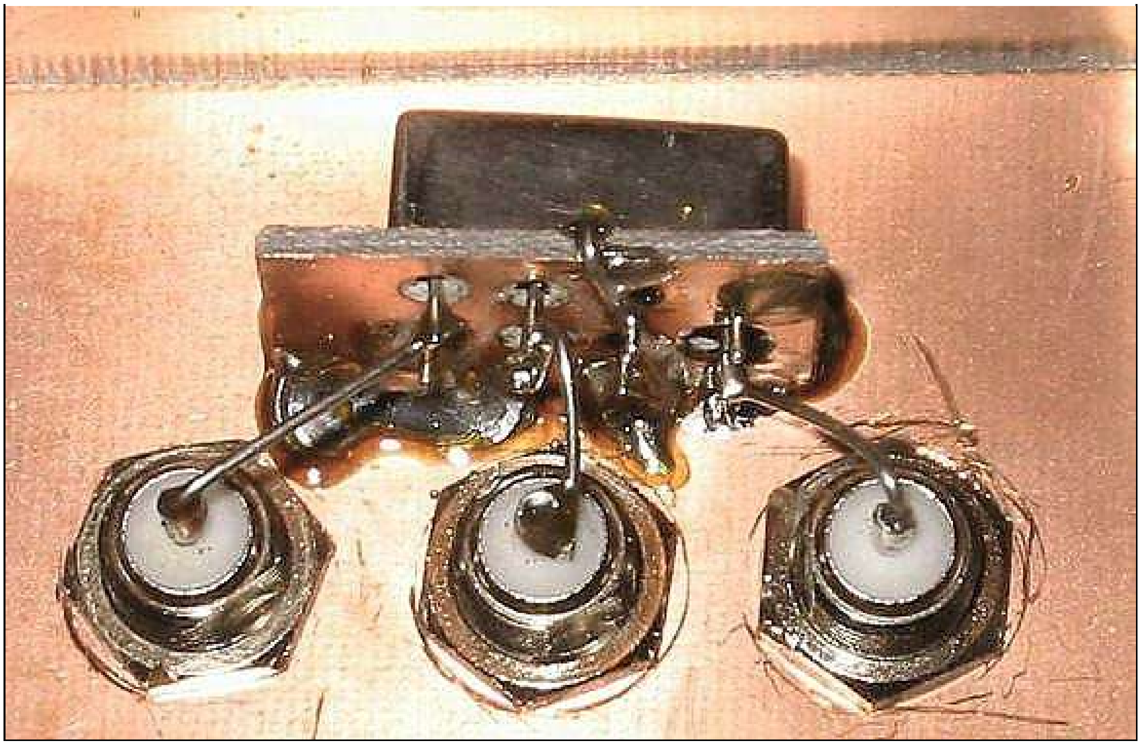

Figure 2. The inside of the module containing a Mini-Circuits SBL-1 diode ring mixer. This is a standard part that is essentially generic.

The first experiment was to set up a pair of 10 MHz signal sources. One was from a homebrew generator, shown below.

Figure 3

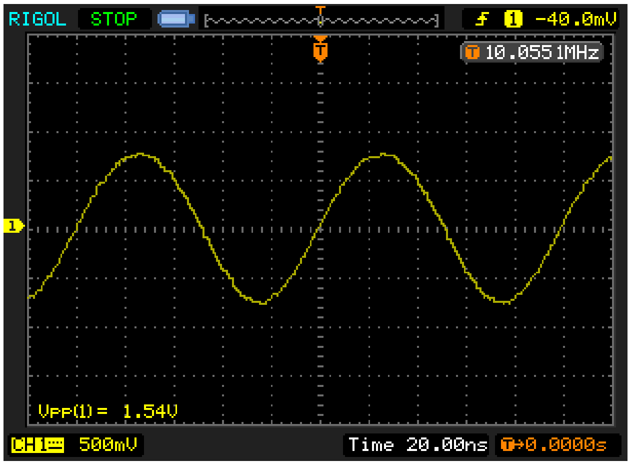
RIGOL
STOP
1 -40.0mV
T 10.0551MHz
1
Vpp(1)= 1.54V
CH1 500mV
Time 20.00ns
T→0.0000s

Figure 3. The 10 MHz signal from a homebrew signal source. The 1.54 volt peak to peak signal is applied to a 50 Ohm terminator at the oscilloscope. The delivered power is then +7.7 dBm.

This signal was filtered with a 14 MHz low pass circuit. This caused the amplitude to drop by 0.2 dB. The source was then attached to the LO (local oscillator) port of the SBL-1 mixer.

Figure 4

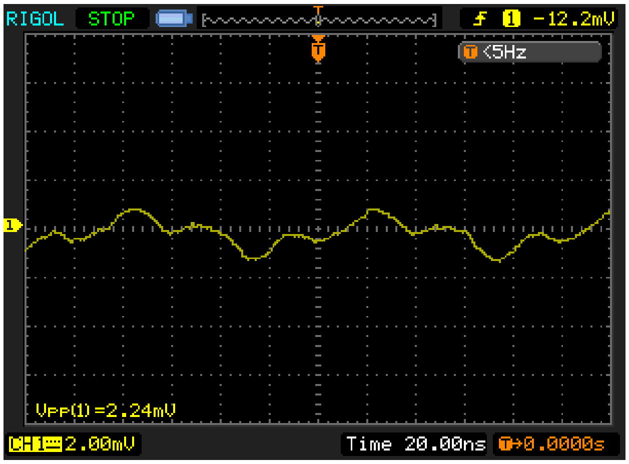

Figure 4. The IF output from the mixer when there is nothing attached to the RF port. Note the scope sensitivity of 2 mV/div.

Next, we attached a 50 Ohm terminator to the R mixer port.

Figure 5

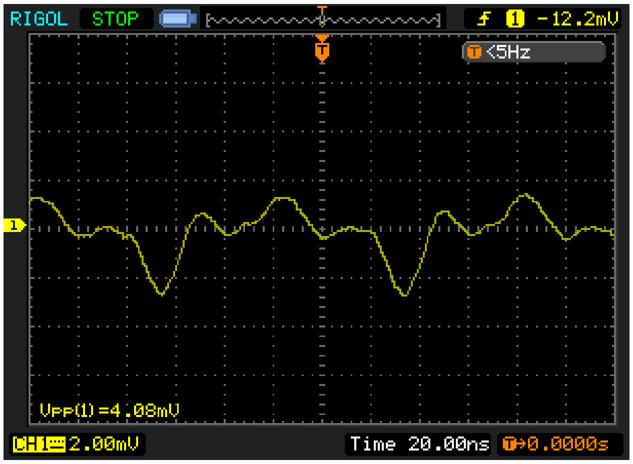

Figure 5. The IF output with LO drive, but without an R signal. But the R port is now terminated. This waveform, when compared with Fig 4, shows just how sensitive the mixer can be to termination.

In the next experiment, a -20 dBm signal was applied to the R port. The frequency was very close to the 10 MHz LO that is still present.

Figure 6

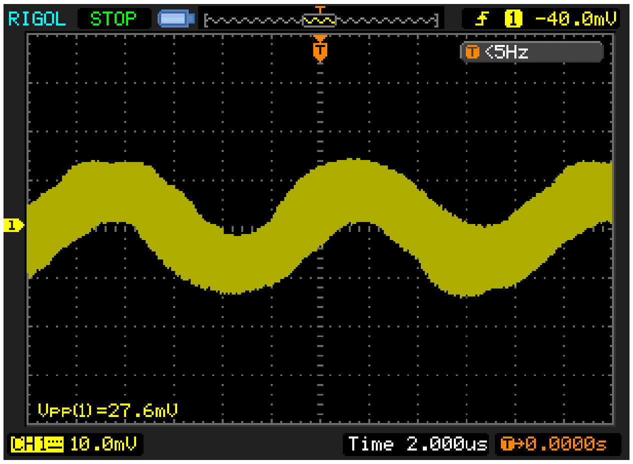
RIGOL
STOP
1 -40.0mV
T <5Hz
VPP(1)=27.6mV
CH1 10.0mV
Time 2.000us
T→0.0000s

Fig 6. There are two dominant signals from the mixer. One is a low frequency at 100 kHz. But this is accompanied by a high frequency of about 20 MHz. These two outputs, a sum and difference frequency, are expected from any mixer.

A filter can isolate the two dominant outputs. This is shown below where a 500 kHz low pass filter is inserted in the line between the mixer and the oscilloscope. A 6 dB pad is between the mixer and the filter, for direct insertion would upset the termination of the mixer.

Figure 7

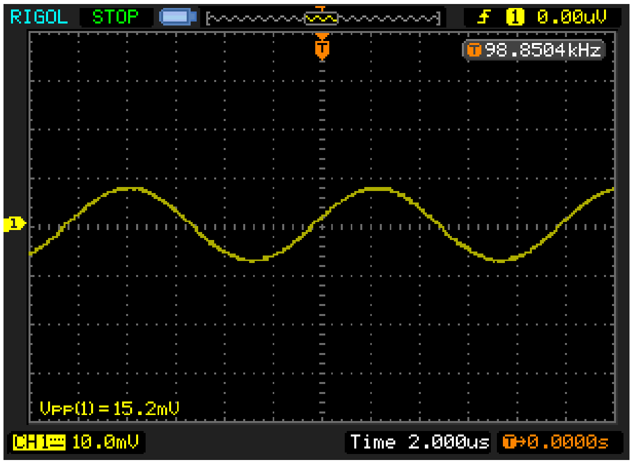
RIGOL
STOP
1 0.00uV
T 98.8504kHz
VPP(1)=15.2mV
CH1 10.0mV
Time 2.000us
T→0.0000s

Figure 7. The output of about 100 kHz after a low pass filter is inserted in the mixer output.

The next experiments emulate a SSB transmitter. We start with a signal at 11.06 MHz with strength -20 dBm. (This is a common IF used in homebrew SSB transceivers such as the BITX-20.) This is applied to the mixer R port. The L port is driven with a +7 dBm signal at 3.19 MHz. The LO signal is low pass filtered to attenuate harmonics, a measure that is probably not necessary, but the filters were there. The IF output is shown below.

Figure 8

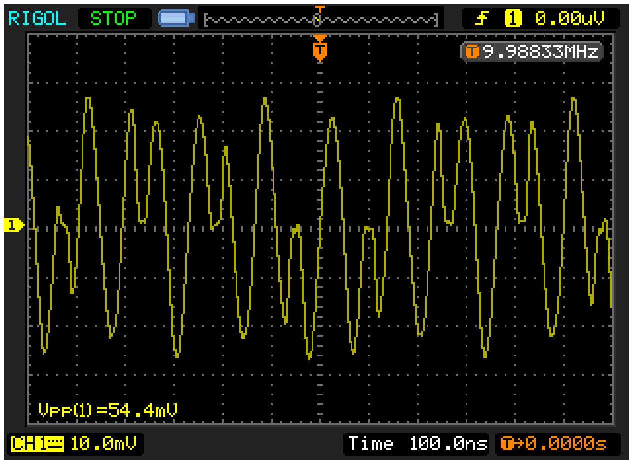

Figure 8. Time domain output of a SBL-1 set up as the “transmit” mixer in a SSB rig. This is far from the “perfect sine wave” that some folks tell us we should observe. This waveform contains many different frequency components. The counter output should not be interpreted to have any meaning. (I should have turned it off.)

The signal of Fig 8 can also be viewed with a spectrum analyzer. This is shown below. This measurement was taken with the August 1998 QST Spectrum Analyzer and not the FFT routine in the DSO. The Rigol scope has a nice display for an analyzer.

Figure 9

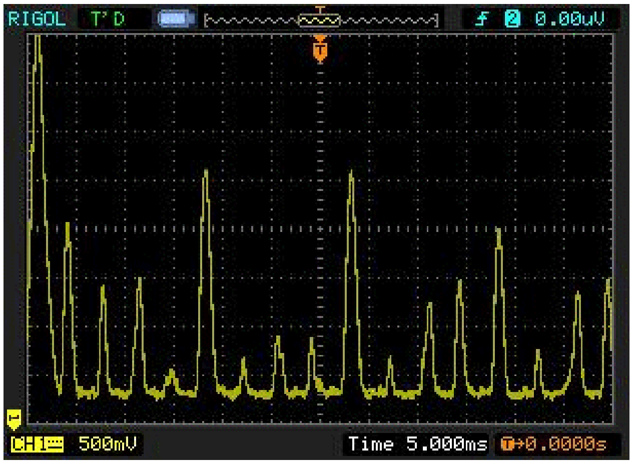

Figure 9. Spectrum of the signal shown in Fig 8. The largest signal on screen is that at the left, which is the spectrum analyzer zero spur. This is a spurious output that is typical of most SA systems. The desired signal at 11.06+3.19=14.25 MHz is just to the right of center. But the image is also present at the different frequency of 7.87 MHz at about 3 major divisions from the left edge.

Alas, I didn’t find a 14 MHz bandpass filter in the junk box. Such a filter would have allowed selection of the dominant 14 MHz component while attenuating all the rest of the junk shown. The many other signals are the result of harmonic mixing. That is, we observe IF outputs at N x FLO +/- M x FRF where N and M are integers. Some of these spurious outputs can be quite strong with diode ring mixers. They are best avoided with high frequency LO signals. In this case, a LO at 14.25+11.06=25.31 MHz would produce a much cleaner output spectrum. It is much easier to obtain LO stability with an oscillator built at 3.19 MHz.

Bottom Line

It is not reasonable to have a well defined, predictable time domain (i.e., normal oscilloscope) output from a mixer. The exact results depend upon too many variables. A spectrum analyzer can be used to garner much more information.

Low Noise Crystal Oscillators

Introduction

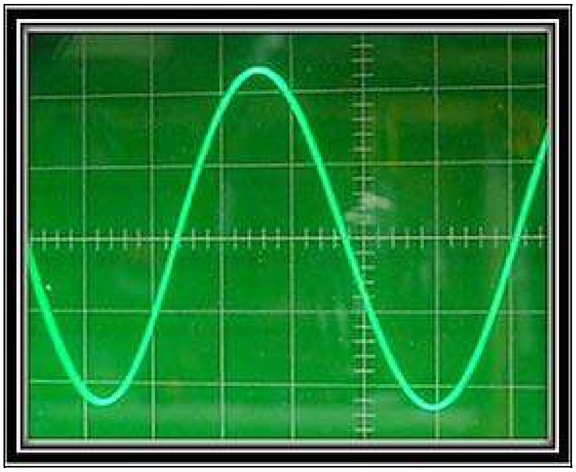

Some experiments were conducted to build a low noise crystal oscillator with 50 ohm output at 7 and 14 MHz for the test bench.

Some readers might wonder why build such an oscillator? Although a variable RF signal generator is an important bench tool, it is also nice to have a fixed frequency signal source on your favorite HAM band. This RF source can be connected to other 50 ohm modules such as band-pass filters, diode ring mixers or feedback amplifiers to conduct experiments at a whim. The project goal was a low noise oscillator with 2 outputs so it could drive a divide by 2 flip-flop for 40 meter band digital mixer work, or fundamental frequency use on the 20 and 40 meter bands.

Presented are some experiments carried out to realize this goal. Only some of the better experiments and circuits are shown. An additional circuit was added Jan 31, 2010.

First Steps

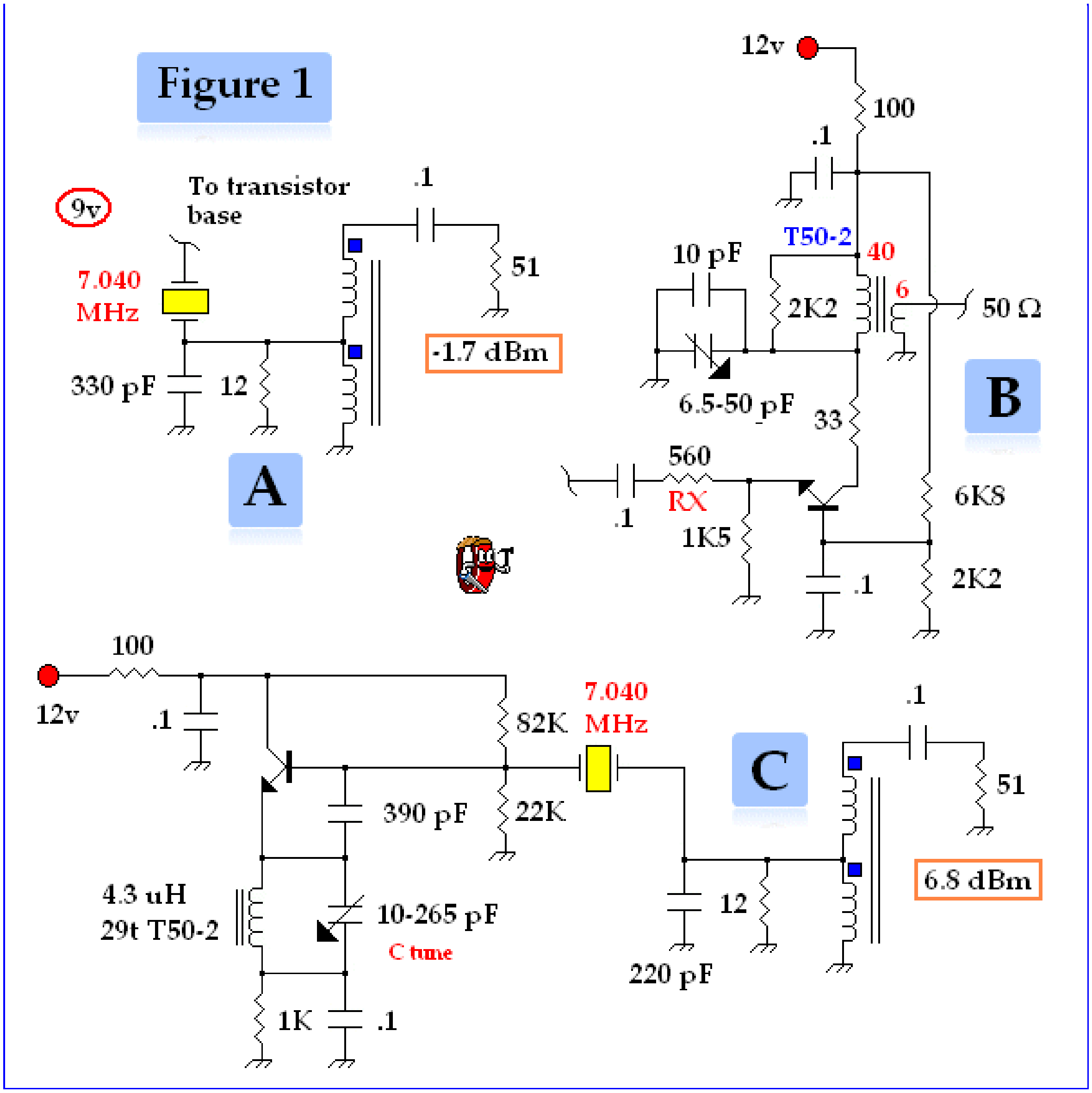
Figure 1
9v
To transistor base
.1
51
7.040 MHz
-1.7 dBm
330 pF
12
A
12v
100
.1
T50-2
40
10 pF
2K2
6
50 Ω
6.5-50 pF
33
B
560
RX
.1
1K5
6K8
.1
2K2
100
12v
.1
82K
7.040 MHz
.1
C
51
390 pF
22K
6.8 dBm
4.3 uH
29t T50-2
10-265 pF
C tune
12
220 pF
1K
.1

The 7.040 MHz crystal used is an AT cut, HC6-U holder part made locally by West Crystal. I reviewed the information concerning crystal oscillators in Chapter 4.5 of EMRFD and then started melting solder. The oscillator output was extracted as described in EMRFD Figure 4.24. The output is low distortion, low impedance and low gain. Like in most experiments, I built it, measured the DC voltages and then looked at the AC voltages in the oscilloscope.

In order to measure the output, a 51 ohm load was transformer coupled as shown in Figure 1a. I am uncertain if this was a good method for power measurement, however, it allowed comparison of the experimental circuits. As shown, the output with a 9 volt regulated power supply is low; -1.7 dBm. Still, a circuit like EMRFD Figure 4.24 can be used with a variety of crystal frequencies and has great utility.

Figure 2

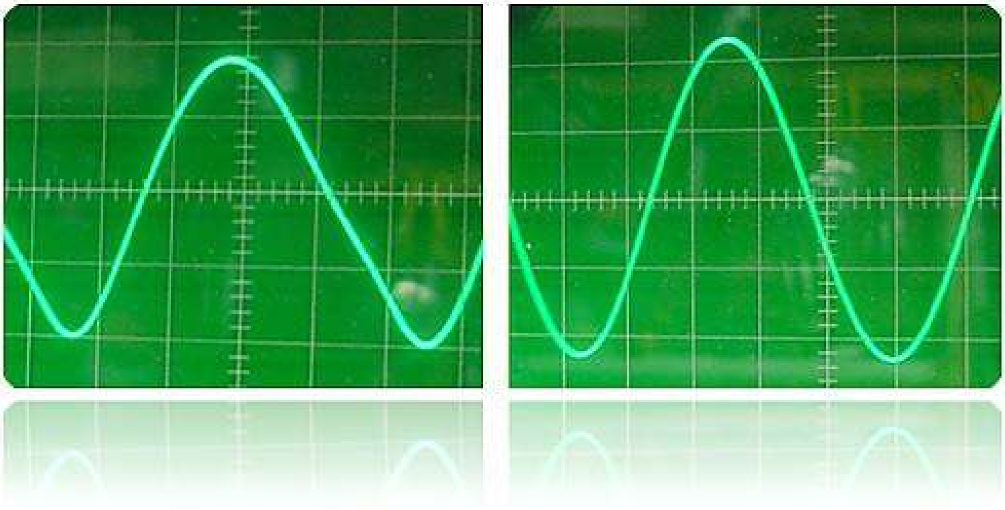

Signal from BJT emitter　Signal from between cap and crystal

Figure 2 shows two output waveforms taken from my crystal oscillator. When a signal is taken from the emitter of the main oscillator transistor (what we typically do) harmonic distortion occurs as shown in the above left. Actually, the distorted waveform photo above left looks better than most do. Typically, they look like this. Many builders will just place a low pass filter on such an oscillator's output and be very satisfied with the harmonic content in their signal. Certainly this is a good, common and practical way to go. However, for some builders, the experimenter's journey is what counts. That is, the fun and learning occurs during designing/building/testing and not just operating home built gear.

When the output is taken from between the shunt capacitor and the crystal per Figure 1A, a much cleaner sine wave is available. The photo above right tells this story. The focus of all of the experiments on this web page is boosting this lower distortion, lower phase noise signal into something useful. To increase the base oscillator output voltage, the VCC was raised to 12 VDC and the BJT emitter circuit was tuned per Figure 1C. Resident on my work bench are a few potentiometers and a 10-254 pF air variable capacitor with short attached leads. These parts are inserted into test circuits to allow tweaking of R or C as desired. Once the desired tweaking is performed, the potentiometer or variable capacitor is removed and measured. The closest fixed value R or C is then substituted as appropriate. In this experiment, the highest output voltage occurred when the variable capacitor measured 181.7 pF. Thus in my version, 33 pF and 150 pF capacitors were placed in parallel and are shown later in Figure 3.

The Figure 1A output transformer circuit was again used and the power output was 6.8 dBm. Being tuned to 1 crystal frequency is the biggest drawback of the 1C circuit. Tuning a crystal oscillator as described earlier is easy to do however.

The next task was to design and build a buffer/amplifier. To match the low impedance of the Figure 1C crystal oscillator, a common base amp was built. The circuit was morphed over time, however, the initial design is shown as Figure 1B. It was interesting to note that a series resistor (RX) is required to keep the waveform pure. Any RX value less than 470 ohms compromised the sine wave purity. The 560 ohm R shown was perfect, however, as expected, attenuated the oscillator signal. In order to get a decent output voltage, the common base amp had to be run at 5 mA or greater current and ultimately collector tuning was added to try and realize an output voltage greater than 4 dBm.

Through experimentation I learned that adjusting Fig 1B's tuning, emitter current and RX value all could distort the output of the main oscillator at certain values or settings. Running high current also invites parasitic oscillations and soon it was realized that common base was perhaps not the best choice (at least for me) as another separate amplifier stage would be required to get a decent output voltage with or without an attenuator pad.. After trying a number of different buffer-amplifiers including a 50 ohm feedback amp, I chose a favorite circuit which I know has excellent gain plus back to front isolation and would not distort the oscillator waveform; a lightly coupled JFET amp.

---

7 MHz Circuit

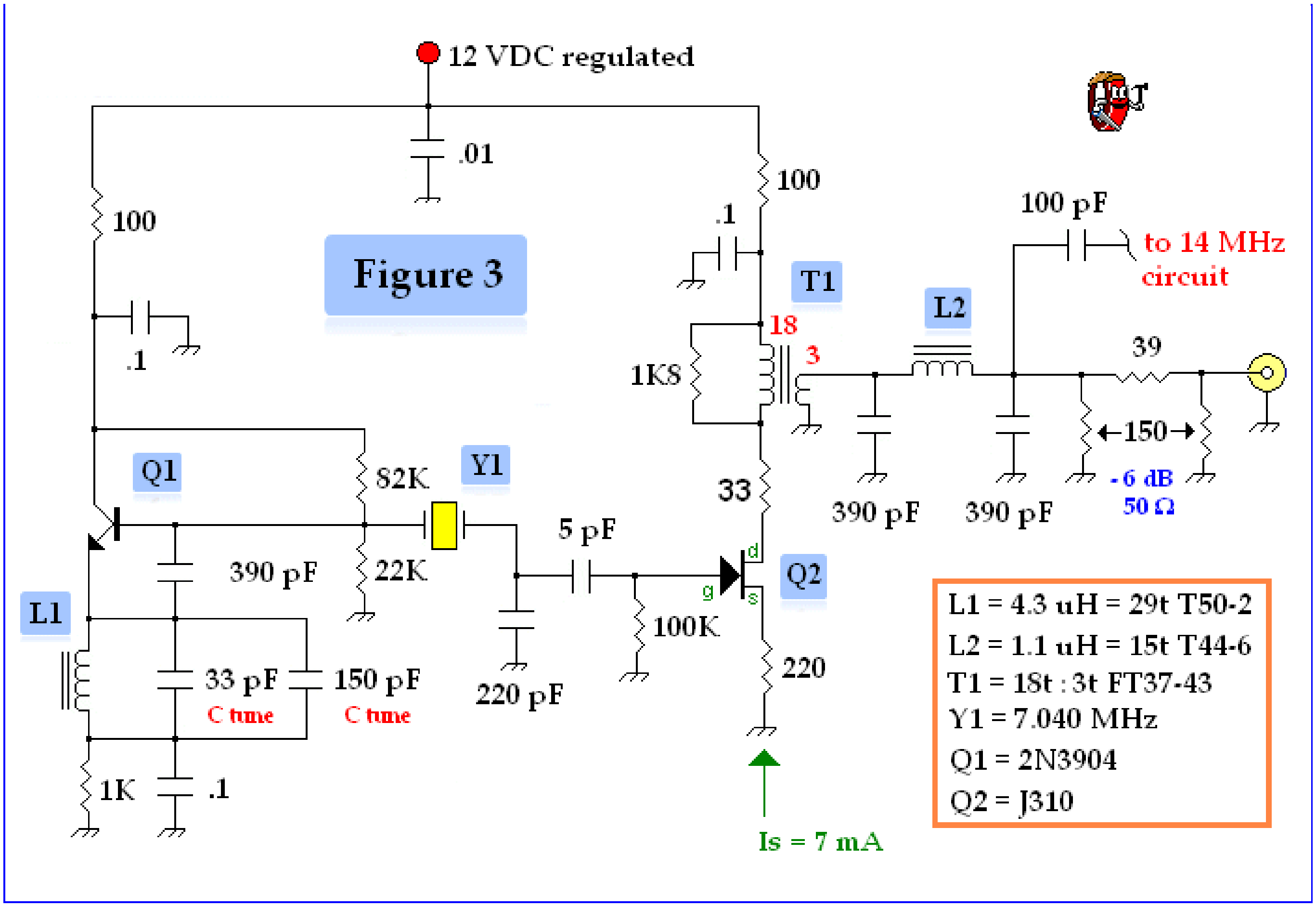
12 VDC regulated
.01
100
100
Figure 3
.1
.1
T1
L2
100 pF
to 14 MHz
circuit
18
3
1K8
39
←150→
- 6 dB
50 Ω
Q1
82K
Y1
33
390 pF
390 pF
5 pF
390 pF
22K
d
Q2
g
s
L1
100K
220
33 pF
C tune
150 pF
C tune
220 pF
1K
.1
Is = 7 mA
L1 = 4.3 uH = 29t T50-2
L2 = 1.1 uH = 15t T44-6
T1 = 18t : 3t FT37-43
Y1 = 7.040 MHz
Q1 = 2N3904
Q2 = J310

In Figure 3 is the complete schematic of the 7 MHz portion of the low distortion crystal oscillator. The output was 0.52 dBm. A 6 dB 50 ohm pad ensures a well-buffered 50 ohm termination. This aids in calculating gain or loss in circuits it drives. You could easily decrease this to a -3 or -4 dB pad. This reflects the wisdom handed down by our mentors who encourage building RF stages in 50 ohm impedance blocks. A Q2 source bypass capacitor was not placed as it increased harmonic distortion in the output signal. The output is filtered with a simple pi filter. The 100 pF coupling capacitor connect this circuit to Figure 6; the 14 MHz circuit.

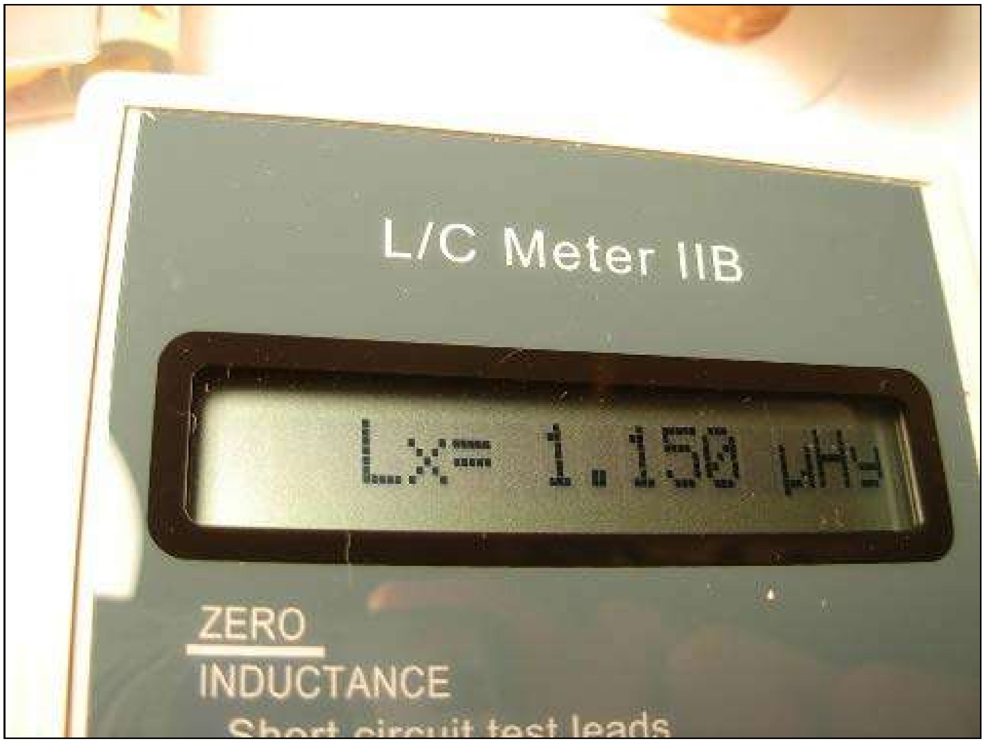
L/C Meter IIB
Lx= 1.150 µH
ZERO
INDUCTANCE
Short circuit test leads

L2 was wound and measured. It is desirable (but not absolutely necessary) to perform measurement on powdered iron toroids to compensate for variations in wire spacing and toroid permeability. I used a T44-6 core; use whatever appropriate powdered iron toroid you want.

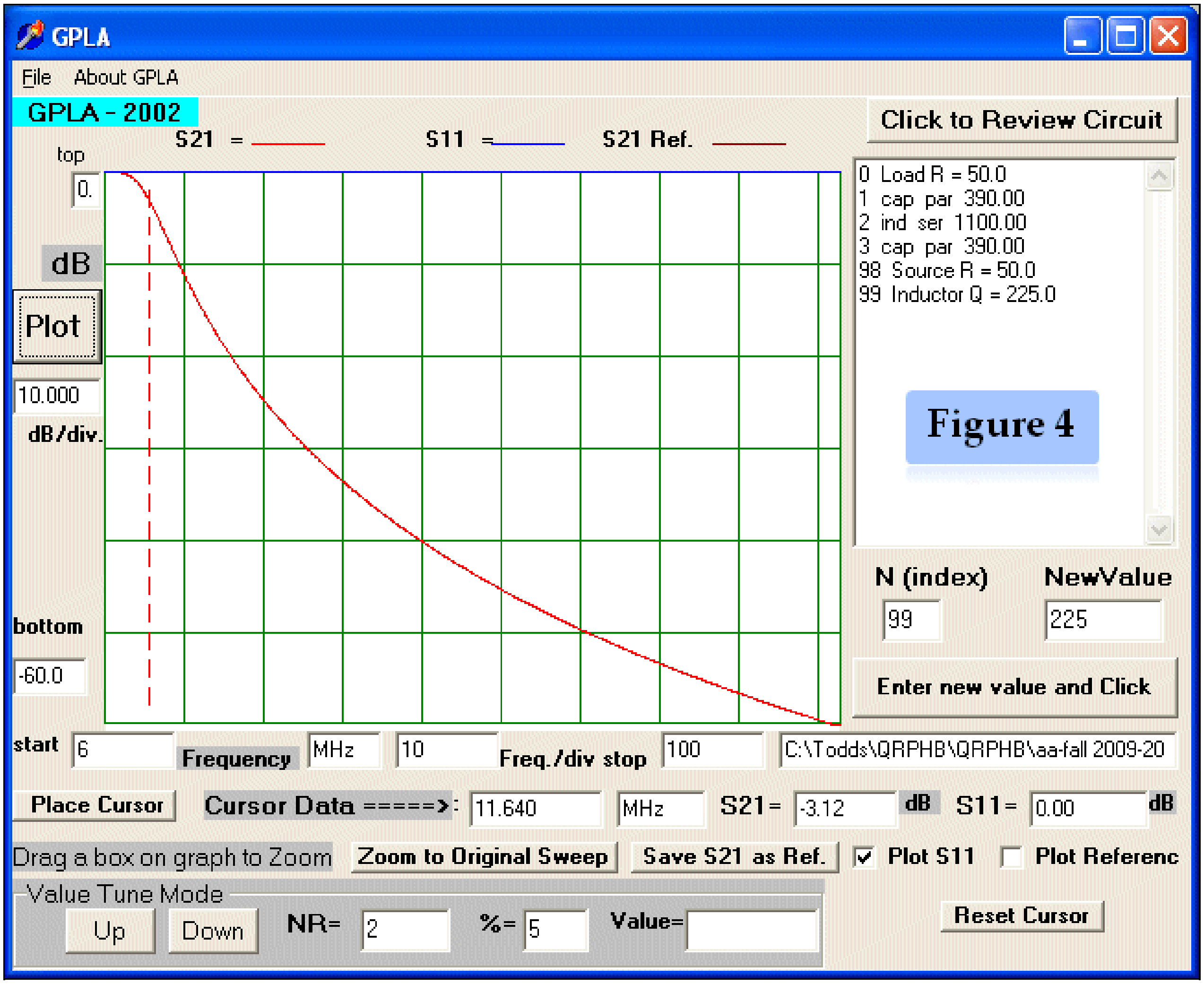
GPLA
File
About GPLA
GPLA - 2002
Click to Review Circuit
S21 =
S11 =
S21 Ref.
dB
Plot
Frequency
Place Cursor
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Enter new value and Click
Value Tune Mode
Up
Down
Reset Cursor

0 Load R = 50.0
1 cap par 390.00
2 ind ser 1100.00
3 cap par 390.00
98 Source R = 50.0
99 Inductor Q = 225.0

Figure 4

top 0.

10.000 dB/div.

bottom -60.0

start 6 MHz

10 Freq./div

stop 100

C:\Todds\QRPHB\QRPHB\aa-fall 2009-20

Cursor Data =====>: 11.640 MHz

S21= -3.12 dB

S11= 0.00 dB

☑ Plot S11

☐ Plot Referenc

N (index) 99

NewValue 225

NR= 2

%= 5

Value=

A GPLA analysis of the Pi low-pass filter is shown in Figure 4. The basic circuit was designed with PI Filter Designer on this page and tweaked in GPLA. You may wish to omit this filter or perhaps, design a better one yourself.

Figure 5

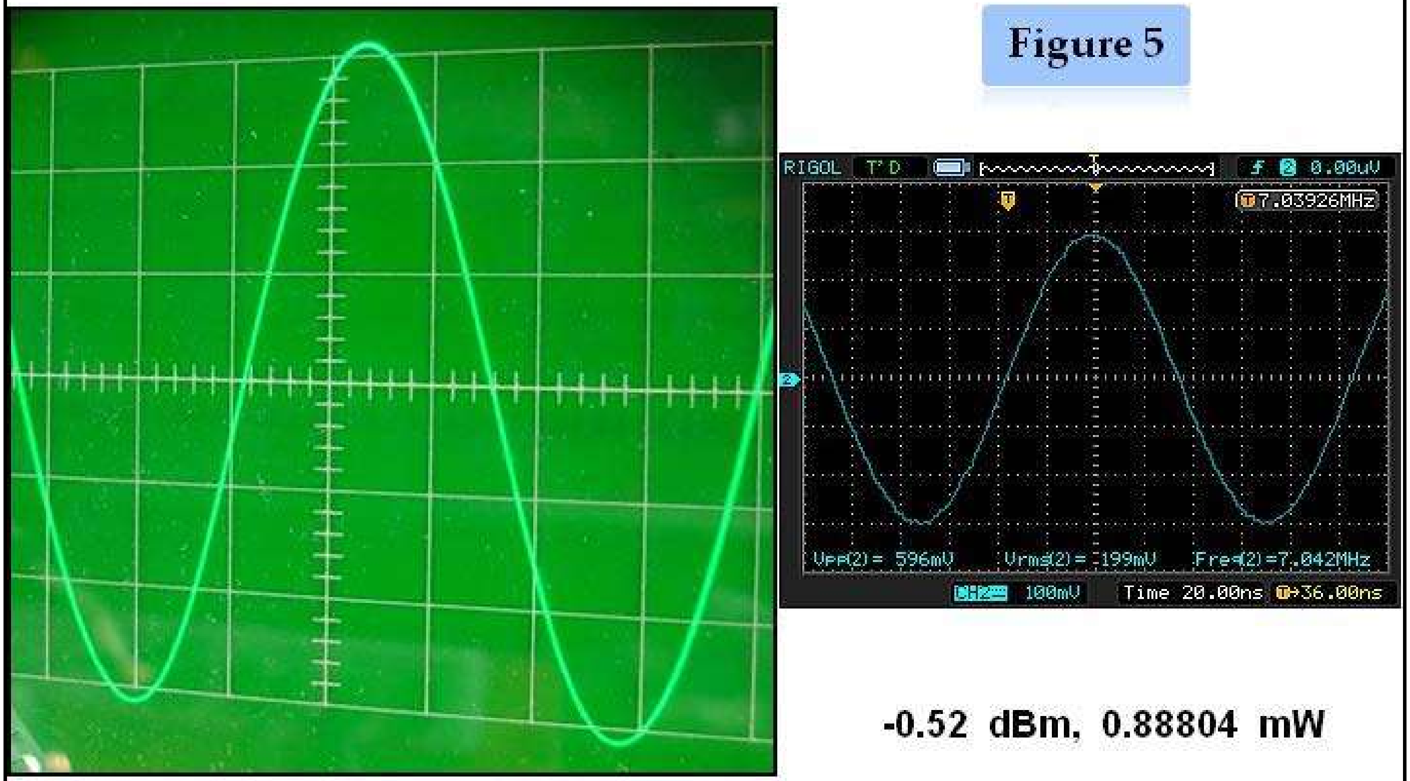

-0.52 dBm, 0.88804 mW

The 7 MHz output waveform on the Tektronix (left) and the Rigol oscilloscopes. On the Tek scope, the output power was 0.50 dBm and on the Rigol scope it was 0.52 dBm. The Rigol is an amazing oscilloscope, but only has 256 horizontal lines of resolution, therefore cannot replicate the stellar and beautiful waveform tracings of the ancient Tek scope. In reality, probably, no other modern scope can. I have received many positive comments about the old Tektronix oscilloscope waveforms. It is important to mention, that Rigol waveform viewing is not bad, just very different. The visual display is incredibly accurate and its triggering options, bandwidth, sampling rate and waveform display tools are fantastic. фантастический !

During these experiments, the 40 year old (plus) Tektronix scope was distorting frequencies greater than 10 MHz and breaking into oscillations. After 3 major repairs in 2009, the scope replaced with a Rigol DS1052E. Signal viewing will certainly different; that is for sure. The decision to move from a cathode ray oscilloscope (CRO) to a digital storage scope (DSO) was not taken lightly.

DSO versus CRO - some comments from the workbench

Choosing a CRO versus a DSO is an individualized process. It is your decision alone. Questions to ask yourself may include: What are my needs? What is my budget? Do I have weight and/or space constraints? Carefully weigh the advantages and disadvantages of each.

Proponents of CROs state that these scopes cannot generate artifacts, nor distort the signal. This of course is true as long as the scope bandwidth is adequate. Further, some people feel that aliasing or artifact generated in DSOs due to undersampling (taking too few samples of a waveform) is unacceptable. They may even feel that DSOs are not precision measurement instruments as a result. Limited horizontal screen resolution in DSOs is also a bugaboo for some experimenters and provides further evidence of DSO inferiority in the minds of these folks. These concerns are indeed valid; however, black and white thinking is a little out of fashion in a world more containing shades-of-gray.

The DSO takes a series of samples and stores them in memory. When sufficient samples are present, they are assembled and displayed. The sampling rate of a DSO is variable and depends on the time base setting used. Modern DSOs, like the Rigol, Tektronix Oscilloscopes and Agilent Oscilloscopes have better sampling rates and larger memories than their predecessors and hence aliasing is less of a problem than before; although in some measurement situations, undersampling can occur. One must always well consider and interpret whatever you are measuring and if you use a DSO, always keep undersampling in mind. When first using a DSO, you are on the bottom of a learning curve, however, with attentiveness and practice, one can learn to look for and possibly mitigate undersampling should it occur. In certain cases, a CRO will be superior to a DSO. In my discussions with others about Rigol signal viewing , only 1 significant "aliasing" problem has been noted by a builder when he tested a balanced modulator. The display did not give the expected result (was not filled in as expected) and a CRO was pulled out and the problem was verified. The builder knew there was a problem and could understand why it occurred. This builder also wrote that this was not so much a problem, as a reality of using a DSO.

Some techniques I have gleaned from the Internet about detecting aliasing may include the following:

Vary the time base over several ranges. Events occurring near the time base should be reproducible and if they are not, undersampling might be occurring.

As possible, use a single sweep and dot display. The 'dots' will indicate just where the scope took each sample. If the dots are far apart relative to the waveform timing, aliasing is a possibility.

Some techniques to minimize aliasing:

Choose linear interpolation when using math functions.

Use bandwidth limiting in low level measurements (The Rigol seems to automatically use B/W limiting in these situations).

Use trace averaging for low level measurements as possible.

The Rigol weighs 50 times less than my old scope and fits on a small shelf in my small workspace. As a hobbyist, it meets my needs and budget plus has some very cool features. Undersampling is considered and in some cases, such as low level measurement, an analog scope might be a better choice. Happily, a CRO is available to me if I really stress out over it. Signal viewing was taken for granted with my old scope. In some ways, this DSO has prompted me to dig deeper; to become more vigilant and thoughtful about signal measurement and display. If you have a spectrum analyzer, and use a DSO for signal viewing, the ability to perform slow sweeps while maintaining a perfect display is quite enjoyable. Again, please decide the CRO versus DSO issue for yourself (or maybe get a hybrid). The DSO is not a perfect solution to every signal viewing situation, but their constraints are quite livable considering their numerous modern features.

DSO's: “They are not your father’s oscilloscope”; that is certain!

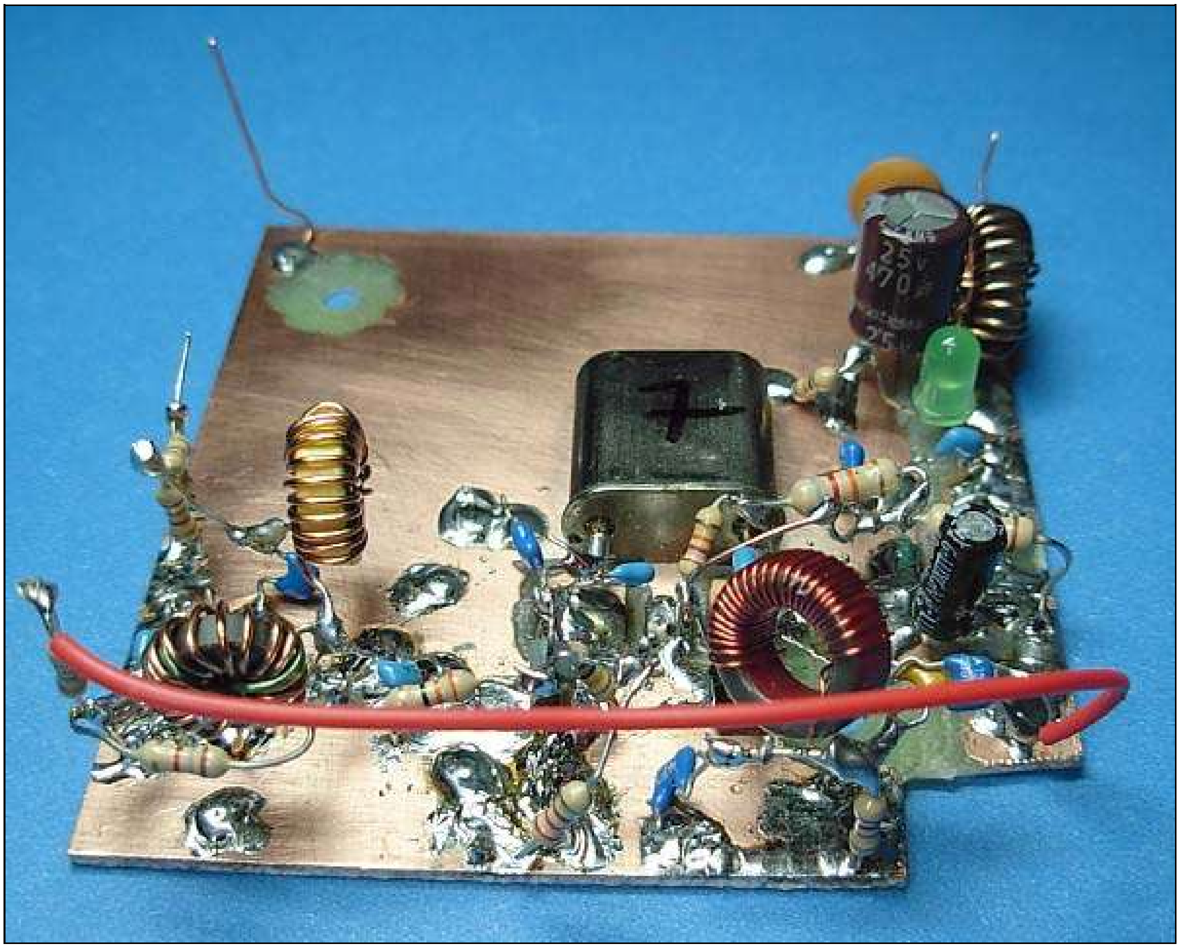

The Figure 3 breadboard. AWG 24 to 26 gauge wire was used in the various inductors and transformers to better secure or anchor these parts.

---

14 MHz Circuit

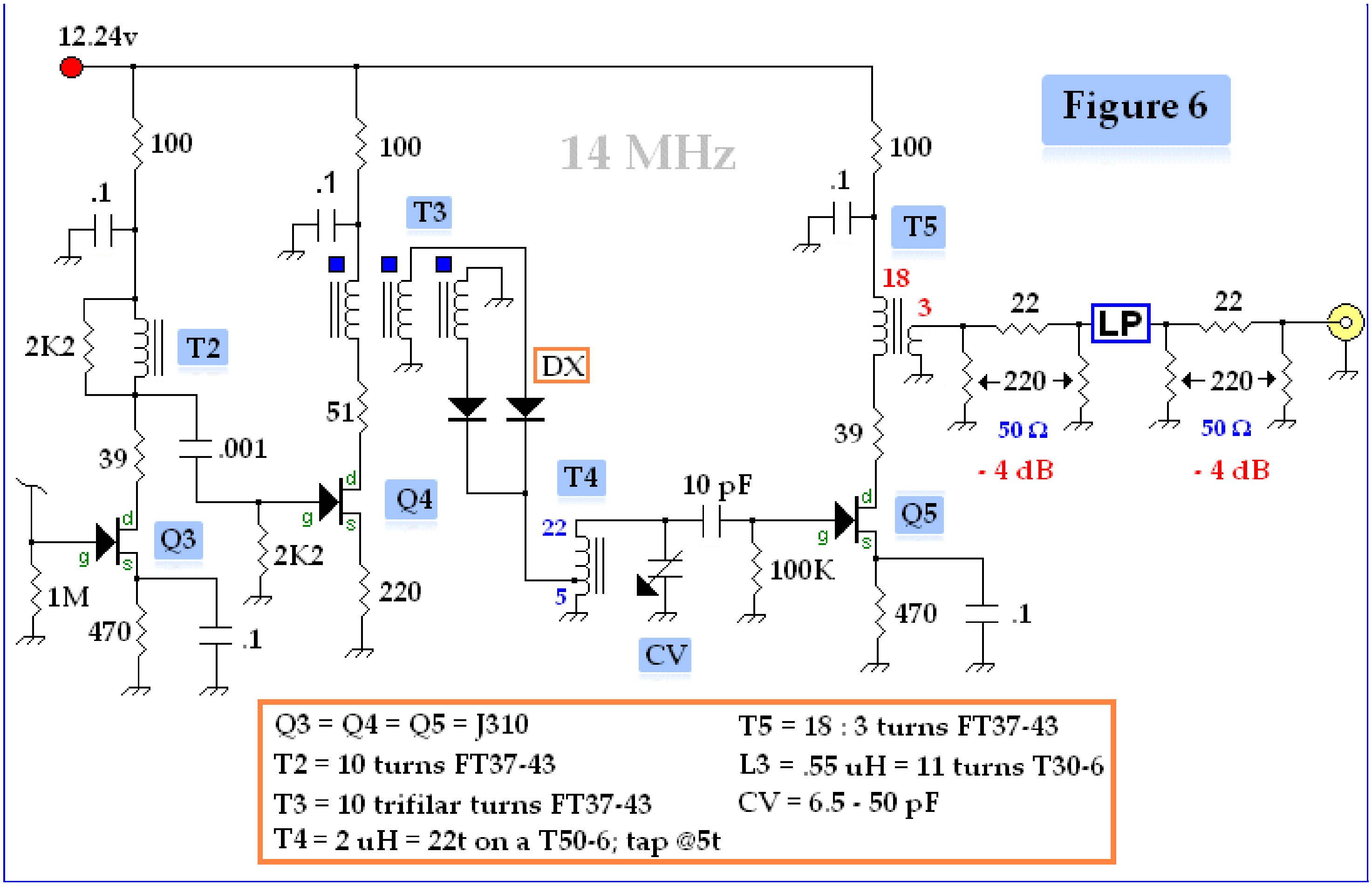
Figure 6
14 MHz
12.24v
100
.1
2K2
T2
39
.001
Q3
d
g
s
1M
470
.1
100
.1
T3
51
Q4
2K2
220
DX
T4
22
5
CV
10 pF
100K
Q5
39
470
.1
100
.1
T5
18
3
22
LP
22
← 220 →
← 220 →
50 Ω
50 Ω
- 4 dB
- 4 dB

Q3 = Q4 = Q5 = J310
T2 = 10 turns FT37-43
T3 = 10 trifilar turns FT37-43
T4 = 2 uH = 22t on a T50-6; tap @5t
T5 = 18 : 3 turns FT37-43
L3 = .55 uH = 11 turns T30-6
CV = 6.5 - 50 pF

Figure 6 is the final schematic for the 14 MHz circuit as developed on the workbench. Your design might look very different than mine. In the first version of this circuit, there were only 2 JFET amplifiers and the resultant output voltage was too low (even without a 50 ohm attenuation pad). To compensate, I ran the source current of the JFETs above 15 mA, placed a source bypass capacitor on Q4 and also used a 1000 pF capacitor to couple the input to Figure 2. Some fairly bad harmonic distortion was measured at the output and it seemed crazy to run so much current. Therefore it was decided to run a third JFET amplifier and use only modest current in the trio of JFET amplifiers.

CV tunes very sharply and required some care when peaking the output voltage.

Figure 7

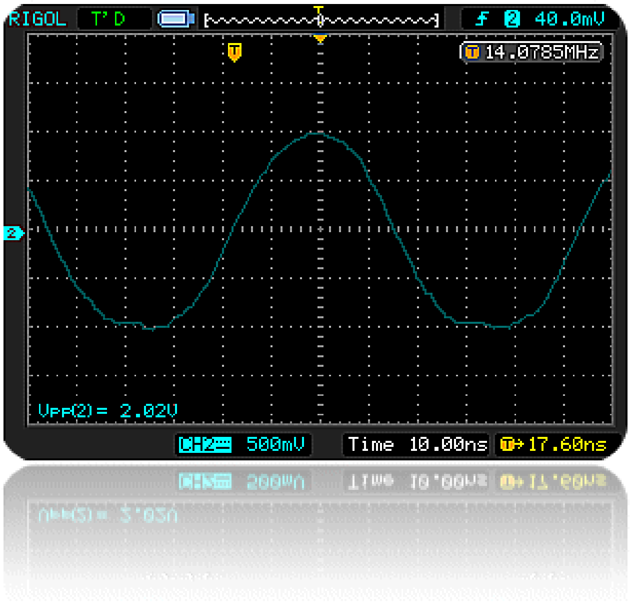
RIGOL
T'D
2 40.0mV
T 14.0785MHz
2
Vpp(2)= 2.02V
CH2 500mV
Time 10.00ns
T→17.60ns

Figure 7 is a screen capture of the 14 MHz circuit output measured using a sensitive 50 ohm terminated oscilloscope. This was with no low-pass filtering. The output is distorted. Presumably this happened in the diode frequency doubler. This is not a low-distortion oscillator.

15.3 MHz Low-pass Filter

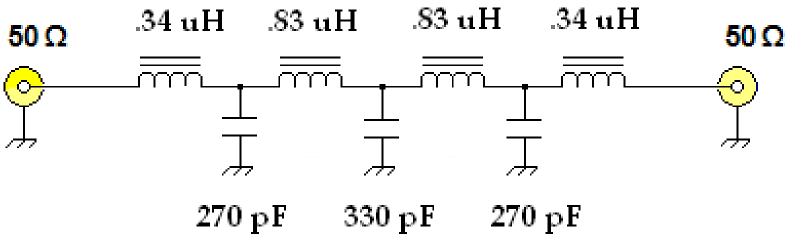
50 Ω
.34 uH
.83 uH
.83 uH
.34 uH
50 Ω
270 pF
330 pF
270 pF

A N = 7 Chebyshev low-pass filter was inserted in Figure 6 at point LP. I checked with a spectrum analyzer and the 2nd harmonic was down 38 dB. There were no other measurable harmonics after that.

Figure 7b

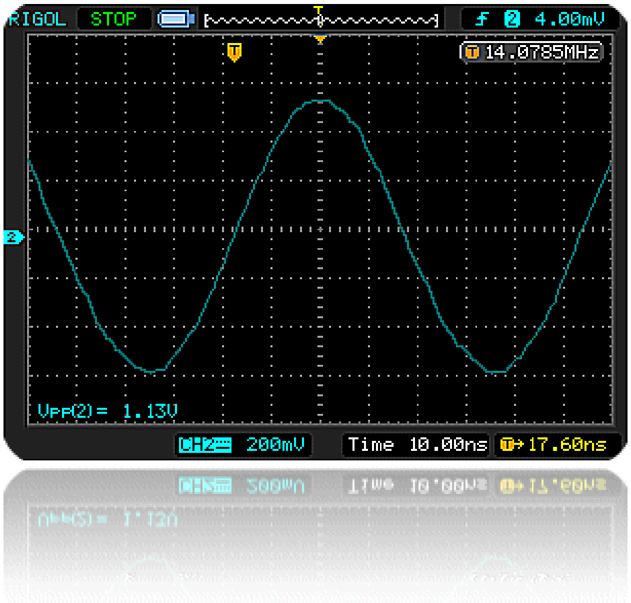
RIGOL
STOP
2 4.00mV
14.0785MHz
Vpp(2)= 1.13V
CH2 200mV
Time 10.00ns
17.60ns

Figure 7b is a screen capture of the 14 MHz circuit output after the low-pass filter and final attenuator pad. Vpp on this graphic = peak to peak voltage = 1.13 volts. The output power is 5.04 dBm, or 3.19 mW . The Q4 and Q5 source resistors and the output attenuator pads are 2 areas of the circuit where you might easily change the output power. In the end, the circuit labeled Figure 6 was chosen. Your output voltage will probably vary, but can be easily adjusted as described.

Figure 8

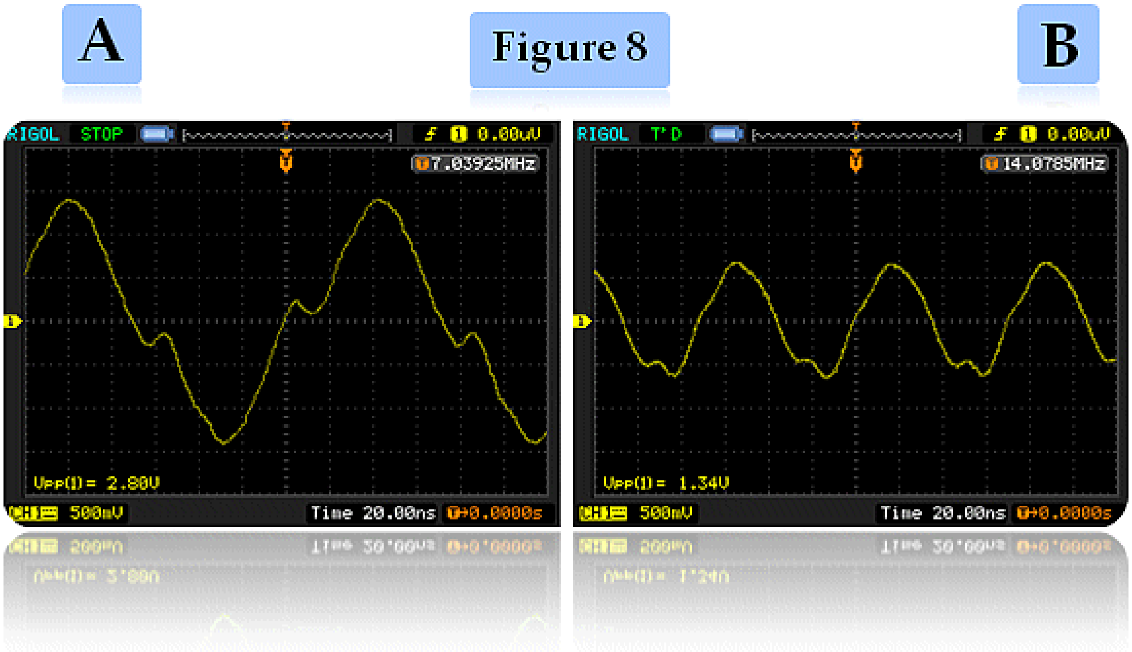
A
B
RIGOL
STOP
1 0.00uV
7.03925MHz
Vpp(1)= 2.80V
CH1 500mV
Time 20.00ns
0.0000s
RIGOL
T'D
1 0.00uV
14.0785MHz
Vpp(1)= 1.34V
CH1 500mV
Time 20.00ns
0.0000s

Figure 8 are Rigol digital screen captures. Figure 8A is a measurement taken from the anode at point DX from Figure 6. Figure 8B is a measurement taken at the cathode of DX and shows distortion caused by the diodes, reduced AC voltage and of course, frequency doubling.

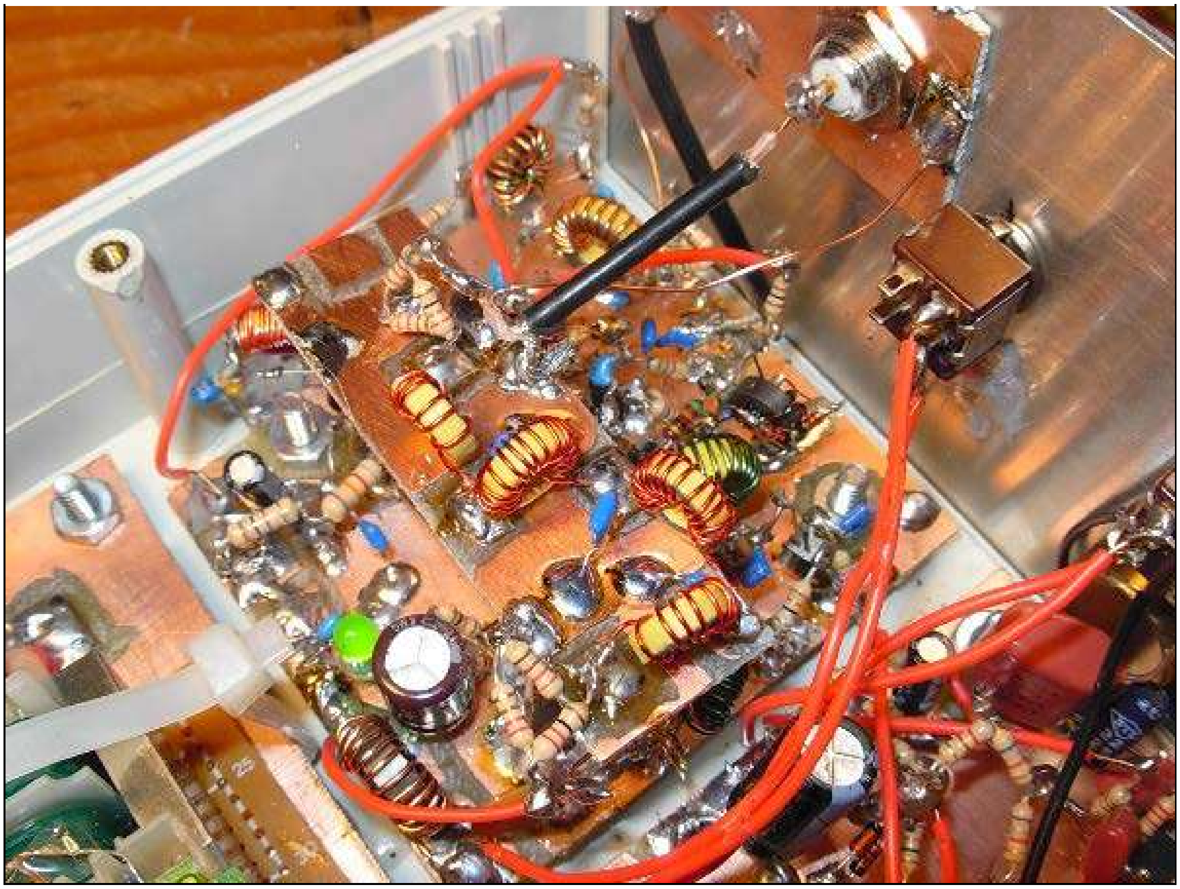

The 7 and 14 MHz circuits bread boarded and mounted in a chassis. The 7.039 MHz and 14.079 MHz outputs are connected to BNC jacks via RG-174 cable. The 15.3 MHz Chebyshev low-pass filter and -4 dB attenuator pads are on a raised Ugly Construction board. Some VCC decoupling parts are also on the bottom RF board.

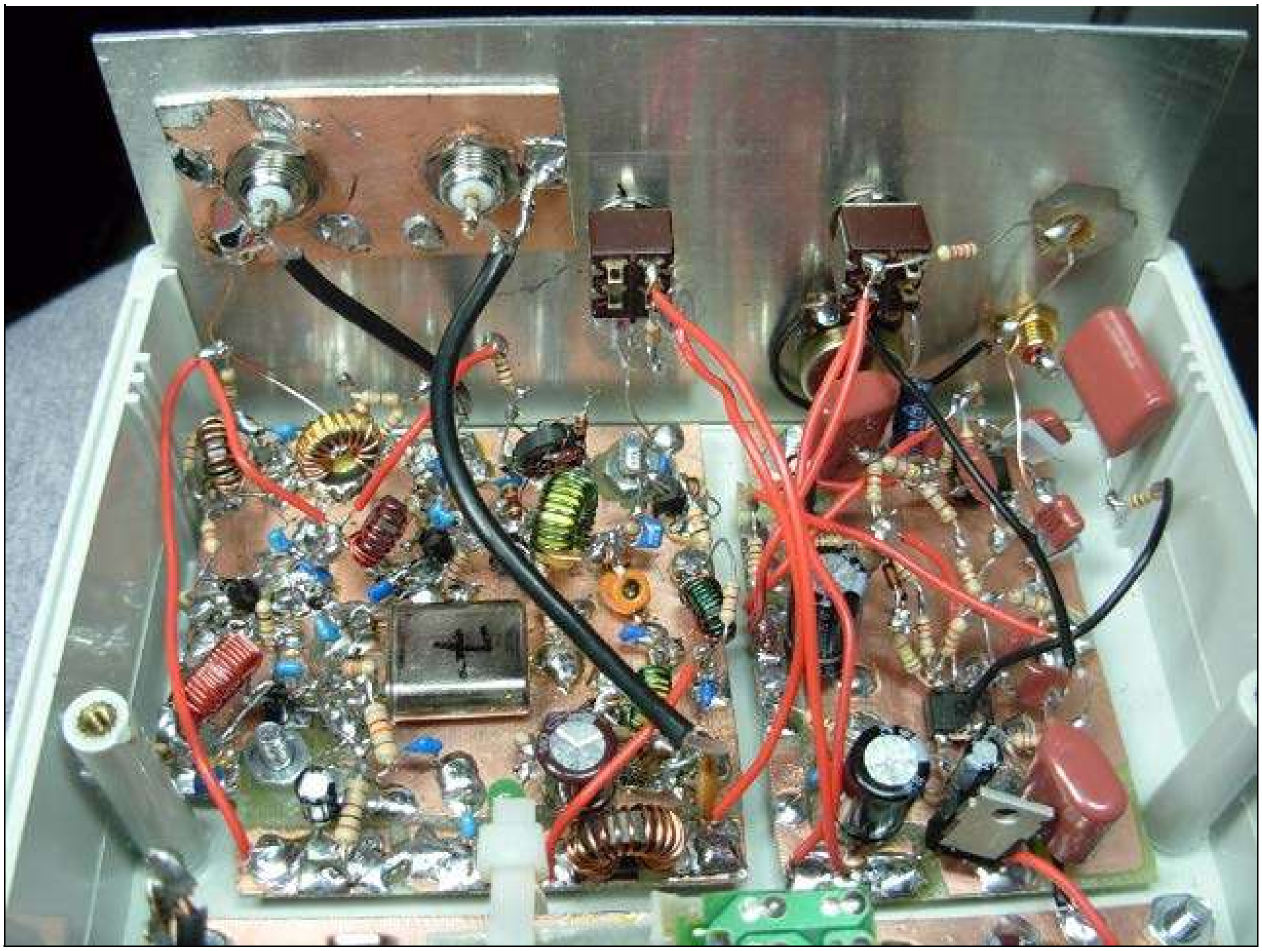

The 7 and14 MHz crystal oscillator board mounted in a project chassis along with a 1 KHz low noise oscillator This photo was a prototype version that did not have a Chebyshev low-pass filter after the 14 MHz stage.

1018 HERTZ
OFF
OFF
GAIN
14 MHZ
7 MHZ

Front view of the .001, 7 & 14 MHz oscillator. It is really fun to build your own test equipment.

Transformer Notes and Conclusion

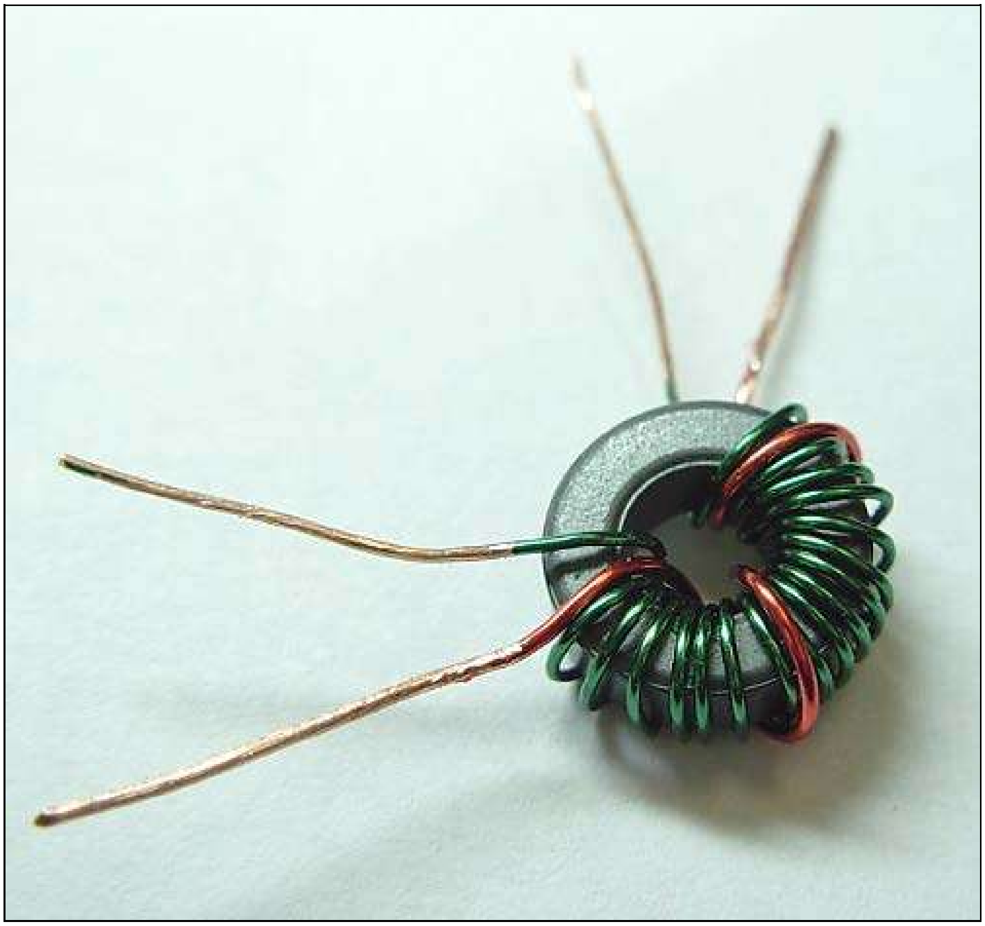

A photo of T5 from Figure 6. Some builders have emailed and stated they do not like to wind inductors/ transformers. I always ask them why? Often these builders were concerned with little details such as wire gauge and spacing, choosing the core size and which magnetic material to use. The Radio Amateur literature is replete with great tutorials on winding coils using toroids. Truly; the more toroids you wind, the easier it gets. Here are some simple points for beginners:

- Powdered iron toroids are generally for tuned circuits. I.e. A capacitor and the inductor are tuned to a center frequency. Powdered irons containing the #2 and #6 material tend to tune sharply and have fairly high Q
- Ferrites toroids are generally for use in broadband or wideband (untuned) applications. #43 material is relatively low Q and lossy as compared to the number #2 and #6 powdered iron toroids
- Wind your inductors with enamel coated magnet wire. Popular gauges include 28, 26 and 24, but this is quite variable.

Minimally, you could get by with just #43 ferrite and #6 powdered iron toroids. For example, FT37-43 ferrites and T50-6 plus T68-6 powdered iron toroids could build a lot of inductors/transformers. In the photo above, I used #24 AWG wire for the 18 turns and #22 AWG wire for the 3-turn link. The 3 turn link is grounded on one end and well anchors the transformer. Thicker wire was chosen because Ugly Construction was used and the part is really anchored with the #22 AWG wire.

Conclusion

It took quite a lot of experimentation to reach the project goal. The experience was pleasant and comparing the old and new oscilloscopes was an added bonus. Perhaps, the circuits are overly complicated, however, they are critical signal generators for my workbench. Well designed signal generators have extensive RF-proof shielding to allow serious attenuation. These will have to do.

Addition - Jan 31, 2010 5 MHz JFET Low Noise Oscillator

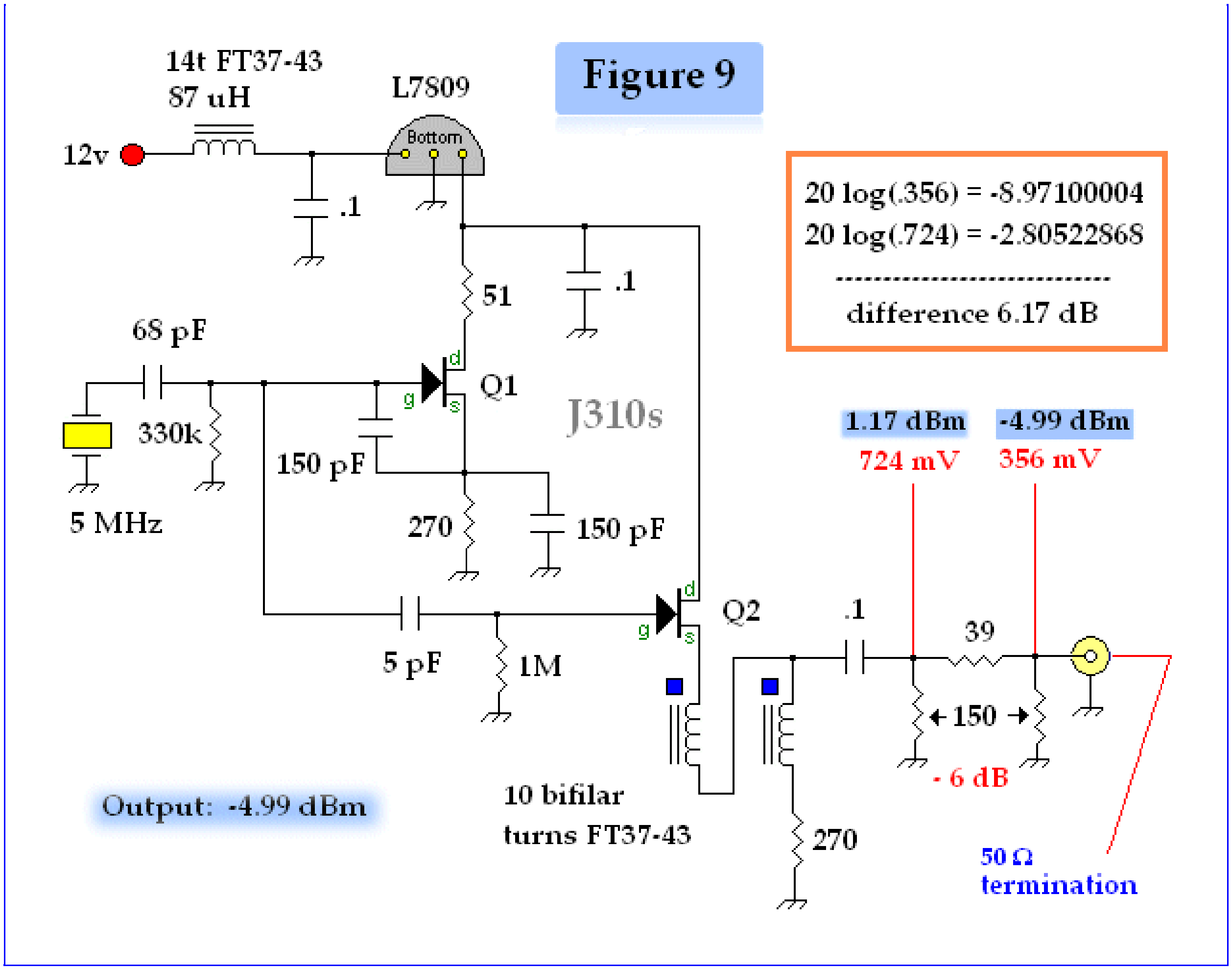
Figure 9
14t FT37-43
87 uH
L7809
Bottom
12v
.1
20 log(.356) = -8.97100004
20 log(.724) = -2.80522868
difference 6.17 dB
51
.1
68 pF
d
Q1
g
s
J310s
330k
150 pF
5 MHz
270
150 pF
1.17 dBm
724 mV
-4.99 dBm
356 mV
d
Q2
g
s
.1
39
5 pF
1M
150
- 6 dB
Output: -4.99 dBm
10 bifilar
turns FT37-43
270
50 Ω
termination

Another experimental low noise oscillator was built for 5 MHz. The desired output power was -5 dBm. The schematic and peak-peak output voltages and powers are shown in Figure 9. The 68 pF capacitor in series with the crystal was found experimentally by using a variable capacitor and measuring the output voltage and observing the scope waveform. When these appeared to be optimal, the trimmer capacitor was removed and measured at 67.17 pF. This circuit can easily be adjusted to give higher output power such as 7 dBm. To increase power, you may consider increasing the VCC to 12 volts, add a 0.1 uF source bypass capacitor in parallel with the 270 Ohm resistor of Q2 and/or adjust the pad attenuation. There are other circuit alterations to increase power, but the aforementioned are a good start.

It is important to measure your output voltages with a 50 ohm load connected to the device. Standard value resistors were used which throw off the value somewhat, but the actual attenuation of the pad is very close to the 6 dB attenuation expected. The math can be done with software such as Applets H and I on the QRP tools web page.

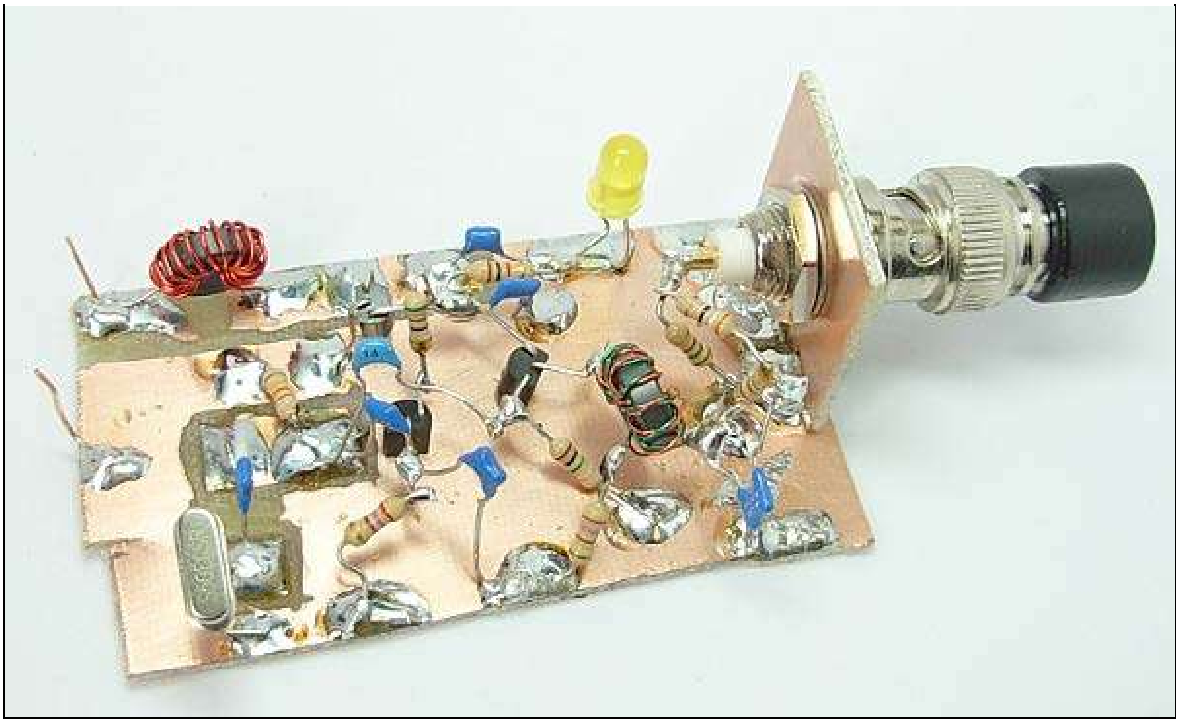

The breadboard of Figure 9 is shown above. The - 5 dBm output was required for some upcoming experiments and to study log and power measurements. The black BNC appliance is a 50 ohm load. These are very handy, but not a necessity. In most cases, a simple 51 ohm resistor to ground is the load.

Figure 10

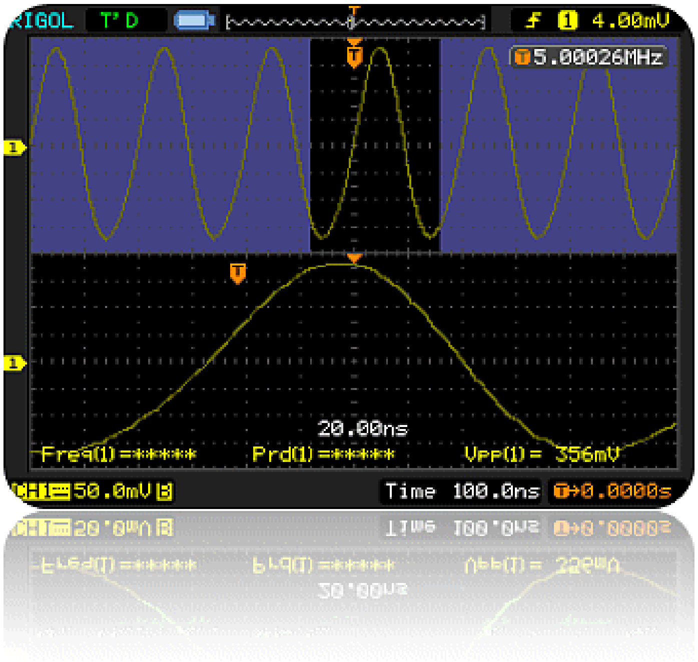
RIGOL
T'D
4.00mV
5.00026MHz
20.00ns
Freq(1)=*****
Prd(1)=*****
Vpp(1)= 356mV
CH1 50.0mV
Time 100.0ns
0.0000s

A screen capture from Figure 9 signal viewing maneuvers.

Figure 11

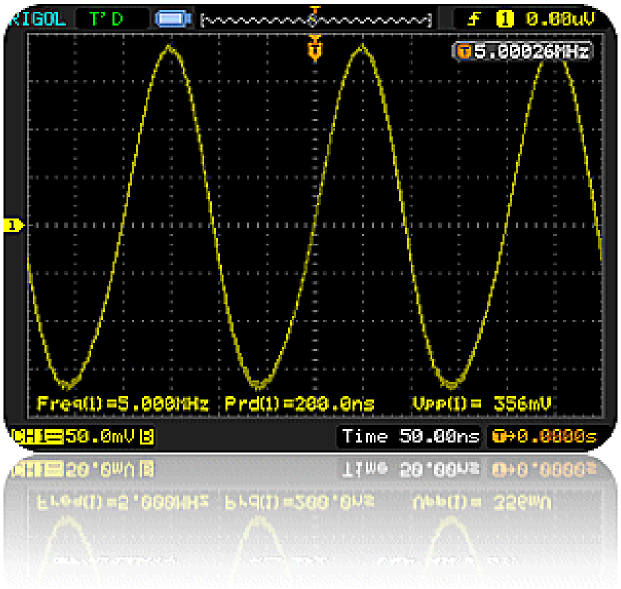
RIGOL T'D
0.00uV
5.00026MHz
1
Freq(1)=5.000MHz Prd(1)=200.0ns Vpp(1)= 356mV
CH1= 50.0mV
Time 50.00ns
0.0000s

The output waveform looks good - even on a DSO!

Spectrum Analysis

Figure 12

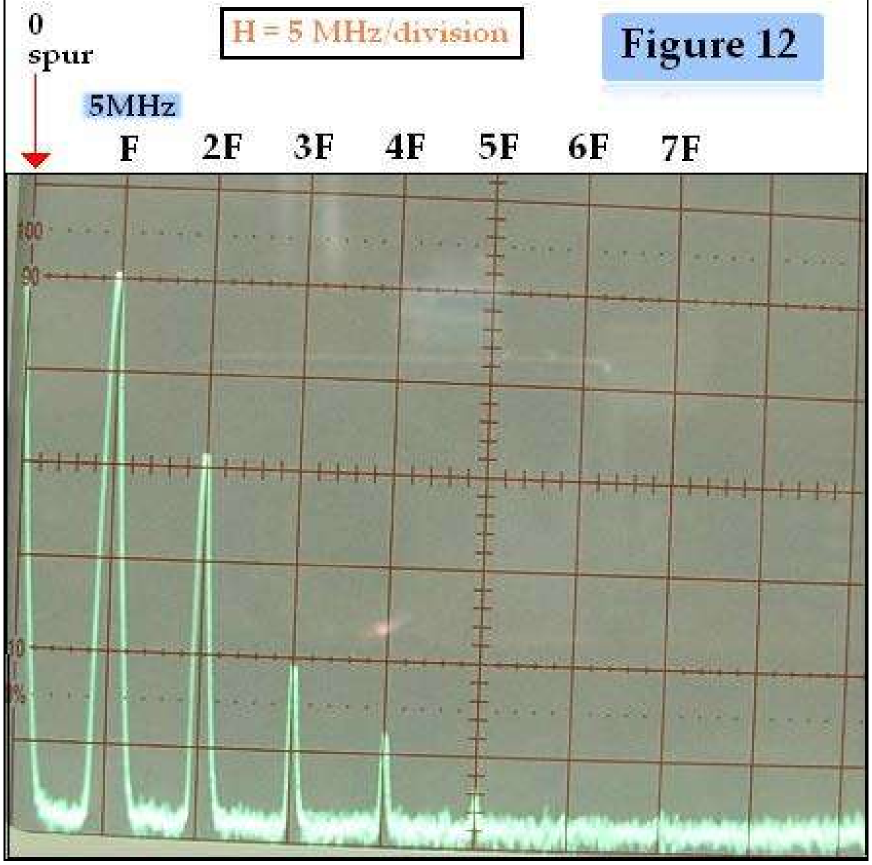

The output of the Figure 9 oscillator was further attenuated -19 dB for analysis in a spectrum analyzer. The -19 dB was chosen as this 50 ohm pad uses near standard resistor values and would well ensure a very safe signal level for the spectrum analyzer. The pad is shown in the Figure 14 schematic. The second harmonic (2F) is down about 20 dB from the fundamental. Each horizontal division is 5 MHz and each vertical division is 10 dB. Compared to other more conventional oscillators that were checked in the spectrum analyzer, this oscillator has low harmonic content.

Figure 13

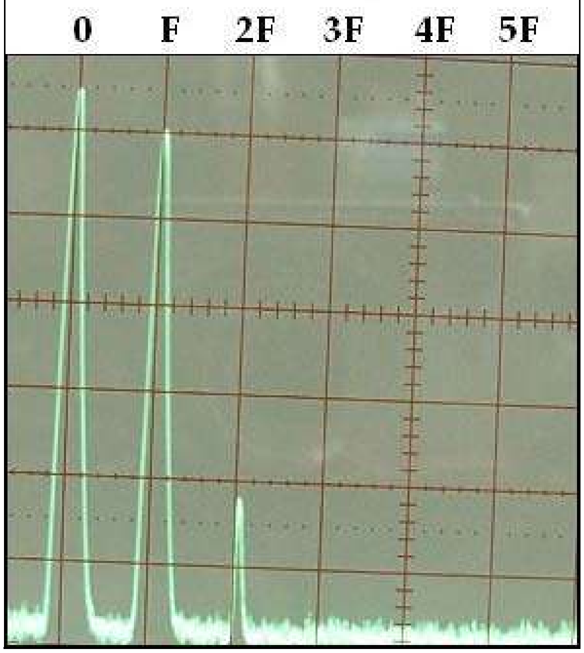

A N=5 or 5 element Chebyshev low-pass filter was placed after the Figure 9 oscillator and connected to the spectrum analyzer to see what happens. The 3 dB down frequency of this low-pass filter was 6.53 MHz. The second harmonic is now ~42 dB down. There are no measureable harmonics after 2F. It is really cool to "see" what a filter does to a signal.

Figure 14

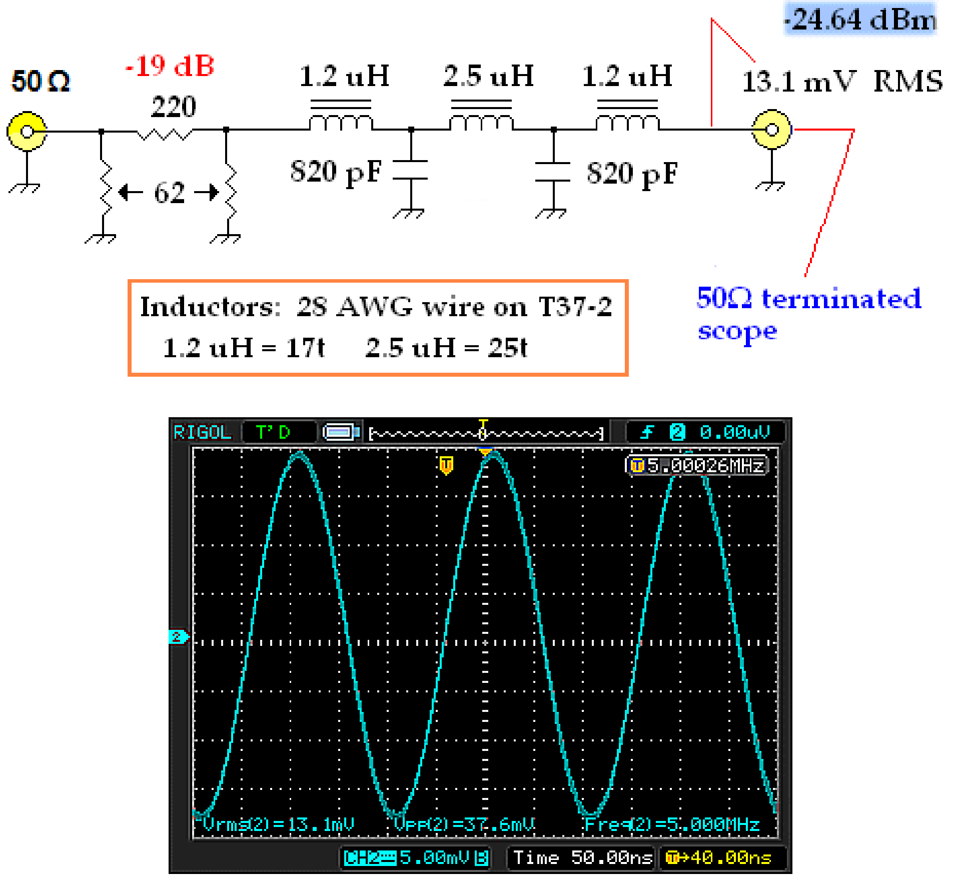
-24.64 dBm
50 Ω
-19 dB
220
1.2 uH
2.5 uH
1.2 uH
13.1 mV RMS
← 62 →
820 pF
820 pF
Inductors: 28 AWG wire on T37-2
1.2 uH = 17t 2.5 uH = 25t
50Ω terminated scope
RIGOL T'D
2 0.00uV
T 5.00026MHz
Vrms(2)=13.1mV
Vpp(2)=37.6mV
Freq(2)=5.000MHz
CH2 5.00mV
Time 50.00ns
T→40.00ns

The schematic of the -19 dB attenuator pad and the low-pass filter is shown. To measure the output, the 10X probe was disconnected. The circuit was connected to the oscilloscope via 50 ohm coaxial cable and the scope input was terminated with 2 paralleled 100 ohm resistors. The 50 ohm scope termination technique will be discussed in a future web page addition. The RMS voltage values were inputted into Applet J on the QRP Tools page to calculate the output power. The RMS output was 0.131 volts which calculated to an output power of -24.6 dBm.

QRP SWL
HOMEBUILDER

RF — Test and Measurement

Crystal Parameter Checker

ABRACON
5.0000
AB 0436

Introduction

This web page is a supplement to JavaScript Applet **G** on this web page. This software does the math using a simpified version of the method to determine motional inductance and capacitance developed by David, G3UUR. This is a very basic tutorial meant as an introductory guide for novice builders.

GND
ABRACON
5.0000
AB 0436
CRYSTAL CHECKER

Shown in the photo above is 1 of my crystal parameter checkers. The schematic may be found in many places including EMRFD Figure 3.35 (See Errata) and on this pdf by Nick, WA5BDU. A power indicator LED has been added, but the circuit is the standard design. In this breadboard, the crystal being measured is tack soldered in.

Many builders just copy other builder's I.F. filter schematics, however, your crystal filters will perform better if your design is based upon the exact parameters of the crystals you have. For the simple design or optimization of a crystal filter, it is necessary to measure crystal parallel capacitance plus take other measurements to calculate motional inductance and capacitance. Determining your crystal parameters is not difficult if you have a capacitance meter, a frequency counter and some math skills. It is easiest to use a program to crunch the math; hence I wrote a stupid-simple JavaScript applet.

Designing filters is another story; it takes knowledge, practice and good software for this. Filter design theory has been extensively covered by Anatoly Zverev, Wes Hayward and others. The work of Nick, WA5BDU is also greatly appreciated. His presentations and references are excellent for those keen on learning more about filter design.

A four crystal 5.00 MHz SSB I.F. filter was desired. 20 crystals were on hand — they were from the same batch. The crystals were all placed in the above oscillator and their frequency was measured. The 4 crystals closest matching in frequency were set aside. The crystal parameters of these 4 were then determined. Typically these values are averaged and this average is used to design or tweak the filter using software.

---

1. Measure Capacitance

The procedure for determining the parameters of 1 crystal is described.

The first step is to measure the crystal capacitance (called parallel capacitance) using a capacitance meter.

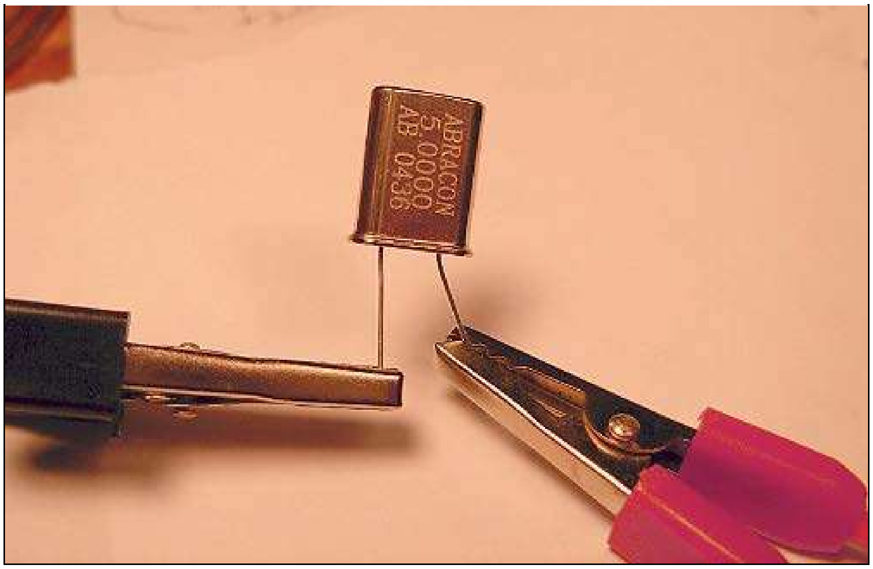

Measuring the crystal parallel capacitance

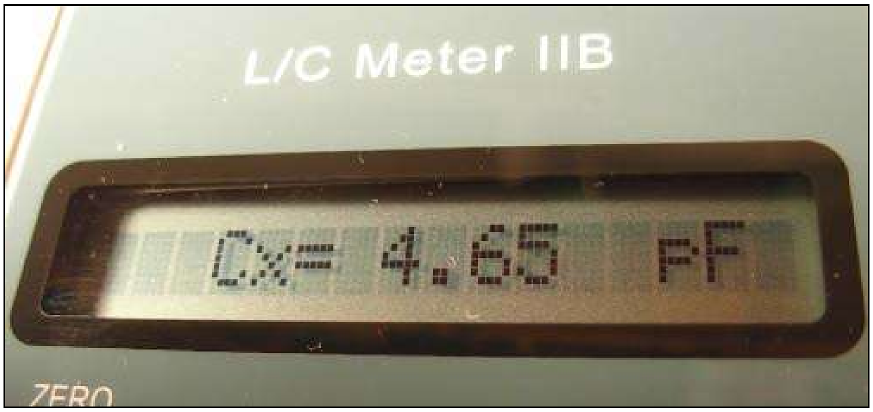

Parallel capacitance of the above 5.0 MHz crystal in an AADE LC meter

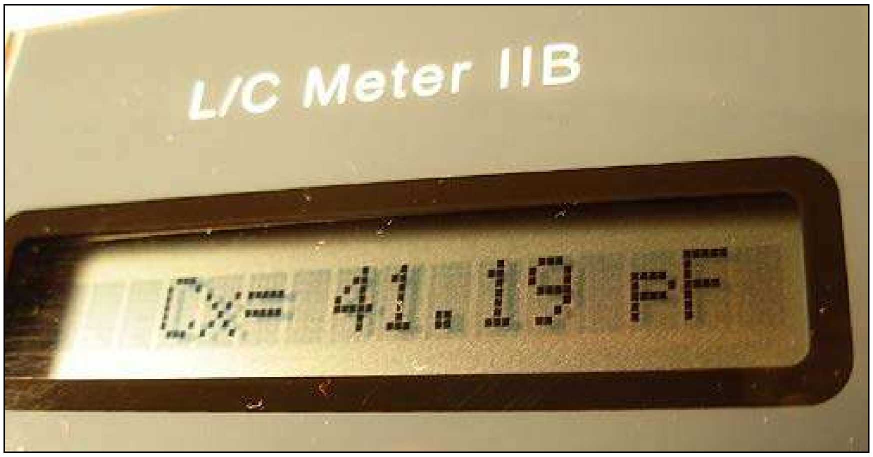

Next, measure the capacitance of the open switch plus the 33 pF fixed value cap wired in-situ. This will give you the total circuit capacitance of the open switch, the 33 pF fixed value capacitor, and any stray capacitance from your crystal holder, wires, etc. The switch itself plus stray wiring will be a few pF so the total should be around 36 to 40 pF or so. In my test oscillator, the result = 41.19 pF as shown. On my other crystal checker with a better switch, it's 36.9 pF.

In the calculation of crystal Lm and Cm, the parallel capacitance and the switch circuit capacitance will be summed.

---

2. Measure Frequencies

A crystal is put in the oscillator with the switch open. Record the frequency. Your counter must have resolution down to 1 Hertz. After recording this value, throw the switch and measure and record this frequency. You now have all the measured values required to calculate motional parameters and adjust or design a filter. Motional parameters are calculated in Applet G.

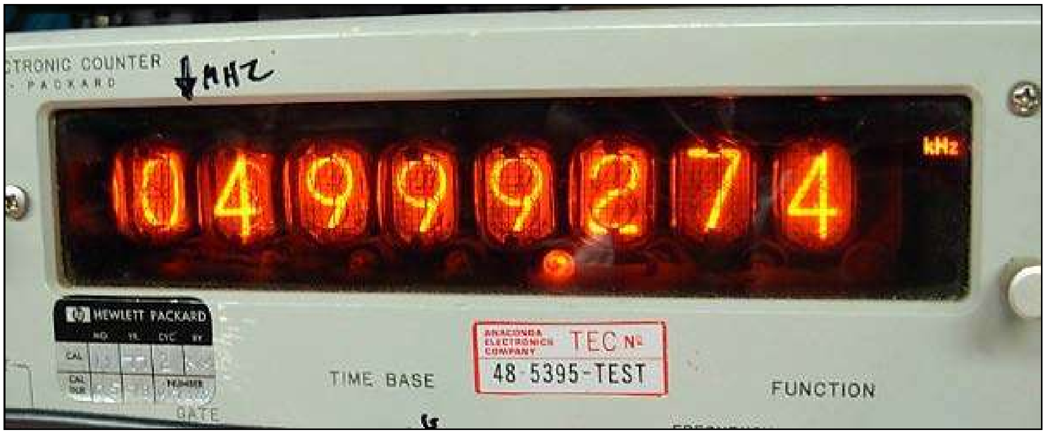
TRONIC COUNTER
PACKARD
MHZ
04999274
kHz
HEWLETT PACKARD
ANACONDA ELECTRONICS COMPANY
TEC №
48-5395-TEST
TIME BASE
FUNCTION
GATE

Frequency measured with the switch open = 4.999274 MHz

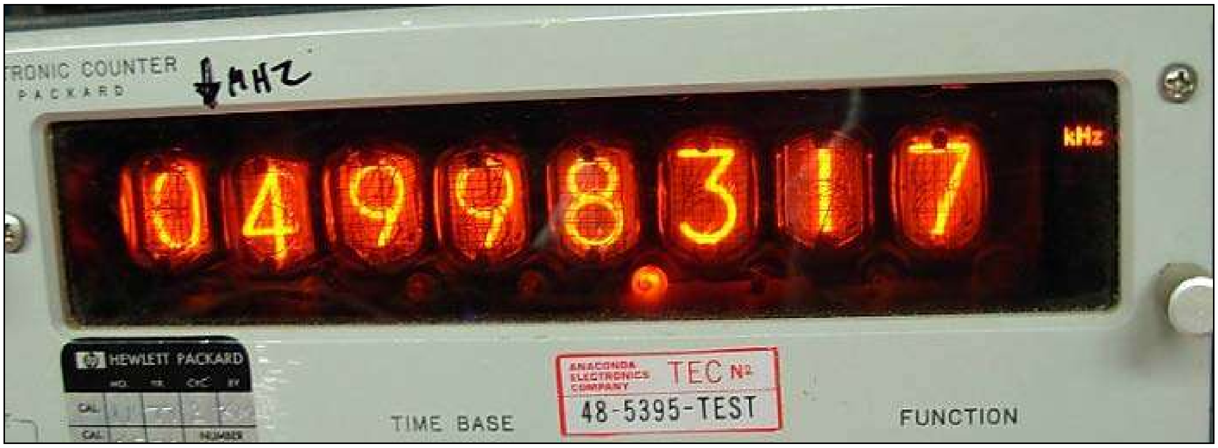
TRONIC COUNTER
PACKARD
MHZ
04998317
kHz
HEWLETT PACKARD
ANACONDA ELECTRONICS COMPANY
TEC №
48-5395-TEST
TIME BASE
FUNCTION

Frequency measured with the switch thrown = 4.998317 MHz

---

3. Do Math by Hand or with Software

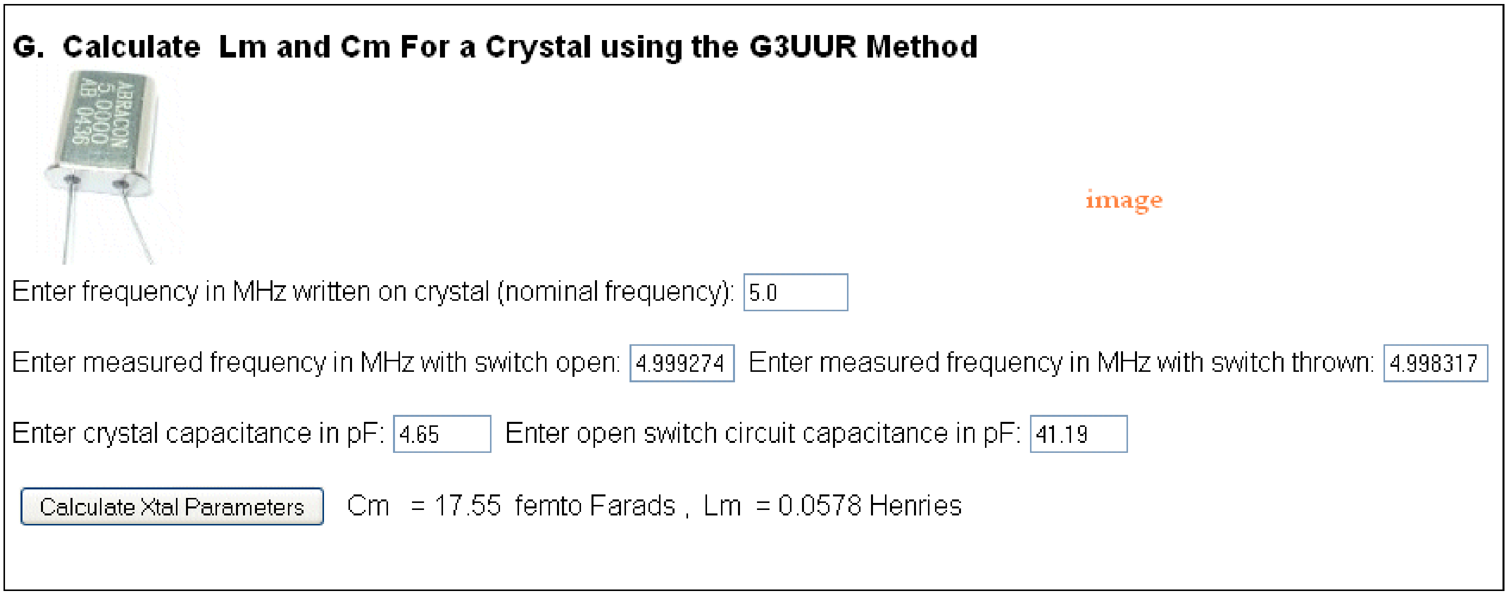

G. Calculate Lm and Cm For a Crystal using the G3UUR Method

ABRACON
5.0000
AB 0436
image

Enter frequency in MHz written on crystal (nominal frequency): 5.0

Enter measured frequency in MHz with switch open: 4.999274 Enter measured frequency in MHz with switch thrown: 4.998317

Enter crystal capacitance in pF: 4.65 Enter open switch circuit capacitance in pF: 41.19

Calculate Xtal Parameters

Cm = 17.55 femto Farads , Lm = 0.0578 Henries

The applet **G** calculation of the crystal parameters using the above measured values

---

4. Example Filter Adjustment

It is assumed that most builders will use software to design or tweak their crystal filters. The only 2 programs tried *(to date)* include AADE Filter Design and the Ladpac software collection that supplements EMRFD. I am more familiar with the Ladpac programs written by Wes, W7ZOI. Only these programs are demonstrated. Please read the instructional file Ladpac2008 Manual.pdf to understand these programs. *The Ladpac software bundle includes GPLA.*

The purpose of this tutorial is not to teach crystal filter design, but to describe a relatively simple method to tweak an existing design using your measured crystal parallel capacitance and its calculated motional inductance.

The first step is to digitally format your filter into a file that can be analyzed in GPLA. In my opinion, the easiest way to do this is to use the ladder circuit editor ladbuild02.exe or better yet, its update - ladbuild08.exe. The model filter follows:

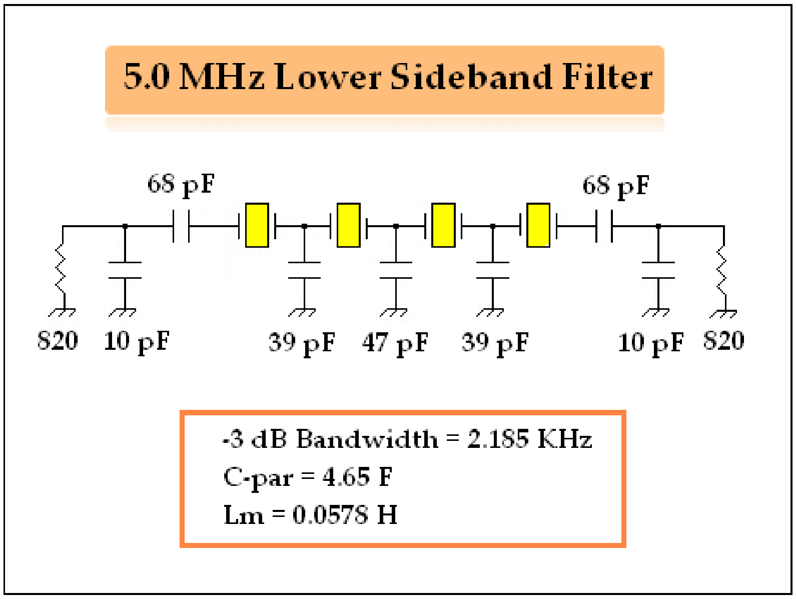
5.0 MHz Lower Sideband Filter
68 pF
68 pF
820
10 pF
39 pF
47 pF
39 pF
10 pF
820
-3 dB Bandwidth = 2.185 KHz
C-par = 4.65 F
Lm = 0.0578 H

The model 5 MHz SSB filter

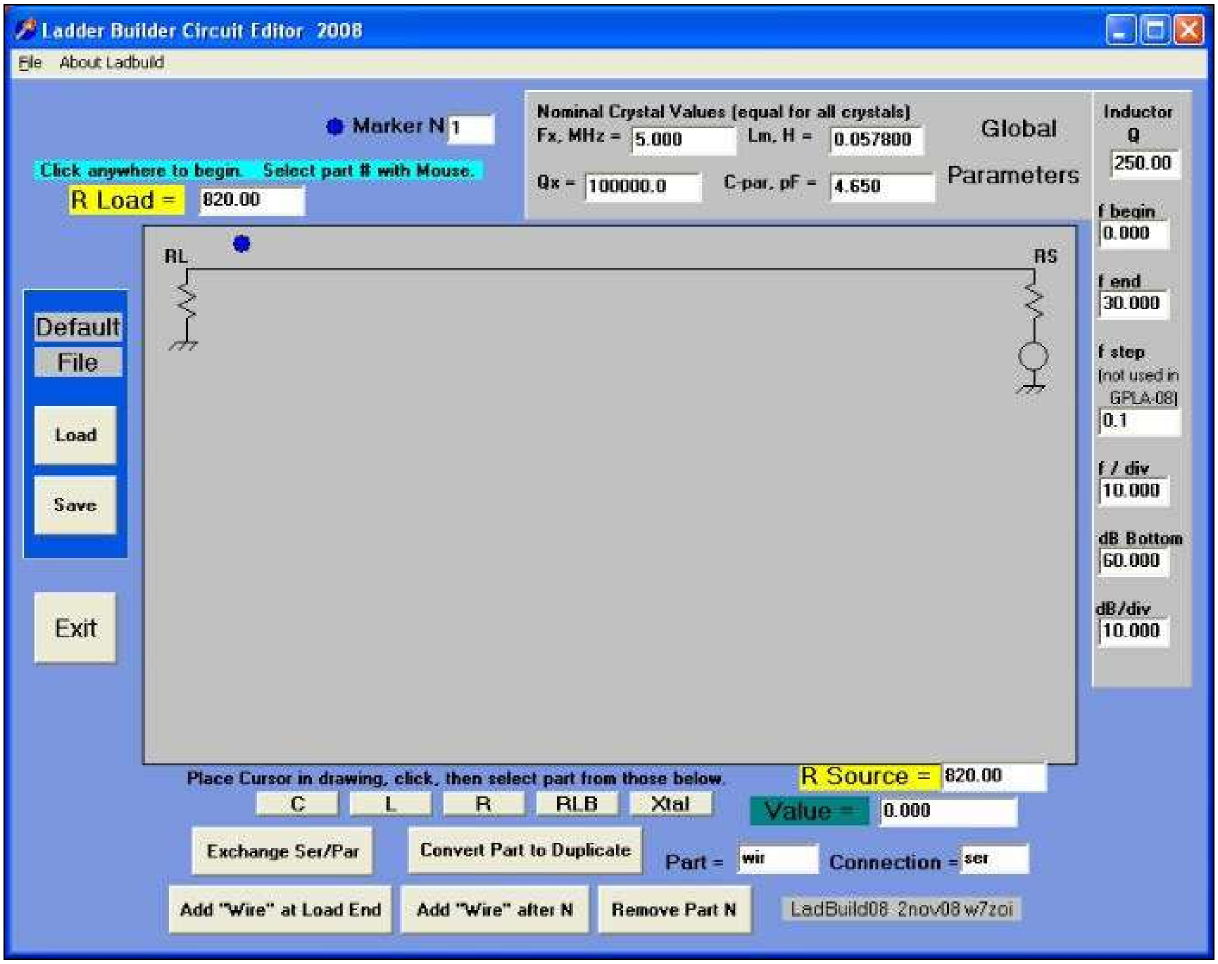
Ladder Builder Circuit Editor 2008
File
About Ladbuild
Global
Parameters
RL
RS
Default
File
Load
Save
Exit
Place Cursor in drawing, click, then select part from those below.
C
L
R
RLB
Xtal
Exchange Ser/Par
Convert Part to Duplicate
Add "Wire" at Load End
Add "Wire" after N
Remove Part N
LadBuild08 2nov08 w7zoi

Marker N 1

Click anywhere to begin. Select part # with Mouse.

R Load = 820.00

Nominal Crystal Values (equal for all crystals)

Fx, MHz = 5.000 Lm, H = 0.057800

Qx = 100000.0 C-par, pF = 4.650

Inductor Q 250.00

f begin 0.000

f end 30.000

f step (not used in GPLA-08) 0.1

f / div 10.000

dB Bottom 60.000

dB/div 10.000

R Source = 820.00

Value = 0.000

Part = wir

Connection = ser

Clear any existing components and enter the termination R, C-par and Lm values. Qx is set at 100000

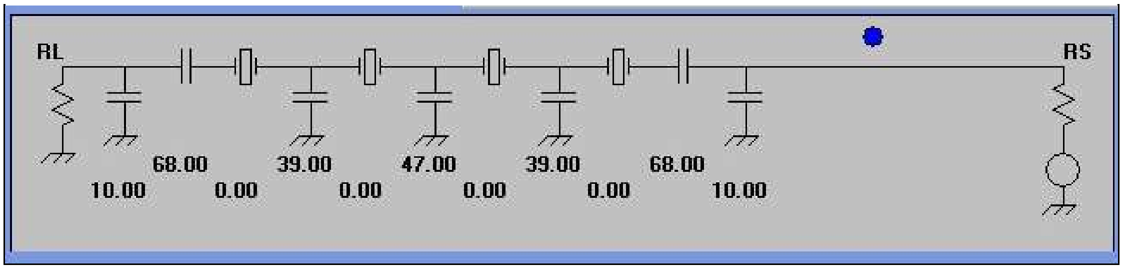
RL
RS
68.00
39.00
47.00
39.00
68.00
10.00
0.00
0.00
0.00
0.00
10.00

Build your filter within the editor. Save your work. Start up GPLA and load your newly saved filter.

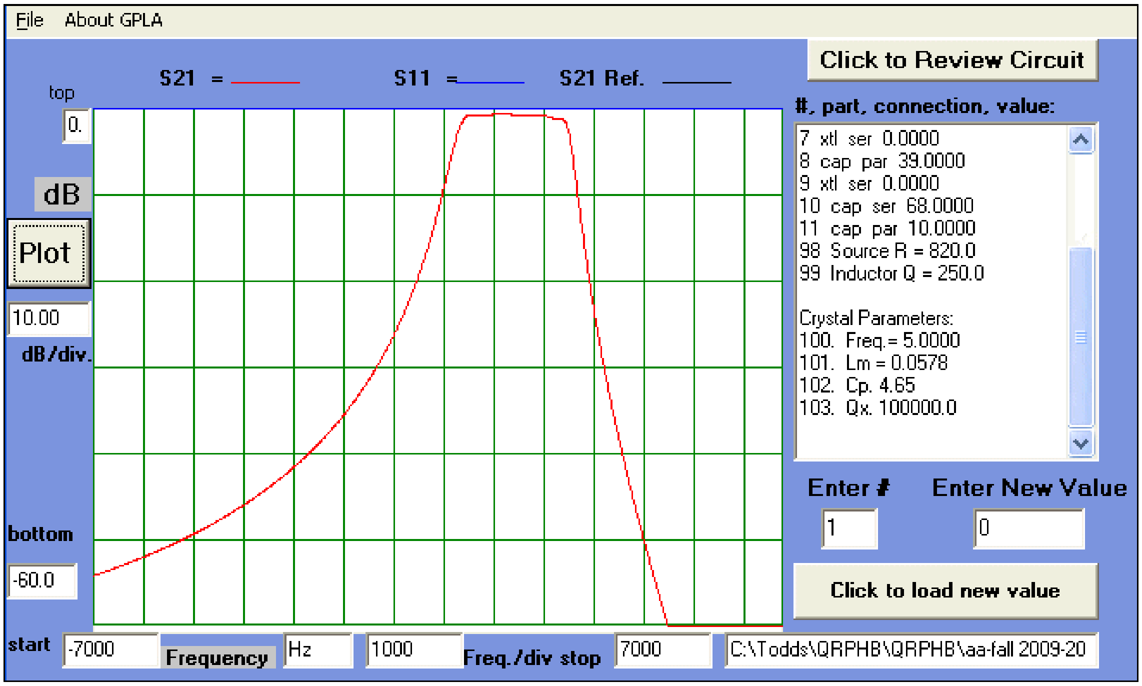
File
About GPLA
S21 =
S11 =
S21 Ref.
dB
Plot
Click to Review Circuit
Click to load new value
Frequency
Freq./div
stop

top: 0.

10.00 dB/div.

bottom: -60.0

start: -7000 Frequency Hz 1000 Freq./div stop 7000

C:\Todds\QRPHB\QRPHB\aa-fall 2009-20

#, part, connection, value:

```
7  xtl  ser  0.0000
8  cap  par  39.0000
9  xtl  ser  0.0000
10  cap  ser  68.0000
11  cap  par  10.0000
98  Source R = 820.0
99  Inductor Q = 250.0

Crystal Parameters:
100.  Freq.= 5.0000
101.  Lm = 0.0578
102.  Cp. 4.65
103.  Qx. 100000.0
```

Enter #: 1

Enter New Value: 0

Set a sweep and **x** axis increment (-7000, 1000 and 7000 in this example). Push the Plot button

Let's say you wanted to use this filter design and have determined the average parameters of the 5.00 MHz crystals in your parts collection. Let's assume that for your crystals, C-par = 3.1 pF, and Lm = 0.098H. Input these values in GPLA.

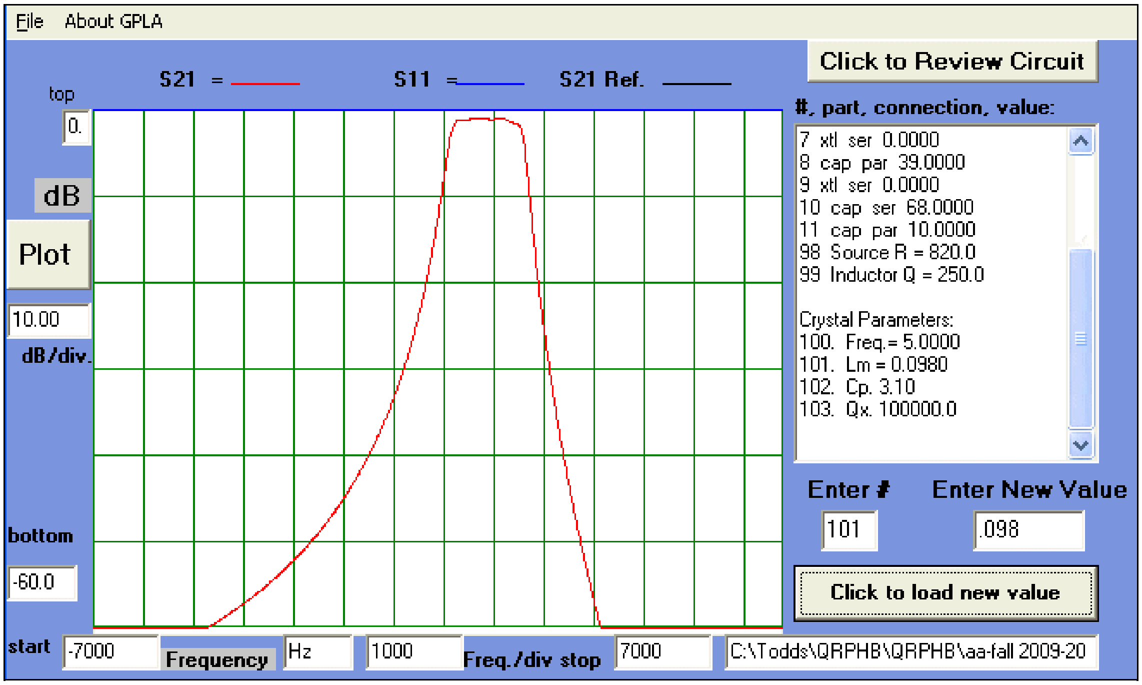
File
About GPLA
S21 =
S11 =
S21 Ref.
dB
Plot
Click to Review Circuit
Click to load new value
Frequency
Freq./div
stop

top: 0.

10.00 dB/div.

bottom: -60.0

start: -7000 Frequency Hz 1000 Freq./div stop 7000

C:\Todds\QRPHB\QRPHB\aa-fall 2009-20

#, part, connection, value:

```
7  xtl  ser  0.0000
8  cap  par  39.0000
9  xtl  ser  0.0000
10  cap  ser  68.0000
11  cap  par  10.0000
98  Source R = 820.0
99  Inductor Q = 250.0

Crystal Parameters:
100.  Freq.= 5.0000
101.  Lm = 0.0980
102.  Cp. 3.10
103.  Qx. 100000.0
```

Enter #: 101

Enter New Value: .098

Look what happened to the crystal filter's bandwidth. Our - 3dB bandwidth is now somewhere around 1464 Hertz. This simple experiment illustrates how important it is to use the parameters of <u>your</u> crystals to obtain a desired filter response.

Experiment with the various functions in GPLA to learn how to use it. Set whatever reasonable sweep you want. This program is best learned by using it repeatedly.

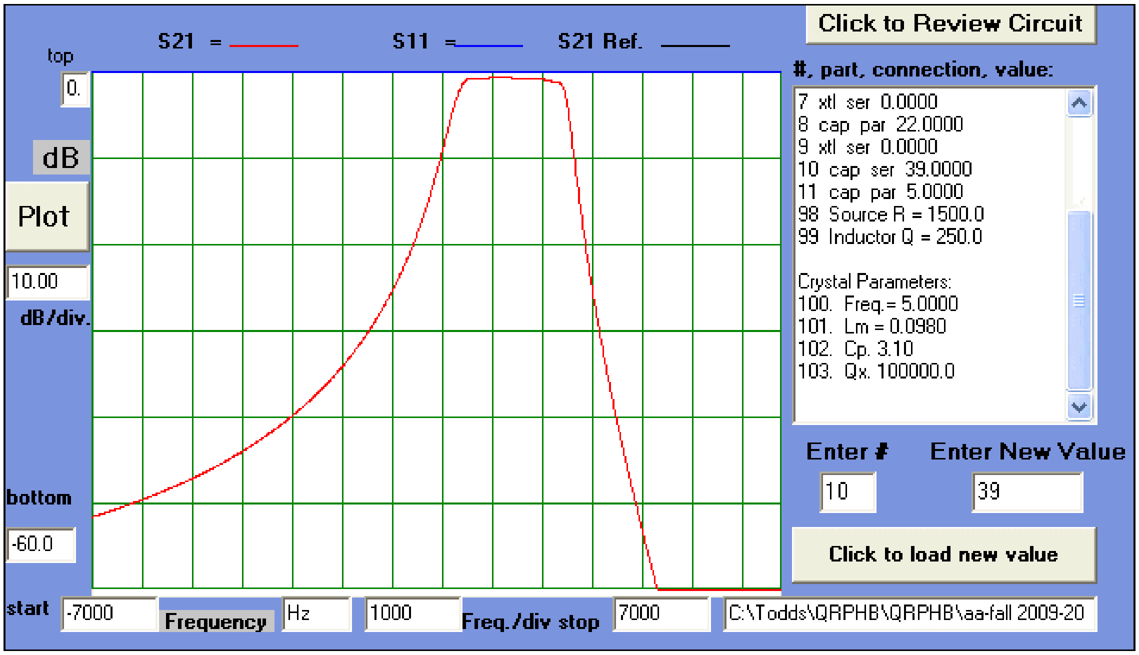
S21 =
S11 =
S21 Ref.
Click to Review Circuit
top
dB
Plot
dB/div.
bottom
Enter #
Enter New Value
Click to load new value
start
Frequency
Freq./div stop

#, part, connection, value:

```
7  xtl  ser  0.0000
8  cap  par  22.0000
9  xtl  ser  0.0000
10  cap  ser  39.0000
11  cap  par  5.0000
98  Source R = 1500.0
99  Inductor Q = 250.0

Crystal Parameters:
100.  Freq.= 5.0000
101.  Lm = 0.0980
102.  Cp. 3.10
103.  Qx. 100000.0
```

top: 0.

10.00 dB/div.

bottom: -60.0

Enter #: 10

Enter New Value: 39

start: -7000 Frequency Hz

1000 Freq./div stop: 7000

C:\Todds\QRPHB\QRPHB\aa-fall 2009-20

In the above screen capture, the above filter was tweaked to "re-establish" a -3dB bandwidth of ~2.172 KHz. All adjustment was performed entirely in GPLA by swapping capacitor values and observing the resultant waveform. When you get an overall pleasing bandwidth plus shape, but there is too much ripple at the top, generally you must increase the terminating R values. This is the brute-force, manual way to tune filters. For this method, you need not understand terminology such as as series resistance, MESH, K or Q values, Butterworth response, or Chebyshev with 0.1 dB of ripple.

Admittedly, at first, this method can be quite time consuming and tedious, however, with practice, you may be able to tweak a filter in only a few minutes. Clearly, the more you dig into understanding crystal filter design, the better your filters can be, however, getting overly complex can scare off builders who are new to this hobby. Note these filters use standard value capacitors and resistors; perfect for popcorn I.F. filters.

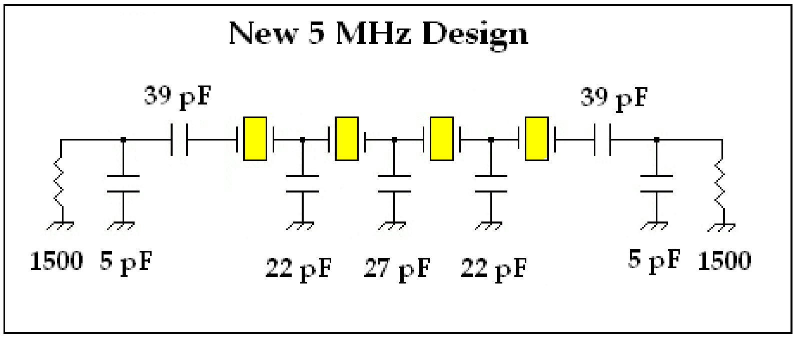
New 5 MHz Design
39 pF
39 pF
1500
5 pF
22 pF
27 pF
22 pF
5 pF
1500

The original 5 MHz Model filter with updated C and R values using Lm = 0.098 and C-par = 3.1

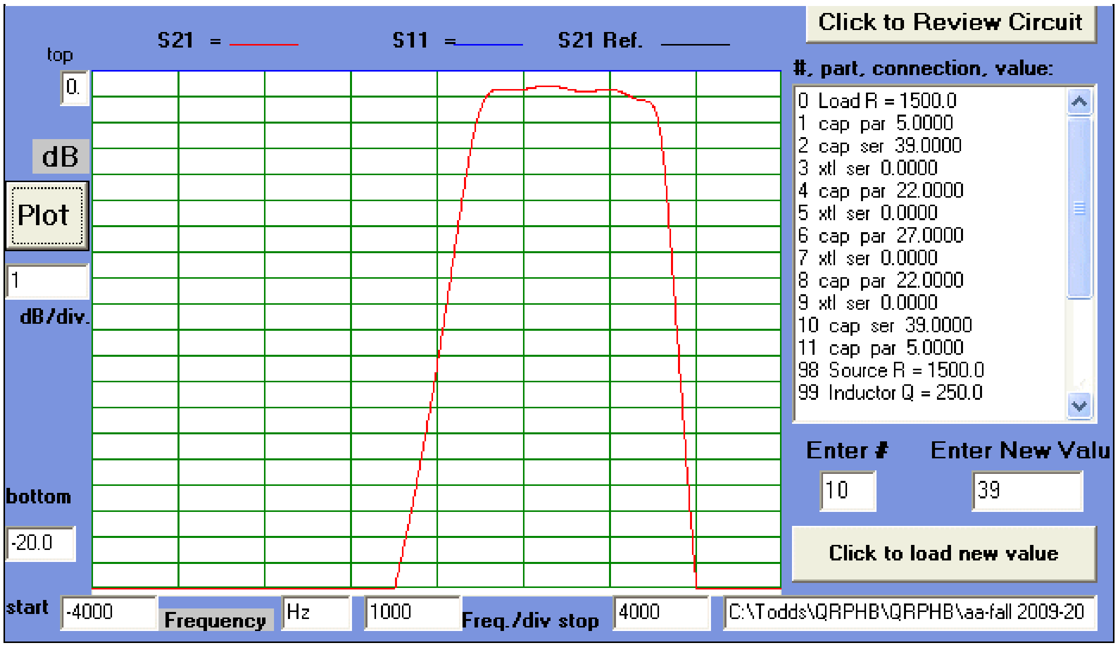
S21 =
S11 =
S21 Ref.
Click to Review Circuit
#, part, connection, value:
dB
Plot
dB/div.
Enter #
Enter New Valu
Click to load new value
Frequency
Freq./div stop

```
0  Load R = 1500.0
1  cap  par  5.0000
2  cap  ser  39.0000
3  xtl  ser  0.0000
4  cap  par  22.0000
5  xtl  ser  0.0000
6  cap  par  27.0000
7  xtl  ser  0.0000
8  cap  par  22.0000
9  xtl  ser  0.0000
10 cap  ser  39.0000
11 cap  par  5.0000
98 Source R = 1500.0
99 Inductor Q = 250.0
```

top 0.

1 dB/div.

bottom -20.0

start -4000 Frequency Hz 1000 Freq./div stop 4000

C:\Todds\QRPHB\QRPHB\aa-fall 2009-20

Enter # 10

Enter New Valu 39

GPLA zoom of the **Y** axis showing the first 20 dB of attenuation.

---

5. The Model 5 MHz SSB filter Design

Although this page is not about crystal filter design, an example follows for reference purposes.

For designing filters, the application xlad08.exe is a good choice. The following 3 screen captures show the raw design process and GPLA analysis of the model 5 MHz Lower Sideband filter shown earlier with C-par = 4.65 pF and Lm = .0578. There are some great articles in print and on the Web to study if you want to learn about filter design. The Ladpac software from EMRFD is excellent. My special thanks to Wes, W7ZOI for answering my questions about his software. From this information, I was able to make this web page.

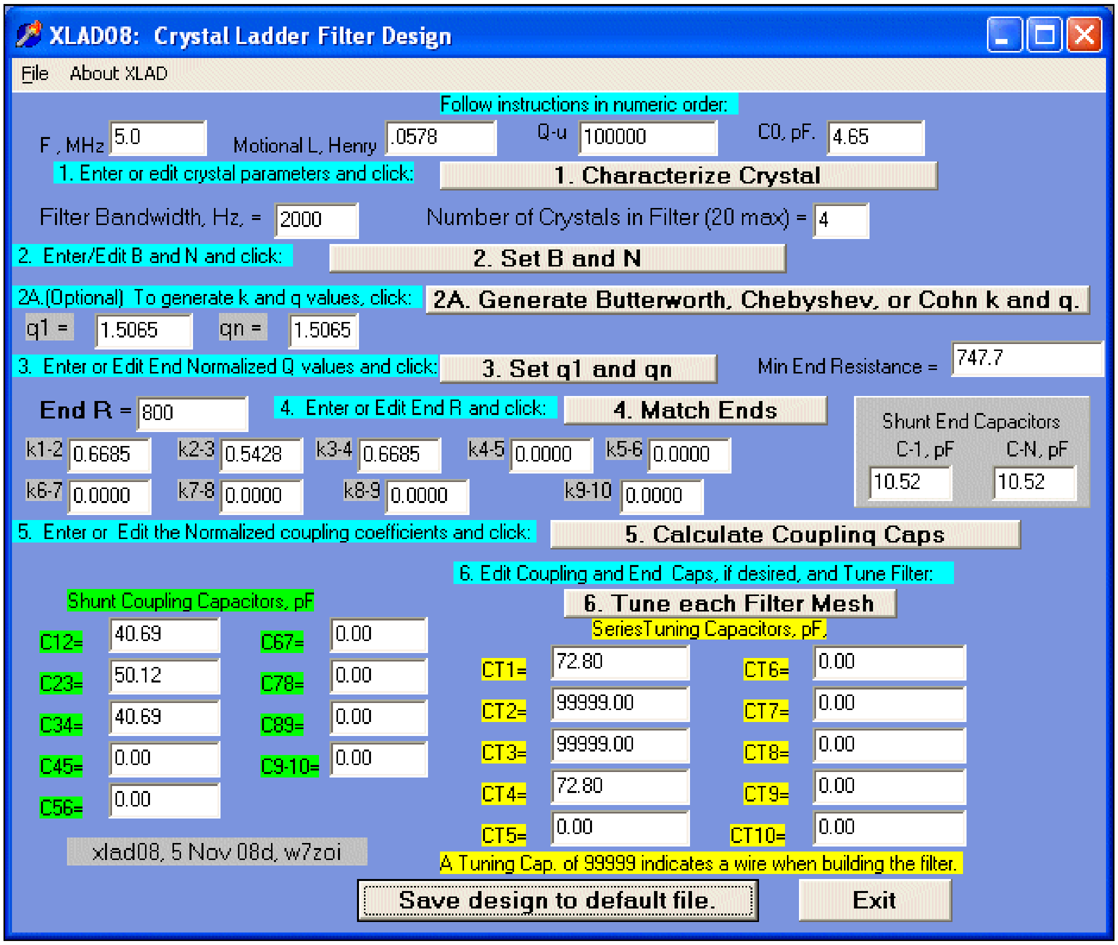
XLAD08: Crystal Ladder Filter Design
File
About XLAD
1. Characterize Crystal
2. Set B and N
2A. Generate Butterworth, Chebyshev, or Cohn k and q.
3. Set q1 and qn
4. Match Ends
5. Calculate Coupling Caps
6. Tune each Filter Mesh
Save design to default file.
Exit

Follow instructions in numeric order:

F, MHz 5.0

Motional L, Henry .0578

Q-u 100000

C0, pF. 4.65

1. Enter or edit crystal parameters and click:

Filter Bandwidth, Hz, = 2000

Number of Crystals in Filter (20 max) = 4

2. Enter/Edit B and N and click:

2A.(Optional) To generate k and q values, click:

q1 = 1.5065

qn = 1.5065

3. Enter or Edit End Normalized Q values and click:

Min End Resistance = 747.7

End R = 800

4. Enter or Edit End R and click:

Shunt End Capacitors

C-1, pF 10.52

C-N, pF 10.52

k1-2 0.6685

k2-3 0.5428

k3-4 0.6685

k4-5 0.0000

k5-6 0.0000

k6-7 0.0000

k7-8 0.0000

k8-9 0.0000

k9-10 0.0000

5. Enter or Edit the Normalized coupling coefficients and click:

6. Edit Coupling and End Caps, if desired, and Tune Filter:

Shunt Coupling Capacitors, pF

C12= 40.69

C23= 50.12

C34= 40.69

C45= 0.00

C56= 0.00

C67= 0.00

C78= 0.00

C89= 0.00

C9-10= 0.00

SeriesTuning Capacitors, pF.

CT1= 72.80

CT2= 99999.00

CT3= 99999.00

CT4= 72.80

CT5= 0.00

CT6= 0.00

CT7= 0.00

CT8= 0.00

CT9= 0.00

CT10= 0.00

xlad08, 5 Nov 08d, w7zoi

A Tuning Cap. of 99999 indicates a wire when building the filter.

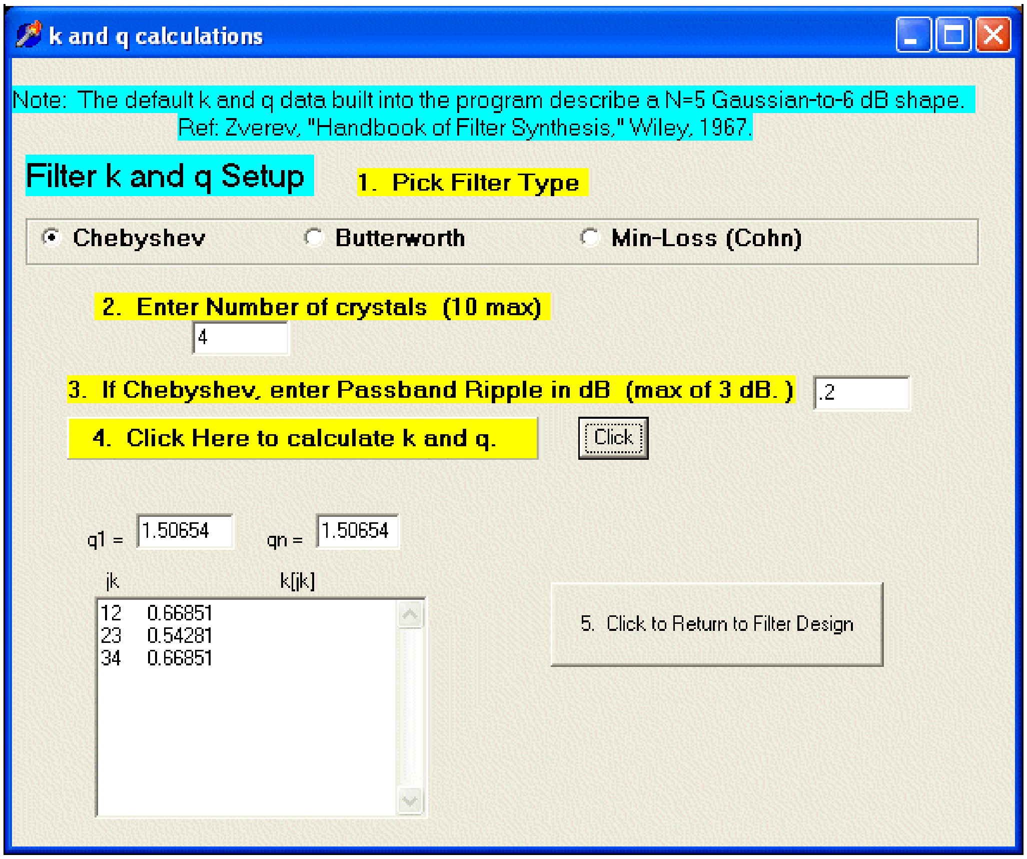

Note: The default k and q data built into the program describe a N=5 Gaussian-to-6 dB shape.
Ref: Zverev, "Handbook of Filter Synthesis," Wiley, 1967.

Filter k and q Setup

1. Pick Filter Type

● Chebyshev ○ Butterworth ○ Min-Loss (Cohn)

2. Enter Number of crystals (10 max): 4

3. If Chebyshev, enter Passband Ripple in dB (max of 3 dB.): .2

4. Click Here to calculate k and q.

q1 = 1.50654 qn = 1.50654

```
jk      k[jk]
12   0.66851
23   0.54281
34   0.66851
```

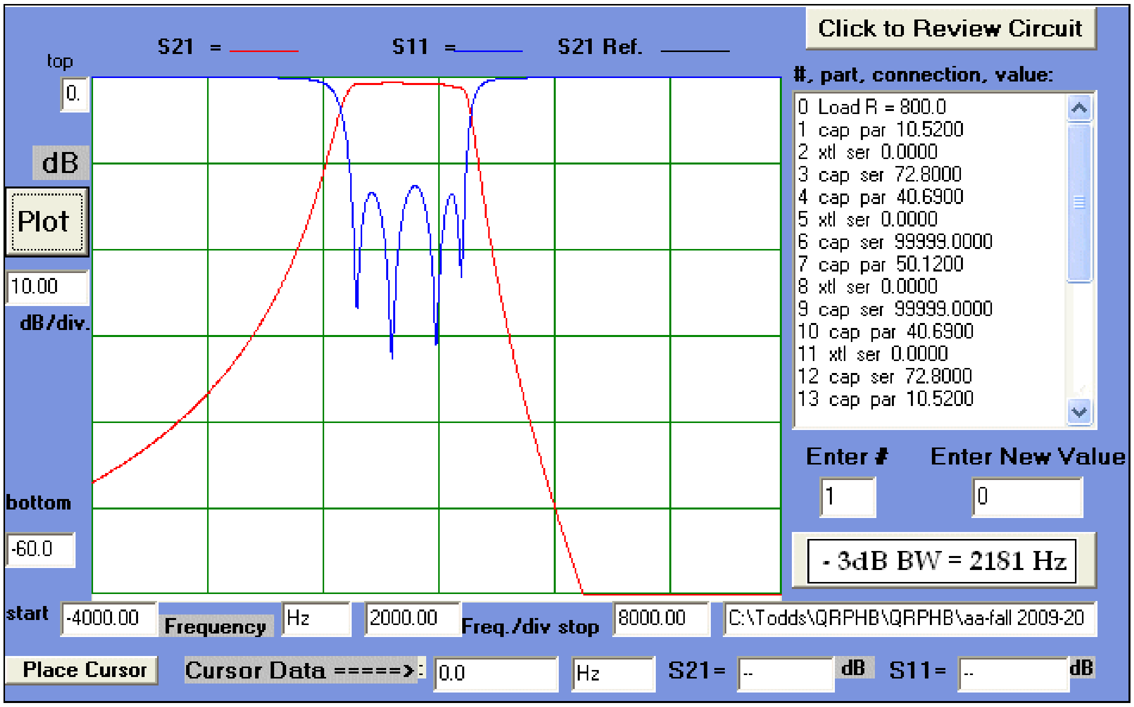

top: 0.
10.00 dB/div.
bottom: -60.0

#, part, connection, value:

```
0 Load R = 800.0
1 cap par 10.5200
2 xtl ser 0.0000
3 cap ser 72.8000
4 cap par 40.6900
5 xtl ser 0.0000
6 cap ser 99999.0000
7 cap par 50.1200
8 xtl ser 0.0000
9 cap ser 99999.0000
10 cap par 40.6900
11 xtl ser 0.0000
12 cap ser 72.8000
13 cap par 10.5200
```

Enter #: 1 Enter New Value: 0

- 3dB BW = 2181 Hz

start -4000.00 Frequency Hz 2000.00 Freq./div stop 8000.00 C:\Todds\QRPHB\QRPHB\aa-fall 2009-20

Cursor Data =====>: 0.0 Hz S21= -- dB S11= -- dB

---

6. Conclusion

This web page presents a brief method to calculate crystal motional parameters and as required, to adjust crystal filter circuits to function optimally. This approach like my Java-script applet are simplistic to avoid the fear factor associated with crystal ladder design.

Listening tests are also valuable for assessing crystal filter function. Is the bandwidth as you expected? Does the filter ring excessively? Does it sound tinny? In the recent past, the crystal filters in 2 kit receivers/transceivers were tweaked as a favor to friends. *Please note, I have total respect for people who sell kits and appreciate the contribution they make to our hobby.*

The crystals of these Cohn type filters were removed and analyzed and the bandwidth was not as specified. In 1 case, the filter shape looked terrible in GPLA. Clearly, the kit sellers provided crystals which had markedly different parameters from those used by the original circuit designer. The I.F. filter capacitors were replaced with appropriate values and the R values were adjusted via either resistors and/or transformer ratios and the improved filters sounded pleasant.

It is a real treat to listen to a receiver with a well designed crystal filter — sadly, I don't enjoy this experience that often.

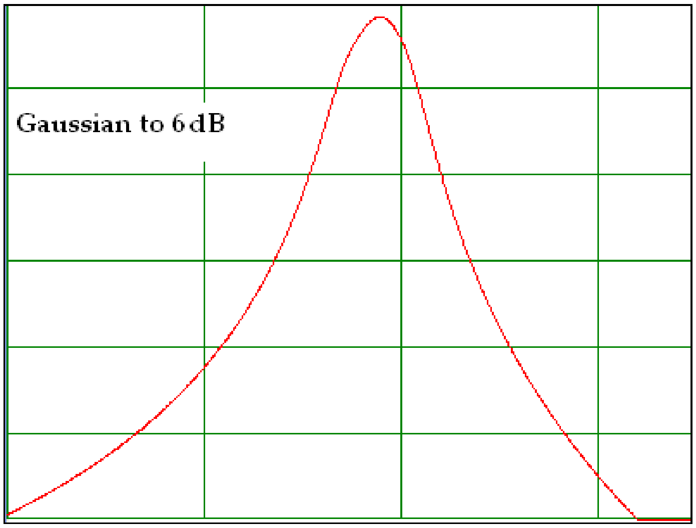

In my opinion, the best sounding CW crystal filter design is the Gaussian to 6 dB. Some operators would never use such a filter in a contest-grade receiver as the filter skirts are not steep enough for them. There are tricks to make the stop band better (more like a Chebyshev response, but without the ringing), however, this topic is out of scope.

I sincerely ask for your feedback on the G. JavaScript Applet. Does it work correctly? How could it be improved? Can you contribute better code? Thanks and 73.

QRP — Posdata for January 2012 — More on Crystal Ladder Design

Important to both superheterodyne receivers and single sideband transmitters, crystal ladder filter design lies juxtaposed as both a favorite and feared RF design topic. Newer builders may lack math skills, and/or become paralyzed by the terminology — or lack the ambition to learn or apply good bench practices. Even a cursory Internet search returns many fabulous files to read — witnessing a crystal ladder filter design article explosion.

The difficulty characterizing and building filters has progressively declined since the advent of the first handheld computers — improvements in personal computers and filter design software allows astute builders to pursue even complex xtal filter response shapes in 2012.

In QST for July 1987. Wes, W7ZOI wrote **Designing and Building Simple Crystal Filters**. This article promoted Cohn or Min-loss filters and its intent was to transcend the math and measurement associated with xtal ladder filters of the day and allow builders to just frequency match some crystals, and then go experiment.

In my estimation, this article proved revolutionary — soon after, builders around the globe, Elecraft, and other kit companies embraced this technique/topology and the rest is history. *(I call it the paper that launched a thousand kits).* If you're a new builder and feel overwhelmed by the material on or referenced by this web page, please consider first obtaining this article.

Learning about crystal ladder filter design is time well spent.

In 2011/2012 I explored the works of 3 builders who share their work via the web and/or journals.

Horst Steder, DJ6EV and Jack Hardcastle, G3JIR

The Steder and Hardcastle works emphasize that we need to measure/calculate crystal L and C parameters, plus the coupling and tuning capacitances (not just frequency). Through emails with Horst, DJ6EV, I learned many things, but 3 stand out:

1. It's better to design a good filter than fix a bad one.
2. Careful measurement of your crystal parameters plus software design and simulation = the best chance for getting your desired performance.
3. Deriving motional parameters from a 3 dB bandwidth measurement remains a great way to characterize multiple xtals. *In my opinion, the G3UUR method is the easiest way to characterize a small batch of xtals.*

Some of the earliest references to modern crystal ladder design I've found were written by Jack Hardcastle and published in RadCom and QST — I later confirmed this by reading work by Wes, W7ZOI and others. Hardcastle's and Haywards' work proved foundational for the experimenters that followed including David Gordon-Smith, Chris Trask, Jim Kortge and many others.

Steder and Hardcastle's combined experience assessing and/or documenting crystal ladder design spans decades.

Their QEX article **Crystal Ladder Filters for All** may be legally downloaded from the ARRL website here.

Program download URL: http://www.arrl.org/qexfiles (select 2010, "11x09_Steder_Hardcastle.zip")

The QEX article describes Steder's Microsoft Windows™ program, methods to derive motional parameters, plus cites many important references. The main program calculates practical lower-sideband crystal ladder filters based on the exact equations published by M. Dishal in 1965. Hardcastle transformed these equations into a computer useable form in 1983 and Steder incorporated these equations into a modern, easy-to-use and interactive application with nice graphing and table displays.

For simplicity, the program assumes lossless crystals, however, the calculated values can easily be transferred into another simulation program such as GPLA to add or refine parameters such as loss resistance.

The main "Dishal" application calculates filters with Butterworth, Chebychev and constant-k (Cohn) properties and the so-called "QER" filter type by G3UUR (a low ripple version of the Cohn filter). Further; sub-programs in the top-level menu calculate xtal parameters (by both the G3UUR and 3 dB method), plus L-C impedance matching and ladder termination networks. An extensive help file well explains the program.

Iacopo Giangrandi — Introduction to Crystal Ladder Filters

Link: http://www.giangrandi.ch/electronics/crystalfilters/xtalintro.html

Iacopo (Jack) uses a transmission measurement to infer the motional parameters — inserting a series capacitance and measuring the series resonant frequency shift was also described in 1998 by Rolf-Dieter Mergner, DJ9FG.

His web article/applications provides what is likely needed by most builders — simple filter synthesis while avoiding expensive test gear. His filters plots/figures are spectrum analyzer measurements that I really like. Although his program can generate aymmetrical filters that some builders might not be used to, the frequency domain plots indicate proper function.
Giangrandi's filter design programs appear to be based on Jack Hardcastle's work and possibly content published in a paper by Patrick Magnin, F6HYE and Bernard Borcard, F3BB in Radio REF for April 1990.

I encourage you to try all the methods and applications mentioned to discover what works best for you. Don't lose heart, for characterizing crystals with a vector network analyzer is also a time-consuming endeavor and simple often = best on the QRP Workbench.

QRP — Posdata for August 2012 — Measuring Crystal Q

Prior to July 26, 2012, I could not measure Xtal Q. Why? I tried to measure crystal Q with the shunt-series tuned method (essentially the crystal acts as a trap and a step attenuator is used to calculate the insertion loss the xtal exhibits when centered in the notch) but failed because I could not precisely set the frequency with my homebrew L-C oscillator.

You really need a DDS or a Si570 based signal generator and preferably a spectrum analyzer to exact the measurement with the "trap method". The DDS is critical, while the SA only preferable — a power meter, or a 50 ohm terminated 'scope can work as the detector if a low-pass filter is placed just after the xtal.

On July 26, 2012 Wes, W7ZOI developed a simplified method and wrote a pdf file called **Simplified Measurement of Crystal Q** after feedback from myself and John Larkin about Q measurement difficulties without a digital-based signal generator. Unfortunately, this pdf file is no longer offered for download by W7ZOI. His method works well and I'm glad that as of now, I can *completely* characterize any crystal I own.

I present an experiment showing how I measured the Q of a 10 MHz crystal applying the new method developed by Wes, W7ZOI.

The crystal is evaluated as a N=1 low-pass filter resonated by a shunt capacitor at each end. I stuck with Wes' suggestion to try 1000 pF at 9 to 10 MHz. For lower frequency xtals he recommended trying larger value shunt capacitors. Just experiment with the shunt capacitance — if you use too high a C for a given crystal, your xtal low-pass filter bandwidth might get too narrow to measure with an L-C based signal generator.

The following diagrams employ 2 programs from the Ladpac software that ship with EMRFD

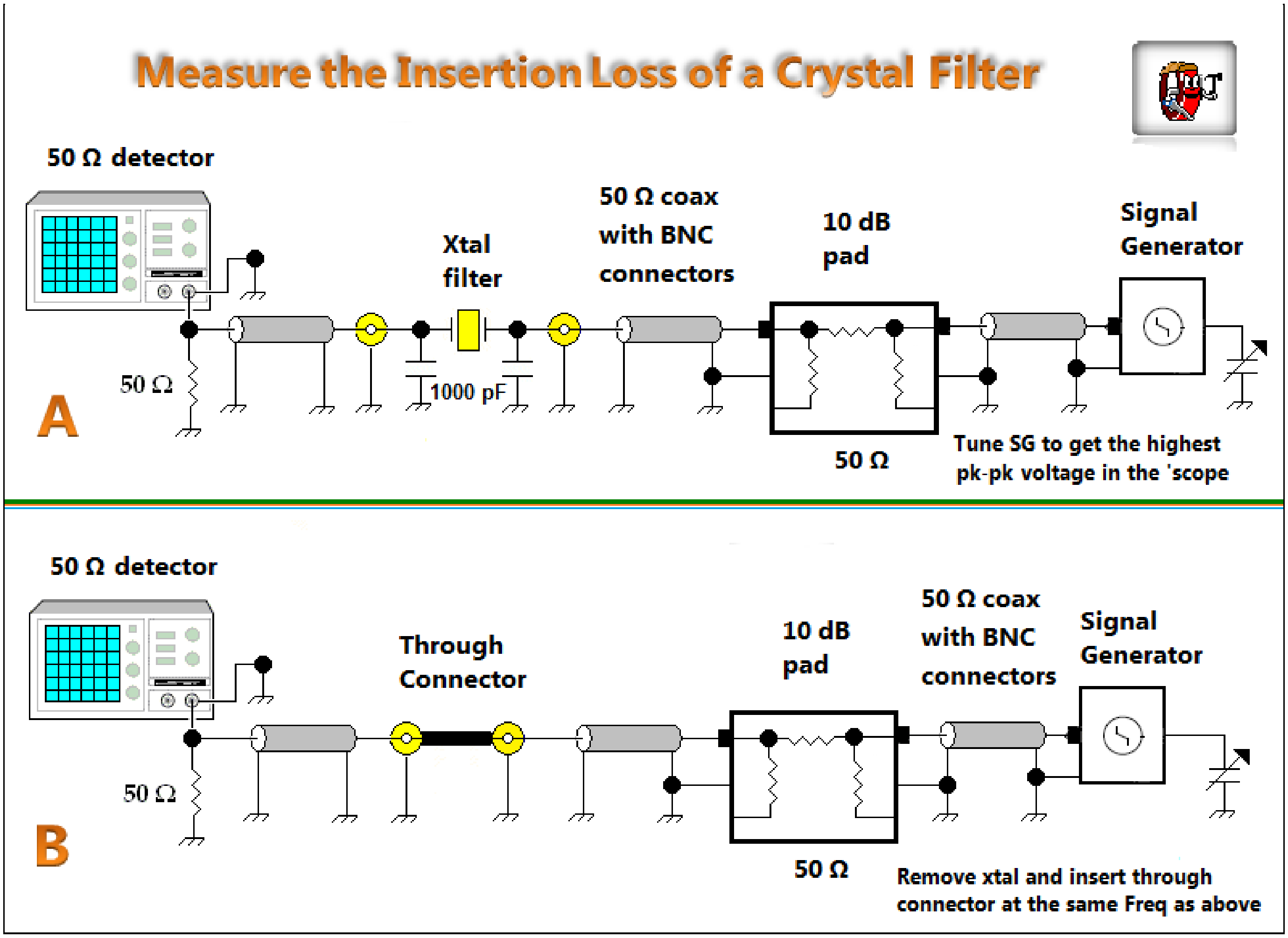

Measure the Insertion Loss of a Crystal Filter

50 Ω detector
Xtal filter
50 Ω coax with BNC connectors
10 dB pad
Signal Generator
50 Ω
1000 pF
A
50 Ω
Tune SG to get the highest pk-pk voltage in the 'scope
50 Ω detector
Through Connector
10 dB pad
50 Ω coax with BNC connectors
Signal Generator
50 Ω
B
50 Ω
Remove xtal and insert through connector at the same Freq as above

Above — Measurement of crystal filter insertion loss. In Part A, I carefully tuned my signal generator to get the highest peak-peak voltage in my 50 Ω terminated oscilloscope. I recorded this AC voltage as V Fil. In Part B, I removed the crystal filter board and replaced it with a BNC clad RF through-connector. I recorded this AC voltage as V Cal.

I discuss this standard method to measure the insertion loss or gain of a *device under test* in RF Workbench 2.

Even with the simplified method, you'll need a signal generator with good tuning resolution. My generator is shown on VFO 2011 as the 2.8-10.5 MHz Signal Generator. This is my version of the EMRFD Figure 7.27 generator.

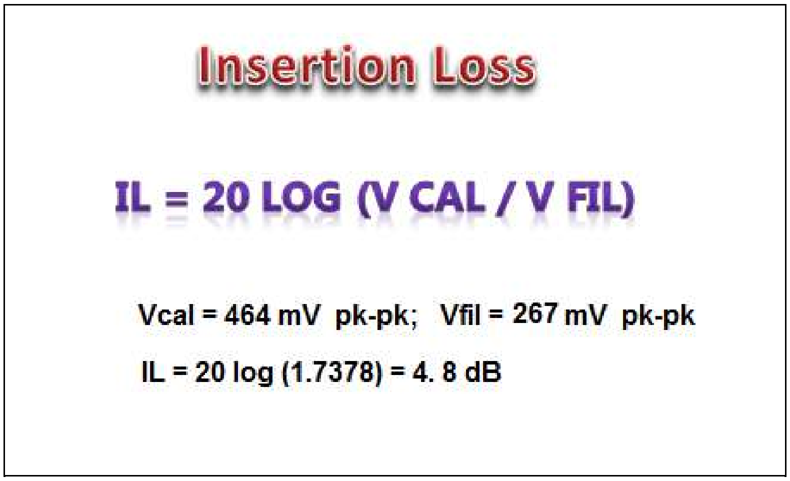

Insertion Loss

IL = 20 LOG (V CAL / V FIL)

Vcal = 464 mV pk-pk; Vfil = 267 mV pk-pk

IL = 20 log (1.7378) = 4. 8 dB

Above — The formula for insertion loss using peak-peak voltages. With my 10.0 MHz crystal, V Cal = 464 mV pk-pk and V Fil = 267 mV pk-pk. The insertion loss of my crystal = 4.8 dB.

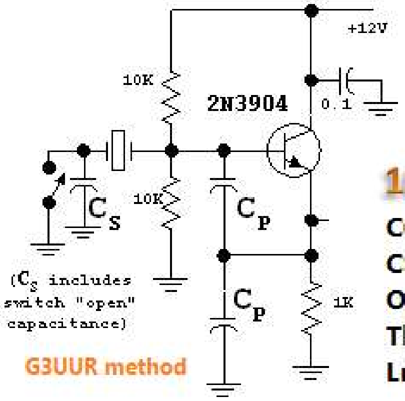
+12V
10K
2N3904
0.1
C_S
10K
C_P
C_P
1K
(C_S includes switch "open" capacitance)
G3UUR method

$$C_M = 2\cdot C_S\cdot\frac{\Delta F}{F}$$ Old Equation, p 3.19.

$$C_M = 2\cdot\left(C_S + C_0\right)\cdot\frac{\Delta F}{F}$$ Updated Equation

10 MHz Xtal

C0 = 4.87 pF

CS = open switch + cap = 37.4 pF

Open switch F = 9.998710 MHz

Thrown switch F = 9.996104 MHz

Lm = 0.0115 H

Measure C0

Calculate Lm

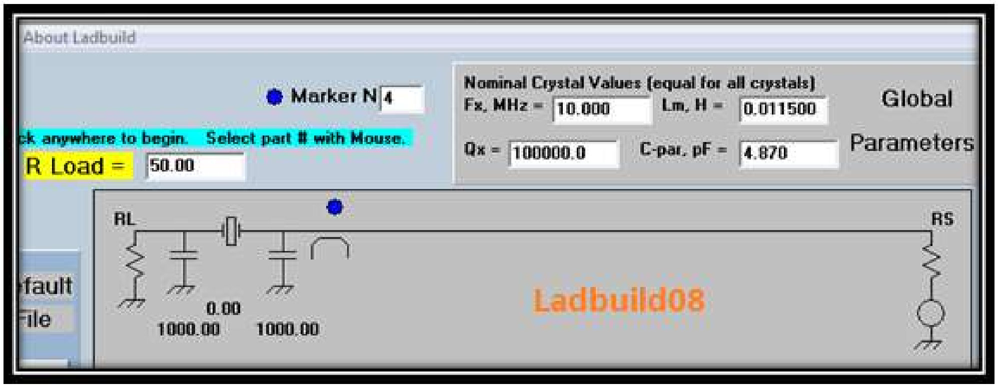
About Ladbuild

● Marker N 4

ck anywhere to begin. Select part # with Mouse.

R Load = 50.00

Nominal Crystal Values (equal for all crystals)

Fx, MHz = 10.000 Lm, H = 0.011500

Qx = 100000.0 C-par, pF = 4.870

Global Parameters

RL
RS
0.00
1000.00
1000.00
Ladbuild08
fault
File

Above — First I measured C0 (C-par) and then with the G3UUR method, calculated the motional inductance of my xtal. Finally, I entered all the needed variables into Ladbuild08 to make a filter to analyze in GPLA.

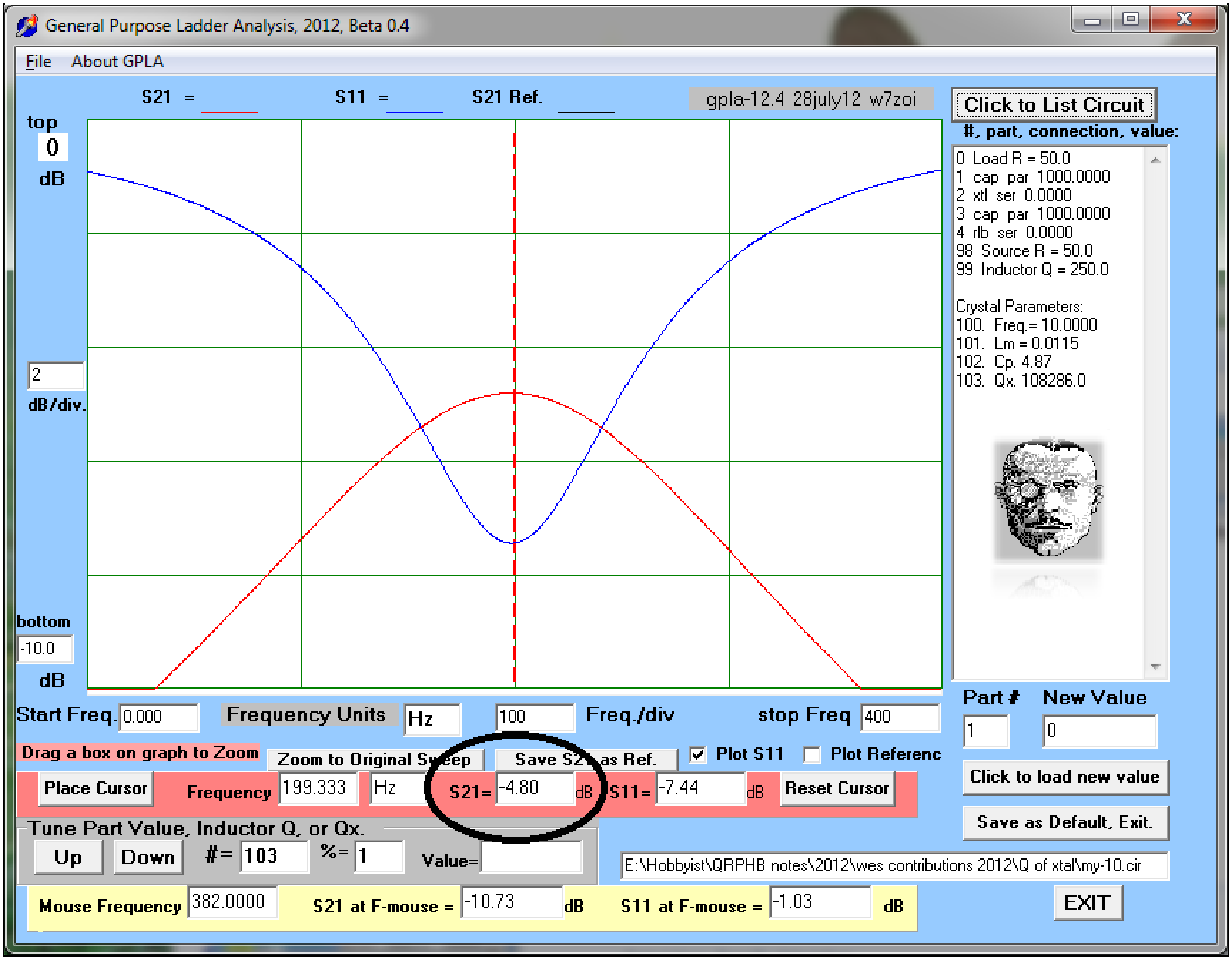
File
About GPLA
S21 =
S11 =
S21 Ref.
Click to List Circuit
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Place Cursor
Reset Cursor
Click to load new value
Save as Default, Exit.
Up
Down
EXIT

General Purpose Ladder Analysis, 2012, Beta 0.4

gpla-12.4 28july12 w7zoi

#, part, connection, value:

0 Load R = 50.0
1 cap par 1000.0000
2 xtl ser 0.0000
3 cap par 1000.0000
4 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 250.0

Crystal Parameters:
100. Freq.= 10.0000
101. Lm = 0.0115
102. Cp. 4.87
103. Qx. 108286.0

top 0 dB

2 dB/div.

bottom -10.0 dB

Start Freq. 0.000

Frequency Units Hz

100 Freq./div

stop Freq 400

Part # 1

New Value 0

☑ Plot S11 ☐ Plot Referenc

Frequency 199.333 Hz

S21= -4.80 dB S11= -7.44 dB

Tune Part Value, Inductor Q, or Qx.

#= 103 %= 1 Value=

E:\Hobbyist\QRPHB notes\2012\wes contributions 2012\Q of xtal\my-10.cir

Mouse Frequency 382.0000 S21 at F-mouse = -10.73 dB S11 at F-mouse = -1.03 dB

Above — With GPLA, I adjusted the value for *Qx* up or down until my *centered* S21 value indicated -4.8 dB (the insertion loss of my crystal determined earlier). My crystal Q = Qx = 108286.

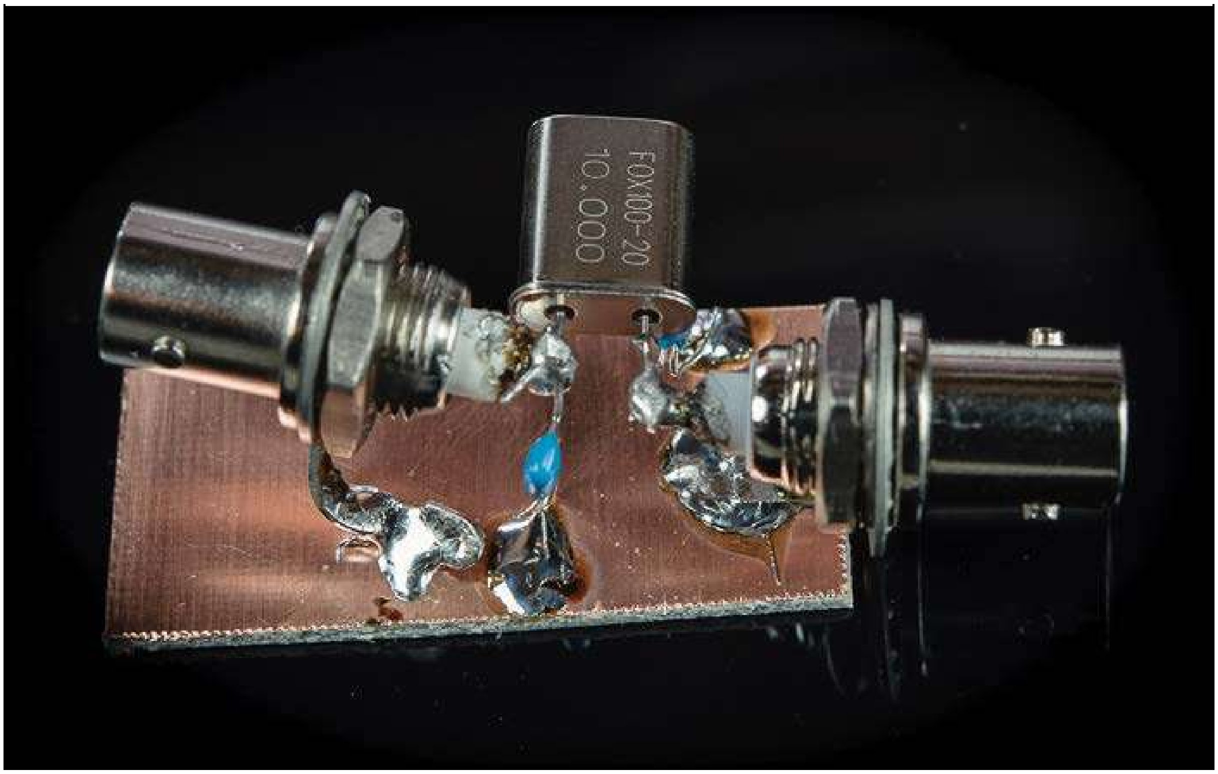
FOX100-20
10.000

Above — 10 MHz Crystal filter breadboard.

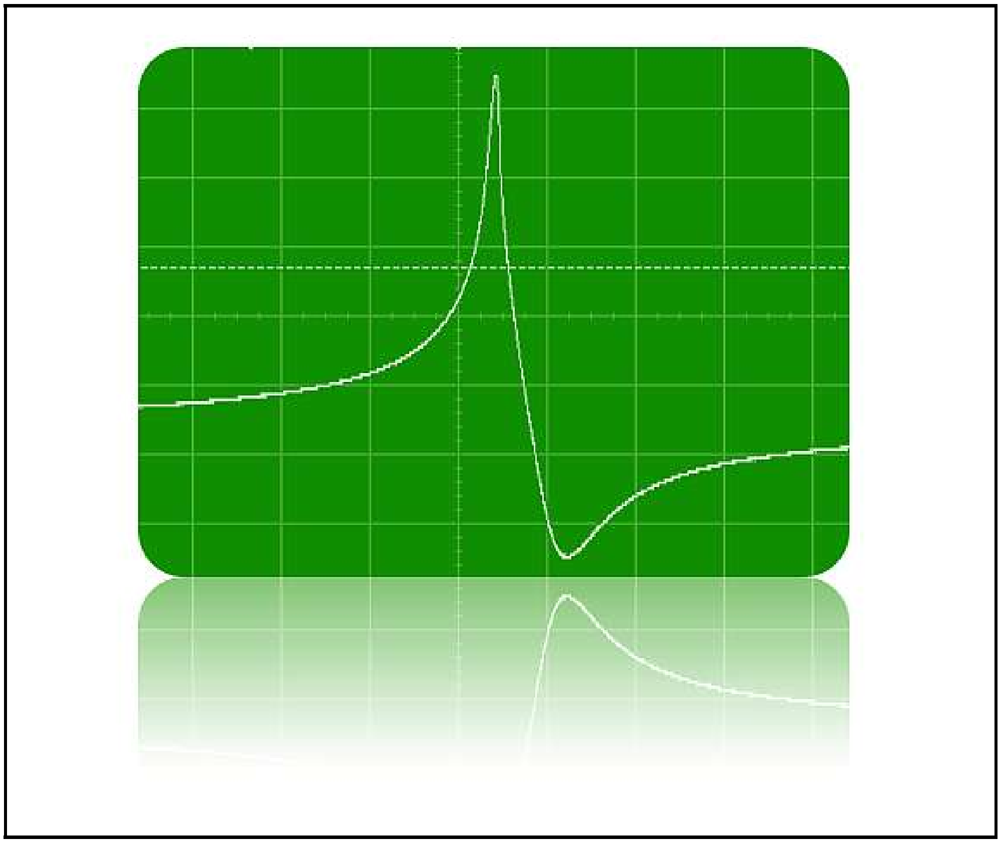

Above — A sweep of the 10 MHz Crystal filter used to determine the Crystal Q

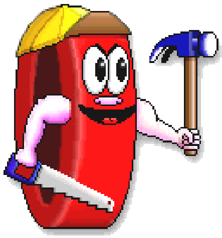

Electronic Hobbyist Circuits 2010

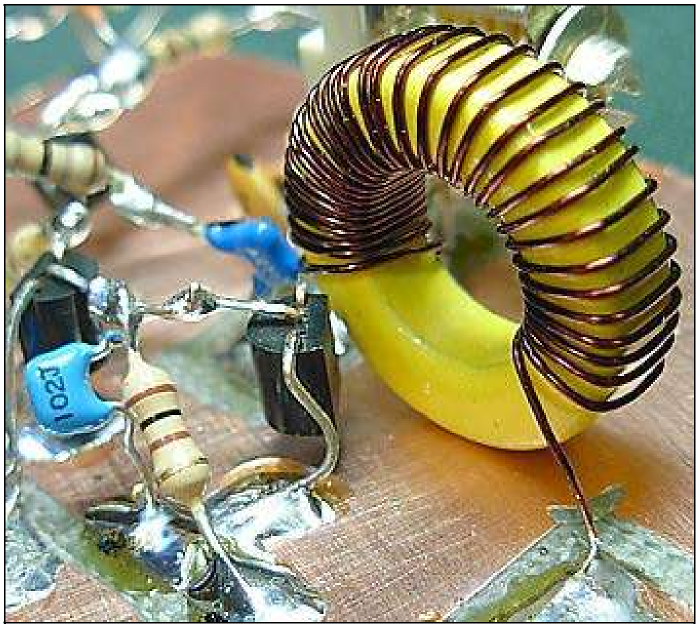

This page houses a collection of brief hobbyist experiments.

1. LM380 Power Examination

M9324
LM380N
BR

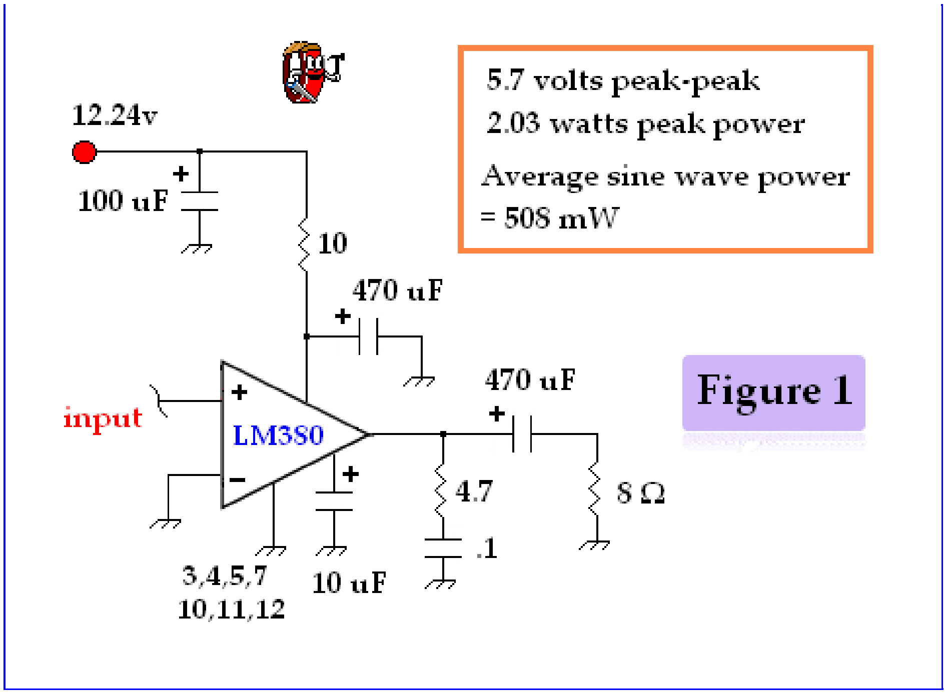
5.7 volts peak-peak
2.03 watts peak power
Average sine wave power = 508 mW
Figure 1
12.24v
100 uF
10
470 uF
470 uF
input
LM380
+
−
4.7
8 Ω
.1
3,4,5,7
10,11,12
10 uF

Figure 1 shows the test set up. This is a good part with an input impedance of 150kΩ. The gain is internally fixed at 34 dB. The average clean power was 508 mW. The test input frequency was 1018 Hertz.

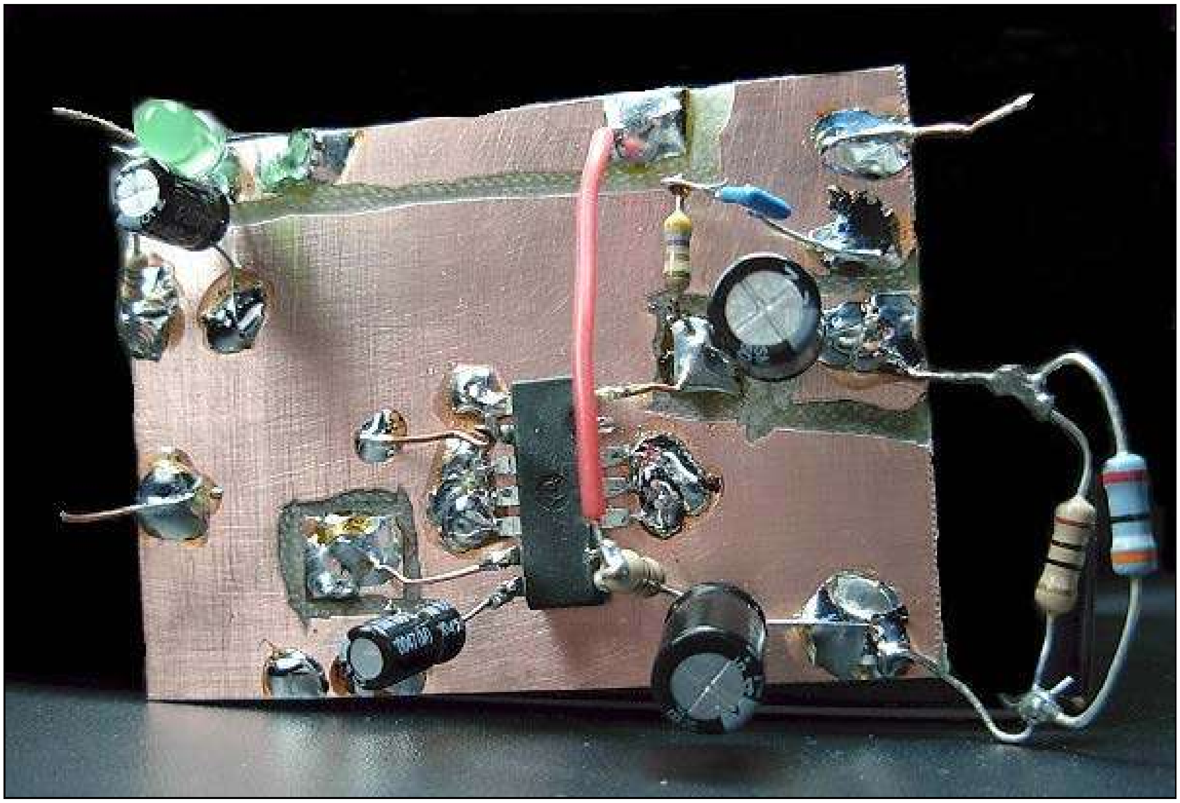

The breadboard of Figure 1 is shown above.

2. Wide Range L- C Oscillator

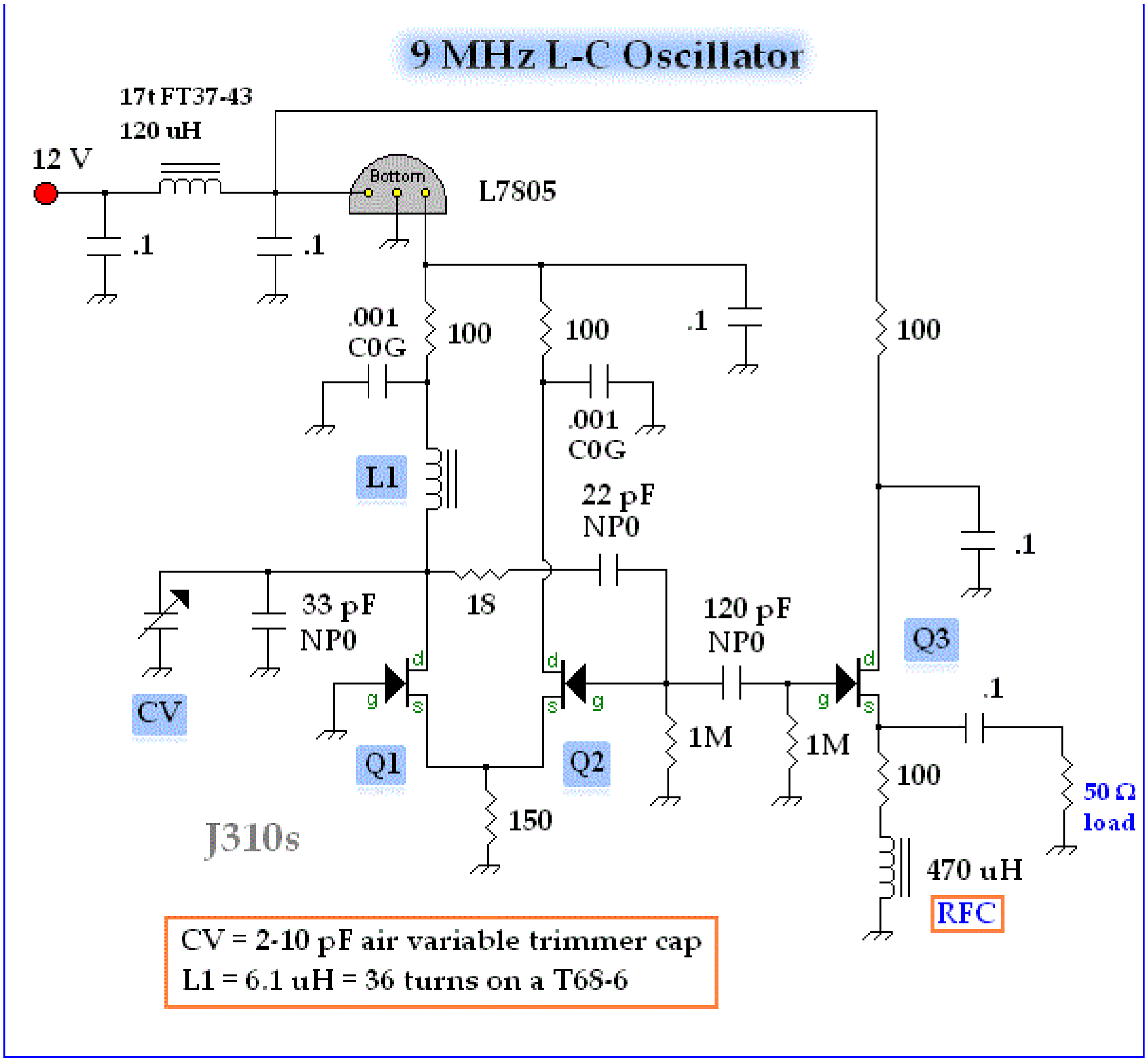
9 MHz L-C Oscillator
17t FT37-43
120 uH
12 V
.1
.1
Bottom
L7805
.001
C0G
100
100
.1
100
.001
C0G
L1
22 pF
NP0
.1
33 pF
NP0
18
120 pF
NP0
Q3
d
g
s
d
s
g
d
g
s
CV
1M
1M
.1
Q1
Q2
100
50 Ω
load
150
J310s
470 uH
RFC
CV = 2-10 pF air variable trimmer cap
L1 = 6.1 uH = 36 turns on a T68-6

Shown above is a single frequency version of a VFO topology which allows a wide frequency range when additional switched inductors and/or capacitors plus a tuning variable capacitor are used. One good usage example would be a to use such a VFO to drive a bridge to make a wide range antenna analyzer. Q1 is essentially a common gate amplifier. The source is driven and the output is taken off the drain. This FET exhibits no signal phase shift. Q2 is a source follower that is AC coupled through that 22 pF capacitor The 18 ohm resistor is used to kill UHF parasitic oscillations. The Q2 follower also has no phase shift. Connecting the output of Q2 back to stage Q1 gives zero phase shift. The L-C tank will select the frequency where 0 phase shift is obtained. The tank will show phases other than 0 away from its resonance.

Q3 is an AC-coupled source follower to further buffer the VFO from its load. The RFC can be anything from your junk box, although it should likely be low Q. The low-pass decoupling filter on the the 12 volt supply path can also be anything reasonable. I wound mine using 17 turns on an FT37- 43 ferrite toroid. Its purpose is to keep RF from traveling down the 12 volts DC voltage wire to other parts of your circuit.

Any component connected to the L-C tank (at the Q1 drain, or the cold end of L1) can affect VFO tuning and drift. Temperature compensation will be necessary to achieve perfect stability. I use NP0 and C0G caps interchangeably. In the design shown, stability was good and the output had low measured distortion. This VFO will pretty much oscillate with any reasonable L and C values in the tank circuit. I found frequency stability was a little better with a higher L to C ratio. This is a great experimenter's circuit. One version built oscillated at 150 MHz.

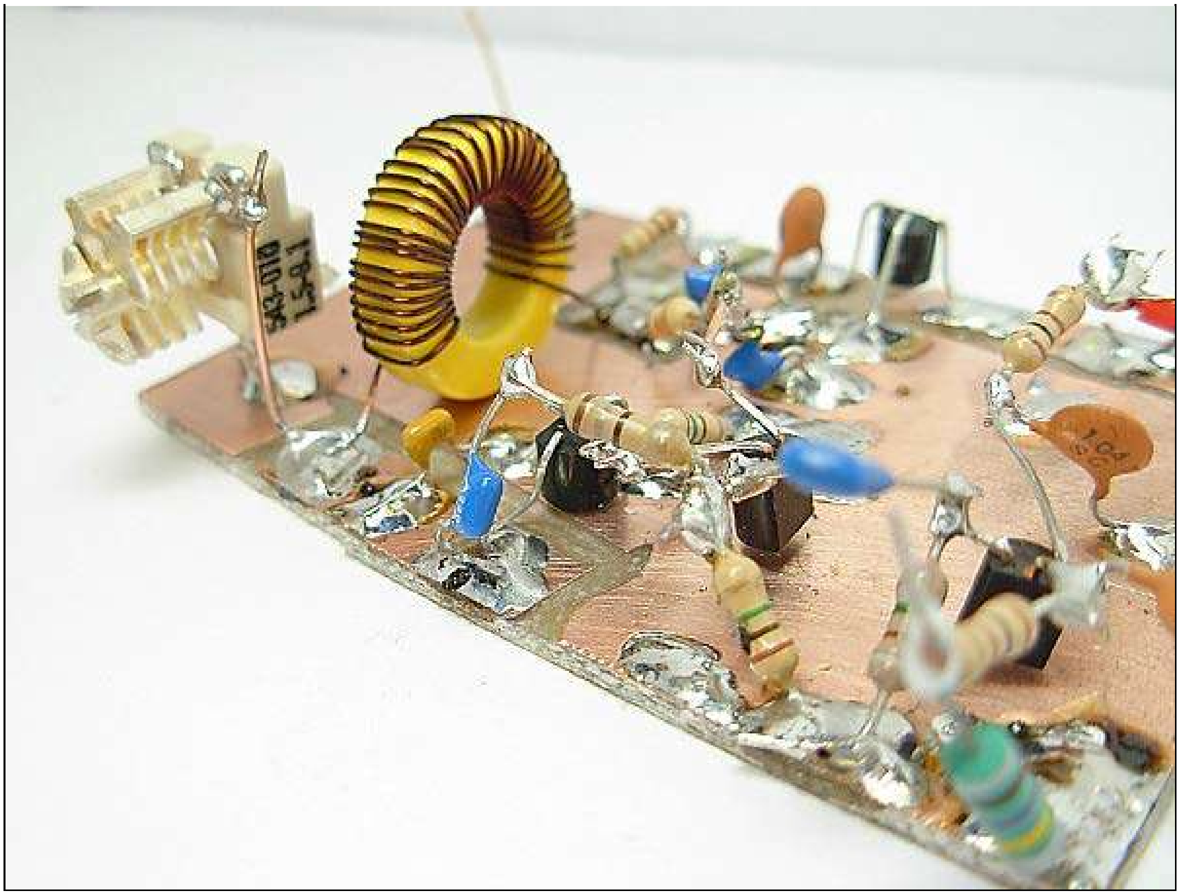

The breadboard of the above schematic. Pull the wire on your #6 powdered iron toroids tight to prevent air gaps between the toroid and the wire. Number 26 gauge wire was used on L1 as shown. High Q tank parts will garner the best results.

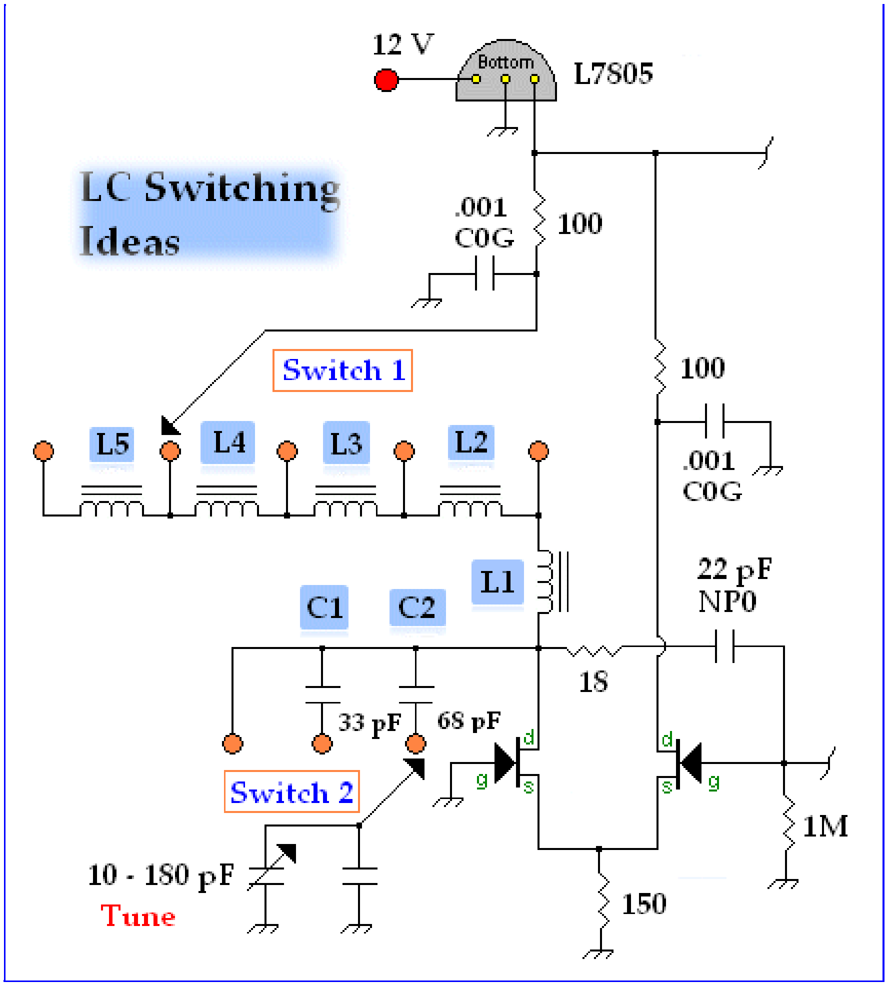
12 V
Bottom
L7805
LC Switching Ideas
.001
C0G
100
Switch 1
100
L5
L4
L3
L2
.001
C0G
L1
22 pF
NP0
C1
C2
18
33 pF
68 pF
d
g
s
d
s
g
Switch 2
1M
10 - 180 pF
Tune
150

Some potential switching ideas are presented above. The builder is in total control of the tuning range and must calibrate the L and C values according to needs and the parts on hand. Output power will vary according to the L-C ratios and some designs include automatic signal amplitude leveling and/or RF gain controls.

3. FET Matching

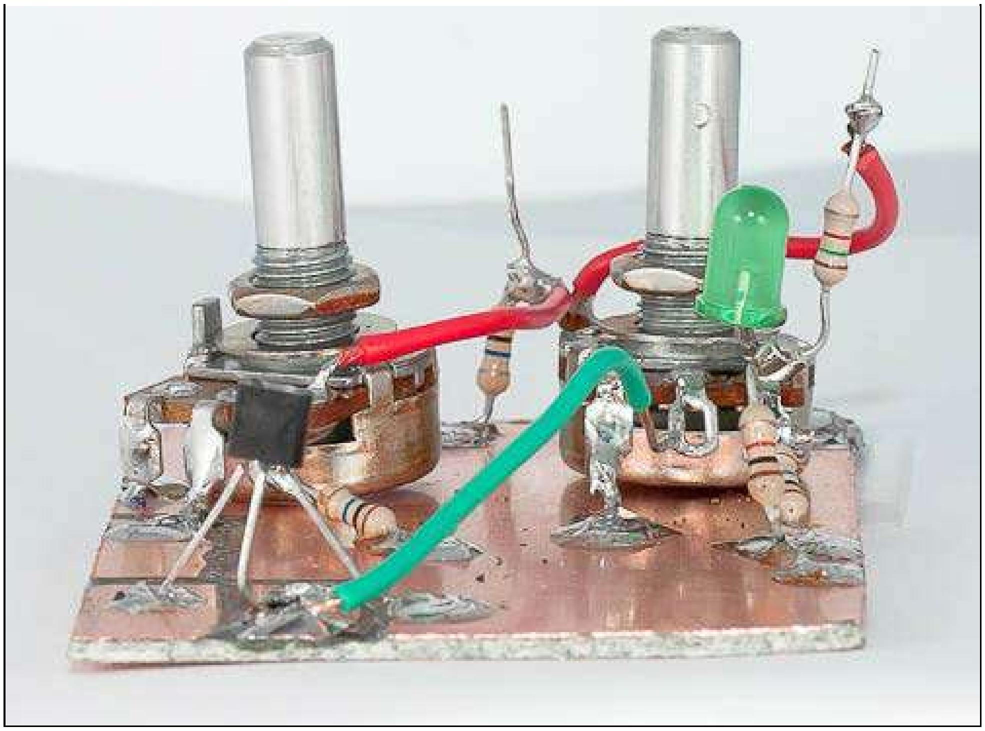

I find matching high IDSS FETs like the J310 to be a pain. I generally matched them for IDSS and occasionally for IDSS and VP. Observations that when the IDSS of 2 or more FETs match, their pinch-off voltage (VP) also matches, led me to not measure VP. In addition, the variability of VP measurements causes me distress. Click here for a tutorial if you don't understand the terms IDSS and VP.

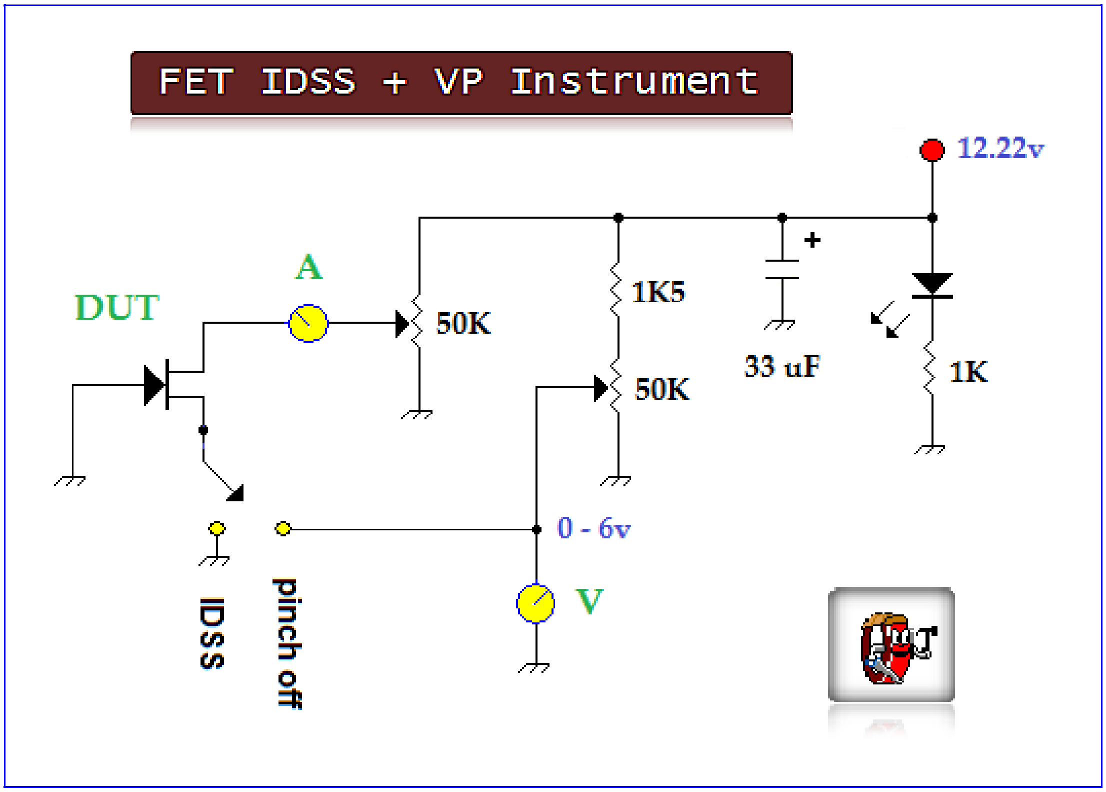

Above — The device I use to measure IDSS and VP. From Ken Kuhn's web site.

Conceptually IDSS and VP aren't difficult to understand — measuring them is another story. With the above device, first IDSS is measured; the final drain voltage potentiometer setting is left and then I measure VP. While measuring IDSS in high IDSS FETs, heating can occur and you may actually see current start to drop as you increase the drain voltage (negative temperature coefficient). On J310 specification sheets, the manufacturers state they pulse the current during measurement to prevent heating. While performing IDSS measurements, I am fearful of destroying the FET I am trying to characterize! Measuring VP is also problematic.

I have tried 3 methods to quantify VP:

Adjust the 0 - 6 volt supply until I think the current goes to 0. Serial measurements 1 day apart can vary by a variation of as much as 0.5 volts; it's quite subjective.

Adjust the 0 - 6 volt supply and measure the gate voltage that produces a drain current somewhere between 0.1 and 1 percent of IDSS and declare that to be VP.

Adjust the 0 - 6 volt supply so the ammeter reads ¼ IDSS and multiply this voltage by - 2.0. Refer to Ken Kuhn's site for details. Although reasonably accurate, the second order math is only a rough approximation — the real math is impossible to do by hand as it involves fractional exponents and these exponents and other factors vary as a function of the physical JFET geometry.

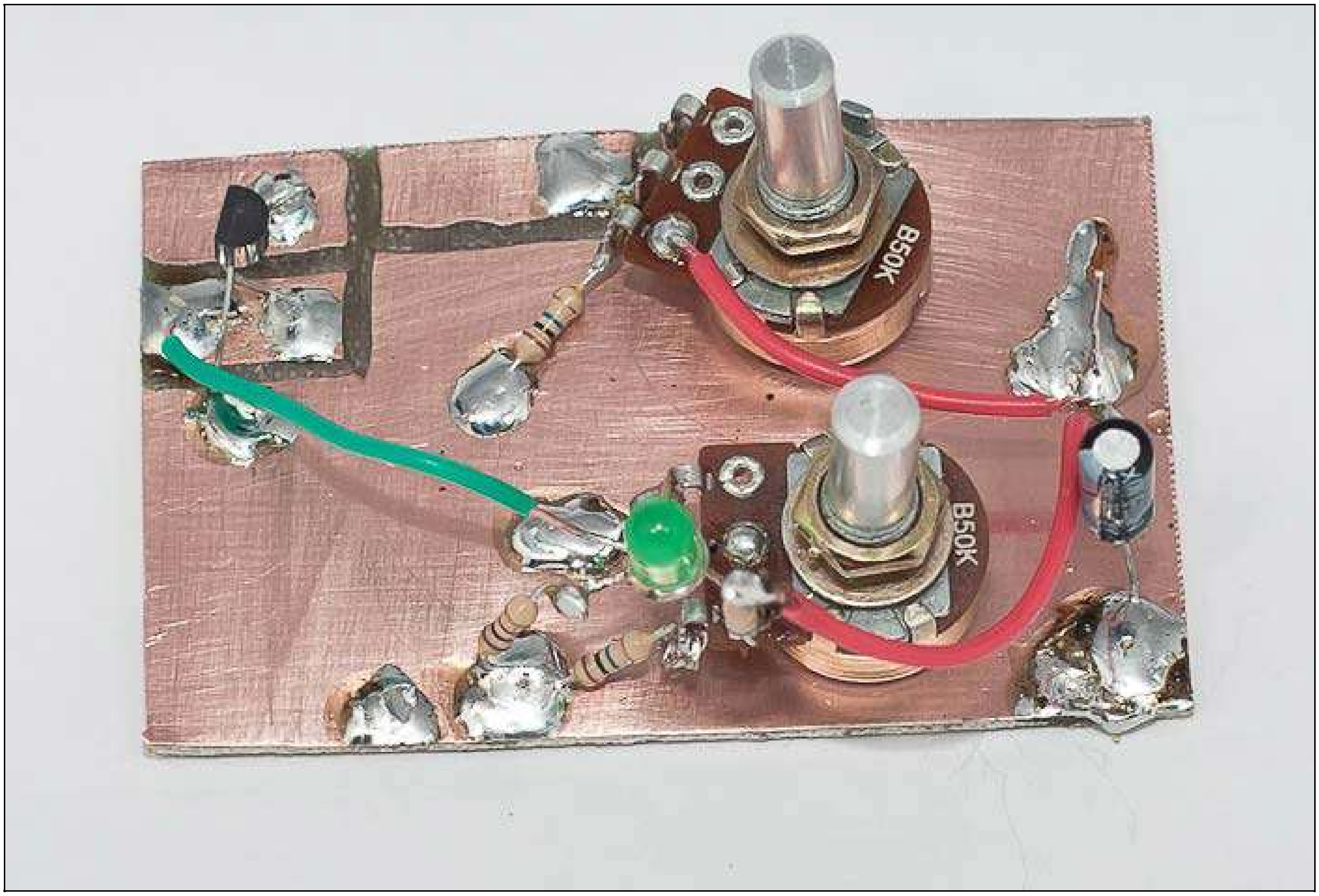

Above — The breadboard of the device I use to measure IDSS and VP. There is no actual switch, I either ground the green wire to the copper board, or tack solder it to the 0 - 6 volt potentiometer wiper. 10 megohm resistors plus the pot ground wire anchor each pot to the copper clad board.

All 3 methods to quantify VP frustrate me. There must be a way to match J310s or other FETs without characterizing them. I frequently collaborate with readers to problem solve and learn. A potential solution contributed by a supportive reader follows:

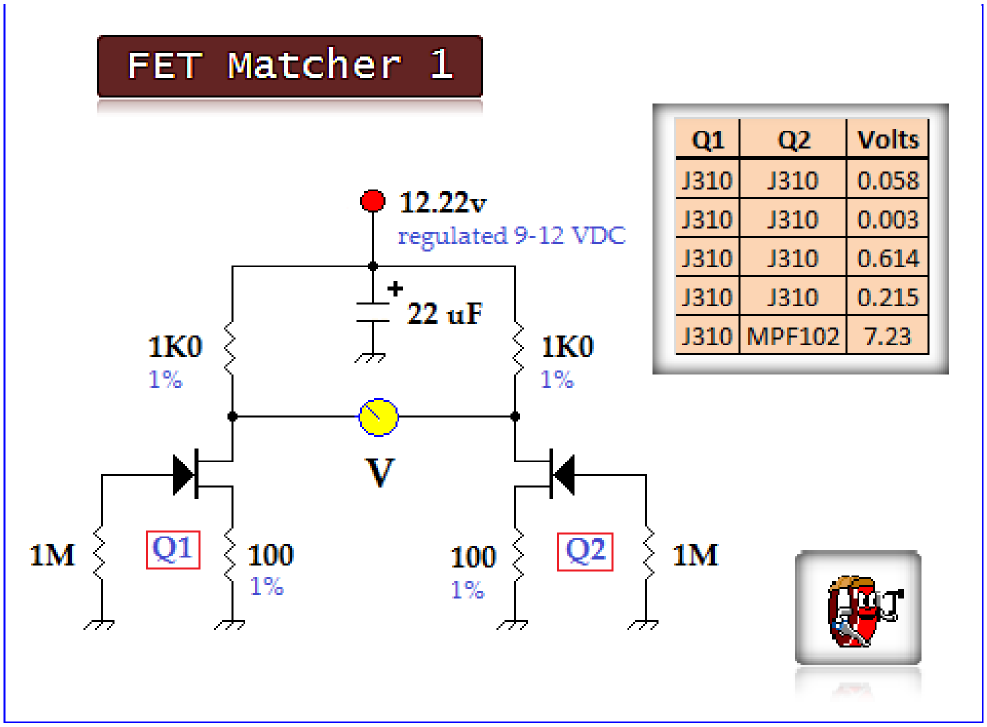

Q1	Q2	Volts
J310	J310	0.058
J310	J310	0.003
J310	J310	0.614
J310	J310	0.215
J310	MPF102	7.23

Above — A bridge is used to match a pair of JFETS. It's often best to match devices in a circuit that closely resembles the one that you intend to use them in. The differential output of each drain is measured by placing a DVM lead on each drain and recording the voltage. Generally, I stick a FET in the Q2 slot and put FETs from my parts bin in the Q1 slot to match it. The results of 5 different FET pairs are tabled above. A match <= 50 mV is probably acceptable and in 1 case, I found a match of 3 mV! You can match 1 FET with many using this device.

Note the poor match when an MPF102 and a J310 were tested. 1% tolerance resistors are recommended for the bridge.

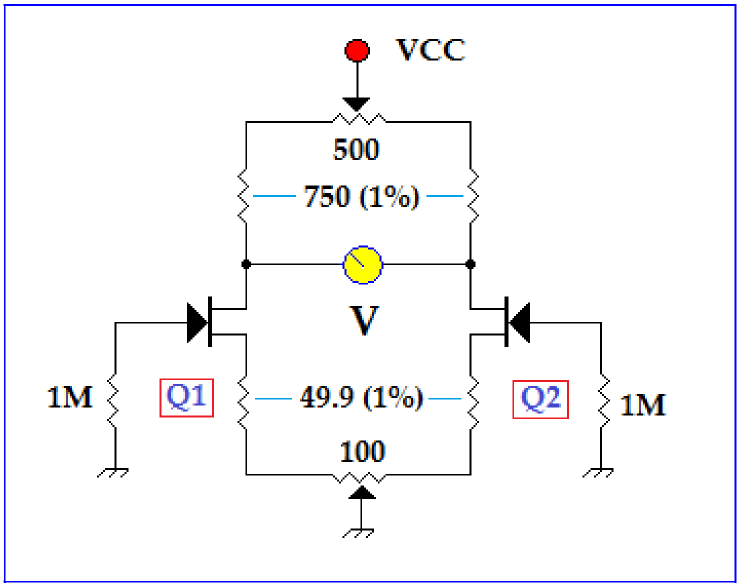

Above — A set and forget precision bridge using trimmers to establish a perfect DC match on both halves. If you don't have 1% parts, the trimmer resistors offer a solution. You can place a trimmer at either the drain or source end as shown and just use 5% resistors. Calibrate each half of the bridge with your ohm meter. I cover bridges on this web page if you need more information on them.

Above — A differential FET matcher breadboard. This version had a 9.1M gate resistor on Q1 by accident, although it made no difference to the experiment, as no DC gate current flows.

4. BJT and Diode Matching

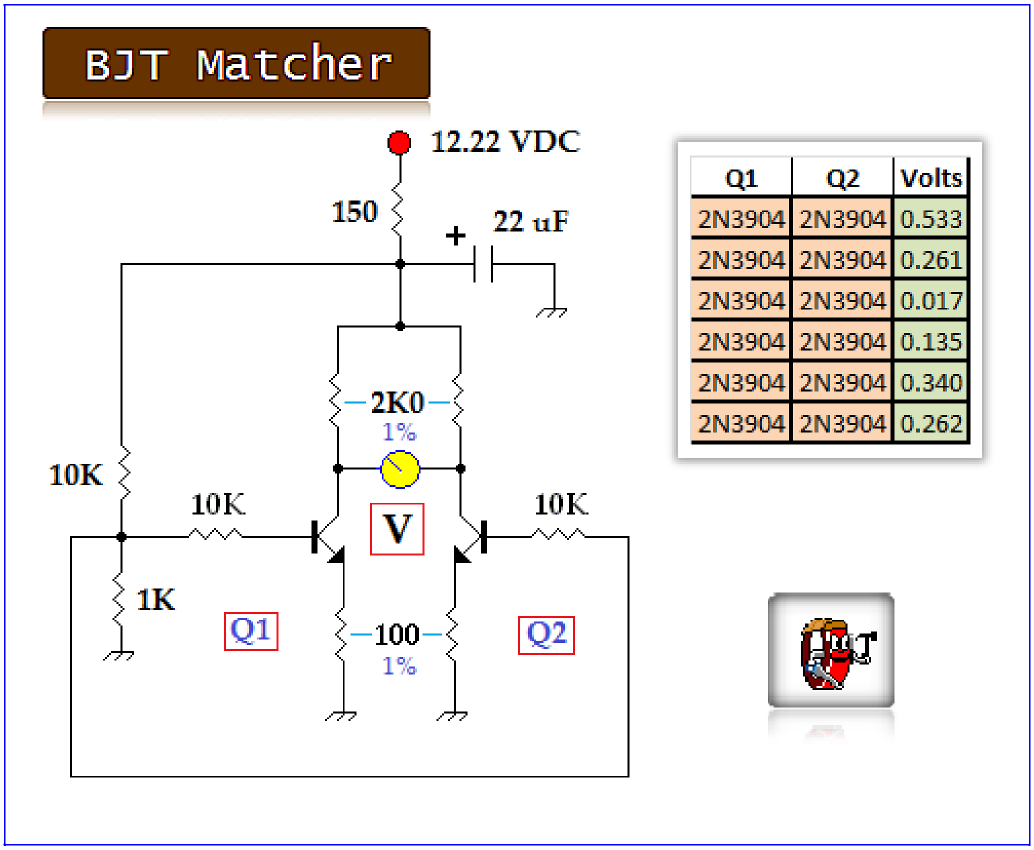

Q1	Q2	Volts
2N3904	2N3904	0.533
2N3904	2N3904	0.261
2N3904	2N3904	0.017
2N3904	2N3904	0.135
2N3904	2N3904	0.340
2N3904	2N3904	0.262

Above — Differential BFT matcher. Differential voltage matching works for bipolar junction transistors and diodes also. PN junctions are thermally sensitive — let them stabilize before testing. I measured 6 2N3904 transistors and the tabled results remind us why this transistor isn't the best choice for matched BJT circuits.

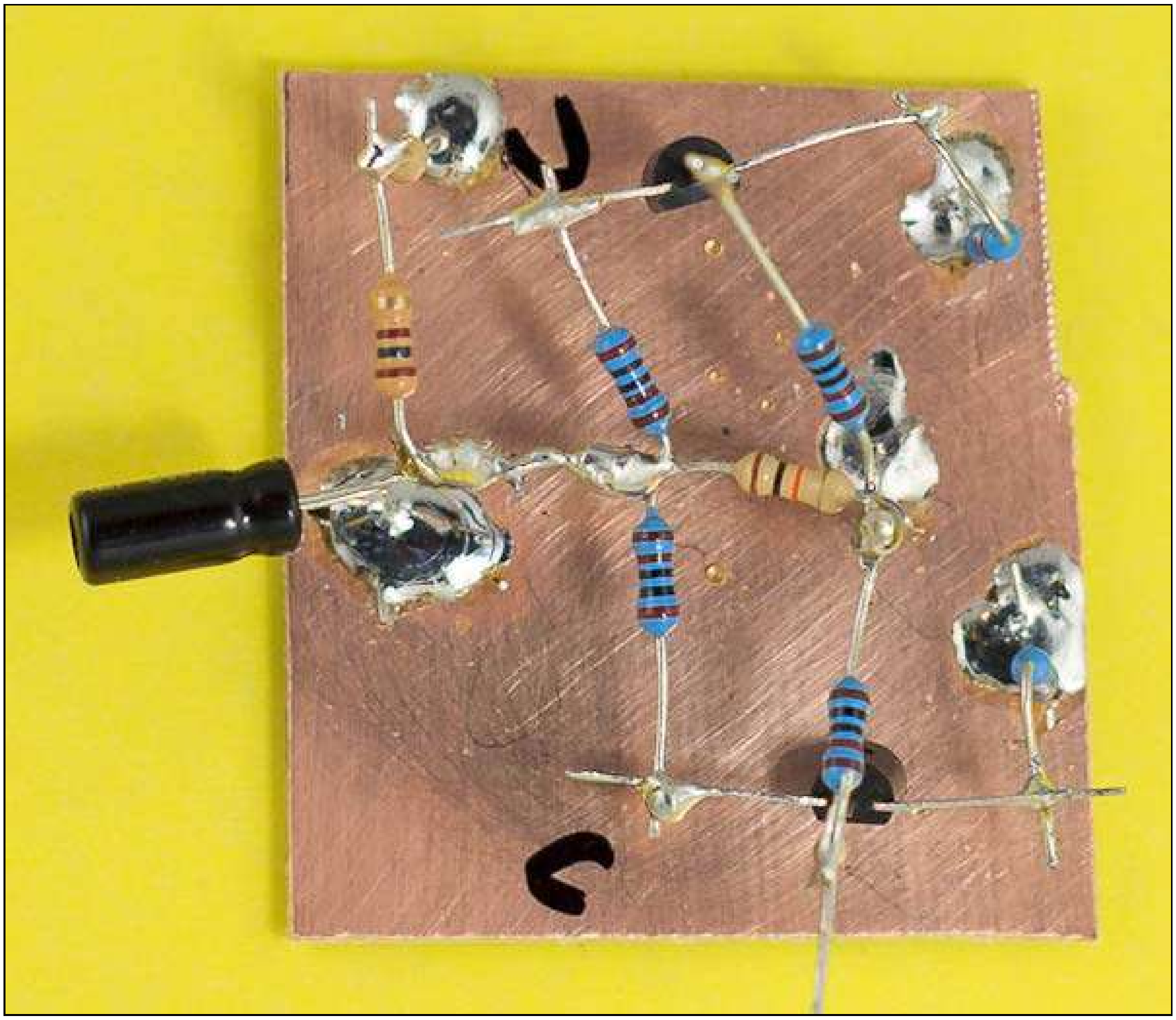

Above — The breadboard of the differential transistor matcher. Any reasonable resistor values will work fine, but don't run the current too high — my circuit has a nominal emitter current of 2.25 mA.

Differential Diode Matcher

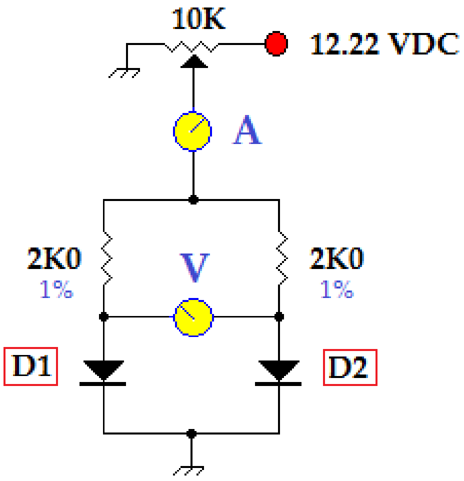

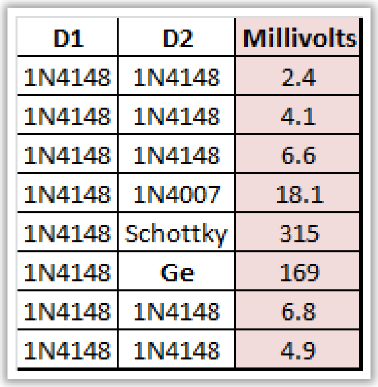

D1	D2	Millivolts
1N4148	1N4148	2.4
1N4148	1N4148	4.1
1N4148	1N4148	6.6
1N4148	1N4007	18.1
1N4148	Schottky	315
1N4148	Ge	169
1N4148	1N4148	6.8
1N4148	1N4148	4.9

Measurements with each diode current @ 77.3 ųA

Above — The differential diode matcher. The 10K pot allows you to vary the current to suit your needs. The ammeter reads double the diode current. Considering only 1 diode; the current ranges from 0.01 to 5.82 mA. You can drop the 1% metal film 2K0 resistors down as low as 470 ohms or so if you need serious current. Tabled are some measurements performed with the 10K pot dialed to give 0.636v (77.3 uA per diode). This provided excellent sensitivity. Measurement was performed on my DVMs 200 mV scale. Builders might experiment with diode current to assess measurement sensitivity and linearity or even to match the diodes across a range of current values.

Clear glass passivated diodes can be affected by light — photons will pass through the glass and knock electrons through the barrier. Ensure each D.U.T is exposed to the same amount of ambient light. Some microwave detector diodes can be damaged with as little as 1 mA of current. Replace the 2K0 resistors with 10-15K resistors, or lower VCC to prevent damage .

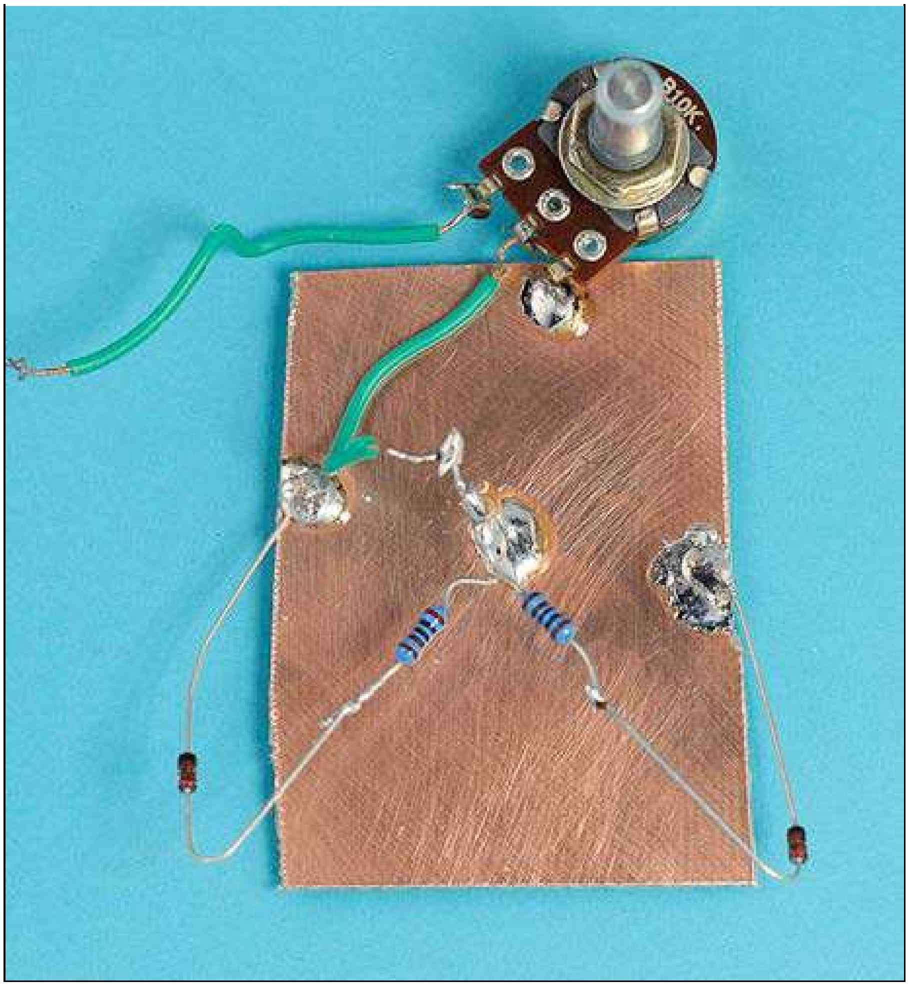

Above — Breadboard of the differential diode matcher. The 2K0 resistors are suspended by a vertically mounted 10M stand-off resistor.

5. Bipolar Junction Transistor Beta Tester

Bipolar Transistor Beta Tester

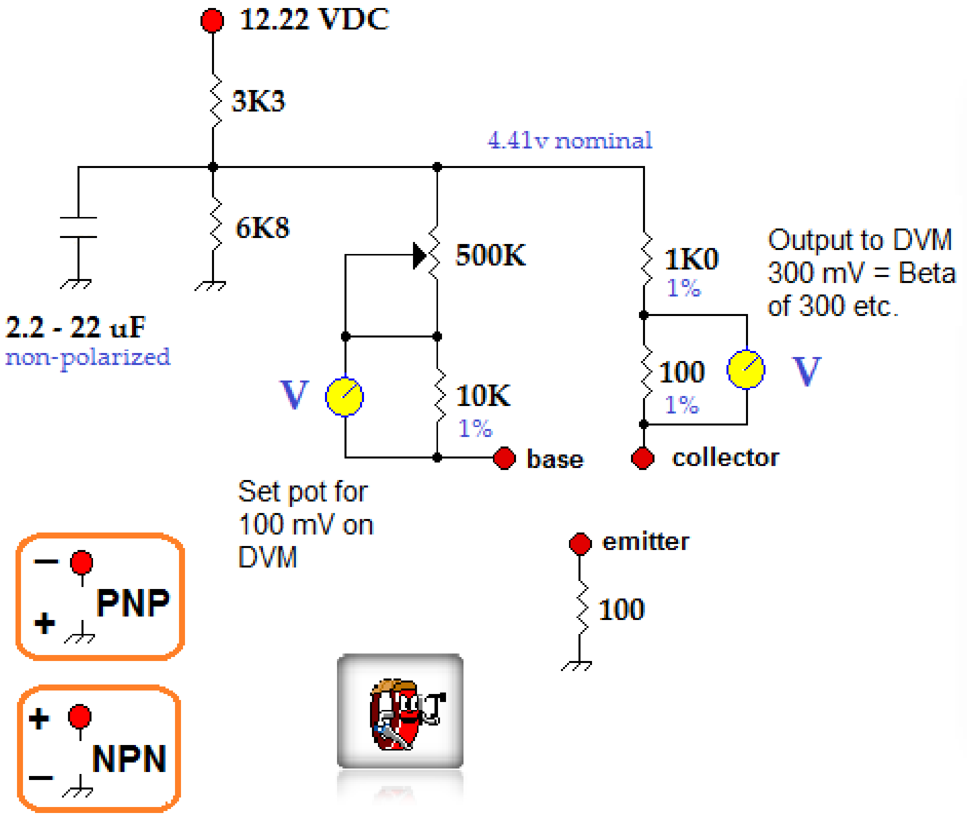

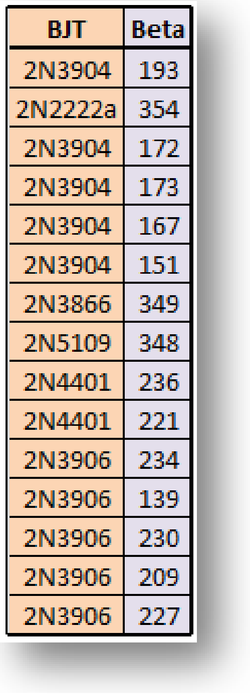

BJT	Beta
2N3904	193
2N2222a	354
2N3904	172
2N3904	173
2N3904	167
2N3904	151
2N3866	349
2N5109	348
2N4401	236
2N4401	221
2N3906	234
2N3906	139
2N3906	230
2N3906	209
2N3906	227

Above — Disappointed with the transistor beta testers in our common, low-cost digital multimeters, we did the logical thing; designed and built our own. This collaborative project was more an experiment with BJTs than anything else. It's about as simple a beta measurement device as you can make and still get good results. Preventing damage to our parts inventory underpins this design — the 100 Ω emitter resistor plus ~ 10 microamps of base bias keeps the IC low to help avoid smoke since most new small signal transistors have a beta of 100-400.

Ensure the correct polarity for PNP versus NPN transistors. The voltage divider targets 5 volts using a standard ~12 volt supply; I just used whatever resistors were handy and ended up with the 6K8 — 3K3 pair. VCC should be regulated. Perform the measurements with a single multimeter allowing time for stabilization.

To use: Set the potentiometer so that the voltage drop across the 10K resistor is 100 mV. Then move your DMM leads to the 100 Ω resistor and measure the beta. This device measures beta, the static gain at DC.

Measuring beta is a bit inexact since beta is affected by so many variables as follows:

Beta tends to be low at low operating currents and rises and plateaus for medium currents and then falls at higher currents.

Beta tends to increase with temperature.

Beta is affected by the voltage between the collector and emitter -- this is a weak effect except when the voltage is very small.

The beta can vary as the battery depletes in DMM beta testers.

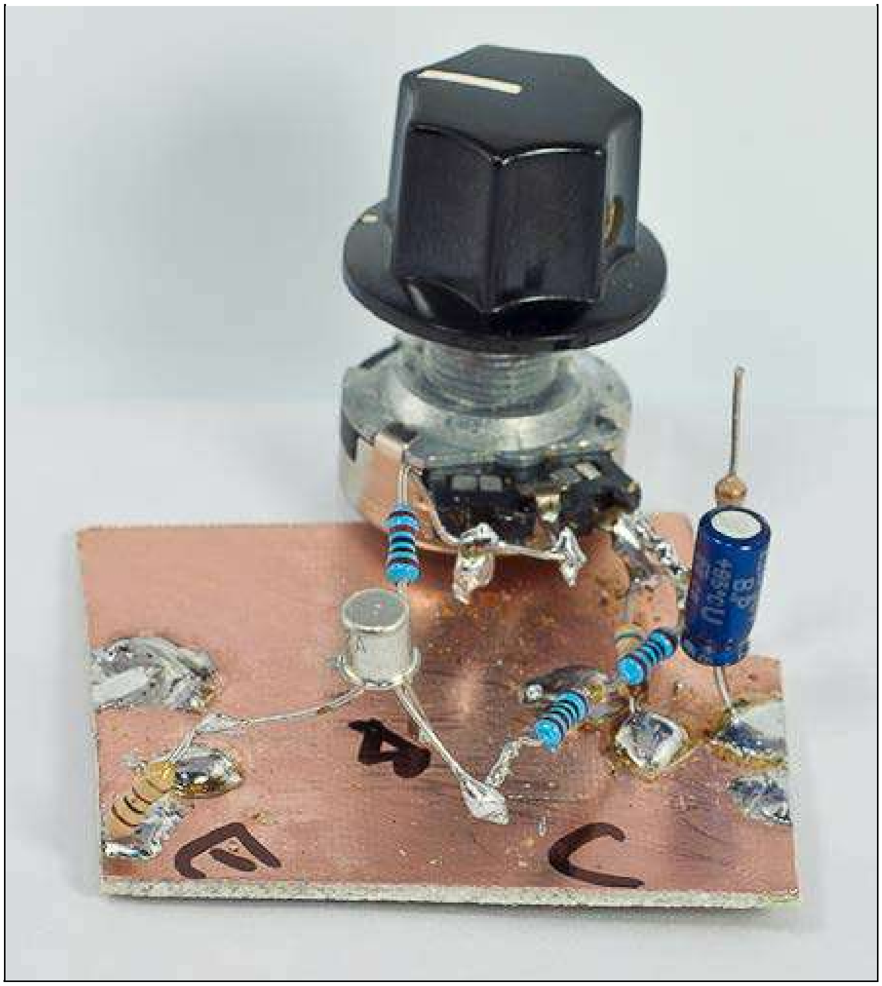

Above — Breadboard of the QRP beta tester. We hope you have as much fun with this circuit as we did.

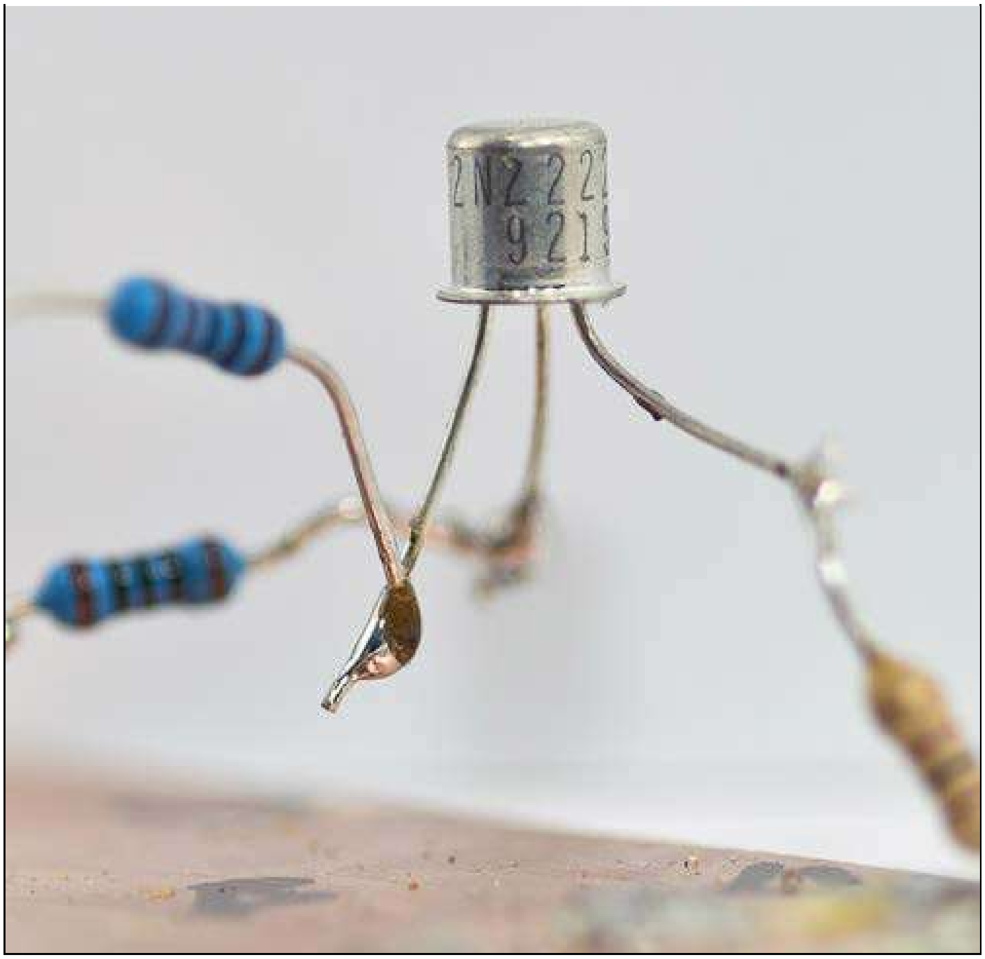

Above — Testing a 2N2222a. We found measuring transistors to be very instructive; comparing our results to specification sheets, handling the different BJTs, gaining experience — all part of the exciting world of bench electronics. Such circuits, although simple, are great for both learning and design.

6. FET Tester - VGS Measurement Tool

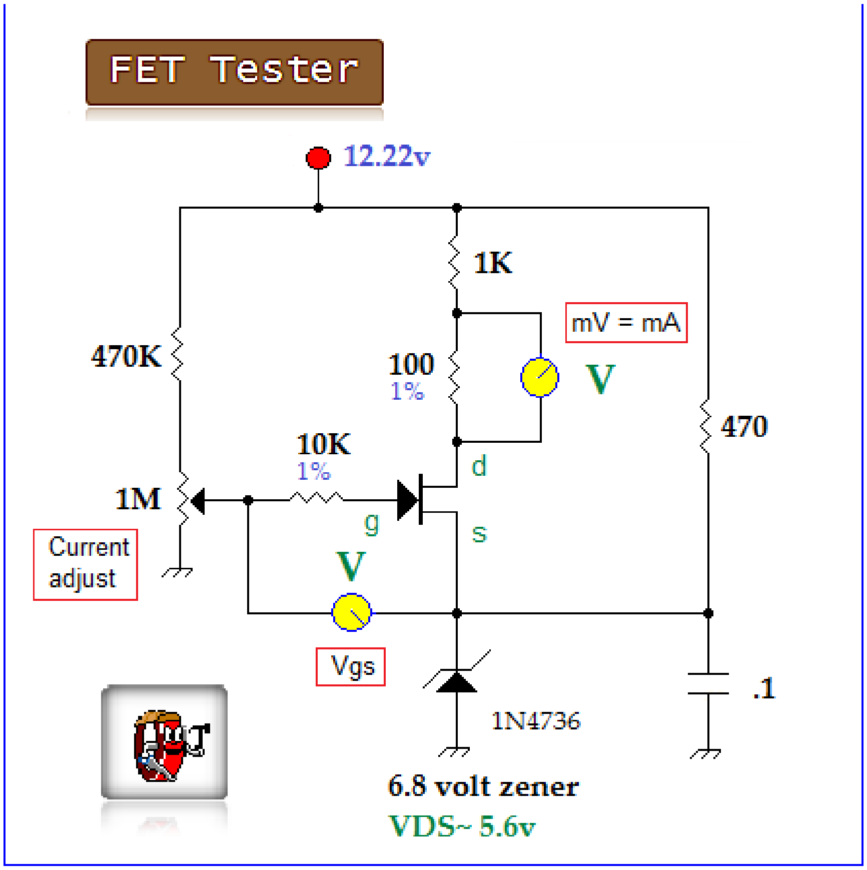

Above — A FET VGS Tester. The final collaborative output from late December 2010: a device to examine VGS. A zener diode value from 6 -7.5 volts or so should work okay. VDS should be greater than the nominal VP for greatest accuracy. The VDS using 7.5 volt zener diode was 4.85 volts. The range of VGS on my breadboard was roughly 6 to -6 VDC without the 470K range limiting resistor; it's up to you if you want to limit the VGS range.

Using the mV scale on your DVM, potentiometer adjust the current until the meter just reads 0. Then move your leads to the VGS test points to read the pinch-off voltage.

The 0 current point will only ever be close. The interesting issue with the 0 point as defined is it never goes to true zero — just zero enough — that point may be hidden if your meter scale truncates lower current measurements to zero. If you're able to repeatedly get consistent VP measurements with whatever method your using, then it's likely accurate.

We thought about measuring gate leakage current across the 10K gate resistor, but accurate measurement is impossible with this device. The gate leakage current is going to be in the low nanoampere region at the very highest unless the JFET is bad. The expected voltage across the 10K resistor would be a few 10s of microvolts. A better way to measure gate leakage is to use the 10M resistance of a DVM instead of this resistor. The DVM then acts as a current meter by measuring the voltage across the 10M resistance. A reading of 200 mV would mean a current of 20 nA.

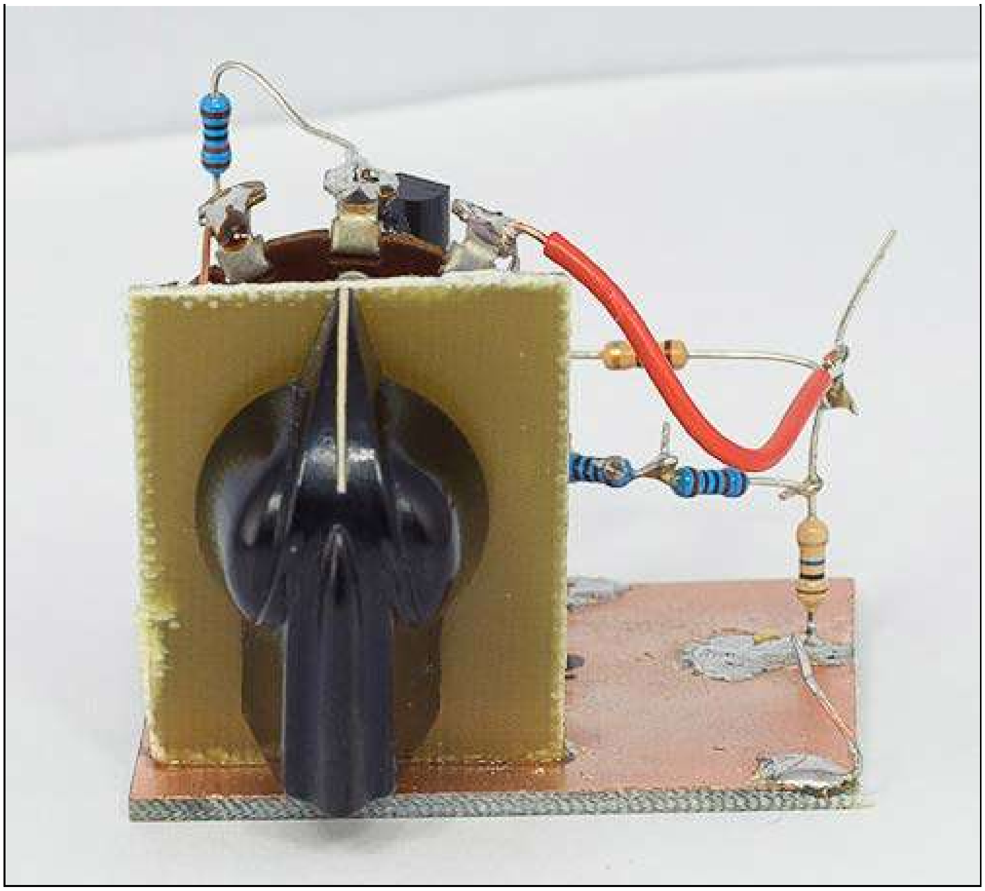

Above — Front view of the FET tester with a "chicken head" knob on the potentiometer.

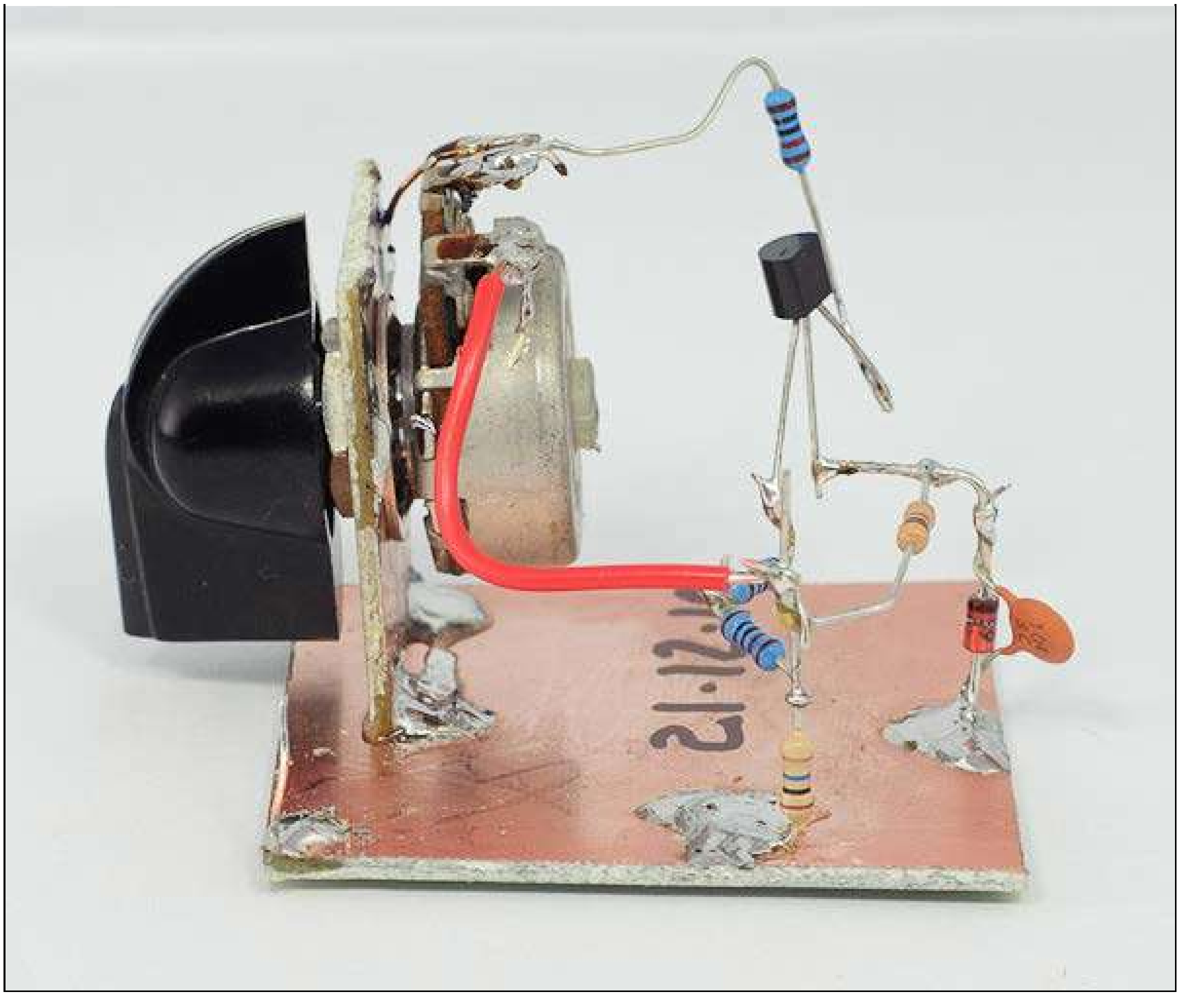

Above — Measuring a J310 with the FET tester prototype breadboard.

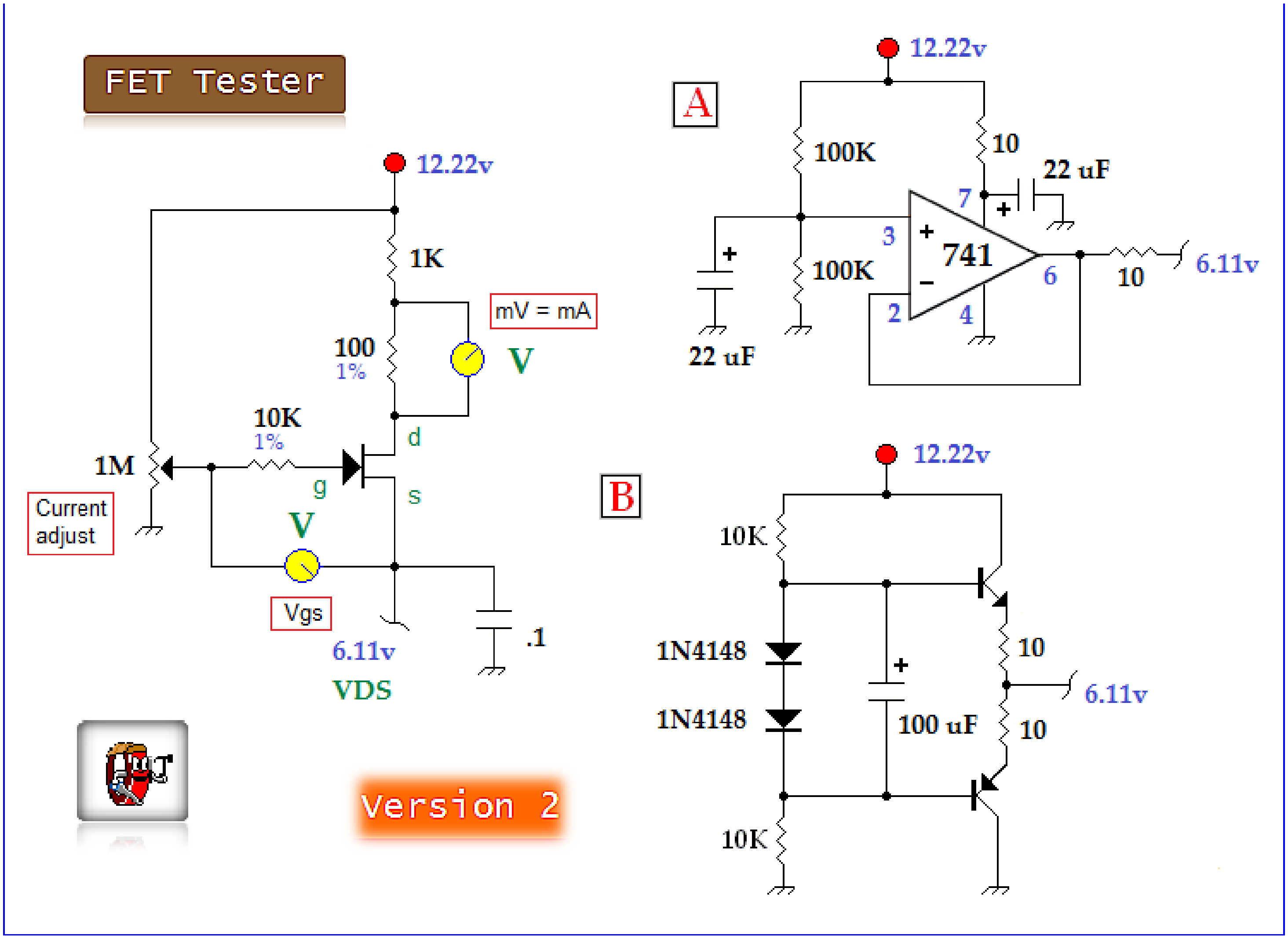
FET Tester
12.22v
1K
mV = mA
100
1%
V
10K
1%
d
1M
g
s
Current adjust
V
Vgs
6.11v
VDS
.1
Version 2
A
12.22v
100K
10
22 uF
7
+
3
+
741
−
100K
6
10
6.11v
2
4
+
22 uF
B
12.22v
10K
1N4148
+
10
6.11v
1N4148
100 uF
10
10K

Above — Instead of a zener diode, low impedance VCC/2 splitters are shown as A and B; the zener diode is easier.

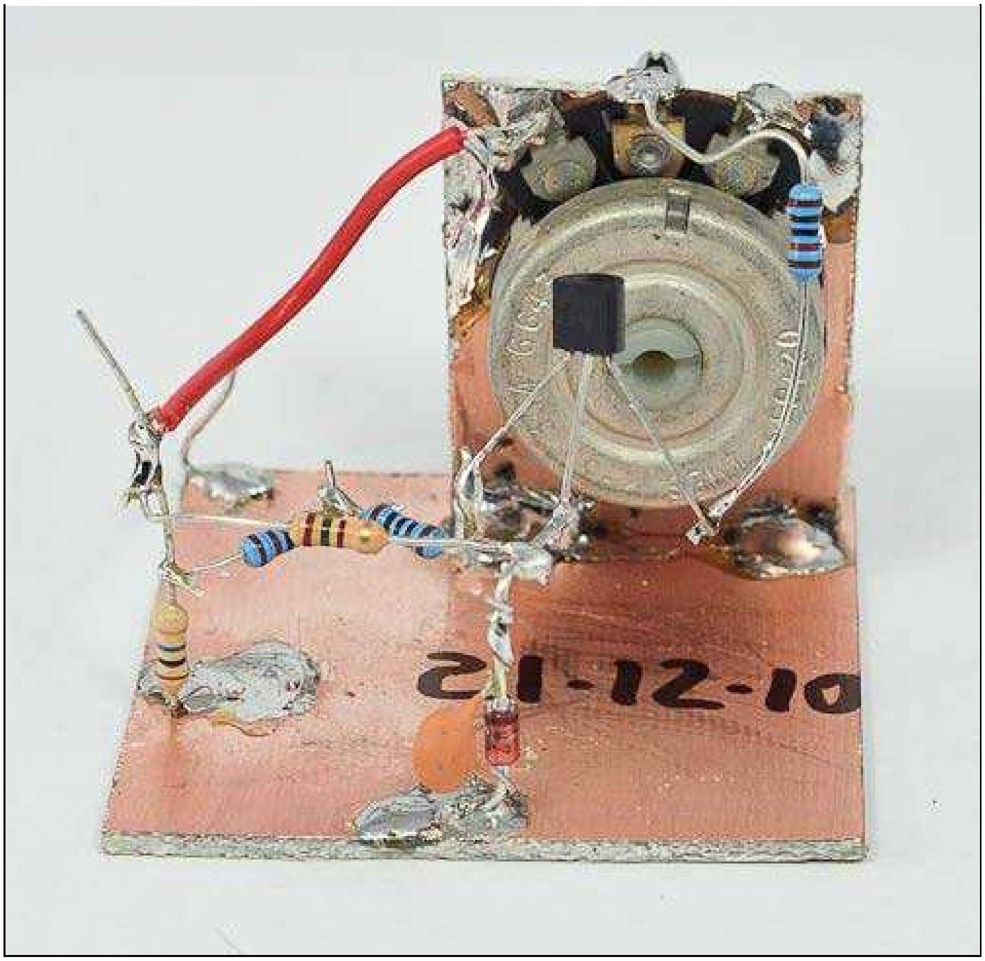

Above — Reverse view. I didn't bother with the 470K VGS range limiting resistor in my breadboard.

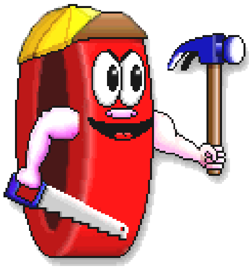

RF — Test and Measurement

RF Workbench Page 1

The first installment of a multi-part series exploring basic 50 Ω RF circuit measurement. This web series borrows heavily from the work of Wes, W7ZOI.

RF Workbench 1 shows a way to measure AC signals and quantify power. As circuit builders, knowing the power gain *or loss* in dB of stages like filters or amplifiers; or the absolute power in dBm of our RF signal sources dominates our bench work.

Ignite your bench measurement — Better measurement fidelity inspires confidence, creativity and fun!

Calculate Power Gain from your Oscilloscope Peak-Peak Voltage

An oscilloscope peak-to-peak voltage provides a popular way to determine power gain or absolute power.

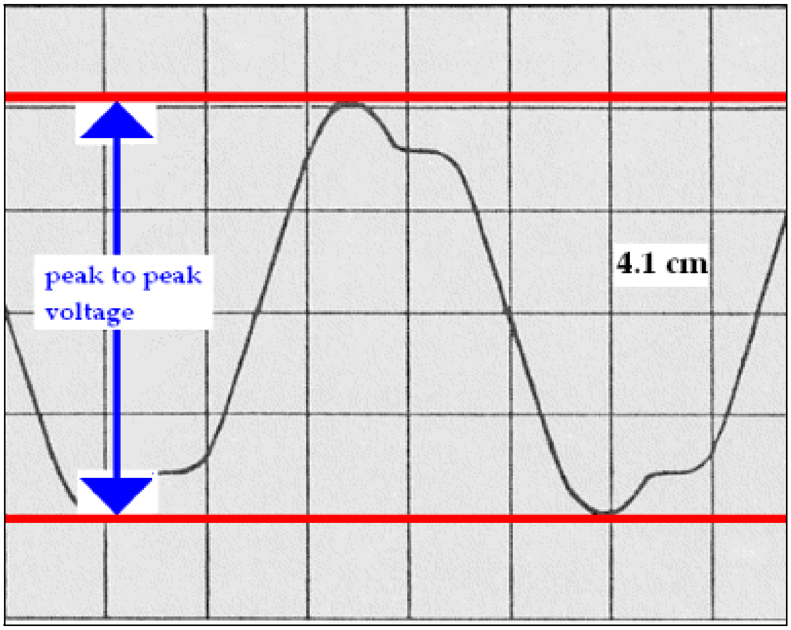

To calculate the power of a sinusoidal waveform, you measure its AC voltage. Assuming your 'scope is calibrated, the first step is to measure the amount of vertical deflection on the screen.

In the figure shown left, the vertical deflection = 4.1 cm. Multiply this measurement by the volts/cm setting of your scope. Lets assume your scope was set to 0.1 volts/cm. Thus the result = 0.41 volts. One final multiplication is required; you must multiply the resultant voltage by the attenuation ratio of the probe. In most cases, a 10X probe is used. Therefore, the measured peak-to-peak voltage is 0.41 x 10 = 4.1 volts. DSOs output numeric voltage readings in addition to that shown — a nice feature.

This signal exhibits a major problem; it's distorted. To calculate the power from peak-to-peak voltage, you require a sine wave. To obtain a sine wave, this distorted signal must be low-pass filtered. Let's examine this topic with some real experiments.

Signal Viewing versus Power Experiments

Measuring 'scope signals takes some skill to get accurate, reproducible results. Consider the following signals taken from an 8 MHz VFO.

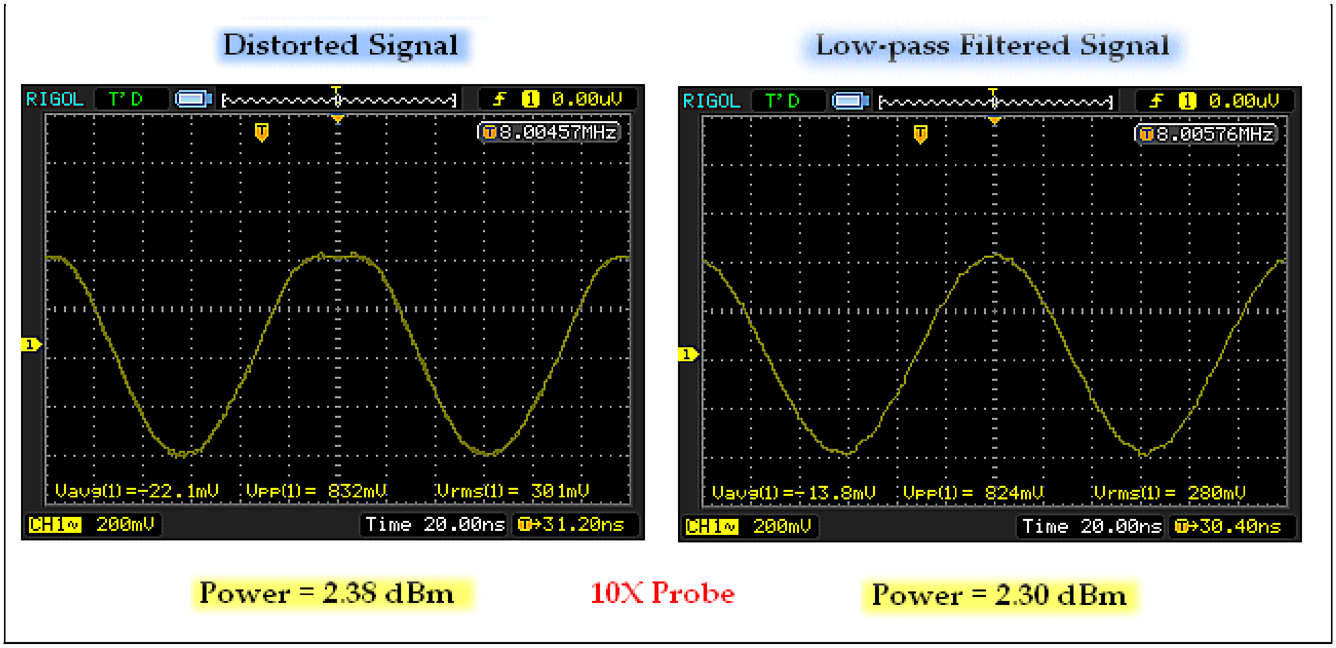
Distorted Signal
Low-pass Filtered Signal
RIGOL T'D
1 0.00uV
8.00457MHz
Vavg(1)=-22.1mV Vpp(1)= 832mV Vrms(1)= 301mV
CH1 200mV
Time 20.00ns
T→31.20ns
8.00576MHz
Vavg(1)=-13.8mV Vpp(1)= 824mV Vrms(1)= 280mV
T→30.40ns
Power = 2.38 dBm
10X Probe
Power = 2.30 dBm

Examine the distorted signal on the left and compare it to same signal after low-pass filtering. I calculated the power in dBm from the peak-peak voltages (Vpp) shown as 832 mV and 824 mV: left to right respectively.

Only the sinusoidal power proves accurate. I'll discuss the formula to assess power soon.

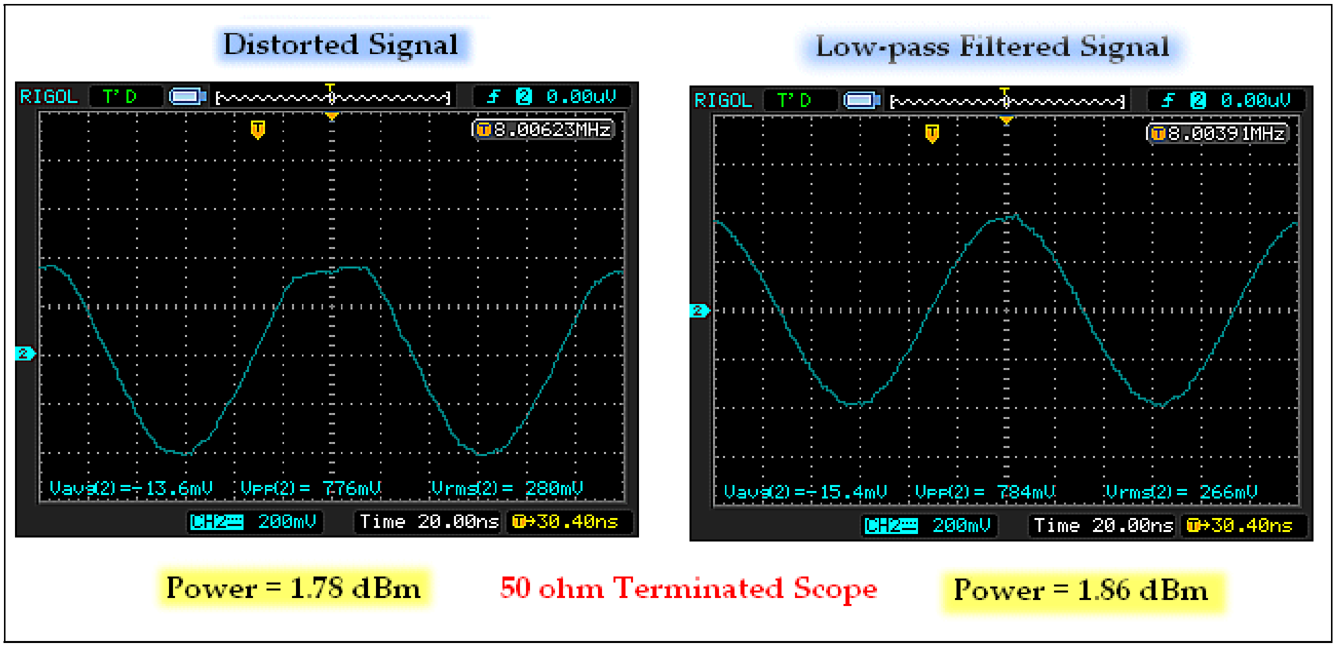
Distorted Signal
Low-pass Filtered Signal
RIGOL T'D
2 0.00uV
8.00623MHz
Vavg(2)=-13.6mV Vpp(2)= 776mV Vrms(2)= 280mV
CH2 200mV
Time 20.00ns
T→30.40ns
8.00391MHz
Vavg(2)=-15.4mV Vpp(2)= 784mV Vrms(2)= 266mV
Power = 1.78 dBm
50 ohm Terminated Scope
Power = 1.86 dBm

The same 8 MHz VFO examined with a 50 Ω terminated oscilloscope; a superb measurement technique that offers greater sensitivity. In all 4 cases shown, the vertical scale = 0.2 volts/cm.

To better compare the distorted and filtered signals, I attenuated the output of the VFO to allow safe examination with a spectrum analyzer. A spectrum analyzer graphs the power (in dBm) of all measured frequencies on its Y axis against a user defined frequency range on the X axis.

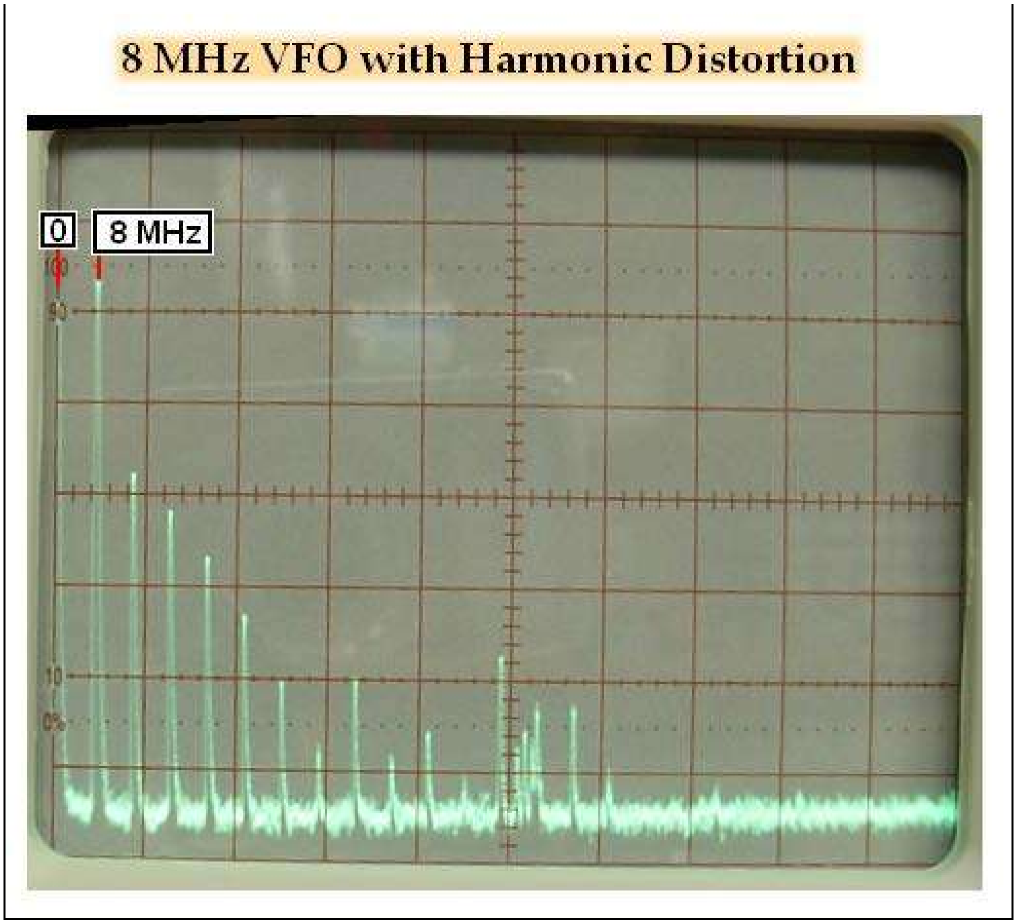

Spectral analysis of the distorted 8 MHz signal. The second harmonic (16 MHz) is about 22 dB down from the fundamental. The signal is rich in harmonics that causes error in the calculation of the output power. Each vertical square is 10 dB. Each horizontal square = 20 MHz. The harmonics go 2x fundamental, 3x fundamental, 4x fundamental and so on.

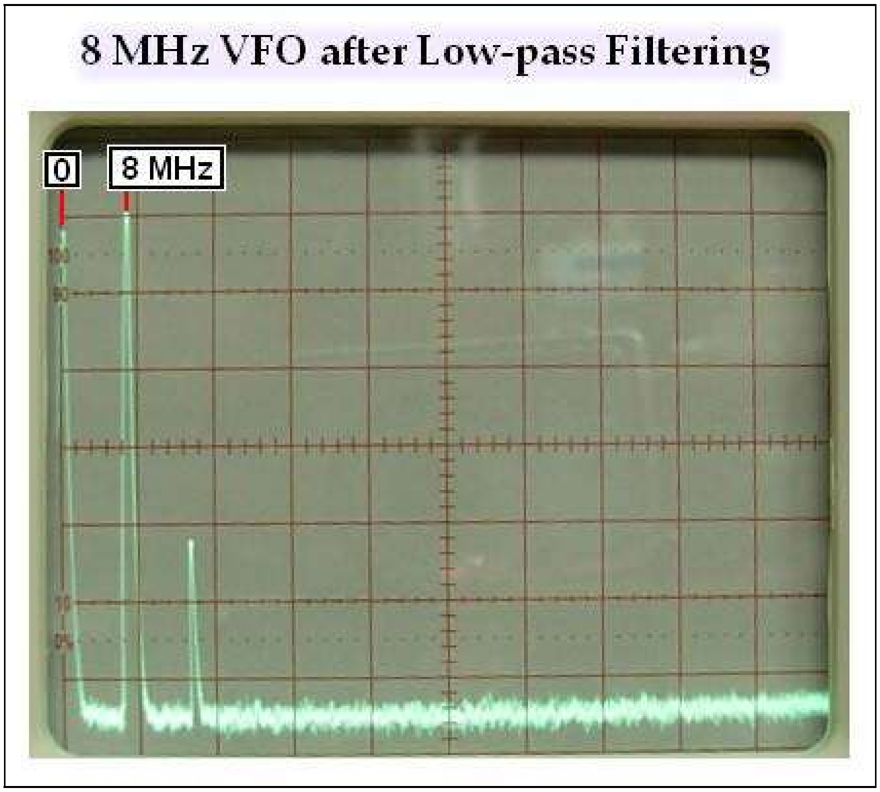

Spectral analysis of the 8 MHz VFO after passing through an N = 5 Chebyshev low-pass filter. The second harmonic now lies about 40 dB down from the 8 MHz carrier (-40 dBc) and the 3rd harmonic is almost down in the noise. Each horizontal square = 10 MHz.

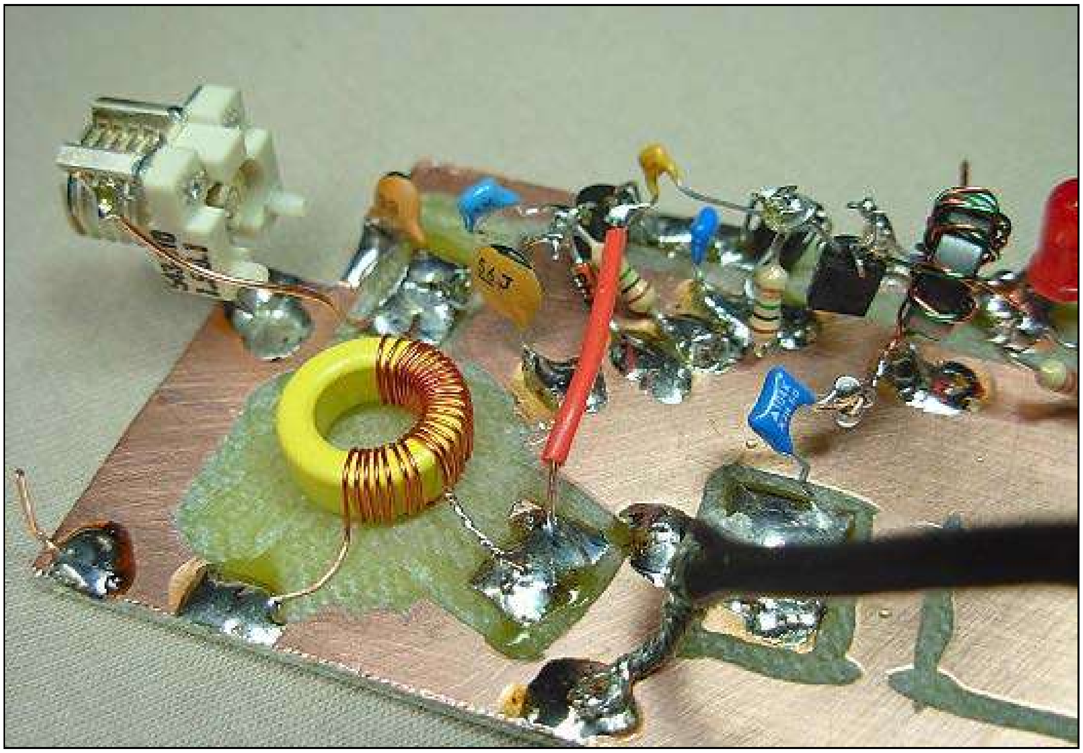

The breadboard of the 8 MHz oscillator oscillator from the above experiments. The output drove a BJT amp biased to give distortion and a 50 Ω output. I adjusted the frequency with the high Q air-variable trimmer capacitor seen to the left.

Low-pass Filter

Some builders wonder why I only employ sine wave signal generators on this web site. To calculate power, they require no low-pass filtering — now you know why. If you're calculating power from a distorted signal, a stiff low pass filter helps ensure measurement fidelity. All of my signal sources feature a 2nd harmonic response of at least -30 dBc, but -50 dBc is typical. To filter receiver front ends, signal generators, or mixer outputs, I keep several 7 element low-pass filter bench modules on hand that cover several 3 dB cut-off frequencies between 3 and 60 MHz.

Although, any old low-pass filters might work fine, Wes, W7ZOI suggests an N = 5 Chebyshev with 0.2 dB of ripple at about 1.2x the signal frequency as a starting point for designing a test-bench low-pass filter. If you don't know how to design low-pass filters choose a pre-designed filter from a filter table. For the experiments above, I selected a filter from an ARRL Handbook. See the schematic below:

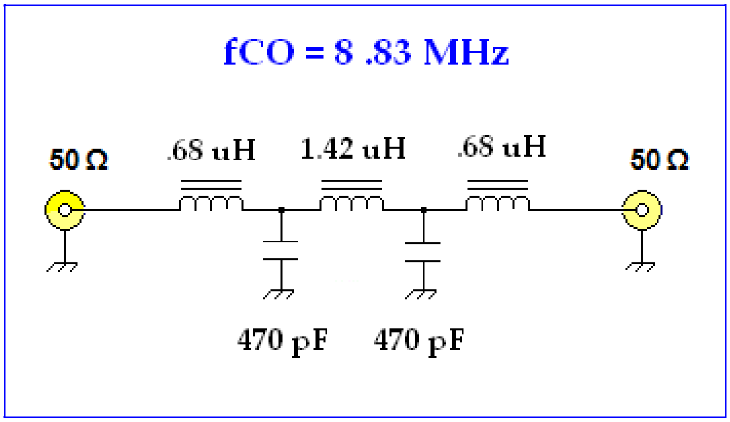

Calculating Power (dBm and mW) from Peak-to-Peak Voltage

To calculate the power from peak-peak voltage, the load impedance (Z) must be known. In RF design, the standardized impedance value = 50 Ω. For CATV and video, 75 Ω is common, and in audio and telecommunication design, a 600 Ω impedance dominates. Although we can technically employ any Z, this web site conforms to the 50 Ω RF impedance standard.

The SI unit of power is the watt. In radio, we might see the term dB used, however, dB is a decibel comparison between 2 signals and not an absolute value like the watt. On the bench, dBm serves as the most common and useful term — dBm is the measured power ratio in decibels referenced to 1 milliwatt.

dBm represents an absolute power — so useful because both large and small signals are quantified with 1 number. Some important bullets follow:

- 0 dBm = 1 mW
- 3 dBm = ~2mW; so doubling the power from 0 dBm equals a 3 dB increase in power
- Increasing the power from 0 dBm to 10 dBm boosts power by 10 dB. The power is now 10X baseline or 10 mW
- 20 dBm = 100 mW
- -27 dBm means that the output has ~500 times less power than 1 milliwatt. -27 dBm = 0.002 mW or 2 microwatts

Hopefully over time, you' ll ingrain the concept of logarithmic power gain or loss (in dB) and power referenced to 1 mW (in dBm). This is bread and butter radio design information you must know.

50 Ω Measurement Virtues

You build a VFO, measure it with your 'scope; calculate the output power into a 49.9 Ω resistor and then record this power in dBm. Let's say it's 6 dBm. VFO output power = 6 dBm.

Next, you place a 6 dB attenuator pad on the VFO output. VFO output power now = 0 dBm.

Finally you connect a 10 dB gain RF amplifer to your VFO. Your VFO output now = 10 dBm. What a beautiful system ! ..... it really gets fun when we measure down at - 30 dBm and so forth.

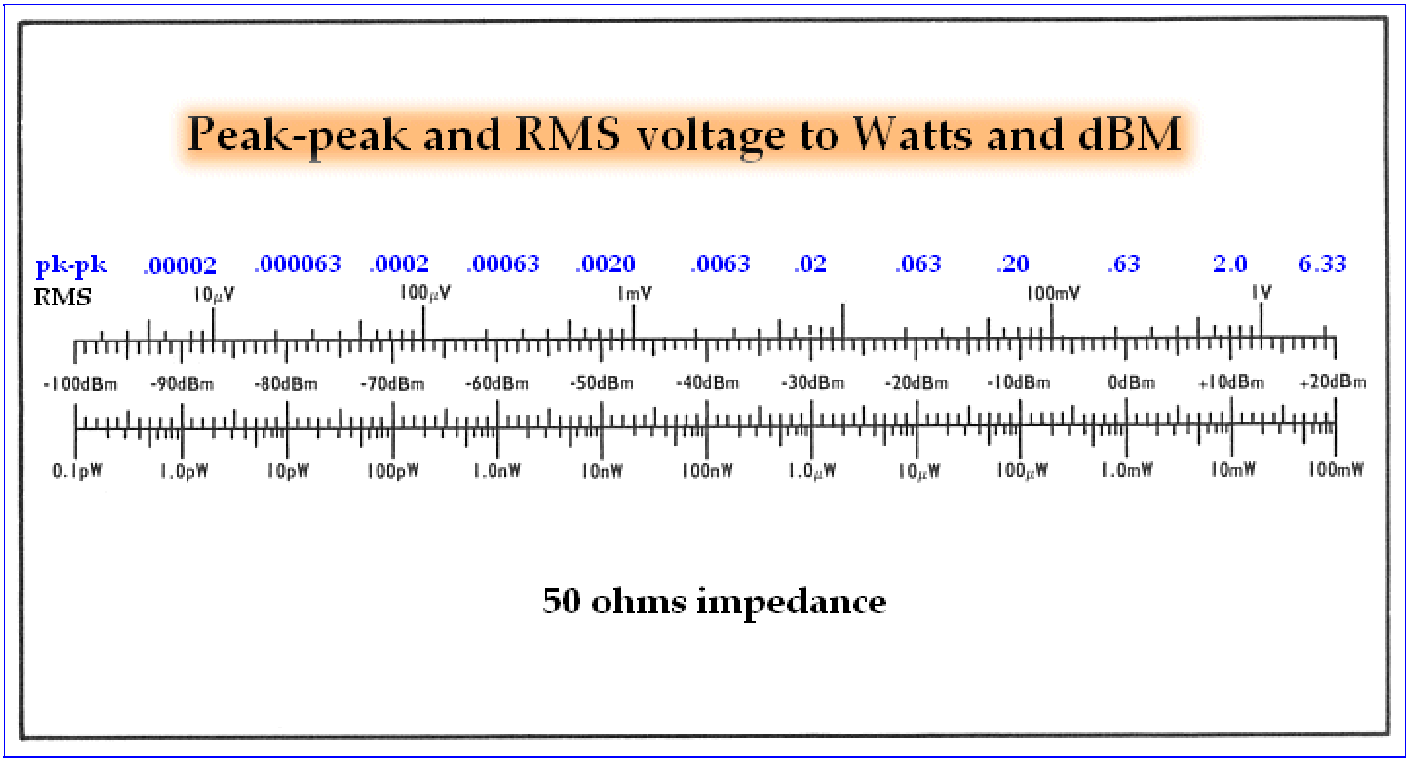

The chart above really helps you visualize the relationships of mW, AC voltages and dBm

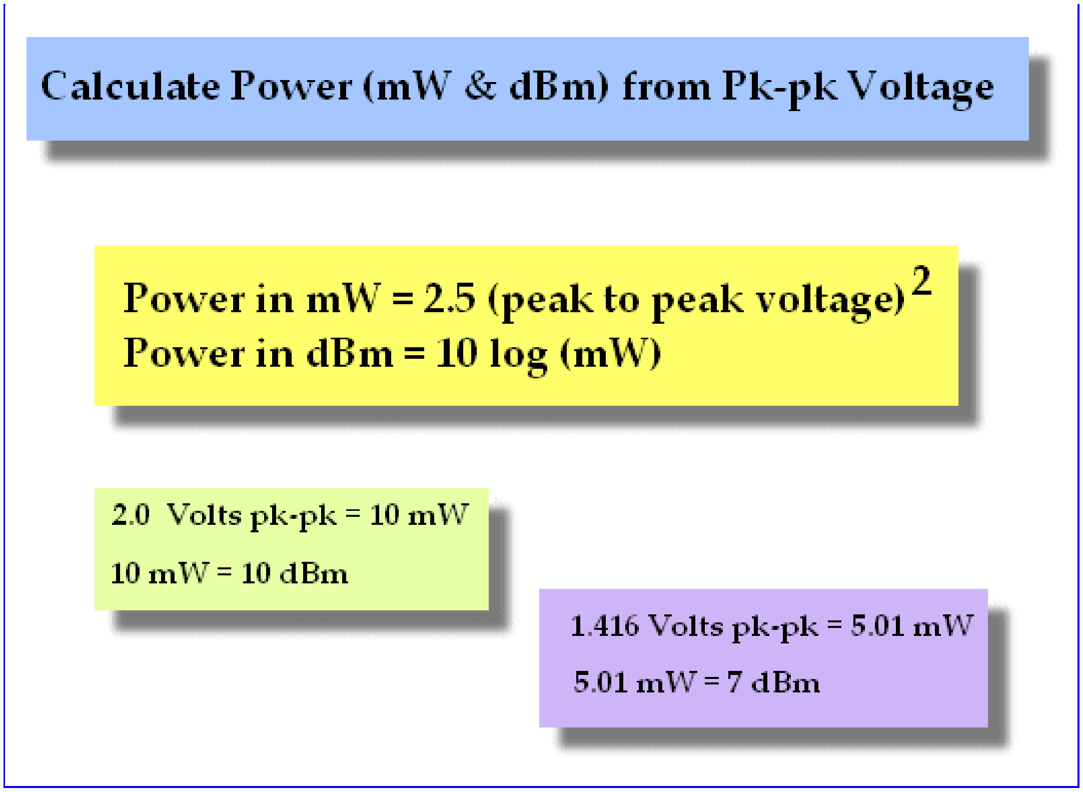

Calculate Power (mW & dBm) from Pk-pk Voltage

Power in mW = 2.5 (peak to peak voltage)^2
Power in dBm = 10 log (mW)

2.0 Volts pk-pk = 10 mW
10 mW = 10 dBm

1.416 Volts pk-pk = 5.01 mW
5.01 mW = 7 dBm

In order to get from peak-to-peak voltage to power, math is required. I show the formula above. You may elect to skip the math and calculate dBm or mW from peak-to-peak voltage with software. A number of programs are available; I wrote 1 here as Applet **F**.

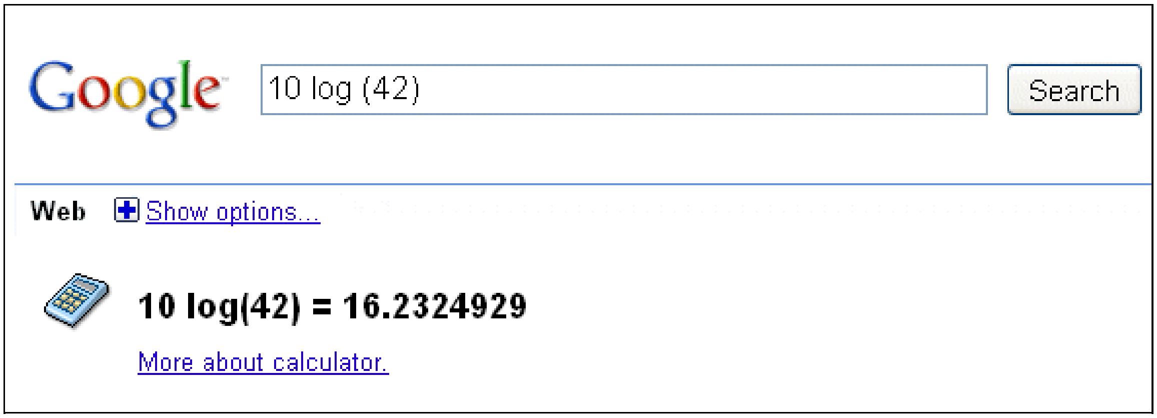

10 log(42) = 16.2324929

If you lack a scientific calculator, Google has math functions. Shown above = a logarithm calculation

In Search of 50 Ohms

How do I establish a 50 Ω output impedance in my RF amplifier? I get this question often.

From my experience, in **simple** amplifiers lacking negative feedback, a 50 Ω output impedance must be created by inserting a resistor somewhere in the circuit that forces a 50 Ω output impedance.

For example, we might use a 50 Ω collector resistor in a BJT amplifier (or a 50 Ω drain resistor in a JFET amplifier), or place a fixed resistor in parallel with the collector/drain transformer and use a secondary winding to establish the 50 Ω output impedance. Sometimes we'll place a series resistor (say from 22 to 51 ohms) on the output of an emitter or source follower to bring the low output impedance up to 50 Ω. These form basic explanations and usage examples may be found on many schematics on the QRP / SWL website.

We experimenters also employ negative **feedback** with or without output transformers to establish a 50 Ω input and/or output Z and I show many examples on this site.

The following diagram explores 1 method to get a 50 Ω output impedance in a simple amplifier. It doesn't matter if the transistor is a JFET or a BJT, the principle is the same. This diagram and tutorial are simplistic and meant to help novice builders learn to design their own amplifier stages. You may connect any resistor value across the output of a transformer to calculate power, however, this web site only considers 50 ohms.

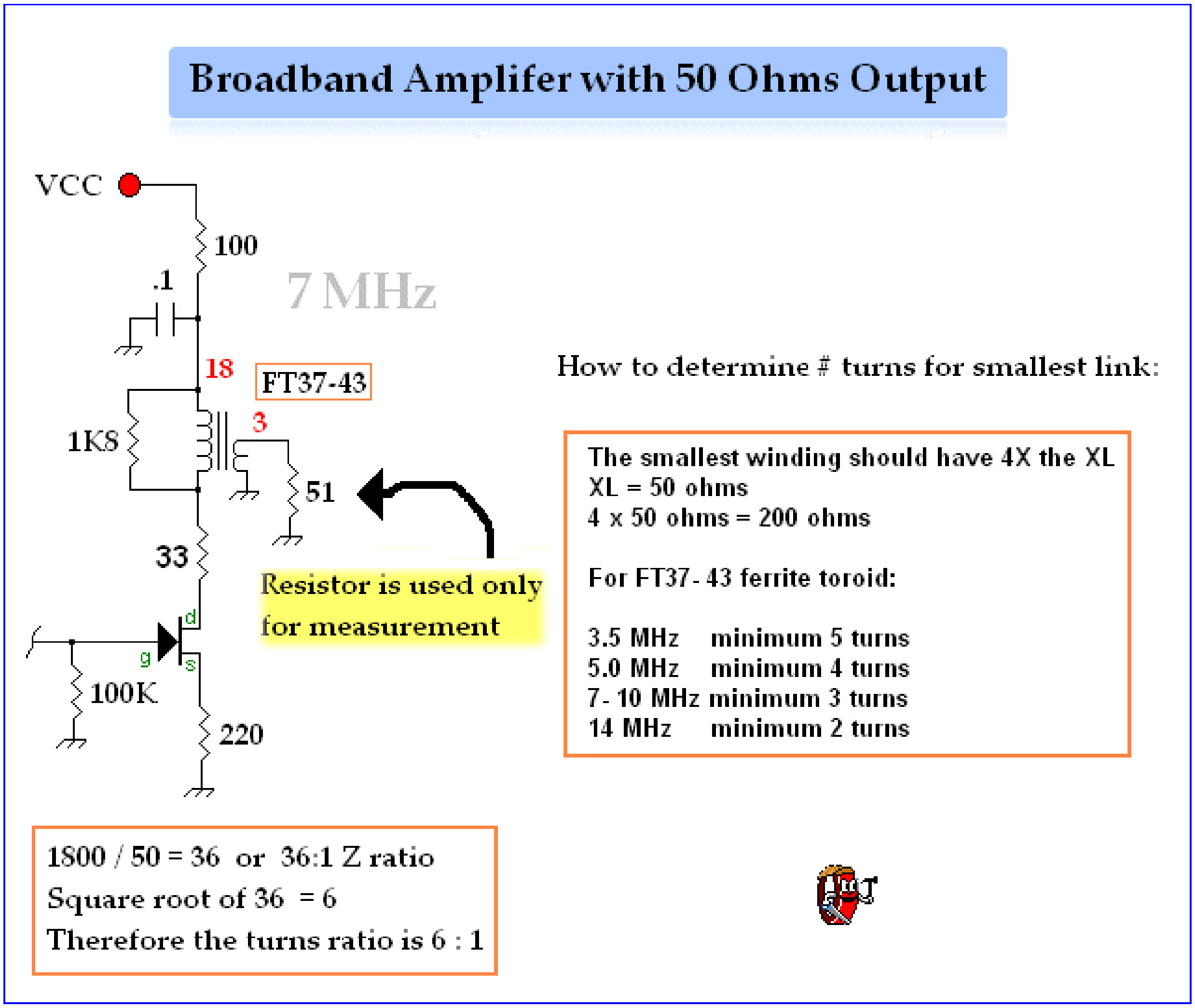

Broadband Amplifer with 50 Ohms Output

How to determine # turns for smallest link:

The smallest winding should have 4X the XL
XL = 50 ohms
4 x 50 ohms = 200 ohms

For FT37-43 ferrite toroid:

3.5 MHz	minimum 5 turns
5.0 MHz	minimum 4 turns
7-10 MHz	minimum 3 turns
14 MHz	minimum 2 turns

1800 / 50 = 36 or 36:1 Z ratio
Square root of 36 = 6
Therefore the turns ratio is 6 : 1

The above diagram describes a broadband (untuned) amplifier. I employed a FT37-43 ferrite toroid: a common part. Other ferrite toroids may be substituted, however the table depicting the minimum number of turns won't apply.

Consider the BJT amp shown. The transformer primary winding is shunted with a parallel 1800 Ω resistor. The 1K8 resistor "forces" a 1K8 ohm collector output resistance in the primary winding.

To transform the 1800 Ω primary impedance to 50 ohms; use a 3 turn secondary link. Calculate the primary to secondary turns ratio as follows:

1800 ohms divided by 50 ohms = 36. The impedance ratio = 36:1.
The turns ratio is the square root of the impedance ratio; thus the turns ratio is 6:1. The primary winding must have 6X the number of turns of the secondary winding. In the 3rd RF Workbench web page, you'll see that the above explanation pertains to the "ideal transformer", however, the concept is useful — especially to the target audience of this website.

New builders might ask — why not wind 6 turns for the primary winding and 1 turn for the secondary winding? We avoid this because the smaller or secondary winding should have have a minimum inductive reactance (XL) of 4X the impedance it is connected to. Thus for a 50 ohm circuit, the minimum XL = 200 Ω at the design frequency.

This design rule serves only as a rough guide. We employ the minimum 4X rule because employing an XL less than 4X may create unwanted signal losses and affect the smaller winding's impedance. The table to the right of the amplifier shows the minimum numbers of secondary turns for a few common frequencies with the FT37-43 ferrite toroid.

Thus for our 7 MHz amplifier, we need at least 3 secondary turns and multiply this number by the turns ratio to give a 18:3 turns ratio. You might also choose 24:4.

For ferrites other than the FT37-43, calculate the minimum number of turns with the XL= 2 pi * F*L formula and detemine L from the turns versus AL toroid data, or measure L with an inductance meter.

For AC measurement a 50 Ω purely resistive load should be temporarily connected between the output link and ground. This might be a 51 ohm resistor, a 49.9 ohm 1% metal film resistor, 2 parallel 100 ohm resistors, or some other "50 Ω" load. We measure peak-to-peak voltage across the load and then calculate the power in dBm or mW. After measurement, the temporary 50 Ω load is removed and the circuit connected to the succeeding stage.

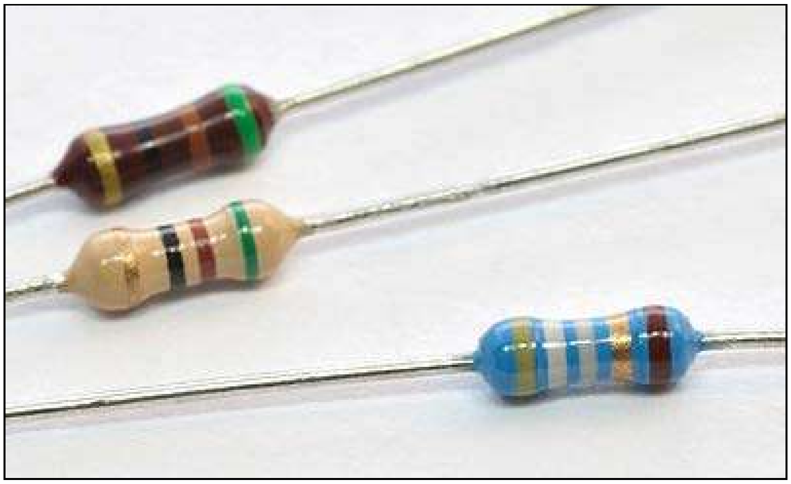

51 and 49.9 Ω resistors

We measure the peak-peak voltage across this 50 ohm resistor with a 10X 'scope probe; or alternately may connect the device output to a 50 Ω terminated scope to measure peak-peak voltage to calculate power.

I normally measure in a 50 Ω measurement environment and temporarily solder a BNC connector onto my breadboard and connect this port to a 50 Ω terminated scope with coax. After testing and voltage measurement, I remove the RF connector and then build and test the next stage.

Case Study

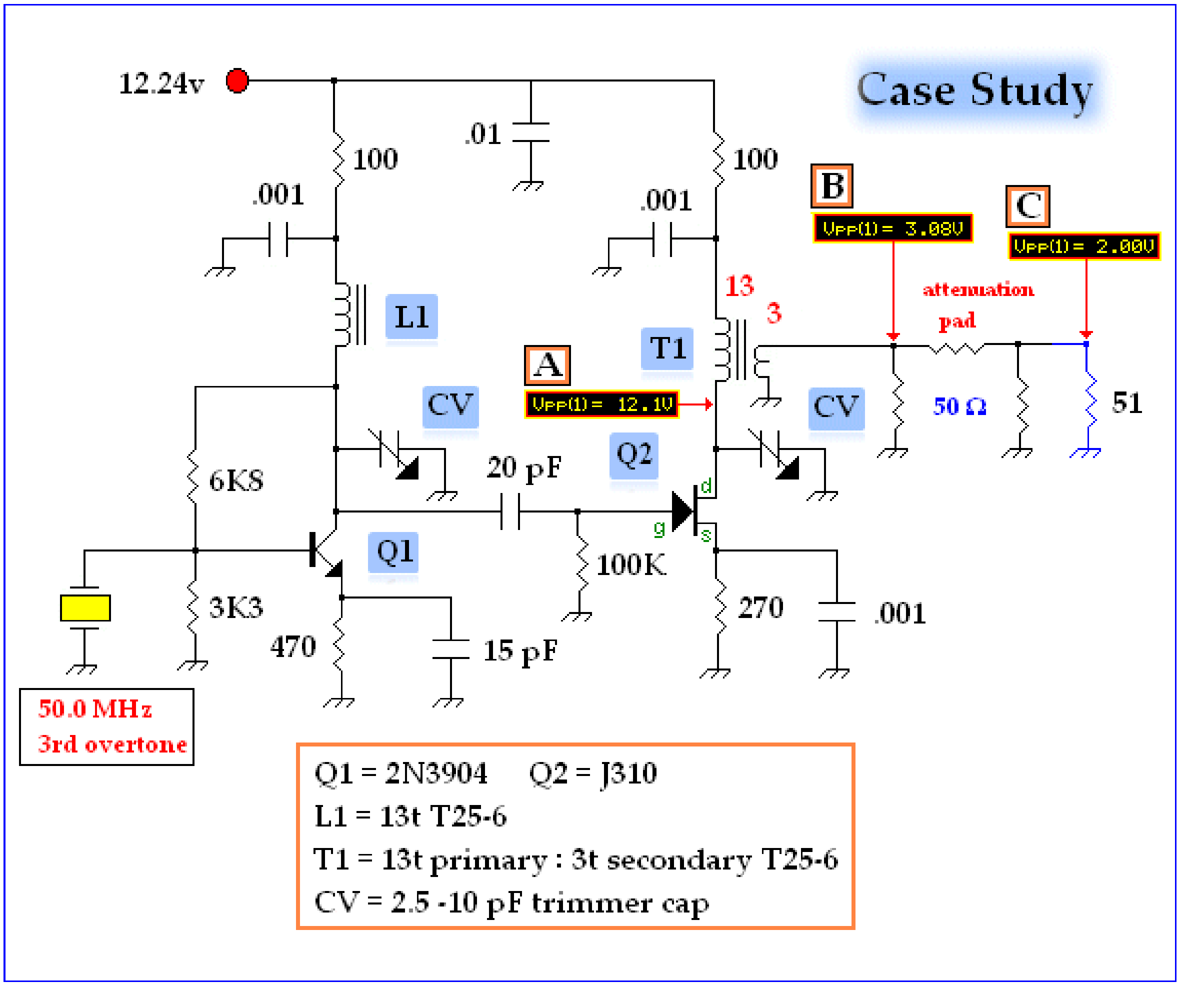
Case Study
12.24v
100
.01
100
.001
.001
B
Vpp(1)= 3.08V
C
Vpp(1)= 2.00V
13
3
attenuation
pad
L1
T1
A
Vpp(1)= 12.1V
CV
CV
50 Ω
51
6K8
20 pF
Q2
d
g
s
Q1
100K
3K3
270
.001
470
15 pF
50.0 MHz
3rd overtone
Q1 = 2N3904 Q2 = J310
L1 = 13t T25-6
T1 = 13t primary : 3t secondary T25-6
CV = 2.5 -10 pF trimmer cap

Pretend that you breadboarded the above circuit entitled "Case Study". This is a 50 MHz crystal oscillator and buffer. The crystal fundamental frequency = 16.7 MHz, but the L1 tank is tuned to its 3rd overtone; 50.0 MHz. You measure and record the peak-to-peak voltages at the points labeled A, B and C.

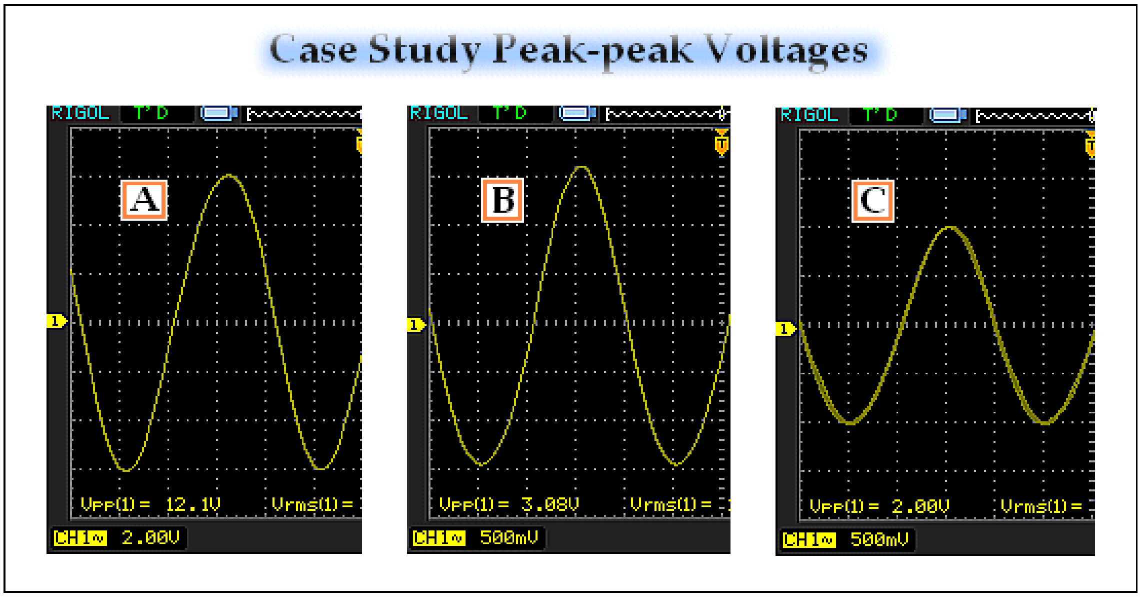

The peak-to-peak voltages are shown as Vpp. The vertical scale (volts/cm) is shown on the bottom of each figure.

Examine Point A. The AC voltage = 12.1 volts peak-peak. Compare this to the peak-peak voltage at Point B. Note the difference. Some builders emailed me after they measured similar differences on the primary and secondary transformer windings of their circuits with a 10X probe. These builders felt something must be wrong. All normal; you can expect the peak to peak voltage to roughly decrease (or increase) by the transformer turns ratio.

The 12.1 volts peak-peak decreased by a factor of 4.3 (13 / 3 turns ratio) which is 2.8 volts peak-peak. In our case, the measured secondary peak-peak voltage was 3.08 volts — in the ballpark. Please remember this serves as a **coarse** guide only. It helps you to know what to reasonably expect during signal viewing.

Peak-to-peak voltage changing in accordance with the transformer turns ratio represents a simplistic explanation describing the "ideal transformer". To understand real world transformer function, you must contemplate factors such as Ohm's law for AC, conservation of energy (this is what causes the voltage to drop while preserving power) and basic transformer behavior. These principles are explained in publications such as The ARRL Handbook for Radio Communications, or the RSGB Radio Communication Handbook. An old high school physics text book might prove a better reference.

Here are the case questions:

1. Calculate the power in dBm at point B
2. Calculate the power in dBm at point C
3. What is the attenuation in dB of the 50 ohms attenuation pad?
4. What is the output power in mW of this stage?

Click on this link for the answers and to see the actual resistor values of the attenuator pad.

Finally, placing a 10X probe at Point **A** will de-tune the L-C tank circuit somewhat and thus alter the AC voltage. In real-world building; to tune Q2, tweak the variable capacitor (CV) with your 10X probe connected to Point C.

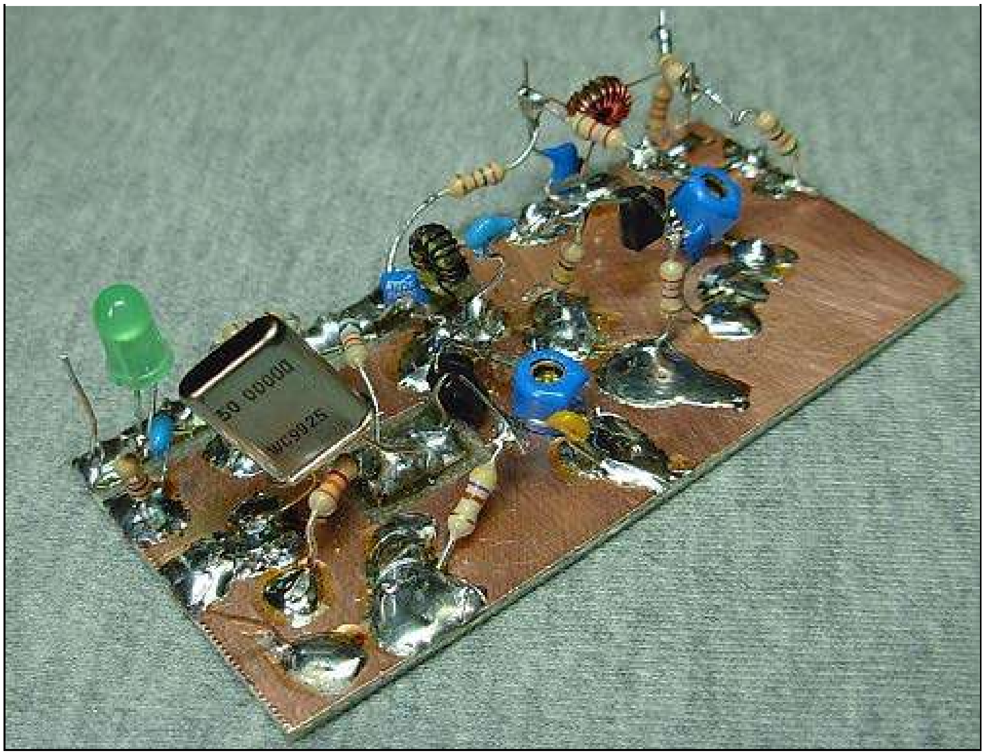

The breadboard of the 50 MHz oscillator prototype.

Oscilloscope Probing

10X Oscilloscope Probe

Please refer to EMRFD Chapter 7 for great information about measuring power in RF circuits. The 10X oscilloscope probe is one of the most important measurement tools to have on your bench. There are countless web articles concerning the 10X probe, so I don't have much to add.

Take care of your 10X probe: don't solder components you've clipped your probe to; avoid setting heavy objects on the cable; store it carefully and inspect it frequently.

When do you use a 10X probe ? Measure with a 10X probe for in-situ ("in place", or "in circuit") voltage measurement and in situations where you can afford a 10X reduction in sensitivity. In low level measurements such as millivolt level measurements, the reduced sensitivity of a 10X probe may reduce or disallow accurate measurement. Additionally, the 20 pF or so capacitance of a 10X probe can detune resonant circuits; especially at VHF on up.

Close up of the Rigol oscilloscope probe 10X and 1X switch.

50 ohm Terminated Oscilloscope

At RF, we generally work with (or try to work with) circuits with 50 Ω impedances. If possible consider performing your measurements in a purely 50 ohm environment.

That is — instead of using a 10X probe, shunt the oscilloscope input port to ground through a 50 Ω resistor and connect your test circuit to the 'scope with 50 ohm impedance coaxial cable. On my 'scopes, I have Channel 1 set up for the 10X probe work and Channel 2 set up for a 50 ohm environment.

I asked Wes about the benefits of performing measurement in a 50 Ω environment. I learned the main advantage of a 50 ohm approach is a well defined port impedance. The second virtue; a 10X greater voltage sensitivity — the increased sensitivity for low level measurement amazes me. In some cases, small signals that I couldn't accurately measure with a 10X probe, gave an excellent scope tracing with more consistent voltage readings in a 50 Ω environment.

You also may enjoy improved signal viewing. For example, in a few cases I have observed harmonic distortion with a 50 ohm terminated scope unseen with a 10X probe I confirmed this distortion with a spectrum analyzer.

If you have never performed measurement in a 50 Ω environment, consider trying it out — you'll enjoy it. You may buy commercial 50 Ω feed-through devices that connect to your oscilloscope input, or homebrew your own, but try to keep the 50 ohm termination as close to the oscilloscope input as possible.

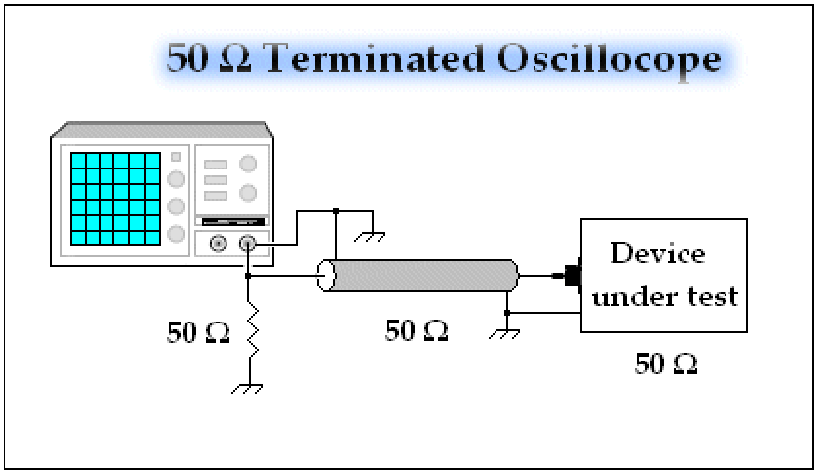

Establishing a 50 Ω impedance measuring environment. The oscilloscope input is terminated with a 50 Ω resistance and connected to a device with a 50 Ω output impedance via 50 Ω coaxial cable.

Try not to routinely connect a feed-through attenuator pad to your feedthrough 50 Ω 'scope terminator — error may arise.

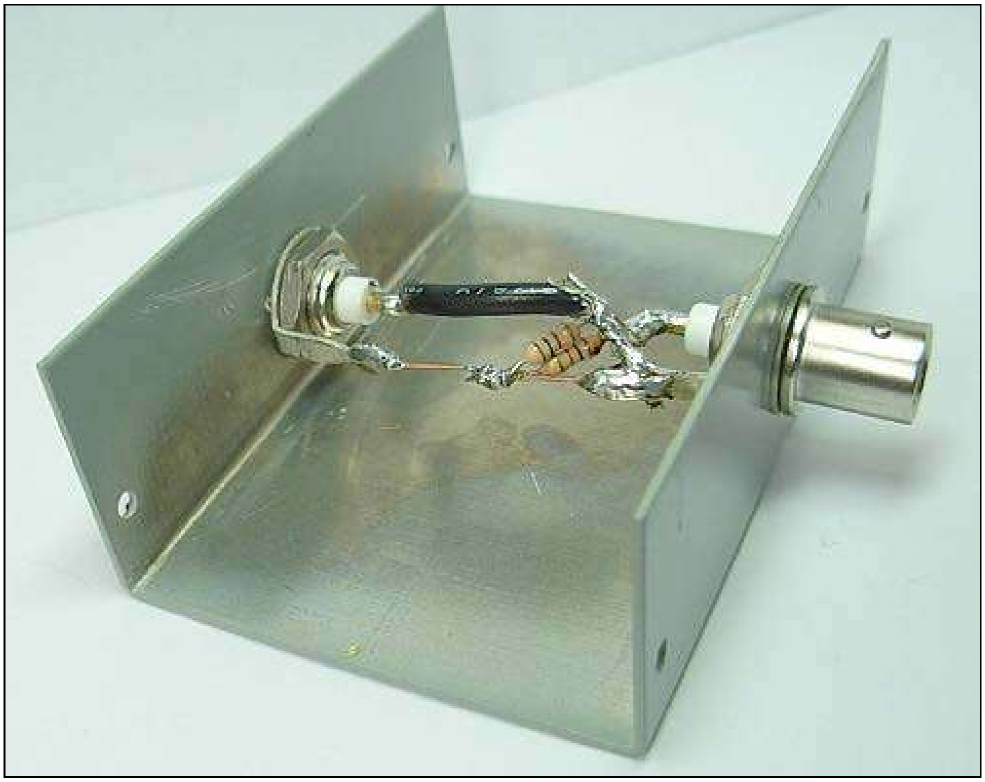

My very first homebrew 50 Ω scope terminator module with 2 female connectors. I connected this module to my oscilloscope input via a commerical 9 cm long 50 Ω coaxial cable with a male connector on each end.

Two parallel 100 ohm resistors formed the 50 Ω load. Ideally, the 50 Ω shunt resistor should be right at the 'scope's female BNC jack — so this homebrew module shown fell short as a stalwart 50 Ω terminator. Inspired to move to a 50 Ω environment and lacking a male BNC connector, it did the job until my commercial version came by mail. You might find oscilloscope feedthrough terminators for sale at Ham festivals.

An ideal homebrew solution — place a male and female BNC connector in a small metal box very close together to allow a very short interconnecting wire. The box would hang off of the oscilloscope. Better still are commercial, shielded 50 Ω feed through terminators which thread right onto the oscilloscope's female BNC input jack.

MATH
REF
SCALE
SCALE
50%
FORCE
EXT TRIG
CH 2
CH 1
X
Y
Feed-through
terminator
DC to 1GHz
50 Ω/1W
Made in China

Above — a commercial 50 Ω feed-through BNC terminator on my oscilloscope input.

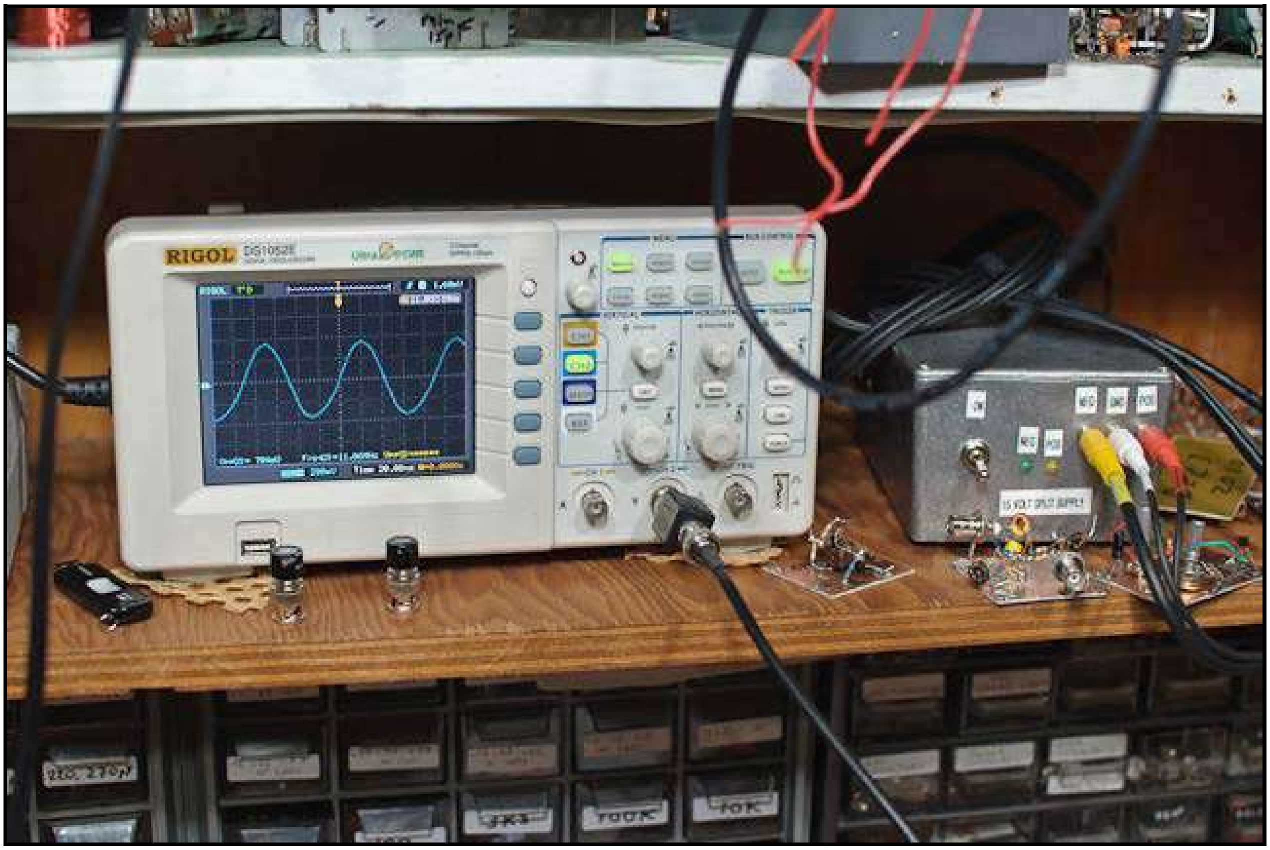

Above — Measuring in a 50 Ω environment. Bliss!   Я люблю это.

RF Current Sampler

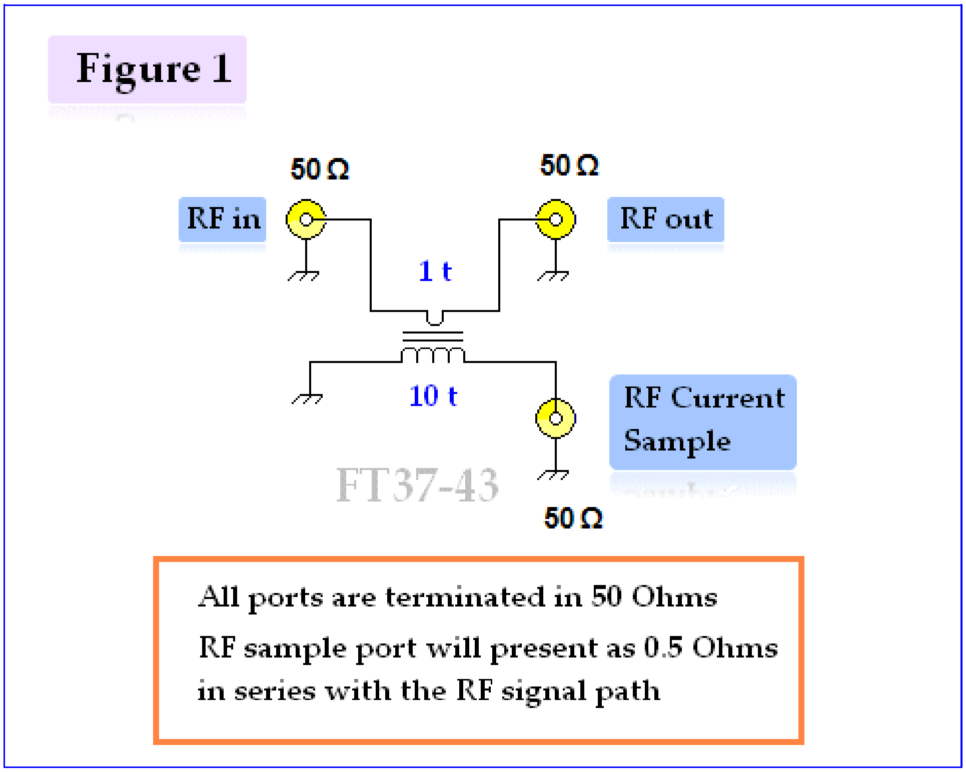

Figure 1

All ports are terminated in 50 Ohms

RF sample port will present as 0.5 Ohms in series with the RF signal path

Figure 1 shows a basic circuit to sample RF current from a power stage such as a QRP transmitter. Many experimenters lack 50 Ω step attenuators rated to handle transmitter-level power. One basic solution is to sample the RF current of the power amplifier using a wideband step-

down transformer. Terminate the RF current sample port with a 50 Ω impedance device. This may include a spectrum analyzer, power meter, receiver with an attenuator, or a 50 Ω terminated oscilloscope.
A usage example = examining a transmitter's spectral purity with a spectrum analyzer. The output power at the sample port will drop by 20-22 dB. A 50 Ω impedance step attenuator can be used to further reduce this power level to whatever you want. For this chore, a typical experimenter's 1-2 watt step attenuator works, since it never "sees" the higher wattage transmitter power.

For example, a 5 watt amplifier 20 dB down is 0.05 watts or 50 mW at the RF current sampler port. 50 mW = 17 dBm. To examine this signal with a spectrum analyzer you may wish to decrease the power down to -27 dBm. The following chart shows the basic process.

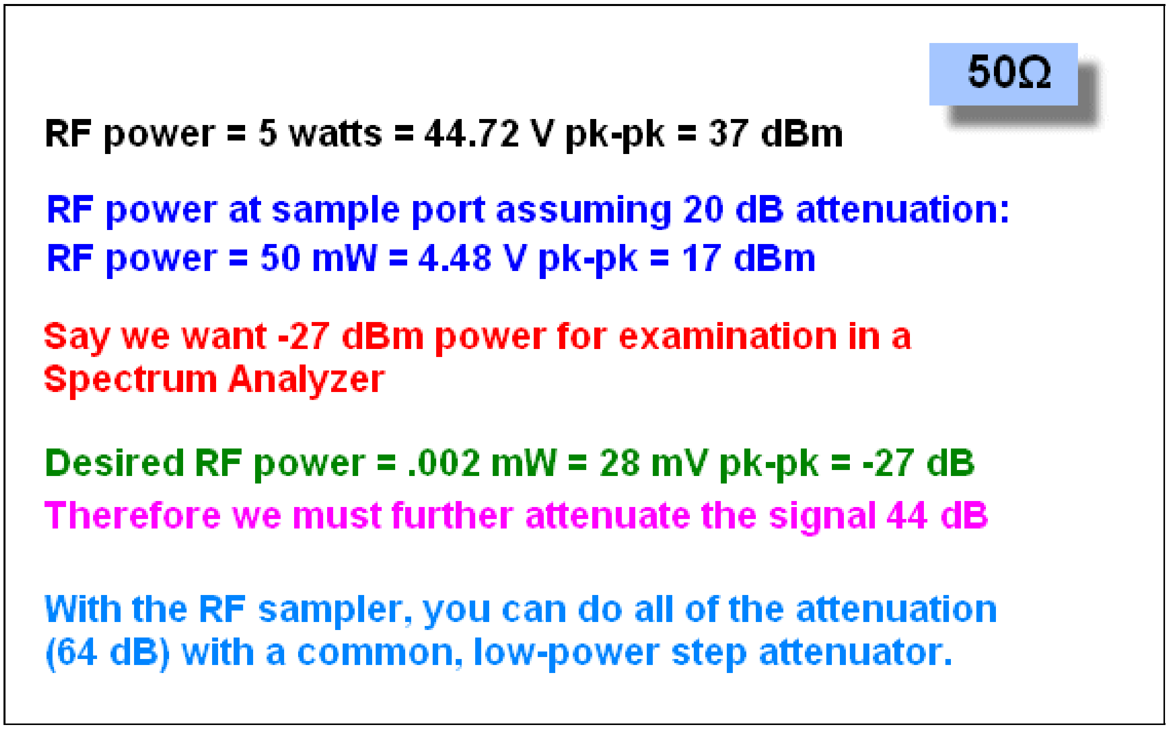

50Ω

RF power = 5 watts = 44.72 V pk-pk = 37 dBm

RF power at sample port assuming 20 dB attenuation:
RF power = 50 mW = 4.48 V pk-pk = 17 dBm

Say we want -27 dBm power for examination in a Spectrum Analyzer

Desired RF power = .002 mW = 28 mV pk-pk = -27 dB

Therefore we must further attenuate the signal 44 dB

With the RF sampler, you can do all of the attenuation (64 dB) with a common, low-power step attenuator.

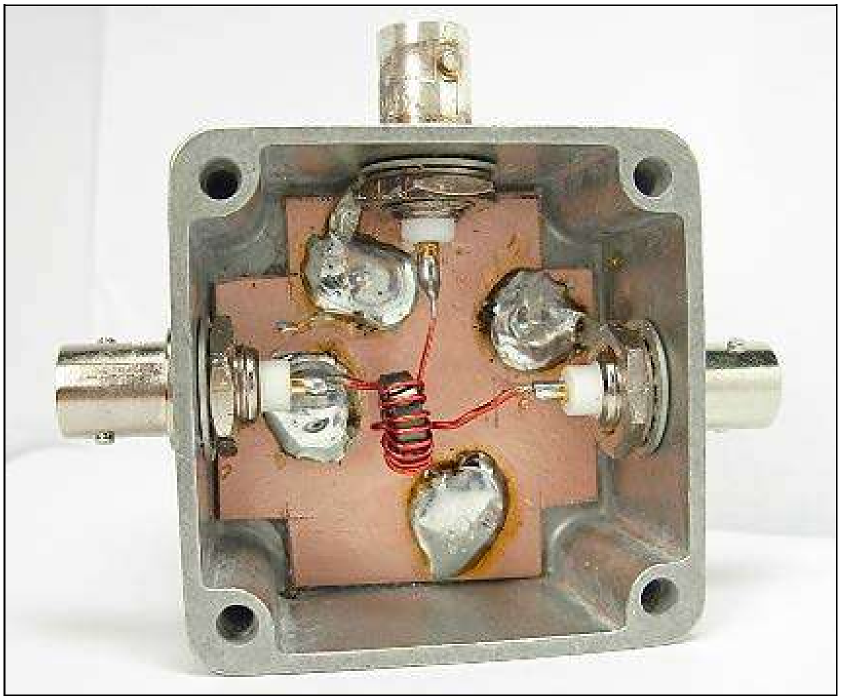

A Hammond chassis shields the RF current sampler used on my bench.

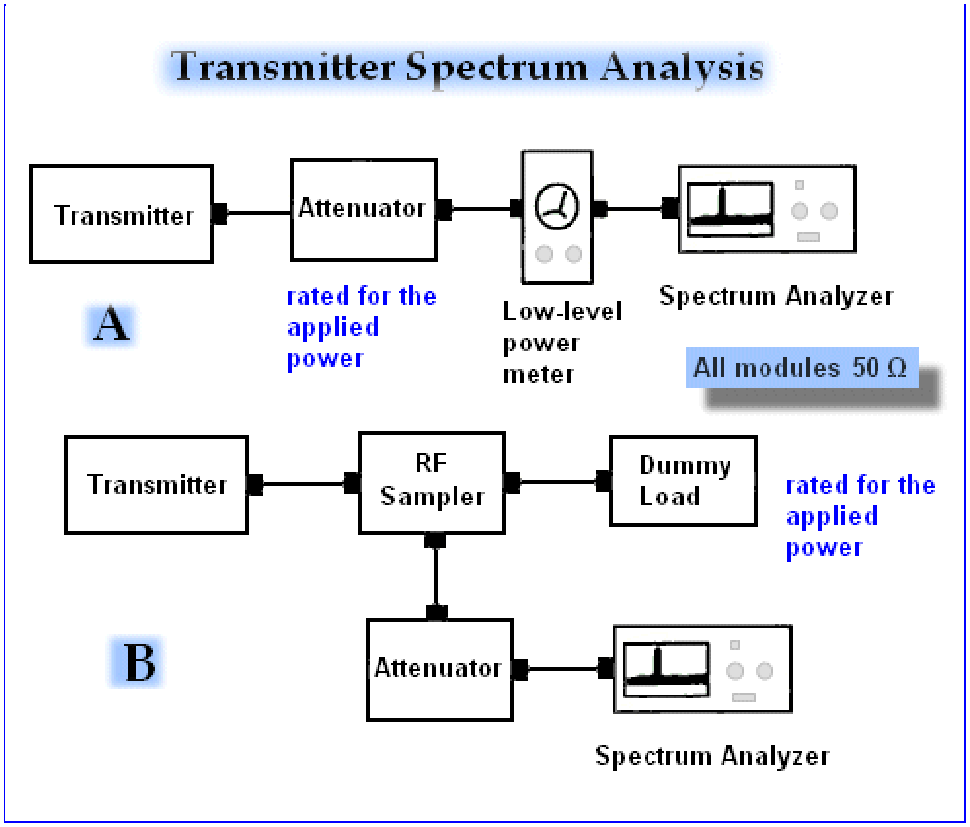
Transmitter Spectrum Analysis
Transmitter
Attenuator
Spectrum Analyzer
A
rated for the applied power
Low-level power meter
All modules 50 Ω
Transmitter
RF Sampler
Dummy Load
rated for the applied power
B
Attenuator
Spectrum Analyzer

The above graphic illustrates 2 methods to examine the output of a transmitter in a spectrum analyzer. Method B is described above. The dummy load must handle the transmitter output power, however a 5 or 10 watt dummy load is easy to make. Method A requires a step-attenuator which can handle the transmitter output power. The low-level power meter promotes the need to quantify the output power before you connect anything to the output of the attenuator. This is also true for Method B.

Always measure the output power at the RF sample port with your oscilloscope or low-level power meter before hooking up any expensive low-level measurement device such as a spectrum analyzer! осторожно!

Miscellaneous Photos

Feed-through
terminator
DC to 1GHz
50Ω/1W
Made in China

TA051
Sales order 62403
BNC feedthrough 50R 1W 1GHz
Pico Technology Limited
www.picotech.com
CE

MALE TERMINATOR 50Ω +1%

Above — a 50 Ω BNC terminator. These are essential QRP work bench items and may be found on eBay for cheap.

Click for a photograph of 4 of my BNC RF port terminators: 27, 50, 75 and 100 ohms from left to right — useful to calibrate and test RF circuitry.

QRP SWL HOMEBUILDER

RF — Test and Measurement

RF Workbench Page 2

Welcome to part 2 of a web series exploring basic RF measurement and bench practices. This installment builds on the information from the RF Workbench Page 1

I share introductory and practical content on attenuation, the return loss bridge, insertion loss or gain and spectrum analyzers. Consult EMRFD and use your favorite web search engine for more information.

The Attenuator Network

Like onions in your kitchen, the importance of attenuators can't be overstated. On the bench, attenuator pads go in nearly every test circuit to deliver correct power and/or help match port impedances.

Think of attenuators as passive networks that intentionally insert power loss between 2 components independent of frequency. For example: you might follow a 50 Ω signal source with an attenuator to decrease its power, increase its return loss, and/or buffer it from downstream impedance changes.

Most attenuator networks have fixed input and output impedances.

The input and output impedances may be the same, or different.

Attenuation may be fixed or variable. Most often, we use simple, fixed resistive pads that function as voltage dividers.

Express attenuation in dB.

Attenuators increase return loss and reduce VSWR.

Attenuators may function as buffers to isolate stages.

All attenuators on this web site feature 50 Ω input and output ports.

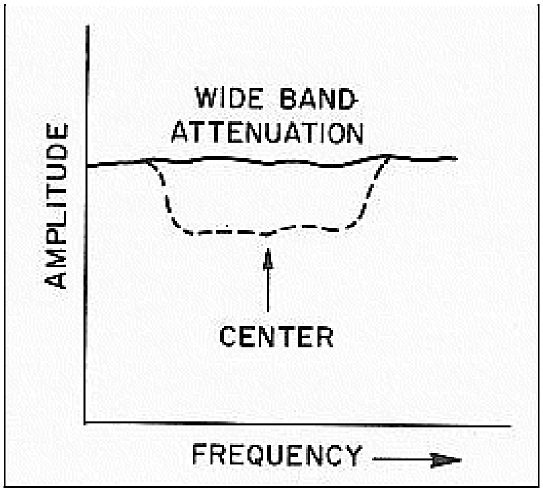
WIDE BAND
ATTENUATION
AMPLITUDE
CENTER
FREQUENCY

The desired response of an attenuator network. Practically speaking, device construction techniques including shielding will limit how high in frequency your attenuator will properly operate.

GENERAL ELECTRIC
AEROSPACE
ELECTRONICS DEPT.
INSERTION LOSS 0DB
Attenuator
MODEL 30-0
Zo: 50 OHMS
IN
20DB
20DB
20DB
20DB
10DB
5DB
3DB
2DB
1DB
KAY ELECTRIC CO.
PINE BROOK, N.J.

A pair of commercial 50 Ω input/output impedance step attenuators from the past. Occasionally, you'll find them for sale at Ham festivals or estate sales. Most work well for HF and perhaps even VHF work depending on their design and condition. Visually examine and test the attenuator before use.

How to Design and Build Attenuator Networks

To design attenuators with a 50 Ω input and output impedance, I recommend viewing a table. Click for a table.

After choosing the degree of precision; solder up 2 or 3 resistors and you're done. Nearest value 5% tolerance resistors offer reasonable precision for our popcorn circuits, but combining 1% and 5% tolerance resistors works too.

Choose low inductance resistor such as carbon film types and strive for short lead lengths. Well consider resistor power dissipation — for example, an 8 dB attenuator pad will dissipate 84% of the RF passed through it. I have seen attenuator pads that were exposed to high power and some or all of the resistors were burnt and turned to charcoal. Clearly the operator did not regard the power rating of the attenuator resistors. Refer to EMRFD Section 7.4 for practical information concerning attenuator design and power dissipation.

Three of the attenuator pads from my bench attenuator drawer. When using Ugly Construction or its variants, you can solder in, change, or remove attenuator pads at a whim. A small stock of these pads speeds up your work flow.

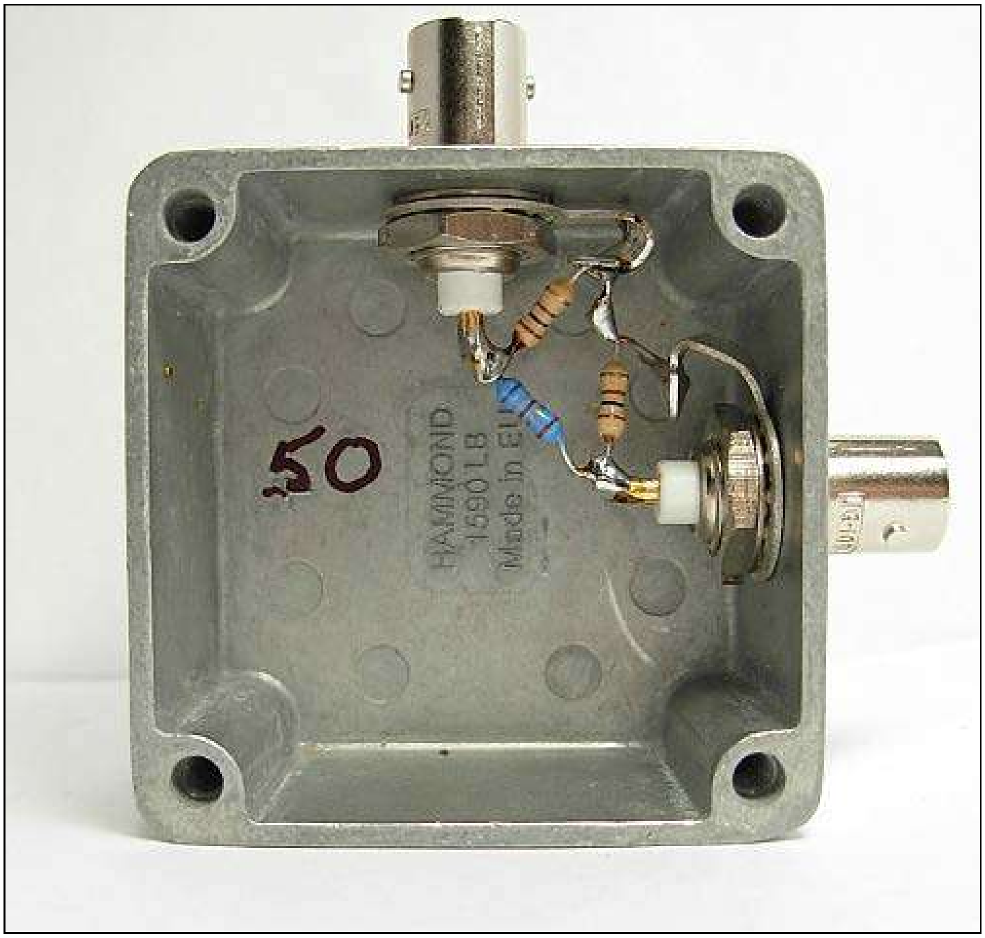

A 10 dB attenuator pad from my collection. This box uses two 100 Ω (5%) resistors and a 68 ohm (1%) resistor for the 96.2 and 71.2 ohm resistances called for. I used the 1% tolerance part because all my 68 Ω resistors are metal film 1% tolerance 1/4 watt types. Perhaps, I'll pursue a closer match to the calculated resistor values 1 day, but this module works okay. You may also stick 2 seperate attenuators in the little Hammond box shown.

6dB ATTEN
FREQUENCY:D
10dB ATTEN
FREQUENCY:D

Two commercial BNC feed-through (in-line) attenuators. I use these *every* day and prefer them over homebrew R networks since they don't require a coaxial cable. I own many: two 6 dB, two 10 dB, one 3 dB and a 20 dB: all were purchased on eBay.

VAT-15
15 dB ATTENUATOR
50 OHM
VAT-6
6 dB ATTENUATOR
50 OHM

Two commercial SMA in-line attenuators for my VHF and UHF experiments: 6 and 15 dB. Click for a marker table with 4 data points derived from sweeping these 2 filters.

Step Attenuators

A step attenuator belongs on every serious RF workbench. They allow in-situ attenuation adjustment with a degree of precision as low as 1 dB. Step attenuators are nothing more than switched calibrated resistances and the switches can be SPDT, relays, rotary or digitally-controlled types. The quality and price of commercial attenuators varies widely. Experimenter concerns include the minimum attenuator insertion loss, power rating, return loss, noise from switch contacts and noise from the resistors themselves.

A homebrew step attenuator makes a great weekend project and almost every radio handbook contains 1. Web linked projects plus commercial kits may be found online — use your favorite search engine to find them.

Some homebuilders prefer 1% metal film resistors to keep resistor noise down. Stick your step attenuators in a metal, RF-proof box and insert quality interfaces such as BNC, N or SMA connectors. Your needs, budget and parts collection determine the outcome when you home build one.

Serebriakova Attenuator - Серебрякова аттенюатор (50 Ом)

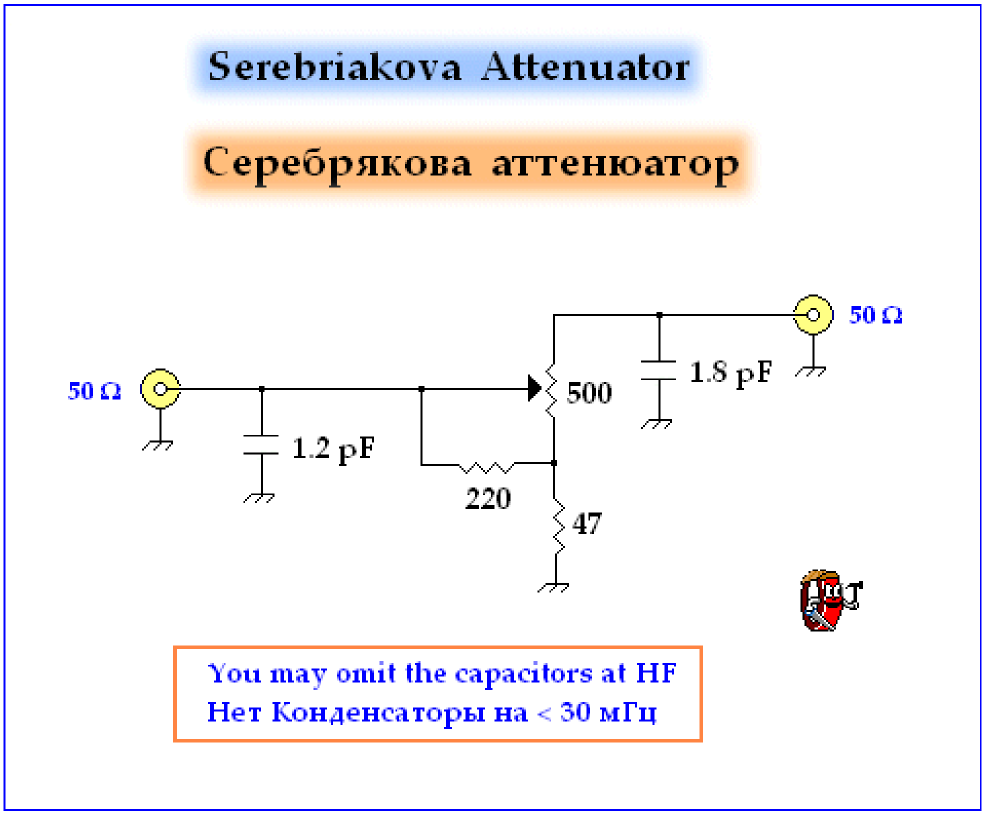

The Serebriakova; a simple, variable attenuator well suited for QRP homebuilding. Filled with gratitude to its Russian designer's family, I share this contribution with my readers. This attenuator network makes signals smaller or larger in a 50 ohm environment via a potentiometer. My analysis indicates acceptable performance considering its simplicity. The input match is close to 50 Ω across the range of the potentiometer. The output match across the potentiometer range is mediocre. Click here for a DC match analysis from Wes, W7ZOI. As shown, you wouldn't place this device on your main bench signal generator output as the output impedance diverges widely during amplitude adjustment.

Add fixed attenuator pads on the input and/or output to improve matching into 50 ohms. This circuit could serve in multiple applications including an RF gain control on a receiver front end, for bench measurement (when adapted) and for a low-level transmitter gain control. The Serebriakova attenuator may function up to 500 MHz in a carefully constructed, shielded box. The input and output capacitors may be omitted below 30 MHz. The attenuation varies a minimum of 20 dB when turning the potentiometer from CCW to CW. Click for a build by he yl.

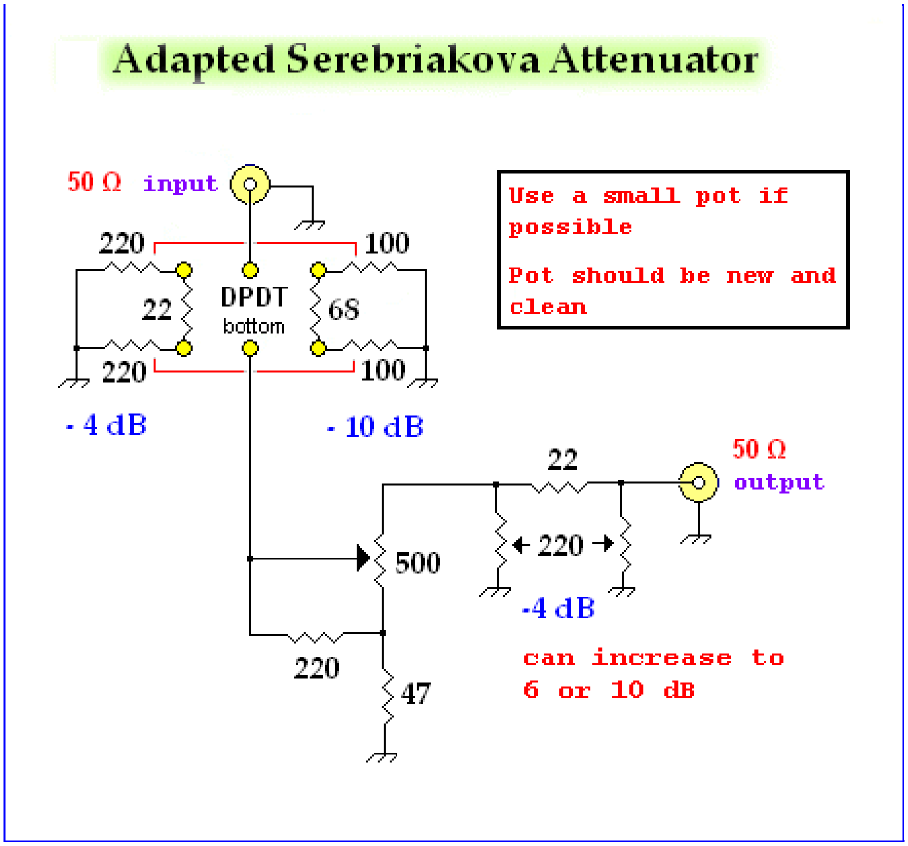
Adapted Serebriakova Attenuator

Use a small pot if possible

Pot should be new and clean

can increase to 6 or 10 dB

A variant of the Serebriakova attenuator is shown above. Input and output matching are enhanced by fixed attenuator pads. The input match into 50 ohms is fine. After testing, I learned that the fixed 4 dB output attenuator pad is likely too low to ensure a wide range output match into 50 ohms. A 6 or 10 dB output pad is preferable, however, if this is your only variable attenuator, the device would then only be usable for very low-level work. You can decide what value of input or output pads to use.

A new, clean and small size 500 ohm pot works best. Store your potentiometer collection in sealed plastic bags to keep out workshop and house dust.

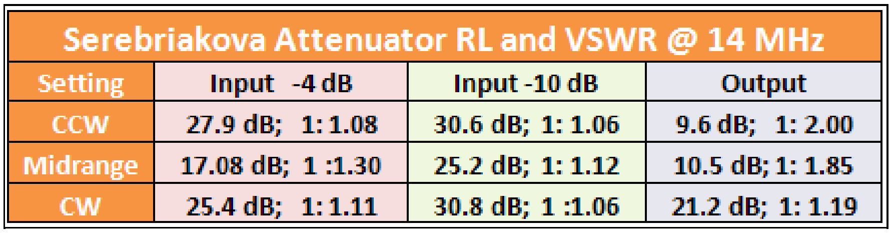

Serebriakova Attenuator RL and VSWR @ 14 MHz			
Setting	Input -4 dB	Input -10 dB	Output
CCW	27.9 dB; 1: 1.08	30.6 dB; 1: 1.06	9.6 dB; 1: 2.00
Midrange	17.08 dB; 1 :1.30	25.2 dB; 1: 1.12	10.5 dB; 1: 1.85
CW	25.4 dB; 1: 1.11	30.8 dB; 1 :1.06	21.2 dB; 1: 1.19

Shown above are return loss (RL) and VSWR measurements performed on the adapted Serebriakova attenuator shown above. Clearly the input match is better than the output match. The output match did not significantly change when the attenuation switch was moved from 4 to 10 dB attenuation or back.

Based upon these values, it might be a better compromise to put a 3 to 4 dB pad on the input and a 10 dB attenuator pad on the output to ensure an output RL of at least 20 dB. Some might argue that the output RL should be higher. Perhaps, but the match is pretty good for such a simple circuit. Let's put it in perspective; a commercial signal generator that sells on the Internet for $450.00 U.S dollars was measured by a builder I know in the UK and he found a best case RL of 10 dB ! Jim later sold it and built a homebrew signal generator with a 35 dB return loss at all frequencies.

Fixed attenuator pads provide a good remedy for mismatched ports and I discuss why and how in the next section.

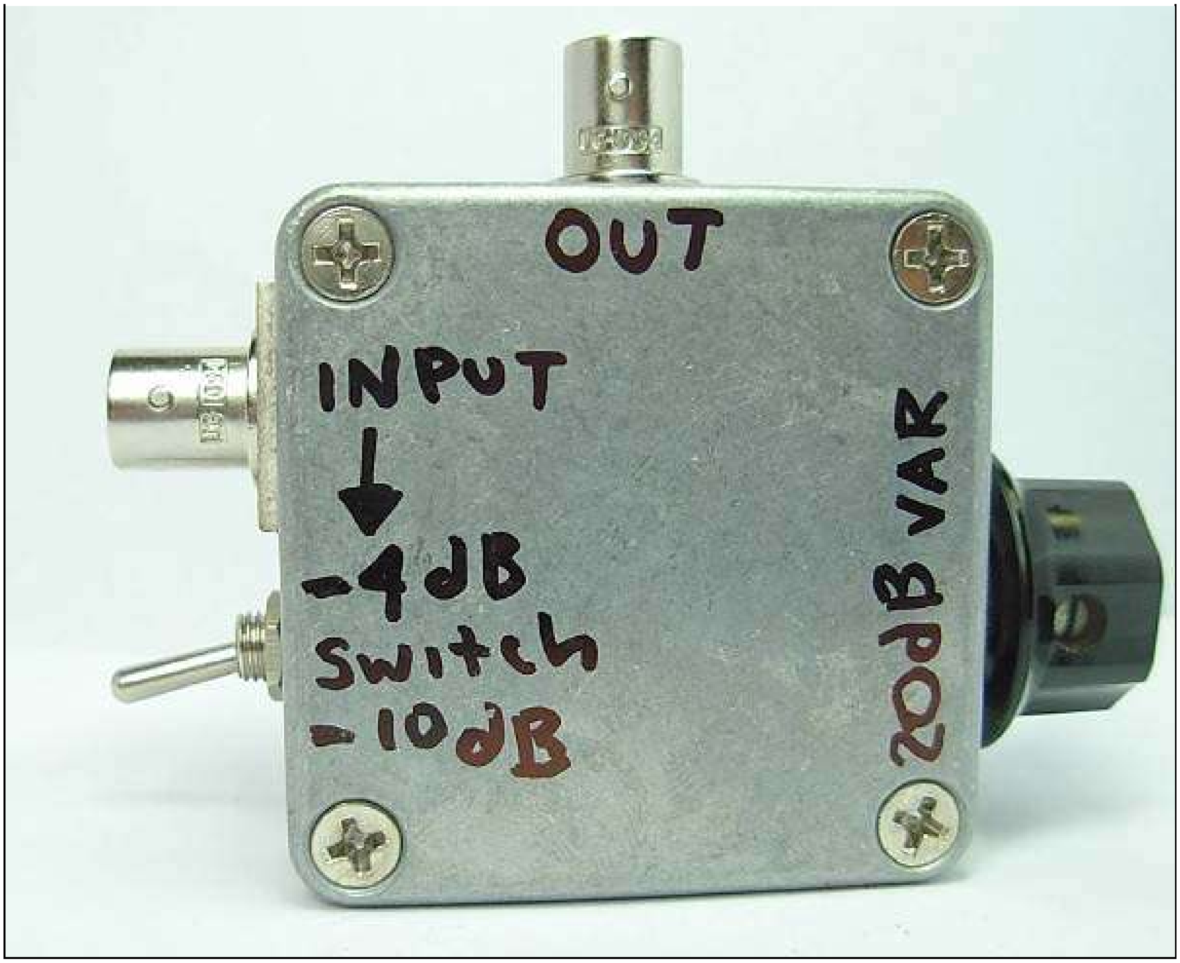

The shielded, adapted, Serebriakova attenuator. When home building your personal version, strive to make the AC connections as short as possible. The above device has nearly 23 dB of variable attenuation at 14 MHz. If you can't build, find, nor afford a precision step attenuator for your QRP workbench, this device may work okay for you.

Impedance Matching, Return Loss and VSWR

We radio folks build, buy and apply lots of gear with a stated *nominal 50 Ω input or output impedance*. In truth, a *pure* 50 Ω impedance occurs rarely and components in an RF system are frequently mismatched. Almost every Ham radio operator matches their antenna impedance to their feed line + radio to maximize transmitter output power — but radio and antenna system matching is often the only case where these Hams match their gear.

In contrast, we experimenters, tirelessly match our 50 ohm RF system components — this work flow avails our **modus operandi** on the bench**.** And so, we builders match the input and/or output ports of *all* our RF stages: signal generators, filters, splitters, antennas and so on. You can easily measure the impedance match of your RF components with a basic device based upon a Wheatstone bridge; the return loss bridge. First, let's discuss matching a little more:

On the RF Workbench, we talk about return loss, reflection coefficient and VSWR to quantify impedance matching. I only consider return loss and VSWR on this web site.

When 2 system components are impedance matched, maximal power transfers from 1 device to the other. If the impedances are different, RF power is reflected back to the signal source. This reduces the amount of power delivered to the load. Transmitted and reflected waves moving along a transmission line superimpose and cause standing or stationary waves. The greater the impedance mismatch between the 2 components, the larger the amplitude of the standing waves. Mathematical formulas compute how much power is lost due to mismatch. Consider reading a great tutorial on SWR, Return Loss, and Reflection Coefficient linked here by Wenzel Associates.

Return Loss

Return Loss = the difference between the outgoing incident power and the reflected power as a result of the mismatch between the the signal source and its load. Return loss is expressed in dB as a positive number on this web page. The higher the return loss, the better the impedance match. An ideal prefect match would have a RL of infinity; that is, no power is reflected back to the signal source and all of the incident power is delivered to the load. If a circuit has no load (open circuit), the RL is 0 dB — all of the power is reflected back to the signal source.

Other terms quantifying return loss are S11 and S22, however S11/S22 are the negative of return loss: RL = 20 dB or S11/S22 = -20 dB. We say S11 as S — one — one and S22 as S — two — two. I discuss these S-numbers, or Scattering Parameters elsewhere.

VSWR

Voltage standing wave ratio is another measure of how well the components in an RF network are impedance matched. Increasing the return loss lowers the VSWR and vice-versa. Most amateur radio enthusiasts are familiar with VSWR and often refer to it as "match" or "SWR". RL and VSWR can be derived mathematically from one other. VSWR = [10^(RL/20) + 1] / [10^(RL/20) - 1]. **Note X ^ Y means X raised to the power of Y therefore 2^3 = 2x2X2 = 8.**

Thus a RL of **10 dB** = 1: 1.92 VSWR and **20 dB** = 1:1.2 VSWR and **30 dB** = 1:1.07 VSWR

In EMRFD, Wes presents a return loss bridge as Figure 7.41. This circuit, shown below is easy to build and use.

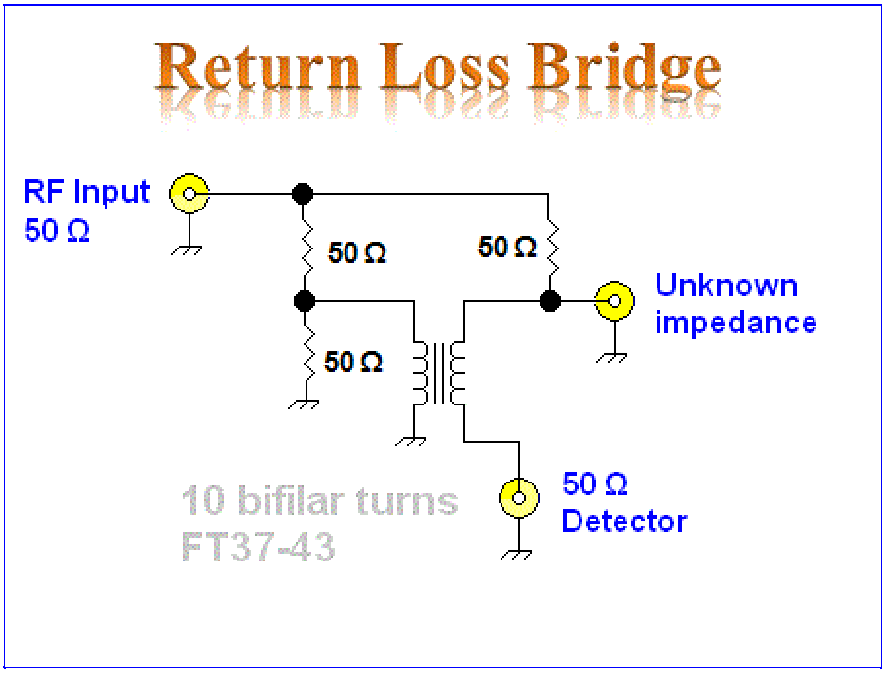

The 50 Ω impedance detector may include a spectrum analyzer, power meter, receiver with an attenuator, or a 50 Ω terminated oscilloscope. On my bench, a 50 ohm terminated scope is favored.

Let's measure the return loss of a 27 Ω resistor to learn how. The procedure with a 50 Ω terminated 'scope follows:

Connect a 50 Ω output impedance signal generator to the bridge RF input port with 50 ohm coax;

Connect a 50 ohm terminated oscilloscope to the detector port via 50 ohm coaxial cable;

Record the peak-to-peak (open circuit) voltage with no load on the end of a short coax cable connected to the unknown Impedance port;

Record the peak-to-peak voltage with "unknown" coaxial cable terminated with the 27 Ω resistor

Calculate the power difference in dBm between these 2 peak-to-peak voltages.

Return loss = the difference in dB between these 2 values calculated by hand or with software. Please refer to the RF Workbench Page 1 for information how to calculate power. I wrote a JavaScript Applet that take these 2 peak-to-peak voltages and calculates RL and VSWR; its labeled **K** on this web page.

Before measuring the unknown RL of a circuit, we usually connect a 50 Ω terminator to the unknown impedance port and calculate the best possible return loss: we refer to this value as bridge directivity — the best possible match for that return loss bridge at that test frequency. I keep a permanent 50 Ω terminator + a barrel connector on my bench for this purpose.

Click for a photo of the gear I use for all RL measurement. *Best viewed at full resolution*

Lets run through the procedure to measure the return loss of a 27 Ω resistor again, but with added photographs and 'scope captures. I tested at 14.070 MHz.

Above — We'll measure the return loss of this device; a 27 Ω resistor soldered on a BNC connector. We call this a resistive terminator and I keep a small collection of 27 - 100 Ω terminators on-hand for calibration purposes.

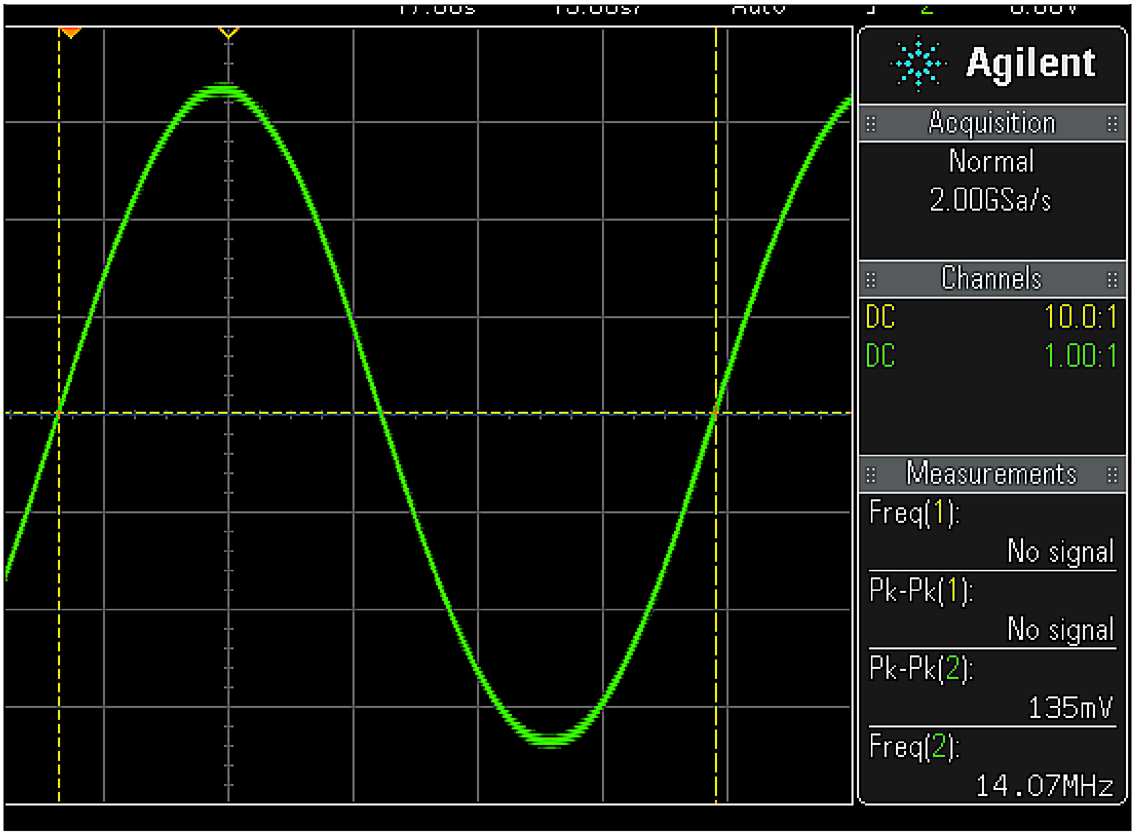

Acquisition
Normal
2.00GSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(1): No signal
Pk-Pk(1): No signal
Pk-Pk(2): 135mV
Freq(2): 14.07MHz

Shown above — The peak-peak voltage with a 14.07 MHz oscillator connected to the RF port; a 50 Ω terminated scope connected to the detector port; and a 20 cm — unterminated — 50 Ω cable connected to the unknown impedance port. **The open circuit measurement.**

Above — Next, I connected the 27 Ω terminator to the unknown impedance coaxial cable with a through-connector interface.

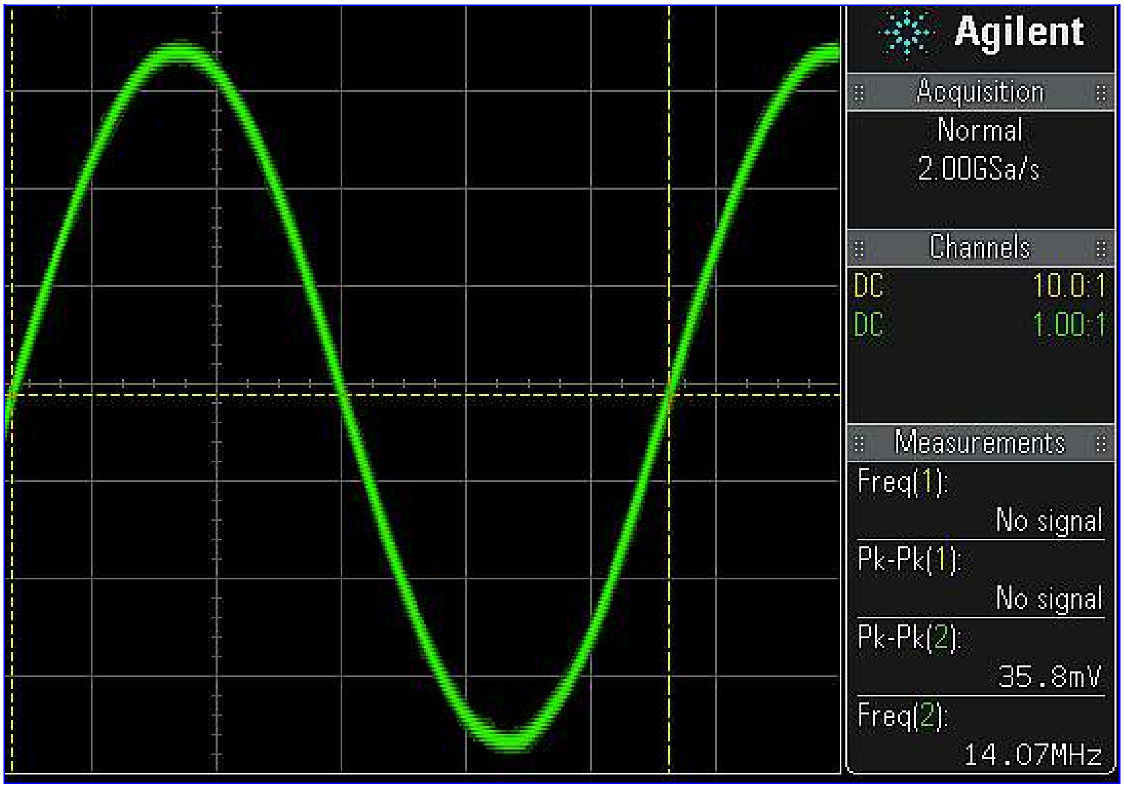

Normal
2.00GSa/s

DC 10.0:1
DC 1.00:1

Freq(1): No signal
Pk-Pk(1): No signal
Pk-Pk(2): 35.8mV
Freq(2): 14.07MHz

Shown above — The peak-peak voltage with a 14.07 MHz oscillator connected to the RF port; a 50 Ω terminated scope connected to the detector port; and a 27 Ω resistive terminator across the end of the unknown impedance cable**. The reflected signal from a 27 Ω resistor.**

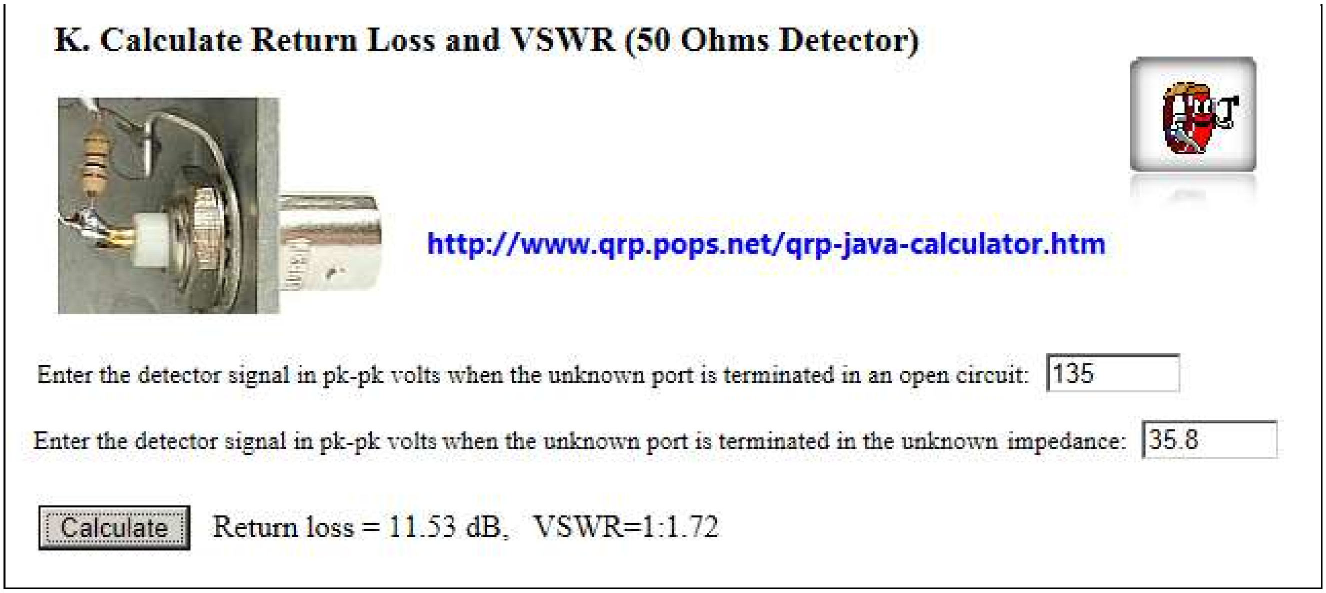

K. Calculate Return Loss and VSWR (50 Ohms Detector)

http://www.qrp.pops.net/qrp-java-calculator.htm

Enter the detector signal in pk-pk volts when the unknown port is terminated in an open circuit: 135

Enter the detector signal in pk-pk volts when the unknown port is terminated in the unknown impedance: 35.8

Return loss = 11.53 dB, VSWR=1:1.72

Shown above — Calculating the return loss and VSWR of the 27 Ω resistor with my JavaScript Applet **K**. mV versus volts peak-peak does not matter since we calculate a ratio of power.

Use the return loss measurement procedure depicted above to measure the return loss of any device you choose. If your D.U.T. has 2 ports, terminate the unmeasured port in 50 Ω. I show further RL measurment examples on RF Workbench 3.

I measured the return loss of some commercial gear in my shack and yard and will share 2 examples: 1) An expensive commercial transceiver I borrowed had an input port return loss of 15 dB (a 1.4:1 VSWR) The return loss of 15 dB indicates that the reflected wave power is 15 dB lower in power than the incident wave. 2) With a borrowed commercial bridge, my tuner-matched antenna revealed a return loss of ~60 dB.

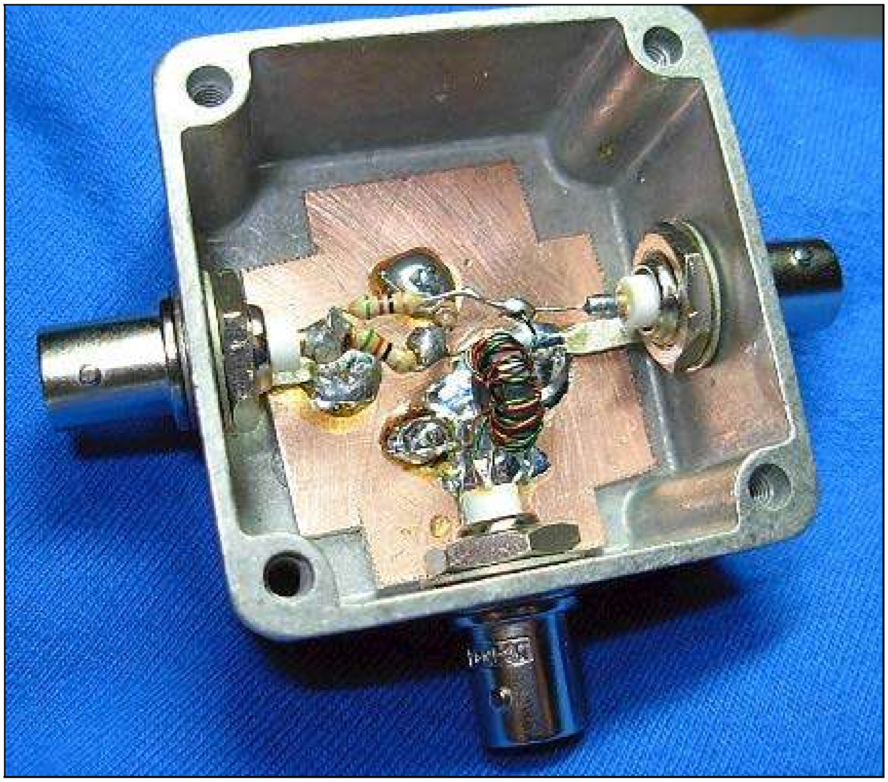

A RL bridge from my bench built with 51 Ω 5% tolerance resistors. I show a better RL bridge and some other experiments on RF Workbench 3.

Return Loss and the Attenuator Network (How Do Attenuator Pads Improve Component Matching?)

We routinely employ attenuator pads to increase return loss in a 50 Ω RF environment. For example, let's say you're testing a signal generator and measure a return loss of 6 dB. If you place a 10 dB attenuator pad after the signal generator, the return loss increases to 26 dB. If we used a 6 dB pad instead, the return loss would now = 18 dB. In both cases the return loss is increased by 2x the attenuator pad value. The doubling of return loss occurs because both the incident wave and reflected signals pass through the attenuator pad — that's how attenuator pads improve matching.

Attenuator pads reduce power, but that is why somebody invented the RF amplifier.

What is the minimally acceptable return loss for a device such as a signal generator? No single answer exists. The minimum return loss depends on the context: are you making precision circuits or just tuning an antenna?

Precision Circuits:

For amateur experimenter bench circuits, aim for a return loss of at least 20 dB. This often means adding an attenuator pad to the ports of your signal generator, amplifier, or other device to get a minimum 20 dB return loss. For an electronic engineer, the minimal return loss is probably higher; maybe 30 dB or so. I have read conflicting opinions about this and for some people — me included — design overkill is normal.

Antenna tuning:

When tuning an antenna for full transmitter output power, the minimal return loss is around 14 dB (a VSWR of 1:1.5). If you measure an antenna system return loss of 14 dB or better, the match is fine. Many Hams will protest a 1:1.5 VSWR and ardently chase a 1:1 VSWR on every frequency with their antenna tuner.

A Method to Measure Insertion Loss or Gain

Often, we want to measure the gain of an amplifier, or the insertion loss of a filter, or attenuator pad. I show how to do this with a 50 Ω terminated scope:

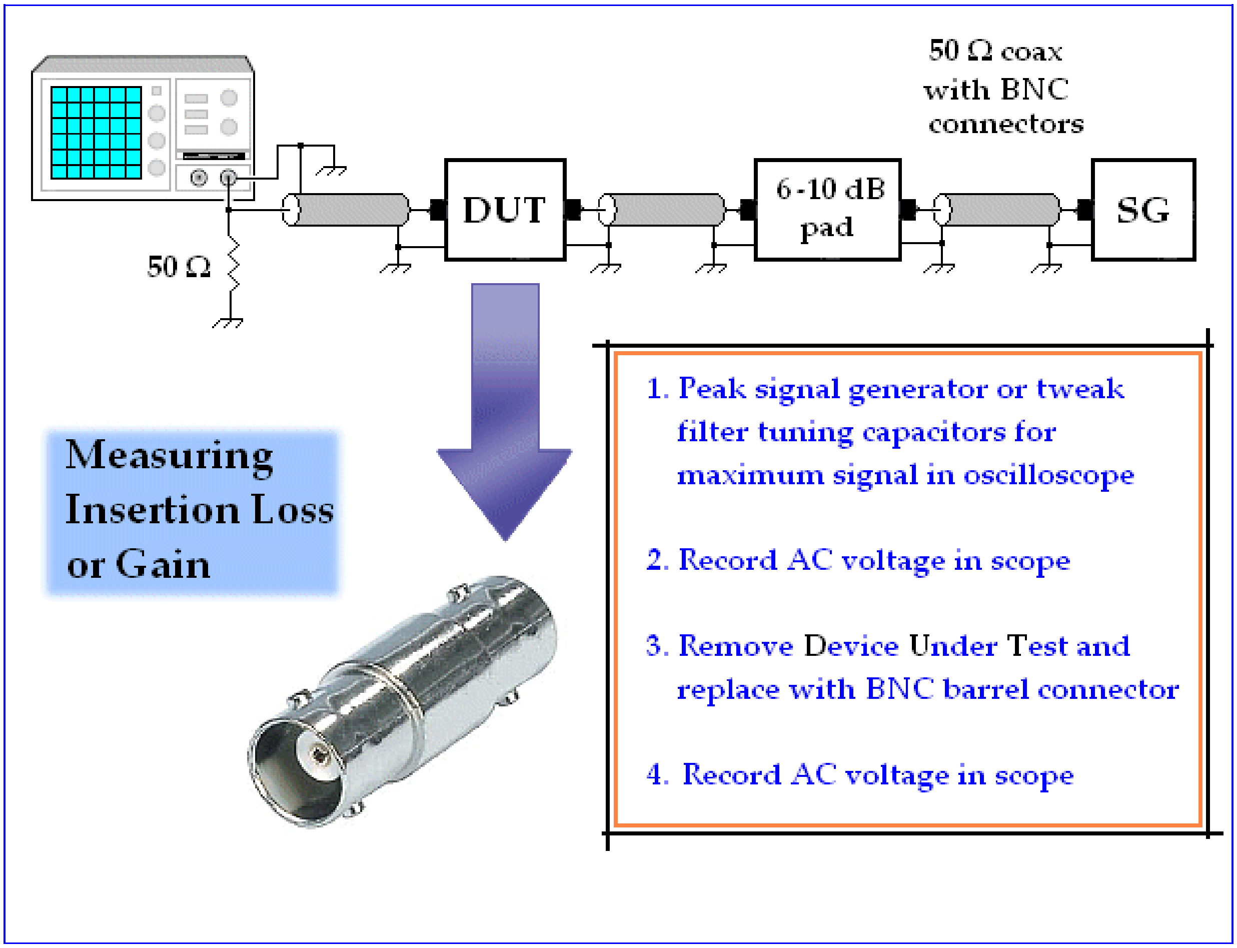

1. Peak signal generator or tweak filter tuning capacitors for maximum signal in oscilloscope

2. Record AC voltage in scope

3. Remove Device Under Test and replace with BNC barrel connector

4. Record AC voltage in scope

The circuit starts with a signal generator set to the frequency of interest. I show an attenuator pad in this diagram to stress that the signal generator output port must have a return loss >= 20 dB.

Connect the input of the 50 Ω **D**evice **U**nder **T**est to the generator output via 50 Ω coax

Connect the 50 Ω output of the D.U.T. to your 50 Ω terminated oscilloscope

Turn on the signal generator and if needed, peak the signal; In the case of a low-pass filter, the signal generator frequency control is tweaked to give a maximum pk-pk voltage in your 'scope. When evaluating a band-pass filter, tweak the filter trimmer capacitors for the maximum signal at the desired center frequency. Signal peaking ensures that losses caused by the filters are not caused by the filter mistuning, or in the case of the low-pass filter, to allow for cutoff frequency deviation caused by component value variations. It may be necessary to increase the signal generator amplitude to view a good quality signal in your 'scope.

Record the peak-to-peak voltage.

Remove the DUT and replace it with a BNC through-connector and record this peak-to-peak voltage.

Calculate the power in dBm of the 2 recorded voltages — their difference equals the insertion loss or gain in dB. I wrote a JavaScript Applet to do this. Click and scroll to **H**

This awesome measurement technique controls the input and output impedance and uses the same coaxial cables with and without the D.U.T. for accuracy. Some builders might choose to terminate the D.U.T. with a 50 Ω resistor and measure with a 10X scope. The capacitance of the probe may alter measurement in some cases. As always, choose your measurement technique based upon whatever gear you own and how exacting your standards are.

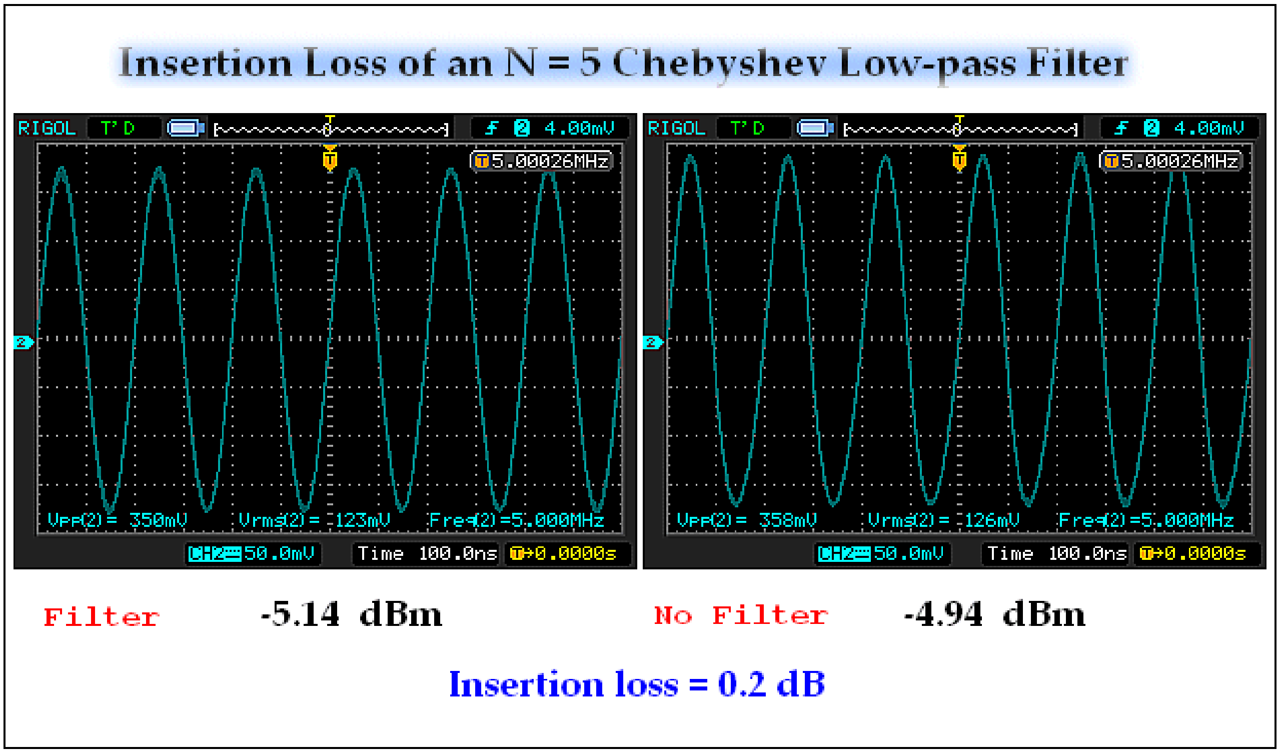

Spectrum Analyzers - Comments from the Workbench

Electronics professionals ruminate that spectrum analyzers are uncommon because experimenters perceive them as esoteric and difficult. My own opinion differs. Spectrum analyzers are relatively uncommon because of one reason - cost. I have watched prices on sites like eBay with amazement. The ads go something like this: 1.5 GHz spectrum analyzer for sale. Built in 1982. Ships in 2 pieces weighing over 22 kilograms. Minimum bid $1850.00. *And...sorry, I live in Florida, U.S.A. and in all likelihood, shipping these 2 heavy pieces is going to cost you a fortune. In the attached ad photos you can see lots of wear and tear, plus some screen burn-in on the display....* Guaranteed to turn on however!

Perhaps I exaggerate or even lampoon the perceived value of old boat anchor spectrum analyzers, but I have bought and sold cars for less money. Be prepared - spectrum analyzers are not cheap. They are however, very cool and open the door into a truly fascinating world. Frequency domain circuit measurement (spectrum analysis) addicts and intrigues. Homebuilding a spectrum analyzer is a serious option, but requires advanced building skills. Click and click for the W7ZOI/K7TAU project.

In recent times, the Rigol DSA-815 spectrum analyzer with tracking generator proved a game-changer to the bloated price of heavy, old and tired gear. Click for a Rigol datasheet. Signalhound also sells spectrum analyzers and tracking generators . A tracking generators *plus* spectrum analyzer allows you to sweep your device under test over a range of frequencies.

Prior to using a spectrum analyzer, I casually considered shielding stages or placing critical pieces in RF-proof boxes. Quickly I learned that RF in our home and community can and does get into your projects. p>

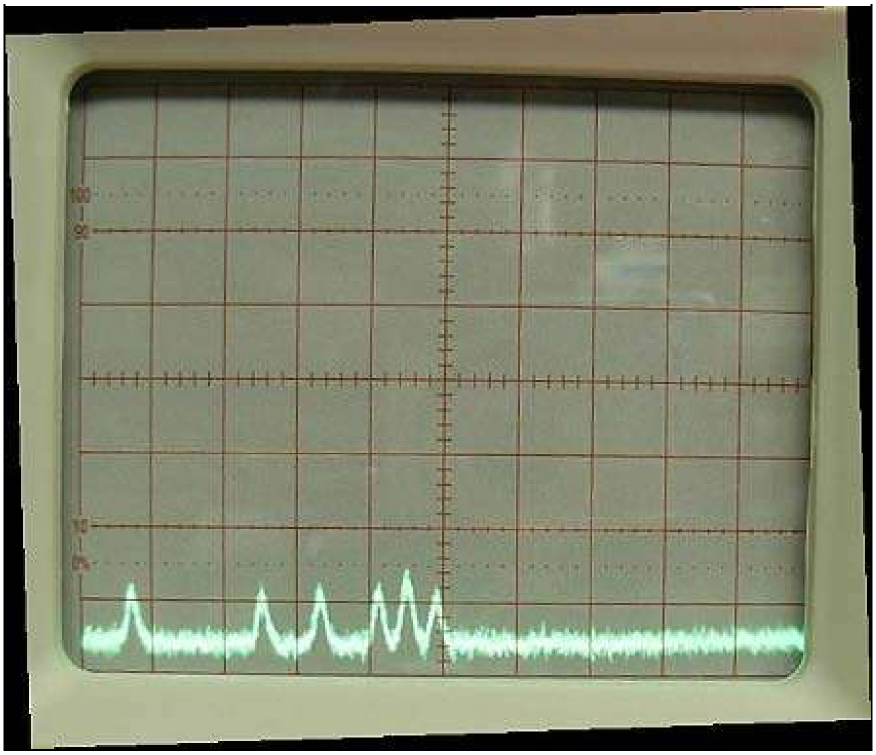

The center frequency of the display = ~150 MHz. The signal spikes appeared and disappeared after 4-9 seconds or so — after a little detective work with my scanner, I learned they were local police and ambulance FM radio conversations. I noticed this interference when I took the lid off a RF-tight band-pass filter — these signals arose in a 28 MHz superhet receiver !! While low in amplitude, experiences like this inform us to watch for lurking RFI.

I found numerous sources of RF in our home with a spectrum analyzer — the clothes washing machine during its spin cycle proved to be the worse RFI generator. RF-tight shielding with SMA, or BNC connectors and DC feed-through capacitors and aggressively decoupling and bypassing DC lines eliminated many of RFI problems during my experiments. I now better appreciate these anti-RFI techniques.

Spectrum Analyzer Calibrator

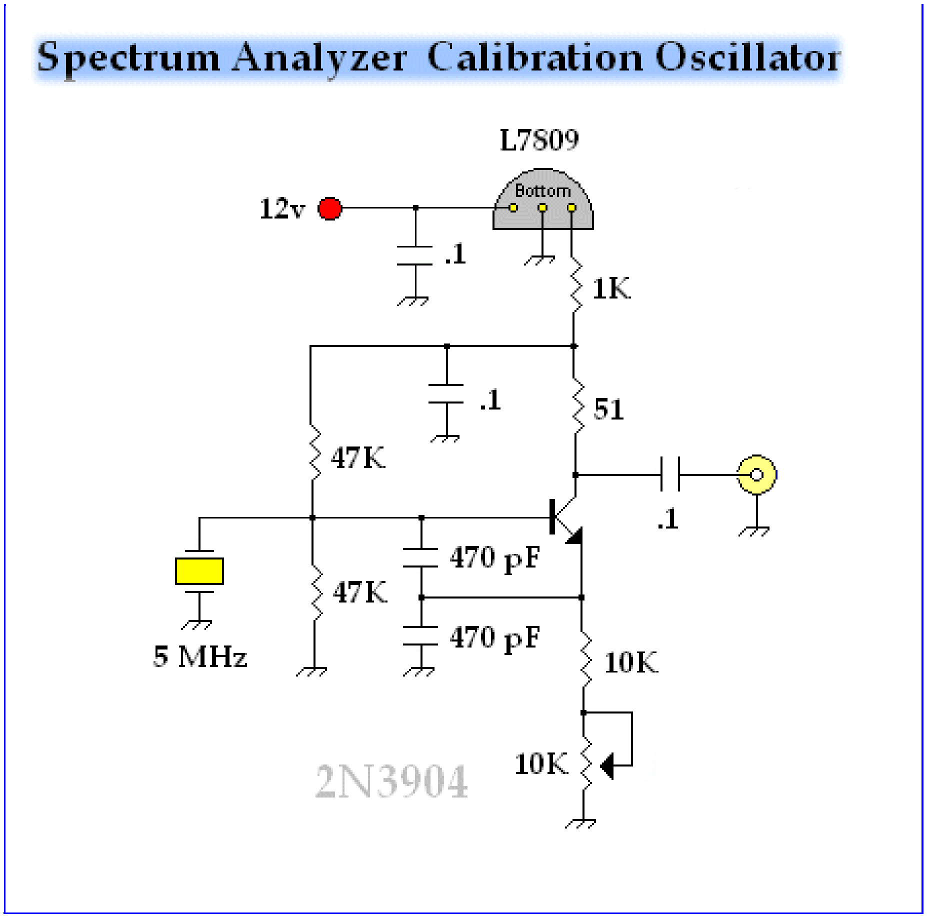
Spectrum Analyzer Calibration Oscillator
L7809
Bottom
12v
.1
1K
.1
51
47K
.1
470 pF
47K
470 pF
5 MHz
10K
10K
2N3904

A harmonic rich, spectrum analyzer calibrator designed by Wes, W7ZOI and displayed with his permission. Adjust the 10K potentiometer to provide the output power needed to calibrate your spectrum analyzer. I set mine to -27 dBm. Be careful when connecting signal generators to your spectrum analyzer, since a higher than rated input power may destroy the mixer/front-end of your spectrum analyzer and cost you dearly.

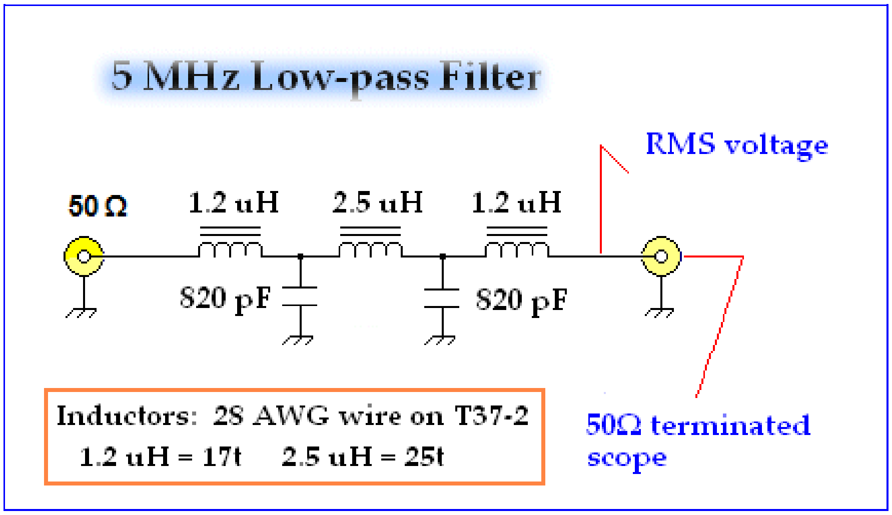
5 MHz Low-pass Filter
RMS voltage
50 Ω
1.2 uH
2.5 uH
1.2 uH
820 pF
820 pF
Inductors: 28 AWG wire on T37-2
1.2 uH = 17t
2.5 uH = 25t
50Ω terminated scope

I used this filter to set the -27 dBm power needed to calibrate my spectrum analyzer.

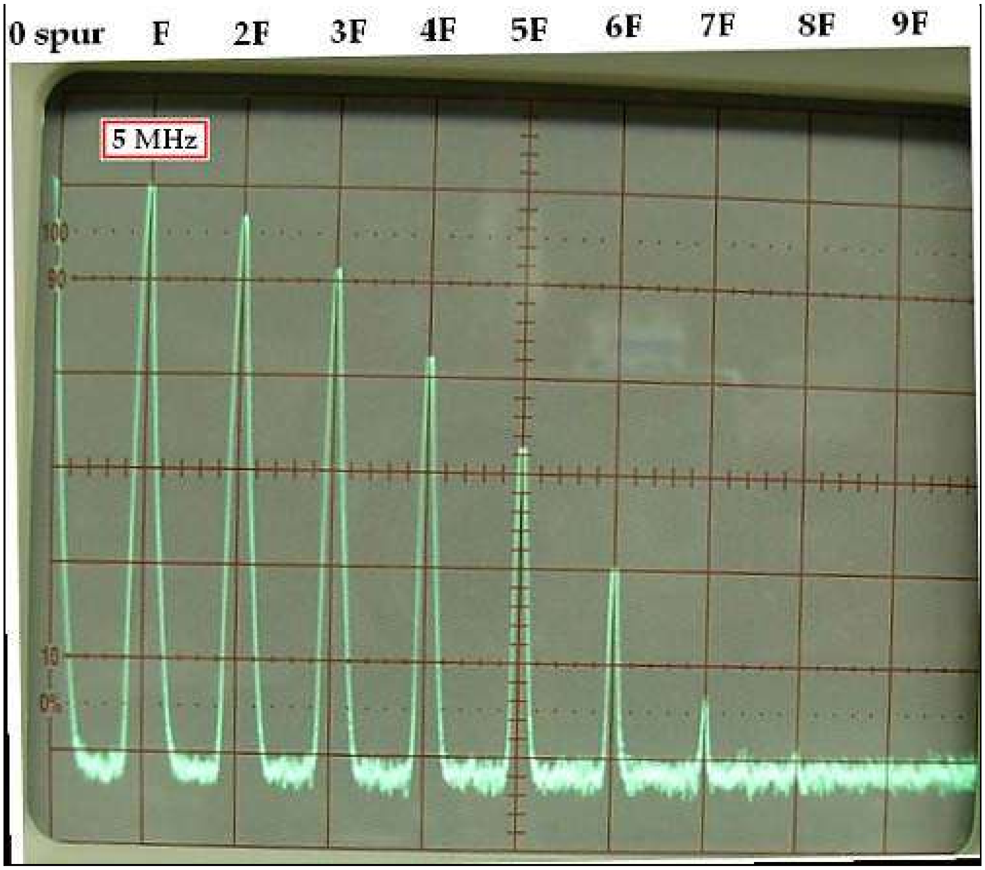

Spectrum analysis of the 5 MHz spectrum analyzer calibrator.

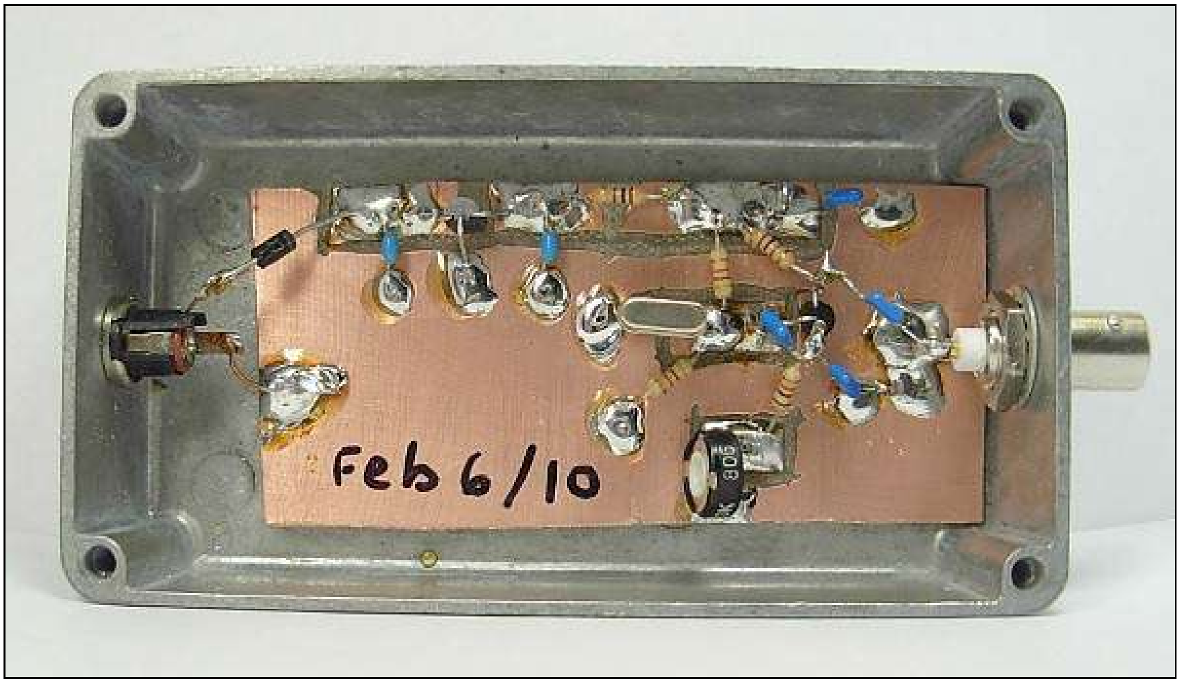

Breadboard of the 5 MHz spectrum analyzer calibrator.

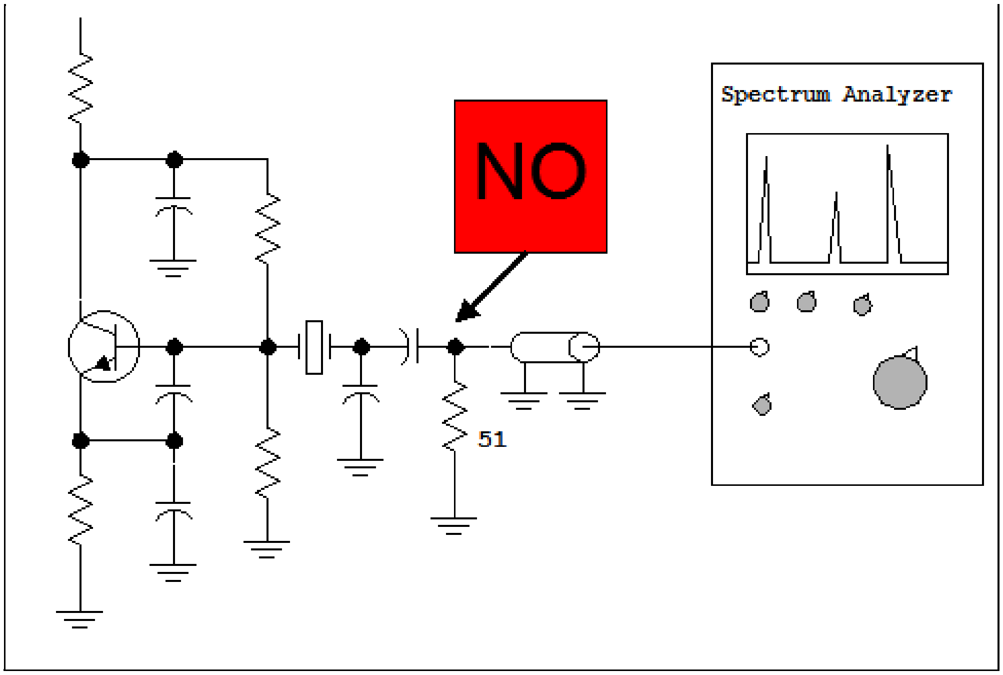
Spectrum Analyzer
NO
51

Don't use a "50 ohm" termination when measuring with a 50 Ω impedance spectrum analyzer.

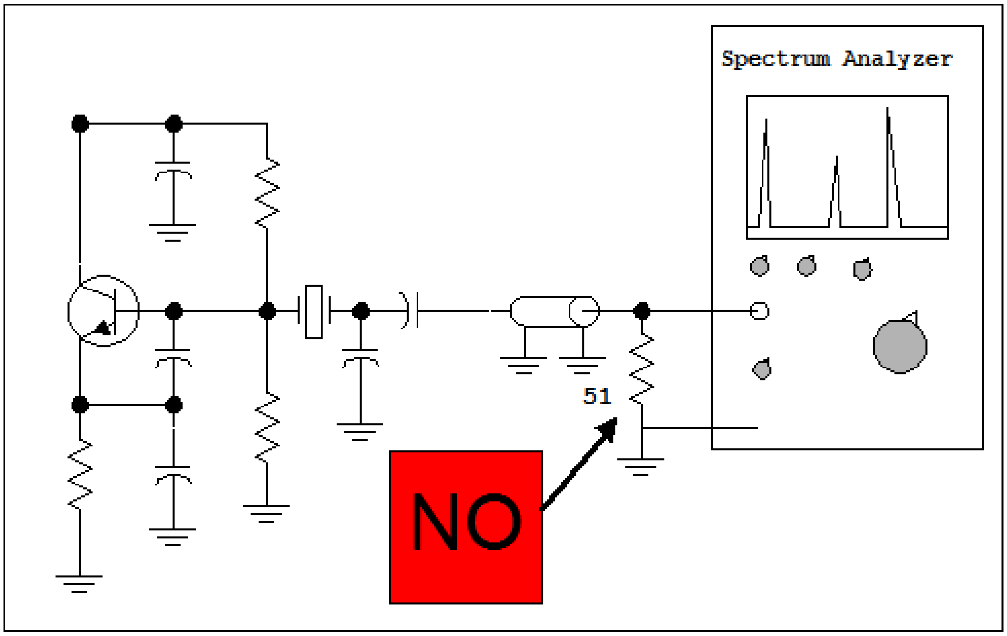
Spectrum Analyzer
51
NO

No resistor is required, as the the input impedance of the SA is 50 Ω.

Miscellaneous Photos and Notes

-10 DB PAD
50 OHMS
RETURN-LOSS BRIDGE
RF
?
DETECTOR
MPLER
S

Some of the 50 Ω modules built during the RF Workbench page 1 and 2 experiments

1
3
6
9
30 dB
10 WATT LOAD

15542
HAT-6+
6 dB ATTENUATOR
50 OHM
15542
VAT–6
6 dB ATTENUATOR

QRP SWL HOMEBUILDER

Amateur Radio Electronic Design

Audio Transistor Input Impedance Experiments

Introduction

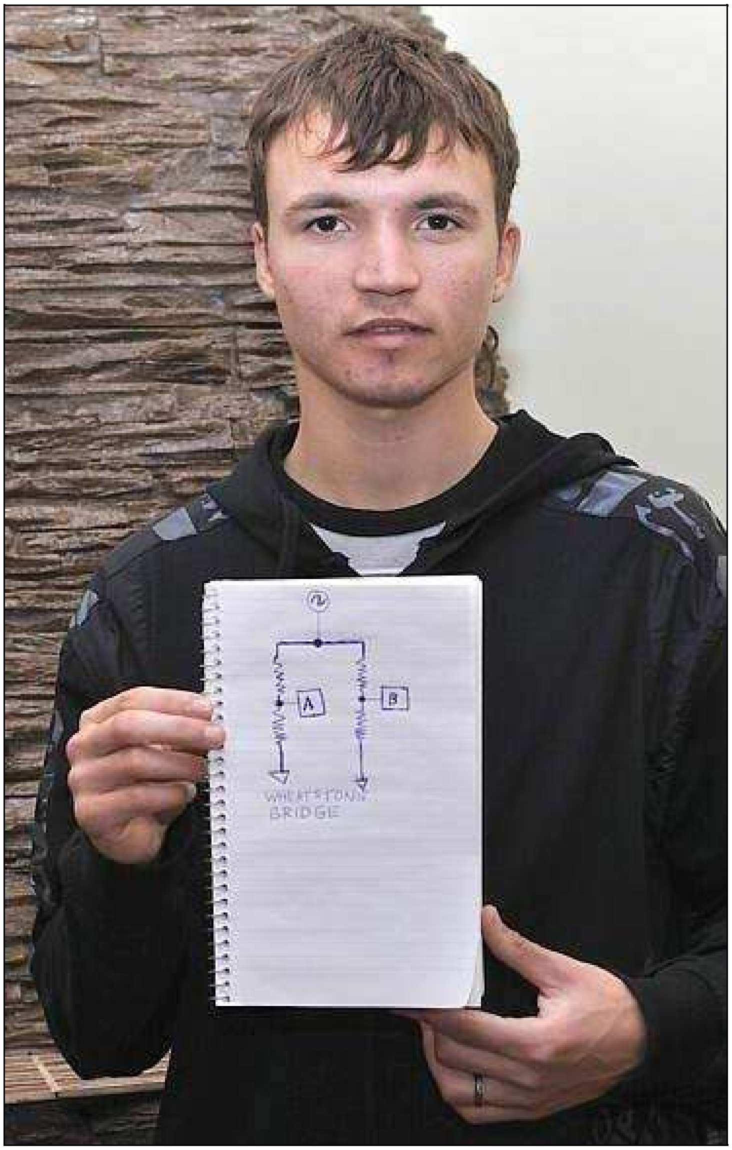

I examined audio transistor amplifier input impedance during Spring — Summer 2010 and generated enough content for a web page.

On this web page, I explore determining AF amplifier input impedance by using network theory and calculation, plus direct measurement with instruments containing a Wheatstone bridge. This content emphasizes learning through performing bench experiments and I hope it sparks your own experiments and research into impedance measurement test equipment and theory.

Many RF circuits require termination with stages containing a well defined input impedance. Consider, for example, amplifiers that follow L-C low-pass filters, diode ring mixers or crystal filters — a known impedance (usually 50 ohms) must terminate these stages to optimize return loss. A special case is the diode ring product detector; which must be followed by a 50 ohm input impedance audio amplifier. How do we design or assess a small-signal audio amplifier that has a 50 ohm input impedance? This question spawned every experiment on this web page — the content grew and evolved along with my understanding of this topic.

Audio transistor input impedance may be calculated with equations or software, however, doing the math or using or affording these programs might be problematic for some amateur builders. Additionally, component variances such as transistor Beta and different power supply voltages can causes significant differences between the theoretical and the actual input impedance realized. Further, amplifiers, such as a feedback pair that involve combinations of series or shunt feedback can be difficult to analyze accurately using equations during small or large signal analysis. It may be easier for experimenters to just measure and tweak amplifier components on the bench — the focus of this web page!

As an rank amateur, I have much to learn, and by no means am an expert in electronic design. If you see an error on this web site, disagree with my analysis, or have suggestions for improvement, please email them — I am an amateur hobbyist, who earns no money from this site, and who relies on the assistance of others to keep the content as accurate and vibrant as possible.

The topics:

Part 1: Some basic transistor network theory and how to calculate input Z

- Part 2: 50 ohm input impedance Wheatstone bridge measurement
- Part 3: Measuring unknown impedances
- Part 4: Miscellaneous circuits, scans and photographs.

Notes:

Small signal analysis refers to modeling or examining an amplifier at a single operating point (its bias point) and applying linear equations which assess the amplifier with no signal applied. In small signal analysis, we assume that the signal is so small that transistor gain, capacitance and other factors are static.

Part 1: Some Basic Transistor Network Theory and How to Calculate Input Z

Some basic network theory plus methods to calculate the input impedance of common emitter and common base audio amplifiers.

Three basic small-signal transistor parameters include *beta,* emitter resistance *Re*, and bulk resistance *REB*.

Beta is the term used to designate the current gain of a common emitter circuit — it's the ratio of collector (output) current to base (input) current.

Small-signal emitter resistance; Re = 26 / IE ,or, 26 divided by the emitter current in millamperes. For example, an emitter bias current of 0.52 mA gives a small-signal emitter resistance of 50 ohms, or visa versa . Re is the resistance seen looking into the emitter whether the stage input is the transistor base or emitter terminal. Re is the dynamic resistance of the input junction due to carrier action.

REB represents the bulk resistance of the semiconductor not arising from contact resistance; in other words, it's the DC resistance of the base and emitter leads plus the pn junction. Typically REB = 2 to 6 ohms and is often ignored (your choice) when the current is low — say, for example, < 9 mA for a typical common-emitter voltage amplifier. In large power transistors or for switching operations, the typical REB value may vary. REB, in part, limits the maximal gain of a transistor.

The constant 26 used when calculating the dynamic resistance of a forward-biased PN junction is derived via calculus. Professor Kuhn's website link containing the math.

There is also a base spreading resistance generally known as 'rbb' that, in effect exists laterally across the transistor. A simple model puts rbb at about 100 ohms in series with the base and it's one of the causes of finite transistor frequency response. While interesting, rbb isn't discussed further.

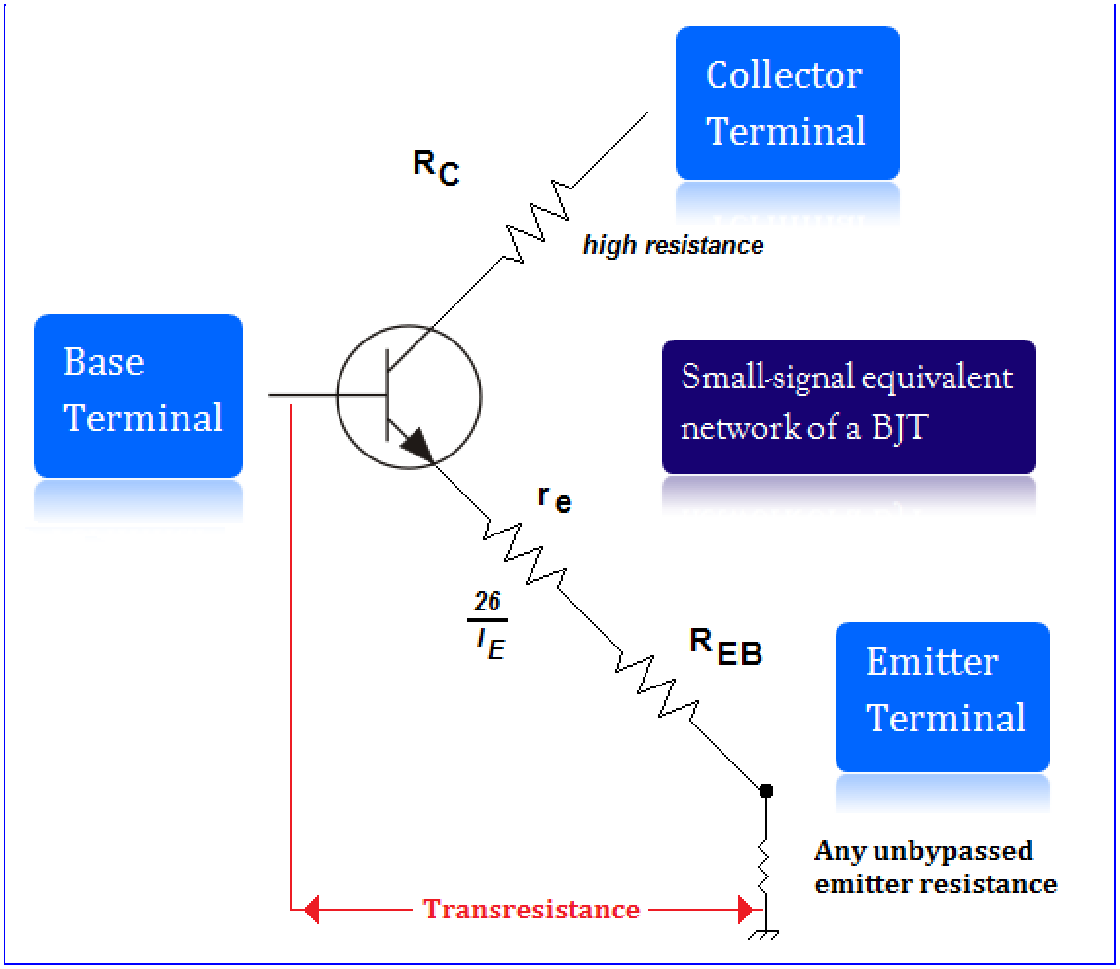
Collector Terminal
R_C
high resistance
Base Terminal
Small-signal equivalent network of a BJT
r_e
$\frac{26}{I_E}$
R_{EB}
Emitter Terminal
Any unbypassed emitter resistance
Transresistance

Above — the small-signal equivalent network of any transistor. re = 26/IE. Also, re + REB + any unbypassed external resistor may be termed the *Transresistance*, a DC ohmic value representing the total resistance of the emitter. The collector resistance RC is high because of its reverse bias. Collector resistance is not considered when calculating input impedance of simple AF transistor stages.

Calculating the input resistance of a common base stage

Calculating the input impedance of a common base amplifier is easy. Input impedance = 26/ emitter current (IE). You can either bench measure or calculate the emitter current using DC analysis. Click for the formula to calculate emitter current . A complete example follows:

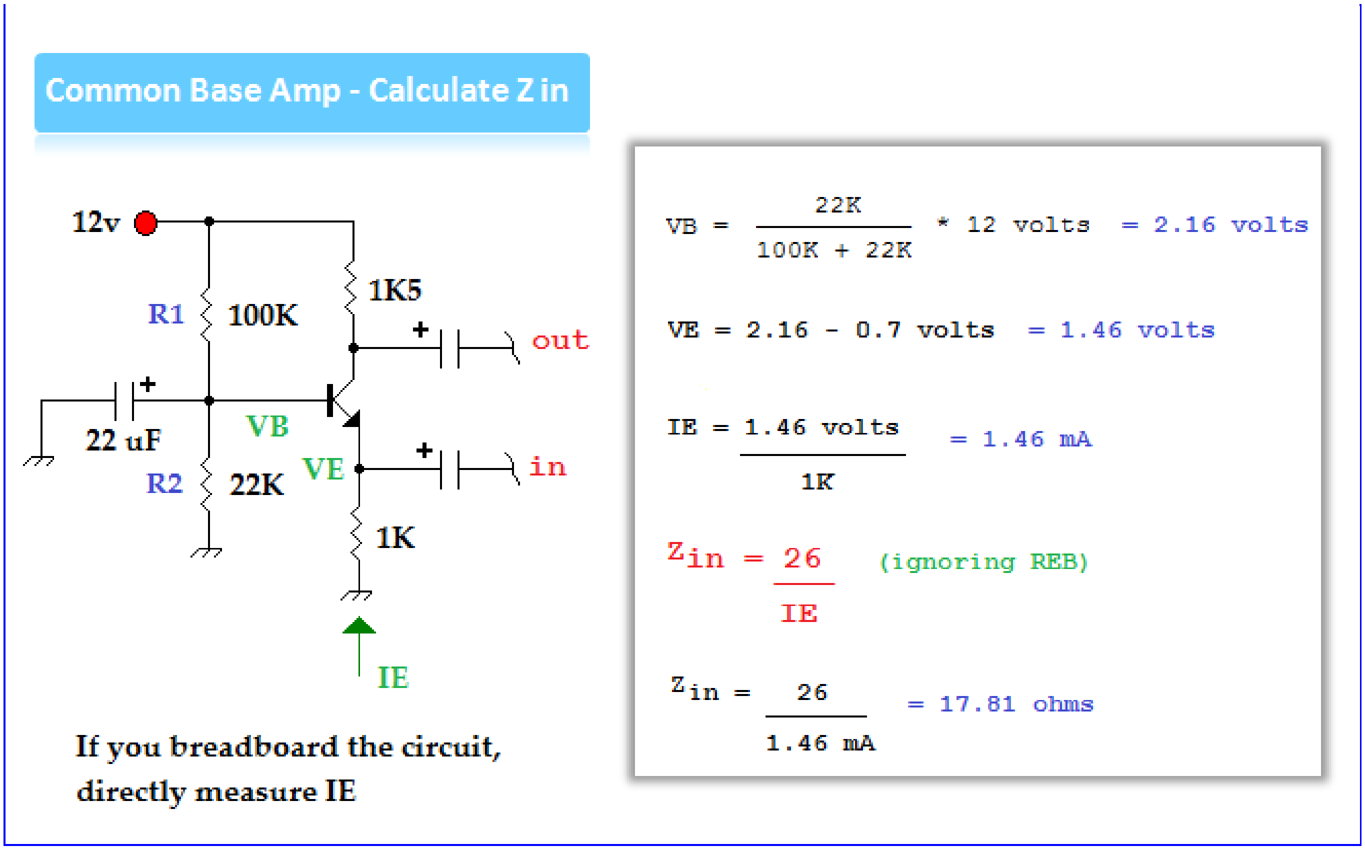
Common Base Amp - Calculate Z in

12v
R1
100K
1K5
out
22 uF
VB
R2
22K
VE
in
1K
IE

If you breadboard the circuit, directly measure IE

$$VB = \frac{22K}{100K + 22K} * 12 \text{ volts} = 2.16 \text{ volts}$$

$$VE = 2.16 - 0.7 \text{ volts} = 1.46 \text{ volts}$$

$$IE = \frac{1.46 \text{ volts}}{1K} = 1.46 \text{ mA}$$

$$Z_{in} = \frac{26}{IE} \quad \text{(ignoring REB)}$$

$$Z_{in} = \frac{26}{1.46 \text{ mA}} = 17.81 \text{ ohms}$$

Above — An example common base amplifier and its input impedance calculation. In this example, emitter current is calculated using DC analysis. On the bench, it's better to un-ground the 1K emitter resistor and connect your ammeter between this resistor and ground to directly measure IE. REB was ignored and = 0.

Consider the 50 ohm input Z common base amplifier we often use after a diode ring product detector plus diplexer:

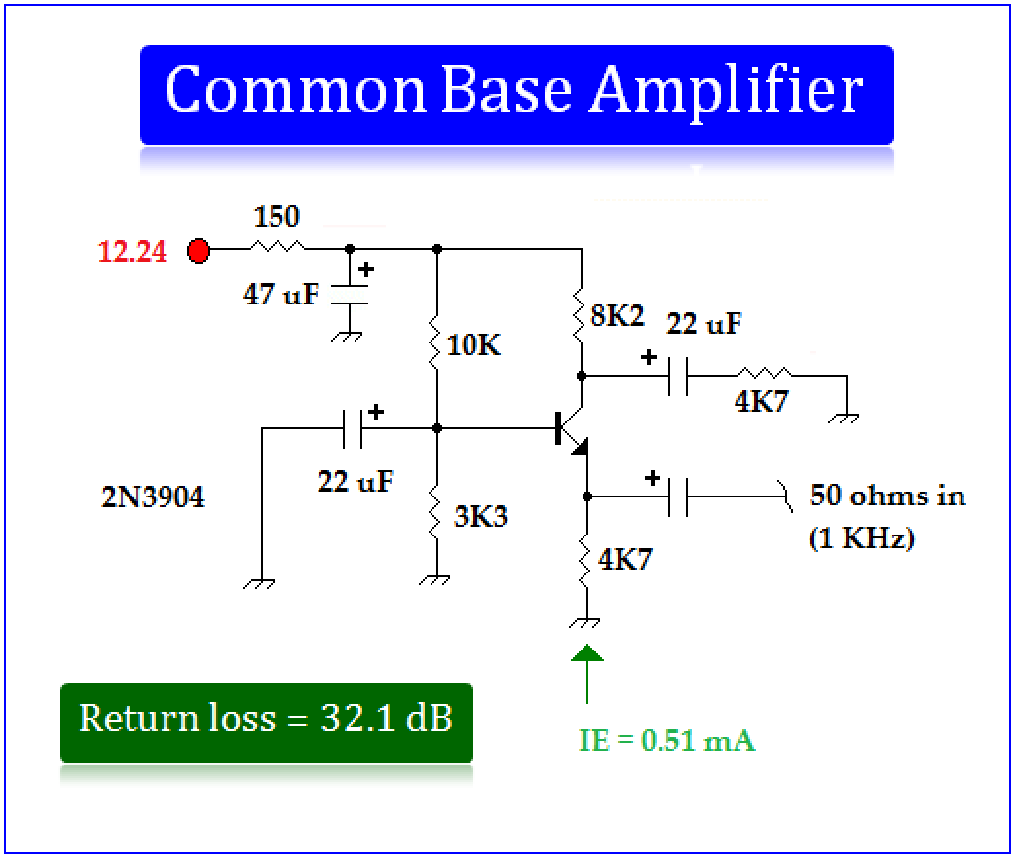
Common Base Amplifier

150
12.24
47 uF
10K
8K2
22 uF
4K7
22 uF
2N3904
3K3
50 ohms in
(1 KHz)
4K7
IE = 0.51 mA

Return loss = 32.1 dB

Above — A common base amplifier built for a direct conversion receiver in Spring 2010. This amplifier is shown in test setup for bench analysis

— with a DC decoupling network and an AC coupled 4K7 resistive load. The emitter current established using 5% tolerance resistors was 0.51 mA. Therefore, the calculated input Z is 26/0.51 = 51 ohms. The return loss of this amplifier as measured with the *active* 50 ohm Wheatstone bridge device described later on this web page was a spectacular 32.1 dB! If a different power supply voltage or biasing/emitter resistors were used, the IE would change and along with IE, the input impedance and return loss.

This amp illustrates that testing and tweaking AF amplifiers on the bench will garner the best results. If I just copy someone else's design; perhaps with a different DC voltage, or decoupling network and don't adjust the emitter current by tweaking the base biasing or emitter resistor resistors, the input impedance could differ significantly. Whenever you build a common base amplifier, measure its DC current and as necessary, tweak resistors to get the current needed for a perfect impedance match. It is good practice to measure all the DC voltages and emitter current on any amplifier you build — you will learn what is normal, what to expect and perhaps detect errors or parts failure(s).

Performing return loss measurement is also a fantastic way to ensure good matching to the 50 ohm impedance diode ring product detector that feeds this amplifier.

Calculating the input resistance of a common emitter stage

Calculating the input impedance of a common emitter amplifier is also straight forward, but not as easy as the common base amplifier.

In the **common base amplifier**, the emitter is the input element, therefore the input signal resistance is 26 / IE + REB. Often we ignore REB. If current of the common base amplifier is for example, 2 mA or so, then the 2-6 ohms of REB may be significant as 26 / 2 mA = 13 ohms. REB may be a factor because 2-6 ohms is a significant percentage of the total input resistance.

For a **common emitter amplifier**, the input resistance looking into the base is Beta (26/IE + REB + RE). Again REB is often ignored. We need to include any transistor DC biasing resistors which are also seen by the input signal as it moves through the transistor base. An example follows:

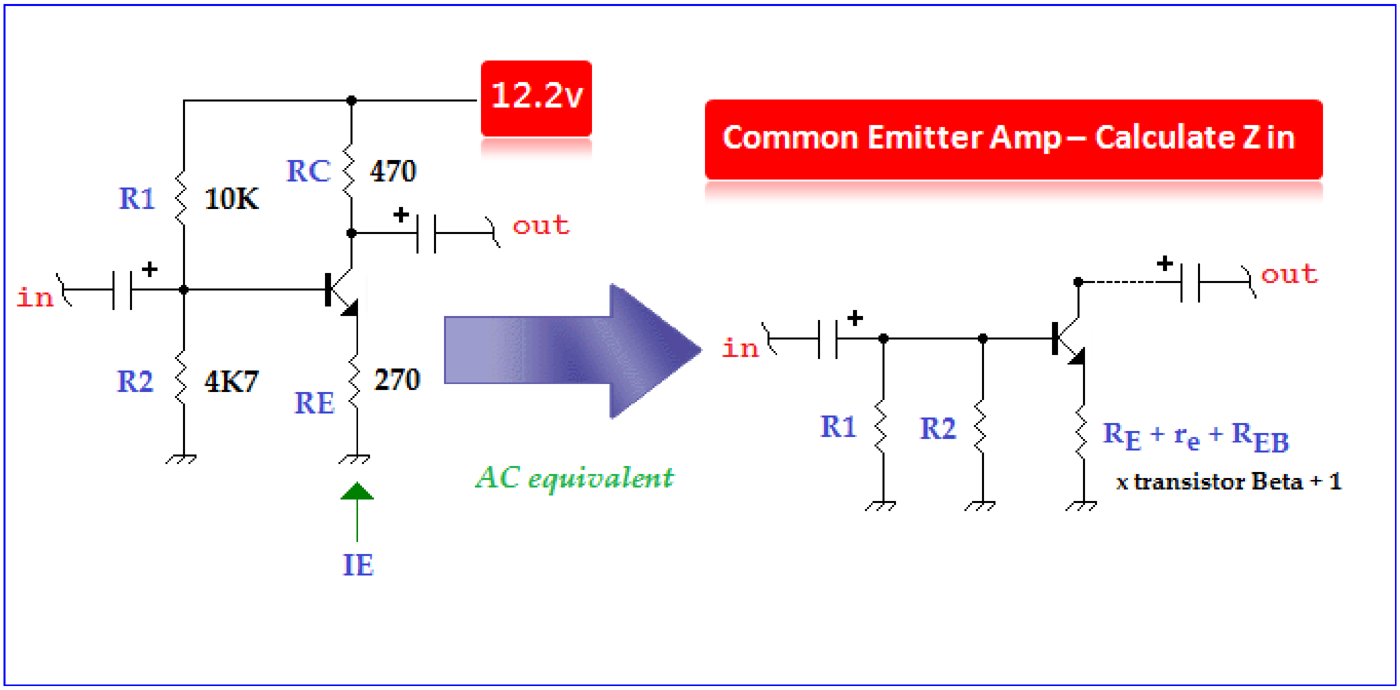

Above — An example common emitter amplifier re-drawn to illustrate how the input resistances combine to provide the AC input impedance. In this case, the 270 ohm emitter resistor RE is un-bypassed. R1, R2 and the components RE, re and REB are in parallel as the DC supply acts as a short to ground for the AC input signal. The components RE, re and REB (if used) must be multiplied by the transistor Beta value (+ 1) since the resistance looking into the base is Beta times that looking into the emitter.

Therefore: Rin = (B+1)*(re + REB + RE'). Normally we ignore REB so practically speaking Rin = (B+1)*(re + RE')

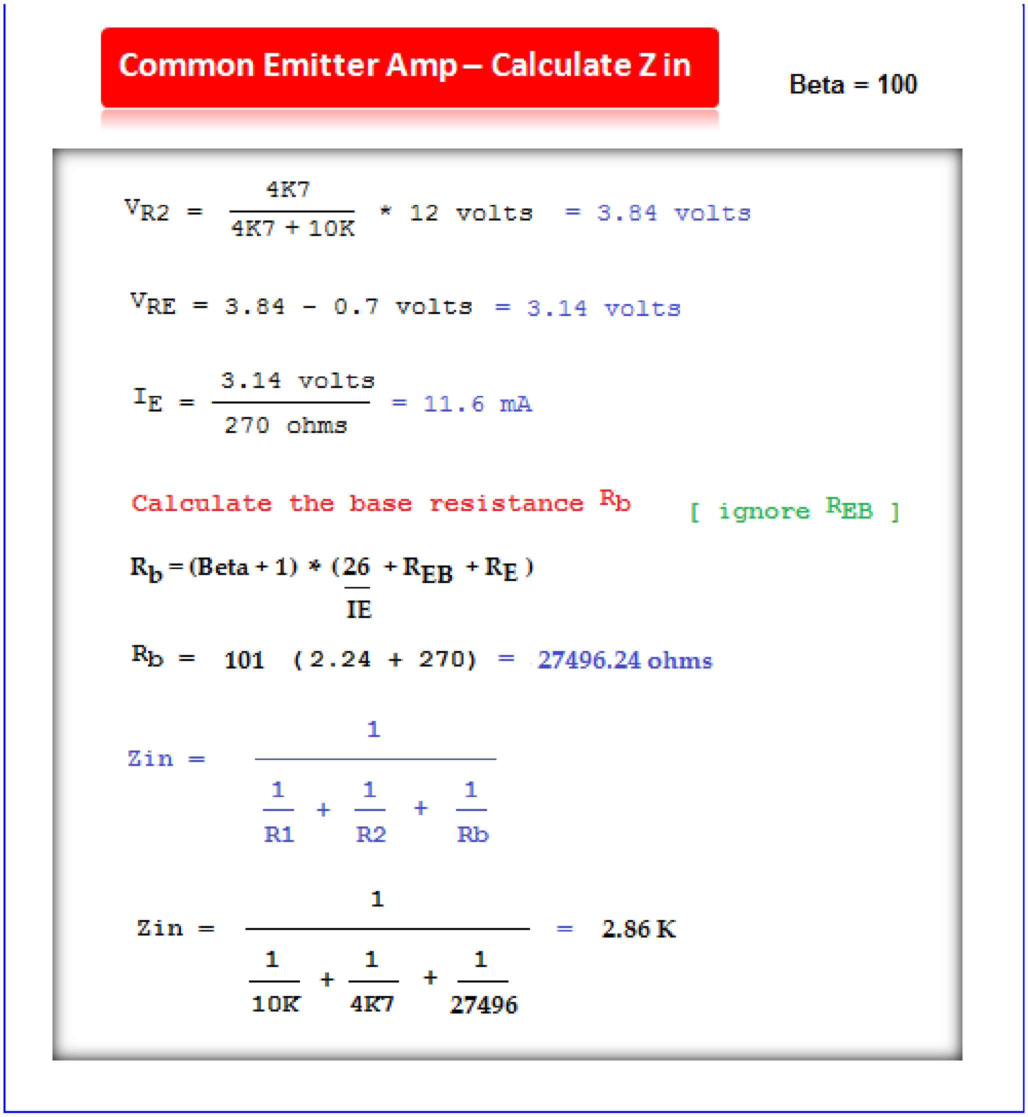

Common Emitter Amp – Calculate Z in

Beta = 100

$$V_{R2} = \frac{4K7}{4K7 + 10K} * 12 \text{ volts} = 3.84 \text{ volts}$$

$$V_{RE} = 3.84 - 0.7 \text{ volts} = 3.14 \text{ volts}$$

$$I_E = \frac{3.14 \text{ volts}}{270 \text{ ohms}} = 11.6 \text{ mA}$$

Calculate the base resistance R_b [ignore R_{EB}]

$$R_b = (\text{Beta} + 1) * \left(\frac{26}{IE} + R_{EB} + R_E\right)$$

$$R_b = 101 \ (2.24 + 270) = 27496.24 \text{ ohms}$$

$$Zin = \frac{1}{\frac{1}{R1} + \frac{1}{R2} + \frac{1}{Rb}}$$

$$Zin = \frac{1}{\frac{1}{10K} + \frac{1}{4K7} + \frac{1}{27496}} = 2.86 \text{ K}$$

Above — The math for the common emitter circuit shown directly above using DC analysis to calculate the current. On the bench, we just measure the emitter current (no need to calculate it). We assume IE = IC for a common emitter amplifier. REB = 0 (when ignored). If the 270 ohm resistor RE was bypassed with an electrolytic capacitor, the 270 ohm resistance would also = 0; and then Rb = Beta + 1 * (26/IE).

Conclusion

This theory explains how to calculate input impedance in 2 basic transistor AF amplifiers. Consult an electronics text for further explanation. Although the arithmetic is simple, quite frankly, it's a little boring. Let's go to the bench and have some fun. I was quite naive about measuring AF amplifier input impedance; however, my experiments yielded some knowledge and a strong appreciation for the Wheatstone bridge network. Onward...

Part 2: 50 ohm Input Impedance Wheatstone Bridge Measurement

Testing for a NULL or measuring the return loss of AF amplifiers with a 50 ohm input impedance.

Refer to the diagram on the right. Redrawn in a way more familiar to builders, the Wheatstone bridge network is just a pair of voltage dividers in parallel. We measure the difference in AC voltage between the ports labeled Out 1 and Out 2. The bridge is said to be balanced and produce NULL or 0 output when Out 1 and Out 2 are equal in voltage. Another description — when in perfect balance, the signal loss due to mismatch between the output ports is infinite. However, if this balance is disturbed by a mismatch between ports Out 1 and Out 2, an AC voltage appears and the return loss decreases in proportion to the mismatch (within limits and providing your instrument can measure accurately).

Let's focus on some practical bench applications in the new millennia: On your bench, you might employ a Wheatstone bridge network to measure return loss (or VSWR) or to simply to detect a NULL indicating a close impedance match between 2 stages. Specific examples include tuning your feed line and antenna, checking the match between a signal generator and a filter, or measuring an audio amplifier input impedance. In my estimation, the Wheatstone bridge lies among the most important test circuits in the amateur designer's arsenal; worthy of study and experimentation.

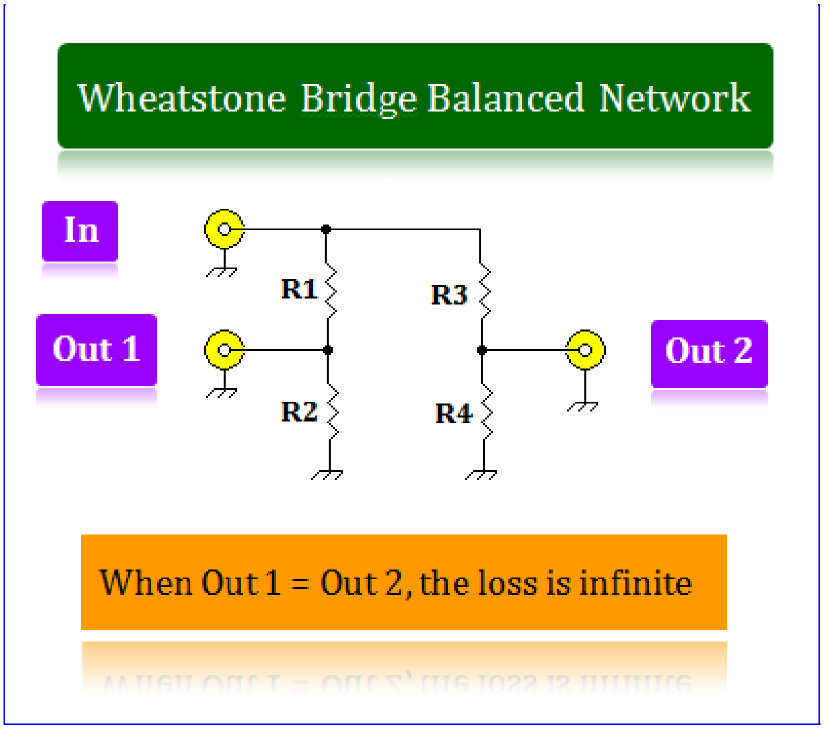
Wheatstone Bridge Balanced Network
In
Out 1
Out 2
R1
R2
R3
R4
When Out 1 = Out 2, the loss is infinite

Notes:

- The input signal can be AC or DC, but all discussion is confined to an AC signal source
- E96 (1%) metal film resistors were used in all Wheatstone bridges
- All bridges were tested at 1 KHz
- Ensure you do not overdrive your bridge; lest distortion occur! When a bridge is overdriven, you might cancel the fundamental frequency when balancing the bridge, but not the harmonics! Therefore, parasitic harmonics appear in the output that skew the the NULL or return loss values. I learned low pass filtering the amplified bridge output is really important.
- Any distortion in the bridge output means you must reduce your input signal drive level; however, this may reduce the accuracy of the RL measurements. There is no free lunch! You generally want just enough input signal to accurately measure the signal with the bridge at NULL.

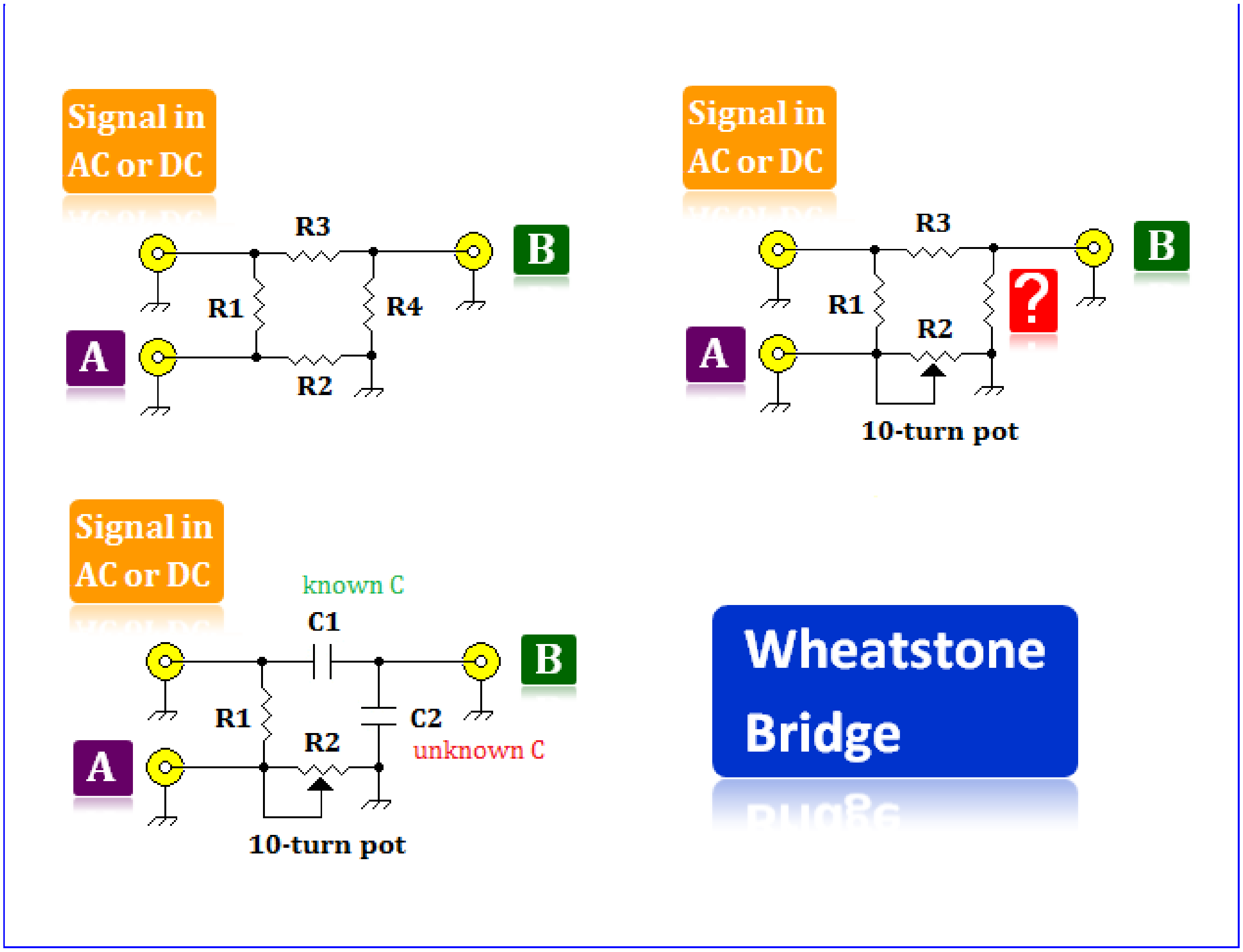

Comments from the Workbench

Above — an evolution of the 4 resistor bridge into a device to measure impedance and capacitance. In its classic form, each bridge leg is a resistor voltage divider with a detector connected across ports A and B. If ports A and B have equal voltages and R1 = R2, then R3 and R4 must also be equal; the bridge lies balanced or in a NULL state. If you remove R4 and measure an unknown resistance, the bridge will return to balance after adjusting pot R2 to equal the unknown resistance. In most cases, R2 is calibrated and the impedance is read directly off the potentiometer dial. Bridges can be arranged to measure unknown capacitance, inductance , frequency and other parameters by using precision 1% fixed components, calibrating the 10 turn pot to indicate the desired parameter, or for deriving the unknown value via equations.

Builders of lore used bridges to quantify many values on the bench. Although we have better ways to measure inductance and capacitance today, the Wheatstone bridge is still the *king* when it comes to simple measurement of network impedances; for example, QRP antenna tuners. Some builders use an LED to indicate bridge imbalance.

Building a passive instrument to measure the return loss of a 50 ohm input Z audio preamplifier.

Non-radio folks don't generally understand this — to properly terminate a diode ring mixer, 50 Ω impedance stages are needed. The inspiration driving *all* the experiments on this web page was to design a 50 ohm input impedance common emitter audio preamplifier to follow a diode ring mixer. I could have just used the familiar common-base amplifier popularized by Roy, W7EL, but of course, wouldn't learn anything. Somehow, I became drawn in by curiosity and generated enough content to fill a whole web page.

I decided to try and build some return loss bridges and test them by using known, fixed-value resistors as the unknown impedance. My first bridge, was an AF version of this RF return loss bridge using a junk box 600 ohm, 1:1 audio isolation transformer. It didn't work until I rearranged it as shown in the schematic below.

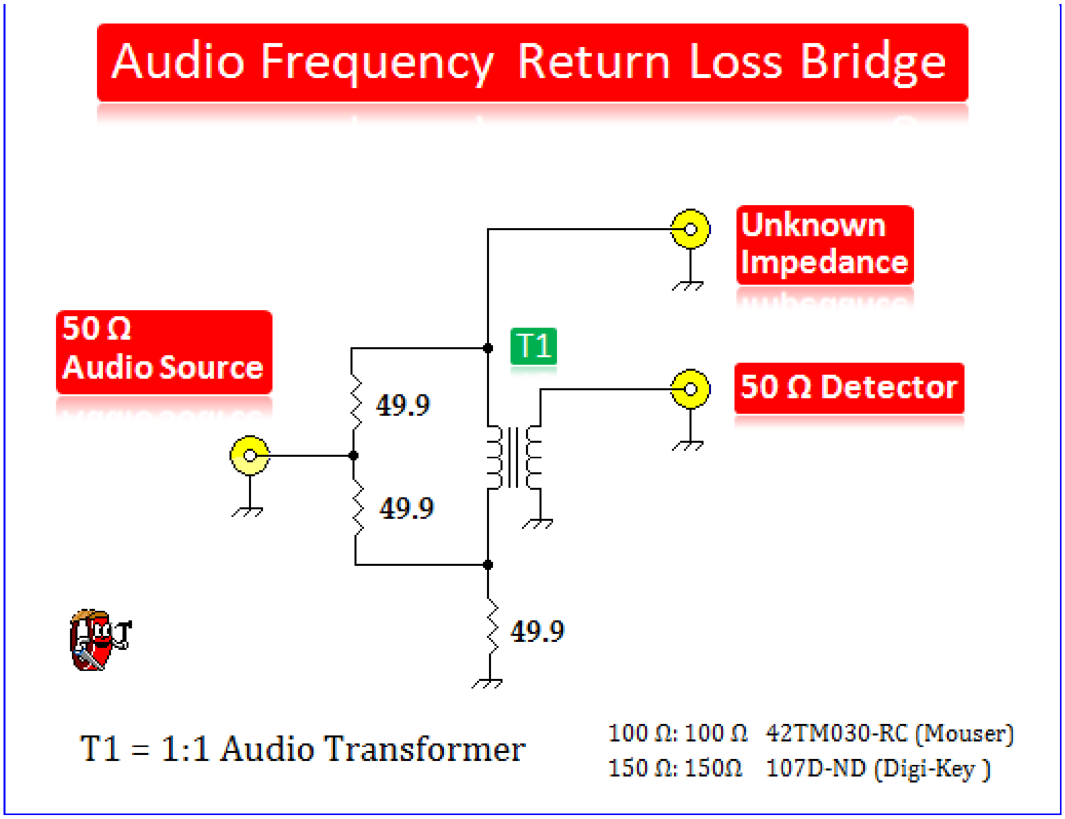
Audio Frequency Return Loss Bridge
Unknown Impedance
50 Ω Audio Source
T1
50 Ω Detector
49.9
49.9
49.9
T1 = 1:1 Audio Transformer
100 Ω: 100 Ω 42TM030-RC (Mouser)
150 Ω: 150Ω 107D-ND (Digi-Key)

Above — A simple return loss bridge using an AF transformer and 50 ohm detector. Suitable detectors are described here in the section covering return loss bridges — I used a 50 ohm terminated scope. Using 20 log (peak-peak voltage) to crunch the 50 Ω AC voltage into dB, the bridge was measured at open circuit, plus with various fixed 5% tolerance resistors terminating the Unknown Z port. Using a junk box 600 ohm 1:1 AF transformer, my results initially seemed good, but upon analysis were fraught with error. Note the suggested transformers in the schematic.

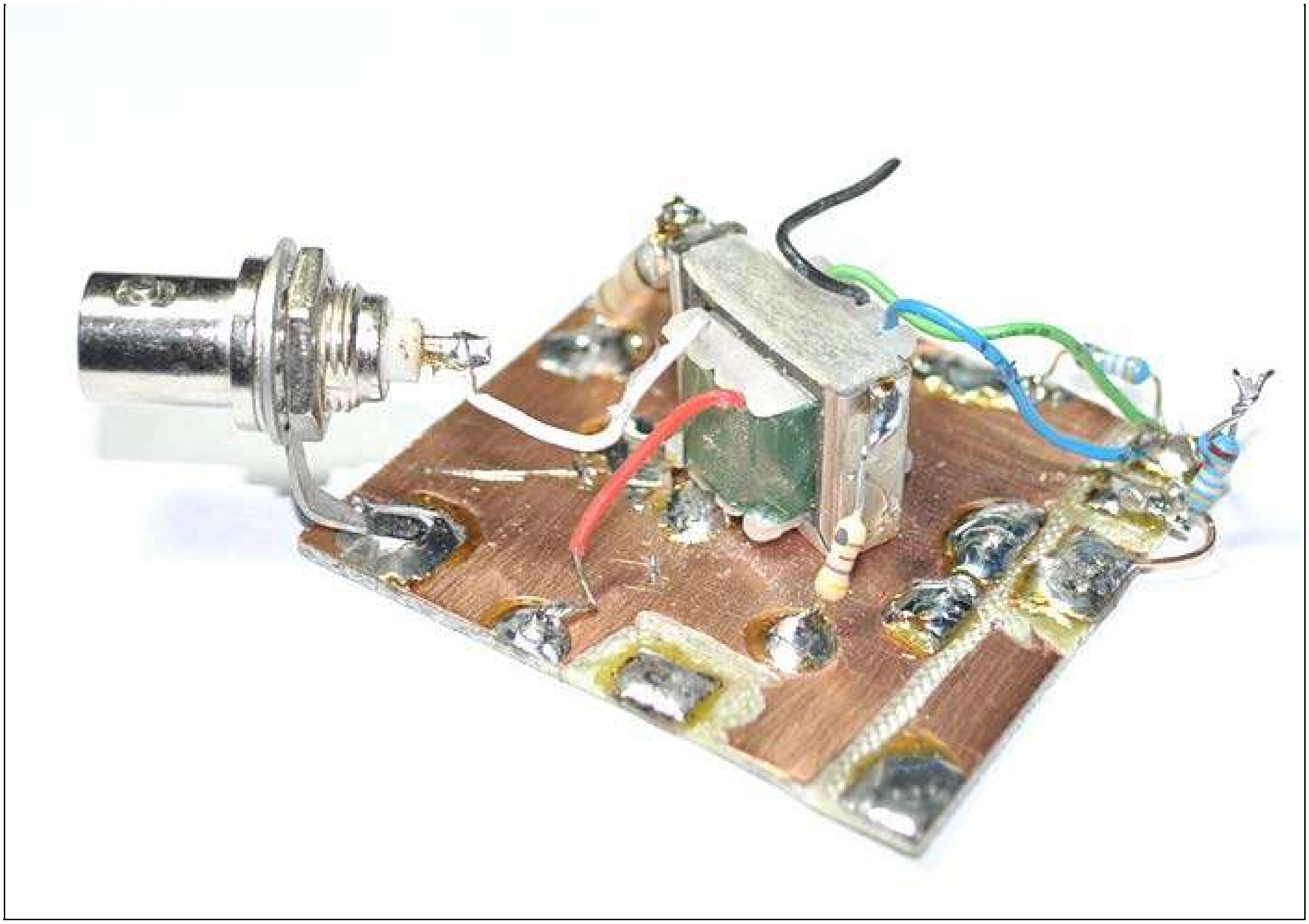

Above — The very first AF return loss bridge built. Anchored to the ground plane with resistors, the transformer was a 600 ohm *junk box* special. Although I was able to achieve a deep NULL using a 49.9 ohm resistor, the return loss was 86 dB; not possible. Additionally, other fixed resistors gave return loss values more than 4-5 dB away from the proper value. Likely, my junk box transformer lacked sufficient inductance for 1 KHz. For testing your bridge, use a formula to inform you of what RL value to expect for a given fixed resistor.

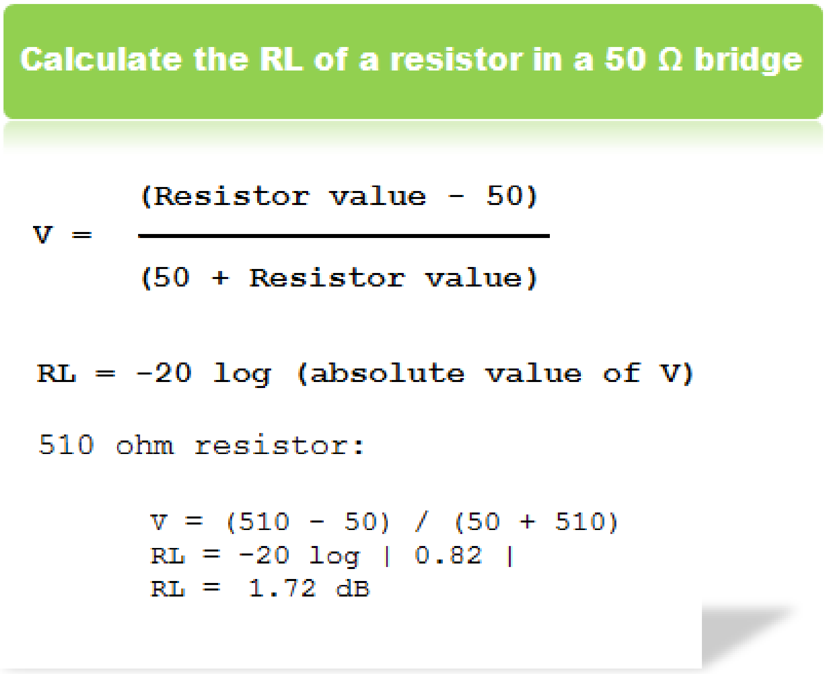

Calculate the RL of a resistor in a 50 Ω bridge

$$V = \frac{(\text{Resistor value} - 50)}{(50 + \text{Resistor value})}$$

RL = -20 log (absolute value of V)

510 ohm resistor:

V = (510 - 50) / (50 + 510)
RL = -20 log | 0.82 |
RL = 1.72 dB

Above — The formula to calculate the expected return loss for a fixed resistor placed in the Unknown Impedance port on a Wheatstone bridge. Click for a table of Return Loss values for some non 50 ohm resistors. Your RL values, will rarely be exact, but should be close to the predicted value. A well functioning bridge should yield a return loss of > 40 dB using a 51 to 47 ohm resistor as the Unknown Impedance.

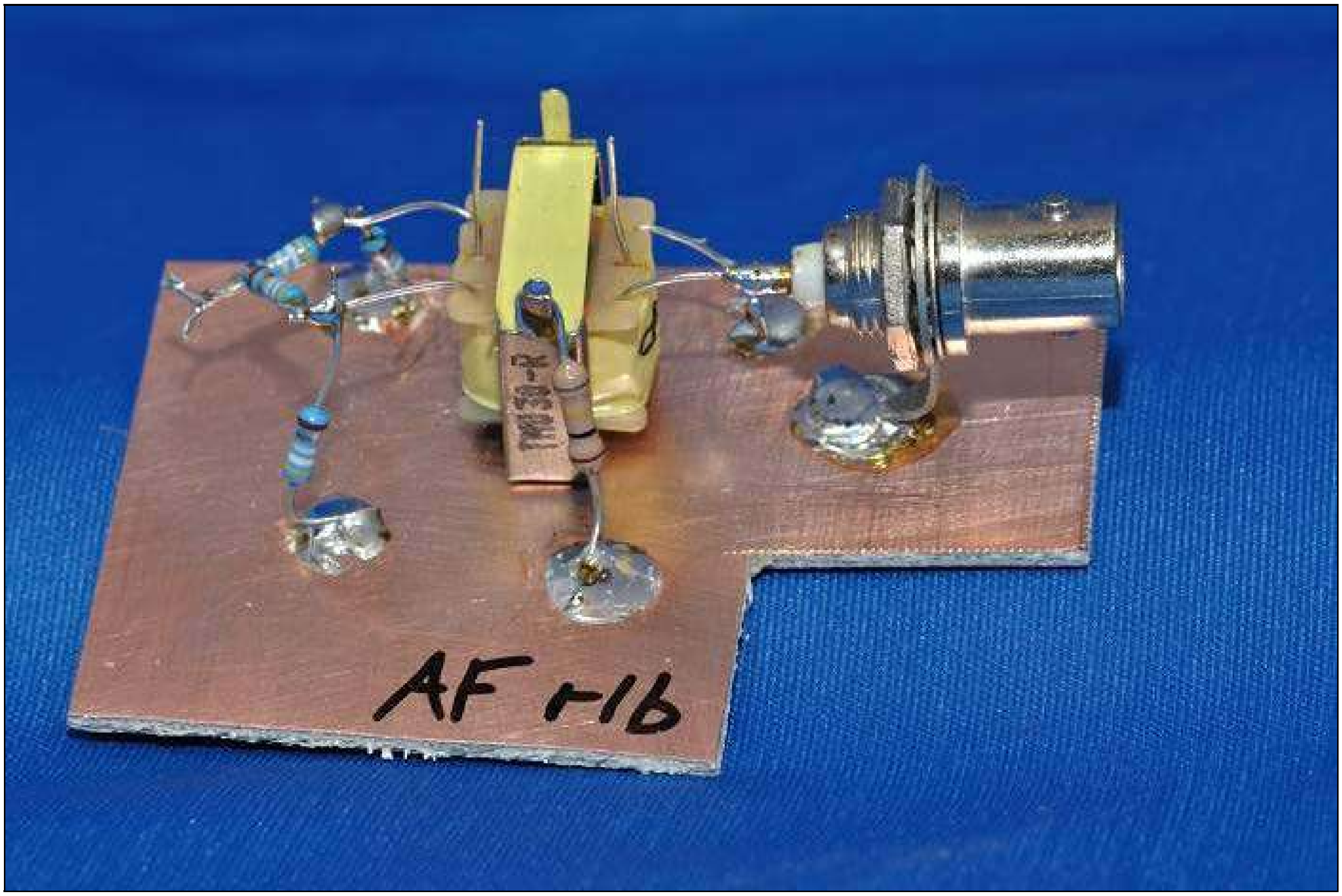

Above — A second bridge was built after obtaining a 100 Ω : 100 Ω AF transformer from Mouser Electronics. This transformer was ideal (each coil has ~ 1H in inductance!). Bench testing indicated good function. My results are tabled below:

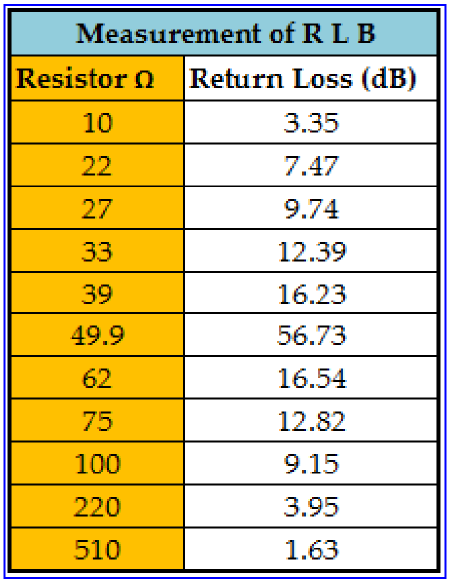

Measurement of R L B	
Resistor Ω	**Return Loss (dB)**
10	3.35
22	7.47
27	9.74
33	12.39
39	16.23
49.9	56.73
62	16.54
75	12.82
100	9.15
220	3.95
510	1.63

Above — A table of the above 50 ohm Wheatstone bridge return loss measurements. These results are acceptable. The NULL with a 49.9 ohm resistor was incredibly sharp and garnered a RL of 56.73 dB. My AF source was a low noise 1 KHz, 50 ohm output impedance signal generator.

If you do not need return loss, and only require a NULL to indicate a match, a common 600 ohm transformer may work okay for you.

Building an active instrument to measure the return loss of a 50 ohm input Z audio preamplifier.

The results of my early experiments with a passive bridge were encouraging. Noting that most builders would have difficulty obtaining a 100 Ω : 100 Ω AF transformer, a version using op-amps was sought. My first 3 designs did not work properly and I became discouraged. Some guidance from Wes, W7ZOI allowed me to problem solve and experience success.

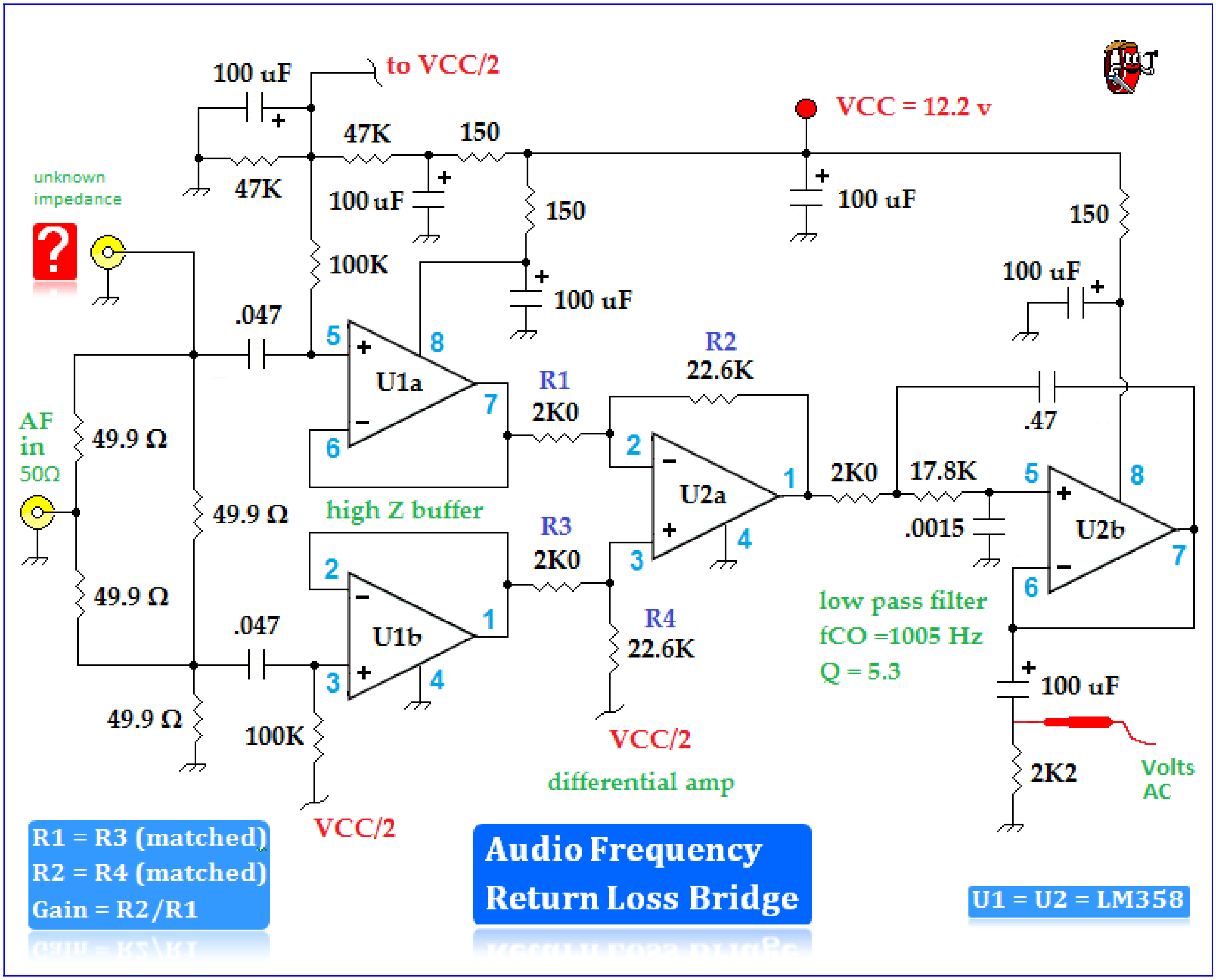
100 uF
to VCC/2
VCC = 12.2 v
47K
150
47K
100 uF
150
100 uF
unknown impedance
150
100K
100 uF
100 uF
.047
5
8
U1a
R2
22.6K
7
R1
2K0
6
AF in 50Ω
49.9 Ω
49.9 Ω
2
U2a
1
.47
2K0
17.8K
5
8
+
high Z buffer
R3
2K0
2
3
4
.0015
U2b
7
6
49.9 Ω
.047
1
U1b
3
4
R4
22.6K
low pass filter
fCO =1005 Hz
Q = 5.3
100 uF
49.9 Ω
100K
VCC/2
Volts AC
2K2
differential amp
VCC/2
R1 = R3 (matched)
R2 = R4 (matched)
Gain = R2/R1
Audio Frequency Return Loss Bridge
U1 = U2 = LM358

Above — Schematic of my active Wheatstone bridge, amplifier and low-pass filter for measuring the Return Loss of 50 ohm input impedance AF amplifiers. I built 3 copies of the above device; best results occurred when careful layout and planning were employed. Optimal performance occurred when encased in a metal box.

The bridge was built from 1% metal film resistors. 0.047 polyester film capacitors lightly couple the bridge to high impedance op-amp buffers labeled U1a + U1b. My experiments informed me that to minimize loading on the bridge is important. The LM358 is an excellent op-amp choice, but almost any other op-amp could be employed successfully. U2a is the differential amplifier and matching R1 + R3 and R2 + R4 with 1% tolerance resistors is critical; 5% resistors did not work well. The gain is non-critical — feel free to choose reasonable resistor ratios based upon the resistors you have in stock. The differential amp promotes the unfortunate side effect of amplifying both the desired AF source plus any common mode signals. Although common mode suppression is an important consideration when designing instrumentation amps, fortunately, performance is fine. A amplifier topology using a differential amp across the bridge was trialed, but functioned identically to the simpler differential amp shown. Consult a textbook for more information on Instrumentation amps. Much information was gleaned from Professor Ken Kuhn's web site.

The output is low-pass filtered by a single stage Sallen-Key low-pass filter with a peak frequency of 1 kHz and a Q of 5. This filter gain at 5 at 1 kHz is 0.328 at 2 kHz and 0.123 at 3 kHz. Thus, the second harmonic is reduced by a factor of (5/0.328) = 15.2 and the third harmonic is reduced by a factor of 40.6. **Do not omit a low pass filter**. I chose a 1 KHz cutoff, but experimentation indicated a low-pass cut-off frequency as high as 10 KHz may work okay if you plan to use the bridge at frequencies other than 1 KHz.

Power supply decoupling proved important. When less DC low-pass filtering (less than the 150 Ω plus the 100 uF capacitors shown) was employed, some low frequency audio noise appeared in the output.

I measured using a X1 oscilloscope probe on the output of U2b.

AF
RLB
JULY
2010

Above — A breadboard of the active Wheatstone bridge schematic located above. When tested with fixed value resistors, the RL @ 49.9 ohms was 55.4 dB and close to predicted value with other test resistors. This instrument will be put in a metal case and become a permanent part of my test equipment arsenal.

Part 3: Measuring Unknown Impedances

Building an instrument to measure the input impedance of an audio preamplifier using a NULL.

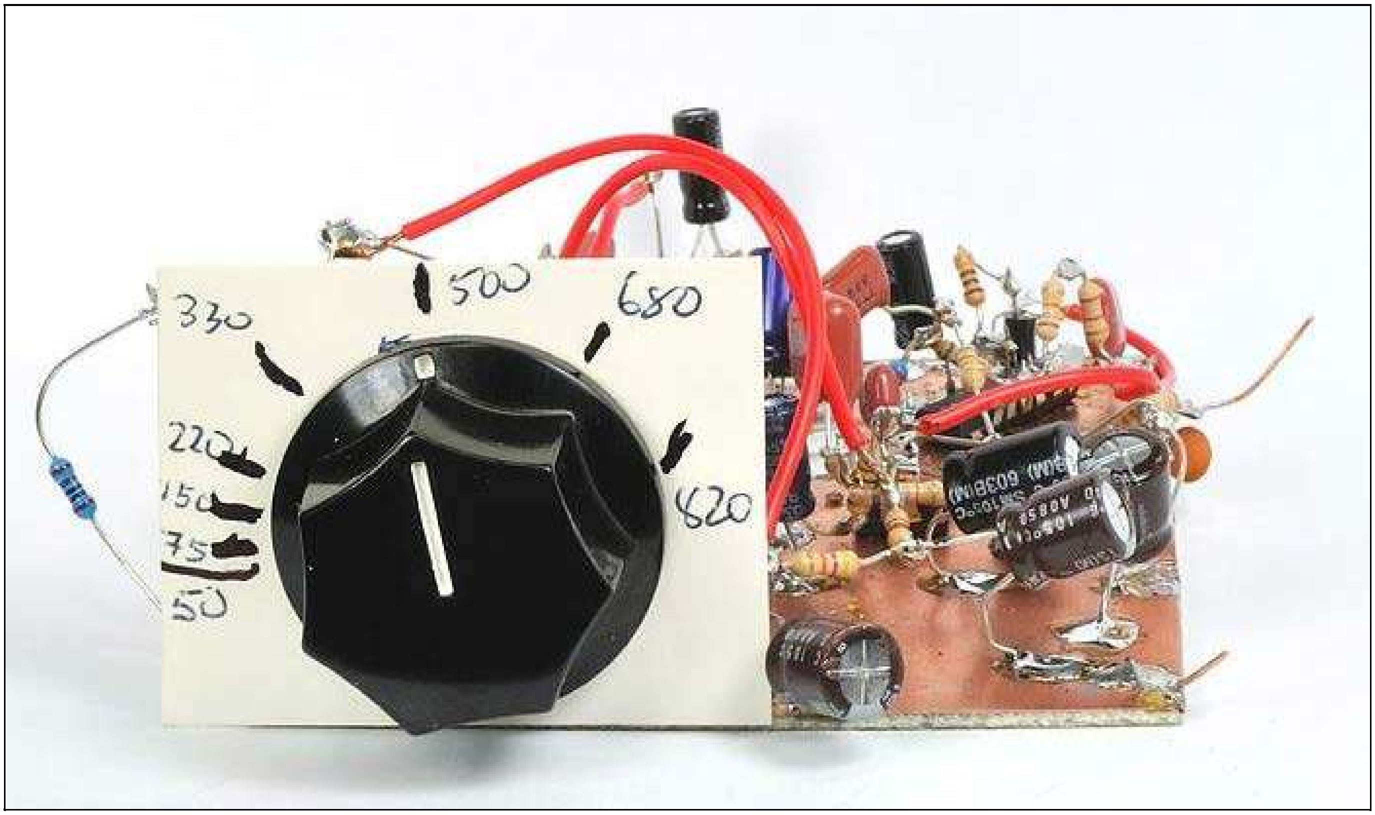

Above — One of several Wheatstone bridge circuits built in the Spring-Summer of 2010. In these bridges, the potentiometer was calibrated and the panel labeled using fixed resistances for calibration. One big challenge is range or resolution; dependent on the bridge resistor values and what impedance you are trying to measure. Greatest accuracy is associated with 5 or 10 turn potentiometers, but these are expensive. Often, I used standard, linear taper pots to save money during my experiments. Over 7 different bridges were built and tested. To save time, I didn't photograph many of my projects from the summer.

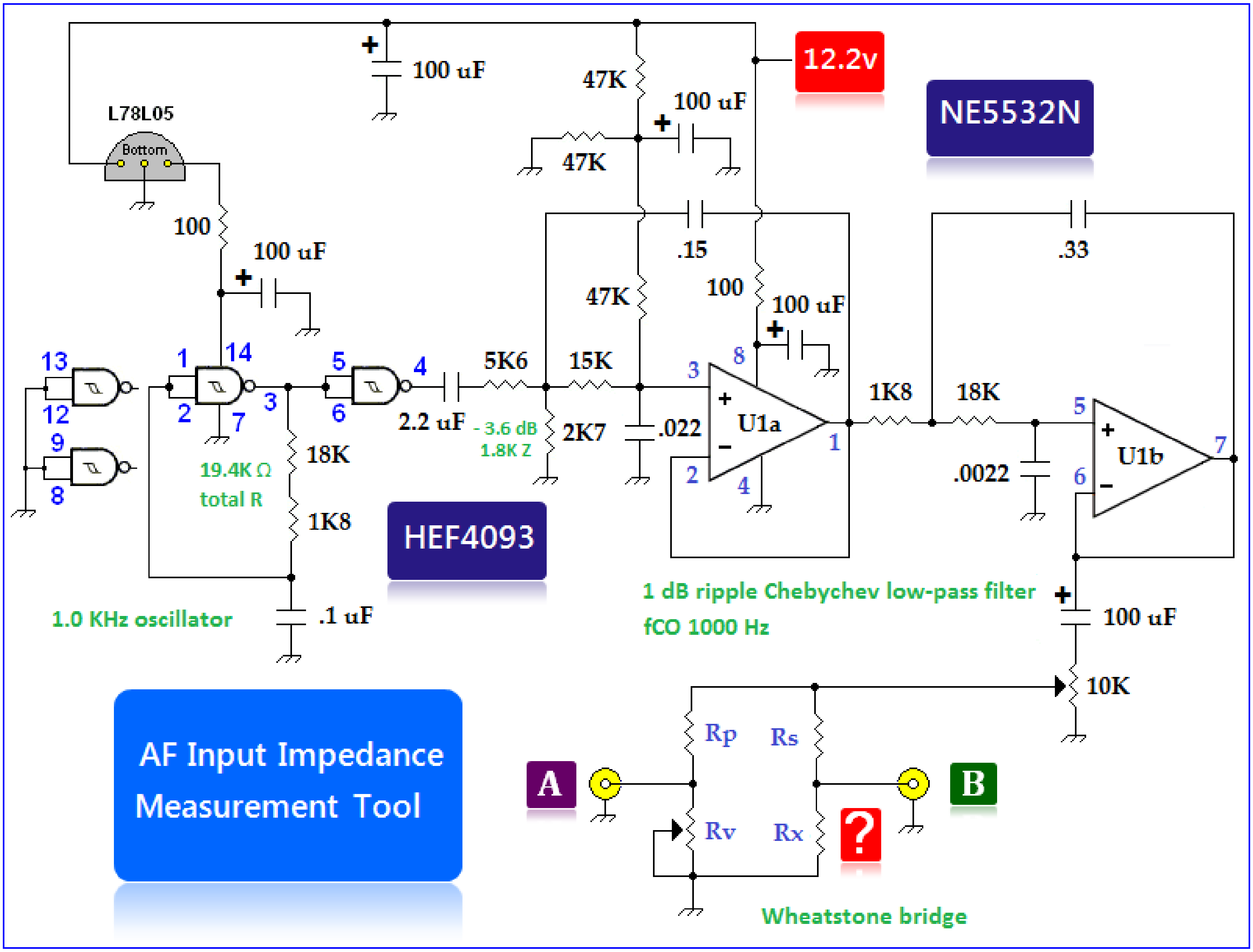
L78L05
Bottom
+
100 uF
47K
100 uF
47K
12.2v
NE5532N
100
100 uF
.15
.33
47K
100
100 uF
13
12
9
8
1
14
2
7
3
5
6
4
5K6
15K
2.2 uF
- 3.6 dB
1.8K Z
2K7
.022
3
8
+
U1a
-
2
4
1
1K8
18K
5
+
U1b
-
6
7
.0022
19.4K Ω
total R
18K
1K8
HEF4093
1.0 KHz oscillator
.1 uF
1 dB ripple Chebychev low-pass filter
fCO 1000 Hz
100 uF
10K
AF Input Impedance
Measurement Tool
Rp
Rs
A
B
Rv
Rx
?
Wheatstone bridge

Above — A complete 1 KHz signal generator, low-pass filter and bridge circuit which became the prototype for most of my experiments in this section. Click for a high resolution photo of 1 of the breadboards during construction.

The 1 KHz signal generator is a digital oscillator built with 2 gates from a 4093. This excellent oscillator uses a single R + C network for tuning and requires a voltage regulator for frequency stability. The output signal is attenuated 3.6 dB and low-pass filtered by 4 poles of active filtering. A 10K pot controls the drive into the bridge circuit. The bridge outputs are labeled **A** and **B** and require buffering, amplification and low-pass filtering similar to the active bridge shown earlier. These functions and some comments on the bridge resistors come later.

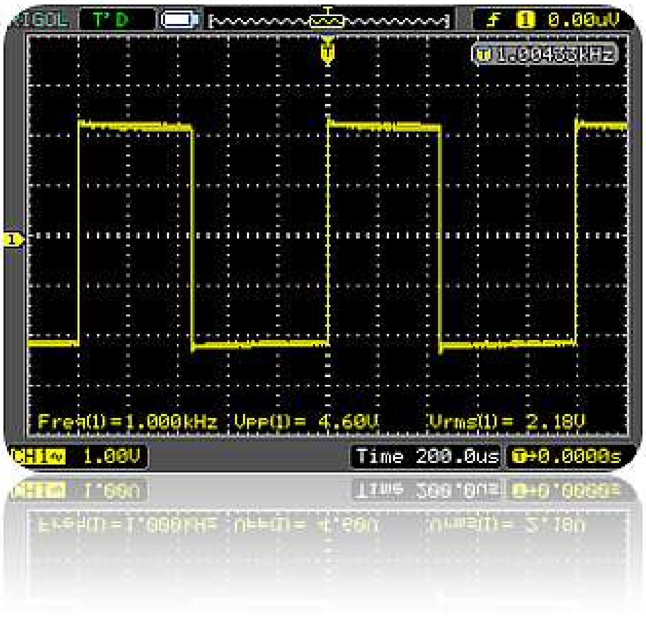
RIGOL
T'D
0.00uV
1.00433kHz
Freq(1)=1.000kHz
Vpp(1)= 4.60V
Vrms(1)= 2.18V
CH1 1.00V
Time 200.0us
0.0000s

Above — The oscilloscope waveform from the digital 1 KHz oscillator. Digital clocks fascinate me and this was an untried design. Initially a CMOS 555 timer was considered, however, I own many 4000 series NAND Schmitt triggers and pressed 1 into service. Another good choice might be the 74HC132. The first NAND gate (inverter) contributes 180 degrees of phase shift, while the RC low-pass filter tank circuit digitally shifts the AC the other 180 degrees. Output noise is filtered by both the low-pass filter and the input Schmitt trigger dead band or hysteresis. The result is a fairly crisp square wave that may rival the 555.

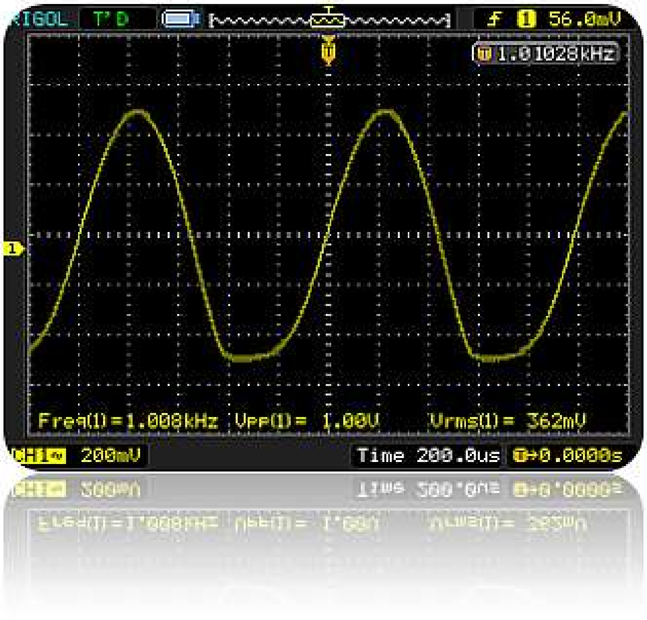
RIGOL
T'D
56.0mV
1.01028kHz
Freq(1)=1.008kHz
Vpp(1)= 1.00V
Vrms(1)= 362mV
CH1 200mV
Time 200.0us
0.0000s

Above — The output of the 1 KHz low-pass filter. A sine wave is desirable, but not critical; suppression of energy in the range of 5-10 KHz, informed the filter design goal. At 8000 Hz, the attenuation is > 80 dB. All good.

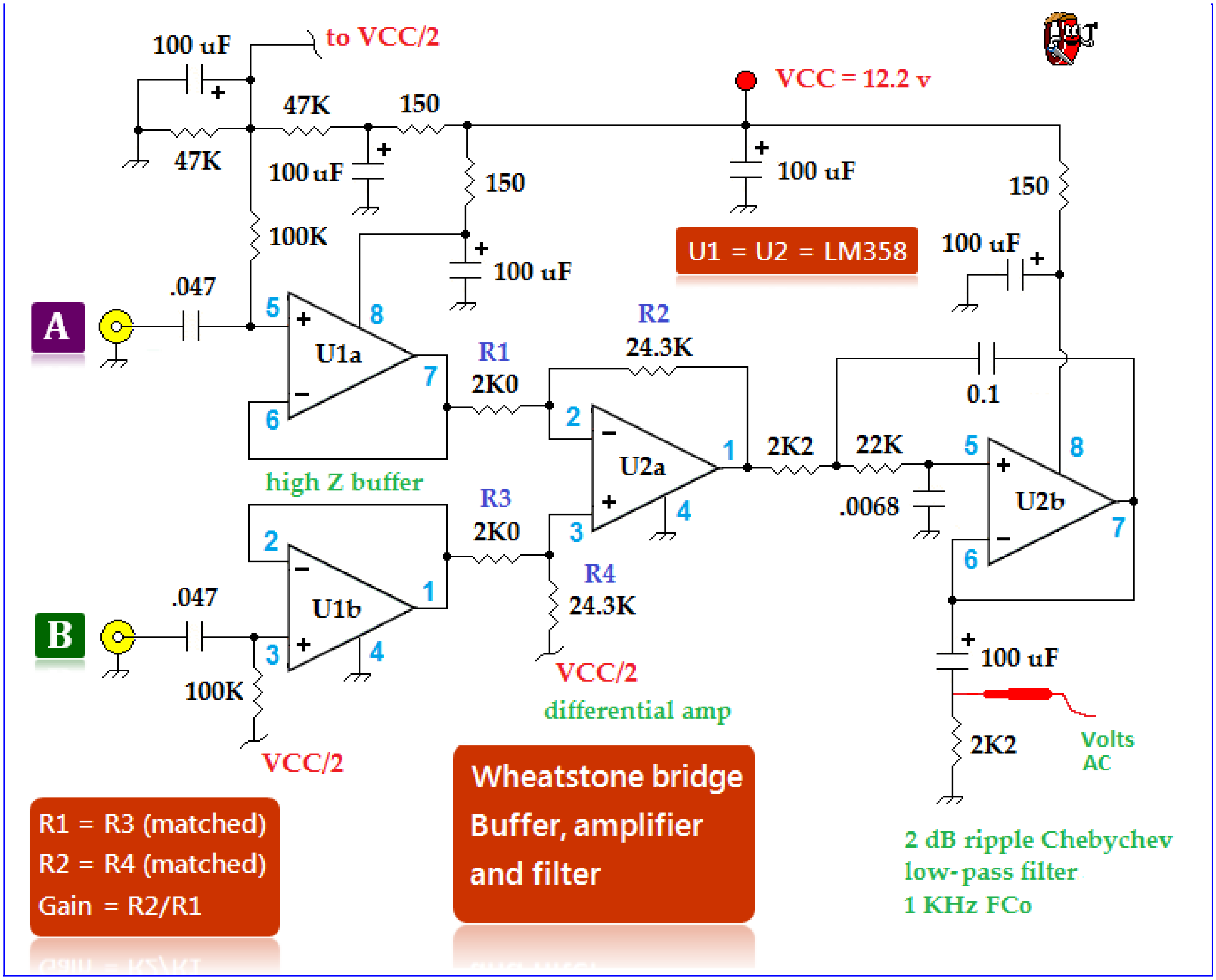
100 uF
to VCC/2
VCC = 12.2 v
47K
150
47K
100 uF
150
100 uF
150
100K
100 uF
U1 = U2 = LM358
100 uF
.047
A
5
8
U1a
7
6
R1
2K0
R2
24.3K
0.1
2
U2a
1
2K2
22K
5
8
U2b
7
6
.0068
4
3
high Z buffer
R3
2K0
2
U1b
1
.047
B
3
4
R4
24.3K
100K
VCC/2
differential amp
100 uF
2K2
Volts AC
VCC/2
R1 = R3 (matched)
R2 = R4 (matched)
Gain = R2/R1
Wheatstone bridge
Buffer, amplifier
and filter
2 dB ripple Chebychev
low-pass filter
1 KHz FCo

Above — The buffer, differential amplifier and low-pass filter employed during this series of experiments. Function is identical to the similar stage described earlier.

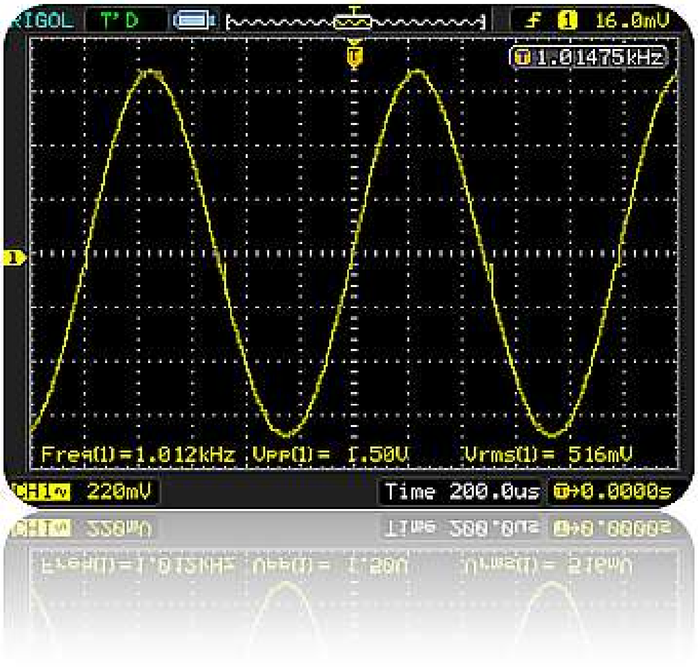
RIGOL T'D
16.0mV
1.01475kHz
Freq(1)=1.012kHz Vpp(1)= 1.50V Vrms(1)= 516mV
CH1 220mV
Time 200.0us
0.0000s

Above — The output of amplified and filtered Wheatstone bridge at open circuit (no resistance at the Unknown Impedance port).

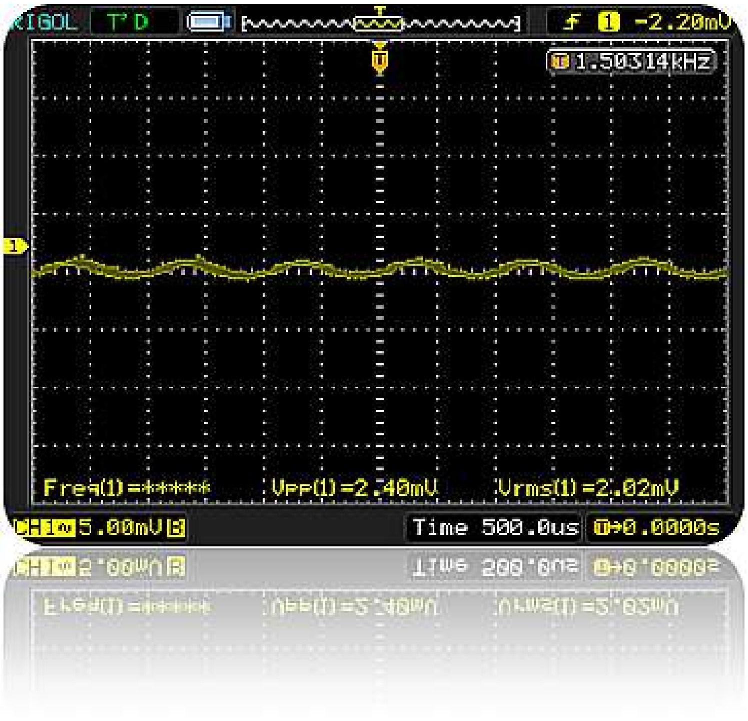
RIGOL T'D
-2.20mV
1.50314kHz
Freq(1)=*****
Vpp(1)=2.40mV
Vrms(1)=2.02mV
CH1 5.00mV B
Time 500.0us
0.0000s

Above — the output of the amplified and filtered Wheatstone bridge at NULL (potentiometer setting balanced to match the resistance at the Unknown Impedance port. Below 2 mV, accuracy is lost.

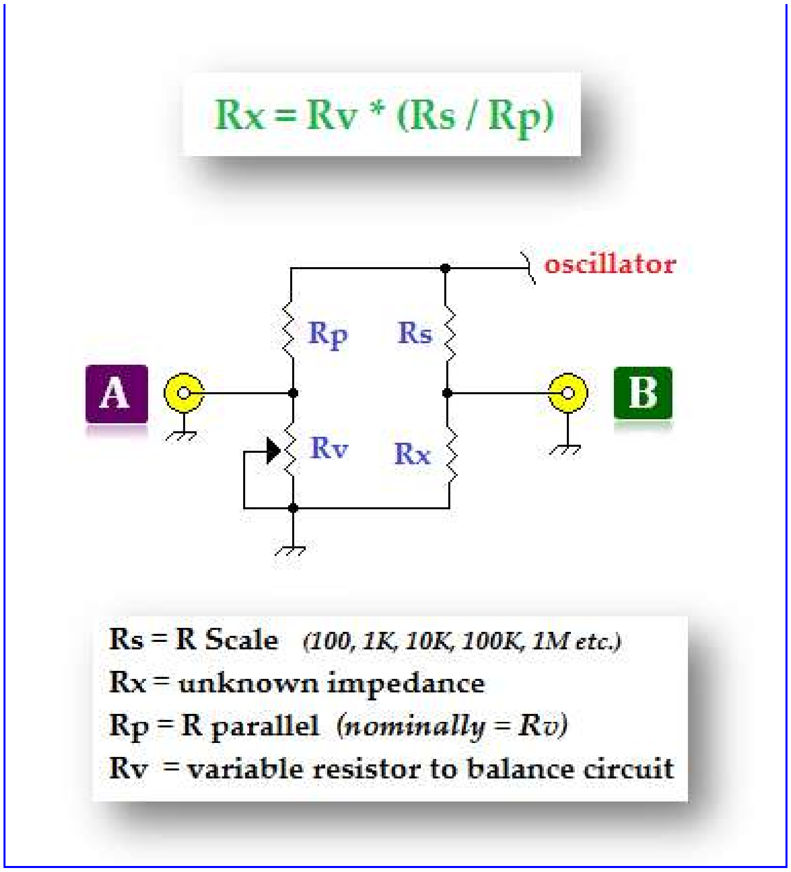

$$Rx = Rv * (Rs / Rp)$$

Rs = R Scale *(100, 1K, 10K, 100K, 1M etc.)*
Rx = unknown impedance
Rp = R parallel *(nominally = Rv)*
Rv = variable resistor to balance circuit

Above — The math behind the bridge. I found when R Variable (Rv) was the same resistance as the fixed resistor R parallel (Rp), reasonable resolution was possible. "Reasonable resolution" means your pot has a good range of rotation as you go from the lowest to highest measureable impedance. Generally, Rv has to be less than the maximal impedance you are trying to measure.

R Scale (Rs) can be switched in decades via a panel mount switch to cover a wide range of resistance with good resolution, or just be 1 or 2 values. It's your design call.

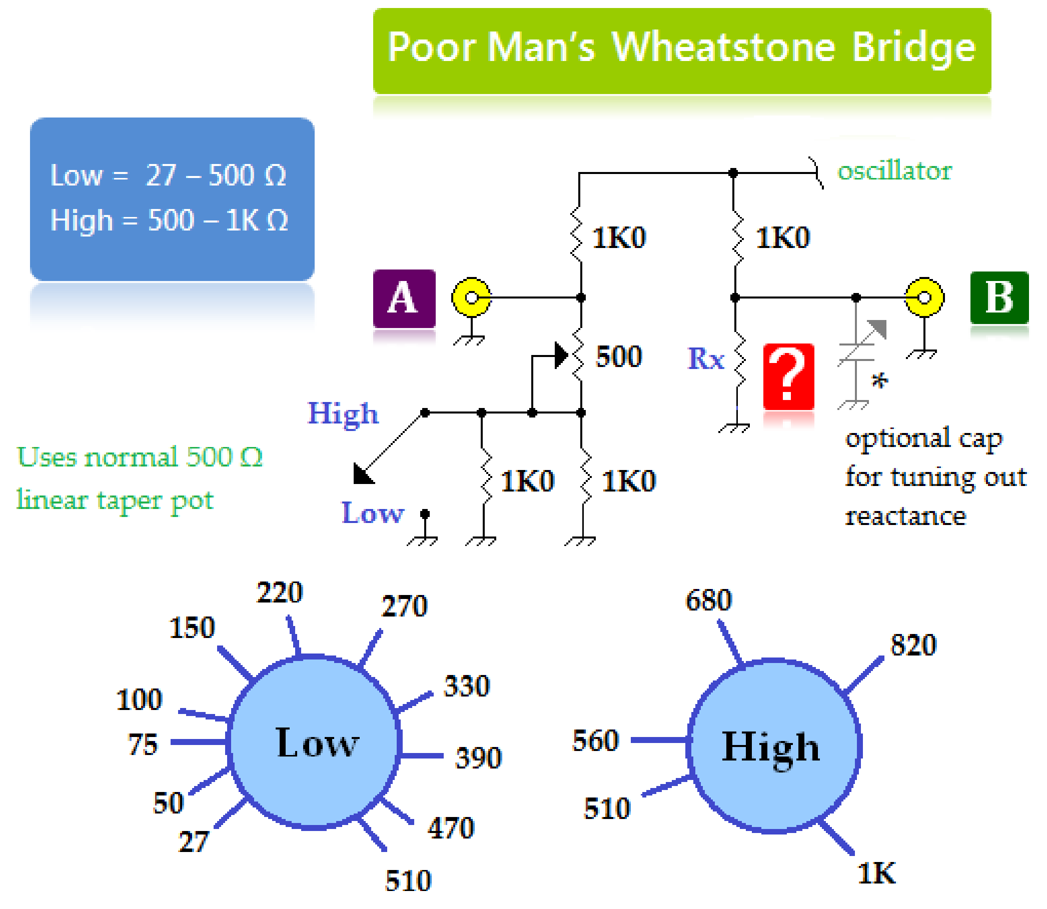
Poor Man's Wheatstone Bridge
Low = 27 – 500 Ω
High = 500 – 1K Ω
oscillator
1K0
1K0
A
B
500
Rx
?
*
High
Low
1K0
1K0
Uses normal 500 Ω
linear taper pot
optional cap
for tuning out
reactance
220
150
270
330
100
75
390
50
27
470
510
Low
680
820
560
510
High
1K

Potentiometer settings in my experiment calibrated with resistors

Above — My poor man's impedance measurement device that uses a common 500 ohm linear taper pot as the balancing resistor. In order to get good pot resolution, the desired range is switched. This bridge measures impedances at the Unknown Port from about 27 ohms up to 1K with decent resolution. The blue circles depict how I calibrated the front panel of my device using 2 colors. This device had an average return loss of 32.5 dB when a NULL was obtained.

Measuring resistors to calibrate a bridge is quite different from real-world measurement of reactive AF amplifier loads — if the unknown resistance has a large inductive or capacitive reactance, obtaining good bridge balance might prove difficult. Your bridge can *only* null the in-phase signal. An extension to the standard bridge involves adding a series or shunt capacitance (depending on the phase of the reactance) to the A or B port. This may allow you to null the reactive part and also provide the reactive impedance value as well. An outstanding reference may be located with your favorite search engine: Look for the manual for the **General Radio GR1650 Impedance Bridge.** I found a copy and the download was very slow, but worth it. This manual may be the greatest reference every published on the Wheatstone bridge and comprehensively covers tuning out the reactance of complex impedances amid a myriad of other topics.

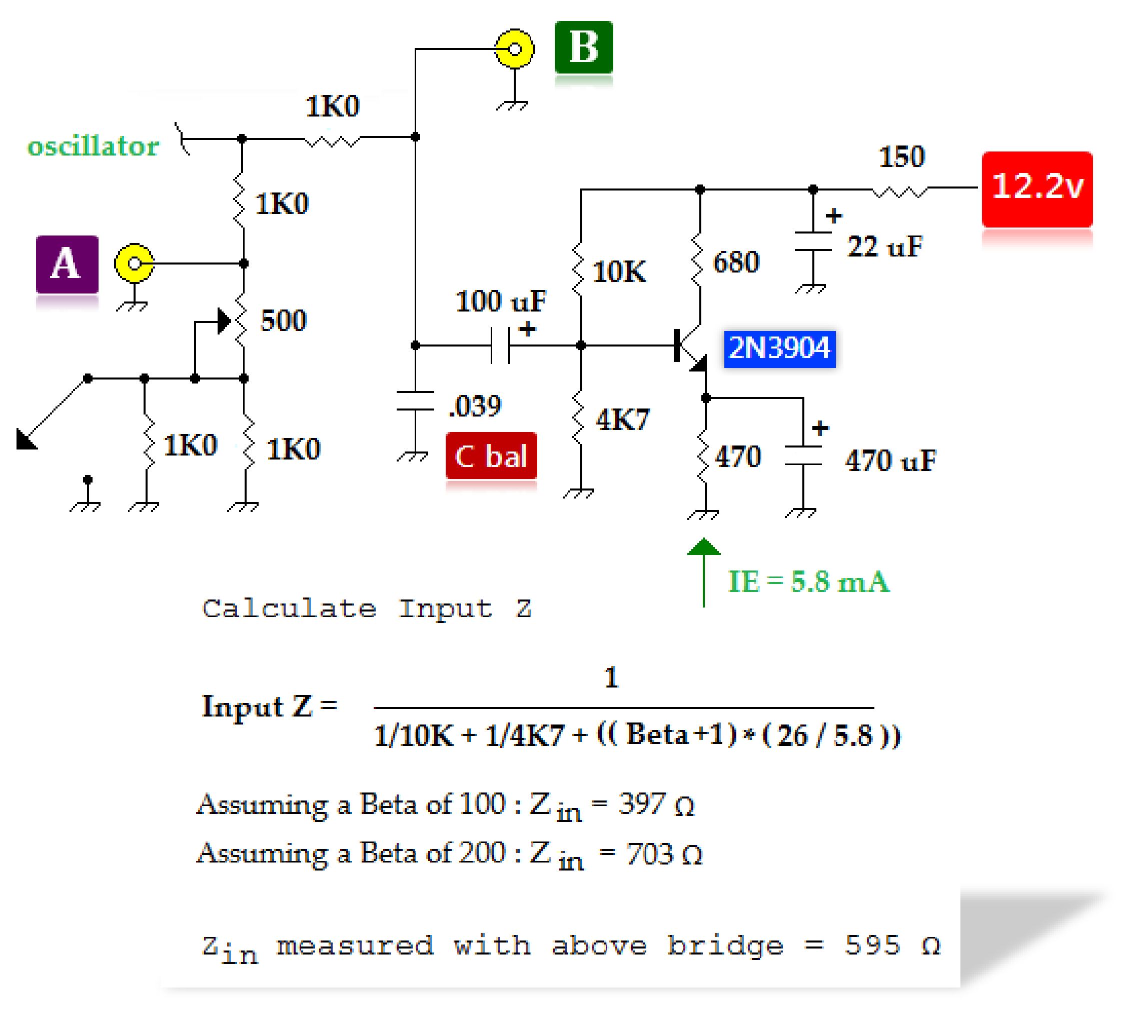

Calculate Input Z

$$\text{Input } Z = \frac{1}{1/10K + 1/4K7 + ((\text{Beta}+1) * (26 / 5.8))}$$

Assuming a Beta of 100 : Z_{in} = 397 Ω

Assuming a Beta of 200 : Z_{in} = 703 Ω

Z_{in} measured with above bridge = 595 Ω

Above — The poor man's bridge measurement of a test AF amplifier on my bench. The reactive component of the amplifier input impedance was minimized using a 0.039 uF capacitor found experimentally. Of particular interest, is the difference between the *calculated* input Z and the *actual* input Z measured with the bridge.

The Beta of the 2N3904's in my collection ranges from about 100 to 225. Calculations with 2 different Beta values are shown (RE is *well* bypassed with a 470 uF capacitor, so, re = (Beta * 26/5.8 mA) . The measured input Z was 595 ohms. I confirmed this by removing the 0.039uF tuning capacitor, plus connecting a fixed 595 ohm resistance to the Unknown Impedance port. I then turned the potentiometer fully clockwise and adjusted it for a NULL. When the bridge was nulled, the potentiometer knob pointed at the same mark as when the amplifier was connected to this Unknown Impedance port.

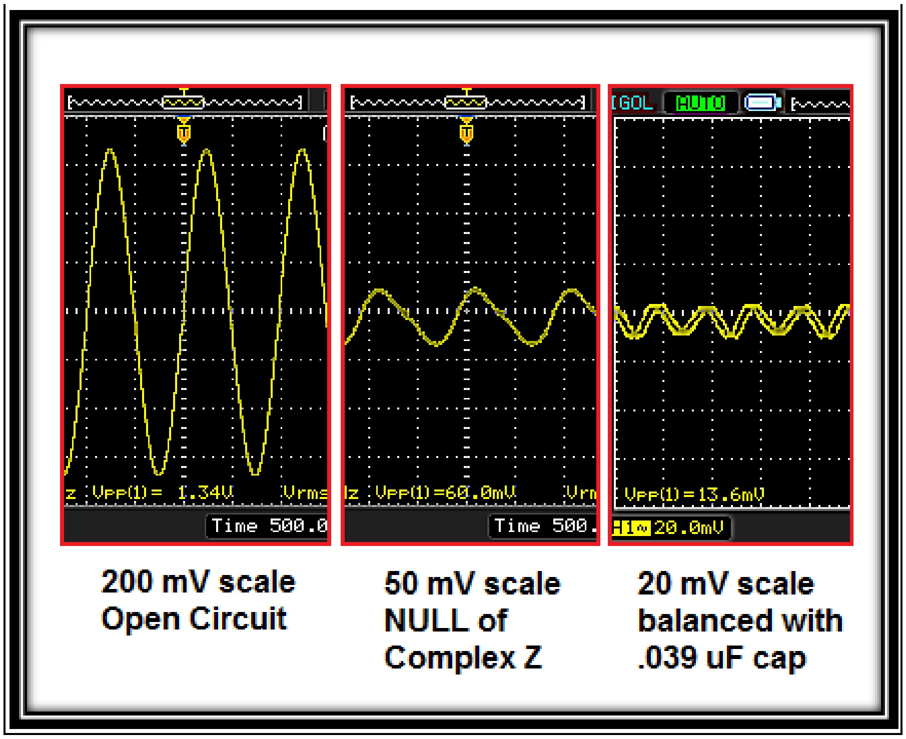
Vpp(1)= 1.34V
Vrms
Time 500.0
Vpp(1)=60.0mV
Vrm
Time 500.
IGOL
AUTO
Vpp(1)=13.6mV
20.0mV
200 mV scale
Open Circuit
50 mV scale
NULL of
Complex Z
20 mV scale
balanced with
.039 uF cap

Above — The oscilloscope output waveform of the amplifier circuit shown above: open circuit, with the potentiometer balanced as well as possible and finally, the potentiometer balanced with the addition of a 0.039 uF capacitor attached to Port B. Although, I was able to get a NULL without the capactitor by just tweaking the potentiometer, slightly better precision was obtained after adding the 0.039 uF balancing capacitor.

Part 4: Miscellaneous Circuits, Scans and Photographs

1
2
3
CW
VISHAY
SPECTROL
PRECISION
POTENTIOMETER
1009 MEX
MOD 534

Above — More accurate results will be obtained with a calibrated 10 turn potentiometer to balance your bridge. A local store sold me this precision 10-turn, 10K pot for about 11 dollars (still expensive for me). They normally sell for twice this price in Canada.

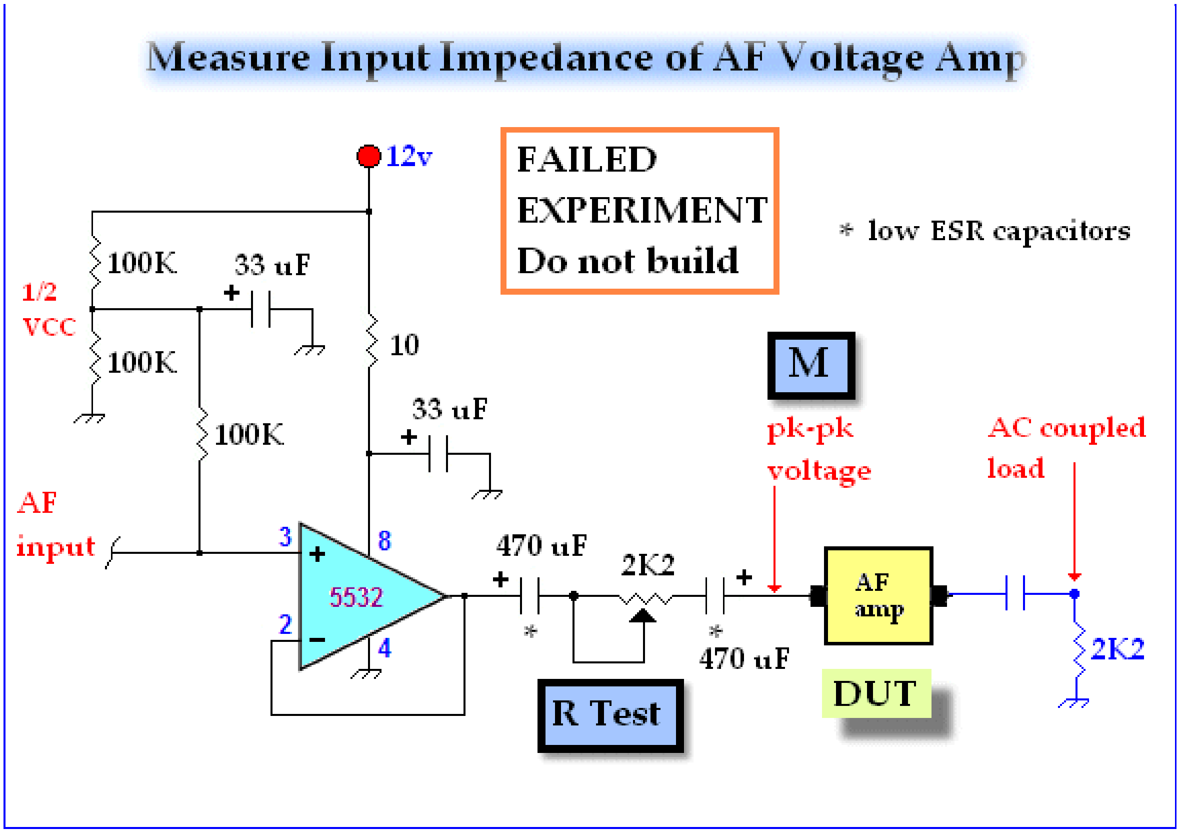

Above — My first input impedance measurement device that didn't work. It turns out my experiments were performed *incorrectly*, however, I'm glad because this failure spurred me to investigate bridge networks. The series resistance method *is* worth understanding and happily, Jeff, AD6MX described the correct procedure in a private email received December 2010. I quote him below:

"The series resistance method for input impedance should start with the variable resistor disconnected from the node to be measured. The open circuit voltage at the end of that resistor is measured (the resistor value doesn't affect the open circuit voltage since there's no current into an open circuit.)

Next the free end of the variable resistor is then connected to the input node and the resistor is finally adjusted for half the open circuit voltage at the same end of the resistor, at the input node being measured. What happens is the variable resistor and the node input impedance form a voltage divider, with equal arms or branches.

The value of the resistor when measured out of the circuit is the same as the input impedance at the measured input node. This scheme has some assumptions: the driving amplifier has negligible output impedance compared to the measured impedance, and the input impedance is purely resistive, with no reactance or V-I phase shift.

The phase shift condition may be checked by taking these 3 voltage measurements: across each branch of the divider separately, and also the driving source voltage (across both branches in series.) The sum of the separate branch voltages should match the source voltage when there is little phase shift.

This 3 voltage scheme is used in some antenna analyzers in order to measure phase shift. For checking for the resistive condition, it's not important the 3 voltage method has a sign ambiguity which needs an additional step to resolve. Your description seemed to suggest starting with zero series resistance, but you see that is not the same as the procedure above. The applied voltage needs to be small enough that the amplifier remains operating in its linear range during the measurement".

Thanks for this info Jeff!

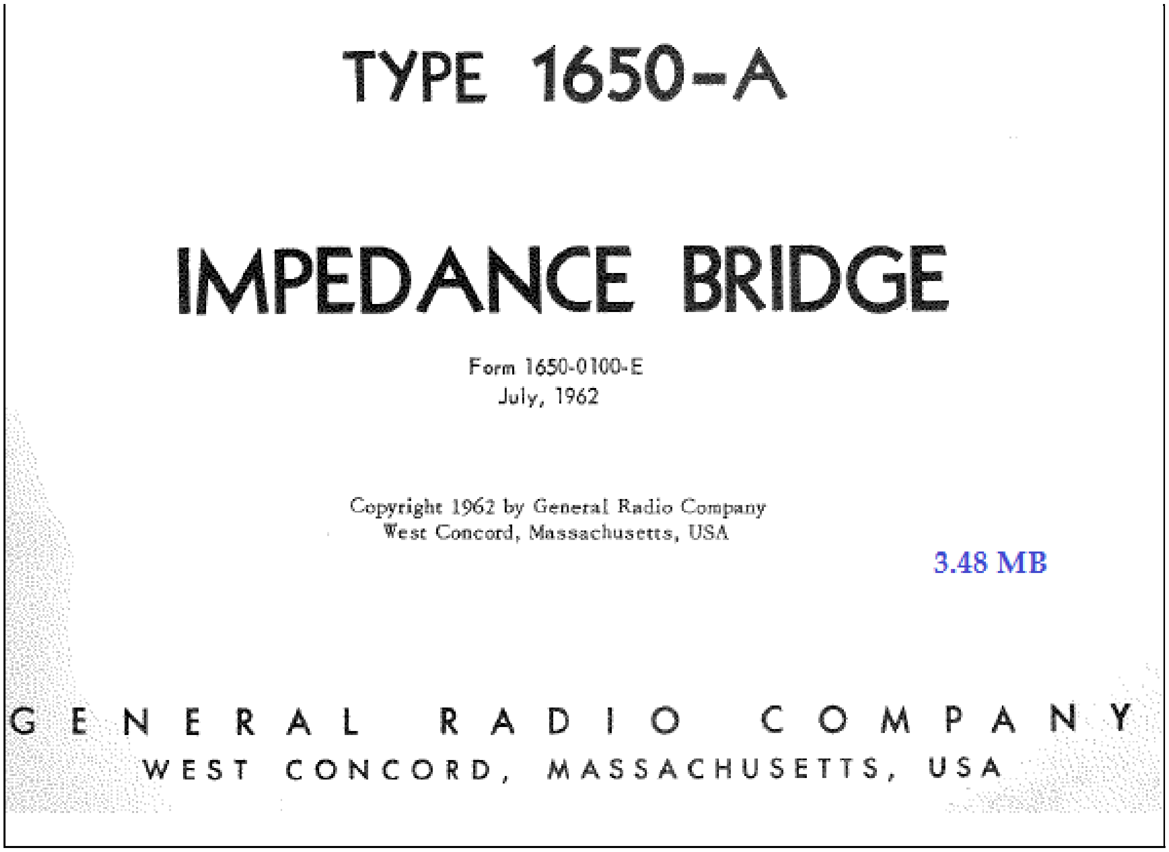

TYPE 1650-A

IMPEDANCE BRIDGE

Form 1650-0100-E
July, 1962

Copyright 1962 by General Radio Company
West Concord, Massachusetts, USA

3.48 MB

GENERAL RADIO COMPANY
WEST CONCORD, MASSACHUSETTS, USA

Above — A modified scan of General Radio's *über awesome* manual for the Type 1650-A Impedance Bridge.

RF — Test and Measurement

RF Workbench Page 3

This web page is the third installment of a 6 part series that explores basic measurement of RF circuits.

Part 3, further examines Return Loss Bridges from a bench-practice viewpoint.

I borrow heavily from the work of Wes, W7ZOI per correspondence, direct contributions and from EMRFD.

I focus on measuring low-level, HF circuits with a return loss bridge — topics such as using the bridge for antenna matching are omitted and readily found on the web. This web page contains minimal text and just relies on simple diagrams and photographs to transfer ideas and knowledge. Information regarding wideband bridge network function may be found elsewhere on this and other web sites and in EMRFD.

More on Return Loss

Gear

Equipping for a 50 Ω measurement environment in 2010 greatly improved my design capacity. The 50 Ω terminated oscilloscope makes a sensitive and accurate detector for return loss measurement. Discussion about using a 50 Ω oscilloscope termination is on the RF Workbench 1 web page.

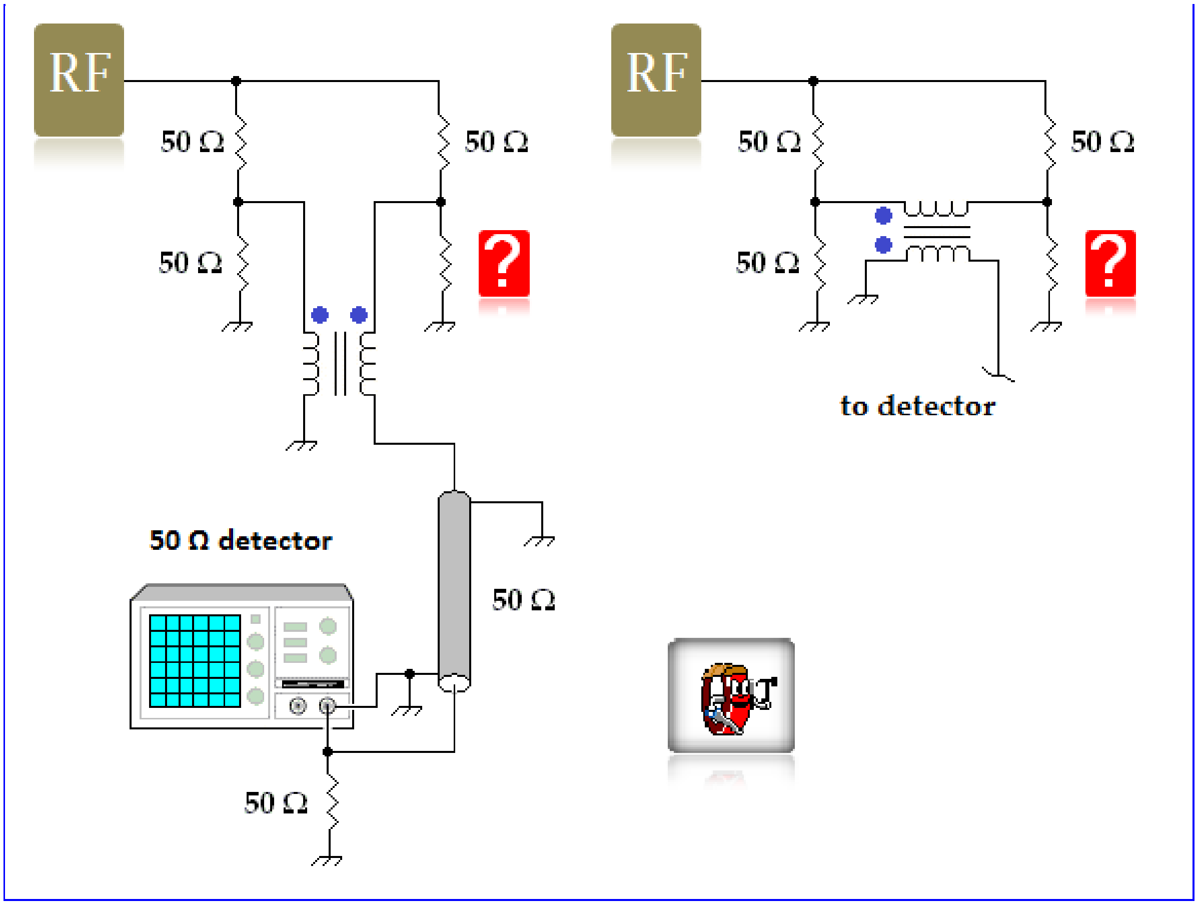
RF
50 Ω
50 Ω
50 Ω
50 Ω
?
50 Ω detector
50 Ω
50 Ω
RF
50 Ω
50 Ω
50 Ω
50 Ω
?
to detector

Above left — The RLB and measurement set up from EMRFD. Occasionally, you may see bridges using a different balun transformer wiring as shown to the right of this figure.

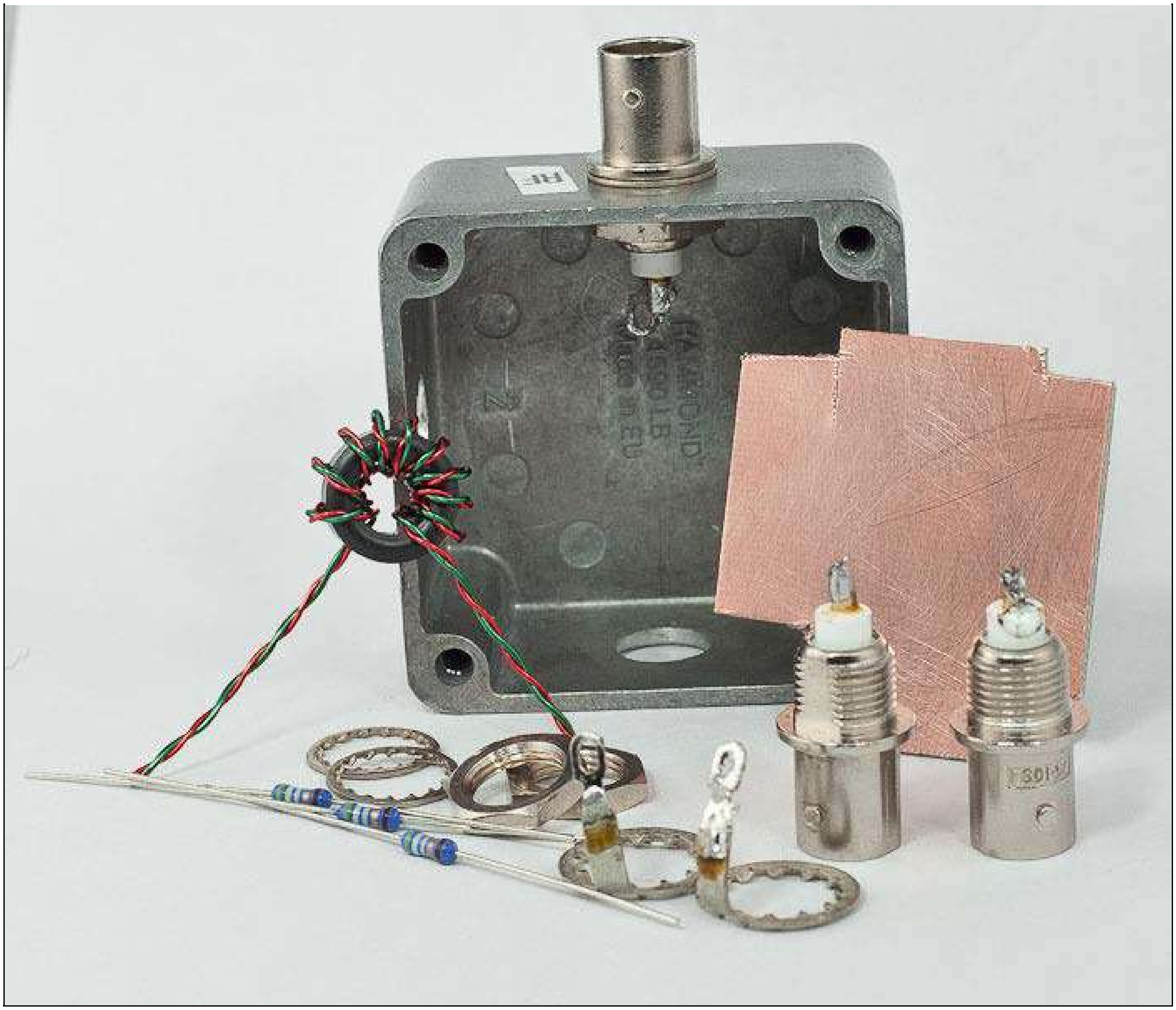

Above — All the needed parts to home build a return loss bridge. For some, the parts investment might seem substantial, but what hobby isn't expensive? If you consider the cost of commercially manufactured bridges, a homebrew solution seems a bargain. Recycled parts and a home-built chassis are inviting cost-containment techniques. See the web site of Jim, K8IQY for an example of a homebrew RLB chassis. Jim, a Manhattan style construction wizard, builds the nicest looking gear — he puts me to shame.

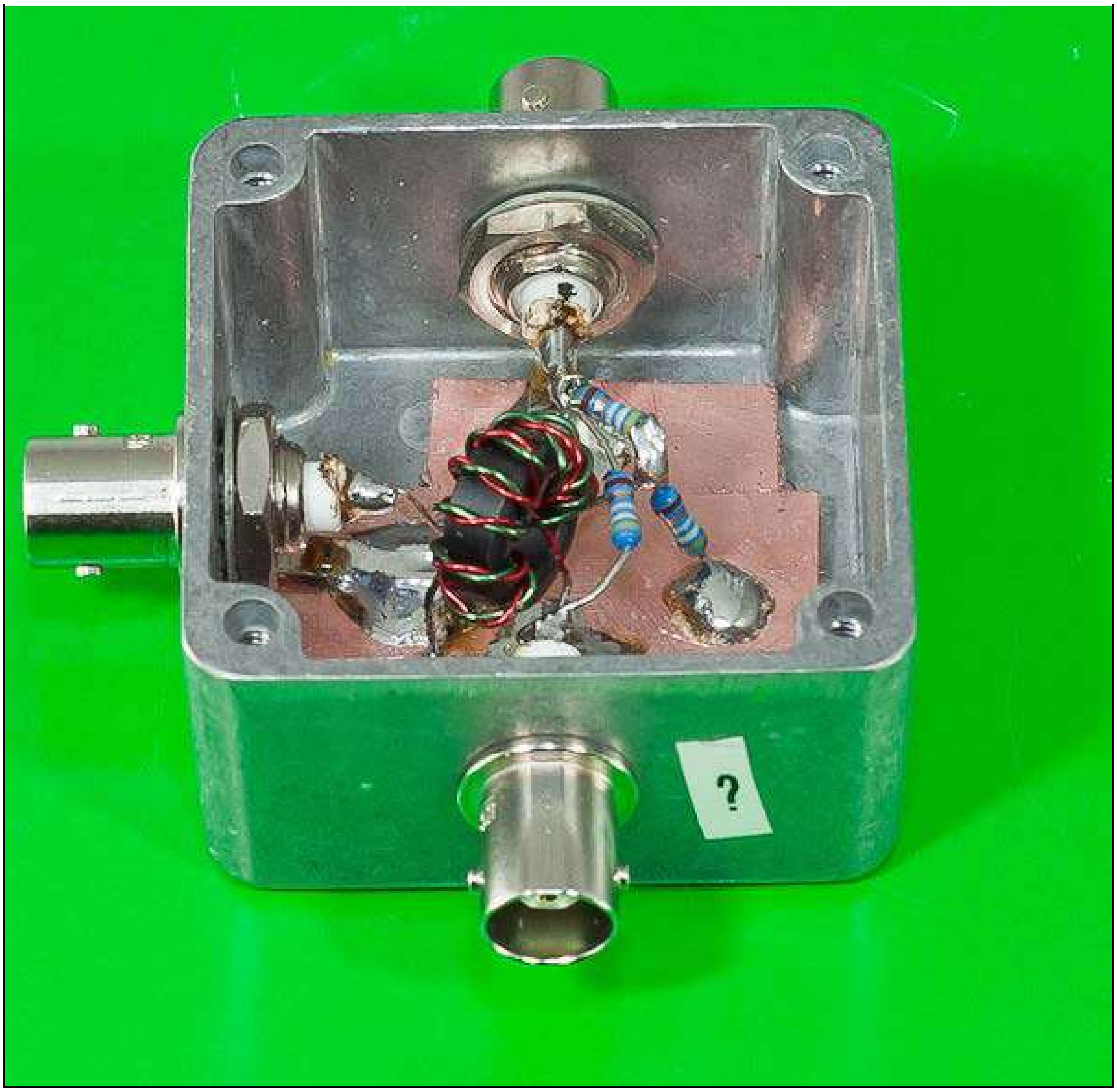

Above — A completed bridge. I used 1% tolerance 49.1 ohm resistors and an FT50-43 ferrite toroid for the bifilar wound transformer. Inductance = 38.4 uH. Many builders use the FT37-43 ferrite core. I prefer using 2 colors of enamel coated wire to avoid confusion when building stuff with transmission line style transformers and all I had in 2 colors was 24 gauge wire The bigger size ferrite toroid better accommodates the 24 gauge wire, plus photographs better.

Bridge directivity of the above RLB was 30 dB at 7 MHz, 34 dB at 14 MHz, 35.6 dB at 21 MHz, 42.1 dB at 50 MHz and 43.4 dB at 100 MHz.

If you build a circuit with a return loss close to 30 dB, it's a good day.

RETURN~LOSS BRIDGE
RF
?
DETECTOR
D

Above — The completed RF-tight bridge. Don't forget to label your network ports.

Measurement Technique

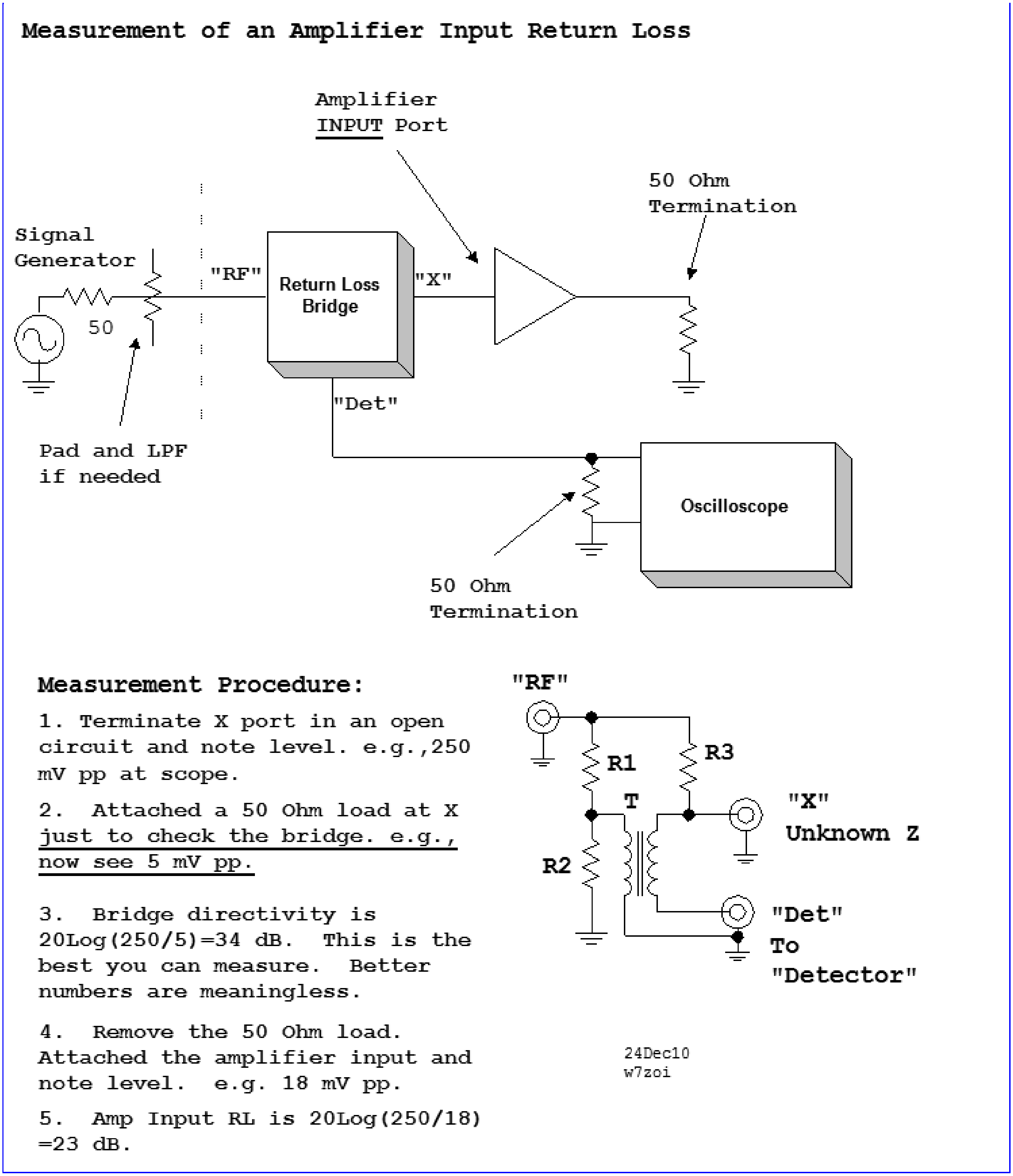
Measurement of an Amplifier Input Return Loss

Amplifier
INPUT Port
50 Ohm
Termination
Signal
Generator
"RF"
Return Loss
Bridge
"X"
50
Pad and LPF
if needed
"Det"
Oscilloscope
50 Ohm
Termination
"RF"
R1
R3
T
"X"
Unknown Z
R2
"Det"
To
"Detector"
24Dec10
w7zoi

Measurement Procedure:

1. Terminate X port in an open circuit and note level. e.g.,250 mV pp at scope.

2. Attached a 50 Ohm load at X just to check the bridge. e.g., now see 5 mV pp.

3. Bridge directivity is 20Log(250/5)=34 dB. This is the best you can measure. Better numbers are meaningless.

4. Remove the 50 Ohm load. Attached the amplifier input and note level. e.g. 18 mV pp.

5. Amp Input RL is 20Log(250/18) =23 dB.

Above — Measure the return loss input of an amplifier. You'll need at least one 50 ohm BNC feed-through terminator on your bench to test amplifiers with wired-in BNC connectors (such as on this amp); else just solder a 47, 49.9 or 51 ohm resistor from the amplifier output to ground. The BNC connectors allow you to quickly and solderlessly interface components such as filters, attenuators, oscilloscopes or **50 Ω** signal generators.

A typical amp measurement work flow may go something like this: Measure gain using a signal generator and the 50 ohm terminated scope; add the bridge and measure input return loss; finally, flip the amplifier around and measure output return loss. All 3 functions can be performed in 2

— 5 minutes including time to drink coffee + perform calculations by computer or with a HP scientific calculator.

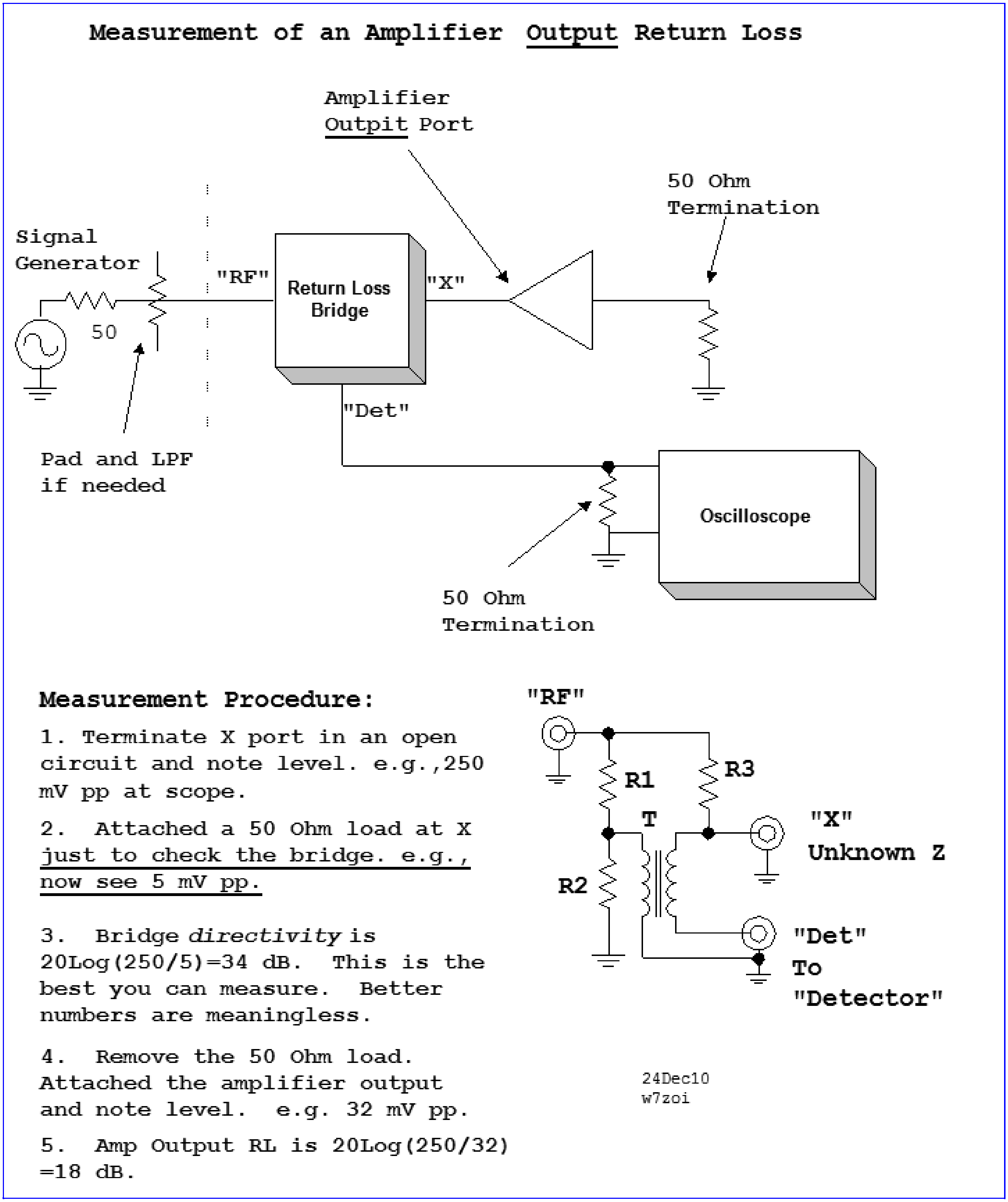

Measurement of an Amplifier Output Return Loss

Amplifier
Outpit Port
50 Ohm
Termination
Signal
Generator
"RF"
Return Loss
Bridge
"X"
50
Pad and LPF
if needed
"Det"
Oscilloscope
50 Ohm
Termination
"RF"
R1
R3
T
"X"
Unknown Z
R2
"Det"
To
"Detector"
24Dec10
w7zoi

Measurement Procedure:

1. Terminate X port in an open circuit and note level. e.g.,250 mV pp at scope.

2. Attached a 50 Ohm load at X just to check the bridge. e.g., now see 5 mV pp.

3. Bridge *directivity* is 20Log(250/5)=34 dB. This is the best you can measure. Better numbers are meaningless.

4. Remove the 50 Ohm load. Attached the amplifier output and note level. e.g. 32 mV pp.

5. Amp Output RL is 20Log(250/32) =18 dB.

Above — Measure the return loss output of an amplifier. The above 2 procedural diagrams provided by Wes, W7ZOI. Many thanks to Wes. These figures are copyrighted © by Wes Hayward, 2010.

Your signal generator should have a return loss of at least 20 dB for greatest accuracy — all of my bench test generators have at least 30 dB of return loss. If you have a signal generator with a low impedance output and place a 10 dB attenuation pad on the output, you'll have at least 20 dB of return loss.

In the above figures, Wes gives an open circuit return loss of 250 mV; I set my signal generator output so the open return loss is somewhere between 170 and 250 mV; this allows you to accurately measure a really good 50 ohm return loss at >= 5 mV or so. Some people may have trouble going any lower than 5 mV due to scope accuracy. This is just something to consider.

Bench Exploration

For me at least, a special case of return loss measurement exists; measuring the return loss of a local oscillator. Since the oscillator under test must be *on* during measurement, it's emitting a signal at the same frequency as the bridge signal generator and interferes with measurement. If some 50 ohm attenuation is added to reduce the local oscillator under test output signal amplitude, this increases the return loss of the local oscillator under test. This is normally a good thing, however, we seek the **raw** output return loss or output impedance of the local oscillator under test.

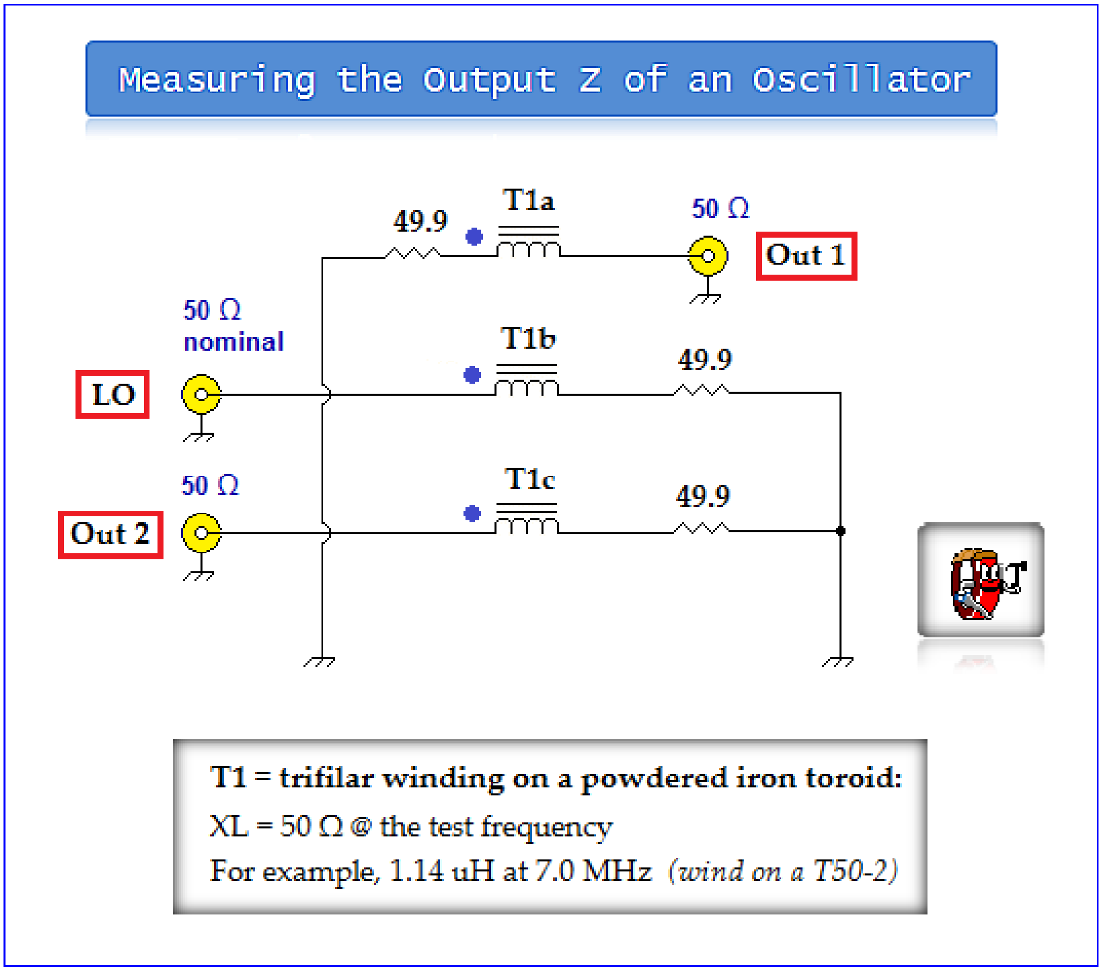
Measuring the Output Z of an Oscillator
49.9
T1a
50 Ω
Out 1
50 Ω
nominal
LO
T1b
49.9
50 Ω
Out 2
T1c
49.9
T1 = trifilar winding on a powdered iron toroid:
XL = 50 Ω @ the test frequency
For example, 1.14 uH at 7.0 MHz (wind on a T50-2)

Above — An initial experiment that a builder from Michigan, USA and I first used to measure the output impedance of a local oscillator consistent with the breadth and scope of this web site. We wanted something simple and wished to avoid building a vector network analyzer or performing ugly algebra. I built a simple crystal oscillator for 7.0 MHz using an output transformer wound to give a low impedance output. The circuit was measured and calculated using the instrument above and the formula and procedure below.

The calculated output Z was 33.2 ohms. I build a standard value resistor 6 dB attenuator pad from this table. After fitting the pad, I re-measured and re-calculated the output impedance at 46.8 ohms. This seemed okay. I built a couple of oscillators for other frequencies and the output impedances were hundreds of ohms! — disappointing. Still, we were on the bench in a solution-focused mode and needed to try something else.

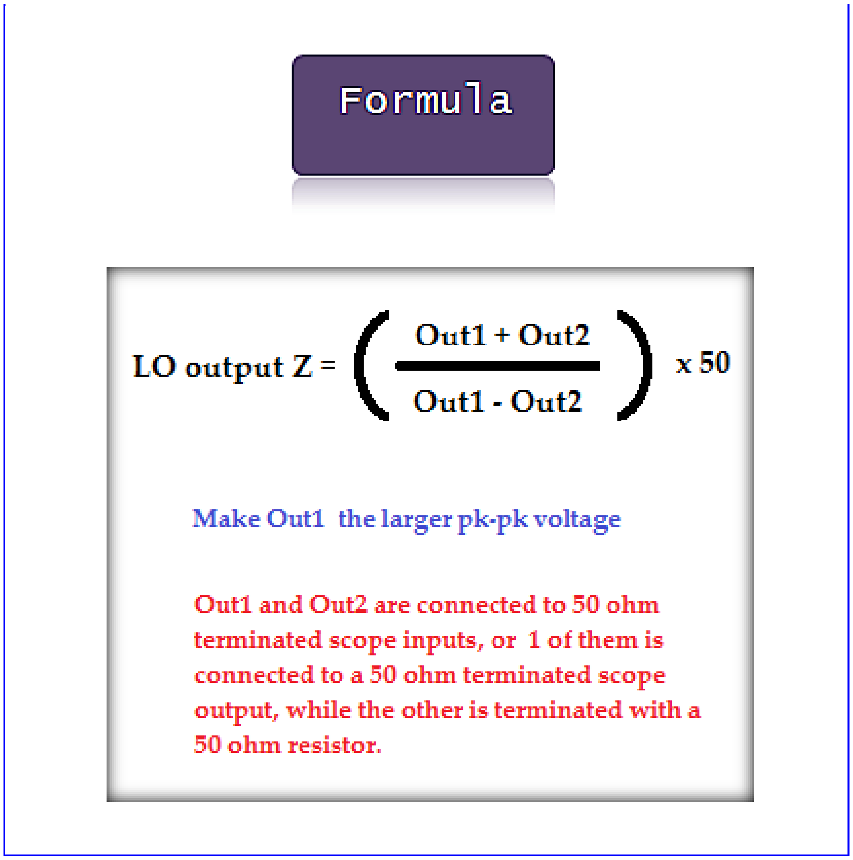

Formula

$$\text{LO output Z} = \left(\frac{\text{Out1 + Out2}}{\text{Out1 - Out2}} \right) \times 50$$

Make Out1 the larger pk-pk voltage

Out1 and Out2 are connected to 50 ohm terminated scope inputs, or 1 of them is connected to a 50 ohm terminated scope output, while the other is terminated with a 50 ohm resistor.

Above — The formula for calculating the output impedance with the *experimental* local oscillator output Z device.

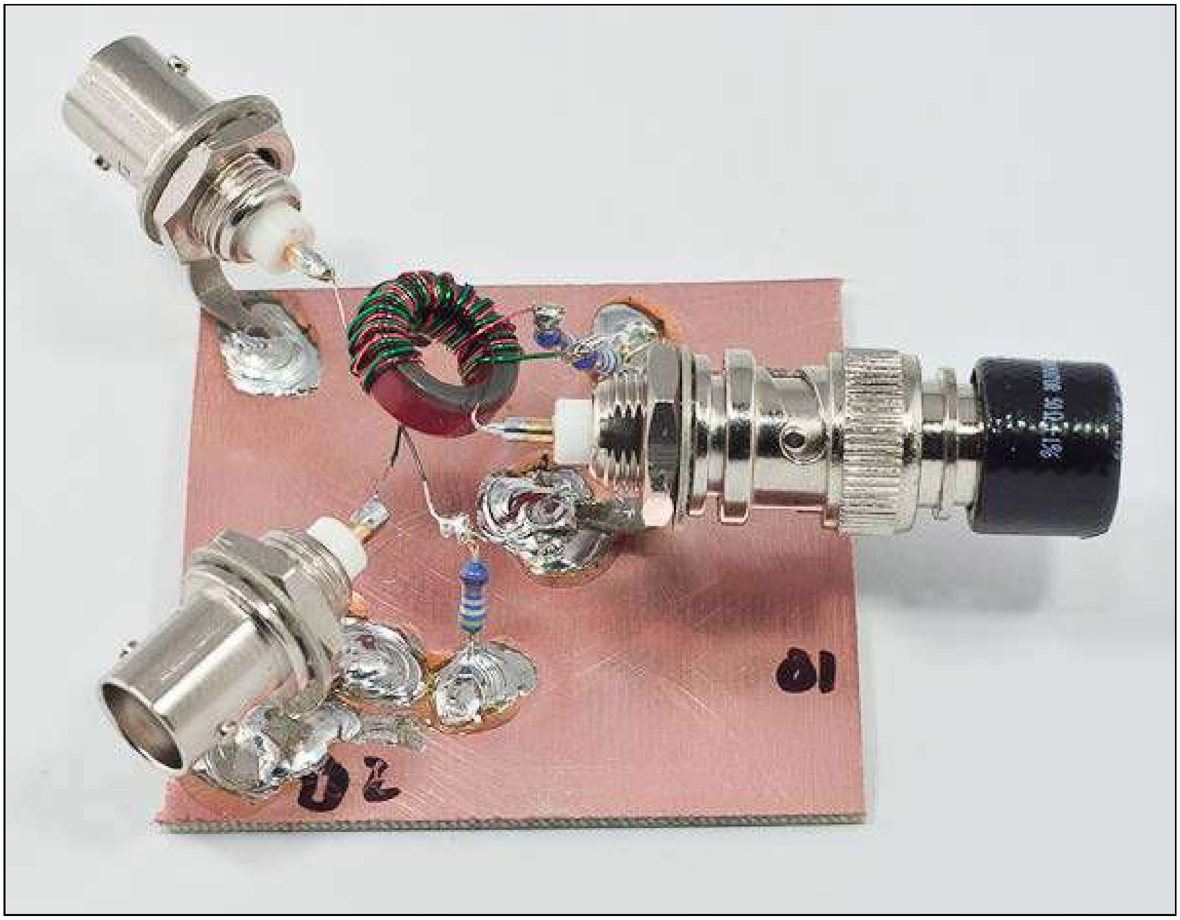

Above — Breadboard of the experimental L.O. output impedance bridge with a 50 ohm feed-through terminator on the Output 1 port. It failed to work as expected. Skillful adult problem solving goes something like this:

1. Identify the problem.
2. Brainstorm to generate some potential solutions.
3. Try out one of your ideas.
4. If that doesn't work, try another idea.
5. If none of your ideas work, wait a while, or ask an expert.

Well, we ran out of ideas and decided to ask experts for some more ideas; Professor Kuhn and Wes, W7ZOI.

I'll share their key messages. First, accurately measuring the output impedance of an RF oscillator can be difficult — measuring the return loss of a buffer amplifier is much easier. For this, some builders run the bridge signal generator on a slightly different frequency than the oscillator under test while using a spectrum analyzer as the 50 ohm RLB detector.

Another way is to short circuit the tank on the oscillator and measure the buffer output in the normal way — a popcorn solution indeed!

We tried calculating oscillator output impedance using different equations and 1 example is shown below. Failing to account for inductive and/or capacitive reactance plus resistance in the output circuit (including the transformer), plus upsetting the circuit during AC voltage measurement adds uncertainty to calculations — measurement seems more reliable.

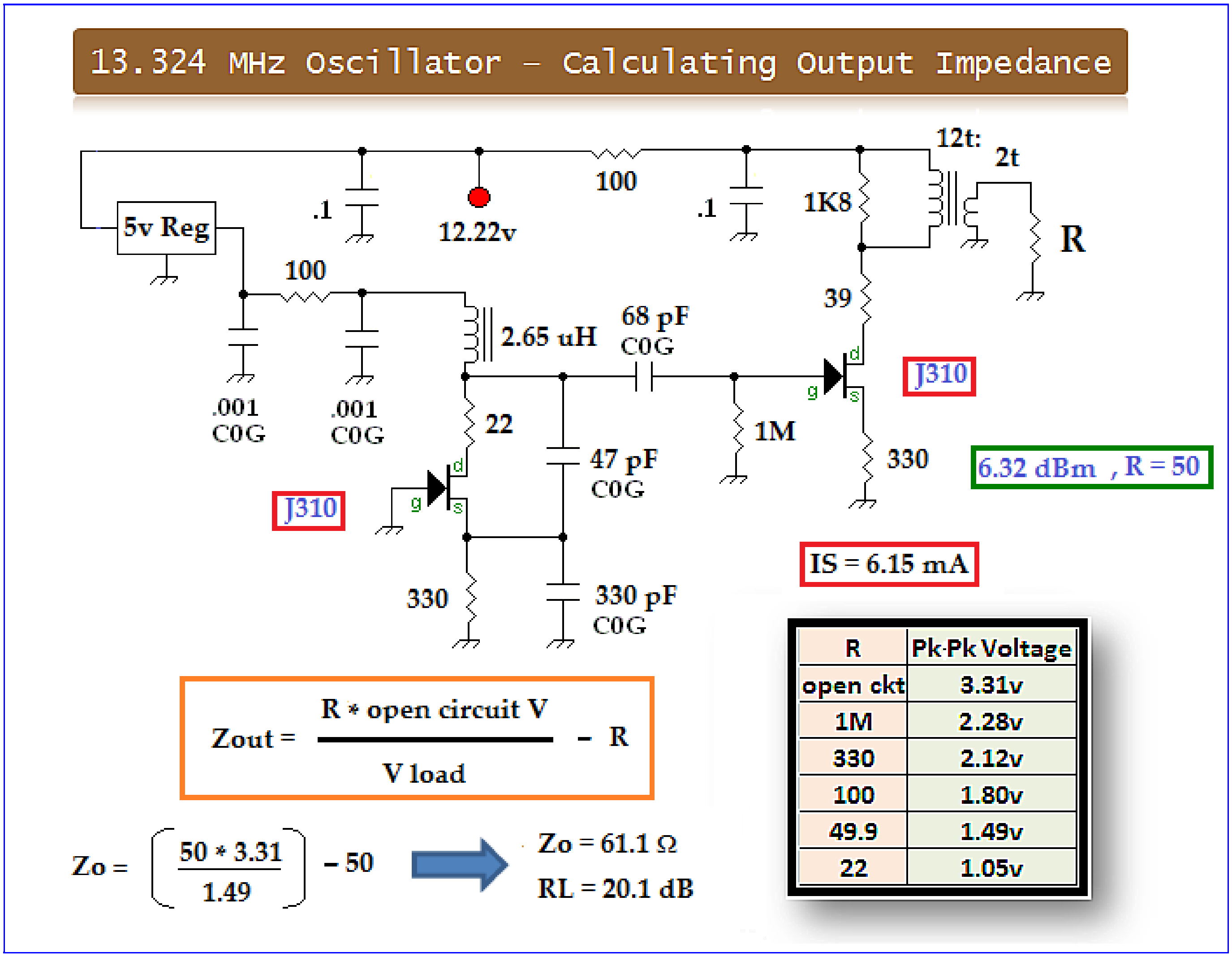

$$Zout = \frac{R * \text{open circuit } V}{V \text{ load}} - R$$

$$Zo = \left(\frac{50 * 3.31}{1.49}\right) - 50 \Rightarrow Zo = 61.1\ \Omega \quad RL = 20.1\ dB$$

R	Pk-Pk Voltage
open ckt	3.31v
1M	2.28v
330	2.12v
100	1.80v
49.9	1.49v
22	1.05v

Above — One method of calculating output impedance. Running the output at open-circuit likely effects the oscillator by changing its load despite having the JFET buffer. Some builders use this equation for calculating the output impedance in their audio amplifiers. This amplifier **should** have a 50 ohm output impedance based upon the transformer turns ratio and the 1K8 resistor across transformer: 1800:50 ohms = a 36:1 impedance ratio and a 6:1 turns ratio.

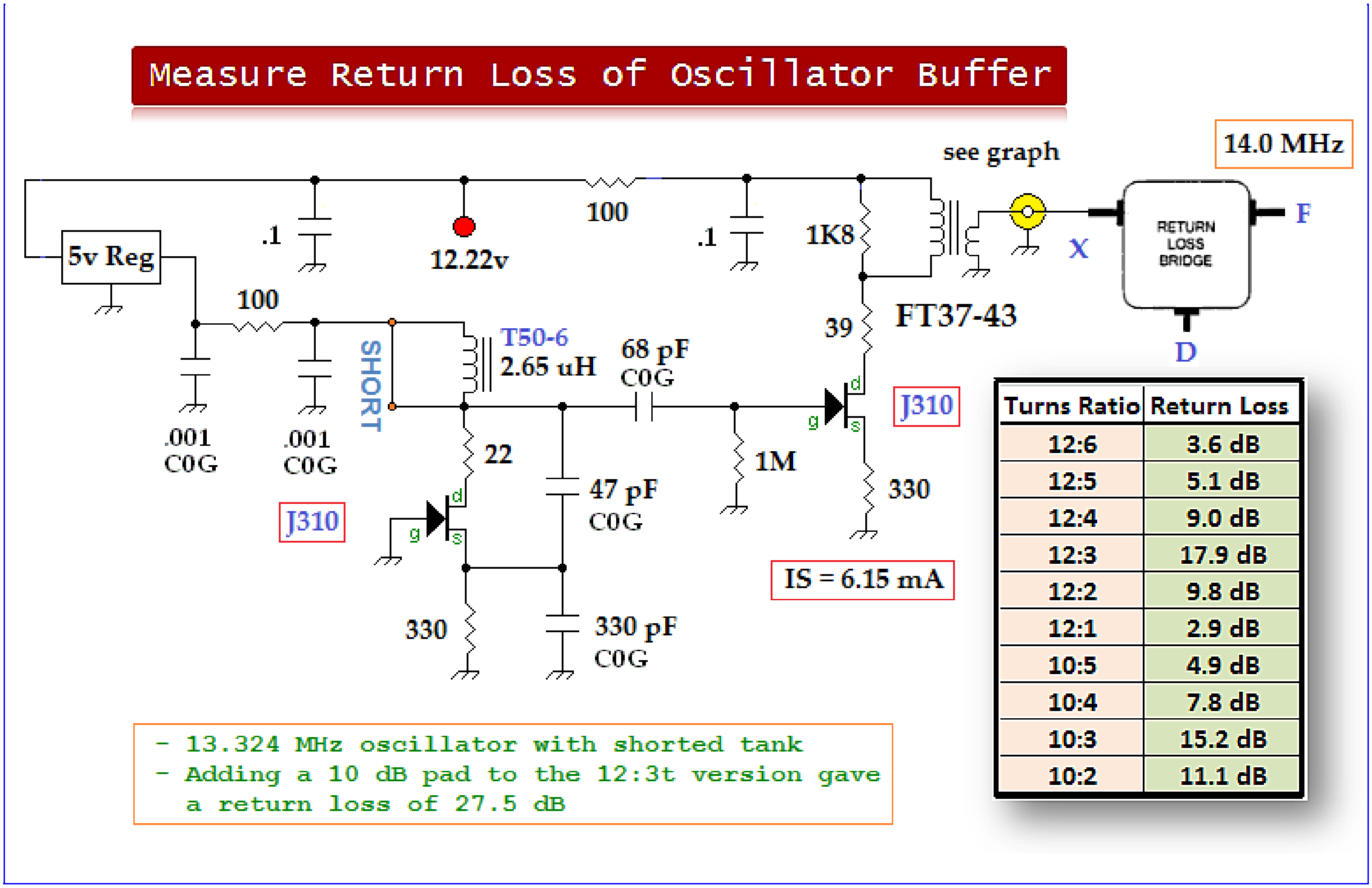
Measure Return Loss of Oscillator Buffer
see graph
14.0 MHz
5v Reg
.1
12.22v
100
.1
1K8
X
RETURN LOSS BRIDGE
F
D
100
SHORT
T50-6
2.65 uH
68 pF
C0G
39
FT37-43
J310
g
d
s
.001
C0G
.001
C0G
22
1M
330
47 pF
C0G
J310
330
330 pF
C0G
IS = 6.15 mA

Turns Ratio	Return Loss
12:6	3.6 dB
12:5	5.1 dB
12:4	9.0 dB
12:3	17.9 dB
12:2	9.8 dB
12:1	2.9 dB
10:5	4.9 dB
10:4	7.8 dB
10:3	15.2 dB
10:2	11.1 dB

- 13.324 MHz oscillator with shorted tank
- Adding a 10 dB pad to the 12:3t version gave a return loss of 27.5 dB

Above — A simple 13.3 MHz L-C oscillator was built and evaluated. After shorting the tank coil, return loss versus turns ratio was measured and tabled as shown. To my surprise, I observed the best match with a 4:1 turns ratio. This suggested that the transformer, wound on a ferrite FT37-43 toroid was exhibiting high resistance and far from the "ideal transformer". The inductance of my 12 turn transformer was 38.3 uH.

The initial secondary winding had 6 turns and then was reduced sequentially by 1 turn. After removing each turn, the 1/2 cm of increased wire length was cut off and the enamel scraped off of the new wire ending to ensure a short connection to the output jack. During measurement, unless the secondary transformer wires were kept tidy, a ~40 MHz oscillation occurred when the 1K8 resistor was disconnected. The 1K8 resistor prevented such oscillations and improved the return loss by 1-4 dB at the various turns ratios. Testing frequency was 14 MHz.

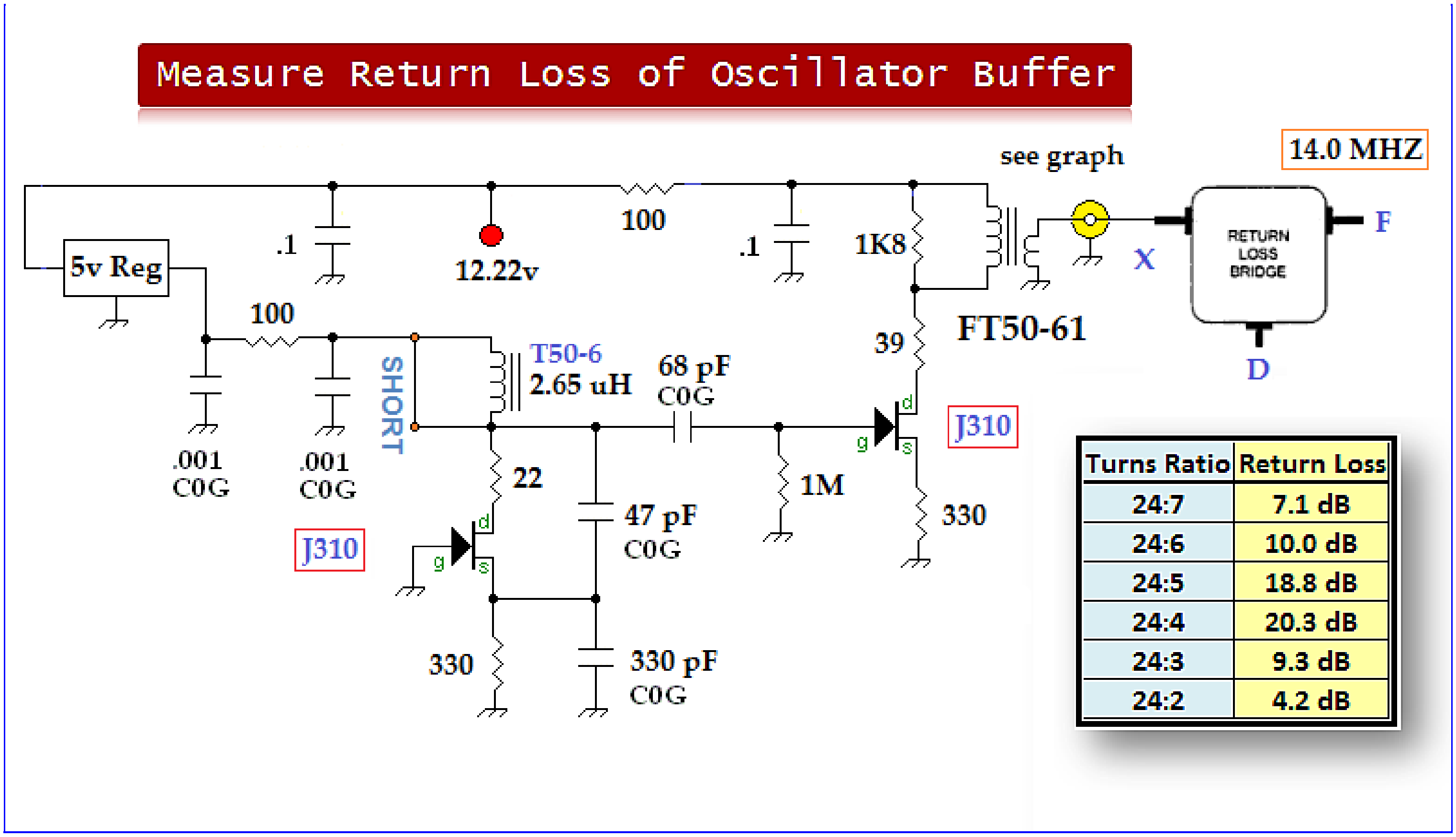
Measure Return Loss of Oscillator Buffer
see graph
14.0 MHZ
5v Reg
.1
12.22v
100
.1
1K8
X
RETURN LOSS BRIDGE
F
D
FT50-61
100
SHORT
T50-6
2.65 uH
68 pF
C0G
39
J310
.001
C0G
.001
C0G
22
1M
330
J310
47 pF
C0G
330
330 pF
C0G
d
g
s

Turns Ratio	Return Loss
24:7	7.1 dB
24:6	10.0 dB
24:5	18.8 dB
24:4	20.3 dB
24:3	9.3 dB
24:2	4.2 dB

Above — The same circuit with a lower-loss, FT50-61 ferrite transformer. I could have used a FT37-61, but prefer the 50-61 as the bigger core allows the use of heavier wire which provides some robustness when performing intensive experiments. The inductance of 24 turns on a FT50-61 measured 36.2 uH. Although lower permeability ferrite toroids require more windings, this transformer is closer to the "ideal transformer" than that wound on a FT37-43 ferrite core — a 6:1 turns ratio gave the best return loss; the output Z is pretty close to 50 ohms.

The information garnered during these tests proved enlightening and reinforces why bench measurement provides the greatest way to learn about and optimize your circuits. I hope this simple web page on return loss measurement fuels your own experiments — the most important experiments will be those *you* do on *your* bench.

Miscellaneous Figures and Photos

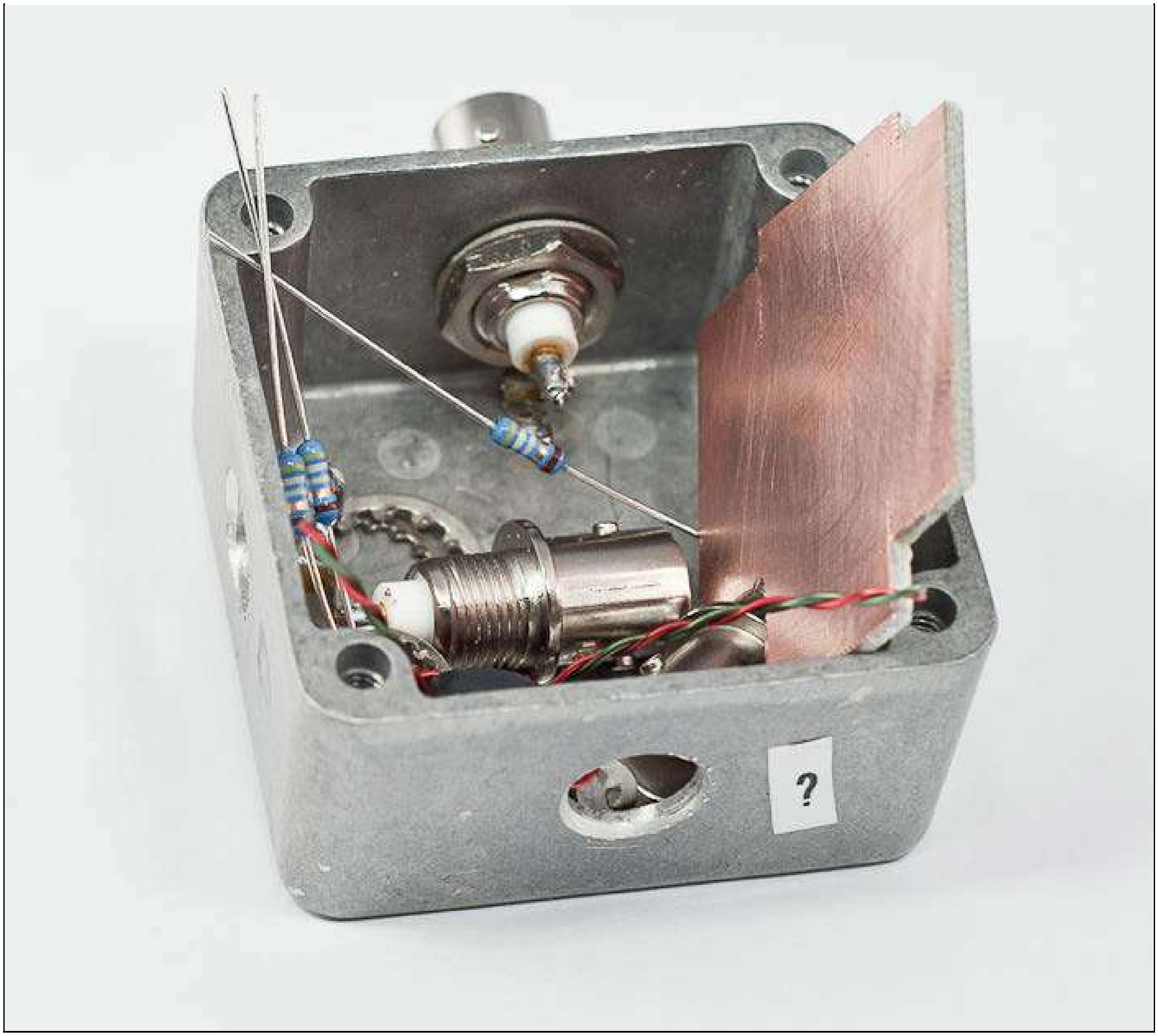

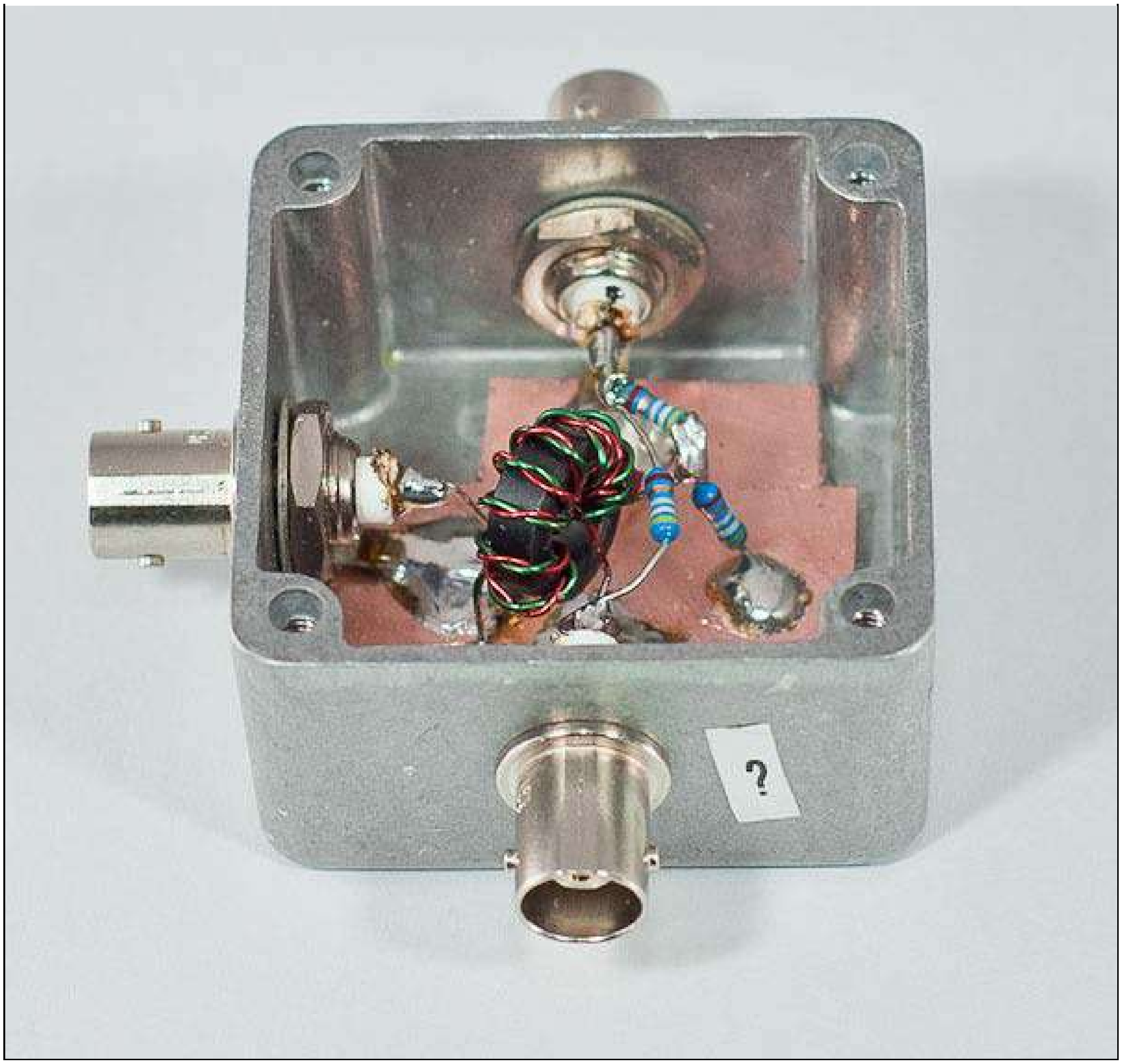

Calculate Output Z of an Oscillator – Experiment 1

Measure the open circuit output amplitude of the oscillator and then add a load so that the loaded voltage dropped to something in the range of 50 to 75 percent of the open circuit voltage. Then use voltage divider theory to calculate the source resistance of the oscillator.

$$V_{\text{loaded}} = \frac{V_{\text{unloaded}} * R_{\text{load}}}{(R_{\text{oscillator}} + R_{\text{load}}}$$

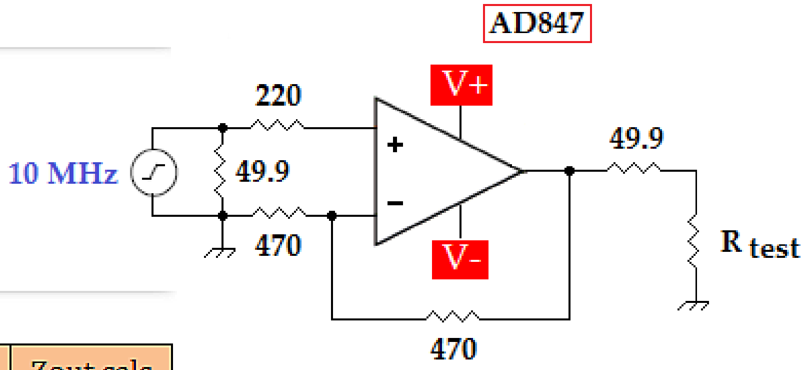

Rtest	Vout	Icalc	Pcalc	Zout calc
1000000	0.9999	9.999E-07	9.998E-07	
220	0.8148	0.0037	0.0030	49.9779
50	0.5003	0.0100	0.0050	49.9300
22	0.3056	0.0139	0.0042	49.9823

RF — Test and Measurement

Miscellaneous RF Topics 2011

Introduction

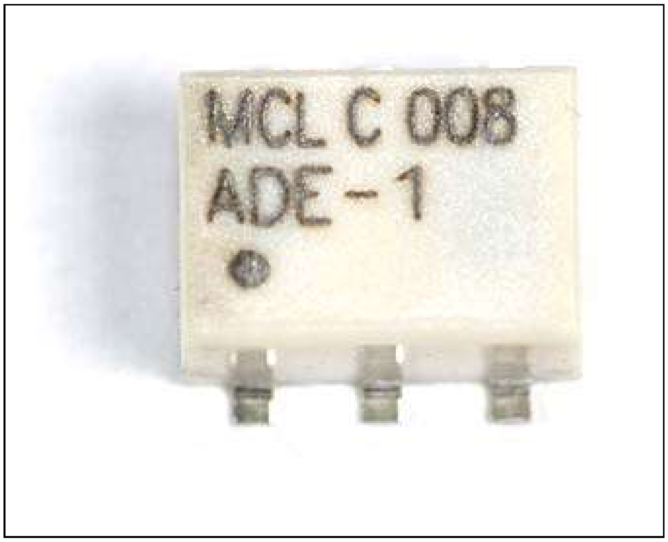

My basic goal for Fall and Winter 2010 was to fearlessly advance my RF design ability — pushing just beyond my comfort zone to impose the psychological stress that promotes focused learning. You may have experienced this in University or when working against a deadline. Cramming, "burning the midnight oil", or locked room brainstorming exemplify this approach.

These circuits feature carefully measured input and output impedances (nominal impedance = 50 Ω), plus voltage gain and DC current.

This web site spawns email with builders worldwide. Our interactions are varied; getting help, giving help — or just chat. On occasion, I design, build and/or test circuits to help struggling builders. After spending considerable time, I email them my work hoping it will help. Often enough, I never receive any acknowledgement from these readers — did the circuit work or did they appreciate I spent 1-2 hours researching their concern?. This is actually normal — we must constantly strive to overcome our innate, self-centered nature; lest it dominate our behavior. все нормальные.

To that point, I wish to gratefully acknowledge the people who support me in this hobby: Wes, Ken, Scott, Peter, Tom and the many others whose email advice and published and private work informs and inspires me.

Topics:

1. Transmit Mixer Experiments
2. Bipolar Transistor Feedback Amplifier Experiments
3. JFET Common Gate Transistor Amplifier Experiments

Navigation and Preamble:

This web page grew into a large monster — and includes a supplemental web page with numbered topics referenced in the text. I apologize for the navigation difficulties this web page poses. Equal time was spent experimenting with the circuit designs and circuit photography. I strive to provide a variety of bitmap and photographic image styles on this web site.

---

1. Transmit Mixer Experiments

Since I've never experimented with transmit mixers, I didn't appreciate how much time goes into their design. Consider, for example, the LO system from the project entitled A Monoband SSB/CW Transceiver in Chapter 6 of EMRFD. The mega low (about -20 dBm) output from a diode ring mixing a VFO and crystal oscillator is triple tuned band-pass filtered and then amplified to +8 dBm. Continuing on, the transmitter chain features more mixing, band-pass filtering and voltage amplification by a feedback amplifier chain boosting the signal to around 300 mW. The circuits needed to mix, filter and amplify this RF chain would challenge most amateur designers — me included.

Contrast this with a typical first transmitter built by a new builder. Likely your first scratch homebrew transmitter consisted of a crystal oscillator, a keyed Class A buffer/amplifier and perhaps a Class C final amplifier. No mixer was needed for we obtained a crystal cut on the frequency of choice. Our focus was **power**— getting 0.25 to 1 watt into our antenna system! A good example was the Tuna Tin 2 transmitter by the late Doug DeMaw, W1FB that only used 2 stages. Although Doug wrote his 1976 article for Hams to build a transmitter from parts found at home, kitted versions are sold today.

Returning to transmit mixers — as amateur designers, we likely need to start on a small portion of the transmit chain and then after developing some competency, slowly extend our experiments all the way to the antenna port. In Fall 2010, I just examined some basic transmit mixing to get a feel for what's involved and what to expect. Mixing signals is a complex affair encompassing topics such as intercept point, conversion gain or loss, image noise suppression, noise figure, spurious/intermodulation products and port isolation. To keep things simple, only mixer port isolation and reducing spurious mixer products were examined.

Before beginning, I express the following concern: We experimenters, as stewards of the airwaves, must build exemplary transmitters with very low spurious outputs. I follow the example of Wes, W7ZOI and others — my transmit chains have spurious frequencies at least 50-60 dB down from the carrier (dBc). As a web author and radio amateur, I never want to directly or indirectly contribute to RFI and hope you agree.

Why Use a Transmit Mixer?

If you plan to design a superheterodyne based transceiver, you'll probably need to use a transmit mixer. Also mixing 2 frequencies permits using cheap microprocessors crystals to target a desired transmit frequency; separate crystal oscillators drive the RF and LO ports of the mixer. For added flexibility, the LO can be converted to a VFO once you have the basic design working well.

I purchased a bag of low cost crystals. By mixing 2 appropriate crystals, output on a Ham band is possible. For example, crystals at 2.048 + 5.0688 MHz = 7.117 MHz; 4.194 + 11.228 MHz = 7.034 MHz; and 3.932 + 11.046 MHz = 7.114 MHz. I frequently operate QRP on 40 Meters in the USA Novice band, so 7.114, or 7.117 MHz is okay. This helps CW operators avoid all the RTTY and QRM down in the traditional 40M band QRP frequency window.

Some Mixer Bullets

Mixers have 3 conventionally named ports; RF, LO (local oscillator) and IF (output).

The diode ring mixers presented are Level 7 mixers. Maximal LO power is 7 dBm.

Many builders limit the maximum RF power into a Level 7 diode ring transmit mixer RF port to between 0 to -3 dBm.

The term *isolation* refers to the amount of LO power that leaks into the RF or the IF ports.

Low-pass filtering the LO *can* significantly reduce harmonic products in a mixer

The top of the spectrum analyzer screen (always the top, and never the bottom) is called the reference level. That is the power at the top. If you have a signal generator with the output adjusted to be -27 dBm and pass this signal into the spectrum analyzer and adjust the attenuation in the analyzer to put that signal at the top of the screen, you then the reference level is -27 dBm. *(Pertains to examining a mixer output in a SA)*

Choosing a Mixer

A number of mixers were considered; passive, active, unbalanced, single-balanced and finally, double-balanced. The diode ring mixer is an obvious good choice commensurate with my goals of reducing spurs, LO feed through and achieving high port-to-port isolation. In future web pages, other mixers may be presented, however this page is focused on the diode ring mixer. Click for a file with a few scanned pages concerning mixers from my "ideas only" notebook from ~2002. I own over 30 notebooks now.

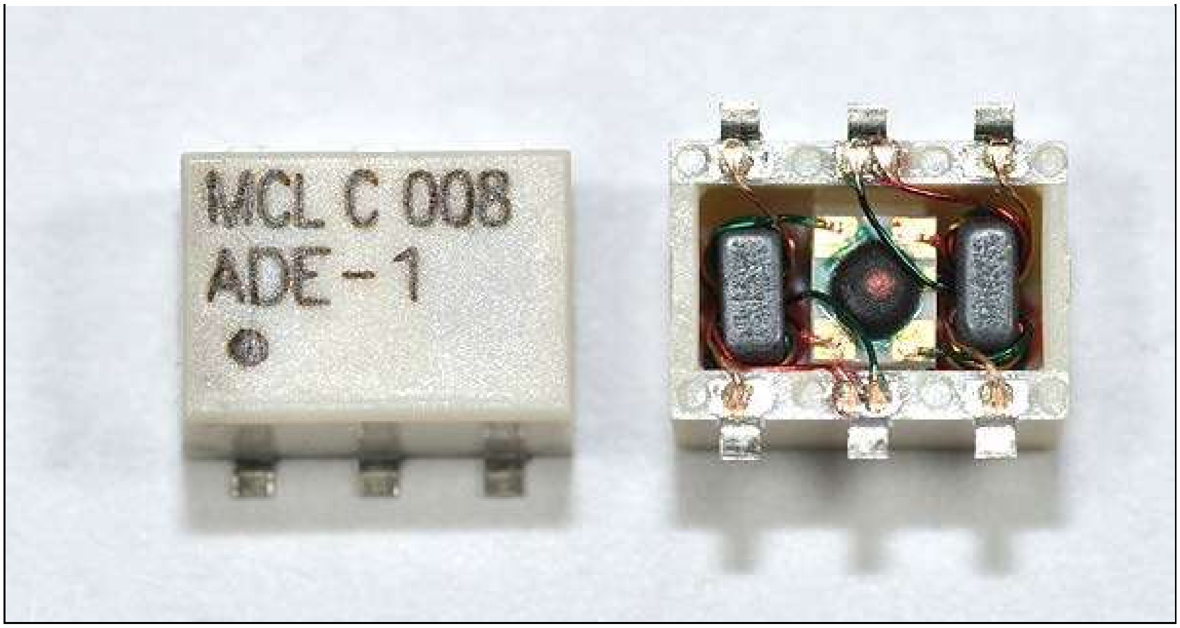
MCL C 008
ADE-1

Above — ADE-1 diode ring mixers. We're using these now as they're cheaper than the SBL-1, TUFF-1 etc. hole-though versions. Although SMT parts, they can be flipped over and wired "normally" with a little effort, steady hands and good vision. Mini-Circuits will sell them in small quantities to Hams; email them and enquire. I feel the diode ring mixer has been misunderstood by some amateur builders — lore and misperceptions that the 7 dBm LO port drive, the need for 50 ohm port terminations, a ~ 5 dB insertion loss and cost make them undesirable. Their excellent performance and design challenges are reasons why we use them; "the journey — not the destination" stuff.

In receiver applications, some builders and kit sellers seem more focused on features such low-battery indicators, digital displays, miniaturization and cost containment than basic receiver performance. Certainly keeping cost down down deserves consideration, however, good mixer performance is king. You'll have to decide what's affordable and important and build accordingly.

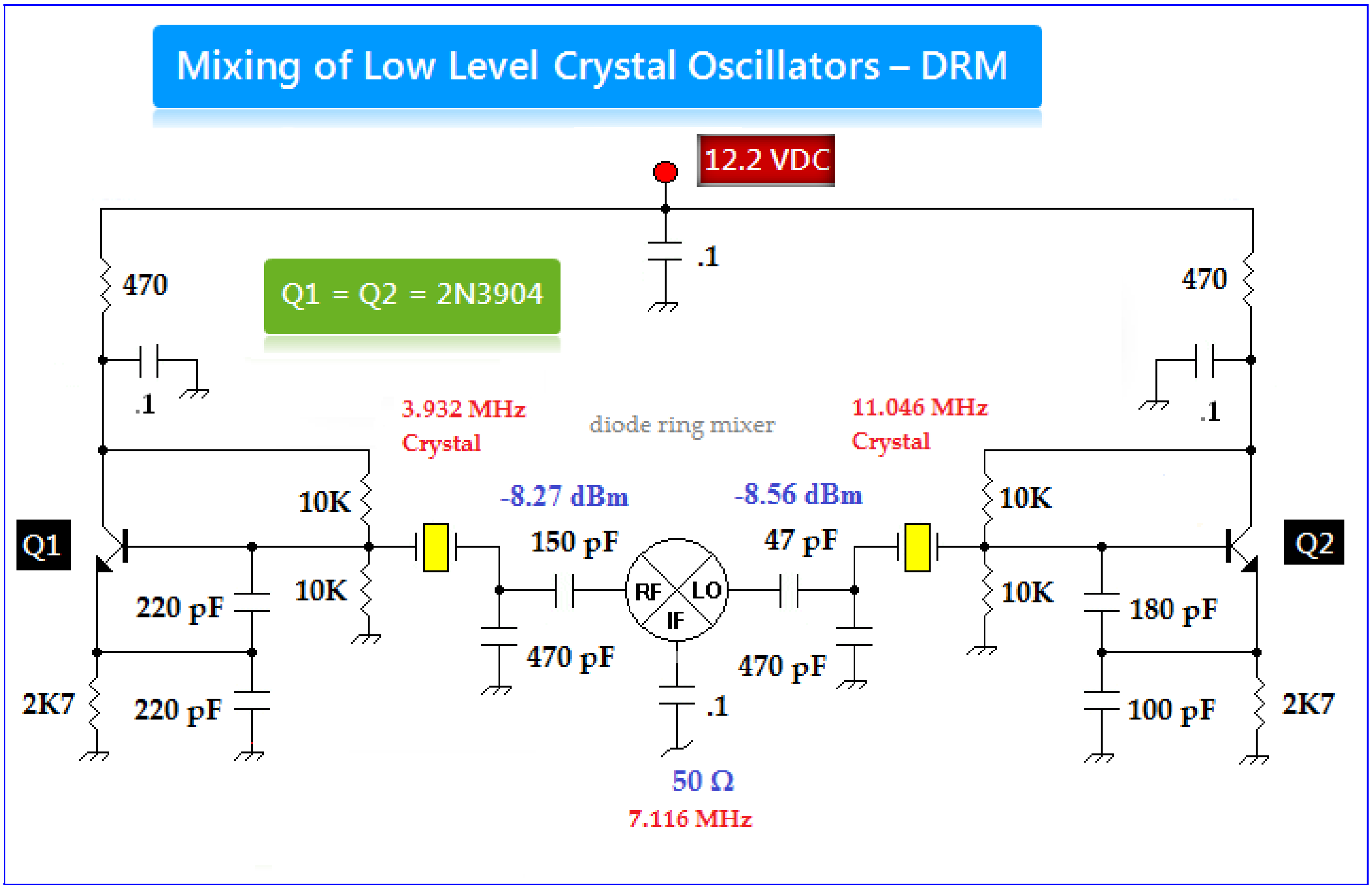
Mixing of Low Level Crystal Oscillators – DRM
12.2 VDC
.1
470
Q1 = Q2 = 2N3904
470
.1
.1
3.932 MHz
Crystal
diode ring mixer
11.046 MHz
Crystal
-8.27 dBm
-8.56 dBm
10K
10K
Q1
150 pF
47 pF
Q2
220 pF
10K
RF
LO
IF
10K
180 pF
470 pF
470 pF
2K7
220 pF
.1
100 pF
2K7
50 Ω
7.116 MHz

Above — My very first transmit mixer experiment. My hope was to build a transmit mixer possessing low spurious output to alleviate the need for stiff, post mixer band-pass filtering such as a triple tuned band-pass filter. Thus, low-level, low distortion output was taken from between each crystal oscillators' shunt capacitor and crystal. The desired output frequency is ~7.114 MHz to build a transmitter for the 40 Meter Ham band.

The mixer output to 50 MHz looked like this in a Spectrum Analyzer. The dominant frequencies are the **sum** and the **difference**: 11.046 **+** 3.932 MHz = **14.98 MHz**; 11.046 **-** 3.932 MHz = **7.114 MHz**. The frequencies realized are *slightly* different since the oscillator output is shifted by crystal variances and from circuit capacitance.

In the experiments that follow, I built some circuits to filter and/or amplify the output of the *above* mixer circuit.

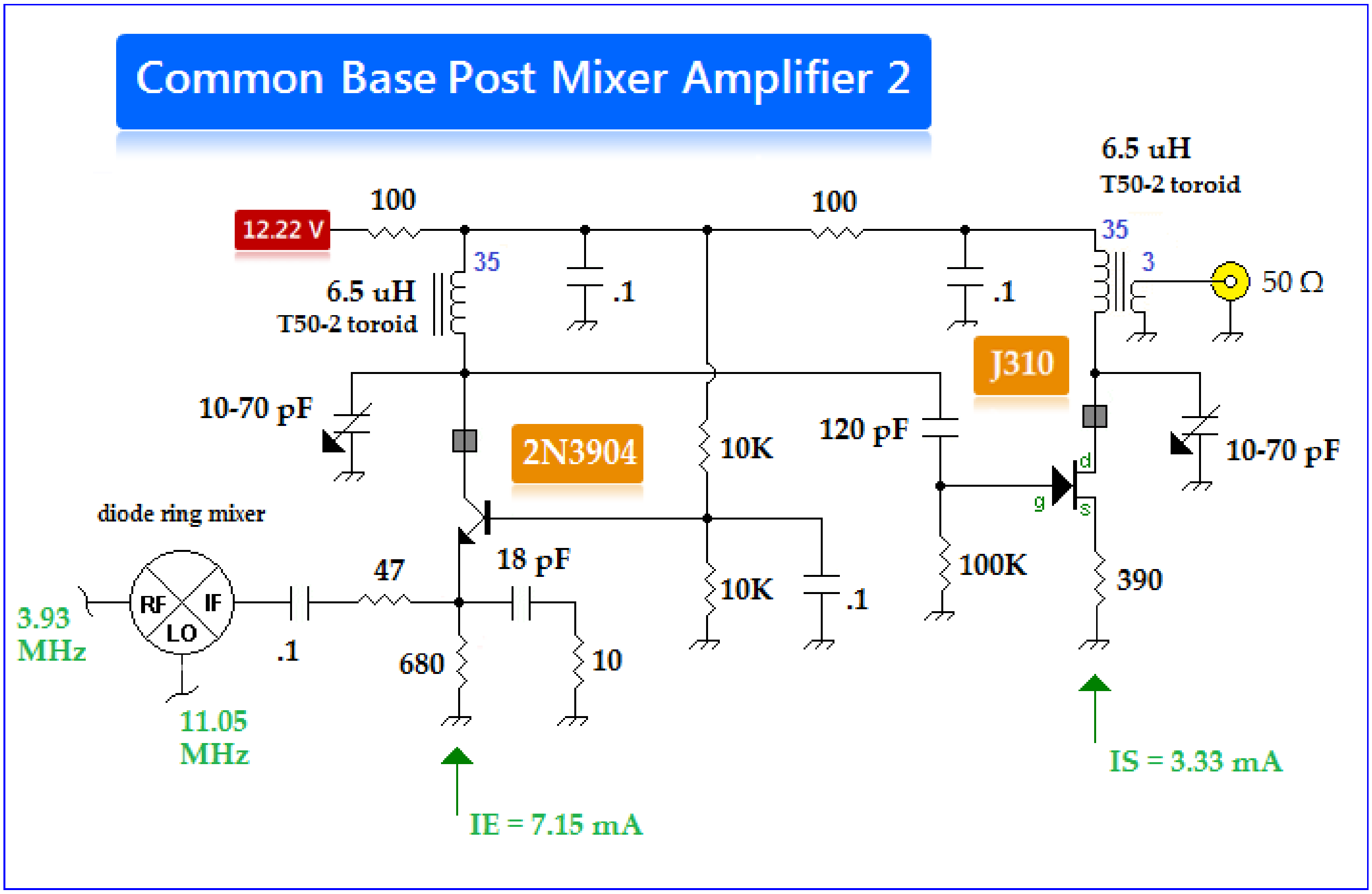

Above — The first post mixer amp; a common base input amplifier that's AC coupled to a common drain FET amp. I hoped that 2 tuned L-C tank circuits could substitute for a passive double or triple-tuned band-pass filter, plus provide some gain.

A broad-band, common base input amp was chosen to properly terminate the diode ring mixer and alleviate the need for a diplexer. A ~50 Ω input impedance is established by a 47 Ω series resistor since the 2N3904s input impedance is quite low due to the moderately high emitter current employed to boost gain and IMD performance.

This amplifier failed to reduce spurious output 50 dB down or greater — my design goal. Here are its scope and spectrum analyzer outputs; please observe that the unwanted 14.98 MHz signal is only 32 dB down from the desire IF of 7.114 MHz. An RC network consisting of a shunt 10 ohm resistor + an 18 pF cap provides additional low-pass filtering above 20 MHz. I attribute this simple filter to Dr. Ulrich Rohde as I have seen it in some of his post mixer, common base RF amplifier designs. Click for a brief supplement regarding his low-pass network ***(#2 RC Low-pass Network on the Supplemental Page)***

The amp design shown above was actually an improved version of this prototype. In the prototype amp, the mixer power at 14.98 MHz was only 23 dB down from the desired intermediate frequency of 7.114 MHz. You can't expect a single L-C tank to well filter a mixer output. Unfortunately, the 1K2 -12K resistor providing DC bias for the emitter follower lowers the Q of the common-base collector tank circuit. 12K is better than 1K2 in this regard. Poor performance sparked the design of the second common base amp shown above.

I use ferrite beads and 51 Ω resistors interchangeably on the collector/drains of amplifiers to snuff out UHF oscillations. According to my experiments, the resistors may work better. I purchased the ferrite beads from Diz. After the common base amplifiers shown above, I decided to try a tuned input + output common gate JFET amp:

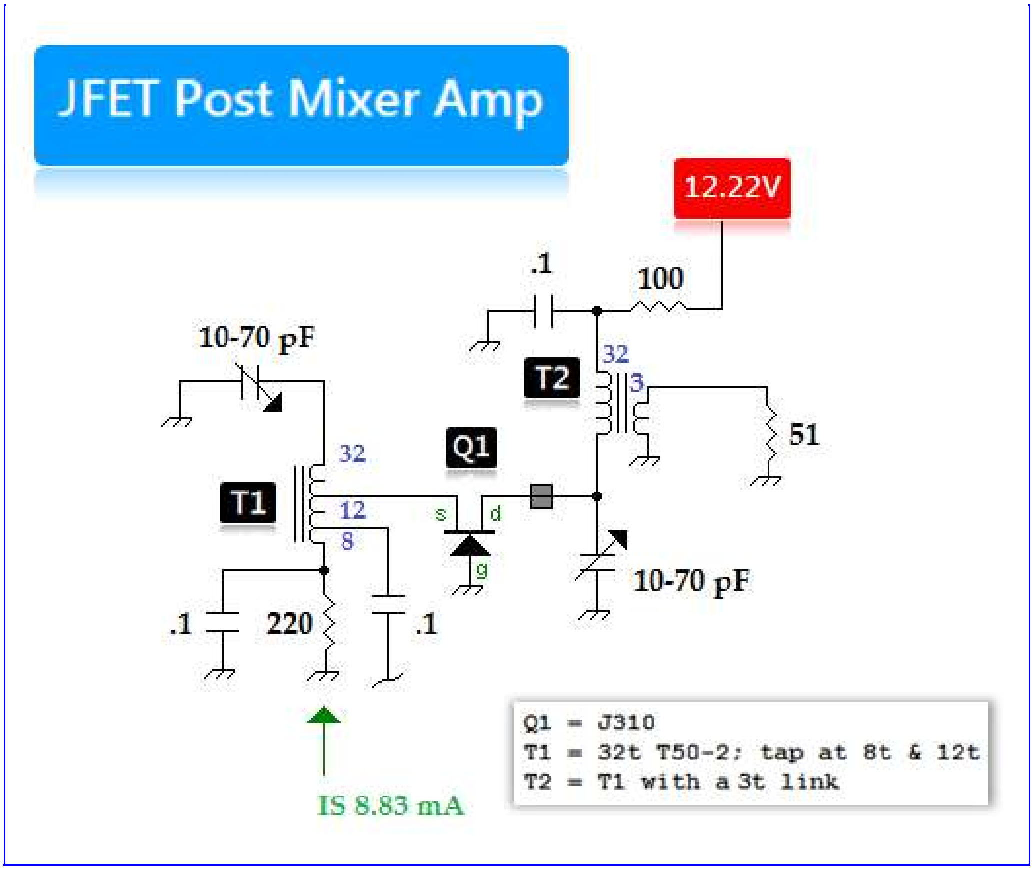
JFET Post Mixer Amp
12.22V
.1
100
10-70 pF
T2
32
3
51
Q1
32
T1
12
8
s
d
g
10-70 pF
.1
220
.1
IS 8.83 mA
Q1 = J310
T1 = 32t T50-2; tap at 8t & 12t
T2 = T1 with a 3t link

Above — A JFET common gate amp with tuned input and output built about 3 years ago. For spectrum analysis. I padded the amplifier output to provide a -28 dBm 7.116 MHz signal. The vertical resolution on the SA is the standard 10 dB/division. As shown, the 14.98 MHz signal is ~ 39 dB down; an improvement over the amplifiers shown previously. This narrow-band amplifier requires a diplexer. I wanted *better* filtering than that offered by amplified circuits with 2 tuned L-C circuits, so I halted this experiment and decided to try a triple tuned band-pass filter.

This JFET circuit experimentation spawned over 8 weeks of experiments concerning common-gate RF amps — some of them appear later.

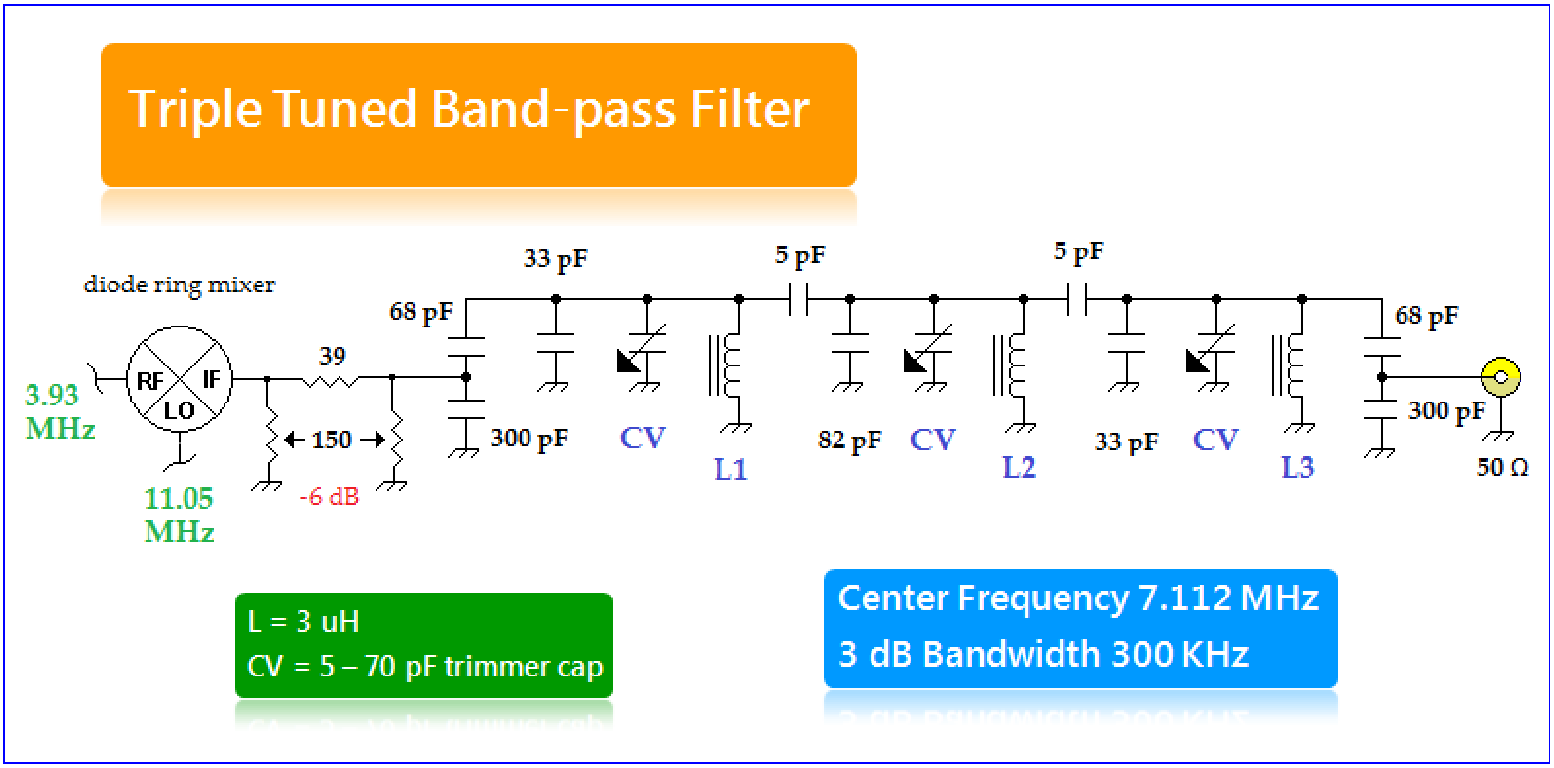
Triple Tuned Band-pass Filter
diode ring mixer
3.93 MHz
RF
IF
LO
11.05 MHz
39
150
-6 dB
68 pF
300 pF
33 pF
CV
L1
5 pF
82 pF
CV
L2
5 pF
33 pF
CV
L3
68 pF
300 pF
50 Ω
L = 3 uH
CV = 5 – 70 pF trimmer cap
Center Frequency 7.112 MHz
3 dB Bandwidth 300 KHz

Above — A triple tuned band-pass filter designed with software from EMRFD called TTC-08. Click for the breadboard photo. The diode ring was connected to the filter input via a 6 dB attenuator pad using short leads. Click for analysis in GPLA. All inductors are 3.0 uH — 23 or 24 turns on a T50-2 powdered iron toroid. I measured all the inductors and obtain the exact desired inductance by expanding or squishing the wire turns, or if necessary, adding (this means rewinding the entire coil) or removing a turn. If you lack an inductance meter, just winding the formula calculated number of turns will be close enough for most applications — I only got a good L- C meter in 2009 and somehow managed.

I learned that the ultimate way to peak a triple tuned filter is by tweaking the tuning capacitors while it's connected to a spectrum analyzer — what a thrill!

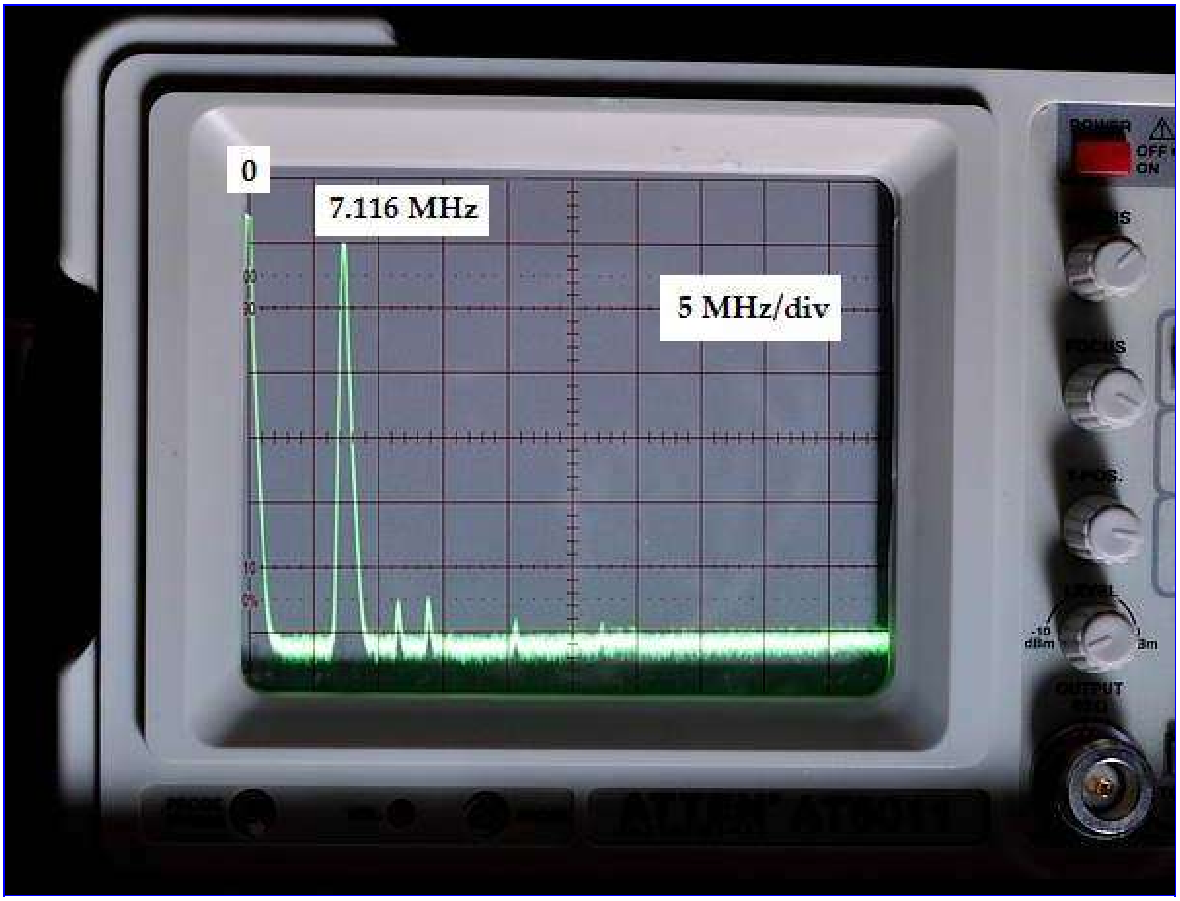

Above — While a little tedious to build and align, the triple tuned filter worked magnificently; the strongest spur is 54-55 dB down and the 14.98 MHz signal is gone. Insertion loss = 2.5 dB. This experiment provided a benchmark of what great post mixer filtering looks like. Post mixer filtering is an important topic worth studying further:

Why do we need filtering on a mixer output?

Let's examine mixer ports more closely. A port is just a pair of wires where signals are applied or removed. There are 2 kinds of mixer *outputs*: 1) the sum plus difference frequencies; 2) spurs.

Further, 2 kinds of spurs occur: One type is straight feed through where 1 signal from the 2 input ports makes it out to a 3rd port. Examples include LO feed through to the RF port, or LO feed through to the IF port.

The other type of spur is a mixing product such as a harmonic.

In general, the mixer output frequencies are numerically described by an equation:
IF (output) = N x L +/- M x R
N and M are both integers, 1, 2, 3, ... L = local oscillator frequency, R = radio frequency

A mixer is said to be balanced when you duplicate some of its functions and then combine them — usually with transformers. Consider, for example, the single diode mixer — they work, but the output contains ++ feed through and spurs. A mixer with 2 diodes or 2 FETs etc. can be much easier to use because the transformer combines the signals in a way that cancels some of the spurs and feed through. The double-balanced diode ring mixer uses four diodes and 2 transformers — producing even less feed through and harmonic output.

In a double-balanced diode ring mixer, the LO and RF ports are balanced and all ports of the mixer are isolated from each other. The double-balanced mixer greatly reduces, but does not stop all LO feed through at the RF and IF ports. A wideband match at 50 ohms is required to maintain mixer balance; hence you will often see attenuator pads on the LO, RF and *especially* the IF ports.

Let's focus on the IF port. Attenuator pads absorb any reflected mixer products and signals coming back into the IF port, thus increasing the match to the IF port. You may have noticed some builders use a diplexer on the IF port. The diplexer presents a wideband match to all IF port frequencies — passing the desired sum or difference frequency and absorbing the unwanted mixer products reflected back into the IF port by subsequent stages.

Since the IF output contains the sum and difference of the LO + RF, LO feed though, and other spurious energy, band-pass filtering is required to launder the IF signal into something useful. Following a transmit mixer, we filter with an L-C band-pass filter — after a receive mixer, crystal band-pass filters dominate. If you choose an unbalanced mixer or single-balanced mixer, filtering becomes more difficult than with a double balanced mixer. Unbalanced mixers are usually reserved for situations where high performance is being sacrificed for cost containment and/or want of a low parts count. There is no free lunch — you either alleviate as many mixer products as you can at a low-level with good practices, or have to deal with them down your signal chain while sacrificing optimal mixer performance.

Double balanced mixers are sensitive to non-resistive IF port terminations. When improperly terminated, the 2 transmission-line transformers work poorly — any reflected power generates high voltage across the diodes and degrades mixer performance. According to Dr. Ulrich Rohde, some proper ways to terminate the mixer include using a diplexer followed by a wideband 50 Ω feedback amp, or a common-gate JFET amplifier. *(Reference 1)*

Improved Local and RF Oscillators

In the earlier experiment, I really should have the run the LO port at 7 dBm. In order to improve my experiments, new crystal oscillators were designed with emphasis on correct LO output power and low harmonic energy.

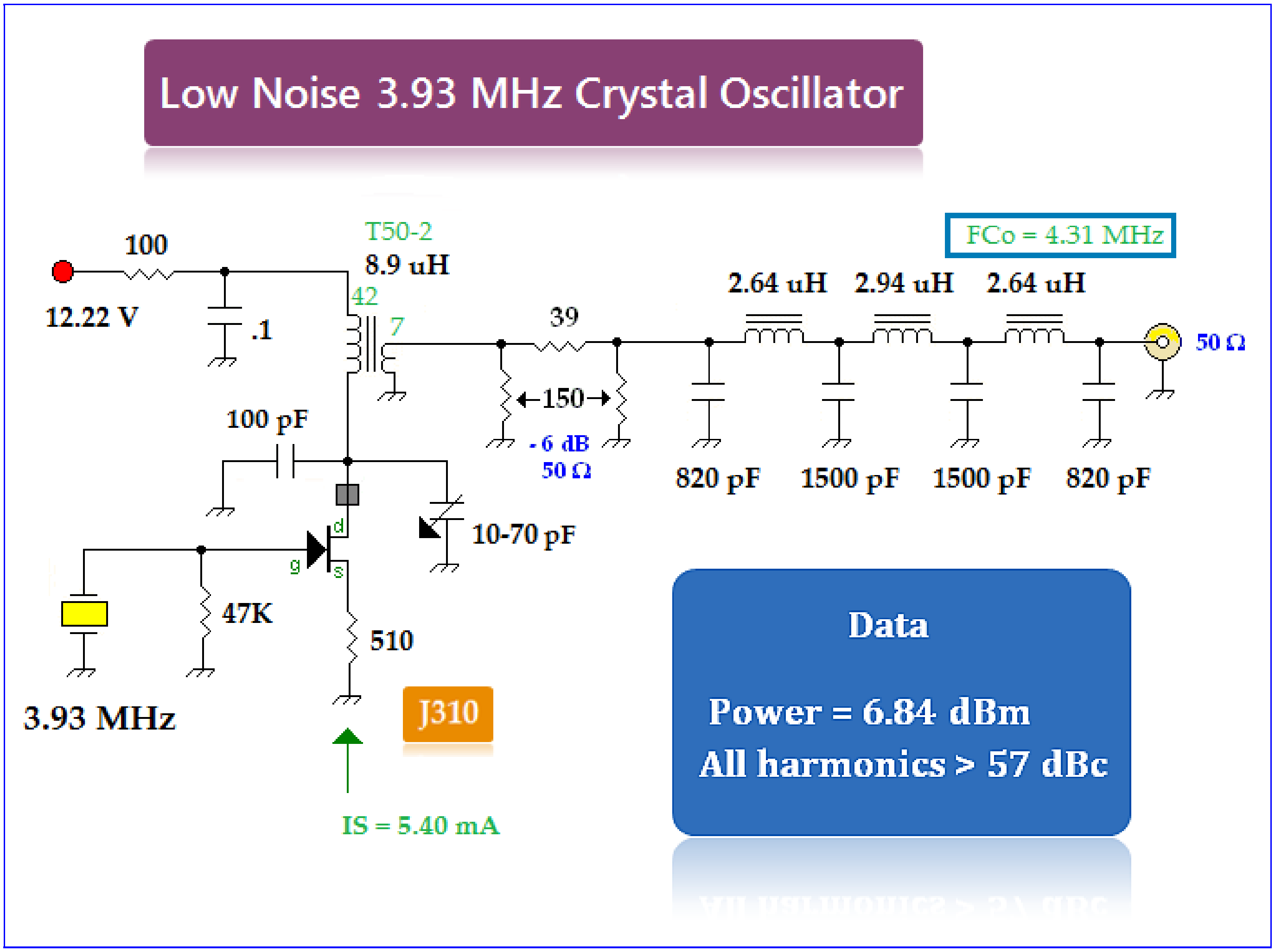

Above — The new LO; a 3.93 MHz crystal oscillator with stiff low-pass filtering. Admittedly, this 7 element Chebychev low-pass filter is overkill, however, I wanted to examine filtering and learn how much is required. On the bench — do whatever you like; even chasing crazy personal goals can be instructive and help you relate to information from texts and articles, or satisfy a whim. I can read something 100 times, but may not understand it well until I actually do it.

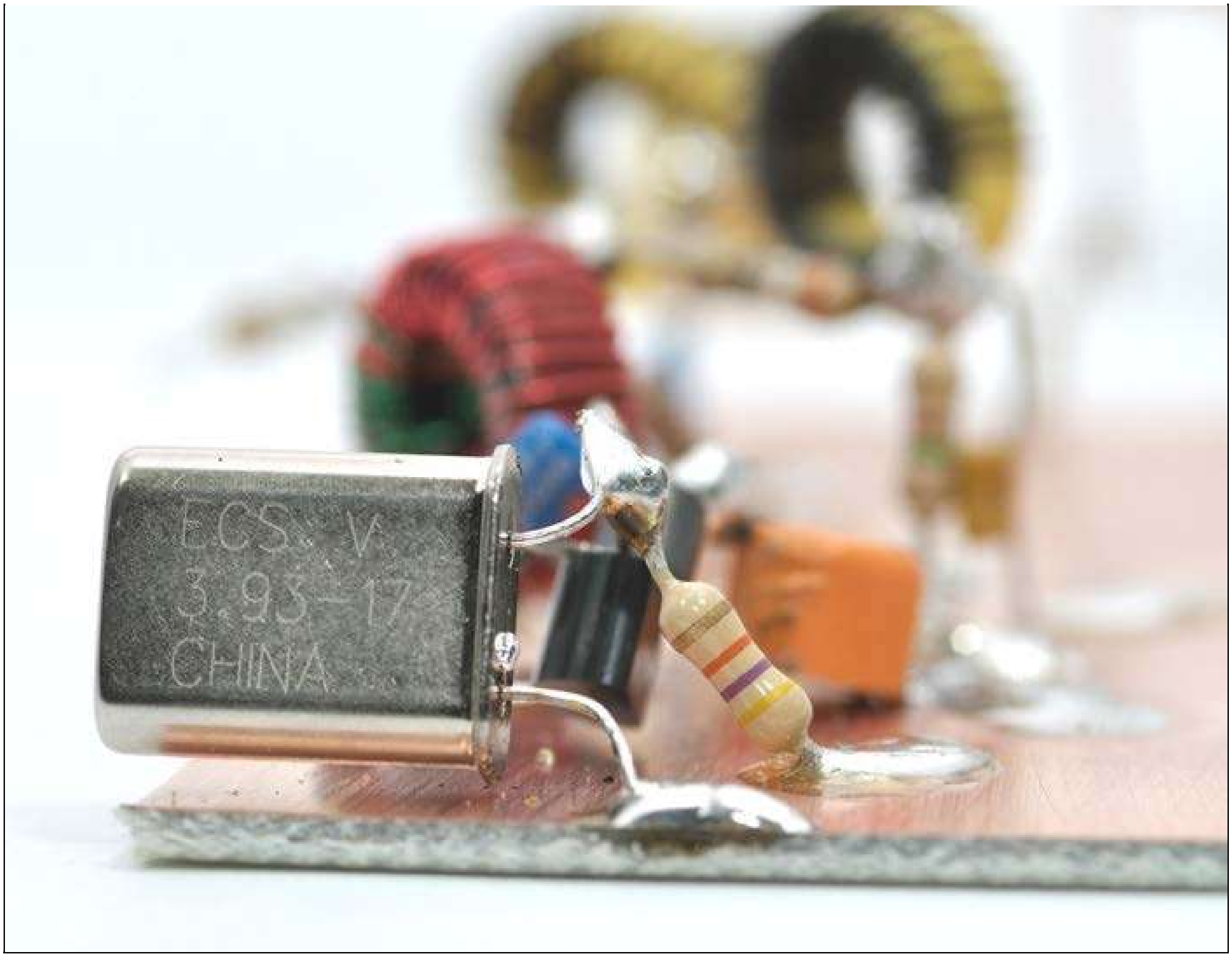

Above —The LO breadboard in close-up using a long focal length lens. Click for a wide angle photograph. The unsoldered end of the 100 Ω resistor in the close up photograph is where I connected the VCC.

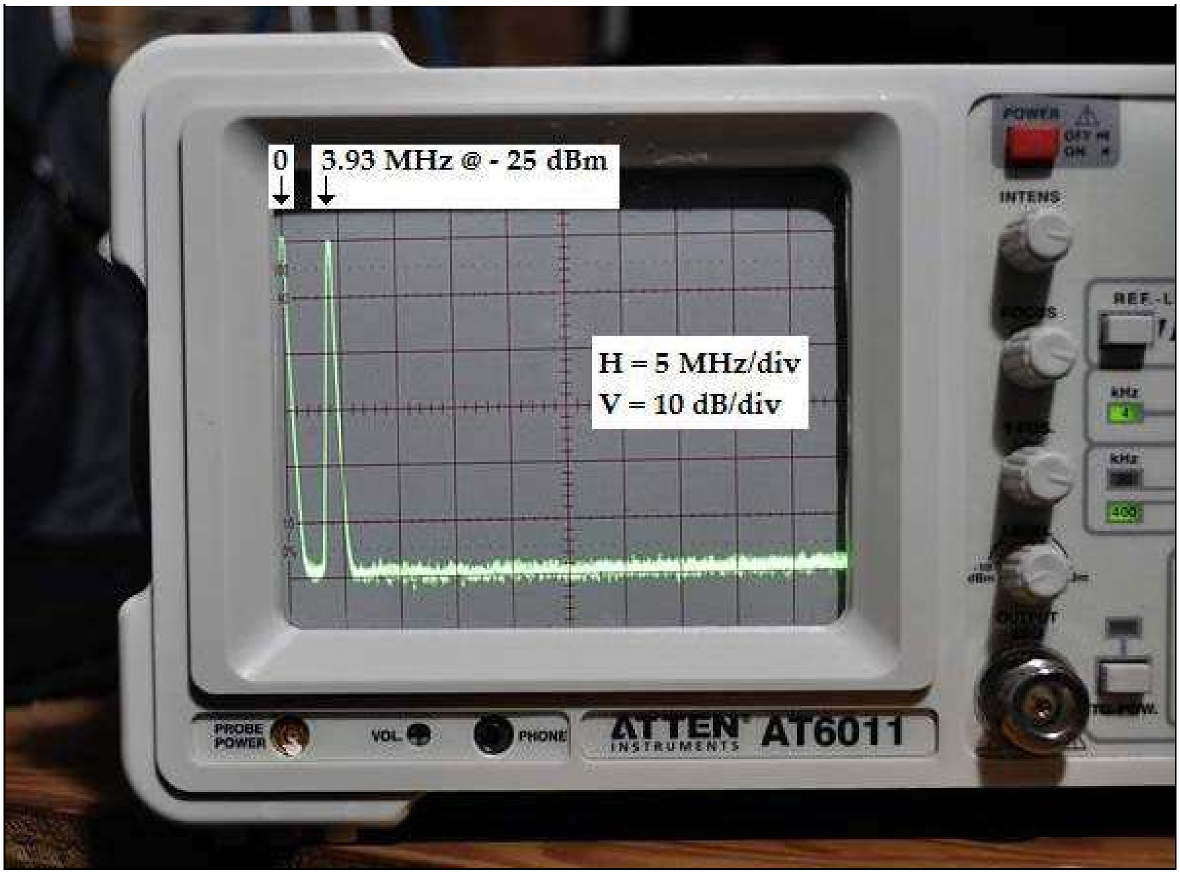
0
3.93 MHz @ - 25 dBm
H = 5 MHz/div
V = 10 dB/div
ATTEN AT6011
INSTRUMENTS
PROBE POWER
VOL.
PHONE
POWER
INTENS

Above — Spectrum analysis of the well-filtered 3.93 MHz LO. Not surprising, no harmonic energy is seen.

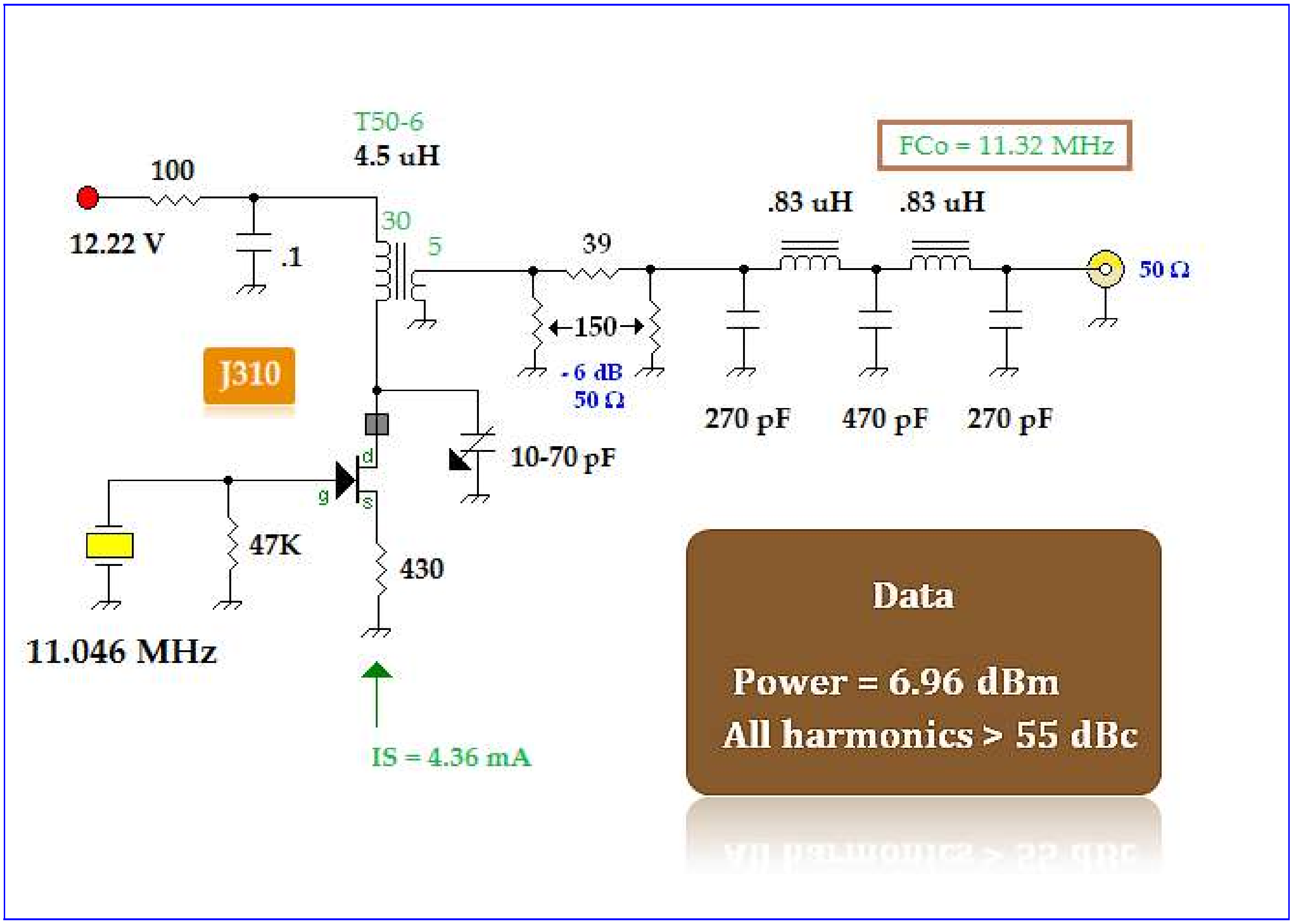
T50-6
4.5 uH
100
12.22 V
.1
30
5
39
←150→
- 6 dB
50 Ω
.83 uH
.83 uH
FCo = 11.32 MHz
50 Ω
270 pF
470 pF
270 pF
J310
d
g
s
10-70 pF
47K
430
11.046 MHz
IS = 4.36 mA
Data
Power = 6.96 dBm
All harmonics > 55 dBc

Above — The redesigned RF port oscillator. Clearly, the 7 element Chebychev low-pass filter isn't needed, so an N = 5 version was tried. Click for the spectrum analysis — again no harmonic energy was seen. In both the RF and LO signal generators, I tried to get as close to 7 dBm output power as possible.

To operate this oscillator at the desired RF port signal level; for example, between 0 and -10 dBm, you might just attenuate the output with a fixed pad or step attenuator. My conclusions echo the work of others experimenters; lowering the RF port down from 7 dBm to as low as - 10 dBm, lowered the amplitude of the spurious mixer products seen in the spectrum analyzer. Click for a sample. Refer to the QST Technical Correspondence citation in the references section for more information.

Above — The 2 re-designed oscillators wired up and connected to an SBL-1 mixer. I actually connected the attenuator pads after the low-pass filters as explained later.

Two different crystal oscillators were then built:

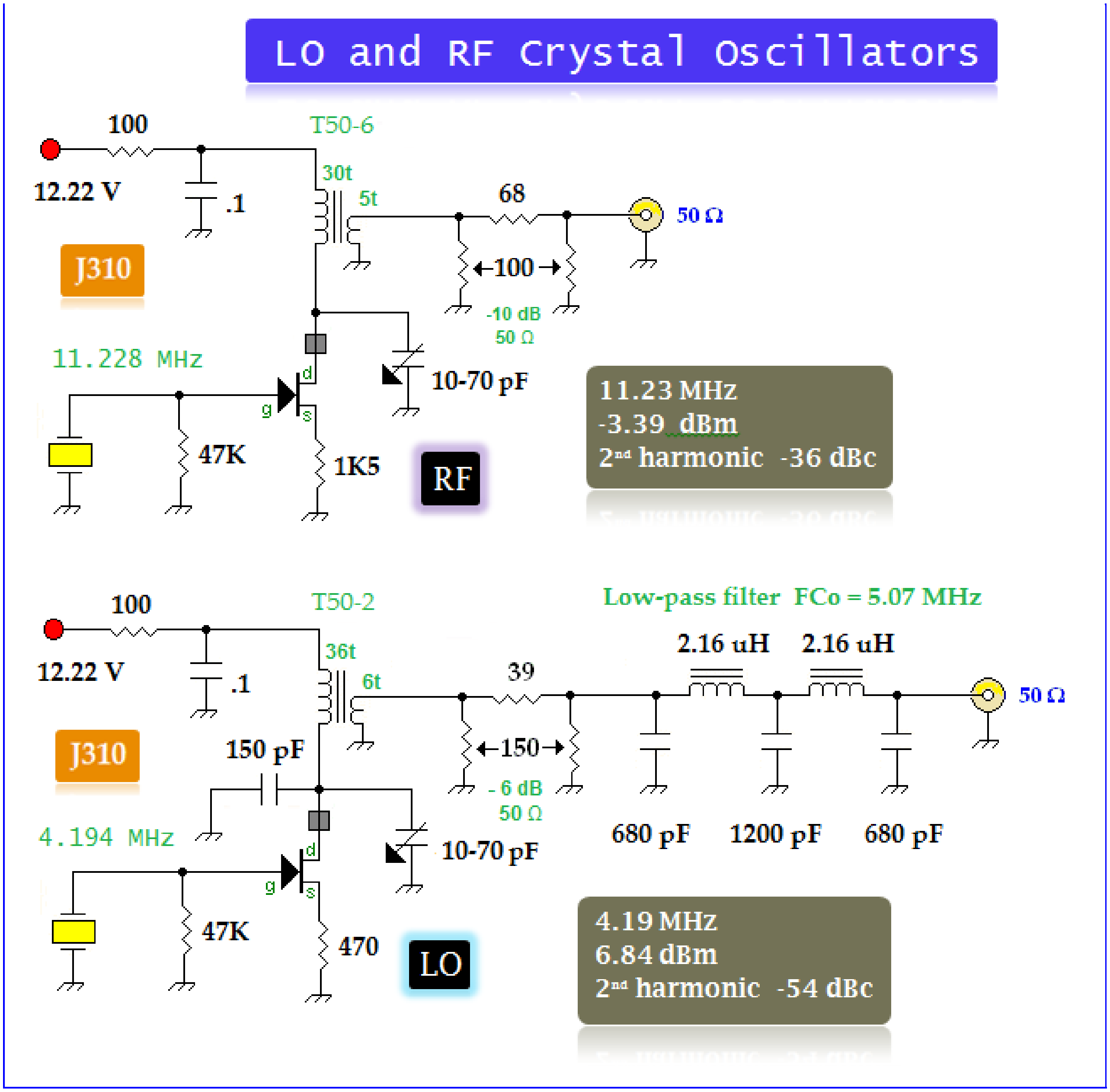
LO and RF Crystal Oscillators
100
T50-6
12.22 V
.1
30t
5t
68
50 Ω
J310
←100→
-10 dB
50 Ω
11.228 MHz
d
g
s
10-70 pF
47K
1K5
RF
11.23 MHz
-3.39 dBm
2nd harmonic -36 dBc
100
T50-2
Low-pass filter FCo = 5.07 MHz
2.16 uH
2.16 uH
12.22 V
.1
36t
6t
39
50 Ω
J310
150 pF
←150→
- 6 dB
50 Ω
4.194 MHz
d
g
s
10-70 pF
680 pF
1200 pF
680 pF
47K
470
LO
4.19 MHz
6.84 dBm
2nd harmonic -54 dBc

Above — Two different crystal oscillators targeting ~ 7.034 MHz were built. Click for the breadboard photo. You can see the crystal frequencies in this photograph. The RF port oscillator power was set to -3.39 dBm by choosing a low value JFET source resistor and attaching a 10 dB attenuator pad. Relatively low harmonic distortion prompted the exclusion of a low-pass filter on the RF oscillator. The LO output power was ~ 7 dBm.

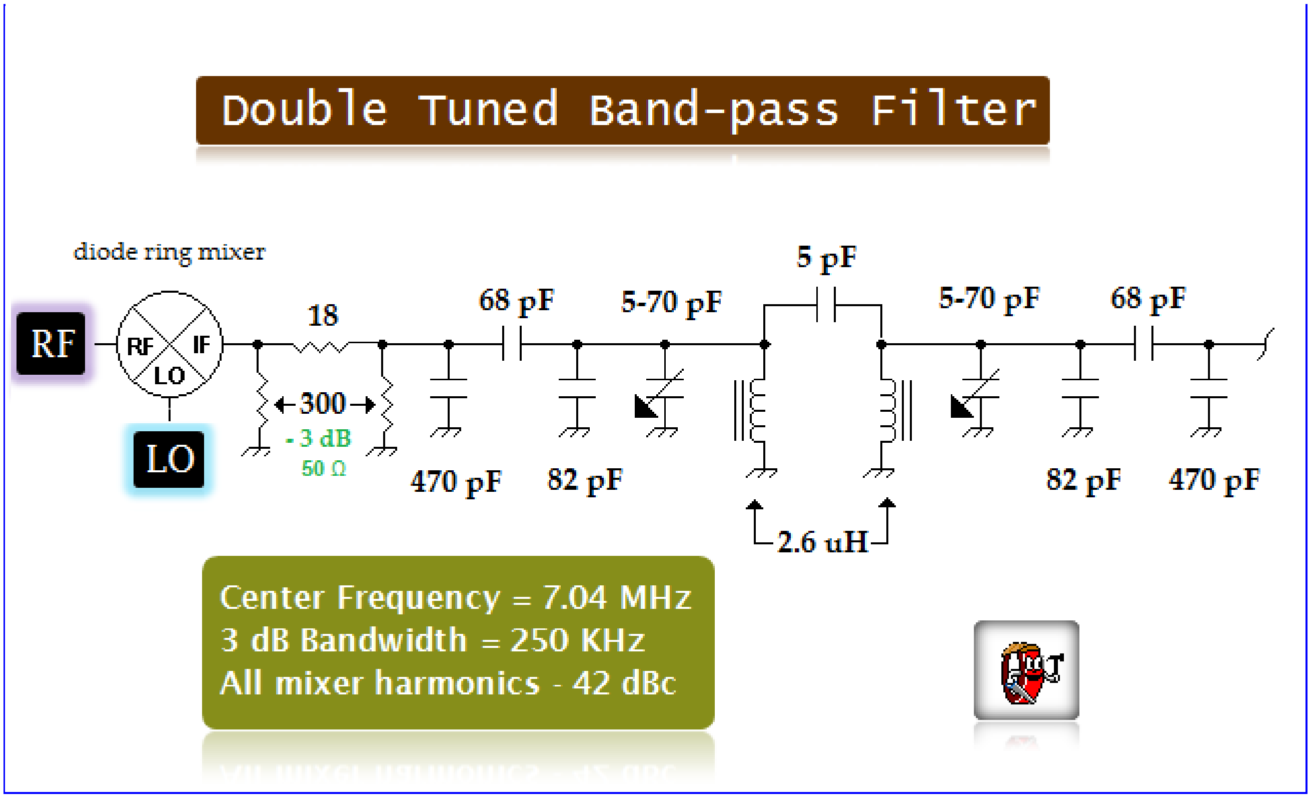

Above — The final experiment; placing a double tuned band-pass filter after the TUFF-1 diode ring mixer with the 2 latest crystal oscillators attached . This filter was in my junk box and I peaked it for the 7.034 MHz IF with a spectrum analyzer. The strongest spur was 42 dB down from the carrier — falling well short of the triple tuned band-pass filter presented before.

Clearly from all these experiments, a strong case for placing a triple tuned band-pass filter after a transmit mixer exists. If you use an unbalanced or single-balance mixer, a double balanced mixer might sufficiently not block feed through and spurious RF to keep your signal chain tidy. I enjoy studying the transmit chains of others to see how they filtered spurious and feed through RF. At the end of the day, as long as the output carrier spurs are low enough to meet your country's regulatory requirements, you're okay. Designing for low spurious emissions is an exciting challenge — one you'll miss if you don't try your hand with RF design.

A realization emerged following these experiments — I couldn't measure the return loss of the local oscillators! It technically could be done, but not by me. After 2 weeks of struggling, I engaged an American colleague with whom I occasionally build experimental circuits. After making some progress, we became stalled again. This time I asked Professor Ken Kuhn and Wes, W7ZOI for some ideas. Eventually a method to measure the RL of local oscillators came together along with enough material for another web page — RF Workbench 3.

When you do experiments, knowledge evolves as you go — for me, I learn mostly from making mistakes. I often think I should repeat most of my experiments over before presenting them, but this would consume too much time. However, footnotes can serve to steer readers for minor issues. If I had to re-build the crystal oscillators from Part 1, I'd build each crystal oscillator with a separate JFET buffer — then the return loss of the oscillator buffers could be measured as shown on RF Workbench 3 (with a shorted oscillator tank). Also, the pi attenuator pads on the crystal oscillators should follow the low-pass filters to garner the best output return loss. The good news is past experiments inform future experiments.

---

2. Bipolar Transistor Feedback Amplifier Experiments

I love making signals bigger — especially while preserving fidelity. It would be nice to become a reasonably competent amplifier designer — hopefully by studying sound schematics, applying software, building circuits and measuring evermore parameters this might occur. The mathematical equations of RF amplifier design seem quite daunting; they're the fodder of electrical engineers with their Hewlett Packard scientific calculators, SPICE software and GHz F-t transistors. With most things technical, as you try to advance, more questions than answers cross your mind; however, somehow this is normal and may actually signal progress.

Abundant amplifier references exist; for example, EMRDF Chapter 2: Feedback Amplifiers. This is essential reading and I won't repeat this information. Rather, I'll just share some ideas developed or reinforced on my bench. In the past, I've preferred amplifiers with narrow-band (tuned circuits) in an attempt to reduce distortion and maximize gain. Now after critically examining these tuned amps with scope and spectrum analyzer, I better appreciate the significant intrinsic feedback of RF transistors ***(the tendency to oscillate)*** and broadband designs are sought. Often you'll spend more time taming a tuned amplifier than building one.

This section focuses on return loss, bias techniques and achieving linear amplification — for example; finding ways to apply negative feedback, match the input, or how to set the collector voltage. All my experiments and thoughts about RF amplifiers are from an amateur designers' perspective and I welcome your feedback. The first amplifier shown is a classic W7ZOI topology that I call the "Beaverton Special".

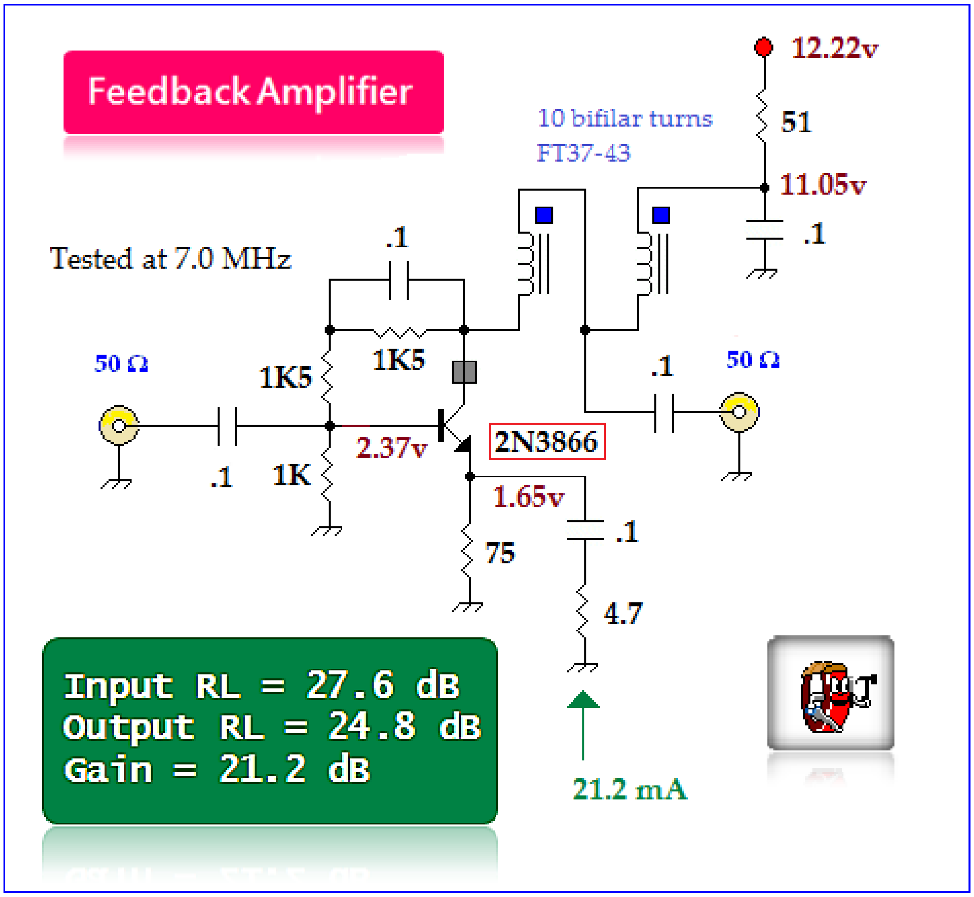

Above — A classic feedback amplifier popularized by Wes, W7ZOI in books like Solid State Design for the Radio Amateur and EMRFD. My respect for this humble design increased after building and testing 4 different versions to get a feel for amps with both shunt and series feedback. Of the 4 built, this particular version became my favorite — providing excellent input and output matching without crazy high emitter current. Employing a low noise / high F-t 2N3866 transistor is icing on the cake — an attempt to maximize impedance matching and performance using this standard, fits most transistors bias/feedback circuit. The humble 2N3904 also worked well in this slot. You don't need the ferrite collector bead with a 2N3904.

Other good experiments include trying different transistors and/or increasing the emitter current while being careful not to exceed the BJT's current rating (plus add heat sinking as required). You might also try the stage at different frequencies or perhaps sweep it to see at what frequency the gain starts to fall off.

Of the 4 BJT feedback amps shown in part 2, only this amp has a true broad-band input and output. What bothers me about broad-band linear amplifiers is that when you chain up 2 or more of them, signal fidelity generally degrades as it passes through each successive amp stage. Solutions include mopping things up with some low-pass filtering after the last stage, leaving it alone, or tuning the amplifiers (i.e. not using broadband stages).

The biggest caveat for feedback amps are variations in input and output impedance caused by source and load mismatches. For example, a 75 Ω resistor was connected to the output of the FBA above. The input return loss degraded to 16.8 dB. Further, the same 75 Ω load was removed and then connected to the input during output return loss measurement— this degraded the output RL to 23.4 dB. Clearly *load* mismatches upset return loss more than *source* mismatching. A 50 Ω attenuation pad should likely follow a feedback amp in situations where high input return loss are desired; for example, after a diode ring mixer.

Noticing a variation of the classis feedback topology in EMRFD Figure 6.140, I asked Wes, W7ZOI about it. It turns out there's another way to "skin the shunt feedback cat". The above RF amp uses a series connection of 2 feedback resistors (1K5 and 1K5 with a bypass cap across one 1K5). The result is a resistance at DC of 3K, but a resistance at RF of just 1K5. You could also use a 3K resistor directly from collector to base that is paralleled by a series connection of a 1K5 resistor plus a 0.1 uF capacitor. That network has the same impedance as my amp shown above. That is; the resistance would be 3K at DC, but 1K5 at RF.

This explanation fueled the next experiment — transistor amplifiers have 2 operating conditions; 1 at DC, the other at AC. Like a carpenter framing a house, you begin design by setting the DC bias — no small design task since bias concerns more than just establishing the base voltage and emitter current. For example, biasing may effect voltage gain, maximum signal handling capability, noise figure, impedance matching, class of operation, the operating point (sometimes called quiescent point or q-point), feedback and temperature stability. Biasing provides much to think about, however, a practical way to explore any topic is to chunk it into small, understandable pieces that become a stepping stone to advancement. Let's focus on biasing for temperature stability. The next amp uses the wrap-around PNP bias — an awesome technique.

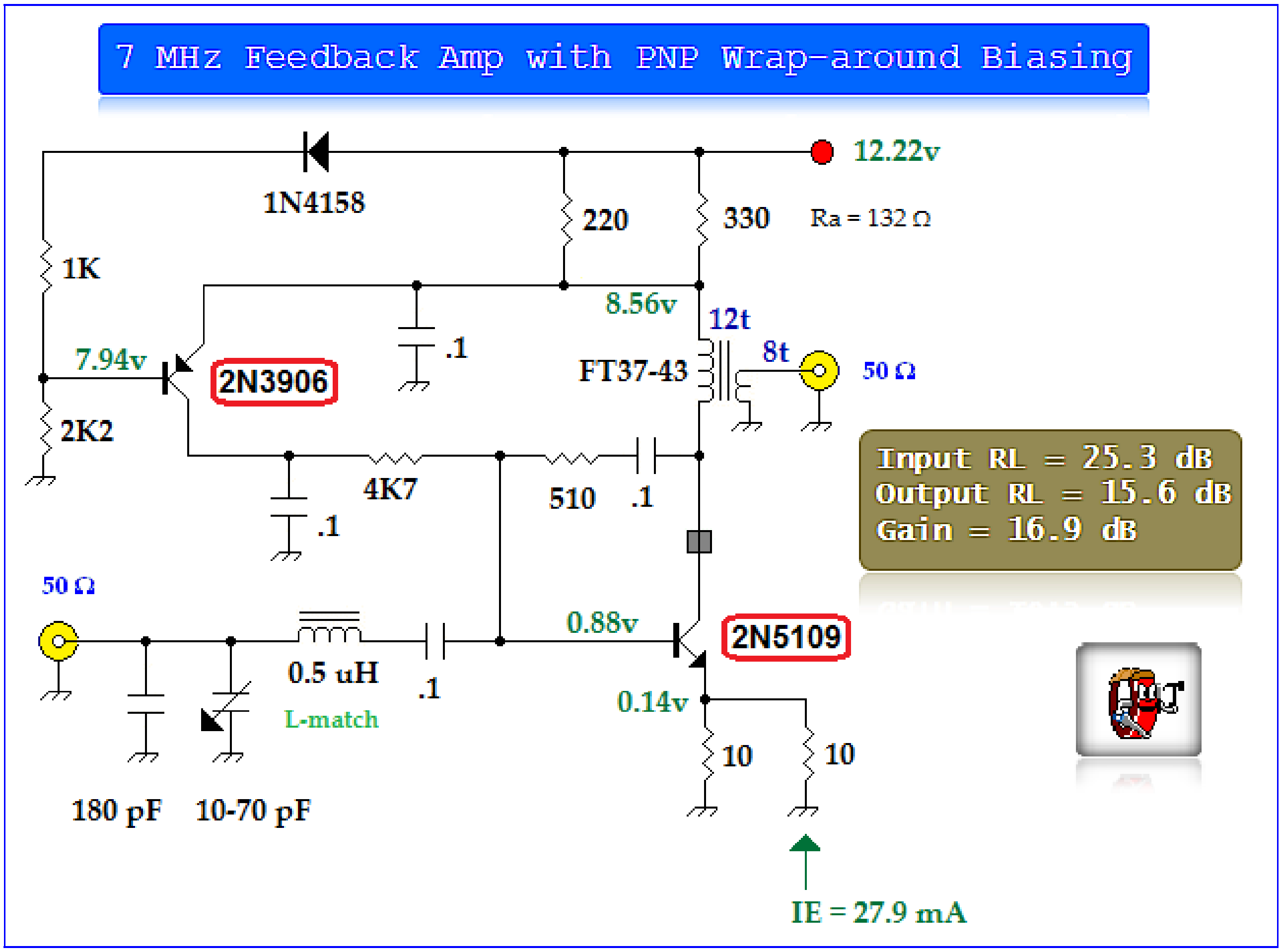

Above — A 7 MHz FBA using PNP wrap-around biasing. I learned about wrap-around biasing from Wes, W7ZOI and share a simple way for new builders to also learn this technique as the **#1 Design Center** on the supplemental web page. Click for a prototype breadboard photograph. This amplifier employs heavy shunt feedback from collector to base. Degenerative (series) feedback from the 2 parallel 10 ohm resistors also enhances temperature stability.
Expanded bias circuit temperature stability discussion follows amplifier number 4.

The wrap-around or feedback bias scheme is good because it's self stabilizing. The diode in the PNP bias network further ensures that the PNP bias remains constant with temperature changes. It really should be to glued to the NPN transistor (or its heat sink) to allow tracking of the NPN's temperature variations. This bias circuit doesn't load the NPN base input impedance. Another great virtue is that the emitter of the amplifying transistor can be directly connected to ground allowing better performance at VHF and UHF.

Noise from the PNP will be amplified by the NPN, so the low-pass network formed by the 0.1 uF capacitor and 4K7 resistor is essential. In some related circuits, you may see an RF choke used instead of a resistor.

The actual 2N5109 input impedance is probably around 40 ohms — easy matching with an L-match network.

Amplifiers with an L-match tuned input shouldn't follow a diode ring mixer unless preceded with a diplexer since the narrow-band L-match tunes only 1 frequency. L-match networks can make an impedance bigger or smaller depending how they're oriented and also provide some low or high-pass filtering depending on the configuration. I design my L-networks on the bench using experience plus trial and error — a better way is to use software. I recommend the program called Zmat08.exe that is included on the CD that accompanies EMRFD. The software will get you close, however, bench tweaking is required since you're often matching a complex impedance, comprised in part, of stray reactance.

Setting up an L-Network for an Input Match

A suggested bench method for optimizing input Return Loss (RL) using an L-Match network.

Your task is find the "perfect" L and C values to get a RL of 20 dB or higher. Start by soldering in an inductor calculated from Zmat08.exe or according to your wisdom. Set up the amplifier for input return loss measurement. The first chore is to find the nominal target capacitance that provides the best match at the design frequency. I use a big range, air variable capacitor for this — with the input circuit connected to a return loss bridge, connect up and tune the big variable capacitor to give the greatest RL. Remove the variable capacitor, measure it, and then solder in an equivalent trimmer capacitor, and as required, fixed capacitor(s) so you can tune at least 25 pF above and below the target capacitance. Often, the target C will be close to whatever the software recommends. In amplifier 2, my C values are the 180 pF + a 10-70 pF trimmer.

Next, determine the optimal inductor. On my bench, I keep a variety of pre-wound #6; and #2 material powdered-iron toroid inductors and choose one close to the calculated or a self-chosen L value. I start with an inductor wound with 4-5 more turns than needed. After soldering it in, the RL is checked. Remove 1 or 2 turns, tweak the trimmer capacitor and again check the RL. If after removing 1 or 2 turns, the RL is going up, you've determined there was enough inductance to get the best RL. *(If the RL goes down, you probably didn't start with enough L to get the best possible RL).*

You can also also squeeze together or spread apart the toroid windings to vary inductance — the maximal inductance variation varies due to factors including wire gauge and total turns. Compressing the windings with thumb and forefinger *increases* the inductance and widening the gaps between windings *reduces* inductance on a toroid.

Assuming the RL increased after removing 1 or 2 turns, remove another turn, tweak the trimmer capacitor and check the return loss, and so on. Repeat until your return loss starts to decrease. Then add back a turn or 2 to find the absolute best match. *This procedure allows you to find the optimum inductance in-situ.* Once, you've figured out the best inductance, cut the inductor leads short, solder it in, tweak the trimmer capacitor, and then consider further tweaking the coil by expanding or squishing the windings on the toroid while looking at the RL in a bridge detector.

In summary, to get the best possible RL — design a prospective L-match with software, and then bench test to determine the optimal in-situ L and C by using values above and below the calculated L and C values while observing the results in a return loss bridge. This method seems tedious, but emphasizes that repeated bench practice and patience pays off. You can always just use the calculated L-network values and/or develop your own method to set them up.

Consider mitigating the stray inductance caused by the long lead that occurs after removing wire turns by cutting the lead and scraping off the enamel insulation every couple of turns or so. This is a gamble — If you cut the lead and need to add back a turn, you'll have to rewind the coil from scratch, or add in and solder another turn (messy). I'm often able to get a L-network RL of 22-26 dB using my method and feel it's worth the the time and effort.

When bench tweaking the L and C values, your actually looking at the peak-to-peak AC voltage with the amp input connected to the unknown port of the RL bridge. Tune for the lowest, stable peak-to-peak voltage. Test it against the open circuit peak-to-peak voltage to calculate the RL. Since the open circuit doesn't change, you know the return loss is improving when the peak-to-peak voltage of the amplifier under test is going *down*. I store the open circuit voltage in my scientific calculator and calculate the RL from time to time as I'm tweaking the L and C values. After awhile, the whole procedure becomes automatic and quick. Once you 're done and everything's tidy, measure the open circuit and connected amp peak-to-peak voltages and calculate RL a final time. This is your reportable return loss.

You can scale matching networks from other builder's schematics by calculating the XL and XL and then applying these reactances to your desired frequency. Bench tweaking is still required. I also hope the person whom I'm copying didn't make a bench or drafting error. Be discerning about whatever your find on the Internet "Misinformation Highway" — this site included. Although I'm no philosopher, I know at least 3 things about people. They: 1. are often biased; 2. can lie; and 3. can make errors.

Above —Some toroids and the *air variable capacitor* I sometimes use to **coarsely** bench tune L-C circuits to determine the "ballpark" tuning capacitance. This capacitor features built-in reduction drive and varies from 15pF to 428 pF. When using an external capacitor connected to your circuit with short copper wires, expect some signal distortion and watch out for hand and body caused capacitance variations. The connecting wires also have reactance which won't be there when you swap in a small trimmer cap plus any fixed value capacitors.

Next up is a common emitter amp using "noiseless" feedback - this means the AC feedback is achieved with transformers instead of "noisy" resistors.

3

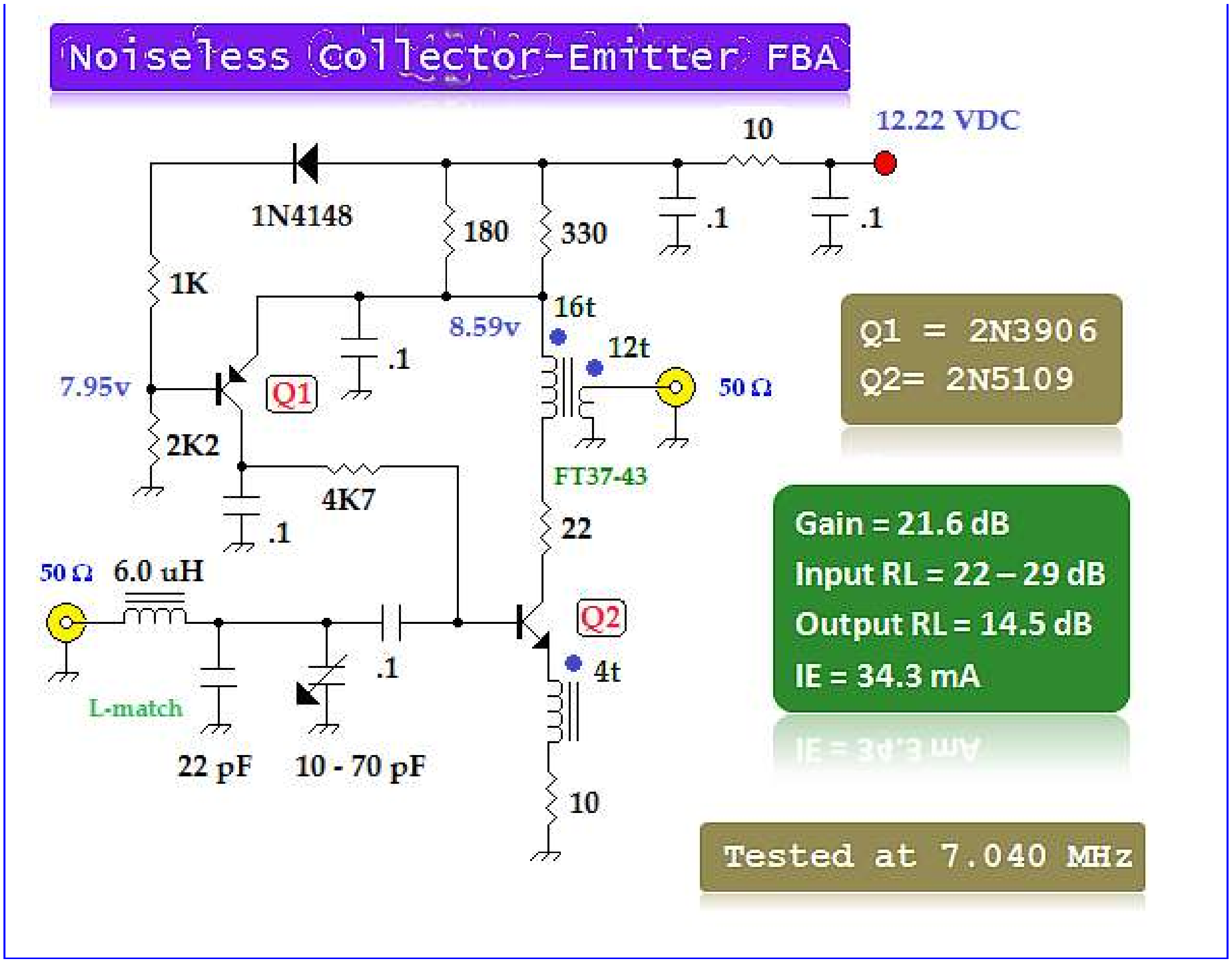
Noiseless Collector-Emitter FBA
10
12.22 VDC
1N4148
180
330
.1
.1
1K
16t
8.59v
12t
.1
7.95v
Q1
50 Ω
2K2
FT37-43
4K7
.1
22
50 Ω
6.0 uH
Q2
.1
4t
L-match
22 pF
10 - 70 pF
10
Q1 = 2N3906
Q2= 2N5109
Gain = 21.6 dB
Input RL = 22 – 29 dB
Output RL = 14.5 dB
IE = 34.3 mA
Tested at 7.040 MHz

Above — Schematic of a 7 MHz collector-emitter "Griffiths" feedback amp. I ran substantial emitter current through this NPN. RC = 116.5 ohms — I paralleled 2 resistors for RC because I lack resistors between 100 and 150 ohms. The basic design is by Bruce Griffiths, who has a great web site. I thought I put up *big* schematics!

The input L-network was designed on the bench and provides a good input match peaked at 7.040 MHz — this pumped up the gain 3-4 dB. In my amp, a T50-2 powdered iron toroid inductor forms the L-match coil. Matching for the best possible input return loss is touchy and best done on the bench. For example, if the 6 uH inductor is decreased to 5.8 uH, the match could fall by 2-4 dB. With patience and careful tweaking return losses approaching 29 dB are possible, but likely too time consuming for most builders. The procedure as described earlier is pragmatic: connect a RL bridge to the input and adjust the L and C values until the lowest return loss is discovered. Even squishing or expanding the toroidal inductor windings can squeeze out a final dB or so of input matching.

Output matching proved interesting. Although I tried, the best output RL I could muster was 14.5 dB. Lowering the 10 Ω degeneration resistor or increasing the current could increase the output return loss. An output attenuator pad might be considered — a 3 - 6 dB pad would increase the output RL to over 20 dB.

All 3 output transformer windings were wound on a FT37-43 with care to keep the phasing correct. Amplifier gain is not dependent on collector current. For example, substituting an Ra of 180 ohms (clipping out the 330 Ω resistor) yielded a gain of 19.5 dB, an emitter current of ~ 20 mA and an output return loss of 12.6 dB, while the input match changed very little.

The oscillation snuffer 22 Ω collector resistor was 15 ohms in another version, however, parasitic oscillations were discovered at ~175 MHz and snuffed out by raising this resistor from 15 to 22 ohms. I sometimes go as high as 51 Ω; especially in JFET circuits.

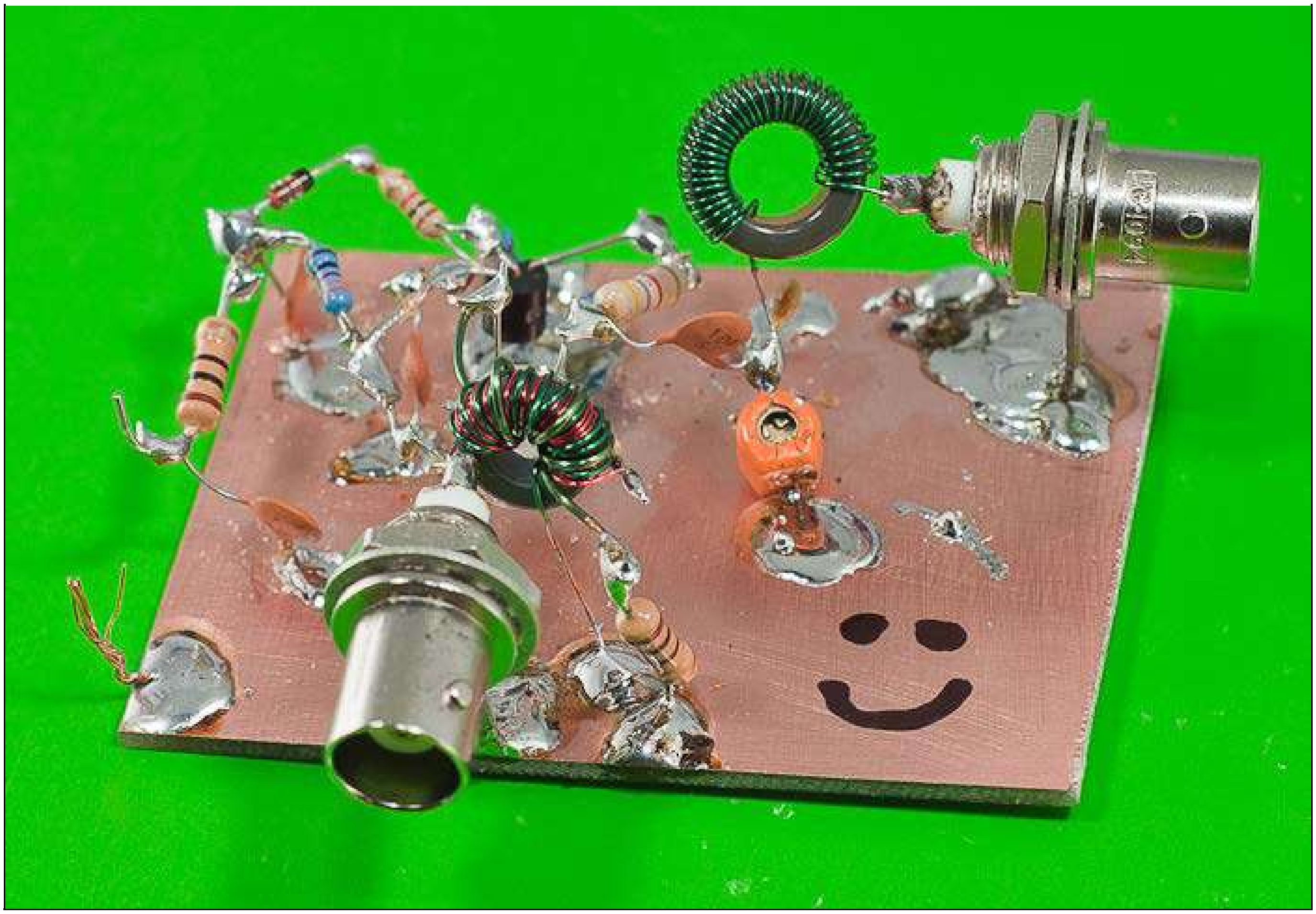

Above — The breadboard of the noiseless collector-emitter 7 MHz feedback amp. Click for a photo of another version. The hot "modern" replacement for the 2N5109 is this SMT part. I also like the BFG135 T/R BJT.

The final FBA experiments *below* use a standard voltage divider bias, tweaked for temperature compensation. The AC feedback is base to emitter — a rarely used topology in North America; although I'm not sure why.

4

Base-Emitter FBA -- DC Bias Using a 2N2222a

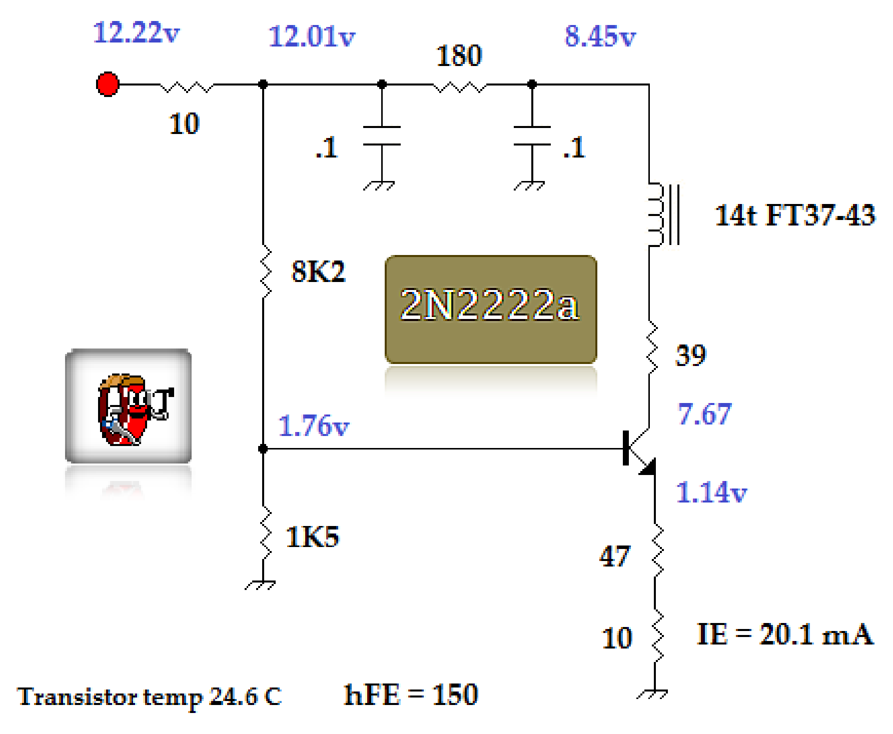
12.22v
12.01v
180
8.45v
10
.1
.1
14t FT37-43
8K2
2N2222a
39
7.67
1.76v
1.14v
1K5
47
10
IE = 20.1 mA
Transistor temp 24.6 C
hFE = 150

Above — The DC bias resistor values for a 2N2222a with a DC Beta or hFE of 150 and a emitter current of 20.1 mA. Almost every text author writes about voltage divider bias temperature stability, but some builders get bogged down in the details. Since the bipolar junction transistor is a voltage controlled device *(see section 4: QRP-POSDATA for an explanation)*, you must set up some DC voltages — I created a design center presenting an easy approach to design reasonably temperature stable BJT amps. See **#5 Design Center** on the supplemental web page. After getting the bias, the AC parts were added, and the completed schematic is shown below.

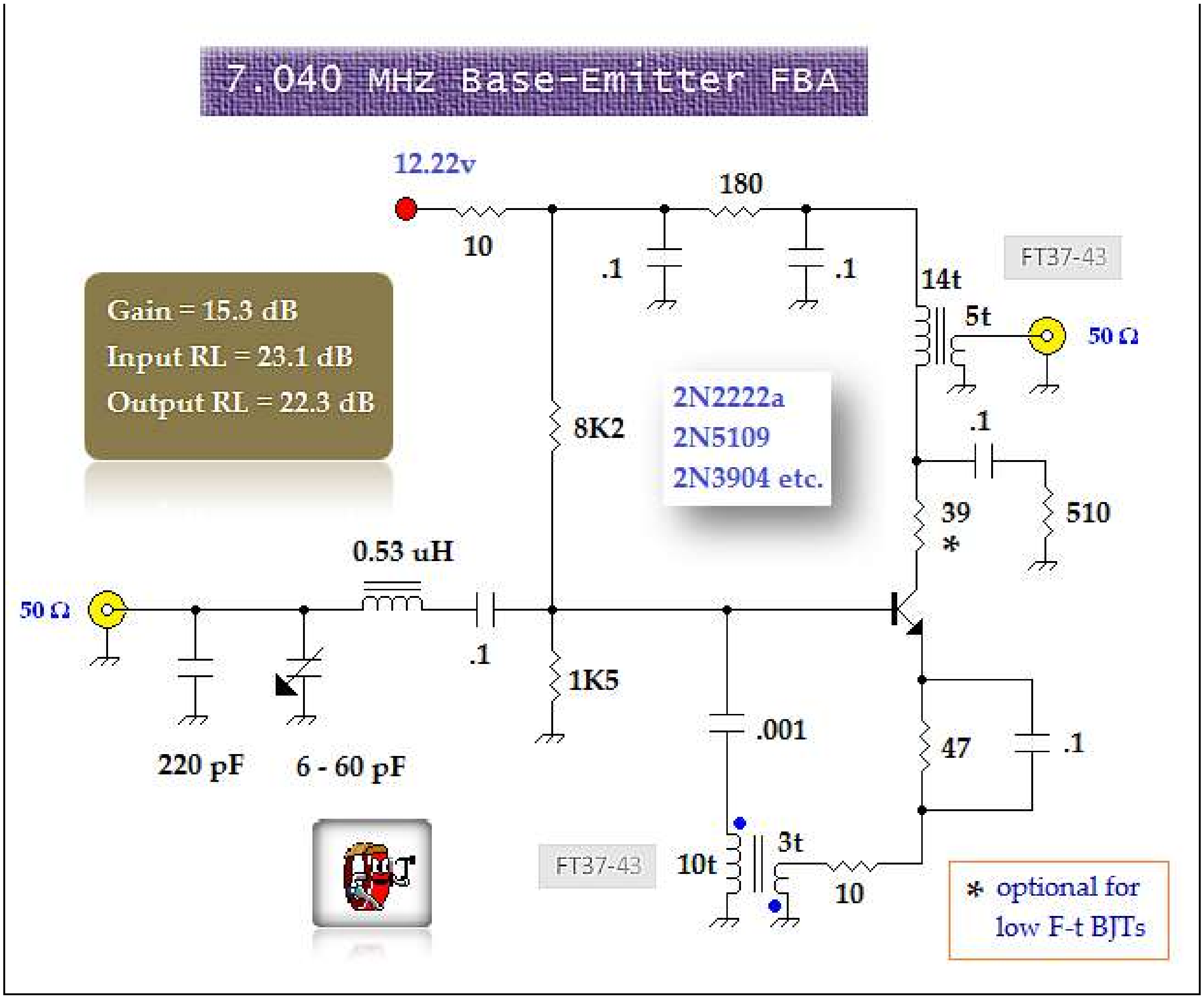

Above — A base-emitter feedback amp built Dec 21, 2010. I read about base-emitter feedback in Dr. Rohde's book *(Reference 1)*. He had some discussion, a small signal model and lots of difficult math, but no circuit examples. After searching on the web I found 1 example in the HBR-2000 transceiver; a project designed and built my respected Canadian colleague Marcus, VE7CA. Click for his web site. I decided to build my own design using a L-match to tune the input to 50 Ω.

The above amp was built around around a 2N2222a. The 39 ohm resistor is not really required with the 2N2222a. For high F-t transistors like the 2N3866, 2N5109 or microwave transistors, ferrite bead(s) or the resistor are *not* an option. Low F-t transistors like the 2N2222a or 2N3904 don't need the UHF oscillation snuffer resistor since they lack real gain at these frequencies. With the design center, you should be able to bias your own amp according to the emitter current you want — choose a BJT, measure or choose its hFE and then choose IE.

Missing from this web page is how to choose an operating point + discussion about DC load lines and related topics. I may tackle these topics on a future web page. I'm not sure anyone cares about this anymore.

The most difficult part was the output transformer. Lacking a base to collector connection, the collector impedance runs quite high and finding a good match into 50 Ω proved impossible — even with a shunt resistor across the primary coil. I saw a strategy in Marcus' amp; AC couple the collector to ground via a 510 Ω resistor. I did this. From then on, it was just trial and error to identify the optimum turns ratio for the collector transformer. An interesting experiment might be to figure out the turns ratio using a lower loss output transformer such as a FT-37-61.

The turns ratio of the various collector and drain transformers on most of these amplifier designs were determined by placing the amp in an output RL measurement setup and adding or removing secondary turns to get the highest possible RL. See the procedures for RL measurement on the RF Workbench pages.

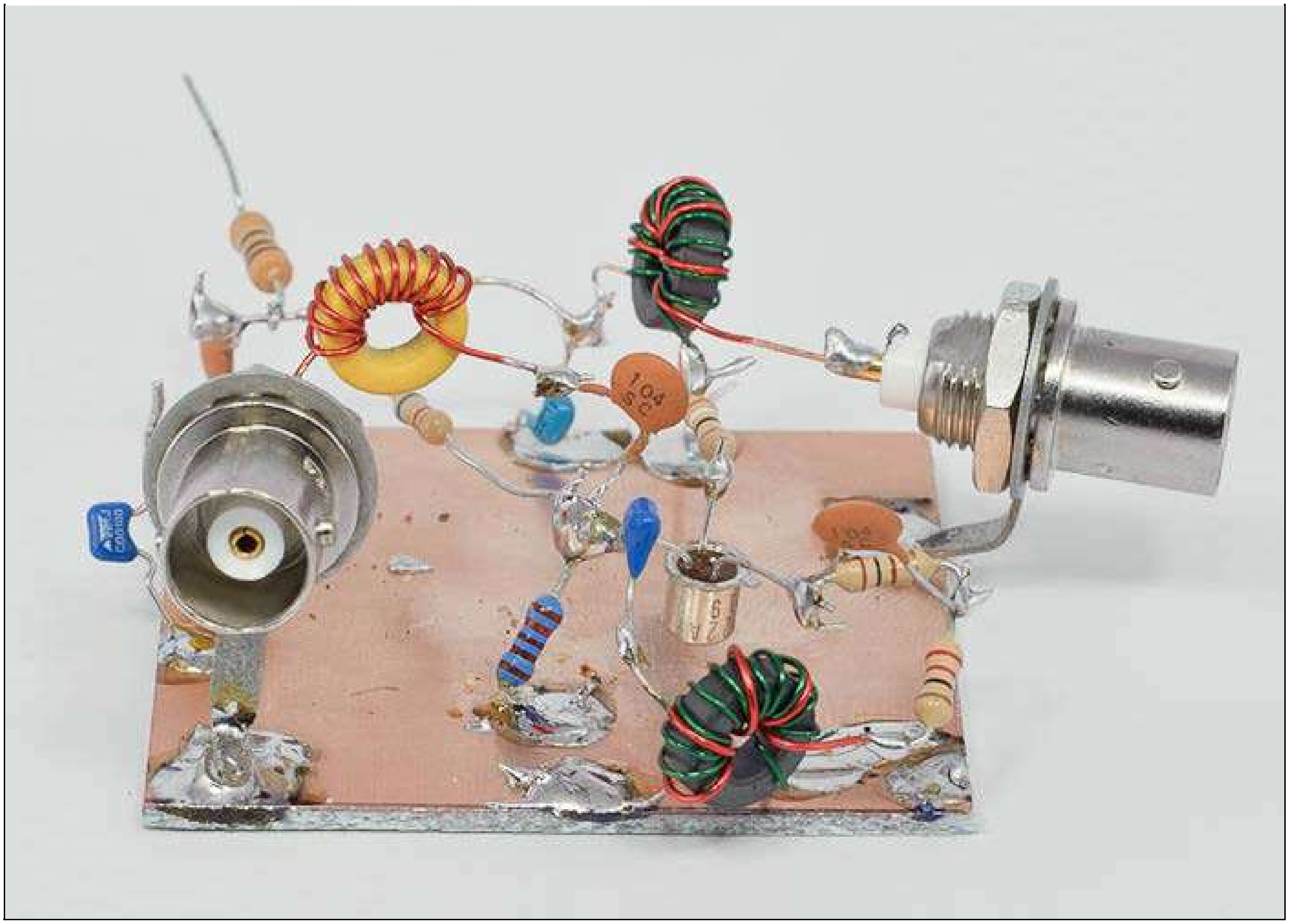

Above — The breadboard of the first version of the base-emitter FBA.

---

3. JFET Common Gate (CG)Transistor Amplifier Experiments

These experiments focus on setting up a desired input return loss and getting a reasonable output return loss in the CG amplifier. My expectation of an easy set of experiments proved wrong — assumptions never substitute for actually building and measuring.

I like motorcycles. The difference between riding a motorbike versus *driving* a car parallels learning on the bench versus learning by just simulating or calculating component "ideal" values on paper or computer. In the car you're isolated from wind, smells, temperature changes and subtle road traction and camber differences that you fully sense on the motorcycle. Bench experiments prove equally visceral and experiential — the sensory input from learning as you build and test circuits imprints deeply in your mind.

1

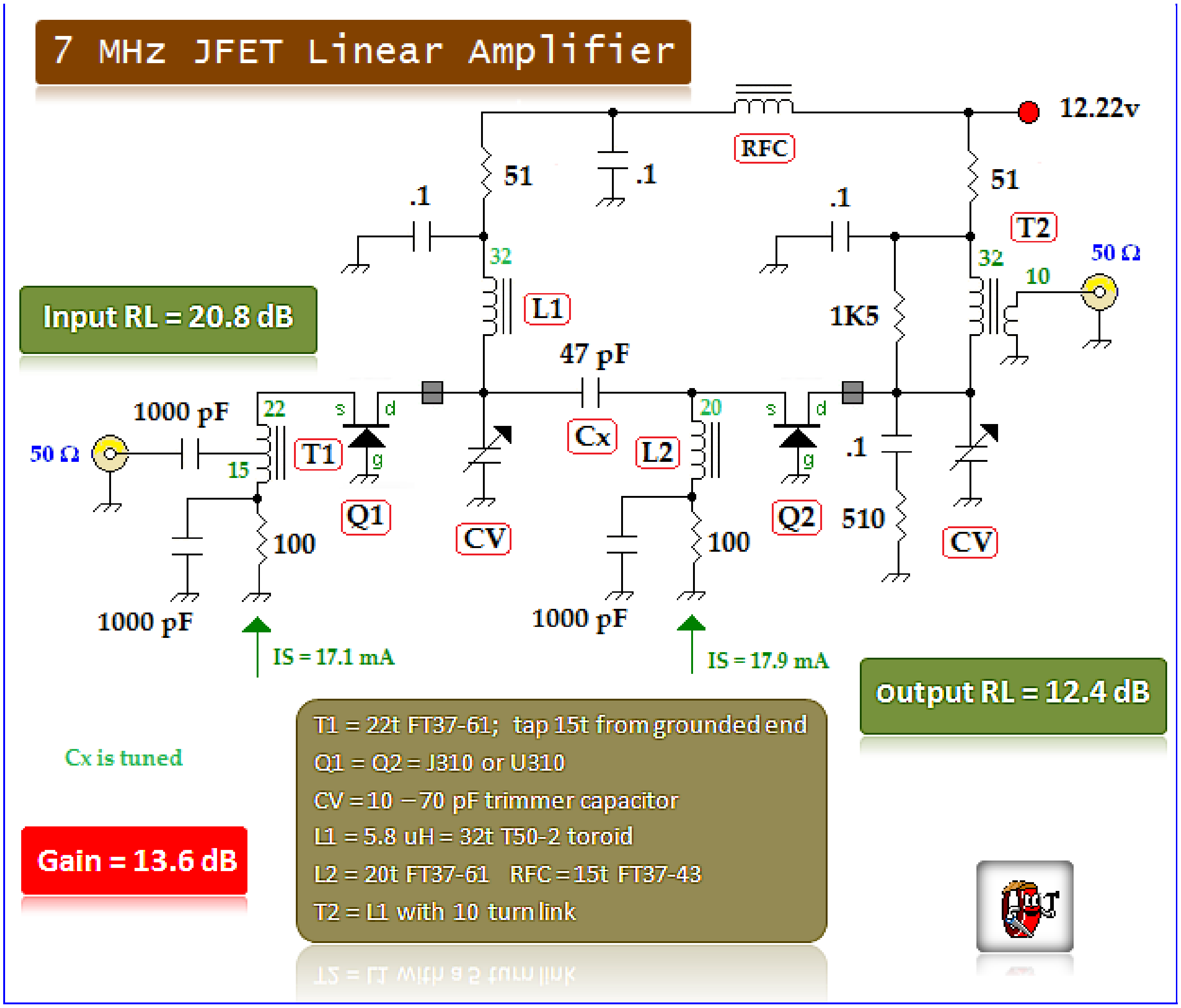
7 MHz JFET Linear Amplifier
12.22v
RFC
51
.1
.1
51
.1
T2
32
50 Ω
10
Input RL = 20.8 dB
32
L1
1K5
47 pF
1000 pF
22
s
d
20
s
d
50 Ω
T1
g
Cx
L2
g
.1
15
Q1
CV
100
Q2
510
CV
100
1000 pF
1000 pF
IS = 17.1 mA
IS = 17.9 mA
Output RL = 12.4 dB
Cx is tuned
T1 = 22t FT37-61; tap 15t from grounded end
Q1 = Q2 = J310 or U310
CV = 10 – 70 pF trimmer capacitor
L1 = 5.8 uH = 32t T50-2 toroid
L2 = 20t FT37-61 RFC = 15t FT37-43
T2 = L1 with 10 turn link
Gain = 13.6 dB

Above — A 7 MHz JFET "linear" amplifier built only for testing ideas — *do not build*. It went through several incarnations and prompted many experiments. The input return loss was deliberately set to 20.8 dB, although I set a RL from 10.0 - 28.6 dB during my experiments. Bypassing the JFET source resistor increases gain, but of course changes input RL.

The output transformer represents a terrible design, but shows the length I went to to try an obtain a decent output return loss. Working with this circuit, led me to abandon tuning the output transformer in situations where a high return loss was desired since the low value resistors required kill the tank Q significantly.

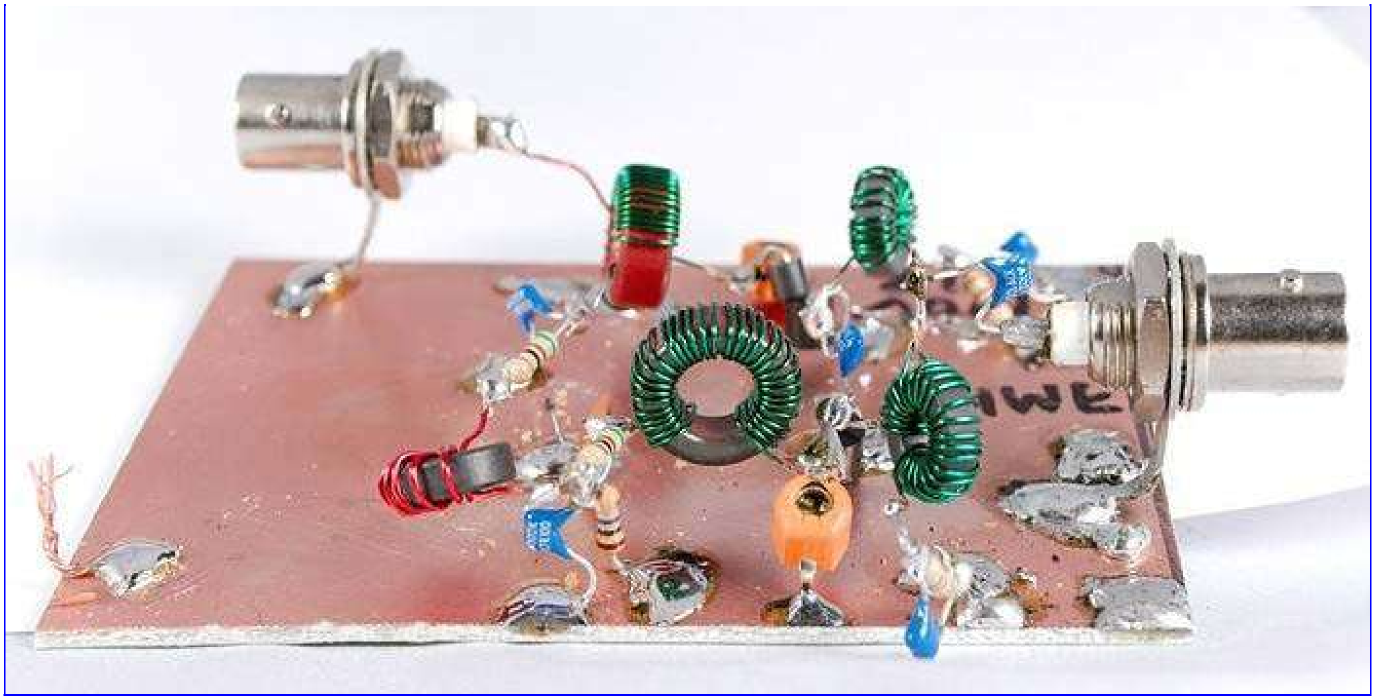

Above — Breadboard of 1 version of the prototype low-level JFET "linear" amplifiers for 7 MHz. Click Click . Cx is tuned with a variable cap and a nearest standard value substituted; in my case 46 pF was the measured value of the variable cap at point Cx.

Setting Input Impedance

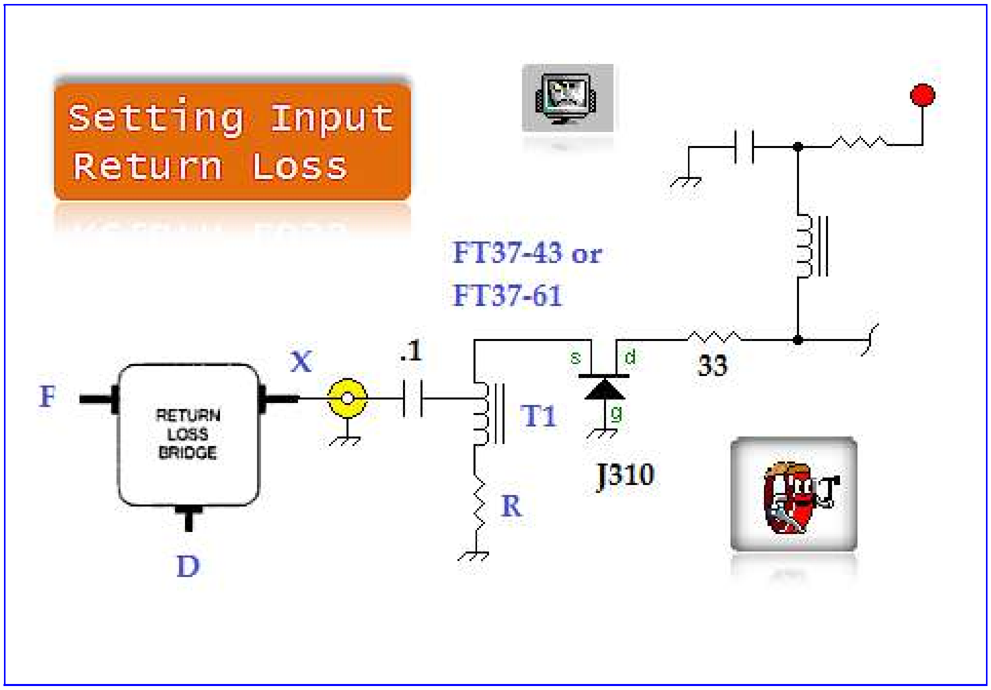

Above — The procedure used to set a desired CG amplifier input Return Loss. Numerous factors influence the input impedance and I discuss them in **#4 Some Factors Affecting Common-Gate Amplifier Input Impedance** on the supplemental web page. I keep some tapped inductors on my workbench such as this FT50-43 or these FT50-61 core inductors. To find the best return loss using such a coil, you can change tap points, remove windings and even wind more turns and solder the 1 end of your new windings to 1 end of the existing wire. Some builders omit inductor taps and manipulate the input return loss other ways as described in the supplemental article.

Normally we set the input match after establishing the output match since the output impedance dramatically affects the input impedance.

Further, you might notice that the tap point may vary between different JFETs. Most of my "real world" coils have at least 2 tap points and I choose the tap that gives the best return loss. More often than not, I bias for 14 - 18 mA and leave off the source bypass capacitor; it's your call.

The input return loss that gives the lowest noise figure is often chosen by engineers.

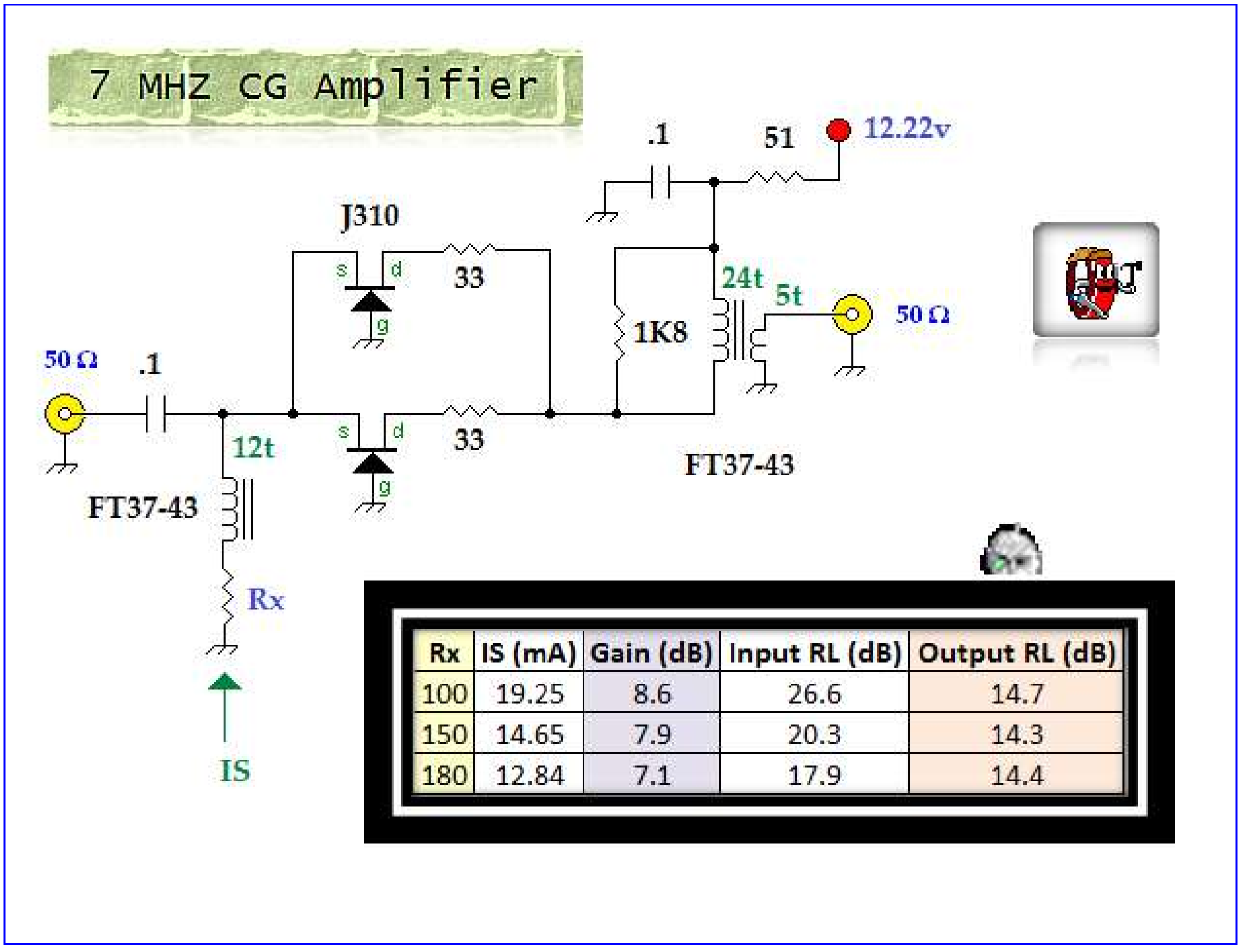

Rx	IS (mA)	Gain (dB)	Input RL (dB)	Output RL (dB)
100	19.25	8.6	26.6	14.7
150	14.65	7.9	20.3	14.3
180	12.84	7.1	17.9	14.4

Above — An experimental 7 MHz common gate amp designed to terminate a diode ring mixer. The best thing about using 2 JFETs is that you don't have to determine the tap point in the decoupling inductor (12 turns on a FT37-43 in this amplifier). I put up to 4 in parallel during my various experiments. It's faster to match just 2 JFETs, so 2 were favored.

The output RL wasn't great at ~ 14 dB, however is probably normal or better than most published amateur projects. I set the output match by adding a shunt 1K8 resistor across the primary winding and then finding the turns ratio to give the best output return loss. Without the resistor, the best output RL will be ~5 dB or worse. The resistor reduces power.

I learned that putting JFETs in parallel in a common gate amplifier reduced the output return loss in circuits using an output transformer like in the schematic above; this is unfortunate.

I wanted an output RL of 20 dB or greater — this is no small request; over a week was spent investigating transformer behavior and finding ways to improve output return loss when you really *want* to.

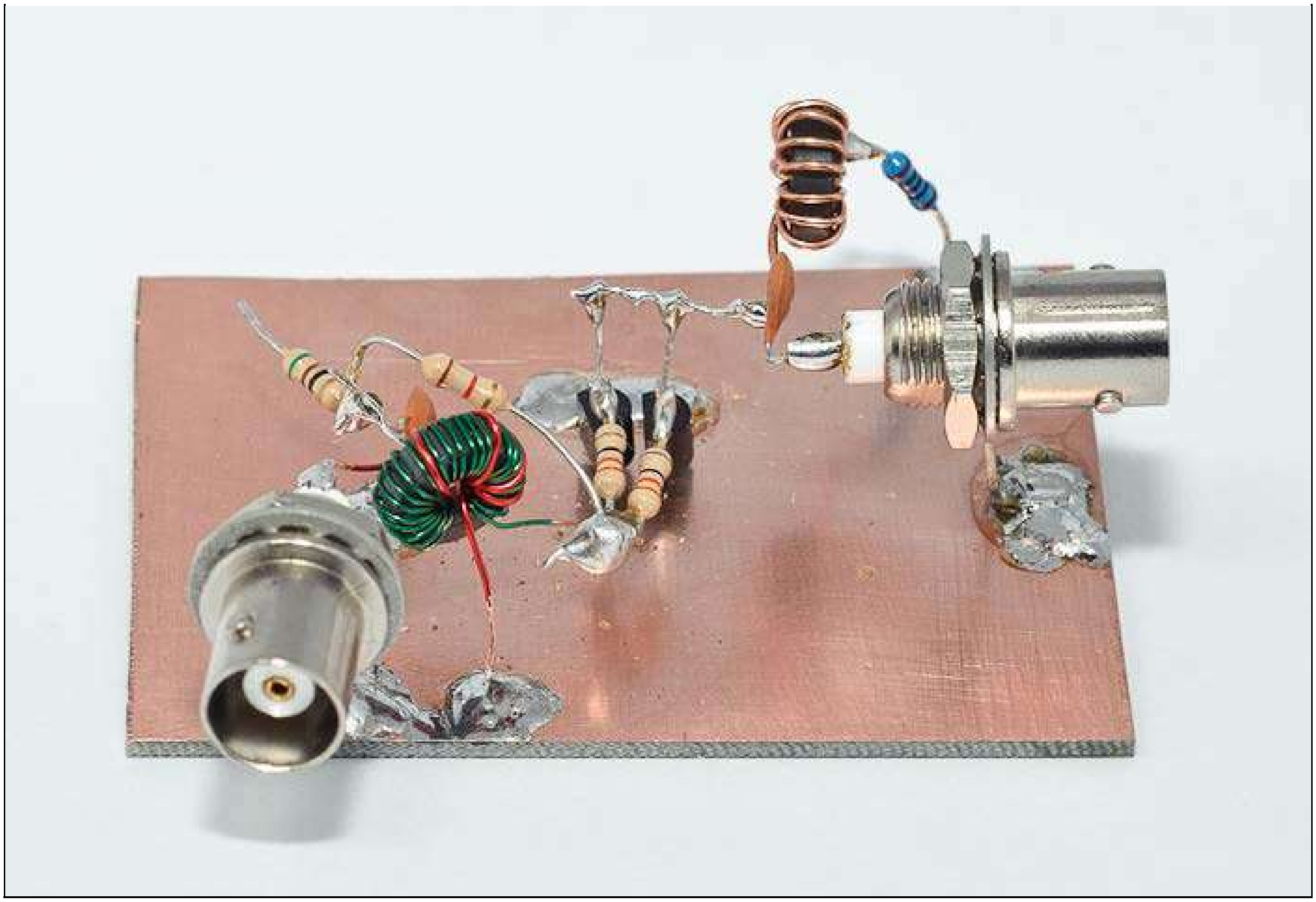

Above — The breadboard of the above 7 MHz CG amplifier.

3

Output Impedance Experiments

For some reason, I assumed that when using an arithmetically correct turns ratio, the output transformer will end up at 50 Ωs. For example, if I wish to transform 450 ohms to 50 ohms, I'd use a 9:1 impedance ratio (3:1 turns ratio) and get 50 ohms. Sadly, it isn't this simple — impedance transformation is complicated and whole books have been written about it. I'll share some of my experiments that might inform yours.

The first task was to built a simple jig to evaluate primary and secondary coupling, turns ratios and return loss.

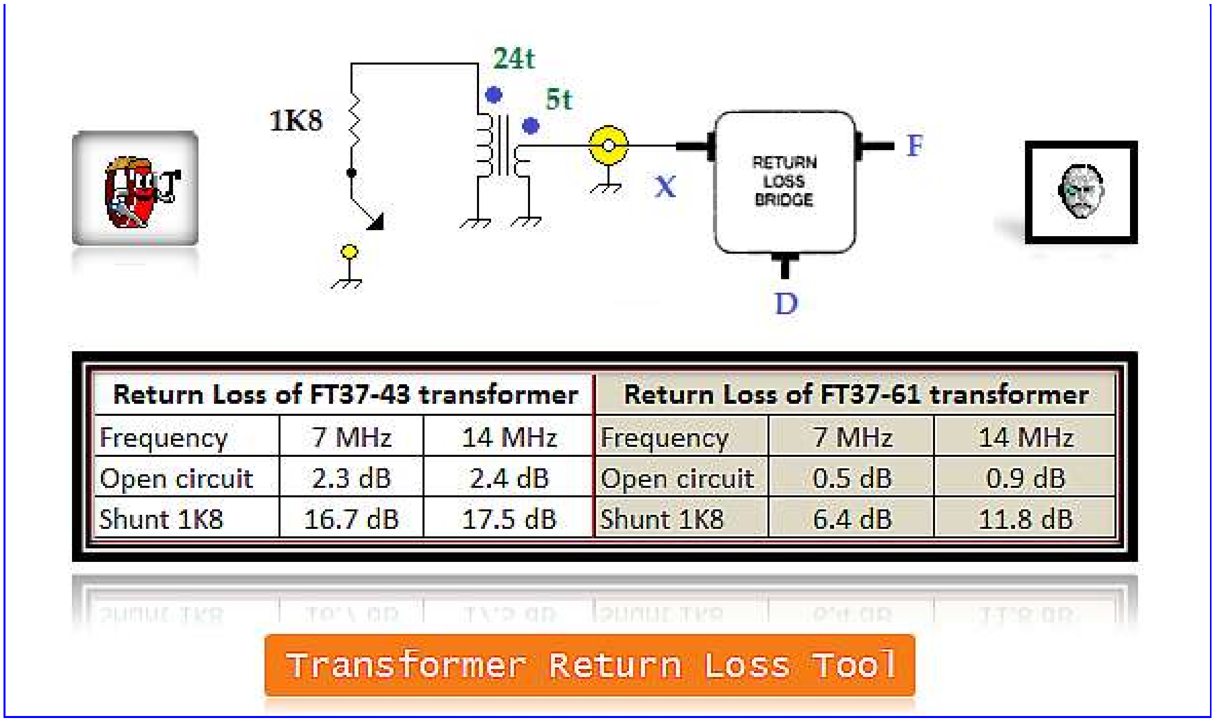
24t
5t
1K8
X
RETURN LOSS BRIDGE
F
D
Transformer Return Loss Tool

Return Loss of FT37-43 transformer			Return Loss of FT37-61 transformer		
Frequency	7 MHz	14 MHz	Frequency	7 MHz	14 MHz
Open circuit	2.3 dB	2.4 dB	Open circuit	0.5 dB	0.9 dB
Shunt 1K8	16.7 dB	17.5 dB	Shunt 1K8	6.4 dB	11.8 dB

Above — The simple tool built to evaluate the return loss of a transformer out-of-circuit. In this case, I examined the 24t : 5t transformer of the 7 MHz CG amplifier shown earlier. The table shows the best possible return loss when the 1K8 resistor is across the primary coil. Additional experiments were completed and follow below.

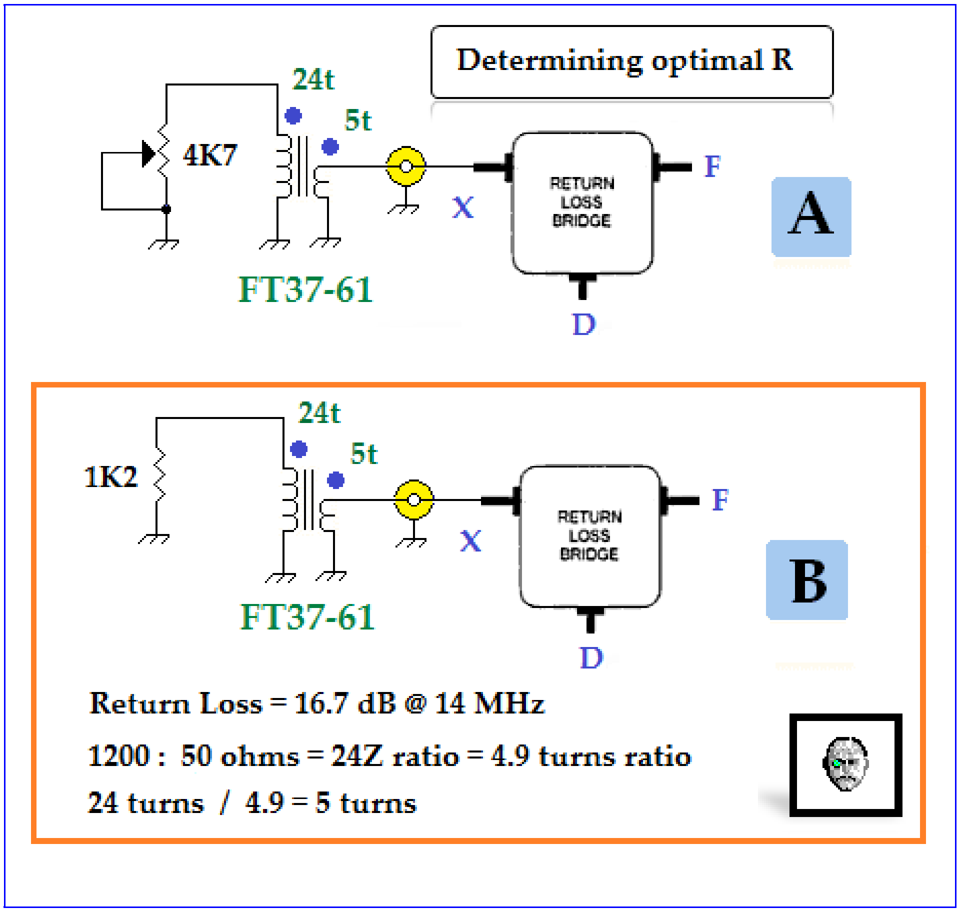
Determining optimal R
24t
5t
4K7
X
RETURN LOSS BRIDGE
F
D
FT37-61
A
24t
5t
1K2
X
RETURN LOSS BRIDGE
F
D
FT37-61
B

Return Loss = 16.7 dB @ 14 MHz

1200 : 50 ohms = 24Z ratio = 4.9 turns ratio

24 turns / 4.9 = 5 turns

Above — An experiment to see if changing the shunt resistor can improve return loss; yes it can. The shunt resistor was a 4K7 potentiometer — Using the potentiometer, I was able to determine the optimal resistance needed to increase the return loss @ 14 MHz of the **FT37-61** ferrite-base 24t : 5t transformer. The pot was removed, measured and replaced with the nearest standard value; a 1K2 resistor. The best possible RL was 16.7 dB using a 1K2 shunt resistor. At 7 MHz, the FT37-61 didn't work well. Five turns on a FT37-61 based transformer doesn't have enough inductive reactance to get a good return loss.

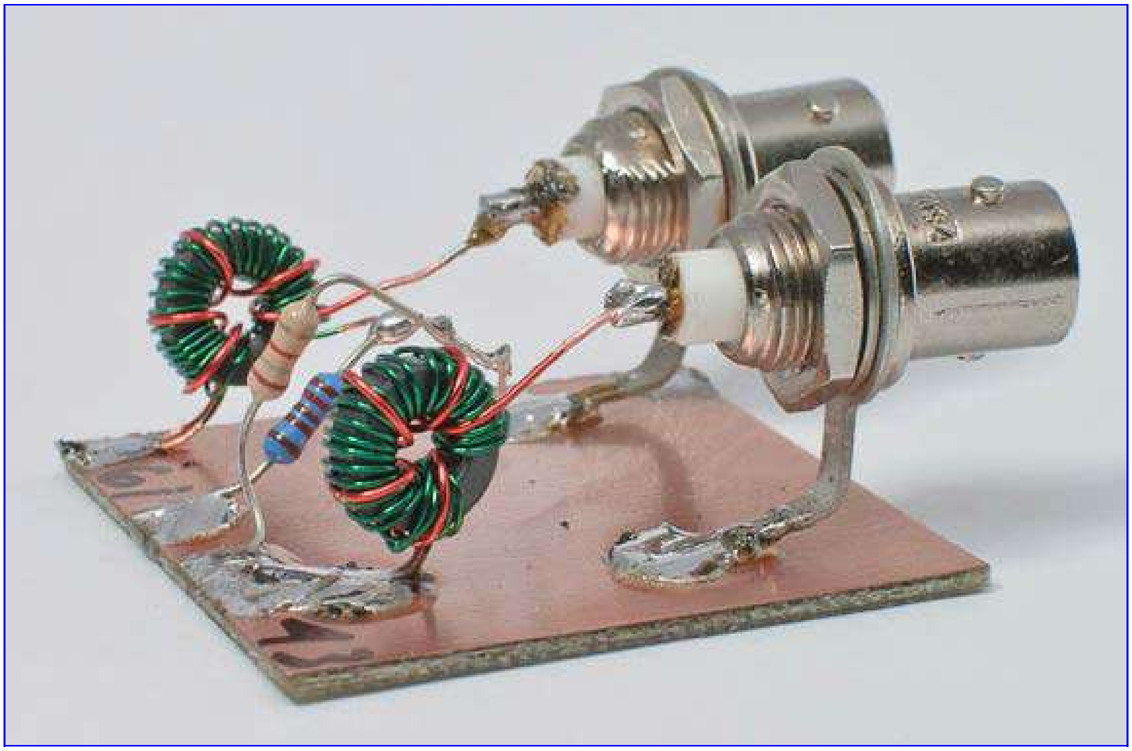

Above — The transformer testing jig. I omitted the switch shown in the schematic above and just soldered the shunt resistor across the primary winding.

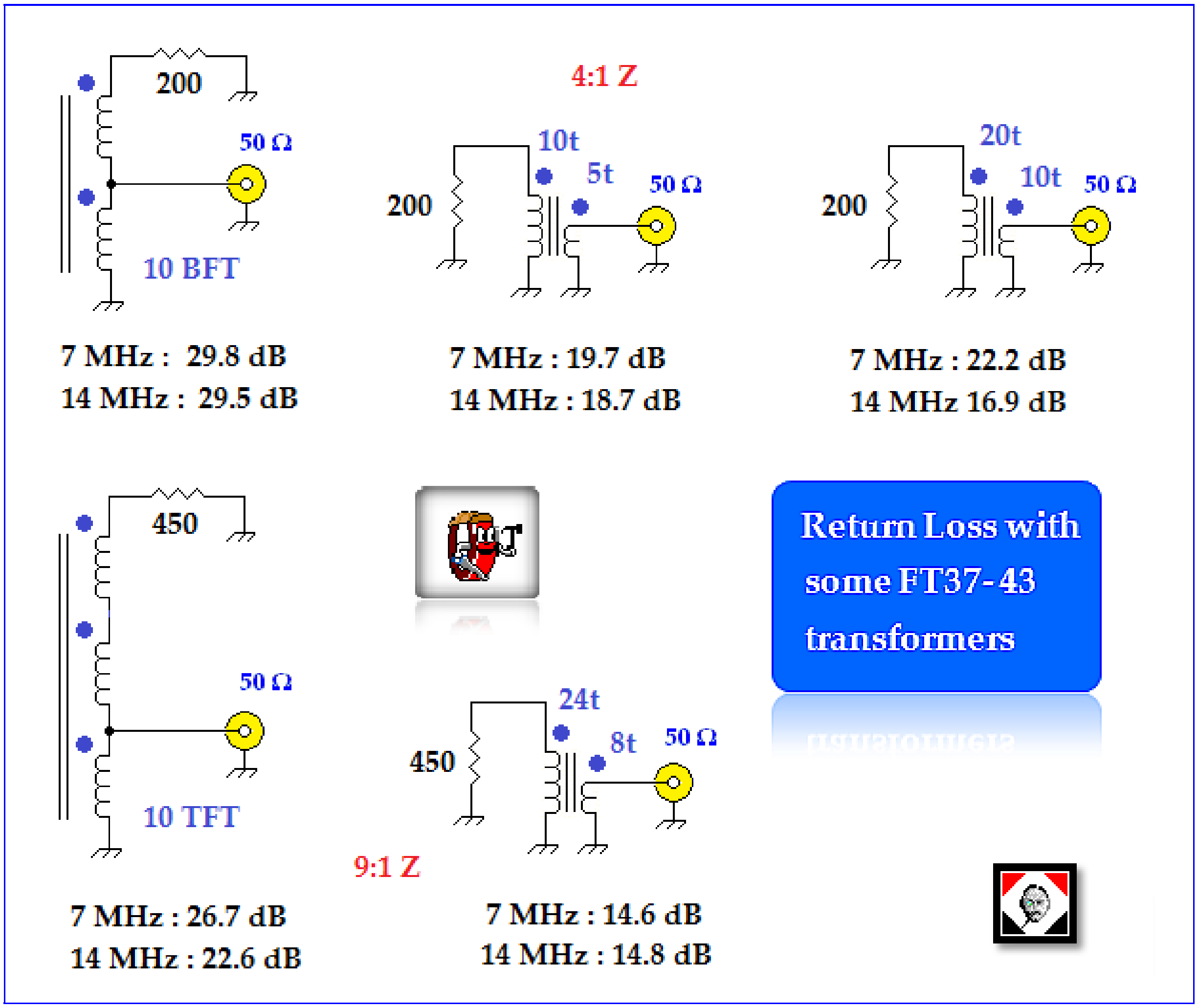
200
50 Ω
10 BFT
7 MHz : 29.8 dB
14 MHz : 29.5 dB
4:1 Z
10t
5t
200
50 Ω
7 MHz : 19.7 dB
14 MHz : 18.7 dB
20t
10t
200
50 Ω
7 MHz : 22.2 dB
14 MHz 16.9 dB
450
50 Ω
10 TFT
7 MHz : 26.7 dB
14 MHz : 22.6 dB
Return Loss with some FT37-43 transformers
24t
8t
450
50 Ω
9:1 Z
7 MHz : 14.6 dB
14 MHz : 14.8 dB

Above — Some of the outcomes using the transformer jig pictured above. While I basically understood that transformer efficiency tends to fall as the turns ratio increases, I never thought this would also happen with return loss. By *no* means do these crude experiments constitute science, but the following themes emerged:

1. The better coupling of transmission line transformers (bifilar, trifilar etc.) translates into improved RL over conventionally wound transformers
2. Limiting the turns ratio to 3:1 or less generally improved the return loss. As the turns ratio moves above 3:1, the best possible return loss tends to decrease.
3. The smaller or secondary winding should have 4-10 times the inductive reactance of the impedance it's connected to. For a 50 ohms impedance this means a minimal XL of 200 - 500 ohms. I noticed a weak trend towards better return loss with higher XLs. This means that to use a FT37-61 at 7 MHz, the secondary winding should be 9 -14 turns or so.

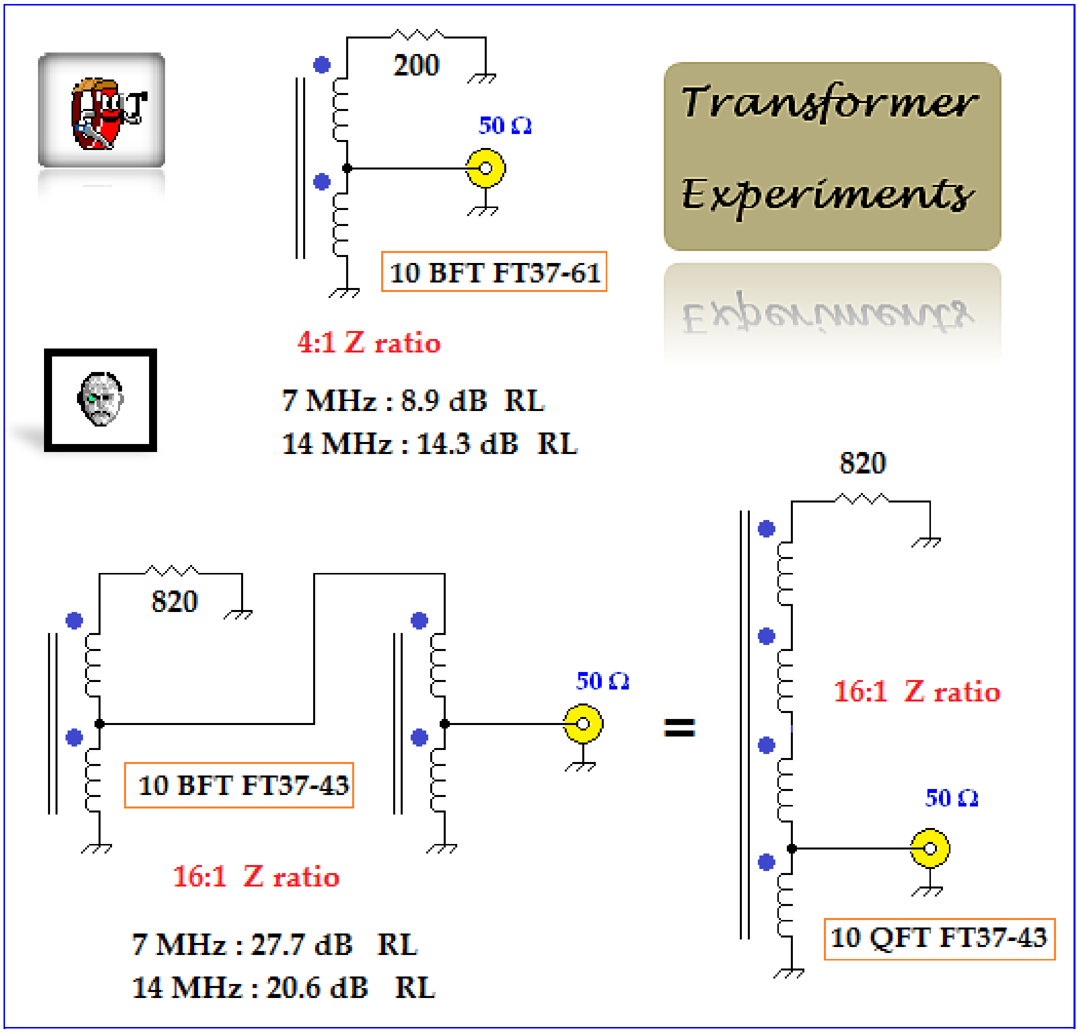
200
50 Ω
10 BFT FT37-61
4:1 Z ratio
7 MHz : 8.9 dB RL
14 MHz : 14.3 dB RL
Transformer Experiments
820
820
50 Ω
=
10 BFT FT37-43
16:1 Z ratio
7 MHz : 27.7 dB RL
14 MHz : 20.6 dB RL
16:1 Z ratio
50 Ω
10 QFT FT37-43

Above — Further transformer experiments. For a 4:1 impedance transformation at 7 and 14 MHz, a FT37-43 ferrite toroid gave a better out-of-circuit RL than the FT37-61 *The comparison transformer with a FT37-43 ferrite core was shown earlier*. It's possible to transform a big impedance such as 16:1 by cascading 2 bifilar transformers, or by using a quadrifilar transformer. I didn't build the quadrifilar transmission line transformer, but show it for completeness sake.

Of course, once you connect the transformers to a real circuit, things will change — still it's great to be able to examine transformer return loss in a controlled environment.

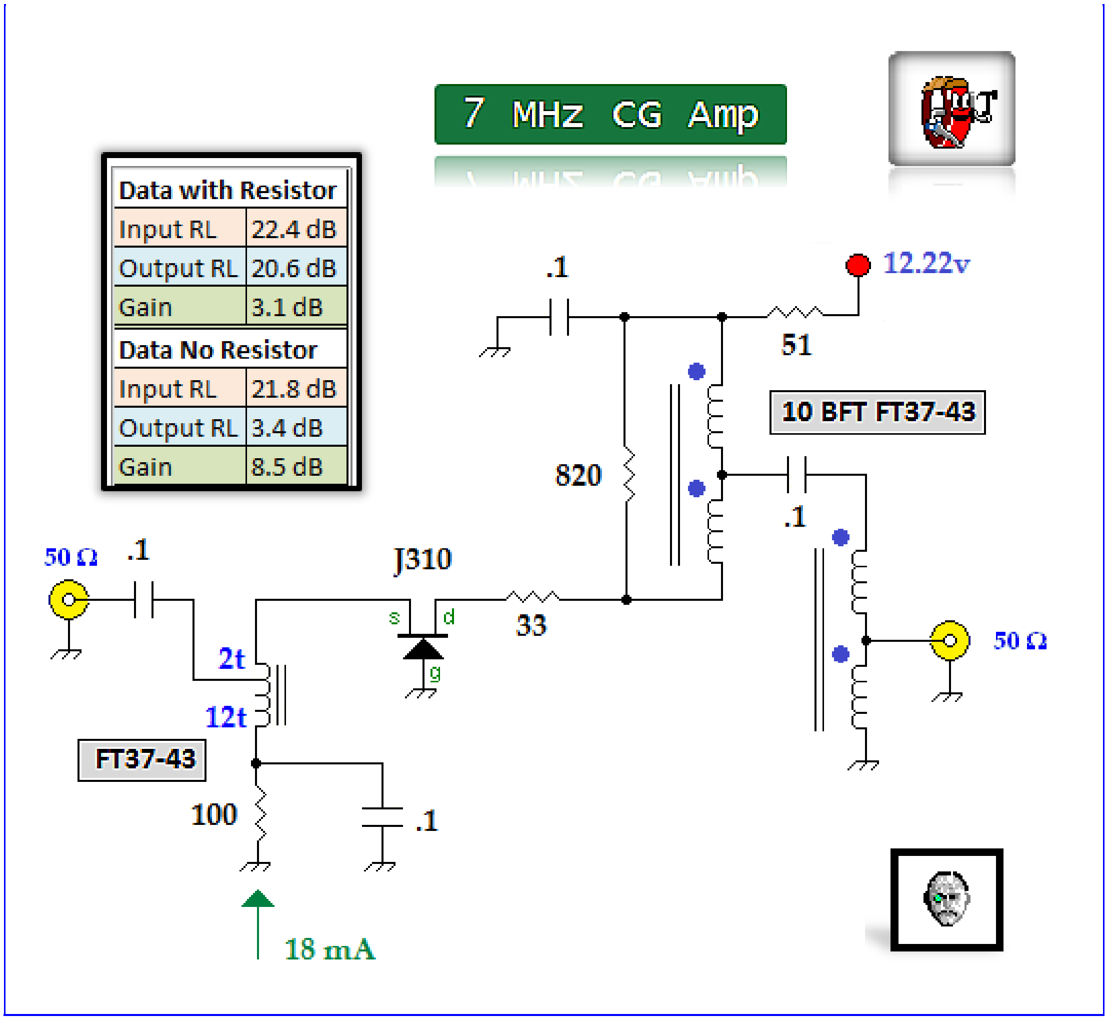

Data with Resistor	
Input RL	22.4 dB
Output RL	20.6 dB
Gain	3.1 dB
Data No Resistor	
Input RL	21.8 dB
Output RL	3.4 dB
Gain	8.5 dB

7 MHz CG Amp
12.22v
.1
51
10 BFT FT37-43
820
.1
50 Ω
.1
J310
s
d
g
33
50 Ω
2t
12t
FT37-43
100
.1
18 mA

Above — a common gate amplifier experiment using 2 cascaded 4:1 Z transmission line transformers. Data with and without the 820 Ω resistor shows that while the resistor gives a great output RL; it eats a lot of power. In cases where I've seen cascaded transmission line transformers used, the resistor was omitted. The Ugly Weekender transmitter by Wes, W7ZOI provides a good example.

In many cases, it's prudent to sacrifice gain for return loss, however, when you see a builder (like the former me), put a 32:3 turns ratio on a 5 MHz amplifier output transformer and label the secondary windings **"50 Ω",** we'll know better.

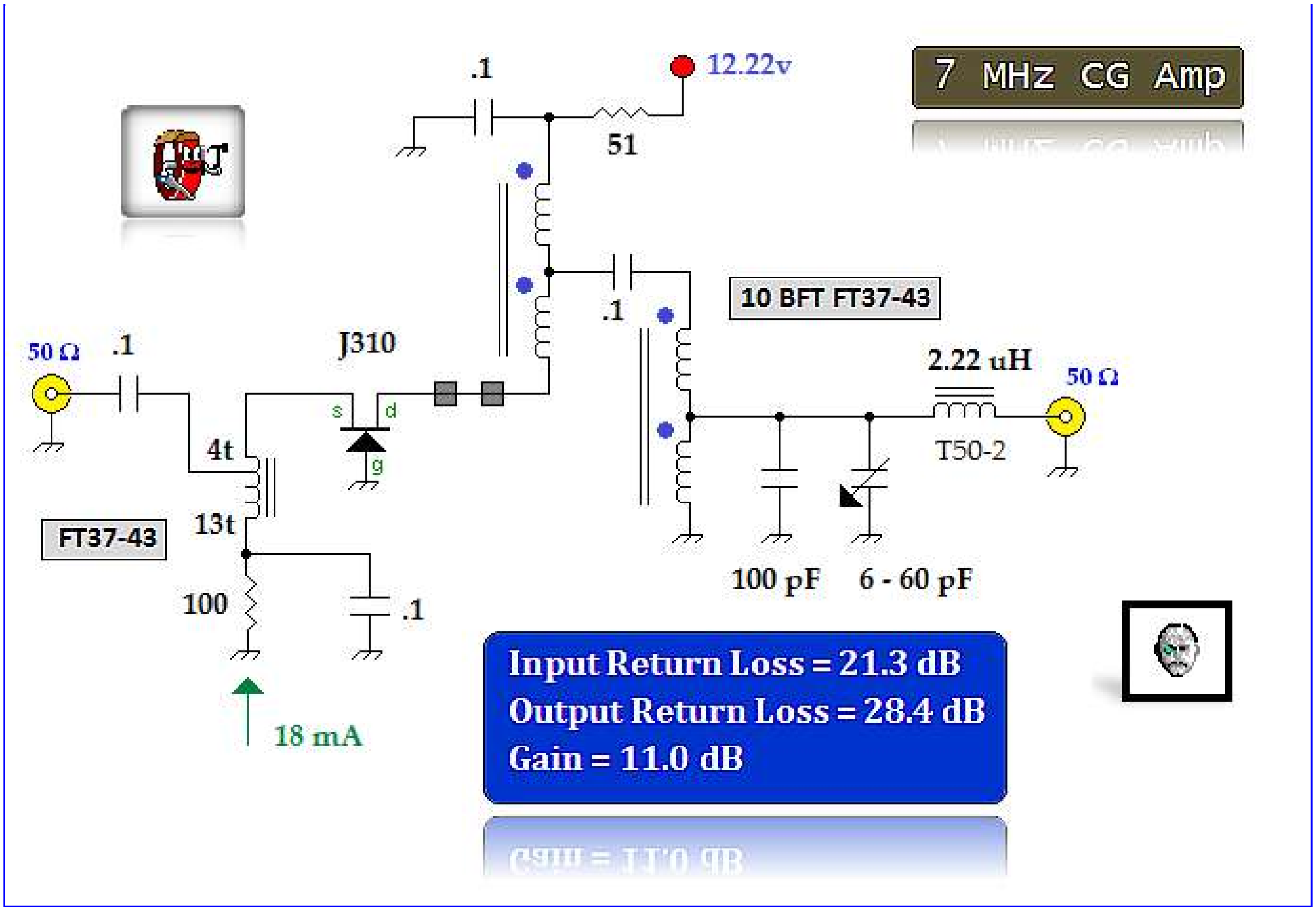

Input Return Loss = 21.3 dB
Output Return Loss = 28.4 dB
Gain = 11.0 dB

Above — An evolution of the amplifier above to get the best possible output RL. I omitted the 820 ohm resistor and matched the output with an L-network. The return loss on the output of the second transmission line transformer (measured before the L-match was added) was 3.4 dB.

The L-match values were *roughly* determined by using this chart (you can also do the math). According to the chart an *(output)* RL of 3.4 dB, is either 10 ohms or ~250 ohms or so. Ten ohms is unlikely, so I designed my L-match to match 250 to 50 ohms. This provided some starting values for the L and C parts and the rest was done on the bench using trial and error with an RL bridge. At the time, this was the highest output RL I'd ever achieved.

RF engineers use math to calculate impedance (they always do). I sent the schematic to Wes, W7ZOI for his analysis and summarize his return email comments as follows: At 7 MHz, the XL of the 2.22uH inductor is 97.6 Ω, therefore the impedance looking into that with 50 Ohms as the load is 50+j97.6. A complex inversion of this value gives a complex admittance that has a real part: 0.0041. Flipping that gives 240 Ω. The equivalent reactance is inductive with a value that would be tuned by a 184 pF capacitor; a bit more than you have there — so there is some reactance presented by the center tap of the second transformer. Neglecting these details, the L net generates about 240 Ohms. The two transformers then kick the Z up by 16 to 3856 Ohms.

I was pleased that my simple chart gave a value close to his calculation. Test it out — the chart may work okay for you.

The input match is over 20 dB and reasonable. More time could have been spent on the input autotransformer by tapping and such to increase the input RL, however, time is the 1 resource we all seem to lack.

The final amplifier experiments employ an L-match to set output return loss. When reading electrical engineering books you'll often see all sorts of matching networks on both the input and output of FETs and BJT amplifiers. The networks look simple, but in practice, aren't. They tune sharply, have a low bandwidth and in the case of the CG amp, harbor a big problem — tuning the output for the best output return loss, **dramatically** affects the *input* return loss and potentially, your return loss measurement by the reactance affecting the RF signal in your bridge.

When tuning the output, you're actually changing 2 complex impedances — this is not trivial.

Also if you're off by a few pF or tens of uH in your network C and L values respectively, you can wreak havoc with the measurements. At this point, I don't possess all the skills needed to tune both the output and the input network to a RL of 20 dB or greater; especially with a broadband input.

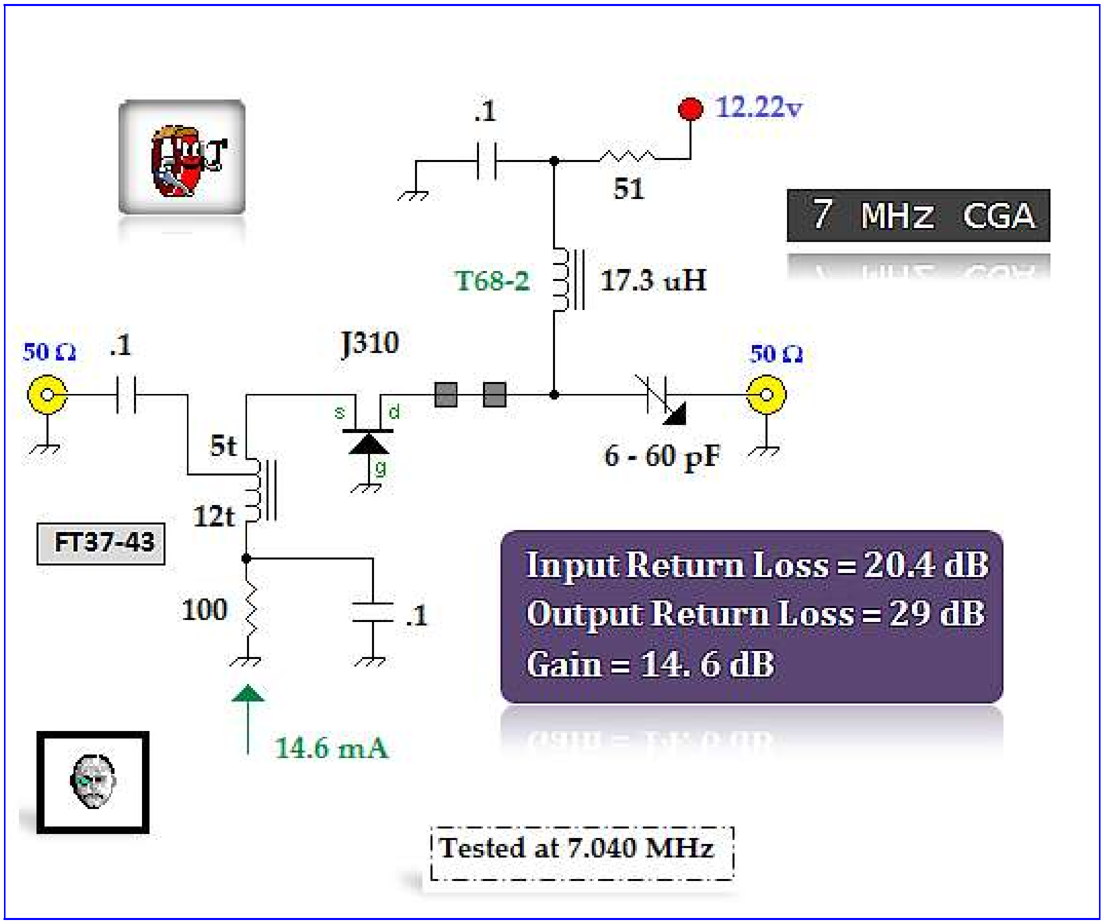

Above — A common gate amp employing a high-pass L-network to match the output. Miraculously after 2 hours of tweaking, I obtained a good input and output match; however this amp isn't reproducible. The inductor was wound on a T68-2 using 28 gauge wire — always a pain. Through trial and error, I learned that the output impedance of the drain was around 11800 ohms. Starting with 18.4 uH on theT68-2, I removed 2-3 turns at a time until a reasonably low return loss was obtained; then I removed 1 turn at a time. I went too far and had to add back a turn. I clipped the excess lead every second turn which made it tedious, but exacting. It seems that the L value is very critical – it would be nice to use a variable inductor to figure these things out. Compressing and expanding the windings also provided a simple way to vary inductance.

In several other circuits, the best possible input return loss was only 14 dB. Mistuning also caused oscillations to occur in one 14 MHz amp with an output network inductor of 7.4 uH. I also tried a 14 MHz amp with an L-match on both the input and output, however, was unable to match both the input and output due to the interplay between them.

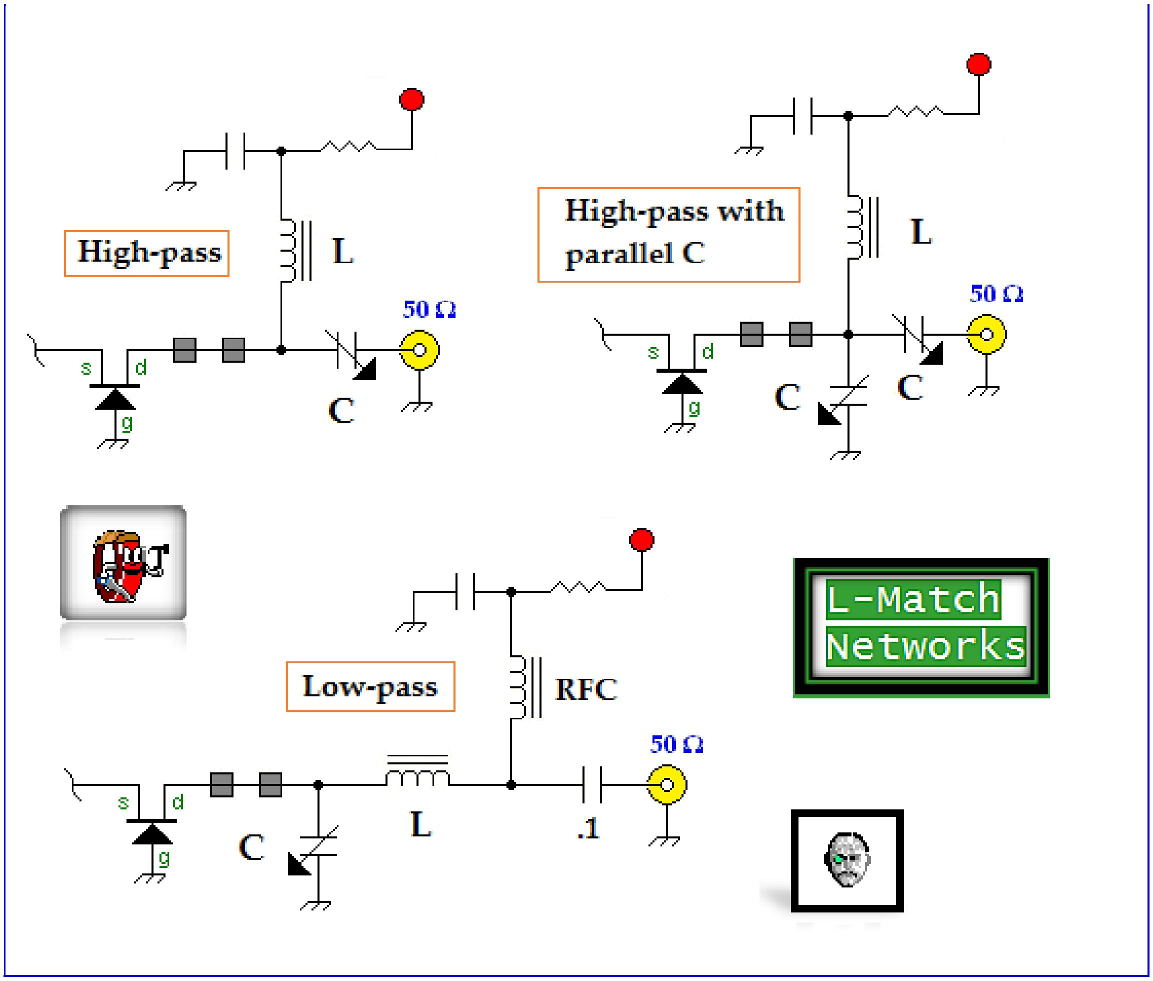
High-pass
L
50 Ω
s
d
g
C
High-pass with parallel C
L
50 Ω
s
d
g
C
C
Low-pass
RFC
50 Ω
s
d
g
C
L
.1
L-Match
Networks

Above — Here are 3 possible L-network configurations for tuning a CG amplifier output. They can be used in other circuits and are worth studying. The L-match with 2 variable capacitors generally requires lower inductance than the others.

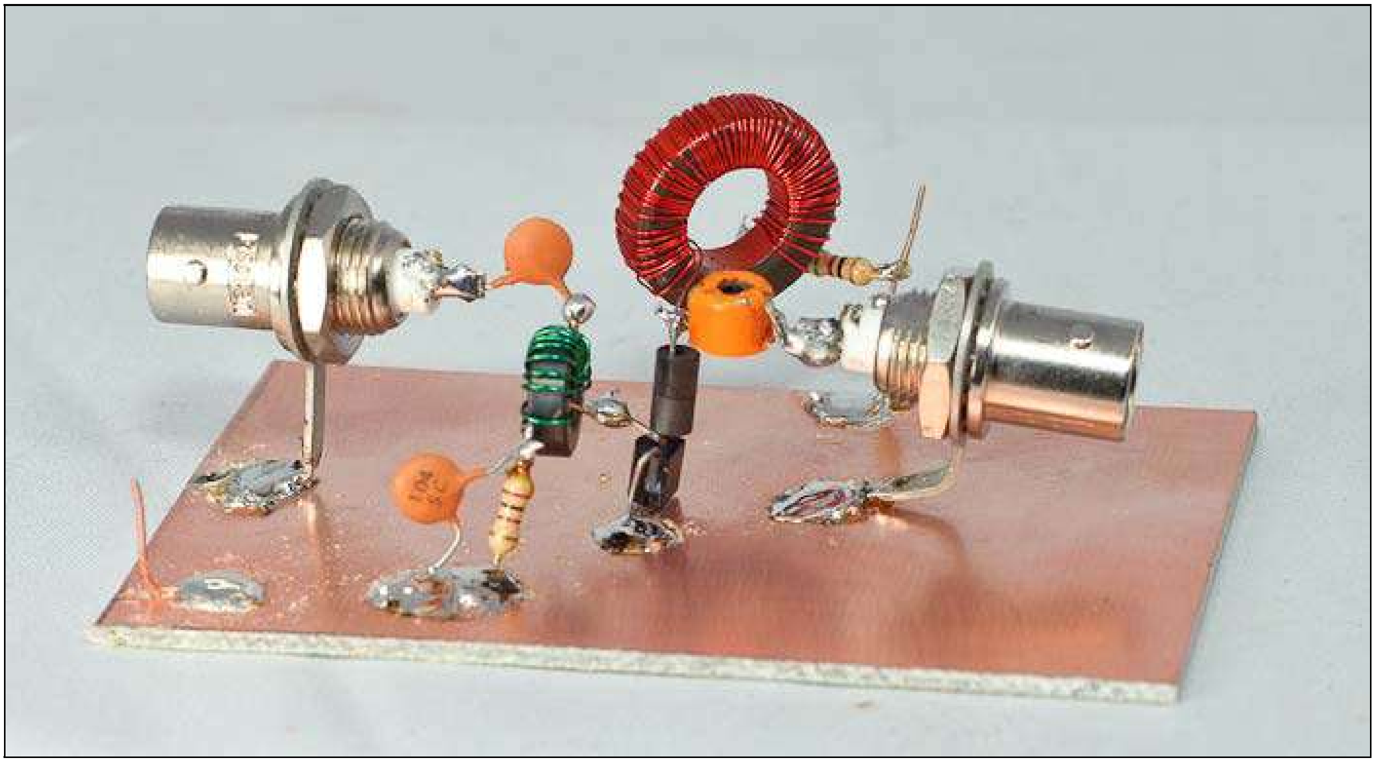

Above — A breadboard of 1 of the high-pass tuned CG amps. The gate lead on this transistor is too long — the inductance will likely cause UHF oscillations. 2 ferrite beads were placed on the drain to mitigate these, but a better construction technique is recommended and shown below.

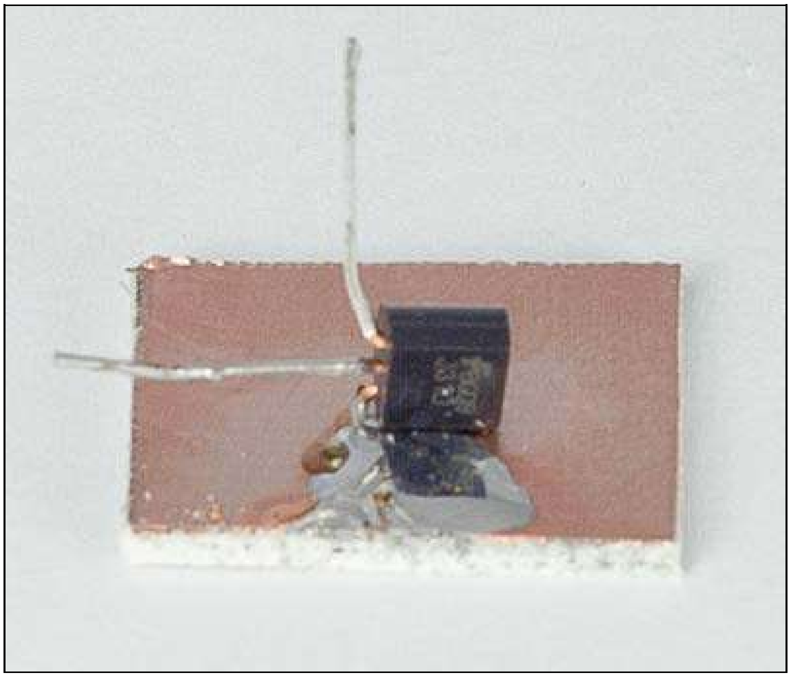

Above — the preferred way to ground the gate with the JFET on its side. The hole-through version of the U310 JFET has a metal case that is connected to the gate that makes it ideal for grounded gate amplifiers. Some suppliers only sell SMT versions of the U310 now.

4. QRP — Posdata for January 2013: Transistor Bias Model

This discussion concerns setting up the DC bias point for linear BJT operation.

Earlier I stated that a bipolar transistor is a voltage controlled device. A few readers thought I made a typo: something I frequently do, but not in this case, since I purposely made that statement. In reality, the argument could go either way since collector–emitter current is controlled by the base-emitter current (~a current controlled device) and by the base–emitter voltage (~a voltage controlled device).

Stated using the correct physical model, a transistor is a **current controlled current source**. With external circuitry we can manipulate this physical model into a voltage controlled current source, or a voltage controlled voltage source, or even a current controlled voltage source. Whether you model the transistor with current or voltage, the math tells the truth when properly examined.

Please view the following two 2N3904 SPICE models generated by Wes, W7ZOI for me many years ago when I began to learn small signal analysis using impedance and hybrid parameters, plus set out to learn ways to establish DC bias and temperature stability in BJTs.

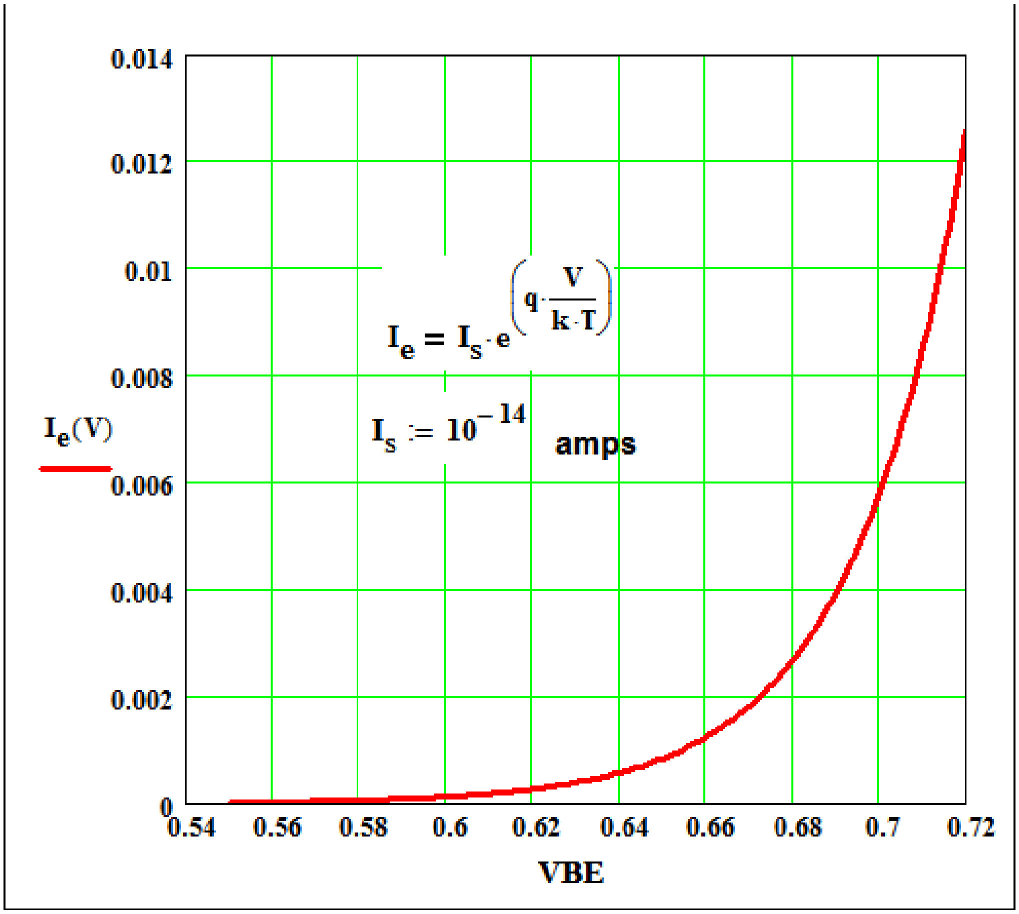

Above — the Y axis shows how changing base-emitter voltage or current changes the VBE. We tend to assume a VBE of 0.7, however, the math shows the truth. Whether we plot voltage or current for the Y axis data, the graph slope remains similar. The greater the applied DC voltage placed on the base-emitter port, the more current will flow.

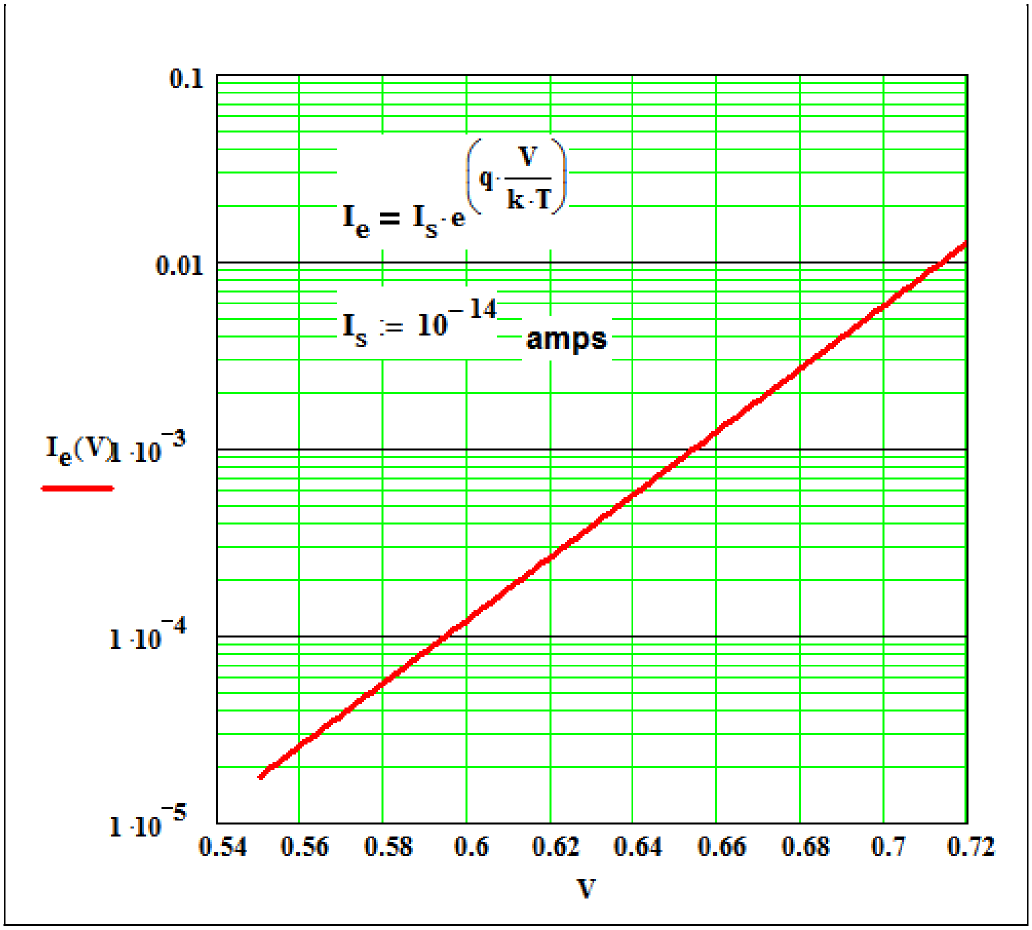

Above — Logarithimic base current plotted against VBE. If we want this current to increase, we need to put more DC voltage on the BE junction. On the bench, we may easily measure base voltage to confirm our calculations — measuring base current proves more difficult. Whether I'm setting up amplifer bias with voltage dividers, a current source, or even biasing it with a downstream AGC voltage, I prefer to think in terms of voltage control — although I get that V and I truly just coexist.

Current or voltage modelling — it's your choice and the math will guide you. Look for these equations on the Web, or in second hand bookstores. I've got 8 or 9 transistor theory books now and they're really timeless.

---

5. Miscellaneous Figures and Photographs

ECS
11.04
KOREA

RF
Shunt Feedback
Decreases input Z
Affects current gain

FBA

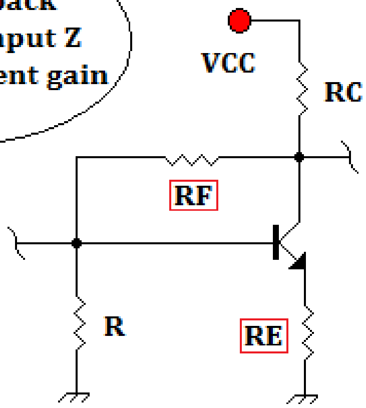

RE
Series Feedback
Increases input Z
Affects voltage gain
emitter degeneration

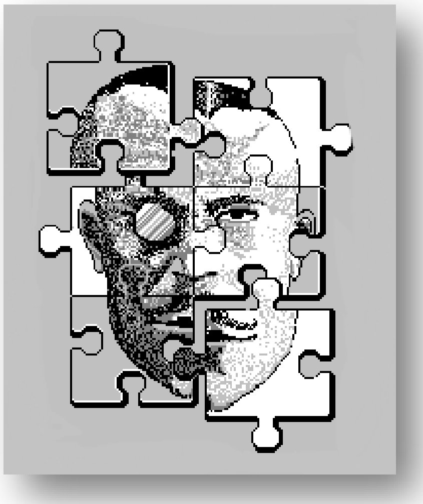

6. References

1. Communications Receivers - Principles & Design. Rohde, L.R and Bucher, T.T.N. 1988. McGraw-Hill.
2. Technical Correspondence QST Magazine (ARRL) Aug 1990. Hayward W.
3. Experimental Methods in RF Design (ARRL) 2003. Hayward W. Campbell R. Larkin B.
4. Microwave Handbook Vol 1. Components and Operating Techniques. 1989 RSGB.
5. Emails with Wes, W7ZOI and Professor Ken Kuhn, Winter 2010-2011 — Thanks! Никогда не забуд.

RF — Test and Measurement

Hobby and Fun 2011

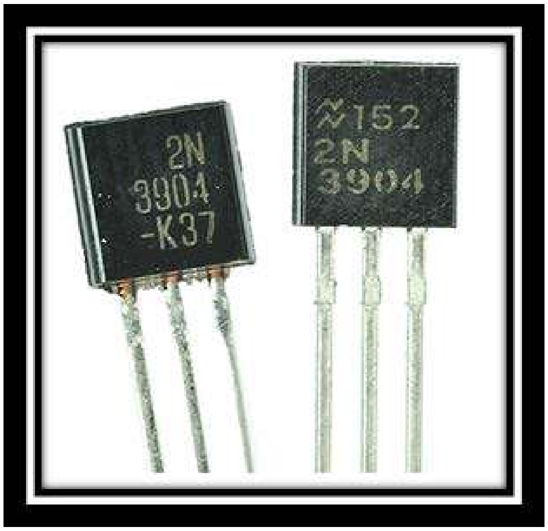

Introduction

Hobbies are supposed to be for recreation. Electronics should be fun, not stressful — heavy math, big parts counts and complexity are more likely to scare away experimenters than recruit them. This page avoids the measurement focus of my latest stuff and simply promotes fun and discovery.

You might be interested to know that my simple experiments/projects garner the most emails. Many wrote "I'm rediscovering electronics", or, "I want a simple and fun hobby". Hobby and fun are my goals too.

Simple Regenerative Receiver Experiments

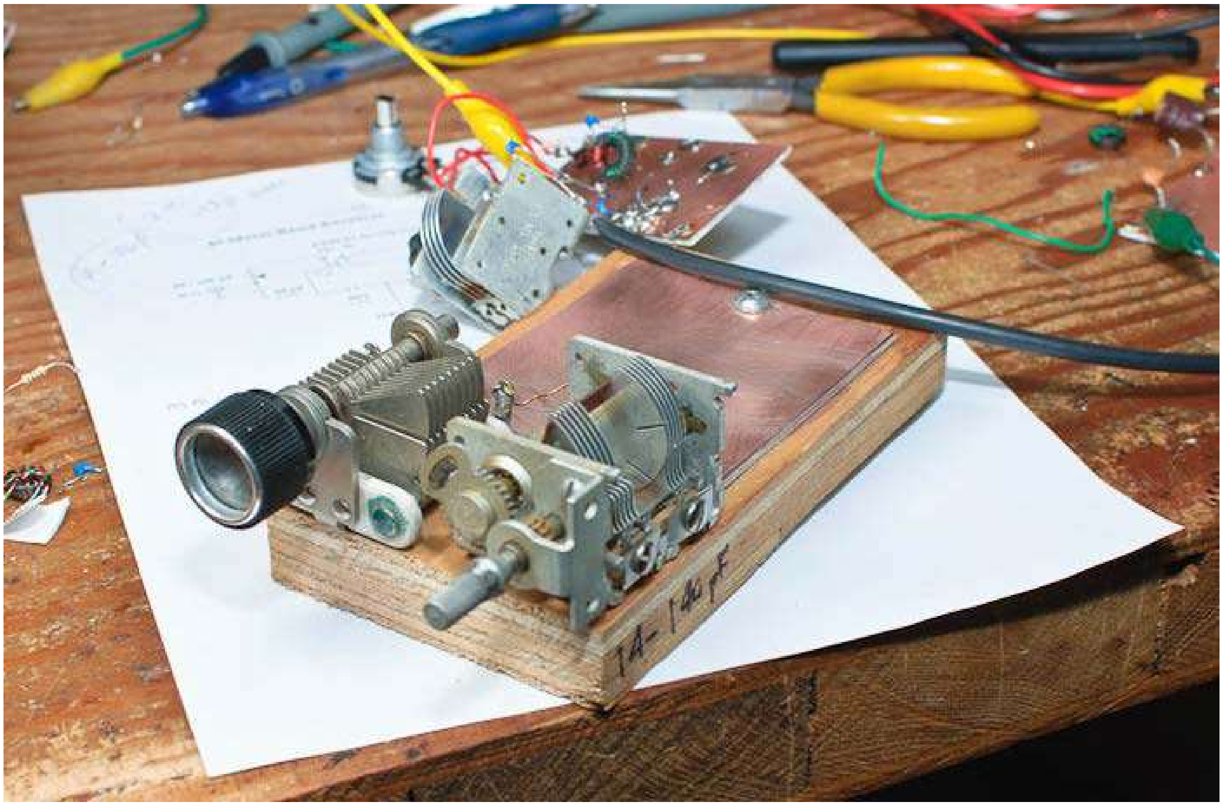

Above — 2 air variable capacitors and a copper clad board screwed onto a piece of wood for my bench musing. Regenerative receivers delight and amaze — some builders take them very seriously. I respect this, but to me; they should be as simple as possible. I wanted a 2 stage "genny" receiver for this page and present 2 different receivers; 1 is my design, the others is a JFET variant of a favorite W7ZOI circuit. Quoting Wes, W7ZOI "feedback your imagination".

Some builders place a simple common gate or a common base RF preamp on the input to boost gain and reduce antenna radiation of the RF oscillator, while others place an RF gain control on the input — usually a potentiometer; to prevent overloading the RF stage. I won't prescribe what to do — that's up to you. There are countless example of regenerative receivers on the web and you can many spend hours viewing them. Some of the most intriguing are those built by Russian speaking experimenters. Example link.

My circuit ideas are meant as fodder for your own experiments.

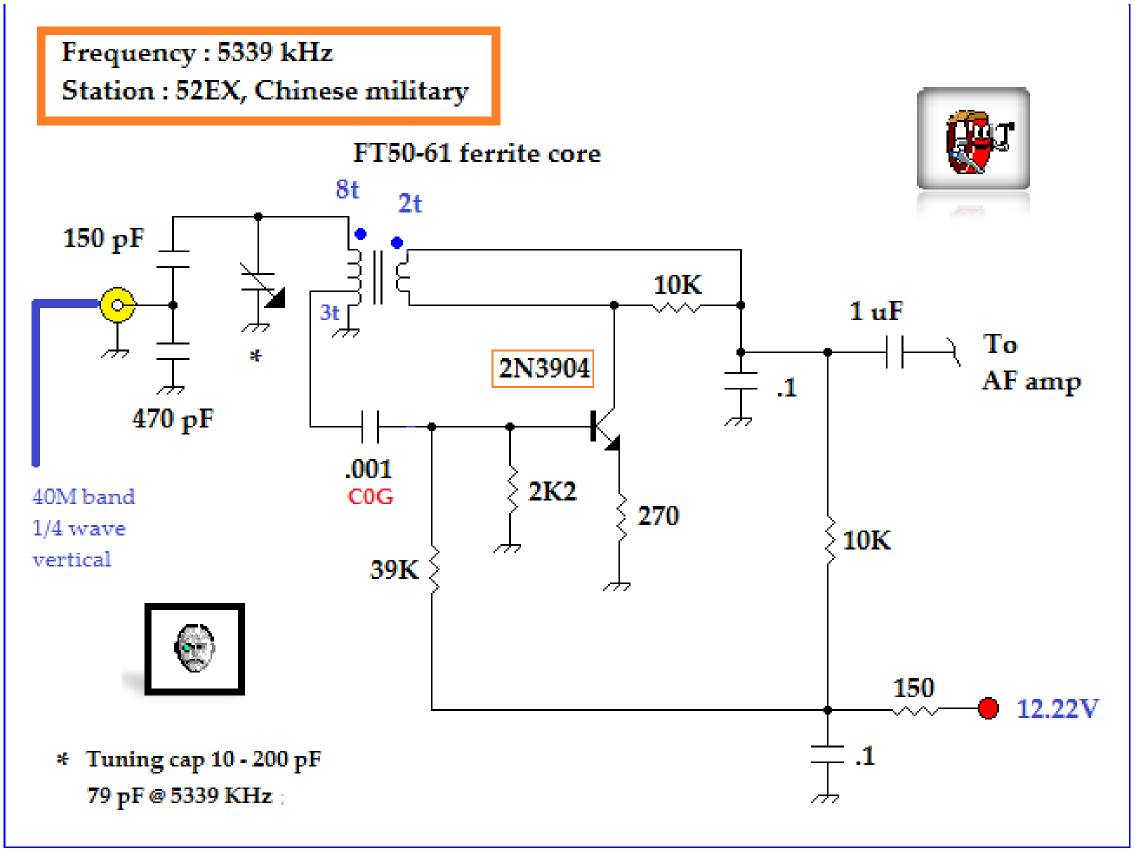
Frequency : 5339 kHz
Station : 52EX, Chinese military
FT50-61 ferrite core
8t
2t
150 pF
3t
10K
1 uF
To
AF amp
2N3904
*
.1
470 pF
.001
C0G
2K2
270
10K
40M band
1/4 wave
vertical
39K
150
12.22V
.1
* Tuning cap 10 - 200 pF
79 pF @ 5339 KHz

Above — An experimental, ultra-simple "CW" receiver. At 5339 KHz I hear strong Morse code each night. It's suggested to be from China, but I'm unsure. Connected to my 1/4 wave 40M band vertical - a simple matching network and trimmer tuning capacitor were fitted to the input. Here's some audio. I like the beat note of this receiver - it has no regeneration control and is fixed for CW. Minimalist circuits are fun — some hardcore regen builders might freak out; no voltage regulator (here's a version with that + a T68-6 inductor), no regeneration control (here's a version with that), a relatively low Q coil etc. I wanted to try my hand at design and not just copy someone else's receiver.

I call it the Stupid— Simple receiver and although it emits crisp, warm audio, some bench work is required to get the correct bias and appropriate amount of positive RF feedback — an experimenter's circuit that explores DC bias and AC feedback. The sort of thing a father can build with his son. We need more circuits fitting this profile.

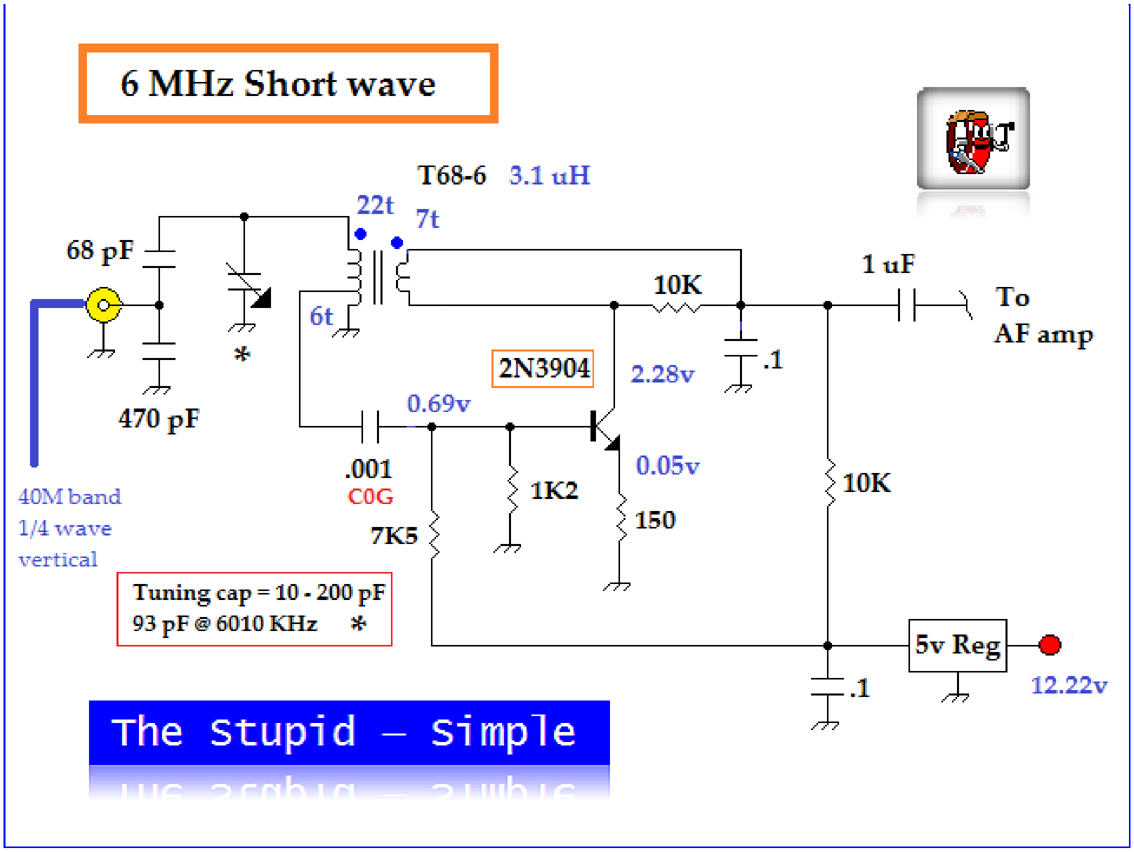
6 MHz Short wave
T68-6 3.1 uH
22t
7t
6t
68 pF
470 pF
1 uF
To
AF amp
10K
.1
2N3904
2.28v
0.69v
0.05v
.001
C0G
1K2
7K5
150
10K
40M band
1/4 wave
vertical
Tuning cap = 10 - 200 pF
93 pF @ 6010 KHz *
*
5v Reg
12.22v
.1
The Stupid — Simple

Above — The Stupid—Simple experimental receiver set up for ~6 - 7 MHz. 2011 marks the 50th anniversary of Radio Habana Cuba. We tune RHC at 6010 KHz, and as long as I've been listening to SW radio, I've tuned this station. Here are 3 sound bytes from around 6 MHz recorded from 2:30-2:40 GMT on Feb 28, 2011, including an old repeated episode of DX'ers Unlimited by Arnie Coro which had faded out by the time I located and turned on my audio recorder. Audio1 Audio2 Audio 3. The audio stage in these recording was a discrete transistor AF amp I designed, however, an LM386 was chosen for the final amp to keep the parts count and difficulty down. I got a little too close to the receiver a couple of times during the recording and made the open circuit breadboard squeal.

Arnie Coro talked about a "regenerodyne" receiver in Sound Byte *Audio3*. Very cool. Here's a link. Now this is radio!

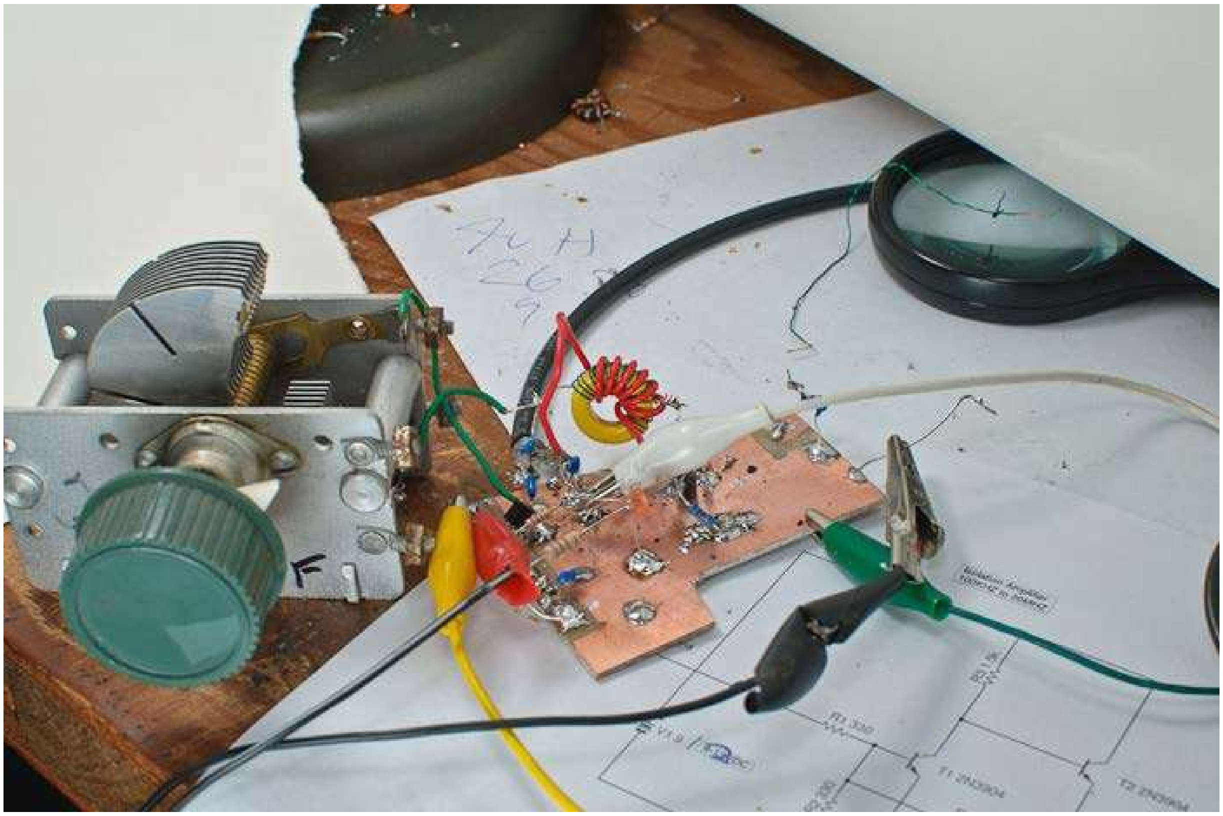

Above — The Stupid-Simple regenerative receiver breadboard for 6 MHz. You can see the T68-6 — the red secondary windings are wound in the same direction as the 22 turns of primary. I started with 9 turns and unwound a link and tested sequentially until I had the right amount of feedback for AM. I just used normal hook up wire for the secondary winding. The white colored cable goes to the audio amp. This is a prototype experimental layout — a regenerative receiver should have short connections around the tank circuit and be in metal box for best results.

On some stations, my little 1 RF + 1 AF stage receiver sounded better than my superheterodyne receivers. The 2N3904 is just barely turned on — I determined that a base bias of 0.66-0.69v provided maximal sensitivity. The 150 ohm emitter resistor can be a 500 ohm pot and used to fine tune the regeneration. 150 ohms gave the best compromise gain and feedback + current for the 2N3904.

Audio Amplifier

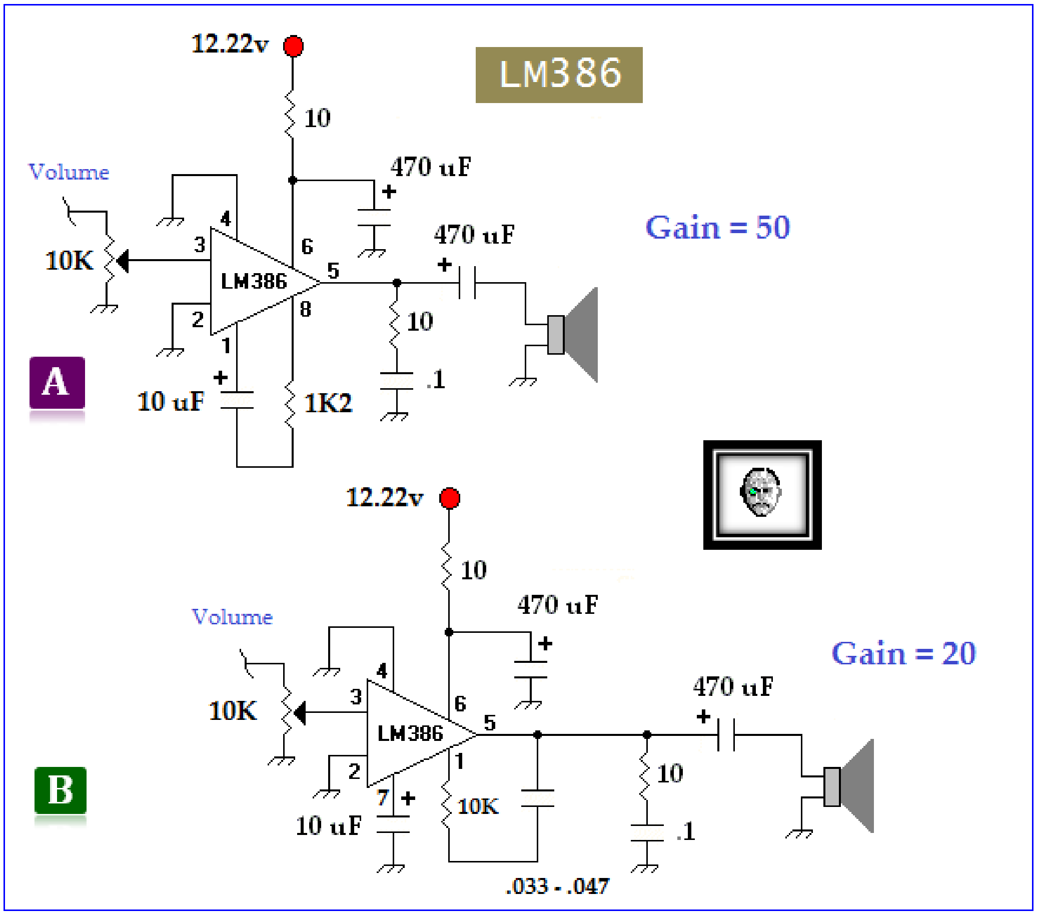

Above — I chose an LM386-N for my AF power amplifier. The LM386 exhibits less peak signal distortion when run in the low-power (X20) gain mode and a higher VCC such as 12 volts. My receiver used the schematic denoted B — a 10K and 0.033 to 0.047 uF RC network is used to reduce the amplifier high frequency response. Click for a sound byte of me tuning around 6 MHz with a 10K + 0.047 uF RC network between pins 1 and 5. Figure B is my favorite way to use the LM386 and comes right off the National Semiconductor LM386 data sheet. Look for this data sheet with your favorite search engine.

Because I have a big antenna, Figure B provides adequate volume to a speaker. Connecting pins 1 and 8 via a 10 uF capacitor bypasses some emitter resistance and gives X200 gain. A resistor in series with the capacitor pin 1 and 8 will reduce the gain. Figure A shows a gain = 50 configuration. You'll have to choose the LM386 set gain to suit your particular regenerative receiver, however, the greater the gain setting, the greater the chance of distortion, unwanted noise and audio feedback.

I generally build my audio power amps around op amps or discrete transistors, but the LM386 exalts this web page's theme. Distortion in all these small power amplifiers is dependent on input signal amplitude as much as anything else.

LM386 Motorboaters

Expect AF feedback motorboating via your DC power lines as you increase the AF gain pot in *many* LM386 circuits. If this happens, try better bypass and decoupling on the LM386 power line: 100 to 470 μF shunt bypass on the DC line + a 10-22 series decoupling resistor, plus 100 - 470 μF shunt bypass on pin 6. Click for 1 example. ypically 470 μF bypass is required if you suffer 60 hertz hum. You may model your simple RC filter with application E on the javascript applet page to see the 3 dB cut off frequency of your particular DC line low-pass filter.

Stupid—Simple Notes

The Stupid—Simple circuit really needs the adjustable 10K regeneration control if you wish to tune both AM and CW. The number of turns on the feedback winding varies with factors including transistor beta, how you wind the primary and secondary windings (greatly affects the coupling between the primary and secondary windings) and whether you want AM, or CW reception — or both. Experiment with the number of turns to figure it out.

You can try "matching" the tank circuit to your antenna by decreasing the 470 pF cap to as low as 68 pF. This will affect the tuning capacitor range. For an air variable tuning capacitor, use anything you can find. Consider connecting fixed parallel and/or series capacitors to reach or limit the desired capacitance. Many good examples are published on the web.

Some builders float the tuning capacitor across the inductor so the cap is ungrounded. The stator (body) of the capacitor should be connected to the circuit ground to help minimize the effects of hand capacitance. A grounded metal case further helps.

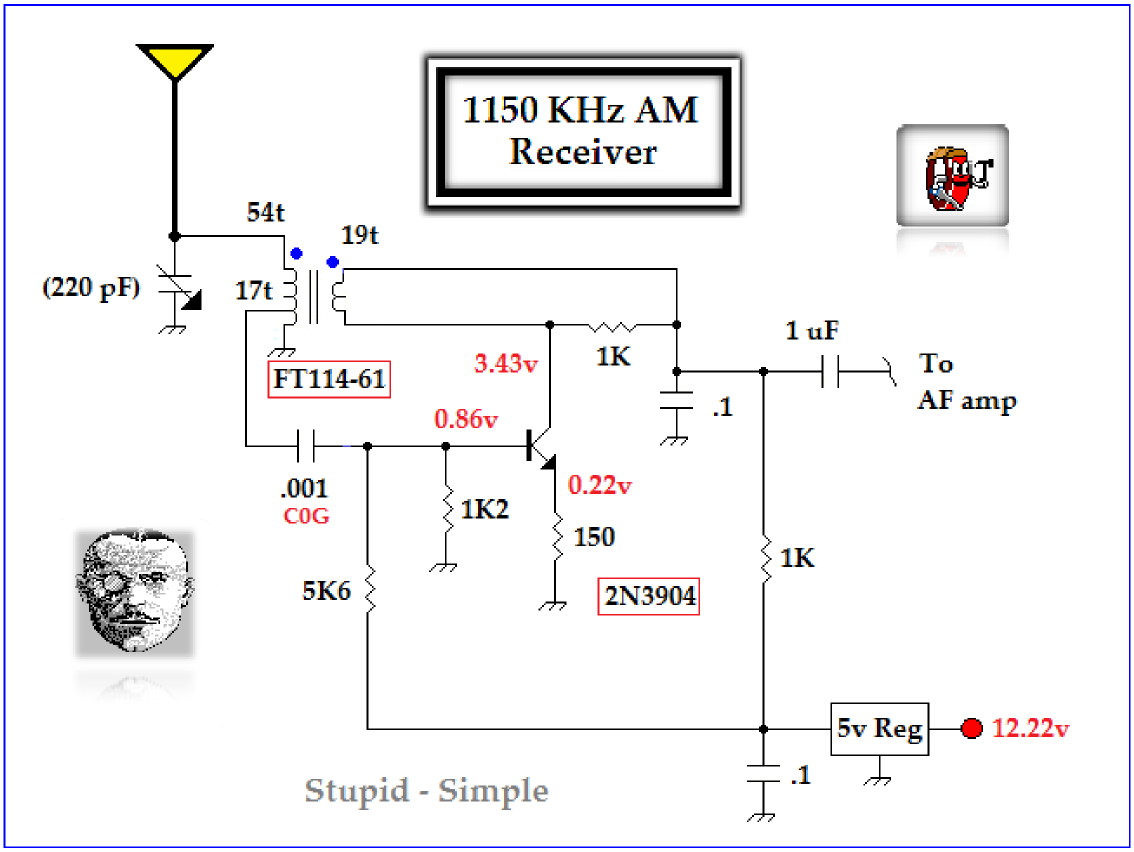

Above — My popcorn regenerative design "The Stupid-Simple" set up for broadcast band radio at 1150 KHz.

A reader from Brazil enquired about putting the Simple Stupid on MW. I had some time for a couple of experiments but only wanted 1 frequency — 1150 KHz, the local 10 KW sports radio station. Using 28 gauge wire, I wound 54 turns (about 230 uH) on a A FT-114-61. Most builders won't have this toroid, however ferrite rods from AM radios are plentiful and a great substitute.

This design relaxed the regeneration to improve audio quality (no whining or hissing). The bias and feedback loop were wound for the best sounding audio. For example, at the bias shown, if you increased the 19 turn link to 21 turns, the bass response increases; decreasing to 17 turns reduces the bass response.

As a result of lowered positive feedback, the selectivity is down, however, a variable capacitor is needed to peak the station. After peak tuning, I removed and measured the air variable cap at ~ 200 pF and then substituted a 220 pF fixed capacitor to simplify things.

For audio, I used a bench AF power amp into a speaker. It sounds nice for 1 transistor. Audio sample

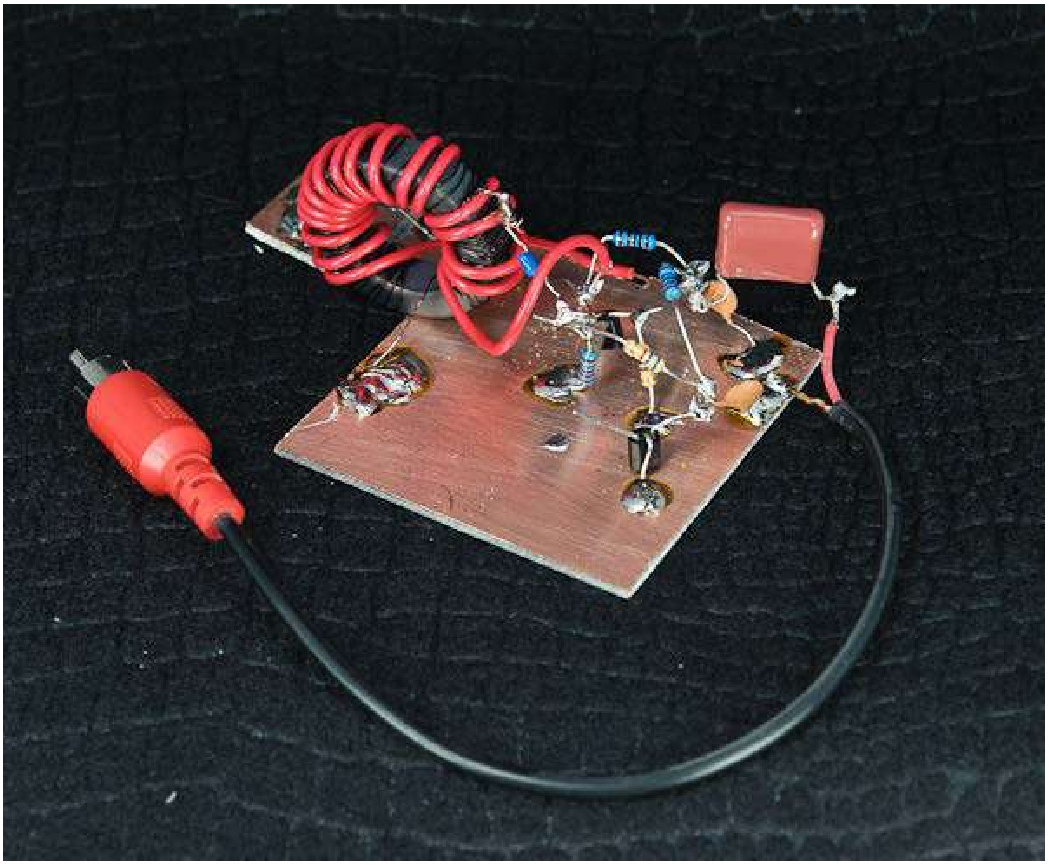

Above — The breadboard of the 1 channel receiver for my workshop - I'll use it to keep track of the Canadian Football League statistics. My test antenna was a long piece of outside wire. Red hook-up wire forms the 19 turn feedback winding.

For cities with multiple AM stations (AM stations are dying out in Canada), you'll have to add more regeneration and probably move to a better design. This radio is simple, but not extraordinary.

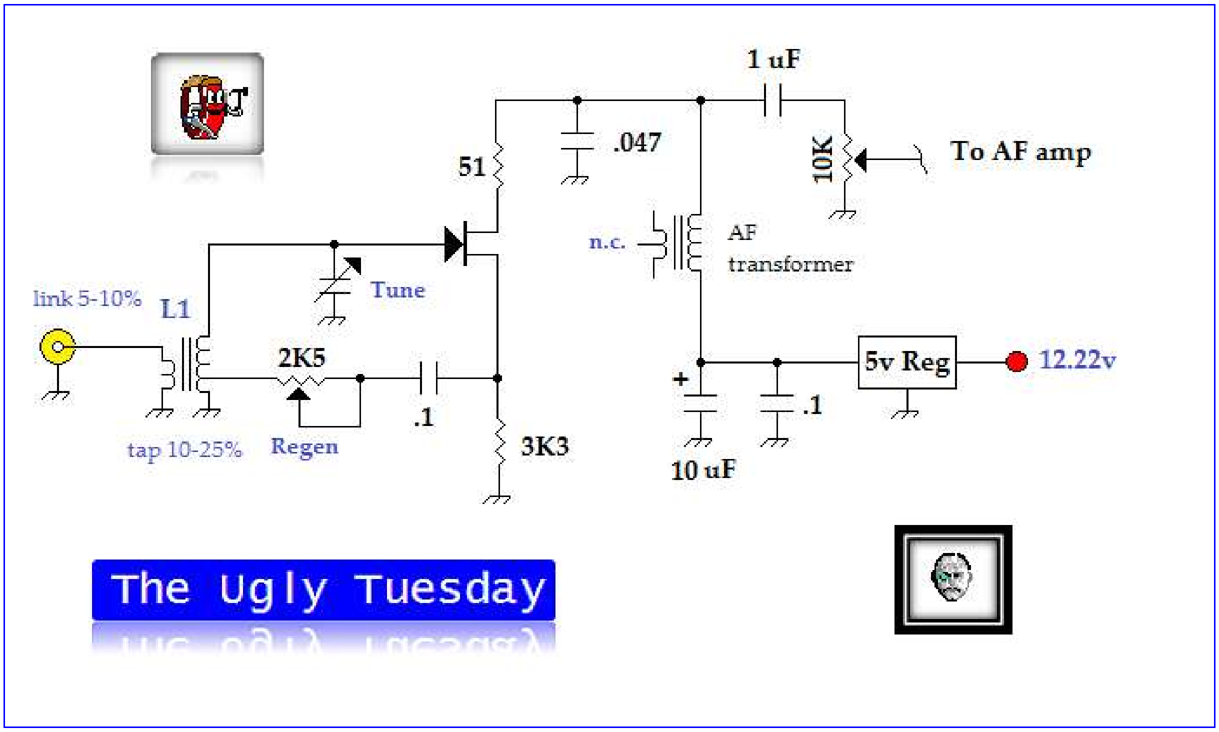

Above — Another regenerative design that tuned AM, SSB and CW from 5-11 MHz with different toroid coils wound on a T50-6. It's based on a favorite design by Wes, W7ZOI. I suggest tapping L1 at 10 - 25% of the total number of turns. The secondary link for the antenna connection depends on the impedance of the antenna, but 5 - 10% of the total number of L1 primary turns worked well at my QTH. Please experiment with the secondary link to determine the optimal coupling to your antenna. My L1 inductance ranged from 1.5 to 5.6 uH. You may have to add a fixed capacitor in parallel with your air variable capacitor when using low inductance coils such as 1.5 uH.

The 51 ohm resistor suppresses UHF parasitic oscillations. The AF transformer is a transistor radio output (1000 : 8 ohm) junk box special and serves as an RF choke. I tried various AF transformers harvested from old transistor radios in this slot and they all worked fine. Nothing's really critical on this receiver — that's why I like it. Truly junk box radio.

Simple Active Antenna Experiments

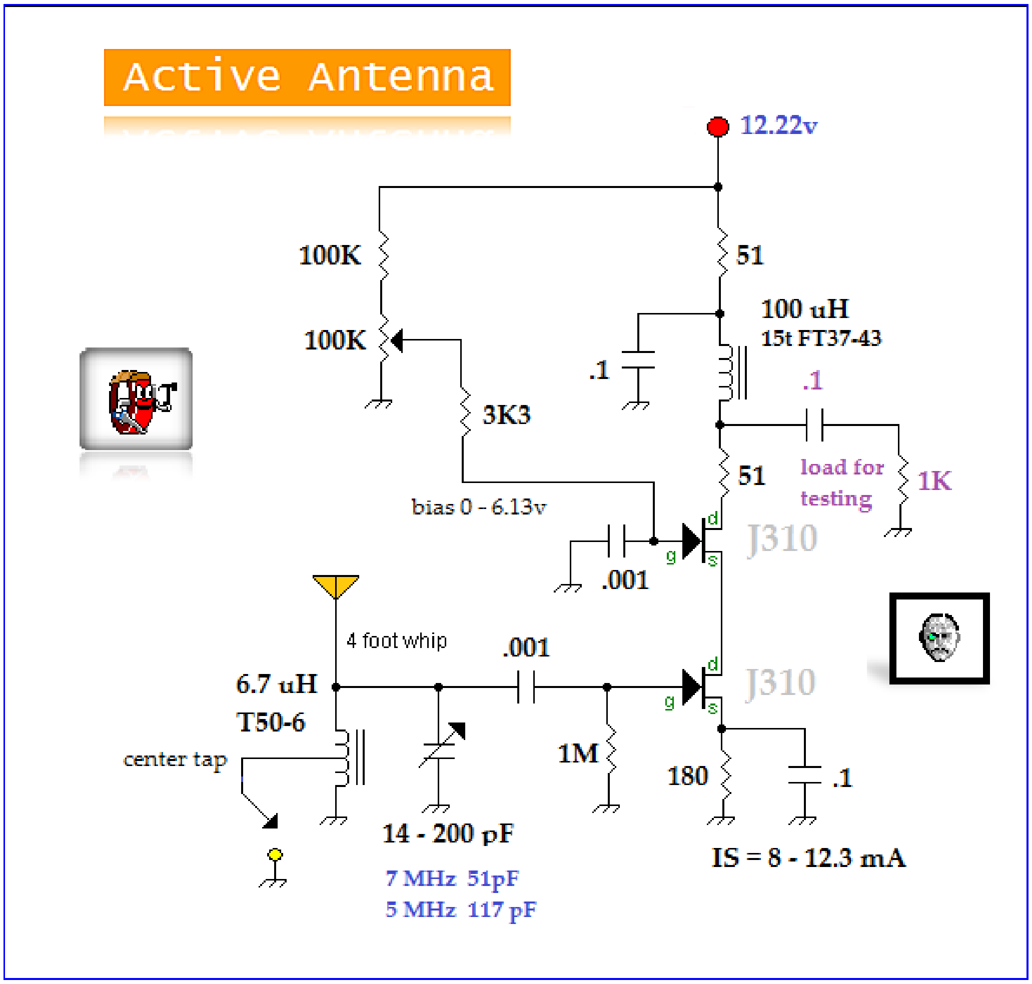

Above — A voltage probe or active antenna using a telescopic whip. It's been awhile since I built one in keeping with a minimalistic circuit theme. I tested this VPA from 5 to 14 MHz. The center tap on the coil allows the peaking at ~10 MHz and higher. The L value is non-critical; choose a value that will work with your tuning variable capacitor or varactor. The L - C values can be *roughly* determined from a chart like this, or just do the math (XL = XC at the desired frequency). Account for stray inductance. If you wish to perform return loss measurements on this circuit, you'll have to short the 6.7 uH inductor as the whip antenna can tune in RF from the RF signal generator used for the return loss bridge.

Without the 1K load, the circuit will oscillate. I thought about some ways to match the output transformer to a regenerative receiver tank circuit. The 100 uH drain choke could be replaced by a (bifilar) 2:1 transmission line transformer or two. Transformer experiments this Winter clearly illustrated the superior coupling of transmission line transformers and mandates using them over conventional transformers whenever possible.

Using a conventional transformer with a shunt resistor across the transformer would also work, but the resistor reduces gain. I built and tested this output circuit with an 8:1 transformation using two 2:1 transmission line transformers. The output impedance at the JFET drain is somewhere around 4300-4500 ohms at 7-14 MHz. The transformed output impedance is somewhere around 250-330 ohms at 7- 14 MHz. Connecting the VPA output to a tap in the regenerative main tuning inductor might work — being careful not to load down the regenerative tank coil.

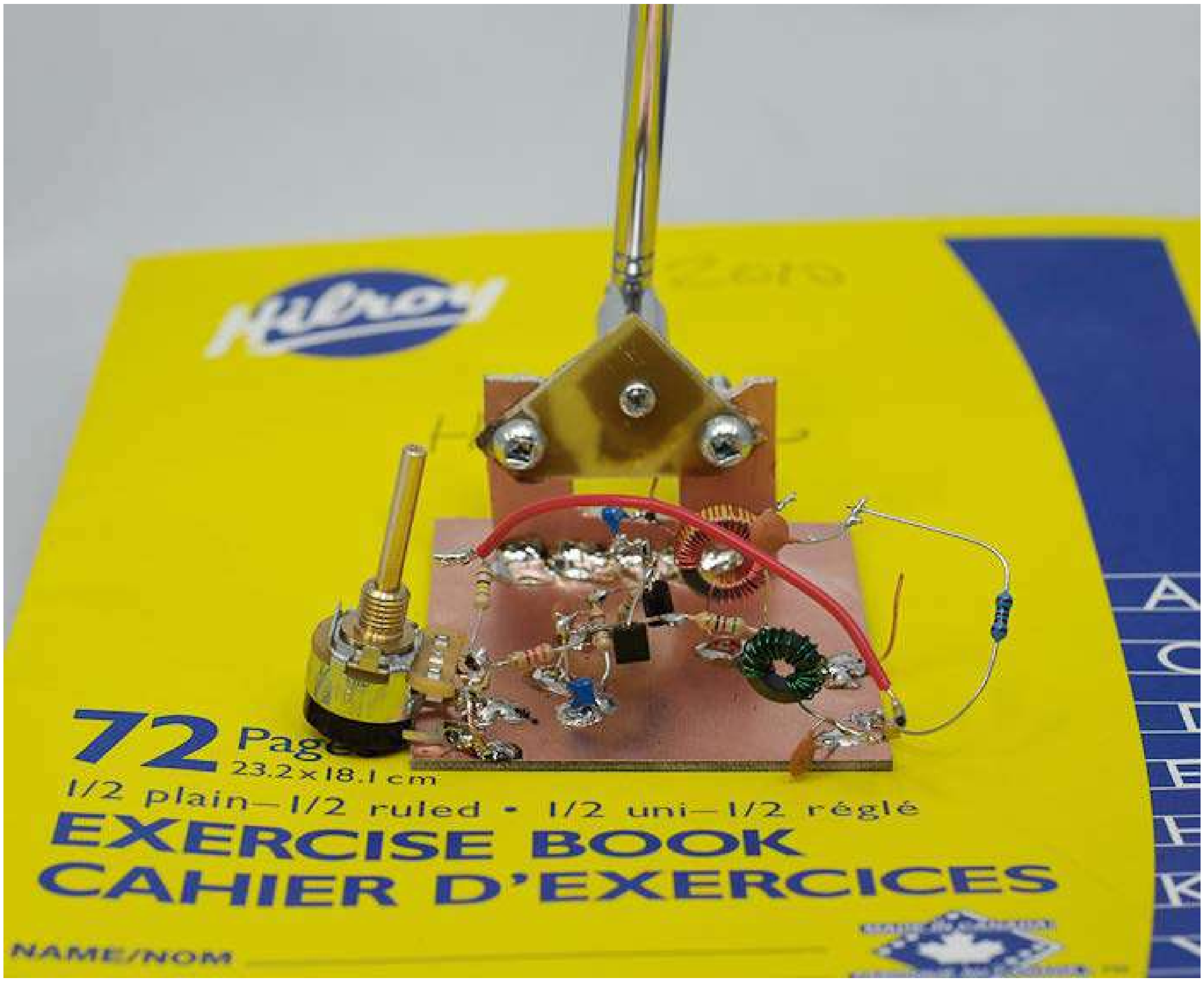

Above — The VPA breadboard on my latest notebook. The 100K pot sets the stage gain. While simple, it works okay.

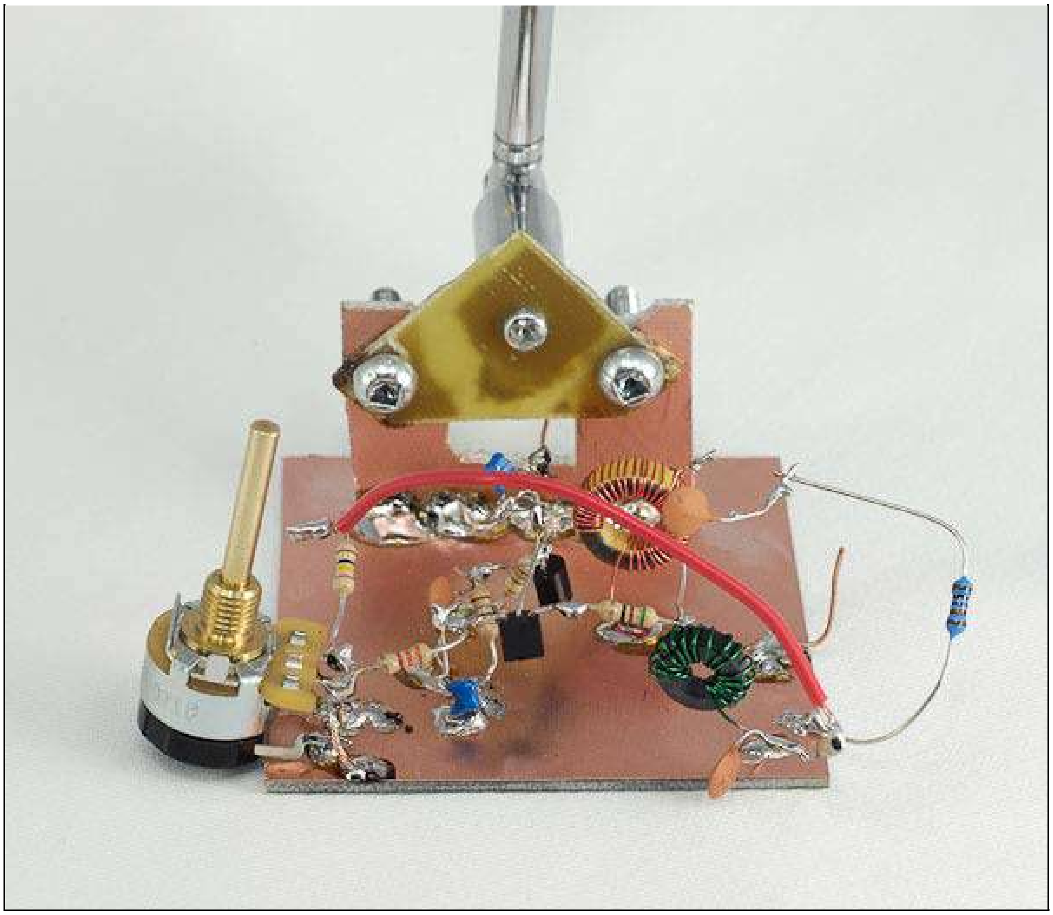

Above — the VPA built March 12, 2011 (the 14-200 pF air variable cap is not shown). It took about 45 minutes to design, build and test it. The 1K load used for testing is the blue resistor to the extreme right.

Low-pass Filter for 21 MHz

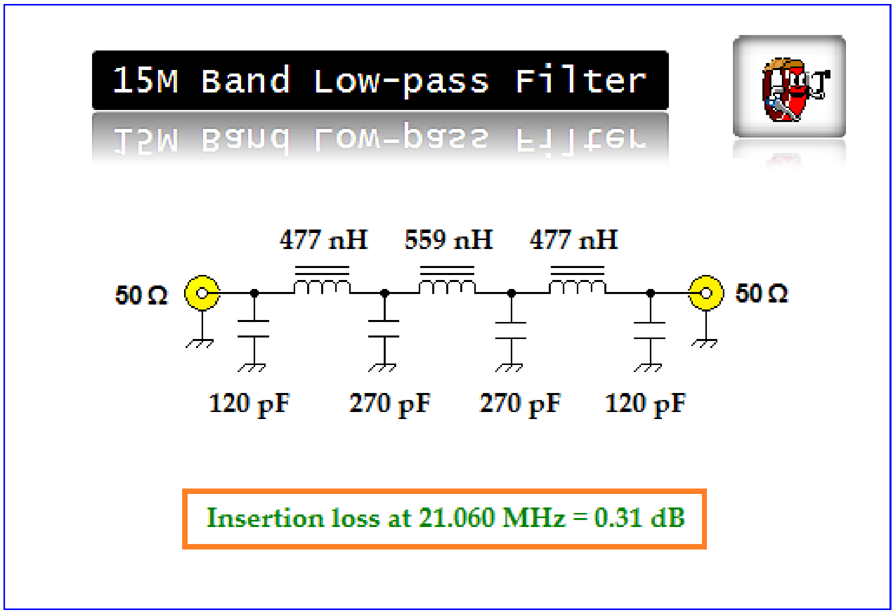
15M Band Low-pass Filter
477 nH
559 nH
477 nH
50 Ω
50 Ω
120 pF
270 pF
270 pF
120 pF
Insertion loss at 21.060 MHz = 0.31 dB

Above — A 7 element Chebyshev low-pass filter for the 15 Meters Ham band (fCo = 25.03 MHz to allow the use of standard value capacitors). A builder requested a band-pass filter design for his 15M band receiver. In order to accurately test my design, I decided to make a permanent, low-pass filter module to follow my signal generator. 15 Meters is a favorite Ham band, so I'm certain to use it in the future.

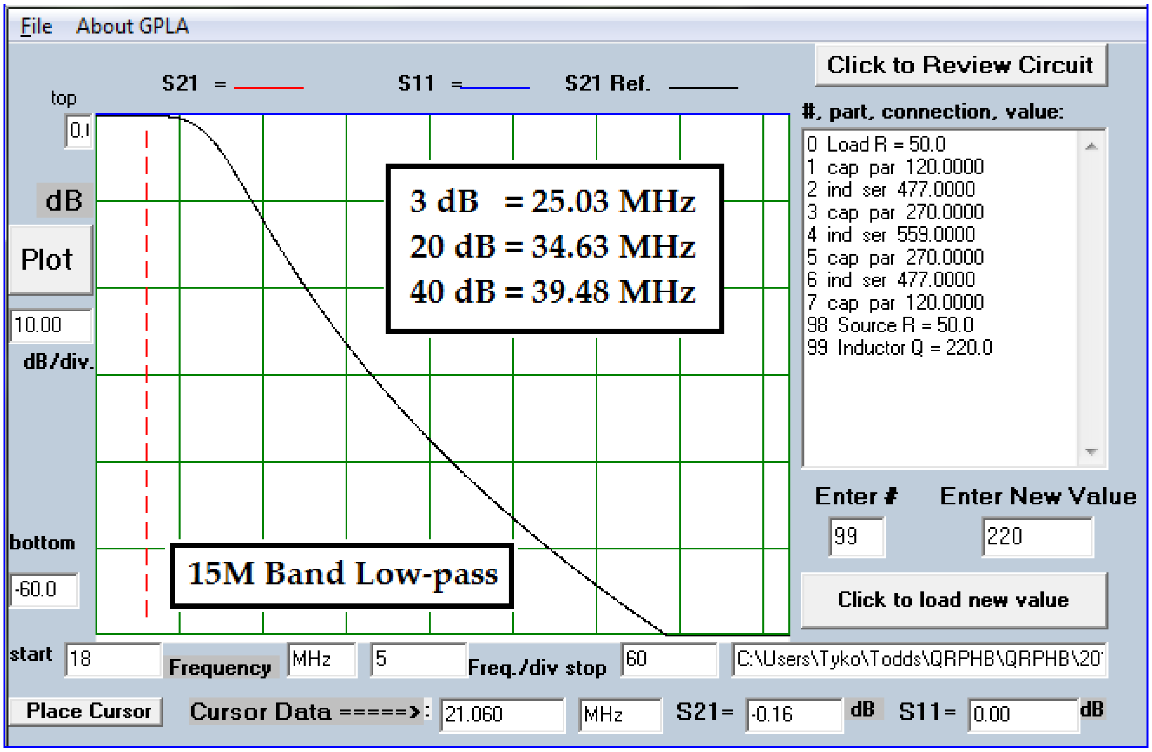
File
About GPLA
S21 =
S11 =
S21 Ref.
Click to Review Circuit
top
dB
Plot
dB/div.
bottom
start
Frequency
Freq./div
stop
Place Cursor
Cursor Data =====>:
Click to load new value

3 dB = 25.03 MHz
20 dB = 34.63 MHz
40 dB = 39.48 MHz

15M Band Low-pass

top: 0.0
dB/div.: 10.00
bottom: -60.0
start: 18 MHz
Freq./div: 5
stop: 60
C:\Users\Tyko\Todds\QRPHB\QRPHB\20

#, part, connection, value:
0 Load R = 50.0
1 cap par 120.0000
2 ind ser 477.0000
3 cap par 270.0000
4 ind ser 559.0000
5 cap par 270.0000
6 ind ser 477.0000
7 cap par 120.0000
98 Source R = 50.0
99 Inductor Q = 220.0

Enter #: 99
Enter New Value: 220

Cursor Data =====>: 21.060 MHz S21= -0.16 dB S11= 0.00 dB

Above — A GPLA plot of the filter. The frequency cut-off at -3, -20 and -40 dB are shown. Perhaps this filter is overkill, but I had all the parts on hand and love a serious low-pass filter. Click for the bread board photo.

Fine-Tuneable 1 KHz Wein Bridge Oscillator

FINE TUNING
AF GAIN

In 2010, I wanted a fine-tuneable Wien bridge oscillator to drive a notch filter in an AF distortion analyzer. Ken Kuhn drew me up a schematic on his coffee break and emailed it the same morning.

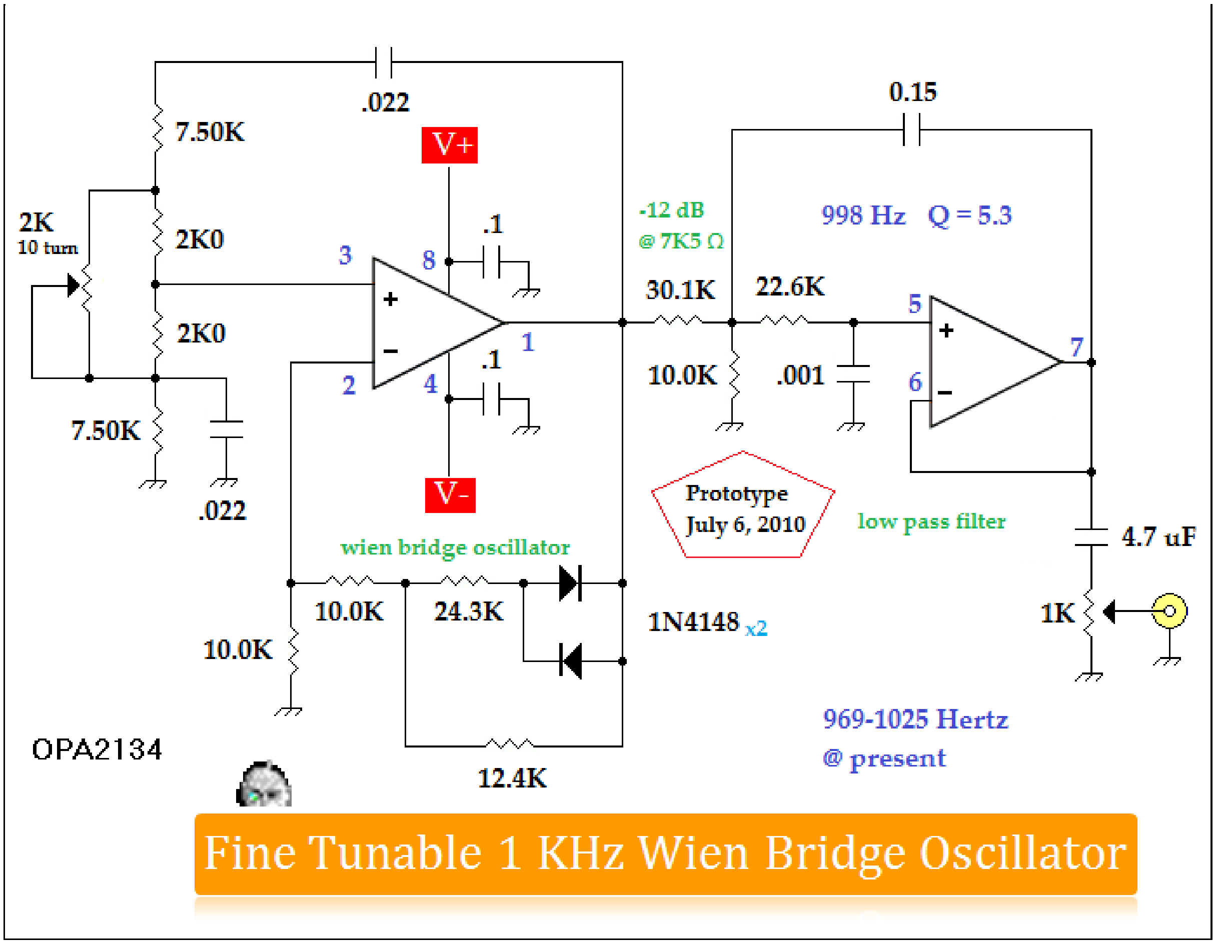
.022
7.50K
V+
0.15
2K
10 turn
2K0
3
8
.1
-12 dB
@ 7K5 Ω
998 Hz Q = 5.3
30.1K
22.6K
5
+
−
1
2
4
.1
2K0
10.0K
.001
6
7
7.50K
.022
V-
Prototype
July 6, 2010
low pass filter
4.7 uF
wien bridge oscillator
10.0K
24.3K
1N4148 x2
1K
10.0K
969-1025 Hertz
@ present
OPA2134
12.4K
Fine Tunable 1 KHz Wien Bridge Oscillator

I applied anti-parallel diodes instead of the classic incandescent bulb for amplitude stabalization in the feedback loop; probably a mistake leading to higher distortion. Ken's fine-tuning circuit works perfectly. I matched the 7K5 + 2K0 + 0.22 uF components on each filter half. 1% parts go in this circuit.

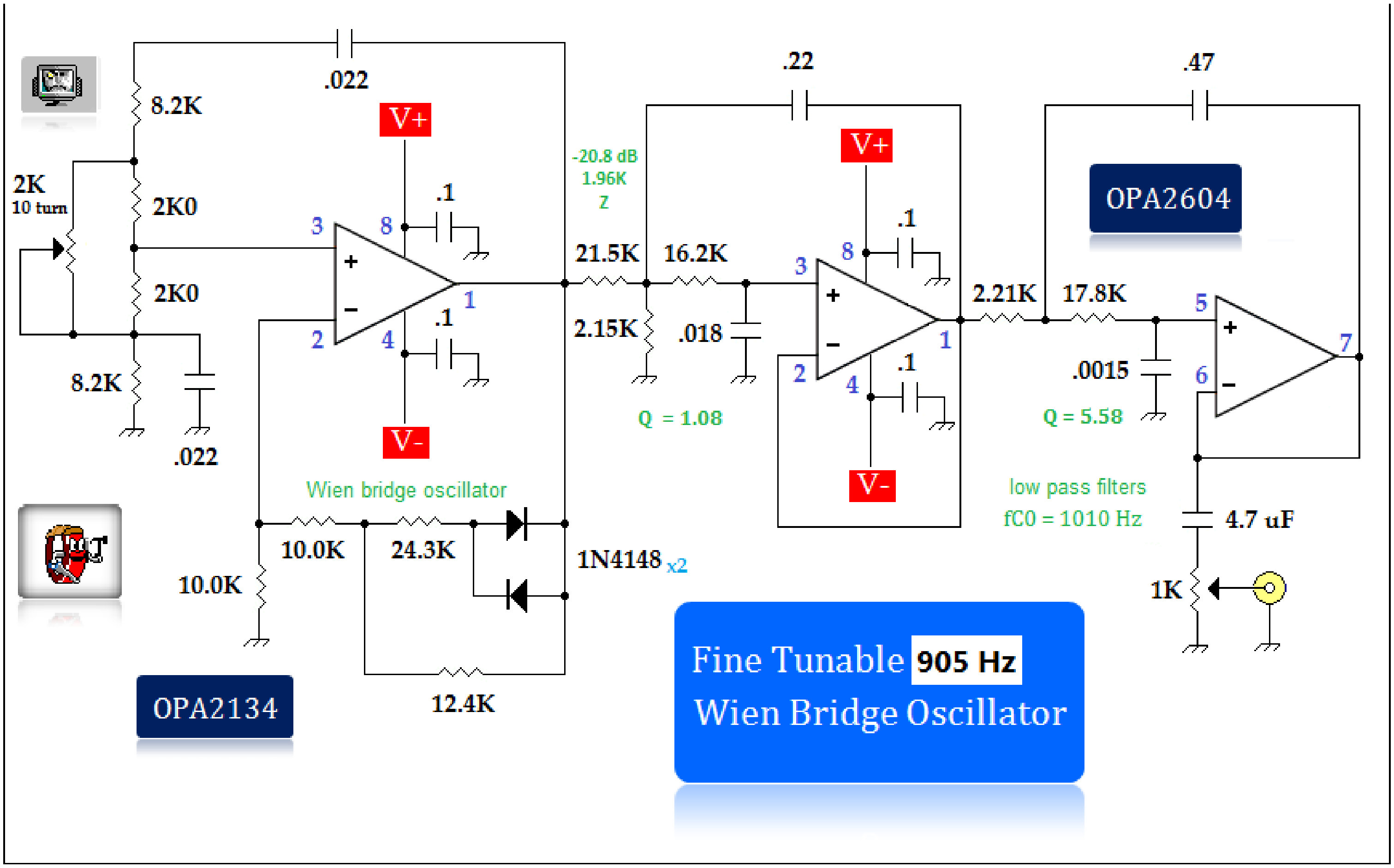
.022
8.2K
V+
2K
10 turn
2K0
3
8
.1
2K0
+
−
1
2
4
.1
8.2K
V-
.022
Wien bridge oscillator
10.0K
24.3K
10.0K
1N4148 x2
12.4K
OPA2134
-20.8 dB
1.96K
Z
21.5K
16.2K
2.15K
.018
Q = 1.08
.22
V+
.1
8
3
1
2
4
.1
V-
2.21K
17.8K
.0015
Q = 5.58
low pass filters
fC0 = 1010 Hz
.47
OPA2604
5
6
7
4.7 uF
1K
Fine Tunable 905 Hz
Wien Bridge Oscillator

A version for 905 Hertz built with crazy expensive op-amps. I chose 905 Hz to match my notch filter frequency.

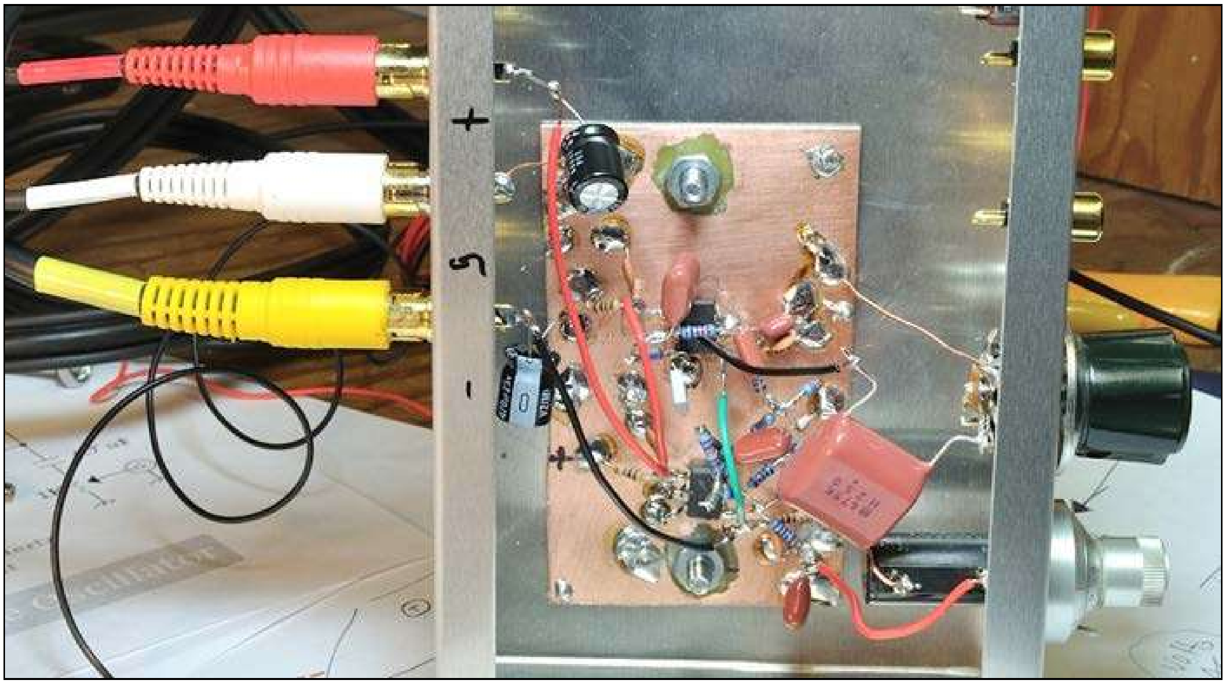

My breadboard.

RF — Test and Measurement

QRP Modules 2011

Introduction

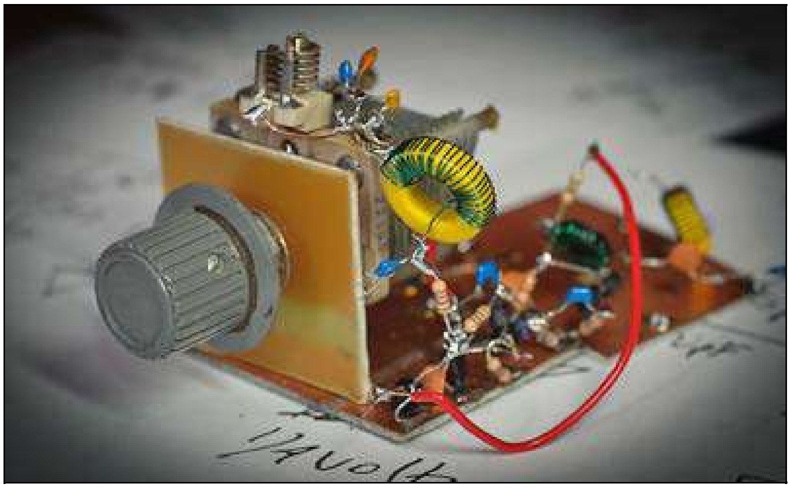

As experimenters, we rebuild core circuits over time. I decided to increase my collection of stock modular circuits to avoid re-inventing the wheel. This web page serves as a module repository for the website.

Since our needs differ, I've shared these circuits more for interest sake and really not as schematics to copy. All modules were carefully built and tested.

1 great virtue of the metal encased module is strong shielding. RF modules use a 50 ohm port impedance and BNC connectors. RCA jacks interface the AF modules.

---

40 Meter Band-pass Filter

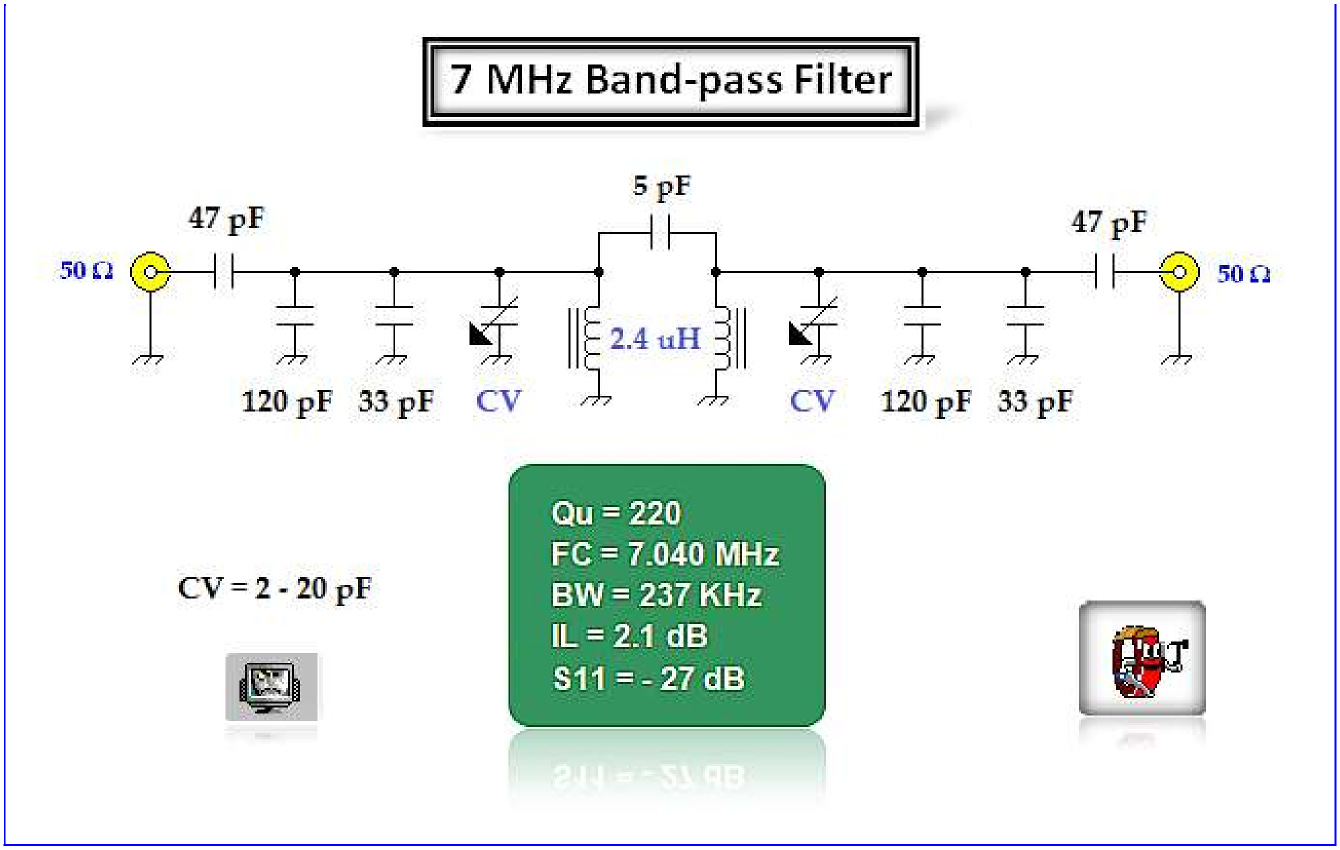
7 MHz Band-pass Filter
5 pF
47 pF
47 pF
50 Ω
50 Ω
2.4 uH
120 pF
33 pF
CV
CV
120 pF
33 pF
Qu = 220
FC = 7.040 MHz
BW = 237 KHz
IL = 2.1 dB
S11 = - 27 dB
CV = 2 - 20 pF

Above — A 40 Meter Ham band double-tuned band-pass filter. I designed this circuit using 2 programs that came with EMRFD and describe the process on this web page. The 2.4 uH measured coils were wound using #22 AWG wire on T68-6 powdered iron toroids and all fixed caps were ceramic C0G type. I centered my filter at 7.040 MHz. You should be able to peak it anywhere on the 40 Meter CW sub-band by tweaking the variable capacitors.

I peaked the trimmer capacitor while looking at the peak-to-peak voltage on a 50 ohm terminated oscilloscope. The filter input was connected to a 7.040 MHz signal generator with a 30 dB return loss, low harmonics (-55 dBc) and 50 ohm cables.

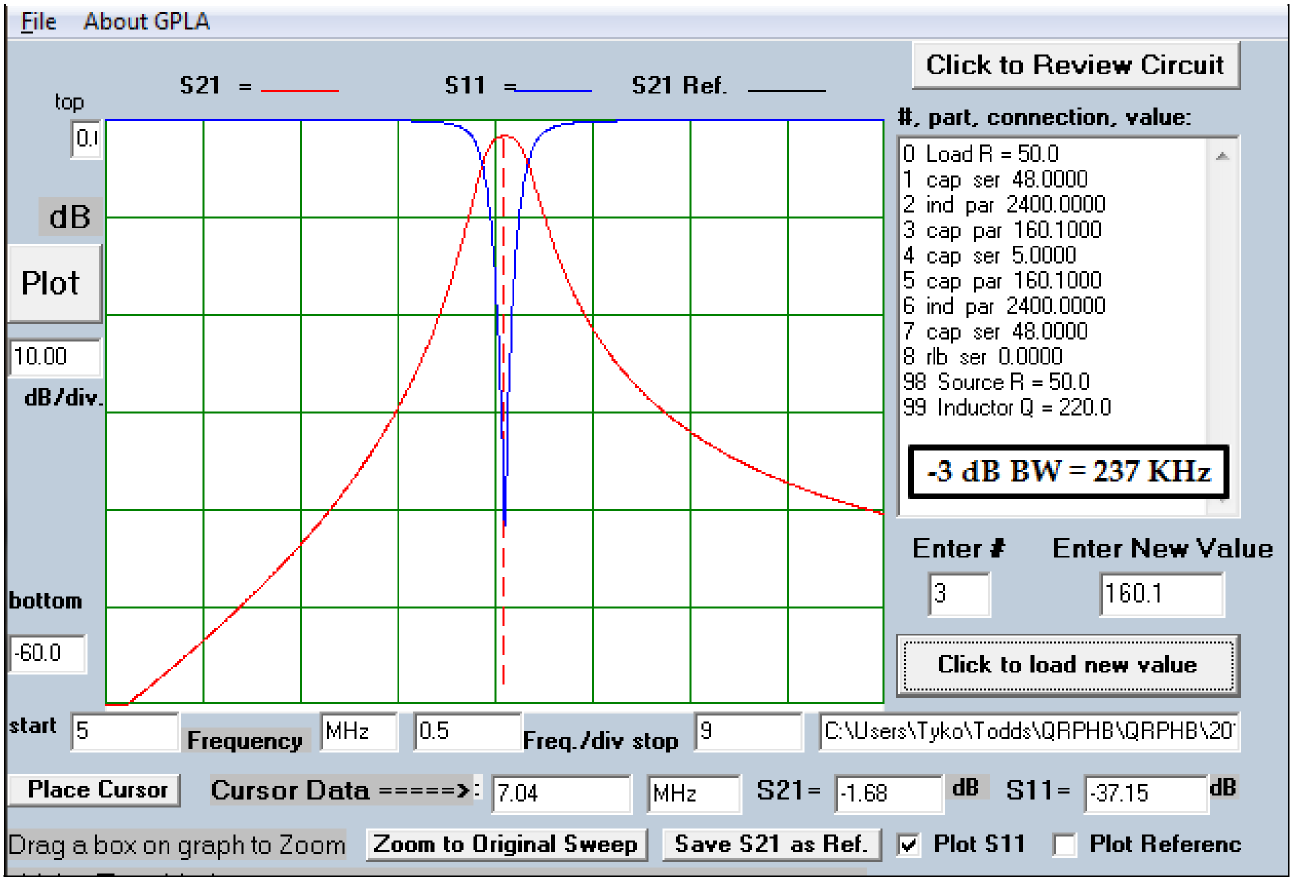
File
About GPLA
S21 =
S11 =
S21 Ref.
top
dB
Plot
dB/div.
bottom
Click to Review Circuit
Enter #
Enter New Value
Click to load new value
Place Cursor
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.

top 0.0
10.00 dB/div.
bottom -60.0

#, part, connection, value:

```
0  Load R = 50.0
1  cap  ser  48.0000
2  ind  par  2400.0000
3  cap  par  160.1000
4  cap  ser  5.0000
5  cap  par  160.1000
6  ind  par  2400.0000
7  cap  ser  48.0000
8  rlb  ser  0.0000
98  Source R = 50.0
99  Inductor Q = 220.0
```

-3 dB BW = 237 KHz

Enter # 3
Enter New Value 160.1

start 5 Frequency MHz 0.5 Freq./div stop 9
C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====> 7.04 MHz S21= -1.68 dB S11= -37.15 dB

☑ Plot S11 ☐ Plot Referenc

Above — A simulation of my filter design in in GPLA08. The calculated IL was 1.68 dB, I measured the IL at 2.1 dB The calculated return loss or S11 was 37.2 dB; I measured 27 dB. A good filter.

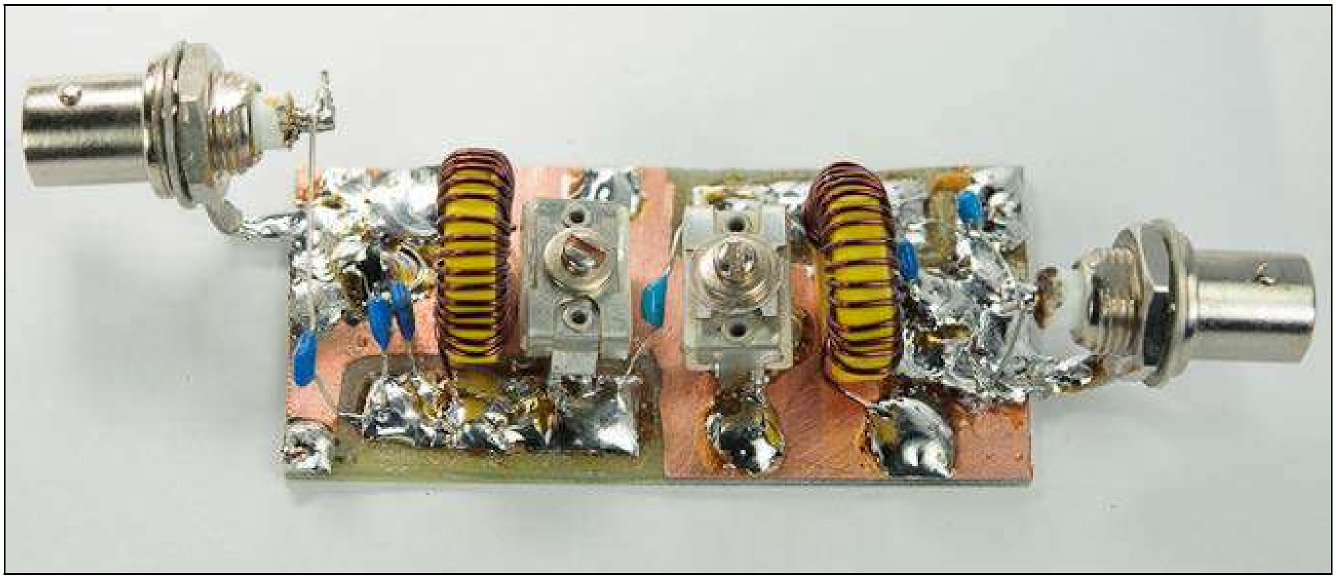

Above — the 40 meter double tuned band-pass filter breadboard with temporary BNC connectors and series caps. Since this filter will serve as my main front-end filter for all future 40M band receiver bench design, I blinged out and put in big toroids and high Q, air-variable trimmer capacitors. While I could have just use a single 150 pF tank capacitor and a wide range trimmer cap such as common, ceramic 10-70 pF, the small range, high Q trimmer capacitors offer better performance and fine tuning.

Click for a spectrum analyzer +tracking generator sweep where the center frequency = 7.040 MHz. Graticules: Horizontal = 1 MHz per division, Vertical = 10 dB per division. You can see why it tunes so sharply.

After testing the bread-board, I removed the temporary BNC connectors and series caps, I stuck it in a Hammond box and wired in permanent, short leaded 47 pF capacitors. Final testing in the sealed box varied minimally from the open bread board. This board looks especially ugly because it held a previous filter and contained lots of remnant solder.

Broadband Feedback Amp

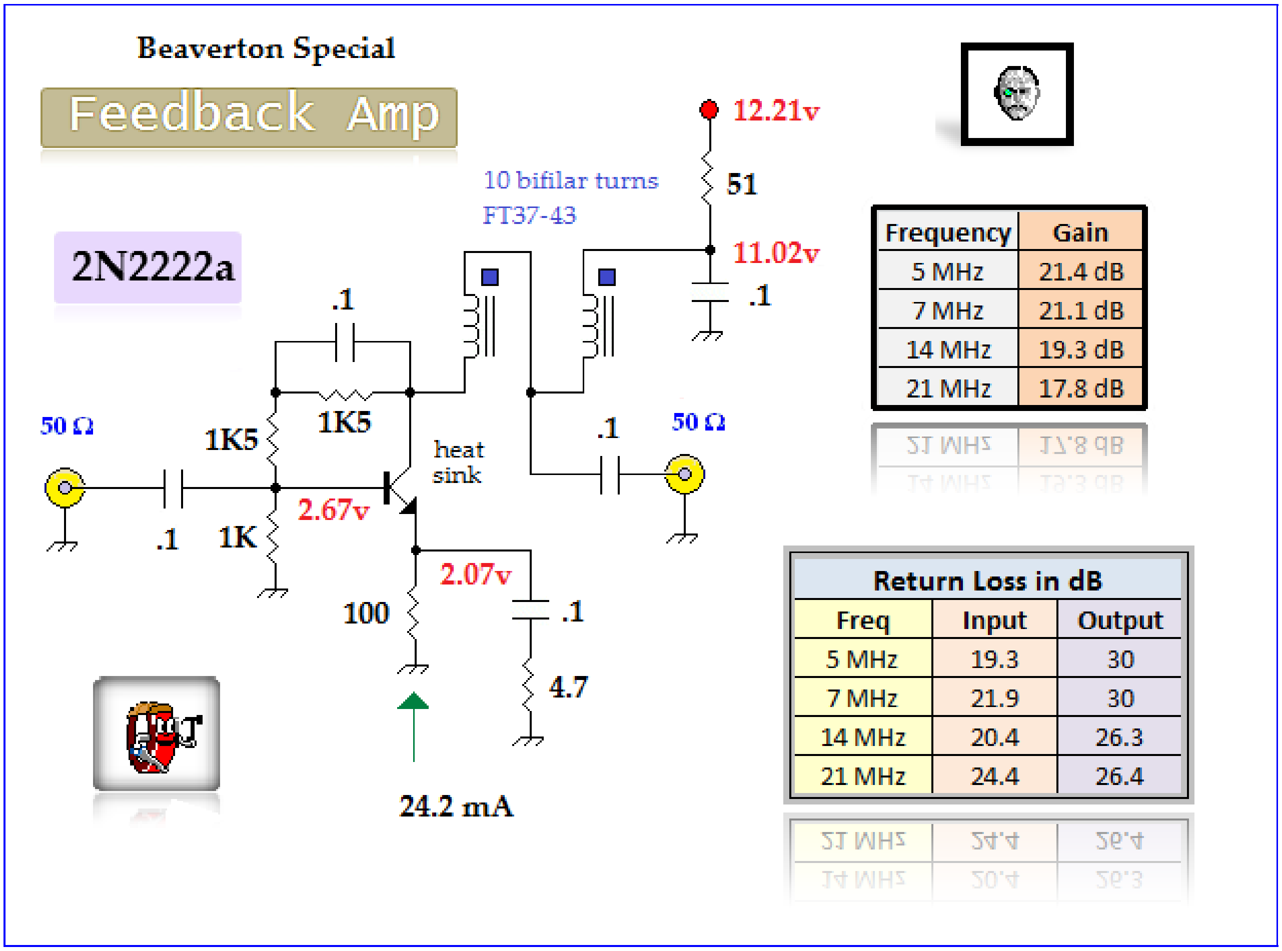
Beaverton Special
Feedback Amp
2N2222a
12.21v
51
10 bifilar turns
FT37-43
11.02v
.1
.1
1K5
1K5
heat
sink
50 Ω
50 Ω
.1
2.67v
.1
1K
2.07v
100
.1
4.7
24.2 mA

Frequency	Gain
5 MHz	21.4 dB
7 MHz	21.1 dB
14 MHz	19.3 dB
21 MHz	17.8 dB

Return Loss in dB

Freq	Input	Output
5 MHz	19.3	30
7 MHz	21.9	30
14 MHz	20.4	26.3
21 MHz	24.4	26.4

Above — A "Beaverton Special" feedback amp with analysis. As experimenters, we often need a go-to, broadband 50 ohms input and output RF amplifier. This is it!

Popularized by Wes, W7ZOI and Doug DeMaw, W1FB, this amp has stood the test of time and fits perfectly into the 50 ohm module concept. A bevy of transistors were tried — a 2N4401, 2N5179, 2N3904, 2N3866 or 2N5109 all worked fine. For the greatest return loss and signal handling possible, current over 21 mA is required and thus a 2N3904 isn't the best choice. Collector current = heat, so heat sink the BJT as appropriate.

I found that a 2N2222a biased with over 22 mA emitter current gave a stellar output return loss and low distortion. Within reason, for different transistors, keep the bias and feedback resistors constant and change the emitter resistor (100 ohms in my amp) to set the current you want or need.

Many builders follow this amp with an attenuator pad to preserve the input return loss.

Diode Ring Mixer

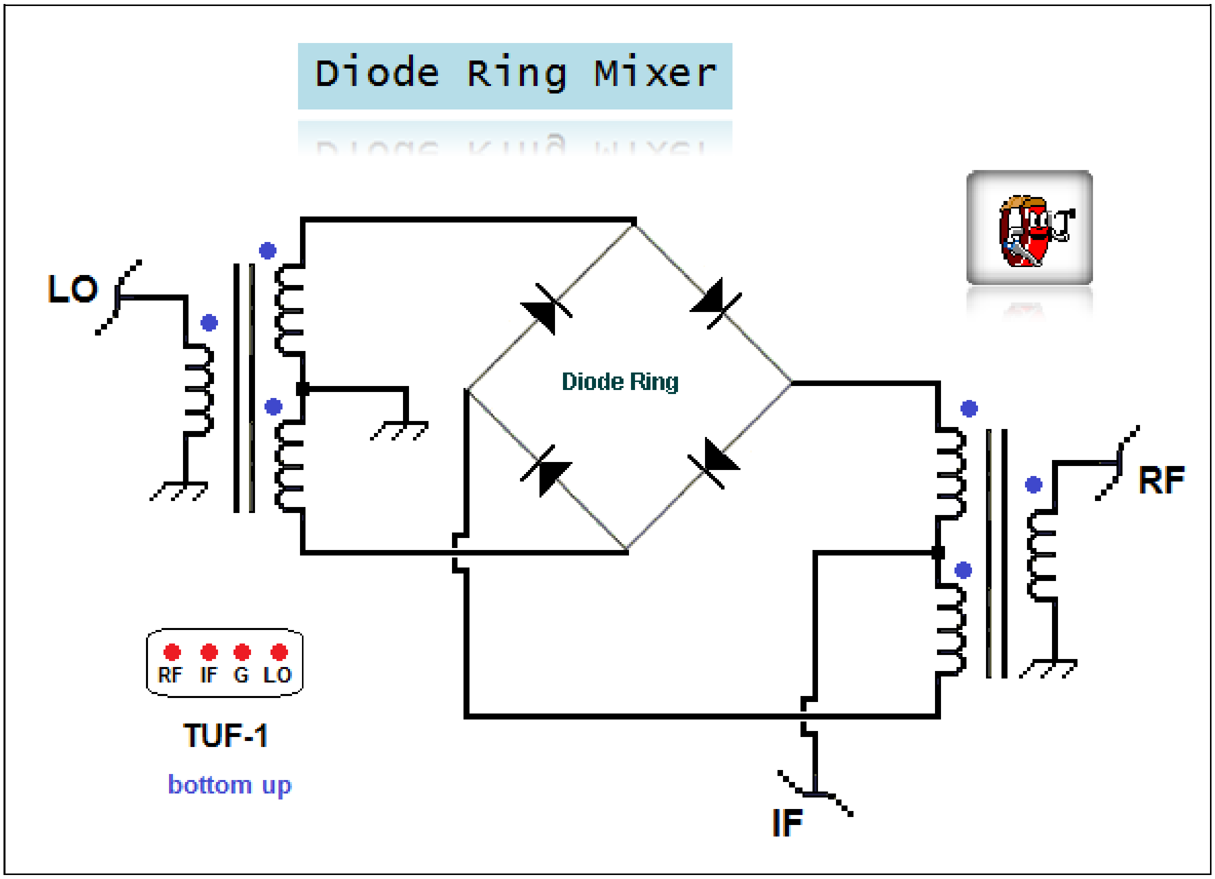
Diode Ring Mixer
LO
Diode Ring
RF
RF IF G LO
TUF-1
bottom up
IF

Above — A Minicircuits TUF-1 diode ring mixer was used in standard configuration. 7 dBm LO drive.

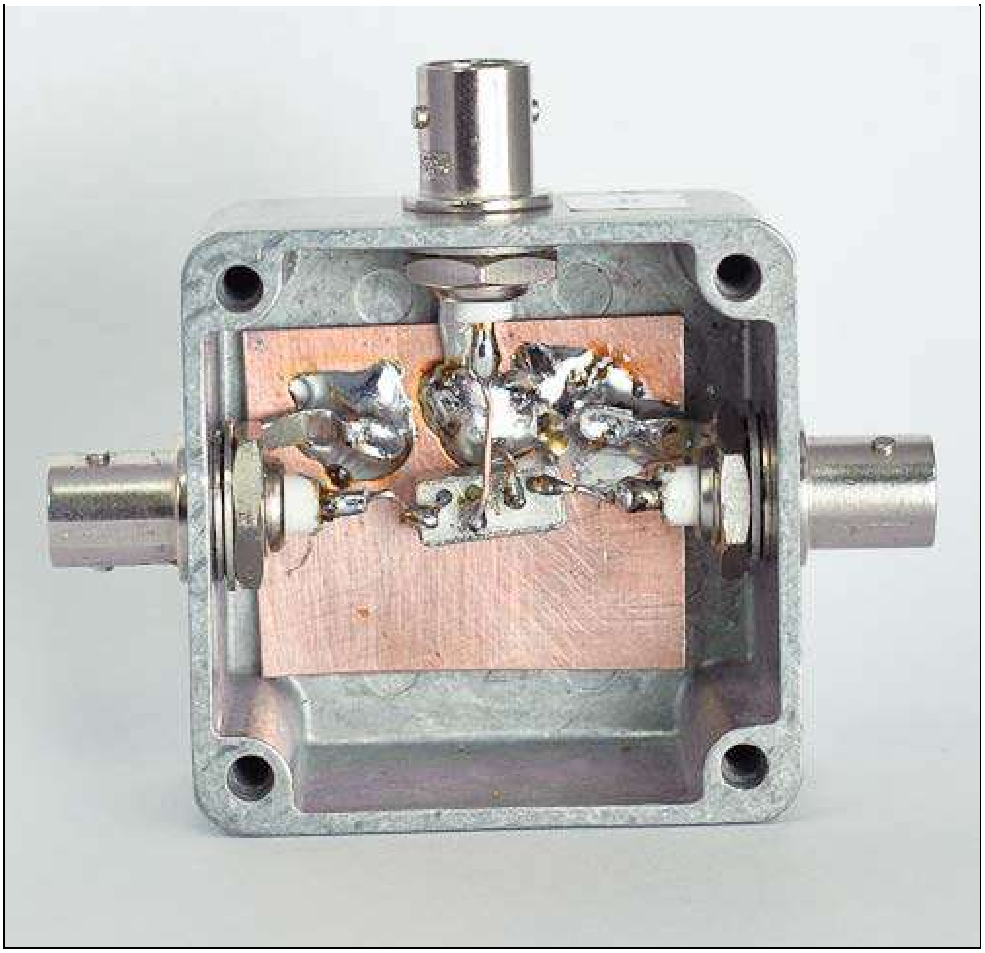

Above — The DRM module. It's hard to photograph inside a solder laden chassis. Connections are short.

Popcorn Audio Frequency Power Amplifiers

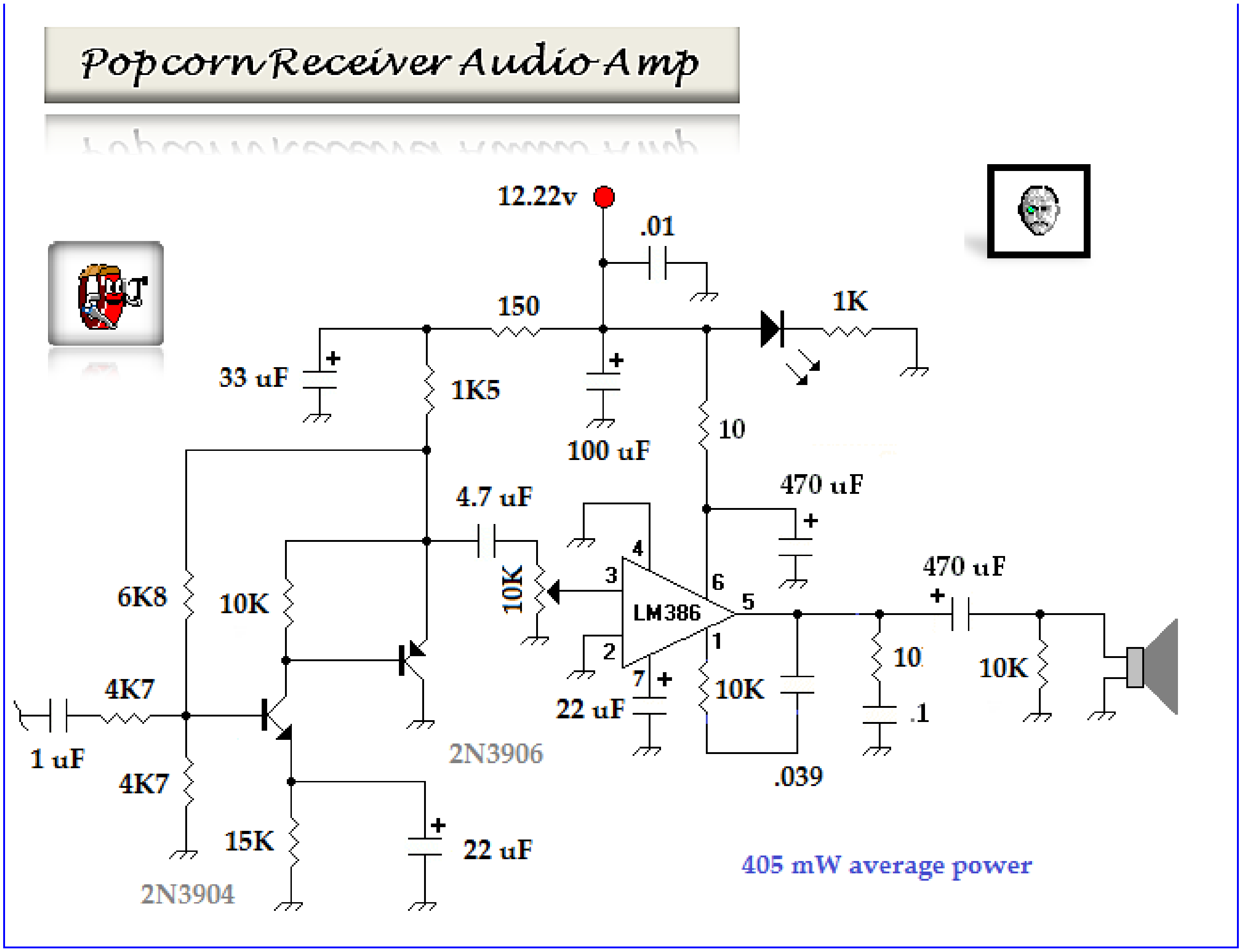
Popcorn Receiver Audio Amp
12.22v
.01
150
1K
33 uF
1K5
100 uF
10
470 uF
4.7 uF
4
3
6
5
LM386
2
1
7
10K
470 uF
6K8
10K
10K
10
10K
4K7
22 uF
10K
.1
1 uF
4K7
2N3906
.039
15K
22 uF
2N3904
405 mW average power

Above — Popcorn receiver audio power amp. I wanted a simple audio stage for testing popcorn receivers (to follow a high output impedance preamp device). Completing this module means never having to build such an amp again. The voltage gain is provided by BJTs to keep the noise down, but the popcorn factor up. The preamp impressed me with its strong signal handling capacity via feedback and careful biasing. The NPN is center biased so that when its intentionally distorted during testing, the positive and negative halves of the AC waveform distort equally — it provides a nice, big, AC voltage swing. An LM386 in X20 gain mode with some bass boosting comprises a reasonable power amp section. The 10K resistor on the output discharges the 470 uF cap when no speaker is connected to avoid a loud pop. The 4K7 series input R can be lowered, or omitted for more sensitivity.

Some builders might employ the LM386 in a higher gain mode at the expense of fidelity, or just wire up a TDA7052. I think in popcorn circuits, what really matters is that you understand what you're doing and try to *design* rather than just copy the "usual circuits".

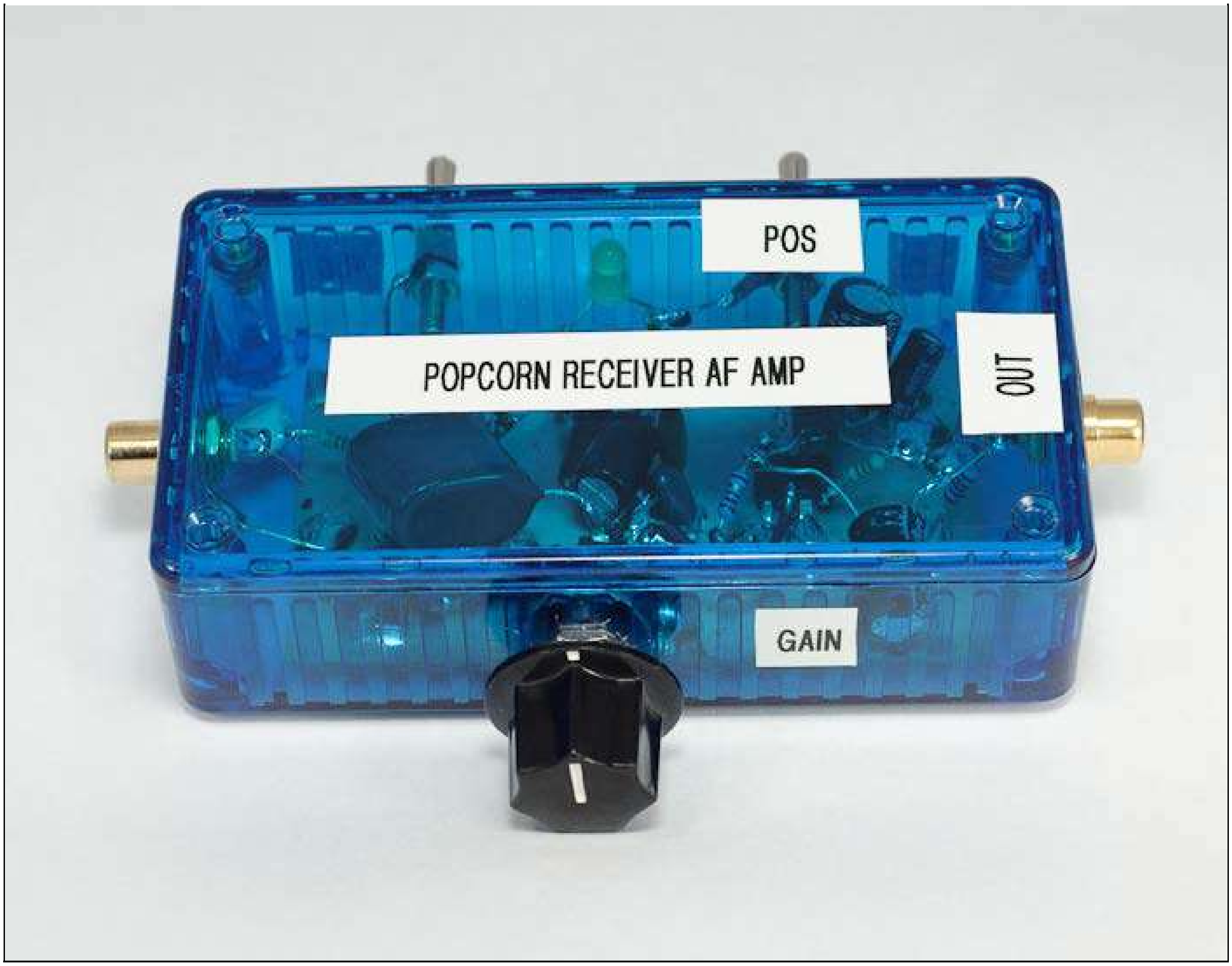

Above — The popcorn AF amp in a clear blue chassis. Phono jacks provided a connection for the input and output — they 're inexpensive and readily available. The DC supply is connected to uninsulated banana jacks on the rear; it's well decoupled (resistor) + bypassed (capacitor) to help stop parasitic AF feedback. This amp is pretty quiet, considering its junk box legacy.

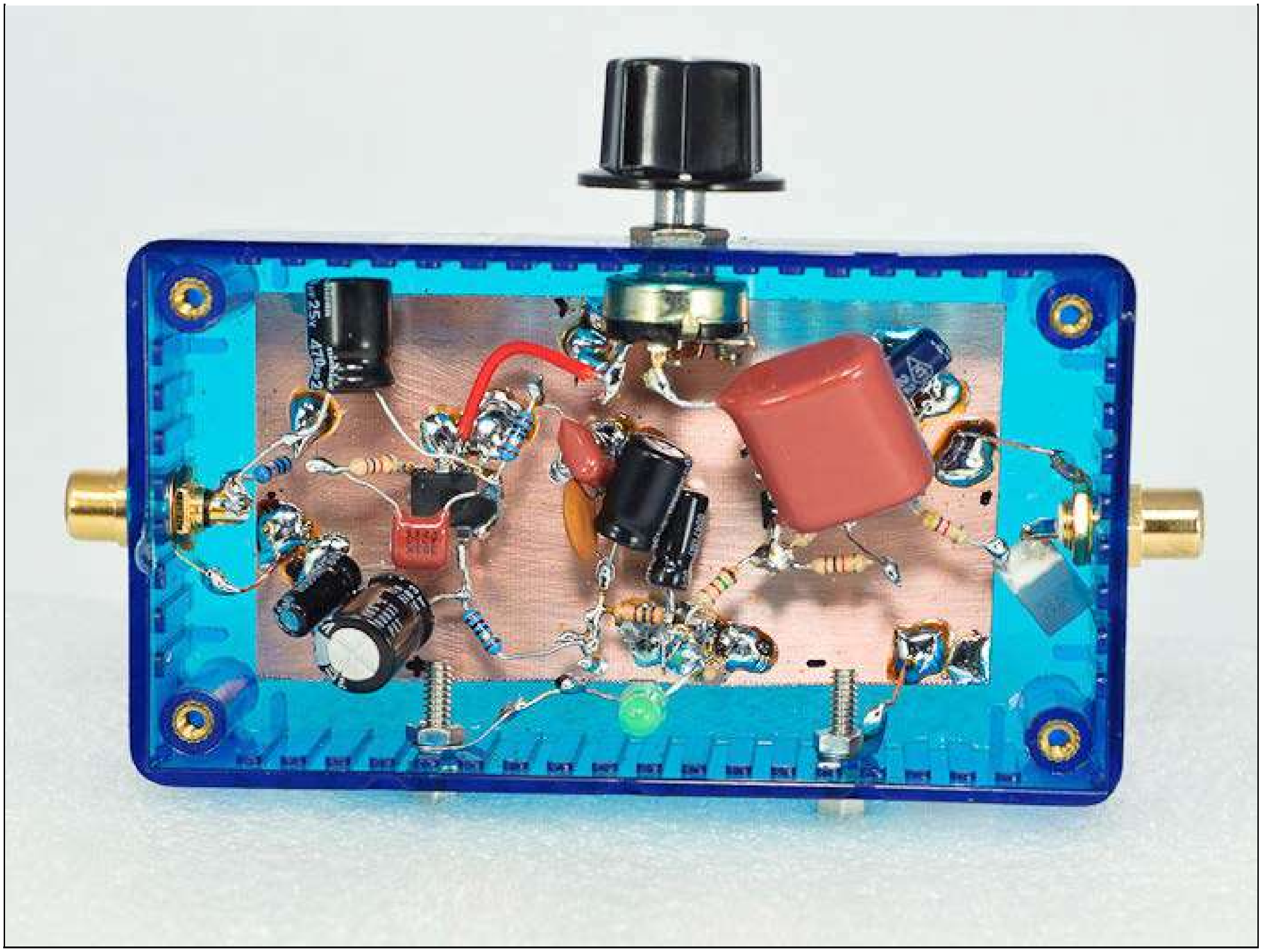

Above — The project with the top cover removed. The board is secured by the ground wires connecting it to the pot, jacks and DC voltage posts. The input is on the right.

I like a relatively simple, lower gain AF amp on the bench for receiver development. You can use such an amp to decide on how much overall AF gain is needed, how you'll distribute it, and not have to deal with unwanted AF feedback.

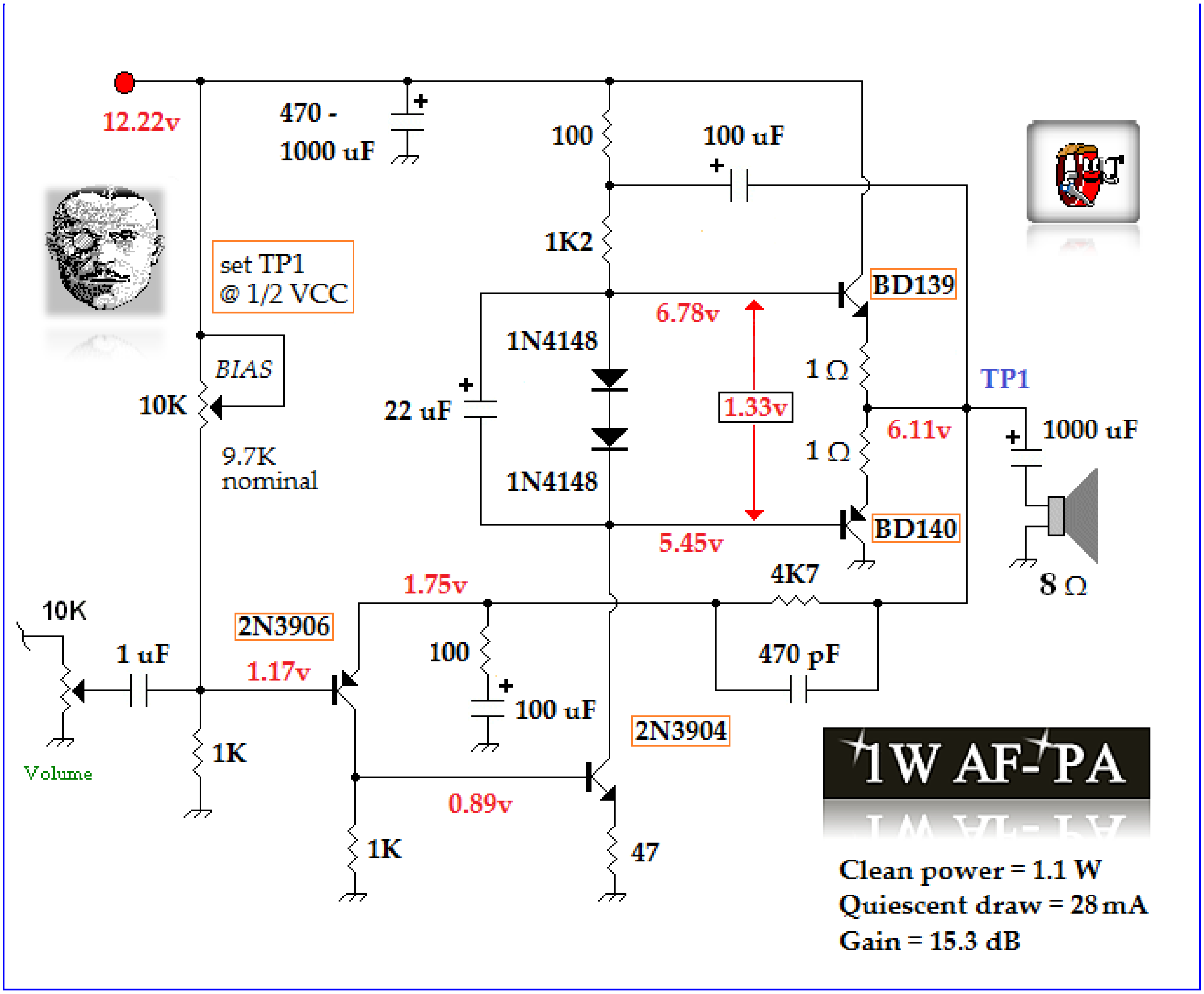
12.22v
470 -
1000 uF
100
100 uF
1K2
set TP1
@ 1/2 VCC
BD139
6.78v
1N4148
1 Ω
TP1
BIAS
10K
22 uF
1.33v
6.11v
1000 uF
9.7K
nominal
1 Ω
1N4148
BD140
5.45v
4K7
8 Ω
1.75v
10K
2N3906
1 uF
1.17v
100
470 pF
100 uF
2N3904
1K
Volume
1W AF-PA
0.89v
1K
47
Clean power = 1.1 W
Quiescent draw = 28 mA
Gain = 15.3 dB

Above — 1 watt popcorn audio power amplifier. Built around the BD139/140 complimentary pair - I achieved a clean I KHz sine wave at 1.1 Watts power after testing + tweaking my prototype design. I chose the familiar series diode pair to bias the power followers into Class A/B; an amplified diode (transistor level shifter) might be a better choice.

Bootstrapping the 2N3904 voltage amp pumps up the clean signal power capacity. The 2N3906 establishes the bias for the 2N3904 and the BD139-140 pair. Set the 10K bias pot so that the DC voltage at TP1 is 1/2 of the VCC. During testing, the AC voltage was centered perfectly between the DC rails and when pushed into clipping, the positive and negative AC waveform distorted nearly equally. Quiescent current = 28 mA ; not meant for a field-portable receiver.

Click for a photo of my breadboard. Copper clad board serves as heat sinks for the power followers. The BD139 and BD140 make great complimentary transistors for audio frequency power amplifiers. With an Ft of 190 MHz, the BD139 can work okay as a driver or even the final in modest power QRP transmitters.

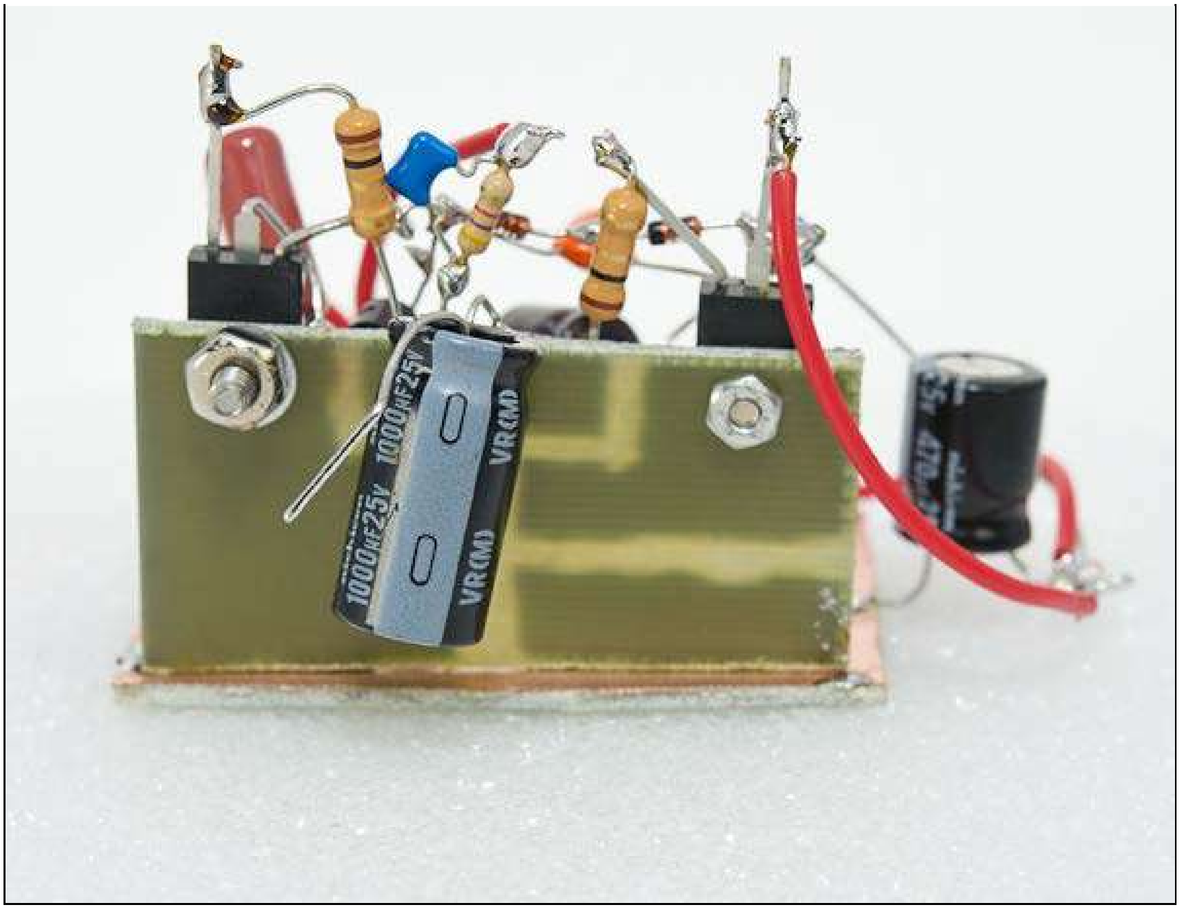

Above — 1 watt Audio PA (reverse view).

Pop DC2 — Popcorn Direct Conversion Receiver Main Frame

These circuits update the Popcorn DC receiver from 1998 and includes all components from the product detector through to the speaker, minus the VFO and band-pass filter.

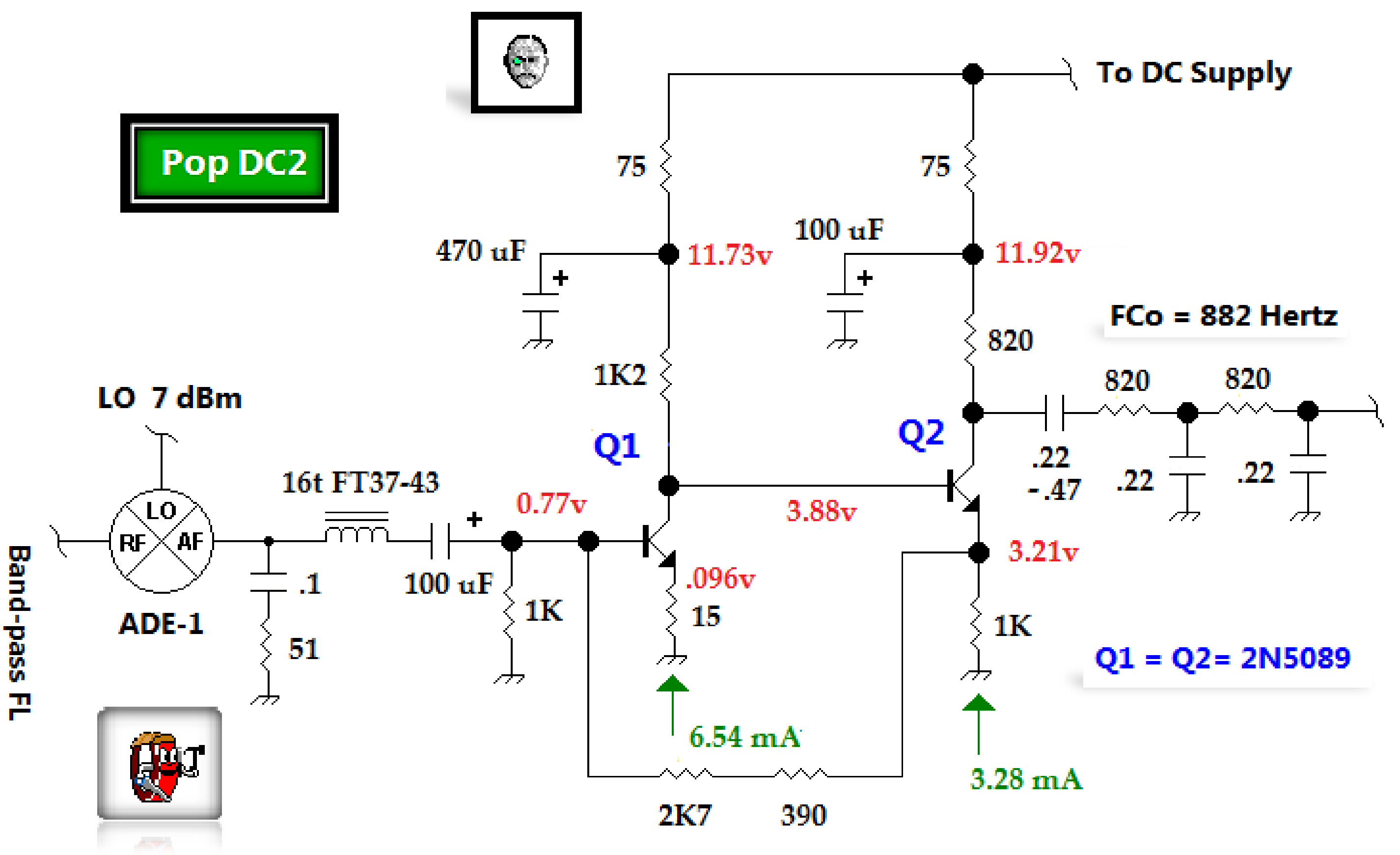
Pop DC2
To DC Supply
75
75
470 uF
11.73v
100 uF
11.92v
FCo = 882 Hertz
820
1K2
820
820
LO 7 dBm
Q1
Q2
.22
- .47
.22
.22
16t FT37-43
0.77v
3.88v
LO
RF
AF
100 uF
.1
.096v
3.21v
Band-pass FL
ADE-1
1K
15
1K
51
Q1 = Q2= 2N5089
6.54 mA
3.28 mA
2K7
390

Popcorn Direct Conversion Receiver Main Frame

Above — The mixer and first audio preamplifier The 0.22 uF to 0.47uF cap connecting Q2's collector to the low-pass filter network exerts a high-pass response to remove low frequency noise and potentially any hum. I heard no hum, although a 470 uF filter capacitor on Q1 helps ensure that. Increase the 100 uF filter capacitor filtering Q2 if you hear hum or motor boating. The diode ring mixer exhibits AF that's hard to beat — very dynamic, vibrant and lively. I enjoyed the low microphonics with the double balance + a return loss of over 25 dB on all of 3 of its ports. Alternate photo.

I've read negative comments about my use of "those big filter capacitors" — 1 thing radiophiles can learn from audiophiles is that to adequately decouple and bypass means we need to stop fooling around with the usual 22 - 47 uF capacitors and *really* bypass. Viewing well designed AF amplifiers informs us so; these designers *really* filter their amplifiers from the DC supply. You can always increase the decoupling resistor value to allow use of a smaller capacitor value, however, we only have a single power supply at around 12 VDC, and I dislike giving up too much of it for DC filtering purposes. Do what ever amuses you.

Click for some analysis of the preamp. The MPSA18 went obsolete in 2011, so I chose the low-noise 2N5089 for Q1 and Q2.

The Popcorn DC2 receiver keeps the format of the earlier version; discrete transistors for all but the power amp and R-C low-pass filtering. The filter still allows you to listen to SSB, as there aren't many poles and the cutoff is nearly 900 Hertz —it just removes the ice-pick in the ear often heard in unfiltered DC receivers. You can change the capacitor values for a different cutoff frequency. Applet E performs this function.

Second Preamplifier Stage with TDA7052 final

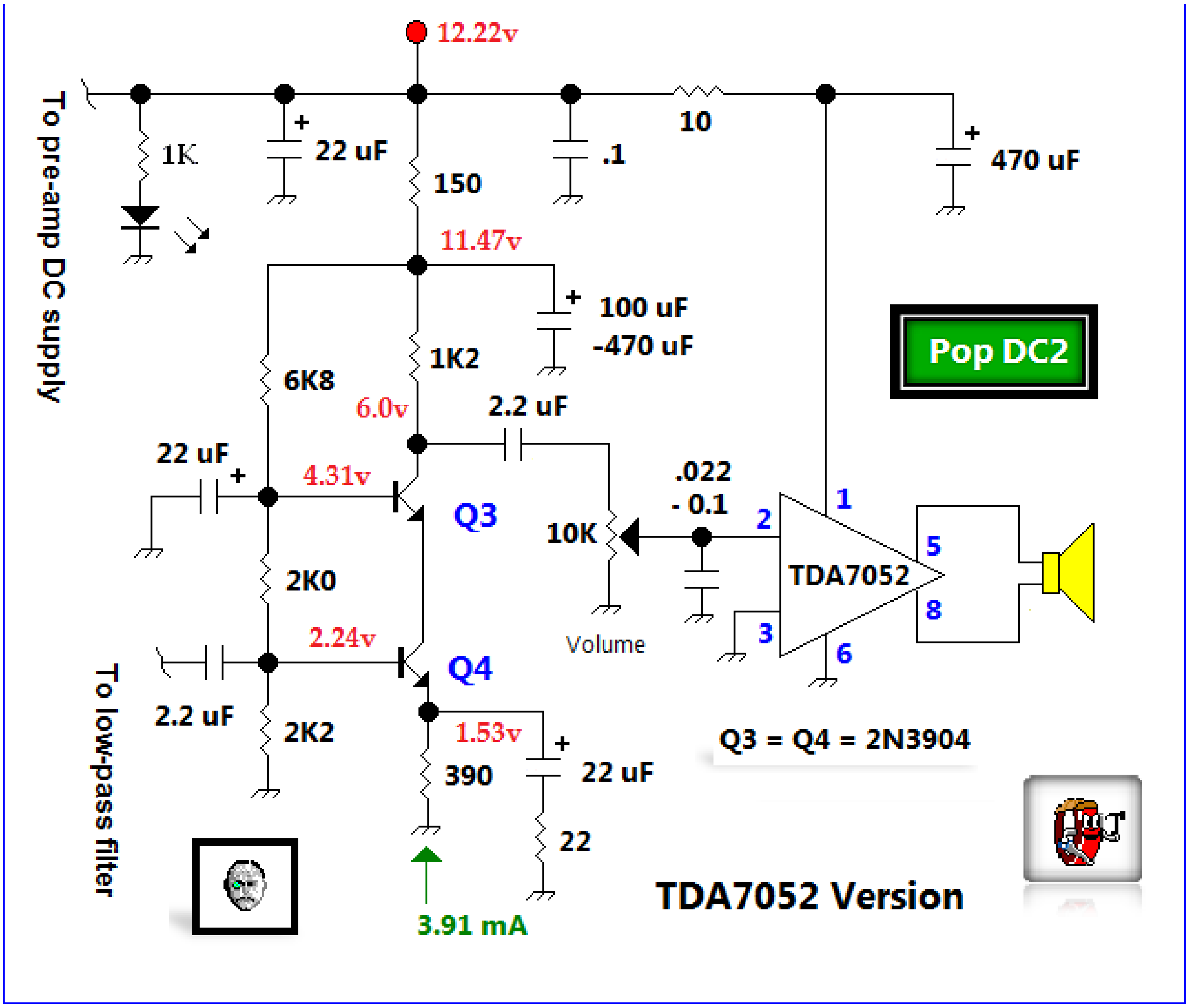

Above — The 2nd pre-amp and AF power amp. Experimenting with a number of audio stages, I decided on this cascode common emitter / common base amp biased to provide temperature stability, high gain, low distortion + proper termination of the low-pass filter. (The input Return Loss = 19 dB in my 820 ohm bridge set up). Increase the 100 uF filter capacitor on Q3 up to as high as 470 uF if you hear motor boating (low frequency thumping). This stage is prone to feedback since it's directly connected to the power amp. Photograph.

The simple and effective TDA bridged amp has a fixed gain of 40, so this receiver isn't crazy loud, however, it sounds okay. The bypass capacitor on Pin 2 filters hash noise and *can* remove some of the high frequency din from off frequency stations. Experiment to find the best value for your ears; even 0.015 uF might be your preference. I chose a 0.047 uF for my final version.

The 2 uF coupling caps between the power amp and Q3 can be lowered or raised to suit your parts collection. All the audio path coupling or bypass capacitors were "polysomething" types in my bread boards.

This is a base station receiver since the quiescent current draw listening to noise = 37 mA.

For low parts count or beginner's receivers, IC audio power amps make sense; 1 chip and you're done. Consider, for example, the TDA7052 — 2 bridged amplifiers supply reasonable power and headroom in an 8 pin DIP package. A good, but imperfect part. Depending on your goals and abilities, the limitations of the 7052's fixed 40 dB gain and/or the inability to drive grounded loads or insert additional feedback networks may constrain your designs.

Sound Bytes on 40 Meters:

I recorded these sound bytes prior to adding a 0.047 uF bypass capacitor to pin 2 of the 7052 chip and increasing the coupling cap on Q2 from 0.22 to 0.47 uF.

For a control — An ICOM superheterodyne receiver with digital IF filtering set to wide (2.2 KHz) was recorded immediately after recording the Pop DC2 receiver (although I pressed the middle (900 Hz) and narrow (600 Hz) filter selection briefly, but they made the noise worse), The antenna is a 1/4 wave vertical in a city lot with noisy conditions. I don't believe in artificially making my stuff sound better than real, and present warts-and-all audio files. I compressed these files heavily so you'll hear the noise phase shifting a little. Normally with this antenna, a "quiet" QRN level is S9; it doesn't bother me. Icom

Pop DC2 — I was tuning through a pile up to hear how the receiver copes with all the signals (twice as many with a DC receiver!)

Pop DC2 — More QRN, QSB and pile ups.

SSB - After this, I changed the .22 coupling cap between Q2 and the R-C filter .47 uF to add a little more bottom end.

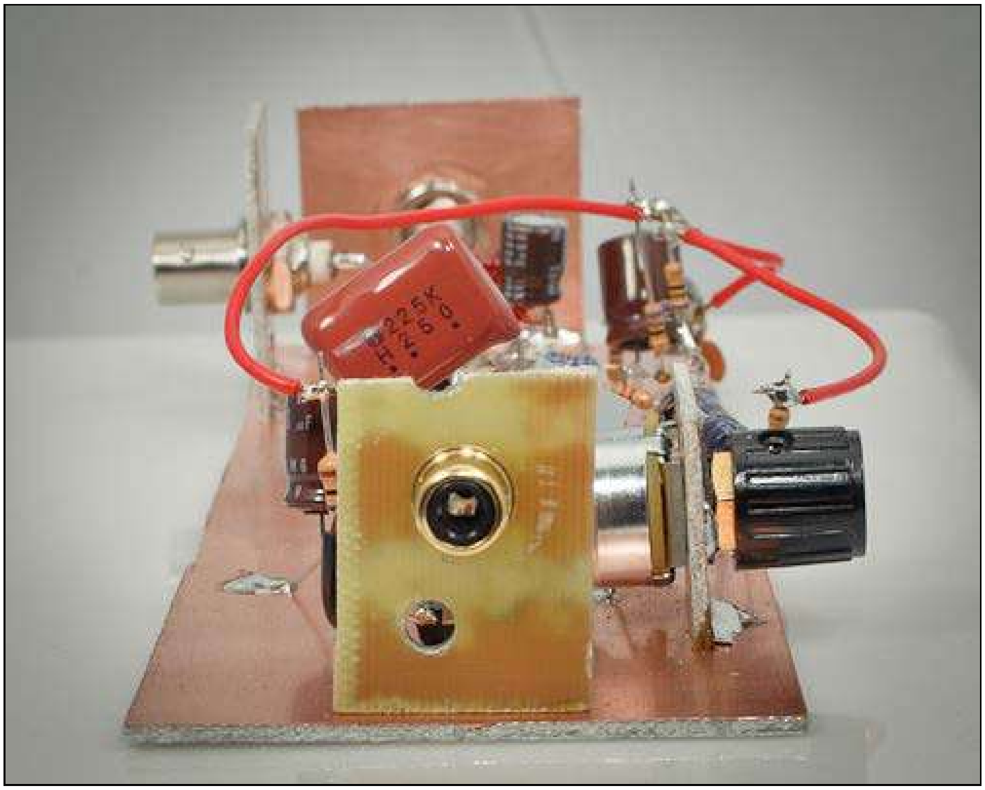

Above — Speaker terminal (an RCA jack isolated from ground on the TDA7052 version). Volume pot at right.

Second Preamplifier Stage with LM386 final

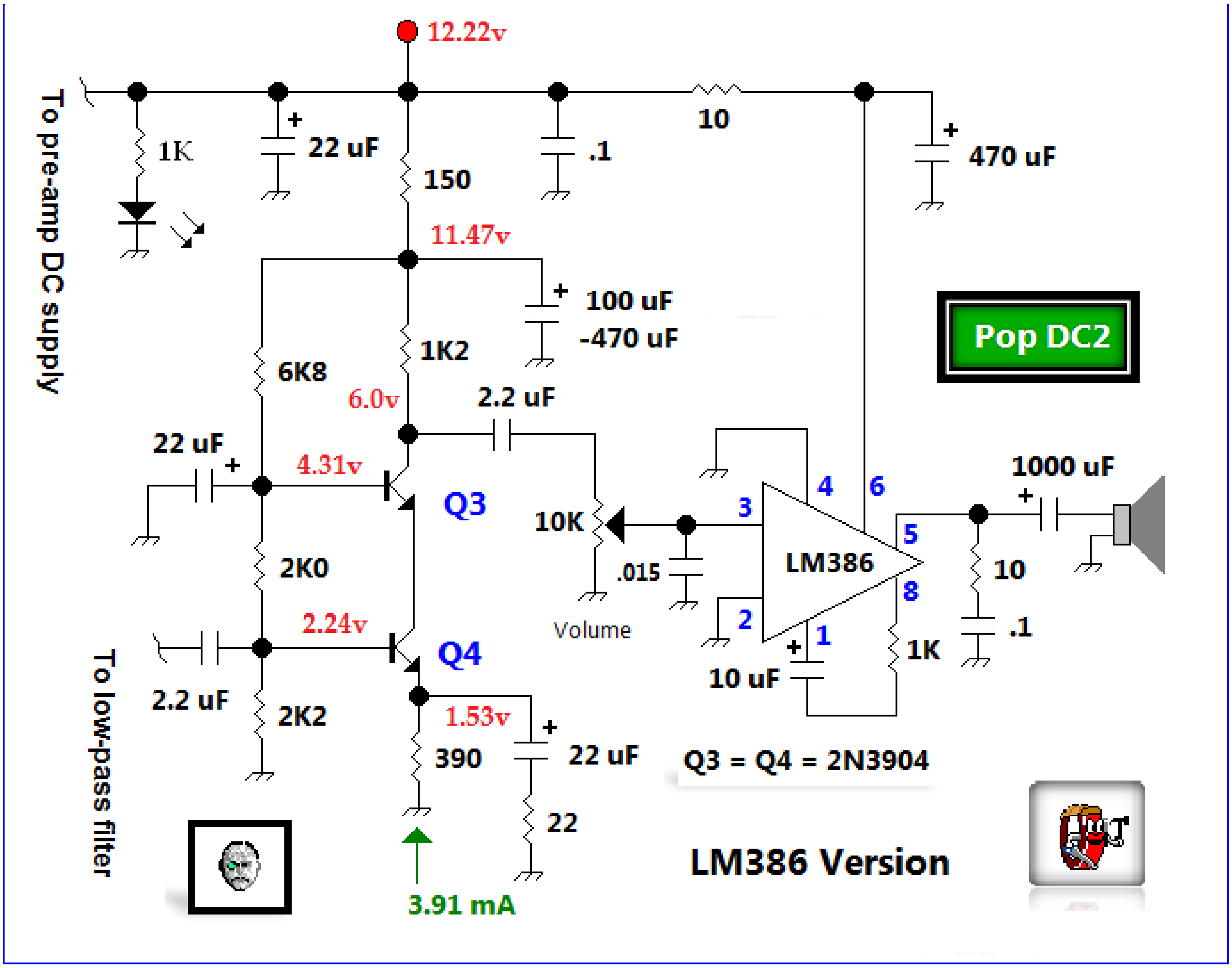
12.22v
To pre-amp DC supply
1K
+
22 uF
150
.1
10
+
470 uF
11.47v
+
100 uF
-470 uF
Pop DC2
6K8
1K2
6.0v
2.2 uF
22 uF
+
4.31v
Q3
10K
3
4
6
1000 uF
+
5
LM386
2K0
.015
10
8
.1
2.24v
Q4
Volume
2
+
1
1K
10 uF
To low-pass filter
2.2 uF
2K2
1.53v
+
22 uF
390
Q3 = Q4 = 2N3904
22
LM386 Version
3.91 mA

Above — The 2nd preamp and final with an LM386 set for just under a gain of 50. Click for a Canadian SSB sound byte.

Alternate Final Amp Stage that connects to the Q3 volume potentiometer

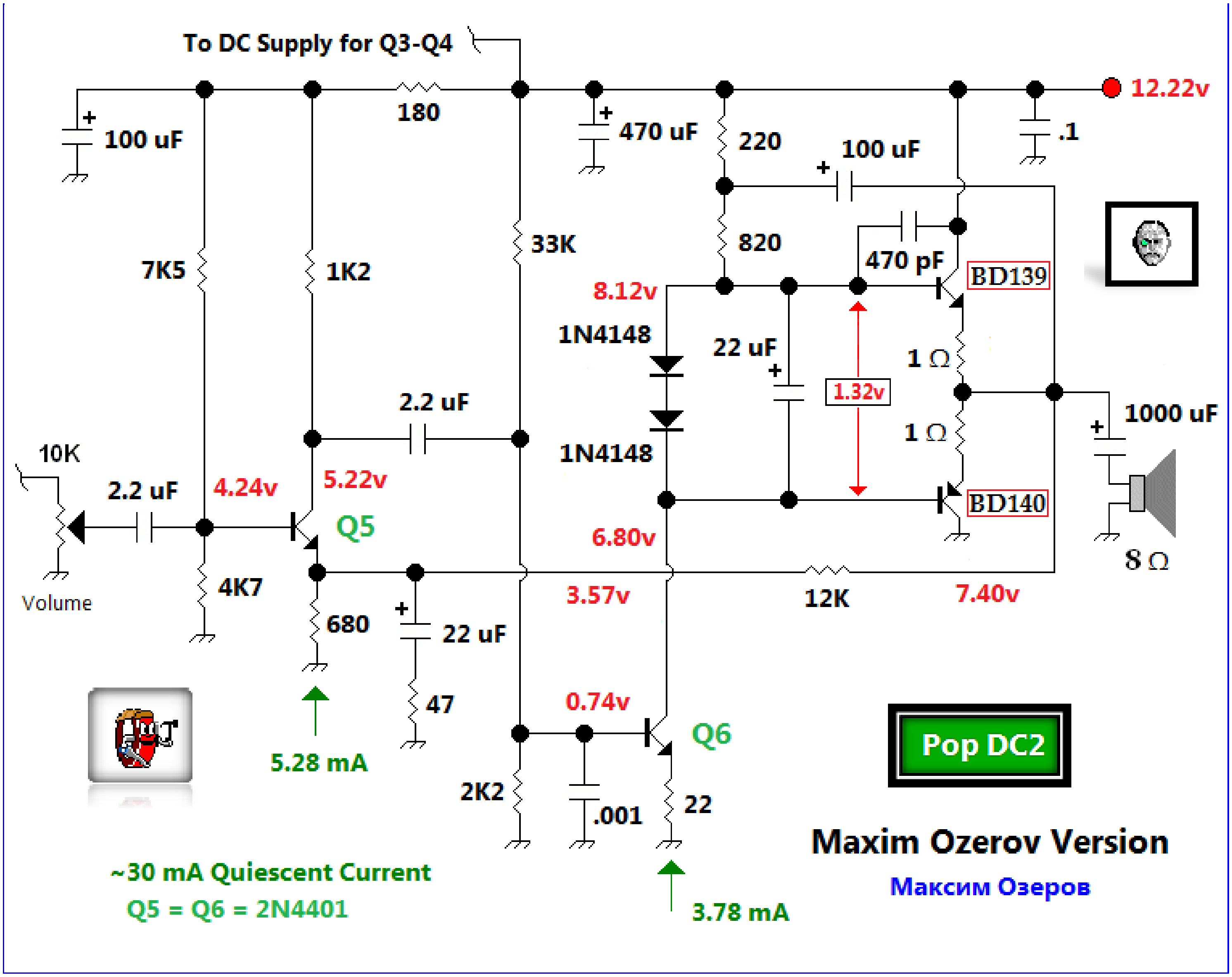
To DC Supply for Q3-Q4
12.22v
100 uF
180
470 uF
220
100 uF
.1
820
33K
7K5
1K2
470 pF
BD139
8.12v
1N4148
22 uF
1.32v
1 Ω
1 Ω
2.2 uF
1N4148
1000 uF
10K
2.2 uF
4.24v
5.22v
Q5
6.80v
BD140
8 Ω
Volume
4K7
680
22 uF
3.57v
12K
7.40v
47
5.28 mA
0.74v
Q6
Pop DC2
2K2
.001
22
Maxim Ozerov Version
Максим Озеров
~30 mA Quiescent Current
Q5 = Q6 = 2N4401
3.78 mA

Above — A reader called Maxim Ozerov requested a discrete semiconductor version of the power amp after I posted the 7052 version on my blog. Placing 2 voltage amps inside the negative feedback loop proved challenging, since I'm no expert and learn on the bench. The gain = ~42 and the maximum pure sine wave power before clipping begins to occur = ~625 mW. This amp is louder and sounds warmer than the 7052 version.

2N3904s work fine for Q5 and Q6, but I found that the 2N4401 had a consistently higher DC beta and this helps ensure the bias and collector resistors shown will provide the widest possible, pure AC signal swing.

This amp replaces the earlier IC power amps (connects to 10K volume pot after Q2, however a 2.2 uF coupling cap is required after the volume potentiometer). If you need more **voltage** gain, increase the value of the 12K negative feedback resistor. Above 75K, the gain will approach 50 and *greatly* increase the possibility of distortion.

My dummy load for development and testing = three 1/2 watt resistors in parallel: 75, 82 and 10 ohms.

Certainly you can craft better - louder - quieter audio stages with low noise op-amps, however, my readers write that they enjoy building up discrete transistor designs, and for popcorn receivers; I do too.

Sound bytes from November 1, 2011

40 Meters - QRN is lower tonight. Some audio from a lineout tape deck (no tone controls nor equalization). I have only 2 cassette tapes, This one is Russian language from 1983 - the 1 strong accented syllable generates good peaks for AF listening tests. Audio recorded from an 8 ohm, 18 cm (7 inch) speaker mounted in a wooden frame with no back. Speaker choice and cabinets are critical and often overlooked; again we may look to audiophiles for guidance

Click for some 50 ohm AF preamplifier experiments cut from this page.

7 MHz VCO Experiments

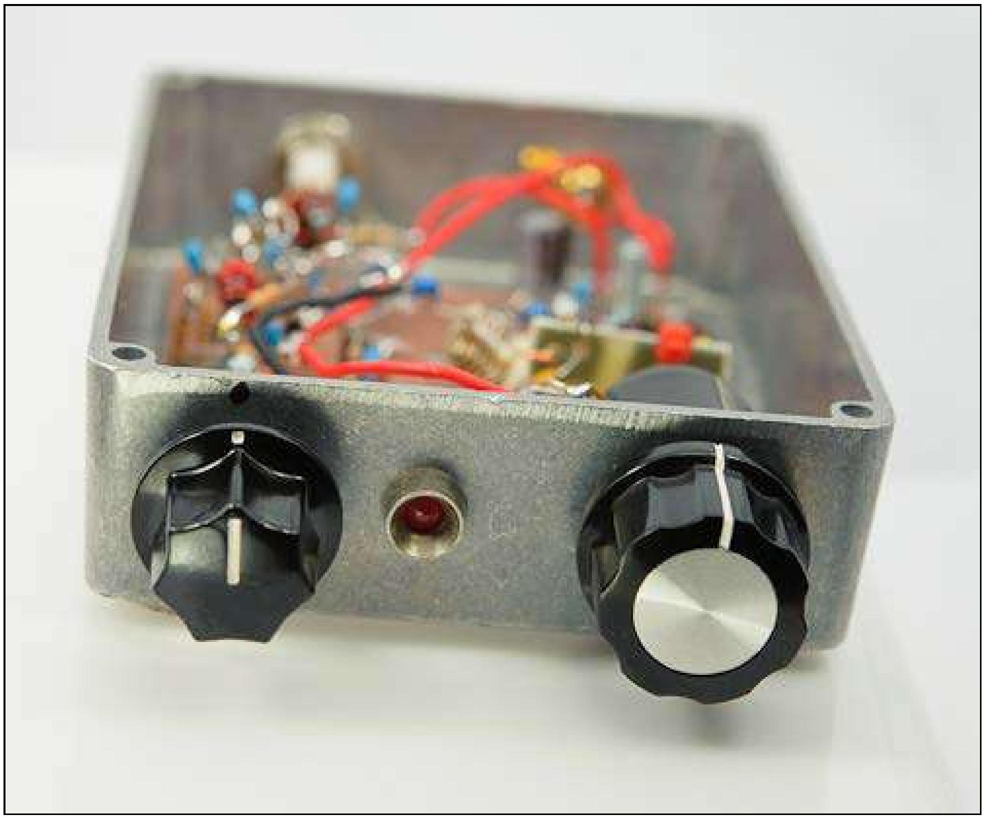

As RF designers and builders, we rely on signal generators for nearly every experiment. I sought a reliable 7 MHz voltage controlled oscillator and built 1 after some effort. I'll describe and critique a VCO I rapidly designed for a reader and then present a better VCO with some design ideas.

7 MHz VCO Experiments: A rapidly developed Popcorn 7 MHz VCO

A reader needed a 7 MHz VCO in a hurry (3 hours); he only had 1 MVAM109 varactor and wanted to cover the bottom 60 KHz of the 40 Meter Ham band using a linear taper 10K potentiometer for tuning. He planned to use a dual-gate MOSFET cascode buffer (good choice), so I didn't have to bother with a buffer.

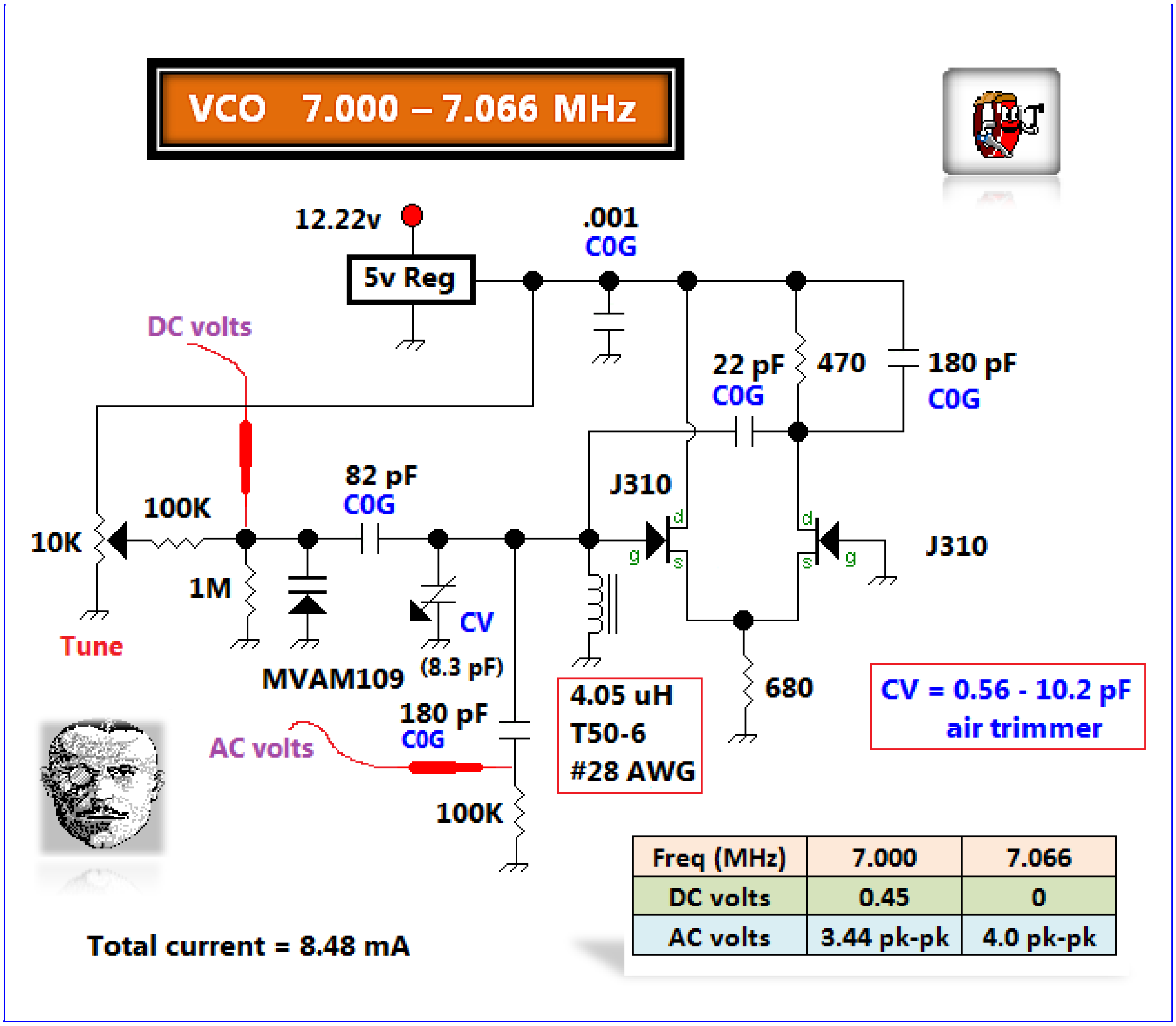

Freq (MHz)	7.000	7.066
DC volts	0.45	0
AC volts	3.44 pk-pk	4.0 pk-pk

Above — The VCO with a 100K resistor as the temporary buffer. He'll use a 100K resistor on G1 of the 2-gate MOSFET buffer. With a Q of 150 at 1 MHz; high noise level and a hyper-abrupt capacitance-versus-voltage curve designed for tuning AM radios, the MVAM109 varactor ranks poorly. The C of my MVAM109 with no reverse DC voltage was 725 pF.

Still, this VCO tuned in a linear fashion, showed a nice sinusoidal output and proved frequency stable. I wanted the AC voltage at the varactor anode at under 1 volt pk-pk (it was 752 mV) to help reduce forward conduction during the positive AC voltage swing. I was bad and ran the tuning DC voltage from 0 to 0.45 volts which greatly increases the potential for forward conduction in a varactor. To mitigate this somewhat, an 82 pF couples the varactor to the tank and drops the AC voltage and reactance seen by the varactor.

In VCOs on the web and print, you'll often see builders connect their varactor to a high Z, and high AC voltage point in the VFO tank; whoa!

At HF, if a varactor is forward biased by the positive half of the AC signal, varactor leakage current and voltage-source loading increases momentarily and lowers Q + broadens tuning. Further, serious harmonic energy and phase noise might be generated as the varactor is biased positive and negative alternately. You can sometimes see distortion in your scope during experiments with extreme AC voltage swings across the varactor. The varactor coupling capacitor should be as low as possible.

Balanced varactor tuning (anode to anode) provides another way to reduce AC signal effects at the cost of reduced maximum capacitance since the 2 varactors are in series. With back-to-back varactors, as the AC signal swings, the varactors are driven into high and low capacitance alternately, but the net capacitance remains constant. Thus applied reverse DC voltage sets the varactor capacitance rather than AC signal amplitude.

The reader for whom I made this impromptu circuit can lower the AC tank voltage by decreasing the VCC or increasing the 680 ohm source resistor after installing the buffer and tweaking things for a 7 dBm output voltage. This topology suffers from an amplitude versus frequency issue — at 7.0 MHz, the output = 3.44 volts pk-pk and at 7.066 MHz the output rises to 4.0 volts pk-pk.

Stuck with an MVAM109 constraint and 3 hours to design/build a VCO, I share this circuit as a raw experiment; not an example of good design because it is not. I took the signal off the gate to derive the best sine wave; this requires a lightly coupled, high impedance buffer with strong reverse isolation to prevent the pulling of the VCO frequency by downstream changes.

A lower L + higher C in the tank, and/or a higher Q varactor could turn this VCO into something reasonable. Popcorn versus high performance? You choose!

7 MHz VCO Experiments: A Suitable 7 MHz VCO

PART 1: Introduction

During my Fall 2011 VCO experiments I studied books including EMRFD and built versions of EMRFD Figures 4.33 and 4.34. Figure 4.33 is a common-base Colpitt's Oscillator using a hyperabrupt varactor. On Q1, the 33 ohm resistor in series with the 0.1 uF cap "de-Q" the 2N3904 to reduce UHF oscillations. Wes also employs current limiting with a 1K5 emitter resistor.

The temperature drift compensation circuit involving a temperature sensitive reference diode + op-amp fascinated me — astute temperature compensation design. I built and tested the whole circuit; the VCO has some amplitude versus frequency and phase noise issues, but it's okay for general use and *great* for varied environments. After tackling Figure 4.33, I built and tested the JFET Colpitts oscillator in Figure 4.34 and share my experiences developing this VCO with an alternate buffer.

These circuits are not cookie-cutter / carbon-copy: they show raw design ideas from the bench.

PART 2: The Voltage Controlled Oscillator

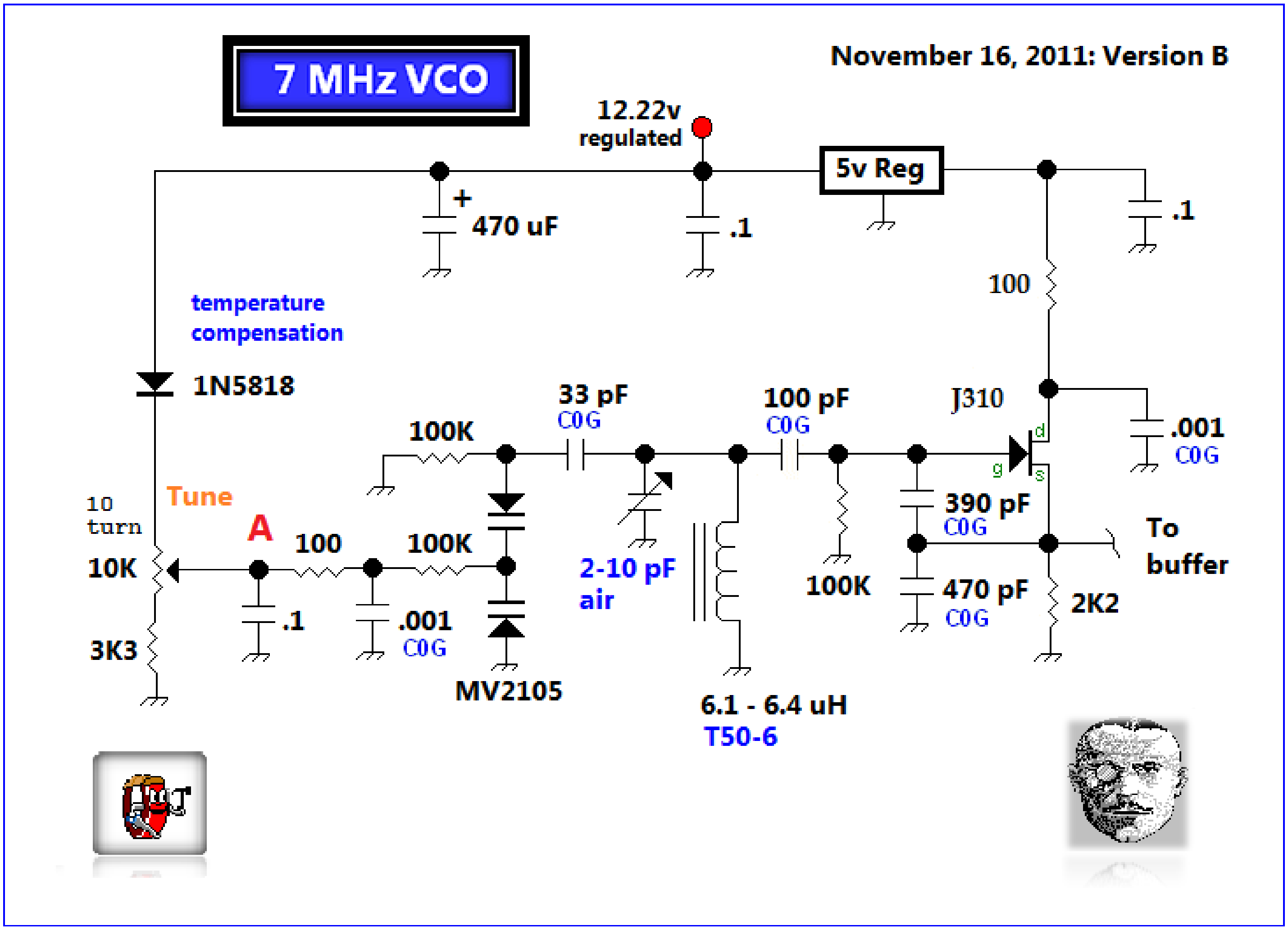
7 MHz VCO
November 16, 2011: Version B
12.22v
regulated
5v Reg
+
470 uF
.1
.1
100
temperature
compensation
1N5818
33 pF
C0G
100 pF
C0G
J310
100K
d
g
s
.001
C0G
10
turn
Tune
A
100
100K
10K
390 pF
C0G
To
buffer
.1
.001
C0G
2-10 pF
air
100K
470 pF
C0G
2K2
3K3
MV2105
6.1 - 6.4 uH
T50-6

Above — A JFET Colpitts VCO picked after after trying 5 different topologies. This VCO is my version of EMRFD Figure 4.34; originally designed by Wes, W7ZOI.

This JFET Colpitts oscillator exhibits a flat output versus frequency, low noise, scales easily to other frequencies and accomodates a wide variety of varactors. For example, you may scale it to other frequencies by changing the L and tweaking the "Colpitt's capacitors" up or down as needed.

I employed a small air variable trimmer capacitor to set the lower band edge and this meant experimenting with the inductor to find 1 that allowed me to set the band edge with such a small trimmer capacitor. I built 2 versions; in 1 the required L= 6.09 uH and in the other, L= 6.4 uH. It would be much easier to use a trimmer cap with a larger capacitance range as it makes chosing the inductor less exacting.

With the trimmer shown set to half its range, I started with a 6.6 uH coil and remove 1 turn at a time until the output in a counter was close to 7.00 MHz. After permanently fixing the inductor, I tweaked the trimmer cap so the lower band edge was 7.000 MHz with the chassis lid on.

To further drop phase noise, you could reduce the 33 pF coupling cap, add another pair of anti-parallel varactors, run a higher C to L ratio, or perhaps decrease the source resistor to increase the current limiting. Also low resistance, high Q, SMT varactors would help lower phase noise — SOD parts are tiny, but test your hand steadiness and vision.

Tuning Range

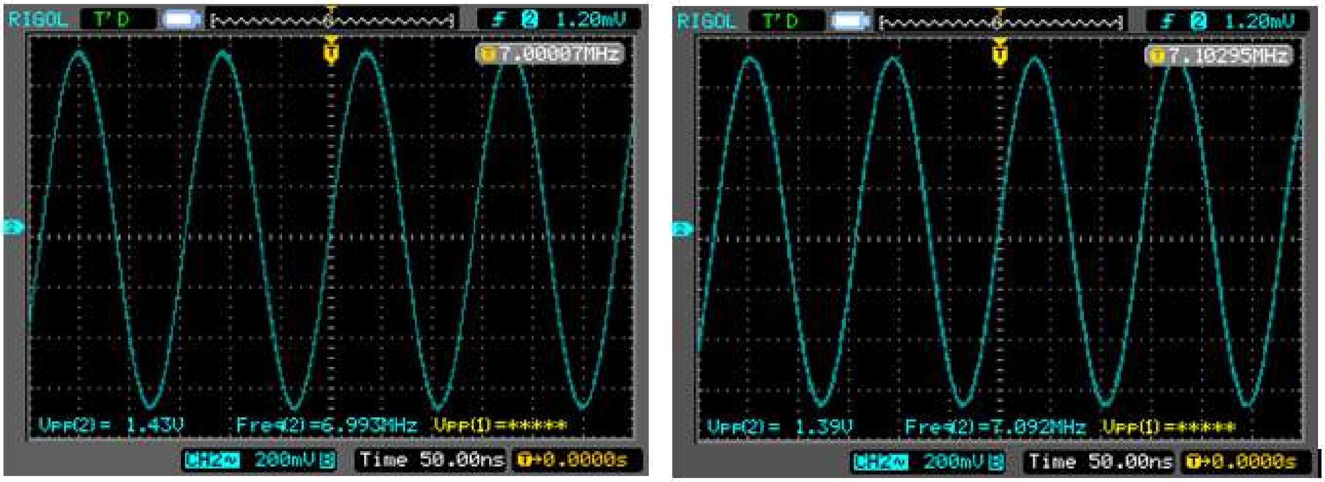
RIGOL T'D
1.20mV
7.00007MHz
Vpp(2)= 1.43V
Freq(2)=6.993MHz
Vpp(1)=*****
CH2 200mV
Time 50.00ns
0.0000s
RIGOL T'D
1.20mV
7.10295MHz
Vpp(2)= 1.39V
Freq(2)=7.092MHz
Vpp(1)=*****
CH2 200mV
Time 50.00ns
0.0000s

DC @ Point A = 3.0 volts **DC @ Point A = 12.21 volts**

Above — When tuning from the minimum frequency and tuning voltage (7.0 MHz / 3.0 VDC) to the maximum tuning voltage and frequency (7.103 MHz / 12.21 VDC) the signal amplitude only changes 0.04 volts peak-peak.

I kept a minimum of 3.0 VDC on the varactors at the minumum frequency to provide reasonably linear tuning, keep the applied reverse voltage away from 0, and improve temperature stability. All were bench determined and are not factors you can generalize to all VCO circuits. Change the minimum DC voltage on your VCO control by adjusting the resistor on the grounded end of the pot; 3K3 in my case.

Click for a moderate resolution photograph of the VCO and buffer prior to adding the temperature compensation parts.

PART 3: The Buffer/Amplifer

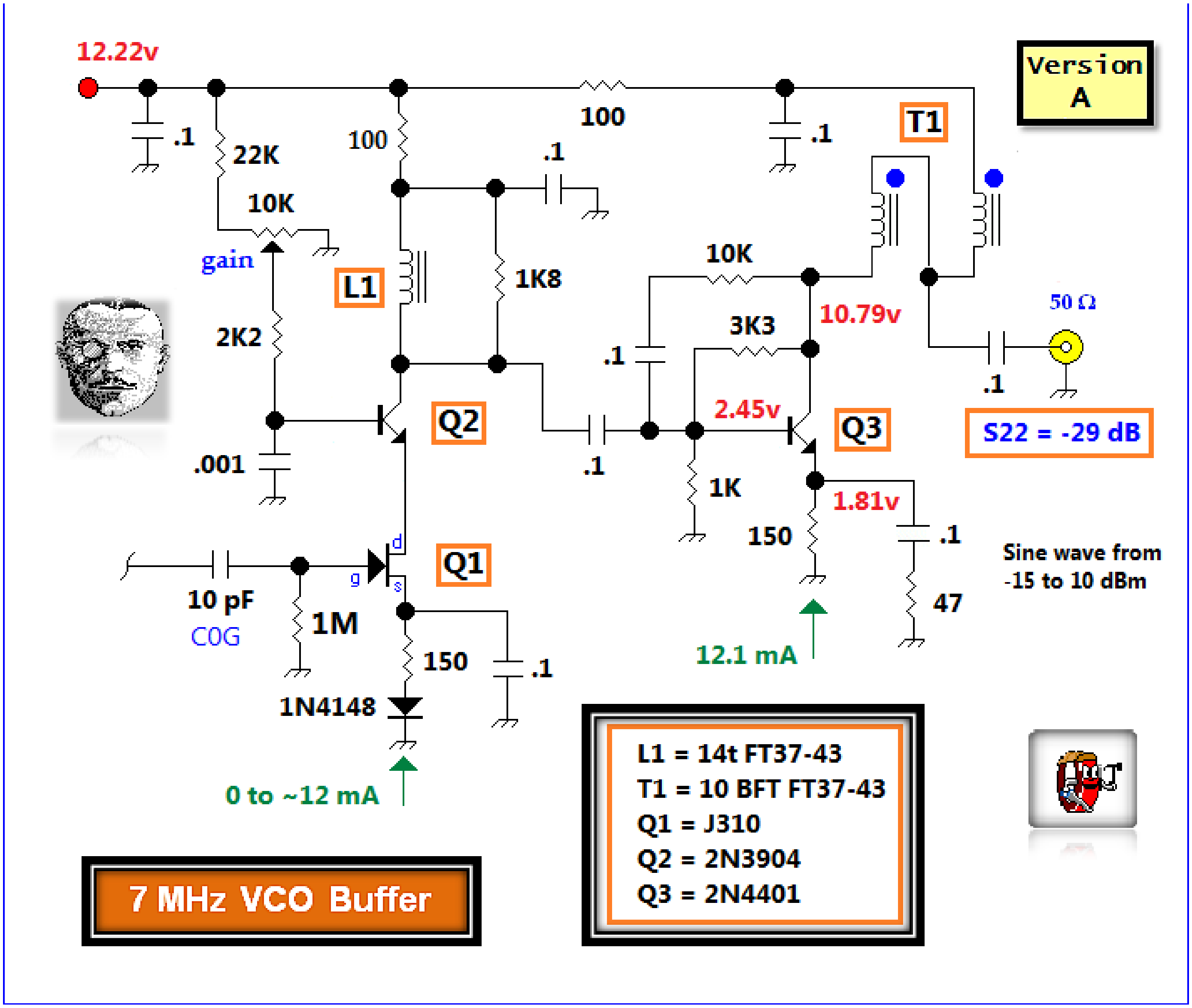
12.22v
Version A
.1
22K
100
100
.1
.1
T1
10K
gain
L1
1K8
10K
2K2
3K3
10.79v
50 Ω
.1
Q2
2.45v
Q3
.1
S22 = -29 dB
.001
.1
1K
1.81v
150
.1
d
Q1
g
s
Sine wave from
-15 to 10 dBm
10 pF
C0G
1M
47
150
.1
12.1 mA
1N4148
0 to ~12 mA
L1 = 14t FT37-43
T1 = 10 BFT FT37-43
Q1 = J310
Q2 = 2N3904
Q3 = 2N4401
7 MHz VCO Buffer

Above — The Q1-Q2 hybrid-cascode amp gives strong reverse isolation (nearly 70 dB) and front panel gain control. You could also employ a dual gate MOSFET or JFET cascode with either fixed bias, front panel control, or a trimmer resistor to adjust the bias on Q2

I enjoyed designing the Q3 final amp amp and matching its input impedance to the output Z of Q2. One way to establish a fixed + known output impedance in order to to get a strong return loss without tuned circuits/networks is to feedback some signal from the collector to the base. The difficulty lies in finding how much negative feedback to apply, while still DC biasing the amplifier for good temperature stability. I set up a crude experiment to determine the Scattering Parameter S22. The goal is to set up a good Q3 output return loss using feedback + matching the Q3 input impedance by tweaking the inductor resistor across L1 and adjusting Q3's emitter degeneration.

The return loss in my first prototype without any attenuator pad = 29 dB; some of this was pure luck.

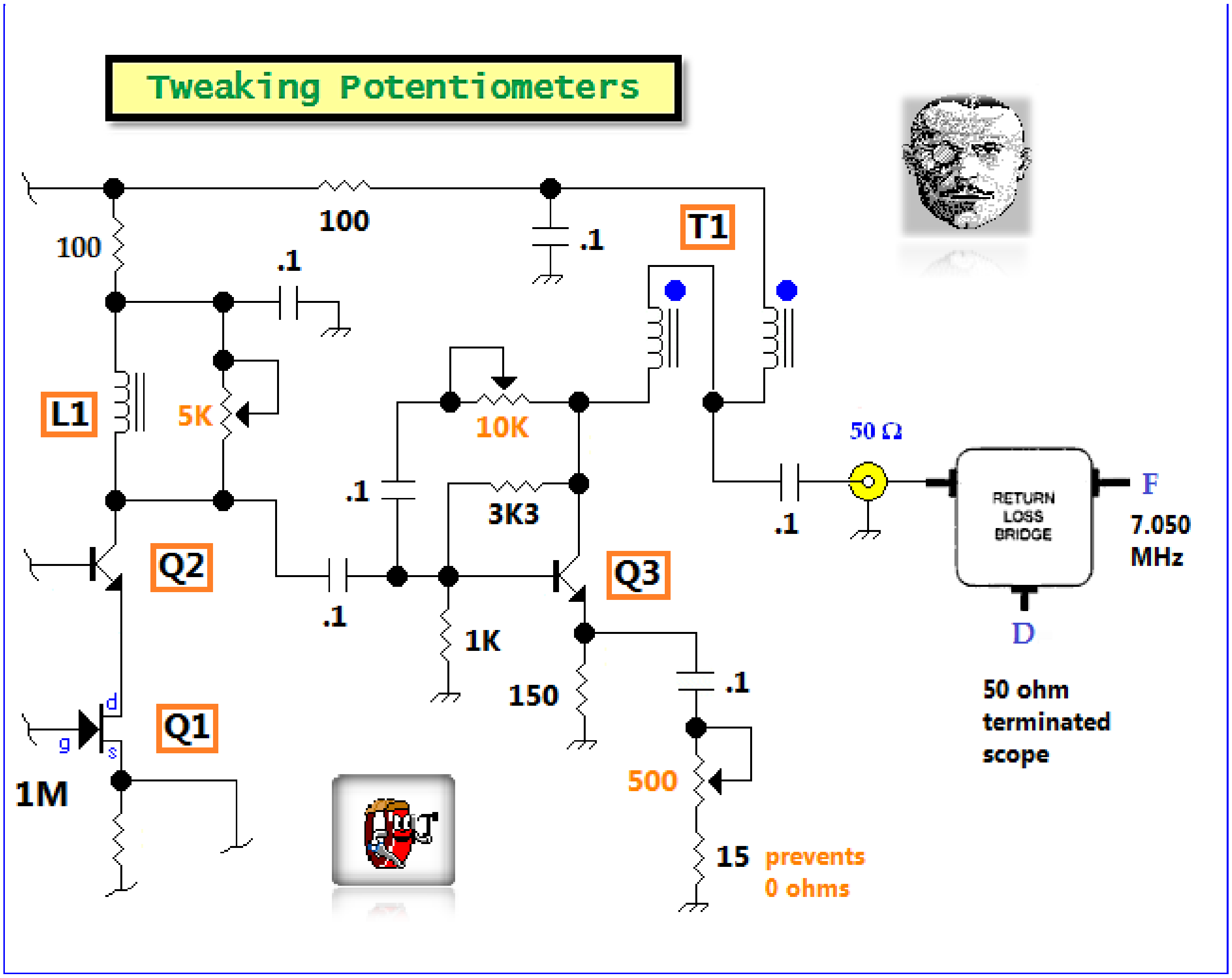
Tweaking Potentiometers
100
100
.1
.1
T1
L1
5K
10K
50 Ω
.1
.1
3K3
Q2
Q3
RETURN LOSS BRIDGE
F
7.050 MHz
.1
1K
D
150
.1
50 ohm terminated scope
d
g
s
Q1
1M
500
15
prevents 0 ohms

Above — Q2 and Q3 with 3 variable **orange colored** resistors in-situ and a Return-Loss bridge connected to the output. The potentiometers are tweaked while watching the detected output in an oscillocope. Adjust all the pots for the lowest peak-peak voltage and then carefully remove each pot and measure its resistance with an ohm meter. Replace all 3 pots with the nearest equivalent standard value resistor. Then measure and calculate the return loss (negative of S22). Watch the Q3 emitter resistance — too little R might bring distortion.

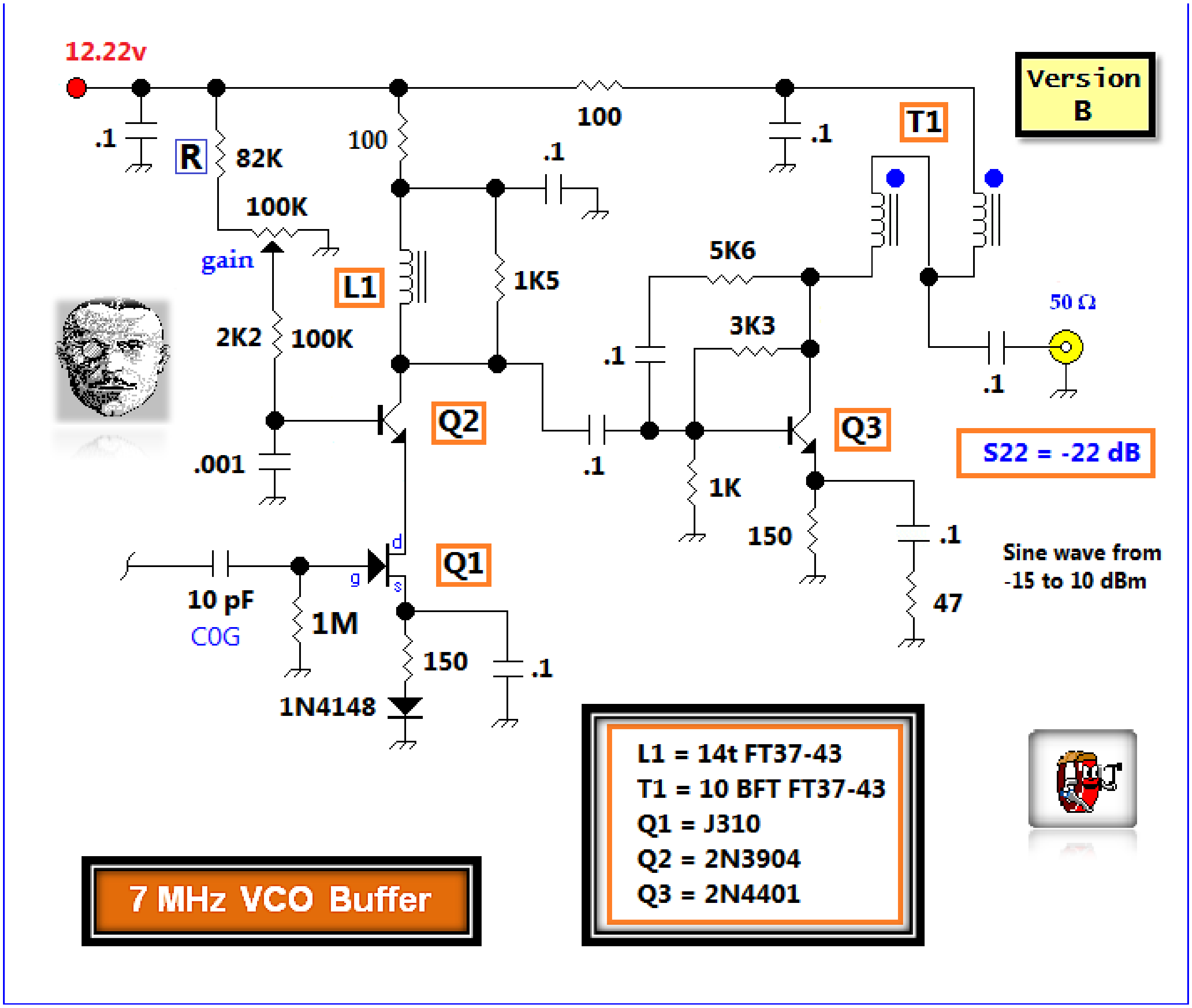
12.22v
.1
R
82K
100
100
.1
.1
T1
Version B
100K
gain
L1
1K5
5K6
2K2
100K
3K3
.1
50 Ω
.1
Q2
.001
.1
Q3
S22 = -22 dB
1K
150
.1
Sine wave from -15 to 10 dBm
d
Q1
g
s
10 pF
C0G
1M
47
150
.1
1N4148
L1 = 14t FT37-43
T1 = 10 BFT FT37-43
Q1 = J310
Q2 = 2N3904
Q3 = 2N4401
7 MHz VCO Buffer

Above — A seperate buffer built with 100% different parts that required different AC feedback plus shunt resistor across L1. The parts in this circuit weren't as hot as Version A, and the maximum output voltage was only 1.8 volts pk-pk. In order to get the AC output voltage to just above 2 volts, I had to tweak the resistor labelled **R**.

To keep the heat and current down in the final amp, I decided to keep the maximum clean output to 2 volts peak-peak (= 10 mW = 10 dBm) with an emitter current of ~ 12 mA. If you want higher clean output than 10 dBm, you'll have to run more Q3 emitter current and maybe choose a different BJT, plus apply a heat sink.

When cranked to maximum DC voltage, the Q2 gain pot allows a peak output AC voltage of ~2.2 volts pk-pk into 50 ohms and distortion is evident. At or below 2 volts pk-pk all is well — I'll use this VCO mostly from 0 to 7 dBm.

Since the circuits uses 2 BJTs and a JFET and many 5% tolerance resistors , the Q3 output will vary according to your parts. Tweak the resistor labelled **R** to provide a maximum AC signal just over 2 volts peak-peak into 50 ohms. This translates to around 3.8-5.5 volts DC bias for Q2 with your gain pot cranked fully clockwise.

Return loss variations. You probably noticed the return loss in Version B = 23 dB, while Version A = 29 dB.

Version B originally had the 1K8 shunt resistor across L1 and the 10K + 0.1 uF AC feedback arm just like version A and I measured a return loss of 22 dB. I stuck in 2 tweaking potentiometers (did not bother tweaking the the emitter series feedback element). After pot tweaking, the best return loss I could obtain with 5% tolerance resistors was 23 dB and this probably represents what the average builder will obtain. An S22 of - 22 to -23 dB works fine for the QRP work bench.

If you don't plan to do any potentiometer tweaking, I recommend building circuit A since it has a little more gain due to the slightly higher shunt resistor, and also I built 3 versions of Version A with an S22 of -22 dB or higher.

PART 4: Temperature Compensation

Before temperature compensation, my VCO slowly drifted down in frequency and was unusable.

If you look through the Ham Radio VFO/VCO literature, you will see that many builders use polystyrene caps as the Colpitt's capacitors, and/or in parallel with other NP0/C0G tuning capacitors. Negative temperature compensation caps like an N750, or the polystyrene types temperature compensated the oscillator. Negative temperature co-efficient caps are hard to obtain for many builders; especially in small quantities, however, they are worth their weight in gold.

Diode Compensation

Stabilize your VCO as much as possible with compensating capacitors and by following prudent temperature stability techniques before adding diode compensation. *See the VFO 2011 web page and EMRFD.* Temperature compensation is best performed in a homebrew oven (see EMRFD) and normally takes an incredible amount of time and patience.

Temperature compensating diodes are far from static — a diodes temperature co-efficient is dynamic and may vary with current and also unfortunately, with temperature and even while tuning your VCO !

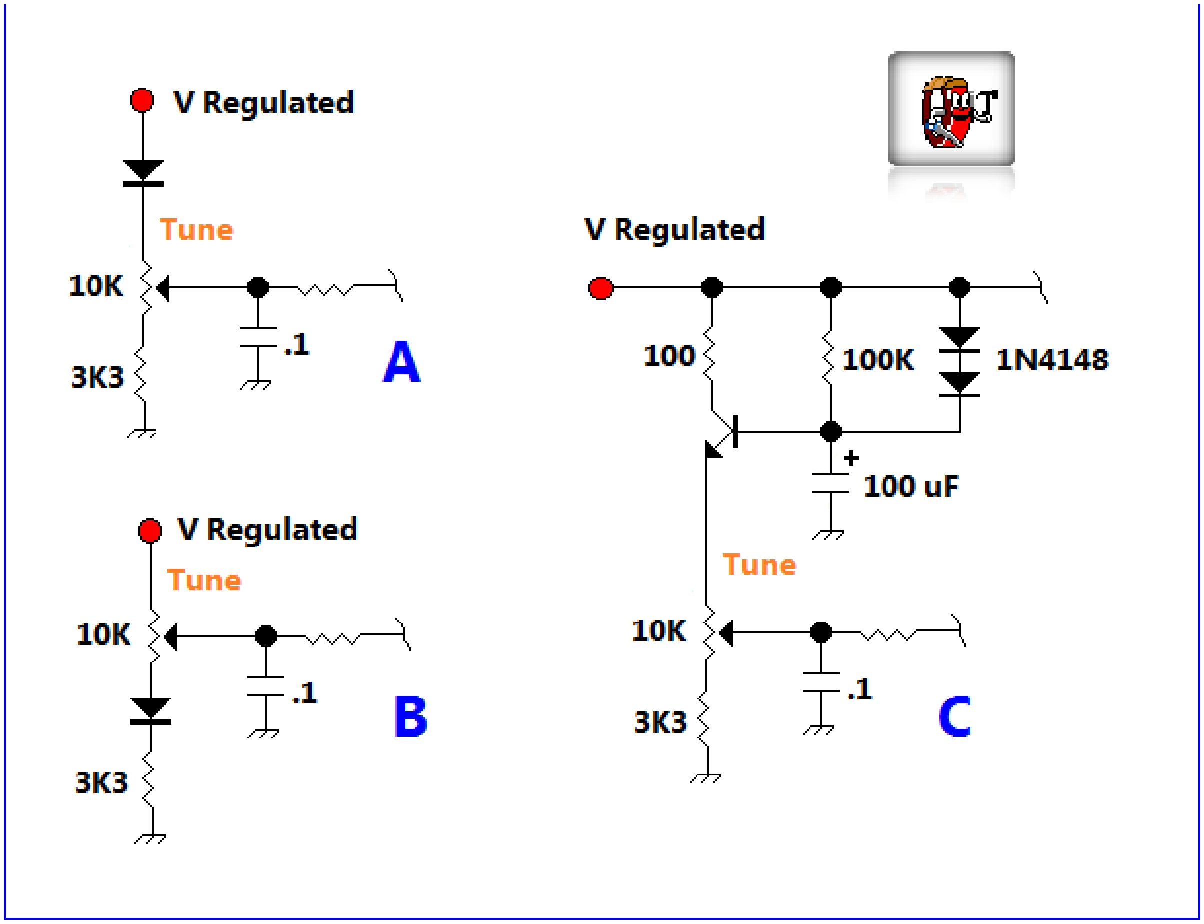

Above — Simplistic diode temperature compensensation schemes.

The late, great, Doug DeMaw advocated sticking a 2N3904 or 2N2222a (wired as a diode) between the control potentiometer and the varactor decoupling network since the forward biased P-N junction exhibits a negative temperature co-efficient and should stop the decrease in frequency. It can help, however, as you tune and swing the control DC voltage from minimum to maximum the forward bias on the diode increases and the diode temperature coefficient decreases.

I've never had success using a transistor in this way; the BJT caused the VCO frequency to increase in an erratic manner that varied along with the DC control voltage. When watching drift in a frequency counter set to sample every second or so, a stable design will slowly change frequency in 1 or occasionallly 2 Hertz increments — some people call this "linear drift". if you see your VCO dropping down frequency in 10 - 20 Hertz jumps per second, you'll have a bad time temperature compensating.

I experimented with the above 3 designs that keep a constant current on the diode. Figures A and B work. I tried both and confirmed that a given diode compensated slightly differently when in circuit A or B. This gives you a bit of a tweaking room for your chosen compensation diode. I tried Figure C, but it had too much negative temperature coefficient and sent the VCO drifting upward about 1-2 Hertz each second.

I settled on circuit A and then tried some diodes: the initial best was a grubby old Germanium from my junk box. The best choice turned out to be a Schottky barrier rectifier (1N5818). I connected the VCO to a receiver and could listen to CW QSOs without groaning. My VCO now drifted up in linear, 1 Hertz hops at about 105 Hz per hour. It took a long time to tack solder in and wait 10 or 15 minutes for each diode to stablize before I finally settled on the 1N5818.

The better solution is to choose a suitable diode and vary its current to tweak the temperature compensation. Wes did this in EMRFD Figure 4.33.

Advanced designs may use a reference voltage + a temperature dependent voltage that is applied to op-amps in a proportional way to temperature compensate the DC control voltage. Then, too, some builders ovenize their VCO container to maintain a very stable environmental temperature.

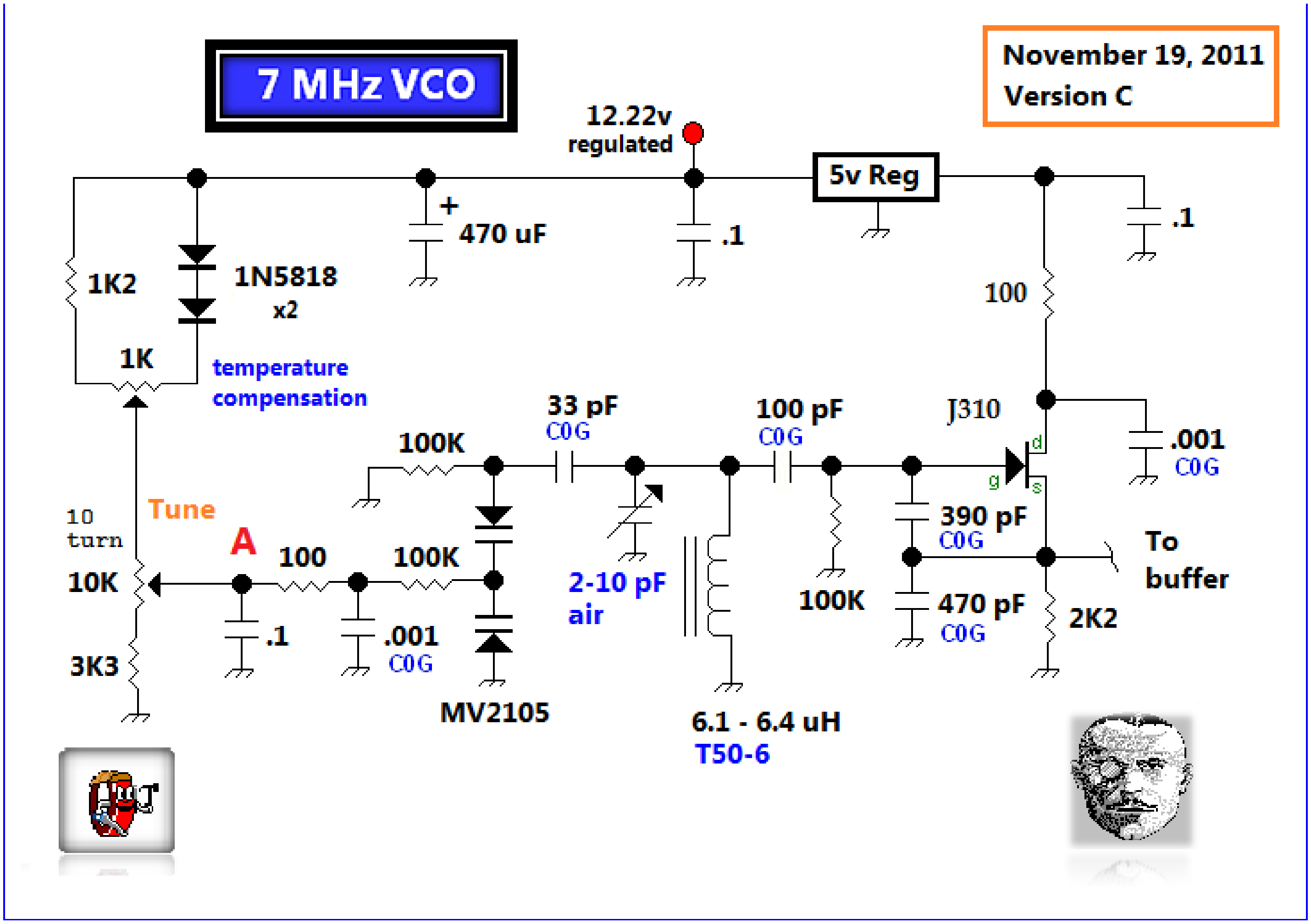
7 MHz VCO
November 19, 2011
Version C
12.22v
regulated
5v Reg
470 uF
.1
.1
1K2
1N5818
x2
100
1K
temperature
compensation
33 pF
C0G
100 pF
C0G
J310
d
g
s
.001
C0G
100K
10
turn
Tune
A
100
100K
10K
3K3
.1
.001
C0G
MV2105
2-10 pF
air
100K
390 pF
C0G
470 pF
C0G
2K2
To
buffer
6.1 - 6.4 uH
T50-6

Above — A simple and elegant diode compensation scheme proposed by Ken Kuhn. Basically, it lets you tweak the degree of compensation to what is really needed rather than accept what you get from a diode. Adding more diodes will increase the effect — but the 1.2 K resistor should be increased accordingly to roughly match the overall voltage drop of the diodes. Hopefully there is a point on the 1K potentiometer where temperature compensation can be very good at a tuning point of interest. The diodes should be located to thermally match the rest of the oscillator circuitry.

Set the band edge *after* finding the sweet spot on the 1K potentiometer since it will affect the tuning frequency. This experimental circuit cannot be casually copied and it took a while to converge to the desired operating point on the 1K potentiometer. Generally you start with the 1K pot towards the 1.2 K resistor and then adjust for the best stability after warm-up. Then repeat and adjust as necessary over time.

Temperature compensatiing an oscillator like this is a challenge as all parts have some temperature drift and it takes a lot of measurements (and often, some dumb luck) to determine the overall compensation curve that is needed.

My 1 hour drift up in frequency is now ~ 60 Hertz per hour at various tuning frequencies across the tuning range. I stuck with the 1N5818 diode, and probably should have tried other diodes and also changed the 1K2 resistor to observe any effects, however, I have spent an inordinate amount of time on this circuit and leave it to others, or future bench work to improve. *See QRP — Posdata below.*

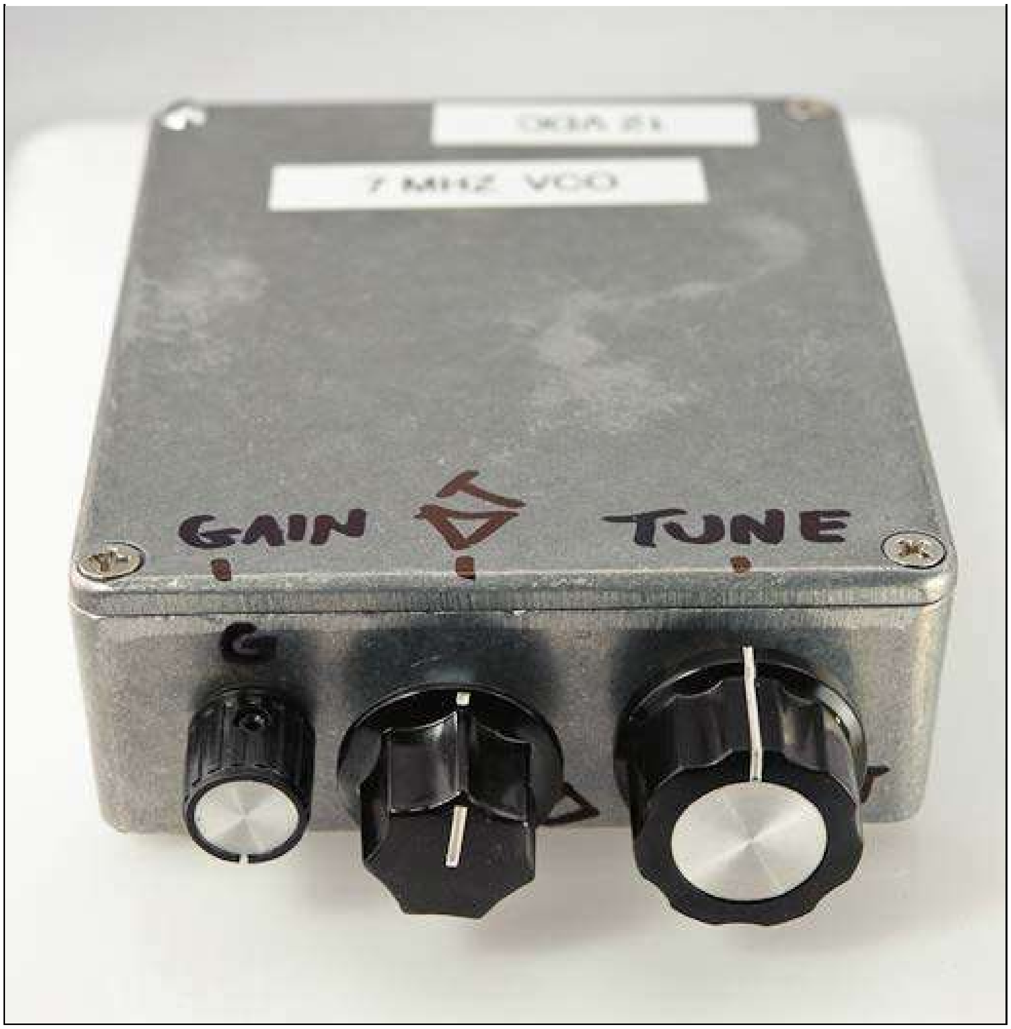

Above — Version C of the VCO. When I built the first versions, I drilled a hole to accomodate a third potentiometer, but filled it with a LED holder that was temperature sealed with epoxy glue. The 1K pot *jus*t fit into my chassis.

QRP— Posdata for January 2012

In late 2011, I shopped on eBay to build up a small quantity of 10 - 270 pF polystyrene, plus some 56 pF N750 ceramic temperature compensation capacitors. After 2 simple, but time-consuming experiments, I temperature stablized my 7 MHz VCO frequency drift to under 10 Hz per hour in the relatively constant temperature of our basement.

I didn't feel like re-doing the whole resonator circuit and thus focused on the tank to FET coupling capacitor. Placing a N750 capacitor in parallel with a fixed NP0/C0G cap to make 100 pF resulted in over-compensation and no amount of tweaking on the "adjustable diode" circuit worked. A few hours late, I swapped in a 100 pF polystyrene capacitor and after further hours of waiting and tweaking, I nailed the frequency stability sought.

An oven provides the best way to temperature compensate, however, whether you choose the oven method, or the bench method like I did, great patience is required to see if a change to your TC circuitry works or not.

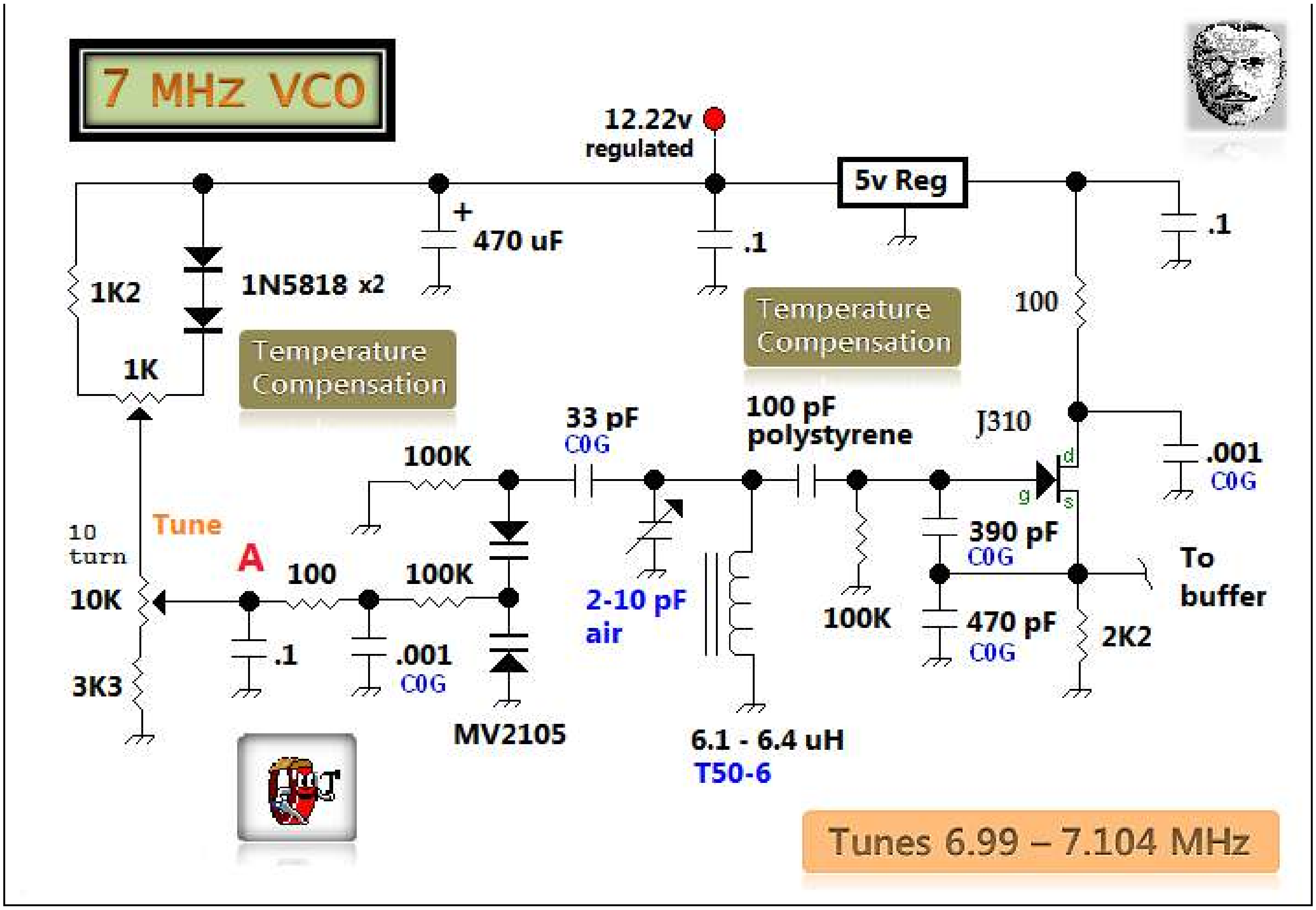
7 MHz VCO
12.22v
regulated
5v Reg
.1
470 uF
.1
1K2
1N5818 x2
1K
Temperature Compensation
Temperature Compensation
100
33 pF
C0G
100 pF
polystyrene
J310
100K
.001
C0G
d
g
s
10 turn
Tune
A
100
100K
390 pF
C0G
To buffer
10K
2-10 pF
air
100K
470 pF
C0G
2K2
.1
.001
C0G
3K3
MV2105
6.1 - 6.4 uH
T50-6
Tunes 6.99 – 7.104 MHz

Above — Final version of the 7 MHz VCO. I changed the 100 pF capacitor coupling the JFET to the resonator circuitry from C0G to polystyrene and slowly tweaked the 1K temperature compensation pot to find the point of convergence.

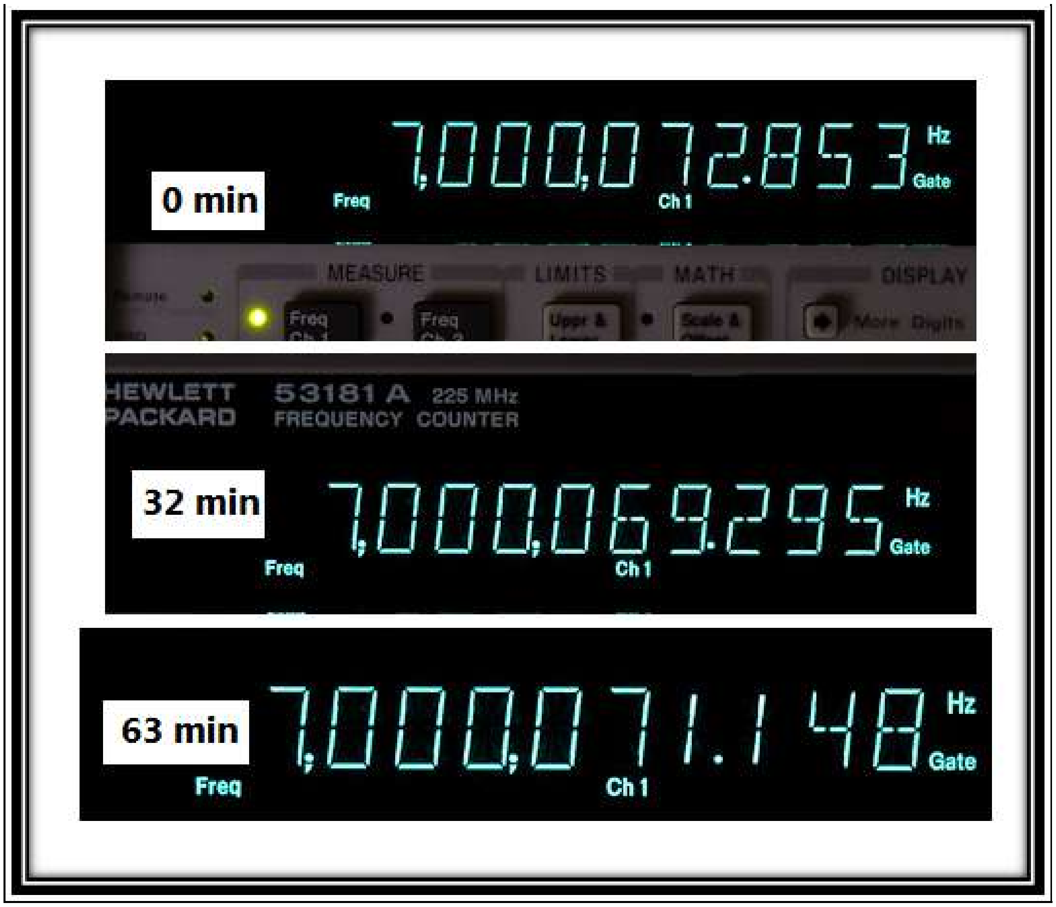
0 min
7,000,072.853 Hz
32 min
7,000,069.295 Hz
63 min
7,000,071.148 Hz
HEWLETT PACKARD 53181 A 225 MHz FREQUENCY COUNTER

Above — The ~1 hour drift after a 30 minute warmup period for the 7 MHz VCO. Love this.

TEMPCO
CAPS

Above — My new temperature compensation capacitor parts drawer. I'll keep an eye out for further bargain temperature compensation parts on eBay and at Ham Radio festivals.

Fearless Leader and Hero   Храбрый вождь и герой

Above — Professor Vasily Ivanenko ( ), fearless leader (ТЫ МОЙ ГЕРОЙ)

He's my hero because he's humble, fallible, well-intentioned and moral. Professor Ivanenko lives for learning — fame is filler — hollow and distracting. His current *ego* lags his voltage by 90 degrees. Is he part inductor / part human?

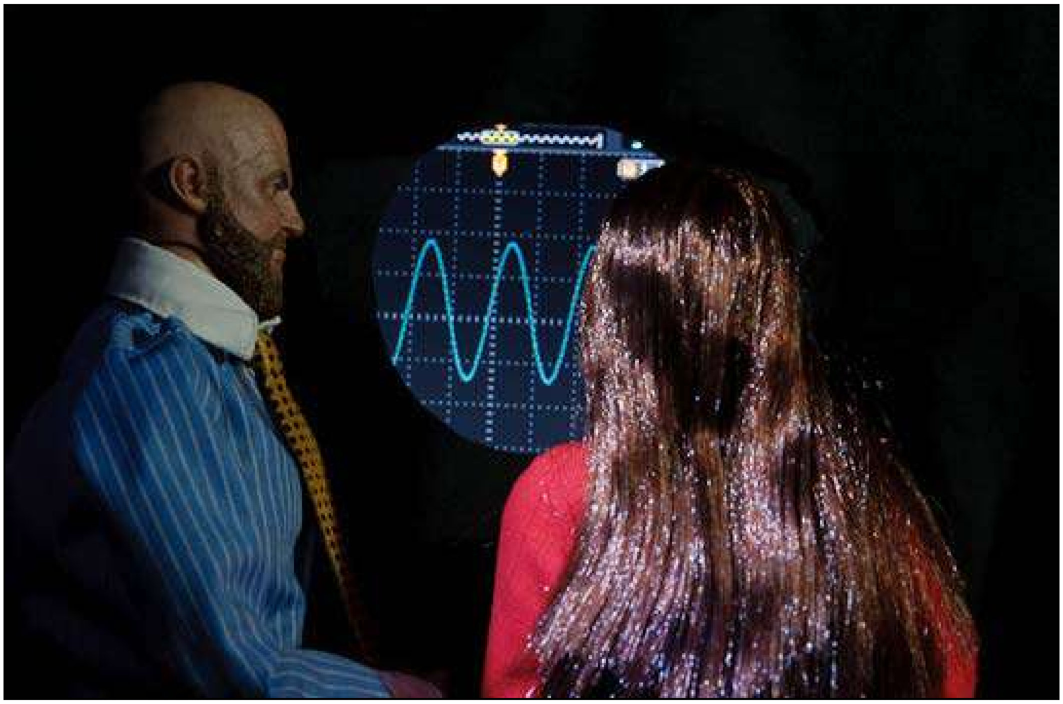

Miscellaneous Photos and circuits

Above — 3 types of adapters. A BNC male to SMA female, a BNC male to PL-259, and a BNC female to S0-239 allow the RF modules to be connected to a variety of equipment.

SCOPE
RF
RF
RF
GAIN
D
4
-10

Click

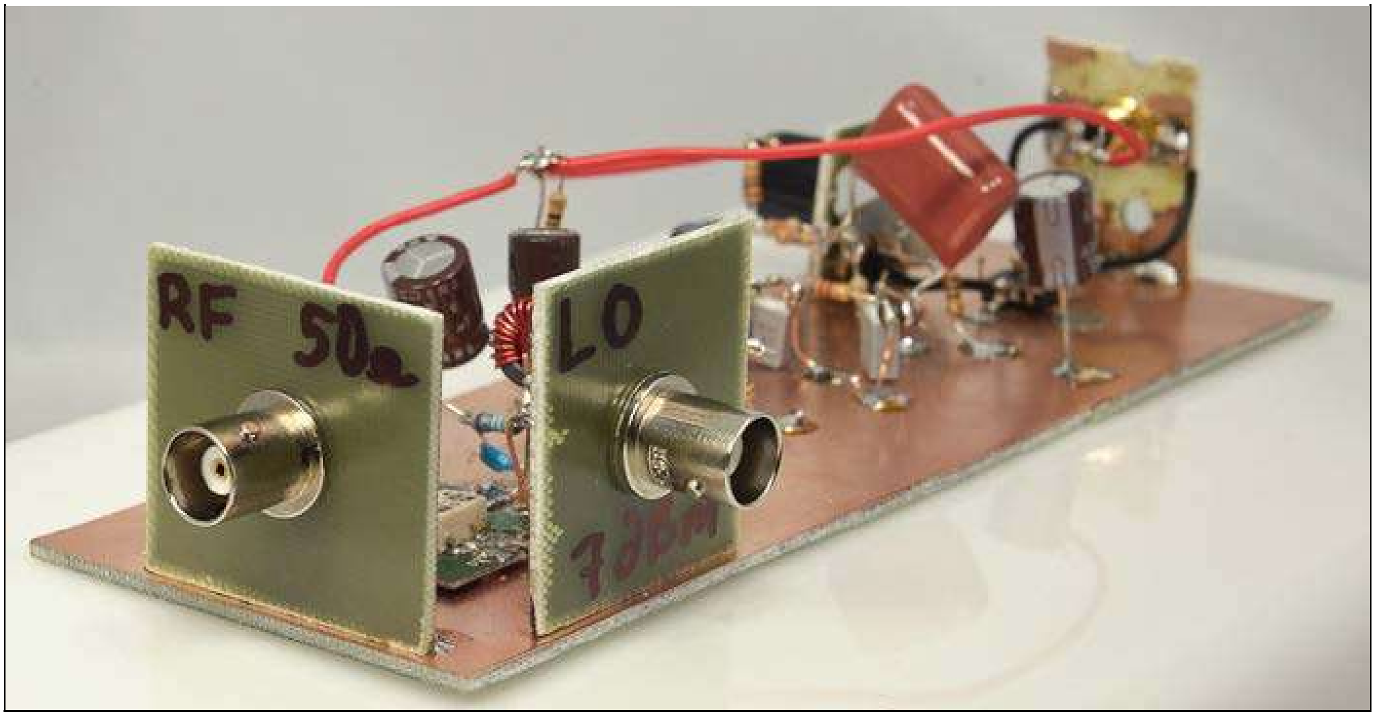
RF 50Ω
LO
7dBm

RF — Test and Measurement

Double Tuned Receiver Band-pass Filter Design Center

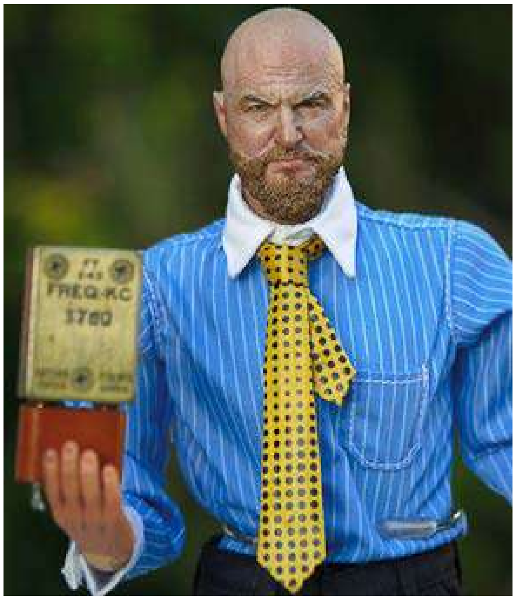

This web page is for builders who own EMRFD. Assisted by 4 of the Ladpac programs from the EMRFD compact disk, and the information presented in EMRFD Chapter 3, I share some experiments building popcorn receiver band-pass filters. Prior to diving into this material, please read the help file ***Ladpac2008 Manual.pdf*** and a file on the EMRFD compact disk called The Double Tuned Circuit: An Experimenter's tutorial by Wes, W7ZOI.

Preface

Derived from experiments, my web content reflects the efforts of a lay-person, hobby-level designer — I make mistakes. I say this not to make excuses or avoid accountability, but to share the truth. My hope is that my experiments inform yours and we all improve over time. I correct reported mistakes and rely on your eyes to see them.

Arduous and requiring good math skills, filter design is out of reach for many builders. Software changes this and learning to apply computer programs in real-world situations is part of our hobby. This web page shares some bench experiences, plus my thoughts about using some programs written by Wes, W7ZOI. I present suggestions and examples based more on empiricism and from reading about band-pass filter design than scientific methodology.

From email regarding my VFO and RF Workbench pages, I have become aware that I've lead many builders to think that a perfect sine wave and a high return loss are "must have" bench outcomes. This is false. A clean sine wave proves useful for accurate measurement, but is not a *de rigueur* bench outcome. A desire for high return loss reflects my own personal *obsession*; in simple QRP rigs, this may represent folly. Please don't overestimate the importance of return loss from my bias; decide for yourself.

Part 1: Experiments with 2 coupled L-C tanks.

Goal: A 15 Meter band band-pass filter with an insertion loss < 4 dB and a return loss of >= 20 dB.

Software: Ladbuild08 and GPLA08.

The simplest band-pass filter is an L-C tank. To get a decent stop band we generally couple 2 or 3 tanks together with series capacitor(s). Other filter topologies were ignored. In Part 1, I just connected up a couple of tanks on the bench without the use of software. Some attempts at impedance matching via transformer links were also trialed.

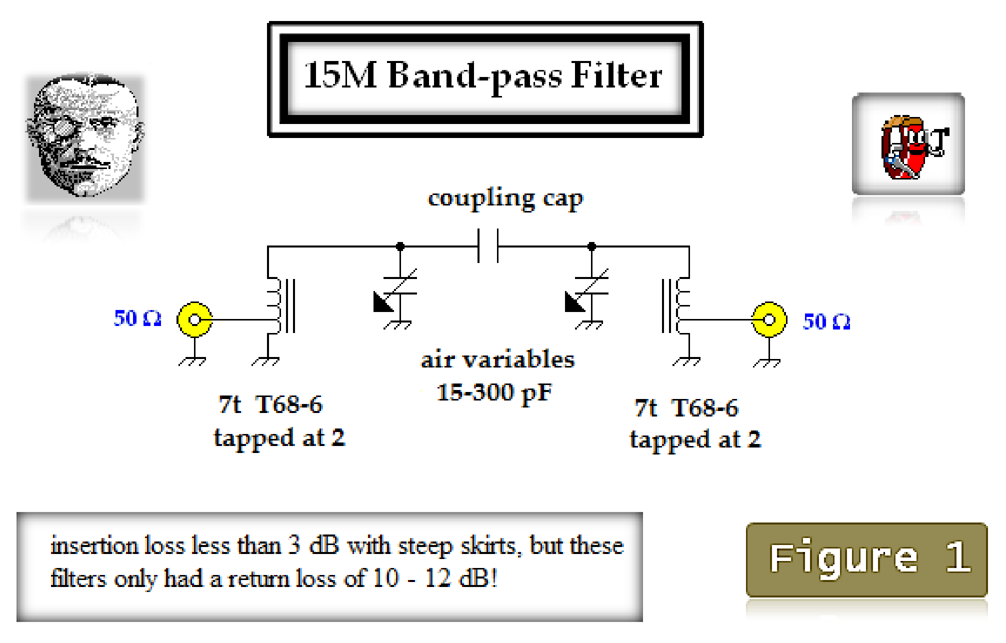
15M Band-pass Filter

insertion loss less than 3 dB with steep skirts, but these filters only had a return loss of 10 - 12 dB!

Figure 1

Above — I built a 5 component filter for the base experiment. Inductors = 7 turns of #22 AWG on a T68-6; tapped at 2 turns from ground. The inductors turns were expanded or compressed until L= 300 nH. Tuning capacitors = large 15 to 300 pF air variable capacitors. Coupling capacitors trialed = 2 pF, 3.3 pF, 5 pF, and 7.5 pF.

After soldering in a coupling capacitor, each tank (also called resonator) was tuned to resonance by looking at the peak-peak output voltage in a 50 ohm terminated oscilloscope. After tuning, I measured insertion and return loss and then swept each filter with a tracking generator + spectrum analyzer. On the bench I determined that the greatest return loss occurred with 2 transformer taps from ground; the result — a dismal 10-12 dB.

What effect does changing the coupling capacitor have?

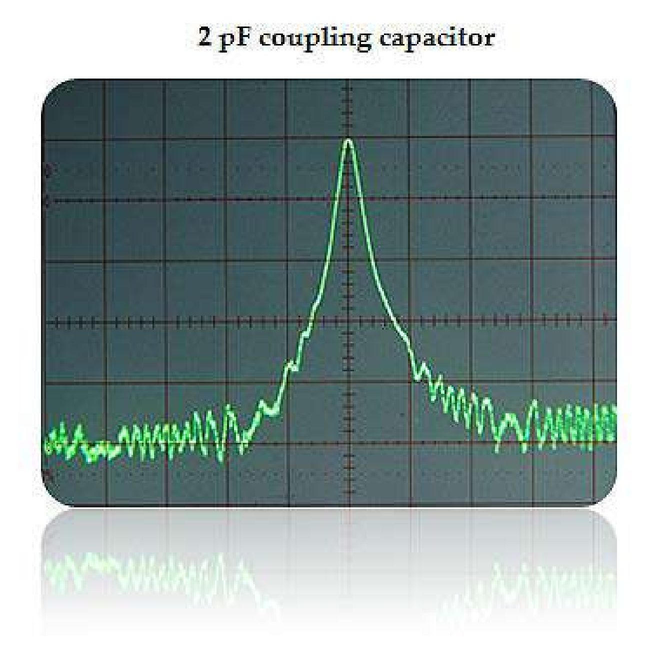
2 pF coupling capacitor

Above — A spectrum analyzer + tracking generator sweep of the filter response with a 2 pF coupling capacitor between the inductors. Graticules = 2 MHz per horizontal division and 10 dB per vertical division. Click on this zoom to better see the - 3 dB bandwidth. The sweep revealed a sharp peak response with steep skirts and a 3 dB down BW of ~220 KHz or so. Some of the noise arose from the big air variable caps connected to each tank with short hook up wires, plus no shielding.

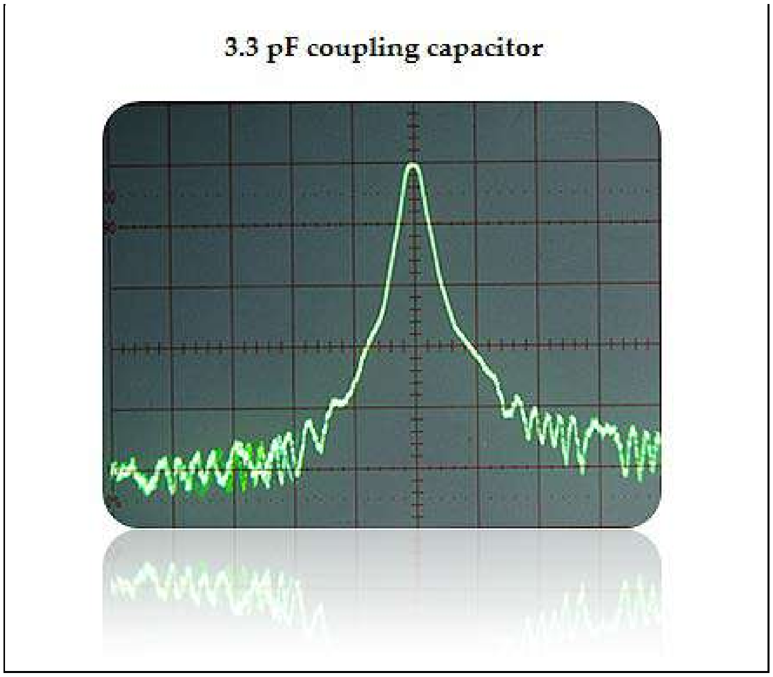

3.3 pF coupling capacitor

Above — The SA + TG sweep with a 3.3 pF coupling capacitor. The peak isn't as sharp, but still looks good. As shown, increasing the coupling capacitor value increases the 3 dB filter bandwidth with all other components equal.

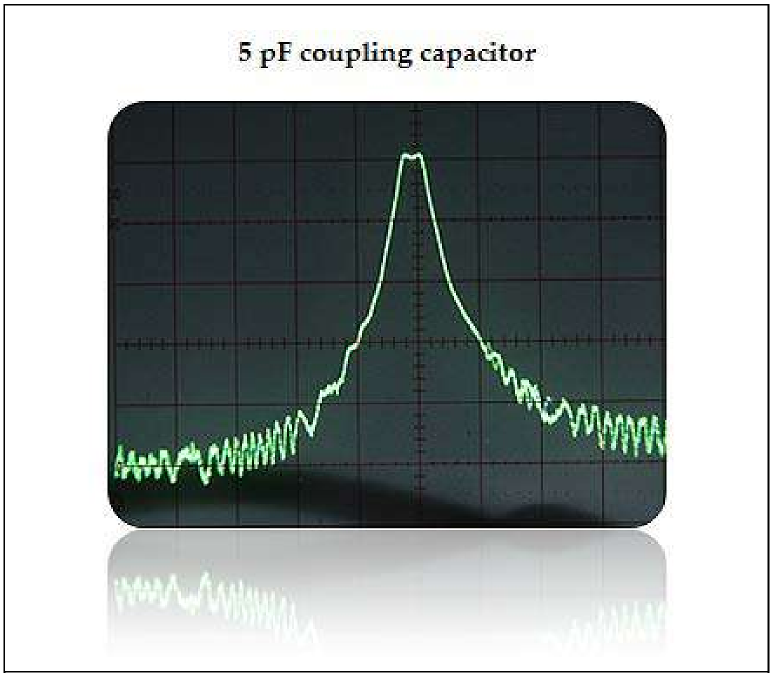

5 pF coupling capacitor

Above — With a 5 pF coupling capacitor, a double humped response appeared. The bandwidth further increases.

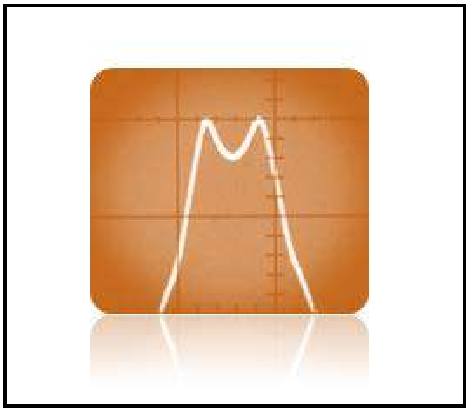

Above — A zoom of the double humped filter response employing a 7.5 pF coupling capacitor. Imagine the difficulty tuning this band-pass filter in a receiver by listening to band noise. Tuning in either peak skews the filter bandwidth. Additionally, the 3dB bandwidth now = ~ 1.6 MHz — Not a good filter!

Optimizing Return Loss

Despite trying, I could not obtain a better return loss than 10 -12 dB by changing the tap point on the 7 turn inductors. In part this was due to limited potential autotransformer ratios on a 7 turn coil. I emailed Wes, W7ZOI and he sent this file. I learned that adding a series capacitor to each end will tune the filter to 50 ohms impedance. What capacitor value should we use?

The answer can be found purely experimentally, or with Ladbuild08 to make a digital file of your filter and GPLA08 to analyze it.

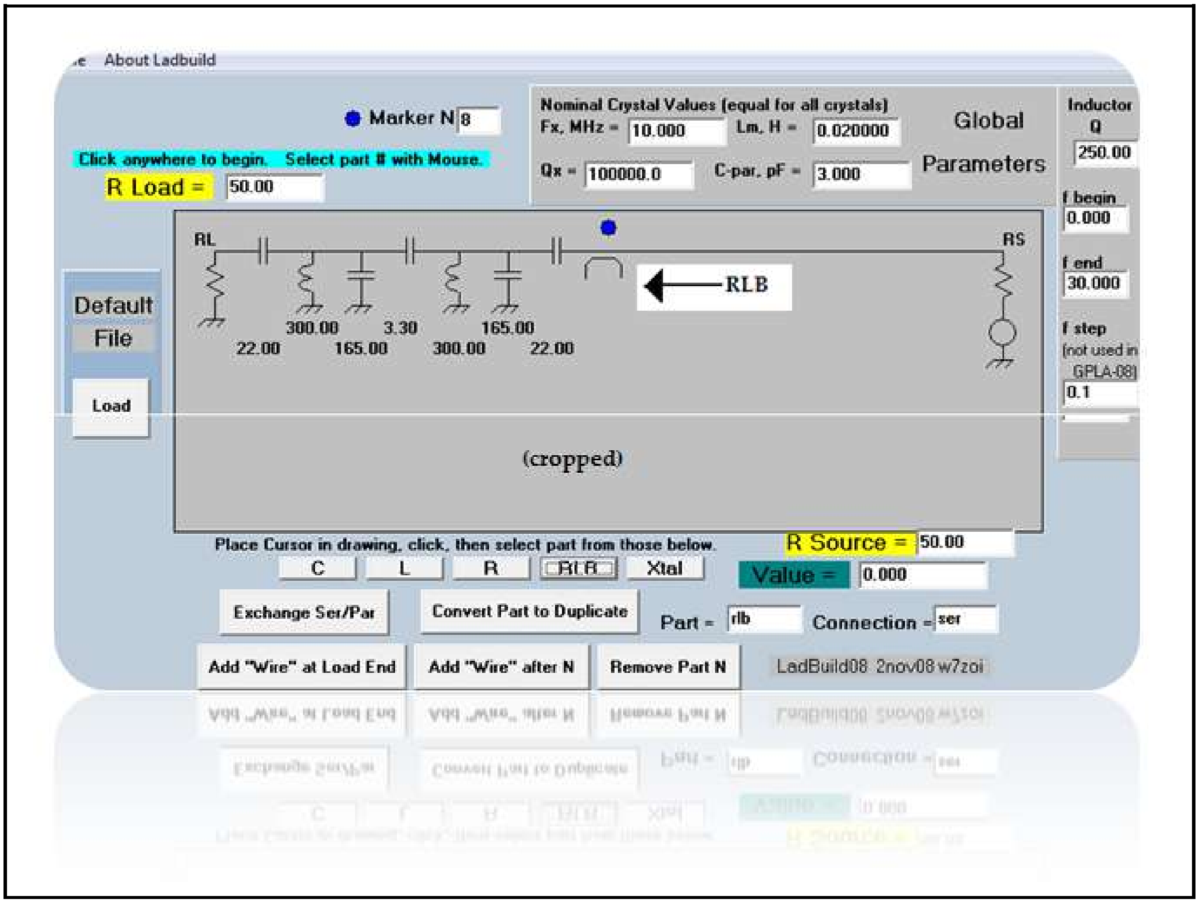
Marker N 8
R Load = 50.00
Nominal Crystal Values (equal for all crystals)
Fx, MHz = 10.000
Lm, H = 0.020000
Qx = 100000.0
C-par, pF = 3.000
Inductor Q 250.00
f begin 0.000
f end 30.000
f step (not used in GPLA-08) 0.1
R Source = 50.00
Value = 0.000
Part = rlb
Connection = ser
About Ladbuild
Click anywhere to begin. Select part # with Mouse.
Global Parameters
Default File
Load
RL
RS
300.00
3.30
165.00
22.00
165.00
300.00
22.00
RLB
(cropped)
Place Cursor in drawing, click, then select part from those below.
C
L
R
RLB
Xtal
Exchange Ser/Par
Convert Part to Duplicate
Add "Wire" at Load End
Add "Wire" after N
Remove Part N
LadBuild08 2nov08 w7zoi

Above — I "built up" my Figure 1 filter in Ladbuild08 with a 3.3 pF coupling capacitor. Initially I guessed at the values for the series end capacitors and knew my tuning capacitor were ~ 165 pF because I removed and measured them from the peaked filter from Figure 1 and added a few pF for stray capacitance. Any of these values can be changed in GPLA, so educated guessing is okay.

For size 50 to 68 toroidal inductors, many builders choose a Qu value from 200 - 250 with # 6 material. Qu affects insertion loss and to some extent, return loss. Click for a tutorial from Wes', W7ZOI site and consult EMRFD for more information.
In order for GPLA08 to display an S11 plot (return loss), a return loss bridge (RLB) must be added as shown. Also check the **Plot S11** check box in GPLA08.

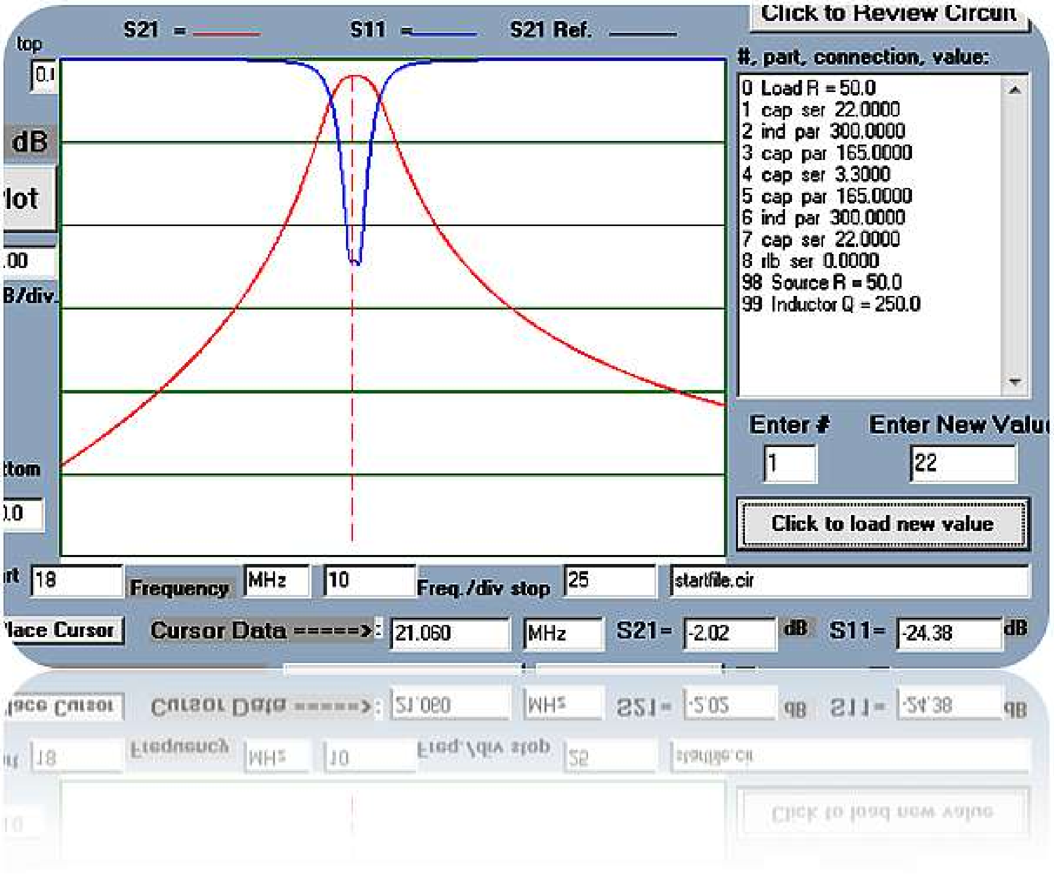
Click to Review Circuit
S21 =
S11 =
S21 Ref.
top
dB
lot
B/div.
ttom
#, part, connection, value:
Enter #
Enter New Value
Click to load new value
Frequency
Freq./div stop
lace Cursor
Cursor Data =====>:

0 Load R = 50.0
1 cap ser 22.0000
2 ind par 300.0000
3 cap par 165.0000
4 cap ser 3.3000
5 cap par 165.0000
6 ind par 300.0000
7 cap ser 22.0000
8 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 250.0

top: 0.1
.00 dB/div.
ttom: .0
Enter #: 1
Enter New Value: 22
rt: 18 Frequency MHz 10 Freq./div stop 25
startfile.cir
Cursor Data =====>: 21.060 MHz S21= -2.02 dB S11= -24.38 dB

Above — The GPLA 08 filter simulation of the filter "built" with Ladbuild08 above.

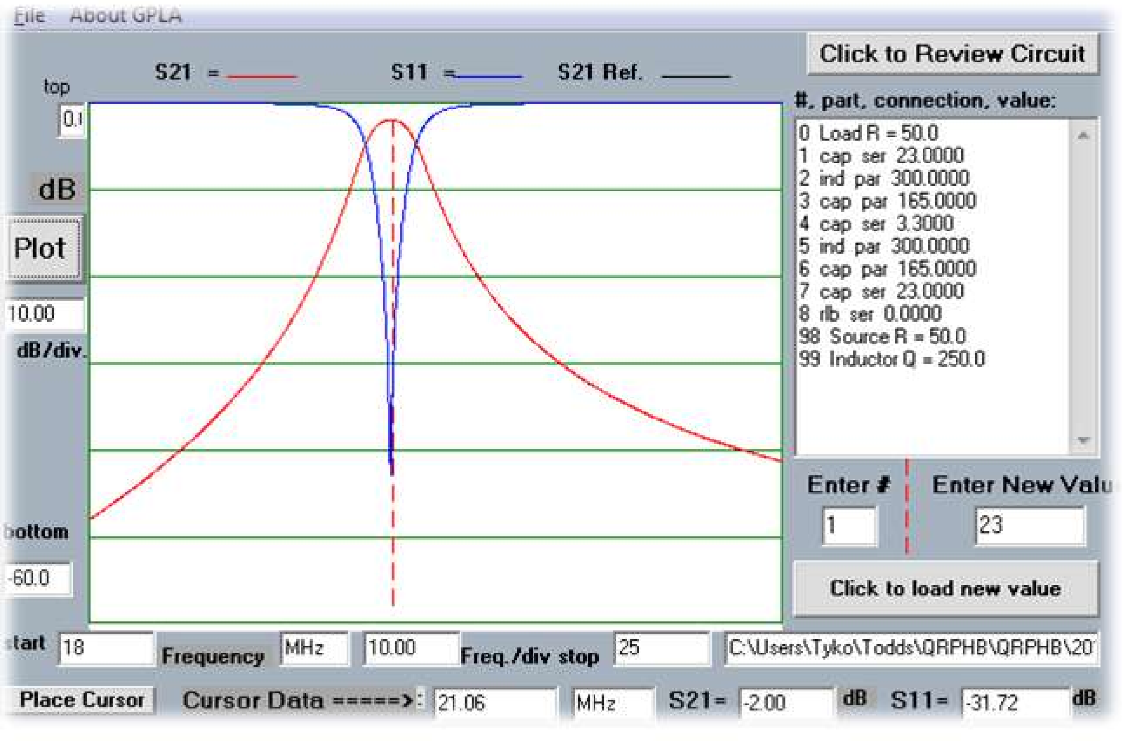
File
About GPLA
Click to Review Circuit
S21 =
S11 =
S21 Ref.
top
dB
Plot
dB/div.
bottom
#, part, connection, value:
Enter #
Enter New Valu
Click to load new value
Frequency
Freq./div stop
Place Cursor
Cursor Data =====>:

0 Load R = 50.0
1 cap ser 23.0000
2 ind par 300.0000
3 cap par 165.0000
4 cap ser 3.3000
5 ind par 300.0000
6 cap par 165.0000
7 cap ser 23.0000
8 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 250.0

top: 0.1
10.00 dB/div.
bottom: -60.0
Enter #: 1
Enter New Valu: 23
start: 18 Frequency MHz 10.00 Freq./div stop 25
C:\Users\Tyko\Todds\QRPHB\QRPHB\20
Cursor Data =====>: 21.06 MHz S21= -2.00 dB S11= -31.72 dB

Above — In this filter simulation, I tweaked the end capacitors (parts #1 and #7) from 22 to 23 pF and watched the return loss (S11) improve by 7.34 dB — if wanted, you can optimize the end capacitor values to improve the match into 50 ohms. To re-establish the center frequency, slight retuning of parallel capacitors #3 and 6 is required when changing the series end capacitors; although I specifically didn't change them for this example.

Increasing the 2 end capacitors to increase S11 renders an option only; you don't have to go for the best S11 in your filters. Increasing the series end capacitors to bump up return loss tends to increase the 3 dB bandwidth and reduce insertion loss.

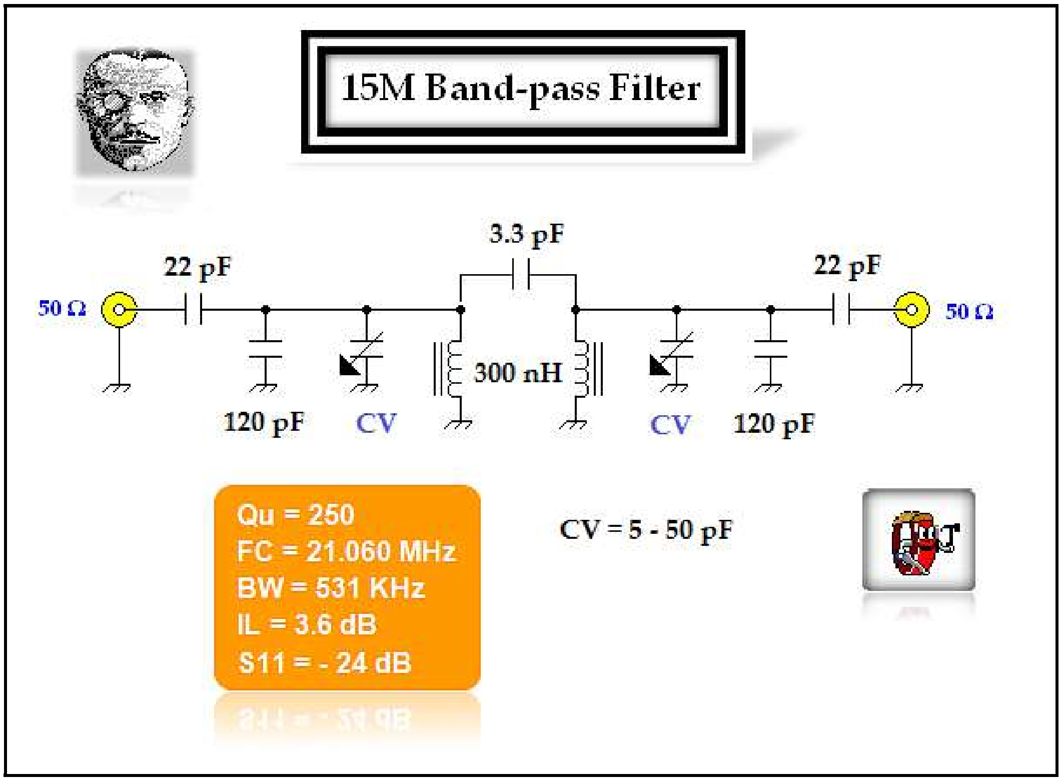

Above — I built and measured the filter with 22 pF end capacitors since these are common, standard values. In another experiment, a 1 pF cap was soldered in parallel with each series capacitor and the return loss increased by about 4 dB. Click for a bench photo of an alternate version of the above filter. Clearly, GPLA08 simulation furnishes us popcorn builders with a starting point to make top-notch band-pass filters.

Click for another simulation of a filter employing a 2 pF coupling capacitor, with the end capacitors tweaked for the best S11. *S-11 is just the negative of the return loss.* I would certainly use this filter in the front end of a popcorn direct conversion receiver.

An easier way to design your band-pass filters involves using DTC08 to design a raw filter and GPLA08 to substitute in standard value capacitors and tweak your filter. That's part 2. The material presented in this section supports the discussion in Part 2 and 3.

Part 2: Band-pass Filter Design using DTC08

Prior to using these Ladpac programs, some numbered design points and a preamble follow.

More than anything else, our parts collection dictates what filter parameters we choose and end up with. For example, if you want filters with a low bandwidth such as 150 KHz and under, you'll require inductors and capacitors that provide really high Q, or you might suffer from punishing insertion loss.

The following are general starting points only — *your* needs, parts and abilities drive *your* filter design. Example variances include: if a low noise amplifier follows a filter, a higher insertion loss might be okay; a high return loss is not always required for a low noise figure; especially in popcorn receivers. Also, it's a viable choice to trade off insertion loss for steep skirts in some filters.

1. A reasonable 3 dB bandwidth = 100 to 500 KHz, but this depends on the purpose of the filter.
Numerous considerations challenge us. Will this be a whole band (CW + SSB) filter, or a CW only filter? As a CW op who uses simple equipment, I tend to design moderate bandwidth (200-300 KHz) CW-only filters. If you need CW + SSB, then a bandwidth of 350 KHz or greater might suit you. It's really up to you.

Other factors affecting bandwidth choice include whether the filter drives a superheterodyne or a direct conversion receiver. In superheterodyne receivers, your intermediate frequency informs your filter bandwidth choice. Consider the following 2 diagrams:

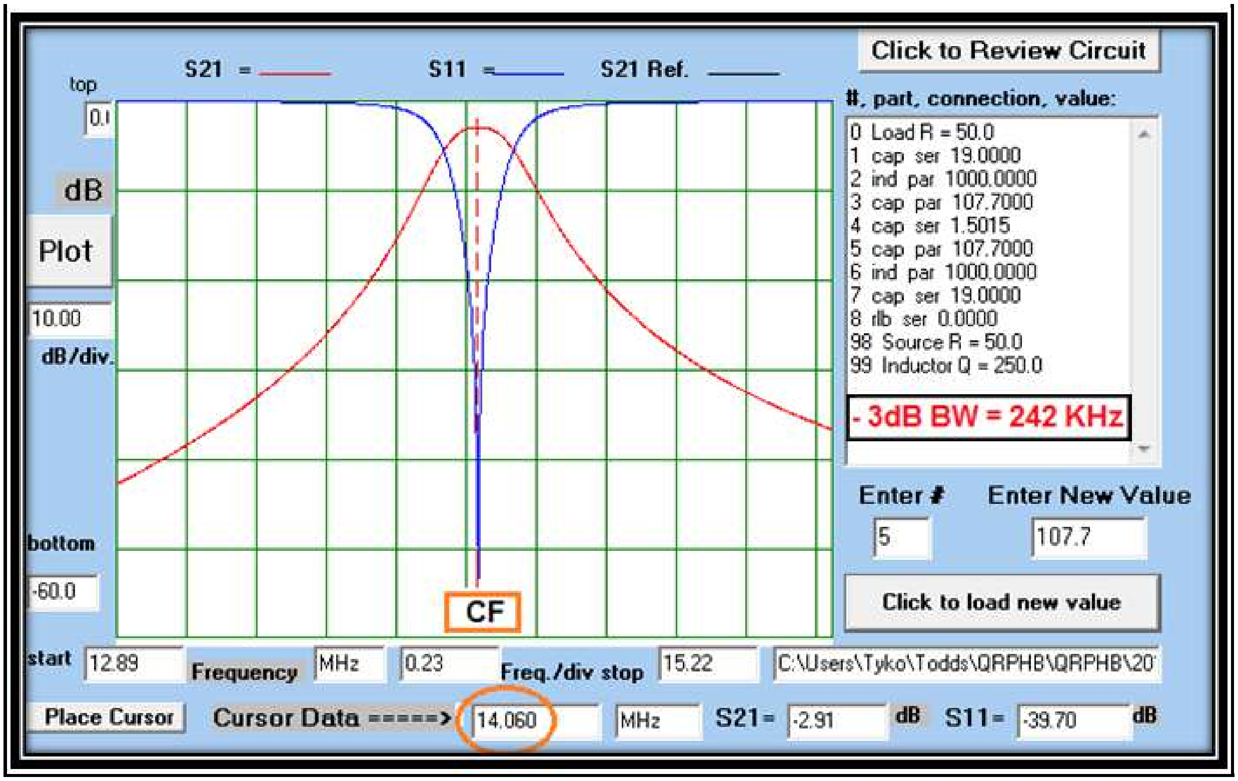
Click to Review Circuit
S21 =
S11 =
S21 Ref.
top
0.0
dB
Plot
dB/div.
bottom
CF
Enter #
Enter New Value
Click to load new value
Frequency
Freq./div
Place Cursor
Cursor Data =====>

#, part, connection, value:

0 Load R = 50.0
1 cap ser 19.0000
2 ind par 1000.0000
3 cap par 107.7000
4 cap ser 1.5015
5 cap par 107.7000
6 ind par 1000.0000
7 cap ser 19.0000
8 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 250.0

- 3dB BW = 242 KHz

Enter # 5 Enter New Value 107.7

dB/div. 10.00 bottom -60.0

start 12.89 MHz Freq./div 0.23 stop 15.22 C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====> 14.060 MHz S21= -2.91 dB S11= -39.70 dB

Above — Using DTC08 and GPLA08, I designed an example filter for the front end of a 14 MHz superheterodyne receiver with an 11 MHz IF. BW = 242 KHz. CF = 14.060 MHz; a frequency some QRP operators favor.

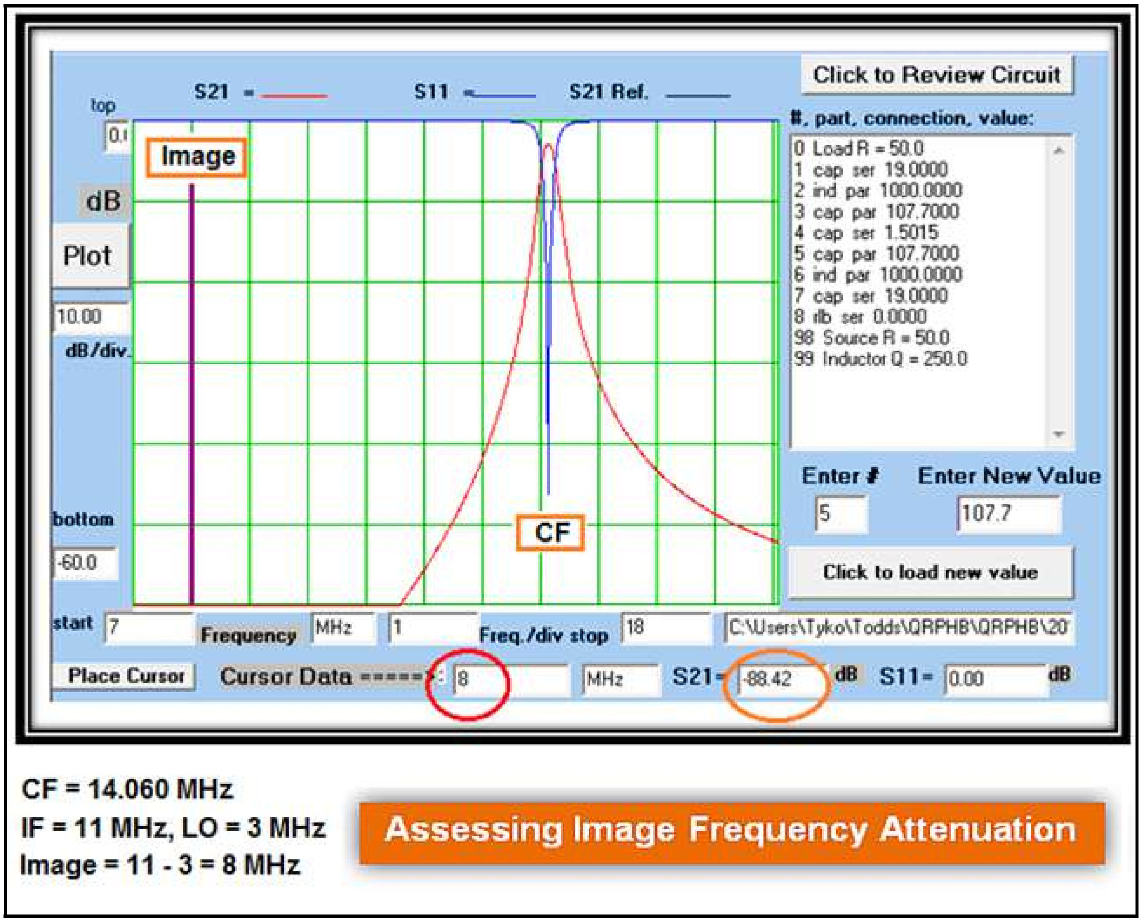
Click to Review Circuit
S21 =
S11 =
S21 Ref.
top
0.0
Image
dB
Plot
dB/div.
bottom
CF
Enter #
Enter New Value
Click to load new value
Frequency
Freq./div
Place Cursor
Cursor Data =====>
Assessing Image Frequency Attenuation

#, part, connection, value:

0 Load R = 50.0
1 cap ser 19.0000
2 ind par 1000.0000
3 cap par 107.7000
4 cap ser 1.5015
5 cap par 107.7000
6 ind par 1000.0000
7 cap ser 19.0000
8 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 250.0

Enter # 5 Enter New Value 107.7

dB/div. 10.00 bottom -60.0

start 7 MHz Freq./div 1 stop 18 C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====> 8 MHz S21= -88.42 dB S11= 0.00 dB

CF = 14.060 MHz
IF = 11 MHz, LO = 3 MHz
Image = 11 - 3 = 8 MHz

Above — Assessing filter attenuation at the image frequency using GPLA08. *To keep the arithmetic simple, I employed a frequency of 14 MHz for the image frequency calculation.* As shown, the simulated attenuation of my 8 MHz image frequency is 88.42 dB. Since I personally target an image frequency suppression of 60-70 dB; at 88 dB, if I wanted, I could increase the bandwidth of this filter for broader coverage and reduced insertion loss.

How much image frequency rejection is needed for superheterodyne receivers? I'm uncertain, for I have seen competent authors choose between 50 and 100 dB. I feel a good target = 60 - 70 dB, and 50 dB is the bare minimum. To realize image attenuation above 50 dB, shielding is usually required.

Three or more L-C tank band-pass filters may be required when your image frequency is close to the IF frequency. Choose both your intermediate frequency and your bandwidth wisely.

2. After selecting your bandwidth, *tweak the inductance* and only if necessary, make minor adjustments to your set 3 dB bandwidth to give standard, or near-standard value coupling capacitors that you *own.* Obviously, you can place fixed capacitors in series or parallel, or even couple your resonators with a variable capacitor.

3. I favor size 50 to 80 powdered iron toroids with number 2, 6, or 10 material for a reasonably high Qu.

4. I aim for an insertion loss of of 3-4 dB; especially above the 40 Meter band; *consider the variances discussed earlier*

5. I aim for a return loss of at least 20 dB; *consider the variances discussed earlier*

6. If you can, measure your bread boarded filter bandwidth to confirm or improve the GLPA simulation. Insertion and return loss are easily measured — see EMRFD and the RF Workbench web pages on this site for methods.

I provide no graphic tutorial of DTC08; however, some work flow suggestions follow:

Open up DTC08, choose your center frequency, Qu, inductance and bandwidth and then press the **Calculate** button. Adjust the L until you get close to a standard value coupling capacitor from your parts bin. If required, you may also tweak the bandwidth value to get the needed coupling capacitor. It's wise to change the L before BW since changes in inductance don't cause too many complications within limits.

Name and save your filter to a specific file system directory or folder; or simply save it as the default file.

Open GPLA08 and load your recently saved filter file. Press the **Plot** button and then the **Click to Review Circuit** button. In some cases, you will have to type the CF in the *Cursor Data* text box and press **Plot** to set the cursor at your center frequency.

Change the coupling capacitor(s) to a standard value using the **Enter New Value** data entry controls.

Adjust the series end coupling capacitors to standard values. and if S11 is an issue for you, tweak them up and down while observing S11. Re-establish your center frequency by tweaking the parallel tank tuning capacitors and then re-plot to ensure the CF is lined up with the center of the plot.

In Part 3, I provide 3 filter design examples. Your own filter designs will be the most important examples to study.

Part 3: Band-pass Filter Examples

Example 1: An 80 Meter Band Filter

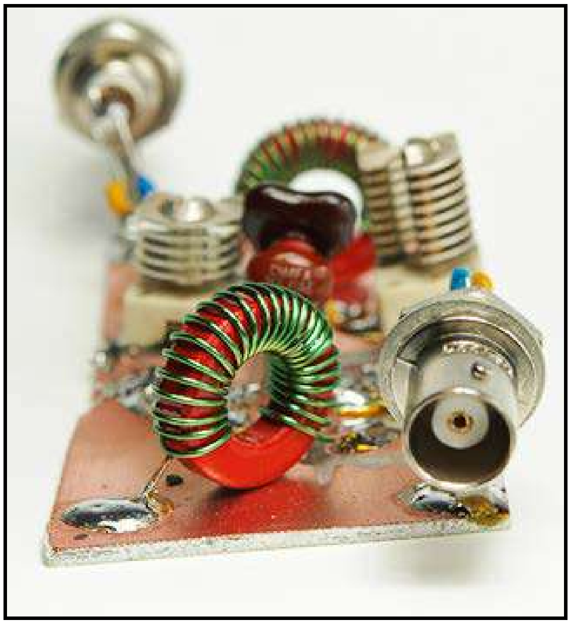

Above — Breadboard photograph of the 80M filter. This example filter may hit home for you — I like listening to CW at and below the 3560 KHz QRP calling frequency, however, another local Ham likes to talk on 75 meters SSB at or above 3790 KHz. This situation calls for a narrow band-pass filter. With my filter, the attenuation at 3790 KHz = ~ 24 dB; had I built a wide bandwidth filter, for example, 350 KHz BW; the attenuation at 3790 KHz, would only be ~4 dB. Perhaps a 3 resonator filter with even steeper skirts would be better?

I'll show the design process from start to test.

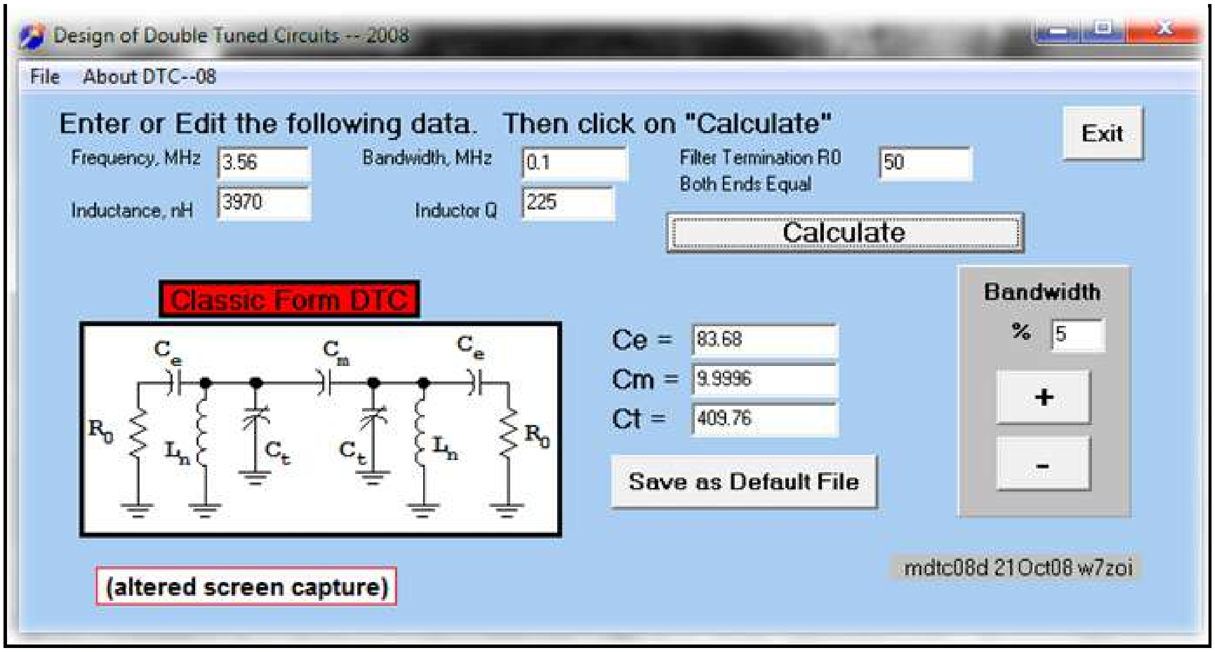
Design of Double Tuned Circuits -- 2008
File
About DTC--08
Exit
Calculate
Classic Form DTC
Save as Default File
Bandwidth
+
-
(altered screen capture)
mdtc08d 21Oct08 w7zoi
C_e
C_m
C_e
R_0
L_n
C_t
C_t
L_n
R_0

Enter or Edit the following data. Then click on "Calculate"

Frequency, MHz 3.56

Bandwidth, MHz 0.1

Filter Termination R0 Both Ends Equal 50

Inductance, nH 3970

Inductor Q 225

Ce = 83.68

Cm = 9.9996

Ct = 409.76

% 5

Above — The basic DTC08 data entry fields were populated. I chose a 100 KHz bandwidth and tried different L values until Cm = 10 pF, since I have a whole drawer of 10 pF capacitors. I believed my Inductor Q would be at least 225 and wound 25 turns of #22 AWG on a T68-2 toroid and expanded or contracted the windings until I measured 3970 nanohenries. In reality, we should measure the inductor Q and in future I will, however, my sense is that few builders do.

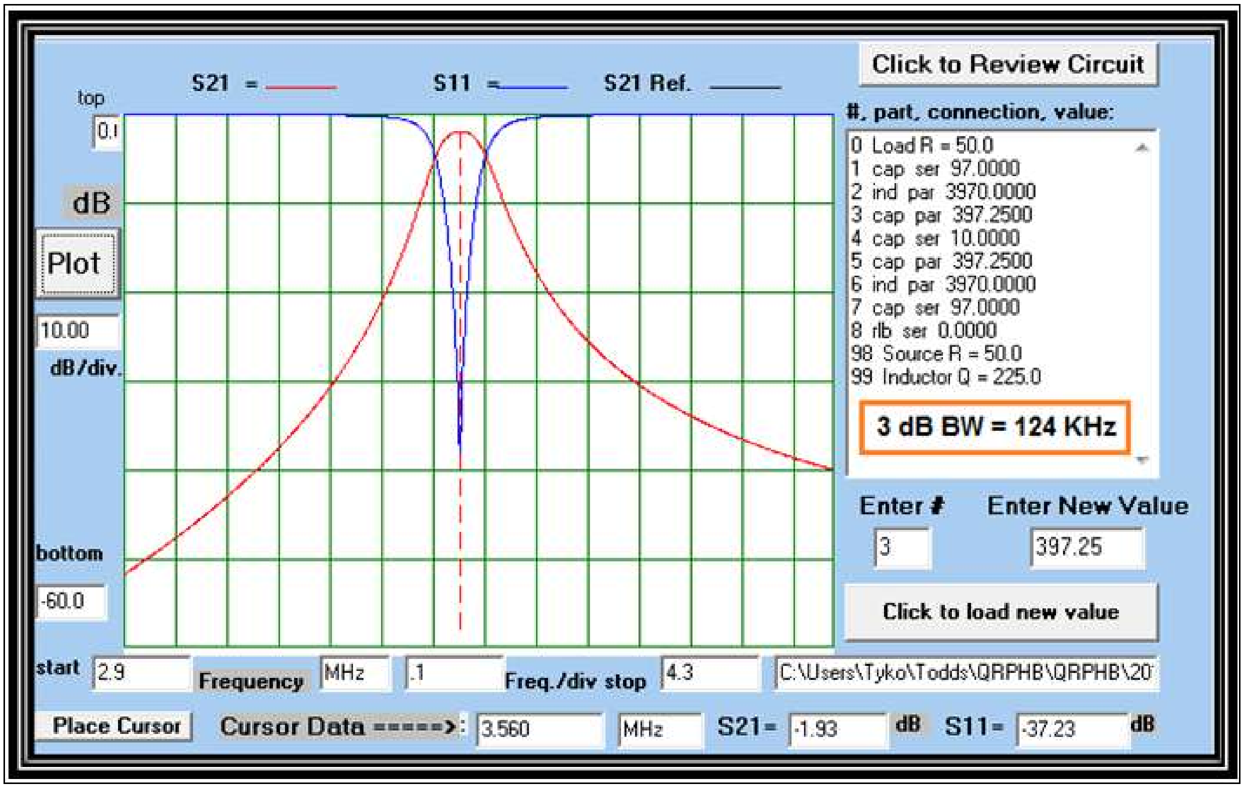
Click to Review Circuit
S21 =
S11 =
S21 Ref.
top
dB
Plot
dB/div.
bottom
Click to load new value
Place Cursor
Frequency
Freq./div
stop

top 0.0

10.00 dB/div.

bottom -60.0

#, part, connection, value:

0 Load R = 50.0
1 cap ser 97.0000
2 ind par 3970.0000
3 cap par 397.2500
4 cap ser 10.0000
5 cap par 397.2500
6 ind par 3970.0000
7 cap ser 97.0000
8 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 225.0

3 dB BW = 124 KHz

Enter # 3

Enter New Value 397.25

start 2.9 Frequency MHz .1 Freq./div stop 4.3 C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====>: 3.560 MHz S21= -1.93 dB S11= -37.23 dB

Above — After saving my filter, I opened it up in GPLA08 as above. I replaced #4 with a standard 10.0 pF value, and started tweaking the end caps; parts #1 and #7 to gain a better S11 per my obsession with return loss. Retuning #3 and #7 re-establishes the center frequency and allows the S21 and S11 values to be interpreted. I settled on this filter and headed for the bench.

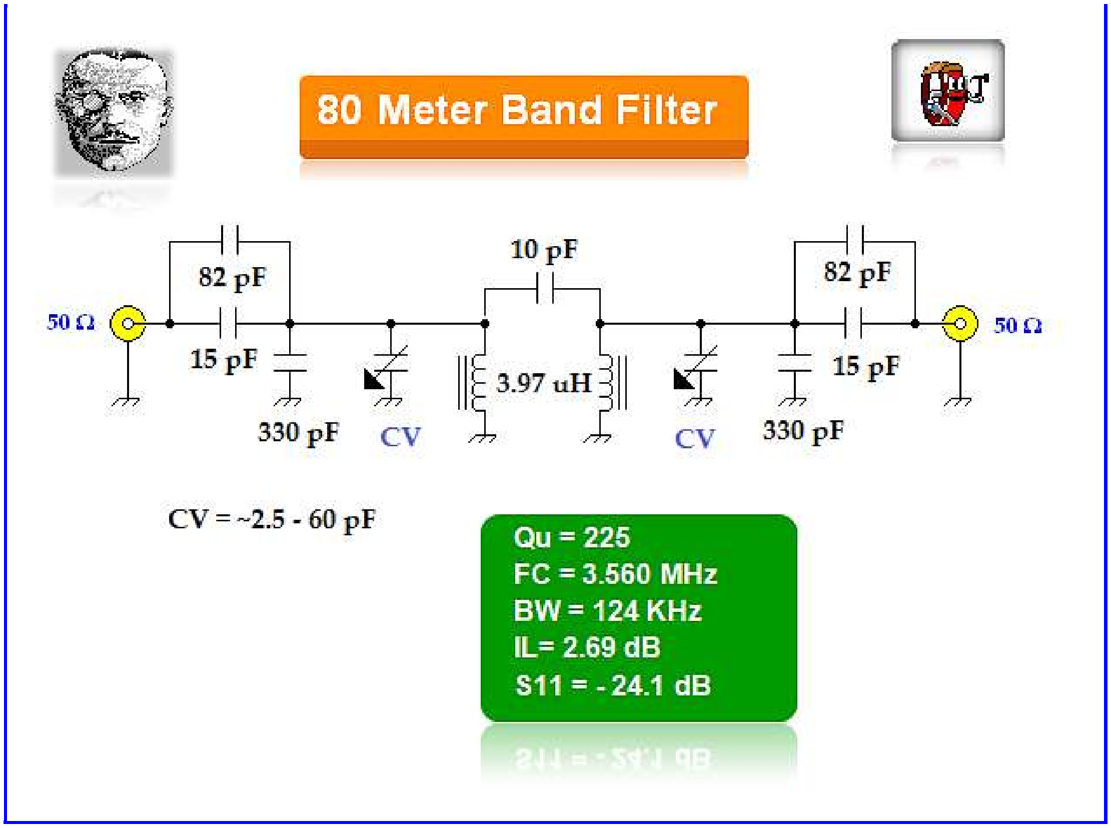

Above — Schematic and analysis of the breadboard. Click for another photograph. In reality, I bench determined the exact capacitance needed to tune each tank at 3.56 MHz with 2 large air variable capacitors that I removed and measured after peaking the filter. For each tank, I try to get just below this value with fixed value capacitors and add a small (2.5 - 22 pF, or so) air variable trimmer capacitor for peaking.

You need to test with capacitance under and over that required to ensure you properly tuned each L-C tank to resonance. Your parts collection, stray capacitance, mistakes or inductance variations in the toroids necessitate custom tuning of your tanks on the bench. I give capacitance values that should work, but it's up to you to ensure resonance of each tank. I find narrow BW filters require a steady hand to tune.

After peaking the tank in your oscilloscope, record the peak-to-peak voltage. Remove the filter and connect your signal generator to your scope with an RF barrel connector and again record the peak to peak voltage. The difference between the 2 is your insertion loss. You can calculate IL with Applet H on this page . Next, perform return loss measurements. If you can, determine the true 3 dB bandwidth of your filter by sweeping it with SA plus a generator. My filter 3 dB BW = 124 KHz.

Example 2: 20 Meter Band Superheterodyne Receiver Filter

A fictitious builder wants a superheterodyne receiver that covers 14.0 - 14.350. His IF = 2 MHz. The local oscillator = 12 MHz. The image frequency = (12 - 2) = 10 MHz. He centers his filter at 14.020 MHz. In this simulation-only example, we'll go from 1 resonator to 3.

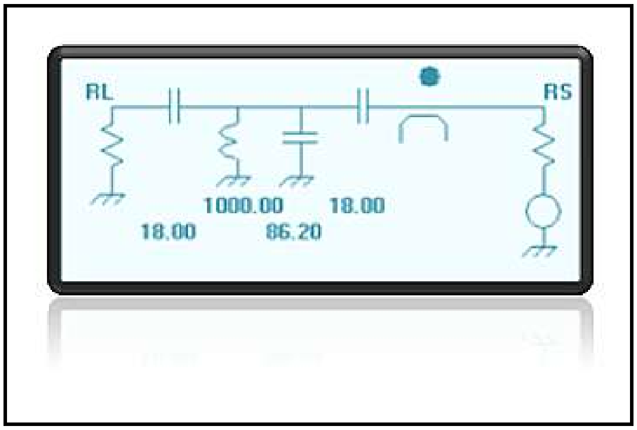

Above — A single resonator with series matching capacitors "built up" in Ladbuild08.

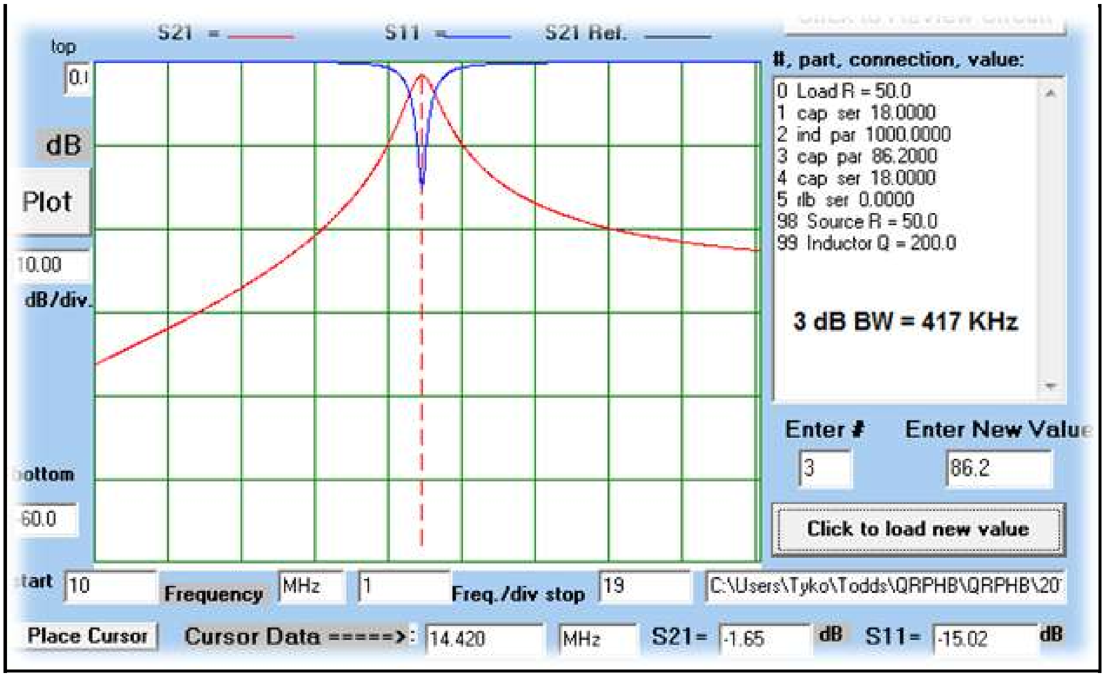
S21 =
S11 =
S21 Ref.
dB
Plot
Place Cursor
Click to load new value

#, part, connection, value:
0 Load R = 50.0
1 cap ser 18.0000
2 ind par 1000.0000
3 cap par 86.2000
4 cap ser 18.0000
5 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 200.0

3 dB BW = 417 KHz

top 0.
10.00 dB/div.
ottom -60.0
tart 10 Frequency MHz 1 Freq./div stop 19
C:\Users\Tyko\Todds\QRPHB\QRPHB\20
Enter # 3 Enter New Valu 86.2
Cursor Data =====> 14.420 MHz S21= -1.65 dB S11= -15.02 dB

Above — The GPLA08 plot of the single tank filter. The bandwidth = 417 KHz. Increasing the end capacitors to 22 pF to try to increase return loss increases the 3 dB bandwidth as shown here, so we better stick with the original design.

In GPLA simulations with a perfectly centered filter, S21 = the insertion loss and S11 is negative of the return loss.

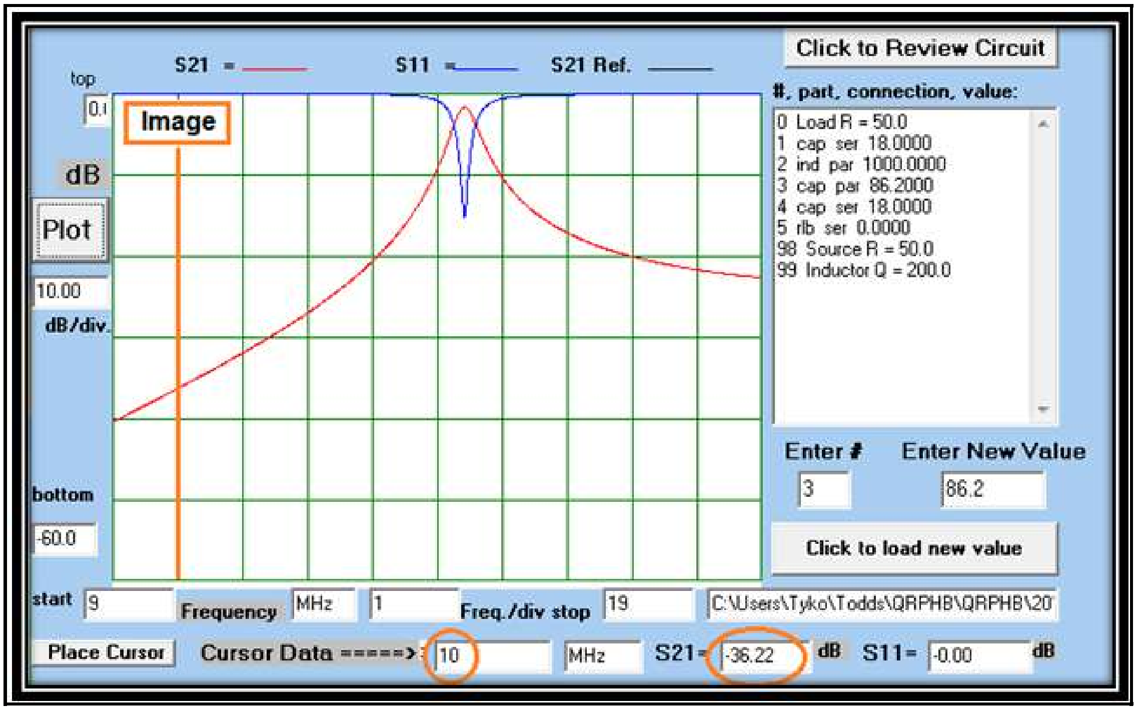
Click to Review Circuit
S21 =
S11 =
S21 Ref.
Image
dB
Plot
Place Cursor
Click to load new value

#, part, connection, value:
0 Load R = 50.0
1 cap ser 18.0000
2 ind par 1000.0000
3 cap par 86.2000
4 cap ser 18.0000
5 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 200.0

top 0.
10.00 dB/div.
bottom -60.0
start 9 Frequency MHz 1 Freq./div stop 19
C:\Users\Tyko\Todds\QRPHB\QRPHB\20
Enter # 3 Enter New Value 86.2
Cursor Data =====> 10 MHz S21= -36.22 dB S11= -0.00 dB

Above — Assessing image frequency attenuation in GPLA08; this sucks — only 36.2 dB down. We need to add a tank.

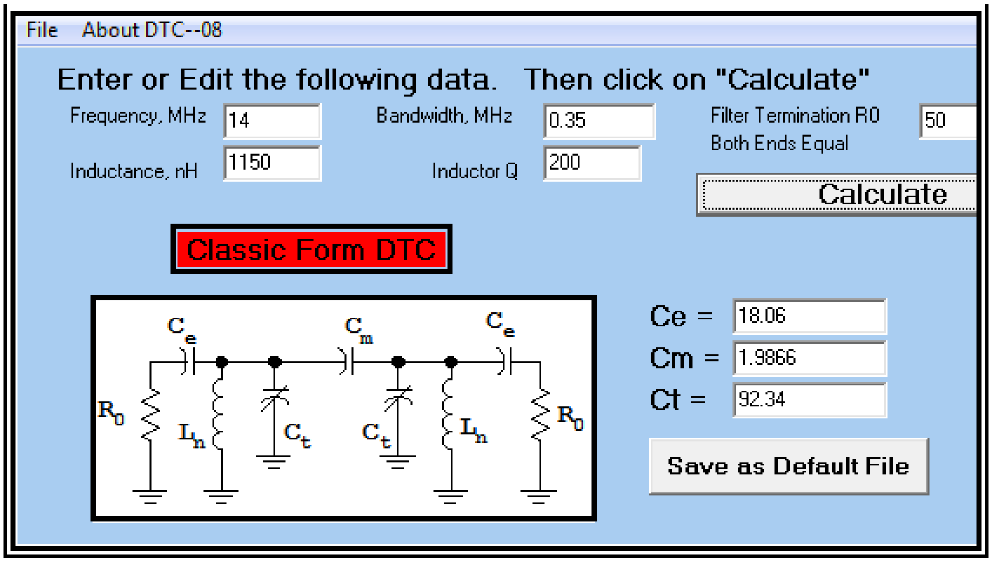
File
About DTC--08
Calculate
Classic Form DTC
Save as Default File
Ce
Cm
Ce
R0
Ln
Ct
Ct
Ln
R0

Enter or Edit the following data. Then click on "Calculate"

Frequency, MHz 14

Bandwidth, MHz 0.35

Filter Termination R0 Both Ends Equal 50

Inductance, nH 1150

Inductor Q 200

Ce = 18.06

Cm = 1.9866

Ct = 92.34

Above — Building up a filter in DTC08, I increased the L from 1000 to 1150 nH to give a Cm near to a standard value.

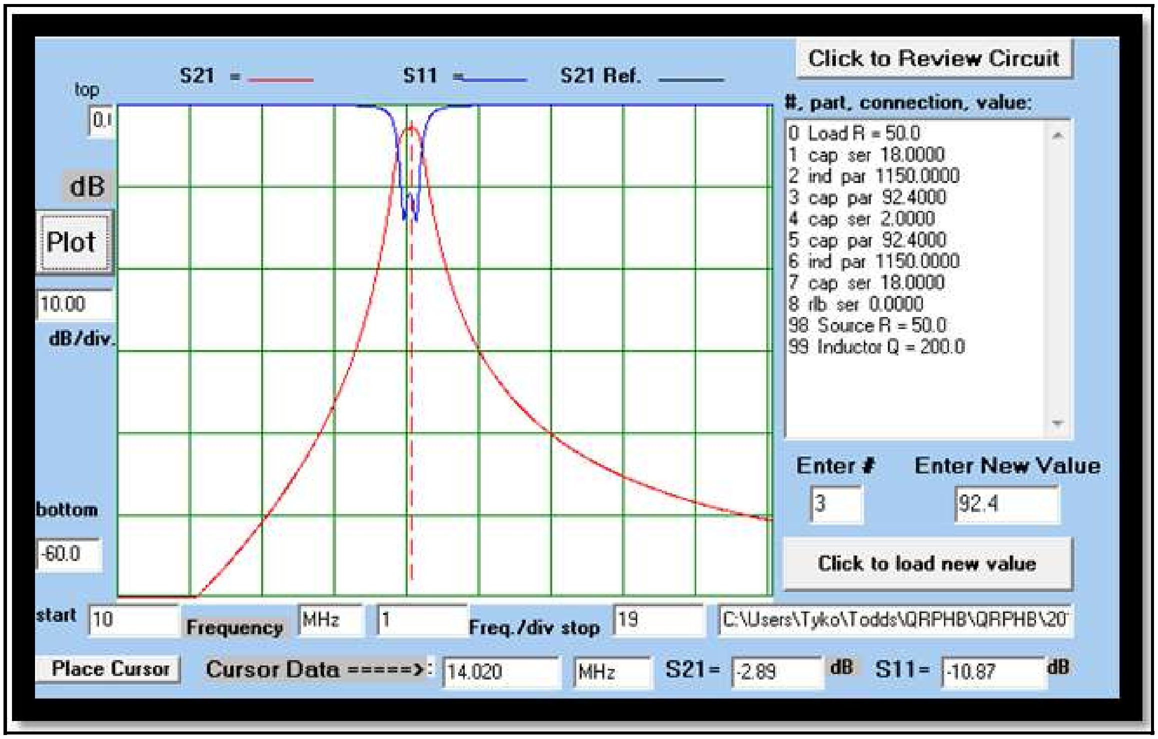
S21 =
S11 =
S21 Ref.
Click to Review Circuit
dB
Plot
dB/div.
Place Cursor
Click to load new value

top 0.(

10.00 dB/div.

bottom -60.0

start 10 Frequency MHz 1 Freq./div stop 19 C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====> 14.020 MHz S21= -2.89 dB S11= -10.87 dB

#, part, connection, value:

```
0  Load R = 50.0
1  cap  ser  18.0000
2  ind  par  1150.0000
3  cap  par  92.4000
4  cap  ser  2.0000
5  cap  par  92.4000
6  ind  par  1150.0000
7  cap  ser  18.0000
8  rlb  ser  0.0000
98  Source R = 50.0
99  Inductor Q = 200.0
```

Enter # 3

Enter New Value 92.4

Above — The GPLA08 plot of the double tuned filter. I set #4 to 2 pF and #1 and #7 to 18 pF (nearest standard values). #3 and #5 were slightly tweaked to center the filter. The simulated IL is only up 1.24 dB from the single resonator version. You are probably wondering why I didn't design the filter for a CF = 14.020 MHz in DTC08 above to keep consistency. I probably should have, but wanted to illustrate the versatility of GPLA08 to center filters "on the fly".

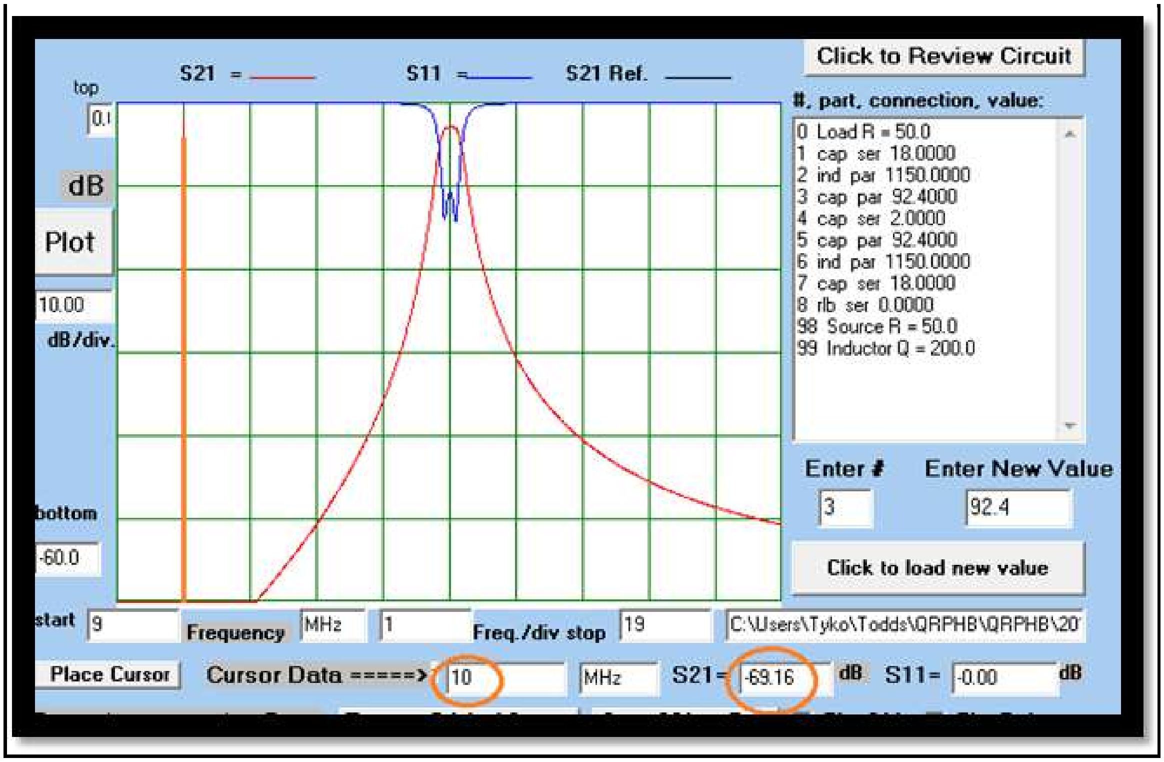
Click to Review Circuit
S21 =
S11 =
S21 Ref.
dB
Plot
Click to load new value
Place Cursor
Frequency
MHz

top 0.

10.00 dB/div.

bottom -60.0

start 9 Frequency MHz 1 Freq./div stop 19 C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====> 10 MHz S21= -69.16 dB S11= -0.00 dB

#, part, connection, value:

0 Load R = 50.0
1 cap ser 18.0000
2 ind par 1150.0000
3 cap par 92.4000
4 cap ser 2.0000
5 cap par 92.4000
6 ind par 1150.0000
7 cap ser 18.0000
8 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 200.0

Enter # 3

Enter New Value 92.4

Above — Assessment of the 10 MHz image — now it's 69.2 dB down. Although this filter will work well for his particular receiver specifications, this fastidious builder wants even greater image attenuation and decides to add a third resonator!

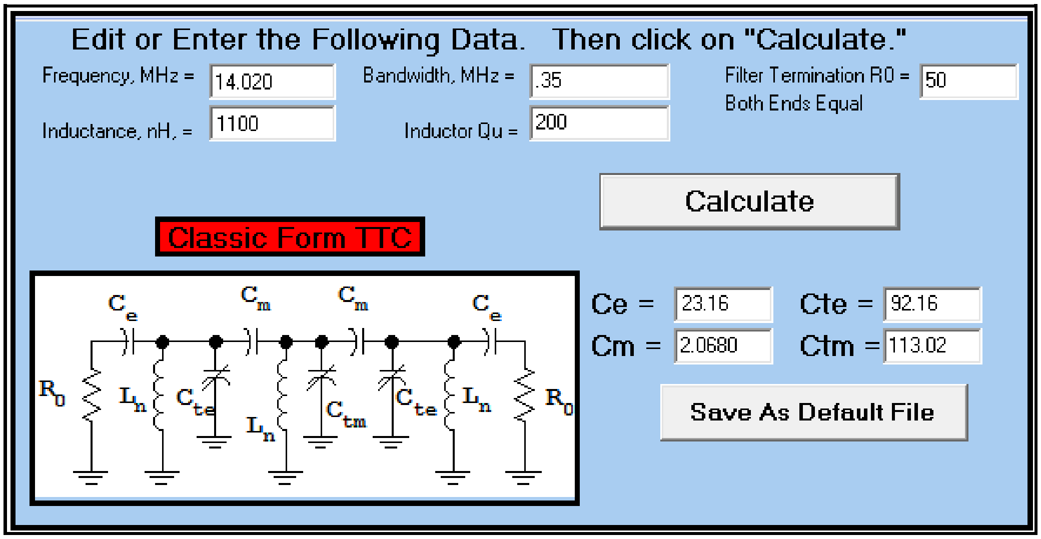
Calculate
Classic Form TTC
Save As Default File
C_e
C_m
C_m
C_e
R_0
L_n
C_{te}
L_n
C_{tm}
C_{te}
L_n
R_0

Edit or Enter the Following Data. Then click on "Calculate."

Frequency, MHz = 14.020

Bandwidth, MHz = .35

Filter Termination R0 = 50
Both Ends Equal

Inductance, nH, = 1100

Inductor Qu = 200

Ce = 23.16

Cte = 92.16

Cm = 2.0680

Ctm = 113.02

Above — Building up a filter in TTC08, I chose an L of 1100 nH to give a Cm close to a standard value.

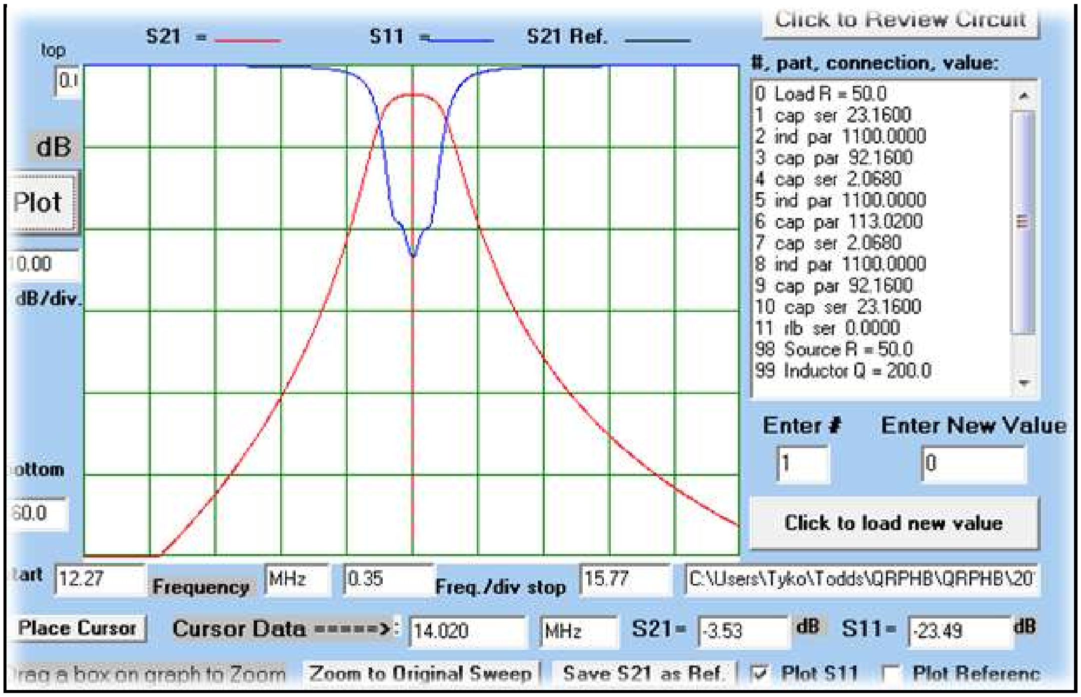
S21 =
S11 =
S21 Ref.
Click to Review Circuit
top
0.0
dB
Plot
dB/div.
ottom
Enter #
Enter New Value
Click to load new value
Frequency
Freq./div
Place Cursor
Zoom to Original Sweep
Save S21 as Ref.
☑ Plot S11
☐ Plot Referenc
rag a box on graph to Zoom

#, part, connection, value:

0 Load R = 50.0
1 cap ser 23.1600
2 ind par 1100.0000
3 cap par 92.1600
4 cap ser 2.0680
5 ind par 1100.0000
6 cap par 113.0200
7 cap ser 2.0680
8 ind par 1100.0000
9 cap par 92.1600
10 cap ser 23.1600
11 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 200.0

Enter # 1 Enter New Value 0

0.00 dB/div.

ottom 60.0

tart 12.27 Frequency MHz 0.35 Freq./div stop 15.77 C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====> 14.020 MHz S21= -3.53 dB S11= -23.49 dB

Above — The GPLA plot of the 3 tank filter. I performed no parts tweaking — it's up to you from here on in. The simulated IL remains quite reasonable.

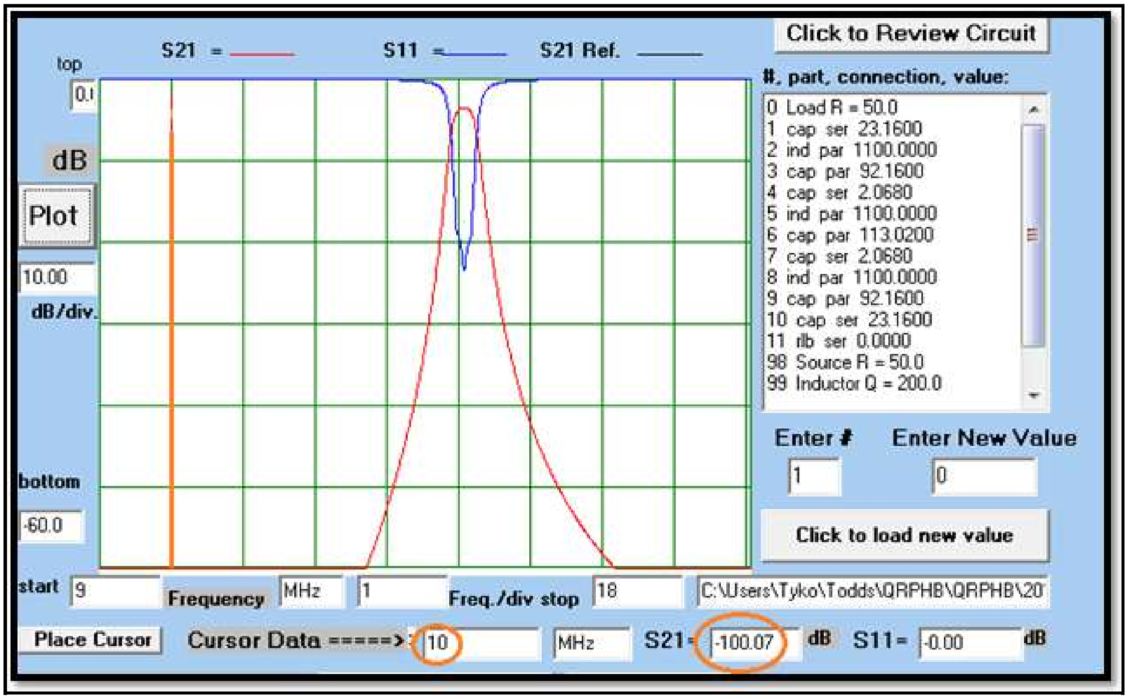
S21 =
S11 =
S21 Ref.
Click to Review Circuit
top
0.0
dB
Plot
dB/div.
bottom
Enter #
Enter New Value
Click to load new value
Frequency
Freq./div
Place Cursor

#, part, connection, value:

0 Load R = 50.0
1 cap ser 23.1600
2 ind par 1100.0000
3 cap par 92.1600
4 cap ser 2.0680
5 ind par 1100.0000
6 cap par 113.0200
7 cap ser 2.0680
8 ind par 1100.0000
9 cap par 92.1600
10 cap ser 23.1600
11 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 200.0

Enter # 1 Enter New Value 0

10.00 dB/div.

bottom -60.0

start 9 Frequency MHz 1 Freq./div stop 18 C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====> 10 MHz S21= -100.07 dB S11= -0.00 dB

Above — The GPLA08 assessment of the 10 MHz image frequency. Now 100 dB down!

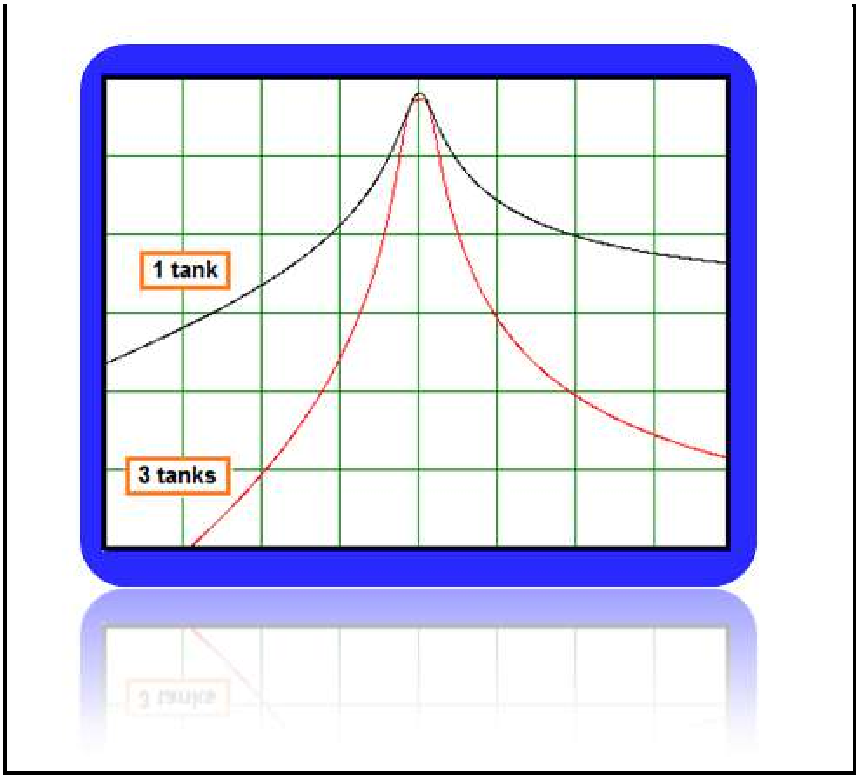

Above — The 1 tank and 3 tank filters superimposed to show the skirt action. The 3 dB bandwidth is the same!

Example 3: A 20 Meter Band-pass Filter for a Builder from Argentina

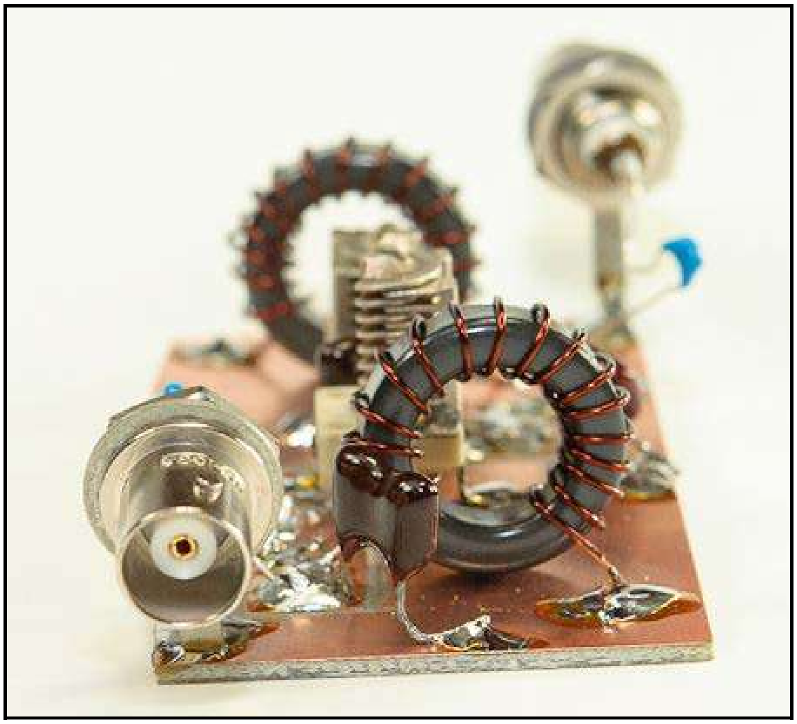

Above — An Argentinean builder emailed that he wanted a band-pass filter optimized for 14.070-14.095 MHz RTTY but also usable for the CW sub-band and lower SSB frequencies. He wanted a center frequency in the RRTY sub-band and I chose 14.079 MHz. Tuning this filter to a center frequency as low as 7.030 MHz for CW should be possible with the variable capacitor value shown, but as mentioned, you really need to do this carefully on your bench. I employed T80-10 toroids and scrunched or expanded the 16 turns of # 22 AWG wire until they measured exactly 1000 nH.

The best return loss will only occur when your filter is perfectly tuned to the test frequency, so tune carefully.

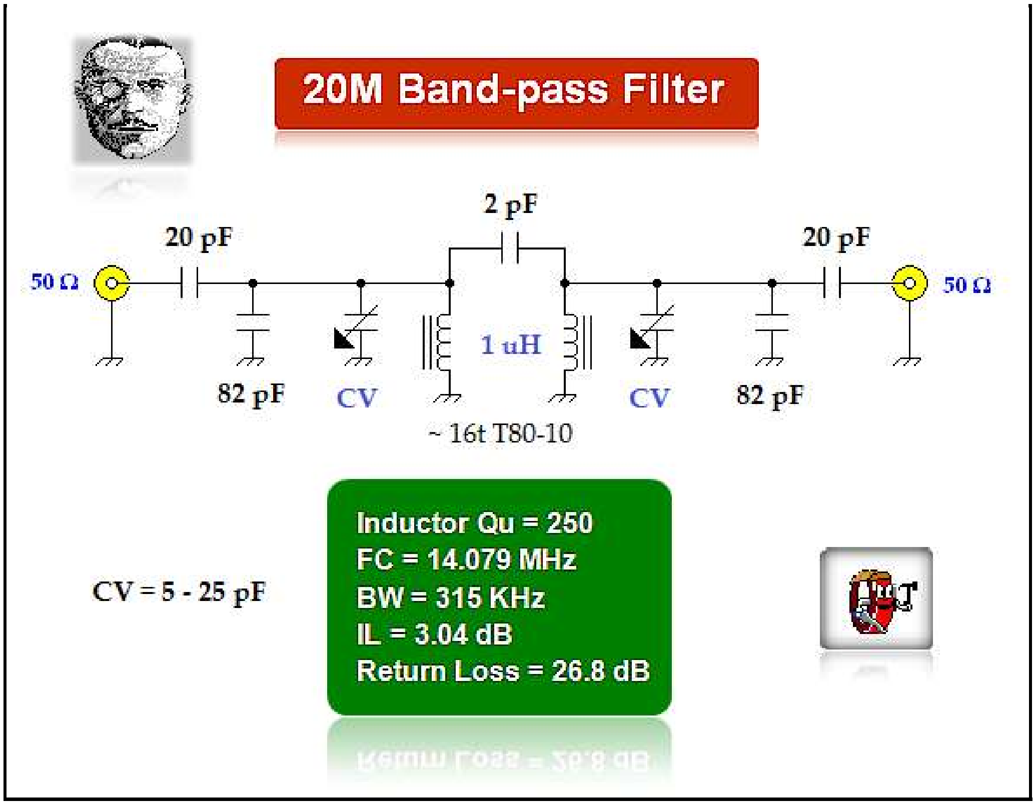

Above — The schematic + bench analysis for the 20 Meter band double tuned band-pass filter. My original design called for 22 pF series end capacitors to get a decent return loss. After building and measuring the circuit, the results were disappointing: insertion loss = 3.7 dB and a return loss = 17 dB. I wanted a better S11 and IL, so I decreased the end capacitors to 20 pF and savored the measured data shown in the schematic.

Simulating this tweaked design in GPLA08 unveiled a lower return loss than the original design simulation with 22 pF end capacitors; exactly opposite to my bench observations.

Bench work reveals the truth — The filter *you* get is dependent on factors such as parts types + tolerances, stray reactance, layout, test gear and any bench errors. For example, I don't know the Qu of my 1 uH inductors, but suspect that the Qu is greater than the 250 specified. Also my intended - 3 dB bandwidth was 350 KHz, yet my filter = 315 KHz; in part, because I lowered the series end capacitors, but also due to other fore mentioned factors.

Many popcorn builders can't easily measure their filter bandwidth. Does it really matter? Probably not, however, the big realization for me is that unless you measure, you won't actually know your data like insertion loss, return loss, or bandwidth — simulations are great, but don't obliterate the need for bench testing as possible.

Consider this; with the SPICE program you can design a circuit with a 2N3904 and run 400 mA of current through it — the transistor won't smoke 1 bit ! Project outcomes depend on understanding and employing best practices, experience and measurement on the bench. Finding best practices proves difficult in a day and time when general scientific literacy, the number of expert mentors and interest in analog electronics are all waning.

Click for another photo of the filter. On my actual filter, I used high Q, air variable trimmer caps that only had a capacitance variation of 15 pF or so. I soldered in fixed capacitors to get close to the capacitance needed to tune each tank. If possible, I think its better use smaller value trimmer caps because they permit finer tuning. The air variable trimmer offers high Q plus you can see when the capacitor is fully meshed (maximum C). This signals that you need to add more fixed capacitance to that tank for peaking.

Conclusion

To repeat; our parts collection dictates our band-pass filter outcomes. Size 50 to 68 #6 material toroids will work fine for most HF frequencies above 3 MHz. Don't stress out too much if your insertion or return loss is a little higher than you wanted; in all likelihood your filter will work fine and you'll be glad you didn't just copy some else's design and rob yourself of the design experience.

I am hopeful, this web page will inspire a few builders to experiment with band-pass filters for their receivers and other applications. My sincere thanks to Wes, W7ZOI for his guidance with filter design.

QRP — Posdata for August 2012 — NE612 Mixer Band-pass Filters

I designed some band-pass filters for NE612 based front-ends with LadBuild and GPLA and show my 75 Meter band filter design figures below. I'm not a fan of employing an NE612 as a receiver mixer since it easily overloads and spews harmonics when mixing strong input signals. The 1K

'RF gain' pot found in many receivers, or the more conventional switchable attenuator pad prove essential when receiving 'booming' signals with a NE602/NE612 mixer in your front-end.

Still, for field-portable tranceivers/receivers, the NE612 mixer keeps the current and radio size down nicely.

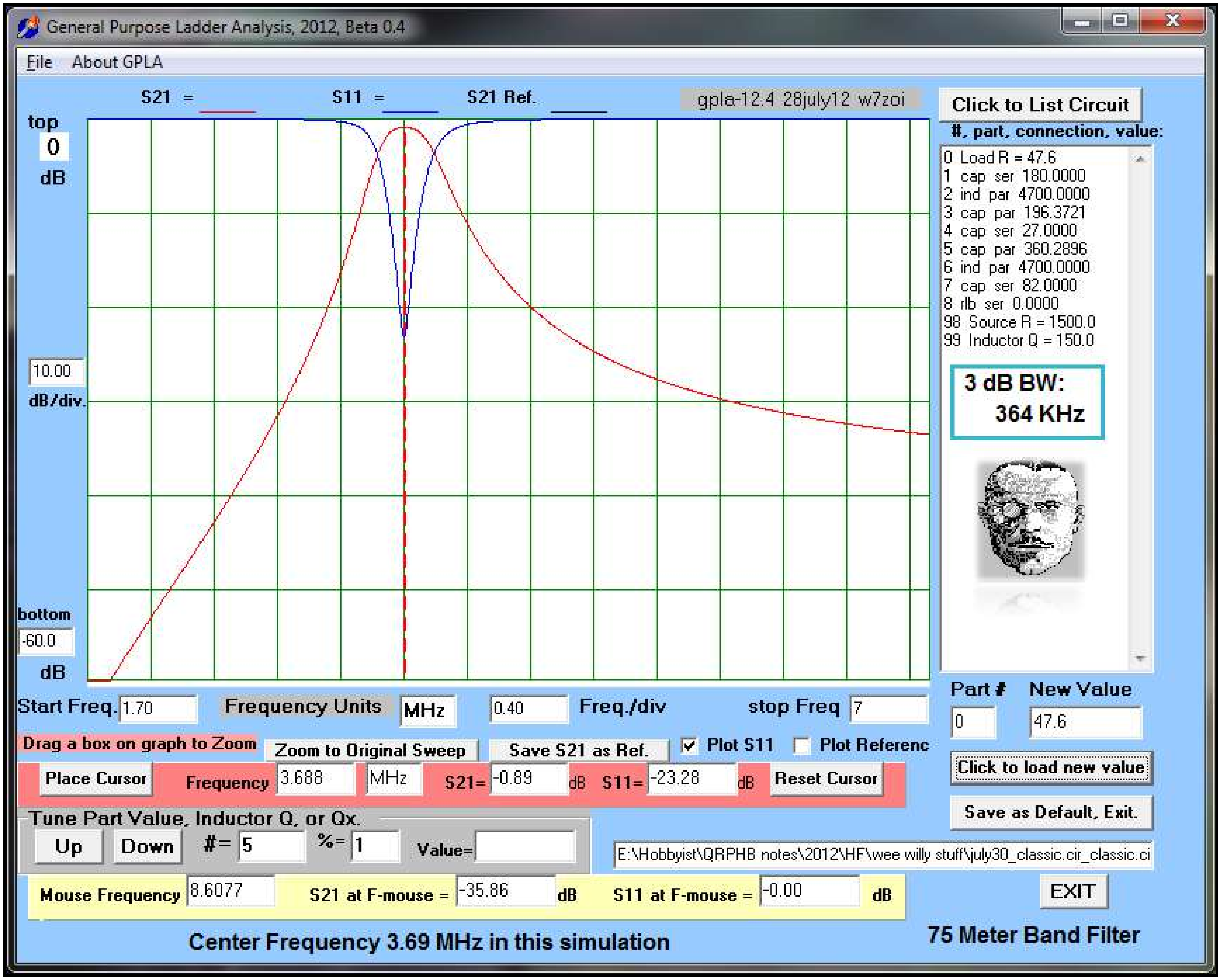
General Purpose Ladder Analysis, 2012, Beta 0.4
File
About GPLA
S21 =
S11 =
S21 Ref.
gpla-12.4 28july12 w7zoi
Click to List Circuit
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Place Cursor
Reset Cursor
Click to load new value
Save as Default, Exit.
Up
Down
EXIT

#, part, connection, value:
0 Load R = 47.6
1 cap ser 180.0000
2 ind par 4700.0000
3 cap par 196.3721
4 cap ser 27.0000
5 cap par 360.2896
6 ind par 4700.0000
7 cap ser 82.0000
8 rlb ser 0.0000
98 Source R = 1500.0
99 Inductor Q = 150.0

3 dB BW: 364 KHz

top 0 dB
10.00 dB/div.
bottom -60.0 dB

Start Freq. 1.70
Frequency Units MHz
0.40 Freq./div
stop Freq 7
Part # 0
New Value 47.6
☑ Plot S11
☐ Plot Referenc
Frequency 3.688 MHz
S21= -0.89 dB
S11= -23.28 dB
Tune Part Value, Inductor Q, or Qx.
#= 5
%= 1
Value=
E:\Hobbyist\QRPHB notes\2012\HF\wee willy stuff\july30_classic.cir_classic.ci
Mouse Frequency 8.6077
S21 at F-mouse = -35.86 dB
S11 at F-mouse = -0.00 dB

Center Frequency 3.69 MHz in this simulation

75 Meter Band Filter

Above — My filter centered at 3.69 MHz. I set the 3 dB bandwidth higher than my usual 200-300 KHz so I could tune a good chunk of the 80-75M band without losing too much signal.

To establish some starting L and C values, I built a classic form filter in DTC08 with a 50 Ω input and output impedance centered at 3.69 MHz. After some tweaking, I settled on L = 4700 nH, Ce = 148 pF, coupling capacitor Cm = 27 pF and about 220 pF (Ct) to resonate each tank.

Next I changed the termination R's to 1500 Ω in DTC08 to simulate the right half of the filter matched into the 1K5 input impedance of an NE612. This gave me some Ct and Ce values to start with. I started Ladbuild 08 and built up a schematic.

In a seperate experiment, I determined that the resonator Qu of a 3.7 MHz L-C tank with an L of 4700 nH wound on a T50-2 core, was ~150.

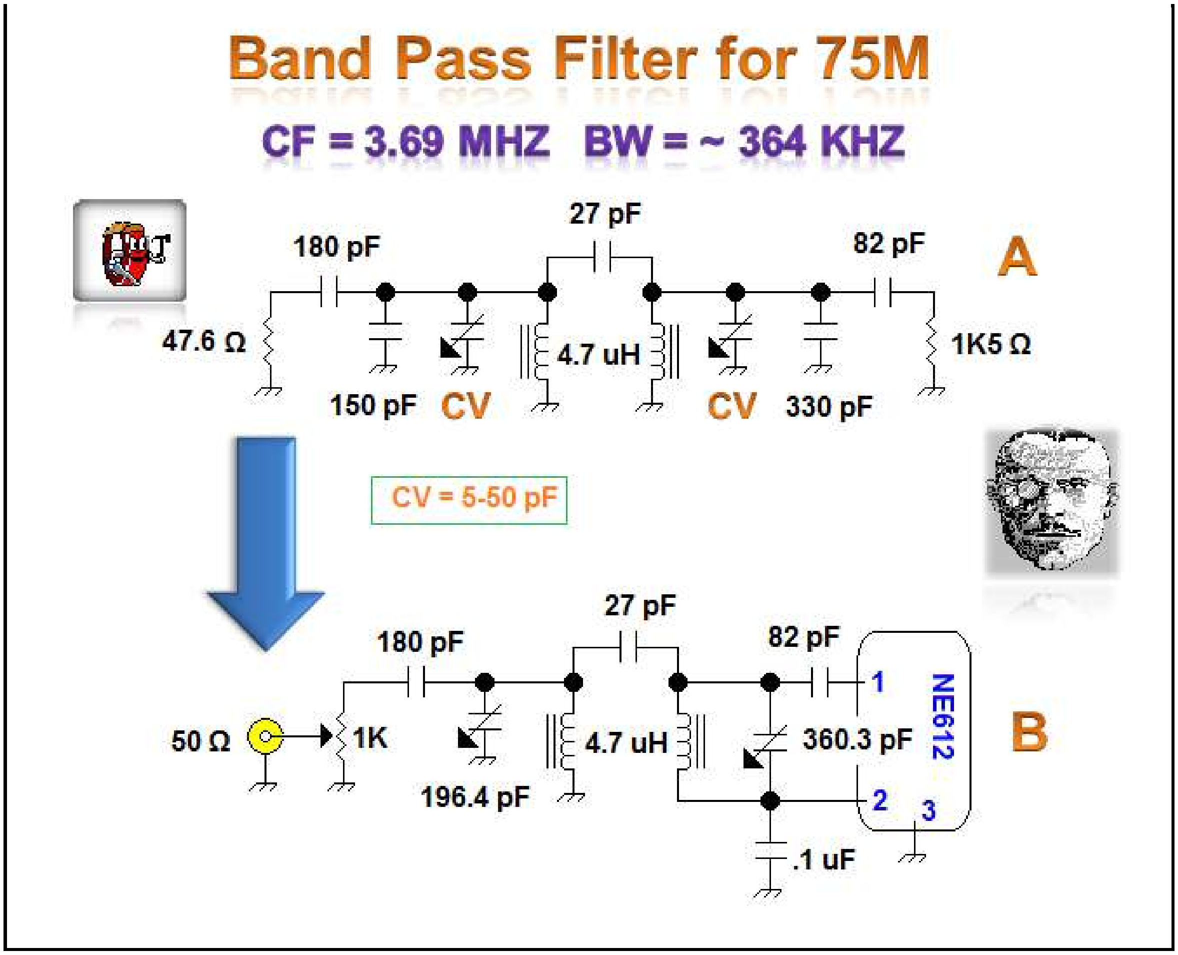

Above — The completed schematic. My filter exhibits an attenuation of ~70 dB at the top of the AM radio broadcast band (1500 KHz) providing I shield it in an RF-tight box. I took the 47.6 Ω input Z from the 1K pot and my 50 Ω antenna in parallel. An input Z of 50 Ω would work just as well in simulation and on-bench.

In Part A, I show a possible way to resonate each tank with 1 fixed C and a trimmer capacitor. In Part B, I omitted this detail and just show the calculated C needed for resonance as a variable capacitance.

To make this filter with GPLA, I tweaked the capacitor values to nearest standard value parts and tuned the filter with the GPLA Tune Part Value controls while looking at the waveform and my 3 dB bandwidth. I love tweaking values in GPLA and over the years have designed several hundred RF filters for readers.
The rubber hits the road on the bench however! You can get an E.E. degree without melting solder in this day and time — but only bench measurements tell the truth.

Please tune each tank carefully like I mentioned earlier... For example, say a tank needs 180 pF for resonance, *but you don't know this*. You solder in a 100 pF cap and a 5-50 pF *enclosed* ceramic trimmer capacitor into the L-C tank. While watching the 'scope this tank will "peak" since the tank will exhibit its highest peak-peak voltage when the trimmer cap is set to 50 pF and fall off as you decrease the C of the trimmer. You might think you peaked the tank, however you' re actually under by 30 pF!

While leaving the trimmer set to maximum C (peak-peak voltage) in this theoretical example, if you tack solder in another 10 pF cap your 'scope will show an even greater pk-pk voltage. , If you remove this 10 pF cap and then place in a 27 pF cap, the pk-pk voltage will go even higher since you're almost at the target 180 pF. If you removed the 27 pF cap and tack soldered in a 47 pF cap, the pk-pk voltage in the 'scope will go down since your now at 197 pF. Thus you know that resonance is somewhere between 177 and 197 pF.

Of course you could decrease the trimmer cap C and stil use the 47 pF cap, however, my description isn't a prescription to follow, just some things to think about. Sometimes I remove a trimmer cap and measure it to ensure the cap is not set to maximum C; that would tell me I need to add more fixed capacitor(s) to the tank. Air variable trimmer caps give visual indication since maximum C occurs with maximum mesh. Unfortunately they are rare and expensive.

On my bench I keep a pair of small 12 to 400 pF air variable caps and temporarily solder them into my tanks. After peaking, I remove and measure them — then I have a good idea of what capacitance is needed to resonate the tank at my test frequency.

It's all an experiment.

Click for another low loss, well matched example: CF = 5.17 MHz, 3 dB BW = 196 KHz. 50 ohm version. NE612 final version.

QRP — Posdata for August 2012 — More NE612 Receive Mixer Band-pass Filter Experiments

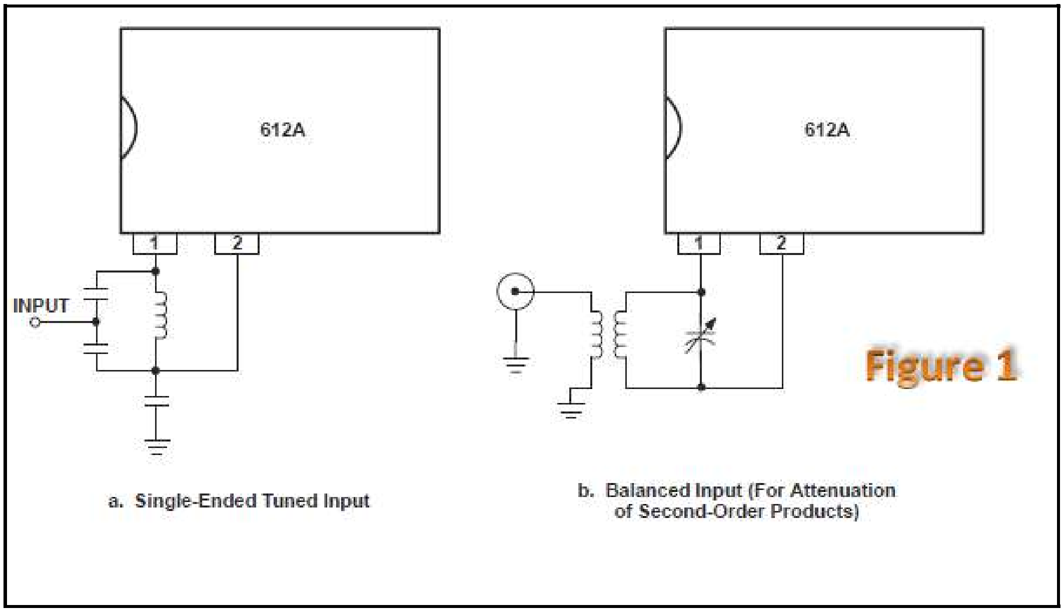
612A
612A
1
2
1
2
INPUT
Figure 1
a. Single-Ended Tuned Input
b. Balanced Input (For Attenuation of Second-Order Products)

Above — NE612 input circuits. The NE612 datasheet specifies a 3 pF input capacitance + a 1K5 input resistance.

If you look around the Web, many builders just run a single tank for band-pass filtering. While okay for novelty-grade rigs, the poor filter stopband may unleash some ugly problems in the mixer and on down the receiver chain.

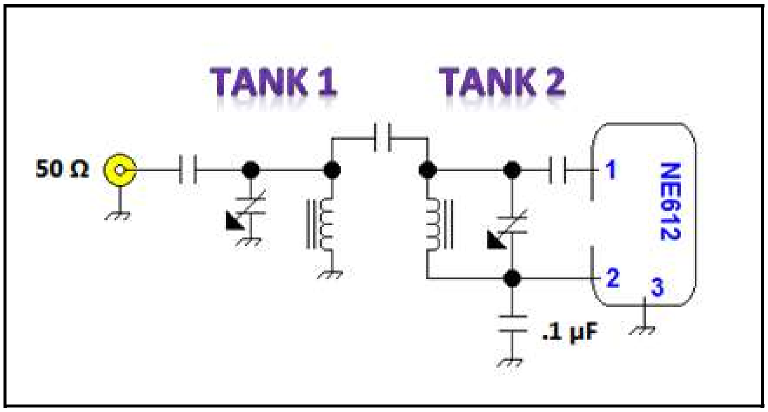
TANK 1
TANK 2
50 Ω
NE612
1
2
3
.1 µF

Above — A double tuned circuit with the L-C tanks named 1 and 2 and a series capacitor to match Tank 2 to the NE612.

Most builders match Tank 1 into its 50 Ω source with a capacitor divider, or a matching transformer. For Tank 2, some enthusiasts just connect the Tank 2 coil directly to pin 1 as shown in Figure 1a. Without the matching series capacitor, unfortunate side effects may arise...

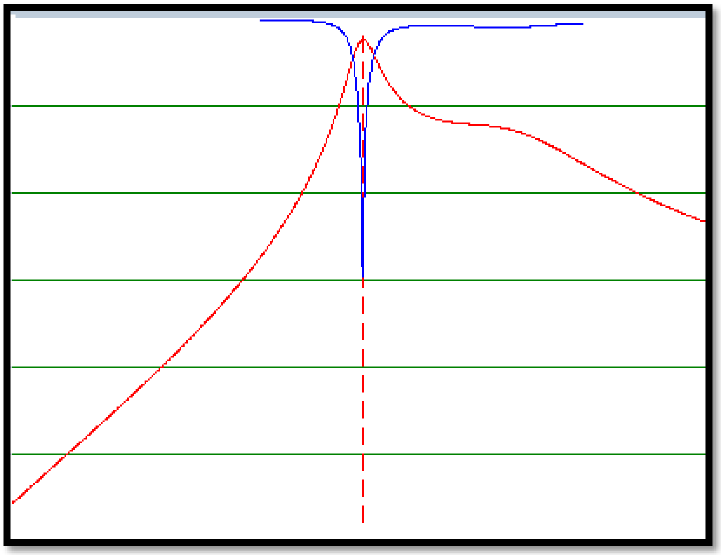

Above — The low-pass skirt of a double-tuned filter may attenuate higher frequencies poorly when no series capacitor (or other network) matches Tank 2 to the NE612 input. I perfectly matched Tank 1 to its 50 Ω source just connected Tank 2 to a 1500 Ω resistive load in this simulation. I wonder how bad things get in the real world when a complex impedance is involved?

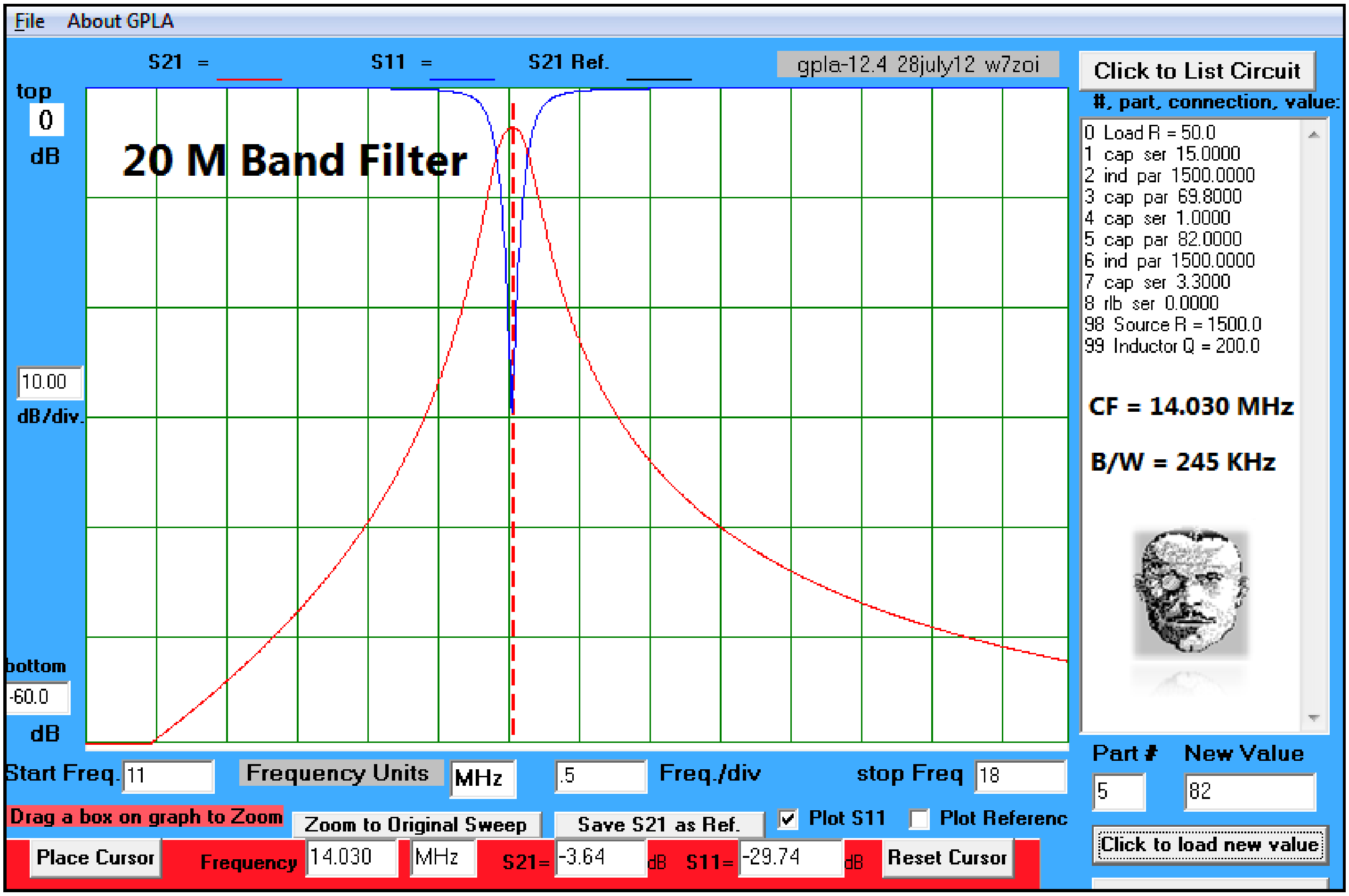
File
About GPLA
S21 =
S11 =
S21 Ref.
gpla-12.4 28july12 w7zoi
top
0
dB
20 M Band Filter
dB/div.
bottom
dB
Frequency Units
Freq./div
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Plot S11
Plot Referenc
Place Cursor
Reset Cursor
Click to List Circuit
#, part, connection, value:
Part #
New Value
Click to load new value

0 Load R = 50.0
1 cap ser 15.0000
2 ind par 1500.0000
3 cap par 69.8000
4 cap ser 1.0000
5 cap par 82.0000
6 ind par 1500.0000
7 cap ser 3.3000
8 rlb ser 0.0000
98 Source R = 1500.0
99 Inductor Q = 200.0

CF = 14.030 MHz

B/W = 245 KHz

dB/div. 10.00
bottom -60.0 dB
Start Freq. 11
Frequency Units MHz
.5 Freq./div
stop Freq 18
☑ Plot S11
☐ Plot Referenc
Frequency 14.030 MHz
S21= -3.64 dB
S11= -29.74 dB
Part # 5
New Value 82

Above — I designed a filter for a 20 Meter band CW receiver centered at 14.030 MHz with DTC and GPLA. The design 3 dB bandwidth = 245 KHz. NE612 filter design was discussed in QRP — Posdata 1.

I then breadboarded the filter with T68-6 inductors, but common lower Qu ceramic trimmer capacitors.

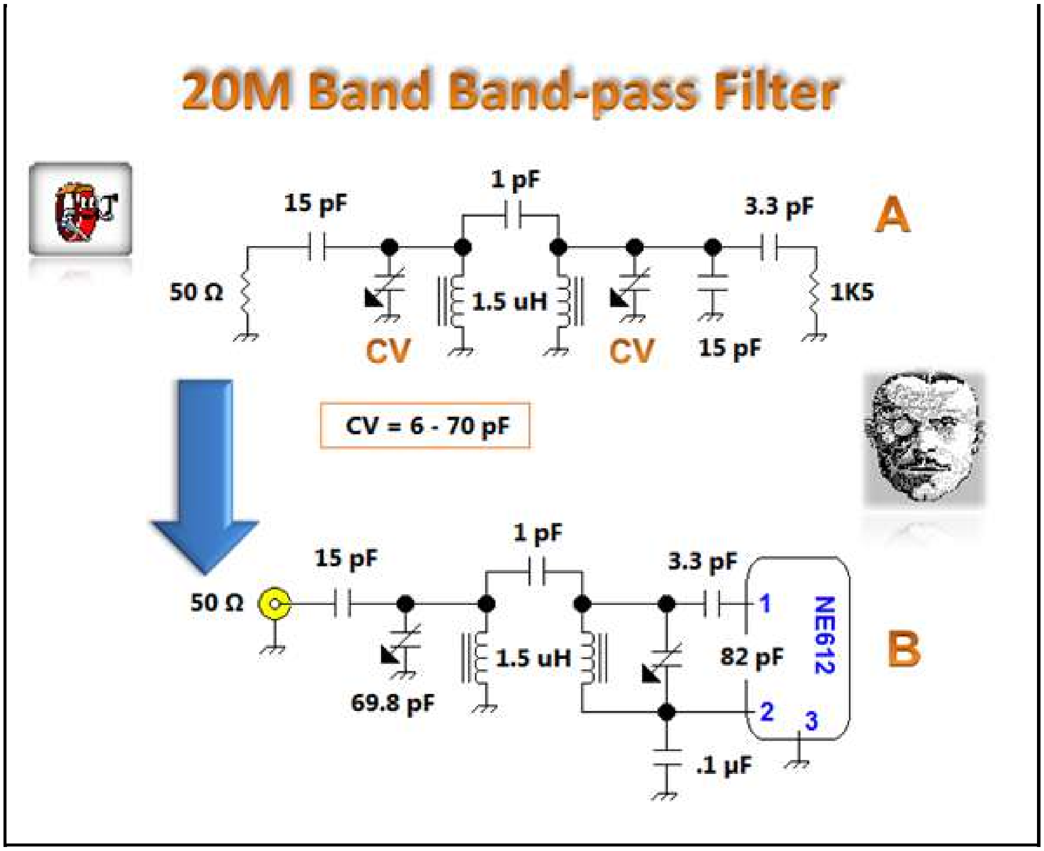
20M Band Band-pass Filter
15 pF
1 pF
3.3 pF
A
50 Ω
1.5 uH
1K5
CV
CV
15 pF
CV = 6 - 70 pF
1 pF
15 pF
3.3 pF
50 Ω
1
NE612
1.5 uH
82 pF
B
69.8 pF
2
3
.1 µF

Above — The DTC/GPLA filter design with Tank 2 evolved to provide single-ended input for the NE612.

I wanted to test 2 questions:

1. Does the 0.1 µF coupling cap connected to the cold end of Tank 2 and Pin 2 change the bandwidth or filter skirt shape?
2. Will the 3.3 pF cap really match the NE612?

I expect that worldwide, the NE612 input impedance may vary slightly from part to part; different breadboards will exhibit different reactances and that although the datasheet specifies 1500 Ω , we may be dealing with a complex impedance that varies with the aforementioned factors plus perhaps, input frequency. I simply want a good filter with clean skirts.

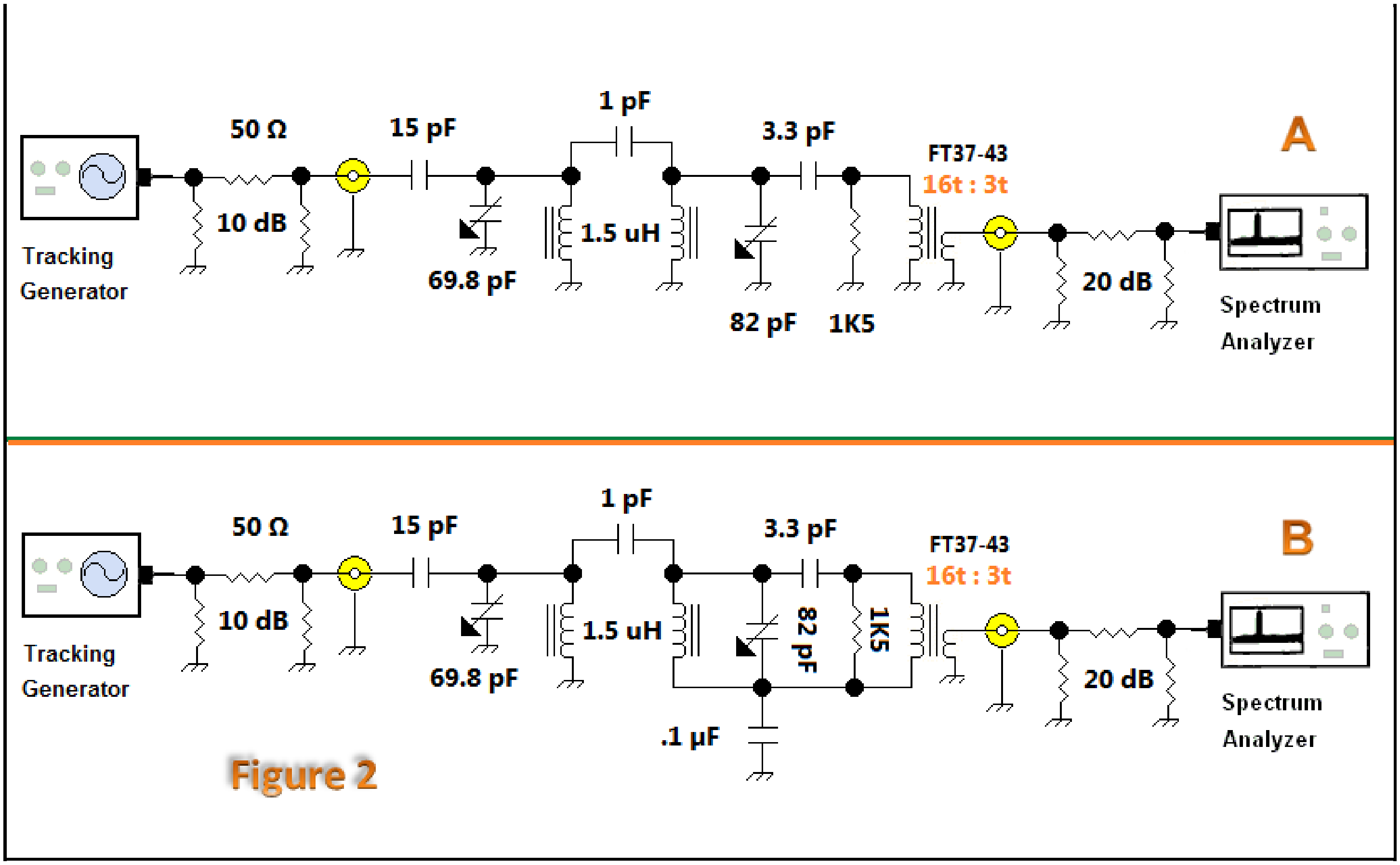
A
Tracking Generator
50 Ω
10 dB
15 pF
69.8 pF
1 pF
1.5 uH
3.3 pF
82 pF
1K5
FT37-43
16t : 3t
20 dB
Spectrum Analyzer
B
Tracking Generator
50 Ω
10 dB
15 pF
69.8 pF
1 pF
1.5 uH
3.3 pF
82 pF
1K5
.1 µF
FT37-43
16t : 3t
20 dB
Spectrum Analyzer

Figure 2

Above — Experiment #1. I compared spectrum analysis of circuit A with B. I built and measured A and then cut away some copper to isolate the copper board grounding the Tank B parts. This "island" was AC coupled to the rest of the ground plane with a short leaded 0.1 µF ceramic capacitor.

I saw no significant difference between Circuit A and B — it appears the 0.1 µF capacitor does not affect the filter parameters to any extent.

To simulate a NE612, I soldered a 1K5 Ω 5% resistor across Tank 2. Tank 2 was transformer coupled to the 50 Ω Z required for spectrum analysis. A 20 dB pad ensured a strong return loss and a safe input amplitude for the SA.

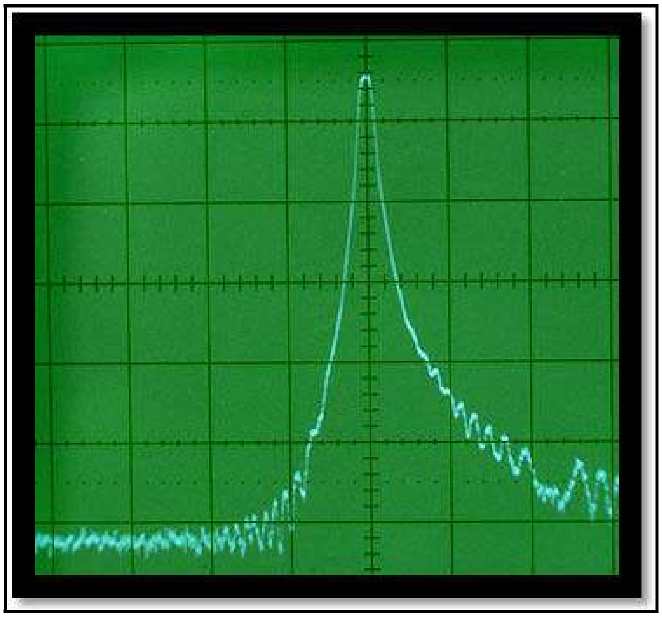

Above — Spectrum analysis of Figure 2A or B. I saw flat topping of the waveform – almost double humping with a higher than wanted 3 dB bandwidth; this bothered me. Fixing this problem was Experiment #2.

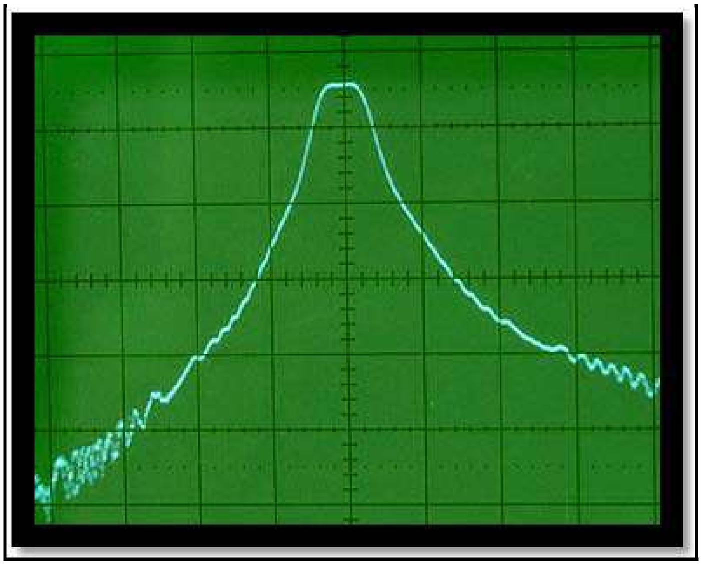

Above — A zoom of the poor coupling of Figure 2A or B. From my experience building filters, I suspected a termination resistance mismatch in Tank 2 ; exactly what I was trying to avoid!

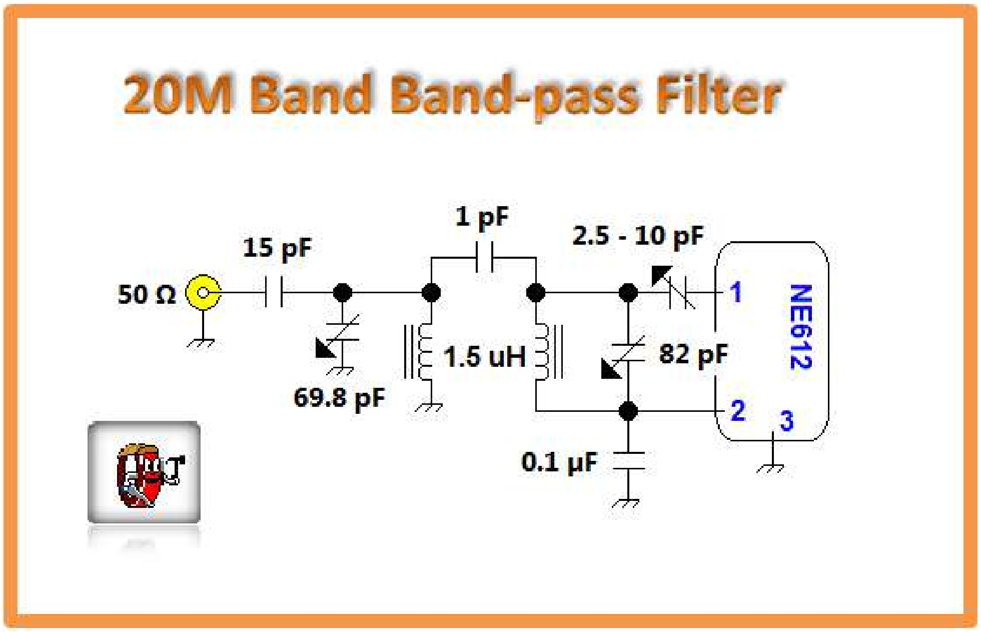
20M Band Band-pass Filter
50 Ω
15 pF
69.8 pF
1 pF
1.5 uH
2.5 - 10 pF
82 pF
0.1 µF
NE612
1
2
3

Above — I swapped a trimmer capacitor for the series 3.3 pF fixed cap in Tank 2. Then I reconnected the circuit into my test set up.

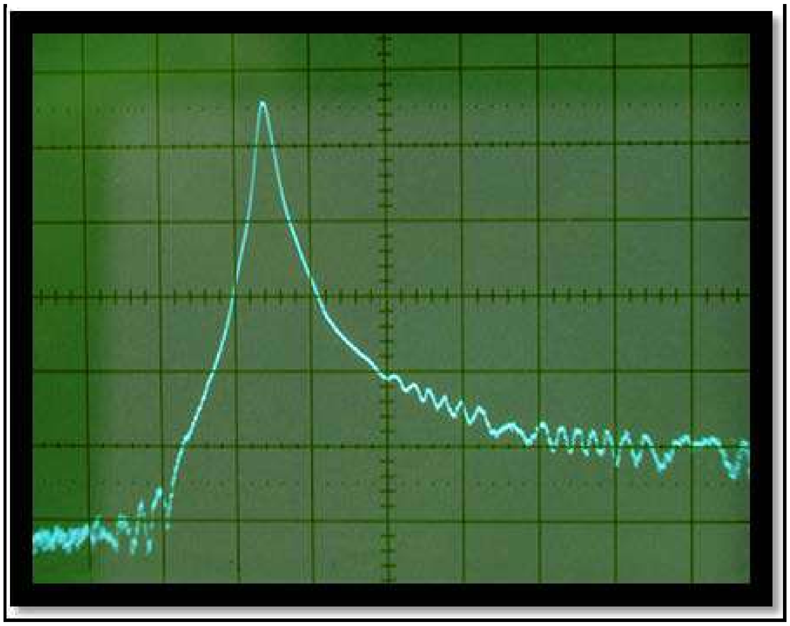

Above — A photo of my TG + SA measurement after tweaking the newly added trimmer capacitor and peaking each tank. I shifted the SA screen center over so the tracing could be seen without all the hash marks in the center. Wonderful.

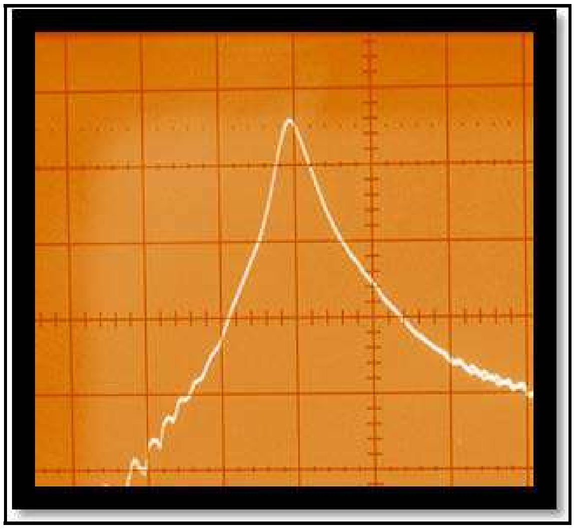

Above — A zoom of the now matched Tank 2. Due to the bench-altered Tank 2 match, my low-Q variable capacitors and other factors, the 3 dB bandwidth now is just over 300 KHz.

I'm not sure if these experiments reflect what actually happens with a NE612 input band-pass filter, however, I plan to match my second tank with a series trimmer capacitor in future NE602/NE612 work. I also want to explore balanced input.

RF — Test and Measurement

RF Workbench Page 4

The 4th installment of a QRP/SWL HomeBuilder series exploring *basic* RF measurement

Part 4 describes a method to calculate reverse isolation in the **50 Ω environment** after converting measured peak-to-peak AC voltages to dBm. I tested 2 common amps at ~7 MHz to show the concepts and calculations.

In this series, I gratefully borrow from the work of Wes, W7ZOI per correspondence, direct contributions and from EMRFD.

Tools Needed

1. 50 Ω terminated scope (or a spectrum analyzer) and a 50 Ω signal generator
2. 50 Ω RF cables with RF connectors (such as short cables with female BNC connectors)
3. 6 dB 50 Ω attenuator pad; plus an adjustable attenuator if you use a fixed output signal generator.
4. BNC through-response connector(s)

Procedure

1. Measure the amplifier forward power gain. This is S21.
2. Measure amplifier reverse power gain. This is S12. Like S11, express S12 as a negative value.
3. Reverse isolation using dB values = S21 - S12.

Step 1 : Measure the forward power gain

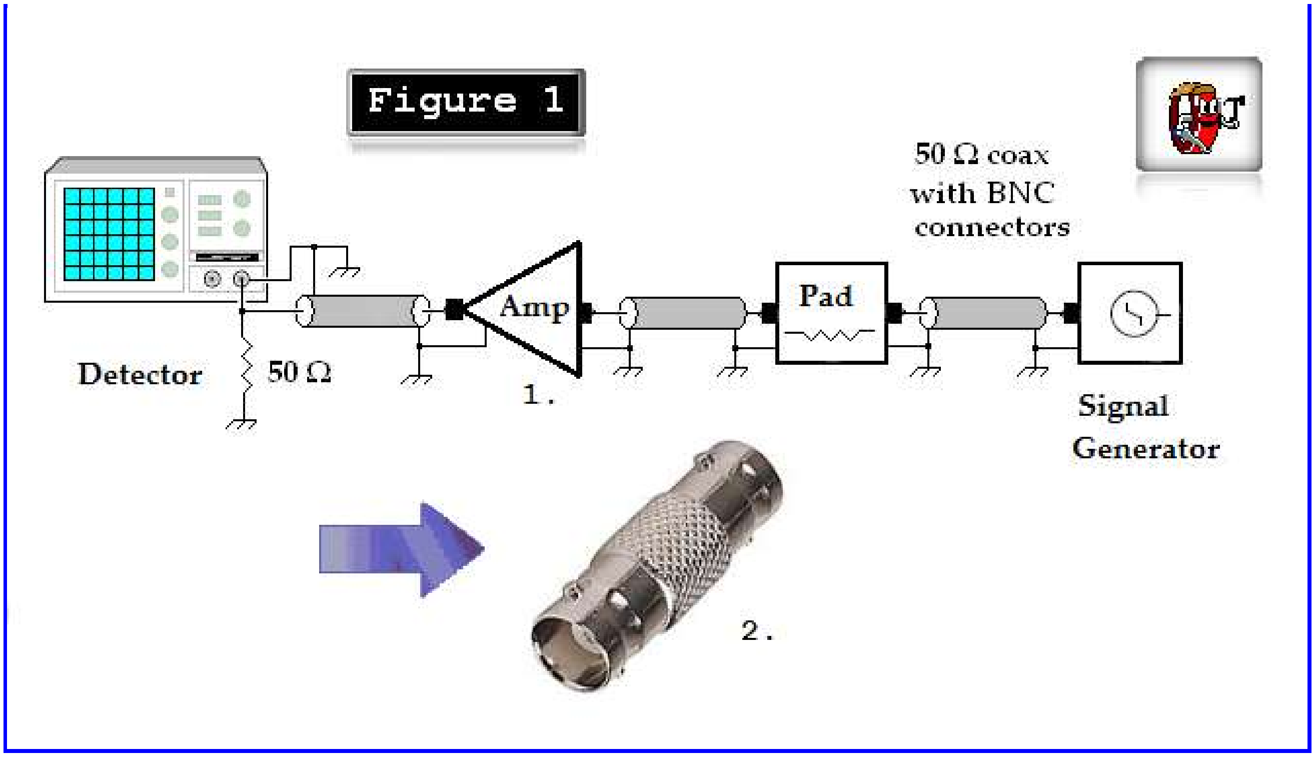

Figure 1 shows how to measure forward gain in a 50 Ω environment — **1**. Convert the measured AC voltage to dBm, **2**. Disconnect the amplifier, insert a through-response connector and convert this measured AC voltage to dBm. The difference between the 2 values = S21. Applet H will do these calculations from the peak-peak voltages. *The attenuation pad following the signal generator in Figures 1 and 2 signify that the signal generators have a 50 ohm output impedance and is optional.*

Choose a signal generator level that ensures the output of your amplifier is linear while providing a good signal to noise ratio for measurement. With an oscilloscope, I generally test amplifiers with an input power of between 0 and -11 dBm; although choose whatever level that works for you consistent with linear amplification.

After measuring the forward gain of your amp, a good way to test for linearity is to add a fixed 6 dB pad between your signal generator and your amplifier to drop the applied signal by half. *[A 6dB pad drops the peak-peak voltage by 1/2 . A 3 dB pad drops the power in dBm by 1/2]*. The power gain should be equal or nearly equal to the measured dB value obtained before you added the 6 dB pad. If they vary significantly, you are likely driving the amp too hard and causing some non-linear output products. S21 = complex *linear* gain.

Step 2 : Measure the reverse power gain

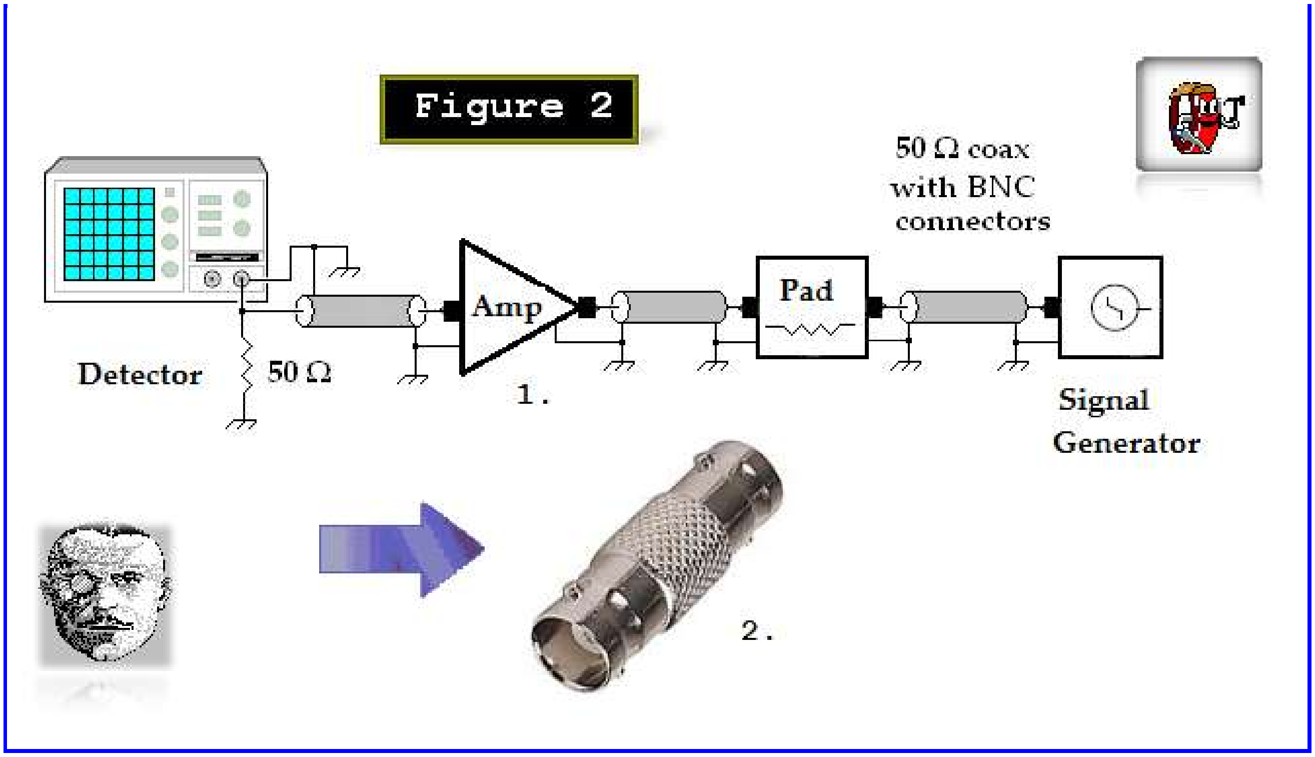

Figure 2 shows how to measure reverse gain in a 50 Ω environment — **1**. Convert the measured AC voltage to dBm, **2**. Disconnect the amplifier, insert a through-response connector and convert this measured AC voltage to dBm. The difference = S12. Ensure that you employ the same drive level used to measure the forward gain.

Measuring reverse gain may be tough. When the amplifier under test requires a low drive level and/or has strong reverse isolation, you may not have enough signal to accurately measure with your oscilloscope. The tool of choice for low reverse voltage measurment is a spectrum analyzer (SA) — a narrow band SA may be required to distinguish the weak signal from random noise with low signal voltages.

Summary

A **practical** bench work flow goes something like:

1. Measure the through-response peak-peak voltage.
2. Measure and record the peak-peak AC voltage while driving the amplifier input port (forward gain set-up).
3. Reverse the amp so you're driving the output port and then measure the peak-peak AC voltage (reverse gain set-up).
4. Calculate S21 and S12.
5. Calculate reverse isolation.

Example 1: Feedback amplifier

In the first example, I measure the reverse isolation of a Beaverton Special feedback amp. The schematic is shown below. Tested with a 7.039 MHz signal generator possessing a return loss of 30 dB.

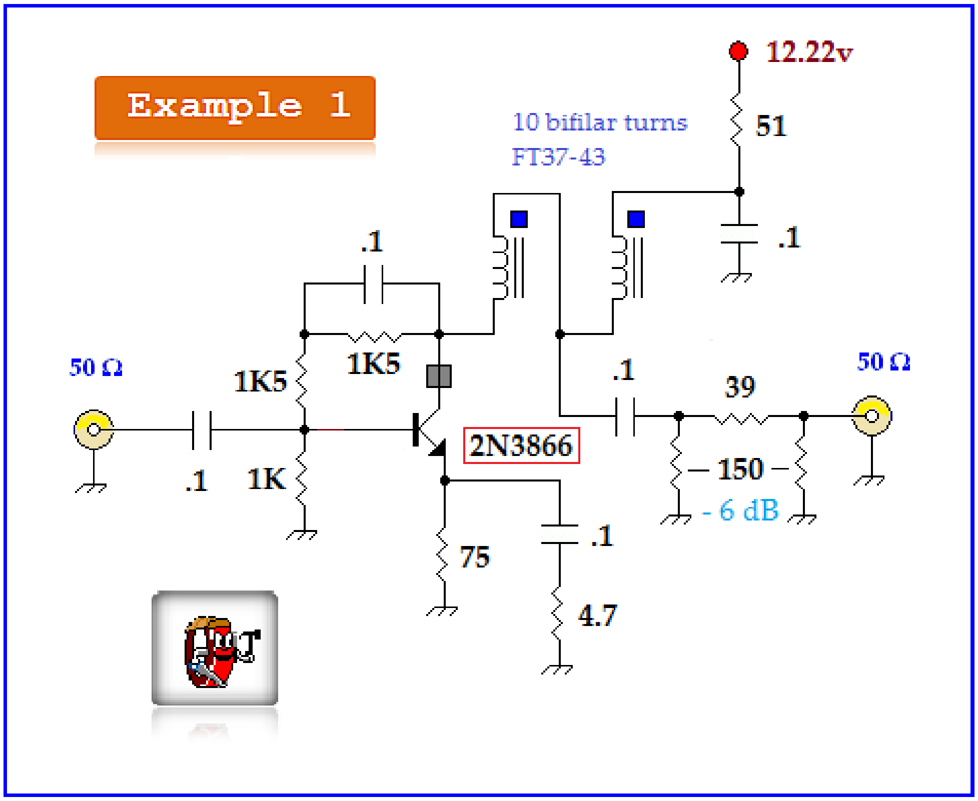

Example 1
12.22v
10 bifilar turns
FT37-43
51
.1
.1
1K5
1K5
50 Ω
50 Ω
.1
39
.1
1K
2N3866
150
- 6 dB
75
.1
4.7

50 ohm Voltage Measurements:

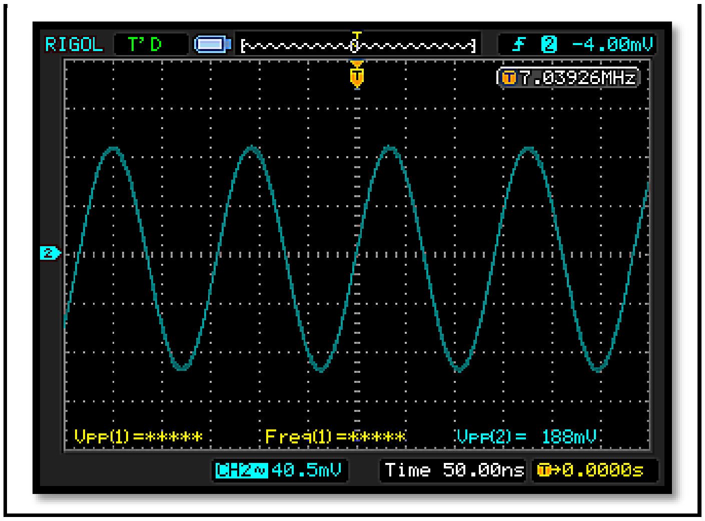
RIGOL
T'D
-4.00mV
7.03926MHz
Vpp(1)=*****
Freq(1)=*****
Vpp(2)= 188mV
CH2 40.5mV
Time 50.00ns
0.0000s

Above — AC voltage with a through-response connector in-situ: 188 mV pk-pk

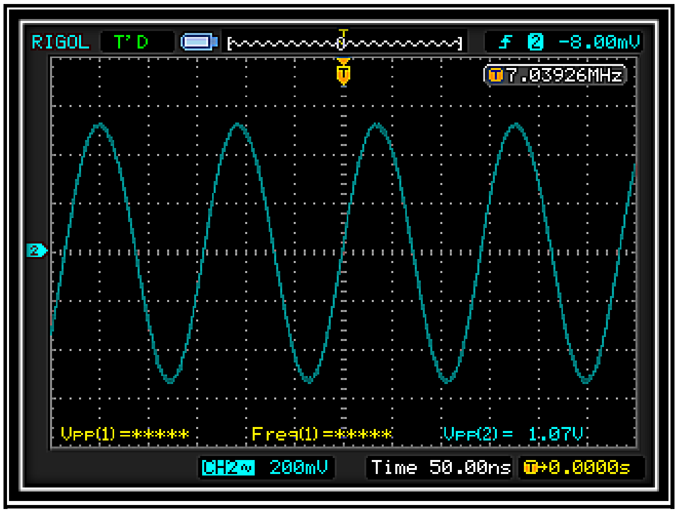
RIGOL
T'D
-8.00mV
7.03926MHz
Vpp(1)=*****
Freq(1)=*****
Vpp(2)= 1.07V
CH2 200mV
Time 50.00ns
0.0000s

Above — AC voltage with forward gain set-up: 1.07 V pk-pk

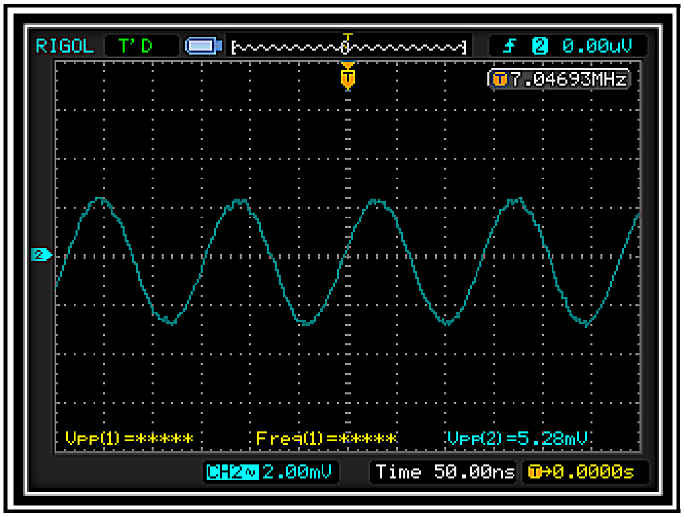
RIGOL
T'D
0.00uV
7.04693MHz
Vpp(1)=*****
Freq(1)=*****
Vpp(2)=5.28mV
CH2 2.00mV
Time 50.00ns
0.0000s

Above — AC voltage with reverse gain set-up: 5.28 mV pk-pk.

S21 = 15.1 dB. S12 = -31 dB. I tested my driven amp's linearity by adding a scrap, standard value 6 dB pad in between the signal generator and the amplifier — the S21 was 15.2 dB with the pad and 15.1 dB without the added 6 dB pad. It's linear.

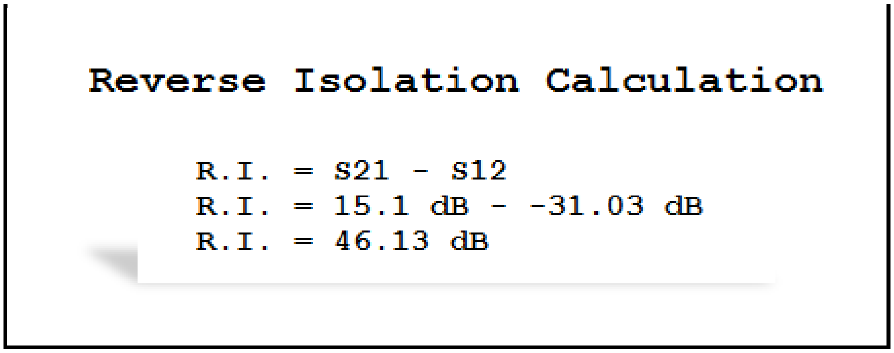

Reverse Isolation Calculation

R.I. = S21 - S12
R.I. = 15.1 dB - -31.03 dB
R.I. = 46.13 dB

Example 2: Common Base Amplifier

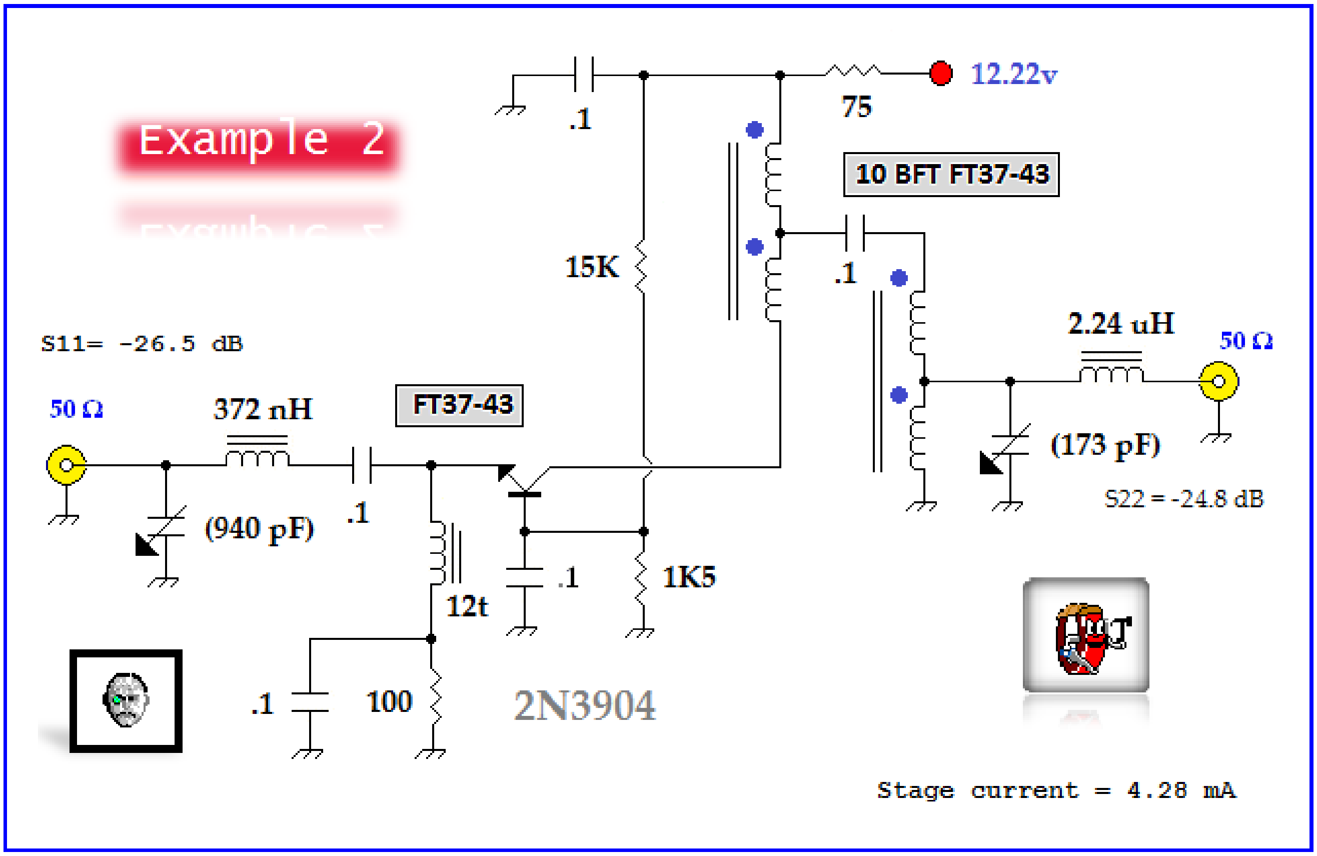
Example 2
12.22v
.1
75
10 BFT FT37-43
15K
.1
2.24 uH
50 Ω
S11= -26.5 dB
50 Ω
372 nH
FT37-43
(173 pF)
(940 pF)
.1
S22 = -24.8 dB
.1
1K5
12t
.1
100
2N3904
Stage current = 4.28 mA

Above — A common base amp employing L- networks for a strong return loss in and out at 7.039 MHz. S11 and S22 = the negative of return loss.

Through connector voltage = 192 mV peak-peak.

S21 = (1.88 volts peak-peak AC voltage) = 19.82 dB.

S12 = (1.68 mV peak-peak AC voltage) = -41.16 dB.

Reverse isolation at 7.039 MHz = (S21 - S12) = 61.42 dB

I confirmed the linearity of my S21 using the aforementioned 6 dB attenuator pad — I couldn't increase my signal generator output level above the indicated -10.35 dBm, since gain compression emerged in the common base amplifier (reduced AC voltage was measured).

My reverse AC voltage was under 2 mV; about the threshold where my oscilloscope waveform becomes rather ugly and sits in the noise. Clearly, the limitations of measuring reverse isolation with an oscilloscope must be factored. Still, you got to love 60 dB + of reverse isolation in a popcorn circuit.

1 local professional EE told me if you can measure it with a 'scope, you don't have spectacular reverse isolation - no doubt, a spectrum analyzer

pumps up the measurement quality in circuits with high reverse isolation, but more amateur designers have 'scopes than spectrum analyzers, so just do your best.

Involving Scattering Parameters as possible on your workbench can only lead to circuit improvement. You can't better your outcomes if you can't or don't measure them — applying and more importantly, understanding test equipment is just 1 component of our hobby. A hobby unto itself; test equipment activity complements amateur radio design. I have met test equipment focused builders who make radio gear just as an excuse to apply their test gear!

Improved bench practices are the corollary of striving to learn more about measurement techniques and increasing our collection of measurement devices.

QRP — Posdata 1: Hycas Amplifier

I love the hybrid cascode (hycas) as a general purpose RF amplifier. What's not to love about a using a common source FET followed by a common base bipolar amp? I attempted to measure the reverse isolation of a version just using an oscilloscope.

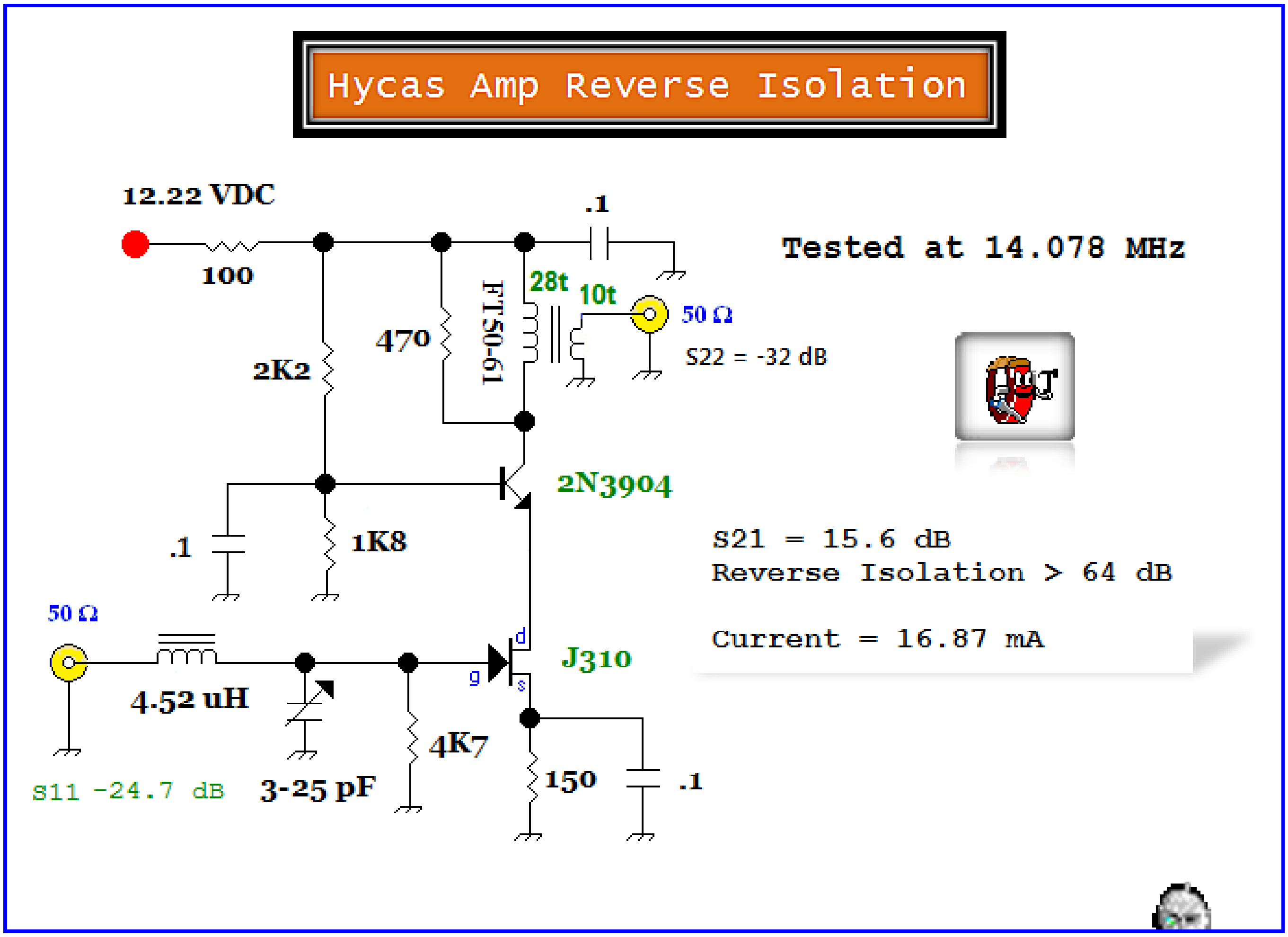

Above — A hycas amp set up for high return loss on both ends at 14.078 MHz. Too much current may cause gain compression and harmonic distortion, so please test your hycas amps for both. I tested using a signal generator with a 30 dB return loss driving a 50 Ω terminated oscilloscope. Since the hycas amp contain a high impedance input JFET and a common base amp, the reverse isolation should be reasonably high, or at least as good as a common base amplifier.

My testing failed — The reverse isolation was too high to measure with an oscilloscope. Using proper bench techniques (linear amplification + honest scope reporting), I determined the highest reverse isolation I could measure = 64 dB. In fact, injecting a whopping signal of 1.08 volts peak-peak into the hycas output port only gave an S12 of 1.84 mV — whoa!

The problem is such a strong signal (1.08 volts peak-peak) at the input port results in severe limiting and distortion; so valid reverse isolation measurement isn't possible. Even a 350 mV peak-peak signal may give some gain compression during S21 measurement depending on your matching. Thus, I can only accurately say that the reverse isolation of my hycas amp is greater than 64 dB.

Strong reverse isolation is 1 reason I favor hycas amps as VFO buffers. They make pretty good I.F. amps also.

The hycas IF amp system by Wes, W7ZOI and Jeff, WA7MLH offers amazing performance and features an excellent JFET bias scheme. I built 1 in 2008 — amazing design.

QRP — Posdata 2: Doesn't S12 = Reverse Isolation?

Many web sites, books and people report that reverse isolation = S12, yet above, I depart from this argument. In truth, I think reverse isolation equals S12, but reverse isolation may also equal S21 - S12.

I'll let you decide what to do, but explain why I enjoy the latter.

S12 is a negative value.

I prefer to turn that negative value into a positive 1 — RF Workbench 4 concerns measuring and applying amplifiers with the goal of high reverse isolation and not just measuring S12.

The main purpose to quantify amplifer reverse isolation is to strive to *improve reverse isolation* and an amplifier is but 1 component in our 50 Ω block. I believe in a creative, systems approach; open minded and positive (pun intended).

The whole RF Workbench series attempts to present 50 Ω bench measurements in a vibrant way devoid of excessive + boring engineer-style content that could blank the eyes of the budding Hams/SWLs designers that visit my site. I imagine this web site bores more advanced RF designers *to tears*.

Our goal is to obtain high reverse isolation while applying a 50 Ω systems approach.

Break away from strict + "stodgy" math-driven methods to fuel creative thought and experimentation.

I posit that an appropriate figure of merit in a well designed isolation amplifier is the difference between S21 and S12; and therefore, the term reverse isolation can be more than just S12.

Ham/SWL component-level experimentation by commoners like me is ~~slowly~~ dying and being replaced by a new generation of skilled, code writing experimenters. Although, some builders just copy other people's code and then apply it to kitted hardware. Reverse isolation impacts both our analog and digital designs.

Please consider this example:

I'm switching a level 7 diode ring mixer with 12 MHz and per normal, create lots of internal harmonic energy. With my spectrum analyzer connected to the mixer RF port I measure my 12 MHz LO signal at 50 dB below a +7 dBm signal, or -43 dBm. That's a 50 over S9 signal — very high amplitude in context!

I require strong LO isolation in my circuit and thus stick in a 50 Ω input/output amplifier. I measure this amp: S21 = 15 dB and S12 = -31 dB.

So, the signal at the amplifier input is -43 dBm plus -31 dB = -74 dBm. But, alas, -74 dBm isn't good enough me — I want to use that amplifier to elicit greater isolation. However, I don't want any gain in my system, so I insert a 15 dB attenuator pad after the amplifier. For this pad, both S12 and S21 = -15 dB. For my amp, the net cascade is S21 = 0 and S12 = -46 dB.
Since S21 = 0, the block has 0 impact on the signal amplitude applied to the mixer, but the signal at the input of my isolation amplifier is -43 dBm plus - 46 dB, or -89 dBm. This isolation I like — it also illustrates a systems approach that gets you thinking about measurement in your own 50 Ω blocks.

The figure of merit for making a good isolation amplifier is now the difference between S21 and S12. If you want, go ahead and just use S12 for reverse isolation, but you'll probably measure S21 plus S12 anyway and that's what this web page is about! Onward.

QRP — Posdata 3

Comments From the Workbench

I'm no amateur electonics expert — I'd like to be one, but this is a tough field; RF and AF *design* is quite scientific, under-resourced and a bit overwhelming. How do we experimenters advance and stay motivated? Reading works by professionals like Chris Trask, N7ZWY, Bob Larkin, W7PUA, Doug Self, Rod Elliott and others may highlight our lack of knowledge and scientific methodology — a realization which can distress and demotivate us lay-designers. To a degree, this is irrational thinking; personal growth is always about hard work, problem solving and overcoming barriers.

Unlike the white belted Karate student, who studies and practices under the guidance of a master to attain black belt skill level, most amateur designers, excluding electrical engineering students, can't access good teachers. As a lay-person, with few face-to-face mentors (nobody in

Canada), I try to learn by experimenting and incorporating whatever knowledge, advice or schematics I can find. Fortunately, some Electrical Engineers give me advice by email and in turn I'm able to share this information via experiments on QRP / SWL HomeBuilder.

Our dusty, analog hobby fades palpably — the number of analog electronics gurus dwindles each decade and modern electronics embraces miniature circuitry often involving digital ICs controlled by lines of code.

Current electronics hobbyist magazines rightfully focus on topics that are contemporary or important to their advertisers; for example, promoting mixed-signal ICs, DSP, microcontrollers and the kits they describe and then sell for income. Nuts and Volts is 1 example. Both analog RF and AF design increasingly lies in the hands of a small group of specialists, enthusiasts and students.

Yet, we *persevere*. Sharing our knowledge, circuits, experiences and references on the Internet helps sustain our small global community. That's the site purpose— sharing the (warts and all) experiments + basic information of a lay-person.

The Emitter Choke in Common Base RF Amps

This web page covers reverse isolation — a really important topic. 2 principle amplifiers we employ for strong reverse isolation are the common base BJT and common gate JFET alone or in cascode with other amplifier topologies.

Some comments regarding using a radio frequency choke in the common base amplifier follow.

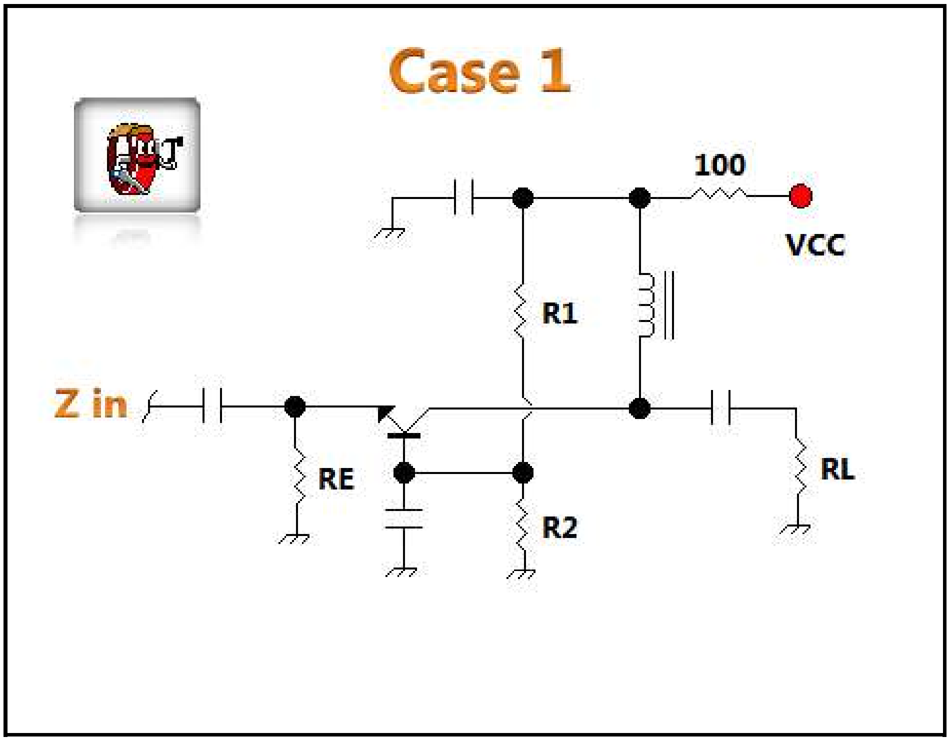

Above — Case 1: Emitter resistor only.

Apart from providing DC bias along with R1 and R2, emitter resistor RE plays another important role. Despending on its value, a portion of the input AC signal may pass through RE to ground instead of going through the transistor — degrading signal amplitude and noise figure. To minimize this, the resistor value should be many times (~10X or more) than the input impedance of the amplifier.

Although we might bias a common base amp to give an input Z of 50 Ω, often we'll choose a much lower input Z to get higher voltage gain. Input Z = 26 / Ie where Ie = mA; so if you bias for 5 mA, you are looking at an input Z of ~5 Ohms. In that case, a low value bias resistor such as 100 Ω won't shunt much of the input signal to ground, nor will it likely contribute much noise.

For most common base RF amps, a correctly chosen emitter resistor is all that's needed to decouple the AC signal and using an emitter choke proves hard to justify. However, it's important to understand how to apply an emitter choke since the basic principle also extends to the common gate JFET amplifer and other circuits.

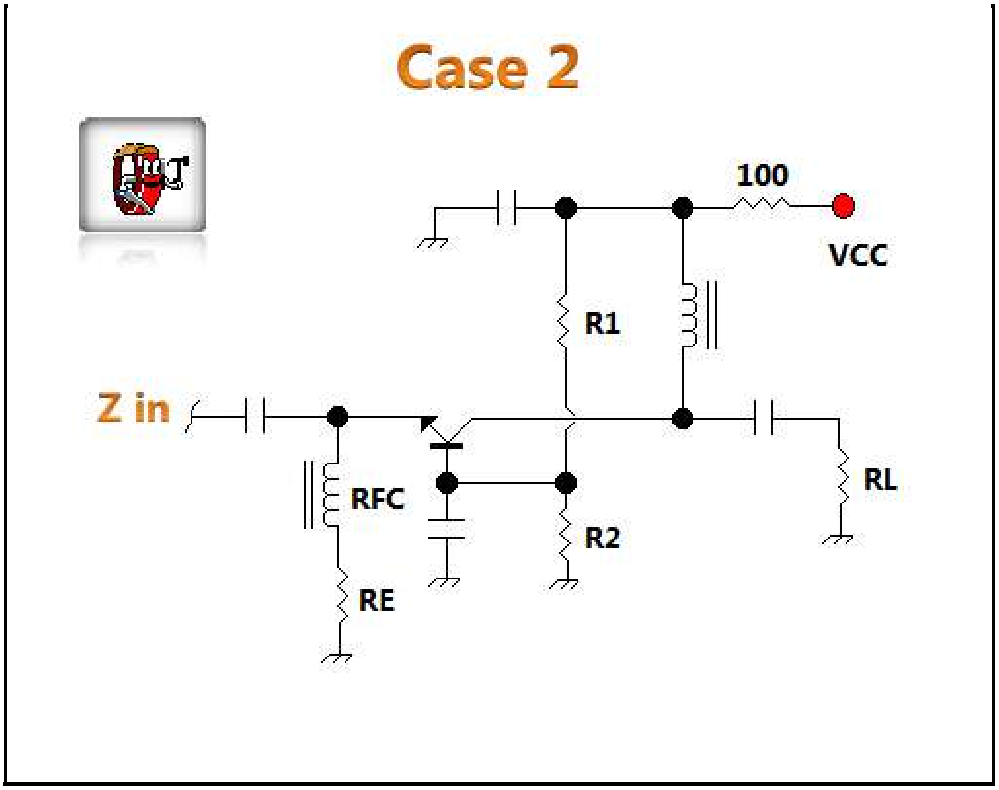

Above — Case 2: Emitter resistor plus a choke.

The choke’s main purpose is to block or choke RF from passing to ground. The ideal choke would present infinite impedance to AC signals, plus 0 resistance to DC voltage. In reality, "ideal" = fantasy electronics and you can simply estimate a choke's inductive reactance using the classic formula (XL =2*PI * Freq * L).

Using a coil (and not just a resistor) is generally better for decoupling — although how much better might be debatable. If the inductive reactance (XL) of the coil is significantly higher than the input impedance of the transistor, then all of the input signal power goes to the transistor.

By convention, a minimum choke XL should be at least 3 times the input resistance, however, the self-resonant frequency of the coil must be significantly higher than the applied frequency. Thus, an ideal range of inductive reactance exists, and too little or too much can degrade performance. Many builders target an XL around 10 times higher than the transistor input impedance at the lowest operating frequency.

Example: For a common base amp biased for an input Z of 50 ohms, the minimum inductive reactance (XL) for the choke = 500 ohms. To calculate the inductance of an emitter choke for this amp at 50 MHz, we re-arrange the formula to solve for L.

L = XL / 2* Pi * F
L minimum = 1.59 uH.
Winding the choke on a ferrite core, or possibly a bead for VHF often means less turns, less winding capacitance and a higher self-resonant frequency.

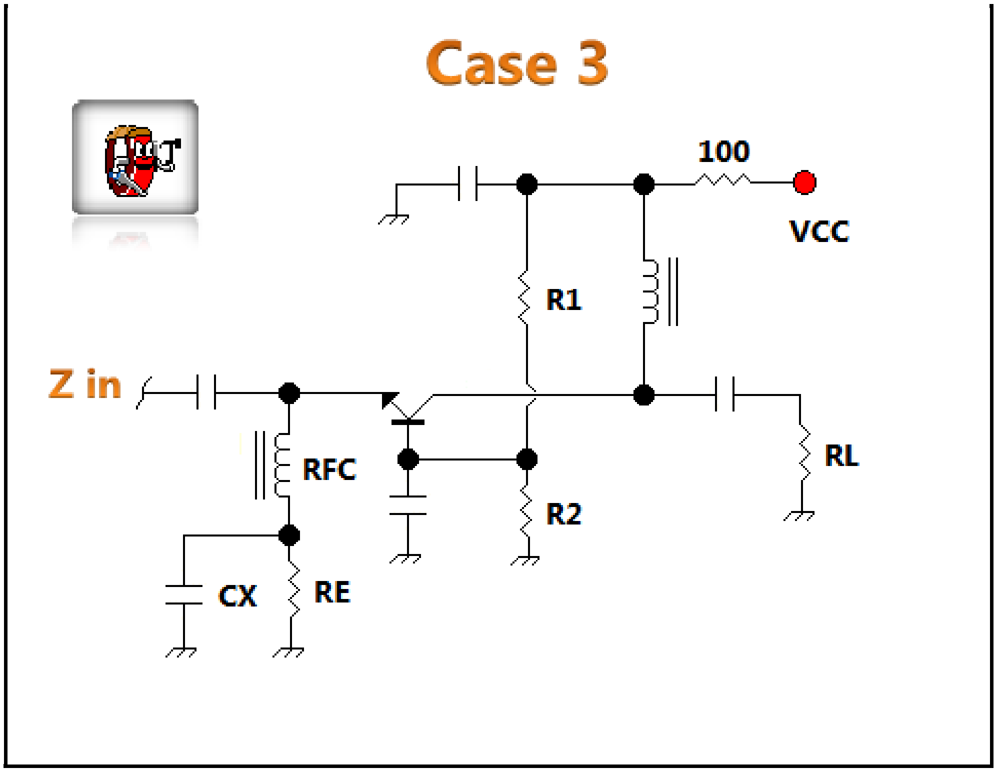

Above — Case 3: Bypassed emitter resistor plus choke.

The primary purpose of the capacitor across RE is to filter resistor noise — but that is only an issue well below the frequency of interest and it should not be relevant at high frequencies where the choke reactance is significant. There may be some useful effect if the self-resonant frequency of the capacitor Cx is above the frequency of interest.

You can only use bypass capacitor **Cx** when a choke is implemented.

A 0.1 uF may be useless at high frequency. In error, I've used this value previously on the site; after 14 years of experimenting, I've learned a lot from my design mistakes.

The case of the common gate JFET amplifier

This discussion also informs common gate JFET amplifer design. The JFET source requires signal decoupling similar to the emitter of the bipolar transistor discussed above.

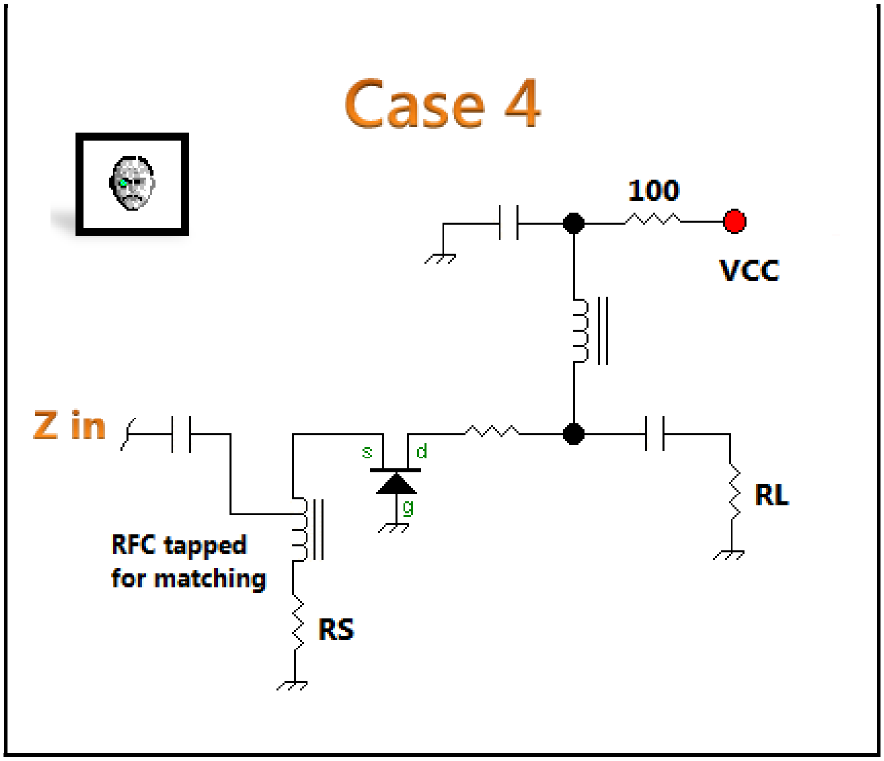

Above — Case 4: A choke plus source resistor will commonly be the "go to" design. Things get a bit more complicated with some, but not all JFET circuits — engineers often match the JFET input for a low noise figure rather than just the "correct" input impedance.

A good example follows: We might place a common gate JFET amp after a diode ring mixer because of the wideband load it presents to the mixer's RF port. The best noise match may occur with a hypothetical input Z of ~70 Ω (this argument represents an advanced topic).

After measuring the JFET pinchoff voltage and Idss, you would likely find that a source bias resistor of ~100 Ohms would be needed. This R value is so close to the JFET input Z that signal losses to ground would occur — demanding a choke for signal amplitude preservation, plus impedance and noise figure control.

RF — Test and Measurement

VFO - 2011

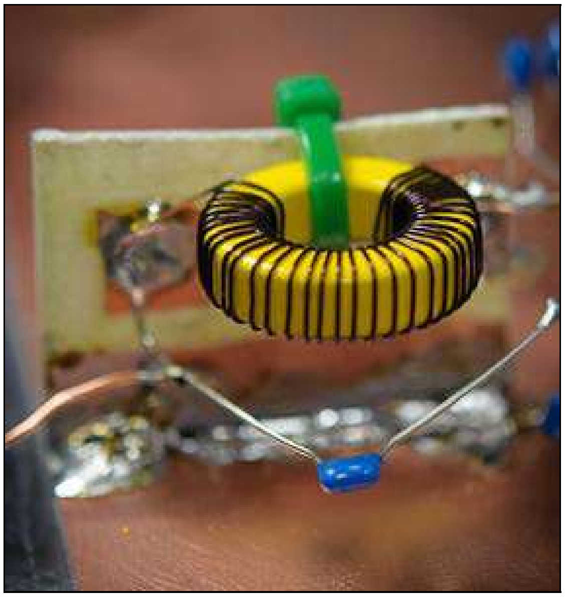

Building VFOs in 2011 might seem an irrelevant exercise given the move to and evolution of digital signal generators laden with bells and whistles like memories and audio or video frequency displays.

A successful L-C VFO requires skill, patience and some good parts to pull off — else, a "drift monster" may result. Despite their limitations, it's possible to build L-C VFOs with low frequency drift, distortion and phase noise; our typical VFO performance markers. L-C VFOs don't require programming skills or equipment to encode a microprocessor — making them a good choice for people who don't build or can't afford kit oscillators. Most of all, they kindle creativity, problem solving and pride when your oscillator actually works as planned. Junk box radio; my passion.

This material reflects lots of empiricism; not pure science. It's really *your* VFO design odyssey; a chance to think creatively and critically to sort out what works and what's folly.
Countless web pages discuss VFO design and I encourage you to search for and read them. Wes' EMRFD oscillator and temperature compensation notes = essential reading. Only your first 25 VFOs will prove difficult — it gets easier after that.

VFO 2011 Topics:

1. Frequency Stability Notes
2. Vackar VFO Experiments
3. HF Signal Generator
4. Miscellaneous Bits

---

1. Frequency Stability Notes

Building an oscillator that stays on frequency purports our greatest challenge and goal in L-C VFO design. Since drifting VFOs pose a source of frustration, I cover some topics that may help your VFO stay on frequency — do they help?

What is good drift parameter?

I'm uncertain, for after warm-up, I've measured kits that drifted 50-150 Hertz per hour, built L-C VFOs that drifted under 20 Hz per hour and every once and a while, build a drift monster VFO that sweeps upward at 2 - 8 hertz per minute! Likely under 20 Hertz per hour after warm up = a gold standard to compare against. You should be able to listen to a 10-20 minute QSO with no re-tuning, however, this assumes the transmitting stations are locked on frequency.

1. Unloaded Q and Frequency Stability

The number 1 reason to employ high resonator Q in oscillators is to obtain low phase noise. Secondly, the very steep phase slope through high Q resonance minimizes the effect of amplifier phase shifts caused by temperature changes and this in turn, minimizes any amplifier-induced frequency instability.

Long term frequency stability is chiefly dependent on the temperature, environmental and age stability of the resonator components regardless of Q.

I often see designs featuring high Q inductors wound on powdered iron toroids complimented with trashy, low Q variable and/or fixed capacitors, If you design for a high Q tank to minimize phase noise, consider using a high Q coil plus appropriately temperature stable, high Q capacitors.

2. Temperature Stable Inductors

Knowing that I'm venturing into a topic of great debate and lore, the inductor is 1/2 of the VFO resonator and thus a major source of temperature drift in L-C VFOs. Since MF and some HF VFO designs *may* preclude using the *inherently more temperature stable air wound inductor*, powdered iron toroids dominate our evermore compact designs. Many builders choose #6 material, although the lower temperature coefficient of #7 material theoretically should be better — however, my experiments have failed to measure a significant difference between these 2.

Some builders prefer size 68 inductors, for the bigger core is less affected by heating than smaller size toroids. My experience suggests that providing the VFO amplifier current is kept low, both size 50 and 68 are both suitable and the inductance needed should inform the core size.

I used to think that heavier gauge wire created greater frequency stability than smaller gauge wire until Wes, W7ZOI, woke me up. As it turns out, smaller gauge wire is often better for thermal stability because smaller gauge wire lies closer against the toroid core. Winding stiffer, heavier gauge wire creates more air gaps than smaller gauge wire and air gaps expand and contract during temperature changes. Smaller gauge wire will have a reduced Q, but it won't be as significantly lower as you might guess. As possible, I prefer tightly wound number 28 wire. 26 gauge wire tends to be my maximum size wire for VFO coils, however I suggest you make your own conclusions.

Wash your hands before winding and use both hands to actively move both the toroid and wire for tight turns. Take your time, ensure steady wire pressure and avoid kinking your wire. Taps increase the likelihood for air gaps — mitigate this by stripping the 2 tap forming wires as close to the toroid as possible and twist them into 1 wire right down tightly to the toroid edge to reduce any air gap.

The thermal stability characteristics of wire can be mitigated somewhat by annealing the wire with temperature cycling or by dunking it in boiling water. Roy, W7EL first reported annealing coils in 1980 and this has been confirmed during experiments by builders using temperature controlled ovens. I don't boil my coils any more.

3. Double Stacked Toroids

I noticed a new trend in VFO design is to stack 2 powdered iron toroid inductors. This allows the builder to double the inductance per number of windings over a single toroidal inductor. In an L-C VFO, the goal of these builders possibly is to reduce heating effects, increase unloaded Q, or perhaps to reduce core magnet flux density. For me the goal is far simpler, I just want to make compact, large L value inductors for 3 MHz and less.

Above — A T68-6 hamburger. The two T68-6 cores were epoxy glued together and compressed lightly in a vice for several hours. One of the initial tests I performed was to see if boiling the stacked coil affected the epoxy glue. The glue was not effected by annealing wire on a stacked coil with 5 or even 10 minutes of boiling in water. As mentioned, I stopped boiling my VFO inductors as tightly winding them with 26 gauge wire seems to work well.

I hold concern that stacked toroids may create more wire-air gaps when compared to a single toroid and stay with 1 toroid as possible. In compact antenna tuners and other non VFO projects, this isn't an issue.

4. VFO Tank Capacitors

We choose VFO tank capacitors to avoid temperature change caused frequency drift, or to counter drift during our temperature compensation process.

Many authors have published guidelines for long term temperature stability. It's important to consider these guidelines, but also try whatever works. I believe the following arguments are accurate based upon my experiments:

1. Multiple NP0 or C0G (0 temperature-compensation) tank caps: Most builders minimally use 4 or more C0G or NPO capacitors to reduce heating effects and to average out temperature coefficient variations.

2. No VFO tank capacitors from online surplus parts stores; buy new stock from known and reputable manufacturers. Grab bags and musty, old, surplus parts can obviate good design.

3. Trimmer and tuning caps need to be temperature stable. Air variable capacitors = my favorite, as possible.

4. Varactor, or diode tuning generally = more drift and a greater need for temperature compensation.

5. Employ short, stiff capacitor leads. I use 100 volt or higher voltage C0G tank caps as they tend to have thicker leads that stay put — perfect for Ugly, Manhattan, or Chuck Adam's MUPPET construction.

5. Temperature Compensation

The goal of temperature compensation is to cancel the tendency of the VFO to drift in 1 direction — easier said than done + very time consuming. A web search for VFO temperature compensation will yield many good write-ups. I feel it's partly art, partly luck and partly science. Your net VFO temperature coefficient can be affected by so many variables, so no 1 recipe will ensure a low drift VFO. Experiment, allow a lot of time to assess your changes and be patient — you'll figure it out.

The simplest way to test for drift involves watching a frequency counter, but if you don't have one, you might use a commercial, frequency stable (synthesized) receiver set in the SSB/CW mode. I use both. Experienced builders often employ an oven to test their temperature compensation at different, controlled temperatures. Wes, W7ZOI employs a styrofoam cooler housing a light bulb heat source controlled by a Variac. See EMRFD for more details and a photograph.

In 2011, I decided to build up a supply of temperature compensation capacitors and keep them in their own parts bin.

TEMPCO
CAPS

Above — "Tempco caps". A parts drawer containing polystyrene capacitors from 10 to 270 pF plus some 56 pF ceramic N750 capacitors for negative temperature compensation. I purchased these capacitors on eBay.

For capacitors other than NP0 (which use 0 instead of a ppm value), the temperature coefficient = **P** for positive and **N** for negative, followed by a 3-digit value specifying ppm/°C. For example, N220 is - 200 ppm/°C. and P100 is +100 ppm/°C.

I use NP0 and C0G ceramic capacitors interchangeably for both tuning and RF bypassing the VFO tank resonator. For C0G/NP0 *temperature compensation bypass*, I normally apply 0.01 or 0.001 μF caps, however, the more expensive 0.1 μF COG ceramic capacitors are still sold if you need C0G/NP0 bypass <= 7 MHz.

If your VFO is drifting upward you might insert 1 or more positive coefficient capacitor(s). If your VFO drifts downward, then try using negative coefficient value(s). Sometimes just 1 capacitor will do the job.

Since I don't stock any positive coefficient capacitors for positive coefficient compensation, I might try a adding a silver mica capacitor. ***Caution*** silver mica capacitors are extremely non-predictable and can't be universally recommended in temperature compensation schemes. You might also try swapping out 1 or more of your main tank NP0 or C0G capacitors in case they are bad; sometimes it gets frustrating. I provide some temperature compensation examples on the QRP Modules 2011 web page in the 7 MHz VCO section.

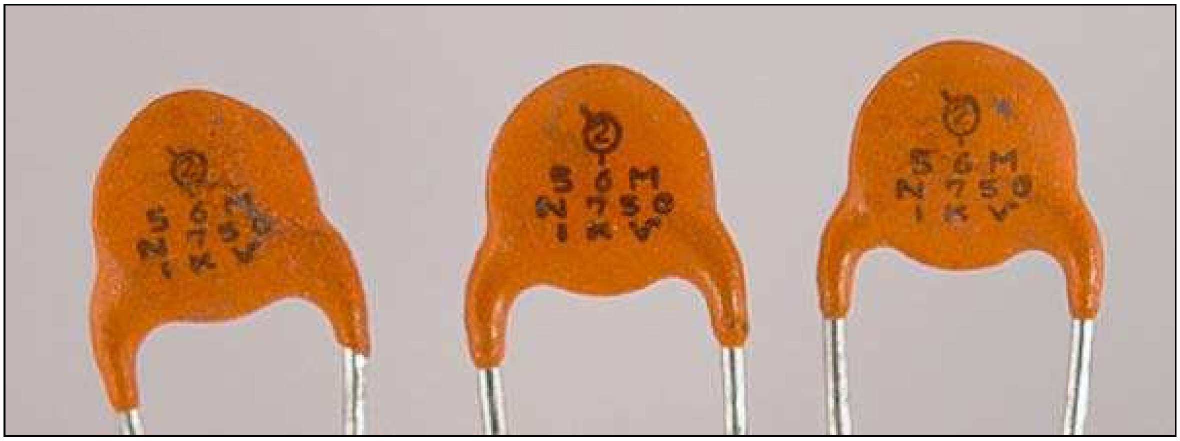

Above — 56 pF N750 ceramic capacitors rated at 1KV

6. Mechanical Rigidity

Movement of your VFO tank parts may lead to frequency instability. For example,

1. Well secure your single-sided only copper board. I use at least 4 number 8 bolts — 4-40 hardware is too light. Boards can warp over time if not lashed down properly. Aggressively bolt down any variable capacitors. No tank parts should move.

2. Anchor your inductor so it cannot budge: nylon bolts, zap-straps, glue - whatever.

3. Consider placing the VFO in a strong chassis with rubber feet.

4. Buss wires should be made from thicker gauge, well anchored wire.

7. Miscellaneous Points

1. Regulate the VFO amplifier DC voltage and wideband filter it. Voltage regulators require RF and often AF bypass to attenuate any noise or ripple riding on the DC.

> A decoupling resistor with a bypass cap on either side will widen your DC supply filtering bandwidth and deserves strong consideration. A poorly filtered DC supply can easily transmit the VFO tank energy to other stages along your DC lines and also may allow noise on the DC supply to modulate your VFO and increase phase noise.

2. You should have the buffer + a load resistor connected to your VFO when testing. **Do your temperature stability work after the buffer is built and the VFO is in its case.**

3.Stick your VFOs in an air tight, RF tight case to minimize air temperature changes and RF leakage respectively. Sometimes a VFO will drift once in a case because any radiated buffer amplifier heat will warm up the inside of the chassis. This usually levels off after warm up.

4. Modern voltage regulators may significantly reduce noise compared to a zener diode regulator. Specific low noise and low temperature coefficient voltage regulators are available, but maybe overkill for you. Whatever you use — filter it well —
The Micrel MIC5209-5.0BS in SOT223 sits in a couple of my reference oscillators.

5. JFET Gate clamping diodes may increase phase noise, but not prohibitively so in most popcorn designs.

6. When winding toroid inductors, wind 2 extra turns. When finished, unwind the first 2 turns since they are usually loosely wound and prime culprits for air gaps.

7. Since magnet wire comes off small spools, wire has a natural curve or radius — ensure you wind your coils according to the natural curve of the wire.

8. The need to secure powdered iron windings with dope, wax, goop, etc. is over-emphasized and usually unnecessary.

---

2. Vackar VFO Experiments

Some builders proclaim the Vackar as the "King of VFOs". I built a couple and became impressed by the low distortion and less than 5 Hertz per hour long-term drift achieved in my 2 designs. Inspired by work from Iulian, YO3DAC entitled Very Low Phase Noise Vackar VFO for HF Transceivers *(link and reference used by permission of Iulian)*, I crafted my version from his notes and schematic.

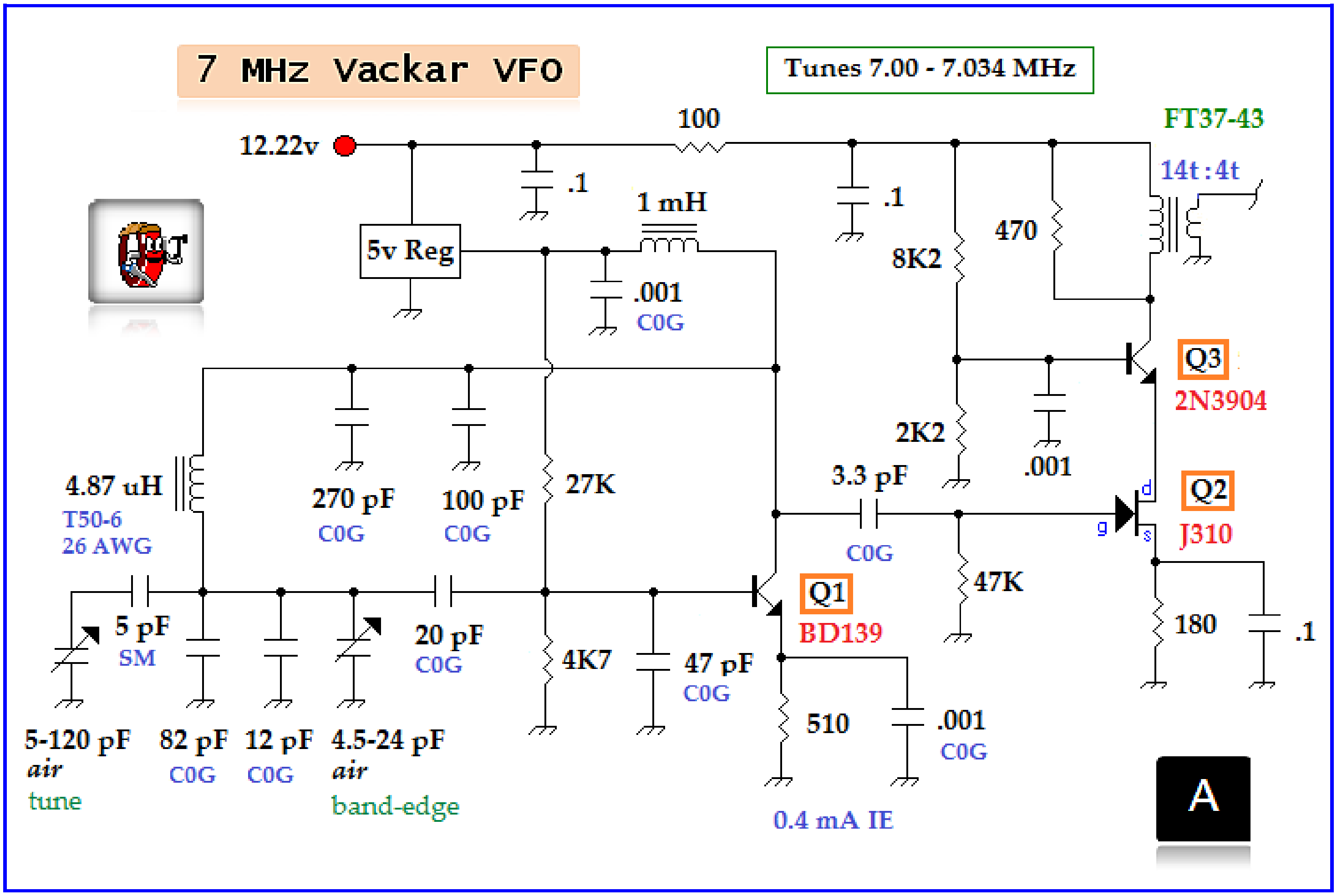
7 MHz Vackar VFO
Tunes 7.00 - 7.034 MHz
12.22v
100
FT37-43
14t:4t
.1
1 mH
.1
470
5v Reg
8K2
.001
C0G
Q3
2N3904
2K2
.001
3.3 pF
C0G
Q2
J310
d
g
s
4.87 uH
T50-6
26 AWG
270 pF
C0G
100 pF
C0G
27K
47K
Q1
BD139
180
.1
5 pF
SM
20 pF
C0G
4K7
47 pF
C0G
510
.001
C0G
5-120 pF
air
tune
82 pF
C0G
12 pF
C0G
4.5-24 pF
air
band-edge
0.4 mA IE
A

Above — Schematic of the Vackar VFO employing a BD139 — a large area transistor, to reduce 1/f noise. Iulian shared many design pearls in his paper and I won't repeat them. I ran Q1 with 0.4 mA emitter current to reduce heating and flicker noise. It's difficult to measure flicker noise, so no objective comments can be made.

I limited the tuning range to 34 KHz since the tuning capacitor lacked reduction gear and I was born with fumble fingers. As a CW operator - you'll find me down at the bottom of the band away from the RRTY anyhow. Increasing the 5 pF cap coupling the tuning capacitor to the tank increases the tuning range as expected.

Temperature compensating my VFO with the 5 pF silver mica capacitor proved a gamble since SM caps are unpredictable and often best avoided. In *my* VFO, it worked perfectly, however. This circuit is difficult to replicate and *not recommended* for new builders. Temperature compensation provided the sublime frequency stability.

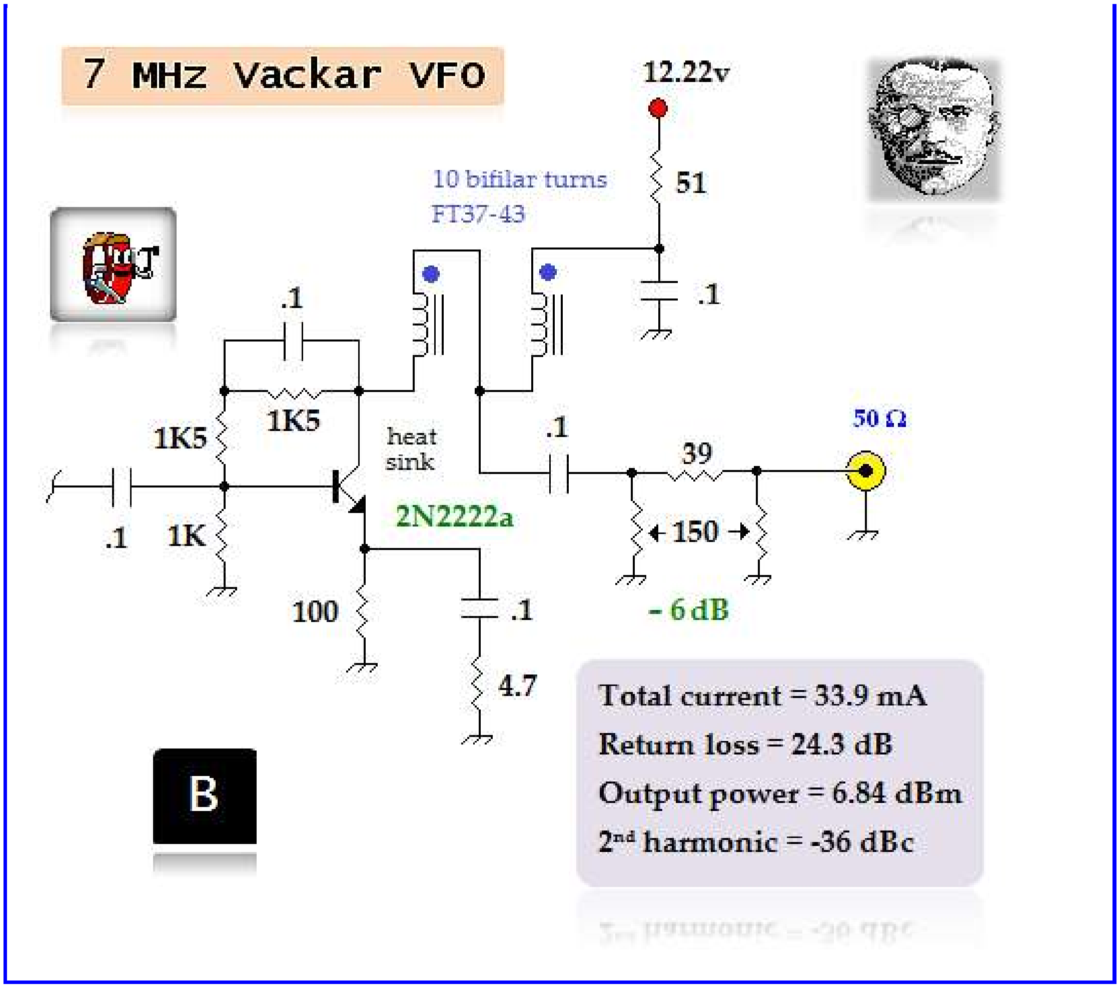
7 MHz Vackar VFO
12.22v
51
10 bifilar turns
FT37-43
.1
.1
1K5
1K5
heat
sink
.1
39
50 Ω
.1
1K
2N2222a
150
100
.1
- 6 dB
4.7
B
Total current = 33.9 mA
Return loss = 24.3 dB
Output power = 6.84 dBm
2nd harmonic = -36 dBc

Above — The final amp and measured output data. Measuring the 2nd harmonic 36 dB down without any tuned circuit or low-pass filter rocked. I ran nearly 22 mA of emitter current to bump up the return loss and spectral purity. A 2N5109 or 2N3866 would likely do a better job with less current. Total current = the entire VFO current. I glued a drilled copper penny on the 2N2222 to dissipate heat.

A photo of my version of a Vackar VFO. My design goals included low phase noise, low distortion, a return loss over 20 dB, good reverse isolation and ~7 dBm output power. I believe all VFOs are experimental; you build to suit whatever tuning capacitor or varactors you have, plus design around constraints such as total current, tuning range and other personal criteria.

Unlike harmonic distortion, oscillator phase noise, being close to the oscillation frequency, cannot be removed by filtering nor limiting — you must design for low phase noise. Modern digital VFOs are well harmonically filtered, and any phase noise depends on the DDS clock employed, so check the DDS specifications carefully if you go the DDS VFO route.

I'll be the first to state I'm no expert with VFOs, however, likely the only way to become expert is to build many and learn from your mistakes.

Above — A 7 MHz Vackar VFO with the lid off

Sound Test?

Although this technique raises the ire of some builders, I test my VFOs in a nearby receiver. The VFO output was terminated with a 51 ohm resistor that was also attached to my frequency counter via alligator clips and wire. I tuned a nearby CW superheterodyne receiver to 7.00 MHz with the audio beat note centered in its 600 hertz wide I.F. filter and watched the counter plus listened to the receiver.

Click for a 1 minute 32 second audio file of the result (it stayed perfectly on frequency for ~5 hours of testing before I got a bad headache from listening to it and shut it off). *You can initially hear a station in the back ground despite only having a 45 cm piece of wire as the receiver antenna.* The VFO slowly drifts down to 6999996 Hz and then slowly back up to 7000000 Hz. You can hear the signal amplitude decrease as the VFO drifts down. So it doesn't stay perfectly on frequency, but slowly cycles up and down a few hertz. This VFO is my lab temperature stability benchmark for an L-C VFO.

A badly drifting VFO will move out of the test receiver I.F. pass band and sound like a Theremin as it does. Testing in a receiver; places the VFO in the exact circumstance it will be used — beating RF to mix to another frequency; in this case, base band audio.

The temperature stability and compensation of any VFO schematic are rarely reproducible since there are just too many variables. Try your best to get the drift out of your VFO using low temperature coefficient capacitors (NP0/C0G) and then after that, temperature compensate. Even today, I occasionally build a drift monster VFO and become frustrated. VFO design is not for the faint of heart and it's no wonder that many builders make a VXO, or cave in and build or buy a DDS signal generator.

---

3. HF Signal Generator

50 OHM
GAIN
HI
LOW
FINE

Above — I built a general purpose ~2.8 to 10.8 MHz signal generator (SG) for my lab in 2011. The first VFO topology tested was the Vackar. In my version, while employing a 100K ohm resistor as the buffer, the VFO only tuned from about 4 to 8 MHz and suffered from *extreme* amplitude variation as I changed the frequency across its range. For sweeping filters or measuring Q, a signal leveling circuit would be needed as normally we like our SG output to be flat across its frequency range. I later changed to a Hartley VFO because of its flatter output and the wider available frequency range with any given resonator.

This initial Vackar VFO experiment wasn't a total waste as I learned a way to accurately sweep a **D**evice **U**nder **T**est with an unlevel amplitude SG. Measure the peak-peak voltages of the DUT with a signal generator and an oscilloscope in the same manner we measure insertion loss or gain in a 50 ohm system: Measure the peak-to-peak voltage with the DUT in line; disconnect the DUT, insert a barrel connector and then re-measure.

The dBm difference between the 2 becomes the dB value to plot for *that* frequency. To a sweep a filter, say for example, a band-pass filter, find the center frequency and then sweep below and above that CF while plotting the dB versus frequency.
This "in and thorough" measurement described takes time, but resolves any frequency versus amplitude issues and can be used to test signal generators. We tend to ignore things like cable loss versus frequency and scope or spectrum analyzer ripple.

Still, it will be easier to just use a Hartley VFO where our sweeps are assumed to be level due to the flatness of the amplitude versus frequency for **small excursions** such as 3 dB band-pass filter sweeps. Do not expect amplitude flatness over wide excursions however — this requires additional circuitry.

To reduce noise and boost fidelity, this SG runs modest current and was not designed for battery use.

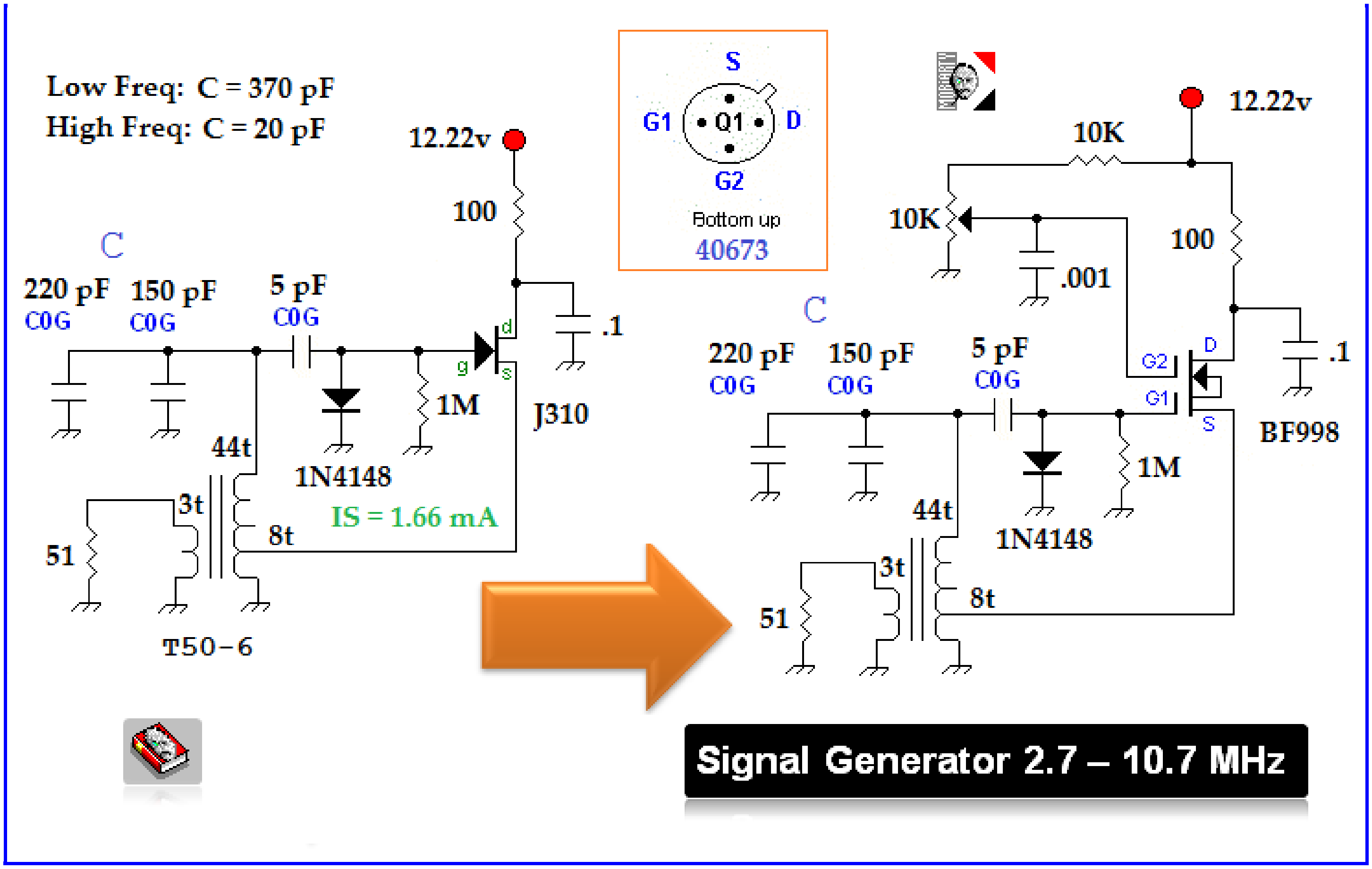
Low Freq: C = 370 pF
High Freq: C = 20 pF
12.22v
100
C
220 pF
C0G
150 pF
C0G
5 pF
C0G
d
g
s
.1
1M
J310
44t
3t
8t
51
1N4148
IS = 1.66 mA
T50-6
S
G1
Q1
D
G2
Bottom up
40673
10K
10K
.001
12.22v
100
C
220 pF
C0G
150 pF
C0G
5 pF
C0G
G2
D
G1
S
.1
BF998
1M
44t
3t
8t
51
1N4148
Signal Generator 2.7 – 10.7 MHz

Above — My Hartley VFO is morphed into a double-gate MOSFET VFO; this was a mistake and I make lots of them.

When venturing out, it's often best to confirm a proven design is working before morphing it to something untried. Shown above left is the Hartley oscillator from EMRFD Chapter 7 sans buffer. Fixed "tuning" capacitors; either 20 pF (not shown) or 370 pF (150 pF + 220 pF) represent the intended high and low frequency swing of my air variable tuning capacitors.

I wanted a variable amplitude VFO and thus replaced the JFET with a double-gate MOSFET using a simple variable voltage divider to control gate 2. I showed this to Wes, W7ZOI and he informed me that the flicker noise of MOSFETs precludes their use in oscillators. I have always wondered why I've never seen MOSFET VFOs in any radio literature.

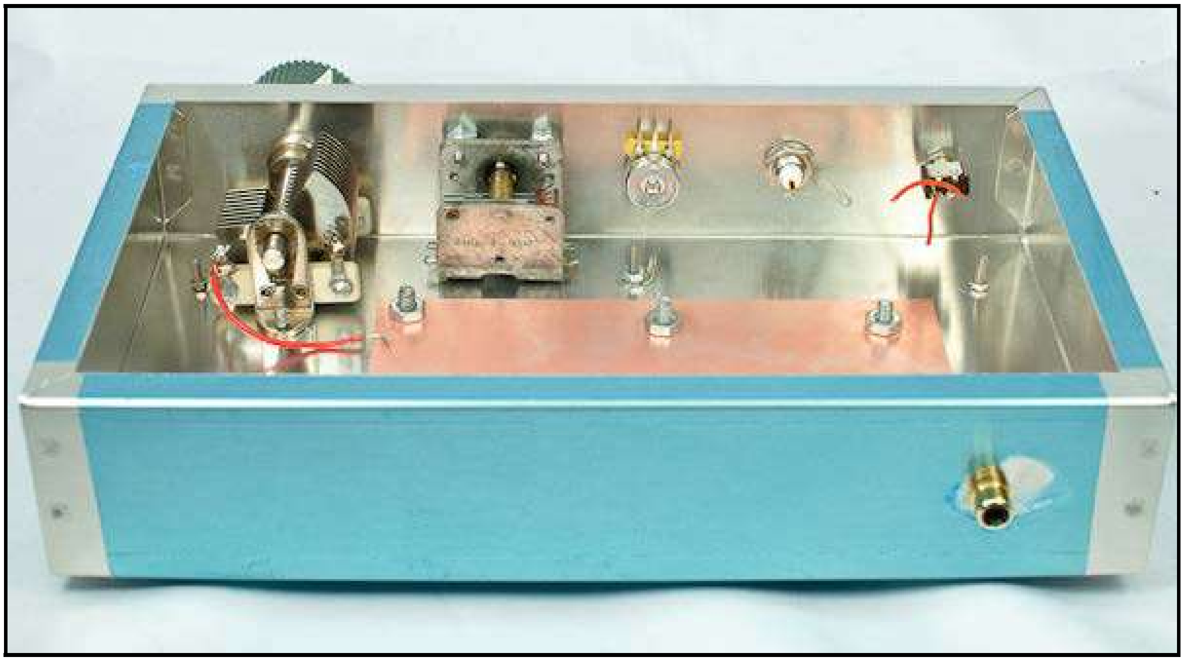

Above — My project chassis fitted with hardware. I employed 2 air variable tuning capacitors — the fine tuning capacitor ranges 13.6 to 27.5 pF, features built in 6:1 reduction gear and was purchased from Doug DeMaw many years ago. I secured the main copper board with 6 number 8 bolts. Rubber feet provide a stable, shock resistant base for the sheet metal box.

Final Build

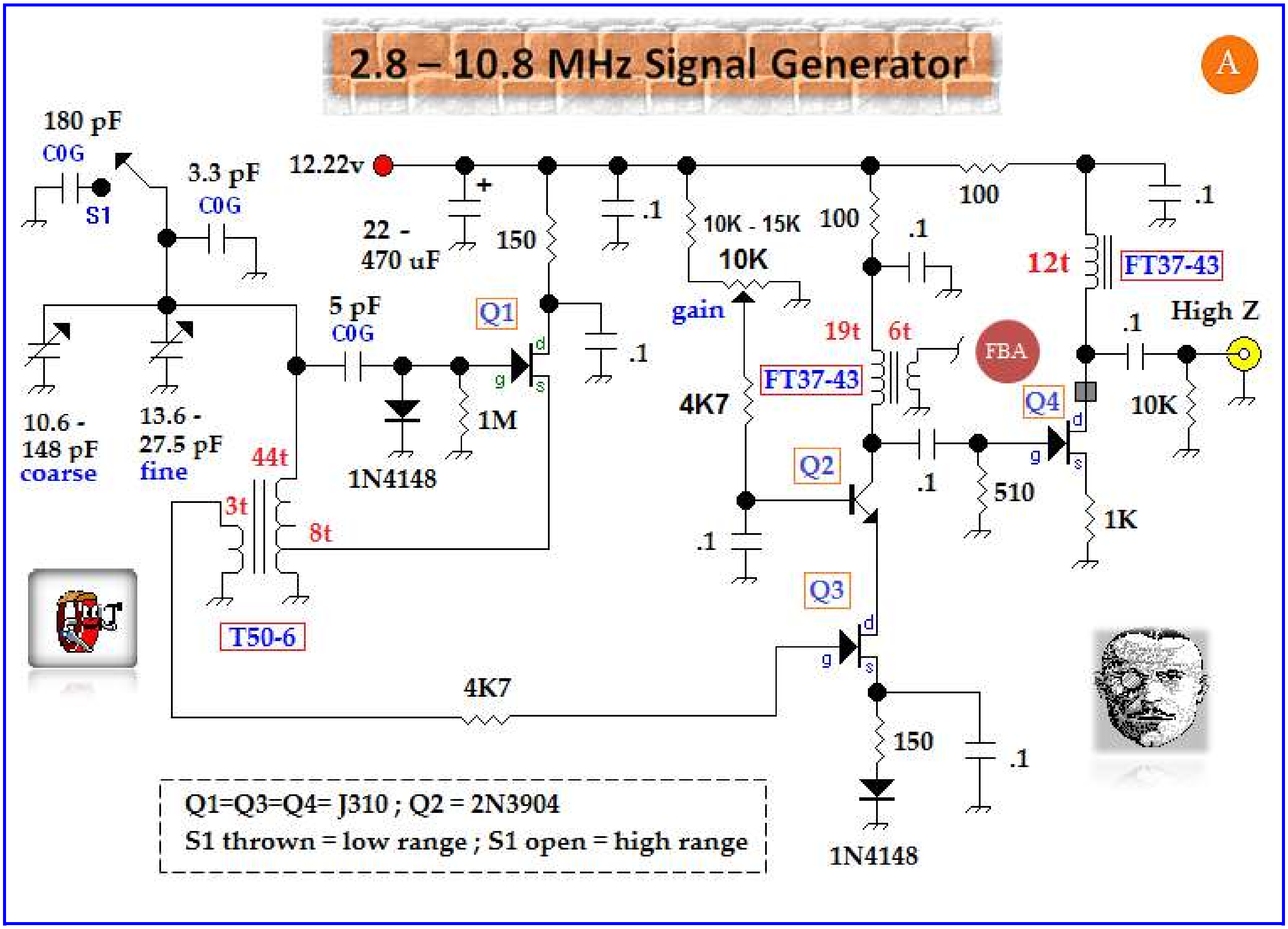
2.8 – 10.8 MHz Signal Generator
A
180 pF
C0G
S1
3.3 pF
C0G
12.22v
+
22 -
470 uF
150
.1
10K - 15K
10K
gain
100
.1
100
.1
12t
FT37-43
High Z
10.6 -
148 pF
coarse
13.6 -
27.5 pF
fine
5 pF
C0G
Q1
d
g
s
.1
1M
1N4148
19t
6t
FBA
FT37-43
4K7
.1
Q4
d
g
s
10K
44t
3t
8t
T50-6
Q2
.1
510
1K
.1
Q3
d
g
s
4K7
150
.1
1N4148
Q1=Q3=Q4= J310 ; Q2 = 2N3904
S1 thrown = low range ; S1 open = high range

Above — Oscillator + buffer schematics of the latest version of my signal generator. I spent 1 evening playing with VFO designs and settled on the simple Hartley from EMRFD, Figure 7 .27. The 3 turn link provided lower distortion than coupling the oscillator to its buffer by the JFET source or gate.

Regulated 12.2 VDC powers the oscillator; avoiding the typical 5-9 VDC voltage regulator we normally use. A 22 to 470 μF cap should be employed to filter any voltage regulator noise from our DC supply. Mine has a 470 μF capacitor.

1 hour drift lies under 40 Hertz when averaged from 15 different frequency points between minimum and maximum. The Q2/Q3, Q4 and Q5 transformer inductances were optimized to allow good signal and/or matching performance in the ~ 2.5 to 10.8 MHz frequency range.

A hybrid cascode (hycas) buffer with variable base bias on Q2 forms the amplitude control for both the high impedance and low impedance outputs. The 510 ohm gate resistance on Q4 terminates the hycas amplifier and sets up a known output impedance to drive the 50 ohm feedback amp. I measured a greater than 22 dB return loss on the output of the 19:6 turn transformer from 4 to 14 MHz — indicating it drives the feedback amp reasonably well.

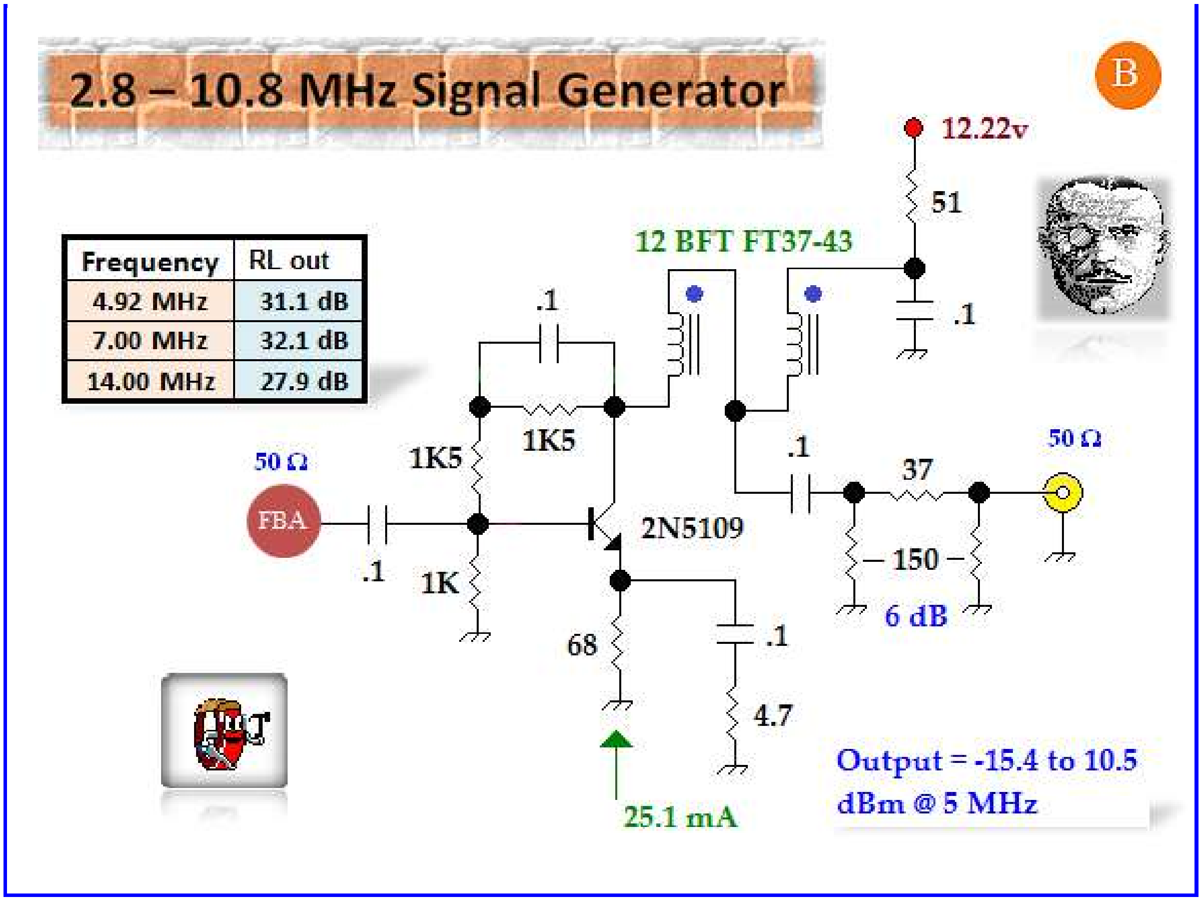
2.8 – 10.8 MHz Signal Generator
B
12.22v
51
12 BFT FT37-43
.1
.1
1K5
1K5
50 Ω
FBA
.1
1K
2N5109
68
.1
4.7
25.1 mA
.1
37
50 Ω
150
6 dB
Output = -15.4 to 10.5 dBm @ 5 MHz

Frequency	RL out
4.92 MHz	31.1 dB
7.00 MHz	32.1 dB
14.00 MHz	27.9 dB

Above — The 50 ohm impedance feedback amp. Running 25.1 mA current allowed a clean sine wave output up to 2.12 volts-peak to peak into a 50 Ω terminated oscilloscope, plus an output return loss of over 30 dB across the SG tuning range. 3 tabled output return loss measurements are shown; including an out-of-range 14 MHz measurement.

Two series resistors made up the 37 ohm "resistor" depicted in the 6 dB pad, although a 39 Ω resistor would work fine.

At low output amplitudes, I typically stick an external 6, 10 or 20 dB attenuator on the output since the hycas amp can distort the signal a little when the gain control is set to a really low bias voltage on Q2. Then I fine tune the output power with the gain control.

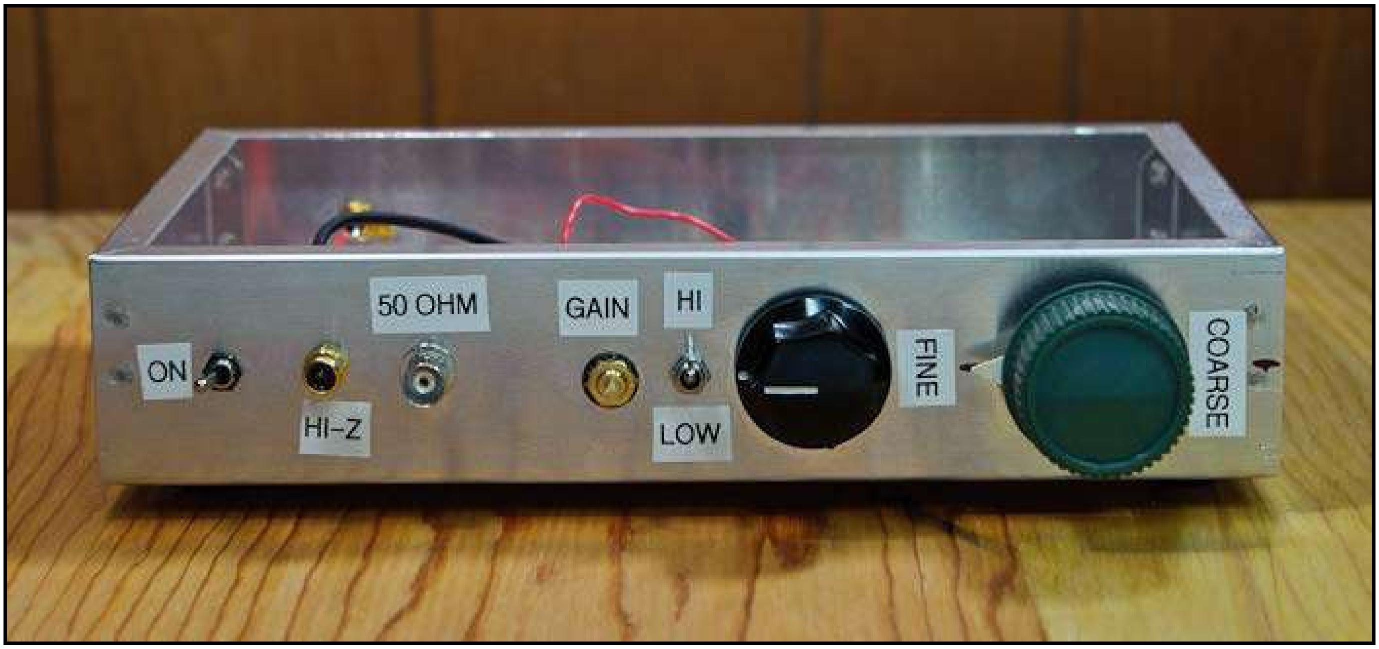
ON
HI-Z
50 OHM
GAIN
HI
LOW
FINE
COARSE

Above — A "lid off" front panel photograph. Click or click for other photos. I'm now using miniature pots with a shaft diameter of 3.18mm. The potentiometer shaft lacks a knob and I'll purchase some on my next parts order.

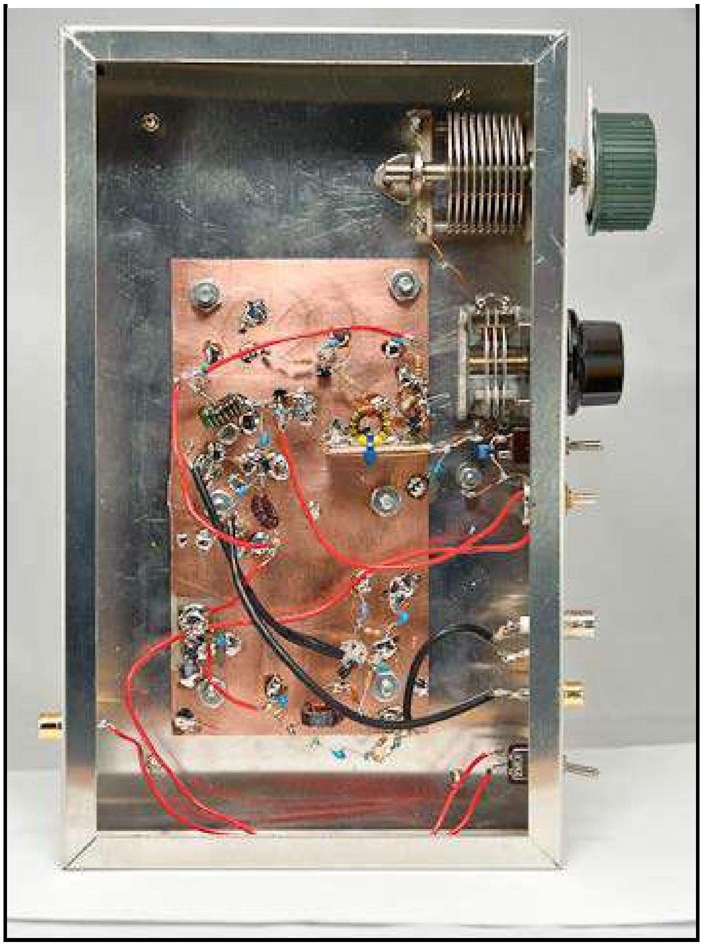

Above — The completed signal generator.

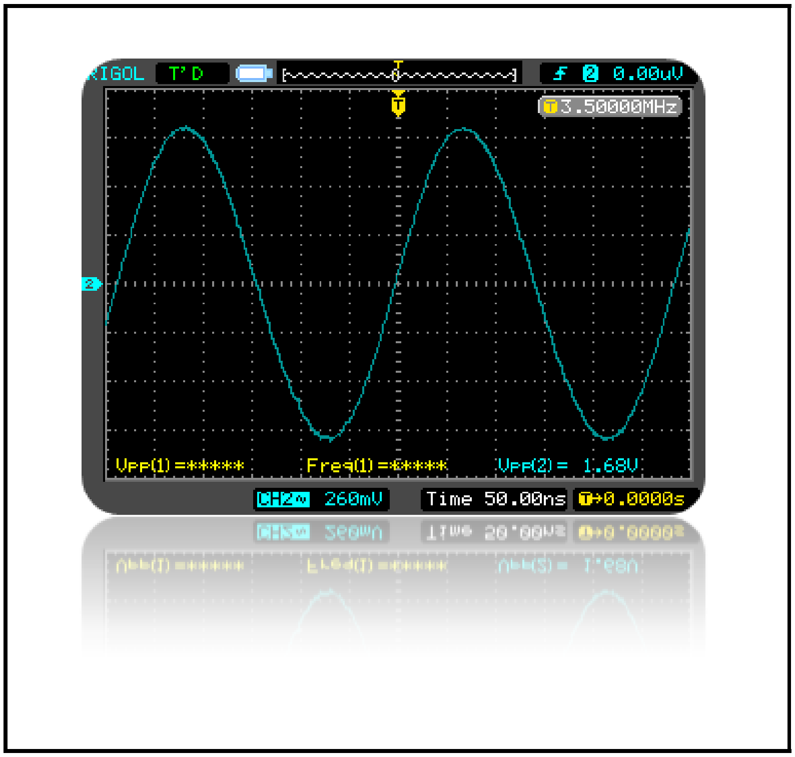
RIGOL T'D
0.00uV
3.50000MHz
Vpp(1)=*****
Freq(1)=*****
Vpp(2)= 1.68V
CH2 260mV
Time 50.00ns
0.0000s

Above — Signal generator output at 3.5 MHz.

I appreciate that the VFO tank would be difficult to replicate since the 2 air variable capacitors are unique, however this is true of most VFOs. Wes wrote some great notes in EMRFD Chapter 7 regarding copying signal generators and the versatility of the Hartley VFO. I hope this project furnishes some ideas that spawn you to build something better than I did. *(I've received over 350 related emails since posting this page in 2011* *and many readers have built really great VFOs — Way To Go !)*

QRP — PosData for December 17, 2013

I slightly boosted the tuning range from 2.8 to 10.8 MHz by dropping the 5 pF resonator capacitor to 3.3 pF in October 2013. The schematics now reflect this changes plus clarify 1-2 stumbling blocks readers had. For example: my use of a 100K gain pot plus a 150K maximum voltage limiting for the Q2 bias. My build still has these, however, I altered the schematic to show a common 10K gain pot, plus a fixed 10 to 15K resistor used to limit the Q2 bias to between 5 and 6 VDC maximum.

In reality, any reasonable pot and resistor will do since they function as simple voltage dividers. With a 12 volt supply, we don't want to drive the Q2 bias with more than ~ 5 to 6 VDC since this will just distort the AC signal as the hycas stage saturates. Measuring with a voltmeter, solder either a 10K, 12K, or 15K resistor to limit the maximum Q2 bias with the 10K potentiometer turned fully clockwise.

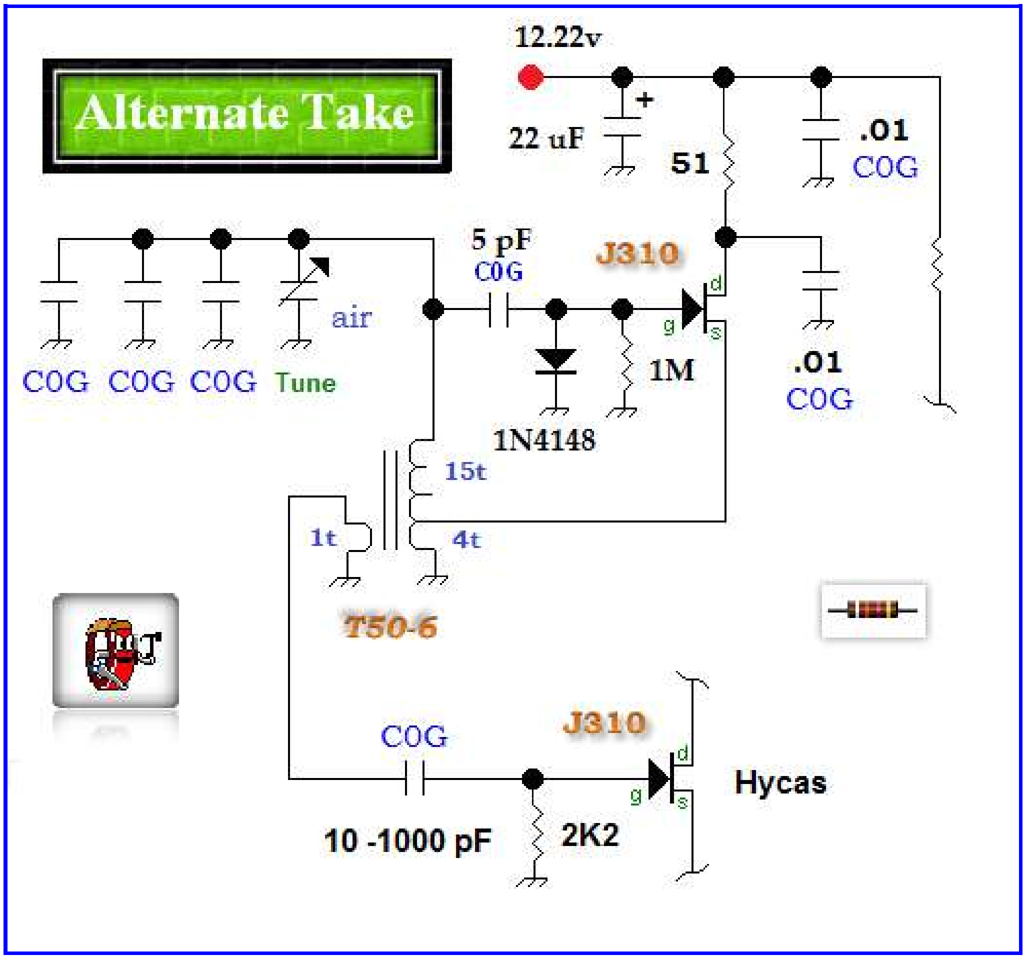

Above — Alternate way to couple the Hartley oscillator to the hycas buffer. Ground the JFET gate with a shunt resistor and lightly AC couple the JFET gate to the Hartley secondary coil with a series capacitor. You choose the capacitor value to limit the signal amplitude as needed. Our main goals are to lightly couple the Hartley tank to its buffer and avoid overdriving the hycas stage. Numerous examples of this "Alternate Take" circuit may be found on the QRP / SWL HomeBuilder web site.

---

4. Miscellaneous Bits

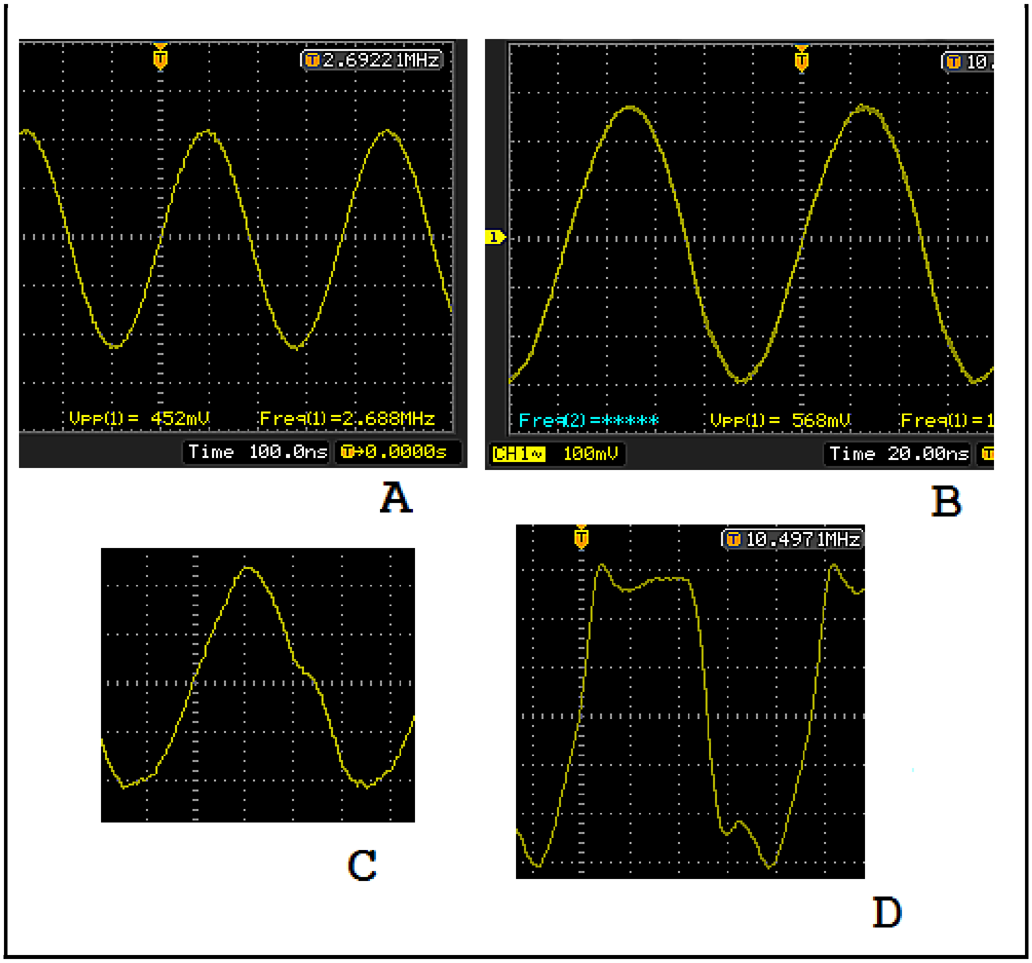

Above — Scope traces of an unbuffered Hartley oscillator with a X10 scope probe across a 51 ohm resistor across the 3 turn link. The unbuffered Hartley sine wave isn't harmonic free, but cleans up when properly buffered with a higher impedance amplifier. Figure A = the lowest frequency (2.7 MHz) — the distortion increased with frequency (Figure B was measured at 10.5 MHz). Figure C illustrates how slightly stronger output coupling with a 6 turn link trashes the output waveform — the strategy of using 2-3 links over the center of the main inductor works well.

Figure D is the 6 link coupled Figure C oscillator with the gate clamping diode removed; yikes! I spent 4 hours studying what different current, voltages, coupling and so forth do to the Harley oscillator. I recommend the Hartley topology because it's simple, always starts and versatile.

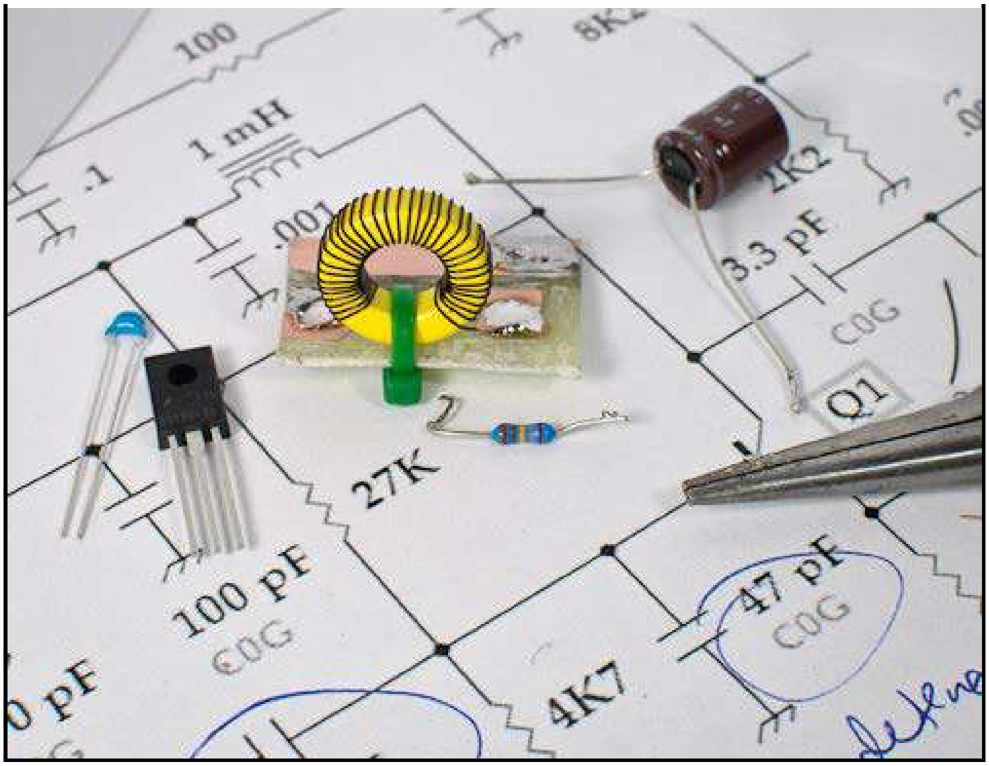
100
1 mH
.1
.001
8K2
2K2
3.3 pF
C0G
Q1
27K
100 pF
C0G
0 pF
47 pF
C0G
4K7

Above — In my various signal generator experiments, I zap strapped the toroid to a small piece of thick copper board that was later soldered to the main board. The L seems robustly secured.

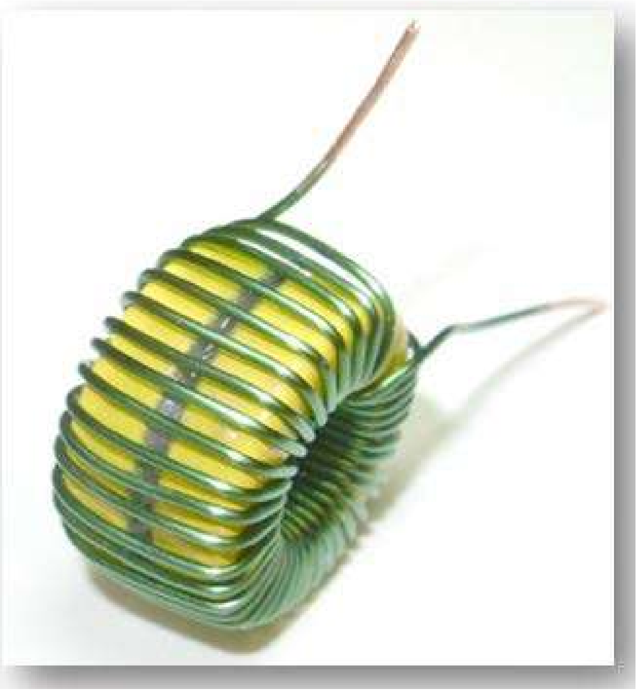

Above — A stacked toroid from the deleted VFO-2008 web page. I incorporated some of the information from the VFO-2008 page into this web page

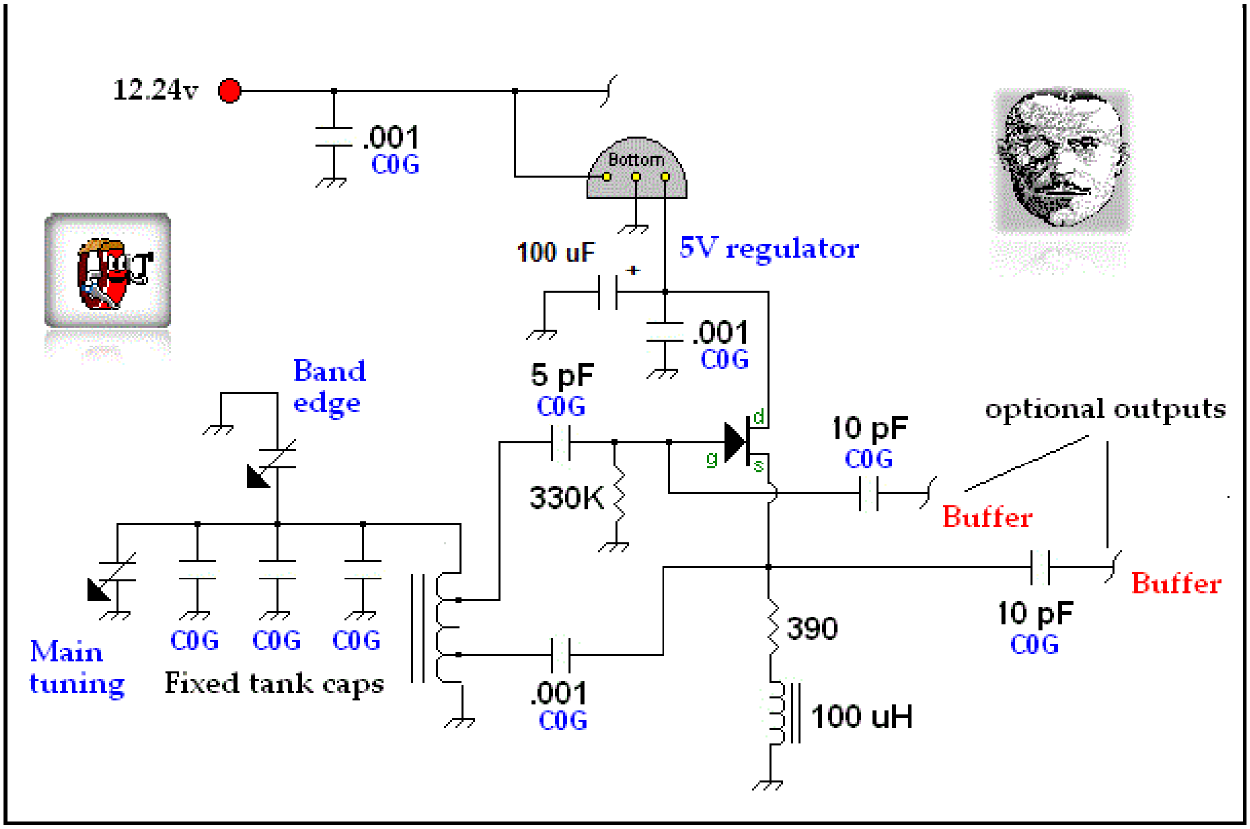
12.24v
.001
C0G
Bottom
100 uF
+
5V regulator
.001
C0G
Band
edge
5 pF
C0G
d
g
s
10 pF
C0G
optional outputs
Buffer
330K
Buffer
390
10 pF
C0G
Main
tuning
C0G
C0G
C0G
Fixed tank caps
.001
C0G
100 uH

Above — In order of preference, 3 ways to couple a Hartley oscillator to its buffer. From now on, I'll couple with a 1-3 turn link since it gave a lower distortion signal than with source or gate coupling. This figure omits the gate clamping diode seen earlier— tapping the inductor as shown keepings the FET gate AC voltage at a reasonable level when not using a gate clamping diode. Some builders leave off the gate clamp diode that clips positive signal peaks for lower phase noise. The diode acts as an AGC and offers benefit. Reverse biasing this diode was suggested by Dr. Ulrich Rohde: see — Key Components of Modern Receiver Design - Part 2: Dr. Ulrich Rohde, KA2WEU , QST for June 1994.

A formula to use for the inductor taps: Divide the total turns by 1.45 to get the first tap and by 7.25 to get the second tap (near ground link).

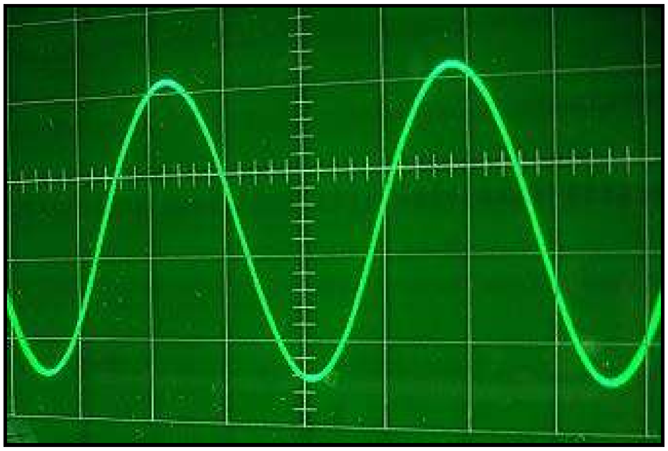

Enjoy your VFO experiments.

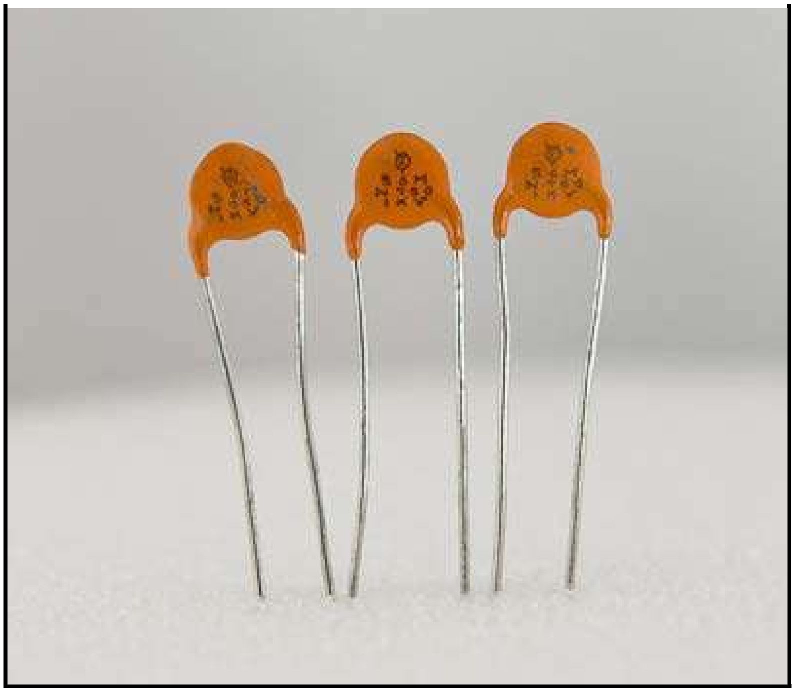

50 OHM
GAIN
HI
ON
HI-Z

RF — Test and Measurement

The Butler Did It ! - First VHF Experiments 2011

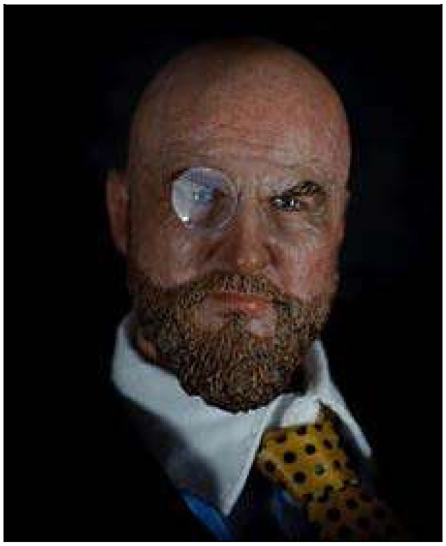

Venturing into VHF, I felt like a beginner with no experience or confidence — however, the excitement and allure of new parts and circuits kept me going.

After performing a literature review, talking to some colleagues and renovating my QRP work bench, my first tasks involved buying some VHF parts + exploring the Butler crystal oscillator. Specifically, I'll cover my experiences with the common base version of the Butler oscillator.

Like HF, the VHF knowledge base contains ever-present lore. Consider the Butler oscillator — I have read arguments stating that the emitter follower version of the Butler oscillattor is vastly superior to the common base version because the latter is prone to UHF and other spurs. These comments seem to have originated from a good book entitled Crystal Oscillator Circuits. *Revised Edition* by Robert J. Matthys published in 1992 by the Krieger Publishing Company.

While examining the schematics of professional/world class gear using a Butler, the common base version clearly dominates. Spectrum analysis and other measurements indicate that when common UHF oscillation management techniques are applied, common base Butler crystal oscillators work well.

Suppressing UHF oscillations with ferrite beads (and small value resistors), feedback, neutralization, limiting gain, etcetera are routine practices for us experimenters applying active devices that have strong gain into UHF on up. This is vanilla, or matter-of-fact construction for us; no worries. While fun and often convenient, lore ultimately stifles our progress.

Increasingly, I'm adopting the philosophy of Bob, K3NHI; "TMITK" — to measure is to know. Consider, too, you have to know what to measure and possess the required gear.

For BJTS, I purchased PN5179, MPSH10, 2SC3355, BF199, 2SC3357-RE, 2N5109/2N3866 and NE46134. The PNP part = MPSH81.

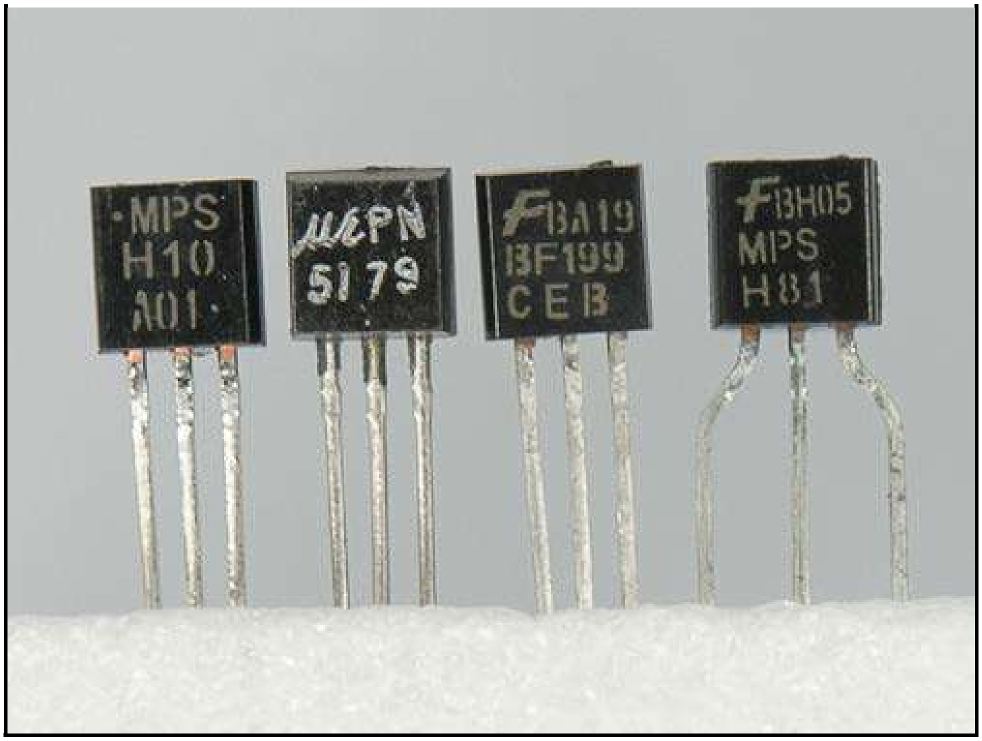

For JFETs, the J310 in TO-92 and SMT will remain my workhorse FET part along with a couple of other *low noise* JFETs and 2-gate MOSFETs. At VHF, the noise figure in a receiver chain is established by the first amplifier so a low noise preamp ranks important.

A collection of 100 volt NP0 capacitors ranging from 1 pF to 22 pF were added along with some chip and SMT caps as low as 0.5 pF. You might need a few air trimmer capacitors with a minimum capacitance ~ 2 pf — I applied 2 - 20 pF trimmer capacitors in most of the circuits that follow.

---

1. Experiments with a Butler Oscillator with a 23.3 MHz Fundamental Crystal

Fundamental Frequency

Long ago, I pulled a crystal marked 70.00000 MHz from a Drake Transmitter. I keep a fundamental oscillator based upon EMRFD Figure 4.23 on hand and verifed the fundamental frrequency at 23.3 MHz.

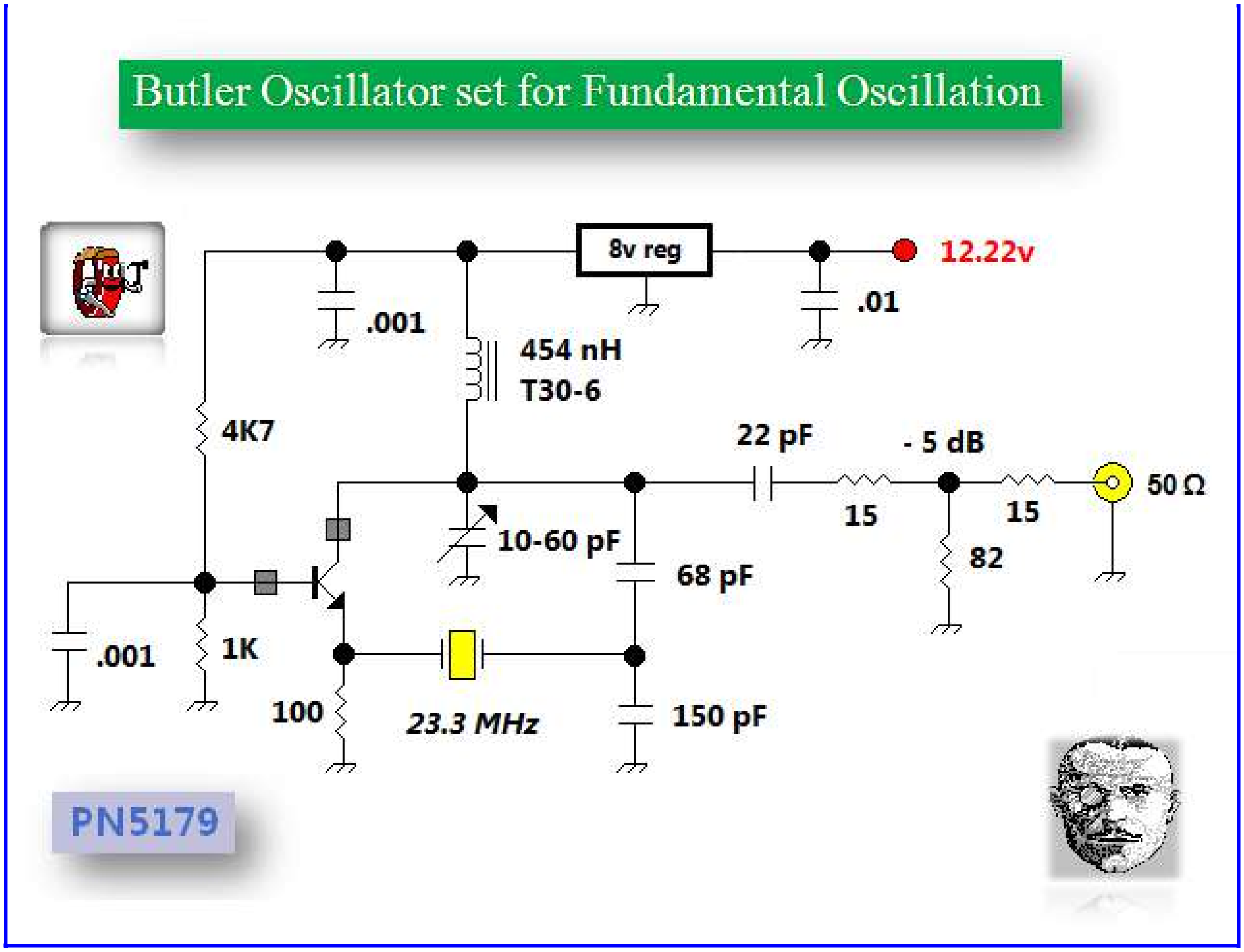
Butler Oscillator set for Fundamental Oscillation
8v reg
12.22v
.001
.01
454 nH
T30-6
4K7
22 pF
- 5 dB
15
15
50 Ω
10-60 pF
68 pF
82
.001
1K
100
23.3 MHz
150 pF
PN5179

Above — A Butler oscillator arranged for output at the fundamental crystal frequency. While commonly arranged as an overtone oscillator, the Butler is a good oscillator for any application. Consider, for example, EMRFD Figure 7.32. Wes applied the Butler at a 14 MHz fundamental because he wanted the lowest phase noise and IMD prone signal source possible. Tellingly, his buffer circuitry also conforms to this high standard.

My initial waveform looked distorted and prompted a solution. My experience yields that the L value needs to be adjusted for the best looking waveform in the Butler circuit. The inductor wire, wrapped around a T30-6 toroid was either scrunched to increase the inductance, or expanded to decrease the L while re-peaking the trimmer cap. Eventually, with patience, a beautiful sine wave emerged on my 'scope. I removed and measured the L with an ADE inductance meter. Consider all of my reported inductance values as nominal — gentle expansion or contraction of the inductor coils might be required to get an agreeable sine wave.

Click for the oscillocope tracing at 23.3 MHz.

The Butler at Overtone Frequencies

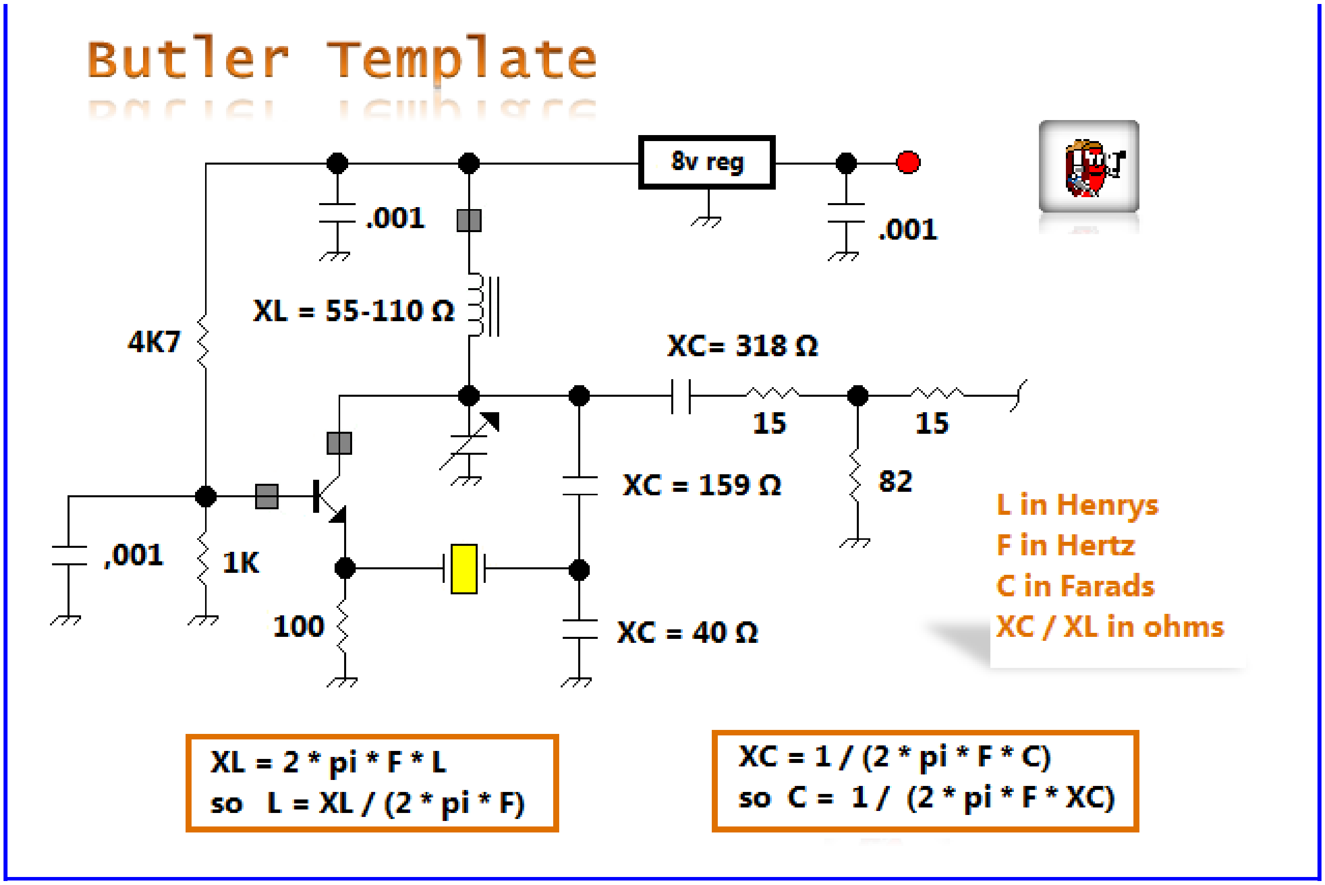
Butler Template
8v reg
.001
.001
XL = 55-110 Ω
4K7
XC= 318 Ω
15
15
XC = 159 Ω
82
,001
1K
100
XC = 40 Ω
L in Henrys
F in Hertz
C in Farads
XC / XL in ohms
XL = 2 * pi * F * L
so L = XL / (2 * pi * F)
XC = 1 / (2 * pi * F * C)
so C = 1 / (2 * pi * F * XC)

Above — A template (of sorts) for calculating Butler capacitor and coil values. The concept, rather than the absolute value matters most. I examined some well-designed Butler oscillators from professional equipment and determined their average XL and XC values. From the reactances shown, calculate the L and C values for the overtone frequency of interest with the 2 formulas in orange boxes. *Remember these XL or XC values just serve as starting values for experiments.*

For example : At 50 MHz with an XL of 108 Ω : L = 108 Ω / (6.28 * 50000000 Hz) = 0.000000344 H or 344 nH.

Fine tuning of the capacitor and inductor values might be required since factors including buffer input resistance +/- reactances, the overtone frequency and/or your breadboard layout may affect your Butler oscilator function.

In the experiments from the 3rd to 9th overtone frequency, the L = an air inductor wound with 21 to 22 AWG wire on a bolt.

Third Overtone Frequency

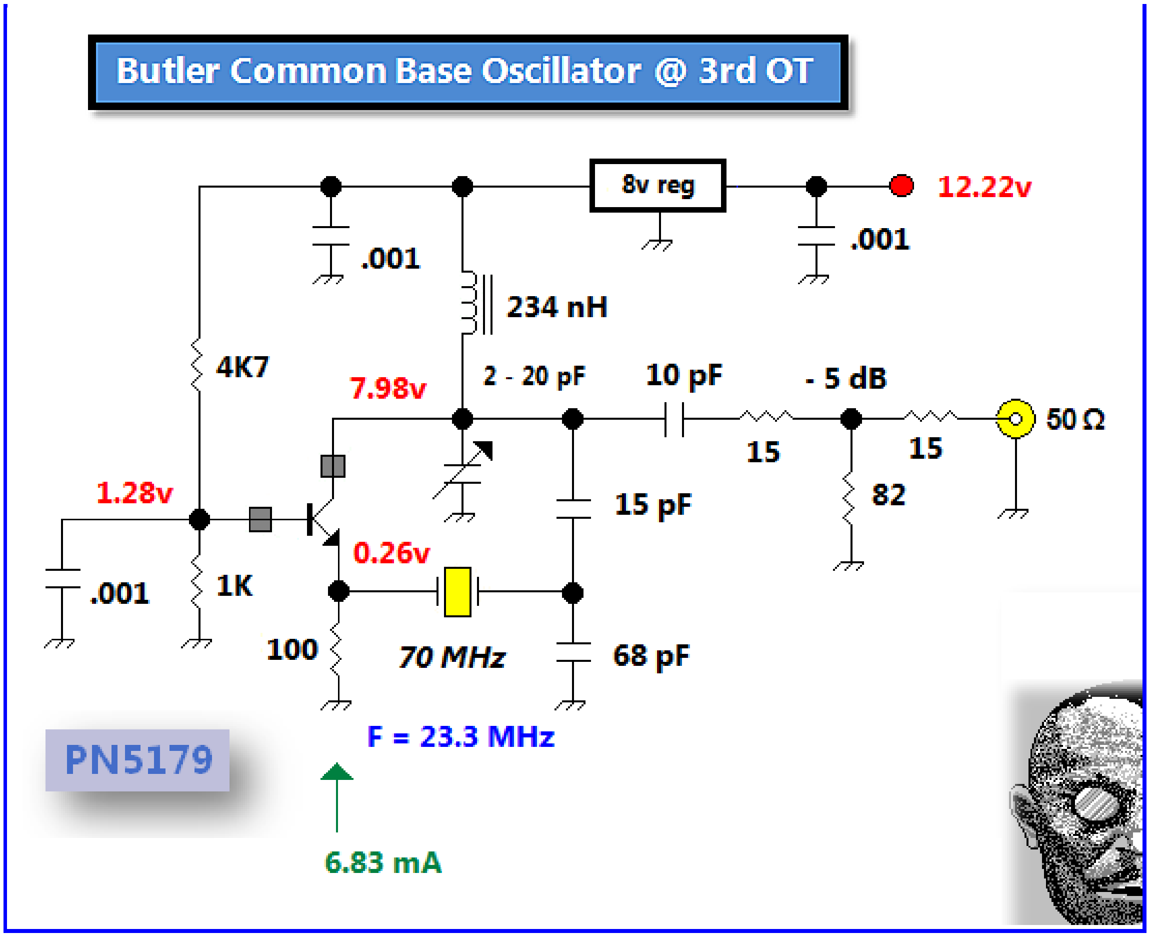
Butler Common Base Oscillator @ 3rd OT
8v reg
12.22v
.001
.001
234 nH
4K7
7.98v
2 - 20 pF
10 pF
- 5 dB
50 Ω
15
15
82
1.28v
15 pF
0.26v
.001
1K
100
70 MHz
68 pF
PN5179
F = 23.3 MHz
6.83 mA

Above — The Butler oscillator now arranged for output at the third overtone. Click and click for the 'scope outputs at the third crystal overtone. The first scope tracing was slightly mistuned. Once again, the inductor had to be gently squished or contracted to obtain a pristine sine wave.

Fifth Overtone Frequency

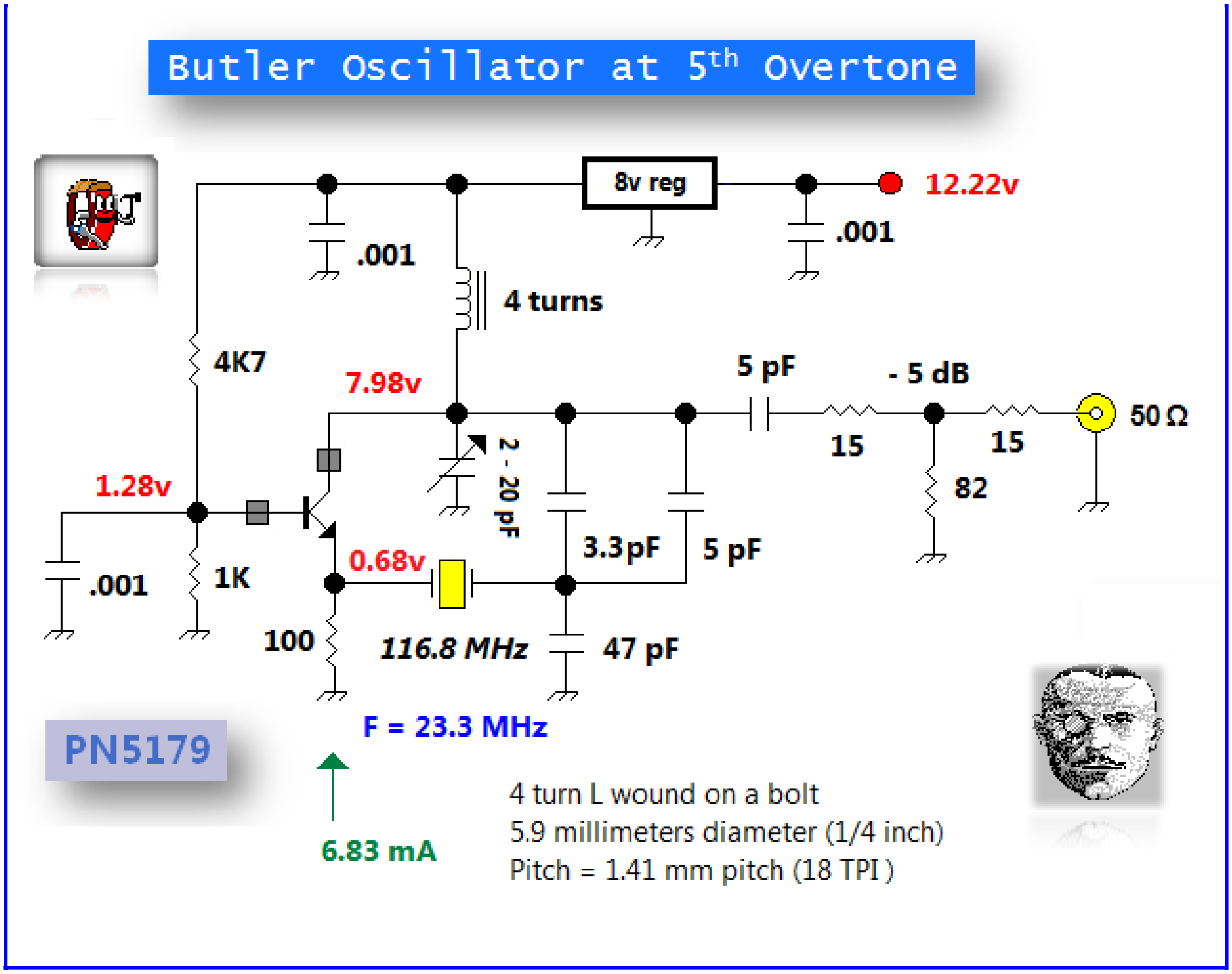
Butler Oscillator at 5th Overtone
8v reg
12.22v
.001
.001
4 turns
4K7
7.98v
5 pF
- 5 dB
50 Ω
15
15
82
1.28v
2 - 20 pF
3.3pF
5 pF
.001
1K
0.68v
100
116.8 MHz
47 pF
PN5179
F = 23.3 MHz
6.83 mA
4 turn L wound on a bolt
5.9 millimeters diameter (1/4 inch)
Pitch = 1.41 mm pitch (18 TPI)

Above — The Butler oscillator at the 23.3 Mhz crystal's 5th overtone. Here's the 'scope tracing and an Ugly Construction photograph. The output = 7 dBm — very cool.

Some authors directly connect the attenuated output to the Local Oscillator port of a diode ring mixer. The signal is adjusted to the desired 7 dBm power by tweaking the 50 Ω pad attenuation, and/or the NPN's current. A good example = Single-Conversion Microwave SSB/CW Transceivers by Rick Campbell in QST for May, 1993.

The circuit above was measured with a 50 Ω terminated oscilloscope however, I also tested it with a 10X probe attached to termination resistors from 51 to 1 Meg ohms. When changing the 51 ohm termination resistor to a higher value such as as 47K, a previously working Butler may stop oscillating. The buffer input impedance and capacitance greatly affected the oscillator in my experiments.

When I wired up a Butler, 1 of 3 things happened: it did not oscillate, it gave a distorted output waveform, or it wowed me with a nice sine wave. Tuning the L-C tank is critical + finicky and may test your patience.

Adding an Inductor Across the Crystal

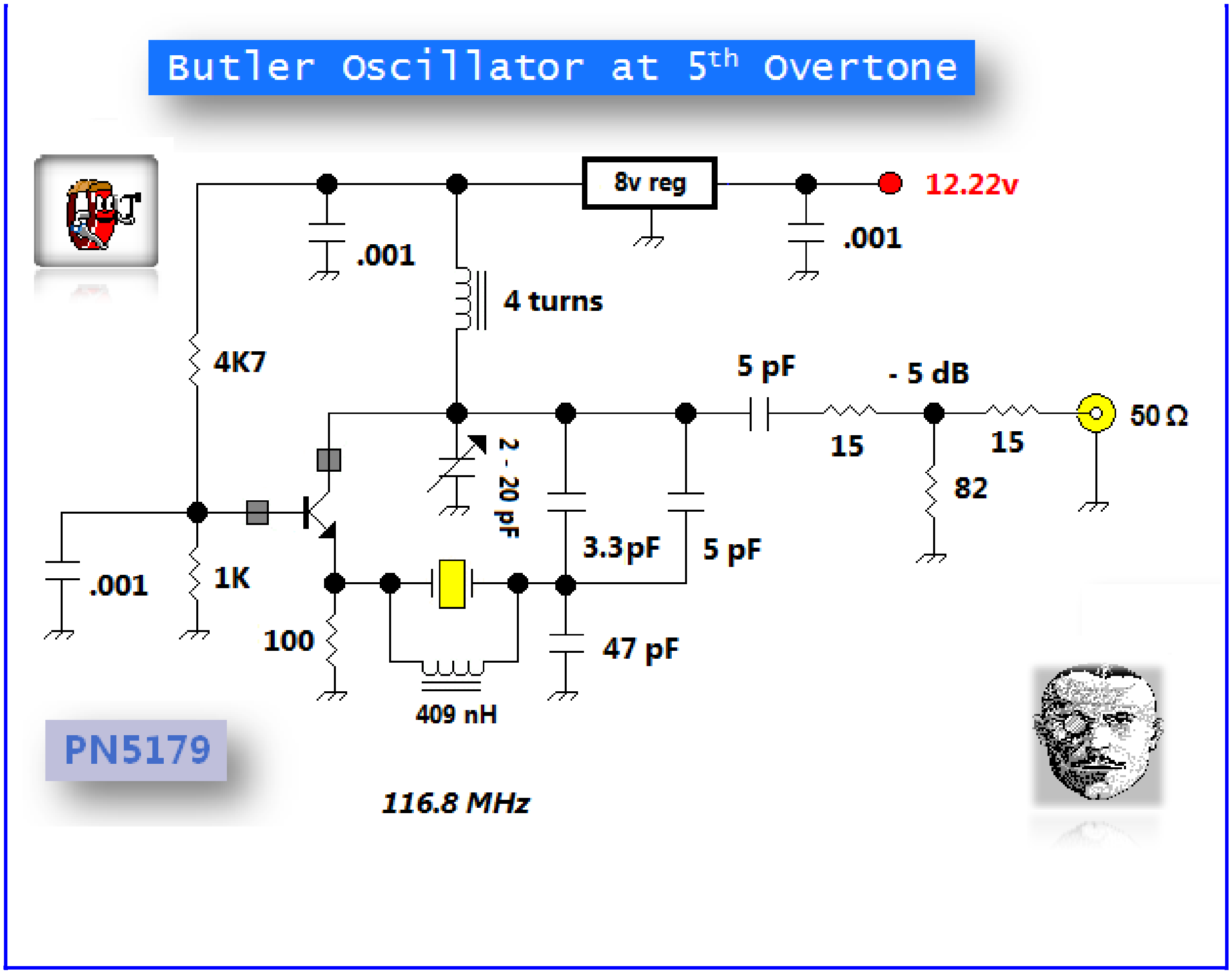
Butler Oscillator at 5th Overtone
8v reg
12.22v
.001
.001
4 turns
4K7
5 pF
- 5 dB
15
15
82
50 Ω
2 - 20 pF
3.3pF
5 pF
.001
1K
100
47 pF
409 nH
PN5179
116.8 MHz

Above — The 5th overtone Butler with an inductor in parallel with the crystal

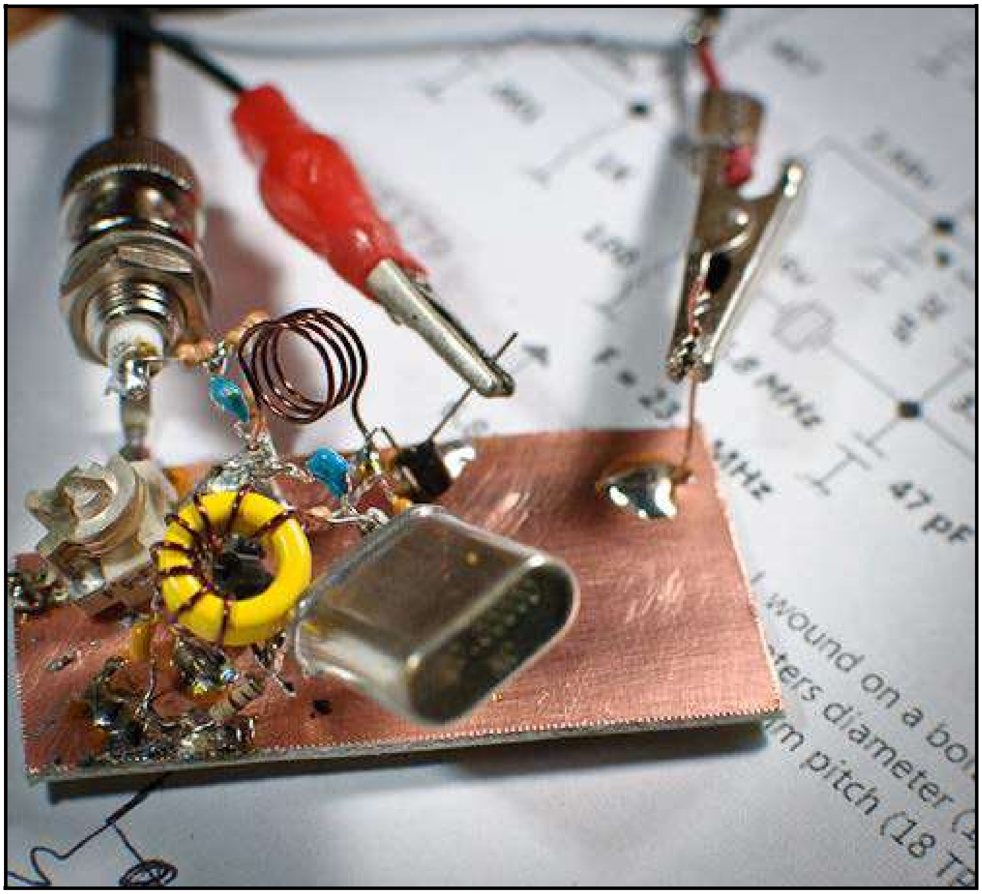
47 pF
wound on a bo
ters diameter
m pitch (18 T

Above — My 116.8 MHz oscillator breadboard with an inductor wound on a T50-6 with wide spacing to allow scrunching and expanding of the windings across the crystal. I roughly determined my L should be ~ 400 nH and wound this on a # 6 toroid. While observing the output in my scope, I scrunched and then expanded its windings and adjusted the trimmer capacitor. The goal was to find a clean signal that snuffed out immediately when the trimmer cap was tuned off resonance. After finding the optimal L, I later removed and measured the coil. There is very little "wiggle room" — the oscillator tunes up and then dies very sharply as you tweak the trimmer cap. No sidebands were observed.

The inductor across the crystal is optional — some suggest it might only be needed above ~ 70 MHz.

At frequencies above ~70 MHz, the parallel capacitance of the xtal (C par) approaches the internal series resistance of the xtal and this provides an alternate path around the crystal for the signal and may short-circuit the crystal. The parallel inductor resonates with the crystal's parallel capacitance and tunes it out, so the crystal remains unbypassed.

Our teacher, Wes, W7ZOI published a document covering the Butler parallel crystal inductor here.

Simply put — the inductor allows clean tuning and output. That is, when you tune the trimmer to one side or another, the oscillator just dies and doesn't produce the sidebands that are shown in Wes' web article.

Since many of us choose computer, or other surplus crystals, a high C par + low Q crystal might give you tuning woes depending on your overtone frequency. In this case, adding the L to your circuit may improve tuning and ward off any unwanted sidebands.

With my particular crystal the parallel inductor is not needed, however, I can report that even slight mistuning just snuffed out my oscillator with the added inductor.

Seventh and Ninth Overtone Frequency

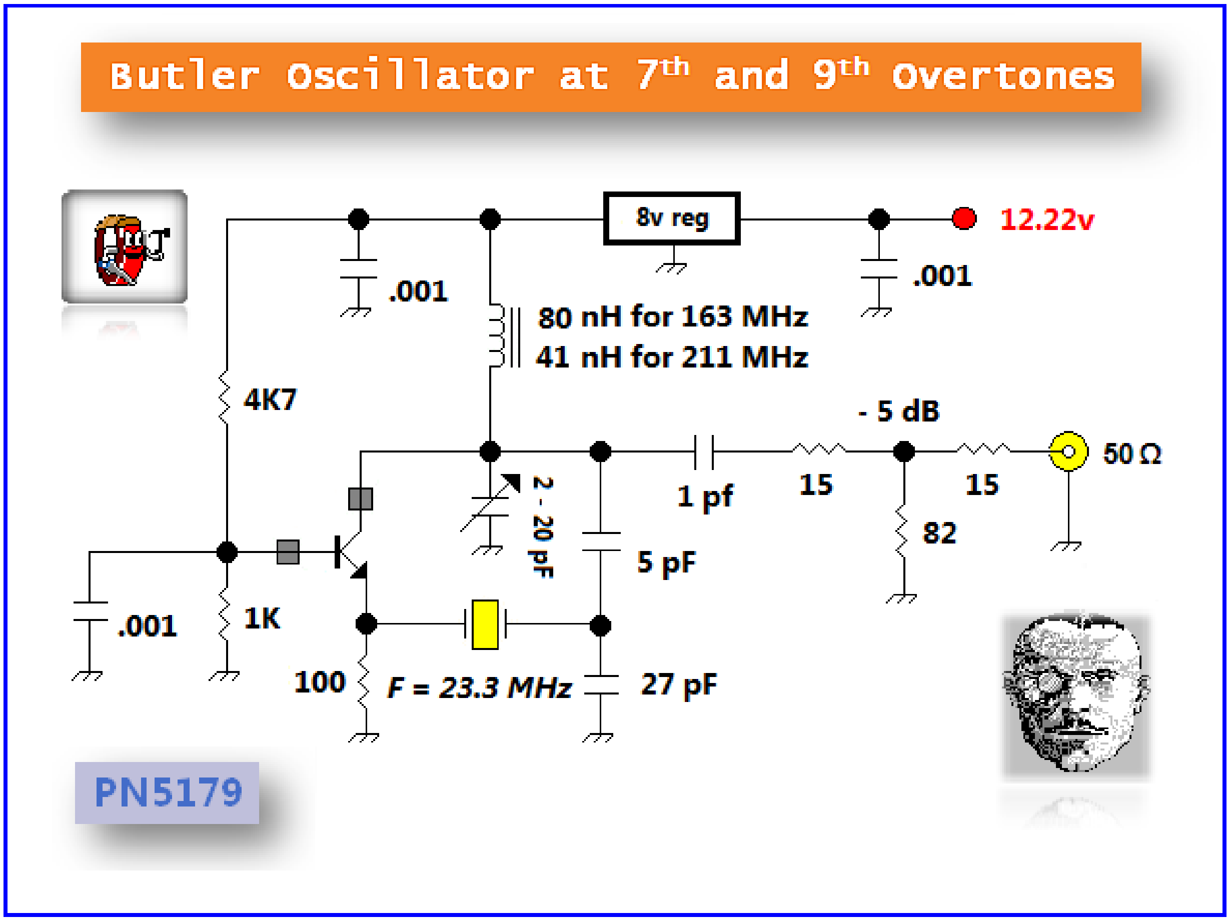

Above — I decided to take the Butler up to the 7th and 9th overtones. The 1 pF coupling capacitor proved the most critical part; for example, if I raised it to 5 pF, the oscillator would not tune above the 5th overtone. I wound the inductors on bolts and then compressed or stretched the links to get the perfect inductance. At these frequencies, stray inductance becomes quite significant and my coils were 1 to 1.5 turns less than indicated on a spreadsheet coil inductance calculator. Click for the seventh overtone 'scope shot. Click for the ninth OT 'scope tracing.

I'm confident that if my 'scope bandwidth was higher, I could have resonated the 11th overtone.

---

2. Butler Oscillator-based 50 MHz Signal Generator

Lacking a 6 Meter band signal generator, I decided to build a 1 frequency device employing a Butler oscillator. Some may laugh at a 1 frequency signal generator — I won't since I'll use it to design and align amplifers, filters, a new 6 Meter band VCO and measure scattering parameters. Besides, you can double, triple, VXO or mix single frequency generators with another variable oscillator — this web site has roots in humble, simple test equipment.

In my bag, I found a crystal labelled 50.0000 MHz and measured its fundamental at 16.67 MHz in a simple Colpitt's oscillator — perfect . Design goals included variable amplitude, strong reverse isolation + reasonable shielding and return loss. Since, I lack another 6 Meter band signal generator I could not measure return loss, however, choosing proven circuits + a terminal attenuator pad will help.

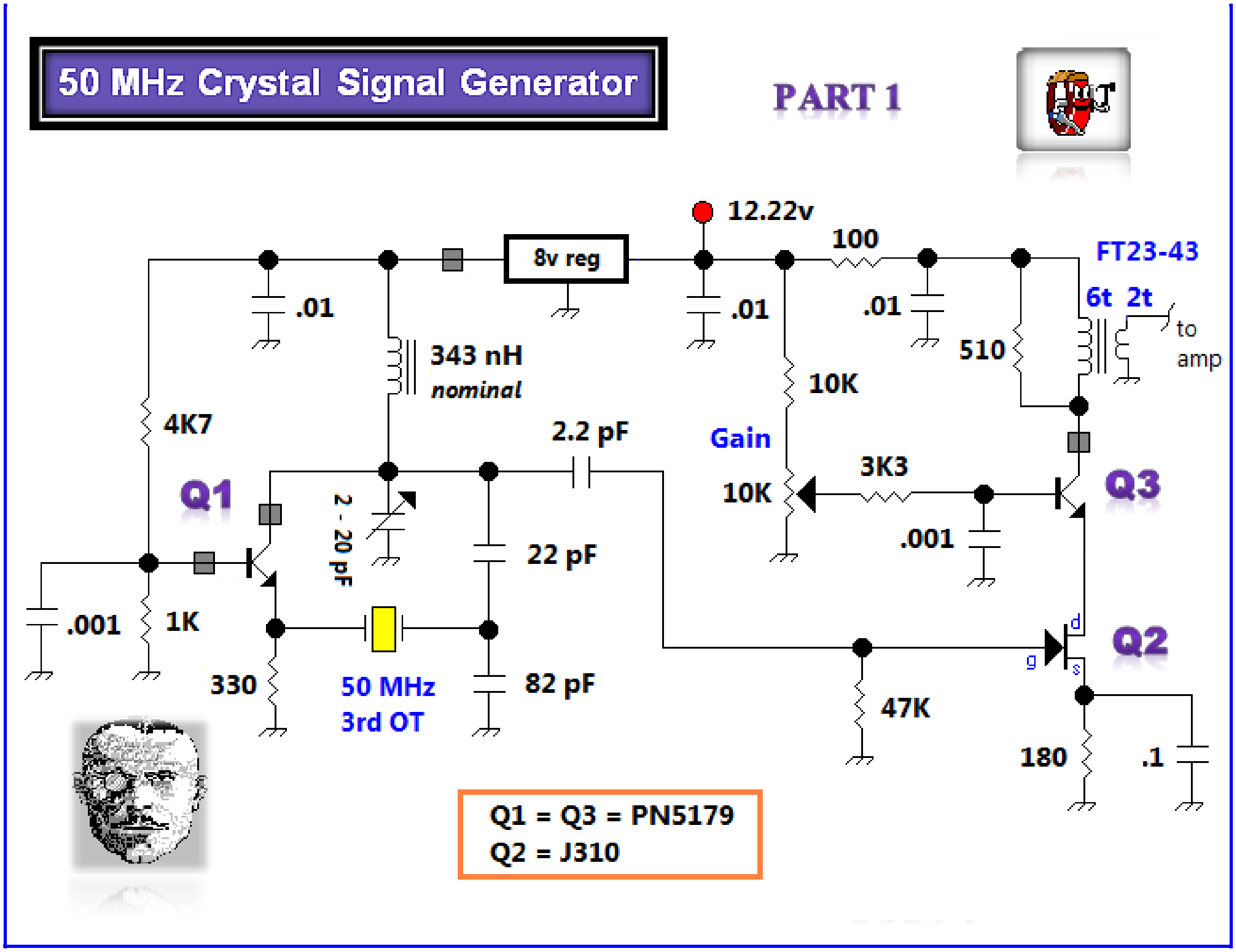
50 MHz Crystal Signal Generator
PART 1
12.22v
8v reg
.01
.01
100
.01
FT23-43
6t 2t
to
amp
510
343 nH
nominal
10K
4K7
2.2 pF
Gain
Q1
2 - 20 pF
3K3
10K
Q3
.001
22 pF
.001
1K
d
Q2
g
s
330
50 MHz
3rd OT
82 pF
47K
180
.1
Q1 = Q3 = PN5179
Q2 = J310

Above — Butler oscillator + hycas buffer. The 22 pF cap in the Butler was originally 15 pF, but when tested with various coils, oscillations proved a litle sluggish, so the 22 pF was substituted. I experimented with the coil, but ran out of time, so I wound a few turns of wire on a T50-10 toroid and soldered it in. After some careful manipulation of the windings, a glorious sine wave arose. I removed the L, measured it at 343 nH and then re-soldered it in place. An air coil or other size 6 or 10 powdered iron inductor should work fine.

Click for a moderate resolution photograph of the entire project.

50 MHz Crystal Signal Generator

PART 2

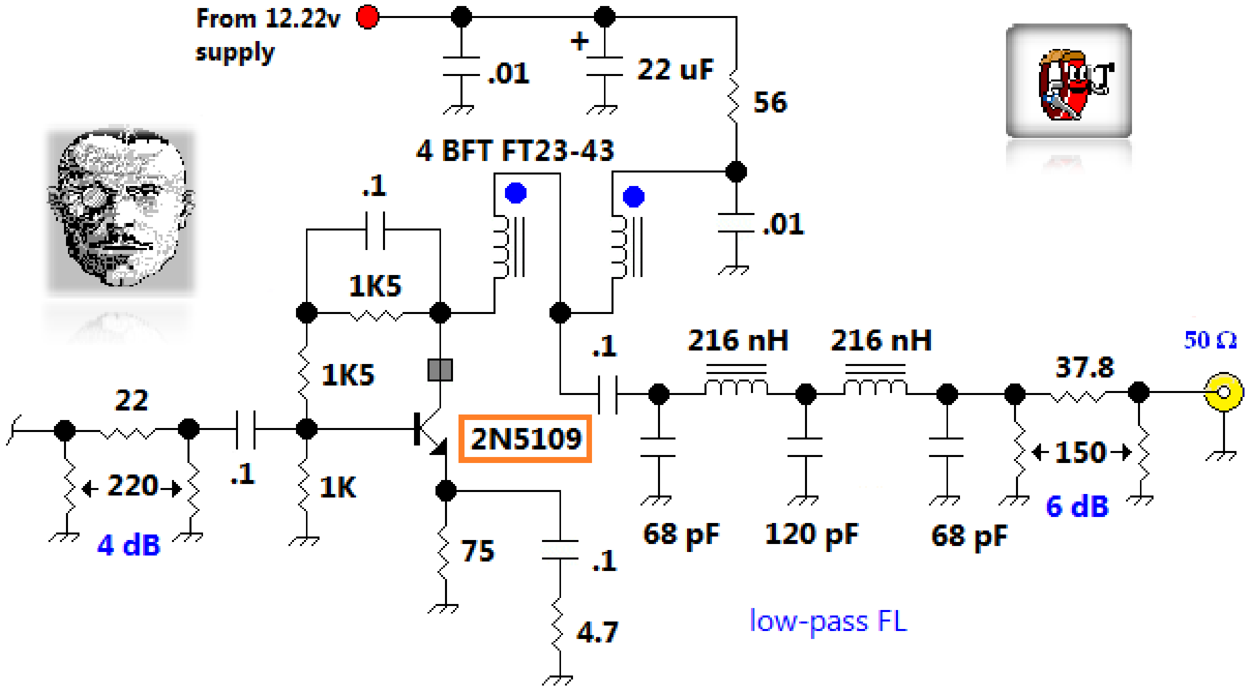
From 12.22v supply
.01
+
22 uF
56
4 BFT FT23-43
.01
.1
1K5
1K5
22
220
4 dB
.1
1K
2N5109
75
.1
4.7
.1
216 nH
216 nH
37.8
50 Ω
68 pF
120 pF
68 pF
150
6 dB
low-pass FL

Above — The final amplifier and low-pass filter. The maximum output of the Q2/Q3 amp is hot and can overdrive the 2N5109, therefore a 4 dB pad was added. This pad also improves the input return loss of the 2N5109. An alternate technique might be to further reduce the maximal DC bias voltage into the Q3 base and remove the 4 dB pad.

Click for the minimum and maximum output voltage waveforms from my 'scope

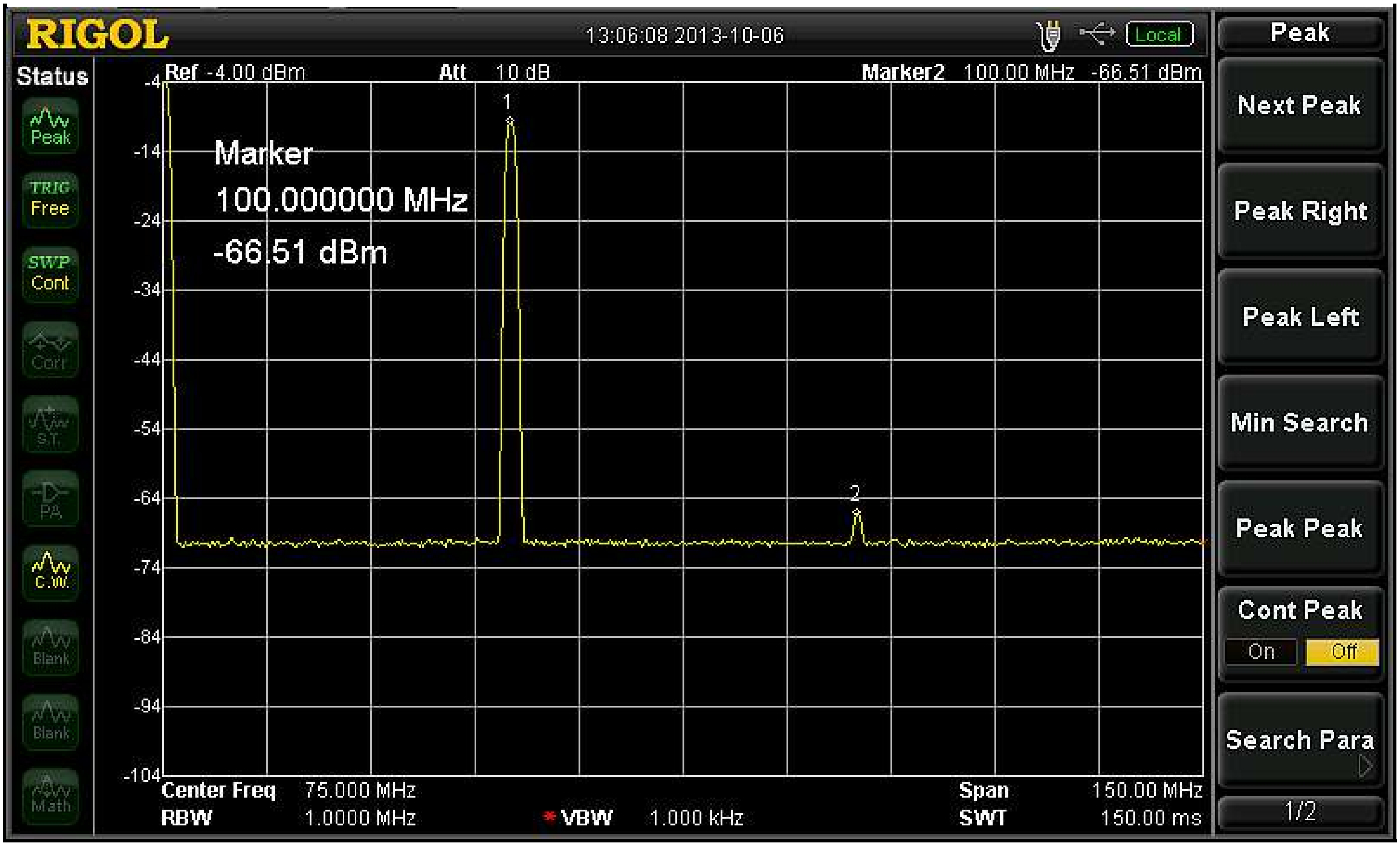
13:06:08 2013-10-06

Ref -4.00 dBm   Att 10 dB   Marker2 100.00 MHz -66.51 dBm

Marker
100.000000 MHz
-66.51 dBm

Center Freq 75.000 MHz   Span 150.00 MHz
RBW 1.0000 MHz   VBW 1.000 kHz   SWT 150.00 ms

RIGOL
Local
Status
Peak
TRIG Free
SWP Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
-4
-14
-24
-34
-44
-54
-64
-74
-84
-94
-104
1
2
Peak
Next Peak
Peak Right
Peak Left
Min Search
Peak Peak
Cont Peak
On
Off
Search Para
1/2

Above — Spectrum analysis with the generator output at -10 dBm. I placed Marker # 2 on the 2nd harmonic peak (100 MHz) and measured the power. The low-pass filter only attenuates the 2nd harmonic by ~ 25 dB, however, with the nice sine wave from the Butler, the 2nd harmonic measured -66.51 dBm = -56.51 dBc.

The 37.8 resistance indicated in the pad = 2 resistors in series. Many just use a 37 or 39 Ω R.

I look forward to advancing my VHF skills with this little signal generator.

50 MHZ XTAL OSCILLATOR
PART 1
Generator
FT23-43
6t
2t
to amp
12.22v
100
.1
510
Q3
3K3
.001
Q2
47K
180
.001

3. Doubler for the 50 MHz Signal Generator

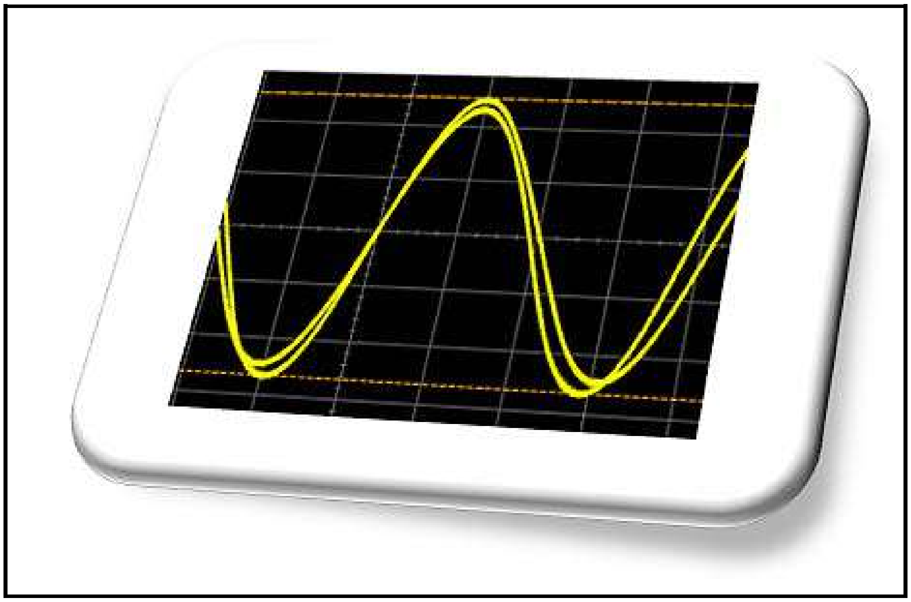

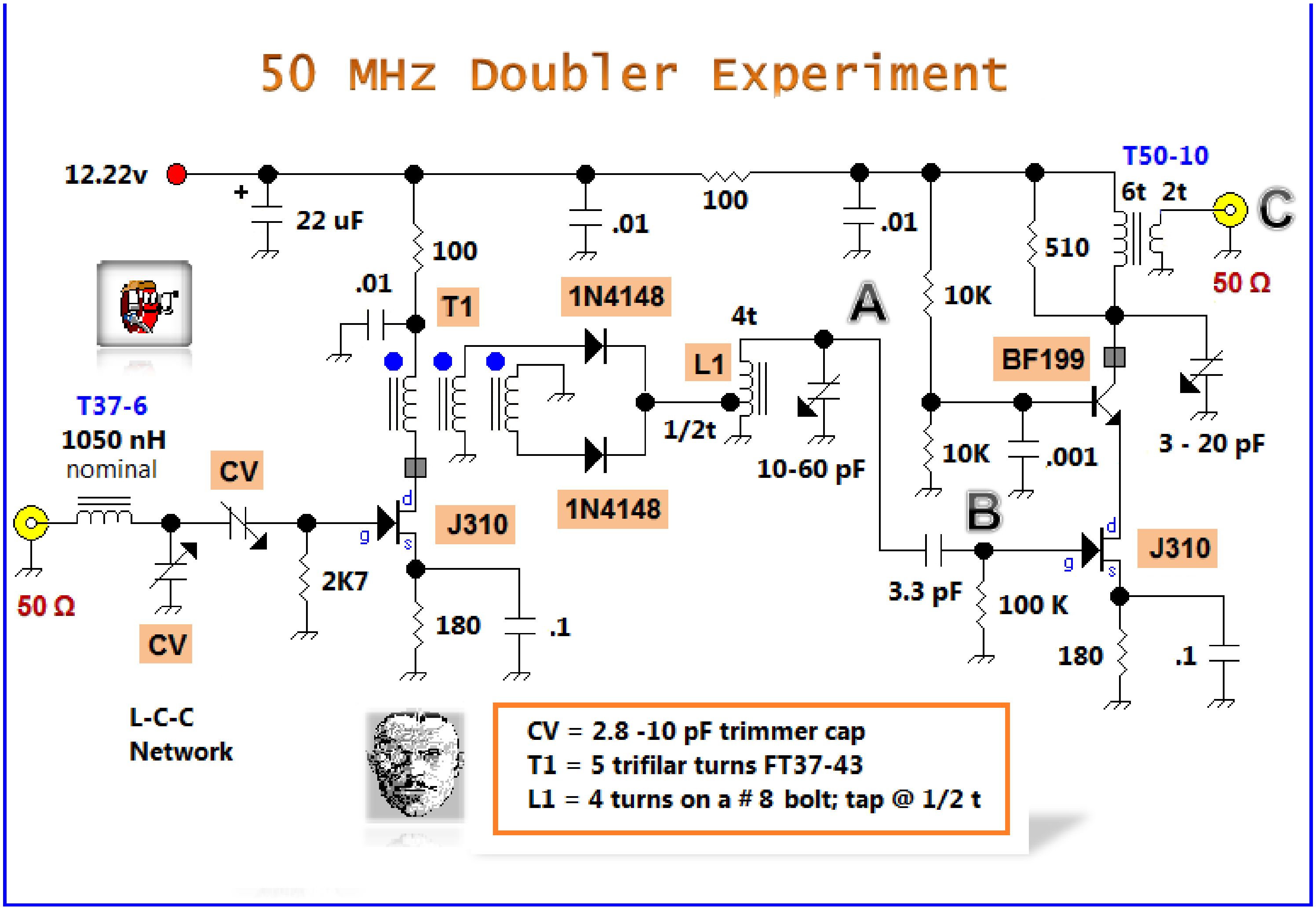
50 MHz Doubler Experiment
12.22v
+
22 uF
100
.01
100
.01
T50-10
6t 2t
C
50 Ω
510
10K
A
4t
1N4148
T1
.01
L1
1/2t
10-60 pF
BF199
3 - 20 pF
10K
.001
B
T37-6
1050 nH
nominal
CV
50 Ω
CV
d
g
s
J310
1N4148
2K7
180
.1
3.3 pF
100 K
d
g
s
J310
180
.1
L-C-C
Network
CV = 2.8 -10 pF trimmer cap
T1 = 5 trifilar turns FT37-43
L1 = 4 turns on a # 8 bolt; tap @ 1/2 t

Above — An experimental frequency doubler for the Butler Oscillator-based 50 MHz Signal Generator shown above.

After trying a few circuits, I settled on a simple full-wave doubler featuring matched 1N4148 diodes driven by a JFET amplifier. Chapter 3 of Solid State Design for the Radio Amateur by Hayward and DeMaw for the ARRL = my key reference. Matched diodes (and a little luck), may suppress the 50 MHz signal up to 60 dB so only a single-tuned circuit follows the diodes.

The first JFET amp drives the diodes to improve harmonics + output voltage — an L-C-C Tee network matches the 2K7 Ω input to the 50 MHz signal generator output. I designed this matching network on the bench with the 2 diodes disconnected to avoid distortion during signal measurements. Using a 10X probe, I peaked the capactors for the greatest signal amplitude after finding the optimal L by educated trial and error. The inductor wound on a T37-6 was scrunched a little to further peak the L-C-C network.

I wound L1 with bare copper 26 AWG wire on a #8 bolt with coarse threads. To make enough space to solder on the 1/2 turn tap, I stretched the last turn with the other 3 turns still on the bolt to prevent stretching these coils. Click for a 'scope tracing at Point A. My 10X scope probe has ~ 15 pF capacitance and this affected the tuning — you can see some harmonics in the signal. With the 10X probe at Point B, I was nearly able to turn the circuit to 100.0 mHz, but still the 10:1 probe upsets the circuit somewhat.

This L - C tank tunes sharply and best with a non-conductive screw driver. I final tuned the L1 tank when the hycas amp was completed and connected to a 50 Ω terminated 'scope — eliminating the earlier tuning problems caused by the 10X probe. It seems that VHF requires more thought and care than HF when tuning resonators (tanks).

Point C is the maximum output (10.34 dBm) into a 50 ohm terminated 'scope. The 100.0 MHz signal is okay, but some builders might want to add low-pass filtering +/- an attenuator pad; or perhaps drive another feedback amp?

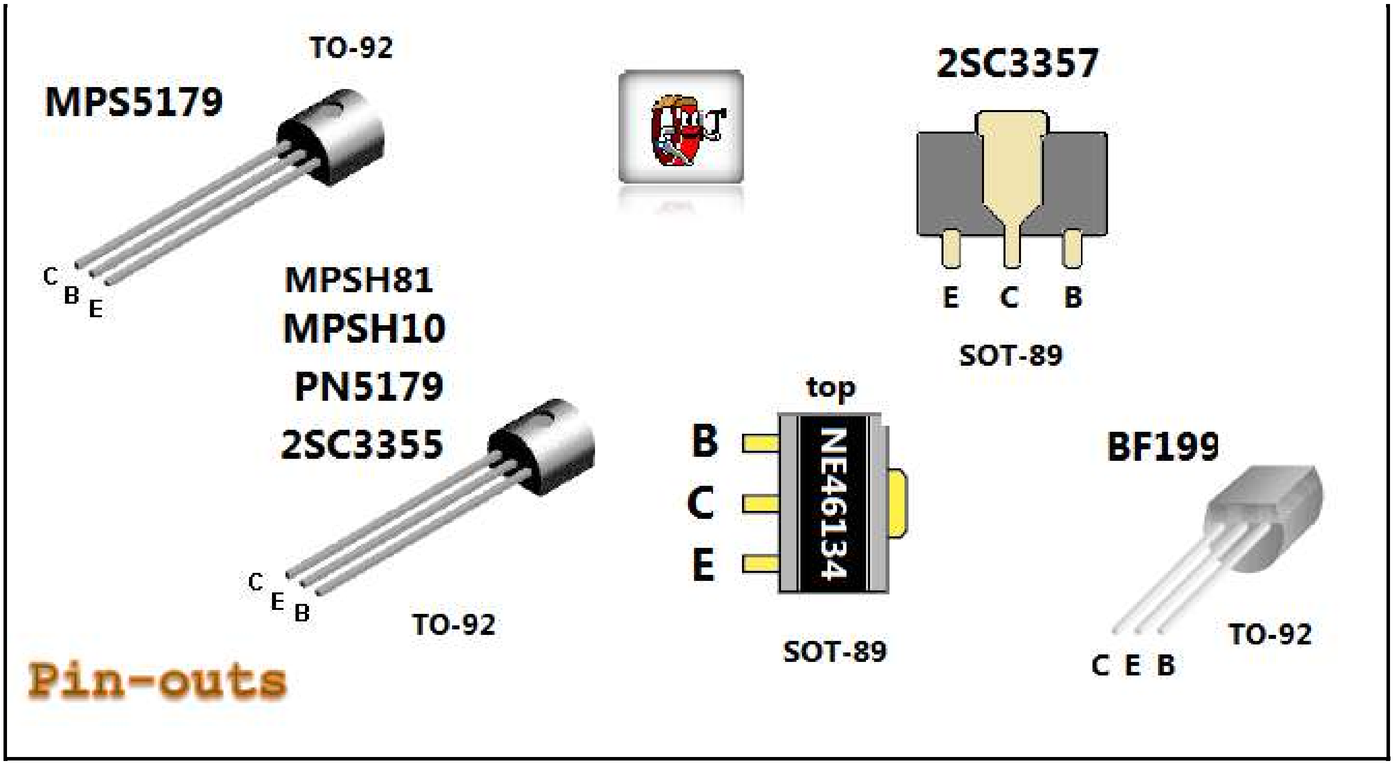
TO-92
MPS5179
C
B
E
2SC3357
E
C
B
SOT-89
MPSH81
MPSH10
PN5179
2SC3355
C
E
B
TO-92
top
B
C
E
NE46134
SOT-89
BF199
C
E
B
TO-92
Pin-outs

RF — Test and Measurement

VHF to the Max — Miscellaneous 2012 Experiments

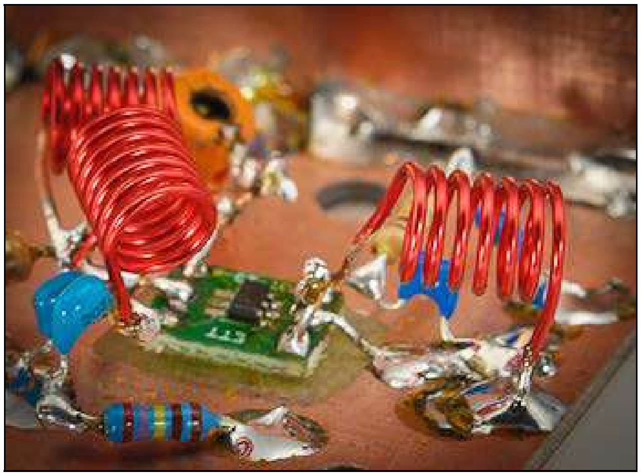

As a VHF newcomer, I need to make lots of circuits, measurements, mistakes and maybe — I might advance. My literature review revealed a big gap between popular, "for-fun", novelty-grade projects and the blinged-out circuits such as ultra low-noise LNA's featuring GaAs, MESFET, and pHEMT devices. Where do we find the middle ground projects? Likely on our own RF work benches.

Numerous questions arose — Will Ugly Constuction work? Do I have the right test gear? Am I measuring the proper things? What about noise figure? All a bit overwhelming — but even improbable discoveries began as simple questions and observations. We solve far more difficult problems each day.

I'm learning that whether your a beginner, or a master, bench fundamentals rank supreme. Want to avoid oscillations in your high fT BJT VHF amplifier? Work towards high reverse isolation, strong S11 and S22, careful layout and employ techniques that suppress instability — nothing earth shattering huh? I'm told that acquiring the needed skills takes time and practice.

In early Winter 2011-2012, I built some VHF signal generators, amplifiers and a lots of junk. This web page documents a few of these experiments.

Section 1. Early MAX2606 VCO experiments
Section 2: A 2-band signal generator based upon the MAX2606 VCO
Section 3: 50 MHz VCO
Section 4: 50 MHz Receiver Pre-amp and Filter
Section 5: QRP — POSDATA: Z-Communications VCO Experiments

Section 1. Early MAX2606 Experiments

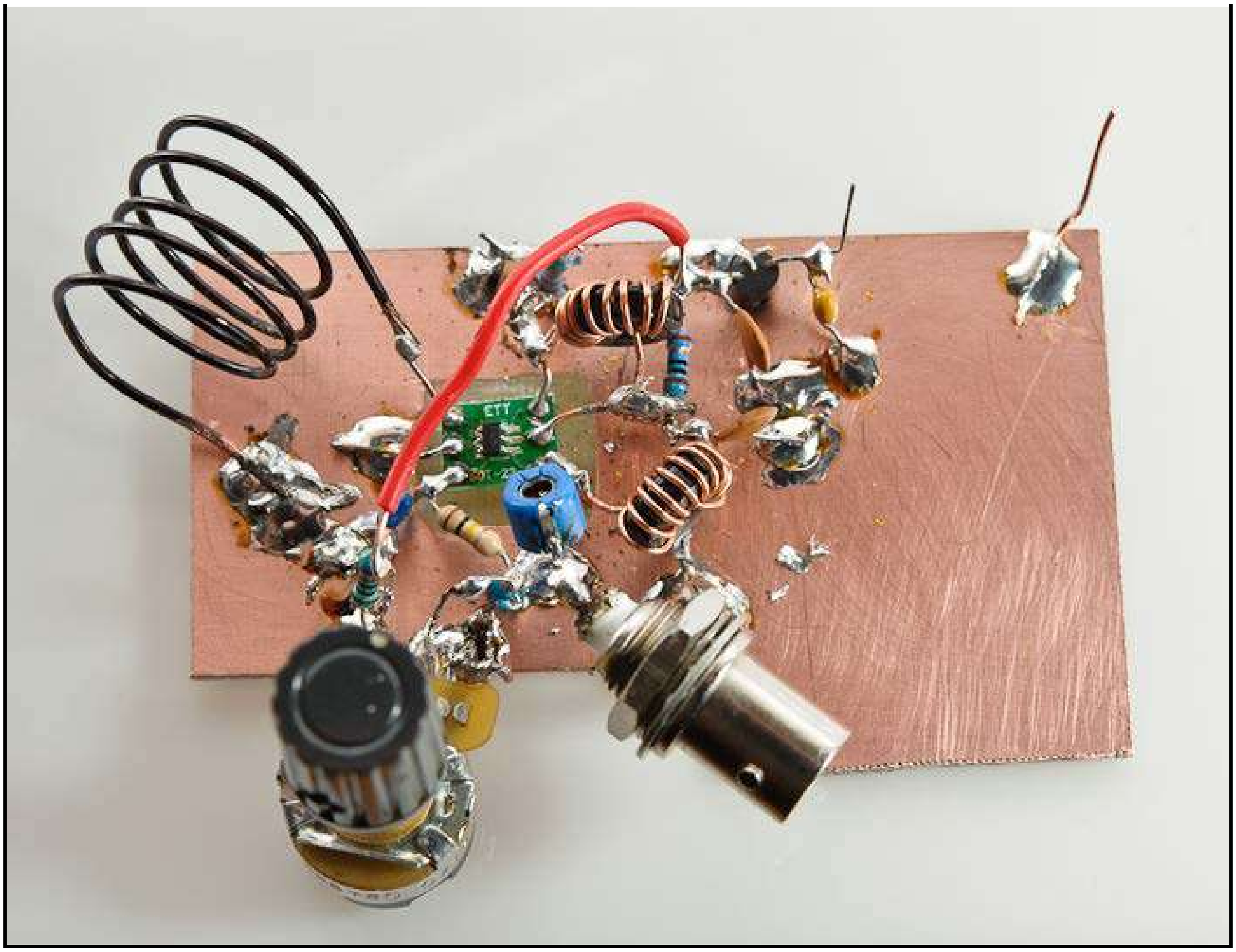

Above — My first of many MAX2606 VCO Experimental breadboards.

Maxim makes a series of cool SOT23-6 VCOs ranging from 45 to 650 MHz. Wanting a VCO to cover from ~100 to 106 MHz, I just had to try this chip. The datasheet provided all the online information I could find; save for a few homebrew, flea-powered FM transmitter projects that proved unhelpful.

After soldering the IC on a breakout (prototype) board using about 20X magnification, the proto-board was dropped onto a copper clad board with some of the copper ground away.

Max2606 VCO Experiments

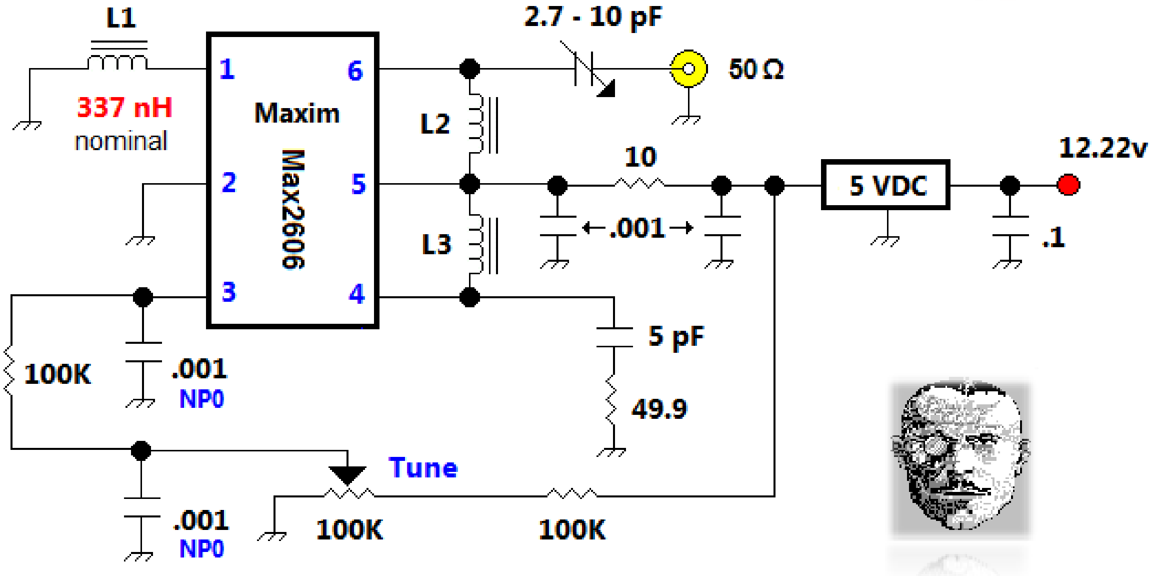

L1 = 5 turns on a 5/8 inch bolt (15.9 mm)
L2 = L3 = 411 nH

Above — My first MAX2606 VCO experiment. A coil lying on the workbench was soldered in as L1. I stretched it a little to set the lower band edge. Going from CCW to CW on the 100K tuning pot, I measured from 86.3 — 106.3 MHz. I'll write about L1 a little later, for making a good output network consumed my initial experiments.

The Maxim datasheet shows a simple output network consisting of 2 pull-up resistors with a maximum R of 1K. All the online FM transmitter designs I saw used 1K pull-up resistors and made no attempt at matching or even employing inductors in place of the resistors. I applied two 560 ohm pull-up resistors in place of L2 and L3 and the output looked distorted and low in amplitude. Click for a 50 Ω terminated 'scope tracing and click for a tracing with a 10X probe connected to a 10K load resistor. Unacceptable for even us scratch-builder RF experimenters.

Wanting to match the output into a 50 Ω load with high-pass L-network instead of just resistors, I calculated that an L of 250 to 800 nH, plus a capacitor from 2 to 10 pF might work, however, as a VHF newbie, my hopes weren't high. I went for a single-ended output and not knowing what to do, terminated the unused port with an AC-coupled 49.9 Ω resistor.

Choosing a 100 MHz test frequency (where I own a doubled 50 MHz xtal oscillator), I removed L1 to snuff the oscillator and measured an output return loss of ~ 17 dB with L2 = L3 = 298 nH and the variable cap set to 7.3 pF (I removed it and measured it). It took about 1 hour of trying different L values to get this return loss. I also learned that the L value and L to C ratio of L2 and its series capacitor greatly affects the purity of the output waveform.

Like all L-networks, the L and C *must* be correct to obtain a strong, clean output. Since C is variable, I tried various inductor values and further fine-tuned them by scrunching or expanding the L2 windings. *Compress or expand the L2 windings + trim the variable capacitor to peak the L-network at your desired frequency.*

From my experiments, the L3 inductance should be close to L2 for the best output voltage and symmetry, however, don't bother scrunching or expanding L3 because this won't greatly affect the output signal as long as the L3 inductance is close to L2.

Click or click for some 'scope captures of poorly matched output L-networks — low gain and/or distortion appeared with mistuning.

Click for the output where L2 = L3 = 298 nH. I had the tuning set to 100.6 MHz for this screen capture, however, you can see the nice waveform and strong peak-peak voltage when compared to the pull-up resistor only versions.

The problem with a high Q L- network = low bandwidth. I wound L2 and L3 on T37-10 powdered iron toroids and later tried T37-6 toroids. The higher Q of the number 10 material seemed to translate into higher output voltage, but narrower bandwidth over the number 6 material toroidal inductors. Alternate breadboard. When tuning the MAX2606, signal amplitude changes with tuning frequency and a single L-network peaked somewhere near the middle of the tuning range further worsens this amplitude issue.

The *unbuffered* VCO waveform distorts somewhat as you move farther away from the peak L-network frequency. After trial and error, I settled in a compromise of L2 = L3 = 411 nH. This gave the best overall signal purity + amplitude from about 98 to 106 MHz with a set peak at ~ 102 MHz. Tuning below 98 MHz tended to really distort the waveform. Click for the *unbuffered* minimum and maximum frequency 'scope tracings.

Placing an attenuator pad plus a buffer amp after the VCO dramatically reduced this distortion in my experiments, assuming the L-network components are correct and tuned.

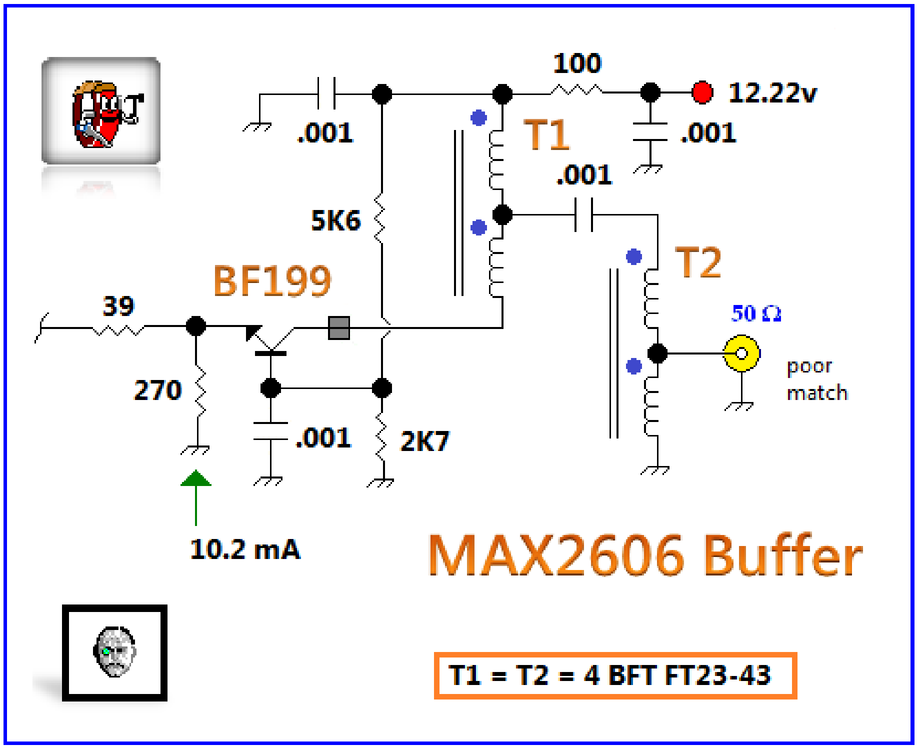

Above — A trial buffer first connected to output of the VCO schematic above. Two series-connected 2 transmission line transformers dropped the output impedance down for examination in my 50 Ω terminated scope — unfortunately the output return loss was poor (less than 10 Ω). Unlike the common base stage in a hybrid cascode amp, varying the bias to adjust gain worked poorly and lower applied bias caused distortion. I found it preferable to just fix the bias to get the greatest voltage gain + lowest harmonic distortion.

This buffer was discarded since the return loss was too low for my needs.

Testing a Hybrid-Cascode Buffer

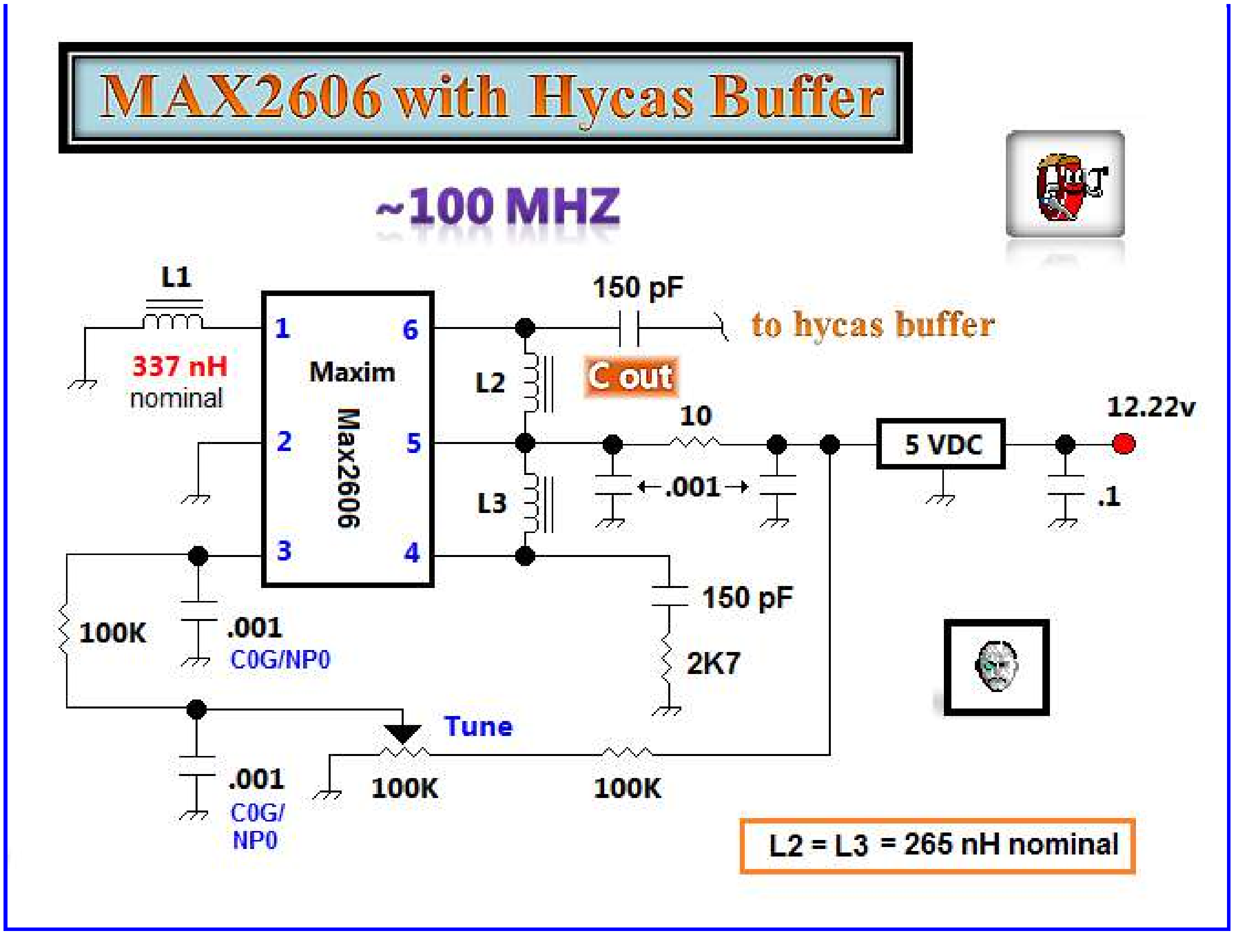
MAX2606 with Hycas Buffer
~100 MHZ
L1
337 nH
nominal
1
2
3
6
5
4
Maxim
Max2606
150 pF
to hycas buffer
L2
C out
10
L3
←.001→
5 VDC
12.22v
.1
150 pF
2K7
100K
.001
C0G/NP0
Tune
.001
C0G/
NP0
100K
100K
L2 = L3 = 265 nH nominal

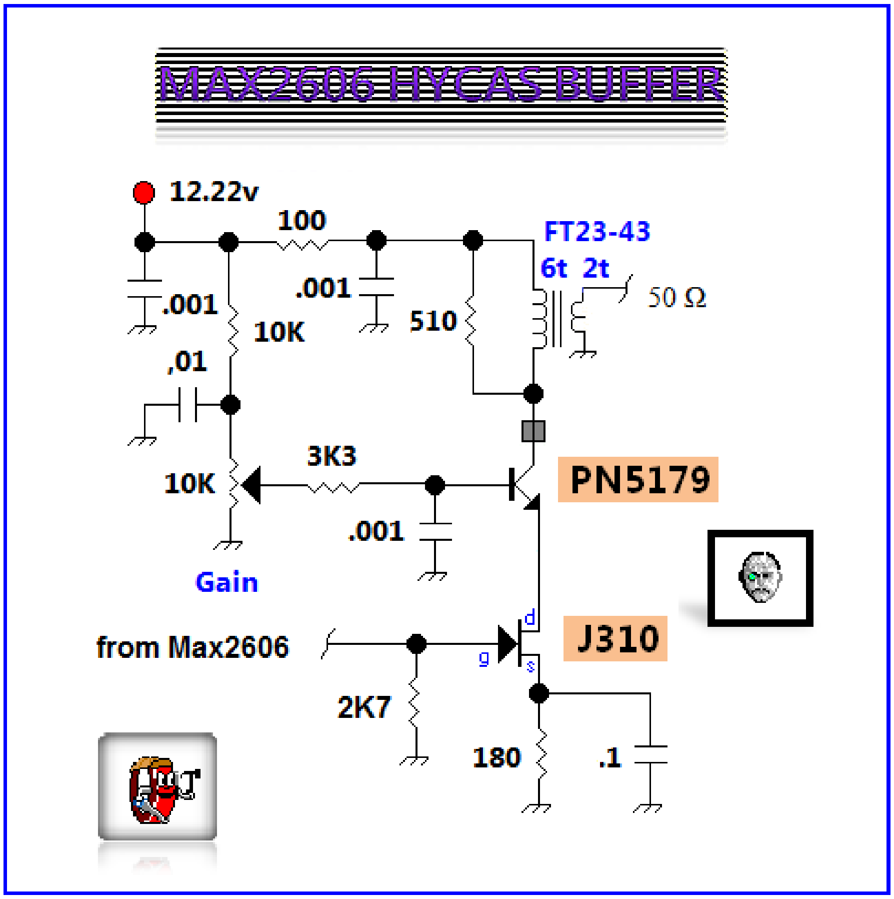
MAX2606 HYCAS BUFFER
12.22v
100
FT23-43
6t 2t
50 Ω
.001
.001
10K
510
,01
3K3
10K
PN5179
.001
Gain
d
from Max2606
g
s
J310
2K7
180
.1

Above — MAX2606 VCO into a hybrid cascode buffer at ~ 100 MHz. I also tried a buffer with a cascode of PN5179 transistors, however, the input Z of the buffer changed when adjusting the gain potentiometer and wrecked the input matching, so I stuck with the hycas buffer shown above.

L1 and L2 = ~ 265 nH wound on T37-10 toroids. To peak the L-match at ~102 MHz, I originally placed a 100 pF in parallel with a 10 - 70 pF trimmer cap for *C out*. After tuning, I removed, then measured these 2 caps to discover that *C out* = ~150 pF; so I just substituted a 150 pF capacitor. The L1 windings were expanded a little to re-peak the output network — expanding the coils raises fCo, while compressing the windings lowers the cut-off frequency. I don't recommend omitting a trimmer capacitor unless your confident with your measurements.

Click for the maximum power 'scope tracing of the hycas amp with the L-match peaked at 102.2 MHz (3.44 dbM)

Once again, I terminated the secondary output of the MAX2606 with an load resistance equal to the main output with an approximately equivalent fixed-value capactor — I'm not sure if it's needed, but it works okay and I stuck with it.

Up at ~ 144 MHz

Click for another experiment with the fore mentioned MAX2606 into the hycas buffer shown above at ~ 144 MHz. I peaked the L-match for 142.2 MHz and the output power = a surprising 7 dBm. The output is a clean sine wave from 135 to 169 MHz, although the signal amplitude varies widely. I also peaked it at ~144 MHz. I wound L1 - L3 with #21 AWG on a #10 bolt. Photo 1 Photo 2.

Let's build something useful with the MAX2606 based upon the experiments thus far...

Section 2.   A 2-band Signal Generator Based Upon the MAX2606 VCO

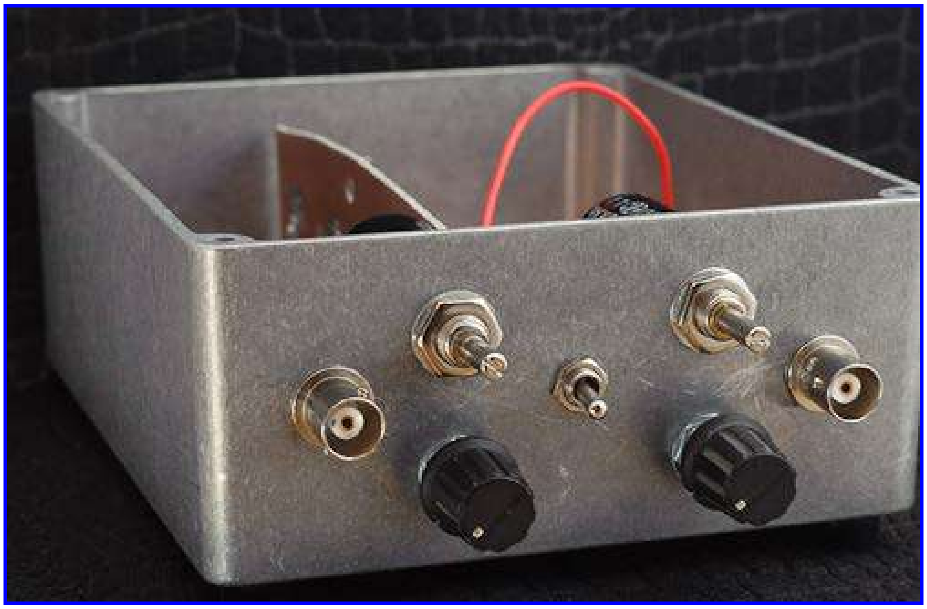

Above — My dual VCO based on the MAX2606. This general-purpose signal generator will start my VHF circuit development in these 2 frequency bands. The importance of owning good signal generators can't be overstated — while not engineer grade, this box features a clean sine wave, strong output return loss and 1 KHz or better tuning resolution.

The center ON-OFF-ON toggle switch only turns on 1 oscillator at a time. The top 2 (ten-turn) tuning potentiometers lack knobs *(I'll get some later)* and the black knobs below them are attached to *unused* potentiometers. Initially, I planned to employ front panel gain control and drilled holes and fitted 10K pots in the chassis — I later decided to control the output power with outboard variable attenuators and skipped front panel power control for simplicity sake. Click for a side view of the unfinished project. Click for an early photo of the VCO **A** breadboard — I attached a shield to the copper clad board to help isolate the 2 VCOs.

I chose the frequency band **A** (138.5 - 172 MHz) to include the Ham 2 meter band + local commercial/service VHF segment and band **B** (98.919 - 109.06 MHz) to capture the FM broadcast band above 98 MHz.

Band A: 138.5 - 172 MHz

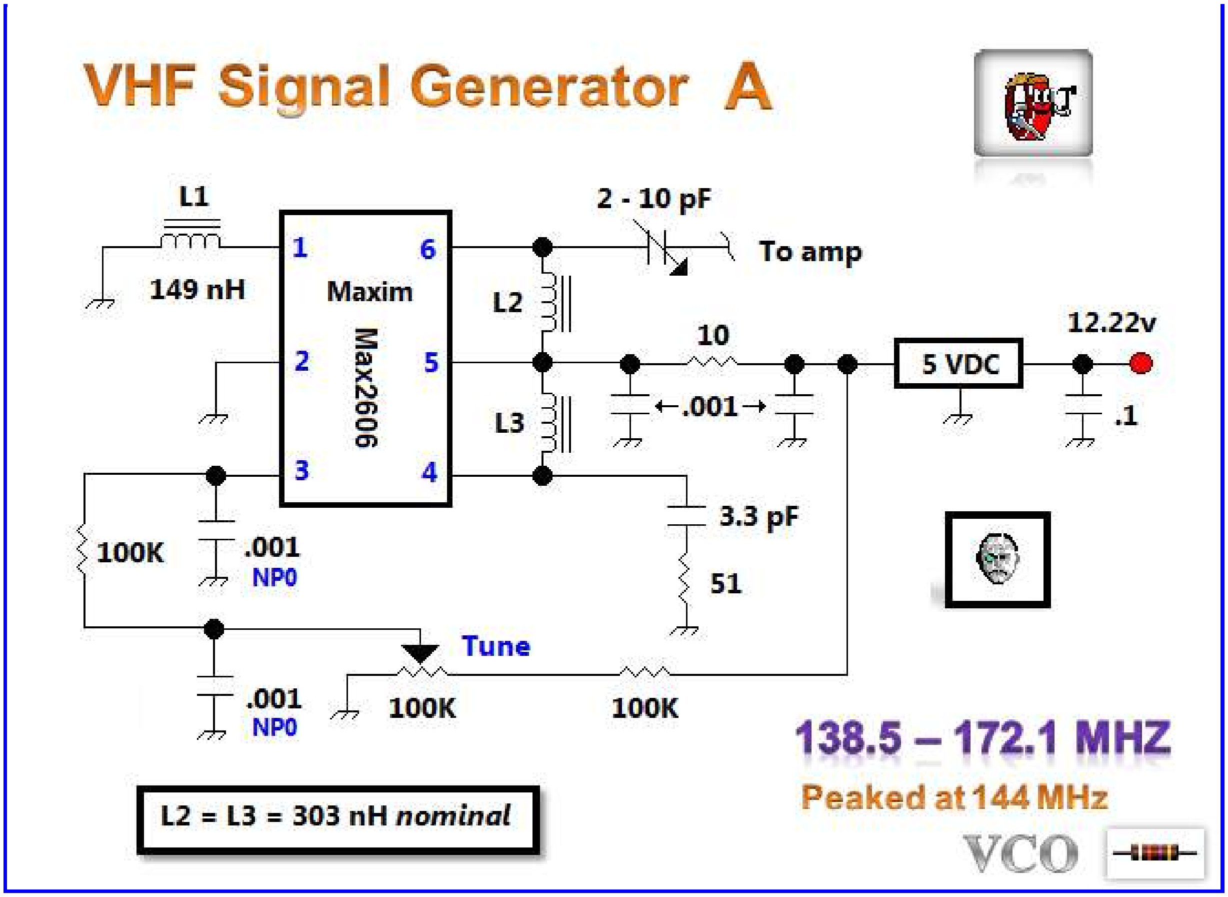
VHF Signal Generator A
L1
149 nH
1
6
Maxim
Max2606
2
5
3
4
2 - 10 pF
To amp
L2
L3
10
←.001→
5 VDC
12.22v
.1
3.3 pF
51
100K
.001
NP0
Tune
100K
100K
.001
NP0
138.5 – 172.1 MHZ
Peaked at 144 MHz
L2 = L3 = 303 nH nominal
VCO

Above — The VCO schematic **A**. An L-match peaks the output at 144 MHz into 50 Ω. The beauty of this circuit = simplicity; just 2 active devices give low distortion and a strong output return loss (S22) on the 2 Meter Ham band.

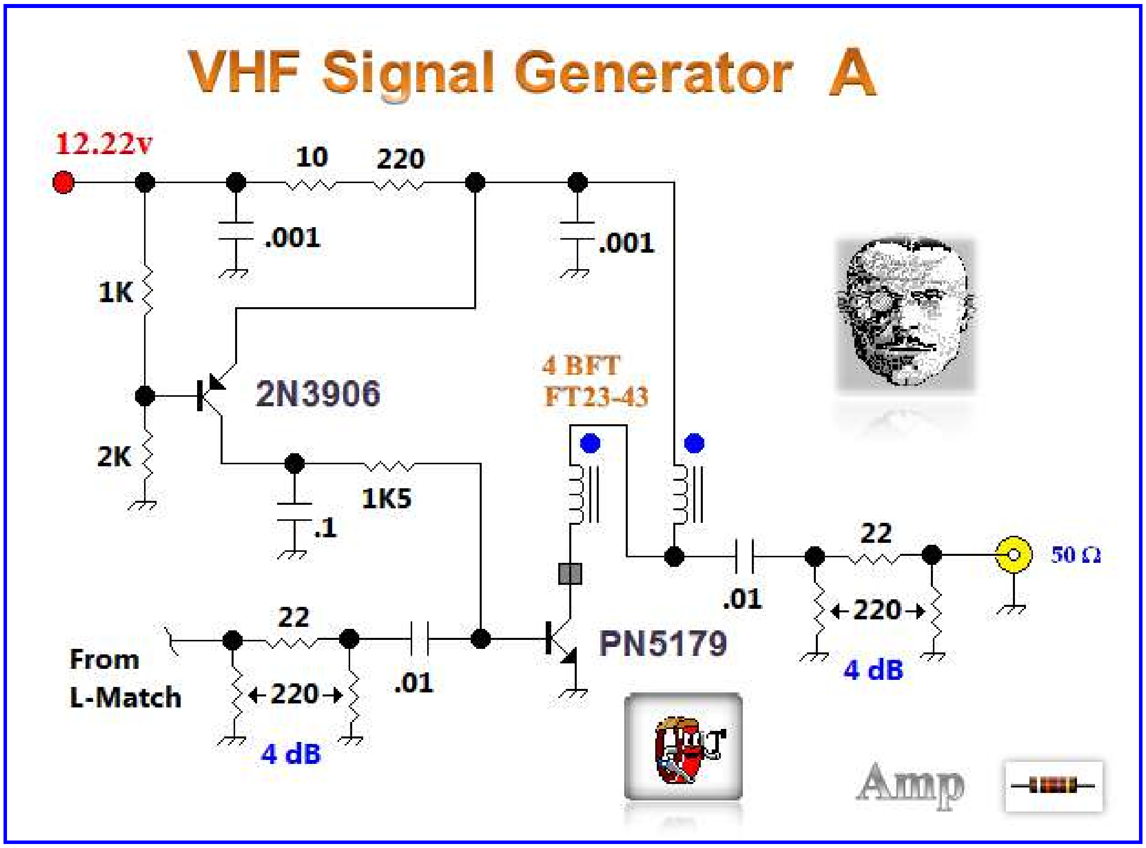
VHF Signal Generator A
12.22v
10
220
.001
.001
1K
2N3906
2K
.1
1K5
4 BFT
FT23-43
22
50 Ω
.01
←220→
4 dB
From
L-Match
22
←220→
.01
PN5179
4 dB
Amp

Above — The output amplifier for VCO "A". Click for a schematic with some analysis at 144 MHz (the frequency I'll use the most). I spent a lot of time trying to develop a 50 Ω output Z voltage amp up at ~ 144 MHz. My attempts to employ shunt and series feedback gave generally poor results — stray reactances plague the standard FBA designs that work great under ~70 MHz.

Since the MAX2606 has an output L-Match, getting a high output return was my only goal — grounding the PN5179 emitter, employing a 4:1 transmission line transformer and biasing for ~15 mA emitter current did the trick. A 4 dB pad on the NPN input further establishes a strong S22 (output return loss).

As possible, I attach additional 50 Ω outboard attenuators on my signal generators, however, the S22 on the stock generator should be okay across the VCO range due to the two 4 dB pads.

I measured the S11 on a prototype NPN amplifier by using the MAX2606 VCO shown to drive a MMIC with a 16 dB 50 Ω attenuator pad on the output and connecting it to the RF port of my return loss bridge. Thus, the actual VCO helped me design the final amplifier which buffers it in my final build. I kept the PN5179 and all other leads short as possible.

The output power looks like a sine wave when tuned from CCW to CW: With the 10-turn tuning potentiometer set to CCW [138.5 MHz], the output power ranges from -1.9 dBm; hits a peak of 0 dBm at 144 MHz and then gradually drops to -11.2 dBm at CW [172.1 MHz].

I moved away from the hycas buffer amp to simply my design and increase reproducibility. A MMIC might also work well, however the high current drain, potential for instability and biasing considerations introduce new problems — a simple wrap around PNP-biased NPN amplifier works okay. Sometimes the best solution = the simplest.

Band B: 98.919 - 109.06 MHz

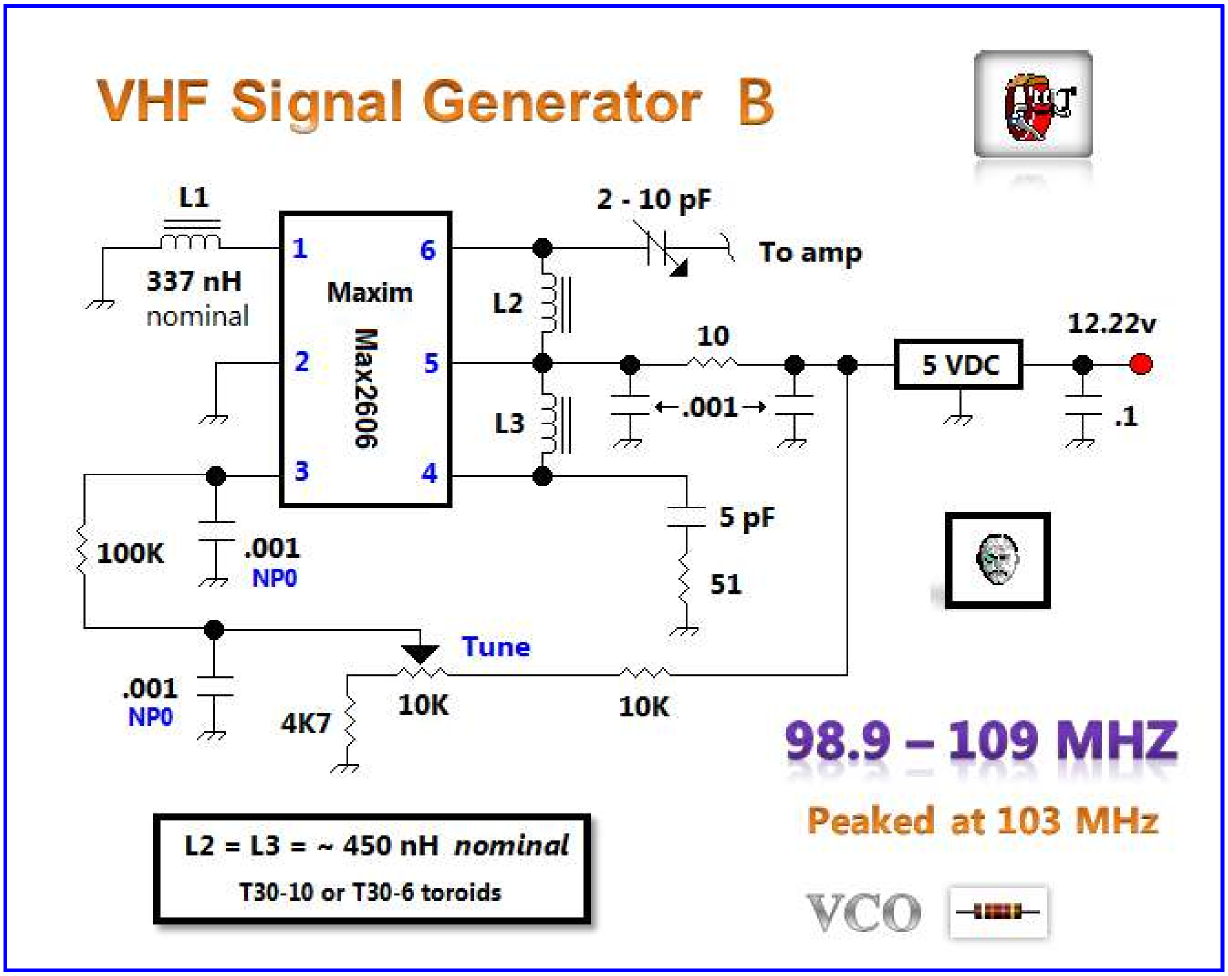

Above — The VCO schematic **B**. An L-match peaks the output at 103 Mhz, although compressing or expanding L2 and tweaking the 2 -10 pF capacitor can peak the L-network anywhere in the tuning range. I limited the tuning range by adding a 4K7 R to the tuning pot to enable better matching and fine tuning.

In all cases L1 = an air inductor wound on a coarsely threaded bolt using 20 or 21 guage enamel-coated magnet wire for mechanical stability. I set the lower band edge of the VCO by setting the tuning pot to CCW and compressing or expanding the L1 inductor to get the frequency shown.

For L1, I found that excessively long leads can create unwanted oscillations and my coils are just a few mm above the copper board. Some builders properly mount their coils in a upright "smokestack" fashion and/or well away from the nearby metallic chassis or copper PCB to minimize Q losses + inductance changes. My coils lie well away from the metal chassis walls.

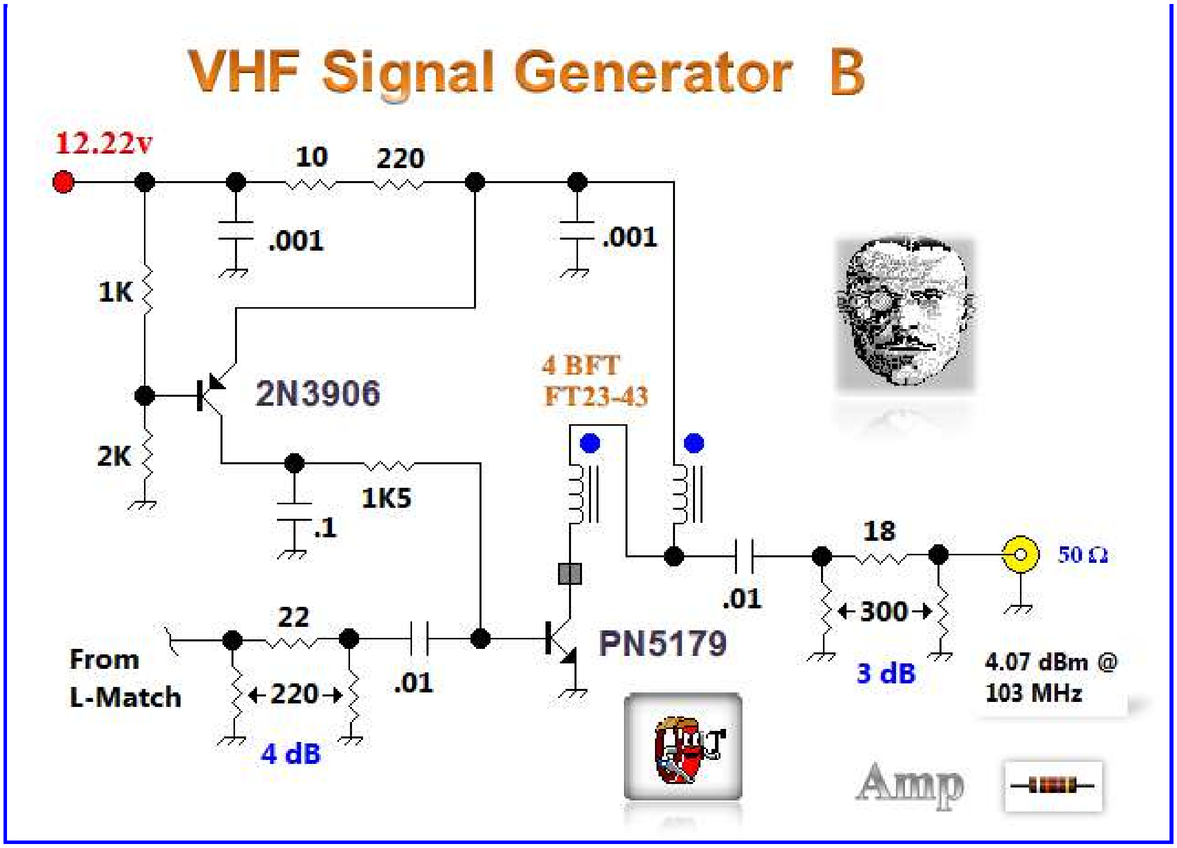
VHF Signal Generator B
12.22v
10
220
.001
.001
1K
2N3906
2K
4 BFT
FT23-43
1K5
.1
18
50 Ω
.01
22
From
L-Match
300
3 dB
PN5179
220
.01
4 dB
4.07 dBm @
103 MHz
Amp

Above — The VCO **B** buffer/amplifier schematic. A clean sine wave appears across the entire tuning range — output return loss tuning from minimum to maximum was > 21 dB before I added the 3 dB pad.

Click for a photo of the completed, partially labelled project during the final tune up with all the boards bolted in.

Section 3. 50 MHz VCO

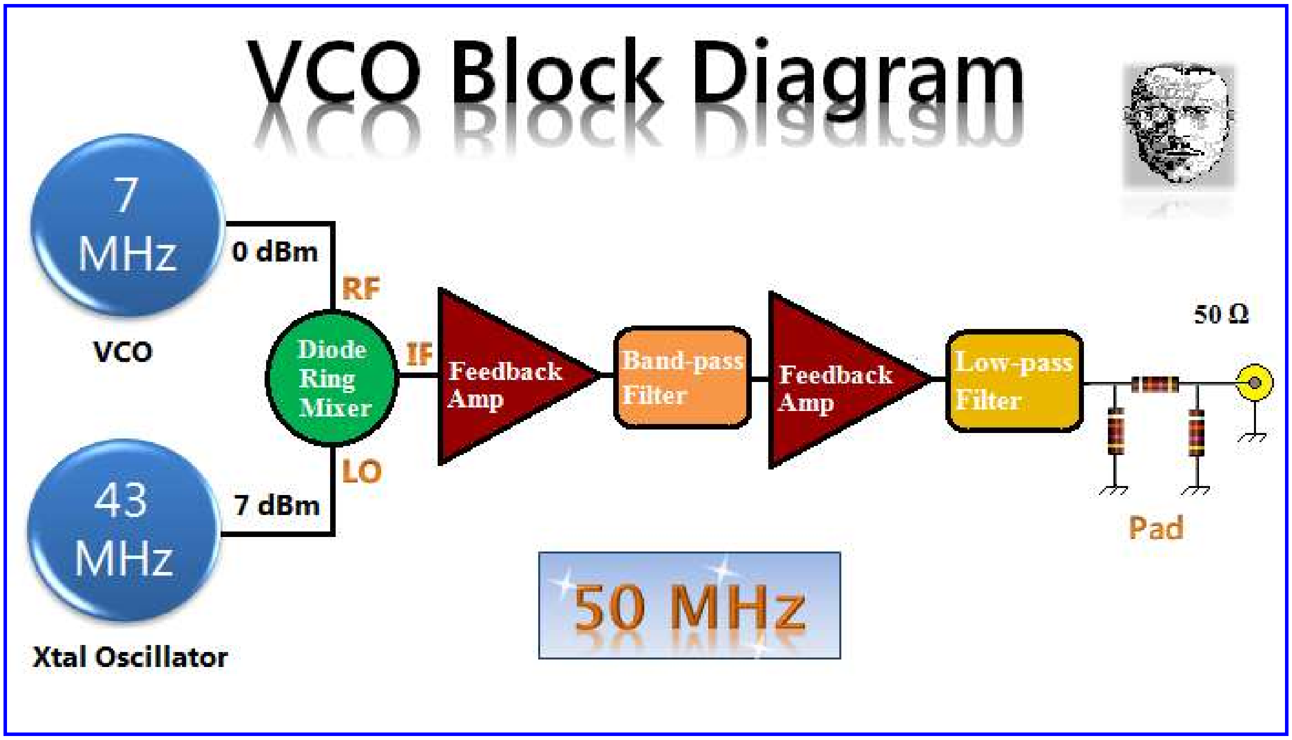
VCO Block Diagram
7
MHz
VCO
0 dBm
RF
Diode
Ring
Mixer
IF
Feedback
Amp
Band-pass
Filter
Feedback
Amp
Low-pass
Filter
50 Ω
LO
43
MHz
7 dBm
Pad
Xtal Oscillator
50 MHz

Above — Block diagram of the 50 MHz VCO I designed and built in February-March 2012. Click for a photo.

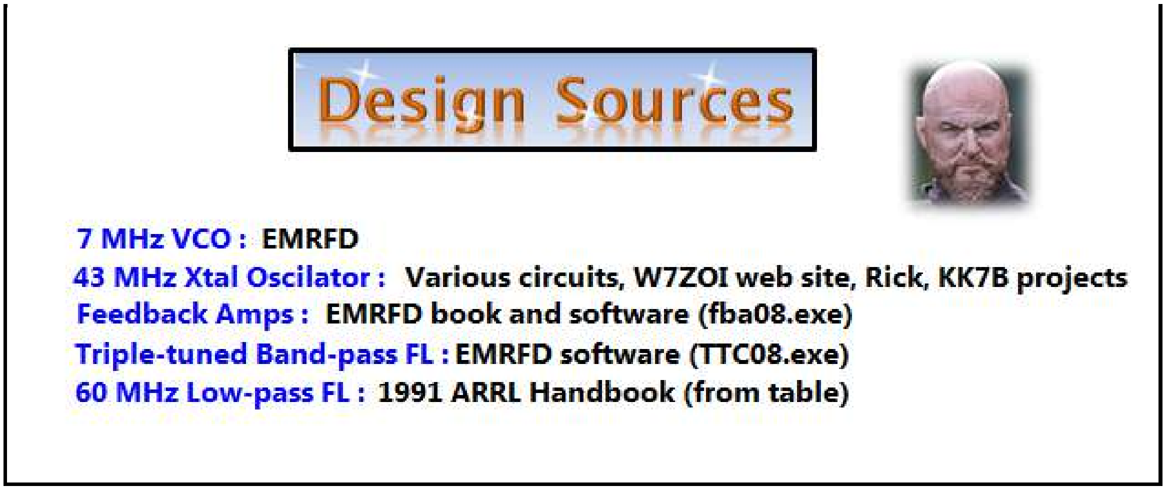

Design Sources

7 MHz VCO : EMRFD

43 MHz Xtal Oscilator : Various circuits, W7ZOI web site, Rick, KK7B projects

Feedback Amps : EMRFD book and software (fba08.exe)

Triple-tuned Band-pass FL : EMRFD software (TTC08.exe)

60 MHz Low-pass FL : 1991 ARRL Handbook (from table)

Above — Like most of you, I'm just an amateur designer who relies on others for example circuits, design procedures and inspiration. These cited references plus hard work drove my experiments. This project succeeds the Miscellaneous RF Experiments web page from 2011 — QRP SWL HomeBuilder evolves as I do.

7 MHz VCO + Buffer Amplifier [0 dBM output power]

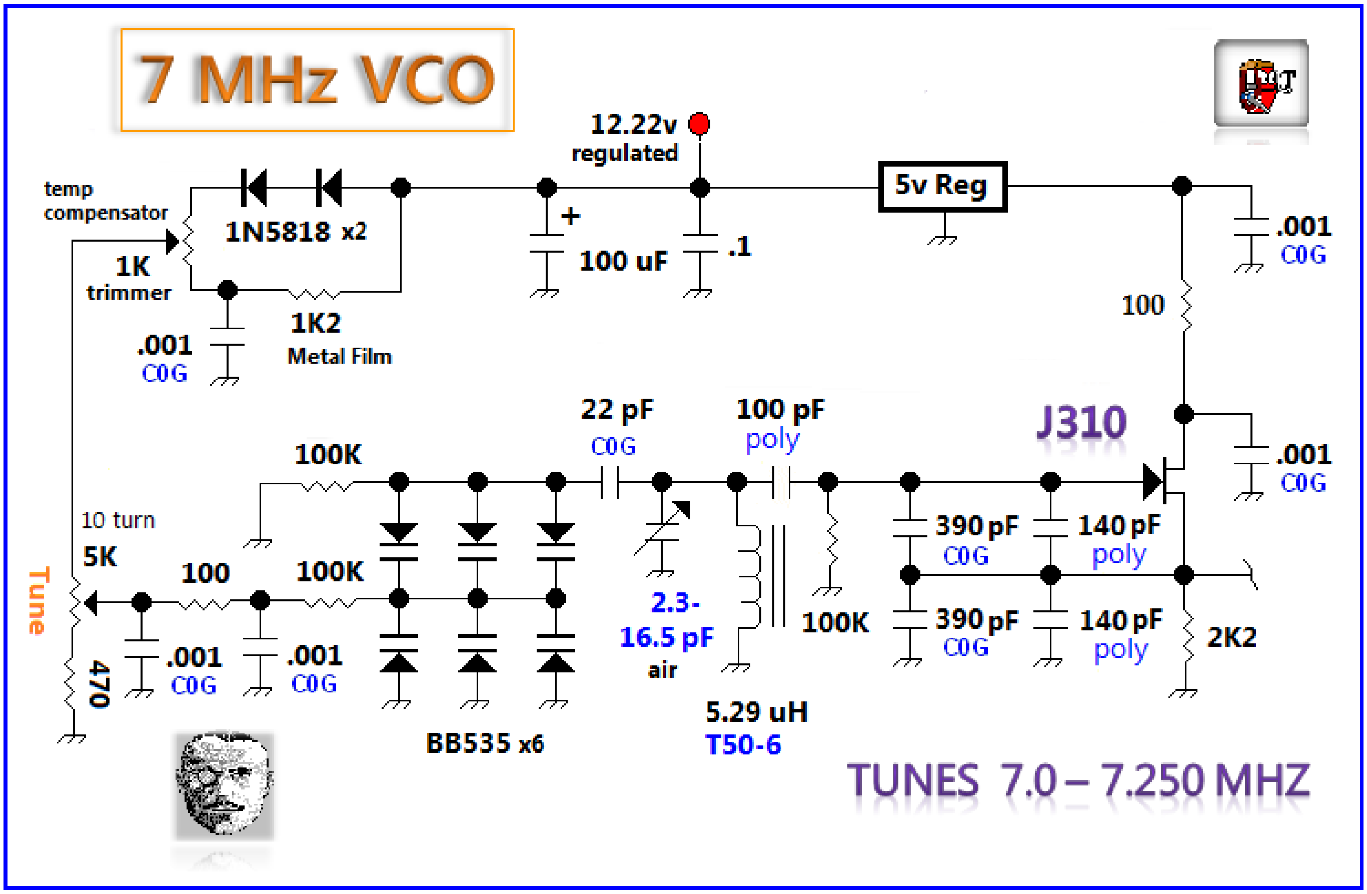

7 MHz VCO
12.22v
regulated
temp
compensator
1N5818 x2
1K
trimmer
.001
C0G
1K2
Metal Film
+
100 uF
.1
5v Reg
.001
C0G
100
J310
.001
C0G
22 pF
C0G
100 pF
poly
100K
10 turn
5K
Tune
470
100
100K
.001
C0G
.001
C0G
BB535 x6
2.3-
16.5 pF
air
5.29 uH
T50-6
100K
390 pF
C0G
390 pF
C0G
140 pF
poly
140 pF
poly
2K2
TUNES 7.0 – 7.250 MHZ

Above — 7 MHz Colpitts VCO schematic.

This VCO tunes ~7.00 to 7.250 MHz, although a wider tuning range occurs if you allow the tuning diodes to drop to 0 VDC with the 5K tuning pot cranked CCW. A 470 ohm resistor keeps about 1 VDC on the varactors at the lowest tuning frequency/applied reverse DC voltage.

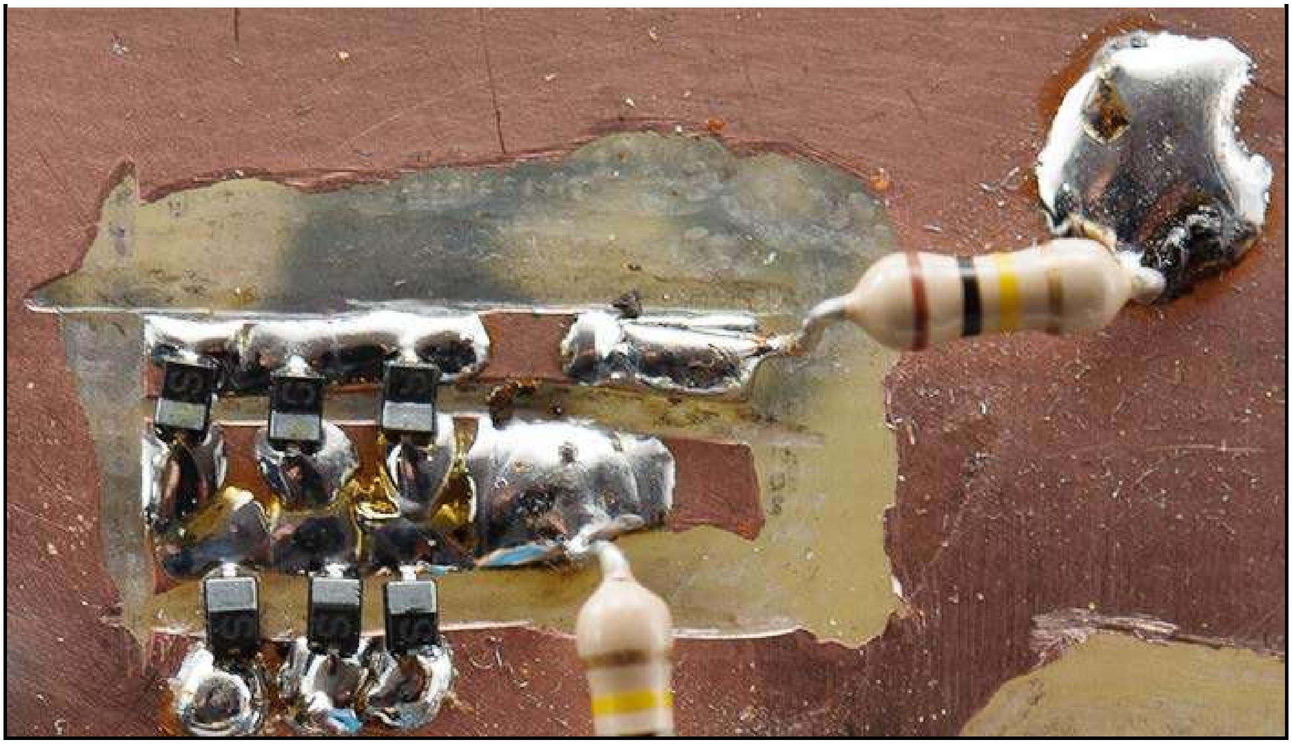

Above — A macro photograph of the six BB535 varactors soldered on the VCO breadboard. With 0 applied reverse DC voltage, their total C = 43.5 pF. I left room for up to 4 more diodes, but didn't need them.

Macro photography provides an excellent way to inspect SMT parts — apart from all the fiberglass dust on the board, no shorts or other problems arose when soldering. Next to pF-value chip capacitors, these SMD varactors proved the most difficult surface-mount parts I've breadboarded to-date. Using clear tape, I tape my SMT parts to the PC board when soldering. With tape, you can still make tiny device placement adjustments with a pick or tweezers and yet the device holds steady enough to solder. I recently obtained a microscope for SMT work, although didn't need it for these diodes.

Striving for lower phase noise meant properly applying high Q tank parts — I soldered in 3 pairs of high-grade BB535 varactors and arranged them anti-parallel to avoid forward conduction + even harmonics. I also limited the AC voltage swing they "see" by connecting them to the L with a 22 pF capacitor. Tight windings of #28 gauge magnet wire on a T50-6 toroid formed the inductor. 4 number 8 bolts anchor the 7 MHz VCO board to the chassis and prevent board warp + movement.

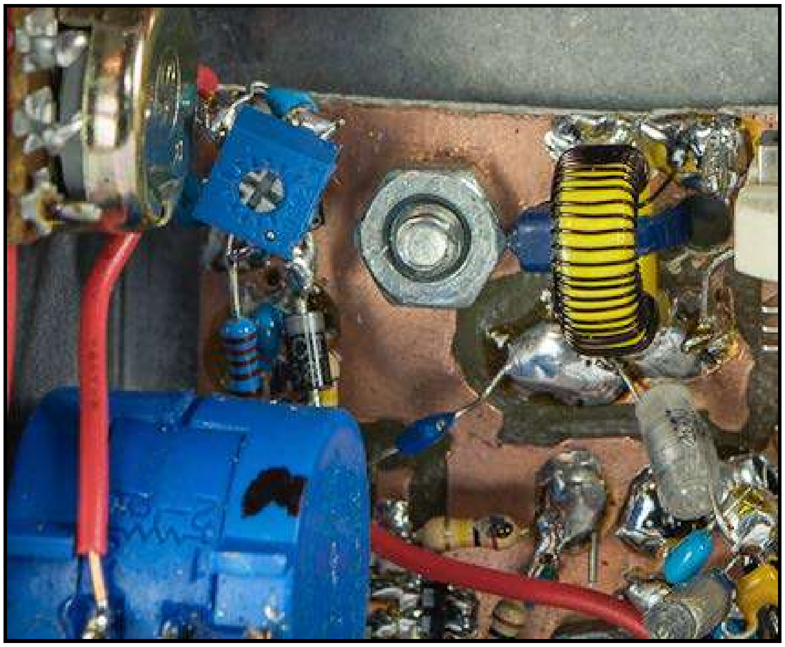

Above — a view of the square blue 1K temperature compensation (tempco) trimmer potentiometer.

To aid temperature compensation, I included 3 polystyrene capacitors in the base VCO — the tempco circuitry represents about 16 hours of work from December 2011. Click or click for photos of the bread board before the tempco parts were soldered on — the temporatry BNC connector lies in the background was removed after testing.

With care and patience my lid-on 1 hour temperature drift = ~10 Hz. My temperature compensation strategy worked because I took the time to measure and then determine how to cancel temperature drift in this 1 circuit — your results will vary and *experimentation* remains the key to temperature compensating VCOs and VFOs.

See EMRFD, and the VFO-2011 + QRP Modules 2011 web pages for more tempco information.

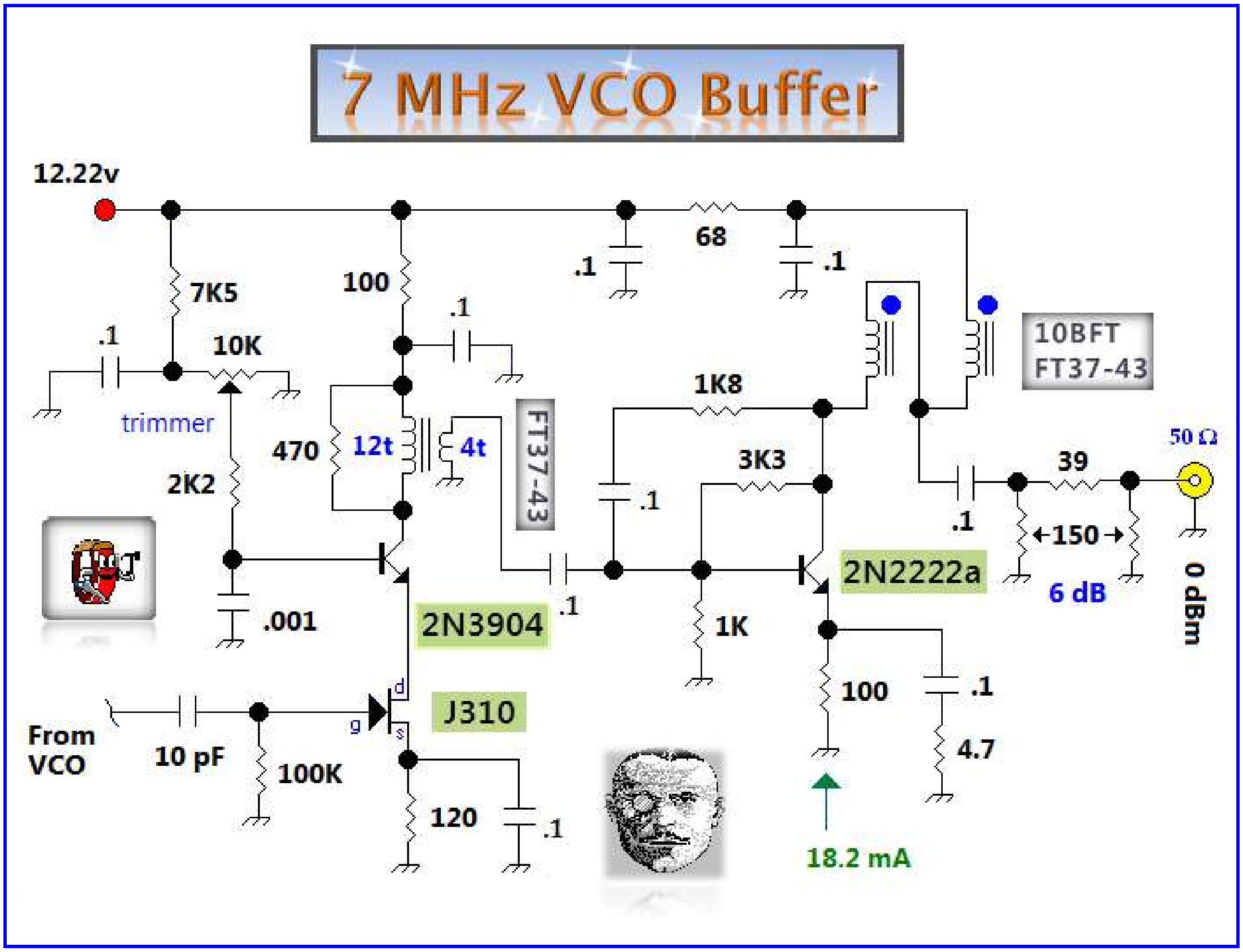
7 MHz VCO Buffer
12.22v
7K5
.1
10K
trimmer
2K2
.001
100
.1
470
12t
4t
FT37-43
2N3904
.1
From
VCO
10 pF
100K
J310
d
g
s
120
.1
68
.1
.1
1K8
3K3
1K
2N2222a
100
.1
4.7
18.2 mA
10BFT
FT37-43
.1
39
150
6 dB
50 Ω
0 dBm

Above — 7 MHz VCO buffer/amplifier. I adjusted the 10K trimmer pot on the hycas buffer for exactly 0 dBm drive. Even before adding the 6 dB attenuator pad, the output return loss = 23.8 dB @ 7.0 MHz.

Originally, I wanted a VCO output of 7 dBm and applied ~18 mA emitter current in the final amp to preserve signal fidelity and eliminate the need for a low-pass filter. This buffer works great up to an output power of ~10 dBm: above 10 dBm or so, distortion occurs and you'll need to add a low-pass filter.

Adjust the hycas trimmer pot for whatever output power you seek, but If you're ever using this buffer for 7 -10 dBm output power, drop the 6 dB attenuator pad to 3 dB. This drops the drive level to maintain low harmonic distortion (2nd harmonic down > 35 dBc).

43 MHz Butler Xtal Oscillator [6.4 dBm output power]

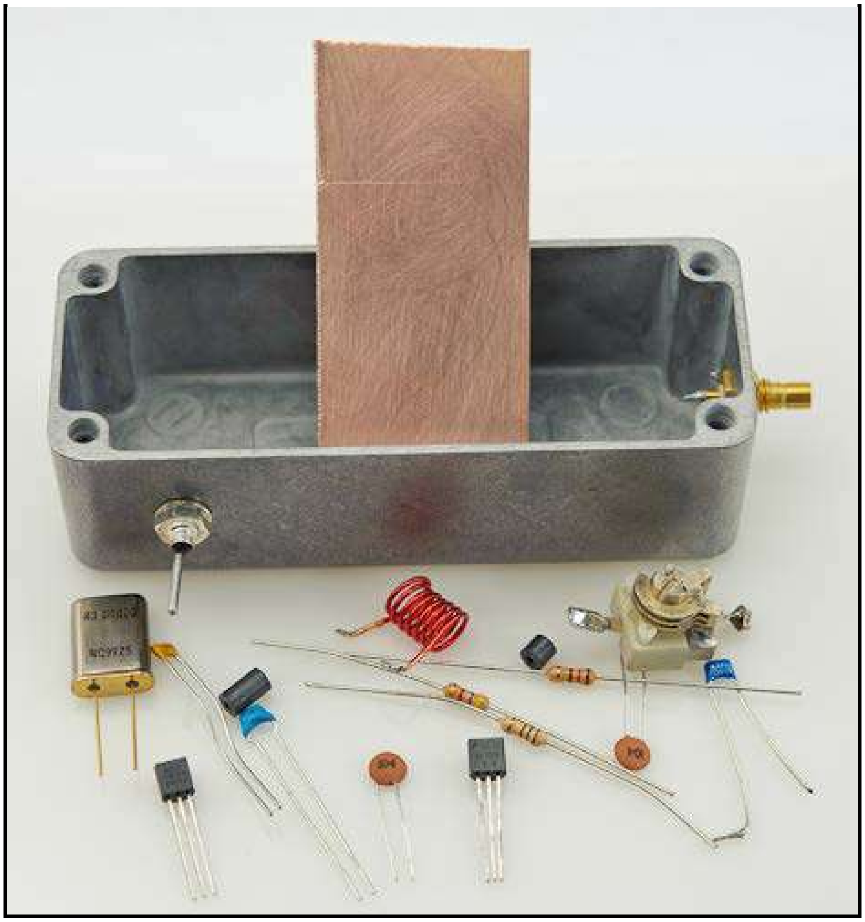

Above — Some Butler crystal oscillator parts prior to the build. Since this Butler will go inside a box containing a VCO and some high gain amplifiers, it would be foolish to not stick it in an RF-tight box. On the Hammond chassis above, you'll notice a feedthrough capacitor for the 12 VDC and an gold colored SMA connector for the output.

Since the required L = over 400 nH, I opted for a toroidal inductor wound on a T30-10 instead of the air coil shown in the photo.

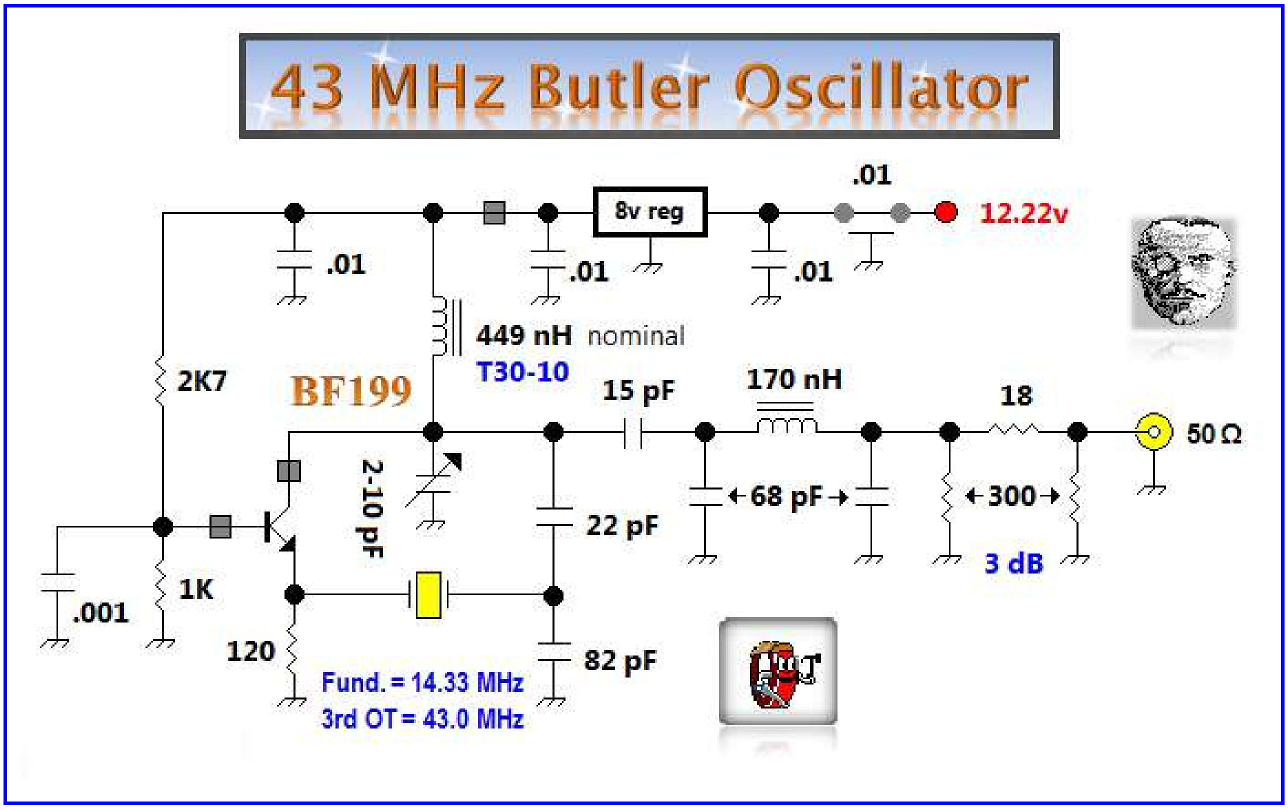
43 MHz Butler Oscillator
.01
8v reg
12.22v
.01
.01
.01
449 nH nominal
T30-10
2K7
BF199
15 pF
170 nH
18
50 Ω
2-10 pF
68 pF
300
22 pF
3 dB
.001
1K
120
82 pF
Fund. = 14.33 MHz
3rd OT = 43.0 MHz

Above — 43 MHz Butler overtone oscillator schematic. The highest power I could muster = 6.4 dBm (close enough to 7 dBM).

This Butler looks good on FFT. Click, Click or click for 3 'scope captures. Despite trying to milk maximal power, the 2nd harmonic is over 40 dBc down. Click for a snap shot of the completed oscillator.

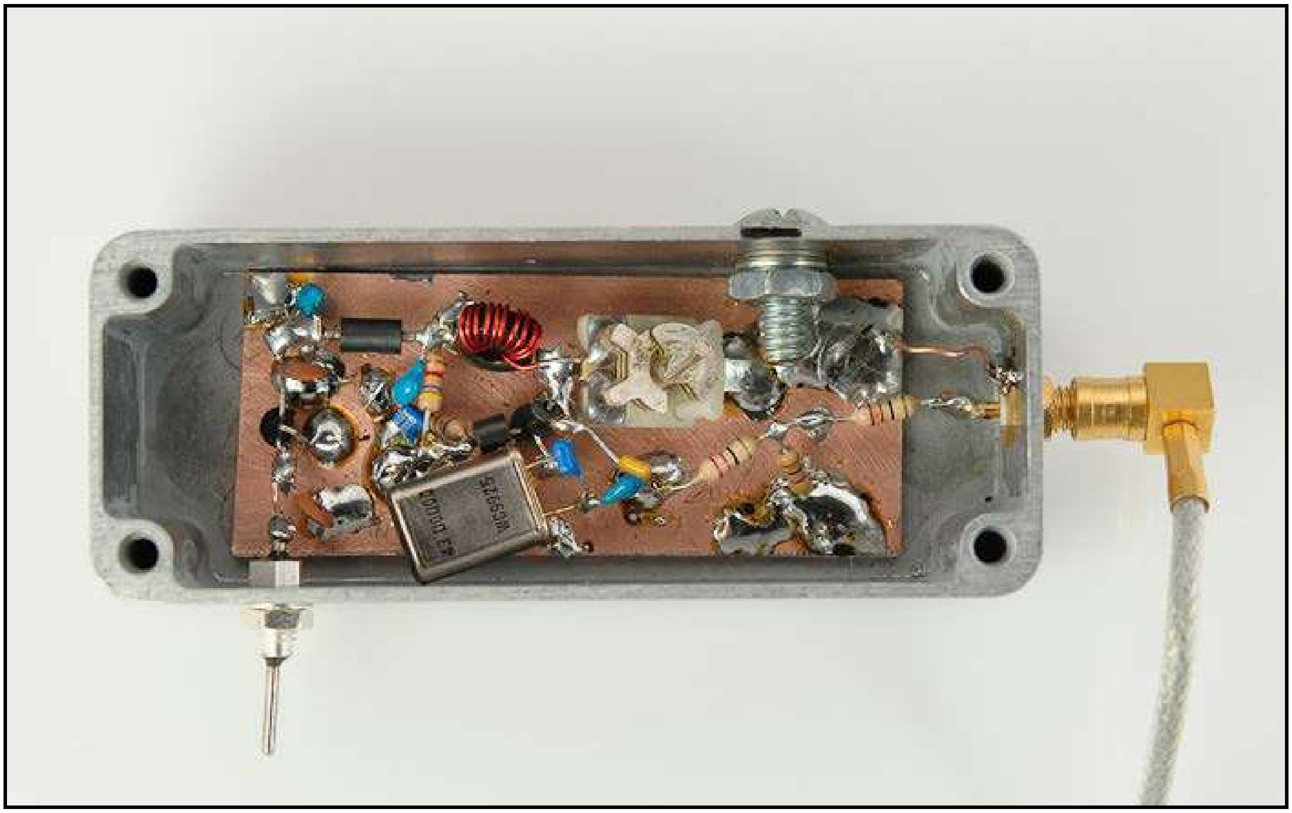

Above — The original Butler oscillator before adding the pi low-pass filter. The bolt (seen at top right) will also pass through the outer VCO chassis to hold this sub-chassis in place. Click for a bigger photograph.

Post-Mixer Amplifer and Triple-Tuned Filter

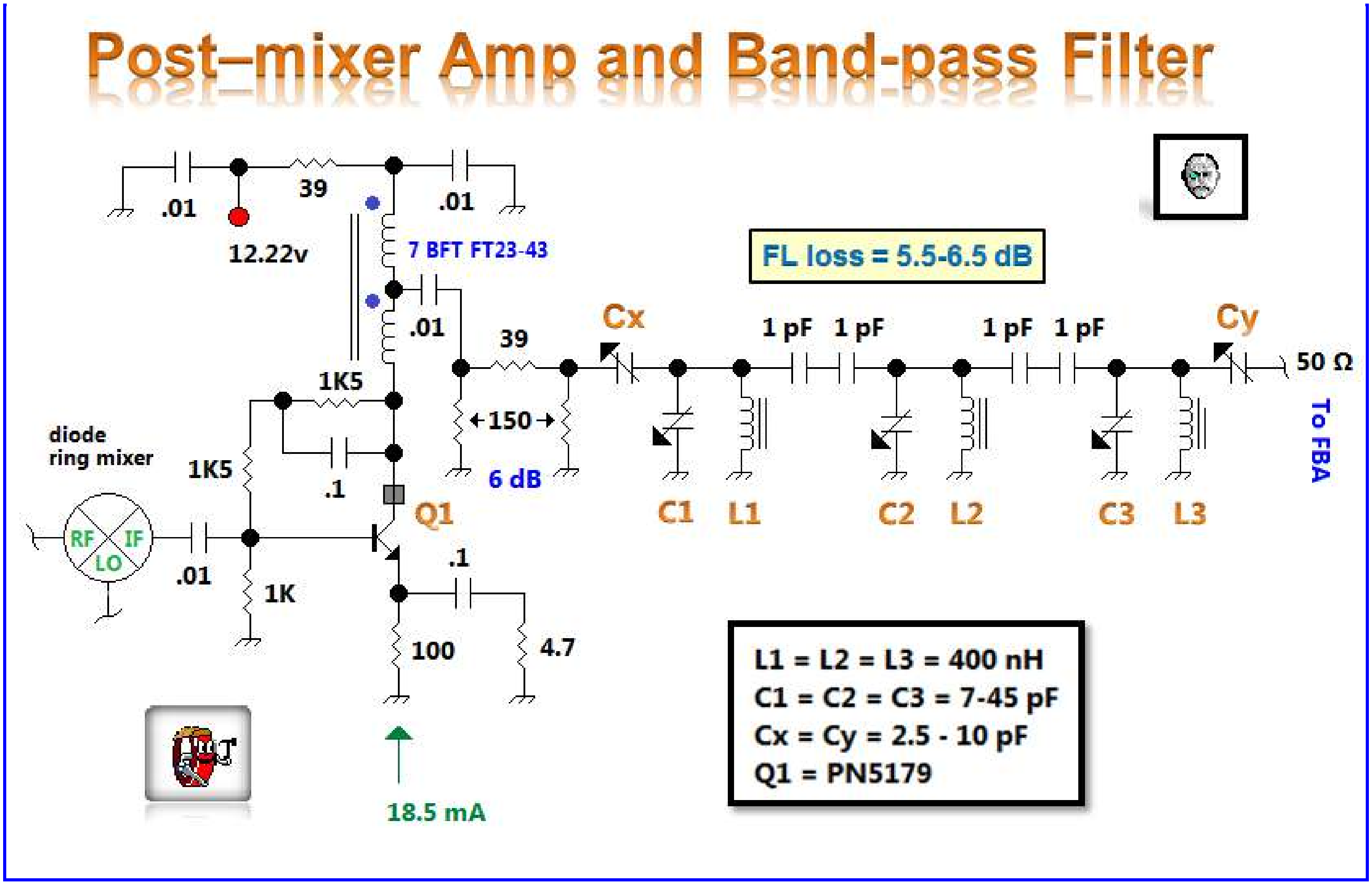

Above — Schematic of the diode ring mixer, Q1 feedback amp and the triple-tuned filter. *I used a MCL SBL-1 mixer*. L1 - L3 were wound on T30-10 toroids. I bought my #10 and some #12 toroids from the great folks at Debco Electronics.

The post-mixer feedback amp data at 50.0 MHz (isolated from the mixer and pad + filter) : Emitter current = 18.5 mA, S21 =18 dB, S11 = -24.4 dB, S22 = - 21.5 dB. *(S21 = power gain; S11 = negative of the input return loss; S22 = negative of the output return loss).*

The 6 dB pad helps absorb signal reflections from the filter caused from stray reactance plus capacitance variations caused by coupling the 2 tanks with only 0.5 pF (2 series 1 pF capacitors with a +/- 0.25 pF tolerance!)

Preliminary filter alignment: Peak your filter however you want — but here's how I peaked my filter with a crystal controlled 50.0 MHz signal generator connected via a temporary BNC connector tack soldered onto the copper board and wired to the Q1 input. Terminate the filter with a ~50 Ω resistor, or a temporary BNC connector plugged with a 50 Ω resistive terminator.

Connect the signal generator to Q1 and peak C1, C2 and C3 (in that order) using a 10X 'scope probe. It's better measure with your probe at point C2 when tuning C1 since this reduces mistuning caused by probe capacitance — measure at point C3 when tuning C2 etc.

Then peak C3, Cx and Cy with the probe touching the terminating 50 Ω resistor. *It easier to perform the first tune-up with a 10X probe going sequentially from C1 to C3 since these peaking capacitors tune pin sharp.*

After the preliminary tune-up, if possible, connect a temporary BNC connector to the output and re-peak all the caps with a 50 Ω terminated scope; this boosts sensitivity and eliminates 'scope probe capacitance.

Perform pentultimate 50.0 Mhz alignment after you add the post-filter amp, low-pass filter and the 3 dB pad. Capacitor Cy critically sets the output return loss of Q2 and when properly matched, establishes a 50 Ω termination for the triple-tuned band-pass filter.

You can also match Cx by connecting a return loss bridge to the input of Q1 and terminating the RF chain with a 50 Ω resistor, although tuning Cx only changes the input S11 a little. In 1 experiment, I replaced Cx with 6 pF and it worked okay.

I wonder how I ever managed before making a return loss bridge: the workhorse of the QRP workbench.

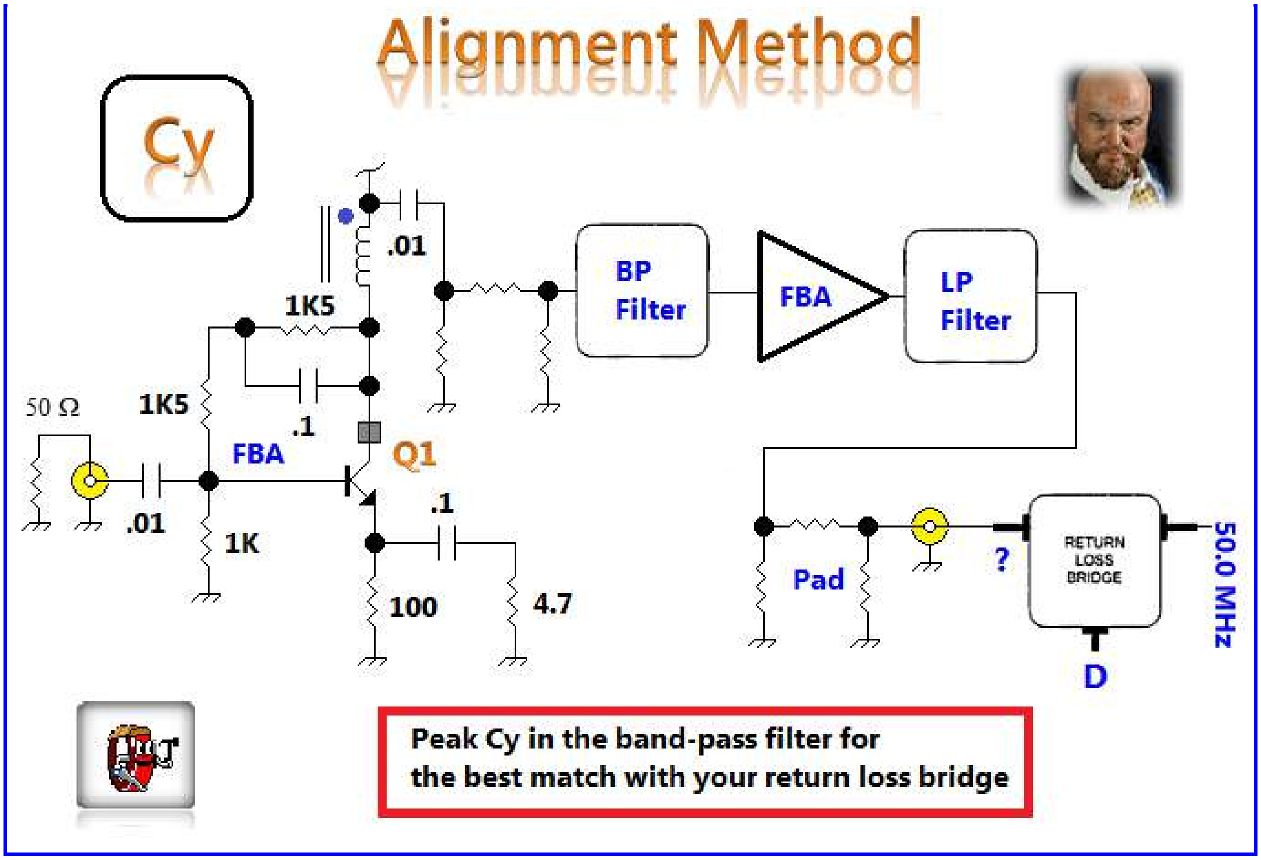
Alignment Method
Cy
.01
1K5
1K5
50 Ω
.1
FBA
Q1
.01
1K
.1
100
4.7
BP Filter
FBA
LP Filter
Pad
?
RETURN LOSS BRIDGE
50.0 MHz
D

Peak Cy in the band-pass filter for the best match with your return loss bridge

Above — Schematic depicting how to tune up Cy.

Tuning Cy matches the band-pass filter output to the Q2 input impedance — it's fascinating to examine the interdependence of these 50 Ω stages. After setting Cy, I connected the 50 MHz signal generator to the Q1 input and a 50 Ω terminated 'scope to the output and re-peaked Cx, C1, C2 and C3 — finally I tweaked Cy 1 last time with the whole stage in a return loss measurement set-up.

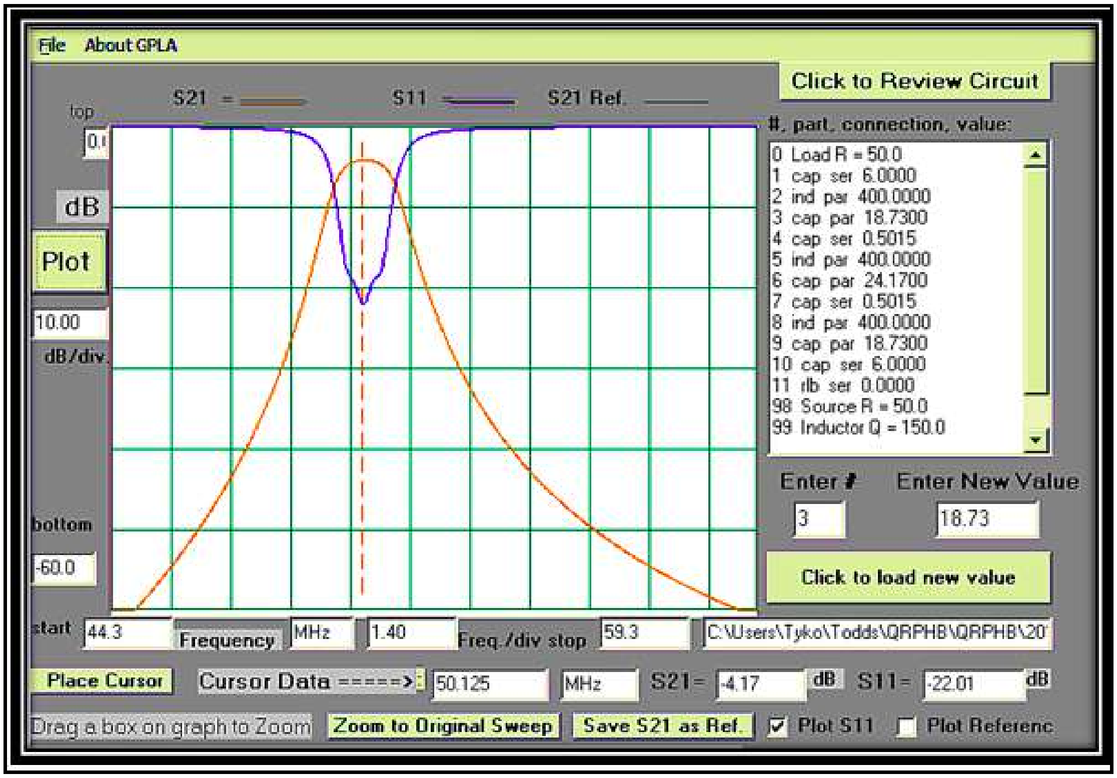
File About GPLA
S21 =
S11
S21 Ref.
top
dB
Plot
dB/div.
bottom
Click to Review Circuit
Enter #
Enter New Value
Click to load new value
start
Frequency
MHz
Freq./div stop
Place Cursor
Cursor Data =====>
MHz
dB
dB
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Plot Referenc

#, part, connection, value:

```
0  Load R = 50.0
1  cap  ser  6.0000
2  ind  par  400.0000
3  cap  par  18.7300
4  cap  ser  0.5015
5  ind  par  400.0000
6  cap  par  24.1700
7  cap  ser  0.5015
8  ind  par  400.0000
9  cap  par  18.7300
10  cap  ser  6.0000
11  rlb  ser  0.0000
98  Source R = 50.0
99  Inductor Q = 150.0
```

top: 0.0
10.00 dB/div.
bottom: -60.0
Enter #: 3
Enter New Value: 18.73
start: 44.3 Frequency MHz
Freq./div: 1.40
stop: 59.3
C:\Users\Tyko\Todds\QRPHB\QRPHB\20
Cursor Data: 50.125 MHz
S21= -4.17 dB
S11= -22.01 dB
☑ Plot S11
☐ Plot Referenc

Above — A GPLA simulation of the triple-tuned band-pass filter. CF = 50.125 MHz. I substituted 6 pF (the nominal value) for the 2 series end capacitors in my simulation. отлично!

Post-filter Feedback Amplifer, Low-pass Filter and Pad

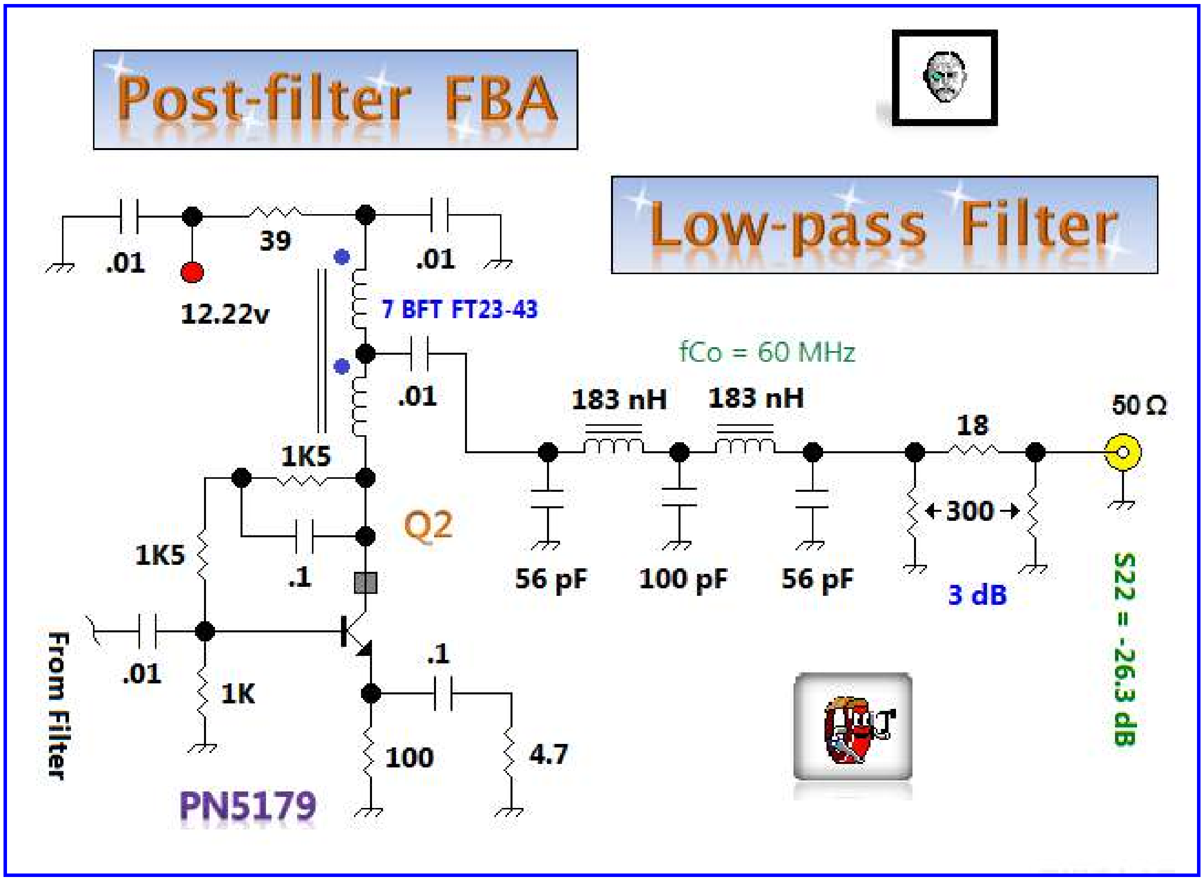

Above — Q2, the post filter feedback amp (FBA), an N=5 Chebyshev low-pass filter plus a 3 dB pad.

For Q2, I copied Q1 to deliver a strong input and output return loss. In many circuits employing cascaded FBAs, you increase emitter current in each successive FBA to reduce distortion, however, increasing emitter current affects both the input and output impedance and may trash your amplifier's S11 and S22.

I spent days studying, simulating + bench testing different amplifier designs in the Q2 slot — I generated enough material for another web page and plan to show this work in an update to my Popcorn superhet receiver some day.

It's possible to overdrive Q2 depending on your amplifer power and stage matching. If so, you might consider placing a 3-4 dB pad after the band-pass filter. Some might opt for a 7 element low-pass filter; experiment — as always.

Low-pass filter inductors = turns on T30-10 toroids, although #6 material toroids, or air coils will work fine.

Outputs

After bolting down the boards, wiring the DC and RF and confirming it worked, I finalized alignment. Using a frequency counter, I tuned the VCO to 50.125 MHz (the half-way point) and peaked C1, C2, C3 for the maximum peak-peak voltage into my 50 Ω terminated 'scope.

Click for the output at 50.125 MHz — 10.09 dBm. I normally hang an outboard 50 Ω attenuator on the output of my signal generators and keep 3 dB, 6 dB, 10 dB and 20 dB BNC-connected pads handy. With a 3 dB pad, the output power = 6.84 dBm — perfect for switching Level 7 diode ring mixers. Click for the 'scope shot with a 3 dB pad applied.

Click or click for an FFT of the output signal. The second harmonic is > 50 dB down. What fun!

The "vestigial" RF gain control shown on the chassis remains unused; wastage.

Miscellaneous Photos and Figures

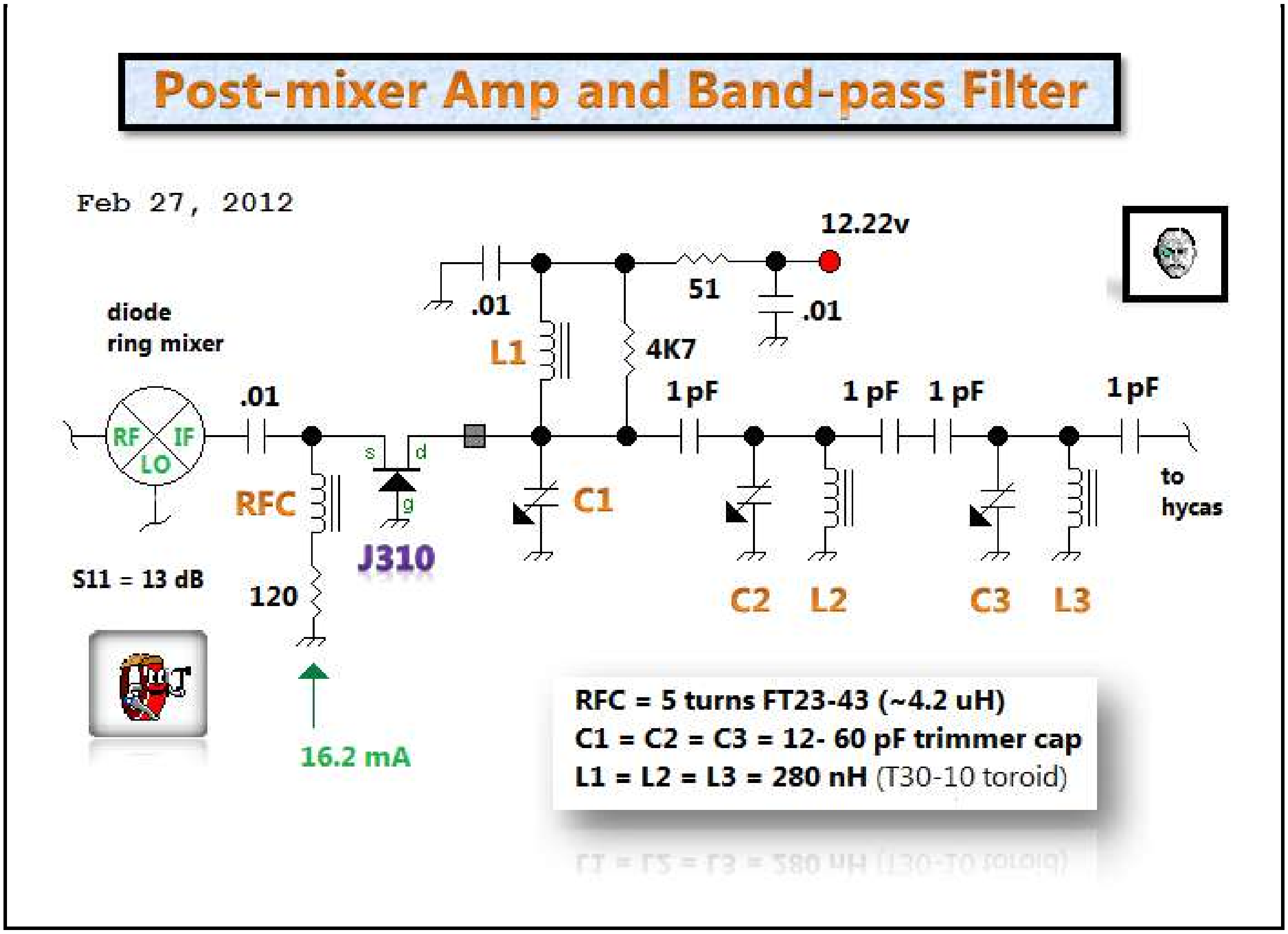
Post-mixer Amp and Band-pass Filter
Feb 27, 2012
12.22v
diode
ring mixer
.01
51
.01
L1
4K7
1 pF
1 pF
1 pF
1 pF
.01
RF
IF
LO
s
d
g
RFC
J310
C1
to
hycas
S11 = 13 dB
120
C2
L2
C3
L3
16.2 mA
RFC = 5 turns FT23-43 (~4.2 uH)
C1 = C2 = C3 = 12- 60 pF trimmer cap
L1 = L2 = L3 = 280 nH (T30-10 toroid)

Above — A **failed** experimental JFET post-mixer amp with tuned output driving a double-tuned filter.

Click for a GPLA simulation of the double-tuned filter. The common gate JFET amp provides a great way to terminate a diode ring and obviates the need for a diplexer network. Click for a breadboard photo of the above stage.

The amplifier input match @ 50 MHz is only ~ 13 dB, however, we're not interested in a narrow band match — the tuned output network makes strong input matching at 50 MHz impossible *(for me at least)* without additional L and C (narrow band components that we don't want!). I tried a few tapped inductor schemes, however, at VHF, adding turns added significant capacitance and things got ugly fast.

The common gate JFET amp/filter goof-up shattered my expectations. The 4K7 input/output impedance drove instability through unwanted coupling between the inductors. I learned my lesson: at or above 50 MHz, *stick to 50 Ω stages* for stability.

OFF
50
ON
GAIN
TUNE

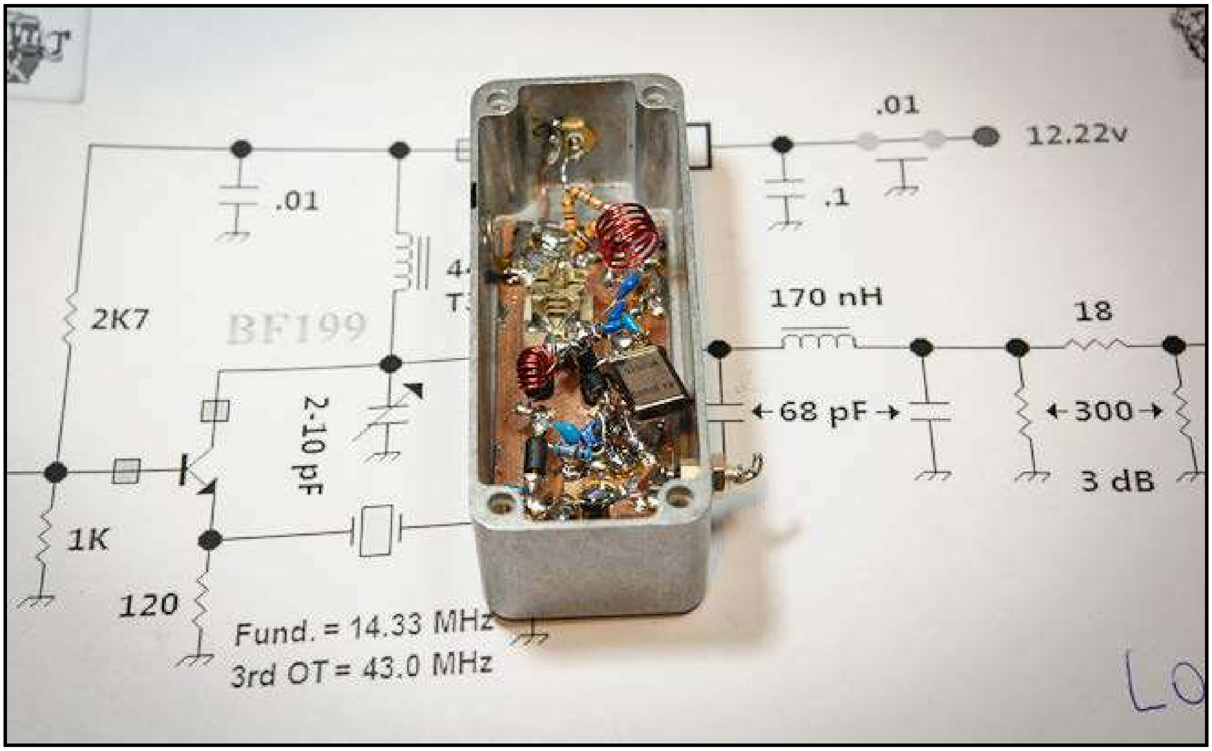
.01
12.22v
.01
.1
2K7
BF199
170 nH
18
2-10 pF
68 pF
300
3 dB
1K
120
Fund. = 14.33 MHz
3rd OT = 43.0 MHz
LO

Section 4. 50 MHz Receiver Pre-amp and Filter

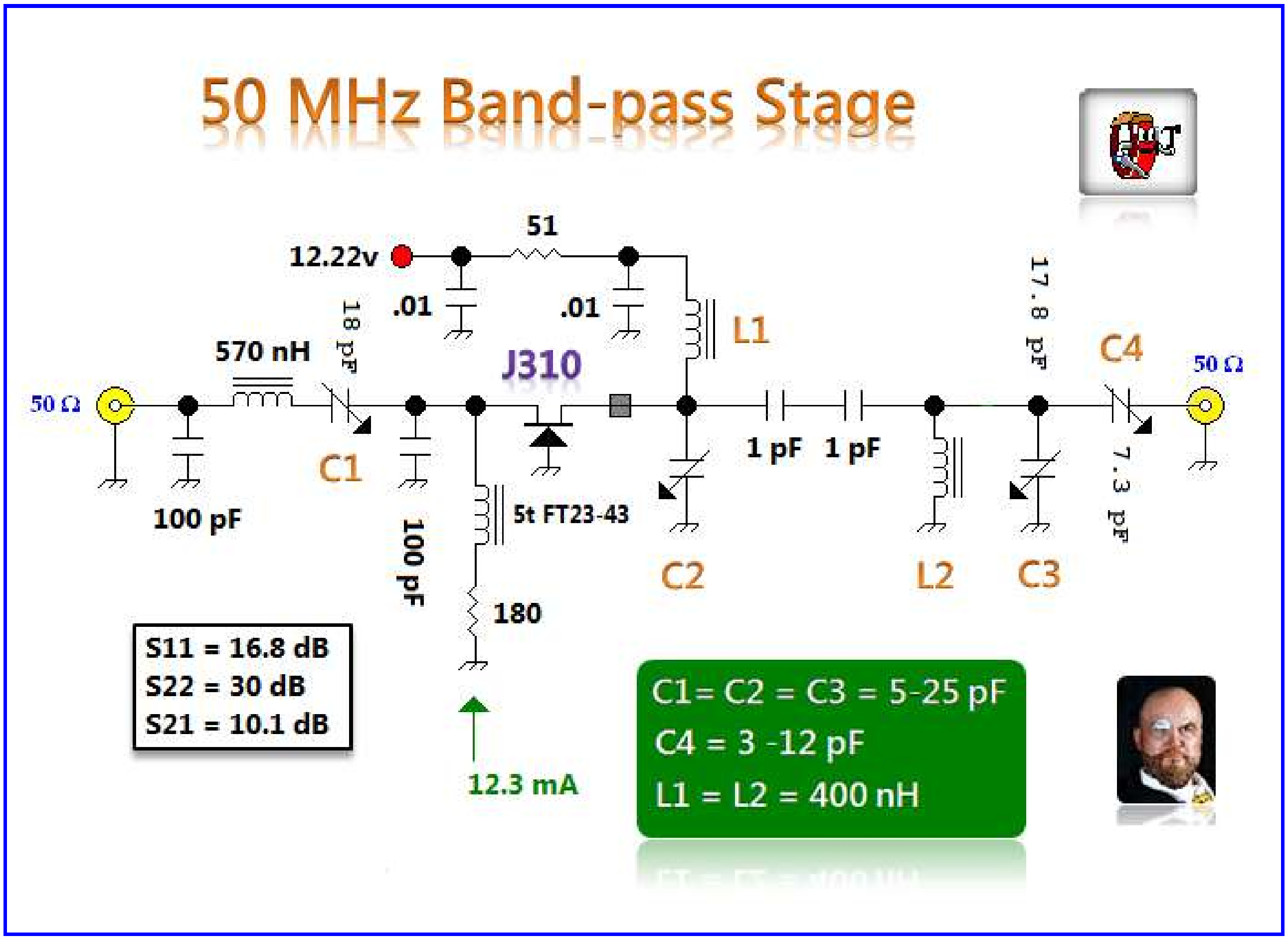
50 MHz Band-pass Stage
12.22v
51
.01
.01
18 pF
570 nH
J310
L1
17.8 pF
C4
50 Ω
50 Ω
100 pF
C1
1 pF
1 pF
7.3 pF
5t FT23-43
100 pF
C2
L2
C3
180
S11 = 16.8 dB
S22 = 30 dB
S21 = 10.1 dB
12.3 mA
C1= C2 = C3 = 5-25 pF
C4 = 3 -12 pF
L1 = L2 = 400 nH

Above — A 50 MHz receiver front end filter with embedded common gate amplifer.

Inspired by the General Purpose Monoband Receiver Front End from Figure 6.69 in EMRFD, I applied inductive and capacitive reactance modeling, DTC08, Ladbuild08 and GPLA08 from the EMRFD ladpac series and built a 50 MHz equivalent.

Connect an antenna to the input and a 50 Ω impedance mixer to the output.

I tested the stage at 50.0 MHz and wound my inductors on T30-10 toroids, although #6 material cores would work okay. You'll find all the measurement techniques in my RF Workbench series 1-4 available though the top-level menu.

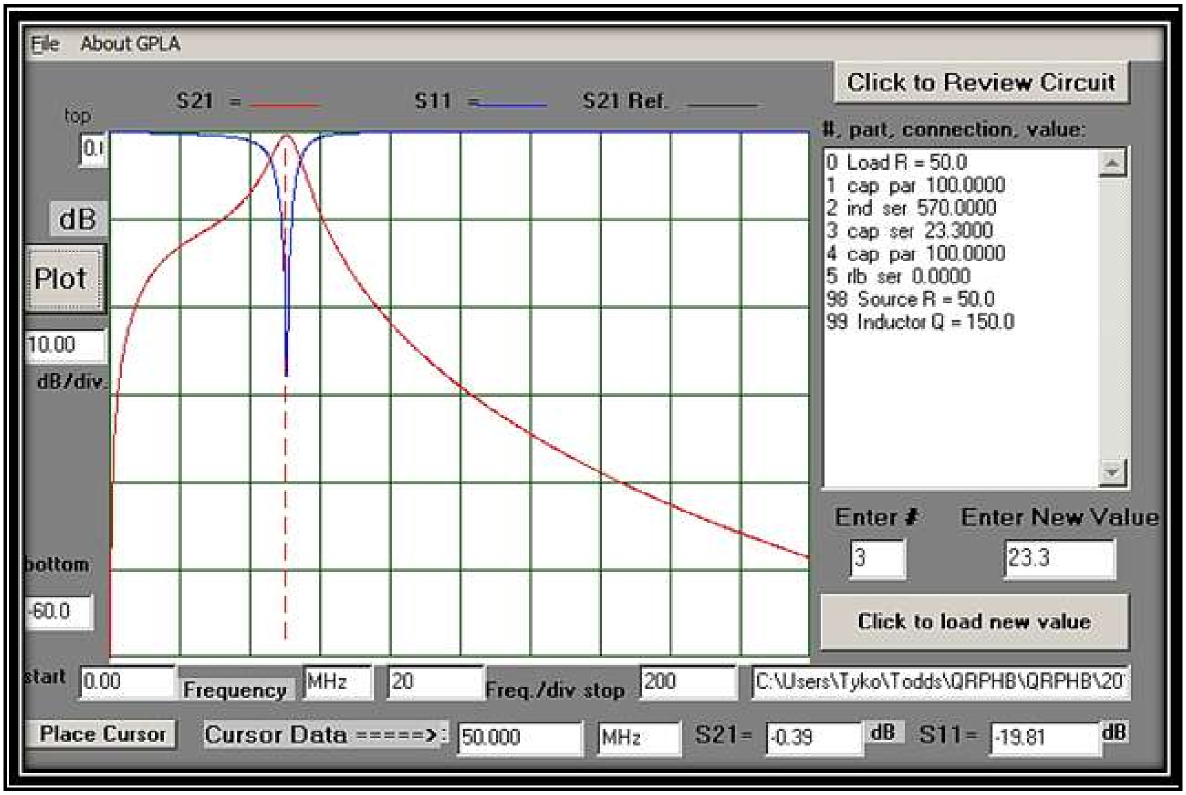

#, part, connection, value:

0 Load R = 50.0
1 cap par 100.0000
2 ind ser 570.0000
3 cap ser 23.3000
4 cap par 100.0000
5 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 150.0

top 0.(
10.00 dB/div.
bottom -60.0
Enter # 3
Enter New Value 23.3
start 0.00 MHz
Freq./div 20
stop 200
C:\Users\Tyko\Todds\QRPHB\QRPHB\20
Cursor Data 50.000 MHz S21= -0.39 dB S11= -19.81 dB

Above — GPLA simulation of the peaked low-pass filter "built" in Ladbuild08.

Wes often employs a peaked low-pass filter and after studying his work, I can see why — way better attenuation than a simple 3 element low-pass filter. The FM broadcast band runs from about 87.5 to 108 MHz and in Russia, they call it "УКВ" (Ультракороткие волны) or ultra-shortwave. At 87.5 MHz, attenuation = 25 dB; pretty good for such a simple filter. At 144 MHz, filter attenuation rises to ~ 40 dB.

This peaked low-pass filter acts as a preselector for the JFET amp that follows it. *Please read the text describing Figure 6.69 in EMRFD for some great notes by Wes.*

In the simulation above, a 50.0 MHz peak response occurred with C1 at 23.3 pF, while in my real circuit, the capacitor was set to ~ 18 pF. Stray L and C + the input Z of the JFET amp caused this variance, but assuredly; GPLA gets you close.

To peak the low-pass filter, I connected a return loss bridge to the input port and tweaked C1 for the lowest possible peak-peak voltage (tuned for the the best return loss which = 16.8 dB in my circuit). *You may also compress or expand the 540 nH inductor to aid tweaking.*

Since common gate amplifiers often exhibit a lower noise figure with a slight mismatch, an S11 of -16.8 dB works fine .I wish I had the gear to set the input match for the lowest possible noise figure — perhaps 1 day I will.

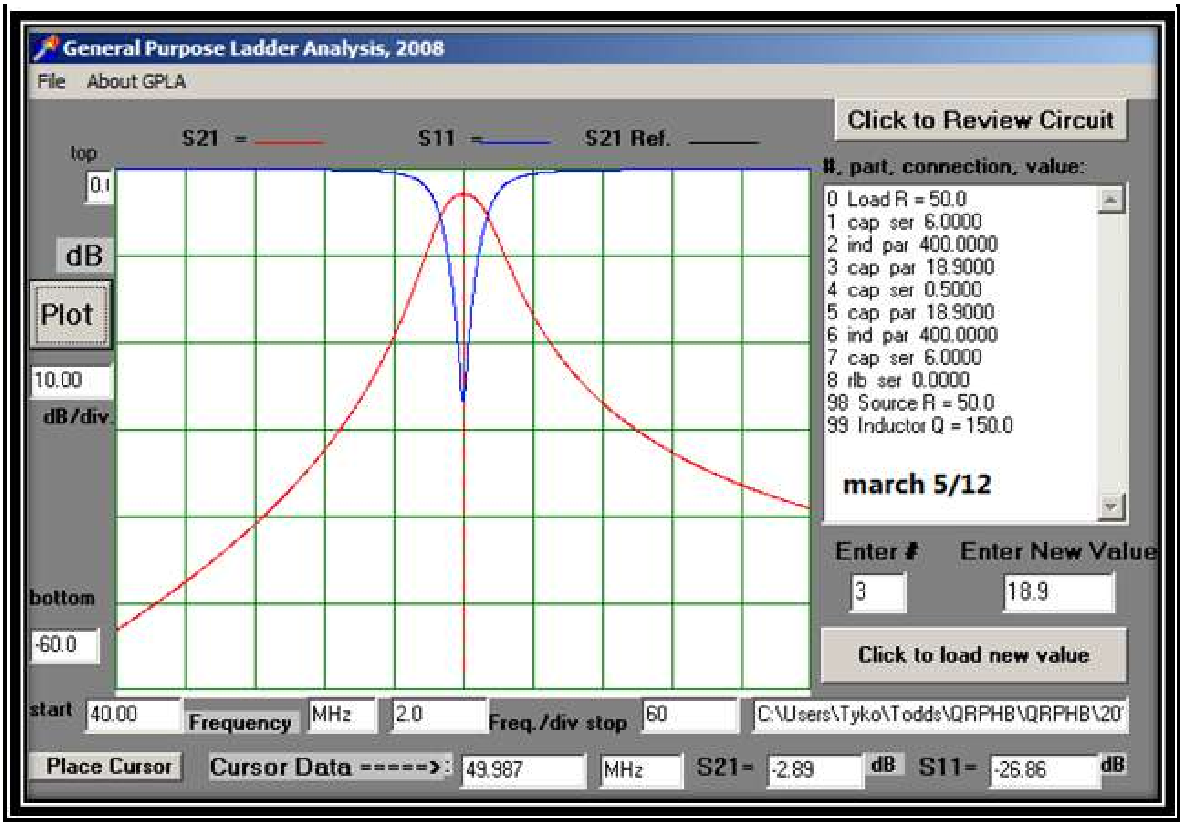

#, part, connection, value:
0 Load R = 50.0
1 cap ser 6.0000
2 ind par 400.0000
3 cap par 18.9000
4 cap ser 0.5000
5 cap par 18.9000
6 ind par 400.0000
7 cap ser 6.0000
8 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 150.0

march 5/12

Enter # 3
Enter New Value 18.9

top 0.1
10.00 dB/div
bottom -60.0

start 40.00 MHz 2.0 Freq./div stop 60 C:\Users\Tyko\Todds\QRPHB\QRPHB\20

Cursor Data =====> 49.987 MHz S21= -2.89 dB S11= -26.86 dB

Above — A GPLA simulation of the 50 MHz double tuned band-pass filter "built" in DTC08. The bandwidth = ~1.8 MHz and varies slightly with the tuning of C4.

I peaked both C2 and C3 with a 50 Ω signal generator and a 50 Ω terminated scope connected to the input and output ports respectively.

Next, I connected my return loss bridge to the output and tweaked C3 and C4 for the lowest peak-peak voltage — the best return loss — and since you tweak 2 capacitors, a strong output return loss delights you.

Finally, I measured the peak-peak voltage with the amp in-line, and after removed the amp and reconnected the 50 Ω cables with a through-connector. Inputting the 2 pk-pk voltages into Applet H on the Design Center web page gave a gain *or S21* of 10.1 dB. I repeated all of the steps above a couple more times to ensure I had set C2, C3 and C4 perfectly.

I found tuning the resonators difficult due to the sharp tuning and wide capacitance range of C1-C3. Assuming your tanks are peaked, the best amplifier gain correlated to the highest input and output port return loss. Have I stressed the importance of a return loss bridge enough?

10.1 dB gain should be enough gain for listening to terrestrial 6 Meter band signals with my 5 element Yagi antenna.

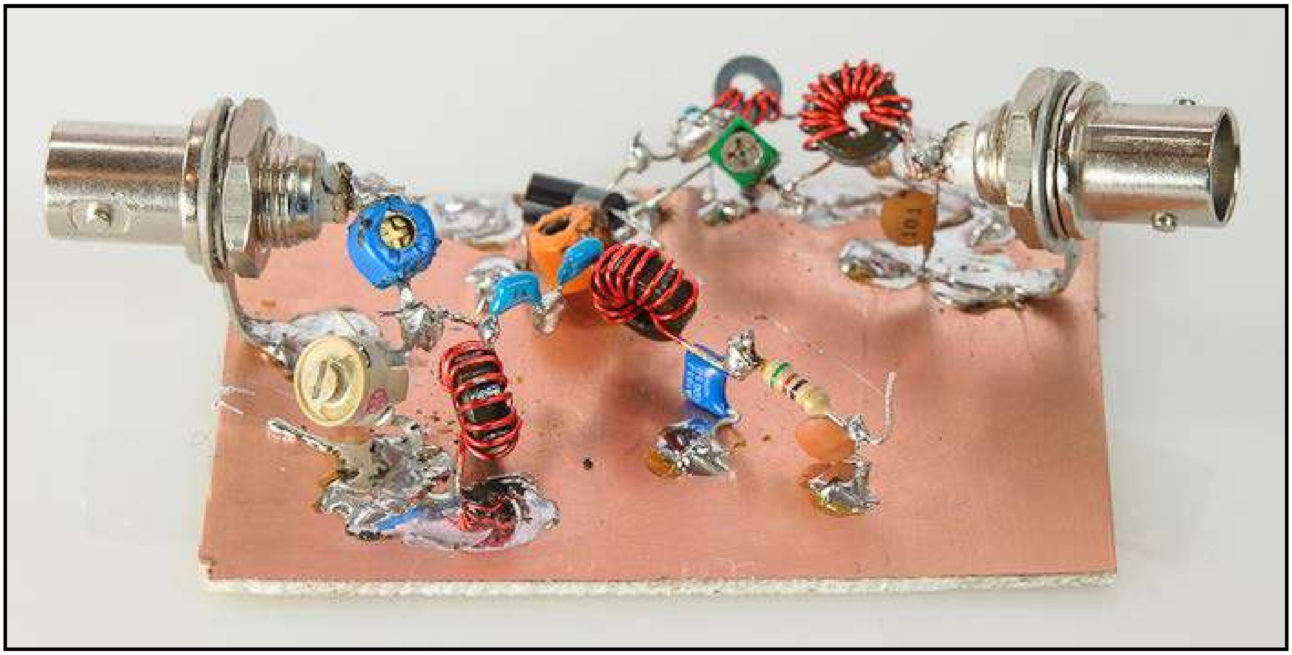

Above — A photo of my protoype 50 MHz pre-amp breadboard. In my "keeper" version, I'll swap in a U310 JFET and bias it for ~15 mA.

Section 5. QRP — POSDATA: Z-Communications VCO Experiment

Looking on eBay, sellers list numerous VCOs, although most are surface mount and go well above VHF. My favorite VCO comes from Mini-Circuits Labs: the POS series. Click for an example: the POS-75. These "plug- in" VCOs come in same package as the SBL-1 mixer and are likely obsolete, but still for sale. If you're building a frequency synthesizer with low phase noise requirements, MCL VCOs seem hard to beat. You can still order them from MCL, but the high product and shipping costs might alarm you.

I've looked for cheaper alternatives and the Z-Comm VCO raises 1 possibility. Last year, I purchased a V149MEM1 device for 5 dollars including shipping. Some experiments follow:

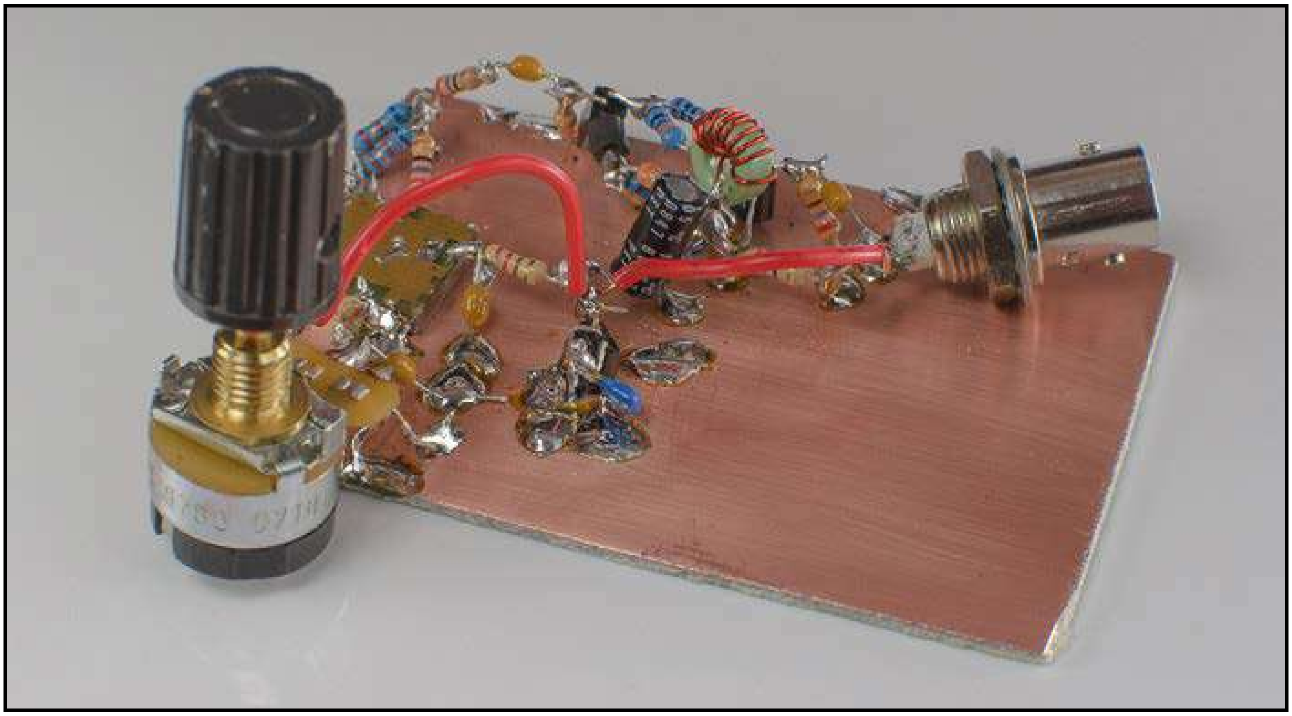

Above — My first breadboard. Lacking the MINI-16 receptable, like with MCL POS VCOs: I turned it upside down and soldered the metal case to my ground plane. If I were to keep this circuit. I wound solder all 4 sides to the copper clad board, plus run some copper de-soldering brade from the bottom to the ground plane, or even cut a square hole and flush mounted the VCO on is back.

While mounting it upside down deviates from the recommendations found on the Z-Comm mounting datasheet, I fiigured that for VHF at least, it might work okay. We desire low inductance grounding, but creativity might allow dead bug construction techniques to work.

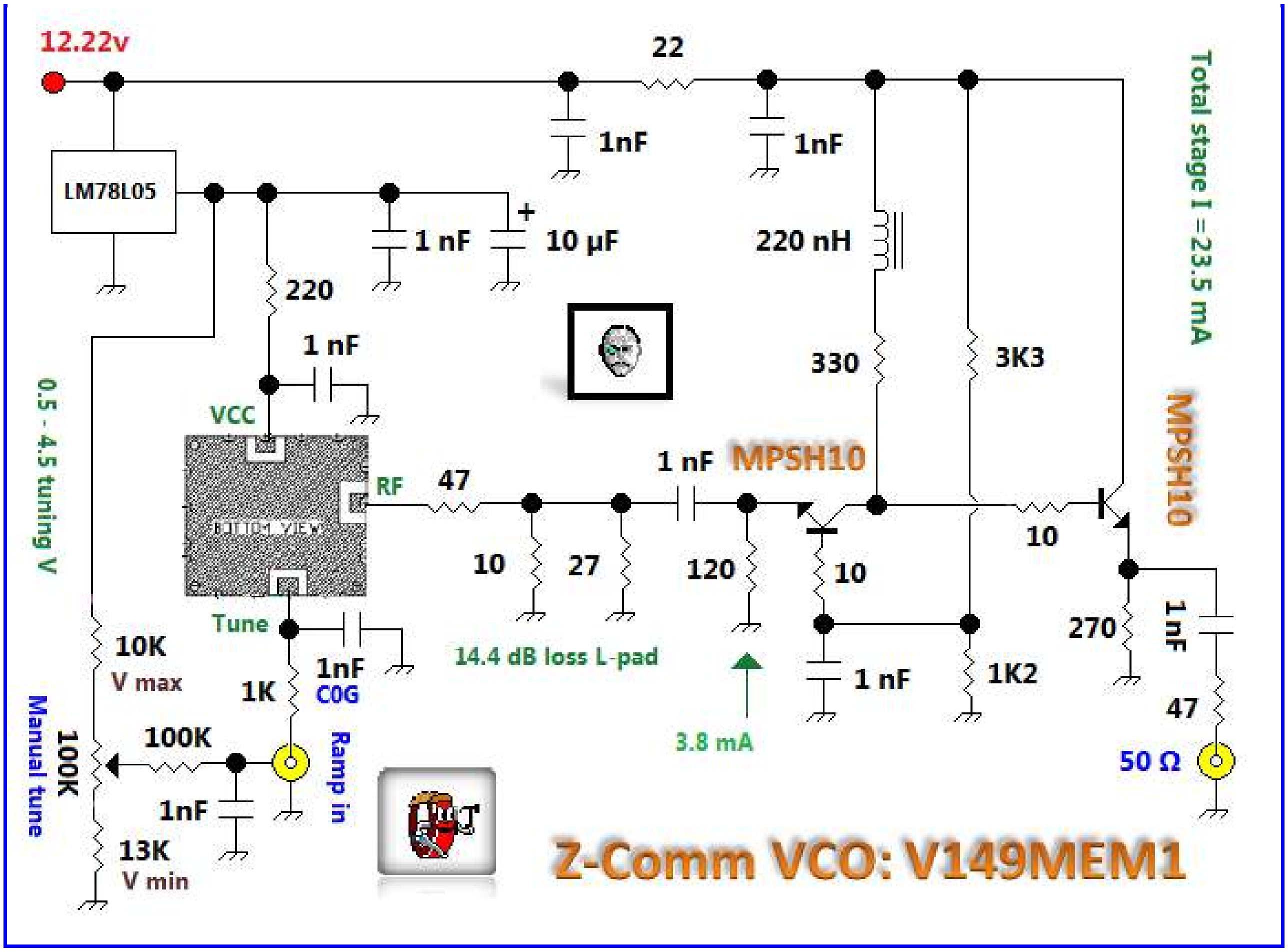
12.22v
22
1nF
1nF
LM78L05
1 nF
+
10 µF
220 nH
220
1 nF
330
3K3
Total stage I = 23.5 mA
0.5 - 4.5 tuning V
VCC
RF
47
1 nF
MPSH10
MPSH10
BOTTOM VIEW
10
27
120
10
10
270
1nF
Tune
1nF
C0G
14.4 dB loss L-pad
10K
V max
1K
1 nF
1K2
3.8 mA
47
100K
Ramp in
50 Ω
100K
Manual tune
1nF
13K
V min
Z-Comm VCO: V149MEM1

Above — My complete VCO. The Z-comm VCOs require at least a 10 dB pad on the output to keep port return loss high. Without a pad, you might see something like this plus boost the phase noise. In my circuit, I applied a resistor L-network with ~ 14.3 dB loss to pad the output and provide a match into a common base amp with an input impedance of ~ 6.8 Ω.

The 2-stage buffer is the brainchild of Bob, K3NHI and I love it. This buffer features a common base stage driving a emitter follower yielding high bandwidth and great reverse isolation. Normally, at VHF, the buffer is followed by more such stage(s), or a MMIC. The 220 nH inductor wound on a T30-12 toroid improves the high frequency response of the common base amplifer — experiment with this L to suit whatever VCO you wish to buffer. The gain of the 2 amp buffer is typically around 9 dB and the return loss at the input and output ports lies under 11 dB, so apply attenuator pads to boost S11/S22 as required.

Click for the scope tracing at 0.5 VDC tuning voltage. Click for the 4.5 VDC tuning voltage 'scope tracing. The harmonic distortion at the lowest tuning voltage = ~ -19 dBc and decreases to -28 dBc at the highest tuning voltage; better than specified. Notice that power decreases as frequency increases. All the commercial VCOs I tested do this. A higher fT amp like the PN5179 or other BJT might be a better choice to offset the power change versus frequency *contribution* of the buffer/amp.

For a sweep circuit, I would mix this VCO with another low level, single frequency VCO with its current controlled by a downstream leveling circut to derive a flat amplitude over the range of the VCO. I plan to try the Z- Comm V150S015 in such an arrangement to make a 70 - 150 MHz VCO for sweeping.

Please refer to the datasheet for the pin out on the Z-Comm VCO: I chose the pinout shown in the schematic to make an efficient drawing. The two 10 Ω resistors in the buffer/amp snub UHF oscillations first measured by Bob and confirmed by me. Ferrite beads might work as alternates.

Section 6. Miscellaneous Photos

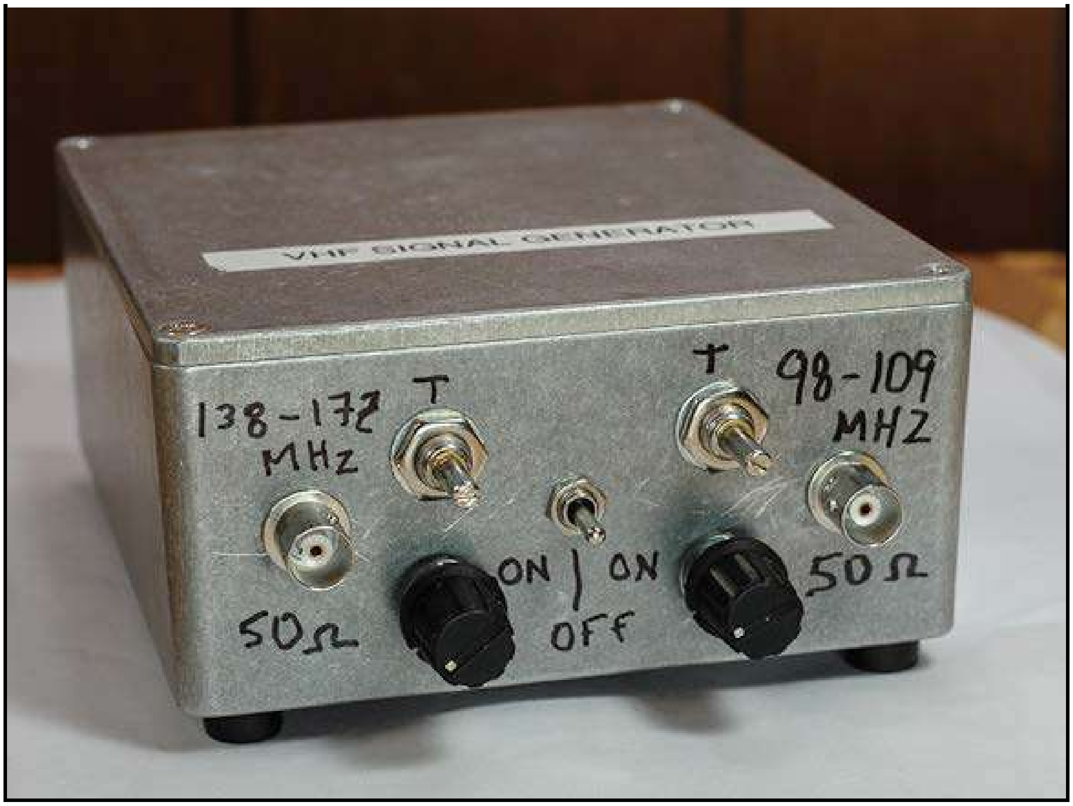
VHF SIGNAL GENERATOR
138-172 MHz
T
50Ω
ON | ON
OFF
T
98-109 MHZ
50Ω

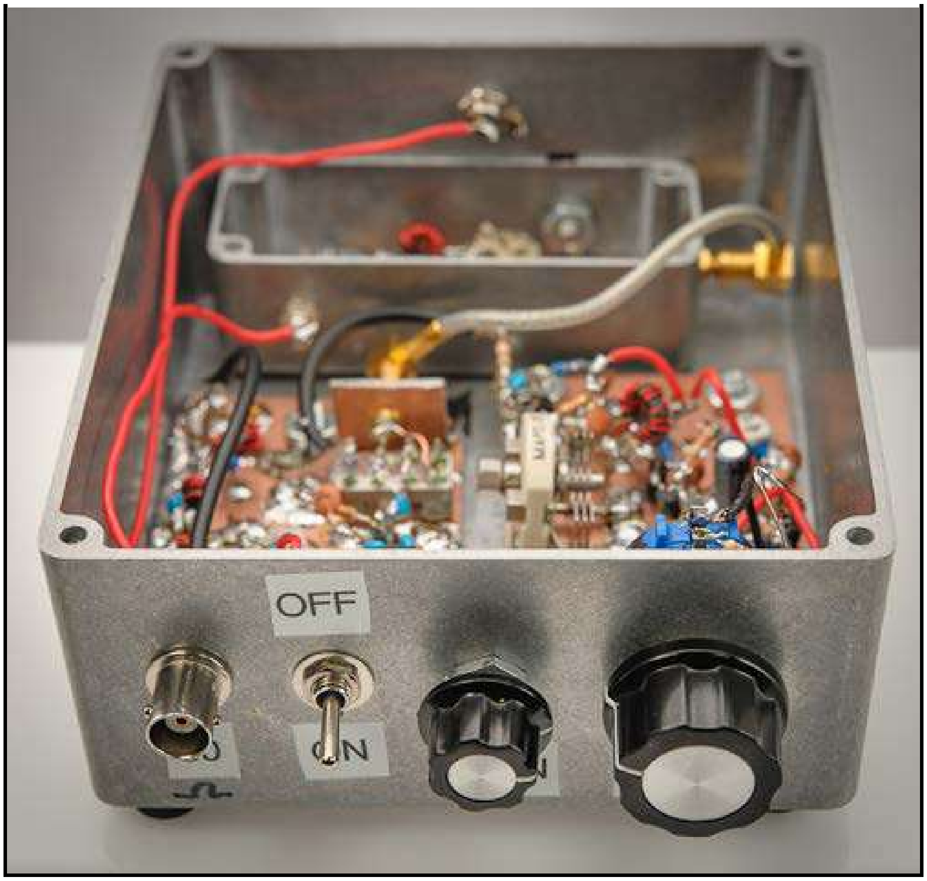
OFF
ON

MALE TERMINATOR 50Ω

RF — Test and Measurement

HF Ragbag

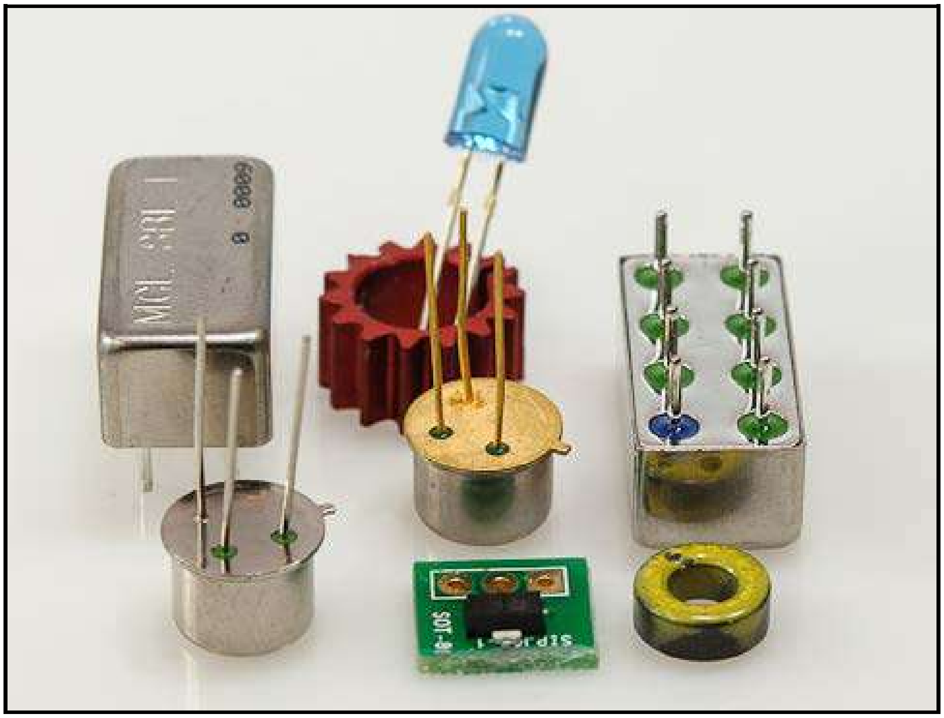

It's easier to present short topics on catch-all web pages — HF Ragbag shows some 2012 non-VHF experiments in no particular order. I also share thoughts on circuit building and writing: we can think and work better.

1. Comments from the Workbench - The Need for Clarity

On Building

In 2012, I boosted my circuit and writing quality to improve your experience: a genuine, return-to-basics approach in amateur, component-level electronic design. As possible, RF circuits will feature 50 Ω input and output ports — totally adopting a 50 Ω environment — for I'm convinced this is the best way to go. The 50 Ω building and measuring standard offers much: an easy-to-interface modular approach; 10 dB improved sensitivity over a 10X 'scope probe and if wanted, measurement with commercial or homebrew test equipment such as a spectrum analyzer, network analyzer or RF power meter.

Like many, I started out by collecting and copying circuits with little emphasis on true understanding. I wanted a completed circuit — quickly as possible — failing to develop my design skills. Without design skills honed by studying and properly measuring our circuits, we bide in hit-and-miss electronics — a frustrating repetition of trial and error, over and over again. We ought to adopt the attitude and thinking of engineers while keeping our design work —including the math — fun. However, embracing scratch-homebrew electronics with the overall goal of trying to understand each stage takes effort. "There is no substitute for hard work" wrote Thomas A. Edison. Scratch homebrew involves reading, simulating, collecting parts, mastering new techniques and building or buying test equipment.

This is more than knack, an abused noun that often means "hack". Our key tasks: to measure, analyze and understand the circuits we copy or create takes patience and practice. Dissecting circuits to understand their function means to hypothesize and reflect — to apply science on paper, with software, and finally, through careful bench experiments. Often we lack the math skills or test equipment to fully investigate some aspects of our circuits, but try our best: measure what we can measure, seek help and grow. I hope this site shows our hobby can be less about making stuff and more about the rewards of actual design: an authentic, personal journey to get better at something you love.

I've never been much of a kit builder; it's too much like Max Klein's Paint by Number for my tastes. But kits dominate HF QRP homebrew and may offer a cost effective way to make gear; especially test gear. Stuffing parts in a printed circuit board won't teach you much about design, but might get your feet wet. Some people remain perfectly happy building kits or madly copying circuits — all the power to you! Do whatever you want. One day you might awaken, but don't worry; I won't try to goad, or convince you.

My favorite builders include people over 50 who suffer the often crippling symptoms of 'appliance apathy' — an epiphany reminds them why they first got into radio: homebrew experiences. Maybe a crystal radio set, or a simple superhet receiver they breadboarded long ago. *Then they come back full circle*; like a loop antenna. Oh-boy — "Bob" rediscovered his radio roots and needs to unleash his creativity and passion to learn and improve. I write for people like Bob. Heck; I am Bob.

On Writing

You'll notice improved narrative writing too: I prefer to read and write crisp statements in short sentences and paragraphs. Brief, yet descriptive text accompanied by ample white space, clear headings and bulleted lists invites you to read on. Plain language writing — simple, clear, writing that is easy to read and understand — signals a refreshing move away from the turgid, word-filled claptrap I learned in grade school. Making your prose easier to read requires greater effort writing and re-writing. My first..to...fourth drafts always suck.

Passive verbs, or nouns and adjectives that function as verbs with no clear subject confuses readers and boosts wordiness: I employ active verbs to invigorate my writing — active verbs connote me or some else performing an activity you can visualize or feel. Actions that may inspire, persuade, or even vex you! Ours' is an emotional hobby.

RF electronics contains rich amounts of jargon. Of course, we must learn some jargon to communicate our ideas as hobbyists, but writing jargon to impress, or to place yourself above others lacks humility and alienates people. Do you know anyone who likes being talked down to? The first step towards becoming humble is to admit you're not humble and then work on it — and I'm working on it.

Although I enjoy writing about electronic experiments, I'm not sure it's worthwhile — Does anyone actually design circuits anymore? Well, back to my 1970's-style analog experiments...

2. Magnitude Only Scattering-Parameters

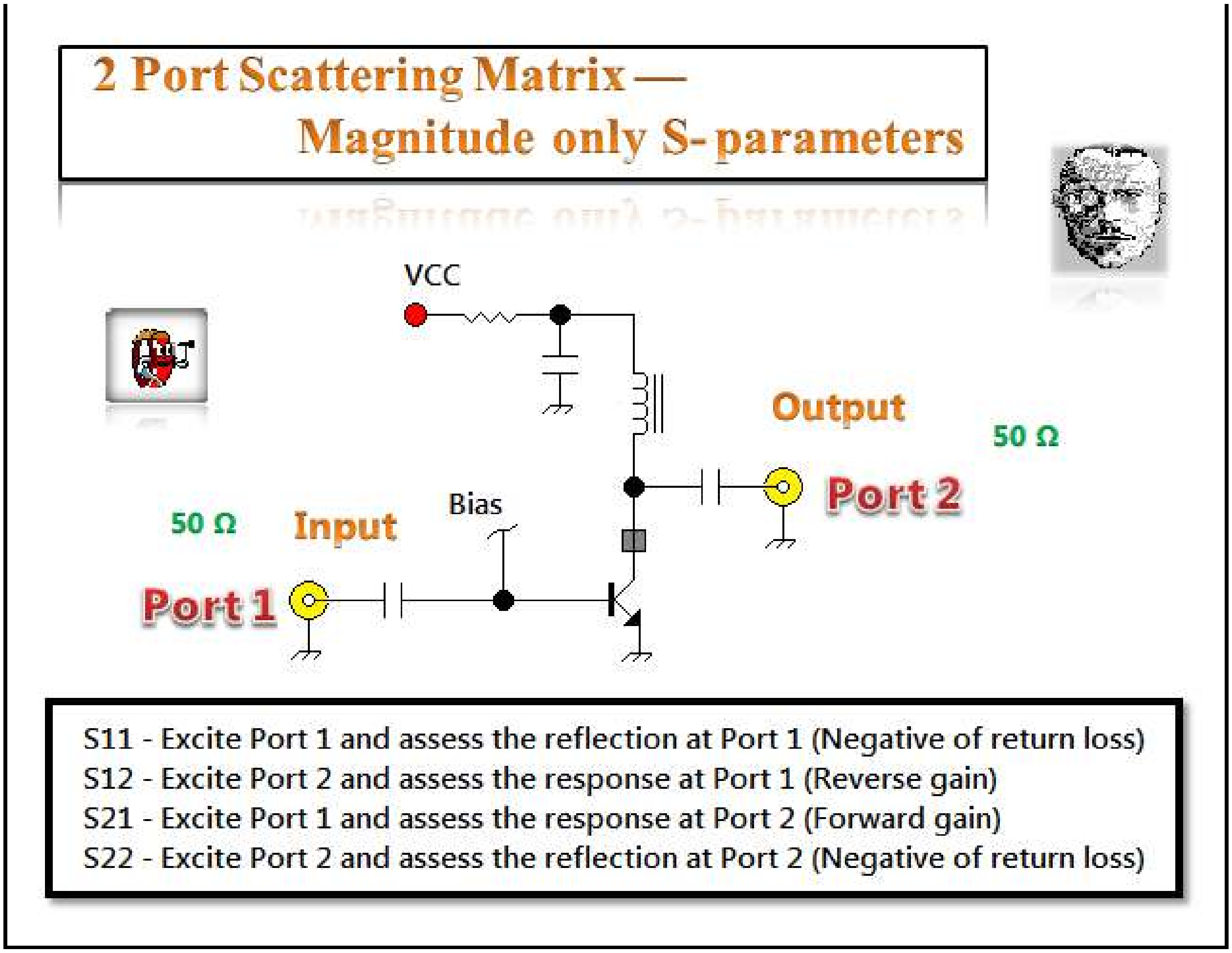
2 Port Scattering Matrix —
Magnitude only S-parameters

S11 - Excite Port 1 and assess the reflection at Port 1 (Negative of return loss)
S12 - Excite Port 2 and assess the response at Port 1 (Reverse gain)
S21 - Excite Port 1 and assess the response at Port 2 (Forward gain)
S22 - Excite Port 2 and assess the reflection at Port 2 (Negative of return loss)

Above — A simple model describing the S-parameters displayed on QRP / SWL HomeBuilder in a Class A amplifer with 50 Ω ports.

S-Parameters

Any device with 2 connectors may be modelled at AC for a specified frequency with just 4 scattering parameters: forward gain, reverse gain + input and output impedance (match or VSWR).

S-parameters address voltage ratios: comparing the amplitude of different signals at the 2 ports. For example, S21 is the magnitude of forward gain and equals the ratio of output voltage to input voltage.

S-parameters are vectors; a mathematical quantity that may be visualized as an arrow anchored at 1 end that pivots around its base. The length of the arrow represents magnitude while the angle it makes with another vector or its base line decribes its phase in degrees. In addition to phase and magnitude, S-parameters allow analysis of gain, stability, complex impedance (resistance + reactance), admittance and other vector quantities.

Measure S-parameters with all ports terminated in a 50 Ω impedance.

Some of us worry only about the gain, losses or "match" in our 50 Ω circuits and could care less how the signal phase changes as it passes through our amplifiers or attenuators. I express only S-parameter magnitude in logarithmic form (dB) and take this *Über simplified* approach because builders can easily measure S11, S12, S21 and S22 on a 50 Ω test bench with a small staple of bench accessories + a 50 Ω 'scope or detector.

Topics like matrix theory, vector math, the "jay" operator, converting S-parameters into other matrices, Smith charts etc. may turn *off* the average amateur designer. You advanced readers, may raise your 2 port network skills by visiting better web sites + reading books, simulating with SPICE, or better yet, measuring your port parameters with a vector network analyzer.

---

3. More on Feedback Amplifiers (FBA)

Many builders (myself included) copy feedback amps rather than design their own. By tweaking the emitter current, shunt and series feedback while measuring S11 and S22, plus simulating with a program called FBA08.exe, I've learned it's possible to design good feedback amps FBA08 is 1 of the Ladpac programs that ships with EMRFD.

I wanted a FBA with ~35 mA emitter current for improved IMD and low distortion on strong signals. Such an amp might follow a diode ring mixer in a receiver I.F. chain.

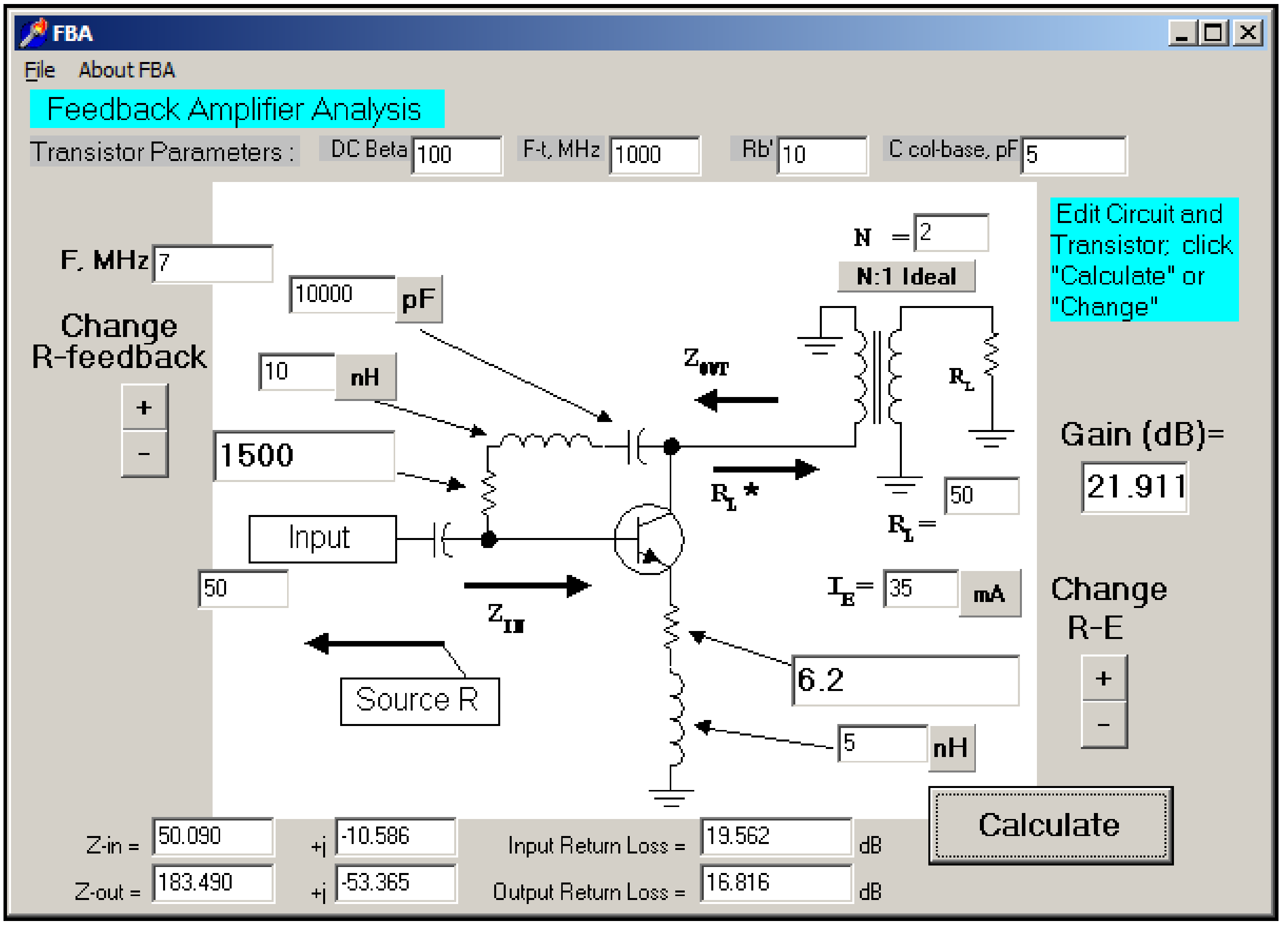
FBA
File
About FBA
Feedback Amplifier Analysis
Edit Circuit and Transistor; click "Calculate" or "Change"
Change R-feedback
+
-
N:1 Ideal
Z_OUT
R_L
R_L *
Z_IN
Input
Source R
Change R-E
+
-
Calculate

Transistor Parameters : DC Beta 100 F-t, MHz 1000 Rb' 10 C col-base, pF 5

F, MHz 7

10000 pF

10 nH

1500

Input 50

N = 2

R_L = 50

I_E = 35 mA

6.2

5 nH

Gain (dB)= 21.911

Z-in = 50.090 +j -10.586 Input Return Loss = 19.562 dB

Z-out = 183.490 +j -53.365 Output Return Loss = 16.816 dB

Above — My 7 MHz FBA set up. Wes, W7ZOI suggested using 5 nH as the default emitter inductance and 10 nH for the default collector to base inductance in FBA08. These represent stray inductances in your circuit breadboard. Emitter inductance affects the input impedance more.

Zin = input impedance. Zout = output impedance.

Explore this program to learn how changing the emitter resistor, feedback resistor and emitter current affect the input and output return loss.

Adjusting the transformer N and load values only affect the calculations for Zin because this app wasn't really designed to crunch output transformer Z ratios for Zout manipulation. The default output Z = 200 Ω and thus for the N parameter with a 50 Ω RL, RL is multiplied by N^2 to set the amplifer load impedance.

From FBA08 simulations: with an emitter current of 35 mA, my series feedback = 6.2 Ω and shunt feedback = 1500 Ω.

I chose a simple voltage divider bias network to set up the ~35 ma and ensure reasonable temperature stability.

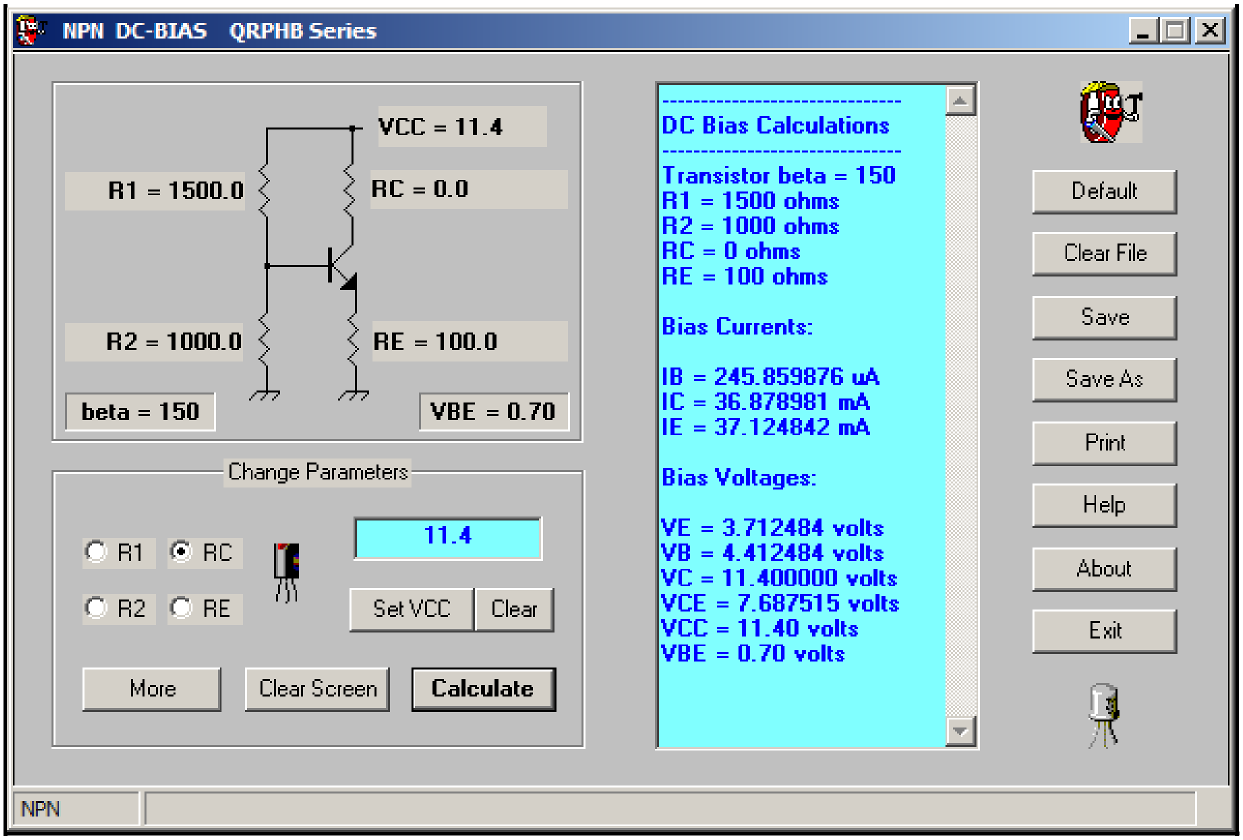

VCC = 11.4
R1 = 1500.0
RC = 0.0
R2 = 1000.0
RE = 100.0
beta = 150
VBE = 0.70

11.4

```
--------------------------------
DC Bias Calculations
--------------------------------
Transistor beta = 150
R1 = 1500 ohms
R2 = 1000 ohms
RC = 0 ohms
RE = 100 ohms

Bias Currents:

IB = 245.859876 uA
IC = 36.878981 mA
IE = 37.124842 mA

Bias Voltages:

VE = 3.712484 volts
VB = 4.412484 volts
VC = 11.400000 volts
VCE = 7.687515 volts
VCC = 11.40 volts
VBE = 0.70 volts
```

Above — Choosing the emitter and nearest standard value bias network resistors to set up ~ 35 mA emitter current with a program. Actual biasing requires you to set up the correct emitter current + establish reasonable temperature stability.

Click and scroll to #5 for some basic transistor biasing notes. While this supplement shows a simple method for stable bias networks, it probably understates that **Beta** bias stabiility is a function of the ratio of RB to RE, where RB = the 2 base resistors in parallel. The lower the ratio the better, but then more input power is lost in those resistors. A higher ratio reduces stability but wastes less input power — another trade off we must negotiate! See Ken Kuhn's web site for thorough, expert-level information on voltage divider biasing your BJT amplifiers.

I use NPN DC BIAS, a program I wrote, however, Wes included 1 in the Ladpac software called Biasnpn08.exe that's also good. Determine the VC for the program by first multiplying the value of your decoupling resistor by the emitter current in Amperes to learn the voltage drop across the R. Then, subtract that voltage drop from your power supply voltage: 12.22V - (.0371 A X 22 ohms) = 11.4 VDC.

Our software allows you to pick approximate base and emitter resistor values to set up a desired current in your amplifier breadboard, but you must still choose reasonable values for temperature stability. Tweak them as needed, or choose some other bias method such as a current source. Let's move to the bench...

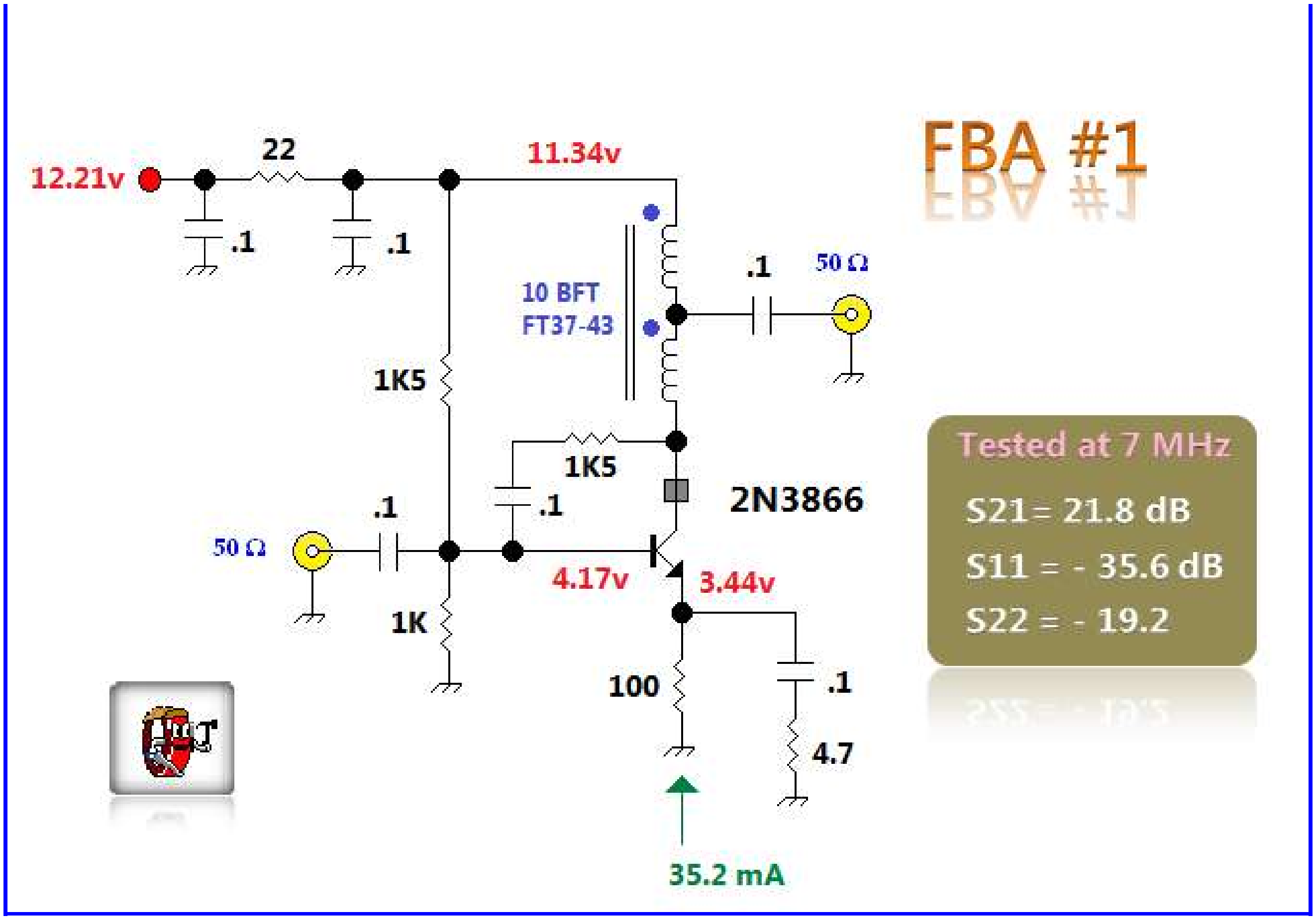
FBA #1
12.21v
22
11.34v
.1
.1
10 BFT
FT37-43
.1
50 Ω
1K5
1K5
.1
2N3866
50 Ω
.1
4.17v
3.44v
1K
100
.1
4.7
35.2 mA
Tested at 7 MHz
S21= 21.8 dB
S11 = - 35.6 dB
S22 = - 19.2

Above — My 7 MHz FBA with some measured S-parameters. On the bench, I lowered the 6.2 Ω series resistor to 4.7 Ω because the voltage divider bias network also affected Z in. I tried 3.3, 4.7 and 5.7 Ω resistors for series feedback and settled on 4.7 Ω since an S11 of -35.6 dB wins the prize!

The S22 of -19.2 dB bettered the value predicted by FBA and seems quite acceptable considering we normally follow a FBA with a 6 dB pad that raises the output return loss another 12 dB. FBA08 gets you close, however, only bench experiments will realize the amplifer you want, and sometimes, a decent S11 and/or S22 may elude you.

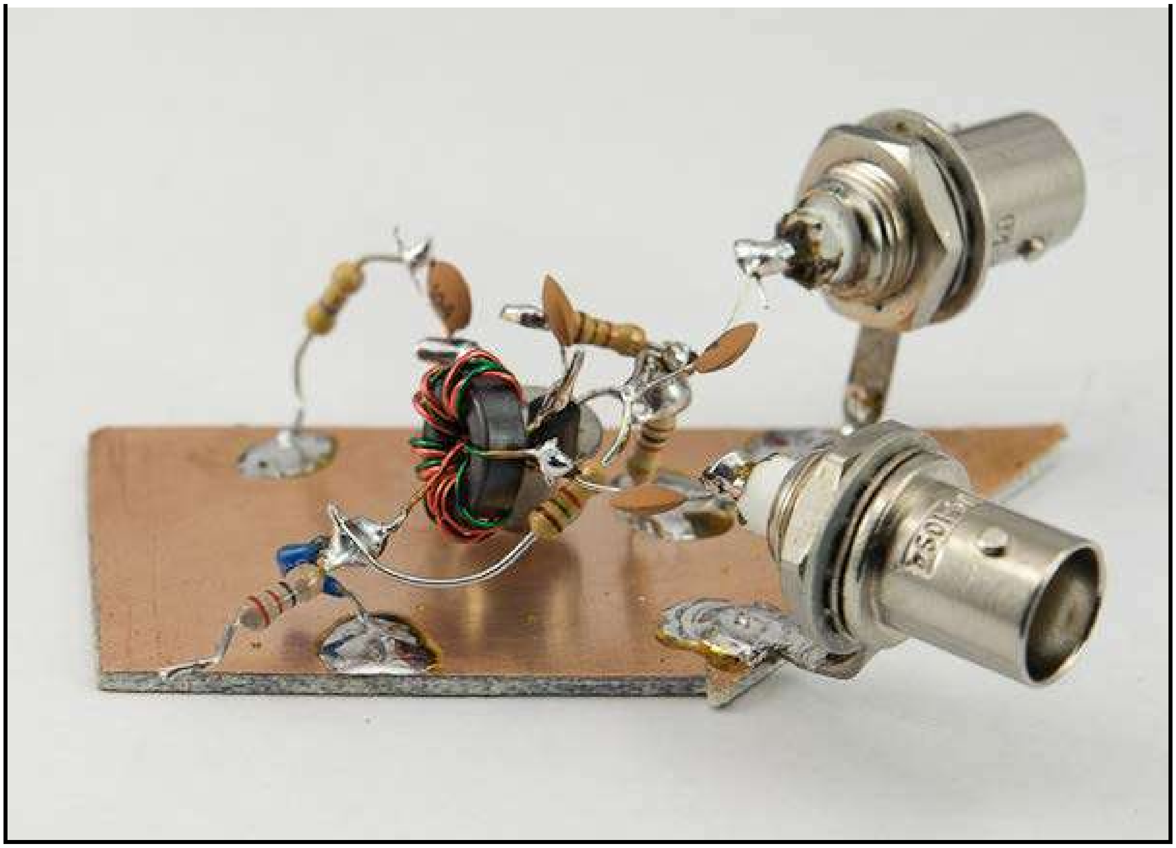

Above — A photo of the 35 mA feedback amp built on scrap of copper clad board.

Parallel Transistor Feedback Amp

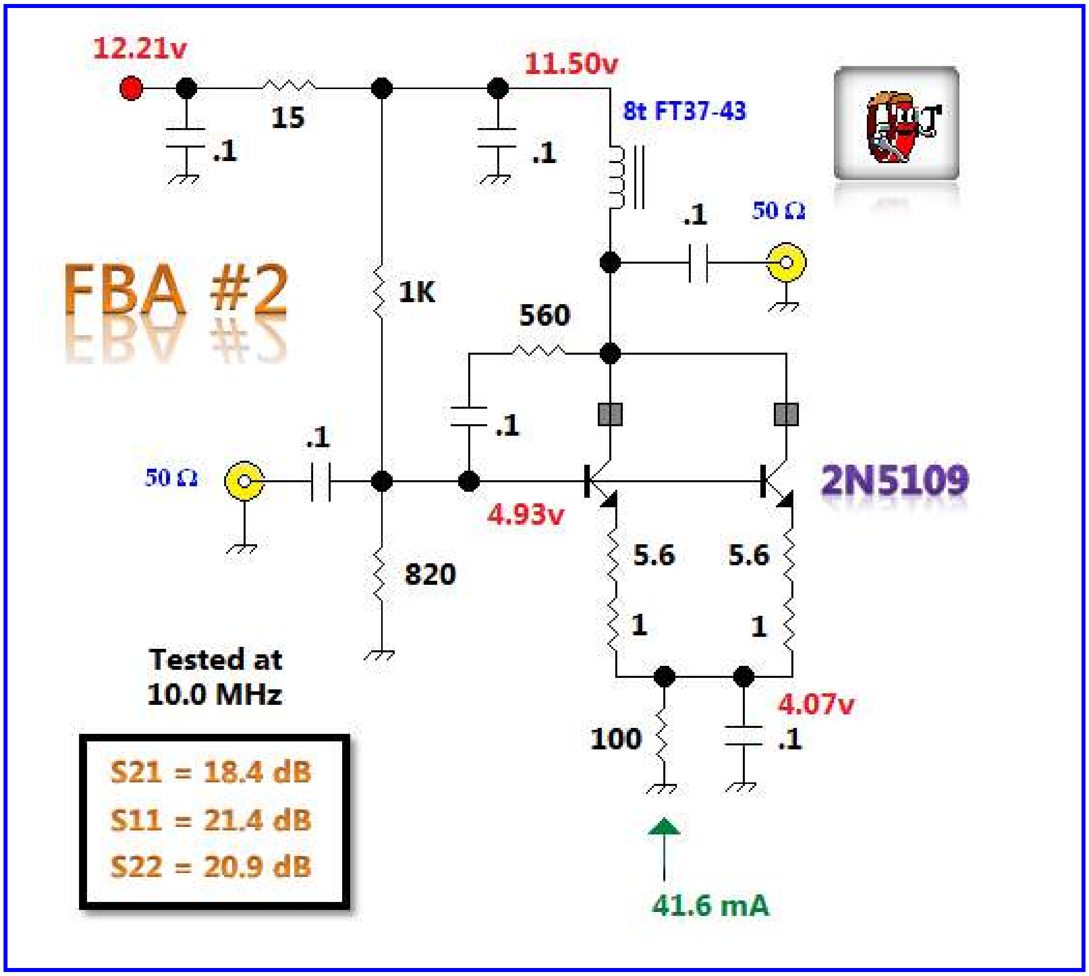

12.21v
15
.1
11.50v
.1
8t FT37-43
.1
50 Ω
FBA #2
1K
560
.1
.1
50 Ω
4.93v
2N5109
5.6
5.6
820
1
1
Tested at
10.0 MHz
4.07v
100
.1
S21 = 18.4 dB
S11 = 21.4 dB
S22 = 20.9 dB
41.6 mA

Above — A feedback amp with two 2N5109 transistors wired in parallel. Click for a photograph of this prototype. I lacked 6.8 Ω resistors and placed 1 Ω + 5.6 Ω to make the needed R for a strong S11.

Originally, I built FBA #2 with a 4:1 Z transmission line transformer, but measurements of S22 disppointed me. Later, a L wound with 8 turns around an FT37-43 ferrite toroid drove an S22 of 24 dB, but S11 was only 18.5 dB.

With the amp set up to measure return loss on the input port, I placed a 500 Ω potentiometer in series with a 100 Ω resistor between the collector and base terminals and tweaked the pot to obtain the lowest peak-peak voltage in my 'scope (lowest return loss). After, I removed the pot and measured its resistance at 572 Ω. Finally, I soldered in a 560 Ω resistor and re-checked S11. Perfect. With my goal of at least 20 dB for S11 and S22 obtained, I powered down my bench and took some photos.

Wes, W7ZOI displayed parallel transistor FBAs in EMRFD and other works and recently I noticed Lyle, KK7P employed a parallel NE46134 FBA as a post-mixer amplifer in the Elecraft K3.

Wes wired 2 parallel 2N3904s to avoid using an expensive medium power BJT like the 2N5109. Doing so splits the heat between 2 devices, but does not deliver better IMD performance beyond what is offered by increasing the emitter current. In a typical FBA bias setup, you may measure as much as 10 volts between the collector and emitter terminals and with a supply of 12 VDC + a standing current of 20 mA, the collector dissipation = ~200 mW. This is about maximum for a TO-92 device likes a 2N3904, but only half of maximal dissipation for 2 in parallel.

Then, too, the K3 applies 2 parallel medium power BJTs get power dissipation with an SMT transistor. For strong IMD performance, Lyle and crew are throwing 80 mA or so into the pair — hard to do with SMT parts, so they overcome heat and power dissipation issues with 2 devices. Cool (literally).

Heat sink BJTs when you crank up the emitter current.

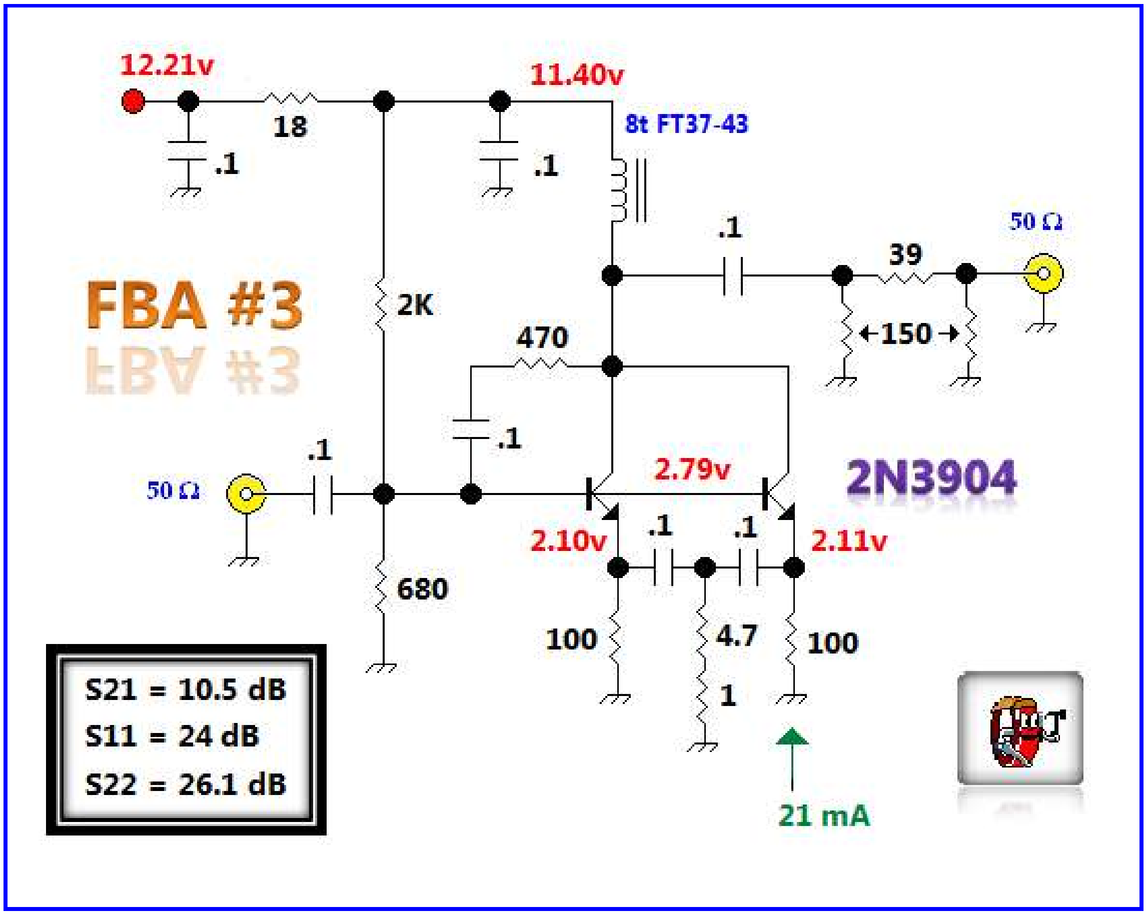

Above — An attempted 2N3904-based parallel feedback amp. Each BJT draws ~21 mA emitter current. Without the 6 dB output pad, the output return loss = 14 dB — I *failed* to realize both a strong (raw) S11 and S22.

The power gain including the 6 dB pad = 10.5 dB. I'll discard this design since it's substandard — without failures, victory may taste bland.

For bench designers, making a parallel FBA where both the raw S11 and S21 are > 20 dB is difficult and bench failures may either frusturate you, or enhance your resolve to succeed. With success, great satisfaction arises and I'm addicted to that feeling.

An FBA bench triumph means you managed to establish the perfect combination of series + shunt feedback, emitter current and the correct output transformer ratio for *that* transistor plus its biasing circuitry — no small task.

A well matched amplifier = a thing of beauty! The fetching trio of high S11, S22 and S21 rewards your efforts and boosts your confidence to experiment further. And so it goes...

Sadly, only a fraction of hobbyists create and evaluate their own circuits.

4. Microphonics in Direct Conversion Receivers

LO = local oscillator or VFO. DC Receiver = direct conversion receiver.

Microphonics are induced electrical responses that arise from a mechanical vibration on the DC receiver chassis or circuitry. The audio amp, acting like a transducer, makes a clicking, or popping noise when you do things like tap the chassis, or unplug components — the disturbance throws out a burst of DC voltage that's amplifed by the AF chain and pops the speaker.

We may read or hear inexperienced builders tell us to expect microphonics in our DC receivers — *de trop* folklore strikes again! As a student of EMRFD and those wise designers who live in and around Beaverton, Oregon, I share some of their best tips to decrease microphonics in your DC receiver projects. " Keep Your LO From Radiating to the Outside World and Keep Unwanted RF from the Outside World Getting Into Your Receiver" seems the appropos title for the bulleted notes that follow:

Read EMRFD pages 8.7 to 8.11 and then build or apply the presented examples. Wisdom is experiential; it comes by doing, not just reading. It's no accident that Chapter 8 author Rick, KK7B mentions microphonics and hum in the same section. I've never read more thorough notes regarding DC receiver nuances anywhere; for example, did you consider that an ungrounded air variable capacitor shaft poking outside the LO box will radiate LO signal per Figure 8.18 ? I didn't in my early days.

Stick your LO in a RF-tight enclosure with RF-grade connectors and coax to patch the AC signal to the product detector. Bypass RF with feedthrough capacitors on any DC voltage lines that pass through the LO chassis wall. Many enthusiasts have only operated kitted or homebrew DC receivers where the LO and receiver guts lie on the same circuit board — this *ensures* microphonics. Wes and Roger built the historic Ugly Weekender VFO, transmitter and receiver in seperate boxes — resulting in low microphonics and no pulling of the VFO when keying the transmitter. Nothing in that 2 part QST series was done by accident. Read these articles to "go to school".

Reciprocally important; keep unwanted outside world RF from getting inside your DC receiver! Apply resistors plus capacitors, or inductors plus capacitors to decouple and bypass RF from moving along on your DC voltage lines, key line, microphone cables etc.

Keep product detector port-to-port isolation high. Typically, we employ double balanced mixers to obtain high port-to-port isolation. I cover mixer balance on this page . For diode ring mixers, measure the return loss of the circuits that you connect to the product detector LO, RF and AF ports — I aim for 20 dB or greater return loss on my LO output, RF output and AF amp input circuits to help preserve the product detector balance and keep port isolation as high as possible. Along with 50 Ω amplifers, attach attenuator pads, AF diplexers, or whatever to help increase port return loss as required.

LO-RF port isolation: Consider a common gate amp with an output matching network to get a high output return loss (S22). The common gate amp provides strong reverse isolation without adding much noise.

Avoid end-fed wire antennas where there is a strong antenna field right next to your radio.

I favor sturdy chassis/cabinets with rubber feet. Homebrew copper clad board or die-cast aluminum cabinets may work best as joints and screwed connections won't corrode. *This is a weak recommendation.*

Double the LO frequency or apply a heterodyne VFO. Often microphonics arise in the VFO tank. EMRFD cover this well. If the VFO operates at a significantly different frequency than any of the signals reaching the balanced mixer, leaked LO won't cause as much havoc as when a LO tank is tuned to the mixer RF port frequency.

Despite proper techniques, RF can exit via the antenna port and make its back into our rig through power supply cables (often modulated by our house AC electricity). In some cases, we require special power supply decoupling to decrease hum and microphonics. We might need to add a common mode choke (+/- capacitors) for common mode noise suppression in addition to the usual differential mode choke(s) and capacitors. In my main shack power supply, I run a common-mode choke plus I soldered a 0.01 uF capacitor across each bridge rectifier diode to bypass RF.

Some radio operators just run battery power supplies.

Above — Feedthrough capacitors. I prefer hole mount over solder mount parts, however, quality feedthrough capacitors of any kind tend to be expensive. As a hobbyist, I'm constantly searching for bargains and when I find 1, I'll purchase a bunch to meet my current and future needs.

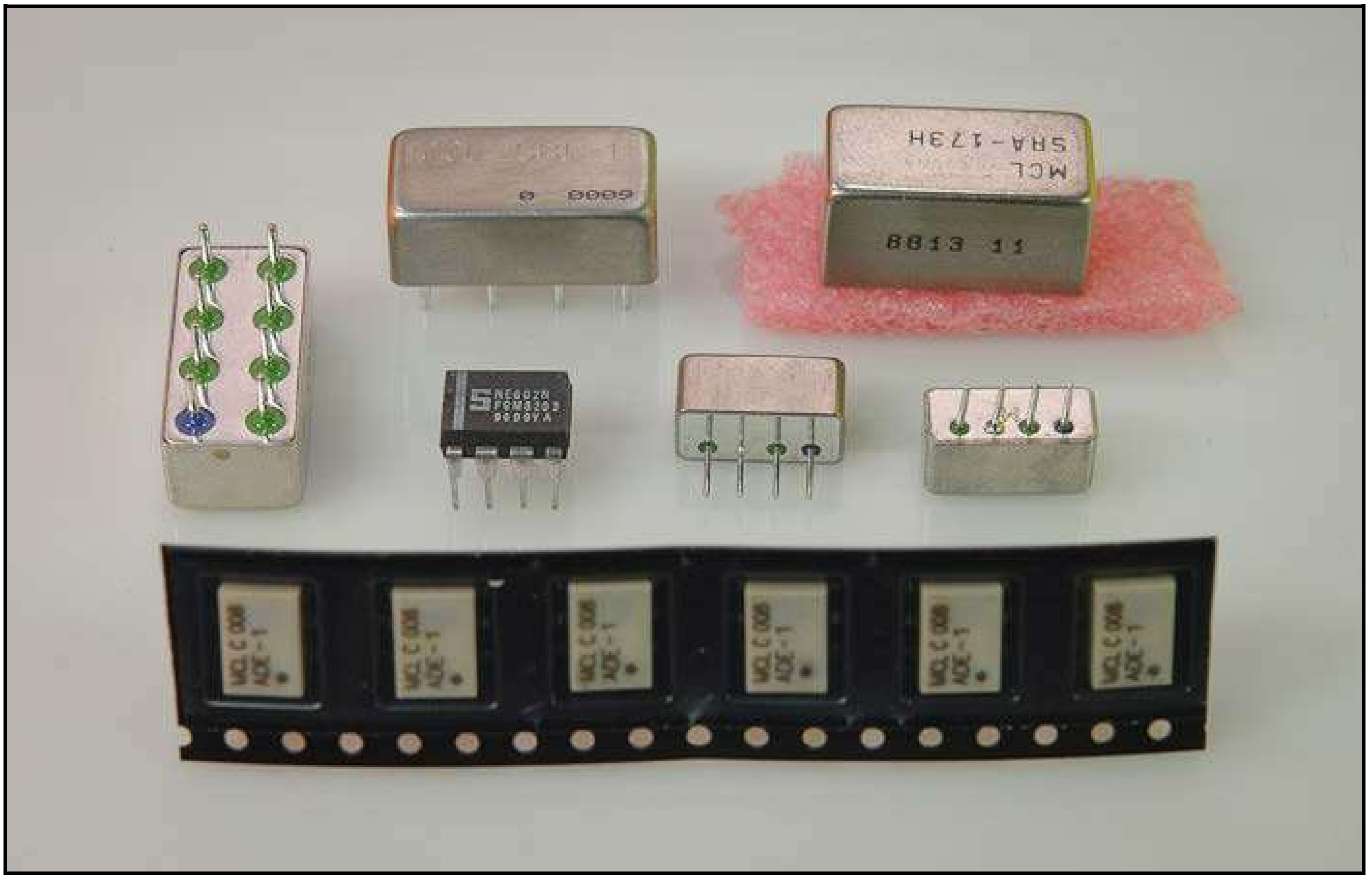

Above — Some double balanced mixers from my collection: ADE-1, NE602, TUF-1, TUF-2, SBL-1 and a SRA-173H; a MiniCircuit Lab's Level 17 diode ring mixer.

You owe it to yourself to listen to a DC receiver designed and built to reduce microphonics — music to our ears.

5. Some Experiments with RF Bypass Capacitors

Introduction

Bypass implies a low impedance path to ground for RF at 1 or more frequencies. After reading EMRFD pages 2.28 - 2.31, I decided to explore this subject for the first time. My bench measurements from Spring 2012 punctuated how little I knew about RF bypass and I share these notes as something for me and others to build on.

In these experiments, I

1. observed the self resonant frequency of MuRata RPE Series, 50v, 5% capacitors with X7R temp compensation at 0.1, 0.01 and 0.001 µF.

2. examined a wire short, plus 1 and then 2 Johanson Dielectric 0.01µF, 50v, X7R, size 1206 chip capacitors.

3. tested a 0.1 µF RF cap plus a parallel 2.2 µF electrolytic capacitor to look at parallel resonance side effects.

4. attempted to reduce the Q of some parallel capacitors to reduce unwanted high impedance peaks.

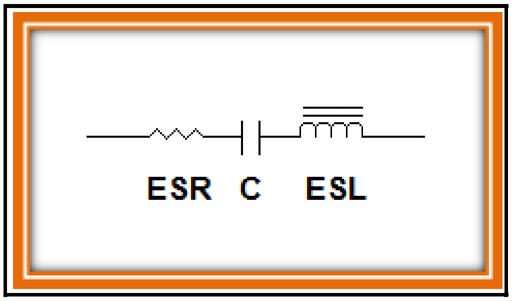

Above — The frequency dependent components of a capacitor are shown in this capacitor equivalent circuit schematic; essentially an RLC network. Engineers use mathematical formulae to describe the components of a capacitor along with reactance and with this math, you might derive an unknown variable from available data so it's worth diving into on your own.

ESR or equivalent series resistance = the sum of all of a capacitors' resistive components. Expressed in ohms, ESR acts like a resistor in series with the capacitor. Normally we desire capacitors with an ESR as low as possible. Consider reading the capacitor datasheets for those your stock and/or searching for information regarding low ESR capacitors on the Internet.

ESL refers to the equivalent series inductance; the sum of all the capacitor's inductive components. This includes lead length in hole-through parts.

In a given capacitor, the series resonant frequency is the frequency where the inductive reactance from the ESL = the capacitive reactance, but since the 2 reactances are 180 degrees opposite in phase, they cancel to drop the impedance to 0 and the capacitor acts like a resistor at its ESR.

The series inductance of a capacitor may be determined using a network analyzer and unfortunately this in unattainable by most average builders. When designing RF bypass with network analysis, we strive for a low impedance over a wide frequency range, although small ripples typically occur.

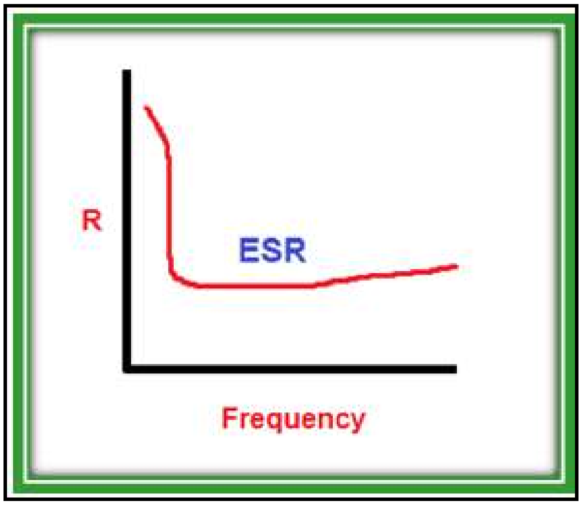

Above — A plot of equivalent series inductance. ESR tends to increase with frequency.

Methods

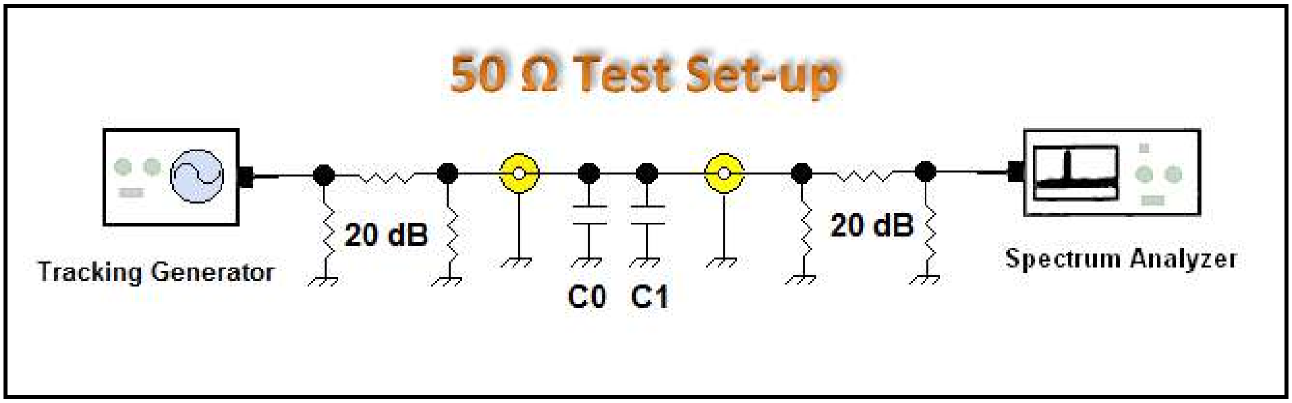

Above — My test set up. I performed all analysis with a tracking generator plus spectrum analyzer. The 50 Ω system used short coax patch cables fitted with BNC connectors with 20 dB attenuator pads before and after the capacitors under test. The capacitors shown as C0 and C1 were soldered on a copper board with short leads and BNC connectors. C1 is omitted when evaluating only 1 capacitor.

You may also perform capacitor self resonant frequency testing with a vector network analyzer, a signal generator plus a 50 Ω terminated scope, or with a sweep generator ramp-driving the oscilloscope X input while simultaneously driving a VCO with logarithmic output to the Y oscilloscope input. SPICE simulations may also yield insight.

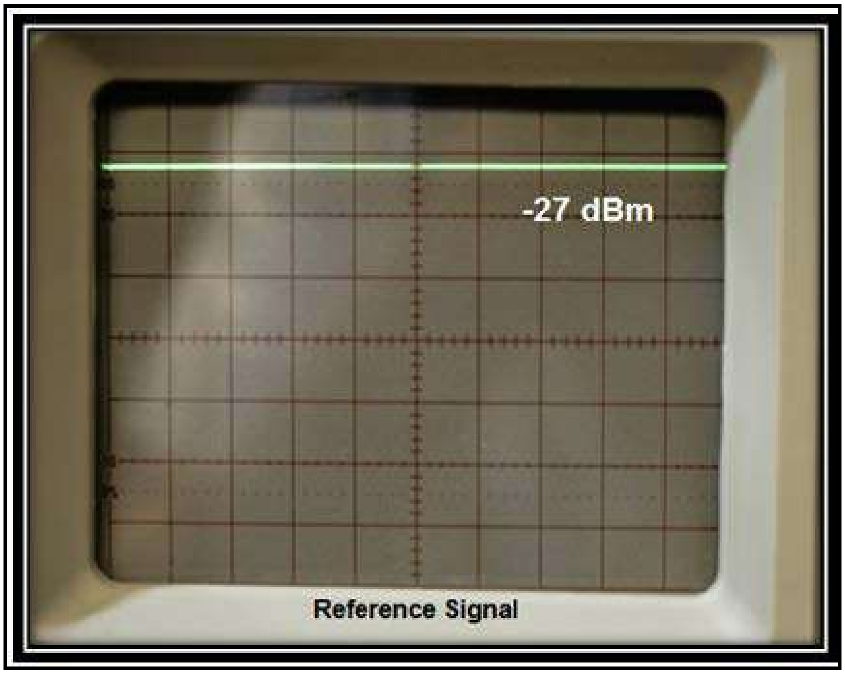

Above — The -27 dBm reference with a through-connector between my 2 coax patch cables (C0 + C1 board removed).
To save time I shot these SA photos handheld and prefer a slower shutter to capture a nice CRT tracing, so some of the photos show a little hand jitter.

Single Shunt Capacitors

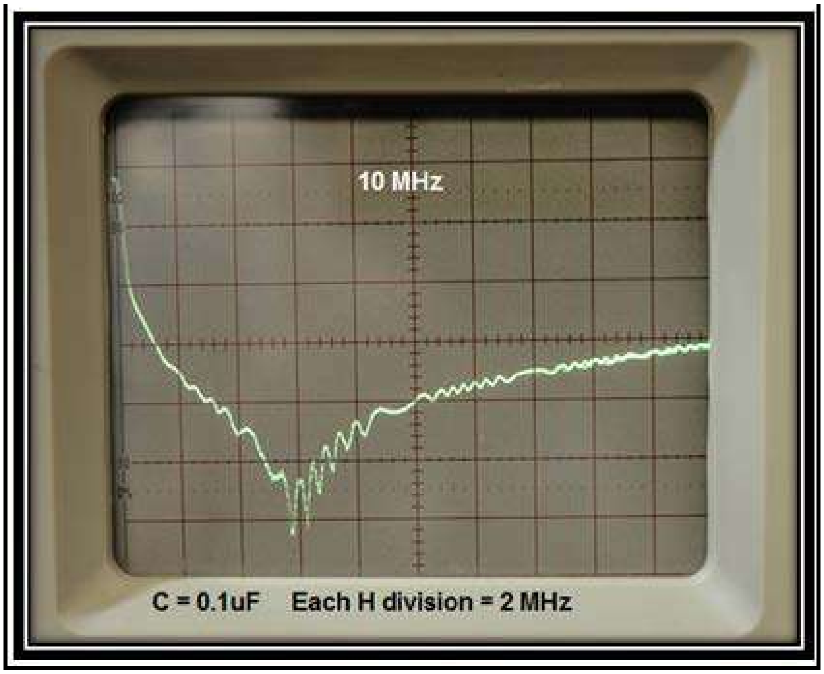

Above — C0 = 0.1 µF. I view the capacitor like a trap. At almost 5.8 MHz lies the peak attenuation, or lowest impedance — this is C0's self resonant frequency. The peak bypass frequency lies ~ 60 dB down. At 20 MHz, the attenuation is only ~ 30 dB.

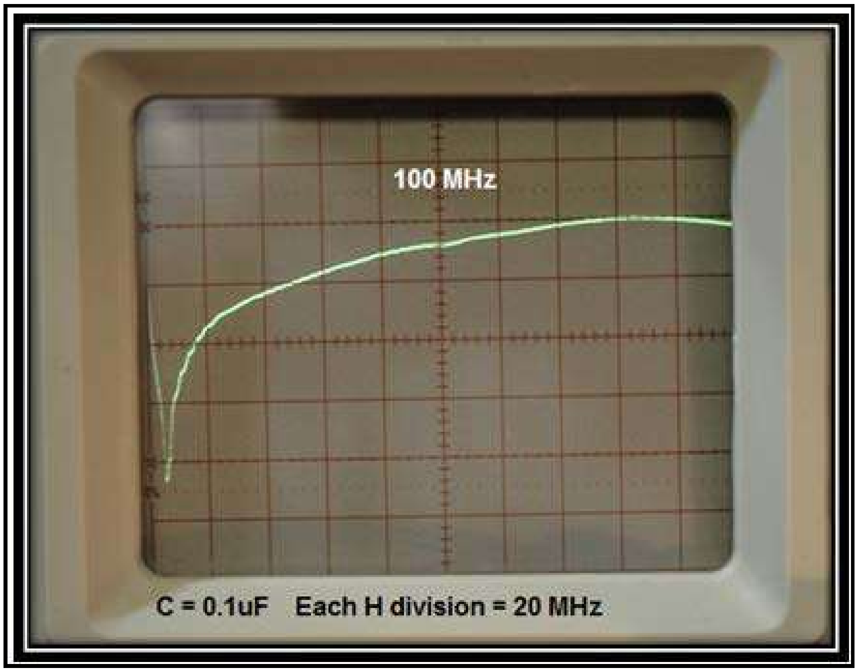

Above — Another shot of the 0.1 µF bypass cap with a 200 MHz span. At 50 MHz, the reference signal lies only ~ 17 dB down. At 100 MHz, the attenuation is only ~ 11 dB — this hardly qualifies as “bypass” much above the self resonant frequency. Above the self resonant frequency, a capacitor's XL affects impedance more than the ESR and XC of the capacitor.

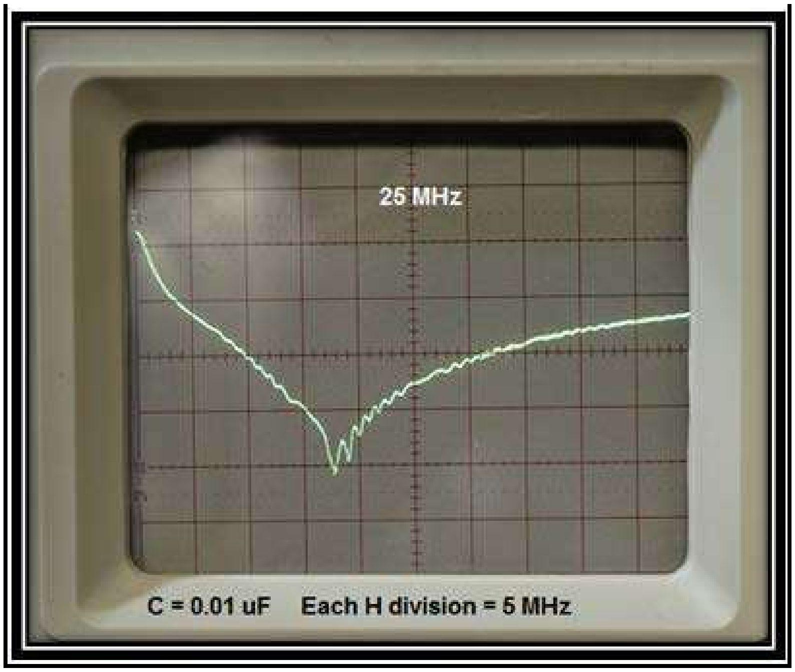

Above — C0 = 0.01 µF. The peak bypass frequency (capacitor self resonant frequency) is centered at 17.5 MHz and is ~ 50 dB down; not as deep as with the 0.1 µf cap. At 50 MHz, the attenuation is ~ 21 dB.

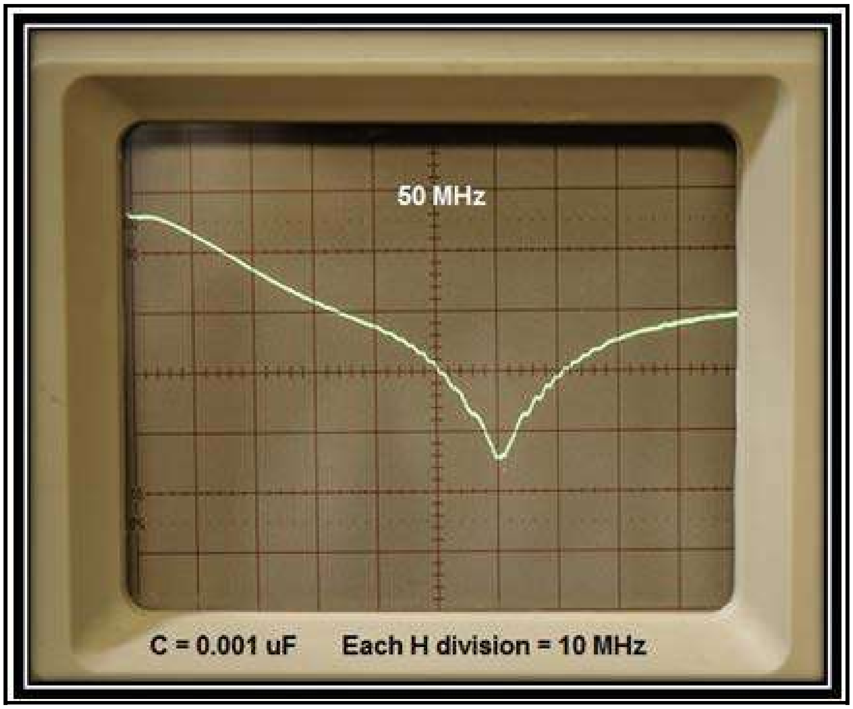

Above — C0 = 0.001. The response is peaked at 62 MHz with an attenuation of ~42 dB. At 100 MHz, the signal is 17 dB down. Again, the peak attenuation looks diminished compared to that of the 0.1 µF and the 0.01µF caps.

Capacitors in Parallel

Now I placed 2 caps in parallel (C0 + C1) as some builders do to try and garner a wider attenuation bandwidth.

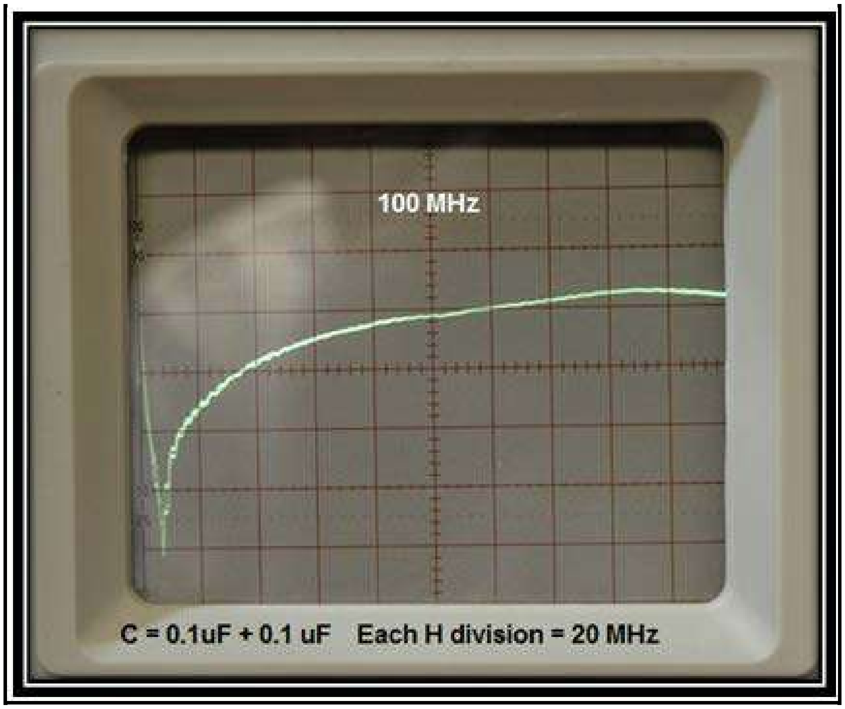

Above — C0 = 0.1 µF + C1 = 0.1 µF. The peak attenuation = 60 dB at 8 MHz; up 2 MHz from that of the single 0.1 µf capacitor. At 100 MHz, attenuation = ~ 19 dB — better than a single 0.1 µF but still low.

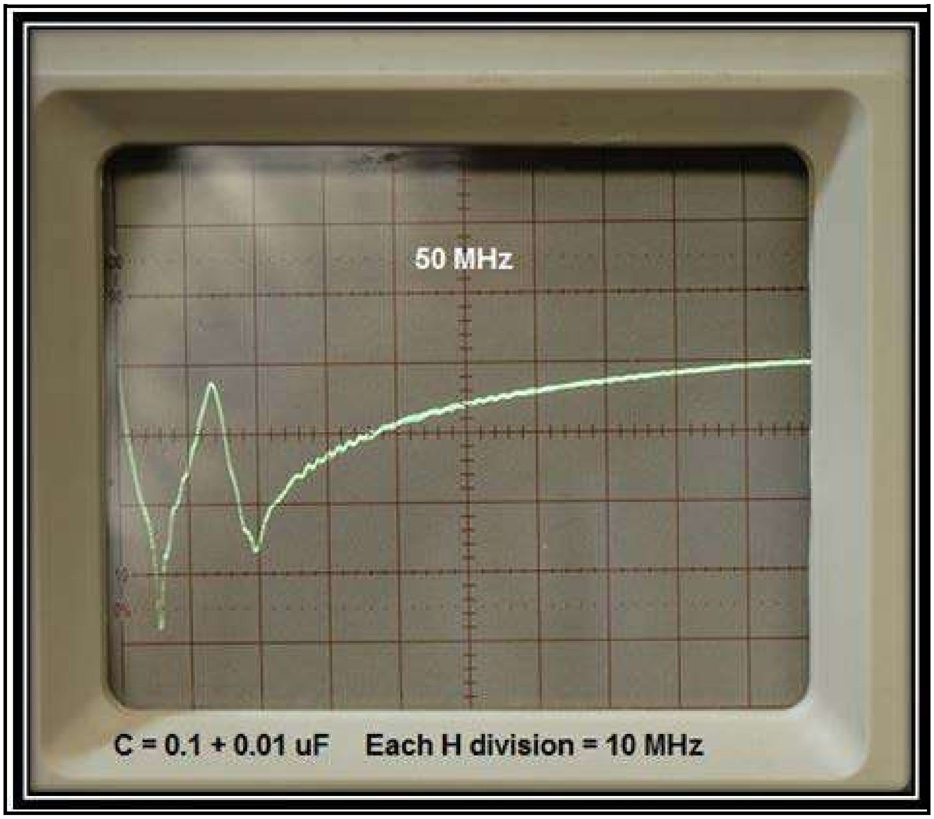

Above — C0 = 0.1 µF + C1 = 0.01 µF. Yikes! With the 2 different cap values, we get an unfortunate high impedance blip peaking at 13 MHz. Each capacitor exerts its self resonant frequency, but in between these self resonant frequencies, lies a disaster.
When placed in parallel, the inductance of 1 capacitor resonates with the capacitance of the other to form a parallel resonance — leading to a high impedance — that blocks RF bypass and peaks at a specific frequency.

But wait. Things can get worse:

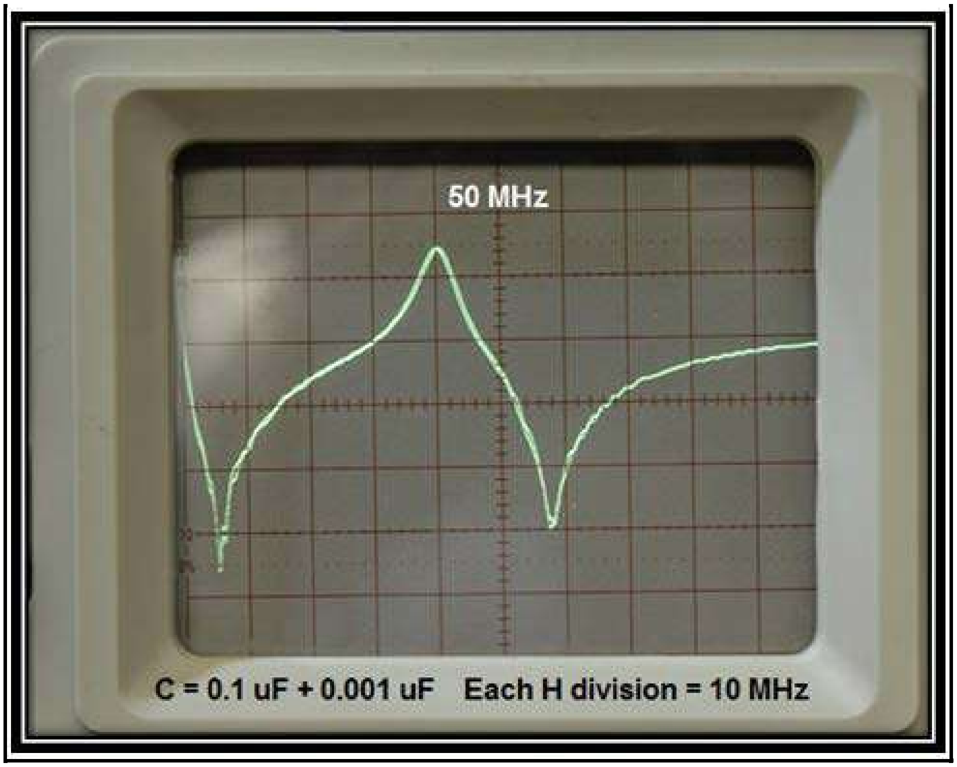

Above — C0 = 0.1 µF + C1 = 0.001 µF. The wide value variance between these 2 capacitors creates a huge, high impedance spike where the attenuation is only about 6 dB at 40 MHz. Catastrophic bypass indeed. катастрофа.

A 7 mm Length of Copper Wire

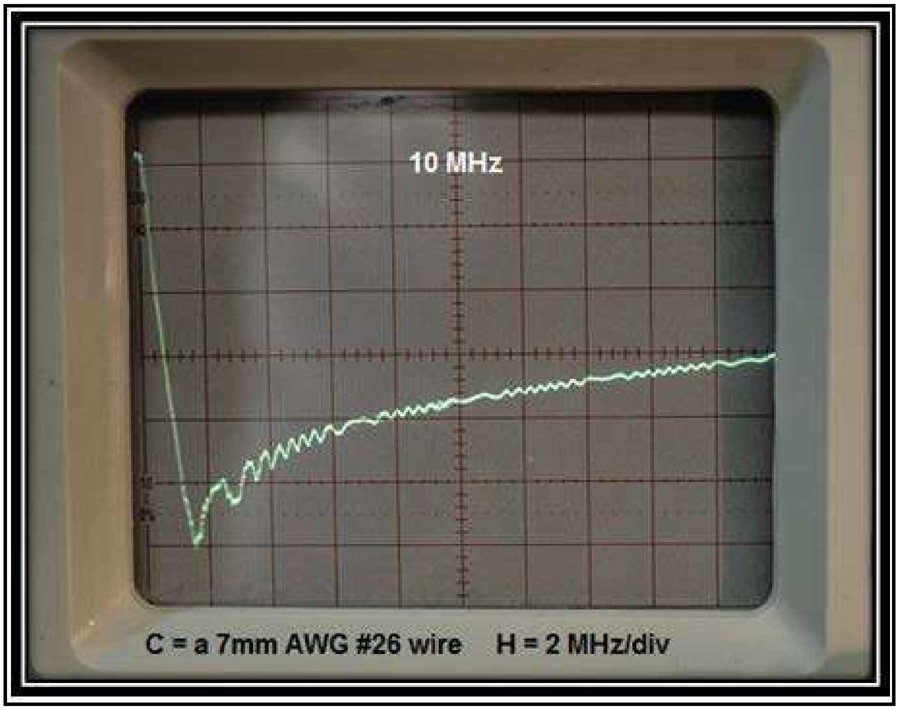

Above — A 7 mm piece of 26 gauge copper wire was shorted to ground instead of C0. This wire measured at ~7 nH of inductance and I saw that attenuation decreases with frequency from 62 dB at 8 MHz to ~33 dB at 20 MHz. Even a short piece of wire doesn't exhibit a flat, wideband bypass.

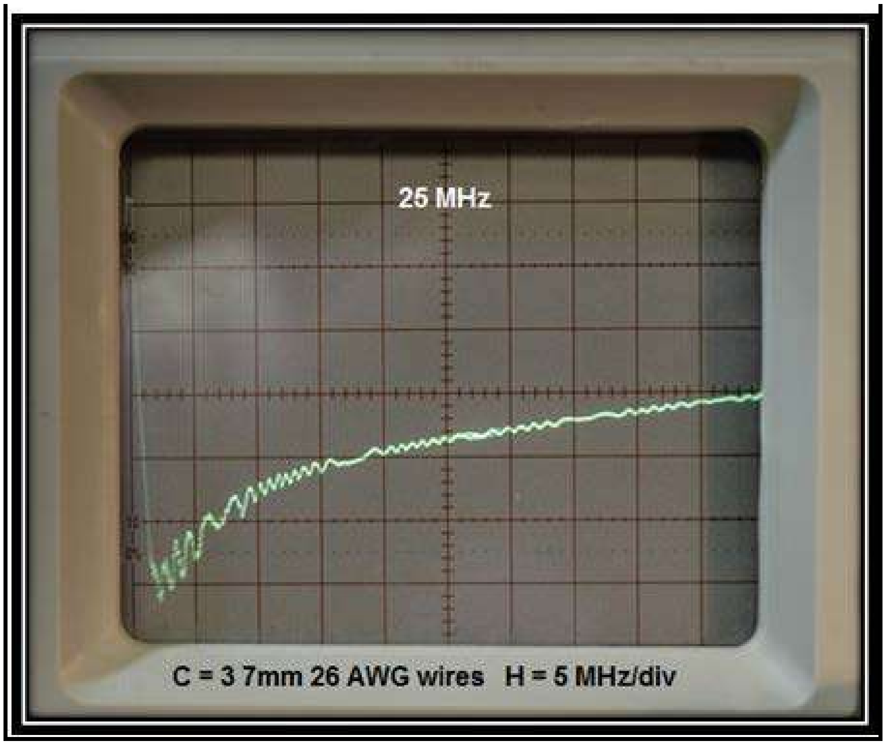

Above — This spectrum analysis shows three 7 mm wires shunted to ground – not much different than 1 wire.

0.01 µF Chip Capacitor(s)

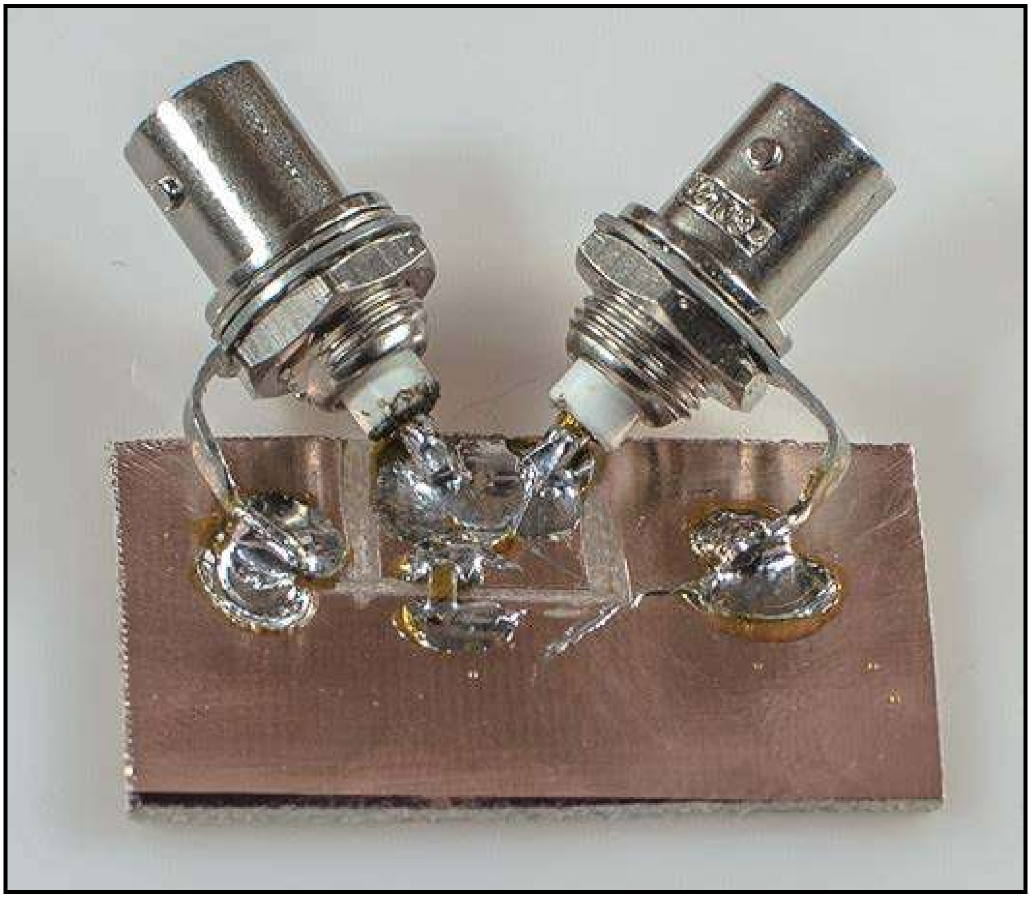

Above — The magnified copper board that I tested one or two 0.01 µF chip capacitors. You can see 1 capacitor soldered in.

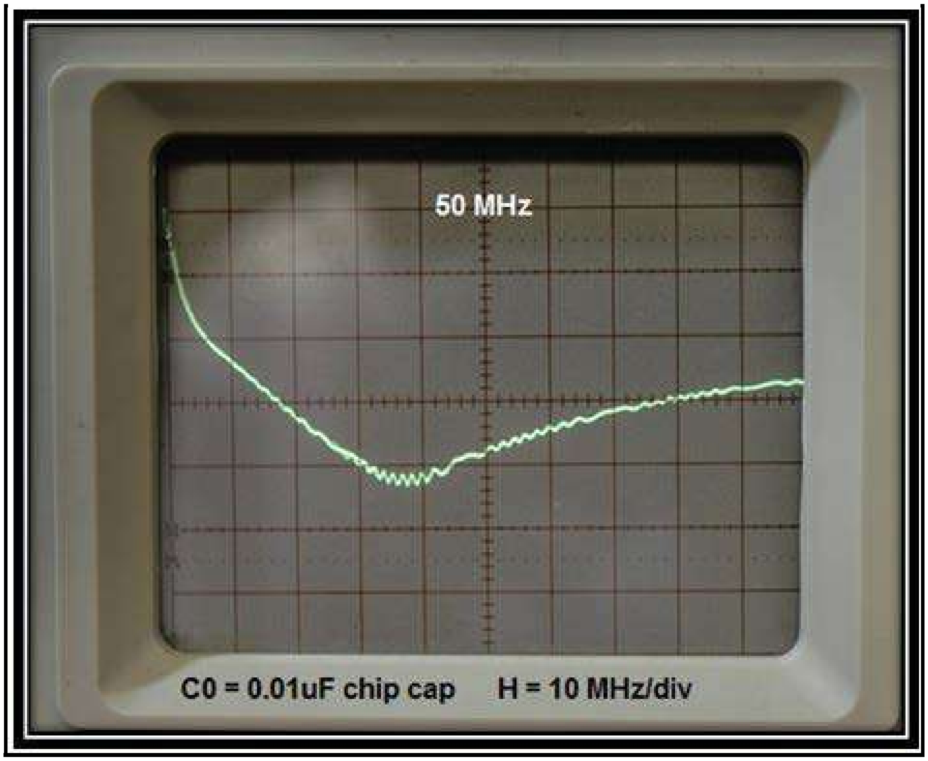

Above — C0 = 0.01 µF SMT cap. The SMT parts exhibited a peak attenuation of 45 dB at ~37 MHz. The attenuation dip lacks the sharp peak of the hole-through 0.01 µF cap shown eariler and exhibits a somewhat wider bandwidth. The self resonant frequency of the chip capacitor is 5 MHz higher than the particular hole-through capacitor I measured. Click for a side by side photo.

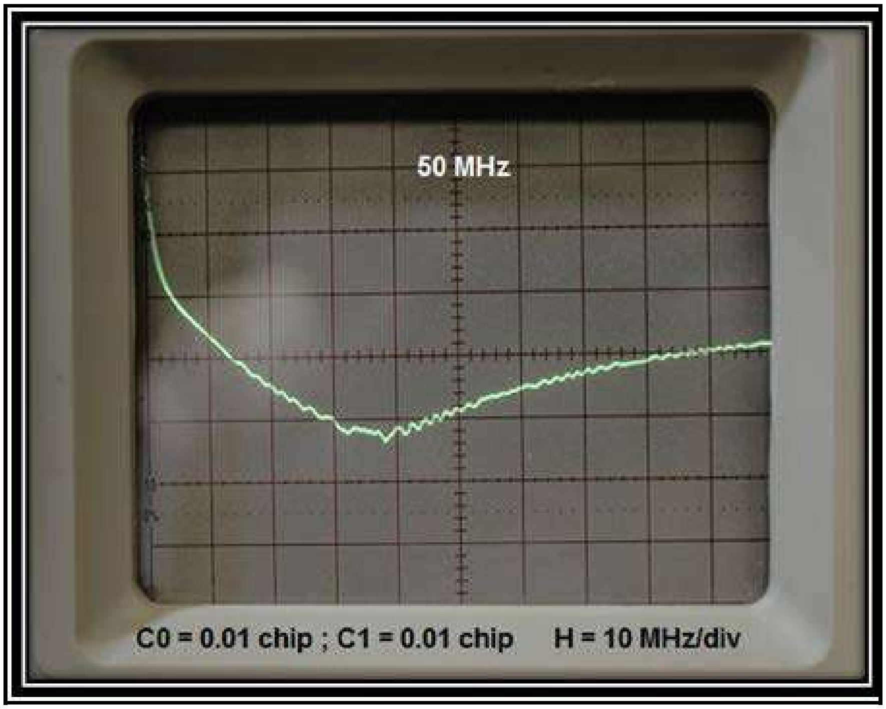

Above — C0 = 0.01 µF + C1 = 0.01 µF. The SMT parts exhibited a peak attenuation of 45 dB at ~37 MHz; similar to the single 0.01 µ F chip cap, but with a few more dB attenuation between 10 and 20 MHz.

0.1 µF Ceramic + a 2.2 µF Electrolytic Capacitor

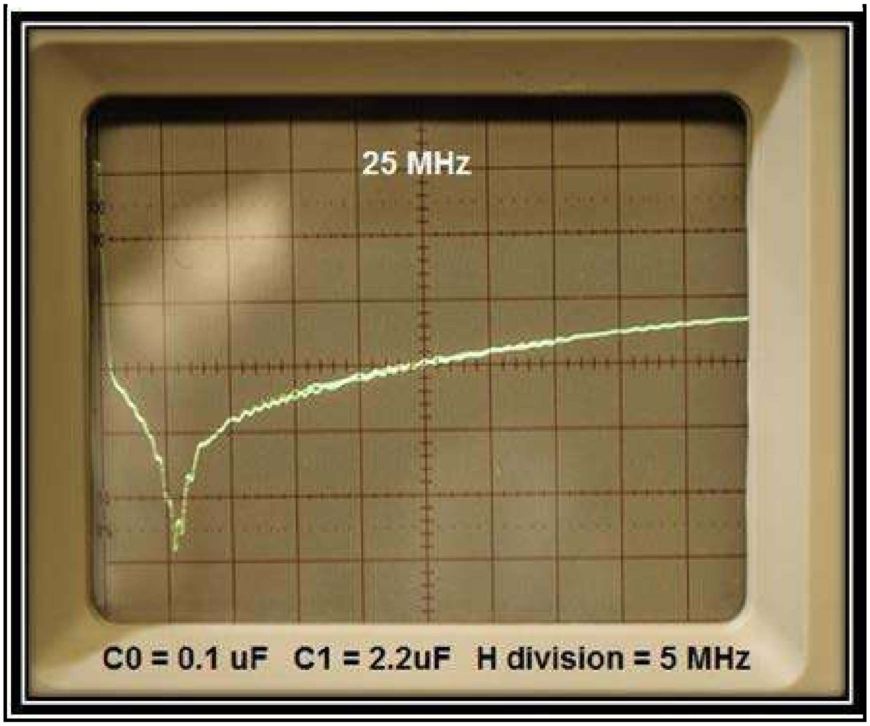

Above — C0 = 0.1 µF and C1 = 2.2 µF. The low Q 2.2 µF C1 electrolytic cap did not create the a parallel resonance with C0. Shaky photo — sorry. I also tested a 10 and 22 µF cap in parallel with C0 and saw no disturbance caused by a parallel resonance between a big AF capacitor and C0 (an RF value cap) with my RF spectrum analyzer.

Some additional experiments applying a low Q *A*F capacitor plus a ceramic *RF* capacitor for wideband bypass yielded some interesting results and I'll present these in a future project.

Capacitors in Parallel with a Series Resistor to Lower Q

In previous experiments, placing 2 RF capacitors in parallel led to the formation of a peaked high impedance blip between the low impedance peaks set by the self-resonant frequency of the 2 capacitors. If multiple capacitors are soldered in parallel, the series inductance of each capacitor will resonate with the capacitance of the next smaller C value.

One solution is to put a resistance in series with all but 1 of the parallel capacitors so that the Q of resonance formed by this capacitor's series inductance and the capacitance of the next smaller capacitor is low. If capacitors exhibited 0 inductance then putting capacitors in parallel would be fine, however, since capacitors exhibit inductance, a parallel resonant frequency may occur with capacitors in parallel.

I found applying a series resistance to lower Q may flatten the impedance versus frequency response of the bypass network, but didn't decrease the impedance at any 1 frequency. Optimal bypassing or achieving the lowest impedance over a wide frequency range presents a complex topic that might even challenge some engineers.

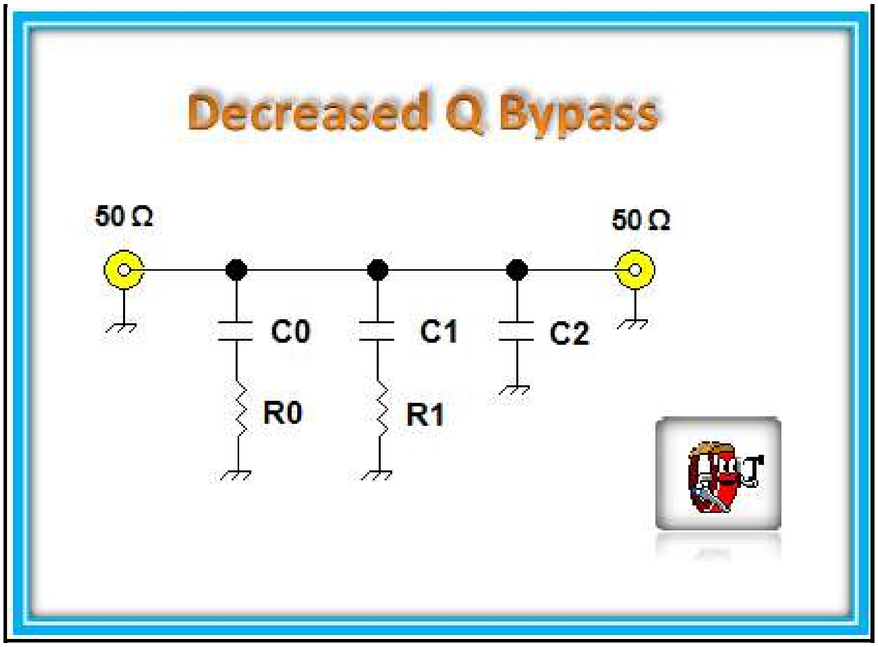
Decreased Q Bypass
50 Ω
50 Ω
C0
C1
C2
R0
R1

Above — A method to exact wideband bypass.

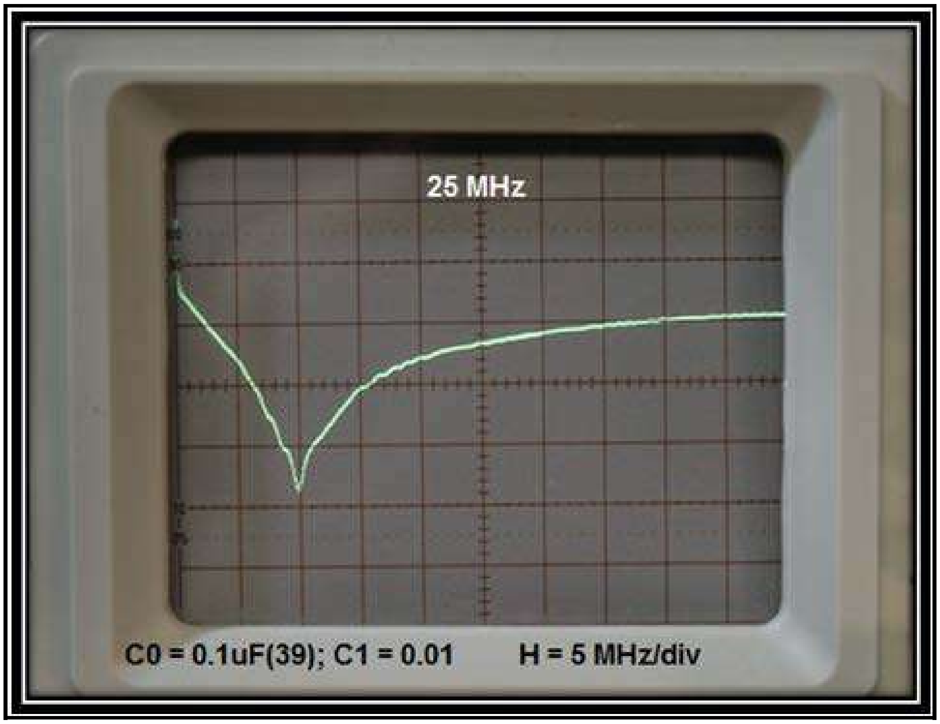

Above — My first try with C0 = 0.1 µF, R0 = 39 Ω and C1 = 0.01 µF. I arbitrarily placed the 39 Ω resistor in the R0 slot and saw that the high impedance peak seen earlier disappeared. This gave me the confidence to try 3 capacitors. I had no idea what R value to use and really just wanted to see what happens.

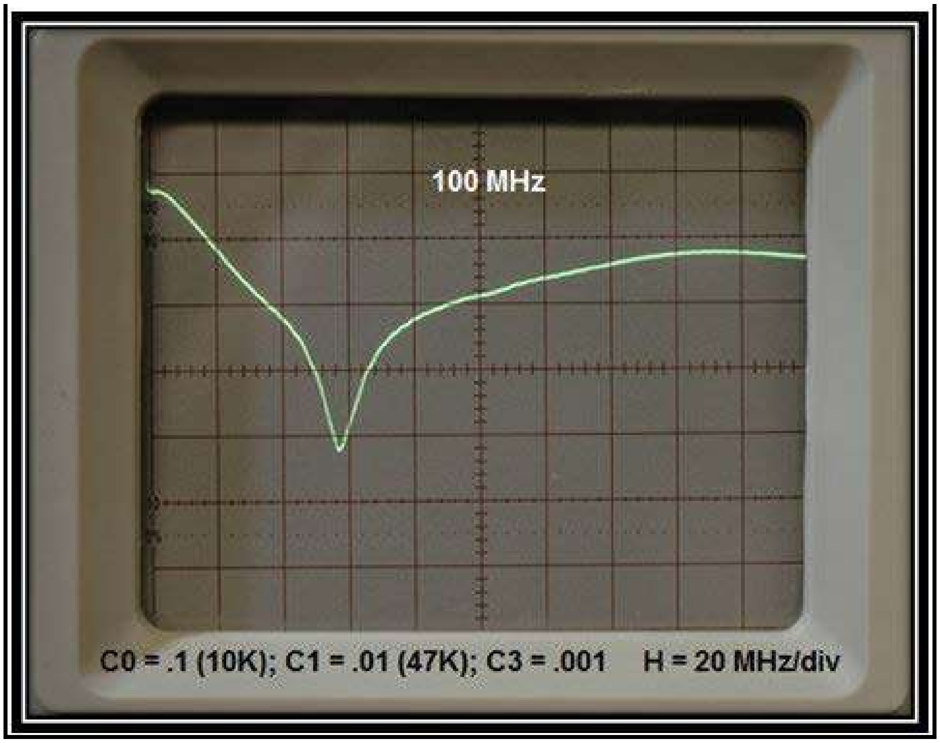

Above — The spectrograph with C0 = 0.1 µF, R0 = 10K, C1 = 0.01 µF, R1 = 47K and C3 = 0.001 µF.

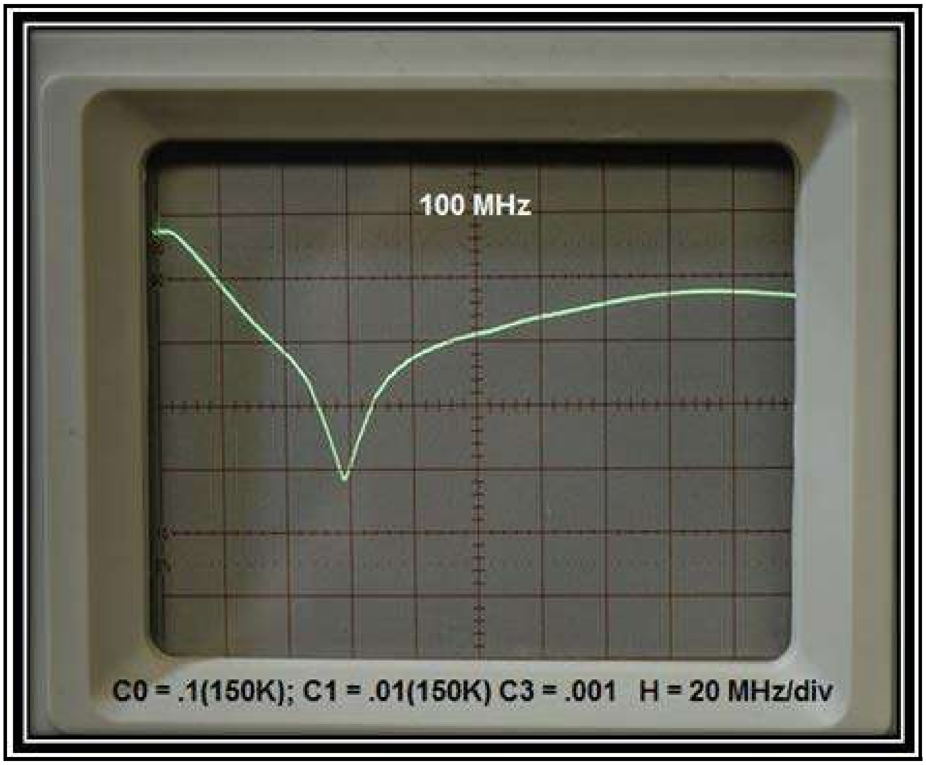

Above — C0 = 0.1 µF, R0 = 10K, C1 = 0.01 µF, R1 = 47K and C3 = 0.001 µF. Again, no high impedance peak response; the self resonant frequency is close to that measured with a single 0.001 uF earlier, however, the peak bypass frequency moved from to 57 MHz from 62 MHz. Changing the resistor values moved the self-resonant frequency and the peak attenuation value a little, but I fell kilometers short of setting a wide band bypass. My approach lacks any real science and I need to step it up.

I hope to learn what capacitor values and types, plus R values to apply. This sounds like a job for simulation as well as further on-bench experiments? After writing this material, I learned that Ken Kuhn wrote an Excel spreadsheet to examine the net impedance of up to 3 capacitors in parallel Click to download.

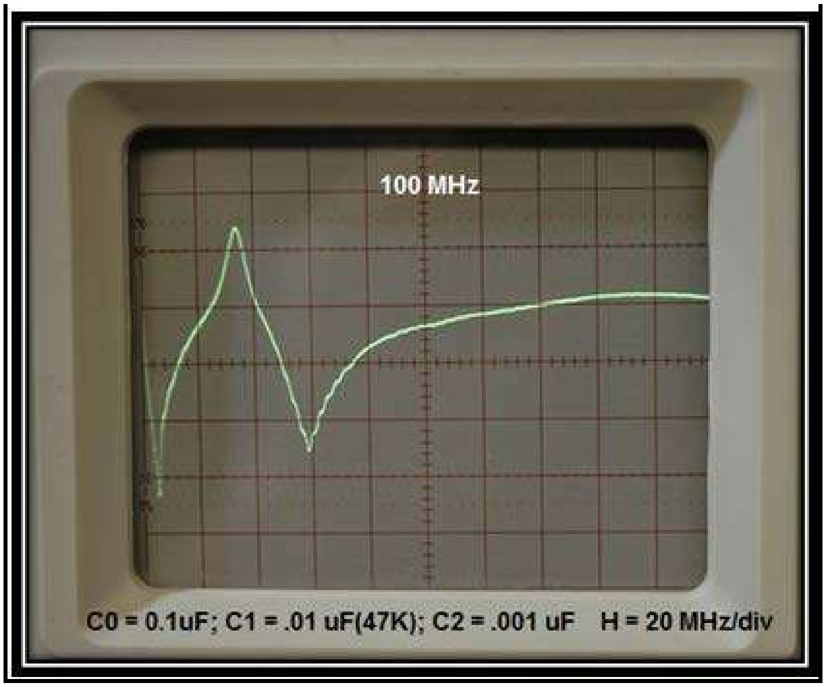

Above — Just as a gag, I removed R0 from capacitor C0 in 1 circuit and then hooked up the board. The high impedance peak re-emerged.

A Commerical Example

I found a wideband MMIC employing (R1 + C1 and C2) as part of a bypass strategy. Cick for the datasheet excerpt. Note the size of the SMD capacitors; 0603 — tiny caps! My experiments showed some high gain MMICs require careful low inductance grounding and correct part choices or crippling oscillations and other bypass issues might arise.

My Learnings

When we think bypass, we really should think frequency dependent attenuation. The bypass cap is actually a network where impedance versus frequency varies significantly. At its self resonant frequency, a capacitor will exhibit the lowest possible impedance making a single capacitor a relatively narrow-band bypass device. Intuitively, we might want to choose a capacitor with a series resonant frequency at the frequency we wish to bypass, however, if we require a wideband bypass, the need to evaluate our bypass capacitor(s) increases.

In short, above the series resonant frequency of a capacitor, its bypass is basically useless and we should likely ensure that the self-resonant frequency of the particular capacitor we're using is above the highest frequency to be bypassed.

Bypassing with 2 or more unmatched RF caps will lead to an attenuation gap with peak(s) determined by the parallel resonance of these capacitors. Going above a 10:1 capacitor ratio, for example, greater than a 0.1 and a 0.01 μf, may cause a **severe** gap in attenuation at the parallel resonant frequency generated by the 2 capacitors.

Mine and work from more reputable authors clearly shows we should avoid applying parallel **RF** bypass capacitors of *different* values unless we apply a Q-reducing resistor to the capacitor(s) in parallel with a given RF bypass capacitor. Please read EMRFD page 2.3 for more information and watch out for abundant folklore concerning RF bypass.

The need for measurement and analysis challenges us; in some cases, you may realize good attenuation in the radio band of interest, while poorly bypassing the frequencies above it and compromise an otherwise good design.

Capacitor lead length may affect self-resonance at RF.

Future Work

It would be awesome to learn more about getting a wide-band bypass. I want to order some low or ultra-low ESR caps and measure them. My MuRata RPE Series caps specify low inductance; low is relative — how low is low? Should we apply chip capacitors for bypass in our critical circuits such as low noise VHF amps or MMICs?. Am I fussing about nothing? Lots of questions that folklore just won't answer.

Per EMRFD page 2.3, bypass is only half the equation — we need to *decouple* + bypass to filter RF from moving along our DC lines and so forth.

---

6. Some Experiments with Chokes plus Decouple and Bypass Filters

Introduction

SRF = self resonant frequency; XL = inductive reactance; XC = capacitive reactance. L = inductor; C = capacitor; R = resistor.

Like the capacitor, inductors are networks with R, L and C and possess a SRF. R, L and C may vary with factors including the number of windings, frequency, or whether the L is wound on a ferromagnetic material, or air wound.

Considering R, L and C:

at frequencies below the SRF, XL dominates;
at frequencies above the SRF XC dominates;
at the SRF, the magnitude of XL and XC are equal but 180 degrees out of phase leaving resistance to dominate.

I encourage you to learn more by visiting the fabulous web site of David, G3YNH.

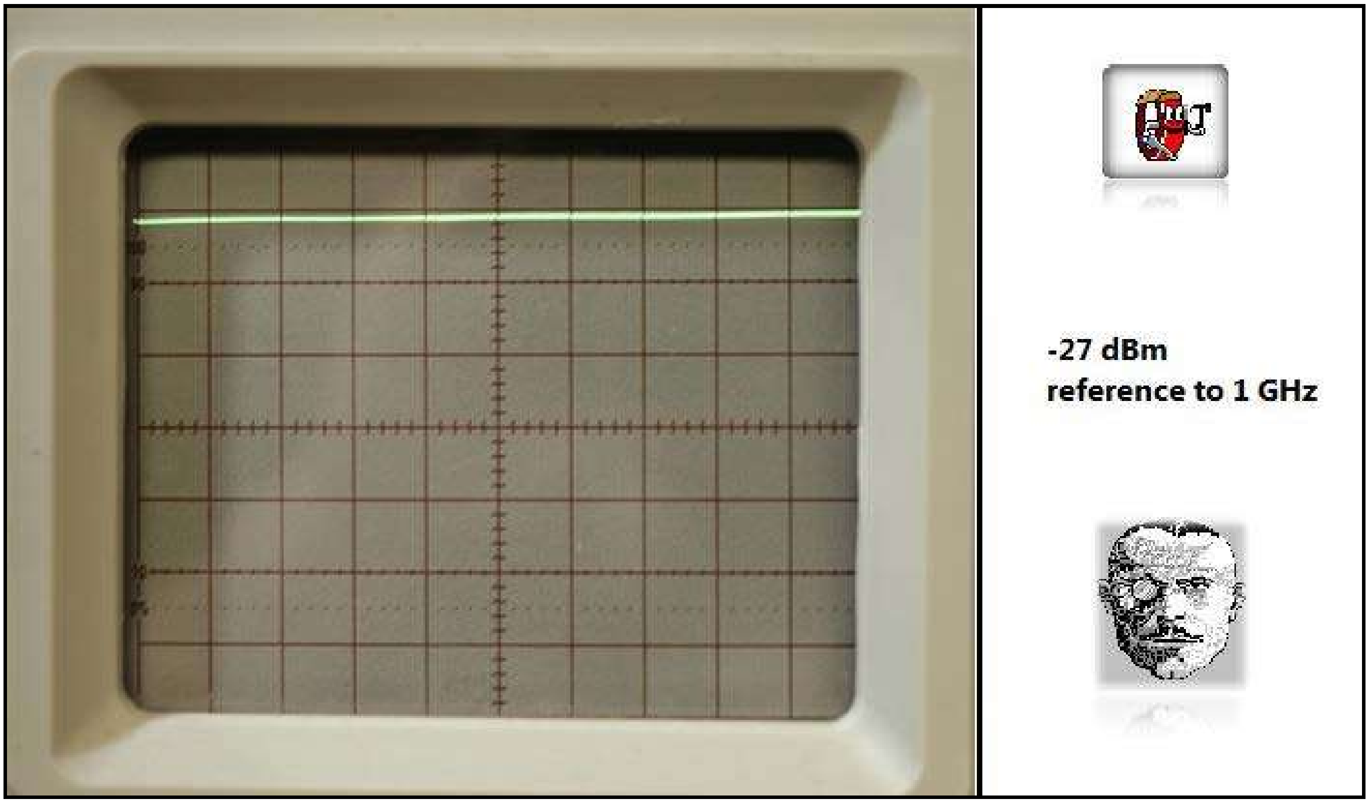

Above — Reference signal at -27 dBm. I used the exact test method shown in Section 5.

For those unaware, the spectrum analyzer screen is divided into 10 by 10 graticules. Each vertical division represent a 10 dB change; read down from the reference -27 dBm to measure the attenuation of the reference signal in dB. Horizontal divisions represent frequency; start at 0 on the left hand side and increment as specified on each figure.

A Few Inductors

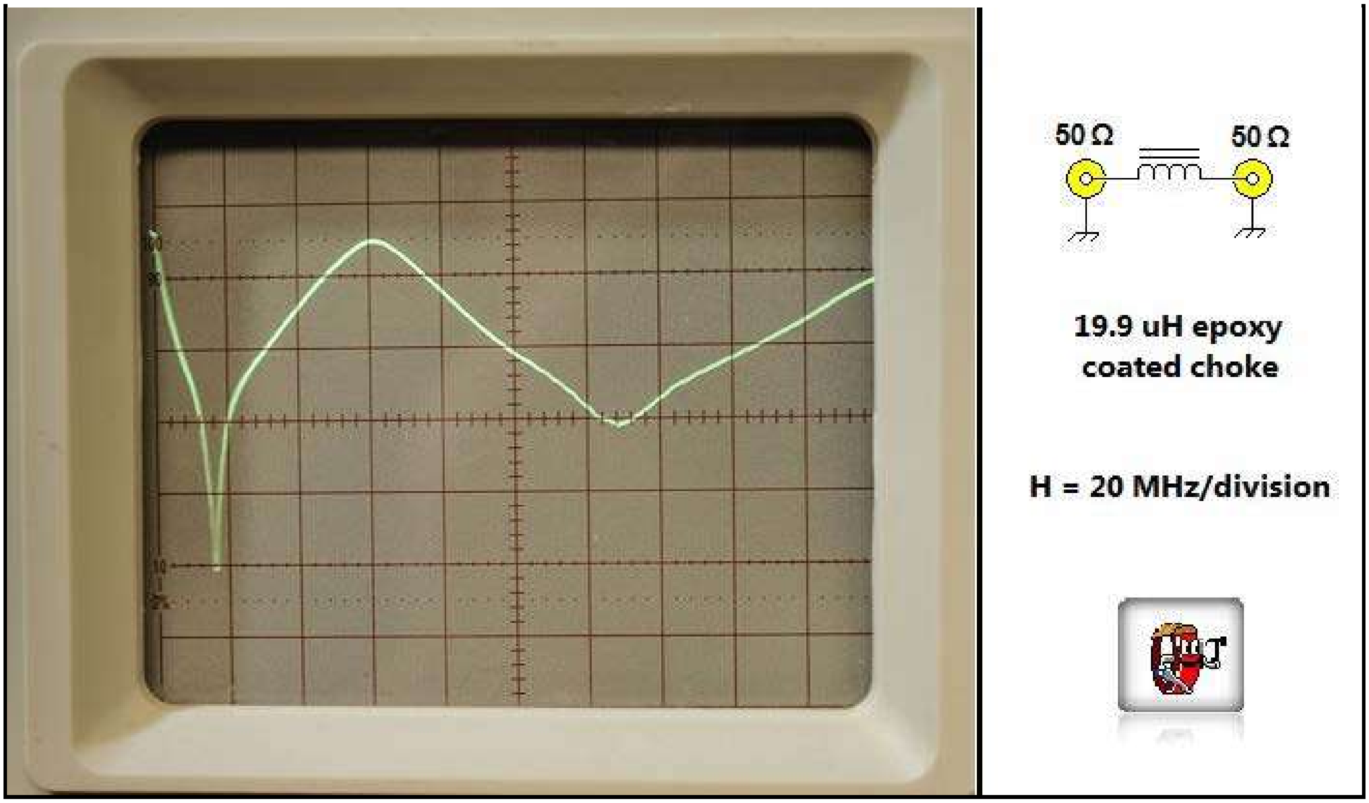

Above — A 19.9 uH epoxy coated choke that exhibits a primary SRF at 18 MHz and a second, smaller SRF at ~128 MHz. This wretched L gave me grief at 63 MHz. After measurement, I tossed it in the garbage can.

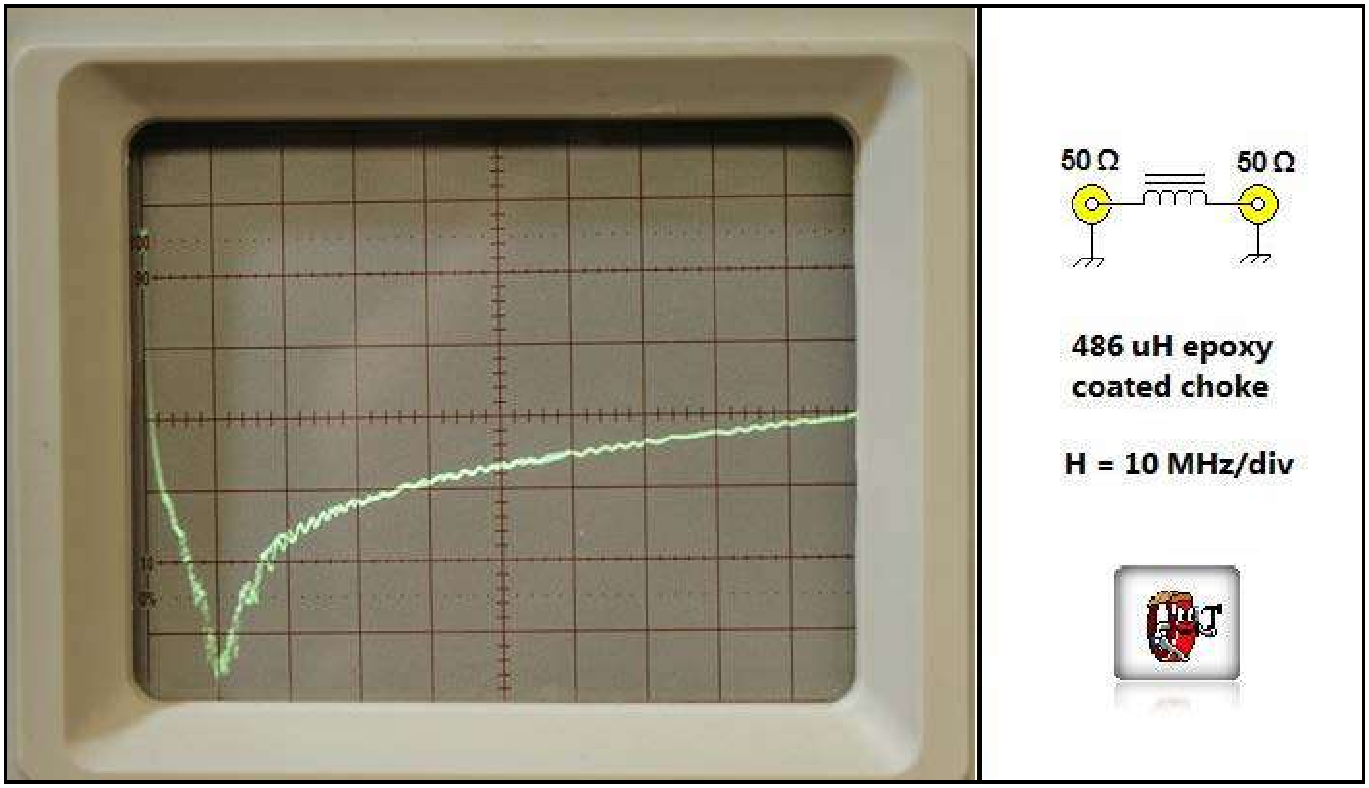

Above — A large, junk box choke with an SRF at about 10 MHz.

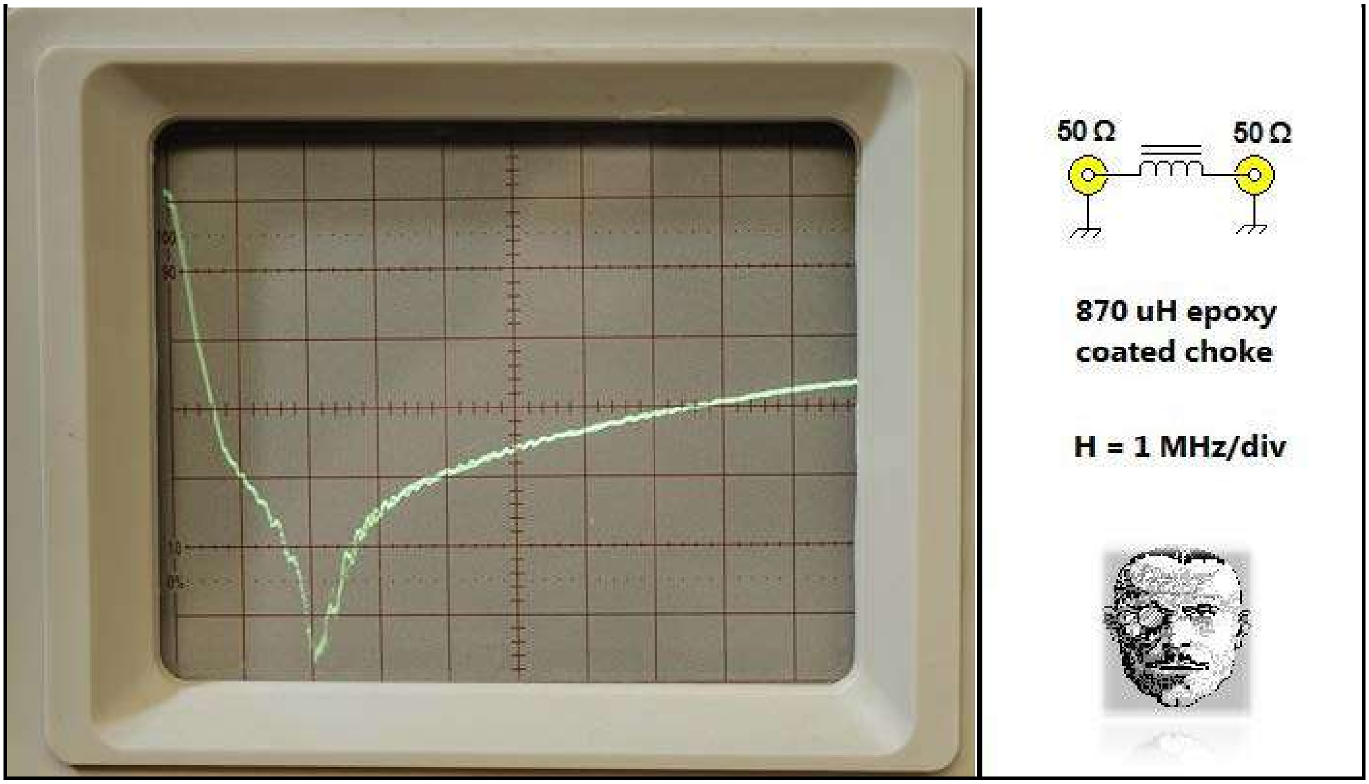

Above — I rarely use these big chokes: 870 µH with a SRF at about 2 MHz.

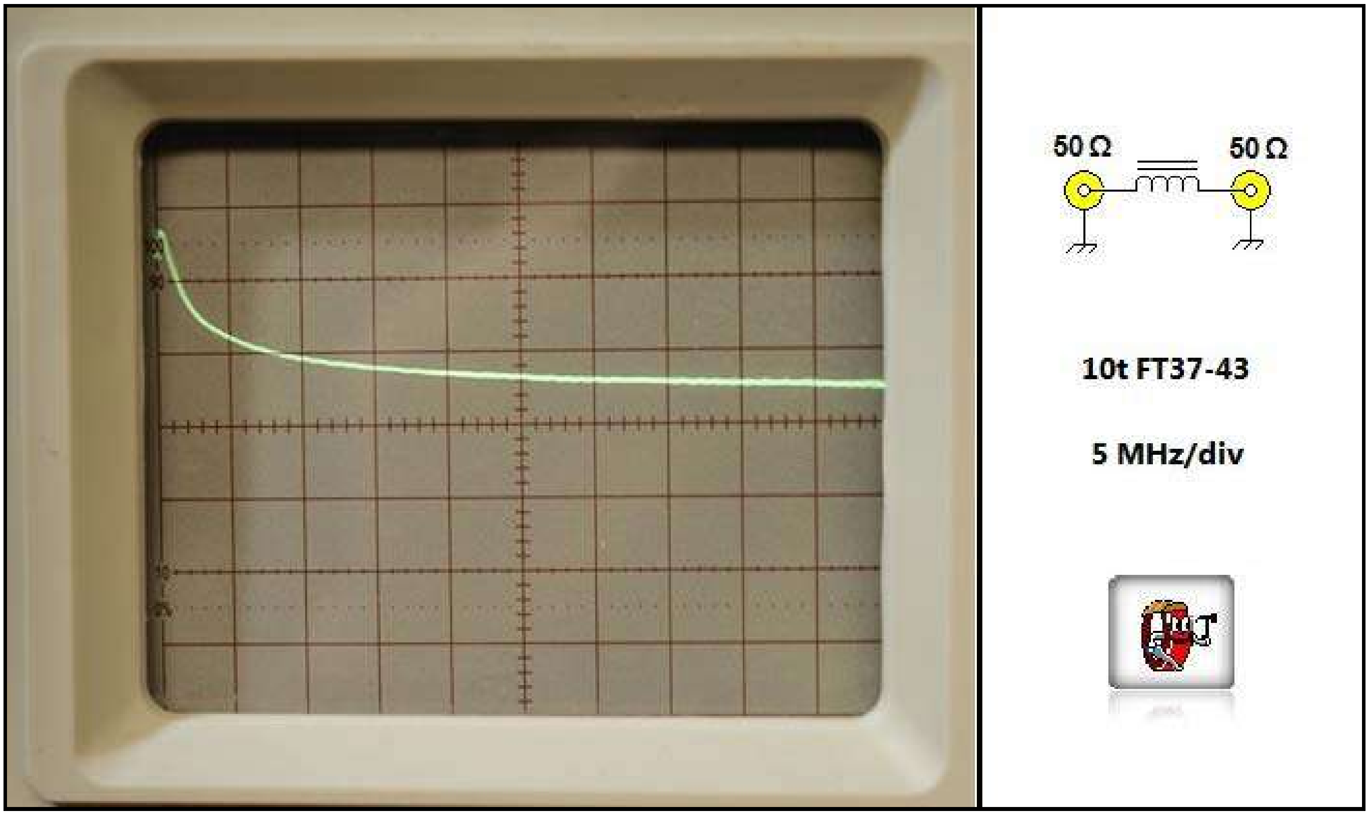

Above — A common L on our benches — 10 turns of #26 AWG on a FT37-43 ferrite toroid.

I couldn't measure the SRF with any span on my spectrum analyzer. I expect that a parasitic capacitance lies in parallel with the inductance, but the #43 material, with its low Q and high losses blankets the usual deep notch we see when the L exhibits a higher Qu.

10 Turns on a FT37- 43 with a Bypass Capacitor Shunting Each End

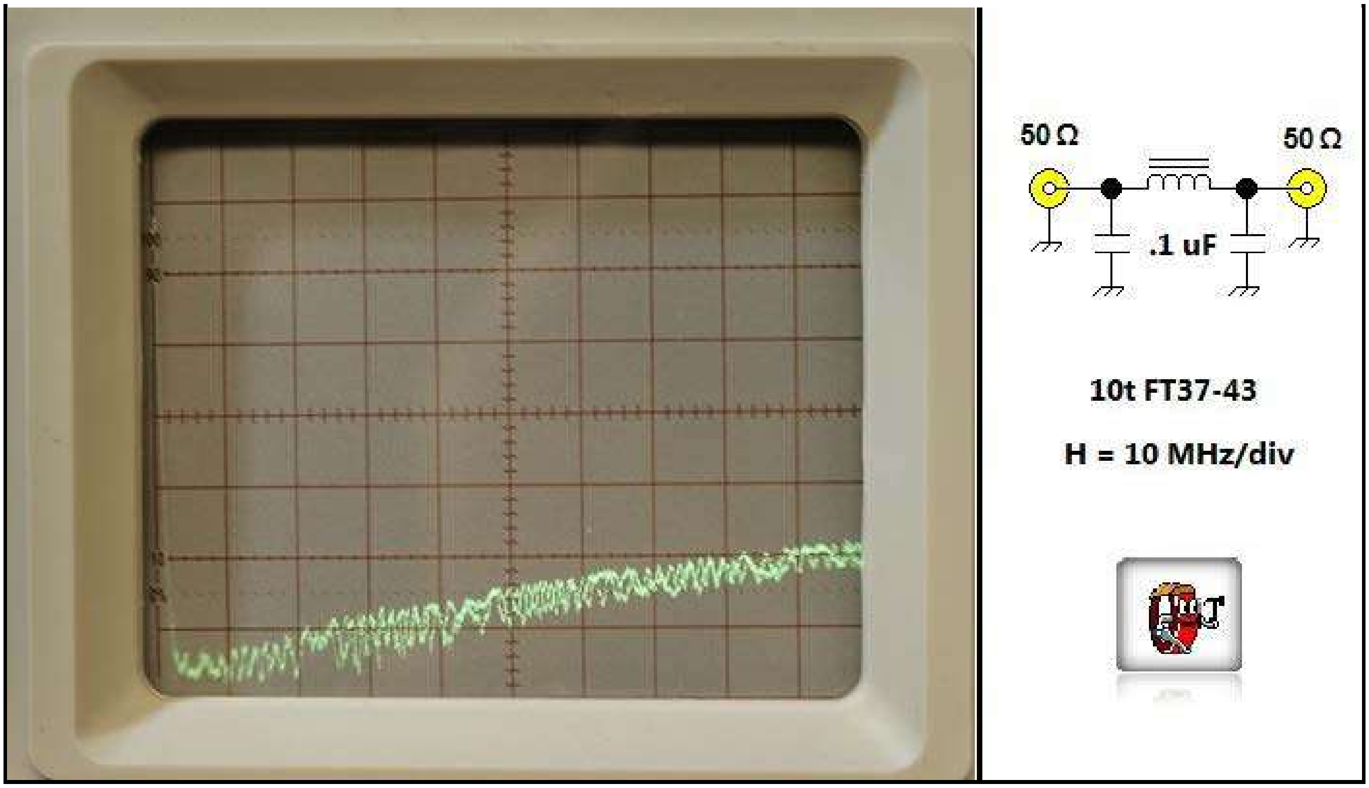

Above — Look at the big difference after adding shunt capacitors to a 10 turn FT37-43! Even at 100 MHz the attenuation lies nearly 50 dB down. Now I understand why Wes says **decouple plus bypass** when filtering our DC lines and so forth.

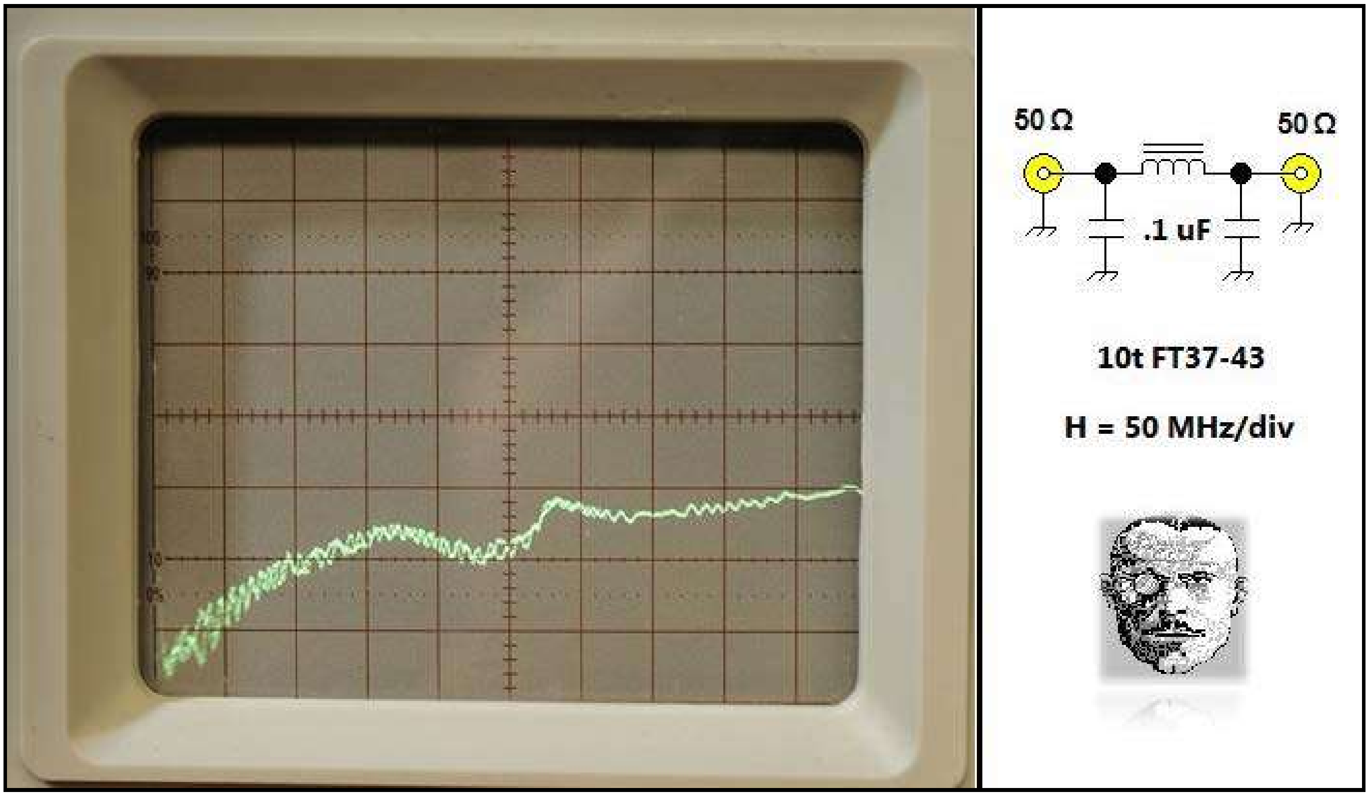

Above — The 10 turn FT37-43 coil with 0.1 μF shunt caps measured out to 500 MHz. Pardon the camera shake; I took all the photos hand held to save time.

A Resistor with a Bypass Capacitor Shunting Each End

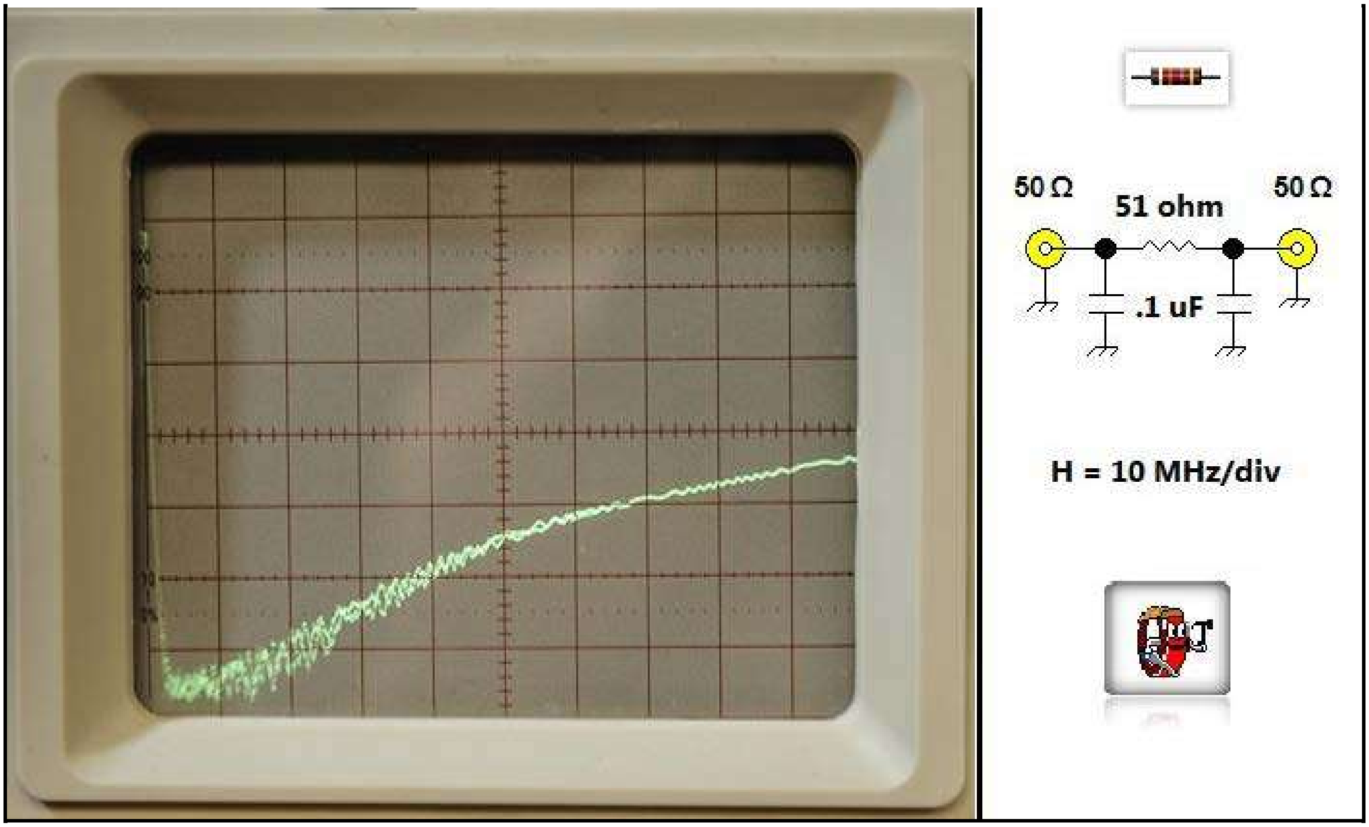

Above — A 51 Ω resistor bypassed with 0.1 μF capacitors at each end spanned out to 100 MHz. Even at 30 MHz, the attenuation looks stellar.

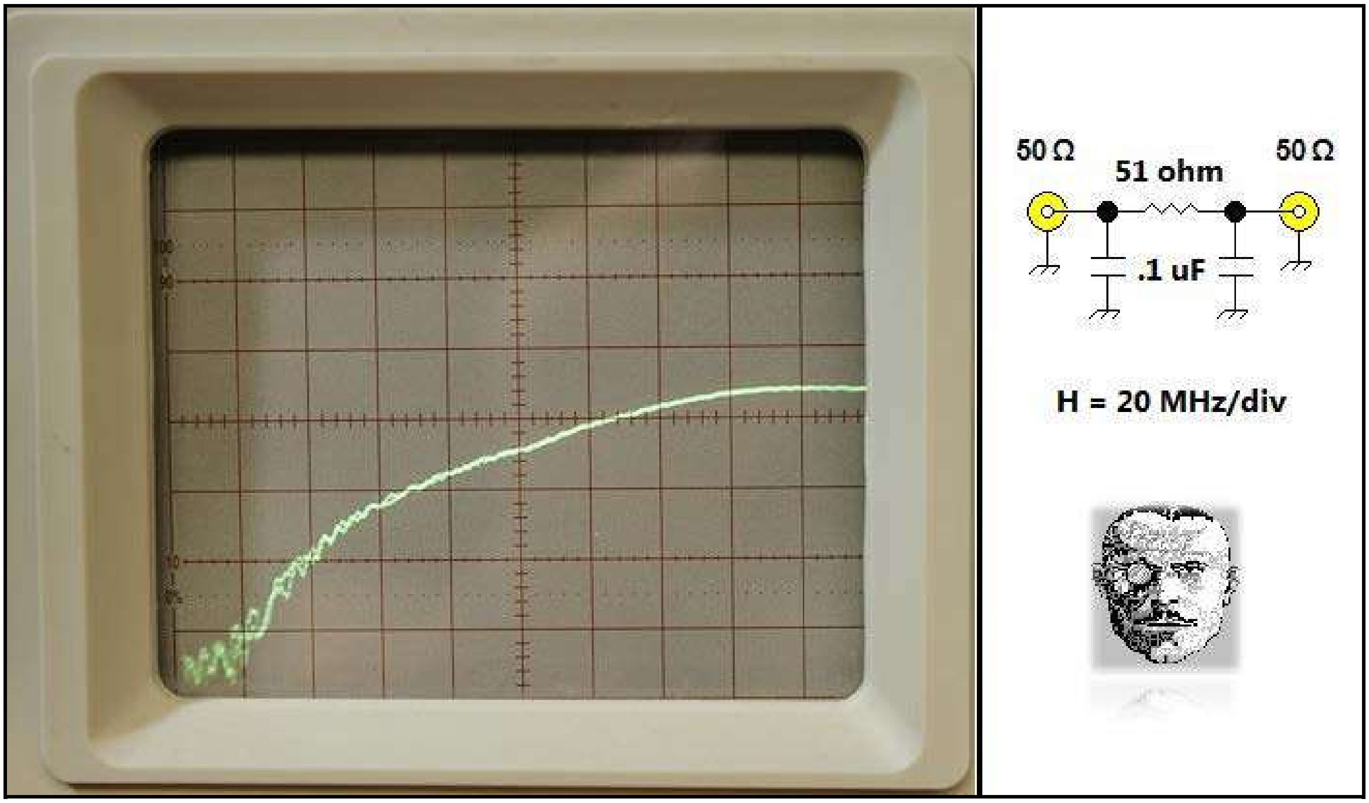

Above — The 0.1 μF bypassed 51 Ω resistor out to 200 MHz. I often use a 51 Ω decoupling resistor with appropriate capacitor values in active circuits that draw from 10 - 18 mA.

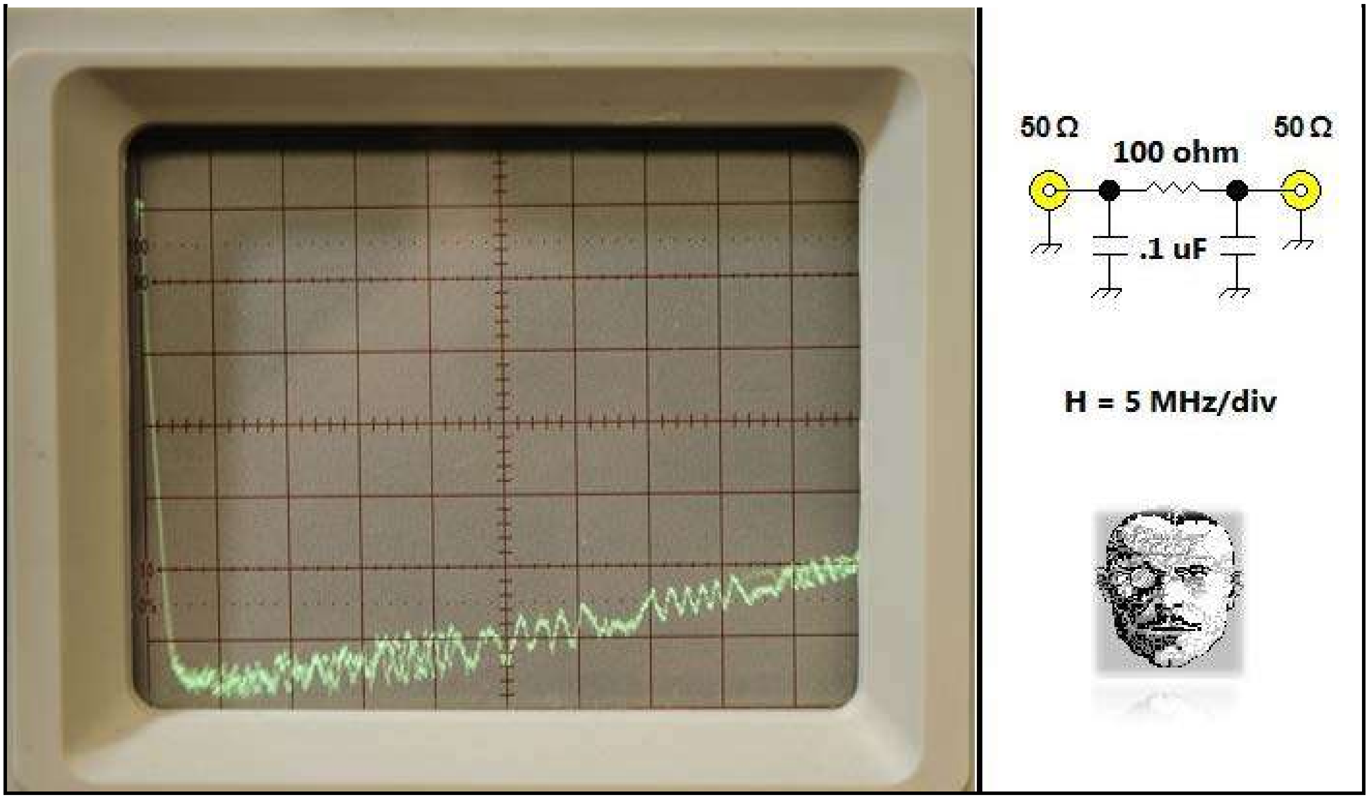

Above — The "bench standard"; a 100 Ω R with a shunt 0.1 μF at each end. We use this all the time. Even at 25 MHz, the attenuation looks around 55 dB down.

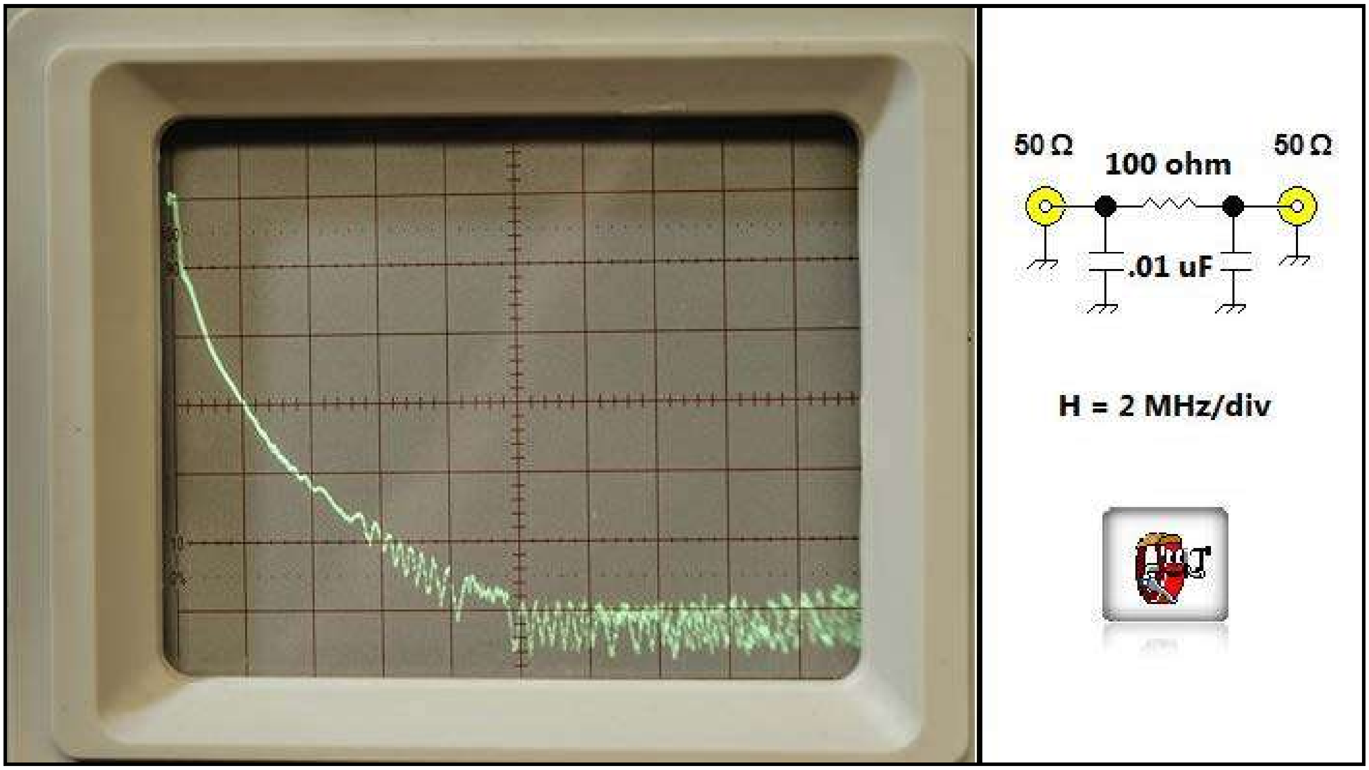

Above — The 100 Ω R with the shunt capacitors decreased to 0.01μF. At 6 MHz, we're about 50 dB down. From 10 to 20 MHz, the attenuation is about as high as I can measure.

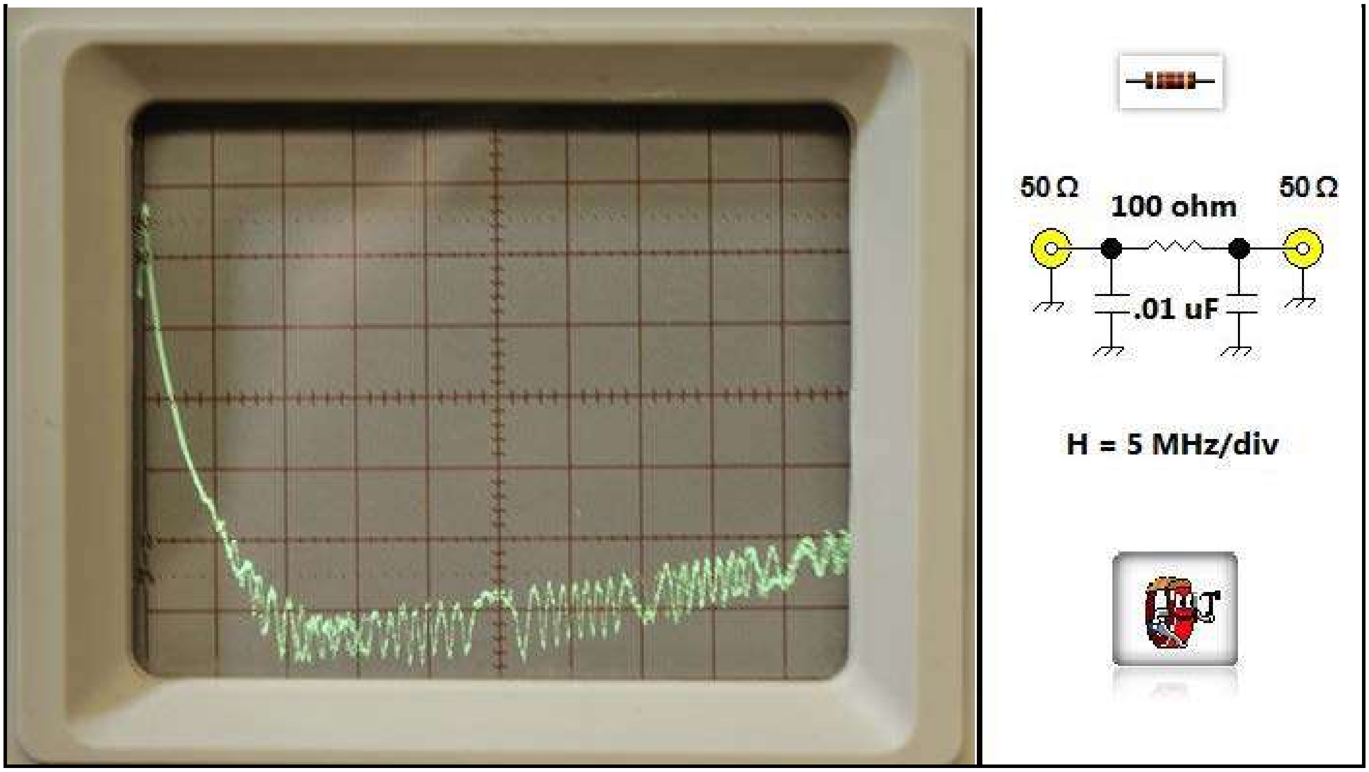

Above — 100 Ω R plus 0.01 µF caps out to 50 MHz. I've used this combination of R and C for filtering at 50 MHz a lot.

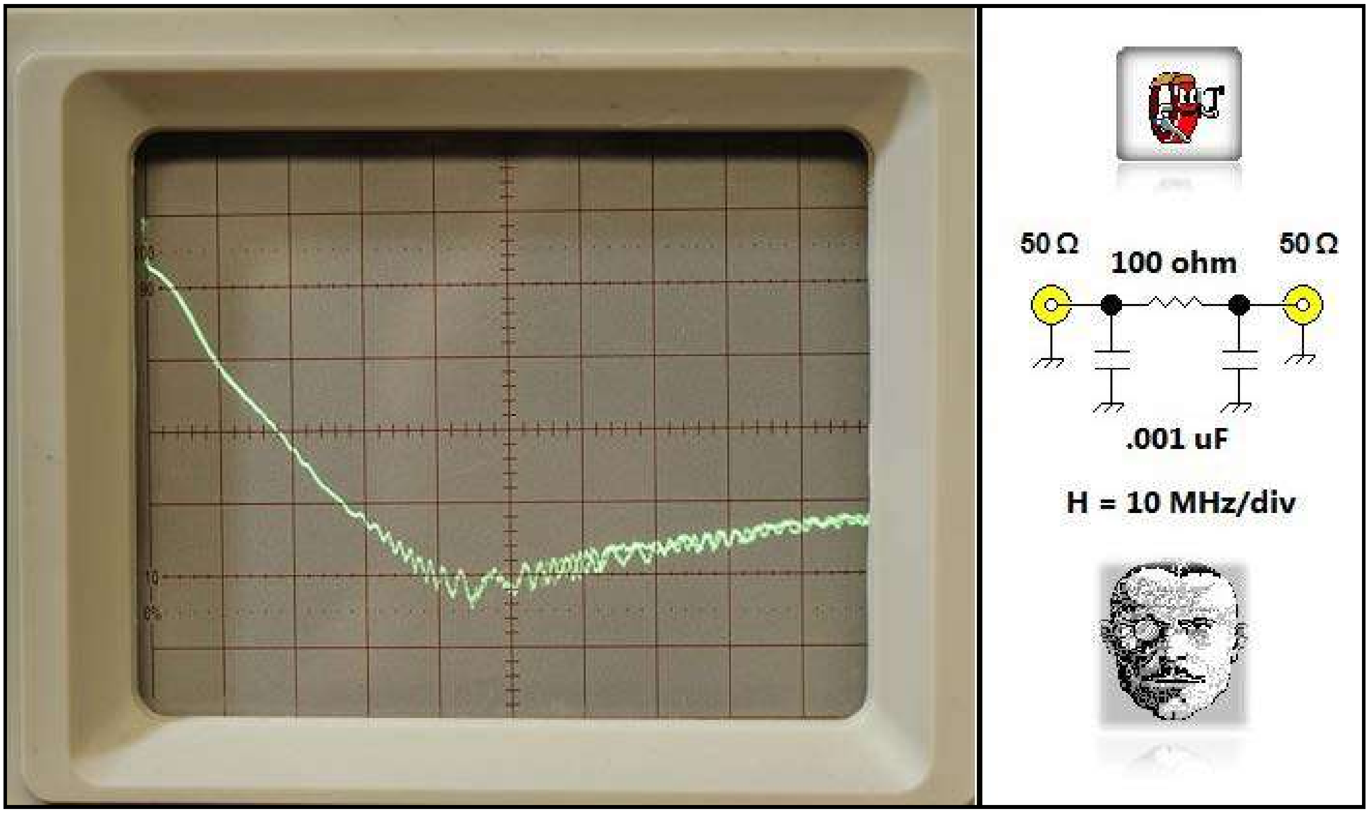

Above — 100 Ω R plus 0.001 µF capacitors out to 100 MHz. In my particular circuit, the attenuation at 50 MHz equals that of the 100 R + 0.01 µF C low-pass filter shown directly above.

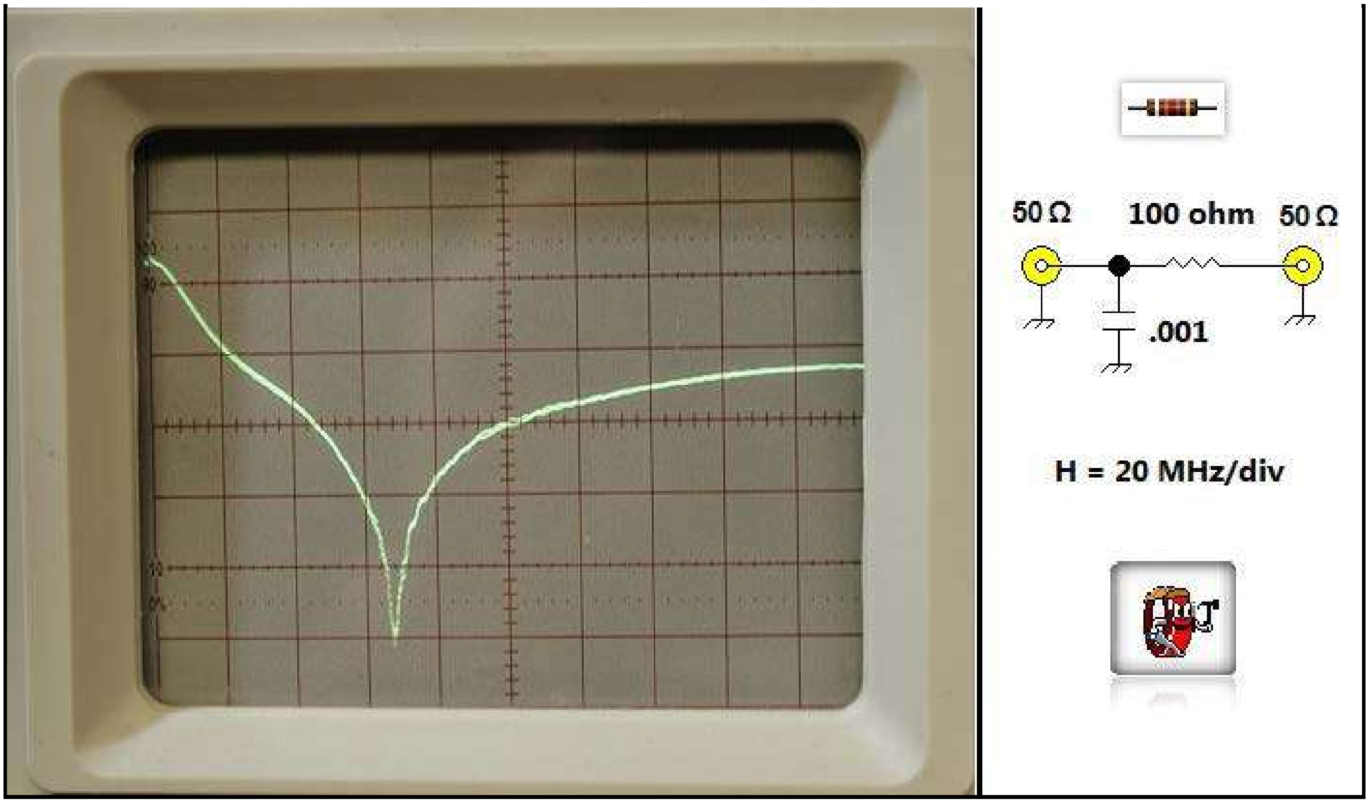

Above — A 100 Ω R plus a single 0.001 µF capacitor. If you leave off 1 capacitor, a serious notch appears at ~ 68 MHz. If you flip the filter around so the bypass cap is on the right hand side, the tracing appears the same. This problem occurred with all the filters tested in all experiments. As possible, solder a suitable bypass capacitor on *both* sides of the R or L.

I encourage you to experiment with the SRF of coils and wideband decouple + bypass filters on your own.

QRP Posdata for Oct 2013 — SRF of some common bypass capacitors

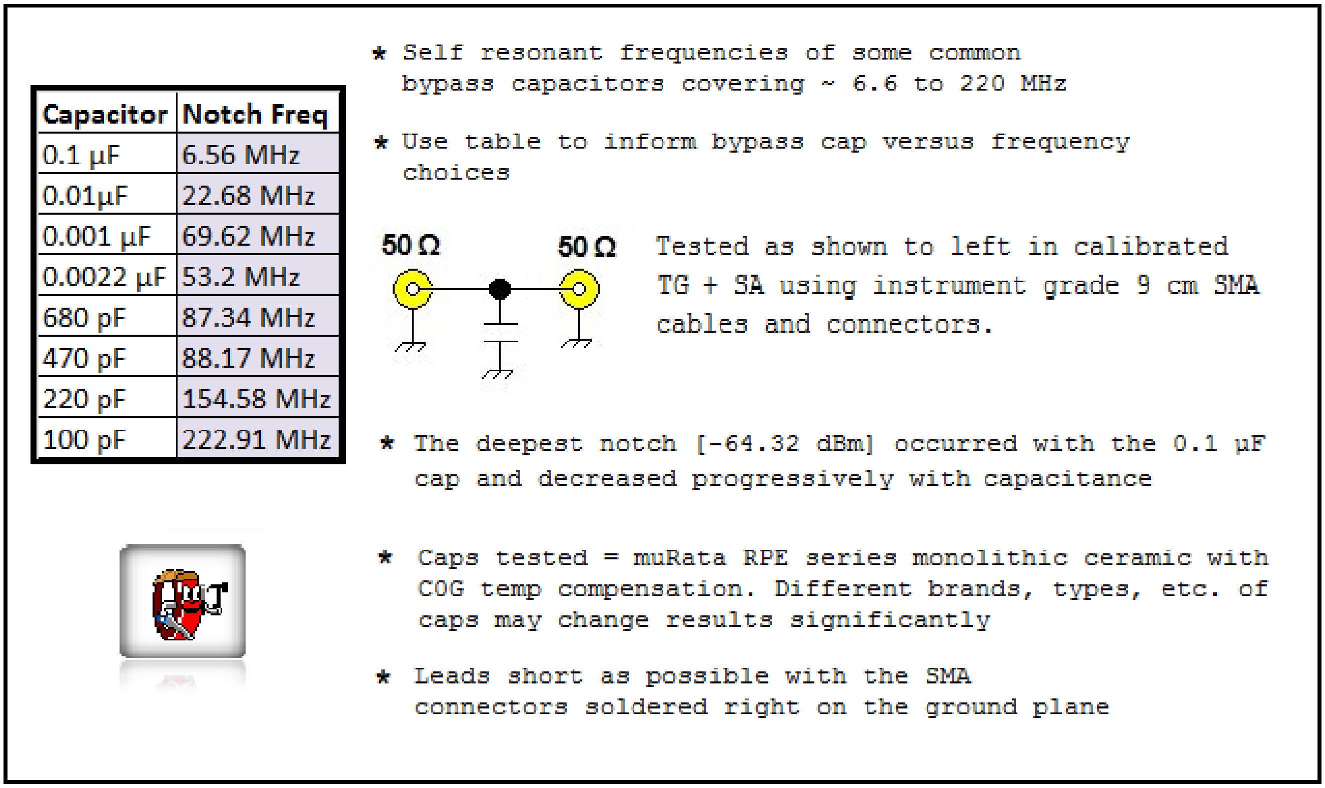

- Self resonant frequencies of some common bypass capacitors covering ~ 6.6 to 220 MHz
- Use table to inform bypass cap versus frequency choices

Capacitor	Notch Freq
0.1 µF	6.56 MHz
0.01µF	22.68 MHz
0.001 µF	69.62 MHz
0.0022 µF	53.2 MHz
680 pF	87.34 MHz
470 pF	88.17 MHz
220 pF	154.58 MHz
100 pF	222.91 MHz

Tested as shown to left in calibrated TG + SA using instrument grade 9 cm SMA cables and connectors.

- The deepest notch [-64.32 dBm] occurred with the 0.1 µF cap and decreased progressively with capacitance
- Caps tested = muRata RPE series monolithic ceramic with C0G temp compensation. Different brands, types, etc. of caps may change results significantly
- Leads short as possible with the SMA connectors soldered right on the ground plane

Above — A reference table showing the self resonant frequency of several comon value bypass capacitors in my parts collection. For example, if

I'm making a 21 MHz circuit, the best bypass capacitor choice from the table above = 0.01 µF. If possible, sweep the capacitors in your own collection to determine their SRF; or whether they're even suitable.

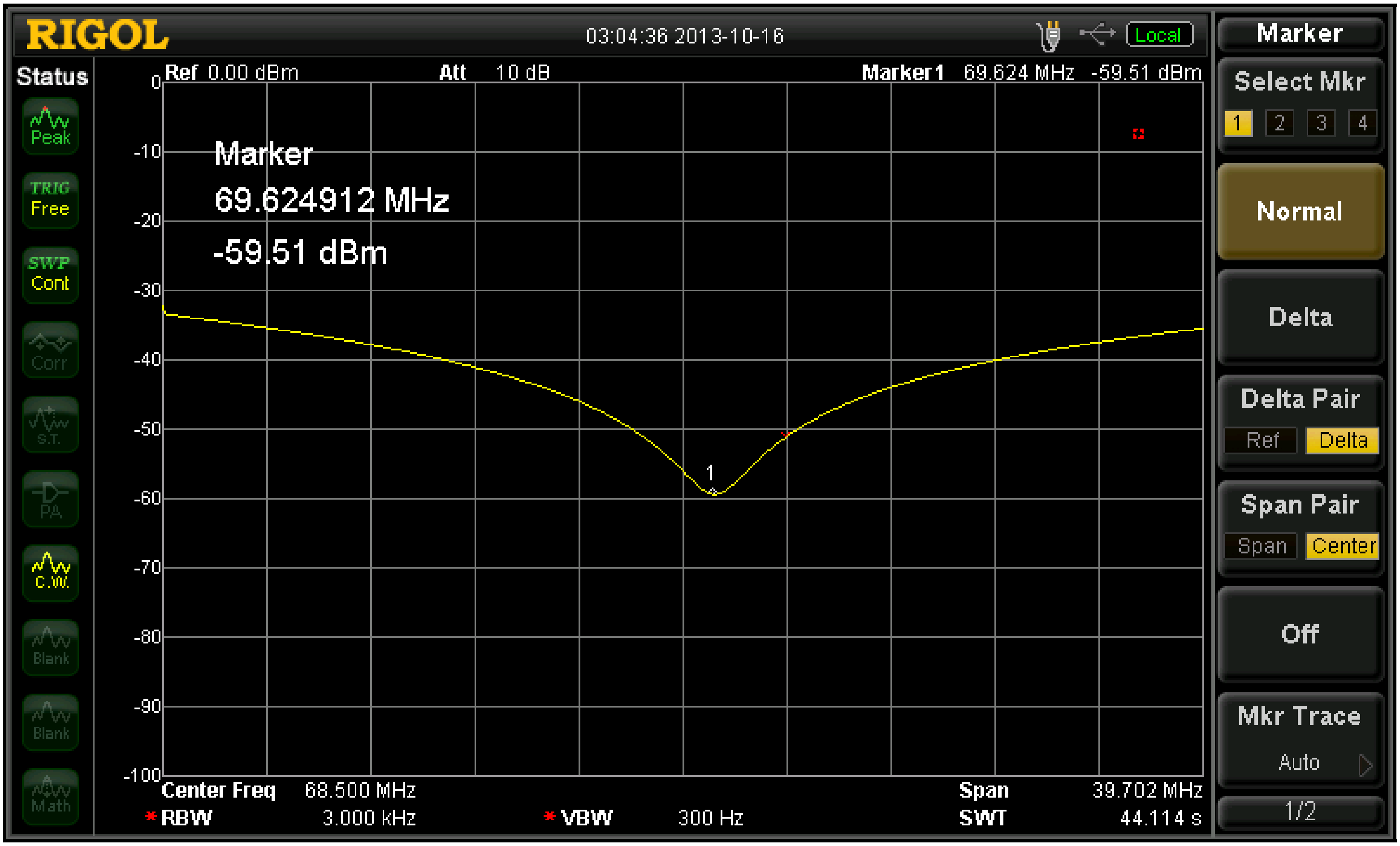
RIGOL 03:04:36 2013-10-16

Ref 0.00 dBm Att 10 dB Marker1 69.624 MHz -59.51 dBm

Marker
69.624912 MHz
-59.51 dBm

Center Freq 68.500 MHz Span 39.702 MHz
RBW 3.000 kHz VBW 300 Hz SWT 44.114 s

Local
Status
Peak
TRIG Free
SWP Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
1
Marker
Select Mkr
1 2 3 4
Normal
Delta
Delta Pair
Ref Delta
Span Pair
Span Center
Off
Mkr Trace
Auto
1/2

Above — The close-in sweep of the 0.001 µF capacitor tabled above.

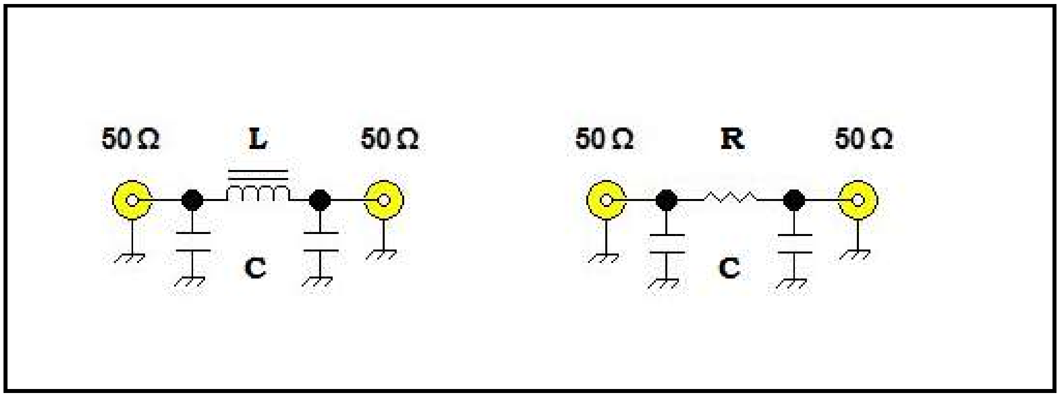
50 Ω
L
50 Ω
C
50 Ω
R
50 Ω
C

Above — As possible, stick 2 of your bypass cap values in a pi filter network with a series decoupling L or R to derive wideband filtration. For example, to filter your DC power lines.

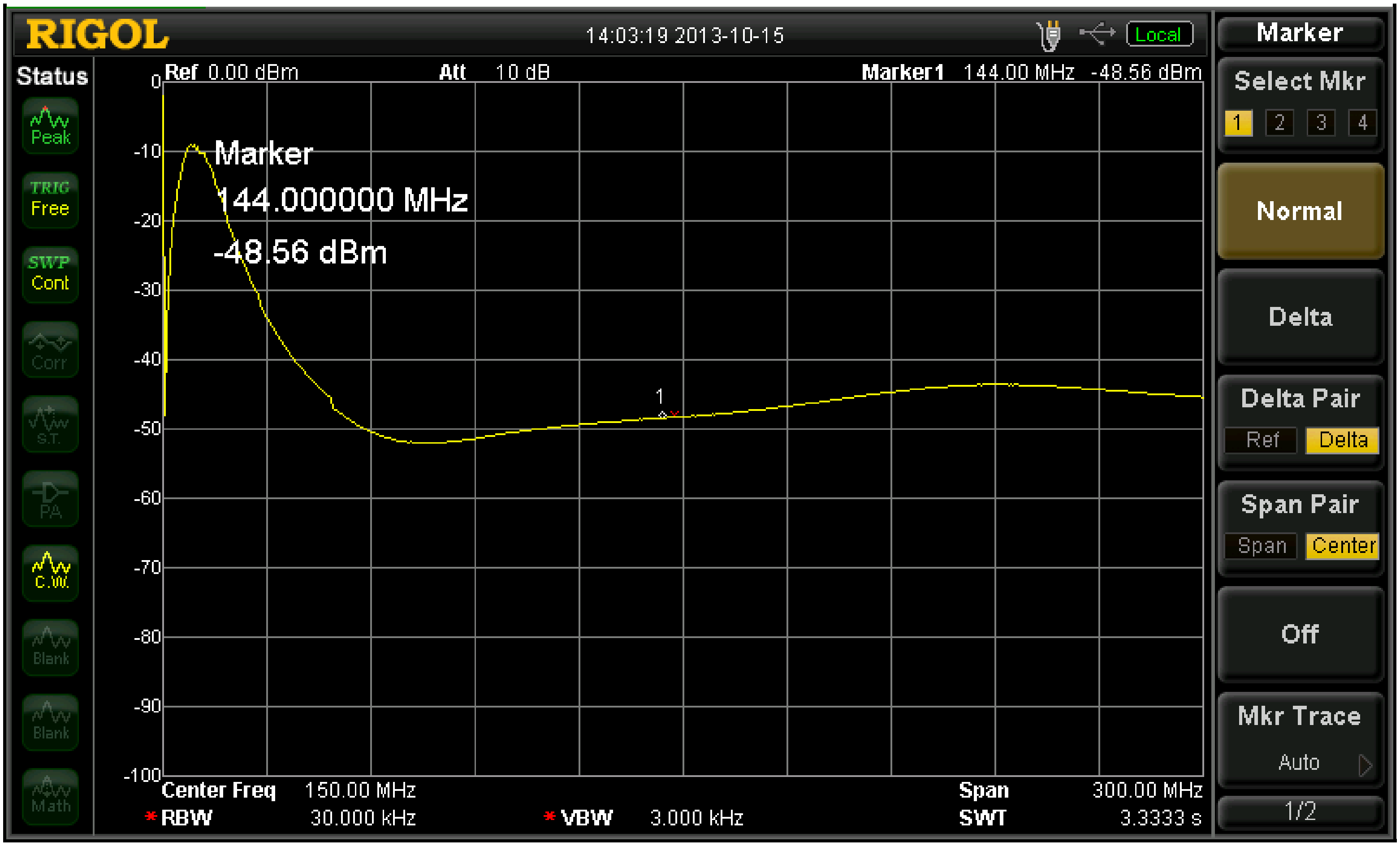
RIGOL
14:03:19 2013-10-15
Local
Status
Peak
TRIG Free
SWP Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
1
Marker
Select Mkr
1
2
3
4
Normal
Delta
Delta Pair
Ref
Delta
Span Pair
Span
Center
Off
Mkr Trace
Auto
1/2

Ref 0.00 dBm
Att 10 dB
Marker1 144.00 MHz -48.56 dBm

Marker
144.000000 MHz
-48.56 dBm

Center Freq 150.00 MHz
Span 300.00 MHz
RBW 30.000 kHz
VBW 3.000 kHz
SWT 3.3333 s

Above — A 300 MHz sweep of a pi filter [220 pF + 1.2 µH + 220 pF] for the DC supply line of a 150 MHz oscillator. The SRF peak lies at 76.69 MHz, but this filter works okay out to about 200 MHz. I placed a marker at 144 MHz and could use this filter for the 2M band as well.

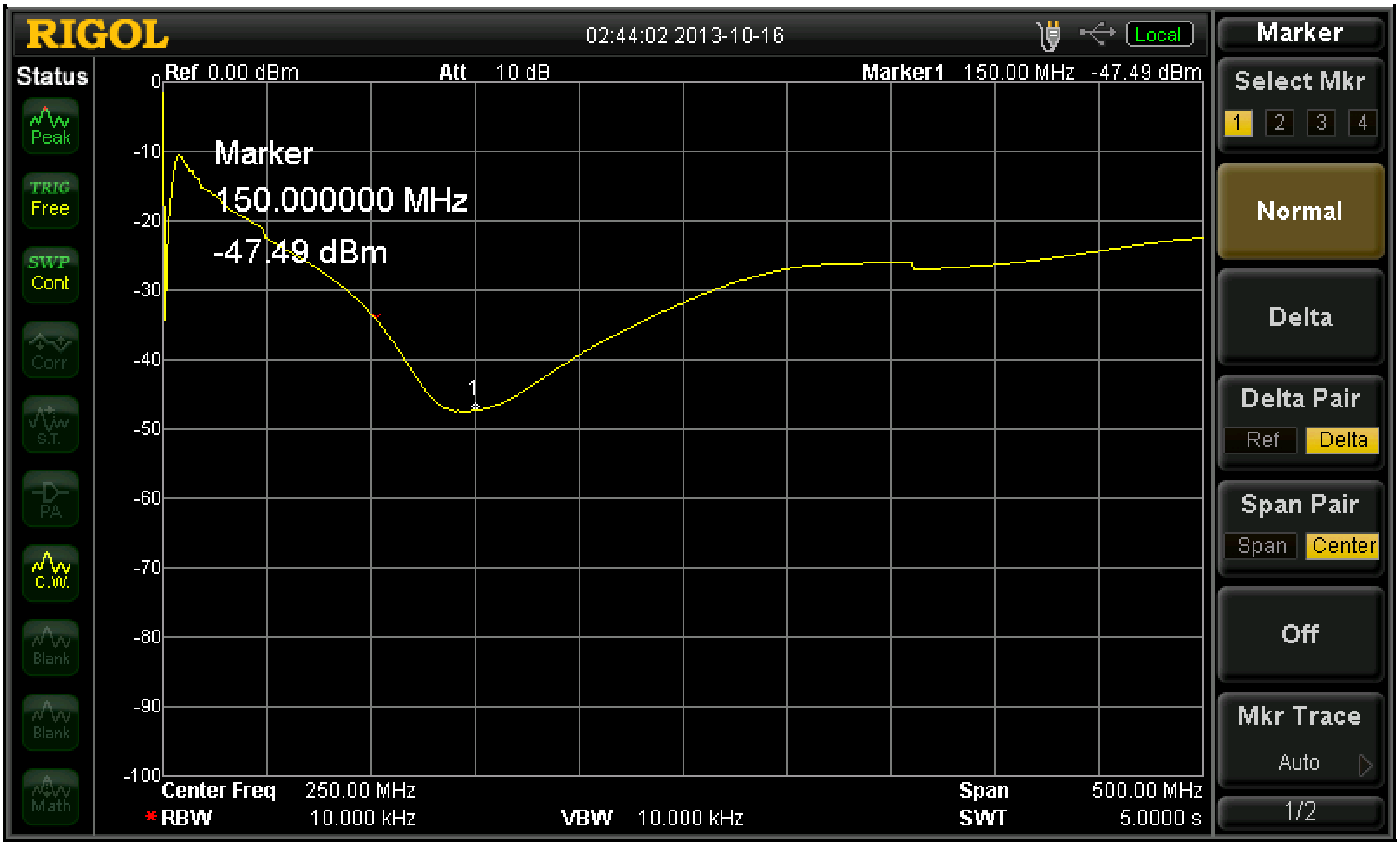
RIGOL
02:44:02 2013-10-16
Local
Status
Peak
TRIG Free
SWP Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
Marker
150.000000 MHz
-47.49 dBm
1
Marker
Select Mkr
1
2
3
4
Normal
Delta
Delta Pair
Ref
Delta
Span Pair
Span
Center
Off
Mkr Trace
Auto
1/2

Ref 0.00 dBm Att 10 dB Marker1 150.00 MHz -47.49 dBm

Center Freq 250.00 MHz Span 500.00 MHz

RBW 10.000 kHz VBW 10.000 kHz SWT 5.0000 s

Above — The network described above, except I replaced the 1.2 µH L with a size 0805 10 Ω resistor [220 pF + 10R + 220 pF] and swept to 500 MHz. I set Marker 1 on 150 MHz; the frequency of the oscillator I wanted to DC powerline filter. The resistor gives a bit more filter bandwidth around 150 MHz. A 51 or 100 Ω resistor will further increase the bandwidth while decreasing the attenuation depth somewhat. Although resistors incur a DC voltage drop, they avoid the potential of an unwanted SRF in your filter arising from a renegade inductor— and so, a resistor may pose a better choice for pi filtering DC lines and so forth. It's your call.

At HF and lower VHF, I've found a hole-through capacitor may sometimes filter better than a "garden variety" SMT counterpart. Click for a graphic that shows this. Presumably, the SMT cap exhibited a lower Q than the equivalent hole-though part. At some frequency above 200 MHz, the lead inductance of the hole-through capactor may cause the opposite effect.

Further, on the VHF — Véronique web page, Section 6: I swept 3 capacitors including an ultra-high Q SMT part.

RF — Test and Measurement

RF Workbench Page 5

Welcome to part 5 of a web series exploring basic RF measurement and bench practices. This installment builds on the information from RF Workbench Parts 1 - 4.

In RFWB #5 I share a hodge-podge of thoughts and circuits concerning power measurement on the beginner-level RF workbench. Consult EMRFD for more support. Big thanks to my mentors: Wes, W7ZOI. Bob, K3NHI and John, K5IRK for their support as I advance to the basics.

Power Measurement Empowers You

Before embracing the 50 Ω RF environment, I misjudged the need to quantify small signal power — now I get that we measure lots of low-level signals on the 50 Ω RF workbench. Whether you're driving a mixer RF port with -30 dBm to reduce spurs, or tweaking an amplifer-under-test to exact the best S21, low-level RF power measurement is fundamental to fruitful RF design.

1 way to measure low-level RF power includes building a log linear RF power meter (PM) based on the Analog Devices AD8307. The basic circuit I show posits that most of you measure from MF to HF and don't need a PM that reads flat into UHF and further; a simple, 2 chip circuit might even prompt you to actually build a barebones PM for your QRP workbench.

Search for and download the Analog Devices AD8307 Revision D datasheet — it's definitely worth a read. Kudos to the design team that brought us a truly milestone device for low-cost power measurement.

Analog Devices offers a whole family of log-amps at different frequency ranges — for example, the AD8311 Log Amp/Detector covers from ~100 MHz to 2500 MHz. A sister product, the AD8302 2.7 GHz RF / IF Gain Phase Detector looks amazing.

AD8307
AR 0512
45536

---

1. A Barebones RF Power Meter

Lacks the usual input frequency compensation network needed to keep power measurement flat over several hundred MHz — since most input compensate by attenuating HF, this simple version offers stronger sensitivity at HF than usual.

Good from MF up to about 100 MHz.

4.5v B+ to eliminate power supply decoupling issues and to give greater battery life than a 9v battery.

Minimum input power about -70 dBm. Maximum input power +15 dBm. Requires a metal box.

Big thanks to Wes, W7ZOI for letting me present his simple RF power meter. I found *numerous* AD8307-based RF power meter designs in periodicals and on the web and the writers devised simple to elaborate input compensation networks to establish a flat response out to 500 MHz — involving parts such as nH-level inductors, chip resistors and capacitors. This PM goes the opposite direction; a plain circuit for you builders who measure under 100 MHz; particularly at HF.

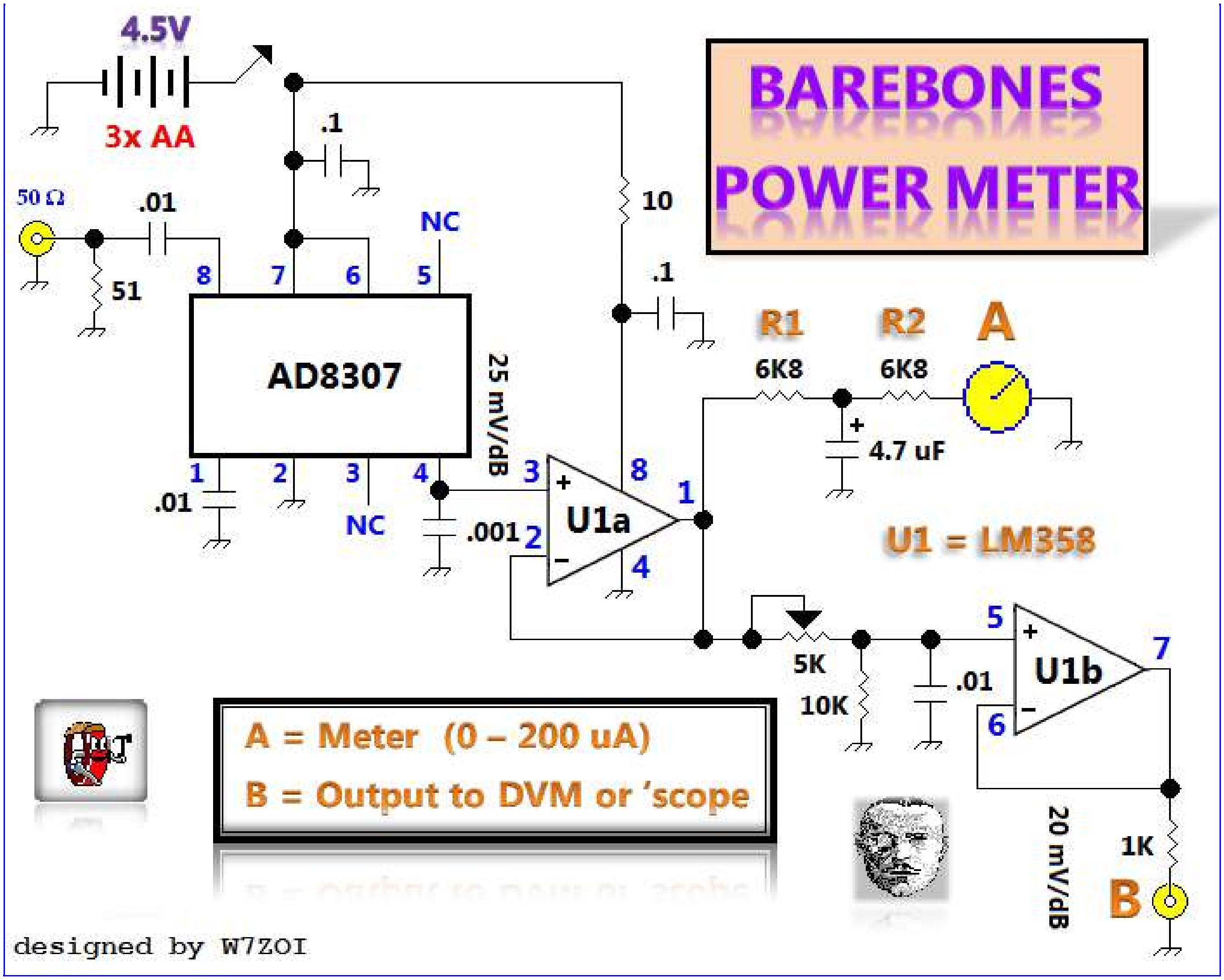
4.5V
3x AA
BAREBONES
POWER METER
.1
50 Ω
.01
NC
10
51
8
7
6
5
.1
AD8307
R1
R2
A
6K8
6K8
25 mV/dB
4.7 uF
+
1
2
3
4
3
8
+
1
.01
NC
.001
2
U1a
−
4
U1 = LM358
5
7
5K
+
10K
.01
U1b
6
−
A = Meter (0 – 200 uA)
B = Output to DVM or 'scope
20 mV/dB
1K
B
designed by W7ZOI

Above — the complete Barebones PM schematic. Presented with the permission of Wes, W7ZOI.

AD8307 Input Pin 8: Some Graphs and Notes:

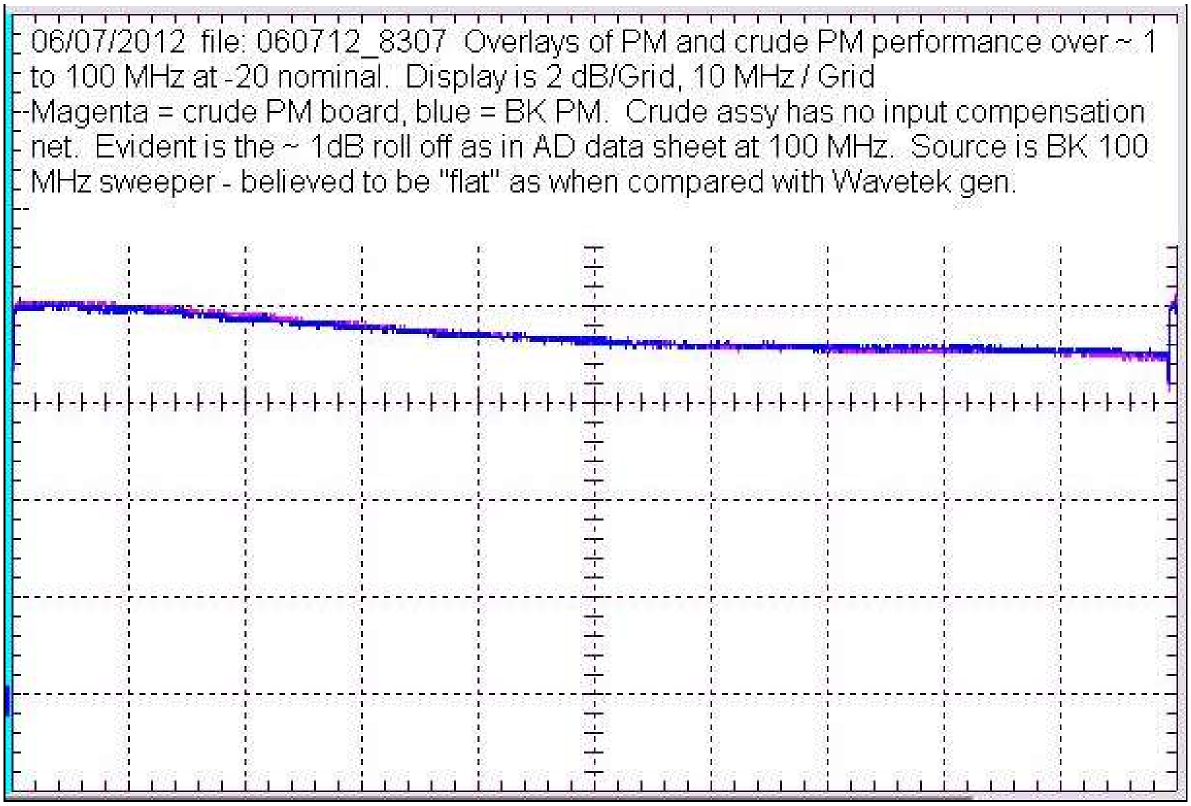

06/07/2012 file: 060712_8307 Overlays of PM and crude PM performance over ~ 1 to 100 MHz at -20 nominal. Display is 2 dB/Grid, 10 MHz / Grid
Magenta = crude PM board, blue = BK PM. Crude assy has no input compensation net. Evident is the ~ 1dB roll off as in AD data sheet at 100 MHz. Source is BK 100 MHz sweeper - believed to be "flat" as when compared with Wavetek gen.

Above — The lack of input compensation might raise a few eyebrows! A sweep of the Barebones PM by Bob, K3NHI shows a reasonably flat response out to 100 MHz that's comparible to some of the 0-100 MHz range of wider sweeps in the published compensated circuits I've read.

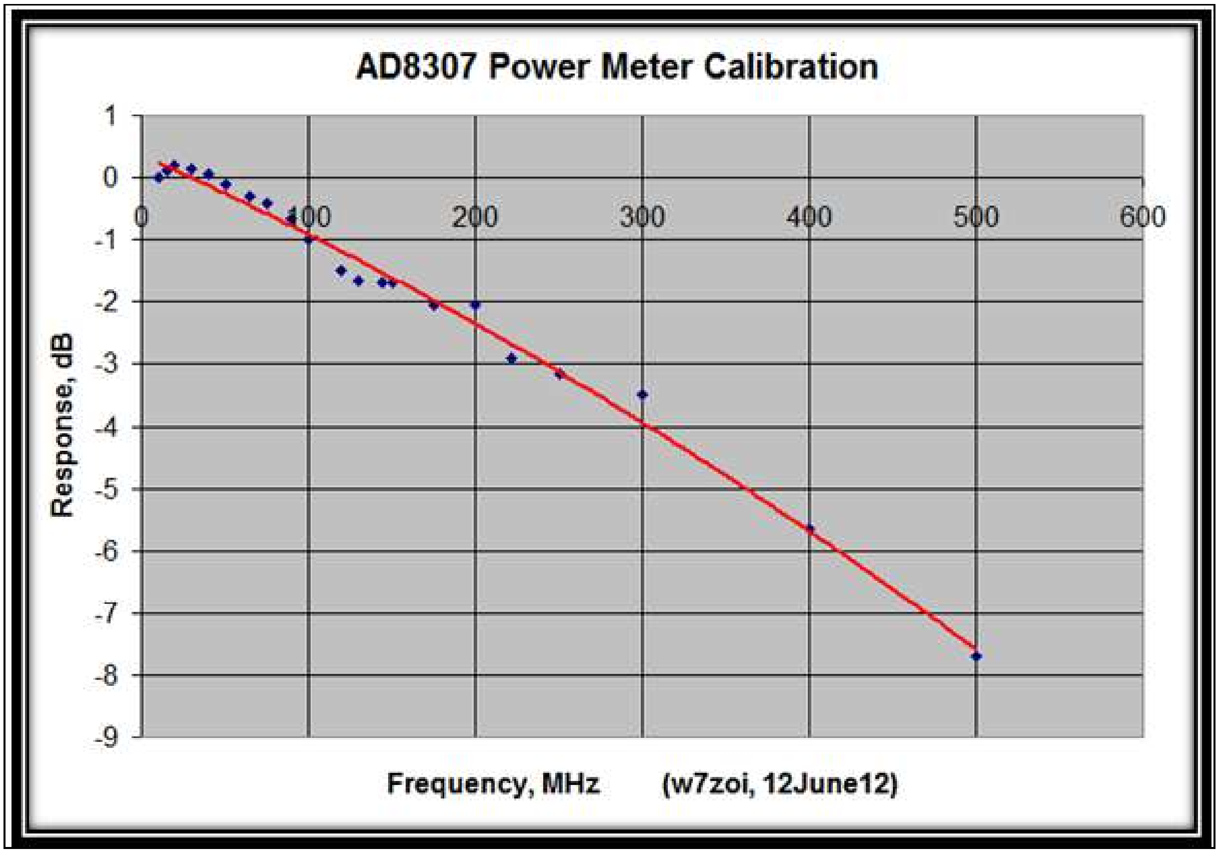

Above — A different Barebones PM plot from Wes, W7ZOI that goes out 500 MHz. For HF work, this power meter proves adequate for QRP HomeBuilders. By all means, add some input compensation if you want to — good circuit examples abound. For example, EMRFD Figure 7.13, or Bob Kopski's — An Advanced VHF Wattmeter referenced in Section 5.

AD8307 specifications allow a typical +/- 0.3 dBm "ripple" in the input to output transfer characteristic. Bob, K3NHI verified this ripple in his lab.

Therefore; depending on the position of a signal in the transfer curve, a move between 2 different power levels may yield as much as 0.6 db peak difference error. Figure 8 of the Rev D datasheet shows this. Figure 24 plus associated text tells why this occurs: the transfer characteristic is a chunk or segmented realization of a log transfer characteristic. It's really good — just not perfect.

If you consider the cost of thermal sensor power meters from Agilent, Techtronix, or even the Mini-Circuits PWR-6GHS, the AD8307 seems a bargain.

Measuring U1a Pin 1 With a Panel Meter

The first output at Pin 1 drives nearly any junkbox DC panel meter you might own The schematic specifies a 0-200 uA movement, however a 1 mA meter movement also works by tweaking R1 and R2 to establish the correct current. In the circuit shown the R = 13.6K. This R establishes the current required to drive the 200 uA meter Wes used in his design — nothing more. A wide variety of panel meters work because the drive current comes from an op-amp. Don't order an expensive panel meter just because you want to employ a 100 or 200 uA model — choose your meter because it affords good needle movement and resolution to allow accurate power measurement.

Although elementary for some, the following diagram shows you how to measure the maximum current for your meter: 100 uA, 200 uA, etc.

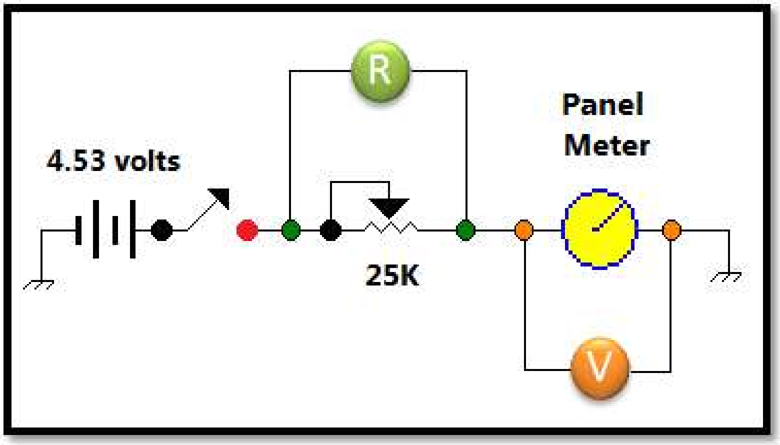

Calculate the maximum current and internal resistance of a panel meter.

1. Adjust 25K pot to get a full needle deflection.
2. Measure the pot resistance with your DVM *R Pot*
3. Measure the voltage across the meter *V Meter*

Example calculation with V Battery = 4.53v
V Meter = 0.24v and R Pot = 22.5K Ω

Current at Maximum Deflection

Formula: (V Battery – V Meter) / R Pot

4.53v – 0.24v / 22500 Ω = 0.0001906 A

or 191 µA

Ohm's Law

Above — In my shoebox sat 4 different panel meters. I measured 1 meter and show the math above — a 191 µA meter.

DIRECT CURRENT
0
.2
.4
.6
.8
1
Simpson

Above — An antique 1 mA panel meter driven with 1 mA to achieve full deflection.

Meter resolution presents the biggest problem for junkbox panel meters and if you look around the web, you'll see some great examples how builders calibrated and/or marked the scale on their AD8307 circuit panel meters.

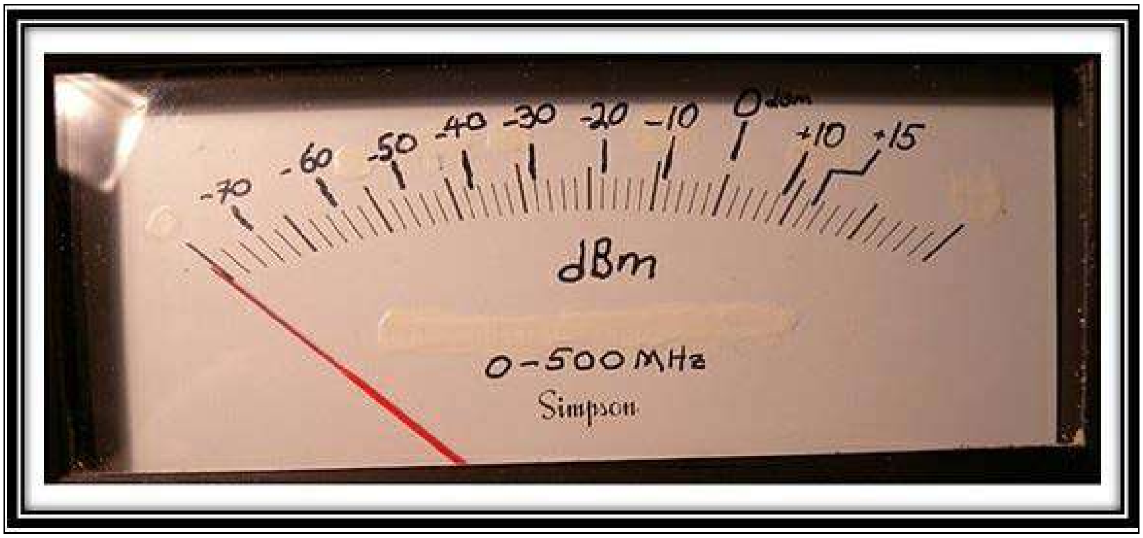

Above — Panel meter markings on my friend Peter's power meter.

I personally like my panel meters to read about 75% of full deflection at 7 dBm, but it's really your choice. Adjust R1 and/or R2 up or down to give the desired amount of meter deflection for whatever panel meter you own.

You may apply Ohm's law to figure out the maximum in-situ current for any panel meter. In the schematic the total R is 13.6K, so when you apply +10 dBm applied to the power meter input, the current in the panel meter will be 2.0v /13.6K = 147 uA. The maximum 200 uA will occur with a Pin 1 voltage of 2.72v. The op-amp won't go all the way to the positive rail, but If it did go all the way to 4.5v, you would hurl 331 uA into the panel meter. This exceeds the meter's uppermost scale, but likely won't destroy it.

The 4.7 uF cap can be any small uF capacitor value and low-pass filters the DC to smooth out the meter movement. I used a 10 uF in my breadboard.

I normally view the panel meter to tune resonant circuits and observe trends; but not to precisely measure power — it's often more accurate to quantify power readings at Port B since this eliminates panel meter resolution issues.

How to Measure Power at Port B

We measure the DC voltage at port B and use equations or graphs to translate this voltage into an actual power reading after reference calibration. U1a and U1b are unity-gain voltage followers to buffer the AD8307 output. U1B features a 5K potentiometer in series with a 10K shunt resistor to form a voltage divider that changes the 25 mV/dB AD8307 output to 20 dB/mV. Tweak the 5K pot to calibrate to get as close to 20 dB/mV as possible, although since the LM358 buffers just pass on the DC voltage changes of the AD8307, the pot does not technically alter log linearity.

Wes published some essential notes on his web site. On page 8, you find his formula to convert a measured DC voltage into dBm. I wrote a program that incorporates his formula: Applet L on my Design Center web page; except that it takes calibration power at -10 dBm and -20 dBm.

Let's run through 1 example:

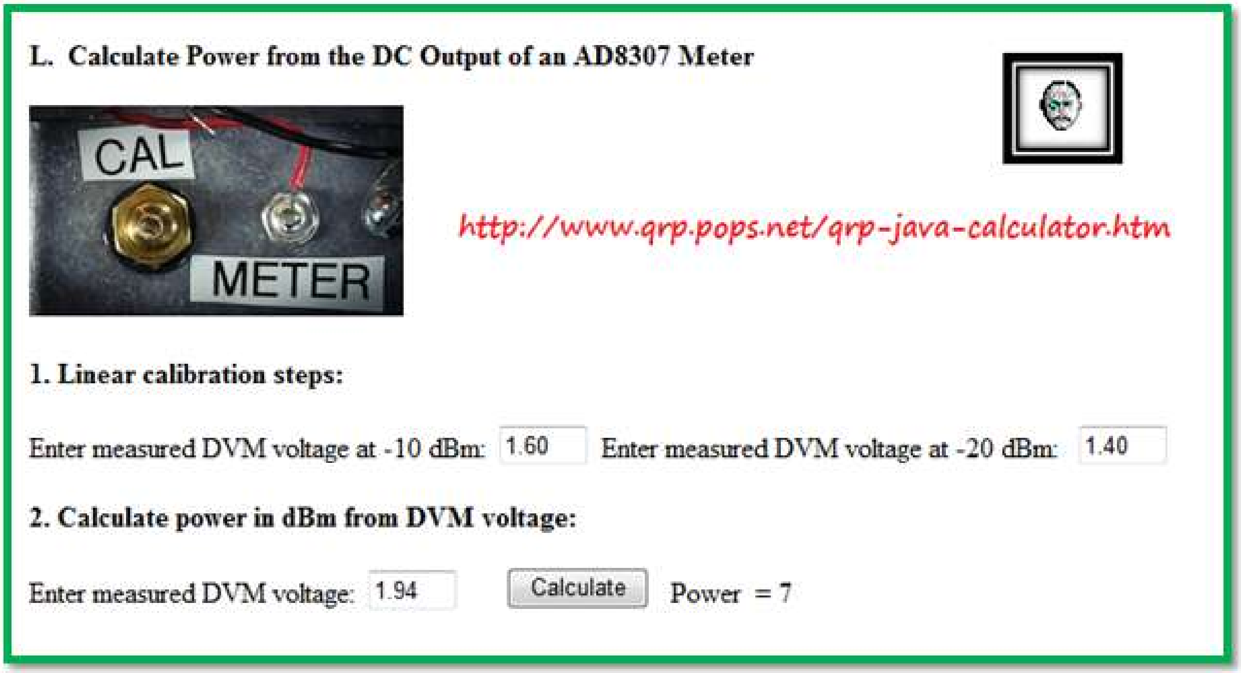

L. Calculate Power from the DC Output of an AD8307 Meter

http://www.qrp.pops.net/qrp-java-calculator.htm

1. Linear calibration steps:

Enter measured DVM voltage at -10 dBm: 1.60 Enter measured DVM voltage at -20 dBm: 1.40

2. Calculate power in dBm from DVM voltage:

Enter measured DVM voltage: 1.94 Power = 7

Above — A screen capture of JavaScript Applet L. I calibrated my instrument by connecting a 10.0 MHz sine wave calibrator to the 50 Ω input port of my PM. At -10 dBm input power, my Port B reading was 1.6 VDC; at -20 dBm, I measured 1.4v at Port B. These 2 points establish the log linearity per Wes' notes. After this calibration, I measured the output power of a 4 MHz sine wave VFO that I designed to run at 7 dBm output power to drive the LO port of a diode ring mixer. Port B read 1.94 volts and when I entered this into Applet L and pushed <<Calculate>>, the VFO output power was indeed 7 dBm. More on this later...

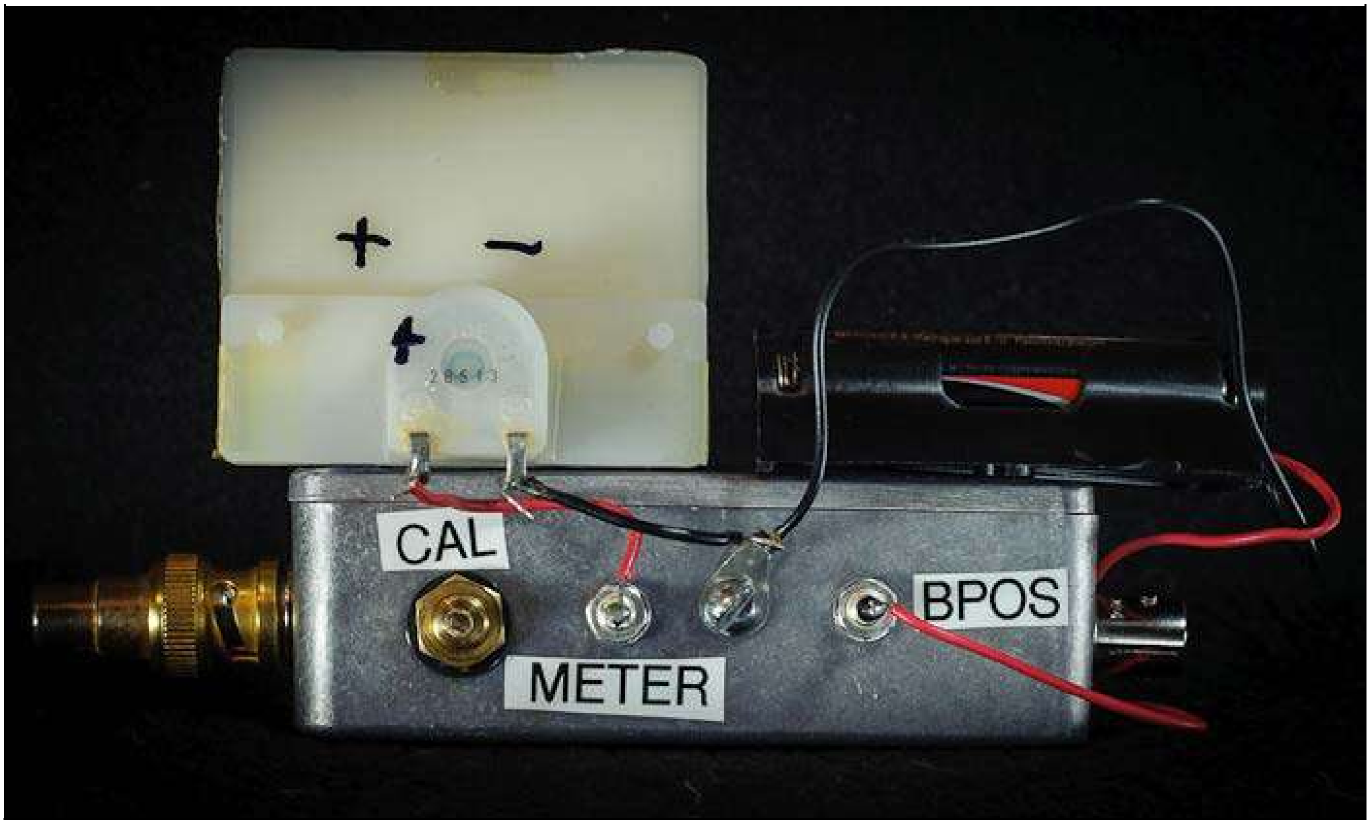

Above — My prototype RF-tight breadboard of the Barebones PM. I ran my 'trademark' die case box, BNC RF input and Port B connectors, a common ground lug and feedthrough capacitors for the B+ and panel meter connections. I placed a BNC to RCA adaptor on Port B to allow the insertion of a standard-type positive DVM probe. A black alligator clip terminates the distal end of the negative probe on all my DVMs. I just clip it onto the ground lug.

−
20 10 7 5 3 0 3
+
20 50 100%
VU

Above — A view of my breadboard showing my 100 uA panel meter and the 4.5v battery pack. I raised R2 in the original schematic up to 8K2 to set my preferred needle movement in this 100 uA panel meter. A meter with linear markings might be better?

Some Further AD8307 Notes

After power up, you'll notice a DC output voltage around 0.22v or so with no input signal; this arises from wideband noise caused by resistors and amplifiers in the AD8307— all normal.

If possible, verify a 25 mV per dB power change with a manual sweep using a sine wave signal generator on the input and a DVM to AD8307 Pin 4. If you don't have the test gear to perform this function check, no problem.

The AD8307 input resistance = 1100 Ω , so we must place a resistor in parallel with the input to establish 50 Ω. Many builders just shunt a short-leaded 51 Ω resistor from input to ground like Wes did, however, you might also see builders place a 1% tolerance 53.2 Ω R in that slot to derive an input Z of 49.9 Ω. I did this in another "blinged-out", frequency compensated AD8307 meter I built for future UHF circuit experiments.

With 20 and/or 40 dB taps, attenuator pads and some 50 Ω cables, a Barebones AD8307 PM can measure everything HF you might build or buy for your *QRP* workbench. I posit that a simple AD8307 power meter may form the heart of a basic, first QRP workbench. Lacking a oscilloscope when I started in radio electronics, I measured RF with a germanium diode RF probe and a DVM — I would have enjoyed a simple log power meter plus a basic calibrator, however, the AD8307 did not exist back then.

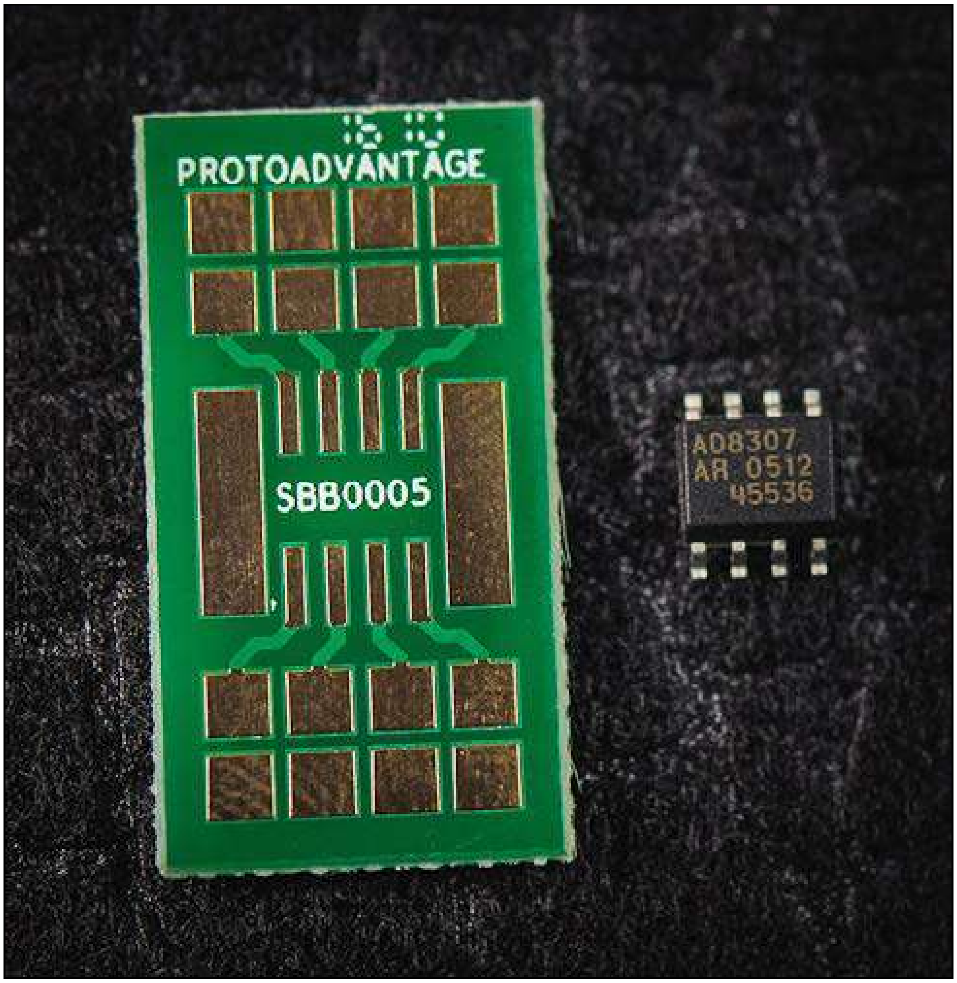

Above — Some builders lament because an AD8307 (in DIP) costs around $10.00. I bid for and bought the AR version chip shown above for $4.42 USD. The above green Proto Advantage breakout board cost $1.00. To compare; some people spend $5.00 for a boutique coffee in Canada — it's all good.

The SMD packaged AD8307AR may offer better performance above 100 MHz with its lower lead inductance.

QRP — PosData for Dec 3, 2013

Realize your bench acumen — adapt and build some test equipment to suit *your* needs. Steve, VK2SJA crafted a version of the AD8307 PM. Click for his beautiful box featuring a 1mA meter movement. Aided by the Tonne software "Meter" program and the AD8307 datasheet, his ranks among the best I've seen. Kudos Steve.

---

2. Power Meter Calibrators

1. CMOS Clock Oscillator RF Calibrator @ -10 dBm

Bob, K3NHI designed a 10.0 MHz reference oscillator for -10 dBm that might help equip the beginner bench. This CMOS signal source does not need an AC power reference to calibrate it — just a DVM to make a DC measurement and your good to go!

In addition to calibrating an AD8307-based PM, you might use this -10 dBm reference to calibrate other gear including homebrew sine wave signal generators; that's what I'll do later.

MX045
10M0000
CTS 3C
1030 12A
MX045
10M0000
CTS 3C
1030 12A

Above — The recommended CMOS clock oscillator for the CMOS RF calibrator. Digi-Key part # CTX772-ND. Data sheet.
If you substitute another CMOS clock, it must swing nearly rail-to-rail for accuracy. Bob intended this calibrator for those who lack the bench instruments needed to precisely calibrate RF devices since only a DVM is needed for calibration. This generator also offers a range of calibrated harmonics as Bob described in his January-February QEX article A Simple RF Power Calibrator — great when you want to examine a spectrum analyzer over a limited span.

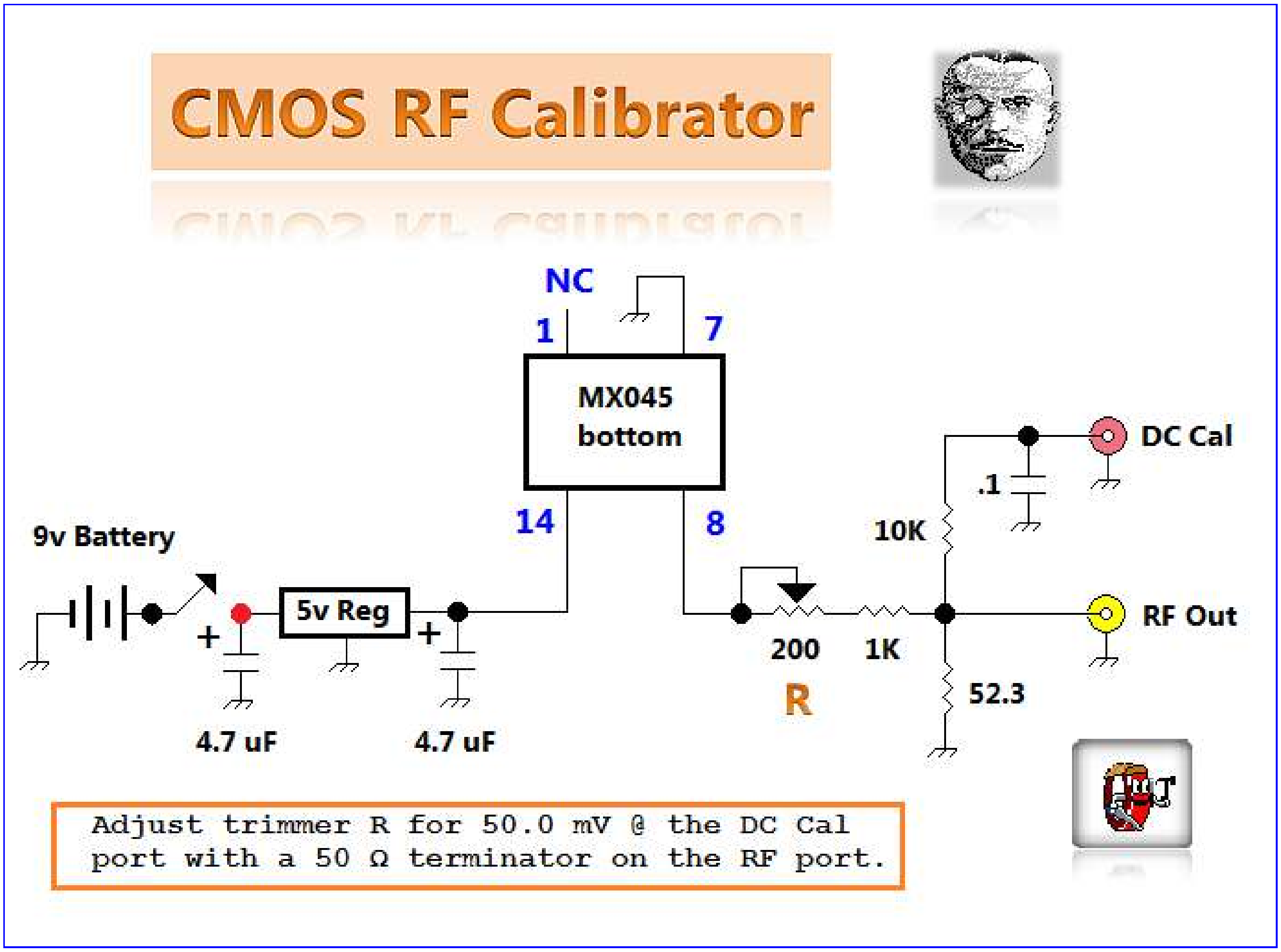
CMOS RF Calibrator
NC
1
7
MX045
bottom
14
8
9v Battery
5v Reg
+
+
4.7 uF
4.7 uF
.1
10K
DC Cal
RF Out
200
1K
R
52.3

Adjust trimmer R for 50.0 mV @ the DC Cal port with a 50 Ω terminator on the RF port.

Above — My version of the K3NHI CMOS signal source (presented with the permission of Bob, K3NHI). The 52.3 Ω resistor is a standard 1% part and I bought 5 for all my AD8307 projects and this little signal generator. In my first version, the trimmer resistor was 500 Ω and worked okay, but the 200 Ω trimmer improved calibration. I ordered 10 Bourns 200 Ω and 10 Bourns 500 Ω trimmers on eBay for a few dollars and after studying Bob's published work and applying his influences to my own, I now love to precisely calibrate or bias circuits with a 200 or 500 Ω trimmer R as appropriate.

QRP — PosData for November 22, 2012

If required, you may substitute a 49.9 to 51 Ω resistor for the 52.3 Ω specified.

Per Bob's QEX 2010 Tech Notes and his emails specifically about calibrating the AD8307 with a CMOS *square wave***:** Modern AD8307 chips are better calibrated with a -10 dBM square wave. To test log linearity after CMOS signal generator calibration, apply a sine wave signal generator to your power meter and adjust its output to get the same power meter DC output voltage as with the CMOS generator. Then insert attenuator pads on the now calibrated sine wave generator to assess the mV/dB change with different power levels.

Bob's original CMOS calibrator outputted - 20 dBm, however, he updated it to output -10 dBm in 2010 as reported in his QEX 2010 Tech Notes.

Some readers have asked why calibrate the AD8307 at 10 MHz? Calibrate at whatever frequency you want, or more than one. However, at 10 MHz, the AD8307 exhibits its best log performance compared to other frequencies. Click or click for datasheet graphs.

EQUUS
#4320
.050
www.iequus.com
AUTOPOWER-OFF
10 MegOhm Input Impedance
DCV
500
20
2
200m
OFF
500
200
ACV
BAT.
12V
6V
1.5V
2m
DCA
mA
20m
200m
10A
20M
200K
2K
OHM
Ω
200
©IEC 2002

Above — The DVM calibration port reading (2 volt scale) from my CMOS clock calibrator after calibration. That was easy!

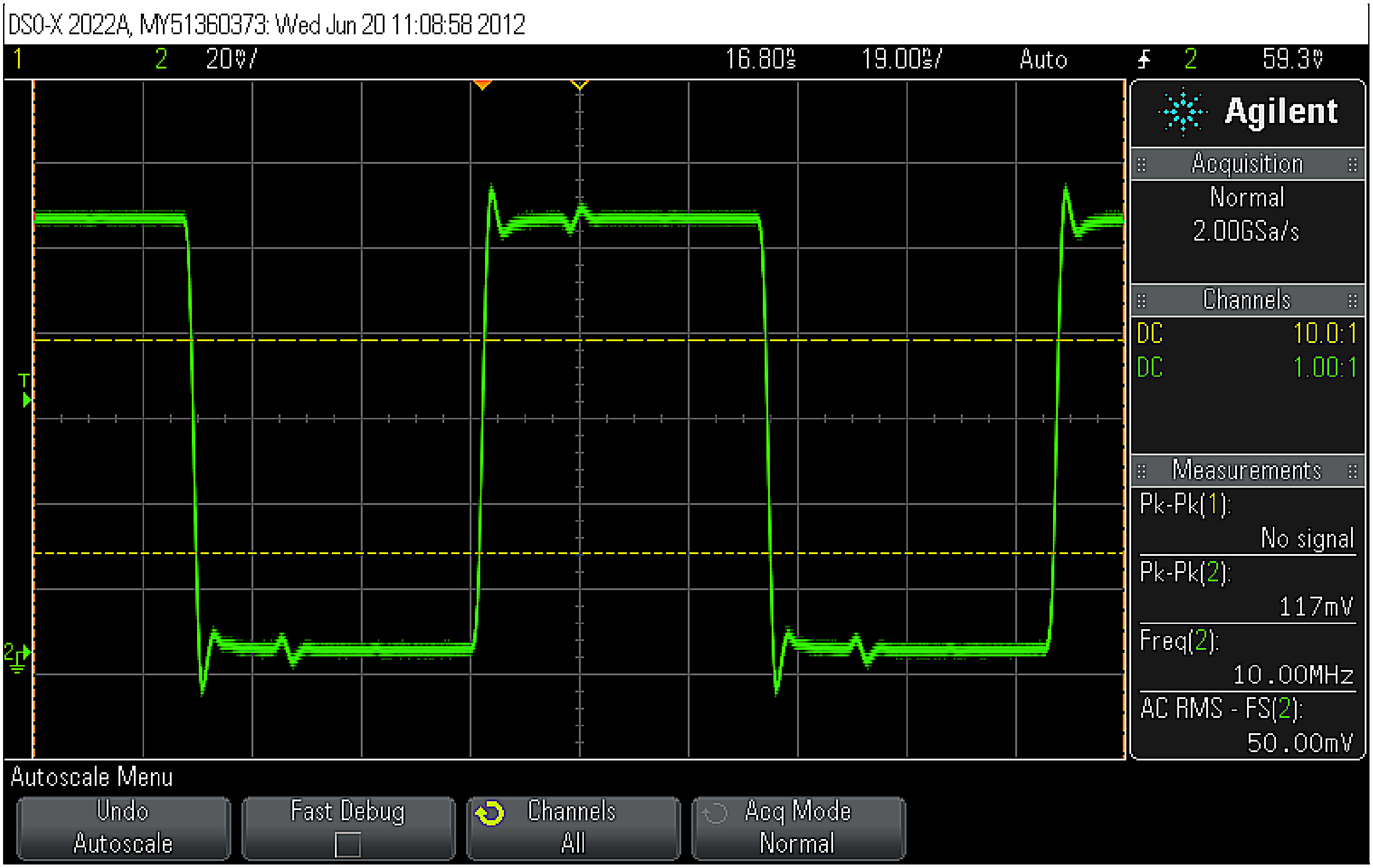

DSO-X 2022A, MY51360373: Wed Jun 20 11:08:58 2012

1 2 20mV/ 16.80ns 19.00ns/ Auto 2 59.3mV

Acquisition
Normal
2.00GSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Pk-Pk(1): No signal
Pk-Pk(2): 117mV
Freq(2): 10.00MHz
AC RMS - FS(2): 50.00mV

Agilent
T
2
Autoscale Menu
Undo Autoscale
Fast Debug ☐
Channels All
Acq Mode Normal

Above — The calibrated CMOS clock RF calibrator in my 50 Ω terminated scope.

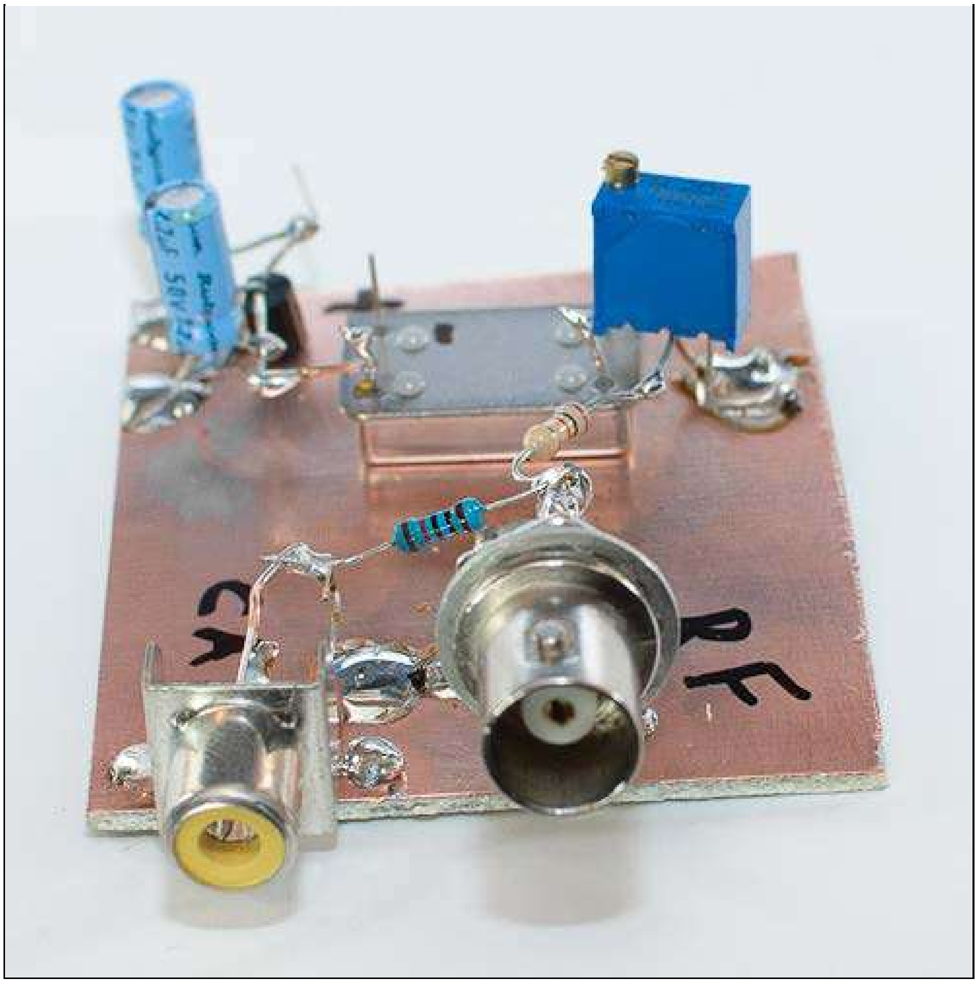

Above — The breadboard of my version of the K3HNI CMOS -10 dBm RF calibrator. For best results, stick it in a shielded box. When correctly calibrated, it outputs -10 dBm on an AD8307 PM, -14 dBm on a spectrum analyzer and -13 dBm on a conventional, thermal-sensor power meter.

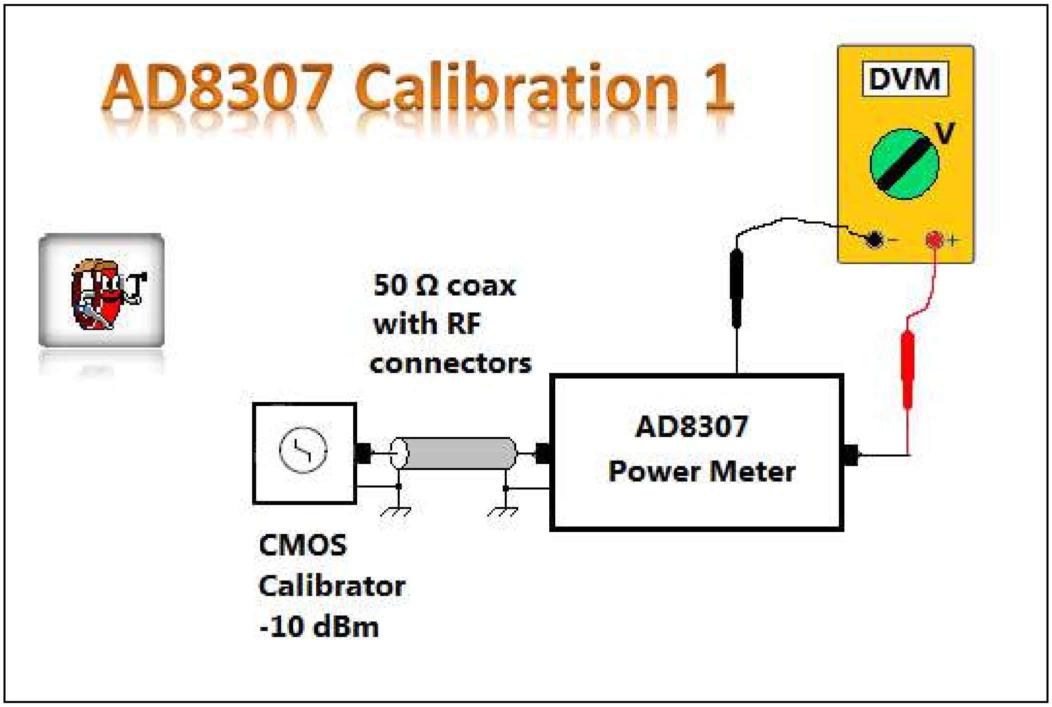
AD8307 Calibration 1
DVM
V
-
+
50 Ω coax
with RF
connectors
AD8307
Power Meter
CMOS
Calibrator
-10 dBm

Above — The procedure for calibrating an AD8307 PM with the -10 dBm square wave CMOS signal source.

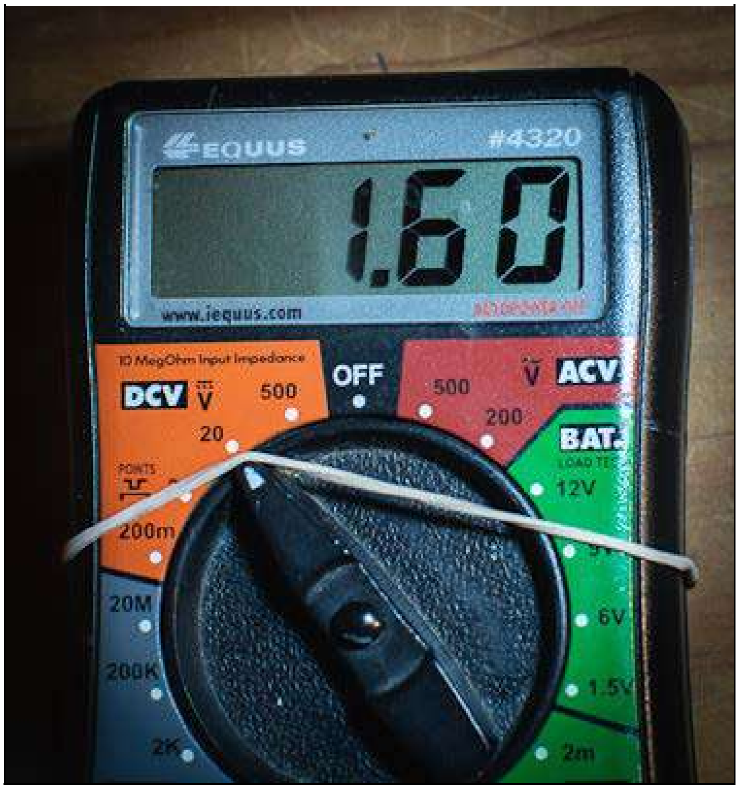

Above — When I connected the calibrated CMOS clock RF calibrator to my build of the Barebones PM, I measured 1.60v with my DVM; ***the -10 dBm reference voltage.*** Since the log-linear power changes 20 mV/dB; for my calibration reference voltage: 1.40v = -20 dBm, 1.20v = -30 dBm, 2.0v = 10 dBm etc.

Remember that the Barebones PM runs on a battery pack and over time the B+ will change. *Each tim*e I measure power with the meter I first calibrate it to establish the -10 dBm reference voltage.

Here's a simple formula that only works for 20 mV/dB @ my particular 1.60v calibration voltage, but gives you the general idea:

Power in dBm = 50 x (V - 1.8)

So if I measure 1.94v: 50 x (1.94 - 1.8) = 7 dBm.

2. Sine Wave Oscillator for Calibration @ -10 dBm and -20 dBm.

You may also calibrate your AD8307 PM with a calibrated sine wave signal generator. Advanced builders who own the gear needed to measure RF power tend to use a sine wave for calibration.

It's easy to calibrate the AD8307 with a sine wave signal source. Normally we calibrate our sine wave signal generators with instruments such as a 50 Ω terminated 'scope, a spectrum analyzer, a calibrated power meter, or a 49.9 to 51 Ω terminating resistor plus a 10X probe etc., but if you lack these instruments, your stuck.

No problem. You may calibrate any appropriate sine wave oscillator at -10 dBm with your Barebones PM and the CMOS RF Calibrator shown earlier. Let's examine the procedure:

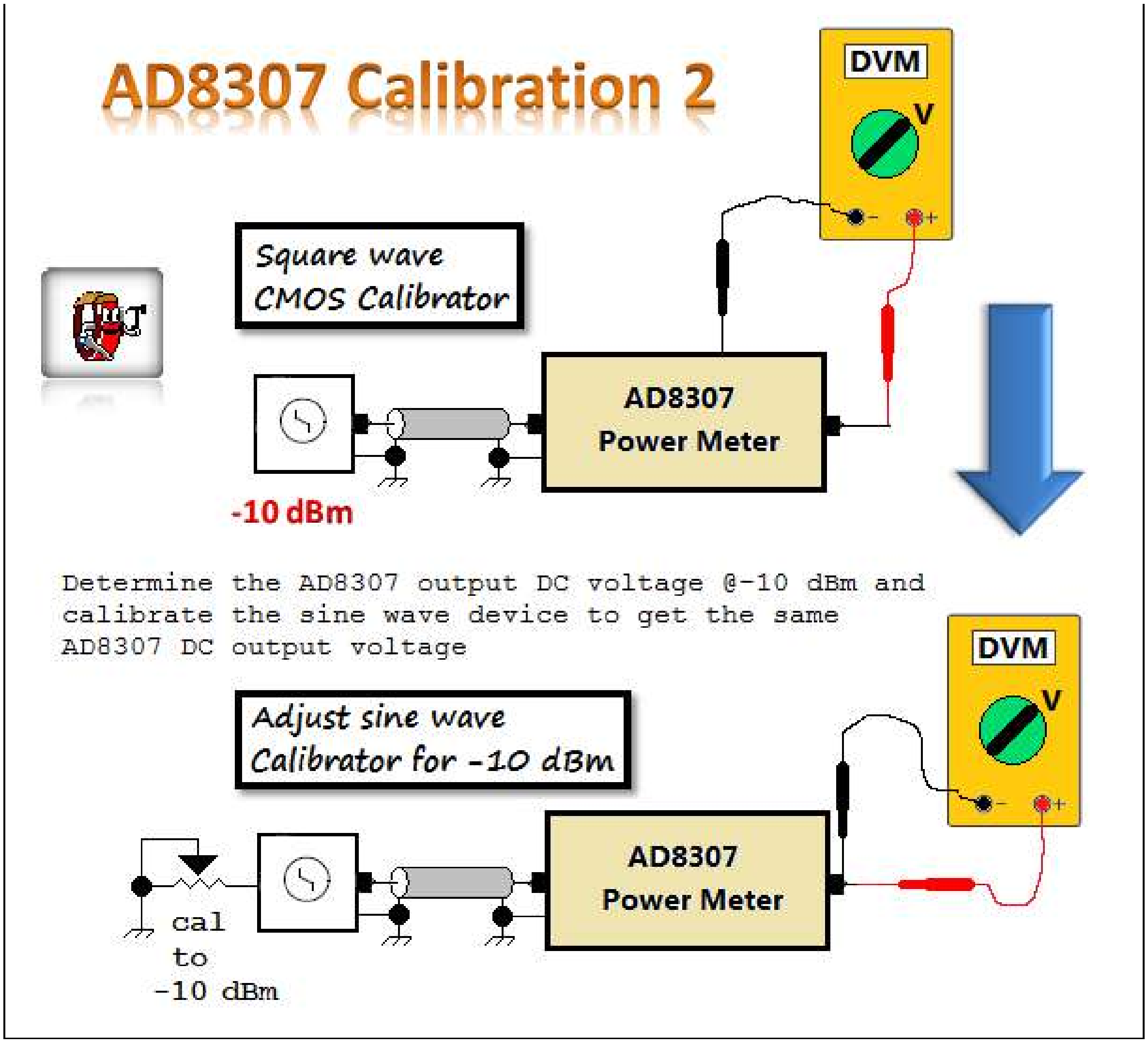

AD8307 Calibration 2

DVM
V
Square wave
CMOS Calibrator
AD8307
Power Meter
-10 dBm

Determine the AD8307 output DC voltage @-10 dBm and calibrate the sine wave device to get the same AD8307 DC output voltage

Adjust sine wave
Calibrator for -10 dBm
DVM
V
AD8307
Power Meter
cal
to
-10 dBm

Calibrate the AD8307 PM with the CMOS square wave reference and record the DC voltage at the output of Port B. Connect up your sine wave generator and adjust its output until you get the same Port B DC voltage as the reference CMOS RF signal generator — your sine wave oscillator should now be calibrated to -10 dBm.

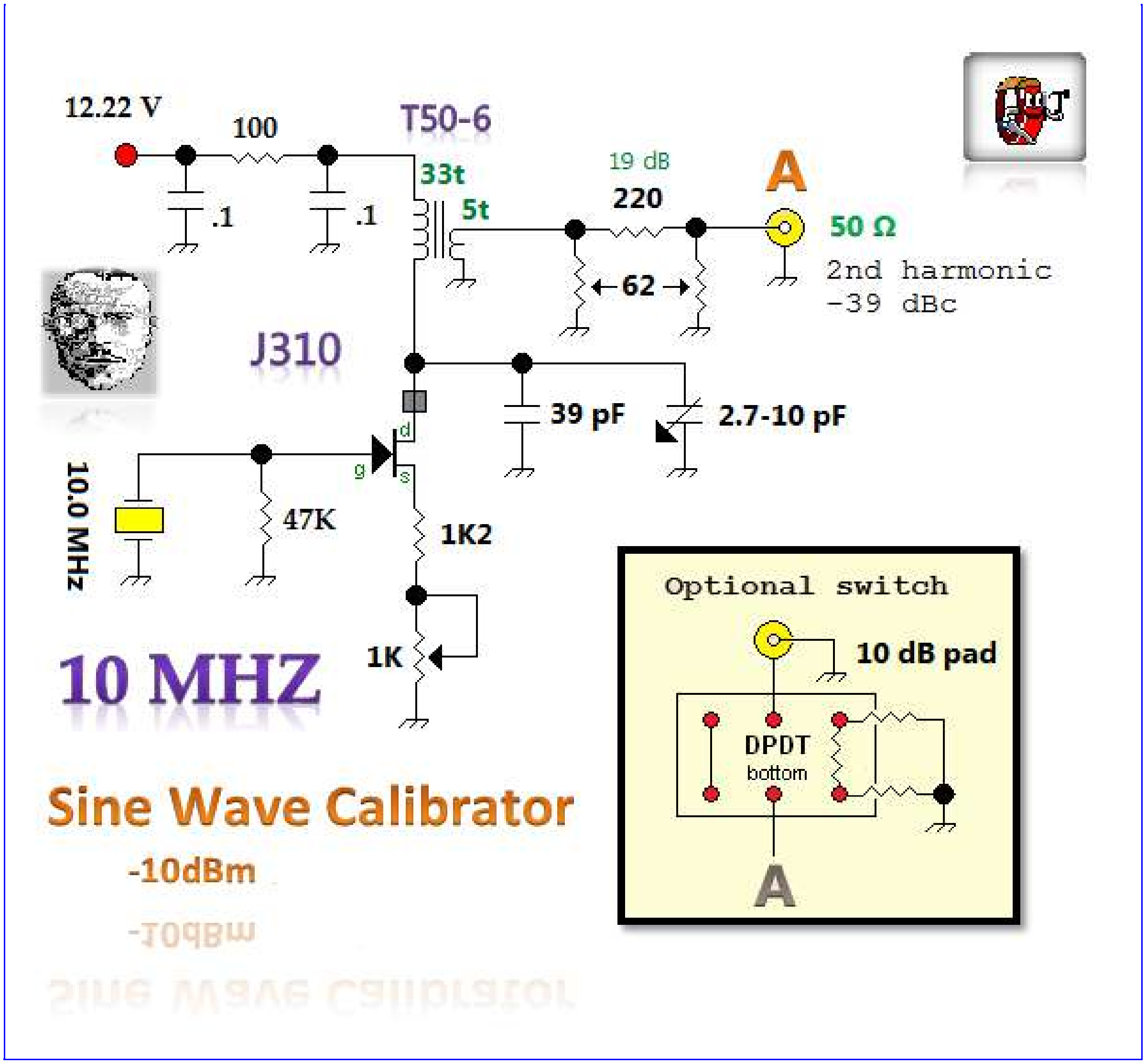
12.22 V
100
T50-6
.1
.1
33t
5t
19 dB
220
A
50 Ω
62
2nd harmonic
-39 dBc
J310
d
g
s
39 pF
2.7-10 pF
10.0 MHz
47K
1K2
1K
10 MHZ
Sine Wave Calibrator
-10dBm
Optional switch
10 dB pad
DPDT
bottom
A

Above — I designed this simple 10.0 MHz sine wave calibrator to evaluate Bob's CMOS square wave RF calibrator and serve as an example sine wave reference oscillator. When running a regulated VCC of at least 12v and the *tank is perfectly* tuned, low distortion arises. I measured the second harmonic @ 39 dBc down. The L with 33 turns = 4.43 uH.

The initial tuning procedure goes like this:

1. Terminate the output with a 50 Ω resistor terminator, or a 50 Ω terminated 'scope, or your AD8307 PM.
2. Connect an ammeter between the VCC node and the regulated power supply.
3. Adjust the emitter trimmer R so that the circuit draws around 2.7 mA — then disconnect the ammeter leads.
4. Adjust the trimmer cap for the highest pk-pk voltage (and/or or best looking waveform) in the 'scope, or highest power in the AD8307 PM. *Nominal total C to resonate my particular circuit was ~ 45 pF.*

If not already done, connect the tuned-up sine wave signal source to a AD8307 PM. Adjust the trimmer potentiometer so the Port B DC voltage = the reference voltage measured during your CMOS RF generator power meter calibration. Your sine wave signal source is now calibrated to -10 dBm.

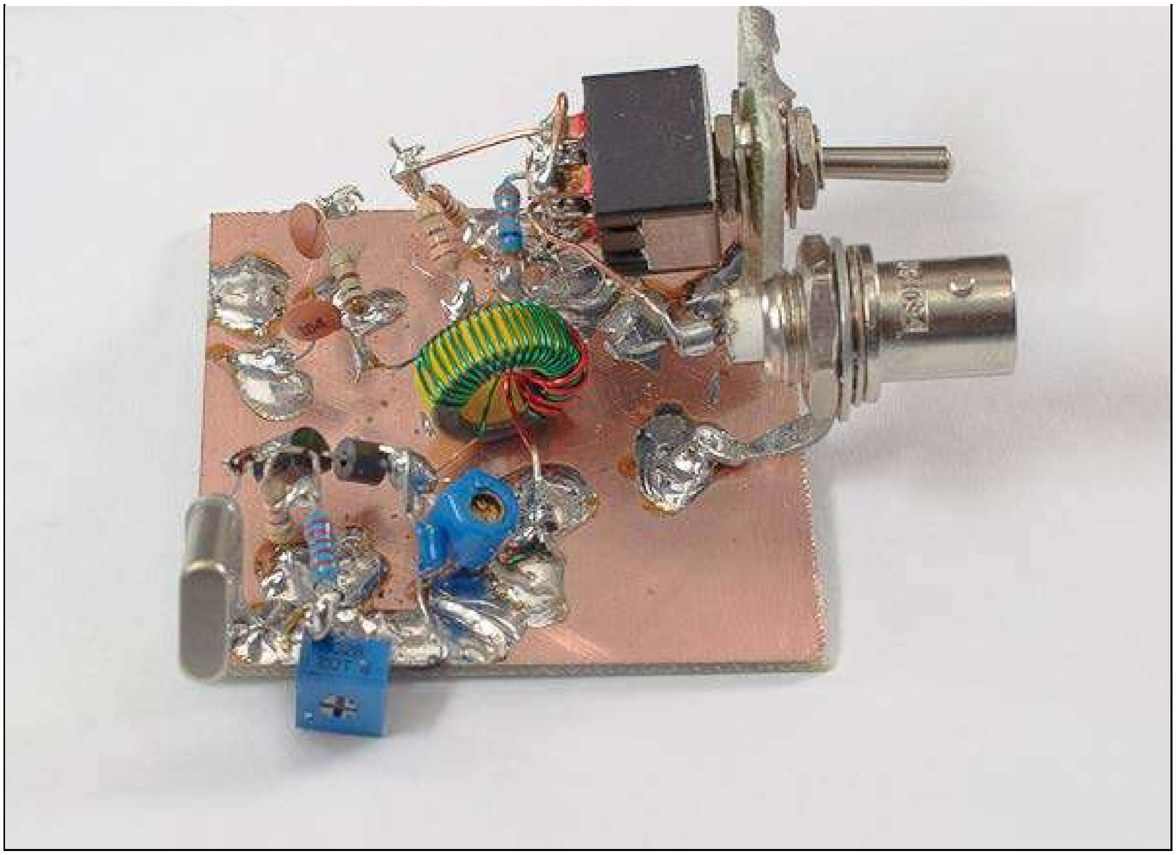

Above — The breadboard of my 10 MHz sine wave signal source. I included the optional switched 10 dB attenuator shown on the schematic inset. In the end I decided to just stick a removable **10 dB** pad like this 6 dB pad in-line for my -20 dBm measurement.

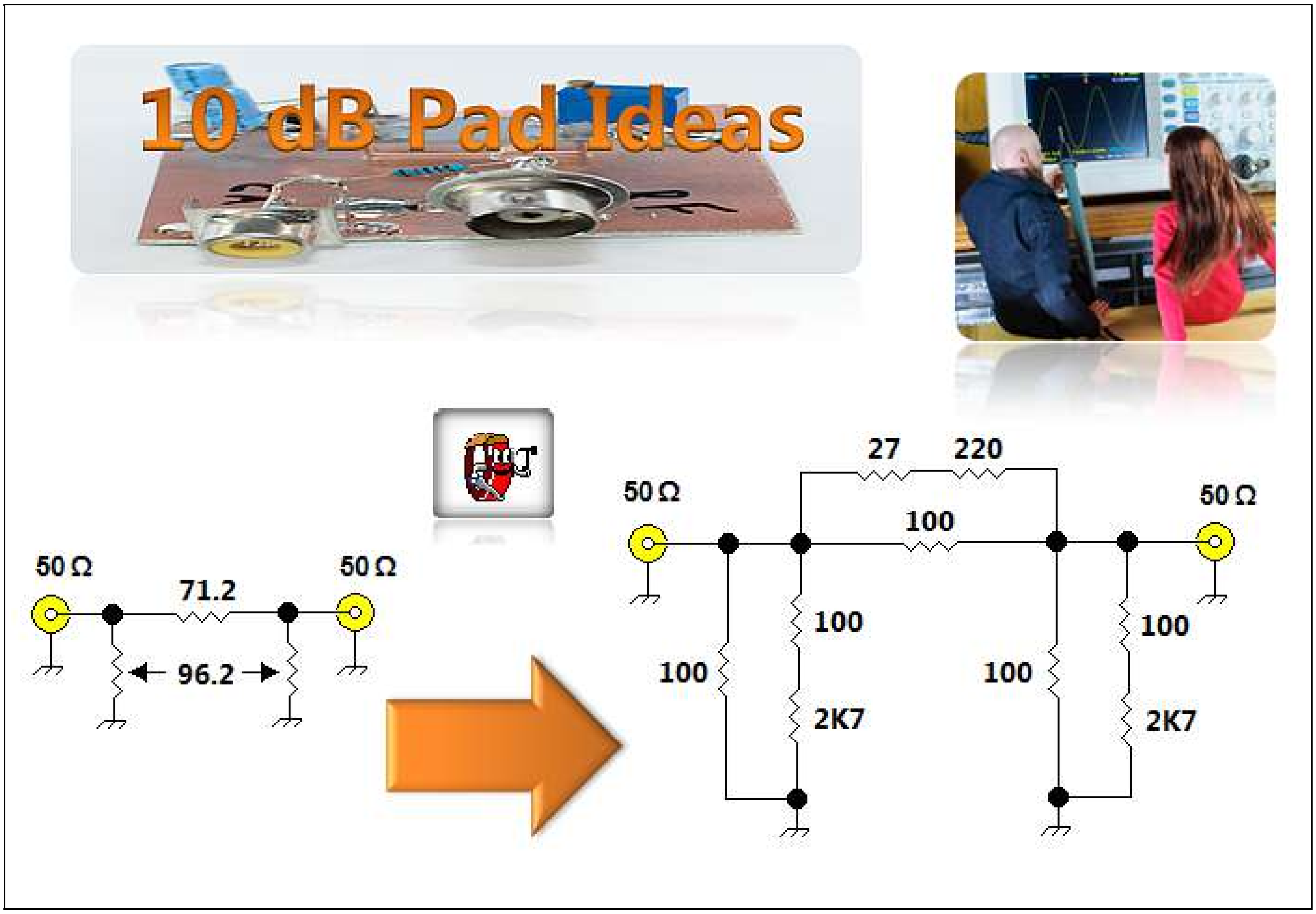
10 dB Pad Ideas
50 Ω
71.2
50 Ω
96.2
27
220
50 Ω
100
50 Ω
100
100
100
100
2K7
2K7

Above — Craft accurate attenuator pads with parallel + series resistors and some creative energy. Lacking the proper 1% parts, I hand selected some 5% resistors among the values shown to the right and built a pad that *precisely* gave a 10 dB power drop at 50 Ω.

Now let's check the calibration of this little sinusoidal RF generator...

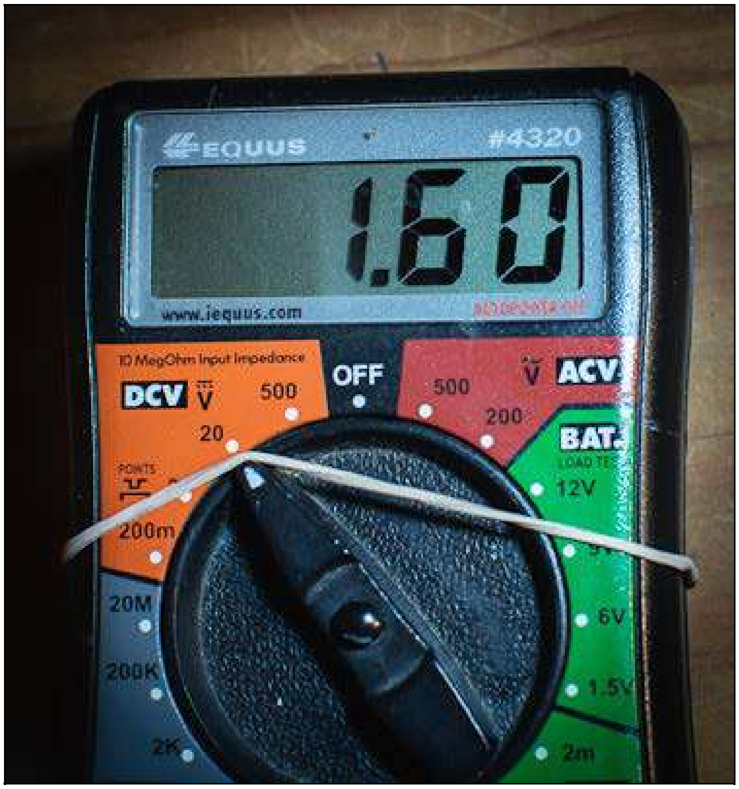

Above — When connected to the Barebones PM, I calibrated the output power of the sine wave generator by tweaking the trimmer pot to give an output of 1.60v; the same Port B voltage yielded by the CMOS RF calibrator.

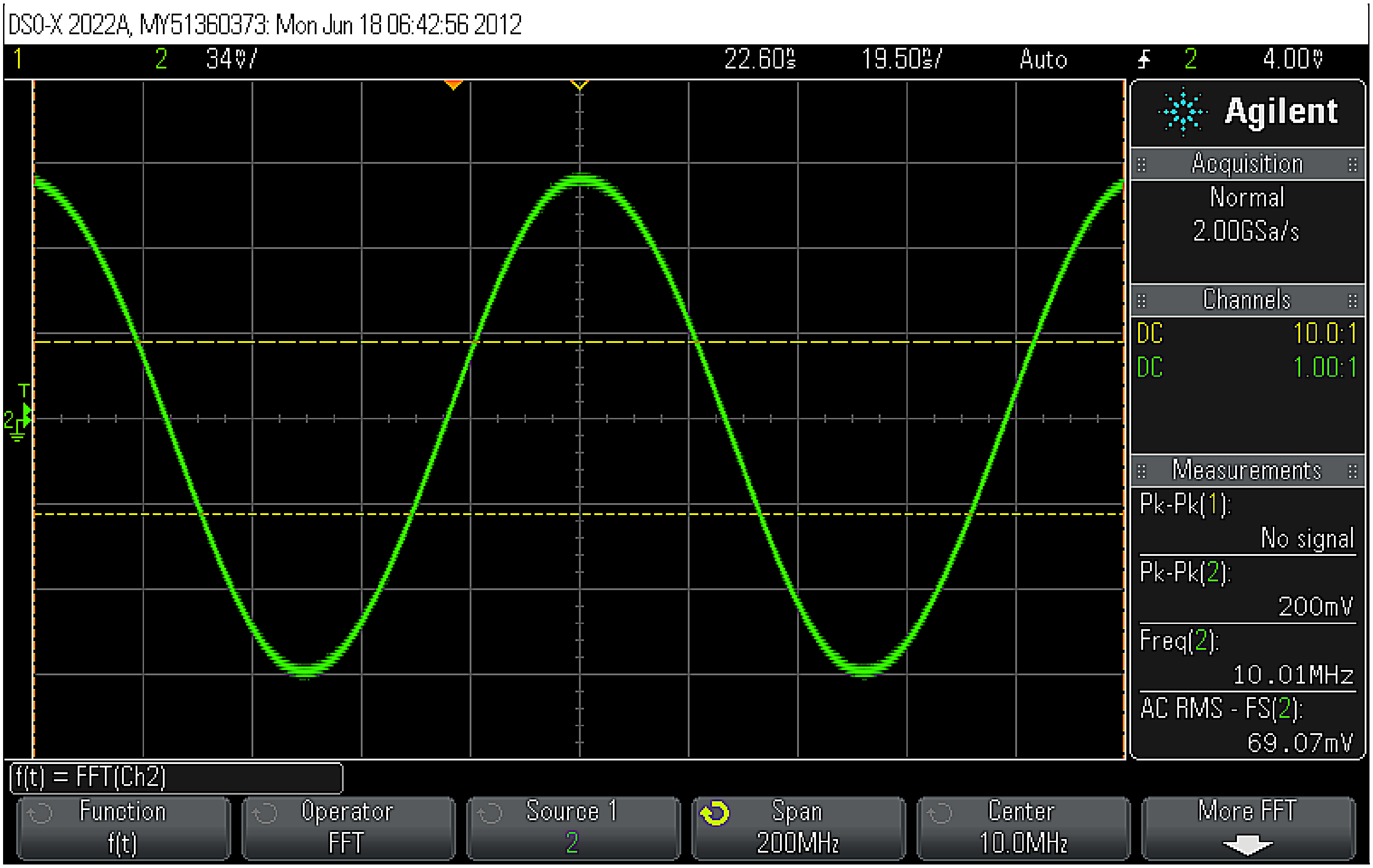

DSO-X 2022A, MY51360373: Mon Jun 18 06:42:56 2012

1
2 34mV/
22.60ns
19.50ns/
Auto
2 4.00mV
Agilent
Acquisition
Normal
2.00GSa/s
Channels
DC 10.0:1
DC 1.00:1
Measurements

Pk-Pk(1): No signal

Pk-Pk(2): 200mV

Freq(2): 10.01MHz

AC RMS - FS(2): 69.07mV

f(t) = FFT(Ch2)

Function f(t)
Operator FFT
Source 1 2
Span 200MHz
Center 10.0MHz
More FFT

Above — The 50 Ω terminated 'scope verification of my calibrated 10 MHz sine wave signal source: 200 mV pk-pk = -10 dBm. Wow, thanks for this Bob!

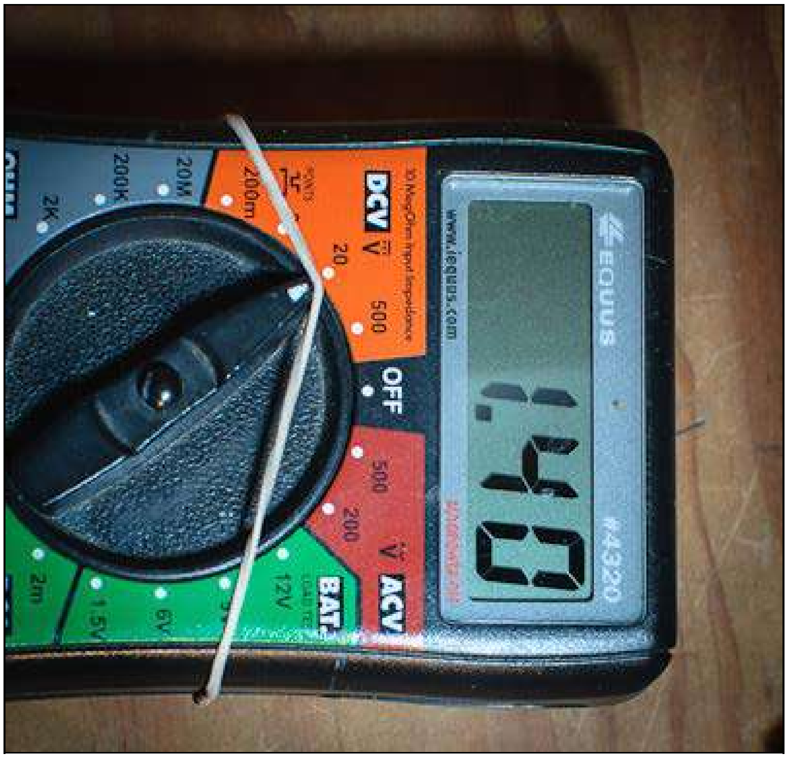

EQUUS
#4320
www.iequus.com
DCV
OFF
ACV
BAT
200m
20
500
200
2K
200K
20M

Above — The Port B voltage when a 10 dB attenuator pad was connected to the calibrated -10 dBm sine wave signal source: Measured power = -20 dBm. Notice the 20 mV/dB power drop — right on specification.

Refer to the section titled **<How to Measure Power at Port B>** Recall Wes' calculation needed 2 points to set the log linearity. I chose -10 dBm and -20 dBm instead of 0 dBm and -10 dBm so the CMOS RF calibrator could be used to calibrate any sine wave signal generators on hand.

The -10 dBm calibration reference power serves as 1 of the calibration points in Applet L while we derive the other by adding a 10 dB attenuator pad to a *sine wave* signal source output port. ***Don't connect an attenuator pad to the CMOS square wave calibrator —*** error in the AD8307 arises. Why? Analog Devices mysteriously changed the AD8307 at some point after 2004 and altered its crest factor — this disallows us performing log linearity calibration with a square wave.

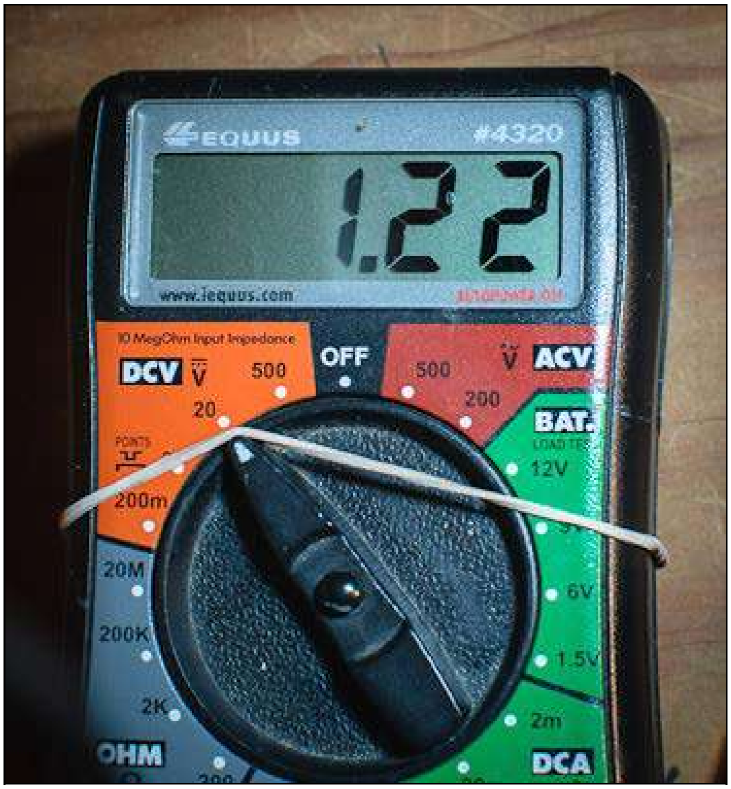

Above — A 20 dB pad connected to the -10 dBm sine wave signal source yielded 1.22v at Port B in the Barebones PM. See the power calculation with Applet L below:

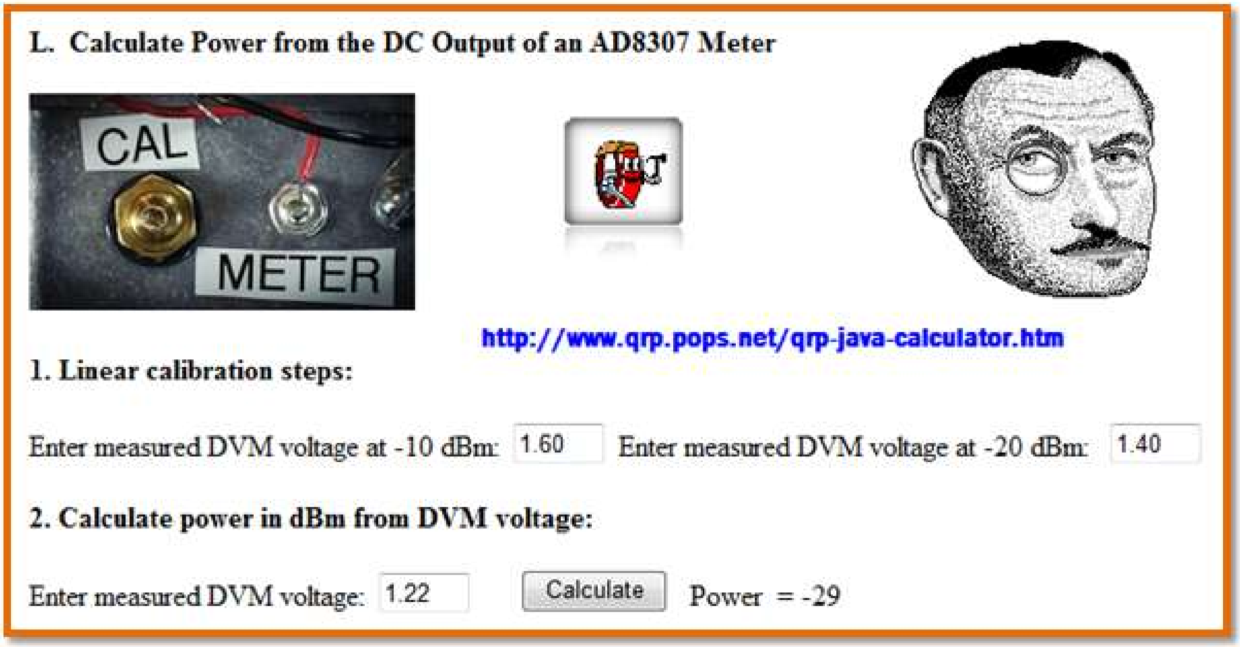

L. Calculate Power from the DC Output of an AD8307 Meter

http://www.qrp.pops.net/qrp-java-calculator.htm

1. Linear calibration steps:

Enter measured DVM voltage at -10 dBm: 1.60 Enter measured DVM voltage at -20 dBm: 1.40

2. Calculate power in dBm from DVM voltage:

Enter measured DVM voltage: 1.22 Calculate Power = -29

Above — The calculation of the output power with a 20 dB attenuator pad on the sine wave signal source output. We're off by 1 dB since it should have calculated a power of -30 dBm. Could this be error caused by my attenuator circuit, or non-linearity by the AD8307 PM, or a bit of both? We have to live with such problems. Still, the Barebones power meter seems quite impressive for a simple circuit that I scratch built and calibrated in about 25 minutes.

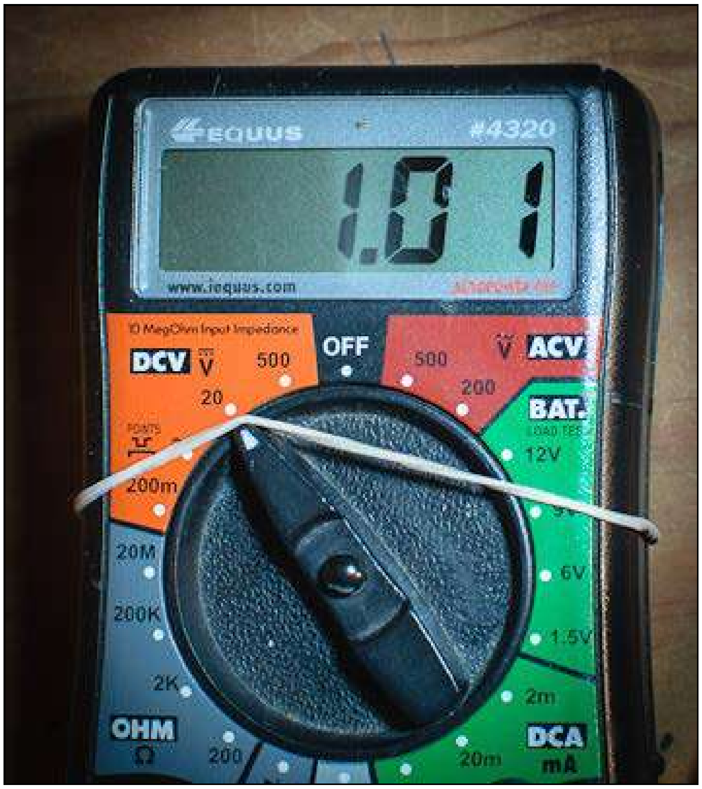

Above — The Barebones PM DC output voltage with a 30 dB attenuator pad on the output of my -10 dBm, 10 MHz sine wave signal generator: Measured power = -39.5 dBm. Again; a very good — but not perfect power meter.

A Barebones power meter allows builders with modest equipment to measure power, gain/loss, and with a return loss bridge, return loss at HF. I wish I owned this little gem back when I started out.

Richard, "Dick", N4HAY, posted some great notes on his blog. I recommend following his blog since he really digs deep and likes math — thanks Dick!

Best!

---

3. A Basic RF Workbench

Since January 2012, a handful of readers asked what I consider a good basic *RF* work bench. Again, I'm just an amateur hobbyist, so my opinion might show my ignorance.

A stand-alone 50 MHz oscilloscope with at least one 10X probe. *More bandwidth if you plan to work above 50 MHz*

If no oscillocope, an AD8307-based power meter *The modern version of the diode RF probe*.

3-4 50 Ω coaxial cables with BNC connectors; a 50 Ω scope feed-through terminator, *1-2* 50 Ω BNC port terminators, a though-connector and some BNC connectors to solder onto temporary circuit boards or mount in a chassis.

A homebrew return loss bridge.

3, 6, 10 and 20 dB BNC connector equipped attenuator pads, or a step attenuator.

Signal generator(s) that cover most of HF; +/- VHF signal generators described in the next section.

AADE L/C Meter IIB. Click

12 volt regulated power supply good for at least 1 amp.

Digital multimeter. *I use 2 and keep 1 set up for current measurement only.*

Frequency counter: homebrew or commercial. *I ran a 40 year old, ovenized, accurate HP counter until 2012.*

With these devices, as possible, you can work in a modular, 50 Ω environment and measure gain or loss in dB, return loss in dB and absolute power in dBm. Starting small and expanding your bench around 50 Ω input and output impedance devices will provide a lifetime of challenge and excitement in RF design.

Later, the big toys can follow: spectrum analyzers, VNAs, commercial signal generators and other lab quality stuff.

Equipping an RF bench presents quite a financial burden. I started small and slowly added pieces over time. Many pieces such as my L/C Meter IIB were gifts for holidays or my birthday. Other pieces were old, inexpensive equipment that I restored and calibrated.

3dB ATTE
FREQUENC
Mini-Circuits
15542
HAT-3

Above — RF tools of the trade. We're RF experimenters! As scratch homebrew builders, gear like BNC, SMA and through-connectors, 50 Ω terminators and inline attenuators lie scattered on our benches; our fodder. Alternate photo.

---

4. VHF Signal Generators

Having only started at VHF in November 2011, my knowledge suffers, however, a search for accessible, affordable, good quality VHF signal generators disheartened me. Ten year old or newer signal generators covering the VHF band work up to several GHz and cost a small fortune.

Lamenting old timers often recommend the vaccum tube HP-608 series that covered ~10-480 MHz. These heavy, glowing beasts sometimes come up in estate sales or on eBay for $200-400.

Then, too, the HP8640 series seems attractive, however, they are full of decaying parts. Ken Kuhn and others restore old HP gear as a hobby and this direction certainly gives us a valid option.

I've investigated 1 or 2 new, low-cost, commercial signal generators that work into VHF, but they failed to excite me; especially after I downloaded the schematics and sat in disbelief over their poor design.

Some minimum commercial signal requirements might include stable, linear tuning, a metal chassis, 50 Ω output with a return loss greater than 20 dB and low harmonic distortion at all frequencies. Like the rest of our lives, our budget usually determines what we buy.

I decided to build my own VHF signal generators and document them on this web site. I've learned that home building signal generators between 50 and 200 MHz requires skill and care, but can be done.

What about digital clocks?

At VHF, DDS spurs get extreme as you get closer to the maximum clock frequency .

The Si570 looks intriguing, however, still requires an MCU + components, I haven't read any lab quality evaluations of the Si570 as part of a engineer-grade VHF-UHF signal generator and if you know better, please email me.

---

5. L - C Meters

If you search for opinions about which L/C meter to get, you'll find an abundance of super write-ups including those that cover measuring with Kelvin probes, SMD tweezers; or statements suggesting that if you really need maximum accuracy, purchase a VNA. I encourage you to research this yourself and find the best L/C meter for *your* bench. Here are my 2 cents worth of opinion and please remember — I'm often wrong.

I use an AADE L/C IIB meter to test inductors and capacitors for HF and even some VHF work. Yes, the AADE L/C IIB doesn't measure large-value electrolytic caps and so forth, however, considering the cost versus performance — it's accurate enough for the popcorn RF workbench.

The AADE L/C meter uses the method described by William Carver, W7AAZ in an article called The LC Tester published in Communications Quarterly, Winter 2003 . EMRFD page 7.12 briefly examines Bill's circuit and shows his original oscillator along with an extended range Colpitts oscillator designed by Wes. For brevity sake, I'll just discuss inductance measurement with Carver-style meters. With care, an inductance resolution of 5 - 20 nH might be realized with such a device.

We normally don't consider that our inductor is actually a network with L, a parallel C and losses that might be modelled as R in series with the L, or a R in parallel with the C depending on our model. The inductor also exhibits a self-resonant frequency and for our design purposes, we usually *ignore* all these details and just consider it a "pure L". My L/C meter's oscillator runs from a few tens to a few hundred KHz and generally lies below the self-resonant frequency of the inductors I measure with it. I've learned by sweeping/analyzing my completed filters, that as long as you avoid the coil's SRF, *the low frequency* Carver-style meter proves a stalwart inductance meter for most HF and some VHF applications.

Often, we popcorn builders want to make a filter, an oscillator, or a pi, or L match and we apply software or tables to calculate the L and C values needed to resonate our filter tank(s) or matching networks, or to synthesize a low-pass filter. These math-driven programs/tables assume the *pure* "pixie dust" L described eariler — disregarding the stray C and R. So getting all worked up about whether our coil is 4.50 or 4.59 microHenries seems moot.

Further, we man-handle our inductors [changing the inductance somewhat] into a breadboard laden with 5% (or greater tolerance) capacitors, copper board pads/paths that exhibit C, active devices and so forth. Then, too, we connect these resonators, or filters to other blocks with sometimes reactive ports, plus or minus shielding. Despite all these variables, miraculously, we make the filter with our "measured" L work!

As we move up in frequency, the effects of stray L effect magnify and at some point our filter networks may behave poorly.

For band-pass filters with Carver-style device measured inductors, we need only adjust each trimmer capacitor to get the highest possible peak-peak voltage, or RF power with our filter between a signal generator and a 50 Ω terminated scope, power meter *respectively*, or whatever. After sweeping these peaked filters, rarely do I need to compress/expand, or add or remove windings to tweak the L get the desired filter response when the filter input and output ports are well matched.

In the case of single frequency matching networks like the L-match, we might need to tweak up or down the L to derive strong port matching. In all cases, our software and the L/C meter can get us into the ballpark, but in-situ bench measurement with other instruments will garner the home run.

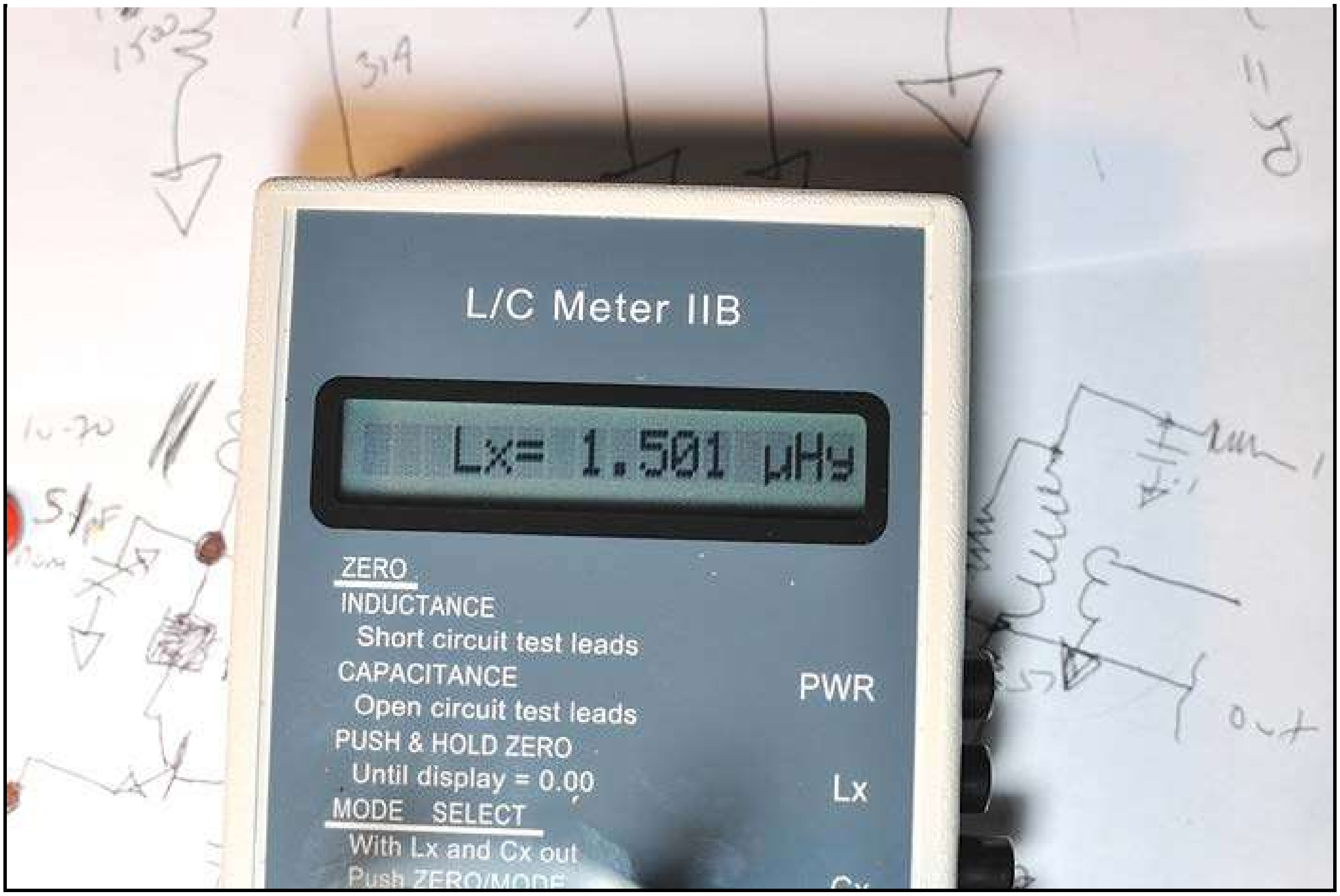
L/C Meter IIB
Lx= 1.501 µHy
ZERO
INDUCTANCE
Short circuit test leads
CAPACITANCE
Open circuit test leads
PUSH & HOLD ZERO
Until display = 0.00
MODE SELECT
With Lx and Cx out
PWR
Lx

---

6. Bob, K3NHI — RF Power Meter Follow-on

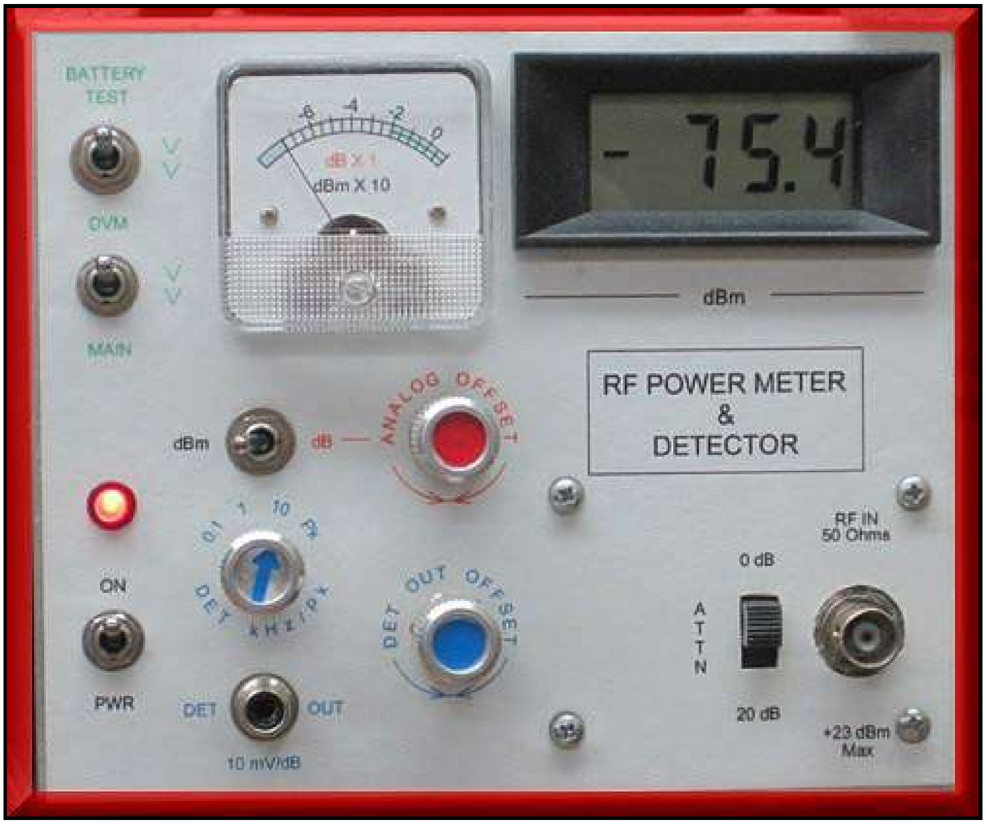
BATTERY TEST
DVM
MAIN
-6
-4
-2
0
dB X 1
dBm X 10
- 75.4
dBm
RF POWER METER & DETECTOR
dBm
dB
ANALOG OFFSET
.01
1
10
PK
DET kHz / pk
DET OUT OFFSET
ON
PWR
DET
OUT
10 mV/dB
RF IN 50 Ohms
0 dB
ATTN
20 dB
+23 dBm Max

Bob built a follow-on detector and power meter to his 2002 QEX power meter *[Reference 3 in Section 7].* Bob gave me the green light to share his creation on my site. Thanks awfully Bob!

Click for the schematic in pdf format. Per typical K3NHI fashion, it's laden with trimmer pots allowing precise calibration — Bob's stuff contains loads of tweaks and wiggles! The new PM offers more flexibility + features including bigger battery supplies that won't quickly die when you fail to turn it off. He also included a means to measure battery voltage on the analog meter. Click for a photo showing the batteries. Low noise, high-speed, rail to rail, CMOS op-amps for IC1 and IC3, allow you to capture PEP during sideband transmitter measurement with appropriate attenuation.

The TS922 op-amp might be hard to find in DIP since that package went obsolete, however, it's available in SMD. If you can't get any TS922, likely other modern, high speed, rail to rail, CMOS op-amps will work fine — consult datasheets to ensure you meet or beat the TS922's performance for IC1 and IC3. Some of the latest design op-amps offer truly sublime specifications and evoke joy in our breadboards.

Bob critically *isolated* the entire RF sub assembly **outlined in green on the schematic** with metal shielding and feed through capacitors for the B+ and DC output. This helps ensure stable and accurate AC voltage measurement. Click or click or click for more of Bob's photographs.

---

7. References

[1] Roger Hayward, KA7EXM — A PIC-based HF/VHF Power Meter, QEX May/June 2005

[2] Wes Hayward, W7ZOI, Bob Larkin, W7PUA, Simple RF-Power Measurement, QST, June 2001

[3] Bob Kopski, K3NHI — An Advanced VHF Wattmeter, QEX, May/June 2002.

[4] Bob Kopski, K3NHI — A Simple Enhancement for the Advanced VHF Wattmeter, QEX, Sept/Oct 2003.

[5] Bob Kopski, K3NHI — A Simple RF Power Calibrator, QEX, Jan/Feb 2004 **+** Tech Notes article, QEX, May/June 2010.

RF — Test and Measurement

VHF FM

Yes - I could purchase a FM radio-in-an IC for $2.00 and be done in 35 minutes, but what would I learn?

Repository for FM superhet receiver experiments conducted from 2012 to 2014.

1. 10.7 MHz IF Filter Experiments
2. A Basic Colpitts VCO
3. DC-DC Converter for VCOs
4. Supplemental Page #1 - it's time to make some receivers
5. Miscellaneous Photos or Figures

---

1. 10.7 MHz IF Filter Experiments

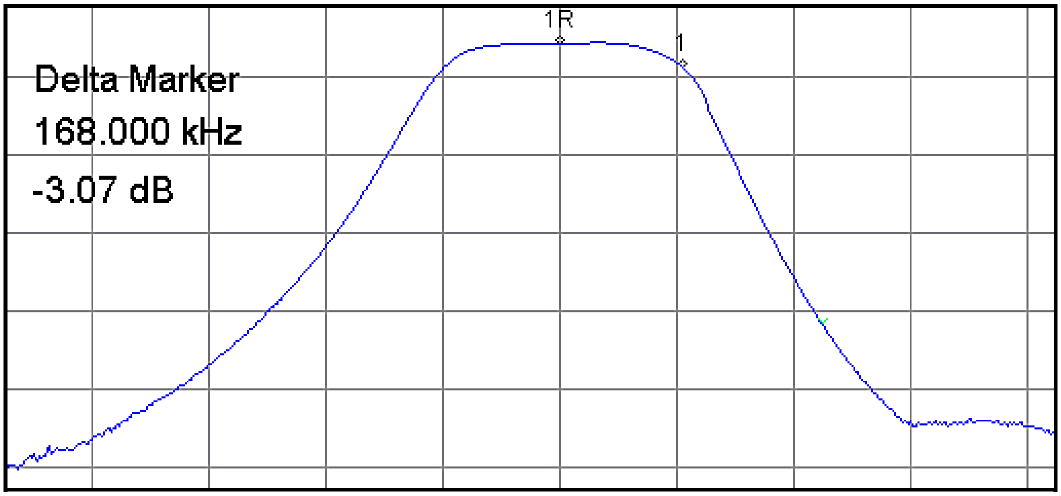

As a FM receiver design newbie, I read about and experimented with some 10.7 MHz IF filters to learn common practices, what's available and which measurements might help me to reach my goals. Over time, I've collected a variety of crystal and ceramic filters for hopeful future work. Click for 2 exotic examples. IF filters might be purchased at Ham festivals, surplus electronic parts stored and/or online. Prior to paying for a filter, I've found it useful to politely request a sweep of the filter, or, better yet, perform this task myself. To sweep a filter in your lab, you'll need a tool such as a spectrum analyzer with tracking generator, a VNA, or some other analog/digital sweep system.

I'll homebrew some crystal filters for narrow band FM in future installments — I ordered some 20 MHz xtals.

Sometimes a filter in your junkbox will state the IF and perhaps the 3 or 6 dB bandwidth, but not the input/output port termination impedance. How do we determine this impedance? I've learned we can figure this out by testing differerent termination resistors with this simple test jig:

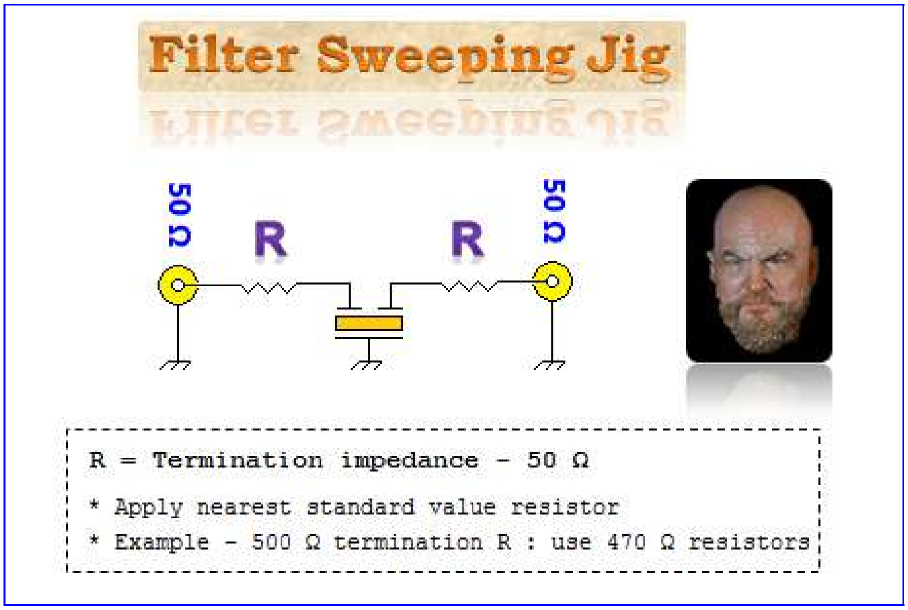

R = Termination impedance - 50 Ω

* Apply nearest standard value resistor

* Example - 500 Ω termination R : use 470 Ω resistors

Above — A simple crystal or ceramic IF filter sweeping jig. Since the series resistors attenuate the signal, losses occur; but the shape should look clean with minimal ripple. Normally, we builders will also place (or switch in) 50 Ω attenuator pads on both the signal source and detector within our sweep system to buffer impedance mismatch. Comparisons of this simple jig with more precise and complicated matching methods suggest that for many filter sweeps, it might work fine.

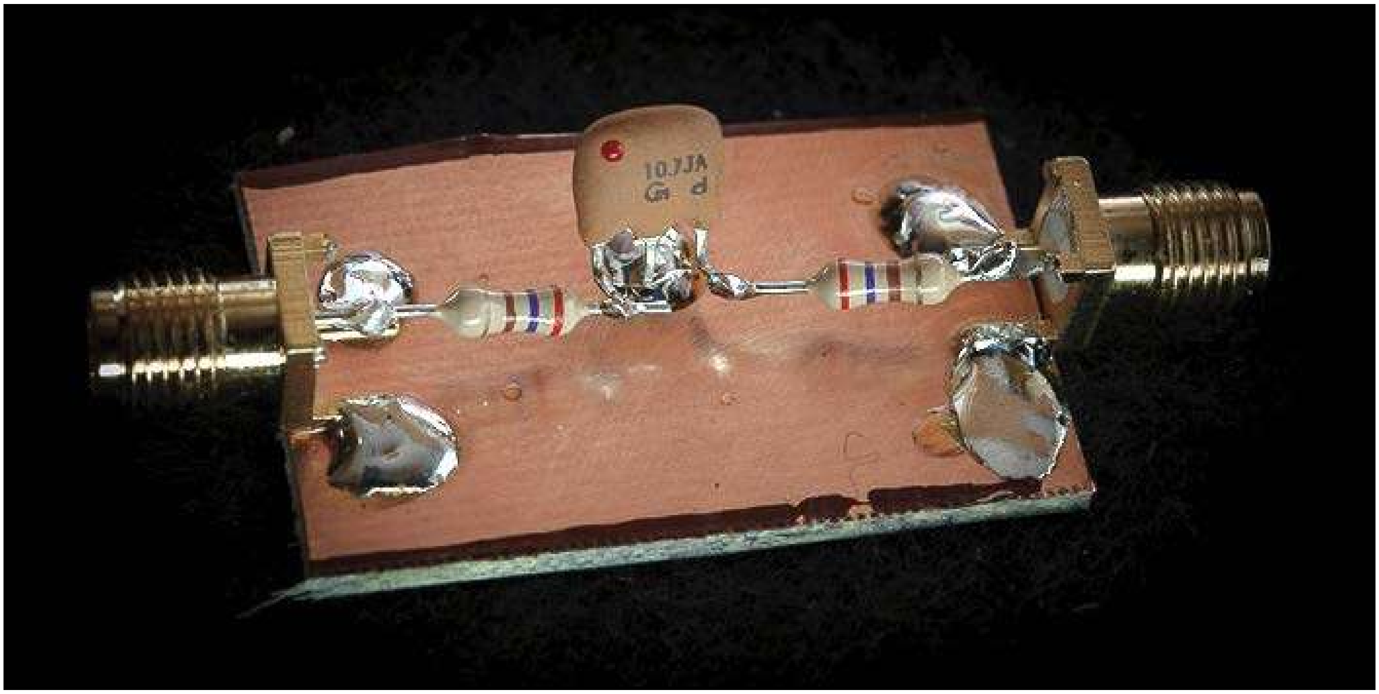

Above — My test jig with a Murata ceramic filter soldered in-situ. Keep the resistors *close* to the board. I've pretty much moved to SMA connectors in my lab: they're cheaper than BNC, plus we can buy a wide variety of quality 50 Ω patch cables donning various connectors for low cost. For example, a 30 cm cable with a male BNC and SMA connector on either end.

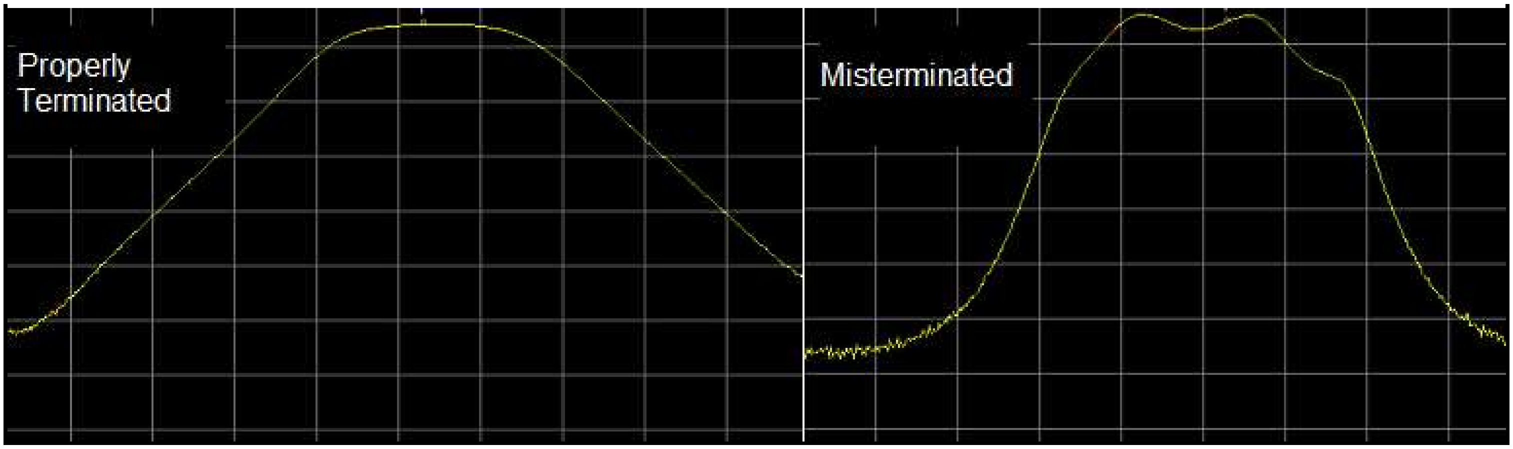

Above — A poor termination may result in improper bandwidth and ripple — easy to spot in this trace. Click for a trace from a 'gone bad' instrumentation crystal filter: 10.7 MHz @ 30 KHz with 2200 Ω Z in/out. Not really usable with ~ 10 dB ripple.

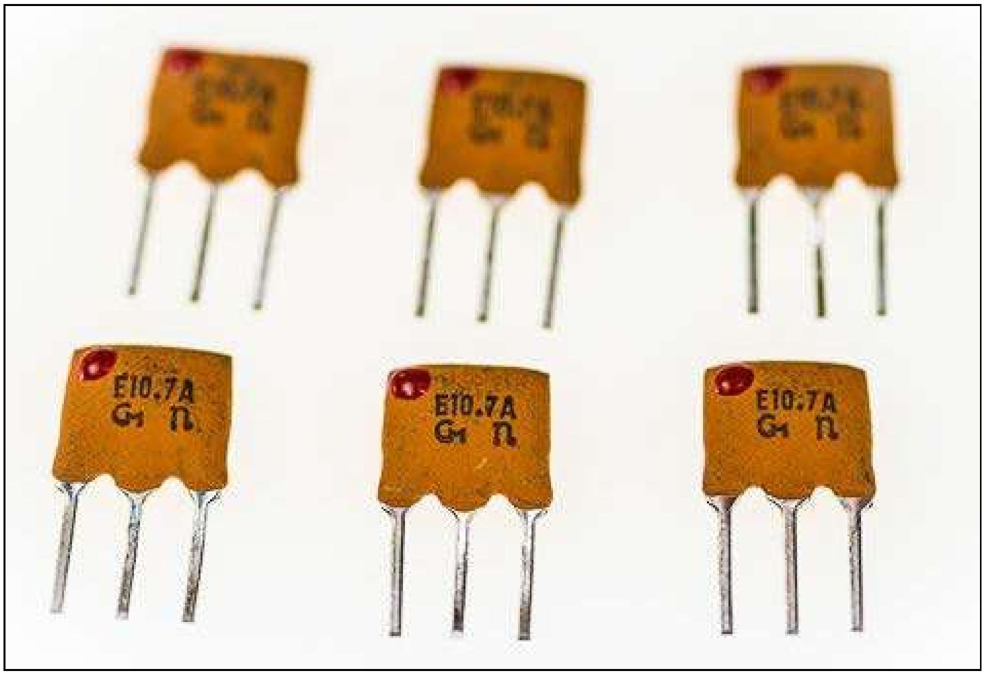

Above — Older 280 KHz Murata 10.7 MHz IF filters purchased long ago. Low cost = their main attraction, although they too will suffer total obsolescence and a price increase.

I bought some newer, lower insertion loss ceramic filters in the following bandwidths: 280 KHz, 230 KHz, 150 KHz, 25 KHz and 20 KHz. Check their datasheet — most Murata ceramic filters require a 330 Ω termination (preferably resistive) and I keep a filter sweeper jig with 270 Ω resistors as a regular bench tool. 280 KHz was a popular WBFM filter bandwidth in many older high-end FM receivers including my 1980's T-85 Yamaha receiver; my benchmark FM receiver.

Many of us hopeful FM builders, smitten by modern digital gear, fail to recognize the fantastic design achievements made by FM receiver engineers back in the day. All those air-variable, ganged band-pass preamp stages, low noise amplifers and often incredibly complicated and great sounding FM multiplex circuits just blow me away. Perhaps I'm a hopeless analog nostalgic? My T-85 sports 5 ceramic filters [280 KHz and 230 KHz B/W Muratas] and the narrow filters are listener switchable for narrow band Dx.

Ways to Match These Filters with Amplifers

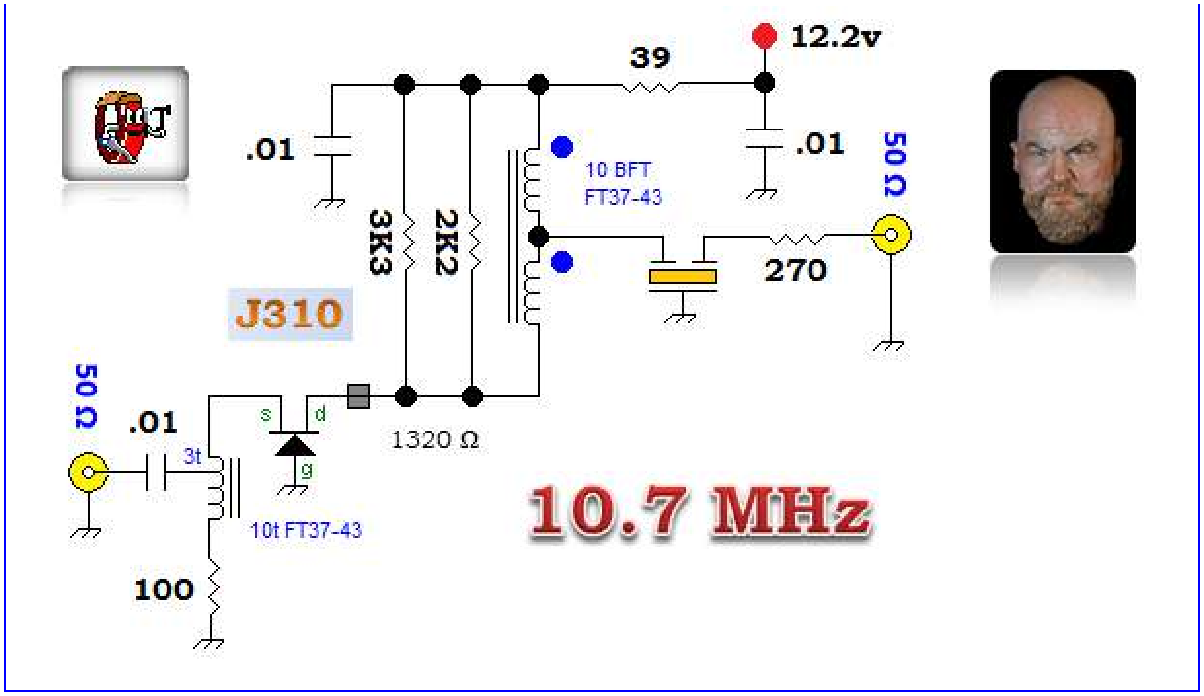
39
12.2v
.01
.01
50 Ω
10 BFT
FT37-43
3K3
2K2
270
J310
50 Ω
.01
3t
s
d
g
1320 Ω
10t FT37-43
100
10.7 MHz

Above — A common gate JFET amplifer drives a 330 Ω ceramic filter. I placed 2 resistors in parallel to get the needed shunt R of 1320; my 2 resistors measured 1316 Ω. The bifilar transmission line transformer provides the 330 Ω Z to drive the filter. Details of the JFET amp come in a later schematic, but the input return loss at 10.7 MHz = 23 dB. I swept this circuit and it looked similar to the tracing with the same ceramic filter in my 270 Ω filter sweeping jig. This particular filter exhibited 6.4 dB of insertion loss.

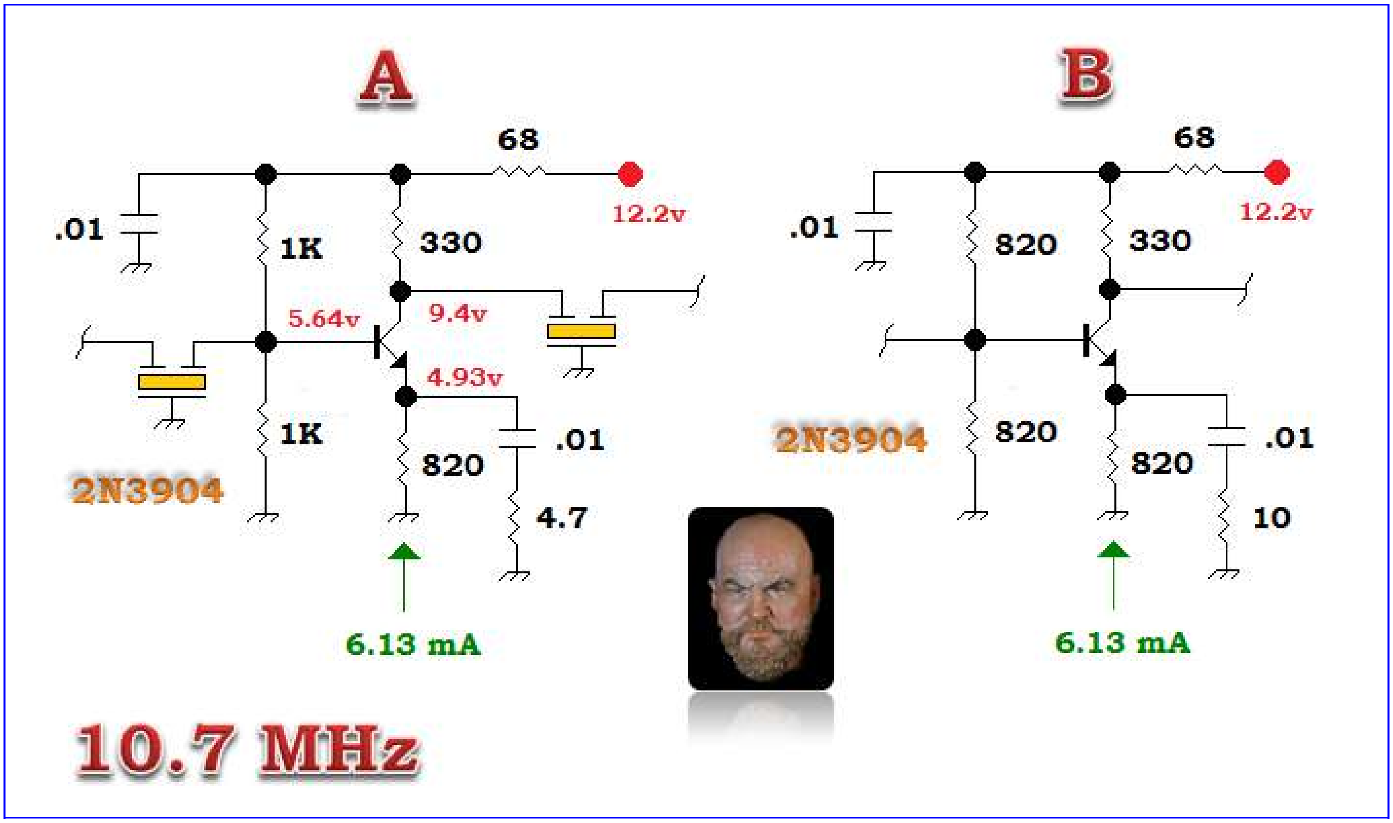
A
68
.01
1K
330
12.2v
5.64v
9.4v
4.93v
1K
820
.01
2N3904
4.7
6.13 mA
10.7 MHz
B
68
.01
820
330
12.2v
2N3904
820
820
.01
10
6.13 mA

Above — 2 versions of a BJT amp with a 330 Ω input and output impedance. If you read schematics of good FM receivers, often the designers drive the filter with a 330 Ω collector resistor. Click for an example. By keeping the bias and degeneration resistors low and the current moderate, an amplifer with 330 Ω input Z is easy to design [although the input Z will vary with Beta].

I felt surprised that version **A** exhibited a voltage gain of 11.7 despite those low bias and collector resistors. You can stick a filter on either side as shown. Murata recommends a buffer amp between cascaded ceramic filters and you'll see this often in FM receiver schematics from the 1980s or so. Resistors provide wideband termination. Version B is the same amp with a little more degeneration to lower the gain and serves as a design example. I've got the procedure documented here under '*Calculating the input resistance of a common emitter stage'.*

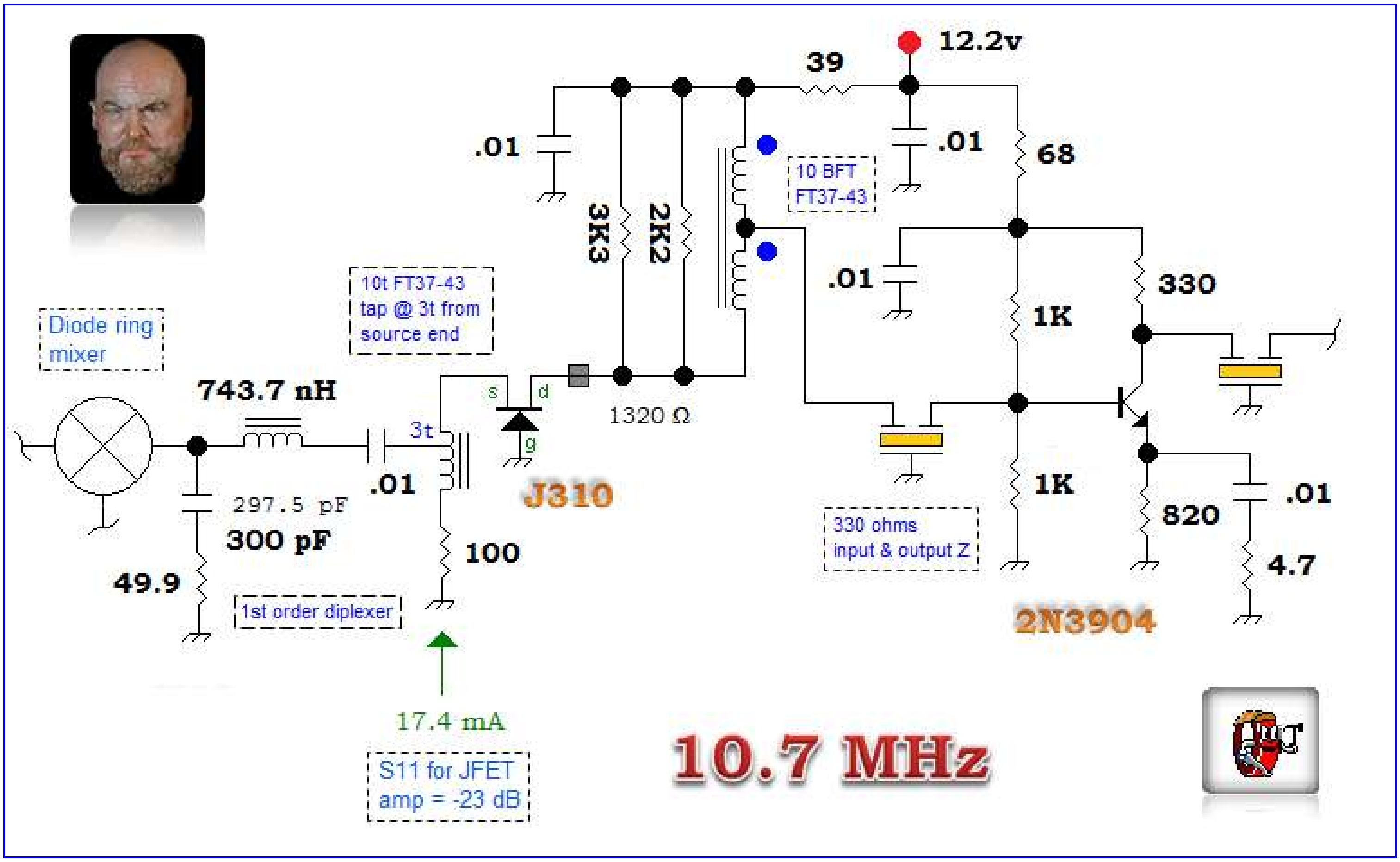
12.2v
39
.01
.01
68
10 BFT FT37-43
3K3
2K2
.01
330
1K
10t FT37-43 tap @ 3t from source end
Diode ring mixer
743.7 nH
3t
s
d
g
1320 Ω
.01
J310
297.5 pF
300 pF
49.9
100
1st order diplexer
330 ohms input & output Z
1K
820
.01
4.7
2N3904
17.4 mA
S11 for JFET amp = -23 dB
10.7 MHz

Above — An IF block using the designs shown earlier. I terminated this stage with a 270 Ω resistor and of course removed the mixer and diplexer. 2 sweeps lie below. I'm tempted to tune the JFET drain and couple the transformer with a few links as needed to get a 4:1 impedance ratio. Anyhow — food for thought.

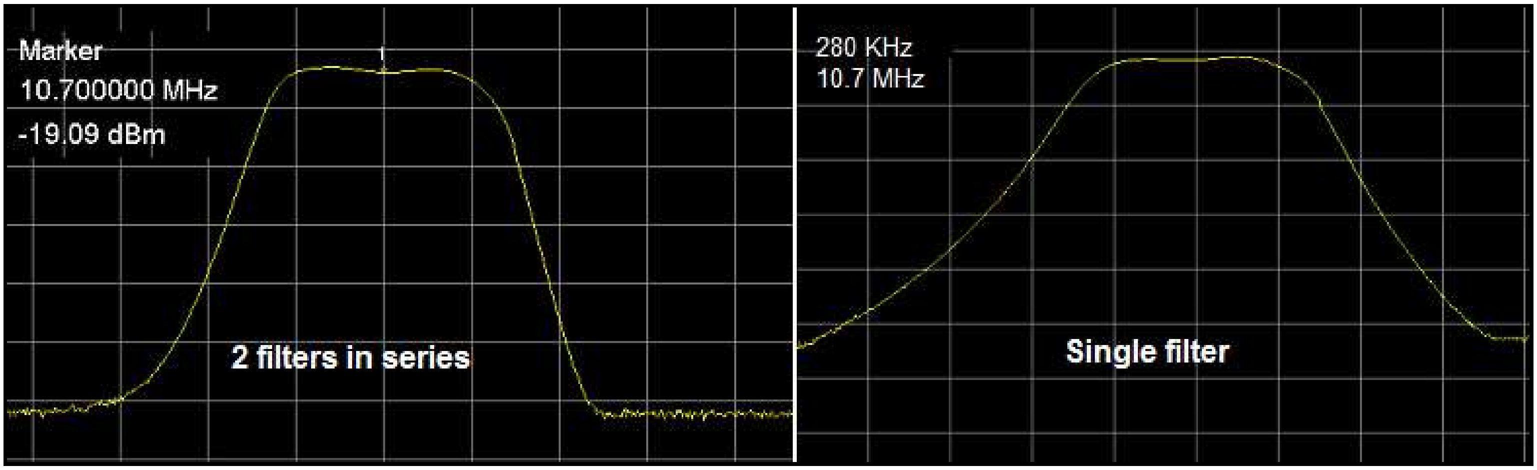

Above — A sweep of the IF block shown above left sans the mixer + diplexer (*two* 280 KHz ceramic filters). On the right lies the trace of the common gate amp driving a *single* 280 KHz filter with its output terminated with a 270 resistor. The advantages of 2 cascaded filters seems apparent, although the slight downward dimple at the center frequency might represent some capacitive loading at the output of the common gate amplifier.

I built a number of other amplifers and swept all of them Click or click for 2 early examples that use active devices instead of series matching resistors on the output. In these circuits, R Term was changed and then the circuit was swept. The tracings looked good.

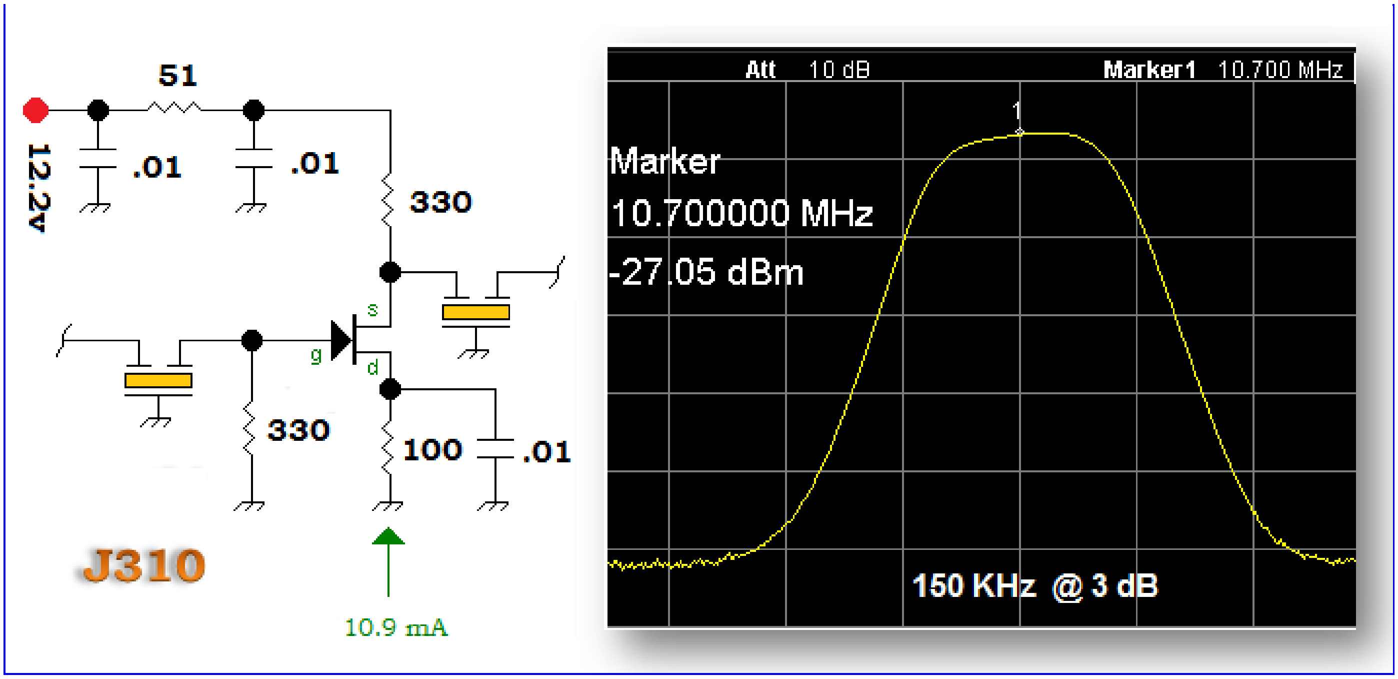

Above — I built Brian, K6STI's nJFET IF amp. He used it to offset the losses associated with 2 ceramic filters. Click for Brian's fabulous website. I placed 150 KHz 3dB BW filters before and after a J310 and swept — my circuit exhibited a 2 dB net loss which seems quite reasonable.

The 330 input resistor is a load/termination on the input filter and will dissipate some energy and lower the AC input voltage to the gate compared to the usual high Z input resistors we apply in our JFET common source amps — from open circuit to full termination would incur a 6 dB voltage drop. Still, for simplicity versus performance, Brian's circuit looks hard to beat.

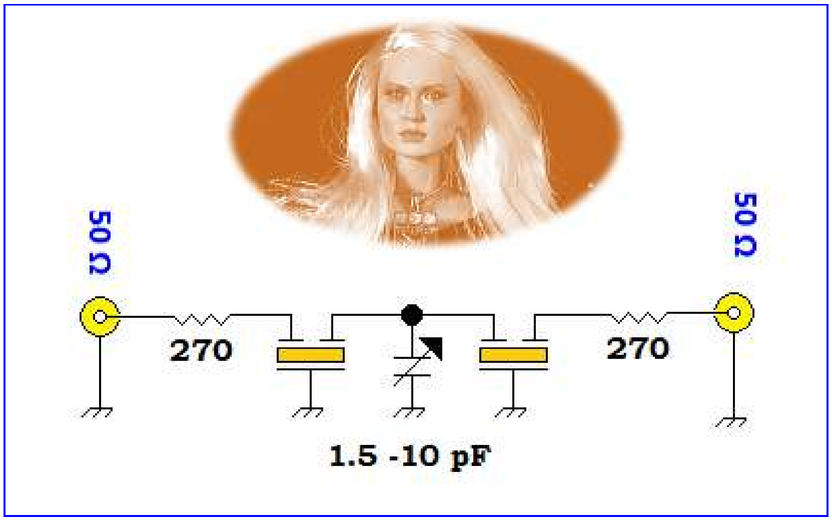

Above — 2 ceramic filters in series. I added a small trimmer between the pair in hopes to mitigate any filter skirt distress or ripple. Click for a tracing with and without the trimmer capacitor. You might experiment with the filter coupling and the filter block termination impedances to better their skirts and passband The losses of the above filter block may reach 12-14 dB.

If you don't have a sweep system, I was able to crudely test the amplifers + filters with my 10.7 MHz signal generator and a DSO.

Resistance Bridge

If you go with a BJT IF amplifer, it's possible to measure the input impedance with a bridge and tweak the emitter current and/or degeneration resistor to get very close to a 330 Ω Zin. I keep a drawer with through-hole resistors rated between 1 and 10 Ω for tweaking my emitter resistor values to change series feedback in my common emitter amps.

I first designed a simple 330 Ω bridge for measurement with my DVM. It worked, but the null lacked the depth and resolution we need. Later I improved the sensitivity by adding another coil and changing to a 'scope or SA detector, but after building EMRFD Figure 7.36, I abandoned my bridge. Figure 7.36 just blew me away. The null of a 330 resistor was only a few 10s of microvolts during calibration.

I placed a small 500 Ω pot in parallel with a 120 Ω resistor for the variable resistance. After some basic testing, I calibrated it with a 330 Ω resistor; adjusting the pot for the deepest null and just left it there for testing my 330 Ω IF amps @ 10.7 MHz.

I plan to make Figure 7.36 for VHF and maybe UHF with chip caps plus a small screwdriver adjustable trimmer pot [to get the lowest possible L] calibrate it and make it a part of my test bench arsenal. After getting a null, we measure the pot's resistance with an ohm meter to learn the impedance at the ? port

Considering that our predecessors measured just about everything RF with a bridge, this little circuit suddenly become relevant. A series L and C "add-on" circuit shown as Figure 7.39 may be placed in series with the ? port and device under test to deepen the null in the face of reactance. Bridge circuits form the very essence of RF measurement. Yes Bobby, we can measure impedance without a VNA.

---

2. A Basic Colpitts VCO

FM TUNER VCO
TUNE

Above — My completed Colpitts VCO. *I installed the unlabelled, left-sided pot in case a potentiometer is required for future AFC circuitry changes. It's not hooked up.*

I reviewed some 1970's FM receiver schematics to learn that before PLL-locked VCOs dominated, often Colpitts VCOs were locked onto a strong frequency with Automatic Frequency Control (AFC). Local oscillators tanks often employed a inductor plus an air variable capacitor that tuned from ~77 to 119 MHz with a varactor for AFC. All the tuning and front-end filter air variable capacitors were ganged together and I'm sure alignment took some skill.

Some VCOs tuned with varactor(s) instead of an air variable cap — this is what I wish to do. Varactor tuned VCOs usually suffer more thermal drift than air variable capacitor versions.

AFC compensates for VCO thermal drift by a seperate varactor with its control voltage line DC coupled to the FM detector through an R-C low-pass filter. Any difference between the VCO frequency and the desired FM frequency produces a proportional DC voltage. The DC control voltage changes the oscillator to the desired frequency by re-tuning the AFC varactor within this feedback loop, albeit over a limited range. AFC is unsuitable for weak signal DXing, since it may pull the receiver onto a strong adjacent signal. Many 1970's FM receivers supplied an AFC defeat switch.

I remember 1 old FM receiver in my parent's home that stayed locked on 1 frequency for years thanks to AFC.

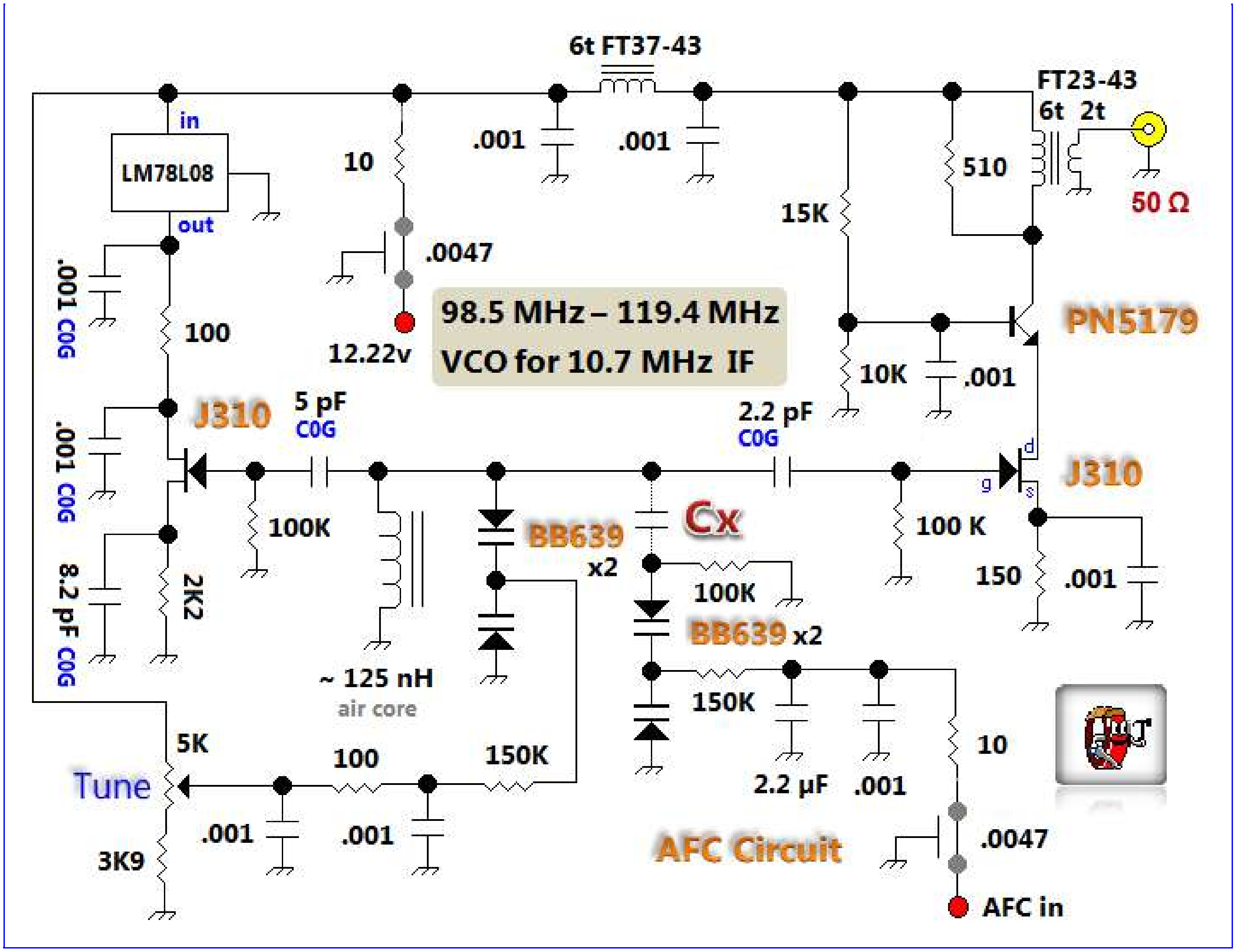
6t FT37-43
FT23-43
6t 2t
in
LM78L08
out
10
.001
.001
510
50 Ω
15K
.0047
.001 C0G
100
12.22v
98.5 MHz – 119.4 MHz
VCO for 10.7 MHz IF
PN5179
10K
.001
J310
5 pF
C0G
2.2 pF
C0G
d
g
s
J310
.001 C0G
100K
BB639
x2
Cx
100 K
150
.001
8.2 pF C0G
2K2
100K
BB639 x2
~ 125 nH
air core
150K
5K
Tune
100
150K
10
2.2 µF
.001
3K9
.001
.001
AFC Circuit
.0047
AFC in

Above — The schematic of my version of a JFET Colpitts VCO (with AFC) that lacks the standard gate to source feedback capacitor; the intrinsic capacitance from the J310 gate to source provides the feedback needed for oscillation.The 8.2 pF bypass cap was determined on the bench — too little, or too much C decreases output voltage, or snuffs out the oscillator.

I just couldn't bring myself to make a VCO with a BJT, since on my bench at least, they suffer more thermal drift than JFET-based oscillators. I built with a mixture of SMT and hole-through capacitors and resistors. The anti-parallel arranged hyperabrupt varactors were found on eBay. Click for a rear photo of the project chassis. The gold colored jack is an SMA connector.

I bench designed this VCO and it took many hours to find the correct amount of L and C for the resonator to give a low distortion, sine wave output across the ~21 MHz tuning range. This meant soldering in and removing these tank components frequently. Click for the lowest frequency output. Click for the highest.

In the example local oscillators I reviewed, the engineers made no attempt to level off the signal that normally increases in AC voltage as you increase frequency. I also ignored levelling. Presumably the designers didn't worry with leveling the oscillator output in their superhet receiver as long as the output voltage sufficiently drove the mixer into complete switching. Levelling would add cost and complexity. This isn't a lab grade RF signal generator — that's for sure.

At present, the AFC varactor pair is disconnected since I won't know how strongly to couple it with Cx until I have a working detector. Also I will need to experiment to determine the best R-C time constant for the low-pass filter; likely the 2.2 uF capacitor will need an increase in value.

With the 3K9 Ω resistor under the 5K tuning pot, I keep at least 5 VDC on the tuning varactors or the VCO would stop running as I tuned the pot towards CCW. The coil = about 3 turns of 16 gauge wire on a 5/8 inch bolt. (Despite Canada going metric in ~1975, they still sell nuts and bolts in inches at our hardware stores). The stiff wire prevents the inductor from turning into a "microphonic" spring when the VCO is bumped. Click for a photo. The nominal L = ~ 125 nH, although I bent and manipulated the coil so it sat attached to the copper clad board with no tension and then squished or expanded the turns to establish my lower band edge.

In many FM receivers, either a single or balanced dual-gate MOSFET mixer was driven by a high impedance buffer/amplifer. If I mix with a 2-gate MOSFET, I'll insert a common gate JFET amplifier on the IF strip to boost the LO output impedance and AC voltage.

The feedthrough capacitors are 0.0047 µF - they were on sale so I bought them. To prevent a parasitic high impedance when placed in parallel with my standard 0.001 µF bypass caps, I placed a series 10 Ω resistor.

I enjoyed this crazy design; trying to replicate a relic, but popular local oscillator idea from decades ago. Let's hope I did it justice. Perhaps future VHF stuff on the FM and even 2 meter band will involve an Si570 and PIC, Arduino or other microcontroller? This simple VCO will do for now. My greatest passion lies in designing and building the front end.

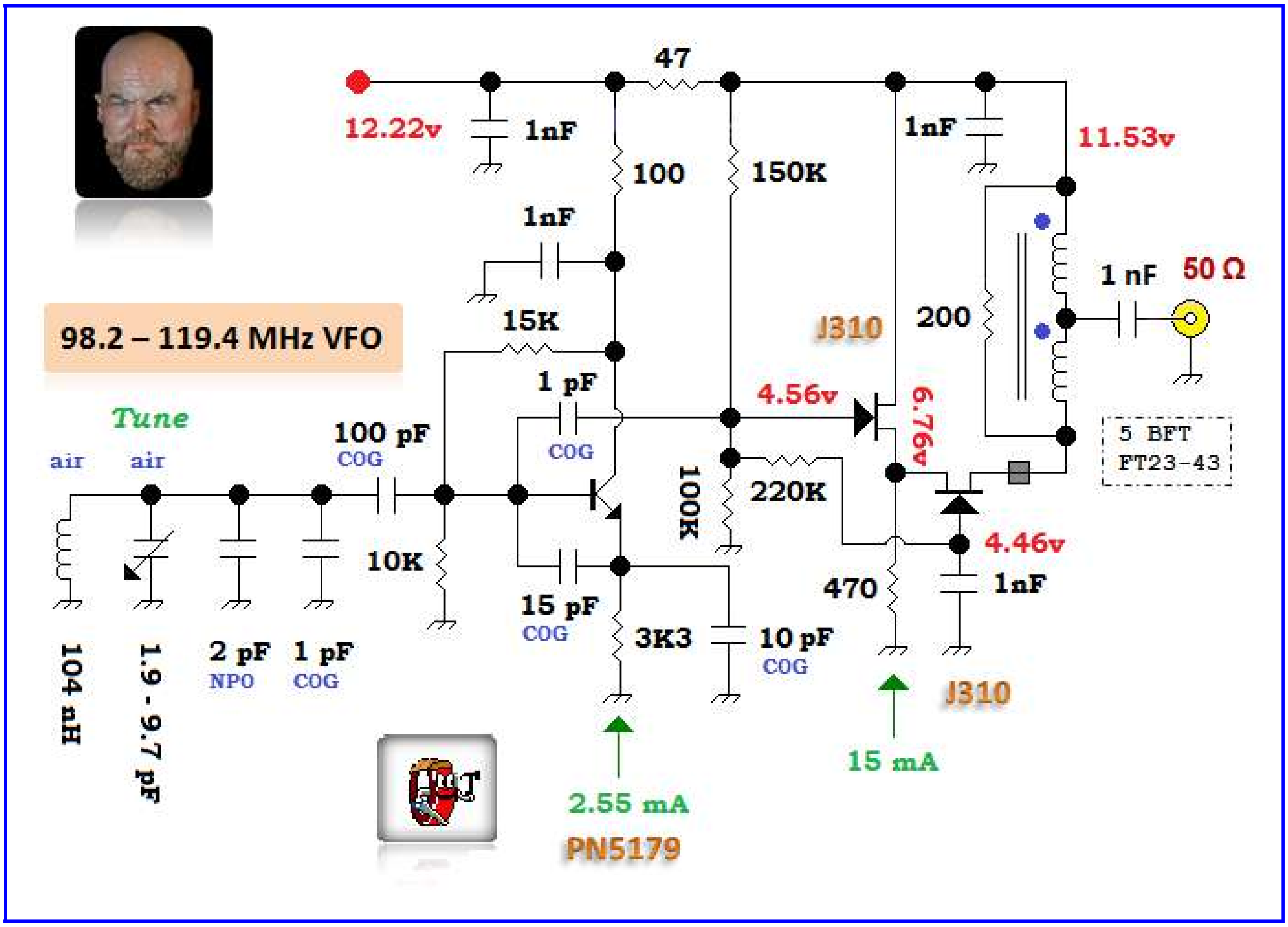
98.2 – 119.4 MHz VFO
12.22v
1nF
47
100
150K
1nF
11.53v
1nF
1 nF
50 Ω
15K
J310
200
1 pF
COG
4.56v
6.76v
Tune
air
air
100 pF
COG
100K
220K
5 BFT
FT23-43
10K
4.46v
1nF
15 pF
COG
470
3K3
10 pF
COG
104 nH
1.9 - 9.7 pF
2 pF
NPO
1 pF
COG
J310
15 mA
2.55 mA
PN5179

Above — A well-buffered, bench-module, high-side VFO I sometimes use for broadcast FM band mixing into a 10.7 MHz IF. The output at 98.5 MHz = -5.35 dBm, perfect for switching Gilbert cell mixers with a little padding or amplitude tweaking. Click for the output of the Colpitts only with a 10X probe @ 120 MHz. With care, you can see the second harmonic in the 'scope tracing — click for the SA tracing that shows the 2nd harmonic 27.5 dB down from the carrier. Click for a 'scope tracing with my MMIC bench module amplifier from VHF Veronica connected; the amp exerts some low-pass filtering that cleans up the signal somewhat.

---

3. DC-DC Converter for VCOs

Until now, I ran a maximum reverse DC voltage of ~12 volts in my varactors. For wider VCO or L-C filter tuning, builders may chose 28 volt varactors such as the BB535 or BB149A and boost the 12v supply up to 28v with a DC-DC converter. Some build inductorless converters pulsed from 555 timers, or use CMOS voltage converters like the CL7662, or Si7661 to make a doubler. As an RF constructor, I like working with coils and built the following circuit:

DC–DC CONVERTER
SET DC
28V
OUT

Above — Bench module: 28v DC to DC converter. While containing no tuning control pot, my build places the zener diode regulator control potentiometer on the front panel to allow fast-tweaking of the output voltage from ~21-30 VDC depending on the load. Click for the breadboard photo.

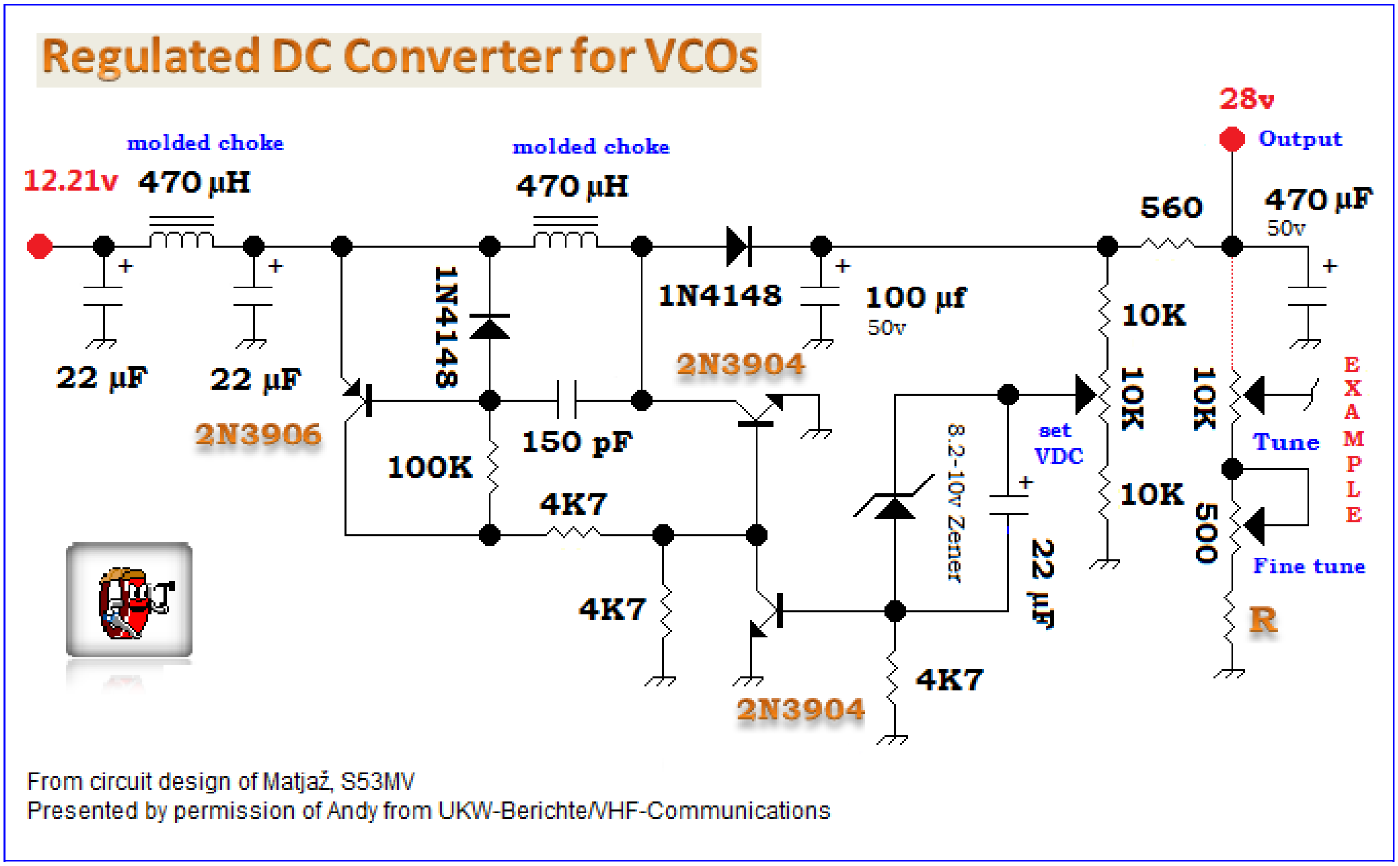
Regulated DC Converter for VCOs
12.21v
molded choke
470 µH
22 µF
22 µF
2N3906
molded choke
470 µH
1N4148
100K
150 pF
4K7
4K7
1N4148
100 µf
50v
2N3904
2N3904
8.2-10v Zener
4K7
set
VDC
22 µF
28v
Output
560
470 µF
50v
10K
10K
10K
10K
500
Tune
Fine tune
R
EXAMPLE

From circuit design of Matjaž, S53MV
Presented by permission of Andy from UKW-Berichte/VHF-Communications

Above — My regulated 28v converter for varactor tuning adapted from a design by Matjaž, S53MV. I pulled this circuit from his *amazing* 2-part article with circuits that span from 11 GHz RF to DC. *See the reference articles below*. I filtered heavily and at switch-on, my circuit draws ~ 50 mA, but then drops to ~ 11 mA after the capacitors charge. The 10K *[set VDC]* trimmer pot allows you to dial in your desired output voltage and thus this converter may work over a wide range of DC power supply voltages.

The tuning control(s) might be a single potentiometer, or even seperate pots for tuning 2 different VCOs. In the above schematic, I show 1 possible tuning scheme: a 10K coarse tuning in series with a 500 Ω fine tuning potentiometer. R keeps some minimal reverse DC on the varactor(s) and is optional. Again, my bench module DC converter omits any tuning controls — these are built into the circuit containing the varactor diode(s).

The oscillator frequency varies slightly with the set output voltage. Click for a screen capture at ~32 VAC with a 10X probe placed on the PNP emitter. In another test, with no load, I watched the coil's magnetic field collapse and ring in this cool 'scope capture. This is why I love deep memory DSOs so much.

VHF COMMUNICATIONS

TV Satellite Receive System Part 1 : Low Noise 11 Ghz Down Converter ---- issue 4/86

TV Satellite Receive System, Part 2 : Indoor Unit ------- issue 1/87

http://www.vhfcomm.co.uk/

Above — Andy, G8ATD who owns VHF Communications magazine granted me permission to show the DC converter circuit. His magazine archives provide a treasure trove of useful circuits from VHF to Microwave and it's clear Andy passionately spent lots of time publishing the magazine until 2013, plus scanning and organizing the archived material.
Although VHF, UHF and microwave focused, much of the concepts and learning can also enrich your HF exploits.

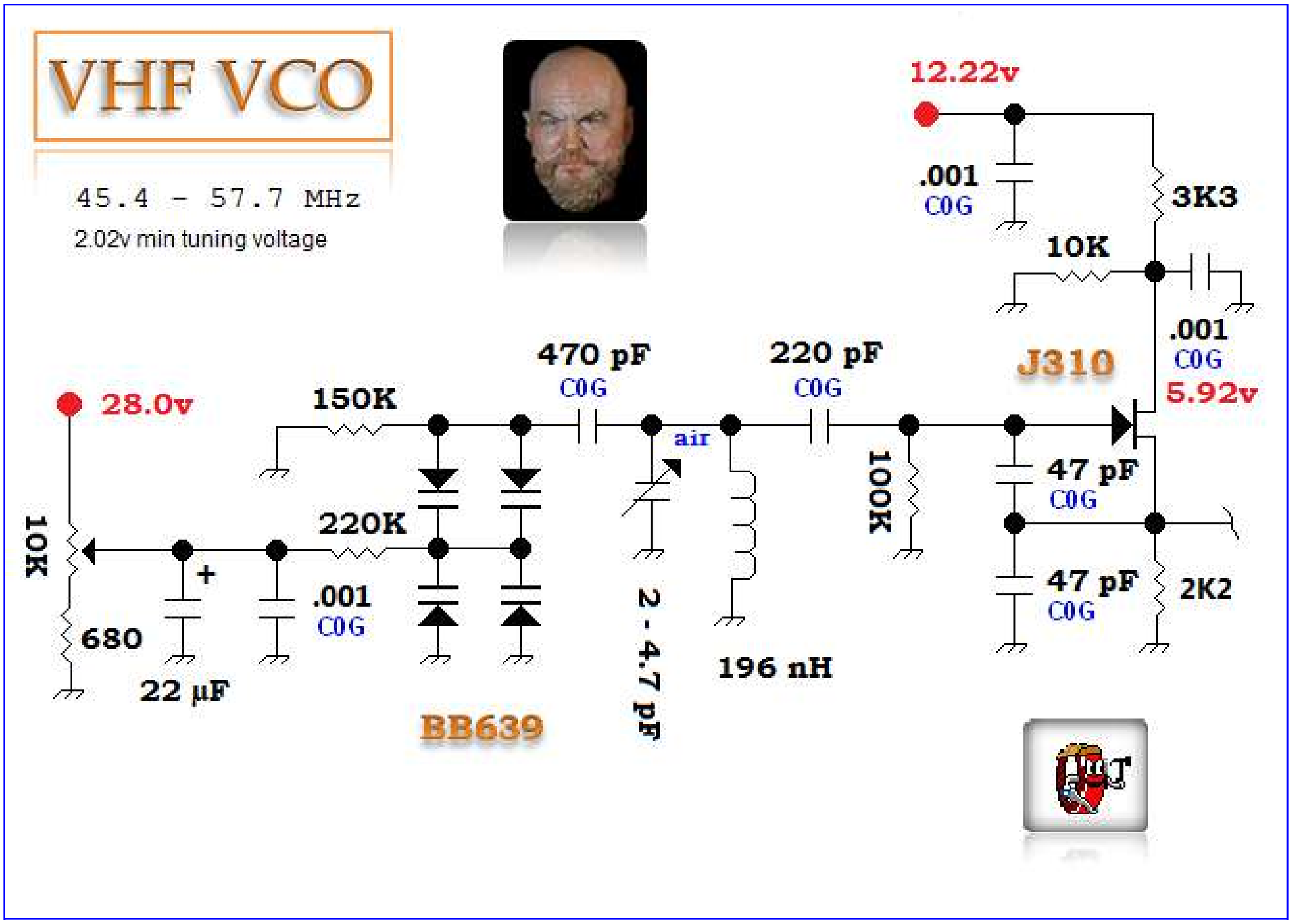
VHF VCO
45.4 - 57.7 MHz
2.02v min tuning voltage
12.22v
.001
C0G
3K3
10K
.001
C0G
J310
5.92v
28.0v
150K
470 pF
C0G
220 pF
C0G
air
100K
47 pF
C0G
10K
220K
+
.001
C0G
2 - 4.7 pF
47 pF
C0G
2K2
680
22 µF
196 nH
BB639

Above — A "quicky" VCO thrown together to test the DC-converter. Click and click for the output with the tuning pot set to fully CCW and then

CW. The 680 R keeps about 2 volts on the varactors with the tuning pot set to CCW.

In the reference articles cited above, you'll find 2 HF-VHF Hartley VFO designs that tune over a 20 MHz span thanks to 28 volt varactors and careful design. In yet another UHF circuit, a 1 octave tuning span is realized with the author's specially designed VCO. Truly hardcore design from a great teacher — I crave exposure to the work of such authors.

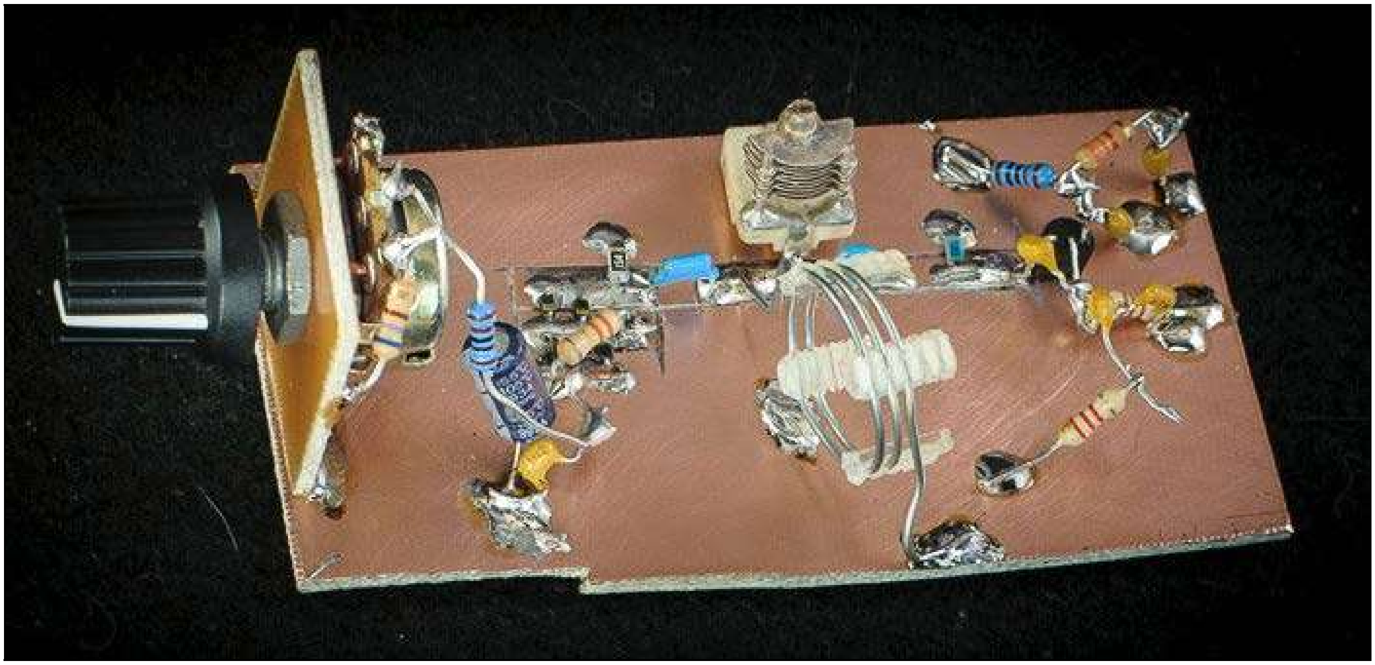

Above — A photo of my "quicky" VHF VCO. 73!

4. Supplemental Web Page #1

Click for the first supplemental web page.

5. Miscellaneous Photos or Figures

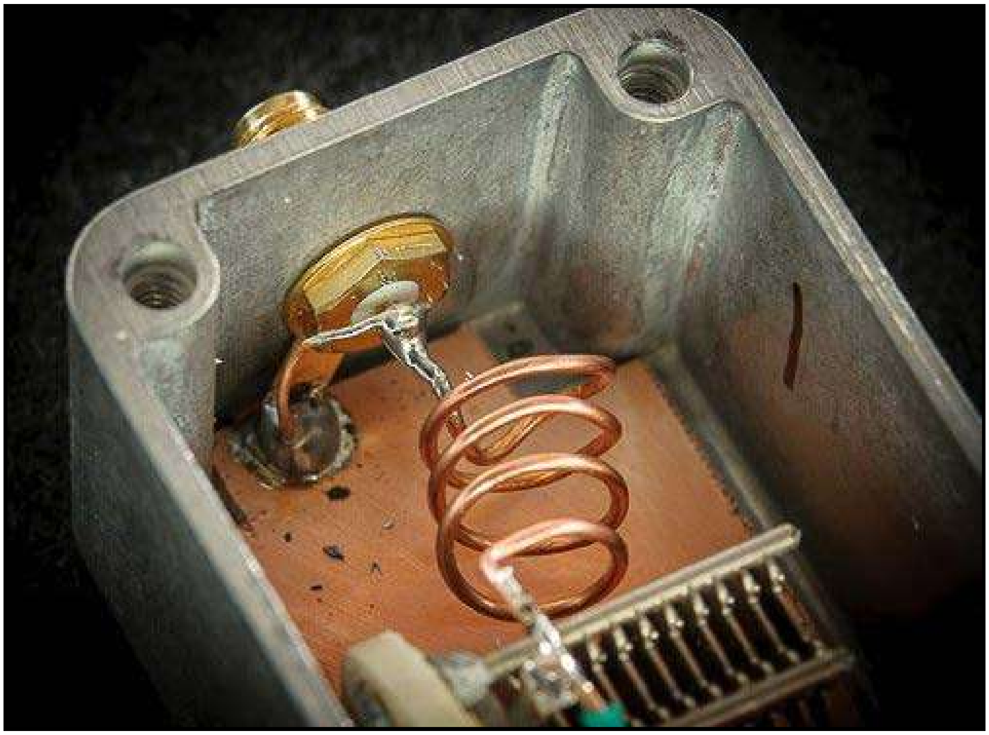

Regency
TRANS
OFF | VOL
SQUELCH
MIC
F1
F2
F3
F4
BTL-304
VHF-FM
EIVER

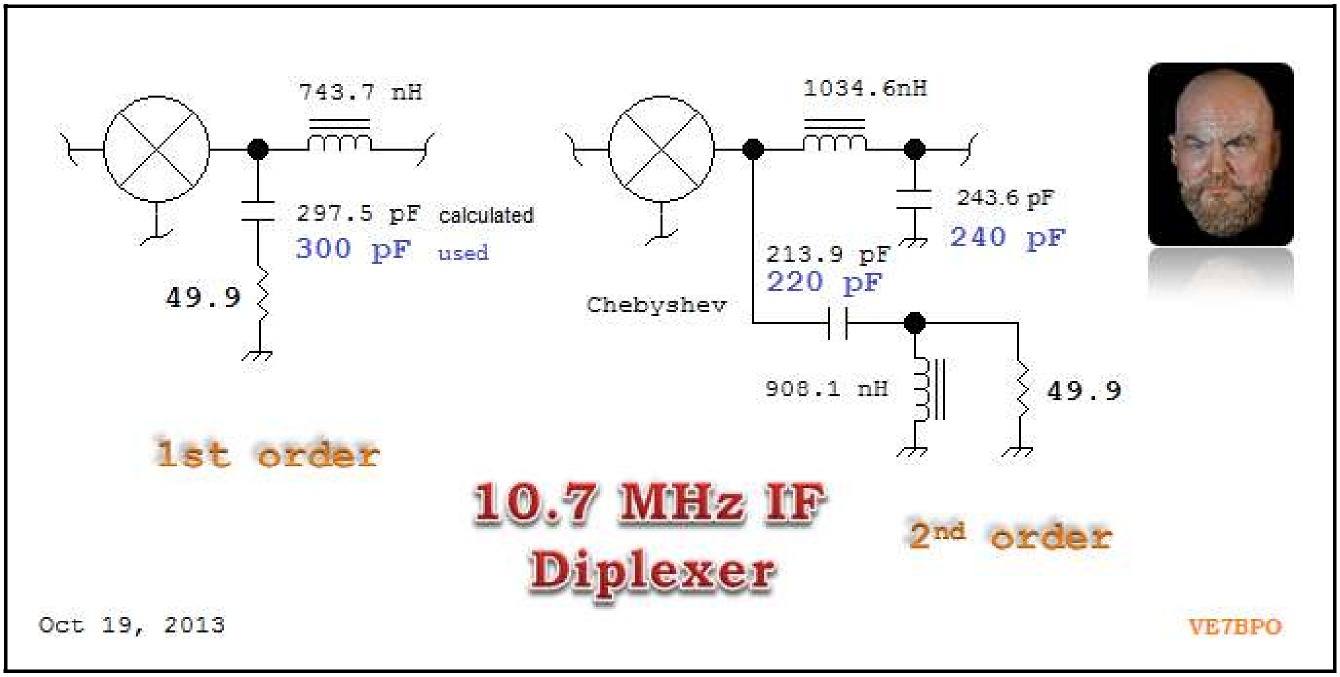
743.7 nH
297.5 pF calculated
300 pF used
49.9
1st order
1034.6nH
243.6 pF
240 pF
213.9 pF
220 pF
Chebyshev
908.1 nH
49.9
2nd order
10.7 MHz IF
Diplexer
Oct 19, 2013
VE7BPO

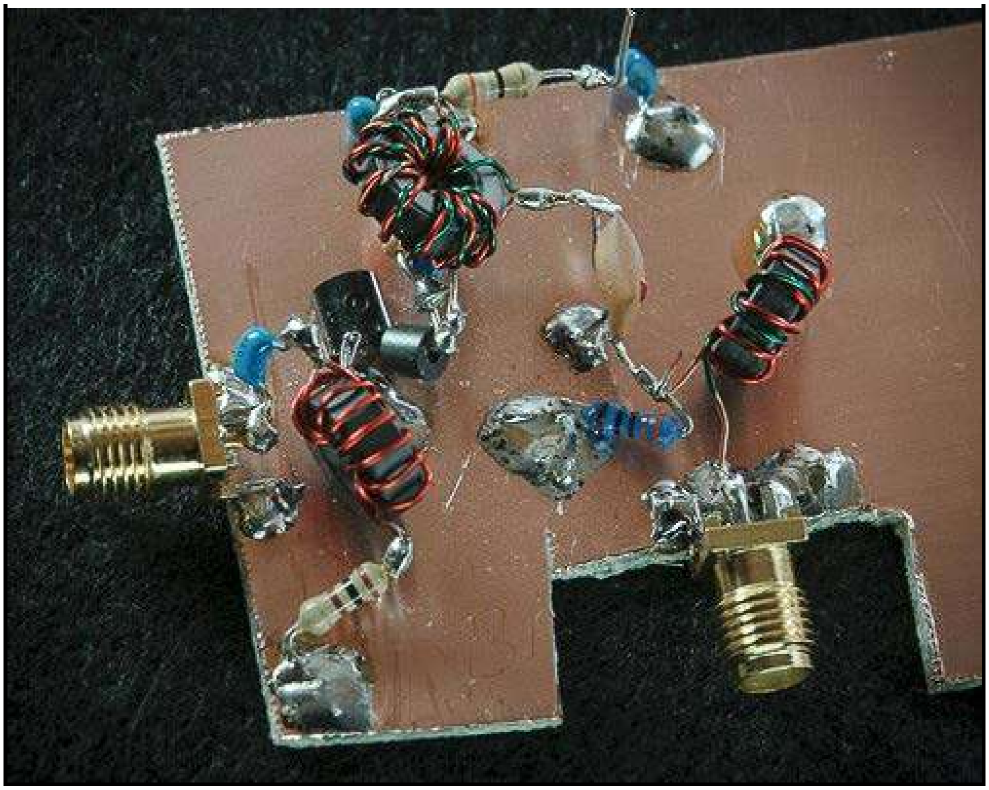

DELL

Sundry Experiments 2012 - 2013

This page shows some of my better non-VHF experiments for 2012-2013.

Although VHF and UHF excite me greatly, It's always fun to build at HF, or even AF.

Section 1: I explain why you might see sweeps that look like oscilloscope tracings on QRPHB: they're devices swept with equipment designed by Bob, K3NHI. Bob's work simply amazes me — full-on, creative precocity.

Section 2: An LM1875 AF power amplifer test.

Section 3: Three Questions with Jason, NT7S.

Section 4: EMRFD Experiments — A 1-on-1 Tracking or Offset Phase-locked Loop.

Section 5: Boot-strapped Popcorn AF Feedback Pair.

Section 6: Non-Mechanical Iambic Paddle.

Section 7: A Journey Above HF.

Section 8: Popcorn AF Amplifier — Reprise

Section 9: The Progressive Receiver by John, K5IRK and Wes, W7ZOI

Section 10: Miscellaneous Pictures and Figures

---

1. Analog Sweep System

Today, advanced experimenters might build a network analyzer/sweeper incorporating a microcontroller, a DDS, or Si570 based frequency synthesizer, plus the needed analog RF circuitry. I went another direction: the Bob Kopski, K3NHI sweep system — all analog, no lines of code and probably 4X the bench work. I show some photos, traces and text in hope it might inspire you to pursue your own sweep system — digital-based or otherwise.

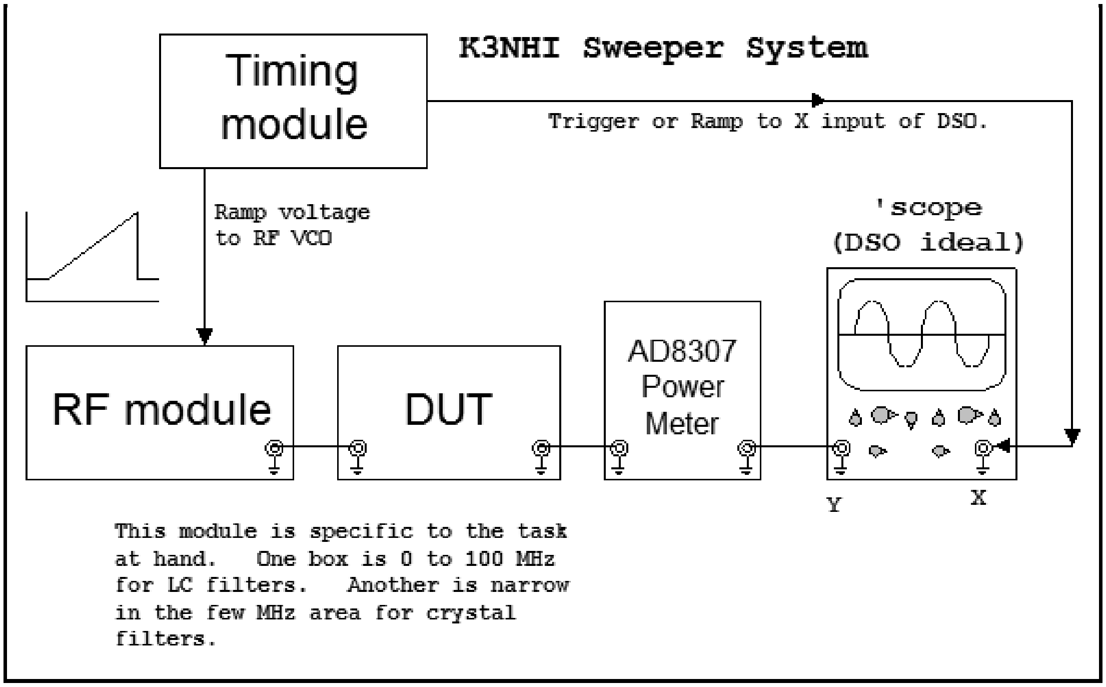
K3NHI Sweeper System
Timing module
Trigger or Ramp to X input of DSO.
Ramp voltage to RF VCO
'scope (DSO ideal)
RF module
DUT
AD8307 Power Meter
Y
X
This module is specific to the task at hand. One box is 0 to 100 MHz for LC filters. Another is narrow in the few MHz area for crystal filters.

Above — The K3NHI Sweep System macro diagram drawn by Wes. The resultant trace looks like the output of a tracking generator plus spectrum analyzer. What I like most is that I'm measuring with my "tough" 'scope and need not worry about input power and so forth like we do with expensive RF test gear. I simply love measuring signals with my oscilloscope. Testing circuits with Bob's sweep system compels me to treasure component-level analog design and renews my passion afresh.

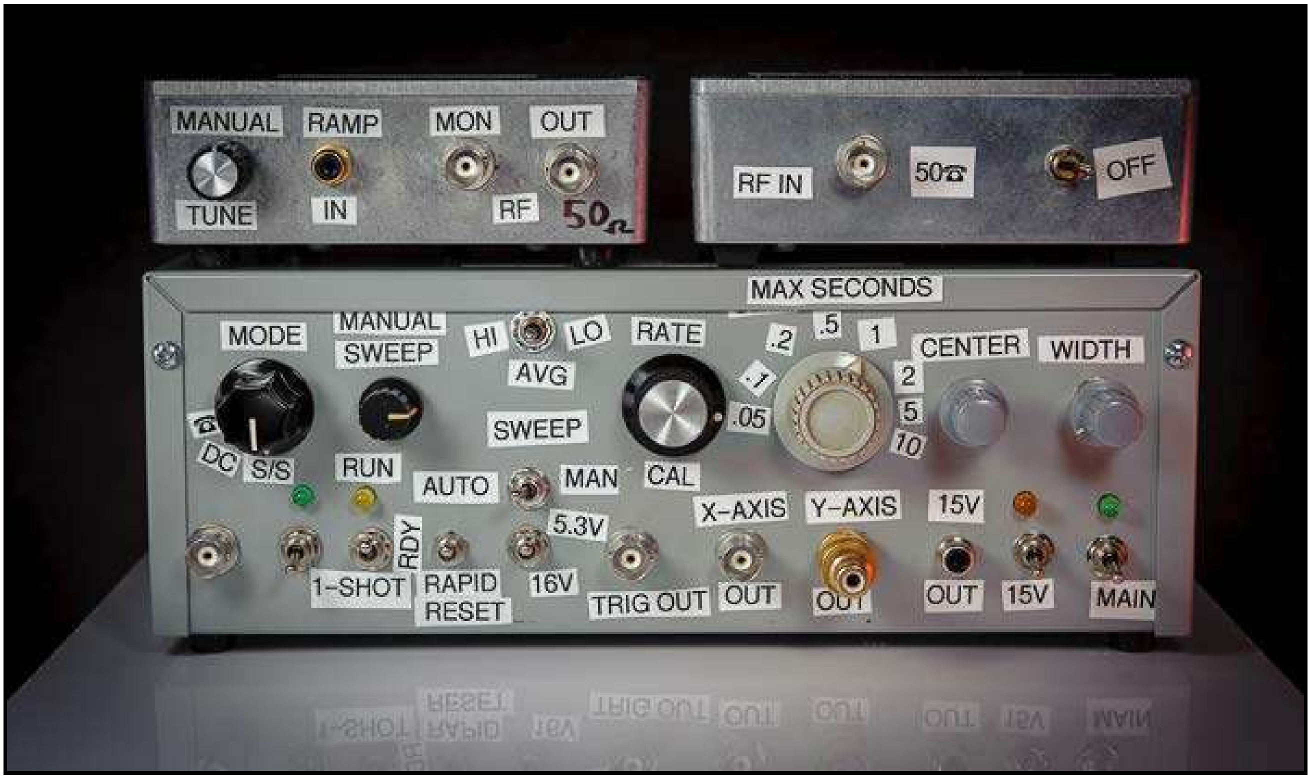
MANUAL
RAMP
MON
OUT
TUNE
IN
RF
50Ω
RF IN
50Ω
OFF
MAX SECONDS
MODE
MANUAL SWEEP
HI
LO
AVG
RATE
.2
.5
1
CENTER
WIDTH
.1
2
.05
5
10
SWEEP
DC
S/S
RUN
AUTO
MAN
CAL
X-AXIS
Y-AXIS
15V
5.3V
RDY
1-SHOT
RAPID RESET
16V
TRIG OUT
OUT
OUT
OUT
15V
MAIN

Above — 3 components of my K3NHI sweep system. Click for a higher resolution photo. To date, I've made the Utility Sweep Generator (time base) *on the bottom*, a 1-118 MHz VCO with clean and level output *top left*, an AD8307 power meter optimized for sweeps *top right* and a crystal filter VCO *shown later*.

Building Bob's Utility Sweep Generator proved difficult. Although technically just a ramp generator, this 1 is calibrated, provides high isolation between the X and Y channels and will sweep *anything*. The power supply has 8 different regulated DC voltages including ~ - 3V. I may use it as the time base for a spectrum analyzer project 1 day.
Bob gave me permission to post the schematics: One Two Three IC1-5 is an LM324 op-amp. Doc 1 Doc 2

Email me for some build notes.

Click, click, or click for photos of the 1- 118 MHz VCO. Click for the schematics. One Two

I show a bare-bones AD8307-based Power Meter (PM) on the **RF Workbench 5** web page, however, to augment the PM for sweeping, I added input compensation, plus some tweaks from Bob's QEX articles: Bob Kopski, K3NHI — An Advanced VHF Wattmeter, QEX, May/June 2002 *and* Bob Kopski, K3NHI — A Simple Enhancement for the Advanced VHF Wattmeter, QEX, Sept/Oct 2003.

I strongly recommend you build Bob's power meter (referenced above) if you're contemplating a power meter build.

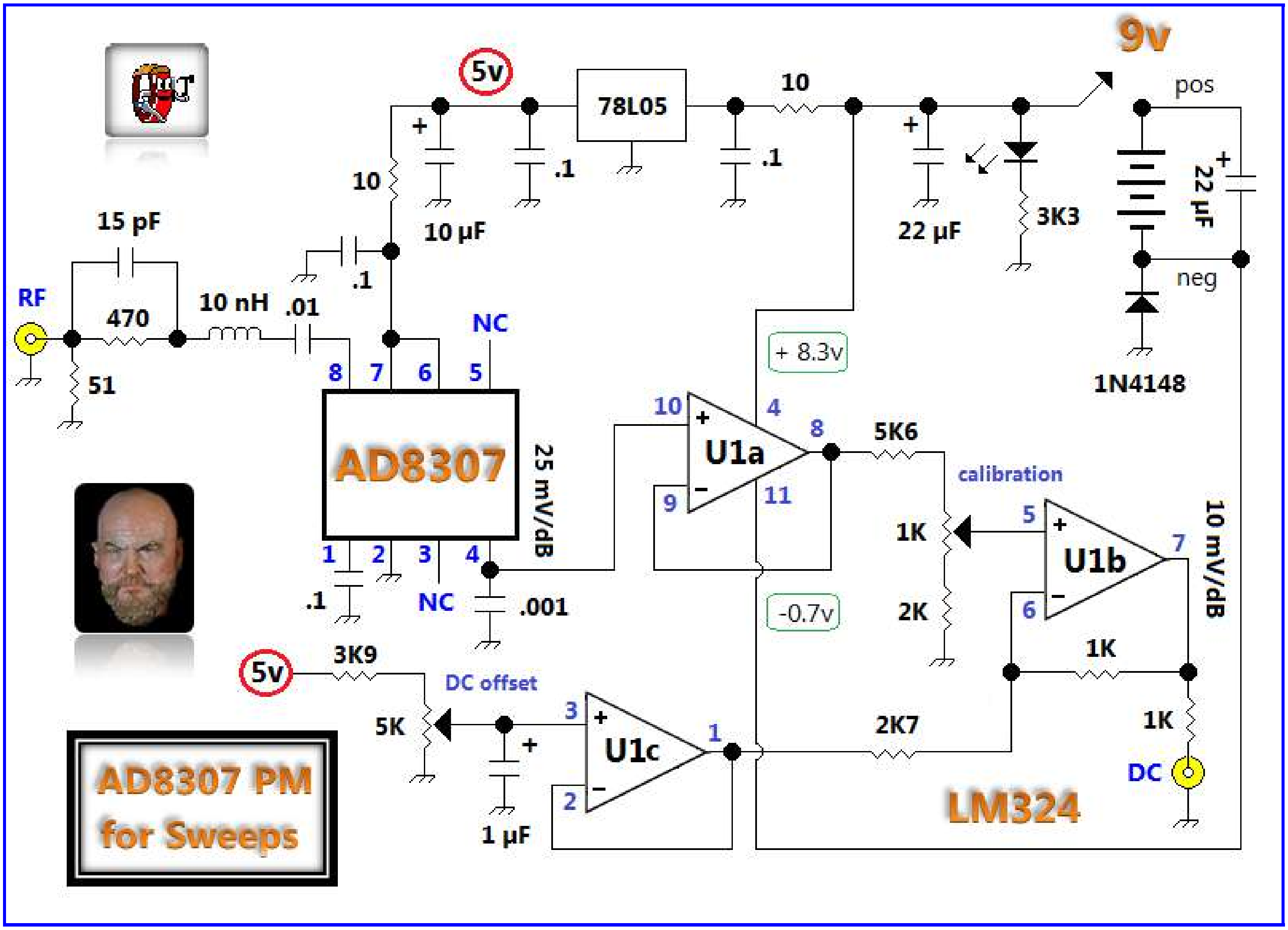
9v
5v
78L05
10
pos
+
.1
.1
10
10 µF
22 µF
3K3
22 µF
15 pF
.1
neg
RF
470
10 nH
.01
NC
1N4148
8
7
6
5
+ 8.3v
51
AD8307
25 mV/dB
10
U1a
4
8
5K6
9
11
calibration
1K
5
U1b
7
10 mV/dB
1
2
3
4
.1
NC
.001
-0.7v
2K
6
1K
5v
3K9
DC offset
5K
3
U1c
1
2K7
1K
2
1 µF
DC
LM324
AD8307 PM
for Sweeps

Above — My AD8307-based power meter. Click for a photo of an early version lacking the level shifter. The level shifter, or DC offset control allows precise Y axis control to enable a resolution up to 1mV/dB when set up properly in the DSO. With this resolution, it's possible to see filter ripple.

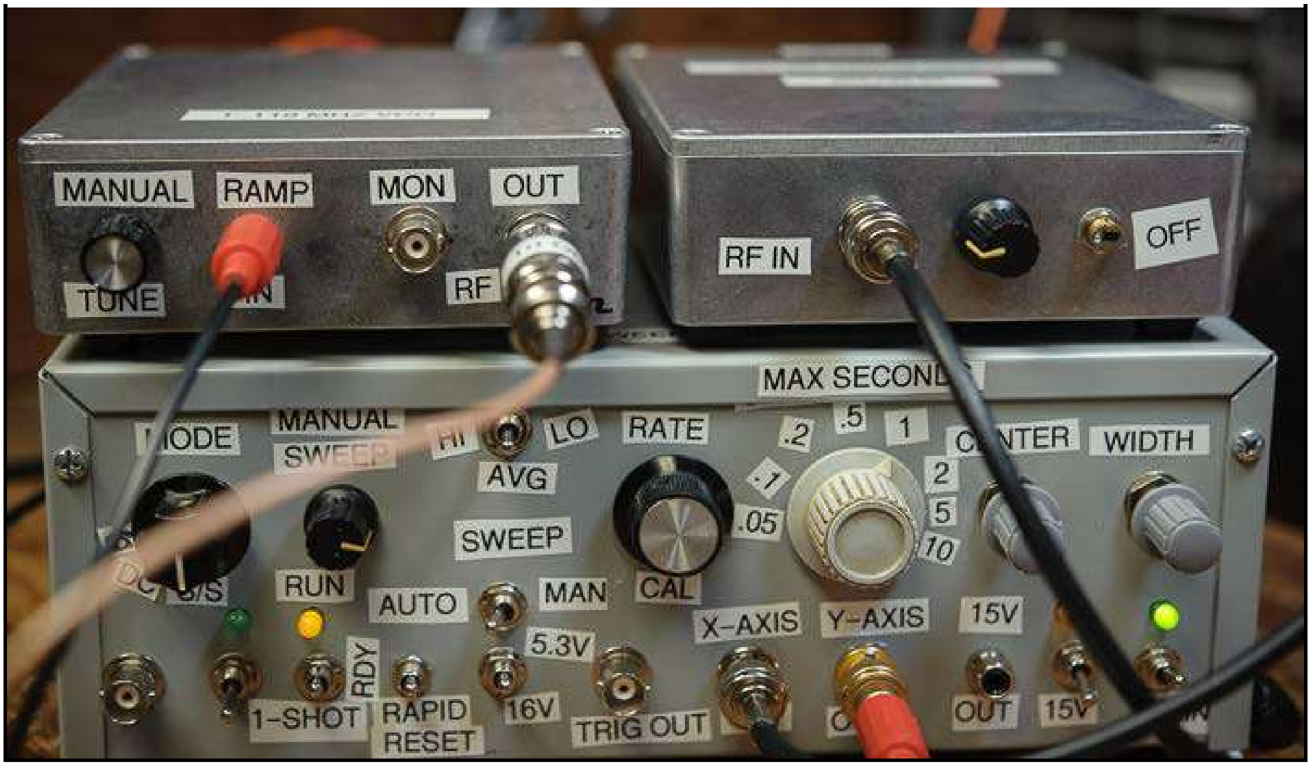
MANUAL
RAMP
MON
OUT
TUNE
IN
RF
RF IN
OFF
MAX SECOND
MODE
MANUAL SWEEP
HI
LO
AVG
RATE
.2
.5
1
CENTER
WIDTH
.1
2
5
.05
10
SWEEP
D/S
RUN
AUTO
MAN
CAL
X-AXIS
Y-AXIS
15V
5.3V
RDY
1-SHOT
RAPID RESET
16V
TRIG OUT
OUT
15V

Above — The K3NHI sweep system in action. This photo shows the power meter with the (later) added DC offset control. Using the offset control potentiometer, I'm able to examine the top of a signal peak at 1 dB per division with various spans.

7 MHZ BANDPASS
BW-237 KHZ IL-2.1 DB

Above — I swept a 7 MHz band-pass filter bench module.

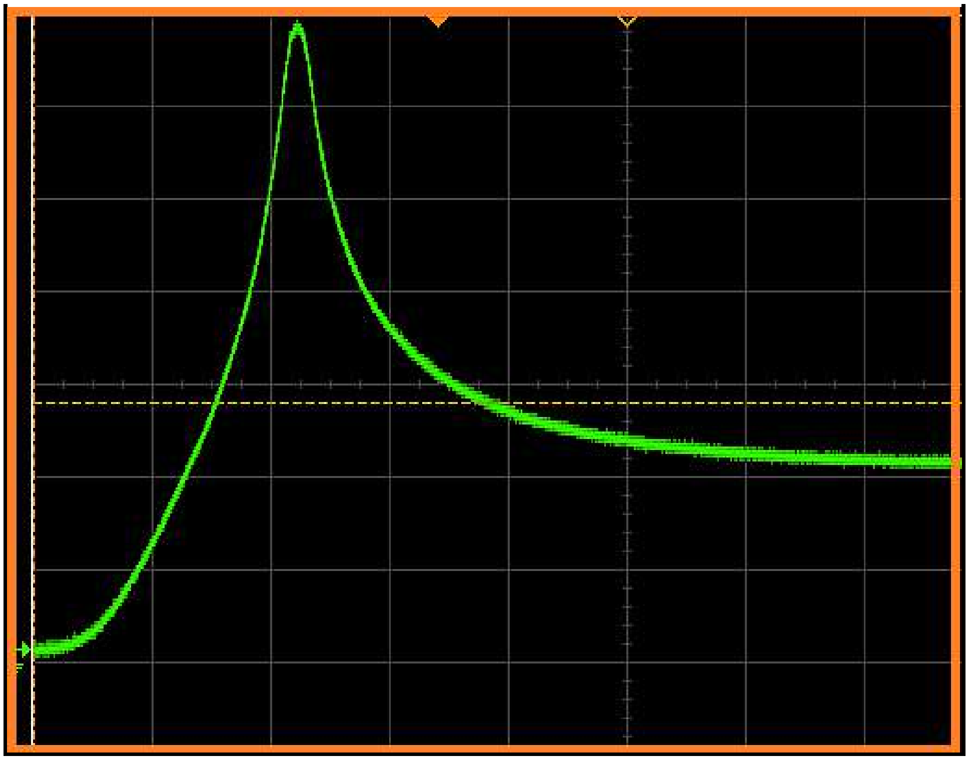

Above — A sweep of the above 7 MHz band-pass filter. At the time I was still learning system calibration and remember feeling blissful that I made such a cool sweep system.

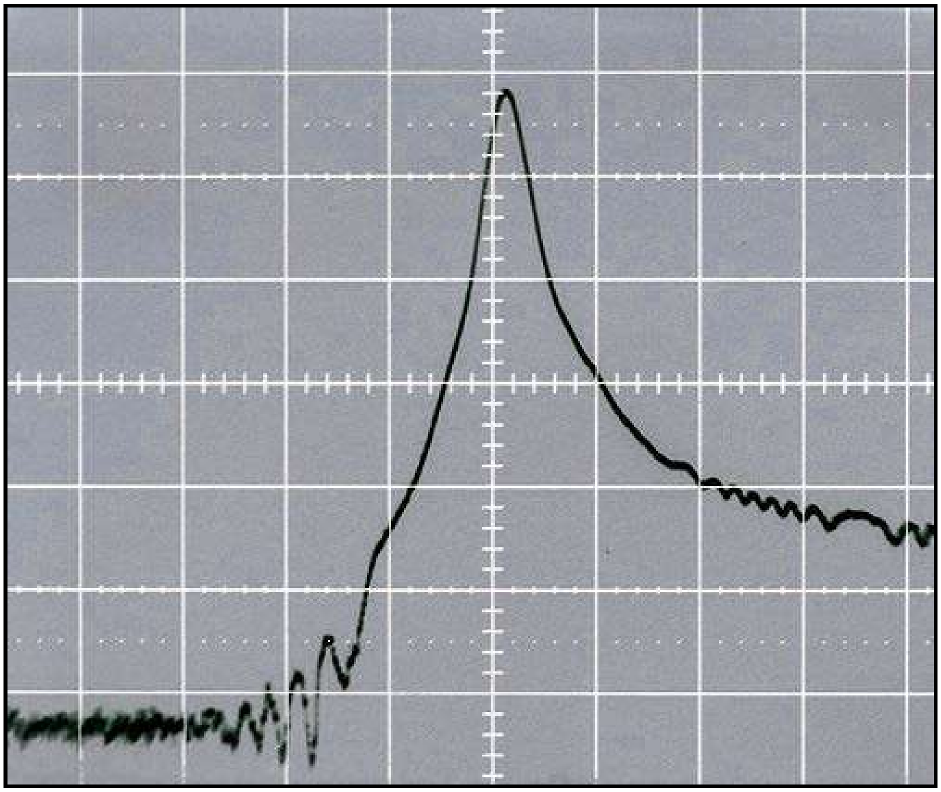

Above — A sweep of the 7 MHz band-pass filter with a tracking generator plus spectrum analyzer for comparison. The span differed from that used with the K3NHI sweep system, however, I'm sure I'll get better with the K3NHI system over time.

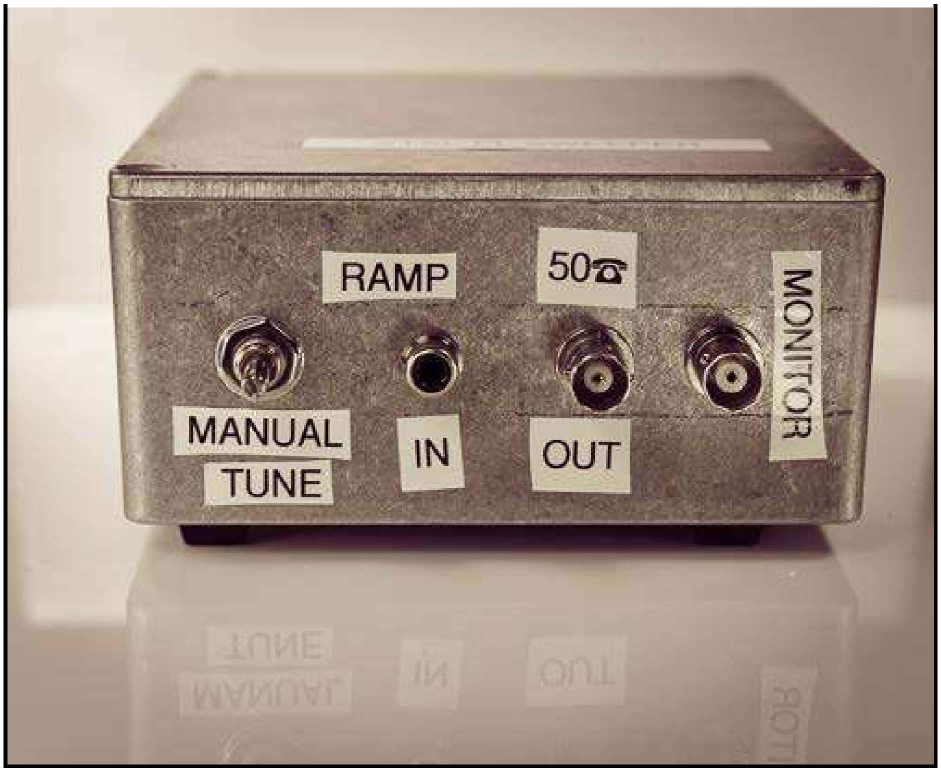

Above — My build of the K3NHI *Hartley* VCO for sweeping crystal filters. I ordered a 4 mm tuning knob for the 10-turn "Manual Tune" potentiometer used to help center the sweep in my 'scope. This 10-turn pot, a DIP switch with C's and a secondary L, plus two MV209 varactors allow narrow resolution sweeps (< 5 Hz nominal) within a ~2.5 - 18 MHz range.

I monitor the sweep frequency on a counter via the VCO monitor port and all the tweaks on this VCO and the Utility Sweep Generator allow easy filter centering. Click for a breadboard of the VCO with the first buffer and a temporary BNC connector for testing. Click for the whole project. The secondary wideband buffer provides strong signal fidelity, reverse isolation and output return loss (22.3 dB) — it draws 61 mA. Click for the VCO schematic courtesy of Bob, K3NHI. Click for a side-by-side of a xtal filter as measured with Bob's sweep system plus an N2PK VNA.

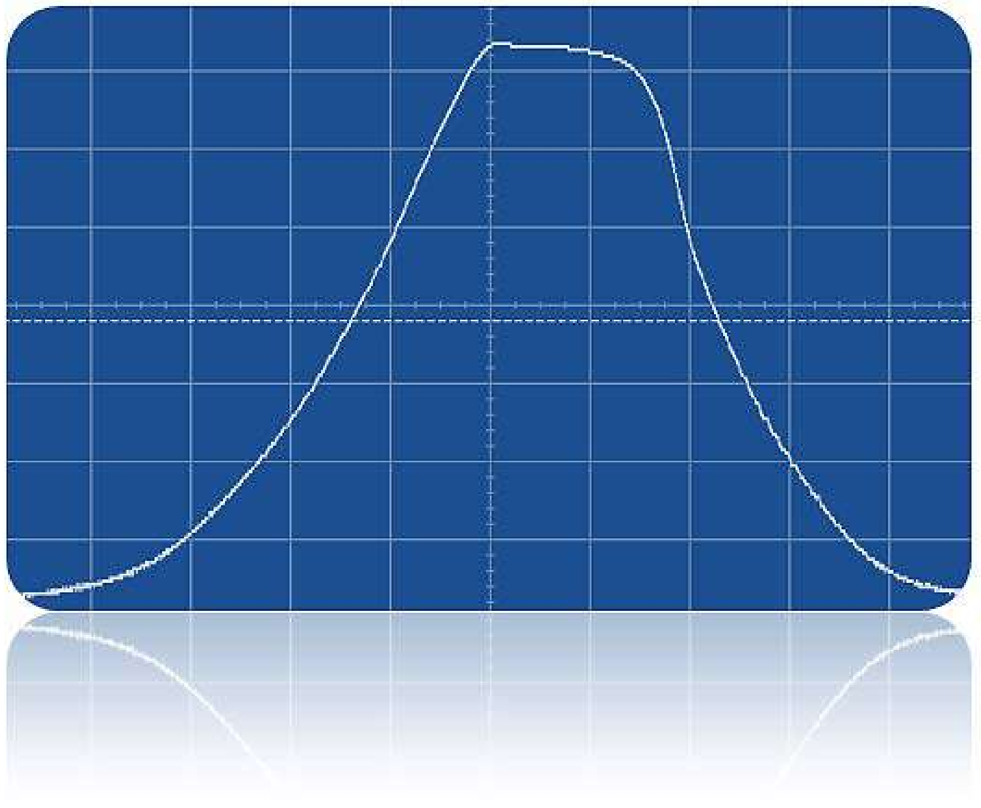

Above — I made and then swept a simple ~500 Hertz wide 4.9152 MHz Cohn or Min-Loss filter using 4 crystals.

So now, if you see sweeps on the site that look like oscilloscope tracings, you'll understand how they were created. I've learned so much from Bob's work and his mentorship last Spring. John, K5IRK coached me also.

QRP-POSDATA for October 2013

3 builders incorporated Bob's sweep system circuitry into projects including a receiver, a spectrum analyzer and the following build of the Utility Sweep Generator (USG) by Jay: Jay built some PC boards to simplify the wiring challenge this USG presents. Click 1 Click 2 Click 3. Great stuff — thanks for sharing.

QRP-POSDATA for March 2014 — Poor Hams Scalar Network Analyzer (PHSNA) —

Above — The PHSNA built by Mikey, WB8ICN.

Jerry W5JH, along with Jim, N5IB and Nick, WA5BDU developed this low-cost sweeper/ lab toolkit.

The Poor Hams Scalar Network Analyzer consists of an Arduino UNO R3, plus an AD9850, or AD9851 DDS, a W7ZOI/W7PUA Power Meter and a MS Windows OS based computer. Builders can scan and plot L-C filters, crystal filters, RF amplifiers and such — much like Bob's sweep system — or an HF tracking generator + spectrum analyzer. The PHSNA also measures crystal parameters with little fuss. Connect a return loss bridge to easily sweep return loss measures of the input or output ports of filters, amplifiers, antennas and more. The total cost to build the PHSNA is approximately $50-60 USD.

Mikey graciously sent me some photographs of his PHSNA build. Complete system in his lab with the chassis lids removed for these photos. Monitor photo showing menu choices. Power meter. Mikey's jig to examine crystals with a 12.5 Ω termination. Once you own a calibrated sweep system, you'll wonder how you ever managed without 1.

Thanks again to Mikey for the photos. I built and tested the return loss bridge using the PCB from the PHSNA Yahoo group. Click for a 613 KB pdf file of my build. Nick, WA5BDU web site.

Hats off to Jerry and crew for this open-system project! A Yahoo group called PHSNA serves as the communications hub and houses superb, detailed documentation. You need to join Yahoo to access this group. Then search for PHNSA and while your at it, also sign onto the EMRFD group.

---

2. LM1875 Audio Power Amplifier

I tested the LM1875 AF power amplifer because its specifications look great: 20 watts into a 4Ω or 8Ω load on ±25V supplies and a TO-220 package for easy heat sinking. Of course, for this web site, I tested it with a typical radio experimenter bench power supply; a single-supply at ~ 12 VDC.

This is probably not a great part for Ugly Construction and I *attempted* to return the load ground, the output Zobel RC filter network, feedback loop and input grounds to a central grounding point through separate paths cut paths into my copper board. A better breadboard method might include the so called "star grounding".

I saw RF oscillations on the 'scope and removed them by soldering a 0.001µF bypass capacitor across the input. A 470 pF bypass capacitor did not work well enough. The datasheet describes specific causes and cures for RF oscillations and I've learned they must be heeded. I once found similar problems with an LM380.

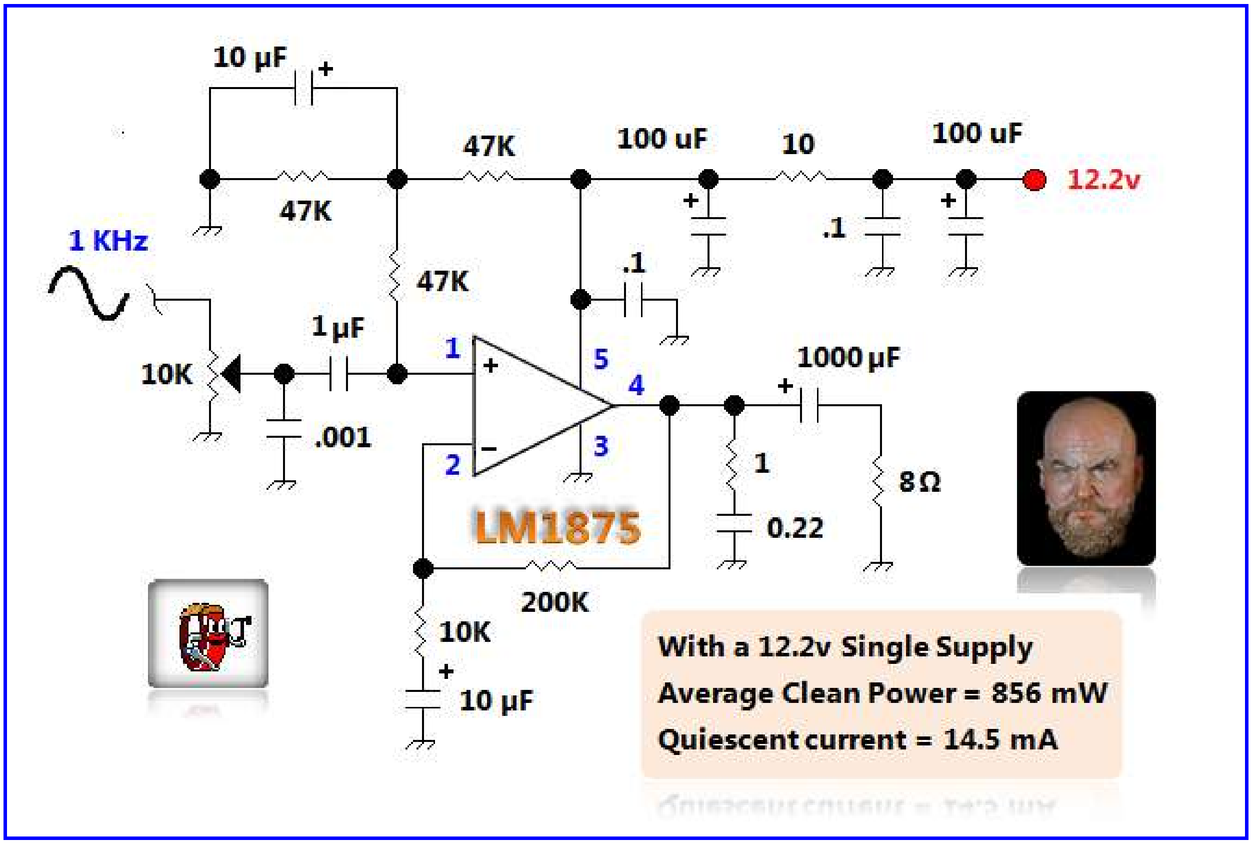
10 µF
+
47K
47K
100 uF
10
100 uF
12.2v
+
.1
+
1 KHz
47K
.1
10K
1 µF
1
+
5
4
1000 µF
+
.001
−
3
2
1
8 Ω
LM1875
0.22
200K
10K
+
10 µF
With a 12.2v Single Supply
Average Clean Power = 856 mW
Quiescent current = 14.5 mA

Above — The LM1875 in the ~suggested datasheet, single-supply set up. This amplifer reminds me of setting up an op-amp. Unlike the LM380, within limits, you may choose the gain to suit your needs. As shown the gain = 25.6 dB. Dropping the feedback R to 100K dropped the gain down to 20.5 dB. For clean output power capacity; it blows away the LM380.

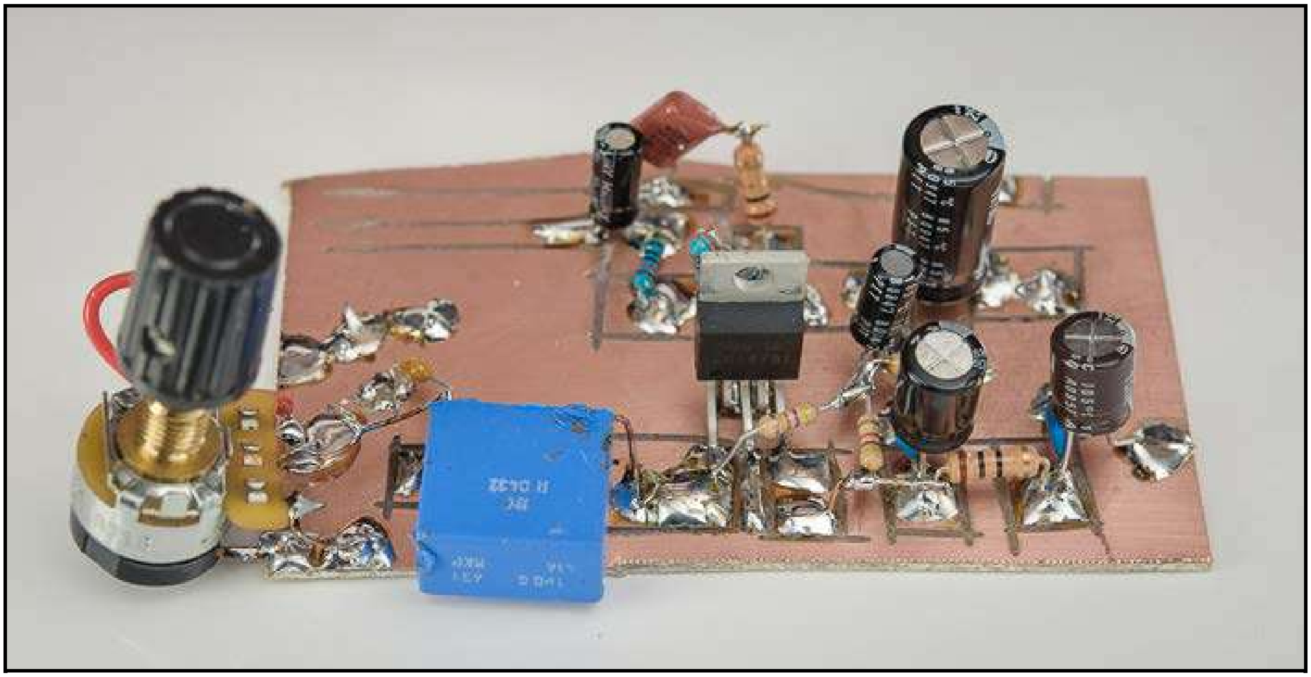

Above — My first LM1875 test breadboard. After the photo, I moved the 0.1 µF RF bypass cap right onto lead 5 — we should carefully RF bypass device power supply leads, but I got sloppy.

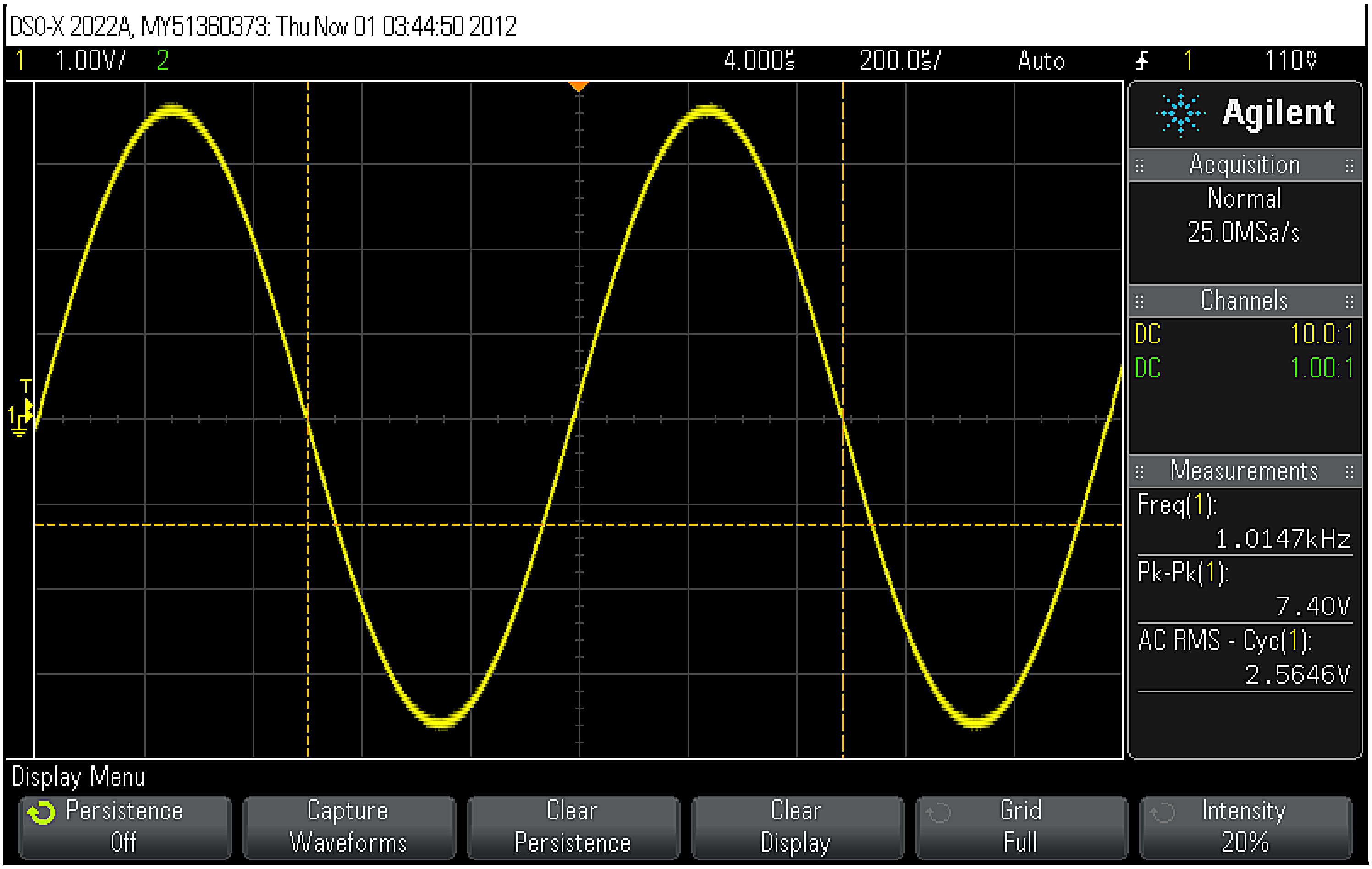

DSO-X 2022A, MY51360373: Thu Nov 01 03:44:50 2012

1 1.00V/ 2 4.000µs 200.0µs/ Auto 1 110mV

Acquisition
Normal
25.0MSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(1):
1.0147kHz
Pk-Pk(1):
7.40V
AC RMS - Cyc(1):
2.5646V

Agilent
Display Menu
Persistence
Off
Capture
Waveforms
Clear
Persistence
Clear
Display
Grid
Full
Intensity
20%

Above — At the maximum power before visible sine wave distortion appears; 856 mW. I listened to this amp while connected to a line-level tape player and an 8 Ω speaker load: very nice. I want to try it with a split +/-15 VDC supply and a star grounded breadboard since a 12v single-supply limits the output power so much. Still, at 12 volts single-supply, this IC yielded the highest clean, average output power of any AF Power chip I've tested: a worthy consideration for a high-grade receiver.

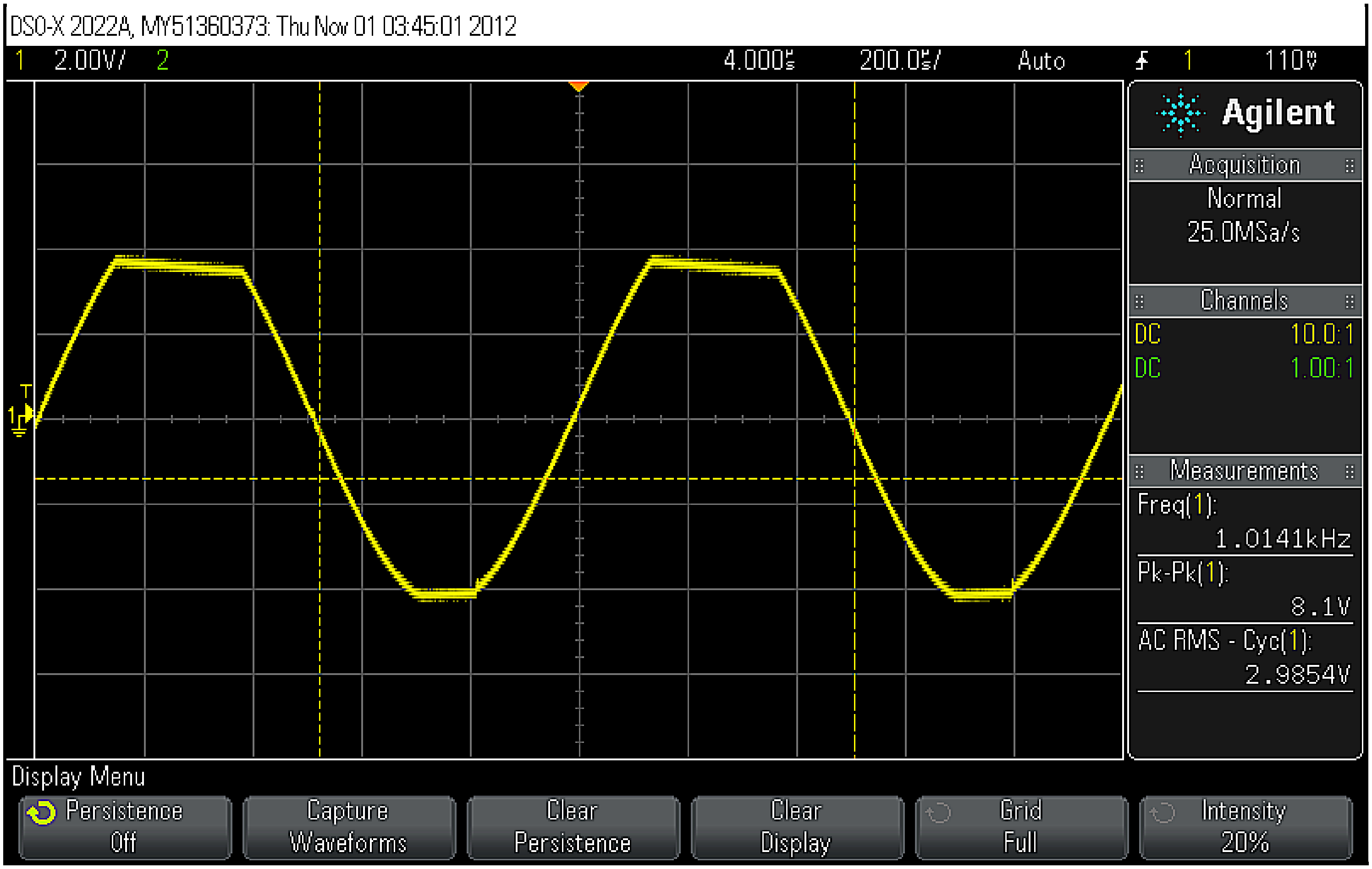
DSO-X 2022A, MY51360373: Thu Nov 01 03:45:01 2012

1 2.00V/ 2 4.000s 200.0s/ Auto 1 110V

Acquisition
Normal
25.0MSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(1):
1.0141kHz
Pk-Pk(1):
8.1V
AC RMS - Cyc(1):
2.9854V

Display Menu
Persistence Off
Capture Waveforms
Clear Persistence
Clear Display
Grid Full
Intensity 20%

Above — 'Scope tracing; I advanced the 10K volume pot to drive the amp into clipping.

---

3. Interview — Three Questions with Jason, NT7S from Etherkit

I follow 2 English language amateur radio blogs — 1 is Ripples in the Ether by Jason, NT7S

Possessing a modern flare, Jason, the blogger, gently but resolutely challenges some of the cliquish, dogmatic thought and behavior that tarnishes amateur radio, or just blogs about fun stuff. He writes well — creating an emotional dialog that stimulates thought and reflection. We get a sense that he cares about our radio hobby and wants it to grow and improve.

Jason, the man behind Etherkit, champions a modern, open-source vision that I find both positive and refreshing.

1. Tell me about your decision to embrace the open source software philosophy for your hardware in a time where proprietary code, copyrights and patents still hold strong. How do Ham Radio equipment sellers benefit from code sharing?

I believe that the open sharing of knowledge has always been one of the cornerstones of our amateur radio community, going back to its earliest days. So the open source/open hardware ethos has always resonated with me in regard to our hobby. I started Etherkit with the intention of providing a small bit of income to my family and as a way to promote the idea of open hardware within the ham radio community at large. I have no illusions of becoming the next Elecraft, but I hope that I can build up a stable of affordable and fully-open ham radio kits that will be "hackable" and extensible for the motivated experimenter. I do this by providing the full source code for my microcontroller firmware, all of the PCB design files, Creative Commons licensed documentation, and programming ports for my products. I've already seen some neat examples of customers extending my first product (the OpenBeacon MEPT kit) by doing things such as adding in WWVB time discipline and pairing it up with a Raspberry Pi for cheap automation. I hope that others will take my code or my circuits and re-purpose them in their own work, even if they don't buy my products.

I am not an open source zealot and do not begrudge the large majority of vendors who choose to keep their intellectual property closed. However, most of what us smaller companies do is not on the cutting edge of radio. We leverage the knowledge and works of those who came before us. Perhaps if I created something wholly-new that would be patent worthy, I would consider keeping it closed, but that's not the kind of products that I'm able to develop as a one-man operation. We do not copy the designs of others, but we do take concepts that are for the most part well-tested and come with new ways to implement them. Because of that, it's my personal opinion that I have a duty to keep my designs open.

In the open hardware world at-large, there is a discussion about whether open source hurts your own business prospects. There are still some debatable points in that discussion, but I think it has been shown that if you look at the entire balance, open hardware is a good thing for smaller companies. One of the largest concerns is that under most open source licenses, a competitor can just clone your hardware and undercut your sales. That is a genuine concern, but I think that products such as Arduino have shown that if you make a quality product, most folks will recognize that quality and stick with the original.

Even if others buy a clone of your hardware, in all likelihood, that may be strengthening your brand identity (as long as that vendor isn't stealing your name). Another concern is that a customer can just copy your product for themselves. To that, I say good!. Because of the work and costs involved (economy of scale), it's going to be time and/or money consuming to make that copy. It's probably cheaper and faster to just buy the kit. The reason you copy it for personal use is because you love working with the technology. Which is exactly what I want to encourage. You may lose a small bit of sales, but I think it gives you more name recognition in the end.

2. What’s it like being a vendor at Dayton?

To be clear, I wasn't a vendor at Hamvention in 2012 (hopefully I will be there by 2014), but I was a vendor at Four Days In May at the Fairborn Holiday Inn. It was a wonderful experience to get to sit with the big names in the QRP world, selling my wares. I got the opportunity to meet tons of QRPers and build up some good relationships. Online sales are wonderful for the ultra-small operations such as myself, but nothing beats actually meeting your customers face-to-face, especially when you are at the world's most well-known QRP convention.

3. In industry, SMT parts are normal and hole-thru might better be called “hold-over”; what’s your view on kitting products with SMT parts?

We've seen some SMT kits within the QRP world, but they still are more of an oddity than anything else. I understand the concerns that people have with SMT assembly, but I think that there is still a lot of trepidation that builders needn't have. It's my opinion that SMT construction with "larger" components such 1206 or 0805 is well within the capabilities of the average kit builder. I also believe that once you are comfortable with SMT construction, it is probably faster and more efficient than through-hole construction.

OpenBeacon is a through-hole product, but I have had a QRP CW rig in development for the last two years that is a SMT design. In beta testing, I've found that one of the biggest challenges in kitting is that I have to clearly identify each and every component. With a through-hole kit, you can just throw all of the resistors or all of the capacitors together because they are clearly marked. Not so with SMT. You have to have a system to keep each value separated from the others and marked with a value. SMT resistors and semiconductors have a laser-etched value, but it's nearly impossible to see by naked-eye, and SMT capacitors generally have no markings at all.

So I have had to compartmentalize each strip of components of the same value cut from a reel, and mark them with a sticker. That is pretty costly and time-consuming. I'm hoping to find ways to streamline this process so that I can release SMT kits without the large time investment that it currently takes.

---

4. EMRFD Experiments — A 1-on-1 Tracking Phase-locked Loop

I built the 1-on-1, or offset phase-locked loop circuitry described on EMRFD page 4.22 and share these schematics in faith you'll create your own. Rich with wisdom and reason, this section lies among the best topics from EMRFD. Please read Wes' notes since I won't repeat his narrative — only supply a few ideas and measurements. In the article closer, Wes suggests some modern parts to raise performance and I applied all of them with the exception of the 14 MHz VCO.

Rather than building the main VCO with divide by N circuitry to allow multiband use, I copied the original 14 MHz oscillator verbatim. Why? Well, I wanted to test this VCO: a design that wisely doesn't expose the varactor to high impedance or signal amplitude and thus avoids forward-biasing the single tuning diode. I've discussed this before on the QRP Modules 2011 web page under **7 MHz VCO Experiments**. Also, I really just wanted to learn about PLL circuitry. The Figure 1 macro schematic below illustrates this project.

In my circuit, a frequency stable 14 MHz VCO = **the goal**; the rest of the circuitry supports this.

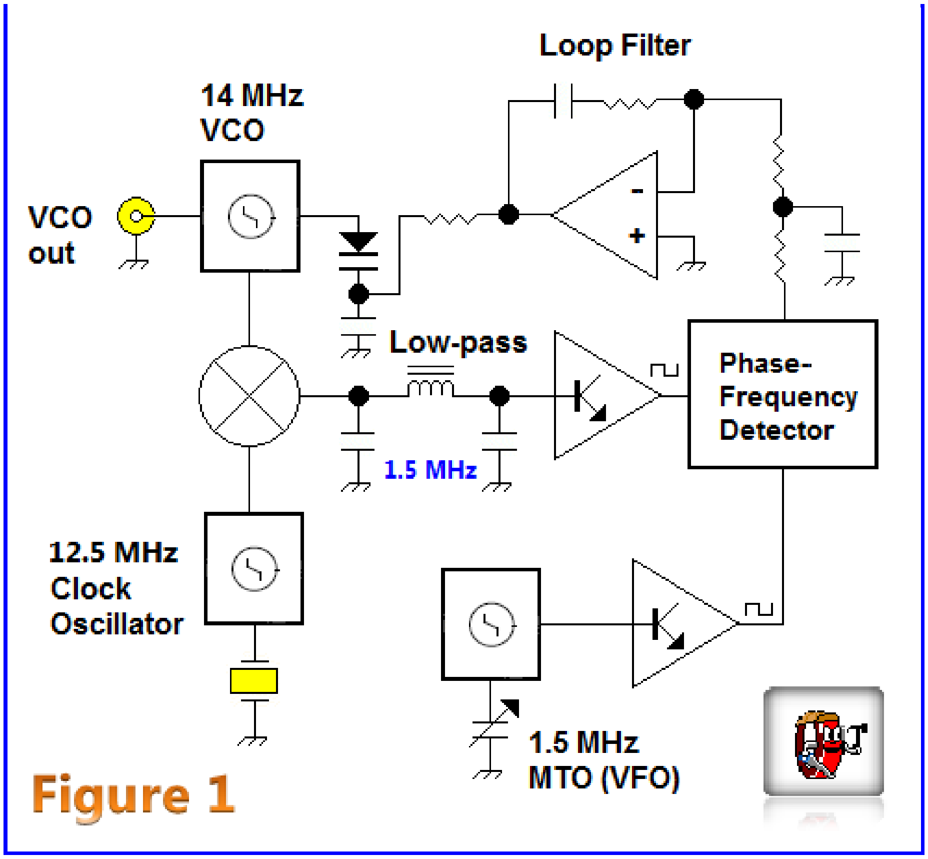
Loop Filter
14 MHz
VCO
VCO
out
-
+
Low-pass
Phase-
Frequency
Detector
1.5 MHz
12.5 MHz
Clock
Oscillator
1.5 MHz
MTO (VFO)

Figure 1

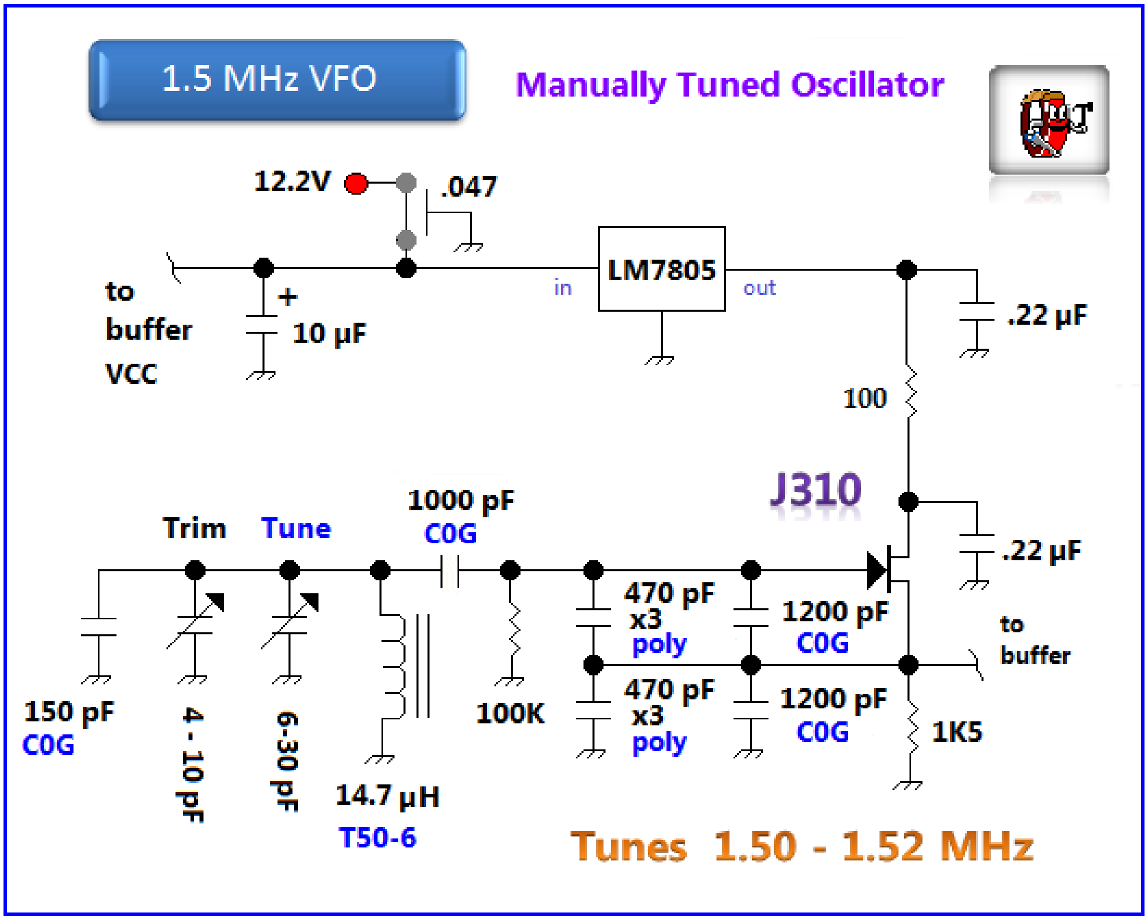
1.5 MHz VFO
Manually Tuned Oscillator
12.2V
.047
to
buffer
VCC
+
10 µF
in
LM7805
out
.22 µF
100
J310
.22 µF
Trim
Tune
1000 pF
C0G
470 pF
x3
poly
1200 pF
C0G
to
buffer
470 pF
x3
poly
1200 pF
C0G
1K5
150 pF
C0G
4 - 10 pF
6-30 pF
14.7 µH
T50-6
100K
Tunes 1.50 - 1.52 MHz

Above — The 1.5 MHz VFO. In his modern writing, Wes calls this the MTO, or Manually Tuned Oscillator *in the context of a tracking PLL.* I wound the L with # 30 AWG wire on a T50-6 toroid.

My MTO exhibits a low tuning range (only 1.50 to 1.52 MHz) since I built in a box with a small air-variable capacitor that swung only 24 pF and I ran the "Colpitts capacitors" at 2610 pF to keep phase noise low. This box normally holds a VHF oscillator and I just removed the main board and swapped in a 1.5 MHz equipped copper board. I won't keep this PLL and thus sticking the 1.5 MHz MTO in an existing oscillator chassis with a grounded tuning shaft and feedthrough capacitor helped save money and time.

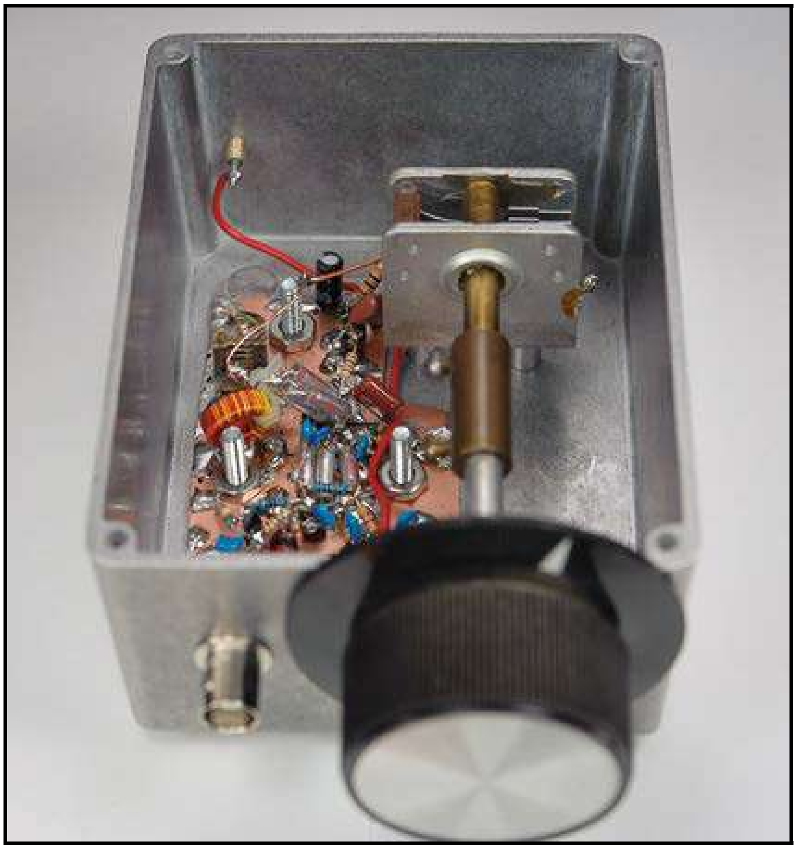

Above — The built 1.5 MHz MTO. With temperature compensation from 6 stiff-leaded, 600 VDC, 470 pF polystyrene capacitors, my frequency drift measured between 3 and 4 hertz per hour upward at room temperature. Properly designed + built + temperature compensated L-C oscillators at 1.5 to 3.5 MHz may exhibit stellar temperature stability. See the VFO - 2011 web page for some tips.

Since this VFO was sublimely frequency stable, I didn't possess the guts to change up the L-C ratio to garner a wider tuning range from the small air-variable tuning capacitor. A 100 pF, or greater delta-F air-variable tuning capacitor would stretch the VFO (MTO) tuning range nicely.

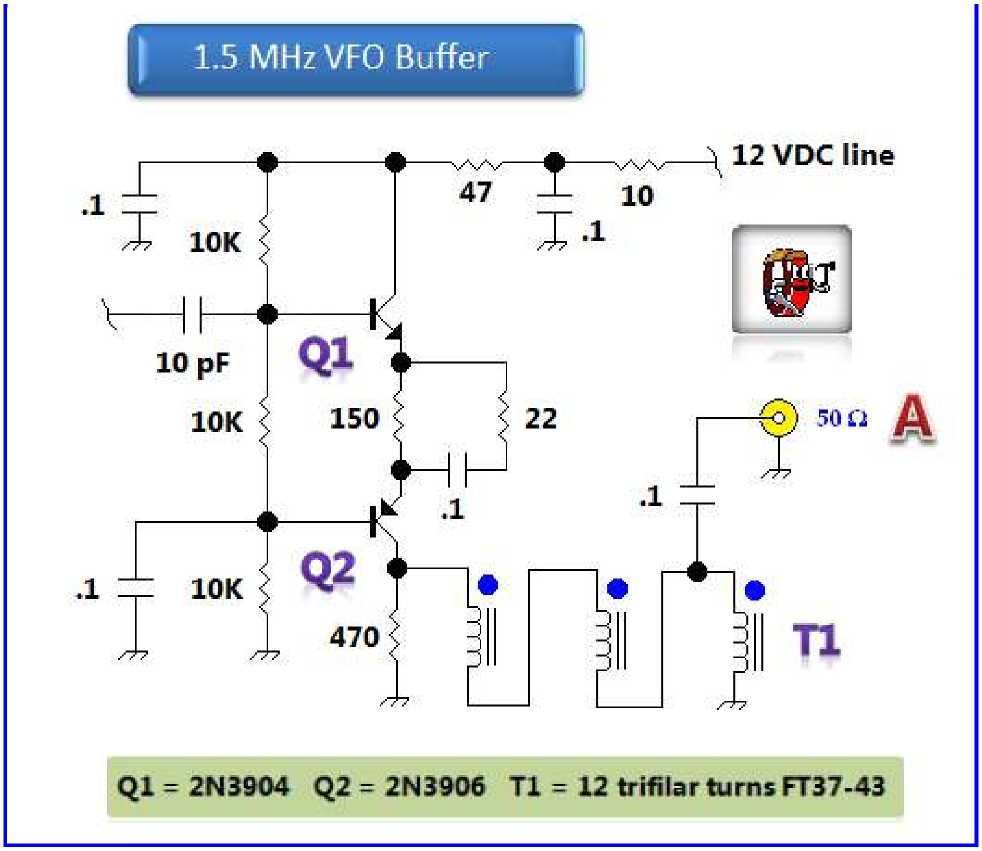
1.5 MHz VFO Buffer
12 VDC line
47
10
.1
.1
10K
10 pF
Q1
10K
150
22
.1
50 Ω
A
.1
.1
10K
Q2
470
T1
Q1 = 2N3904   Q2 = 2N3906   T1 = 12 trifilar turns FT37-43

Above — I designed this buffer last year and it's my new favorite. Click for the original. A 10 pF C0G/NP0 capacitor lightly couples the MTO output to the high impedance of Q1, an emitter follower. Further, a common base amp provides gain and *essential* reverse isolation. You may adjust Q2 gain by changing the degenerative feedback offered by the 22 Ω resistor and 0.1 μF capacitor. For example, decreasing the R to 18 Ω may provide 7 dBm output for a diode ring mixer.

MTO output power = 6.71 dBm.

I transformed the 470 Ω collector resistor impedance to 50 Ω with a transmission line transformer. Even though part of the PLL circuitry involves logic gates, or is at DC; as possible, my circuits employ a 50 Ω input or output impedance to allow measurement with my 50 Ω modules and/or instruments, plus transmission via 50 Ω cables.

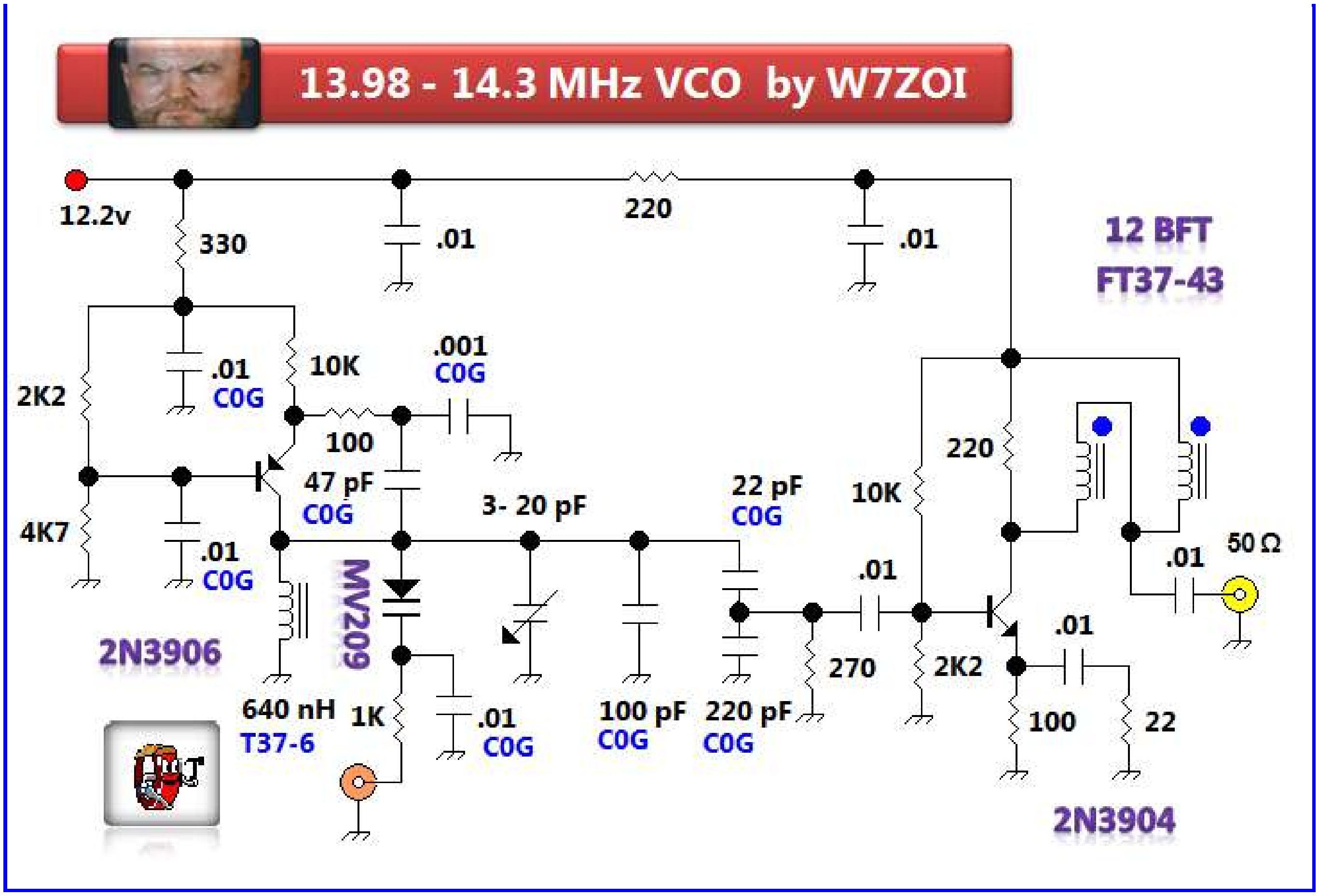
13.98 - 14.3 MHz VCO by W7ZOI
12.2v
330
.01
220
.01
12 BFT
FT37-43
2K2
.01
C0G
10K
.001
C0G
100
47 pF
C0G
3- 20 pF
22 pF
C0G
10K
220
4K7
.01
C0G
MV209
.01
50 Ω
2N3906
640 nH
T37-6
1K
.01
C0G
100 pF
C0G
220 pF
C0G
270
2K2
.01
100
22
2N3904

Above — The 13.98 to 14.3 MHz VCO by Wes, W7ZOI (Figure 4.43 in EMRFD). The connector in series with the 1K varactor resistor was an RCA type. Output power = 1.62 dBm. I employed a 3 - 20 pF air variable for the trimmer.

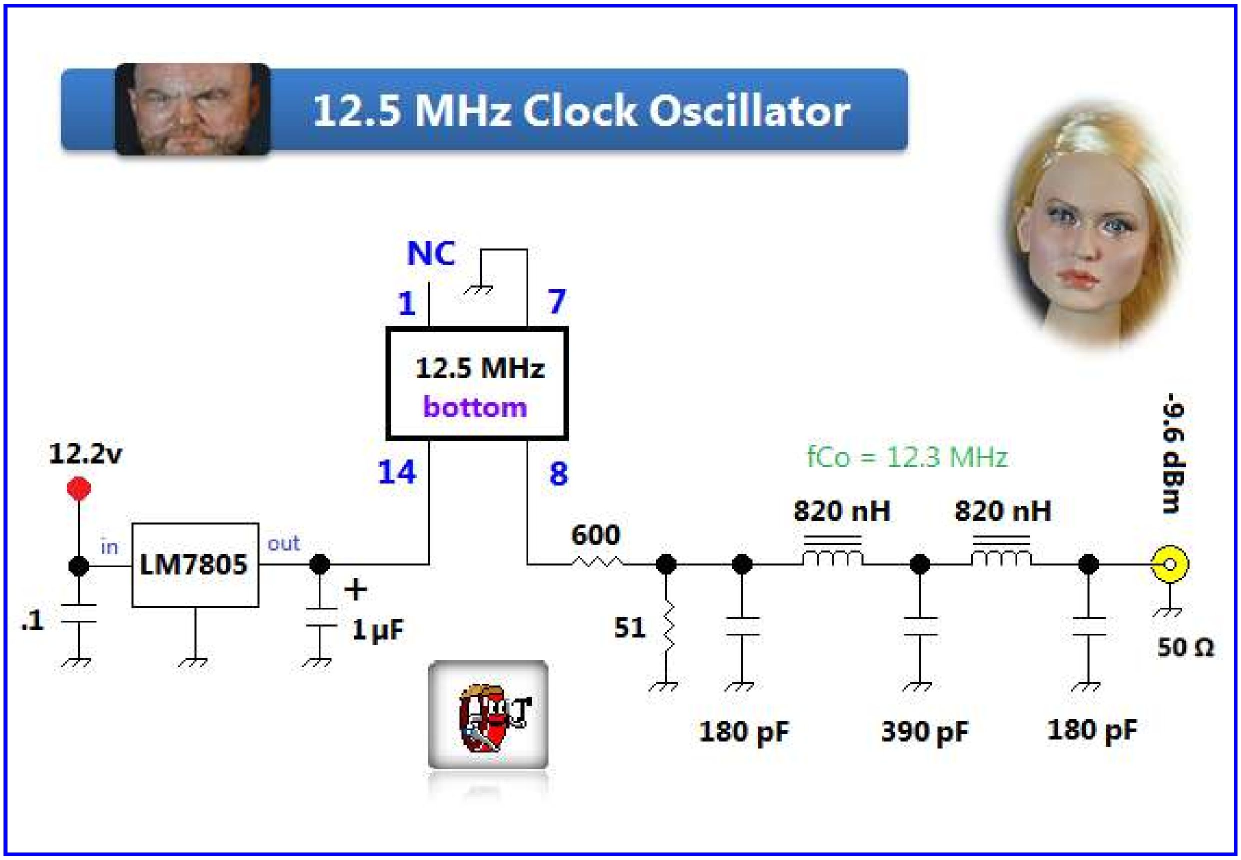
12.5 MHz Clock Oscillator
NC
1
7
12.5 MHz
bottom
14
8
12.2v
in
LM7805
out
+
.1
1 µF
600
51
fCo = 12.3 MHz
820 nH
820 nH
180 pF
390 pF
180 pF
-9.6 dBm
50 Ω

Above — In the original circuit, Wes built his 12.5 MHz crystal oscillator with a single 2N3904. Lacking a 12.5 MHz crystal, I built my xtal oscillator from an old, junkbox 12.5 MHz clock oscillator. A resistor L- matching network drove a low-pass filter to scrub off harmonics, Click for the clean output 'scope tracing at 211 mV pk-pk in my first version. Later, some tweaks gave a final power of -9.6 dBm (208 mV pk-pk). Many authors switch their NE612 mixers with a peak-peak voltage of ~200-300 mV. An AC-coupled 51 Ω resistor on the NE612 pin 6 properly terminates the oscillator to establish the desired drive power and filtering.

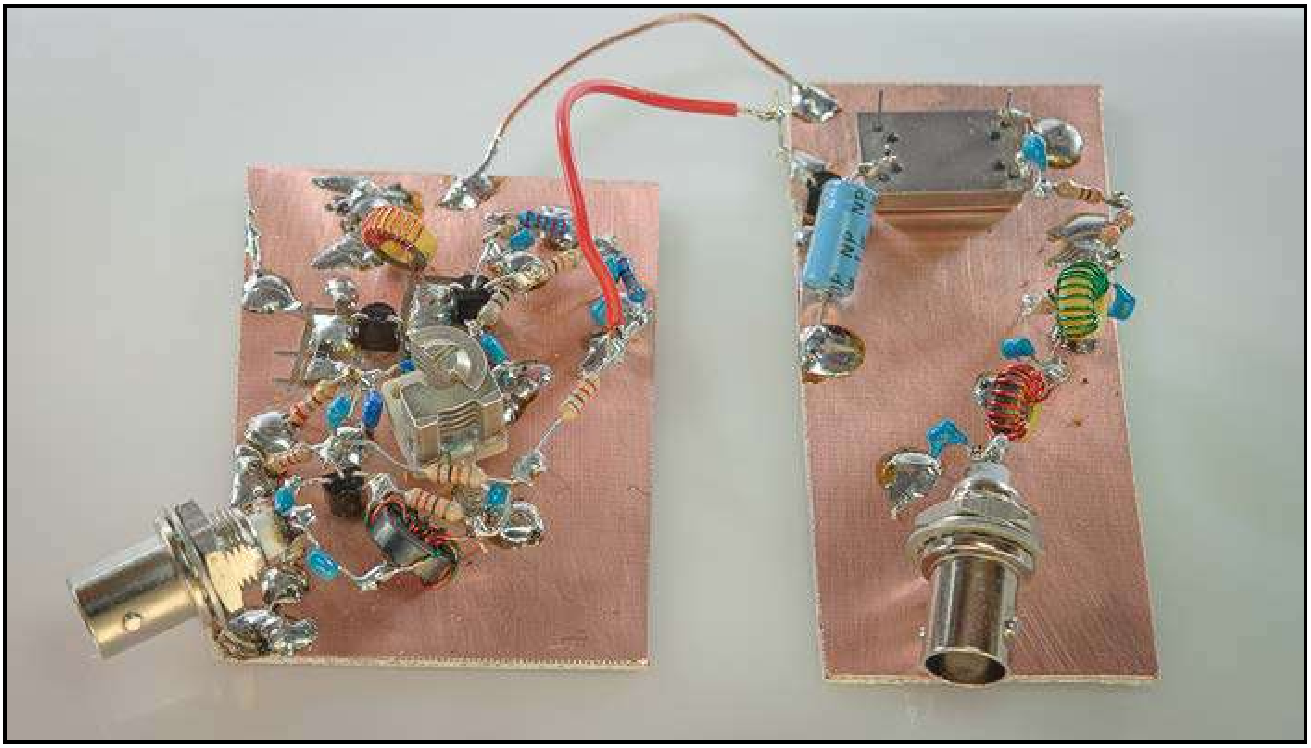

Above — Oscillator breadboards: 14 MHz VCO (left) and the 12.5 MHz clock oscillator (right).

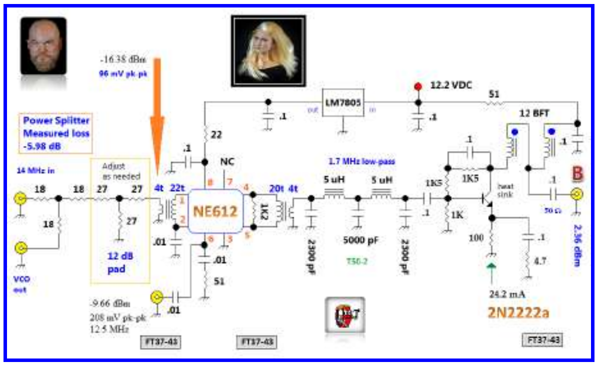
-16.38 dBm
96 mV pk-pk
Power Splitter
Measured loss
-5.98 dB
14 MHz in
18
18
18
VCO out
Adjust as needed
27
27
27
12 dB pad
4t
22t
.1
22
NC
NE612
1
2
8
7
6
3
4
5
.01
.01
51
-9.66 dBm
208 mV pk-pk
12.5 MHz
.01
FT37-43
FT37-43
FT37-43
.1
LM7805
out
in
12.2 VDC
.1
51
12 BFT
1K2
20t
4t
1.7 MHz low-pass
5 uH
5 uH
2300 pF
5000 pF
T50-2
2300 pF
.1
1K5
1K5
1K
.1
heat sink
100
.1
4.7
24.2 mA
2N2222a
.1
.1
B
50 Ω
2.36 dBm

Above — Click to view the power splitter, mixer, low-pass filter and amplifier schematic. Click for an FFT of the clean 1.5 MHz output sine wave. At this point, the 14 MHz VCO has no DC voltage connected to its frequency compensation varactor. As shown, the mixer products are seriously attenuated by the simple, low-pass filter + keeping the mixer RF port signal amplitude low. The power splitter provides the input for the mixer and also the main output for the 14 MHz VCO. The main VCO output requires 50 Ω buffer/amplifer(s) to drive a receiver mixer, transmitter chain, or whatever.

I inserted the 12 dB attenuator pad to keep my mixer RF port signal low to drop the mixer products amplitude down; further losses occur in the transformer. You can change this pad to whatever is required. I belong to the camp of builders who drive their transmit mixers with low-level RF signals to avoid messy outputs at the IF port.

Click for a breadboard photo of my initial bench tests with the mixer board. A 50 Ω resistive terminator shunts the main VCO output port during this testing. I temporarily insert BNC connectors along my development breadboards to measure output signals with my 50 Ω terminated 'scope, spectrum analyzer, or power meters and *rarely* measure RF circuits with a 10X probe.

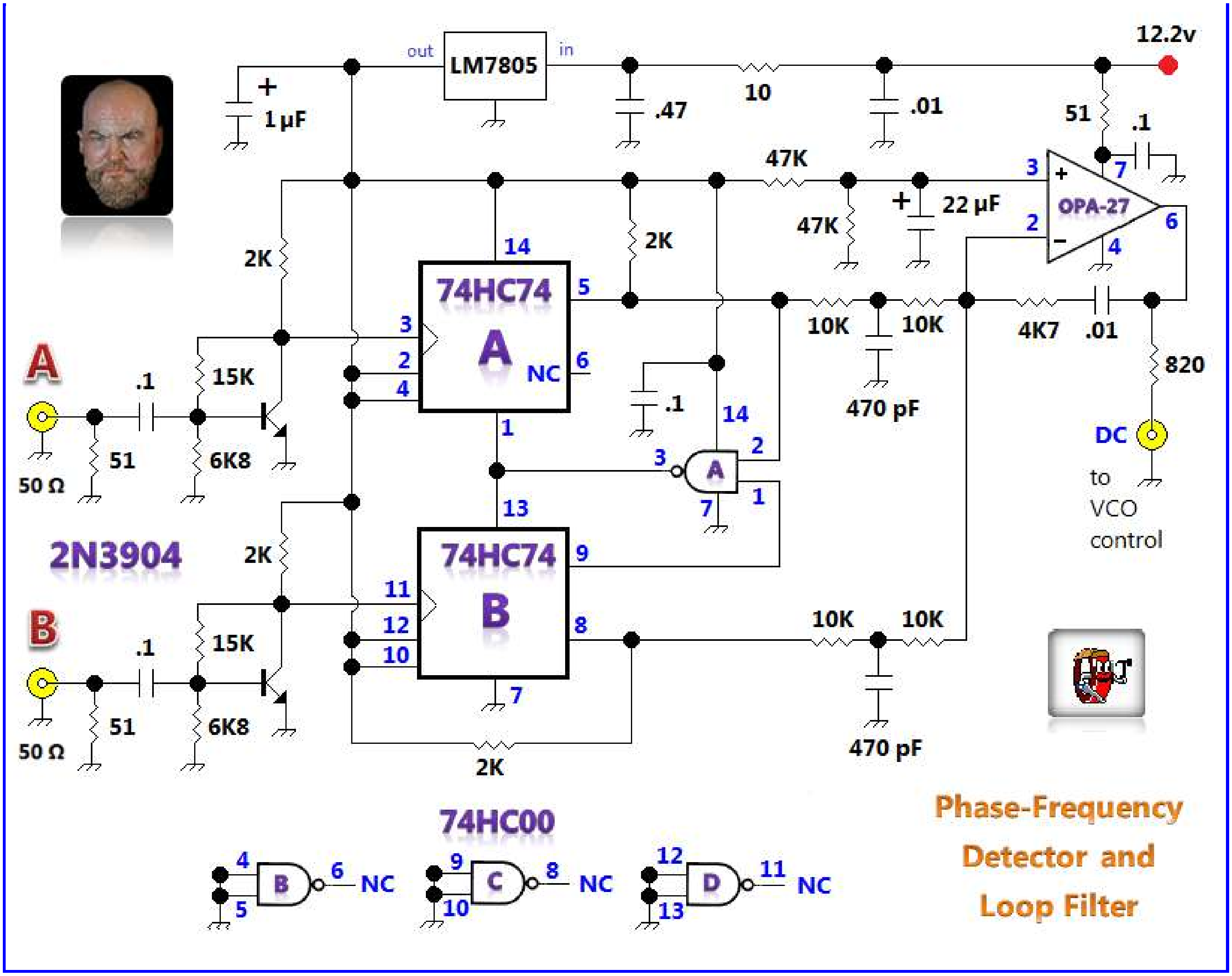
12.2v
LM7805
out
in
1 µF
.47
10
.01
51
.1
47K
47K
22 µF
OPA-27
3
2
7
4
6
2K
14
74HC74
A
5
6
NC
3
2
4
1
13
2K
15K
6K8
51
.1
50 Ω
A
B
2N3904
.1
14
3
A
2
1
7
10K
10K
470 pF
4K7
.01
820
DC
to
VCO
control
74HC74
B
9
11
12
10
8
7
10K
10K
470 pF
2K
74HC00
4
5
B
6
NC
9
10
C
8
NC
12
13
D
11
NC
Phase-Frequency
Detector and
Loop Filter

Above — Phase-frequency detector and loop filter schematic. 51 Ω resistors terminate each 1.5 MHz input and drive two 2N3904 switches per EMRFD Figure 4.41. The loop filter design from EMRFD works as described, however, if you make a loop filter for a different circuit, casual copying goes out the window. Engineers design their loop filters according to factors including the crossover frequency, VCO gain, the N-division for the loop, etcetera *with software.* Some people and companies offer such software on the Web.

My loop filter 0.01 µF cap was a 1% polyester capacitor, although Wes specs a 10% tolerance in EMRFD. No cheapo ceramic bypass caps here please.

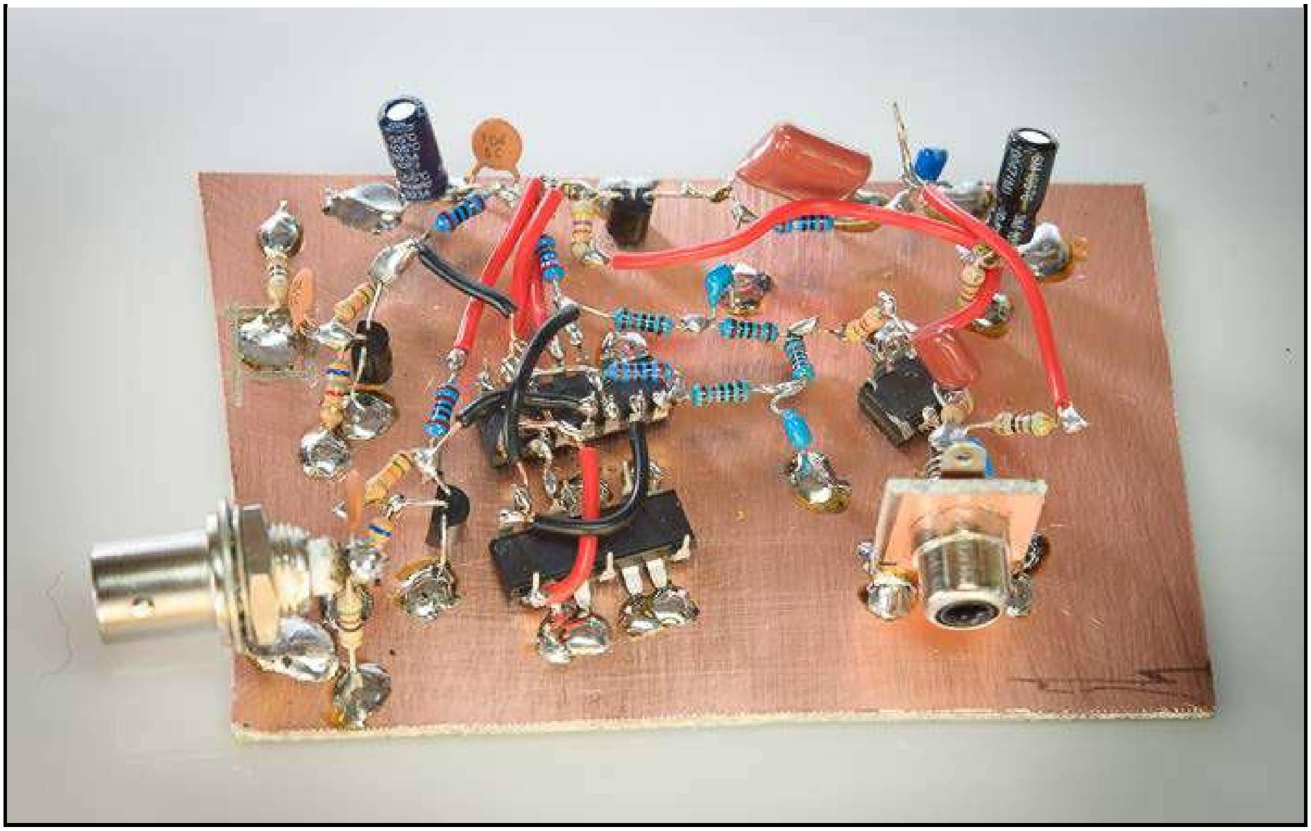

Above — Phase-frequency detector and loop filter breadboard. Click for a photo of the scattered, ugly, working boards on my workbench. Many prototypes look like this on our benches, however, sometimes, they work perfectly until we stick them in a box! Do you relate? Each oscillator belongs in its own metal box with strong bypass and decoupling networks (feedthrough caps reign supreme here) since the 3 oscillators might decide to party together and create havoc.

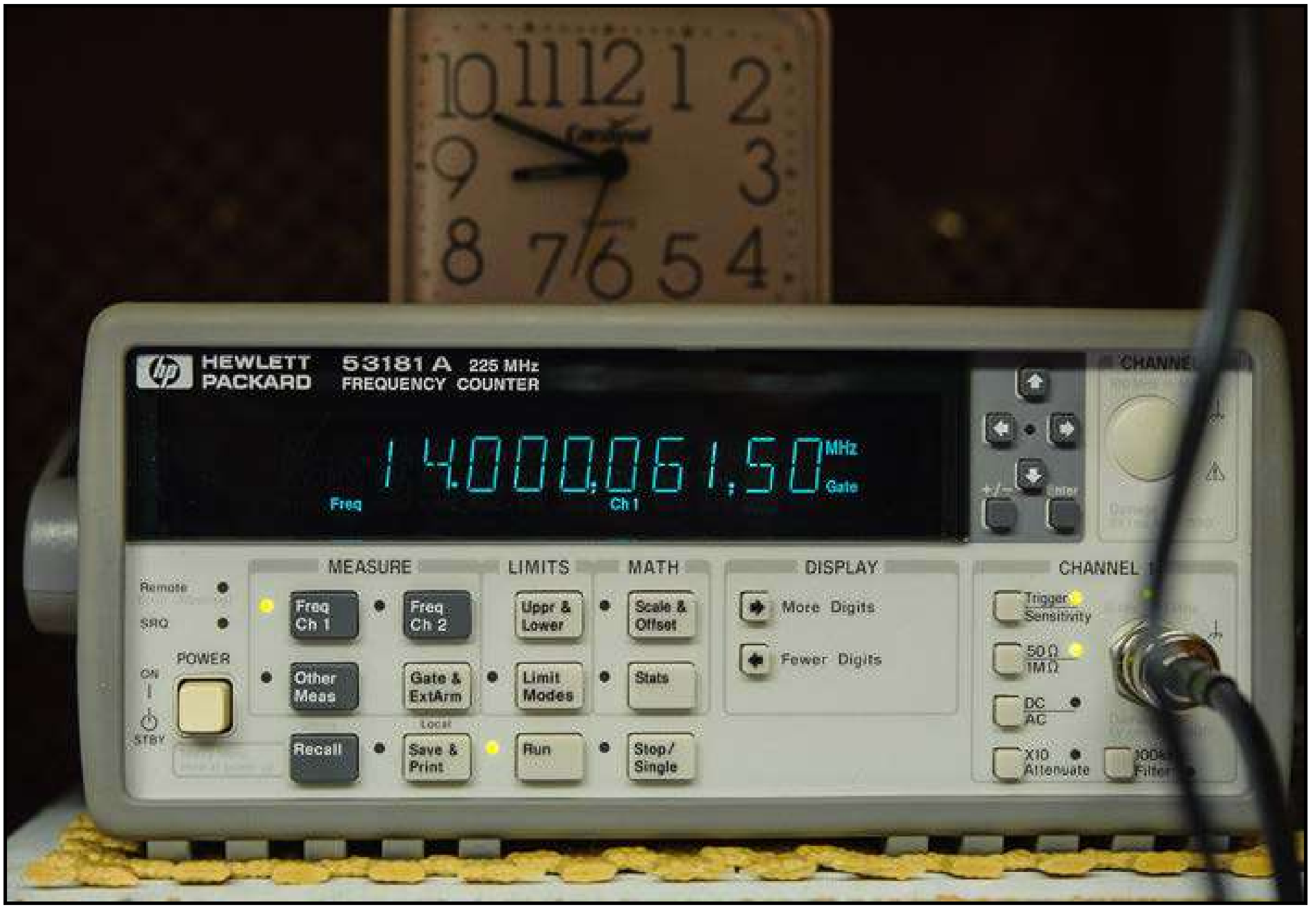
HEWLETT PACKARD
53181A 225 MHz FREQUENCY COUNTER
14.000.061,50 MHz
Freq
Ch 1
Gate
MEASURE
LIMITS
MATH
DISPLAY
CHANNEL 1
Freq Ch 1
Freq Ch 2
Uppr & Lower
Scale & Offset
More Digits
Fewer Digits
Other Meas
Gate & ExtArm
Limit Modes
Stats
Recall
Save & Print
Run
Stop/ Single
Remote
SRQ
POWER
ON
STBY
Local
Trigger Sensitivity
50 Ω 1M Ω
DC AC
X10 Attenuate
100kHz Filter
Enter

Above — My VCO frequency with the 1.5 MHz MTO set at full mesh. Since my MTO only tuned from 1.50 to 1.52 MHz, my VCO only tuned from

14.00 to 14.020 MHz, but that's easily fixed as I've stated. I'm very happy — it locked perfectly and my 14 MHz VCO stayed on frequency at the exact previously measured frequency drift of the 1.5 MHz MTO. When I connected the 14 MHz VCO to my counter without the PLL circuitry, it drifted willy-nilly.

Conclusion

The sense of awe and joy arising from locking a VCO on frequency won't be understood by many. The concepts and circuitry offers many possibilities. If the MTO and VCO exhibit low phase noise, short-term oscillator stability may be fantastic.

The 14 MHz VCO could be a 56 MHz VCO with sequential division by flip-flops to provide output at 28, 14 and 7 MHz with the 14 MHz portion going to the offset mixer. In EMRFD, Figure 4.44, Wes offers 14 and 7 MHz output by dividing the 14 MHz signal from the main power divider output port. The 7 MHz band is low-pass filtered to remove harmonic energy.

Wes extended this circuit by dividing the MTO by a hardware programmable frequency divider so that the difference from the mixer and low-pass filter is 170 kHz nominal. He uses this 'Almost Synthesizer' on the air for his QRP adventures.

While most builders will sensibly jump from an L-C VFO to a kit containing a programmed microcontroller plus a DDS or Si579, it's also fun to play with hardware to learn and ingrain synthesizer concepts + gain bench wisdom.

QRP — PosData for April 17, 2013

For a good read on the offset PLL, consider studying Wes' book **Introduction to Radio Frequency Design**, ARRL, 1994, page 320 and on. *This book is now out of print*. Wes ported the PLL active loop filter design program he wrote for IRFD from DOS to Windows in April 2013. Click and scroll for it. Thanks for this Wes!

---

5. Boot-strapped Popcorn AF Feedback Pair

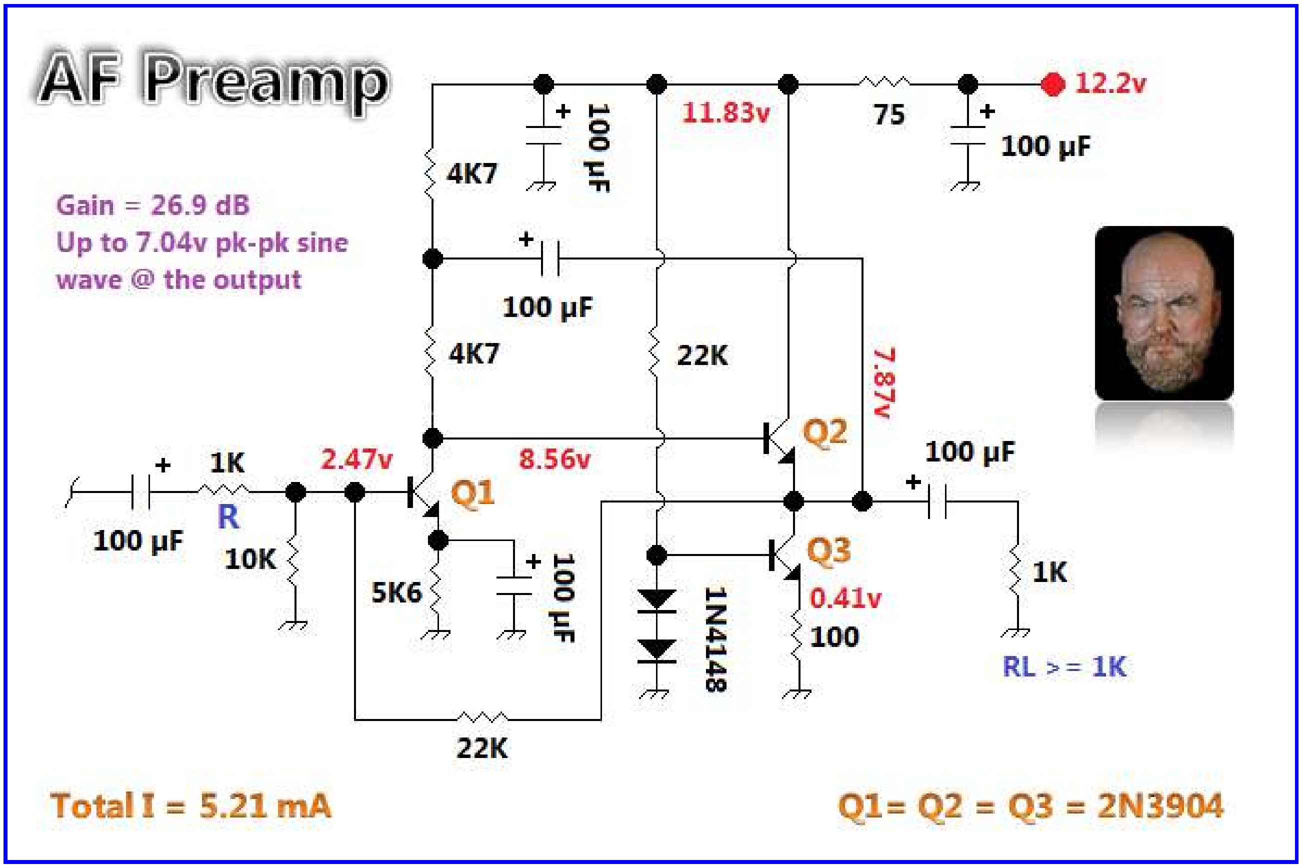

Above — I designed this AF stage for a builder from Indonesia; a popcorn AF shunt feedback amp based on the work of Douglass Self. Despite only drawing ~ 5 mA, this amp stayed clean until the output voltage exceeded 7.04 volts peak-peak on my test bench. Boot-strapping increases gain and lowers distortion in Q1.

Q2 buffers the Q1 voltage amp from external loading and increases gain.The Q3 current source boosts the load-handling capacity of the Q2 emitter follower. The input R can be raised to reduce sensitivity. The 1K output R could be a 5-10 K volume pot.

---

6. Non-Mechanical Iambic Paddle

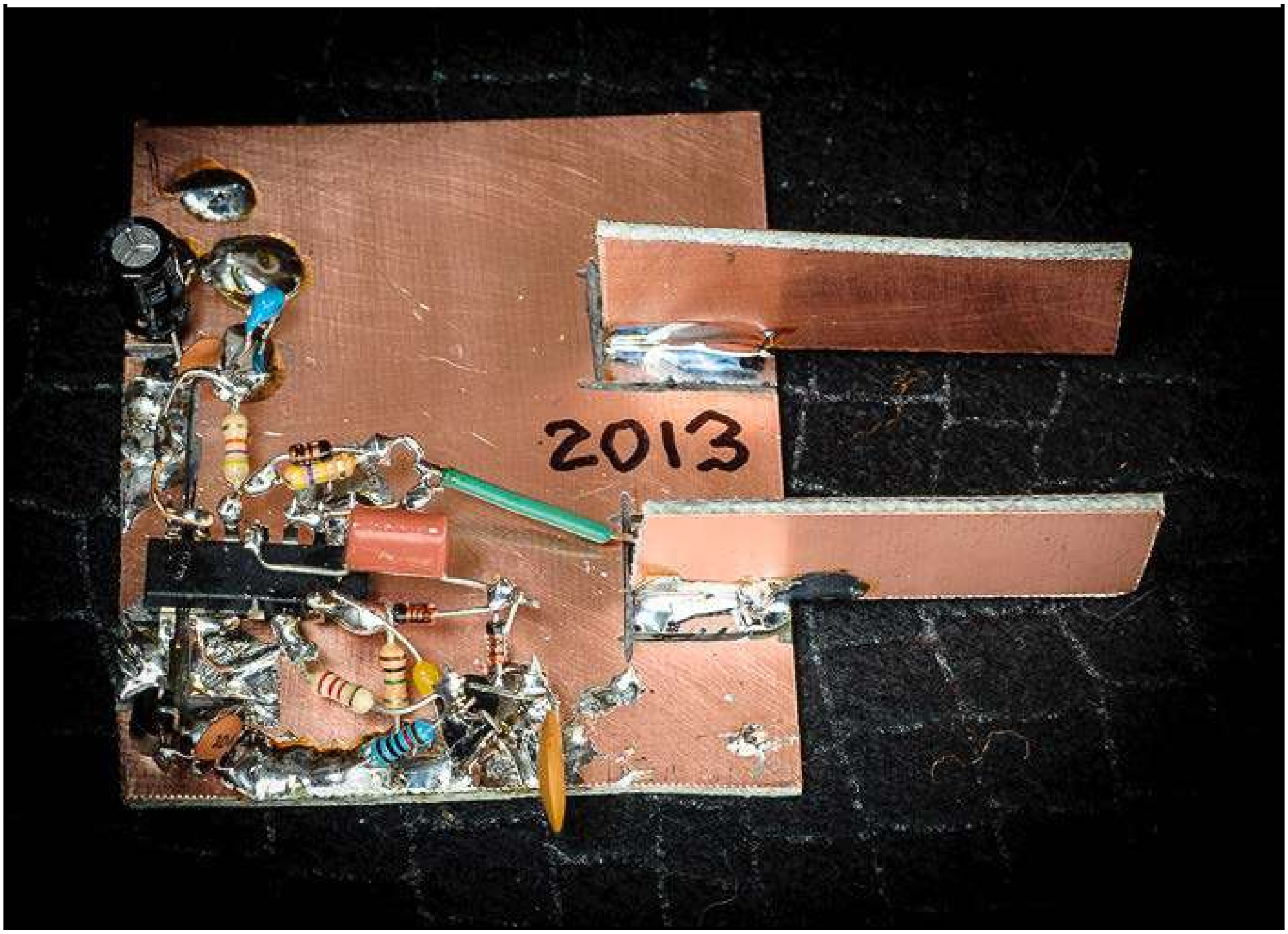

Above — The very ugly development proto-board of my half-done non-mechanical Iambic paddle. At some point I'll build the other half (the dah paddle switch) and press it into service. You might also use this circuit as a non-mechanical straight key.

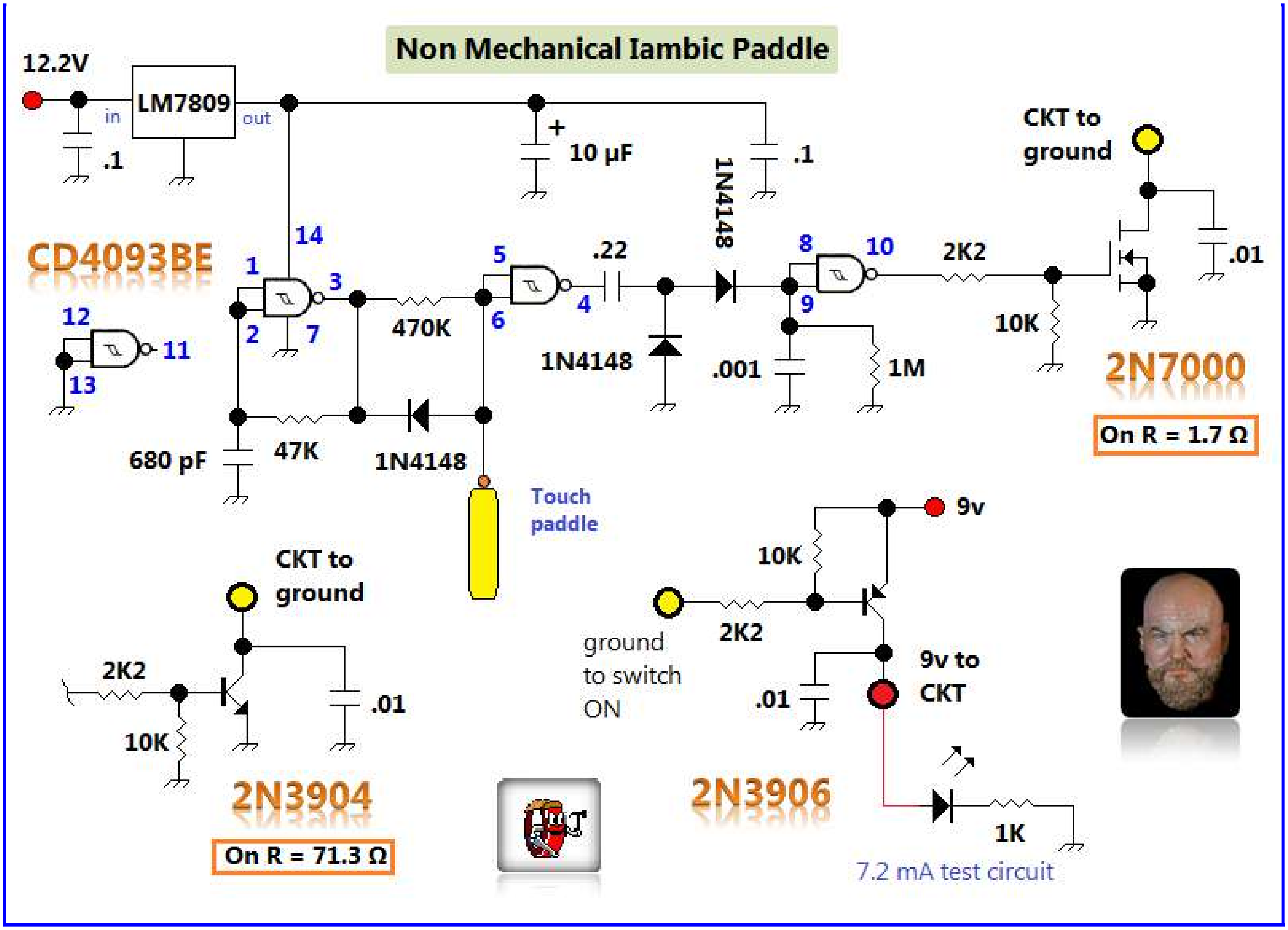
Non Mechanical Iambic Paddle
12.2V
in
LM7809
out
.1
+
10 µF
.1
1N4148
CKT to ground
CD4093BE
14
1
3
2
7
12
11
13
5
6
4
.22
470K
8
10
9
2K2
10K
.01
2N7000
On R = 1.7 Ω
1N4148
.001
1M
680 pF
47K
1N4148
Touch paddle
9v
10K
CKT to ground
2K2
10K
.01
2N3904
On R = 71.3 Ω
ground to switch ON
2K2
9v to CKT
.01
2N3906
1K
7.2 mA test circuit

Above — The schematic for 1 of the paddle circuits. I compared the ON resistance of the BJT switch with the enhancement mode FET and the 2N7000 won: only the FET could key the continuity tester on my DVM.

You may extend this circuit with a 2N3906 switch for paddle-switched 9 volts (or whatever VCC you want). In the bottom right, I connected the PNP collector to an LED and flashed it for fun. The 0.01 µF capacitor on the switch drain or collector bypasses any RF to ground. With higher power RF, you may have to place a similar bypass cap in parallel with the shunt 10K resistor on the 2N3904/2N7000 base and gate respectively.

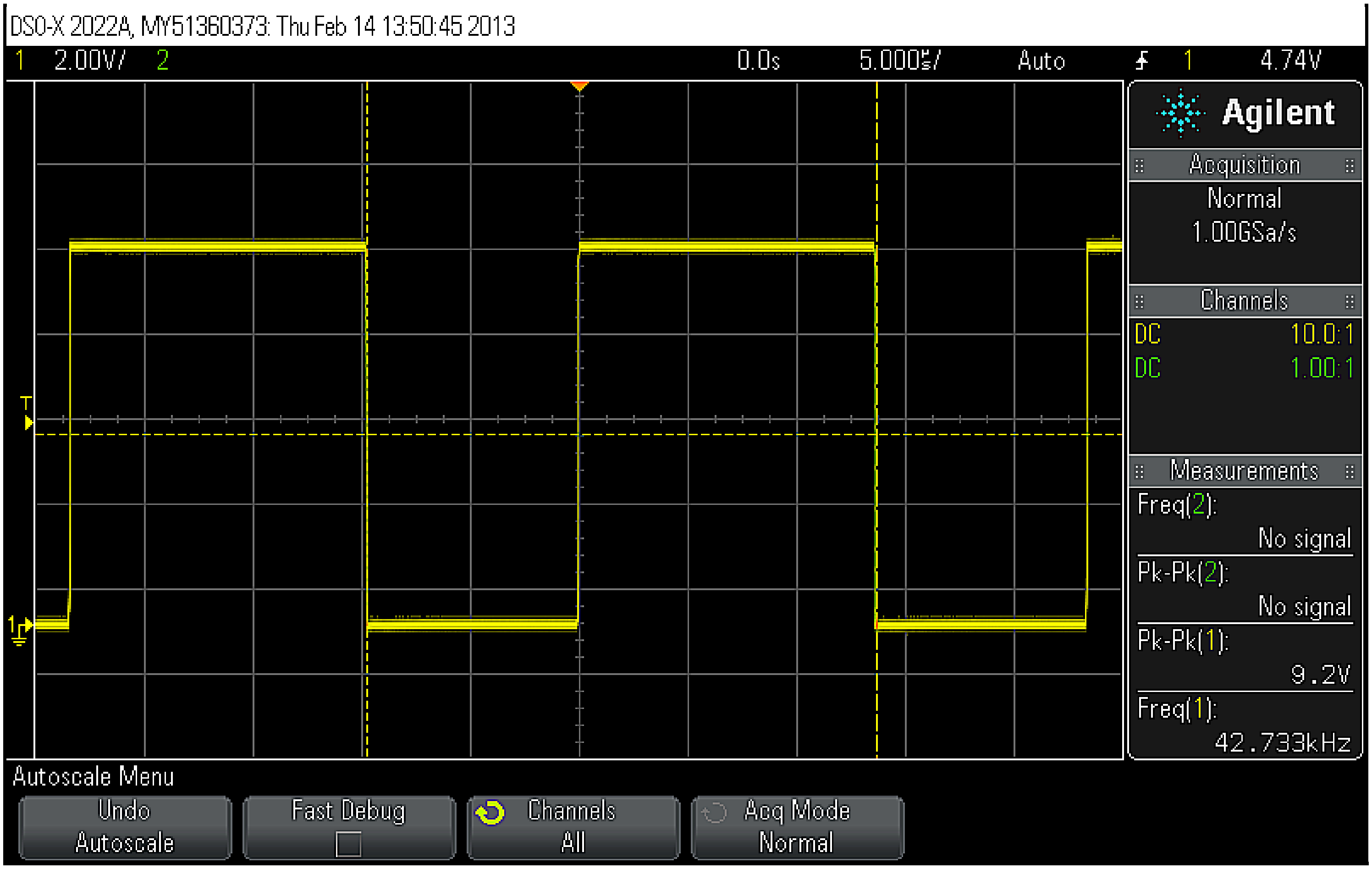

DSO-X 2022A, MY51360373: Thu Feb 14 13:50:45 2013

1 2.00V/ 2 0.0s 5.000µs/ Auto 1 4.74V

Acquisition
Normal
1.00GSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(2): No signal
Pk-Pk(2): No signal
Pk-Pk(1): 9.2V
Freq(1): 42.733kHz

Agilent
Autoscale Menu
Undo Autoscale
Fast Debug ☐
Channels All
Acq Mode Normal
T
1

Above — A 'scope shot of the ~ 43 KHz oscillator generated in the first 4093 Schmitt trigger NAND gate.

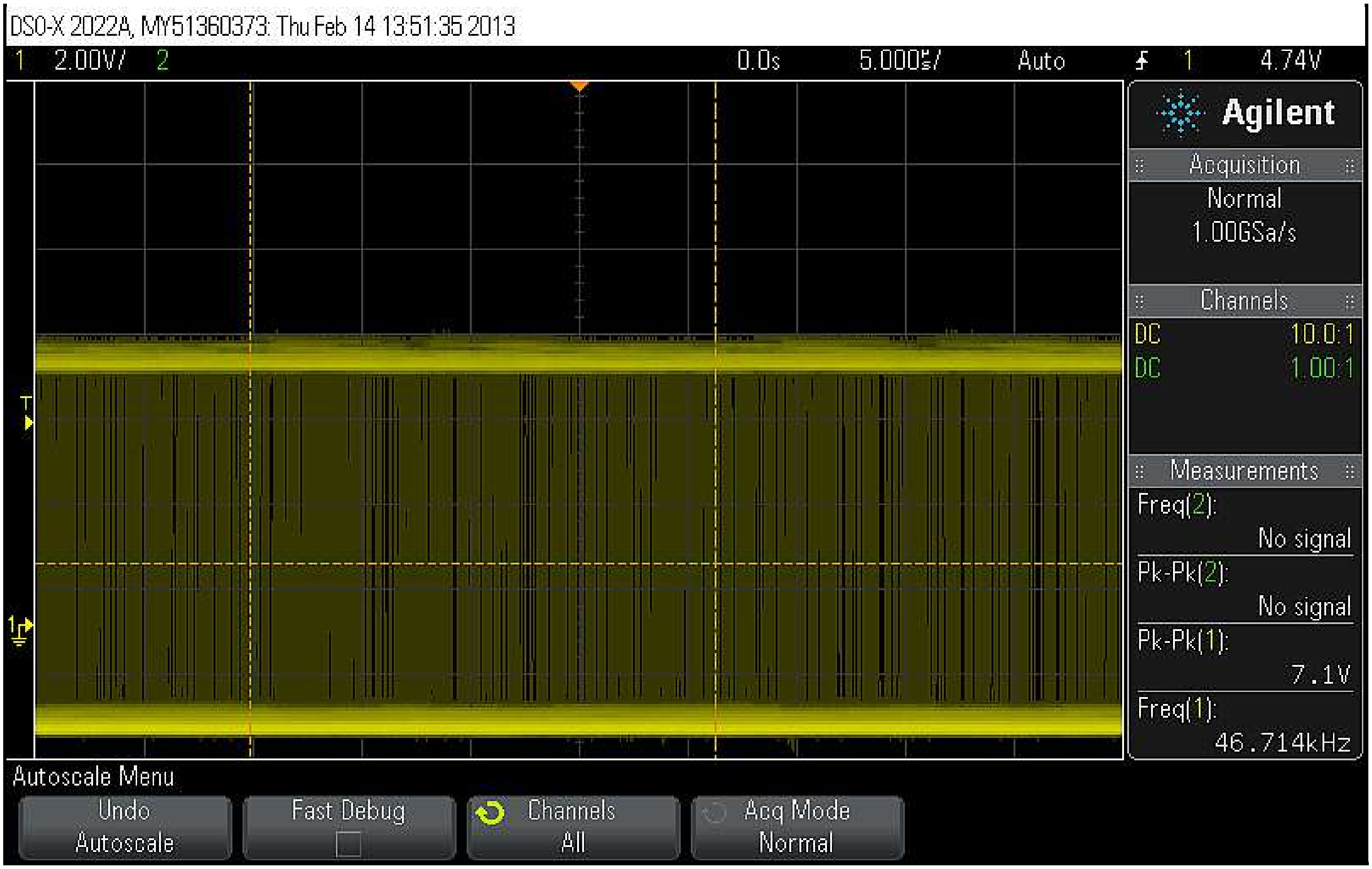
DSO-X 2022A, MY51360373: Thu Feb 14 13:51:35 2013
1 2.00V/ 2
0.0s 5.000S/ Auto 1 4.74V
Agilent
Acquisition
Normal
1.00GSa/s
Channels
DC 10.0:1
DC 1.00:1
Measurements
Freq(2):
No signal
Pk-Pk(2):
No signal
Pk-Pk(1):
7.1V
Freq(1):
46.714kHz
T
1
Autoscale Menu
Undo Autoscale
Fast Debug
Channels All
Acq Mode Normal

Above — Here's the disturbed oscillator output *just* after the paddle is touched: this stops the signal at pin 5 and 6 of gate 2 and kills the AC output at pin 4. The DC voltage across the 0.001 μF cap discharges through the 1 megohm resistor pushing pin 10 HIGH to turn on the 2N7000 (or the 2N3904). Normally, pin 10 is LOW since the rectified output of the undisturbed ~43 KHz oscillator goes to both pin 8 and 9.

A fun circuit for a Saturday afternoon...

---

7. A Journey Above HF

This project began as a 14 MHz low-noise amplifier build, but ended up with me learning more about SMT breadboard techniques and suppressing spurs. A short exploratory/descriptive account of my bench journey plus some photos follow.

VAT-6
6 dB ATTENUATOR
50 OHM

Above — I'm slowly adding SMA connectors and pieces. Since modern consumer digital network engineers use them, they're abundant and often rated from DC to 18 GHz; more bandwidth than I'll ever need.

I'm also building with evermore SMT components and just love it. Through-hole (I prefer to say hole-through) stuff continues to disappear like lemonade on a hot August afternoon.

Above — The schematic of my version of Victor 4Z4ME's feedback amp (FBA) as tested at 14 MHz: he emailed me a paper and provided some online support. Click for another version from December 2012. Typical noiseless FBAs suffer from poor reverse isolation, however, Victor runs the collector to base feedback through an asymmetrical 3 dB power combiner/splitter that boosts port isolation, defines the gain, plus sets the input and output impedance.

Strong virtues of the asymmetrical power splitter — fully utilized in this design include a very low loss on the input side (the 1 turn side) and a much higher loss ont the feedback side that allows the feedback to defines the input impedance on one side while exerting a negligible impact on noise figure and dynamic range on the other side. Victor measured a noise figure of 1.5 dB using a MRF586 BJT. Thanks to Victor for the information and design.

For strong IMD properties, I ran 50.6 mA total stage current into a gorgeous, low-noise, NE46134 NPN transistor with a fT of 5.5 GHz. Using VHF-UHF techniques, I built with mostly SMT parts on 2–sided board using copper wire vias to connect the 2 copper surfaces. I discussed the wrap-around bias technique in 2011 as number 1.

Above — My prototype breadboard with dremel cut islands for soldering the size 1206 or 0805 SMT parts, plus a few hole-through items. Carving an island for the SOT89 transistor package proved difficult, but even I (a challenged dremelist) did it.

Woe to Oscillations: Like misplaced car keys, oscillations may remain hidden unless you search for them. Often, the only difference between a proper oscillator and a regular amplifier is we *want* the former to oscillate. To check for instability, we might use our high bandwidth scope, or a spectrum analyzer, but many will have to find spurious RF with basic, DC - HF bandwidth test equipment. In any case, just do your best. To some extent, unwanted oscillations are the *elephant in the room* that few talk about. Well, it's okay to think, talk and feel some emotions about them.

Sure enough, when I connected a 14 MHz signal to the amp's input and a 50 Ω terminated 'scope to the output to measure gain, my sine wave had 2 or 3 others on top of it. In the 4Z4ME amp, the PNP bias transistor can be a source of AF to HF oscillations.

Victor wrote: "The circuit has a low frequency amplifying loop that goes through both transistors. The PNP transistor does not invert the signal (it is a common base amplifier) and the RF transistor inverts so it is a loop with 180 degrees phase shift (negative feedback). The various decoupling and RF coupling capacitors in this loop add phase shift on this low frequency loop. If the accumulated phase shift adds to an additional 180 degrees and gain is larger than 1 you have oscillations. The simplest way to solve it is to make one of the capacitors very large so it will add only 90 degrees phase shift but it will drop the gain at the higher frequencies where the other capacitors start to add phase shift to be less than 1 so there are no conditions for low frequency oscillations. This technique is called "Dominant Pole". That's the reason that I suggested to connect a very large capacitor to the PNP transistor".

I found my oscillations disappeared with a 0.1 μF collector bypass cap on the PNP (Cx on my schematic). The 0.1 μF cap on the PNP collector was critical – a 0.22 μF failed to work, as did a .001 μF --- but a 0.1 μF held it stable. In another 2N2222a-based 4Z4ME amp with 0.01 μF input and output caps, it took a 10 μF capacitor on the PNP collector to snuff out some ~766 Hz oscillations.

We don't use a wrap-around PNP bias with our RF oscillators — that's asking for trouble.

I aso measured oscillations at ~ 372 MHz with my spectrum analyzer. A collector 10 Ω R killed these UHF oscillations and after that I saw no spurs from .001 to 1 GHz. (I should have made the dremel cut right close to the NPN collector for the 10 Ω resistor. I hoped there were no oscillations above 1 GHz because I can't measure them.

Finding oscillations:

Many builders lack a spectrum analyzer, let alone 1 that goes up into UHF bandwidth. I'll share a few tips I've learned on the bench that don't require expensive test gear:

A 10X scope probe on the drain or collector of an amp may sometimes reveal oscillations up to the maximum 'scope bandwidth — set your 'scope vertical scale for high sensitivity. This provides direct measurment of oscillations.

Indirect methods to infer unwanted oscillations also lie in our armatorium. I learned this trick from Wes: Place the circuit under test in your normal gain measurement set up with an oscilloscope. Then vary the DC power supply voltage slowly and smoothly — your measured 'scope voltage changes should also track slowly and smoothly. You may see an AC voltage jump as the amplifier goes into and out of oscillation with the DC power supply tweaking. After finding this oscillation caused AC voltage spike, you work to remedy it with a variety of means such as better bypassing, changing bias voltages, shielding and locating breadboard errors.

Sometimes if you put your finger near the active part while watching the bias voltage or current you may see the bias jump around *if* oscillations are present.

My final indirect oscillation busting technique: If you measure the specified/expected gain and return loss on the input/output port, this *may* signal your device is stable — I've noticed this with MMICs where I saw oscillations on my SA, stabilized them and *only then*, measured the expected S21, S11 and S22.

Sometimes, eliminating a hot part proves the best fix! In 2012, a new builder wrote to say that he soldered in a Mini-Circuits DC - 6GHz MMIC; the ERA-2SM in SOT-86 as a buffer for his 3.5 MHz VFO. Anyhow, in the photo were long leads plus no decoupling resistors etc. It sounds like the circuit behaved hyperreactively and *vibrat*ed in spasm. The cure was to eliminate the microwave part and put in a hycas amp built with a J310 + a 2N3904 — we encounter risk when plying the latest, hottest, super-high fT amplifiers sold on eBay with casual abandon.

Practice makes perfect. if you believe learning is experiential and build to learn, you'll learn to build.

Finally, as an amateur, I struggled to choose a SMT ferrite bead and after reviewing many datasheets and application notes I ordered a size 0805, 800 mA part with 120 ohms Z at 100 MHz and its peak impedance at 340 MHz. I'll let you know how that works out.

RF Bypass on our DC lines: As possible, we ought to provide a broadband RF bypass to provide a low impedance to RF from low frequency up to the maximum frequency wherever our FET, BJT, MMIC, etc. operates. For example, you can't just swap a higher gain BF998 (1 GHz) for a 40673 (VHF) dual-gate MOSFET and expect the same stability and bypass requirements can you? At the very least, I bypass G2 of the BF998 with a size 0805 0.01 or .001 µF SMT capacitor and the drain with RF bypass good for 1 GHz.

Wideband RF bypass may solve oscillation issues too.

I tried to apply a broadband bypass in my breadboard, although it gets extremely difficult to think about bypassing RF at > 1 GHz for the QRP homebuilder. Our hobby should include reflection and proper intention at the very least.

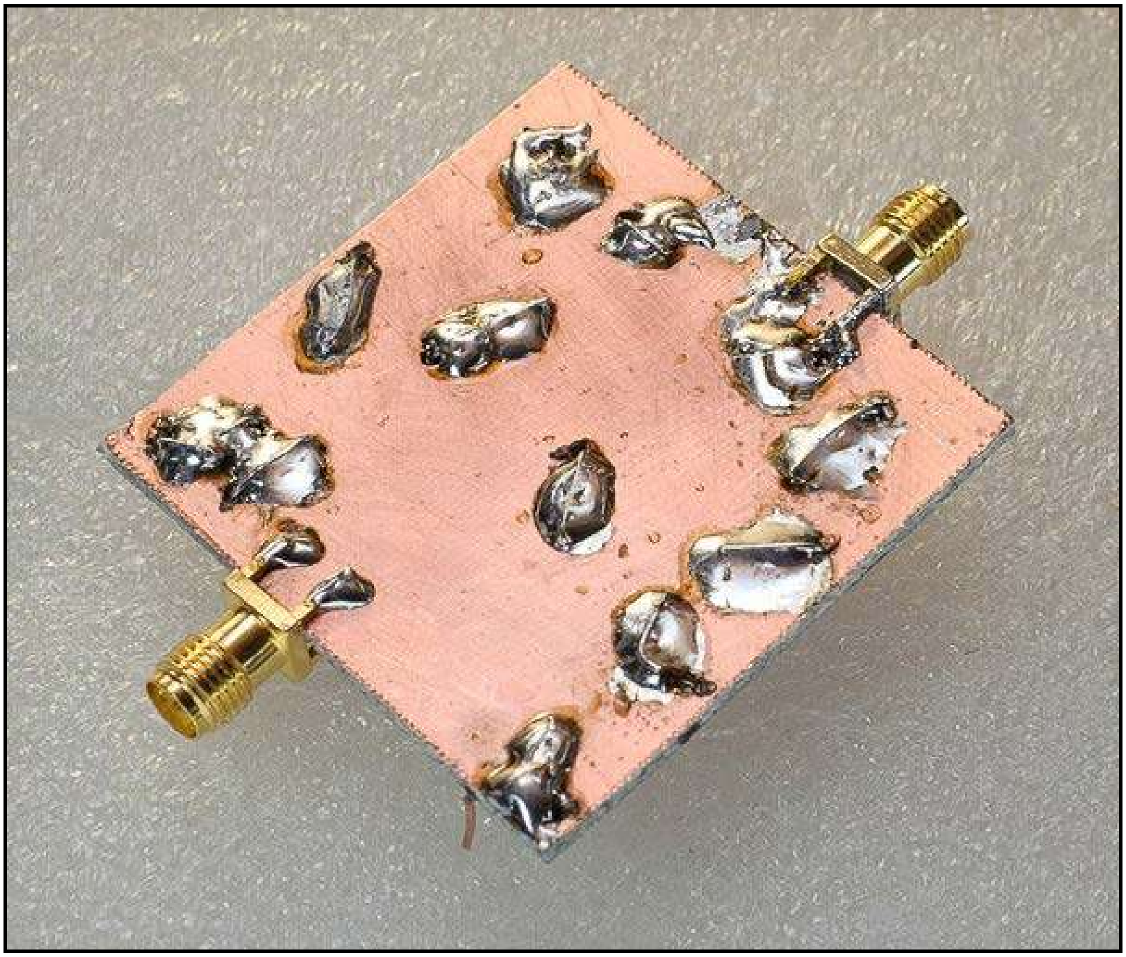

Above — A photo of the bottom of my breadboard showing the vias. I made mistakes: we should try to keep the via holes as close as possible to all bypass caps, my 10 Ω collector snubber resistor, collector port, or whatever we need to put at RF ground. The vias connect circuit areas to the large area, low impedance ground plane to minimize inductance. We should also try to place bypass capacitors as physically close to the pins of whatever we're bypassing.

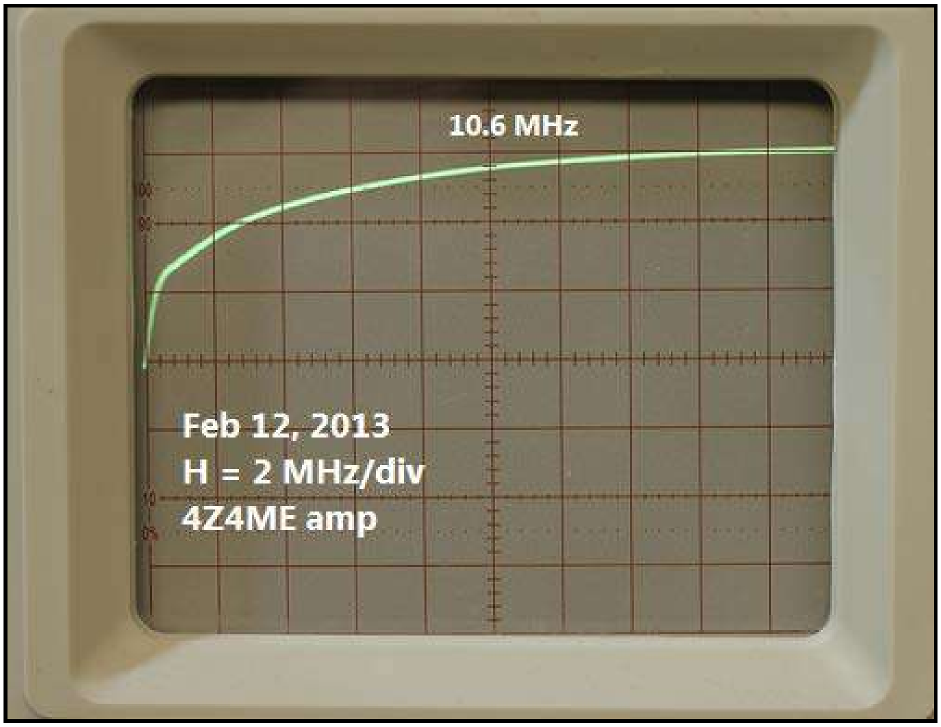

Above — A TG + SA sweep of the gain of my 4Z4ME NE46134 FBA from ~1 to 20.6 MHz. Each vertical quare denotes 10 dB. Each horizontal square = the value specified in the photo. In this case; 2 MHz per division. Maximum gain was ~16 dB.

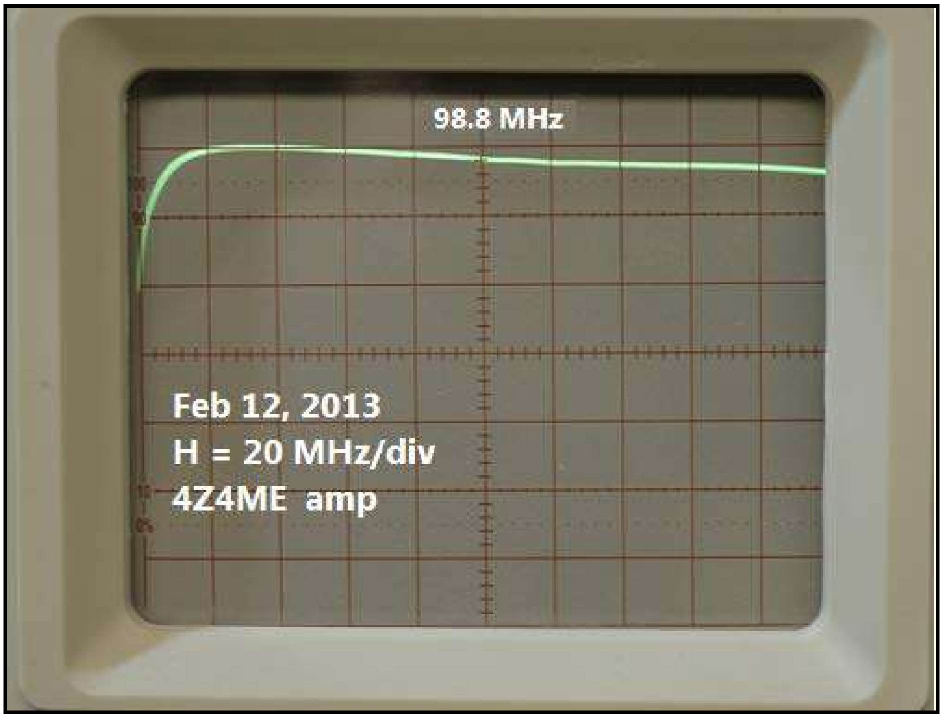

Above — A sweep from ~ 1 to 200 MHz. This would also make a good 6 meter band amplifer or ????.

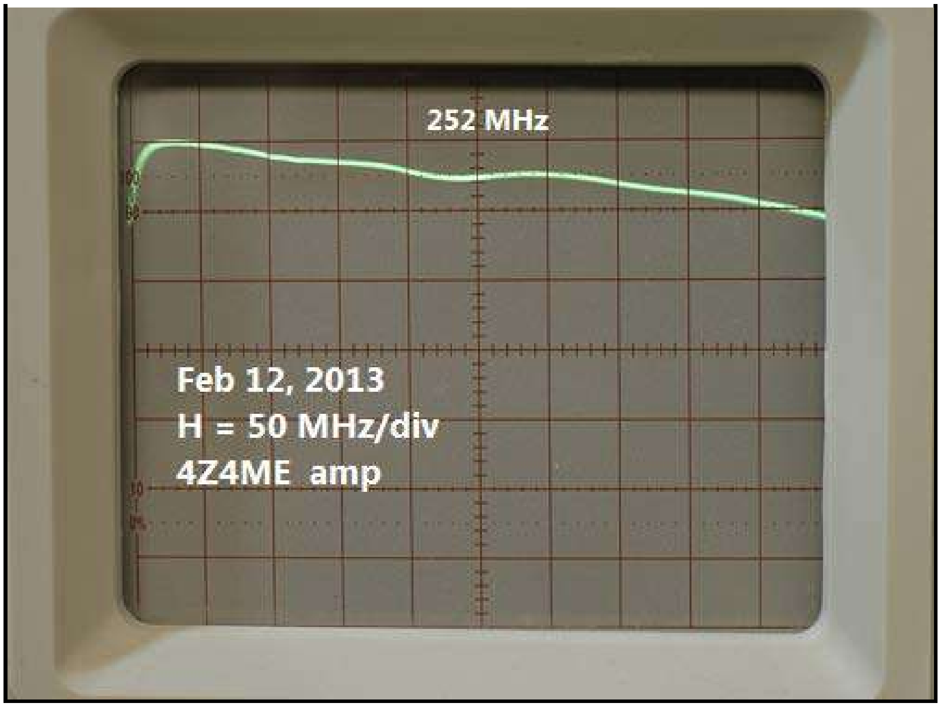

Above — A sweep from ~ 1 to 500 MHz.

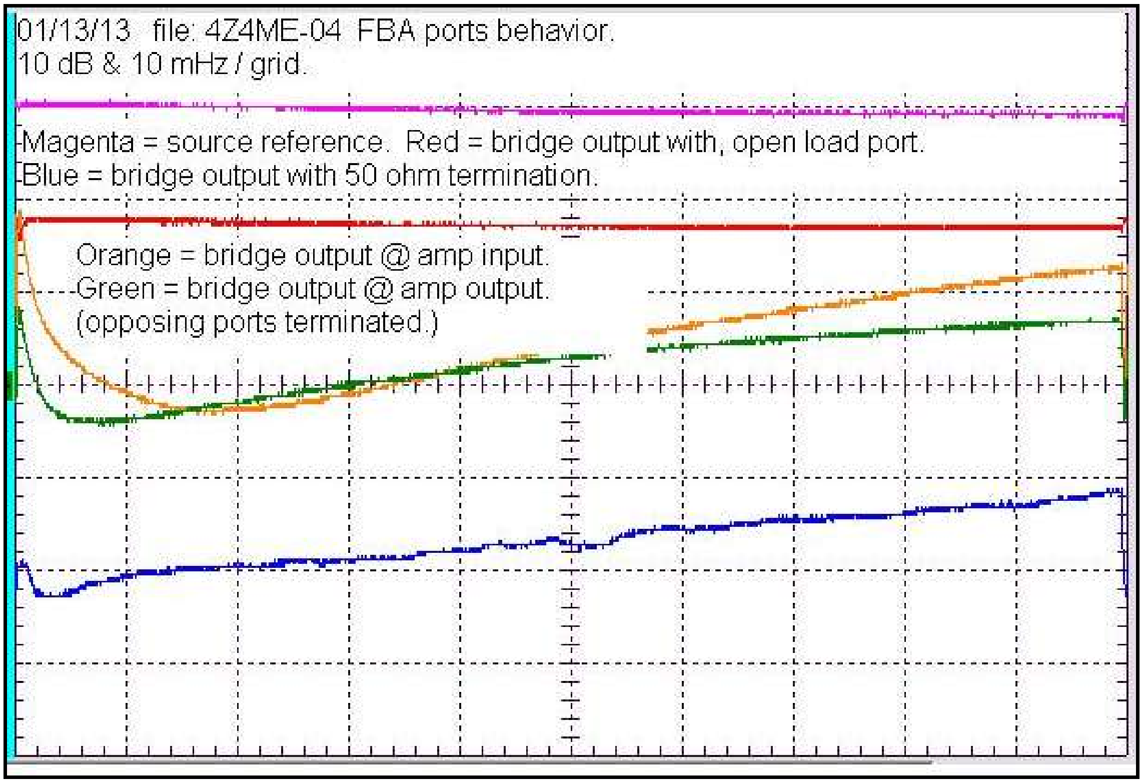

Above — Bob K3NHI made and swept a 2N5109 version of the 4Z4ME amp biased for ~47 mA emitter current. Here he swept return loss at the amp's input and output from about 1 to 100 MHz. The output return loss of my build was down, however, I didn't own many SMT resistors between 33 and 68 Ω. For example, if the output impedance at the collector is 10 Ω, then the series resistor should be 40 (39) Ω.

I've found that in my FBAs, changing the current and also the transistor type (2N5109 , 2N2222a etc.) also affects the input and output return loss. At HF, it's possible to measure RL with a simple bridge, so optimization is possible.

I learned a lot by building just 1 amplifer and discussing my findings by email with friends. Hopefully the next version I make will show improved understanding and skill.

---

8. Popcorn AF Amplifier for Receivers — Reprise

I've worked on a popcorn audio power amplifier (PA) since 2008 and offer my latest experiments. There's only so much you can do with a single-supply 12 volt AF power amp, but I enjoy improving my circuit.

My power measurement technique is shown as Figure 4 here. To enhance versatility, the following PA's may be coupled to whatever preamplifer you choose. In all cases, I drove the power amp stage with a 5532 op-amp voltage amplifer. The power followers were biased with a 2N3904 amplified diode (also called NPN shifter bias amplifier, or DC level shifter) rather than just a pair of series diodes, since this allows you to dial in just the right amount of bias as you watch the AC signal in your 'scope. I wrote a tutorial that explains how to bias complimentary-symmetry power followers in 2008: Click for the link.

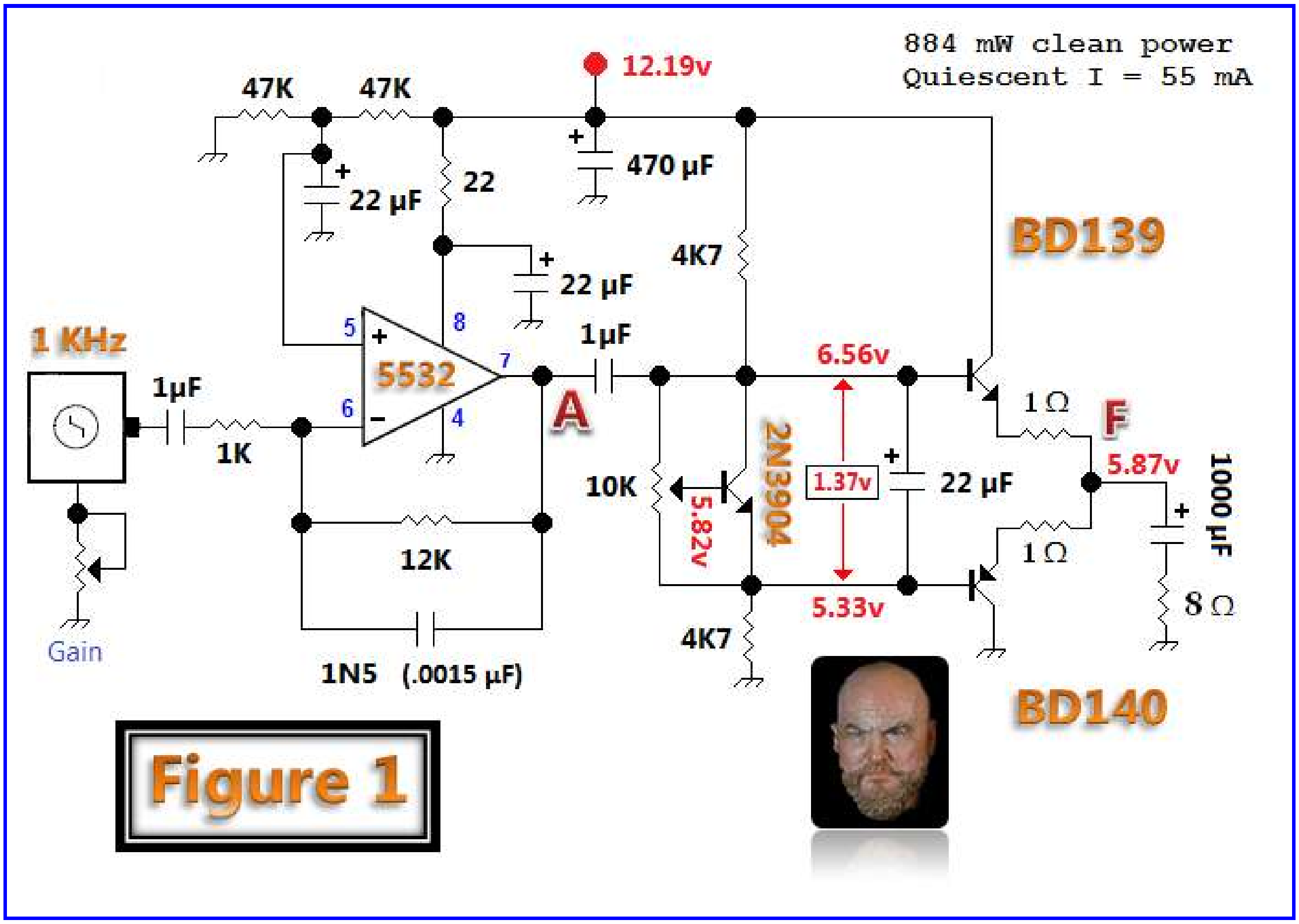
884 mW clean power
Quiescent I = 55 mA
12.19v
47K
47K
22 µF
22
470 µF
22 µF
4K7
BD139
1 KHz
5
8
5532
7
6
4
1µF
1µF
1K
A
6.56v
1Ω
F
5.87v
1000 µF
10K
5.82v
2N3904
1.37v
22 µF
1Ω
8 Ω
12K
5.33v
4K7
Gain
1N5 (.0015 µF)
BD140
Figure 1

Above — Figure 1: A popcorn AF power amplifer in full bench *test* mode. Measure the AC with a 10X 'scope probe across the 8 Ω resistor and the DC voltage and current with a multimeter. A distortion analyzer proves useful, but not essential for popcorn circuitry. I also listened to each amplifer connected to a line-level cassette player and an 8 Ω, 15 cm speaker. A 4 Ω speaker doubles the maximal clean power, but I don't own any and stuck to 8 Ω.

Containing no negative feedback, the power amplifer stage runs from the red-colored designator points **A** to **F**. You can AC or DC couple point F to your preamplifer stage as required to apply negative feedback.
As mentioned, you can use the 5532 preamp shown with any reasonable gain (i.e. change the 12K resistor), or opt to replace it with your own design. A low output impedance amplifer best drives the power stage.

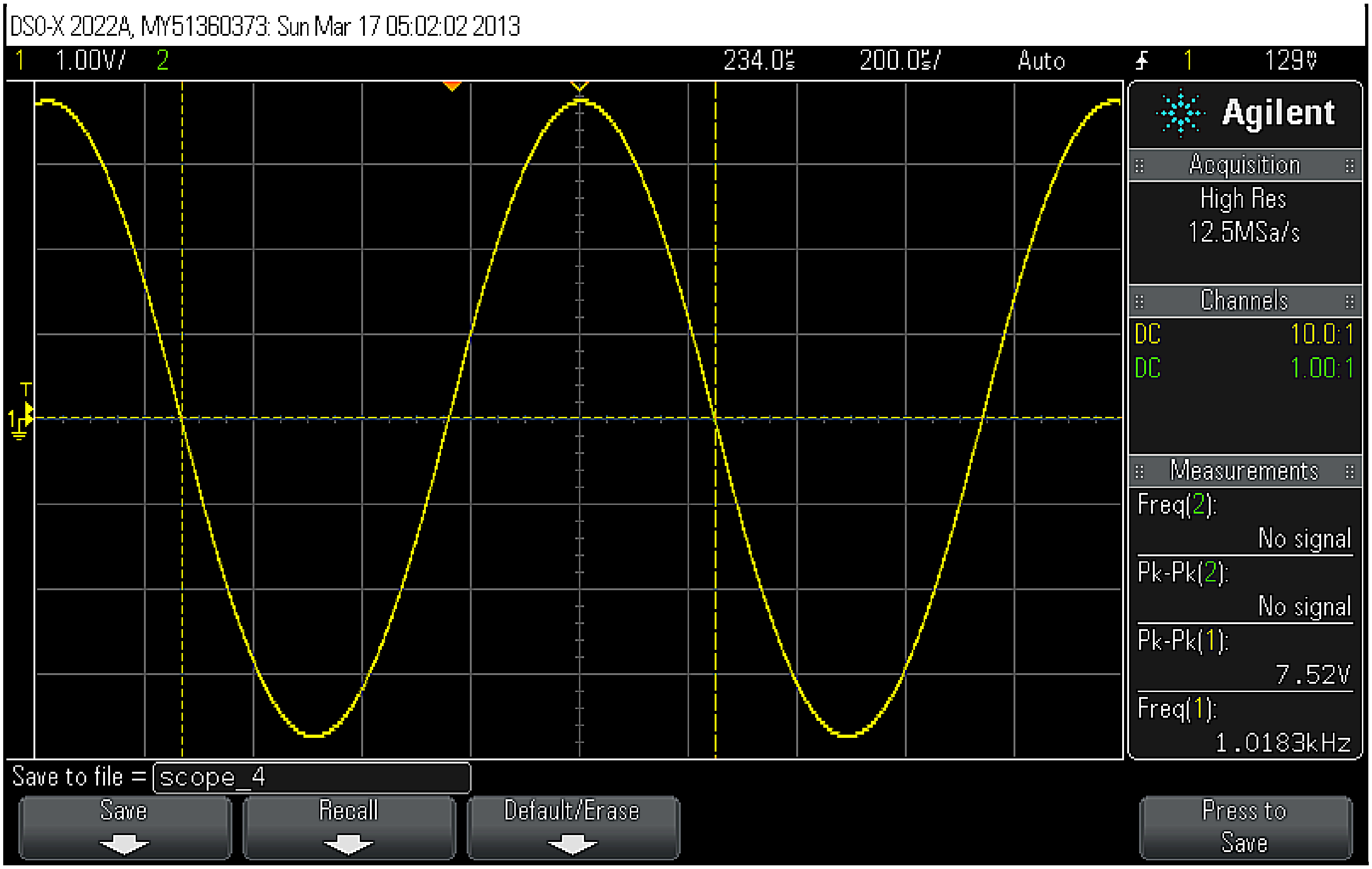

DSO-X 2022A, MY51360373: Sun Mar 17 05:02:02 2013

1 1.00V/ 2 234.0µs 200.0µs/ Auto 1 129mV

Acquisition

High Res

12.5MSa/s

Channels

DC 10.0:1

DC 1.00:1

Measurements

Freq(2): No signal

Pk-Pk(2): No signal

Pk-Pk(1): 7.52V

Freq(1): 1.0183kHz

Save to file = scope_4

Above — The output of Figure 1 in my 'scope driven to the maximal pk-pk voltage just before distortion begins to appear. Obviously this task is somewhat subjective, however, allows comparison of the amps you build on your bench.

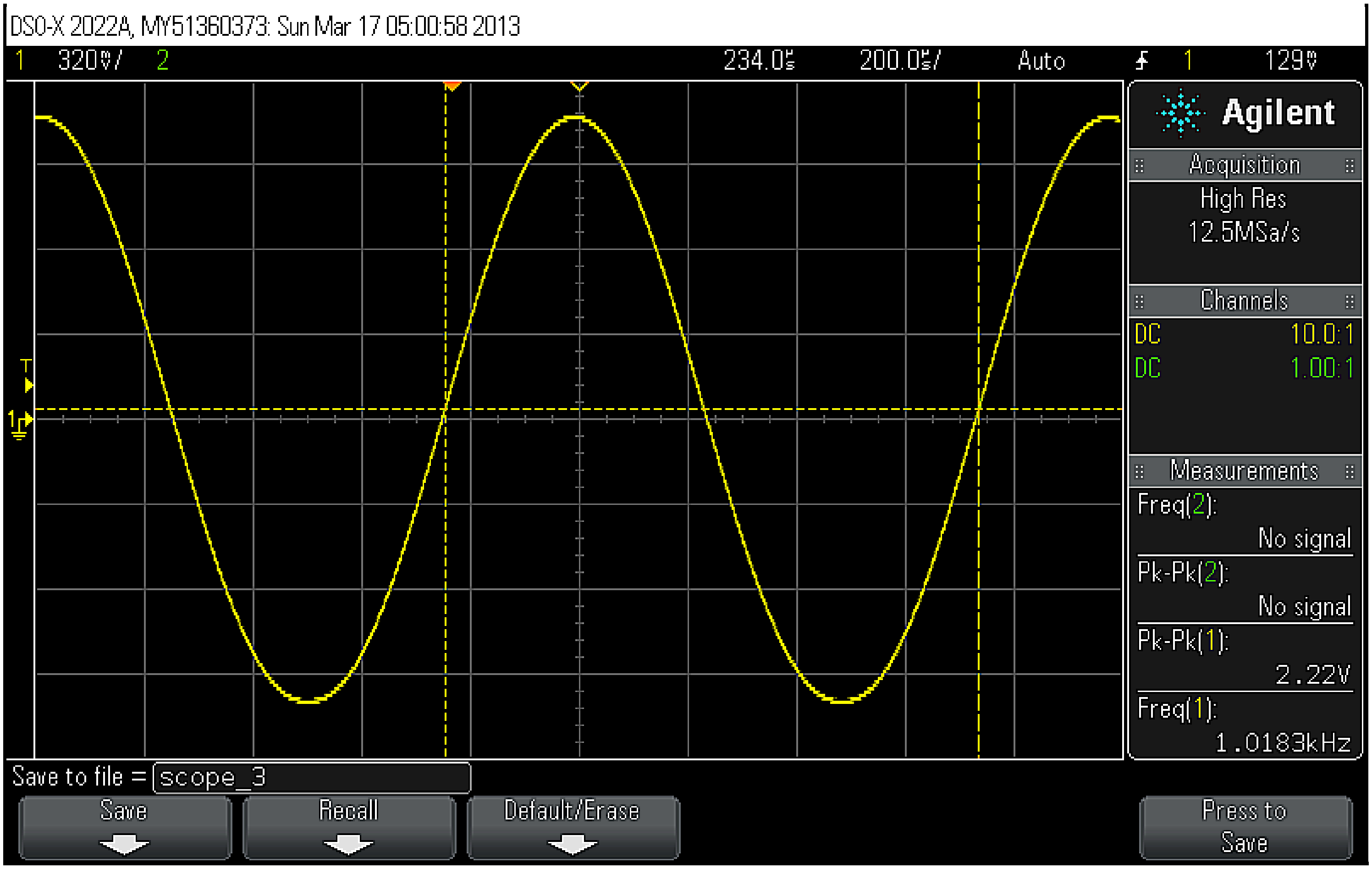

DSO-X 2022A, MY51360373: Sun Mar 17 05:00:58 2013

1 320mV/ 2 234.0µs 200.0µs/ Auto 1 129mV

Acquisition
High Res
12.5MSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(2): No signal
Pk-Pk(2): No signal
Pk-Pk(1): 2.22V
Freq(1): 1.0183kHz

Save to file = scope_3

Above — A 'scope screen capture with the 22 µF level-shifter filter capacitor from Figure 1 removed. Look what happened; the maximum clean signal fell from 7.52v pk-pk to 2.22v pk-pk. That capacitor is essential to get the maximal possible headroom.

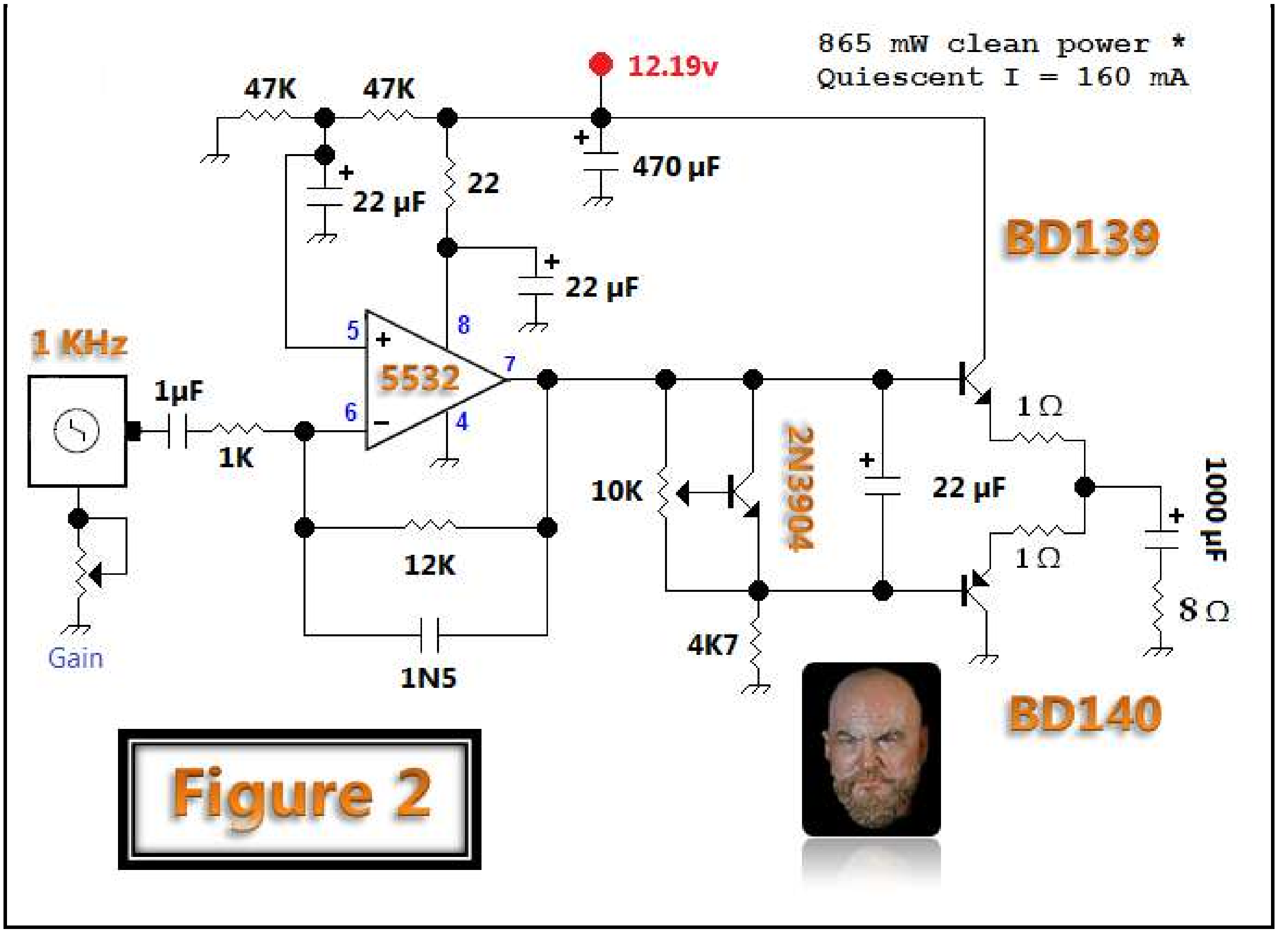
865 mW clean power *
Quiescent I = 160 mA
12.19v
47K
47K
22 µF
22
470 µF
BD139
22 µF
5
8
1 KHz
5532
7
6
4
1µF
1K
10K
2N3904
22 µF
1 Ω
1000 µF
1 Ω
8 Ω
12K
Gain
4K7
1N5
BD140
Figure 2

Above — Figure 2 is Figure 1 with the op-amp DC coupled to the level-shifter. I tested the circuit with and without the 4K7 resistor connecting the base of the 2N3904 to the DC supply: it didn't boost the amplifer headroom, nor reduced crossover distortion, so I removed that R.

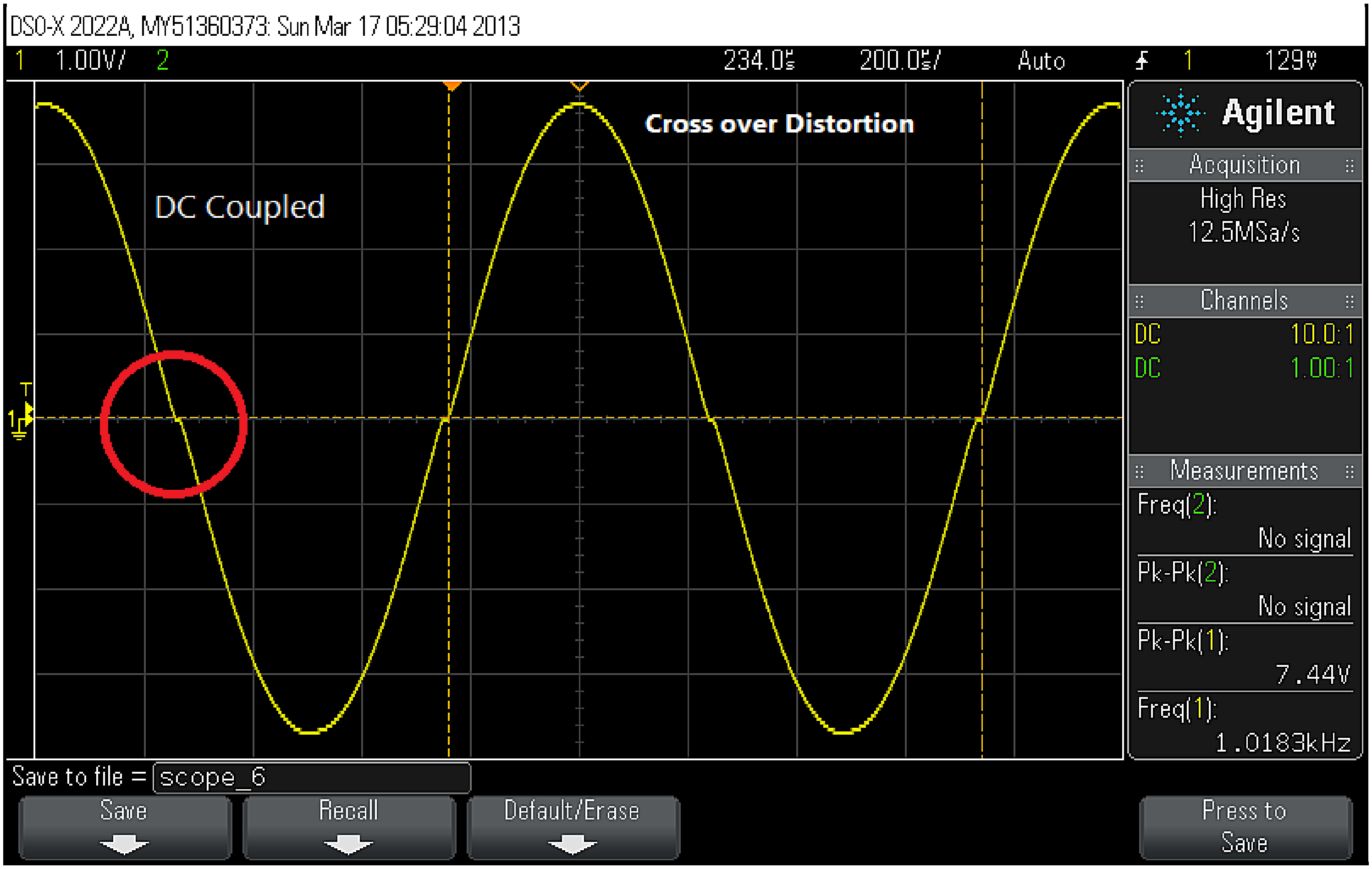
DC Coupled
Cross over Distortion
Agilent
Save
Recall
Default/Erase
Press to Save

DSO-X 2022A, MY51360373: Sun Mar 17 05:29:04 2013

1 1.00V/ 2 234.0µs 200.0µs/ Auto 1 129mV

Acquisition
High Res
12.5MSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(2): No signal
Pk-Pk(2): No signal
Pk-Pk(1): 7.44V
Freq(1): 1.0183kHz

Save to file = scope_6

Above — The Figure 2 amplifer 'scope tracing. At *maximum power*, crossover distortion appeared and I've seen this before. Likely, there is not enough base drive to keep the power followers forward biased. By adjusting the level shifter, I almost removed the crossover distortion, but never eliminated it. This drove the quiescent current up to 160 mA. Yikes!

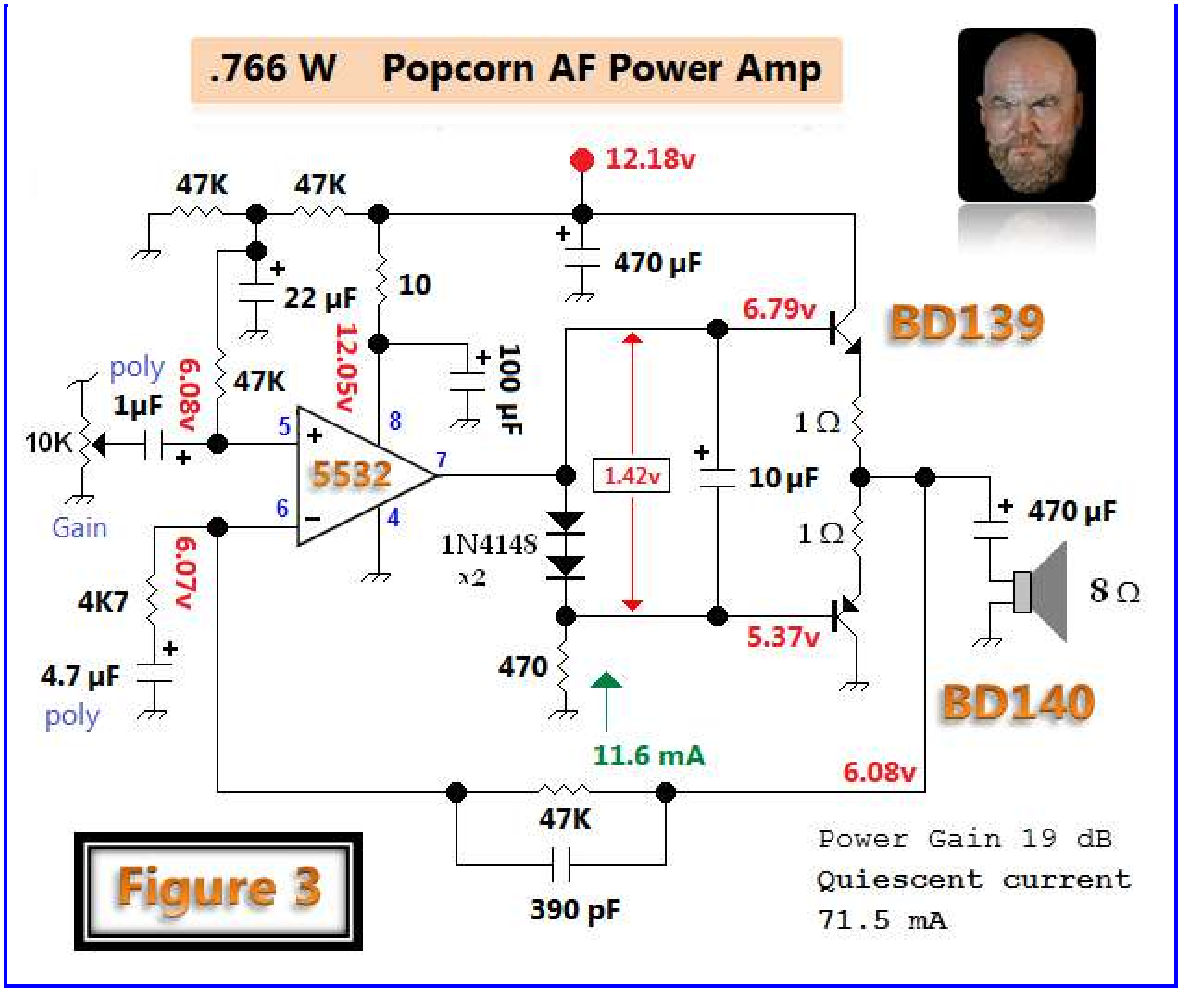
.766 W Popcorn AF Power Amp
12.18v
47K
47K
22 µF
10
470 µF
12.05v
100 µF
6.79v
BD139
poly
1µF
6.08v
47K
10K
5
8
5532
7
1.42v
10 µF
1 Ω
1 Ω
Gain
6
4
1N4148
x2
470 µF
8 Ω
4K7
6.07v
5.37v
4.7 µF
poly
470
11.6 mA
BD140
6.08v
47K
390 pF
Figure 3
Power Gain 19 dB
Quiescent current
71.5 mA

Above — A variation of Figure 2 employing diodes instead of an NPN level-shifter. To kill the cross-over distortion, I lowered the 4K7 resistor by a magnitude of 10. This gave a maximum clean power of 766 mW with a quiescent current of nearly 72 mA. Head room and quiescent current are inferior to the Figure 1 circuit.

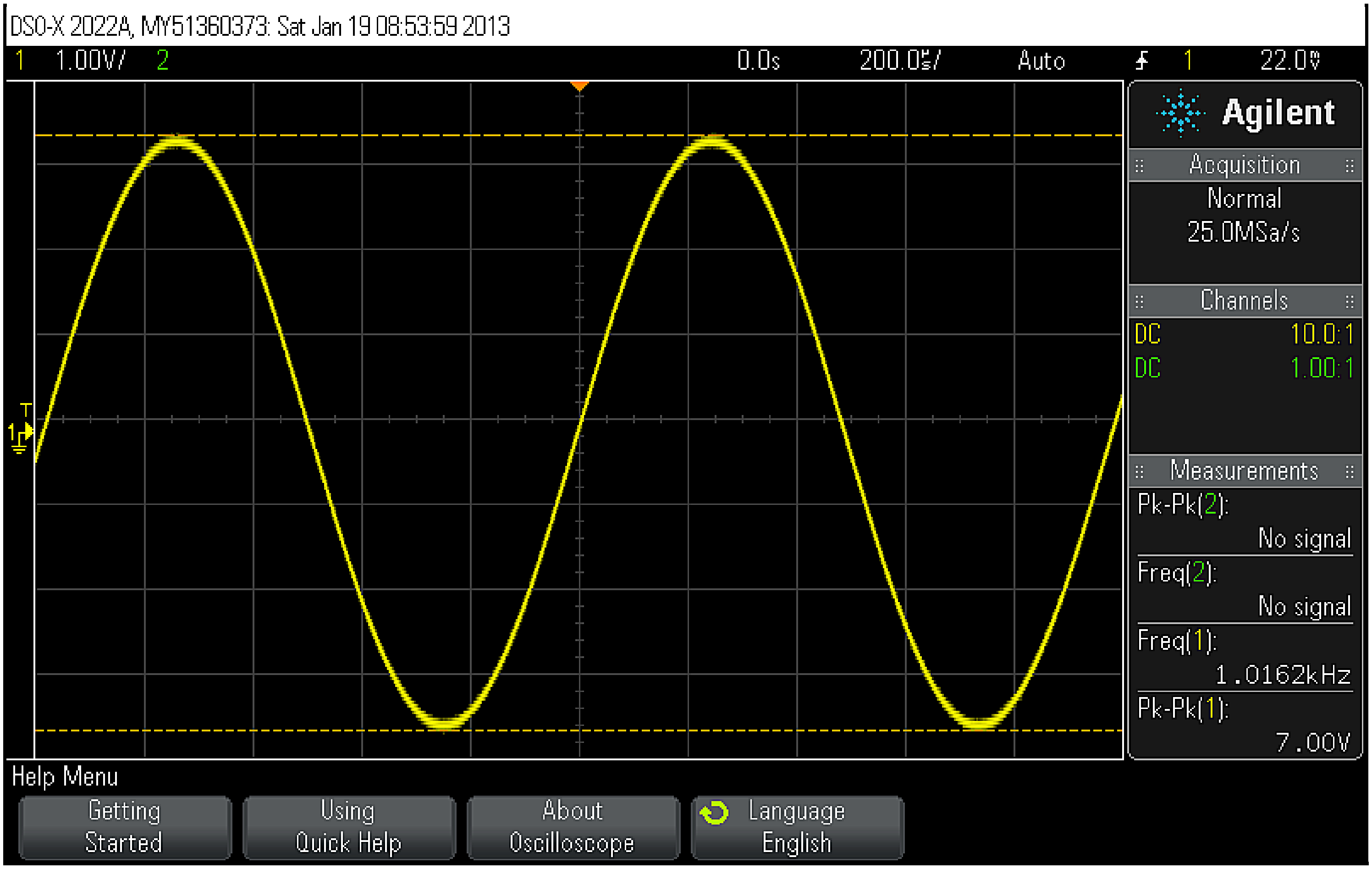

DSO-X 2022A, MY51360373: Sat Jan 19 08:53:59 2013

1 1.00V/ 2 0.0s 200.0us/ Auto 1 22.0mV

Acquisition
Normal
25.0MSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Pk-Pk(2): No signal
Freq(2): No signal
Freq(1): 1.0162kHz
Pk-Pk(1): 7.00V

Agilent
Help Menu
Getting Started
Using Quick Help
About Oscilloscope
Language English

Above — The Figure 3 'scope tracing. Click for a 'scope tracing with the signal generator amplitude increased *slightly* to push this amp into clipping.

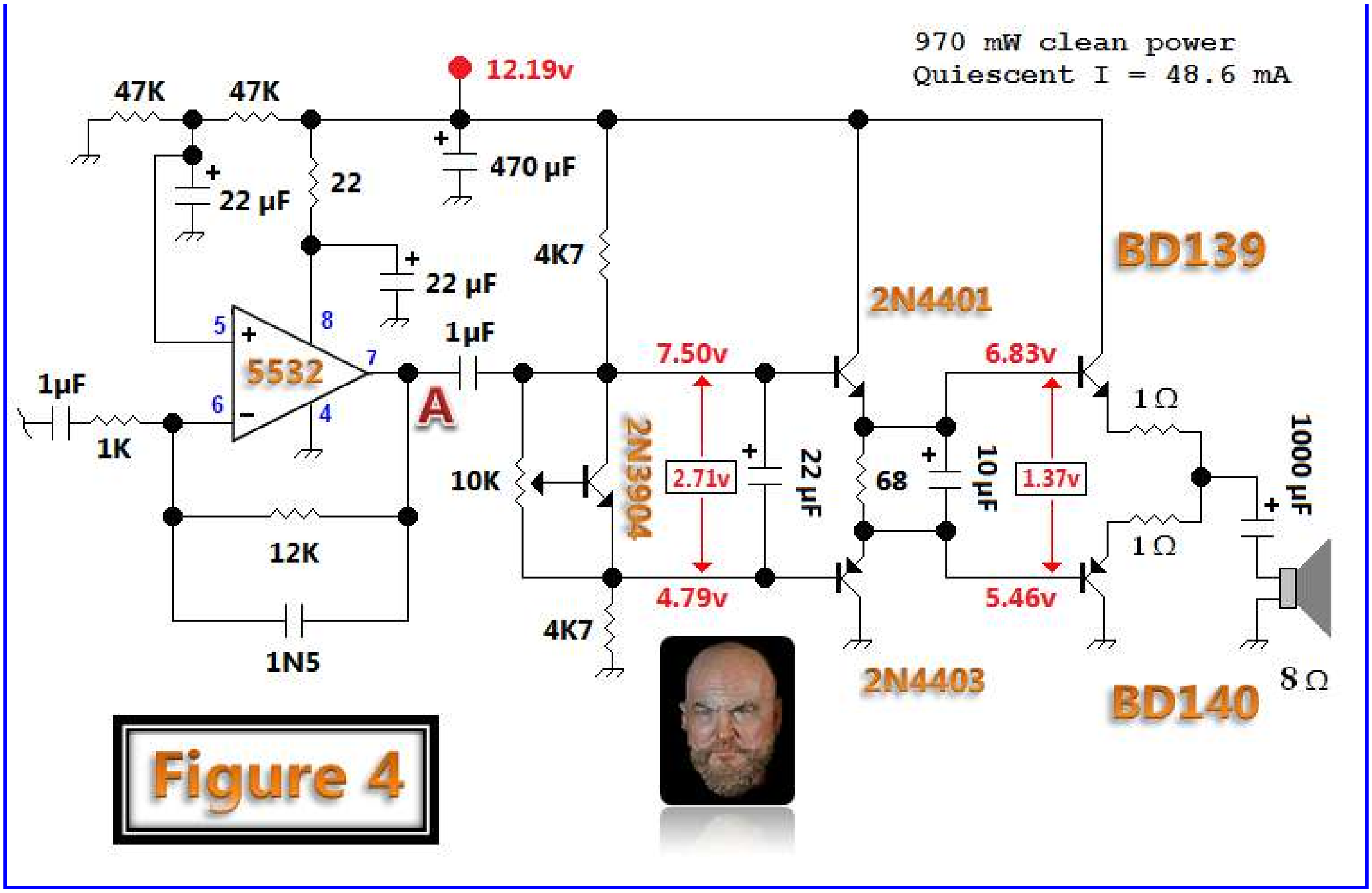
970 mW clean power
Quiescent I = 48.6 mA
12.19v
47K
47K
22 µF
22
470 µF
4K7
22 µF
BD139
2N4401
5532
5
8
7
6
4
1µF
1µF
A
7.50v
6.83v
1Ω
1K
10K
2N3904
2.71v
22 µF
68
10 µF
1.37v
1000 µF
1Ω
12K
4.79v
5.46v
1N5
4K7
2N4403
BD140
8Ω

Figure 4

Above — Back to an AC coupled power amplifer like Figure 1. I added a set of intermediate followers built with a 2N4401/2N4403 pair. The clean output power now lies at 970 mW with a quiescent current under 50 mA. Adjusting the trimmer potentiometer on the level-shifter even a tiny amount may change the quiescent current dramatically.

I found a bias of 1.37v across the BD139/140 pair removed all trace of cross-over distortion at maximum clean signal power. Just tweak the 10K trimmer potentiometer while looking at your 'scope and decide what bias you prefer. I lower the bias until crossover distortion appears and slowly tweak it to find the sweet spot. Then measure the DC voltage across the power follower base terminals, plus the total stage quiescent current with the signal generator switched off. You might have to repeat this procedure a few times, since trimmer pot adjustment is quite sensitive.

Jerry, W5JH made a PC board. Click for front photo. Click for rear photo.

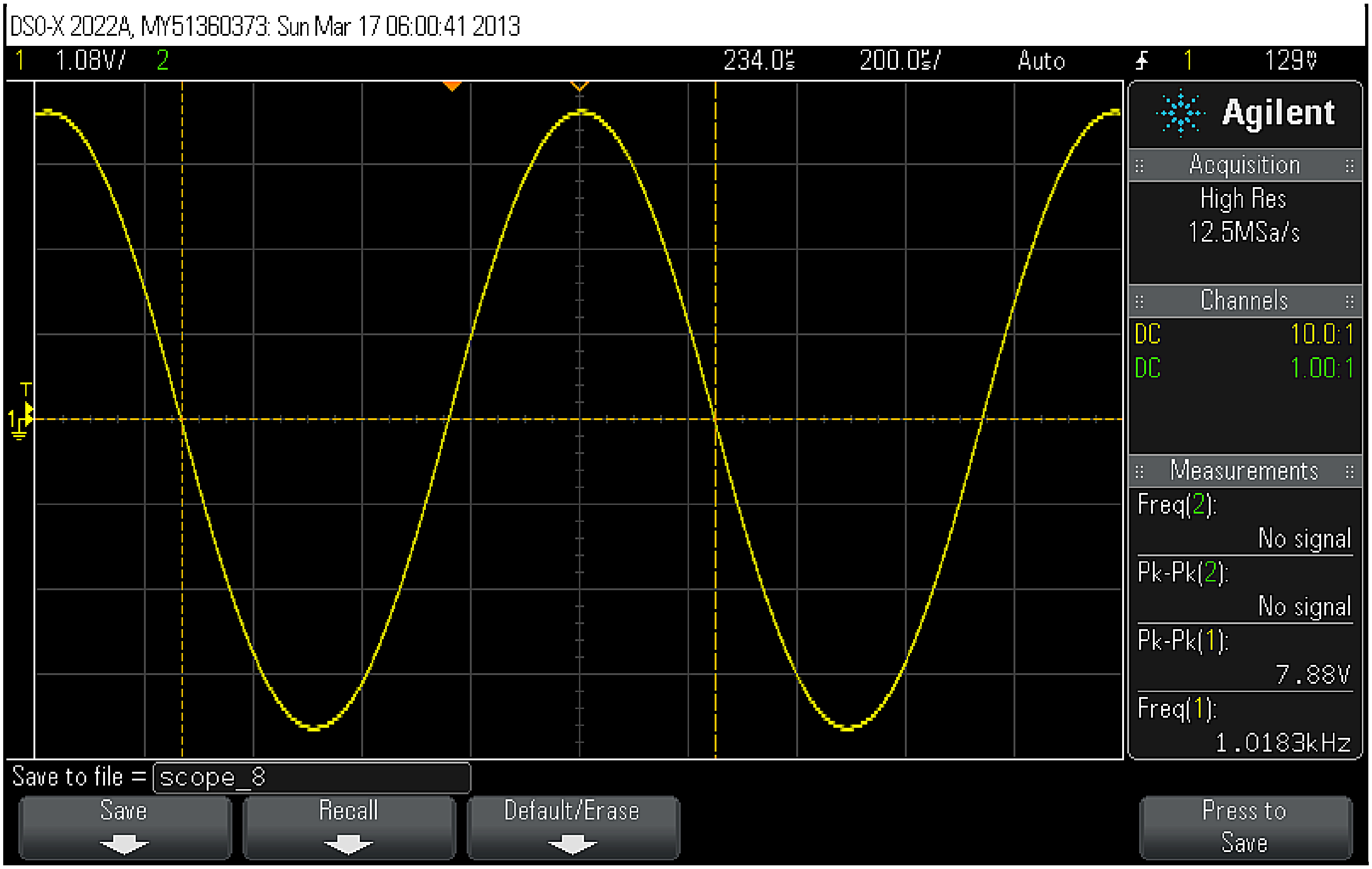

DSO-X 2022A, MY51360373: Sun Mar 17 06:00:41 2013

1 1.08V/ 2 234.0µs 200.0µs/ Auto 1 129mV

Acquisition
High Res
12.5MSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(2): No signal
Pk-Pk(2): No signal
Pk-Pk(1): 7.88V
Freq(1): 1.0183kHz

Save to file = scope_8

Agilent
Save
Recall
Default/Erase
Press to Save

Above — The Figure 4 'scope tracing.

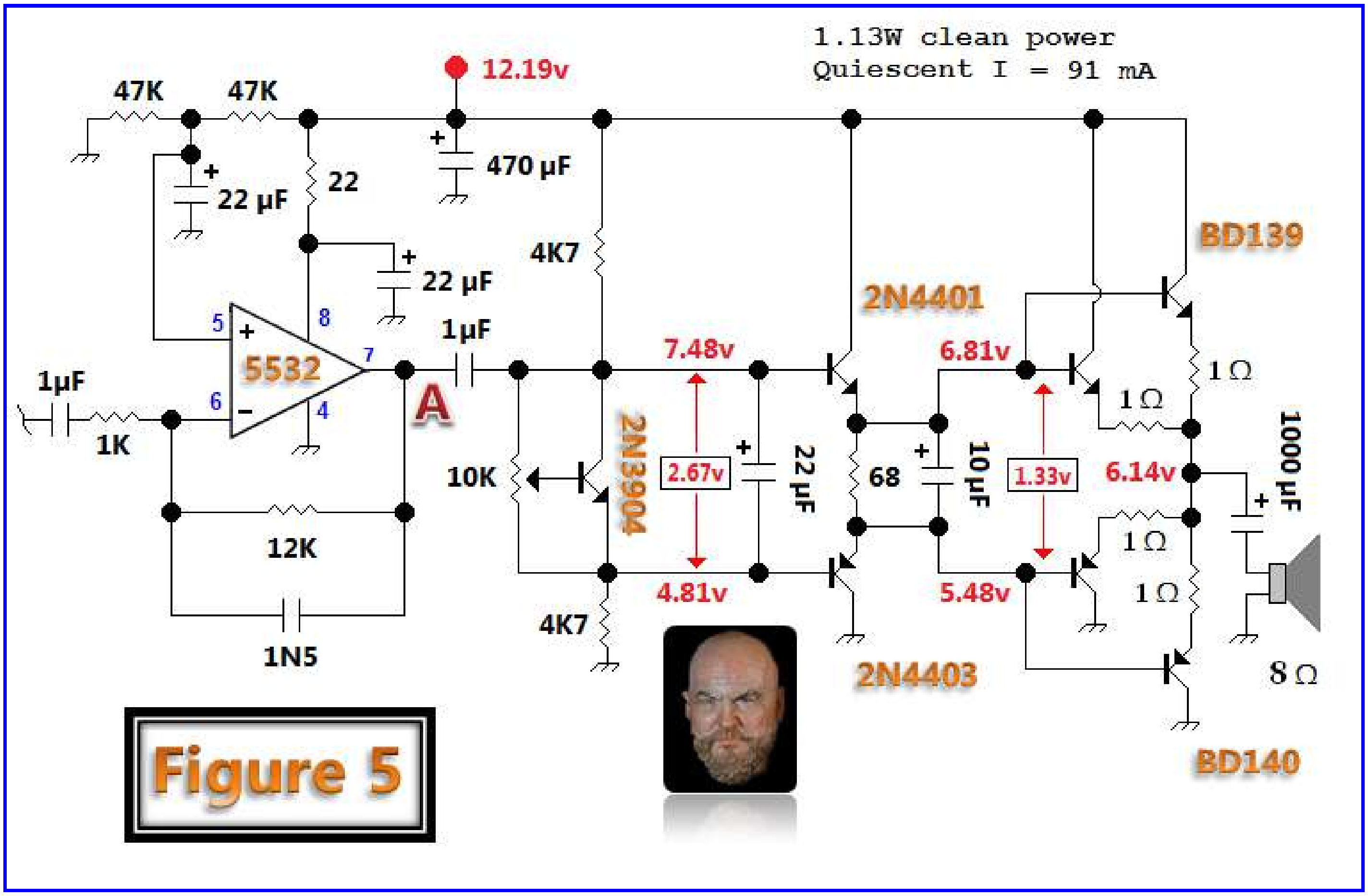
1.13W clean power
Quiescent I = 91 mA
12.19v
47K
47K
22 µF
22
470 µF
22 µF
4K7
BD139
2N4401
5
8
7
6
4
5532
1µF
1µF
A
1K
7.48v
6.81v
1 Ω
1 Ω
2N3904
10K
2.67v
22 µF
68
10 µF
1.33v
6.14v
1000 µF
12K
1 Ω
1 Ω
4.81v
5.48v
1N5
4K7
2N4403
8 Ω
BD140

Figure 5

Above — I added another BD139/140 power follower pair in parallel. The boost in headroom over Figure 4 was small, but it was nice to break the 1 Watt barrier. This amp sounded great and blew away an LM386 set up for a gain of 20 — more headroom, less noise and boosted warmth.

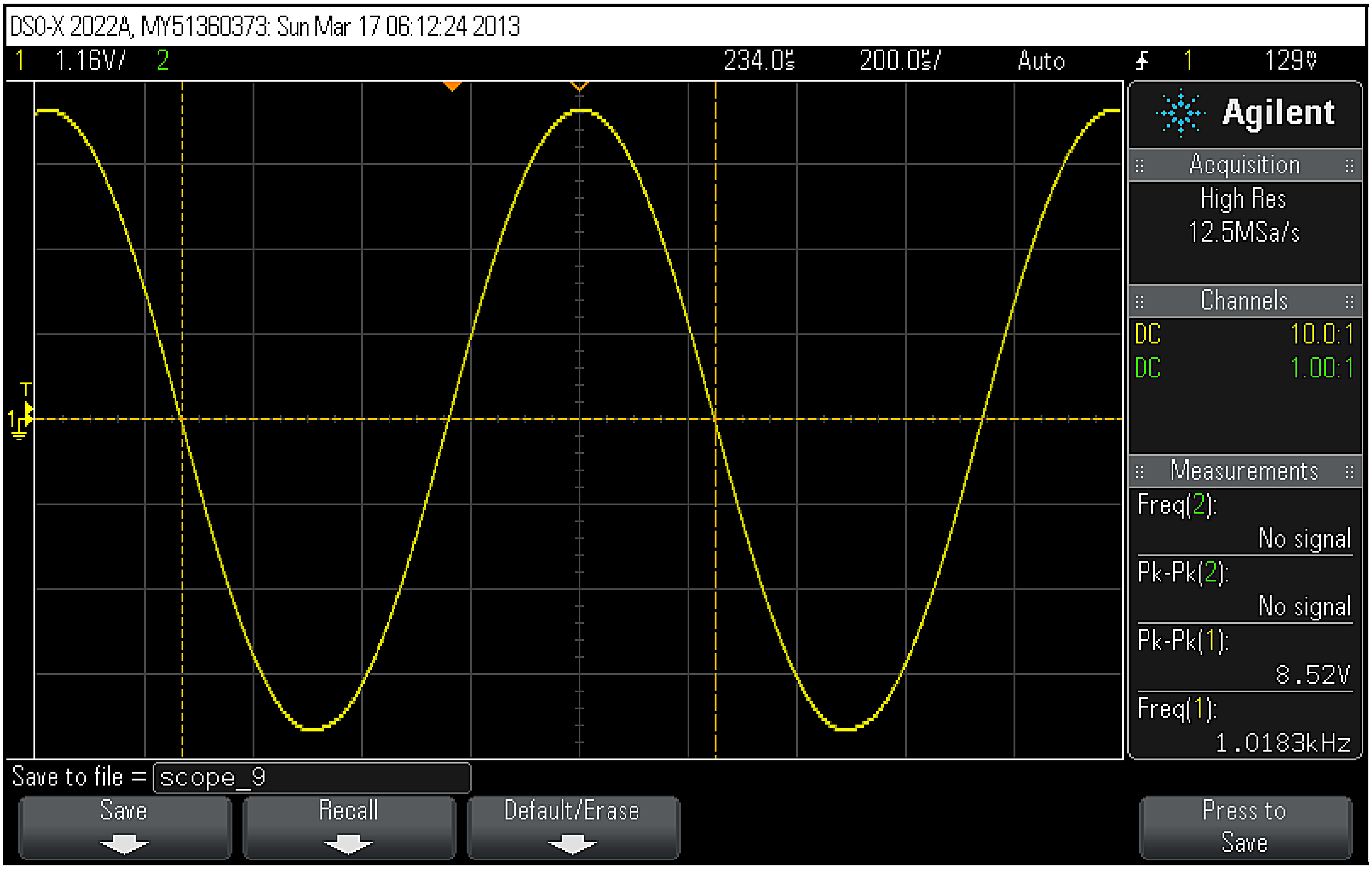

DSO-X 2022A, MY51360373: Sun Mar 17 06:12:24 2013

1 1.16V/ 2 234.0µs 200.0µs/ Auto 1 129mV

Acquisition
High Res
12.5MSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(2): No signal
Pk-Pk(2): No signal
Pk-Pk(1): 8.52V
Freq(1): 1.0183kHz

Save to file = scope_9

Above — The Figure 5 'scope tracing.

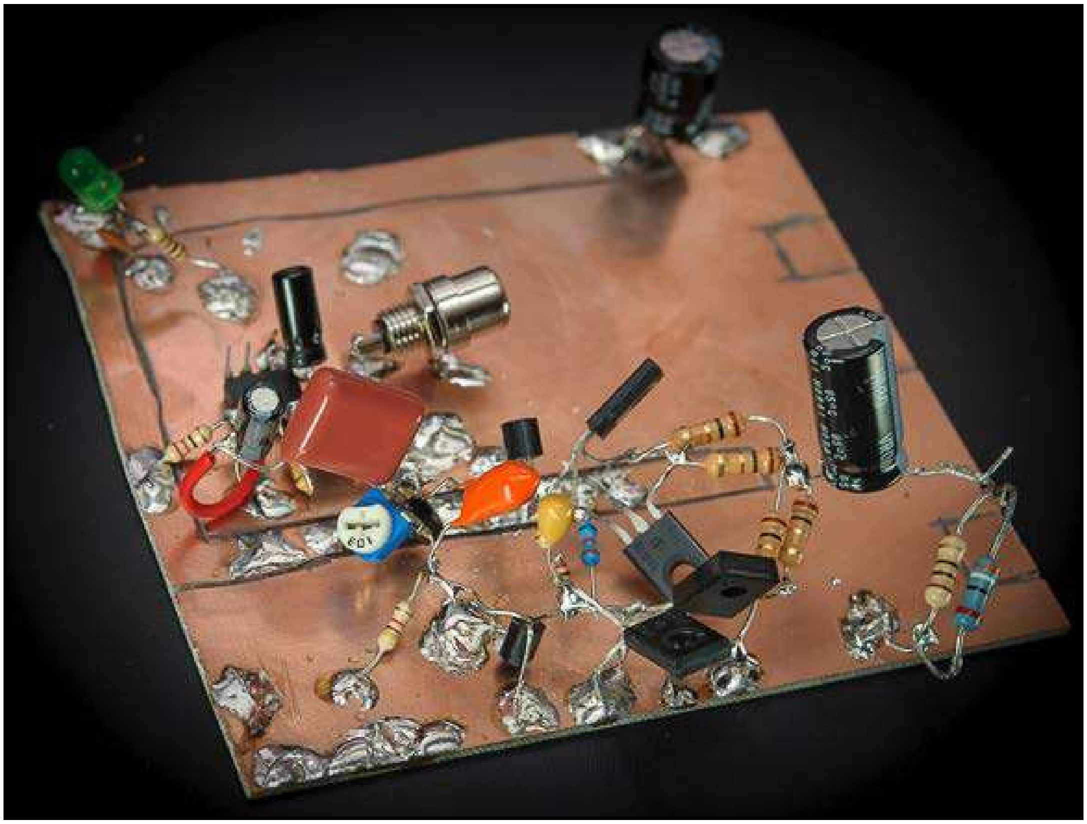

Above — The Figure 5 breadboard. I built all the AF power amps on this board. Signal caps <= to 1 µF were polyester film, while I employed 10 or 22 µF tantalum caps for the level shifters. Electrolytic caps work fine; especially for the level shifter capacitors. The green power indicator LED drew 10 mA and I subtracted this from the quiescent current measurements.

Depending on the AF gain of your receiver, you might wish to add the familiar Zobel filter; a 10 Ω R in series with 0.1 µF C from the positive end of the output capacitor to ground, or more AF bypassing/ decoupling to the circuitry.

No component values were critical — imbuing the spirit of homebrew radio, substitute parts and measure outcomes.

QRP — PosData for August 18, 2014

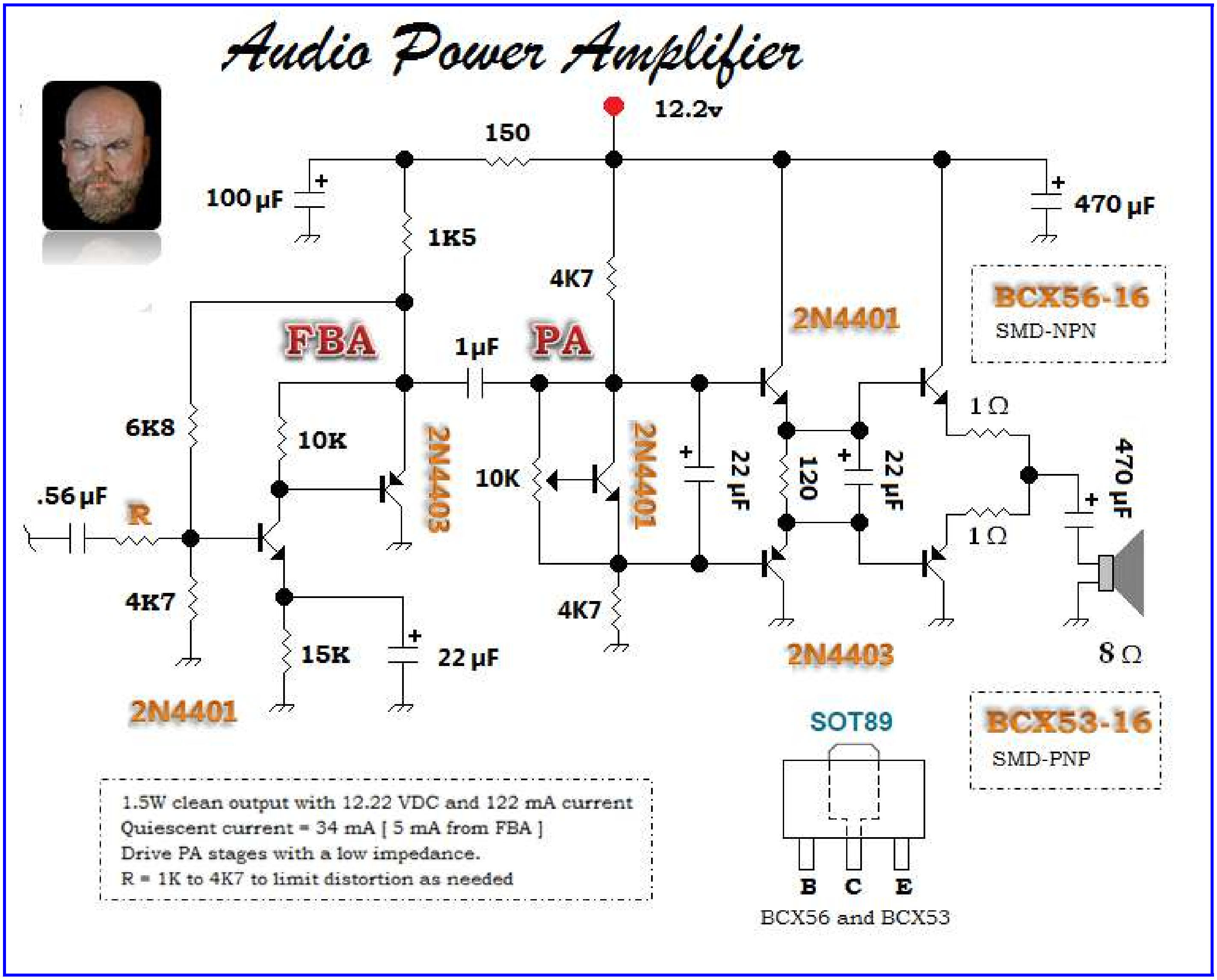

1.5W clean output with 12.22 VDC and 122 mA current
Quiescent current = 34 mA [5 mA from FBA]
Drive PA stages with a low impedance.
R = 1K to 4K7 to limit distortion as needed

Above — I built a 100% discrete version of Figure 4 power amplifer with a 2N4401/2N4403 feedback voltage amp that drives the PA stage with a low impedance. Without a low impedance drive, the PA [all the components right of the 1 µF capacitor where you see **PA**] the headroom of the PA section falls down. Thus an op-amp or emitter follower drive works great.

I moved to SMD power transistors: the NXP BCX56 + BCX53. These give a little more output than the BD139/140, and take much less board room. I ran no heat sink other than 8 mm by 4 mm PC board traces for the collectors. A better choice might be the related BCP56 + BCP53 pair in SOT223 since the bigger package of this version better sinks heat.

I determined the optimum emitter resistor for the 2N4401/2N4403 followers = 120 Ω [not 68 Ω as shown in Figure 4]. This drops the the PA quiescent current in half and yields the same performance. If you build Figure 4, change to 120 Ω.

I meant Figure 4 and the QRP POS-Data amp as easy-to-make power amplifiers with a *little* gain. You will have to add voltage gain in some AF projects. For Figure 4, it's easy to change up the 12K resistor, or apply the other half of the 5532. Decouple and bypass better if you get motor boating or hum. 73 de VE7BPO

---

9. The Progressive Receiver by Wes, W7ZOI and John, K5IRK

When introduced in QST for November 1981, the Progressive Communications Receiver (PR) by Wes, W7ZOI and John, K5IRK set a dynamic range benchmark for dual-conversion homebrew receivers: 94 dB in CW mode. 31 years later, few other home-built radios have ever reached this benchmark.

The PR is a single conversion superheterodyne 80m receiver; or perhaps it's a direct conversion receiver with an extra mixer ahead of the product detector? It's both and that's the point. Further, Wes and John added other bands with another mixer, plus a crystal oscillator, RF filter +/- an RF amp for each band.

You'll see the PR listed as A High-Performance Communication Receiver in ARRL Handbooks from the early 90s or so. ARRL staffers built and enjoyed the PR in their test lab for many years.

After purchasing the 1991 ARRL Handbook and reading about this project, I slowly adopted and entrenched the PR's progressive (modular) approach. When I built and web published my 2 TRF WWV receivers with crystal IF filters, a couple of people wrote that I "grossly over-designed" them, however, other, more astute builders placed a mixer on the front end and turned their TRF into a multi-band superhet like I later did. These builders understood how progressive circuit building works — all PR inspired.

I hope my introduction renews your interest in the PR. Wes and John's article is thorough and complete — I can't add to it, however, I'll share my thoughts along with those from Wes and John all these decades later.

The IF Stage gain comes from 2 dual gate MOSFET amplifiers. The final MOSFET amp drives a BJT differential pair providing 9 MHz RF to the product detector at 50 Ω and signal for the AGC circuit. Wes recalls using a 3N211 in the original MOSFET slots since Doug DeMaw owned a pile of them and contributed some. In some areas, the 40673 was the dominant 2-gate MOSFET soldered by radio home builders back in the day.

An evolved version of the PR Intermediate Frequency stage appeared in EMRFD as Figure 6.50. Wes replaced the now hard-to-find and *expensive* leaded dual-gate dual gate MOSFET with a cascode of J310s. I built this stage and it performed well with my 12.2 VDC power supply.

Jeff, WA7MLH built the general purpose IF system and noticed if the power supply dropped below 12 VDC, the gain control and maximum gain fell off and he wrote to Wes about his findings. Wes later confirmed dysfunction with a lower DC supply. I asked him to recollect this time:

"…. Frankly I don't remember if I went immediately to computer simulations or if I built a single stage. I think I built. Anyway, Jeff was correct, depending upon the FETs that were used. Then I got to plowing into the details of the cascode, this time with SPICE. I was using PSPICE for some simulations, but was in the process of switching over to LT-SPICE. That makes no difference, for I used the same models in both.

Anyway, playing with a single stage showed that the mechanism for gain reduction was that when the voltage on the upper gate was reduced, it did nothing to the upper stage, but it compressed the supply on the lower FET. So far as signal goes, the upper part was nothing more than a common gate stage that passed whatever signal current was there in the drain of the lower part on into the source of the upper part.

But these were depletion mode FETs, the normal thing for most of the JFETs we use. As such, you have to get the gate down pretty low to get the source voltage low enough to be effective in reducing the gain of that lower FET. That's when it became clear that one could use other parts in the upper slot. An enhancement mode FET such as a MOSFET would work well. And just a common bipolar would do the job nicely too..."
Wes, W7ZOI per an email - January 2013.

In QST for 2007, a refined version of the EMRFD Figure 6.50 IF stage appeared with a hybrid cascode amplifier instead of the cascode JFETs (and the 2 original dual gate MOSFETs of the PR). Further, Wes added a third amplifier to widen gain control. The hybrid cascade stage has since gone in 100s of receivers across the globe; ensuring the PR legacy lives on as the modernized hycas version.

Roger, KA7EXM sold kits of the hycas amp for years, however, these kits sold out. PC boards may still be available via his web site. Roger received permission to publish the QST hycas article online:

For contents info / errata:
www.ka7exm.net/hycas
ARS KA7EXM
Roger Hayward
HYCAS-IF-AMP
WA7MLH
W7ZOI

Above — The now *sold out* hycas IF System kit once offered by Roger, KA7EXM.

To underscore my love of the original PR IF amp, I built a version using SMT dual gate MOSFETs in 2013 and feel the SMT version might be perfect for builders who prefer to manually control IF gain in most situations, but want AGC control now and then. Versions built with the BF998 MOSFET may suffer parasitic ocillations at UHF and so meticulous attention to decoupling and bypassing out to UHF and in some cases double-sided copper board may be required to prevent unwanted ringing at UHF.

If you build the entire PR with the BF998, the VFO output runs about 10 dBm and requires attenuation. Again, please consider UHF oscillation precautions.

Wes and others have built both leaded and all-SMT versions of the hycas amplifer.

John, K5IRK Recalls

I wanted John's recollection of the PR design and build and received the following narrative in late March 2013:

"…One night in late 1979 I had telephoned Wes to chat about QRP and building rigs. During the conversation he told me he had a project and related QST article in mind and asked me if I wanted to participate. He was in the process of writing a book (IRFD) and didn't have much free time, so he needed some help. I told him yes. A few weeks later I received the first schematics and began to gather parts.

Recall that the project started as direct conversion receiver. But it then progressed (hence the title of the article) to a superhet. I choose to build my superhet with a SSB filter as I had already built Wes' CW Competition Grade RX from SSD. During the next year he sent me schematics and a few critical parts thru snail mail. (The Internet and e-mail would come later.) I would build the circuits and report the results back to him. As I recall, we only had about three phone conversations throughout the whole project.

The design was an iterative process. For example, the IF circuitry began with a single 40673, but grew to include two stages. Wes bread boarded this in a mono band version of the receiver at his end. An early audio derived AGC system was replaced by the IF derived circuit that appears in the article. The VFO was also designed twice, as were the Front end Filters and the BFO. More is said below about the design procedure.

I designed the circuit board layouts. The traces on my boards were drawn by hand and etched at home. I sent hand drawn sketches of each of the boards to Doug DeMaw; Circuit Boards Specialists (CBS) then created the commercial boards for those folks who wanted boards. I recall that we sent out layouts to any readers for a SASE. All of my boards were designed and built on double sided PC material with the exception of the VFO board and the Audio Filter boards. This can be seen in the photos. The boards eventually sold by CBS were single sided. All layouts and functionality were confirmed prior to publication.

The traditional design process for a homebrew ham receiver in 1980 started with schematic sketches based upon the intuition and experience of the designer. The total circuit would then be built. Measurements were merely things that were done afterward, something to characterize the result. The PR was different. Individual stages were designed, built, and measured. Negative feedback was used in the critical amplifiers to guarantee that the gains were high enough for reasonable noise figure, but low enough to preserve input intercept.

Only after the individual stages were operational, were they assembled to form a working receiver. Some stages were further modified during system assembly and evaluation. We had no computers available for circuit simulation, although the gain distribution was optimized with a hand calculator. The goal was not just a receiver that sounded good, but a box with good two tone dynamic range.

Once my receiver was completed, I sent it to Wes for MDS and DR measurements. A goal was to compare these measurements with those already done in Oregon. Roger (KA7EXM) took many (if not all) of the photos. Wes then forwarded the receiver to ARRL. They returned it to me when they were finished. Wes refined the article during the final months of the project before it was finally published. I received copies of his drafts, and then offered my feedback.

We had hoped that a "few" experimenters would enjoy building the receiver, but had no idea that it would be as successful as it was. We eventually learned why this occurred: First, the receiver held up well when it was measured at ARRL Headquarters, exceeding the performance of most of the appliances being evaluated at the time. This prompted some League staff members to build the receiver for their own use. This, in turn, prompted them to include it in the Handbook for several years.

My original receiver plays as well today as it did in the beginning. The electrolytic caps on the audio board have been replaced, for they were beginning to go south on me, but that's the only change. I do have a second version that is used for experiments..."

Some PR photos taken by John in March 2013

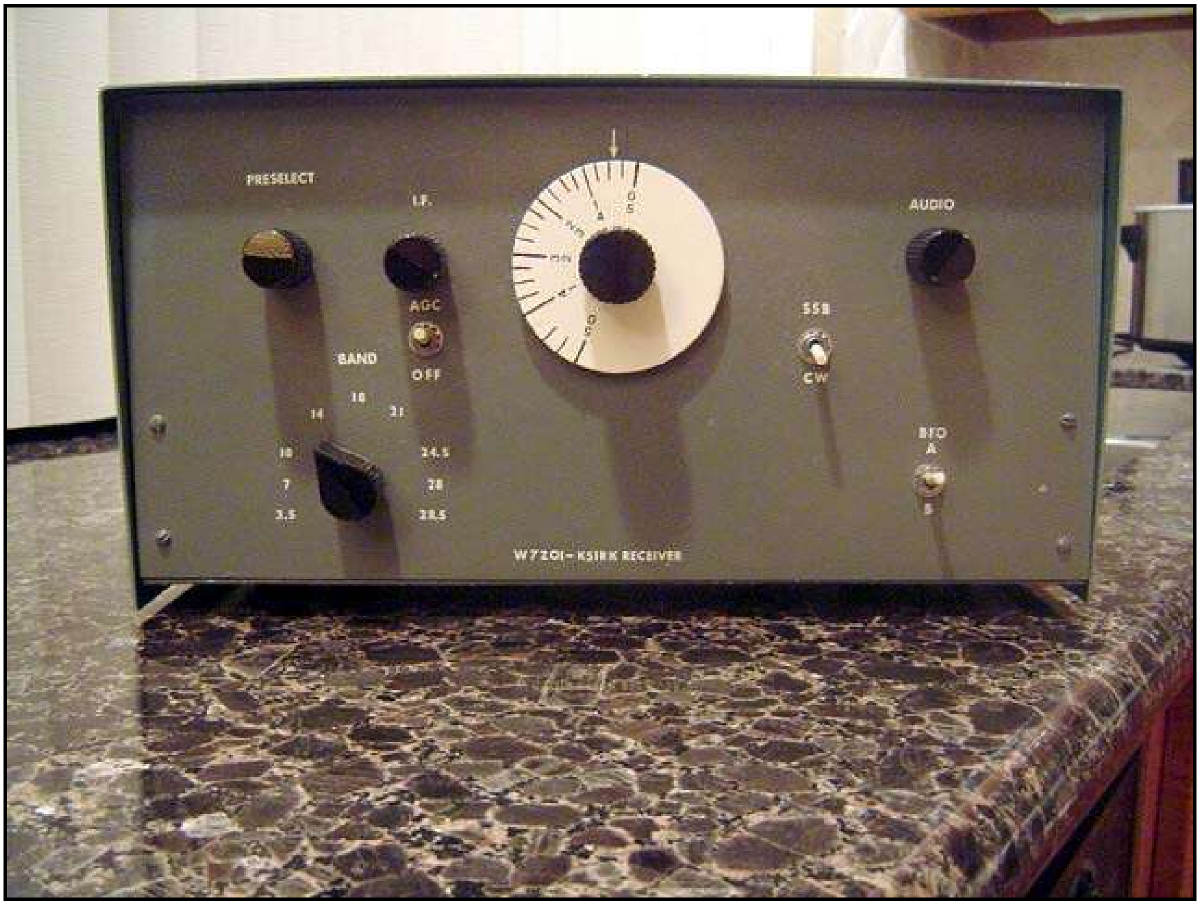

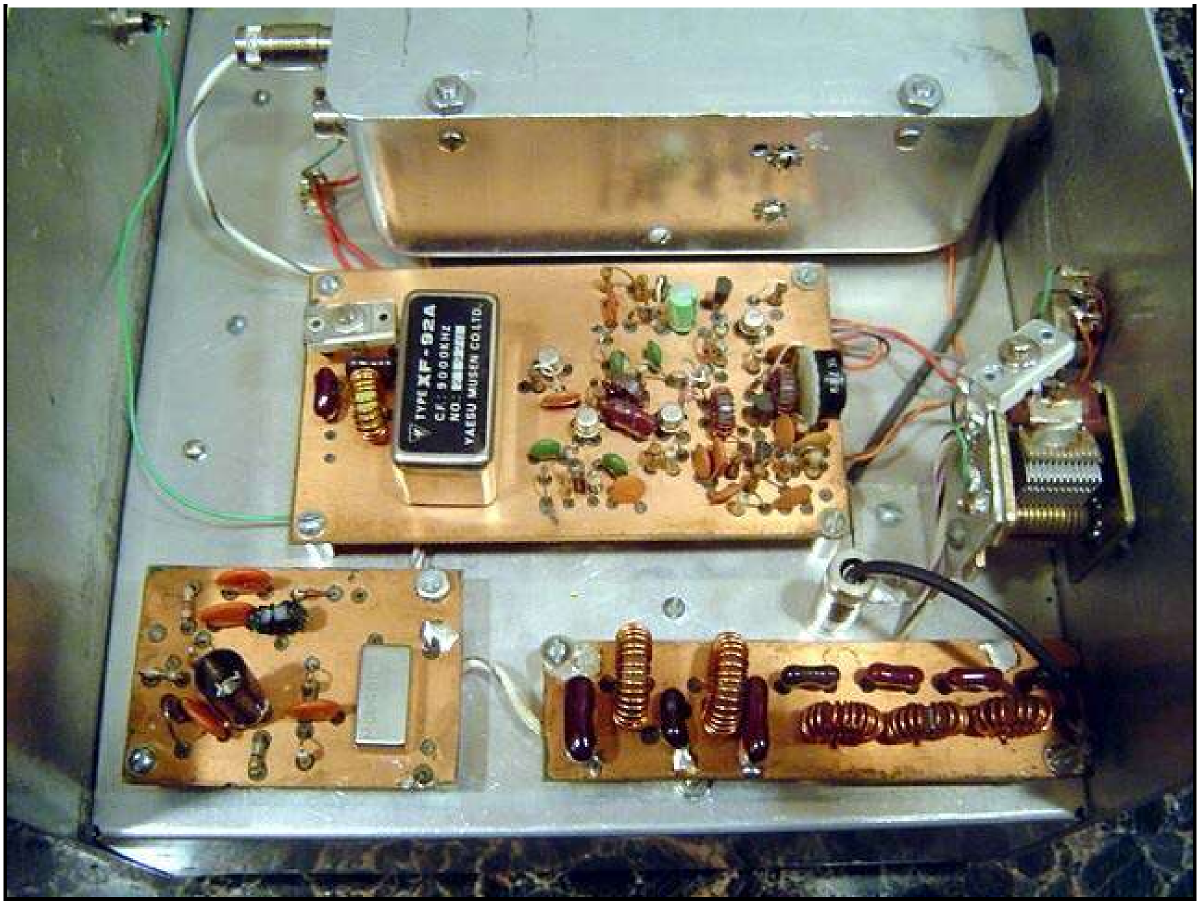
TYPE XF-92A
C.F. 9000KHZ
NO
YAESU MUSEN CO.LTD.

"...You will notice a couple of things different in these photos than in the photos in the article. First, tacked on to the Front Panel 365 pF variable cap is a mica trimmer cap.....this is of no significance as I was just messing with lowering the BW upper and lower frequencies to see what difference it would make....None is the answer..." *John, K5IRK.*

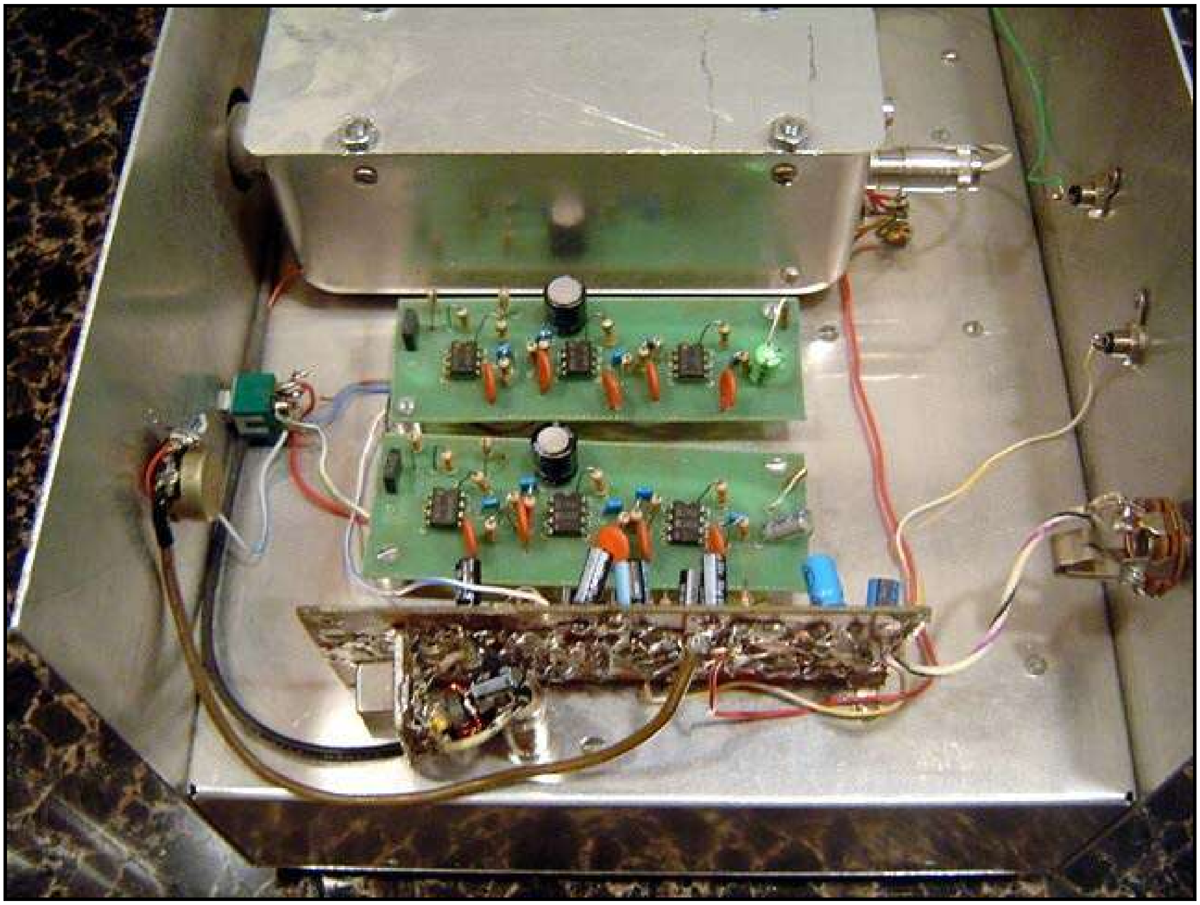

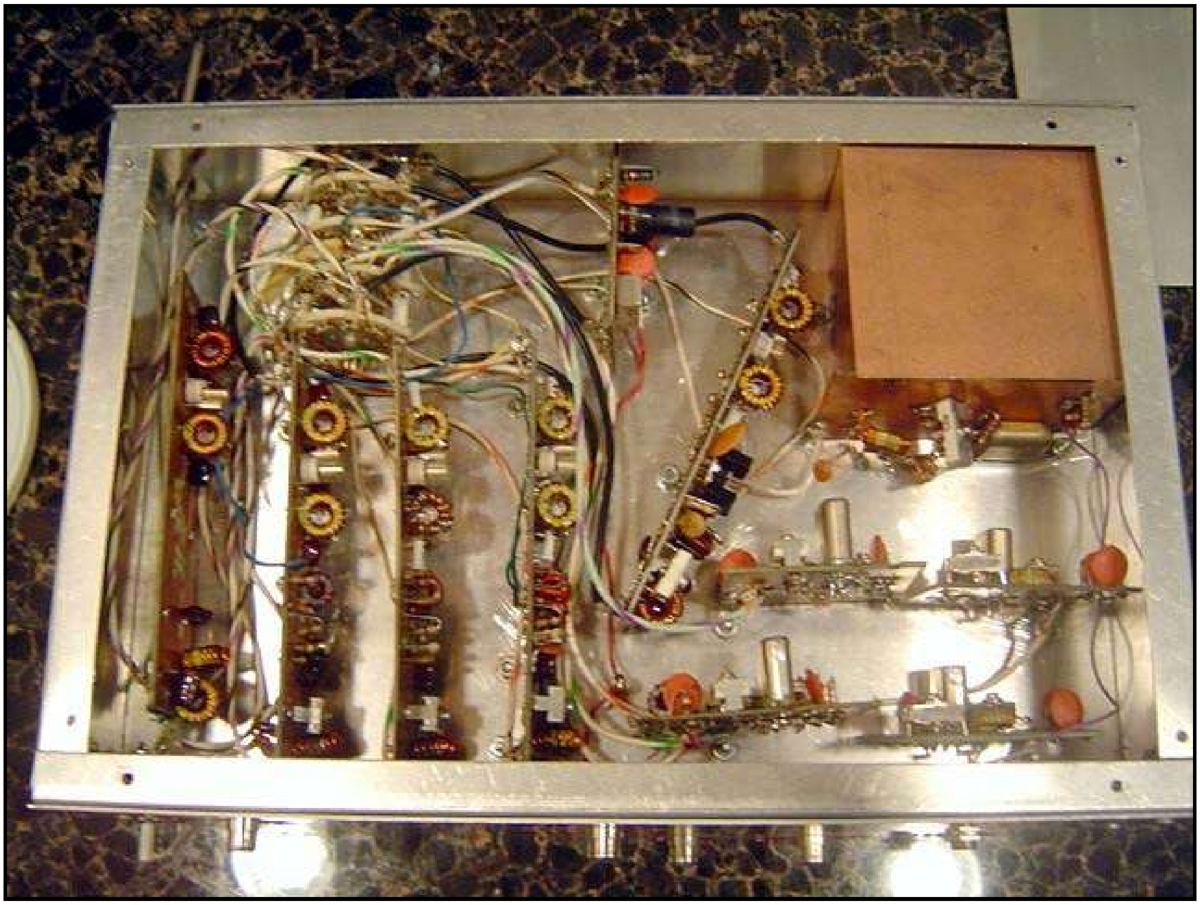

"...In the above photo you will see the addition of the 30 meter band to the RX with the diagonal front end filter board and the Xtal Oscillator attached to the BFO Box's wall that was added some years after the article was published…." *John, K5IRK.*

High-pass plus Peaked Low-pass Filter

If you know me, you know I love the peaked low-pass filter both at RF and AF. This adoration came from studying the PR and other work published and shared privately to me by Wes. *A good reference = The Peaked Lowpass: A Look at the ultraspherical filter by Wes for Ham Radio, June 1984.*

While the high-pass + peaked low-pass filter placed ahead of the second mixer in the PR wouldn't likely go in a modern W7ZOI design, its narrow-band LC filtration generates a roofing action when the receiver uses converters for bands other than 80 meters. The result is that much of the DR obtained in the single conversion version is retained in the dual conversion receivers.

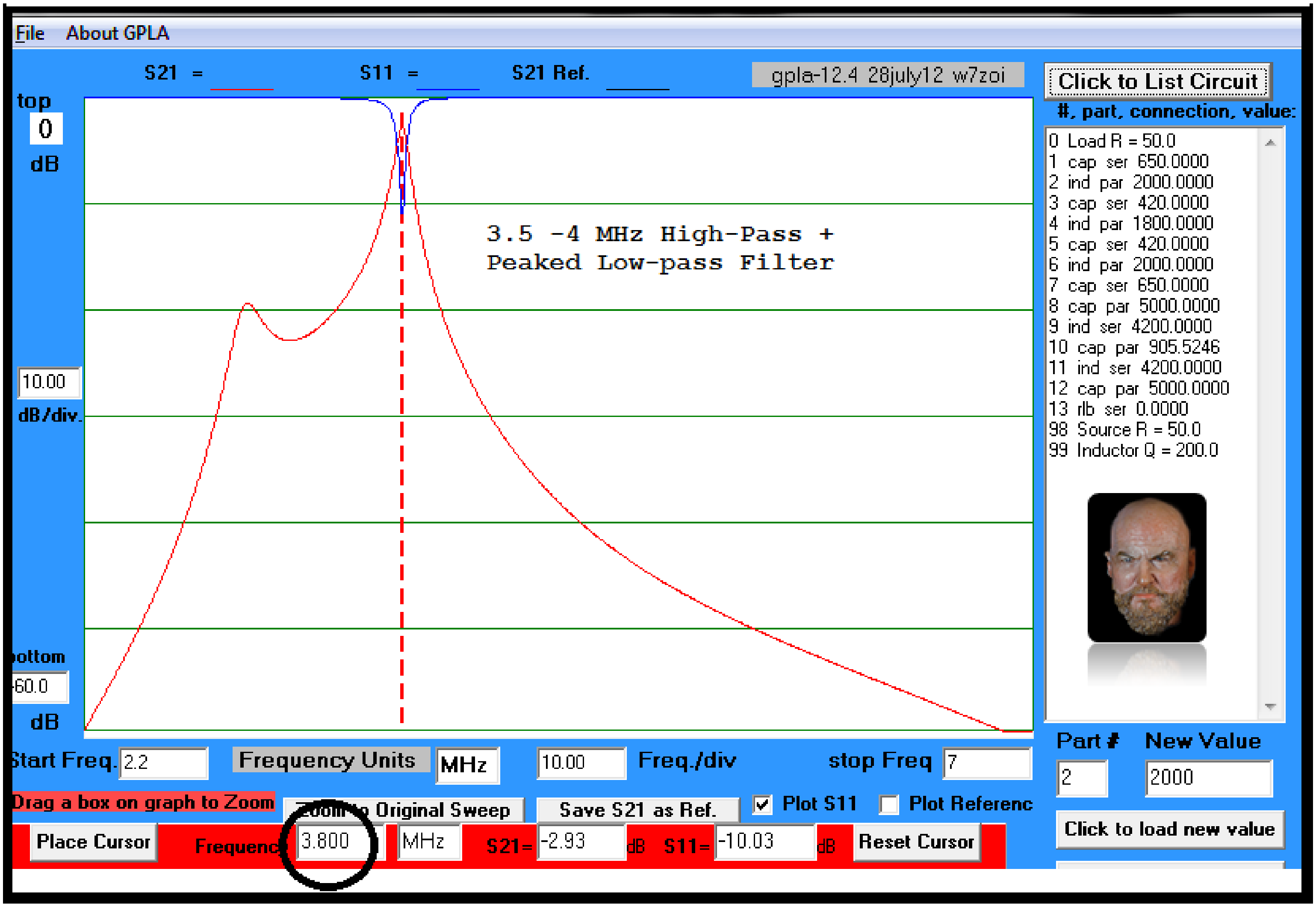
File
About GPLA
S21 =
S11 =
S21 Ref.
gpla-12.4 28july12 w7zoi
Click to List Circuit
top
0
dB
10.00
dB/div.
ottom
60.0
dB
3.5 -4 MHz High-Pass +
Peaked Low-pass Filter
Frequency Units
Freq./div
stop Freq
Drag a box on graph to Zoom
Zoom to Original Sweep
Save S21 as Ref.
Plot S11
Plot Referenc
Place Cursor
Reset Cursor
Part #
New Value
Click to load new value

#, part, connection, value:

```
0 Load R = 50.0
1 cap ser 650.0000
2 ind par 2000.0000
3 cap ser 420.0000
4 ind par 1800.0000
5 cap ser 420.0000
6 ind par 2000.0000
7 cap ser 650.0000
8 cap par 5000.0000
9 ind ser 4200.0000
10 cap par 905.5246
11 ind ser 4200.0000
12 cap par 5000.0000
13 rlb ser 0.0000
98 Source R = 50.0
99 Inductor Q = 200.0
```

tart Freq. 2.2 Frequency Units MHz 10.00 Freq./div stop Freq 7

☑ Plot S11 ☐ Plot Referenc

Frequenc 3.800 MHz S21= -2.93 dB S11= -10.03 dB

Part # 2 New Value 2000

Above — A GPLA simulation of the front end high-pass/peaked low-pass filter centered for 3.8 MHz by tweaking capacitor #10 in the software. Click for another simulated filter that covers 3-4 MHz. Look at the sublime low and high-pass skirt action with a 3 dB bandwidth of 73 KHz. Love this!

Click for a screen shot of an entire 40 Meter band version shown with the low-pass filter section centered at 7.003 MHz. I also made 1 for WWV 10 MHz — by scaling the original version's XL and XC and tweaking with GPLA, you may build 1 for any HF band.

AF Amplifier

The post product detector AF chain sounds great. Many of us later replaced the Q5 mute switch with something quieter, however, even today — this AF block holds its own against most discrete component headphone-level circuits. I particularly love the crisp fidelity of the Class A feedback pair Q3 and Q4. Today, I would substitute a BD139 for Q4.

Wrap up

I could go on, but this web page is already too long. Studying + building stages from the PR; a receiver designed more than 3 decades ago, might raise your game today. Go team!

---

10. Miscellaneous Pictures and Figures

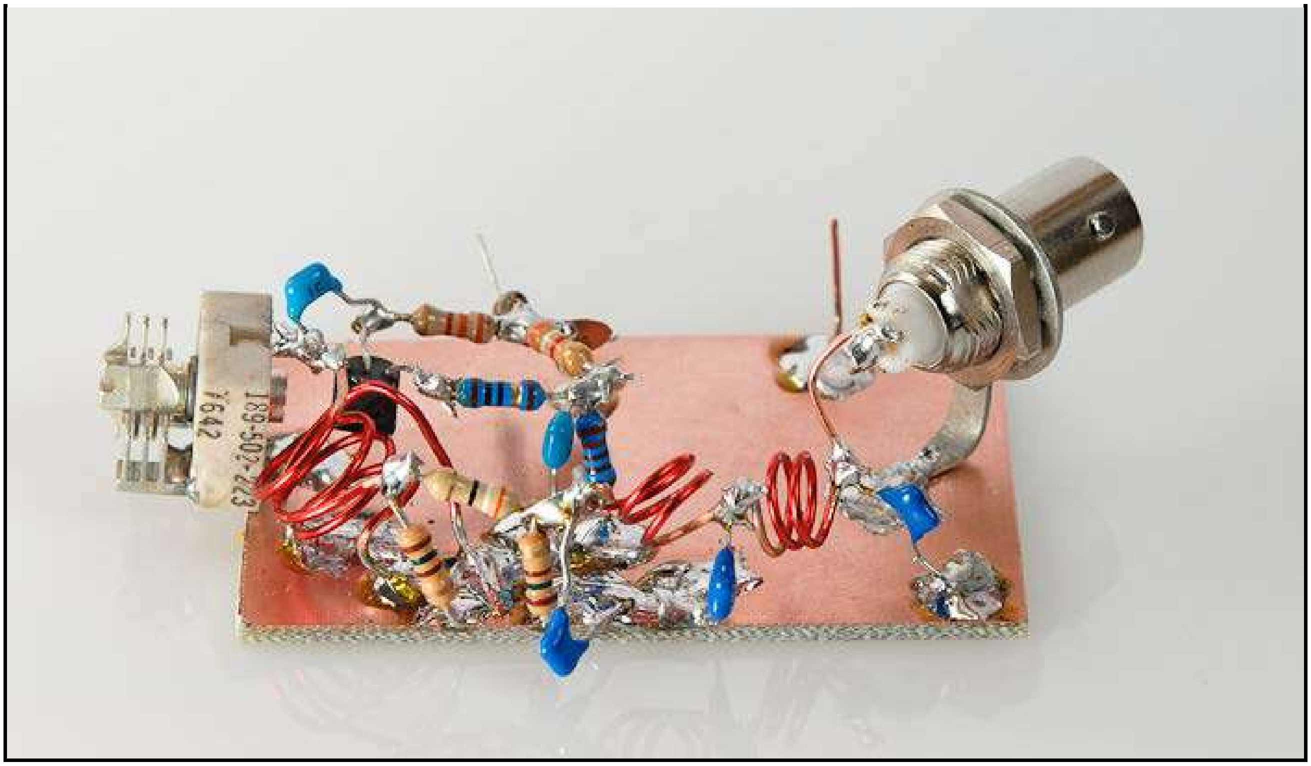

Above — I built a prototype 200 MHz oscillator with trailing low-pass filter before making the 210 MHz version placed in the 1-118 MHz VFO for the K3NHI sweep system. (Section 1). The secondary coil floats and may be positioned between any of the 4 primary links to change coupling and thus output power.

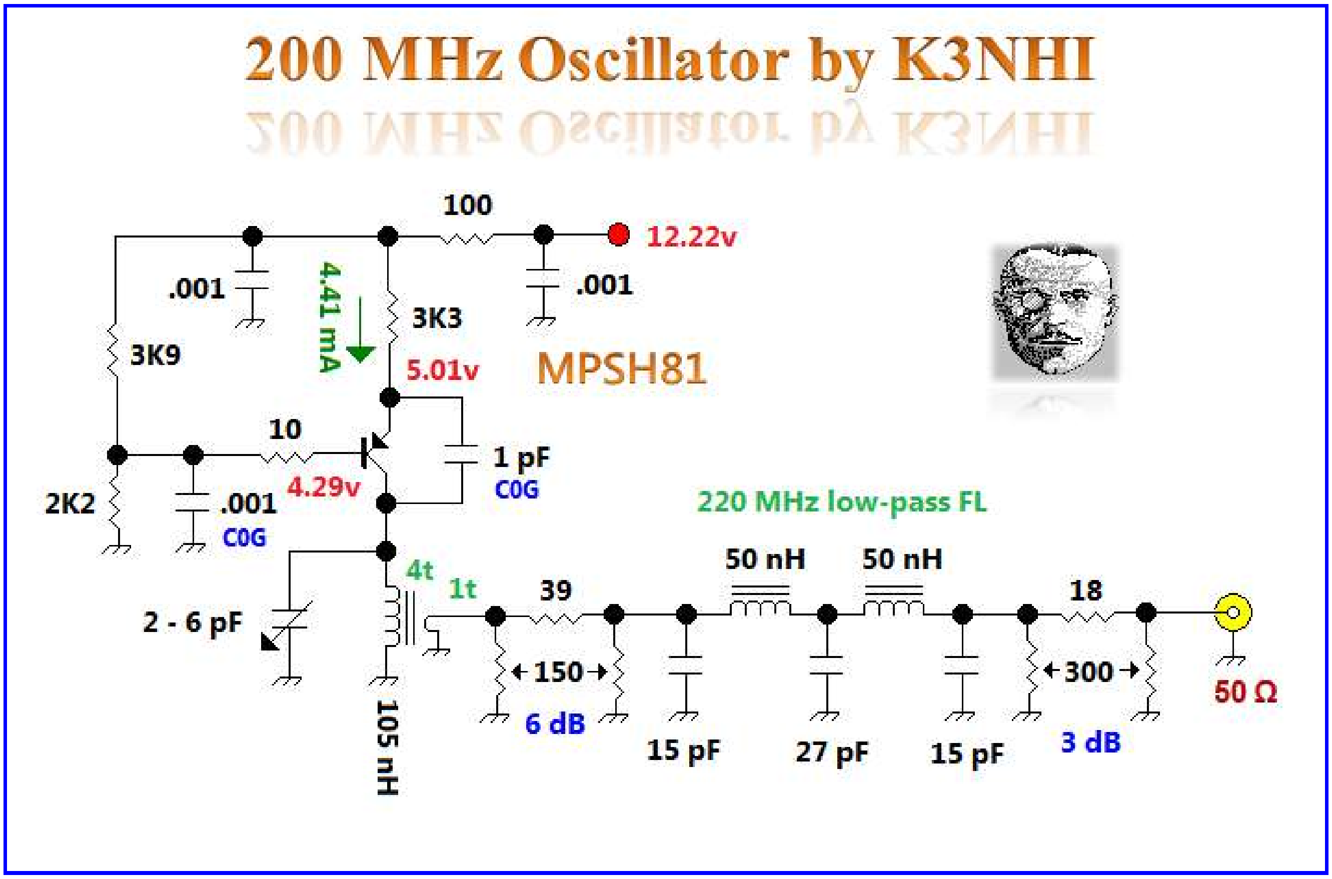
200 MHz Oscillator by K3NHI
100
12.22v
.001
4.41 mA
3K3
.001
3K9
5.01v
MPSH81
10
1 pF
C0G
2K2
.001
C0G
4.29v
220 MHz low-pass FL
4t
1t
50 nH
50 nH
2 - 6 pF
39
18
150
300
105 nH
6 dB
15 pF
27 pF
15 pF
3 dB
50 Ω

Above — The schematic of my 200 MHz local oscillator. Click for another version I designed that tuned from ~135-208 MHz.

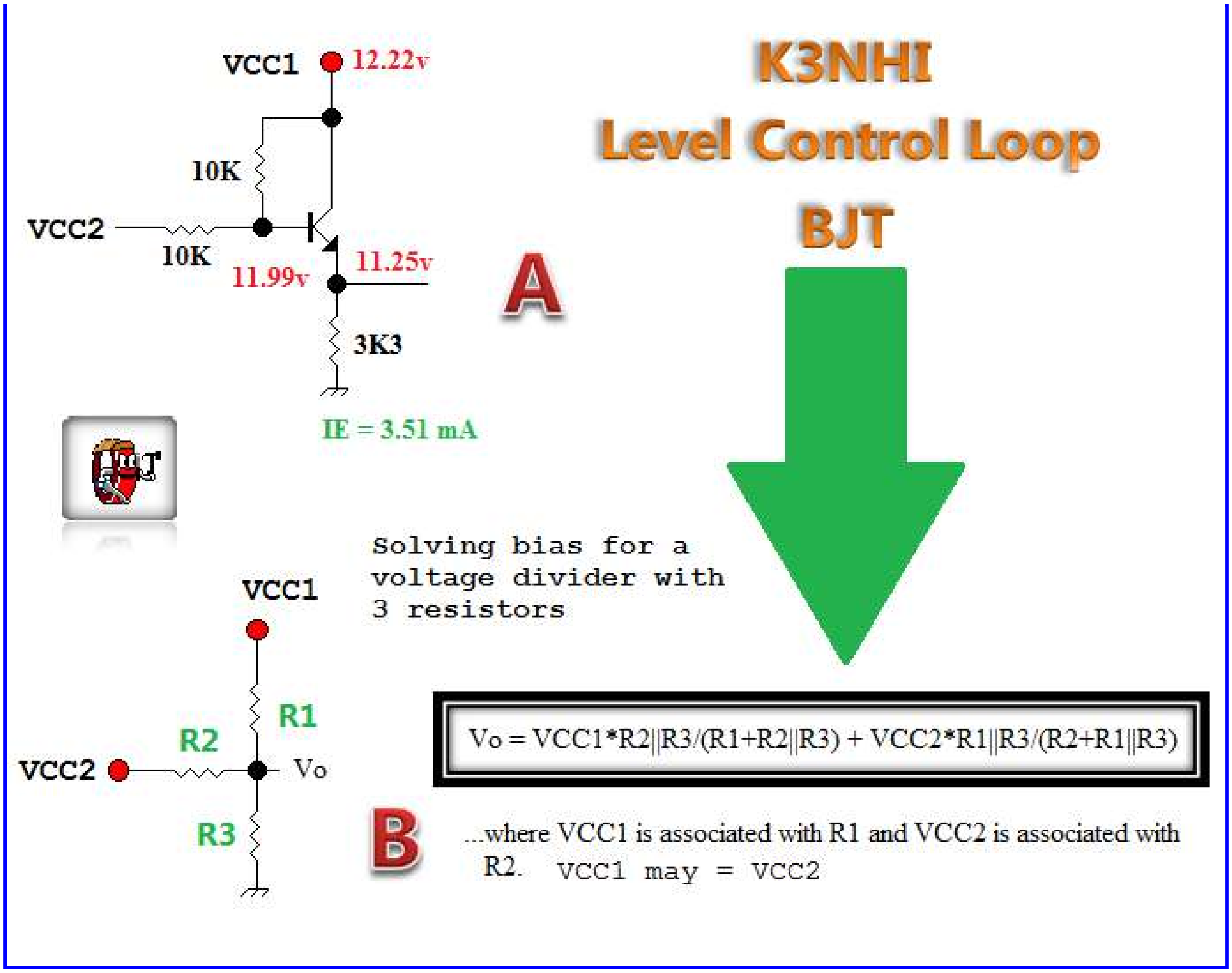

K3NHI
Level Control Loop
BJT

Solving bias for a voltage divider with 3 resistors

$$V_o = VCC1*R2\|R3/(R1+R2\|R3) + VCC2*R1\|R3/(R2+R1\|R3)$$

...where VCC1 is associated with R1 and VCC2 is associated with R2. VCC1 may = VCC2

Above — In the K3NHI 1-118 MHZ VCO lies a fabulous leveling circuit that involves a CMOS rail-to-rail op-amp controlling a BJT level shifter. To understand the bias of this BJT, I made a breadboard (A) and then a simple model (B) and developed the equation shown that involves 3 bias resistors. Bias or Vo is determined by superposition. VCC1 and VCC2 may be equal or not. VCC2 comes from the op-amp. I also learned it's a good thing to surround myself with smart people.

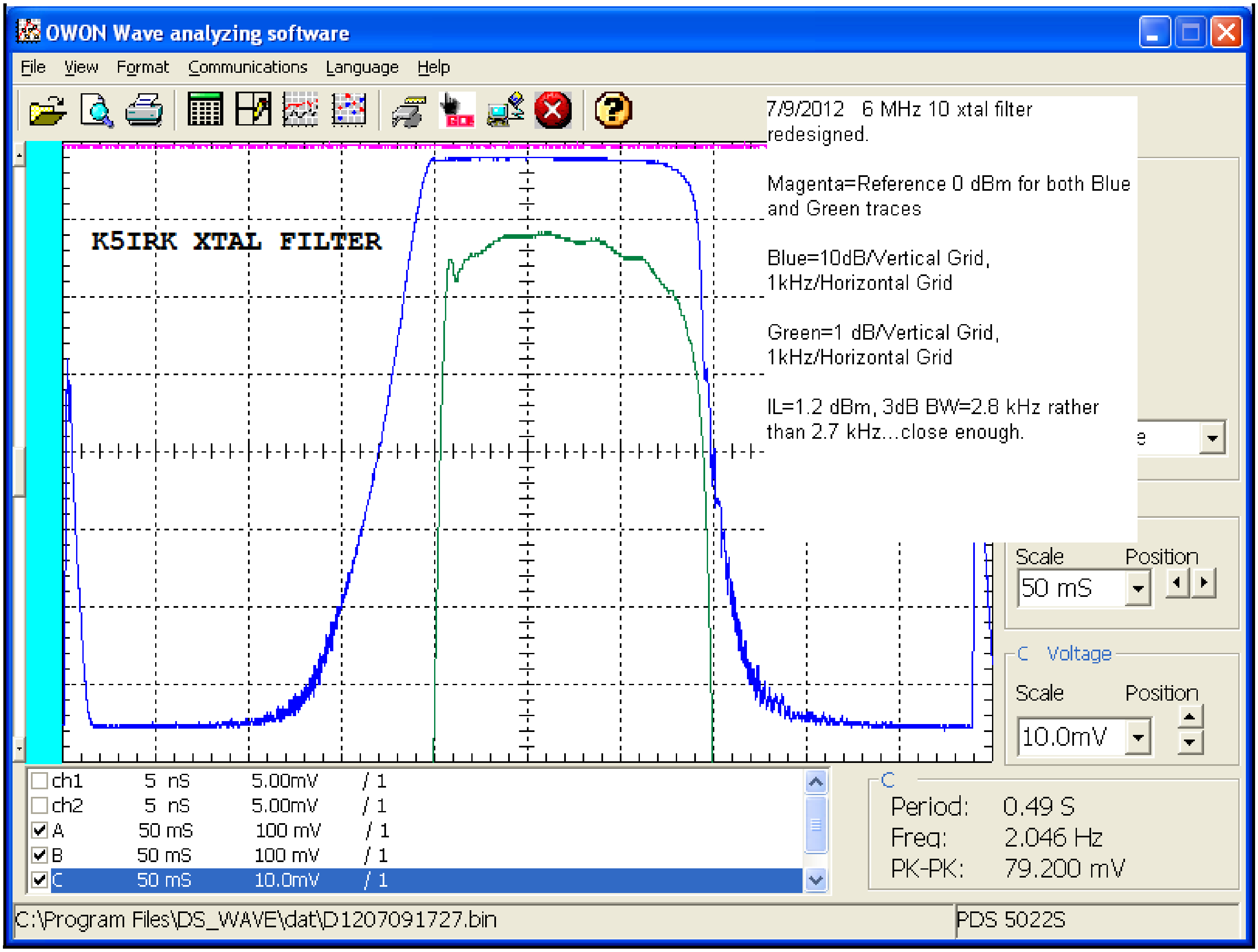
OWON Wave analyzing software
File View Format Communications Language Help
K5IRK XTAL FILTER
Scale Position
50 mS
C Voltage
Scale Position
10.0mV

7/9/2012 6 MHz 10 xtal filter redesigned.

Magenta=Reference 0 dBm for both Blue and Green traces

Blue=10dB/Vertical Grid, 1kHz/Horizontal Grid

Green=1 dB/Vertical Grid, 1kHz/Horizontal Grid

IL=1.2 dBm, 3dB BW=2.8 kHz rather than 2.7 kHz...close enough.

☐ ch1	5 nS	5.00mV	/ 1
☐ ch2	5 nS	5.00mV	/ 1
☑ A	50 mS	100 mV	/ 1
☑ B	50 mS	100 mV	/ 1
☑ C	50 mS	10.0mV	/ 1

C

Period: 0.49 S

Freq: 2.046 Hz

PK-PK: 79.200 mV

C:\Program Files\DS_WAVE\dat\D1207091727.bin

PDS 5022S

Above — John, K5IRQ designed and swept a 10 crystal SSB filter with an insertion loss of 1.2 dB in the K3NHI sweeper that I just had to show you. Click for his build up in Ladbuild08 and then GPLA08 from the EMRFD CD. Bob's xtal sweeper from Section 1 and Wes' software arm us experimenters with solid tools. Nothing can replace measurement and reason.

RF — Test and Measurement

VHF — Véronique

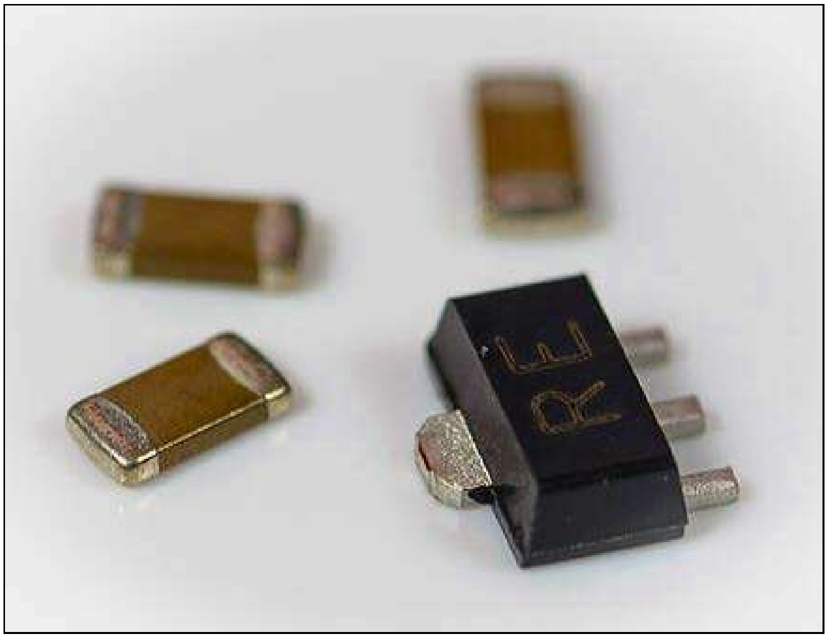

Repository for VHF experiments conducted in 2013 and 2014.

1. Ferrite beads
2. BFS17A transistor
3. 50 Ω MMIC Bench Amplifer
4. Breadboarding Double-Sided Copper Boards
5. Simple Preamp Filter
6. VHF Band-pass Filter Experiments
7. NE612 Mixer Diddy
8. Miscellaneous Photographs

---

1. Ferrite Beads

I toiled to choose an SMT ferrite bead for my junk box. Selecting suitable chip beads not only troubles us, but even some engineers I've read — which to choose? Chip beads act as high frequency resistors that present a low impedance to DC plus our desired RF signal while impeding and dissipating UHF through resistive losses caused by minute heating of the ferrite material from eddy currents. SMT beads are typically rated by their maximum current, resistance at DC, and the impedance they present at 100 MHz.

Since a ferrite bead's impedance is essentially resistive to parasitic UHF oscillations, I remembered that the model for a bead is actually an inductor paralleled by a resistor. It's the "resistor" we're after, for that's the extra load that tends to stabilize the amplifier. From past success in UHF supression with low-value (10-51 Ω) resistors, I chose my SMT bead to have a relatively low Z at 100 MHz. Consider, too, that many chip beads sold today serve as RFI suppressors for high speed digital lines where the Z at 100 MHz and maximum current lie well above my requirements.

I found a chip bead that appealed to me on eBay and bought 50: size 0805; 800 mA; with a Z at 100 MHz of 120 Ω.

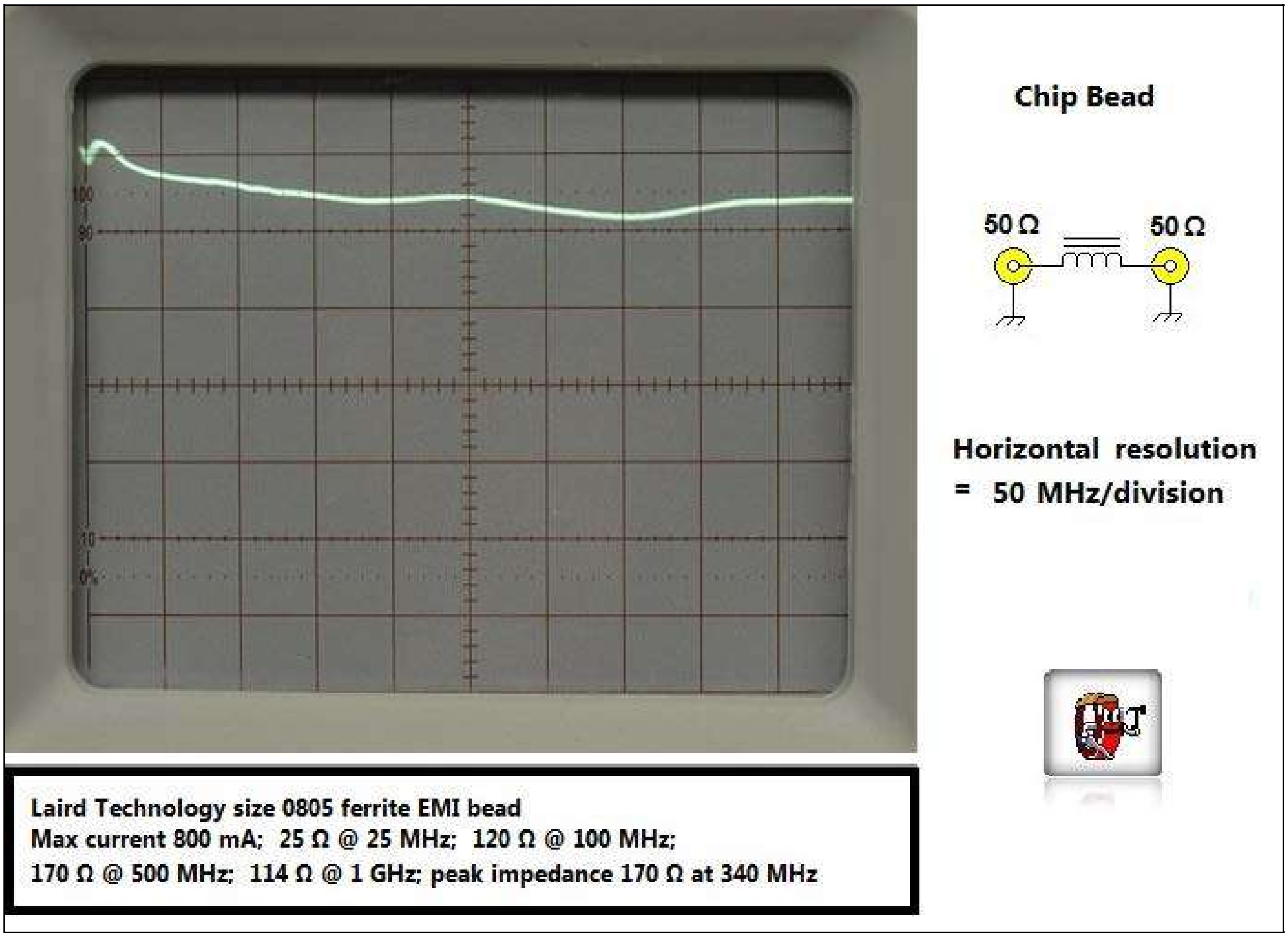

Laird Technology size 0805 ferrite EMI bead
Max current 800 mA; 25 Ω @ 25 MHz; 120 Ω @ 100 MHz;
170 Ω @ 500 MHz; 114 Ω @ 1 GHz; peak impedance 170 Ω at 340 MHz

Above — A sweep of my bead from 1- 500 MHz. Sadly, I did not perfectly center the sweep at 250 MHz, but the peak impedance occurred close to the specified 340 MHz.

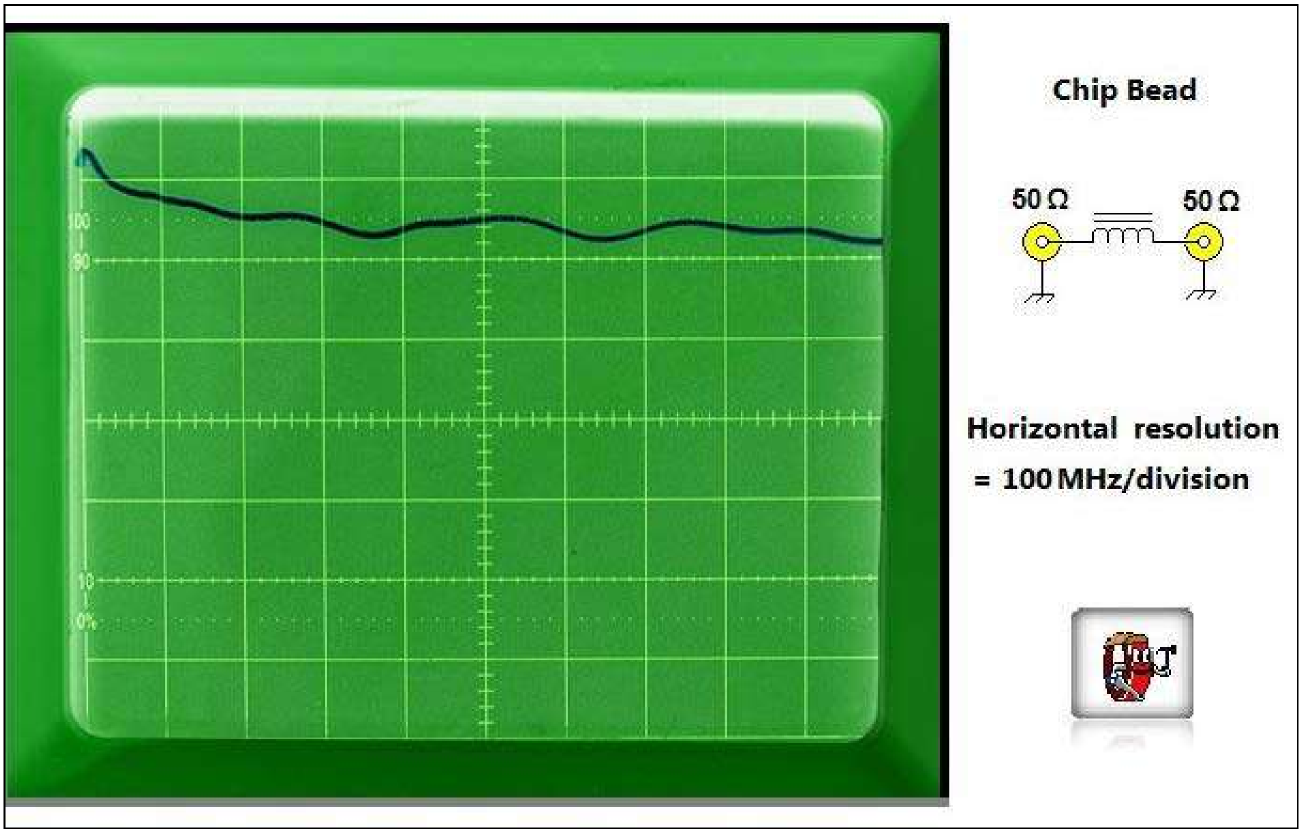

Above — I swept my Laird techlogies chip bead out to 1 GHz. Ultimately, I'll have to verify its function with in-situ experiments. For example, before and after chip bead analysis of a low noise amplifier that's oscillating at UHF.

FB43-101

If you sweep the floor under my QRP work bench, you'll find 2N3904s, J310s and a few FB-43-101s in the dust pan. In our hobby, FB-43-101s are common as fleas on a dog; many builders like to stick them on the drain, collector, or base/gate lead(s) of an active part to squash UHF oscillations. I've never seen a datasheet for this part and decided to sweep 1 for reference purposes.

2 sweeps of the FB-43-101 follow:

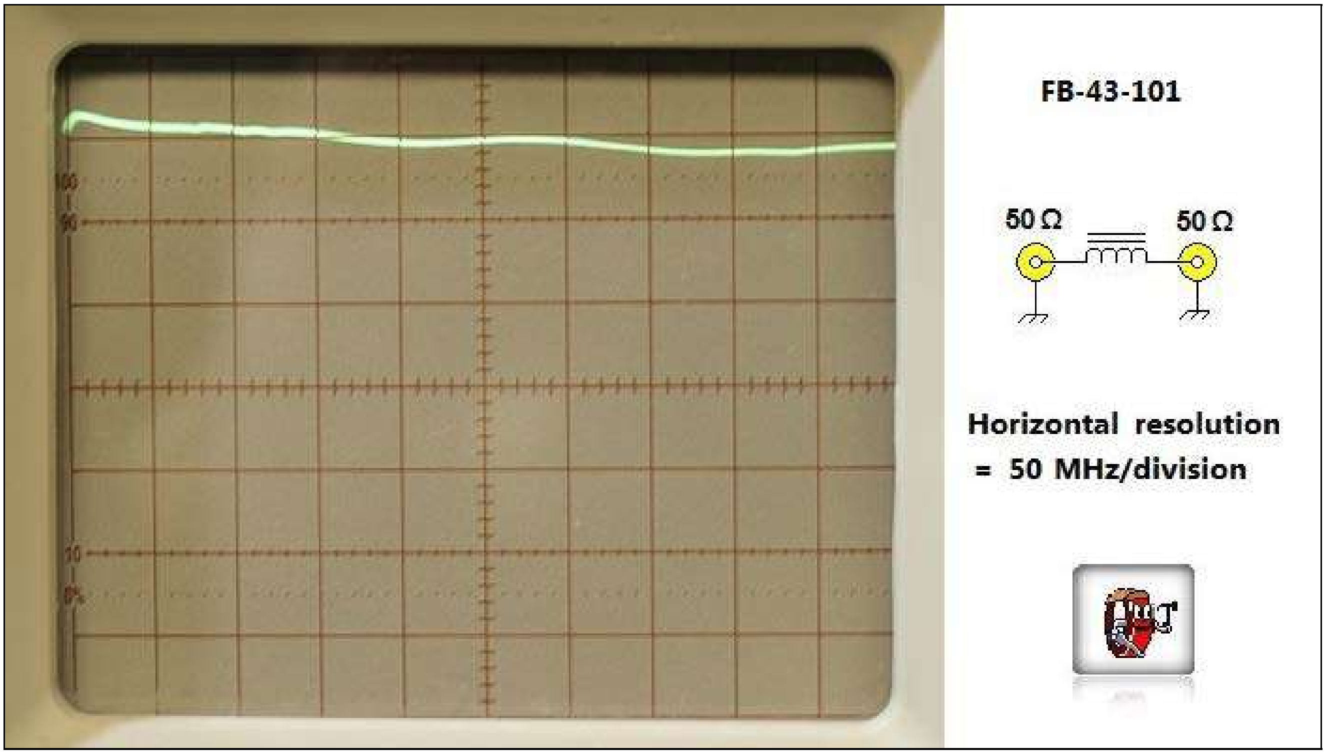

Above — A 1 - 500 MHz sweep of the FB-43-101 slide-on bead. This tracing shows very subtle attenuation that peaks between 400 and 450 MHz. Probably this minimal attenuation at UHF provides the reason we often need to put 2 on the drain or collector of our amplifiers?

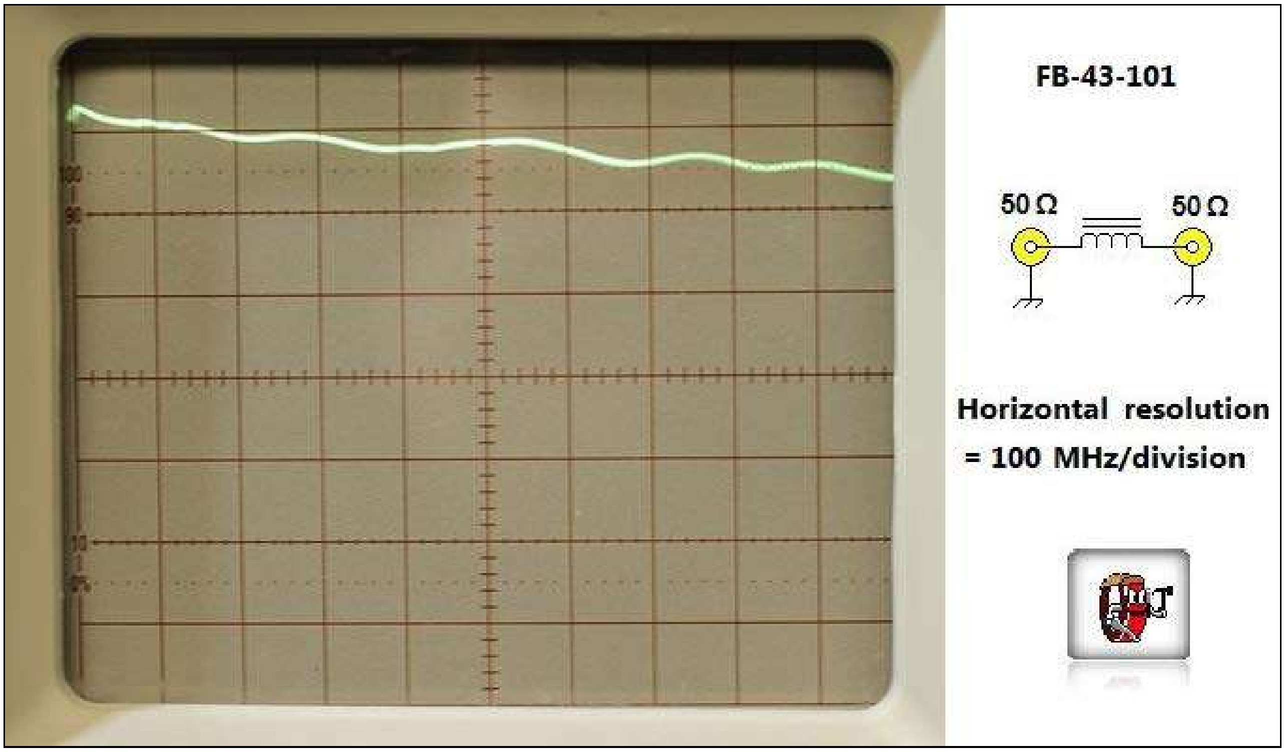

Above — A 1 - 1000 MHz sweep of the FB-43-101.

---

2. BFS17A Transistor

I sought a surface mount NPN transistor to supplant my leaded transistors such as the PN5179, BF199 or MPSH10. Serving as a general purpose transistors biased for ~5 to 20 mA collector current — they'll work as buffers + oscillators + as the BJT for hybrid-cascode amplifiers built with a SMT J310 on the bottom.

My requirements were SOT-23 (3 well *seperated* leads allow easily carved islands in a copper board with a dremel tool); an fT that's not too high to help reduce higher UHF oscillations; and a noise figure < 3. Other factors included price and availability. We enjoy many choices with such criteria, but I settled on the BFS17A. Click for the data sheet . In summary, it's a 2.8 GHz part with a NF of 2.5 dB at 800 MHz.

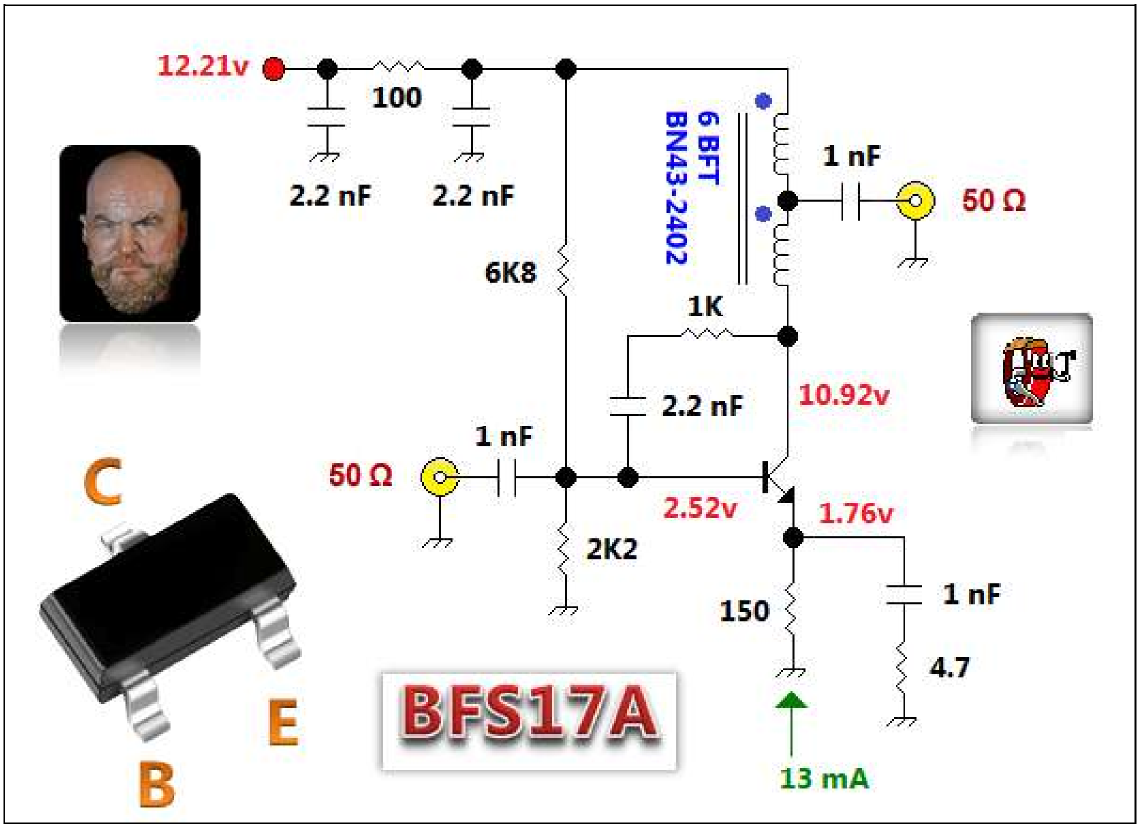

AAbove — My test schematic; a simple feedback amp with capacitors oriented to 50 MHz and above. A limited selection of SMT resistors constrained my experients, but the results seem okay. I strove for the best possible input and output return loss at 50 MHz. I wanted a emitter current of ~ 12 mA and got 13 mA with the bias resistors on-hand.

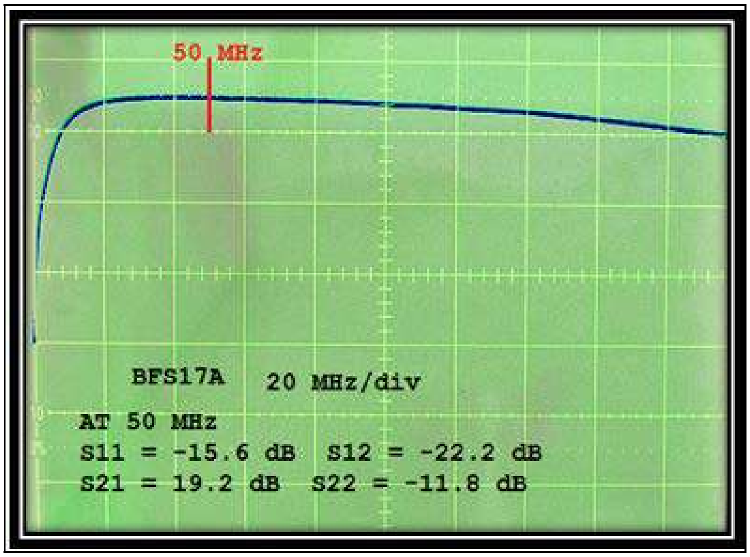

AT 50 MHz
S11 = -15.6 dB S12 = -22.2 dB
S21 = 19.2 dB S22 = -11.8 dB

Above — A 200 MHz wide sweep of my breadboard with tracking generator + spectrum analyzer. Please view the various 2-port measurements at 50 MHz in the lower half. Despite some input and output port mismatching, a gain of 19.2 dB was measured at 50 MHz.

I wound my output transformer within a BN43-2402 binocular ferrite, but ordered some #61 material equivalents for assessment at VHF.

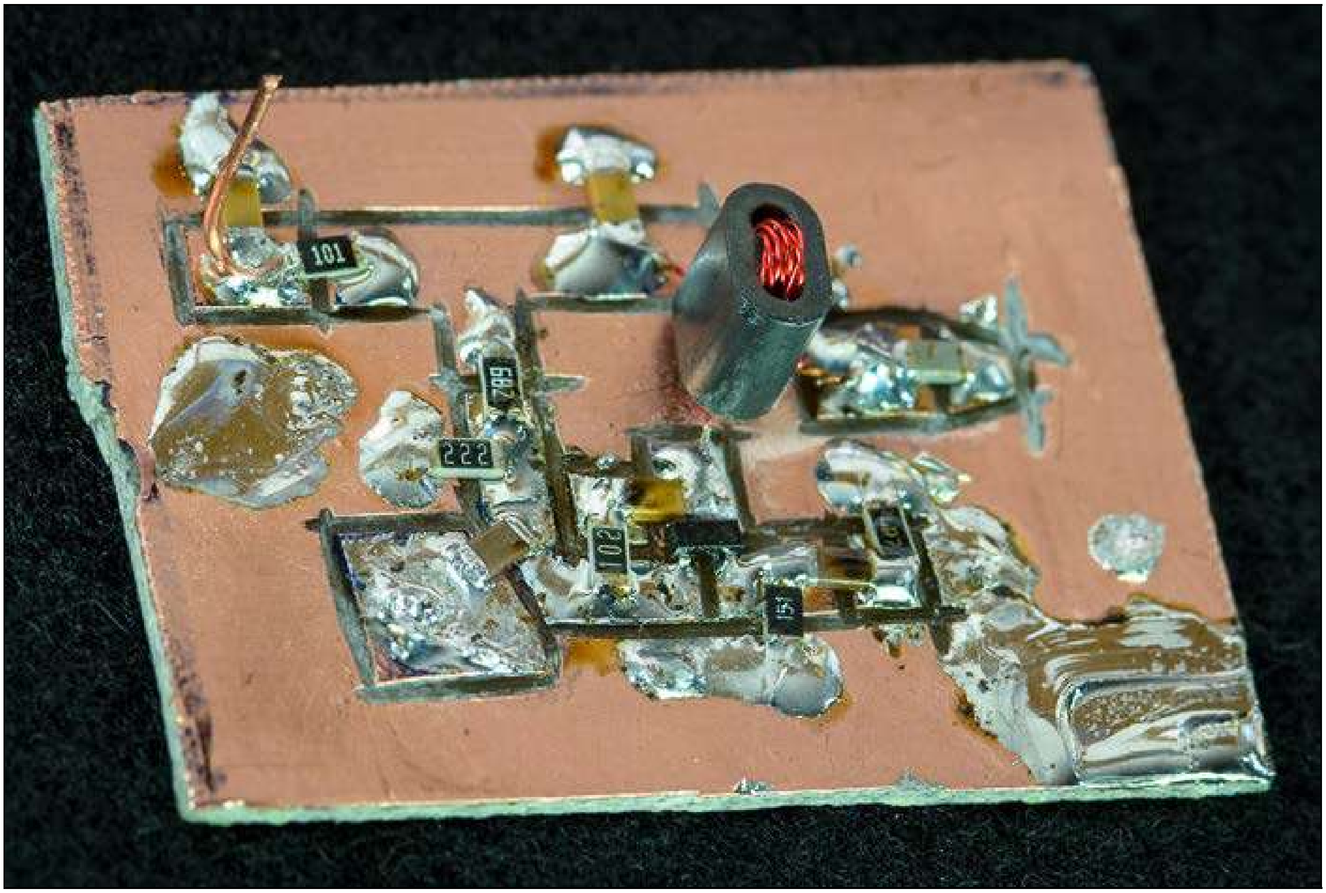

Above — My magnified breadboard looking messy after many parts substitutions from experiments to find the best possible S11 and S22. I removed the shiny RF connectors prior to photography since they often create blown highlights. Alternate image. Despite a limited RF bypass before and after the 100 Ω decoupling resistor + crude bread boarding, I measured no parasitic oscillations from 0.001 to 1 GHz — likely due to the heavy shunt feedback. I liked this transistor, found it easy to breadboard and ordered 100. You'll see the BFS17A in my experiments hereafter.

By all means, order whatever transistor *you* want. Past emailers wrote to say they enjoy reading how other experimenters think about and assess their parts such as RF beads or transistors.

Surface Mount Part Soldering

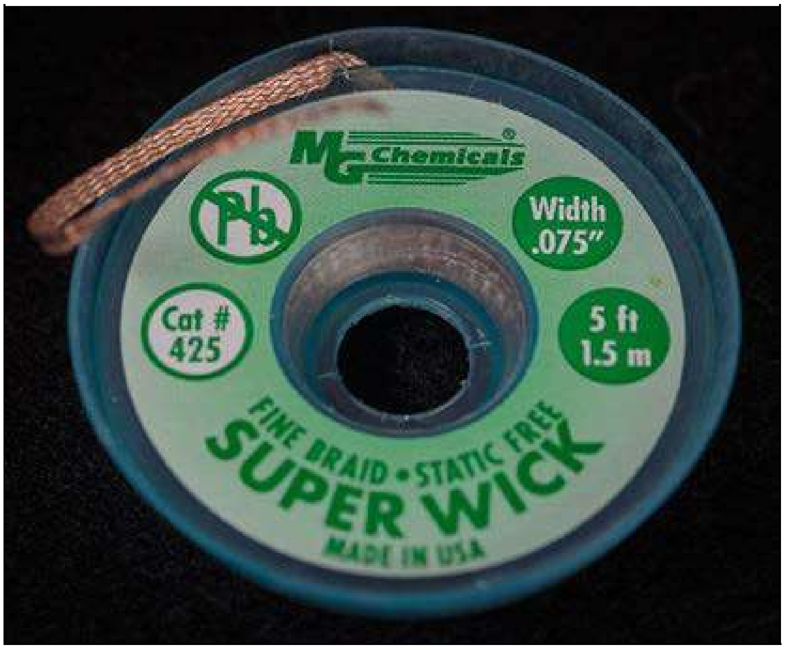

Above — Solder wick. On my SMT workbench, wick proves the most important tool period. A piece of desoldering wick can transform a monster mistake into a perfect circuit in under 2 seconds.

I've read many accounts of how builders anchor the parts they're soldering. We need to keep the SMT part flat on the board to avoid mechanical tension and poor solder joints — a system is required. For size 0805 and 1206 parts, I often just use my index finger fingernail to hold the part in place while I solder 1 lead with my other hand. After a few seconds of cooling, I solder the other lead and then usually touch up the first joint. For tricky formats like SOT-23, or SOT-143, I usually tape down the part while ensuring even pressure on all leads and just solder 1 lead. After cooling, I remove the tape and easily solder the other leads. Again, the intial lead may require some retouching.

SMT Size

Above — A 5X magnified roll of 50 BFS17A in SOT-23. Compact or what?

Above — BFS17A in SOT-23

---

3. 50 Ω MMIC Bench Amplifer

To boost low-level signals during experiments, every 50 Ω workbench needs a utility RF amp module. With strong wideband gain + RF port return loss (S11 and S22), MMICs make a good choice. Since I own 15 pieces, I built an amp around the MCL MAR-3, a modest gain + noise figure, 2 GHz plastic part in the Micro-X package. We enjoy a bevy of MMICS to choose from and as they increase in fT and gain, so does the need for solid, low inductance grounding techniques.

IN
12 V
OUT
MMIC

Above — The completed amplifer module in a Hammond box with an RCA jack for the DC voltage.

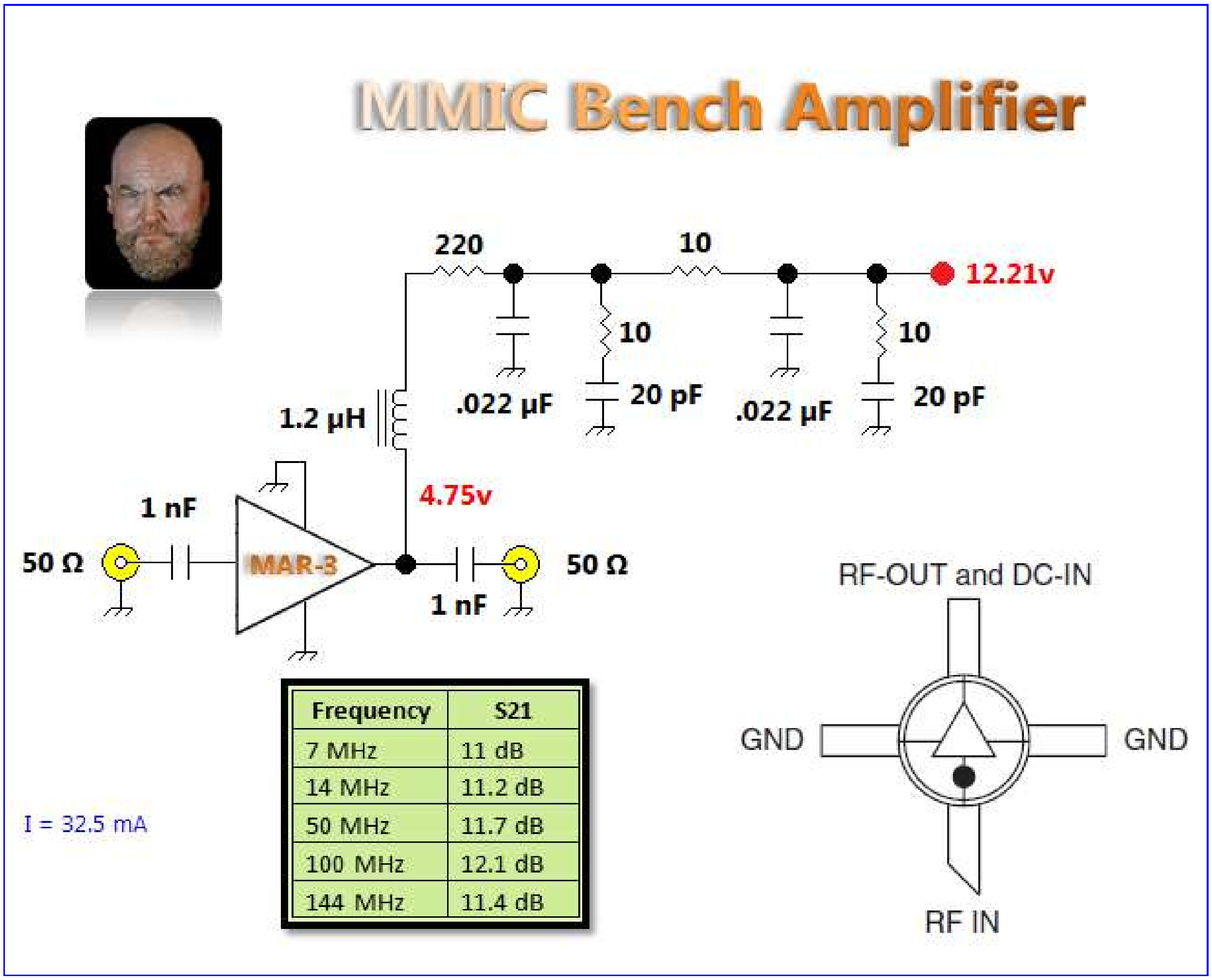

Frequency	S21
7 MHz	11 dB
14 MHz	11.2 dB
50 MHz	11.7 dB
100 MHz	12.1 dB
144 MHz	11.4 dB

Above — Schematic with a table showing gain versus frequency. I built all SMT (size 0805 mostly) except for *short* leaded 1 nF input and output coupling capacitors. I applied sturdy wideband DC filtering since I can't risk RF flowing on my DC lines during bench experiments. I measured no oscillations from 100 KHz to 1.5 GHz.

I chose a 1.2 µH SMT choke in an attempt to boost gain from ~66 to 150 MHz. Many builders leave off the choke, but some RF will flow through the 220 Ω bias resistor to ground and thus, to get maximum gain, a high impedance RFC might help. Recall this RFC should exert an inductive reactance of at least 500 Ω (>= 10 X the load impedance) at the lowest amplifer operating frequency; in my case this occurs at ~66 MHz. A better choice for the RFC might be 1.6 µH which exerts ~ 500 Ω at 50 MHz, but I just own size 0805 1.0 and 1.2 µH chip inductors.

My particular amp delivers ~10 to 12 dB gain from about 5 to 180 MHz. Since the coupling capacitor value determines the low frequency response, if you want a MMIC amp for mostly HF, try reducing your input coupling caps from 0.001 to 0.01 µF or so. You mighty increase the RFC to say 10 µH as well — that's why we don't just build kits; design to suit *your* needs.

The 220 Ω bias resistor should really be 200 Ω for a VCC of 12.2v, however, I had to substitute the 220 Ω R due to a low selection of SMT parts. This dropped the current and also my gain at 100 MHz = 0.4 dB less than that specified in the MAR-3 datasheet — no big deal.

I built the bread board on 2-sided FR4 copper clad board. I soldered thin copper foil around all 4 edges to adjoin the top and bottom copper surfaces. Around each MMIC ground lead, I drilled 4 via holes and soldered copper wire from top to bottom. 2 vias were positioned near every bypass capacitor and all along the input and output paths. 23 total via wires.

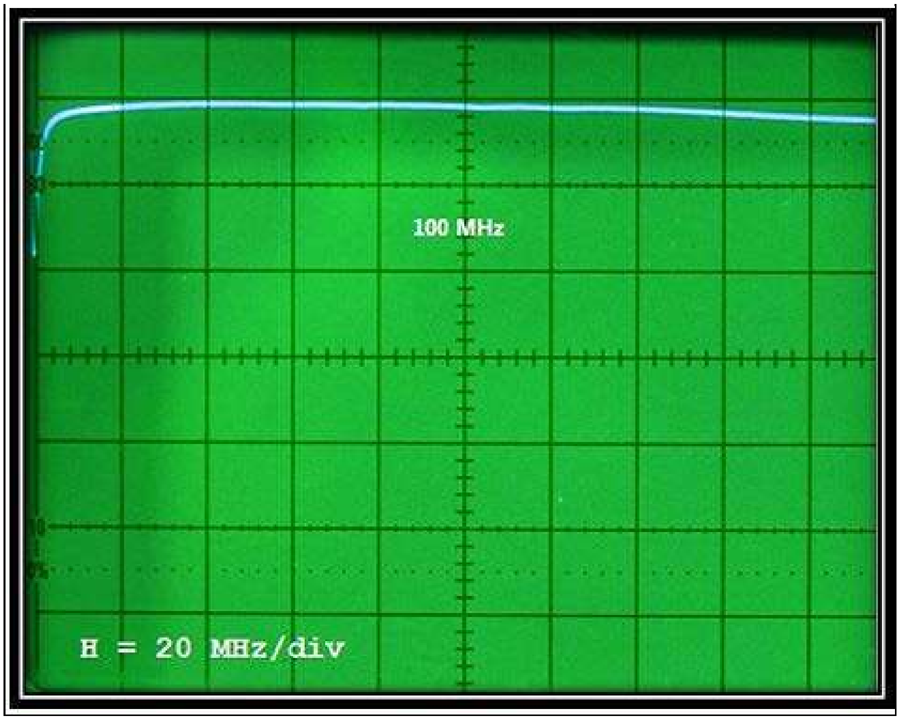

Above — Tracking generator plus spectrum analyzer sweep out to 200 MHz. Click for a sweep from 1 to 10 MHz showing a rapid fall off below 5 MHz. This little amp will serve me well for most HF and especially my VHF experiments. A UHF amplifier module is planned and all knowledge gained from VHF circuit building will flow forward.

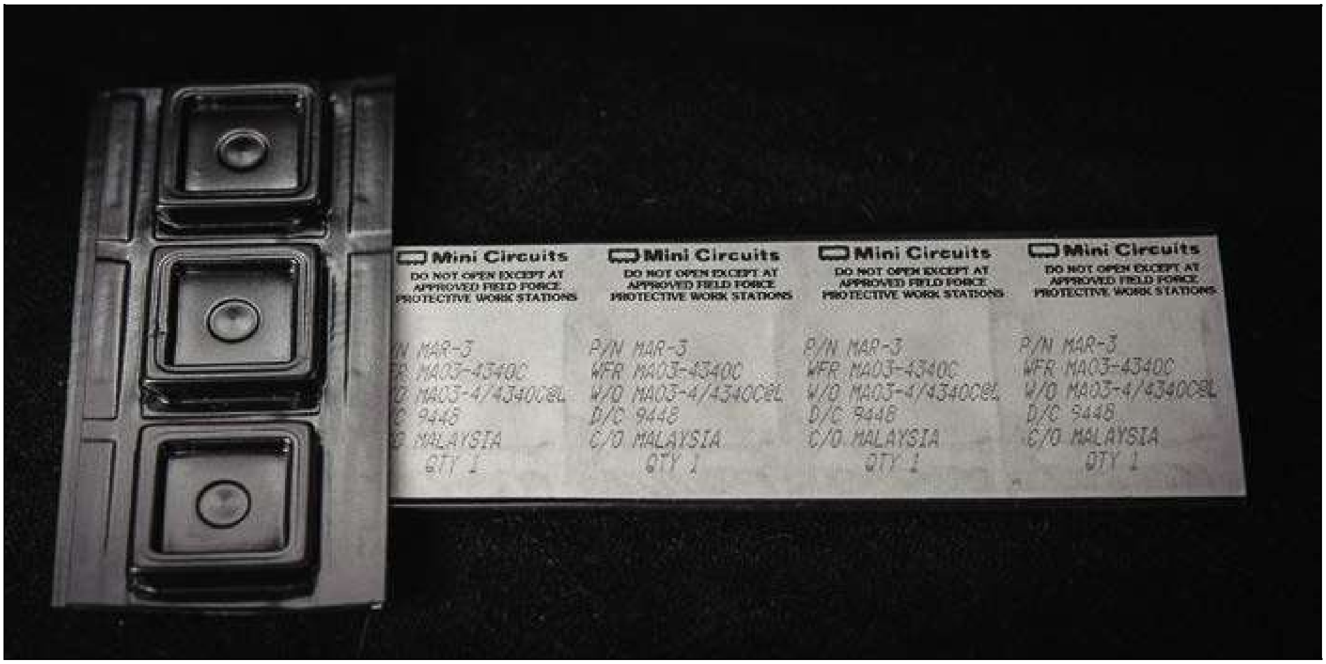
Mini Circuits
DO NOT OPEN EXCEPT AT APPROVED FIELD FORCE PROTECTIVE WORK STATIONS
P/N MAR-3
WFR MA03-4340C
W/O MA03-4/4340C8L
D/C 9448
C/O MALAYSIA
QTY 1

Above — Part of a strip of MCL MAR-3 amplifiers.

Feedback Amps (FBAs)

MMICs like the MAR-3 use a Darlington feedback pair. Still, too, we shouldn't write-off discrete component FBAs wielding transformers + shunt and series feedback at VHF. These amps; staples of countless W7ZOI and W1FB projects since the 1970s; evolved commercially into GaAs FBAs built on a tiny wafer by companies like TriQuint Semiconductor.

In broadband amplifiers, negative feedback permits a wideband (flat) gain response, reduces input and output VSWR (S21 and S22) and controls performance changes from S-parameter and other variations from transistor to transistor. In 2013, I needed an amp chain with >=25 dB gain at 144-165 MHz [~150 MHz mostly] for low level stuff : -25 to -10 dBm input power. Inspired by Wes' 144 MHz CW/DSB transmitter FBA chain *(**SSD** -- Chapter 8 -- Figure 30 in Solid Stage Design for the Radio Amateur; ARRL published in 1986 and out of print)*; I cascaded 3 PN5179 BJTs as FBAs:

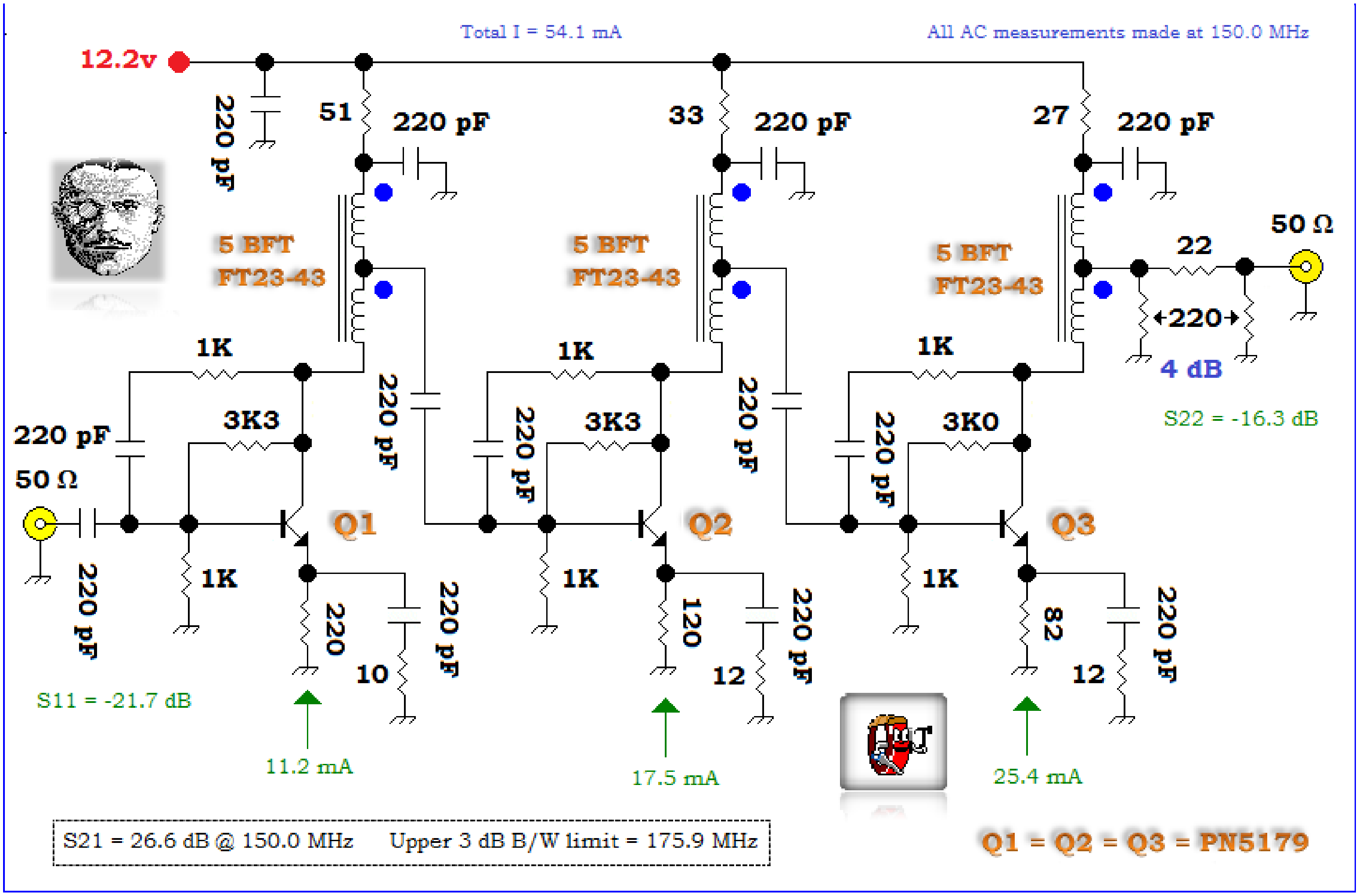
Total I = 54.1 mA
All AC measurements made at 150.0 MHz
12.2v
220 pF
51
220 pF
33
220 pF
27
220 pF
5 BFT
FT23-43
5 BFT
FT23-43
5 BFT
FT23-43
50 Ω
22
220
4 dB
S22 = -16.3 dB
1K
1K
1K
220 pF
220 pF
220 pF
220 pF
3K3
3K3
3K0
220 pF
220 pF
50 Ω
Q1
Q2
Q3
220 pF
1K
220
10
220 pF
1K
120
12
220 pF
1K
82
12
220 pF
S11 = -21.7 dB
11.2 mA
17.5 mA
25.4 mA
S21 = 26.6 dB @ 150.0 MHz
Upper 3 dB B/W limit = 175.9 MHz
Q1 = Q2 = Q3 = PN5179

Above — My 150 MHz (design frequency) FBA module. Since the fT of the PN5179 = 900 to 2000 MHz @ a 100 MHz test frequency, gain drops steadily as we move up above 100 MHz. Centered at 150 MHz, my upper 3 dB drop off point was 175.9 MHz — thus this amp works okay from 144-165 MHz where I need it.

I built and tweaked each stage to derive the best S11 and S22 . *Click for a snapshot taken of my test of Q1 with a BNC connector attached to both ports — I did this for each stage.* It's a trade off since if you boost the emitter degeneration R, the S11 will improve but the gain decreases. Typical stage S11 = -18.5 dB and S22 = -17 dB. I progressively bumped up the current in each stage to reduce gain compression. I get ~12-13 dBm power out of the whole amplifer before compression/distortion occurs.

Securing the best S11 and S22 for each stages might seem silly, but invoked learning. The S11 and S22 of the *total amplifer* counts the most. An interesting thing happened to S22 – with a RF connector soldered right on the board without the pad, S22 = -23.8 dB. When I added the 4 dB pad and stuck it in the box and then ran a ~ 5mm copper wire to connect an SMA output jack, the S22 decreased to -16.3 dB! At 150 MHz — life differs from 7 MHz.

Placing the 4 dB pad between stage 2 and 3 will stabilize + enhance S22, gives a 4 dB higher 1 dB compression point and output intercept, which may make a FBA chain more useful for experiments like transmit driver applications. I can report I measured no oscillations out to 1.5 GHz and this amp works okay for my intended purpose with gain of 26.6 dB at 150 MHz.

I employed 220 pF caps because the plus the self resonant frequency of my 220 pF Murata caps = 154.6 MHz. Therefore, gain at lower HF is low.

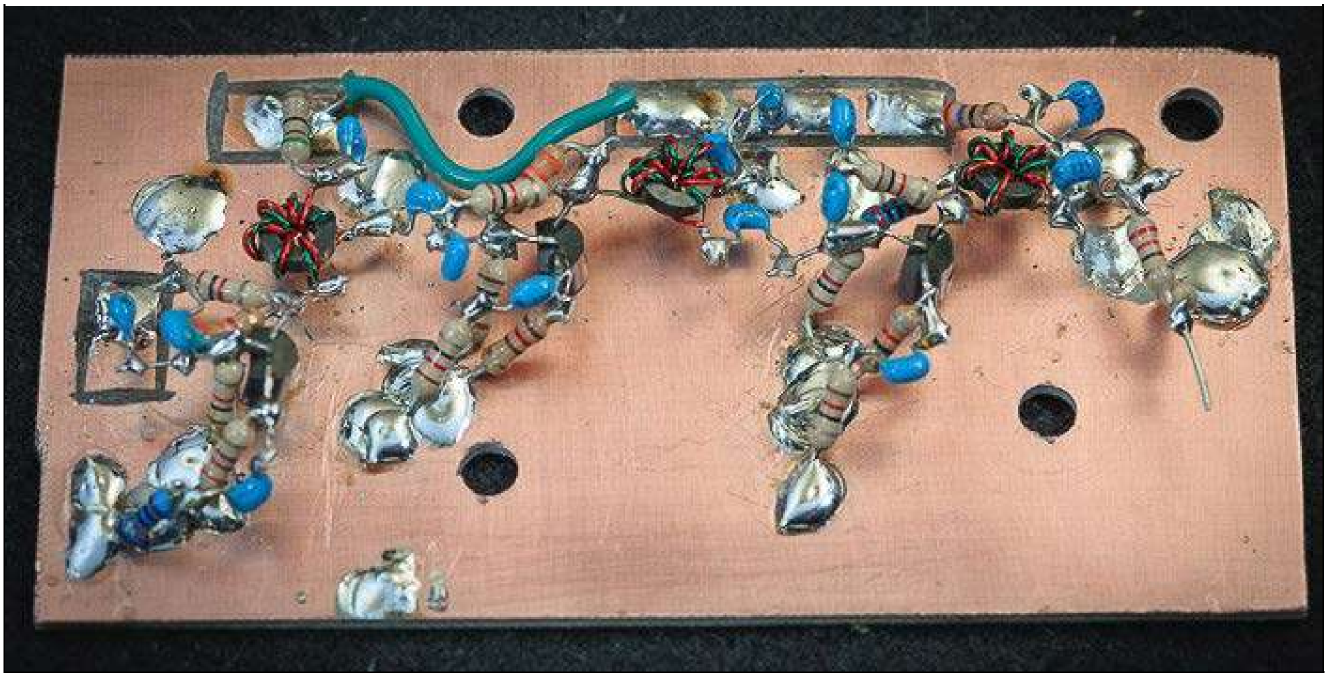

Above — FBA breadboard photograph.

QRP — Posdata for October 27, 2013

I first learned radio electronics from SSD, EMRFD and the ARRL Handbook and still remember the joy when my copy of SSD arrived in 1990. Feedback amplifiers litter this book — and for good reason — they offer stable 50 Ω blocks for building up our RF signal. Wes' 144 MHz CW/DSB transmitter still intrigues me and inspired the circuit shown above.

I asked Wes if he kept that 2M transmitter, and if so, could he snap a couple of photos for me? Wes opened it up to see that the VXO and the frequency multiplier chain were missing. Click Click Click . Thanks to Wes, W7ZOI for the photos.

---

4. Breadboarding Double-Sided Copper Boards

Seeking a low inductance AC ground for some of my VHF and UHF FR-4 prototype breadboards, I join the top and bottom copper surfaces with solder wick along its 4 edges and copper wire vias in the main board area.

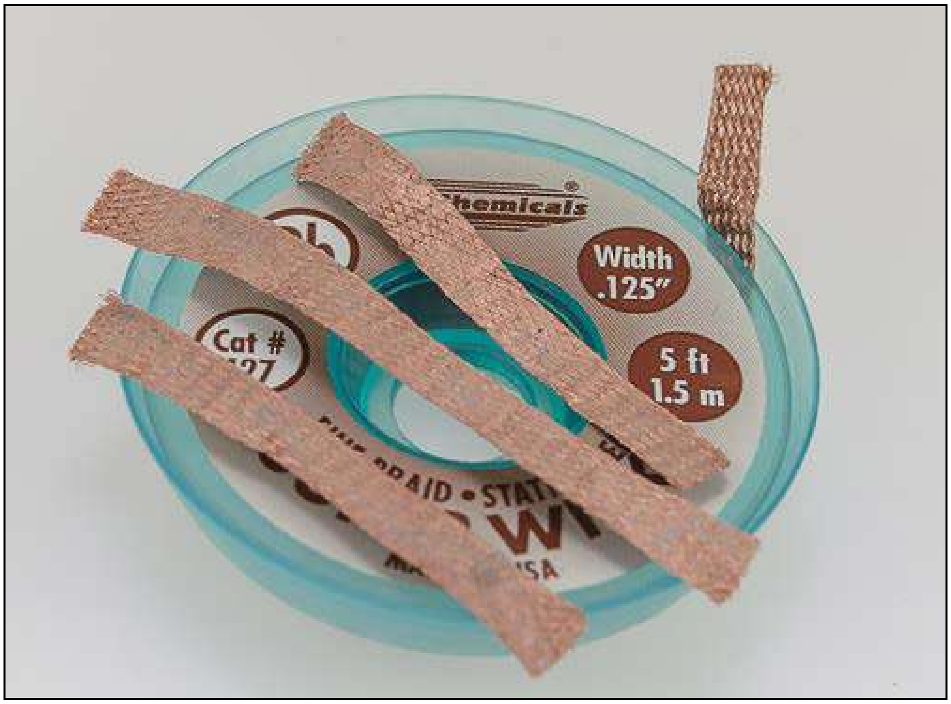

Above — A bench staple, solder wick, joins the 4 edges in true Ugy fashion. Many prefer copper foil for this task, but it's not cheap, nor readily available prompting a pragmatic approach. I take a hammer and pound the solder wick [0.125 inch minimum] so it becomes wider and thinner.

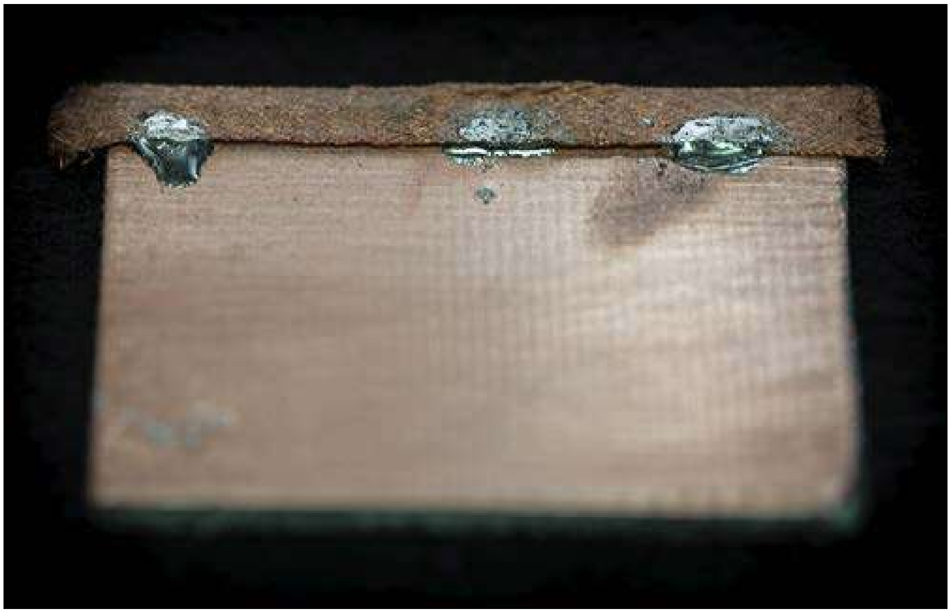

Above — Take a piece of wick and tack it in 2-4 places along the top surface of the copper board. The center solder shown above looks perfect — too much solder will flow into the wick and make a difficult bend job around the board edge. Tack just enough wick so you have enough left to go over the edge and to solder on other side of the copper board.

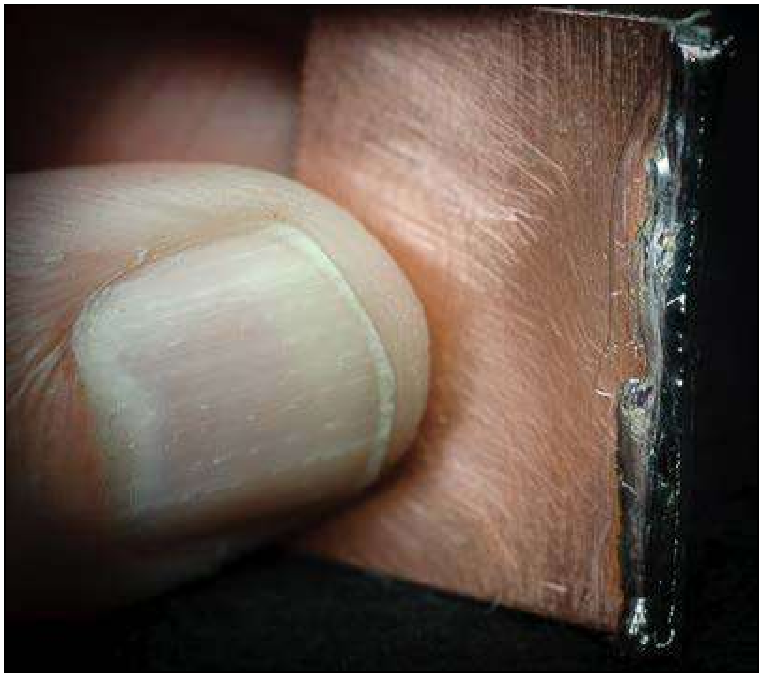

Above — With hand and/or pliers manipulate the wick around the other side to completely bridge the top to bottom. Finally, liberally solder the wick on both surfaces, move to the next board edge and repeat. Click for a rather messy UHF board with copper wick on 2 edges and numerous vias around a prescaler chip, some input + DC lines and a switch.

Above HF, most don't make pretty prototype boards to admire — rather, builders strive for good AC grounding and generally stick their board in a metal box for shielding and this hides the breadboard. No one's ever told me I make pretty circuits for any frequency and I'm okay with it.

Onto wire vias:

DRILLCO
Cutting Tools
Baton Rouge, LA
200A103 13 ~~12~~ Each
SIZE 3/64 JOBBER DRILL
GENERAL PURPOSE
DRILLCO CUTTING TOOLS

Above — Your drill chuck will need to accommodate small bits such as those shown above (3/64 inches = 1.2 mm) . I keep these and others that vary from 0.5 mm to 1.2 mm for making via holes.

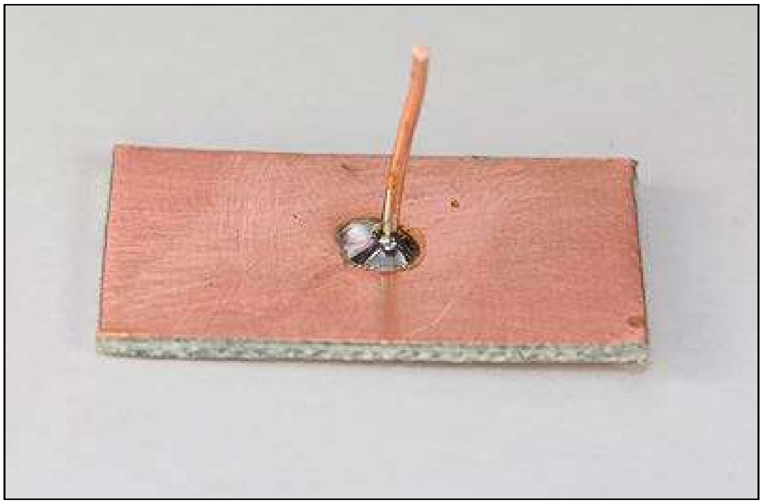

Above — Magnified via wire soldered at the top side. I found making a hole close to the same size of the via wire will hold the wire in place for rapid soldering. This wire is normally clipped flush at this point. Click for a photo of my flush cutters.

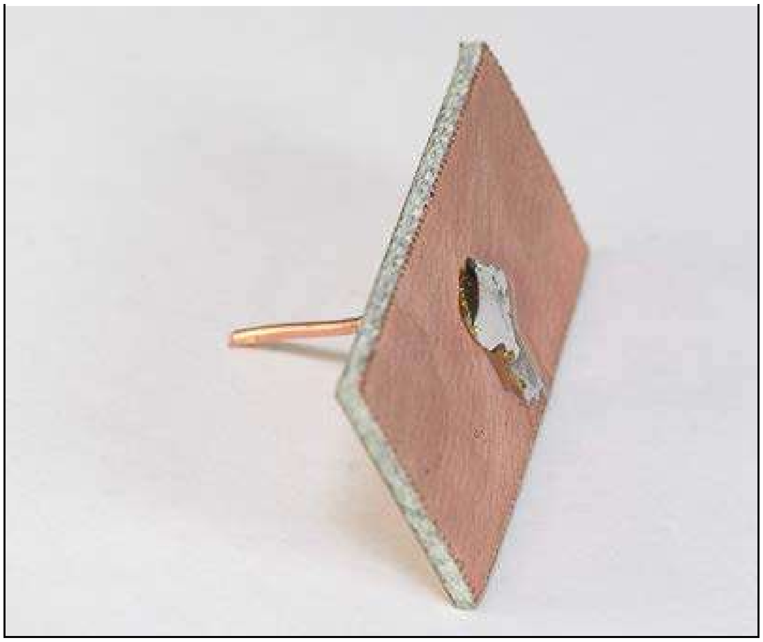

Above — The board shown above was flipped over to solder the bottom copper surface. I try to drill and solder in all vias prior to soldering active devices, since the board heat may damage some parts.

---

5. Simple Preamp Filter

Some RF filtering should proceed our receiver low noise amplifer (LNA) input — but what? Low-pass, high-pass, band-pass, 2 poles, many poles? Factors informing this decision might include our receive frequencies, QRM, selection of IF and thus our image stripping requirements, plus maximum tolerable signal loss before the LNA. We might not want image rejection from this filter since the post LNA filter can tackle this function.

I examined an input filter presented by Joe Reisert, W1JR in the November 1986 Ham Radio. Joe wrote a fantastic column called VHF/UHF World and his context was a high dynamic range 2M receive converter to a 28 MHz IF.

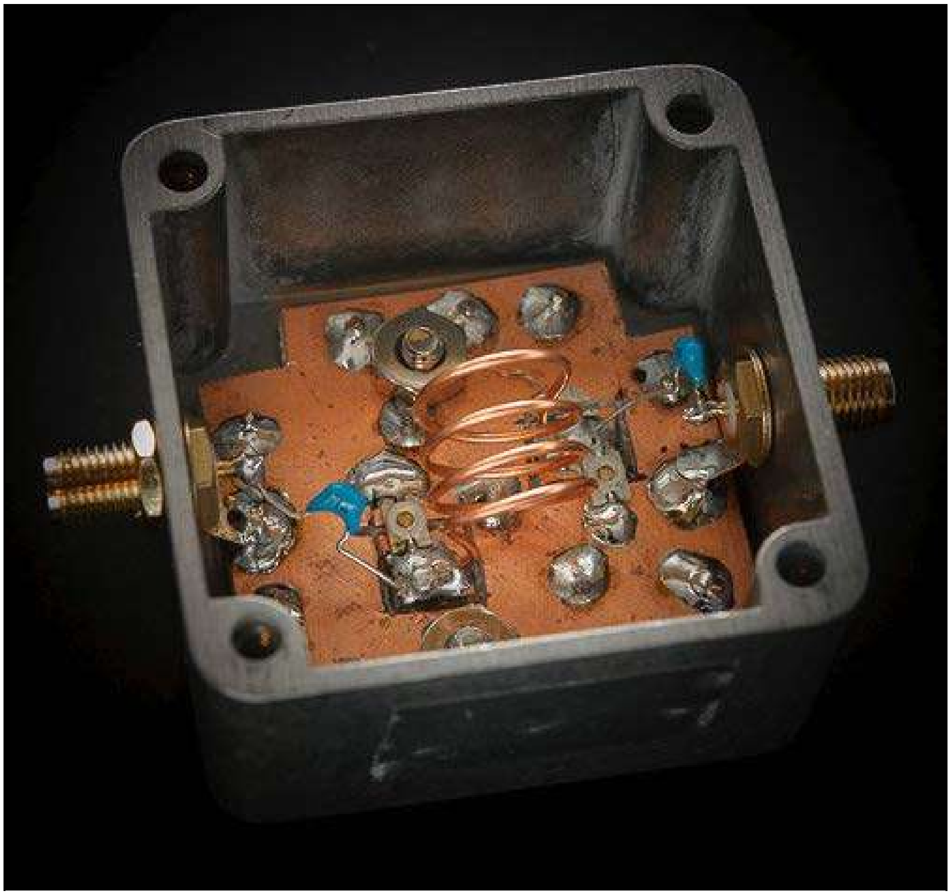

I'm uncertain how to classify this filter: it looks like a standard pi low-pass filter except the input and output series capacitors exert a high-pass skirt. Joe built his filter for 144 MHz. After "building" Joe's filter in Ladbuild08 I tweaked the filter in GPLA *[software from EMRFD]* to center it at our local NOAA weather channel: 162.55 MHz. Click for the GPLA plot.

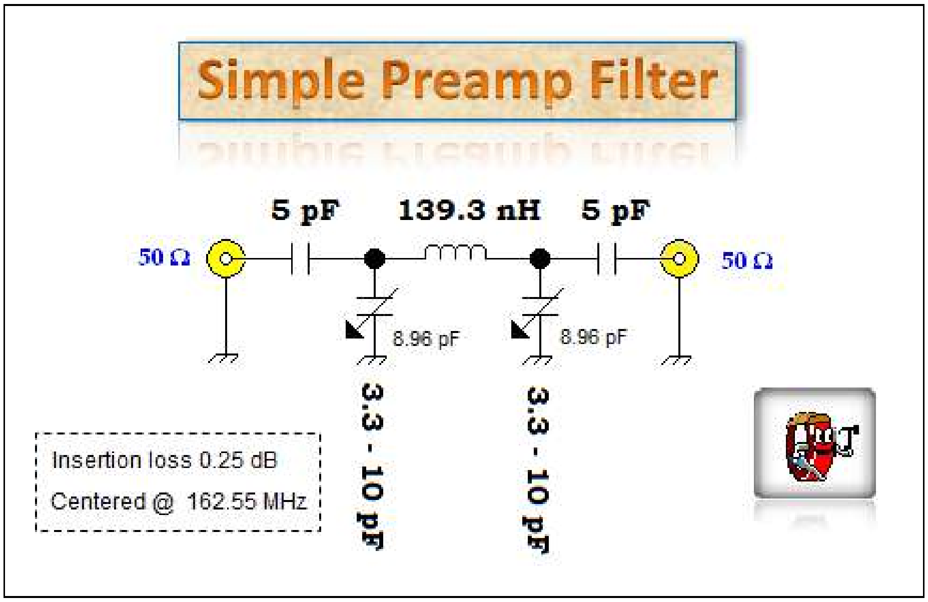
Simple Preamp Filter
5 pF
139.3 nH
5 pF
50 Ω
50 Ω
8.96 pF
8.96 pF
3.3 - 10 pF
3.3 - 10 pF
Insertion loss 0.25 dB
Centered @ 162.55 MHz

Above — Input filter schematic with values shown to center this rather symmetrical filter at 162.55 MHz. The half power (3 dB down) bandwidth = 18 MHz. You can easily scale it to other other frequencies — by tweaking the trimmer caps and perhaps squishing or expanding the inductor links some, this filter will tune widely including the 2 Meter Ham band. The IL blew me away (better than calculated) — the resonator Q was high due to the air wound L and Q >= 700 trimmer caps along with good port matching.

Click to view a larger photo of the breadboard. I wound the coil on a 3/8 inch diameter bolt but then spread the turns to get the L and length I wanted. Built on double-sided copper clad board, a few copper wire vias connect top to bottom. I joined the LC circuit to its ports with leaded 5 pF ceramic caps.

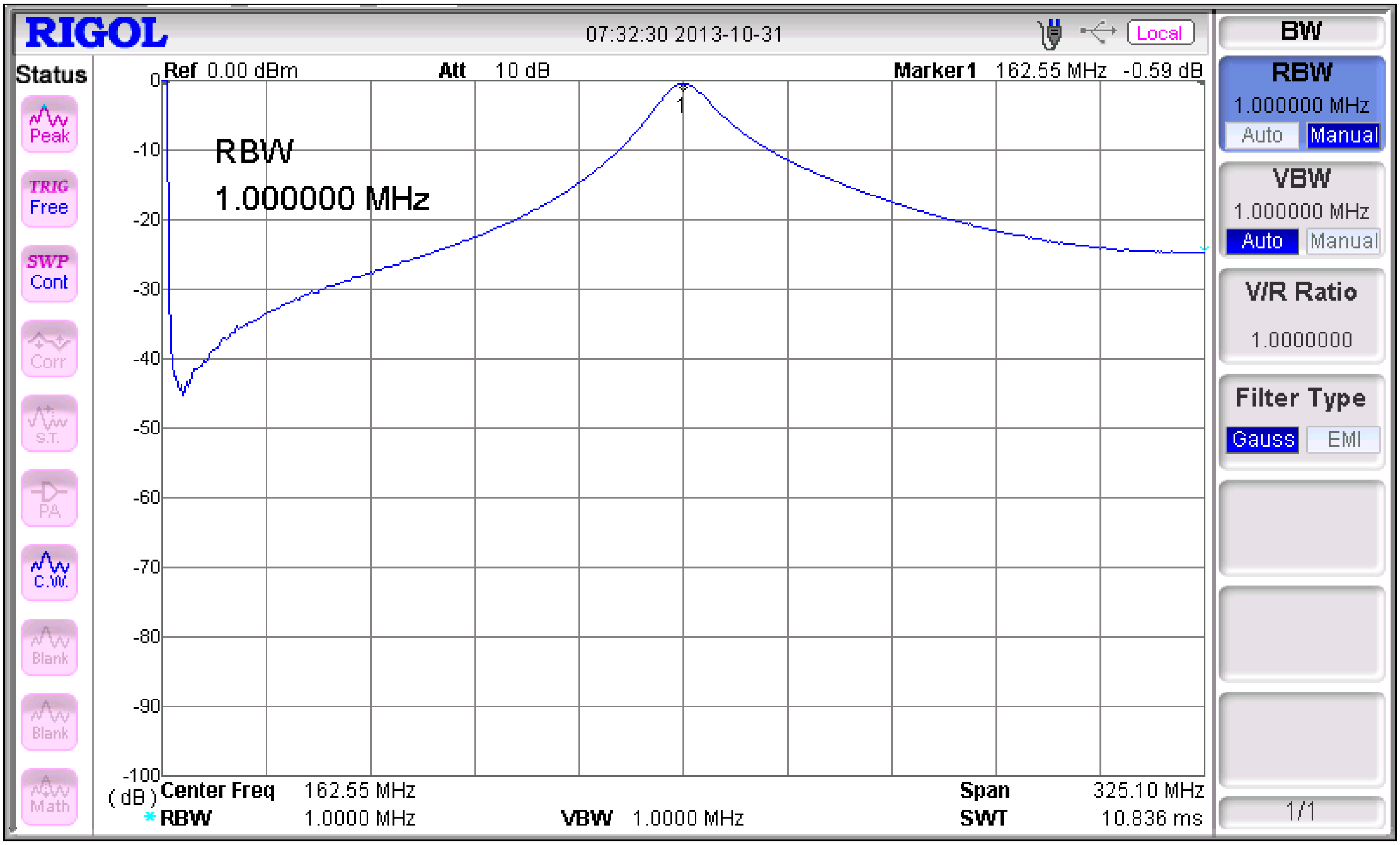
RIGOL 07:32:30 2013-10-31 Local

Ref 0.00 dBm Att 10 dB Marker1 162.55 MHz -0.59 dB

RBW 1.000000 MHz

Center Freq 162.55 MHz Span 325.10 MHz

RBW 1.0000 MHz VBW 1.0000 MHz SWT 10.836 ms

BW

RBW 1.000000 MHz Auto Manual

VBW 1.000000 MHz Auto Manual

V/R Ratio 1.0000000

Filter Type Gauss EMI

Status
Peak
TRIG Free
SWP Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
1
1/1

Above — Tracking generator plus spectrum analyzer sweep. Under sweep set up, I minutely squished or expanded the L and tweaked the caps to peak the filter at 162.55 MHz while obtaining the lowest insertion loss. Further, I had to tweak the trimmer capacitors to center the filter with the Hammond lid on since putting the lid on changed the center frequency slightly. A *TG equipped* spectrum analyzer proves the ultimate bench tool for VHF and UHF in my humble opinion.

Above — Zero calibration of my AADE L/C IIB meter. By sweeping numerous LC circuits after measuring the L with this tool, I've learned it's reasonably accurate with low inductance coils. I zero the meter prior to *each* measurement and zero calibrate it with the alligator clips butted end-to-end while gripping a small piece of copper wire. Try to keep the same relative alligator clip position during coil measurement. I'm uncertain if this is protocol, but it seems to work for me.

Thanks to Joe, W1JR for this circuit and the opportunity to learn more about component-level VHF design and construction.

---

6. VHF Band-pass Filter Experiments

Build Notes

1. All filters designed with the EMRFD ladpac programs DTC08, or TTC08 and then tweaked in GPLA.

2. Seeking a band-pass filter for my experiments with the local NOAA weather channel at 162.55 MHz, I designed, built and tested 3 filters. With a 10.7 MHz IF superhet receiver, the image frequency = 141.15 MHz.

3. My VHF sweep system = a tracking generator plus a spectrum analyzer.

Filter 1

To avoid the need for proper electromagnetic/electrostatic shielding, I kept my HF experimenter hat on and built my first filter with powdered iron toroids (T30-12) knowing I would pay a Qu penalty that may wreck filter insertion loss, bandwidth and port matching. I hoped that the passive electromagnetic shielding offered by toroidal inductors would reduce resonator coupling.

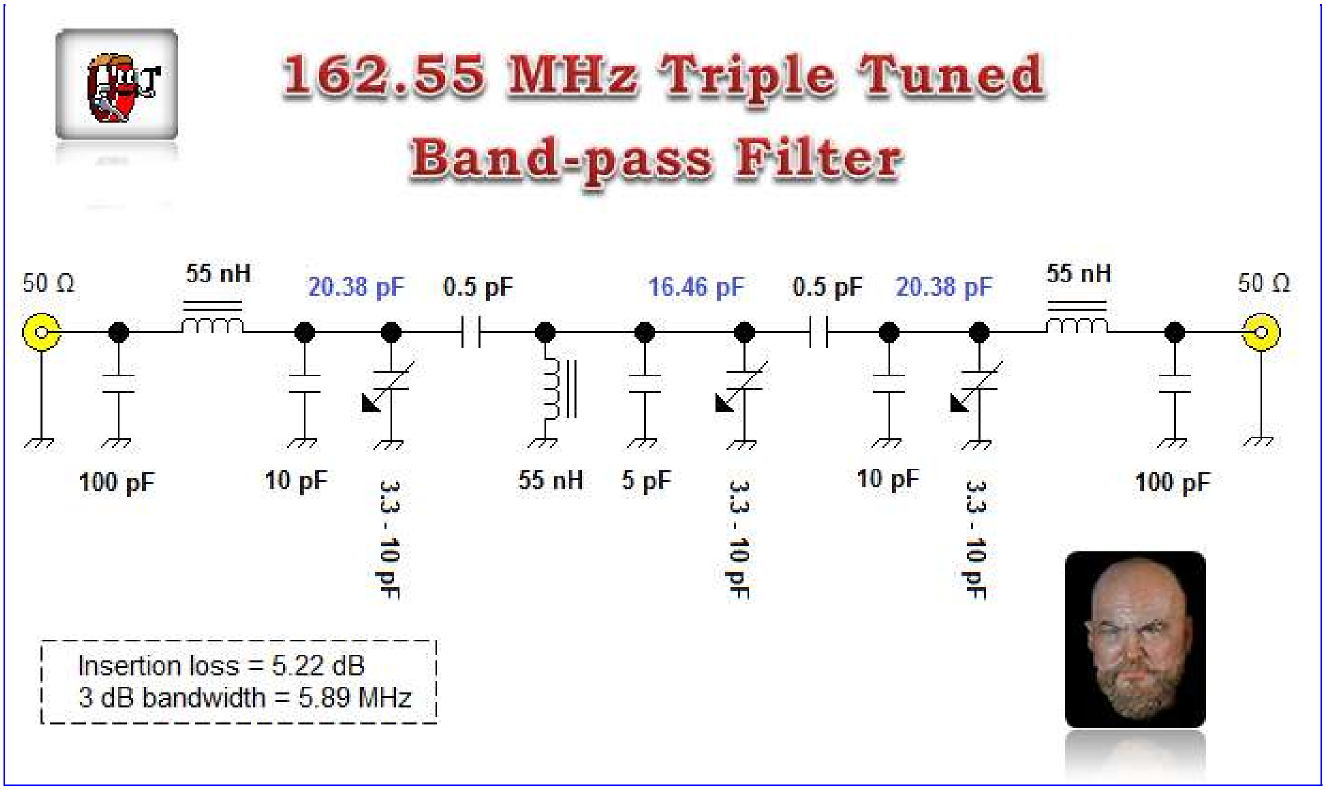
162.55 MHz Triple Tuned
Band-pass Filter
50 Ω
55 nH
20.38 pF
0.5 pF
16.46 pF
0.5 pF
20.38 pF
55 nH
50 Ω
100 pF
10 pF
3.3 - 10 pF
55 nH
5 pF
3.3 - 10 pF
10 pF
3.3 - 10 pF
100 pF
Insertion loss = 5.22 dB
3 dB bandwidth = 5.89 MHz

Above — Mixed-mode triple-tuned filter schematic. Click for the GPLA simulation. When we want to go up in frequency plus desire a narrow bandwidth, we face 2 options: reduce our inductance and/or coupling capacitor(s). This taxes our parts collection and breadboard skills since capacitors less than 1 pF are relatively uncommon and require SMD breadboard techniques. I adjusted my design bandwidth and inductance in TTC08.exe to allow the 0.5 pF capacitors plopped in my parts bin to work. For a change of pace, I chose a mixed-mode filter topology.

I strive to place a single high Q variable capacitor in my band-pass filter tanks, however, my Q = 700 SMD trimmers only ranged 3.3 to 10 pF so I stuck a leaded ceramic capactor in parallel with each trimmer.

Common SMD capacitors exhibit low to medium Q and I try to avoid them in my at HF band-pass filters: often I'll just solder a short leaded though-hole caps instead. I also applied this logic to my VHF filter, however, this might prove foolish at VHF since the self-resonant frequency of a given leaded cap is lower than that of an equivalent SMD cap and capacitor XL becomes more significant as we move up in frequency.

Click for a pdf file showing a simple experiment with 3 capacitors . I'll occasionally use ultra-high Q SMD caps in my future experiments based upon these results.

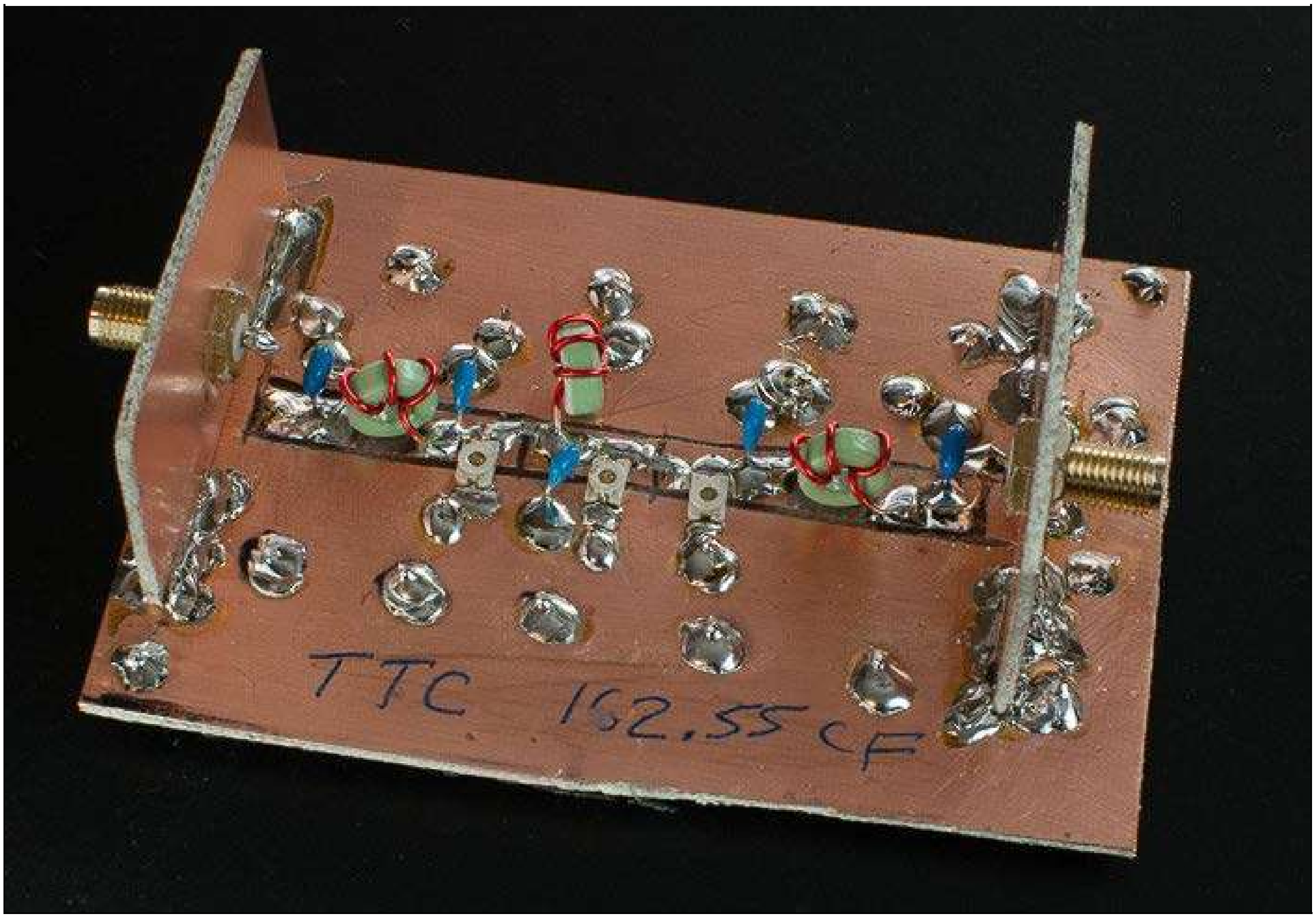

Above — My TTC breadboard. Click for a larger photo. Click for a bench sweep to view a feculent skirt peak in the filter spectogram. Some resonator coupling occurred despite spacing and placing the inductors at right angles. This circuit possessed 6 tweaks: the 3 trimmers and the 3 Ls — I just kept tweaking them until I got the best shape and lowest IL. That's 21 gauge wire on the toroids. My image frequency = 47.28 dB down. Since the sweep yielded such a wide discrepancy from my GPLA simulation, I discarded this filter and moved on to design #2.

Filter 2

Attempting to obtain a resonator Qu of at least 300-400, I moved to air core inductors wound with bare 22 gauge copper wire *plus* air-variable capacitors. Double-tuned classic topology. Without stout shielding, air wound inductors will couple and trash the stopband.

Create With Metal
Quality Metal Sheets
Metal edges may be sharp. Use gloves and safety glasses when handling. Children should have adult supervision.
K·S
6917 W. 59th St.
Chicago, IL 60638
773-586-8503
www.ksmetals.com

Above — Copper and brass sheets I bought for making filter boxes after Wes recommended using metal sheets or strips for shielding filters. In North America, builders can purchase a sheet metal hand brake plus a rotary cutter/shear for under $100 and make all sorts of boxes for radio gear. I'll equip my QRP workshop with such equipment over the next year. This was my first experience man-handling sheet metal or strips.

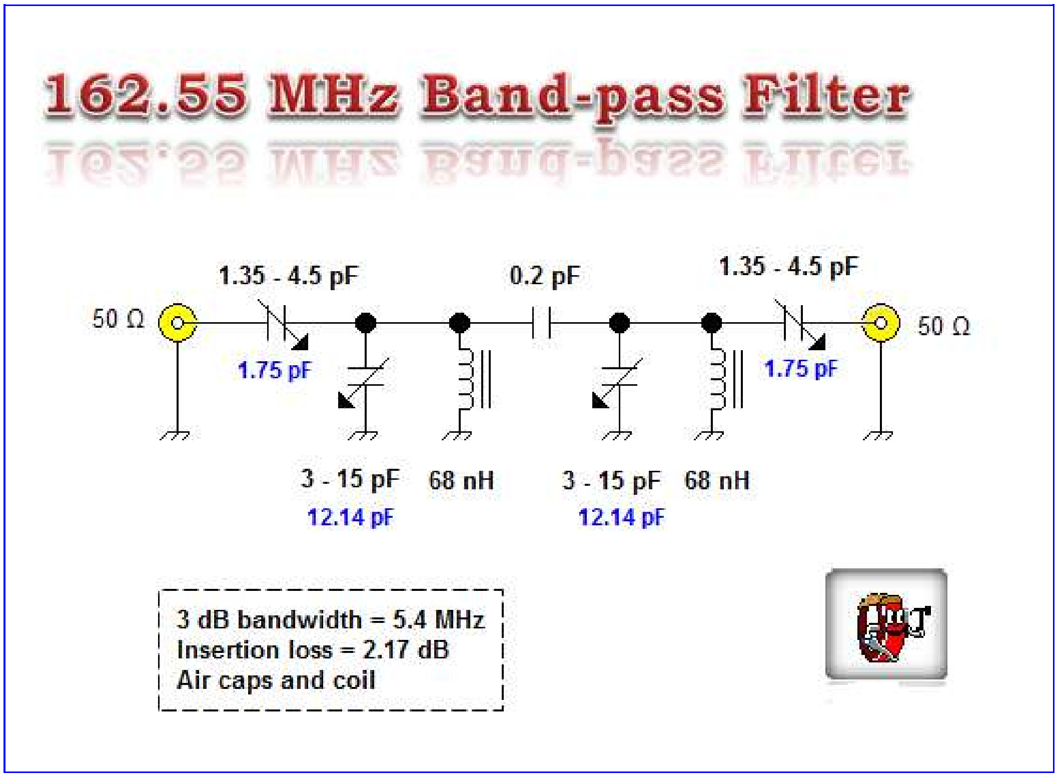
162.55 MHz Band-pass Filter
50 Ω
1.35 - 4.5 pF
1.75 pF
0.2 pF
1.35 - 4.5 pF
1.75 pF
50 Ω
3 - 15 pF
12.14 pF
68 nH
3 - 15 pF
12.14 pF
68 nH
3 dB bandwidth = 5.4 MHz
Insertion loss = 2.17 dB
Air caps and coil

Above — Schematic diagram of Filter #2 with tabled IL and half-power bandwidth. Click to view a macro photograph of the filter. I soldered the brass box onto a double sided copper-clad FR4 bread board laden with some copper wire vias to enhance the ground plane. The 0.2 μF coupling cap = an ATC microwave cap.

Click and click for the sweeps. Notice the splendid skirt symmetry at the filter top and also the fantastic peak attenuation in the low-pass skirt

(~81 dB from the filter peak to the lowest point in the low-pass filter skirt). I'd never seen deeper attenuation on my bench before: quite gratifying. My double-tuned filter had essentially matched the 3 dB BW of triple-tuned filter #1 with less insertion loss.

Not all appeared perfect however; attenuation of frequencies above the passband looked so-so and I was unable to get a lower 3 dB bandwidth by manipulating L and C in each tank.

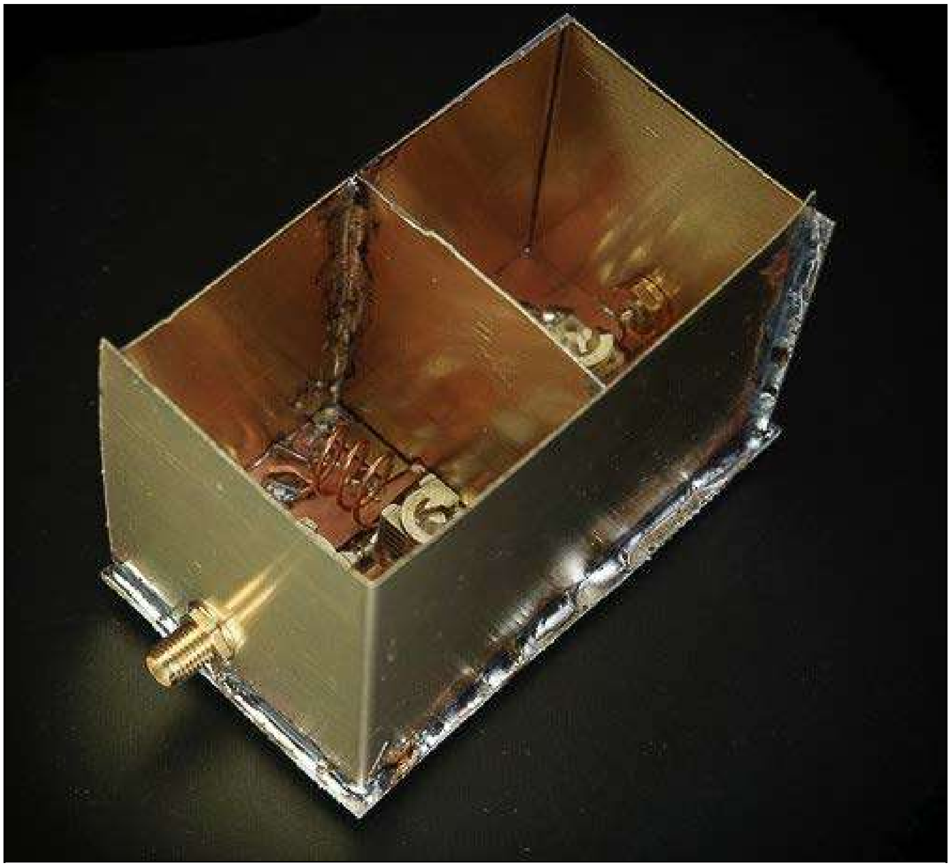

Above — My first ever brass box soldered on a 2-sided copper clad board: Filter #2. Lacking proper tools and also the knack for making beautiful chassis like Dave, AA7EE does, I just did my best. My only concern was getting tank isolation. I chiseled out a small cut in the center brass divider just high enough to clear the SMD 0.2 pF chip capacitor and center copper strip. Overlaps allowed soldering of brass walls After tweaking the L and C parts, I placed a brass lid on top during sweeps. Okay, onto filter # 3:

Filter 3

When coupling caps fall below 1 pF, some builders place their resonators in close proximity to couple the tanks, however, these usually involves slug-tuned powdered iron or ferrite inductors with a generally low Q. The coil consists of cup core and a threaded center core which combine to give a magnetic shield. A tin plated copper can surrounds the core for electrostatic shielding. I think after these experiments — placing each resonator in its own sealed box seems a better option.

I noticed another approach in The Double Tuned Circuit: An Experimenter’s Tutorial in QST for December 1991 by Wes, W7ZOI. Wes drilled a hole in the tank compartment divider and passed a wire soldered to 1 tank through the hole and positioned it near the variable capacitor of the other resonator. We change coupling by adjusting this wire.

The wire may be thought of a "tweakable" capacitor, however, the key point is we need to couple energy from 1 resonator to the next; whether by wire, capacitor or inductor, proper resonator coupling will give the required skirt shape within the the limits imposed by Q and the matching to the 50 Ω ports. When establishing a filter passband, the end Q and resonator coupling pretty much dictate the outcome.

Filter #3 involved more mechanics than electronics. I made a U shaped box from a small sheet of copper.

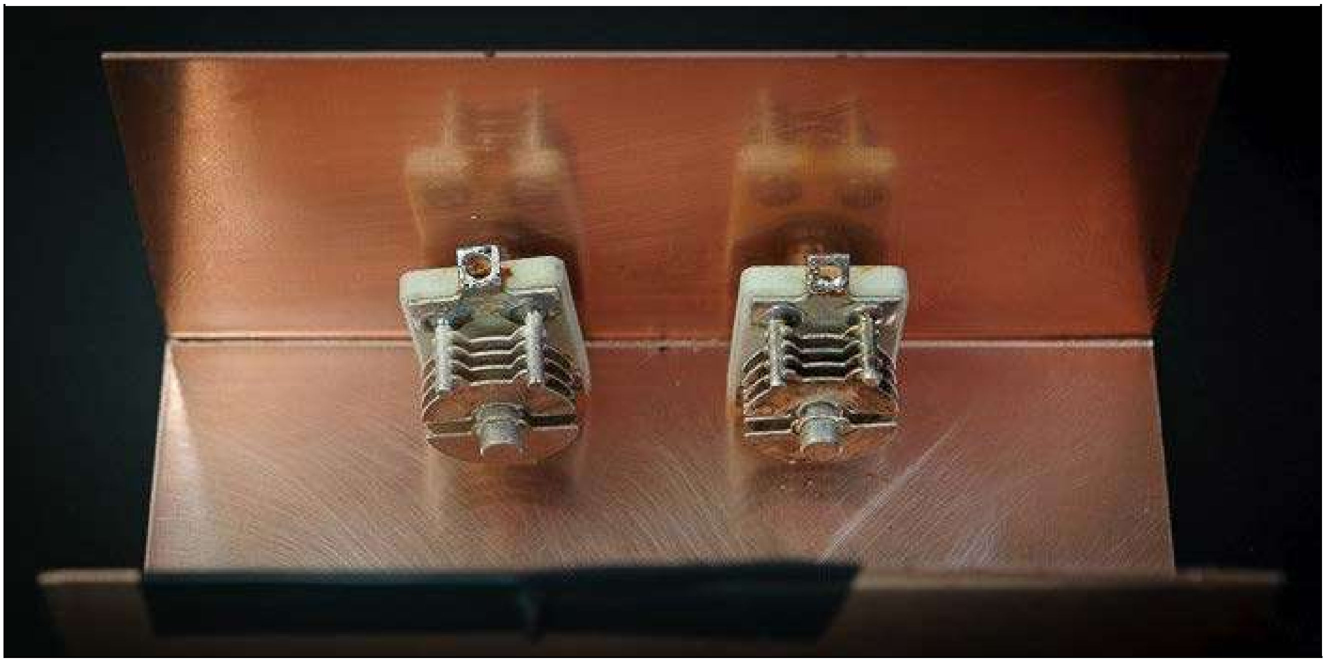

Above — My copper box drilled for the 2 air-variable tuning capacitors. I built the walls, top and divider from thinner gauge brass strips. Rosin flux aids soldering your metal sheets together — I soldered mine with a Weller 80 Watt iron and began with the center divider. Click for a side shot after completion.

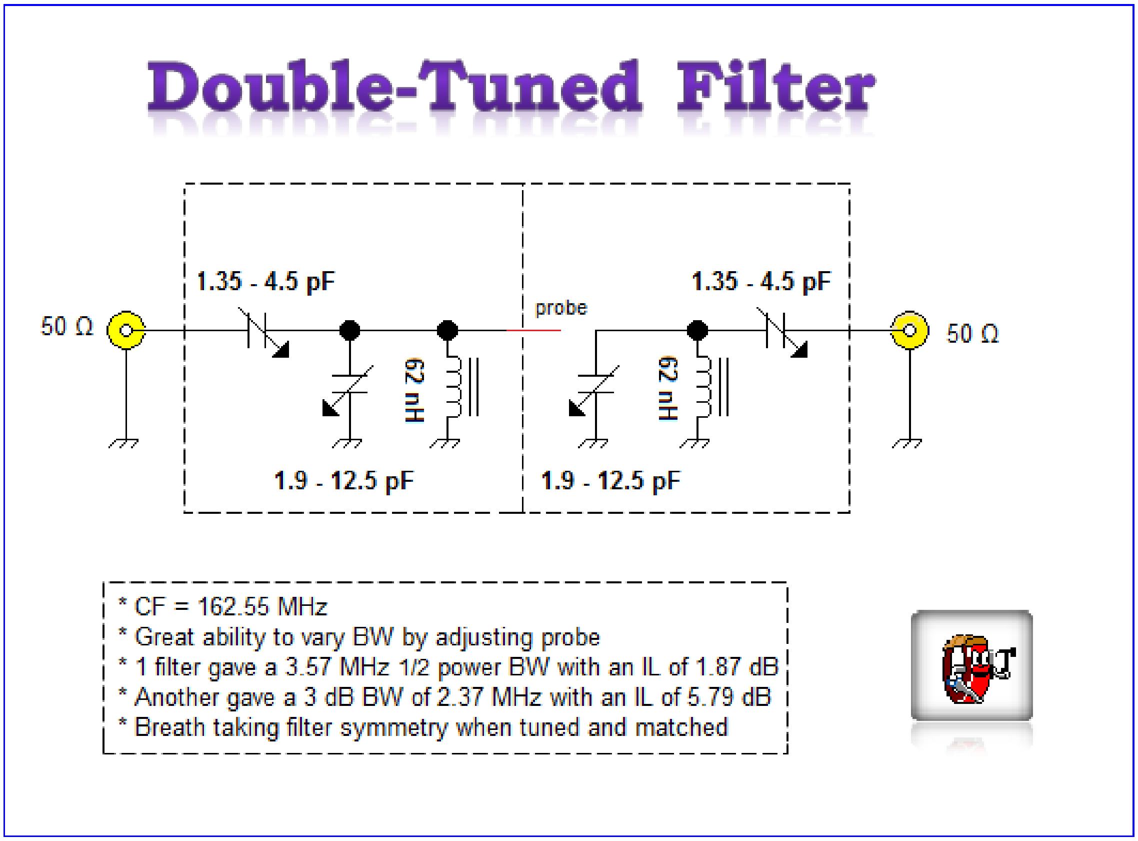

- * CF = 162.55 MHz
- * Great ability to vary BW by adjusting probe
- * 1 filter gave a 3.57 MHz 1/2 power BW with an IL of 1.87 dB
- * Another gave a 3 dB BW of 2.37 MHz with an IL of 5.79 dB
- * Breath taking filter symmetry when tuned and matched

Above — My filter schematic with some tabled data. Click for a top off photograph where you can easily see the black wire probe. To change coupling, you change the wire length and/or its distance from the second tank capacitor. I really enjoyed moving the wire and watching the outcome on the screen. Per usual, I tweaked the coupling, Ls and end caps to derive many 3 dB bandwidths. In 1 arrangement with a 3 dB bandwidth of 3.57 MHz, the insertion loss was only 1.87 dB — I had evidently found a perfect ratio + combination of L and C yielding a low IL.

I'll increase the coil wire gauge in future experiments to lower solenoid resistance and try to garner even more Qu.

This is the *best* filter I've ever built. After peaking the tanks to 162.55 MHz, resonator adjustments weren't really needed. By moving the wire probe I saw that BW and IL change inversely. While tweaking the series caps and moving the probe, it felt like déjà vu from my experiences simulating these exact changes in GPLA over many years — except now they were alive and kicking. With some inductor manipulation (or as needed, changing the inductors) this filter can be centered from ~140 to well over 200 MHz. A high-octane, dramatic and versatile filter indeed.

Click for a sweep where I moved the wire very close to the neighboring tank capactor and over coupled the filter.

Click for a screen shot of another sweep.

Conclusion

With new awareness that resonator coupling and other issues may *also* occur in HF band-pass filters, making these 3 filters surely boosted my future band-pass filter construction standards at any frequency. Over the years, builders have emailed me sweeps of HF band-pass filters (usually VNA sweeps) with also sorts of extra resonances +/- poor insertion loss. Since I'm just an amateur, often nimrod RF enthusiast, I won't pretend I possess all the knowledge to critique their filter woes.

However, photos of their HF filters usually look messy and rushed: opposite to our needs — eschewing important factors like symmetrical layouts, inductor spacing, a first-rate ground plane, larger gauge + well spaced wire wound on big as possible toroid cores Yikes, I better stop now lest I venture into filter folklore.

Today, designing a filter with software is merely half the battle — construction poses the harder task. Unsurprisingly, measurement and experimental methods will elevate your outcomes.

Filter analysis with a sweep system is the ultimate, however I built and then tested many band-pass filters with a frequency counter, a signal generator, a 'scope and a return loss bridge during the first 12 years of my RF life.

Assuming you have a filter that peaks and isn't overcoupled: to measure the 3 dB bandwidth, first peak the filter to resonance for the highest peak-peak voltage in your 'scope. Here's the rest of the procedure:

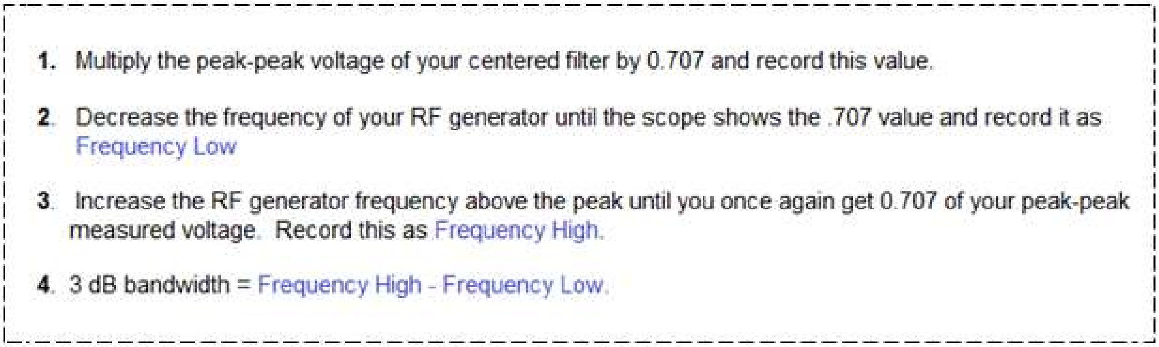

1. Multiply the peak-peak voltage of your centered filter by 0.707 and record this value.
2. Decrease the frequency of your RF generator until the scope shows the .707 value and record it as Frequency Low
3. Increase the RF generator frequency above the peak until you once again get 0.707 of your peak-peak measured voltage. Record this as Frequency High.
4. 3 dB bandwidth = Frequency High - Frequency Low.

You may also wish to measure the insertion and return loss as described in EMRFD, Radio Society Handbooks, or the RF Workbench series on this web site.

Whether you run superhet, low IF, SDR, or Zero IF (DC) receivers, or need to filter a transmit chain, well made band-pass filters just might boost your project performance. VHF poses additional challenges but offers great learning opportunities.

QRP — Posdata for Nov 12, 2013

A. Notes for Filter #2

Victor, 4Z4ME wrote to explain why the attenuation above the passband in Filter #2 looked mediocre. Filter topology and component values yielded the poor high-pass attenuation and not unwanted coupling as I originally shared.

Victor wrote: "The problem is that all the couplings are of one nature, in this case capacitive.
At very low frequencies the inductors short the signal to ground while the capacitors impedance gets higher so it is clear that the output signal goes to zero.

At high frequencies the inductors may be regarded as disconnected because their impedance get very high, however although the capacitors impedance get lower , you have them both at the series arms and at the parallel arms, so the signal gets to be attenuated by their impedance ratio and not get shorted to ground.

This can be shown in analytic calculation, or easier by simulation. See the attached file (LTSpice simulation). You can see that the filter (similar to your filter) gets closer to a finite attenuation at high frequencies, getting closer to the attenuation of a similar circuit with the inductors *removed*.

You can get better high frequency results by trading the center 0.2pF capacitor coupling to an inductive coupling. Remove the coupling capacitor and insert a wire through the center shield, ground its ends close to the main inductor. Shape the wire to a small loop near the grounded end of each of the resonator inductors to get small inductive coupling, and you get again a good BPF but with better high frequency attenuation. Instead of loops you can connect the coupling wire at a tap at the bottom of the main resonator inductors, close to the ground".

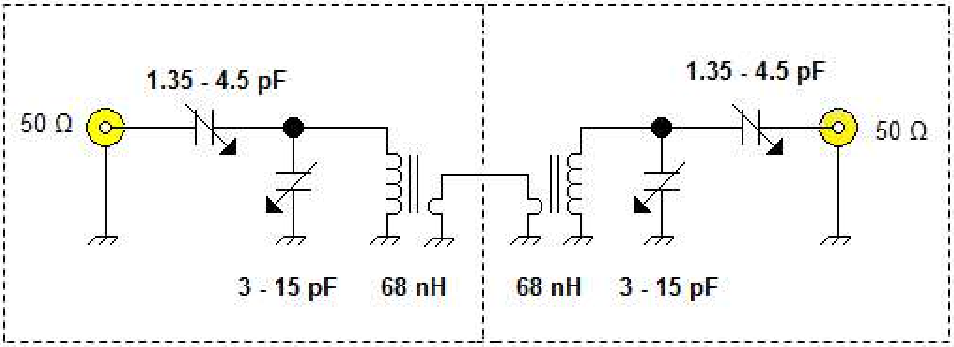

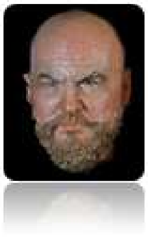

Filter 2 Pos-data Experiment

Above — Following Victor's suggestion, I inductively coupled the Filter #2 resonators with a wire looped near the grounded end of each tank inductor. Please view the schematic above. High-pass attenuation above the passband improved significantly. Click for a sweep out to ~ 511 MHz that shows the peak high-pass attenuation much closer to the low-pass skirt peak reponse.

At 2X center frequency (325.1 MHz) the high-pass skirt peak attenuation closed within 14 dB of the low-pass peak attenuation compared to the original Filter #2 response where the delta was around 30 dB. Further improvement occurs as we move above ~400 MHz. Big thanks to Victor for his support.

B. Sheet Metal Safety

A cut finger while working sheet metal prompted the following remarks: Jagged edges from cut sheet metal (especially brass stock it seems) may lead to deep cuts. Filing a burred edge with the metal clamped in a vice between some thicker metal stock with its raspy edge barely showing will reduce the "knife edge effect". When drilling, clamp the metal piece to your bench on a wooden backing board . If you jam your hand drill in an unclamped piece of metal it could spin and really gore you. Safety first.

---

7. NE612 Mixer Diddy

The Signetics NE612 and ilk enjoys wide use by the amateur radio community in novelty-grade Ham band receivers, or in somewhat better receivers laden with abounding front end filters + *stout,* switchable RF attenuators. Do you think of the NE612 as an ersatz receiver mixer? Some only consider them for the transmit mixer slot. Read this link why.

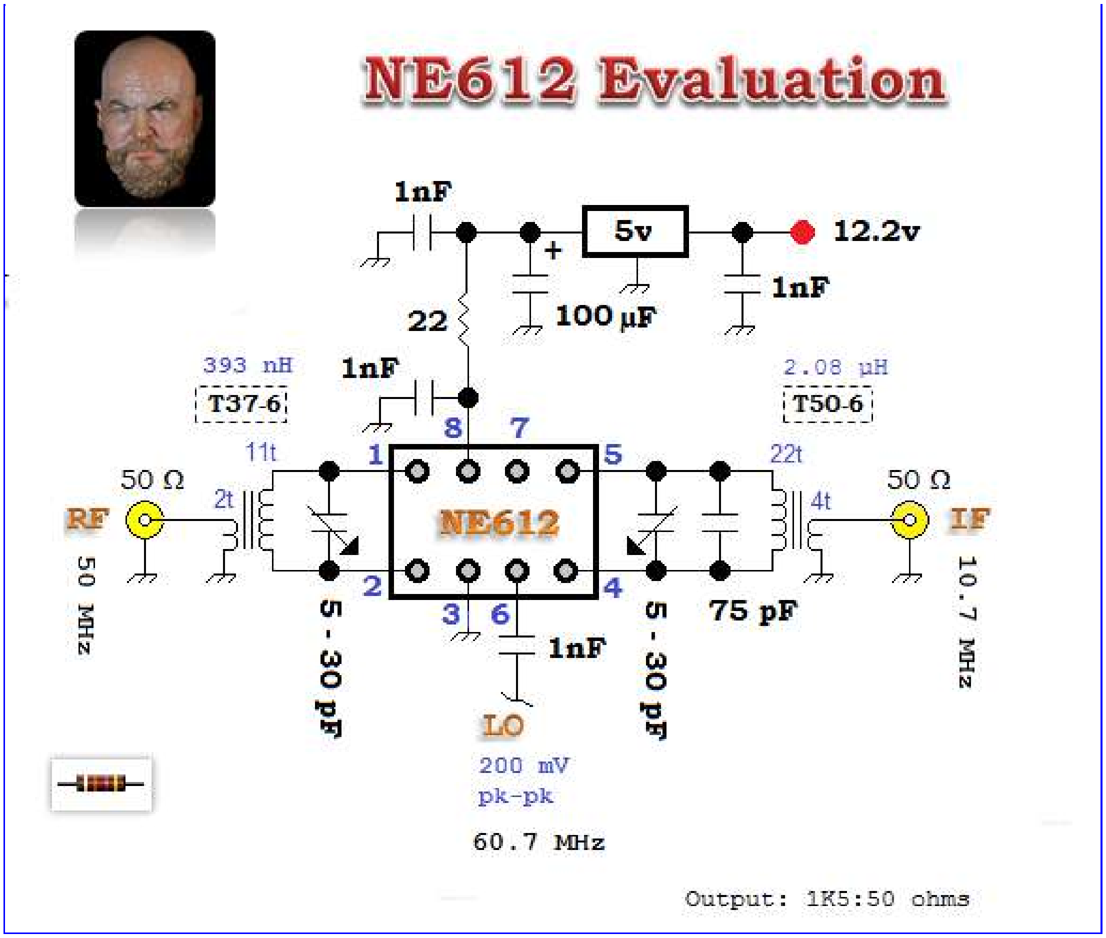
NE612 Evaluation
1nF
5v
12.2v
+
100 µF
1nF
22
393 nH
T37-6
1nF
2.08 µH
T50-6
8
7
1
5
11t
22t
50 Ω
2t
50 Ω
4t
RF
NE612
IF
50 MHz
2
4
10.7 MHz
3
6
5 - 30 pF
75 pF
1nF
5 - 30 pF
LO
200 mV
pk-pk
60.7 MHz
Output: 1K5:50 ohms

Above — My base schematic. I applied differential input to cancel 2nd order mixer products and balanced output to obtain maximum conversion gain. Click for the breadboard. Powdered iron toroids, Q = 1500 trimmers and a fixed C on the output formed the resonators. I chose a 10.7 MHz IF to drive low-cost ceramic filters in further (unshown) experiments. Normally, we don't bother tuning the input since our bandpass filter should launder the input signal sufficiently.

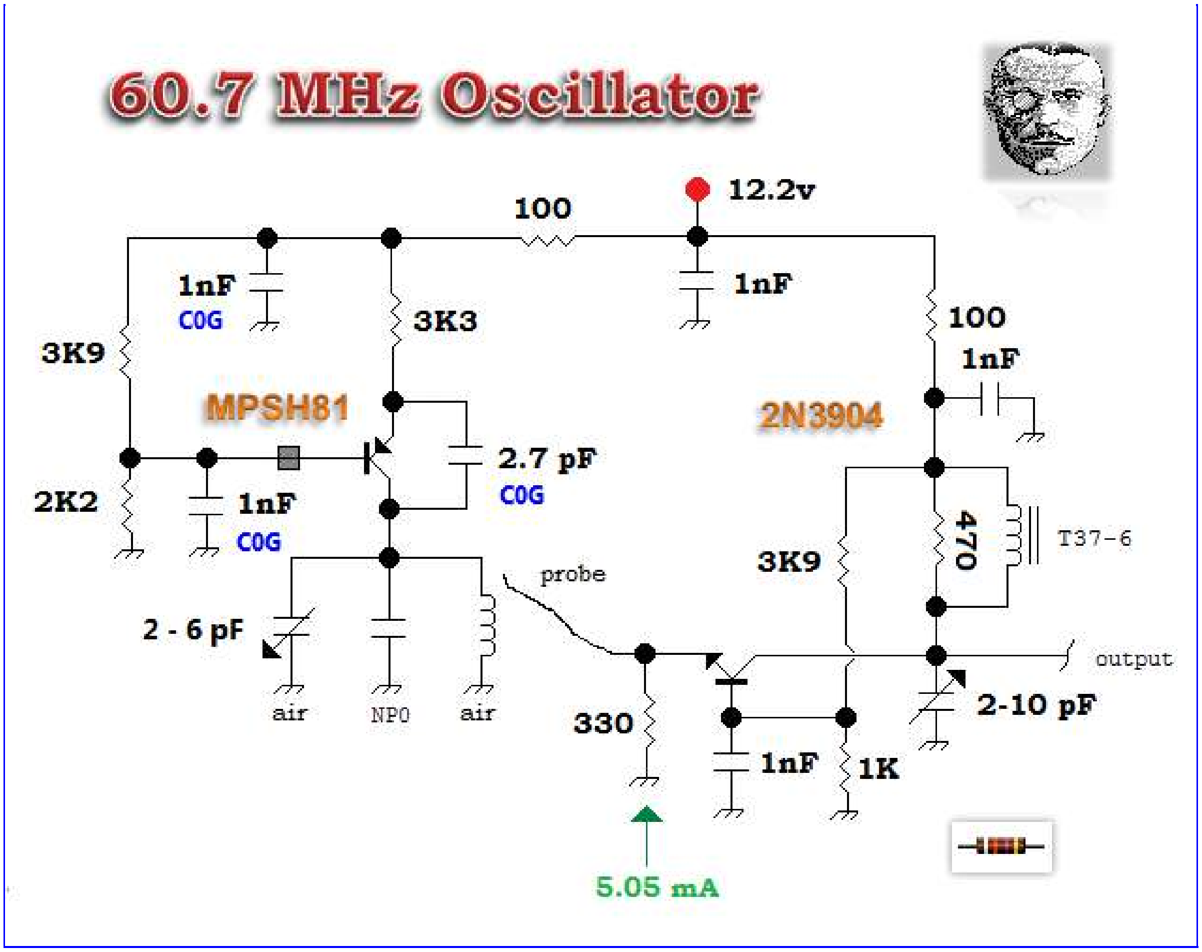
60.7 MHz Oscillator
12.2v
100
1nF
COG
3K3
1nF
3K9
100
1nF
MPSH81
2N3904
2.7 pF
COG
2K2
1nF
COG
470
T37-6
3K9
probe
2 - 6 pF
output
2-10 pF
air
NPO
air
330
1nF
1K
5.05 mA

Above — My 'thrown together' 60.7 MHz local oscillator. A wire on the common base buffer amp input avoids the series R and C usually required and I moved the wire probe to set the output power between 180 and 250 mV peak-peak into a 1K resistor load.

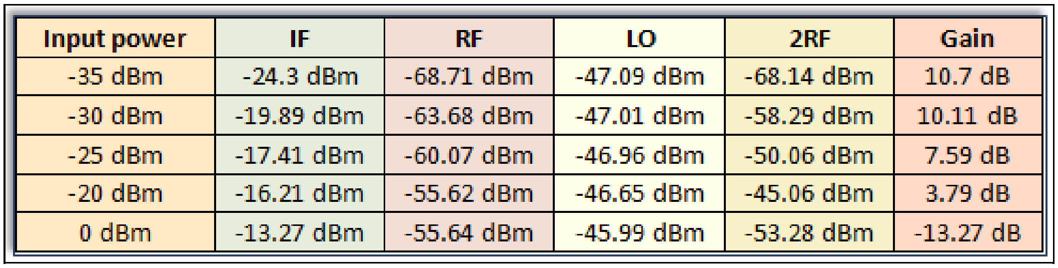

Input power	IF	RF	LO	2RF	Gain
-35 dBm	-24.3 dBm	-68.71 dBm	-47.09 dBm	-68.14 dBm	10.7 dB
-30 dBm	-19.89 dBm	-63.68 dBm	-47.01 dBm	-58.29 dBm	10.11 dB
-25 dBm	-17.41 dBm	-60.07 dBm	-46.96 dBm	-50.06 dBm	7.59 dB
-20 dBm	-16.21 dBm	-55.62 dBm	-46.65 dBm	-45.06 dBm	3.79 dB
0 dBm	-13.27 dBm	-55.64 dBm	-45.99 dBm	-53.28 dBm	-13.27 dB

Above — A table showing the mixer input, IF, RF, LO, 2xRF [the worst spur] power readings, plus calculated conversion gain. This table echos the datasheet + work by others — input power >= -25 dBm drives this mixer into compression. The NE612 runs low current plus low voltage across the output transistors — thus the maximum output power is limited to somewhere between +2 dBm to +5 dBm from my experiments.

Even with -35 dBm input power, I measured a conversion gain much lower than the typical 14-18 dB enjoyed by others who applied differential output at ~ 50 MHz. Turns out, I goofed — each output transistor contains a 1K5 Ω resistor, so when wired differentially, the output Z = 3K. Back to the bench!

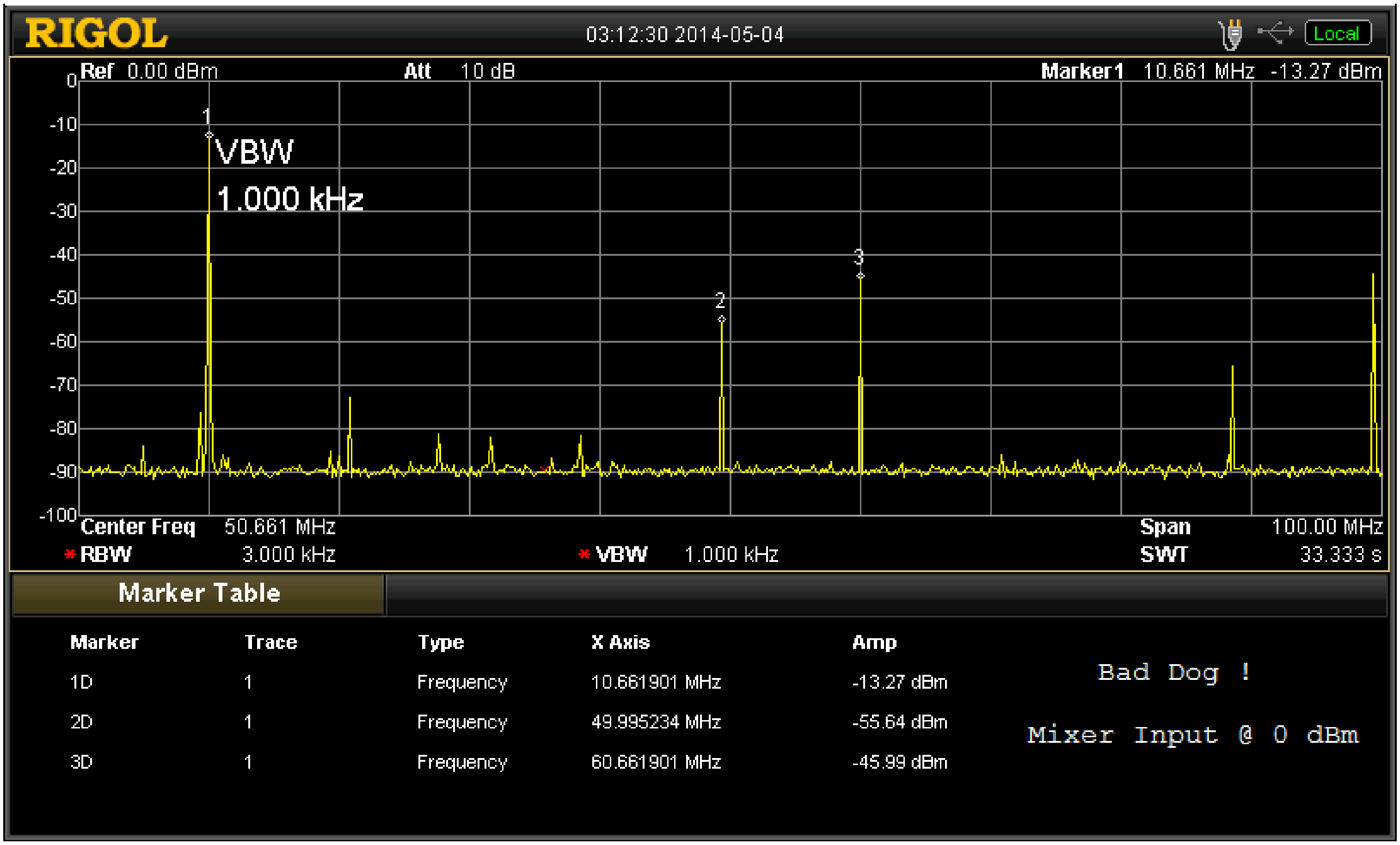
RIGOL
03:12:30 2014-05-04
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
1
2
3
VBW
1.000 kHz

Ref 0.00 dBm

Att 10 dB

Marker1 10.661 MHz -13.27 dBm

Center Freq 50.661 MHz

Span 100.00 MHz

* RBW 3.000 kHz

* VBW 1.000 kHz

SWT 33.333 s

Marker Table

Marker	Trace	Type	X Axis	Amp
1D	1	Frequency	10.661901 MHz	-13.27 dBm
2D	1	Frequency	49.995234 MHz	-55.64 dBm
3D	1	Frequency	60.661901 MHz	-45.99 dBm

Bad Dog !

Mixer Input @ 0 dBm

Above — SA of the mixer output with 0 dBm RF input. Something we would never do in real-life, but I wanted to see the spurs with such a drive: the spurious output actually looks great! The NE602, NE612, MC1496 and SO42P [all Gilbert cell mixers] output cleanly; especially when compared to diode-based mixers.

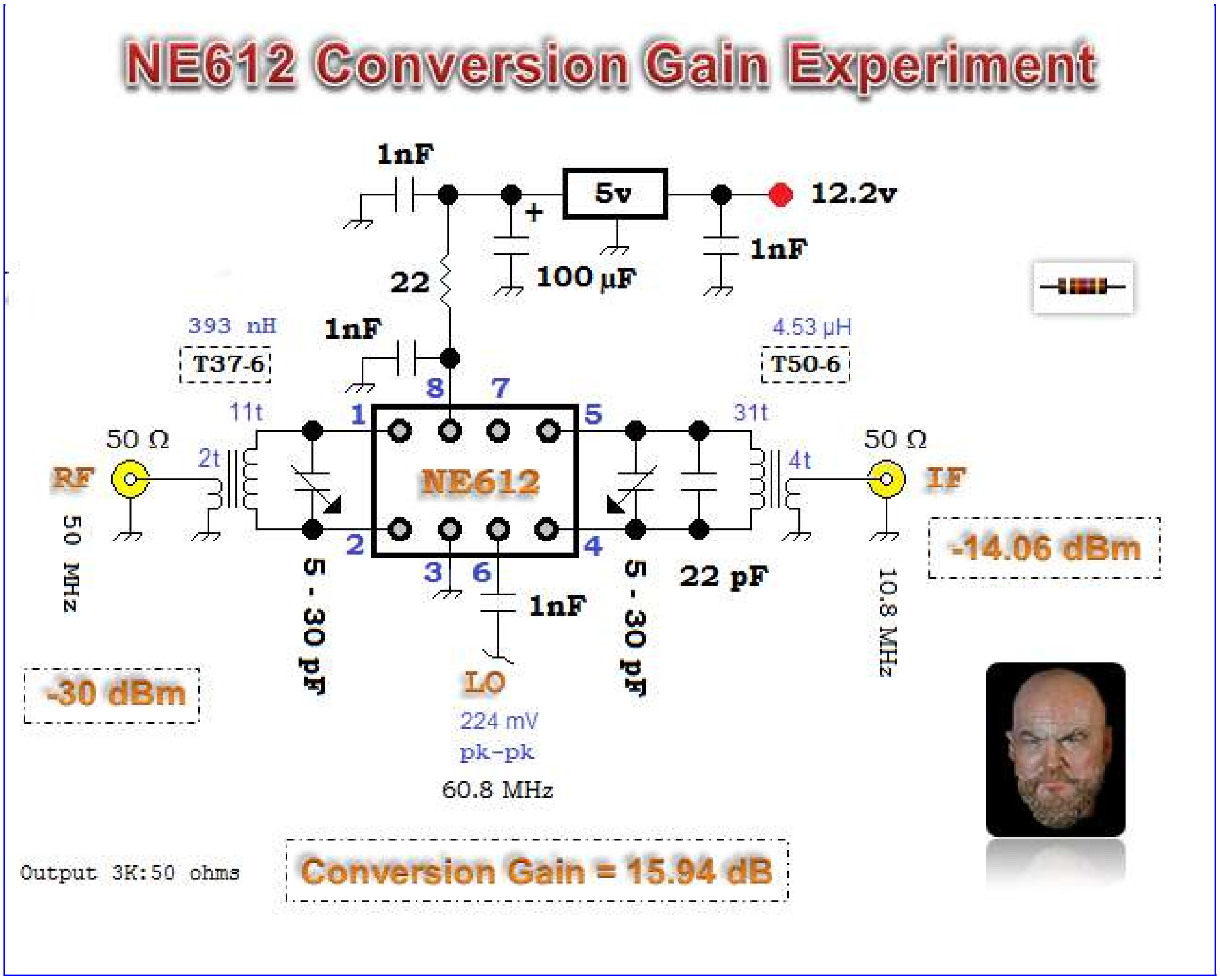
NE612 Conversion Gain Experiment
1nF
5v
12.2v
+
100 µF
1nF
22
393 nH
T37-6
1nF
4.53 µH
T50-6
11t
2t
50 Ω
RF
50 MHz
8
7
1
5
NE612
2
4
3
6
31t
4t
50 Ω
IF
10.8 MHz
-14.06 dBm
5 - 30 pF
5 - 30 pF
22 pF
1nF
LO
224 mV
pk-pk
60.8 MHz
-30 dBm
Output 3K:50 ohms
Conversion Gain = 15.94 dB

Above — My mixer with output Z corrected at 3K Ω. Wow, a conversion gain of 15.94 dB with -30 dBm drive. All good now. Click for the IF measurement on my SA. The LO had drifted up to 60.8 in this experiment, however, I left it and peaked the output resonator for maximum smoke. I also extracted another 0.5 to 0.8 dB gain by tweaking the LO drive. In my breadboard, a drive of 224 mV pk-pk gave the highest conversion gain.

Hereafter, I'll drive my NE612 transmit mixers with between -26 to -30 dBm on the RF port with differential input and output.

---

8. Miscellaneous Photographs

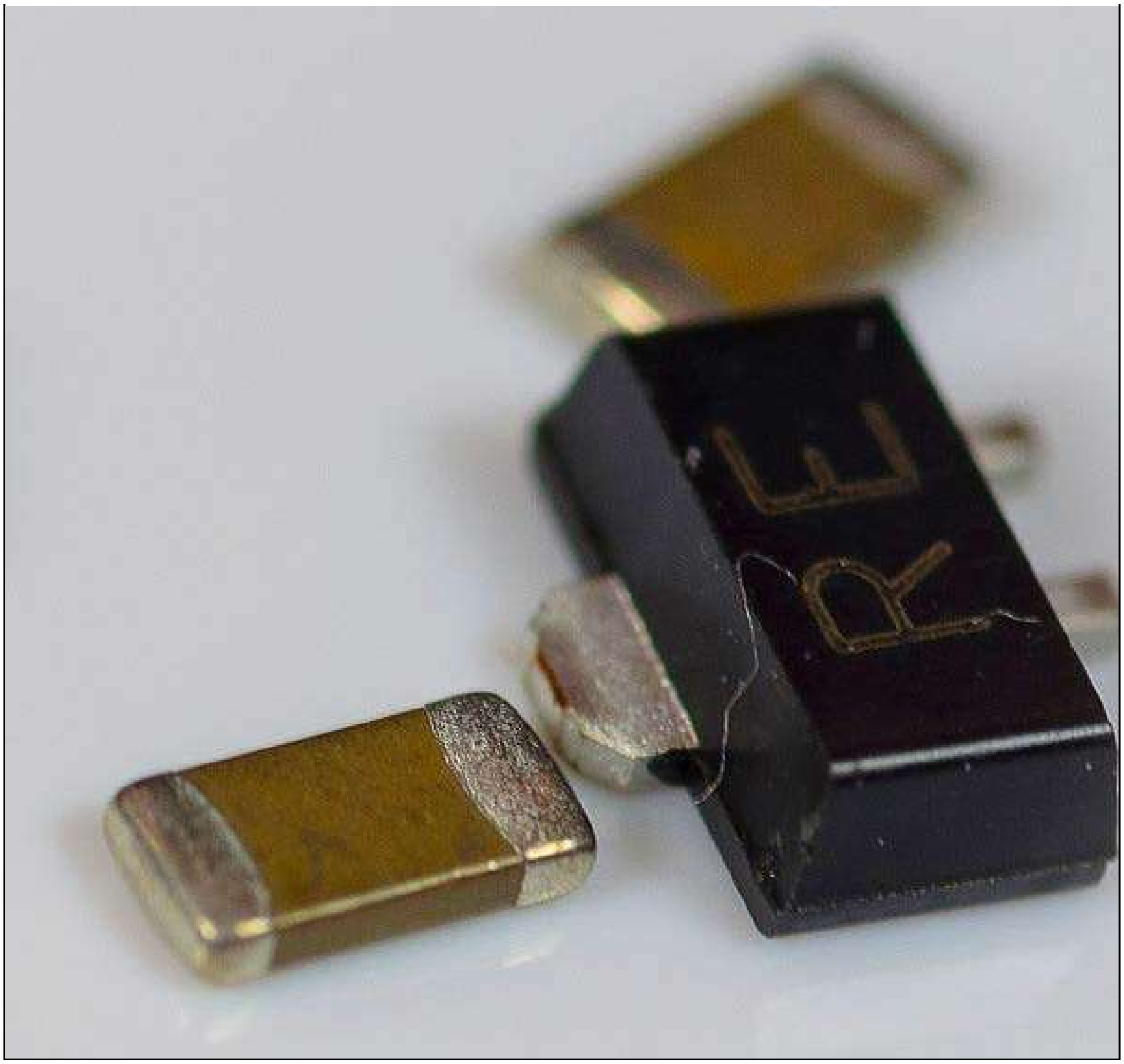

Above — Hair or lint in my SMT macro photo

Rosenberger
osenberger

Above — Instrument grade N-male to SMA female

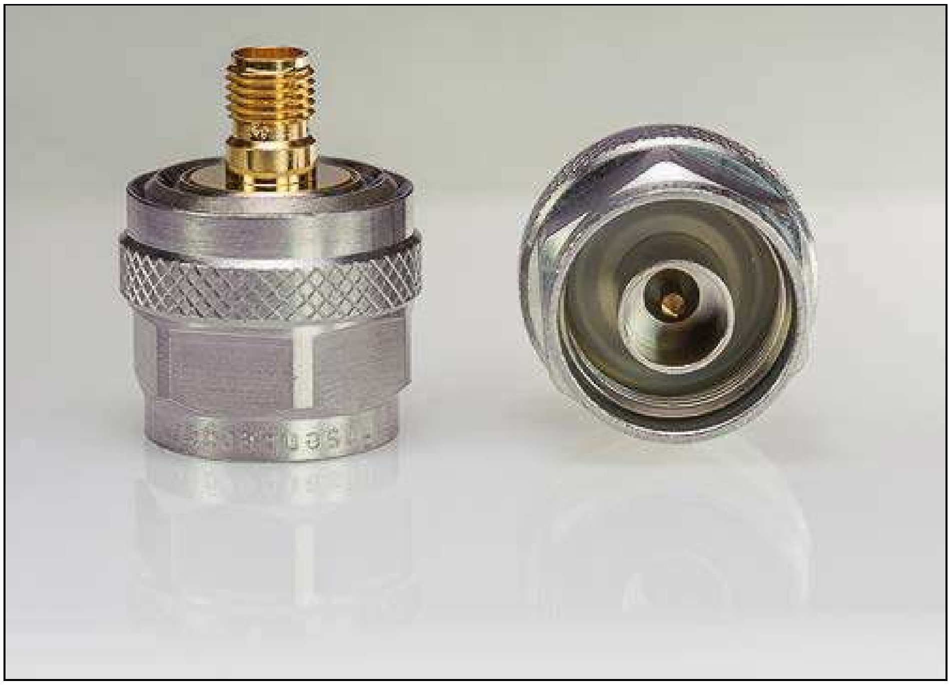

Above — Instrument grade N-male to SMA female; gorgeous machine work.

RF — Test and Measurement

RF Workbench Page 6 — QRP Transmitters

RF Workbench 6 explores some questions regarding basic transmitter power amplifer (PA) design. Since I write for both Ham and non-Ham readers, I feel badly that RF Workbench 6 excludes pure radio listeners; please consider getting your amateur radio licence to join the fun.

For Class A, AB and C transmitters in current source operation, radio builders knew most everything presented on this web page back in the 1980's — humbled and guileless, I've but a skosh of new ideas to further elevate this topic. Sometimes I feel bound in a creative straightjacket because puzzling design questions paralyze my understanding and progress. This web page explores a few transmitter questions born on my QRP work bench and I hope you'll relate and go design your own high-octane transmitters. As always, I hope to share insightful + actionable info for the DIY scratch builder; albeit from a lay-person perspective.

I borrow heavily from the work of Wes, W7ZOI per correspondence, direct contributions, from EMRFD and the references listed in Section 5. My sincere thanks to Wes for his support and inspiration.

Preface

You require an amateur radio operator licence to operate a radio transmitter and it's up to you to know and ensure your transmitter spurious emissions meet or beat the regulations in your country.

Although I measure my circuits with calibrated equipment, expect your measured voltage and current to differ from mine. **Why?** Variables may include your DC power supply voltage and current rating, dummy load resistance, stray L and C, cables and wire effects, bench errors, parasitic oscillations, temperature effects and parts tolerances.

I unabashedly declare my love for experiential learning: to build and measure, discover and develop — and best of all, to tune, tweak, and transform my RF power amps to unleash the "smoke" and power we all crave.

Contents

Section 1: 21.06 MHz Bipolar Transistor Transmitter in Class C: Dig out your old transistors and explore some Class C topics.

Section 2: PA Measurement with Examples: Measure collector V and I

Sections 3, 4: Incomplete and coming in 2013-14. I'll post any additions on the QRP Log -- Thanks

Section 5: I'll list all the references for this web page

Section 6: Photographs and Miscellany. I'm told at least ~~3~~ (now 4) people enjoy my photos.

Section 1: 21.06 MHz Bipolar Transistor Transmitter in Class C

Despite all the buzz about PA efficiency and the trend towards more efficient switched-mode power amplifiers, my email inbox fills with content regarding those Class A to C power amps. Consider too, most Hams, including QRP buffs, power their rigs with an AC connected DC supply that makes battery life moot. Perhaps our love of PAs in current source operation harkens from our historic tube days: do we simply enjoy making heat and wasting power? Like me, you might just love making and measuring bipolar transistor circuits.

A Class C power amplifer excited by a crystal oscillator offers both new and reawakening homebrewers the chance to make a simple but relatively powerful, 1 frequency transmitter with only 2 - 4 RF stages. Keying a Class-C PA embodies the Ham glory days: put some RF into the sky with basic, analog circuitry.

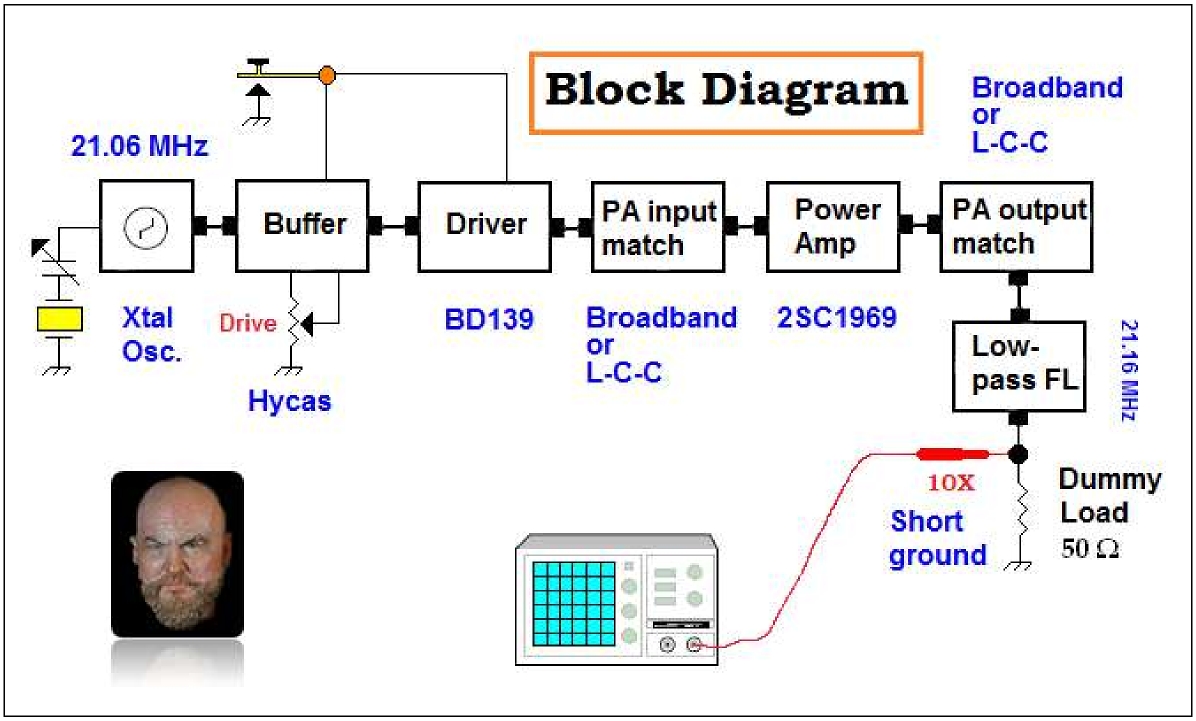
Block Diagram
21.06 MHz
Buffer
Driver
PA input match
Power Amp
PA output match
Broadband or L-C-C
Xtal Osc.
Drive
Hycas
BD139
Broadband or L-C-C
2SC1969
Low-pass FL
21.16 MHz
10X
Short ground
Dummy Load 50 Ω

Above — The block diagram for all the experiments in Section 1. Each stage involving transistors was built and tested from left to right before making the next stage.

Crystal Oscillator, Keying Switch and Hycas Buffer Amplifer

As more a listener than a talker, I wanted a low-level crystal oscillator to keep my *key up* current draw low. In the class C power amplifers based on the above block diagram, I measured only 7.5 mA. *key up* and several hundred mA *key down*. I also wanted my back wave to be > 40 dB when I eventually converted the PA to Class AB during another set of experiments.

Design Question 1: What backwave should I strive for?

Backwave = the dB difference in output power between key up and down. Opinions differ and context may partly explain why. If you're working DX, backwave is less important than when your working nearby ops on 80 meters. From asking respected authors, experiments and operating, I feel 40 dB should be the minimal backwave to strive for. Operating Class C, your final(s) are turned off during key up and a >= 40 dB backwave proves an easy goal.

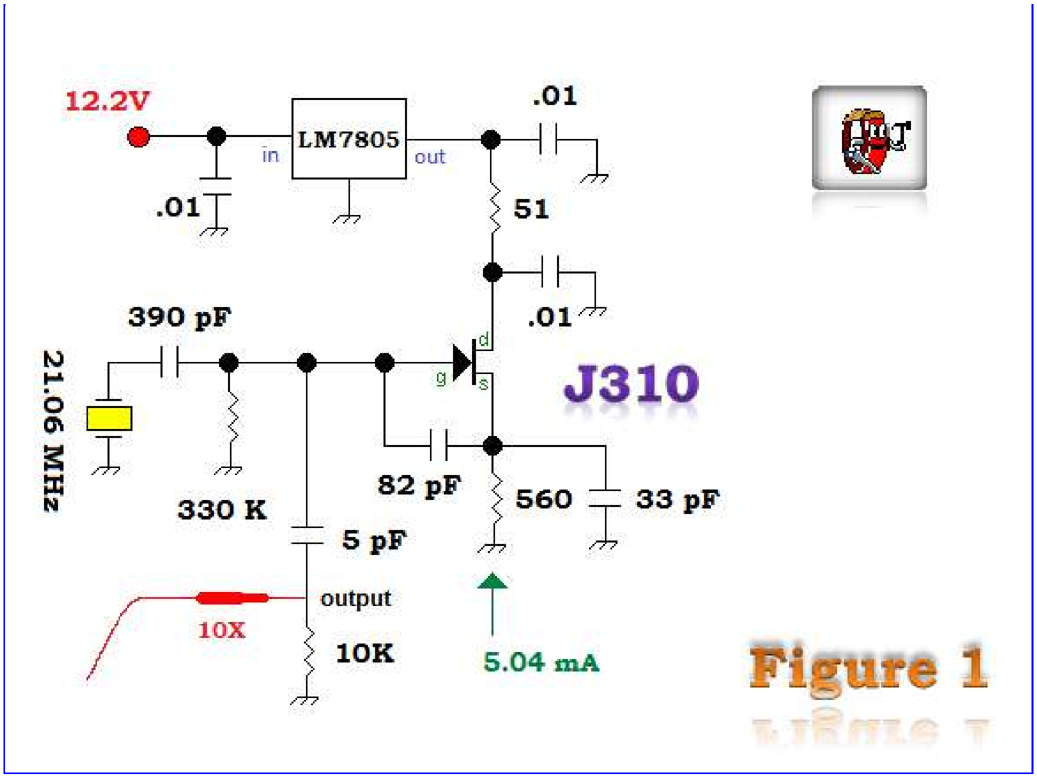
12.2V
LM7805
in
out
.01
.01
51
.01
390 pF
21.06 MHz
d
g
s
J310
330 K
82 pF
560
33 pF
5 pF
output
10X
10K
5.04 mA
Figure 1

Above left — The 21.06 MHz crystal oscillator. Click for the output voltage into 10K Ω. Click for the FFT. Seeking low distortion and output power, I placed variable capacitors in the feedback and source slots and also played with the 560 Ω source resistor. While watching in my 'scope and tweaking , I chose the the lowest distortion signal under 1 Vpp. The variable capacitors were removed, measured and replaced with nearest standard 5% parts. The 10K load of Figure 1 corresponds to the gate resistor of the hycas amplifer input.

21.058,824,1
MHz
Gate
Freq
Ch 1

Above — Without the 390 pF capacitor, the crystal oscillator vibrated at 21.058 MHz.

21.060,316,8
MHz
Freq
Ch 1

Above — The crystal oscillator measured frequency after experimentally adding a 390 pF cap. Should I have tried a 470 pF capacitor and attemped to get 21.0600 MHz?
My oscillator looks unconventional and in some cases you might want 1 with higher output voltage. Check EMRFD and the other W7ZOI references in Section 5 for xtal oscillator design ideas.

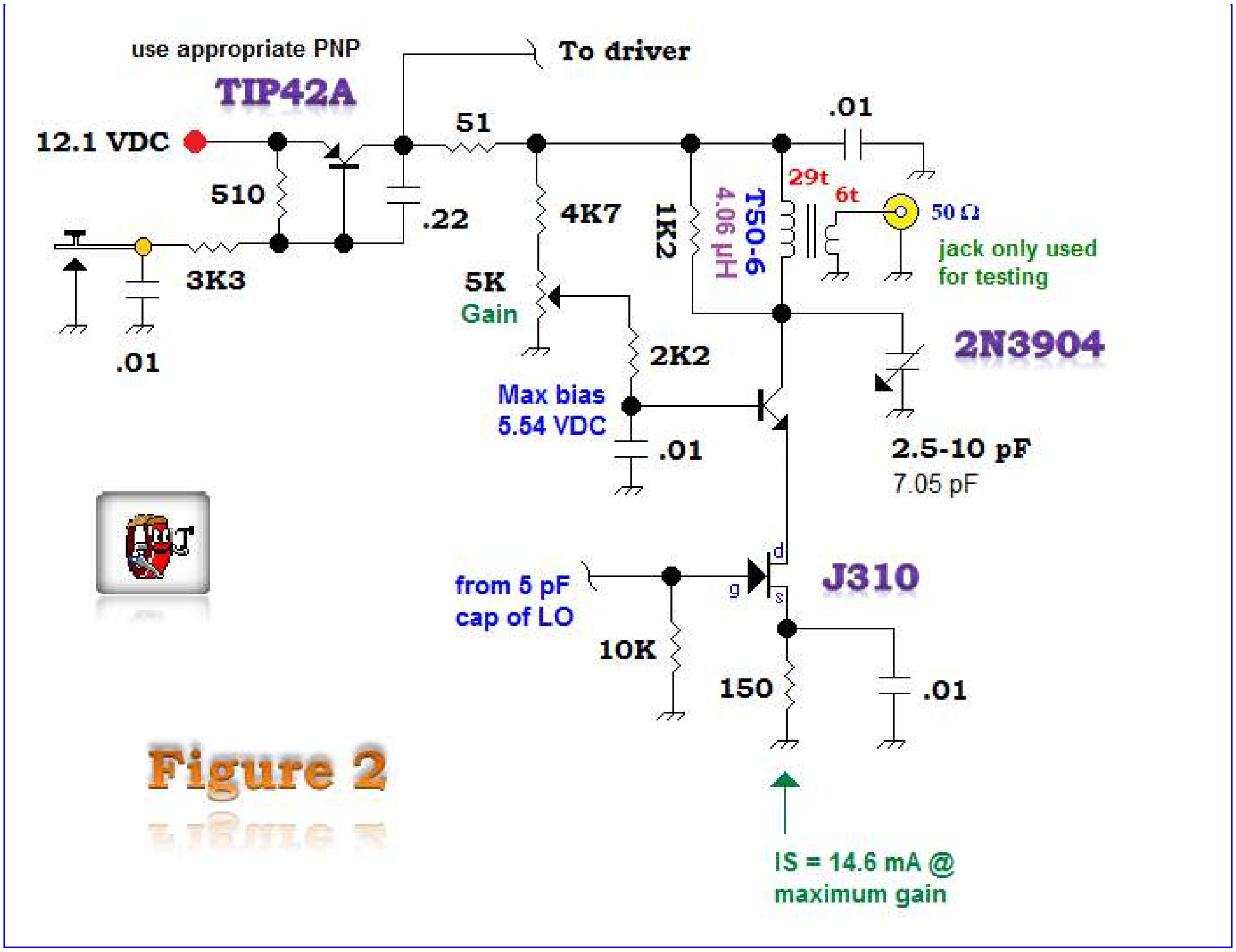
use appropriate PNP
TIP42A
To driver
12.1 VDC
51
.01
510
.22
4K7
1K2
T50-6
4.06 µH
29t
6t
50 Ω
jack only used
for testing
3K3
.01
5K
Gain
2K2
2N3904
Max bias
5.54 VDC
.01
2.5-10 pF
7.05 pF
from 5 pF
cap of LO
d
g
s
J310
10K
150
.01
IS = 14.6 mA @
maximum gain
Figure 2

Above — Figure 2, the hybrid cascode buffer. I've never tuned, nor keyed a hybrid cascode amp before and it works well. Click for the oscilloscope output. Click for the FFT measured with a long wire grounded 10X probe. I applied a junkbox TIP42A switch lying on my bench since I keyed a number of *unshown* high current stages during my experiments. Only about 10% of my experiments ever make it onto the QRP/SWL HB web site. For the PNP switch, just use whatever can handle the *measured* current with key down: 2N3904, 2N4403, 2N2907, BD140, etc.,etc..

Yes dear reader, measure everything.

The Driver Amplifer

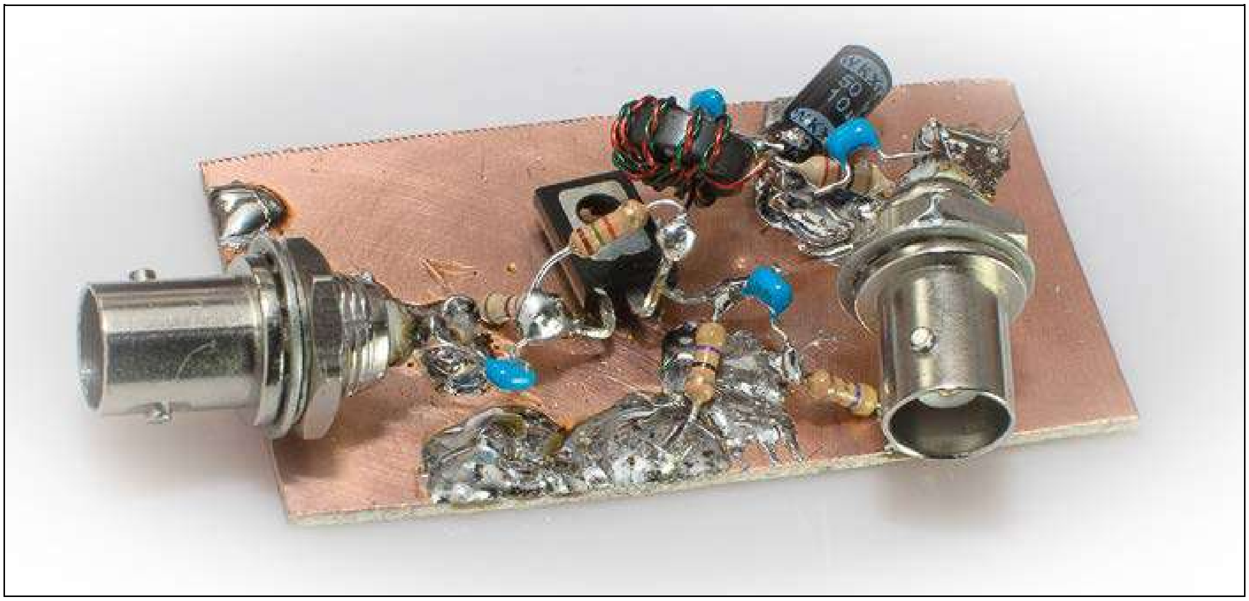

Above — A BD139 feedback amp breadboard. Click for the schematic. Wanting an output power of at least 22 dBm to drive my final, I built 3 varieties with both the BD139 and a 2N3866. Normally, for receiver amplifers, I focus on a strong input and output return loss and work hard to maintain linearity — but mostly keep the emitter current under 50 mA. At 21.06 MHz, the biggest *clean* signal I could muster with such a broadband feedback amp was +16 to +18 dBm — well short of my goal. Click for a DSO screen shot of the 41.3 mA amp at 6.42 Vpp or ~20.13 dBm power; quite distorted.

I needed to try something else and saw the Figure 3 amplifier in some of Wes' work; wielding bold emitter current to boost signal handling and power.

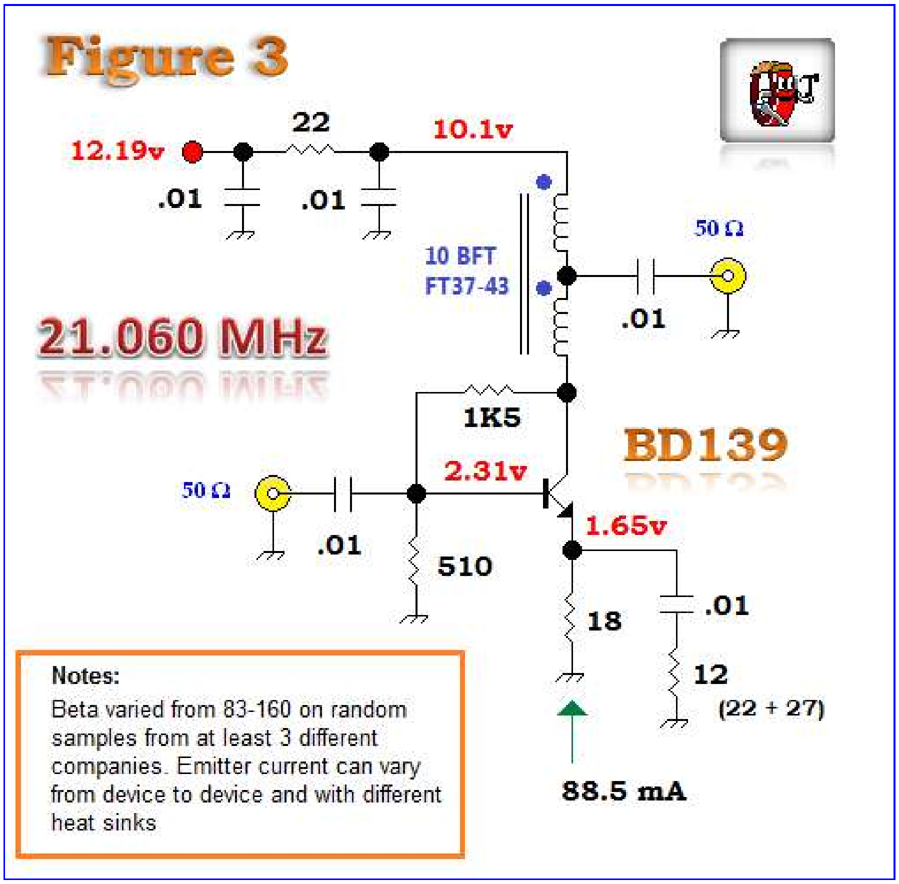

Notes:
Beta varied from 83-160 on random samples from at least 3 different companies. Emitter current can vary from device to device and with different heat sinks

Above — The W7ZOI designed feedback amp used for a driver in all Section 1 power amp experiments. Love it! Many builders feel tempted to abandon series and/or shunt feedback and just "go for gain". My experiments suggest this approach often leads to instability within the driver circuit and possibly on down the RF chain. We should strive for linearity and stability; 2 big reasons we apply feedback. Flirt with a driver *lacking* feedback if you dare.

Click for the maximum output power attained by turning my hycas bias pot fully clockwise: 7.77 Vpp, or 21.79 dBm. You can see distortion, but since this stage will drive a Class-C transistor PA (i.e. it will switch "a diode" on and off and generate clipping), I decided it will work fine. A VHF-UHF transistor such as the 2N3866 or 2N5109 may yield as much as 3-4 dBm more power, however, you'll see more clipping at higher driver levels. I stuck with the BD139 and tried several pieces from at least 3 different manufacturers (The BD139 was made by at least 10 companies at 1 point). A Philips version purchased in 2008 yielded the 21.8 dBm, so I bolted it to a home built copper clad heat sink and soldered it in place.

Calculating Transistor Output Power Off the Bench

Design Question 2: How do I mathematically estimate my Class A amplifier output power?

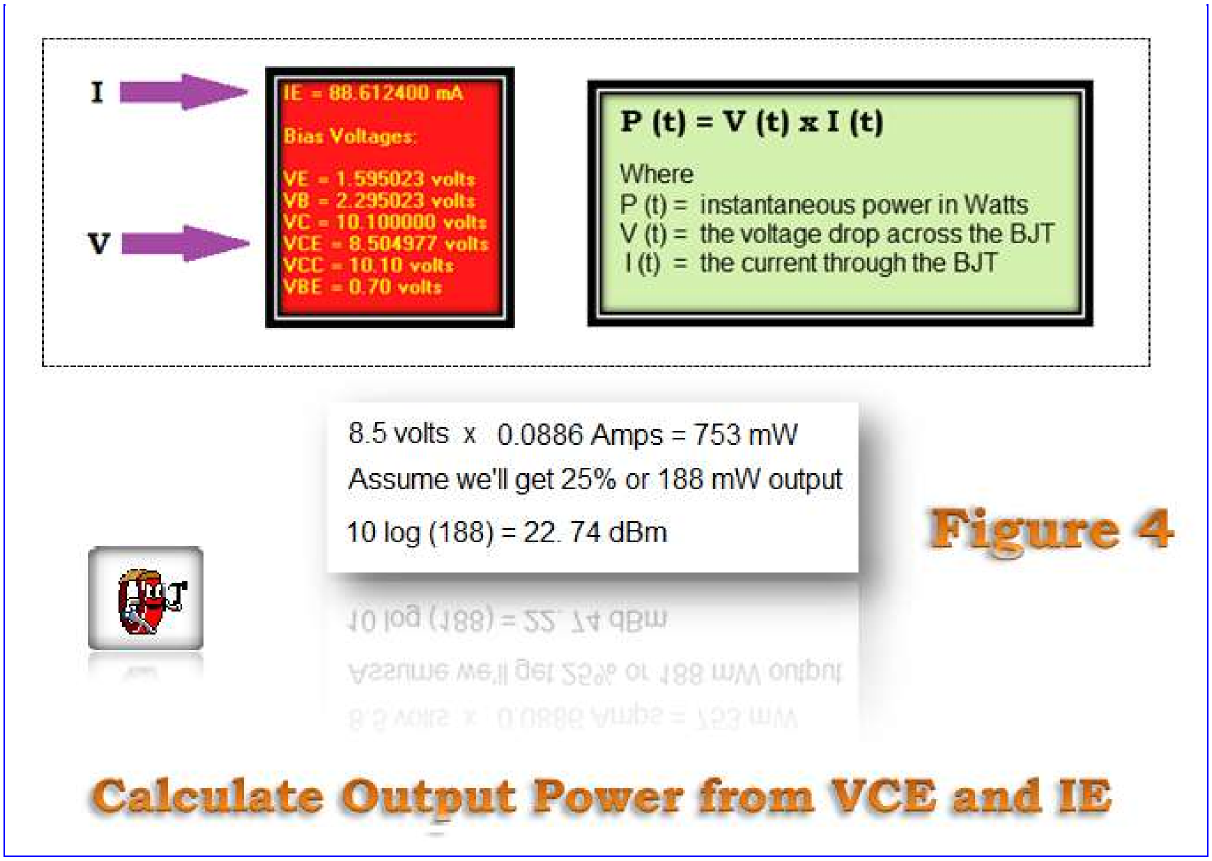

```
IE = 88.612400 mA
Bias Voltages:
VE = 1.595023 volts
VB = 2.295023 volts
VC = 10.100000 volts
VCE = 8.504977 volts
VCC = 10.10 volts
VBE = 0.70 volts
```

P (t) = V (t) x I (t)

Where
P (t) = instantaneous power in Watts
V (t) = the voltage drop across the BJT
I (t) = the current through the BJT

8.5 volts x 0.0886 Amps = 753 mW

Assume we'll get 25% or 188 mW output

10 log (188) = 22. 74 dBm

Figure 4

Calculate Output Power from VCE and IE

Above — Estimating output power from my Figure 3 amplifier. I measured a static DC Beta of 125 as shown in Section 5 on this page. Then I calculated VCE (the voltage across the transistor) and emitter current with my program NPN-DC Bias. Wes wrote a similar and likely better program called Biasnpn08.exe that's included on the EMRFD compact disk.
You may also measure your VCE and IE directly and improve the above calculation with measurements. I do this frequently.

Class A amps exhibit about 25% efficiency and also imagine that Figure 3 will bear some gain compression.

My measured power = 21.8 dBm — it amazes me how close a non-bench calculation using DC analysis can get to a AC bench measurement. The Figure 4 procedure is apparently a bread and butter RF design tool.

The Power Amplifer Input

Design Question 3: How do I match my driver output to the Class-C power amp input to maximize power?

This question provided a few nights of reading and experiments; time well spent. We have 2 main ways to match our driver output to the PA input: broadband or tuned. First, lets explore the problem theoretically:

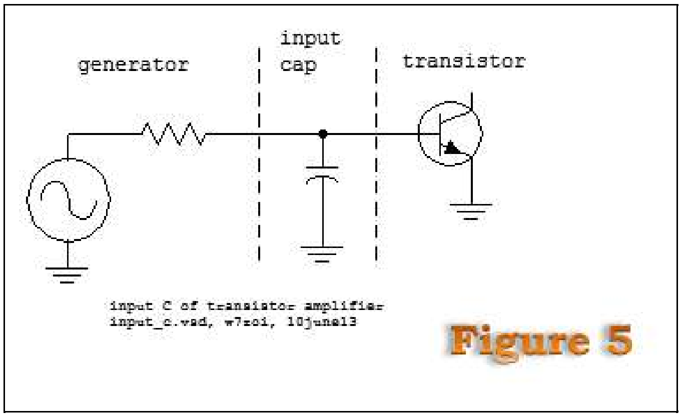

Figure 5

Above — A simple model sent by Wes: driving an amplifier with a power source. As amateur designers, simple models help us understand complex concepts.

A pure voltage source doesn't really exist. We don't apply a voltage to the base of the PA; rather, we apply a power source (our signal generator) — a signal voltage followed by a series impedance that's usually just a source resistance for a load. The power source + source impedance together *source* the load that includes the load capacitance represented as an input cap in Figure 5. You may wish to read about Thévenin's theorem and its dual, Norton's theorem to learn more about linear network analysis.

The load presents an impedance where the power is delivered and either dissipated in its resistance, or reflected back to the source when the load contains a reactance . At our PA input, this capacitive load reflects some of the signal back and induces a pole that corresponds to a "capacitor" time constant.

If you consider our power source, its source resistance, the PA input capacitance and just omit the effects of Q you'll see that the the voltage across the capacitor decreases as frequency increases. If you insert a 2:1 turns ratio transformer after the driver, you'll drop the AC voltage by 2 and its resistance by 4. For the same capacitance, the 3 dB cutoff frequency of this single pole filter increases. Also, the lower output impedance drops the driver amplifier voltage gain and may boost stability.

In short, we need a low driver source impedance since the Class C bipolar amplifier base-emitter junction has lots of capacitance To get current flowing into the PA base, we drop our driver output source impedance down *low*.

I thought — couldn't we just neutralize this capactance with an inductive output impedance? Thus our PA driver stage would drive a resistive load and everything would be "ducky". I consulted Wes, plus viewed some math and circuitry.

You can't effectively neutralize a bipolar transistor amplifier, since the capacitance isn't constant and behaves more like like a varactor diode rather than the C shown in Figure 5.. Even when forward biased, diode capacitance depends on the instantaneous voltage that appears across the diode. In contrast, 'FETs prove less troublesome and thus it's possible to achieve some neutralization with a 'FET amplifier. Please view the power stage from EMRFD Figure 11.14. Check out the push-pull pair of IRF511s. C5 and C6, at 22 pF each, are neutralizing capacitors.

In some cases, the capacitive reactance may be absorbed as part of a resonant circuit on the PA input. For example, a low-pass filter might soak up some of the C.

In all cases, our goal is drive in some power and switch the PA transistor solidly on and off (i.e. technically approaching class B but definitely class C because conduction will be less than 180 degrees) — a low output impedance can drive that input capacitance.

Driving final(s) with a very low output Z rings true even if your driving the latest (super-costly) ultra-low loss, high breakdown voltage, silicon GaN FETs to output >100 Watts at 1296 MHz in Class D, E, F, or Inverse Class F. Just get it done!

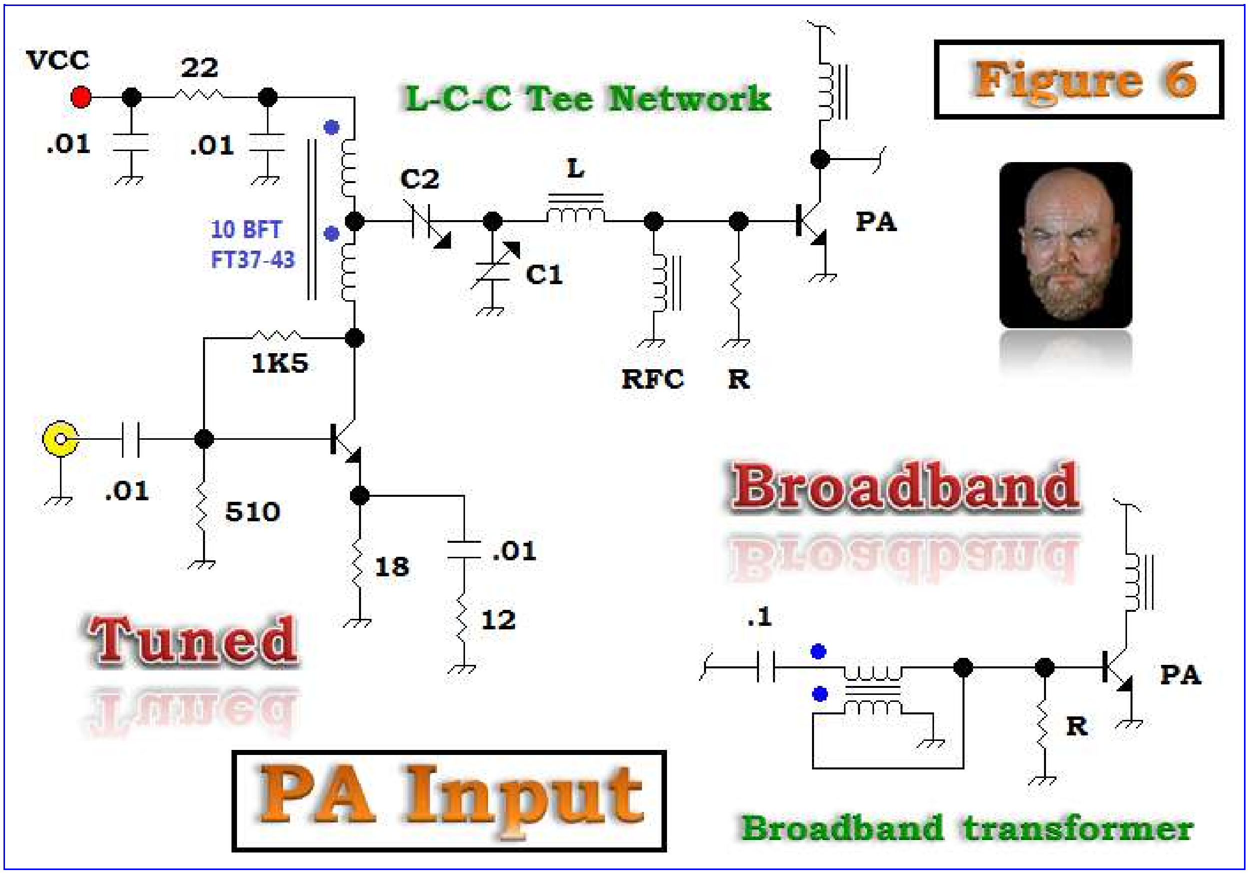

Above — Figure 6 shows 2 basic methods to drive the PA with a low impedance: single frequency network or broadband step-down transformer. I'll discuss both methods in detail later. Let's move away from theory and lighten things up for a couple of minutes.

Above — A short 10X probe ground lead boosts Tx measurement accuracy; in particular, for FFT analysis with your DSO. You can read or see **why** by searching the web for text and/or videos concerning oscilloscope probes.

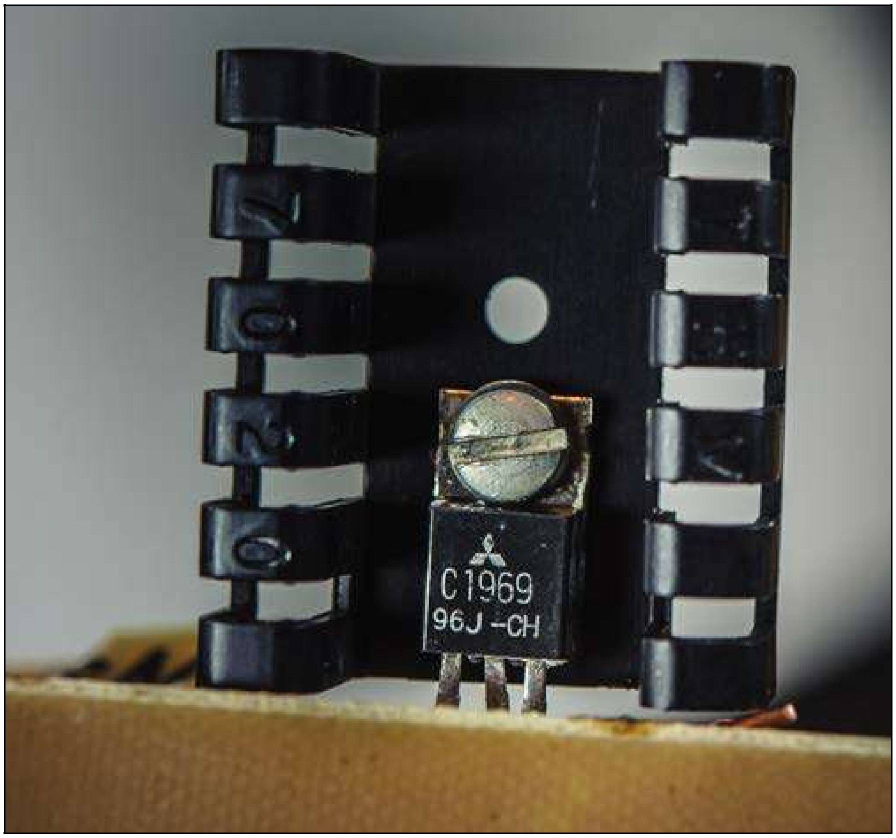

Above — The long gone 2SC1969 works great. Beware of counterfeits however.

Bipolar transistor selection troubles us since most of the 'old standbys' went obsolete way back when. I've been collecting dusty, old *QRP* driver and PA transistors for decades and my bag includes BD139, MJE181, 2SC2509, 2SC1969, 2SC2166, NTE123, 2N3053, 2N4037 and 2N2219A. For VHF, apart from the 2N5109/2N3866 standards, I've kept a few MRF607, 2SC1971 and 2N3553.

VHF/UHF transistors exhibit tremendous gain at lower HF, but they might oscillate without due care.

Some rules I've gleaned: The fT of a BJT should be at least 3-5 times the operating frequency; the VCEO should be at least double the VCC (more is better) ; don't exceed the maximum heat dissipation, nor maximum current. In short, read the datasheet. I won't cover heat sinking, nor gain compensation for multi-band amplifers on this web page; for it's been done by far better authors.

Lets make a basic Class C power amplifier based upon the RF blocks presented thus far.

Class C PA with Broadband Input and L-C-C Tuned Output

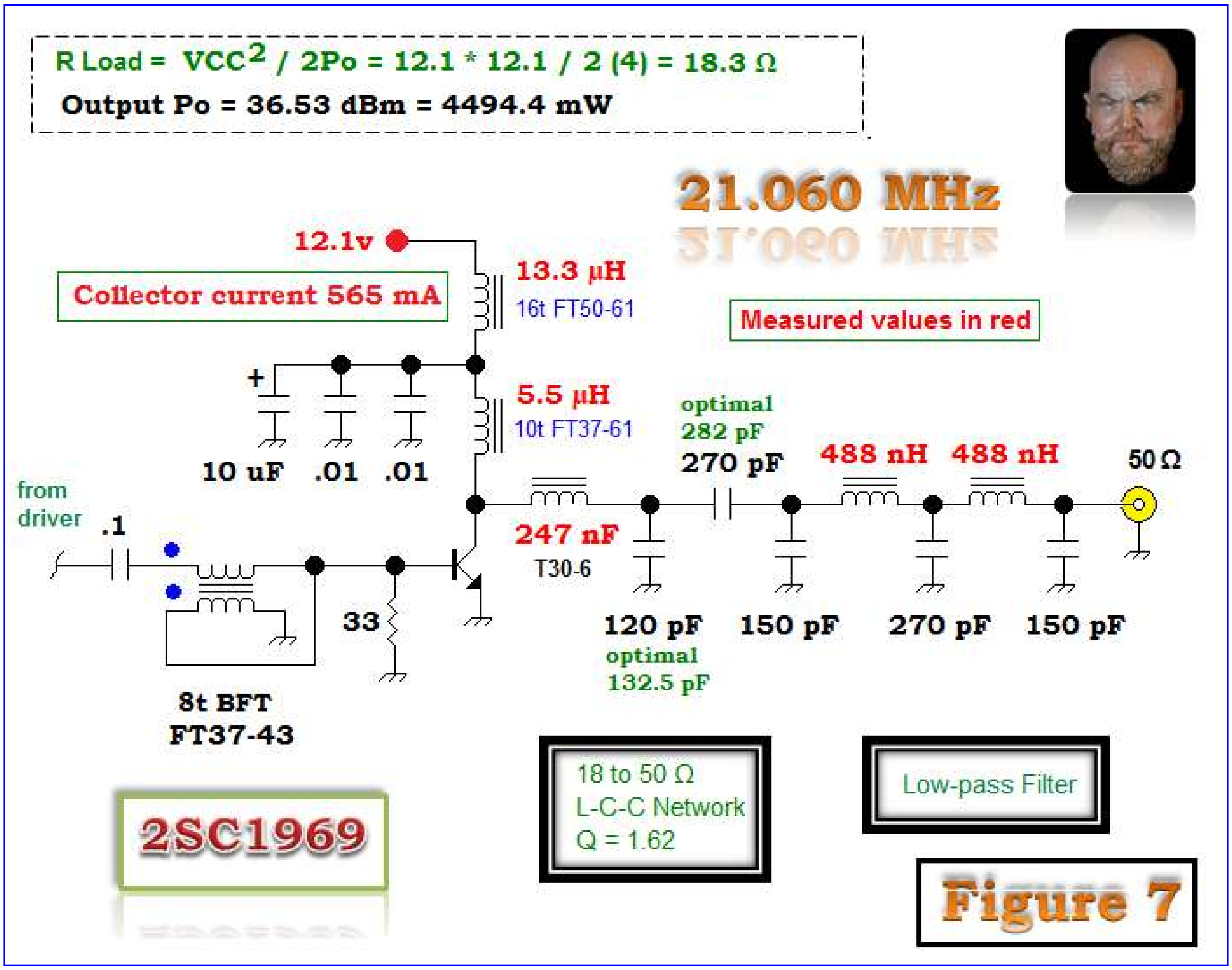
R Load = VCC2 / 2Po = 12.1 * 12.1 / 2 (4) = 18.3 Ω
Output Po = 36.53 dBm = 4494.4 mW
21.060 MHz
12.1v
Collector current 565 mA
13.3 µH
16t FT50-61
Measured values in red
+
10 uF
.01
.01
5.5 µH
10t FT37-61
optimal
282 pF
270 pF
488 nH
488 nH
50 Ω
from
driver
.1
247 nF
T30-6
33
120 pF
optimal
132.5 pF
150 pF
270 pF
150 pF
8t BFT
FT37-43
18 to 50 Ω
L-C-C Network
Q = 1.62
Low-pass Filter
2SC1969
Figure 7

Above — Input matched PA via a broadband transformer (a bifilar wound transmission line transformer with a 4:1 Z ratio). A great example of this design includes the Figure 2 of Ugly Weekender transmitter by Wes. The 50 Ω output Z of the driver is transformed to 12.5 Ω. to drive the PA input. While a conventional link-coupled transformer works okay, the closer coupling of a twisted wire transmission line transformer often affords a better impedance transform and lower losses. You might experiment with a 9:1 Z transformer instead of the 4:1 shown. Click for a highly compressed photo of 1 of the full breadboards.

The 33 Ω resistor helps lower gain and stabilize the amp. I ran 33 to 100 Ω applying experimental methods. Read more about the PA base resistor and coil XL requirements in the text following Design Question 6.

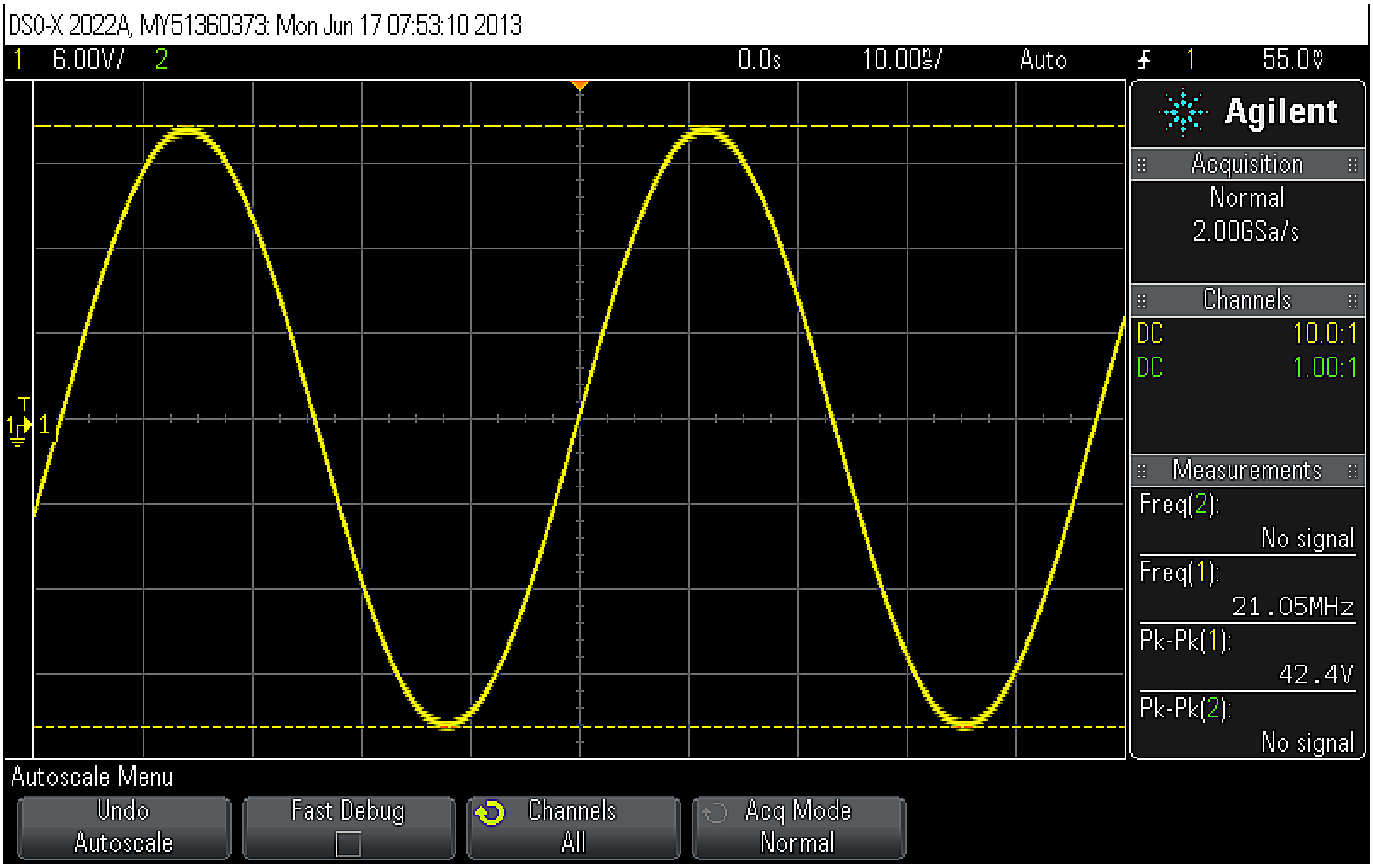

DSO-X 2022A, MY51360373: Mon Jun 17 07:53:10 2013

1 6.00V/ 2 0.0s 10.00ns/ Auto 1 55.0V

Acquisition: Normal 2.00GSa/s

Channels: DC 10.0:1, DC 1.00:1

Measurements:

Freq(2): No signal

Freq(1): 21.05MHz

Pk-Pk(1): 42.4V

Pk-Pk(2): No signal

Above — The 'scope output from Figure 7. 42.4 Vpp = 36.53 dBm = 4.49 Watts.

L-C-C Output Matching Network

3 discussion points arise from Figure 7: [1] the output matching network. [2] the collector choke + decoupling circuitry, and [3] the low-pass filter. I'll discuss choosing a collector choke and decoupling in Section 2. We design our low-pass filters with tables or software; I used an old ARRL Radio Amateurs Handbook and then tweaked it in GPLA to get standard value capacitors. Click for a GPLA simulation of my filter. The L-C-C network also provides some low-pass filtering.

To calculate the collector output resistance, we use the formula R(load) = Vcc^2/ (2*Pout) as shown in Figure 7 for 4 Watts. For 1.5 Watts at 12 VDC --- R(load) = 12 * 12 / 2(1.5) = 48 Ω --- no output matching network is required; making 1.5 watts a popular transmitter power for new builders eager to learn about Class C.

Design Question 4: How do I make an L-C-C matching network?

Let's stick with 4W power and make an L-C-C Tee network for a 18 to 50 Ω transform. I prefer the L-C-C network over the L-network because you get 2 capacitors to tweak and I find it much easier to match a wide range of impedances compared to the L-network. Since we may pinch, or expand the inductor windings, in reality, the L-C-C gives you 2.5 possible tweaks.

Normally I wind the L on a number 6 powdered iron toroid with 26-28 gauge magnet wire in size T25, T30, T37, or T50. 26 gauge wire allows for good coil squishing and expansion with thumb and forefinger. For design cycle capacitors, I mostly solder in those horrible, low Q, orange, green, or red muRata trimmer caps such as the TZ03Z500F169B00. Although rare, ceramic-mica trimmers work well.

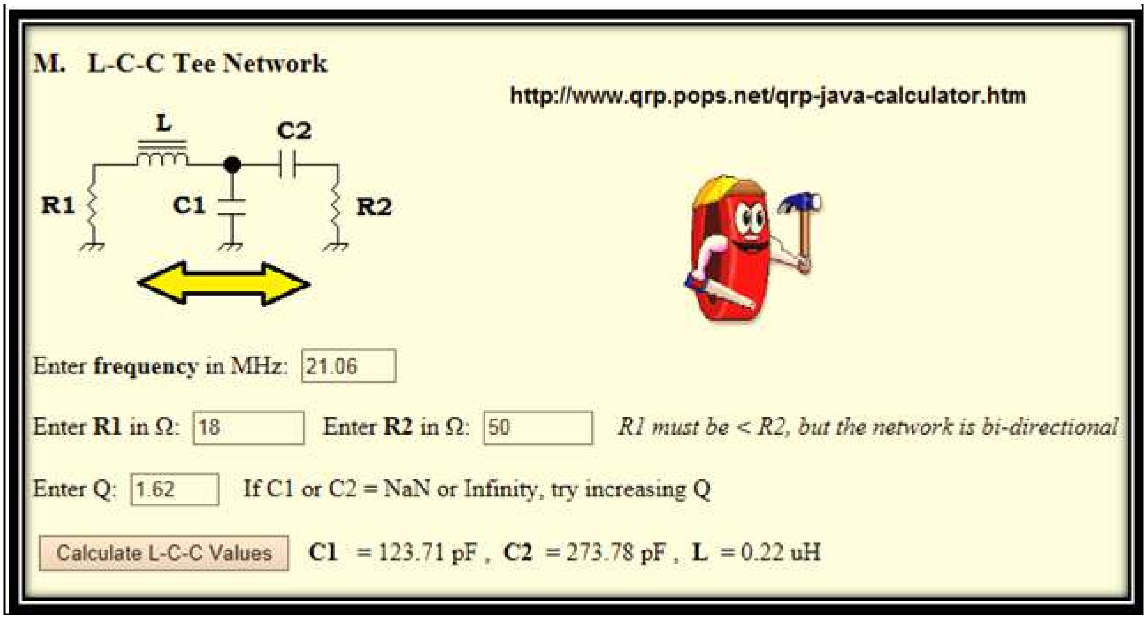

M. L-C-C Tee Network

http://www.qrp.pops.net/qrp-java-calculator.htm

Enter **frequency** in MHz: 21.06

Enter **R1** in Ω: 18 Enter **R2** in Ω: 50 *R1 must be < R2, but the network is bi-directional*

Enter Q: 1.62 If C1 or C2 = NaN or Infinity, try increasing Q

Calculate L-C-C Values **C1** = 123.71 pF , **C2** = 273.78 pF , **L** = 0.22 uH

Above — To calculate L, C1 and C2 we may employ software to obtain starting L and C values. I recommend Wes' ZMAT08.exe, a Ladpac program that ships on the EMRFD CD. I wrote my own java-script applet with less prompting and error checking and put in on the Design Center web page as **M**. Note I used 18 Ω as the load resistance in my calculations.

During actual bench assessment of your network, output your PA to a 50 Ω dummy load rated for that power and preferably measure the peak to peak voltage with a short-grounded 10X probe.

For Figure 7, I wound and tack soldered in a measured L of about 230 nH and placed a 100 pF cap plus 6-50 pF trimmer for C1 and a 220 pF + 22 pF cap + 6-50 pF trimmer cap for C2. The tuning procedure goes like this:

Tune the capacitors for maxmum power. Try squishing the L, then re-tweak C1 and C2 and see what happens. Did the Vpp go up or down? We want it to go up, so if the Vpp went up you needed more L, and perhaps still need more. If you squish the coil again (or remove it, add 1-2 turns and then replace it) and the Vpp goes down, you increased the L too much and need to go back down. If the original Vpp goes up when you expand the coils, you had too much L and need to do the opposite procedure just described. Re-tweak C1 and C2 after every inductance change and calculate the power in dBm and/or watts on your bench.

Through trial and error you'll find the optimal L. If nothing works, suspect a bench error, or device failure. Occasionally, I add too much, or too little fixed C in parallel with the trimmer cap. Another time, my transistor was fried.

Once you've extracted the maximum power, remove the cap(s) that make up C1 and measure them. Do the same for C2. I replace C1 and C2 with the nearest equivalent standard value capacitor or (parallel capacitors) and then repeat the tuning procedure to find the optimal L for these 2 new capacitors. Normally, after optimizing the Vpp, I'll remove and measure L to confirm my L-C-C network design. You can use the L just removed or make a tidier version to permanently solder in.

Once you solder in the "keeper" L, you'll need to expand or compress the links a tad to peak your Vpp once again. Although it sounds like a chore, I love measuring a higher Vpp with each successful manipulation of the L-C-C network; it's quite a rush actually.

In the end, my optimal measured values were: L = 247 nH, C1 = 132.5 pF and C2 = 282 pF. In my final, I substituted C1 = 120 pF and C2 = 270 pF as shown in Figure 7. Although I designed for 4W, I squeezed out 4.49W with my L-C-C network.

Design Question 4: How do I choose my L-C-C matching network Q?

The easy answer is the degree of impedance transformation sets the minimum Q and thus choosing the best Q becomes the central question.

The theory regarding loaded versus unloaded Q + network losses and such may bog down many, but in context, a lower Q leads to a more efficient network with less loss. Low Q values also means we tweak larger value capacitors that minimize the effects of stray capacitance. In the collector output network of Figure 7, the Q of 1.62 gave more power than an earlier Q = 5 network. Further, during design, I sometimes slightly adjust the Q in software so I get closer to standard value caps for C1 and C2.

Don't be afraid to remove C2 and just make an L-network. Again, just experiment on your bench to gain confidence.

I tried making this circuit with a BD139 in the PA slot and managed to squeeze a mere 3.1 Watts power at 12.1 VDC. Others, such as Wes and Jason, NT7S delivered more power into 50 Ω at and even above 21 MHz with a BD139 PA.

Factors such as transistor specifications, frequency, VCC, and driver power may affect the output power. In the end, the 2SC1969, a final I harvested from an old Cobra CB more than 20 years ago delighted me. I constantly scrounge for old driver and PA transistors in my travels. Sadly, many old CB radios given to me over time suffered from fried finals.
You might try checking eBay, but watch out for fakes.

The BD139 grew to notoriety because it's widely available, cheap and generally does the job at or below ~28 MHz, depending on what output power your seek. I found wide variation in Beta (or hFE) and fT in the random batch I tested. Also remember we're hitting them with lots of current and they heat up and specs change. At 7 MHz, the BD139 works great. The BD139 reminds me of the MPF102 JFET; you never know what your going to get. The best I've tested were Fairchild BD139s and woefully, they're all gone.

Class C Power Amp with Broadband Input and L-C-C Tuned Output

Let's make a PA with an L-C-C Tee network tuned input: Figure 8.

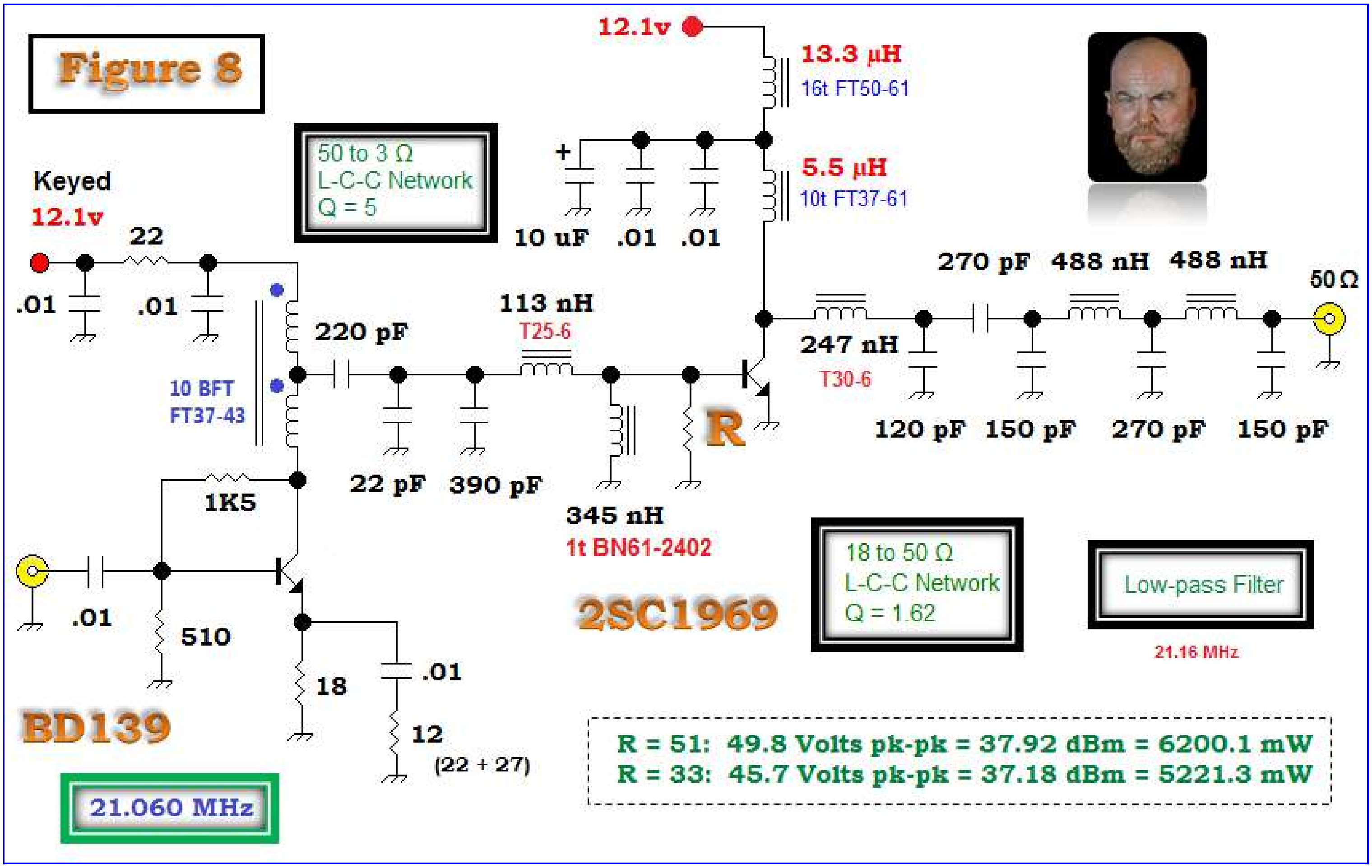
Figure 8
Keyed
12.1v
22
.01
.01
50 to 3 Ω
L-C-C Network
Q = 5
10 BFT
FT37-43
220 pF
22 pF
390 pF
113 nH
T25-6
1K5
.01
510
18
.01
12
(22 + 27)
BD139
21.060 MHz
12.1v
13.3 µH
16t FT50-61
+
10 uF
.01
.01
5.5 µH
10t FT37-61
345 nH
1t BN61-2402
R
2SC1969
247 nH
T30-6
270 pF
488 nH
488 nH
50 Ω
120 pF
150 pF
270 pF
150 pF
18 to 50 Ω
L-C-C Network
Q = 1.62
Low-pass Filter
21.16 MHz
R = 51: 49.8 Volts pk-pk = 37.92 dBm = 6200.1 mW
R = 33: 45.7 Volts pk-pk = 37.18 dBm = 5221.3 mW

Above — Figure 8. The Figure 7 PA with a tuned input network smashed my world with 6.2W power. Click for the FFT showing the second harmonic down at least 55 dB. I later confirmed this with spectrum analysis.

Design Question 5: How do I choose my PA input impedance?

PA Input Impedance

Calculating the input Z of a transistor with 100's of mA emitter challenges us.The familiar calculations like Zin = (Beta +1) x 26 /I e (mA) don't work out perfectly since our transistor is hot and Beta changes, and in my case I'm way up in fT. I asked Wes and he suggested I assume an input Z of 3 Ω (with about 200 pF of parallel C). Thus , I chose a 50:3 Ω transform to design my L-C-C network.

From my experiments, bench tweaking an L-network, or L-C-C network will accommodate anything even close to 3 Ω input impedance — 3 Ω seems quite actionable. Tune your L-C-C network in the manner described for the PA output match. Since your transforming 50 Ω to 3 Ω, the minimum Q will rise.

Design Question 6: How do I choose my PA input choke and resistor when fed with an L-C-C network?

The RFC serves as a DC return path for the PA base just like in the transformer in the wideband version. We normally apply a rule that the minimum inductive reactance of a choke should be 10x the impedance it's connected to, so that's 30 Ω XL.
Since F = 21.06 MHz, the minimum choke value will be 30 Ω / 21.06 MHz x 6.28) = 227 nH. In order to keep the SRF up and for ease, I ran a single turn of 26 gauge wire through a BN61-2402 binocular core and measured 345 nH. Good enough. The # 61 mix will exhibit some Q. Likely # 43 material will work fine too.

The resistor labelled R serves to reduce the Q of *my* particular RFC, lowers gain and boosts stability. With R = 33 - 51 Ω, the amp held stable, but if I tried to increase R above 51 Ω spasm erupted. The resistor affected output power as shown in Figure 8. Again the resistor should minimally be about 10X the input Z or so. I tried an R of 10 Ω and the output power dropped down to a little under 3W as more of my driver output just flowed through the resistor to ground.

To conclude, I executed the time honored tradition of choosing parts and circuits to extract power from a crystal oscillator-driven analog RF amplifer chain. Much of this information extends to linear operation (Class A, AB) and general RF design. Have fun designing your own transmitters and as always, apply experimental methods!

Section 2: PA Measurement with Examples

In this section, I posted some PA measurement ideas + experiments and comments about PA collector chokes.

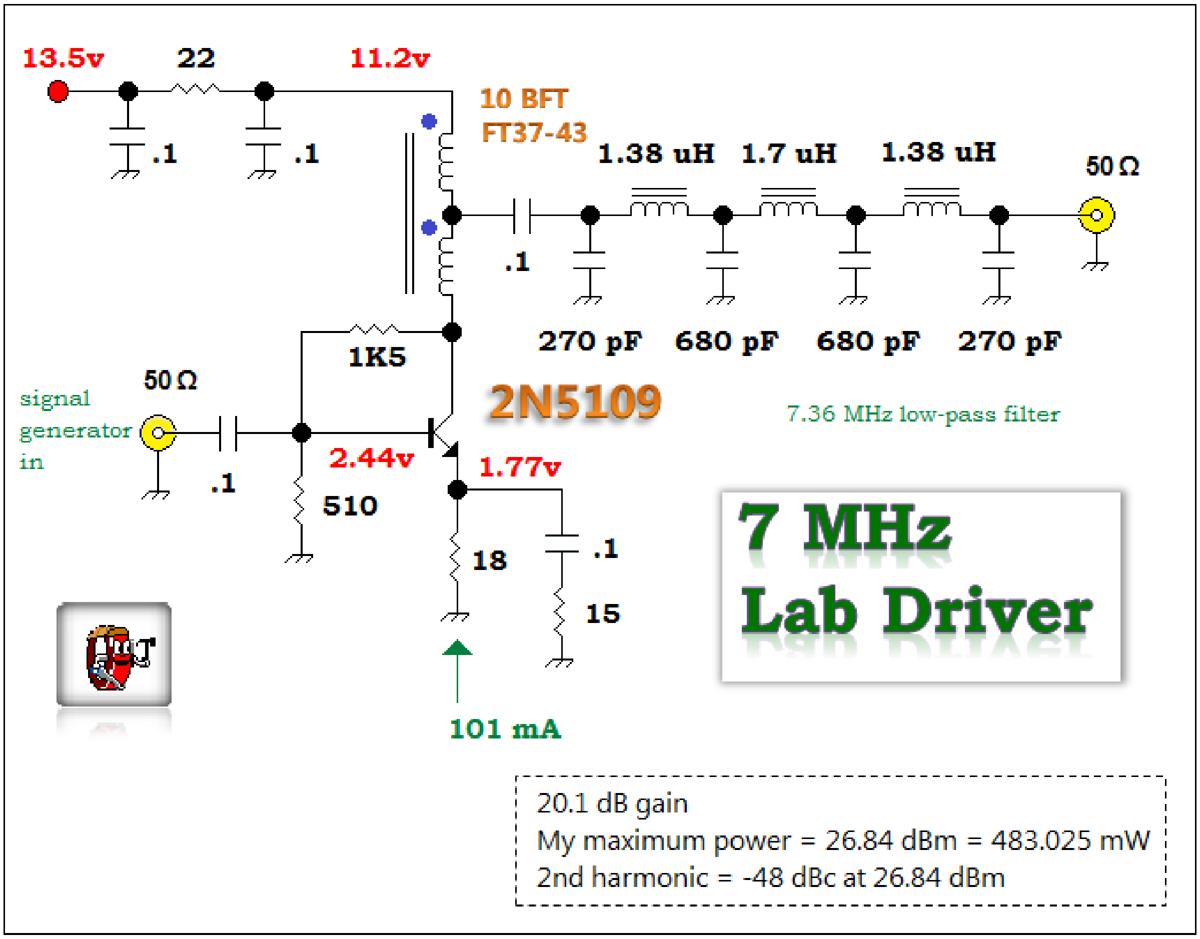
13.5v
22
11.2v
.1
.1
10 BFT
FT37-43
1.38 uH
1.7 uH
1.38 uH
50 Ω
.1
270 pF
680 pF
680 pF
270 pF
1K5
50 Ω
signal
generator
in
.1
2N5109
7.36 MHz low-pass filter
2.44v
1.77v
510
18
.1
15
101 mA
7 MHz
Lab Driver

20.1 dB gain
My maximum power = 26.84 dBm = 483.025 mW
2nd harmonic = -48 dBc at 26.84 dBm

Above — To boost my signal generator output for driving *any* experimental PA up to about ~20W output, I built a high-current, sine wave driver. I excite this amplifer with my variable gain homebrew HF signal generator — in the version shown above, a 7 MHz low-pass filter is soldered on the amplifier breadboard. I'll stick this amp in a metal box sans filter and then apply the correct bench module low-pass filter via a BNC connector. Click for a photo of 3 of my HF low-pass filter bench modules.

Normally the low-pass filter output is pad attenuated to buffer and reduce the output voltage since it's rare I need up to 483 mW drive on the QRP workbench.

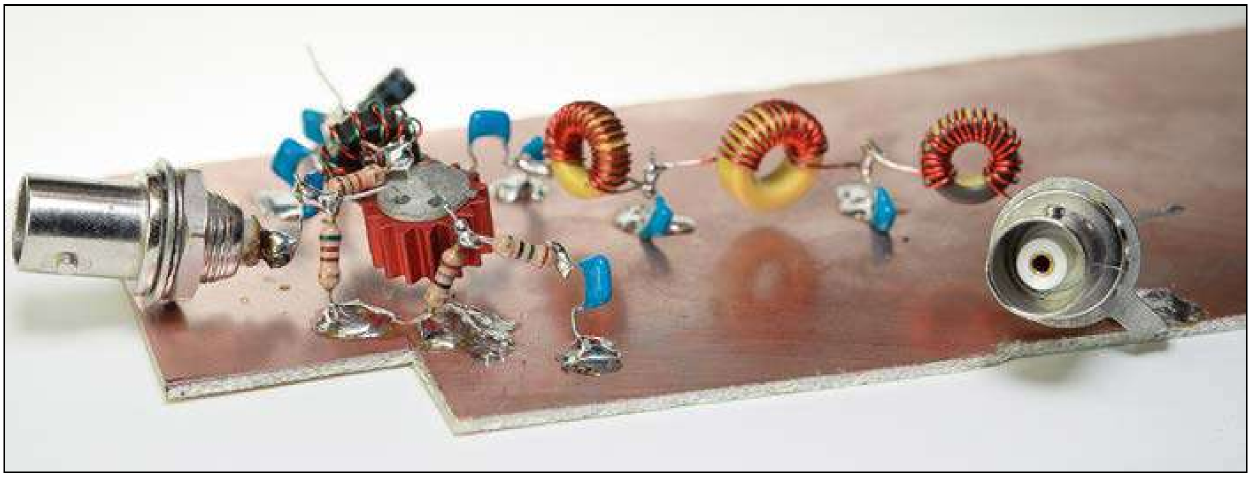

Above — A photo of the PA driver amplifer breadboard.

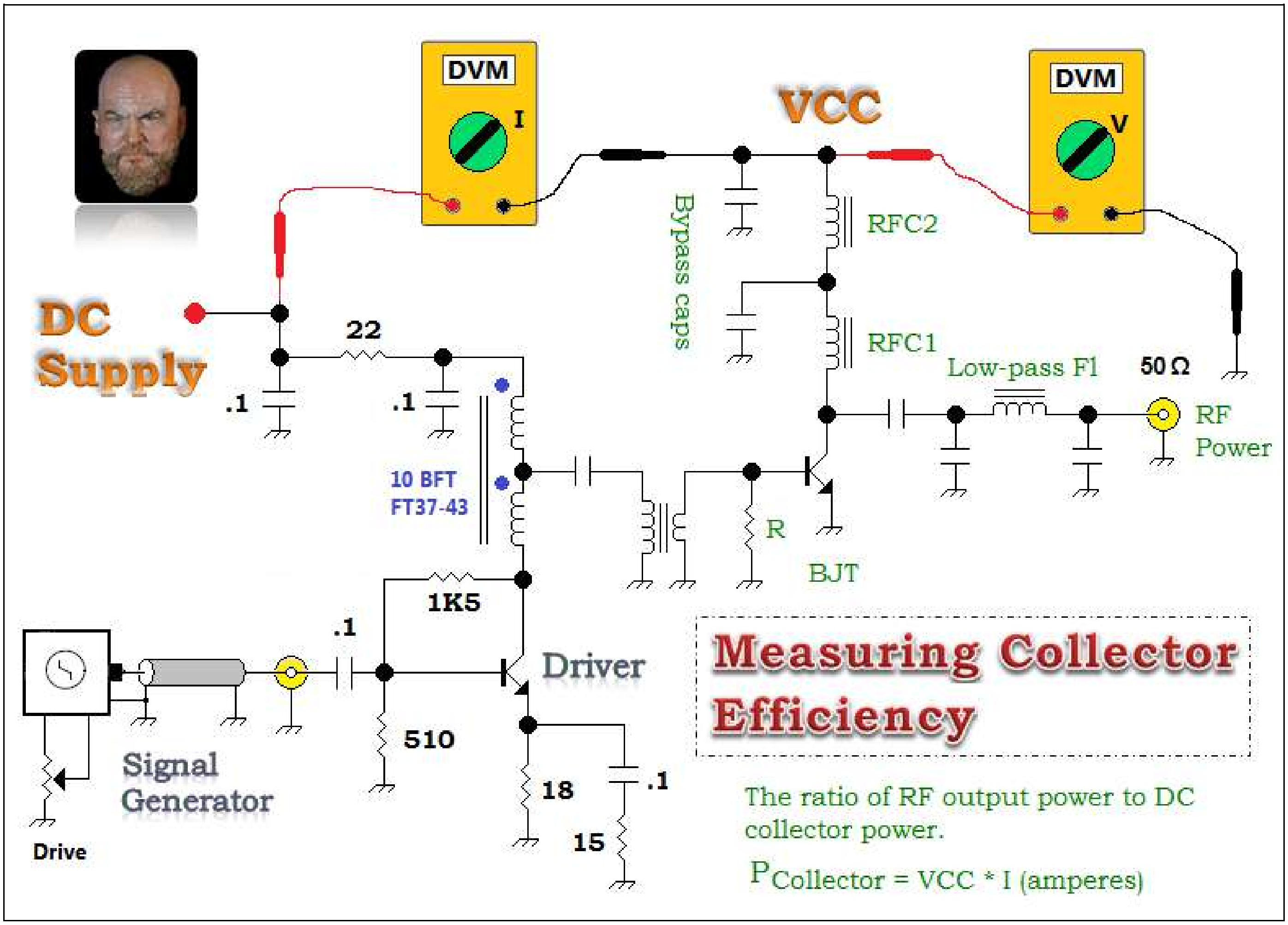
DVM
I
VCC
DVM
V
Bypass caps
RFC2
RFC1
DC Supply
22
.1
.1
Low-pass Fl
50 Ω
RF Power
10 BFT
FT37-43
R
BJT
1K5
.1
Driver
Measuring Collector Efficiency
The ratio of RF output power to DC collector power.
$P_{Collector}$ = VCC * I (amperes)
Signal Generator
510
18
.1
15
Drive

Above — A method to simultaneously measure collector **V** and **I** to calculate DC collector power. Two DVMs boost measurement fidelity since the current measuring ammeter's internal resistance will drop the DC supply ~ 0.2 to 0.5 volts depending on the device and current. Collector efficiency = RF output power / (V * I) .

This method provides a great excuse to purchase another multimeter.

Measure VCC where shown since the decoupling and bypass filters help prevent the AC signal from affecting the DC measurement. We only touch our 10X 'scope probe on the collector. Exert due care and wear eye protection in case you blow up the transistor and create shrapnel.

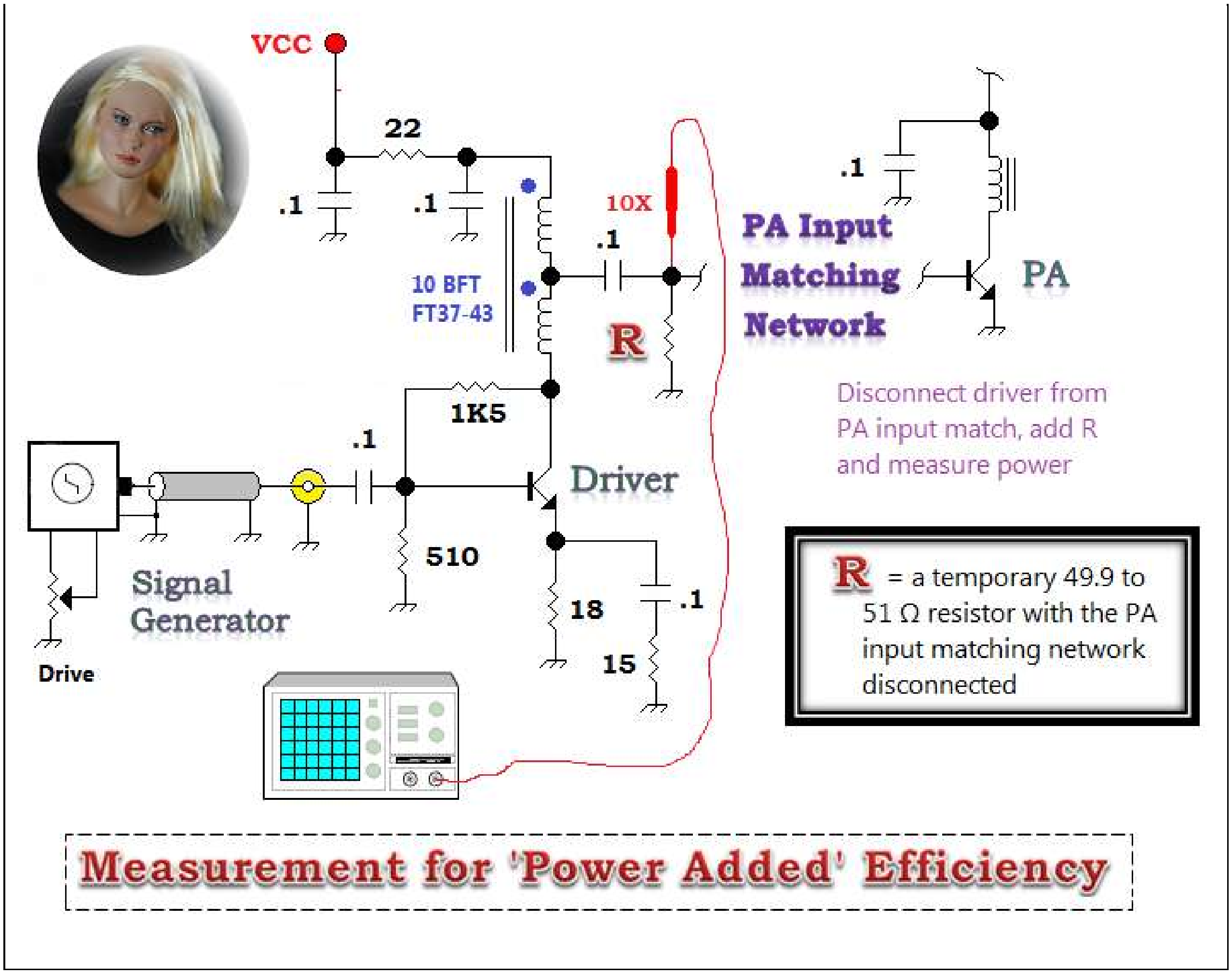
VCC
22
.1
.1
10 BFT
FT37-43
10X
.1
R
PA Input
Matching
Network
.1
PA
1K5
.1
Driver
510
18
.1
15
Signal
Generator
Drive
Disconnect driver from
PA input match, add R
and measure power
R = a temporary 49.9 to
51 Ω resistor with the PA
input matching network
disconnected
Measurement for 'Power Added' Efficiency

Above — Power Added Efficiency considers the base drive power in addition to the DC collector power. The formula = RF output power / (DC Power + RF drive power). Although uncommon, measuring drive power helps you to assess and compare both input matching and PA efficiency metrics.

At the very least — I like knowing my PA drive power and so employ driver amplifiers with 50 Ω output impedance to allow this power measurement. I often permanently solder a 49.9 or 51 Ω resistor to the breadboard ground plane and temporarily connect this R ; plus disconnect the input matching circuit to measure driver power when I've finished tweaking my collector to antenna port circuitry.

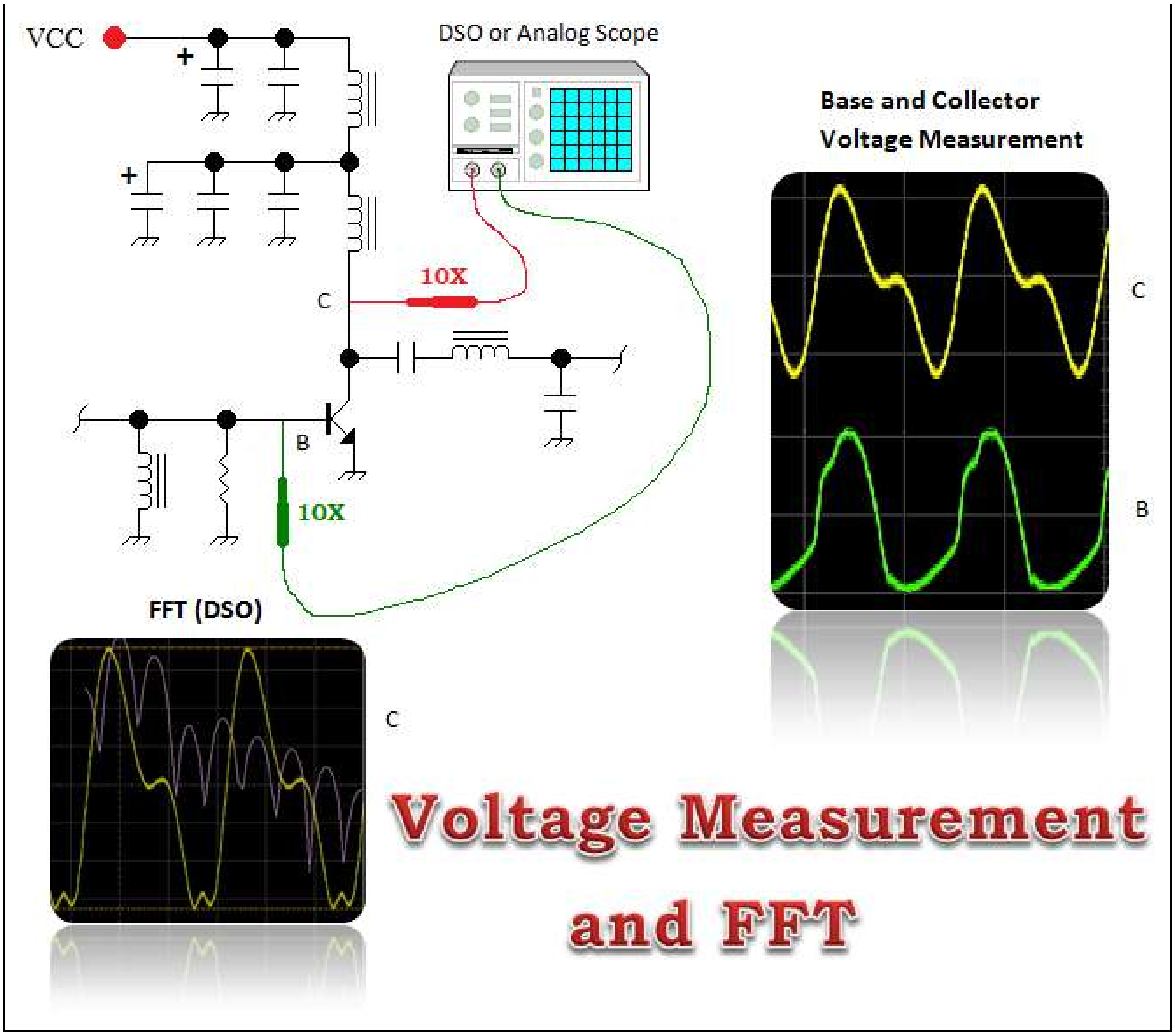
VCC
+
+
DSO or Analog Scope
Base and Collector
Voltage Measurement
10X
C
B
10X
C
B
FFT (DSO)
C
Voltage Measurement
and FFT

Above — Oscillocope measurements of the base and collector voltages. Most of us enjoy simultaneously running 2 channels on our 'scopes. With a good quality DSO, FFT analysis of the collector waveform may yield insight, however, this knowledge might also confound us since interpreting our measures is fully half of the task.

I often struggle to properly interpret measures. Some builders blog about or email me that they enjoy an innate gift at understanding electronic circuitry — I don't. I guess I lack the knack? For some of us, problems and complications block progress, but more often than not, we eventually cross the finish line.

Above — 1.5 Watt, Class C breadboard used for the experiments that follow. I changed the collector choke (RFC1) and applied various filters to examine some of the measurement techniques aforementioned. Click for a bigger photo.

The Collector Choke

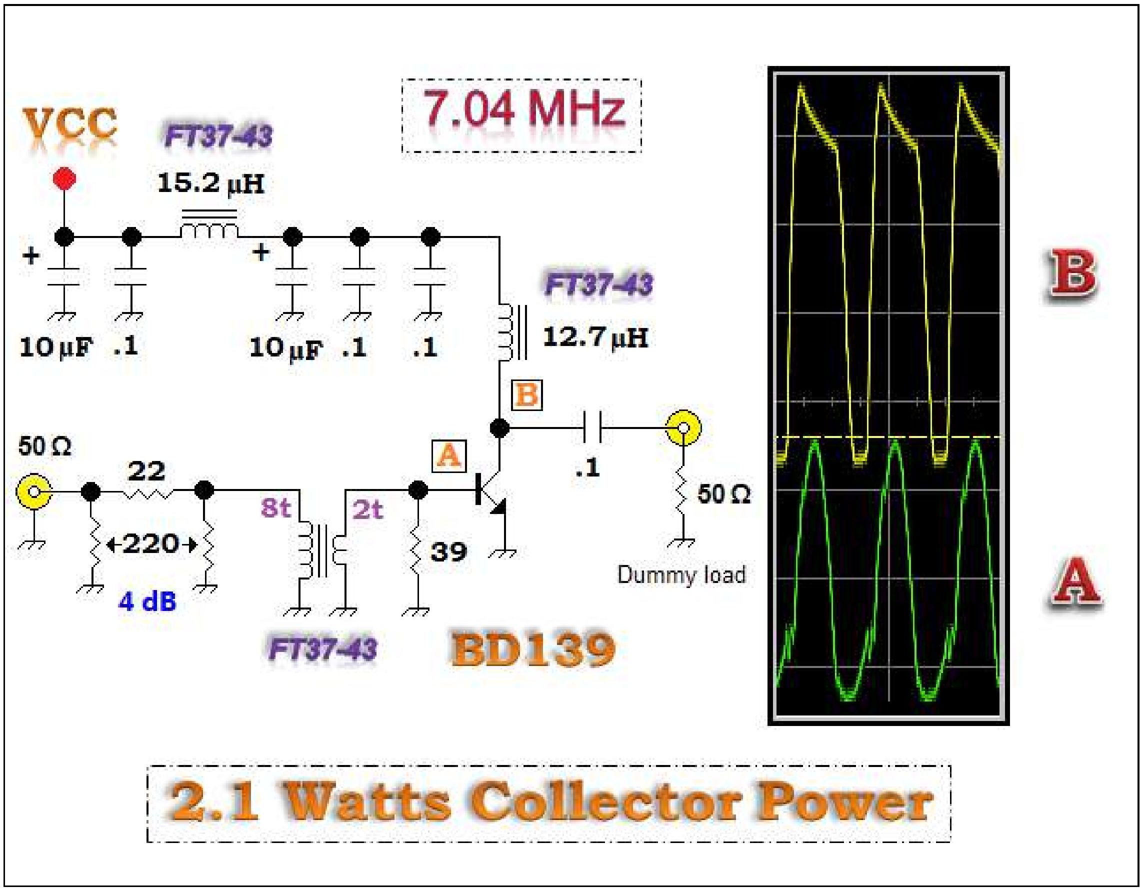
VCC
FT37-43
15.2 μH
7.04 MHz
+
10 μF
.1
+
10 μF
.1
.1
FT37-43
12.7 μH
B
50 Ω
22
220
4 dB
8t
2t
FT37-43
39
A
.1
50 Ω
Dummy load
BD139
B
A
2.1 Watts Collector Power

Above — A Class C amplifer with it's collector AC coupled to a 50 Ω resistor load. *No filter.*

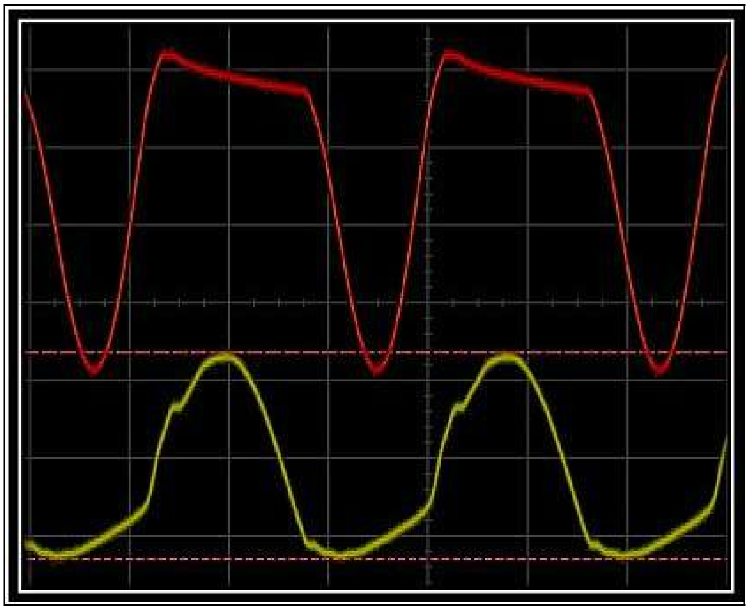

Above — Base and collector waveforms. The unfiltered collector waveform (top) rich in even and odd harmonics ~resembles a square wave. The maximum collector voltage occurs when the current is 0 and in concert, the maximum collector current flows where collector voltage falls to near 0. As the BJT shuts off, the collector circuit kicks and rings so that the DC collector voltage may rise as high as twice VCC.

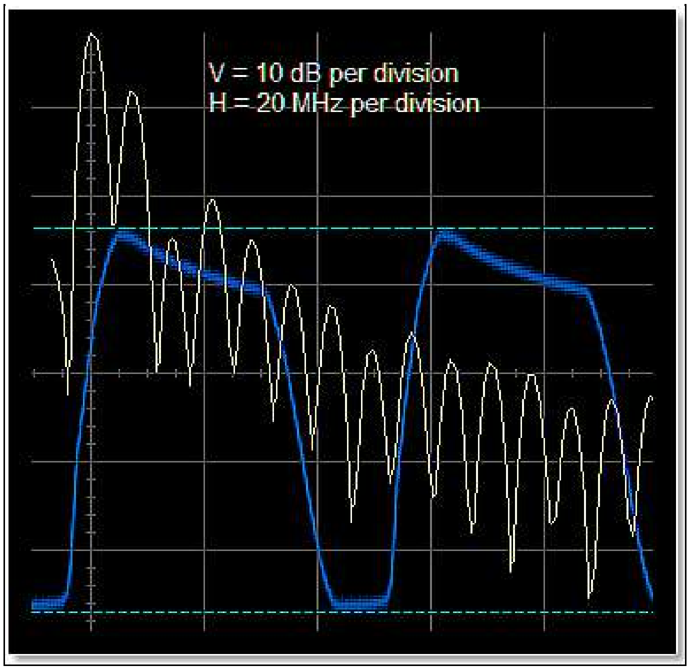

Above — An FFT of the unfiltered collector showing F1, F2, F3, F4, F5, etc. Distortion city!

The Collector Choke

I've read web pages suggesting we might boost PA efficiency by decreasing the collector choke inductance. While that's possible, in reality, it's not that simple. Experiments involving factors like collector load resistance, angle of conduction, RFC inductance/Q, matching networks, base drive configuration and output filtration are ripe with complicated math and theory that I struggle to understand — entire books and research papers explore these topics as author(s) configure their PA into Class D, E, etc. to boost efficiency. The published design work and PA devices out today are truly breath taking.

We should likely think of the BJT, collector current + voltage, the RFC, base drive and output networks as a system of parts that interact — changing 1 part might change collector current, impedance, efficiency, harmonics, etcetera. As such, the experiments that follow do not construe science; only some experiments by a neophyte to ponder and build on.

The ideal RFC passes only DC — i.e. exhibits infinite impedance to AC, but a short circuit to DC. The primary purpose of the choke is to block our high-power, distorted AC signal from flowing on our DC power lines. Some authors suggest that the minimum inductance for the inductor should be 5-10X the collector resistance and I'll leave that up to you. At HF, I wind my RFCs on #43 or #61 material ferrites while keeping the turns number low as I remember that the coil will exhibit a self resonant frequency that might cause problems if it lies near a frequency I need to filter.

1.5 Watt Amp with Three Different Collector Chokes

The 1.5 Watt obviates the need for a matching network since the collector load resistance = 48 Ω with a VCC of 12v. I rebelled and ran 13.5 V supply for a calculated 60.8 Ω collector load R. I probably should have this matched this collector load to 50 Ω at 13.5v, but wanted a "real world" amp without a matching network. At the very least, these amplifiers illustrate the PA measurement techniques shown above. I wound the RFC on a common FT37-43.

In all cases, I adjusted the base drive so that the sine wave PA output read ~24.5 Vpp while measuring collector V and I. I made DSO measures of the the base and collector voltage and collector FFTs, although forgot to save some of them as screen captures.

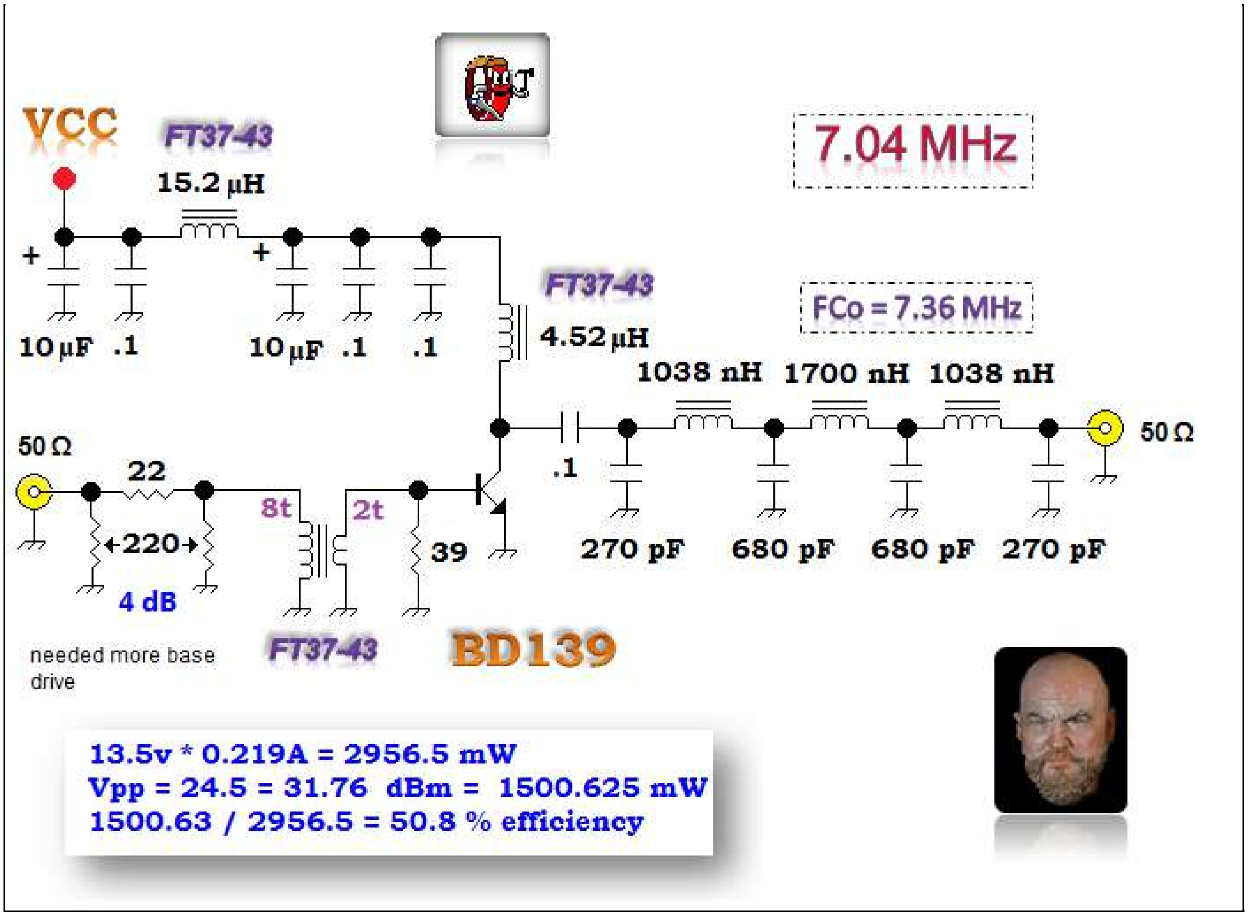
VCC
FT37-43
15.2 µH
+
10 µF
.1
+
10 µF
.1
.1
7.04 MHz
FT37-43
4.52 µH
FCo = 7.36 MHz
1038 nH
1700 nH
1038 nH
.1
50 Ω
22
220
4 dB
8t
2t
39
270 pF
680 pF
680 pF
270 pF
50 Ω
FT37-43
BD139
needed more base drive

13.5v * 0.219A = 2956.5 mW
Vpp = 24.5 = 31.76 dBm = 1500.625 mW
1500.63 / 2956.5 = 50.8 % efficiency

Above — The collector choke = 4.52 µH or an XL of ~ 200 Ω @ 7.040 MHz. My measurements indicate an efficiency of 50.8%. With a choke XL of only ~ 200 Ω, some of my collector AC might be flowing to ground through the bypass caps and that'll kill efficiency. Click for an FFT of the collector. This circuit required more base drive than the other 2.

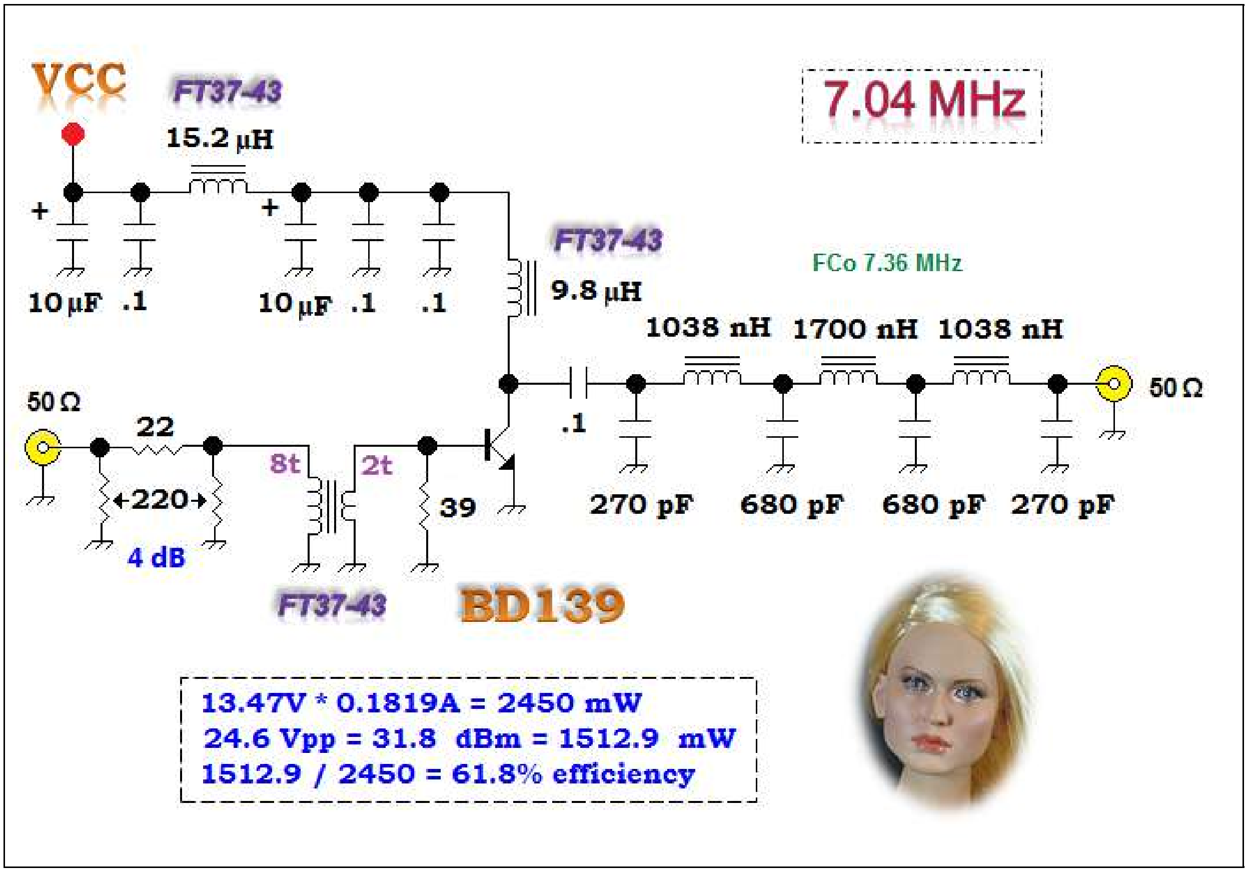
VCC
FT37-43
15.2 µH
+
10 µF
.1
+
10 µF
.1
.1
7.04 MHz
FT37-43
9.8 µH
FCo 7.36 MHz
1038 nH
1700 nH
1038 nH
.1
50 Ω
22
220
4 dB
8t
2t
39
270 pF
680 pF
680 pF
270 pF
50 Ω
FT37-43
BD139

13.47V * 0.1819A = 2450 mW
24.6 Vpp = 31.8 dBm = 1512.9 mW
1512.9 / 2450 = 61.8% efficiency

Above — The collector choke = 9.8 µH = an XL of ~ 434 Ω. Efficiency now 61.8%. Base and collector tracing. Collector FFT.

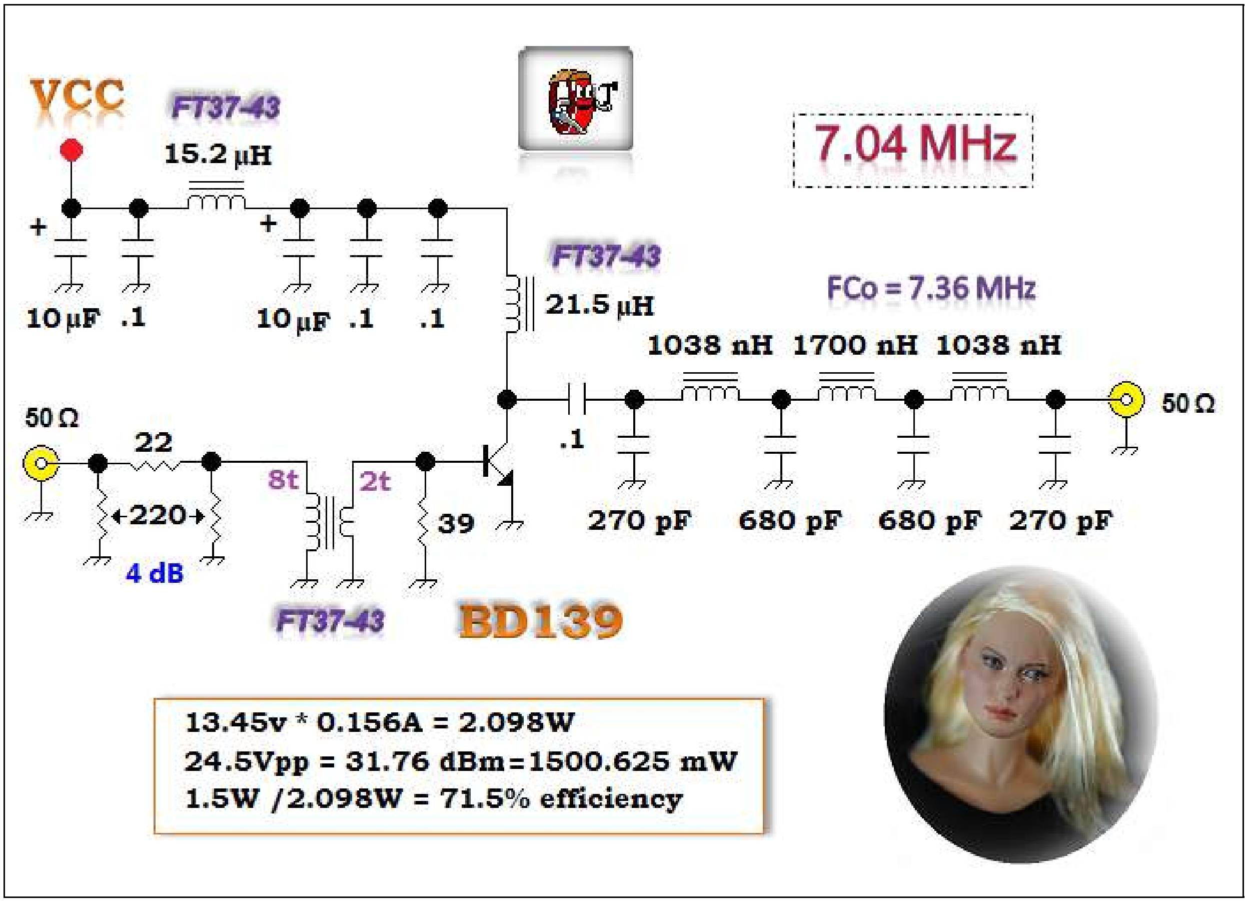
VCC
FT37-43
15.2 µH
+
+
10 µF
.1
10 µF
.1
.1
7.04 MHz
FT37-43
21.5 µH
FCo = 7.36 MHz
1038 nH
1700 nH
1038 nH
50 Ω
50 Ω
22
.1
8t
2t
220
39
270 pF
680 pF
680 pF
270 pF
4 dB
FT37-43
BD139
13.45v * 0.156A = 2.098W
24.5Vpp = 31.76 dBm=1500.625 mW
1.5W /2.098W = 71.5% efficiency

Above — The collector choke = 20.5 µH = an XL of ~ 951 Ω. 71.5% efficiency. Base + collector. Collector FFT. Sine wave FFT. To date, 71.5% is the best Class C efficiency I've measured.

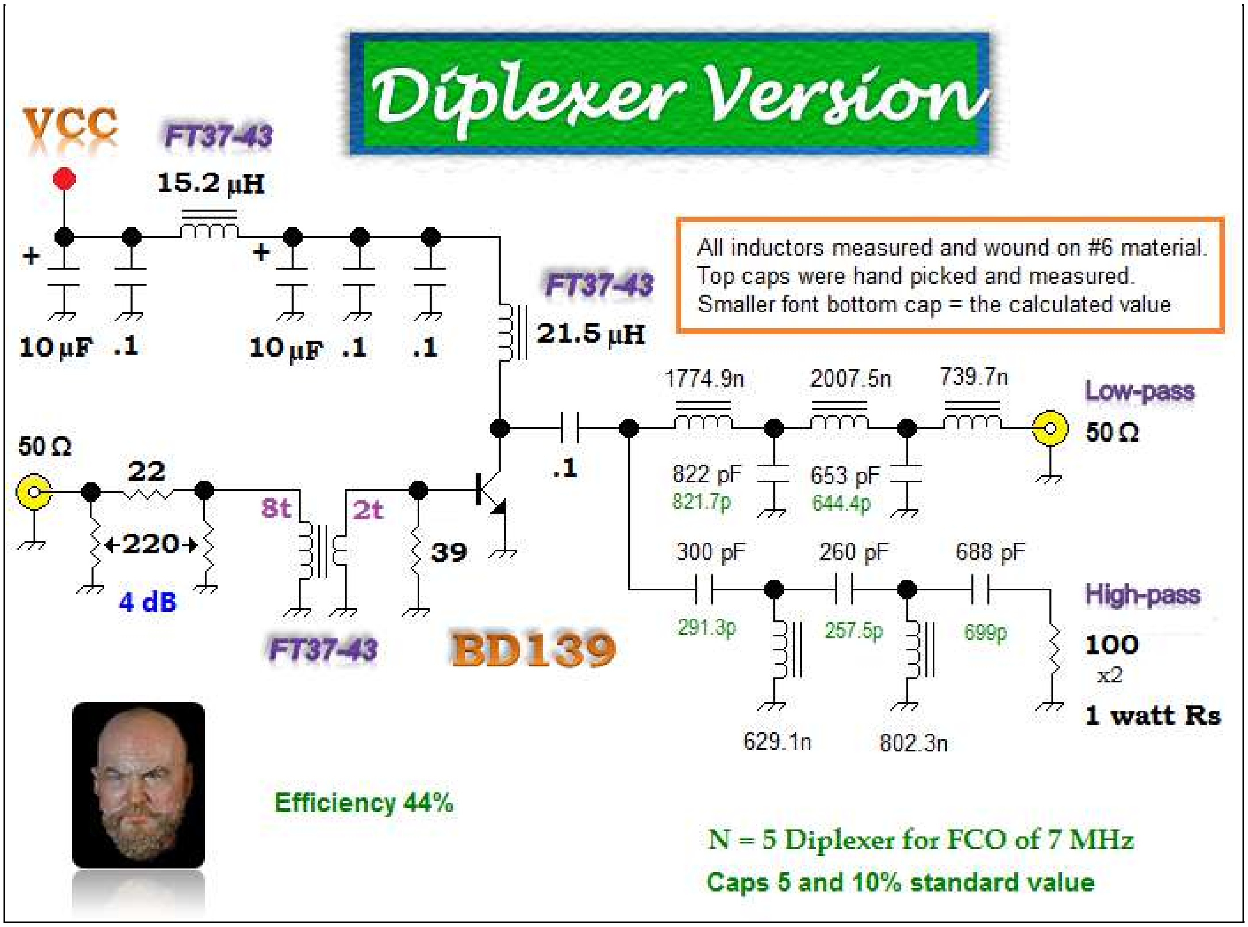
Diplexer Version
VCC
FT37-43
15.2 µH
+
10 µF
.1
+
10 µF
.1
.1
FT37-43
21.5 µH
All inductors measured and wound on #6 material.
Top caps were hand picked and measured.
Smaller font bottom cap = the calculated value
1774.9n
2007.5n
739.7n
Low-pass
50 Ω
822 pF
821.7p
653 pF
644.4p
50 Ω
22
.1
220
4 dB
8t
2t
39
FT37-43
BD139
300 pF
260 pF
688 pF
291.3p
257.5p
699p
High-pass
100
x2
1 watt Rs
629.1n
802.3n
Efficiency 44%
N = 5 Diplexer for FCO of 7 MHz
Caps 5 and 10% standard value

Above — A PA terminated in a N= 5 diplexer for 7 MHz. Click for the FFT of the collector. The collector waveform looks telling — resembling the 50 Ω resistor terminated version shown earlier. Unshown; I lowered the signal generator to 6.99 MHz and drove it to a sinusoidal 1825 mW to obtain a true 50 Ω collector load resistance plus get the greatest low-pass output signal and efficiency (44%) . Although diplexers will 50 Ω terminate the collector at all frequencies, they don't afford high collector efficiency.

Rogues Gallery

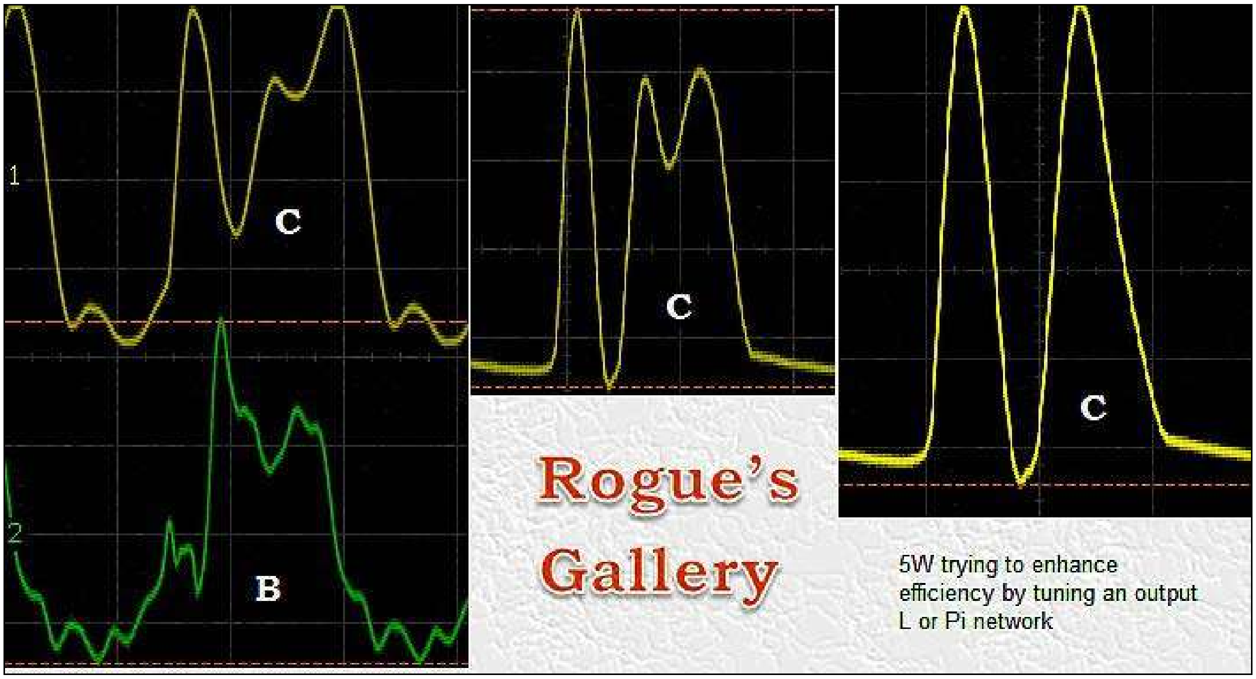

Above — Some 5W CW amplifier collector waveforms as I tried to boost efficiency by manipulating a matching network to drive up the ringing; all failures. These seemingly simple Class C PA's aren't so simple after all. Although failures — I felt thrilled by these experiments and vow to get better at it.

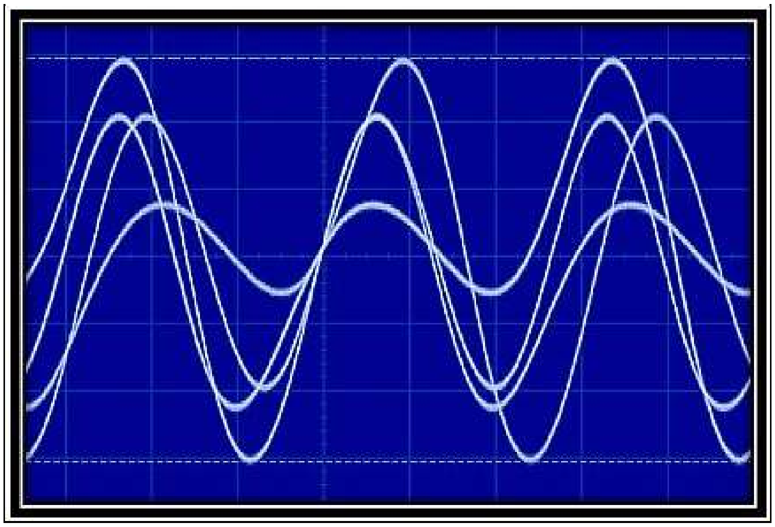

Above — A sine wave output of a very unstable PA.

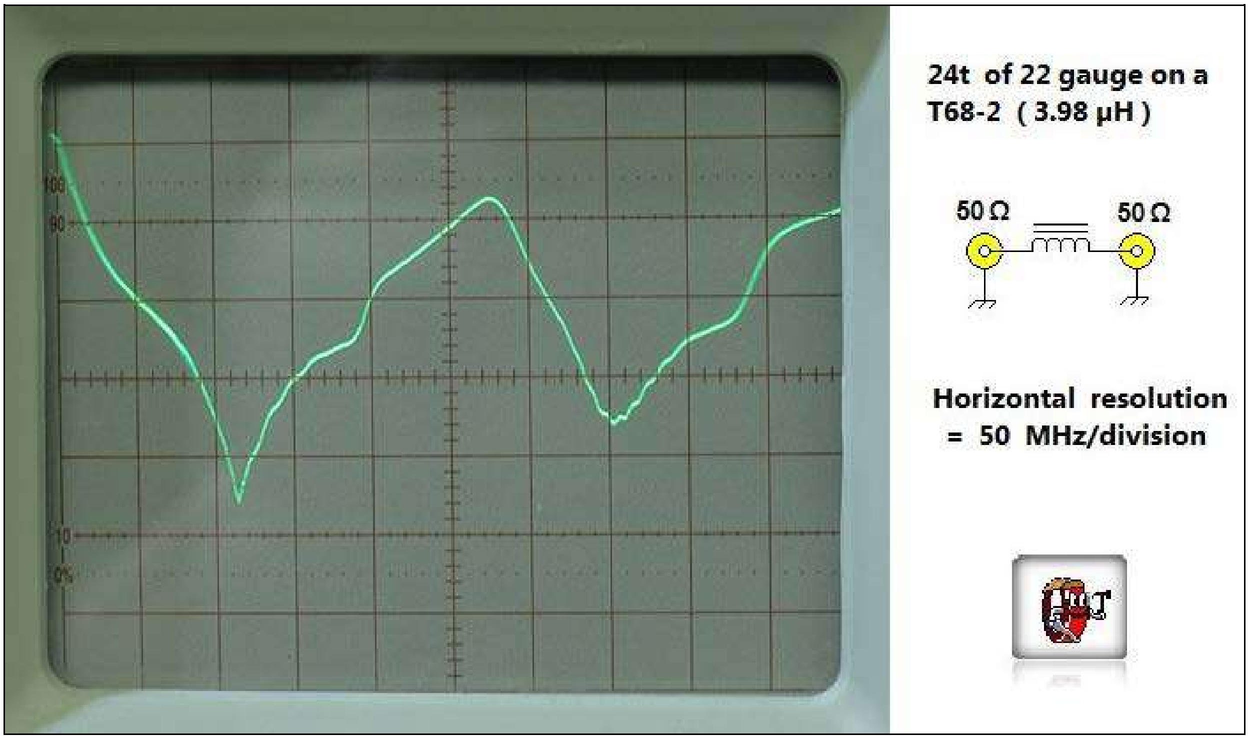

Above — To observe a renegade series resonant frequency, I wound a choke for a 21 MHz amplifer on a T68-2 powdered iron toroid: a casual 24 turns of wire. Sweeping it shows 2 SRFs at roughly 115 and 350 MHz. 2 SRFs jarred me.

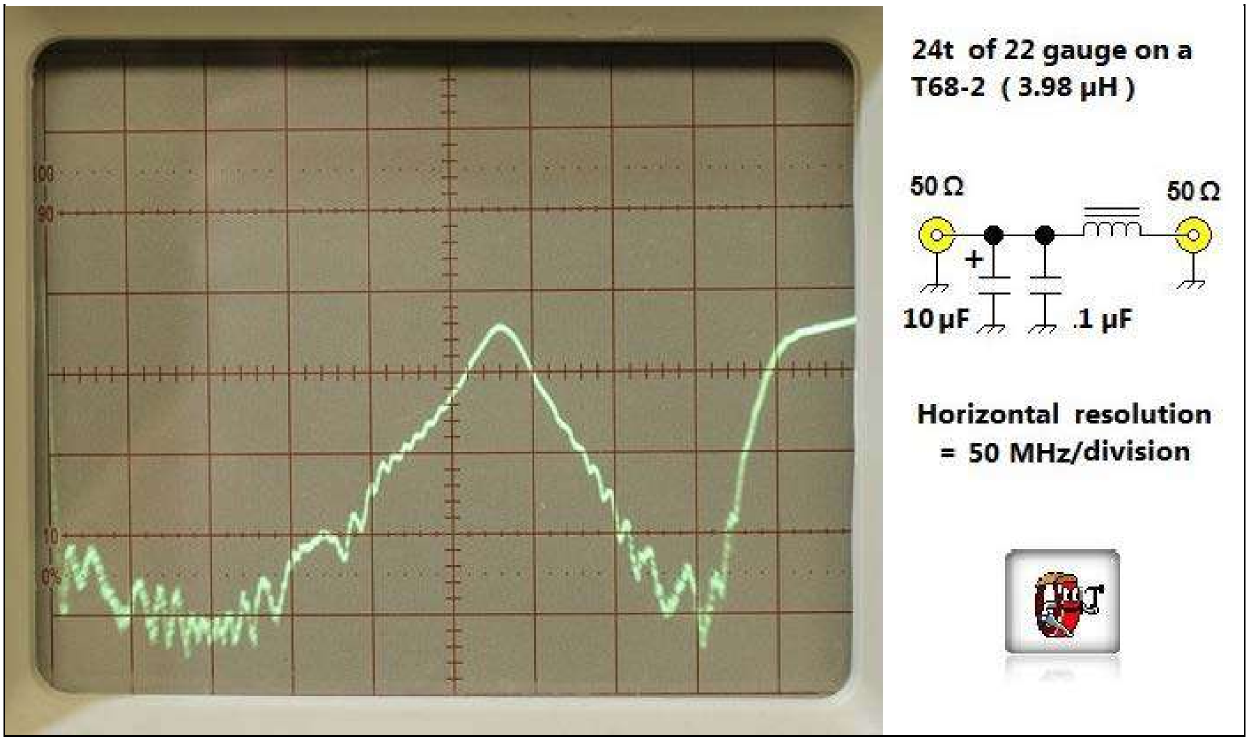

Above — I placed a couple of caps on the cold side like we normally do for our PA chokes. Likely I should have used a 0.01 µF cap instead of the 0.1 µF shown for 21 MHz, but the result would be the same: 2 SRFs remain and poor wideband bypass arises. We clearly need to consider SRF in our chokes and how it might affect our PA function and DC line filtering.

Conclusion to PA Measurement

With an oscilloscope and 2 DVMs you can easily explore boosting your Class C PA efficiency. I've seen precious little info on this topic in the amateur RF experimenter literature. 1 reputable reference = a section in EMRFD called **A Look at Some High Efficiency Amplifers**. For now, I'll temporize. Likely you'll advance faster than me, since I'm more a receiver guy and don't devote enough time to study transmitter design.

The math and concepts of efficient PA design within Class C - E flummox me. Most have just switched to driving FETs with a square wave built up from parallel inverters (high-speed logic gates in the now obsolete 74AC series) — but hopefully our experiments tweaking BJTs in Class C will carry us forward into other Tx modes.

Measuriing collector power with 2 DVMS while watching the sine wave output in your 'scope gives you incredible feedback: you can even observe thermal runaway. After purposely advancing the base drive to "heat up" a BJT, I saw the Vpp progressively drop in the 'scope, the DC voltmeter drop somewhat and then the DC current meter 'take off' : 400, 500, 890, 1200 mA and suddenly, poof, the BJT was on fire, exploded and flung 2 leads into the air. Normally, they just smoke and stink up your lab. I put a picture of a couple of blown BD139s in the photo section below. Expect to burn up some BJTs during experiments and please work safely.

Section 4: Class E Amplifier

This section unposted.

Section 5: References

1. Solid State Design for the Radio Amateur by Wes Hayward, W7ZOI and Doug DeMaw W1FB. American Radio Relay League 1986.

2. An Updated Universal QRP Transmitter by Wes Hayward, W7ZOI. American Radio Relay League: QST for April 2006.

3. Class E Amplifiers Part 1: Class E Basics. Dan Tayloe, N7VE / Ozarkcon Class E Presentation, 1994. *Search for it online*

4. The Handiman's Guide to MOSFET "Switched Mode" Amplifiers by Paul Harden, NA5N Part 1 Introduction to Class C,D,E and F. First Published in the journal "QRPp". Paul's Site

---

Section 6: Miscellaneous Figures and Photos

8047
C3039
M0H4

2N3553
K0350
2N3553
K0350

2I 06000
WC9838

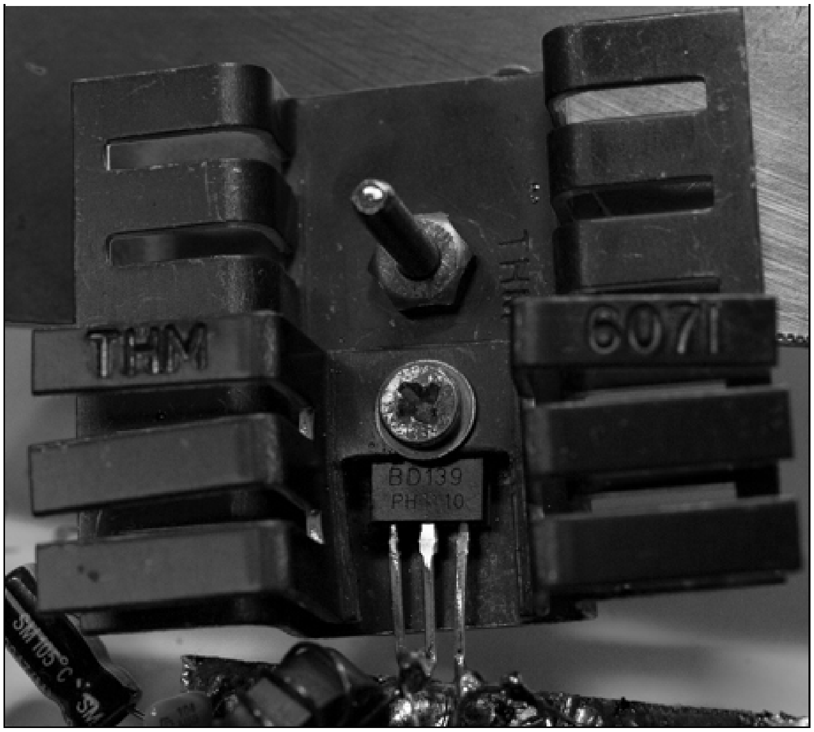
THM
6071
BD139
SM 105°C

RF — Test and Measurement

HF - Embarcadero

Respository web page for HF experiments performed in Fall 2013 and Winter 2014.

1. IMD Oscillator Quest
2. Almost Popcorn Superhet — an older, variant and incomplete HF project added for sake of interest
3. VXO — VCXO Notes
4. Funster Line — A Simple 40M Band QRP Trans-receiver.
5. Miscellaneous Photos and Figures

---

1. IMD Oscillator Quest

To measure IMD and calculate IP3 in receivers and components, we need 2 oscillators that typically lie 20 KHz apart, although some apply other frequency spacings. We join these oscillators with a 6 dB (or 3 dB) hybrid combiner and often follow the combiner with an RF attenuator and/or low-pass filter before connecting to a DUT and detector. I sought an oscillator possessing low phase noise, low distortion, strong return loss + high reverse isolation and documented my experiments to fuel your own. Nestled half-way within the HF band, 14 MHz is a popular IMD oscillator frequency choice. I chose 14.0 MHz and 14.020 MHz for classic 20 KHz spacing.

Wes, W7ZOI well covers intercept measurements in EMRFD and in part, popularized receiver DR measurements with his QST article entitled Defining and Measuring Receiver Dynamic Range in QST for July 1975. Click for the article header only. After publishing his article, receiver DR testing by the major Ham transceiver manufacturers became standard.

If you own 2 crystals spaced 20 KHz apart, you're set. A Butler oscillator set up for vibration at the xtal fundamental frequency works well. See EMRFD Figure 7.32 for a great example — A Butler carefully biased to stand reasonable current for low phase noise with a low impedance in the Colpitt's tank. Wes placed a 100 Ω resistor on the link between base terminals of the differential pair. The low impedance and the bias levelling provides current limiting and prevents the oscillator BJT from going into saturation.

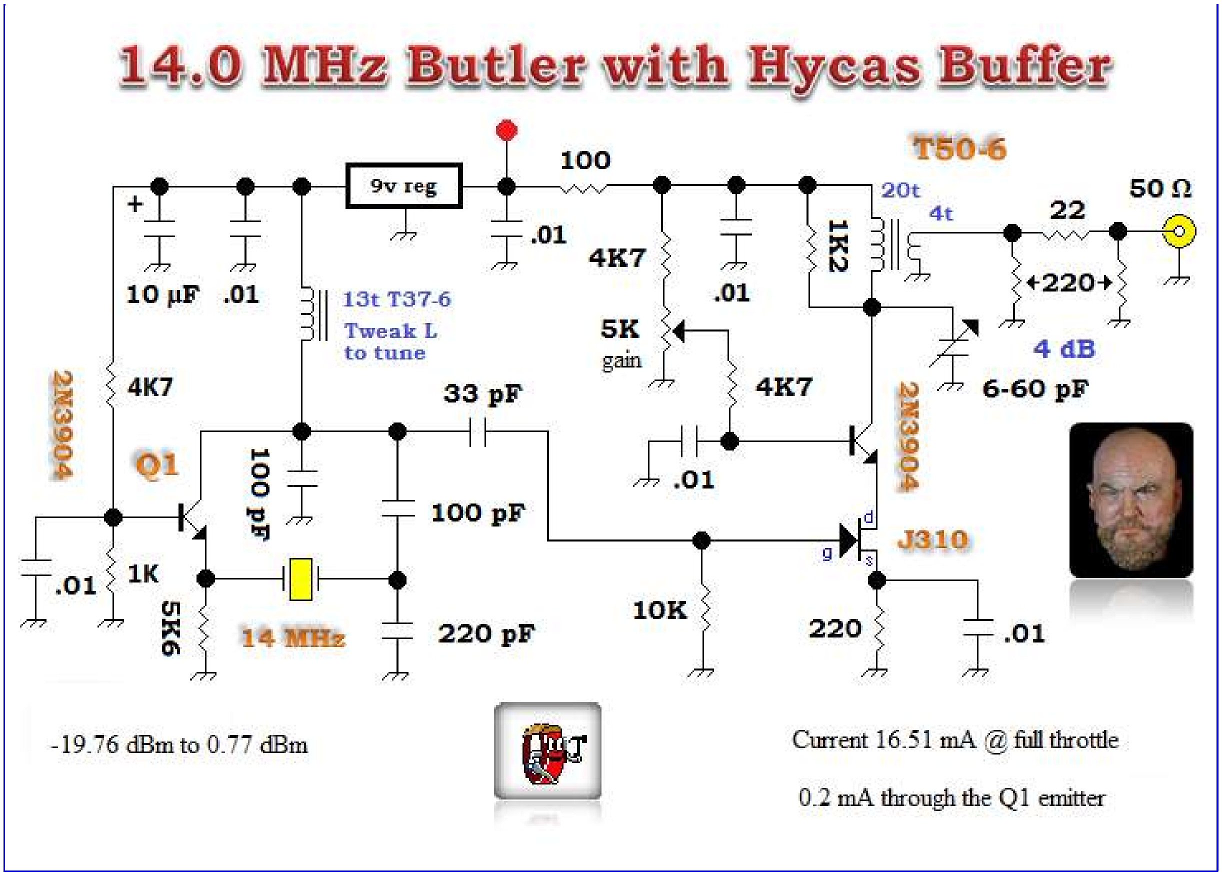
14.0 MHz Butler with Hycas Buffer
T50-6
9v reg
100
50 Ω
20t
4t
22
+
.01
4K7
1K2
10 µF
.01
13t T37-6
Tweak L
to tune
.01
220
5K
gain
4 dB
2N3904
4K7
4K7
6-60 pF
Q1
33 pF
2N3904
100 pF
.01
100 pF
d
J310
g
s
.01
1K
5K6
14 MHz
220 pF
10K
220
.01
-19.76 dBm to 0.77 dBm
Current 16.51 mA @ full throttle
0.2 mA through the Q1 emitter

Above — A fundamental frequency Butler with hybrid-cascode (hycas) buffer I built for 14.0 MHz some years ago. Rather than tuning the oscillator tank with the usual low-Q trimmer capacitor, I designed for a fixed C of ~100 pF and thumb and finger tuned the tank to preserve the Q.

Alternately, a standard Colpitt's crystal oscillator with the crystal connected to ground via a shunt capacitor of >= 200 pF provides a great way to get a low phase noise, low distortion oscillator. Take the oscillator output from the junction of the xtal and shunt capacitor and ensure you lightly load this node. See EMRFD Figure 4.24 or my Low Noise Crystal Oscillator web page for examples.

A problem arose. I couldn't find 2 crystals 20 KHz apart, so running 2 fundamental Butler or Colpitts oscillators for IMD measurement seemed out of reach. A web search yielded 1 solution; pull the xtal oscillators to a desired spacing around 14 MHz and crystal filter the output to drop phase noise.

My next experiments mirrored the work of Christian, DL9NL on this web site. It appears he based his design on that of Stuart, KI6QP who authored A Precision Two-Tone RF Generator for IMD Measurements for QEX for April 1995. Stuart used crystals cut 20 KHz apart to make 2 Colpitt's oscillators running moderate current with output filtering by 1 xtal filter tuned to the oscillator frequency to reduce phase noise.

In his project, Christian, DL9NL wired up two 14.318 MHz microprocessor filters in a super VXO topology to cleverly move the xtals 5 or 20 KHz apart. He also output filtered his 2 oscillators with a xtal filter.

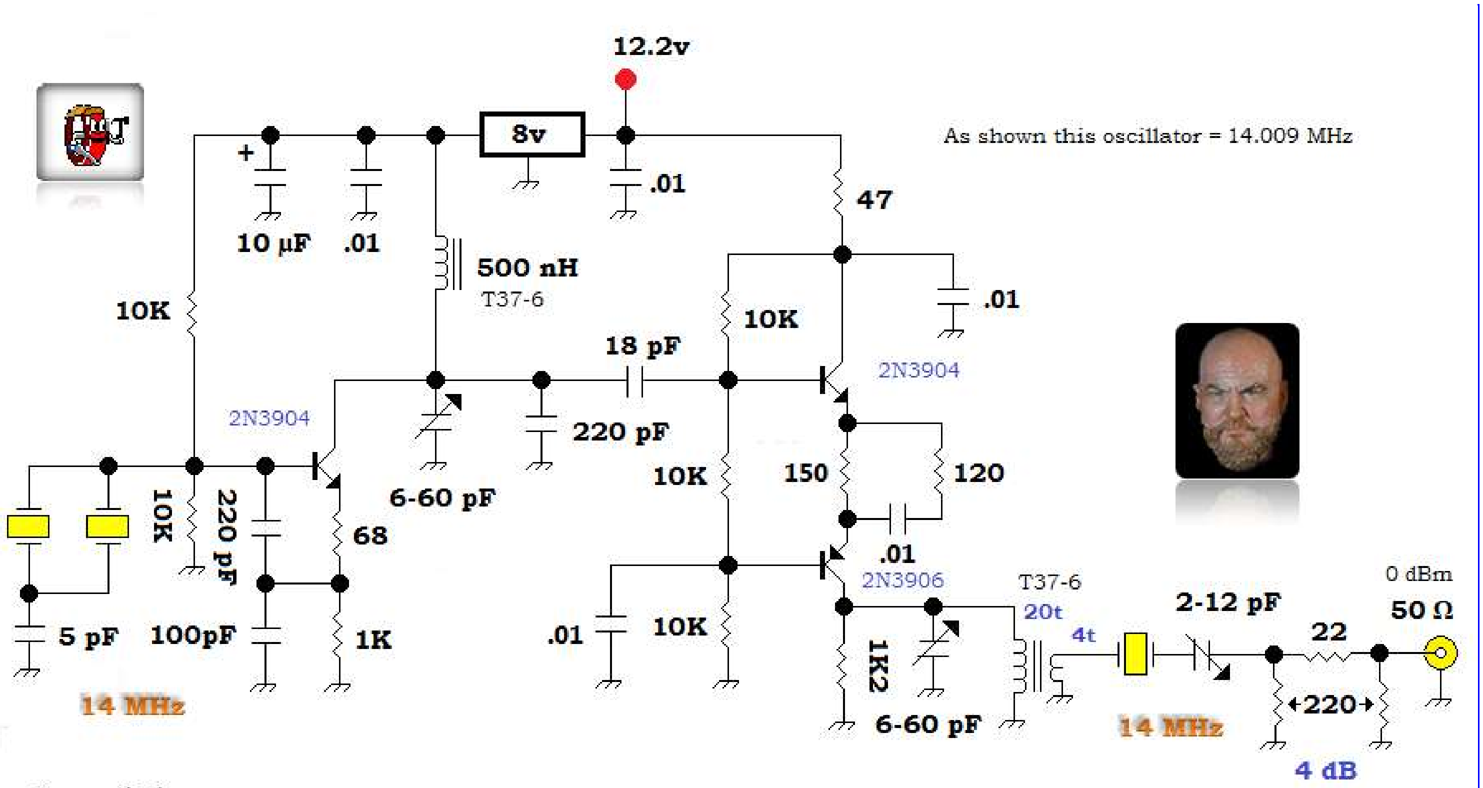
12.2v
8v
.01
47
As shown this oscillator = 14.009 MHz
10 µF
.01
500 nH
T37-6
10K
.01
18 pF
10K
2N3904
2N3904
220 pF
6-60 pF
10K
150
120
10K
220 pF
68
.01
2N3906
T37-6
20t
4t
2-12 pF
0 dBm
50 Ω
22
5 pF
100pF
1K
.01
10K
1K2
6-60 pF
220
14 MHz
14 MHz
4 dB

General Idea:

1. Use a super VXO with added L and/or C on 2 different oscillators to pull them 20 KHz apart
2. Apply an external xtal filter to lower phase noise

Above — 1 of my rudimentary experiments involving a super VXO plus output xtal filter. Not a great design. In this experiment, the 2 crystals were connected to ground via a 5 pF capacitor to yield 14.009 MHz. I learned that xtal frequency and xtal parameters determined the available frequency spread and output filter function.

After thought and the above bench work, and also because I need to make oscillators for IMD measurements at 50-100 MHz, I decided to go with an LC oscillator. They work fine for IMD measures with proper design and construction. For example, the signal generator shown in EMRFD Fig 7.27 sports a phase noise approaching -140 dBc/Hz @ a 10 kHz spacing; getting awfully close to the cherished vintage HP-8640B signal generator.

With LC oscillators, I learned that a metal enclosure with a feedthrough capacitor for the DC line proved important.Strong reverse isolation also ranks 'very important'. A case in point follows:

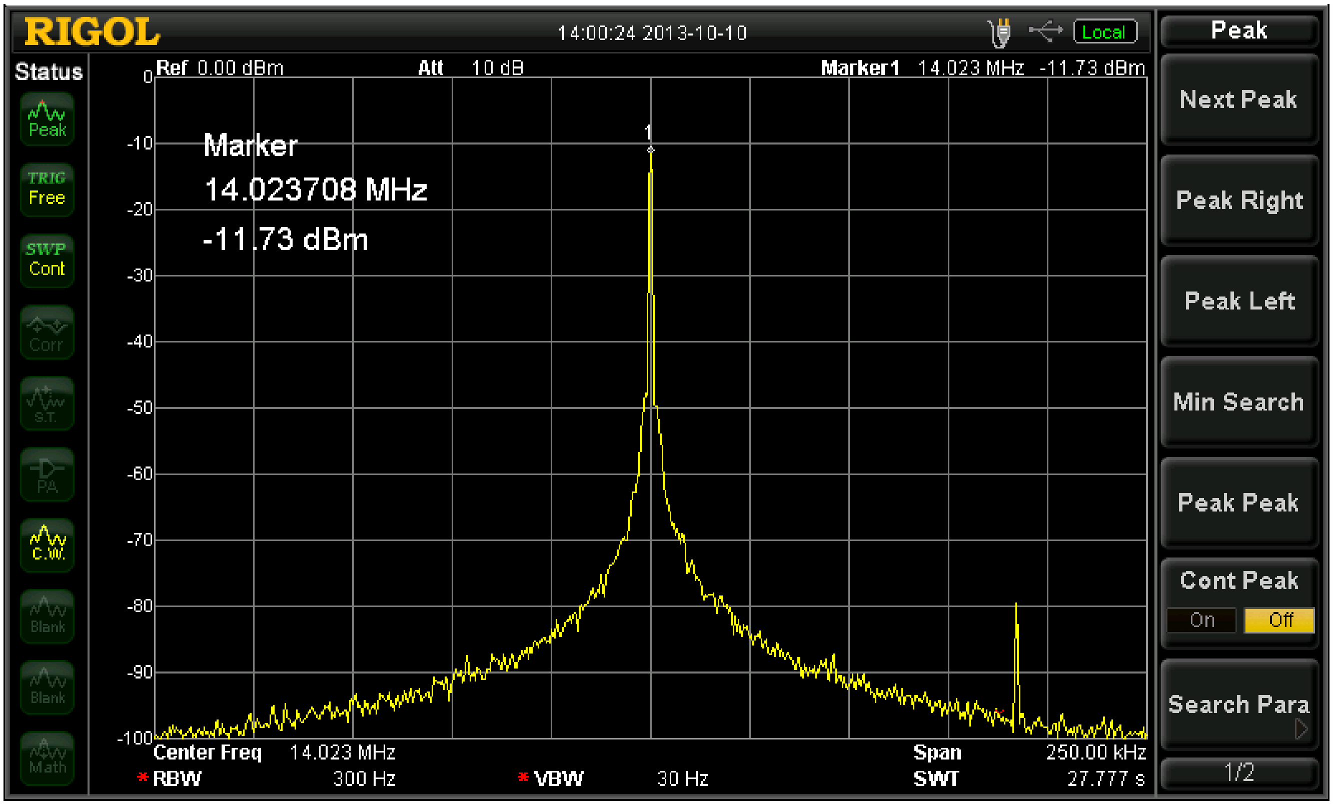
14:00:24 2013-10-10

Ref 0.00 dBm

Att 10 dB

Marker1 14.023 MHz -11.73 dBm

Marker
14.023708 MHz
-11.73 dBm

Center Freq 14.023 MHz

Span 250.00 kHz

RBW 300 Hz

VBW 30 Hz

SWT 27.777 s

RIGOL
Local
Status
Peak
TRIG Free
SWP Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
1
Peak
Next Peak
Peak Right
Peak Left
Min Search
Peak Peak
Cont Peak
On
Off
Search Para
1/2

Above — A close-in spectrum analysis of 1 of my 14.02 MHz LC oscillators. Span = 250 KHz; RBW = 300 Hz and my video bandwidth = 30 Hz to "cut the grass". You can see a spur about 92 KHz right of center; a signal from my SA or environment is getting back into my VCO and low-level modulating it. Yikes! Another RF-tight LC oscillator with a common-base isolation amp showed no such spur.

Close in spectrum analysis may allow you to measure phase noise with your spectrum analyzer, however the wide-range, first oscillator of a Rigol spectrum analyzer operates in the range of 2- 4 GHz and is varactor tuned; as such, the best phase noise we can measure with it = -88 dBc/Hz @ 10 KHz spacing.

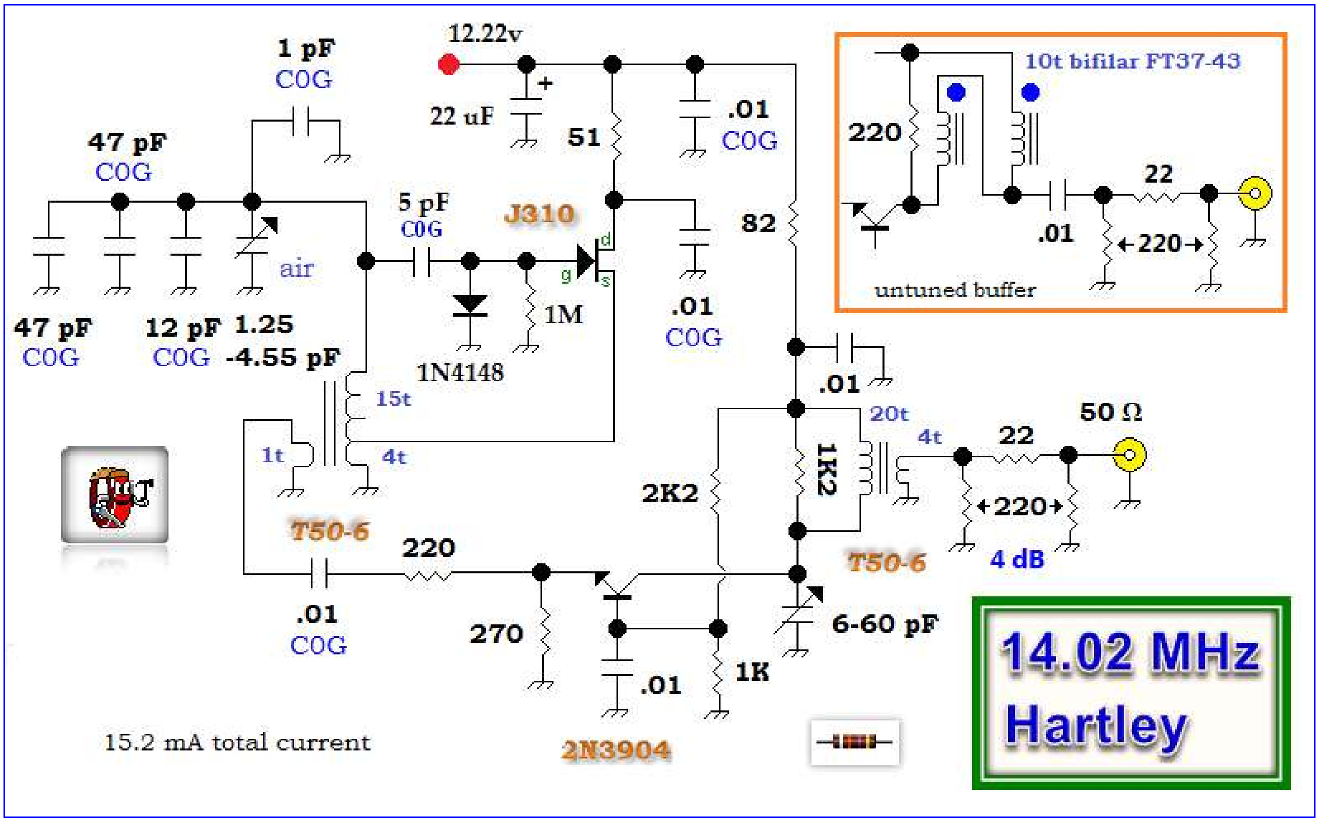
1 pF
C0G
12.22v
22 uF
51
.01
C0G
47 pF
C0G
5 pF
C0G
J310
82
air
d
g
s
1M
.01
C0G
47 pF
C0G
12 pF
C0G
1.25
-4.55 pF
1N4148
15t
1t
4t
T50-6
220
.01
C0G
270
.01
2N3904
2K2
1K
1K2
20t
4t
22
50 Ω
220
4 dB
T50-6
6-60 pF
.01
10t bifilar FT37-43
220
.01
22
220
untuned buffer
15.2 mA total current
14.02 MHz
Hartley

Above — 1 of my experiments: a Hartley with common-base buffer amp right out of EMRFD Figure 7.27 sans the final amp. I made 2 versions: 1 with a tuned buffer amp and another with a wideband output (see orange inset). The output power of the tuned amp and wideband amp = -6.7 dBm and -11.7 dBm respectively.

You may want to add a final amplifer — I've got a couple of amplifer modules if I need to get to 0 dBm or above. Prior to tweaking the lid-on frequency to 14.02 MHz, I measured the 2nd harmonic at 55 -dBC in the tuned output version— yay! Building with solid temperature drift techniques, the frequency drift in my counter went just a few hertz *up* over time. The Figure 7.27 oscillator remains a go-to circuit for us builders.

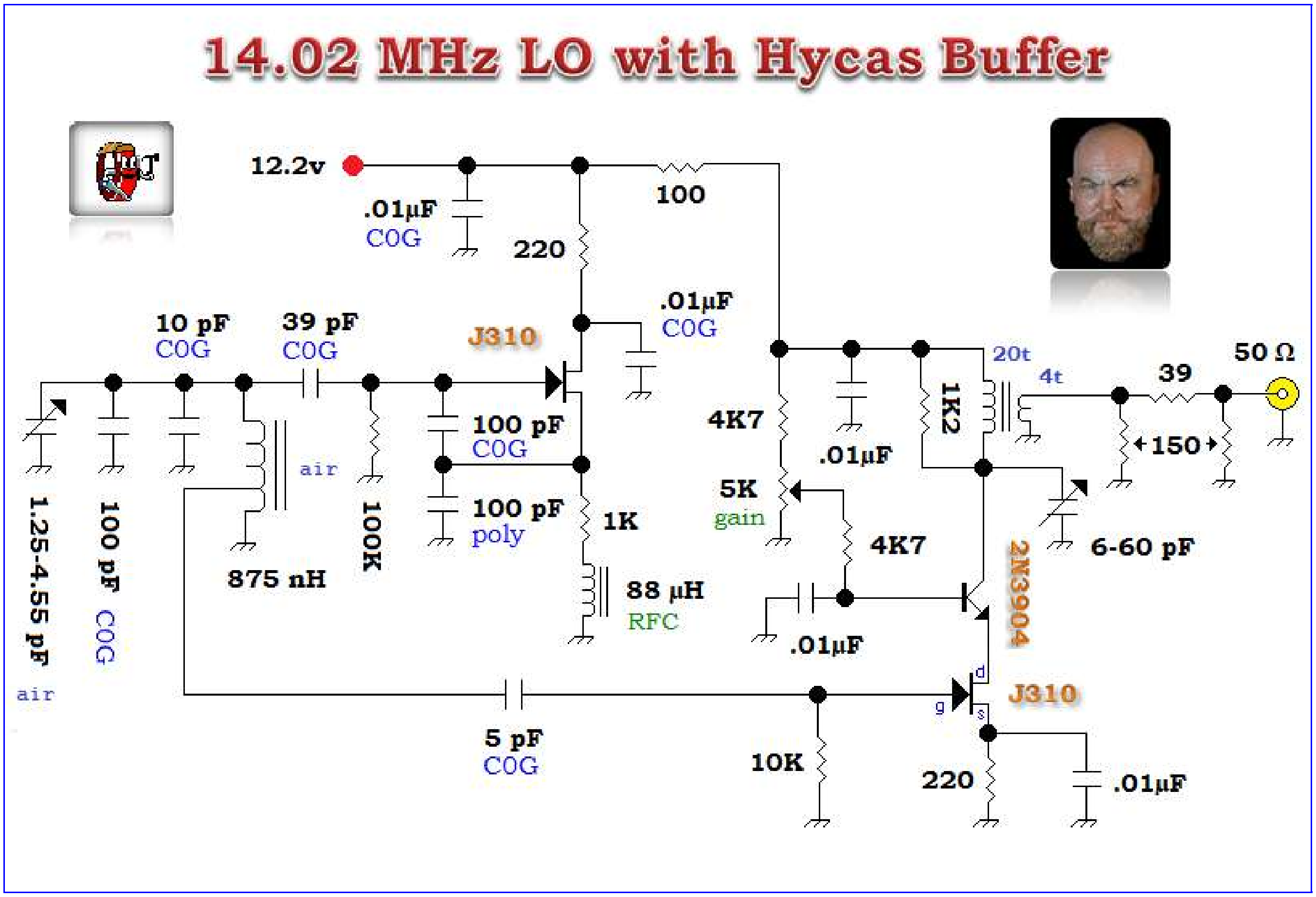
14.02 MHz LO with Hycas Buffer
12.2v
.01µF
C0G
220
100
.01µF
C0G
10 pF
C0G
39 pF
C0G
J310
1.25-4.55 pF
air
100 pF
C0G
air
875 nH
100K
100 pF
C0G
100 pF
poly
1K
88 µH
RFC
4K7
.01µF
1K2
20t
4t
39
50 Ω
150
5K
gain
4K7
6-60 pF
2N3904
.01µF
5 pF
C0G
d
g
s
J310
10K
220
.01µF

Above — A Colpitts with hycas buffer allowing variable power output. During measurement of IMD, the 2 oscillators need to exert the same amplitude, so only 1 needs to have variable output power. A wideband bench generator plus a 1 frequency oscillator like this might be all you need.

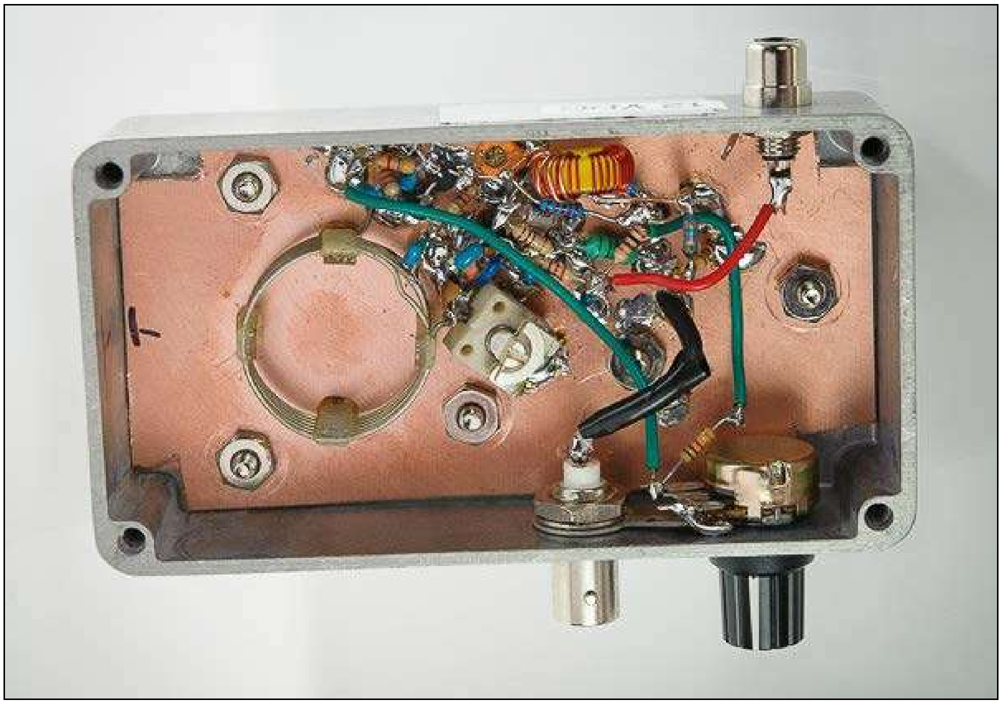

Above — A photo of the completed Colpitts LC oscillator. Running a phono jack for the DC input was a mistake. A feedthrough capacitor = the best choice. The epoxy glued air wound coil and air trimmer cap plus the regular temperature stability techniques and a 100 pF polystyrene capacitor yielded a temperature drift under 40 Hertz per hour. 5 bolts anchor the copper clad board.

I connected the 5 pF coupling cap to the lowest turn on the air coil. Mounting an air coil (solenoid) in smokestack fashion over a copper board probably upsets its Qu, but it worked okay.

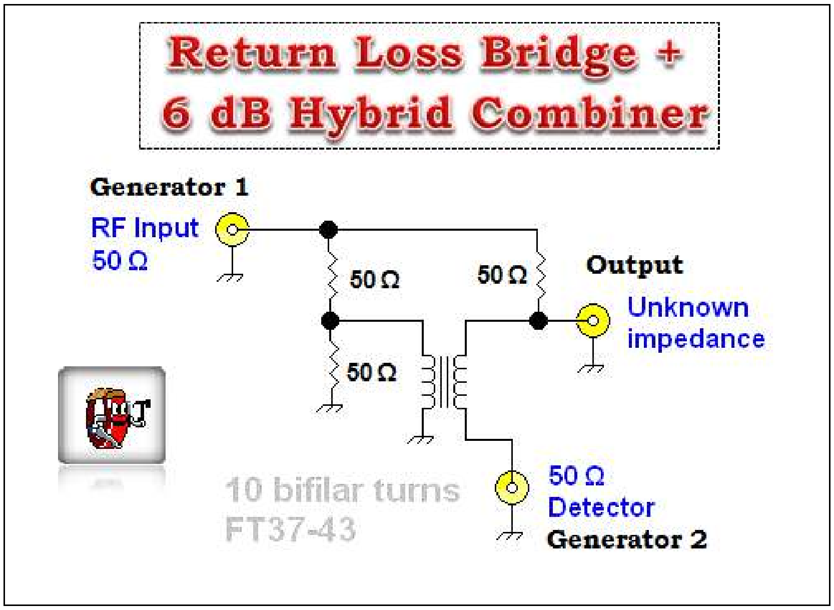

Above — Our old friend the Return Loss Bridge is also known as a 6 dB hybrid combiner and 1 serves double-duty on your QRP workbench as an isolator and a return loss bridge.

Phase Noise and Commercial Signal Generators

The worst of today's ham transceivers transmit a CW carrier with noise that is down by 117 dBc/Hz @ 10 kHz spacing. This is bad enough that if you had a Ham operating this rig 0.8 km away you would hear this noise, even with a perfect receiver. In a typical oscillator, the phase noise @ 100 kHz spacing lies ~20 dB lower than @ 10 kHz out to a noise floor that is around -140 dBm/Hz or better. The Elecraft K3 impresses us with phase noise at -138 dBc/Hz @ 10 KHz [Sherwood].

We either build, borrow or buy our signal generators. Second hand commercial products range from good old HP/Agilent gear or LC boxes once used by TV/FM receiver repairmen. The TV service equipment covers ~300 kHz to 150 MHz and tends to be inaccurate with a limited output amplitude range, while the vintage HP stuff often is heavy, decaying and crowds your bench.

Most of hobbyists can't afford nor justify spending $8-18k on boxes that go down to -127 dBm amplitude and -146 dBc phase noise at 1 GHz @ 20 kHz offset [Agilent N5181B MXG X-Series RF Analog Signal Generator]. These generators feature components like a YIG oscillator phase locked to an ovenized crystal reference. Professional, high-end gear.

Another choice = the Arbitrary/Function Generator or ARB that sell for a fraction of the cost of the aforementioned boxes. A few friends own them, in particular, products made by Rigol and GW Instek. These DDS generators range in frequency and price: higher price generally = better phase noise.
For example, the popular Rigol DG1022A phase noise = -108 dBc/Hz @ 10 KHz offset and the DG4162 = ≤-115 dBc/Hz @ 10 KHz offset; not good. Often ARBs will feature a host of modulation functions, sweeps, frequency counters and the 2 output channel models may offer I/Q outputs and the ability to perform 2-tone IMD tests. Search around the T Equipment website to read some equipment datasheets

My question to 2 informed EEs — Is the phase noise of a -115 dBc/Hz ARB *good enough* to reliably measure IMD? Their answer: yes, if you have a high end, modern spectrum analyzer with a minimum RBW of 100 Hz, deft analyzer skill and understand the measurement. Often us hobbyists have to make due with whatever test equipment we can build or buy.

You might just build a low noise oscillator with a crystal at 14.0 MHz and provide the other tone with your bench signal generator or some other combination of 2 devices. This is test equipment junkie fodder — band-of-brother's stuff. Thanks for reading.

2. Almost Popcorn Superhet

I published this rogue web page with hope you might find it interesting. It houses notes and photos for an incomplete update to my 1998 Popcorn 7 MHz, CW, superhet receiver.

Click for the Almost Superhet web page.

Click for the Almost Superhet Supplement page

3. VXO — VCXO Notes

VXOs and VCXOs provide a useful main or reference oscillator for HF and when multiplied, in VHF-UHF designs. Ignorant of such circuits, I executed some crude experiments to discover how they work and find ways to improve my designs. Increasingly clinical on my bench, I'm losing patience with the radio mystique — although the limits of my lab equipment, brain power and bench skills humble and tumble me into deeper understanding why educated people might submit to electronics folklore and go in blind faith.

I'm not sure my bench work accomplished much, but if I can add 1 factoid to the RF enthusiast VXO knowledge base, then, I'll feel happy. I'm officially adopting **System D**; the slang term that originated in the French-speaking African continent and Caribbean Plate. In Down and Out in Paris and London, George Orwell called these System D practitioners "débrouillards" — low-level kitchen help, who achieve culinary success in the worst of circumstances. Television hero MacGyver practiced System D.

System D in the context of the QRP Homebuilder describes an effective, practical and resourceful builder who skillfully adapts to any situation and *gets it done*. System D builders don't just kludge together crap — these science-informed débrouillards innovate and shine. Sounds hopeful doesn't it?

Nearly all my bench adventures begin with a web search and the key word *VCXO* yielded many espléndido files + presentations by people and companies who passionately share their knowlege and experiences. Thanks to all!

Experiments

It started with an email from Paol — what's the S12 of the source follower in his 14 MHz VXO buffer amplifier? I built 1 for 10 MHz and tested it since I didn't know.

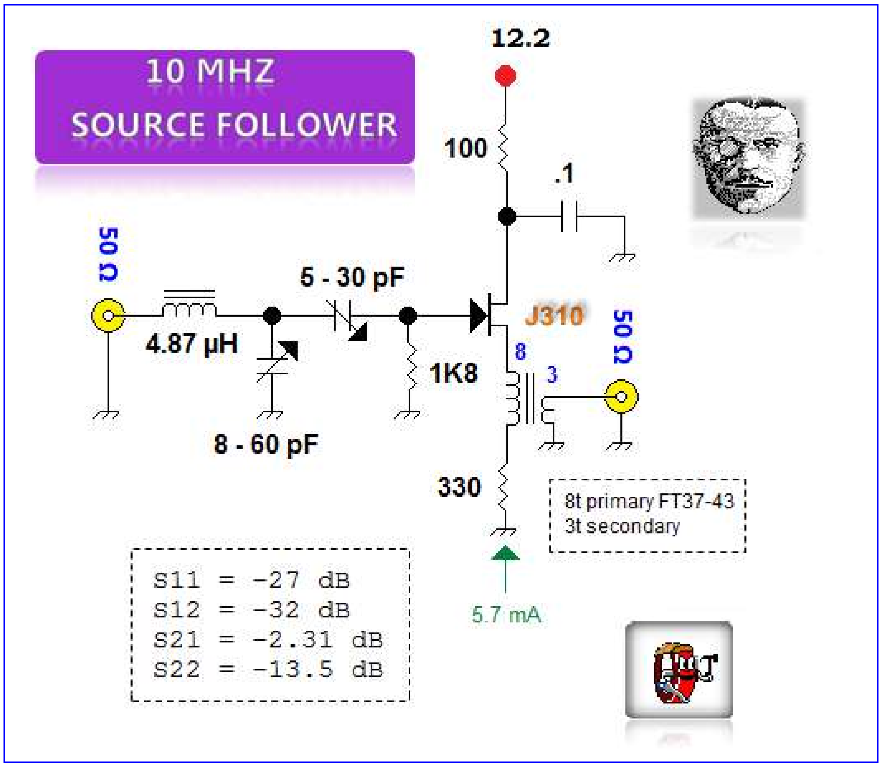

S11 = -27 dB
S12 = -32 dB
S21 = -2.31 dB
S22 = -13.5 dB

Above — Device Under Test. Paol copied a source follower he found on a web site (complete with an über kitsch 1 mH choke on the source). I tweaked the current and replaced that big, old MF choke with a tapped FT37-43 giving an output return loss of ~ 13.5 dB. Although mediocre, it's probably better than the match provided by the series resistor we typically stick between the JFET source and its 50 Ω load. Further, for my S12 measurement, I matched the input with a L-C-C network to get a strong return loss @ 27 dB. Measured with a spectrum analyzer, S12 = -32 dB; paltry reverse isolation. The matching networks incurred some losses but raised measurement fidelity.

-32 dB — not great, but okay if followed by another amp. For a diode ring mixer, I might couple it to a common base amp, or eliminate the source follower altogether and replace it with a hycas buffer amp. Paol drove a NE612 transmit mixer with his VXO + a source follower and it worked okay in his bare-bones TX. He AC-coupled a 51 Ω resistor to ground on pin 6 of the NE612 that well terminated the LP filter. Intrigued, I then helped him with his VXO.

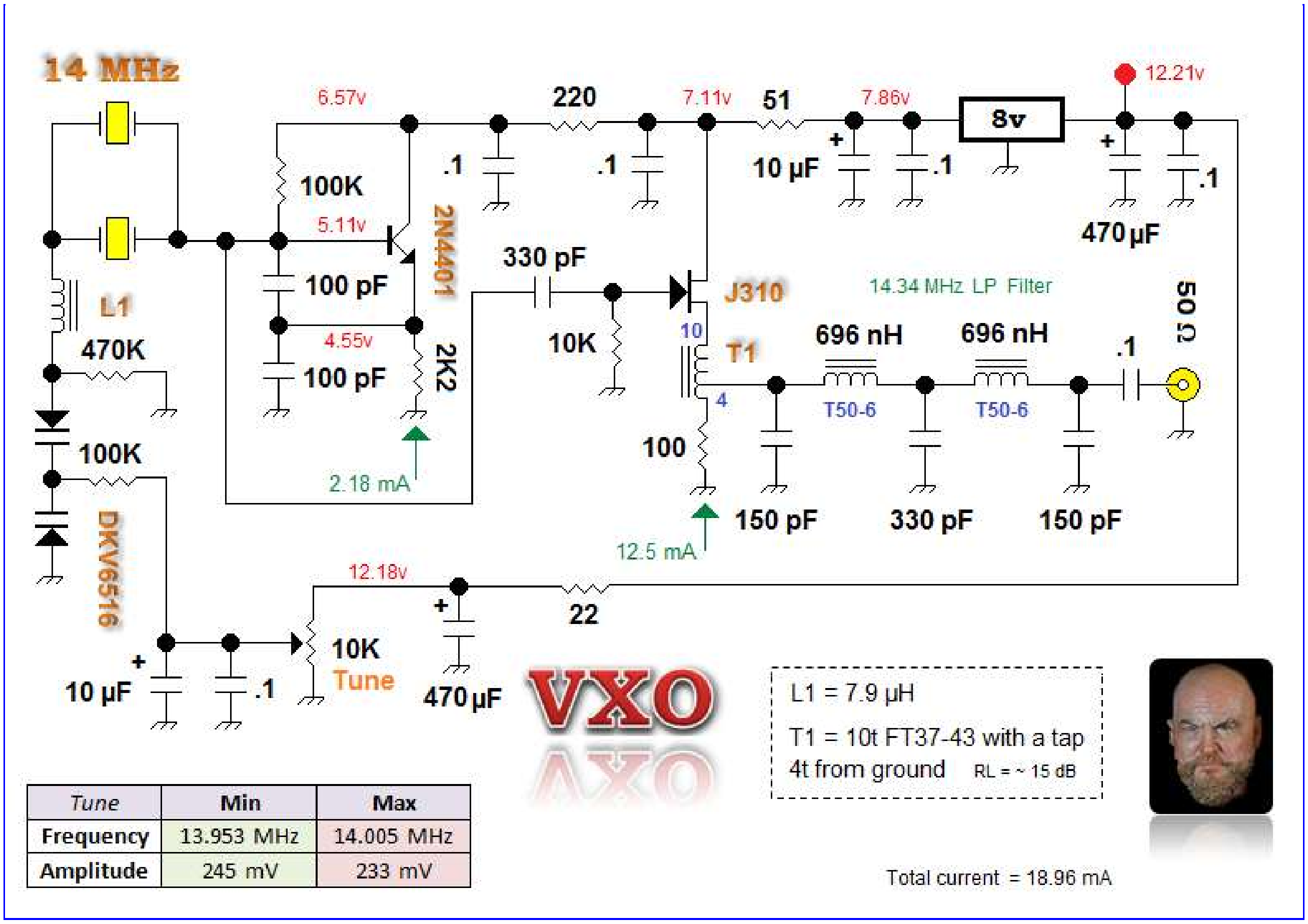
14 MHz
6.57v
220
7.11v
51
7.86v
8v
12.21v
100K
.1
.1
10 µF
.1
.1
5.11v
2N4401
330 pF
470 µF
100 pF
J310
14.34 MHz LP Filter
50 Ω
L1
4.55v
10
T1
696 nH
696 nH
.1
470K
2K2
10K
100 pF
T50-6
T50-6
4
100K
2.18 mA
100
DKV6516
12.5 mA
150 pF
330 pF
150 pF
12.18v
22
10K
Tune
10 µF
.1
470 µF
VXO

Tune	Min	Max
Frequency	13.953 MHz	14.005 MHz
Amplitude	245 mV	233 mV

L1 = 7.9 µH

T1 = 10t FT37-43 with a tap 4t from ground RL = ~ 15 dB

Total current = 18.96 mA

Above — My first VXO experiment with varactors instead of an air variable cap that technically creates a voltage-controlled crystal oscillator (VCXO). The tabled values were measured with L1 wound on a FT50-61 ferrite toroid.

Paol had varactors but no air-variable capacitors. Both of us owned some Q = 200, DKV6516 silicon varactor diodes that with a reverse DC voltage of 2 to 10 volts give a nice tuning range of 18 to 140 pF for a single diode. We cheated and ran the reverse DC voltage from 0 to 12.2 volts, although later increased the 22 Ω R to limit the peak tuning voltage to 10v. Paol also put a fixed resistor on the cold end of his 10K tuning pot to keep a minimum 2 VDC on the varactors: I didn't bother, since my "this is only a test" breadboard will be dismantled and junked.

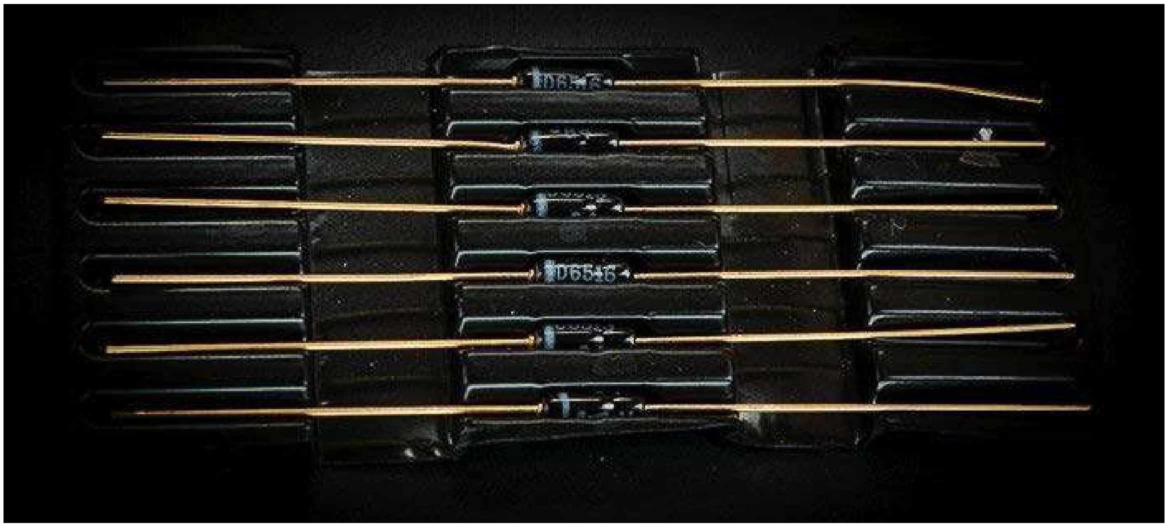
D6516
D6516

Above — DKV6516

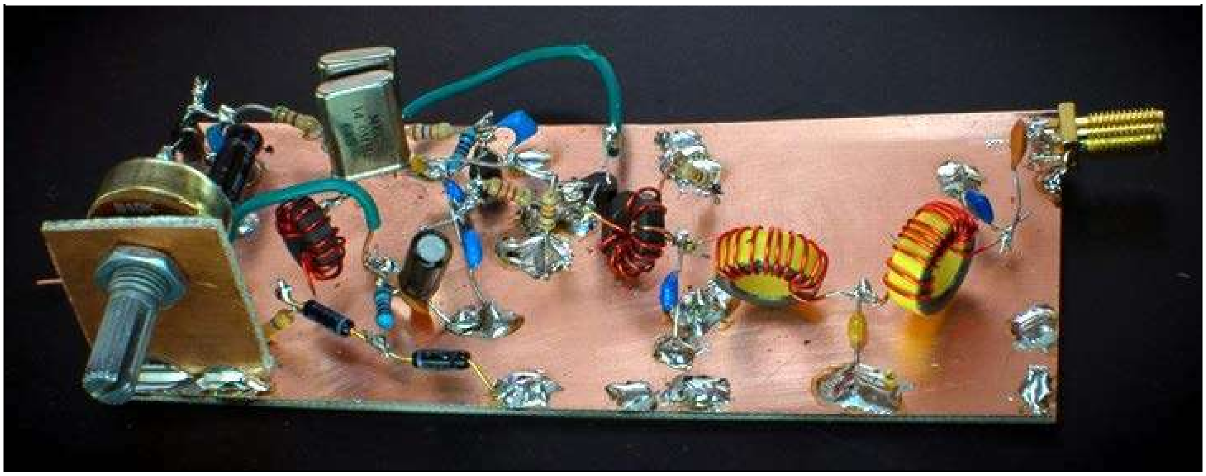

Above — VCXO breadboard. Paol and I each made a VCXO and he with 14.060 MHz crystals. Lacking these, I just used 2 junk box 14 MHz crystals with a measured Qu of ~80K. The tapped FT37-43 gave an output return loss of ~ 15 dB thus providing a reasonable match into a simple 50 Ω low-pass filter. Click for a sweep showing the 2nd harmonic suppression.

My initial build didn't contain the two noise filtering 470 µF caps on the 12 volt supply: I originally soldered a solitary 22 µF cap cap close to the 8 volt voltage regulator input. My 12.2 volts comes from a regulated DC supply. More on filtering noise later.

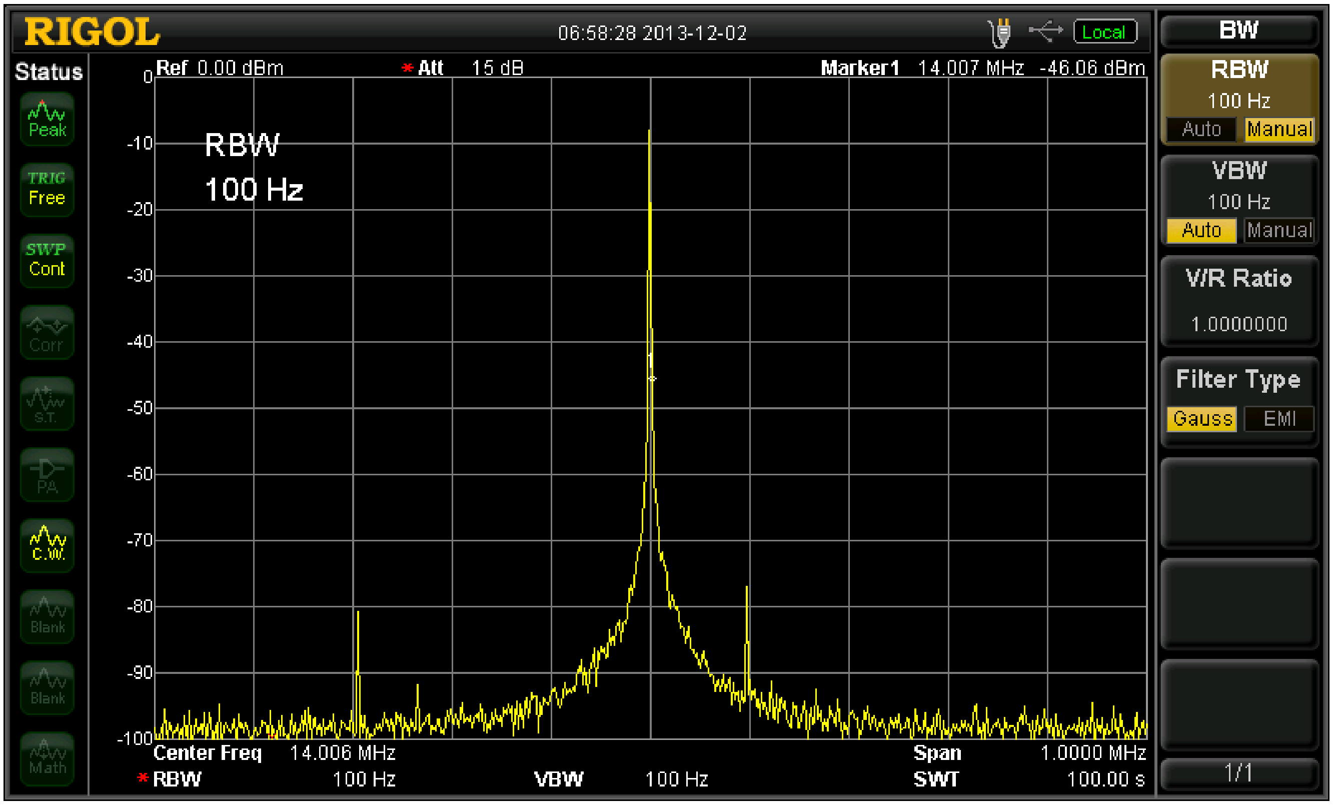

06:58:28 2013-12-02

Ref 0.00 dBm

Att 15 dB

Marker1 14.007 MHz -46.06 dBm

Center Freq 14.006 MHz

Span 1.0000 MHz

RBW 100 Hz

VBW 100 Hz

SWT 100.00 s

RBW 100 Hz Auto Manual

VBW 100 Hz Auto Manual

V/R Ratio 1.0000000

Filter Type Gauss EMI

Above — My VCXO spectrogram showed 2 big spurs: 300 KHz down and 100 KHz up from center frequency (CF). I exchanged the varactors with an elegant, Q >= 1000 air variable, plus changed to a higher Q inductor, but could not get rid of these spikes. I assumed these spurs came from the crystals, or the BJT. Then, too, might they arise from my spectrum analyzer?

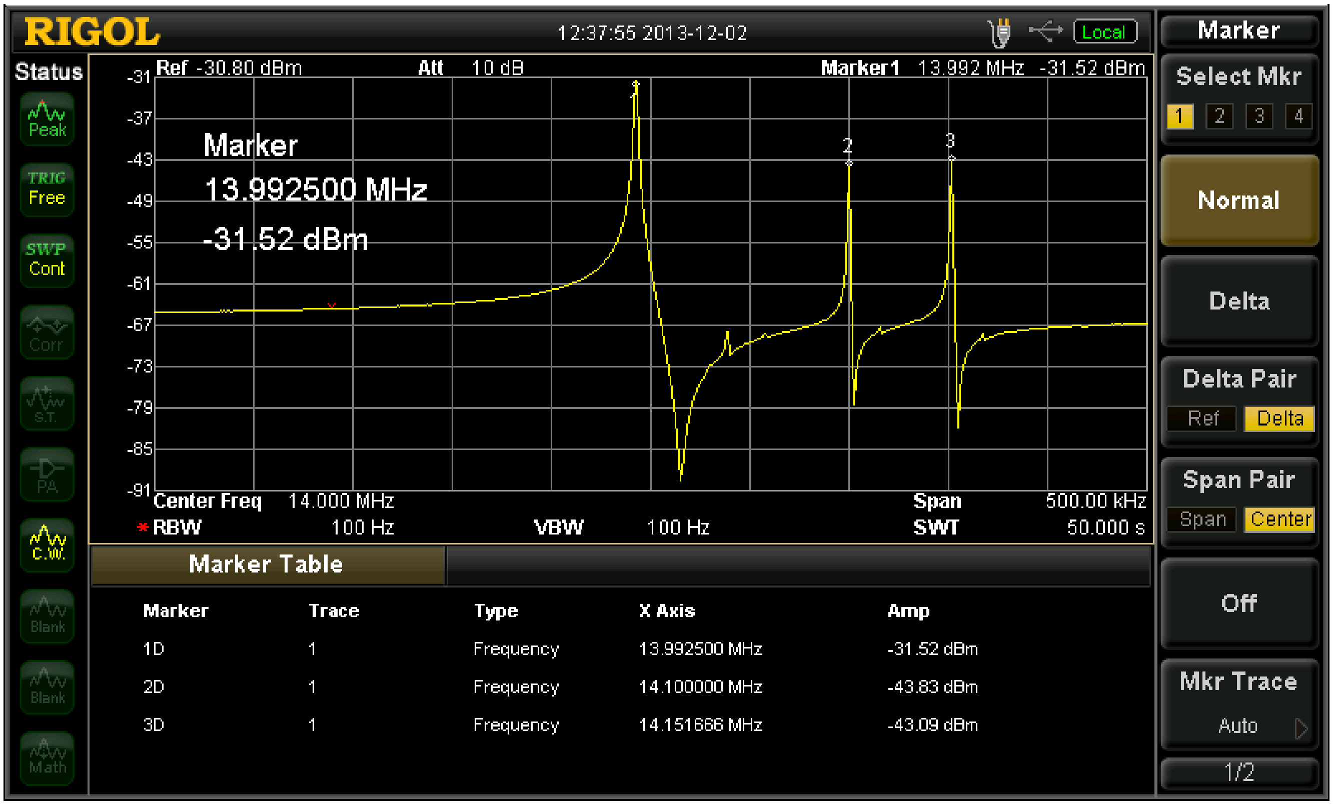
RIGOL
12:37:55 2013-12-02
Local
Status
Peak
TRIG Free
SWP Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
-31
-37
-43
-49
-55
-61
-67
-73
-79
-85
-91
2
3
Marker
Select Mkr
1 2 3 4
Normal
Delta
Delta Pair
Ref Delta
Span Pair
Span Center
Off
Mkr Trace
Auto
1/2

Ref -30.80 dBm Att 10 dB Marker1 13.992 MHz -31.52 dBm

Marker
13.992500 MHz
-31.52 dBm

Center Freq 14.000 MHz Span 500.00 kHz
RBW 100 Hz VBW 100 Hz SWT 50.000 s

Marker Table

Marker	Trace	Type	X Axis	Amp
1D	1	Frequency	13.992500 MHz	-31.52 dBm
2D	1	Frequency	14.100000 MHz	-43.83 dBm
3D	1	Frequency	14.151666 MHz	-43.09 dBm

Above — 14 MHz crystal sweep spectogram. To find out why these spurs occurred, I swept 1 crystal at a time from my batch and saw spurs on all of them. However, after some reading, I discovered that the mere presence of a crystal spurious response doesn't signal causality since this is almost a normal happening. Wes, W7ZOI explains this on this web page. Crystals usually have a spurious response and from Wes, the spur is more often than not about 80 or 100 kHz above the main signal. This happens with fairly good crystals, less so with crappy, cheap crystals and even less often with good quality crystals. Since many of us buy computer-grade crystals from Mouser-Key, or online auction sites, who really knows what we're getting? 1 day I hope to test a Q = 1 million crystal — that will be fun.

However, renograde crystals could still be the cause. Sometimes when pulling a crystal(s), we're living on the edge — the main resonance is pulled so far that the circuit balks at the resonant frequency, or lies near the edge of such behavior and spurs start to show up. *Back to System D*. Simple substitution with 2 other crystals from another batch could tell me if these particular crystals were behaving poorly. It turns out, they were.

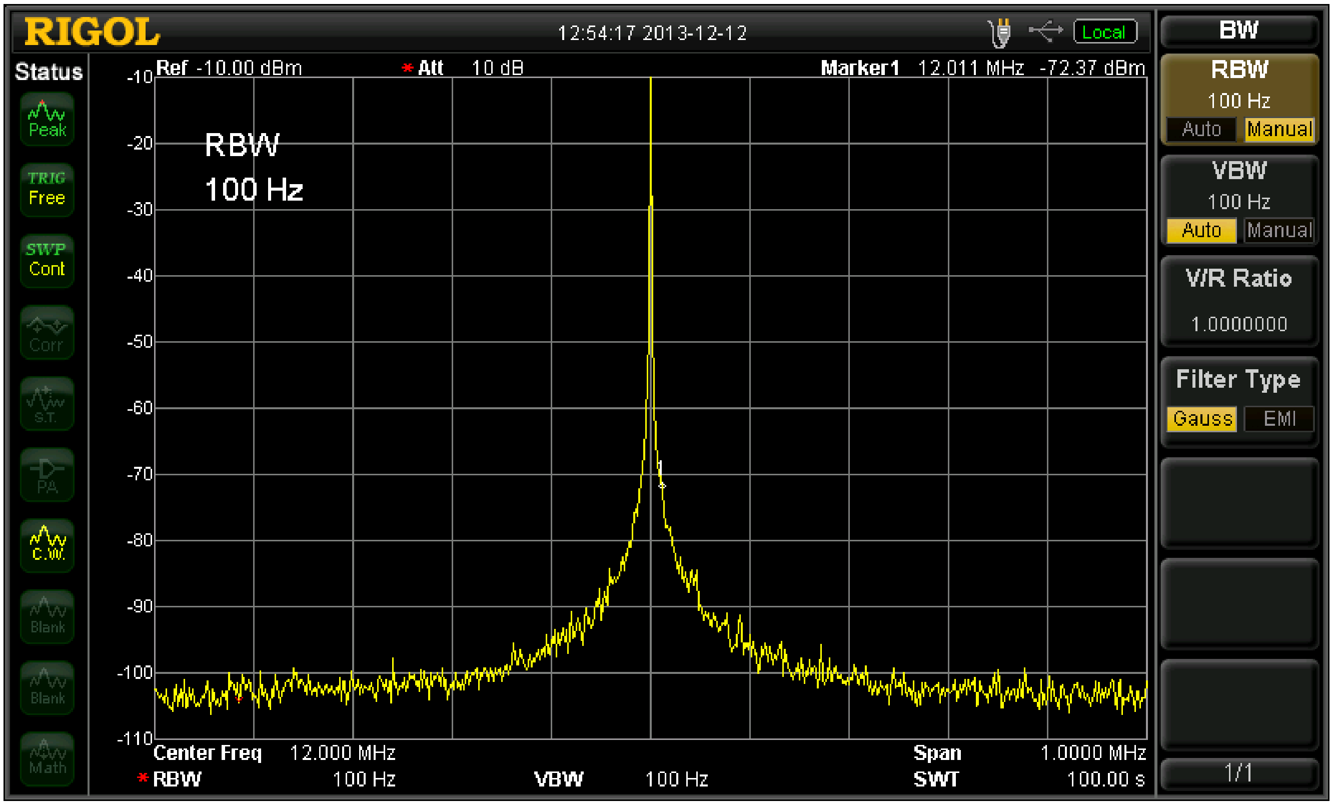
RIGOL
12:54:17 2013-12-12
Local
Status
Peak
TRIG Free
SWP Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
BW
VBW
V/R Ratio
Filter Type
1/1

Ref -10.00 dBm Att 10 dB Marker1 12.011 MHz -72.37 dBm

RBW
100 Hz

RBW 100 Hz Auto [✓ Manual]
VBW 100 Hz [✓ Auto] Manual
V/R Ratio 1.0000000
Filter Type [✓ Gauss] EMI

Center Freq 12.000 MHz Span 1.0000 MHz
RBW 100 Hz VBW 100 Hz SWT 100.00 s

Above — Spectogram of the VCXO with 12 MHz crystals. I wanted to try other 14 MHz crystals, but didn't own any. My nearest value crystals were 12 MHz jobs with a Qu of 120K. Eh voila! No spurs — yay! Let's talk about noise:

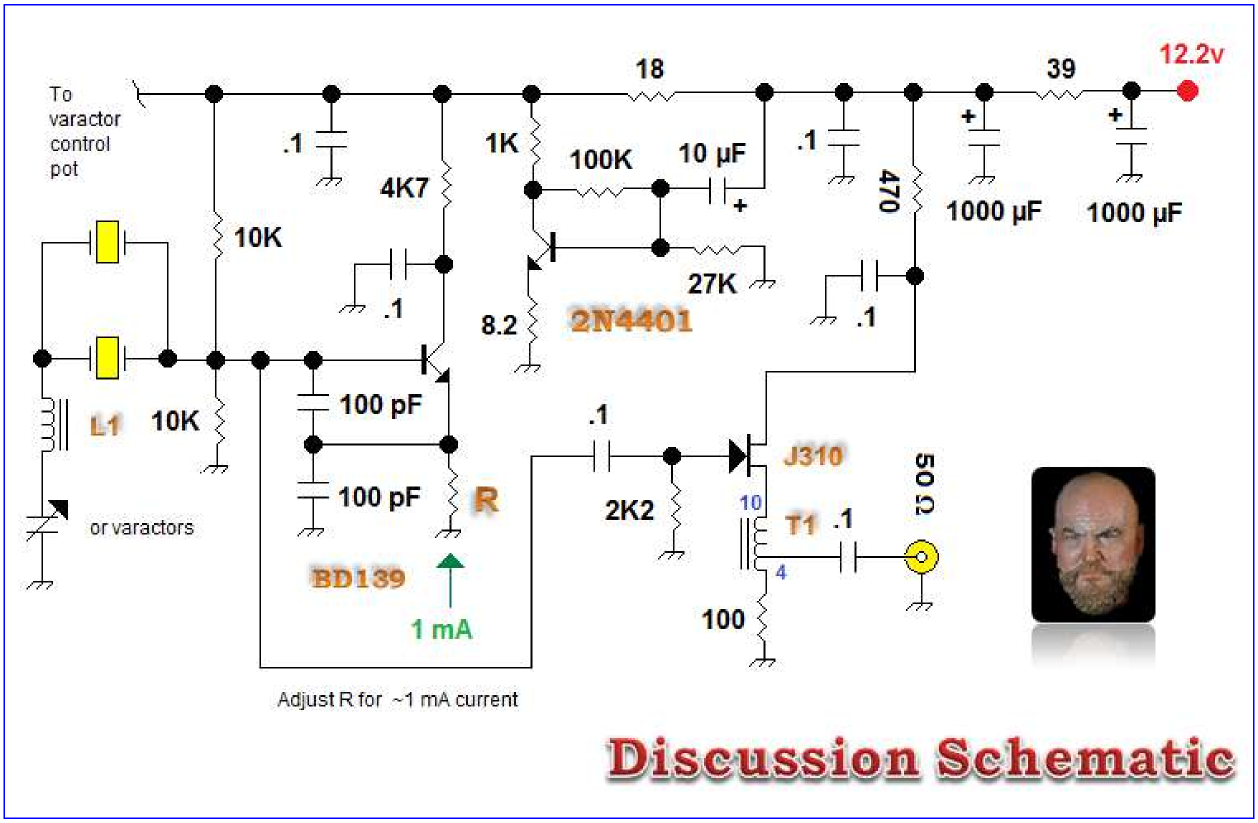

Above — Schematic for sideband noise discussion

System D design poses a real challenge with repect to noise. Measuring noise proves an advanced technique beyond the scope of this popcorn RF web site. We can't measure noise with our spectrum analyzer since the oscillator in the analyzer usually exhibits more noise than the VXO we want to test.

On page 4.12 of EMRFD version 1 is a small photo of 2 spectograms that show a clean crystal trace contrasted with a really bad oscillator trace. Baffled, because I could not replicate this experiment, I emailed Wes and asked him how he did it. Wes wrote we're able to see the noise in those traces because his SA featured a multiple pole, wideband filter with steep skirts that allows us to see the baseline noise close-in to the carrier. In contrast, most modern SAs feature synchronously tuned resonator type, or digital filters that produce a narrow peak with skirts that spread way out. The noise gets swept under the skirts.

By all means, advanced builders may apply notch filter techniques to measure oscillator sideband noise down low, however, a consideration for others might be to apply techniques proven to reduce noise and spurs — and just build the best oscillator possible plus measure what you can.

Some basic techniques to make a low-noise VCXO are well described elsewhere and may include: a high Q resonator, high Q varactors [with low series resistance], low power supply noise [including varactor tuning voltage], a low flicker noise amplifer, avoid overdriving the crystal and try to swing the highest AC voltage in the tank without reaching saturation and/or causing reverse breakdown in the varactor(s). Even the ratio of the Colpitts capacitors can affect phase noise, but without a method to measure phase noise, this gets esoteric. Some of these techniques also apply to VXOs.

Power Supply Noise

Bob, K3NHI and others have measured significant noise from the LM317 or related voltage regulators that we amateurs typically use. Power supply noise can FM our resonator and boost oscillator sideband noise. We may filter power supply noise with *large* value capacitors plus/minus series decoupling. See the work of Leif, SM5BSZ on this web page under the heading "The +12V supply is carefully decoupled like this". In addition, Charles Wenzel, published this web page.

I employed both techniques in my discussion schematic above. I went with the pi filter shown above 1000 µF - 39R - 1000 µF at the voltage input. A inductor might be used instead of the R if the voltage drop is too high for you. The Wenzel circuit fascinates me. The series resistor (18 Ω in my version) samples the current that flows into the oscillator. A voltage is then developed that is applied to the amplifier to generate a voltage that is out of phase with the noise current.

With good DC noise filtering I could measure small improvements in sideband noise in my SA despite only able to measure down to -88 dBc/Hertz at 10 KHz offset. In contrast, you rarely ever see even 1 electrolytic capacitor in the VXO schematics featured on many Ham radio builder sites.

Another trick is to avoid a regulator altogether. Since my DC supply already has a voltage regulator set to 12.2v, placing a second 5 to 8 volt regulator after it might just add another noise source.

Overdriving the Crystal

I heed the advice of Konstantinos, SV3ORA. He found that gently driving his crystal generated less noise and spurs compared to overdriving the crystal. Thus I adjusted my emitter resistor to about 1 mA current. For the schematic above, a typical emitter resistor starting value = 1K8. Just measure your emitter current.

I employed a large area, slower BJT, the BD139 to hopefully drop flicker noise. Expensive, super high Q crystals (Q = 500K to 1 million) may be driven with up to several mA of current, however, it's likely not so for the computer-grade xtals we typically use. I can't measure the outcome of these interventions, but System D thinking might just prevail.

Varactors

Varactors tend to be low Q and their Q deteriorates as we go up in frequency. I've seen very low phase noise VCXOs employing varactors, however, these were engineering marvels. Still, as a VHFer and hopeful UHF enthusiast, this is the direction I wish to go. We ideally need to keep some reverse DC voltage on the varactors at all times and not go to 0 — this also limits tuning range in our simple 12v supply radios. Read your datasheet. Air variable capacitors (a VXO) might be a better solution for your application.

How Does a VXO work?

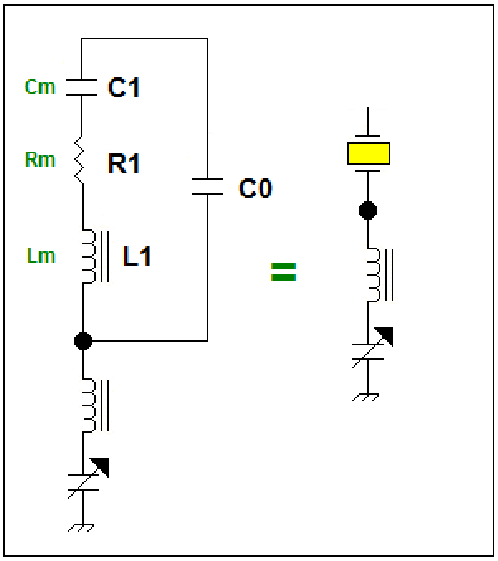
Cm
C1
Rm
R1
C0
Lm
L1
=

Above — The crystal equivalent electrical circuit has 2 arms: The impedance of the motional arm Zm lies in parallel with Zo, the impedance resulting from the parallel capacitance C0 (formed by the 2 crystal leads.) We measure and/or calculate these parameters when designing our crystal filters. In the above schematic I added the external L and C used to move the crystal’s series resonant frequency around in our VXO, VCXO, or Super VXO circuits.

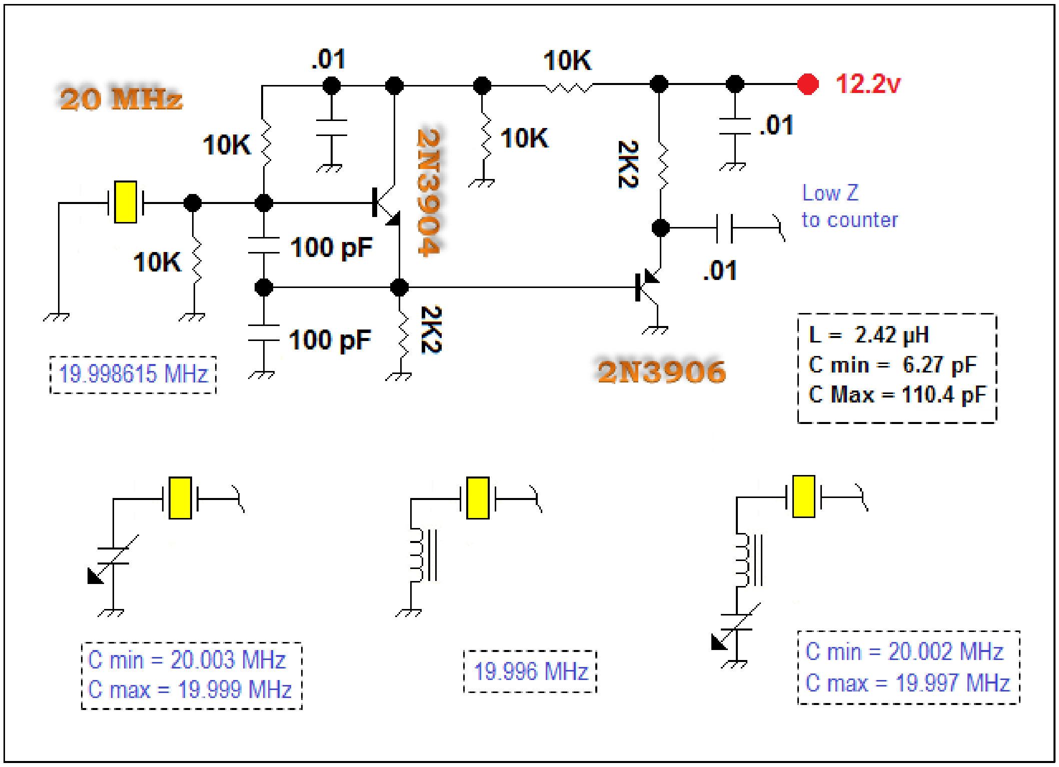
20 MHz
.01
10K
12.2v
10K
10K
2N3904
2K2
.01
Low Z
to counter
10K
100 pF
.01
100 pF
2K2
19.998615 MHz
2N3906
L = 2.42 μH
C min = 6.27 pF
C Max = 110.4 pF
C min = 20.003 MHz
C max = 19.999 MHz
19.996 MHz
C min = 20.002 MHz
C max = 19.997 MHz

Above — I made some simple experiments to explain why we place both a L and C in series with the crystal and how it changes the crystal frequency in our oscillator. From these experiments, some theory from EMRFD and wisdom from Ken Kuhn, I made the following simple conclusions:

Electrically a crystal looks like a large inductance in series with a small capacitance and in parallel with C0. An external inductor adds to the crystal's series inductance and *lowers* the resonant frequency.

The external variable capacitor adds in series with the crystal internal capacitance to lower the net capacitance and *raise* the resonant frequency.

The purpose of adding an external inductor is so that we can tune both lower and higher than the natural crystal frequency. In contrast, if we only apply an inductor, the oscillator frequency would always be lower and if we just apply a capacitor, the oscillator frequency would always be higher.

Since the external inductance has much lower Q than the effective internal inductance, it adds to noise and signal losses. The same goes for the external C.

As we raise the inductance ever higher to boost the delta F, at some point, the parallel C of the crystal resonates with this "too high" L and the series resonant frequency has nothing to do with the series resonance of the crystal. This kills the normally sublime frequency stability of a crystal-based oscillator.

Q of the VXO/VCXO Inductor

I experimented to learn that the Q of that L doesn't need to be that high to allow the frequency to be moved up or down by an air variable cap or varactor.

For example, if I wound a coil on a T68-2 and shunted it with a resistor no less than 1K8 — it worked fine (sometimes the frequency shift would even improve 2-3 KHz for a given L). A #61 ferrite toroid wound inductor also worked well and requires much less turns allowing thicker wire for mechanical stability. Also, the lower number of turns helps avoid significant SRFs. Temperature stability remains a question with the this and all ferrites.

An FT37-43 or FT50-43 ferrite inductor worked poorly – I measured high signal losses plus the maximum frequency swing dropped when compared to the same inductance value wound on a T-68-2, or FT-50-61. Avoid #43 ferrites in any resonator circuit unless you're purposely building a noisy oscillator for learning purposes.

Some guys use those little molded chokes with a Q of ~ 50 for the L and they also work well. Predictably, as I lowered the Q of the external L, or went from an air variable C to a varactor, signal losses increased resulting in lower signal output amplitude.

Decreasing resonator Q increases phase noise. For a popcorn radio, I'm not sure the change in phase noise would be a deal breaker in some homebrew projects, but as possible, I think it's wise to employ high Q resonator components: crystal, inductor and the variable capacitance. Thus I'll wind my VXO/VCXO inductors on powdered iron toroids hereafter.

Since a quartz crystal resonator is driven by the piezoelectric effect, fundamental plus overtone frequency modes (3/5/ 7/11) may be derived. I unbypassed the collector and placed a collector choke on it to examine the output with my 'scope. The waveform rung with many harmonics. Lots to learn.

Go débrouillards!

---

4. Funster Line — A Simple 40M Band QRP Trans-receiver

Click for the Funster transmittter notes.

---

5. Miscellaneous Photographs and Figures

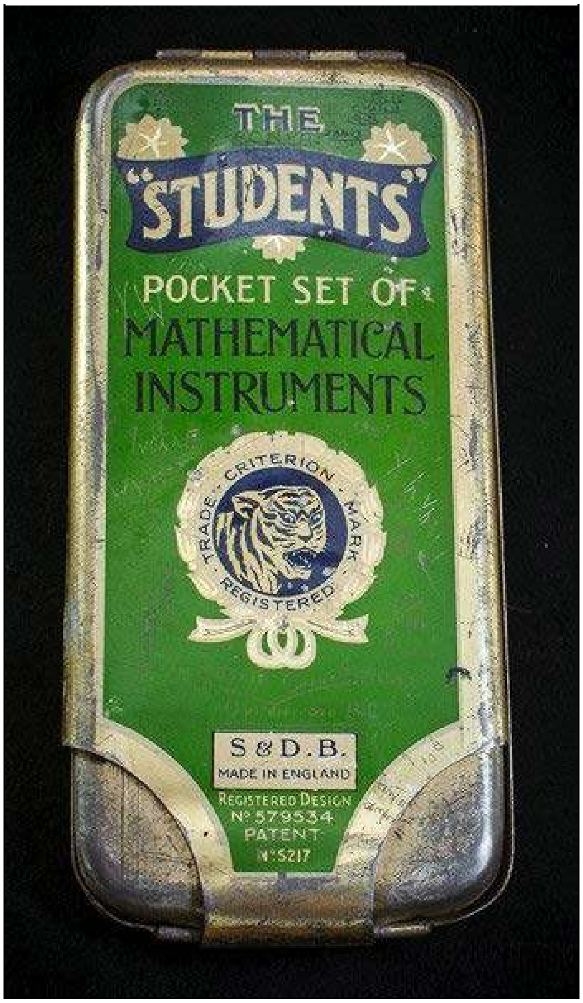
THE
"STUDENTS"
POCKET SET OF
MATHEMATICAL
INSTRUMENTS
CRITERION
TRADE MARK
REGISTERED
S&D.B.
MADE IN ENGLAND
REGISTERED DESIGN
Nº 579534
PATENT
Nº 5217

Above — My handed-down geometry set — purchased in 1938. Click for larger picture.

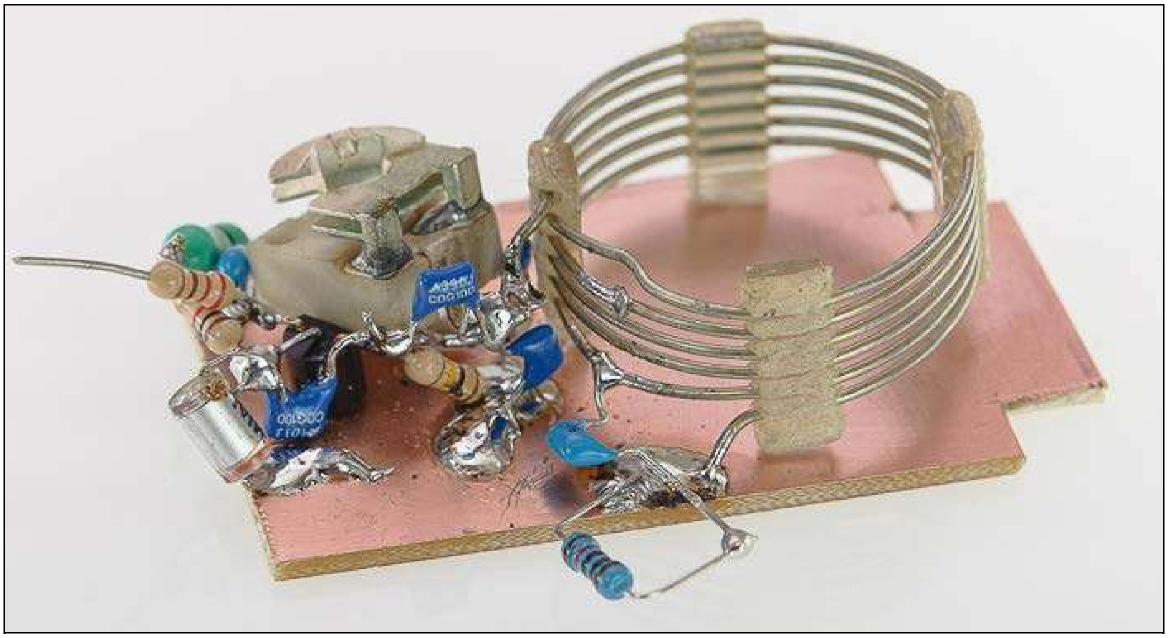

Above — My prototype Colpitts LC oscillator with only a 10K resistor as the load.

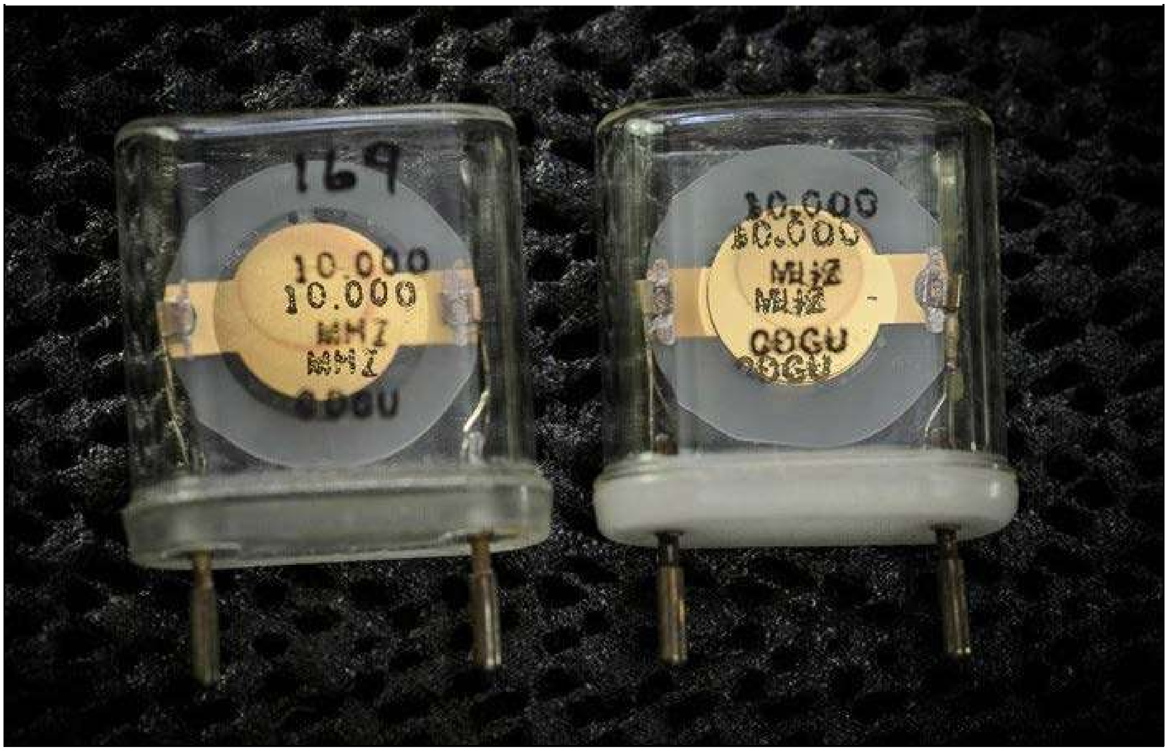
169
10.000
MHZ
COGU
10.000
MHZ
COGU

Above — Hi Q glass encased crystals. Click for another photo.

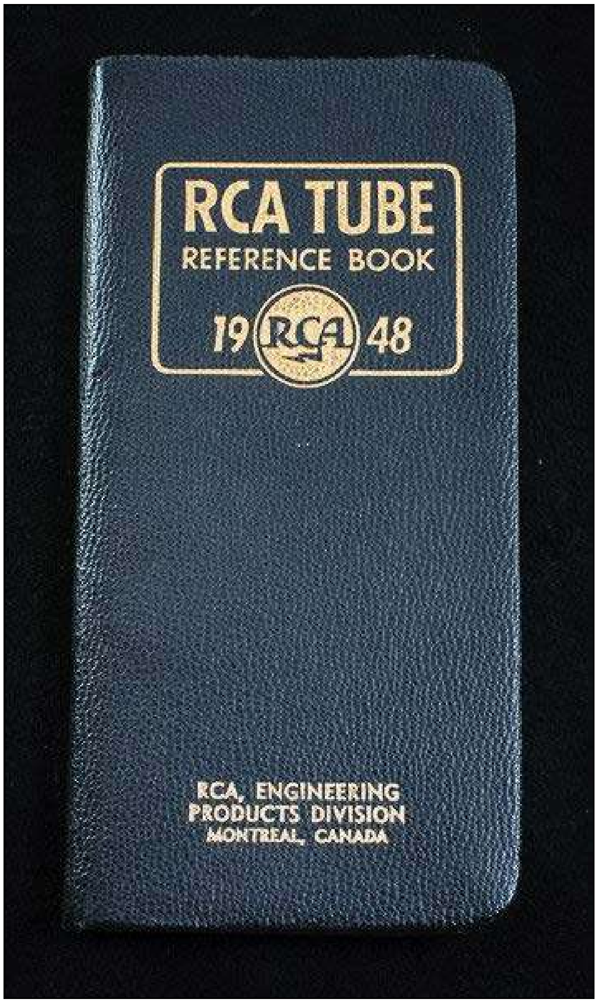
RCA TUBE
REFERENCE BOOK
19 RCA 48
RCA, ENGINEERING
PRODUCTS DIVISION
MONTREAL, CANADA

April 2014: New player on the scene — Bar Bar Kitty — chief science officer and QC consultant.

April 2014: Other new pet: Munchkin. Clear-spoken and direct — pet me, or feed me... and oh yah, Q2 is oscillating.

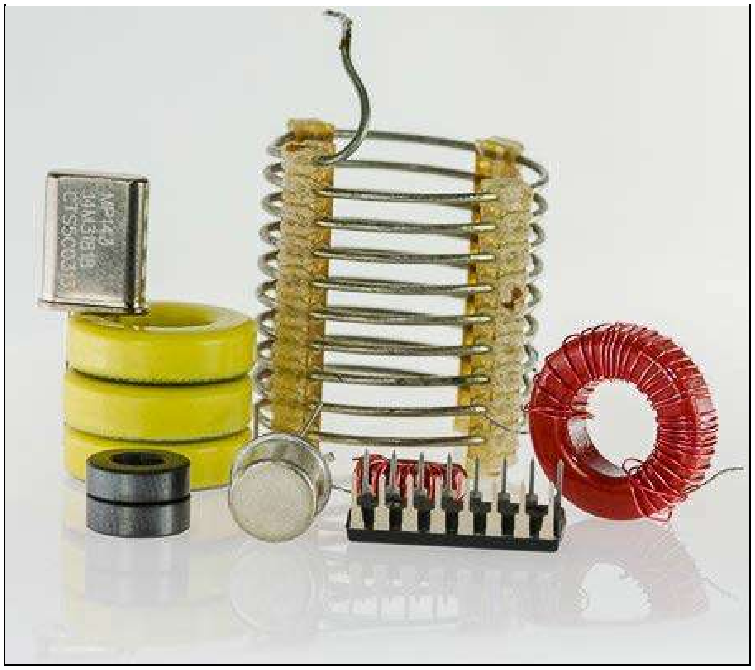

RF — Test and Measurement

Caitlyn 310 — UHF Beginnings

Respository web page for beginner UHF experiments in 2014-2015.

1. References
2. Transistors, RFCs, Capacitors and a 50 Ω Detector
3. Return Loss Bridge Experiments
4. VCO Experiments
5. Frequency Counter Kit

---

1. References

After a modest quest, I found an excellent microwave primer called Microwaves Made Simple. Principles and Applications by W. Stephen Cheung and Fred H. Levien. Published in 1985, the authors brilliantly teach the basics in plain language (no small task). My copy was printed in December 2013 by Artech House who print single copies of many classic texts on demand. Click for a photo of my book.

The ARRL published 2 UHF compendium books entitled: The ARRL UHF / Microwave Experimenters Manual [1990] and The ARRL UHF / Microwave Projects Manual [1994]. While out of print, used copies seem prevalent and might serve as good resources for builders who hunger for circuits and Ham-related UHF circuit building information. Click and click for photos.

Instrument companies provide continuing education and materials to educate & support their customers. With luck you'll find helpful free application notes and other tools. Two recent RF library additions include an Agilent CD containing all their current application notes and webcasts [my image is truncated and the CD includes applying SA's, 'scopes and signal generators], and this old Spectrum Analyzer Fundamentals booklet by Tek.

---

2. Transistors, RFCs, Capacitors and a 50 Ω Detector

I want to grasp circuits from 300 to 1296 MHz. This meant acquiring some UHF parts, making some bench tools and buying a 50 Ω detector to sweep and/or measure power in my UHF stages. Click for a snapshot of some of my parts cabinets.

BJTs

I'll employ the BFR93A, 2SC3356, and 2SC3583 as my mainstay NPN bipolars; at least to start with. Numerous transistors will work — price, availability, NF and shipping costs informed my choice. I'd also feel happy with MRF901, BFR92 and BFR96 NPNs and it's always fun reading datasheets. My PNP choice = BFT93 by NXP. Later on, I'll add in discrete GaASFET and other non-BJT species. Only 2 MMICs lie in the parts cabinet: MAR-3 and MGA-61563.

Radio Frequency Chokes (RFC)

Although, I'll also wind, or perhaps ply carefully dimensioned copper board traces to make RF chokes, I built up a stock of lower Q chip inductors in size 0805 and 0603. These include 8.2 nH, 33 nH, 68 nH, 220 nH, 330 nH, 470 nH and a few other values. For RFCs we need to consider L, tolerance, SRF, Q , maximum current and temperature stability — it's probably better to only buy clearly specified parts so you can study their datasheet to glean this information. I've swept nearly all my RFCs and recorded their SRF.

Choosing RFCs — SRF (series resonant frequency)

An RF choke exerts its greatest impedance at its SRF and in the case of a first-order RFC such as a single inductor, we should aim the SRF at or near the frequency we wish to attenuate the most. Consider a choke in the emitter lead of a BJT Colpitts oscillator with a frequency range from 375 to 400 MHz: The choke's ideal SRF should lie somewhere above 400 MHz. As possible, the SRF of a first-order RFC should exceed the operating frequency of the device it's choking.

In higher-order filter applications such as a C-L or C-L-C (pi filter) applications, engineers tend to target the choke's SRF to >= 10 times the operating frequency. Challenges such as limited inductor choices, needing to match stages, or coping with stray reactances will arise, however, as amateur UHFers, we just try our best to fit an RFC.

UHF Capacitors

I swept numerous size 0805 + some 0603 and 1206 capacitors in a microstrip line jig to determine their SRF. Math formulas enable us to calculate and plot attentuation, Q and ESR at self-resonance if we wish. Click for 1 example. Although we rarely apply capacitors at their SRF in practical circuits, considering Q, ESR and attenuation gets us thinking that capacitor SRF is a dynamic value that may vary widely with part size, capacitance and datasheet specified Q.

For instance, if you take a size 1206 capacitor, measure its SRF and then replace it with a same-value size 0805 capacitor, the SRF of these two caps might vary by as much as 200 MHz or more. Size does matter! Even the same value capactor by a different manufacturer may be a SRF game changer. Bob, K3NHI tells it correctly: to measure *is* to know.

Datasheet specified capacitor Q is usually measured at a low frequency like 1 MHz. Since Q decreases as frequency increases, a capacitor that specs a Q of 2500 at 1 MHz may only sport a Q of 87 at its SRF way up at 512 MHz. During sweeps, I measured a deeper attenuation notch and often a higher SRF for "very high Q" caps when compared to the common, low-medium Q, popcorn caps sold as mystery parts by online stores or auctions. I'm now hypervigilent about finding very high Q SMD capacitors for sale at a low price.

I wrote the SRF of all my SMD caps in my notebook and use this data to inform my capacitor choice for RF bypass and tuning. Click for a photo of 3 UHF parts drawers. I've pretty much gone to size 0805 caps to try and garner a higher SRF at VHF and UHF. Still, too, we might have to solder down 0603 parts up above 1 GHz. Yikes!

Did you notice that when you drop a 0805 part on the floor, you never ever find it again? Even a little cough can expunge a size 0603 part from your QRP workbench.

50 Ω Detector

From 0.5 - 200 MHz, my 50 Ω terminated input DSO forms the heart of my test bench. An analog, or digital storage oscillocope with a bandwidth >= 1 GHz, plus a suitable wideband probe costs dearly and for many, won't be an option. UHF bandwidth, thermistor-based 50 Ω power meters are also expensive and lack signal viewing.

Thus, a spectrum analyzer seems the best choice for the amateur UHF enthusiast. A spectrum analyzer with a built-in tracking generator adds serious value — giving us a MF-UHF sweep system with profound utility. One popular, modern TG + SA choice is the Rigol DSA-815. I'm fortunate to have access to gear that measures up to 3 GHz.

3. Return Loss Bridge Experiments (RLB)

I'd be lost without a return loss bridge at UHF. Directivity marks the touchstone of a return loss bridge — and up at UHF grabbing a minimum 20 dB directivity may pose quite a challenge. Experiments to find a bridge with good directivity from 250 to 1750 MHz follow:

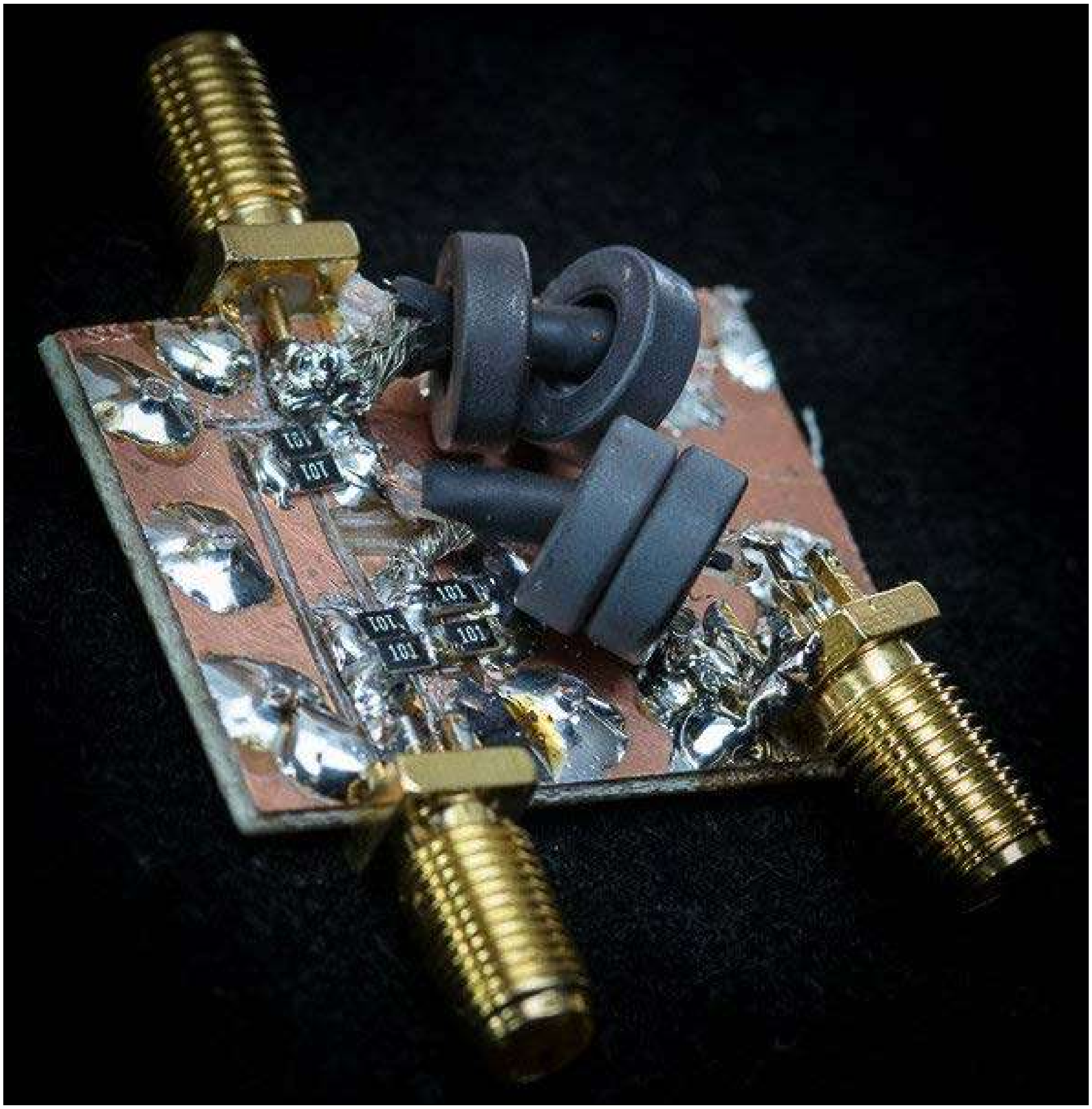

Above — A macro photo of RLB#1: 1 of the 3 "string of pearls" type return loss bridge built this Winter. I tried different numbers, sizes and mixes of ferrites. This particular bridge had 3 and then 2 FT37-61 toroids as shown, although I did not shorten the RG-174 coax when reducing to 2 toroids. The 61 mix gave poor directivity and was abandoned.

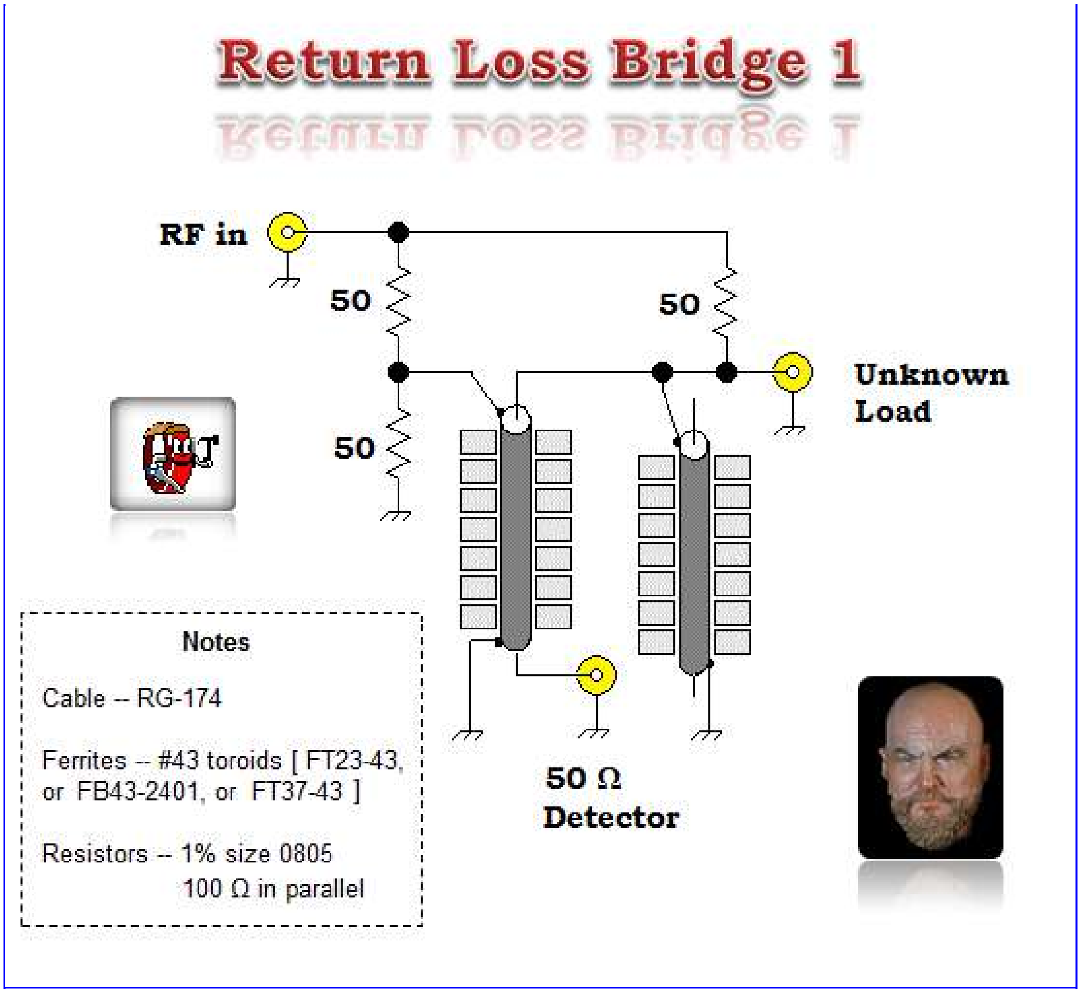
Return Loss Bridge 1
RF in
50
50
50
Unknown Load
50 Ω Detector
Notes
Cable -- RG-174
Ferrites -- #43 toroids [FT23-43, or FB43-2401, or FT37-43]
Resistors -- 1% size 0805
100 Ω in parallel

Above — Return loss bridge #1: the "string of pearls" variant. Several web sites feature great "string of pearls" RLB experiments for you to review and many place enough ferrite to extend from HF to UHF. I swept each of my builds in a Rigol tracking generator + spectrum analyzer with 4 frequency markers set. I made an open-circuit sweep [nothing on the unknown impedance port] followed by a sweep with a Mini-Circuits Labs DC - 18 GHz, 50 Ω load threaded on the unknown Z port. The difference between the open-circuit and 50 Ω measurement = the directivity at that frequency.

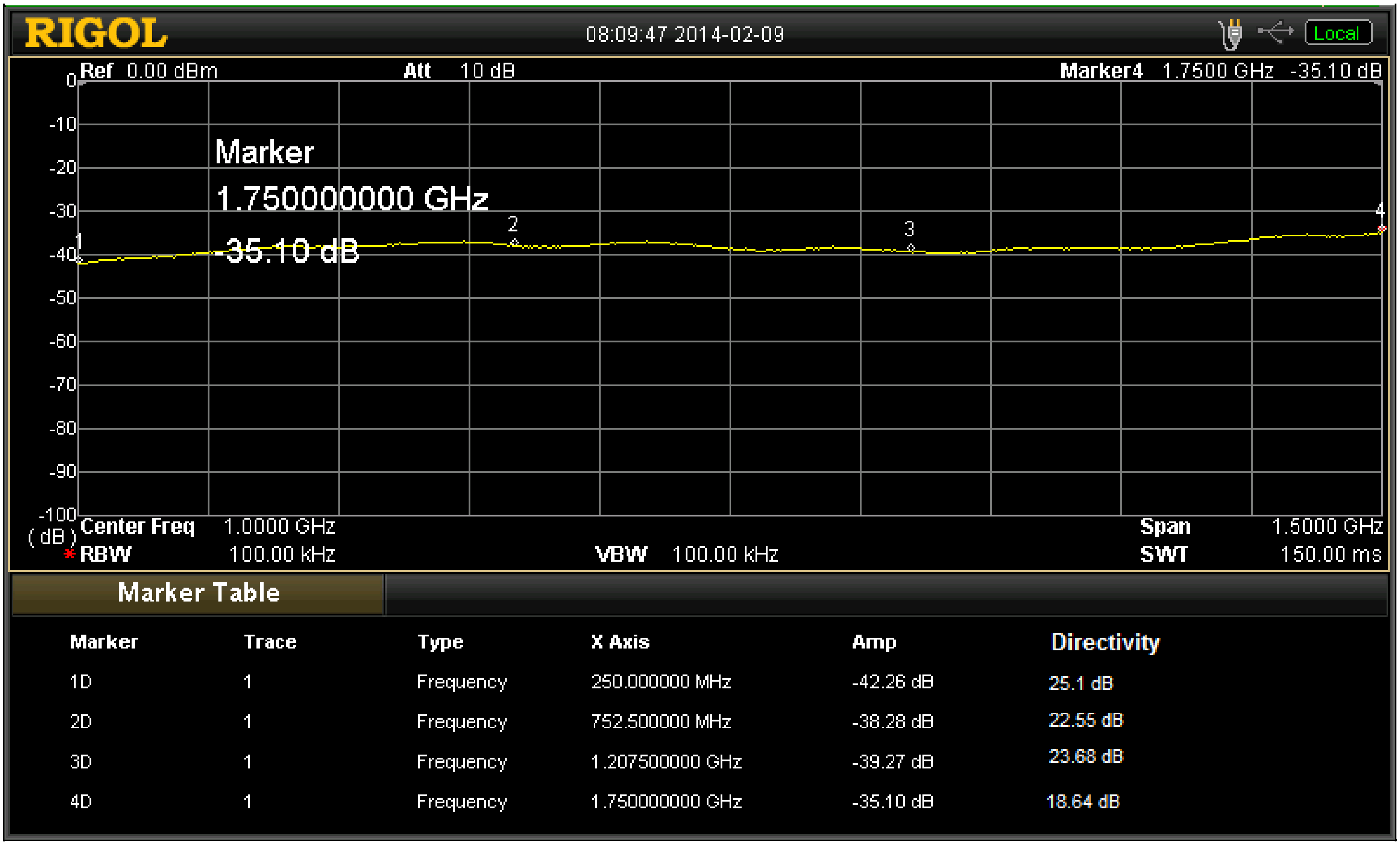
RIGOL
08:09:47 2014-02-09
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
1
2
3
4

Ref 0.00 dBm

Att 10 dB

Marker4 1.7500 GHz -35.10 dB

Marker
1.750000000 GHz
-35.10 dB

Center Freq 1.0000 GHz

Span 1.5000 GHz

RBW 100.00 kHz

VBW 100.00 kHz

SWT 150.00 ms

Marker Table

Marker	Trace	Type	X Axis	Amp	Directivity
1D	1	Frequency	250.000000 MHz	-42.26 dB	25.1 dB
2D	1	Frequency	752.500000 MHz	-38.28 dB	22.55 dB
3D	1	Frequency	1.207500000 GHz	-39.27 dB	23.68 dB
4D	1	Frequency	1.750000000 GHz	-35.10 dB	18.64 dB

Above — The sweep of my best "string of pearls" bridge from 250 MHz to 1.75 GHz showing the directivity at 4 frequency points. Click for the open circuit sweep. I felt encouraged by this sweep with five FB43-2401 beads over each piece of coax cut just long enough to fit the beads. I built on 2-sided FR4 copper clad board with at least 6 via wires connecting the 2 copper surfaces.

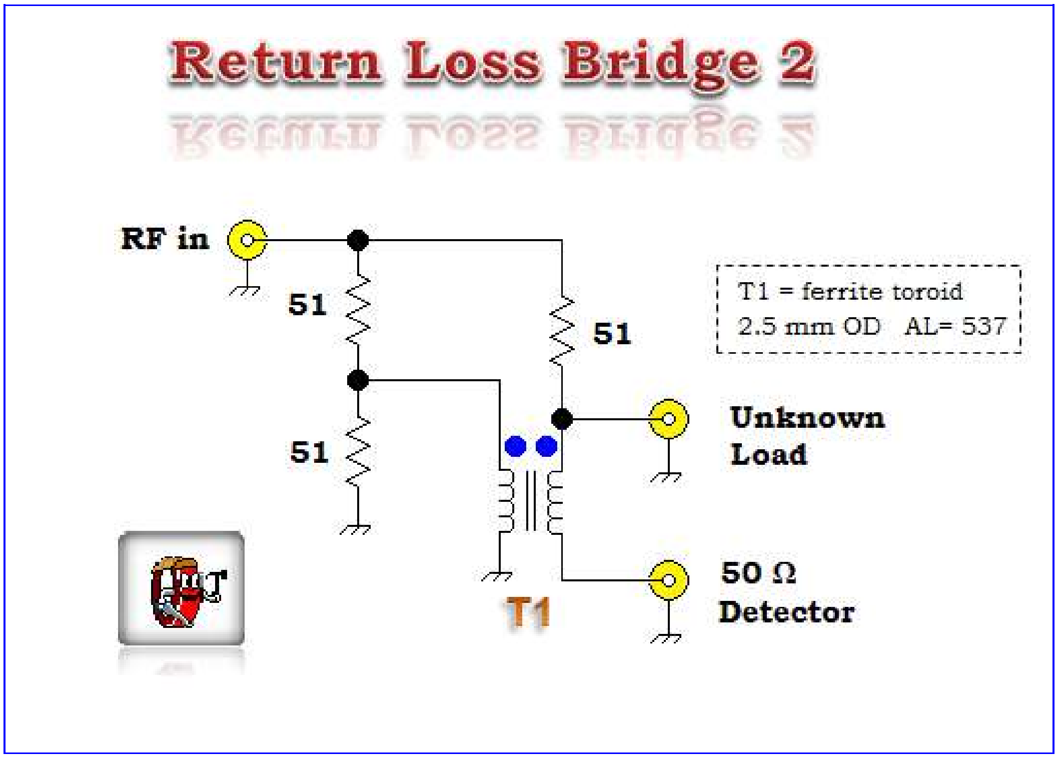
Return Loss Bridge 2
RF in
51
51
51
T1 = ferrite toroid
2.5 mm OD AL= 537
Unknown
Load
50 Ω
Detector
T1

Above — Schematic of RLB #2: the bridge topology featured in EMRFD. T1 = 5 bifilar windings of 32 gauge wire wound on a *tiny* ferrite toroid boasting a strong AL. I pulled this ferrite from a blown up mixer removed from an old 1.5 GHz spectrum analyzer. I started with 5 toroids, but during winding, got spastic and dropped 2 of them on the floor and concluded that the SMD part gnome who lurks beneath my bench swallowed them up, since I couldn't find them after an intensive search.

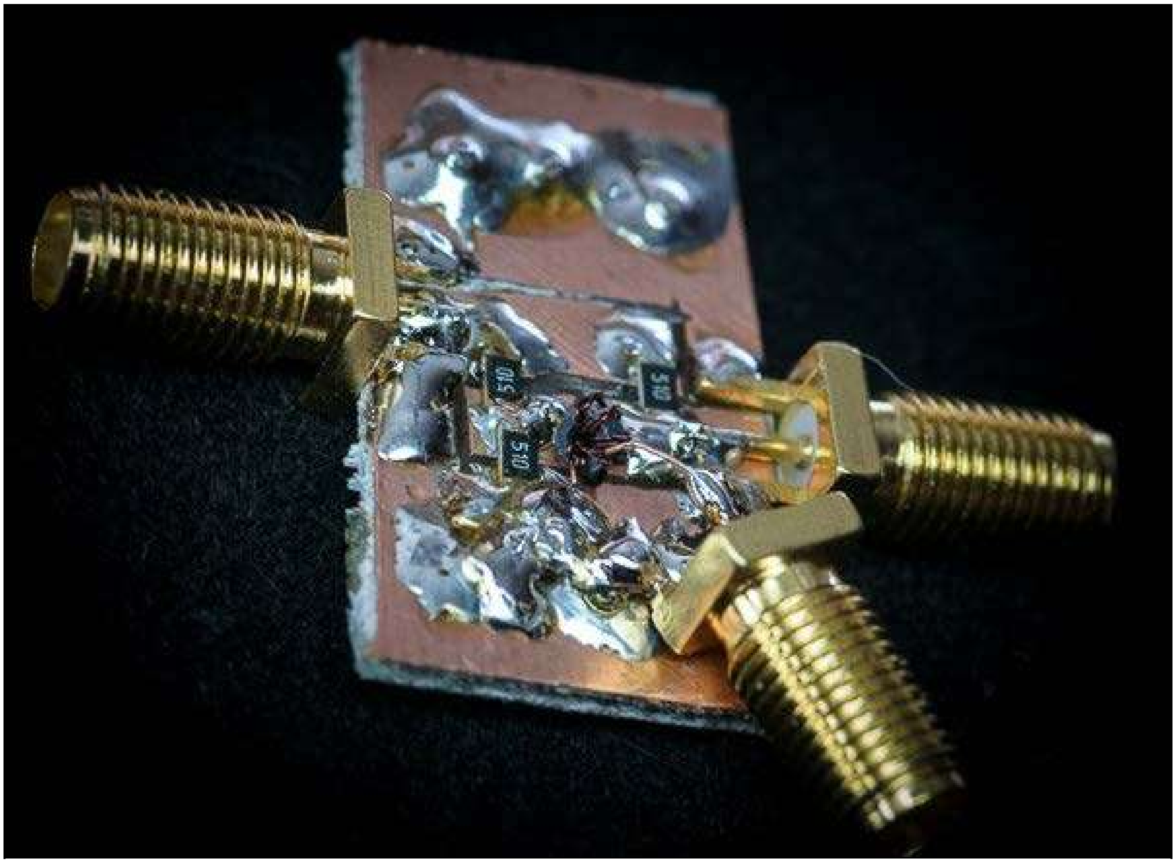

Above — Magnified return loss bridge #2 breadboard. Serious UHF enthusiast Paul Wade, W1GHz found that the smaller your breadboard, the better the UHF directivity in his bridge experiments: My findings support his. In RLB #2, I went with plain, size 1206, 5% resistors. I just ordered some size 0603 49.9 Ω 1% resistors for future experiments in my quest to make my ultimate UHF RLB.

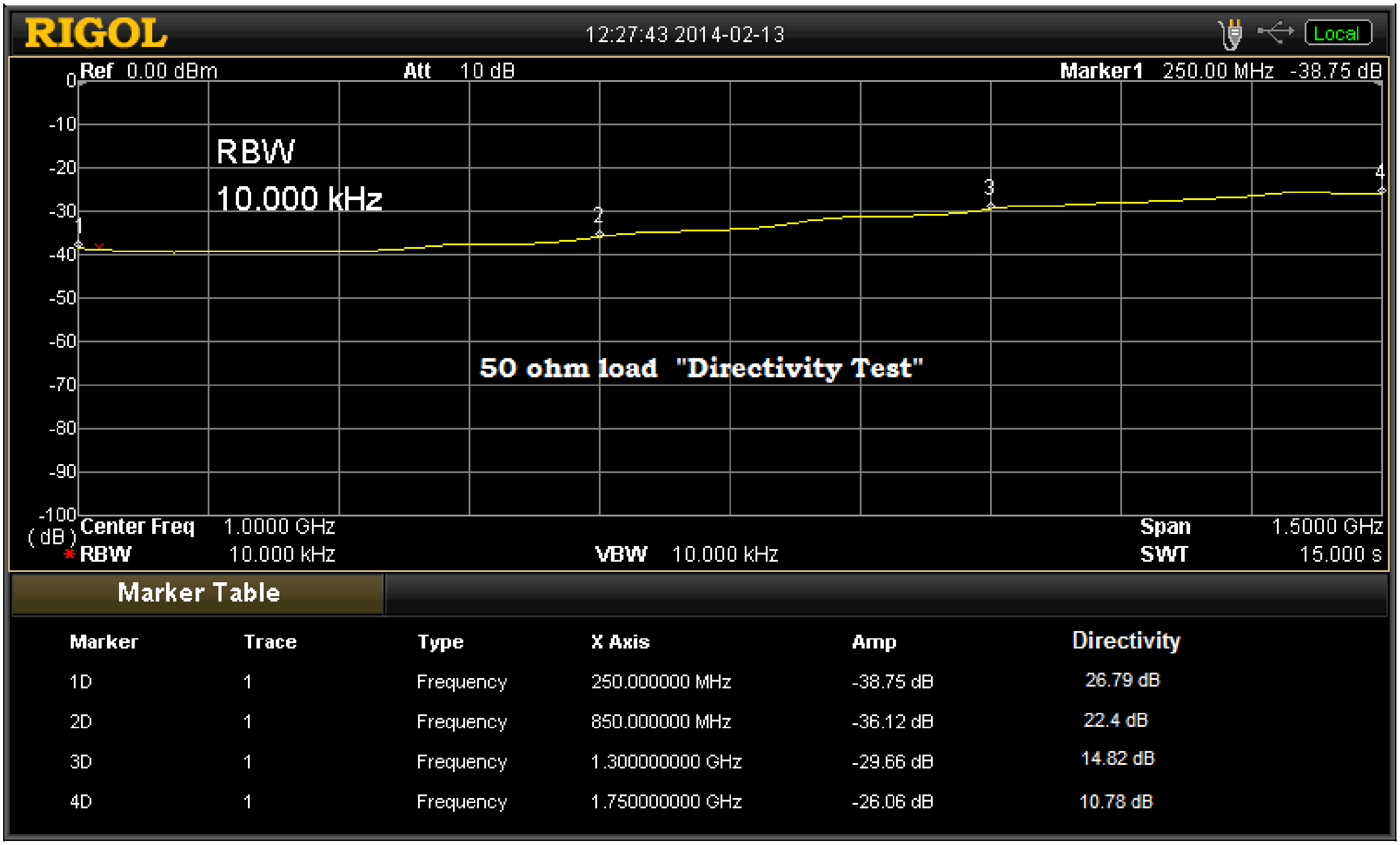
RIGOL
12:27:43 2014-02-13
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
1
2
3
4
RBW
10.000 kHz
50 ohm load "Directivity Test"
Marker Table

Ref 0.00 dBm
Att 10 dB
Marker1 250.00 MHz -38.75 dB
Center Freq 1.0000 GHz
Span 1.5000 GHz
RBW 10.000 kHz
VBW 10.000 kHz
SWT 15.000 s

Marker	Trace	Type	X Axis	Amp	Directivity
1D	1	Frequency	250.000000 MHz	-38.75 dB	26.79 dB
2D	1	Frequency	850.000000 MHz	-36.12 dB	22.4 dB
3D	1	Frequency	1.300000000 GHz	-29.66 dB	14.82 dB
4D	1	Frequency	1.750000000 GHz	-26.06 dB	10.78 dB

Above — The sweep of RLB #2 from 250 MHz to 1.75 GHz showing the directivity at 4 frequency points. Above ~900 MHz, directivity falls right off, so this bridge won't cut it for a serious UHF bench. I tried reducing T1 to 3 bifilar windings but this worsened the directivity. Click for a different sweep of RLB # 2 from 50 MHz to 1.5 GHz. Directivity = 27.5 dB at 50 MHz, so this bridge would work great for VHF. I imagine that SMD 49.9 Ω 1% resistors might boost directivity a little.

Of course, we're always impressed by the work of Kostas , SV3ORA.

Return Loss Bridge 3

Some experiments with the 3-bead RLB from Sam Wetterlin follow:

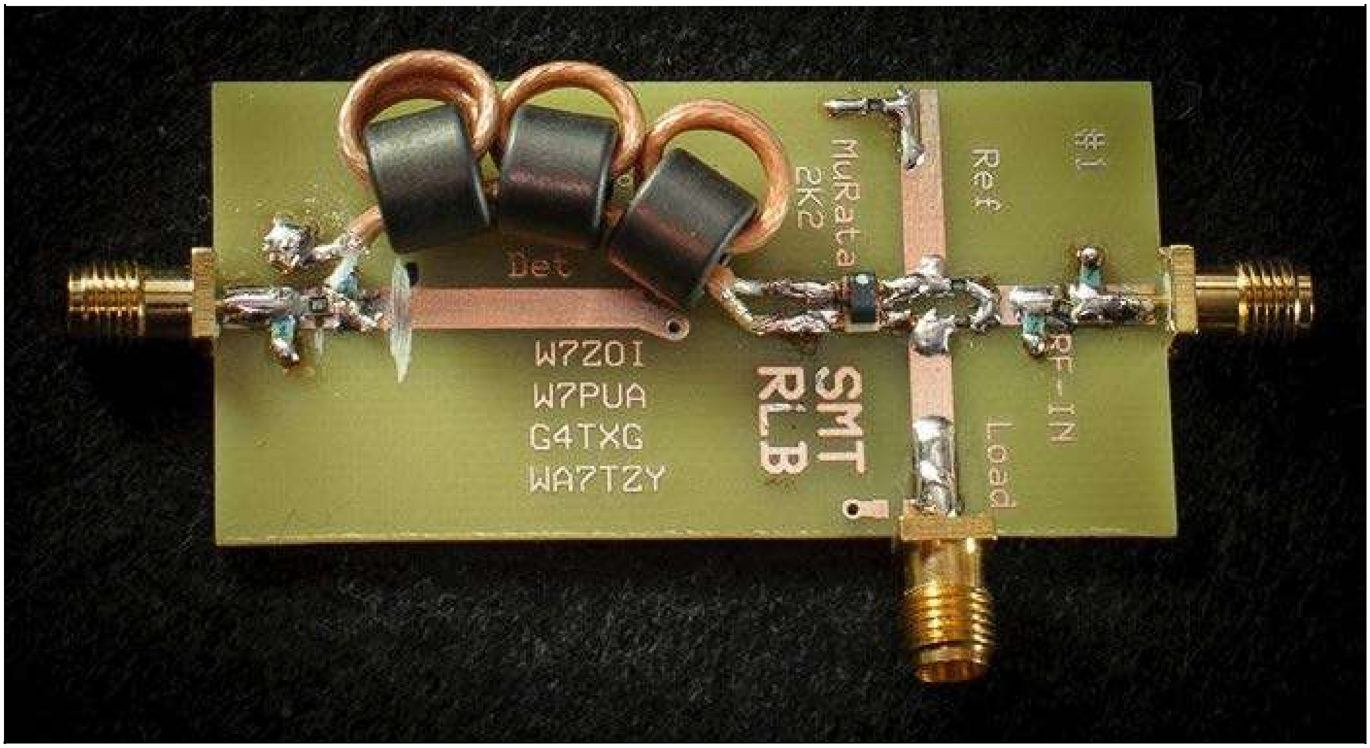
MuRata
2K2
Ref
#1
Det
W7ZOI
W7PUA
G4TXG
WA7TZY
SMT
RLB
RF-IN
Load

Above — My build of the Wetterlin 3-bead return loss bridge. This PC board was the only 1 left over from a 2006 joint project between the 4 Hams whose calls are etched on the PCB. Wes sent me this board and some of the parts and I'm grateful. My 3 dB pads = 294 Ω 1% resistors plus an 18 Ω 5% in size 0805. My 49.9 Ω 1% resistors = size 0603. The transformer = a Minicircuits Labs TC1-1-1-13M+. Click for the ferrite description. I used 24 gauge copper via wires to connect grounded nodes to the ground plane on the back of the PC board. I chose to permanently solder a 49.9 Ω resistor on the reference port.

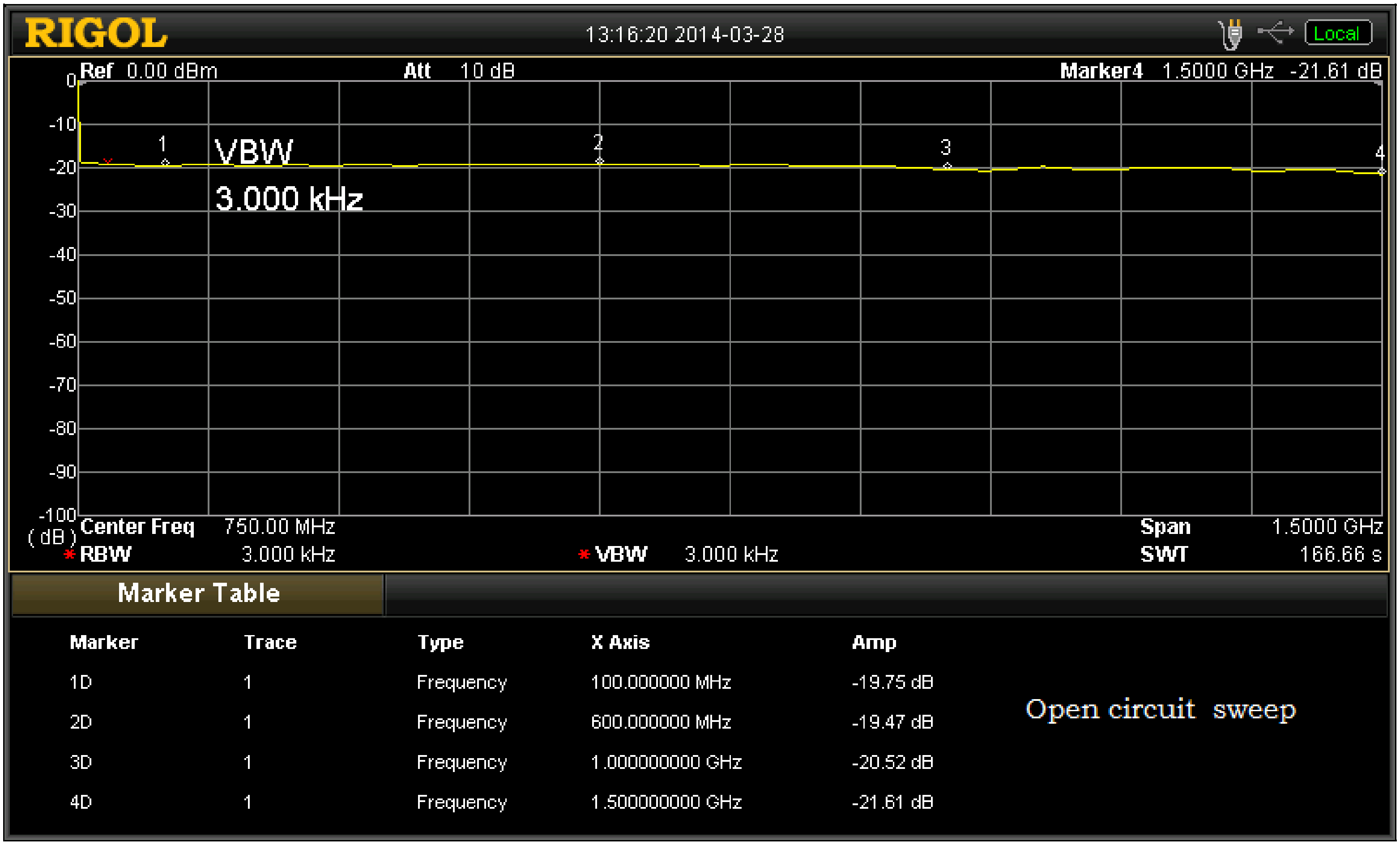
RIGOL
13:16:20 2014-03-28
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
1
2
3
4
VBW
3.000 kHz

Ref 0.00 dBm
Att 10 dB
Marker4 1.5000 GHz -21.61 dB
Center Freq 750.00 MHz
RBW 3.000 kHz
VBW 3.000 kHz
Span 1.5000 GHz
SWT 166.66 s

Marker Table

Marker	Trace	Type	X Axis	Amp
1D	1	Frequency	100.000000 MHz	-19.75 dB
2D	1	Frequency	600.000000 MHz	-19.47 dB
3D	1	Frequency	1.000000000 GHz	-20.52 dB
4D	1	Frequency	1.500000000 GHz	-21.61 dB

Open circuit sweep

Above — Open circuit sweep with 4 markers set. Span = 1.5 GHz.

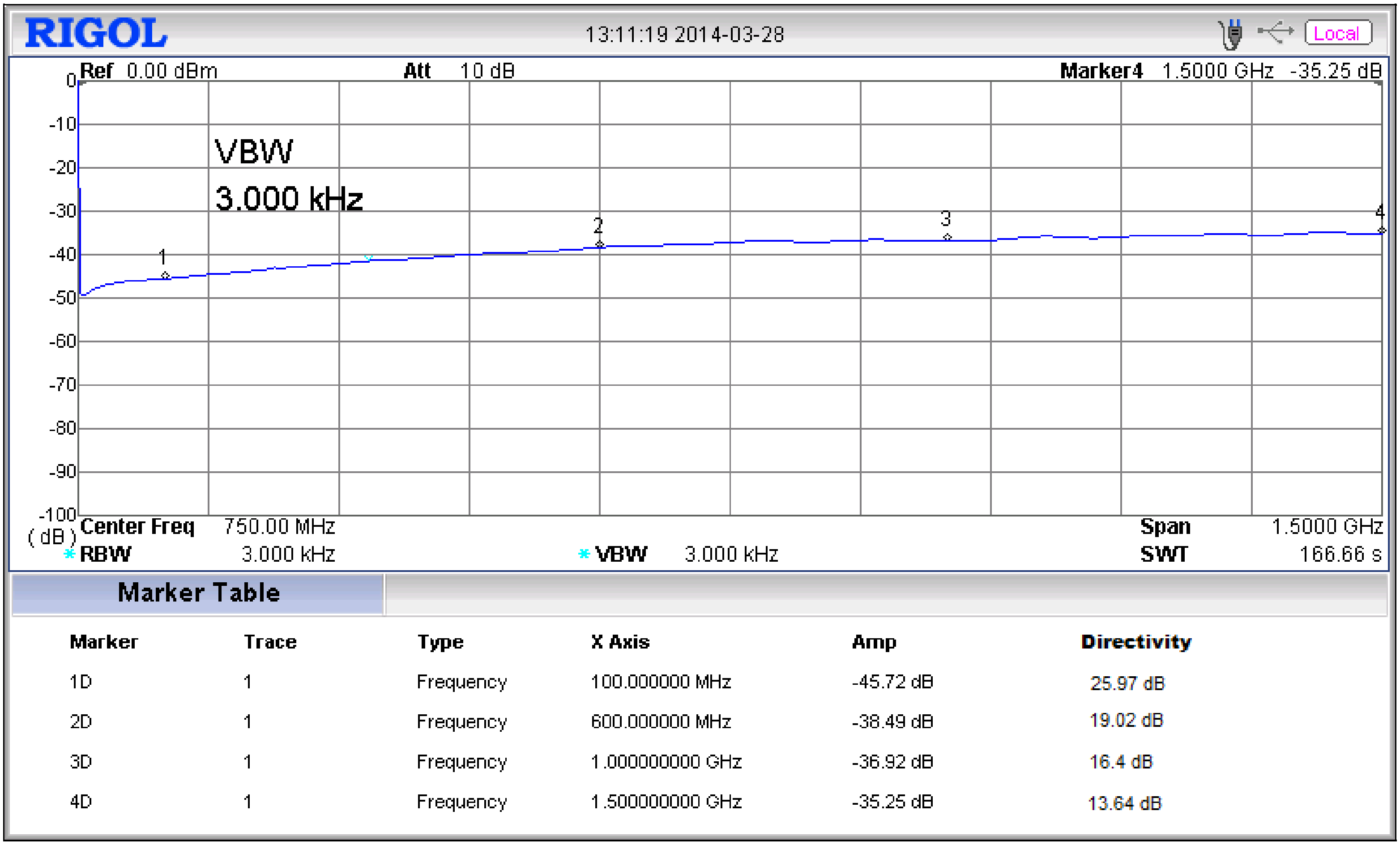
RIGOL
13:11:19 2014-03-28
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
VBW
3.000 kHz
1
2
3
4

Ref 0.00 dBm **Att** 10 dB **Marker4** 1.5000 GHz -35.25 dB

Center Freq 750.00 MHz **Span** 1.5000 GHz

RBW 3.000 kHz **VBW** 3.000 kHz **SWT** 166.66 s

Marker Table

Marker	**Trace**	**Type**	**X Axis**	**Amp**	**Directivity**
1D	1	Frequency	100.000000 MHz	-45.72 dB	25.97 dB
2D	1	Frequency	600.000000 MHz	-38.49 dB	19.02 dB
3D	1	Frequency	1.000000000 GHz	-36.92 dB	16.4 dB
4D	1	Frequency	1.500000000 GHz	-35.25 dB	13.64 dB

Above — 50 ohm load termination sweep with calculated directivity. Click for the 50 Ω sweep where the span = 200 MHz. I was disappointed with the UHF directivity in *my* build.

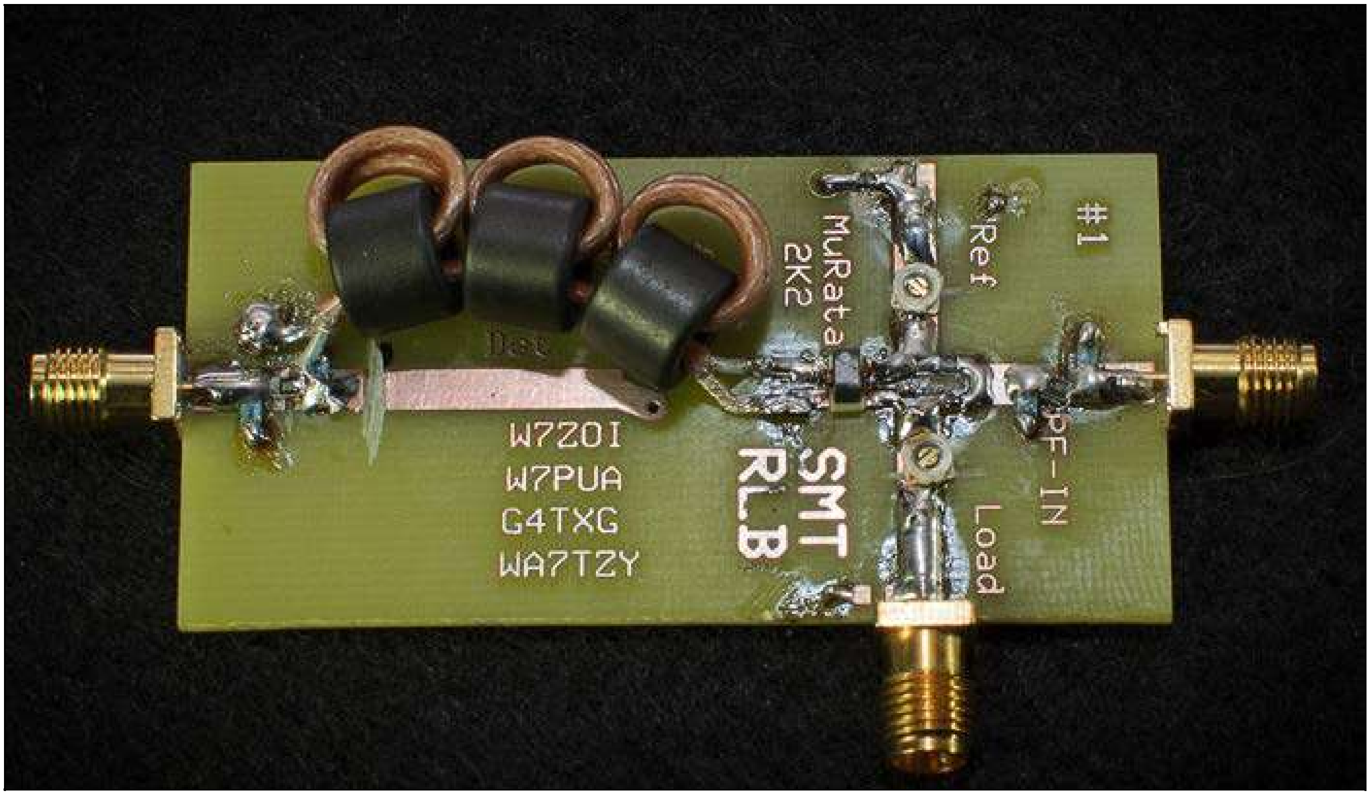
MuRata
2K2
Ref
#1
Det
W7ZOI
W7PUA
G4TXG
WA7TZY
SMT
RLB
RF-IN
Load

Above — The 3-bead RLB with added series trimmer capacitors on the load and reference ports. SMT trimmer caps = Voltronics JR300 5.5-30 pF with a Q of 1500. I tried both shunt and series caps to try and boost UHF directivity.

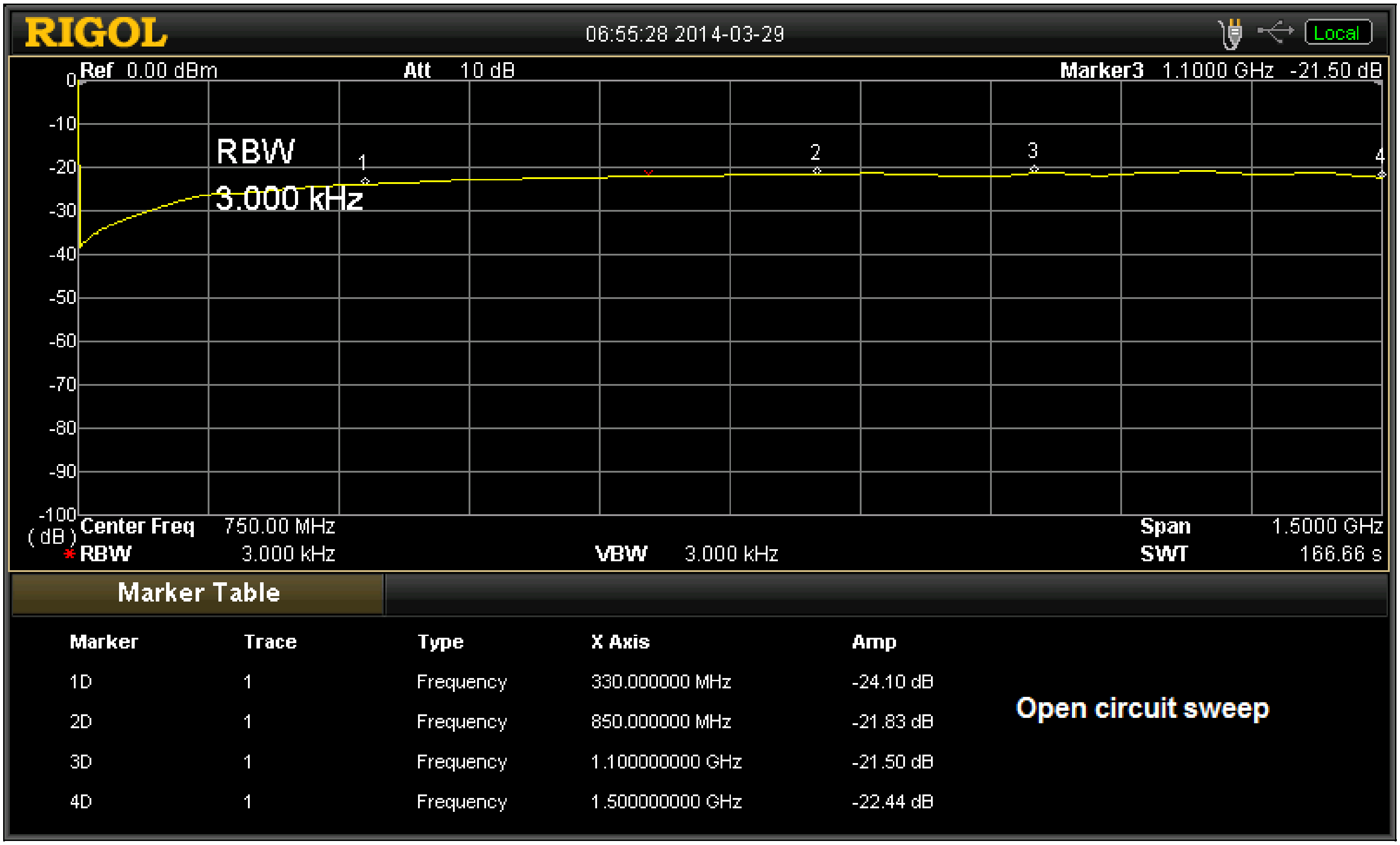
RIGOL
06:55:28 2014-03-29
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
RBW
3.000 kHz
1
2
3
4
Open circuit sweep

Ref 0.00 dBm
Att 10 dB
Marker3 1.1000 GHz -21.50 dB

Center Freq 750.00 MHz
Span 1.5000 GHz
*RBW 3.000 kHz
VBW 3.000 kHz
SWT 166.66 s

Marker Table

Marker	Trace	Type	X Axis	Amp
1D	1	Frequency	330.000000 MHz	-24.10 dB
2D	1	Frequency	850.000000 MHz	-21.83 dB
3D	1	Frequency	1.100000000 GHz	-21.50 dB
4D	1	Frequency	1.500000000 GHz	-22.44 dB

Above — Open circuit sweep of the 3-bead bridge with series capacitors on the load and reference ports.

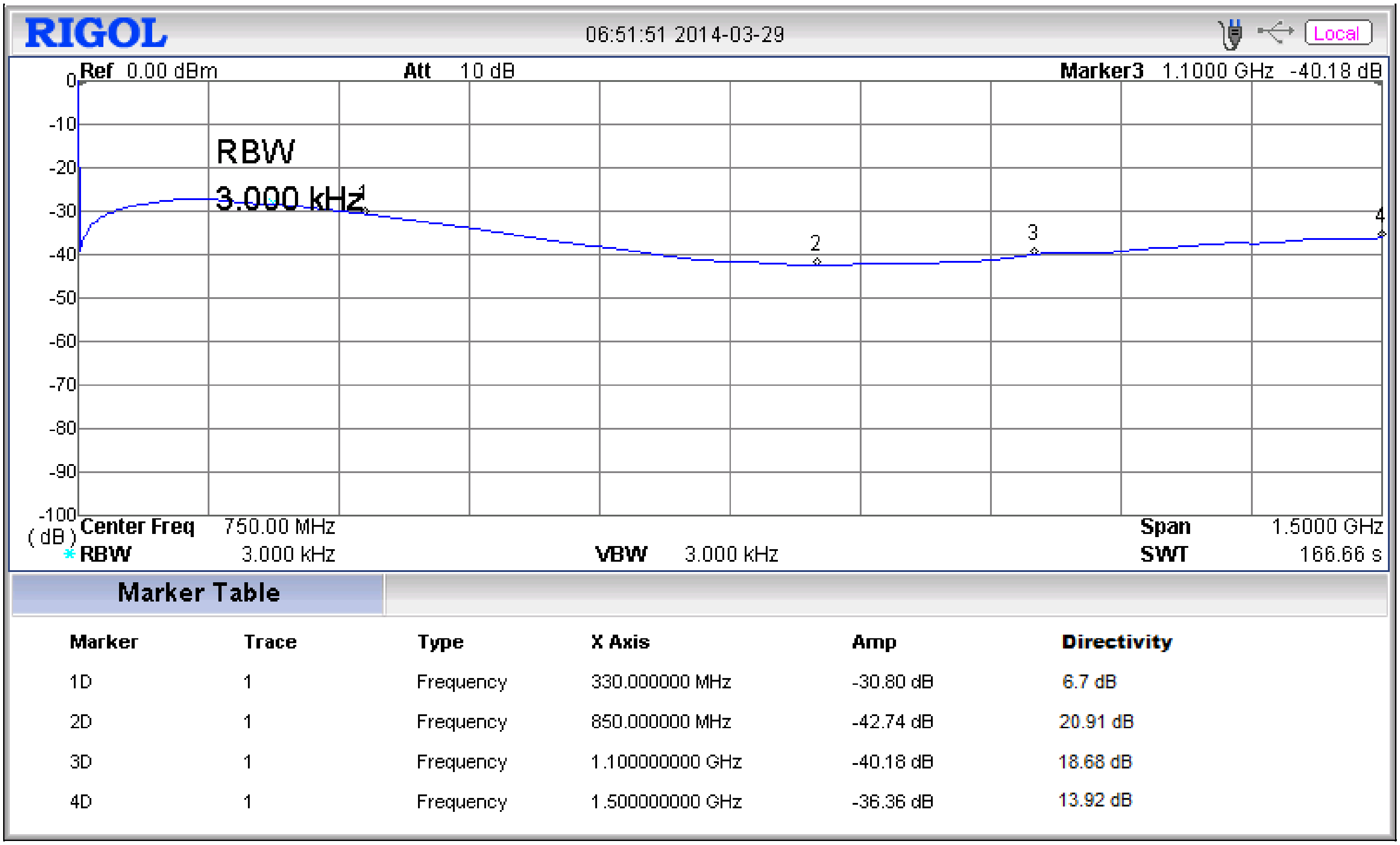

06:51:51 2014-03-29

Ref 0.00 dBm **Att** 10 dB **Marker3** 1.1000 GHz -40.18 dB

Center Freq 750.00 MHz **Span** 1.5000 GHz

RBW 3.000 kHz **VBW** 3.000 kHz **SWT** 166.66 s

Marker Table

Marker	Trace	Type	X Axis	Amp	Directivity
1D	1	Frequency	330.000000 MHz	-30.80 dB	6.7 dB
2D	1	Frequency	850.000000 MHz	-42.74 dB	20.91 dB
3D	1	Frequency	1.100000000 GHz	-40.18 dB	18.68 dB
4D	1	Frequency	1.500000000 GHz	-36.36 dB	13.92 dB

Above — 50 Ω load sweep. By carefully tweaking the 2 trimmer caps, I set the deepest NULL at ~ 850 MHz; although it's not as deep as I want. Low end directivity was compromised.

Return Loss Bridge 4

Since my bench experiments with the "string of pearls" (SOP) return loss bridge showed promise, I kept going and eventually built a version closer to my goal: A bridge for 300 MHz to 2 GHz with decent directivity.

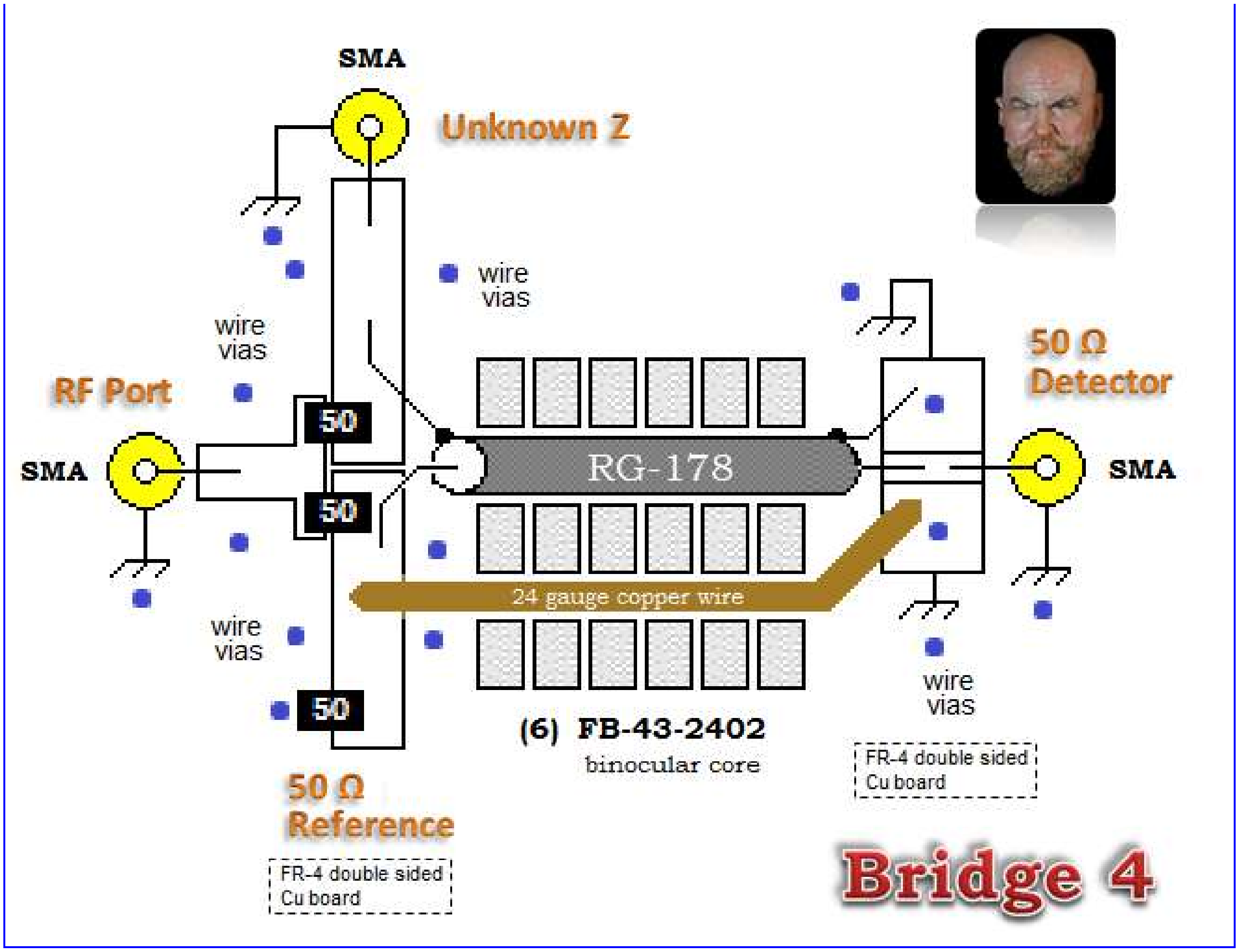
SMA
Unknown Z
wire vias
wire vias
RF Port
SMA
50
50
RG-178
24 gauge copper wire
wire vias
50
50 Ω Reference
FR-4 double sided Cu board
(6) FB-43-2402
binocular core
50 Ω Detector
SMA
wire vias
FR-4 double sided Cu board
Bridge 4

Above — My SOP build diagram of Bridge #4. I cut traces in 2-sided FR-4 Cu board to make the needed microstripline paths.The two 49.9 Ω 1% bridge resistors = size 0603, however, my reference port 49.9 Ω, 1% resistor was size 1206. The braid of the coax goes to the unknown impedance port and the center core to the reference port. Six FB-43-2402 binocular cores just passed the diameter of the RG-178 coax and binocular ferrites boosted balance compared to seperate beads in my experiments.

Rather than run coax through the remaining hole in the ferrite string, I just used a 24 gauge piece of Cu wire for balun symmetry. I left the green insulation on the wire to help keep it in place during soldering — I used the same wire spool to make the via wires.

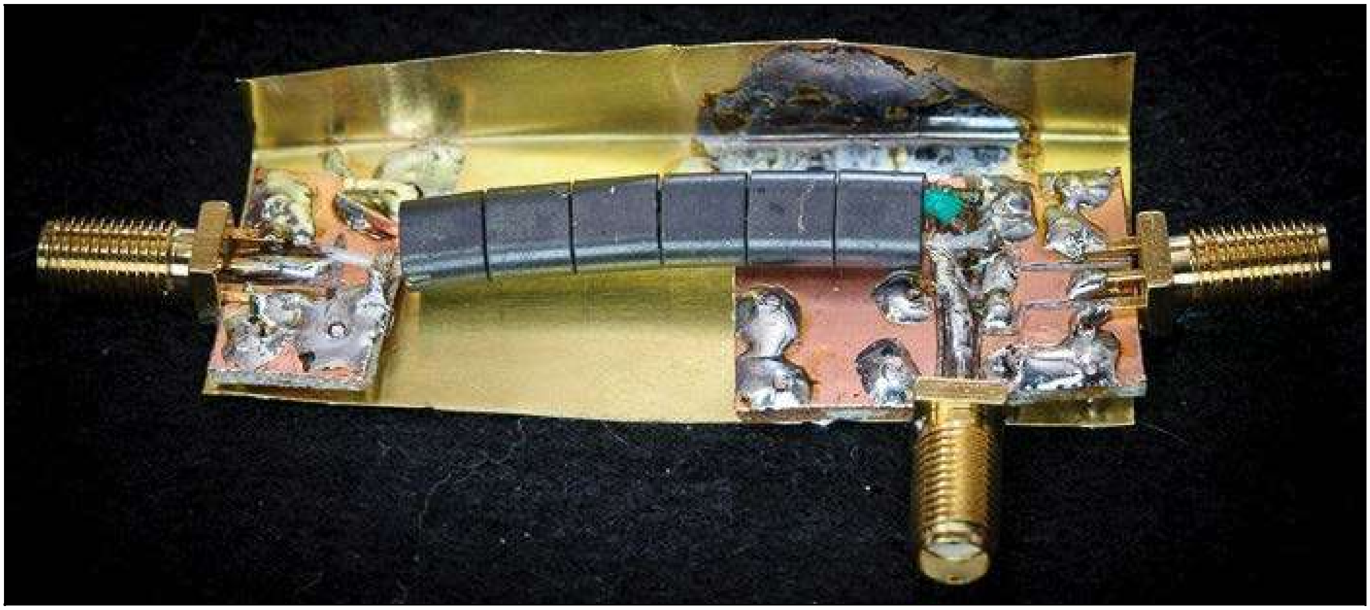

Above — Breadboard photograph. Click for a larger photo. I soldered some thin brass sheet metal to electrically join the ground planes of the 2 carved boards [represents the normal RF-tight metal chassis we apply].The brass sheet boosted directivity 8 dB by improving balun function.

Comments by a prominent figure in our hobby ....."and the thing is 'ugly to the core'. Way to go."

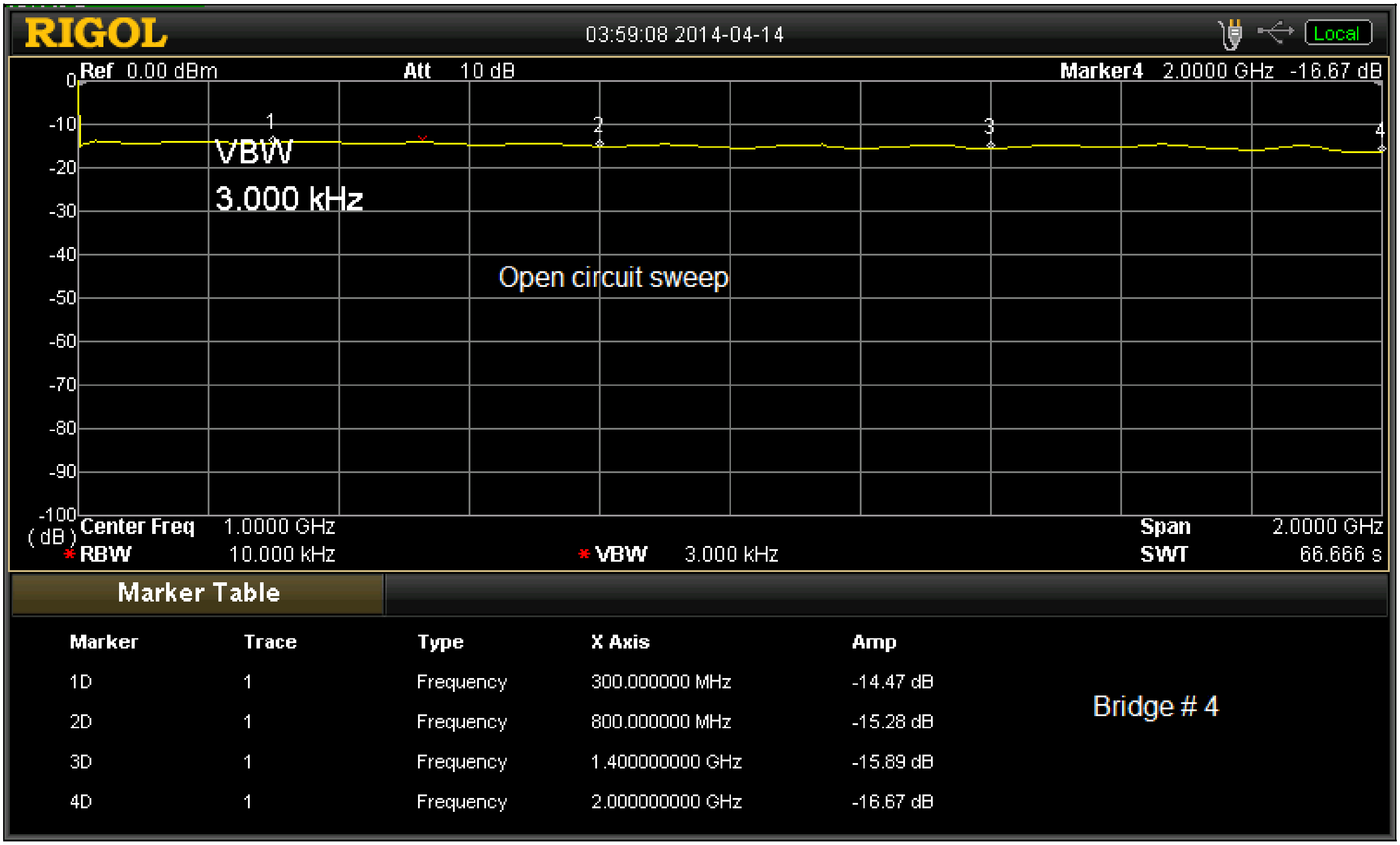

RIGOL
03:59:08 2014-04-14
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
1
2
3
4
VBW
3.000 kHz
Open circuit sweep
Bridge # 4

Ref 0.00 dBm **Att** 10 dB **Marker4** 2.0000 GHz -16.67 dB

Center Freq 1.0000 GHz **Span** 2.0000 GHz

* **RBW** 10.000 kHz * **VBW** 3.000 kHz **SWT** 66.666 s

Marker Table

Marker	Trace	Type	X Axis	Amp
1D	1	Frequency	300.000000 MHz	-14.47 dB
2D	1	Frequency	800.000000 MHz	-15.28 dB
3D	1	Frequency	1.400000000 GHz	-15.89 dB
4D	1	Frequency	2.000000000 GHz	-16.67 dB

Above — Open circuit sweep with a 2 GHz span and center frequency of 1 GHz.

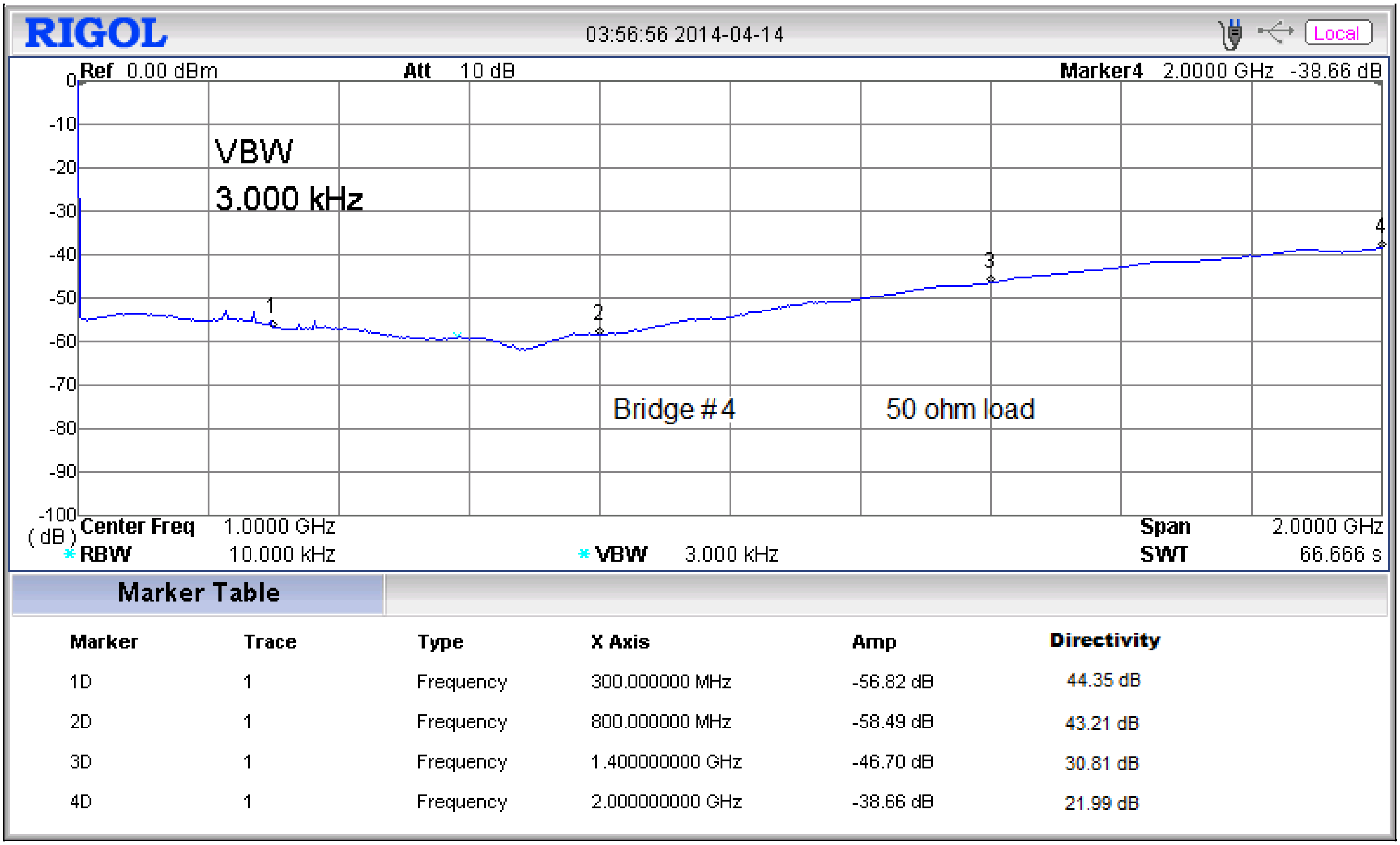
RIGOL
03:56:56 2014-04-14
Local
Ref 0.00 dBm
Att 10 dB
Marker4 2.0000 GHz -38.66 dB
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
VBW
3.000 kHz
1
2
3
4
Bridge #4
50 ohm load
Center Freq 1.0000 GHz
Span 2.0000 GHz
RBW 10.000 kHz
VBW 3.000 kHz
SWT 66.666 s
Marker Table

Marker	Trace	Type	X Axis	Amp	Directivity
1D	1	Frequency	300.000000 MHz	-56.82 dB	44.35 dB
2D	1	Frequency	800.000000 MHz	-58.49 dB	43.21 dB
3D	1	Frequency	1.400000000 GHz	-46.70 dB	30.81 dB
4D	1	Frequency	2.000000000 GHz	-38.66 dB	21.99 dB

Above — Sweep with a 50 Ω load threaded on the unknown impedance port and 4 markers showing directivity at those 4 frequencies. Yay — this is much better! I wanted at least 30 dB directivity at 1296 MHz and got it.

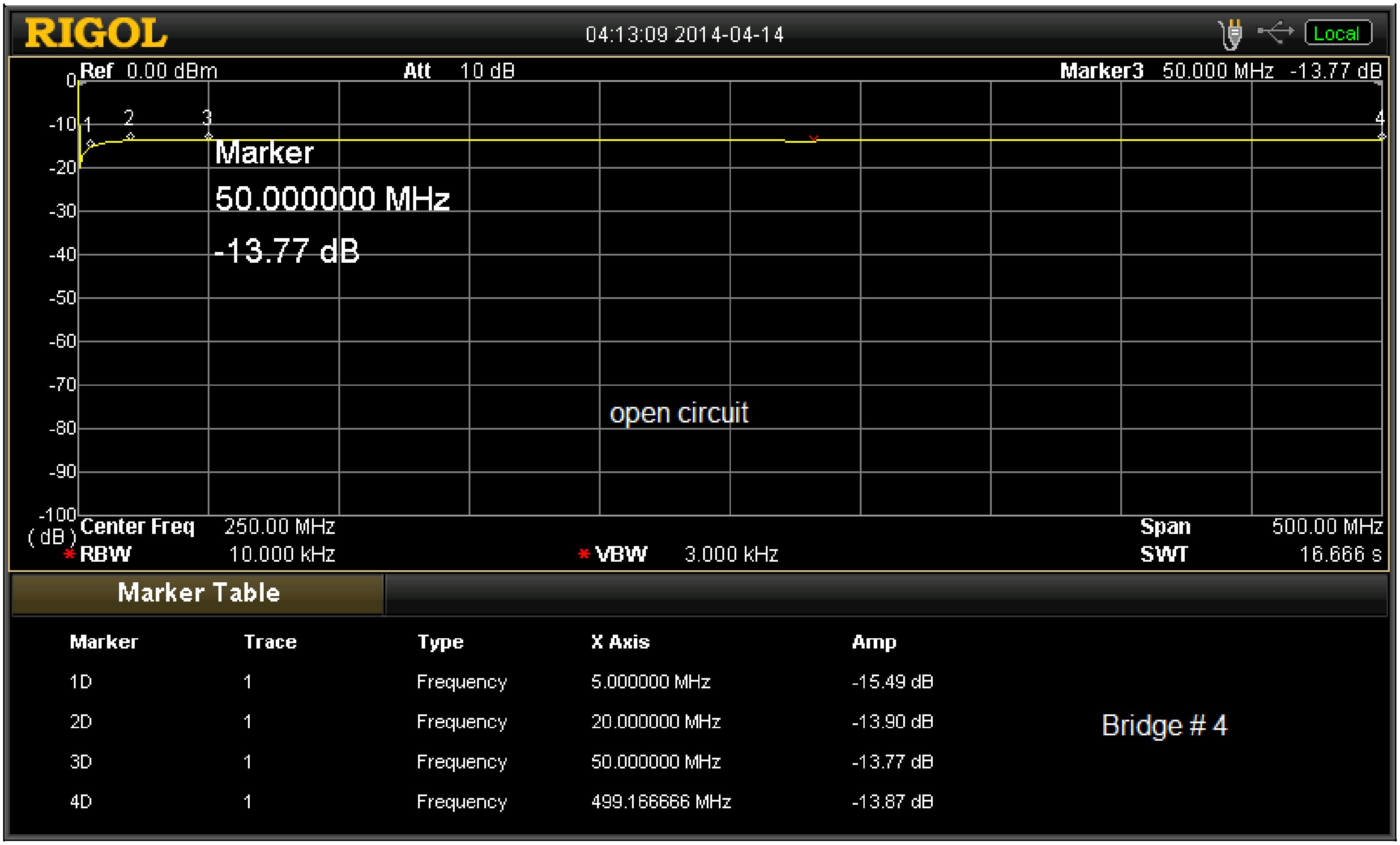
RIGOL
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
1
2
3
4
Marker
50.000000 MHz
-13.77 dB
open circuit
Bridge # 4

04:13:09 2014-04-14

Ref 0.00 dBm **Att** 10 dB **Marker3** 50.000 MHz -13.77 dB

Center Freq 250.00 MHz **Span** 500.00 MHz

* **RBW** 10.000 kHz * **VBW** 3.000 kHz **SWT** 16.666 s

Marker Table

Marker	**Trace**	**Type**	**X Axis**	**Amp**
1D	1	Frequency	5.000000 MHz	-15.49 dB
2D	1	Frequency	20.000000 MHz	-13.90 dB
3D	1	Frequency	50.000000 MHz	-13.77 dB
4D	1	Frequency	499.166666 MHz	-13.87 dB

Above — Open circuit sweep with a 500 MHz span and center frequency of 250 MHz. I set 2 of the markers to evaluate the directivity at HF.

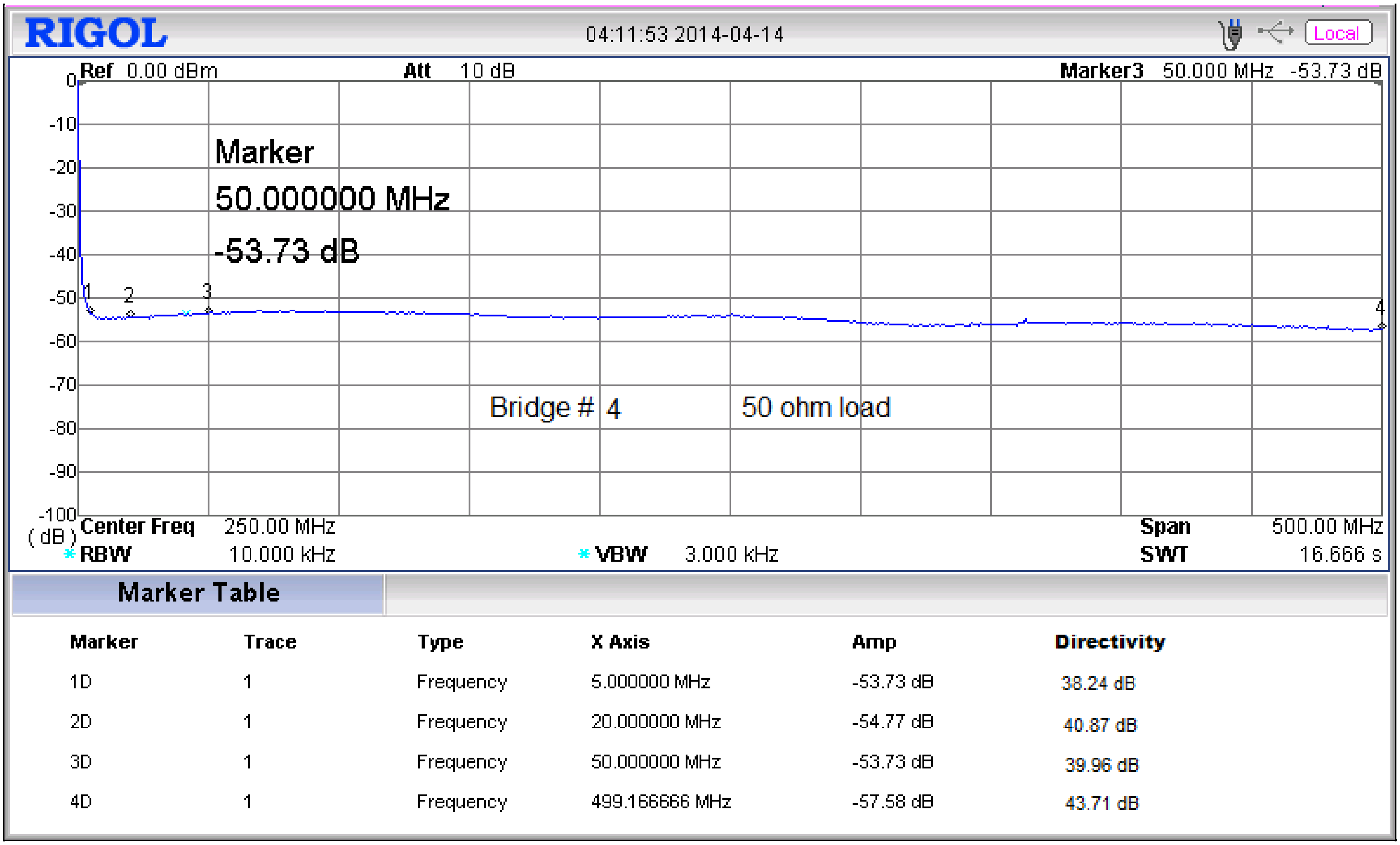
RIGOL
04:11:53 2014-04-14
Local
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
1
2
3
4
Marker
50.000000 MHz
-53.73 dB
Bridge # 4
50 ohm load

Ref 0.00 dBm
Att 10 dB
Marker3 50.000 MHz -53.73 dB
Center Freq 250.00 MHz
Span 500.00 MHz
RBW 10.000 kHz
VBW 3.000 kHz
SWT 16.666 s

Marker Table

Marker	Trace	Type	X Axis	Amp	Directivity
1D	1	Frequency	5.000000 MHz	-53.73 dB	38.24 dB
2D	1	Frequency	20.000000 MHz	-54.77 dB	40.87 dB
3D	1	Frequency	50.000000 MHz	-53.73 dB	39.96 dB
4D	1	Frequency	499.166666 MHz	-57.58 dB	43.71 dB

Above — 500 MHz sweep of Bridge #4 with a 50 load on the unknown impedance port. The directivity at HF surprised me. Although better suited for VHF-UHF, this bridge will work okay from 5 MHz to ~1.5 GHz.

---

4. VCO Experiments

Ask experimenters about homebrew RF oscillators and you might hear about a pleasing paradox — vexing and satisfying in the same breath! At UHF we usually make a VCO and keep it on frequency by phase/frequency locking to a xtal oscillator reference. I sought some wideband, free running VCOs in hopes I'd inch up the learning curve.

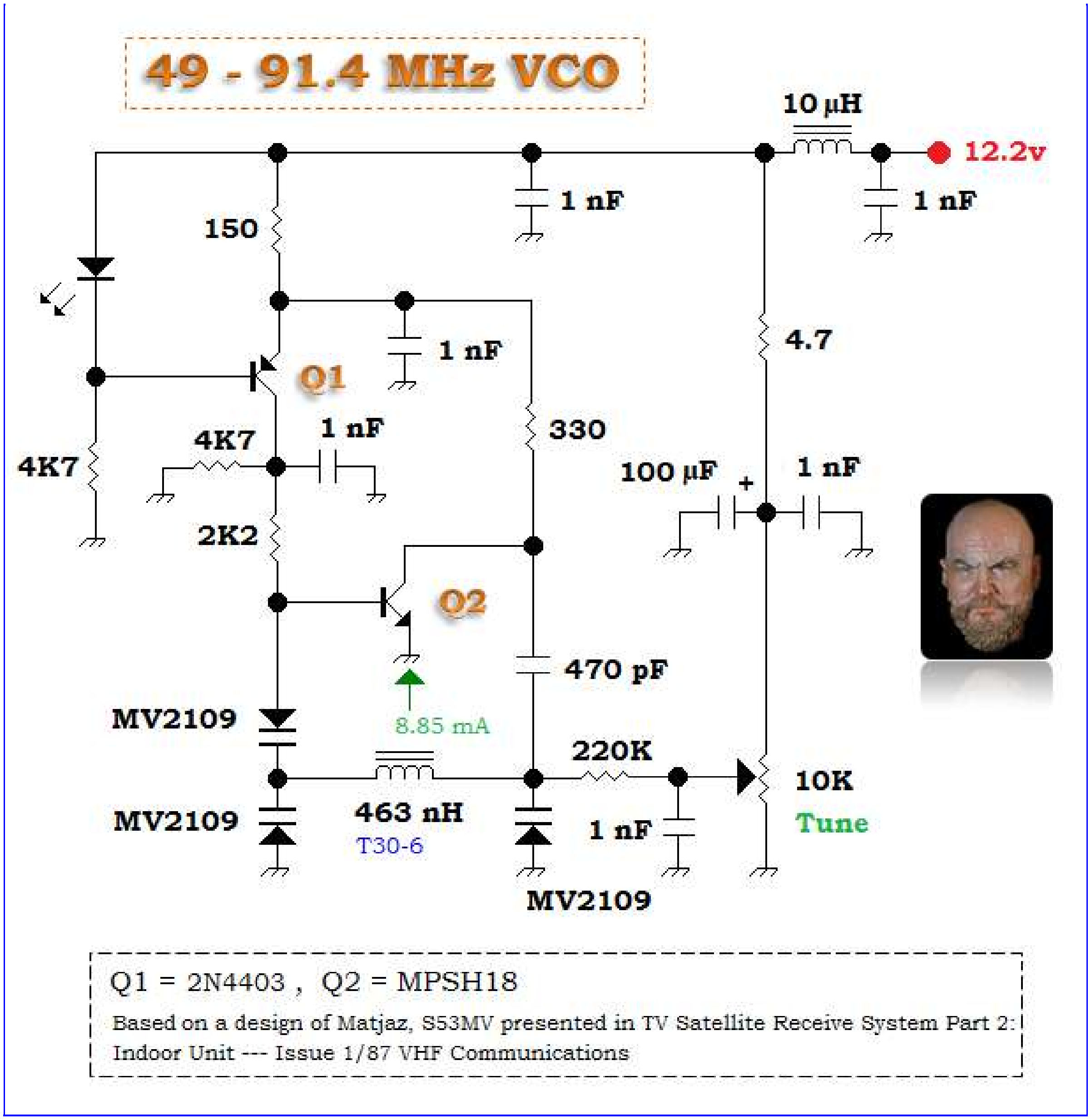
49 - 91.4 MHz VCO

Q1 = 2N4403 , Q2 = MPSH18

Based on a design of Matjaz, S53MV presented in TV Satellite Receive System Part 2: Indoor Unit --- Issue 1/87 VHF Communications

Above — My first attempt at a VCO circuit described by Matjaž, S53MV in a 2-part article published in VHF Communications magazine. The original circuit contained a tunable RF filter, mixer, buffer amplifer + more and I recommend you purchase the 2 articles for your own analysis. Shown with permission from Andy Barter of VHF Communications magazine, I ordered all issues from the 1980s which came on this DVD.

With 28v varactors, this VCO will tune 1 octave minimum.

To bench investigate this stunning common emitter Colpitts variant, I first built it at VHF using through-hole parts with a standard 12 volt tuning voltage that tuned from 49 to 91.4 MHz. The output came from 1 link around the 463 nH inductor terminated in a 1K resistor at the hot end + ground at the cold end to allow 10X probe measurement with my DSO.

Matjaž, S53MV describes his circuit function here. A variable C on the base in combination with the input capacitance of Q2 operates as a tuned impedance matching network and preserves a steep phase slope (which means you have higher Q) — a requirement for low phase noise. In essence, the inductor with a varactor at each end forms an adjustable pi-network to provide correct phase shift tracking.

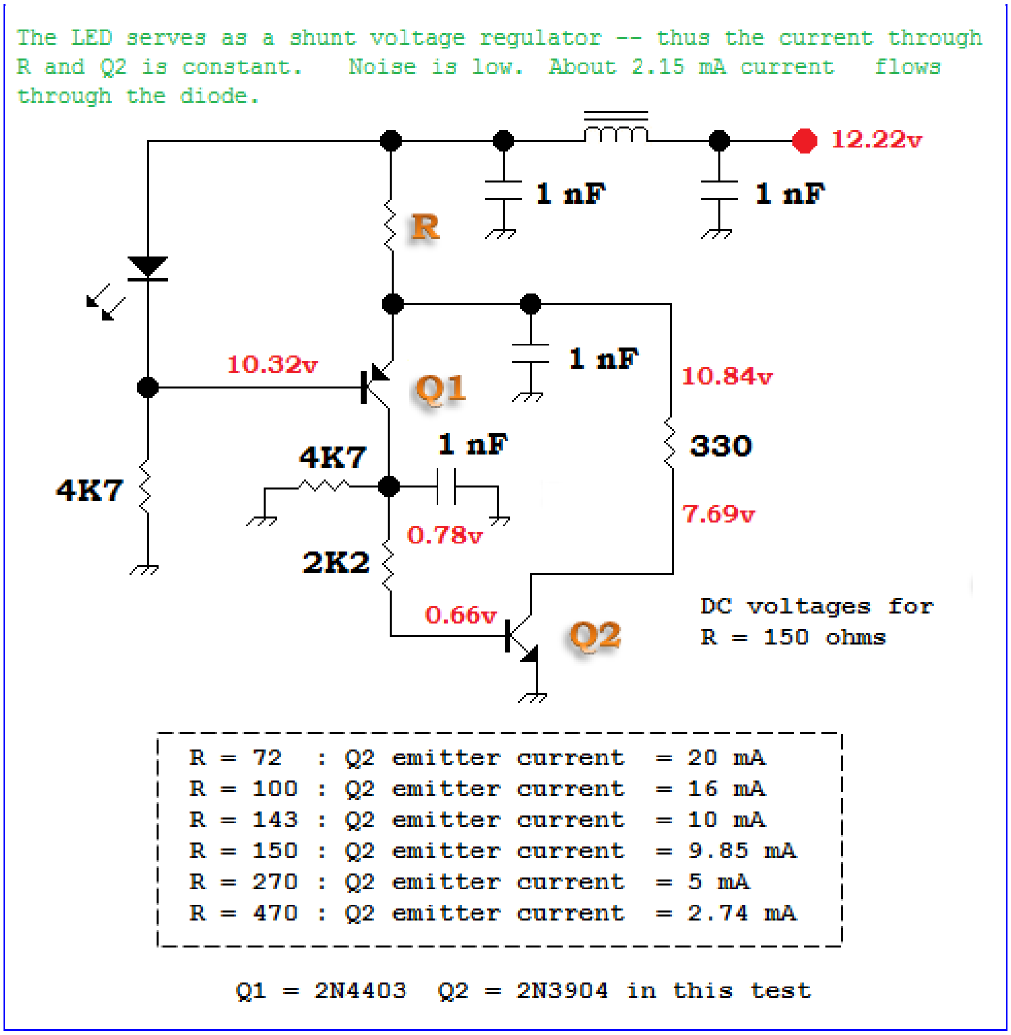

The LED serves as a shunt voltage regulator -- thus the current through R and Q2 is constant. Noise is low. About 2.15 mA current flows through the diode.

12.22v
1 nF
1 nF
R
1 nF
10.32v
Q1
10.84v
1 nF
4K7
330
4K7
0.78v
7.69v
2K2
0.66v
Q2

DC voltages for
R = 150 ohms

R = 72 : Q2 emitter current = 20 mA
R = 100 : Q2 emitter current = 16 mA
R = 143 : Q2 emitter current = 10 mA
R = 150 : Q2 emitter current = 9.85 mA
R = 270 : Q2 emitter current = 5 mA
R = 470 : Q2 emitter current = 2.74 mA

Q1 = 2N4403 Q2 = 2N3904 in this test

Above — I built the low-noise, temperature-stable, PNP current source using though-hole parts. R provides a convenient way to vary collector current, although changing the PNP base bias also works.

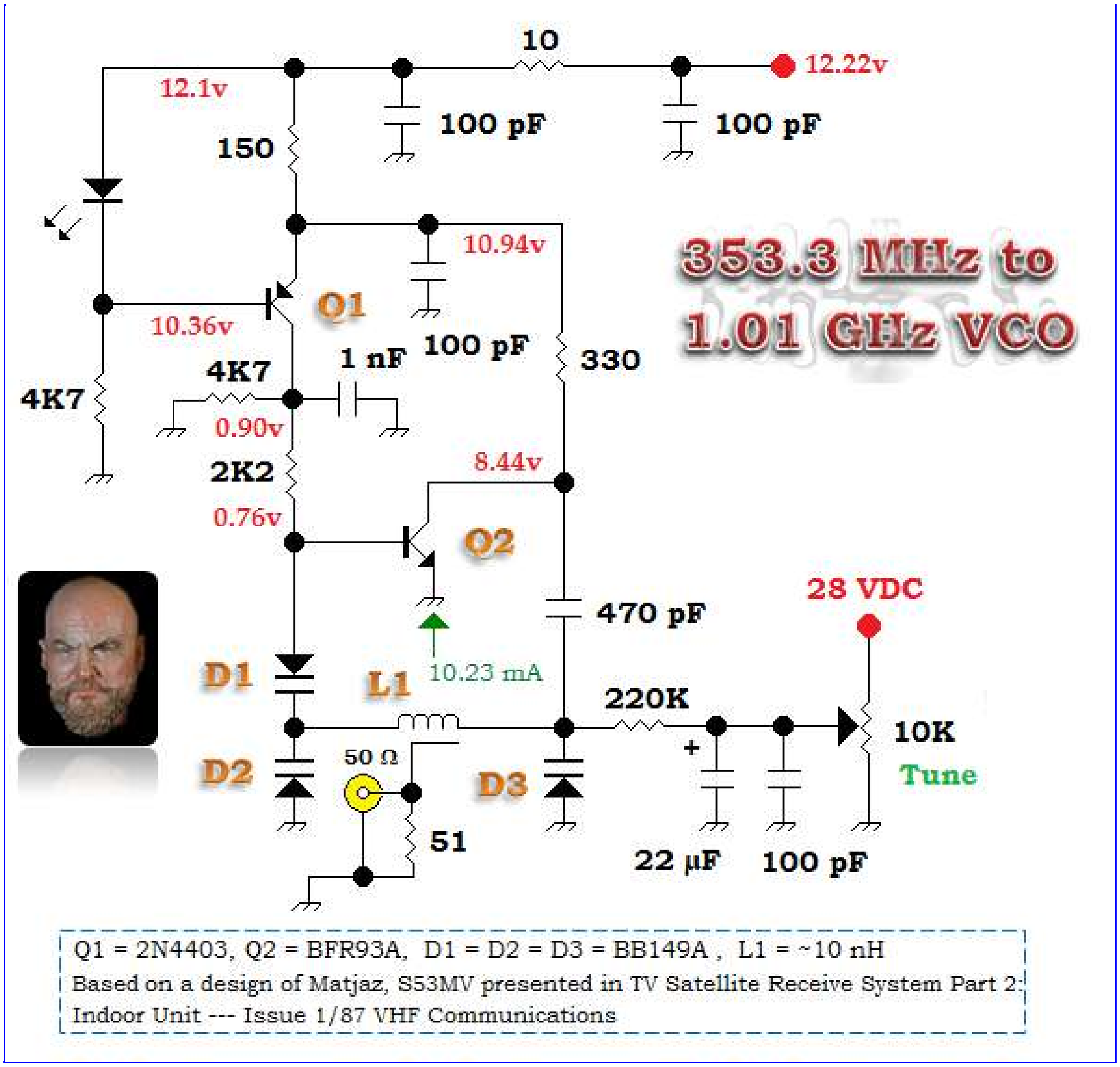

Q1 = 2N4403, Q2 = BFR93A, D1 = D2 = D3 = BB149A , L1 = ~10 nH
Based on a design of Matjaz, S53MV presented in TV Satellite Receive System Part 2:
Indoor Unit --- Issue 1/87 VHF Communications

Above — My UHF VCO schematic. Click for the breadboard photo. A wire positioned near L1 provided the output into my spectrum analyzer. Click for a sweep with the reverse DC voltage set at 0v. Click for the a sweep with the tuning at 28v to establish the lowest possible varactor capacitance. Markers indicate the carrier frequency. As shown, a common emitter VCO exhibits rich harmonic energy and viewing the second harmonic at 2 GHz felt really cool in the 28v sweep. Not to worry — the tracking generator in the Rigol TG + SA units also exhibit strong harmonics and work fine for sweeping.

Click for a pdf file showing how I laid out my board. I cut traces on the top of 2-sided FR4 board and connected any grounded points on top to the bottom ground plane with a 24 gauge copper via wire. A 28v DC-DC Converter to supply the varactors lies on VHF-FM.

I varied L and even tried a 3 mm by 8 mm copper sheet metal inductor but this failed to oscillate. In the end, a single wire with ~10 nH inductance worked to give me the tuning span shown. L1's distance above the ground plane was determined by tuning the VCO for the smoothest transistion across its range. This circuit begs further experiments including Q2 current, L1 and the 470 pF fixed feedback capacitor.

---

5. Frequency Counter Kit

A sensitive UHF bandwidth frequency counter makes an important tool for the UHF homebrew experimenter. With a search engine, you'll find commercial and kit counters for sale along with DIY projects for builders handy with microcontrollers.

If you already own a decent counter, then a crude prescaler tool employing an IC like the MC12079 might work okay if you can tolerate or manage modulus division by 64, 128, or 256. The MC12079 requires an RF signal between -17 dBm to +3 dBm to count properly, so many builders place amplifier(s) along with anti-parallel clamping diodes before the digital IC. The Motorola datasheet shows a basic circuit to build off.

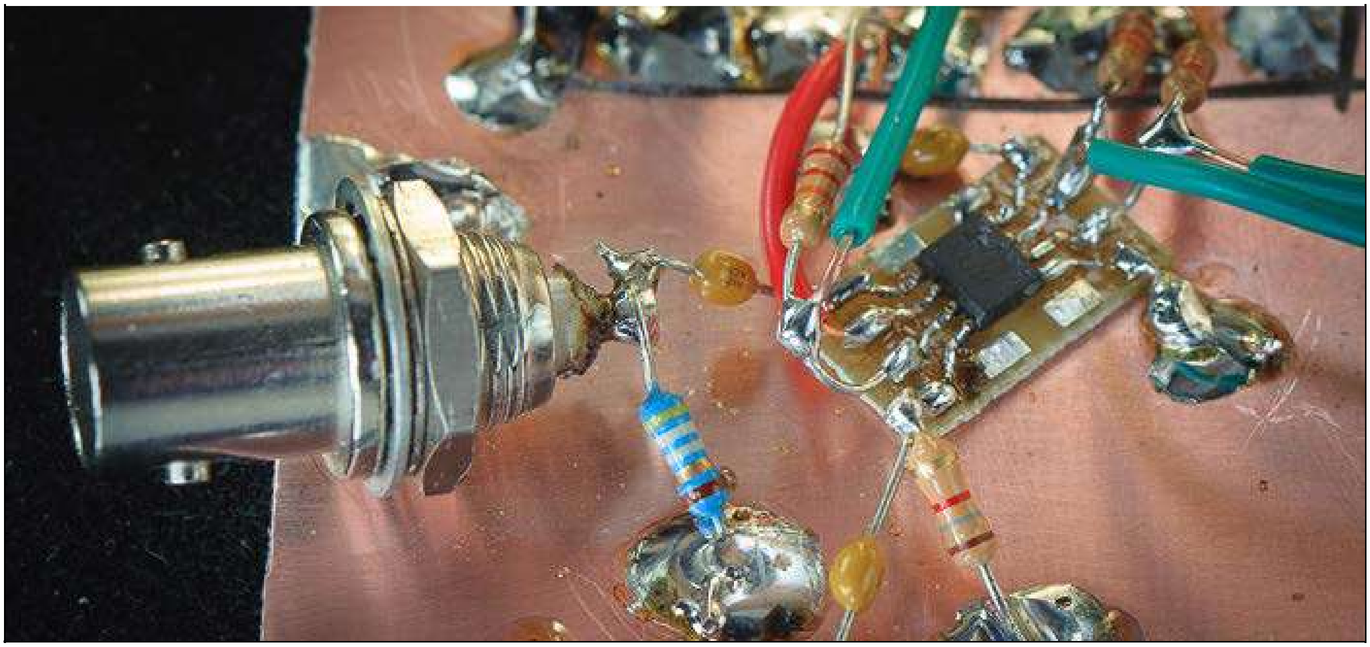

Above — My first UHF counter experiment involved the Motorola MC12079 and 12080 prescalers. Eventually I made a proper double-sided copper board device with anti-parallel diodes, a MMIC, the MC12079 and some 74AC series logic — it worked okay for casual use. The 2 big drawbacks were MC12079 self-oscillation with 0 applied input signal and having to remember to multiply measures by my set prescaler division integer. I sought a long-term solution.

A Solution Appears

In 2012 I bought a second-hand, calibrated HP 53131A with the optional A6 High Stability Timebase from an EE for ½ the going price at the time. 225 MHz bandwidth poses its only limitation; it even came with the original spiral bound manual.

This summer, Shahriar from TheSignalPath.com reviewed a Chinese kit to take the HP 53131A to 3 GHz with a clone of the original HP OPT30 Channel 2 kit. Based on his quantitative review and warm thoughts of using my sturdy HP counter at UHF, I got 1. Mine sold on eBay for $87 Canadian including shipping. Click for TheSignalPath.com video. Click for a larger photo of the kit I bought.

53181A 225 MHz
FREQUENCY COUNTER
CHANNEL 2
Enter
MEASURE
LIMITS
MATH
DISPLAY
CHANNEL 1
Freq Ch 2
Uppr & Lower
Scale & Offset
More Digits
Fewer Digits
Gate & ExtArm
Limit Modes
Stats
Save & Print
Run
Stop / Single
Trigger Sensitivity
50Ω 1MΩ
DC AC
X10 Attenuate
100kHz Filter

Above — My HP counter with the kit installed as Channel 2.

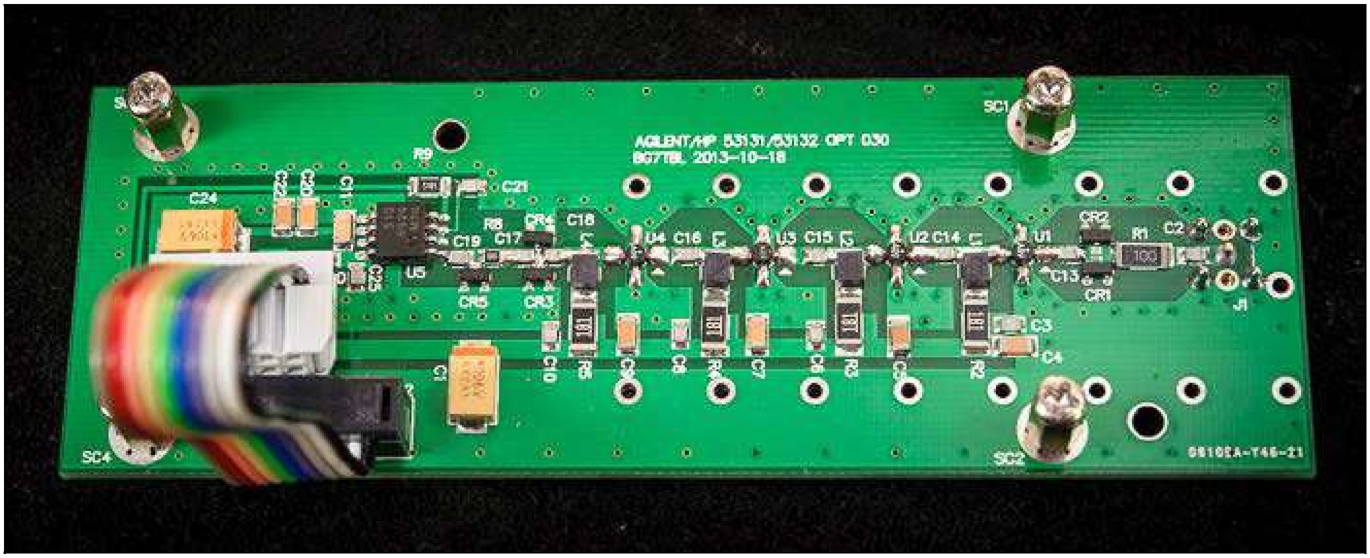
AGILENT/HP 53131/53132 OPT 030
BG7TBL 2013-10-18
SC1
SC2
SC4
U1
U2
U3
U4
U5
J1
0810EA-Y46-21

Above — Top of the circuit board. Anti-parallel diodes, 4 MMICS, passives and then even more anti-parallel diodes go before the prescaler chip.

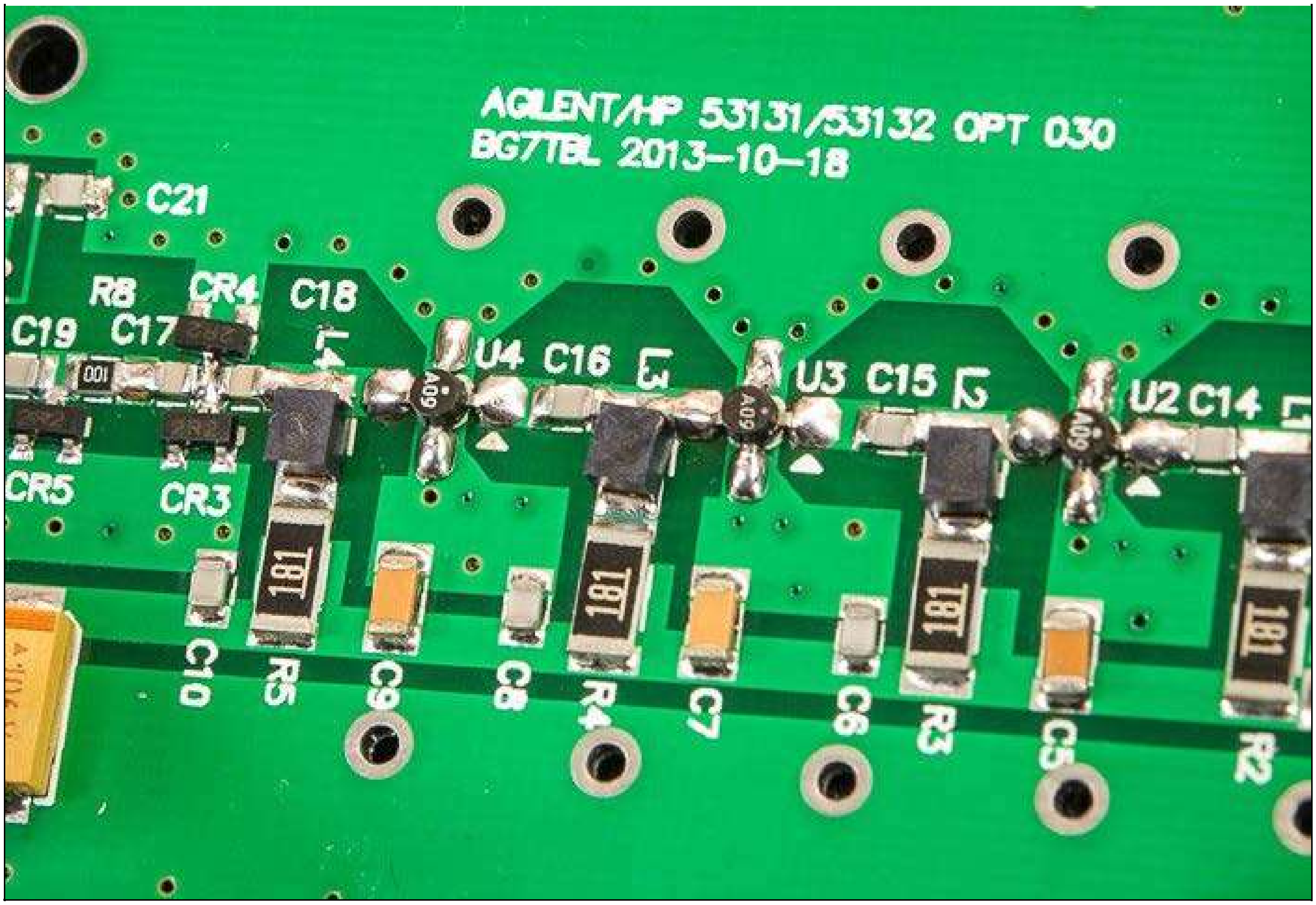
AGILENT/HP 53131/53132 OPT 030
BG7TBL 2013-10-18
C21
R8
CR4
C18
C19
C17
U4
C16
U3
C15
U2
C14
CR5
CR3
C10
R5
C9
C8
R4
C7
C6
R3
C5
R2

Above — Close up of SMD circuitry.

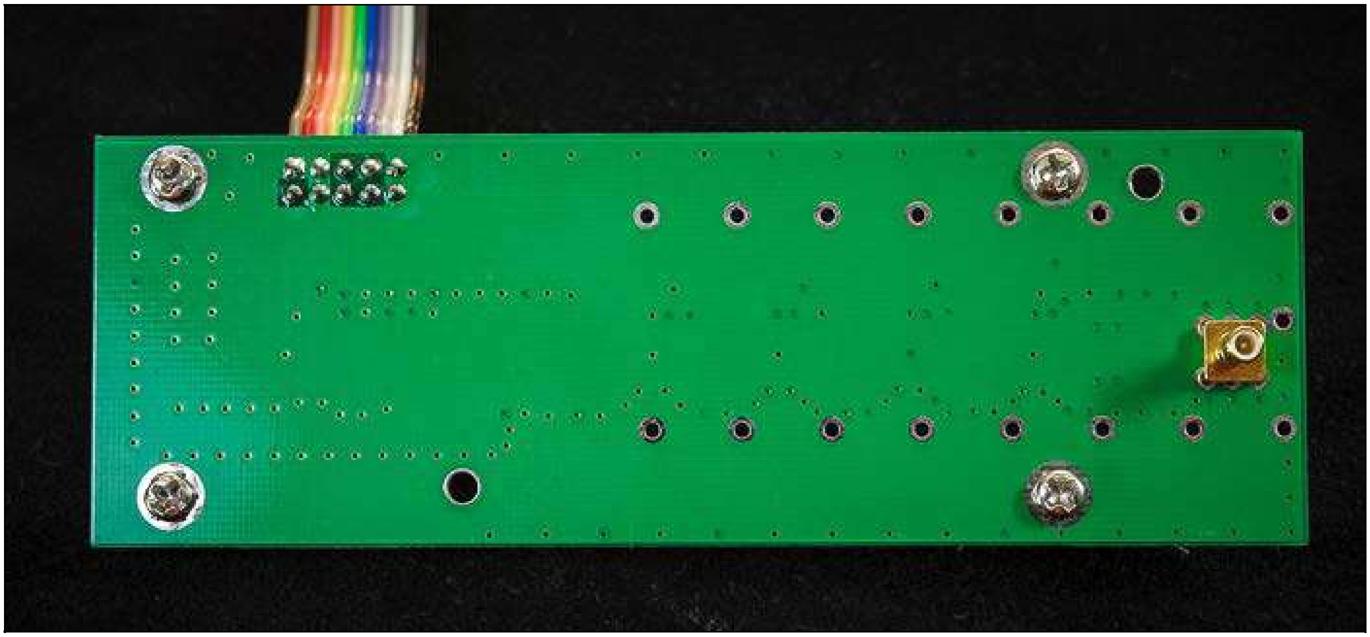

Above — Rear view of the circuit board showing stiching vias that join the top ground plane to the bottom ground plane along the signal path. My tests showed that it works okay to 3.1 GHz. Click for a picture of the installed board. I'm very happy with my HP counter's boosted bandwidth.

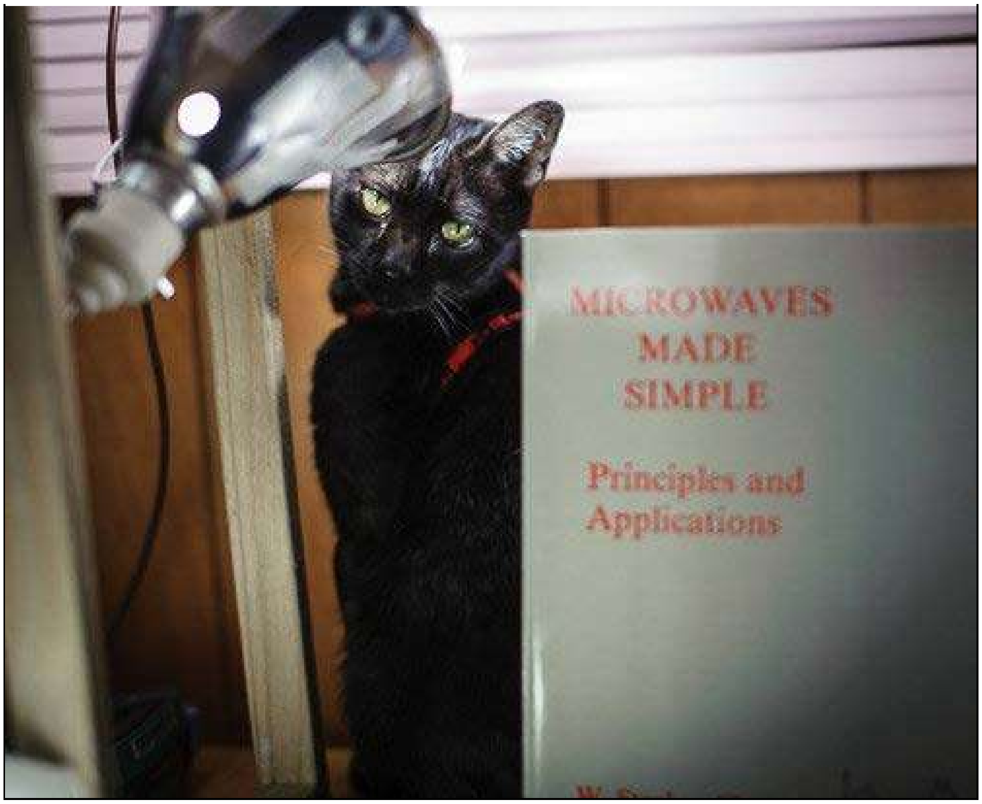
MICROWAVES
MADE
SIMPLE
Principles and
Applications

RF — Test and Measurement

Funster Line — Transmitter

Funster Line
VFO IN
FO
RF GAIN
KEY
XTAL
POWER

Menu Item 4 on the HF Embarcadero web page. Click here to go back.

The receiver will go on its own page and hopefully I'll complete and write it up this Fall.

1. Transmitter
2. References

Introduction

I began the Funster transceiver line in 1995 and show the 2014 version [now back to a CW trans-receiver on the 40M band]. Judging by reader feedback, plus analog scratch-homebrew threads in social networks, email lists and print, many of us feel a nostalgic pull towards old-school radio. Me too.

Now at version 8 — Funster shares my joy in making a simple, old-style radio with all discrete parts. In some ways, this radio set is a twisted version of Wes and Roger's Ugly Weekender from QST for August 1981.

1. Transmitter

Summary

Variable power: 128 to 3312 mW @ 13.6 VDC. Switchable 7.04 MHz xtal oscillator or external VFO input port. Solid state transmit/receive [T/R] circuit with PIN diode. External continuous and switched 12 VDC power jacks. Single-ended BJT Class C power amplifier with 67% efficiency.

Note: I state 12 VDC in my DC circuits, however, this nominal value might better be termed 12 -13.6 VDC.

Key up power current draw @ 13.6 VDC = 17 mA. Key down current @ 3W RF power with 13.6 VDC = 420 mA.

DC Circuitry

It's fun to march off the signal path and focus on DC circuitry — switching, regulating and filtering the DC voltage feeding our RF or AF stages with finesse. Trans-receivers contain DC, RF, AF and external patch cords that benefit from mindful DC line filtering + selective high-isolation amps in the signal path to prevent RF flowing where it's unwanted. You'll see heaps of decoupling with RF + AF bypass and a few strong reverse-isolation amplifers in this radio set.

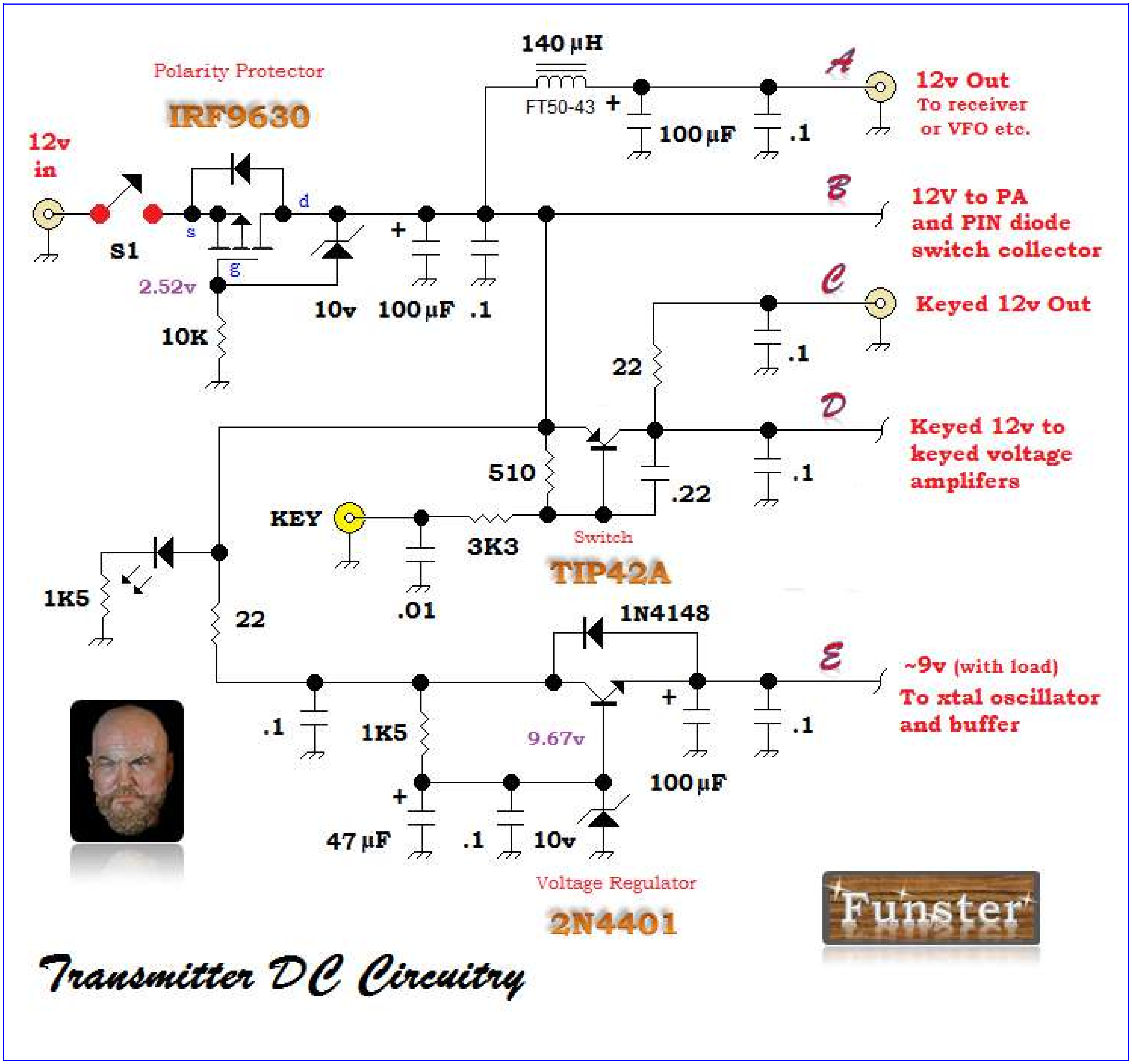
140 μH
FT50-43
Polarity Protector
IRF9630
12v in
S1
s
d
g
2.52v
10K
10v
100 μF
.1
A
12v Out
To receiver
or VFO etc.
100 μF
.1
B
12V to PA
and PIN diode
switch collector
C
Keyed 12v Out
22
.1
D
Keyed 12v to
keyed voltage
amplifers
510
.22
.1
KEY
3K3
Switch
TIP42A
.01
1K5
22
1N4148
E
~9v (with load)
To xtal oscillator
and buffer
.1
1K5
9.67v
100 μF
.1
47 μF
.1
10v
Voltage Regulator
2N4401
Funster
Transmitter DC Circuitry

Above — The DC circuit board. Click for the breadboard photo. A reader sent me 2 IRF9630 power MOSFETS: avoiding the voltage drop of the usual single rectifier diode, I employed 1 as a reverse polarity protector. Another reader sent me a long strip of 2N4401 BJTs, so I ran them instead of the equally good 2N3904. Thanks for the parts guys!

A simple series voltage regulator with a zener diode reference supplies DC to the crystal oscillator and its buffer. The base-emitter voltage drop of the 2N4401 emitter follower lays in series with the load, so load current changes alter the regulator output voltage. Vout = Vzener - VBE : so the regulated voltage is roughly 9 VDC.

Oscillator Circuitry

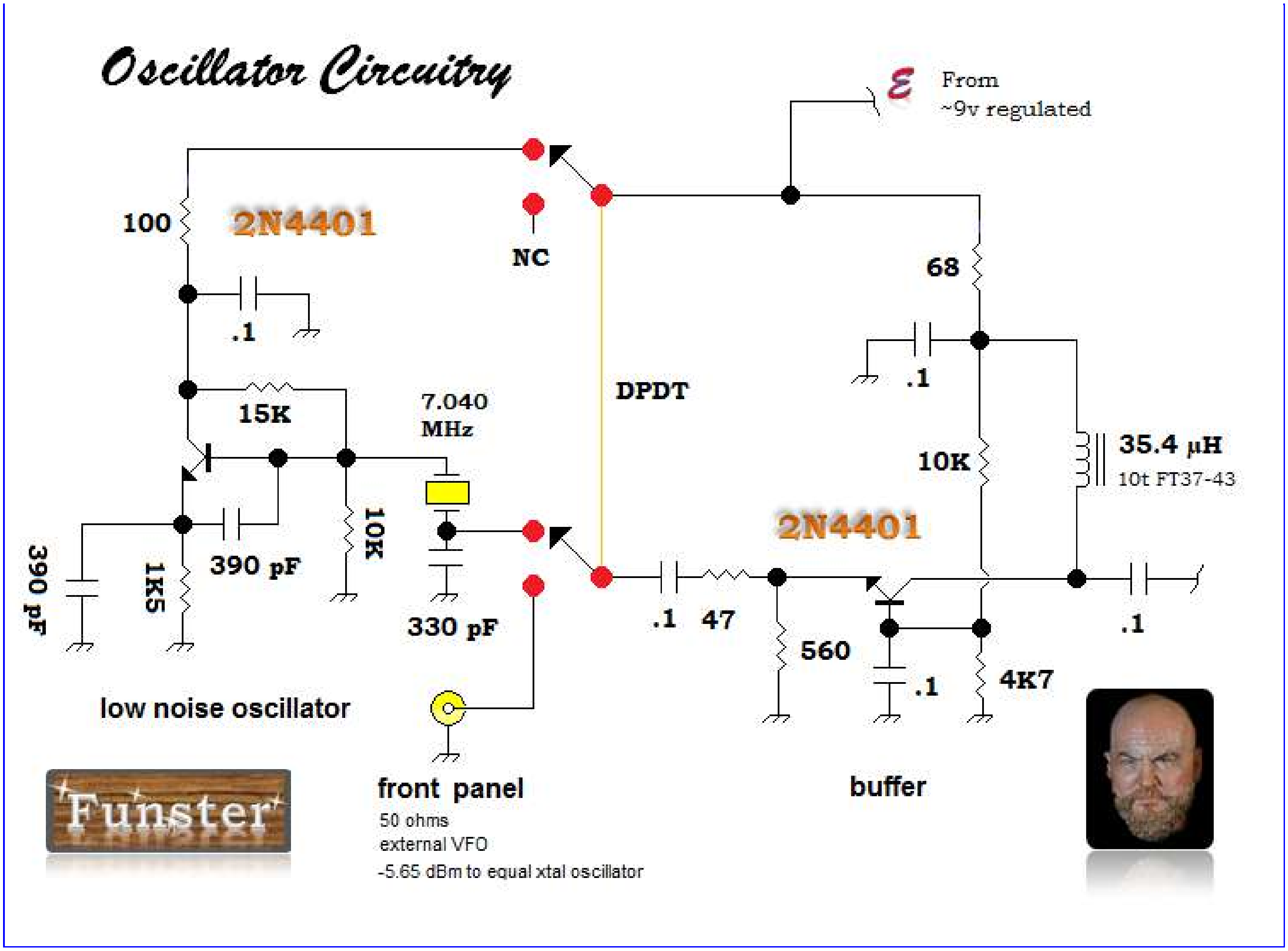
Oscillator Circuitry
From
~9v regulated
100
2N4401
NC
.1
DPDT
68
.1
15K
7.040
MHz
10K
35.4 µH
10t FT37-43
10K
2N4401
390 pF
1K5
390 pF
330 pF
.1
47
560
.1
4K7
.1
low noise oscillator
Funster
front panel
50 ohms
external VFO
-5.65 dBm to equal xtal oscillator
buffer

Above — Oscillator circuitry. I soldered my oscillator circuitry on the DC control board. Click for the breadboard. Click for a close-up of the old NorCal crystal supplied by NT7S — thanks Jason! To equilibrate to the xtal oscillator AC output voltage, the VFO input port should run about -5.65 dBm. I might stick a 6 dB attenuator pad on the front panel switch and use a 7 MHz VFO with 0 dBm output power.

Keyed Voltage Amps

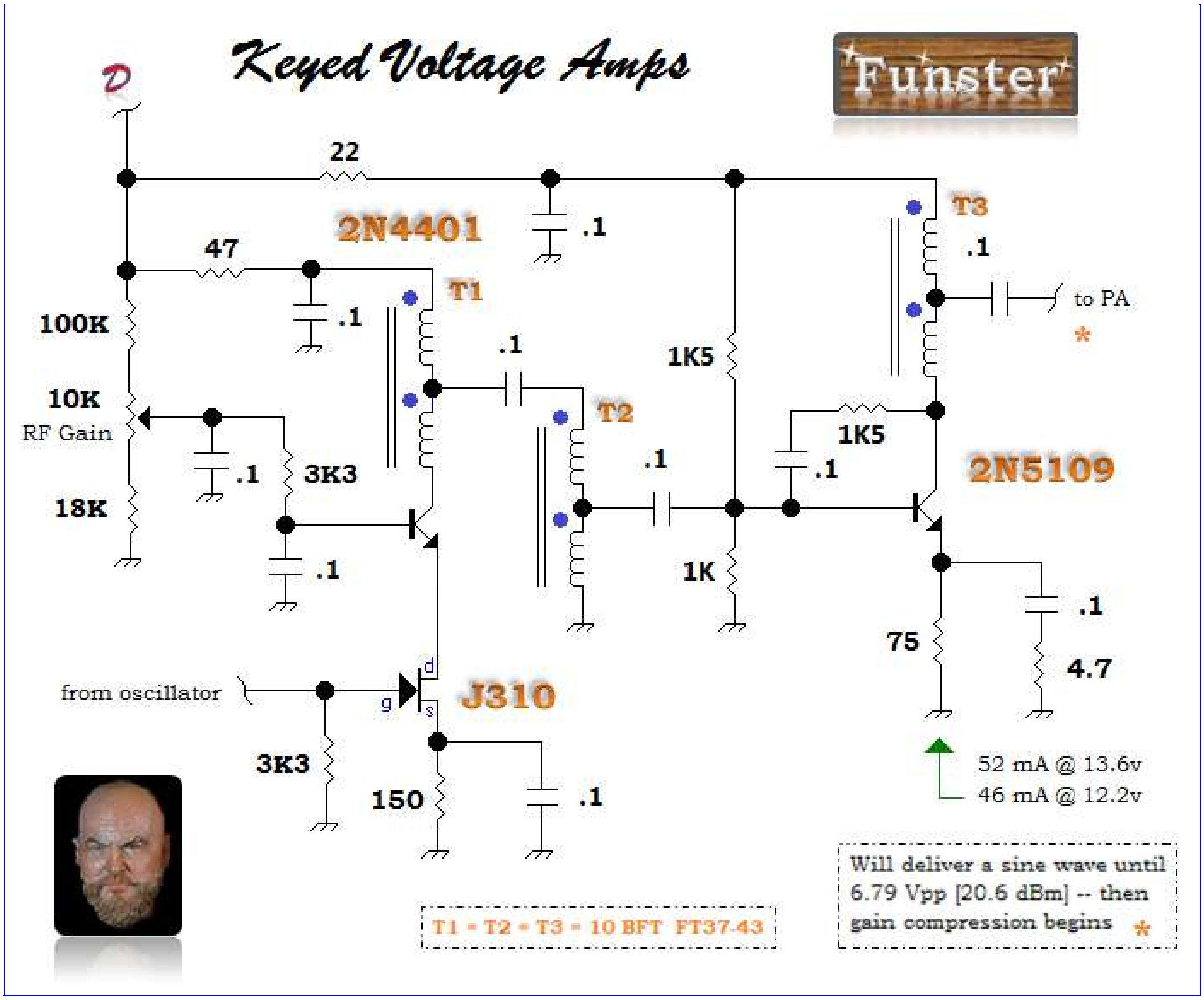
Keyed Voltage Amps
Funster
D
22
47
2N4401
.1
.1
T1
100K
10K
RF Gain
18K
.1
3K3
.1
.1
T2
.1
1K5
1K
T3
.1
to PA
*
1K5
.1
2N5109
.1
4.7
75
from oscillator
g
d
s
J310
3K3
150
.1
52 mA @ 13.6v
46 mA @ 12.2v
T1 = T2 = T3 = 10 BFT FT37-43
Will deliver a sine wave until
6.79 Vpp [20.6 dBm] -- then
gain compression begins *

Above — The 2 keyed voltage amplifers. Click for a breadboard photograph. I set the value of the RF gain pot series resistors by placing a 100K pot before and after the 10K gain pot and tweaked them back and forth to find close to 3W maximum and 125 mW minimum RF power with the gain pot set @ CW and CCW respectively. The 100K pots were removed, measured and nearest standard value resistors substituted: 100K and 18K.

By carefully choosing the 75 Ω emitter resistor, the 2N5109 feedback amp output remains a sine wave at the drive levels needed for a PA output at 3 Watts. At 3 Watts PA output, the feedback amp output voltage = 5.91 volts pk-pk [19.4 dBm]. The emitter current shown will cleanly drive a PA for more power if desired. Boost it as needed.

Power Transistor and T/R Switching

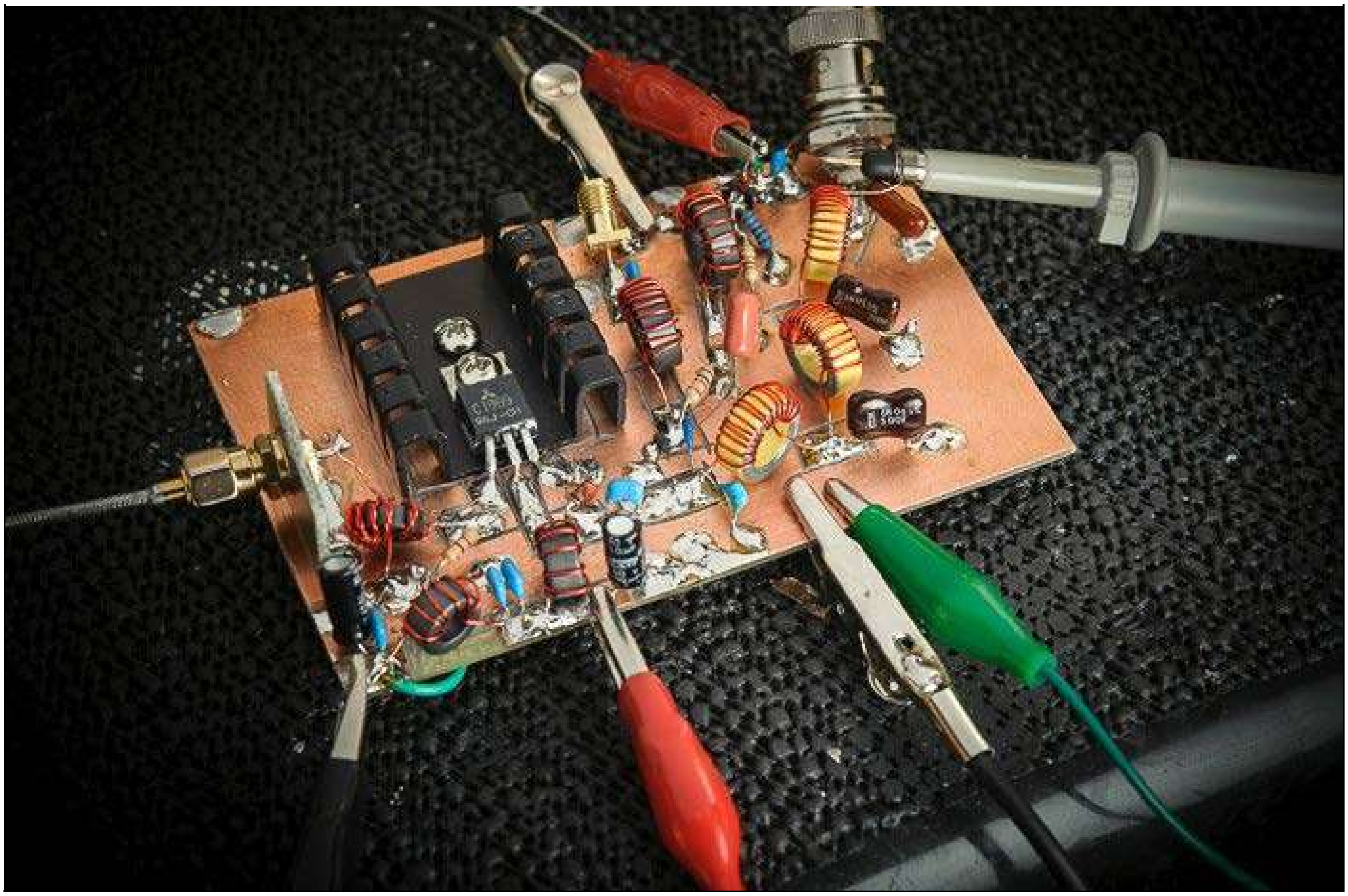

Above — Final testing the PA and T/R circuitry before installing it into the Hammond chassis. Click for a bigger photo of my test setup.

Three temporary RF connectors were soldered on this board for bench measurement. I connected the (SMA) receiver port to Channel 1 of my 50 Ω terminated 'scope to measure the RF not getting shunted to ground by the PIN diode. The SMA cable/connector to the far left is a temporary cable connected to the keyed voltage amp board output in the Tx chassis — thus I tested the PA stage with its own xtal oscillator and RF amplifers. The right-sided BNC RF port connects the PA low-pass filter to a 10W, 50 Ω dummy load via a 22 cm 50 Ω patch cable.

Further, I measured collector voltage + collector current with 2 DVMs and also PA output power in my oscillocope (DSO) Channel 2 with a 10X probe employing a tiny ground lead.

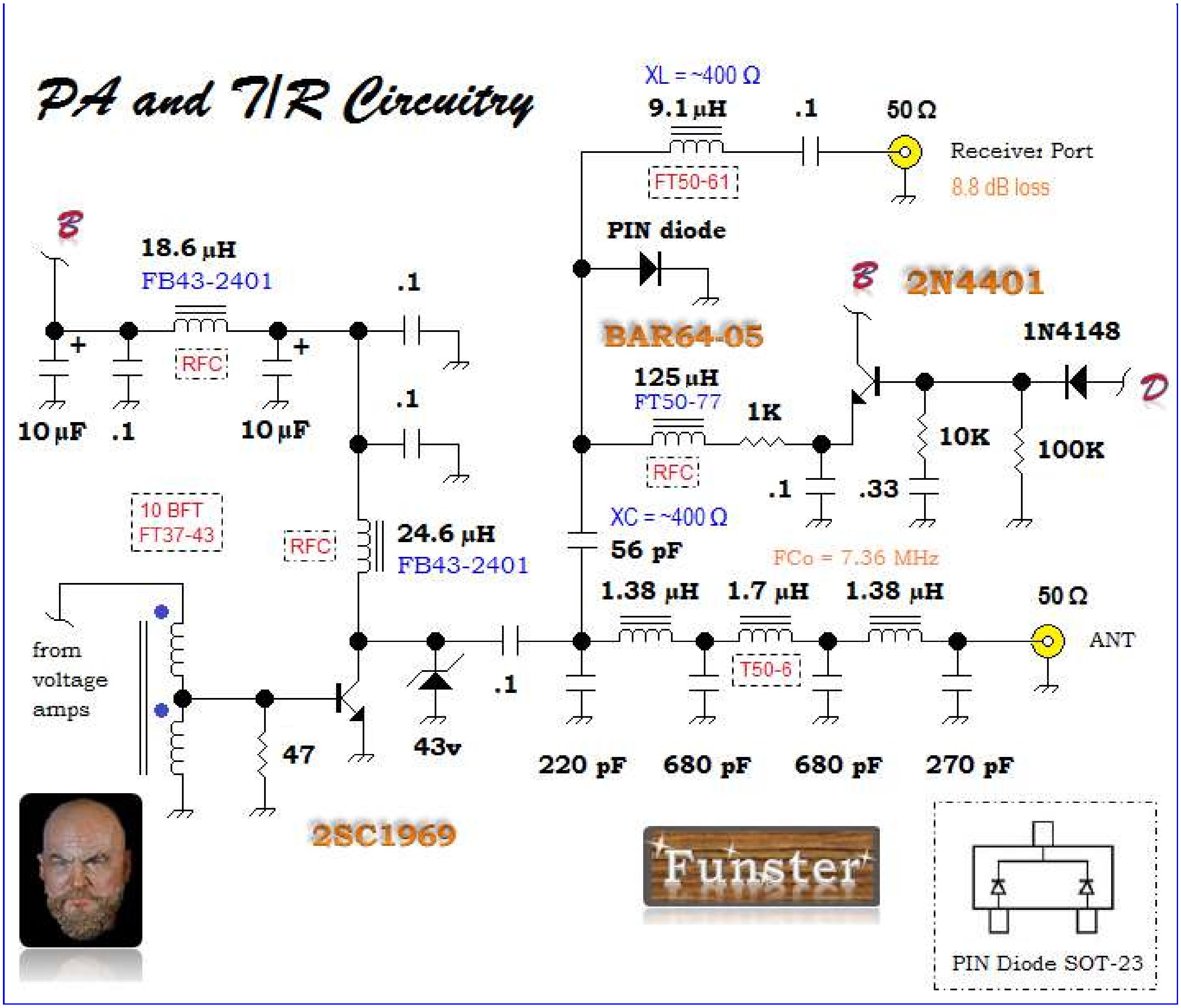
PA and T/R Circuitry
XL = ~400 Ω
9.1 µH
.1
50 Ω
Receiver Port
8.8 dB loss
FT50-61
PIN diode
B
18.6 µH
FB43-2401
.1
RFC
+
+
B
2N4401
BAR64-05
1N4148
125 µH
FT50-77
1K
D
10 µF
.1
10 µF
.1
10K
100K
RFC
.1
.33
10 BFT
FT37-43
RFC
24.6 µH
FB43-2401
XC = ~400 Ω
56 pF
FCo = 7.36 MHz
1.38 µH
1.7 µH
1.38 µH
50 Ω
ANT
T50-6
from
voltage
amps
.1
47
43v
220 pF
680 pF
680 pF
270 pF
2SC1969
Funster
PIN Diode SOT-23

Above — PA and T/R circuitry schematic. Click for another breadboard photograph taken before the 2N4401 switch was added. Although we might run a BD139 or other cheap BJT for the final, I chose my favorite HF classic, the 2SC1969. This transistor will survive a high VSWR, however a 43 volt - 500 mW zener diode further protects it from such.

In 1 version with a 1:4 Z collector matching network and no zener diode, I measured 6.25W RF power after tweaking the 100K and 18K RF gain limiting resistors in the preceeding hycas voltage amp to drive the PA harder. At ~3W maximum power in my final experiment, the 2SC1969 barely warms up.

Collector Matching

After a number of experiments, I learned that from tens of mW up to ~ 3.3 Watts RF power, we may omit collector matching and still get reasonable efficiency if you run a collector choke >= 15 µH. Thus, I made the maximum power ~ close to 3W — easy peasy.

Ask QRP enthusiasts what QRP is all about and they might extol the thrill of exacting the most kilometers per Watt — in keeping with that spirit, I love turning down my Tx power and the Funster front panel RF gain control makes it a breeze.

The PA circuit is standard fodder. I wound the 24.6 µH and 18.6 µH RFCs on FB43-2401 ferrites [a single hole bead that looks like a fat FT37-43]. An FT37-43 or FT50-43 makes a great substitute.

PA Measures

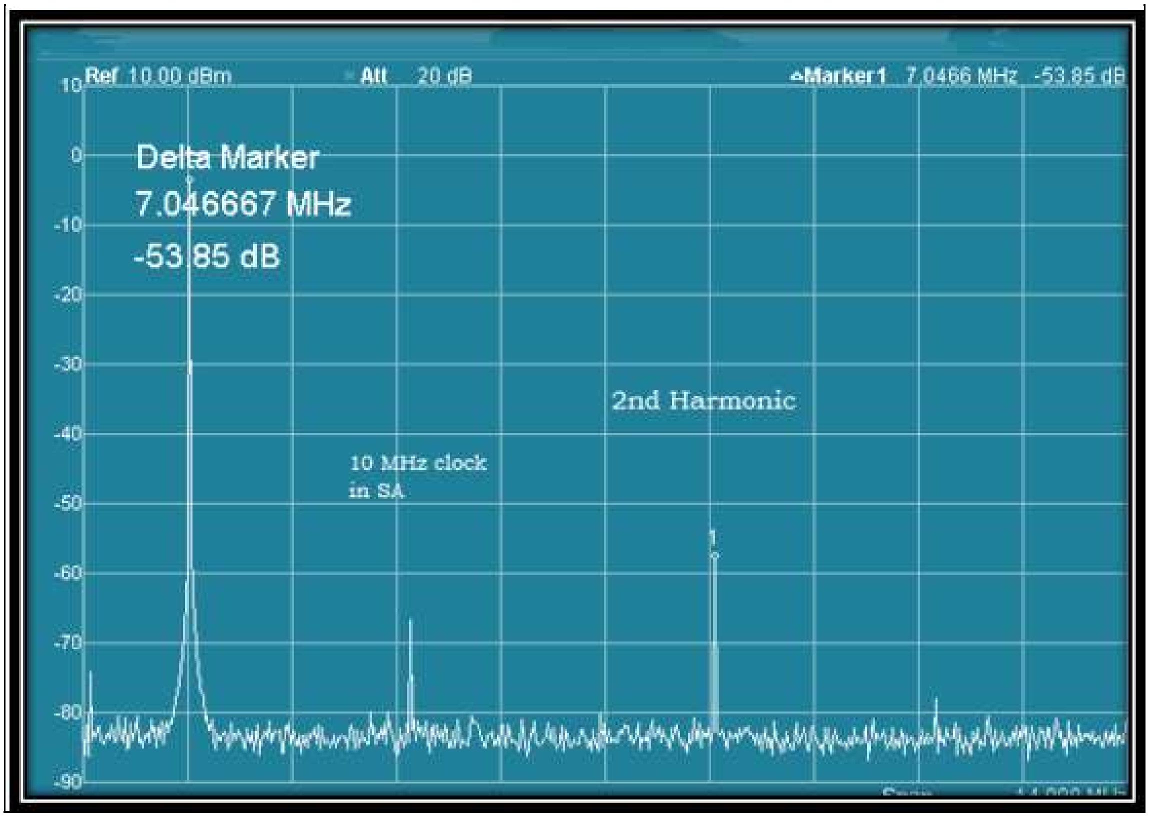
Ref 10.00 dBm
Att 20 dB
Marker1 7.0466 MHz -53.85 dB
10
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
Delta Marker
7.046667 MHz
-53.85 dB
10 MHz clock
in SA
2nd Harmonic
1

Above — Spectrum analysis of the Tx output. The 10 MHz clock signal from my spectrum analyzer lays between the fundamental and 2nd harmonic. The 2nd harmonic = -53.85 dBc @ 3W power output.

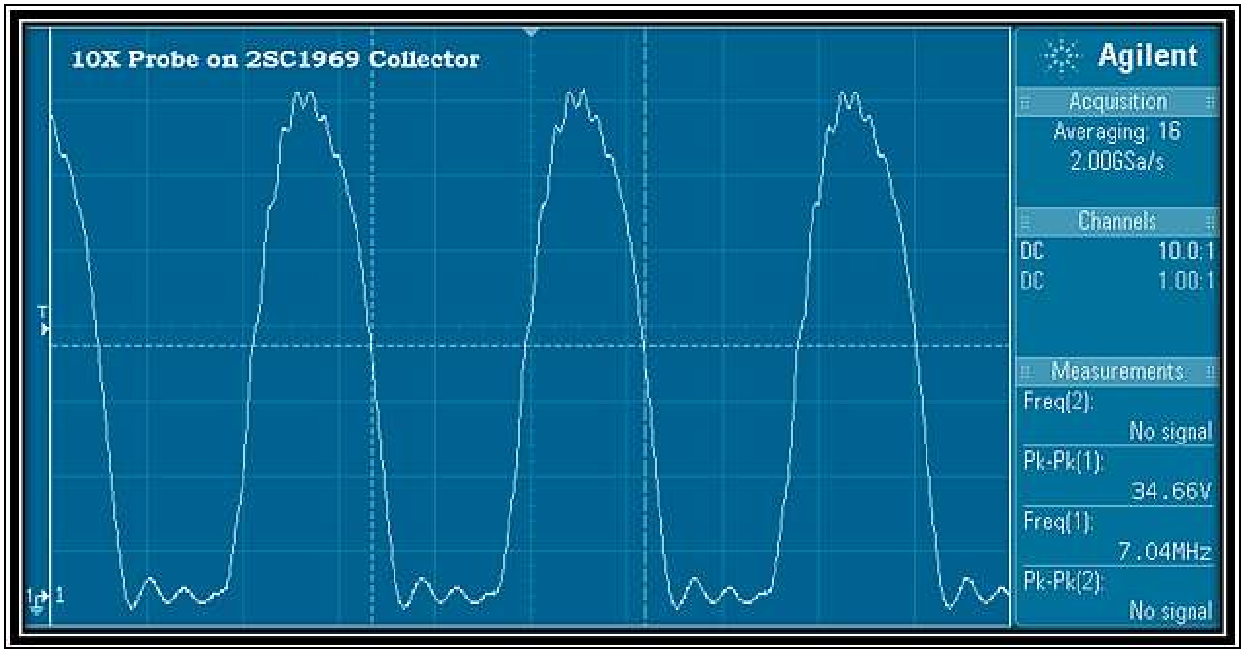
10X Probe on 2SC1969 Collector
Agilent
T
1

Acquisition
Averaging: 16
2.00GSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(2): No signal
Pk-Pk(1): 34.66V
Freq(1): 7.04MHz
Pk-Pk(2): No signal

Above — PA collector waveform at 3W output measured with my DSO's 10X probe.

VICHY
VC99
APD
DC
AUTO
0.331
A
#4320
13.58
DCV
OFF
ACV
REL
HOLD
RANGE
Hz
°C
DC/AC
hFE

Above — Set your meters on stun! DC measures to calculate PA efficiency [See RF Workbench 6]: DC power = 13.58V x 0.331A = 4.49W. RF power = 3W. Efficiency = RF power / DC power = 3W / 4.49W = 66.8%.

T/R Discussion for Single Band Transceivers

Previous Funsters ran a simple QSK T/R circuit developed by Wes, W7ZOI in the early 1980s: a series X = 500 Ω resonator with anti-parallel diodes between the L and C. The series capacitor presents 500 Ω reactance to the 50 Ω transmit line so that only a small portion of the transmitter power flows to ground via the diodes and even less reaches the receiver input. In short, the 500 Ω capacitive reactance + clamping action of the anti-parallel diodes protect the receiver front end.

For 1 band, this simple circuit sizzles. Two potential concerns:

~6-9 dB loss of the receive signal making it better for lower HF, although I've seen builders add a receive preamp and use this T/R system on 12 and 10 meters. Since both band noise and signals get attenuated simultaneously, I've rarely had readability issues on 160-30 meters in my popcorn transceivers.

The anti-parallel diodes may cause IMD. The intercept for the 500 Ohm series reactances with shunt diodes is around +6 or +7 dBm. This might exert problems in a high performance receiver front end, but poses a non-issue with typical NE612 front ends. Many builders, including Wes, reduce IMD potential with 2 or 3 series diodes on each leg. Another solution includes running a PIN diode as the shunt element. That's what I did.

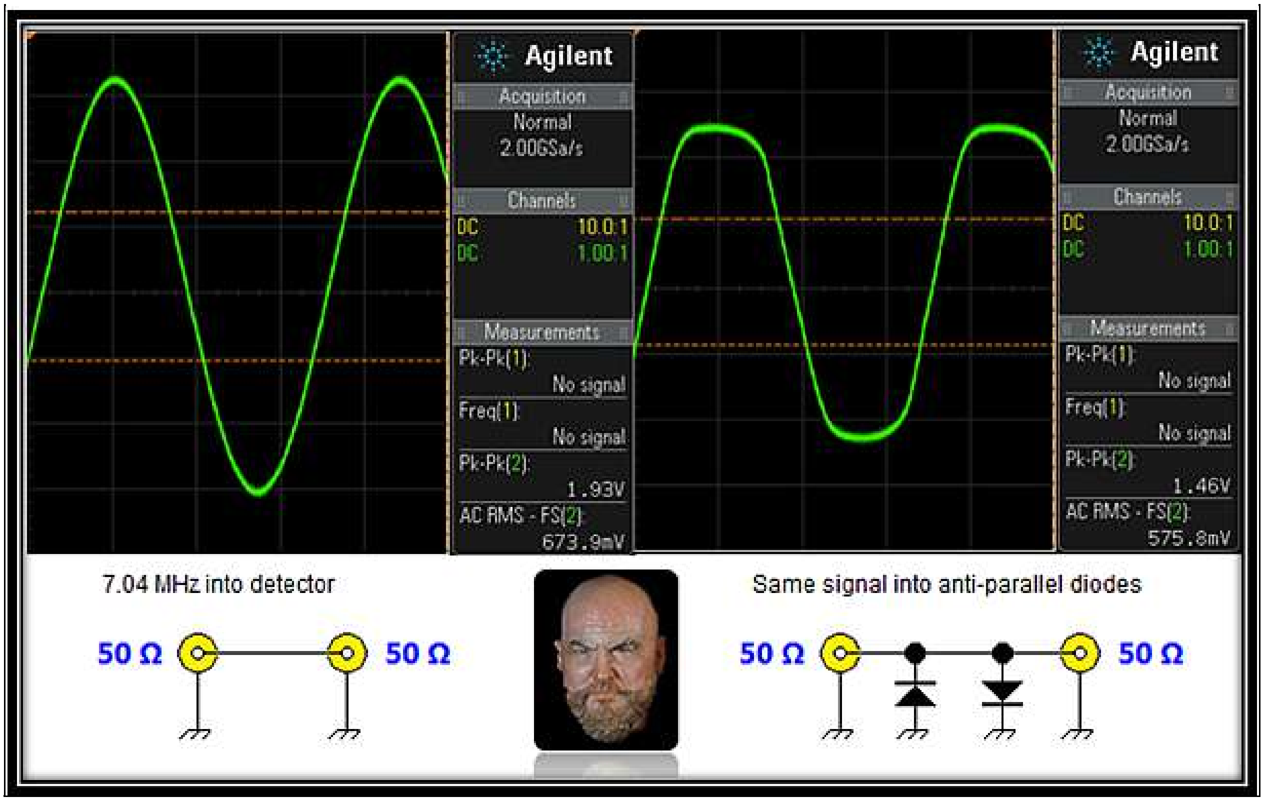
Agilent
Acquisition
Normal
2.00GSa/s
Channels
DC 10.0:1
DC 1.00:1
Measurements
Pk-Pk(1): No signal
Freq(1): No signal
Pk-Pk(2): 1.93V
AC RMS - FS(2): 673.9mV
Agilent
Acquisition
Normal
2.00GSa/s
Channels
DC 10.0:1
DC 1.00:1
Measurements
Pk-Pk(1): No signal
Freq(1): No signal
Pk-Pk(2): 1.46V
AC RMS - FS(2): 575.8mV
7.04 MHz into detector
50 Ω
50 Ω
Same signal into anti-parallel diodes
50 Ω
50 Ω

Above — The DSO tracing of an experiment showing the before and after effect of 2 anti-parallel diodes. The voltage across a silicon diode varies with the current going through it. Silicon diodes don't conduct much current until the voltage across them reaches a threshold [typically ~0.6 to 0.7 volts].

In situ, the forward threshold voltage of each diode limits the circuit voltage by clipping off AC voltage above its threshold. Clamping refers to the limiting of voltage in *2* directions using anti-parallel pair(s) of clipping diodes. For example, if a negative-going AC dips more than 1 diode drop below ground, the left hand diode will begin to conduct and clip off any voltage above its threshold. This, of course, describes the "ideal diode" — in real life, factors like switching speed, current-handling and temperature are all in play.

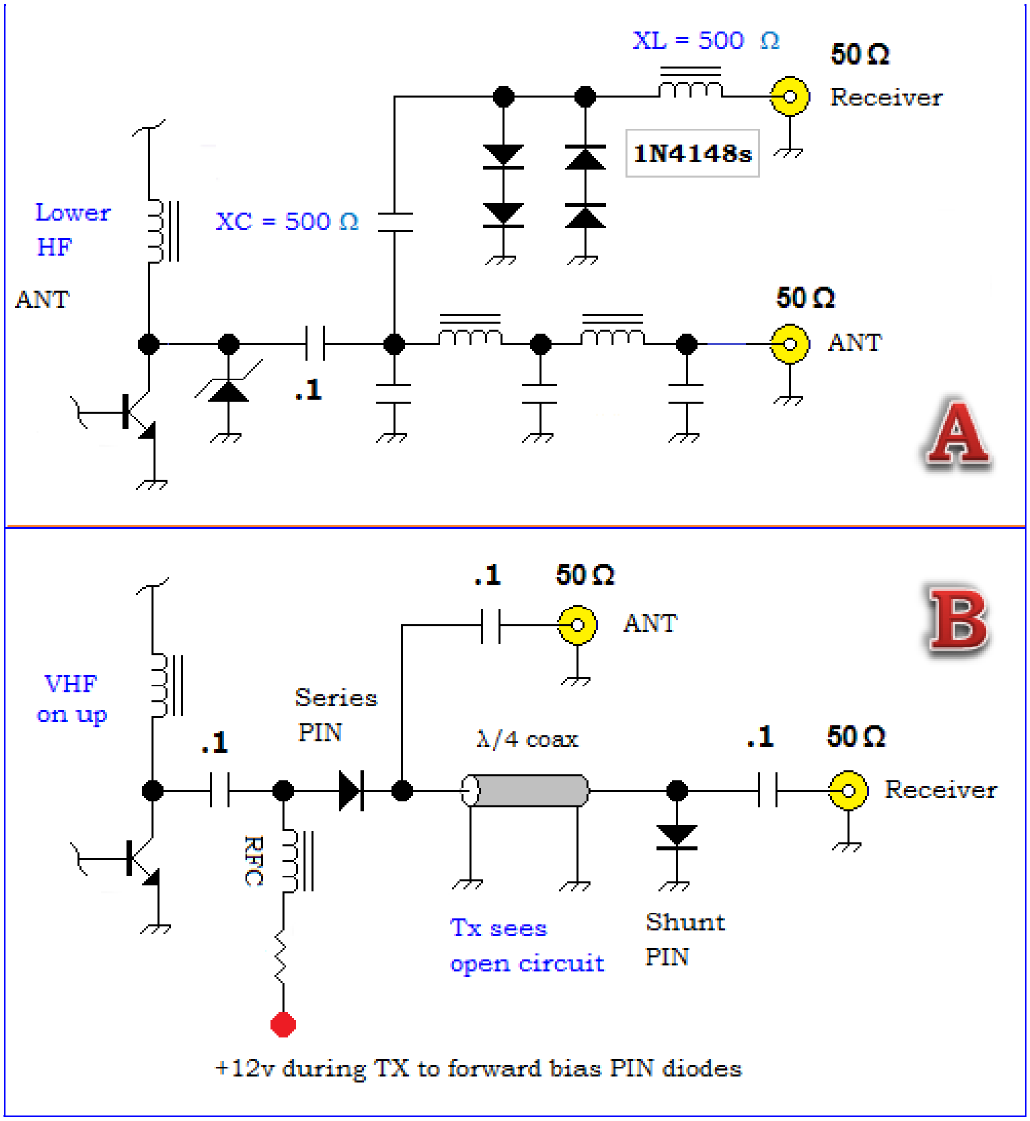

Above — 2 possible T/R schemes. Figure A shows the classic W7ZOI T/R system with 2 diodes per leg. See EMRFD Fig 2.112, for an example with 3 diodes per leg. For 1 band, it's often our go-to circuit.

Figure B shows a scheme well suited for VHF on up: Both PIN diodes get forward biased during transmit mode, however, a 1/4 wave transmission line lays between the series and shunt PIN diode. 1/4 wave transmission lines exert magic [actually it's pure physics that may be verified with a Smith chart]. When the shunt PIN diode gets biased *on* and grounds the receiver port end of the transmission line, the other end looks like an open circuit and thus power flows to the antenna port and not into the forward biased shunt PIN diode to ground. I hope to apply this system in future VHF and UHF projects and sadly, it's impracticable at HF.

Funster Version 8 T/R

As it happens, I seek experience with PIN diodes and therefore shunted a PIN diode instead of running anti-parallel diodes in the classic W7ZOI T/R scheme. This meant adding a transistor switch to quickly bias on the PIN diode @ *key down* and run a little hang time @ *key up*.

Please refer to the PA-T/R schematic. During key down, keyed 12 volts saturates the 2N4401 switch and 11 ma current biases *on* the PIN diode. At key up, the BJT switch's base voltage is held on by the 10K plus series 0.33 μF capacitor just long enough to allow the stored energy in the large PA collector choke to collapse — protecting the receiver.

Feel free to experiment with the timing capacitor value. Previously, Wes, W7ZOI suggested connecting the T/R receiver port to a 50 Ω teminated 'scope for general measurement and also for clocking up RC time constants in DC switching circuits. With a 2 channel 'scope you can monitor the receiver port switch function in 1 channel while watching the PA collector AC voltage in the other. Now that's experimenting! Click for a sample from my QRP workbench.

I chose an XC / XL of 400 Ω for the series L-C network. The 9.1 μH coil shouldn't be wound on a 43 material ferrite [the Qu is just too low]: a bigger powdered iron toroid like size 68, or even a 61 ferrite material toroid works okay from my experiments [Q's as high as 162 were measured on a FT50-61 @ 1 MHz in 2013]. Although this coil tunes coarsely, try tweaking the 9.1 μH inductor with thumb and forefinger during actual signal testing in a receiver.

I wound the 125 µH choke connected to the pin diode cathode on a 77 material ferrite toroid to get it done with just a few turns and avoid a self resonant frequency close to 7 MHz.

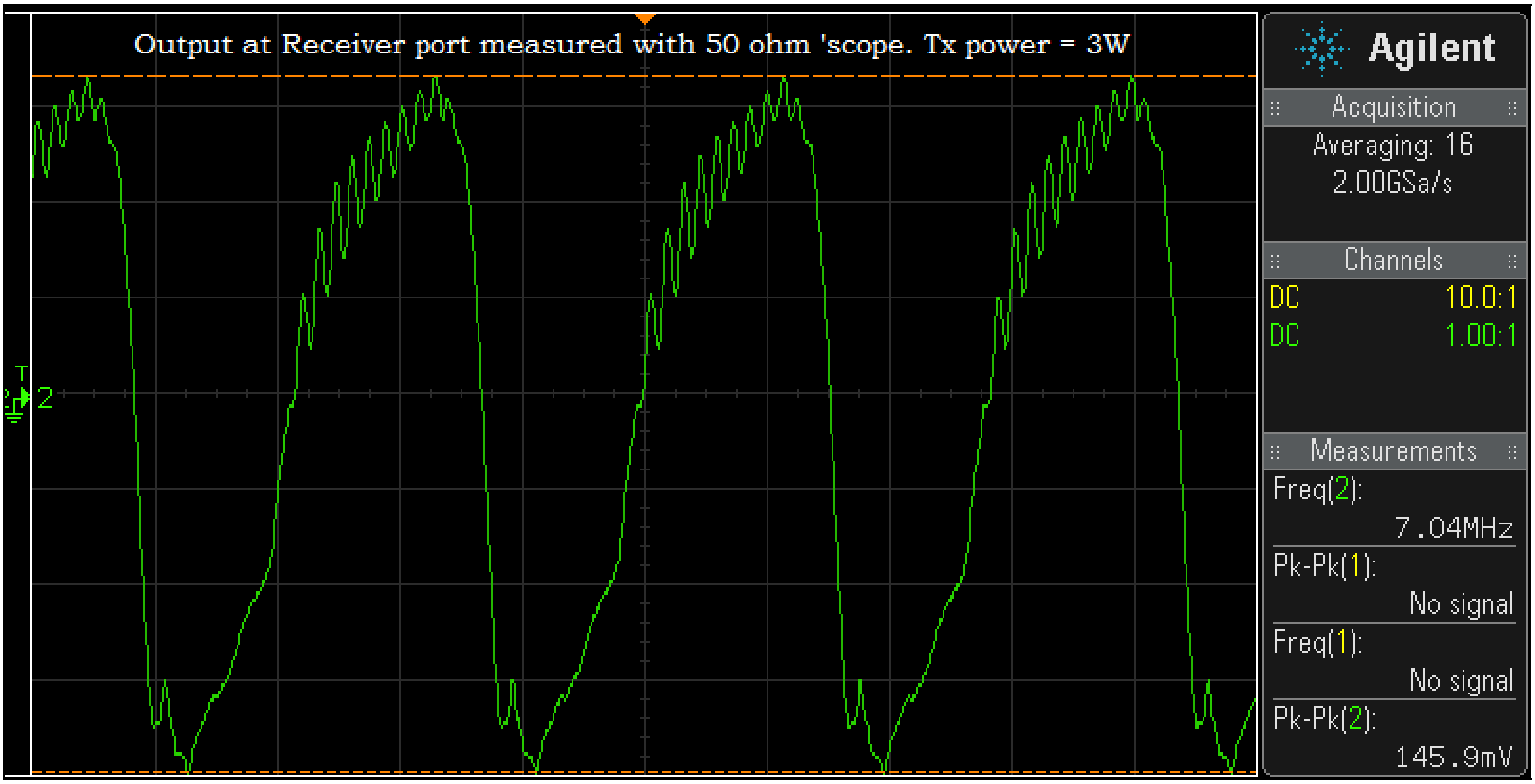
Output at Receiver port measured with 50 ohm 'scope. Tx power = 3W

Acquisition
Averaging: 16
2.00GSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(2): 7.04MHz
Pk-Pk(1): No signal
Freq(1): No signal
Pk-Pk(2): 145.9mV

Above — 50 Ω terminated DSO measurement of the receiver port @ 3W transmitter power. Theoretically all power should shunt to ground through the PIN diode, however, the diode exhibits a little resistance and the series L-C circuit contains reactance, so isolation is never perfect.

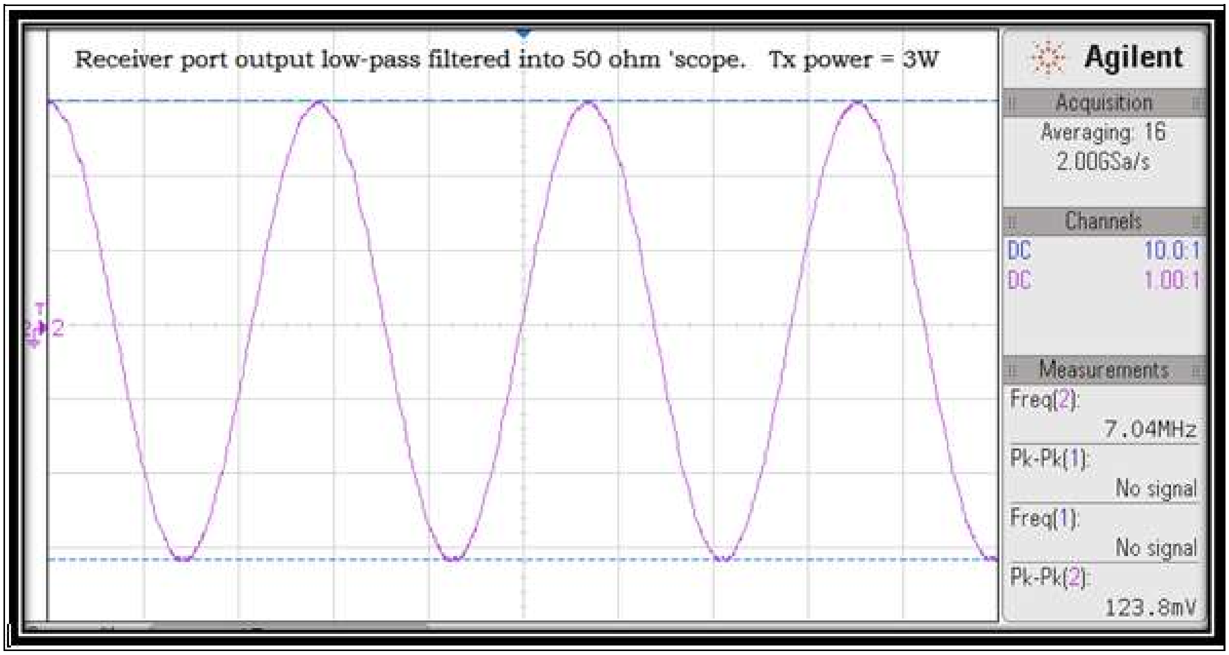
Receiver port output low-pass filtered into 50 ohm 'scope. Tx power = 3W

Acquisition
Averaging: 16
2.00GSa/s

Channels
DC 10.0:1
DC 1.00:1

Measurements
Freq(2): 7.04MHz
Pk-Pk(1): No signal
Freq(1): No signal
Pk-Pk(2): 123.8mV

Above — A low-pass filtered measure of the receiver port while in transmit mode. The beauty of a 50 Ω measurement environment shines: I placed a 7.36 MHz-bench module low-pass filter between the receive port and my DSO to quantify the leaked power to the receiver port as -14.17 dBm when the transmitter delivers 3W. Safe for my receiver.

My sincere thanks to those who offered advice and feedback during the build of this transmitter.

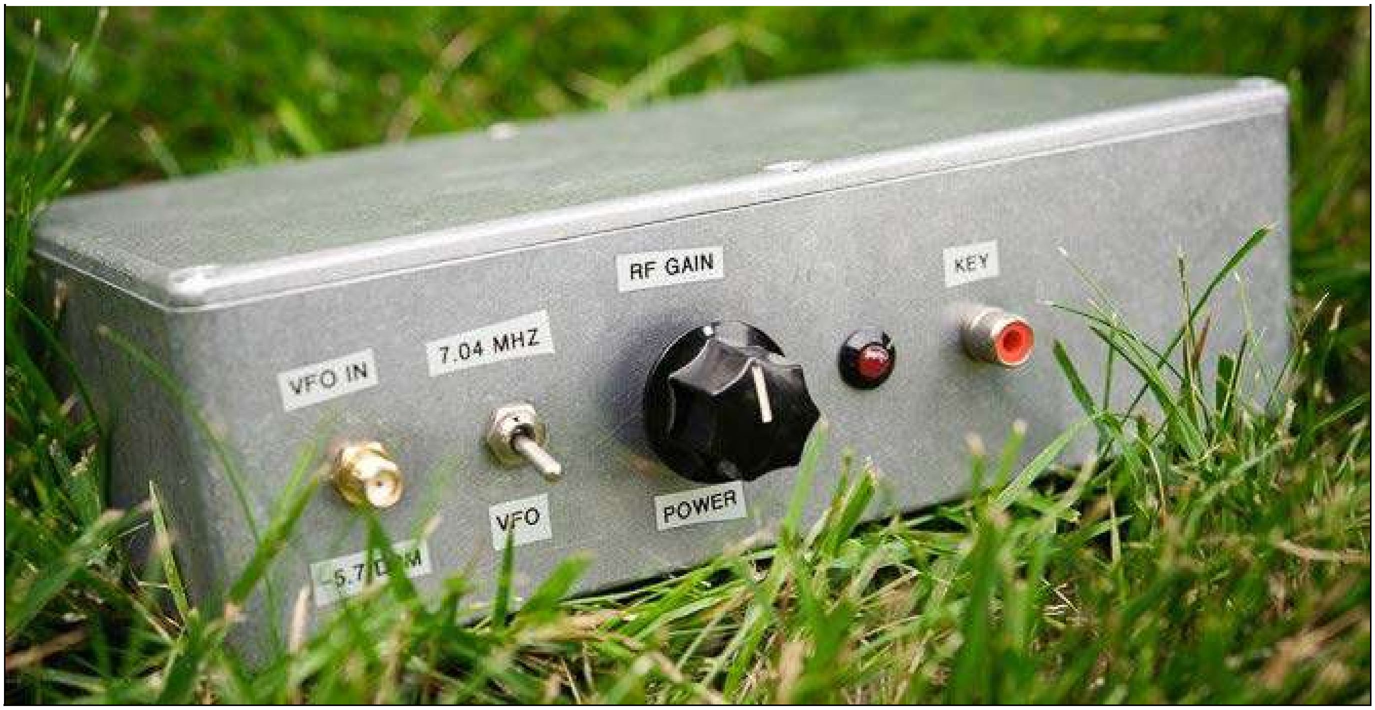
RF GAIN
KEY
7.04 MHZ
VFO IN
VFO
POWER
-5.7 DBM

Above — Funster in the field.

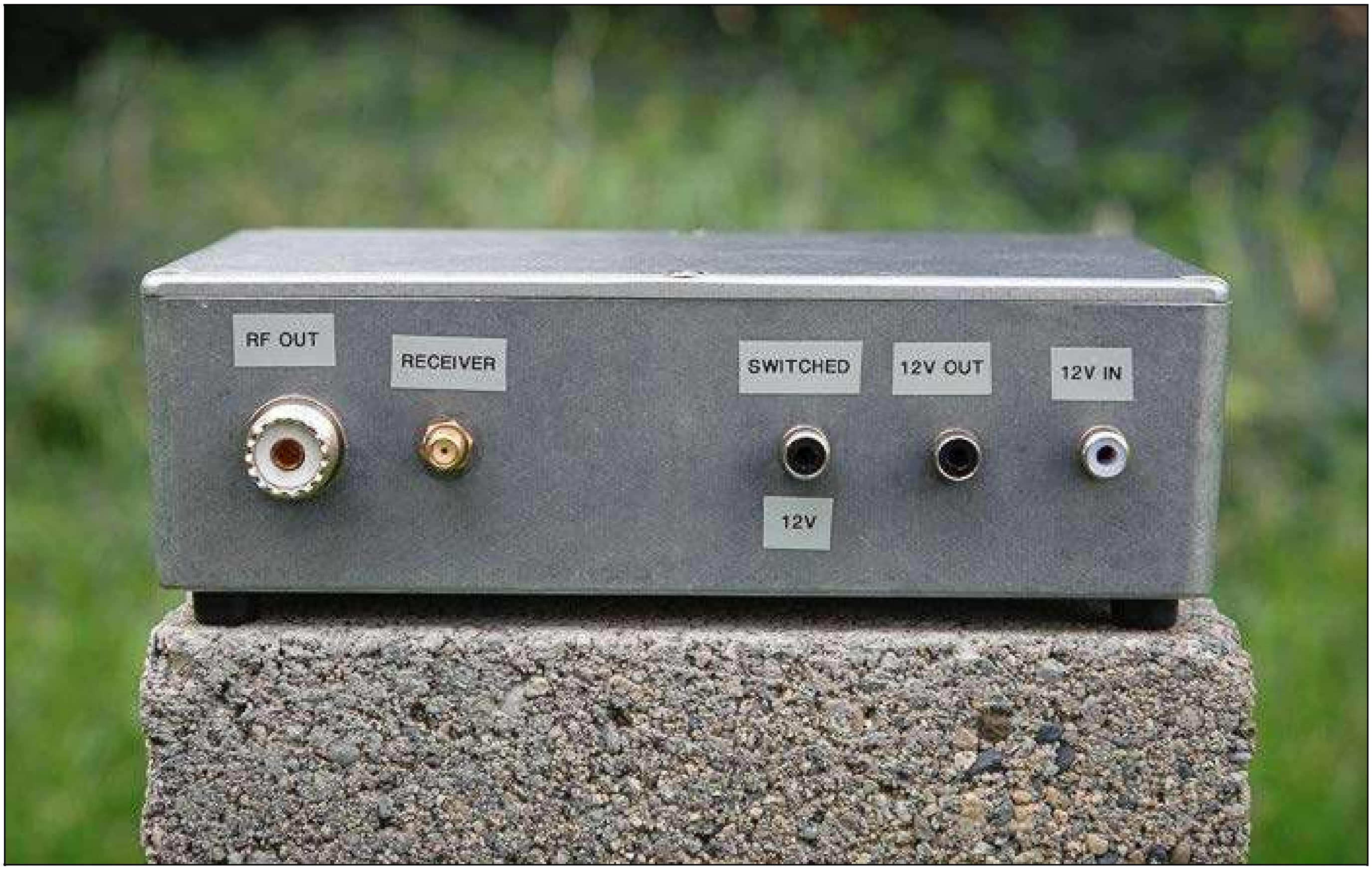
RF OUT
RECEIVER
SWITCHED
12V
12V OUT
12V IN

Above — Rear view of the Funster Tx.

VFO IN
VFO
XTAL
RF GAIN
POWER
KEY

---

2. References

T/R References

Perfecting a QSK System. QEX for Mar/Apr 2006. Markus, VE7CA. Click for pdf.

Electronic Switching and How it Works, QST for September 1984. Doug, W1FB

7 MHz receiver -- discussion begins on page 215 -- On page 214 in Fig 1 see the 33 pF input cap plus receiver protecting anti-parallel diodes. Also view the 555 timer and other T/R control circuitry in Figure 2 on page 215. *This receiver realizes the first-ever printed version of the W7ZOI T/R system described earlier.* Solid State Design for the Radio Amateur. ARRL,1986 [out of print] by Wes, W7ZOI and Doug W1FB.

Transmit - Receive Antenna Switching. EMRFD. First published by the ARRL in 2003. Page 6.68.

Schematic for the 30-W, 7 MHz power amplifier. Figure 2.112 [page 2-42] EMRFD. Design by Wes, W7ZOI

RF — Test and Measurement

VHF FM — Supplement 1

This page supplements VHF FM

1. 10.7 MHz IF Strip - with example broadcast band FM receiver
2. NOAA Receiver Experiments @ 162.55 MHz

---

1. 10.7 MHz IF Strip

I'm inexperienced with making FM receivers and lack confidence. I began with a utility module to allow basic assessment of the circuit gain required for different detectors, ceramic filter bandwidths, and more so — to get something working to boost my confidence.

The FM receiver 10.7 MHz IF standard began soon after World War 2 and 10.7 MHz still makes sense for the primitive radio builder — surplus IF filters still seem available in a variety of bandwidths. Although vexxing, if we suffer from receiver images at 10.7 MHz, we can always filter the input better, or jump up the IF and make a crystal filter for say 20 MHz. Above 20 MHz good, inexpensive, fundamental-mode crystals get hard to find.

Mixer

I chose a diode ring mixer; again. After trying other mixers such as balanced FETs, or an NE612 , I kept coming back to the diode ring. Partly I'm biased and partly it's from from measures like this one. Although the oscillators weren't perfectly set for a 10.7 MHz IF, the LO and RF signals are both > 40 dB down from the desired IF — love this.

Since I designed a 50 Ω bench module, the IF output contains a source follower to drive the IF port. In an actual "keeper" receiver, you probably wouldn't bother since you might just directly drive the detector, or perhaps another 330 Ω input/output IF amp before the detector.

Although I humbly build simple/primitive circuits, my design goal is never for a *low parts count* — each part serves its purpose. A low parts count ideology plus minimalism exploits the frugal and primitive homebrew nature of our brother hobbyists, however, often gets out of hand. Some builders just cite a low parts count *design goal* to rationalize a poor design. Worse yet, inexperienced minimalists often exclude vital parts such as DC supply decoupling and bypassing, temperature stablizing bias networks, or adequate band-pass filtering. To each his own. Please discern all you see or read in homebrew radio; we're builders, not lemmings.

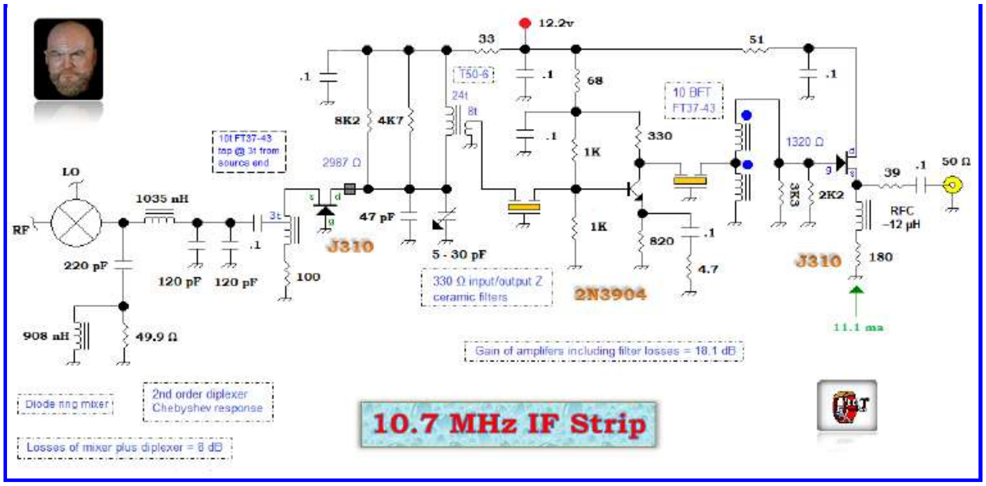
12.2v
33
51
.1
.1
.1
68
T50-6
24t
8t
8K2
4K7
10 BFT
FT37-43
10t FT37-43
tap @ 3t from
source end
2987 Ω
330
1K
1320 Ω
LO
1035 nH
RF
3t
s
d
g
47 pF
J310
5 - 30 pF
1K
820
.1
4.7
3K3
2K2
g
d
s
39
.1
50 Ω
RFC
~12 µH
180
J310
220 pF
120 pF
120 pF
.1
100
330 Ω input/output Z
ceramic filters
2N3904
11.1 ma
908 nH
49.9 Ω
Gain of amplifers including filter losses = 18.1 dB
Diode ring mixer
2nd order diplexer
Chebyshev response
10.7 MHz IF Strip
Losses of mixer plus diplexer = 8 dB

Above — Click on the IF strip schematic to view it. Most all the stages were discussed on VHF-FM. My net module gain = 10.1 dB. The 2 ceramic IF filters = 230 KHz Muratas with a measured average loss of 6.15 dB each.

10.7 MHZ IF
STRIP
RF
LO

Above — The boxed up circuit contains 3 SMA jacks. Click for a photo of the completed breadboard.

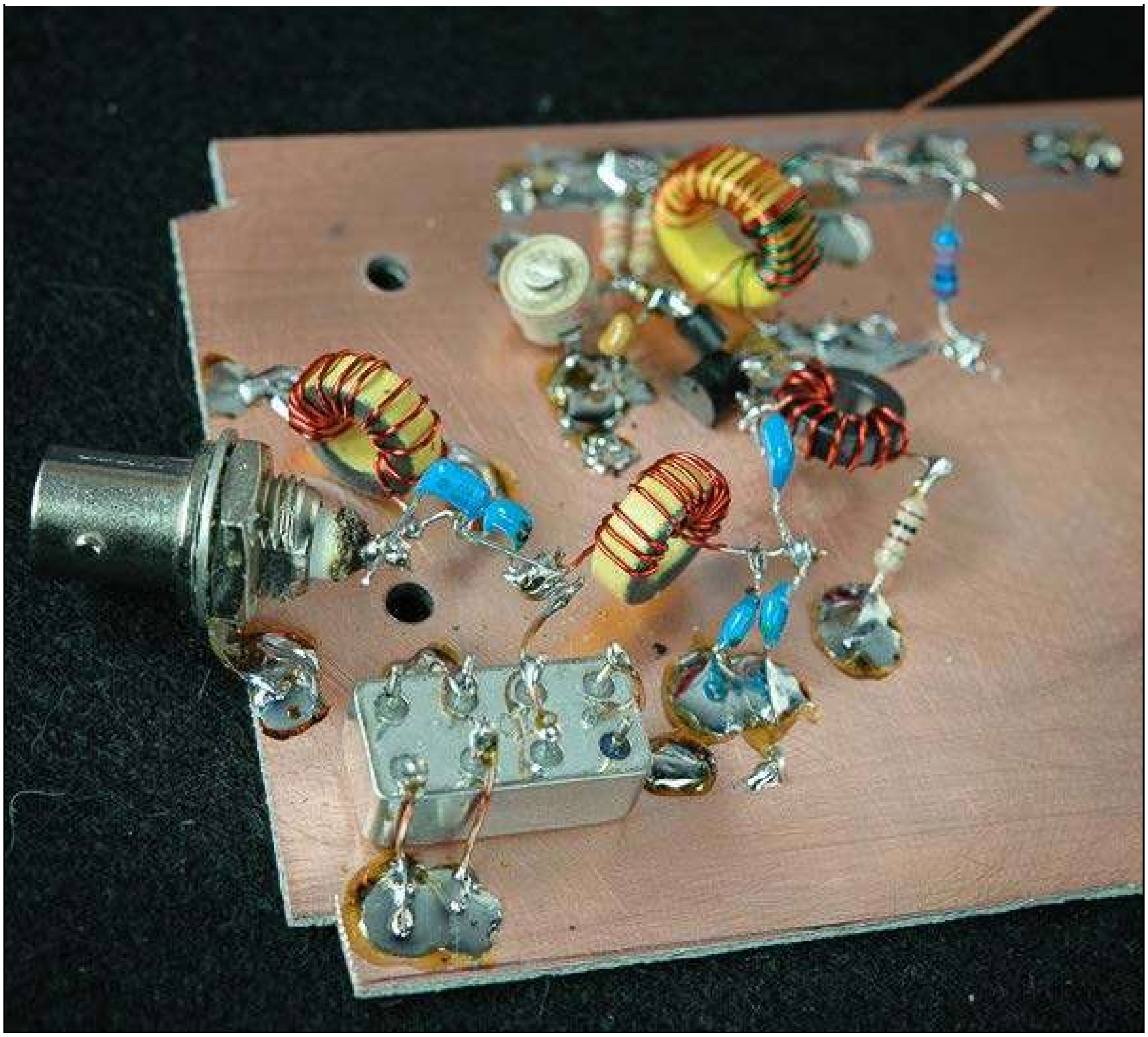

Above — A close up of the mixer, diplexer and JFET amp, although the SBL-1 mixer is disconnected. This board contains a temporary BNC jack to allow testing of the non-mixer parts such as JFET gain, diplexer loss, return loss, to confirm the tuning cap choice and also to look for any UHF oscillations eminating from the FET, a UHF part. Click to zoom in on the common gate amplifer with its gate lead soldered as close to the plastic case as possible.

Detector

I tried to homebrew my first IF amp +/- limiter and detectors , but felt unhappy with their performance, failures, circuit complexity and/or noise issues. Click click or click for 3 raw, miserable design examples from Winter 2014. Discouraged; after 2 months of lamenting, I opted for an IC detector subsystem. 3 relic part choices might include the CA3089, SA604A or LA1150.

In my case, the CA3089 suited me best and offers a lot: a double-balanced quadrature FM detector with limiter + IF ampifiers; squelch, AFC, AGC, AF preamp and the potential for high-fidelity FM reception. All that's required is an outboard detector resonator — easy to build with a powered iron toroid plus a fixed + trimmer cap.
Price + shipping varies widely, but last year, I got 5 shipped to me from Europe for $7.25 USD and these will make 5 receivers — not bad at all. What hobby isn't expensive?

I've even read that some builders used the CA3089 for an S-meter amplifer, although its logarithmic compliance is far from stellar. You'll need to download the CA3089 datasheet.

Let's build a receiver:

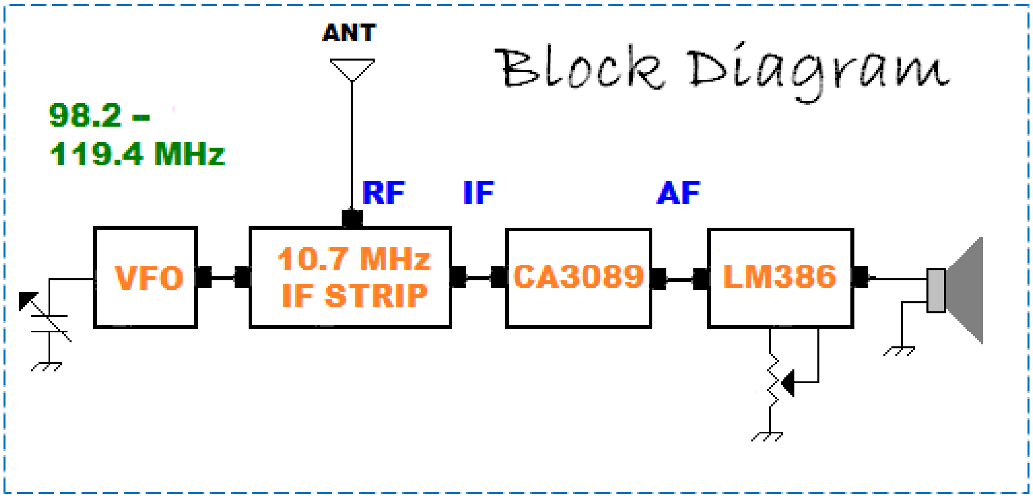
Block Diagram
ANT
98.2 – 119.4 MHz
RF
IF
AF
VFO
10.7 MHz IF STRIP
CA3089
LM386

Above — The block diagram. The antenna = 1 meter piece of wire soldered to a SMA jack that was threaded into a brass sheet to fashion a crude ground plane. No LNA or input band-pass filtering was employed.

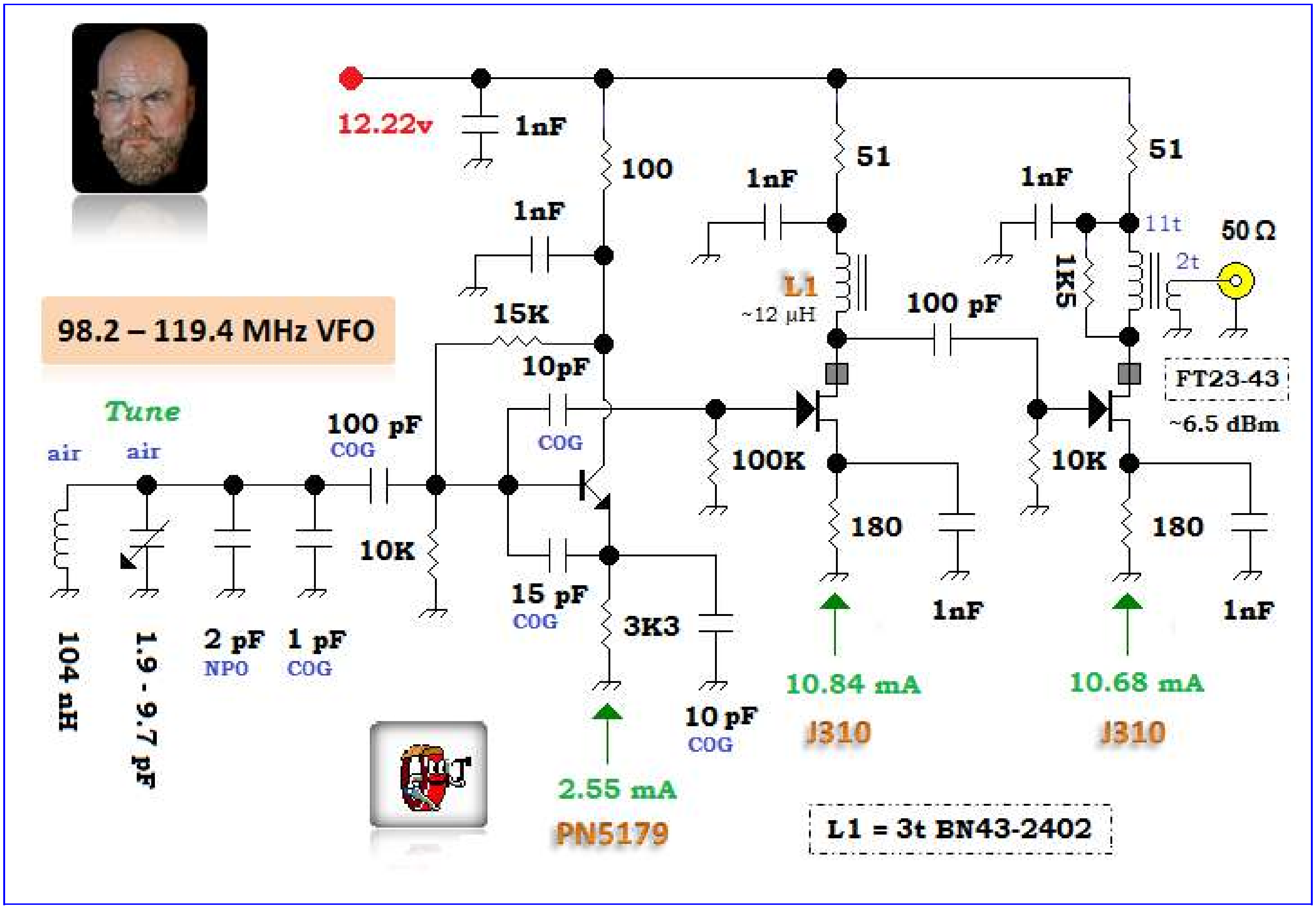
12.22v
1nF
100
1nF
51
1nF
51
1nF
11t
50 Ω
2t
L1
~12 µH
100 pF
1K5
FT23-43
~6.5 dBm
98.2 – 119.4 MHz VFO
15K
10pF
COG
Tune
air
air
100 pF
COG
100K
10K
10K
180
180
15 pF
COG
3K3
1nF
1nF
104 nH
1.9 - 9.7 pF
2 pF
NPO
1 pF
COG
10 pF
COG
10.84 mA
10.68 mA
J310
J310
2.55 mA
PN5179
L1 = 3t BN43-2402

Above — The temperature compensated, free running VFO. Normally we should run a VCO-PLL, digital contraption, or a VFO/VCO with ALC feedback, but this thing worked fine to learn on.

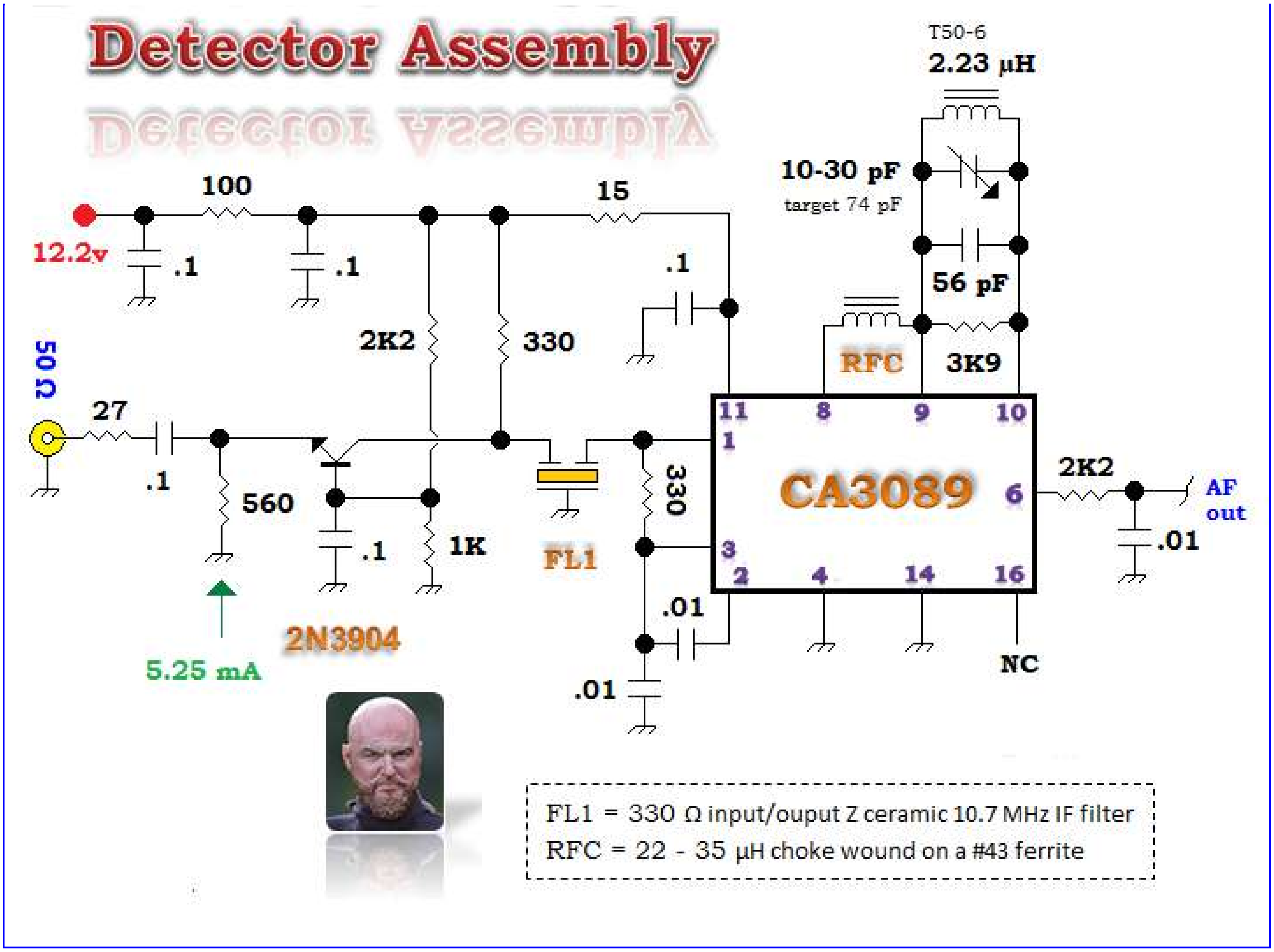

FL1 = 330 Ω input/ouput Z ceramic 10.7 MHz IF filter
RFC = 22 - 35 µH choke wound on a #43 ferrite

Above — The detector assembly. To match to the 50 Ω IF strip port, I employed a common base amp with a 330 Ω collector resistor driving a final 230 KHz wide "tail end or mop up" filter. In many designs you could simply drop this amp as the gain isn't really needed. On the other hand, the CA3089 contains a limiter and it won't hurt at all.

To make the detector coil, a small square island was carved on the copper clad board. The island was bisected so that pin 9 soldered to 1/2 the island and pin 10 to the other half. The 3K9 R, 56 pF C and a small, high Q SMT trimmer capacitor bridged the 2 islands. Finally, a 2.23 µH, T50-6 toroidal inductor was soldered across the 2 islands.

Tune a station and peak the trimmer cap for the most crisp sounding audio — it's set and forget. When driven with enough signal for full quieting, the CA3089 delivers very low distortion — the FM receiver in my truck sounds worse.

An RFC provides a DC short across pin 8 to 9 and keep ~5.5 VDC bias on the pin 9 transistor base. The minimum inductance should be 10X the detector coil. I wound my RFC with 6 turns on a BN43-2402 [35 µH] , but any old #43 ferrite wound choke >= 22 µH should work. For my chokes, I prefer a low # of turns on a binocular core to avoid any SRF issues.

Since the CA3089 contains a powerful audo preamplifer, only an AF power amplifer with low gain is required. I opted with the old LM386, but in future, better, "keeper rigs" I'll add an op-amp tone control stack driving a pair of power followers for good measure.

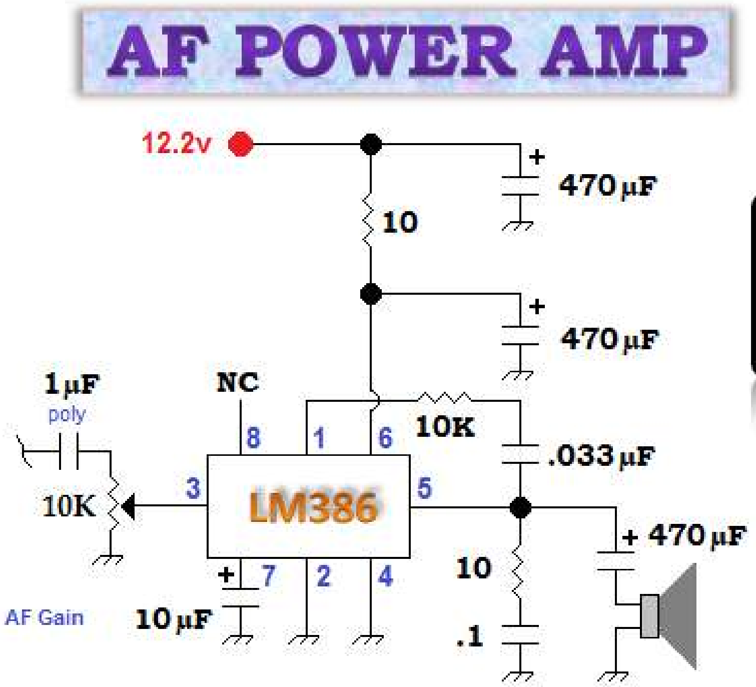

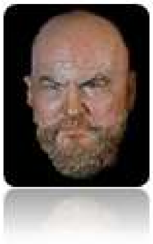

Above — Final audio power amp: an LM386 set for a voltage gain of 20 with some low-pass filtering and stout DC line filtering.

Click for a photo of the detector and LM386 assembly. Click for a better view of the outboard resonator and nearby RFC.

Outcomes

After connecting the VFO, antenna, IF strip, detector assemby, speaker and powering them up, loud FM radio blared through the loudspeaker. I tuned every local FM station with full quieting and felt surprised that despite no band-pass filtering, plus no RF preamp, it worked and sounded so well.

Further experiments showed that inserting as much as 15 dB attenuation between the IF stip and the detector assembly still kept full quieting on all the stronger stations. In some cities with multiple, strong FM stations, input band-pass filters will be needed. I'll cover that in a later installment.

I connected this VFO to the IF strip and tuned in my local NOAA weather station at 162.55 MHz. Click for a photo of the VFO built on a piece of bent copper sheet metal.

Future experiments may include changing ceramic filter bandwidth, trying different mixers, applying varactor-tuned input band-pass filters for strong FM environments and for FM Dxing. Finally, I'll design a low-noise preamp with a common gate JFET and/or a dual-gate MOSFET. Utimately, I hope to make a dedicated NOAA weather receiver, a broadcast FM Dx receiver and then some narrow band FM jobs for 2M or 6M. The future looks good.

Audio sample one : a few local stations [mostly recorded before I added the negative feedback on the LM386]
Audio sample two: a weak signal [our provincial Highway Report] that doesn't reach full quieting.

Thanks.

---

2. NOAA Receiver Experiments @ 162.55 MHz

I document some *raw* experiments and hopefully 1 day will make a good receiver for my local NOAA FM broadcast.

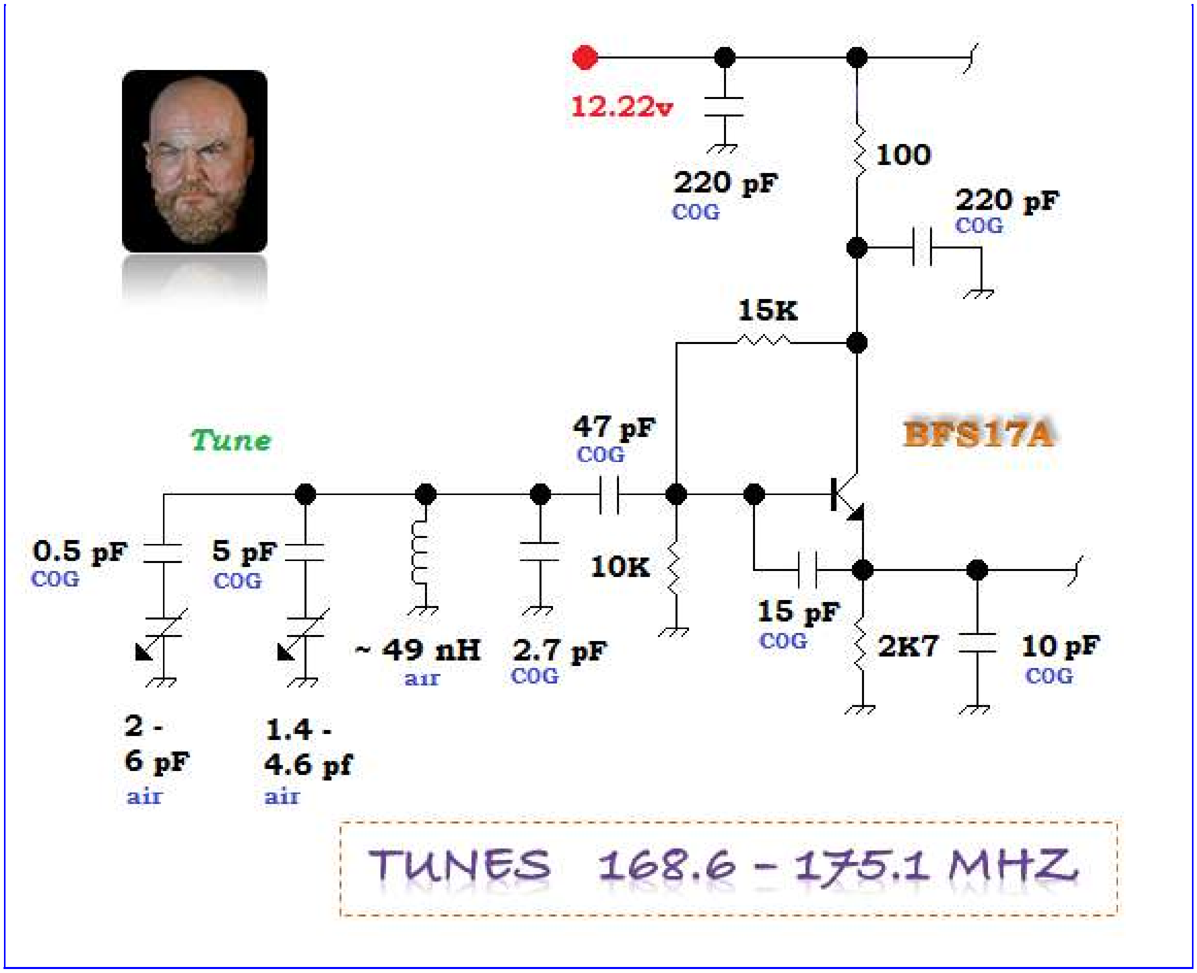
12.22v
220 pF
COG
100
220 pF
COG
15K
Tune
47 pF
COG
BFS17A
0.5 pF
COG
5 pF
COG
~ 49 nH
air
2.7 pF
COG
10K
15 pF
COG
2K7
10 pF
COG
2 -
6 pF
air
1.4 -
4.6 pf
air
TUNES 168.6 – 175.1 MHZ

Above — I built a VFO for 173.25 MHz with fine tuning for my local, NOAA weather report receiver @ 162.55 MHz. The fine tuning 2- 6 pF variable capacitor in series with an ATX 0.5 pF chip cap resolves 10's of kilohertz to permit tuning narrow band signals. You'll especially need fine resolution tuning when applying ceramic IF filters <= 150 KHz.

Above — The 2 VFO air variable tuning capacitors. Click for the circuit breadboard with a tuned, common emitter buffer, that I'd rather not show [because it sucks]. I ran the VFO into a bench module to boost the power up between 5 and 10 dBm in my experiments.

Since @ VHF, my VFO went on 2-sided FR4 board with several wire vias joining the 2 surfaces. We normally avoid double sided boards for VFOs since doing so makes a low Q, potentially moisture absorbant, undefined temperature coefficient "capacitor" that may worsen temperature drift. For bench learning and in particular, FM reception, temperature stability tested adequate and I can listen for hours without retuning.

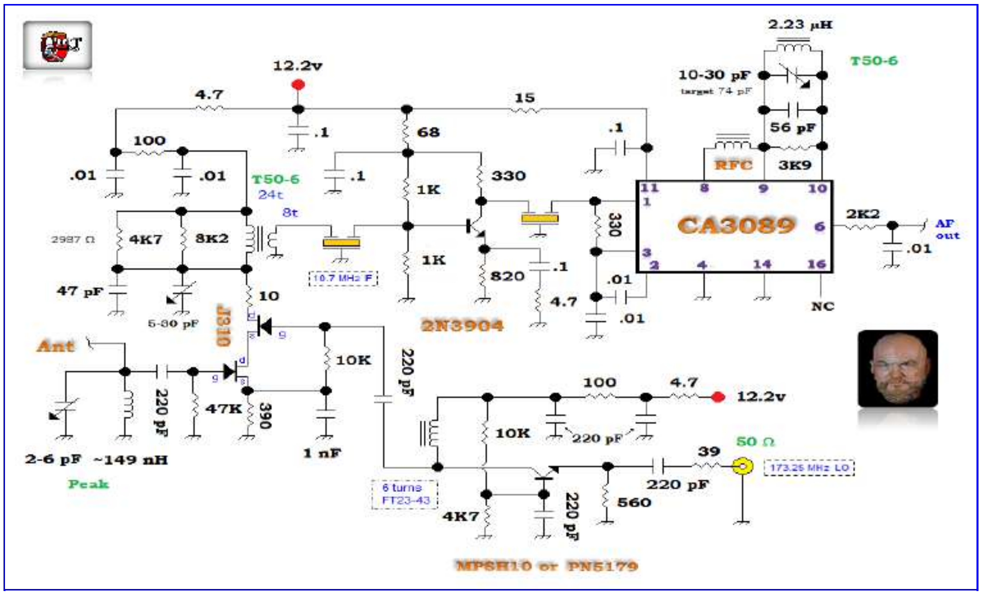
12.2v
4.7
15
.1
68
100
.01
.01
.1
T50-6
24t
8t
1K
330
2987 Ω
4K7
8K2
10.7 MHz F
1K
820
.1
4.7
47 pF
5-80 pF
10
J310
2N3904
Ant
10K
220 pF
220 pF
47K
390
1 nF
2-6 pF
~149 nH
Peak
2.23 µH
T50-6
10-30 pF
target 74 pF
56 pF
.1
RFC
3K9
11
8
9
10
1
CA3089
6
330
3
2
4
14
16
.01
.01
NC
2K2
AF out
.01
100
4.7
12.2v
10K
220 pF
50 Ω
39
173.25 MHz LO
6 turns FT23-43
220 pF
4K7
220 pF
560
MPSH10 or PN5179

Above — The entire receiver with a cascode mixer from EMRFD. In experimenter's mode; I tried out many ideas. For example: different bandwidth IF filters [settled on 230 KHz] , 1 versus 2 ceramic filters, or eliminating the 2N3904 amp [2 better drove the CA3089]. One salient learning — IF filters may possess a wide variation in center frequency and thus if you have a sweeper, matching them might be useful. In this receiver topology, you don't need to match 230 KHz or 280 KHz ceramic filters; the mismatch became noticeable with a <= 150 KHz 3 dB filter bandwidth.

I boosted the VFO signal with a common base amp — I varied the power into this LO amplifier by applying different strength attenuator pads. Between 1.5 to 2 Vpp on the mixer gate gave a reasonable compromise of conversion gain, noise and easy-to-get VFO power. From EMRFD, maximum conversion gain occurs when driven with 5 Vpp.

This unbalanced mixer worked fine, however, without a metal case, I heard a little hum from my DC supply and noises from household appliances such as the clothes washer and fridge. This receiver suffered from microphonics as well. The AF amp was a LM386 set for gain = 50.

A simple LC resonator tuned the mixer RF input and proved essential. I tried different ways to couple a 25 cm long wire antenna such as tapping the coil and an L- network, however, my main antenna peeve was severe hand capacitance effects. Although this receiver does not require a preamp for RF gain, the isolation seems needed. If you put your hand near the telescopic whip antenna of my Realistic Pro 2020 scanner, no noise or RF disturbance arises.

From studying scanner and FM receiver schematics, most rigs contain a dual-gate MOSFET preamp for a low noise figure and high input isolation. These MOSFET amplifiers feature input and output tuning, although with slug-tuned inductor cans and the oh-so-typical varactors, the resonator Q runs low.

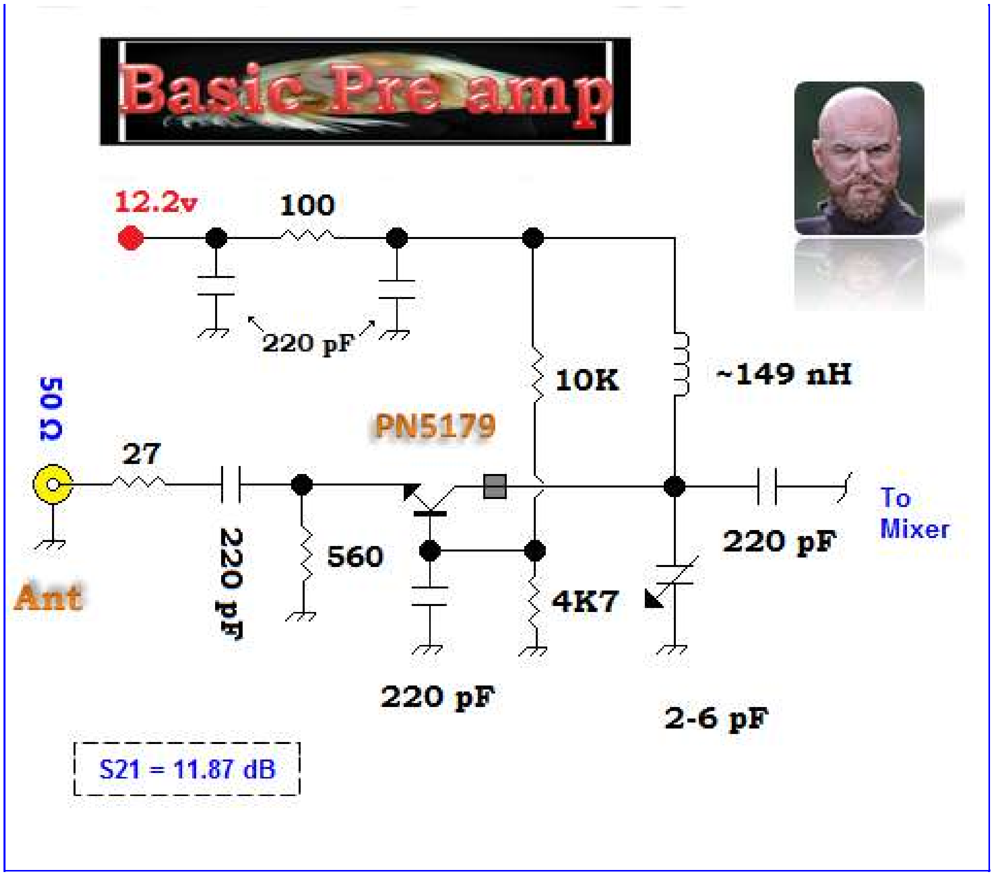
Basic Pre amp
12.2v
100
220 pF
10K
~149 nH
50 Ω
27
PN5179
To
Mixer
220 pF
560
220 pF
Ant
4K7
220 pF
2-6 pF
S21 = 11.87 dB

Above — I quickly built a tuned-output common base amp to study front end isolation. I removed the existing resonator and connected this to the JFET mixer gate. It eliminated antenna hand capacitance effects. Click for a sound byte of the receiver with this preamplifier and a 20 cm wire antenna attached to its input. [206 KB download]. I reckon some of my receiver problems arose from the mess of unshielded boards, modules and wire on my bench.

Tuned Input with a 2-Gate MOSFET [162.55 MHz]

A dual gate MOSFET preamp concluded this day's experiments. Tests of a simple input network follow:

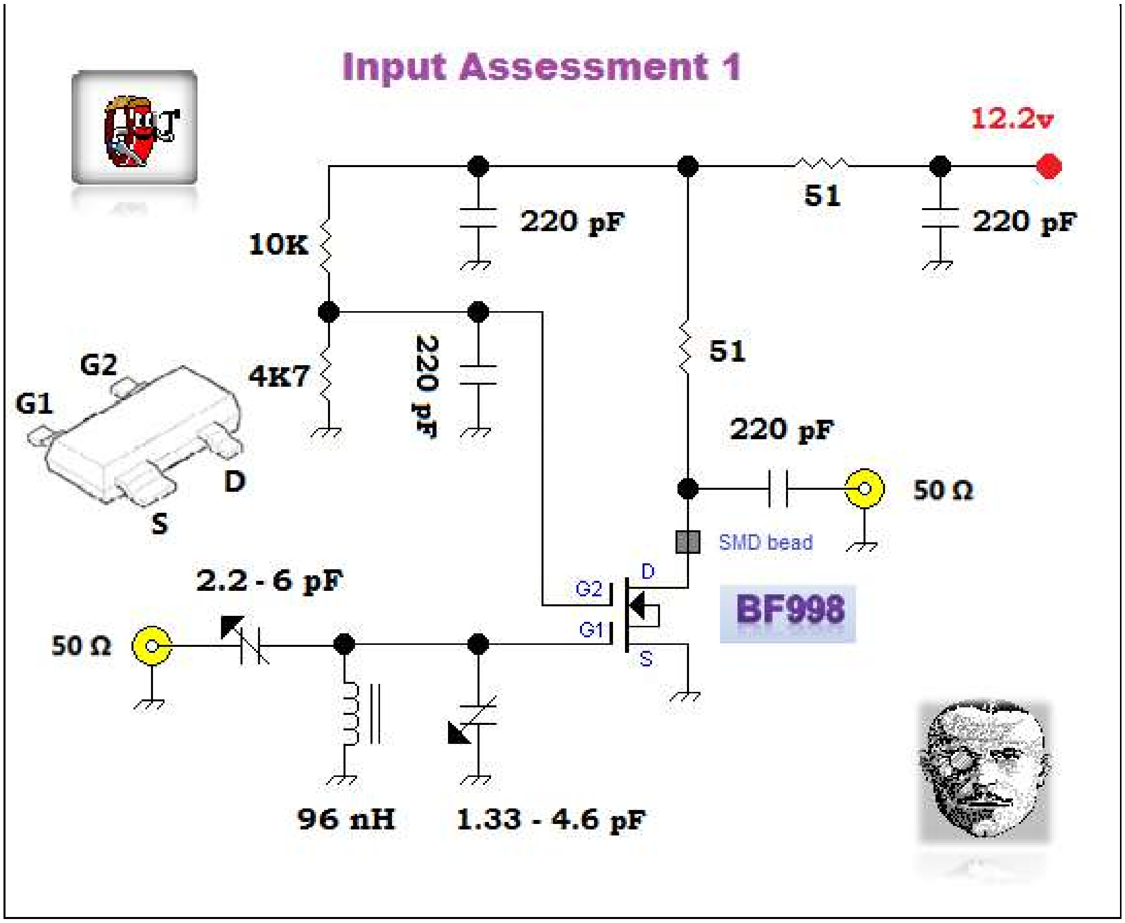
Input Assessment 1
12.2v
51
220 pF
220 pF
10K
4K7
220 pF
51
220 pF
50 Ω
G2
G1
D
S
SMD bead
BF998
2.2 - 6 pF
50 Ω
G2
G1
D
S
96 nH
1.33 - 4.6 pF

Above — Input Assessment 1. To examine the input without any output network effects, I loaded the BF998 drain with a 51 Ω resistor and matched the input parallel resonator for the greatest input return loss [-S21] by tweaking the series SMD trimmer capacitor. After setting the best input return loss [21 dB], I swept this amp.

The XL of the 96 nH air wound coil = 98 Ω and the XL and XC of this tank may be scaled to other frequencies by keeping the reactance values shown or choosing your own. For example; a coil for 50 MHz where XL = 98 Ω = ~312 nH. I discussed and swept the ferrite bead soldered on the FET drain in Section 1 of the VHF Veronica web page.

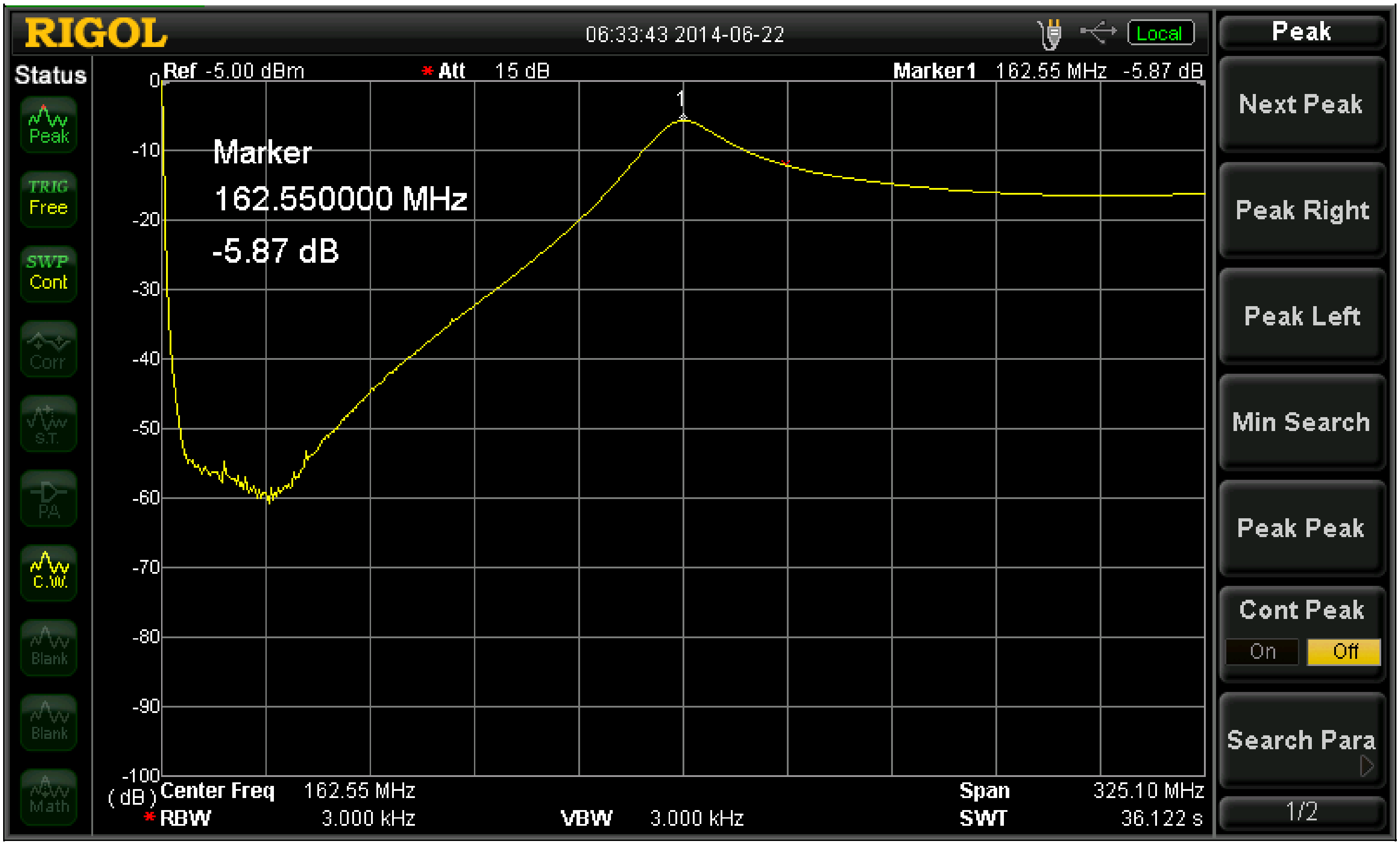
RIGOL
Local
Status
Peak
TRIG
Free
SWP
Cont
Corr
S.T.
PA
C.W.
Blank
Blank
Math
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
(dB)
1
Peak
Next Peak
Peak Right
Peak Left
Min Search
Peak Peak
Cont Peak
On
Off
Search Para
1/2

06:33:43 2014-06-22

Ref -5.00 dBm

Att 15 dB

Marker1 162.55 MHz -5.87 dB

Marker
162.550000 MHz
-5.87 dB

Center Freq 162.55 MHz

Span 325.10 MHz

RBW 3.000 kHz

VBW 3.000 kHz

SWT 36.122 s

Cont Peak: On [✓ Off]

Above — A sweep of the Input Assessment 1 amplifer. Although easy peasy to adjust, series matching the resonator to its input load results in a slipshod right-hand [low-pass] skirt. Click for a screen shot showing the amp's 3 dB bandwidth.

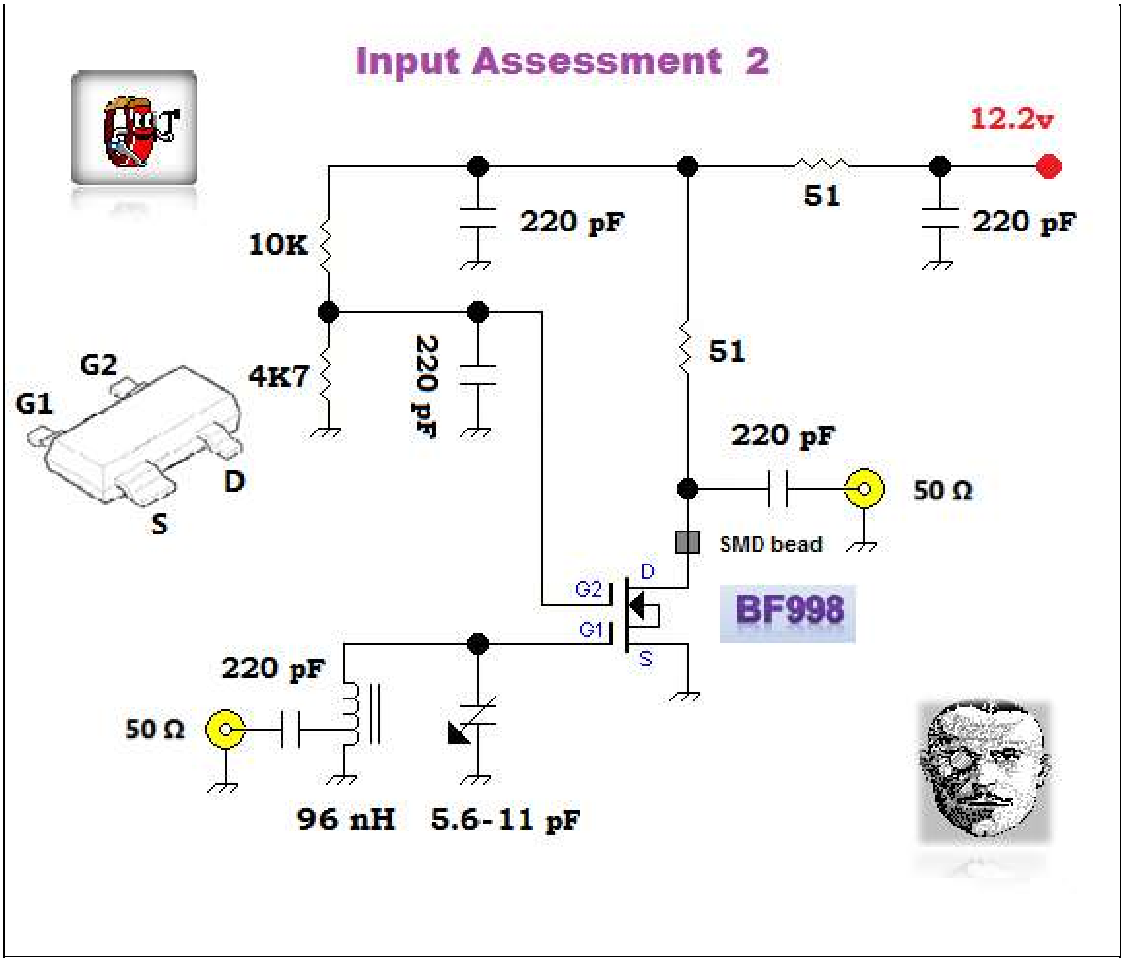
Input Assessment 2
12.2v
51
220 pF
10K
220 pF
4K7
220 pF
51
220 pF
50 Ω
G2
G1
D
S
SMD bead
G2
G1
D
S
BF998
220 pF
50 Ω
96 nH
5.6-11 pF

Above — Input Assessment 2. After measuring a poor low-pass skirt in Input Assessment # 1, I opted for the conventional tapped transformer to steepen the low-pass skirt response versus frequency. My sweeps *did* show a better low-pass skirt shape . Finally, I built a "real" amp and soldered in a broadband output network, tweaked the bias scheme + limited the VDD with a series 150 Ω resistor. A higher value input resonator trimmer cap is required for this amplifier versus that in Input Assessment # 1.

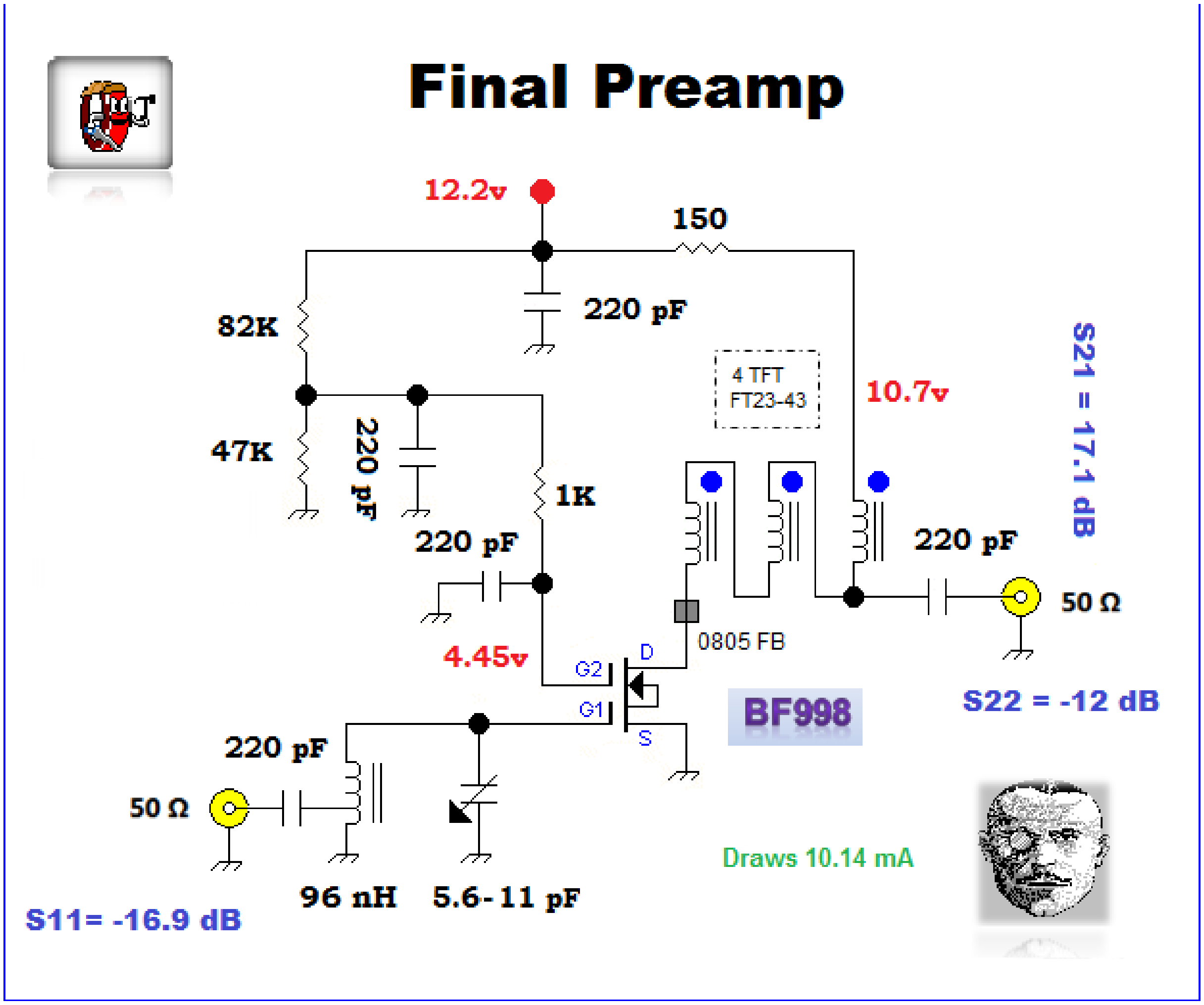

Above — The final 162.55 MHz preamp schematic + some measures. Click for a breadboard photo. Click for the gain measurement with a TG & SA. I measured the same S21 [within 0.3 dB] using my VHF signal generator plus my 50 Ω input DSO. I didn't push too hard with the gate bias or VDD, since neither maximal gain nor parasitic oscillations were sought. Click for a zoom of the filter peak response. Click for a wider sweep to see the right-hand skirt response.

Built on 2-sided FR4 with short cut paths to fit mostly size 0805 SMD parts, I placed via wires on every ground point and detected 0 unwanted oscillations from HF to 3 GHz.

A bare copper wire coil allows us to set the input impedance tap point anywhere along the wire — unlike the *integer only* tap points provided by magnetic toroid-based transformers. In LNA design, many builders set the input tap point for the lowest NF with the amp in noise figure test mode. Another vanilla flavored option is to set the tap with the amplifier input connected to a return loss bridge while adjusting the tap point for the best input return loss — that's what I did. My S11 = - 16.9 dB with my final tap around 1.2 turns of the total 7 links on my 96 nH coil.

The output transformer presents a poor match to 50 Ω with a S22 of just -12 dB. I plan to put a 4 dB pad after it for my NOAA receiver, so it's okay. Some might chase a better output return loss and also perhaps tune the output for better band-pass filtration.

Your FM receiver preamp might need better input filtration, however, if a simple design works, then simple rules! Have fun with *your own* experiments.

473
270
472

RF — Test and Measurement

Pin Outs

Mini-Circuits Labs SBL-1 and TUF-1 diode ring mixers

...more later

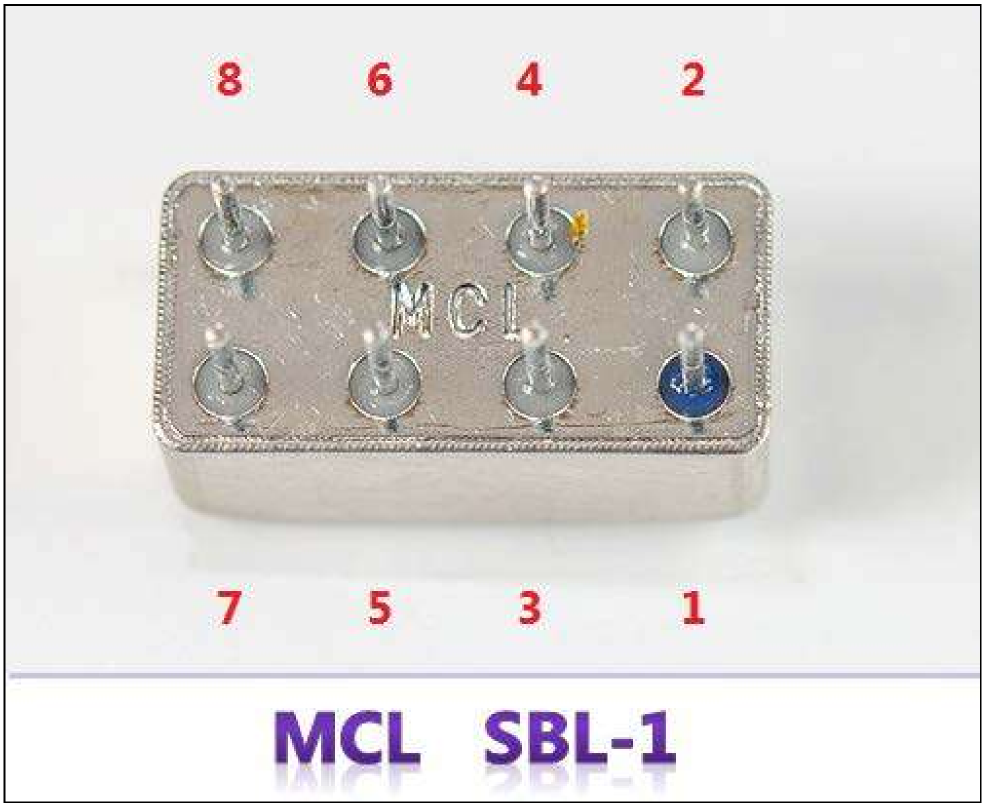
8
6
4
2
MCL
7
5
3
1
MCL SBL-1

Above — Normal set-up: **Ground** -- Pins 2, 5, 6 and 7. **LO** -- Pin 8, **IF** -- Pins 3 and 4, **RF** -- Pin 1

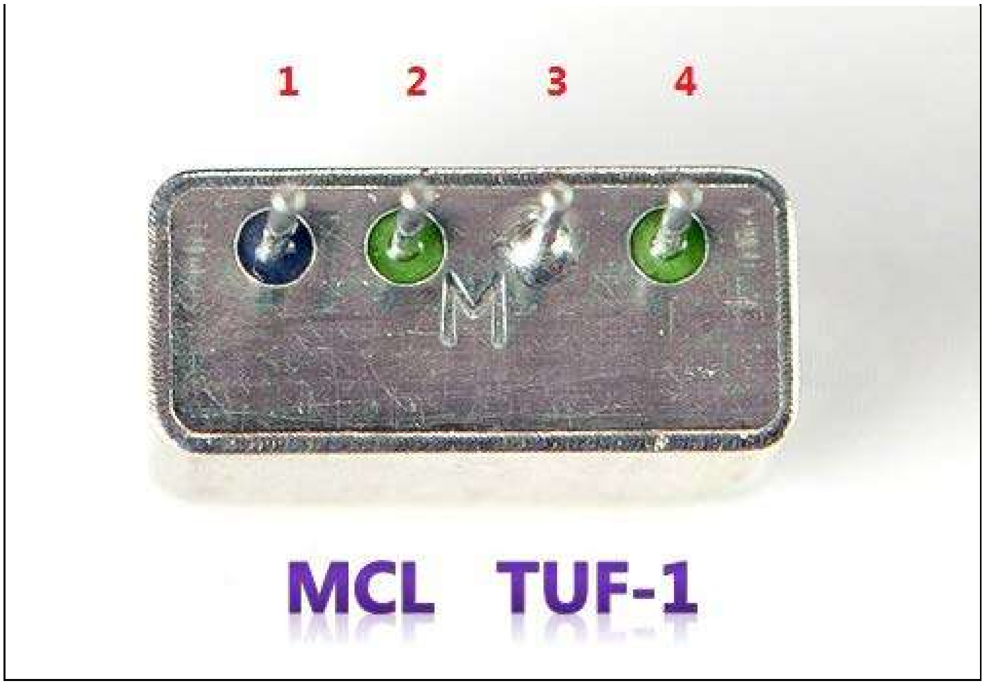
1
2
3
4
M
MCL TUF-1

Above — **RF** -- Pin 1, I**F** -- Pin 2, **Ground** -- Pin 3, **LO** -- Pin 4.

RF — Test and Measurement

VHF-FM Supplement — Amplifiers

Deleted October 26, 2013

Amateur Radio Electronic Design

KL7R Memorial Receiver Experiments

Introduction

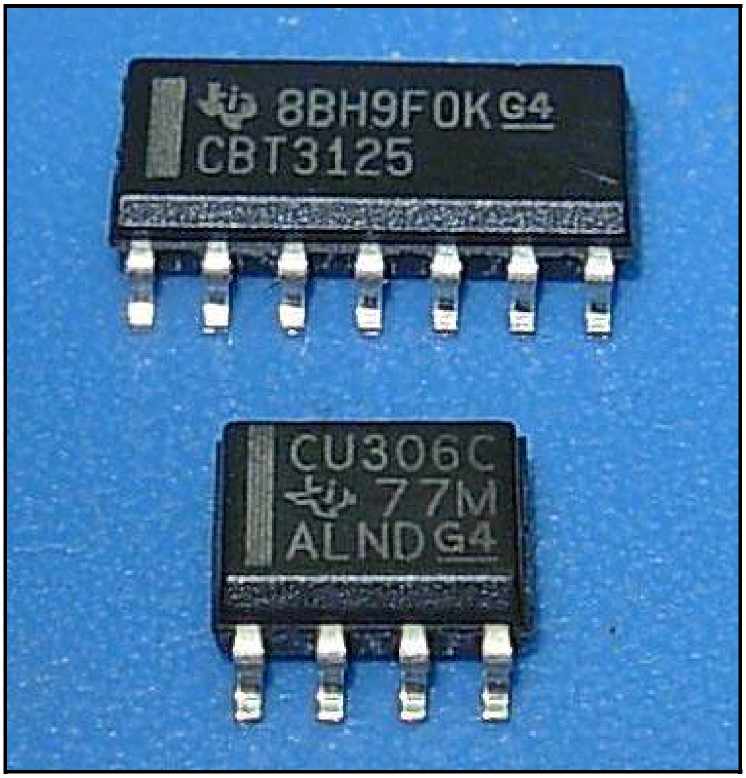

This web page is a memorial to Mike Caughran, KL7R, who died suddenly in January 2007. Mike was a passionate experimenter who was embraced by the homebrew radio electronics community. He was best known as the co-creator of and sidekick to Bill Meara, M0HBR on the podcast Solder Smoke. Mike was a hardcore science and technology buff. His knowledge of general science and curiosity about minimalist RF designs was amazing. I worked Mike on 40 and 80 meter CW and later by voice on eQSO. Mike was the first person to refer to this web site as the "popcorn site". He held an interest in digital circuits. In tribute to Mike, a series of receiver experiments which includes some digital circuits are presented. Mike Caughran will be remembered as a remarkable, kind and passionate homebuilder.

Navigation

There are 3 linked web pages associated with this KL7R tribute web page.

1. Mike's personal web page His own web page memorializes him best
2. Supplemental web page Additional schematics and photos which supplement this main web page
3. VFO web page Describes the VFOs used for these experiments

My special thanks to Wes, W7ZOI for his coaching and suggestions to improve many of the circuits on this web page. This web page borrows heavily from his designs as presented in EMRFD.

Receiver Description

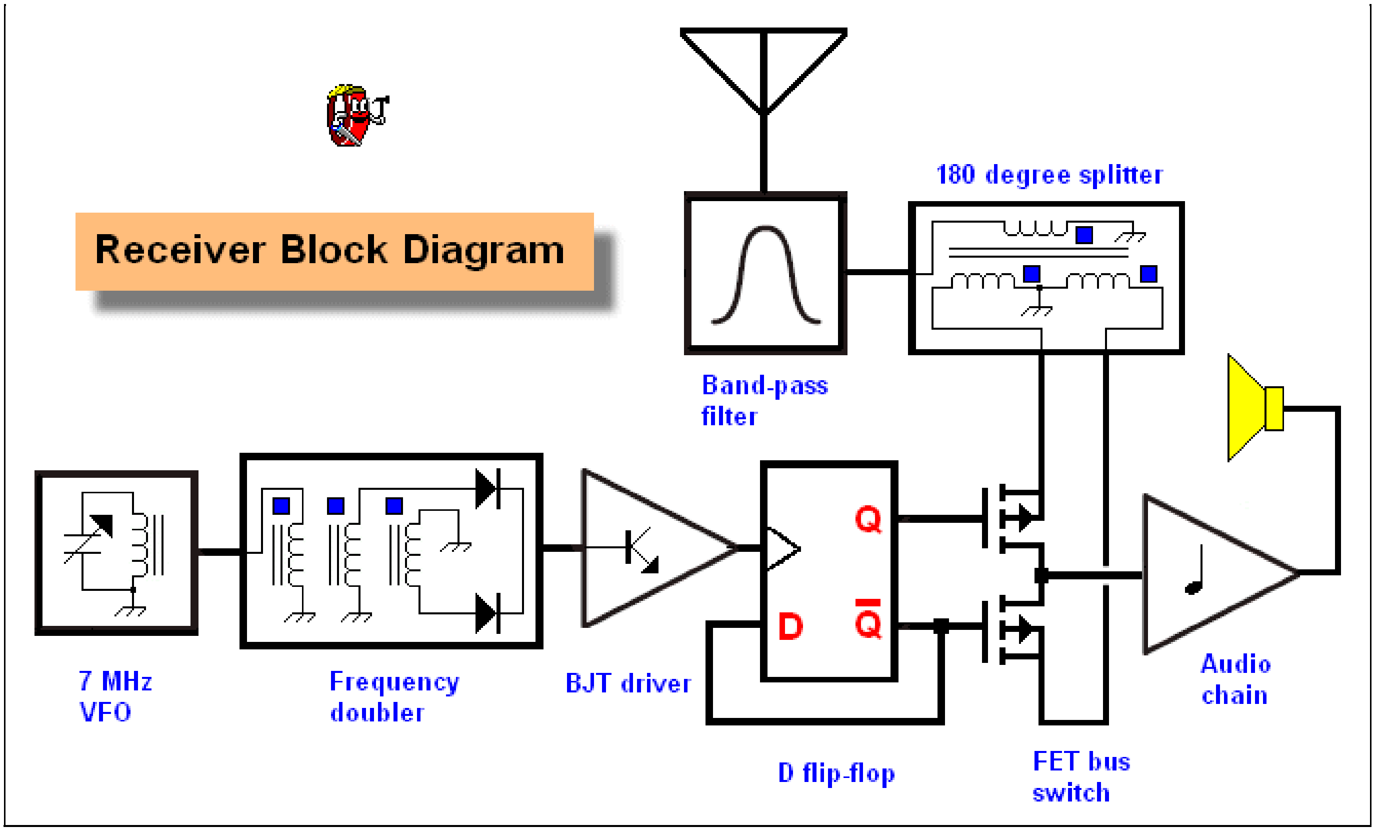
Receiver Block Diagram
180 degree splitter
Band-pass filter
Q
D
Q̄
7 MHz VFO
Frequency doubler
BJT driver
D flip-flop
FET bus switch
Audio chain

Shown above is the receiver block diagram. The KL7R memorial receiver depicted on this web page is the final output of many hours of experimentation. Most of the circuits or circuit ideas originated in EMRFD.
From our conversations, Mike was always challenging himself; experimenting, testing and pushing his knowledge threshold. The joy of discovery motivated him. Fired by his spirit of inquiry, I explored methods to build a receiver containing at least 1 digital circuit. On many days, I accomplished nothing. The circuits did not work and little to no progress was made. These were the difficult dry spells all experimenters must endure. Design and circuit failures can be very disheartening. I also wasted a lot of parts. However, I kept going and slowly successes occurred and my confidence rose. The end result was a little more knowledge and a cool, popcorn, direct conversion receiver which I hope will provide ideas and inspiration for your own experiments.

Double Tuned Band-pass Filter

Figure 1

40 Meter Band-pass Filter

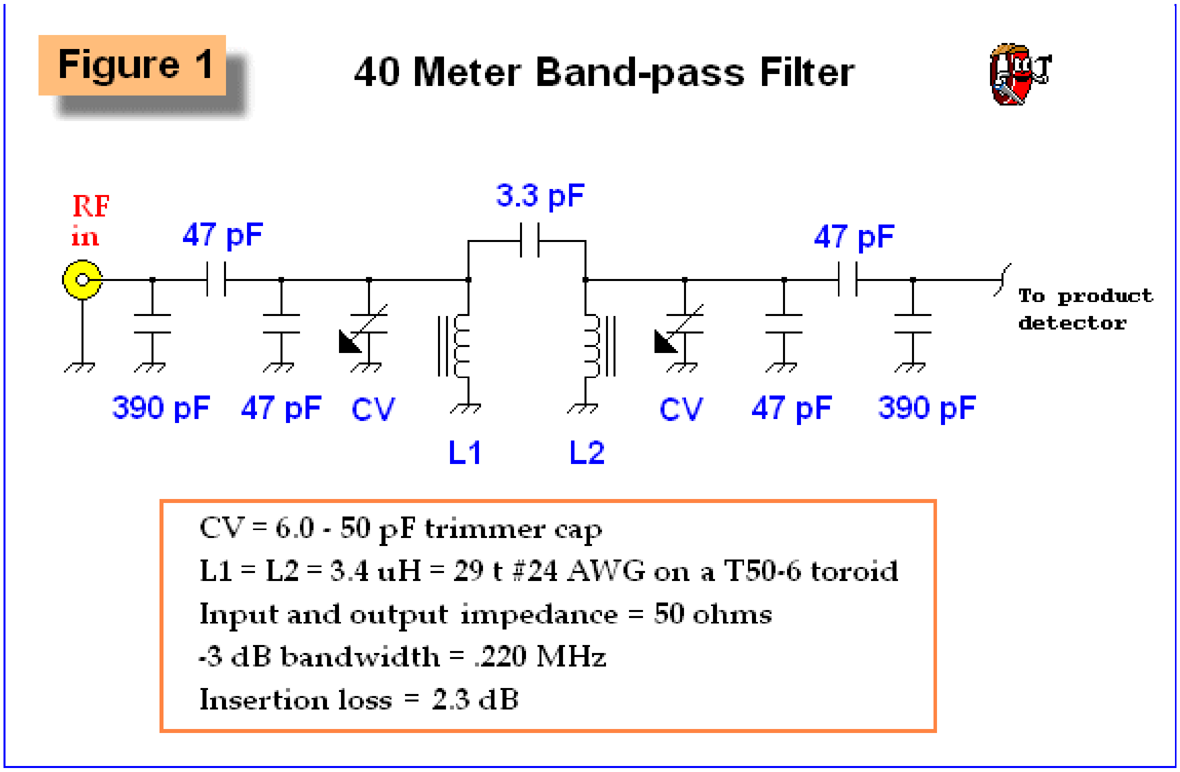

CV = 6.0 - 50 pF trimmer cap
L1 = L2 = 3.4 uH = 29 t #24 AWG on a T50-6 toroid
Input and output impedance = 50 ohms
-3 dB bandwidth = .220 MHz
Insertion loss = 2.3 dB

Figure 1 shows the front end band-pass filter. If you can't obtain a 3.3 pF coupling capacitor, try this other 7 MHz band-pass filter circuit or perhaps just design your own. See the Webmaster's page for information concerning many of the parts used on this web page.

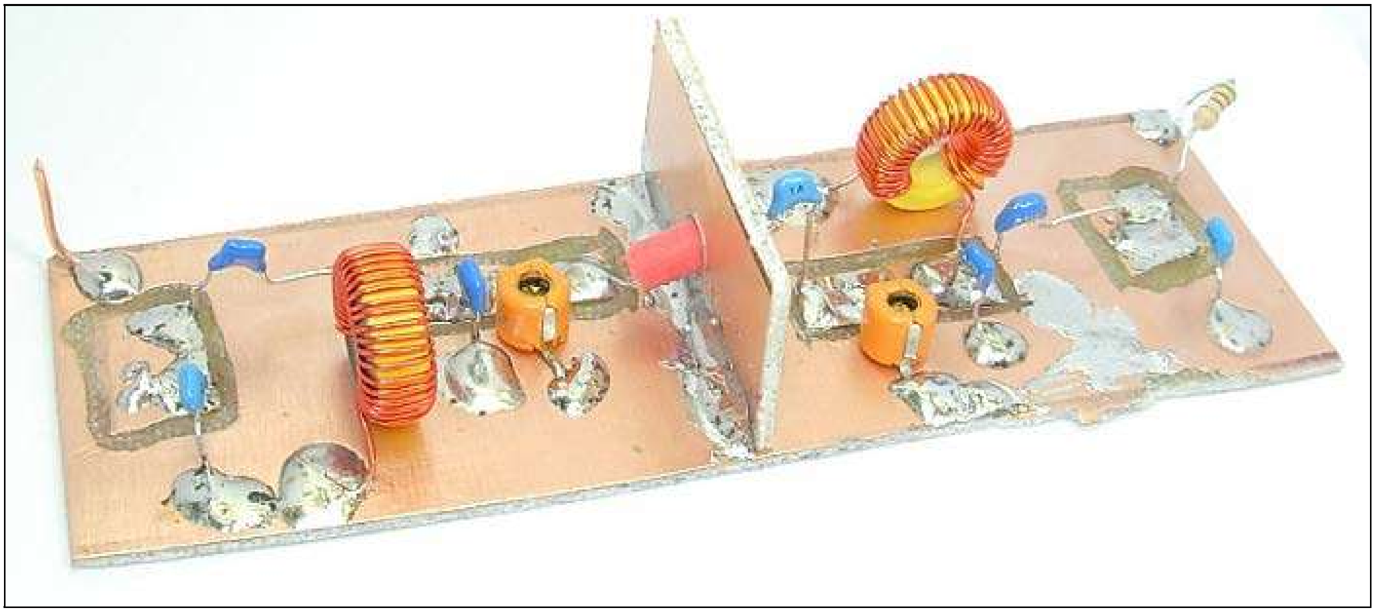

Shown above is a photograph of the Figure 1 breadboard. The inductors were spaced apart at right angles to reduce unwanted coupling. The copper clad board L-C tank divider is not necessary. The 51 ohm load resistor seen to the right was removed after testing.

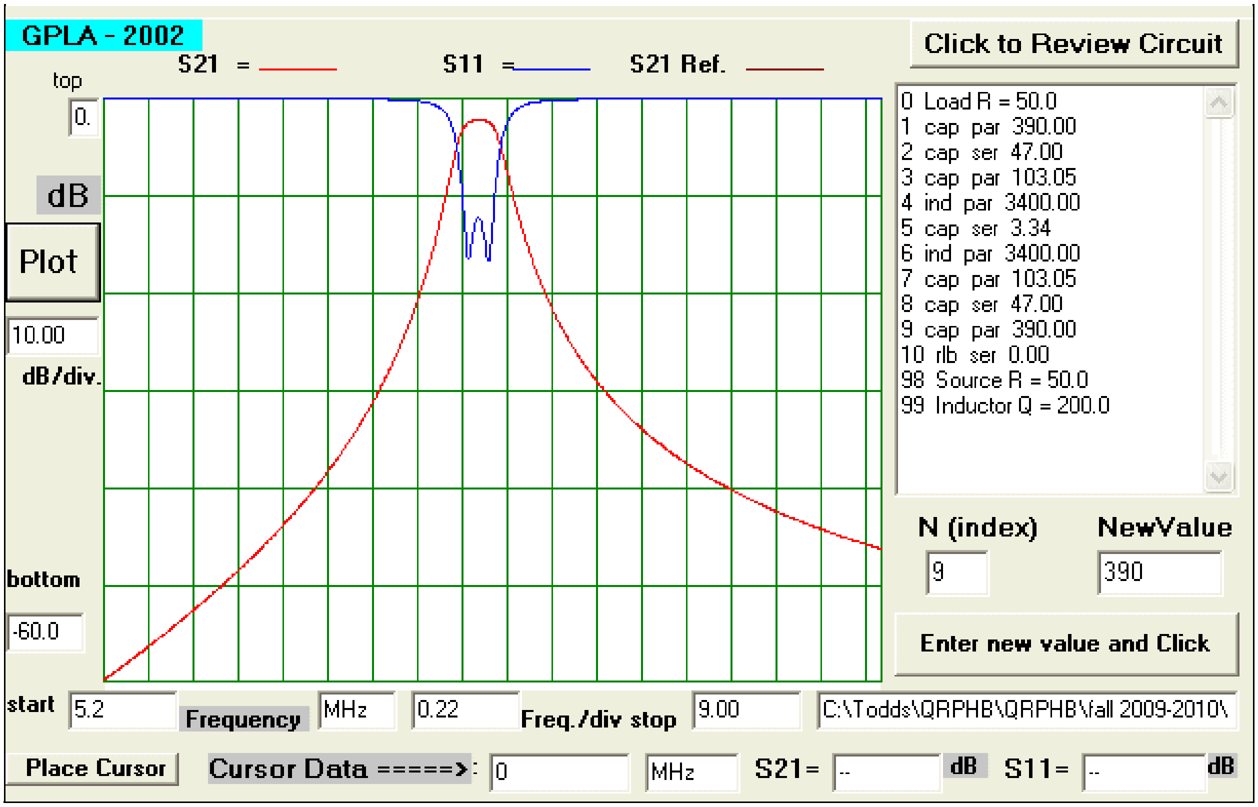
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
Plot
Enter new value and Click
Place Cursor
Cursor Data =====>:

```
0  Load R = 50.0
1  cap  par  390.00
2  cap  ser  47.00
3  cap  par  103.05
4  ind  par  3400.00
5  cap  ser  3.34
6  ind  par  3400.00
7  cap  par  103.05
8  cap  ser  47.00
9  cap  par  390.00
10  rlb  ser  0.00
98  Source R = 50.0
99  Inductor Q = 200.0
```

top: 0.

dB

10.00 dB/div.

bottom: -60.0

start: 5.2 Frequency MHz

0.22 Freq./div

stop: 9.00

C:\Todds\QRPHB\QRPHB\fall 2009-2010\

N (index): 9

NewValue: 390

Cursor Data: 0 MHz

S21= -- dB

S11= -- dB

Shown above is a GLPA simulation of the Figure 1 band-pass filter. Хорошо.

Product Detector Experiments

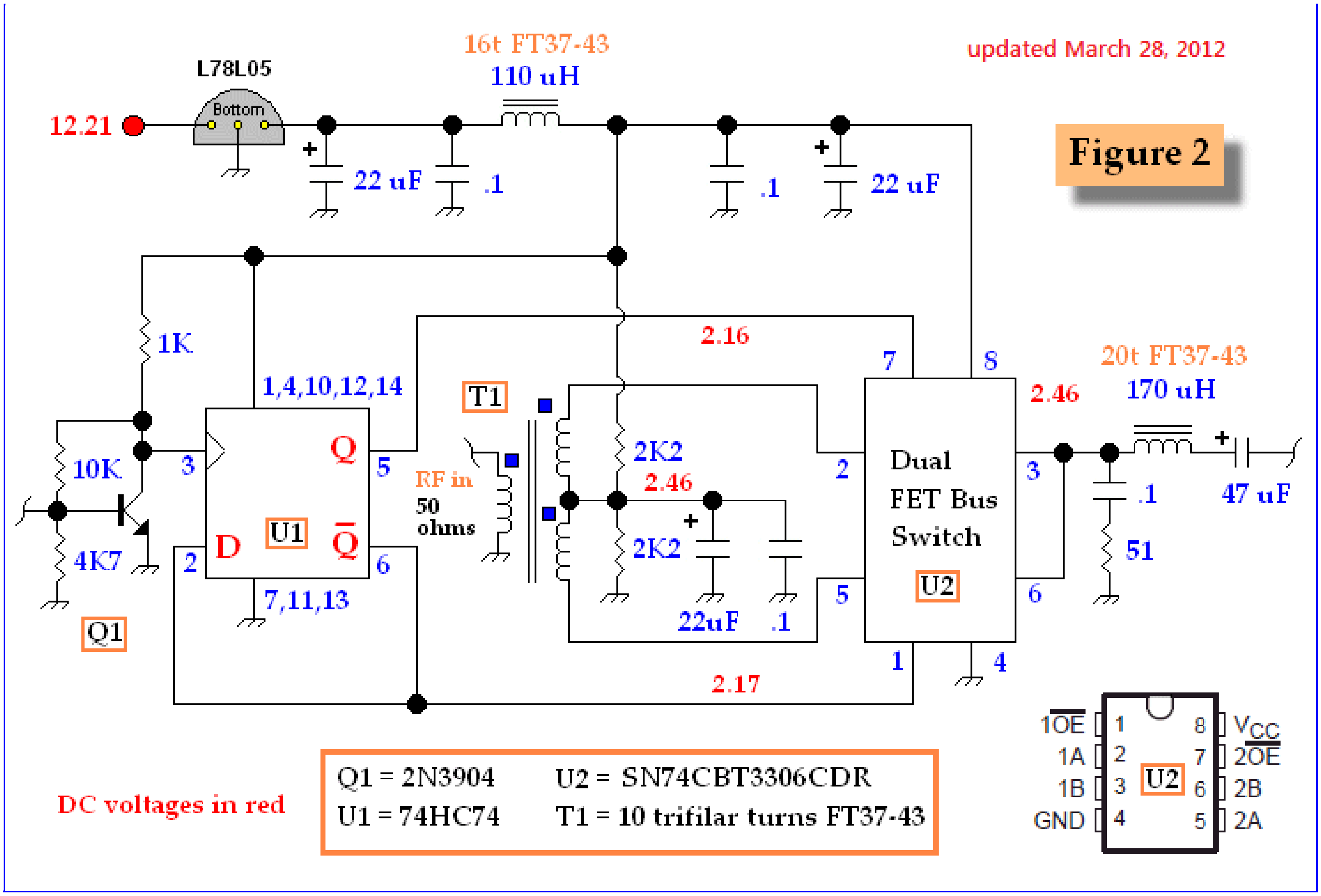
updated March 28, 2012
Figure 2
L78L05
Bottom
12.21
16t FT37-43
110 uH
22 uF
.1
.1
22 uF
1K
10K
4K7
Q1
3
1,4,10,12,14
Q
U1
D
Q̄
5
6
2
7,11,13
T1
RF in
50 ohms
2K2
2.46
2K2
22uF
.1
2.16
7
8
2
5
1
4
3
6
2.17
Dual FET Bus Switch
U2
2.46
20t FT37-43
170 uH
.1
47 uF
51
DC voltages in red
Q1 = 2N3904
U1 = 74HC74
U2 = SN74CBT3306CDR
T1 = 10 trifilar turns FT37-43
1OE
1A
1B
GND
1
2
3
4
8
7
6
5
VCC
2OE
2B
2A
U2

Shown above in Figure 2 is the BJT driver, D flip-flop and the CMOS switch product detector. The 14 MHz VFO connects to Q1 via a 0.1 uF coupling capacitor that is shown on the VFO schematic. A dual FET bus switch (CBT3306) serves as the product detector. The on-resistance of this switch is only around 3 ohms! If you are ambitious, you might try using a 14 pin SOIC switch such as the QUAD FET SN74CBT3125DR with 2 pairs of the 4 switches wired in parallel. I tried 3 different CMOS switches in the U2 slot. The other switches were the MAX4066CPD and a 74HC4053 (wired up appropriately using their datasheets). The insertion loss and performance of these 2 switches was disappointing. My bench standard for comparison was a 7 dBm diode ring mixer. Numerous experiments were performed. For example, I tried running the 4066 at 12 volts VCC to minimize its on-resistance and had to modify most of Figure 2 as well. Being new to digital, blending 5 volt and 12 volt logic IC required great effort to get it working properly. These experiments consumed the better part of 2 days. My conclusion was that if you are going to go to the trouble of make a CMOS switch work, you might as well use a part that has a low on-resistance. Hence, I have since abandoned using DIP IC CMOS switches (4053, 4052, 4066 etc.) as mixers and product detectors. They may still be a good choice in your own context.

The CBT3306 is outstanding and very similar to the diode ring mixer with respect to insertion loss and audio quality as a product detector. There are some other good CMOS switches you might try. I chose the CBT3306 because it was SOIC (the largest of the common SMT topologies), costs only 77 cents (Canadian dollars) and only has 8 pins to deal with. Pragmatism on the workbench is always good!

SMT versus VE7BPO

The difficulty with using the CBT3306 was that I had to learn about and equip for SMT. I ordered the switches and after their arrival, hesitated to do anything with them for 5 days. I managed to borrow a magnifier and bought a SOIC prototype board and some flux. Still, I seemed to be paralyzed with fear about soldering U2. I was stuck. For inspiration, I went back to Mike's web site and found this web page. Here Mike was working with a 18 pin SOIC chip and I was worried about a mere 8 pin IC! Part of the problem was my relatively poor eyesight. I learned this can be managed with a magnifying visor. I bought mine here , but only after U2 was soldered in. The borrowed table magnifier was okay, but constrained arm movement and reduced lighting. The visor seems to be a better choice.

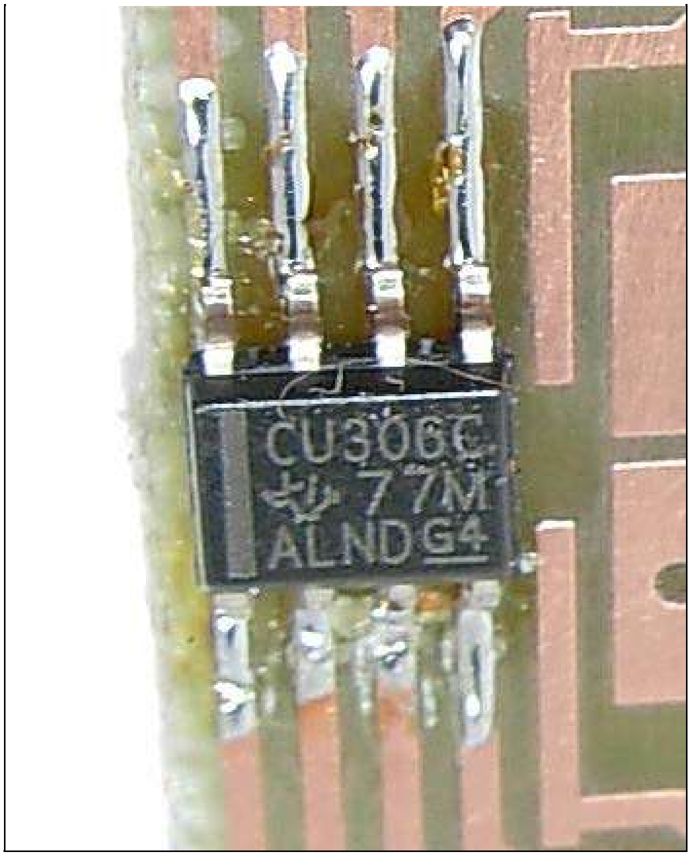

Shown above is the CBT3306 soldered on my prototype board. I put flux on the board traces and then tinned them. Following that, I lined up U2s leads on the traces and began soldering. The bottom pins (pins 1-4) did not go well, but they were soldered all the same. Following that, the soldering of pins 5-8 went very well. I was very happy; the SMT monster had been tamed! It was learned from Wes, W7ZOI that many builders use Surfboards for SOIC applications. I will get some for future SMT IC work. Using SMT parts in your Ugly Construction allows you to use parts which are unavailable otherwise. Increasingly, good old DIP ICs are disappearing from catalogs and some new parts are appearing as SMT only. It makes sense to jump in and use SMT parts when it is advantageous for spec reasons, or if miniaturization is required.

With some effort, the Figure 2 circuit could be morphed into Colin, G3SBI's H-mode mixer format. Not on this web site though! If you Google "H-mode mixer" , many good websites will be returned.

The product detector's baseband audio output is at the 47 uF capacitor which connects to Q1 of Figure 3. A simple 51 ohm, low pass network is used to terminate U2. This is from W7EL's An Optimized QRP Transceiver from QST for Aug 1980. I have some more information regarding AF termination circuits on this web page. You may wish to increase the 170 uH inductor value somewhat to get more low pass filtering. The product detector circuit is the number one potential source of hum and noise in this receiver. Take the time to plan your layout to minimize wire length and crossing and provide some physical symmetry. Decouple well.

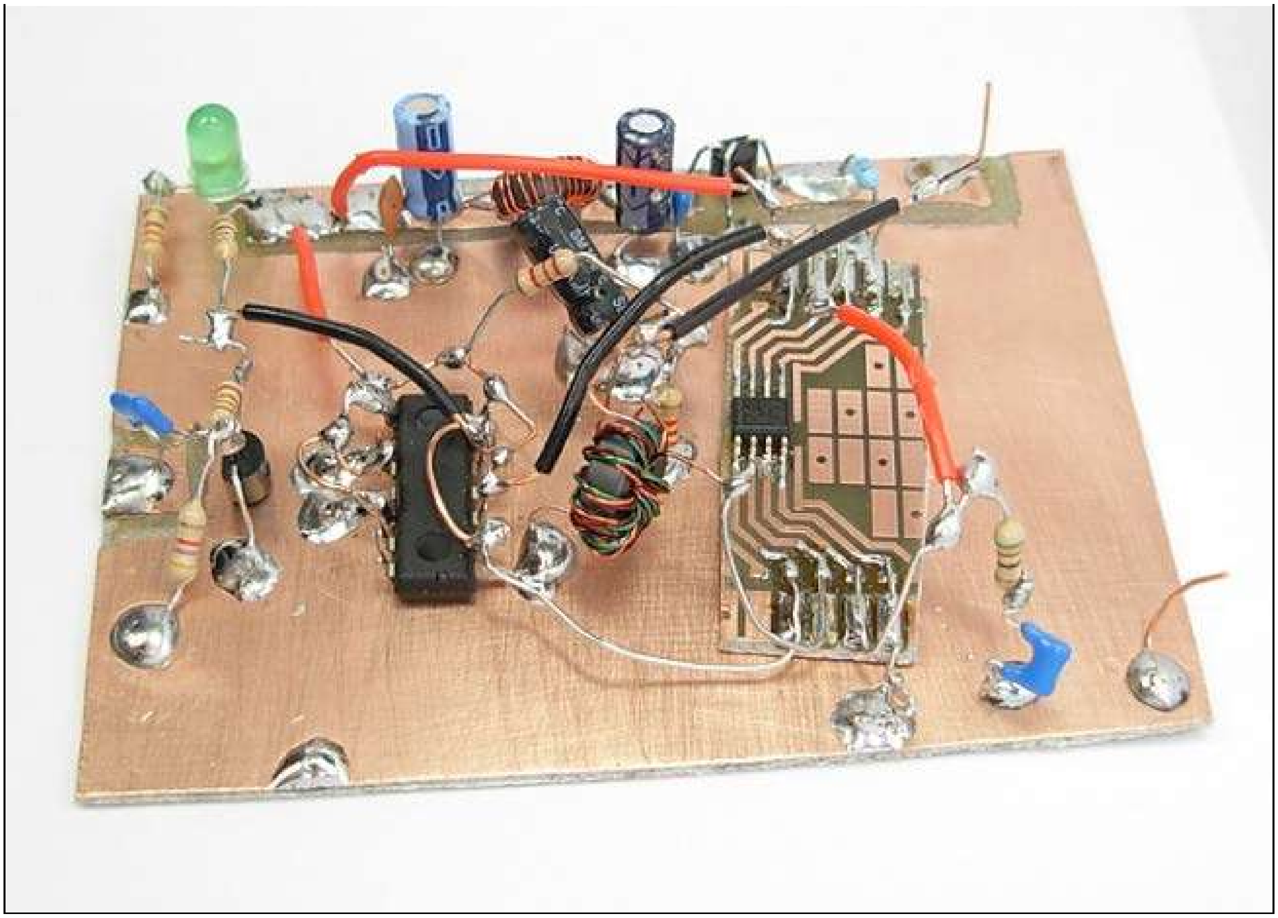

Shown above is the product detector. At this point, I had not decide how to terminate U2 and had soldered a shunt 0.1 uF cap and 51 ohm resistor to the switch output. U2 is dwarfed by the FT37-43 toroid; a part we normally think of as small in size.

Figure 3

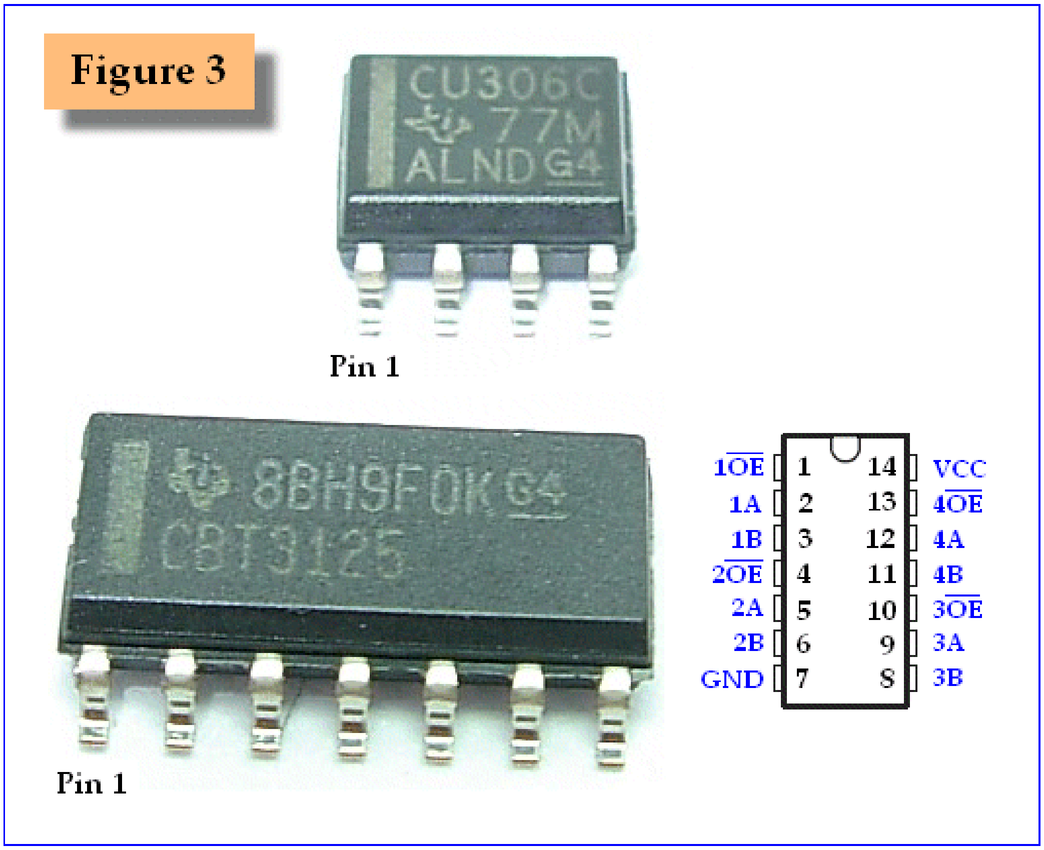
CU306C
77M
ALND G4
Pin 1
8BH9F0K G4
CBT3125
Pin 1
1$\overline{OE}$ 1 14 VCC
1A 2 13 4$\overline{OE}$
1B 3 12 4A
2$\overline{OE}$ 4 11 4B
2A 5 10 3$\overline{OE}$
2B 6 9 3A
GND 7 8 3B

Figure 3. Photographs of the FET bus switches I have in my collection and the pin-out for the 14 pin CBT3125

A diode ring mixer was used as a reference mixer in the product detector experiments.

Audio Amplifier Chain

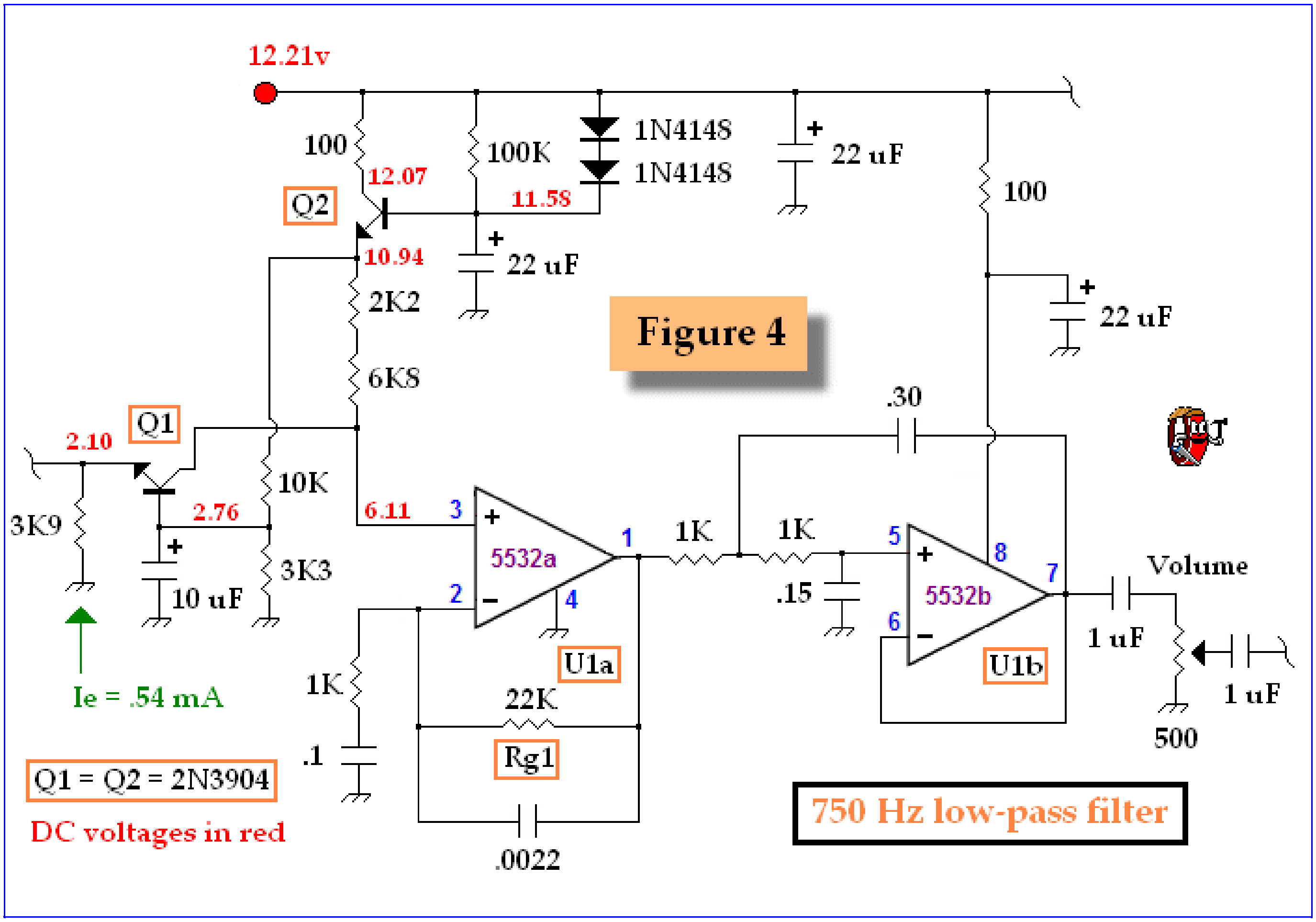
12.21v
100
12.07
Q2
100K
1N4148
1N4148
11.58
22 uF
22 uF
100
10.94
2K2
6K8
22 uF
Figure 4
.30
Q1
2.10
3K9
2.76
10K
10 uF
3K3
6.11
3
2
1
4
5532a
1K
1K
5
8
7
6
.15
5532b
U1a
U1b
Volume
1 uF
1 uF
500
Ie = .54 mA
1K
22K
.1
Rg1
.0022
Q1 = Q2 = 2N3904
DC voltages in red
750 Hz low-pass filter

Shown in Figure 4 is the audio preamp and first low pass filter. Direct conversion receivers are all about audio. High performance DC receivers have become more common since Rick, KK7B unleashed his R1 on the world in 1992. High performance receivers are out of scope for this popcorn website, however, this design is welcome. The entire audio chain uses low-cost 5532 op amps, cheap BJTs, plus fairly common resistor and capacitor values. Poly"something" capacitors were used for all audio AC coupling and shunting capacitors of 4.7 uF or less.

For hum immunity, the familiar BJT capacitive multiplier popularized by Roy, W7EL was used at Q2. Using a high-end spectrum analyzer, Wes, W7ZOI demonstrated that this circuit can oscillate. He detected oscillations at UHF. Thus, Wes recommended using a 100 ohm snubber resistor on the Q2 collector to alleviate these parasitic oscillations. Additionally, 2 diodes were added in parallel to the 100K base resistor. This provides an instant-on feature for the audio preamp, as normally there is a time-constant delay when you switch on the receiver. I like this feature, but it is purely optional.

Q1 is a common base amp biased for 0.54 mA. Therefore, the input impedance is 26/0.54 or about 48 ohms. The 6K8 + 2K2 collector resistors were paralleled so that the quiescent collector voltage was close to 6.1 volts. You could also substitute a single 9K1 resistor, although this is an uncommon value. This voltage provides the approximate VCC/2 bias needed for U1a and U1b. Connecting the Q1 collector directly to the pin 3 op-amp input allows the exclusion of a coupling capacitor and the usual VCC/2 resistor network used to bias the 5532 op-amp from a single power supply. I borrowed this from EMRFD.

The gain of U1 is set by the resistor labeled Rg1. If after testing, the AF gain is too high (for example if this AF stage was used in a superheterodyne receiver), simply lower this resistor value. I chose a 22K ohm resistor to allow enough gain for weak signal listening. On louder stations, you will need to lower the volume control to prevent distortion in the stages that follow as overall, there is a lot of gain in this AF chain. Adjust the Rg1 value to suit your needs. If you use a switch such as a 4052 for U2, or your receive antenna is small, you may want to increase this resistor a little for more gain. This is an experimenter's receiver after all. The 0.0022 uF capacitor in the op-amp feedback loop provides a single pole of low pass filtering. I ran this capacitor as high as 0.0082 uF. This gave a theoretical 3 dB cutoff frequency of ~ 880 Hz, but the receiver lost its sizzle. Experimentally, I learned that using a 0.0022 uF feedback capacitor on both U1A and U2A dropped some of the high frequency noise while preserving some sparkle in the received audio.

The audio preamp chain is concluded with a 750 Hz low-pass filter. Resistor values were kept low to minimize noise. All the active low pass filters are low Q, Sallen-Key designs with a Butterworth response. They are stable, easily scalable and brain-dead simple. These filters are fatigueless due to their gentle cut off slope. I actually had a 0.30 uF capacitor in my parts collection, although it was a 600 volt rated part from my tube amp parts bag. You could use two 0.15 uF capacitor in parallel or just substitute a 0.33 uF capacitor. If you can't find a 500 ohm volume control potentiometer, swap in whatever you have, but try to keep the resistance low.

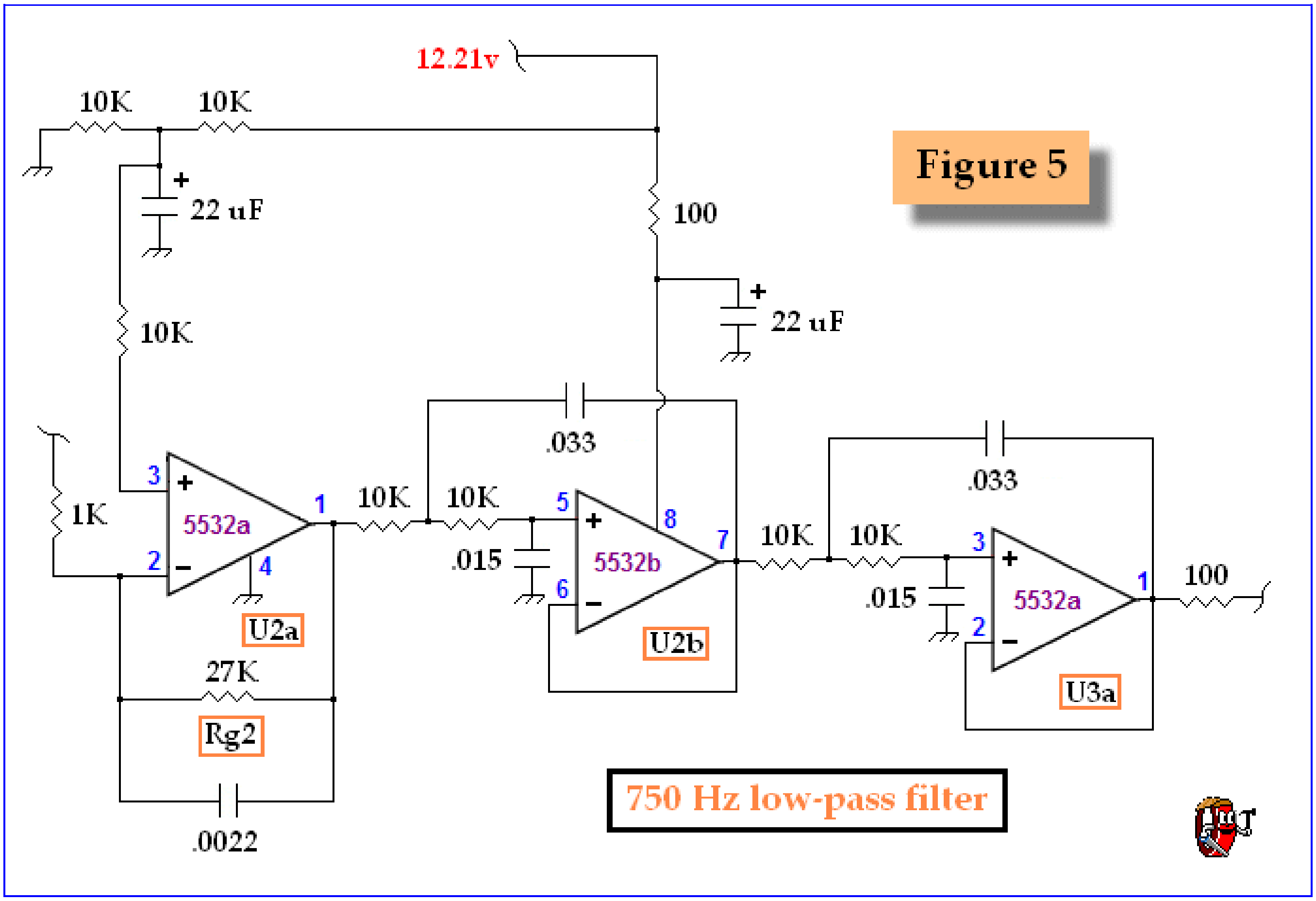
Figure 5

In Figure 5, the remaining preamplifier stage and low-pass filters are shown. Rg2 was chosen for the same reason as Rg1 in Figure 4. Please experiment with these values. The low pass filter stages are scaled up by a factor of 10 as resistance values are less critical at this point in the audio chain. You could use the same resistor and capacitor values used in the first low-pass filter or use the Figure 5 values for all of the low-pass filters. Very often, our parts collections dictate how we experimenters design and build circuits.

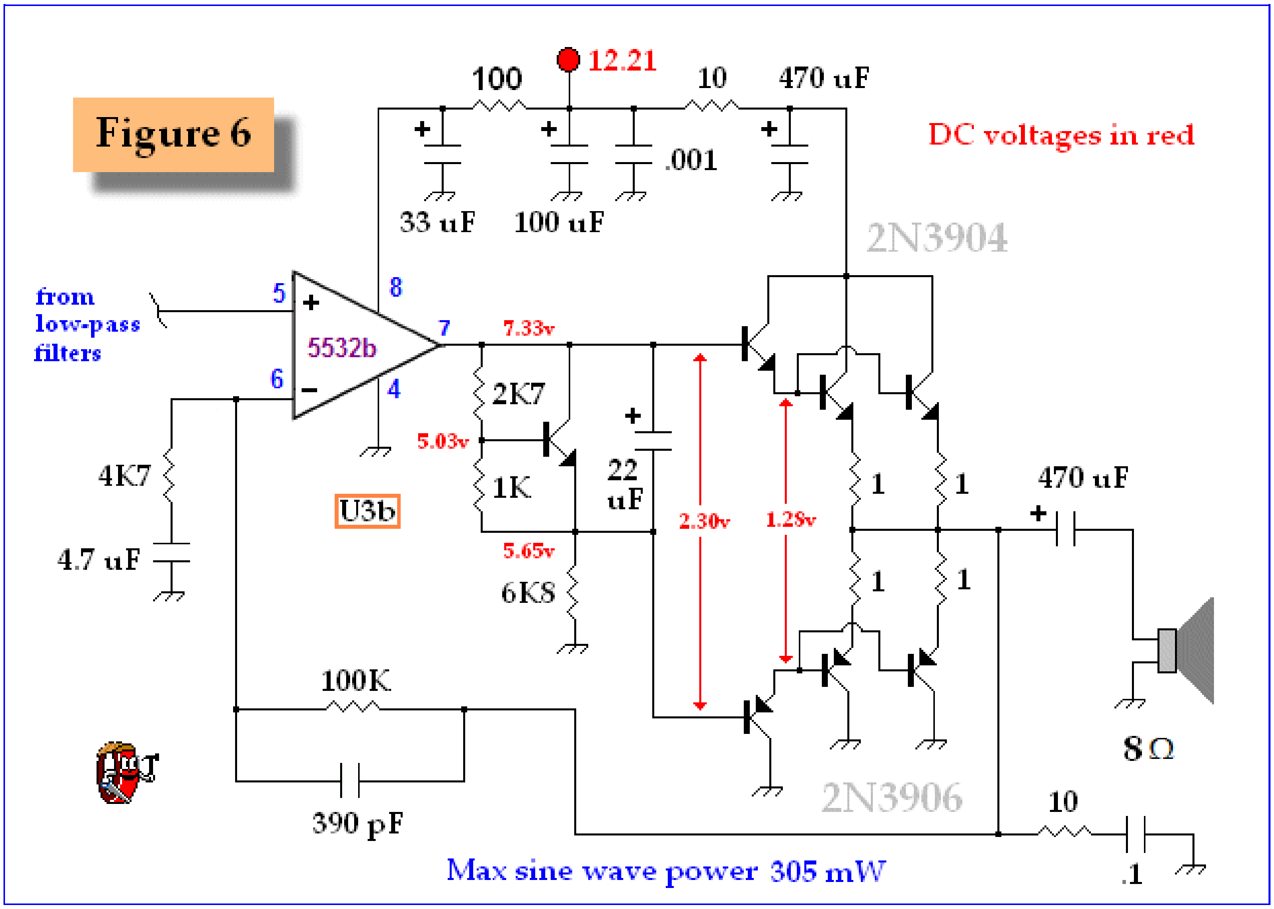
Figure 6
12.21
100
10
470 uF
DC voltages in red
33 uF
100 uF
.001
2N3904
from
low-pass
filters
5
6
8
7
4
5532b
7.33v
2K7
5.03v
1K
22
uF
2.30v
1.28v
1
1
470 uF
4K7
U3b
4.7 uF
5.65v
6K8
1
1
100K
8 Ω
390 pF
2N3906
10
.1
Max sine wave power 305 mW

Figure 6 is the power amplifier schematic. Apart from field portable transceivers where headphones are used to save battery power, I exclusively listen to my receivers via a loudspeaker. This safely allows the exclusion of AGC circuits. You may have noticed that AGC circuits have not been presented on this web site. I rarely use them. The extensive VCC decoupling in Figure 6 is needed in this high gain AF chain. The 1 ohm emitter resistors in the finals invite instability, but were tamed with the 10 ohm / 0.1 uF low pass filter. My original design called for 2.2 ohm emitter resistors, but none were available. This final amplifier is moderately loud and very quiet. When it was first built (the AF chain was built backwards), you could not tell it was on when no signal was applied. With no antenna connected to this receiver and the volume on full, there is only a little noise. This is a good way to test a receiver AF chain for noise. Speaker choice is also important. Speakers of a greater power rating and size sound better; especially when mounted in a wooden cabinet.

Biasing power amps has been discussed extensively on this web site. Ensure you measure and record your quiescent DC voltages as shown in Figure 6. If you hear cross over distortion or the quiescent voltage between the bases of the paralleled final transistors is less than 1.25 volts, try decreasing the 6K8 resistor to 5K6 ohms.

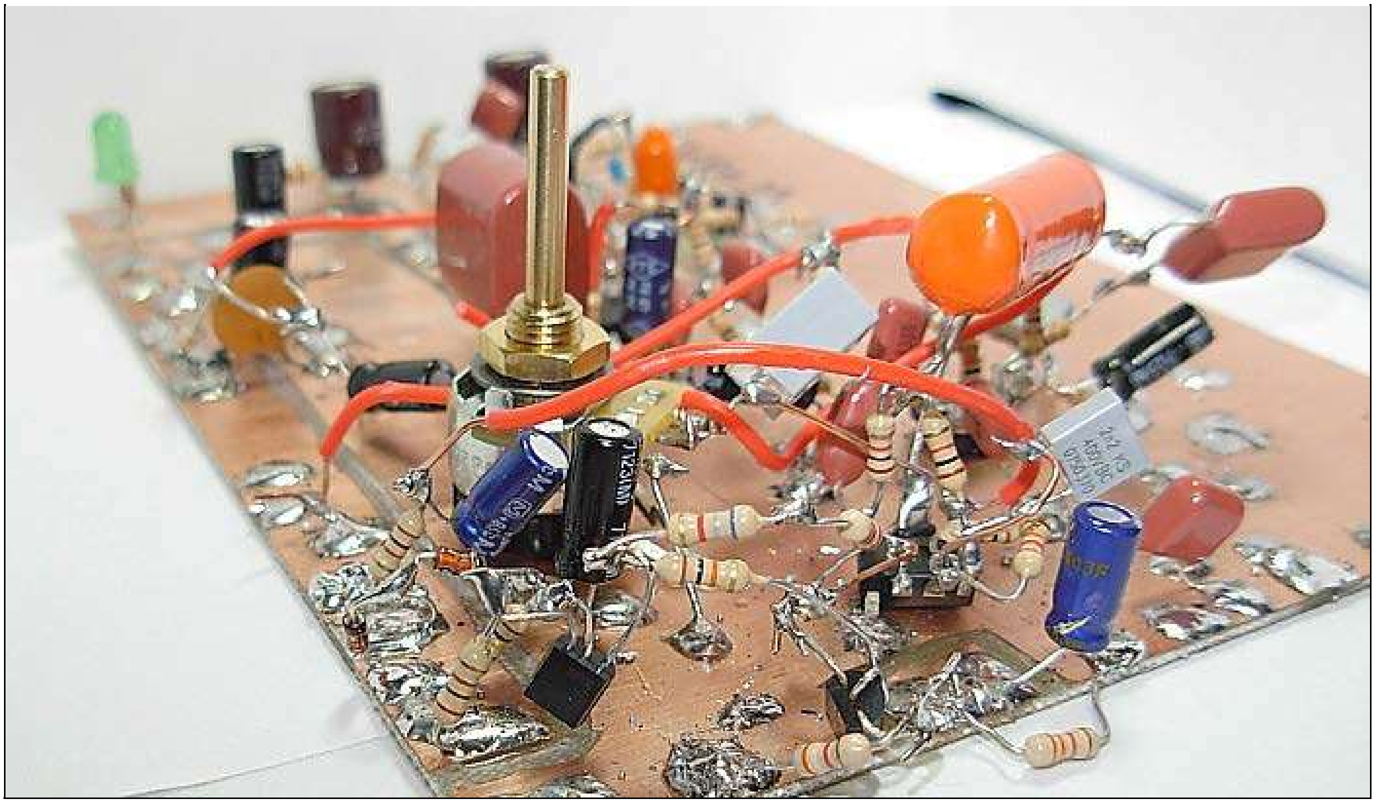

The final preamp breadboard is shown above. You can tell by all the grunge and the solder marks, that many different configurations were trialed. Despite looking haggard, the AF chain is quiet and does not hum nor detect broadcast radio.

Shown above is an early breadboard of Figure 6. In this version, the volume control was at the input and a voltage divider network was wired to pin 5 to provide VCC/2 bias. Testing with an audio signal generator, a tape player and other sources were performed. Later, the biasing resistors and the potentiometer were removed and the Figure 5 stage was added to the copper clad board and tested.

Final Thoughts

We are remembered best by those whom we affect. Certainly Mike's unique perspective and enthusiasm inspired me to dig deeper into this hobby. I asked Bill, M0HBR, for a quote to conclude this web page. Bill wrote this: "Just last week somebody was asking me for background info on Mike's DSB modification of the Heath HW-8... I think it is a real tribute to Mike that years after his passing, hams around the world are still talking about him fondly, still visiting his web site, still following his lead on homebrew radio projects. Certainly among the 3,000 + listeners of the SolderSmoke podcast, Mike is present in spirit every time a soldering iron is heated up." Thank you Bill.

My heartfelt regards to Mike's family.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

VFO Experiments 2009

QRP SWL HOMEBUILDER

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Supplement to the VFO 2009 Web Page

This web page is a supplement to the VFO 2009 Web Page

Figure 10 Bipolar Transistor Notes

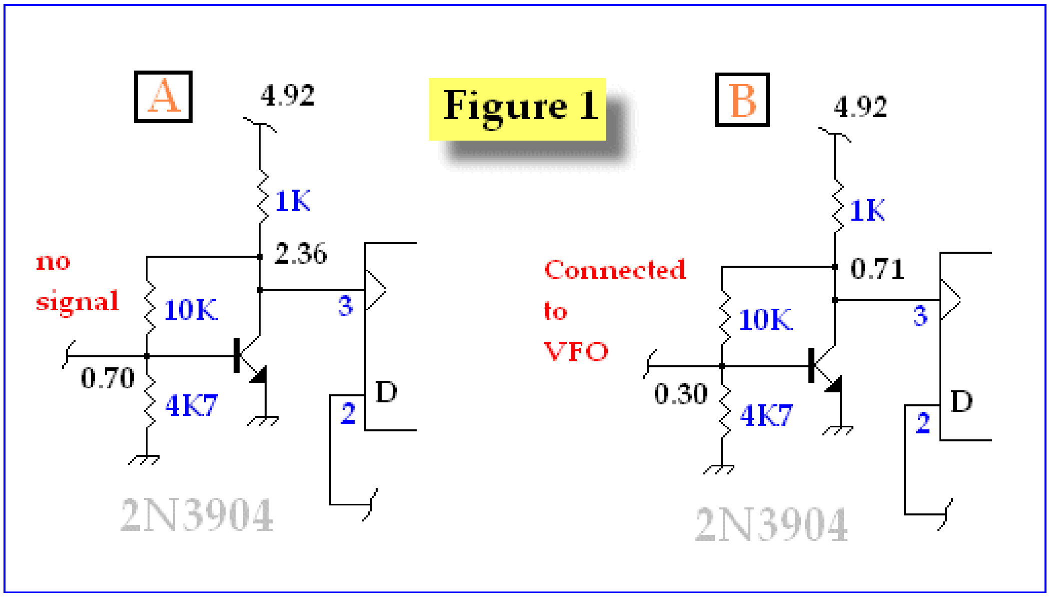

This circuit is from EMRFD and is an analog to digital interface for the VFO. I did not understand the circuit, so I asked Wes, W7ZOI to explain his design criteria and the basic circuit function. This transistor circuit is a driver. The purpose of this transistor driver is to positive edge trigger a D flip-flop. The paramount criterion was that the transistor collector has a quiescent bias voltage that fell between the minimal acceptable logic high and logic low DC voltages on the 74HC74 clock input. See this chart for details about this topic. The correct DC collector voltage is set via feedback from a voltage divider (in this case, a 10K and 4K7 resistor). The collector DC voltage (shown above in Figure 1A) was 2.36 and lies perfectly between the logic low and high state of the 74HC device that it drives.

This transistor circuit is also a high gain amplifier, however, its primary function is a driver. Refer to figure 1B. When connected to the VFO output, the base voltage drops to 0.30 volts. This is an average DC voltage because it fluctuates as the AC waveform swings up and down. More or less AC drive on the input will change this DC voltage.
When connected to the VFO output, the collector voltage dropped to 0.71 VDC. Again, this is just an average, as when the AC swings positive, the the collector voltage will drop down to saturation (where collector voltage is less than base voltage) and then increase towards the positive supply. It does not have to go into cutoff; just to 3.2 volts or so. It also does not have to go into saturation; just to below the 1.3 volts or so. Thus, the transistor remains between cutoff and saturation when appropriately driven with a sine wave.

Our VFO serves as the clock and data on the D flip-flop inputs are only transferred to the flip-flop's outputs on the positive edge of a clock pulse. Positive edge triggering means that the output only responds to the input changes when the clock signal transitions from logic low to logic high. What a fantastic little circuit! You can find additional information and some practical examples in EMRFD. My special thanks to Wes, W7ZOI for helping me to better understand his design.

Positive Edge Triggering

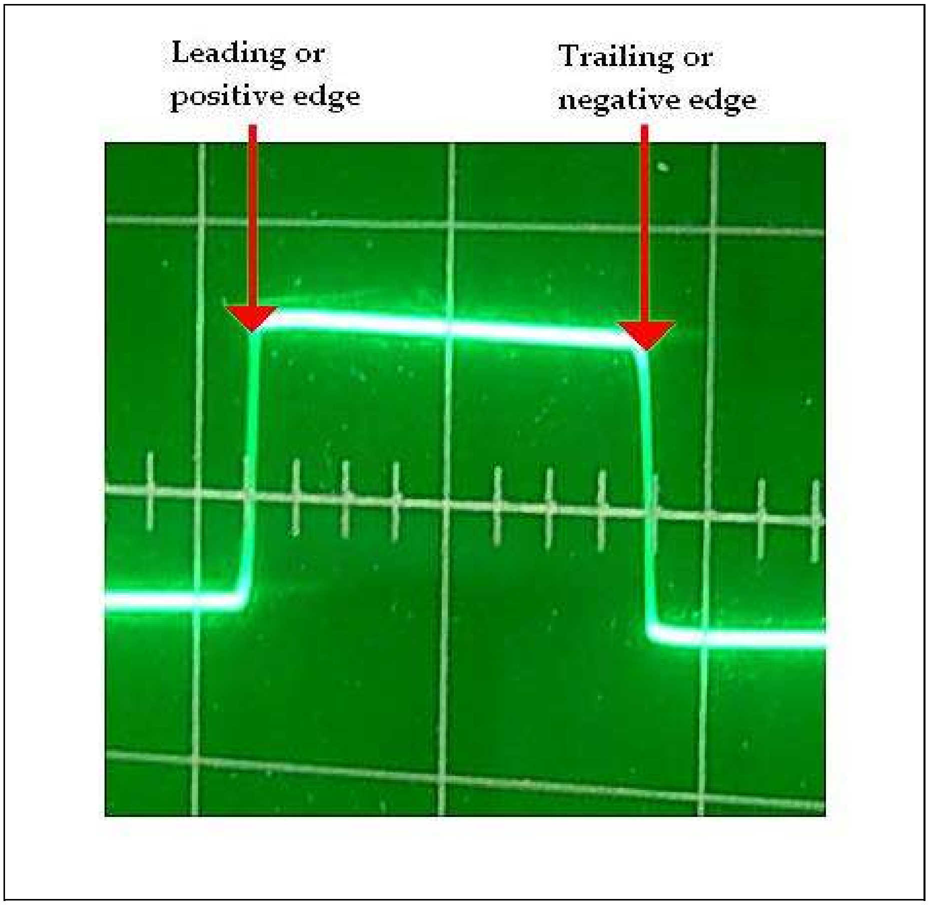
Leading or positive edge
Trailing or negative edge

The triggering of a positive edge triggered flip flop is better understood viewing a graphic like above. The instant the waveform goes positive, the flip flop is triggered. It is easy to understand why a square wave is desirable for clean triggering.

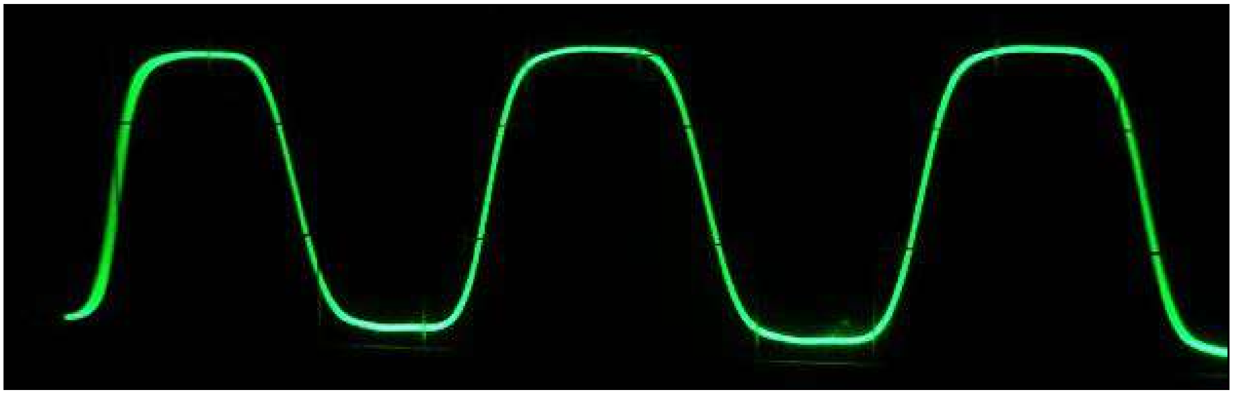

RF — Test and Measurement

Radio Amateurs of Canada Support Page

Commentary for Canadian Radio Amateurs

A significant number of licensed Canadian amateur radio operators do not support their national amateur radio society: Radio Amateurs of Canada Typically support is by membership, but may also include volunteering time to perform the various functions required to run a large, non-profit organization. It has been demonstrated repeatedly, that a small percentage of people are doing most of the work in modern volunteer organizations.

Some thoughts concerning volunteerism

- Volunteerism is down in many sectors including service clubs, meal delivery programs, church groups, volunteer firefighters radio clubs, etc. etc.
- The population is aging and our current volunteers are retiring; Younger folks have so many distractions and things they can do today. We now live in a world with 200 plus TV channels! The spirit of volunteerism seems to be decreasing in the western world.
- There is some evidence that people who donate their cash to charities also tend to participate in groups, associations and organizations. Is this a characteristic or trait?
- People volunteer for causes they believe in. What do people believe in now? Is our culture too self-absorbed, individualistic and focused on our own personal agendas and pleasure?
- Economics. Are young to middle aged people less able to volunteer and contribute because they are spending more of their time working?

RAC Membership

If you haven't, please consider joining the RAC. I joined out of respect to the many RAC volunteers who have devoted their personal time and continue to toil to keep amateur radio viable in Canada. We cannot take our currently allocated radio frequency bands for granted. There are groups who for profit, want to apply technologies which will consume or interfere with the radio frequency bands we enjoy. Canadian amateur radio needs a strong, united voice to survive into the future.

Apathy and lack of awareness may significantly decrease the rights and privileges we enjoy as radio amateurs today. The least we can do is support our national amateur radio society. It would be even better if more of us contributed by volunteering our personal time towards our wonderful radio hobby and serving the RAC.

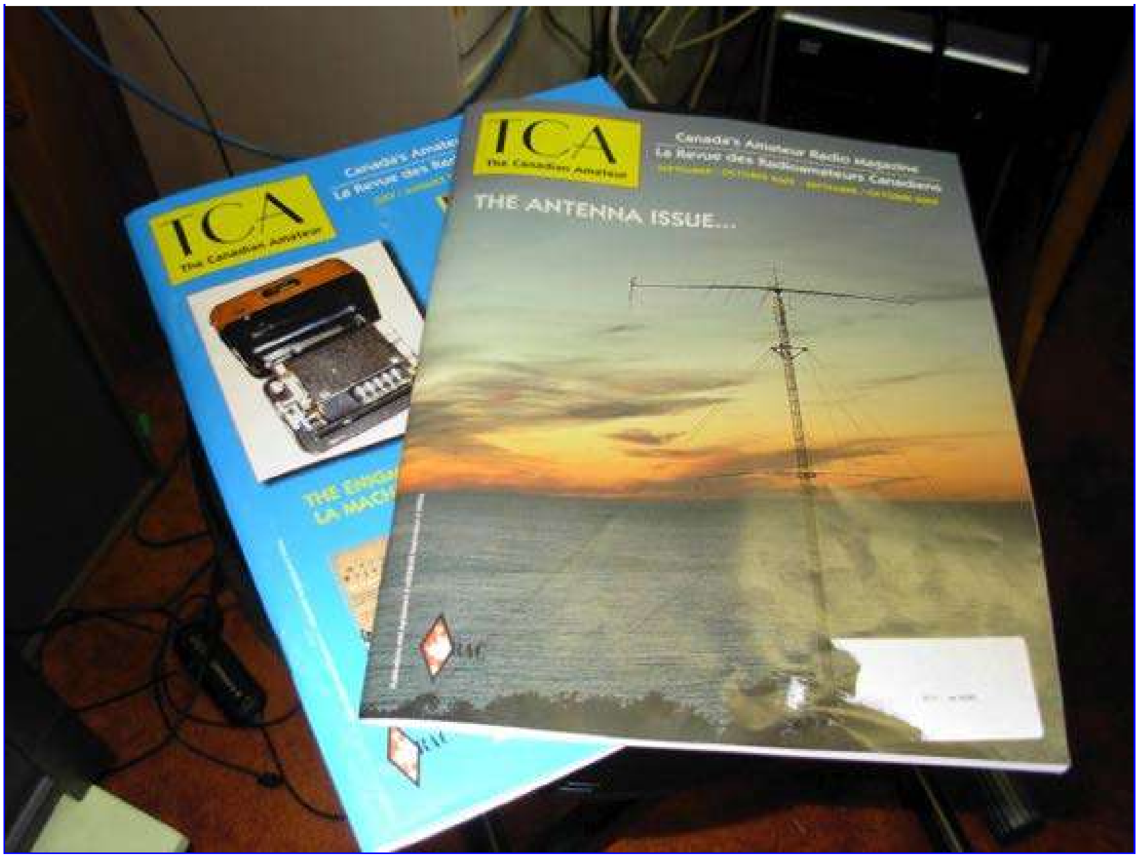
TCA
The Canadian Amateur
Canada's Amateur Radio Magazine
La Revue des Radioamateurs Canadiens
THE ANTENNA ISSUE...

The RAC Journal The Canadian Amateur is an excellent resource. Sept-Oct 2009 marked the first ever antenna issue.

Old QRPHB Web site graphics...

Junk Box Blog

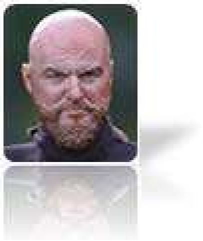

QRP
HOMEBUILDER

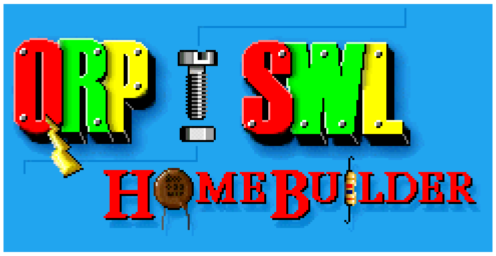
QRP
SWL
HOMEBUILDER

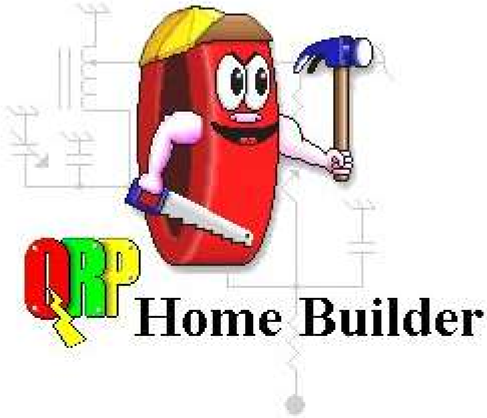
QRP
Home Builder

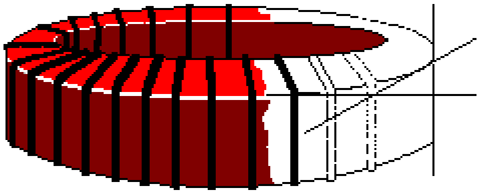

L1
C1
QRP
Homebuilder Homepage
C2
L2

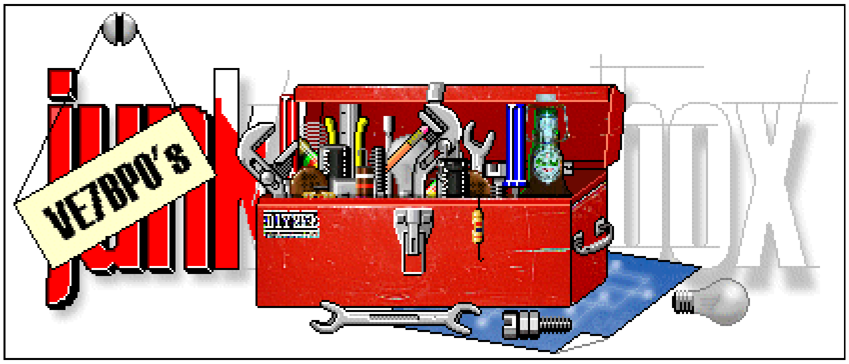
VE7BPO's
junk box

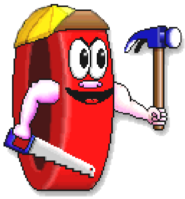

This page last updated: September 21, 2013

Info on the Popcorn Superhet RF Amp

Discussion

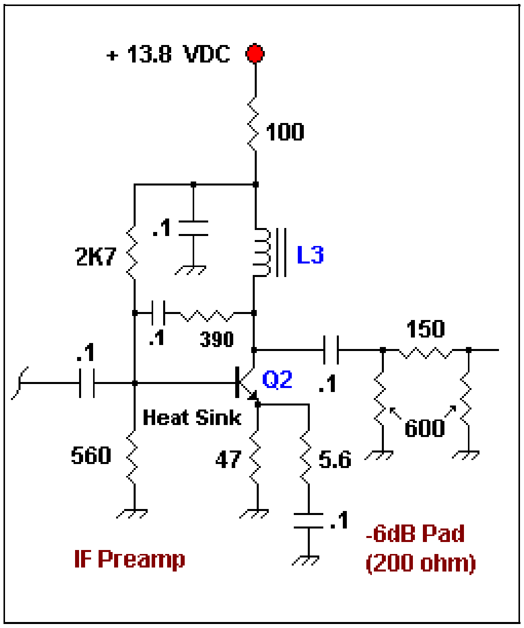

I have received a lot of emails regarding this RF amp from the popcorn superhet receiver. This amp was designed using a calculator and as it turns out, my input and output impedance was off by several ohms from the target 50 ohms input Z and 200 ohms output Z. Today we have computer programs to calculate the resistor values, so crunching the math is not a factor any more. This RF amp is preceded by a 50 ohm diplexer and followed by a 200 ohm pad, so it actually matches the preceding and following stages very well considering that it is clearly depicted to be a popcorn receiver. If I had to redesign this amp in 2008, I would use software to design a CE amp using "noiseless" Norton feedback, Still, this was a good sounding receiver at the time.
Probably the weakest section of this receiver is the audio preamp from Q5 on through to the LM386. I would likely never use these noisy stages in 2008-9. Instead I would use low noise op amps. This web site like my own abilities has grown over time. In the following section, I have paraphrased some writing from Wes, W7ZOI concerning the Q2 RF amp.

An Analysis by W7ZOI

The input impedance of a common emitter bipolar transistor amplifier is related to the standing current, the emitter degeneration resistance, the transistor beta, the F-t, and the operating frequency. Most of those details can be well modeled with the hybrid pi model. If you put quite a bit of emitter degeneration in such an amplifier, the gain drops while the input impedance increases. However, it is much less dependent upon beta and frequency. The emitter degeneration is one form of negative feedback. But now we introduce a second form of negative feedback by applying some of the collector voltage to a resistor that generates a current that is fed back to the base. This feedback tends to reduce gain just as does emitter degeneration. But it also tends to reduce input impedance. By playing a careful balancing act, you can design an amplifier that has a 50 Ohm input and output impedance with a reasonable amount of gain. This particular circuit has a gain of 21.7 dB at 10 MHz and an input return loss of 14 dB. The output return loss is even better at 18.6 dB. Z-in is 39-j14 while Z-out is 172-j34.

These are calculated values using that hybrid pi model. I have devoted a few pages to a discussion of this feedback amplifier topology in the ARRL 2003 book, Experimental Methods in RF Design. This was co-authored by KK7B and W7PUA. You can get the book on line from the ARRL. The book includes computer programs that will handle the design chores. One deals with bias issues while another does the RF design and analysis.

The RF amplifier under discussion may be found on this web page

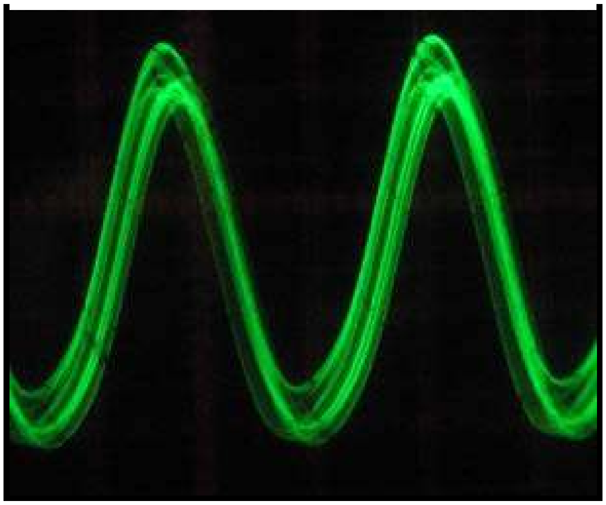

Amateur and Short Wave Radio Electronics Experimenter's Web Site

3 MHz VFO for 40M Popcorn Superhet Receivers

Discussion

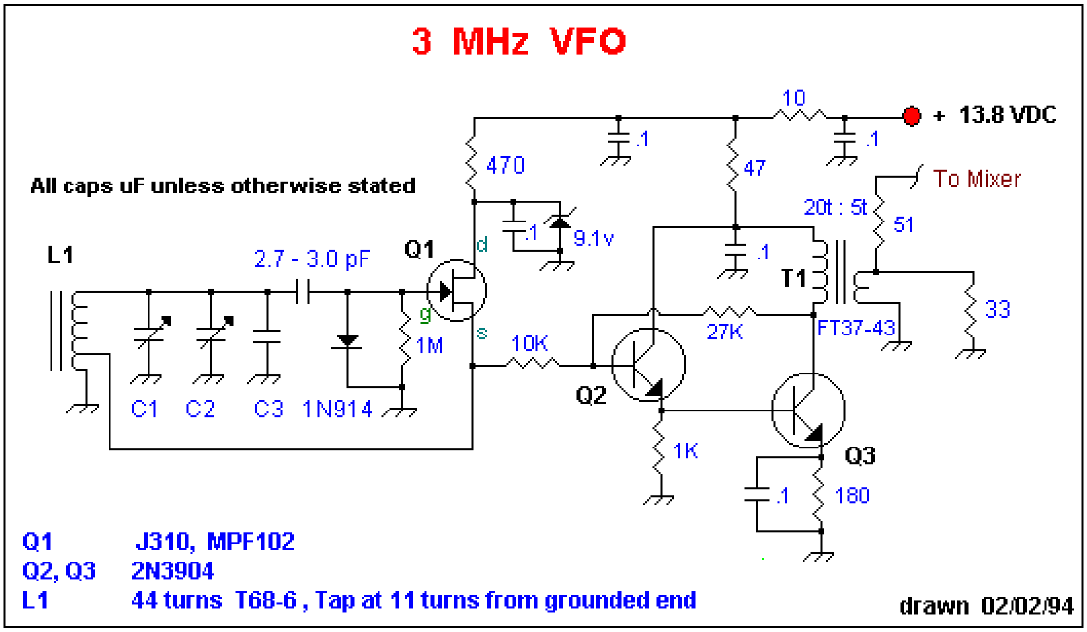

This VFO was first popularized by Roy Lewallen, W7EL and has been used in countless homebrew and commercially sold QRP rigs. For the 40 meter ugly super hets, you have a choice of a VFO frequency of 3 MHz or 11 MHz and I chose the former for this website. For L1, you can use the T68-6 as shown or redesign your own inductor using a T50-6 core. The T68-6 core allows the builder to wind the coil using # 24 AWG wire for enhanced Q and stability. The main tuning capacitor, C1 that was used had a built in reduction drive and went from 4 - 19 pF. If you use a tuning capacitor with a greater capacitance swing, you may have to connect it to the top of L1 via a small-value NP0 capacitor to reduce the tuning frequency range. This can be done using math or by just plain experimentation. C2 is a small ceramic air-variable trimmer from my junk box used to set the lower band edge of the VFO. It can be omitted if the user wants to go frugal and experimentally set the lowest frequency of the VFO using small-value ceramic NP0 caps. Air variable caps for C1 and C2 are mandatory for minimal drift VFO operation. C3 actually refers to 4 NP0 ceramic caps which were used to place the VFO on the correct frequency. Four caps were used to minimize heating and thus drift in the VFO. For the prototype VFO which tunes something ~ 3.00 to 3.67 MHz , I used the mentioned air-variable, four NP0 ceramic caps (20pF,100pF,5pF,100pF) and C2 the trimmer cap that went from 2 - 50 pF. These values should be used only as a guide for prospective builders as there are many variables at play. Following a 10 minute warm up period, the VF0 frequency stability is excellent. Again, keeps all component leads as short as possible. T1 is a broadband transformer wound by using 20 turns of # 26 AWG over a FT37-43 toroid core and then distributing 5 secondary turns spaced evenly over the primary windings. Do not omit the 33 ohm load resistor. The 2.7 to 3.0 pF coupling cap should also be of the NP0 ceramic type. Q1 can be the J310 (my favorite) , 2N4416 , MPF102 or other correct substitutes.

This VFO should be in a separate sealed, shielded box from the rest of the receiver

.

VFO Stability Hints

Numerous tips for enhancing VFO stability have been generated from Wes Hayward, W7ZOI, Roy Lewallen, W7EL, Doug DeMaw, W1FB and others. Here are a few summarized:

1. Use air wound or powdered-iron toroidal inductors made from number 6 material.
2. Use the heaviest gauge of wire possible to wind your inductor.
3. Anneal the inductor by boiling it in water for around 5 minutes after winding.
4. Use ugly construction above a SINGLE-SIDED copper side up ground plane.

5. Capacitors in the L-C circuit should be NP0 ceramic type.
6. Use air variable capacitors for tuning that have double bearings and no backlash.
7. The VFO should be operated at a lower regulated voltage.
8. Do not use cheap low-Q trimmer caps.
9. Completely encase the VFO to prevent RF leakage and to minimize environmental temperature changes.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

4 MHz BFO for 40M Popcorn Superhet Receivers

Discussion

Shown is a 2 stage BFO for use with the Ugly 40 meter superhets with a 4.00 MHz IF. This is a design by W7ZOI which I have used from 3.5 to 9 MHz by only changing the crystal and the Q2 output coil and capacitor values to suit the frequency of choice. The output coil is 59 primary and 12 secondary windings on a T68-2 toroid core. You will need in the order of 79 pF to resonate this coil, thus a 68 pF plus a 5-20 pF variable works well. If you do not have a T68-2 core, a T50-2 could be used with 30 AWG wire which is a bit difficult so you could opt for an XL of 200 ohms which would be 40 primary and 8 secondary turns on a T50-2 core. You would need ~200 pF to resonate such a coil and the BFO would have reduced output power, but would still work okay. If you design your own tank, you probably should not use an XL/XC less than 190 ohms for the output stage at 4 MHz. Use a 5:1 turns ratio for primary to secondary windings.

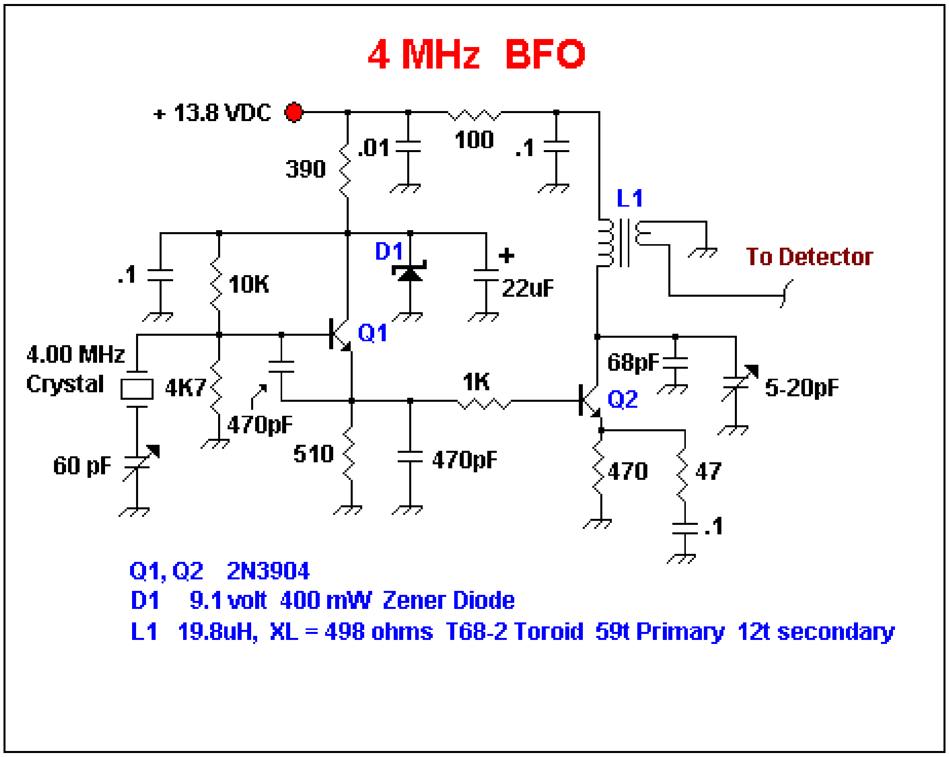
4 MHz BFO
+ 13.8 VDC
390
.01
100
.1
L1
To Detector
D1
+
22uF
.1
10K
Q1
4.00 MHz
Crystal
4K7
470pF
510
1K
470pF
68pF
Q2
5-20pF
60 pF
470
47
.1
Q1, Q2 2N3904
D1 9.1 volt 400 mW Zener Diode
L1 19.8uH, XL = 498 ohms T68-2 Toroid 59t Primary 12t secondary

It wouldn't hurt to follow the Popcorn Superhet BFO with a low pass filter. A simple pi-type 3 element filter is suitable. Since I had a couple of 470 pF caps left over, I made a low pass filter using a 6.8 MHz cutoff frequency with a 1.17 uH inductor. A 1.17 uH inductor using a T37-2 core requires 17 turns while on a T50-2 core requires 15 turns. Caps were inexpensive monolithic ceramics from Digi-Key. The filter is shown below, L1 = 1.17uH , C1 and C2 = 470 pF. If you want, you may use other cutoff frequencies to suit any junk box caps you have on hand and use an XL/XC of 50 ohms.

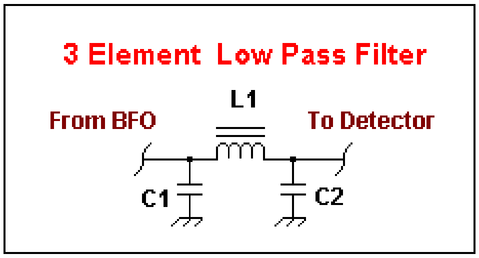
3 Element Low Pass Filter
From BFO
L1
To Detector
C1
C2

Construction

Keep component leads as short as possible to promote stability. Do not omit the zener diode D1. I thought of using a small ferrite bead on the base of Q2 and will try it on the next version of this BFO that I make. The BFO, if unstable can break into all sorts of weird AF sound effects that are very annoying. The BFO should ideally be in a shielded box, but many amateurs have good results by just placing the BFO strategically in the receiver chassis. When winding L1, make sure you wind the secondary coil so that the secondary's grounded end is on the cold side of the primary coil that connects to the 100 ohm resistor and positive voltage. I just pick one end of the primary coil and close wind the secondary coil from this point in the same direction for the correct number of windings. Then the starting point of the secondary can be grounded and the other end connected to the detector or a low pass filter. After building the coil, temporarily place a 51 ohm resistor from the secondary output lead to ground. Then you can test and peak the tuning of the output tank into a load. Leave the resistor in place for all your testing work such as crystal matching or RF stage testing until you are ready to connect it to the product detector. The 60 pF variable cap connected between the crystal and ground is used to set the beat frequency of the BFO. Adjust it to get a good sounding beat note when listening to a station. You will notice that the beat frequency can be tuned roughly by just listening to 40 meter band noise in the headphones and then tweaked on an active QSO.

QRPHB Archive Schematics

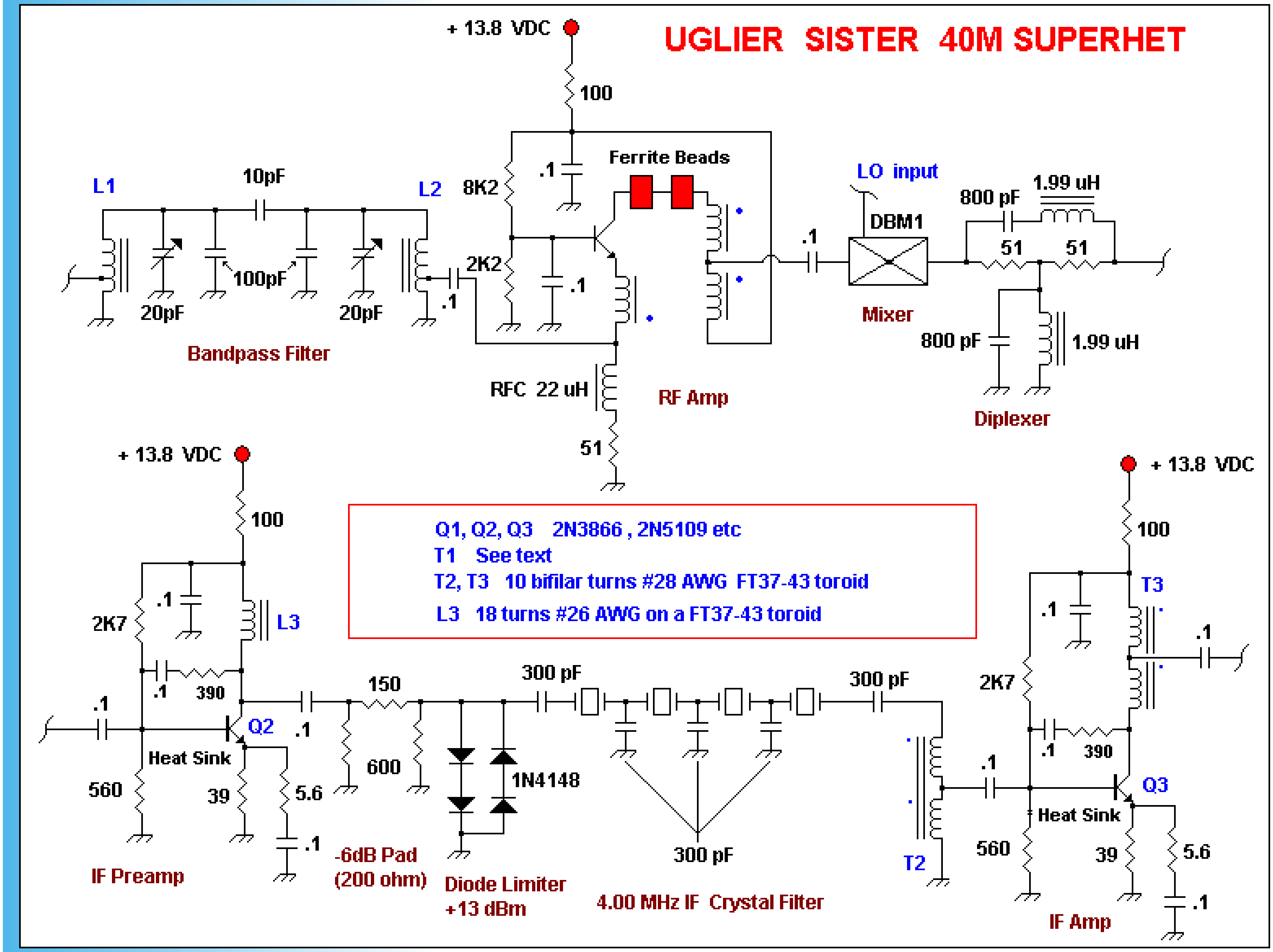
UGLIER SISTER 40M SUPERHET
+ 13.8 VDC
100
L1
10pF
L2
8K2
.1
Ferrite Beads
LO input
800 pF
1.99 uH
DBM1
.1
51
51
2K2
.1
100pF
20pF
20pF
.1
Mixer
800 pF
1.99 uH
Bandpass Filter
RFC 22 uH
RF Amp
Diplexer
51
+ 13.8 VDC
100
+ 13.8 VDC
100
T3
.1
2K7
.1
L3
.1
2K7
.1
390
300 pF
150
300 pF
.1
Q2
.1
Heat Sink
600
1N4148
.1
390
560
39
5.6
.1
Q3
.1
Heat Sink
-6dB Pad
(200 ohm)
300 pF
T2
560
39
5.6
IF Preamp
Diode Limiter
+13 dBm
4.00 MHz IF Crystal Filter
.1
IF Amp

Q1, Q2, Q3 2N3866 , 2N5109 etc
T1 See text
T2, T3 10 bifilar turns #28 AWG FT37-43 toroid
L3 18 turns #26 AWG on a FT37-43 toroid

<\p>

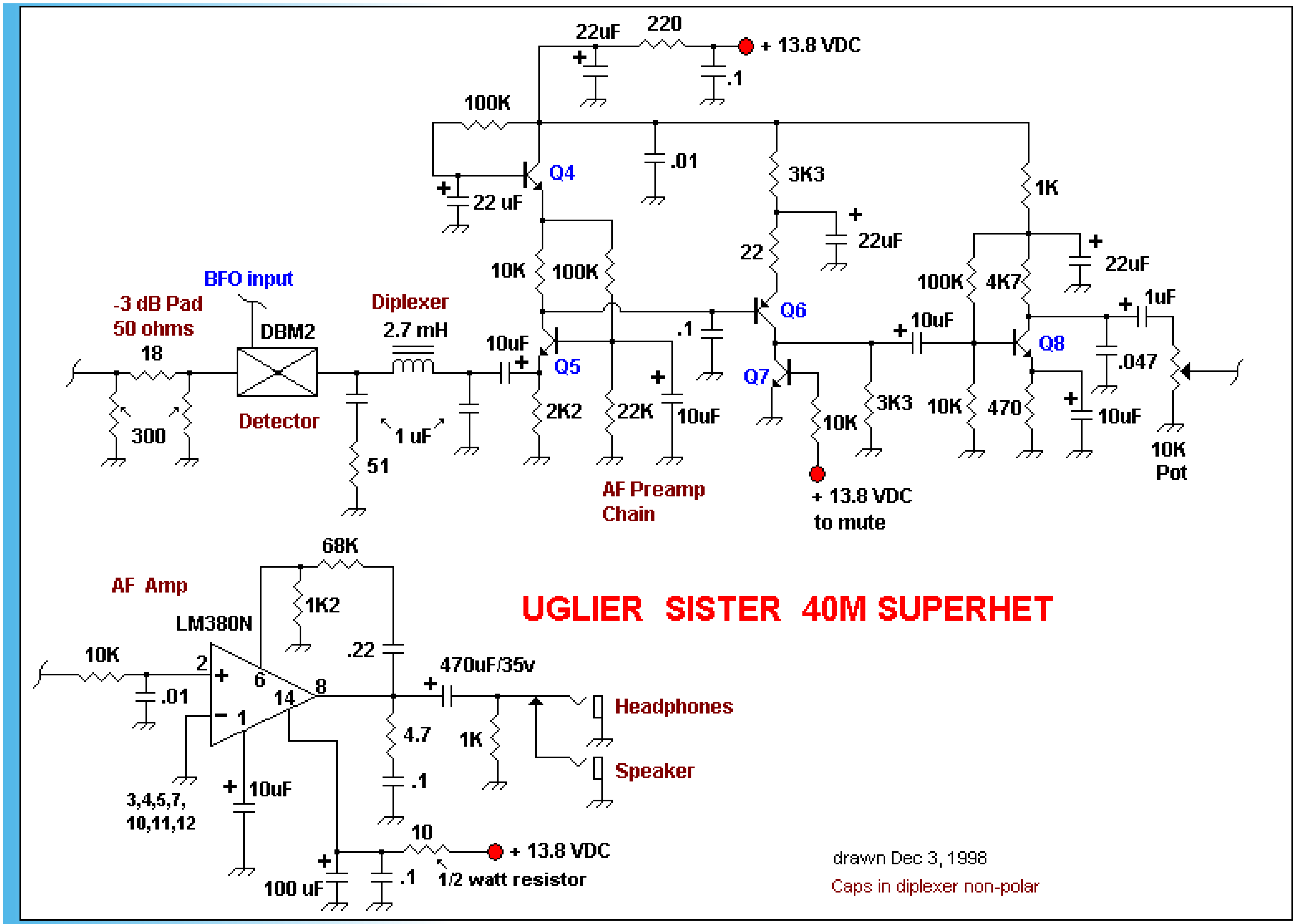

UGLIER SISTER 40M SUPERHET
+ 13.8 VDC
22uF
220
.1
100K
Q4
22 uF
.01
3K3
1K
22
22uF
BFO input
-3 dB Pad
50 ohms
18
DBM2
Diplexer
2.7 mH
10K
100K
Q6
100K
4K7
1uF
10uF
Q8
10uF
Q5
.1
Q7
.047
300
Detector
1 uF
2K2
22K
10uF
10K
3K3
10K
470
10uF
10K
Pot
51
AF Preamp
Chain
+ 13.8 VDC
to mute
68K
AF Amp
1K2
LM380N
10K
.22
470uF/35v
.01
2
6
14
8
1
Headphones
4.7
1K
Speaker
.1
3,4,5,7,
10,11,12
10uF
10
+ 13.8 VDC
100 uF
.1
1/2 watt resistor
drawn Dec 3, 1998
Caps in diplexer non-polar

<\p>

Uglier Sister Superhet Receiver

This is a 40 M superhet that I built a few years ago. While I was happy with this receiver, it has been discontinued from the main listings on the QRPHB website. I have had a couple requests for this schematic, so I have created this archival web page.
I stood a lot of current in the IF preamp and main IF amplifiers, which is not really necessary, particularly in the case of the post-filter amp. Of course raising the emitter resistor values is easy to do in these amps if you do wish to lower the current. There also is no audio filtering in the AF preamp stages, which is something that I often use in better quality receivers that I build now.

VFO and BFO

[VFO schematic]
[BFO schematic]

Construction Information

Refer to this web page for details on the xtal filter, diplexer and construction information.

The Norton RF preamp is discussed on the RF Preamp web page.

Return to the QRP HomeBuilder Homepage

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Diplexer Supplemental Page

More Information on the Bridge-Tee Diplexer

Introduction

No web page on the QRPHB web site has furnished more email and controversy than the Diplexer Web page. This page has received much work to improve its content and accuracy and the email received has been greatly appreciated. Presented is the detailed math and some information regarding the Bridge-Tee Diplexer by Helmut Strickner, VK4STR and Wes Hayward, W7ZOI.

VK4STR

The T-bridge diplexer math calculation (also known as the W1JR diplexer)

The formulae presented and used in the program for the Bridge-T Diplexer uses a Q of 1. Using a Q = 1 resonator is not optimal because we don't want a "broad peaking response" of the series LC resonator in the T-bridge diplexer. Our goal should be to only let the desired IF frequency pass and attenuate all other frequencies as much as possible by properly terminating them into 50 Ohms in the diplexer.

The optimal Q for HF frequencies is around 10, less for higher frequencies. The higher Q of the series resonator effectively increases the attenuation of undesired harmonic frequency components from the mixer IF-port, and the reflected products from the crystal back to the IF-port. The IM performance (IP3) of the receiver is improved.

The possible improvement in IP3 depends mainly on the following stages. The post-mixer amplifier following the diplexer has to handle huge impedance variations at the input of the crystal filter. Crystals do not absorb but reflect any products they cannot handle. These products are then reflected back to the IF-port of the mixer deteriorating the IM performance.

A low Q diplexer lets these reflections from the crystal pass unhindered and most mixers will react with degraded intermodulation performance. It is also important to note that a good termination insensitive post-mixer amplifier should follow the diplexer.

The best results for a high intercept (IP3) receiver can be obtained by choosing:

1. High level mixer
2. T-bridge diplexer with optimised Q (Q=10 for HF)
3. Termination insensitive post-mixer amplifier (see N6NWP - QEX July 1995)

I believe that the intermodulation performance of many receivers can be improved by some simple measures like a properly designed diplexer. The T-bridge diplexer from W1JR is the best one around, go for it.

A brief example of the calculation for a 10.7MHz diplexer follows in Figure 1. The value of K should be 10 for short wave and 2 to 4 for VHF (2 meter band). K is the ratio of L to C.

I hope that you can use this example to calculate the values for other diplexer frequencies.

To fully understand the basis of this calculation please refer to Meinke/Grundlach: "Taschenbuch der Hochfrequenztechnik", Springer-

Verlag Berlin/Heidelberg. Page 127.

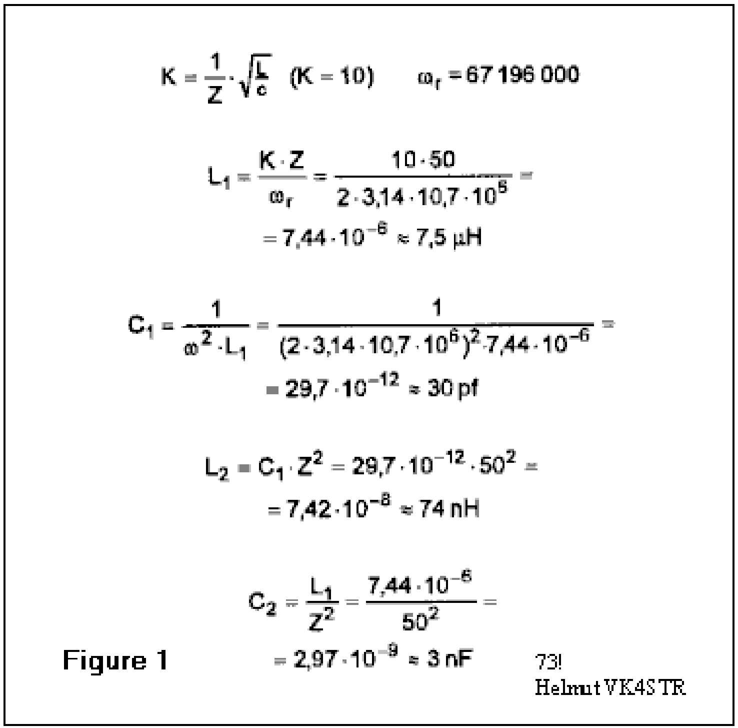

$$K = \frac{1}{Z} \cdot \sqrt{\frac{L}{c}} \quad (K = 10) \qquad \omega_r = 67\,196\,000$$

$$L_1 = \frac{K \cdot Z}{\omega_r} = \frac{10 \cdot 50}{2 \cdot 3{,}14 \cdot 10{,}7 \cdot 10^6} =$$
$$= 7{,}44 \cdot 10^{-6} \approx 7{,}5\ \mu H$$

$$C_1 = \frac{1}{\omega^2 \cdot L_1} = \frac{1}{(2 \cdot 3{,}14 \cdot 10{,}7 \cdot 10^6)^2 \cdot 7{,}44 \cdot 10^{-6}} =$$
$$= 29{,}7 \cdot 10^{-12} \approx 30\ pf$$

$$L_2 = C_1 \cdot Z^2 = 29{,}7 \cdot 10^{-12} \cdot 50^2 =$$
$$= 7{,}42 \cdot 10^{-8} \approx 74\ nH$$

$$C_2 = \frac{L_1}{Z^2} = \frac{7{,}44 \cdot 10^{-6}}{50^2} =$$
$$= 2{,}97 \cdot 10^{-9} \approx 3\ nF$$

Figure 1

73!
Helmut VK4STR

W7ZOI

Attached are some outputs from Mathcad, Version 7.0. The "proper" equations are derived, along with some other results. Feel free to put these on the web, although this is mathematics rather than the usual "formulas" and there might not be any interest in such stuff.

As it turns out, the formula that Todd presented in completely accurate for the case of Q=1. That is, both the series and parallel tuned circuits have inductors and capacitors with reactance of 50 Ohms. Evaluation then shows that S11=0 for all frequency, while S21 peaks broadly at the resonant frequency. If you pick a higher Q, the inductor in the series tuned circuit increases by Q and the capacitor in the parallel one increases by Q. Resonance is maintained in both. Again, you then get a perfect match at all frequencies and a S21 peak at resonance. But I don't know if this is what Joe had in his original work.

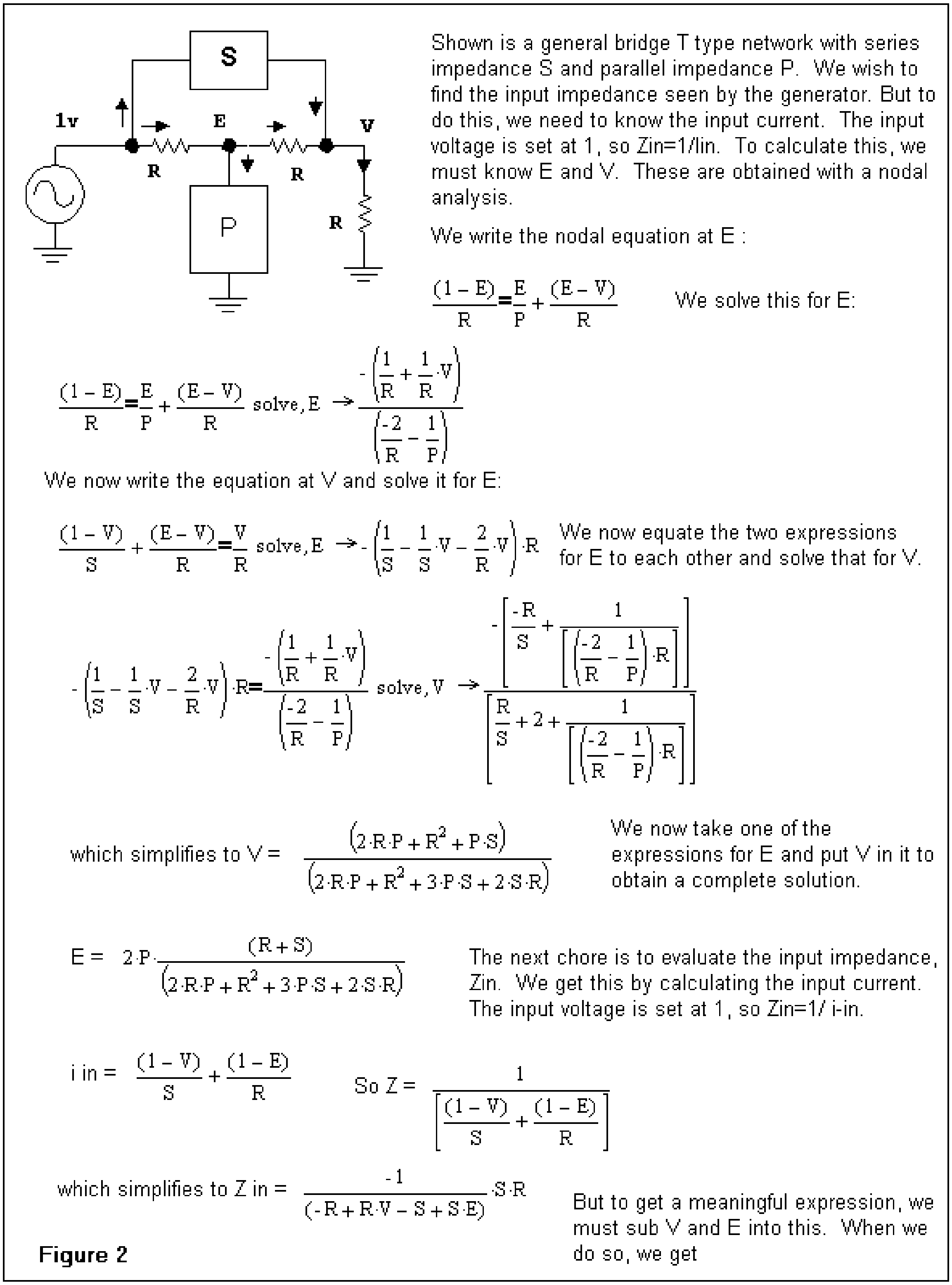

Shown is a general bridge T type network with series impedance S and parallel impedance P. We wish to find the input impedance seen by the generator. But to do this, we need to know the input current. The input voltage is set at 1, so Zin=1/Iin. To calculate this, we must know E and V. These are obtained with a nodal analysis.

We write the nodal equation at E :

$$\frac{(1-E)}{R}=\frac{E}{P}+\frac{(E-V)}{R}$$

We solve this for E:

$$\frac{(1-E)}{R}=\frac{E}{P}+\frac{(E-V)}{R}\ \text{solve},E\ \rightarrow \frac{-\left(\frac{1}{R}+\frac{1}{R}\cdot V\right)}{\left(\frac{-2}{R}-\frac{1}{P}\right)}$$

We now write the equation at V and solve it for E:

$$\frac{(1-V)}{S}+\frac{(E-V)}{R}=\frac{V}{R}\ \text{solve},E\ \rightarrow -\left(\frac{1}{S}-\frac{1}{S}\cdot V-\frac{2}{R}\cdot V\right)\cdot R$$

We now equate the two expressions for E to each other and solve that for V.

$$-\left(\frac{1}{S}-\frac{1}{S}\cdot V-\frac{2}{R}\cdot V\right)\cdot R=\frac{-\left(\frac{1}{R}+\frac{1}{R}\cdot V\right)}{\left(\frac{-2}{R}-\frac{1}{P}\right)}\ \text{solve},V\ \rightarrow \frac{-\left[\frac{-R}{S}+\frac{1}{\left[\left(\frac{-2}{R}-\frac{1}{P}\right)\cdot R\right]}\right]}{\left[\frac{R}{S}+2+\frac{1}{\left[\left(\frac{-2}{R}-\frac{1}{P}\right)\cdot R\right]}\right]}$$

which simplifies to $V = \dfrac{\left(2\cdot R\cdot P+R^2+P\cdot S\right)}{\left(2\cdot R\cdot P+R^2+3\cdot P\cdot S+2\cdot S\cdot R\right)}$

We now take one of the expressions for E and put V in it to obtain a complete solution.

$$E = 2\cdot P\cdot\frac{(R+S)}{\left(2\cdot R\cdot P+R^2+3\cdot P\cdot S+2\cdot S\cdot R\right)}$$

The next chore is to evaluate the input impedance, Zin. We get this by calculating the input current. The input voltage is set at 1, so Zin=1/ i-in.

$$i\ in = \frac{(1-V)}{S}+\frac{(1-E)}{R}\qquad \text{So } Z = \frac{1}{\left[\frac{(1-V)}{S}+\frac{(1-E)}{R}\right]}$$

which simplifies to $Z\ in = \dfrac{-1}{(-R+R\cdot V-S+S\cdot E)}\cdot S\cdot R$

But to get a meaningful expression, we must sub V and E into this. When we do so, we get

Figure 2

$$\text{Z in} = \frac{\left(2\cdot R\cdot P + R^2 + 3\cdot P\cdot S + 2\cdot S\cdot R\right)}{\left(2\cdot R\cdot P + 3\cdot R^2 + P\cdot S + 2\cdot S\cdot R\right)}\cdot R$$

This is now a useful result. We can, for example, let S, P, and R be resistor values. We can then design bridge T attenuators. But let's continue with the analysis. Let's ask the question "What are the conditions that will provide an input impedance of R?" We merely set the Zin expression equal to R, and then solve that equation for S or P. We will solve for P.

$$\frac{\left(2\cdot R\cdot P + R^2 + 3\cdot P\cdot S + 2\cdot S\cdot R\right)}{\left(2\cdot R\cdot P + 3\cdot R^2 + P\cdot S + 2\cdot S\cdot R\right)}\cdot R = R \text{ solve}, P \rightarrow \frac{R^2}{S}$$

Hey, that's a cool result, and it's even rather simple. This is an expression for P in terms of S. If we use it for attenuators, it says that our pad is matched with these conditions. We can combine these results with the equation for V to get the attenuation. A couple of interesting matched pads are shown:

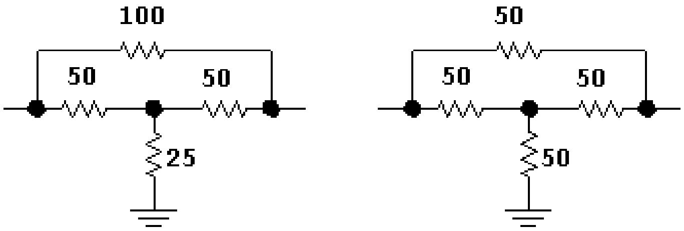

Figure 3

The expression for V was $\frac{\left(2\cdot R\cdot P + R^2 + P\cdot S\right)}{\left(2\cdot R\cdot P + R^2 + 3\cdot P\cdot S + 2\cdot S\cdot R\right)}$ But the conditions for a match were

$P = \frac{R^2}{S}$ So, we put this in for P.

$$\frac{\left(2\cdot R\cdot P + R^2 + P\cdot S\right)}{\left(2\cdot R\cdot P + R^2 + 3\cdot P\cdot S + 2\cdot S\cdot R\right)} \text{ substitute}, P = \frac{R^2}{S} \rightarrow \frac{\left(2\cdot \frac{R^3}{S} + 2\cdot R^2\right)}{\left(2\cdot \frac{R^3}{S} + 4\cdot R^2 + 2\cdot S\cdot R\right)}$$

which simplifies to $V = \frac{R}{(R + S)}$ So the two pads have voltage outputs of 0.5 and 0.333 and the resulting attenuations are 6 and 9.54 dB.

But enough about the attenuators. Let's get back to the original question, that of the diplexer. The circuit of interest is one where the series network is a series tuned circuit while the parallel one is a parallel tuned circuit. But we want to know the conditions that will generate a perfect match. With such a circuit, it is clear that a match will exist at resonance so long as both circuits are resonant at the same frequency. The impedance of the series tuned circuit goes to zero while that of the parallel TC becomes infinite (neglecting loss.)

We write the susceptance (imaginary part of admittance) directly for the parallel tuned circuit as

$B_p = \omega\cdot C_p - \frac{1}{\omega\cdot L_p}$ But the reactance is the recipricol of this, $\frac{1}{\left(\omega\cdot C_p - \frac{1}{\omega\cdot L_p}\right)}$

which simplifies to become $\frac{1}{\left(\omega^2\cdot L_p\cdot C_p - 1\right)}\cdot\omega\cdot L_p$

We know that this is resonant at the same frequency as the series tuned circuit with L and C. So, $L\cdot C = L_p\cdot C_p$ and $L_p = \frac{L\cdot C}{C_p}$ We substitute these in the above expression:

$X_p = \frac{1}{\left[\omega^2\cdot(L\cdot C) - 1\right]}\cdot\omega\cdot\frac{L\cdot C}{C_p}$ But earlier analysis showed us that this reactance was related to the series reactance:

$P = \frac{R^2}{S}$ with $S = \omega\cdot L - \frac{1}{\omega\cdot C}$ This lets us set up an equation that is then solved for Cp.

$$\frac{1}{\left[\omega^2\cdot(L\cdot C) - 1\right]}\cdot\omega\cdot\frac{L\cdot C}{C_p} = \frac{R^2}{\left(\omega\cdot L - \frac{1}{\omega\cdot C}\right)} \text{ solve}, C_p \rightarrow \frac{L}{R^2}$$

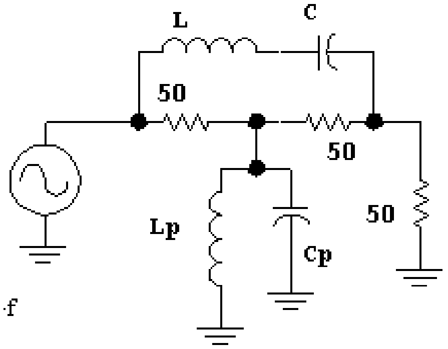

This is a particularly simple, yet useful result

1) Pick a crossover frequency f, Hz. Then, $\omega = 2\cdot\pi\cdot f$

2) Pick a series inductor L. Use one with a reactance of around 50 Ohms if you want a very broad response. Make it larger for narrower bandwidth.

3) Resonate the series L with a cap, $C = \frac{1}{\omega^2\cdot L}$

4) Pick a parallel cap $C_p = \frac{L}{R^2}$

5) Resonate that cap with a parallel inductor $L_p = \frac{1}{\omega^2\cdot C_p}$

Figure 4

This web page is copyrighted 1998-2005

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Astable Multivibrator Projects

Discussion:

Astable or free-running multivibrators have been used in home-built amateur radio equipment for many years. The basic circuit is a two stage amplifier with AC-coupled feedback from output to input. One transistor stage is on (conducting current) while the other is off (not conducting current) until the stages switch conducting states repeatedly at a specific frequency. The oscillation frequency is set by the resistor and capacitor values connected to the base terminal of each stage. This RC network determines how long the transistor stays in the off position.

Presented are two projects which utilize astable multivibrators built using the ubiquitous 2N3904 BJT. The first project is a code practice audio-frequency oscillator while the second is a simple , no-frills electronic keyer for keying a transmitter. Either circuit would be a great first project to learn how to build circuits using Ugly Construction.

Code Practice Oscillator

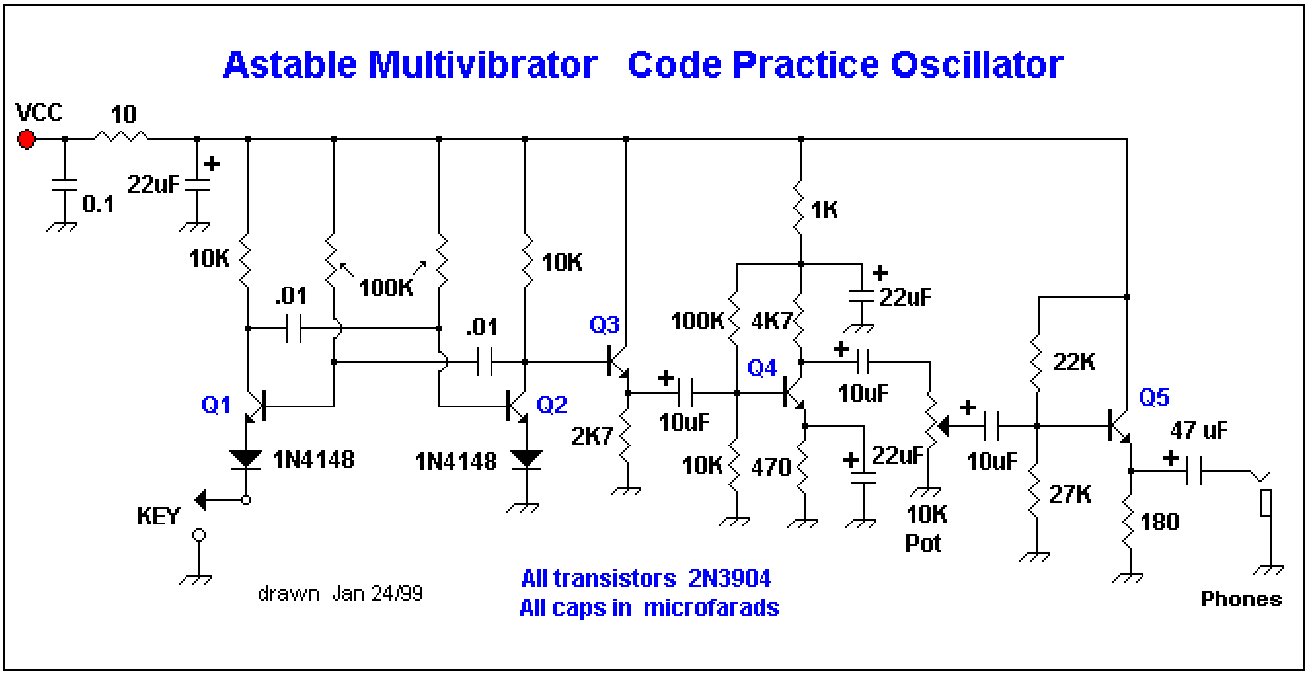
Astable Multivibrator Code Practice Oscillator
VCC
10
0.1
22uF
+
10K
100K
10K
.01
.01
Q1
Q2
Q3
Q4
Q5
1N4148
1N4148
KEY
2K7
10uF
1K
22uF
100K
4K7
10uF
10K
470
22uF
10K Pot
22K
27K
10uF
47 uF
180
Phones
drawn Jan 24/99
All transistors 2N3904
All caps in microfarads

Above is the schematic for a simple Morse code practice oscillator. This circuit was originally built with 2N3904 transistors, however many different NPN transistors could be substituted as required.

Tracing the circuit from left to right first brings us to the multivibrator circuit which is composed of Q1 and Q2. The oscillation frequency of the multivibrator is ~ 700 hertz and is set by the RC network formed by the 100K resistor and the 0.01 uF capacitor connected to each transistor base terminal. The approximate time off for each transistor is given by the following formula:

[Time 0ff = 0.7 * R * C] with R in ohms and C in farads.

It maybe more practical to leave the resistance value fixed and vary the capacitor value to obtain a desired oscillation frequency. Rearranging the above formula allows this :

[Total Time Off = 1 / Frequency] with Total Time Off being the total number of seconds that both transistors are off and Frequency is in hertz.

Once the total time off is known, you must divide that answer by 2 as each transistor is off half of the total time off in this symmetrical circuit. Then you simply solve for the capacitor value:

[Time Half = Total Time Off / 2]
[Capacitor = Time Half / (0.7 * R)] with Capacitor answer in farads.

Lets run the numbers to solve for the capacitor values in the schematic;
R = 100K, desired frequency = 700 hertz.
Total Time Off = 1 / F ----> 1 / 700 = 0.00143 seconds.
Time Half = Total Time Off / 2 ----> 0.00143 / 2 = 0.00071 seconds.
Capacitor = Time Half / (0.7 * R) ----> 0.00071 / (0.7 * 100000) = 0.0000001 farads = 0.01 microfarads.

For 600 hertz, the capacitors would be 0.012 uF and for 400 hertz, 0.018 uF.

As you can see it is maybe necessary to adjust the base resistor value to achieve a specific oscillation frequency. The rule of thumb is that the base resistor should be ~ten times the value of the collector resistor assuming the base-bias resistor is connected to VCC as shown.

The output of the multivibrator is buffered by the high input impedance of an emitter follower Q3. This serves to prevent the oscillation frequency from changing when the output load is changed. The AF stage connected to the emitter-follower is a standard high gain common-emitter amp that has been used in many of the projects on this web site. As the multivibrator is buffered by 2 amplifier stages, good frequency stability is realized and frequency changes are negligible when turning the volume control pot. The final stage is a common-collector amplifier which can drive low impedance headphones with reasonable volume.

Code Practice Oscillator Project Notes

The voltage / time output waveform of the astable multivibrator is largely a square wave which some people find harsh. Many users prefer listening to a sine wave although that is beyond the scope of this web page. The multivibrator shown has real advantages in that, it is both dependable and tolerant with respect to parts substitutions. Keying the oscillator as shown practically guarantees that the multivibrator will start running each time you hit the key.

VCC can be 9 to 13.8 volts DC and the larger the voltage the greater the volume in the headphones. The B+ decoupling circuit at the top left can be omitted for battery operation or if you prefer not to bother with it. The basic multivibrator and emitter-follower circuit can be used in a transceiver as a sidetone for monitoring keying. A series resistor from the emitter-follower maybe necessary to attenuate/match the sidetone to the transceiver's AF amp.

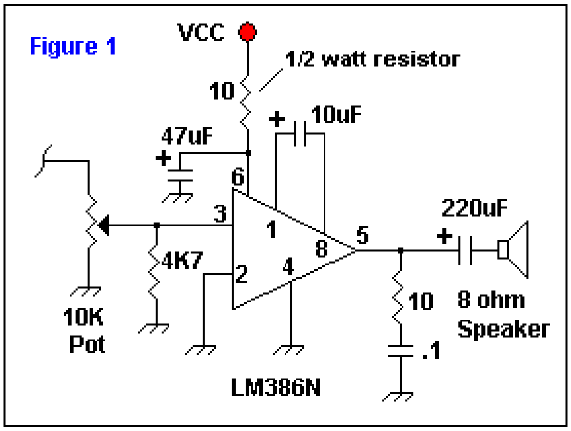
Figure 1

If you desire speaker level output, the Q5 common-collector final can be omitted and the circuit shown below used. This circuit uses the LM386N and provides up to ~ 0.5 watts into an 8 ohm speaker. Connect the Figure 1 circuit as shown to the 10 uF coupling capacitor connected to the collector of Q4. Do not connect the 10 ohm half-watt resistor to the decoupled VCC shown in the schematic. The power supply to the LM386N AF amp should be directly connected to the VCC, not like the common-collector AF stage shown in the schematic to the right.

Simple Electronic Keyer

A simple keyer suitable for a popcorn QRP transmitter can be built around an astable multivibrator and an example of such is shown above. The basic design of this keyer is from notes, however the original author of the circuit is unknown. The notes were written in 1973 . I modernized the circuit, added a variable speed control and designed an additional output driver stage.

Keyer Operation

Keying this circuit generates either dits and spaces or dahs and spaces. When the keyer is idle, Q1 is on and Q2 and Q3 are switched off. When the user sends code, Q1 turns off and Q2 and Q3 switch on and in turn key any device appropriately connected to the collector of Q3. The off-time of Q1 sets the on-time of Q2 and Q3. The off-time of Q2 and Q3 is set by the 22K Q2 base resistor. This off-time is the set time of the spaces and is constant.

The 68K resistor on the base of Q1 is about three times the resistance of the 22K base resistor on Q2 and consequently dahs are ~ three times the length in duration than dits. Spaces and dits are of the same length of time because when sending dits, the 68K base resistor is paralleled with the 33K resistor and effectively the resistance is ~22K ohms. If the optional relay driver transistor Q4 is used instead of Q3, the theory is the same, just substitute Q4 wherever you see Q3.

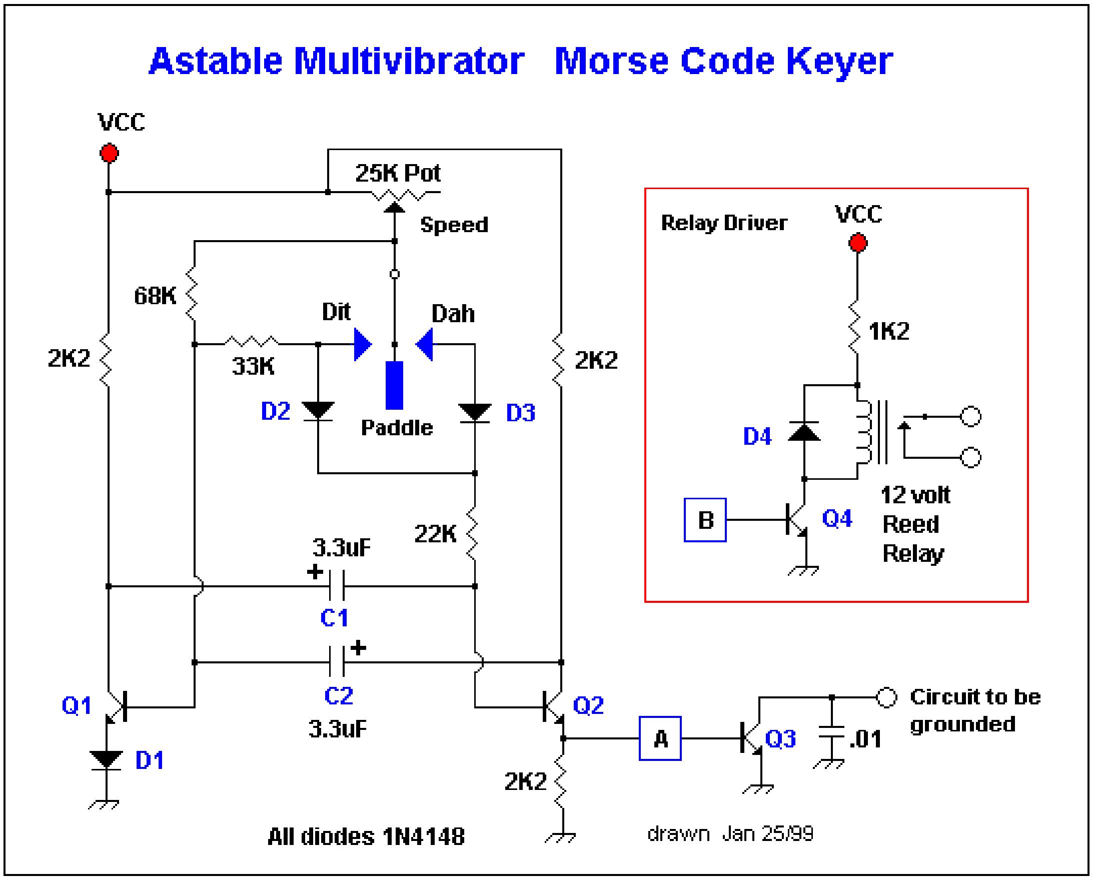

Keyer Speed

The keyer speed is very sensitive to the power supply voltage and any keyer speeds mentioned are ball-park values. Your results may vary depending on your VCC and component tolerances. If the 25K speed control pot is turned to minimum resistance, the actual power supply voltage will be present on both the paddle common and the top end of the 68K base resistor. This will be the maximum speed for the keyer. In fact, the speed control pot could be omitted if you want to economize and the keyer will run at the maximum speed as determined by C1 and C2. If you do not want the speed control feature, connect the paddle common and the 68K resistor to the VCC supply. Another alternative is to build a two speed keyer by using a switch to switch in or out a fixed resistor to vary the voltage instead of using a potentiometer. A trimmer resistor may also be used for "lid-off" speed adjustments.

Varying the base-bias voltage with a pot changes the charging rate on capacitors C1 and C2. Although, I experienced no problems be careful with some resistances/VCCs as the circuit may be unable to provide enough current to saturate the transistors when the base-bias voltage is at its minimum setting (pot set to maximum resistance). Smaller pots such as 10K can also be used with a more limited ability to reduce the keyer speed below the maximum rate.

To set the maximum rate for the keyer, it is necessary to vary the value of C1 and C2. For this circuit to function correctly, C1 must equal C2. An experiment was conducted with the 25K pot removed and the paddle common and the supply end of the 68K resistor connected to the main B+ terminal. VCC was measured at 13.8 volts. C1 = C2. The words per minute were counted for four different standard capacitance values and the results were as follows:

2.2 uF = 27 WPM
3.3 uF = 23 WPM
4.7 uF = 17 WPM
10 uF = 9 WPM

For this project, I settled with the 3.3 uF value, although personally, I use a 2.2 uF capacitor for C1 and C2. Sending speed can be reduced with the speed control pot or by increasing the time interval between characters and words. With 3.3 uf caps for C1 and C2, turning the 25K speed control pot to maximum resistance dropped the sending rate down to 12 words per minute. If you needed to slow down below 12 WPM, an amateur could send code using the Farnsworth method as mentioned above. At any rate this circuit allows you to determine the maximum speed rate by choosing the C1 and C2 capacitance value to suit your needs.

Q3, Q4 Output Stages

Two different output stages maybe used with this keyer and they will be referred to as the Q3 or Q4 stage.The Q3 stage is a simple transistor switch which will ground any component(s) connected to its collector when turned on during code sending. A variety of transistors maybe used here and care must be taken to ensure that you do not exceed the maximum dissipation of a given BJT.

The Q4 stage is a relay driver. The 12 volt relay used during bench testing was a Radio Shack reed relay that I had in my parts collection. Specifications were SPST 1A @ 125 VAC , 1050 ohms DC coil resistance, part number 275-233. My VCC was 13.8 volts so a 1K2 current limiting resistor was placed between the relay and supply voltage. The relay has suffered no harm despite significant torture but be careful when you are using a VCC greater than 12 volts. This resistance may be dropped down to 470 ohms or omitted if you are using lower voltages. Any relay with a DC coil resistance of 500 to 3000 ohms should work in this circuit.

I have never used this keyer with the Q4 relay driver for QRP operation as I prefer solid-state switching. The Q3 stage in turn can be connected to trigger a PNP transistor switch to supply DC voltage to a keyed transmitter driver and/or to pull a VFO to its offset frequency. A great example of PNP transistor switches can be found in The Ugly Weekender article by KA7EXM and W7ZOI. This article is referenced in the recommended reading list on this web site. Another bonus of the Q3 stage is that it draws less current than the Q4 stage.

Sending Code

If you are used to a deluxe iambic keyer, this keyer will take some getting used to. The dahs in particular can be problematic as they do not self-complete like they do on my station homebrew TTL iambic keyer. After some practice, however most people should be able to send some good morse code with it despite the lack of iambic luxury.

Miscellaneous

When you power up this keyer, it sends a dit. In one version of this keyer, used in a 40M transceiver, I had a 10K pot for the speed control with a built-in switch on the 10K pot to turn the keyer on or off during station setup.

The diodes D1 - D4 can be any common switching diode such as the 1N914. Q1 to Q4 can be any NPN transistor with a Beta greater than 50 such as the 2N3904 or 2N2222. In the test keyer, all BJT's were 2N3904.

For fun, I connected the multivibrator code practice oscillator to the keyer and gave the keyer a workout. They both work great together and are very complimentary. Have fun with astable multivibrators !

Update Oct 2007:

Giovanni, IW7EHC has posted a great web page on this schematic including PCB layout files in FidoCAD format

Check out his excellent Italian and English language web site

---

Supplemental Page for the 10 MHz WWV TRF receiver

I have received a tremendous amount of feedback regarding the 10 MHz WWV TRF receiver. Many builders have constructed their own versions and have improved it. This captures the spirit of the QRP/SWL Homebuilder web site; take an idea and evolve it into something better on the bench.

I will post some of the feedback as well as provide additional comments and images regarding the 10 MHz WWV project on this supplemental web page. Please continue to provide project feedback and help me to improve the QRP/SWL HomeBuilder web site.

Terry, K5HP
Hi Todd,

I have just finished (few days ago) your design for the 10 MHz WWV Receiver. I reduced you schematic diagrams to CirCad schematics, drug out the rat-net, positioned and connected everything, and etched the boards. I'll include a photo of the 4 finished boards. They are very roomy for extra tinkering. Except for two solder bridges on two very big pads on the tweeker capacitors in the front end, it all worked perfectly first try. Hmmm, I cannot say that for all my projects in the past. After repairing the two shorts to ground of the RF path, the front end then tuned and worked perfectly also.

I have not taken the project out to the Lab (barn) and put it on my Motorola Service monitor, but the input to the filter board is roughly 500 uV. It can easily hear WWV at this level with a decent antenna connected to the filter input when WWV is fairly strong at my location. Not bad, considering the mismatch at that location.

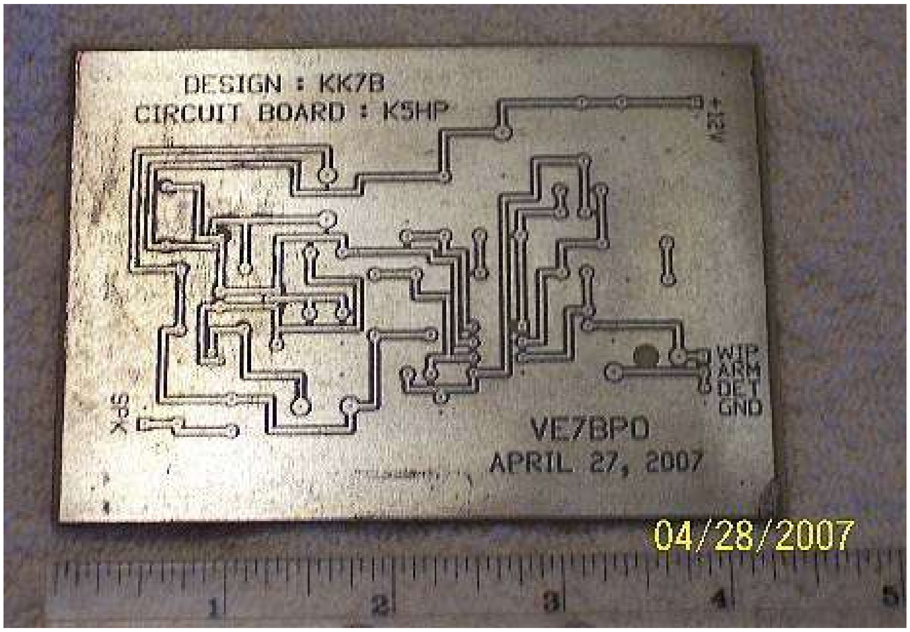

Above. Etched Audio amplifier PC board from K5HP.

Have you or someone else tinkered with an AGC/AVC circuit? At one time I had a nice circuit using an National chip (don't remember the number) which is no longer available that was easy to build and easy to retro-fit into something. It was an AVC which they called an audio AGC.

When I built out the boards I placed proper credit to you and the other developers mentioned on the boards. Should you want my work product it's free for the asking. I'll be glad to send it to you. CirCad in ver 4 is a fairly

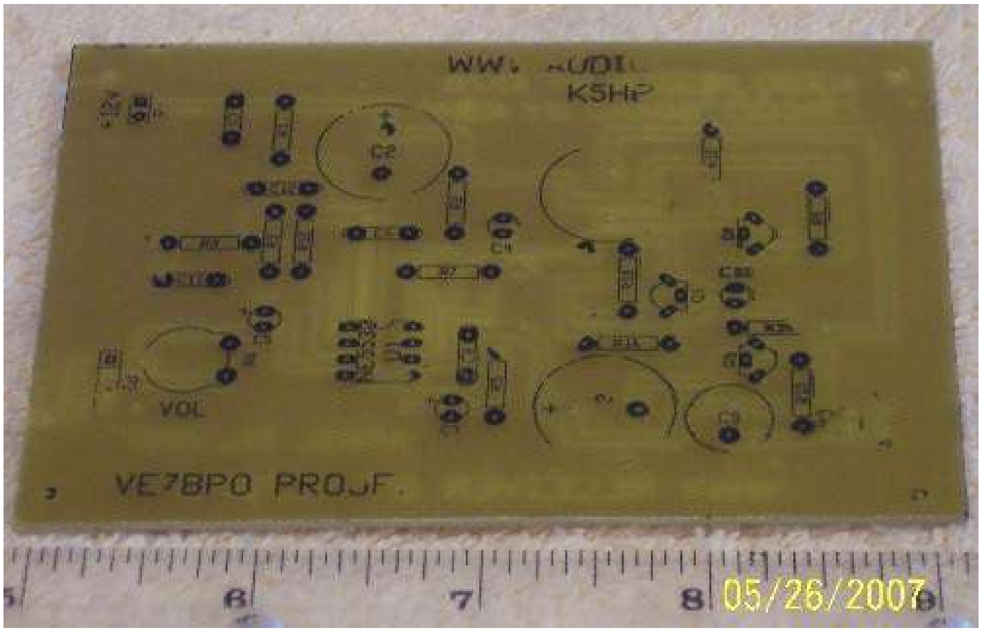

Above. Audio amplifier PC board overlay for parts placement from K5HP.

complete PCB cad program and it's free too. There is a version 5, but I have not tried it out yet.

My sincere congratulation for producing and error free web article. That is a heck of a more than you can say about many other projects in the past. Most of those in the past had errors a plenty, and some just did not flat work as described. Someone just thought they would as it turned out. Thanks for the fine article and I'll check your site over for more interesting thing to build or learn from. Best wishes and good luck.

"I use the CirCAD to produce the artwork, then take the laser printer output and bond the plastic toner onto the PCB. Works good most of the time. Sure easier than trying to do photo etch, and silk screening is pretty expensive. The bottom copper is viewed as you would look down onto the top of a circuit board, through to the copper side. The side you are seeing is bonded to the copper surface. Print one out, hold it up to a light source looking through non printed side and it becomes apparent."

Above. A layout of the 4 completed boards by K5HP.

Terry designed and built printed circuit boards. His photographs, etching and overlay files are all bundled together in convenient zipped files which can be downloaded below.

Download the front end zip file

Download the Crystal filter zip file

Download the Detector zip file

Download the AF amp zip file

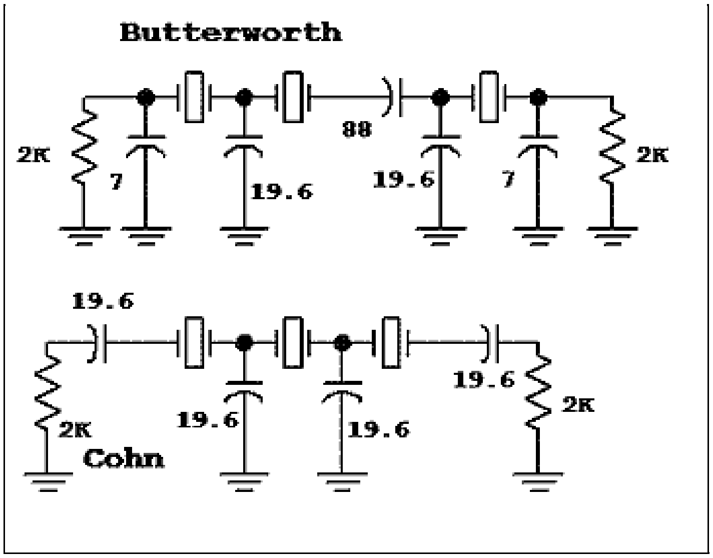
Butterworth
2K
7
19.6
88
19.6
7
2K
19.6
19.6
2K
19.6
19.6
19.6
2K
Cohn

Above. Schematics diagrams of 10 MHz crystal filters by Wes, W7ZOI. I did not experiment with the Butterworth type filter.

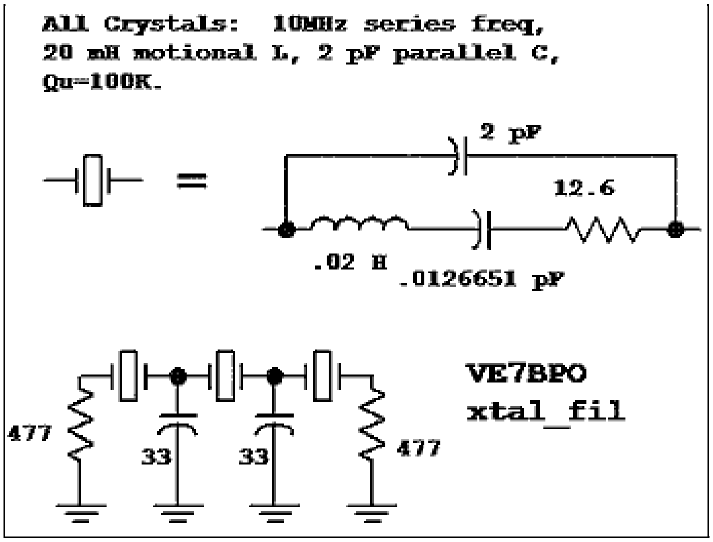
All Crystals: 10MHz series freq,
20 mH motional L, 2 pF parallel C,
Qu=100K.
=
2 pF
12.6
.02 H
.0126651 pF
477
33
33
477
VE7BPO
xtal_fil

Above. Schematics diagrams by Wes, W7ZOI.
A schematic representation of the 10 MHz xtal is shown along with the "popcorn" 10 MHz xtal filter I designed.

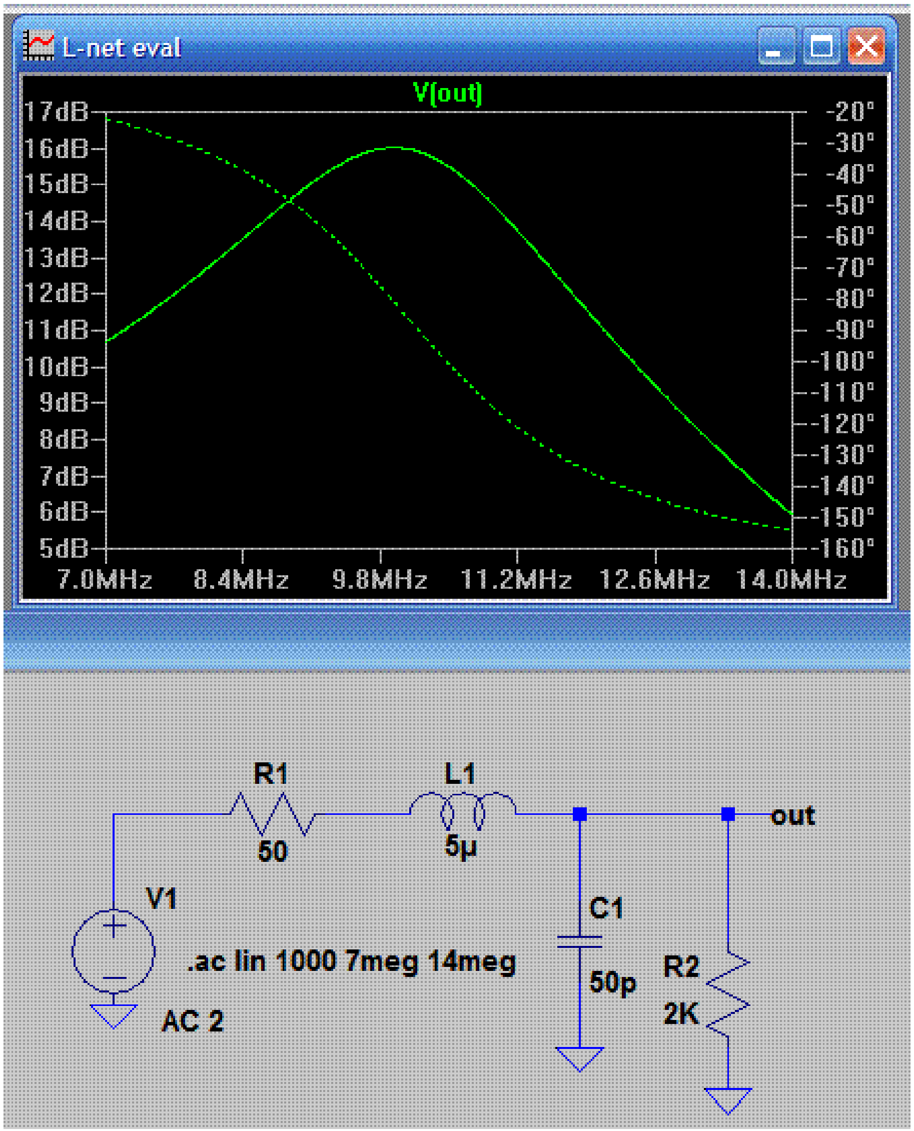
L-net eval
V(out)
17dB
16dB
15dB
14dB
13dB
12dB
11dB
10dB
9dB
8dB
7dB
6dB
5dB
-20°
-30°
-40°
-50°
-60°
-70°
-80°
-90°
-100°
-110°
-120°
-130°
-140°
-150°
-160°
7.0MHz
8.4MHz
9.8MHz
11.2MHz
12.6MHz
14.0MHz
R1
50
L1
5µ
out
V1
AC 2
C1
50p
R2
2K

```
.ac lin 1000 7meg 14meg
```

Above. A simulation of an L network depicted in the L-Match and AM
Detector schematic for those interested in driving this detector with a 50 ohm output impedance.

SWL HOMEBUILDER

Amateur and Short Wave Radio Electronics Experimenter's Web Site

MF TRF Receiver Supplemental Page

Click here to go to the main project web page This supplemental page is under construction

---

5532 Op amp Pin Out

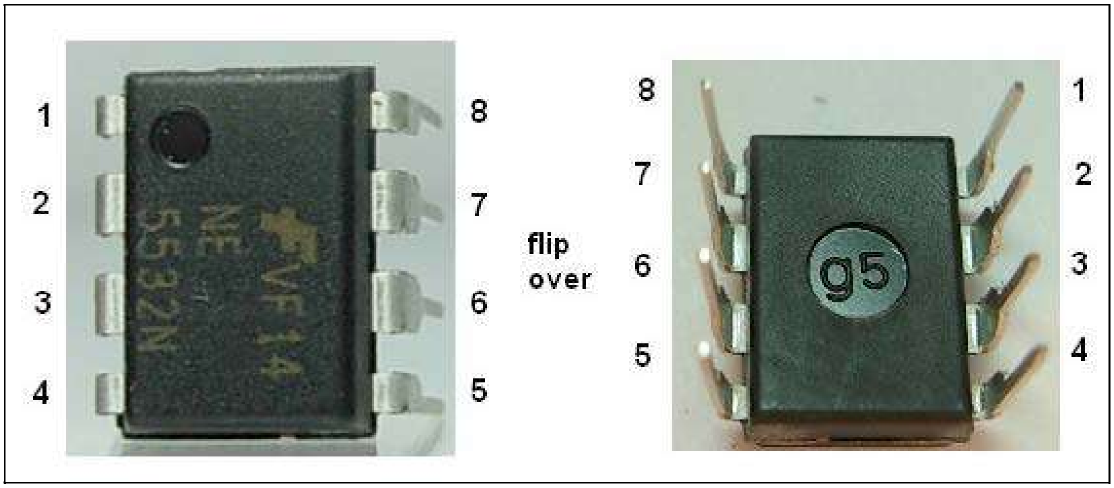

A Version of the Receiver Using Just L1 and C1 as the band pass filter

You can see a red LED and current limiting resistor in the photos. This is a power on indicator. The telescopic whip antenna can be seen in the lower photo.

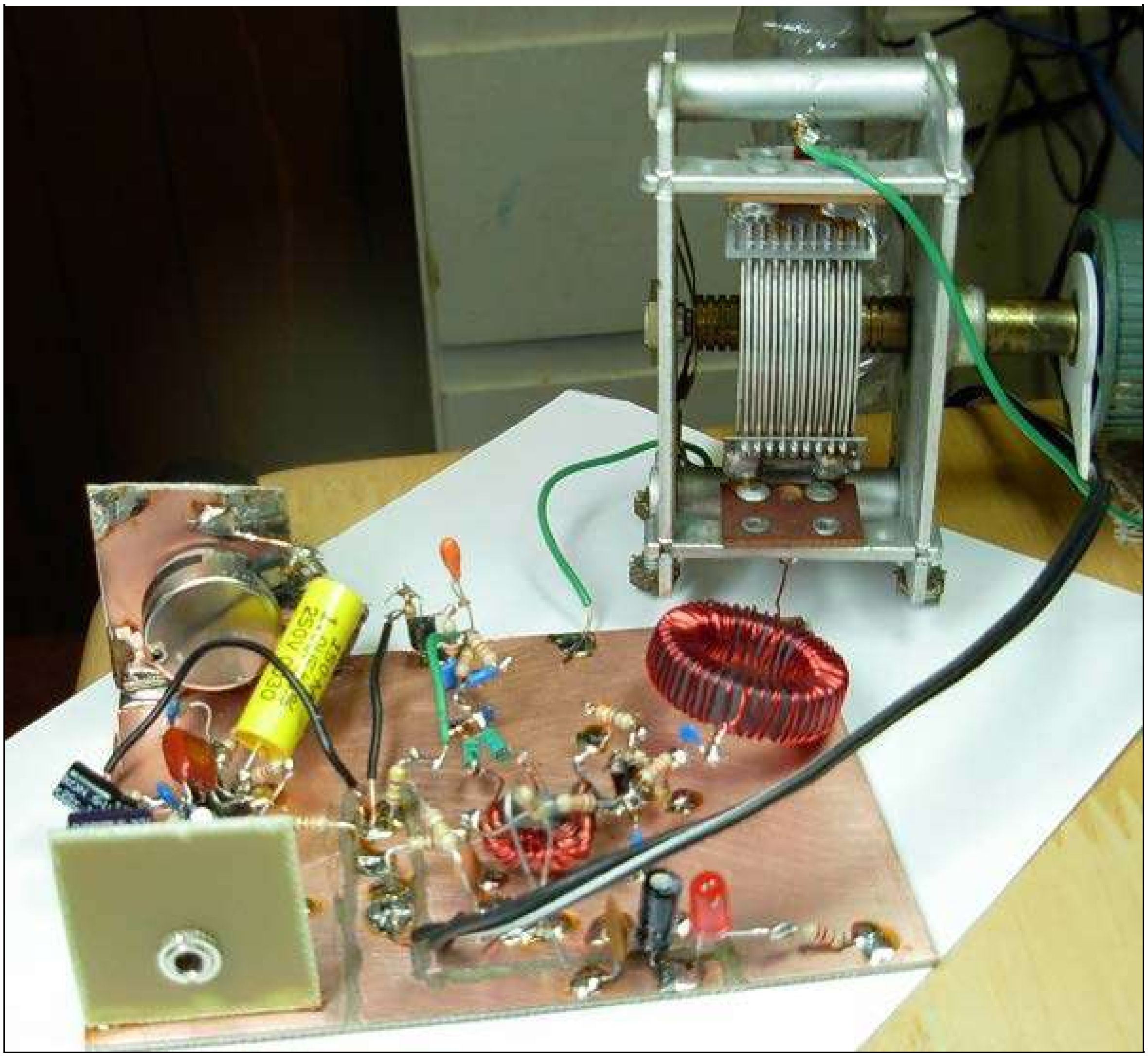

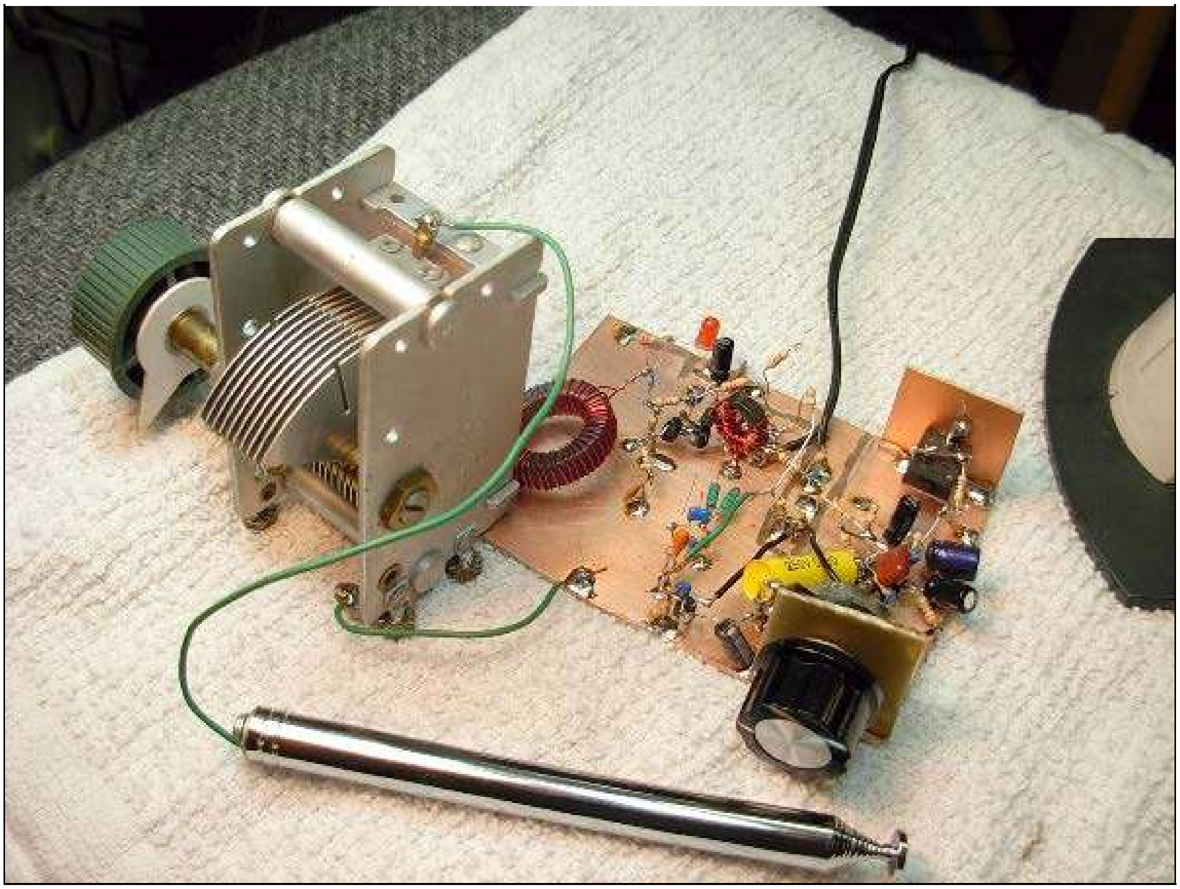

Early Versions of the Project Front End

In section A of the schematic below, a dual-ganged 365 pF air variable capacitor was used for tuning. CV, a small compensating capacitor may be required to adjust for the capacitive differences of the 2 L-C tanks in the front end filter. It may have to be placed on the L1 side of the filter (opposite to what is shown). A JFET source follower was used for Q1 but discarded because stage gain was insufficient with a short whip antenna.

In the version depicted as section B, two 12-110 pF air variable capacitors were used along with switchable 270 pF capacitors to provide 2 ranges of tuning. Although this worked, the switching mechanism was found to degrade selectivity somewhat. The JFET was reconfigured as a common source RF amp. This amp caused some distortion of the RF signal on stronger stations unless the gate resistor was lowered to 47K or less. Ultimately a cascode BJT RF amp was chosen for the project. A photograph of the section A breadboard is shown below the schematic.

Broadcast Radio Band Front End

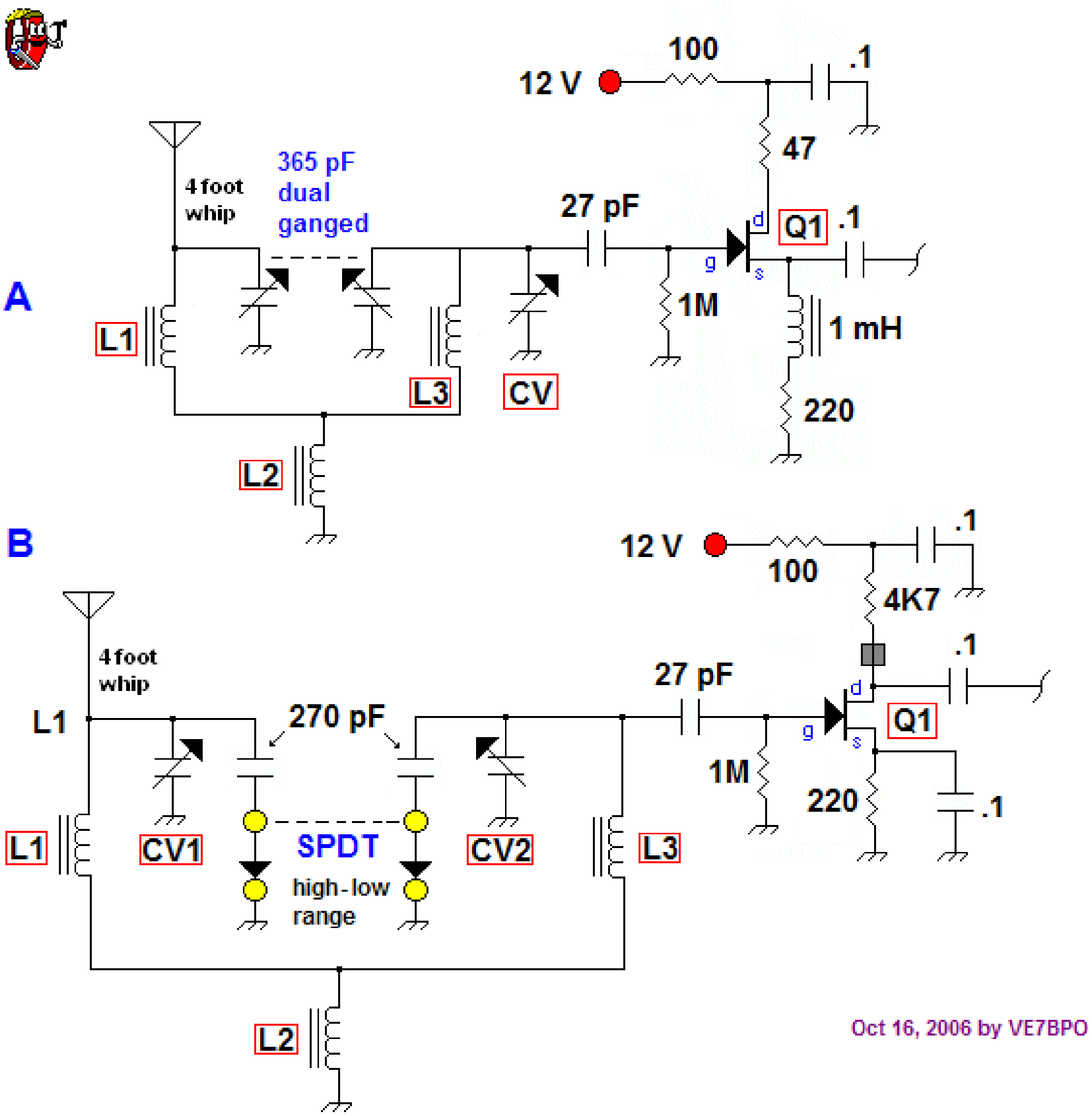
A
12 V
100
.1
47
4 foot whip
365 pF dual ganged
27 pF
d
Q1
.1
g
s
L1
1M
1 mH
L3
CV
220
L2
B
12 V
100
.1
4K7
4 foot whip
.1
27 pF
d
L1
270 pF
g
Q1
s
1M
220
.1
L1
CV1
SPDT
high-low range
CV2
L3
L2

Oct 16, 2006 by VE7BPO

L1 = L3 = 200 uH or 55 turns #28 AWG on a FT50-61 ferrite core
L2 = ~ 5 uH = 9 turns #22 AWG on a FT50-61 ferrite core
CV = small value trimmer capacitor for tracking - 4.5 - 20 pF
CV1 = CV2 = 12 - 110 pF air variable capacitor
Q1 = nJFET; MPF102 or J310

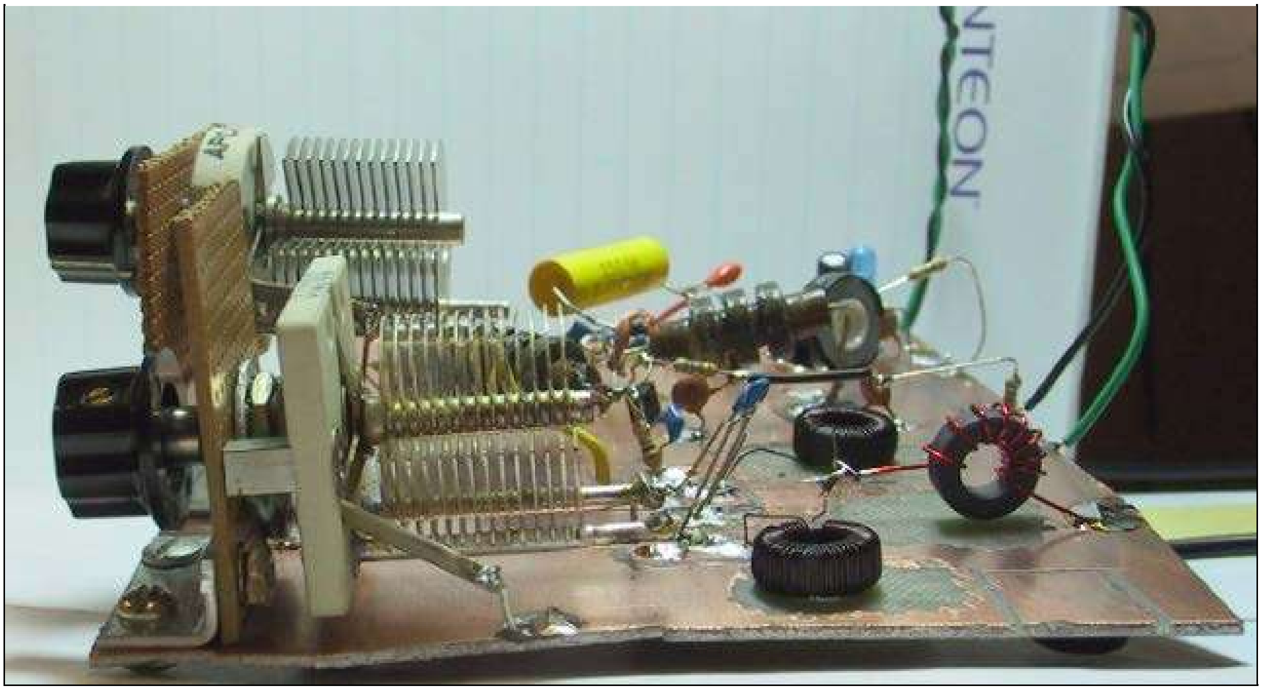

The inspirational detector schematic used for motivation to develop the full wave detector

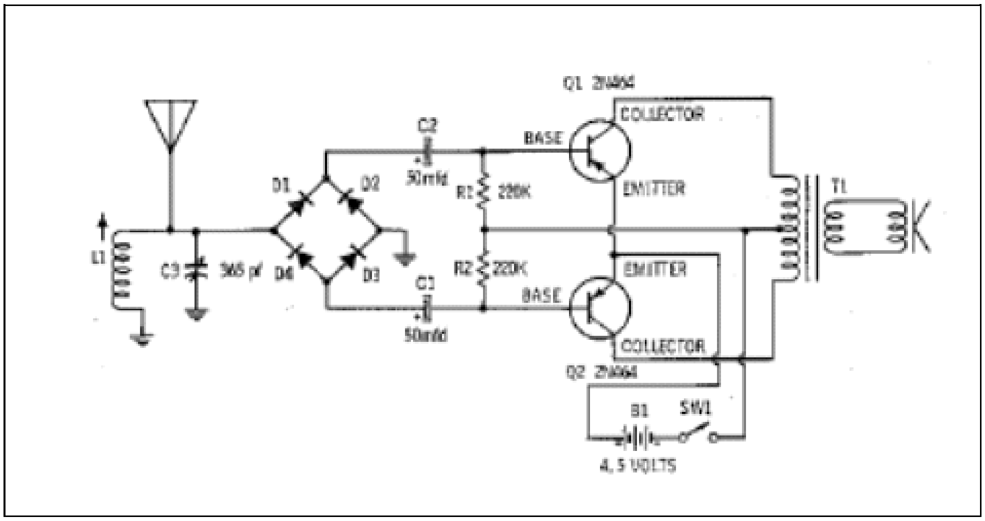
Q1 2N404
COLLECTOR
C2
BASE
50mfd
D1
D2
R1 220K
EMITTER
T1
L1
C3 365 pf
D4
D3
R2 220K
EMITTER
C1
BASE
50mfd
COLLECTOR
Q2 2N404
B1
SW1
4, 5 VOLTS

Varactor Tuned Front-end Band pass Filter

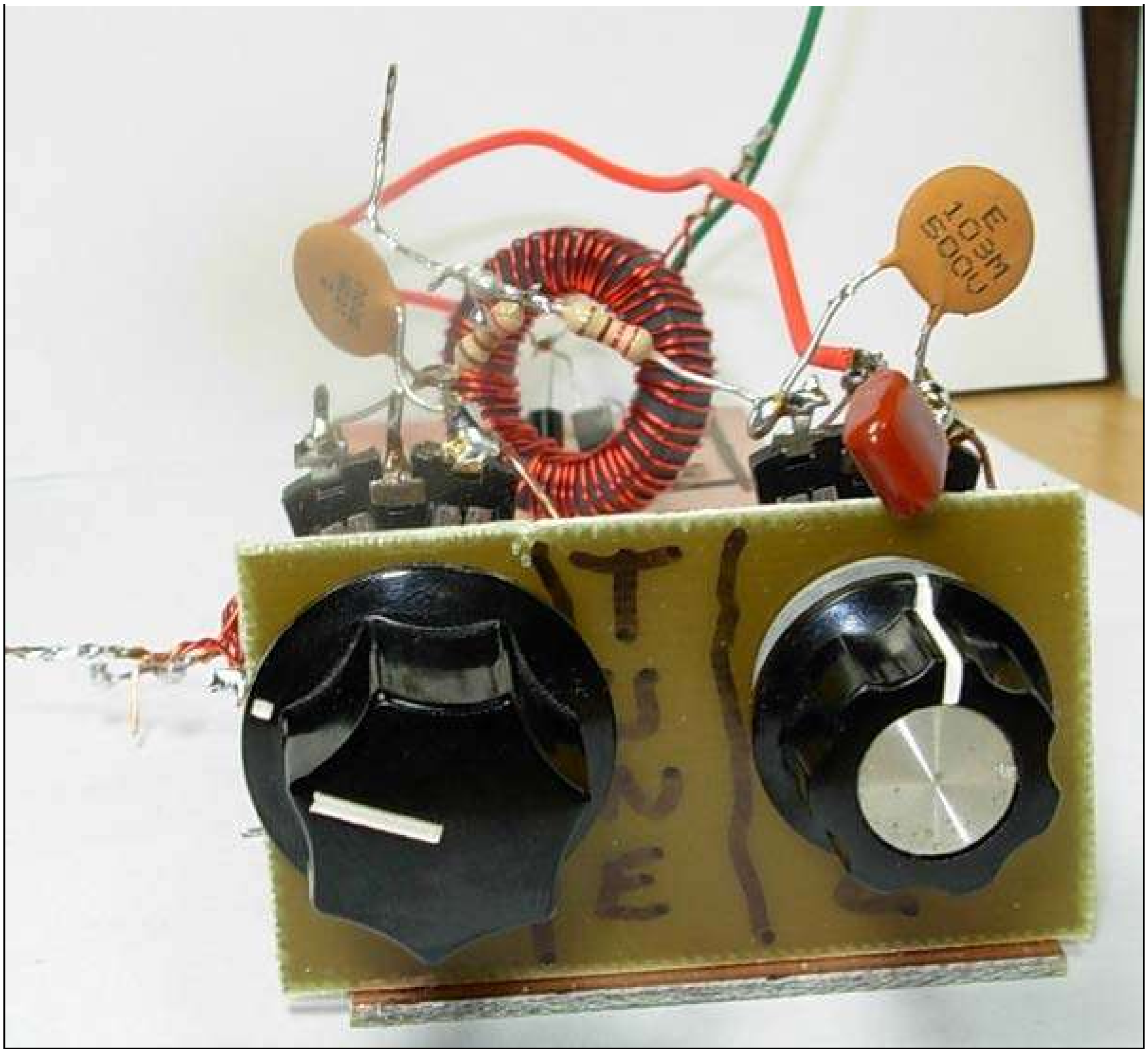
E
103M
500V
TUNE

Above photo. The 2 potentiometer varactor controlled band pass filter.

Amateur and Short Wave Radio Electronics Experimenter's Web Site

Supplemental to the Cascode Hybrid 5 MHz WWV Receiver

DC Voltages of the Cascode Hybrid Amp

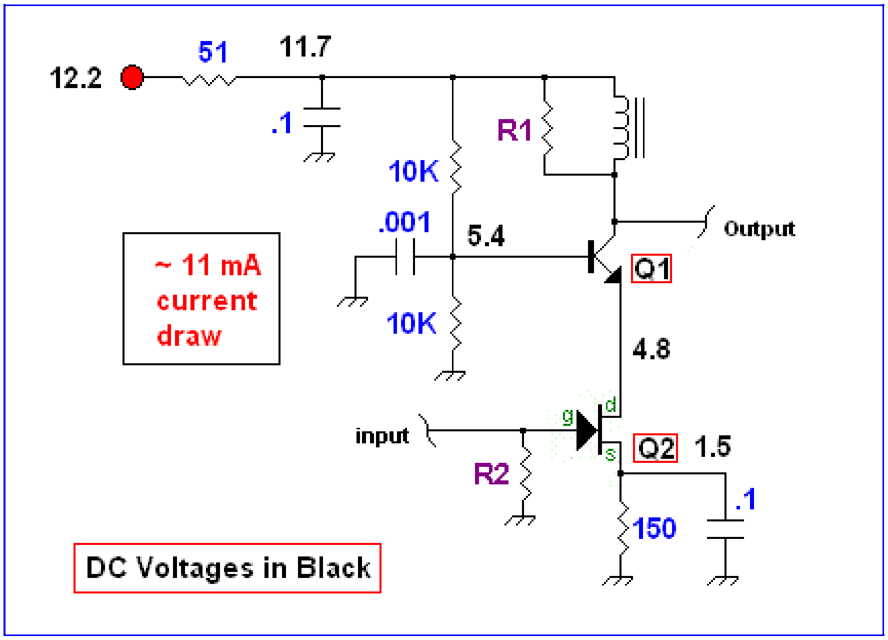

On the right is the basic hybrid cascode amplifier schematic. Note that the Q1 voltage divider bias has two 10K resistors as compared to the 12K and 10K combination used in the final project. A slightly lower Q1 bias voltage was ultimately chosen along with a 150 ohm Q2 source resistor to enhance stability. Early bread boards of this amplifier used a 47 - 100 ohms Q2 source resistor along with the slightly higher Q1 bias, but suffered from parasitic oscillations.

The DC voltages are rounded to 1 decimal point. These are ball park values for reference purposes. The typical stage current draw was around 11 mA. R1 sets the output impedance and is optional. R2 sets the input impedance. This amp is very versatile in that it can be used to match a wide variety of input and output impedances.

The **main 5 MHz WWV receiver web page** which this web page supplements is here.
Click here for a recording of 5 MHz WWV made with the Icom R-75 receiver using the 6 KHz AM filter and the stock speaker. You may wish to compare it to the recorded 5 MHz WWV receiver project audio files linked here , here and here.

---

10 MHz Components

For those who wish to experimentally develop a 10 MHz version, I have included some potential band pass and crystal filter schematics you may wish to consider. In addition, my preliminary experiments indicated that another RF amplifier stage might be necessary for 10 MHz WWV. During development of these band pass and crystal filters, 10 MHz WWV reception was very poor and due to lack of time and the poor conditions, further experimentation was abandoned.
To the left is a 10 MHz band pass filter schematic with a -3dB bandwidth of around 100 KHz.

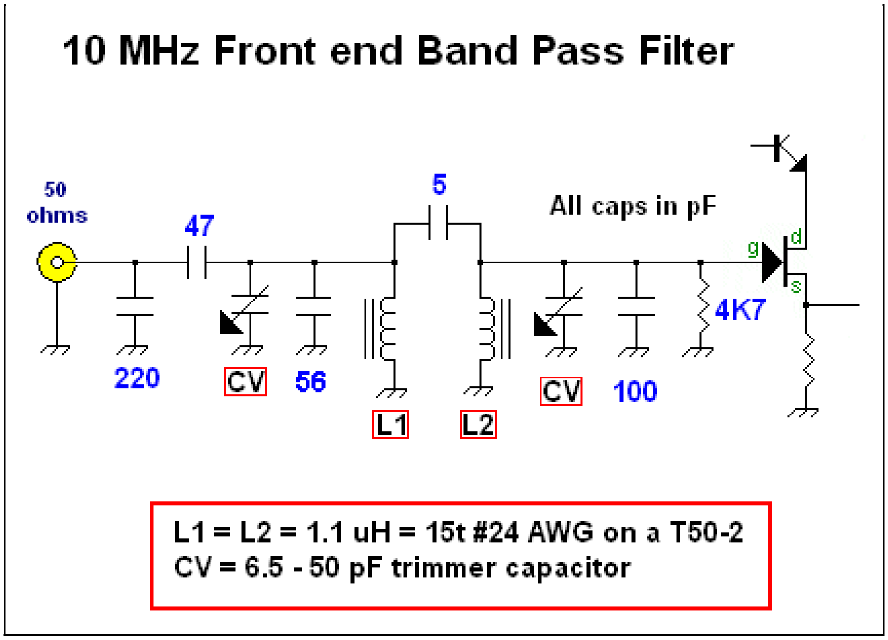
10 MHz Front end Band Pass Filter
50 ohms
47
5
All caps in pF
g
d
s
4K7
220
CV
56
L1
L2
CV
100
L1 = L2 = 1.1 uH = 15t #24 AWG on a T50-2
CV = 6.5 - 50 pF trimmer capacitor

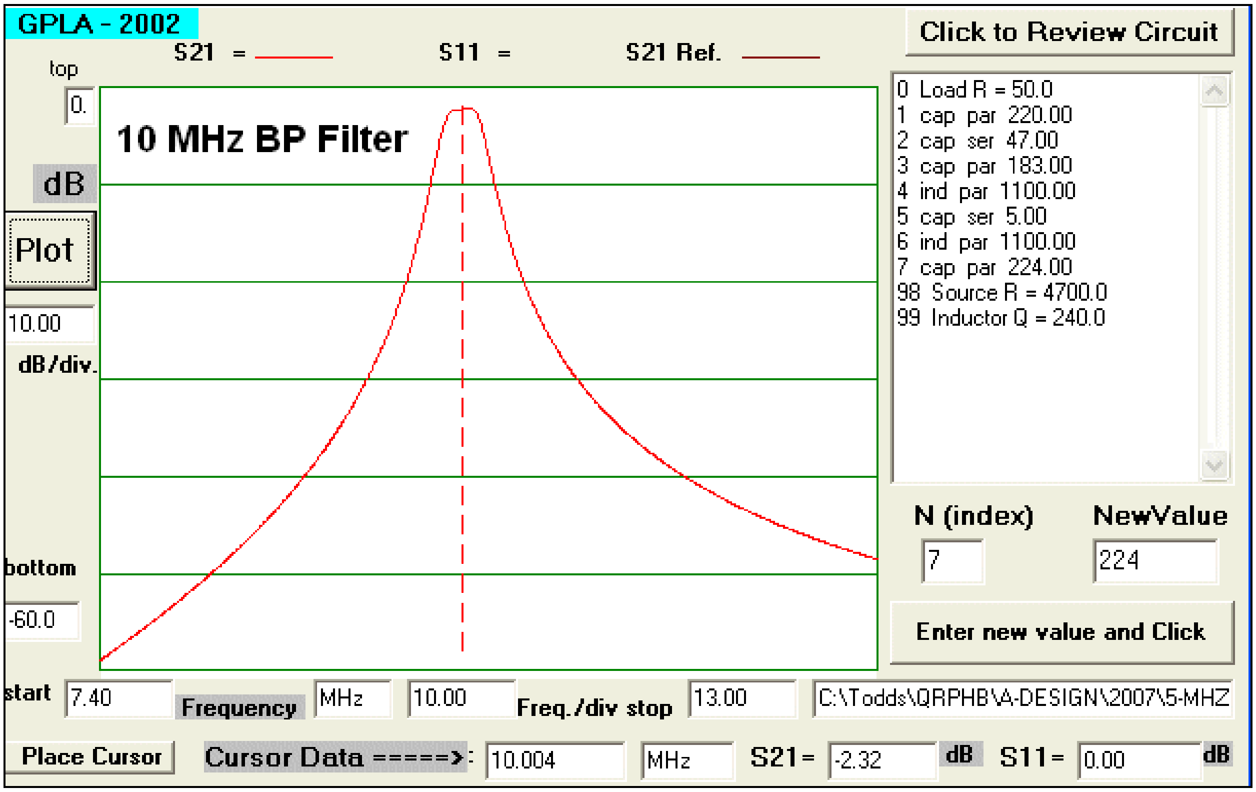
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
top
dB
Plot
dB/div.
bottom
10 MHz BP Filter
N (index)
NewValue
Enter new value and Click
start
Frequency
Freq./div stop
Place Cursor
Cursor Data =====>:
S21=
dB
S11=
dB

```
0  Load R = 50.0
1  cap  par  220.00
2  cap  ser  47.00
3  cap  par  183.00
4  ind  par  1100.00
5  cap  ser  5.00
6  ind  par  1100.00
7  cap  par  224.00
98  Source R = 4700.0
99  Inductor Q = 240.0
```

top: 0.
dB/div.: 10.00
bottom: -60.0
N (index): 7
NewValue: 224
start: 7.40 MHz
Freq./div: 10.00
stop: 13.00
C:\Todds\QRPHB\A-DESIGN\2007\5-MHZ
Cursor Data =====>: 10.004 MHz
S21= -2.32 dB
S11= 0.00 dB

The GPLA simulation of the 10 MHz front end band pass filter. Tuning is very sharp with this filter.

To the left is a crystal filter design for 10 MHz. The realized bandwidth of these simple min-loss crystal filters is dependent on your crystal parameters.

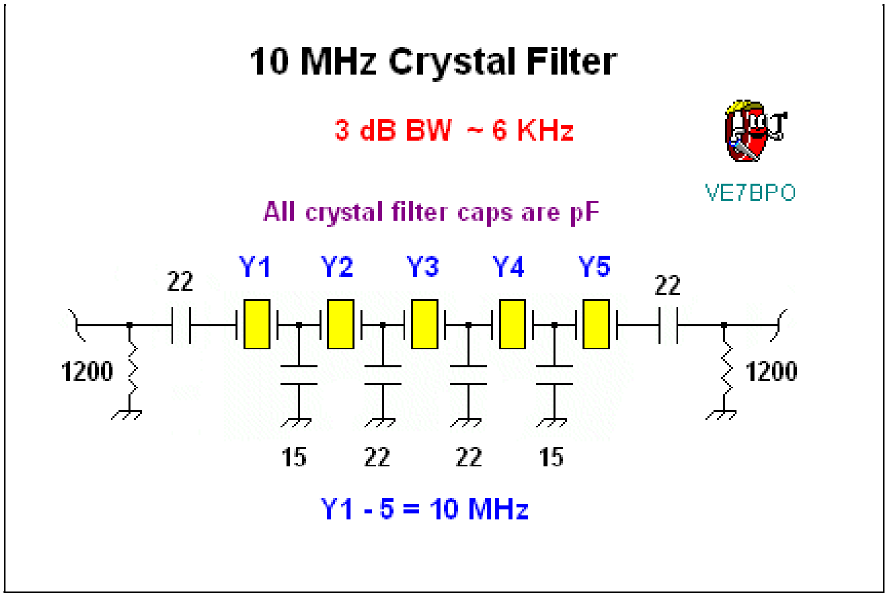
10 MHz Crystal Filter
3 dB BW ~ 6 KHz
VE7BPO
All crystal filter caps are pF
Y1
Y2
Y3
Y4
Y5
22
22
1200
1200
15
22
22
15
Y1 - 5 = 10 MHz

A photograph of the experimental 10 MHz crystal filter is shown to the right. Leads were kept long so that parts could be reused in future experiments.

CY12A
CRYSTEK
10.000

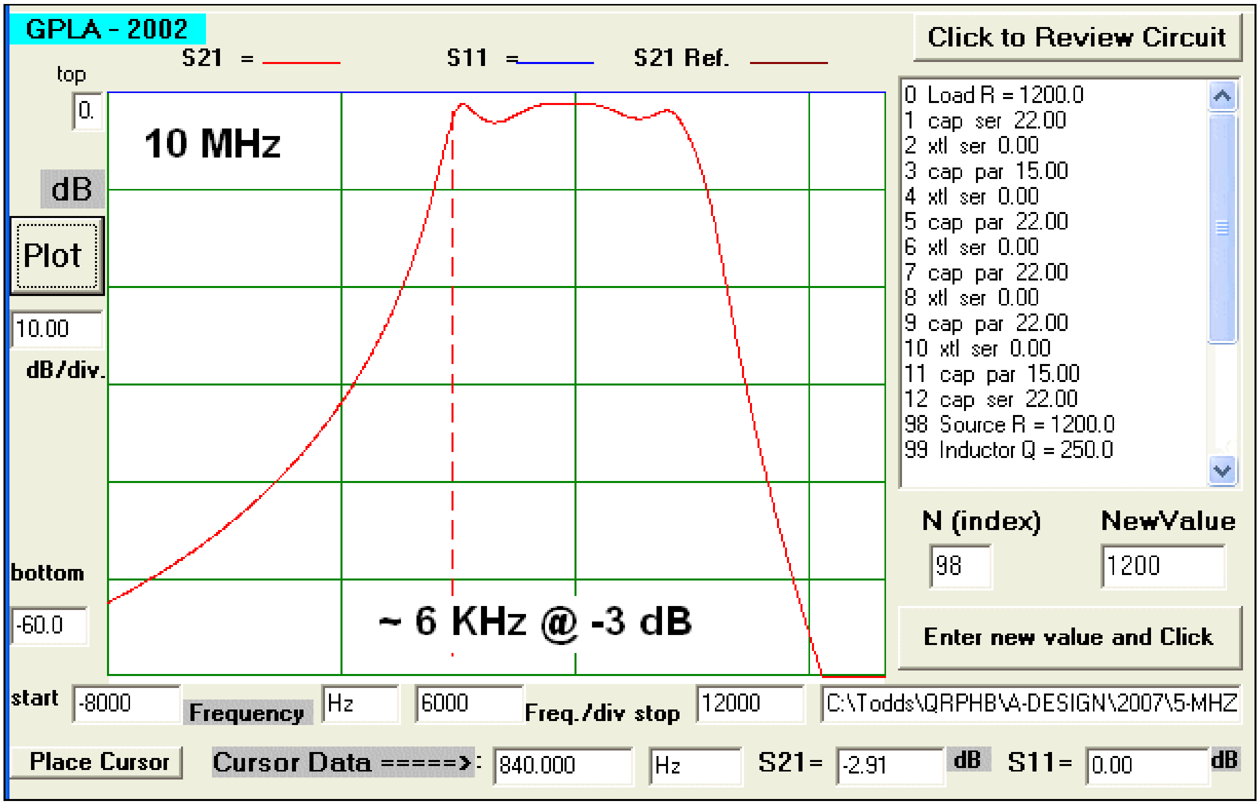
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
top
0.
dB
Plot
10.00
dB/div.
bottom
-60.0
10 MHz
~ 6 KHz @ -3 dB
Enter new value and Click
Place Cursor

0 Load R = 1200.0
1 cap ser 22.00
2 xtl ser 0.00
3 cap par 15.00
4 xtl ser 0.00
5 cap par 22.00
6 xtl ser 0.00
7 cap par 22.00
8 xtl ser 0.00
9 cap par 22.00
10 xtl ser 0.00
11 cap par 15.00
12 cap ser 22.00
98 Source R = 1200.0
99 Inductor Q = 250.0

N (index): 98
NewValue: 1200

start -8000 Frequency Hz 6000 Freq./div stop 12000 C:\Todds\QRPHB\A-DESIGN\2007\5-MHZ

Cursor Data =====>: 840.000 Hz S21= -2.91 dB S11= 0.00 dB

Above. The GPLA simulation of the 10 MHz crystal filter.

6 MHz Components

Since it is easy to find 6 MHz crystals and Radio Habana Cuba is on 6.0 MHz, band pass and crystal filters were designed for reference purposes. To the right is the front end, band pass filter schematic.

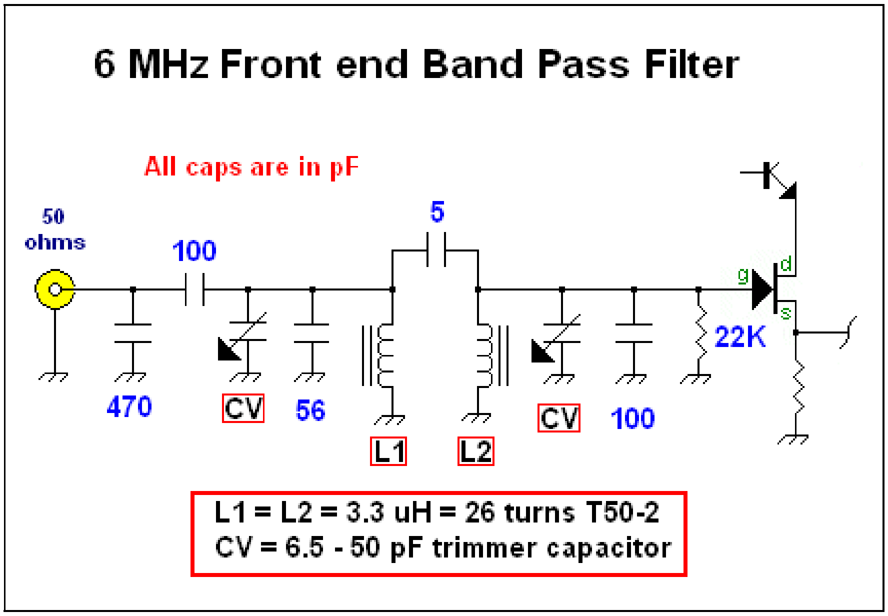
6 MHz Front end Band Pass Filter

All caps are in pF

50 ohms
100
5
470
CV
56
L1
L2
CV
100
22K
g
d
s

L1 = L2 = 3.3 uH = 26 turns T50-2
CV = 6.5 - 50 pF trimmer capacitor

To the left is a potential crystal filter for 6 MHz AM reception. The bandwidth is ~ 5 KHz which is starting to get too narrow for some listeners, however, your crystal parameters will determine your filter's actual bandwidth.

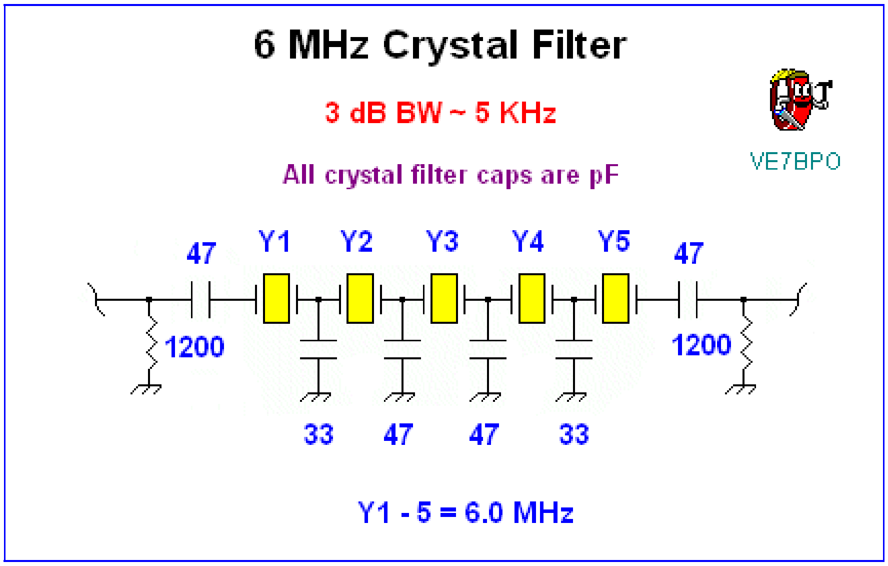
6 MHz Crystal Filter

3 dB BW ~ 5 KHz

All crystal filter caps are pF

VE7BPO
47
Y1
Y2
Y3
Y4
Y5
47
1200
1200
33
47
47
33

Y1 - 5 = 6.0 MHz

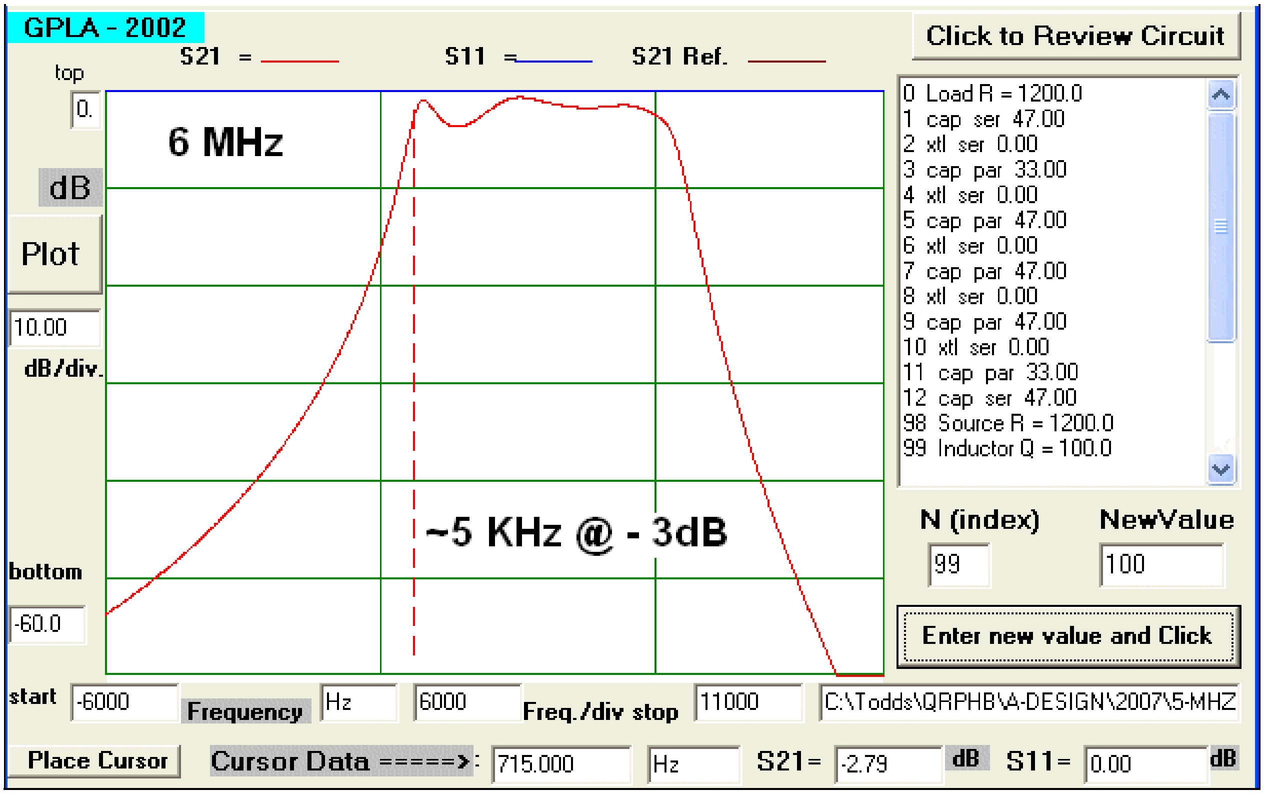
GPLA - 2002
S21 =
S11 =
S21 Ref.
Click to Review Circuit
6 MHz
~5 KHz @ - 3dB
Plot
Enter new value and Click
Place Cursor

top 0.

dB

10.00 dB/div.

bottom -60.0

0 Load R = 1200.0
1 cap ser 47.00
2 xtl ser 0.00
3 cap par 33.00
4 xtl ser 0.00
5 cap par 47.00
6 xtl ser 0.00
7 cap par 47.00
8 xtl ser 0.00
9 cap par 47.00
10 xtl ser 0.00
11 cap par 33.00
12 cap ser 47.00
98 Source R = 1200.0
99 Inductor Q = 100.0

N (index) 99
NewValue 100

start -6000 Frequency Hz 6000 Freq./div stop 11000 C:\Todds\QRPHB\A-DESIGN\2007\5-MHZ

Cursor Data =====>: 715.000 Hz S21= -2.79 dB S11= 0.00 dB

The GPLA simulation of the crystal filter using the parameters from the 6.0 MHz crystals I had in my parts collection.

VE7BPO Envelope Detector

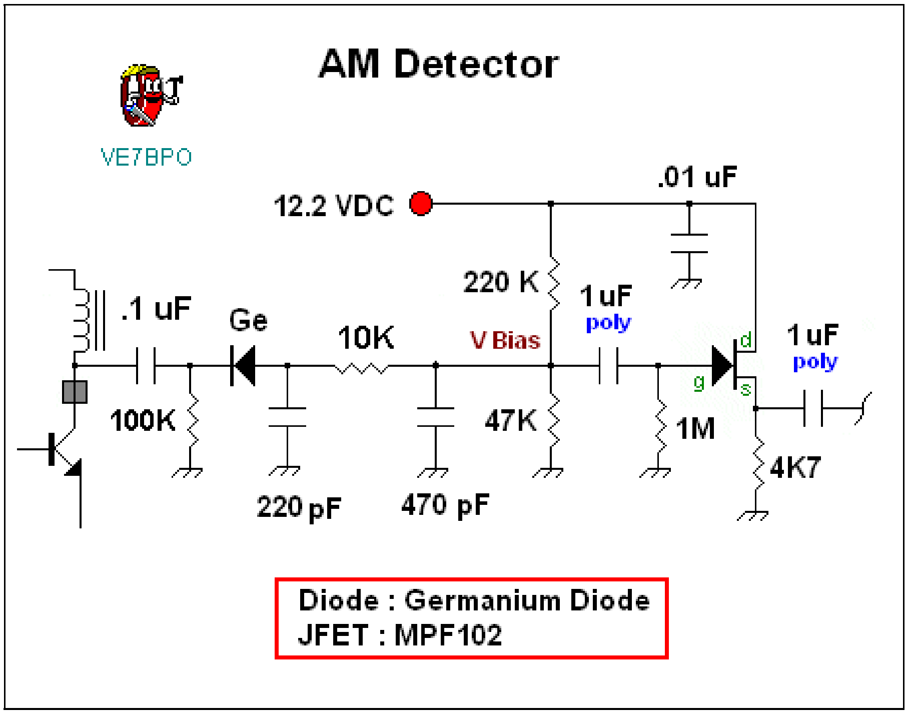
AM Detector
VE7BPO
12.2 VDC
.01 uF
220 K
1 uF
poly
.1 uF
Ge
10K
V Bias
d
g
s
1 uF
poly
100K
47K
1M
4K7
220 pF
470 pF
Diode : Germanium Diode
JFET : MPF102

To the left is the schematic of another detector experimentally developed for the 5 MHz WWV receiver. The germanium diode has incredible sensitivity when biased, although noise tends to increase with applied DC voltage. This basic design also worked well with a hot carrier diodes, although had less sensitivity. V bias can be changed by adjusting the 220K, 47K and 100 K resistor values. Note diode polarity.

Miscellaneous Photos and Circuits

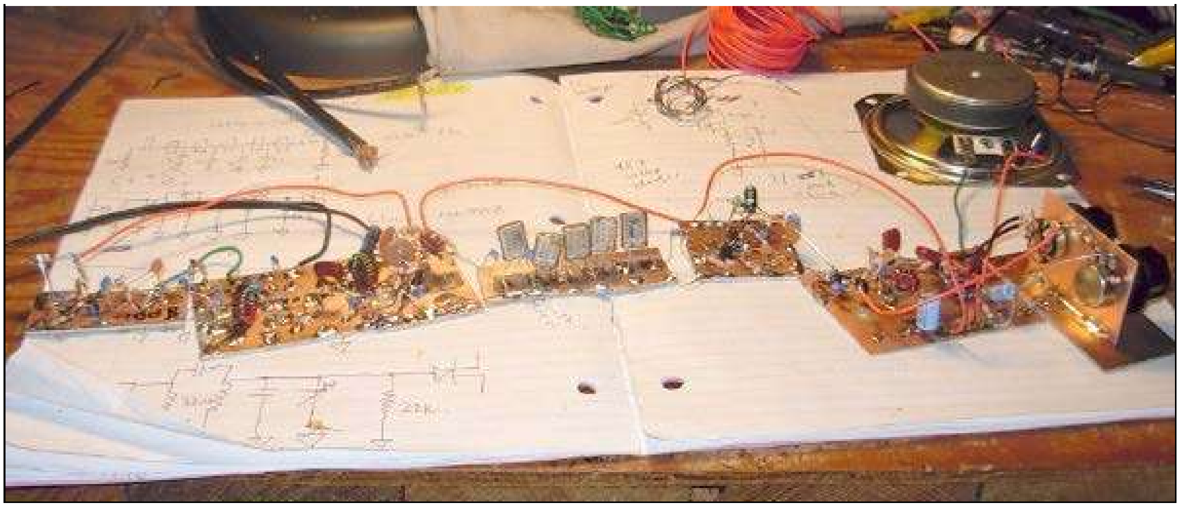

Above. A work bench version of the basic receiver bread board built in distinct stages from the antenna input to the speaker.

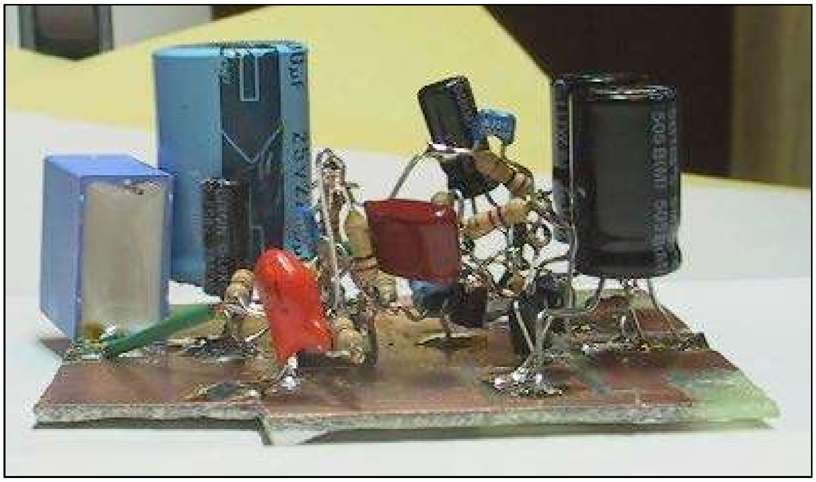

Above, A breadboard of the KK7B audio stage which is preferred to the LM386 stage used in the 5 MHz WWV receiver. The LM386 does take up less room and certainly increases the popcorn factor.

JVC
VGS1001-002
NOM 2.5W
MAX 4W
3.2 Ω
±20%
105°C
07/85

Above. The AF stage shown above in a chassis containing a speaker. This is my lab AF amp for testing receivers on my work bench.

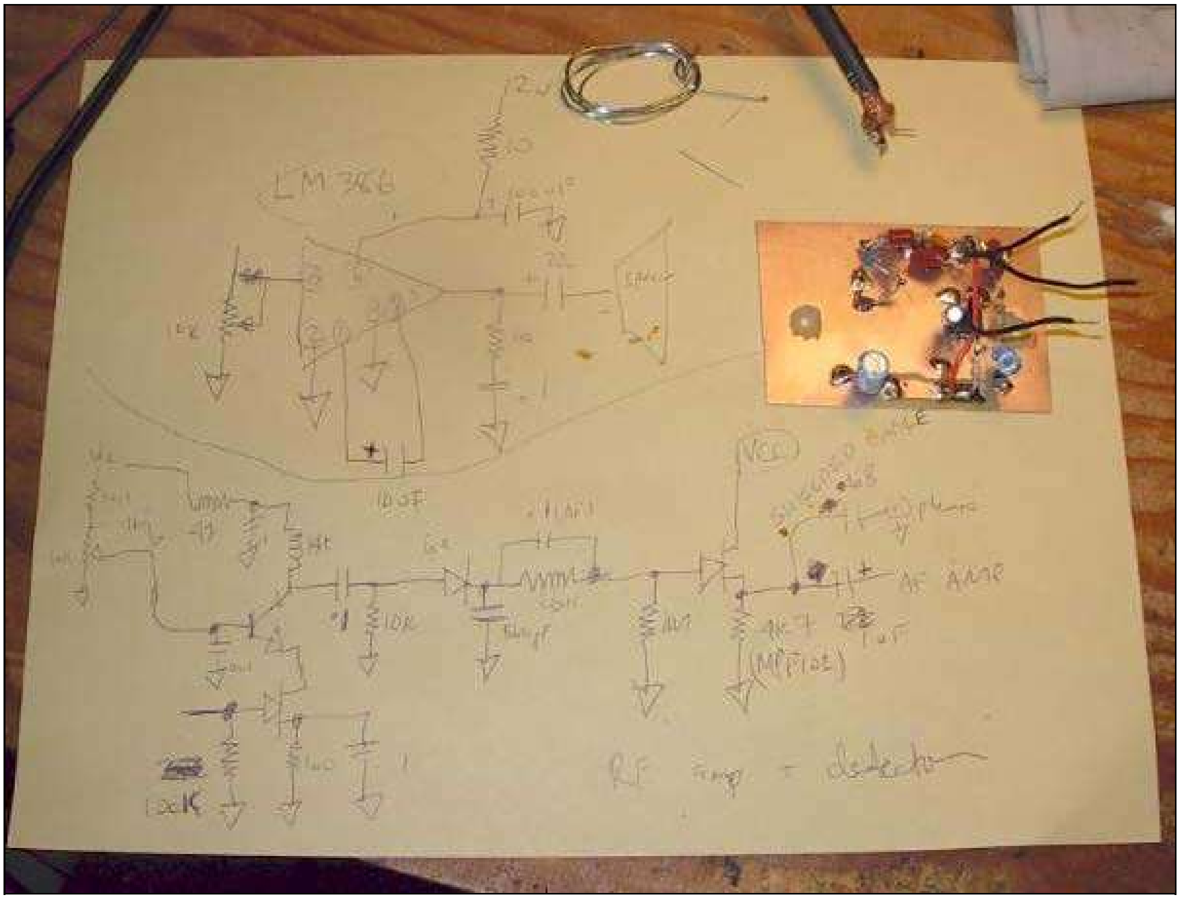
LM386
12V
10
100uF
10K
.1
10uF
10K
.1
4.7uF
10K
VCC
AF AMP
1uF
4K7
(MPF102)
1M
100K
RF amp + detector

A work bench photo of the first version of the W7ZOI detector built in October 2007 is shown above.

RF — Test and Measurement

My Links

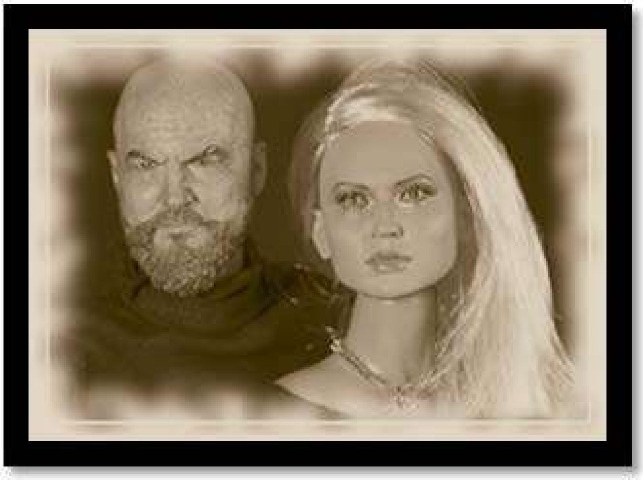

Greetings! I enter the QRPHB / SWL HomeBuilder site here and linked a few of my favorite tools, part suppliers and URLs I like and wish to remember. This page might feel intransigent.

In the following section called **Graphics**, I show the odd newer schematic or photo unpublished as permanent site content. A schematic might just represent my mistaken notion or whimsy: a sense of what's caught my fancy — not that my opinion is better than anyone elses.

Quick Navigator

Graphics

Russian giant sunflower — grew to over 3 meters tall

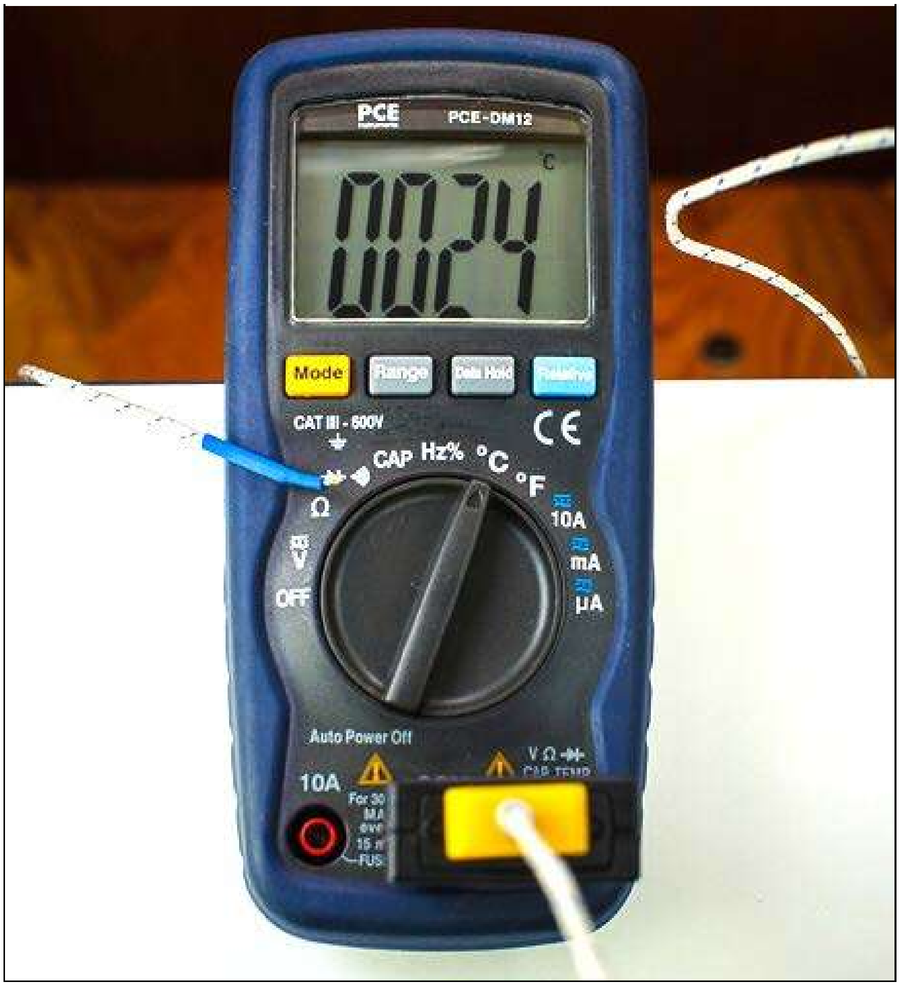

My gorgeous new PCE multimeter just in time for Season 16 experiments starting October 3. Click for another photo. Click for the PCE website — test equipment galore.

RF Transistor Data Spreadsheets

FET

BJT

Test Equipment

Test equipment

USA Rigol Dealer T Equipment

Online Tools and Publications

Standard resistor and cap values

Smith Chart information

Good inductive/capacitive reactance calculator other tools available

Resistive pad attenuator calculator

Ohm's Law 1 pager

VHF Communications website UKW-Berichte/VHF-Communications

The Signal Path EE video blog They review and show the DSA1030A SA + TG

DUBUS Technical magazine for 50 MHz and above

Favorite Experimenters

Esteemed QRP homebrew mentor - Wes Hayward's, W7ZOI web site

🍁 Iulian, YOUDAC-VA3IUL Comprehensive design information and amazing circuits web site

🍁 Markus Hansen, VE7CA A venerable Room 19 Radio School graduate. He has authored for QST. See his receiver featured in QST for March 2006

🍁 Blog of fellow VE7-land experimenter Steve, VE7SL. Look for a link to Steve's web site on his blog

Professor Ken Kuhn's web site. Perhaps start on his site map. A treasure trove of useful information

Ripples in the Ether An always interesting blog from our friend Jason who runs Etherkit

SV3ORA's Site Friend Konstantinos from Greece. Design info galore. Check out his microwave links.

FM Tuner Site Vintage FM Tuner Information

Paul, KE7HR Caver website

Matjaž, S53MV Microwave and RF. My favorite microwave designer. See this page for some inspiration

Paul Wade, W1GHZ authors my favorite QST column *Microwavelengths* His site

Jacques, VE2AZX web site Retired EE with great ideas

Michigan QRP Club Publish *The 5 Watter* Quarterly

Website of Michel, **F6FEO** from France. F6FEO project pages: click and click

Roland, a retired EE who lives in France keeps this wonderful test equipment page

JF1OZL Mega site that first inspired me years ago. He once wrote me a letter - great guy

Brian, K6STI 88-108 MHz web site.

Leif, SM5BSZ Amazing work

KA7OEI Microwave Page web site

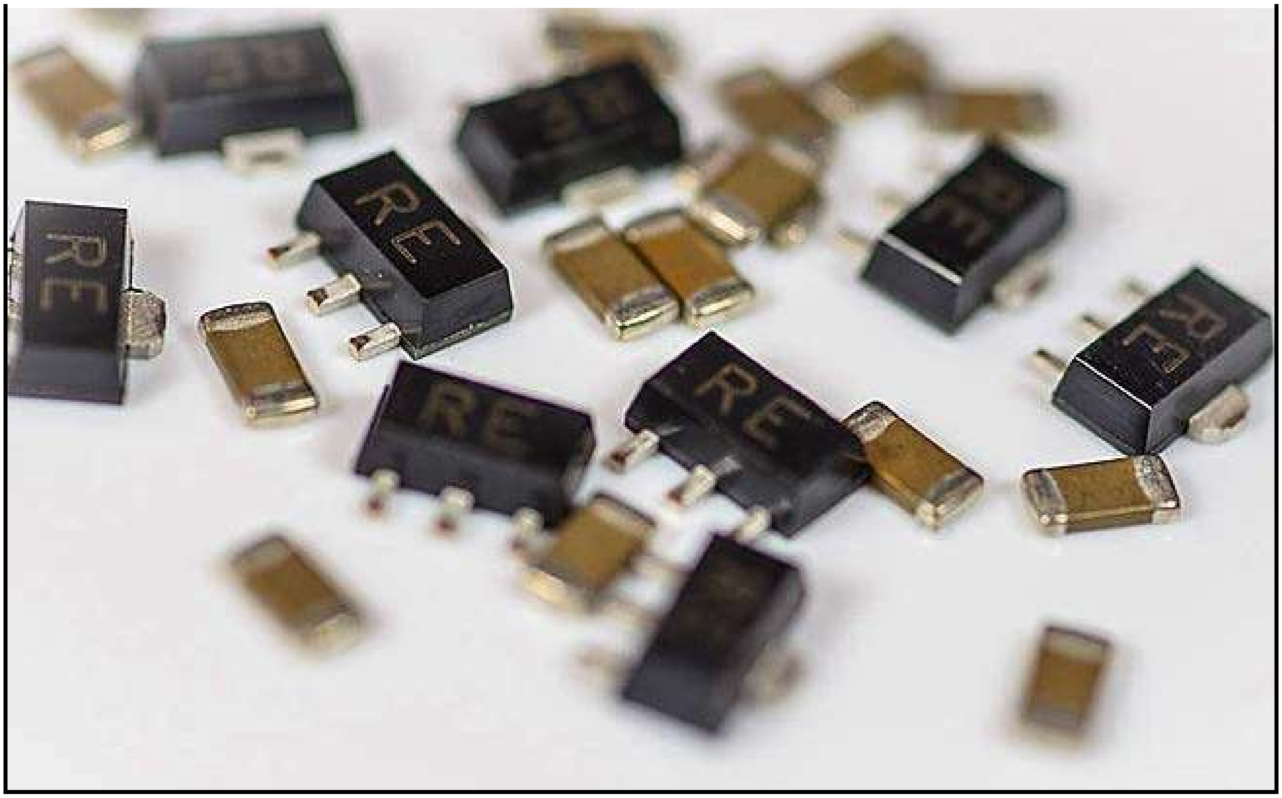
RE
RE
RE
RE
RE
RE

ПОЧЕТНАЯ ГРАМОТА

Memorial Museum of the Leningrad Siege, St. Petersburg, Russia VE7BP0 photos 2006

Государственный музей обороны и блокады Ленинграда

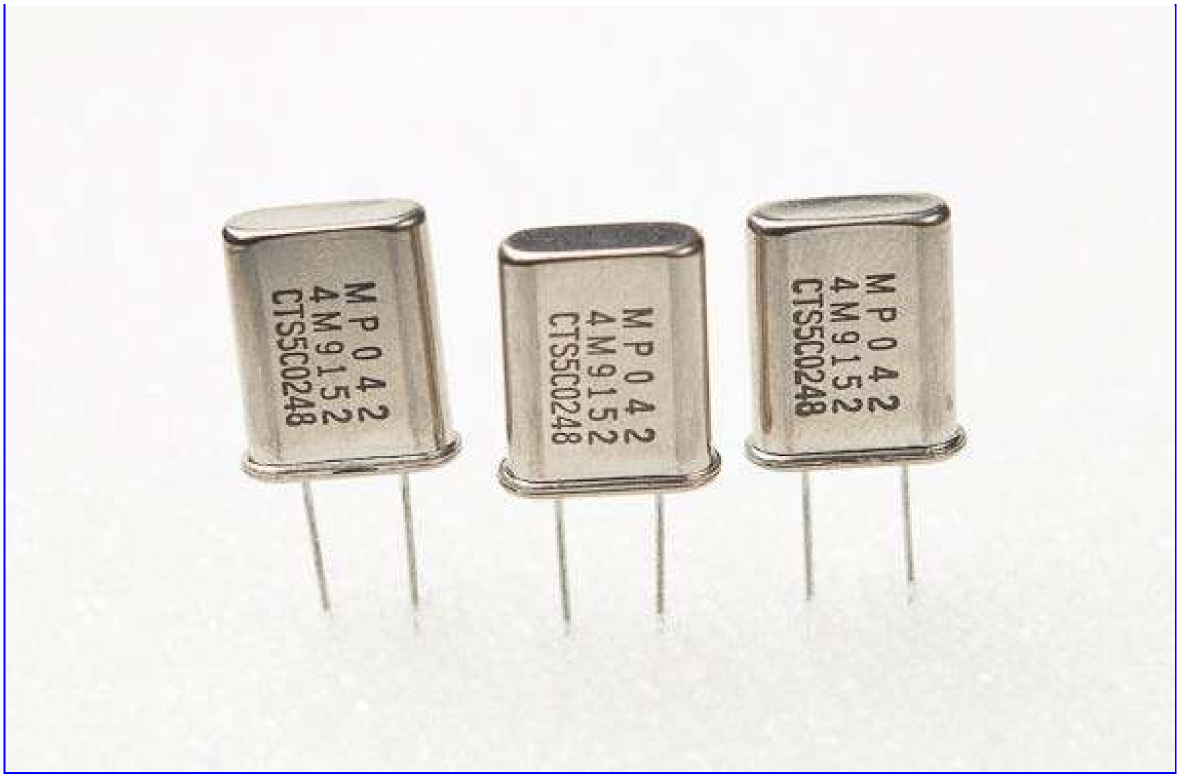
MP042
4M9152
CTS5C0248
MP042
4M9152
CTS5C0248
MP042
4M9152
CTS5C0248

RF — Test and Measurement

RF Workbench Page 1 - Case Study Answers

This web page supplements RF Workbench Page 1

Answers:

1. Calculate the power in dBm at point B **= 13.75 dBm**
2. Calculate the power in dBm at point C **= 10 dBm**
3. What is the attenuation in dB of the 50 ohms attenuator pad? **= -3.75 dB**
4. What is the output power in mW of this stage? **= 10 mW**

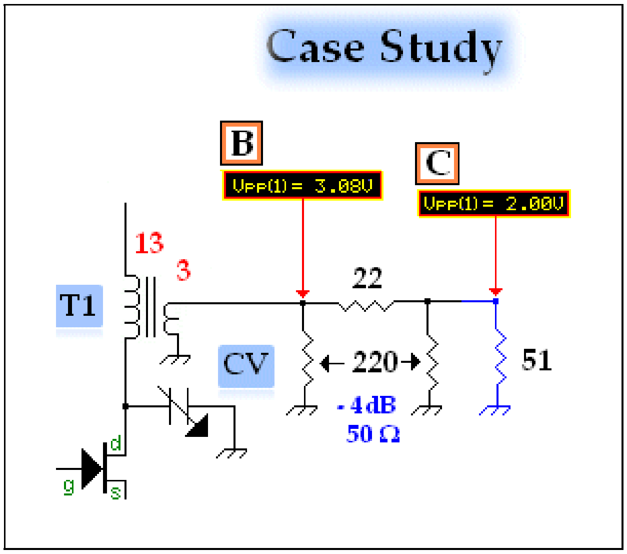
Case Study
B
Vpp(1)= 3.08V
C
Vpp(1)= 2.00V
13
3
T1
22
CV
← 220 →
51
- 4 dB
50 Ω
d
g
s

The design contained a "theoretical" - 4 dB, 50 Ω attenuator pad using standard value resistor substitutions

QRP SWL
HOMEBUILDER

RF — Test and Measurement

50 Ohms Attenuator Pads

This web page supplements RF Workbench Page 2

Table For Pi and T- type pads

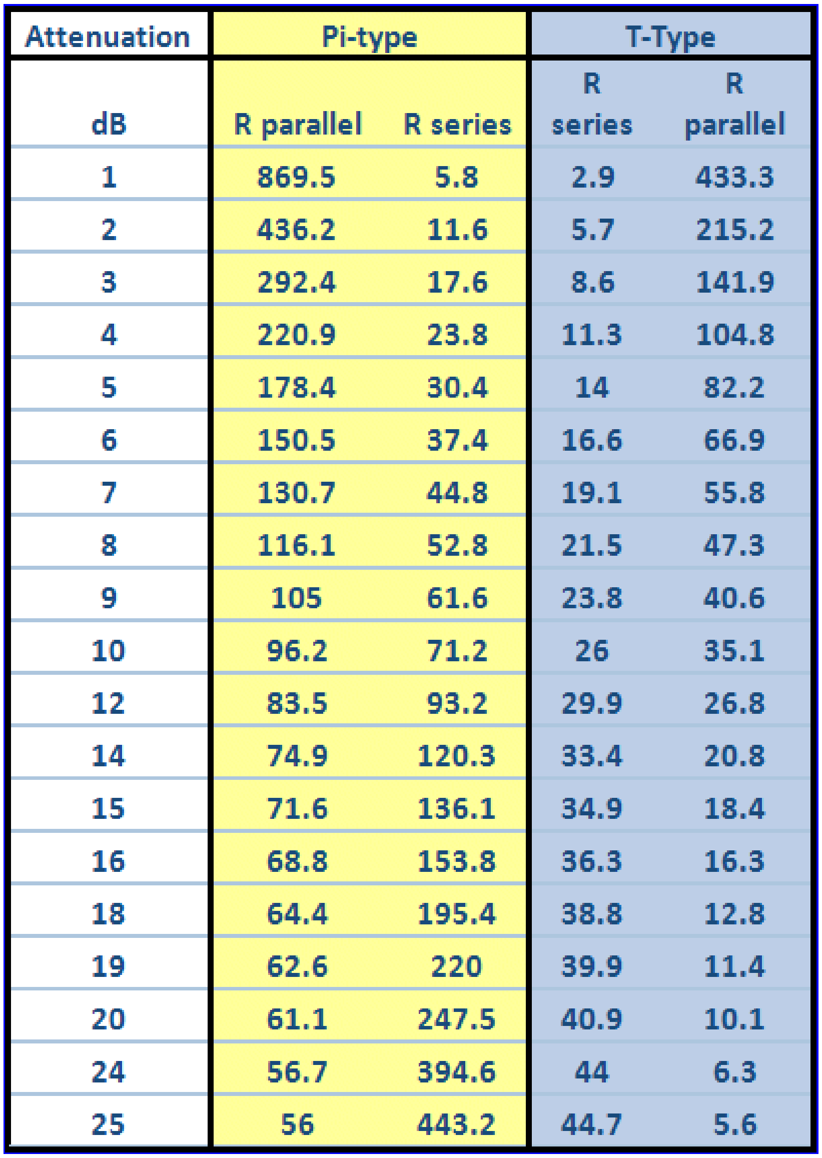

Attenuation	Pi-type		T-Type	
dB	R parallel	R series	R series	R parallel
1	869.5	5.8	2.9	433.3
2	436.2	11.6	5.7	215.2
3	292.4	17.6	8.6	141.9
4	220.9	23.8	11.3	104.8
5	178.4	30.4	14	82.2
6	150.5	37.4	16.6	66.9
7	130.7	44.8	19.1	55.8
8	116.1	52.8	21.5	47.3
9	105	61.6	23.8	40.6
10	96.2	71.2	26	35.1
12	83.5	93.2	29.9	26.8
14	74.9	120.3	33.4	20.8
15	71.6	136.1	34.9	18.4
16	68.8	153.8	36.3	16.3
18	64.4	195.4	38.8	12.8
19	62.6	220	39.9	11.4
20	61.1	247.5	40.9	10.1
24	56.7	394.6	44	6.3
25	56	443.2	44.7	5.6

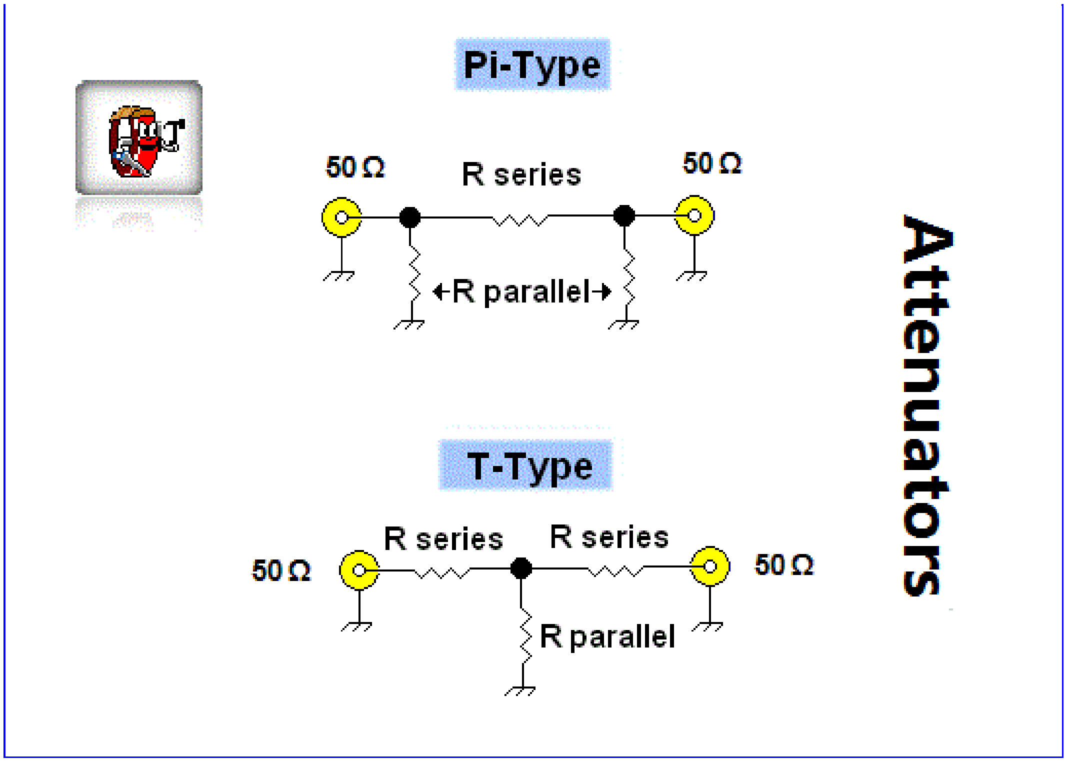
Pi-Type
50 Ω
R series
50 Ω
←R parallel→
T-Type
R series
R series
50 Ω
50 Ω
R parallel
Attenuators

To build attenuator networks, we normally choose use nearest standard resistor values, or parallel/series 2 or more resistors from our resistor collection to approximate the tabled R. If you're building with 5% parts, consider measuring a few resistors with an ohmmeter to find the best match to the tabeled R.

Resistors in the 8 dB T-type or the 4, 14 and 19 dB Pi-type networks are close to 5% standard resistor values.

RF — Test and Measurement

Expected Return Loss Table (50 ohm bridge)

This web page is a supplement to Transistor Input Impedance Experiments

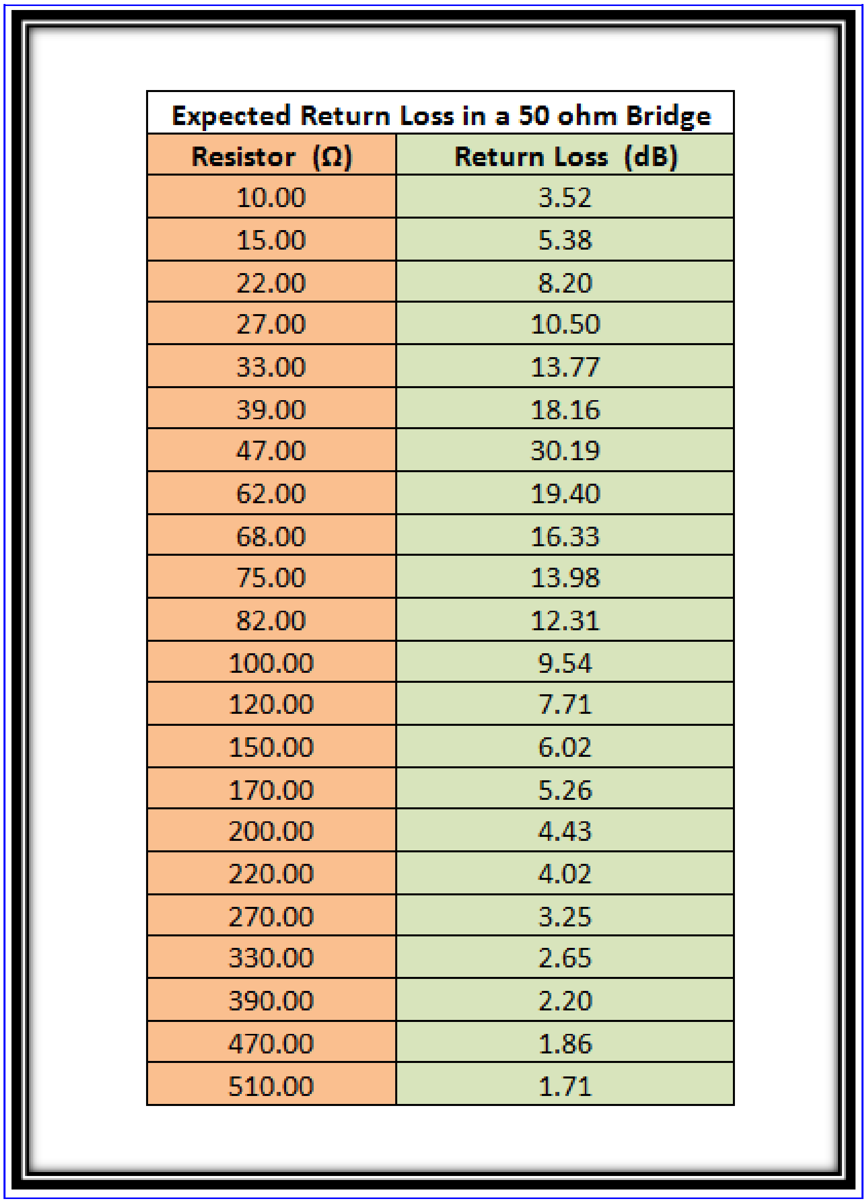

Expected Return Loss in a 50 ohm Bridge	
Resistor (Ω)	**Return Loss (dB)**
10.00	3.52
15.00	5.38
22.00	8.20
27.00	10.50
33.00	13.77
39.00	18.16
47.00	30.19
62.00	19.40
68.00	16.33
75.00	13.98
82.00	12.31
100.00	9.54
120.00	7.71
150.00	6.02
170.00	5.26
200.00	4.43
220.00	4.02
270.00	3.25
330.00	2.65
390.00	2.20
470.00	1.86
510.00	1.71

RF — Test and Measurement

Miscellaneous RF Topics 2011 Supplement

Introduction

This web page supplements the Miscellaneous RF Topics 2011 web page.

1. Design Center - Wrap Around Bias

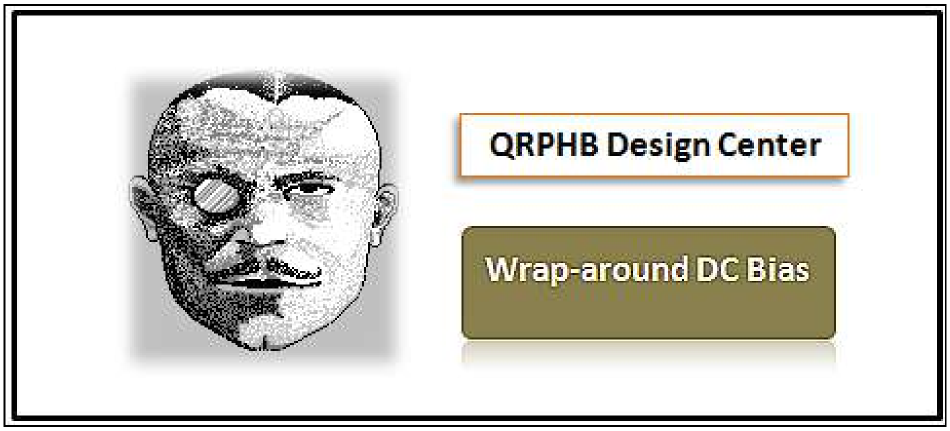

PNP Wrap-around Bias for an NPN RF Amp

There are entire books written on RF power amps. I don’t pretend to be an expert and offer a simple approach to help move novice builders up the learning curve. Big thanks to Wes, W7ZOI for the theory and explanation for EMRFD Figure 2.16; the wrap-around PNP bias of an NPN transistor with grounded emitter. This bias circuit is fantastic and helps prevent you from burning up too many expensive medium power transistors such as the 2N3866 or 2N5109 during experiments and operation. From our emails, the following experiments were performed and information gleaned.

Since the NPN emitter terminal is grounded or near to ground on my web page examples, the current can really run high — Don't forget to use heat sinks and appropriately rated resistors on the NPN circuit when you really go for power.

Wrap Around PNP Bias

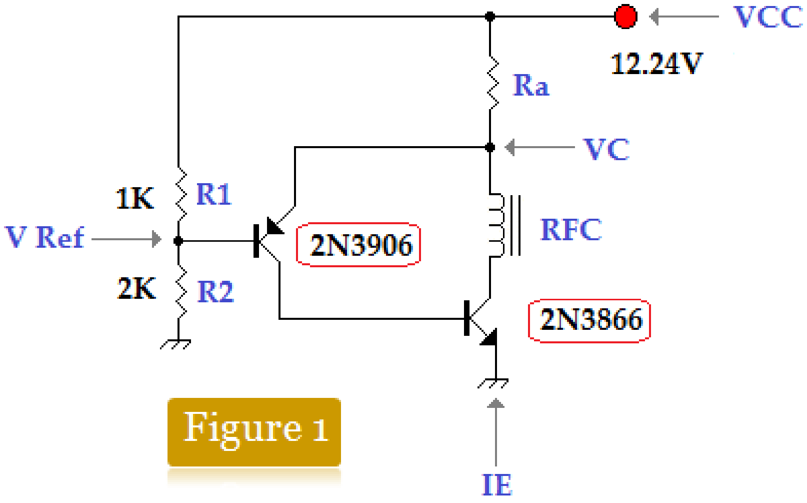

Figure 1

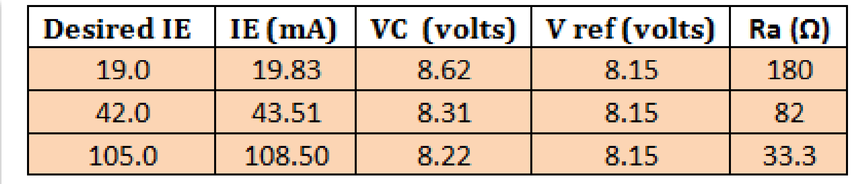

Desired IE	IE (mA)	VC (volts)	V ref (volts)	Ra (Ω)
19.0	19.83	8.62	8.15	180
42.0	43.51	8.31	8.15	82
105.0	108.50	8.22	8.15	33.3

Above — The schematic and data table from some biasing experiments. The normal decoupling parts were removed to isolate the biasing circuitry. All voltage and current values were measured. Another photo. A folded piece of paper is under the wire connecting the NPN emitter to the copper clad board. This is how I measure emitter current. The paper acts as an insulator and allows you to press hard with the ammeter probe to ensure good contact. After current measurement, the paper was discarded and the emitter soldered to the copper ground plane.

Biasing Procedure

Step 1: Choose a Target VC

Using collector resistor Ra, we want to drop about 2/3 of the power supply voltage across the NPN transistor. For a 12 volt supply or VCC, this is 8 volts VC. The formula is Target voltage = VCC x 0.66. Why choose 2/3 VCC instead of ½ VCC like we do for Class A audio amplifiers? In audio amps we just use a resistor as the load and the collector voltage can't go above the power supply unless we insert some reactances in the circuit. In simple AF voltage amps, the half VCC supply allows the greatest undistorted, peak-to-peak AC signal swing.

In RF power amps, a transformer or RFC lies on the collector terminal, so the instantaneous collector voltage (sum of the supply voltage plus the peak AC voltage) can go well above the supply voltage. If you have 8 volts on the collector, VC can easily rise to 16 volts or so on signal peaks. It's often prudent to choose a VC closer to VCC than the classic ½ VCC of audio amplifiers.

Hereafter, Target VC is referred to as **VRef**

Step 2: Choose R1 and R2 to bias the PNP @ VRef

Choose the PNP biasing transistors R1 and R2 so that the PNP bias = VRef. The formula follows:

$$\text{VRef} = \frac{\text{VCC x R2}}{\text{R1 + R2}}$$

I wrote a Java Applet that does the math Applet A ; just plug in a VCC and play with the resistor values until you get your target VRef with the resistors from your parts collection. Try to get a bias as close as possible to the targeted VRef, but don’t stress out over it.

Step 3: Choose Ra to give (VRef + 0.6)

Ra = VCC – (VRef + 0.6) / desired current in amperes. Use or make up Ra with nearest standard resistor value(s) with an appropriate power rating. In a PNP, the emitter terminal is 0.6 VDC higher than the base, so need to we add that to VRef. The desired current is any reasonable current that you want to set up; for example 25 mA or 75 ma.

Example Calculation

I wanted a emitter current of 105 mA from the Figure 1 circuit. VCC = 12.24. Let's run through the steps.

Step 1: Choose a Target VC

Target voltage = 12.24v x 0.66 = 8.1 VDC -> Now called VRef

Step 2: Choose R1 and R2 to bias the PNP at VRef

I just plugged some resistor values into Java Applet A. R1 is normally lower than R2.

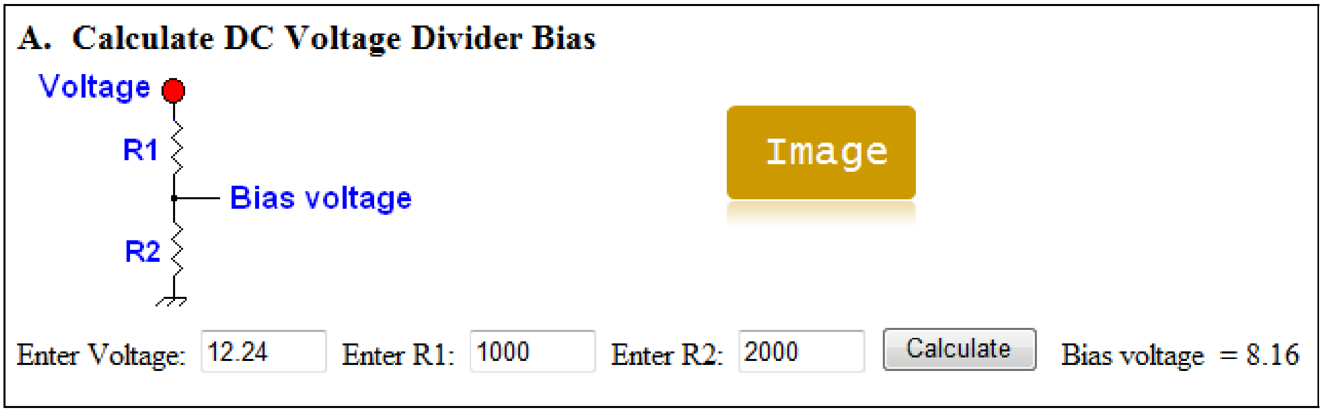

A. Calculate DC Voltage Divider Bias

Enter Voltage: 12.24 Enter R1: 1000 Enter R2: 2000 Bias voltage = 8.16

Above — A practical bias network. I have lots of 1 and 2K resistors. 8.16 v is close to my target VRef.

Step 3: Choose Ra to give (VRef + 0.6)

12.24v – (8.16v + 0.6v) / 0.105 A desired = 33.1 ohms. I put three 100 ohm ½ watt resistors in parallel for Ra. Click for lateral view photo showing the 3 resistors.

The exact measurements of this example and 2 other (19 and 42 mA) are tabled in Figure 1. I used an 82 ohm 1/2 watt resistor for Ra in the 42 mA version. The bench values closely approximate the calculated values. This circuit is quite instructive. In a real circuit (Figure 2) we add in the decoupling parts. Ra performs double duty as the decoupling resistor as well as the current regulator resistor.

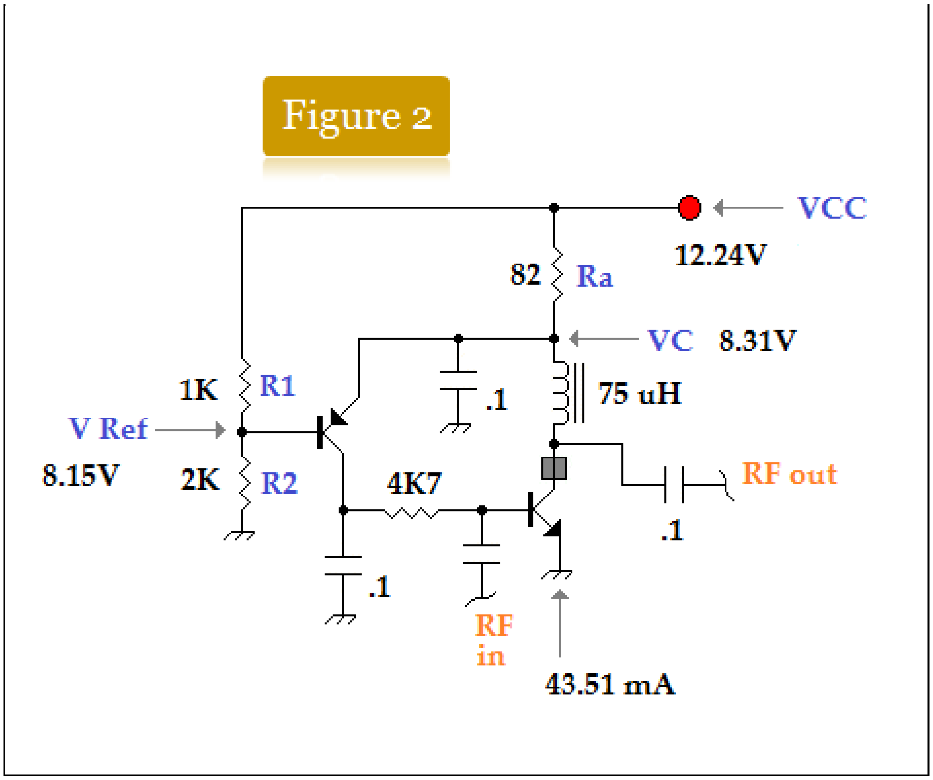

Above — A working version of the desired 42 mA medium power amp from Figure 1. Refer to EMRFD Chapter 2 for further information.

How does the PNP wrap-around bias loop work?

Using the sample conditions from Figure 2: VC = 8.16v calculated, Desired IE = 42 mA and VCC = 12.24v
The main decoupling/current limiting resistor Ra is established at 12.24v – (8.16v + 0.6v) / 0.042 mA = 82.9 Ohms.

No matter what, the PNP VRef = 8.16 volts. All of the desired conditions are established and the current is flowing in the NPN. Let's change these conditions and see what happens.

First, instead of 42 mA, assume that the NPN current drops to something less. Hence, there will not be as much of a drop across Ra. This will cause VC to increase. This is the voltage on the PNP emitter. When this voltage increases, there is more voltage between the base and the emitter of the PNP, which will make it draw **more** current. That current comes out of the PNP collector runs to the base of the NPN. This causes the NPN to draw more current (beta times) which increases its collector current and causes the collector voltage to drop because of the I x R drop in Ra. It drops until it reaches the 8.16 volt level.

Let's now assume the opposite. The NPN draw too much current, more than the desired 42 mA. This means that there is too much I x R drop in Ra and VC goes **below** 8.16 volts. This tends to turn the PNP off. The PNP collector current drops, so the NPN base current also drops, causing the NPN collector current to drop. This causes VC to again increase to reach 8.16 volts. Is this cool or what?

---

2. RC Low-pass Network

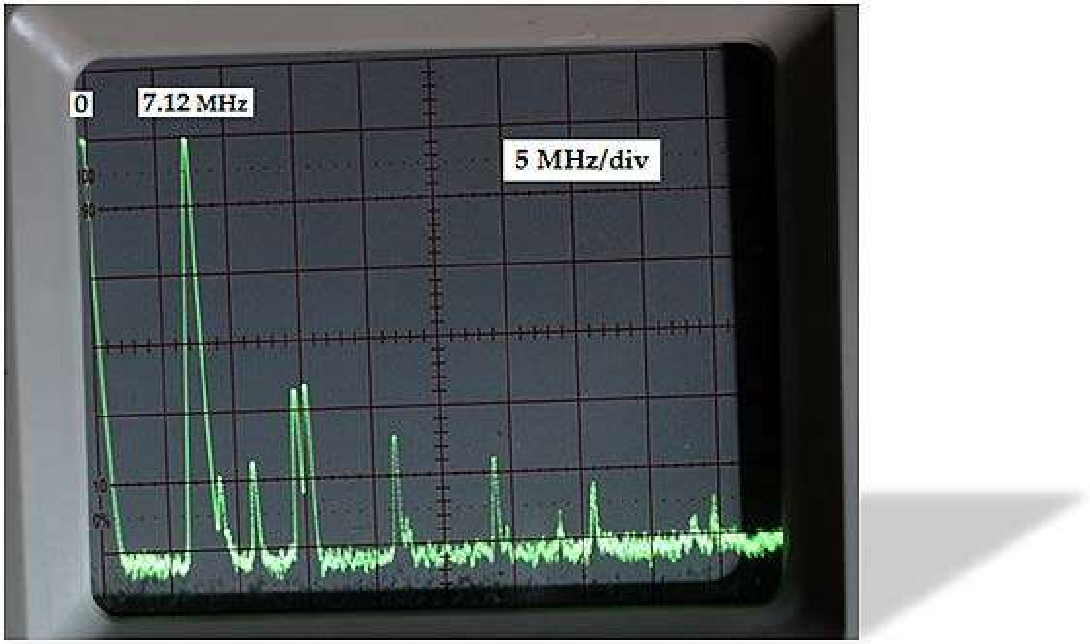

Above — Spectrum analysis without the RC low-pass filter.

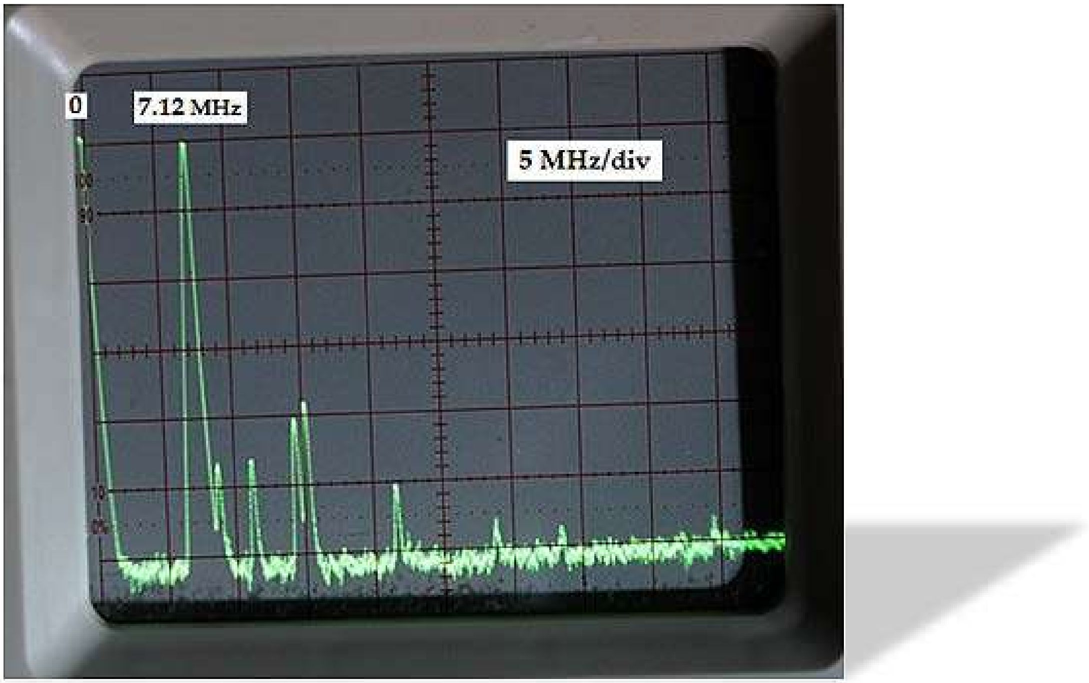

Above — Spectrum analysis with the RC filter in-situ. Above 20 MHz, the spurs are ~5-6 dB down.

3. Transistor Power Dissipation in the 2N3904

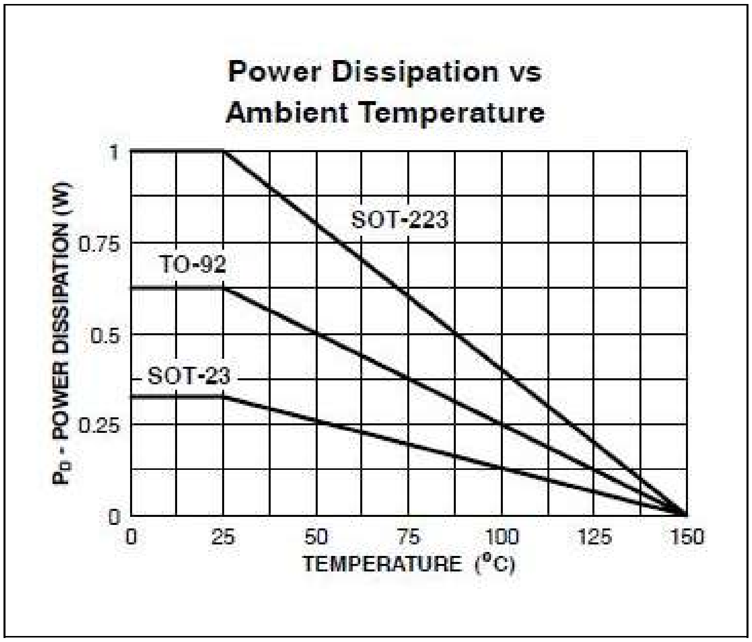

Above — Power Dissipation versus Temperature for the 2N3904 in 3 packages.

Power dissipation in the collector resistor

For example, VC (the voltage across the transistor) = 10 volts and the emitter current = 20 mA

P = I x E = .020 amps x 10 volts = 200 mW

Power dissipation in the emitter resistor

P = I x E = .020 amps x the measured emitter voltage

or P = I x I x R

For example, emitter current = 20 mA with a 5 ohm emitter resistor; P = .020 amps x .020 amps x 5 ohms = 2 mW

We assume IE = IC in the above examples. At 200 mW collector dissipation, the 2N3904 temperature temperature climbs to about 65 C. A metal can transistor such as 2N2222 has more surface area and better heat dissipation. A heat sink also improves heat dissipation. Temperature dissipation charts are referenced to ambient temperature which may be dependent on the air circulation around the transistor. If you plan on running 200 mW or so collector dissipation with a 2N3904, glue on a heat sink. A pinch of the finger and thumb, or a infrared thermometer are the most common popcorn ways to test for the need for more heat sinking, or a part with better power dissipation.

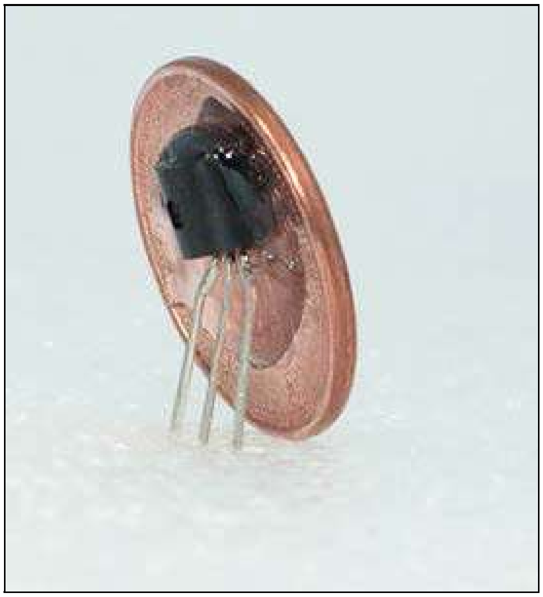

Above — I learned this from a Russian builder. 100 heat sinks for a dollar. Typically, a small blob of epoxy with the transistor positioned to the lower edge of the penny works best. I used too much epoxy and didn't wait for it to completely dry before taking this photograph — the penny slid down.

---

4. Some Factors Affecting Common-Gate Amplifier Input Impedance

A question arises: with all the devices and topologies available, why build a common-gate JFET RF amp in 2011 and on? For the J310 JFET at least: they're cheap, the input impedance is close to 50 ohms and wideband, the amp demonstrates good stability, signal handling capability and linear performance — all without running mega-current. Мне нравится. Many authors state the common-gate noise figure is inferior to that obtainable with the common-source topology, but for the popcorn builder, the common-gate noise figure is probably reasonable.

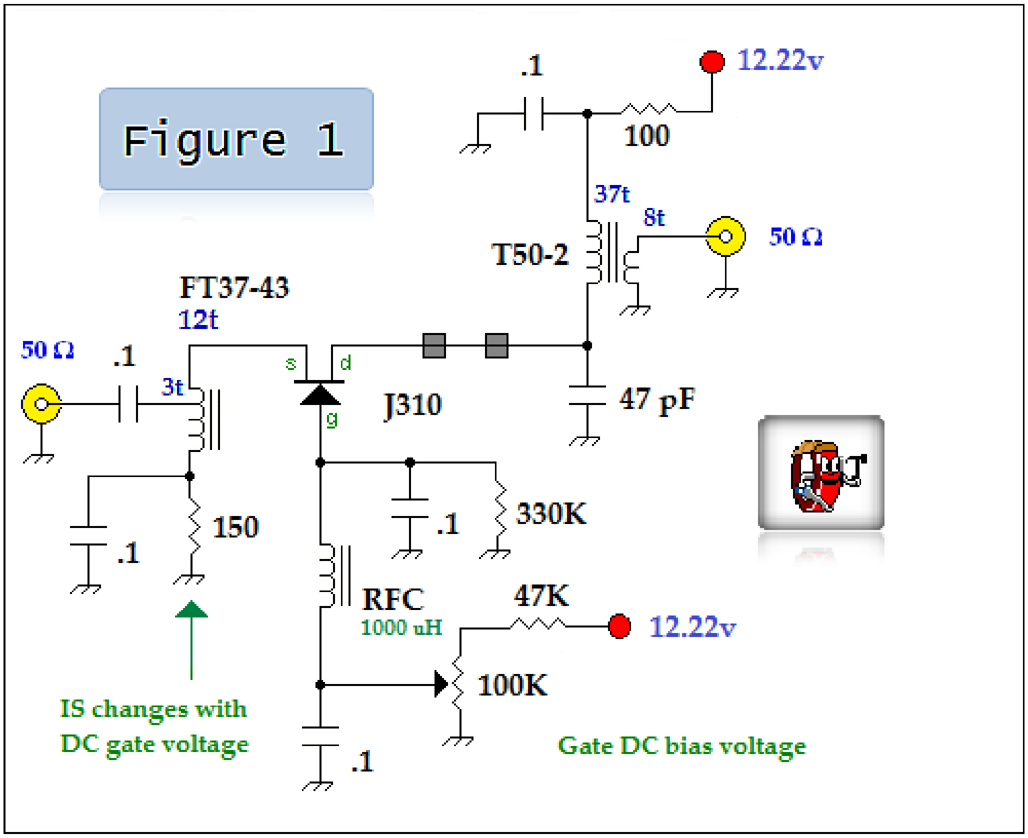
Figure 1
.1
12.22v
100
37t
8t
50 Ω
T50-2
FT37-43
12t
50 Ω
.1
3t
s
d
g
J310
47 pF
.1
330K
150
.1
RFC
1000 uH
47K
12.22v
100K
IS changes with
DC gate voltage
.1
Gate DC bias voltage

Above — A JFET gate biasing experiment. Here I applied 0-7 volts DC to the gate of a common-gate FET amp. The results proved interesting; see the graph below. The drain was tuned with a fixed 47 pF cap and manually "scrunch tuned" by pinching the T50-2 toroid with fingers and thumb. The input RFC serves to decouple the DC from the AC signal and the tap is used to improve the input impedance match.

Figure 2

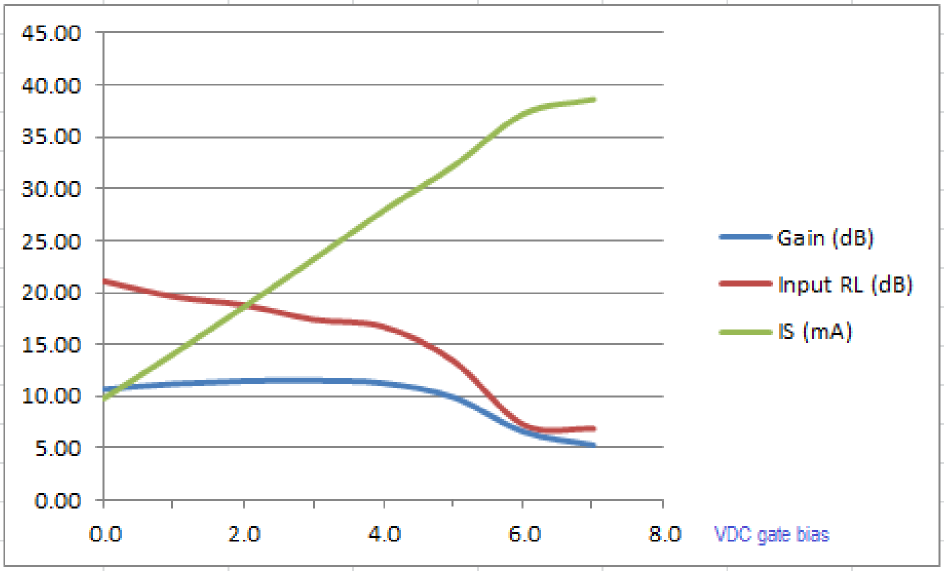

Above in Figure 2 — A graph of the DC gate voltage versus gain, input return loss and source current from Figure 1.

Increasing the DC gate voltage increases the source current as shown in the graph — ranging from 9.8 to 38.6 mA. From a DC gate voltage of 0 to about 4 volts, amplifier gain changes very little. Technically, the gain of this amplifier should increase along with drain current, however, the input resistance simultaneously goes down, creating greater voltage division loss at the input plus reduced return loss from impedance mismatch. This tends to cancel the increased gain as the plot shows.

The AC signal current injected into the FET comes out of the drain with power gain because that same current is now flowing into a higher impedance. The gain is dependent on the ratio of the impedance at the drain to that at the source. Click for a SPICE plot of DC gate voltage versus source current for 0 - 5 volts from Wes, W7ZOI.

From experiments, I learned that at least 6 things effect input return loss: the source resistor value, stacking or paralleling JFETS, the gate DC voltage(all which change current and/or transconductance), whether or not you bypass the source resistor, negative feedback and also the tap on the input RFC. Like other amps, the impedance of the device before and after the JFET amp can also change the input match, but those are external and normal considerations.

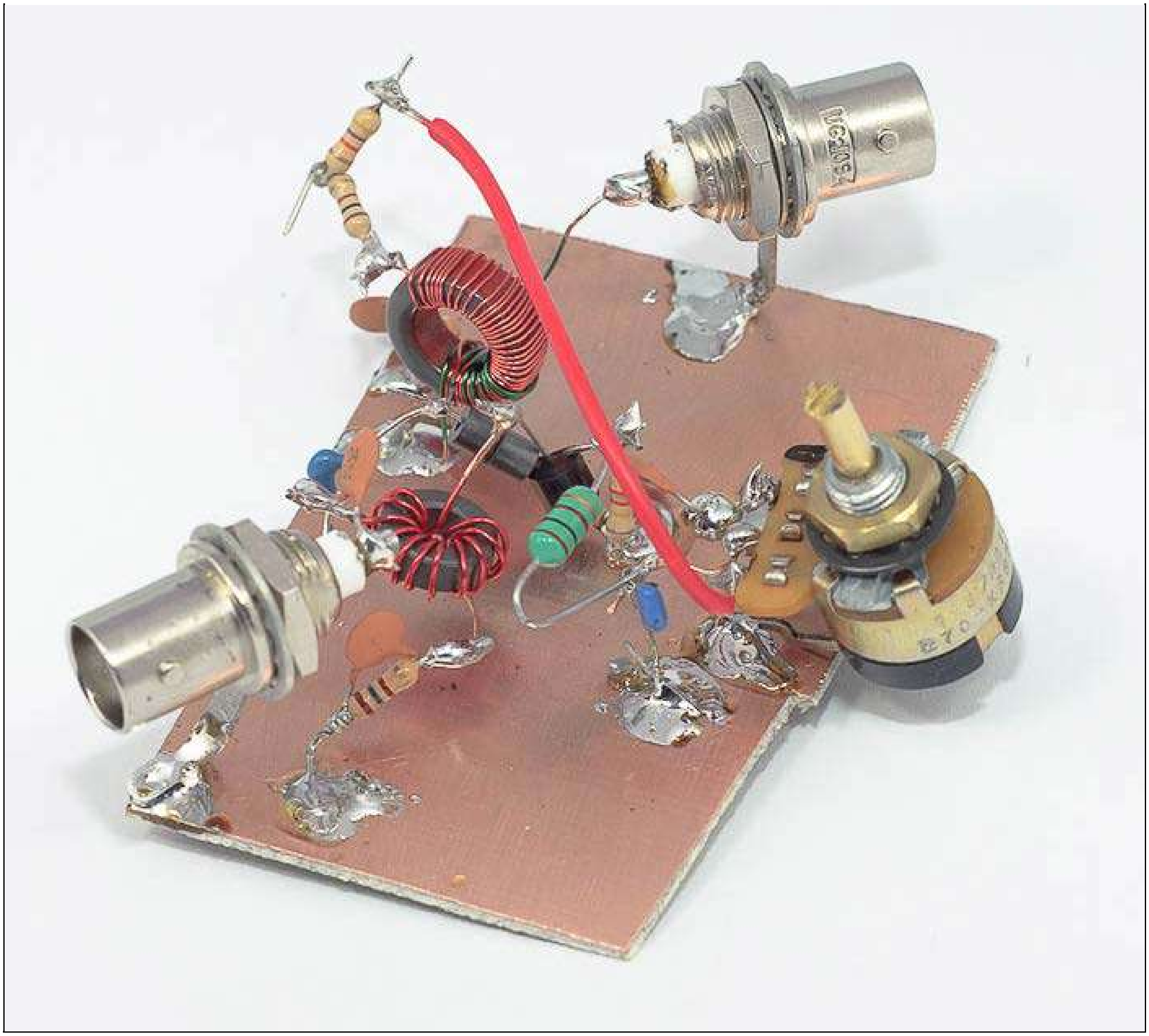

Above — Breadboard of the DC gate bias experiments of Figure1. Properly biased JFETs act like a current source where drain current is controlled by the gate-to-source DC voltage. The input impedance of the common gate amp is the reciprocal of the transconductance. Ken Kuhn posted some great JFET design notes in the yellow JFET section on his web site. EMRFD provides another essential reference for RF JFET design.

Noise Figure and the Common Gate Amplifier

Noise figure is important in low-level RF amplifiers such as a preamp or post-mixer amp because noise generated by these amplifiers contributes to the overall noise figure of a receiver. The input impedance an amp like the common-gate is considered an *active* impedance. Often, with active impedances, obtaining the maximum power transfer and lowest noise figure don't occur simultaneously. That is, the optimum input impedance match for gain is not the optimum input match for low noise performance. With pure resistances, the best noise figure possible is 3 dB; however, when using an active impedance, it's possible to achieve a noise figure less than 3 dB by applying negative feedback or other techniques to manipulate the JFET active input impedance and lower its voltage noise.

A stellar example by Bill, W7AAZI employs DC gate bias, noiseless negative feedback and paralleled JFETs to achieve a noise figure in the 1.2-1.5 dB range. This mind-boggling circuit is described in EMRFD as Figure 6.94. I wanted to explore this topic: can amateur builders design *common-gate* amplifiers for a lower noise figure? The answer is perhaps, however, unfortunately, there is no cookie-cutter approach.

Low noise amplifier design imposes an advanced topic and many builders lack the math skills and/or test equipment to design and generate results. Each design must be approached on a case-by-case basis, although similar fundamentals apply to all cases — using active impedance manipulation techniques and paralleling up JFETs while paying attention to all the resistances in a circuit — even loss resistance in the inductors. I learned from Professor Ken Kuhn that when amplifiers are paralleled the output signal power adds linearly, whereas the internal noise adds statistically (square root of the sum of the squares) and thus noise figure improves. In short, paralleling JFETs reduces voltage noise, but only to a point, as input capacitance also increases which can degrade high frequency performance and noise figure. Ideally, the JFETs paralleled should be matched for IDSS and VP.

Further, all devices have a voltage noise specifications (in volts per root Hz) and a current noise specifications (in amperes per root Hz). The source impedance that produces the lowest noise is the ratio of the two — derived from Ohm's law. The input impedance of the device may be higher or lower; thus the optimum source impedance for low noise is not necessarily the optimum source impedance for maximum power transfer. It's unlikely that these two impedances differ widely, however, only careful bench measurements or closely copying proven designs will ensure the desired results.

---

5. Design Center - Temperature Stable Voltage Divider Bias

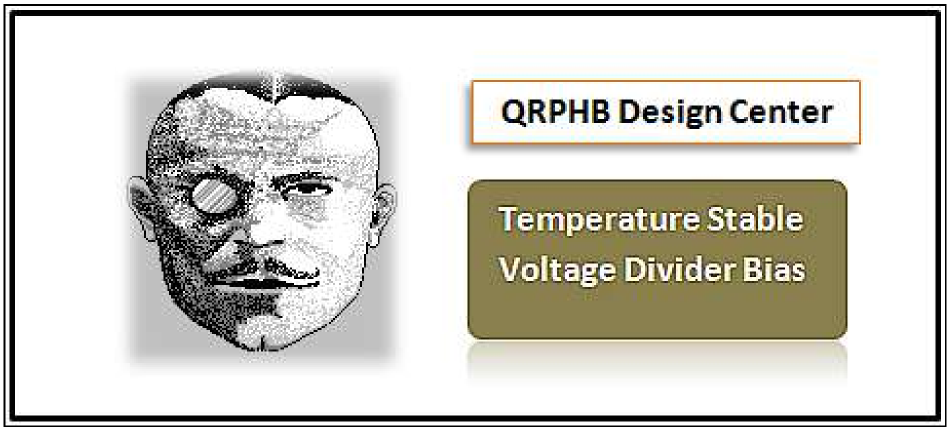

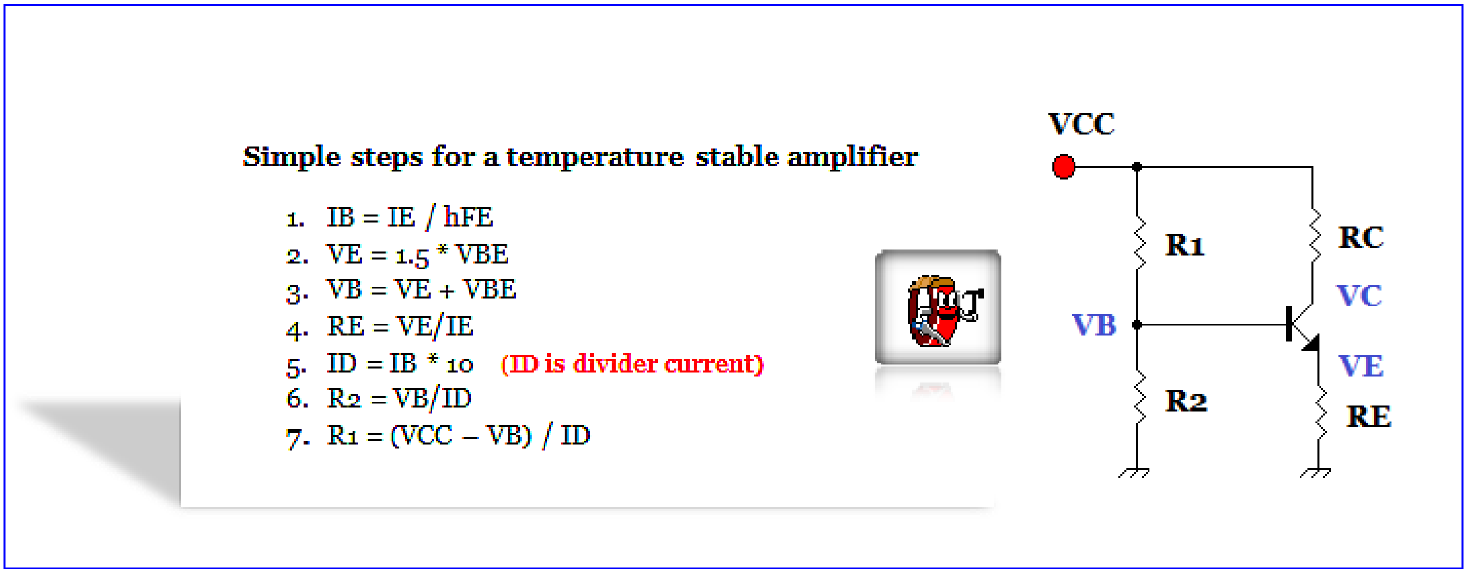

Simple steps for a temperature stable amplifier

1. $IB = IE / hFE$
2. $VE = 1.5 * VBE$
3. $VB = VE + VBE$
4. $RE = VE/IE$
5. $ID = IB * 10$ (ID is divider current)
6. $R2 = VB/ID$
7. $R1 = (VCC – VB) / ID$

Above — A simple way to design a temperature stable BJT voltage divider bias. Assumes that IC = IE. You specify IE.

I sought to develop a simple way for builders to design a temperature stable BJT amp using voltage divider bias. It wasn't so easy. Here is the full math from Ken Kuhn. As shown, it gets complicated and over-the-head of the target readers of this web site.

My simple approach makes assumptions, and certainly an astute builder could improve my algorithm - I developed other formulae, however settled on this one because it well fits the scope of the QRP/SWL HB web site. Most engineers design their bias networks in SPICE and I recommend this.

To choose a Beta or hFE, you can measure the Beta or use the transistor datasheet Beta near to the emiiter current you intend to apply. Many builders debate this and even go as far to state that spec sheet Beta values are useless. This may be true, but you *have* to decide on some approach. Ultimately, you will build the amplifier and can tweak 1 or more of the bias resistor values to set your bias voltage (emitter current). Click for a snippet from a 2N3904 datasheet showing hFE at a range of currents.

Biasing BJTs is a great example of building to suit your design requirements. The average reader likely operates their radios at room temperature where a minimum temperature stability factor is needed. On the other hand, you may operate field portable, or use a crammed chassis where transistors heat threatens stability. As always, your design choices must fit your needs, parts and abilities.
The above 7-step method forms a simple **QRPHB Design Center** method to achieve a reasonable temperature stability factor at normal

operating temperatures. I considered 2 factors: VBE and Beta (hFE) changes.

VBE

We normally attribute the voltage difference from base to emitter (VBE) as 0.7 volts for silicon transistors. VBE changes inversely with temperature and at AF to HF, we mitigate VBE creep by applying an emitter resistor to increase the emitter voltage VE. The emitter resistor provides a feedback loop so as collector current starts to increase due to a rise in temperature, the voltage drop across RE also starts to increase. This voltage tends to reverse bias the base-emitter junction. The net effect is decreased collector current.

How do I choose an emitter resistor value for VBE bias stability asked the little grasshopper (novice)?

The VE value should typically be about 10 to 25% of VCC. VE recommendations really vary from author to author — some say VE should be at least 5X VBE while other authors suggest 2-4 volts as the minimum acceptable VE value for bias stability in typical AF-HF amplifiers. I use 1.5 * VBE (0.7) = a VE of 1.05 volts, but it's up to you. Since you know IE for your transistor, calculate the resistor RE to provide the target VE using ohm’s law: RE = VE/IE.

At VHF on up, emitters are typically directly connected to ground since that path must have as near zero inductance as possible. In these amps, the bias is derived by a feedback circuit that controls the base current to stabilize the collector current at a specific level. Inductance in the emitter lead causes significant loss of gain at higher frequencies and also upsets the input impedance to the base. Similar effects occur in JFETs and even vacuum tubes. 10s of nH of inductance at 144 MHz represents a reactance of 10s of ohms. At UHF on up this becomes very serious — that's why construction methods are so critical for these builders.

In conclusion; to influence VBE stability against temperature changes and to some extent leakage current, at AF- HF, apply series feedback to a raise emitter voltage (VE). Accounting for variations arising from 5% tolerance parts, I chose a minimal VE of 1.05 volts (equation 2 of 7 in my "Simple Steps..." algorithm).

A tradeoff challenges us — increasing the emitter resistance, increases AC degeneration, lowers AC signal gain and raises the transistor input impedance. Bypassing all or part of the emitter resistance offsets these problems.

Beta

Beta variation arising from temperature changes must be mitigated. I briefly discuss the factors that change Beta in the bullets under the #5 Bipolar Junction Transistor Beta Tester on this page.

To design "beta independent amplifiers", the builder alters the resistance ratio RB/RE. Decreasing this ratio improves the Beta stability while decreasing the amplifier current gain. The ratio can be confusing, so the Design Center equations 5, 6 and 7 just make the current in the divider 10X the base current. It's simple — you're done. Feel free to choose whatever ratio you think is best. Good values for the RB/RE ratio lie in the 12 - 20 range.

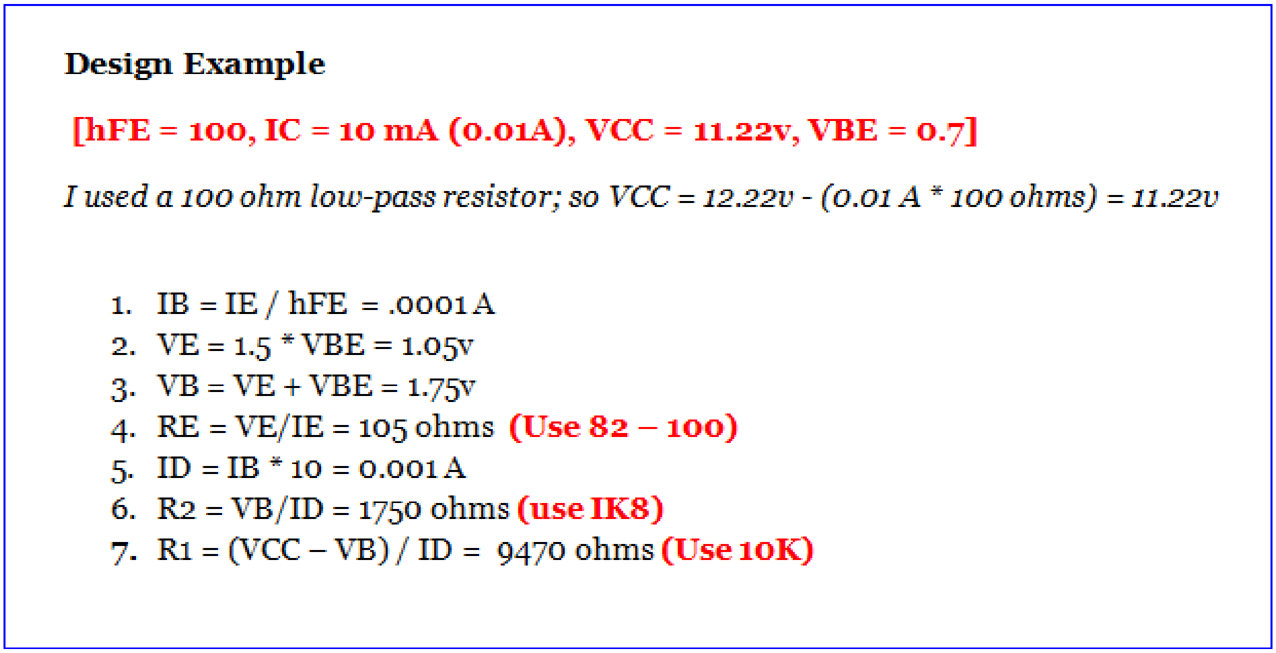

Design Example

[hFE = 100, IC = 10 mA (0.01A), VCC = 11.22v, VBE = 0.7]

*I used a 100 ohm low-pass resistor; so VCC = 12.22v - (0.01 A * 100 ohms) = 11.22v*

1. IB = IE / hFE = .0001 A
2. VE = 1.5 * VBE = 1.05v
3. VB = VE + VBE = 1.75v
4. RE = VE/IE = 105 ohms **(Use 82 – 100)**
5. ID = IB * 10 = 0.001 A
6. R2 = VB/ID = 1750 ohms **(use IK8)**
7. R1 = (VCC – VB) / ID = 9470 ohms **(Use 10K)**

Above — A design example employing my "Simple steps for a temp stable amp" bias algorithm. My VCC = 12.22 volts, but, as always I decoupled and bypassed the VCC — in this case, I applied a 100 ohm /0.1 uF network. At 10 mA emitter current, the voltage drop across the 100 Ω resistor results in a VCC of 11.22 volts.

Big thanks to Wes, W7ZOI and Professor Ken Kuhn for helping me with my math problems and to better understand biasing concepts.

A friend made a Design Center spread sheet for MS Excel. In this application you enter the desired IC rather than IE.

QRP SWL HOMEBUILDER

Amateur Radio Electronic Design

50 Ohm Common Emitter Audio Preamp Experiments

Introduction

This page supplements the Pop DC2 Receiver Experiments from this web page.

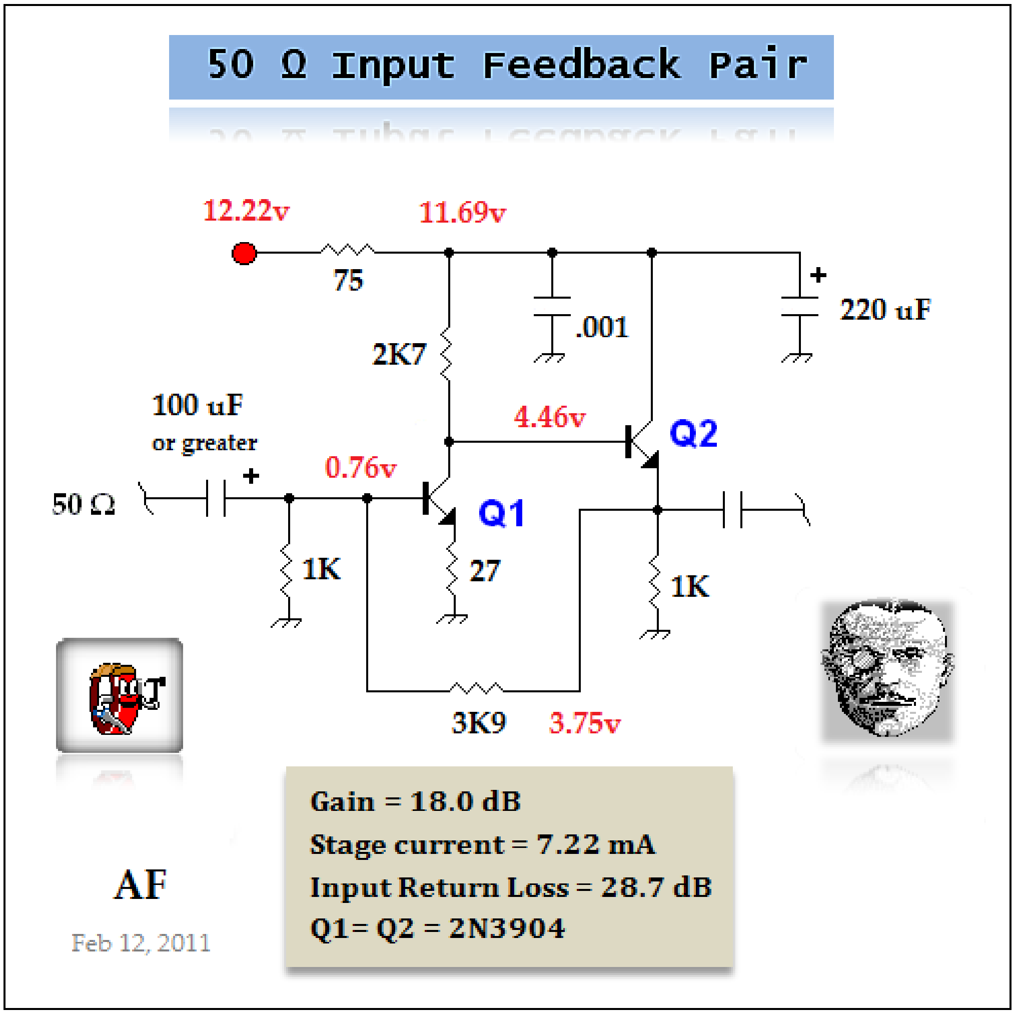
50 Ω Input Feedback Pair
12.22v
11.69v
75
.001
220 uF
+
2K7
100 uF
or greater
4.46v
Q2
0.76v
+
50 Ω
Q1
1K
27
1K
3K9
3.75v
AF
Feb 12, 2011

Gain = 18.0 dB
Stage current = 7.22 mA
Input Return Loss = 28.7 dB
Q1= Q2 = 2N3904

Above — A 50 ohm audio preamp to follow a diode ring product detector.

Many seek an alternative to the common base AF preamp that typically follows a diode ring product detector. Winter experiments yielded some possible ideas. The simplest is shown above.

Using all different parts, I built and measured another version to ensure reproducibility of the input return loss while just using standard 5% resistor values. The second amp proved similar: 18.4 dB gain, 7.27 mA stage current (3 .8 mA is consumed by the emitter follower), a return loss of 27.6 dB and a nice sine wave output. A 2K2 resistor AC coupled to the emitter follower provided a load. Note in the photo, I used 1% resistors in some slots, but the values were standard 5% and not unusual values like 15.4 K etc.. The 3K9 feedback R is the most critical value — it's a 5% part.

About 1/4 of my resistor collection are 1% metal film, as I don't like stocking both 5% and 1% tolerances of the same R value. I envision this amp driving a 5532 amplifier/low-pass filter built with 5% resistors with the op-amp output connected to a 1-10K volume control potentiometer.

The amp above is pure popcorn fluff, yet represents over 8 hours of reading, calculations and bench discovery. In another version, I soldered in low noise transistors (MPSA18 for example) and 1% metal film resistors to try and drop the noise figure. Yet another stage had Q1 collector bootstrapping (3 layers of feedback). I'm not sure what I'll do with these experiments, however, at least 1 of these will end up on the module page as the update to the popcorn DC receiver mainframe.

Glenn, VE7DNL built a landmark common emitter preamp with parallel 2N4401s, shunt feedback and some 1% bias resistors. I built a version of his pre-amp without the diplexer. The input return loss in my version was only 5 dB — likely the diplexer network helps establish the input match. It appears Glenn's goal was low NF and the best match and noise figure don't always correlate. I matched Q2 and Q3 as well as I could since only 6 were available. An Hfe of 220- 240 is not uncommon for the 2N4401 — a good part.

Here's a SPICE plot of the amp performed by Wes, W7ZOI. The two plots are S21 and S11. S21 is the forward gain in dB. S11 is the input reflection coefficient, which is just the negative of the return loss. At 1 kHz, the calculated gain = just under 42 dB and the input match = ~17 dB return loss.

To improve return loss in my breadboard, I need to tweek Q2 and Q3 bias (lower the 6.04K resistor), plus add some emitter degeneration. Emitter degeneration (series feedback) would worsen the noise figure. The 100 uF emitter caps offer some series feedback and capacitor values like 1000 uF are needed if "true" AF bypass is wanted.

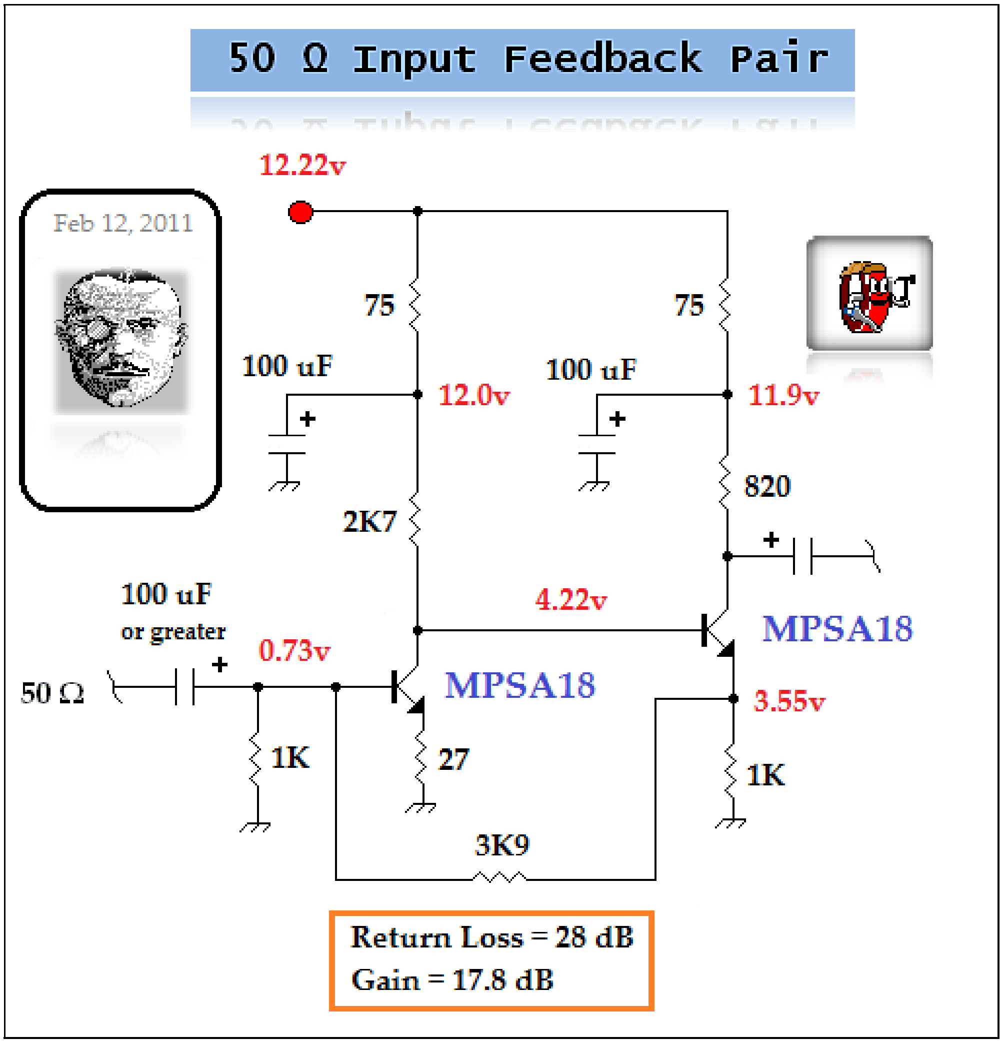

Above — Alternate experimental 50 ohm audio preamp to follow a diode ring product detector. The high impedance output enables small value output coupling capacitors for pseudo high-pass filtering. I simply added a collector resistor plus a decoupling network to Q2. The low noise MPSA18 BJT is a favorite and predictable part. The 2N3904, 2N4401 and others should work fine.

Don't even think about using an input capacitor less than 100 uF at 50 ohms — you'll trash the input Return Loss. I measured the highest Return

Loss with a 470 uF coupling capacitor, but the 100 uF cap shown worked almost as well.

I'm currently decreasing the (Q2) 1K emitter resistor with hopes of squeezing out a little more gain while trying to preserve strong signal handling capacity. I've found that heavy feedback is your friend.

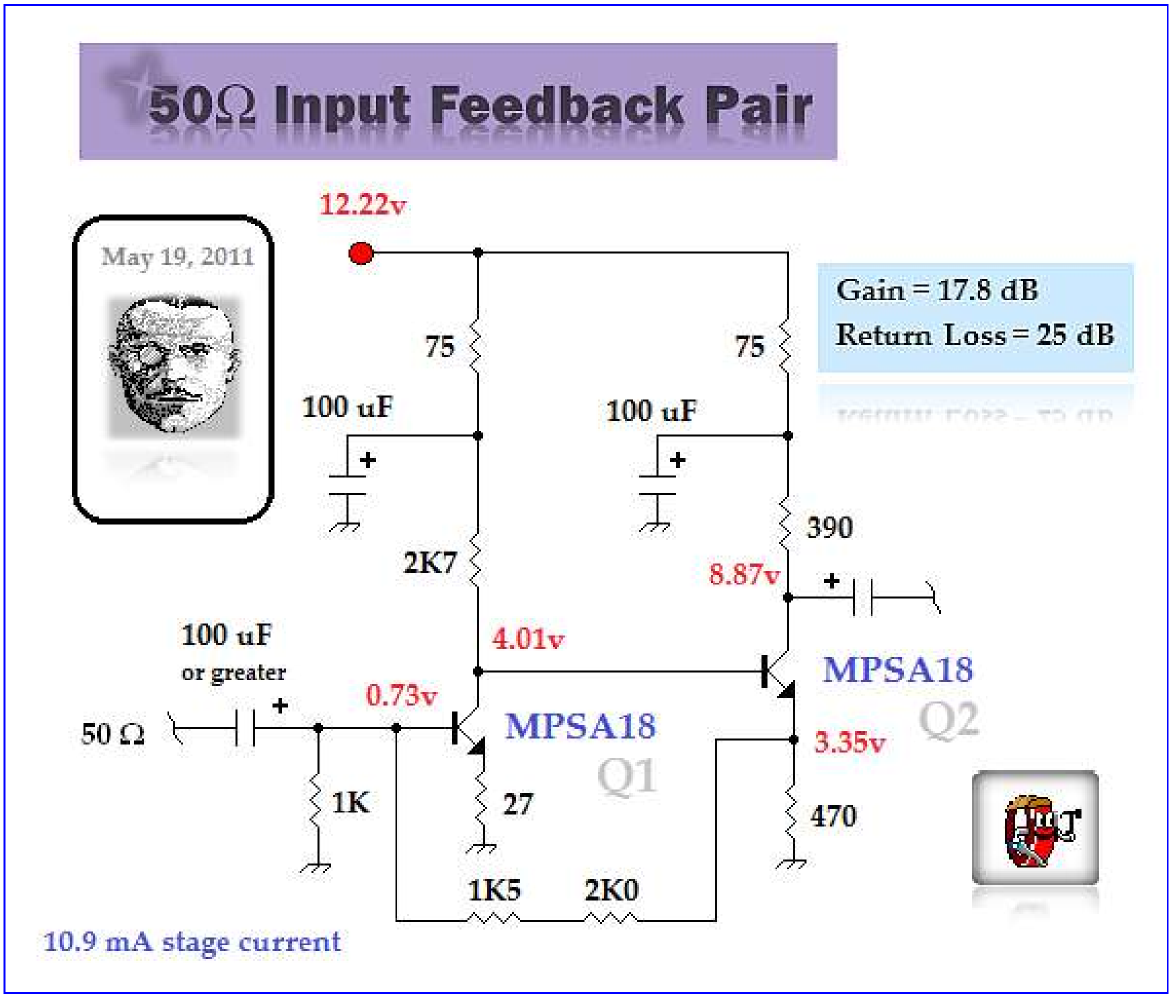
50Ω Input Feedback Pair
May 19, 2011
12.22v
Gain = 17.8 dB
Return Loss = 25 dB
75
75
100 uF
100 uF
390
2K7
8.87v
4.01v
100 uF
or greater
MPSA18
Q2
0.73v
50 Ω
MPSA18
Q1
3.35v
1K
27
470
1K5
2K0
10.9 mA stage current

Above — Another version; trying to increase stage gain. The Q2 collector resistor dramatically affects clean signal handling — lowest distortion occurred with a 367 ohm collector resistance. I substituted a 390 ohm collector R. This dropped the gain a little, but increasing distortion over gain is never my goal.

During experiments, dropping the the Q2 emitter resistor below 470 ohms made the other biasing resistors more critical, decreased return loss and invited distortion. The best return loss occurred when critically biasing with 1% resistors — I'm avoiding that!

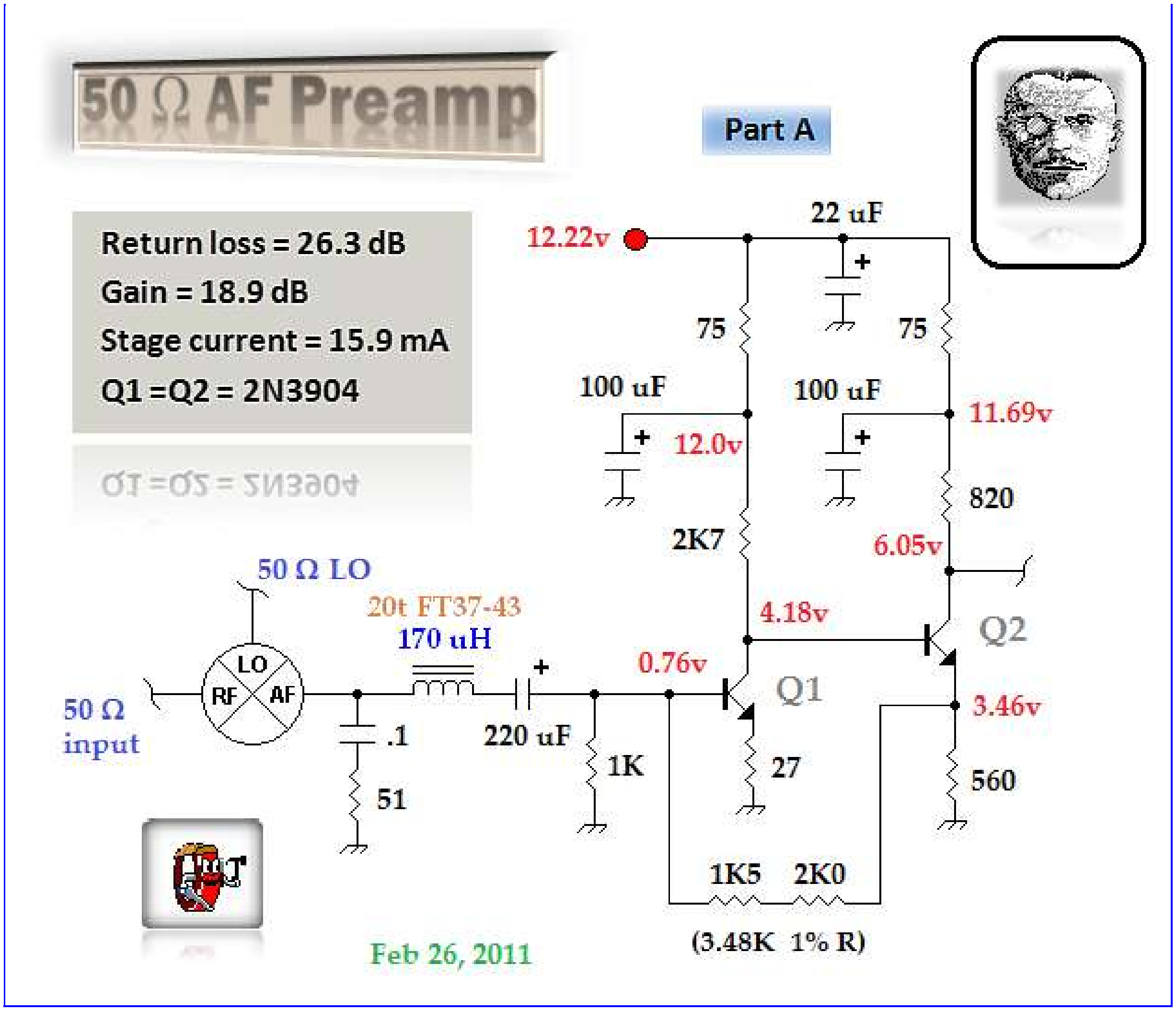
50 Ω AF Preamp
Part A
Return loss = 26.3 dB
Gain = 18.9 dB
Stage current = 15.9 mA
Q1 =Q2 = 2N3904
12.22v
22 uF
75
75
100 uF
100 uF
12.0v
11.69v
820
2K7
6.05v
50 Ω LO
20t FT37-43
170 uH
4.18v
Q2
0.76v
Q1
LO
RF
AF
50 Ω
input
.1
220 uF
3.46v
1K
27
560
51
1K5
2K0
(3.48K 1% R)
Feb 26, 2011

Above — Another popcorn part experiment from February. A low-pass network terminates a diode ring mixer and drives a 50 ohm input feedback pair arranged so the Q2 collector DC voltage is approximately 1/2 VCC — this provides the DC bias for a 5532 op-amp low-pass filter arranged in the Sallen-Key topology.

The Q2 collector voltage removes the need for a separate single supply op-amp bias network and an AC coupling cap. I didn't try this circuit in a receiver; arranging the feedback network and testing the possibility of such a design drove this experiment. A perfect 1/2 VCC Q2 collector voltage might be obtained with a standard 1% value 806 ohm R. I lack this part.

Almost Popcorn Superhet

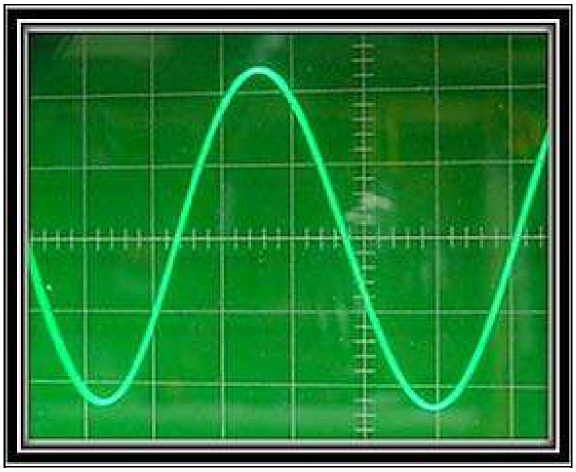

1. Receiver Concept and Macro Diagram

Not done.

Maybe a flight-of-fancy that never crosses the finish line? This page seems to repeatedly find the back burner; a good intention that never sees completion. Sadly this is normal for some of us builders — how many incomplete projects adorn your home and workshop? Still, this web page might fuel your *own* experiments.

For years, I've attempted to update my Popcorn Superhet from 1998 but always seem to lose interest from numerous distractions such as VHF and UHF experiments and other non-radio stuff. So here it is so far — raw and unrefined: more experiments from my analog-centric workbench. To clarify — I love the new digital radio stuff, but that's well covered on the net by numerous code writers and those bleeding edge, frontier-pioneer types. I prefer a 50 Ω lab with all the visceral enjoyment and learning it has to offer. To each his own.

This web page remains unlisted on the main top level menu — and for good reason.

Click for some supplemental lab notes.

2. N = 4, Gaussian-to-6 dB Crystal Ladder Filter

I frequency matched four 11.0592 crystals, characterized them and took an average of those crystal parameters plus some data from Zverev and inputted this into xlad08.exe, a program that ships with EMRFD.

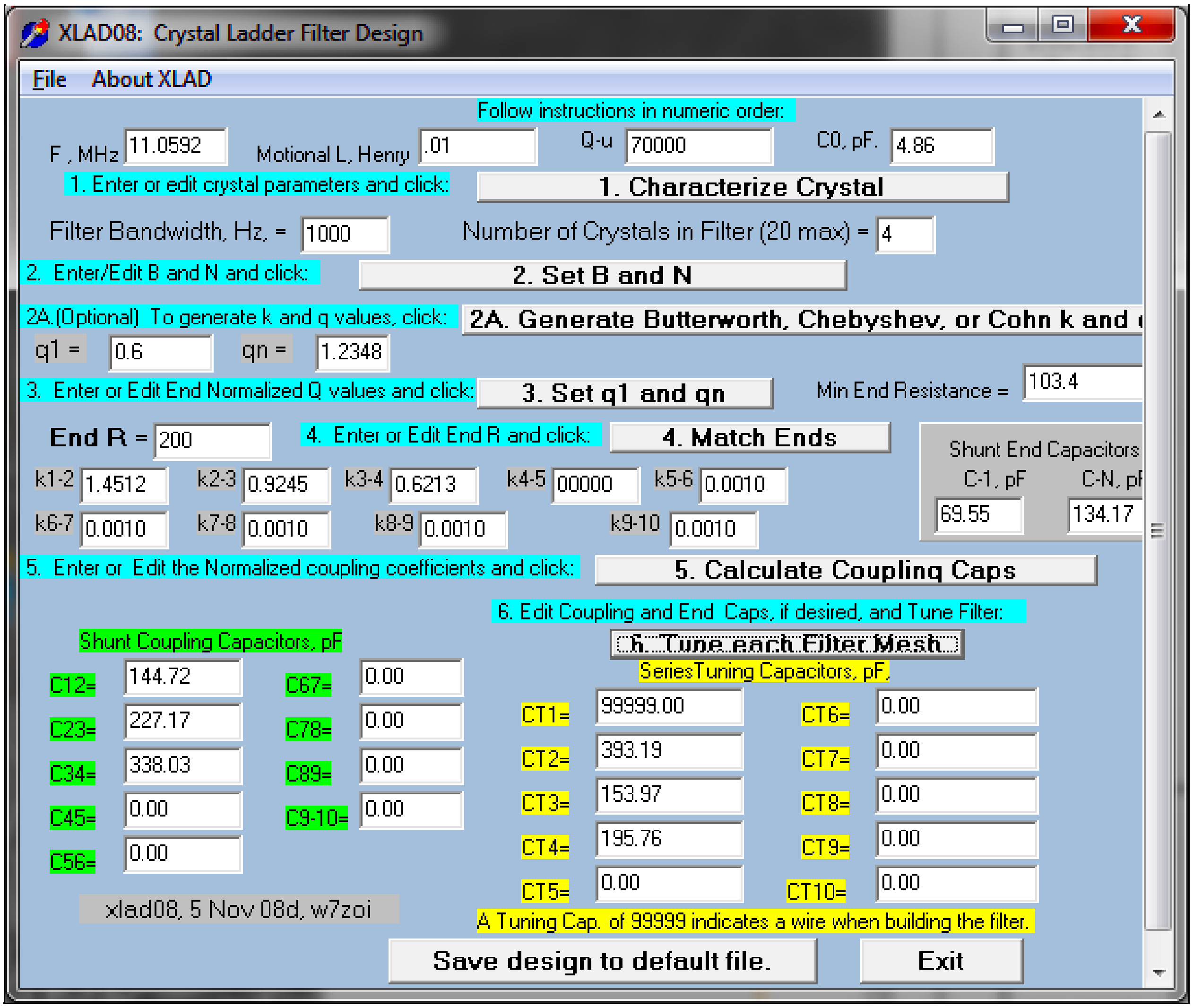
XLAD08: Crystal Ladder Filter Design
File
About XLAD
1. Characterize Crystal
2. Set B and N
2A. Generate Butterworth, Chebyshev, or Cohn k and
3. Set q1 and qn
4. Match Ends
5. Calculate Coupling Caps
6. Tune each Filter Mesh
Save design to default file.
Exit

Follow instructions in numeric order:

F , MHz 11.0592
Motional L, Henry .01
Q-u 70000
C0, pF. 4.86

1. Enter or edit crystal parameters and click:

Filter Bandwidth, Hz, = 1000
Number of Crystals in Filter (20 max) = 4

2. Enter/Edit B and N and click:

2A.(Optional) To generate k and q values, click:

q1 = 0.6
qn = 1.2348

3. Enter or Edit End Normalized Q values and click:

Min End Resistance = 103.4

End R = 200

4. Enter or Edit End R and click:

Shunt End Capacitors
C-1, pF 69.55
C-N, pF 134.17

k1-2 1.4512
k2-3 0.9245
k3-4 0.6213
k4-5 00000
k5-6 0.0010
k6-7 0.0010
k7-8 0.0010
k8-9 0.0010
k9-10 0.0010

5. Enter or Edit the Normalized coupling coefficients and click:

6. Edit Coupling and End Caps, if desired, and Tune Filter:

Shunt Coupling Capacitors, pF

C12= 144.72
C23= 227.17
C34= 338.03
C45= 0.00
C56= 0.00
C67= 0.00
C78= 0.00
C89= 0.00
C9-10= 0.00

SeriesTuning Capacitors, pF

CT1= 99999.00
CT2= 393.19
CT3= 153.97
CT4= 195.76
CT5= 0.00
CT6= 0.00
CT7= 0.00
CT8= 0.00
CT9= 0.00
CT10= 0.00

xlad08, 5 Nov 08d, w7zoi

A Tuning Cap. of 99999 indicates a wire when building the filter.

Above — Xlad08.exe screen shot.

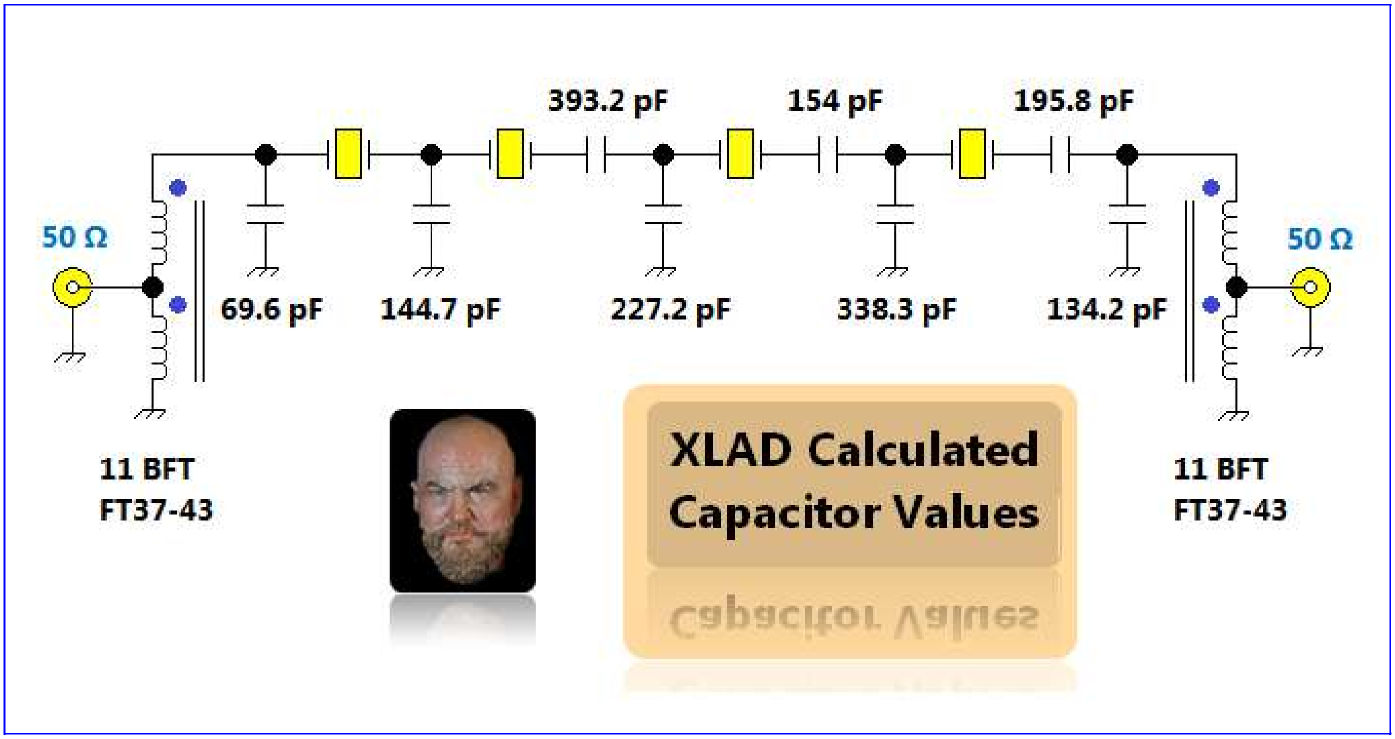
393.2 pF
154 pF
195.8 pF
50 Ω
50 Ω
69.6 pF
144.7 pF
227.2 pF
338.3 pF
134.2 pF
11 BFT
FT37-43
11 BFT
FT37-43
XLAD Calculated
Capacitor Values

Above — The xlad filter expressed as a schematic. Click for a GPLA plot of the filter.

I built and hand tuned this crystal filter with my analog sweep system. Click for a photo of the optimized filter breadboard — I used 1 or 2 parallel capacitors to tune each capacitance to derive the best possible skirt shape. Click for a sweep of this optimized filter in my 'scope. 2 bifilar transmission line transformers provided the required 200 Ω port termination in my 50 Ω sweep system.

I sought a crystal filter based on the xlad08 calculated values using only 1 nearest standard value in each capacitor slot.

Transitional filters like the G-to-6 dB, are generally tolerant of part substitutions and thus I sought a filter with only 1 nearest 5% standard-value capacitor in *every* slot instead of 2 sticking 2 close in value parallel caps to make up the needed capacitance as needed.

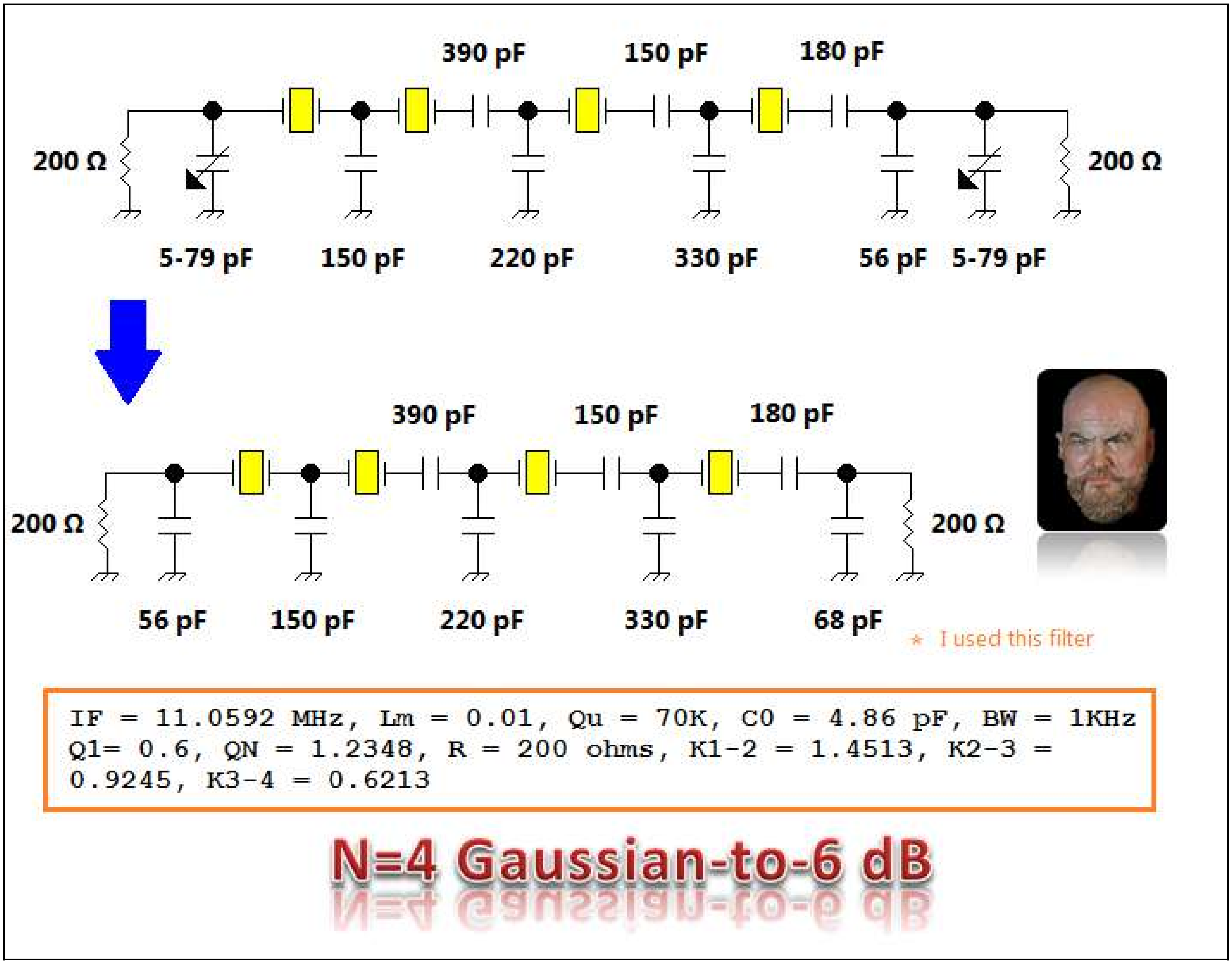

IF = 11.0592 MHz, Lm = 0.01, Qu = 70K, C0 = 4.86 pF, BW = 1KHz Q1= 0.6, QN = 1.2348, R = 200 ohms, K1-2 = 1.4513, K2-3 = 0.9245, K3-4 = 0.6213

N=4 Gaussian-to-6 dB

Above —The 2-step transition to a 1 capacitor, nearest standard value in each capacitor slot. In the top diagram, I placed a trimmer cap on each end to allow filter tweaking while looking at the swept filter in my 'scope. I've learned that adjusting the parallel cap nearest to each port a simple way to tweak a G-to-6 dB filter. After setting the best looking skirt, I removed and measured the end parallel capacitance and substituted the nearest standard fixed capacitor. Manipulating the capacitors values will change filter bandwidth and return loss.

You might compare the lower filter to the xlad calculated values to see how tolerant the circuit is.

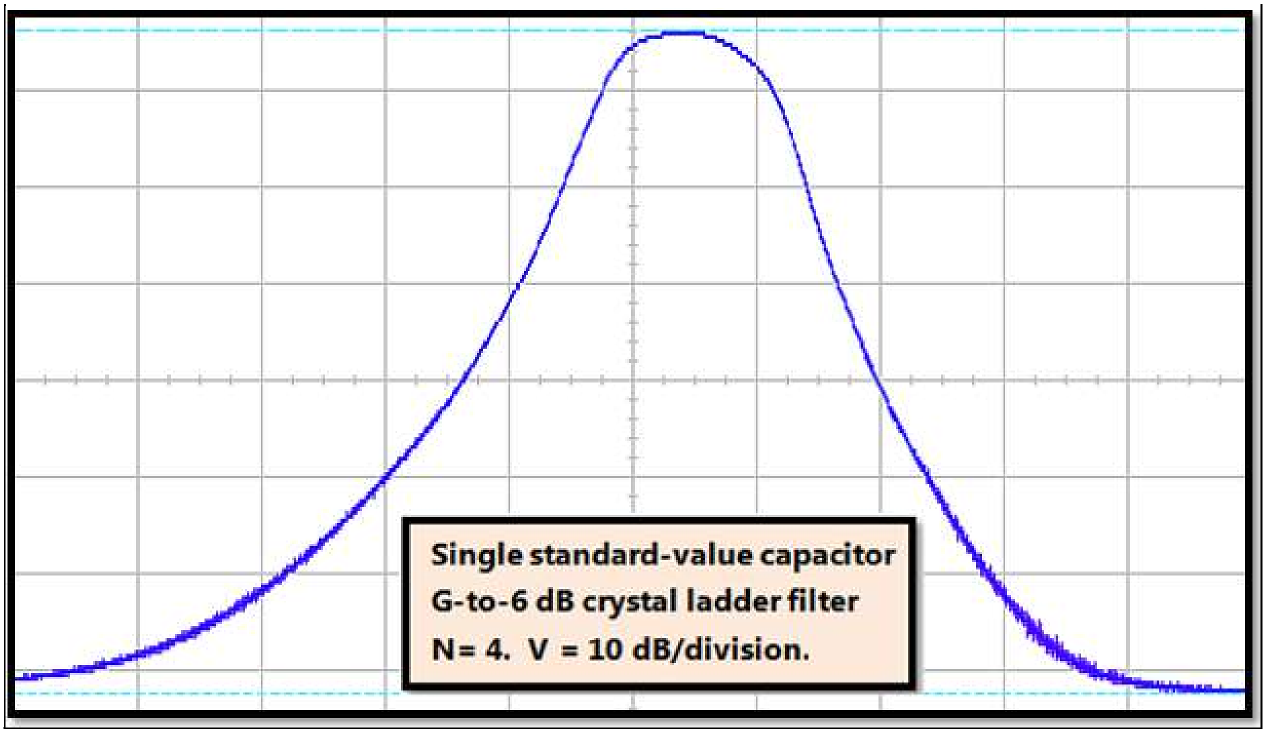

Above — A sweep of the single value capacitor version that's used in the Almost Popcorn receiver. While the 1 KHz filter skirt isn't perfect, it's reasonable. Click for the test bench breadboard. This particular filter measured 1/2 power bandwidth = 1227 KHz and the insertion loss was ~3.6 dB.

---

3. IF Amplifier

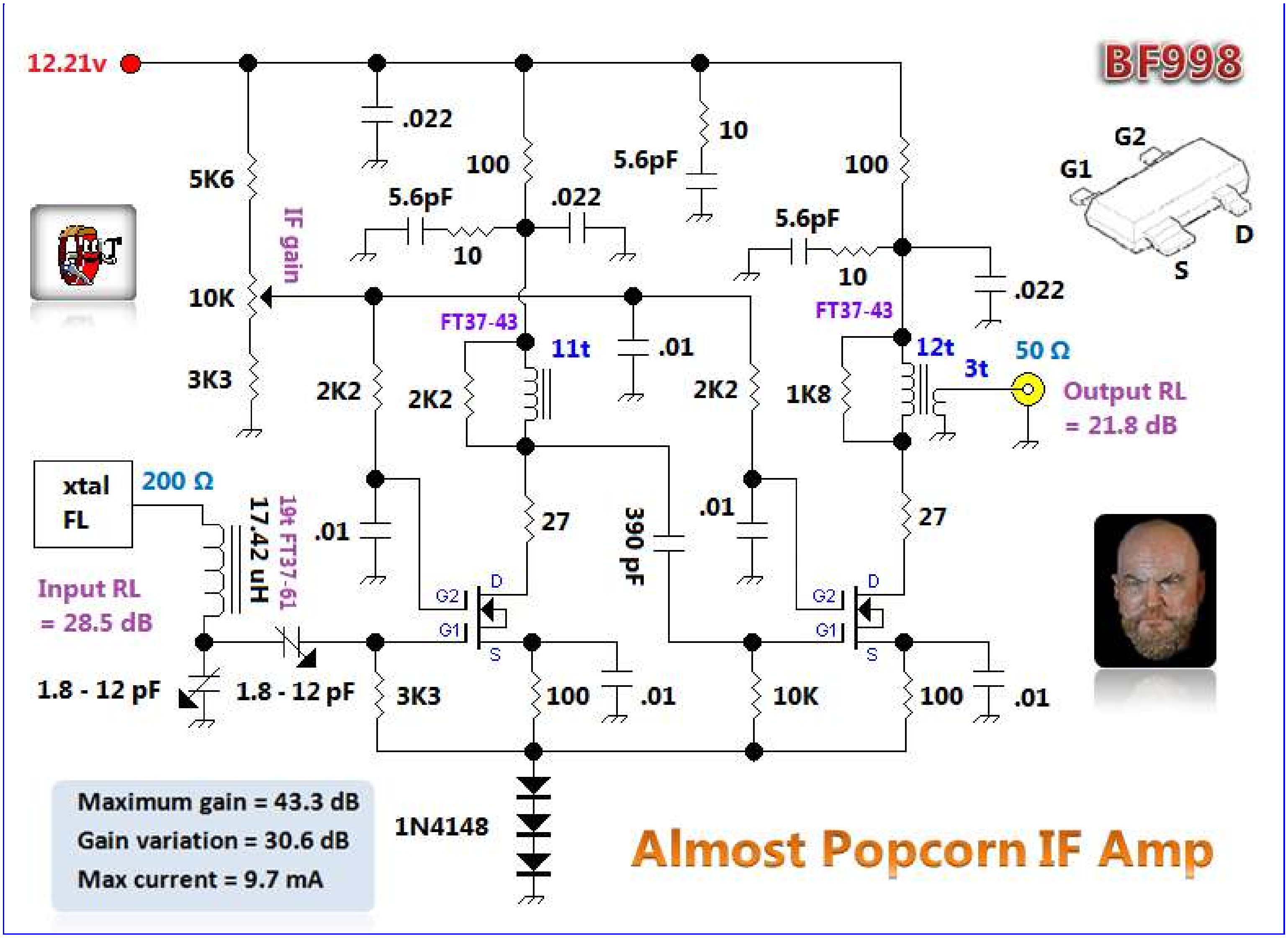
12.21v
BF998
.022
10
5.6pF
100
100
5K6
5.6pF
.022
G2
G1
D
S
IF gain
10
5.6pF
10
10K
.022
FT37-43
FT37-43
11t
.01
12t
3t
50 Ω
3K3
2K2
2K2
2K2
1K8
Output RL
= 21.8 dB
xtal
FL
200 Ω
17.42 uH
19t FT37-61
.01
27
390 pF
.01
27
Input RL
= 28.5 dB
G2
D
G1
S
G2
D
G1
S
1.8 - 12 pF
1.8 - 12 pF
3K3
100
.01
10K
100
.01
Maximum gain = 43.3 dB
Gain variation = 30.6 dB
Max current = 9.7 mA
1N4148
Almost Popcorn IF Amp

Above — Version 2 of a popcorn IF amplifer idea with BF998s. IF = 11.0572 MHz.

I have no use for AGC in my personal receivers since I listen through a speaker and *always* ride the RF gain control with my finger. Since the BF998 has gain into UHF, wideband decoupling + bypass SMT R-C networks filter each MOSFET in case of oscillations. Click and click for photos. This circuit will operate fine with just 0.022 µF bypass caps.

A network matches the 200 Ω xtal filter output impedance to the 3K3 input Z of the IF amp cascade. The L-C-C network has become a favorite matching weapon — 2 tweaks is often better than the 1 offered by a garden variety L-network. The product detector that follows this stage is a Level 7 diode ring mixer soldered onto the audio preamp board. This receiver will feature some modish circuits aimed at low noise + high fidelity but still stay "almost popcorn". Anyhow, that's my hope.

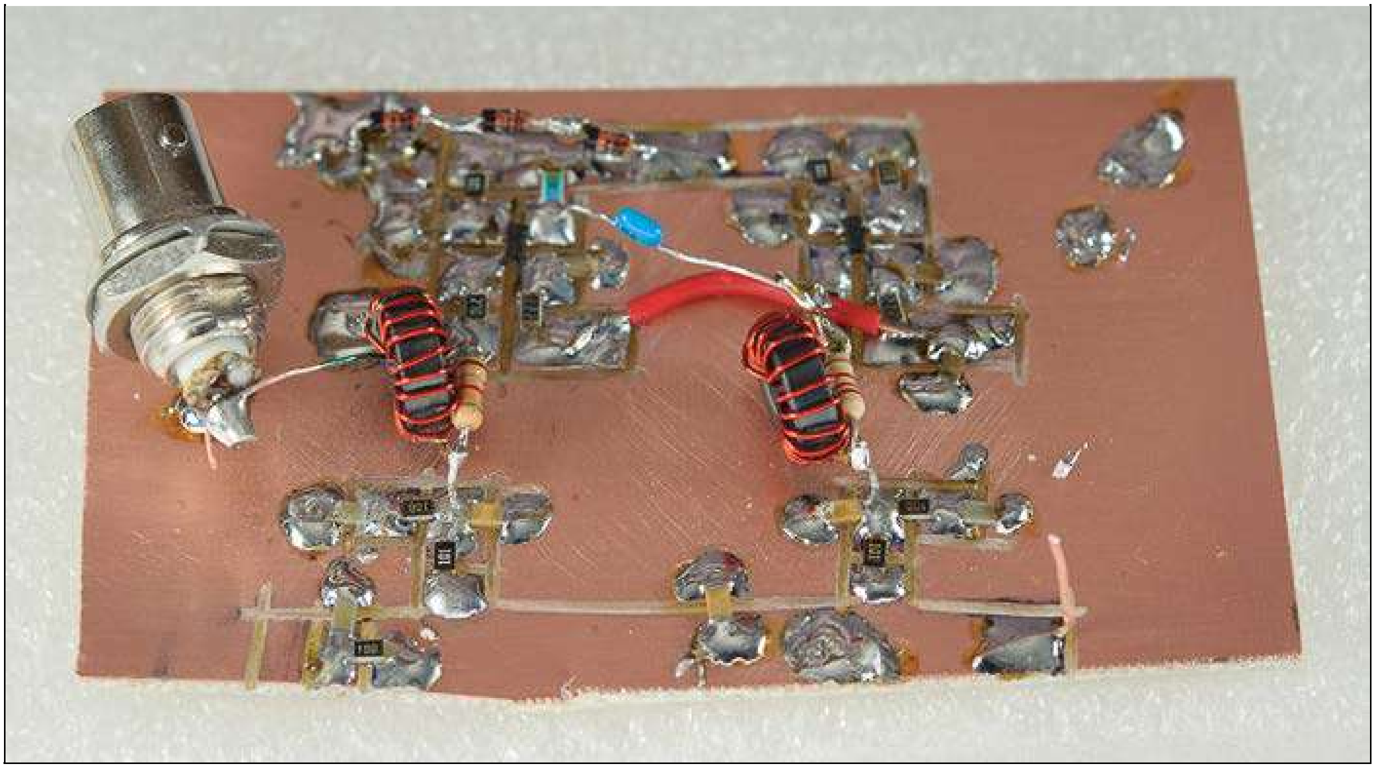

Above — A breadboard of the Popcorn IF amp sans the input network. I tested this stage using an oscillator with a 3K3 output impedance and liked what I saw. Click for a reverse angle view. After this build, I ordered some SMT versions of the 1N4148 diode. The BNC connector was only soldered on the board for testing purposes.

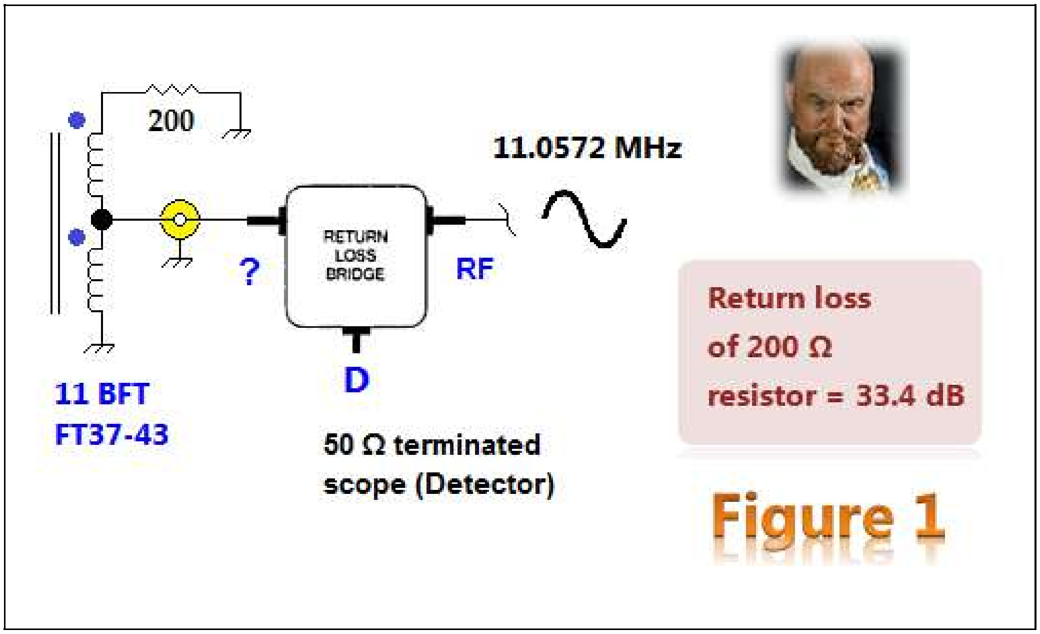
200
11.0572 MHz
RETURN LOSS BRIDGE
?
RF
D
11 BFT
FT37-43
50 Ω terminated scope (Detector)
Return loss of 200 Ω resistor = 33.4 dB
Figure 1

Above — Setting up to measure the return loss of the 200 Ω input Z of my popcorn IF amp. 1 difference between scratch homebrew and kit building is all the measurements you need to make to optimize your scratch receiver. A joie de vivre stems from our descriptive/exploratory analysis before and after making our breadboard. We just learn from mistakes and move forward to improve both our knowledge and gear.

I chose a 200 Ω input impedance for a reason — it's easy to apply a standard 50 Ω return loss bridge via a 4:1 transmission line transformer to establish the correct input impedance for the single frequency matching network.

From the set-up shown above, I found that with a 200 Ω resistor, my return loss at 11.057 MHz = 33.4 dB. This signals the very best we can achieve with our network: directivity. Click for a photo of the above experiment breadboarded on a scap of copper clad board.

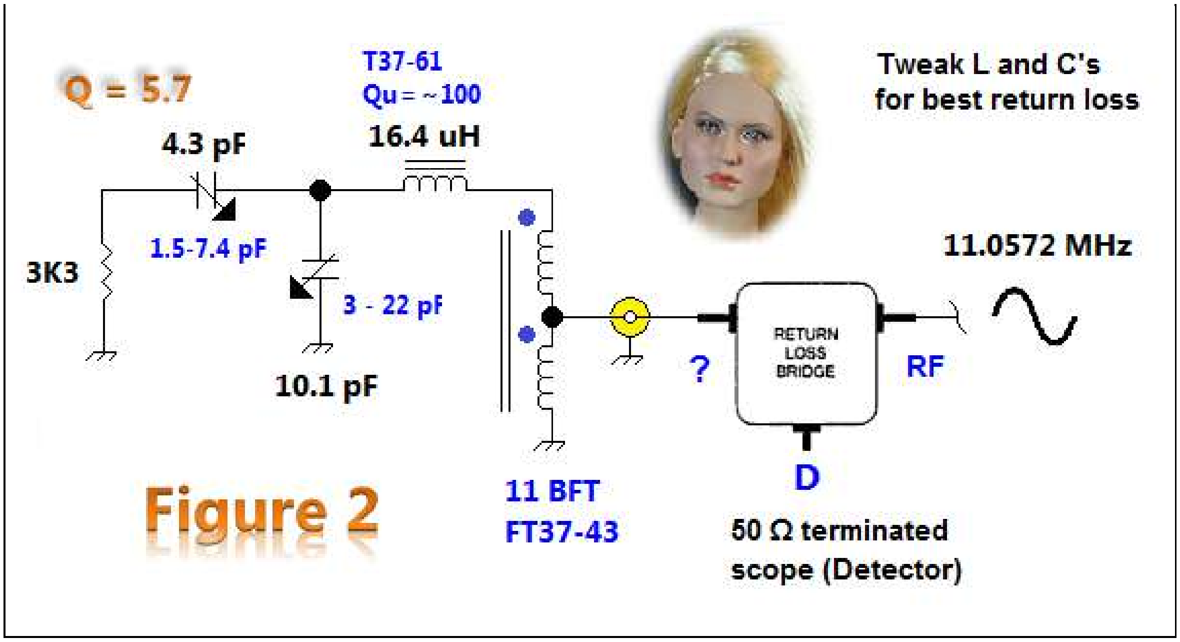
Q = 5.7
T37-61
Qu = ~100
16.4 uH
4.3 pF
1.5-7.4 pF
3K3
3 - 22 pF
10.1 pF
Tweak L and C's
for best return loss
11.0572 MHz
RETURN
LOSS
BRIDGE
?
RF
D
Figure 2
11 BFT
FT37-43
50 Ω terminated
scope (Detector)

Above — The schematic of my 200 to 3K3 Ω matching network. I designed it with zmat08.exe, a program by Wes that I first used as a console app since the days of DOS. *Zmat08 program comes with EMRFD.* Click for a screen shot of from Zmat08.
The calculated values from Zmat08 only consider a pure resistance and not a complex impedance, so on-bench adjustments of the design L and C values are normally required and fun to do. Variable trimmer caps make tweaking the C easy.

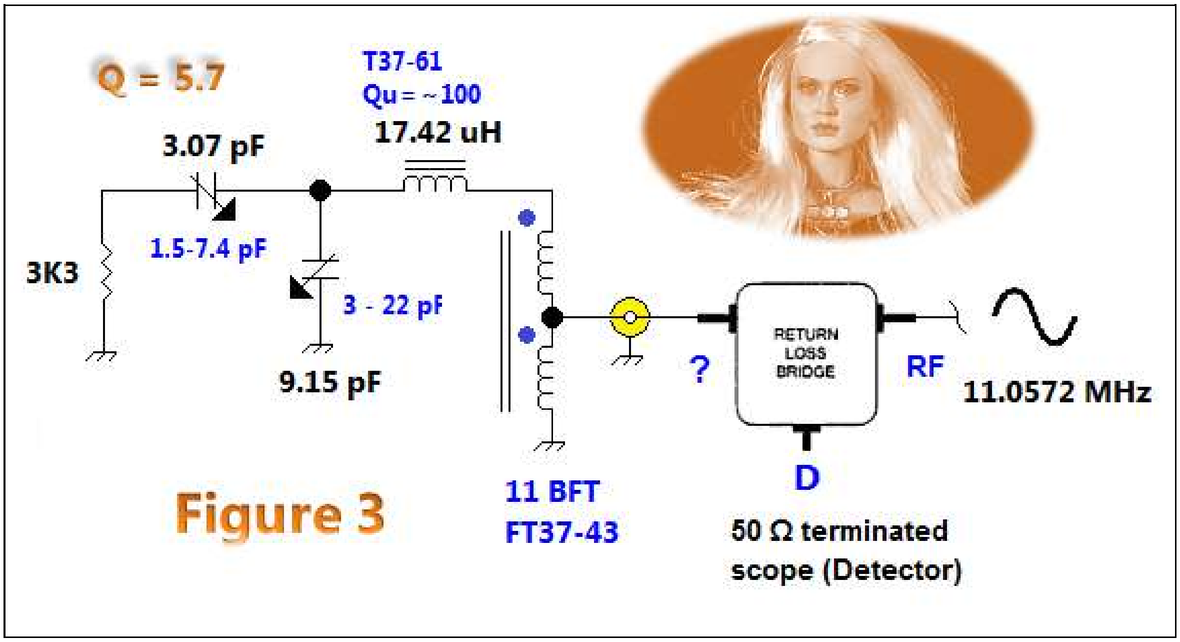
Q = 5.7
T37-61
Qu = ~100
17.42 uH
3.07 pF
1.5-7.4 pF
3K3
3 - 22 pF
9.15 pF
RETURN
LOSS
BRIDGE
?
RF
11.0572 MHz
D
Figure 3
11 BFT
FT37-43
50 Ω terminated
scope (Detector)

Above — The final— *measured* schematic — While watching the peak to peak voltage on my 50 Ω terminated scope, L and both capacitors were tweaked until I found the highest return loss. L is tweaked by scrunching or expanding turns, or adding/removing turns as needed: in this case, I added a turn and expanded the windings slightly. With the L-C-C match, a return loss greater than 25 dB may arise plying good bench practices — mine measured a lovely 29.2 dB. I removed and measured the inductor plus 2 capacitors to get the values shown **black** in Figure 3.

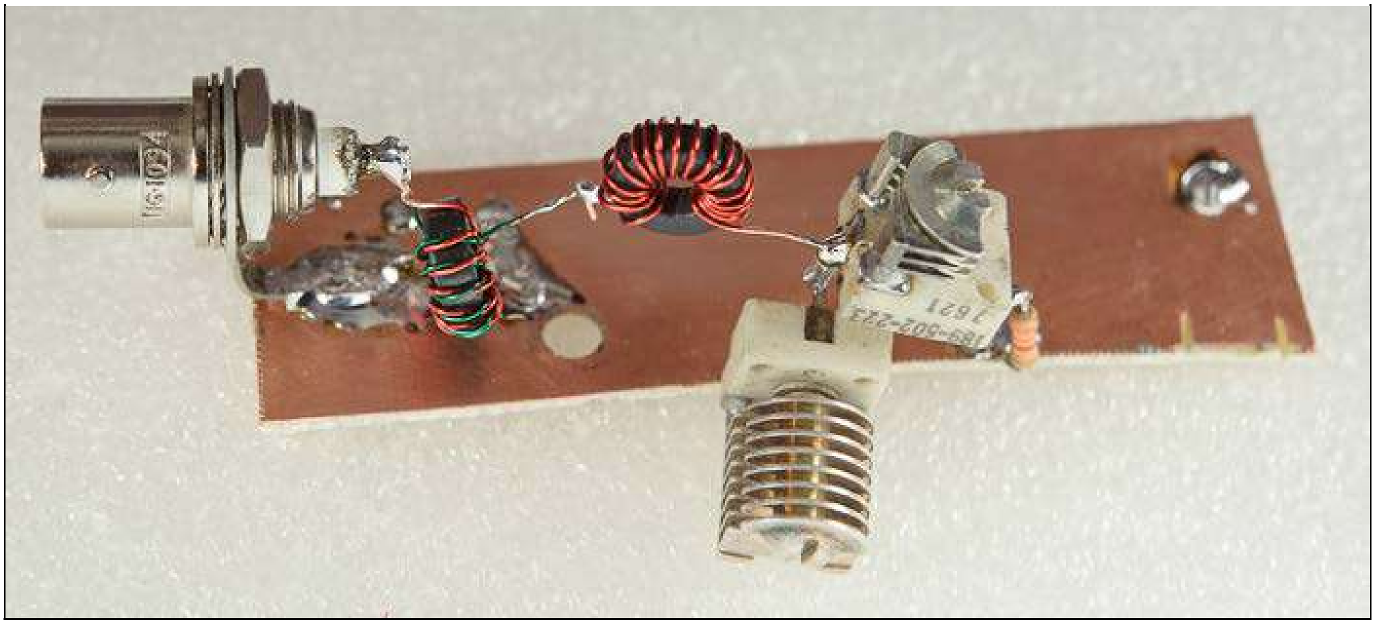

Above — A breadboard of the Figure 3 network before I experimented to find the best L (and shortened its leads).

After I making and tweaking the Figure 3 network, I took the coil along with 2 cheaper trimmer caps and added them to my IF amplifer breadboard. While measuring the return loss of the IF strip input, I had to expand a couple of links on the inductor, but after tweaking the L and the 2 trimmer caps, I measured a 28.5 dB return loss. This ensures my crystal filter will see close to a 200 Ω termination at the IF input. I love well matched stages!

Further, I made a L network and then a L-C-C network to match the output port, however, UHF oscillations erupted. I then went back to a simple broadband transformer and measured a RL of 21.8 dB with a 12:3 turns ratio. Done.

---

4. Mixer and Post-Mixer Amplifer

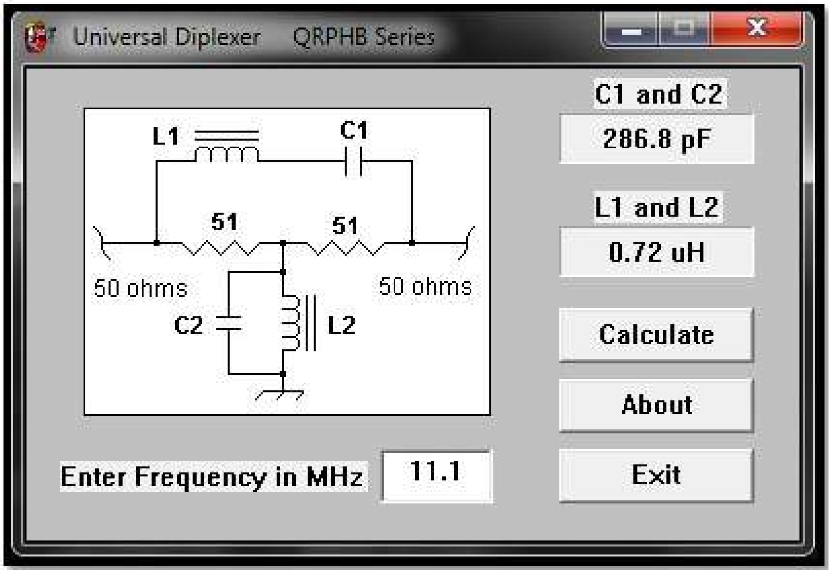

Universal Diplexer QRPHB Series

C1 and C2: 286.8 pF

L1 and L2: 0.72 uH

Enter Frequency in MHz: 11.1

Above — I tested the W1JR (Joe Reisert) bridged-tee diplexer against another design used by Ten-Tek. I'll show the outcomes on the Almost Popcorn receiver page some day, but the W1JR better fullfills its purpose; termination from DC to daylight at the mixer IF port. I calculated the L and C values for a 11.0592 MHz IF with my Universal Diplexer application from many moons ago.

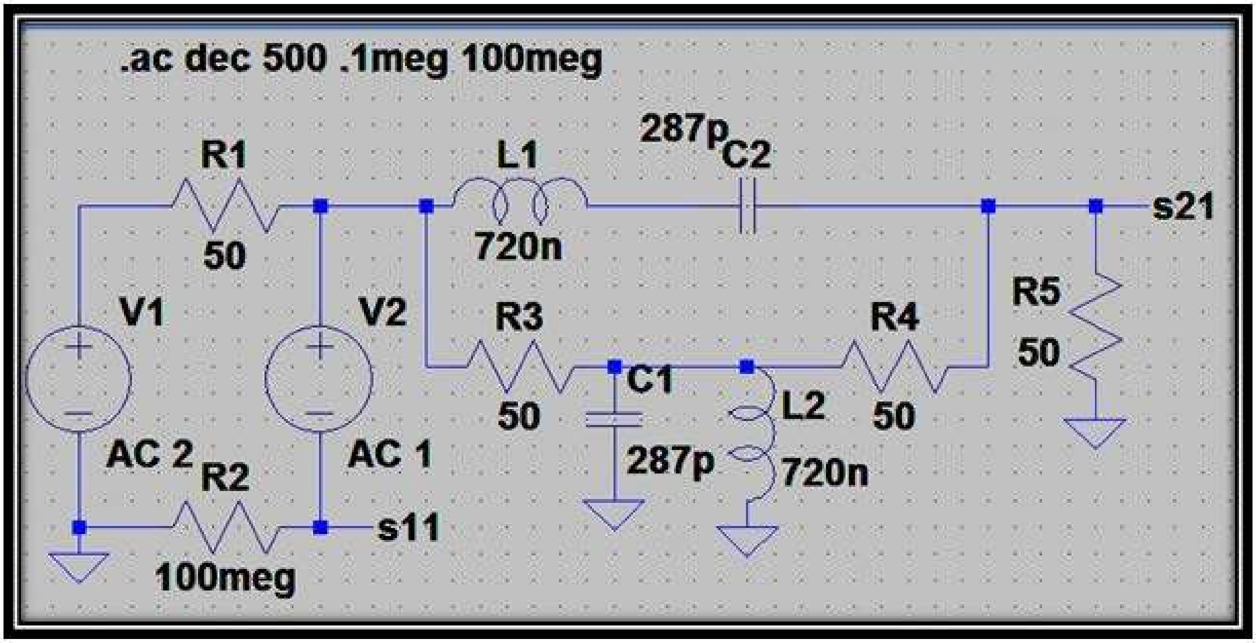
.ac dec 500 .1meg 100meg
R1
50
L1
720n
287p
C2
s21
V1
V2
R3
R4
R5
50
50
50
C1
L2
AC 2
R2
AC 1
287p
720n
s11
100meg

Above — SPICE provides a great way to assess a diplexer since you can plot both S21 and S11 against frequency.

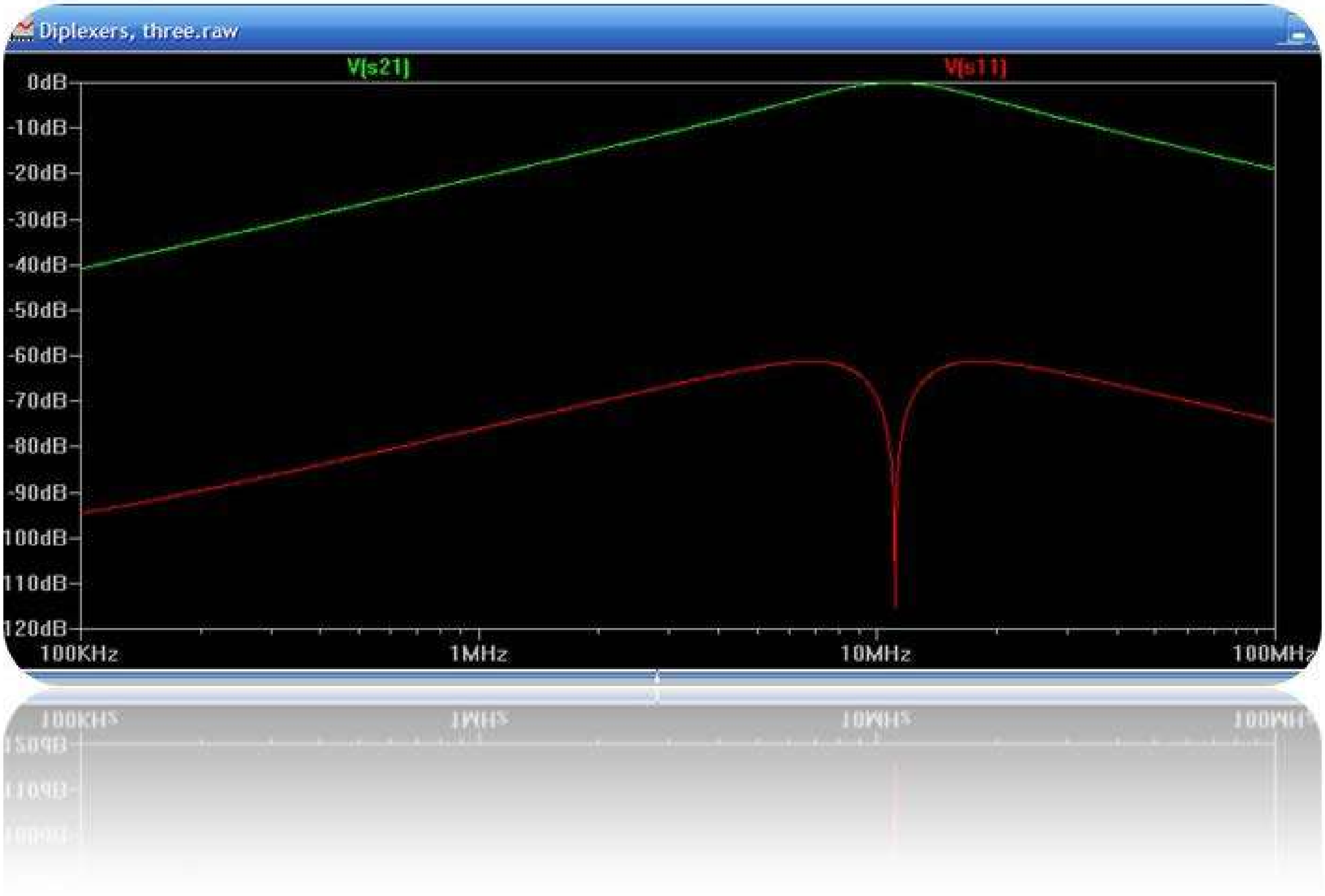

Above — A fantastic match! Thanks Joe for the design and to **Wes, W7ZOI** for his help with diplexers and the SPICE plots through many years of related experiments. Since web publishing the diplexer page, at least 500 diplexer-related emails have come in. Human translated versions of the web page exist in at least 3 languages.

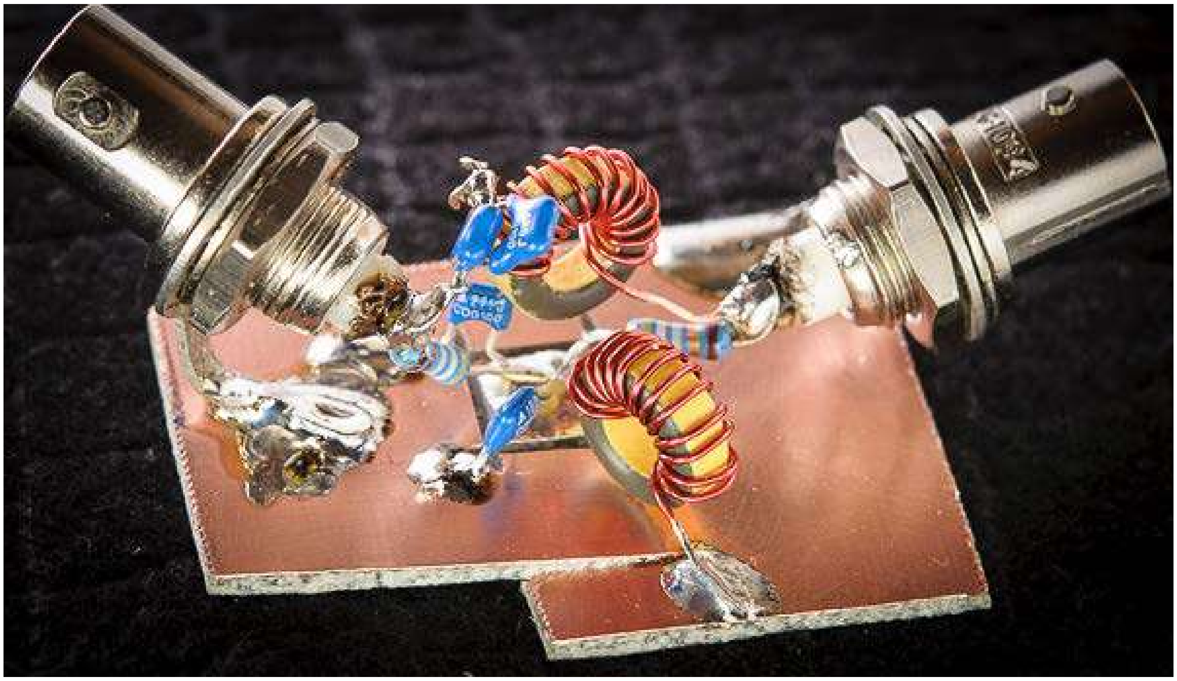

Above — The evaluation breadboard of my W1JR diplexer. Those are T30-6 toroids but they look huge in this photo.

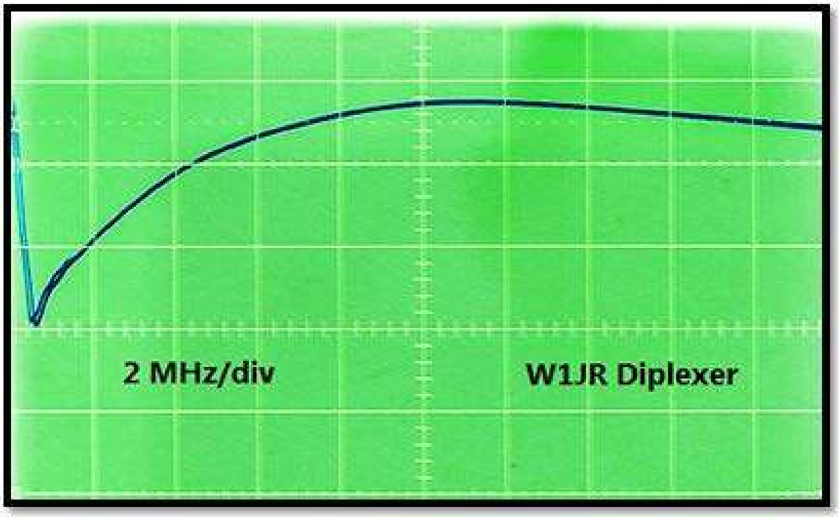

Above — Tracking generator plus spectrum analysis sweep. SPICE does a better of examining the high-pass side.

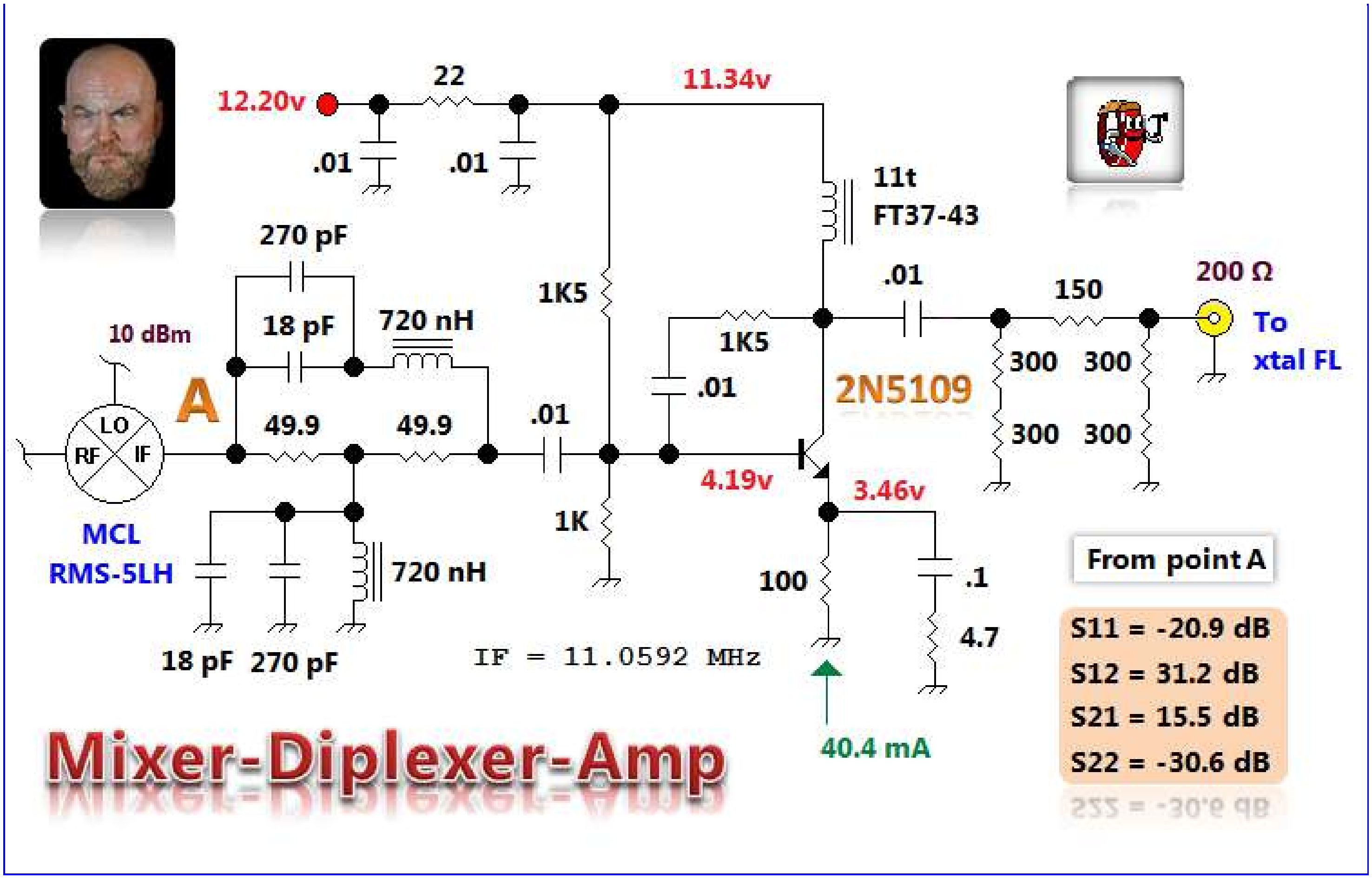
12.20v
22
11.34v
.01
.01
11t
FT37-43
270 pF
18 pF
720 nH
1K5
.01
150
200 Ω
To
xtal FL
1K5
10 dBm
A
.01
2N5109
300
300
LO
RF
IF
49.9
49.9
.01
300
300
4.19v
3.46v
MCL
RMS-5LH
1K
720 nH
100
.1
From point A
S11 = -20.9 dB
S12 = 31.2 dB
S21 = 15.5 dB
S22 = -30.6 dB
4.7
18 pF
270 pF
IF = 11.0592 MHz
40.4 mA
Mixer-Diplexer-Amp

Above — A look at the diode ring mixer, diplexer and post-mixer amp from the latest version of my Almost Popcorn Receiver. I tested the diplexer and RF amp with pad as one unit (see the S-values).

It did not make sense to use a transmission line transformer to get 50 Ω on the amp collector and then use another transformer to build it back up to the needed 200 Ω impedance for the crystal filter, so I just employed a single choke driving a 200 Ω Z attenuator pad.
S parameters are 50 Ω values, thus S22 = the 200 Ω transformed into 50 Ω. After terminating the stages 200 Ω output with a 4:1 Z transmission line transformer, I tested it like any other 50 Ω circuit andmeasured S22 as -30.6 dB — my xtal filter input Z is well established.

Even with the 6 dB *(200 Ω)* attenuator pad and losses from the diplexer, the stage gain from point A = 15.5 dB. Click for to view a bigger picture of the breadboard. Losses from a diode ring mixer are typically 5 - 7 dB.

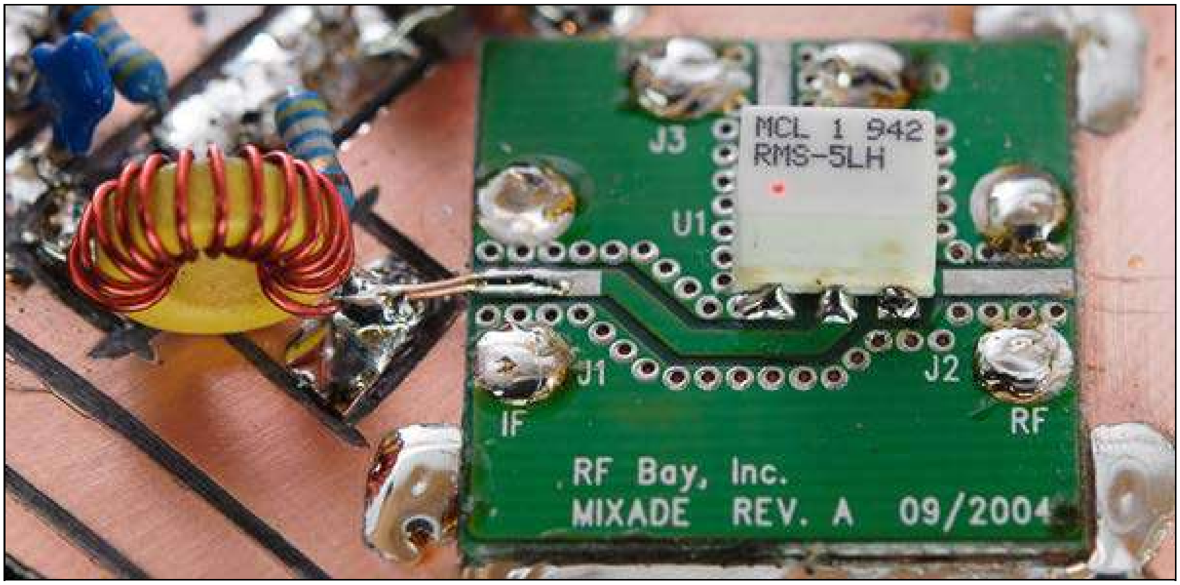
MCL 1 942
RMS-5LH
J3
U1
J1
J2
IF
RF
RF Bay, Inc.
MIXADE REV. A 09/2004

Above — Mixer and diplexer assembly. The RMS-5LH takes 10 dBm LO drive.

5. Product Detector and Post Product Detector AF Amplifer

I kept the diode ring for a product detector: an SBL-1 driven with a xtal oscillator (BFO).

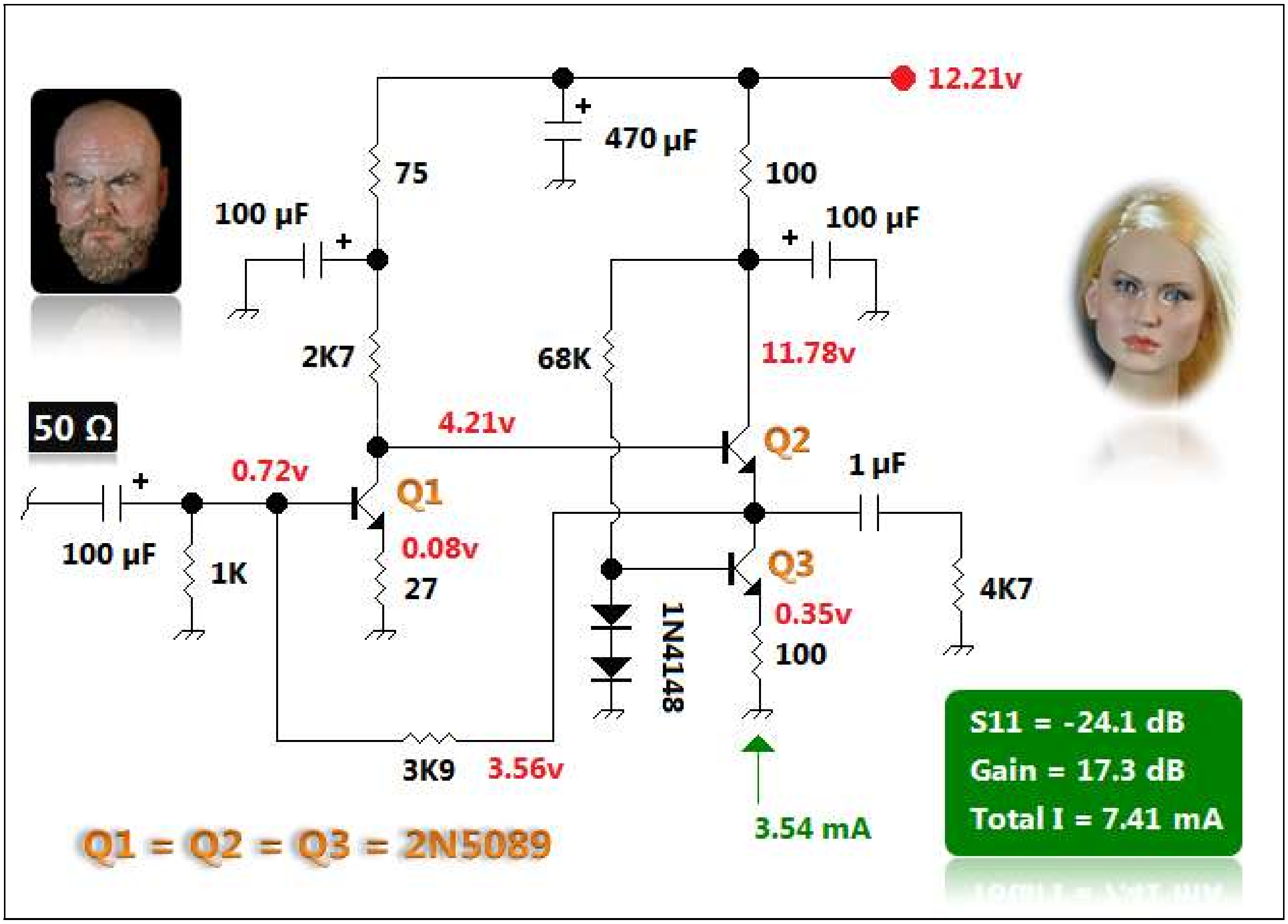
12.21v
470 µF
75
100
100 µF
100 µF
2K7
68K
11.78v
50 Ω
4.21v
Q2
0.72v
Q1
1 µF
100 µF
1K
0.08v
27
Q3
4K7
0.35v
100
1N4148
3K9
3.56v
3.54 mA
S11 = -24.1 dB
Gain = 17.3 dB
Total I = 7.41 mA
Q1 = Q2 = Q3 = 2N5089

Above — My post-product detector amplifer: a common emitter/ emitter follower pair arranged for a 50 Ω input impedance with shunt and series feedback. Getting the biggest, clean sine wave in and out posed my only goal — the 3.54 mA current source on the follower helps that cause. The input Z tested quite insensitive to output load variations and the (S11) return loss = 24.1 dB.

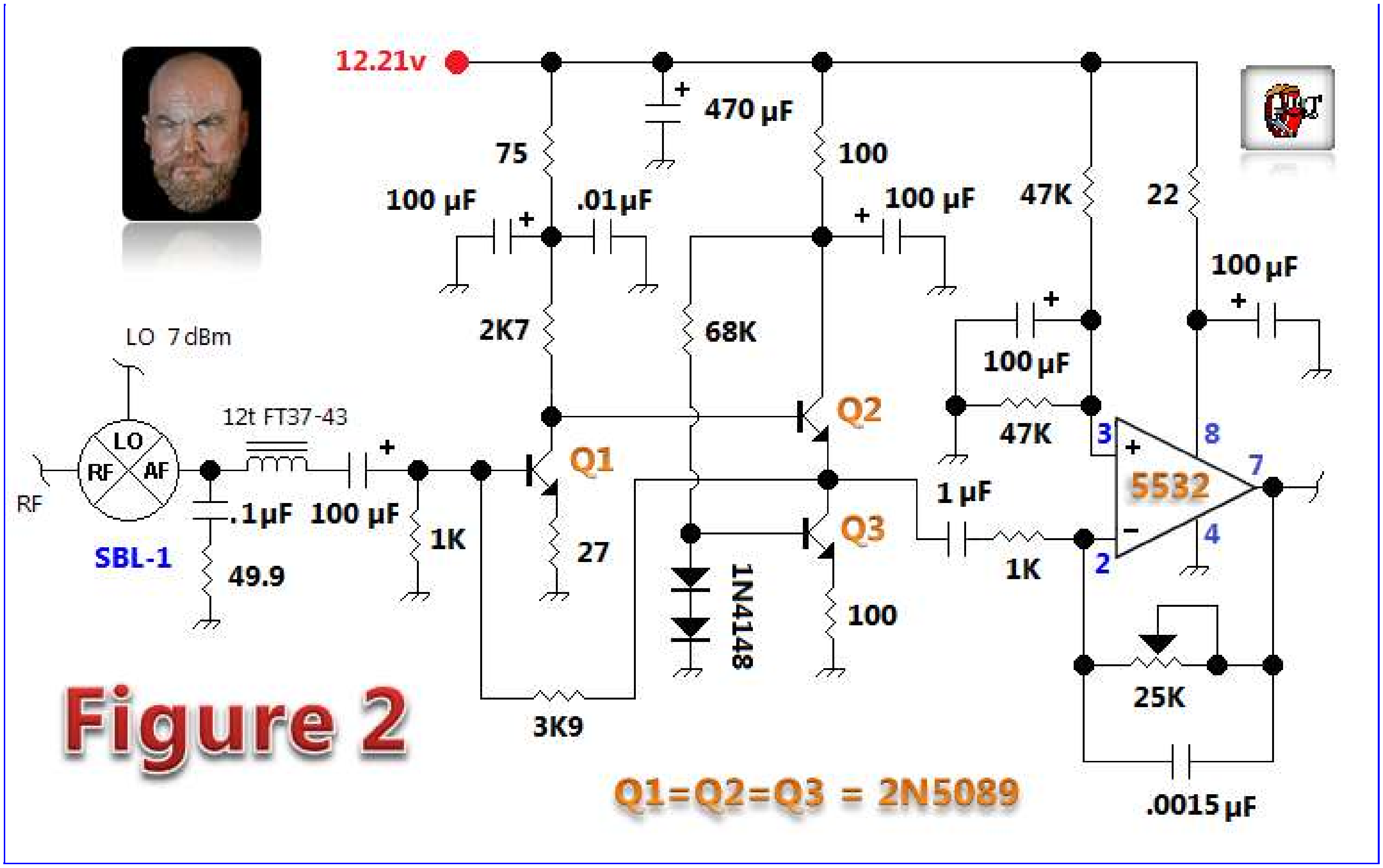
12.21v
470 µF
75
100
100 µF
.01µF
100 µF
47K
22
100 µF
2K7
68K
100 µF
LO 7dBm
12t FT37-43
Q2
47K
3
8
7
LO
RF
AF
Q1
+
5532
1 µF
RF
.1µF
100 µF
SBL-1
1K
27
Q3
–
4
1K
2
49.9
1N4148
100
25K
3K9
.0015 µF
Q1=Q2=Q3 = 2N5089

Figure 2

Above — The "Full Monty product detector to AF preamp". A MCL SBL-1 diode ring mixer drives the 50 Ω input Z feedback pair and finally a 5532 op-amp with adjustable gain. The 25K gain trimmer pot allows the builder to accommodate factors such as whether the stage is a DC receiver, or an RF-AF block for a superheterodyne receiver like in this radio. I'll set the gain as low as possible and hope to build up my AC voltage with active, gain producing low-pass filter op-amp stages.

The output t pin 7 is directly coupled to an op-amp AF filter. With software we can easily design a myriad of audio filters: band-pass, low-pass, high-pass, all-pass etc. I only run low-pass filters in this slot and even here we must decide whether to make our reponse Gaussian to 6 dB (or 12 dB), Bessel, Butterworth, or a Chebyshev with some degree of ripple. I'm testing out a few designs with an eye towards simplicity and 10% parts.

Spurred on by curiosity, with my ears as the main evaluation tool; I currently have a 5532 paper design to deliver 20 dB gain with a 500 Hz pass-band, Gaussian to 12 dB low-pass response. Click for schematic, I chose a 500 Hz cut off frequency since the skirt shape by definition is subtle and the Q low. This trick should better roll off the AF above 1 KHz a little faster without the ugly group delay associated with steeper filters employing more poles. We shall see!

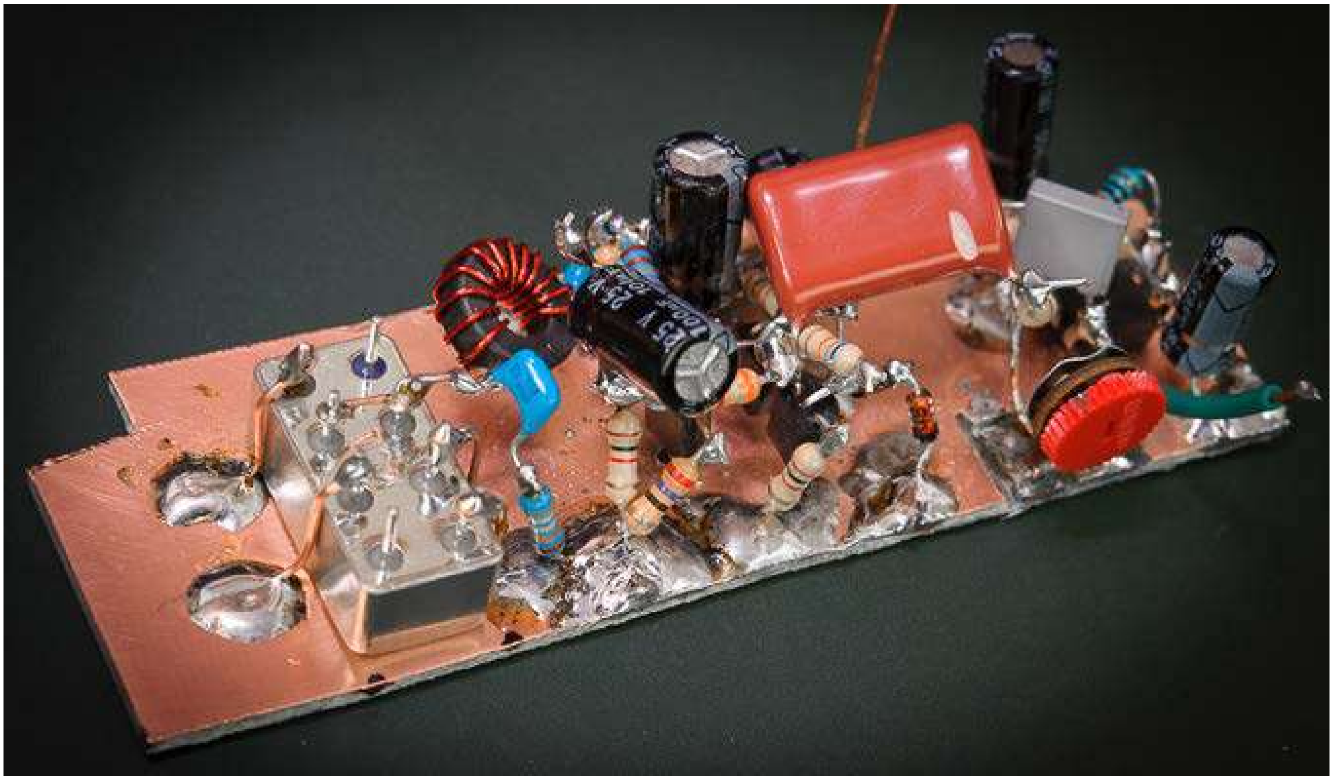

Above — The compact breadboard of Figure 2.

Audio filters and AF power amp are incomplete.....

RF — Test and Measurement

Almost Popcorn Superhet — Supplemental Lab Notes

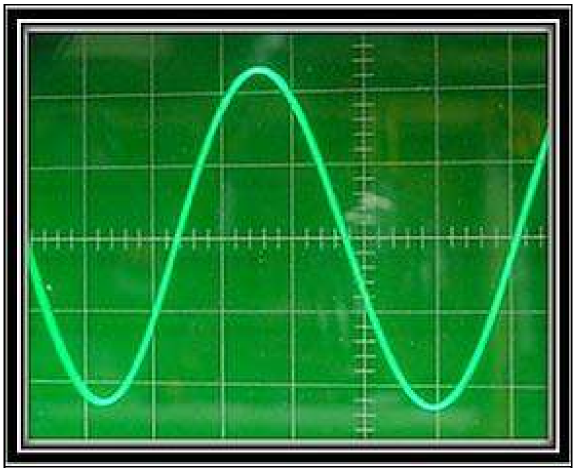

These lab notes supplement the Almost Popcorn Superhet Receiver Web page which has remained incomplete for greater than 10 years. I'm sorry for the rough writing and procrastination.

N = 3 Gaussian-to-6 dB Crystal Ladder Filter

I spent a pleasant Saturday afternoon designing, building and testing Gaussian-to-6 dB crystal filters. I inputted some xtal parameters + constants from Zverev into Xlad08.exe from EMRFD. Here's the N = 3 version

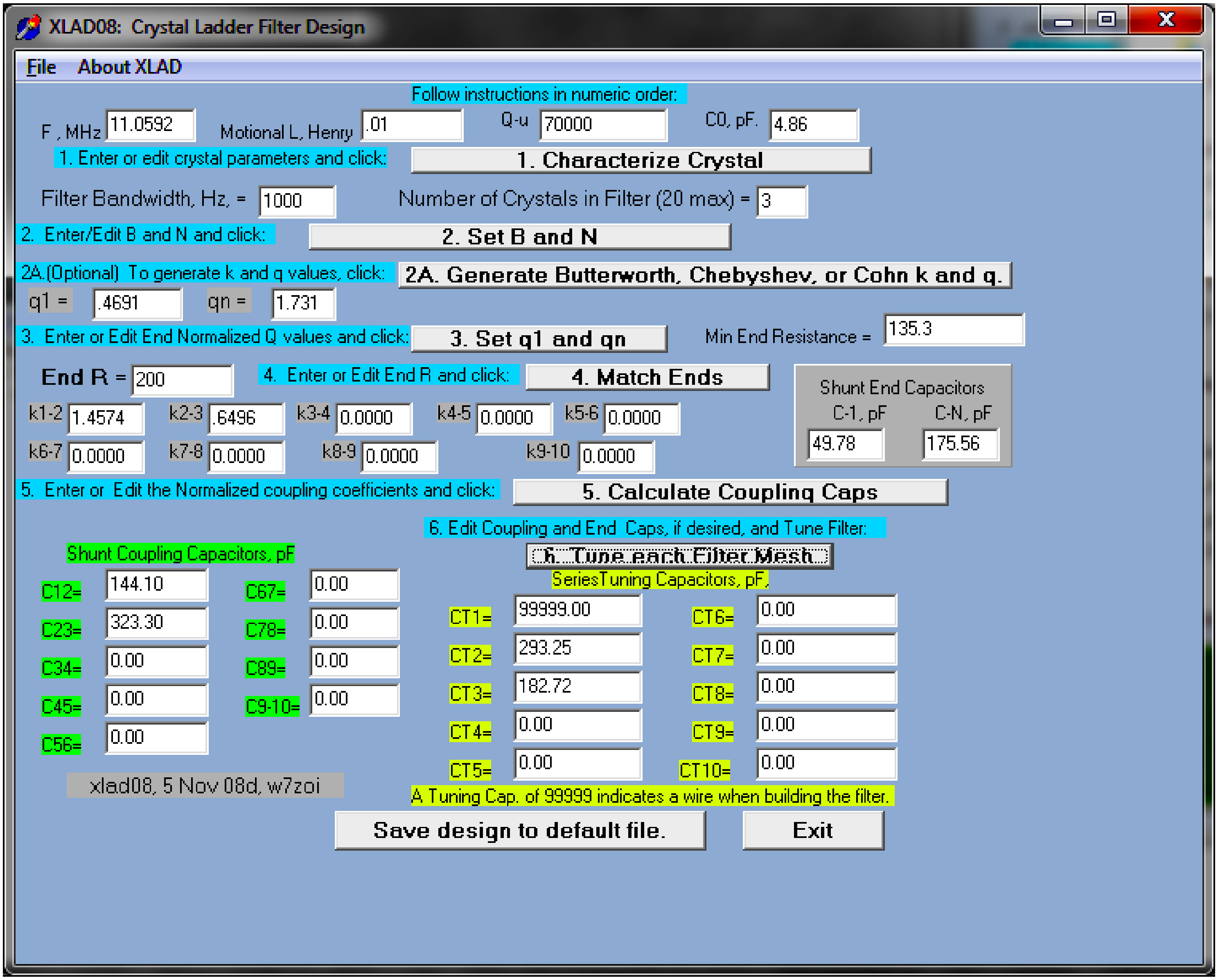
XLAD08: Crystal Ladder Filter Design
File
About XLAD
1. Characterize Crystal
2. Set B and N
2A. Generate Butterworth, Chebyshev, or Cohn k and q.
3. Set q1 and qn
4. Match Ends
5. Calculate Coupling Caps
6. Tune each Filter Mesh
Save design to default file.
Exit

Follow instructions in numeric order:

F , MHz 11.0592
Motional L, Henry .01
Q-u 70000
C0, pF. 4.86

1. Enter or edit crystal parameters and click:

Filter Bandwidth, Hz, = 1000
Number of Crystals in Filter (20 max) = 3

2. Enter/Edit B and N and click:

2A.(Optional) To generate k and q values, click:

q1 = .4691
qn = 1.731

3. Enter or Edit End Normalized Q values and click:

Min End Resistance = 135.3

End R = 200

4. Enter or Edit End R and click:

k1-2 1.4574
k2-3 .6496
k3-4 0.0000
k4-5 0.0000
k5-6 0.0000
k6-7 0.0000
k7-8 0.0000
k8-9 0.0000
k9-10 0.0000

Shunt End Capacitors
C-1, pF 49.78
C-N, pF 175.56

5. Enter or Edit the Normalized coupling coefficients and click:

6. Edit Coupling and End Caps, if desired, and Tune Filter:

Shunt Coupling Capacitors, pF

C12= 144.10
C23= 323.30
C34= 0.00
C45= 0.00
C56= 0.00
C67= 0.00
C78= 0.00
C89= 0.00
C9-10= 0.00

SeriesTuning Capacitors, pF,

CT1= 99999.00
CT2= 293.25
CT3= 182.72
CT4= 0.00
CT5= 0.00
CT6= 0.00
CT7= 0.00
CT8= 0.00
CT9= 0.00
CT10= 0.00

xlad08, 5 Nov 08d, w7zoi

A Tuning Cap. of 99999 indicates a wire when building the filter.

Above — A screen capture from Xlad. This app allows skilled builders to depart from the typical Cohn (Min-Loss) filter and tune filter poles to derive beautiful skirt shapes and low insertion loss if wanted. I'm not a great crystal filter designer, but I've seen work from friends who have a knack for it: imagine making a N= 8 to 10 crystal CW filter with just ~1 dB insertion loss like these builders! Click for the GPLA simulation.

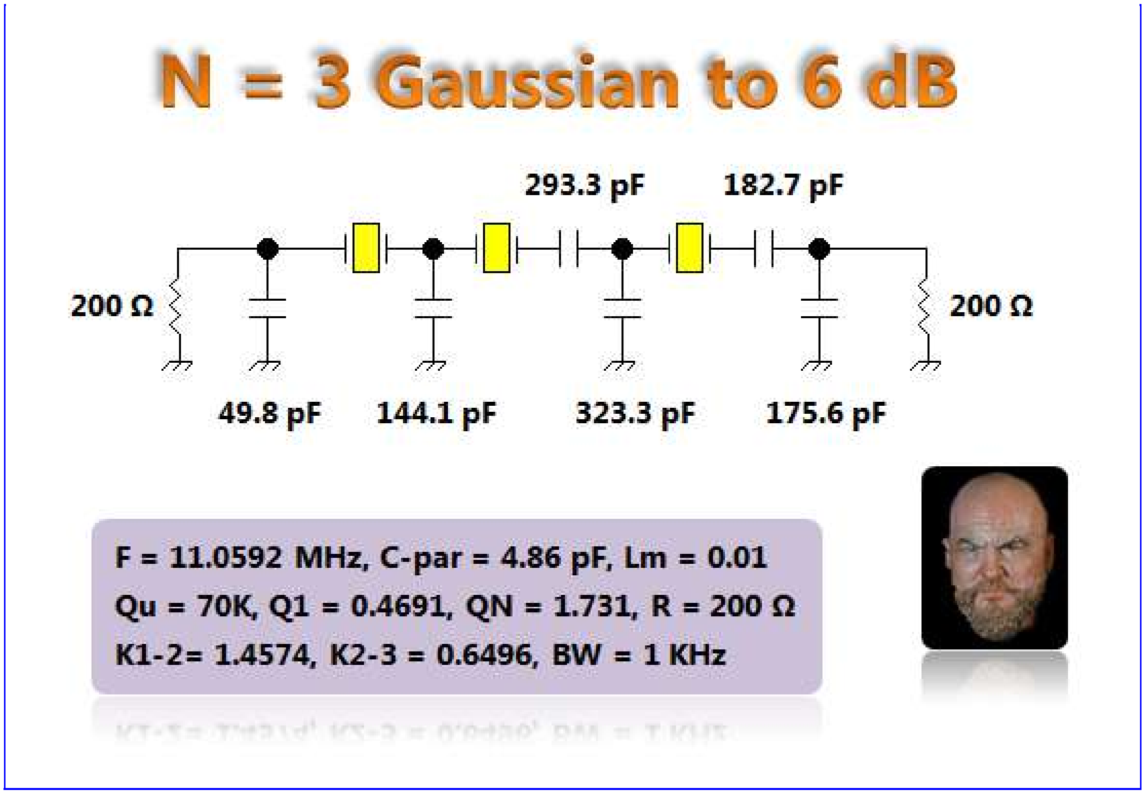
N = 3 Gaussian to 6 dB
293.3 pF
182.7 pF
200 Ω
200 Ω
49.8 pF
144.1 pF
323.3 pF
175.6 pF
F = 11.0592 MHz, C-par = 4.86 pF, Lm = 0.01
Qu = 70K, Q1 = 0.4691, QN = 1.731, R = 200 Ω
K1-2= 1.4574, K2-3 = 0.6496, BW = 1 KHz

Above — Gaussian-to-6 dB xtal filer schematic where N = 3.

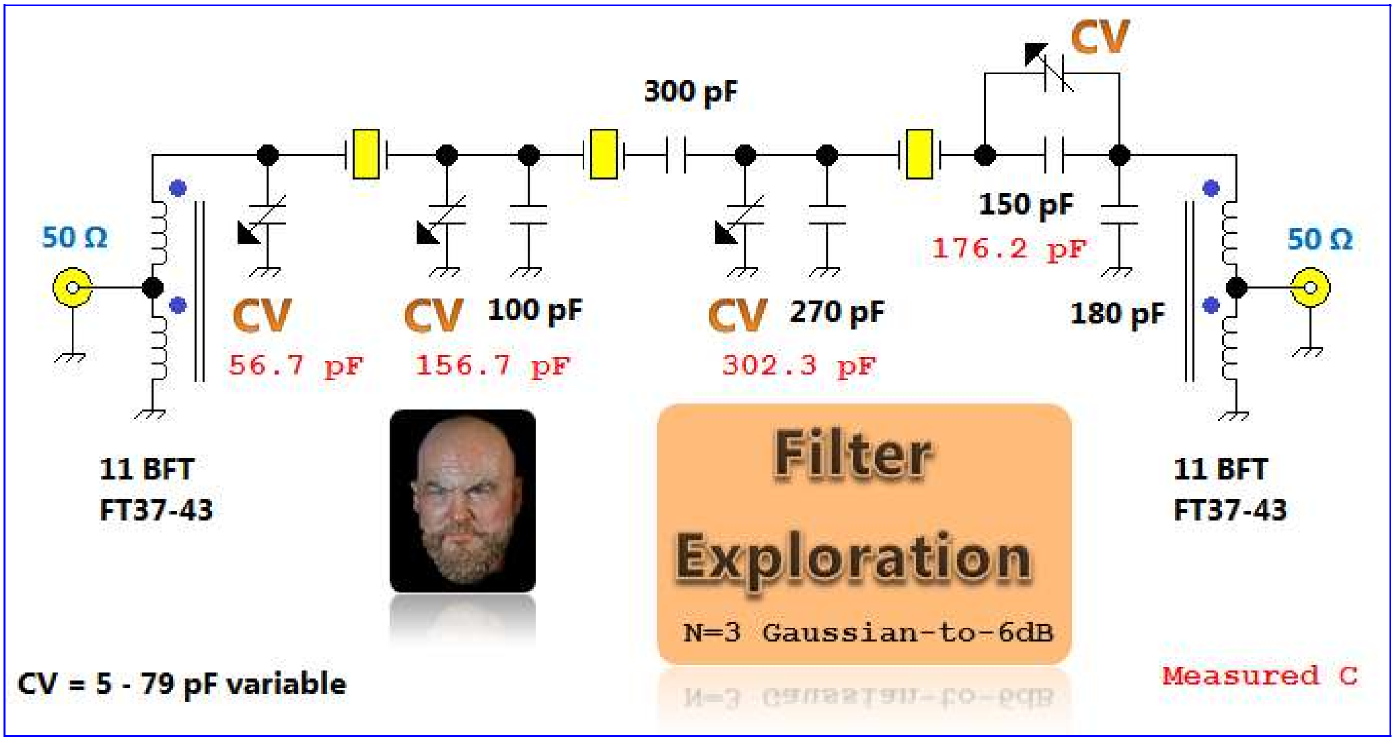
CV
300 pF
50 Ω
50 Ω
150 pF
176.2 pF
CV
56.7 pF
CV
156.7 pF
100 pF
CV
302.3 pF
270 pF
180 pF
11 BFT
FT37-43
Filter
Exploration
N=3 Gaussian-to-6dB
11 BFT
FT37-43
CV = 5 - 79 pF variable
Measured C

Above — Circuit to allow evaluation of the filter in my 50 Ω sweep system. Transmission transformers provide the needed 200 Ω filter termination. I inserted 4 trimmer caps to permit tweaking of the filter under test to establish the very best filter shape. With trimmer capacitor adjustment, I was able to trash the filter or peak it well. Click for 3 different swept bandwidths: Sweep-1 Sweep-2 Sweep-3.

After peaking, I removed *all* the caps and measured them as shown in **red font**. I'll replicate these values with nearest standard-value fixed capacitors. The trimmer caps nearest to the ports proved more sensistive compared to the 2 middle trimmers capacitors during tuning.

This might be the coolest thing I've ever done on my test bench.

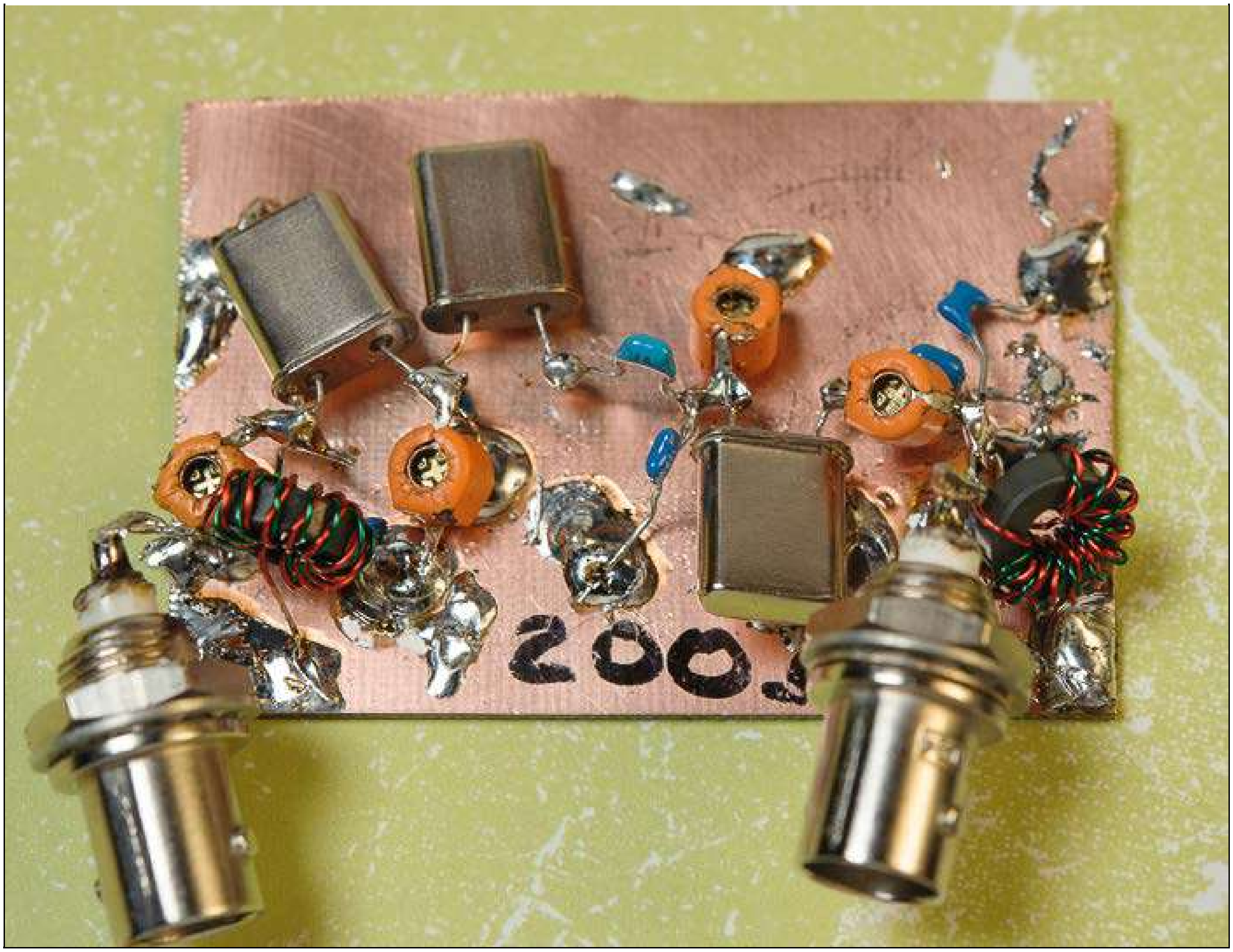

Above — My N = 3 filter breadboard. I left the capacitor leads long for re-use. I consume huge numbers of parts and recycle as possible to keep costs down.

Mixer and Post Mixer Amplifer Development

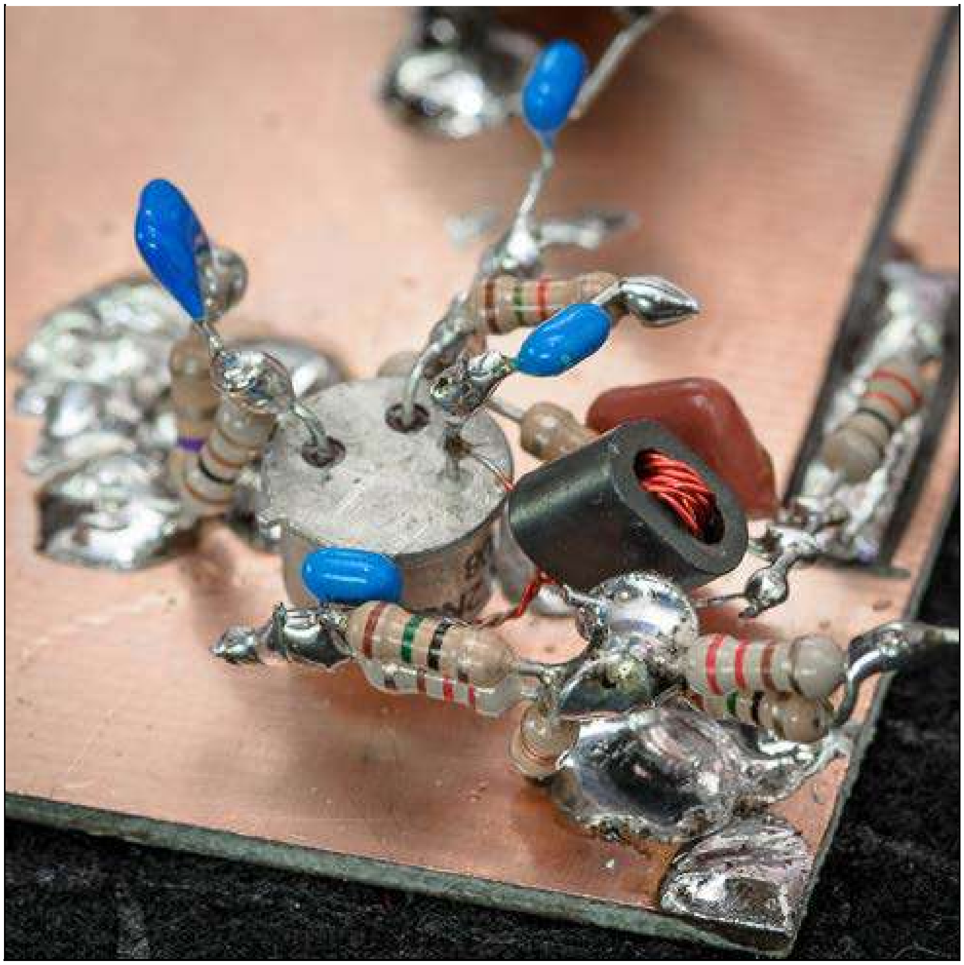

Above — Ugly breadboard shot with my ancient 105 mm lens for a change. It's a Beaverton Special feedback amp (FBA) biased for ~40 mA. For my post-mixer amp. If you need a stout RF amp, this one will do in a storm.

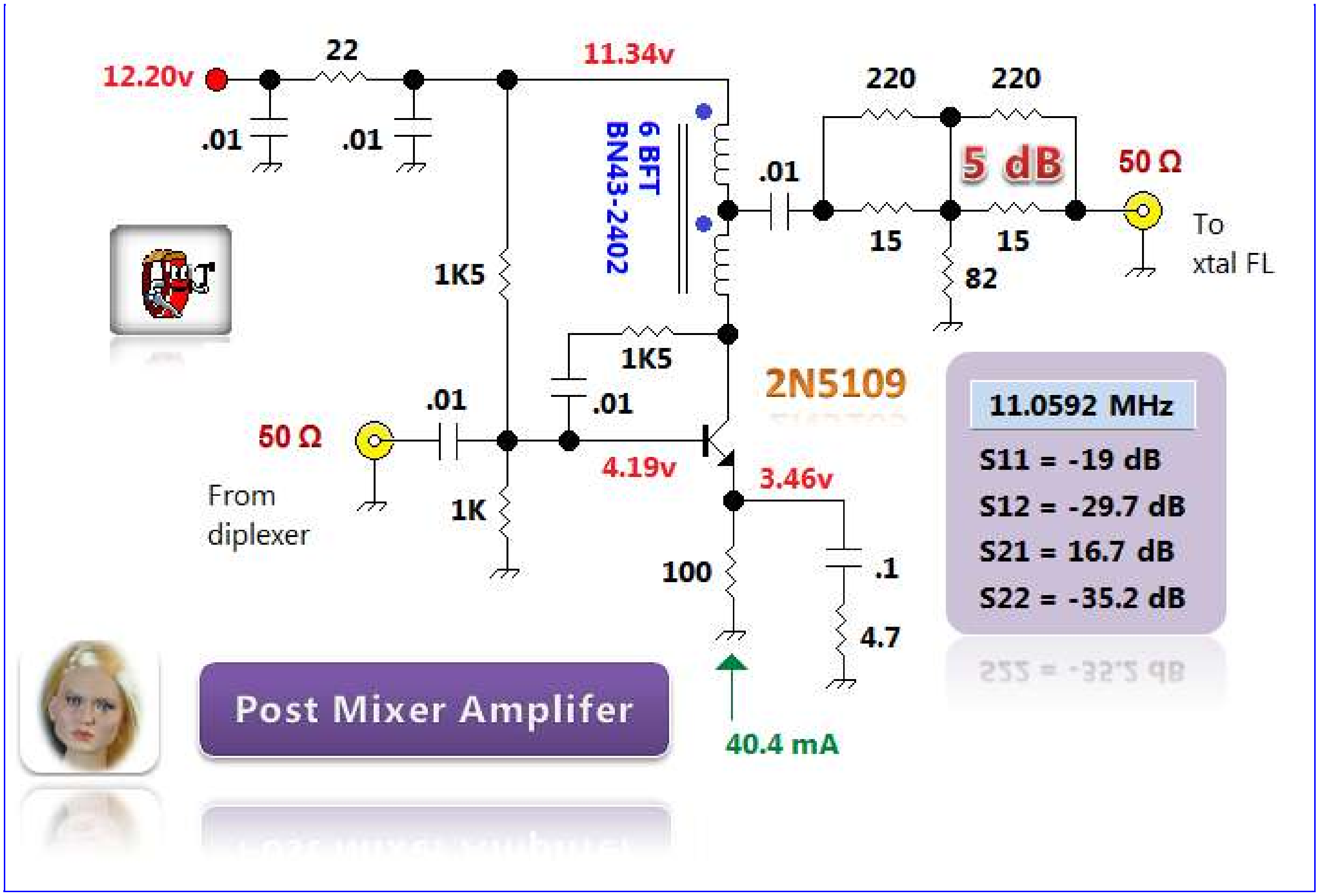
12.20v
22
11.34v
.01
.01
6 BFT
BN43-2402
220
220
.01
5 dB
50 Ω
15
15
82
To
xtal FL
1K5
1K5
.01
.01
2N5109
50 Ω
4.19v
3.46v
From
diplexer
1K
100
.1
4.7
40.4 mA
11.0592 MHz
S11 = -19 dB
S12 = -29.7 dB
S21 = 16.7 dB
S22 = -35.2 dB
Post Mixer Amplifer

Above — The Beaverton Special feedback amp biased for ~40 mA with a 50 Ω output Z. The FBA with no 5 dB pad: S11 -18.5 dB; S12 -34.6 dB; S21 22.1 dB and S22 -31.8 dB. I love this FBA for its simplicity and well defined port impedance. My IF = 11.0592 MHz. Click for a discarded (but cool) photo.

To drive the 200 Ω input Z of my xtal filter I later converted the output to 200 Ω by changing the collector transformer and pad.

Frequency Synthesizers and Logic

I've spent a great deal of time reading about & playing with some HC series logic I bought in 2011-2012. I even purchased a few 74AC74, a D flip-flop you can clock up to 160 MHz or so. Most of my reading concerned frequency synthesizers: PLLs, dividers, prescalers, etc. Since most modern devices run a clock at GHz frequencies, the evolution of frequency synthesis technology from around the mid 1970's to now makes for a fascinating read.

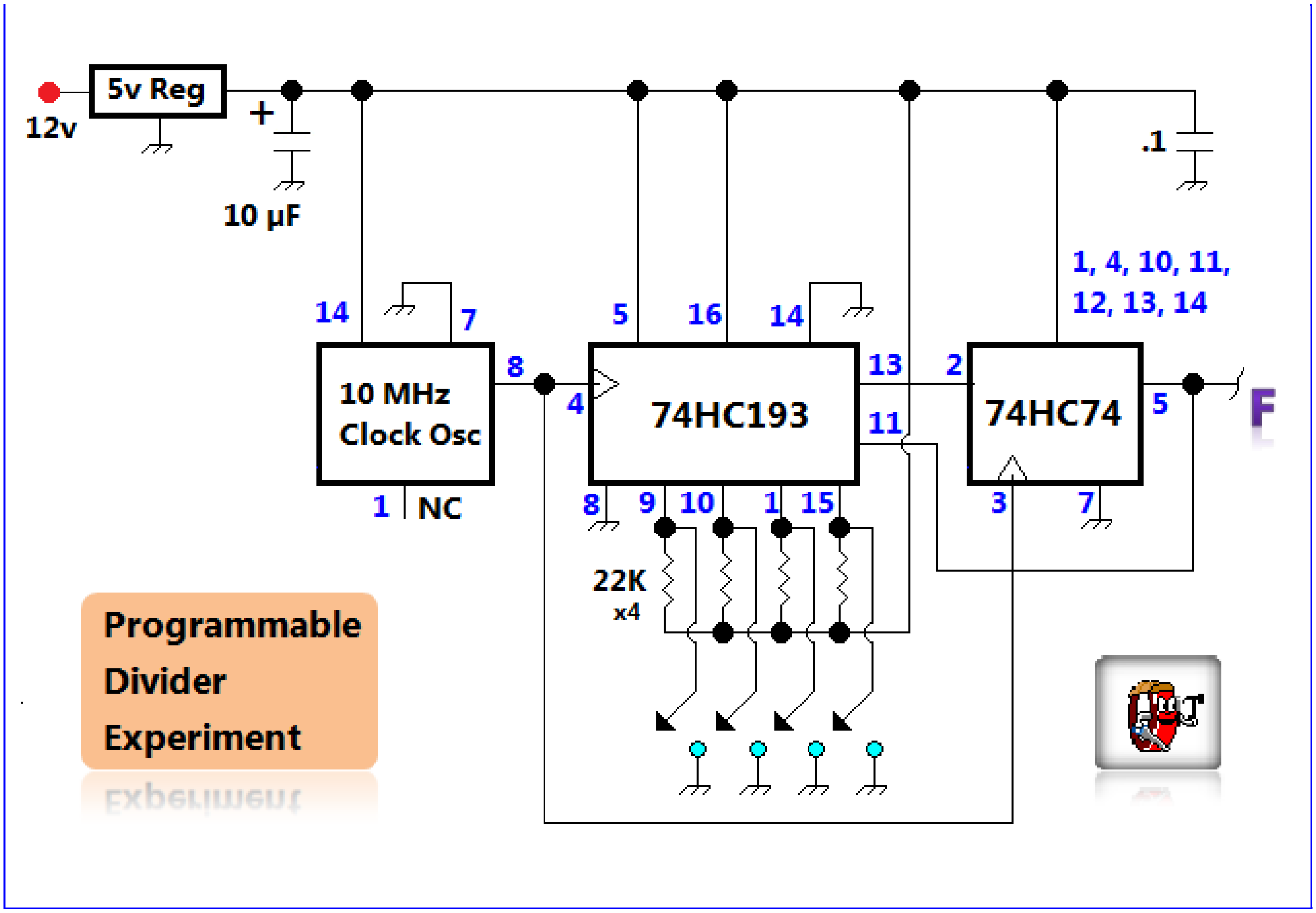
5v Reg
12v
+
10 µF
.1
14
7
10 MHz Clock Osc
1
NC
8
4
5
16
14
74HC193
13
11
8
9
10
1
15
22K
x4
2
74HC74
1, 4, 10, 11,
12, 13, 14
5
3
7
F
Programmable
Divider
Experiment

Above — Using a CMOS 10 MHz clock oscillator, I tested out a positive edge-triggered programmable frequency divider designed by Wes, W7ZOI in 2011. It works perfectly.

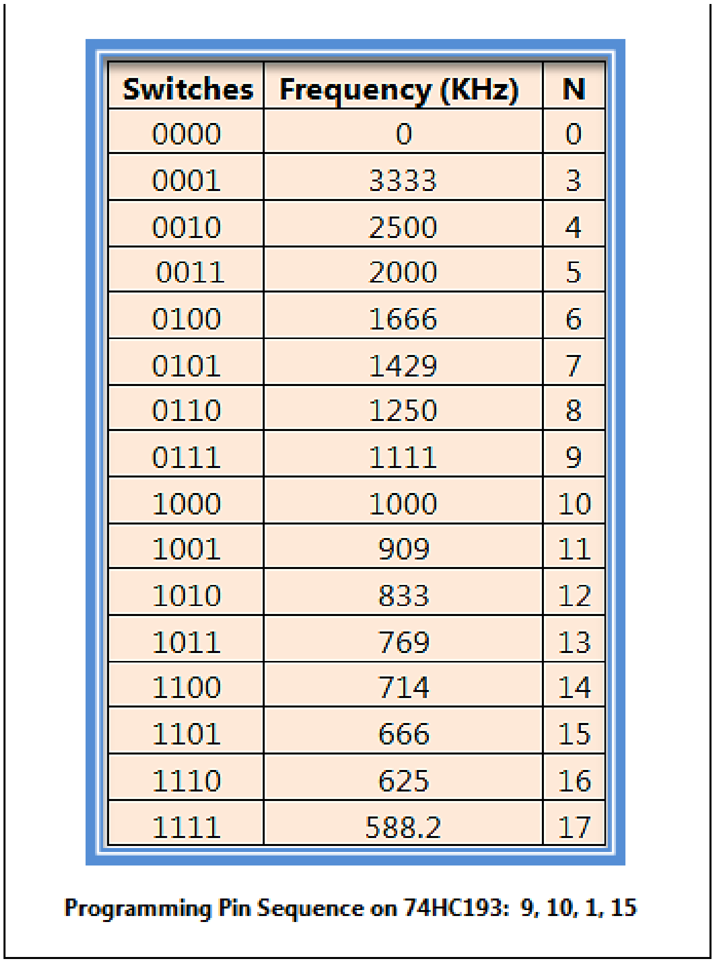

Switches	Frequency (KHz)	N
0000	0	0
0001	3333	3
0010	2500	4
0011	2000	5
0100	1666	6
0101	1429	7
0110	1250	8
0111	1111	9
1000	1000	10
1001	909	11
1010	833	12
1011	769	13
1100	714	14
1101	666	15
1110	625	16
1111	588.2	17

Programming Pin Sequence on 74HC193: 9, 10, 1, 15

Above — A test of all 16 combinations of the 4 programming switches with the 10,000 KHz input. I show the measured output frequency in KHz and the integer generated by dividing the output F into 10,000 KHz. You may leave pin 9 HIGH and just run it as a divide by 10 to 17 counter with 3 panel mounted programming switches — that's what I hope to do.

While working at Tektronixs, Wes designed the frequency synthesizer for the 492 Spectrum Analyzer; a portable SA that went to 21 GHz or so. His experiences informed the design of this simple, low-cost, edge-triggered divider.

Here, Wes applies the 74HC193 as a down counter where a number is loaded and then it count downs until zero is reached. The number gets boosted by 2 — 1 comes from the 74HC74 D flip-flop, while the other arises from the phasing of the signals. The 74HC74 makes the overall output coherent with its clock and avoids the flicker noise that otherwise might be generated in the long divider chain of the ripple counter.

The standard binary to decimal conversion is altered in this circuit.

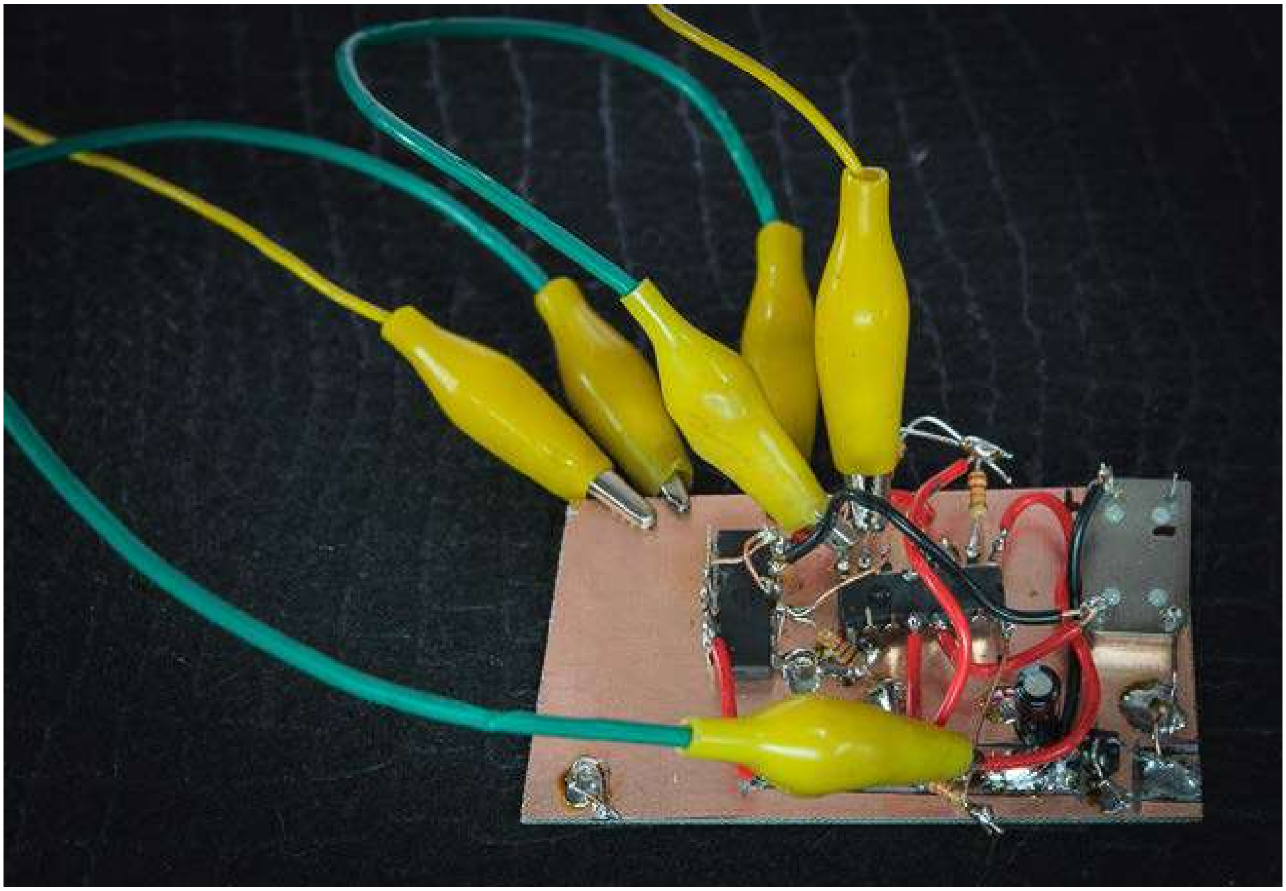

Above — My ugly breadboard. Instead of switches, I relied on bench jumpers to set a pin LOW.

Check out the standard binary switch combinations on this cool website: Math Is Fun. I also enjoyed their hexidecimal drum machine tool - Click. Actually, the whole web site looks great and I will recommend it to our friends with children and maybe even a few radio enthusiasts.

The Local Oscillator

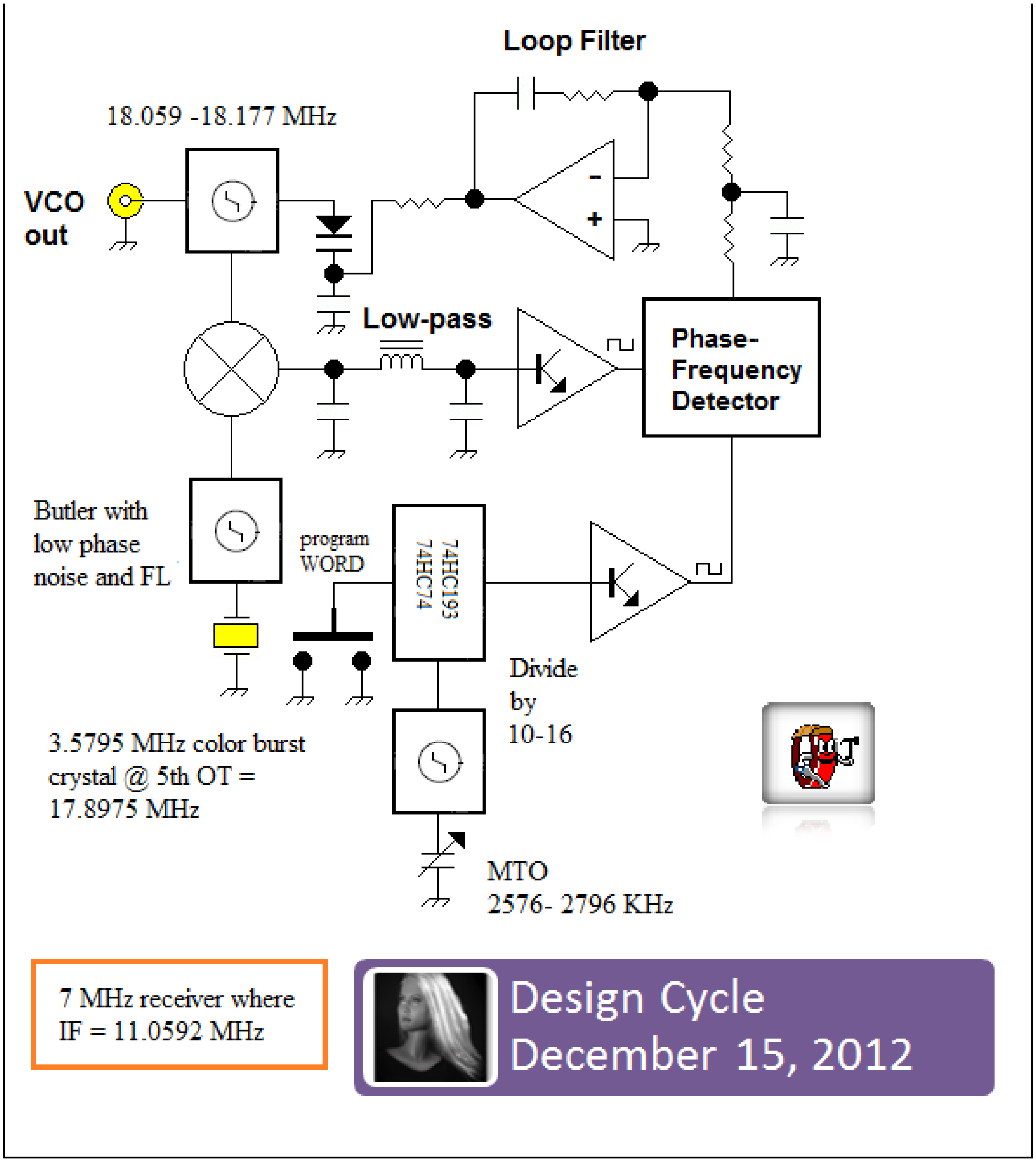
Loop Filter
18.059 -18.177 MHz
VCO out
-
+
Low-pass
Phase-Frequency Detector
Butler with low phase noise and FL
program WORD
74HC193
74HC74
Divide by 10-16
3.5795 MHz color burst crystal @ 5th OT = 17.8975 MHz
MTO 2576- 2796 KHz
7 MHz receiver where IF = 11.0592 MHz
Design Cycle
December 15, 2012

Above — An Almost Synthesizer based upon the design currently in use at W7ZOI.

I picked the MTO, xtal oscillator and VCO frequencies and will design the loop filter once I get the parts and a few small die-cast boxes on hand. The MTO division integer is low and quiet — plus this scheme enhances MTO frequency stability and spectral purity. A 4-bit ripple counter + flip flop perform the division where N = 10 - 16. I might go higher with N , but don't want to get too greedy. For CW, I'm generally in the bottom 20 KHz of the band anyway.

The MTO tunes about a 13-14 KHz range and the N gives 6 different tuning ranges (with some overlap) to span a total range of ~118 KHz. The 4-bit WORD is changed by 3 chassis-mount toggle switches and give a PLL reference of ~~175 to 280 KHz.

This almost synthesis seems a little crude in a time when people use microcontrollers to input a WORD into a chip with with the simple turn of a rotary controller, however this set-up fosters learning, huge fun, and when the oscillators plus loop filter are well designed; low phase noise and reciprocal mixing.

The Crystal Filter

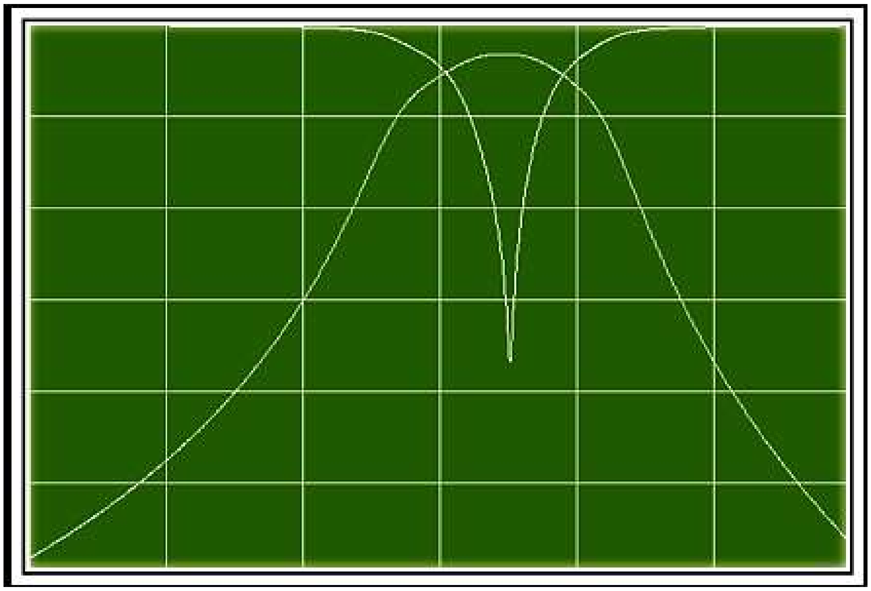

Above — A GPLA sweep of my 4 crystal filter where C parallel = 4.86 pF, Lm = .01 H, IF = 11.0572 MHz, BW = 1 KHz and R term = 200 ohms. This Gaussian to 6 dB filter design was chosen for speaker audio that sounds full + dynamic without the hollow ringing so often heard in homebrew receivers. I filter in my brain and first learned to copy CW across a room while listening to a speaker.

My code mentor Doug, could make coffee, walk around the room, talk to us and still 100% copy CW at 40 WPM. He taught us the ear to brain bandwidth is ~ 100 Hz when copying weak signal CW through a sideband crystal filter. To this day, I prefer listening through wider CW filters with brain band-pass filtering.

KL7R Memorial Supplemental Page

This web page is a supplement to the K7LR Memorial Receiver Experiments Web Page

Audio Preamplifier Focus

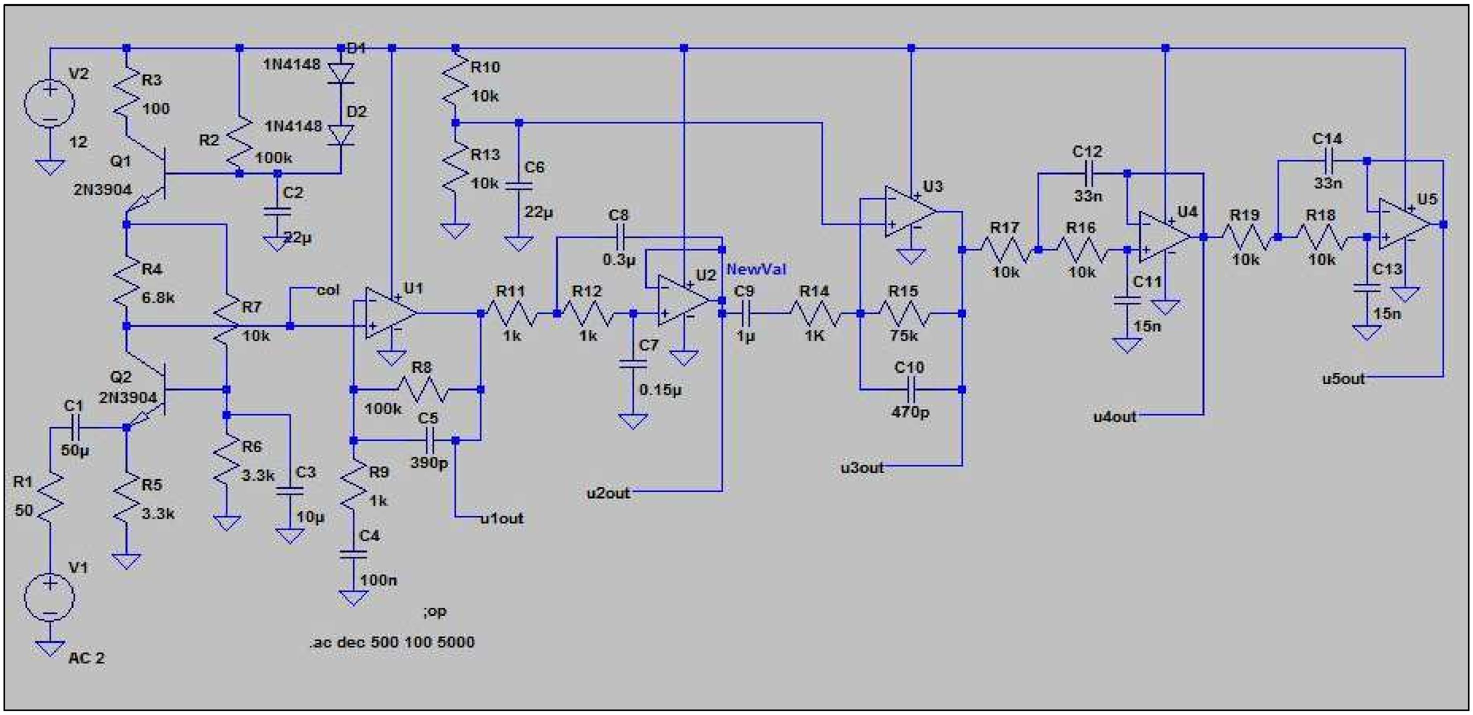
V2
12
R3
100
Q1
2N3904
R2
100k
1N4148
D1
1N4148
D2
C2
22µ
R4
6.8k
R7
10k
col
Q2
2N3904
C1
50µ
R1
50
R5
3.3k
R6
3.3k
C3
10µ
V1
AC 2
U1
R8
100k
C5
390p
R9
1k
C4
100n
u1out
;op
.ac dec 500 100 5000
R10
10k
R13
10k
C6
22µ
R11
1k
R12
1k
C8
0.3µ
C7
0.15µ
U2
NewVal
u2out
C9
1µ
R14
1K
R15
75k
C10
470p
U3
u3out
R17
10k
C12
33n
R16
10k
C11
15n
U4
u4out
R19
10k
C14
33n
R18
10k
C13
15n
U5
u5out

Shown above is a SPICE build of an earlier prototype of the audio preamp chain which was ultimately used in this receiver. This SPICE work was performed by Wes, W7ZOI. Note that at this point, I had 100K plus 390 pF feedback in U1 and a 75K plus 470 pF feedback resistors in the op-amp amplifiers (really massive gain). I spent 4-5 days trying different audio stages before settling on the circuits shown.

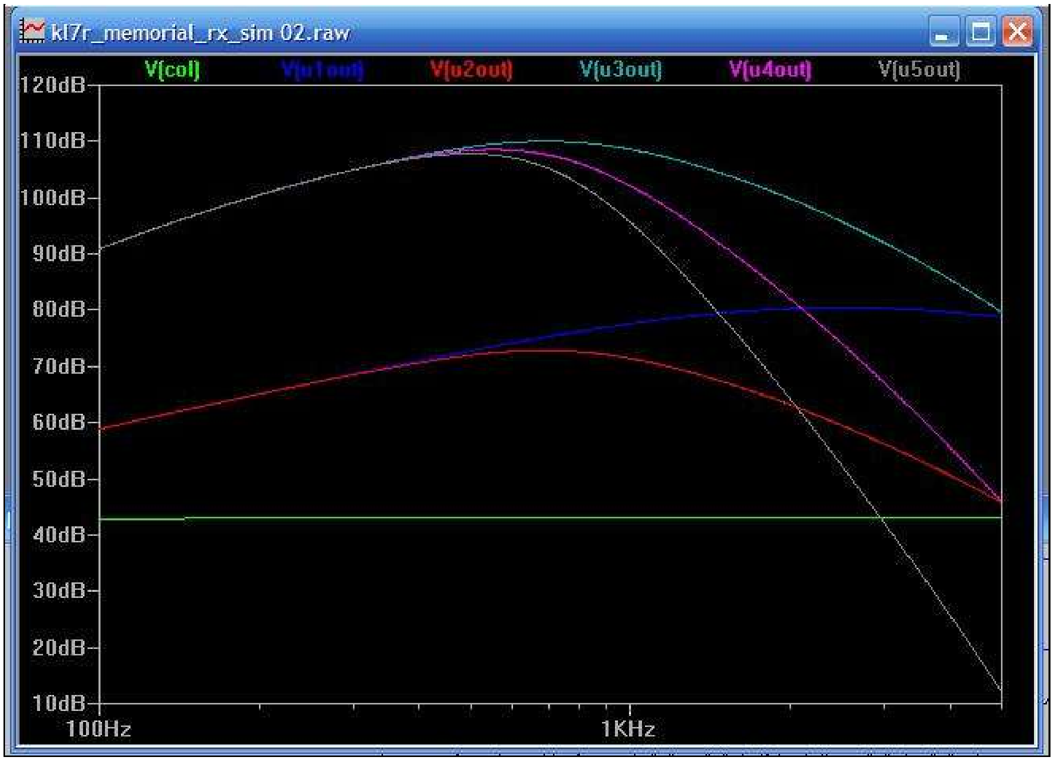

Shown above is the voltage versus frequency plot from a SPICE run of the schematic shown above. The gentle slope (decibels per octave) of the low pass filters can be seen. Certainly Chebychev filters with 1-2 dB of ripple would have provided steeper filtering, but the filters would have odd resistor values and be less easily modified or reproduced by popcorn builders. The V(u5out) shows a reasonable low pass response for the AF preamp.

Audio Power Amp

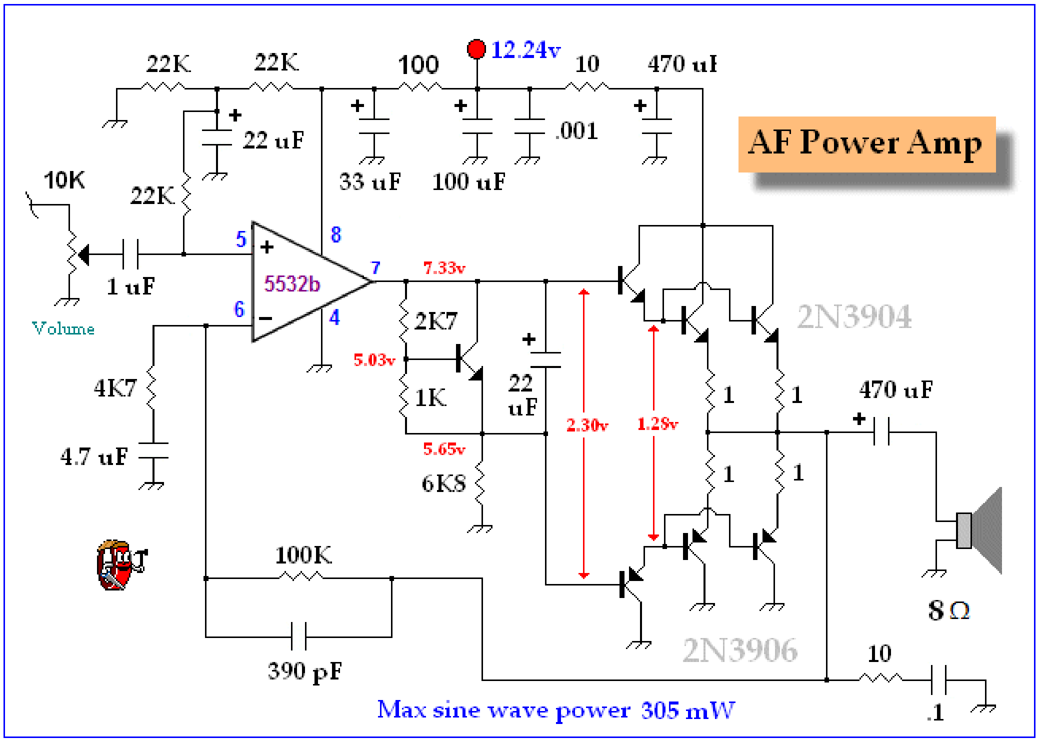
22K
22K
100
12.24v
10
470 uF
22 uF
33 uF
100 uF
.001
AF Power Amp
10K
22K
Volume
1 uF
5
6
8
7
4
5532b
7.33v
2K7
5.03v
1K
5.65v
6K8
22 uF
2.30v
1.28v
2N3904
1
1
1
1
470 uF
4K7
4.7 uF
100K
390 pF
2N3906
8 Ω
10
.1
Max sine wave power 305 mW

This schematic shows a stand-alone version of the audio power amp. If only 1/2 of the 5532 is used, connect pins 3 and 2 of the unused op-amp to the point where all the 22K resistors connect (VCC/2 bias point).

Product Detector Notes

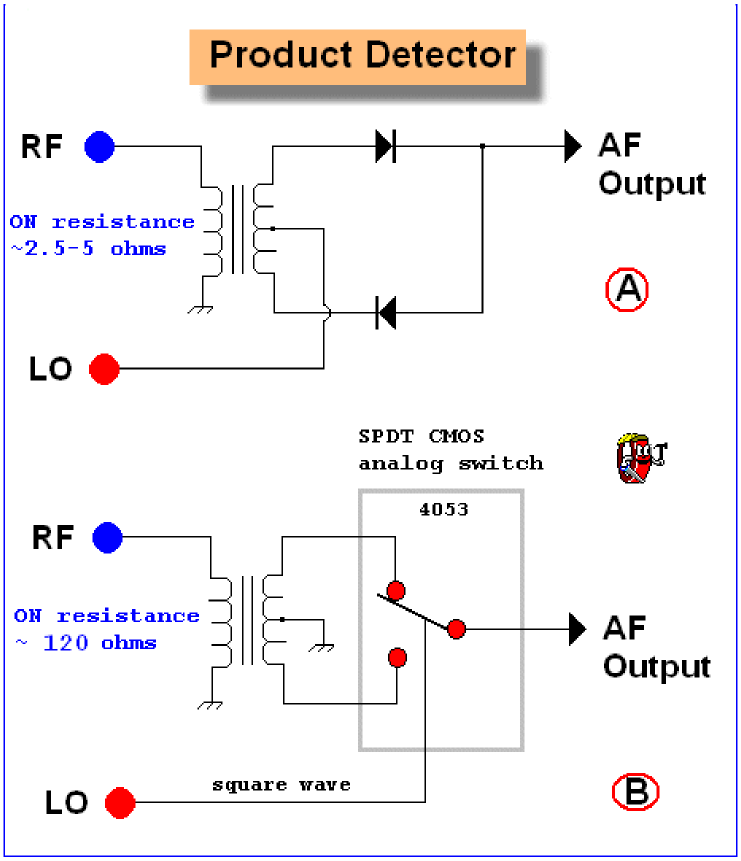
Product Detector
RF
ON resistance
~2.5-5 ohms
LO
AF
Output
A
SPDT CMOS
analog switch
4053
RF
ON resistance
~ 120 ohms
AF
Output
square wave
LO
B

Shown above is the analogy between an SPDT CMOS switch and a single-ended, diode product detector. In both cases (A and B), the LO causes 1 diode (switch) to conduct, while the other is OFF. Then, on the other half cycle, the other diode (switch) conducts. If you wish to actually using the product detector/mixer depicted as "A" go with Figure 5.19B in EMRFD which has the LO applied to the transformer primary. As shown, "A" has poor LO to RF isolation, but it helps model the CMOS SPDT analog switch function. Function and performance are completely different topics... The on-resistance of the 4053 is much higher and thus performance is not identical.

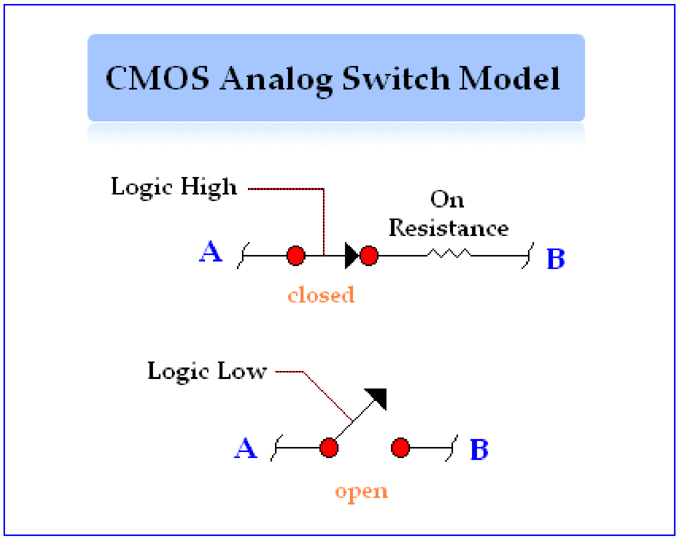
CMOS Analog Switch Model
Logic High
On
Resistance
A
B
closed
Logic Low
A
B
open

CMOS analog switch model. A or B are bidirectional and can be used interchangeably for the input or output. Ideally the on-resistance should be low to reduce propagation delay of digital signals plus resistance and perhaps distortion of analog signals. They can be used to switch digital bus data or analog signals. CMOS switches can replace mechanical reed switches in some low power circuits.

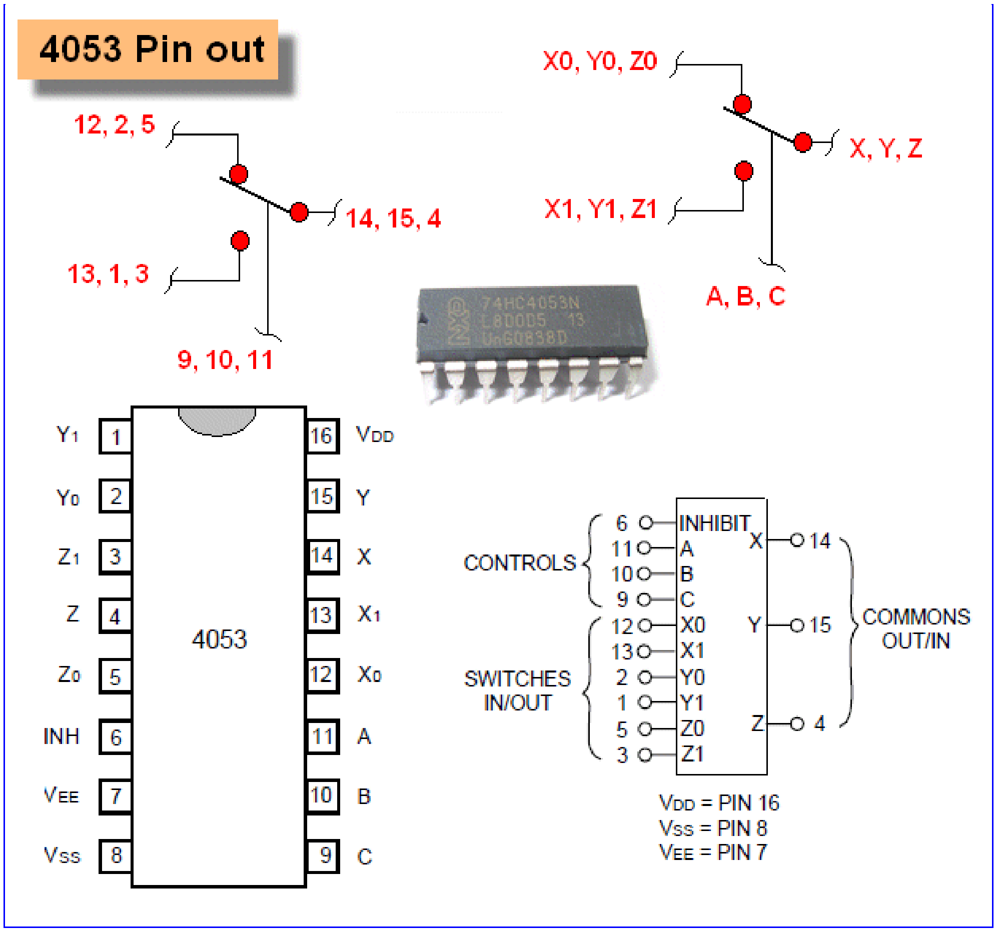
4053 Pin out
12, 2, 5
14, 15, 4
13, 1, 3
9, 10, 11
X0, Y0, Z0
X, Y, Z
X1, Y1, Z1
A, B, C
74HC4053N
L8D0D5 13
UnG0838D
Y1 1
Y0 2
Z1 3
Z 4
Z0 5
INH 6
VEE 7
VSS 8
16 VDD
15 Y
14 X
13 X1
12 X0
11 A
10 B
9 C
4053
CONTROLS
6 INHIBIT
11 A
10 B
9 C
SWITCHES IN/OUT
12 X0
13 X1
2 Y0
1 Y1
5 Z0
3 Z1
X 14
Y 15
Z 4
COMMONS OUT/IN
VDD = PIN 16
VSS = PIN 8
VEE = PIN 7

The CD4053BC is a triple 2-channel multiplexer having three separate digital control inputs; A, B, and C, and an inhibit input. Each control input selects one of a pair of channels which are connected in a single-pole double-throw configuration. A high on resistance makes the 4053 a compromise part for RF mixing. It might be perfectly okay for your design.

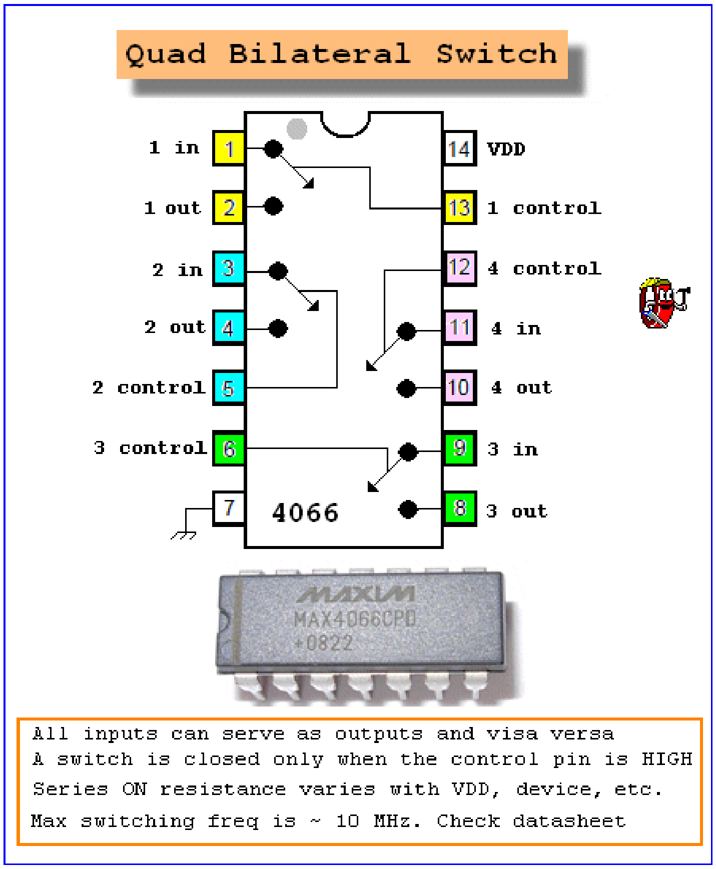
Quad Bilateral Switch
1 in 1
1 out 2
2 in 3
2 out 4
2 control 5
3 control 6
7
14 VDD
13 1 control
12 4 control
11 4 in
10 4 out
9 3 in
8 3 out
4066
MAXIM
MAX4066CPD
+0822

All inputs can serve as outputs and visa versa
A switch is closed only when the control pin is HIGH
Series ON resistance varies with VDD, device, etc.
Max switching freq is ~ 10 MHz. Check datasheet

The 4066 is more easily understood when drawn in a semi-schematic. The on-resistance of this MAXIM version is better than the usual HC4066. About 45 ohms with a VCC of 12 or so.

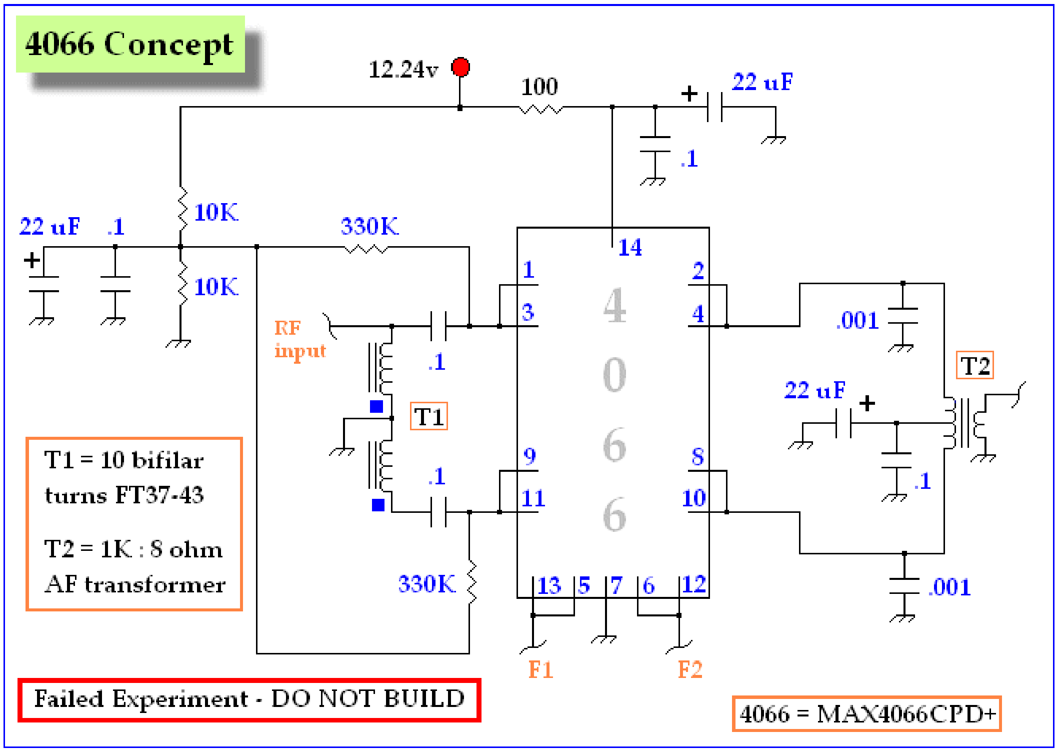
4066 Concept
12.24v
100
22 uF
.1
22 uF
.1
10K
10K
330K
14
1
2
3
4
RF input
.1
T1
.001
T2
22 uF
9
8
11
10
.1
4066
330K
13
5
7
6
12
.1
.001
F1
F2
T1 = 10 bifilar turns FT37-43
T2 = 1K : 8 ohm AF transformer
Failed Experiment - DO NOT BUILD
4066 = MAX4066CPD+

A 4066 experiment that was a total failure. It is both humbling and useful to show your failures and not just the good fruit. This circuit had hum, noise and low output. The input transformer was a bifilar type, which was really cool.

The breadboard of the 4066 Concept schematic shown above.

74HC4053N
L8D0D5 13
UnG0838D
MAXIM
MAX4066CPD
+0822

The DIP IC switches used as product detectors in the K7LR memorial receiver experiments.

My work bench during some final experiments for this project.

K7LR MEMORIAL RECEIVER VFOs

This web page is a supplement to the [K7LR Memorial Receiver Experiments Web Page]

14 MHz VFO Circuit

Figure 1

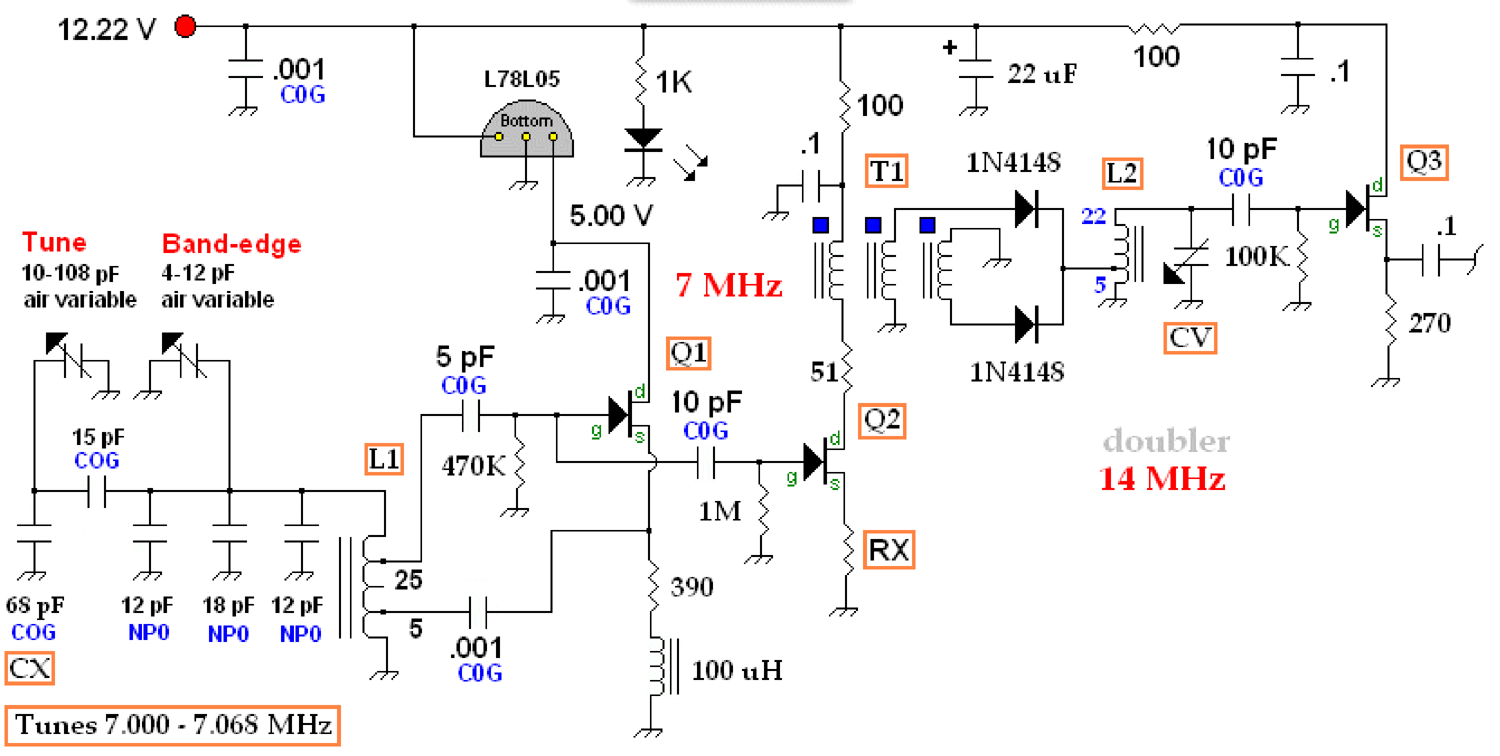
12.22 V
.001
C0G
L78L05
Bottom
1K
5.00 V
.001
C0G
.1
100
T1
+
22 uF
100
.1
1N4148
1N4148
L2
22
5
10 pF
C0G
100K
CV
Q3
g
d
s
.1
270
Tune
10-108 pF
air variable
Band-edge
4-12 pF
air variable
7 MHz
Q1
d
g
s
5 pF
C0G
470K
L1
15 pF
C0G
25
5
.001
C0G
390
100 uH
10 pF
C0G
1M
51
Q2
d
g
s
RX
68 pF
COG
12 pF
NP0
18 pF
NP0
12 pF
NP0
CX
Tunes 7.000 - 7.068 MHz
doubler
14 MHz

L1 = 6.1 uH = 36t on a T68-6 toroid. Tap at 25 and 5 turns. (used 26 AWG)
Q1 = Q2 = Q3 = J310
100 uH RFC that I used was wound as follows: 38t on a FT50-61 ferrite toroid
T1 = 10 trifilar turns on a FT37-43   L2 = 2 uH = 22t on a T50-6; tap @5t
CV = 6.0-50 pF trimmer capacitor
RX = 470 ohms

Figure 1 is the 14 MHz VFO used to switch (clock) the CMOS switch product detector. The tuning range can be easily changed by adjusting capacitor CX. As shown, the the tuning capacitor has a frequency range of about 68 Hz. If you decrease CX to 47 pF, the tuning range increases to around 99 Hz. Increasing CX to 100 pF gave about a 40 Hz tuning range. When you use a larger tuning range, fine tuning can become difficult. A geared reduction dial is one possible solution. For simplicity, I prefer to just keep the tuning range low as reasonably possible and use a big knob on the main tuning variable capacitor. Even the 68 Hertz tuning range shown can prove difficult for fine tuning. As experimenters, we are continually problem solving. Choose the tuning range and/or methods which suit your personal needs.

The base 7 MHz oscillator is doubled since a D flip flop is used to clock the CMOS switch product detector. This oscillator has low harmonic content and this is important for suppressing the 7 MHz fundamental frequency and its harmonics in the frequency doubler. Some astute builders may even match the two diodes used in the doubler. Diode matching is discussed on this page. CV tunes very sharply and the output at Q3 is a clean sine wave. This in turn provides a well balanced square wave at each output of the D flip-flop. The Q3 output voltage was measured at -0.46 dBm (0.6 volts peak to peak into 50 ohms).

Choosing Capacitor Values

There are a number of ways to determine the capacitor values required to tune your VFO L-C tank circuit. It is pretty much essential to have a frequency counter and nice perhaps, to own a capacitance meter. Main tuning capacitors are typically harvested from an old radio or from the junk box. Its tuning range could be anywhere from 30 to 300 pF, or more. To limit the tuning range of this variable capacitor, normally you parallel a fixed-value capacitor and then series connect it to the top of L1 with a small, fixed-value capacitor. I performed this task entirely by trial an error using a frequency counter connected to the collector of Q2. Some fixed value NP0 or C0G capacitors plus a small trimmer capacitor are also required for tuning and to set the lower band edge respectively. Finding the right combination of capacitor values is painstaking, but with practice, gets easier. Once you have the basic capacitor values sorted out and your tuning range set, frequency stability experiments are then performed. This is known as VFO temperature compensation. Some times temperature compensation can be achieved by finding the right combination of NP0 and C0G capacitors. Additionally some negative or positive temperature coefficient capacitors may have to be soldered in and tested. Your final capacitor leads should be short as practically possible to reduce stray lead capacitance and for mechanical rigidity. Temperature compensation is discussed on this web page.

A 7 MHz VFO Circuit for Diode Ring Product Detectors

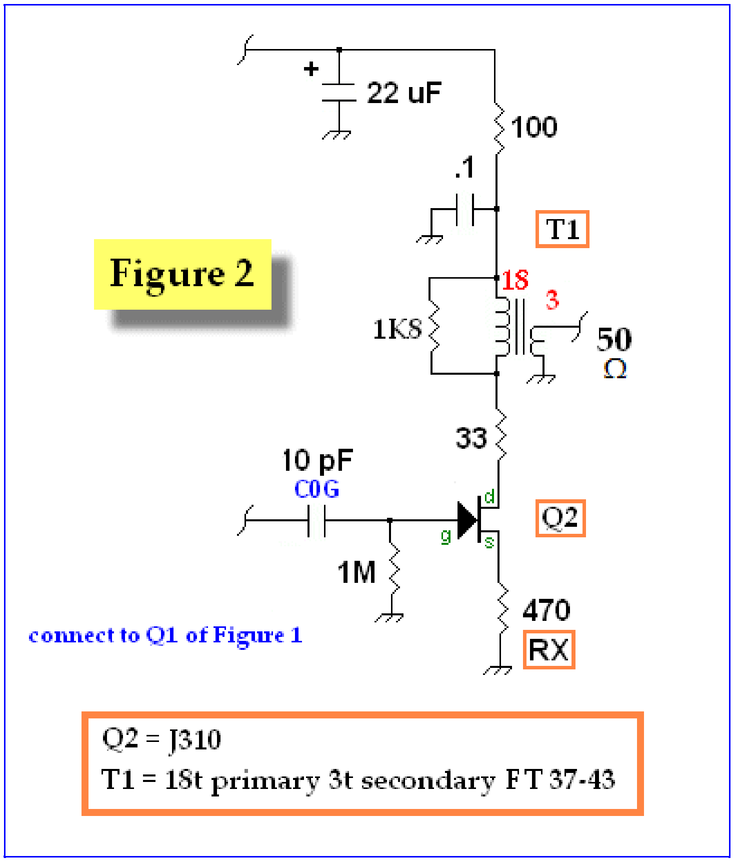
Figure 2
22 uF
+
100
.1
T1
18
3
1KS
50 Ω
33
10 pF
C0G
d
g
s
Q2
1M
470
RX
connect to Q1 of Figure 1
Q2 = J310
T1 = 18t primary 3t secondary FT 37-43

If you want to build a simpler version of the K7LR memorial receiver, a diode ring mixer may be substituted for the digital switch. The VFO requires modification as it will be run at 7 MHz. A different buffer is used and the Q2 to Q3 frequency doubler circuit is excluded. The diode ring product detector version is very nice. You could use a Mini-circuits TUF-1 , SBL-1 or alternate, or perhaps homebrew your own.

In Figure 2 is a VFO buffer configured to drive a 50 ohms input impedance, 7 dBm level, diode ring mixer. Change Rx to change the output voltage. With RX at 470 ohms, the output was somewhere around 5 dBm. This is useable for many situations. If you want exactly 7 dBm, the AC peak to peak voltage with a **51** ohm load resistor connected to the 3 turn link should be 1.43 volts. Adjust RX to achieve this voltage in your oscilloscope.

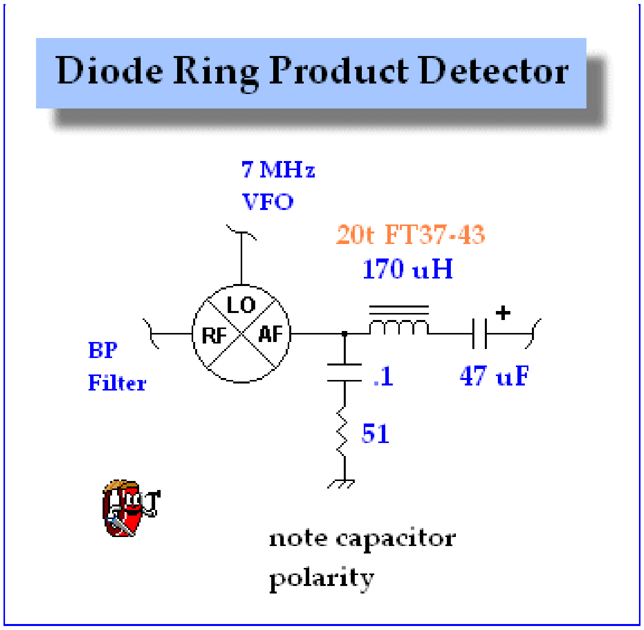
Diode Ring Product Detector
7 MHz
VFO
20t FT37-43
170 uH
LO
RF
AF
+
BP
Filter
.1
47 uF
51
note capacitor
polarity

Note that if you build a diode ring product detector receiver version using the simple W7EL low pass termination network, the polarity of the 47 uF electrolytic capacitor will need to be changed as shown.

Key goals of this website include providing ideas, basic support and encouragement. I am delighted when builders make their own stuff and not copy my circuits. This is why we homebrew; to create, explore and share. Mike, K7LR did all of these very well.

QRP SWL
HOMEBUILDER

Amateur and Short Wave Radio Electronics Experimenter's Web Site

---

Photography Links

Nikkor 50mm lenses

HENRI CARTIER-BRESSON - shot a lot with 50mm - video to watch over and over

The genius of photography -- BBC4 series about the history of photography

http://www.freestylephoto.biz

The Care and Feeding of Transmission Lines

Joel R. Hallas, W1ZR

A radio amateur's
guide to understanding
transmission lines

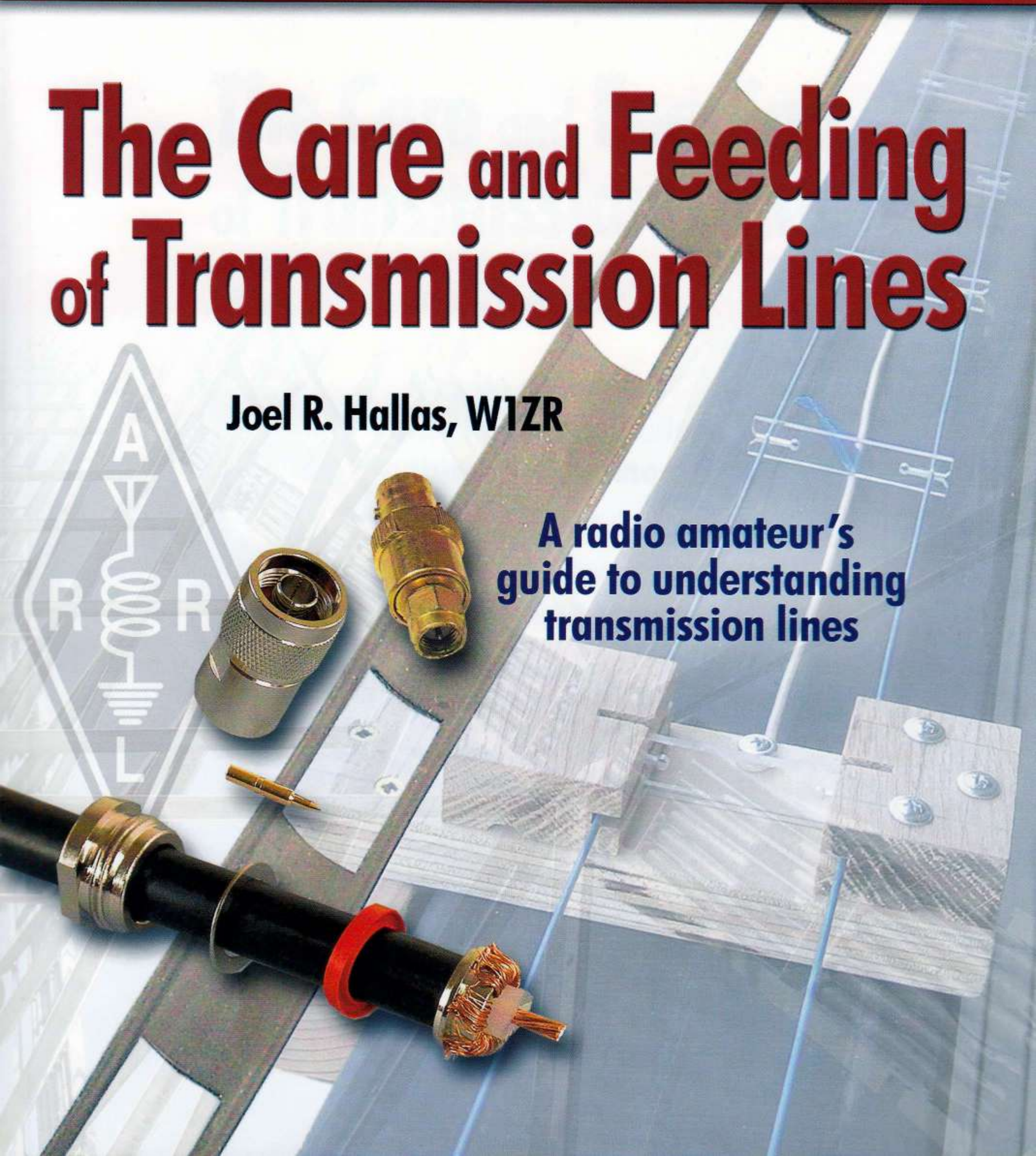

The Care and Feeding of Transmission Lines

Joel R. Hallas, W1ZR

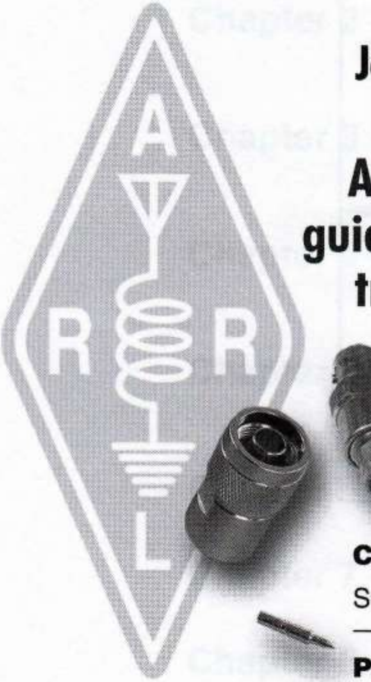

A radio amateur's guide to understanding transmission lines

Cover Design:
Sue Fagan, KB1OKW

Production:
Michelle Bloom, WB1ENT
Jodi Morin, KA1JPA
David Pingree, N1NAS
Carol Michaud, KB1QAW

Copyright © 2012 by

The American Radio Relay League, Inc.

Copyright secured under the Pan-American Convention

All rights reserved. No part of this work may be reproduced in any form except by written permission of the publisher. All rights of translation are reserved.

Printed in USA

Quedan reservados todos los derechos

ISBN: 978-0-87259-478-4

First Edition
First Printing

Contents

Foreword

Chapter 1 — What is a Transmission Line and Why Do We Need One?

Chapter 2 — What Are the Types of Transmission Lines?

Chapter 3 — Let's Examine Coaxial Transmission Line

Chapter 4 — Other Types of Unbalanced Transmission Line

Chapter 5 — Let's Examine Balanced Transmission Line

Chapter 6 — Transmission Line Connectors

Chapter 7 — Installing Coaxial Connectors on Cable

Chapter 8 — Determining Which Line is Best Suited for a Particular Application

Chapter 9 — Application and Installation Notes

Index

Foreword

While radio equipment and antennas get most of the amateur's attention, arguably neither can amount to much unless interconnected by a transmission line. Unless the appropriate line type is selected and properly installed, the amateur is likely to be dissatisfied with the results no matter how heavily invested he or she is in what happens at each end of the line.

This book is intended to introduce readers with a basic understanding of radio equipment and antennas to the ins and outs of transmission lines. Subjects covered include an introduction to the various types of line available, detailed discussions of the key parameters of both coaxial cable and balanced line types as well as the types of connectors and how to install them. With that knowledge, we then go into how to select the best line for each application, then how to install the line in a real environment and make sure it will still work at peak efficiency after it has been in place for some time.

As with all ARRL books, you can check for updates and errata, if any, at **www.arrl.org/notes/**.

Chapter 1

What is a Transmission Line and Why Do We Need One?

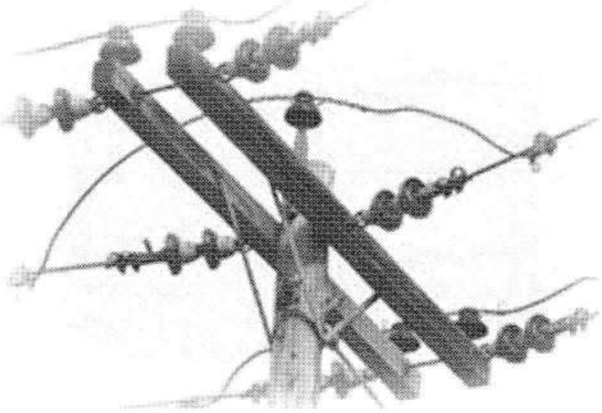

A transmission line is a kind of cable designed to convey power between physical locations. This kind of system is used for transmission of water, gas, sewage, electrical power, telephony signals and, in our case, radio frequency (RF) energy.

As with each of these applications, the desired results are the same: To convey the material or signal from one location to the other with minimal loss and with an output that includes all the characteristics of the input. While the ideal line is easy to imagine, real systems do not ever quite measure up.

In a pipe-based system, for example, material is lost through leaks in the pipe wall or at connectors. Friction of the fluid against the pipe wall results in a reduction of velocity and pressure over distance. Excessive pressure can result in a catastrophic rupture and loss of system functionality.

In a radio frequency transmission line, we have the same type of potential departures from an ideal transmission system, and some other issues that plumbers don't routinely worry about. The system functionality, however, is still very similar.

WHY DO WE NEED TRANSMISSION LINES?

We generally need a transmission line to carry RF between different parts of a radio system. **Figure 1.1** shows locations within a typical Amateur Radio station in which transmission lines would be appropriate. While not every amateur station includes all of these blocks, and some may have multiples of some blocks, the sketch is typical. Some larger pieces of equipment also have transmission lines within the unit for subassembly interconnections.

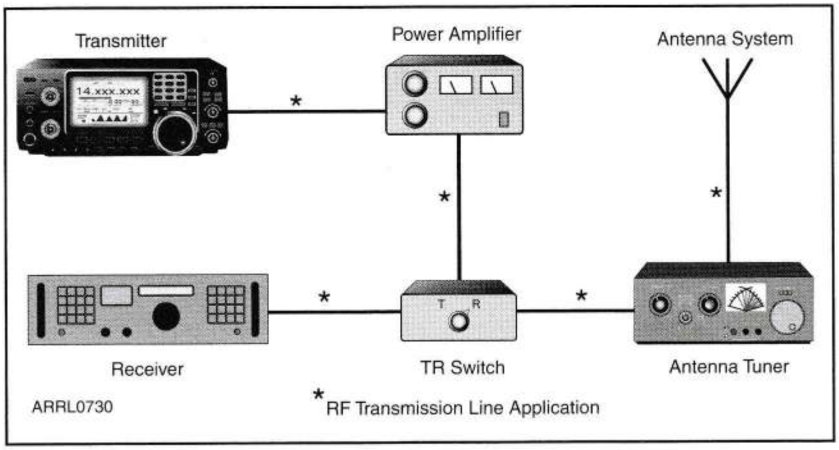

Figure 1.1 – Block diagram of a typical Amateur Radio station, showing places that would be appropriate applications for an RF transmission line.

While we may associate transmission lines with antenna connections, Figure 1.1 indicates that they have wider application. The short interequipment lines are sometimes referred to as *jumper cables*. Note that the dc power interconnections not shown between each piece of equipment and its power source could also be considered a kind of transmission line. Power transmission is a different topic beyond the scope of this book.

The reason transmission lines are used for these applications is that if other types of connection arrangement are used, we are likely to have radiation from the connecting wires. While the radiation does reduce the signal available at the end of the interconnection, it is usually more significant that the radiation will couple to other circuitry. This can result in a number of problems, including oscillation, transmitter lockup, and interference to reception of other systems such as computers.

BUT SOMETIMES WE DON'T NEED TRANSMISSION LINES

Actually, we don't always need a transmission line for a radio system. There are some cases in which all parts of a system are in a single location. In that case, no transmission lines are required. One example of such a system that I once worked with is shown in **Figure 1.2**. The problem was to accurately determine the reception pattern of a complex HF antenna array used as a part of a military over-the-horizon radar system. The solution was to monitor the receiver output as a helicopter rode in a circle around the array with a hanging center-fed dipole suspended below. A battery-powered low power transmitter was located at the center of the antenna and connected directly to each side of the antenna. A technician at the array center took azimuth data so we could tell the azimuth of the helicopter, while an onboard

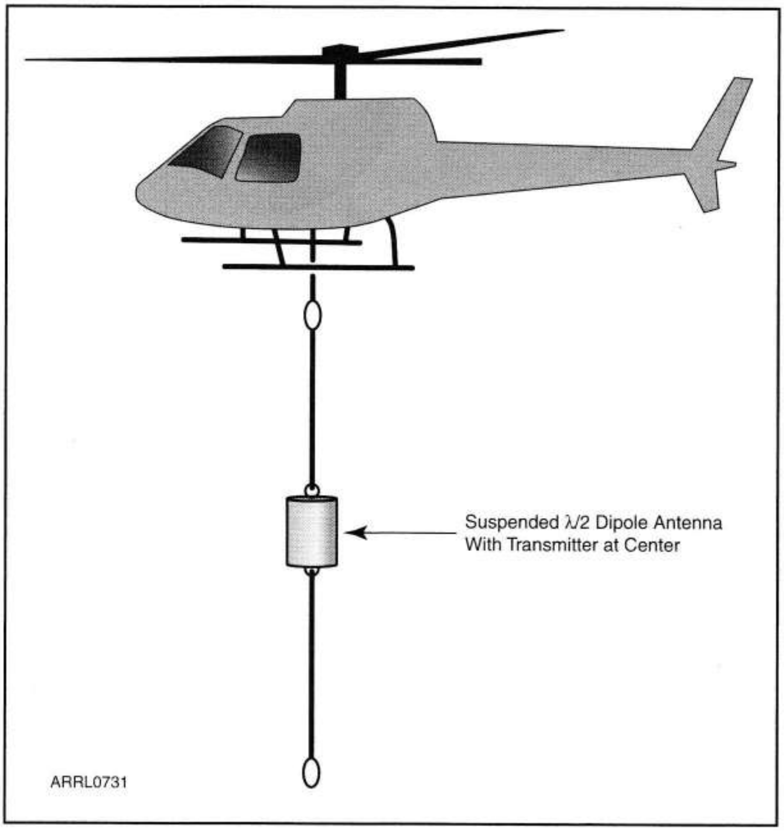

Figure 1.2 –
No RF transmission line required. This small battery-powered transmitter is located at the center of a directly fed dipole. The system was used to test the receive antenna pattern of a vertically polarized directive HF antenna array used in a military over-the-horizon radar.

navigation system was used to maintain a constant distance — worked like a charm!

Another more common application is found in some microwave systems. As we will discuss later, losses in transmission lines go up rapidly with frequency. Radio systems at microwave frequencies are usually designed to minimize transmission line length. Some have the actual microwave portion of the radio system located within the antenna structure. This is common in satellite TV receiver systems, as well as in microwave communication and radar systems. It is often more economical to have the interunit cabling carry the dc and intermediate frequency signals (on transmission lines, but with less loss because the signals are at lower frequencies) than to suffer the transmission line losses at microwave frequencies. **Figure 1.3** illustrates a typical configuration.

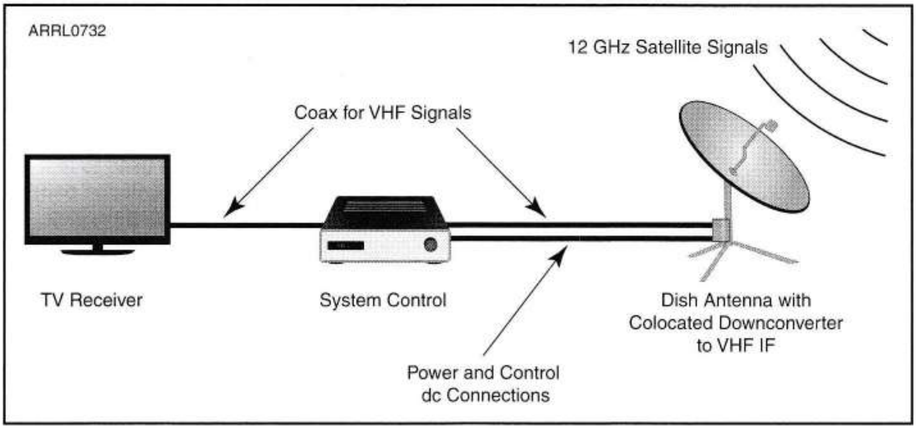
ARRL0732
12 GHz Satellite Signals
Coax for VHF Signals
TV Receiver
System Control
Dish Antenna with
Colocated Downconverter
to VHF IF
Power and Control
dc Connections

Figure 1.3 – In this simplified satellite TV receiver system, the microwave components are located directly at the antenna feed to avoid losses. A downconverter translates the desired channel information from the 12 GHz broadcast range down to a TV VHF channel, at which losses in the transmission line are more manageable.

A similar arrangement is used in many military radar systems. The rotating antenna with transmitter power amplifier and receive preamplifier can be located in one unstaffed vehicle while the signal processing equipment and sometimes the display systems are located in another at some distance. In this way, the operators and signal processing equipment are not part of the target for an antiradiation missile, especially popular as a counter to air defense radars.

Chapter 2

What Are the Types of Transmission Lines?

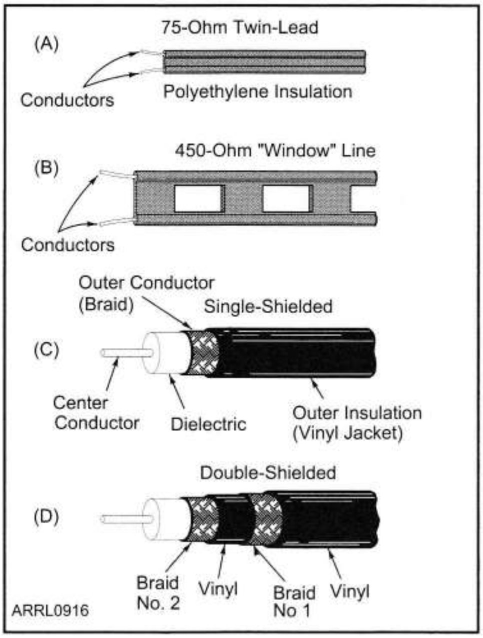
(A)
75-Ohm Twin-Lead
Conductors
Polyethylene Insulation
(B)
450-Ohm "Window" Line
Conductors
Outer Conductor (Braid)
Single-Shielded
(C)
Center Conductor
Dielectric
Outer Insulation (Vinyl Jacket)
Double-Shielded
(D)
Braid No. 2
Vinyl
Braid No 1
Vinyl
ARRL0916

Examples of popular transmission lines.

Transmission lines come in multiple flavors, depending on the application. Here we will broadly categorize lines into their major types. We will also introduce some of the varieties within each category that we will discuss in greater detail in later chapters.

UNBALANCED TRANSMISSION LINE

Transmission lines in which the conductors are at different potentials with respect to ground are referred to as unbalanced transmission lines.

Coaxial Cable

Probably the most frequently encountered type of RF transmission line is coaxial cable. It is found connecting almost every television set to an antenna or to a cable service provider. The cable that cable service providers use is also coaxial cable, at least in the last part of the system, near the subscriber. Coaxial cable is one member of the class of unbalanced transmission lines.

Coaxial cable consists of two conductors, one surrounding the other in a concentric configuration as shown in **Figure 2.1**. In operation, if properly terminated and used, the RF field is contained within the space marked "dielectric" (an insulating material) between the two conductors. At the ends, the outer conductor, also called the *shield*, can be at ground potential. The inner conductor is at a potential above ground, thus the name "unbalanced."

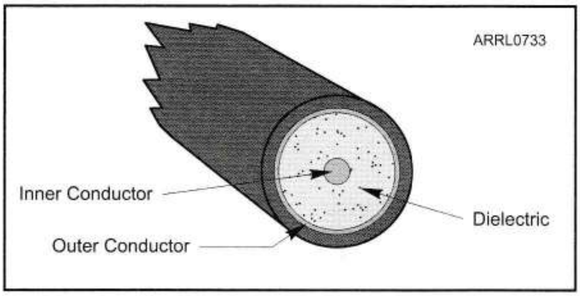

Figure 2.1 – The configuration of coaxial cable. The inner conductor is often wire, sometimes stranded for additional flexibility. Rigid cable may have a center conductor of solid wire or sometimes copper pipe. The outer conductor can be braided copper, wrapped tinfoil, or for rigid cable, copper or aluminum tubing. The lowest-loss rigid coax is made of copper tubing for both conductors with a dielectric mostly of air. Occasional ceramic spacers are employed to maintain the spacing. Other cables have a polyethylene dielectric, or sometimes a foamed poly that acts like something in between.

At RF frequencies, the phenomenon known as *skin effect* keeps the RF shield currents on the inside of the outer conductor. The outside of the outer conductor then acts almost as if it were a separate third conductor at ground potential along its length. This, along with the fact that the net field is within the cable, provides the effect of shielding the signals within the cable from the effect of any signals or other effects outside the cable. This is a major advantage of coaxial cable.

Within the class of coaxial cable ("coax" to its friends), there are a variety of subdistinctions:

◆ Characteristic impedance — The ratio of the conductor diameters and the properties of the dielectric determine the characteristic impedance, a topic that will get a chapter of its own.

◆ Cable diameter — The diameters together may scale up and down from less than ⅛ inch to many inches. The larger sizes are needed for high-power applications, and also have less loss than smaller-sized cable.

◆ Cable flexibility — There are both rigid and flexible coaxial cables, and some semi-rigid, or perhaps semi-flexible, a major factor in selection based on application. The rigid, and some semi-rigid, cables are usually made with a solid tube, rather than a woven outer conductor structure. Solid outer conductors result in coax with the potential for shielding properties closer to ideal than those with woven shields.

◆ Cable dielectric material — The properties of the material between the conductors, the dielectric, is the major factor in the cable loss at higher frequencies and also a factor in determining the cable characteristic impedance.

Unbalanced Microstrip Transmission Line

Imagine slitting the outer conductor along a length of coaxial transmission line and unfolding it so it lies flat. This is the essential configuration of microstrip transmission line. Its major application is found in intraequip-

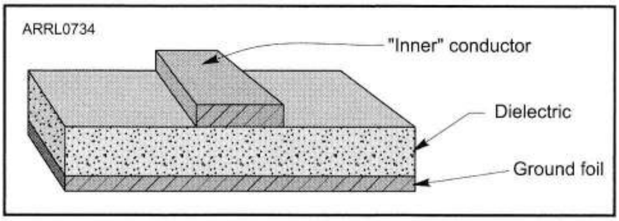

Figure 2.2 – Unbalanced microstrip line. A conductive microstrip left on one side of a piece of two-sided printed circuit board can be operated as an unbalanced transmission line with reference to the foil on the other side. While it acts a lot like coax in some ways, it does not share its shielding properties. In some implementations, a second board or layer of a multilayer board makes a sandwich with additional shielding. In all cases, the connections must be made close to the beginning and end of the conductors. With the sandwich configuration, the ground connection must be made to both ground planes near the location of the start and end of the conductive path.

ment wiring of assemblies made with double-sided printed circuit board. The line can be formed by etching away some of the conductor on one side of the board (see **Figure 2.2**) as part of the process of generating the rest of the wiring on the board.

While microstrip line is efficient from a manufacturing standpoint, it does not offer the shielding qualities of coaxial cable. The fields between the conductors will extend outward from the space between the conductors, requiring care in locating wiring for other parts of the circuit to avoid coupling between circuits. Of course, if coupling is a problem, small coaxial cable is well suited to PC board use, if not quite as easy to implement.

Single Wire Line

Before the days of radio, a single wire was used as a transmission medium for wire-line telegraph systems. The wire was fed against an earth/ground connection and was useful but very lossy. It is not particularly suited for RF systems since it will act more like an antenna, unless it is very close to a highly conductive ground. Think virtual microstrip line.

BALANCED LINE

A transmission line with two parallel conductors can be used as a balanced transmission system — meaning that each conductor is at the same magnitude of potential compared to ground but of opposite sign. This means that at a particular instant, if one wire is at +5 V, for example, the other will be at –5 V, with respect to some common level or ground reference.

Open-Wire Line

The earliest balanced transmission line was a pair of wires at a constant spacing insulated from ground (see chapter photo and **Figure 2.3**). Before the days of radio, and for some time thereafter, such a pair was used for each telephone circuit, making for very imposing utility poles with many cross Ts, almost blocking out the sun in large cities. Current telephone

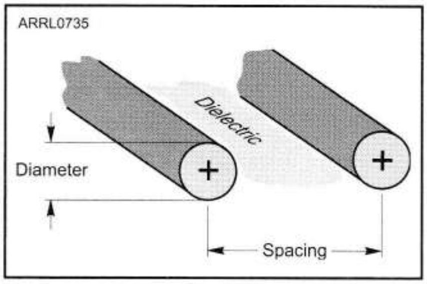

Figure 2.3 – Balanced transmission line. Just two wires, spaced a fixed distance and separated by a dielectric — air, or some other insulating material.

wiring uses insulated twisted pairs, bundled with others in large cables supporting multiple end-user customers. Once individual circuits reach a switching center, they are concentrated with switches and routed between centers on high-density fiber optic links.

Very early radio was conducted on low frequencies typically using large top-loaded monopole antennas fed directly from the transmitter.[1] As operating frequencies became higher, horizontal antennas could be high enough to be useful, and balanced antenna structures, such as a center-fed dipole, became popular. A natural way to interconnect the typical balanced output of the transmitter's tuned transformer to the center of a balanced antenna was to use balanced open-wire line (see **Figure 2.4**).

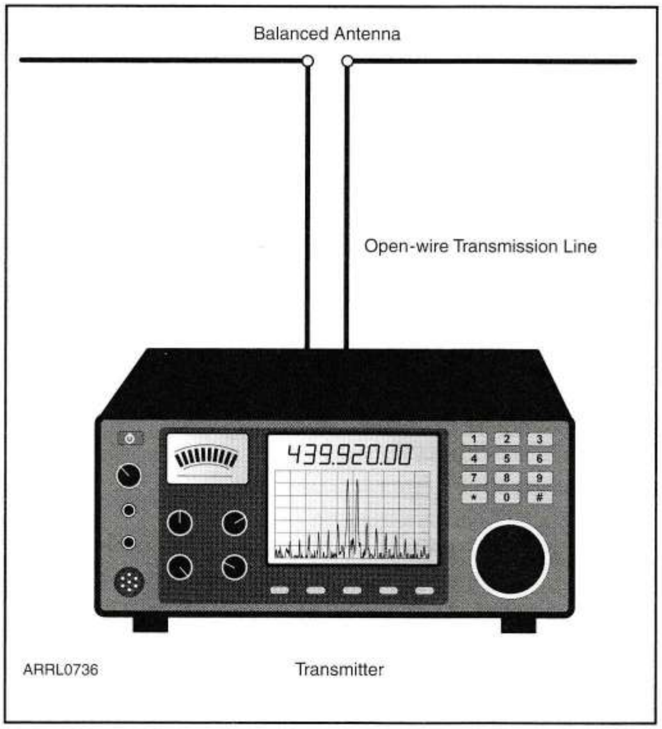

Figure 2.4 – An open-wire transmission line is a natural for interconnecting a balanced output transmitter and a balanced antenna.

The wires of open-wire line require some method to hold them at a relatively constant distance. In early telephone systems, this was accomplished by tying each wire to a glass insulator held to a cross T by a peg. In recent rural travel, I saw quite a few of these still in use on poles along railroad right-of-way, likely for signaling. Early radio transmission lines used paraffin-soaked wood dowels to position the wires. These gave way to glazed molded ceramic insulators made for the purpose. Current amateur practice is to use PVC plumbing pipe or acrylic plastic with slits to hold the wires (see **Figure 2.5**).[2]

Twinlead

After World War II, television started to become popular in the US. The signals were transmitted over the air using a portion of the VHF range. All antenna designs of the time were of balanced feed and used easily installed transmission line developed for TV applications.

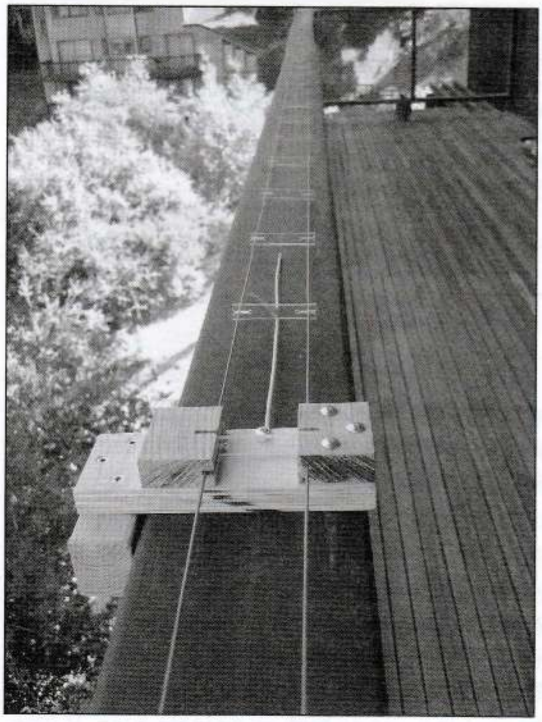

Figure 2.5 – The production line is in full swing as Barry Shackleford, W6YE, fabricates his own open-wire line. [BARRY SHACKLEFORD, W6YE]

The answer was called twinlead. It was a balanced transmission line made from a pair of stranded wires held about half an inch apart and enclosed in a polyethylene web that insulated each wire and filled the space in between with a flexible insulating material (as shown at the top of the chapter photo). This was the line of choice for this application, as well as some amateur antenna systems until it was displaced in TV service by coaxial cable.

Window Line

This derivative of twinlead is made for higher power and is often used in Amateur Radio transmitting applications. Window line is constructed similarly to twinlead, but has windows cut in the webbing along its length. The windows make the line act more like open-wire line in terms of attenuation, without having the complication of discrete insulators.

The wider spacing typically results in a higher characteristic impedance than regular twinlead.

Twisted Pair

A pair of wires twisted together can be used as a makeshift balanced transmission line. The characteristic impedance will be close to 100 Ω. This type of line is not often used for RF applications, although it can serve in a pinch. It is frequently encountered in telephone and local area network (LAN) wiring.

Shielded Twisted Pair

By enclosing a twisted pair in a shield, we can get some of the benefits of balanced line, along with the shielding effect of coax. Of course, we get the disadvantages of each as well, primarily the high attenuation of coax as compared to most balanced lines. The trade name for this cable is Twinax, and it has most recently been used in computer networking, but was once a choice for RF applications as well.

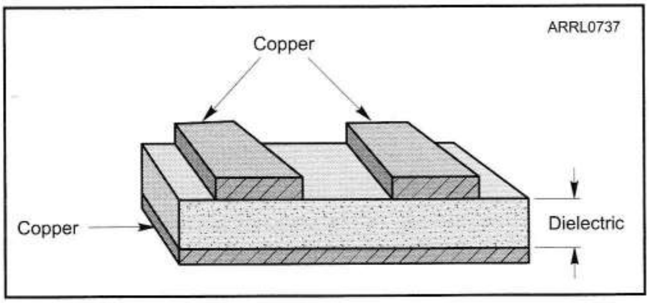

Figure 2.6 – Balanced microstrip line. This works the same way as unbalanced microstrip line, but the two conductors on one side act as a balanced pair.

Balanced Microstrip Line

A pair of etched conductors on PC board material can serve as a balanced line within equipment in the same way that microstrip can provide unbalanced connectivity (see **Figure 2.6**).

Parallel Coax Lines

Two similar coaxial cables of the same length can be operated as a balanced line if the balanced signals are connected to the inner conductors and the shields are tied together at both ends (see **Figure 2.7**). This is not a good choice for many applications of balanced line, since the attenuation is the same as for coax and the cost is twice as high as a single coax. It does work well for balanced line installations in which a small section of the line must be on the ground or in some environment that is unsuitable for open-wire line. Because the fields are entirely within the coax, the two lengths can be run separately as long as they are the same length and the shields at each end are connected together with short connections.

GUIDED WAVE STRUCTURES

UHF and especially microwave (UHF above 1000 MHz) signals are sometimes transported using a kind of mechanism different from a transmission line. The signal loss associated with coaxial cable at UHF and higher frequencies makes them unattractive for all but very short connection paths. A better alternative is to transmit the signals as electromagnetic waves propagating in space. An antenna could make such a transition, but suffers because the signal will spread spatially as it propagates. Instead, the waves are guided by propagating them within a different kind of medium.

The structures described below are a substitute for transmission lines. They are not frequently encountered in Amateur Radio, except when used by some microwave specialists. Beyond the short description below, they will not be covered in this book.

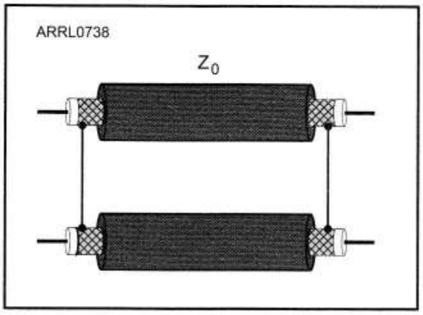

Figure 2.7 – Pair of coaxial cables used as a balanced transmission line. This is particularly useful as a short section of a balanced run to transit an area that is not appropriate for the fields surrounding open-wire line. The attenuation is as high as coax, so short sections are key. It is important that the shields are interconnected at both ends as shown. Since the shields are at reference potential in a balanced system, either may also be connected to structural ground, if appropriate.

Waveguide

The most common guided wave structure is a hollow metal tube called a waveguide. The waves are launched into the tube by an antenna-like probe, propagate within the tube, and are extracted by another probe at the far end. In order to carry the waves without significant attenuation, the guide needs to have a width of at least ½ wavelength and is often of rectangular cross-section. The size restriction limits the usefulness of the technology to the higher UHF frequencies, although I have seen waveguide in use in military radar installations at frequencies as low as 450 MHz — the waveguide tubes looked like very large heating ducts.

Single Wire Transmission Line

It is also possible to guide microwave energy along a wire. This method was originally called Goubau line (after its inventor), or G-line for short. The wire was surrounded by a special dielectric with the appropriate properties for slowing the wave in order to keep it traveling along the wire, rather than radiating as if it were an antenna. Special conical launchers are used at each end of the line, to and from coaxial transmission line. This type of line is easier to deploy in the field than waveguide since it can be rolled up, but it must be used in straight runs. G-line was part of a military microwave communications system that I encountered in the 1960s.

It was later found that waves of a different orientation could be launched onto an uninsulated wire and transmitted with low attenuation using a different configuration of launcher. This method is called E-line.

Notes

[1] J. Hallas, W1ZR, *Basic Antennas*. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9994. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

[2] B. Shackleford, W6YE, "Custom Open-wire Line — It's a Snap," *QST*, Jul 2011, pp 33-36.

Chapter 3

Let's Examine Coaxial Transmission Line

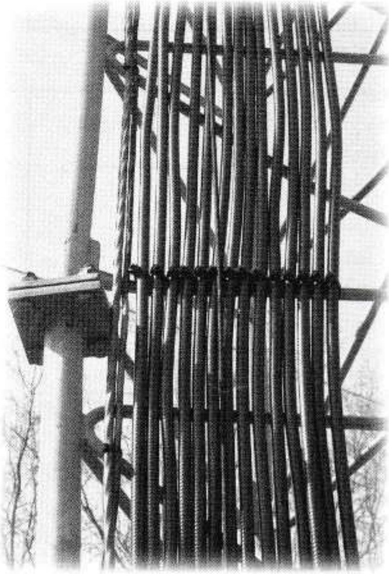

Coaxial cables going up one of the towers at W1AW, the ARRL Headquarters station. An advantage of coax is that the runs can be in close proximity to each other without interference.

Coax cable is the most commonly encountered type of transmission line. Coax is composed of two concentric conductors separated by a dielectric material (see **Figure 3.1**). While some types have an exposed outer conductor, the commonly encountered flexible cables and some rigid types are covered by a protective insulating jacket. Coax is found in many sizes for different applications, with the most common ranging from about 0.075 inches to a bit less than half an inch in diameter.

While individual conductors are occasionally connected directly to circuitry, most often within equipment, in most applications the cables are fitted with one of a number of special connectors on their ends. There are different series of connectors, some appropriate for higher frequencies, some waterproof, and some selected just because of the equipment to which they must connect.

EQUIVALENT CIRCUIT OF A COAXIAL TRANSMISSION LINE

Any two conductors separated by a dielectric will have electrical capacitance between the conductors. The capacitance is a function of the surface area of the conductors, the separation between them, and the dielectric constant of the insulation material between them. Coax cable has these properties, with capacitance created by the insulation between the inner surface of the outside conductor and the outer surface of the inside conductor. Any dielectric between the conductors will increase the capacitance compared

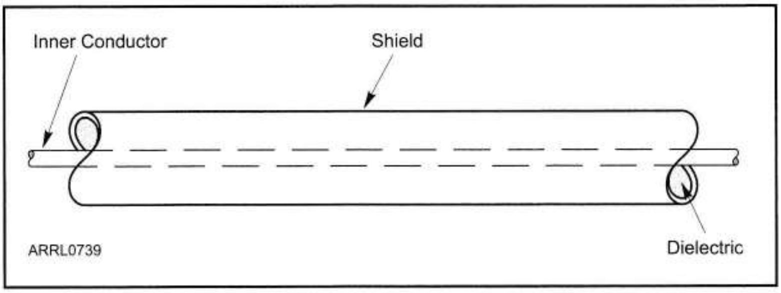

Figure 3.1 – Sketch of a typical coaxial transmission line.

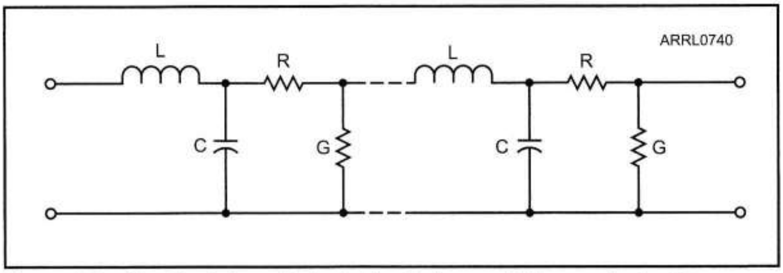

Figure 3.2 – Electrical dc equivalent circuit of a coaxial transmission line. At higher frequencies the series resistance increases because of skin effect. The length of each section can be changed by scaling the values. For many analysis applications, having 10 sections per wavelength will provide reasonable accuracy, thus the one-foot sections might be usable up to around 100 MHz, if the resistance were corrected for skin effect.

to air-insulated cable. The total area of each conductor is a function of the cable length, so the capacitance is usually specified in terms of capacitance per unit length, notated in this country in pF/ft. For example, the usual 50 Ω coax cable (which will be discussed shortly) might have a capacitance of 29 pF/ft.

Similarly, any piece of wire will have a certain inductance, and that includes the conductors making up a coaxial cable. Again, since the inductance is also a function of the cable length, it too is usually expressed in terms of inductance per unit length. For typical lines as above, it will be about 72.5 nH/ft.

The conductors will also have electrical resistance. Because of their relative sizes, this will predominantly be the resistance of the inner conductor. At very low frequencies, the resistance will be around the dc resistance of the inner conductor, perhaps 1.8 mΩ/ft depending on conductor size. There will also be some loss in the dielectric, increasing with frequency. This is represented as the conductance G. We thus could model a transmission line as an electrical circuit such as that shown in **Figure 3.2**.

CHARACTERISTIC IMPEDANCE

The characteristic impedance (Z_0) of a transmission line is one of its most significant parameters. If we consider the line model shown in Figure 3.2 and imagine a dc voltage connected to one end, a current will start to flow, regardless of what is happening at the other end. If the line were just capacitive, the initial current would be very high, as it tried to charge the capacitance. If it were entirely inductive, the initial current would be zero because of the back EMF. Because it is a combination, the initial current will be in between the two values, also limited by the resistance, and not usually a significant portion of the effect. That current will continue until the leading edge of the current pulse gets to the far end and returns with an indication of the effect of the load. The current pulse will travel at somewhat less than the speed of light in free space.

The amount of the current will be the same as if the source were connected to an impedance — generally close to being a resistor — during the initial period. The value of that resistor is called the *characteristic impedance* of the line, usually referred to as its Z_0. If the far end of the line were connected to an actual resistor equal to the Z_0, the same current would continue to flow until the source was disconnected, just as though it were an infinite length of line.

If we use the transmission line for the carrying of an RF signal rather than dc, the same kind of thing happens. For example, using Ohm's law, if we put a 1 V_{RMS} RF signal on one end of a line with a Z_0 of 50 Ω, the initial current would be 1/50 = 0.02 A or 20 mA. If the line were lossless (the R in the model = 0) and terminated at the far end with a 50 Ω resistor, the same 20 mA would flow into the resistor. This is known as a *matched* condition. If the termination were different than a 50 Ω resistor, we could no longer still have 1 V_{RMS} and 20 mA into the load, because it would violate Ohm's law for a different value resistor. Some fraction of the signal would be reflected from the end so that the combination of the forward and reflected signals would result in a voltage and current at the load that satisfies Ohm's law.

The sum of the forward and reflected wave is referred to as a *standing wave*, because the amplitude at any point along the line stays constant, just as it has to stay constant at the termination. This means that the voltage and current along the line will vary, repeating every half wavelength until it is back to the source (as shown in **Figure 3.3**). The impedance at the source will not be the value of the termination nor the line Z_0, but a function of the termination impedance and the distance from it (in wavelengths).

The ratio of maximum voltage (or current) to the minimum voltage (or current) on the line is called the standing wave ratio (SWR), a major consideration in most systems. By having a termination equal to the line

I Know What's Happening at the Shack — What's Happening at the Other End of my Feed Line?

If you want to find out — here's the easy way, using TLW.

Joel R. Hallas, W1ZR
***Technical Editor,* QST**

I'm told that one of the more frequent questions received by *QST*'s "Doctor" has to do with folks wanting to determine the impact of transmission line losses on the effectiveness of their antenna system. These questions are often along the lines of, "I measure an SWR of 2.5:1 at the transmitter end of 135 feet of RG-8X coaxial cable. My transceiver's auto-tuner can tune it to 1.1, but how can I tell what my losses are?" or "How much difference will I have if I have a tuner at the antenna instead of using the built-in tuner?"

These are important questions that almost every amateur operator is faced with from time to time. An approximate answer can be obtained by using the graphs found in any recent edition of *The ARRL Antenna Book* showing the loss characteristics of many transmission line types, plus adding in the effect of an SWR greater than 1:1. The SWR at the antenna end can be determined from the bottom end SWR and the cable loss. Using these graphs requires a bit of interpolation or Kentucky windage, but can result in useful data.

But There's an Even Better Way!

Packaged with each of the last few editions of *The ARRL Antenna Book* is a CD containing the pages of the whole *Antenna Book* as well as some very useful software. The program that I use almost daily is one written by *Antenna Book* Editor R. Dean Straw, N6BV, called *TLW*, for *Transmission Line for Windows*.

TLW provides a very easy-to-operate mechanism to determine everything I usually need to know about what's happening on a transmission line. When you open the program, you are presented with a screen as shown in **Figure A**. This has the values plugged in from the last time you used it, often saving a step. Let's take a quick tour of the inputs:

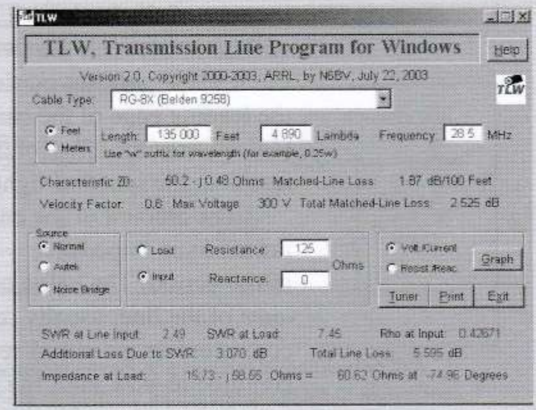

TLW, Transmission Line Program for Windows

Version 2.0, Copyright 2000-2003, ARRL, by N6BV, July 22, 2003

Cable Type: RG-8X (Belden 9258)

● Feet ○ Meters

Length: 135.000 Feet 4.890 Lambda Frequency: 28.5 MHz

Use "w" suffix for wavelength (for example, 0.25w)

Characteristic Z0: 50.2 - j 0.48 Ohms Matched-Line Loss: 1.87 dB/100 Feet

Velocity Factor: 0.8 Max Voltage: 300 V Total Matched-Line Loss: 2.525 dB

Source: ● Normal ○ Autek ○ Noise Bridge

○ Load ● Input Resistance: 125 Reactance: 0 Ohms

● Volt/Current ○ Resist/React

SWR at Line Input: 2.49 SWR at Load: 7.45 Rho at Input: 0.42671

Additional Loss Due to SWR: 3.070 dB Total Line Loss: 5.595 dB

Impedance at Load: 15.73 - j 58.55 Ohms = 60.62 Ohms at -74.96 Degrees

Figure A — The opening screen of *TLW*, illustrating the process described in the article.

Cable Type — This allows you to select the cable you would like to analyze. A drop-down box provides for the selection of one of 32 of the most common types of coax and balanced lines. An additional entry is provided for User Defined Transmission Lines that can be specified by propagation velocity and attenuation.

Length — In feet or meters, your choice.

Frequency — This is an important parameter when dealing with transmission line effects.

Source — This defines the form of the input impedance data. Generally, you can use NORMAL.

Impedance — The impedance can be specified as what you measure, resistive (real) and reactive (imaginary, minus means capacitive). This could come from your antenna analyzer at either end of the transmission line. Note, if you only know the SWR, not the actual impedance, all is not lost — see below.

Now for the Outputs

SWR — The SWR is provided at each end of the cable. This is an important difference that many people miss, important even with a moderate SWR at the transmitter end, as we'll see — the SWR at the antenna will be much higher due to the cable loss. With *TLW*, you instantly know the SWR at both ends, and the loss in the cable itself.

Rho at Load — This is the reflection coefficient, the fraction of the power reflected back from the load.

Additional Loss Due to SWR — This is one of the answers we were after.

Total Loss — And this is the other, the total loss in the line, including that caused by the mismatch.

But Doctor, What if I can Only Measure the SWR — Not the Actual Impedance?

Often the only measurement data available is the SWR at the transmitter end of the cable. Because the losses are a function of the SWR, not the particular impedance, you can just put in an arbitrary impedance with that same SWR and click the INPUT button. An easy arbitrary impedance to use is just the SWR times the Z0 of the cable, usually 50 Ω. For example, you could use a resistance of 125 Ω to represent an SWR of 2.5:1. This is what we've done in Figure A, using 135 feet of popular Belden RG-8X.

The results are interesting. Note that the 2.5:1 SWR as seen at the radio on 28.5 MHz results from a 7.45:1 SWR at the antenna — perhaps this is an eye-opener! Note that of the 5.6 dB loss, more than half, or 3.1 dB, is due to the mismatch. Note that if we used something other the actual measured impedance, we can't make use of the impedance data that *TLW* provides. We can use the SWR and loss data, however, but that's probably what we wanted to find out.

We can now do some "what ifs." We can see how much loss we have on other bands by just changing the frequency. For example, on 80 meters, with the same 2.5:1 at the transmitter end, the SWR at the antenna is about 3:1 and the loss is slightly more than 1 dB. We could also plug in an impedance calculated at the antenna end and see what difference other cable types would make. For example, with the same 28.5 MHz SWR of 7.45 at the antenna and 135 feet of ½ inch Andrew Heliax, we will have a total loss of 1.5 dB at 28.5 MHz. Note that the SWR seen at the bottom will now be 5.5:1 and our radio's auto-tuner might not be able to match the new load.

But Wait, There's More!

You can also click the GRAPH button and get a plot of either voltage and current or resistance and reactance along the cable. Note that these will only be useful if we have started with actual impedance, rather than SWR.

Pushing the TUNER button results in a page asking you to select some specifications for your tuner parts. *TLW* effectively designs a tuner of the type you asked for at the shack end of the cable. It also calculates the power lost in the tuner and gives a summary of the transmitted and lost power in watts, so you don't need to calculate it!

When you've finished, be sure to hit the EXIT button; don't just close the window. Otherwise *TLW* may not start properly the next time you want to use it.

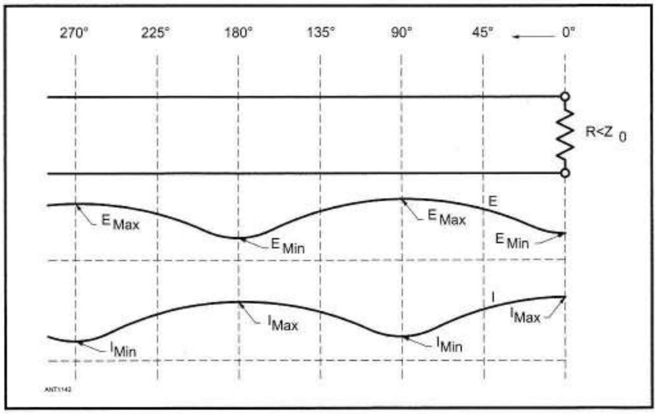

Figure 3.3 – The voltage and current along a mismatched line will vary, repeating every half wavelength.

Z_0, the standing wave doesn't exist and the SWR is 1:1. For a number of reasons, this is a desirable condition and cable with different values of Z_0 are available, including most commonly 35 Ω, 50 Ω, 75 Ω, and 93 Ω.

FIELDS WITHIN COAX

We need to briefly consider both the electric and magnetic fields in coax to understand how and why it works as it does. The electric field, just as with the plates of any capacitor with an applied voltage, exists between the outside of the center conductor and the inside of the outer conductor (see **Figure 3.4**). Because the outer conductor surrounds the inner conductor, all the electric field terminates on the inside of the shield and does not couple to, nor is it affected by, nearby conductors. Thus, the outside of the cable is effectively shielded from the effects of the electric field within the cable — one of the major benefits of coax.

As in any conductor, a magnetic field surrounds the two current-carrying conductors. For properly operating coaxial cable, the two currents will be equal and opposite; thus, the fields will cancel in the region outside the cable (see **Figure 3.5**), so the net magnetic field surrounding the cable will be zero. This, combined with the shielding of the electric field, results in a transmission line that is independent of the electrical conditions of its surroundings. This means it can be coiled, bundled with other cables, or run within metal conduit without any problems — as long as the currents are balanced and the shield is solid.

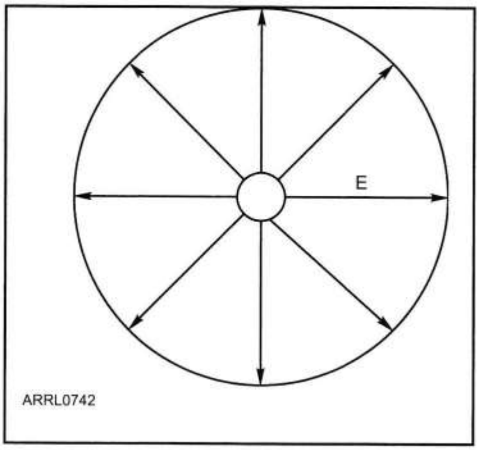

Figure 3.4 – The electric field within a coaxial transmission line. Note that if the shield is solid, it exists only within the cable, resulting in a part of the shielding effect.

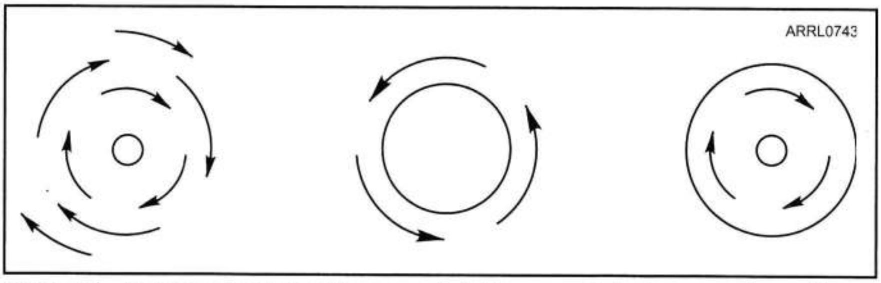

Figure 3.5 – As in any conductor, the magnetic field surrounds the two current-carrying conductors. On the left, the magnetic field surrounding a single isolated conductor is shown. The center image shows the magnetic field surrounding the outer conductor, and at right is shown the magnetic field surrounding the inner conductor. For properly operating coaxial cable, the two currents will be equal and opposite, as shown, and thus the fields will cancel outside the cable.

SHIELDING LIMITATIONS OF REAL COAX

Coax in the real world can be quite close to the ideal in most respects, but there are some limitations. The electric field shielding effectiveness is a function of the shield coverage. The best shielded cables are those with a solid tubular outer conductor. These can be considered a 100% shield, although loose or improperly installed connectors can be leakage points.

Unfortunately, solid tubular coax is not applicable to the many applications that require flexible cable. Probably the next best type, in terms of shielding effectiveness, is the type used by the cable television industry. It usually has multiple shields, generally one of wrapped aluminum foil and another of braided tinned copper. This is also quite effective — and has to be, otherwise cable signals, which occupy most of the radio spectrum, would interfere with other users, including such critical systems as aircraft communications and navigation.

The rest of the coax world uses braided copper shielding. This can be adequate for most applications, although the quality varies considerably among brands and types. In many cases the manufacturer will specify the shielding coverage, usually in terms of a percentage, with 100% meaning fully covered. Some cable types have double shields — two independent braided shields, one around the other, in electrical contact over their length. These can be very close to 100% effective. Less expensive cables can have effectiveness as low as in the 80% range. The less completely shielded cables allow a partial leakage of the electric field, allowing some coupling to and from nearby circuits.

Another coupling mechanism occurs if the shield current is different from the current on the inner conductor. This can occur in a number of ways, typically due to a termination in which the outside of the shield acts like a separate conductor from the inside of the shield. Currents on the outside can also be induced by proximity to an antenna system.

TYPES OF COAXIAL CABLE

Coaxial cable came into common use during World War II, mainly in military radar and radio applications. A set of standards was developed during the period for radio guide (RG) cable, with that designation followed by a series number. The most common in radio use over the years have been 50 Ω characteristic Z_0 RG-8 and RG-58. Each has a copper center conductor, a polyethylene dielectric, a braided copper shield, and an outer jacket, originally made from black vinyl but more often of PVC or polyethylene today. The primary difference between the two is size, with the RG-8 having an outside diameter of 0.405 and RG-58 having one of 0.195 inches. The smaller cable has higher attenuation and lower voltage and power ratings, but offers lighter weight, additional flexibility, and handling convenience.

The next most frequently encountered early cables were those with 75 Ω characteristic impedance, most commonly RG-11 at 0.405 inches and RG-59 at 0.242 inches. A suffix "U" indicates cable that is universal (usable for multiple functions), while other suffixes indicate special characteristics or construction differences. Probably the most common coax in use today is RG-6 type (see below), an RG-59-size 75 Ω cable with a more effective foil shield, used for in-house connection of cable television equipment.

These characteristics were documented in a US military procurement manual that was cancelled in 2001, so the early numbers are no longer subject to official standardization. Thus cable that is called "RG-8U type" is a manufacturer's designation that may mean the cable is somewhat like RG-8, but does not guarantee that it meets either early or current military standards. It may have a less dense shield or lower quality jacket, for example. Later standards include MIL-DTL-17H from August 19, 2005, which superseded MIL-C-17G of March 9, 1990. These standards refer to newer standard cable series that have replaced the earlier types. For example, today one would obtain RG-223 instead of RG-58, and RG-213 instead of RG-8. As in the earlier series, the word "type" is a manufacturer's escape clause that allows them to skirt military specifications.

Table 3.1 is a listing of many of the types of coaxial transmission line, as well as other types that we will discuss later. Note that within a type of cable, there are multiple cables from different manufacturers, with quite different characteristics. Some have a foamed polyethylene dielectric (FPE) instead of solid polyethylene. FPE is partly air and thus has a lower relative dielectric constant and less loss than standard cable. The foam also makes for a less sturdy cable that is more susceptible to water migration and cannot be flexed as tightly as the solid dielectric coax. The construction of a number of coaxial cable types is shown in **Figure 3.6**.

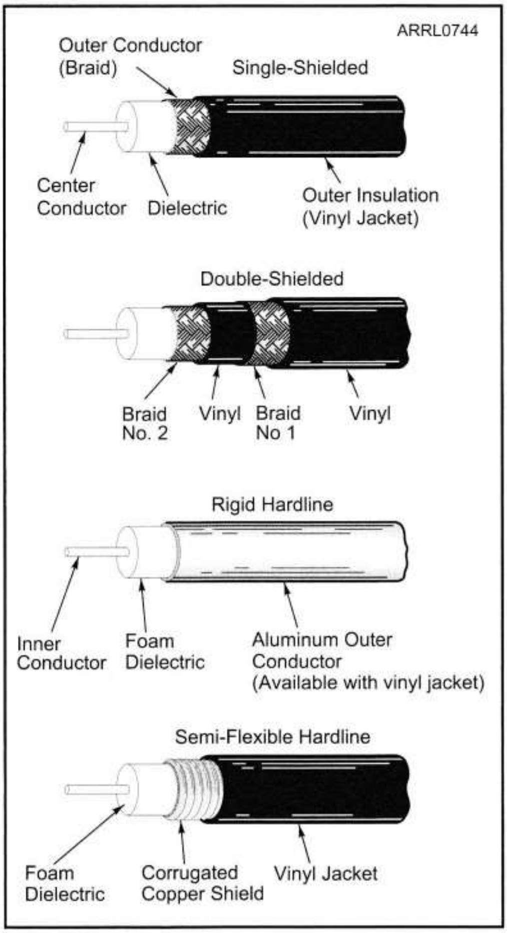
ARRL0744
Outer Conductor (Braid)
Single-Shielded
Center Conductor
Dielectric
Outer Insulation (Vinyl Jacket)
Double-Shielded
Braid No. 2
Vinyl
Braid No 1
Vinyl
Rigid Hardline
Inner Conductor
Foam Dielectric
Aluminum Outer Conductor (Available with vinyl jacket)
Semi-Flexible Hardline
Foam Dielectric
Corrugated Copper Shield
Vinyl Jacket

Figure 3.6 – Construction of a number of common types of coaxial cable.

KEY CABLE PARAMETERS

The headings in Table 3.1 give insight into the differences between cable types. Each is described briefly below.

Nominal Z_0 (Ω) — As was discussed earlier, Z_0 is a primary transmission line parameter. Most systems specify the design impedance for various connections, and typically that will align with a standard cable Z_0. The usual practice is to use a cable with that Z_0 to connect to equipment to a load with the same impedance, although we will discuss exceptions. In most cases, particularly at lower frequencies, the difference between a Z_0 of 50, 51, 52, or 53 Ω, as shown in the RG-58 types, will not be significant. For coaxial cable, the Z_0 can be determined by the dimensions as follows:

Table 3.1

Nominal Characteristics of Commonly Used Transmission Lines

G or	Part	Nom. Z_0	VF	Cap.	Cent. Cond.	Diel.	Shield	Jacket	OD	Max V	Matched Loss (dB/100')			
Type	Number	Ω	%	pF/ft	AWG	Type	Type	Material	inches	(RMS)	1 MHz	10	100	1000
RG-6	Belden 1694A	75	82	16.2	#18 Solid BC	FPE	FC	P1	0.275	600	0.2	0.7	1.8	5.9
RG-6	Belden 8215	75	66	20.5	#21 Solid CCS	PE	D	PE	0.332	2700	0.4	0.8	2.7	9.8
RG-8	Belden 7810A	50	86	23.0	#10 Solid BC	FPE	FC	PE	0.405	600	0.1	0.4	1.2	4.0
RG-8	TMS LMR400	50	85	23.9	#10 Solid CCA	FPE	FC	PE	0.405	600	0.1	0.4	1.3	4.1
RG-8	Belden 9913	50	84	24.6	#10 Solid BC	ASPE	FC	P1	0.405	600	0.1	0.4	1.3	4.5
RG-8	CXP1318FX	50	84	24.0	#10 Flex BC	FPE	FC	P2N	0.405	600	0.1	0.4	1.3	4.5
RG-8	Belden 9913F7	50	83	24.6	#11 Flex BC	FPE	FC	P1	0.405	600	0.2	0.6	1.5	4.8
RG-8	Belden 9914	50	82	24.8	#10 Solid BC	FPE	FC	P1	0.405	600	0.2	0.5	1.5	4.8
RG-8	TMS LMR400UF	50	85	23.9	#10 Flex BC	FPE	FC	PE	0.405	600	0.1	0.4	1.4	4.9
RG-8	DRF-BF	50	84	24.5	#9.5 Flex BC	FPE	FC	PE	0.405	600	0.1	0.5	1.6	5.2
RG-8	WM CQ106	50	84	24.5	#9.5 Flex BC	FPE	FC	P2N	0.405	600	0.2	0.6	1.8	5.3
RG-8	CXP008	50	78	26.0	#13 Flex BC	FPE	S	P1	0.405	600	0.1	0.5	1.8	7.1
RG-8	Belden 8237	52	66	29.5	#13 Flex BC	PE	S	P1	0.405	3700	0.2	0.6	1.9	7.4
RG-8X	Belden 7808A	50	86	23.5	#15 Solid BC	FPE	FC	PE	0.240	600	0.2	0.7	2.3	7.4
RG-8X	TMS LMR240	50	84	24.2	#15 Solid BC	FPE	FC	PE	0.242	300	0.2	0.8	2.5	8.0
RG-8X	WM CQ118	50	82	25.0	#16 Flex BC	FPE	FC	P2N	0.242	300	0.3	0.9	2.8	8.4
RG-8X	TMS LMR240UF	50	84	24.2	#15 Flex BC	FPE	FC	PE	0.242	300	0.2	0.8	2.8	9.6
RG-8X	Belden 9258	50	82	24.8	#16 Flex BC	FPE	S	P1	0.242	600	0.3	0.9	3.1	11.2
RG-8X	CXP08XB	50	80	25.3	#16 Flex BC	FPE	S	P1	0.242	300	0.3	0.9	3.1	14.0
RG-9	Belden 8242	51	66	30.0	#13 Flex SPC	PE	SCBC	P2N	0.420	5000	0.2	0.6	2.1	8.2
RG-11	Belden 8213	75	84	16.1	#14 Solid BC	FPE	S	PE	0.405	600	0.2	0.4	1.3	5.2
RG-11	Belden 8238	75	66	20.5	#18 Flex TC	PE	S	P1	0.405	600	0.2	0.7	2.0	7.1
RG-58	Belden 7807A	50	85	23.7	#18 Solid BC	FPE	FC	PE	0.195	300	0.3	1.0	3.0	9.7
RG-58	TMS LMR200	50	83	24.5	#17 Solid BC	FPE	FC	PE	0.195	300	0.3	1.0	3.2	10.5
RG-58	WM CQ124	52	66	28.5	#20 Solid BC	PE	S	PE	0.195	1400	0.4	1.3	4.3	14.3
RG-58	Belden 8240	52	66	28.5	#20 Solid BC	PE	S	P1	0.193	1900	0.3	1.1	3.8	14.5
RG-58A	Belden 8219	53	73	26.5	#20 Flex TC	FPE	S	P1	0.195	300	0.4	1.3	4.5	18.1
RG-58C	Belden 8262	50	66	30.8	#20 Flex TC	PE	S	P2N	0.195	1400	0.4	1.4	4.9	21.5
RG-58A	Belden 8259	50	66	30.8	#20 Flex TC	PE	S	P1	0.192	1900	0.4	1.5	5.4	22.8

G or	Part	Nom. Z_0	VF	Cap.	Cent. Cond.	Diel.	Shield	Jacket	OD	Max V	Matched Loss (dB/100')			
Type	Number	Ω	%	pF/ft	AWG	Type	Type	Material	inches	(RMS)	1 MHz	10	100	1000
RG-59	Belden 1426A	75	83	16.3	#20 Solid BC	FPE	S	P1	0.242	300	0.3	0.9	2.6	8.5
RG-59	CXP 0815	75	82	16.2	#20 Solid BC	FPE	S	P1	0.232	300	0.5	0.9	2.2	9.1
RG-59	Belden 8212	75	78	17.3	#20 Solid CCS	FPE	S	P1	0.242	300	0.6	1.0	3.0	10.9
RG-59	Belden 8241	75	66	20.4	#23 Solid CCS	PE	S	P1	0.242	1700	0.6	1.1	3.4	12.0
RG-62A	Belden 9269	93	84	13.5	#22 Solid CCS	ASPE	S	P1	0.240	750	0.3	0.9	2.7	8.7
RG-62B	Belden 8255	93	84	13.5	#24 Flex CCS	ASPE	S	P2N	0.242	750	0.3	0.9	2.9	11.0
RG-63B	Belden 9857	125	84	9.7	#22 Solid CCS	ASPE	S	P2N	0.405	750	0.2	0.5	1.5	5.8
RG-142	CXP 183242	50	69.5	29.4	#19 Solid SCCS	TFE	D	FEP	0.195	1900	0.3	1.1	3.8	12.8
RG-142B	Belden 83242	50	69.5	29.0	#19 Solid SCCS	TFE	D	TFE	0.195	1400	0.3	1.1	3.9	13.5
RG-174	Belden 7805R	50	73.5	26.2	#25 Solid BC	FPE	FC	P1	0.110	300	0.6	2.0	6.5	21.3
RG-174	Belden 8216	50	66	30.8	#26 Flex CCS	PE	S	P1	0.110	1100	1.9	3.3	8.4	34.0
RG-213	Belden 8267	50	66	30.8	#13 Flex BC	PE	S	P2N	0.405	3700	0.2	0.6	1.9	8.0
RG-213	CXP213	50	66	30.8	#13 Flex BC	PE	S	P2N	0.405	600	0.2	0.6	2.0	8.2
RG-214	Belden 8268	50	66	30.8	#13 Flex SPC	PE	D	P2N	0.425	3700	0.2	0.6	1.9	8.0
RG-216	Belden 9850	75	66	20.5	#18 Flex TC	PE	D	P2N	0.425	3700	0.2	0.7	2.0	7.1
RG-217	WM CQ217F	50	66	30.8	#10 Flex BC	PE	D	PE	0.545	7000	0.1	0.4	1.4	5.2
RG-217	M17/78-RG217	50	66	30.8	#10 Solid BC	PE	D	P2N	0.545	7000	0.1	0.4	1.4	5.2
RG-218	M17/79-RG218	50	66	29.5	#4.5 Solid BC	PE	S	P2N	0.870	11000	0.1	0.2	0.8	3.4
RG-223	Belden 9273	50	66	30.8	#19 Solid SPC	PE	D	P2N	0.212	1400	0.4	1.2	4.1	14.5
RG-303	Belden 84303	50	69.5	29.0	#18 Solid SCCS	TFE	S	TFE	0.170	1400	0.3	1.1	3.9	13.5
RG-316	CXP TJ1316	50	69.5	29.4	#26 Flex BC	TFE	S	FEP	0.098	1200	1.2	2.7	8.0	26.1
RG-316	Belden 84316	50	69.5	29.0	#26 Flex SCCS	TFE	S	FEP	0.096	900	1.2	2.7	8.3	29.0
RG-393	M17/127-RG393	50	69.5	29.4	#12 Flex SPC	TFE	D	FEP	0.390	5000	0.2	0.5	1.7	6.1
RG-400	M17/128-RG400	50	69.5	29.4	#20 Flex SPC	TFE	D	FEP	0.195	1400	0.4	1.1	3.9	13.2
LMR500	TMS LMR500UF	50	85	23.9	#7 Flex BC	FPE	FC	PE	0.500	2500	0.1	0.4	1.2	4.0
LMR500	TMS LMR500	50	85	23.9	#7 Solid CCA	FPE	FC	PE	0.500	2500	0.1	0.3	0.9	3.3
LMR600	TMS LMR600	50	86	23.4	#5.5 Solid CCA	FPE	FC	PE	0.590	4000	0.1	0.2	0.8	2.7
LMR600	TMS LMR600UF	50	86	23.4	#5.5 Flex BC	FPE	FC	PE	0.590	4000	0.1	0.2	0.8	2.7
LMR1200	TMS LMR1200	50	88	23.1	#0 Copper Tube	FPE	FC	PE	1.200	4500	0.04	0.1	0.4	1.3

G or Type	Part Number	Nom. Z_0 Ω	VF %	Cap. pF/ft	Cent. Cond. AWG	Diel. Type	Shield Type	Jacket Material	OD inches	Max V (RMS)	Matched Loss (dB/100') 1 MHz	10	100	1000
Hardline														
½"	CATV Hardline	50	81	25.0	#5.5 BC	FPE	SM	none	0.500	2500	0.05	0.2	0.8	3.2
½"	CATV Hardline	75	81	16.7	#11.5 BC	FPE	SM	none	0.500	2500	0.1	0.2	0.8	3.2
⅞"	CATV Hardline	50	81	25.0	#1 BC	FPE	SM	none	0.875	4000	0.03	0.1	0.6	2.9
⅞"	CATV Hardline	75	81	16.7	#5.5 BC	FPE	SM	none	0.875	4000	0.03	0.1	0.6	2.9
LDF4-50A	Heliax –½"	50	88	25.9	#5 Solid BC	FPE	CC	PE	0.630	1400	0.05	0.2	0.6	2.4
LDF5-50A	Heliax –⅞"	50	88	25.9	0.355" BC	FPE	CC	PE	1.090	2100	0.03	0.10	0.4	1.3
LDF6-50A	Heliax – ¼"	50	88	25.9	0.516" BC	FPE	CC	PE	1.550	3200	0.02	0.08	0.3	1.1
Parallel Lines														
TV Twinlead (Belden 9085)		300	80	4.5	#22 Flex CCS	PE	none	P1	0.400	**	0.1	0.3	1.4	5.9
Twinlead (Belden 8225)		300	80	4.4	#20 Flex BC	PE	none	P1	0.400	8000	0.1	0.2	1.1	4.8
Generic Window Line		405	91	2.5	#18 Solid CCS	PE	none	P1	1.000	10000	0.02	0.08	0.3	1.1
WM CQ 554		420	91	2.7	#14 Flex CCS	PE	none	P1	1.000	10000	0.02	0.08	0.3	1.1
WM CQ 552		440	91	2.5	#16 Flex CCS	PE	none	P1	1.000	10000	0.02	0.08	0.3	1.1
WM CQ 553		450	91	2.5	#18 Flex CCS	PE	none	P1	1.000	10000	0.02	0.08	0.3	1.1
WM CQ 551		450	91	2.5	#18 Solid CCS	PE	none	P1	1.000	10000	0.02	0.08	0.3	1.1
Open-Wire Line		600	92	1.1	#12 BC	none	none	none	**	12000	0.02	0.06	0.2	0.7

Approximate Power Handling Capability (1:1 SWR, 40°C Ambient):

	1.8 MHz	7	14	30	50	150	220	450	1 GHz
RG-58 Style	1350	700	500	350	250	150	120	100	50
RG-59 Style	2300	1100	800	550	400	250	200	130	90
RG-8X Style	1830	840	560	360	270	145	115	80	50

Legend:

**	Not Available or varies	FC	Foil + Tinned Copper Braid	S	Single Braided Shield
ASPE	Air Spaced Polyethylene	FEP	Teflon® Type IX	SC	Silver Coated Braid
BC	Bare Copper	Flex	Flexible Stranded Wire	SCCS	Silver Plated Copper Coated Steel
CC	Corrugated Copper	FPE	Foamed Polyethylene	SM	Smooth Aluminum
CCA	Copper Cover Aluminum		Heliax Andrew Corp Heliax	SPC	Silver Plated Copper
CCS	Copper Covered Steel	N	Non-Contaminating	TC	Tinned Copper
CXP	Cable X-Perts, Inc.	P1	PVC, Class 1	TFE	Teflon® Systems
D	Double Copper Braids	P2	PVC, Class 2	UF	Ultra Flex
DRF	Davis RF	PE	Polyethylene	WM	Wireman

$$Z_0 = \left(\frac{138}{\sqrt{\varepsilon_r}}\right) \times \log_{10}\left(\frac{D}{d}\right)$$

Where D and d are the diameters of the shield and center conductor respectively in the same units. ε_r is the relative dielectric constant of the dielectric (for air, $\varepsilon_r = 1$, for polyethylene $\varepsilon_r = 2.26$).

VF (%) — This is the velocity factor as a percentage of the speed of light in free space (300,000,000 m/s). This is largely a function of the relative dielectric constant of the insulation material. For polyethylene dielectric, the VF is usually 66%, in the 80th percentile for foamed poly, depending on density, and close to 100% for air-insulated lines. In many applications, the velocity is not particularly important, with two notable exceptions — it is critical if lines are used for delay or phasing lines, and it also directly affects the required length of lines used as impedance transformers (covered later). The velocity factor can be found as shown below.

$$VF = 100/\sqrt{\varepsilon_r}$$

Cap (pF/ft) — This is the capacitance between the inner and outer conductors of the cable. It is a function of the ratio of the diameters and the dielectric constant. Note that the ratio of diameters also determines the Z_0, so it's not surprising that cables of the same Z_0 with similar dielectrics have the same capacitance. Sometimes pieces of transmission line are used as capacitors in circuits; otherwise this is not often important.

Dimensions and Materials — These are often major considerations. It is critical to know the dimensions if determining connector or adapter types. The abbreviations are listed at the bottom of the table. Note that some center conductors are indicated as being flexible — a prime attribute in many applications.

Max V (RMS) — This is the maximum Root Mean Square, or effective voltage that can be safely applied to the cable. Note that this doesn't equate to power, which is listed separately at the bottom and applies for a matched (SWR = 1:1) condition. This voltage might occur briefly for any reason (the maximum peak voltage is 1.4 times the listed voltage), and above that, arcing is possible. The power is generally more of a thermal issue. For example, the listed voltage for RG-58, 300 V_{RMS}, would translate to a power of 1800 W in a 50 Ω system, while the maximum power (at 1.8 MHz) is listed as 1350 W.

Matched Loss (dB/100 feet) — This is the attenuation of 100 feet of the cable when matched to its Z_0. Note that the cable loss increases with frequency and is shown for four frequencies across the spectrum. The loss is linearly related to length, so 200 feet has twice the loss, and 50 feet has half the loss. This is a key differentiator between cable types. Keep in mind that a 1 dB loss represents about a 20% reduction in power, a 3 dB loss is a 50% reduction in power, and 10 dB is a 90% reduction. All things being equal, larger cables have less loss (due to less conductor resistance) as do cables with higher velocity (less dielectric loss). The matched loss of major cable types is shown graphically in **Figure 3.7**.

ADDITIONAL LOSS DUE TO MISMATCH

You may have noticed that all the loss figures stated in Table 3.1 and shown in Figure 3.7 were for the case of a transmission line terminated in its Z_0. The additional loss due to mismatch is a function of both amount of the mismatch (SWR) and the loss if matched. **Figure 3.8** shows the additional loss in dB that occurs as a result of a line not being matched.

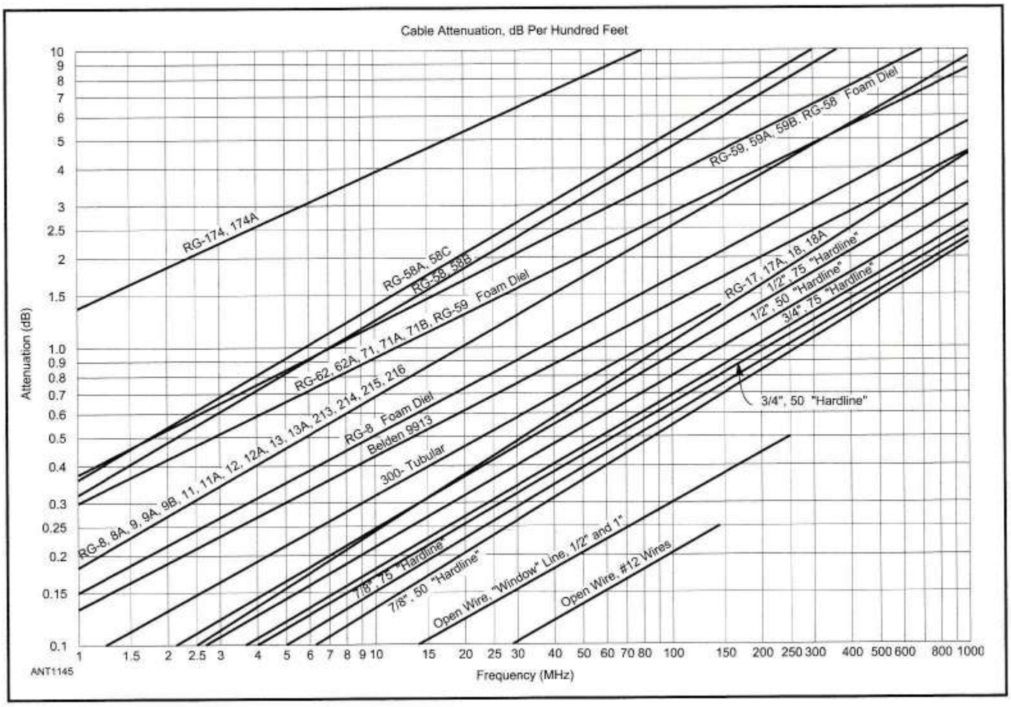

Figure 3.7 – Matched nominal loss of various transmission line types in dB/100 feet of representative transmission line types. As you choose transmission line, it is important to note that this data is representative — the variation between attenuation of different manufacturers and even different part numbers from the same manufacturer can be striking. Check the manufacturer's website for the data sheet of the transmission line you are considering to be sure you know what you are getting. For lengths other than 100 feet, the loss scales linearly. For example, if a line has a matched loss of 2 dB at 100 feet, the same line at the same frequency will have a loss of 1 dB if 50 feet long, and 5 dB if 250 feet long.

Note that the SWR shown is the SWR as measured or calculated at the load, and not the SWR measured at the transmitter end of the cable. This is particularly important in the case of a lossy line, since the loss will reduce both the power reaching the antenna and the power of the reflected wave that is used to determine the SWR. This can give very optimistic and erroneous results.

To give an example of this effect, consider a 100 W transmitter driving 100 feet of coax with a loss of 3 dB (50% loss). The antenna will see 50 W of power. Let's say 20% of the power is reflected due to the antenna mismatch. That will result in 10 W being reflected back toward the source. The 3 dB loss results in 5 W showing up as reflected power at the bottom of the cable. This is quite different than if the 100 W were applied to the antenna on lossless line — in

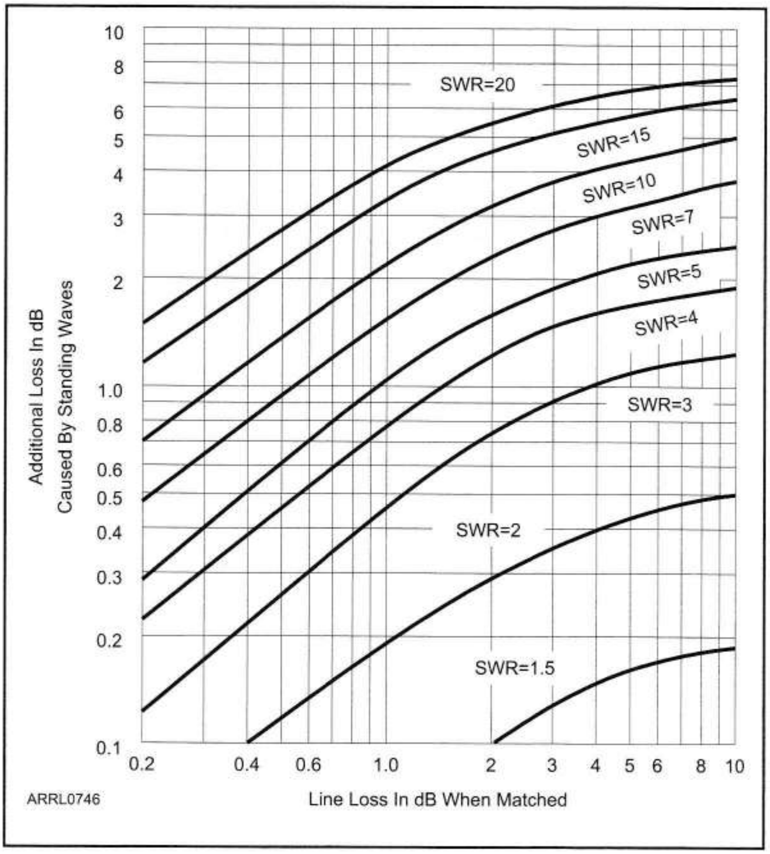

Figure 3.8 – Additional loss of a transmission line when mismatched. This loss needs to be added to the loss in Figure 3.7 for mismatched lines.

that case, a reflected power of 20 W, not 5 W, would show up at the SWR measurement device.

Table 3.2 summarizes what we have at the two locations. Note the rather distressing result. A very acceptable measurement of an SWR of 1.6:1 at the bottom of the coax is the result of an unpleasant SWR of 4:1 at the antenna. In this example, our 100 W of power results in only 40 W radiated from the antenna — yet all of our measurements make us think we're doing well. Unfortunately, this example is not unusual, especially at the upper end of HF into the VHF range. If it happens at higher frequencies, it is usually more evident since nothing much ends up going in or out of the system! The sidebar discusses ways that this can be calculated, and perhaps avoided, through the use of software.[3]

Table 3.2
Forward and Reflected Power and SWR as Seen at Each End of a Transmission Line With 3 dB Loss

Measurement	*Bottom of Cable*	*Top of Cable*
Forward Power (W)	100	50
Reflected Power (W)	5	10
Indicated Reflection Coefficient	0.224	0.447
Indicated SWR	1.6	4.0

Notes

[1] J. Hallas, W1ZR, *Basic Antennas*. Available from your ARRL dealer or the ARRL Bookstore, ARRL order no. 9994. Telephone 860-594-0355, or toll-free in the US 888-277-5289; **www.arrl.org/shop**; **pubsales@arrl.org**.

[2] B. Shackleford, W6YE, "Custom Open-wire Line — It's a Snap," *QST*, Jul 2011, pp 33-36.

[3] J. Hallas, W1ZR, "I Know What's Happening at the Shack — What's Happening at the Other End of My Feed Line?" *QST*, Feb 2007, p 63.

Chapter 4

Other Types of Unbalanced Transmission Line

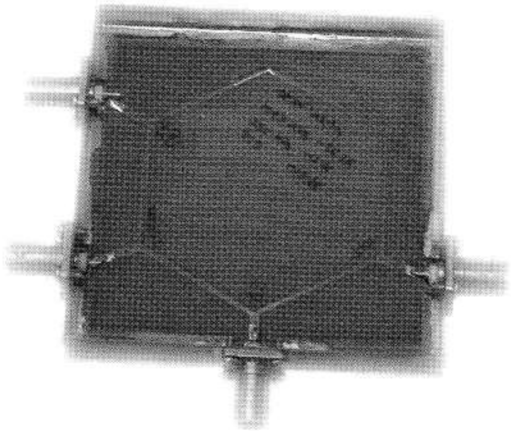

This microwave circulator is constructed of microstrip transmission line. Microstrip is easily implemented as part of PC board fabrication. The lines shown are single-sided stripline with the foil on the other side of the board serving as the "shield."

While coaxial cable is the most frequently encountered type of unbalanced transmission line, it is not the only type. This chapter will briefly introduce some other types of line in that category.

MICROSTRIP TRANSMISSION LINE

Microstrip is a form of transmission line that takes advantage of the properties of printed circuit (PC) boards. PC assemblies are a common and efficient method of circuit fabrication that has largely displaced the hand wiring of most electronic systems. The technique employs a thin piece of insulating material, the *substrate*, that serves as the mechanical strength member of the assembly and that supports all the components. Professional and military equipment generally makes use of a fiberglass-based material, while less critical consumer products are often made of phenolic resin.

A thin layer of copper that will eventually be transformed into component interconnections is bonded to the substrate. More common double-sided PC board has copper on both sides of the substrate. The PC designer (a person or software), working from the schematic of the circuit, lays out the board by specifying where each component will be placed and which pins need to be interconnected. The interconnections are made by removing

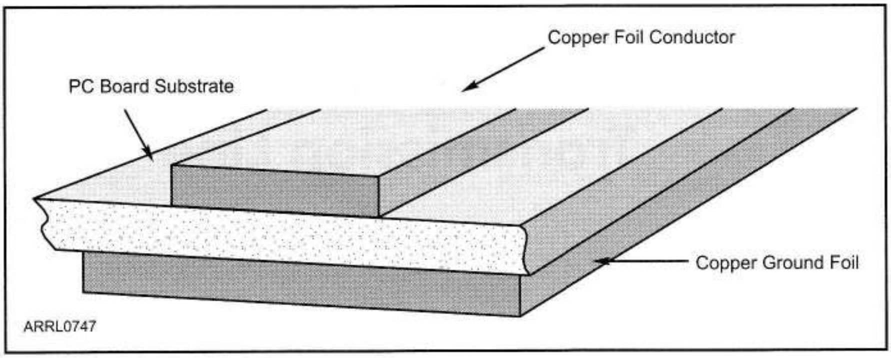

Figure 4.1 – Sketch of a typical microstrip transmission line as etched onto a PC board.

the copper that isn't needed. Often connections are established on both sides of double-sided board, but sometimes one side is used as a ground reference across the board. The unneeded copper is removed by using one of a number of methods of applying "resist" material to the areas that need to be kept. The board is then submerged in an acid bath to remove the excess copper. Following etching, the resist material is removed and the components are mounted to complete the circuit. While this sounds like an arduous process, compared to hand-wiring, it is easily automated and very efficient, especially if multiple boards are required for a production run.

RF circuits within a board, or leaving the board, can be interconnected using miniature coaxial cable with ends connected through holes in the board to the appropriate termination points just as other components with wire leads can be connected. Alternately, a microstrip transmission line can be formed on double-sided board during the printed circuit process. The line must be fabricated above a region that has a ground area on the bottom foil of the board, and merely consists of a conductor etched above the ground foil interconnecting to two end points. This is shown in **Figure 4.1**, along with the electric field between the conductors.

Note that unlike the case of the fields within a coax cable, the fields, while concentrated between the conductors, are not entirely contained there. This is a major limitation of microstrip line — it is not totally shielded. Therefore, care should be taken to make sure that coupling does not occur to sensitive circuits. This can usually be managed at the time of board layout.

In operation it is important that the connections of both input and output ends of the line include direct connections to the end of the "center conductor" and the ground plane directly beneath it.

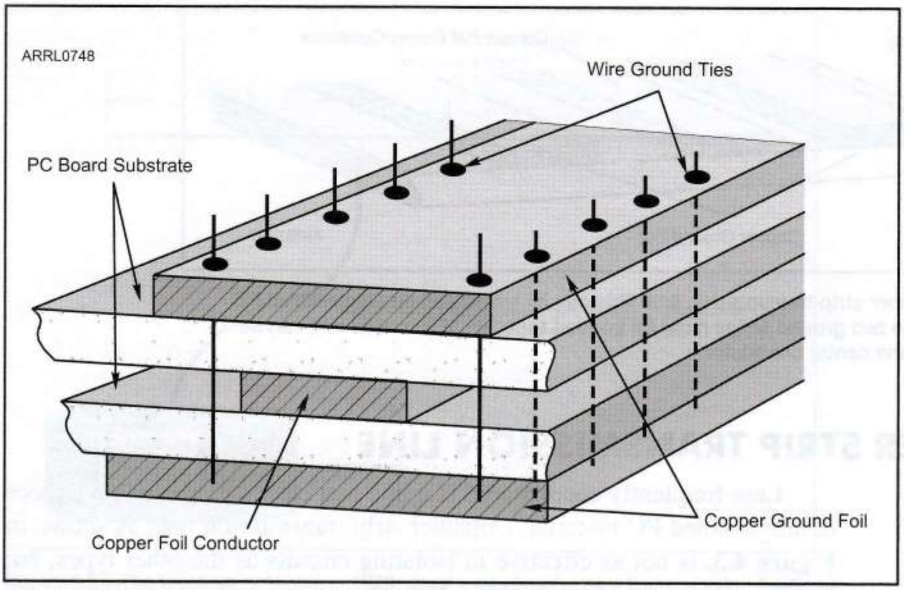

Figure 4.2 – Stripline transmission line, adding another ground layer to microstrip. Note the periodic connections between the two ground strips. These can improve the isolation to close to that of coax.

Characteristic Impedance

The characteristic impedance (Z_0) of microstrip line and the effective dielectric constant of the insulation are determined by the sizes of the conductors and the material used as the dielectric in a manner similar to coaxial cable.

STRIPLINE TRANSMISSION LINE

A variation of microstrip that actually was developed first is called stripline. It is just a microstrip "sandwich," formed by adding another piece of board with a ground foil above the single board in microstrip. Alternately, it can be formed as part of a multilayer board. It can have additional isolation from nearby conductors since the fields are more tightly constrained. It can approach coax in this regard if the two shield foils are tied together along the edge with frequent wires, as shown in **Figure 4.2**. In this case, the ground side of each end termination must connect directly to both ground foils in the area near the inner conductor.

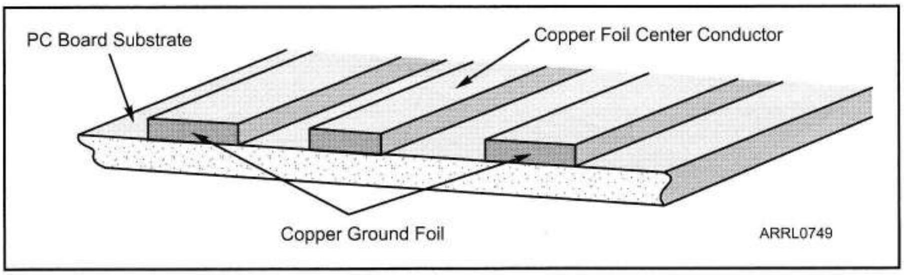

Figure 4.3 – Coplaner strip transmission line. This can be fabricated on single-sided PC board material. The two ground strips must be bonded together at the ends and can be much larger than the center conductor.

COPLANER STRIP TRANSMISSION LINE

Less frequently encountered is a line that can be deployed on a piece of single-sided PC material. Coplaner strip transmission line, as shown in **Figure 4.3**, is not as effective in isolating circuits as the other types, but may be usable in some situations. Note that in this case, the dielectric is a mix of the board substrate below and air above. Again, the ground side of the connection at each end must connect to both ground strips. This can be built into the PC board etching process.

SINGLE WIRE OVER A GROUND PLANE

An insulated wire located close to a metal chassis, or even a piece of PC material with a ground foil, can act like a microstrip line. (See **Figure 4.4**.) This is commonly employed in equipment that is hand-wired in the traditional manner between components on a metal chassis — in many cases without the builder thinking in terms of transmission lines. The effectiveness can be increased by locating the wire in a corner.

As with microstrip, while the end-to-end transmission is predictable and efficient, there can easily be unplanned and undesired coupling to other circuits within the chassis since the fields are not completely contained close to the wire. If there are other sensitive circuits close by, a piece of miniature coax can often fit in the same place and, if properly terminated, can reduce the coupling.

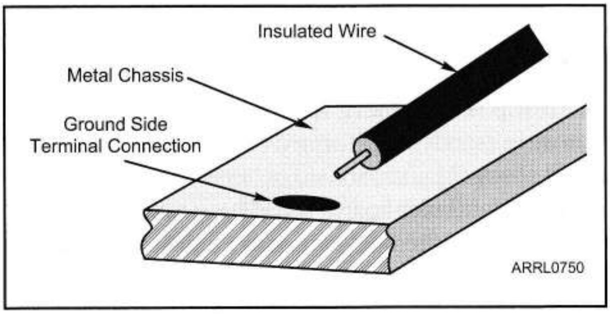

Figure 4.4 – An insulated wire positioned tightly against a chassis or other metal surface can serve as a transmission line similar to a microstrip line.

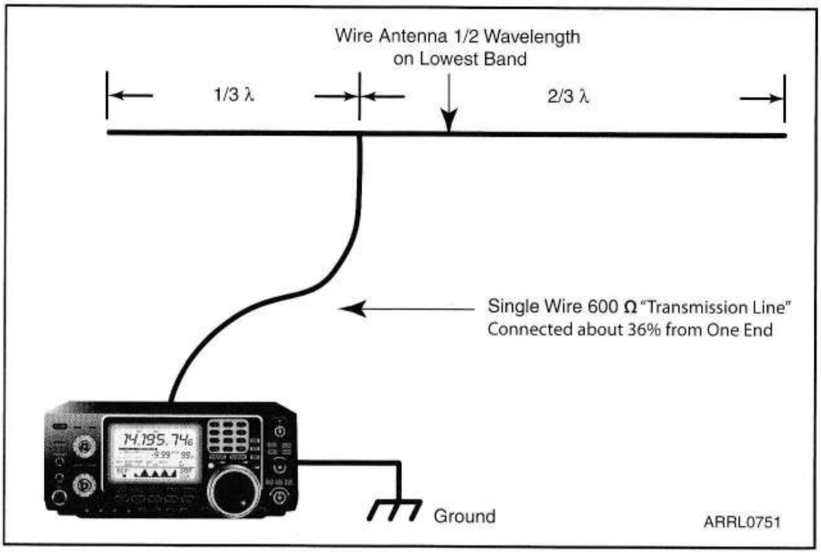
Wire Antenna 1/2 Wavelength
on Lowest Band
1/3 λ
2/3 λ
Single Wire 600 Ω "Transmission Line"
Connected about 36% from One End
14.195.746
Ground
ARRL0751

Figure 4.5 – The Windom antenna, c. 1929, used a single wire as a transmission line. There is controversy even today as to whether it really acts as a transmission line, or as a top-loaded vertical monopole.

SINGLE WIRE ISOLATED FROM GROUND

A popular multiband amateur antenna before World War II and continuing in popularity in different versions today is the Windom, as shown in **Figure 4.5**. It was described in detail in a 1929 *QST* article by L. G. Windom, based on work by others, and still carries his name.[1] The idea was that an antenna a half wavelength long on 80 meters could also be used on the other three HF bands of the time — 40, 20, and 10 meters.

However, there was a wide variation in the impedance at the center of such an antenna as the bands were changed. The solution was to feed it off center, at approximately one third of its length. The antenna would then provide a reasonable match to the line on all four bands.

The current in the "transmission line" has no direct return path to cancel its radiation, so it will radiate as well as (or perhaps better) than the antenna, depending on the band. In the typical configuration, this makes for a combination of vertically polarized radiation from the feed line and horizontally polarized radiation from the antenna. This may or may not be useful, depending on where the signal's desired destination is, but the RF at the bottom of the feed is almost never a good idea for the radio equipment or operators.

Recent Windom Implementations

More recently, Windom-type antennas have substituted high-impedance transmission lines for the single wire line. Some use a transformer at the antenna connection point to transform the impedance to 50 Ω for connection to coax. In both cases the lack of a balanced connection point at the antenna results in common mode current on the transmission line that will radiate. The coax-fed variety uses a choke at the bottom of the feed line to keep the current from the station. In both cases, the vertical section will radiate — commercial versions proclaim this as a selling point and have many happy users. One thing to keep in mind is that in Windom's day, we had four harmonically related bands. We still have those, but we've added five more, making an "all-band" antenna design even more of a challenge now than in 1929.

So while such a single wire may serve in some ways as a low-cost transmission line, and may have met the needs of early amateurs, it is not in the same category as lines in current use. It should thus be used with caution, if at all.

---

[1] L. Windom, W8GZ/W8ZG, "Notes on Ethereal Adornments — Practical Design Data for the Single-Wire-Fed Hertz Antenna," *QST*, Sep 1929, pp 19-23, 84.

Chapter 5

Let's Examine Balanced Transmission Line

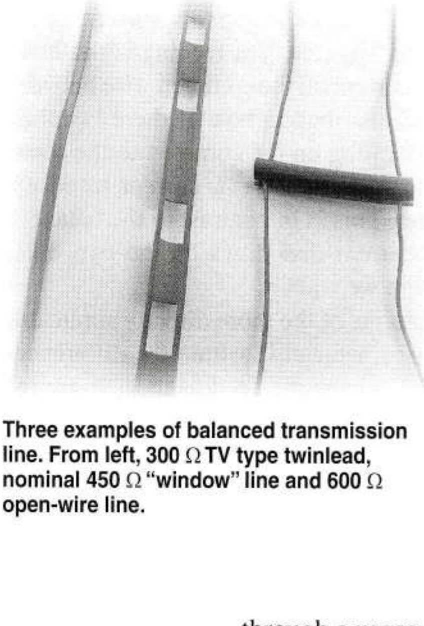

Three examples of balanced transmission line. From left, 300 Ω TV type twinlead, nominal 450 Ω "window" line and 600 Ω open-wire line.

In an unbalanced line of the type we have been discussing, the signal is applied between a conductor and a common conductor, often a shield or a ground plate. The voltage on the *hot* conductor is specified with respect to the common conductor. In a balanced line, there are two conductors, both at a potential with respect to ground. If properly balanced, both signals are of the same magnitude, but of opposite phase, so the total line voltage is twice the voltage of either with respect to ground.

This may be easier to grasp for those familiar with standard US household wiring. The transformer on the pole, which serves perhaps five houses, takes the multi kV distribution voltage and transforms it to 240 V ac. The secondary of the transformer has a center tap that is the *common* connection. The balanced pair and common are connected to the house, usually via a twisted wire cable. At the house end, the cable goes through a usage meter for billing purposes, then to the household distribution and circuit breaker or fuse box. The common connection is tied to the box frame, which is also tied to a ground connection at that point. Consider **Figure 5.1**, in which we portray the usual connectivity from the line to the primary ac distribution panel.

The two outer conductors are used to provide a balanced (with respect to common) 240 V feed for such heavy appliances as ovens, electric clothes dryers, or electric heating systems. Either outer conductor, fed with the common circuit, can be used to provide a 120 V unbalanced source for standard

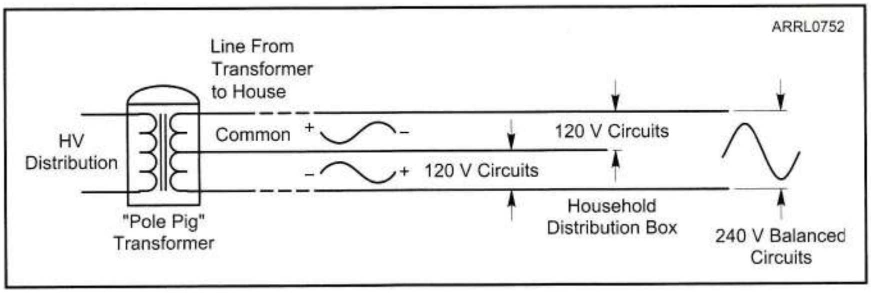
ARRL0752
Line From
Transformer
to House
HV
Distribution
Common
+
-
120 V Circuits
120 V Circuits
"Pole Pig"
Transformer
Household
Distribution Box
240 V Balanced
Circuits

Figure 5.1 – Typical household ac power connectivity illustrates the concepts of balanced and unbalanced transmission lines. The 120 V circuits are unbalanced with a common return, while the 240 V circuits and their transmission system are balanced with respect to the common.

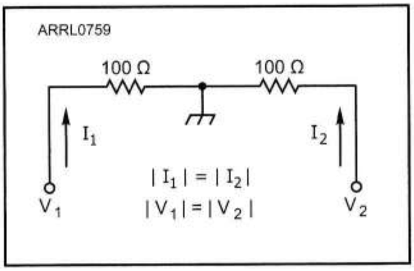
ARRL0759
100 Ω
100 Ω
I_1
I_2
$|I_1| = |I_2|$
$|V_1| = |V_2|$
V_1
V_2

Figure 5.2 – Perfectly balanced load, where both the current and voltage on each side will be in balance.

wall outlets with the common serving as a return. This is exactly analogous to radio transmission lines. Note that while each 120 V circuit returns its current on the common wire, if the two sides are feeding equal loads there will be no net current on the common wire between the transformer and distribution box. If there is unbalance, the current returning on the common to the transformer is referred to as *common mode* current, a concept we will discuss later. This is in contrast to the balanced *differential mode* current that flows on the two outer wires — that's where we want it.

In terms of the more directly applicable radio uses, we can illustrate the differences as well. For example, feeding the non-ground end of two 100 Ω resistors (see **Figure 5.2**) that have the other end grounded will result in a balanced 200 Ω system. In such a case, the voltage on each side will be the same magnitude, but 180° out of phase. The magnitude of the currents on each side will also be the same, since the voltages and resistances are equal.

Note that the system would be balanced even without the ground connection. No current flows in the ground lead of the perfectly balanced system, so it could be removed without changing the operational properties. **Figure 5.3** shows two examples of inherently balanced antennas, one with a central ground

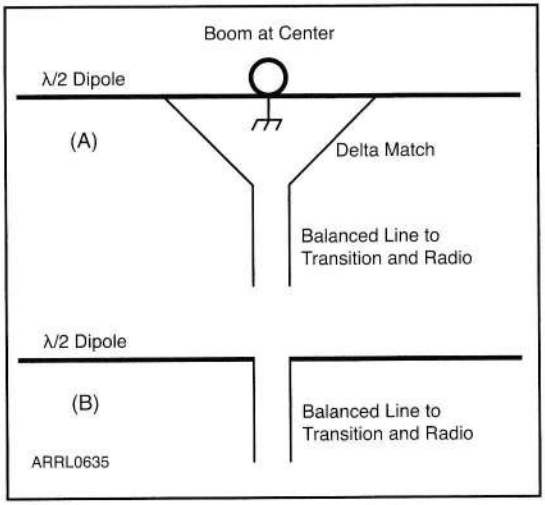
Boom at Center
λ/2 Dipole
(A)
Delta Match
Balanced Line to
Transition and Radio
λ/2 Dipole
(B)
Balanced Line to
Transition and Radio
ARRL0635

Figure 5.3 – Two examples of inherently balanced antennas, one (delta matched dipole) with a central ground (A) and one (split dipole) without (B).

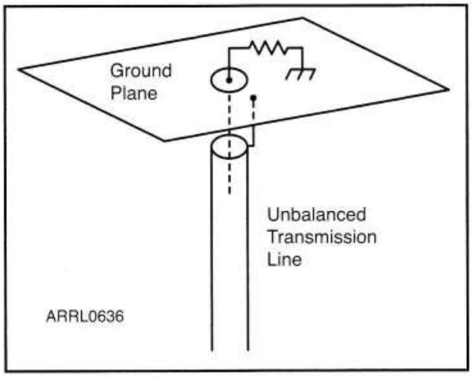
Ground
Plane
Unbalanced
Transmission
Line
ARRL0636

Figure 5.4 – Unbalanced resistive load. The signal is applied with reference to ground.

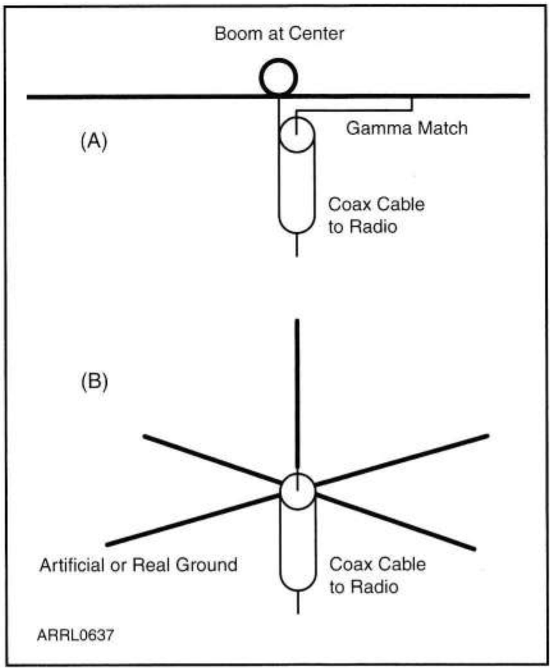
Boom at Center
Gamma Match
(A)
Coax Cable
to Radio
(B)
Artificial or Real Ground
Coax Cable
to Radio
ARRL0637

Figure 5.5 – Two examples of inherently unbalanced antennas, a gamma matched dipole at (A) and a vertical monopole at (B).

(A) and one without (B). The ground in the first is not actually necessary, but can be beneficial for lightning protection purposes. For both cases, some kind of transition is required to shift to an unbalanced system for connection to the radio. The types of transition will be the subject of a later chapter.

An unbalanced system, on the other hand, is fed with respect to ground. That is, one side of the load is at ground potential. **Figure 5.4** gives an example of an unbalanced resistive load, and **Figure 5.5** shows two antennas with inherently unbalanced feed points.

TYPES OF BALANCED LINE

Balanced transmission line is generally of relatively simple construction. Think two-wire lamp cord or speaker wire — both common examples of wire construction that could be used as balanced transmission line. Early receiving balanced line was typically a twisted pair of wires used to connect a dipole antenna to the commonly found balanced input of a receiving set of the time.

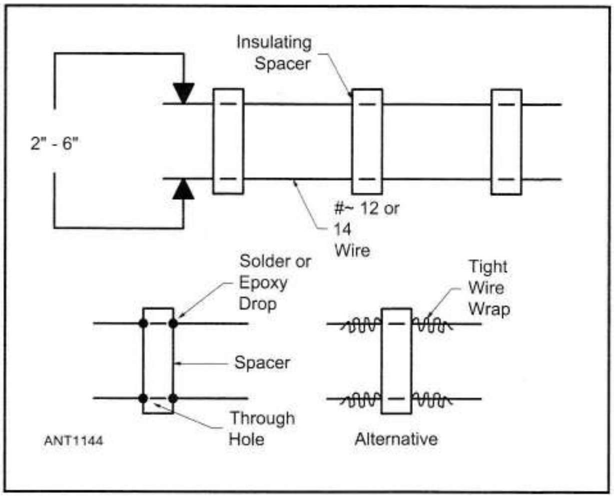
Insulating Spacer
2" - 6"
#~ 12 or 14 Wire
Solder or Epoxy Drop
Spacer
Through Hole
Tight Wire Wrap
Alternative
ANT1144

Figure 5.6 – Construction method of open-wire transmission line. Basically, just two parallel wires with a sufficient quantity of spacers to hold them apart.

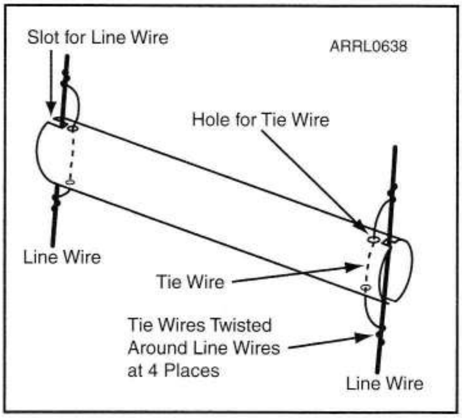
Slot for Line Wire
ARRL0638
Hole for Tie Wire
Line Wire
Tie Wire
Tie Wires Twisted Around Line Wires at 4 Places
Line Wire

Figure 5.7 – Modern homemade open-wire line often uses PVC tubing as the insulating spacers. Easier fabrication is possible if the wires are in slots rather than being pulled through holes. The small tie wires run through holes to secure the line to the spacer.

For transmitting use, open-wire line was usually employed. It is still popular today for many applications and offers simplicity of construction, as shown in **Figures 5.6** and **5.7**. In the earliest days, the spacers were often made of wood that was boiled in paraffin to reduce the penetration of water. During World War II, and for some time after, porcelain spacers were made for the purpose. They are still ideal in many ways, but no longer in production.

Modern commercial balanced line is often made using a molded polyethylene web as both spacer and insulator. As shown in **Figure 5.8**, such line comes in several varieties. The twinlead types have continuous insulation, while the window line (also available in 300 Ω versions) has "windows" punched out periodically to make it almost equivalent to open-wire line. Twinlead at 300 Ω was originally developed as a transmission line for television reception, and was popular for connections to TV antennas until inexpensive low-loss coax overtook it in the marketplace

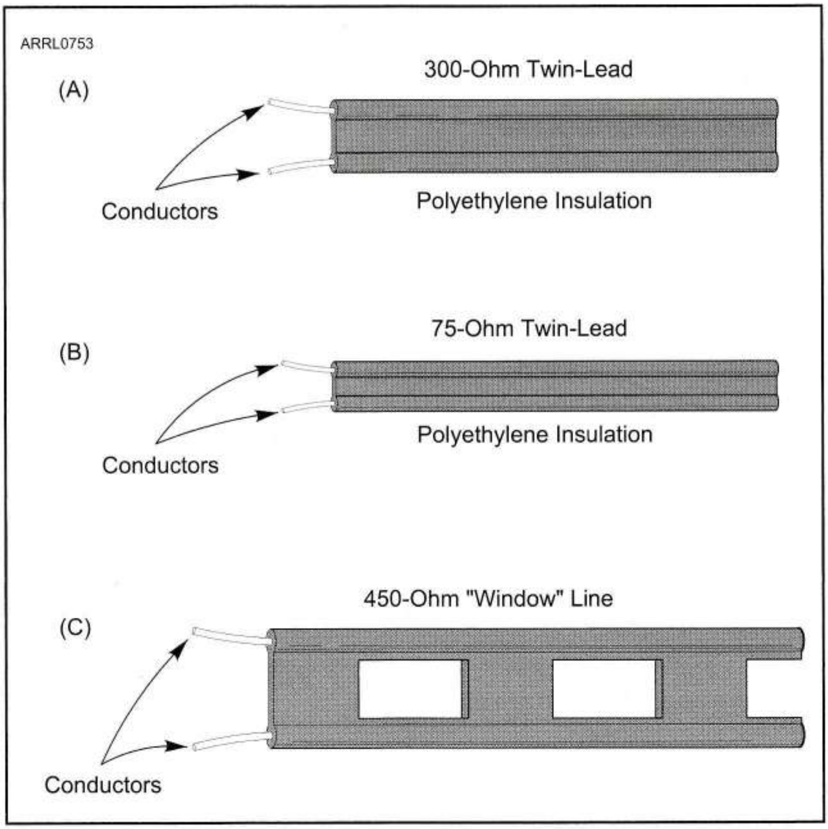
ARRL0753
(A)
300-Ohm Twin-Lead
Conductors
Polyethylene Insulation
(B)
75-Ohm Twin-Lead
Conductors
Polyethylene Insulation
(C)
450-Ohm "Window" Line
Conductors

Figure 5.8 – Examples of balanced transmission line made with a molded polyethylene insulating webbing. Window line is made the same way, except that "windows" are punched out periodically to make the characteristics more similar to open-wire line.

in the 1970s. It is still available at electronic retailers, as are TV and FM antennas designed to use it. Open-wire line, constructed much like that shown in Figure 5.7, is available commercially as well.

There are a few varieties of shielded twinlead around as well. These have been in use for some time, first in military systems with various RG numbers, including RG-22 and RG-97 with a Z_0 of 95 Ω. Later, shielded 300 Ω twinlead was offered for TV antenna use in areas that would not be appropriate for usual balanced lines. Unfortunately, both have the advantages and disadvantages of unbalanced coax. They are easy to route through materials that would give problems to standard twinlead, but because of their construction, they also have loss characteristics similar to coax. If needed, a similar balanced transmission line can be fabricated from two coax cables,

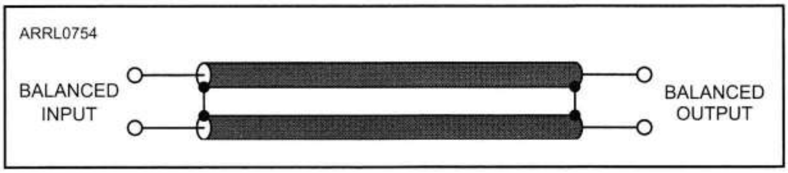

Figure 5.9 – Balanced transmission line made from two equal-length pieces of coax with shields tied together at each end. While it offers the installation flexibility of coax, it also has similar loss.

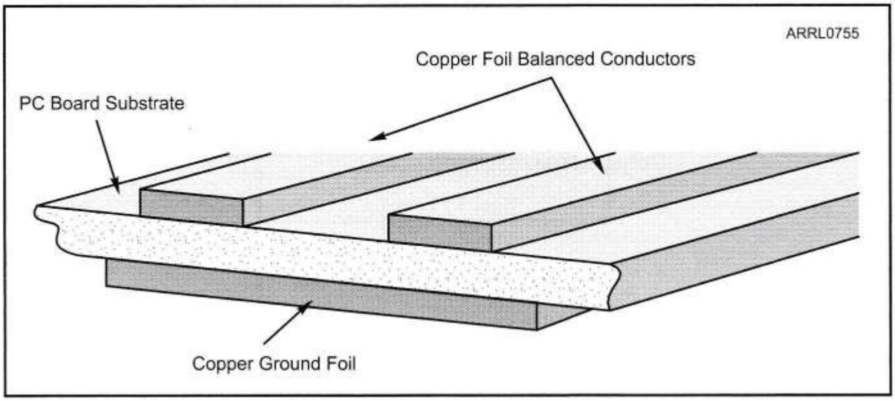

Figure 5.10 – Balanced microstrip line fabricated as part of a printed circuit.

as shown in **Figure 5.9**. The Z_0 is just twice the Z_0 of the individual cables. Note that if the shields are tied together at both ends, there will be no fields between the two cables, so they can even be run along different routes, as long as they are the same length. The loss will be the same as for coax.

Printed circuit board fabrication of stripline, or microstrip line, is also adaptable to balanced systems. The balanced microstrip line shown in **Figure 5.10** utilizes a balanced pair of conductors on the opposite side of the substrate from a ground foil to help contain the fields; however, a balanced pair on single-sided board is also possible.

THE BENEFITS OF BALANCED LINE

Balanced line has two primary benefits in comparison to coaxial cable. Most balanced lines, especially those with large portions of air dielectric, such as window line or ladder line, have a significantly lower loss than most coaxial cables.

Balanced line is often very conveniently connected to balanced antennas and is most often encountered in that application. It is also frequently found in very long runs from unbalanced antennas in which the losses (or cost) of coax would be prohibitive. In that case, it may be worth the effort to transition from unbalanced to balanced near the antenna and then back to unbalanced at the radio end.

Most balanced line is relatively inexpensive when compared to coax — often an important consideration. Excellent homemade open-wire line can be fabricated at low cost from two rolls of wire (stranded works very well, particularly if subject to flexing) and insulators made from inexpensive household PVC tubing.

THE DOWNSIDES OF BALANCED LINE

As might be expected, there are downsides to balanced line. While coax cable, with its fields contained within the shield, can be rolled, buried (if it has a direct burial rating), or installed within or near pipes without impacting performance, this is not true of balanced line. With balanced line, the fields conveying the signal down the line exist in a region around the two lines several times the distance between them. This means that balanced line can't be placed on the ground, run in metal ducts, run through lossy material, or be rolled up without causing additional loss or change in characteristics.

FIELDS BETWEEN CONDUCTORS OF BALANCED LINE

The fields between the conductors of balanced lines are not constrained within a boundary, as is the case with coax. The strongest fields are in the space directly between the conductors, but the total fields between are similar to that shown in **Figure 5.11**. Outside the region between the conductors, the fields will tend to cancel in the plane at the center between and perpendicular to the wire axis. Coupling to other circuits will be strongest near either wire until the distance is large compared to the spacing. The coupling will be minimal for most purposes at a distance of two to three times the conductor spacing.

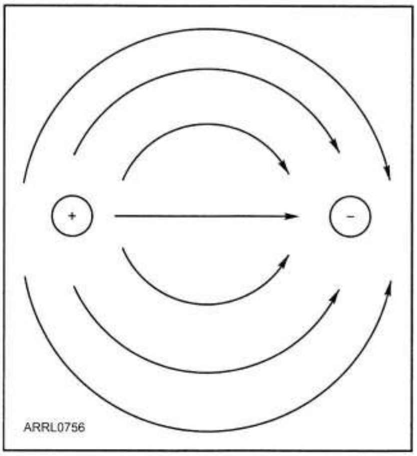

Figure 5.11 – Representation of the electric field between the conductors of balanced transmission line. The magnetic fields that surround each conductor are equal and of opposite sense if the currents are balanced.

In the heyday of twinlead for TV use, tubular twinlead was popular for critical applications. This provided an elliptical cross-section of air space in the region between the conductors in an attempt to minimize losses in the region in which the fields are strongest.

While fields at a distance from balanced line are small, nearby fields can cause interference to other systems, particularly those interconnected by wiring.

In addition, balanced line run too close to computers and other potential RFI-generating systems can pick up interfering signals during reception.

ELECTRICAL CHARACTERISTICS OF BALANCED LINE

As with coax, a lumped circuit model of balanced transmission line can be generated based on the per unit electrical characteristics. A schematic diagram of the result is shown in **Figure 5.12**. According to Kraus, the per unit inductance of a balanced pair of similar wires removed from ground and with an air dielectric is:

$$L = 0.92 \times \log^{10} (D/r)\ \mu H/m, \text{ if } D >> r$$

Where D is the center-center spacing, and r is the radius of each conductor in the same units.[1]

The characteristic impedance, under the same conditions, is:

$$Z_0 = 276 \times \log^{10} (D/r)\ \Omega$$

Since Z_0 also equals $\sqrt{(L/C)}$, we can easily find the capacitance as $C = L/Z_0^2$.

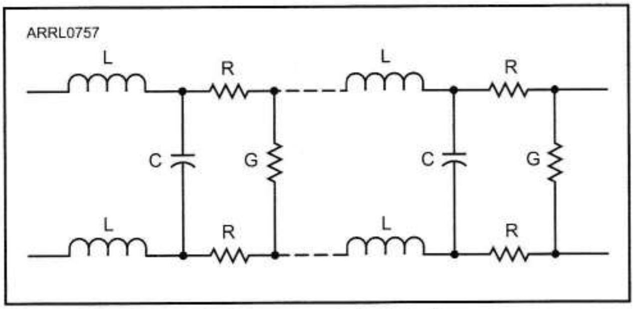

Figure 5.12 – Lumped element model of a balanced transmission line.

The series resistance is just the dc resistance of the wire at low frequencies, increasing at higher frequencies due to skin effect forcing current to the outer region of the conductors with corresponding reduction in effective surface area of the conductors.

The shunt conductance is the result of the losses in the dielectric. This is usually minimal in air dielectric line.

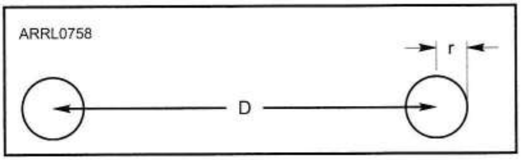

Figure 5.13 – Dimensions used to calculate transmission line parameters.

MATCHED LINE LOSS OF BALANCED LINES

The loss of matched balanced transmission line is provided in Table 3.1 in Chapter 3. It is worth making a note of the difference between the matched loss of coaxial cables shown in comparison to the losses of window line. Looking at the loss per 100 feet column of coax cable at 10 MHz, we note that most are in the range of 0.5 to 1.5 dB/100 feet. There are very large and very small cables outside this range, but typical matched loss would be around 1 dB/100 feet. The balanced lines shown are in the range of 0.06 to 0.3 dB/100 feet, with window and open-wire line all in the 0.06 to 0.08 dB/100 foot range. At the 100 MHz point, the difference is perhaps even more striking. Flexible (non-hardline) coax, in the usual sizes, runs from around 1 dB to more than 5 dB/100 feet, while the window lines are all 0.3 dB/100 feet or less.

MISMATCHED LINE LOSS IN BALANCED LINES

The low matched loss results in dramatically lower total line loss for window line in comparison to the usual coax. Let's compare two lines with an SWR of 10:1. If we had 100 feet of coax with a 5 db loss (see Figure 3.8 in Chapter 3), we would add an additional 4 dB of loss due to the mismatch for a total of 9 dB. For window line with a matched loss of 0.3 dB, we would add only 0.95 dB, for a total of about 1.3 dB.

If we put 100 W of transmit power into the coax, at the far end we will have 12.6 W available to radiate from our antenna — a loss of 87.4 W in the transmission line as heat. With the window line, the antenna will get 74.1 W, a loss of 35.9 W in the transmission line. While 1 dB is usually just perceivable, 9 db can make a big difference.

This is the big strength of balanced line, particularly window and open-wire lines at HF where the matched loss is dramatically small. With window line, antenna systems can be designed to operate with high SWR, making it easy to cover multiple bands, for example. Coax, on the other hand, is generally limited to uses with an SWR of no more than 2 or 3:1 to avoid excessive loss. It is particularly noteworthy that the best window line generally costs less than even inexpensive coax.

POWER HANDLING CAPABILITY OF BALANCED LINE

Table 3.1 does not provide any information about the power rating of the various balanced lines. While the dielectric strength of polyethylene will be well into the megavolt per inch range, the line breakdown will likely be limited by the dielectric strength of air at most interfaces, about 76 kV/inch. The wire will also fuse at some current, around 82 A for #18 AWG copper wire.

Keep in mind that both the maximum voltage and current on the line will go up with the square root of the SWR, compared to what they would be for a matched line. As an example, consider a 450 Ω window line at 1500 W, the US amateur power limit. The voltage if matched would be:

$V = \sqrt{(Z_0 \times P)} = \sqrt{(450 \times 1500)} =$ 822 V, with a 100:1 SWR, a maximum of 8222 V at some locations along the line.

The line current would be:

$I = \sqrt{(P / Z_0)} = \sqrt{(1500 / 450)} = 3.3$ A, with a 100:1 SWR, a maximum of 33 A at some locations along the line.

Neither of these quantities would be a problem for the physical characteristics of the line; however, the components to which they are connected in terminal equipment could be a very different story. Note that the maximum of one parameter in mismatched line will occur at the location along the line of the minimum of the other, with the maximums separated by ½ wavelength. Thus the maximums need not be encountered at all, and the line length can be designed to avoid maximums at sensitive locations, but only if the system will be used at a particular frequency. Multiband systems are likely to hit some maximum on some band.

---

[1] J. Kraus, W8JK (SK), *Electromagnetics*, McGraw-Hill Book Company, New York, 1953.

Chapter 6

Transmission Line Connectors

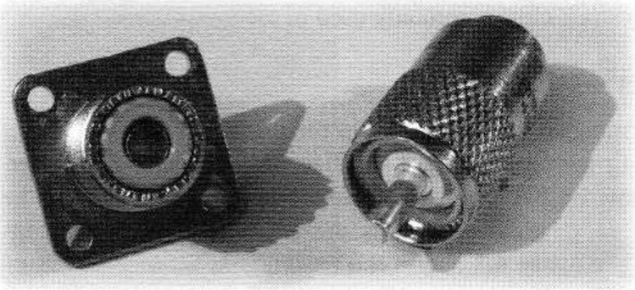

The popular "UHF" connector series is synonymous with "coax connector" for many, yet it is not always appropriate.

The most usual application of transmission lines is interconnecting different subsystems such as radios to antennas, interconnecting video systems or any functions that need to be interconnected in a well-defined way. A transmission line only becomes useful if the ends are properly prepared to interconnect with equipment at each end.

Generally, purchased equipment will include one side of the connector that the designer feels is appropriate for the application as part of the equipment. It is the responsibility of the systems integrator or installer to research the connector type and make sure the transmission line is terminated in a connector that will mate with the one on the equipment at each end to allow the connections to be made. Note that with different equipment types, the connector at each end of the cable may be different.

If you are designing the entire system, you have the flexibility to select the connector types as part of your design. It is important to select connectors that do not degrade system performance. This chapter will discuss the various connector types and their properties to help in that determination.

There are many types of connectors especially designed for termination of coaxial cables. These cables first came into widespread use in radio and video (radar screen data, for example) systems around the time of World War II due to the use of higher frequencies for communications and navigation systems for the military. In subsequent years there have been many new types developed for special purposes, some of which have come into general use.

"UHF" SERIES COAXIAL CONNECTORS

The UHF connector series was defined at a time when the "ultra high frequencies" meant anything above 30 MHz. Since we now consider VHF the range between 30 and 300 MHz and UHF from 300 to 3000 MHz, the designation is no longer appropriate, hence my use of quotation marks in the heading.

The first use of the UHF connector series was for broadband video systems such as radar signals. It allowed a termination of RG-8 sized coaxial cable into a continuously shielded interconnection across an interface. This was a significant improvement compared to the open screw terminals, binding posts, or Fahnestock clips commonly used in the prewar era.

The Basic UHF Connector Series

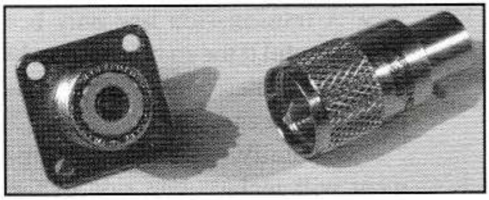

Figure 6.1 – The basic elements of the UHF coaxial connector series. A PL-259 UHF plug is on the right, with a mating SO-239 socket on the left.

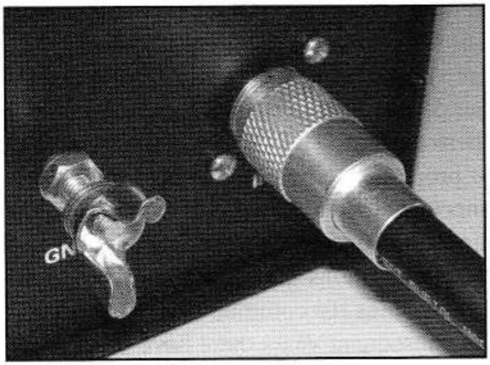

Figure 6.2 – A PL-259 plug attaches a cable to an SO-239 jack on a piece of equipment in the author's station.

The connector consists of a center pin for the center conductor that plugs into a spring tensioned socket the size of a "banana" jack, surrounded by a screw-on backshell that provides the connection for the shield and holds the connection in place. **Figure 6.1** shows the basic elements of the system. The basic cable-terminating plug is still known by its World War II-era nomenclature, PL-259. The matching flange-type socket, nomenclature SO-239 (note difference in center digit), is normally used on the equipment side and mounts to a chassis or panel with four machine screws or other fasteners at the corners of the flange, which also serves as the shield termination.

The original PL-259 design called for the center conductor to be soldered to the inside of the hollow pin while the outer conductor or shield is soldered inside the outer shell through the holes provided. The inside of the outer shell of a PL-259 includes course threads that can help secure the cable jacket if tightly screwed in before soldering. **Figure 6.2** shows a typical application of a UHF connector pair.

Additional UHF Connector Types

Over the years there have been additional types of UHF connectors added to the series, some officially and some designed to meet particular needs.

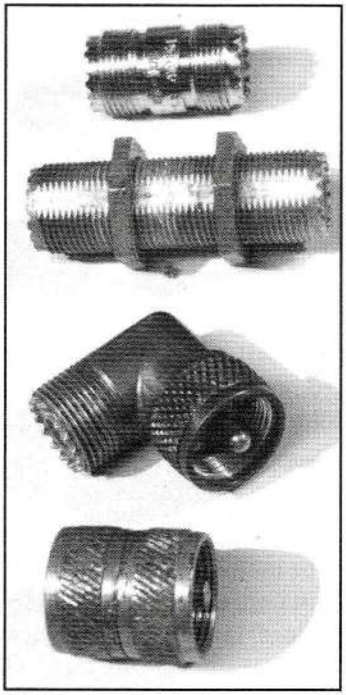

Figure 6.3 – Additional UHF series connector types. At the top is a PL-258 coax splice with an extended version below. Next is a UHF male-female elbow adapter that allows a change in cable direction — handy with heavy cables. At the bottom a double-male adapter.

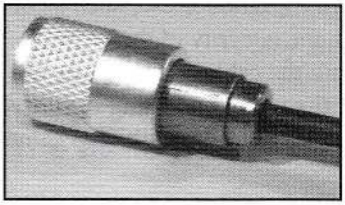

Figure 6.4 – A PL-259 with UG-175 reducer and an installed RG-58 cable.

Double Female Adapter: Probably the first was a double female adapter that allowed the interconnection of two male connectors, effectively splicing two cables together. This has a military designation of PL-258 and is shown at the top of **Figure 6.3**. Because of their shape, they are often referred to as *barrel* connectors. An unofficial expansion of the PL-258 concept is a somewhat similar double female connector that is made longer with a continuously threaded outer surface. These are usually provided with matching nuts (see one below the standard PL-258 in Figure 6.3) designed to secure them going through panels, partitions, or even walls. I have seen them up to 1 foot long.

Elbow Connectors: A male-female elbow connector (below the PL-258s in Figure 6.3) is often useful for changing a cable's angle of entry into equipment. This will allow a cable to hang straight down behind equipment without requiring a tight radius bend.

Double Male Adapter: Not surprisingly, double male adapters (shown at the bottom of Figure 6.3) are occasionally encountered as well, although the need to interconnect two pieces of equipment at close range is less common.

Reducing Adapters: The sizing of both the inside of the pin and the inside of the shell is just right for the 0.405 inch outside diameter of RG-8 or RG-11 and their inner conductors, or for their successor types of cables of similar size.

In many applications, it is desirable to use the smaller, more flexible RG-58 or RG-59 cables, especially for short patch cables. The shields of these cables are too small to be properly soldered to the inside of the PL-259 plug shell.

UHF connector adapters are available that screw into the back of the shell and provide a proper fit to the outer jacket of the smaller cable. The shield is then folded back over the adapter to appear at the plug solder holes. While the center conductor is also smaller, generally the pin can be filled with solder to complete the connection. The adapters, sometimes called reducers, are available in a number of sizes with the most common being the UG-175 for RG-58 size cables and UG-176 for RG-59 size cables including RG-8X. A UG-176 adapter with a PL-259 and an installed cable end are shown in **Figure 6.4**.

T Connectors: There are also triple-connector T connectors, sometimes with all three ports female, sometimes with the center port a male plug. These are useful for bridging cable connections, but consider the implications of parallel impedances to the system SWR. Another variation that is frequently encountered is a jack with a threaded body that mounts through a panel hole much like half of the extended double female.

Jack Hood: In order to address the unshielded nature of the standard SO-239 panel jack, a special hood is also available that provides a solder connection to the shield and a shielded surrounding of the center conductor. While this can improve isolation, it exacerbates the Z_0 issue because of the changing outer conductor diameter of the hood's conical shape.

Jacks Without Flange: Jacks are available that mount into a single ⅝ inch diameter hole. The threads for the backshell are extended to the rear of the connector, and a shoulder is located on the panel side. It looks superficially like a PL-258, but has a center solder terminal.

Limitations of UHF Series Connectors

That the UHF connector is as popular as it is may be a testament to the power of being early on the scene. While it can be useful, it does have many limitations and weaknesses, none of which make it unsuitable, but all of them need to be taken into account. The major limitations, in my view, are listed below.

◆ The characteristic impedance of the connector itself is not the same as that of the usual coaxial cables with which it is used. The connector pair Z_0 is typically 30-40 Ω for perhaps an inch of length. While there actually are 36 Ω cables for which this would be a plus, they are rarely encountered. This results in an "impedance bump" that can cause a change in the impedance at the input to the system, depending on the electrical length.

The significance of this limitation will depend on how critical the cable impedance match, or SWR, is to the system operation. It is also true that it will be much more significant at frequencies with wavelengths above about 10 times that length or 10 inches. This corresponds to just above 1 GHz, within the UHF range. Many elect to use other connector series above 6 or 2 meters.

◆ UHF connectors are not waterproof. This is not usually a problem with their use indoors, but can be a serious problem for outdoor use, such as on antennas. Water not only causes problems with the integrity of the connector itself, but can also result in water migration into the cable. While water in a cable is never a good thing, it gets worse over time, often resulting in corrosion and contamination of the dielectric. This can result in additional cable loss over the years.

Fortunately, the connectors can be waterproofed using various tape-like products intended for the purpose. This is an extra step and one that impedes the connection and disconnection of outdoor equipment.

◆ The shield connection primarily occurs at the perpendicular surface between the two connectors, augmented by the backshell itself. The integrity of both aspects of this connection is dependent on the tightness of the backshell. In mobile systems, or others subject to vibration, this can result in loose, high resistance, or even disconnected shields over time. Such conditions result in excessive loss, noise, undesired coupling, and functionally the loss of all benefit of coax in the first place. Many recommend the use of pliers for an extra bit of torque, rather than just hand-tightening the backshells.

◆ Availability of low-quality products. While this is not a fault of the connector design, it is a fact that the UHF series is relatively easy to duplicate. It has also been around long enough so that there are many substandard UHF connectors in the marketplace, which may present the following issues:

- Inner conductor plug pins that are too narrow to accommodate RG-8 center conductors.
- Inner conductor plug pins or connector sockets that fall out of the body.
- Plug insulation that melts while soldering.
- Poor-quality metal parts, including insufficient or improper material plating.
- Jack spring fingers that lose tension after a few insertions.

The solution is simple — only buy products from manufacturers known to make high-quality products. The one that I know is always of high quality is Amphenol, although I'm sure there are others. All major Amateur Radio dealers offer Amphenol UHF connectors, and the prices are reasonable. I recommend avoiding electronic retail store house brands, as well as any connectors found at flea markets that don't show a known high-quality brand.

TYPE N COAXIAL CONNECTORS

The Type N connectors are of a similar size to their UHF counterpart, but these connectors are seriously designed for UHF operation. They were designed in the 1940s by Paul Neill of Bell Laboratories to overcome the limitations associated with the UHF type, particularly with respect to higher frequency operation. While originally specified to work well up to 1 GHz, they are routinely used at frequencies more than 10 times that high. The nomenclature is in appreciation of Neill's name.

The Basic Type N Connector Series

Only a very close look at a connected pair will indicate that it's not a UHF pair, however, once they are apart, the difference is clear. The biggest difference is that the shield interconnection is primarily made via a spring

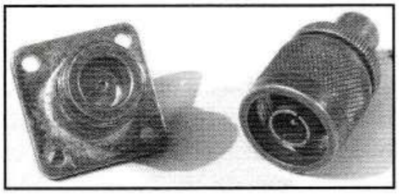

Figure 6.5 – A Type N connector pair — a good choice as an alternate in any application for which a UHF type would be used, but especially beneficial in the current UHF range or in wet conditions.

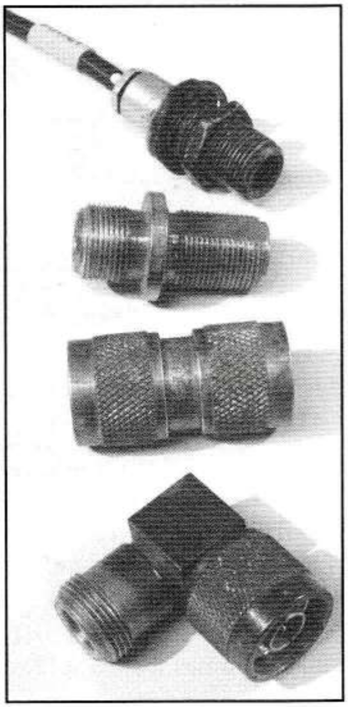

Figure 6.6 –Type N series connector variations. At the top, a Type N cable socket, below double female, double male and right angle adapters. Not shown — T adapters. All of the varieties of UHF connectors are also available in Type N.

fingered sleeve surrounding the male pin that fits snugly into a matching cylinder in the female jack. This means the shield will be connected even if the backshell is loose.

A Type N connector pair is shown in **Figure 6.5**. The concentric shield connection on each side is evident in the photo. The size of the shield connection, the air space in between, and the mated pin/socket combination make for a constant-impedance, low-loss match.

There are two versions available; one with center pin and socket diameters that have a Z_0 of 50 Ω, generally used for RF work, and another at 75 Ω intended for video applications.

The constant impedance avoids the impedance bump associated with the UHF series — the major frequency limiting factor of the UHF series. To put frosting on the cake, properly assembled Type N connectors are also waterproof.

As discussed detail in the *Coaxial Connector Assembly* chapter, the traditional Type N connector is assembled in a different manner than the soldered UHF type. A clamping arrangement is used to assemble the connector with the shield compressed between a washer and the shoulder of both sides of the connector. This avoids the melting cable dielectric insulation problem encountered with the soldering of shields to some kinds of connectors, but at a cost of requiring more careful measurement of trimming dimensions to make all the pieces come together in the right place. The only soldering required is of the low mass center conductor to its pin.

Other Varieties of Type N Connectors

Virtually all the different connector types described in the section under UHF connectors are also available for Type N connectors. A selection of Type N connectors is shown in **Figure 6.6**. In addition, the Type N series also offers a cable jack that looks like a cable plug, but is of opposite gender. This is significant because it allows the fabrication of extension cables that can be spliced together without the need for a barrel, or double female adapter. This is beneficial because each connector interface contributes to system loss and reduction of system reliability.

Limitations of Type N Connectors

Here we have a much shorter section than for UHF connectors. As with all connector types, there is a power limitation associated with the connectors, basically related to its physical dimensions.

Power Limits

The peak power is limited by the voltage breakdown across the air gap between the center pin and the shield connection, with a safe recommendation of at least 1000 V_{RMS} corresponding to a power level of 50 kW, while the average power is based on the heating of the conductors, limited by the current in the center pin. Skin effect concentrates the current in thinner and thinner outer regions of the pin as frequency increases, resulting in a typical average power rating around 5 kW at 20 MHz, decreasing with frequency to around 500 W at 2 GHz. Check the connector manufacturer's (as well as the cable manufacturer's) specifications if you expect to approach these limits.

Impedance Variation

This is a more insidious issue. The two different versions of the Type N connector have exactly the same size backshell and shield connection arrangement. The difference between them is strictly with the diameter of the center conductor, the 75 Ω version being thinner to result in a higher Z_0, as you would expect.

This means that a 75 Ω plug will mechanically fit into a 50 Ω Type N socket without difficulty. Unfortunately, it will not make a good contact. I am aware of one case in which the assembly alignment of the two was so perfect that it didn't make contact at all, leading to a lot of head-scratching.

The problem in the other direction is actually worse, in a way. Inserting a 50 Ω Type N plug into a 75 Ω socket will cause the center female socket to open further than it should. It will likely work in this situation, however, if a 75 Ω plug is ever inserted, it will no longer make good contact.

This problem is exacerbated by the fact that the connector types have no distinguishing marks to identify them. The only reason that this doesn't often come up is that the two varieties are generally used in different domains — 50 Ω in RF work — 75 Ω in video systems. Should you find yourself in a mixed environment — over-the-air television transmission comes to mind — I suggest adopting a color-coded or tagging standard for all Type N connectors so that this is less likely to happen. Better yet, select a different connector type for one or the other application.

Assembly Process

Some would consider it an advantage for UHF connectors that most experienced technicians can strip the outer and inner insulation and terminate a plug by eyeballing the dimensions, while the Type N requires careful measurement to assemble properly. Especially with full size cable, the Type N has an advantage in not needing soldering of the shield connection, a time-consuming and failure-prone part of the operation. I would call it about even — but others will have their own opinion, especially if in the field without a rule or stripping guide.

BNC COAXIAL CONNECTORS

The BNC connector is perhaps the second most popular RF coaxial connector type in Amateur Radio circles. It was developed in 1951 and is related to the Type N, but has a smaller outer structure and a snap-on "bayonet" mount rather than the screw-on backshell of the Type N and UHF (see **Figure 6.7**). Its size is a perfect fit for the smaller (RG-58/59/8X) cables, although versions are available for smaller (e.g. RG-174) and larger (e.g. RG-8) sized cables.

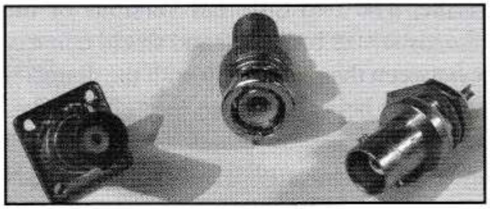

Figure 6.7 – The basic elements of the BNC coaxial connector series. On the left is a BNC flange mounting jack. On the right is a BNC through-hole mount jack and in the middle, a plug.

The connector was invented by Paul Neill of Bell Laboratories (of Type N fame) and Carl Concelman (inverter of the Type C connector, below) of Amphenol. BNC stands for Bayonet — Neill — Concelman. The flexibility of the small cables and the easy to attach mechanism of the bayonet backshell, make BNC cables ideal for patch panel use.

When originally designed, the connector was just made to match 50 Ω systems — in fact the inner pin and shield connecting portion is the same size as that of the Type N (in fact, a Type N plug can be inserted into a BNC socket). Later a 75 Ω version was added, although the 50 Ω is much more frequently encountered. So now there are both 50 and 75 Ω versions of the BNC connector. However, this time they were cleverly designed to be interchangeable without destruction or connectivity issues.

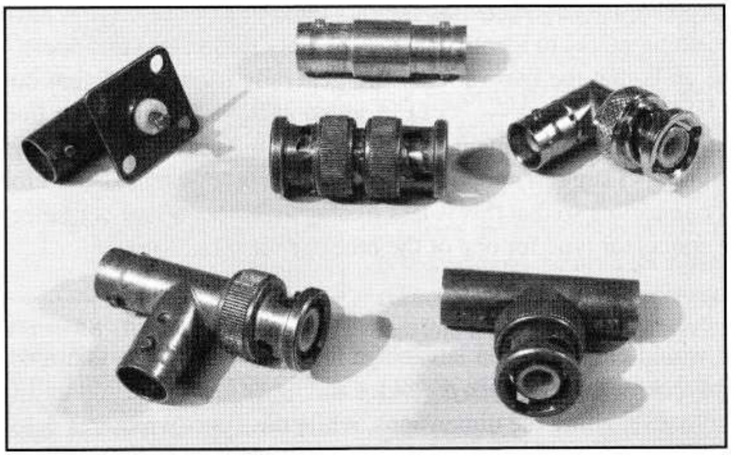

Figure 6.8 – Additional BNC series connector types. All the usual suspects are available.

The ratings of the BNC are quite similar to those of the Type N, since it has the same electrical dimensions, although its shielding is somewhat less effective so its upper frequency limit is around 2 GHz. In fact, the voltage rating may be higher because an insulating sleeve surrounds the center pin, making for a longer air path between conductors. All of the usual variations and special fittings for other connector types are available for the BNC (see **Figure 6.8**)

RCA CONNECTORS

The RCA connector, sometimes called a *phono* (for phonograph, an early music playback system) connector is probably more frequently encountered than any other type in Amateur Radio equipment because it is used for many other purposes besides terminating coaxial cables.

This connector was initially developed as an audio connector by the Radio Corporation of America to connect phonographs to amplifiers. Its structure is inherently coaxial with some jack configurations designed to ground shield connections directly to panels, so it has appeared as an RF connector as well. It is most frequently encountered as a receive antenna connector, especially in lower-priced equipment. As a transmit RF connector, it generally is found at power levels of 100 W or less, also in lower-priced gear.

A notable exception is the series of high-end Collins Radio SSB transmitters, receivers, and transceivers introduced in the late 1950s. These used RCA connectors for audio and RF interconnections. This highlights one aspect of this connector series — they are made in many forms. The RCA jacks on the Collins gear were of very high quality, using ceramic dielectric insulators. Some have been described as "RF rated" — a noble thought, but by no means a standard designation. Most usually encountered RCA connectors are not as well fabricated, but they can do the job if carefully selected and properly installed. A selection of different types is shown in **Figure 6.9**.

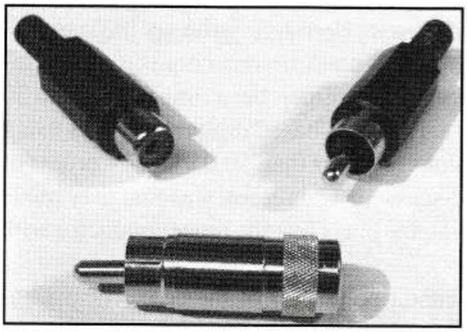

Figure 6.9 – Examples of RCA connectors. The plug at the bottom has a shielded backshell, more suited for RF use.

Unlike any of the previously described coaxial connectors, the RCA pair has no mechanical locking mechanism to retain the connections if subject to any pull on the cable. Since the BNC is similar in size, it may be a good alternative if that is likely to be an issue.

Since RCA connectors are so ubiquitous, it is worth mentioning that they are not, in my opinion, a very good choice for power interconnections. I have what is otherwise high-quality equipment that has RCA connectors as power accessory sockets and companion equipment with a panel jack designed to accept 12 V dc from an RCA plug. Of course if the equipment is on while making this connection, as soon as the pin strays to the outer conductor on insertion, it blows a fuse deep in the source equipment.

OTHER COAXIAL CONNECTORS

While the connectors described in the earlier sections are those most frequently encountered as coaxial cable terminations, there are many other types that will be encountered from time to time. I think it's worth a bit of ink to let you know something about them.

SMA Coaxial Connectors

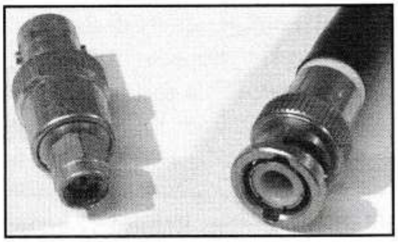

Figure 6.10 – A handheld transceiver BNC connected antenna compared to an SMA connector.

SMA connectors — short for *sub-miniature type A* — are found on a significant number of handheld transceivers as antenna connections (see **Figure 6.10**). These small threaded coax connectors are designed for operation well into the microwave region — typically to 13 GHz, but some varieties go much higher. Outside of amateur handheld antennas, they are mostly found used as interconnections between chassis of microwave equipment, often on the ends of miniature semi-rigid coax.

Their electrical properties are quite suitable for amateur handheld antennas, however, they are typically rated for 500 to 600 connect-disconnect cycles, so care should be taken with their mechanical properties. The alternate connector in this application is the BNC, which is my preference for a number of reasons, especially durability. An SMA handheld antenna is shown in Figure 6.10.

Type F Coaxial Connectors

Anyone who has ever hooked up a modern television set has encountered the Type F connector. This is a great mass-market connector — as inexpensive as possible for millions of users. I remember a time (more than 30 years ago when I was working on early cable-based digital networks using cable TV components) when the plugs cost less than $0.05/unit — orders of magnitude less than some we've discussed. The reason was that they made large quantities and the center conductor pin is just the inner conductor wire of the coax (see **Figure 6.11**).

With a center pin that is the center conductor of the coax, the assembly is trivial. The cable is stripped — fancy tools that do it in one step are available — the shield is crimped to the outer conductor shell of the connector (while soldering is possible with some varieties, anyone using these

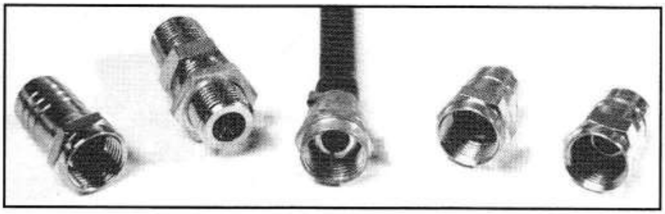

Figure 6.11 – Examples of Type F connectors as used in cable TV installations.

frequently uses a crimp tool), and you're done. The disadvantage is that the inner conductor must be solid wire and that it has no corrosion resistance. Special waterproof versions are available for outdoor use.

Type F connectors are not often encountered in Amateur Radio, with one notable and worthwhile exception. TV-type coax cable, often described as RG-6 type but made with aluminum foil shields, is actually pretty good coax. Modern cable TV systems operate well into the UHF range, so they are designed for low loss. They are often available at low cost, or even prewired within a house begging to be used for amateur antenna access. The problem is that most connectors of the type used in Amateur Radio can't deal with connections to the aluminum foil shield. Often the best approach is to either use the F connectors on the cable, or crimp some on and then use between series adapters (see below) to transition to the connector that will mate with your termination.

TNC Coaxial Connector

The TNC (Threaded — Neill — Concelman) is essentially a BNC with a threaded rather than a bayonet backshell. It thus is essentially a smaller Type N connector, but with the inner connecting surfaces the same size. It extends the 2 GHz frequency limit of the BNC up to 10 GHz or higher, and takes less panel space. It is a natural for heavy duty requirements with smaller cables.

Type C Coaxial Connector

This connector, honoring just Concelman, is essentially a TNC for larger cables. It is about the size of a Type N, but with a bayonet backshell and larger diameter inner components. It would seem to be a natural for patch panels for those running high power, but it has never caught on in the amateur community, that I'm aware of. In fact I hardly see them anywhere, which is too bad.

Motorola Connectors

These are the coaxial connectors designed and almost exclusively used to connect automotive antennas to automotive broadcast radios (see

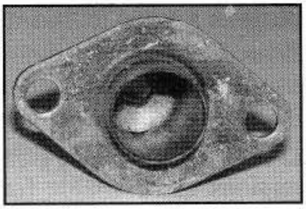

Figure 6.12 – A Motorola jack. These are commonly used in automotive broadcast receivers as antenna connectors.

Figure 6.12). They became popular for amateur use in the 1950s, at which time most HF voice used full carrier amplitude modulation. A number of companies offered tunable converters, designed for steering-column mounting, that translated amateur HF signals to the upper broadcast band.

The auto antenna, with its Motorola connector, plugged into the converter and a patch cable went from the converter to the antenna jack of broadcast radio. Usually, each had a Motorola connector. Apparently because of the converter connection, I have seen these connectors also used in home station VHF converters.

Type HN Coaxial Connector

The Type HN connector is similar in design and layout to the Type N, but is a bit larger. It has higher voltage breakdown and power ratings than the type N. It is a good match to the cables that are somewhat larger than RG-8 size and finds some amateur use in high power stations with long cable runs.

General Radio Connectors

The General Radio type 874 or GR connector is most frequently seen on test equipment made by that company, later GenRad through the 1970s. It is still popular in university physics and engineering laboratories because it is unique (to my knowledge) as a genderless coaxial connector. Both the inner and outer connection arrangements are made of four spring leaves in two pairs, with one pair of each separated further than the other (see **Figure 6.13**). In this way, by rotating 90 degrees, any two connectors can be mated resulting in a solid, matched connection without the need for gender converters (see **Figure 6.14**). The downside is that there is no locking arrangement, so it's not particularly suitable for long-term connections, but it's perfect for lab use.

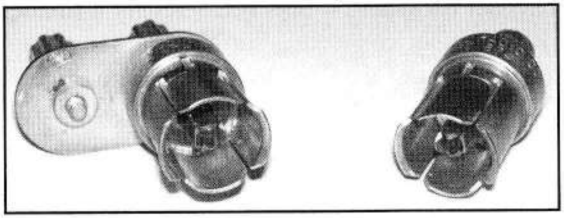

Figure 6.13 – The GR-874 genderless coaxial cable connector — popular for use with laboratory test equipment for many years.

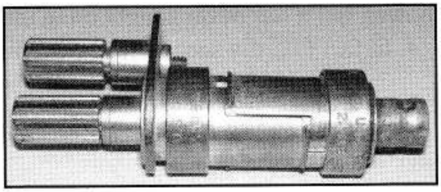

Figure 6.14 – A pair of GR-874 connectors mated.

Between Series Adapters

Virtually every coax connector type of either gender is available in a short adapter with a connector of every other type. These adapters allow equipment with one type of coax connector to be attached to cables

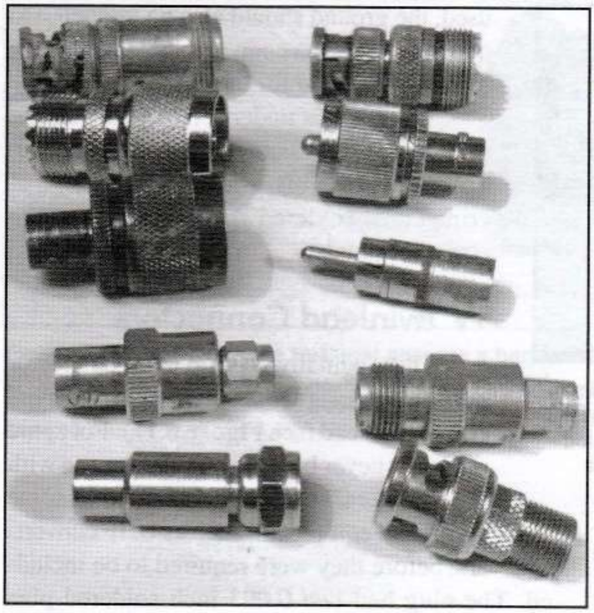

Figure 6.15 – Representative example of coaxial between series adapters.

terminated with another type. A collection of representative examples is shown in **Figure 6.15**.

These are best used for temporary test or measurement applications, although some find use in long-term applications. Every additional part, especially connectors, reduces the overall system reliability by adding another potential failure point.

There are a few exceptions, in my opinion, to this rule. One is that some equipment supplied with UHF jacks makes use of lower-cost parts in which the fingers give out quickly. While the best solution is to replace them with higher-quality units, sometimes that is not feasible and will likely void any equipment warranty. An adapter with a UHF plug and Type N socket is perfect for this situation, especially if the cable can be equipped with a Type N plug. This works best if done before the spring contacts get limp.

BALANCED LINE INTERCONNECTIONS

Considering how much longer balanced transmission line has been in use than coaxial cable, you might think there would be at least as many choices of connector. As it happens, that's not the case, but a few standards have evolved over the years.

Screw Terminals — Most Amateur Radio equipment designed for balanced transmission line offered screw terminals, as did most television receivers until the shift to coaxial cable interconnection made TV twinlead almost obsolete.

Screw terminals have the advantage of being nearly universal, in that virtually any reasonably sized wire size can fit under the screw head and may stay in place and not short during tightening. Still, the connection is not a terribly reliable one.

Most amateur receivers actually had three screw terminals. Two were for a balanced antenna, the third was a chassis ground (see **Figure 6.16**). The chassis ground should be connected to the power system safety ground, since many receivers had two-wire ac plugs. If an unbalanced antenna were

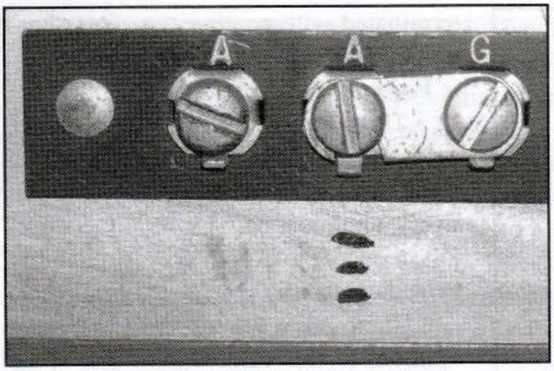

Figure 6.16 – Terminal strip on a 1950s communication receiver. There are two terminals for a balanced transmission line. If an unbalanced antenna is used, the ground terminal is jumpered to one of the antenna terminals as shown.

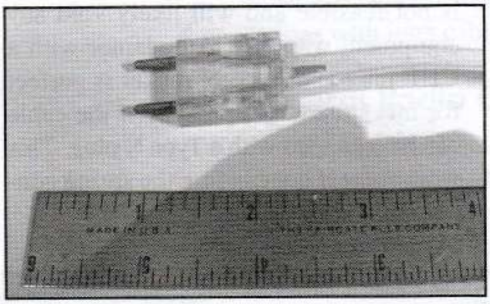

Figure 6.17 – A plug designed to terminate 300 Ω TV type twinlead.

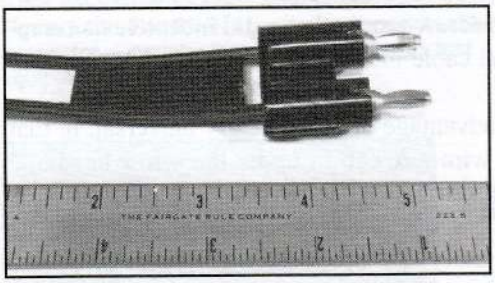

Figure 6.18 – The double banana plug, designed for low frequency balanced test equipment. It is a great fit with 450 Ω window line.

used, the ground should also be jumpered to one of the antenna connections, sometimes a link was provided for the purpose.

While wires under screw heads can be particularly unreliable, in my experience, a better solution is available in the use of crimped or soldered spade lugs under the screw terminals.

TV Twinlead Connectors

There actually was a type of connector especially designed for use with TV type 300 Ω twinlead (see **Figure 6.17**). For some reason, it never became popular for use with TV receivers, but was encountered in some peripheral equipment, such as UHF converters, before they were required to be included. The plug had two 0.093 inch soldered pins spaced at 0.486 inches. This is the same spacing as the pins of an FT-243 crystal holder, so would fit into a crystal socket, or two pins (with one in between) of an octal tube socket.

Double Banana Plug and Jack

A connector type that is not especially designed for the purpose is the double banana plug. A banana plug is a robust spring loaded plug of a size that happens to fit nicely into the center conductor of a UHF socket. A popular test equipment connectivity arrangement is two banana plugs within a common insulating housing spaced 0.75 inches (see **Figure 6.18**). This spacing is very convenient for use with most window-type transmission line.

Jacks are available in various forms, including panel mounting types and even combination plug and jack sets that also have screw terminals to clamp wire leads. Another handy arrangement is a kind of terminal that is sometimes called a "five-in-one." This mounts through a single hole into a panel and has a banana socket surrounded by a nut arrangement that surrounds a threaded surface that includes a hole for wire or tip insertion. Thus it can accept a banana plug, a spade lug, a wire bent around the threads or a wire pushed through the perpendicular hole. If two are installed on ¾ inch centers, it will also accept the double banana plug, but can deal with loose wire ends, or lugs as well (see **Figure 6.19**).

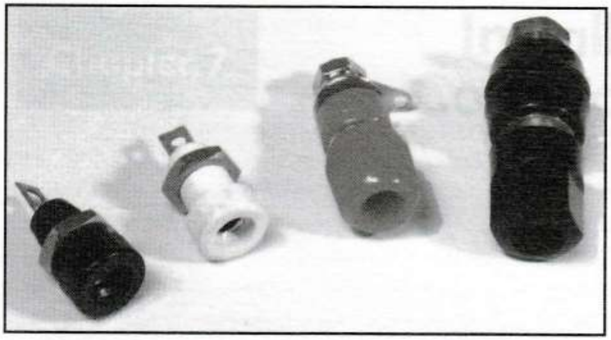

Figure 6.19 – A five-in-one terminal accepts a banana plug, loose wire, or terminal lugs. If two are spaced at 0.75 inches, they will also accept the handy double banana plug.

Notes

[1] J. Kraus, W8JK (SK), *Electromagnetics*, McGraw-Hill Book Company, New York, 1953.

Chapter 7

Installing Coaxial Connectors on Cable

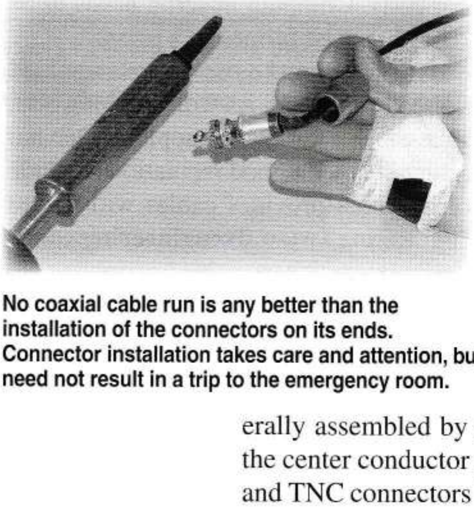

No coaxial cable run is any better than the installation of the connectors on its ends. Connector installation takes care and attention, but need not result in a trip to the emergency room.

The installation of transmission line connectors is looked on by some with dread, since it is both important and requires special care and moderate precision in order to do the job properly. Of course, any connector installation will be no better than the quality of the connector being used. Connector failures are surely the leading cause of problems with coax cables, although there are certainly other ways to ruin a coax run.

Standard or Crimp Connectors?

The "standard" coaxial cable designs are generally assembled by soldering. In the case of the PL-259 UHF plug, both the center conductor and shield are soldered, while standard Type N, BNC, and TNC connectors have a center pin that must be soldered and an integral tightened clamp for shield connection. TheType F television coax connector is a bit different. The center conductor of the coax is used directly as the center pin, while the shield is almost always crimped.

Alternative connector types are available in each series that use crimping for connection. In many cases, both the center conductor and shield are crimped, while some have a crimped shield and soldered inner conductor. I have seen no evidence that a crimped connector is inferior to a soldered one. With either type, poor assembly practices can yield poor results.

The issue with crimped connectors is that they require a crimping tool that is matched to the sizes of the crimped surfaces. Most quality crimping tools are relatively expensive and fit a limited series of cables and connectors. Some have dies that include multiple sizes of crimping holes, making

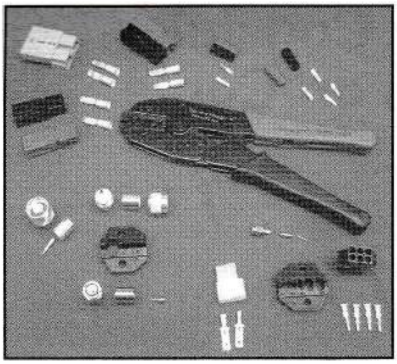

Figure 7.1 — The Andy-Crimp Pro™ tool, showing some of the available interchangeable dies along with a sampling of the different coax and power connectors that it can crimp with the right dies.

them more useful. The Andy-Crimp Pro™ tool — available from High Sierra Antennas (**www.hamcq.com**) or Quicksilver Radio (**www.qsradio.com**) — has available interchangeable dies that will also fit in other similar tools, such as the Powerpole crimper sold by West Mountain Radio (**www.westmountainradio.com**). In addition to the two die sets available for coax connections, there are sets available for the multiple sizes of popular Anderson Powerpole dc power connectors, and another set for Molex connectors and other crimp terminals (see **Figure 7.1**).

The crimping process is very effective if the proper tools and connectors that match the cable are all at hand. Some vendors have offered crimping arrangements for standard connectors, but these can result in fractures of the solid connector body. Crimping material needs to be pliable enough to be able to be formed in a press.

Another possibility is to make use of preassembled coaxial cables with the connectors already installed. As with anything else, the quality will vary depending on the capabilities of the supplier, often but not always indicated by the price. Suppliers with good reputations include ABR Industries (**www.abrind.com**), which provides cables with either UHF or N connectors, and DX Engineering (**www.dxengineering.com**), which provides cables with UHF connectors. These cable assemblies are available in different lengths and made from different types of cable, for a cost that is not much higher than the cost of the cable plus the connectors if you were to make them yourself. Both offer custom lengths and free shipping if a sufficient quantity is ordered.

The sections that follow will focus on installing standard coaxial connectors, with the exception of one representative sample of an Amphenol crimp-type UHF plug. If you decide to use crimp-type connectors and associated tools, the manufacturer should provide the cable and connector specific trimming dimensions. While the crimp diameters have moved toward industry standards, the details, especially of ferrule length, may result in different stripping lengths. Look at the website of LNL Distributors (**www.lnl.com/howto.htm**) to see some representative examples.

INSTALLING UHF SERIES COAXIAL CONNECTORS

UHF series connectors include cable-mounted plugs (the PL-259 type is the original pattern, although there have been others through the years) and chassis-mounted sockets (the SO-239 — note the difference in the center digit — being the standard type, although there are others here as well).

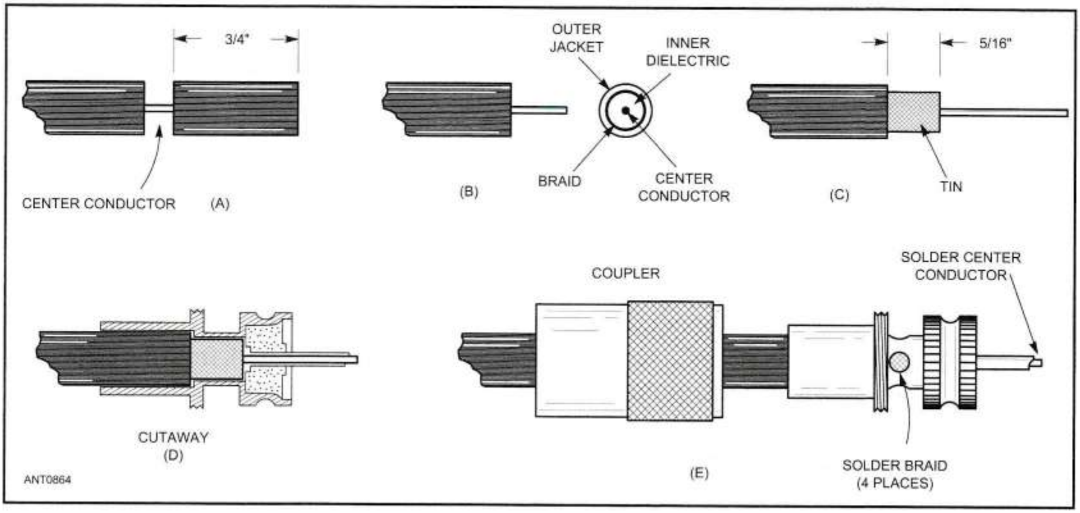

Figure 7.2 — The PL-259 or UHF plug is the most popular coax connector in Amateur Radio. Here the steps required for installation on RG-8/11/213 size (0.405-inch outer diameter) are shown. See the main text for a description of each step.

Most of the problems people have assembling these are with the plugs, so we'll start there.

UHF Plugs with Standard Cable

The standard PL-259 plug is designed to fit the common larger sizes of coax, such as RG-8, 11, 213 and RG-214, with an outer diameter of 0.405 inches. Note that this includes both 50 and 75 Ω types. Adapters are available for smaller cable types, discussed below. The installation of such coax into a PL-259 is illustrated in **Figure 7.2**.

In Step (A), the coax is stripped through the outer jacket, shield, and dielectric down to the center conductor starting ¾ of an inch back from the end, as shown. The hard part is cutting through all those layers and not also cutting some of the center conductor strands. This looks easier than it is with normal hand tools, but it is dramatically easier using a special tool. One by Ripley Brothers, the Cablematic UT-8000, works like a handheld pencil sharpener and makes a remarkably clean and straight cut.

By using a sharp knife, and going slow with a lot of patience, it is possible to work your way through the layers. Special tools are available that make this quite easy. It is also possible to use a tubing cutter with a sharp blade, at least to get through the outer jacket. Often, by the time you get all of that off, you will find that you have lost a strand or two of the center conductor. The purist would start over; you may decide to define your own quality control standard and decide to allow the loss of a strand, as it will not make much difference in the scope of things. (The perfectionist will certainly start over; it's always a good idea — one of many reasons to start with a somewhat longer cable than you think you might need.)

As shown in (B), remove the cut pieces together, if possible. Sometimes a bit of a tug with electricians' pliers is helpful. Once all of the dielectric is removed, carefully check, with a magnifier if needed, to make sure that there are no strands of the shield hanging about and that the center conductor strands aren't nicked.

In step (C), score the outer jacket $\frac{5}{16}$ of an inch back from the earlier cut without nicking the shield, and remove that section of the jacket. Sometimes a longitudinal cut from the score to the end makes it a bit easier to remove. Now *very lightly* tin the exposed shield. This is a very tricky step since:

◆ The shield has to be hot enough for the solder to flow into the wires without melting the dielectric. This is a particular challenge with foam dielectric type cables.

◆ The tinned shield needs to be almost the width of an untinned tightly woven shield in order to fit inside the barrel of the connector. This is a difficult step, since once the solder flows, it tends to wick into the shield. Recheck for any loose strands of shield that could short out the cable later. Some installers prefer to lightly tin the center conductor; this is not a bad idea, but also needs to be done sparingly to make sure the center conductor still fits within the pin. This can be checked from the pin side of the connector.

Now slide the coupling ring onto the cable, making sure that it is oriented in the proper direction. If doing both ends of the cable, this step can be delayed until the second end — just don't forget to do it. There is almost nothing less useful than a PL-259 without a coupling ring (or perhaps even worse, one remembered but installed backwards).

In step (D), the coax cable is inserted into the body of the connector. Again, check that the coupling ring hasn't fallen off. The connector should be installed with a clockwise turning motion in order that the threaded portion of the body screws onto the outer jacket. This can be a bit of an effort. Sometimes inserting the connector into the jack of a T connector will provide additional purchase, if the sharp points are engaged. Alternately, there are tools available, such as the DX Engineering UT-80P, that provide a good grip without chewing up the connector with pliers. The connector is on far enough if the center conductor is protruding from the pin and the shield is beyond the four solder holes. Recheck that the coupling ring is still there and oriented properly — this is your last chance for an easy correction.

Step (E) is the last part of the assembly process — soldering everything in place. First, support the cable and connector solidly, preferably in a way that doesn't draw off the heat from the connector body. Make sure that the cable weight is not pulling it downward, or else the heat from soldering will allow the dielectric to melt and the inner conductor to migrate through.

Use a heavy soldering iron of at least 150 W with a tip that will fit into the groove containing the solder holes. Since heat flows upward by convection, apply the iron to the bottom of the area and start flowing the solder into the top hole when the braid reaches melting temperature. Continue with the other holes, turning the cable carefully to reach the bottom hole. Let the cable and connector cool significantly before moving to the next step.

Check that the cable is not shorted by checking the resistance between the connector body and the inner conductor protruding from the pin. It should indicate infinite resistance, unless there is something providing connectivity at the far end. Now insert the connector in the vise with the open end of the pin pointing below the horizontal. Apply the soldering iron to the outside of the pin and gently apply solder to the junction of the center conductor and pin until it fills the void. When heated, the solder will flow between the inside of the pin and the center conductor by capillary action. If the pin were pointed the other way, it would flow down the outside of the pin and make it too thick it fit.

Let it cool and recheck the resistance. Trim the center conductor flush with the center pin and use a fine-tooth file to remove any solder from the outside of the pin. You're finally done — congratulations!

Another Opinion

Noted Amateur Radio contester Tim Duffy, K3LR, operator of one of the major multi-multi (multiple operator, multiple simultaneous transmitter) contest stations in the US, has another approach to assembling PL-259 plugs with the larger size cable, which he credits to William Maxson, N4AR. Tim has assembled an extreme antenna farm, visible from US Route 80 in Western Pennsylvania, and has terminated more coax cables than most people. Space considerations don't allow inclusion of all of his material, which can be seen on his website at **www.k3lr.com/engineering/pl259/**.

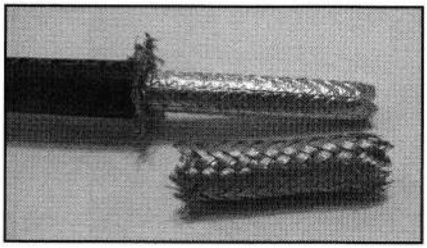

Figure 7.3 — After removing the outer jacket with a sharp knife, Tim carefully removes all but about ¼ inch of braid and fans out the remainder. [ROBERT BASTONE, WC3O]

Tim first measures a connector against the RG-8 or other 0.405-inch cable and removes just the outer jacket with a sharp knife. He then pulls the braid back all the way by fanning it out (as shown in **Figure 7.3**) and wraps three to four turns of Scotch brand 88 black electrical tape around the dielectric up against the fanned-out braid (**Figure 7.4**). Using wire strippers, he removes the remaining dielectric from the center conductor (**Figure 7.5**). Then he installs the fully assembled PL-259 onto the center conductor and over the 88 tape. The back of the PL-259 body should rest on the fanned-out shield. He then solders the center conductor.

This done, he fans out the shield and cuts it to ¼-inch long and folds it over the back of the PL-259 (**Figure 7.6**). He solders the shield all the way around to the back of the PL-259 body. While the back is still hot, he wraps two turns of 88 electrical tape around the soldered shield to seal it all (**Figure 7.7**).

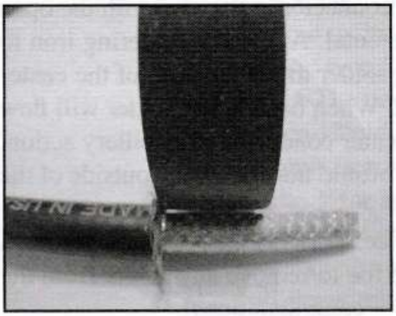
Figure 7.4 — Tim wraps three to four turns of electrical tape around the dielectric up against the fanned-out braid. [ROBERT BASTONE, WC3O]

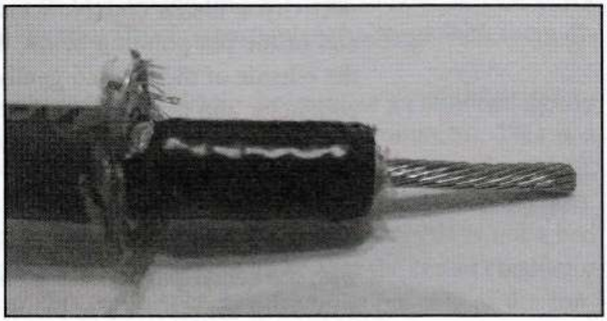
Figure 7.5 — The dielectric around the center conductor is removed from the area in front of the tape. [ROBERT BASTONE, WC3O]

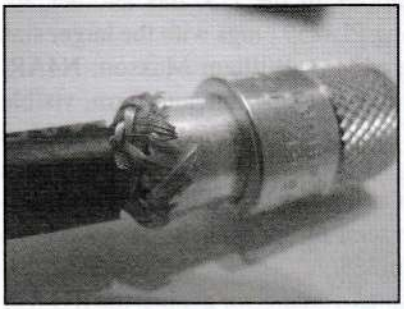
Figure 7.6 — Tim screws the connector onto the cable until it is fully seated, trims the shield, and solders the center conductor. [ROBERT BASTONE, WC3O]

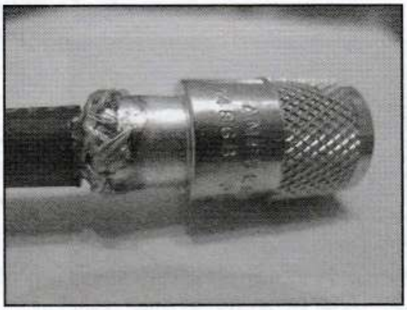
Figure 7.7 — The shield is soldered to the outside of the rear of the connector body and will be wrapped with tape while still warm to provide a seal. [ROBERT BASTONE, WC3O]

UHF Plugs with Smaller Cable

The PL-259 can easily be adaped to smaller cable types such as RG-58, 59 or RG-8X through the use of an adapter (some call it a reducer) that fits into the threaded portion of the plug. In many ways, this is easier to accomplish than the connection of the larger cables, even though the steps are similar. The cable is somewhat easier to handle and easier to strip since the dielectric is thinner and doesn't need to be cut through in one step. The shield also has less thermal mass to deal with while soldering. The process is illustrated in **Figure 7.8**.

Just as with the coupling ring and the larger cables, it is very important to slide both the coupling ring and the adapter onto the cable before progressing very far. It is also very important to support the connector and cable during soldering so it won't bend and allow the center conductor to migrate through the dielectric.

In Step 3 of Figure 7.8, it is important to trim the shield so that it doesn't extend over the threaded portion of the adapter; otherwise strands will jam the threads. Following Step 4, after the connector cools, make sure that the shield is soldered to both the adapter and the connector body. If it isn't properly soldered, you will be able to turn the adapter in the body. If that happens, just reheat the rear of the connector body and apply more solder into the holes when it's ready to flow. As with any connector installation, check for shorts with an ohmmeter.

83-1SP (PL-259) Plug with Adapters (UG-176/U or UG-175/U)

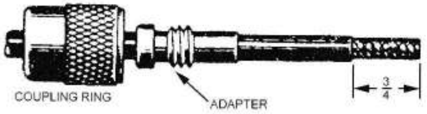

1. Cut end of cable even. Remove vinyl jacket 3/4" - don't nick braid. Slide coupling ring and adapter on cable.

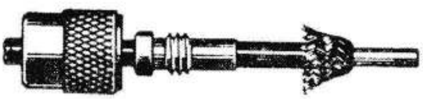

2. Fan braid slightly and fold back over cable.

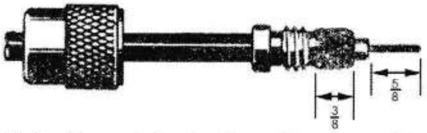

3. Position adapter to dimension shown. Press braid down over body of adapter and trim to 3/8". Bare 5/8" of conductor. Tin exposed center conductor.

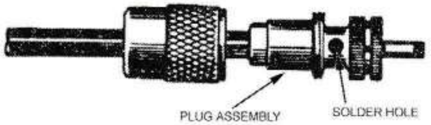

4. Screw the plug assembly on adapter. Solder braid to shell through solder holes. Solder conductor to contact sleeve.

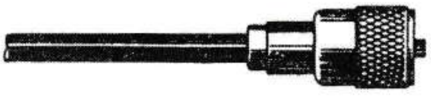

5. Screw coupling ring on plug assembly.

ARRL0908

Figure 7.8 — The procedure for assembly of UHF plugs with smaller coaxial cable. The UG-175/U adapter is for RG-58 size cable with 0.195-inch outer diameter. The UG-176/U adapter is for RG-59 size cable with 0.242-inch outer diameter, including RG-8X and RG-62. [AMPHENOL ELECTRONICS, RF DIVISION]

UHF Plugs with Crimped Connections

PL-259 compatible connectors are available with crimped rather than soldered connections for both large and small cable sizes. Each requires different size ferrules and different crimping tools. Some types have a center conductor that is soldered while the shield is crimped, while others have crimp arrangements for both conductors. If you use a type with a crimped center pin, make sure the center conductor extends all the way to the bottom of the pin, or the connection won't be solid. (See **Figure 7.9** for assembly details.)

UHF Sockets

The standard SO-239 panel socket mated with the PL-259 is a flange mount arrangement with a center pin and only four mounting holes for shield connection. The usual practice is to mount the connector using four machine screws, lock washers, and nuts. This provides a connection to a metal panel. If it is desired to continue with coax inside the panel, the usual practice is to connect the shield, via as short a connection as possible, to a solder lug under one (or more) of the mounting nuts.

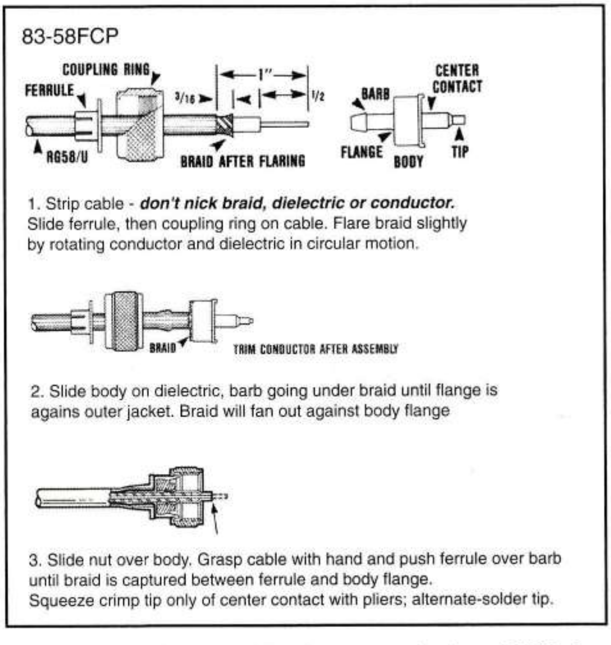

1. Strip cable - ***don't nick braid, dielectric or conductor.*** Slide ferrule, then coupling ring on cable. Flare braid slightly by rotating conductor and dielectric in circular motion.

2. Slide body on dielectric, barb going under braid until flange is agains outer jacket. Braid will fan out against body flange

3. Slide nut over body. Grasp cable with hand and push ferrule over barb until braid is captured between ferrule and body flange. Squeeze crimp tip only of center contact with pliers; alternate-solder tip.

Figure 7.9 — The procedure for assembly of one type of crimped UHF plug with smaller coaxial cable. The 83-58 family of connectors are available in multiple versions for different cable types with different ferrule diameters requiring different crimp dies. Some have a soldered center conductor, some (as shown) have an inner conductor that can be crimped or soldered. [AMPHENOL ELECTRONICS, RF DIVISION]

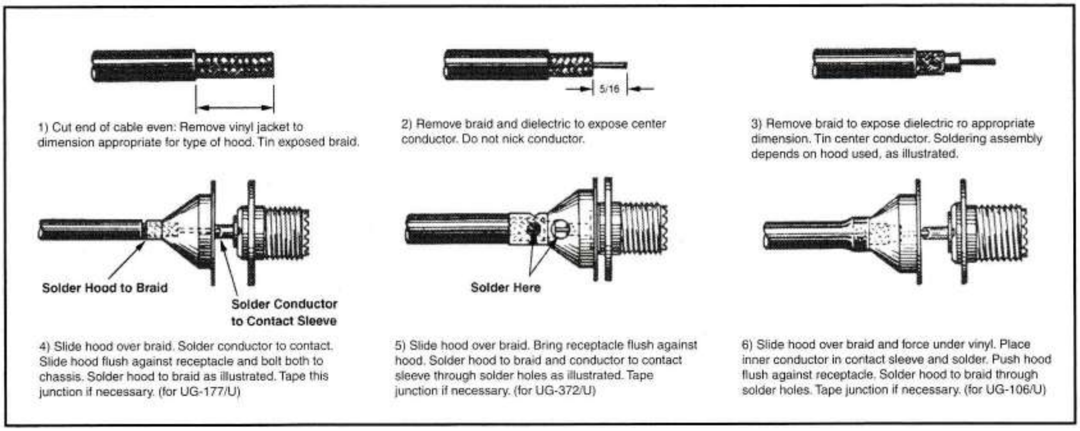

1) Cut end of cable even: Remove vinyl jacket to dimension appropriate for type of hood. Tin exposed braid.

2) Remove braid and dielectric to expose center conductor. Do not nick conductor.

3) Remove braid to expose dielectric ro appropriate dimension. Tin center conductor. Soldering assembly depends on hood used, as illustrated.

4) Slide hood over braid. Solder conductor to contact. Slide hood flush against receptacle and bolt both to chassis. Solder hood to braid as illustrated. Tape this junction if necessary. (for UG-177/U)

5) Slide hood over braid. Bring receptacle flush against hood. Solder hood to braid and conductor to contact sleeve through solder holes as illustrated. Tape junction if necessary. (for UG-372/U)

6) Slide hood over braid and force under vinyl. Place inner conductor in contact sleeve and solder. Push hood flush against receptacle. Solder hood to braid through solder holes. Tape junction if necessary. (for UG-106/U)

Figure 7.10 — The official procedure and dimensions for assembly of a shielding hood for SO-239 flange mount jacks. The different part numbers are for different cable types.

This arrangement is obviously not satisfactory if high frequency shielding is required. Using a shield hood can result in a properly shielded connection. (Details are shown in **Figure 7.10**.)

In addition to flange mounting, one-hole mount sockets are available. These are constructed much like the threaded PL-258 barrel connectors and are secured with nuts that fit the backshell threads. A center pin solder connection is provided inside the panel and if a shield connection is required, it is often made using a large solder lug that fits around the connector body. To extend a shielded connection with a one-hole mount, a PL-258 type double female connector with matching nuts can be used. The coax inside the panel is then terminated with a PL-259 plug.

INSTALLING TYPE N COAXIAL CONNECTORS

The Type N coaxial connector series is superficially similar in size and operation to the UHF series, but it corrects the major deficiencies of the earlier UHF type:

◆ Unlike the UHF, the Type N connectors maintain a constant characteristic impedance through the connector.

◆ The construction is such that the shield connection is not dependent on a tight backshell. Of course, if the backshell is removed far enough, the connectors can be pulled apart and break contact.

◆ The gaskets of the Type N prevent water penetration, a major factor in the deterioration of coax cables.

A Type N plug is actually easier to assemble than the UHF counterpart.

So why don't more amateurs switch over to Type N? Simple: They are more expensive (typically around \$6, compared to \$2 for a high-quality plug) and many installers don't know how to deal with them. The more expensive part is relative — what's it worth to save a trip up the tower in the winter?

Select the Correct Connector

Type N connectors come in 50 and 75 Ω versions. Be sure to select the right series, and get the right connector for your cable type if you are concerned about maintaining impedance. The insidious problem with having 50 and 75 Ω versions is that they look identical, and can be made to mate with each other, to a degree. As with the RG-8 (50 Ω) and RG-11 (75 Ω) type cables themselves, the connectors have the same outside and shield contact diameters. In order to maintain the constant impedance, the 75 Ω N connectors have a smaller diameter center conductor pin.

If a 50 Ω plug is inserted into a 75 Ω socket, the socket's inner conductor will be spread open too far, making it unreliable for use any later use with its proper 75 Ω a plug. While inserting a 75 Ω plug into a 50 Ω socket will not cause damage, it will result in an uncertain center conductor connection. If the pin is in perfect alignment, there will actually be no contact at all.

Installing a Type N Plug or Socket

Unlike the standard UHF connector arrangements that just have cable plugs and panel jacks, the Type N is available with similarly installed cable plugs and cable jacks. This makes coaxial "extension cords" feasible without the need for a double female adapter.

The installation instructions for a Type N plug or cable socket are shown in **Figure 7.11**. To add clarification here, we expand upon these steps using an Amphenol connector, part number 82-202-RFX, designed for a number of 0.405-inch diameter 50 Ω coax cables, including RG-8 and RG-213. Note that the suffix RFX indicates it is a commercial version rather the 82-202, which is the full military type at about three times the price. (Visit their website at **www.amphenolrf.com** to find the right kind for your cable.) A step-by-step installation process follows.

◆ *Lay it out* — Slide the nut, washer, and gasket over the cable jacket — in the correct order.

◆ *Remove the outer jacket* — Measure $\frac{9}{16}$ inch of the outer jacket and carefully cut and remove it with a razor knife, without nicking the braid.

◆ *Comb out the braid* — Using an awl or a metal comb, straighten out the shield wires and slide on the clamp until it stops on the jacket, as shown in **Figure 7.12**.

◆ *Strip the dielectric* — Remove enough of the dielectric to leave

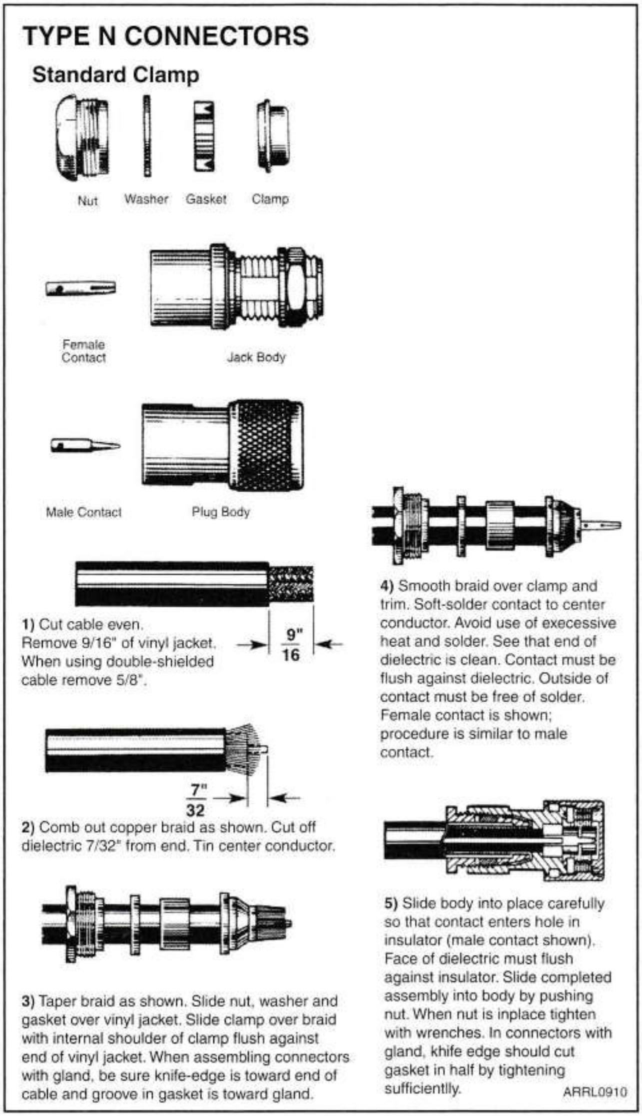

1) Cut cable even. Remove 9/16" of vinyl jacket. When using double-shielded cable remove 5/8".

2) Comb out copper braid as shown. Cut off dielectric 7/32" from end. Tin center conductor.

3) Taper braid as shown. Slide nut, washer and gasket over vinyl jacket. Slide clamp over braid with internal shoulder of clamp flush against end of vinyl jacket. When assembling connectors with gland, be sure knife-edge is toward end of cable and groove in gasket is toward gland.

4) Smooth braid over clamp and trim. Soft-solder contact to center conductor. Avoid use of execessive heat and solder. See that end of dielectric is clean. Contact must be flush against dielectric. Outside of contact must be free of solder. Female contact is shown; procedure is similar to male contact.

5) Slide body into place carefully so that contact enters hole in insulator (male contact shown). Face of dielectric must flush against insulator. Slide completed assembly into body by pushing nut. When nut is inplace tighten with wrenches. In connectors with gland, khife edge should cut gasket in half by tightening sufficientlly.

Figure 7.11 — The official procedure and dimensions for assembly of a standard Type N plug or cable jack.

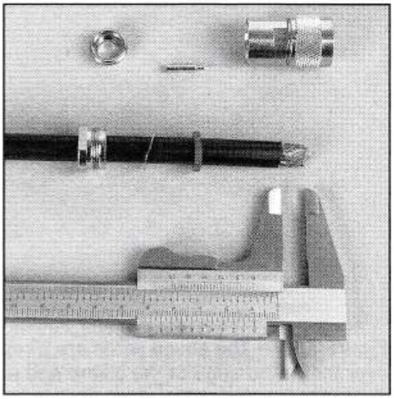

Figure 7.12 — The nut, washer, and gasket are in place on the cable jacket, in the correct order. The right amount of outer jacket is removed, and the shield wires are combed out straight.

$\frac{7}{32}$ of an inch between the end of the dielectric and the end of the inner conductor.

◆ *Trim the shield wires* — Use sharp diagonal cutters to trim the shield wires to make them just the right length to cover the round portion of the clamp (**Figure 7.13**).

◆ *Tin the center conductor and solder the pin* — Lightly tin the center conductor. Check to make sure it fits in the pin, and lightly dress it with a fine file, removing enough solder so that it fits fully into the pin. The bottom of the pin should rest against the dielectric and, with the pieces pushed together (as shown in **Figure 7.14**), the pin should be flush with the end of the housing. Now is the time to adjust as needed, then solder the pin and remove any solder on the outside of the pin with the file or emery cloth.

◆ *Final assembly* — Insert the pin into the connector body dielectric by pushing on the cable. Turn the nut until the threads start. Hold the connector

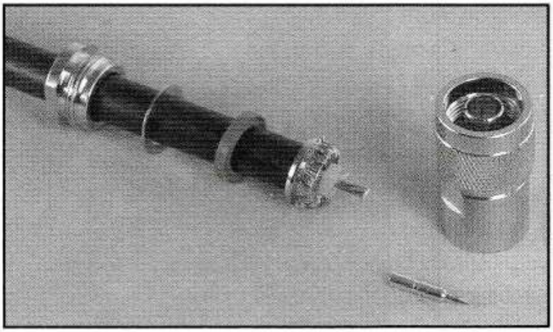

Figure 7.13 — The shield wires are trimmed to fit over the round edge of the clamp and enough of the dielectric is removed to allow the pin to fit flush.

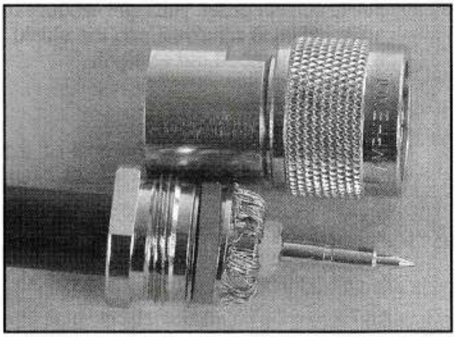

Figure 7.14 — Lightly tin the center conductor, make a final check of dimensions, and then solder on the pin.

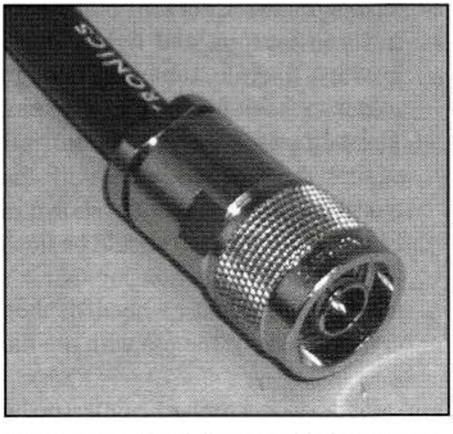

Figure 7.15 — The fully assembled connector.

body with a wrench and turn the nut with another wrench until tight. The pin should be about flush with the end of the coupling, the connector should not easily turn on the cable, and you should not be able to pull it off the cable (**Figure 7.15**).

Type N Jacks

As previously noted, the Type N offers cable jacks that assemble in the same manner as cable plugs. They also offer panel jacks in a few configurations. There are types that are similar to the open frame SO-239 panel jack, but there are also panel jacks that maintain the characteristic impedance and shielding integrity, more appropriate for critical RF applications. These assemble in the same manner as the other plugs and jacks, and are best installed in the panel after assembly.

Some Observations

Note that with the Type N, we have avoided the step of soldering to the connector body. This avoids the risk of melting the cable dielectric, a frequent problem with UHF connectors. Also, it is possible to modify the procedure slightly and solder on the pin, push the cable with pin attached through a small bulkhead hole, and then assemble the connector on the other side. With a UHF PL-259 connector, you need to do the soldering on the far side of the wall.

Unless there is a lot of working room, it is not recommended to replace UHF connectors on equipment with Type N. Adding a UHF to Type N adapter to a UHF connector, however, especially when the UHF is still new, can avoid some of the typical problems of loose coax connections later on.

For cables of the RG-58, 59, and 8X variety, the BNC connector is a good choice. It offers many of the benefits of the Type N in a smaller, bayonet-attached package. Type N connectors are also available for virtually all types of coax the amateur is likely to encounter, although they will be harder to find and may not be available in the less-expensive commercial series. A BNC connector with a BNC to N between series adapter may be a good solution.

Type N Plugs for the Dedicated UHF Plug User

It is no wonder that hams are wedded to the ubiquitous PL-259 UHF series plug, since most HF and a large portion of VHF equipment provides matching UHF sockets for connection to antennas. However, there is a line of Type N plugs that assemble in almost the same way as a PL-259 plug. In fact, as seen in **Figure 7.16**, the shield is connected in exactly the same way, whether the cable is RG-8 or RG-8X. The same UG-176 adapter for RG-59 or similar sized coax that fits a PL-259 also fits in this Type N connector, as does the UG-175 for RG-58.

The same techniques that are used to prepare either size cable for a PL-259 are also used here. The same tools, such as the Ripley UT-8000 jacket and dielectric stripper, can be used to prepare the larger cables, or the traditional knife and pliers approach will work as well for this connector.

The connector comes in two pieces: an inner body where all the soldering takes place that includes a captive pin, and a housing that includes the backshell and the shield connection arrangement. As with a PL-259, this Type N can be disassembled, cleaned out, and reused. The usual Type N includes a gasket that is destroyed upon first assembly, and those can't be found either.

With the backshell removed, the soldered connector interior has a diameter of 0.6 inches, significantly less than assembled Type N at 0.83 inch,

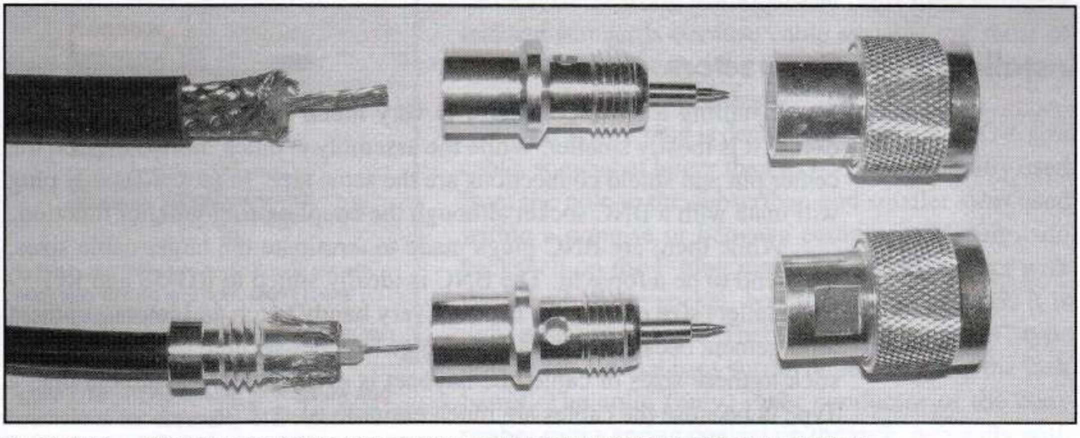

Figure 7.16 — RG-213 and RG-8X cable about to be assembled into UHF-like Type N plugs.

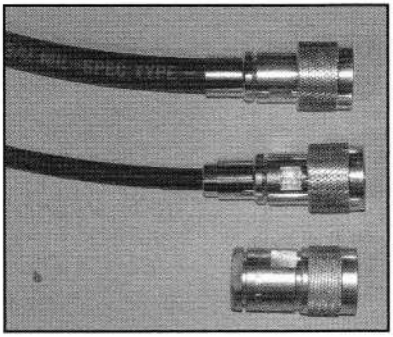

Figure 7.17 — Fully assembled Type N coax connectors on RG-213 and RG-8X cable. A standard clamp-type Type N plug is shown for comparison.

or even PL-259 at 0.74 inch. This can make it easier to route cables through smaller holes in partitions. The standard Type N can be soldered to its pin and routed through holes as small as the cable itself, but the final assembly has to be completed on the far side of the partition. Of course the current connector also needs final assembly, but it's just the threaded backshell piece. This can easily be tightened up using two 9⁄16-inch open-end wrenches, or one wrench and a vise.

The only difference between assembling this connector and a PL-259 is that the center conductor wire has a maximum length before it bottoms out inside the center pin. It must be trimmed until it can be screwed in far enough that the shield appears beyond the solder holes, as the connector is screwed onto the coax or the UG-175/6 adapter. **Figure 7.17** shows the assembled connectors in comparison to the size of a standard solder and clamp Type N plug.

These connectors cost between $5 and $7, about twice what a quality PL-259 costs, but about the same as the standard solder and clamp Type N plugs. They are somewhat longer than the usual Type N, which may be an issue in some installations. The standard N plug is about 1.5 inches long, while the item under discussion is 1.785 inches, 1.96 with a UG-176 adapter.

INSTALLING BNC AND OTHER COAXIAL CONNECTORS

Each coaxial connector type has its own installation requirements, although the BNC is quite similar to the Type N connector described in the last section.

Installing BNC Connectors

Installing a BNC connector is very much like installing a Type N, except it is mostly smaller. While the assembly is much more compact, the center pin and shield connections are the same size. In fact, a Type N plug will mate with a BNC socket although the coupling ring will not function.

While there are BNC plugs made to terminate the larger cable sizes, they tend to be a force fit. The BNC is ideally suited to RG-58 and RG-59 (or smaller) size cables and makes a very handy patch and interequipment arrangement because of its quick-change bayonet-secured backshell. If you stick to these sizes of cable, the bayonet is even easier to assemble than a Type N because the cables are much easier to work with.

As with others, the BNC is available in both crimp and clamp styles.

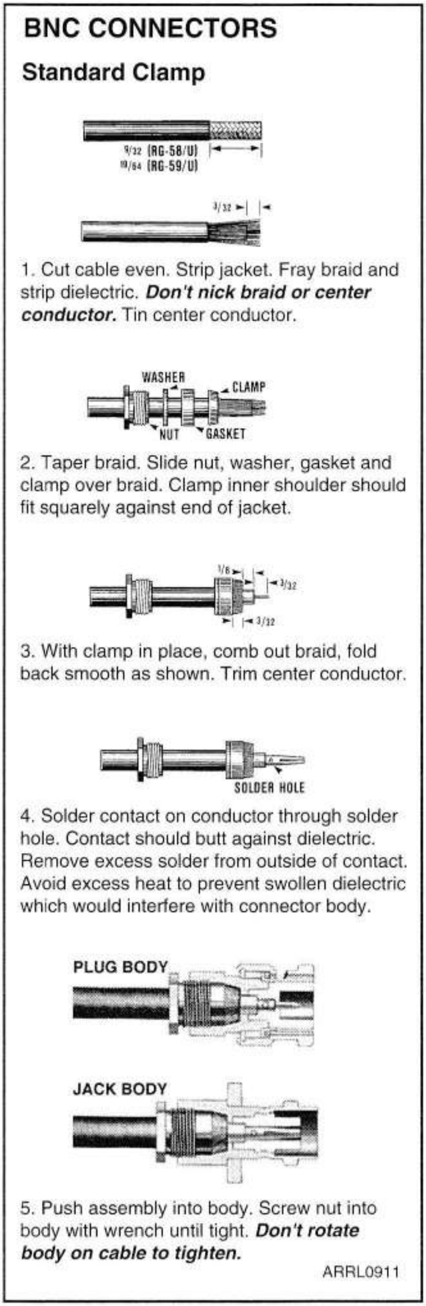

BNC CONNECTORS

Standard Clamp

1. Cut cable even. Strip jacket. Fray braid and strip dielectric. ***Don't nick braid or center conductor.*** Tin center conductor.

2. Taper braid. Slide nut, washer, gasket and clamp over braid. Clamp inner shoulder should fit squarely against end of jacket.

3. With clamp in place, comb out braid, fold back smooth as shown. Trim center conductor.

4. Solder contact on conductor through solder hole. Contact should butt against dielectric. Remove excess solder from outside of contact. Avoid excess heat to prevent swollen dielectric which would interfere with connector body.

5. Push assembly into body. Screw nut into body with wrench until tight. ***Don't rotate body on cable to tighten.***

Figure 7.18 — The official procedure and dimensions for assembly of a standard BNC plug or cable jack.

The solder (pin) and clamp (shield) are so easy to do with smaller cable, and are recommended if you can read a ruler and have power for a soldering iron. A notable exception would be cables with foil (or mostly foil) shields. The foil is somewhat less suitable for the shield clamp and a good match with a crimp arrangement.

Figure 7.18 provides the step-by-step instructions for BNC assembly. As with the Type N, both cable and panel jacks are available that assemble in the same way. There are open-frame jacks available in both flange type and one-hole mount styles.

Beware of electronic retailer "BNC" plugs that have a BNC-size backshell but use the center conductor of the cable as the pin. While they advertise them as "no solder" types, the wire is not the correct size to mate with a BNC jack.

Type F Connectors

Probably the most common coaxial connector ever made is the Type F, described in Chapter 6. While rarely encountered in Amateur Radio equipment, it appears in virtually all television and consumer video equipment made since the 1970s. It may be encountered in amateur video, or in making use of cables originally intended for other services. Because of its high volume, and the nature of the consumer marketplace, it was designed to be inexpensive and easy to install. In spite of this, cable systems using the connector must meet stringent leakage standards because cable systems use most of the spectrum allocated to over-the-air services and they must not interfere with them.

The Type F plug is designed for use with RG-6 size cable, somewhat larger than RG-59 and generally used from the pole to the subscriber, and smaller sizes used within a premise or between equipment. A screw-on backshell on the F plug maintains shield contact with the barrel of the socket, but as with the PL-259, it is only effective if tight. There is also a "push-on" type plug available that uses spring fingers to grasp the jack barrel. The solid (only) inner conductor of the coax pushes into a spring contact on the jack, generally with significant through clearance so the length is not critical, unless it doesn't make the spring contact.

Cable TV-type flexible coax provides a large fractional shield coverage through the use of a wrapped aluminum foil shield, sometimes in combination with a loose wire shield. This arrangement pretty much rules out soldering of the shield connection and provides either a clamp or, more often, a crimp-on connection. Again, a number of different proprietary types of plugs with proprietary tools are available for such applications as waterproof outdoor use, although the very basic types are not very fussy about the crimp tool used. Still, a good fit makes for a better and neater connection.

RCA (Phono) Connectors

The RCA connector (see Chapter 6) was originally employed to connect a low-level phonograph cartridge signal, via a coax-like shielded wire, to an amplifier. As such, it has an inherently coaxial structure with the plug having a hollow pin and the shield connected via four springy petals, making a non-threaded almost, but just barely, backshell.

The connector type is most frequently encountered in Amateur Radio applications as a between-equipment connection arrangement for auxiliary functions, such as carrying reference and oscillator signals as well as switching, control and sometimes low voltage dc lines. In the 1950s and 1960s, it was sometimes also found used in place of a UHF connector as an antenna connection on receivers and transmitters at or below 100 W PEP.

While crimp-type connectors may be available, loose connectors tend to be the solder type. The basic connector consists of a center conductor soldered within the hollow center pin and the shield soldered over the outside of the shield fingers in K3LR fashion. Such connectors provided as part of preassembled audio cables with molded housings may be of a different arrangement.

In addition to the basic type, there are some with a metal screw-on back cover and interior terminal for the shield. These are neater and easier to grasp, but don't offer any electrical advantages over the basic configuration.

Chapter 8

Determining Which Line is Best Suited for a Particular Application

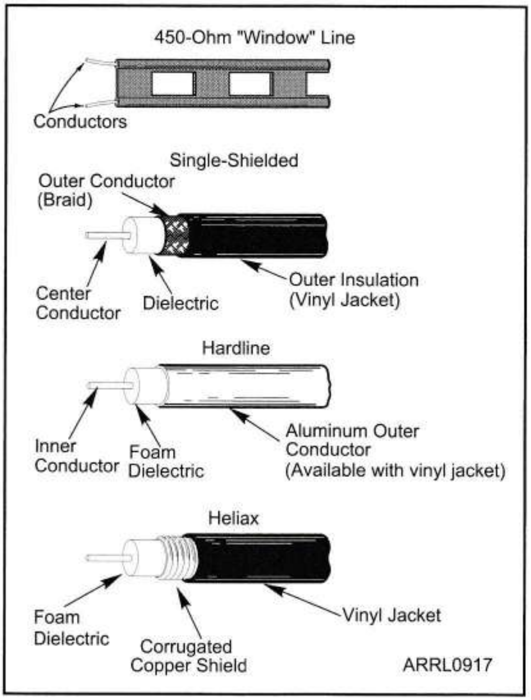
450-Ohm "Window" Line
Conductors
Single-Shielded
Outer Conductor (Braid)
Center Conductor
Dielectric
Outer Insulation (Vinyl Jacket)
Hardline
Inner Conductor
Foam Dielectric
Aluminum Outer Conductor (Available with vinyl jacket)
Heliax
Foam Dielectric
Corrugated Copper Shield
Vinyl Jacket
ARRL0917

Not every line is right for every application — picking the right one may offer improved performance and longer cable life.

The choice of the best transmission line for an application may make a big difference in how well the system works. Some application problems appear immediately, sometimes with dramatic results, while other, more insidious ones manifest themselves gradually over time and can be easily overlooked.

GETTING THE SIGNAL TO THE FAR END OF THE LINE

What we usually want most from our transmission line is to get the signal to come out the far end looking as much as possible like the signal with which we started. In the realm of radio signals, this statement is mostly about transmission line loss. In other applications, such as pulsed data systems, we may have other concerns, including preserving pulse shape without distortion, but here we will focus on transmission of RF energy. As we have discussed, the selection of line type plays a crucial role in how successful we will be.

We spend a lot of money to have a transmitter to generate a radio signal, and we want that signal to arrive at our antenna and leave as radiated RF energy — not be radiated as heat from our transmission line. There are many factors that contribute to line loss, most of which we have already covered in detail. Here we will discuss which is most important for a particular application.

Characteristic Impedance

The characteristic impedance (Z_0) of a transmission line is one of the most fundamental attributes of a transmission line, and often considered first in selecting a line. While the Z_0 does relate directly to loss (higher Z_0 lines generally have somewhat less loss, all other things being equal), the matched loss due to the Z_0 is usually less important than the extent to which the Z_0 matches the load.

The Matched Case — In the most common configuration, the load is of approximately the same impedance that the equipment is designed to drive. That impedance is most often 50 Ω for radio equipment and 75 Ω for video equipment. Not coincidentally, the most common coaxial transmission lines are available with those Z_0 and the obvious choice for Z_0 is the one that matches the actual Z impedance of the equipment at both ends. Any other Z_0 will result in an SWR greater than 1:1, which will increase cable loss and not properly load the equipment.

Once we know the Z_0, we have a number of other considerations — how much loss will our system design tolerate, and what power rating does the cable need to meet. **Table 8.1** shows representative data on these parameters for some of the most popular 50 Ω cables. Note the dependence on operating frequency for both of these parameters.

If losses are of a particular concern, it is possible to change cable types to meet particular requirements. For example, if the load is a rotary antenna, it is usually necessary to feed it with a loop of flexible cable to avoid fatigue as the rotator turns back and forth. Since the lowest loss cable types do not tend to be very flexible, it is possible to change types (within the same Z_0) if the run is long enough to make a difference. At most frequencies, there is

Table 8.1
Attenuation and Power Handling Capability for Various 50 Ω Coax Cables vs Frequency

Cable	*Frequency (MHz)*	*Attenuation (dB/100')*	*Power Rating (W)*	*Minimum Bend Radius (")*
RG-58*	10	1.2	1937	2
	100	4.3	542	
LMR-195**	10	1.1	1430	
	100	3.6	450	
RG-8*	10	0.6		4.5
	100	1.9	800	
LMR-400**	10	0.39	5800	
	100	1.2	1810	

*Belden data from **www.belden.com**
Times Wire and Cable data from **www.timesmicrowave.com.

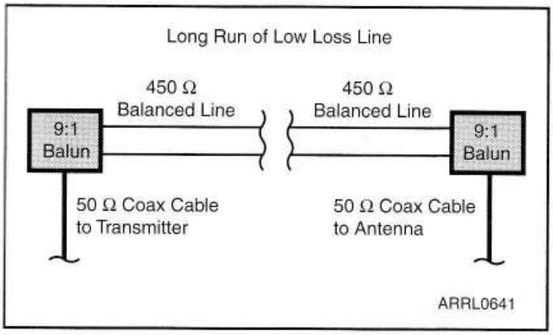

Figure 8.1 – The transition from coax to low loss, air dielectric, balanced transmission line can reduce loss if the loss reduction in the line is greater than the loss in the baluns. Here 9:1 (450: 50 Ω) baluns are used to transform the 50 Ω end impedances to the 450 Ω impedance of the window line.

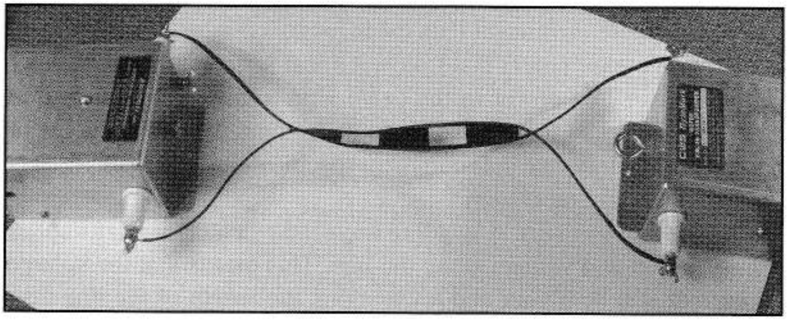

Figure 8.2 – Physical model of the 9:1 baluns from Figure 8.1 and a short window line section set up for laboratory loss testing.

negligible additional loss due to the connectors required; however, they do become additional points of potential failure as well as points of possible water ingress, if precautions aren't taken.

It is also possible to change to a very low loss medium if the length is sufficient. For example, the loss of 450 Ω window line, or 600 Ω open-wire line, is significantly less than that of coax, particularly as the frequency goes up. A balanced to unbalanced impedance transformer (balun) can be used to transform the impedance of the low loss medium to the desired 50 Ω at each end of the run as shown in **Figure 8.1** and **Figure 8.2**. The trade-off here is that the transformers will have some loss.

ARRL Lab measurements during the experiment pictured conclude that the pair of HF baluns had a total loss of less than 1.0 dB. Thus the run needs to be long enough so that the difference in loss between the cable types is greater than 1.0 dB in order for this to make sense. It is also necessary that the balanced line be installed so that it is away from the ground and not subject to other interfering objects.

Mismatched Case — If the transmission line isn't matched to the load or the equipment, a number of additional considerations need to be addressed. The three main areas are:

◆ *Additional Cable Loss.* As noted in Chapter 3, a mismatched transmission line will have more loss than a matched one. This is not generally a problem with a reasonable mismatch, perhaps 2:1 or 3:1, but it all depends on how high the matched loss would have been for the cable type, length and frequency.

◆ *Equipment Limitations*. Many radio transceivers will engage foldback circuitry if the load is different than that specified. This circuitry is designed to protect the output stages of a transmitter from the higher voltages or currents that could be present in a mismatched environment. While some radios are happy to drive loads with 2:1 or higher SWR, some start reducing power at 1.5:1. In terms of getting a signal to the end of the cable, this can be even more significant than cable loss.

Some types of equipment that follow a linear source impedance model will also deliver less power to the cable with a mismatch than if it is matched. The answer to these issues is to understand the requirements of your equipment and translate the equipment specifications into requirements for your transmission line system.

◆ *Stress on Components*. A standing wave ratio greater than 1:1 implies a changing current and voltage along the line. This means that the voltage and the current will be higher and lower along the length of the line than they would be if the line were matched. Fortunately, the locations of high current will be the locations of low voltage, and vice versa.

If the complex impedance of the load and the line propagation velocity and length are known, it is possible to determine the actual voltage and current at the end of the line. However, these values will hold for a single frequency, so in a multiband system, each may appear anywhere. The maximum voltage and current will each be equal to the values in a matched system times the square root of the SWR. If the termination equipment is specified to handle the resulting value, there should be no problem. As an example, with a matched 50 Ω, 1500 W system, the line voltage will be about 274 V, the current about 5.5 A. With a 4:1 SWR, the maximum voltage and maximum current will be twice those values at some points along the line.

OUTDOOR ISSUES

Generally, all modern transmission lines can be successfully operated indoors or outdoors. There have been types of TV twinlead, available years ago, specifically intended for indoor use because they were less visible. After a few seasons in sunlight, they would crack and eventually disintegrate. While modern cables last much longer outdoors, they do suffer degradation to different extents due to sunlight (primarily UV radiation) and moisture. The main degradation mechanism is the breakdown of the outer jacket, the rate of breakdown being dependent on the characteristics of the jacket material.

Jacket Contamination

As part of the jacket manufacturing process, plasticizers and other compounds may be left within the jacket material. These materials can leach and migrate from and within the cable, changing the characteristics and causing degradation of both the shield and the inner dielectric. This process is accelerated by the temperature rise resulting from sunlight exposure. Cables identified as *non-contaminating*, such as those with non-contaminating vinyl (NCV) jackets, are resistant to these chemical effects. High-quality PVC and other materials can provide good protection to both UV and high temperature degradation mechanisms.

Polyethylene is another jacket choice that does not have plasticizers and is inherently resistant to UV, as well as offering additional abrasion resistance. Good-quality PVC jacketed cable can be expected to last 9 to 14 years before it suffers significant degradation, while polyethylene jacketed cable can last about twice as long. Low-quality cable can be expected to degrade significantly faster.

Water Penetration

Coaxial cable jackets are made of different materials suitable for different environments. A major problem with many types of coax, as well as some connectors on the ends, is water penetration. Water within the coax can cause the dielectric to become lossy and result in corrosion of the shield and even the inner conductor.

Water enters coaxial cable in at least two ways. Gaseous water vapor can migrate through the outer jacket. Liquid water can also enter through jacket pinholes left from the manufacturing process, or from abrasion resulting from rough handling.

In most cases, the entry of water into the cable from the connector ends is a more significant issue. While BNC and Type N connectors are designed to be waterproof, PL-259s are not. In order to avoid problems, a PL-259 must be sealed by means external to the connector. In addition, it is particularly critical that a drip loop be included at each end of every cable run that goes downward. A drip loop (see **Figure 8.3**) is any local low point before the connector so that water running down the outside of the cable can drip or run off before it gets to the connector. Without the loop, water will pour onto the connector, sometimes coming right into the equipment if connections aren't tight.

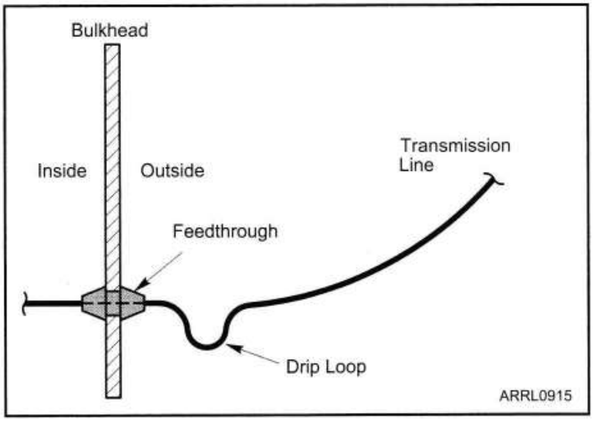

Figure 8.3 – A drip loop is an intended local low point in the cable before the connector so water running down the outside of the cable can run off before it gets to the connector.

There are a number of commercial products available for waterproofing connectors, although some make later connector removal very difficult, so make sure the connection is right before you seal it. High-quality cable will have fewer pinholes and be less subject to vapor migration, but improper connector installation can negate the benefits.

Direct Burial Coax

Most types of coax are designed for above-ground use, but some are rated for direct burial. This is mostly a statement about abrasion resistance. The usual ground near the surface moves with seasonal changes as well as with differences of surface load. Any shifting of the ground will tend to result in jacket abrasion from rocks or other debris in the soil.

Note that direct burial does not mean "submersible." Only special submarine cables, out of financial reach for most amateurs, are designed to spend time in the water. Direct burial cable doesn't promise lack of pinholes. This can result in problems for cables in any kind of ground without good drainage, such as clay. If the older houses in the area are made of brick, check your soil carefully before you bury any coax cable.

Some installers use conduit to avoid abrasion problems with below-ground cable. While this can reduce abrasion, it can result in other problems, such as water accumulation. Water can enter conduit through the ends or, more likely, through condensation. Without proper drainage, the coax ends up in the water. To avoid this problem, implement one of the following:

◆ Use conduit or drainage pipe with drainage holes on the bottom side. Prepare the trench with a layer of gravel above sand or earth with good drainage. Lay the conduit in the trench, hole-side down, and use elbows on the entrance and exit points to avoid direct water entry (see **Figure 8.4**).

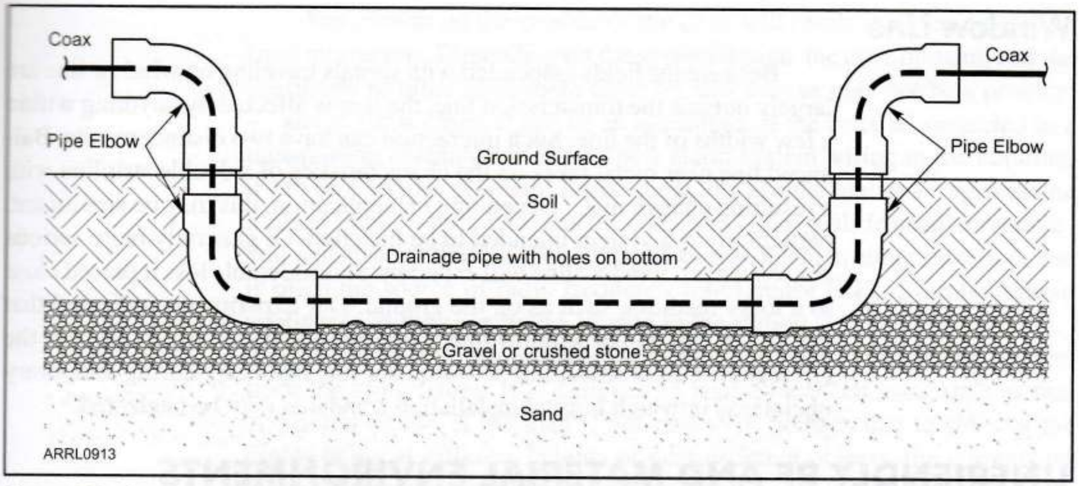
Coax
Coax
Pipe Elbow
Pipe Elbow
Ground Surface
Soil
Drainage pipe with holes on bottom
Gravel or crushed stone
Sand
ARRL0913

Figure 8.4 – A problem with burying coax in conduit is the accumulation of water in the conduit. By using drainage pipe with holes on the bottom side over good drainage material, such as the gravel and sand shown, any water can be moved away from the cable.

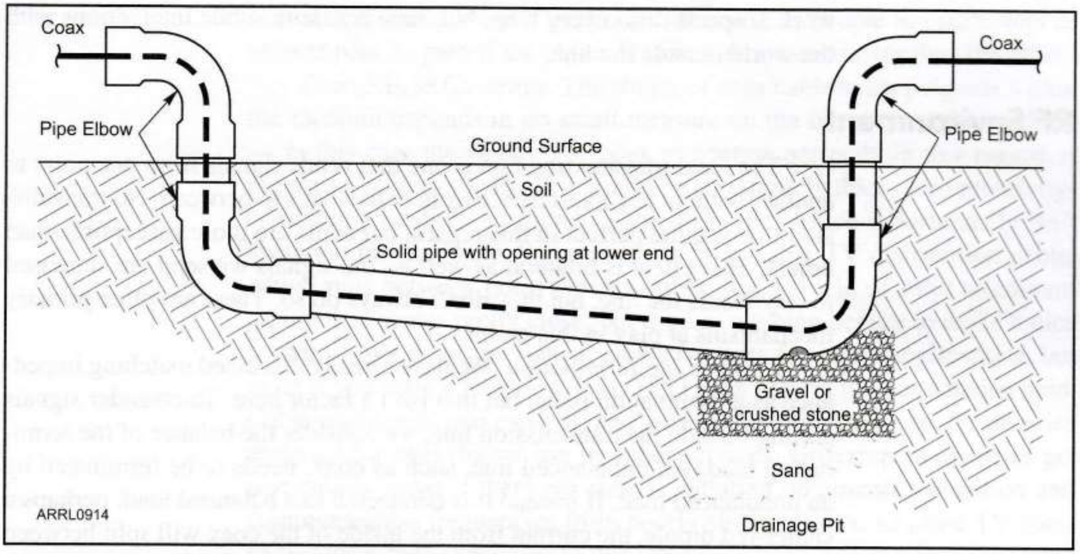
Coax
Coax
Pipe Elbow
Pipe Elbow
Ground Surface
Soil
Solid pipe with opening at lower end
Gravel or crushed stone
Sand
Drainage Pit
ARRL0914

Figure 8.5 – Another solution to the problem of accumulation of water in the conduit is to use a pitched conduit with one or more low points able to drain into drainage pits.

- Lay the conduit with a pitch so that all water will run to one end, or to a midpoint. Have a drain at that point into a pit with gravel and sand in sufficient quantity to absorb the expected water. Follow up with a look at the drainage pit after some wet weather (see **Figure 8.5**).

Window Line

Because the fields associated with signals traveling on window line are largely outside the transmission line, the line is affected by anything within a few widths of the line. Such interaction can have two distinct results. Balanced line near metal takes on the characteristics of a double stripline with a resulting change in Z_0. Depending on application, this may be significant, particularly if a system intended to be matched. Of generally more serious consequences, window line will experience considerable loss if placed close to a lossy medium, such as on the ground. Our experiments indicated that the loss in 100 feet of window line on dry ground was 10 to 20 dB over the HF region. This is something to watch out for, especially during temporary operations in which normal installation standards may be neglected.[1]

UNFRIENDLY RF AND MATERIAL ENVIRONMENTS

To a greater or lesser extent, all transmission lines are subject to effects of their external environment. Some of these are obvious. For example, a transmission line that melts in a high-temperature environment will not work at specification very long, but there are more subtle interactions with the world outside the line.

RF Environment

The ideal transmission line is one that sends signals from one point to another without any interference from or to what's in between. No transmission line is quite perfect in this regard, but some are more susceptible than others. This effect is bilateral as well — the signals we send are supposed to stay inside the line, but they don't always do so. There are three primary mechanisms at play in this realm.

Proper Line Termination. We have already discussed matching impedance as a termination issue, but that isn't a factor here. To consider signals getting outside the transmission line, we consider the balance of the terminating load. An unbalanced line, such as coax, needs to be terminated by an unbalanced load. If instead it is connected to a balanced load, perhaps a center-fed dipole, the current from the inside of the coax will split between the half-antenna connected there and the outside of the coax, acting like another wire in parallel.

The amount of current going on the outside of the coax, which — because of skin effect — acts like a separate conductor from the inside surface of the shield, depends on the relative impedances of the two connection paths. This will depend on the frequency, the length of the coax, and the impedance of the ground termination of the shield at the far end.

Any current on the outside of the coax will result in radiation, just as from an antenna. Depending on the system design, the environment, and the sensitivity of other equipment close by, this may or may not be a problem — and in some cases, it can be an advantage. If the coax is connected to a transmitting antenna, and it runs past alarm system wiring in the building, it can be a serious problem. If you spent lots of time tuning a Yagi antenna for best front-to-back ratio and the coax picks up signals from all directions, you will also not find it a plus. This can be a problem in either direction, and is often the source of radio frequency interference (RFI) from household appliances as the coax works its way to the station.

Balanced transmission line, such as the popular window line, suffers from the same issues. It is designed to feed a balanced load, such as that center-fed dipole. If the load is not balanced, perhaps due to sides of the dipole being at different heights or near different objects, the currents on the two conductors will not be equal and opposite. Any resulting difference current due to imbalance will act like a separate antenna current on the line, and radiate just as will the coax described above, resulting in all the same issues. In addition, the unshielded balanced line, while its balanced currents cancel at distance, even if perfectly balanced can couple to conductors or objects near its path if the object is closer to one conductor than the other.

Coax Shield Coverage. The ability of coax cable to keep signals within the medium depends in no small measure on the integrity of the medium — in this case the shield. For coax to operate perfectly in this regard, it requires a perfect shield. This would be an infinitely thick, zero resistivity-conducting medium surrounding the dielectric, which unfortunately isn't available for purchase. Some coax, such as cable TV distribution cable, which runs between poles down a street, has a shield of solid aluminum tubing that comes pretty close, but it takes a tubing bender to make a nice corner — not what the typical amateur wants for many applications, but great if you want a long, low-loss, 75 Ω coax run out to your antenna field. Cable companies make use of, and must avoid interference to, frequencies assigned for over-the-air use by other services. Government agencies get particularly upset if there are signals radiating on aircraft navigation and communication frequencies from poorly shielded coax, so cable TV companies have been sensitized to this issue.

Most coax used by amateurs has a braided copper shield with an effectiveness related to the percentage of coverage — often a parameter listed by manufacturers. Note that even 100% coverage by braid will not quite equal the shielding effectiveness of tubing. Still, all things being equal, more coverage means more effective shielding. The 75 Ω RG-6 cable, often used as "drop cable" between the street and an end user, typically has a dual shield

— one layer aluminum foil, the other tinned copper braid. While the braid is usually quite sparse, the combination meets the strict cable TV emission requirements. Unfortunately, it doesn't work well with soldered connectors, but is appropriate for crimp-on types.

Connector Continuity. Particularly with UHF and F-type connectors, in which the connection of the shield is dependent on connector backshell tightness, improper tightening of connectors can nullify part of the effectiveness of shield coverage. Any resistance in the shield connection acts like a coupling mechanism between the inside and outside of the shield. This can get worse with time, due to vibration or oxidation of the surfaces.

Material Environment

Coax cable pretty much has the advantage here. The fields between the conductors of coax, except as noted above, stay largely within the cable itself. Thus coax is pretty impervious to normal outside influences — other than water or abrasion as outlined previously. Still, each cable has a temperature rating that should be observed, as well as some sensitivity to some types of chemical vapor environments that can attack the outer jacket or permeate and degrade the shield or inner materials.

Window line, with most of its fields outside the physical location of the wires, is very susceptible to signal degradation of various types due to proximity effects. This should be carefully considered in selecting appropriate line for a particular application. As noted previously, the attenuation of window line closer than a few inches from dry ground is very high — perhaps 10–20 dB in about 100 feet, based on our measurements.

If window line is routed close to metal surfaces, it will have a major change in characteristic impedance, as well as the possibility of coupling signals to undesired objects. Similarly, window line can be a problem in getting through walls or other partitions in which the lossiness of the materials in the partition may be unknown. One solution may be to use short sections of dual coax (see Figure 5.9 in Chapter 5); however, they should be quite short if the coax loss is to be negligible.

Notes

[1] B. Allison, WB1GCM, J. Hallas, W1ZR, "Getting on the Air — A Closer Look at Window Transmission Line," *QST*, Nov 2009, pp 66-68.

Chapter 9

Application and Installation Notes

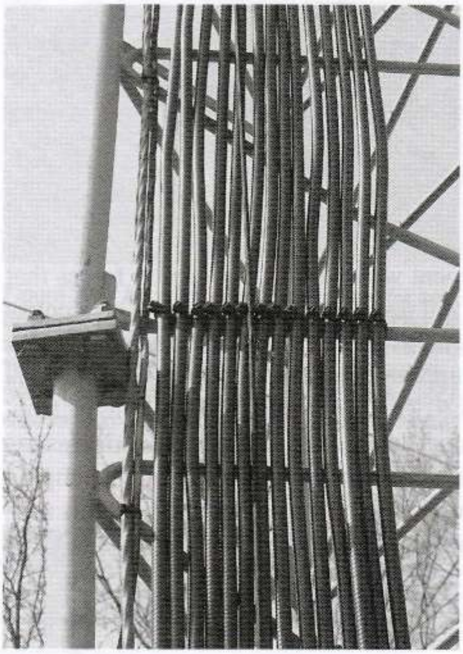

Coaxial cables going up a tower at W1AW, the ARRL Headquarters.

Once you have decided on a transmission line with the features you want, it is important to install it properly. Poor installation practices can result in more rapid degradation of transmission lines due to additional stress or inappropriate environmental factors. Implementing some of the following elements can help extend the life of your transmission line.

SUPPORTING LINE RUNS

Transmission lines are not generally intended to be structural members, yet they are often treated as such. A transmission line, particularly a coaxial cable, is not light and, in typical lengths, can be fairly heavy. **Table 9.1** provides some of the physical characteristics of typical coaxial cables. A quick glance indicates that each type should be able to withstand the tension of at least a 1000-foot vertical run; however, that's not quite the whole story.

Vertical Transmission Line Runs

While the cable itself may be able to withstand the tension of its own weight, whatever is used to secure the upper end will need to exert some kind of force on something to help it defy the forces of gravity. While a connector should also be able to support the weight, (assuming proper installation and no vibration that will cause it to break free in time), hanging a long line on a connector is not generally a great idea. The only thing worse is to have everything hanging directly on the soldered connections between the antenna and the cable.

Table 9.1
Physical Properties of Selected Transmission Line Types*

Cable Type (Belden PN)	*Weight/1000 feet (pounds)*	*Max Tension (pounds)*	*Minimum Bend Radius (inches)*
RG-213U (8267)	104	184	4.0
RG-8U (8237)	104	190	4.5
RG-8U Foam (9913)	97	300	6.0
RG-8U Foam (8214)	106	230	4.0
RG-8X Foam (9258)	35	75	2.4
RG-58 (7807)	22	25.4	1.9

***www.belden.com**

The usual clamping arrangement will exert perpendicular forces that will distort the shape of the cable and result in a combination of concentration of tension forces and reduced ability to withstand the tension.

One good way to provide support to a coaxial transmission line is shown in **Figures 9.1 through 9.5**. This was described by Lyle Nelson, AB0DZ, in a *QST* "Hints and Kinks" column.[1] This method distributes the tension along a length of the cable and doesn't require any tight clamping to hold it securely in place. To further reduce tension on the line, this may be repeated as often as desired so each section of the cable is just holding up itself.

Horizontal Transmission Line Runs

A horizontal run adds the force required to avoid sag in the run. In fact, strictly speaking, it is only possible to have a horizontal run if there is infinite tension on each end to result in zero sag. All practical runs over horizontal spans are actually in the shape of a catenary, the shape that a cable takes on if its weight is supported at each end (see **Figure 9.6**).

To determine the amount of tension needed for a particular amount of sag, use the nomograph shown in **Figure 9.7**. The operation is performed in two steps:

Step 1. Lay a ruler between the two outside scales, with one end on the desired span length, and the other on the cable weight (in the same units as in Table 9.1). Make a dot on the "Work Axis."

Step 2. Rotate the ruler, keeping an edge on the dot from Step 1. At each position there will be a value for the tension and the corresponding half-span sag. Use a tension that is comfortably below the maximum tension shown in Table 9.1.

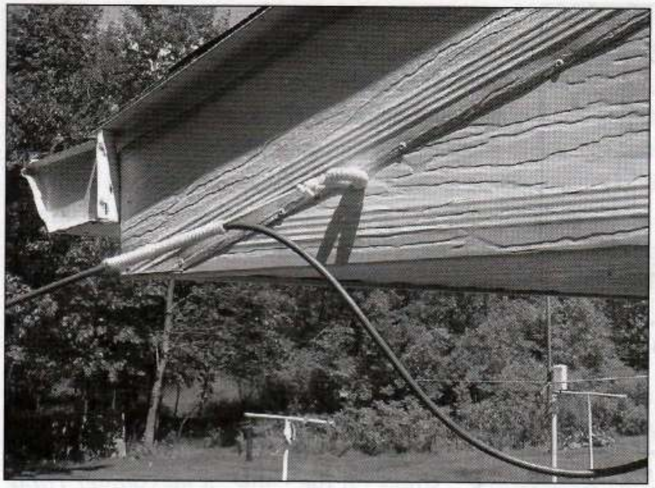

Figure 9.1 – An attachment mechanism for securing coaxial cable that distributes the tension of a length of the line to avoid excessive stretch. This can be used for either horizontal or vertical cable runs. The contributor recommends a 40-inch length of ⅛-inch braided nylon rope for RG-8X size cable, 60 inches for RG-8 size, with an overhand knot at each end and the ends melted to avoid fraying. [LYLE NELSON, AB0DZ]

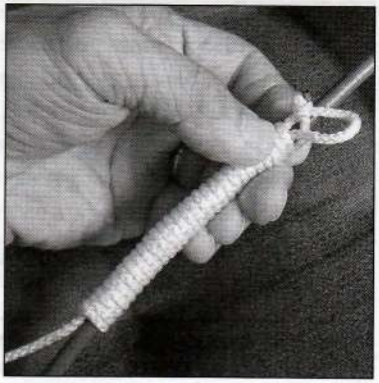

Figure 9.4 – Wind turns until you have about 3½ inches of the loop covered. [LYLE NELSON, AB0DZ]

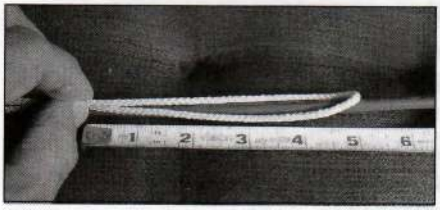

Figure 9.2 – Start by making about a 5-inch loop along the cable. [LYLE NELSON, AB0DZ]

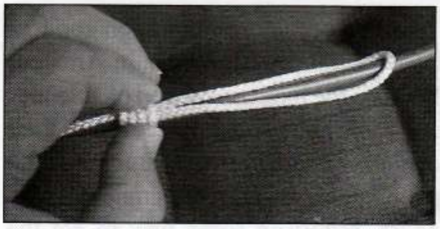

Figure 9.3 – Wind the cable in tight wraps starting at the open end of the loop. [LYLE NELSON, AB0DZ]

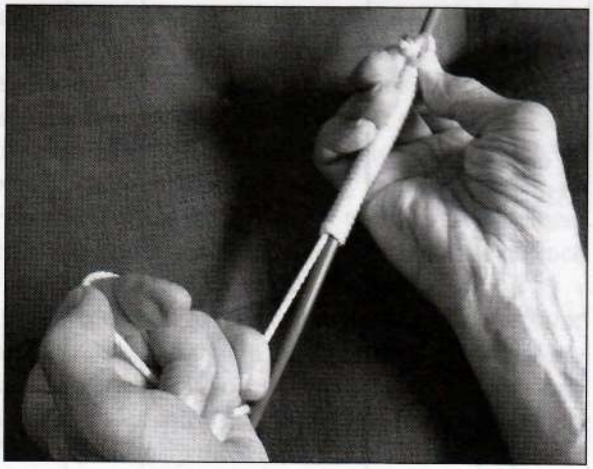

Figure 9.5 – Stick the end through the remaining loop and pull the rope from the other end until you have it tightly secured. The "tail" left out is then used to tie the cable to a screw eye or other support. [LYLE NELSON, AB0DZ]

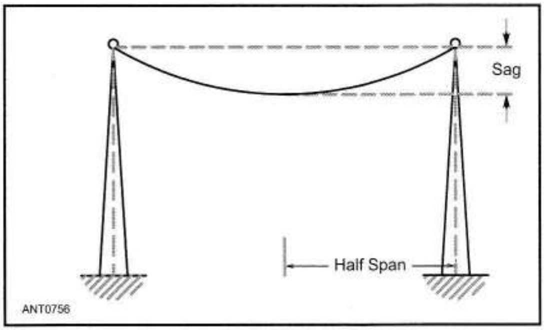

Figure 9.6 – A horizontal cable run with less than infinite tension on the ends will assume the shape of a catenary, as shown.

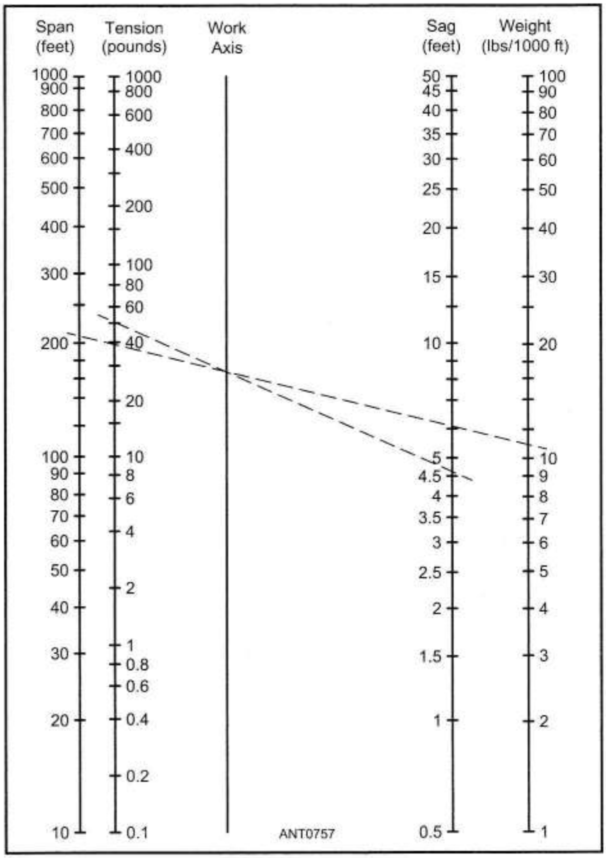

Figure 9.7 – Nomograph for determining sag in a cable run based on cable weight and span distance. [JOHN ELENGO, JR, K1AFR]

The dashed lines on Figure 9.7 are for an example with a cable weight of 11 pounds per 1000 feet and a span of 210 feet. The second line indicates that for a tension of 50 pounds, a sag of 4.7 feet will result.

Note that all of this assumes the end supports are fixed. If the end supports are trees, or other objects that will shift in the wind, the system needs to be laid out as if the trees were both at the position of furthest separation. The result will be additional sag in the rest position.

If the resulting sag is not satisfactory, or if you wish to reduce physical stress on the cable, a *messenger cable* can be used to support the cable run. This is just a kind of line with a high tension capability that is installed in the same path with correspondingly reduced sag. The transmission line(s) are then secured to the messenger cable every few feet so it assumes the shape of the messenger.

Keep in mind that the tension applied to the cable will also be applied to whatever anchor is supporting the end. This doesn't work if the tension pulls out the fasteners, or the wall they are attached to.

BUILDING ENTRANCE ARRANGEMENTS

One of the most common applications of transmission lines in Amateur Radio is the interconnection of antennas and radio equipment. By virtue of their operating characteristics, antennas tend to be outdoors and radio equipment indoors, requiring some way to get the line from one to the other. This can be a bit tricky in the case of housing — it happens that most houses are especially designed to not have holes in their walls for running cables.

Of course, there are already such penetrations provided for utility power, telephone service drop lines and coaxial cable for CATV or satellite TV service, as well as non-electrical facilities such as water, gas, sewage and heating flues. If you are the owner of the property, you have the option of observing the techniques used by those services and doing something similar for radio antenna connecting transmission lines, although some techniques may be much easier to accomplish before the dwelling is completed.

Drilling Holes in Walls

As the owner, or with the permission of an owner, you can usually drill holes in walls. I have done this many times and have learned a few things in the process.

Be Sure You Know What's Between Outside and Inside

There are few things worse for the amateur constructor than drilling through a wall and running through a water supply or water heating pipe. Of course these can be repaired, but usually the damage extends far and wide before you can get the water turned off. If you will need professional help to recover, a good operating rule is to not attempt something like that on a Sunday!

The best way to avoid such problems, be they caused by pipes or wiring, is to make sure you know what's there before you drill. One way that works for me is to make all such penetrations into unfinished space, such as the wall above a basement foundation. Coming in near ground level also has potential benefit in terms of your lightning protection plan, which we will discuss in the next section. My station is usually in the basement, so this is a natural; however, if it's not, it may be easier to make the basement entrance and then go between floors, since most floors either have no such obstacles, or they can be easily seen (more on this later).

If you don't have a basement, or if basement entry won't work for some other reason, you will need to come through in higher ground. Here it is very important to know what you are doing. If you are not familiar with construction techniques, find a house under construction that has similar characteristics. Look it over before they close up the walls. You will see some places that make sense to drill through. Note that (in my experience) every window and door is surrounded by a pair of 2 × 4 or other framing lumber. This means that a hole drilled next to a door or window, or beneath a window will be wood the whole way through — generally no pipes or wiring. In addition, you can probably come through a piece of molding, which is a good idea.

I like to drill so that there is wood on both sides, so the inside hole comes out in a wooden molding. This way the hole is solid and pulling cable won't rip up plaster or sheetrock. It is also very easy to patch after you are done, especially for painted woodwork — a few applications of spackle will usually make even the biggest hole disappear. If you want to be even more solid, drive a dowel in to just below the surface, to be a base for the spackle. Adding a bit of trim paint is much easier than repainting or papering the whole room. If next to a window, you often have the added benefit of having the cable entrance hidden by draperies.

Note that your usual twist drill will not be long enough for the task. Invest in a couple of 12-inch or longer drill bits in sizes that slowly come up to the cable size. An exterior wall will be the thickness of the framing studs, plus the sheetrock or paneling on the inside, plus the inside molding. On the outside you may have underlayment or sheathing and then multiple layers of shingles, for example. Another advantage of drilling through solid wood is that you can then just push the cable through the hole. If you are drilling through a hollow wall, you will have to snake your way through the insulation and then pull the wire through — not the end of the world, but certainly an extra hassle. Make sure you drill the hole at a slight angle downward going outward, so no water will run into the house through the hole or follow the cable in.

If You're Doing More Than One, Combine Them

Somehow it seems that most amateurs eventually have a need for more than one such penetration. Even if you start out with one, your interests and capabilities will likely grow as you get further into Amateur Radio. While one can just keep punching holes in different places, at some point some kind of entrance facility may make more sense. We'll discuss some in the context of lightning protection, but even independent of that, a single entrance facility may be a better idea. For example, interior and exterior electrical boxes on each side of the exterior sheathing and attached on the inside to a stud may make a professional-looking arrangement that is easy to add to just by popping off the covers on both sides. If you're not up to the task, having an hour of discussion time with a contractor or electrician may pay for itself in the long run.

Making a Window Entrance

If walls won't work, another possibility is a window entrance. While I have seen descriptions of ways to drill holes in glass, I've never attempted it. Instead, I've unglazed and removed the pane of glass, and replaced it with a piece of ⅛-inch polycarbonate sheet. The light transmission is just a bit less than through window glass, but not dramatically so. Still, I would avoid windows that highlight the appearance of a major room.

The polycarbonate is easy to work with using the usual tools, such as a reciprocating saw to cut it to size. Holes can be easily drilled, either for transmission line directly, or for PL-258 type or other feed-through connectors. If the holes are near the window edge, the cables may be able to be hidden by curtains or draperies as they go toward the floor. When it's time to move, recovery couldn't be simpler. Just remember where you stashed the pane of glass and swap it out with some glazing points and compound, and no one is the wiser.

Another possibility is to drill through the window sash itself. Carefully drill through a wooden sash in a spot that will be in the clear if the window is closed and that will not try to run through the glass. This can be accomplished with a short drill, is not likely to encounter plumbing or electrical runs and is easy to patch, if the surfaces are painted. While this is not quite in the same "no damage" category as the temporary board described below, it can be repaired to be almost invisible very quickly.

Temporary or "No Damage" Entrance Arrangements

If your circumstances are such that you can make no changes to the structure, or aren't even authorized to do any of this, there are still ways to bring antenna connections into the house. Some avoid the problem by using indoor or attic antennas. While that can keep you on the air, the chances are that you can have better results with an outdoor antenna. Here are a few possibilities.

Make the Connections Just When You Need Them

Most amateurs don't operate all the time, so they don't need an antenna connection all the time either. A transmission line connection can be just outside a window or door with its mate on the other side. When it's time to operate, just open the door, hook things up and leave the door open a crack while operating. Unless you are faced with heavy mosquito activity, this can work well. In fact some amateurs have the antenna, such as a short loaded vertical, just inside a patio door and move it all outside when it's time to operate — and when no one is looking.

Pinch Your Cable Under a Window Sash

Those using single-wire feed or window line can often put the line under a double-hung window and, if there's any gap or slop, just close the window on it, if it can still be latched. This is not ideal for a number of reasons, but is pretty easy to do. Perhaps a bit of "dressing" with a file can take the hard corner off the edge to make it less likely to damage the cable.

Add a Removable Panel Under a Window Sash

An improvement on the above that is almost as easy, is to get a small piece of trim board the width of your window sash and fit it under the lower sash, or above the upper sash so the window closes on it. Now you can notch (especially for cables with connectors) or drill as many holes as needed in your board for coax or other cables, or for feed-through connectors. If you don't want to build your own, MFJ offers a preassembled unit (see **Figure 9.8**) that can be cut to fit vertical or horizontal windows up to 48 inches wide and has feed-through arrangements for both coax and balanced lines. See **www.mfjenterprises.com**, and look for MFJ-4603.

There are two complications that need to be dealt with using this method. The first is that the security latch will no longer operate. This is pretty easy to fix. Either cut a piece of wood that can be jammed between the movable sash and frame, or install a small hardware store angle bracket so the window can't move. The screw hole can be easily spackled when it's time to move. The other problem is that there will be an open passage between panes of the two sashes that will allow insects to come through. Mosquitoes are particularly prone to do so, since they are attracted by the focused emission of CO_2 from people breathing inside the house. This is the same problem that occurs with window air conditioners and yields to the same solution — just a piece of foam strip pushed into the space (get one from an air conditioner dealer, or in a pinch, it's a great use for rolled-up surplus T-shirts).

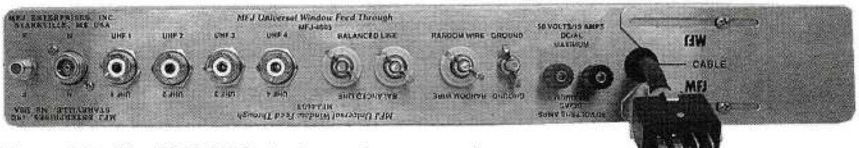

Figure 9.8 – The MFJ-4603 window entrance panel can be obtained with a number of feed-through connector configurations, or you can make your own from a piece of 1 × 2 lumber.

Drip Loops

With any cable entrance arrangement, it is critical that a drip loop be provided. This is easy to do; just make sure that the cable has a lower point before it enters the entrance arrangement. Otherwise, rainwater will follow the cable down and run into the wall or the feed-through mechanism. The drip loop allows a lower path from which the water will drip harmlessly onto the ground.

Extending the Route Between Floors or Rooms

If the entrance arrangement can't be on the same floor or in the same room as the station equipment, it is often fairly easy to continue on. If you can't make permanent holes, look for existing ones that you can follow. A building with central heating usually has ducts or pipes that run from room to room, or between floors. In my experience, there is often space to follow, especially with the thinner varieties of coax. If you can't push it through without losing it in the space between wall surfaces, try pushing a straightened coat hanger through from the other side. Then tape the coax to the hanger with electrical tape, making a point of the tape — along with tapering the coax end in extreme cases — so the coax won't be pushed off, and then pull it back.

If you can drill a path, you may be able to find a good path — again, be alert for the possibility of wiring or plumbing. An unfinished basement ceiling is the easiest to get through since it is only one surface thick (usually subfloor plus finished flooring, perhaps 2 inches total). My favorite technique is to go through under an interior wall with a molding. First checking to make sure you won't end up in a floor joist, start the hole through the molding at an angle to enter the basement slightly offset. This way results in an easy-to-patch hole in the molding, rather than having to deal with patching a hole in a finished floor.

Going between upper floors is a bit trickier to do without making a mess. In many cases, following existing penetrations may be the best way. If not, my next choice is to find closets one above another. A discreet hole between floors in a corner adjacent to the door, or even next to the door molding, is usually not too obvious, especially if the cable is then painted the same color as the wall surface. If you're not comfortable with any of this, a competent residential electrician can likely do it all for you, once he understands exactly what you want. If possible, do it in such a way that you can pull the cables out and replace them or add to them. Ask the electrician to leave a loose piece of cord along the path so you can add additional cables as your needs expand — and they will.

Bringing Balanced Transmission Lines In from the Cold

All the techniques that we discussed for coax can also be applied to balanced lines, with just a few caveats. Balanced lines need to avoid a number of situations that wouldn't bother coax. Conditions to watch for are runs that are within a few inches of metal ducting, pipes, or wiring. Also, lossy material that will absorb energy from the fields around the line must be avoided. Perhaps most critical is inadvertent coupling to and from signal lines, including telephone, network, or alarm systems.

If it is not possible to avoid such environments, it is possible to employ a section of dual coax balanced line, as discussed in Chapter 5 and illustrated in Figure 5.9. While this will be an unmatched section resulting in an impedance bump, that is not at all critical in the usual unmatched balanced line configuration, but should be considered if dealing with a matched condition, such as a 600 Ω feed to a rhombic antenna, or a 300 Ω feed to a folded dipole. The other consideration is that the loss will be that of unmatched coax, likely significant, so the length of this section should be as short as possible. For example, I used about 4 feet on one side of a rhombic feed as it went through a potentially lossy wall sill, just above the foundation line, without any problems.

GROUNDING AND LIGHTNING PROTECTION

While not a "transmission line" issue, per se, transmission lines that bring signals in from antennas can also bring in lightning energy if not properly protected. We will not go completely into lightning protection here, since that is arguably a subject for a book of its own; rather, we will briefly discuss ways to minimize what may be coming down the line as it comes inside.

A key to understanding lightning effects is to appreciate that with peak lightining strike current intensities from perhaps 20,000 A (50th percentile) to 200,000 A (90th percentile), it's not a question of where the current will go, but rather how much is left after you've gotten rid of as much as you can outside, before it heads toward the shack. This requires consideration of tower and antenna grounding with a low enough impedance to drain as much of the current as possible.

Still, it won't all go there — it will divide between paths to ground based inversely on impedance. Some will always go on the coax or other transmission line towards the station. It will show up in two forms; common mode current, essentially on the outside of the coax shield, and differential mode, between the inner conductor of the coax and the shield. In most environments, the common mode is a much larger fraction of the lightning current than the differential mode and can be reduced by low impedance grounding of the shields of coax as it comes into the building.

The Grounded Entrance Panel

The way to achieve this is to have the cable entrance at ground level and have as close to zero length conductivity as possible to a good ground structure at that point. This is not the usual solitary ground rod required by the National Electric Code (NEC), but a network of ground rods interconnected by buried, heavy, bare conductors that help distribute the current to the Earth's surface. The amount and extent needed for a desired grounding effectiveness will depend on the electrical characteristics of the soil at each location. The NEC, written into most local building codes, requires that all

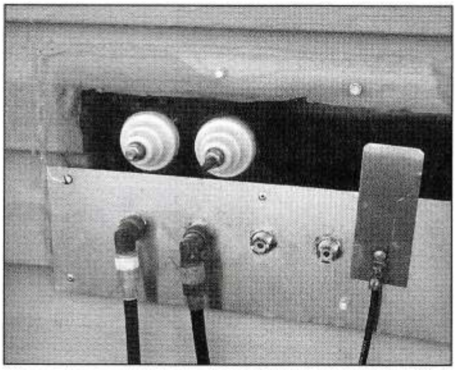

Figure 9.9 – The grounded entrance panel at amateur station K8CH.

grounds be bonded together. Thus your radio ground system should be bonded to the ground used at the electric utility panel — preferably connected via a short heavy conductor (sized per the local building code) running outside the building.

The best way to provide maximum protection, in my opinion, is to have all cables enter through a common ground level entrance panel, as shown in **Figure 9.9**. By having the panel at ground level, it is possible to have the lowest impedance ground connection arrangement by having the ground system start right at the panel. Instead of using feed-through connectors, my ground level entrance panel (shown in Figure 8.3) has commercial lightning arresters that serve double duty and are bonded to an aluminum plate that is grounded to an extensive ground system tied to a rod at the window and extending outward to multiple rods through buried bare ground wires.

The window line entrance (not shown in Figure 8.3, but similar to the two feed-through insulators visible in Figure 9.9) goes to a balanced feed arrestor just inside the panel. It is also possible to use two coaxial lightning arrestors to protect the equipment tied to the balanced feed line, with one conductor on each arrestor. Just make sure you calculate how high the operating voltage will be, and select appropriate arrestors so that they won't fire during normal transmission.

At my station, I have push-on type UHF coax connectors on my coax cables going to the inside connection of the arrestors (available from **www.americanradiosupply.com** and others) and double banana type going to the balanced arrestor. I routinely keep them disconnected when not actually on the air. In addition, I keep my ARRL All Risk equipment insurance policy up to date (see **arrlinsurance.com**).

Above-Ground Entrance Arrangements

A station located on an upper story without ground-level access has a different set of challenges. While it may be tempting to reduce transmission line length by making runs directly to the station, arguably it is the worst case from a lightning protection perspective, since all lightning currents on transmission lines are forced to go through the station. A better arrangement is to run the cables to ground level with an appropriate grounding and surge protection interface at that point, as shown in **Figure 9.10**. It's not bad to have additional arrestors at the station end, but the ground-level ones are the most important, in my view.

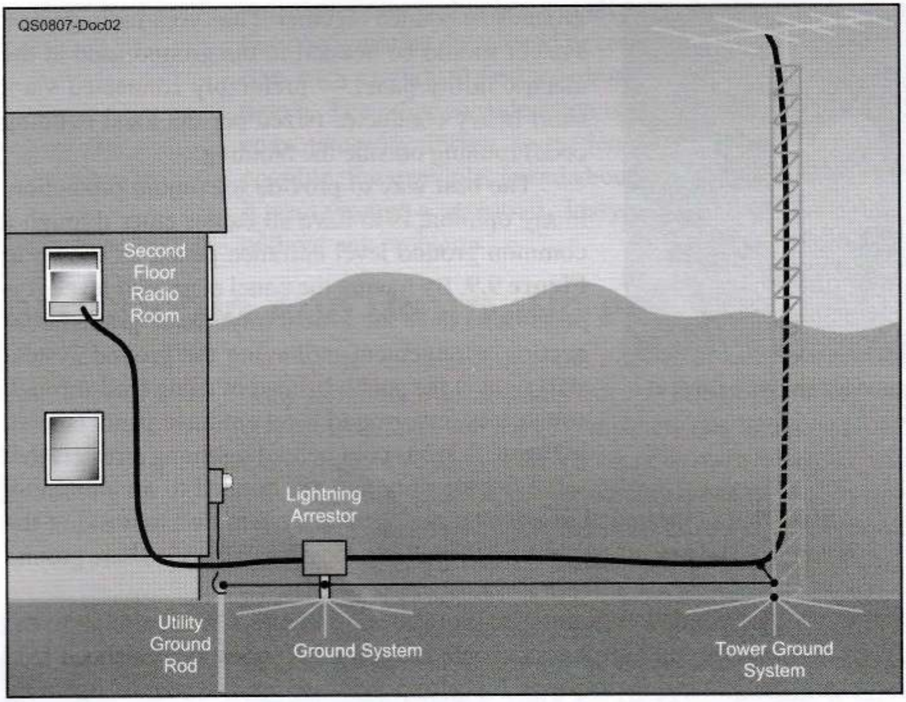

Figure 9.10 – Suggested arrangement for lightning protection grounding of an upper-story cable entrance. By running the cables first to grounding and lightning protection at ground level, the lowest impedance ground connection is provided.

TRANSMISSION LINE CARE AND MAINTENANCE

Fortunately, transmission lines tend to be rather low maintenance, however, that's not quite the same as "no maintenance." Transmission lines can last for many years, but none last forever, with typical expected useful lives of 10 to 20 years. That, of course, assumes proper installation and care.

Premature Line Degradation

Environmental factors are the major cause of early line degradation. Probably the most significant are the result of water penetration that occurs due to insufficient sealing of ends, or through jacket penetration resulting from abrasion. The insidious aspect of this is that such gradual degradation is not easily noticed through normal operation until it goes quite far. While we will pay lots of money for an additional 2 or 3 dB of antenna gain, we are not too likely to notice 2 or 3 dB of additional transmission line loss. In fact, the additional line loss results in lower SWR at the station end, along with easier tune up and wider bandwidth — each of which could be, but probably aren't, noticed due to the gradual change.

Line Inspection

A regular program of physical inspection of transmission lines is most appropriate. Areas that are becoming abraded, show signs of insulation cracking, or from which waterproofing material is becoming unsealed can be readily addressed before they cause serious problems. A good application of electrical tape can provide additional coverage, as well as additional layers in areas subject to abrasion.

Be aware of areas that are bent more tightly than the allowed bending radius shown in Table 9.1. Foam dielectric cable, while lower in loss than cables with solid polyethylene dielectric, is subject to migration of the center conductor and possible shorting, if bent too much.

Another potential problem area is any solder connections that are exposed to the elements. Solder oxidizes in nature and can erode over time, eventually making for intermittent or failed connections.

Routine Measurement

It is a good idea to take and record a complete SWR run across frequencies of each antenna and transmission line system at installation. If nothing degrades, the values should stay unchanged over time. As a general rule, if the SWR goes up, look to the antenna for changes. If the SWR goes down, check for additional transmission line loss, since antennas rarely retune themselves, left to their own devices.

Checking Line Operation

Perhaps the best way to evaluate a line's operational characteristics is to measure the loss and SWR before installation, and then again during regular inspection periods. Just put a signal into a dummy load and wattmeter at the radio end, without the line and repeat the measurement at the far end. Any reduction in power delivered should be noted, and the loss calculated and recorded. Additional loss over time is an indication of degradation.

Checking Line Loss from One End

While measuring loss as described above is the most accurate method, sometimes it is not feasible to get equipment to the far end. In that case we can get a good measure of line loss from one end by measuring the SWR at the frequency of interest with the coax either open or shorted. If we had ideal lossless coax, our measured SWR would be infinite, since all the power would be reflected and would return to the source. With losses

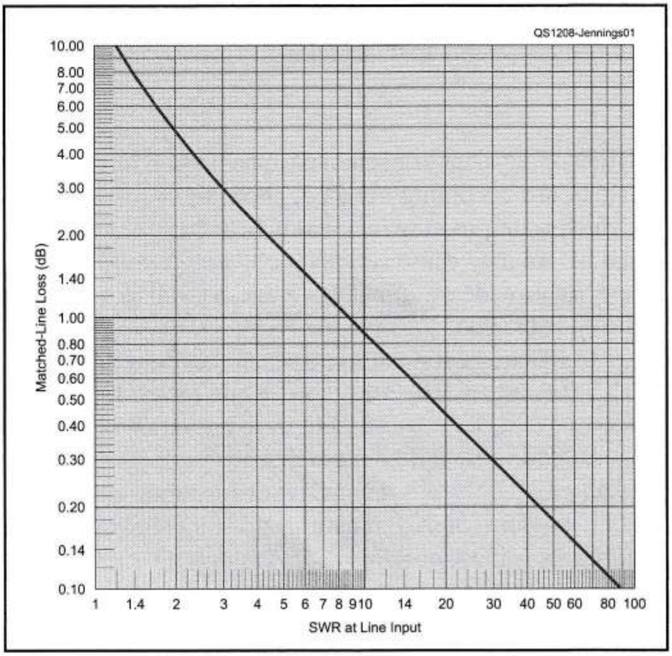

Figure 9.11 – Graph of the apparent SWR versus losses in a coax measured terminated in an open or short. Find the SWR value you measured on the horizontal axis. then go straight up and read the one way matched loss in dB on the vertical axis where the SWR reading crosses the diagonal line. Note that if the line is not matched, the loss can increase significantly.

in the coax, something else happens. Whatever signal power we put into one end will arrive at the other end with the same power, less the amount of line loss. Because the short or open at the far end of the coax provides a total reflection, the signal then reverses its direction and returns to the source end. On the return trip, the power level at the far end of the coax will be again reduced by the losses in the coax.

The reduced reflected power will show up in real coax as an SWR that is less than infinite. The higher the loss in the coax, the lower the measured SWR as read at the test end. Carried to extremes, a very high loss in a length of coax (such as the losses at UHF for a 500 foot long piece of RG-58) will result in an SWR of nearly 1:1 with the far end of the coax open or shorted.

Figure 9.11 shows a graph from the 16th edition of *The ARRL Antenna Book* of the SWR versus matched loss in a coax measured with a short or open at the far end. The section also provides an equation to determine the one-way matched loss from the measured SWR: L_M is the one-way matched line loss in decibels at the measurement frequency.

$$L_M = 10 \times \log_{10} \{(SWR + 1)/(SWR - 1)\}$$

In other words, if you measured an SWR of 3:1 with a length of open or shorted coax, either Figure 9.11 or the above equation indicates that the cable loss will be 3 dB at the frequency of interest.

To make the measurement, first make sure the antenna end of the coax is not connected to anything, and that any shorts you placed there during the ohmmeter tests have been removed. Follow the instructions in the manual for the antenna analyzer to make the SWR measurement.

As an example, let's say we measured the SWR of a 50 foot length of

RG-213 used at a dual-band VHF/UHF repeater site. At 147 MHz the open circuit SWR was 6.6:1. This indicates a loss of about 1.3 dB, compared to 1.4 dB predicted by *TLW* for 50 feet of RG-213 (Belden 8267).[2] At 445 MHz, the measured SWR was 1.4, indicating a loss of 7.78 dB, much worse than the 2.6 dB predicted by *TLW*. Thus this coax, while useful on 2 meters, will not be very good on 70 cm and is probably in the process of degrading.

What if the Antenna is Up in the Air?

The tests described previously are great if you have access to both ends of the coax, but sometimes you would like to find out about that coax without going up the tower. While you don't have control over the far end, you likely know what's there, if it's your antenna. For example, I have a 2 meter Yagi with a T matching section. While that provides a nice 50 Ω termination at 2 meters, it should look like a short at dc, and an ohmmeter continuity test should show a very low resistance, almost as low as if it were a short. A split feed Yagi or dipole should look like an open at dc.

The SWR at 2 meters won't tell us much about loss, in fact the lossier it is, the better the SWR will look. On the other hand, it will look a lot like a short at 1⁄10 the frequency or at 20 meters. A measurement there should give us a good, but not precise, idea of the 14 MHz loss. If the cable is good there, chances are it is also good at 2 meters. Other types of antenna connections may be trickier to make use of, but if you measure the SWR on other frequencies when the cable is new, and store the data in your archives, any change in later years may mean either the antenna or transmission line has undergone a change that merits investigation.

Another trick is to sweep your analyzer frequency looking for the worst SWR. That probably indicates either a very high or very low impedance termination and the loss at that frequency will be no better than that indicated by the SWR.

A Cautionary Note — Note that all of the loss results based on measured SWR are no better than the data provided by the SWR analyzer. *The ARRL Antenna Book* notes: "The instruments available to most amateurs lose accuracy at SWR values greater than about 5:1, so this method is useful principally as a go/no go check on lines that are fairly long." That really is the only question you need answer — do we want to keep this coax, or is it time for a replacement? These tests should give you enough information to make those decisions.

Notes

[1] L Nelson, AB0DZ, "Hints & Kinks — Anchoring Coaxial Feed Line," *QST*, Aug 2009, p 60.

[2] *The ARRL Antenna Book,* 16th Edition.

[3] *TLW, Transmission Line Program for Windows* software is provided on a CD with *The ARRL Antenna Book.*

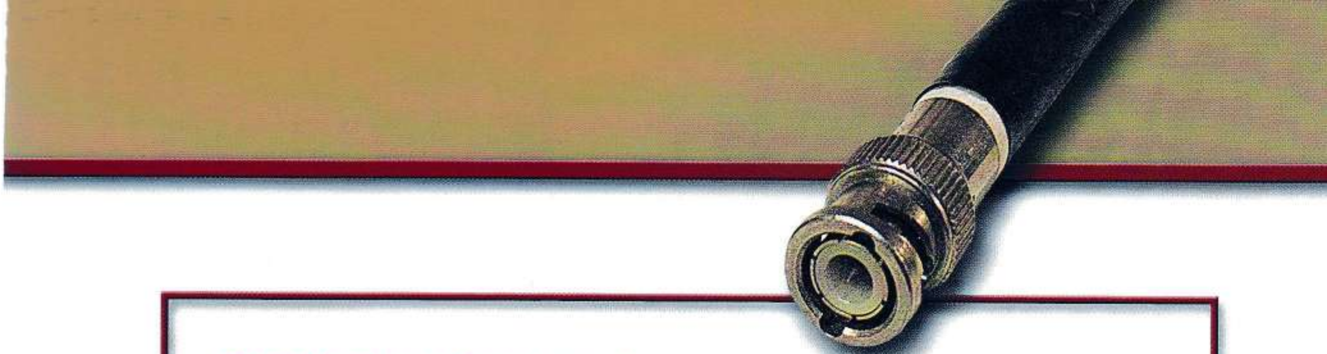

Eliminate Line Loss!

Transmission lines are an integral part of the art of radio. Designed to carry RF power over short to medium distances with minimal loss, properly selected and installed transmission lines can be as important as radio equipment and antennas are for a station's success.

The Care and Feeding of Transmission Lines is a radio amateur's guide to understanding transmission lines. It includes an introduction to the various types of transmission lines, key parameters of coaxial cable and balanced line types, as well as the different types of connectors. With this book, you'll learn how to select the most appropriate transmission line for an application and how to install it, as well as maintenance techniques to help extend the useful life of the line.

Includes:

- **What is a Transmission Line and Why Do We Need One?**
- **Let's Examine Coaxial Transmission Line**
- **Other Types of Unbalanced Line**
- **Let's Examine Balanced Transmission Line**
- **Transmission Line Interconnections**
- **Determining Which Line is Best Suited for a Particular Application**
- **Application and Installation Notes**
- **Transmission Line Care and Maintenance**

About ARRL and the Amateur Radio Service

Amateur (Ham) Radio provides the broadest and most powerful wireless communications capability available to any private citizen anywhere in the world. The principles of this federally licensed radio service include public service, radio experimentation, and international goodwill. ARRL is the national membership association for Amateur Radio operators. ARRL has books, software, online courses and other resources for licensing, operating and training.

ISBN 978-0-87259-478-4
52795
9 780872 594784
USA $27.95 ARRL Order No. 4784

Published by:
ARRL The national association for AMATEUR RADIO®
225 Main Street, Newington, CT 06111-1494 USA
www.arrl.org

ELECTRONIC WARFARE FUNDAMENTALS

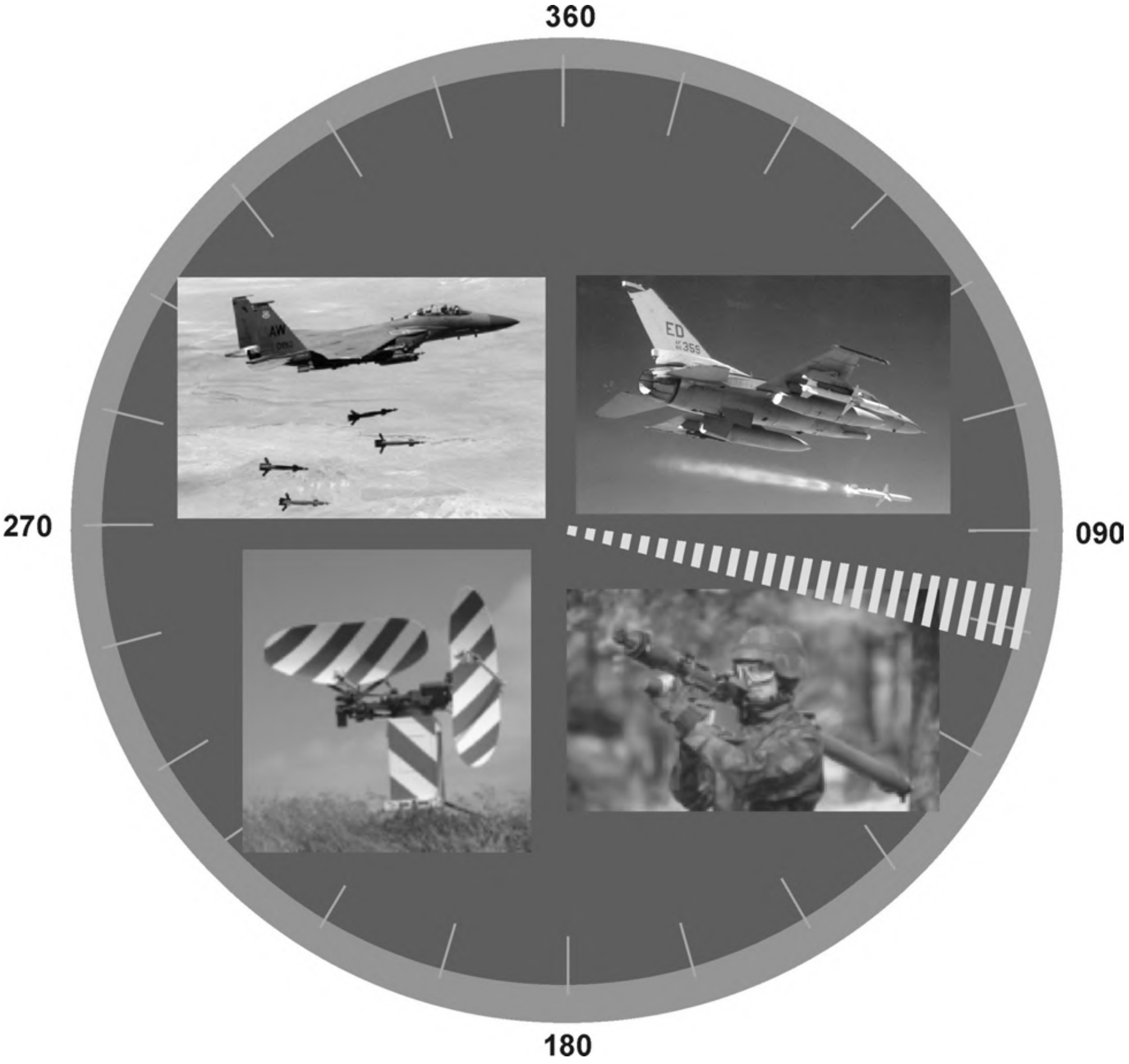
360
270
090
180
AW
ED

NOVEMBER 2000

PREFACE

Electronic Warfare Fundamentals **is a student supplementary text and reference book that provides the foundation for understanding the basic concepts underlying electronic warfare (EW). This text uses a practical building-block approach to facilitate student comprehension of the essential subject matter associated with the combat applications of EW.**

Since radar and infrared (IR) weapons systems present the greatest threat to air operations on today's battlefield, this text emphasizes radar and IR theory and countermeasures. Although command and control (C^2) systems play a vital role in modern warfare, these systems are not a direct threat to the aircrew and hence are not discussed in this book. This text does address the specific types of radar systems most likely to be associated with a modern integrated air defense system (IADS).

To introduce the reader to EW, ***Electronic Warfare Fundamentals*** **begins with a brief history of radar, an overview of radar capabilities, and a brief introduction to the threat systems associated with a typical IADS. The two subsequent chapters introduce the theory and characteristics of radio frequency (RF) energy as it relates to radar operations. These are followed by radar signal characteristics, radar system components, and radar target discrimination capabilities. The book continues with a discussion of antenna types and scans, target tracking, and missile guidance techniques.**

The next step in the building-block approach is a detailed description of countermeasures designed to defeat radar systems. The text presents the theory and employment considerations for both noise and deception jamming techniques and their impact on radar systems. This is followed by a chapter on decoys, both for defeating an IADS as well as for self-protection. Then, the next chapter discusses chaff characteristics, employment, and impact on specific radar systems.

The following two chapters are dedicated to the IR threat. The first covers IR theory, IR target detection and tracking, and advanced IR missile flare rejection techniques. The second chapter presents IR countermeasures to include flare employment, maneuvers, and missile warning equipment.

Electronic Warfare Fundamentals **then addresses an important aspect of EW, specifically electronic protection (EP). This section includes a description of the most common radar EP techniques designed to counter noise jamming, deception jamming, and chaff employment. The book concludes with an overview of the basic components and limitations of a typical radar warning receiver (RWR), current geolocation techniques, and it finally discusses the basic components of a self-protection jamming system.**

In addition to the textual material, the book also contains a detailed glossary of EW-related terms and acronyms for quick reference. A list of references is provided to acknowledge the source material used in the preparation of this text. These sources also provide the interested student with a supplementary reading list to learn more about a specific topic.

For information or comments on this document, please contact:

Det 8, ACC TRSS
4349 Duffer Drive, Ste 437
Nellis AFB NV 89191-7007

DSN: 682-4897
DSN FAX: 682-7391
Comm: (702) 652-4897
Comm FAX: (702) 652-7391

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION TO RADAR 1-1
1. INTRODUCTION 1-1
2. HISTORY 1-1
3. TARGET DISCRIMINANTS 1-2
4. INTRODUCTION TO INTEGRATED AIR DEFENSE SYSTEMS (IADS) 1-6
5. SUMMARY 1-10

CHAPTER 2. CHARACTERISTICS OF RF RADIATION 2-1
1. INTRODUCTION 2-1
2. FREQUENCY 2-1
3. WAVELENGTH 2-2
4. POLARIZATION 2-2
5. DOPPLER EFFECT 2-5
6. ELECTROMAGNETIC SPECTRUM 2-7
7. RF PROPAGATION 2-9
8. SUMMARY 2-14

CHAPTER 3. RADAR SIGNAL CHARACTERISTICS 3-1
1. INTRODUCTION 3-1
2. PULSE WIDTH (PW) 3-1
3. PULSE RECURRENCE TIME (PRT) 3-1
4. PULSE REPETITION FREQUENCY (PRF) 3-2
5. RADAR RECEIVER CHARACTERISTICS 3-3
6. DUTY CYCLE 3-4
7. PEAK POWER 3-4
8. AVERAGE POWER 3-5
9. MODULATION 3-6
10. SUMMARY 3-8

CHAPTER 4. RADAR SYSTEM COMPONENTS 4-1
1. INTRODUCTION 4-1
2. PULSE RADAR SYSTEM 4-1
3. CONTINUOUS WAVE (CW) RADAR 4-7
4. PULSE DOPPLER RADAR 4-9
5. MONOPULSE RADAR 4-10
6. SUMMARY 4-13

CHAPTER 5. RADAR PRINCIPLES 5-1
1. INTRODUCTION....................5-1
2. RADAR RANGE 5-1
3. AZIMUTH DETERMINATION 5-7
4. ELEVATION DETERMINATION....................5-11
5. RADAR RESOLUTION CELL 5-12
6. PULSE DOPPLER VELOCITY DETERMINATION 5-14
7. BASIC RADAR EQUATION 5-18
8. SUMMARY 5-21

CHAPTER 6. ANTENNA CHARACTERISTICS AND SCANS 6-1
1. INTRODUCTION....................6-1
2. PARABOLIC ANTENNA 6-1
3. CASSEGRAIN ANTENNA....................6-3
4. PHASED ARRAY ANTENNA 6-5
5. ANTENNA GAIN....................6-9
6. POWER DENSITY....................6-11
7. CIRCULAR SCAN 6-12
8. LINEAR SCAN....................6-16
9. UNIDIRECTIONAL SCAN 6-16
10. BIDIRECTIONAL SCAN....................6-17
11. CONICAL SCAN....................6-18
12. TRACK-WHILE-SCAN 6-20
13. IMPACT OF TERRAIN ON RADAR SCANS....................6-23
14. SUMMARY....................6-26

CHAPTER 7. TARGET TRACKING....................7-1
1. INTRODUCTION....................7-1
2. RANGE TRACKING 7-2
3. CONICAL SCAN....................7-4
4. TRACK-WHILE-SCAN (TWS) 7-6
5. LOBE-ON-RECEIVE-ONLY (LORO)....................7-11
6. MONOPULSE RADAR 7-12
7. CONTINUOUS WAVE (CW) RADARS....................7-16
8. PULSE DOPPLER RADAR....................7-20
9. SUMMARY 7-24

CHAPTER 8. RADAR MISSILE GUIDANCE TECHNIQUES....................8-1
1. INTRODUCTION....................8-1
2. COMMAND GUIDANCE....................8-3
3. SEMI-ACTIVE GUIDANCE....................8-6
4. ACTIVE GUIDANCE....................8-9
5. SEEKER-AIDED GROUND GUIDANCE/TRACK-VIA-MISSILE GUIDANCE...8-11

6. ANTIAIRCRAFT ARTILLERY (AAA) 8-12
7. SUMMARY 8-15

CHAPTER 9. INTRODUCTION TO RADAR JAMMING 9-1
1. INTRODUCTION 9-1
2. RADAR JAMMING TYPES 9-1
3. RADAR JAMMING EMPLOYMENT OPTIONS 9-3
4. FUNDAMENTALS OF RADAR JAMMING 9-7
5. SUMMARY 9-15

CHAPTER 10. RADAR NOISE JAMMING 10-1
1. INTRODUCTION 10-1
2. RADAR NOISE JAMMING EFFECTIVENESS 10-1
3. RADAR NOISE JAMMING GENERATION 10-4
4. BARRAGE JAMMING 10-5
5. SPOT JAMMING 10-6
6. SWEPT-SPOT JAMMING 10-7
7. COVER PULSE JAMMING 10-8
8. MODULATED NOISE JAMMING 10-10
9. SUMMARY 10-11

CHAPTER 11. DECEPTION JAMMING 11-1
1. INTRODUCTION 11-1
2. FALSE TARGET JAMMING 11-3
3. RANGE DECEPTION JAMMING 11-5
4. ANGLE DECEPTION JAMMING 11-8
5. VELOCITY DECEPTION JAMMING 11-11
6. MONOPULSE DECEPTION JAMMING 11-17
7. TERRAIN BOUNCE 11-23
8. SUMMARY 11-24

CHAPTER 12. DECOYS 12-1
1. INTRODUCTION 12-1
2. SATURATION DECOYS 12-1
3. TOWED DECOYS 12-3
4. EXPENDABLE ACTIVE DECOYS 12-6
5. SUMMARY 12-7

CHAPTER 13. CHAFF EMPLOYMENT 13-1
1. INTRODUCTION 13-1
2. CHAFF CHARACTERISTICS 13-2
3. CHAFF OPERATIONAL EMPLOYMENT 13-9
4. SUMMARY 13-16

CHAPTER 14. IR FUNDAMENTALS 14-1
1. INTRODUCTION 14-1
2. BASIC IR THEORY 14-1
3. IR SIGNATURE SOURCES 14-5
4. IR SEEKER CHARACTERISTICS 14-7
5. IR SEEKER TYPES 14-9
6. IR MISSILE FLARE REJECTION 14-12
7. SUMMARY 14-18

CHAPTER 15. IR COUNTERMEASURES 15-1
1. INTRODUCTION 15-1
2. FLARE CHARACTERISTICS 15-1
3. ADVANCED FLARES 15-3
4. FLARE EMPLOYMENT 15-4
5. IRCM TACTICS 15-5
6. MISSILE APPROACH WARNING SYSTEMS (MAWS) 15-6
7. SUMMARY 15-8

CHAPTER 16. RADAR ELECTRONIC PROTECTION (EP) TECHNIQUES 16-1
1. INTRODUCTION 16-1
2. RADAR RECEIVER PROTECTION 16-1
3. JAMMING SIGNAL AVOIDANCE 16-4
4. JAMMING SIGNAL EXPLOITATION 16-5
5. OVERPOWERING THE JAMMING SIGNAL 16-6
6. PULSE DURATION DISCRIMINATION 16-7
7. ANGLE DISCRIMINATION 16-7
8. BANDWIDTH DISCRIMINATION 16-8
9. DOPPLER DISCRIMINATION 16-9
10. TIME DISCRIMINATION 16-9
11. SUMMARY 16-10

CHAPTER 17. RADAR WARNING RECEIVER (RWR) BASIC OPERATIONS AND GEOLOCATION TECHNIQUES 17-1
1. INTRODUCTION 17-1
2. RWR ANTENNAS 17-1
3. RWR RECEIVER/AMPLIFIERS 17-3
4. SIGNAL PROCESSOR 17-6
5. EMITTER IDENTIFICATION (EID) TABLES 17-8
6. RWR SCOPE DISPLAY 17-8
7. RWR AUDIO 17-9
8. RWR INTERFACE CONTROL UNIT (ICU) 17-10
9. RWR LIMITATIONS 17-10
10. THREAT GEOLOCATION TECHNIQUES 17-15
11. SUMMARY 17-20

CHAPTER 18. SELF-PROTECTION JAMMING SYSTEM OPERATIONS.............. 18-1
1. INTRODUCTION... 18-1
2. RECEIVE ANTENNAS.. 18-2
3. RECEIVER SECTION... 18-2
4. SYSTEM PROCESSOR... 18-3
5. JAMMING TECHNIQUES GENERATOR... 18-4
6. TRANSMIT ANTENNAS.. 18-4
7. C-9492 CONTROL INDICATOR UNIT.. 18-5
8. JAMMING POD CONSIDERATIONS 18-5
9. SUMMARY... 18-7

ANNEX A. GLOSSARY...A-1

ANNEX B. EC-RELATED ABBREVIATIONS AND ACRONYMS.............................B-1

LIST OF REFERENCES

CHAPTER 1. INTRODUCTION TO RADAR

1. INTRODUCTION

The word "RADAR" is an acronym for RAdio Detection And Ranging. As it was originally conceived, radio waves were used to detect the presence of a target and to determine its distance or range (Figure 1-1).

Figure 1-1. Radar System

2. HISTORY

The reflection of radio waves by objects was first noted more than a century ago. In 1903, the reflection of radio waves was employed in Germany to demonstrate detection of ships at sea (Figure 1-2). In 1922, Marconi presented the same idea in Britain but received little official interest. These early experiments used continuous wave, or CW, transmissions and relied on the reflection of a transmitted wave from a target to indicate the presence of a target. CW transmissions can detect the presence of an object and, if the radio wave is formed into a narrow beam, can also provide azimuth information. CW transmissions cannot provide range.

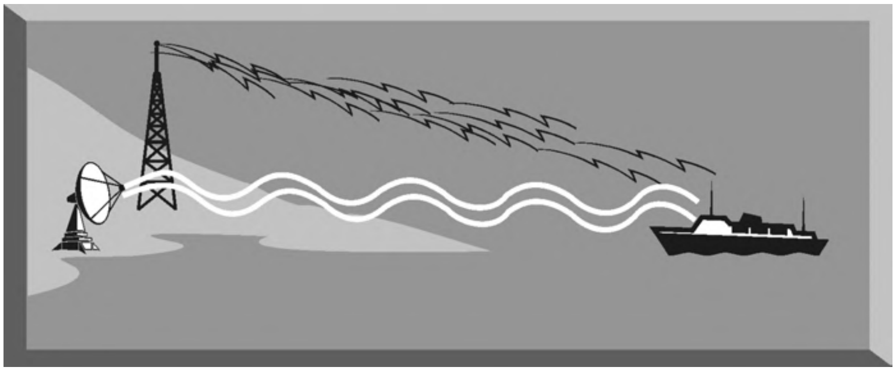

Figure 1-2. Early Continuous Wave Radar Experiments

a. The lack of range information was a serious limitation but was finally overcome by modulating the radio wave transmissions to send out a train of short pulses (Figure 1-3). The time between pulse transmission and an echo return to the receiver provides a direct measurement of range. Practical development of the pulse radar began in the 1930s, principally in the United States, Britain, and Germany. Due to deteriorating relations with Germany and the threat of invasion, the British intensified their efforts to develop a pulse radar in 1935. These efforts culminated in the development and deployment of multiple radar stations which formed the Chain Home System. These radar installations provided critical information to British pilots on the size and location of German bomber formations during the Battle of Britain.

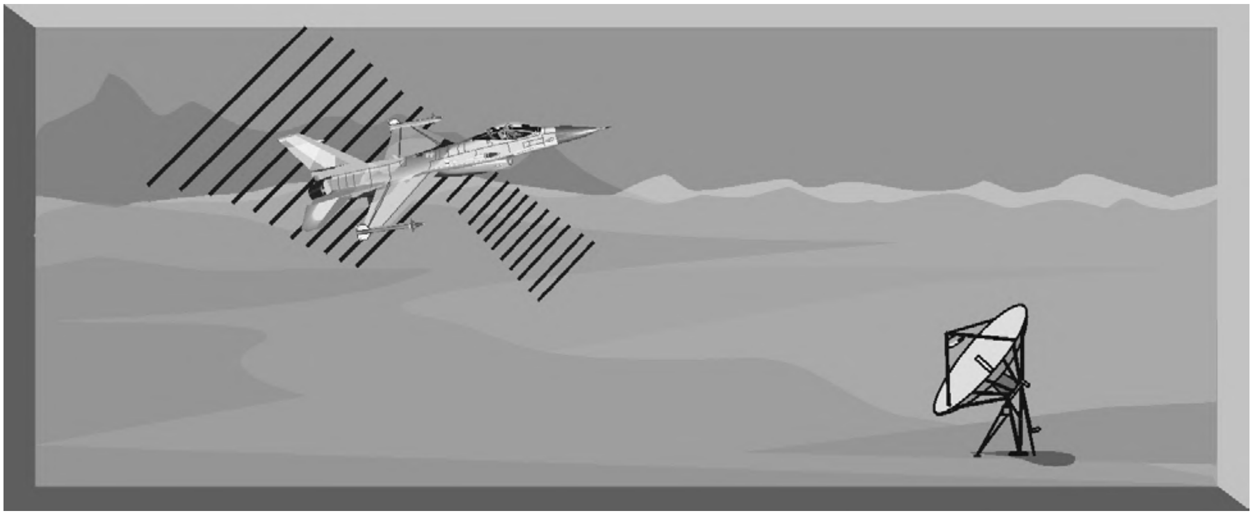

Figure 1-3. Pulse Radar Operation

b. The Chain Home System is considered the first integrated air defense system, or IADS. Radar development, with its obvious military and civilian applications, has continued unabated to the present day.

3. TARGET DISCRIMINANTS

The widespread military and civilian application of radar is based on its inherent advantages over the human eye (Figure 1-4). Radar can "see" farther than the human eye and more accurately assess the range or distance of an object. Radar works well in all-weather conditions and is relatively immune to smoke, haze, and clouds. What's more, radar works 24 hours a day because it can transmit its own energy and does not have to rely on sunshine or ambient radiation. There are some disadvantages of radar when compared to the human eye. First, radar does not have the resolution that the human eye has. While radar can detect the presence of an airplane, the human eye can discern, in great detail, the shape, size, color, and even markings. This can be a serious limitation if positive identification is required prior to engagement.

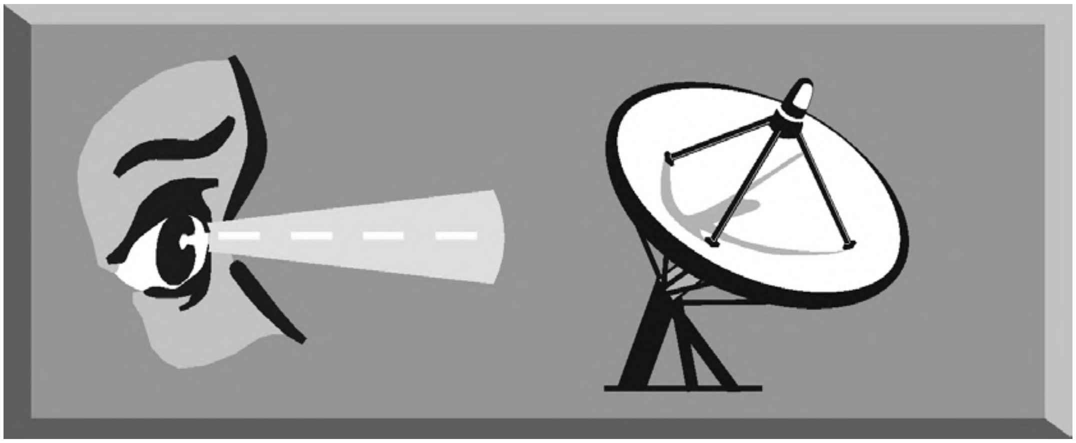

Figure 1-4. Radar Advantages and Disadvantages

Second, the human eye is not bothered by undesirable reflections, called clutter, the way radar sometimes is. Although metal is the best reflector of radio frequency (RF) energy, nearly any material will reflect some RF energy. Mountains, trees, buildings, rain, birds, and chaff all reflect RF energy. Radar systems must use target discriminants to isolate the desired target return from the clutter. These target discriminants include range, velocity, and angle (Figure 1-5).

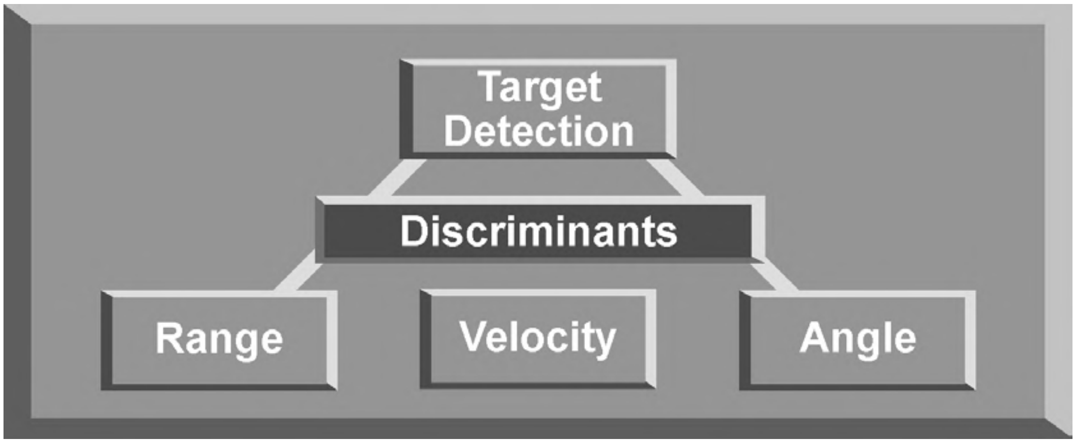

Figure 1-5. Radar Target Discriminants

a. The first target discriminant is range. The time an RF wave takes to go to, and return from, a target allows measurement of the range to that target. We know that RF energy travels at the speed of light, or “c” which is 3 x 10^8 meters per second. Target range can be determined by using the basic radar range equation (Equation 1-1), target range equals measured time, multiplied by the speed of light (“c”), divided by 2.

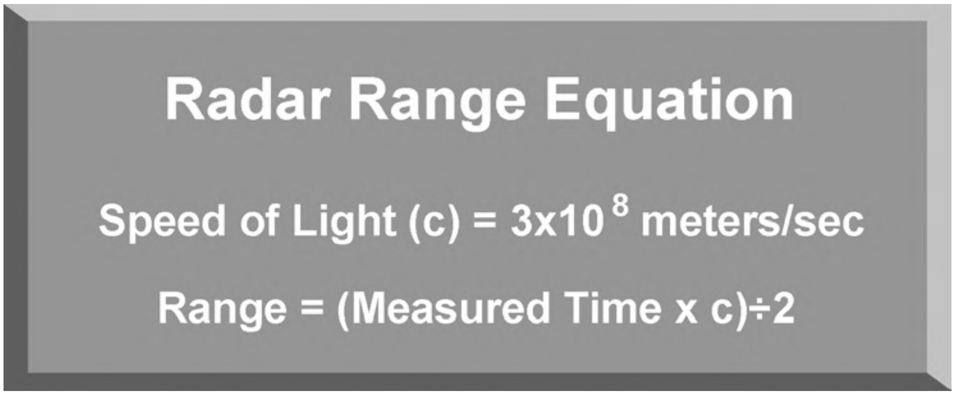

Radar Range Equation

Speed of Light (c) = $3x10^{8}$ meters/sec

Range = (Measured Time x c)÷2

Equation 1-1. Basic Radar Range Equation

b. Target angle discrimination is another critical capability of radar systems. In order for a radar system to detect a target, the antenna must be pointed at the target during the transmission and reception of RF energy. The ability of a radar system to accurately determine angle is a function of the horizontal beamwidth of the antenna. If the radar sweep is referenced to true North, the angle of a radar return can be measured relative to true North (Figure 1-6).

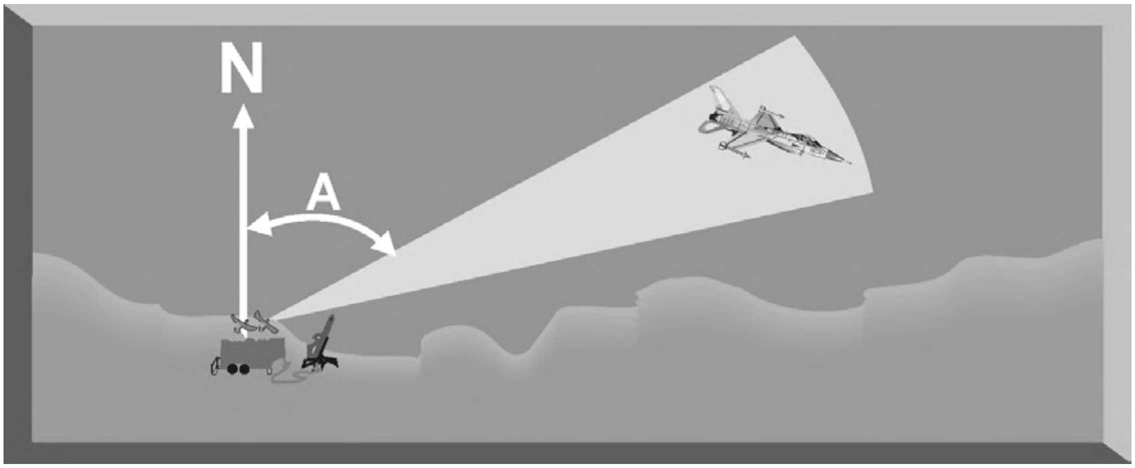

Figure 1-6. Angle Discrimination

c. Velocity discrimination is a specific capability of CW and pulse Doppler radar systems. The transmitters of CW radars send out continuous RF at a specific frequency (Figure 1-7). The reflected signal frequency is changed, or shifted, by a specific amount by a moving target. This frequency shift, called the Doppler effect, allows the measurement of the velocity of that target relative to the radar. The receiver measures this frequency difference which equates to a specific radial velocity. Pulse Doppler radars can measure the Doppler effect while still obtaining the range.

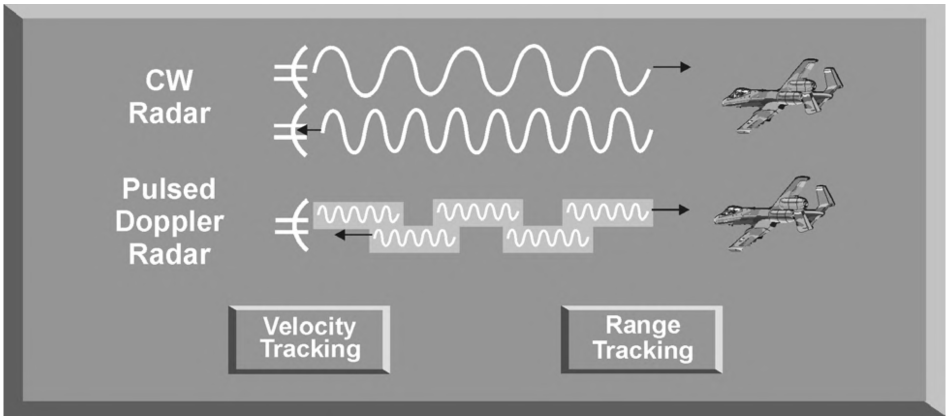
CW
Radar
Pulsed
Doppler
Radar
Velocity
Tracking
Range
Tracking

Figure 1-7. Velocity Discrimination

d. A basic pulse radar system consists of a transmitter, antenna, receiver, and a master timer (Figure 1-8). The transmitter sends electromagnetic energy (RF) to the antenna. This energy is radiated through the atmosphere. When this RF energy is interrupted by any object, such as a plane, ship, or the earth, a portion of the RF energy is reflected back to the antenna and processed by the receiver.

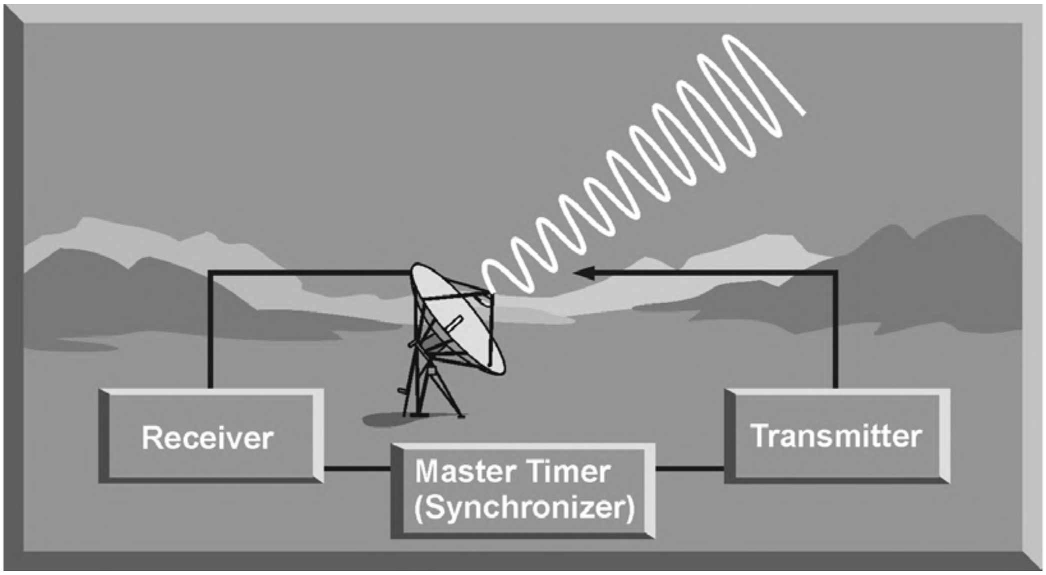
Receiver
Master Timer
(Synchronizer)
Transmitter

Figure 1-8. Basic Radar Operation

The reflection is called an echo, and the object interrupting the RF energy is called a target. The presence of an echo indicates target detection. If the target detected is the desired target, the echo is referred to as a target signal. If the echo is from undesired targets, such as the earth, the echo is referred to as clutter. This is especially true when the undesirable echoes make the detection of the desired target difficult. The capability of the antenna to focus the RF energy

affects the angular discrimination of the radar. The ability of the master timer to determine the time between RF transmission and target echo reception impacts the range determination capability of the radar. The ability of the receiver to analyze the Doppler frequency shift in the target echo determines the velocity discrimination capability of the radar and the ability of the radar to reject clutter.

4. INTRODUCTION TO INTEGRATED AIR DEFENSE SYSTEMS (IADS)

Radar systems have the inherent capability to determine accurate range, azimuth, and/or velocity information on airborne targets. Radar systems can provide this information in nearly all types of weather, day or night, and at distances that far exceed the capabilities of the human eye. Military commanders have taken advantage of these capabilities by employing radar systems to provide air defense for high-value targets. The primary missions of radar systems employed for air defense are attack warning and threat engagement.

a. Radar systems specifically designed to provide attack warning are called early warning (EW) radars (Figure 1-9). These radars are characterized by high power output, large antennas, and low frequencies. These same characteristics limit the accuracy of the target parameters available from early warning radars. The long-range detection of aircraft and the earliest possible attack warning capabilities of early warning radars provide the first line of defense for the air defense system.

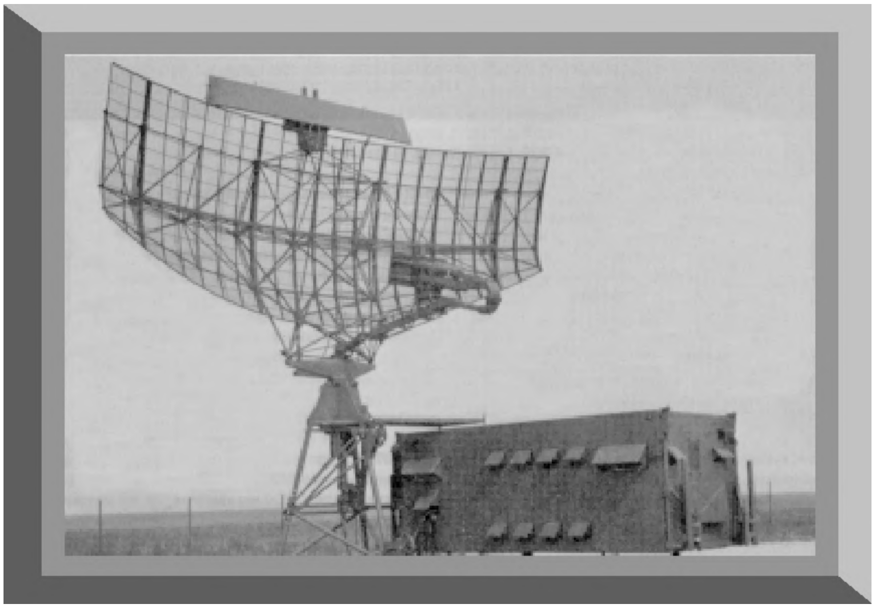

Figure 1-9. Early Warning Radar System

b. Radar systems designed to provide target engagement information include ground control intercept (GCI) radars, acquisition radars, target tracking radars (TTRs), and airborne interceptor (AI) radars.

(1) GCI radars are designed to provide sufficiently accurate target aircraft range, azimuth, and altitude information to vector AI assets to intercept and destroy attacking aircraft (Figure 1-10). To provide this data, early warning radars can be deployed along with specialized height finder radars. This combination of radar systems is commonly referred to as a GCI site. Newer GCI radar systems, employing phased array antennas and Doppler processing, can provide the required 3-dimensional target information. Any radar system, or combination of radar systems, that can determine 3-dimensional target data, and is equipped with the communication equipment to pass this information to AI assets, can act as a GCI site. GCI radar systems can be used to supplement early warning radar systems to provide critical attack warning.

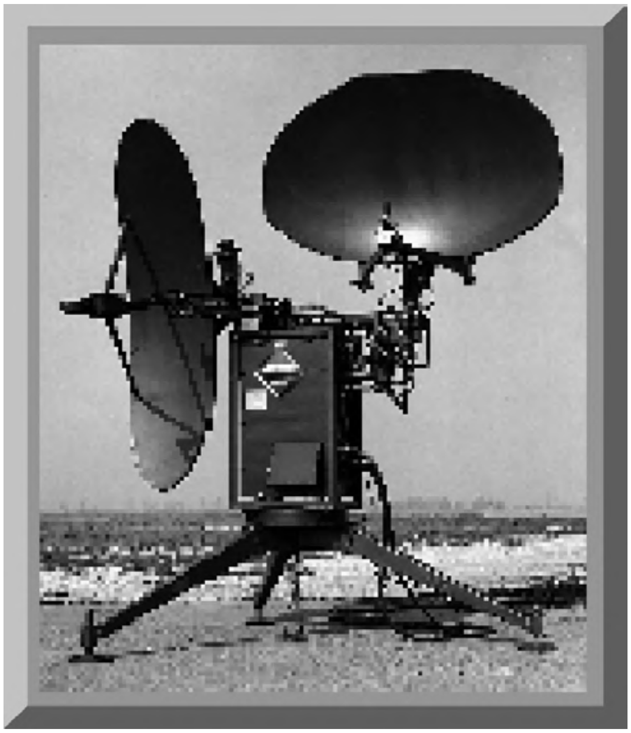

Figure 1-10. GCI Radar

(2) Acquisition radar systems are designed to act as GCI radars for ground based TTRs. Acquisition radar systems generally have shorter range capability than early warning radars and operate at higher frequencies. These radar systems provide accurate target range and azimuth data to TTRs to facilitate target engagement. Acquisition radars can be a distinct radar system (Figure 1-11) or be incorporated as part of the TTR (Figure 1-12).

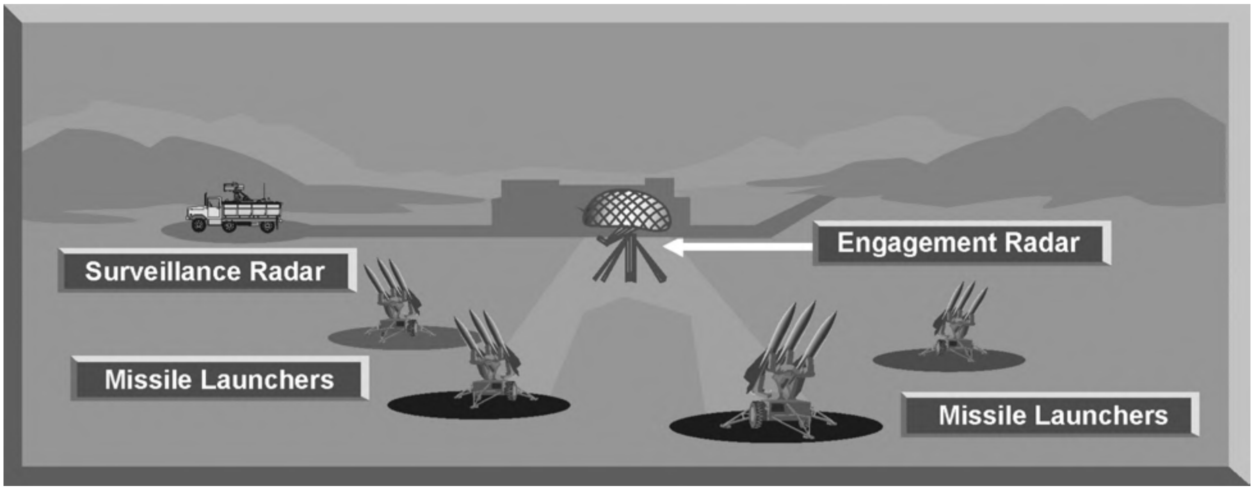
Surveillance Radar
Engagement Radar
Missile Launchers
Missile Launchers

Figure 1-11. Acquisition Radar and TTR

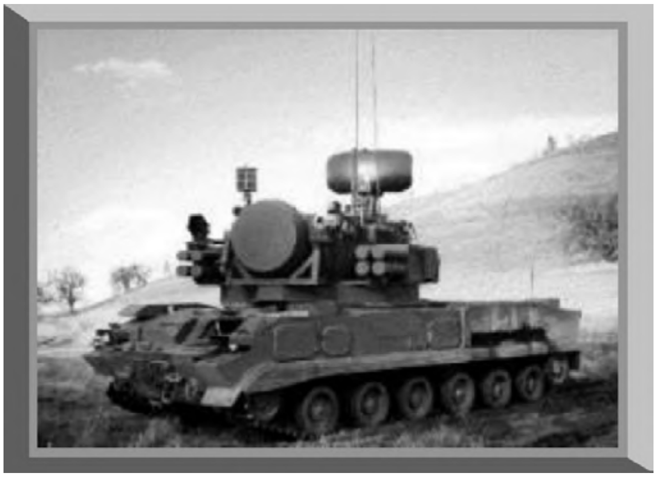

Figure 1-12. TTR with Acquisition

(3) The primary role of TTRs, in support of an air defense system, is to provide continuous and accurate target parameters to a fire control computer. The fire control computer uses this data to guide missiles or aim antiaircraft artillery (AAA) to destroy attacking aircraft. TTRs employ various tracking techniques to continuously update target parameters. TTRs generally employ high frequencies, narrow beamwidths, and computer signal processing to enhance the accuracy of target parameters provided to the fire control computer.

(4) AI radar systems are TTRs employed by fighter aircraft to engage and destroy airborne targets (Figure 1-13). These radar systems are characterized by high frequency, sophisticated computer processing, and accurate target tracking capability. They are designed to allow the AI asset to employ air-to-air missiles and guns/cannons. TTRs and AI radars constitute the highest radar threat associated with an air defense system.

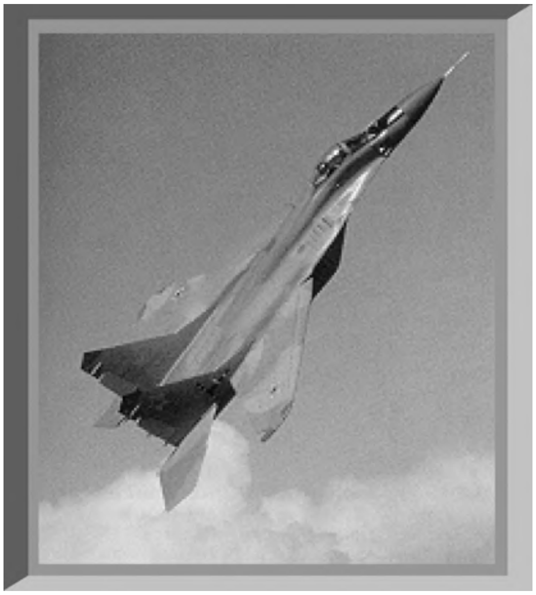

Figure 1-13. MiG-29

(5) Another growing lethal threat associated with an air defense system is infrared (IR) missiles. IR missile systems can be man-portable (Figure 1-14), mounted on vehicles, or employed by AI assets. These missile systems guide on the distinctive IR signature of aircraft. The recent proliferation and enhanced performance of IR systems has increased the contribution of these systems to air defense.

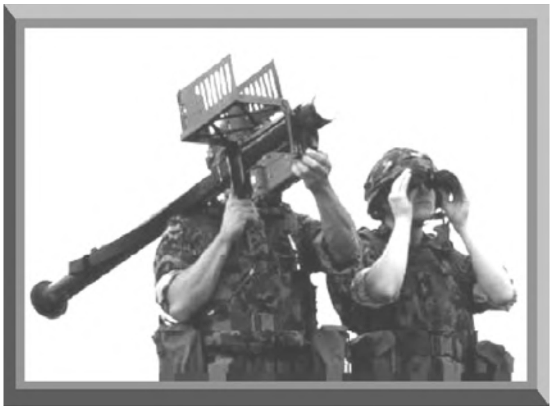

Figure 1-14. IR System

c. All these radar systems can be deployed to provide air defense for a particular country or geographical area. When the employment of these radar systems is integrated by a command and control (C^2) structure, this constitutes an IADS (Figure 1-15). The C^2 structure allows the military commander to take advantage of the threat warning provided by early warning radars. Based on this threat warning, the military commander can allocate specific assets (GCI and AI assets, or acquisition radars and TTRs) to engage airborne targets. This allocation decision is based on the capabilities of these systems and the tactical situation. This allocation process enables the military commander to maximize the capabilities of his forces to engage and destroy attacking aircraft.

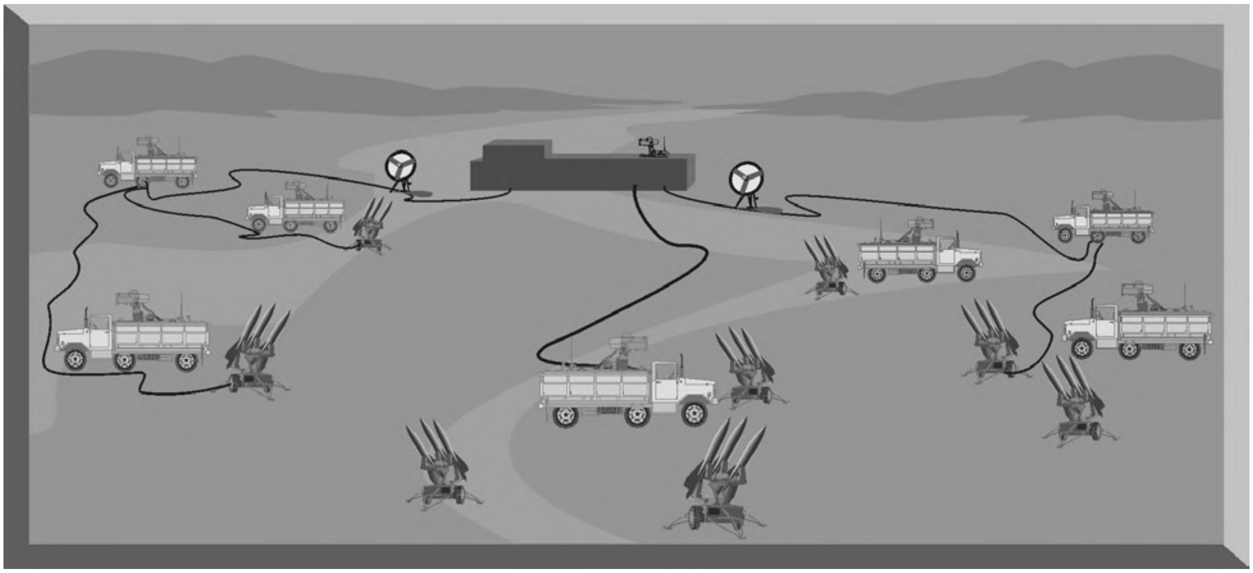

Figure 1-15. A Typical IADS

5. SUMMARY

Radar systems are the cornerstones of a modern IADS. Radar and IR threat systems operate at frequencies that span most of the electromagnetic spectrum. Each system has unique capabilities and operating characteristics that enable it to accomplish assigned tasks in support of the IADS. In order to effectively employ offensive air power on the modern battlefield, the systems that support the IADS must be negated. A basic knowledge of how radar and IR systems operate, their capabilities, limitations, and the available countermeasures is the key to defeating these systems. The purpose of this book is to provide this information.

CHAPTER 2. CHARACTERISTICS OF RF RADIATION

1. INTRODUCTION

In order for a radar system to determine range, azimuth, elevation, or velocity data, it must transmit and receive electromagnetic radiation. This electromagnetic radiation is referred to as radio frequency (RF) radiation. RF transmissions have specific characteristics that determine the capabilities and limitations of a radar system to provide these target discriminants, based on an analysis of the characteristics of the target return. The frequency of transmitted RF energy affects the ability of a radar system to analyze target return, based on time, to determine target range. RF frequency also affects the ability of the transmitting antenna to focus RF energy into a narrow beam to provide azimuth and elevation information. The wavelength and frequency of the transmitted RF energy impact the propagation of the radar signal through the atmosphere. The polarization of the RF signal affects the amount of clutter the radar must contend with. The ability of a radar system to use the Doppler effect in analyzing the radar return impacts the velocity discrimination capability of the radar. These characteristics of RF radiation will be discussed in this chapter.

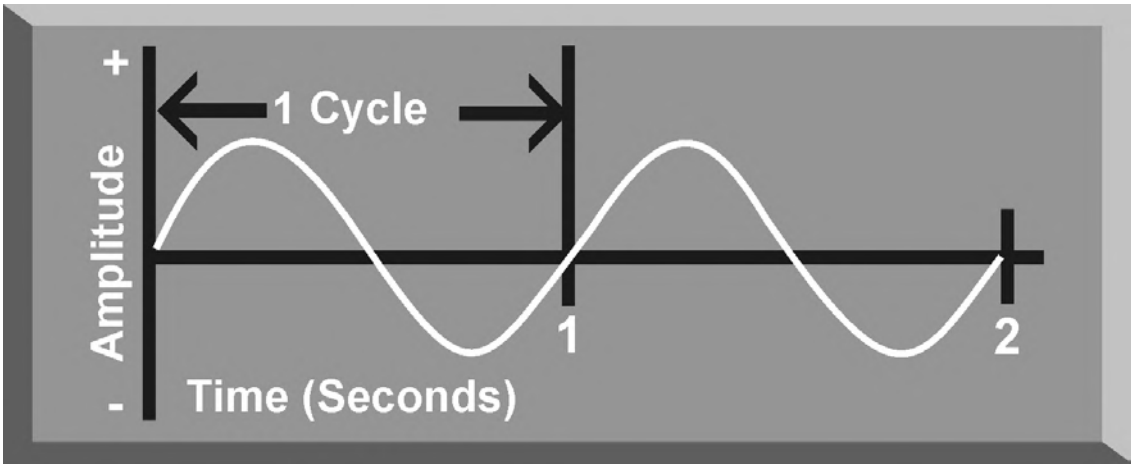

Figure 2-1. Radio Frequency

2. FREQUENCY

The output signal from a typical radio or radar system has several important characteristics that affect the capabilities and limitations of radio or radar systems. The first characteristic considered is usually RF. The frequency of the transmitted signal is the number of times per second the RF energy completes one cycle. The RF signal depicted in Figure 2-1 has a frequency of one cycle per second. The basic unit of measurement is the hertz (Hz). One hertz equals one cycle per second. Most radars have an RF in the millions of hertz, or megahertz (MHz).

3. WAVELENGTH

A characteristic of any RF signal is wavelength. Wavelength is a measure of the physical distance between peaks of a sine wave propagated in space (Figure 2-2).

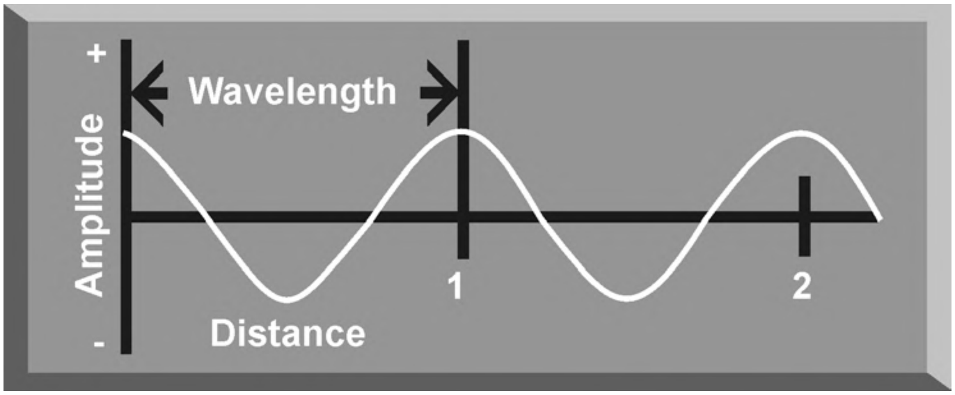

Figure 2-2. Radar Signal Wavelength

Though wavelength is measured in meters, most radar signals have wavelengths measured in centimeters or millimeters. The wavelength of a radar signal can be computed using the equation shown in Equation 2-1. The relationship between wavelength and frequency is inverse: the higher the frequency, the shorter the wavelength. In early radio and radar terminology, wavelength was used instead of frequency to describe operating characteristics of a system. Today, wavelength is used to describe systems operating at very high frequencies, such as millimeter wave, and for describing infrared (IR) systems.

$$\text{Wavelength}\ (\lambda) = \frac{c\quad \text{(Speed of Light)}}{f\quad \text{(Radar Frequency)}}$$

Equation 2-1. Basic Wavelength Equation

4. POLARIZATION

Another characteristic of a radio frequency wave is polarization. Polarization is determined by the radar antenna and refers to the orientation of the RF wave as it travels through space. There are two types of polarization: linear and circular.

a. Traveling electromagnetic energy has two components: an electrostatic field and a magnetic field. These two fields are always perpendicular to each other and perpendicular to the direction of travel. The polarization of the wave is defined in terms of the orientation to the electrostatic field. Many radar antennas are linearly polarized, either vertically or horizontally. The signal depicted in Figure 2-3 is vertically polarized.

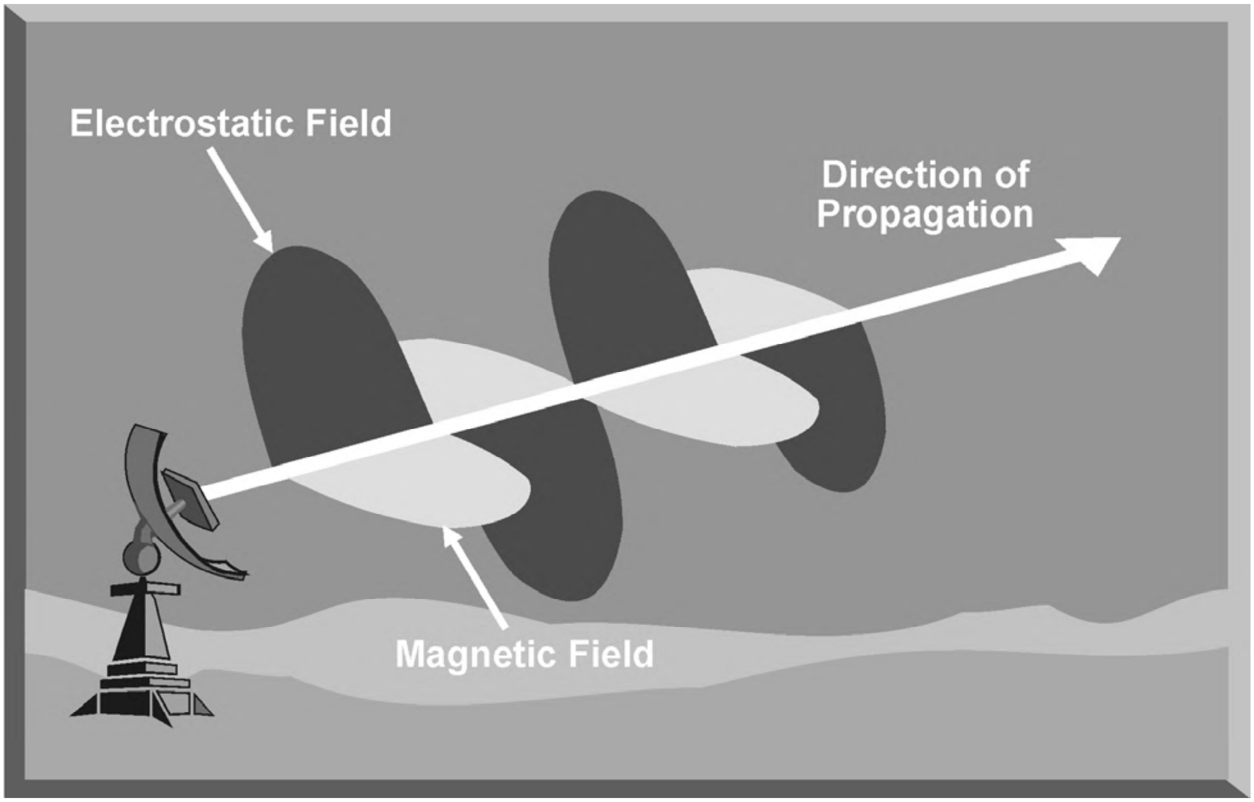

Figure 2-3. Radar Signal Vertical Polarization

b. Some radars use circular polarization to improve target detection in rain. Circular polarization can be right-hand, or left-hand orientation. For circular polarization, the direction of the electrostatic field varies with time and traces a circular locus about a fixed plane perpendicular to the direction of propagation. For a right-hand circular polarized signal, the electrostatic vector appears to rotate in a clockwise direction. For a left-hand circular polarized signal, the rotation is counterclockwise. Circular polarization can be visualized by pointing the thumb of either hand in the direction of propagation and curling the fingers in the direction of electrostatic field rotation (Figure 2-4).

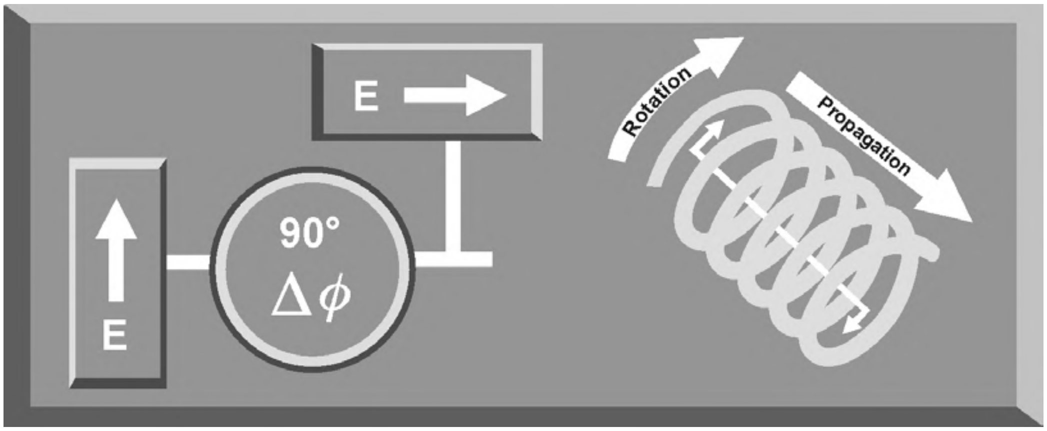
E
90°
$\Delta\phi$
E
Rotation
Propagation

Figure 2-4. Circular Polarization

c. The impact of polarization on receivers and transmitters is fairly straightforward. If an antenna is designed to receive a particular polarization, it will have difficulty receiving a signal with an opposite polarization. This situation is defined as cross polarization (Figure 2-5).

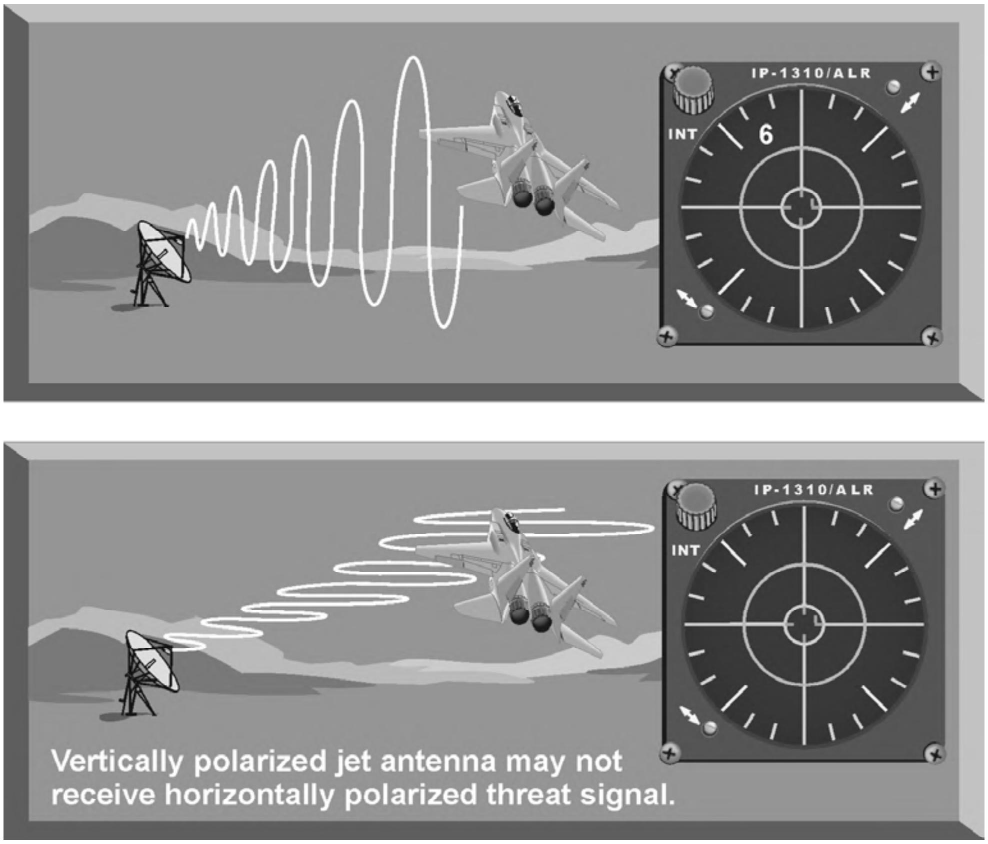
IP-1310/ALR
INT
6
IP-1310/ALR
INT

Vertically polarized jet antenna may not receive horizontally polarized threat signal.

Figure 2-5. Impact of Polarization

The impact of cross polarization on electronic combat can be dramatic. If a radar warning receiver antenna is polarized to receive vertically polarized signals, a threat system employing a horizontally polarized radar signal may not be detected, or may be displayed on the scope well after the threat has acquired the aircraft. In addition, if the jamming antenna on an electronic attack (EA) system is also vertically polarized, it may not be able to jam this system. Fortunately, this potentially lethal situation rarely occurs, but future threat systems may take advantage of this situation. Table 2-1 details the impact of polarization on selected transmit and receive antenna combinations.

Table 2-1. Antenna Polarization Loss

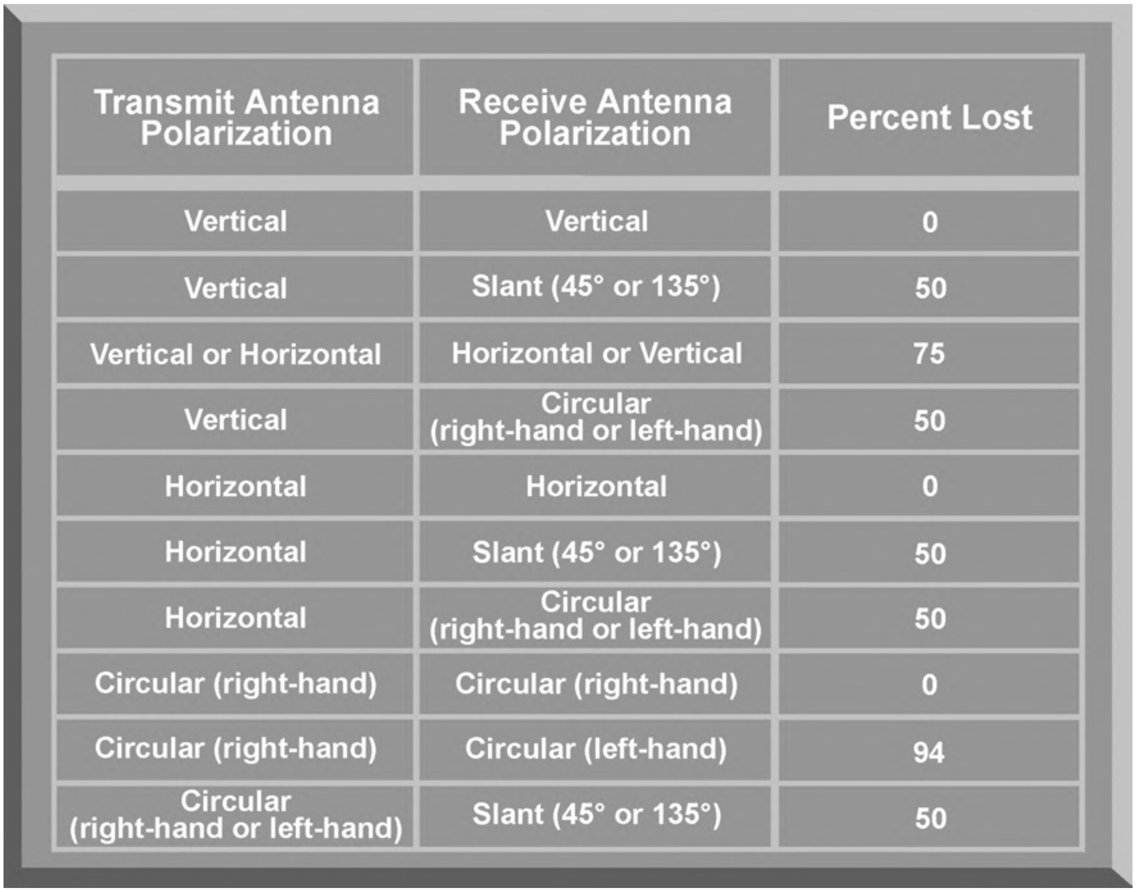

Transmit Antenna Polarization	Receive Antenna Polarization	Percent Lost
Vertical	Vertical	0
Vertical	Slant (45° or 135°)	50
Vertical or Horizontal	Horizontal or Vertical	75
Vertical	Circular (right-hand or left-hand)	50
Horizontal	Horizontal	0
Horizontal	Slant (45° or 135°)	50
Horizontal	Circular (right-hand or left-hand)	50
Circular (right-hand)	Circular (right-hand)	0
Circular (right-hand)	Circular (left-hand)	94
Circular (right-hand or left-hand)	Slant (45° or 135°)	50

5. DOPPLER EFFECT

The “Doppler effect” takes advantage of the fact that the frequency of RF waves will be changed or shifted when reflected from a target moving relative to the radar. The shifted frequency of the returning RF wave depends on the movement of the aircraft in relation to the radar. In Figures 2-6, 2-7, and 2-8, f_o is the transmitted frequency of the radar, and f_t is the frequency of the reflected RF

wave from the target. For a stationary target, the frequency of the reflected signal will equal the frequency of the transmitted signal (Figure 2-6).

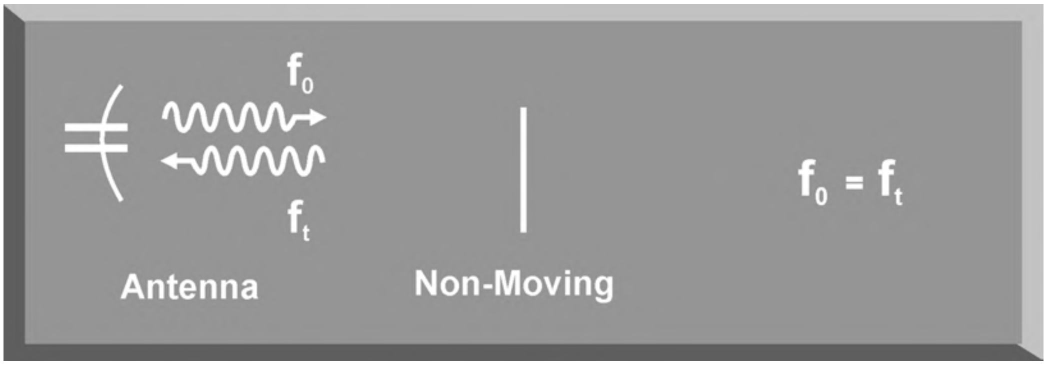
f_0
f_t
Antenna
Non-Moving
$f_0 = f_t$

Figure 2-6. Zero Doppler Effect – Stationary Target

a. For a target moving toward the radar, the frequency of the reflected signal will be higher than the transmitted signal (Figure 2-7).

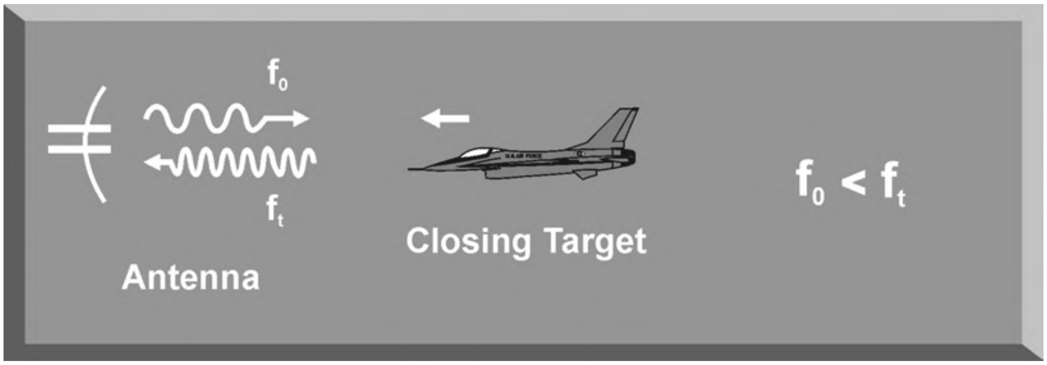
f_0
f_t
Antenna
Closing Target
$f_0 < f_t$

Figure 2-7. Doppler Effect – Closing Target

b. The reflected frequency for a target moving away from the radar will be lower than the transmitted frequency (Figure 2-8).

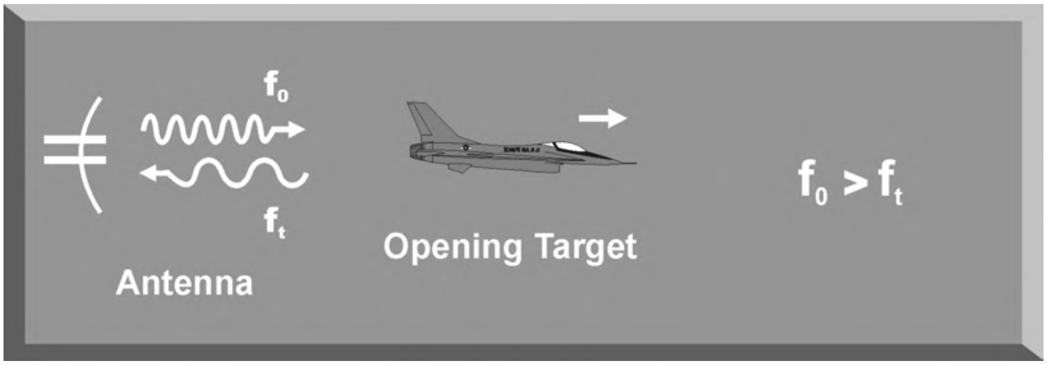
f_0
f_t
Antenna
Opening Target
$f_0 > f_t$

Figure 2-8. Doppler Effect – Opening Target

6. ELECTROMAGNETIC SPECTRUM

The portion of the electromagnetic spectrum (Figure 2-9) that today's electronic combat systems must deal with starts with radio waves and encompasses microwaves, infrared, and a small portion of the ultraviolet region. Communications systems generally operate in the HF, UHF, and VHF regions. Some satellite communications operate in the SHF region. Radars operate in the microwave region, normally from 0.2 – 200 gigahertz. Infrared systems operate in the region just below visible light.

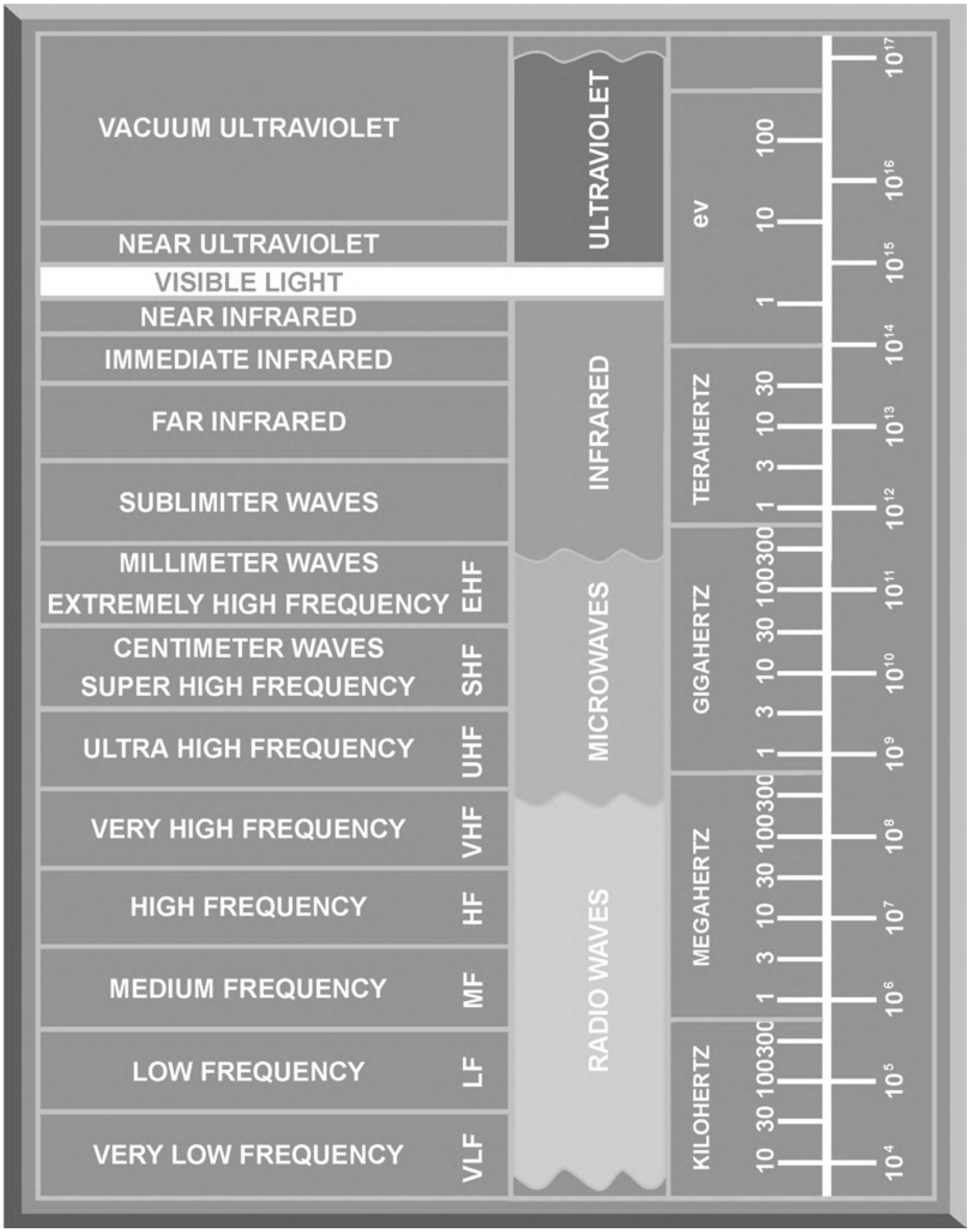
VACUUM ULTRAVIOLET
NEAR ULTRAVIOLET
VISIBLE LIGHT
NEAR INFRARED
IMMEDIATE INFRARED
FAR INFRARED
SUBLIMITER WAVES
MILLIMETER WAVES
EXTREMELY HIGH FREQUENCY
EHF
CENTIMETER WAVES
SUPER HIGH FREQUENCY
SHF
ULTRA HIGH FREQUENCY
UHF
VERY HIGH FREQUENCY
VHF
HIGH FREQUENCY
HF
MEDIUM FREQUENCY
MF
LOW FREQUENCY
LF
VERY LOW FREQUENCY
VLF
ULTRAVIOLET
INFRARED
MICROWAVES
RADIO WAVES
ev
1 10 100
TERAHERTZ
1 3 10 30
GIGAHERTZ
1 3 10 30 100 300
MEGAHERTZ
1 3 10 30 100 300
KILOHERTZ
10 30 100 300
10^{4}
10^{5}
10^{6}
10^{7}
10^{8}
10^{9}
10^{10}
10^{11}
10^{12}
10^{13}
10^{14}
10^{15}
10^{16}
10^{17}

Figure 2-9. Electromagnetic Spectrum

a. Table 2-2 lists prefixes commonly used to indicate scientific notation and their abbreviations. These prefixes are often used when dealing with extremely high frequency radar signals or RF waves with extremely small wavelengths. For example, when dealing with extremely high frequency RF, it is simpler to use gigahertz, or GHz. One gigahertz is one thousand megahertz (MHz). For example, a radar operating at 3150 MHz is also operating at an RF of 3.150 GHz.

Table 2-2. Commonly Used Prefixes for Scientific Notation

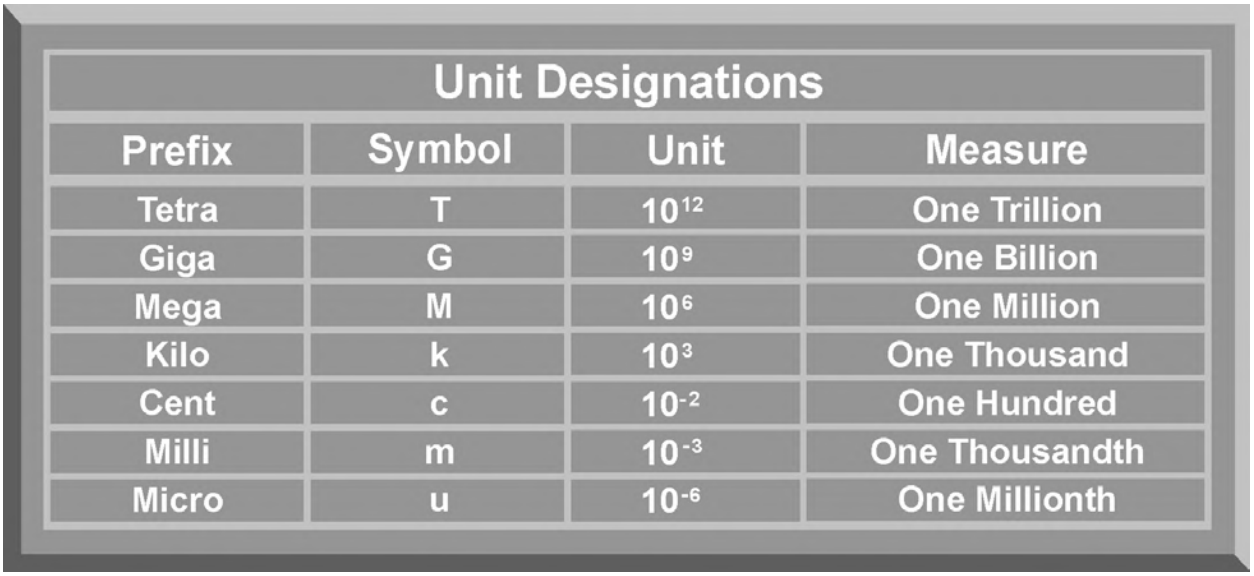

Unit Designations

Prefix	Symbol	Unit	Measure
Tetra	T	10^{12}	One Trillion
Giga	G	10^{9}	One Billion
Mega	M	10^{6}	One Million
Kilo	k	10^{3}	One Thousand
Cent	c	10^{-2}	One Hundred
Milli	m	10^{-3}	One Thousandth
Micro	u	10^{-6}	One Millionth

b. Frequency bands are often used when discussing radars and electronic combat systems. Radar designers use a frequency band designation system entirely different from the one used in electronic combat. Table 2-3 depicts both frequency band designation systems and the corresponding frequency ranges.

Table 2-3. Radar Frequency Band Designations

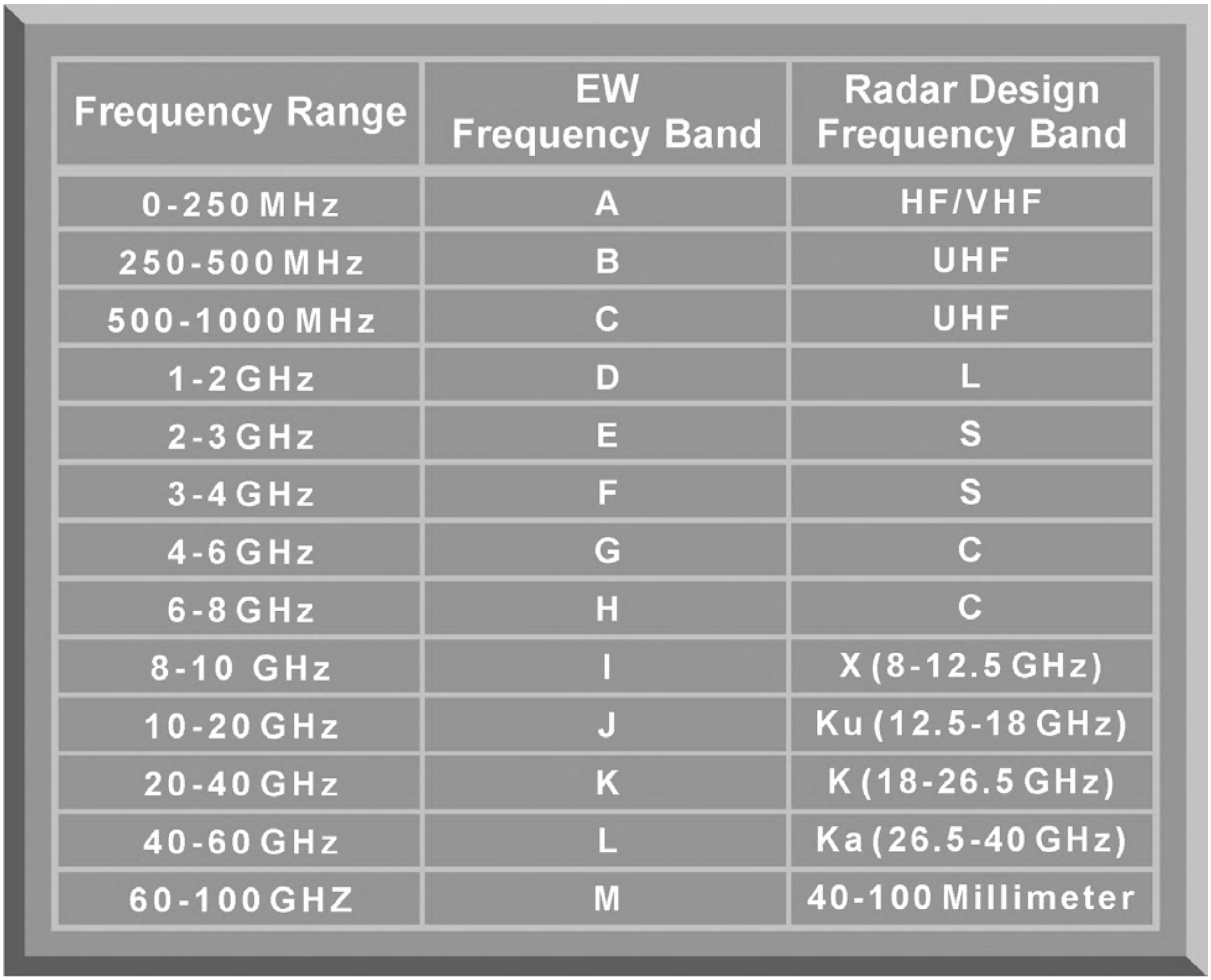

Frequency Range	EW Frequency Band	Radar Design Frequency Band
0-250 MHz	A	HF/VHF
250-500 MHz	B	UHF
500-1000 MHz	C	UHF
1-2 GHz	D	L
2-3 GHz	E	S
3-4 GHz	F	S
4-6 GHz	G	C
6-8 GHz	H	C
8-10 GHz	I	X (8-12.5 GHz)
10-20 GHz	J	Ku (12.5-18 GHz)
20-40 GHz	K	K (18-26.5 GHz)
40-60 GHz	L	Ka (26.5-40 GHz)
60-100 GHZ	M	40-100 Millimeter

7. RF PROPAGATION

Propagation characteristics of RF energy are profoundly affected by the earth's surface and atmospheric conditions. Any analysis of radar performance must take into account the propagation phenomena associated with RF radiation in a "real world" environment. The most important propagation phenomena include refraction, anomalous propagation (ducting), and attenuation.

a. In a vacuum, RF waves travel in a straight line. However, RF waves propagating within the earth's atmosphere do not travel in a straight line. The earth's atmosphere bends, or refracts, RF waves. One impact of the atmospheric refraction of RF waves is an increase in the line of sight (LOS) of the radar. This increase in radar LOS effectively extends the range of the radar system (Figure 2-10).

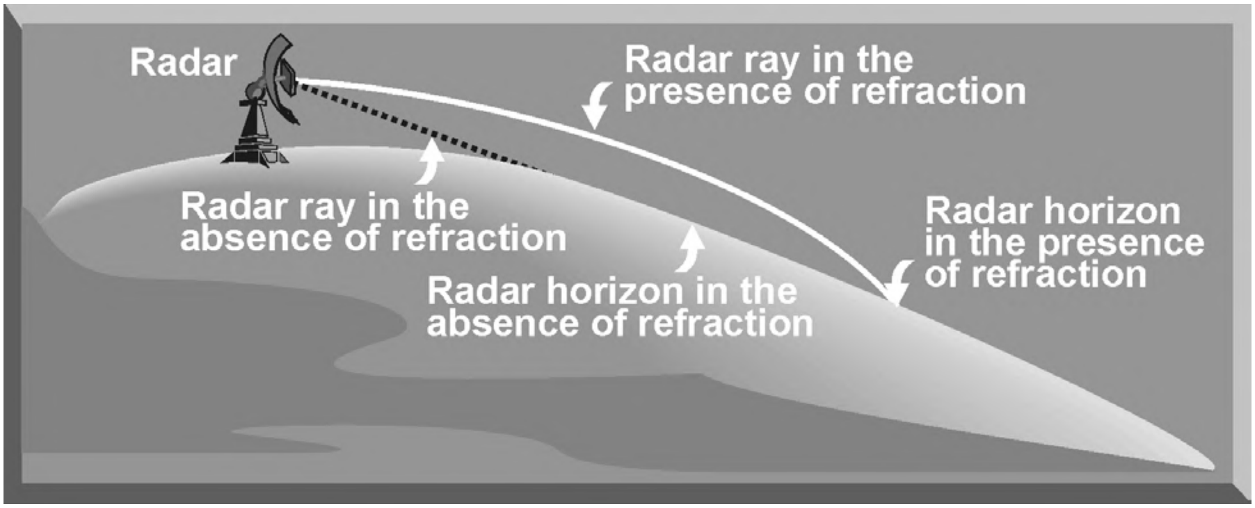
Radar
Radar ray in the presence of refraction
Radar ray in the absence of refraction
Radar horizon in the presence of refraction
Radar horizon in the absence of refraction

Figure 2-10. Impact of Refraction on RF Propagation

Atmospheric refraction of RF energy can also induce elevation measurement errors in radar systems (Figure 2-11).

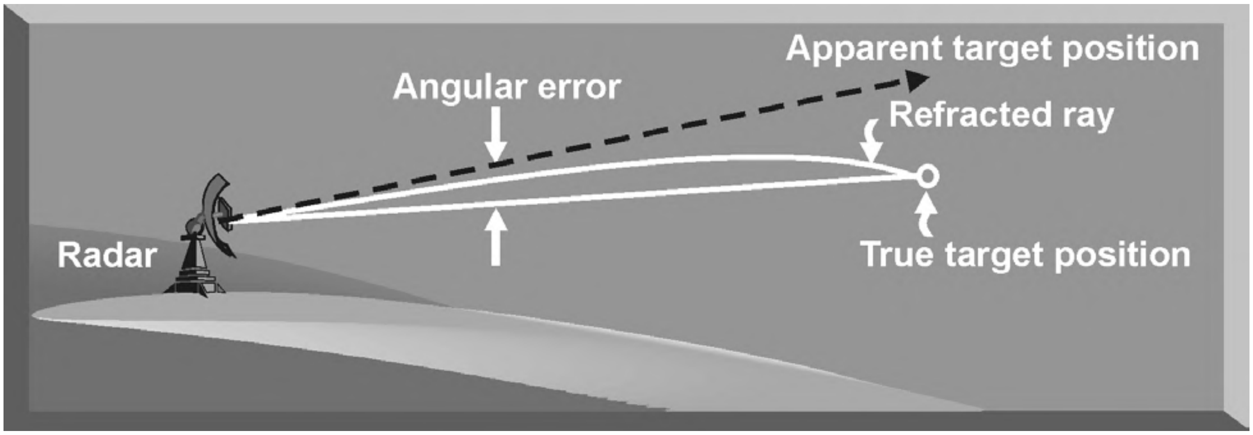
Apparent target position
Angular error
Refracted ray
Radar
True target position

Figure 2-11. Impact of Refraction on Target Elevation Determination

(1) The refraction of RF waves in the atmosphere is caused by the variation in the velocity of propagation with altitude. The index of refraction (n) is used to describe this velocity variation and is defined by Equation 2-2.

$$\text{Index of Refraction } (n) = \frac{\text{Velocity of Propagation in a Vacuum}}{\text{Velocity of Propagation in the Atmosphere}}$$

Equation 2-2. Index of Refraction

(2) The term refractivity (N) is used for predicting the impact of refraction on RF wave propagation. Refractivity is a “scaled up” expression for the index of refraction and is used by radar designers to calculate the impact of refraction on actual radar systems. At normal radar operating frequencies, the refractivity for air containing water vapor can be computed using Equation 2-3.

$$\text{Refractivity } (N) = (n-1) \times 10^{6} = \frac{77.6\,p}{T} + \frac{3.75 \times 10^{5} e}{T^{2}}$$

n = index of refraction

e = partial pressure of water vapor

p = barometric pressure

T = absolute temperature (degrees K)

Equation 2-3. Refractivity of RF Waves (Normal Radar Frequencies)

(3) As altitude increases, the barometric pressure and water vapor content decrease rapidly. At the same time, the absolute temperature decreases slowly based on the standard lapse rate. Using Equation 2-3, it can be seen that the refractivity of the atmosphere decreases with increasing altitude. This decrease in refractivity means that the velocity of RF waves increases with altitude. The result is a downward bending, or refraction, of RF waves as depicted in Figure 2-10. RF wave refraction primarily affects ground-based radar systems at low antenna elevation angles, especially at or near the horizon. For most radar applications, refraction is not a factor at elevation angles above 5 degrees.

b. The term anomalous, or nonstandard, propagation is used to describe atmospheric conditions that extend the propagation of RF waves and increase radar range. The most common anomalous propagation phenomena is called superrefraction, or ducting.

(1) A superrefracting duct is formed when the refractivity of the atmosphere (Equation 2-2) rapidly decreases with altitude. Based on Equation 2-2, this occurs when the temperature increases with altitude and/or the water vapor content decreases with altitude. An increase in temperature with altitude is called a temperature inversion. To produce a duct, the temperature inversion must be very pronounced.

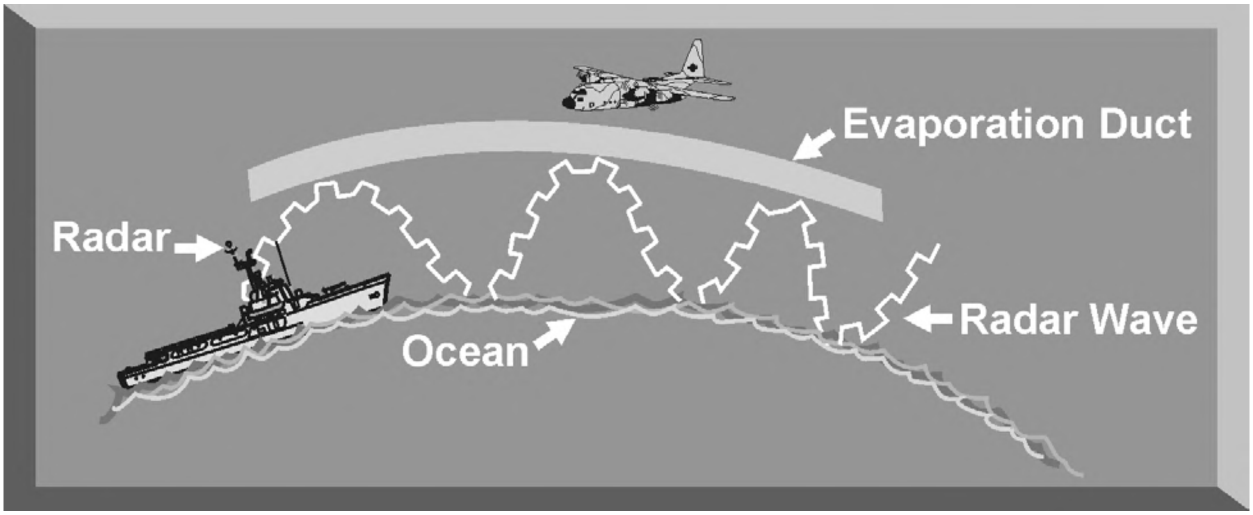

Figure 2-12. Superrefracting Surface Duct

(2) A superrefracting duct acts like a wave guide which traps the RF wave (Figure 2-12). This channels the radar signal and reduces attenuation. In order for an RF wave to propagate within a duct, the angle of the radar signal, in relation to the duct, should be less than one degree. The RF waves trapped by the duct take advantage of the decrease in refractivity and travel much further than normal. This can greatly extend the range of a radar system.

(3) The extension of radar range inside a duct can result in a reduction of radar coverage outside the duct. The area of reduced radar coverage because of ducting is called a radar hole. Due to radar holes, the extended radar range caused by ducting may result in a decrease in radar coverage along other paths of propagation. These holes can seriously degrade the effectiveness of early warning radar systems. For example, a radar system is taking advantage of a duct formed at the surface to extend low altitude radar range (Figure 2-12). Airborne targets flying just above the duct would normally be detected, but because of ducting, these targets may be missed.

(4) Water vapor content is a significant factor in producing ducts. Consequently, most ducts are formed over water and in warm climates. Any atmospheric phenomenon that results in a pronounced increase in temperature and/or a decrease in water vapor content as altitude increases can generate a superrefracting duct, of which there are three types. A superrefracting duct which is formed just above the surface of the earth is referred to as a surface duct. A surface duct formed just above the surface of the ocean is called an evaporation duct. A duct which is formed well above the surface of the earth is known as an elevated duct.

(a) Surface ducts formed over land are usually a result of the nighttime radiation of heat from the earth. Duct formation is especially prevalent during the

summer months when the ground is moist. As the earth loses heat, a temperature inversion is created at the surface coupled with a sharp decrease in the moisture content. These conditions are favorable to the formation of a surface duct. A superrefracting duct can also be produced by the diverging downdraft under a thunderstorm. The cool air that is dispersed creates a local temperature inversion while the water vapor content decreases due to rain. Surface ducts formed in conjunction with thunderstorms are difficult to predict and normally persist for a short period of time.

(b) A superrefracting surface duct that lies just above the surface of the ocean is a result of evaporated water, thus the term evaporation duct. The air in contact with the ocean is saturated with water vapor, while the air several feet above the ocean contains a much lower level. This rapid decrease in water vapor pressure with an increase in altitude creates an evaporation duct. An evaporation duct exists over the ocean almost all the time. The height of this duct varies from 20 to 100 feet based on the season, time of day, and wind speed. One positive aspect of an evaporation duct is the extended range available to a shipborne radar system with a properly aligned antenna. This extended range coverage against surface ships and low altitude aircraft is a definite advantage of ducted propagation.

(c) An elevated duct is generally formed by a temperature inversion in the upper atmosphere. To take maximum advantage of the increased radar range inside an elevated duct, both the radar and the target should be inside the elevated duct. In addition, radar systems operating below an elevated duct may also experience enhanced range performance.

(4) The presence of surface ducts and elevated ducts, especially over land, are extremely difficult to predict and may persist for very short periods of time. The atmospheric conditions favorable to duct formation are difficult to predict using conventional weather forecasting techniques.

c. The attenuation of RF energy in a clear atmosphere is due to the presence of oxygen and water vapor. Attenuation results when a portion of the RF energy strikes these molecules and is absorbed as heat. Figure 2-13 details the RF attenuation loss due to atmospheric gasses based on the frequency of the RF energy.

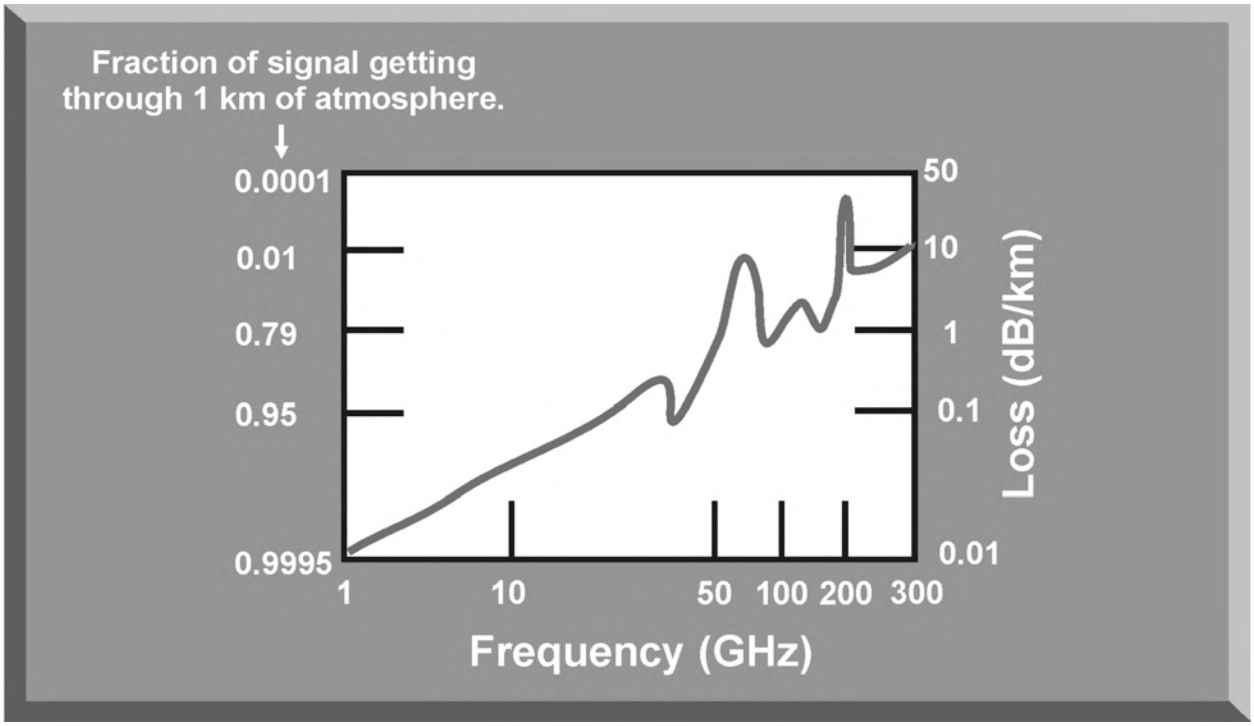

Figure 2-13. RF Atmospheric Attenuation

(1) At frequencies below 1 GHz, the effect of atmospheric attenuation is negligible. Above 10 GHz, atmospheric attenuation increases dramatically. This dramatic signal loss impacts the maximum detection range of radars operating in the millimeter wavelength band.

(2) RF energy attenuation decreases as altitude increases. The RF attenuation experienced by an air-to-air radar will depend on the altitude of the target as well as target range. For a ground-based radar, RF attenuation will decrease as antenna elevation increases.

8. SUMMARY

Since all radar operations depend on the transmission and reception of RF energy, a basic knowledge of RF frequency, wavelength, and polarization provides the basis for understanding the more complex radar characteristics. Since most modern radar systems employ some form of Doppler signal processing, the concept of the Doppler effect is fundamental to understanding modern radar operation. The concepts of refraction, anomalous propagation (ducting), and atmospheric attenuation are key to understanding how RF waves propagate in the atmosphere. The topics in this chapter provide a foundation for understanding radar and jamming system operation.

CHAPTER 3. RADAR SIGNAL CHARACTERISTICS

1. INTRODUCTION

Every radar produces a radio frequency (RF) signal with specific characteristics that differentiate it from all other signals and define its capabilities and limitations. Pulse width (pulse duration), pulse recurrence time (pulse repetition interval), pulse repetition frequency, and power are all radar signal characteristics determined by the radar transmitter. Listening time, rest time, and recovery time are radar receiver characteristics. An understanding of the terms used to describe these characteristics is critical to understanding radar operation.

2. PULSE WIDTH (PW)

Figure 3-1 depicts the output from a typical pulse radar. PW, sometimes called pulse duration (PD), is the time that the transmitter is sending out RF energy. PW is measured in microseconds. It has an impact on range resolution capability, that is, how accurately the radar can discriminate between two targets based on range.

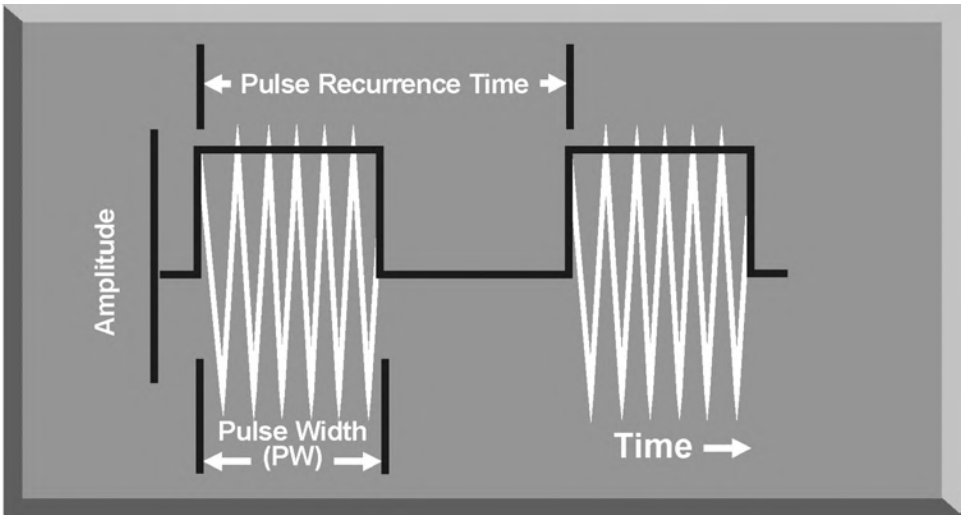

Figure 3-1. Typical Radar Pulse

3. PULSE RECURRENCE TIME (PRT)

Pulse recurrence time is also known as pulse repetition time. PRT is the time required for a complete transmission cycle. This is the time from the beginning of one pulse of RF energy to the beginning of the next. PRT is measured in microseconds. PRT is the same as pulse repetition interval (PRI), which is used in radar warning receivers and other electronic warfare support (ES) assets to discriminate between radar systems. It also affects maximum radar range.

4. PULSE REPETITION FREQUENCY (PRF)

One of the most important characteristics of a pulse radar signal is pulse repetition frequency. PRF is the rate at which pulses or pulse groups are transmitted. Generally, PRF is the number of pulses generated per second and is expressed in hertz (Hz). PRF and PRI are related in that PRI is the inverse of PRF. A word of caution—do not confuse the operating frequency of the radar, which is measured in Hz, with the pulse repetition frequency, which is also measured in Hz. They are entirely different characteristics of a pulsed radar signal.

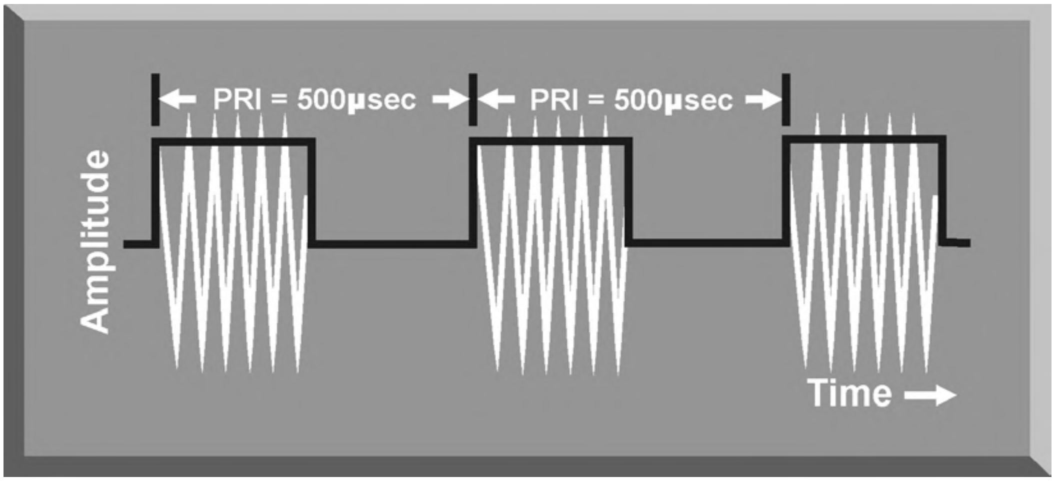

Figure 3-2. Constant PRF Radar Pulse

a. A pulse radar operating at an unvarying PRF is called a constant PRF radar (Figure 3-2). Pulse radar systems can employ PRF stagger or PRF jitter as an electronic protection (EP) technique against repeater or synchronous jammers. The time between each pulse is the PRI.

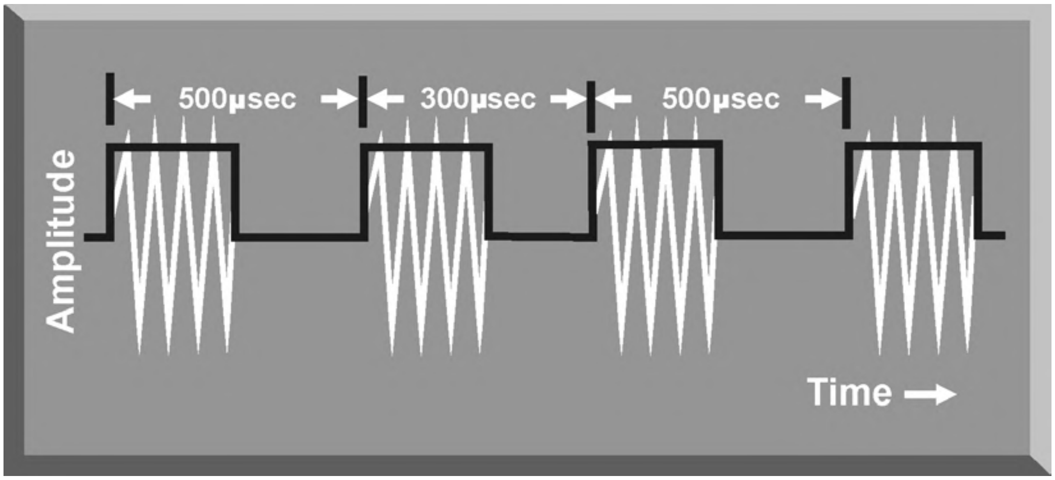

Figure 3-3. PRF Stagger

b. PRF stagger is accomplished by assuring that no adjacent PRIs are equal. The number of different PRIs generated is called the "position" of the stagger. Two-position stagger would have two PRI values, for example, 300 microseconds and 500 microseconds (Figure 3-3).

c. PRF jitter may be considered a random stagger. It is also an EP technique to counter synchronous jammers. PRF jitter has no repeating pattern of PRI values (Figure 3-4).

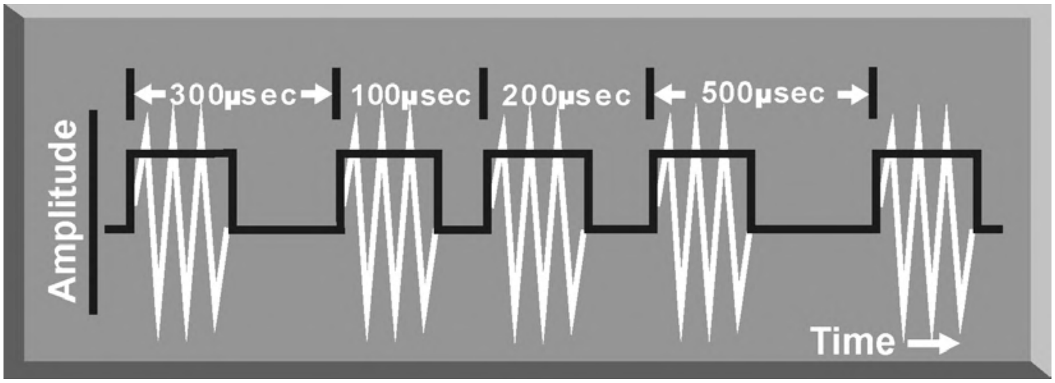
300µsec
100µsec
200µsec
500µsec
Amplitude
Time

Figure 3-4. PRF Jitter

5. RADAR RECEIVER CHARACTERISTICS

Pulse repetition frequency, pulse recurrence time, and pulse width are determined by the transmitter. The pulse radar signal characteristics that relate to receiver operation are rest time, recovery time (RT), and listening time (LT) (Figure 3-5).

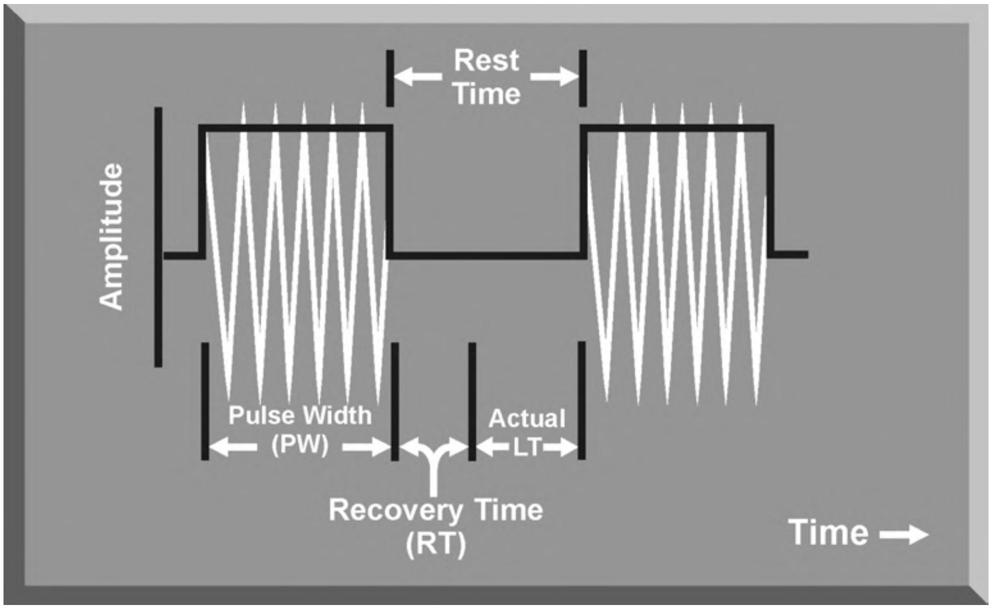
Rest Time
Amplitude
Pulse Width (PW)
Actual LT
Recovery Time (RT)
Time

Figure 3-5. Basic Radar Pulse

a. Rest time is the time between the end of one transmitted pulse and the beginning of the next. It represents the total time that the radar is not transmitting. Rest time is measured in microseconds.

b. Recovery time (RT) is the time immediately following transmission time during which the receiver is unable to process returning radar energy. RT is determined by the amount of isolation between the transmitter and receiver and the efficiency of the duplexer. A part of the high power transmitter output spills over into the receiver and saturates this system. The time required for the receiver to recover from this condition is RT.

c. Listening time (LT) is the time the receiver can process target returns. Listening time is measured from the end of the recovery time to the beginning of the next pulse, or PRT minus (PW + RT). Listening time is measured in microseconds.

6. DUTY CYCLE

Duty cycle is the ratio of the time the transmitter operates to the time it could operate during a given transmission cycle. The duty cycle of a radar can be computed by dividing the PW by the PRT, or by multiplying the PW times the PRF. Duty cycle has no units (Equation 3-1). CW radars have a duty cycle of 100%, while early warning radars may have a duty cycle of around 1%.

$$\text{Duty Cycle} = \frac{PW}{PRT} \text{ or Duty Cycle} = PW \times PRF$$

Equation 3-1. Duty Cycle

7. PEAK POWER

The power output of a radar is normally expressed in terms of peak power or average power (Figure 3-6). Peak power is the amplitude, or power, of an individual radar pulse. It is simply the power, measured in watts or megawatts, that is radiated when the transmitter is on. The power a radar transmits is normally used to determine the maximum detection range of that radar. However, it is the energy in a radar pulse that determines maximum radar detection range. Since power is the rate of flow of energy, the energy in a radar pulse is equal to the peak power multiplied by the time the radar is transmitting, or pulse width (Equation 3-2).

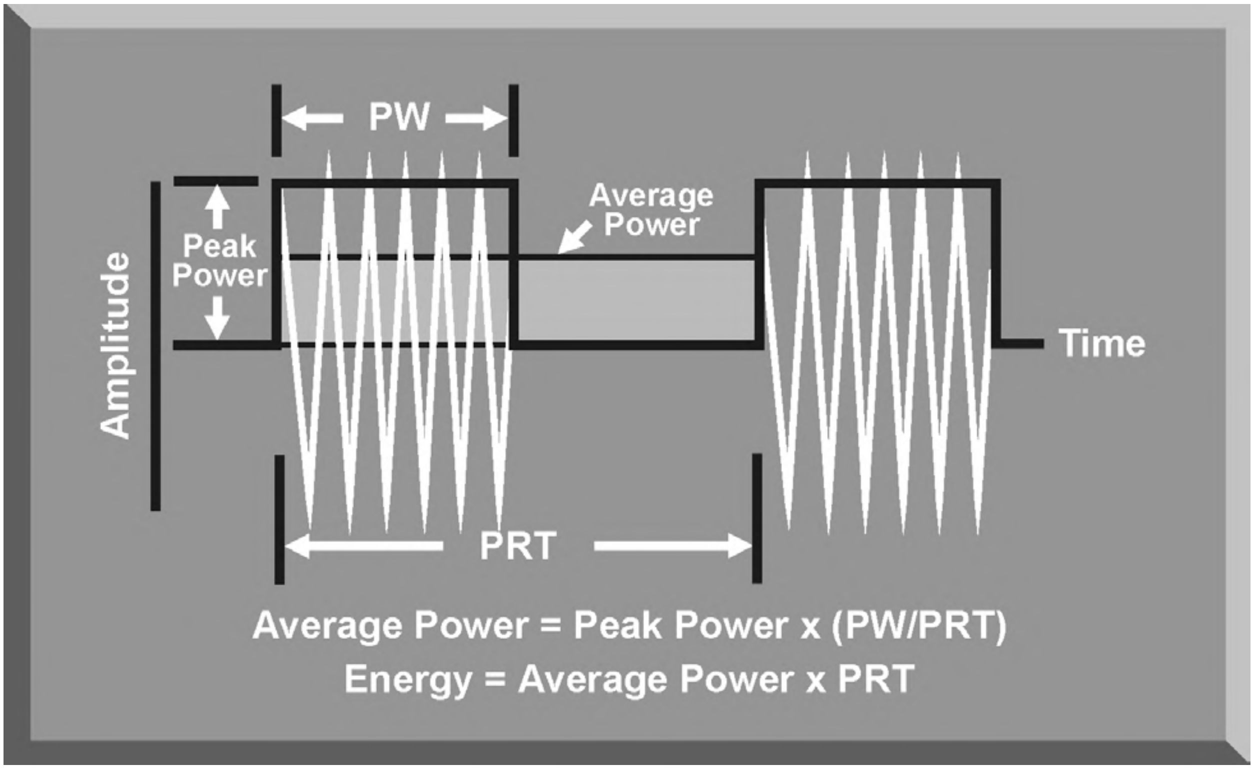
PW
Amplitude
Peak Power
Average Power
Time
PRT
Average Power = Peak Power x (PW/PRT)
Energy = Average Power x PRT

Figure 3-6. Peak Power and Average Power

$$\text{Energy Per Pulse} = \text{Peak Power} \times \text{Pulse Width}$$

Equation 3-2. Energy Per Pulse

8. AVERAGE POWER

Average power is the power distributed over the pulse recurrence time. It can be computed using the formula in Figure 3-6. The energy transmitted by average power can be computed by multiplying average power by PRT. Since the energy in a set of pulses determines detection range, average power or energy provides a better measure of the detection range of a radar than does peak power. Average power can be increased by increasing the PRF, by increasing the pulse width, or by increasing peak power (Figure 3-7).

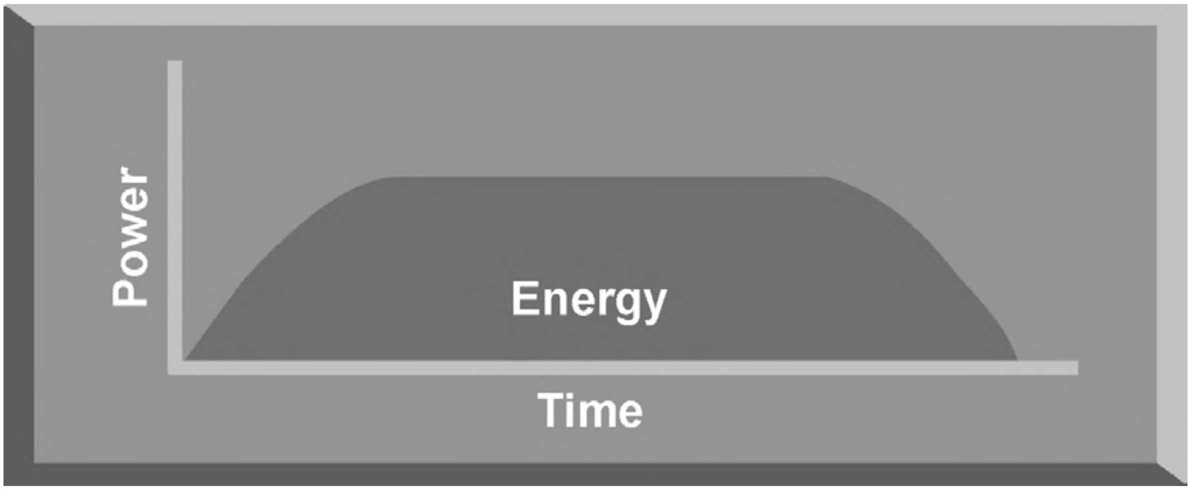

Figure 3-7. Radar Power and Energy

9. MODULATION

The characteristics of an RF signal must be changed in order to transmit information on the signal. This process is called modulation. Modulation is accomplished by combining a basic RF signal, called a carrier wave, with a modulating signal that contains the desired information. The resulting waveform is then used to transmit the desired information.

a. One basic modulation technique is amplitude modulation (AM). The carrier wave is combined with a modulating signal containing information of varying amplitude. Waveforms produced have the same frequency as the carrier wave but with a varying amplitude based on the information from the modulating signal. AM is used extensively in communications and broadcast radio transmissions (Figure 3-8).

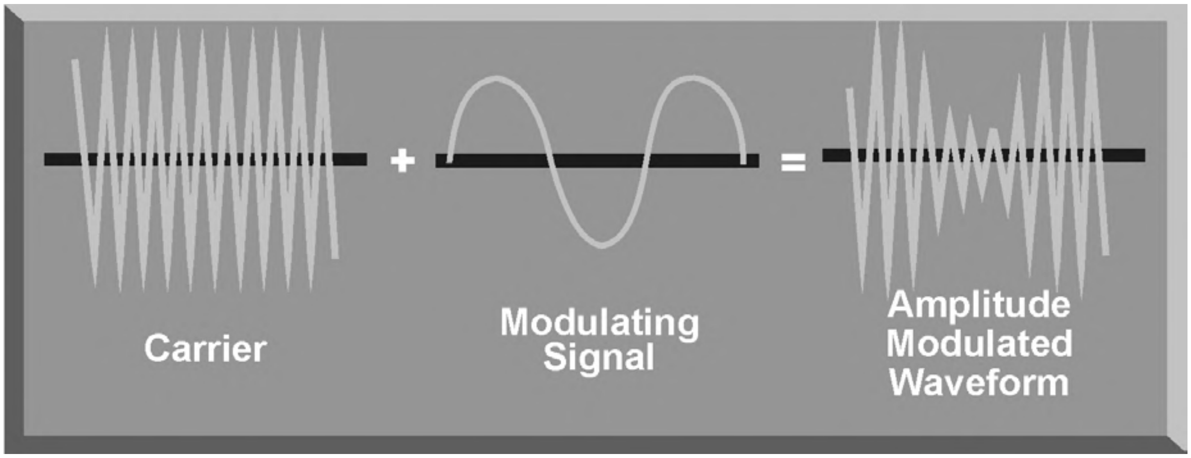

Figure 3-8. Amplitude Modulation

b. Frequency modulation (FM) is another means of impressing information on a carrier wave. Frequency modulation is accomplished by combining the carrier wave with a modulating signal containing information of varying frequency. The waveform produced has the same amplitude as the carrier wave, but the frequency varies based on the information from the modulating signal (Figure 3-9). FM is used extensively in communications and commercial radio. FM is also used with continuous wave (CW) radars to make them more resistant to jamming and to add range determination capability.

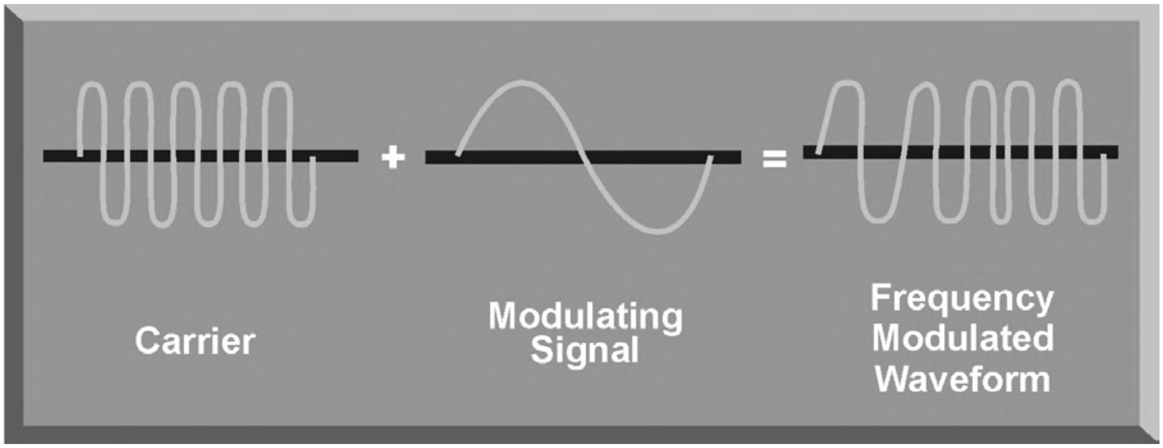

Figure 3-9. Frequency Modulation

c. A type of amplitude modulation known as pulse modulation (PM) is used in pulse radars to produce the short, powerful bursts of RF energy. PM combines the carrier wave with a rectangular pulse that acts like a switch. PM turns the transmitter on, leaves it on for a predetermined time, and then turns it off. The result is a waveform that produces radar pulses that can be used to measure range and angle to the target (Figure 3-10).

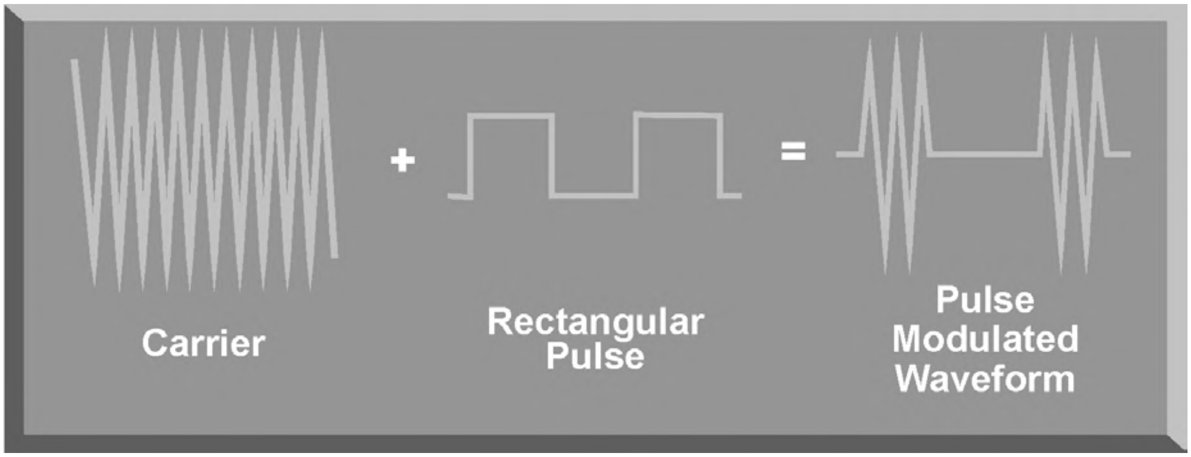

Figure 3-10. Pulse Modulation

10. SUMMARY

The radar signal characteristics of PW, PRI, PRF, and power determine the maximum range and the range resolution capability of a specific radar. When combined with the frequency of the carrier wave of the radar signal, these parameters provide a unique signature to identify a specific radar signal. Modulation is the method used to put information on an RF carrier wave. The primary modulation techniques used in radar signal generation include amplitude, frequency, and pulse modulation. The radar signal characteristics of PRF, PRI, power, and modulation are the keys to understanding radar operation and jamming techniques.

CHAPTER 4. RADAR SYSTEM COMPONENTS

1. INTRODUCTION

The individual components of a radar determine the capabilities and limitations of a particular radar system. The characteristics of these components also determine the countermeasures that will be effective against a specific radar system. This chapter will discuss the components of a basic pulse radar, a continuous wave (CW) radar, a pulse Doppler radar, and a monopulse radar.

2. PULSE RADAR SYSTEM

The most common type of radar design is the pulse radar system. The name describes a process of transmitting discrete bursts of RF energy at the frequency of the radar system. The time that pulses are transmitted determines the pulse repetition frequency (PRF) of the radar system. A pulse radar system can figure out range and azimuth. Range is determined by the time that it takes a pulse to go to a target and return. Target azimuth is determined by the relative position, or antenna orientation, when the pulse strikes the target. Figure 4-1 is a basic block diagram of a simple pulse radar system.

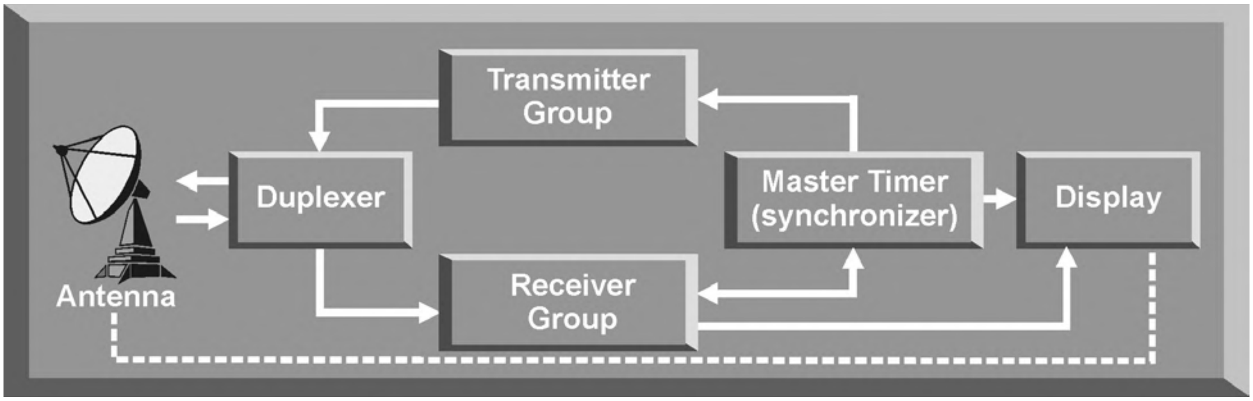

Figure 4-1. Pulse Radar Block Diagram

a. The purpose of the transmitter is to deliver a series of high-energy bursts of radio frequency (RF) energy to the antenna. The transmitter group of a modern pulse radar normally consists of a pulse generator or waveform generator, a modulator, and some kind of power amplifier (Figure 4-2).

(1) The purpose of the waveform generator is to generate the proper waveform or pulse, normally at a low power level, before delivery to the modulator. It is much easier to generate complex waveforms at a lower power level. These complex waveforms are required for coherent systems employing digital moving target indicator techniques and for pulse Doppler radar operations.

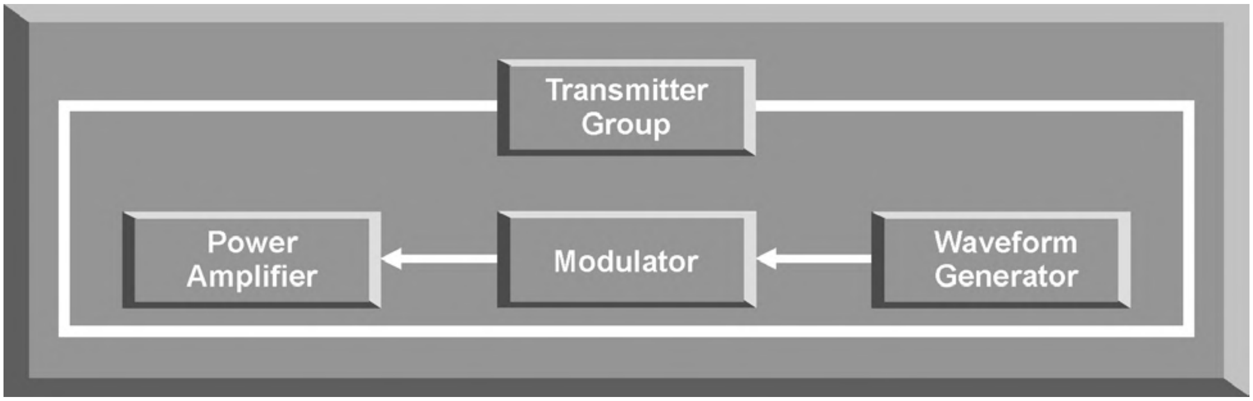

Figure 4-2. Transmitter Group

(2) The modulator is a major portion of the transmitter. The modulator provides an extremely powerful, very short pulse of direct current (DC) voltage to the power amplifier. This is similar to the ignition system of an automobile but with very stringent requirements. The modulator has an energy storage device and a switch. Between pulses, during the resting time of the transmitter, energy is accumulated and stored in the storage device. When keyed by the master timer, all this energy is switched to the power amplifier as a pulse. The waveform of this pulse is determined by the waveform generator.

(3) The power amplifier for a modern radar is normally a klystron, traveling wave tube, cross field amplifier, or solid state amplifier. Most common pulse radars use a klystron power amplifier. No matter what power amplifier is used, the purpose of the transmitter group is to produce a series of pulses at the correct amplitude, at the proper interval, with the exact waveform, and at the operating frequency of the radar.

b. A duplexer is required when both the transmitter and receiver use the same antenna. The duplexer acts as a rapid switch to protect the sensitive receiver from damage when the high-power transmitter is on. When the transmitter is off, the duplexer directs the weak target signals to the receiver. The duplexer's main purpose is to minimize power loss and maximize isolation. Power lost in the duplexer during transmission reduces the maximum detection range of the radar. Isolation refers to the amount of transmitter power that "bleeds through" the duplexer to the receiver during transmission. This "bleed through" must be extremely small to avoid receiver saturation or damage.

c. The capabilities of the receiver group are critical to radar performance (Figure 4-3). The ability of the radar receiver to detect the presence of the target return and extract the required information is limited primarily by noise. Noise can enter the receiver through the antenna along with the target return. This type of noise is called external noise. Noise generated within the receiver is called thermal noise. Radar noise can never be completely eliminated. Minimizing noise

is the most important consideration in the design of the sensitive receivers used with modern radars. In addition, relative immunity to noise makes a radar system more resistant to jamming.

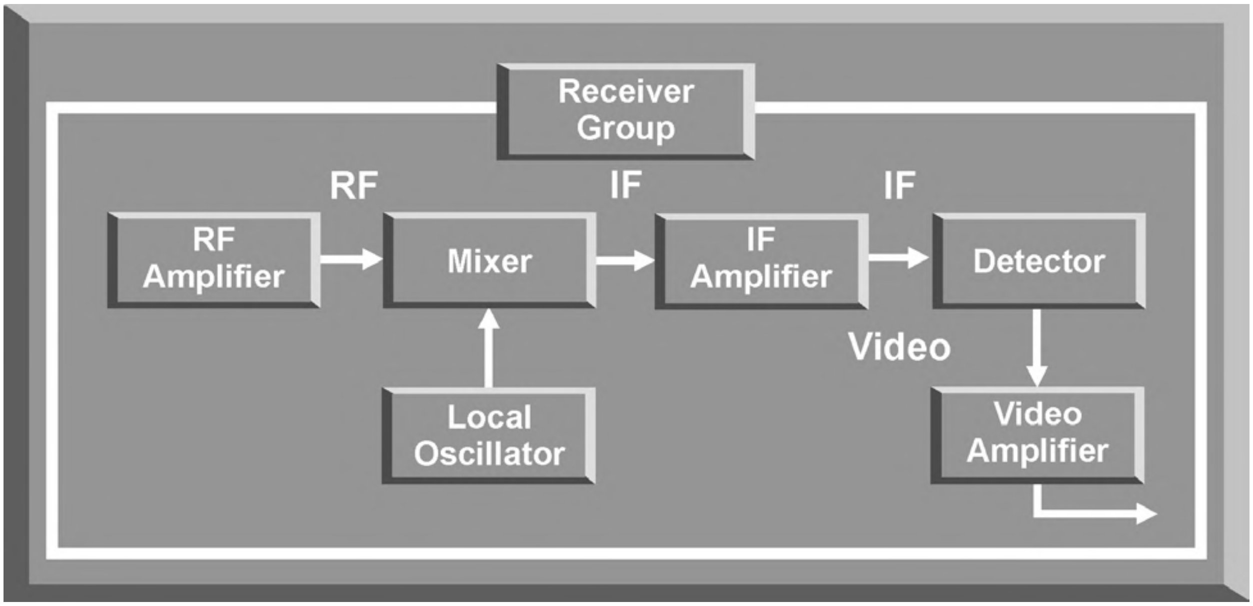

Figure 4-3. Receiver Group

(1) The most common pulse radar receiver is the superheterodyne receiver. A superheterodyne receiver consists of an RF amplifier, a mixer and local intermediate frequency (IF) amplifier, a detector, and a video amplifier.

(2) Radar target returns enter the receiver group via the antenna and duplexer. Since these signals are normally very low power, the RF amplifier boosts the signal gain and filters out as much external noise as possible. The capability of the RF amplifier to minimize noise determines the receiver sensitivity. The boosted RF signal is sent to the mixer where it is converted to a lower IF. This is accomplished by mixing the RF signal with the signal from the local oscillator to produce an IF that is easier to process. The IF amplifier increases the IF signal level and includes a matched filter. The matched filter maximizes the signal-to-noise ratio which enhances detection of the target return. The detector, which is usually a crystal diode, extracts the video modulation from the IF or converts the IF to a video signal.

d. The brain of a basic pulse radar is the master timer, or synchronizer, which coordinates the operation of the various parts of the radar (Figure 4-4). Exact timing within the radar is necessary to get accurate range. The master timer is an oscillator that triggers the transmitter to initiate transmission of a pulse. Simultaneously, the master timer sends a signal to initialize the display to ensure that range and azimuth information is accurately displayed.

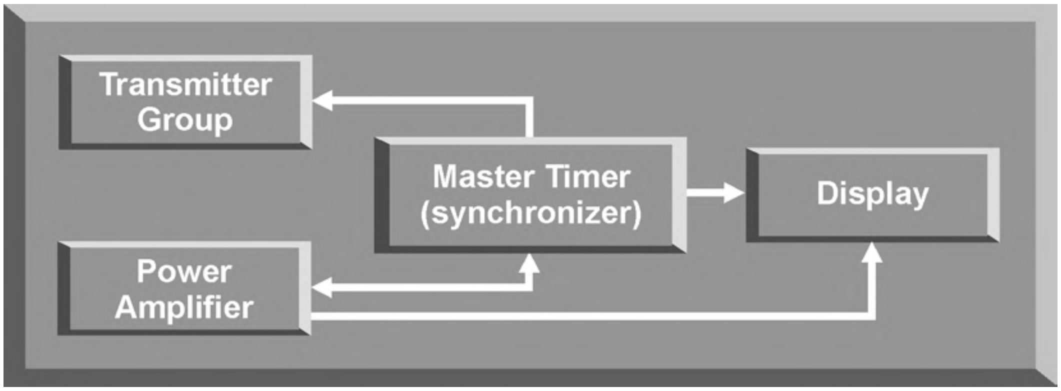

Figure 4-4. Master Timer/Synchronizer

e. The function of the antenna during transmission is to concentrate the radar energy from the transmitter into a shaped beam that points in the desired direction (Figure 4-5). During reception, or listening time, the function of the antenna is to collect the returning radar energy contained in the echo signals, and deliver these signals to the receiver. Radar antennas are characterized by directive beams that are usually scanned in a recognizable pattern.

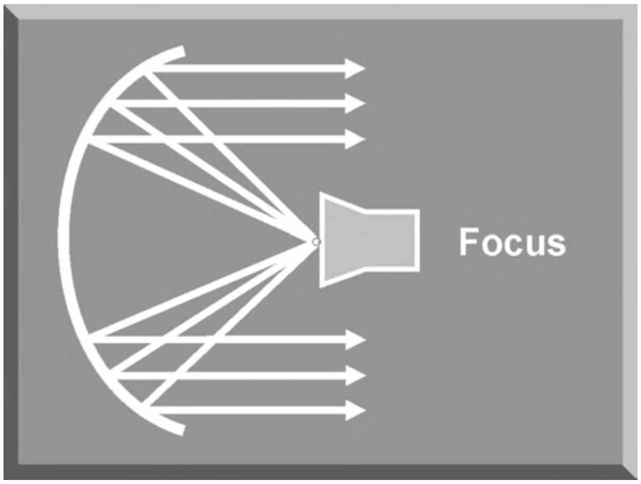

Figure 4-5. Parabolic Antenna

f. The purpose of the radar display is to take the information derived from a radar target in the receiver group and present it to the operator in a usable format. There are many different types of scope displays depending on the purpose of the radar and how the radar information is to be used. There are four basic types of radar displays: the A scope, B scope, range height indicator (RHI) scope, and plan position indicator (PPI) scope.

(1) The A scope is used to display target range or velocity (Figure 4-6). Threat systems using A scope displays include air interceptors (AIs) with range-only radar, surface-to-air missiles (SAMs), and radar-directed antiaircraft artillery (AAA) systems. SAM and AAA systems may use the A scope for range or velocity information, and other radar displays for azimuth and elevation data. The A scope displays range or velocity in relation to amplitude. The operator must distinguish the target return from other returns, including ground return and noise.

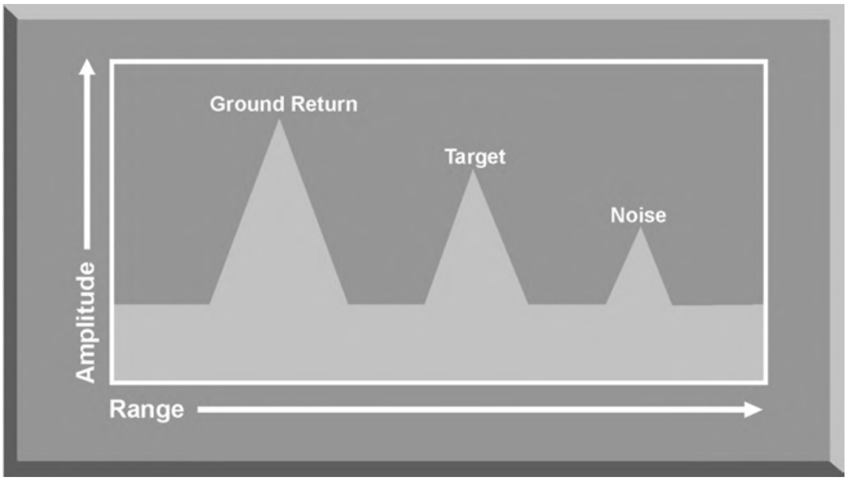

Figure 4-6. A Scope Display

(2) The B scope is used to display target range and azimuth (Figure 4-7). Threat systems using B scope displays include AI and SAM systems. The position of the target return to the right or left of the centerline of the screen shows the azimuth of the target. The position of the target return in relation to the bottom of the display, or zero range, shows target range.

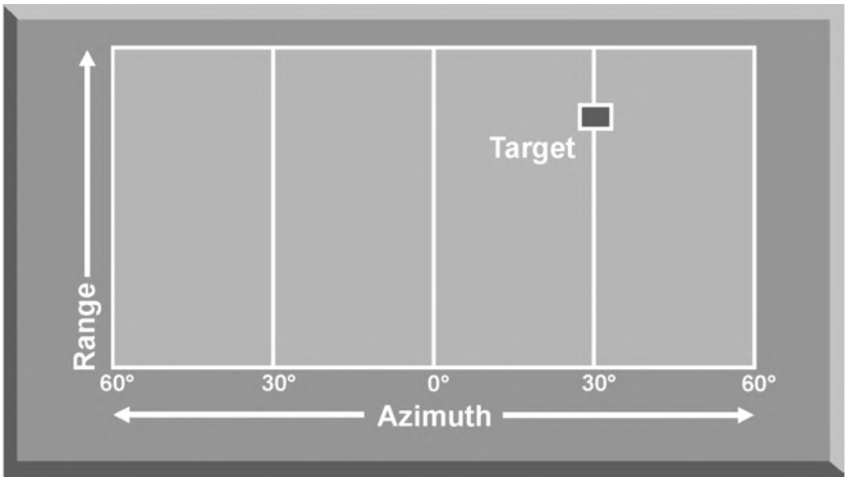

Figure 4-7. B Scope Display

(3) The RHI scope is used to display range and elevation (Figure 4-8). The RHI scope is used with height finder radars, and a modified RHI scope is used for ground-controlled approach (GCA) radars. The sweep trace of the display produces a fan-shaped display with the vertex at the lower left of the scope. The antenna sweeps up and down and is synchronized with the display.

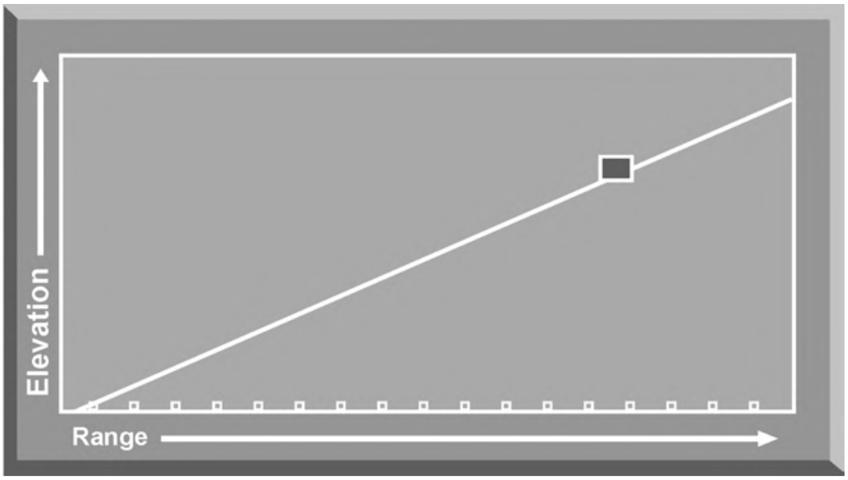

Figure 4-8. RHI Scope Display

(4) The PPI display is probably the best known radar display (Figure 4-9). The display represents a map picture of the area scanned by the radar beam, usually 360 degrees. The PPI display is used by early warning, acquisition, ground-controlled intercept (GCI), and SAM radar systems. The target return's angular position shows target azimuth, while distance from the center of the display shows range.

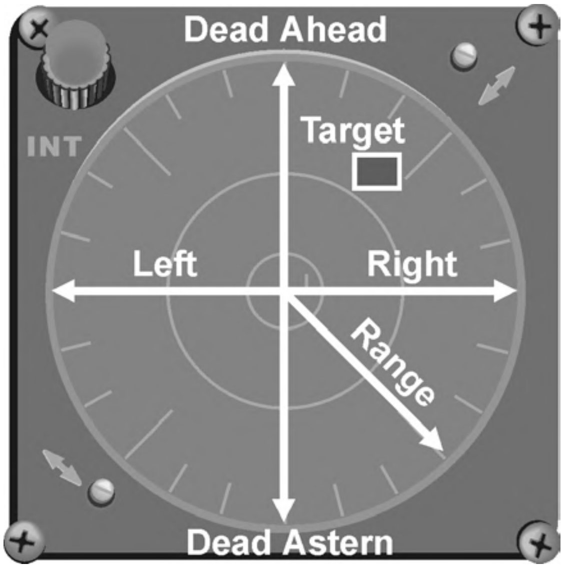

Figure 4-9. PPI Scope Display

3. CONTINUOUS WAVE (CW) RADAR

A continuous radar transmission from the antenna requires that classic CW radars have two antennas, one for transmission and one for reception (Figure 4-10). Since a continuous transmission results in a continuous echo signal, it is impossible to tell what part of the echo is associated with any particular part of the transmission. This makes conventional range determination (based on timing) impossible. However, the simple application of the Doppler principle provides a means for a CW radar to track a target. The Doppler principle deals with the fact that a radar return from a moving target will be shifted in frequency by an amount proportional to its radial velocity relative to the radar site. Using the difference in frequency from the transmitted signal to the received signal, a CW radar can separate the target return from clutter, based on a change in frequency. This type of radar is called a CW Doppler radar.

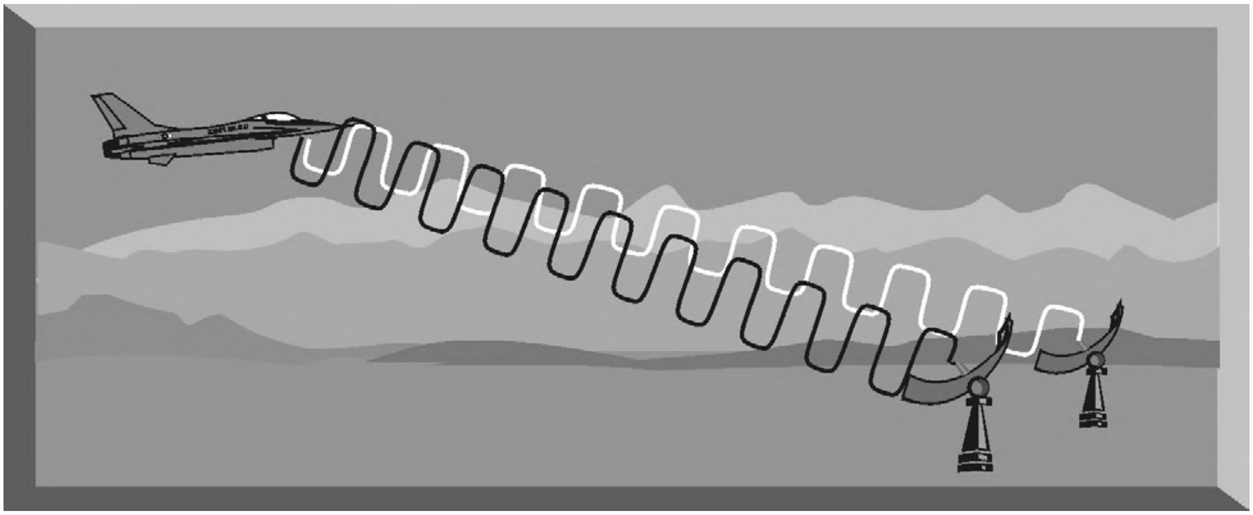

Figure 4-10. Continuous Wave Radar

a. Figure 4-11 depicts a basic CW Doppler radar. In a simple CW Doppler radar, the transmitter transmits a continuous signal at the radar's operating frequency. This signal is reflected by a moving target and travels back to the receiving antenna. The frequency of the reflected signal (f_d) is the frequency change due to the Doppler effect. This target frequency is passed to the detector. The transmitted frequency (f_o) is also fed to the detector as a reference. The detector notes the difference between the transmitted and received frequencies and passes this frequency to the Doppler filters. The Doppler filters only allow Doppler frequencies within a certain range to pass through. A filter is required for each Doppler frequency. The number of Doppler filters determines the number of targets that the radar can resolve in velocity.

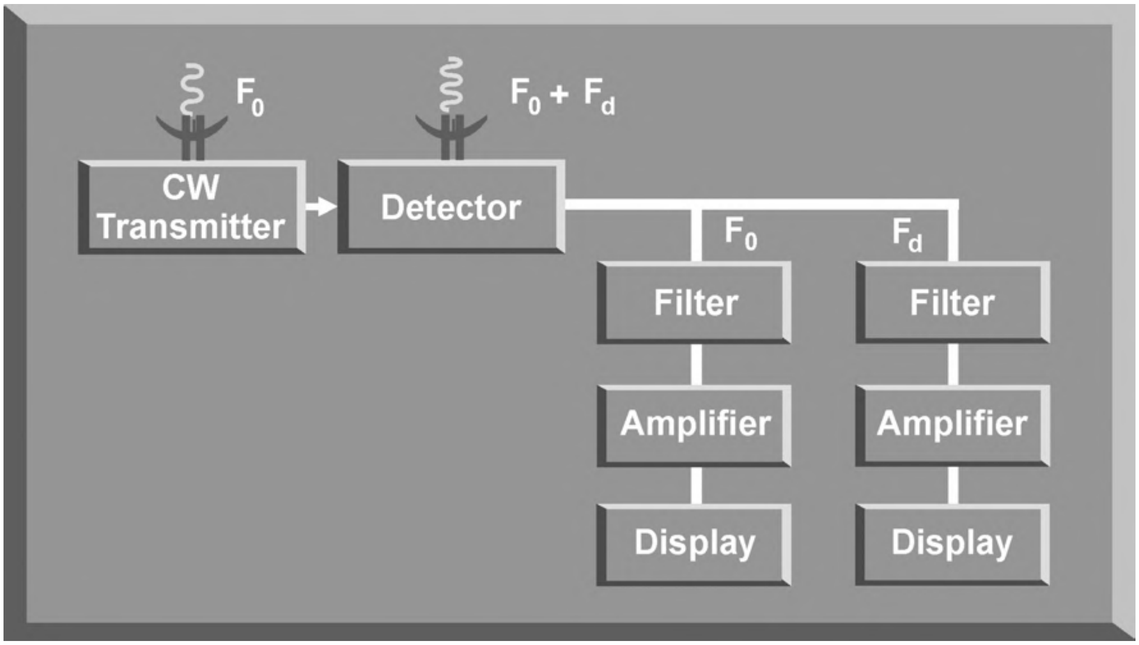
F_0
$F_0 + F_d$
CW Transmitter
Detector
F_0
F_d
Filter
Filter
Amplifier
Amplifier
Display
Display

Figure 4-11. Basic CW Doppler Radar

b. The output of each Doppler filter is amplified and passed to its own display. The display is normally an A scope as shown in Figure 4-12.

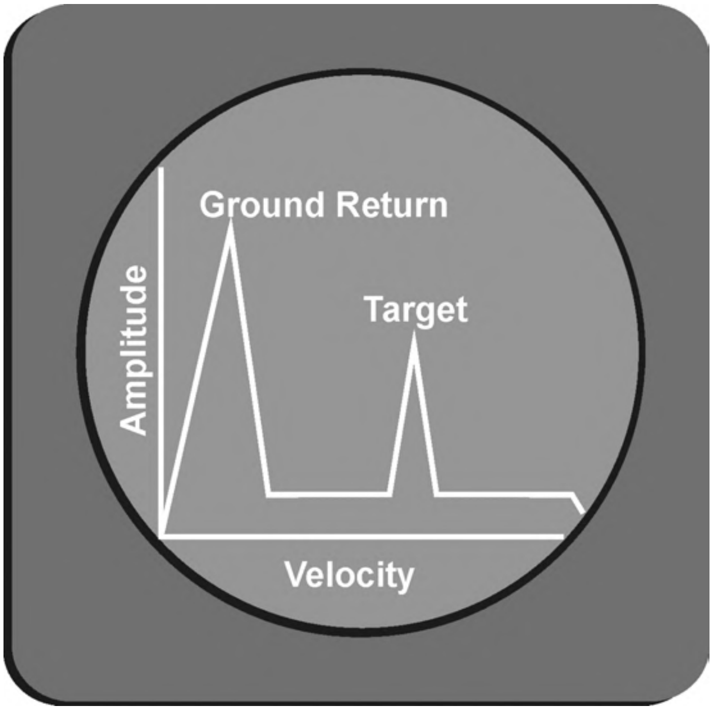
Ground Return
Target
Amplitude
Velocity

Figure 4-12. CW Doppler Radar Display

4. PULSE DOPPLER RADAR

Pulse Doppler radars combine the advantages of both pulse and Doppler radar systems (Figure 4-13). Because the signal is pulsed, the radar can determine range, azimuth, and elevation, similar to a conventional pulsed radar. A pulse Doppler radar can also compute overtake, or rate of closure, relative to the radar system on a pulse-to-pulse basis. Pulse Doppler radars also use multiple PRFs to eliminate target eclipsing and for range determination in medium PRF. The beauty of a pulse Doppler radar is that it eliminates ground clutter and provides range, azimuth, and velocity resolution.

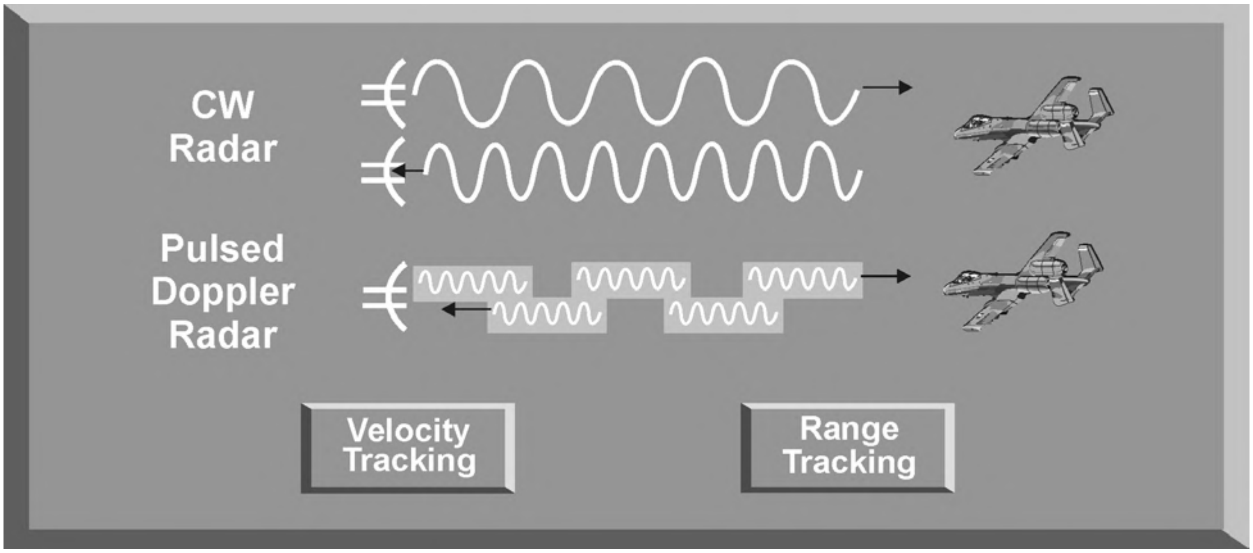

Figure 4-13. CW and Pulse Doppler Radar Comparison

a. A pulse Doppler radar transmits a box, or pulse, of coherent RF energy at the operating frequency of the radar. The frequency inside these boxes reacts the same way as the continuous waves of a CW radar. However, since the RF waves are pulsed, range determination can be accomplished by measuring the time it takes for the reflected pulse to return from the target. Velocity determination and tracking are accomplished by capturing and quantifying the Doppler shift of the frequencies in each box or pulse.

b. The basic block diagram of a coherent pulse Doppler radar (Figure 4-14) is similar to a pulse radar except for the addition of an exciter, a radar computer, and a digital signal processor. The exciter generates a continuous stable low power signal at the desired frequency and phase for the transmitter. It also sends this signal as a reference to the receiver. The digital signal processor performs the adding and subtracting functions required to find, track, and sort targets with respect to velocity and range. The radar computer performs all routine functions of the radar such as changing modes and accounting for aircraft flight parameters.

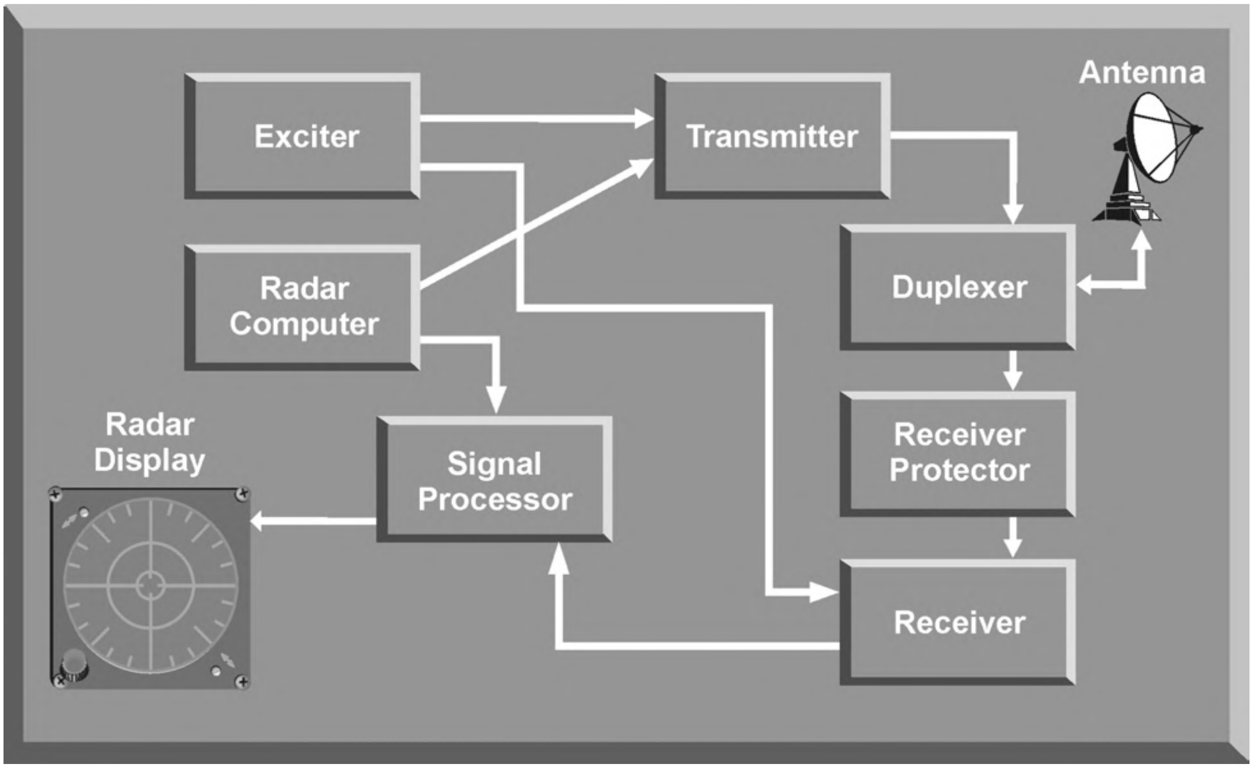

Figure 4-14. Basic Pulse Doppler Radar Diagram

5. MONOPULSE RADAR

The primary functions of monopulse radars include target tracking and weapon guidance. Monopulse radars were developed to overcome the limitations and jamming susceptibility of scanning radar systems. A monopulse radar typically receives returned radar energy in two, three, or four separate receivers, or channels, each looking at a different area. By comparing the returns from each receiver, track errors can be determined. For example, in Figure 4-15, this radar compares the amplitude, or signal strength, in each channel. Obviously, an aircraft in area "A" will produce a stronger return in receiver "A" since that receiver is focused in that direction. Since this type of system uses signal strength to determine position, typical noise jamming will only highlight an aircraft" position. Other coherent radars can measure phase differences verses the amplitude differences used in the last example. If a signal is directly ahead of the antenna, each quadrant will receive the returned signal at exactly the same time and therefore the same exact phase. However, if a return is off-center, the returned signal will strike part of the antenna slightly ahead of the opposite side and each quadrant will measure a different phase. This phase difference is used to determine azimuth error.

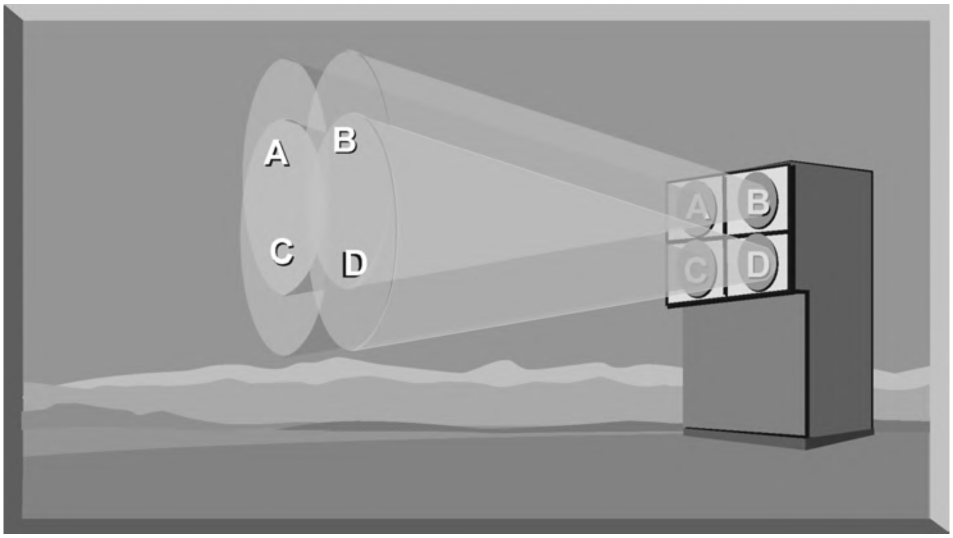

Figure 4-15. Monopulse Radar

a. Monopulse radars operate by comparing the amplitude or the phase of the received signal in each of the transmitted beams. A complex set of comparator circuits, called magic T's, do the addition and subtraction of the received signals based on the orientation of the magnetic field, or H-plane, and the electrostatic field, or E-plane (Figure 4-16). The received signal from antenna A enters the magic T in the H-plane arm while the signal from antenna B enters the magic T in the E-plane arm.

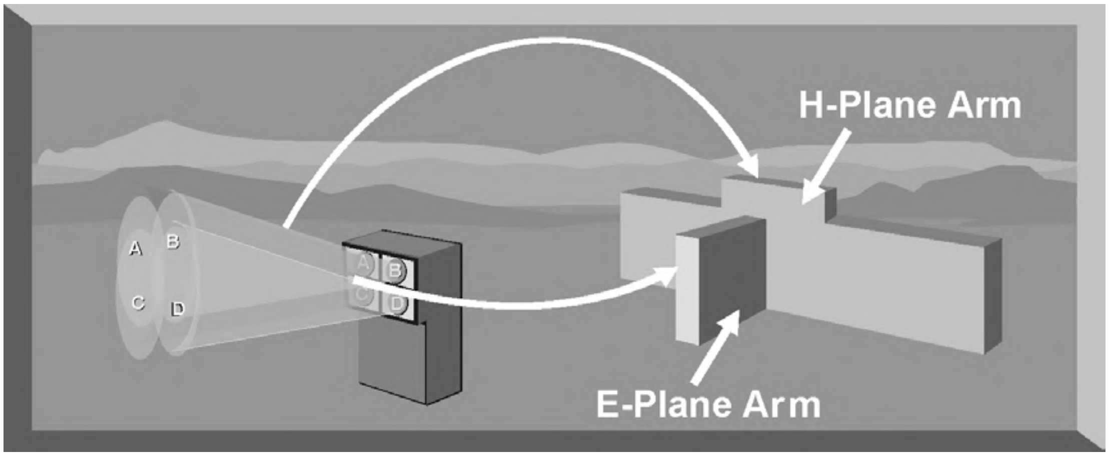

Figure 4-16. Monopulse Magic T’s

b. Figure 4-17 depicts the output of a magic T. Output is the sum and difference of the two input signals. These sum and difference values, in amplitude or phase, are used to generate azimuth and elevation error signals and to compute range.

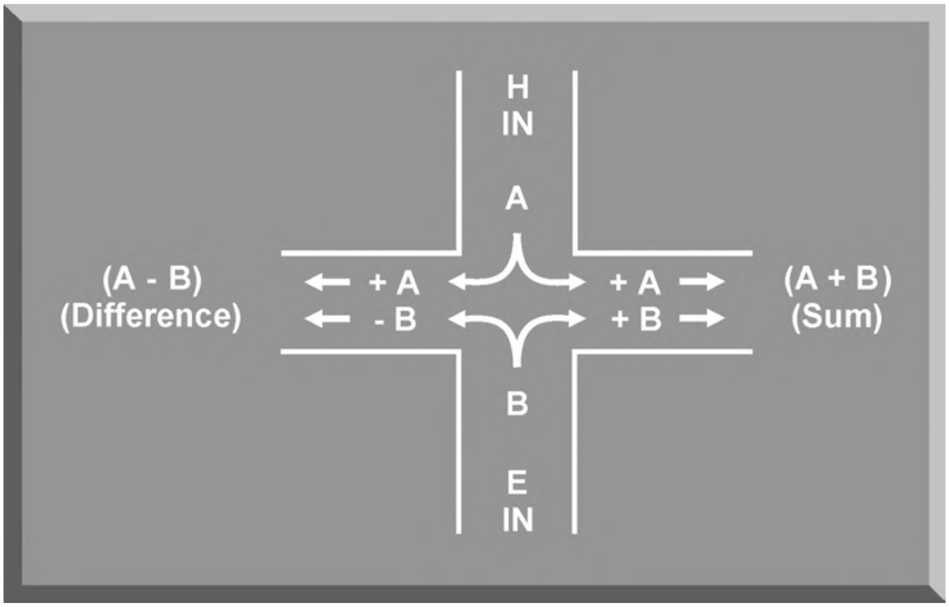
H
IN
A
(A - B)
(Difference)
+ A
- B
+ A
+ B
(A + B)
(Sum)
B
E
IN

Figure 4-17. Magic T Output Signals

c. These sum and differences computed by the magic T's are sent as signals to the monopulse tracking servos to generate both azimuth and elevation corrections to keep the monopulse antenna centered on the target (Figure 4-18). Target range is computed by calculating the difference of signal transmission time with the reception time of the target echo. Range information is displayed to the operator for range tracking.

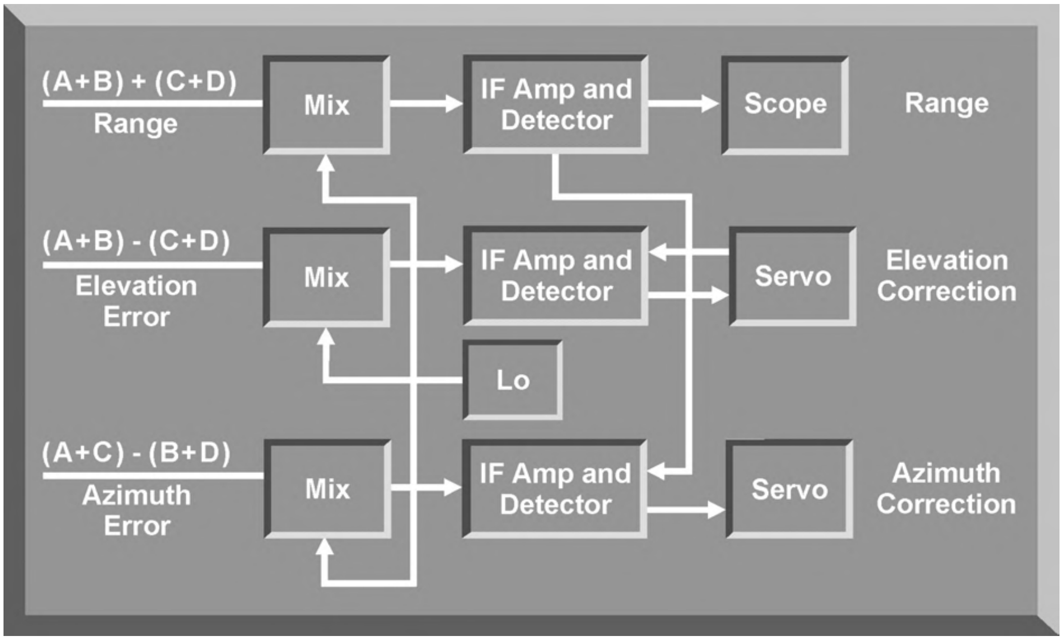
(A+B) + (C+D)
Range
Mix
IF Amp and Detector
Scope
Range
(A+B) - (C+D)
Elevation Error
Mix
IF Amp and Detector
Servo
Elevation Correction
Lo
(A+C) - (B+D)
Azimuth Error
Mix
IF Amp and Detector
Servo
Azimuth Correction

Figure 4-18. Monopulse Radar Track

d. If the signal in antenna A equals the signal in antenna B, equals the signal in antenna C, and equals the signal in antenna D, the target is in the center of the four radar beams (Figure 4-19). This is the condition the monopulse tracker attempts to maintain when tracking a target.

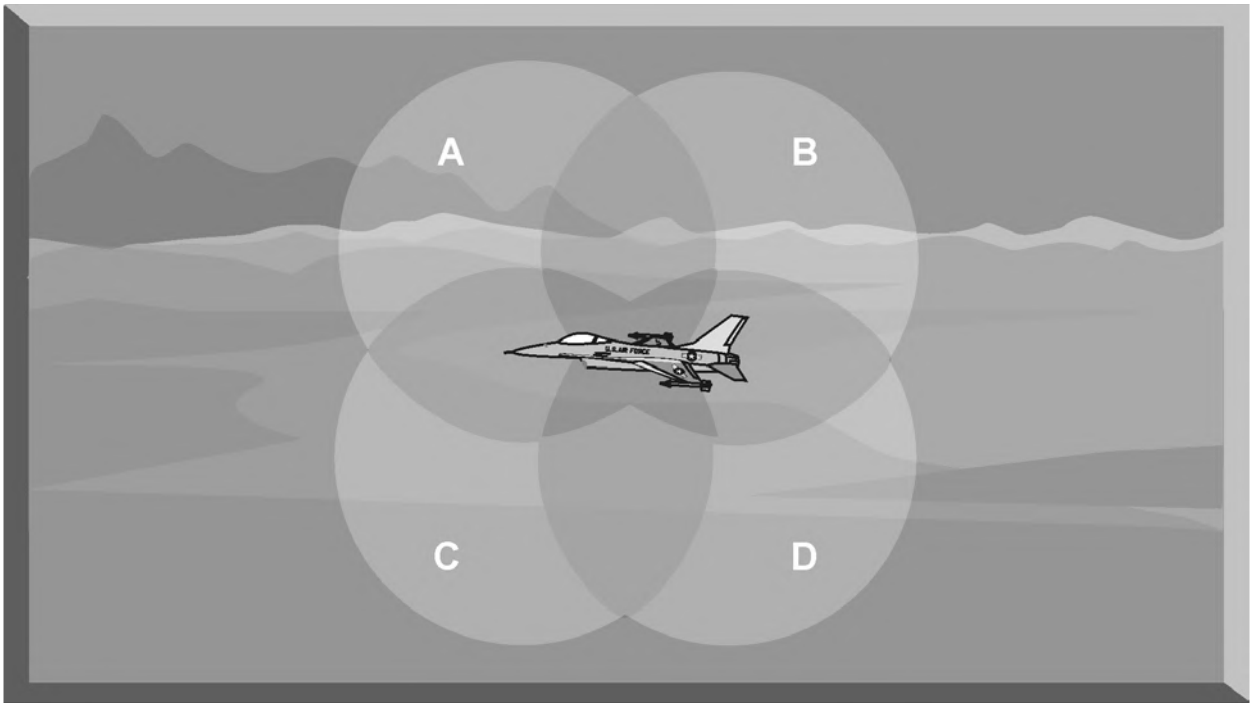

Figure 4-19. Monopulse Radar Track Logic

6. SUMMARY

This chapter has discussed the basic components of a simple pulse radar, a continuous wave radar, a pulse Doppler radar, and a monopulse radar. The characteristics of these components determine the capabilities and limitations of a particular radar system.

CHAPTER 5. RADAR PRINCIPLES

1. INTRODUCTION

The primary purpose of radar systems is to determine the range, azimuth, elevation, or velocity of a target. The ability of a radar system to determine and resolve these important target parameters depends on the characteristics of the transmitted radar signal. This chapter explains the relationship of radar frequency (RF), pulse repetition frequency (PRF), pulse width (PW), and beamwidth to target detection and resolution.

2. RADAR RANGE

A basic pulse radar system consists of four fundamental elements: the transmitter, the receiver, the antenna, and the synchronizer, or master timer.

a. The transmitter, through the antenna, sends out a pulse of RF energy at a designated frequency. The presence of a target is revealed when the RF energy bounces off the target, returns to the radar antenna, and goes into the receiver (Figure 5-1). The master timer measures the time between the transmission of a pulse and the arrival of a target echo.

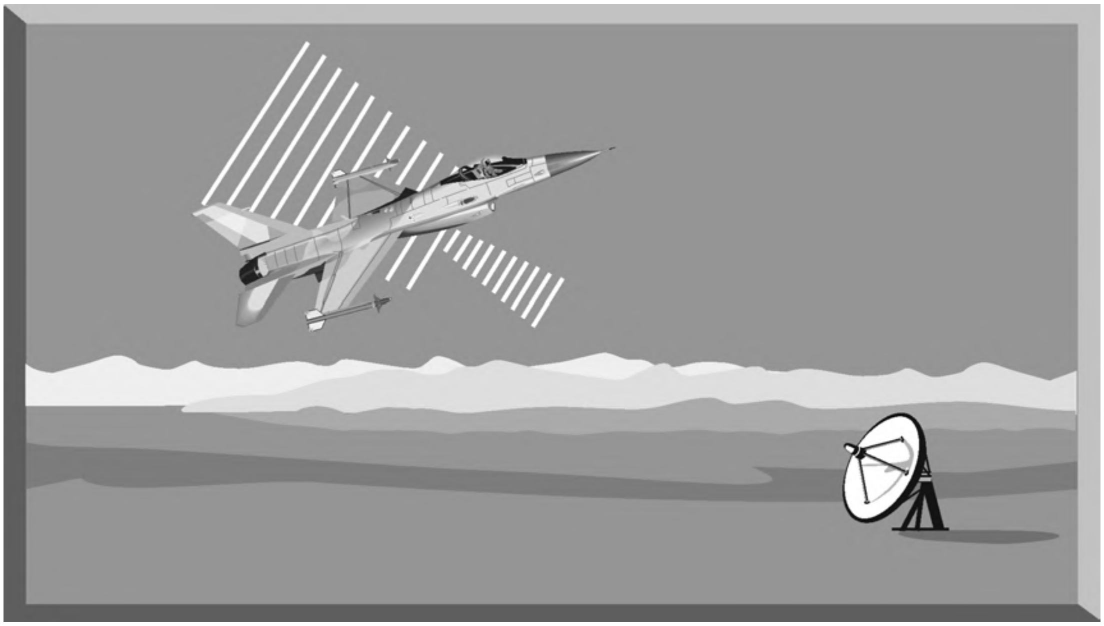

Figure 5-1. Basic Radar Pulse

(1) RF energy travels at the speed of light (c) which is 3 x 10^8 meters per second. Target range can be computed by using the basic radar range determination equation (Equation 5-1).

$$\text{Target Range} = \frac{\text{Measured Time} \times \text{Speed of Light (c)}}{2}$$

Equation 5-1. Basic Radar Range Determination Equation

(2) Another useful measurement is the radar mile, which is the round trip time for an RF wave to travel to and from a target one nautical mile away (Figure 5-2). Solving the radar range equation for time results in Equation 5-2. Substituting the appropriate values into the equation and solving for time gives measured time of 12.4 microseconds for a one nautical mile (1853 meters) round trip.

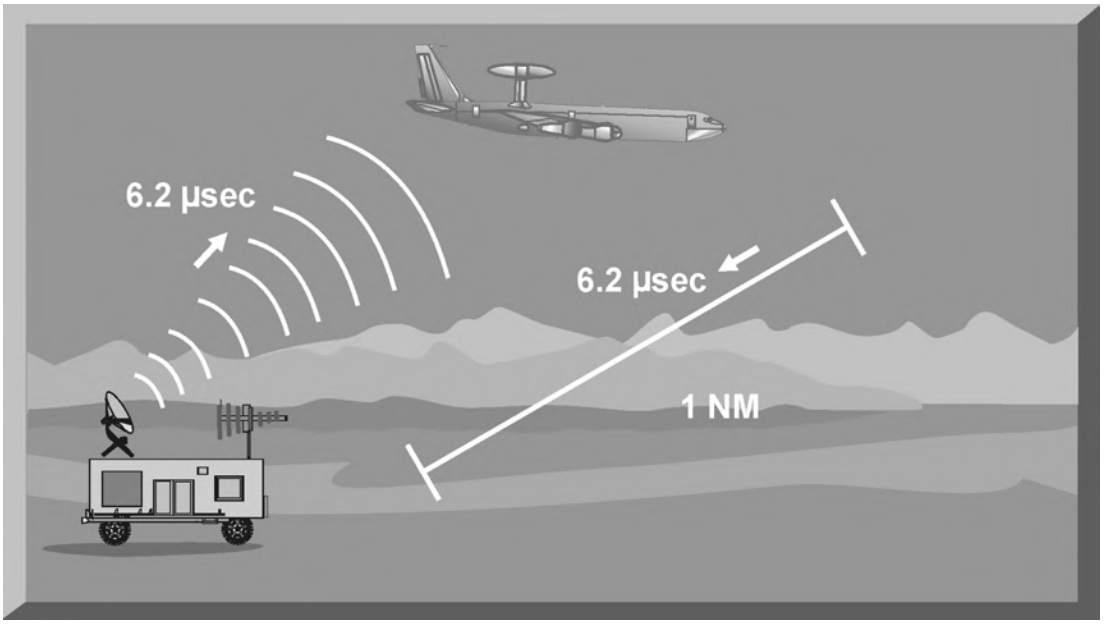

Figure 5-2. Radar Mile

$$\text{Measured Time} = \frac{\text{Target Range} \times 2}{\text{Speed of Light (c)}} = \frac{1853 \text{ meters} \times 2}{3 \times 10^{8} \text{ meters/second}} = 12.4\,\mu\text{sec}$$

Equation 5-2. Radar Mile

(3) A limitation on radar detection range is the concept of a second time around echo. A second time around echo occurs when a target echo associated with a particular radar pulse arrives at the antenna after another radar pulse has been transmitted. The radar master timer always assumes the target echo is associated with the last pulse transmitted. This makes the target echo ambiguous in range. Figure 5-3 depicts a radar signal with a pulse recurrence time (PRT) of 248 microseconds. Radar pulse A takes 372 microseconds to travel to the target and return. Using the range determination equation, actual target range is 30 nautical miles (nm). However, before the target echo returns to the antenna, radar pulse B is transmitted. The master timer associates the target echo of pulse A with radar pulse B, and calculates a target range of 10 nm. This ambiguous and false range is displayed to the operator.

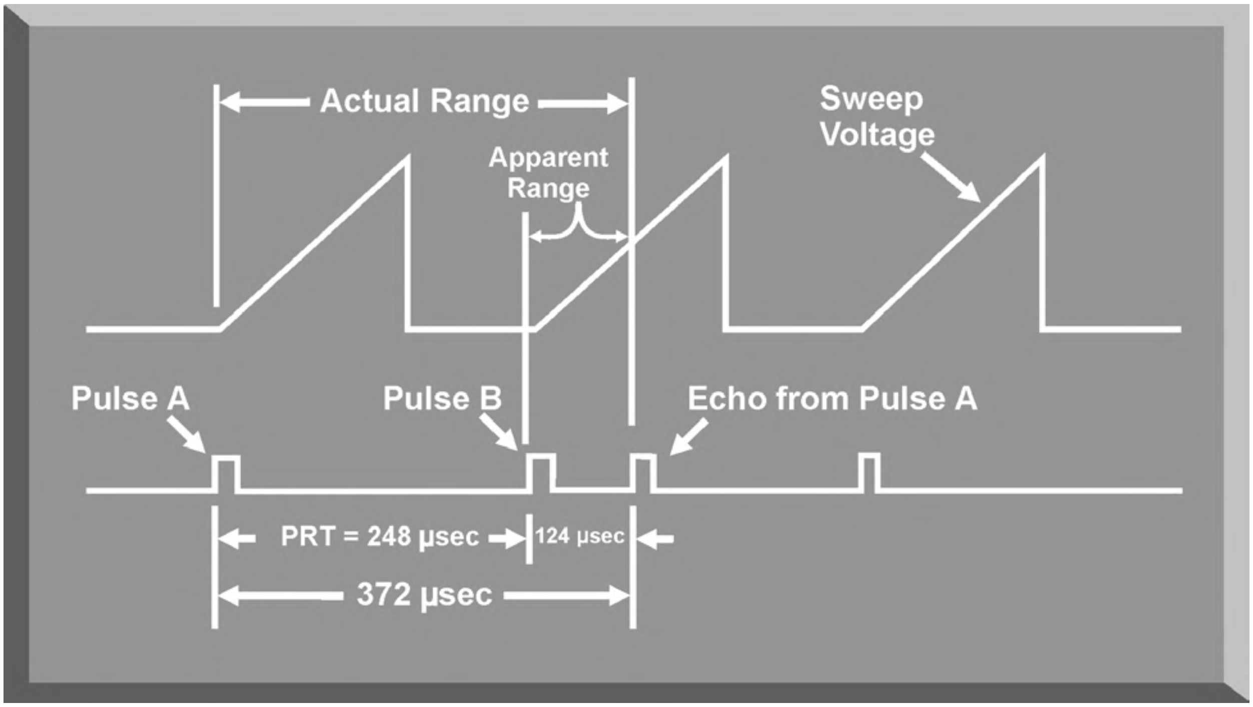

Figure 5-3. Second Time Around Echo

(4) Range ambiguities caused by second time around echoes limit the maximum unambiguous range of a radar system. This important capability can be calculated by using Equation 5-3. An analysis of this equation shows that a radar system designed for long-range detection should transmit a radar signal with a large PRT. In addition, as the PRF of a radar signal increases, the PRT decreases, and the maximum unambiguous range decreases.

$$\text{Maximum Unambiguous Range} = \frac{\text{PRT } (\mu\text{secs})}{12.4\ (\mu\text{secs/nm})}$$

Equation 5-3. Maximum Unambiguous Range

b. A critical aspect of range determination is range resolution. Range resolution is the ability of a radar to separate two targets that are close together in range and are at approximately the same azimuth (Figure 5-4). The range resolution capability is determined by pulse width. Pulse width is the time that the radar is transmitting RF energy. Pulse width is measured in microseconds.

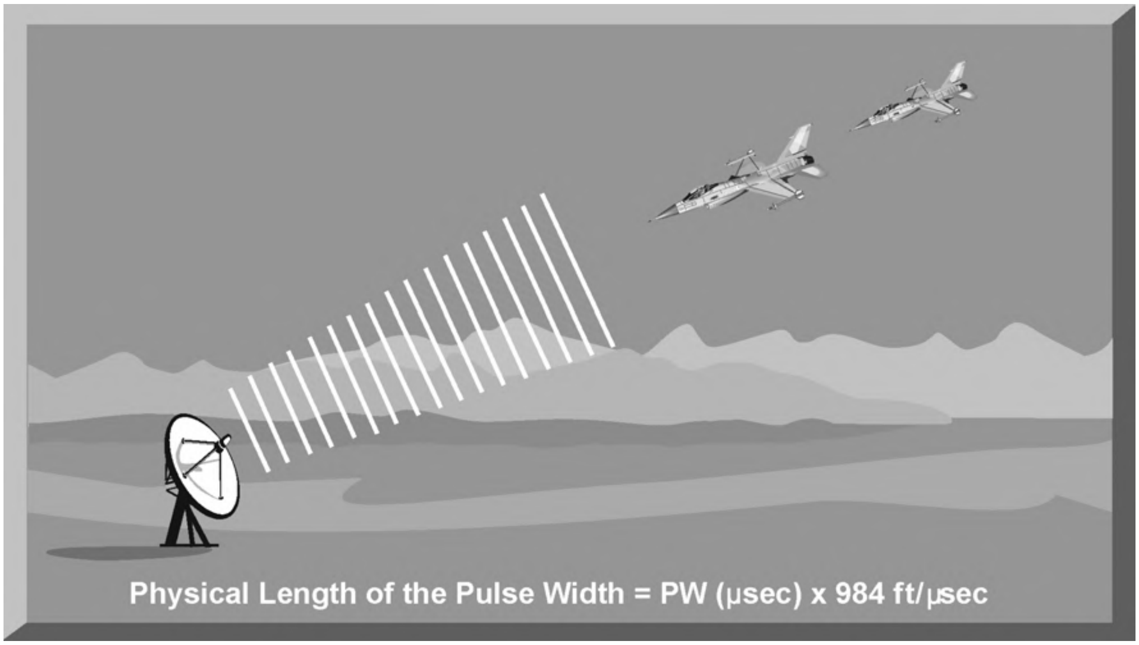
Physical Length of the Pulse Width = PW (µsec) x 984 ft/µsec

Figure 5-4. Radar Pulse

(1) A radar pulse in free space occupies a physical distance equal to the pulse width multiplied by the speed of light, which is about 984 feet per microsecond. If two targets are closer together than one-half of this physical distance, the radar cannot resolve the returns in range, and only one target will be displayed.

(2) To illustrate range resolution, consider Figure 5-5, in which two aircraft are separated by a distance of one-half the pulse width or less. At T1, the leading edge of the radar pulse hits the lead aircraft.

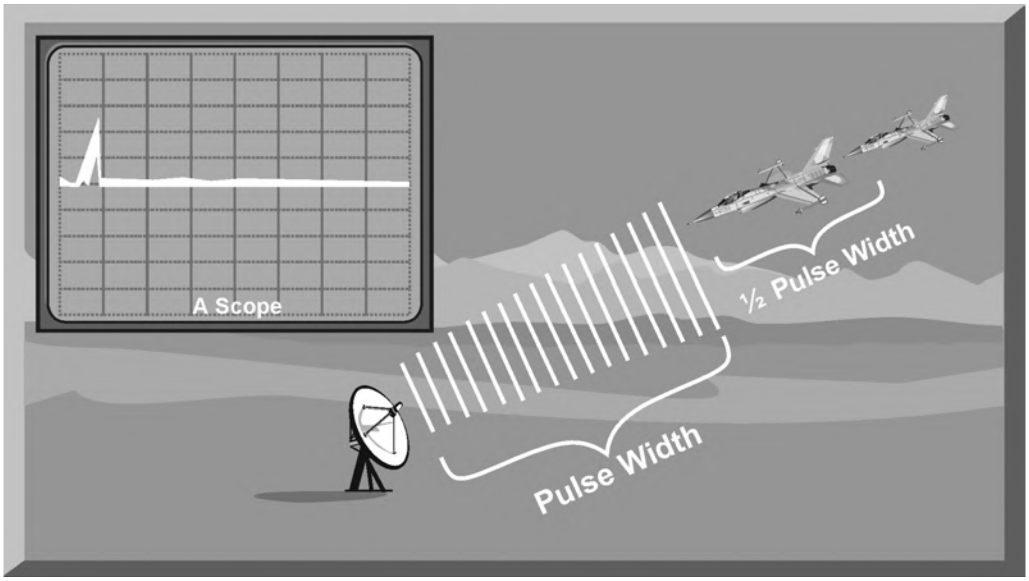

Figure 5-5. Radar Pulse at T1

(3) In Figure 5-6, at T2 the leading edge of this same pulse hits the trailing aircraft. Since the trailing aircraft is less than one-half the pulse width from the lead aircraft, the return echo from the lead aircraft is received by the antenna before the entire pulse has left the trailing aircraft.

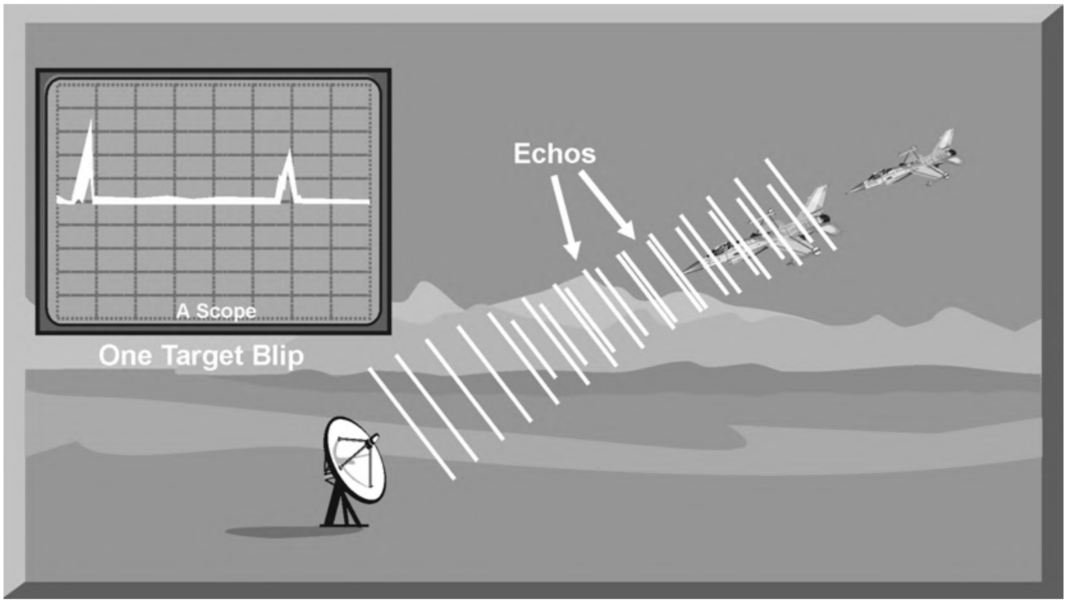

Figure 5-6. Radar Pulse at T2

(4) T3 depicts the merging target echoes at the radar (Figure 5-7). The radar would display only one target in this situation, such as the one shown on the radar scope in Figure 5-6.

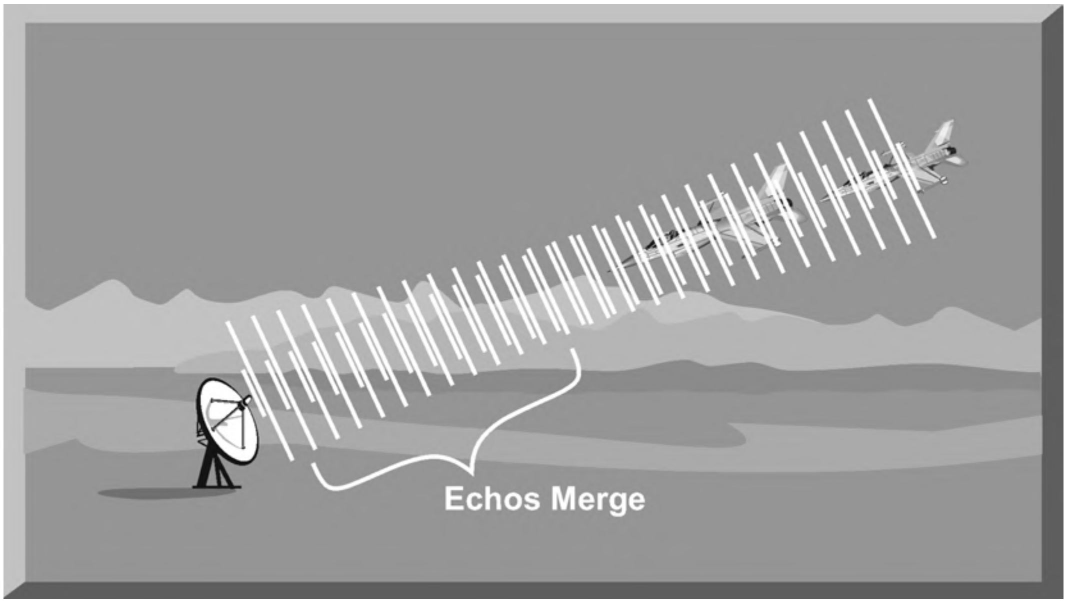

Figure 5-7. Radar Pulse at T3

(5) Two targets separated by more than one-half the pulse width, as in Figure 5-8, will be displayed as two targets. In this case, the transmitted pulse is completely past the lead aircraft before the return echo from the trailing aircraft reaches the antenna.

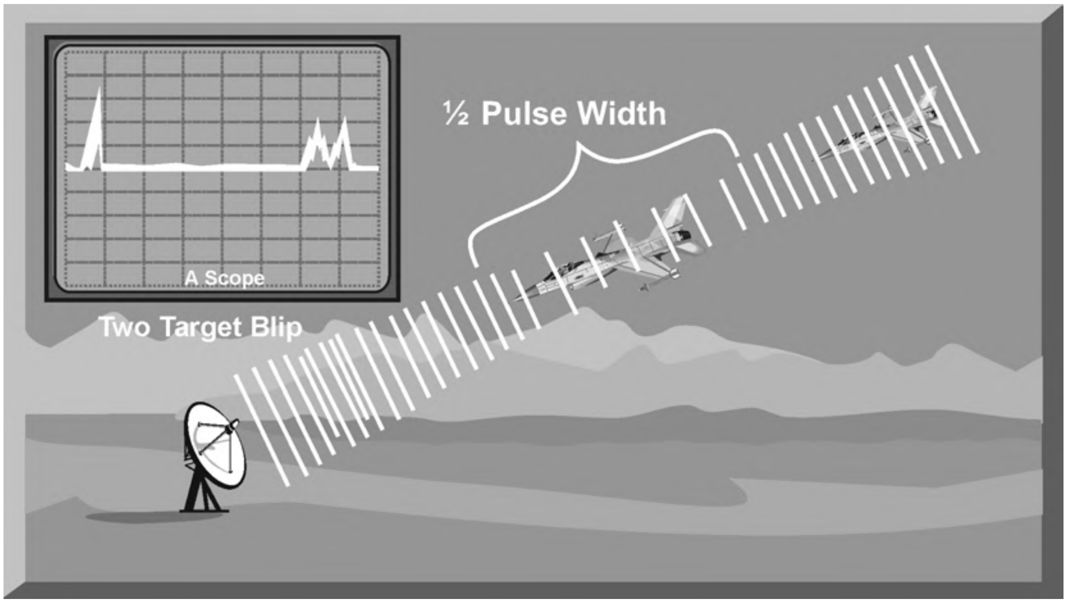

Figure 5-8. Radar Range Resolution

(6) The range resolution of the radar is usually expressed in feet and can be computed using Equation 5-4. It is the minimum separation required between two targets in order for the radar to display them separately on the radar scope.

$$\text{Range Resolution} = \frac{\text{Pulse Width} \times 984 \text{ feet}}{2}$$

Equation 5-4. Range Resolution

3. AZIMUTH DETERMINATION

The beamwidth of a radar system is the horizontal and vertical thickness of the radar beam (Figure 5-9). Beamwidth depends on antenna design and is normally measured in degrees from the center of the beam to the point at which the power drops off by half. This half-power point is -3 dB in power drop-off. Beamwidth governs the azimuth and elevation accuracy and resolution capability of a radar system in the same way that pulse width governs radar range accuracy and resolution.

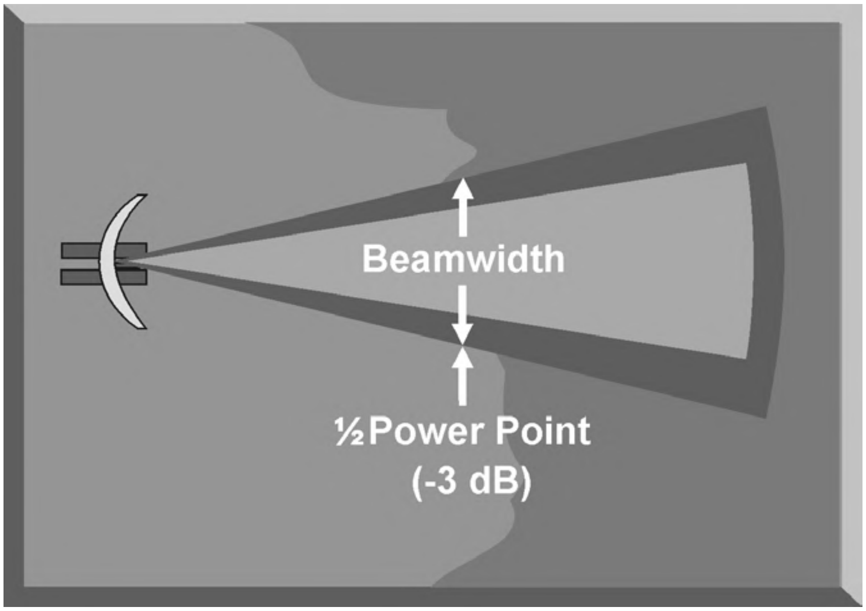

Figure 5-9. Radar Beamwidth

a. In order for a radar system to figure out target azimuth, the antenna must be aligned with a point of reference and pointed at the target during the transmission and reception of several pulses of radar energy. If the antenna is

referenced to true North, the azimuth of the target can be measured relative to true North (Figure 5-10). Azimuth determination is based on the position of the antenna when the target is being illuminated.

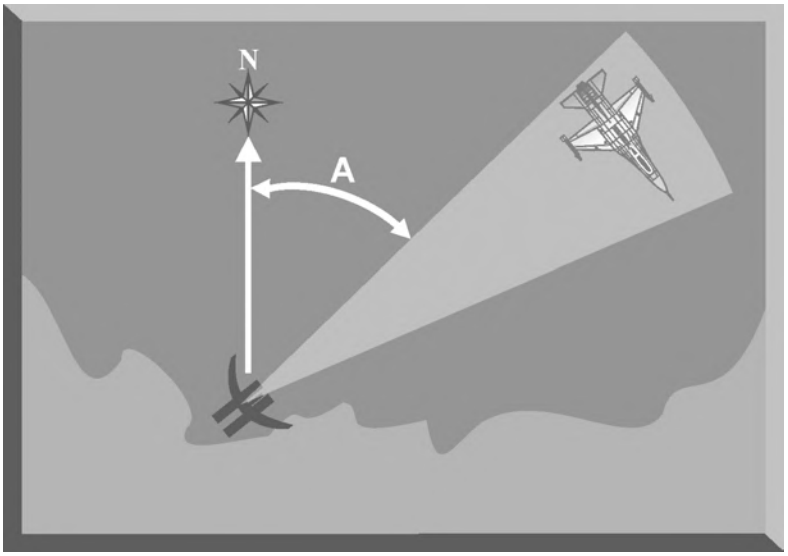

Figure 5-10. Azimuth Determination

(1) To provide accurate azimuth determination over a large area, many radars employ a narrow beam and scan the antenna in a predictable pattern. The most common scan pattern is a 360° circular scan at a constant rate (Figure 5-11). The plan position indicator (PPI) radar scope display is normally associated with this scan pattern. As the radar beam sweeps, a target is detected and displayed. The position of the antenna, when the target is displayed, shows the relative azimuth.

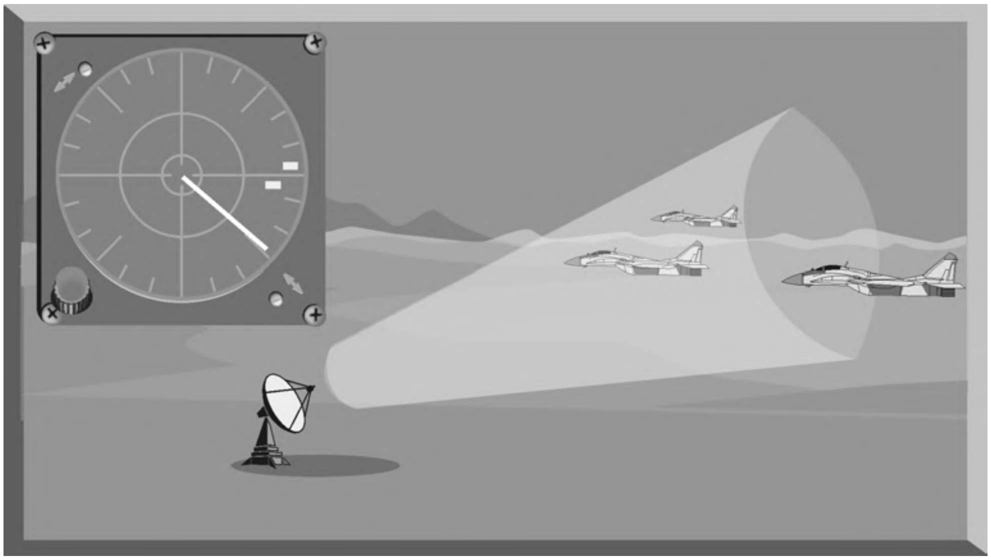

Figure 5-11. Antenna Scan

(2) The azimuth accuracy of a radar system is determined by the horizontal beamwidth (HBW). In Figure 5-12, radar system A has a horizontal beamwidth of 10°. As the beam sweeps, the target is illuminated for as long as it is in the beam. This means that the target covers 10° in azimuth on the PPI scope. Radar system B has a beamwidth of 1°. A target displayed on the PPI scope will cover 1° in azimuth. The narrower the horizontal beamwidth, the better the azimuth accuracy.

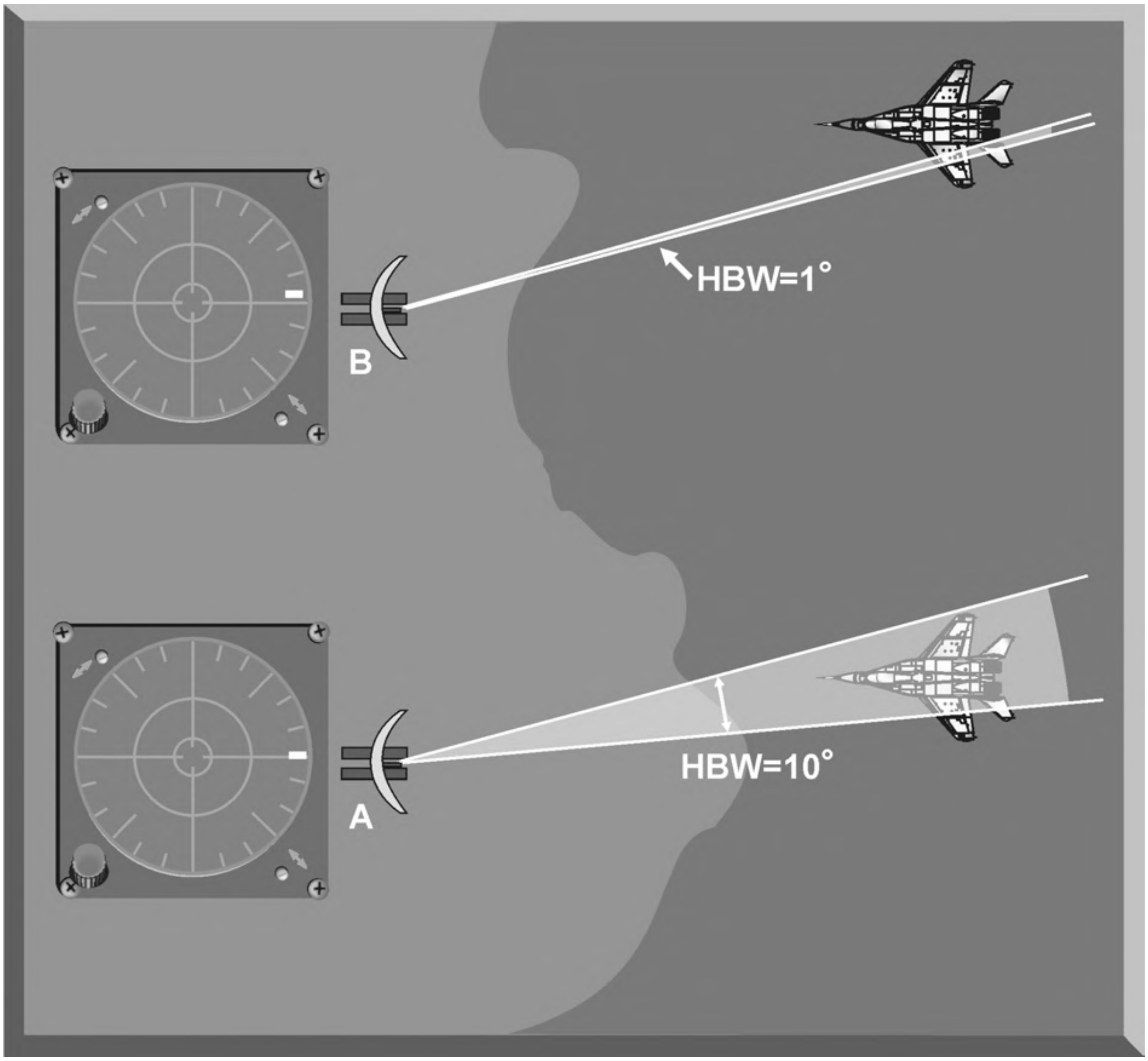

Figure 5-12. Horizontal Beamwidth Comparison

b. Azimuth resolution is the ability of a radar to display two targets flying at approximately the same range with little angular separation, such as two fighters flying line-abreast tactical formation. The azimuth resolution capability is usually expressed in nautical miles and corresponds to the minimum azimuth separation

required between two targets for separate display. Azimuth resolution depends on the horizontal beamwidth of the radar. The radar system in Figure 5-13 has a horizontal beamwidth of 10°. The two targets are so close in azimuth that the return echoes are blended into one return.

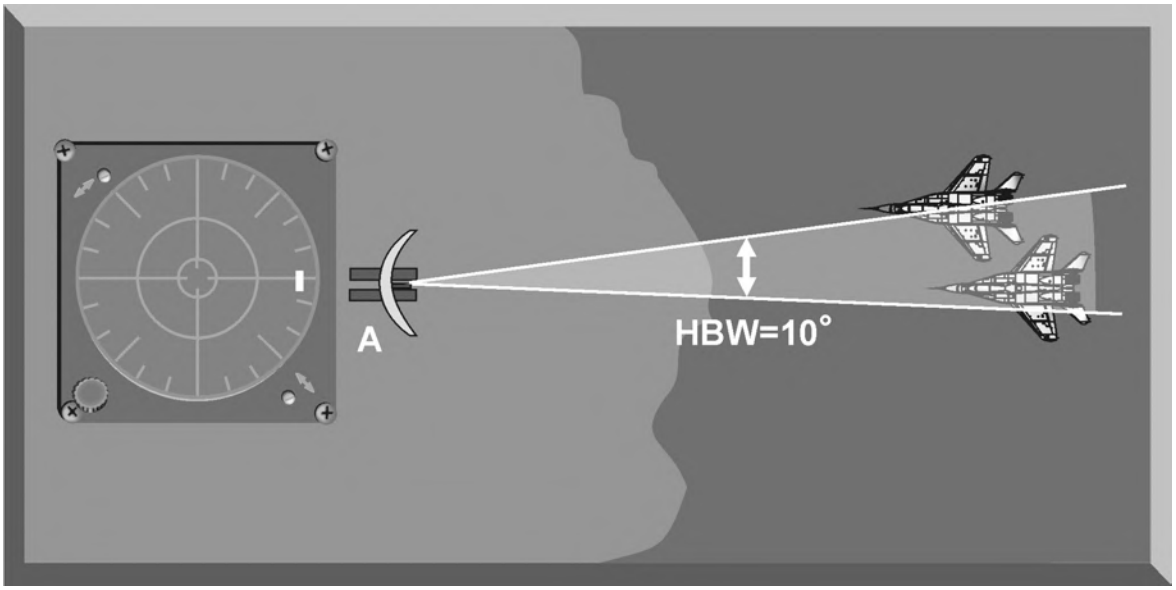

Figure 5-13. Horizontal Beamwidth and Azimuth Resolution

(1) The radar system in Figure 5-14 has a horizontal beamwidth of 1°. The radar beam not only hits the targets, but passes between them without causing a return. This allows the radars to display two distinct radar returns. A small horizontal beamwidth improves azimuth resolution.

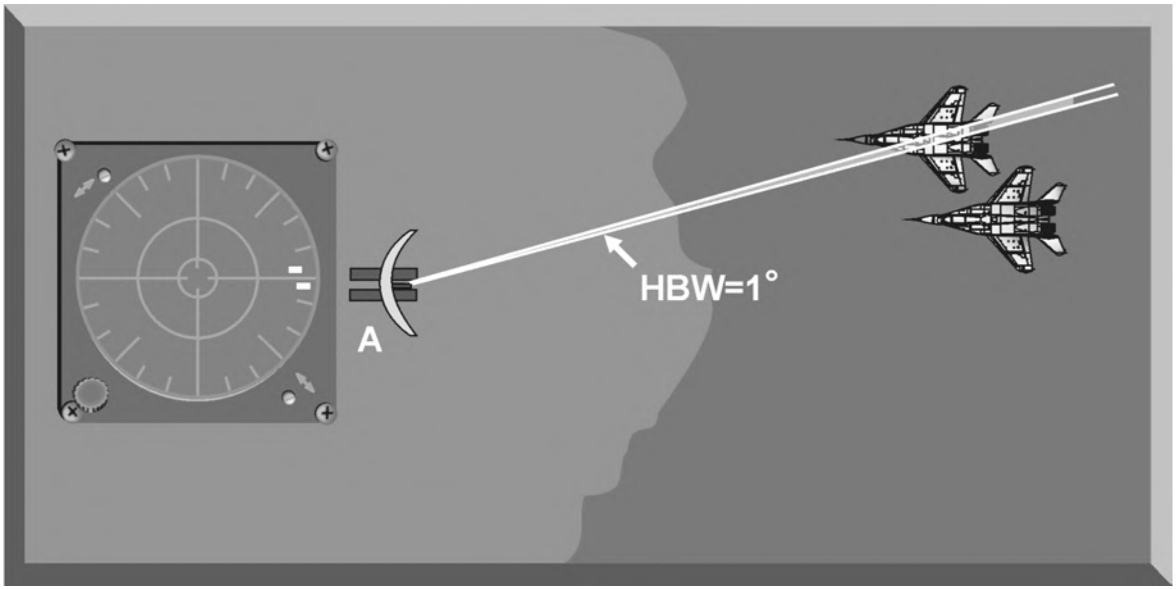

Figure 5-14. Azimuth Resolution

(2) Azimuth resolution, in nautical miles, can be computed using Equation 5-5. Notice that this equation is the "60 to 1 rule" used for navigation. A 1° beamwidth will yield a one-mile-wide cell at 60 nautical miles.

$$\text{Azimuth Resolution} = \frac{\text{Horizontal Beamwidth x Range}}{60}$$

Equation 5-5. Azimuth Resolution

4. ELEVATION DETERMINATION

Since a radar beam is three-dimensional, the vertical beamwidth is the primary factor in determining altitude resolution capability. Altitude resolution is the ability of a radar to display two targets flying at approximately the same range and azimuth with little altitude separation, such as two fighters flying a vertical stack formation. The altitude resolution capability is usually expressed in feet and corresponds to the minimum altitude separation required between two targets for separate display. The radar system in Figure 5-15 has a vertical beamwidth of 10°. The two targets are so close in altitude that the return echoes depicted on the range height indicator (RHI) are blended into one.

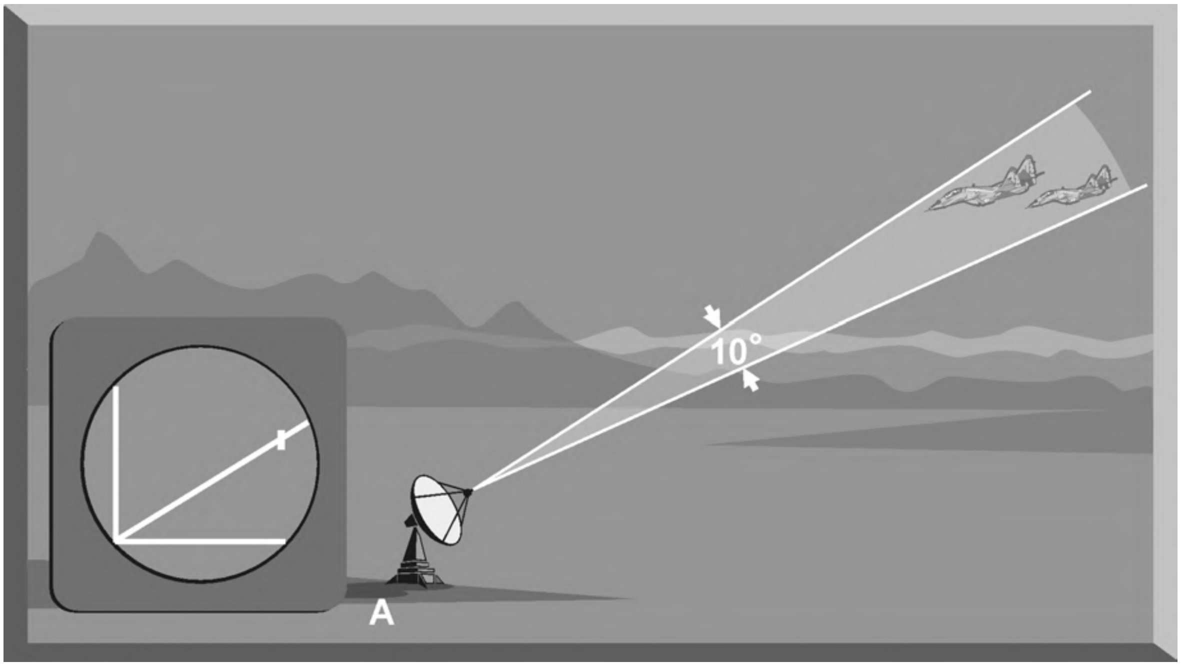

Figure 5-15. Vertical Beamwidth and Elevation Resolution

a. The radar system depicted in Figure 5-16 has a vertical beamwidth of 1°. This small beam not only hits the targets, but passes between them without causing a return. This allows the radar to display two distinct targets.

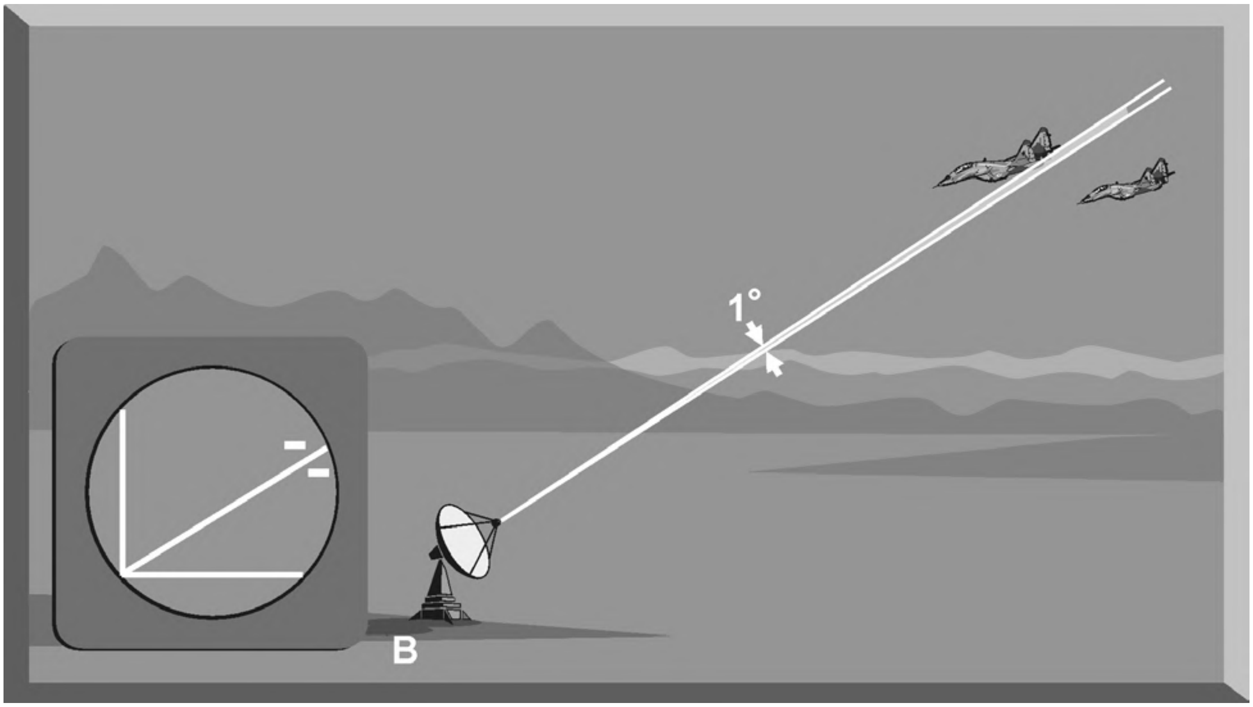

Figure 5-16. Elevation Resolution

b. Altitude/elevation resolution, in thousands of feet, can be computed using Equation 5-6.

$$\text{Altitude Resolution} = \frac{\text{Vertical Beamwidth} \times \text{Range}}{60}$$

Equation 5-6. Altitude Resolution

5. RADAR RESOLUTION CELL

A radar's pulse width, horizontal beamwidth, and vertical beamwidth form a three-dimensional resolution cell (RC) (Figure 5-17). A resolution cell is the smallest volume of airspace in which a radar cannot determine the presence of more than one target. The resolution cell of a radar is a measure of how well the radar can resolve targets in range, azimuth, and altitude. The horizontal and vertical

dimensions of a resolution cell vary with range. The closer to the radar, the smaller the resolution cell.

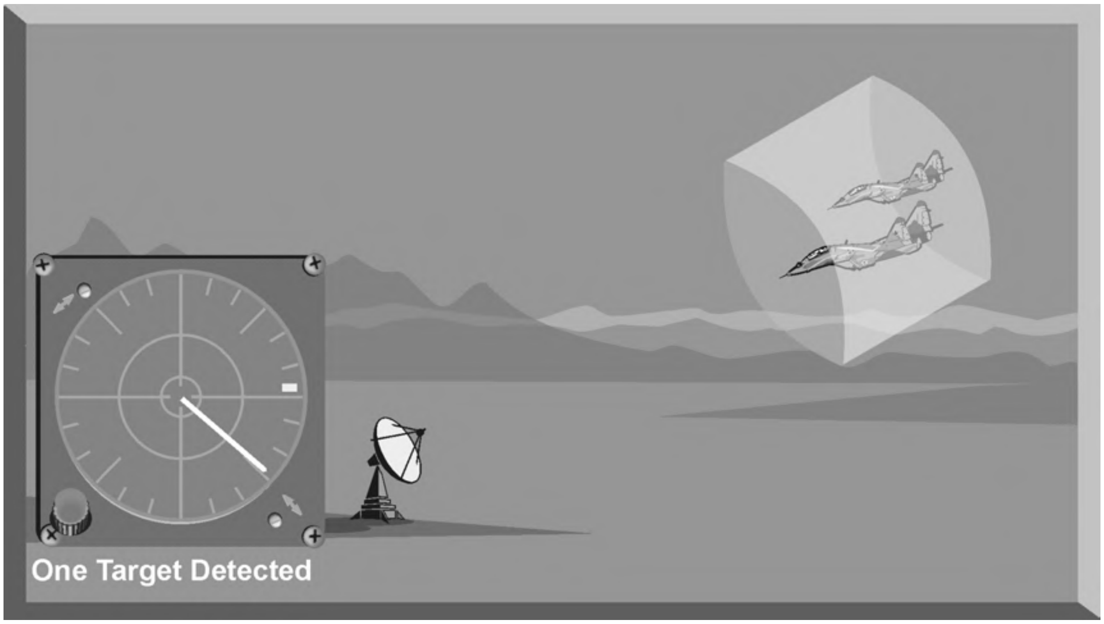

Figure 5-17. Radar Resolution Cell

a. The physical dimensions of a radar's resolution cell can be computed. For a radar with a pulse width of 1 microsecond, a horizontal beamwidth of 1°, and a vertical beamwidth of 10°, the formulas for range resolution, azimuth resolution, and altitude resolution can be used to compute the dimensions of the resolution cell. In the example in Figure 5-18, at a target range of 10 nm, the physical dimensions of the radar's resolution cell are 492 feet in range, by 1000 feet in azimuth, and 10,000 feet in altitude. These figures can be confirmed by using Equations 5-4, 5-5, and 5-6.

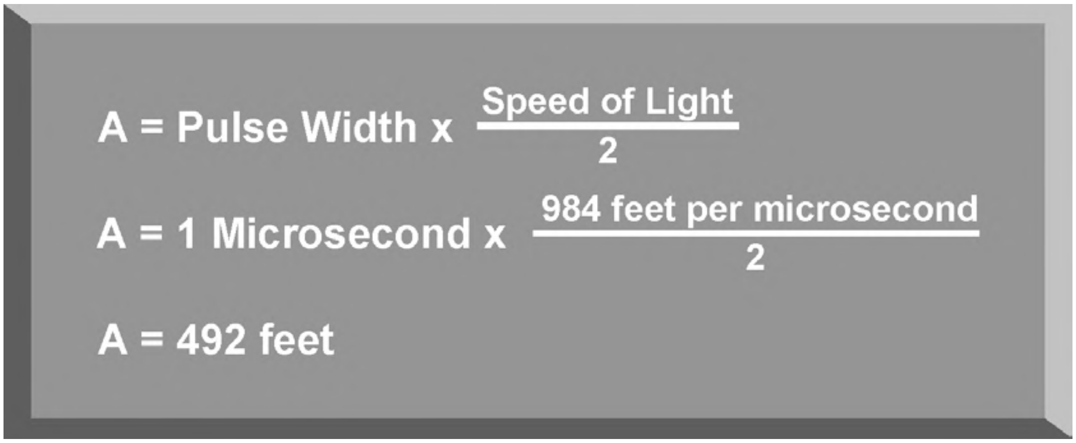

$$A = \text{Pulse Width} \times \frac{\text{Speed of Light}}{2}$$

$$A = 1 \text{ Microsecond} \times \frac{984 \text{ feet per microsecond}}{2}$$

$$A = 492 \text{ feet}$$

Figure 5-18. Radar Resolution Cell Dimensions

b. Based on these computations, two, or more, aircraft flying a trail formation closer than 492 feet would be displayed as a single target. Two, or more, aircraft flying line abreast closer than 1000 feet would be displayed as a single target. Two, or more, aircraft flying a vertical stack closer than 10,000 feet would be displayed as a single target. This also shows that the shorter the pulse width, the better the range resolution capability of a radar system. The narrower the horizontal beamwidth, the better the azimuth resolution capability. The narrower the vertical beamwidth, the better the altitude resolution capability.

c. Another type of resolution is velocity resolution. For a Doppler radar aircraft flying within the conventional resolution cell described above can be distinguished as separate targets if they have enough speed differential. Paragraph 6 below will describe how this is carried out.

6. PULSE DOPPLER VELOCITY DETERMINATION

To fully understand how a pulse Doppler radar determines target velocity, it is necessary to know more about the pulsed waveform. To generate a pulse modulated wave, a continuous carrier sine wave, like the output from a CW radar, is combined with a rectangular wave, like that of a pulse radar, to produce the pulse modulated waveform. Figure 5-19 depicts pulse modulation.

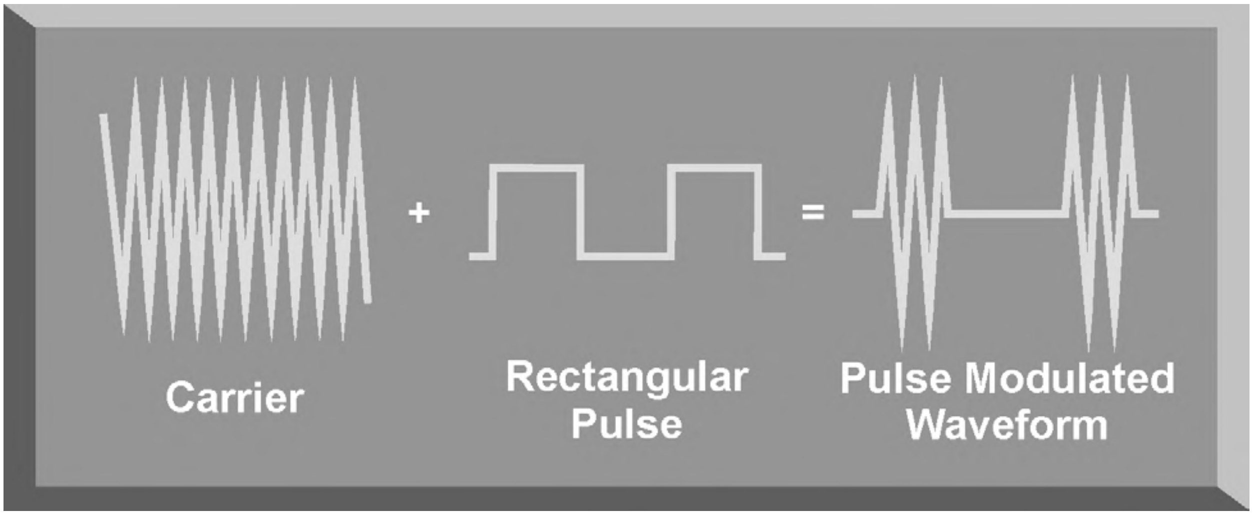

Figure 5-19. Pulse Modulation

a. Mathematically, any waveform other than a sine wave is composed of many different pure sine waves added in the proper amplitude and phase relationships (Figure 5-20). In a pulsed modulated waveform, the sine waves correspond to the fundamental frequency, which is the PRF, and the sum of all harmonics in the proper amplitude and phase. The frequency of the harmonic is the basic frequency plus or minus a multiple of the PRF.

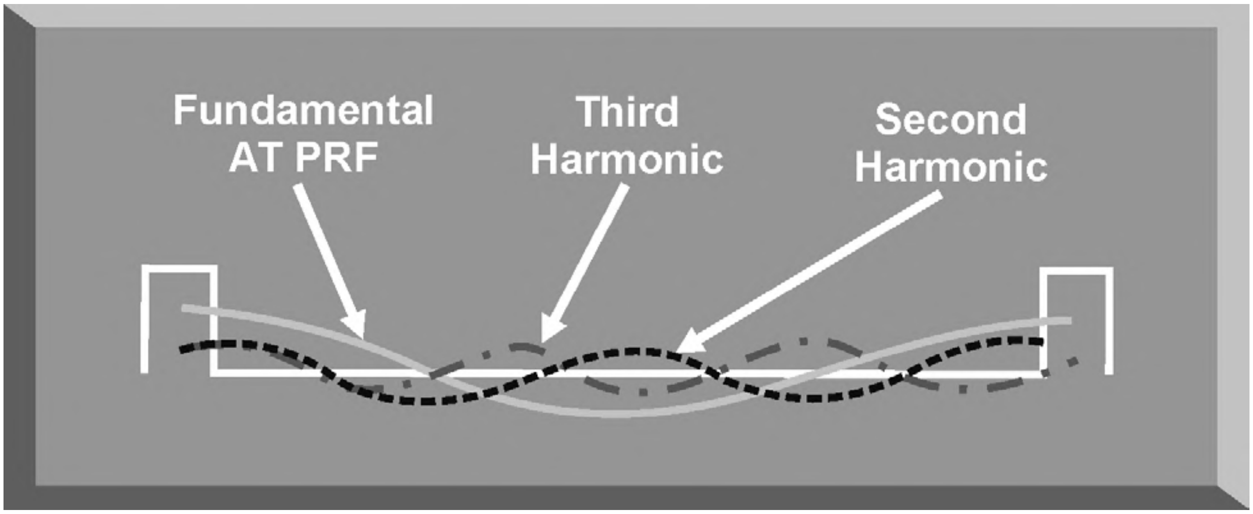
Fundamental AT PRF
Third Harmonic
Second Harmonic

Figure 5-20. Harmonics of a Pulse Modulated Waveform

(1) Figure 5-21 is a plot of the harmonic content of a pulse modulated waveform operating at a carrier frequency of 2800 megahertz (MHz) with a PRF of 1 MHz. Note the loops of frequencies on either side of the carrier frequency. These are the additions and subtractions of all the frequencies in the rectangular pulse to the carrier frequency. The important thing to remember is that there are many frequencies present, and a pulse Doppler radar must deal with a crowded frequency spectrum. This becomes even more important when one considers the fact that every frequency present will experience a Doppler shift when it is reflected by a moving target. The individual frequencies shown are called spectral lines.

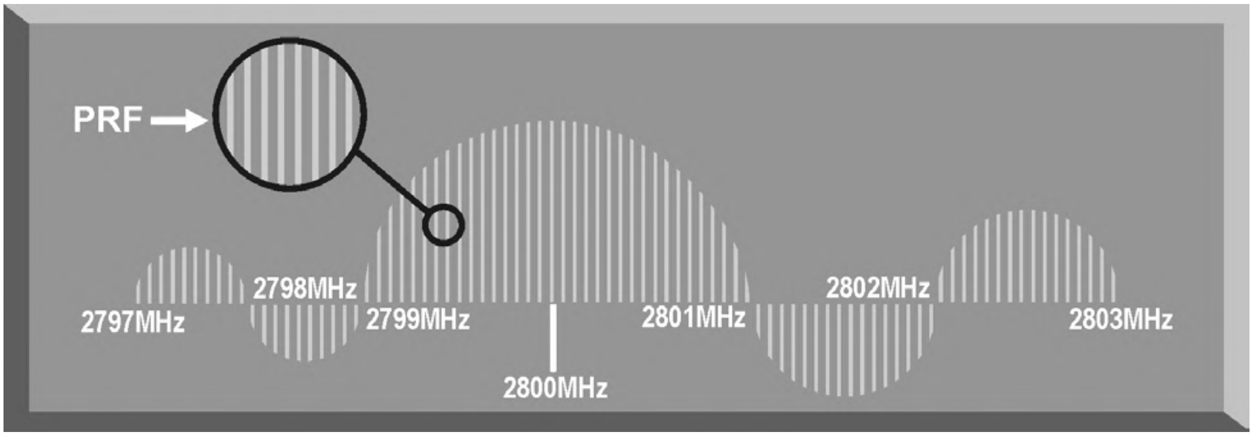
PRF
2797MHz
2798MHz
2799MHz
2800MHz
2801MHz
2802MHz
2803MHz

Figure 5-21. Harmonic Content

(2) For a pulse Doppler radar to accurately measure velocity, it must compare the frequency change, or Doppler shift, between the carrier frequency

and the frequency returning from the target. It is a difficult task for the radar to differentiate between the returning carrier and all the harmonic frequencies (Figure 5-22).

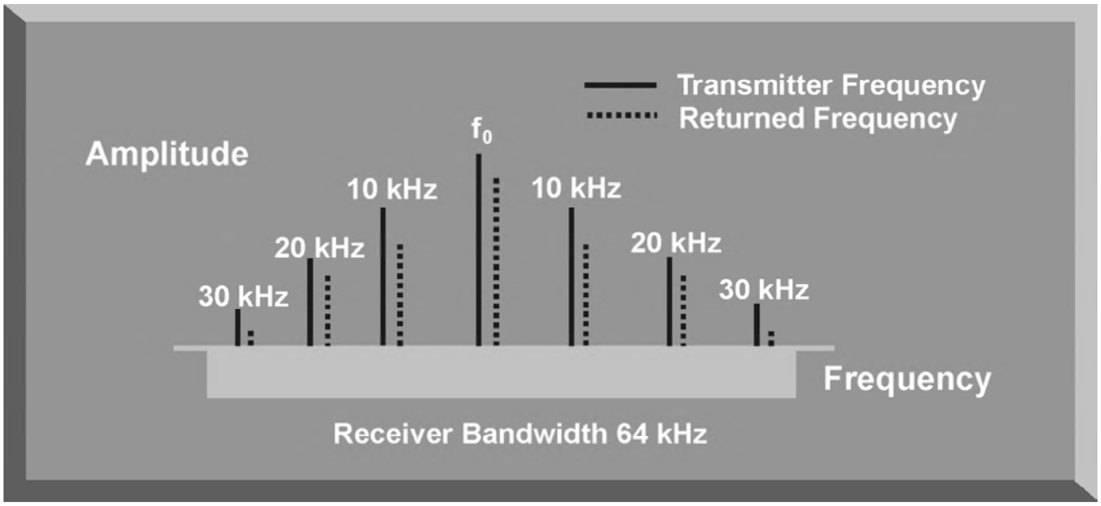
Transmitter Frequency
Returned Frequency
Amplitude
f_0
10 kHz
10 kHz
20 kHz
20 kHz
30 kHz
30 kHz
Frequency
Receiver Bandwidth 64 kHz

Figure 5-22. Spectral Line Frequencies

b. The radar differentiates between the returning carrier frequencies and all other harmonic frequencies by using clutter cancellers, or filters, at the known harmonic frequencies (Figure 5-23). The radar cannot process frequencies cancelled by these filters. The filters create "blind speeds" for the radar. The closer together the spectral lines, the more "blind speeds" the radar will have.

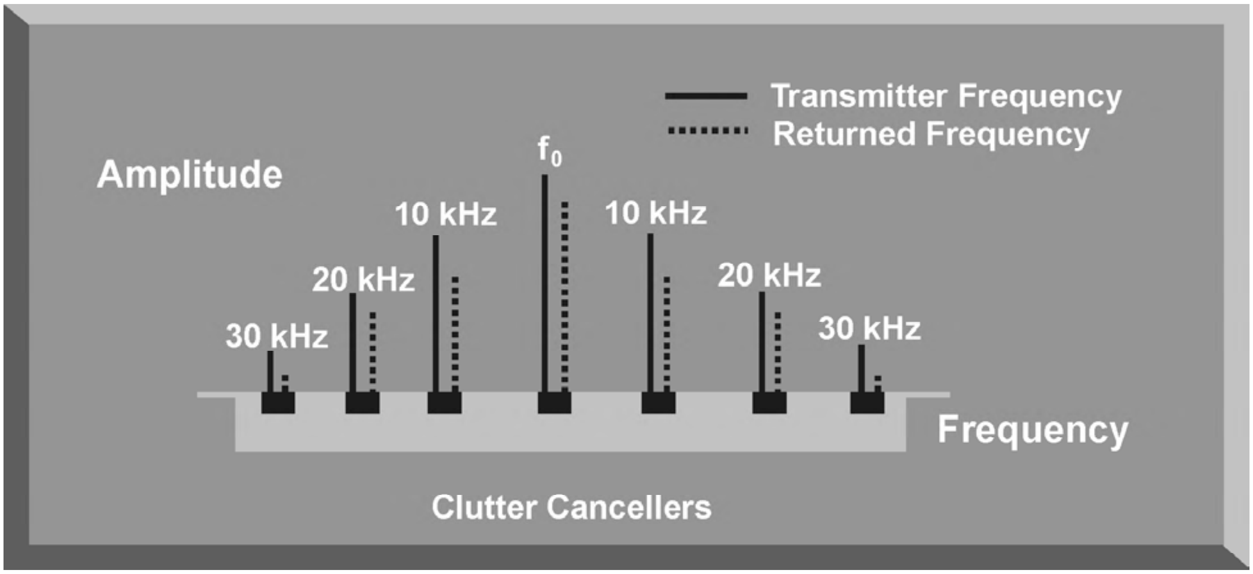
Transmitter Frequency
Returned Frequency
Amplitude
f_0
10 kHz
10 kHz
20 kHz
20 kHz
30 kHz
30 kHz
Frequency
Clutter Cancellers

Figure 5-23. Selective Clutter Canceling

(1) Since the position of the harmonics in relation to the carrier frequency is based on PRF, the number of blind speeds can be reduced by changing the PRF of the radar. The higher the PRF, the wider the spacing of the spectral lines and the fewer blind speeds due to selective clutter canceling. However, a high PRF increases the problem of range ambiguities. Most modern pulse Doppler radars employ a medium and high PRF mode. Medium PRF equates to fewer range ambiguities but more blind speeds. High PRF has fewer blind speeds but more range ambiguities (Figure 5-24).

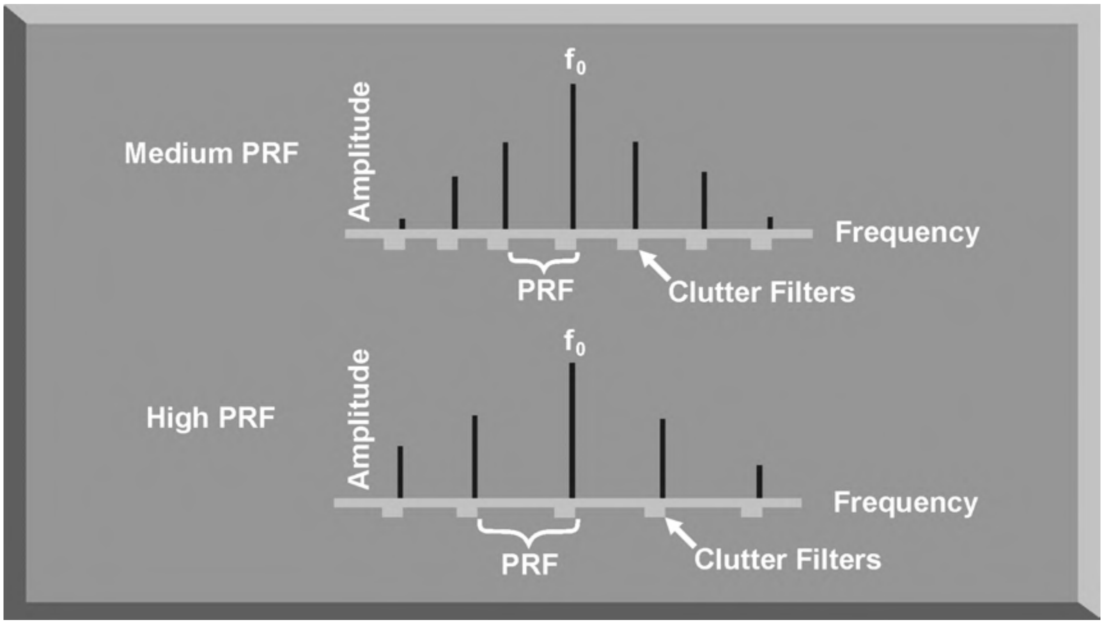

Figure 5-24. PRF and Spectral Lines

(2) To separate the returning target frequency shifts from all other frequencies in the returning waveform, the pulse Doppler radar employs filters to cancel the known harmonic frequency shifts. In addition, the radar cancels out all returns with no frequency shift, which equates to canceling all returns with no movement relative to the radar. However, if the radar has too many clutter filters, this creates multiple blind speeds, and targets will be missed. In Figure 5-25, the detection filters allow target frequencies to be processed, and clutter filters cancel unwanted frequency shifts. Target 1 will be detected, but Target 2 will be canceled.

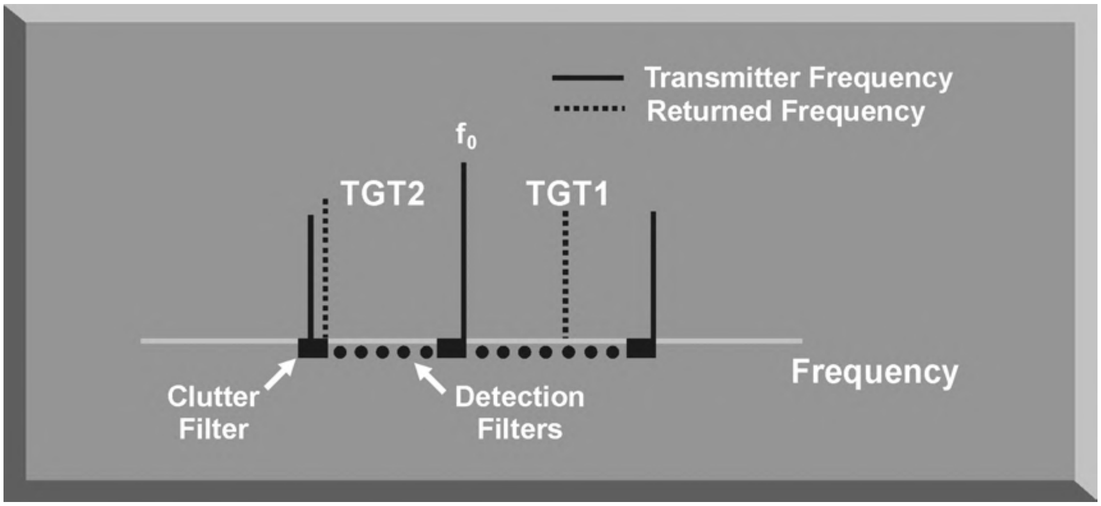

Figure 5-25. Pulse Doppler Filters

7. BASIC RADAR EQUATION

The basic radar equation relates the range of a radar system to the characteristics of the transmitter, receiver, antenna, and the target. The radar equation provides a means not only to figure out the maximum range of a particular radar system, but it can be used to understand the factors that affect radar operation. In this section, the simple forms of the radar equation are developed, starting with the power density of the transmitting antenna to the power received by the receiving antenna.

a. Power density is the power of a radio wave per unit of area normal to the direction of propagation. The power density generated by a practical antenna can be expressed in Equation 5-7.

$$\text{Power Density from Antenna} = \frac{P_T \times G}{4\pi r^2}$$

P_T = transmitted power
G = antenna gain
r = radius of the antenna

Equation 5-7. Power Density From an Antenna

b. As the radar beam propagates through space, it arrives at a target at some range (R) from the antenna. As the radar beam travels through space, the wavefront of the beam expands to a very large cross-sectional area, especially in relation to the target dimensions. The power density of the radar beam, across this wide area, at the target, is detailed in Equation 5-8.

$$\text{Power Density at Target} = \frac{P_T \times G}{4\pi R^2}$$

P_T = transmitted power
G = antenna gain
R = range to the target

Equation 5-8. Power Density at the Target

c. Since the cross-sectional area of the radar beam is so large, only a small portion of the total power in the beam can be reflected toward the antenna. The rest of the radar energy continues through space and is dissipated, absorbed, or reflected by other targets. The small portion of the radar beam that hits the target is reradiated in various directions. The measure of the amount of incident power intercepted by the target and reradiated back in the direction of the antenna depends on the radar cross section (RCS) of the target. Equation 5-9 details the power density of the target echo signal reflected back to the radar antenna.

$$\text{Power Density at Antenna} = \frac{P_T \times G}{4\pi R^2} \times \frac{\sigma}{4\pi R^2}$$

P_T = transmitted power
G = antenna gain
σ = RCS
R = range to the target

Equation 5-9. Power Density at the Antenna

d. As the target echo reaches the antenna, part of the echo is captured by the antenna based on the effective aperture (A_e). Equation 5-10 details the actual signal power received by the radar system. This is one form of the basic radar equation and is the signal strength of a radar return from a specific target at range (R) from the radar.

$$\text{Signal Power Density } (S) = \frac{P_T G \sigma A_e}{(4\pi)^2 R^4}$$

P_T = transmitted power
G = antenna gain
σ = RCS
R = range to the target
A_e = antenna aperture area

Equation 5-10. Signal Power Density

e. A detailed analysis of this equation is not required to draw some basic conclusions about the factors affecting the detection of an aircraft. If any factor in the numerator, such as transmitted power, is increased by a factor of three, the signal received by the radar will increase by only 30 percent. This clearly shows why radar system operation is characterized by the transmission of megawatts of power and the reception of microwatts of returning power. In addition, this equation shows that the most critical factor in determining radar detection is target range.

f. The maximum radar range (R_{MAX}) occurs when the signal power density received just equals the minimum detectable signal (S_{MIN}) for the receiver. Solving Equation 5-11 for range, and substituting S_{MIN}, yields the basic radar equation for R_{MAX} for a specific target. This is another form of the basic radar equation.

$$R_{MAX} = \frac{\left[P_T G \sigma A_e\right]^{1/4}}{\left[(4\pi)^2 S_{MIN}\right]}$$

P_T = transmitted power
G = antenna gain
σ = RCS
A_e = antenna aperture area
R = range to the target

Equation 5-11. Basic Radar Equation

g. It is important to note that the basic radar equations do not consider such factors as meteorological conditions, changes in aircraft RCS, the impact of clutter on gain, or operator abilities. The radar equation does explain why a radar system designed for long-range detection should transmit a very high power signal, concentrated into a narrow beam, collected by a large antenna, and processed by a very sensitive receiver.

8. SUMMARY

This chapter has discussed the methods employed by radar systems to determine target range, azimuth, elevation, and velocity. The relationship between pulse width and range resolution, beamwidth and azimuth/elevation resolution, and PRF and velocity resolution have been explained. In addition, an explanation of the complex radar equation has been presented. The capabilities and limitations of a specific radar system to determine these critical target parameters is the key to understanding the countermeasures designed to defeat this system.

CHAPTER 6. ANTENNA CHARACTERISTICS AND SCANS

1. INTRODUCTION

The function of the antenna during transmission is to concentrate the radar energy from the transmitter into a shaped beam that points in the desired direction. During reception, or listening time, the function of the antenna is to collect the returning radar energy, contained in the echo signals, and deliver these signals to the receiver. Radar antennas are characterized by directive beams that are usually scanned in a recognizable pattern. The primary antenna types in use today fall into three categories: parabolic, Cassegrain, or phased array antennas. Additionally, the method radar antennas employ to sample the environment is a critical design feature of the radar system. The scan type selected for a particular radar system often decides the employment of that radar in an integrated air defense system (IADS). The process the radar antenna uses to search airspace for targets is called scanning or sweeping. This chapter discusses circular, unidirectional, bidirectional, helical, raster, Palmer, and conical scans, and track-while-scan (TWS) radar systems.

2. PARABOLIC ANTENNA

One of the most widely used radar antennas is the parabolic reflector (Figure 6-1). The parabola-shaped antenna is illuminated by a source of radar energy, from the

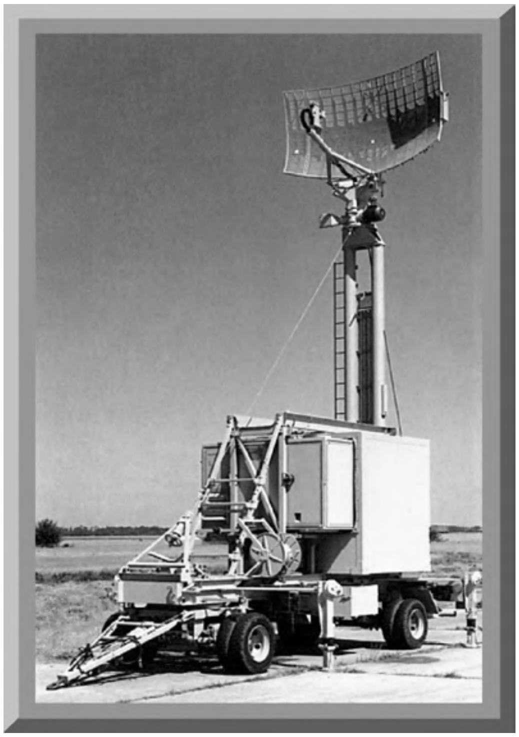

Figure 6-1. Parabolic Antenna

transmitter, called the feed. The feed is placed at the focus of the parabola, and the radar energy is directed at the reflector surface. Because a point source of energy, located at the focus, is converted into a wavefront of uniform phase, the parabola is well suited for radar antenna applications. By changing the size and shape of the parabolic reflecting surface, a variety of radar beam shapes can be transmitted.

a. The antenna depicted in Figure 6-1 generates a nearly symmetrical pencil beam that can be used for target tracking.

b. Elongating the horizontal dimensions of the parabolic antenna creates a radar antenna called the parabolic cylinder antenna (Figure 6-2). The pattern of this antenna is a vertical fan-shaped beam. Combining this antenna pattern with a circular scan technique creates a radar system well suited for long-range search and target acquisition.

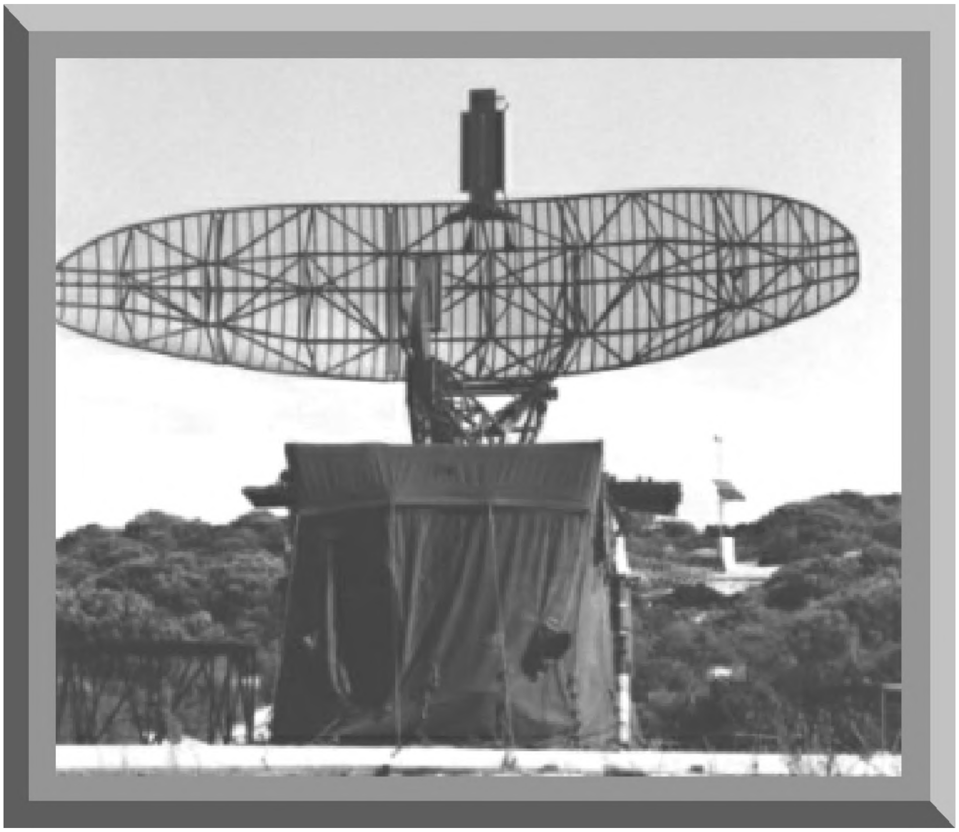

Figure 6-2. Parabolic Cylinder Antenna

c. Elongating the vertical dimensions of the parabola creates a radar antenna that generates a horizontal fan-shaped beam with a small vertical dimension (Figure 6-3). This type of antenna is generally used in height-finding radar systems.

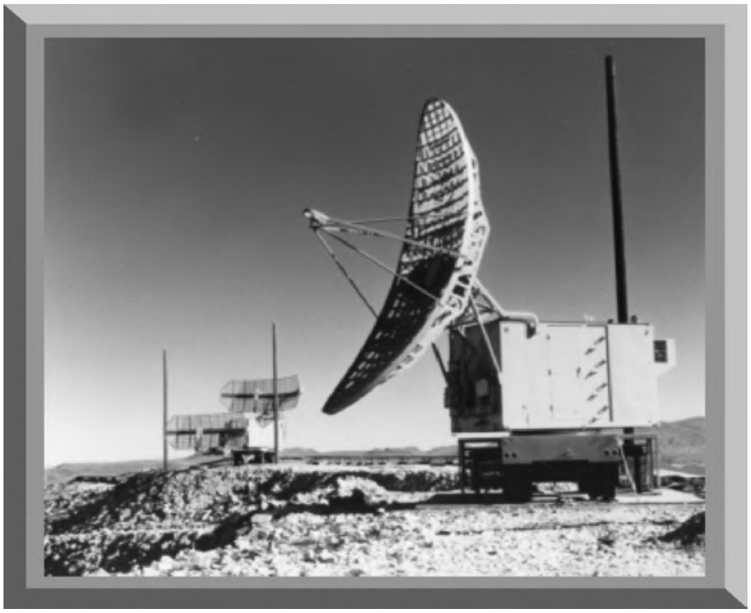

Figure 6-3. Height-Finder Parabolic Antenna

d. Another variation of the basic parabolic antenna includes using an array of multiple feeds instead of a single feed (Figure 6-4). This type of parabolic antenna can produce multiple radar beams, either symmetrical or asymmetrical, depending on the angle and spacing of the individual feeds.

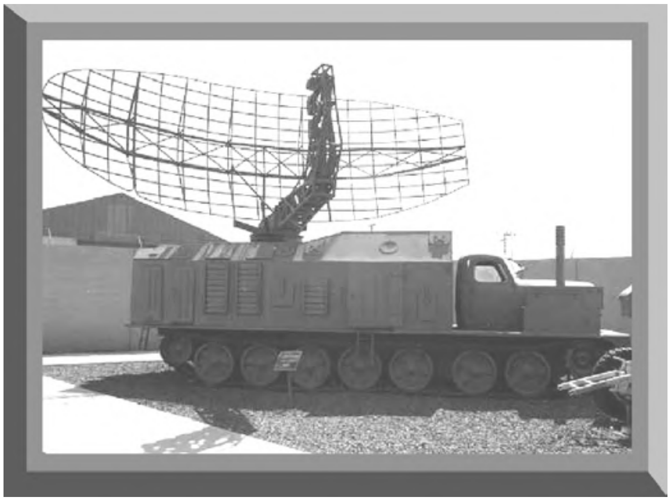

Figure 6-4. Multiple-Feed Parabolic Antenna

3. CASSEGRAIN ANTENNA

A Cassegrain antenna uses a two-reflector system to generate and focus a radar beam (Figure 6-5). The primary reflector uses a parabolic contour, and the secondary reflector, or subreflector, has a hyperbolic contour. The antenna feed

is located at one of the two foci of the hyperbola. Radar energy from the transmitter is reflected from the subreflector to the primary reflector to focus the radar beam. Radar energy returning from a target is collected by the primary reflector and reflected as a convergent beam to the subreflector. The radar energy is rereflected by the subreflector, converging at the position of the antenna feed. The larger the subreflector, the closer it can be to the primary reflector. This reduces the axial dimensions of the radar but increases aperture blockage due to the subreflector. A small subreflector reduces aperture blockage, but it must be positioned at a greater distance from the primary reflector.

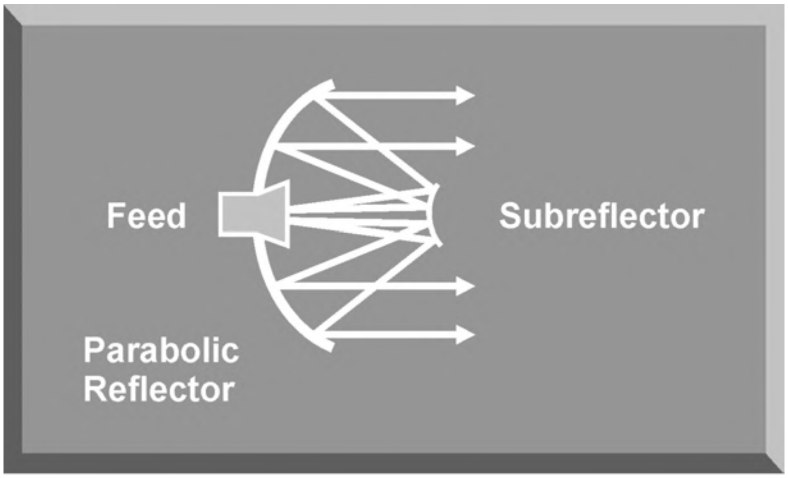
Feed
Subreflector
Parabolic Reflector

Figure 6-5. Cassegrain Antenna

a. To reduce the aperture blockage by the subreflector and to provide a method to rapidly scan the radar beam, the flat plate Cassegrain antenna was developed. This type of Cassegrain antenna is depicted in Figure 6-6.

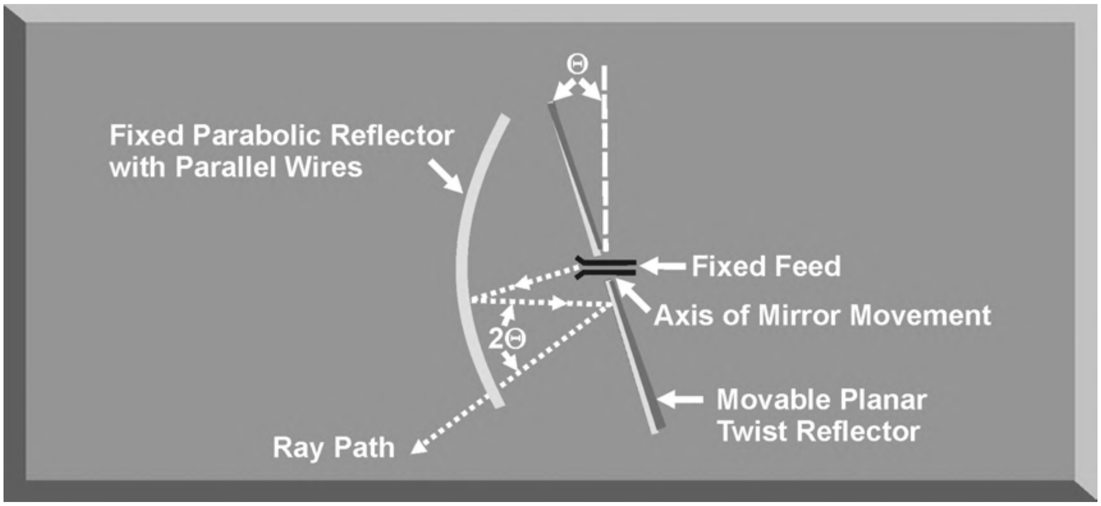
Θ
Fixed Parabolic Reflector with Parallel Wires
Fixed Feed
Axis of Mirror Movement
2Θ
Movable Planar Twist Reflector
Ray Path

Figure 6-6. Flat Plate Cassegrain Antenna

The fixed parabolic reflector is made up of parallel wires spaced less than a half wavelength apart and supported by a low-loss dielectric material. This makes the fixed parabolic reflector polarization sensitive. It will completely reflect one type of linear polarization and be transparent to the orthogonal polarization. The fixed antenna feed, in the middle of the moveable mirror, transmits a radar signal polarized to be reflected by the parabolic reflector. The moveable mirror is constructed as a twist reflector that changes the polarization of the radar signal by 90°. The signal from the feed is reflected by the parabolic reflector to the mirror, which rotates the polarization 90°. This rotation makes the transmitted signal transparent to the parabolic reflector, and the signal passes through with minimal attenuation. The radar beam can be scanned over a wide area by rotating the moveable mirror. A deflection of the mirror by the angle θ results in the beam scanning through an angle of 2θ.

b. The geometry of the Cassegrain antenna is especially well suited for monopulse tracking radar applications. Unlike the parabolic antenna, the complex feed assembly required for a monopulse radar can be placed behind the reflector to avoid aperture blocking.

4. PHASED ARRAY ANTENNA

The phased array radar is a product of the application of computer and digital technologies to the field of radar design. A phased array antenna is a complex arrangement of many individual transmitting and receiving elements in a particular pattern. A phased array antenna can, in effect, radiate more than one beam from the antenna by using a computer to rapidly and independently control groups of these individual elements. Multiple beams and computer processing of radar returns give the phased array radar the ability to track-while-scanning and engage multiple targets simultaneously. Figure 6-7 is a view of the phased array radar antenna.

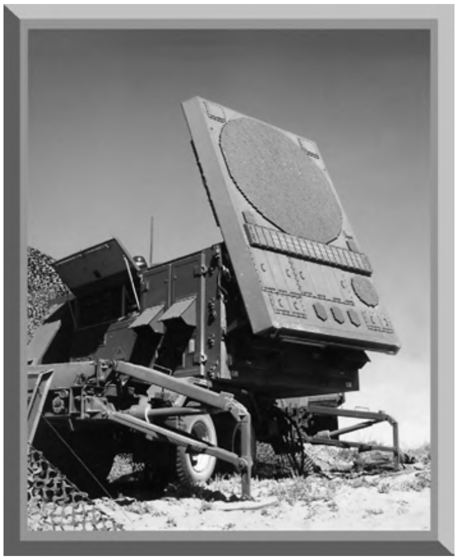

Figure 6-7. Phased Array Antenna

a. A phased array radar uses the principle of radar phase to control the individual transmitting and receiving elements. When two transmitted frequencies are in-phase, their amplitudes add together, and the radiated energy is doubled. When two transmitted frequencies are out-of-phase, they cancel each other. Phased array radars use this principle to control the shape of the transmitted radar beam (Figure 6-8).

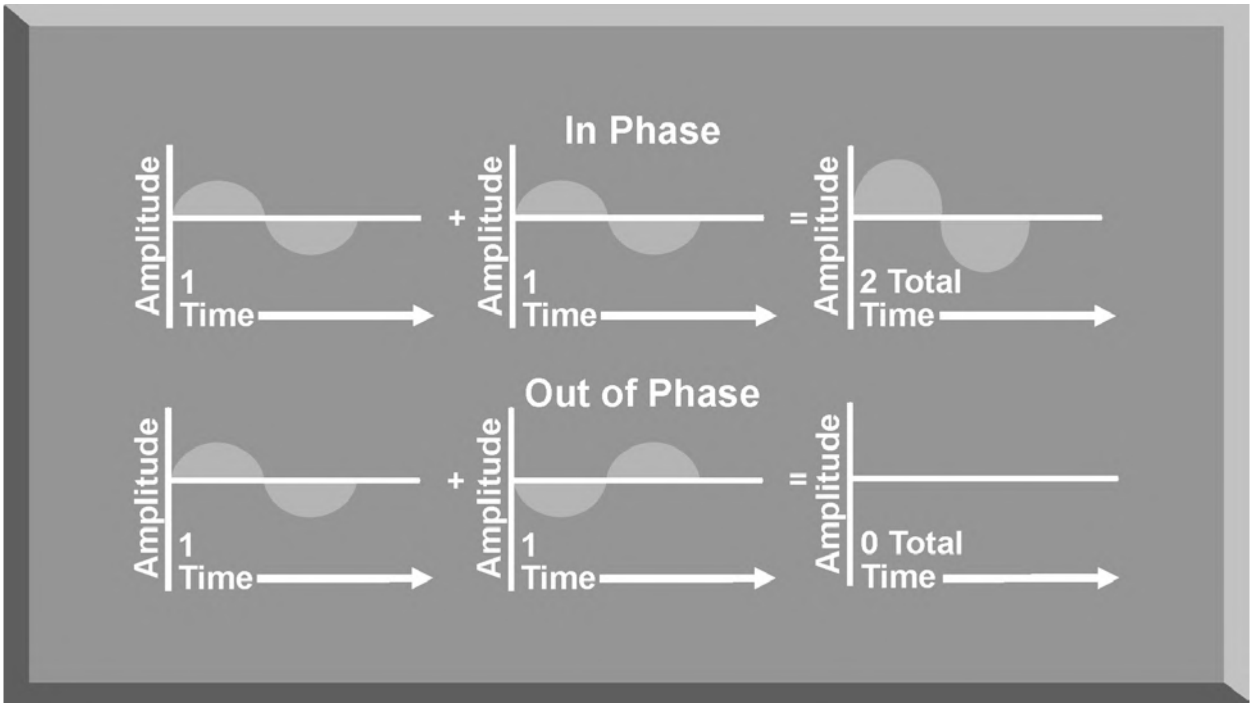

Figure 6-8. Phase Relationships

b. Phase relationships and antenna element spacing determine the orientation of the transmitted beam. In Figure 6-9, antenna elements A and B are separated by one-half wavelength and are radiating in-phase, that is, when one is at the positive peak, the other is also at a positive peak. Since the elements are one-half wavelength apart, when the positive peak radiated by A reaches B, B will be radiating a negative peak. As the peaks propagate along the X axis, they will cancel each other out. The total radiated power along that axis will be zero. Along the Y axis, however, the positive peaks from A will add to the positive peaks from B, causing the total radiation along this axis to be at its maximum value. This type of array is called a “broadside array” because most of the radiation is in the direction that is broadside to the line of the antenna array.

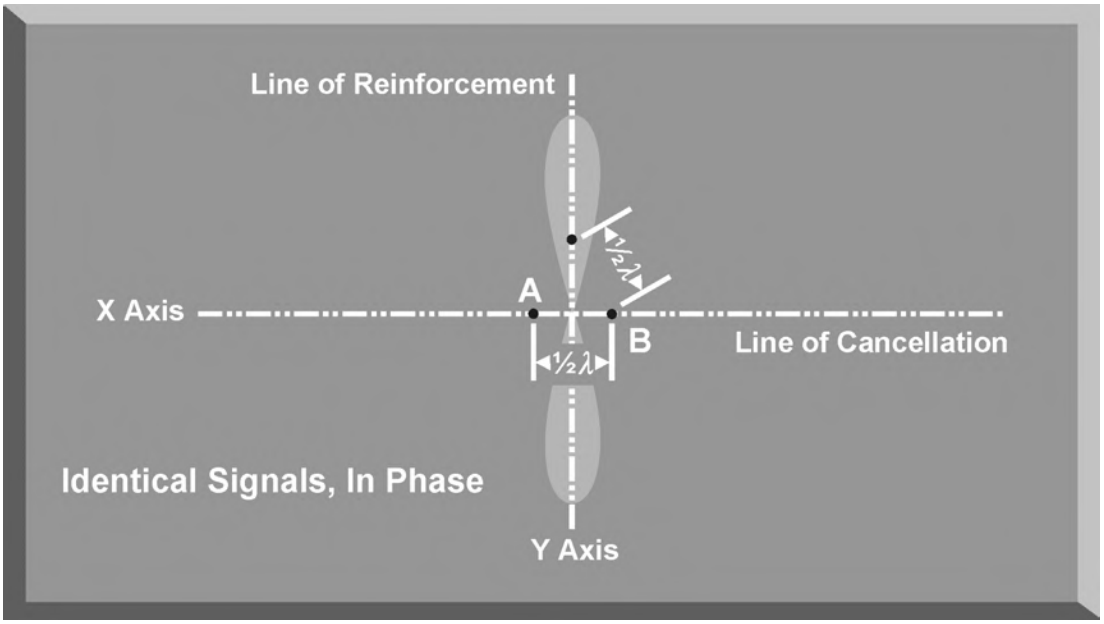
Line of Reinforcement
½λ
A
X Axis
B
Line of Cancellation
½λ
Identical Signals, In Phase
Y Axis

Figure 6-9. Broadside Array

c. If the same antenna elements are fed out-of-phase, the principal direction of radiation will be along the axis of the antenna elements. In Figure 6-10, when the positive peak from A arrives at B, B is now positive also. These energies interact to strengthen the energy being radiated from the ends of the array. Meanwhile, when the positive peak from A, radiating along the Y axis, meets the negative peak from B, they are canceled. This type of array is called an "end-fire array."

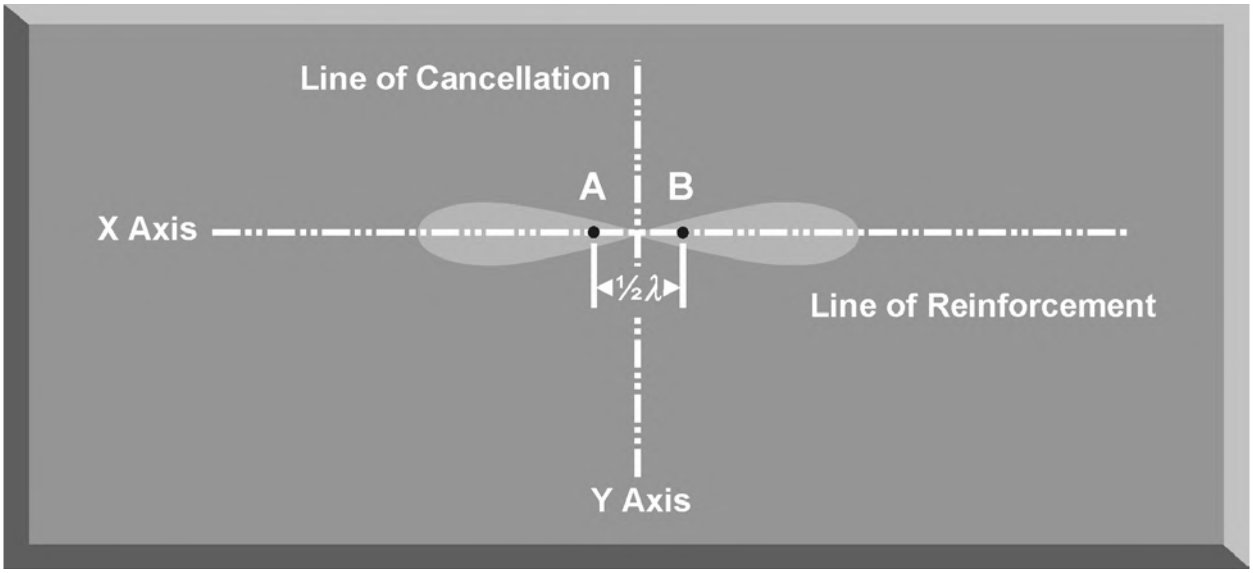
Line of Cancellation
A
B
X Axis
½λ
Line of Reinforcement
Y Axis

Figure 6-10. End-Fire Array

d. The computer controlling the phase of the signal delivered to each transmitting and receiving element of a phased array antenna controls the direction and shape of the radiated beam (Figure 6-11). By shifting the phase of the signals between 0° and 180°, the beam sweeps. This is the basic means of producing an antenna scan. In addition, the amplitude, or power, of the signal applied to each element can be varied to control the sidelobes. This alters the shape of the beam which affects the range capability and angular resolution of the radar.

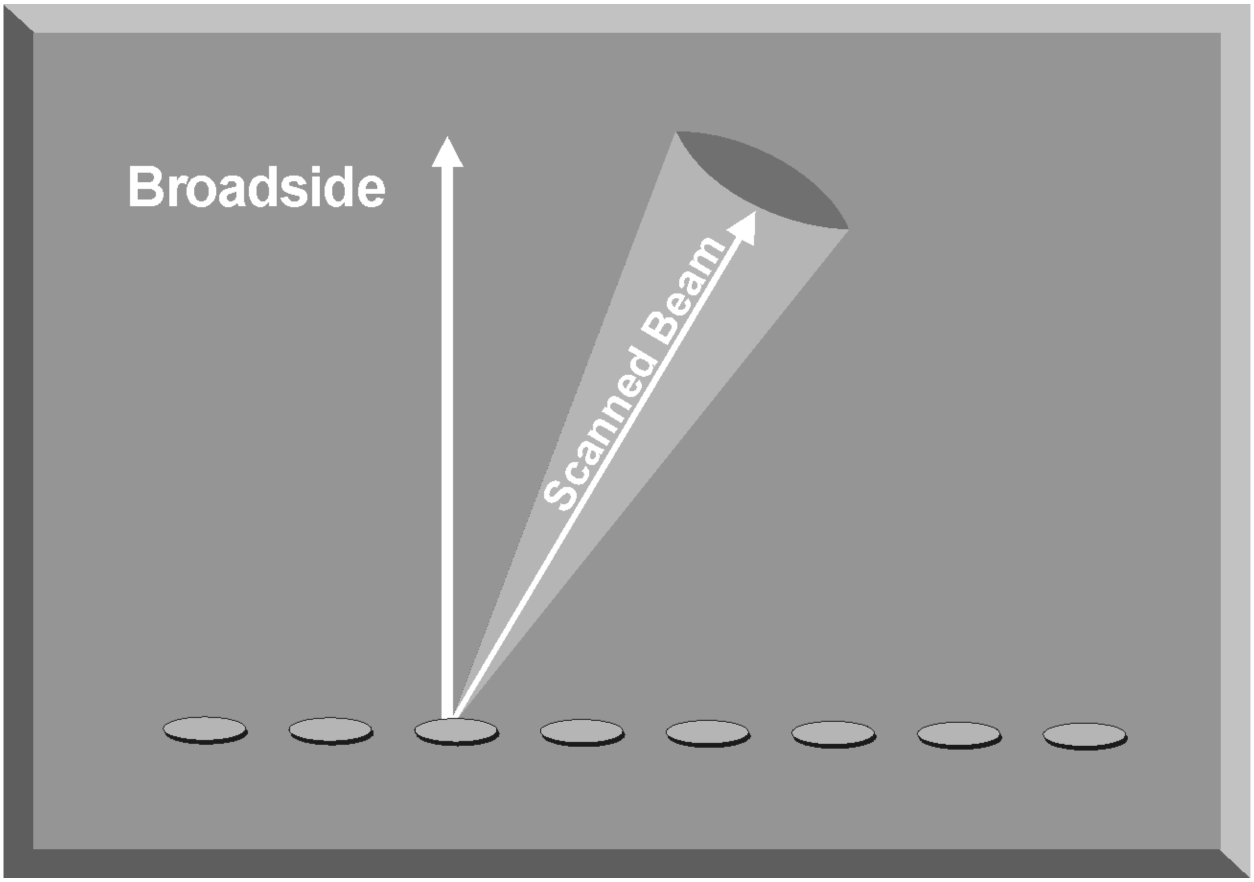

Figure 6-11. Phased Array Antenna Scan

e. Figure 6-12 depicts a variation of the phased array antenna, known as a planar array antenna. A planar array antenna uses transmit and receive elements in a linear array, but, unlike the phased array radar, the elements are smaller and are placed on a movable flat plate. The ability to simultaneously track several targets is one advantage of this type of radar.

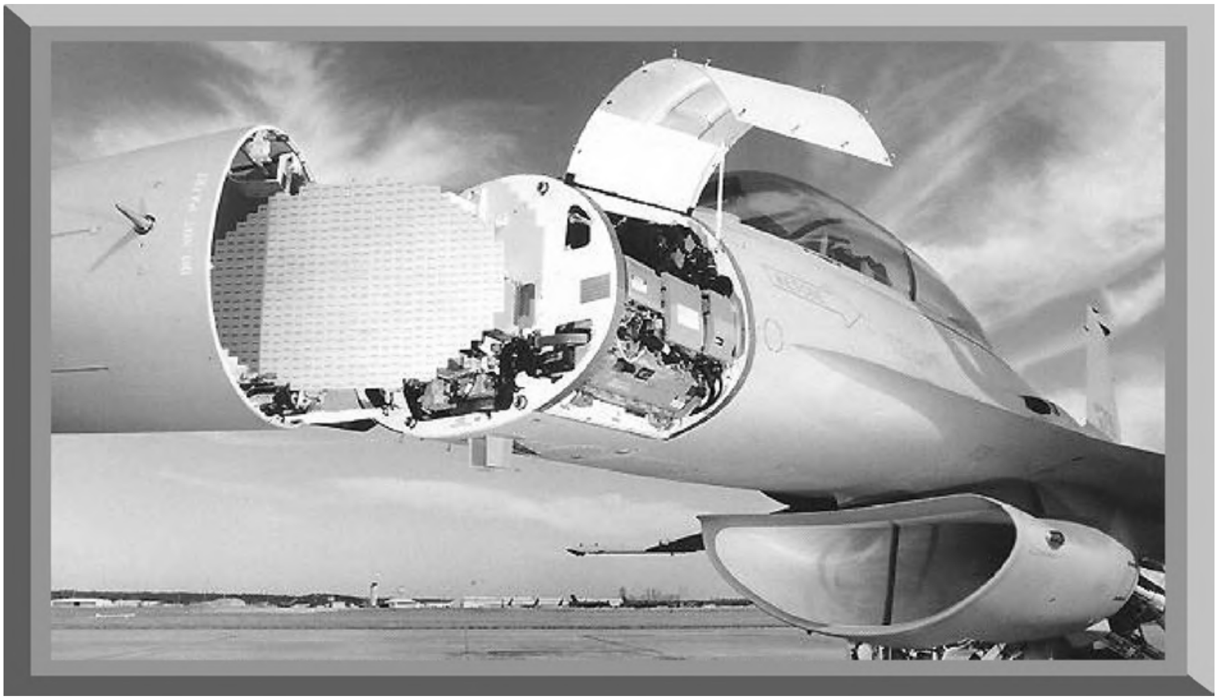

Figure 6-12. Planar Array Antenna

5. ANTENNA GAIN

The most important characteristic of any type of antenna is antenna gain. Antenna gain is a measure of the ability of an antenna to concentrate energy in the desired direction. Antenna gain should not be confused with receiver gain, which is designed to control the sensitivity of the receiver section of a radar system. There are two types of antenna gain: directive and power.

a. The directive gain of a transmitting antenna is the measure of signal intensity radiated in a particular direction. Directive gain is dependent on the shape of the radiation pattern of a specific radar antenna. The directive gain does not take into account the dissipative losses of the antenna. Directive gain is computed using Equation 6-1.

$$G_D \text{ (Directive Gain)} = \frac{\text{Maximum Radiation Intensity (Desired Direction)}}{\text{Average Radiation Intensity}}$$

Equation 6-1. Directive Gain

b. The power gain does include the antenna dissipative losses and is computed using Equation 6-2.

$$\text{G (Power Gain)} = \frac{\text{Maximum Radiation Intensity (Practical Antenna)}}{\text{Radiation Intensity of an Isotropic Antenna}}$$

Equation 6-2. Power Gain

c. The term isotropic antenna describes a theoretical spherical antenna that radiates with equal intensity in all directions. This results in a spherical radiation pattern. The power density for any point on an isotropic antenna is the radiation intensity and can be calculated by dividing the total power transmitted (P_T) by the total surface area of the sphere, as shown in Equation 6-3.

$$\text{Power Density (Isotropic Antenna)} = \frac{P_T\ \text{(Watts)}}{4\pi r^2\ (\text{Centimeters}^2)}$$

Equation 6-3. Power Density for an Isotropic Antenna

d. The radiation pattern of an isotropic, or spherical, antenna would provide neither azimuth or elevation resolution and would be unusable for radar applications. To provide azimuth and elevation resolution, a practical antenna must focus the radar energy. The power density of a practical antenna differs from the isotropic antenna only in terms of antenna gain (G). Solving Equation 6-3 for the power density of a practical antenna yields Equation 6-4.

$$\text{Power Density (Practical Antenna)} = \frac{P_T\, G}{4\pi r^2}$$

Equation 6-4. Power Density for a Practical Antenna

e. The actual power gain (G) of a practical antenna can be calculated by using Equation 6-5.

$$G = \frac{4\pi A_e}{\lambda}$$

A_e = effective area of aperture
λ = wavelength of the radar

NOTE: The effective area of aperture (A_e) is the effective antenna area presented to the returning radar echo.

Equation 6-5. Power Gain of a Practical Antenna

6. POWER DENSITY

The power density and gain of an antenna are a function of the antenna pattern of a radar system. Figures 6-13 and 6-14 illustrate the antenna pattern of a typical parabolic antenna. Most of the power density of the radar is concentrated in the main beam. However, since the radar is not a perfect reflector, some radar energy is transmitted in the sidelobes. In addition, there is spillover radiation due to the energy radiated by the feed that is not intercepted by the reflector. Finally, the radar has a back lobe caused by diffraction effects of the reflector and direct signal leakage. Sidelobes and backlobes are all undesirable radiations that adversely affect the maximum radar range and increases the vulnerability of the radar to certain jamming techniques.

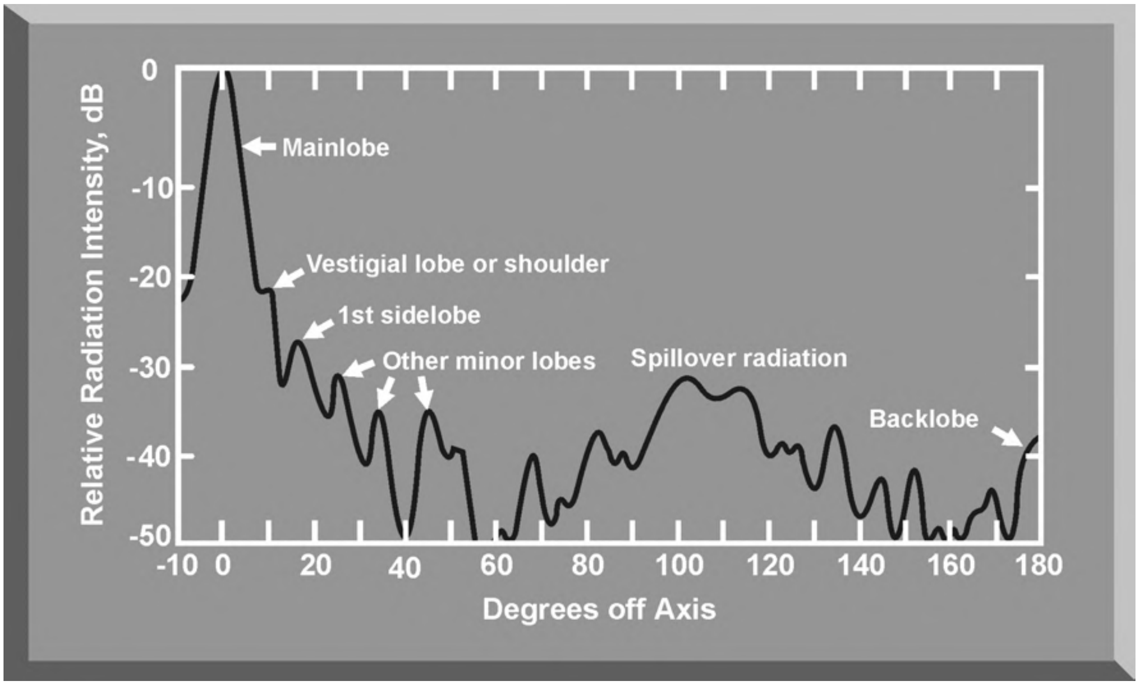

Figure 6-13. Radiation Pattern for a Parabolic Antenna

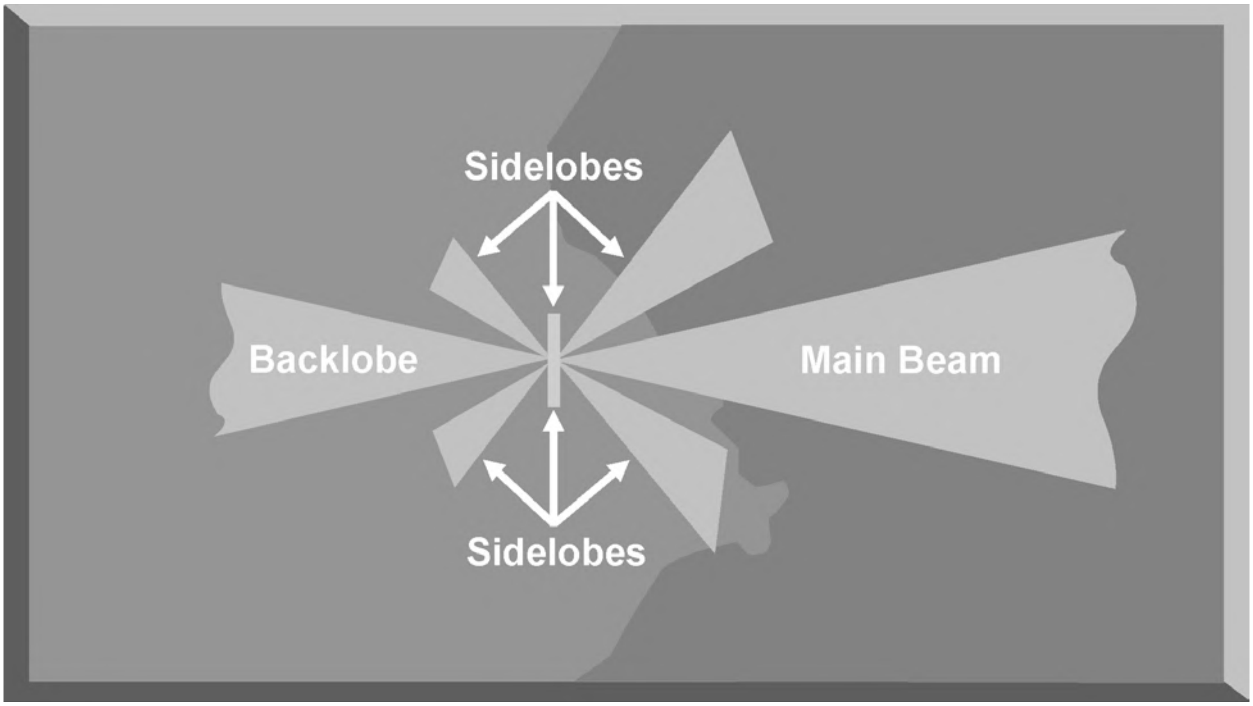

Figure 6-14. Radar Antenna Pattern

a. All radars have a primary main beam, which is where the radar has the most power and where target detection usually occurs. The dimensions of this main beam are highly dependent on the design of the antenna.

b. Besides the main beam, all radars have what is called a backlobe. This lobe is directly opposite to the location of the main beam. The sensitivity and signal strength associated with the backlobe is significantly less than that in the main beam.

c. Sidelobes add another dimension to the radar pattern. As with the backlobe, sidelobes do not have the signal strength or sensitivity associated with the main beam. Normally, the sensitivity associated with the sidelobes is 40-50 decibels (dBs) less than the main beam. The radar signal weakness in the backlobe and sidelobes of the main beam make these areas of the radar signal vulnerable to jamming. It is much easier to introduce jamming into these areas because of the reduced jamming-to-signal ratio needed to be effective. It is difficult for jamming to be effective in the main beam because the radar signal is very powerful in that region.

7. CIRCULAR SCAN

A circular scanning radar uses an antenna system that continuously scans 360° in azimuth (Figure 6-15). The time required for the antenna to sweep one complete 360° cycle is called the scan rate. Scan duration is the number of “hits per scan,”

or the number of pulses, reflected by a target as the radar beam crosses it during one full scan. Most pulse radars require 15 to 20 hits per scan to obtain sufficient information to display a target. The factors that determine the number of hits per scan the radar receives include pulse repetition frequency (PRF), antenna beamwidth, and scan duration.

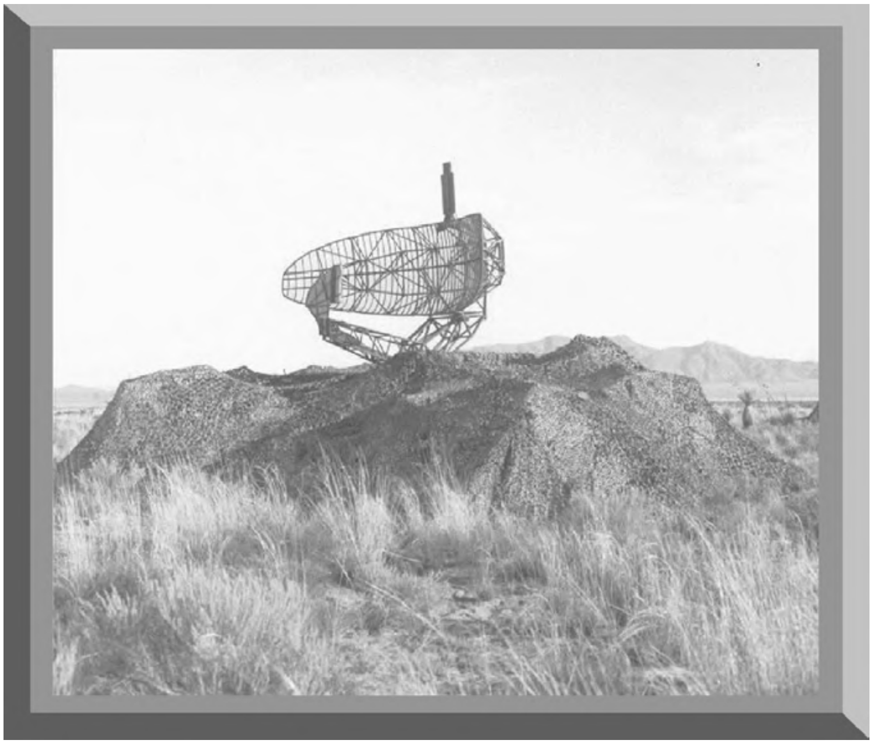

Figure 6-15. Circular Scan Radar

a. Circular scan radars provide accurate target range and azimuth information. This makes these radars ideal for the roles of early warning and initial target acquisition. To accomplish these missions, the antenna generates a fan beam that has a large vertical beamwidth and a small horizontal beamwidth. Since elevation information will normally be provided by height finder radars, the size of the vertical beamwidth is not a limitation. This antenna scan allows the radar to scan large volumes of airspace for early target detection. Since early detection is the primary goal of early warning radars, accurate altitude and azimuth resolution are secondary considerations.

b. Circular scan radars designed for early warning transmit a radar signal with a low PRF. A low PRF allows sufficient time for the radar pulse to travel long distances, and return, before another pulse is transmitted. This gives the radar system a long, unambiguous range capability. Circular scan radars with low PRFs generally use long pulse widths in order to increase their average power and long-range detection capability. The scan durations of early warning radars are relatively long to provide the required "hits per scan" for long-range target

detection. The plan position indicator (PPI) scope display is normally used with a circular scan radar (Figure 6-16).

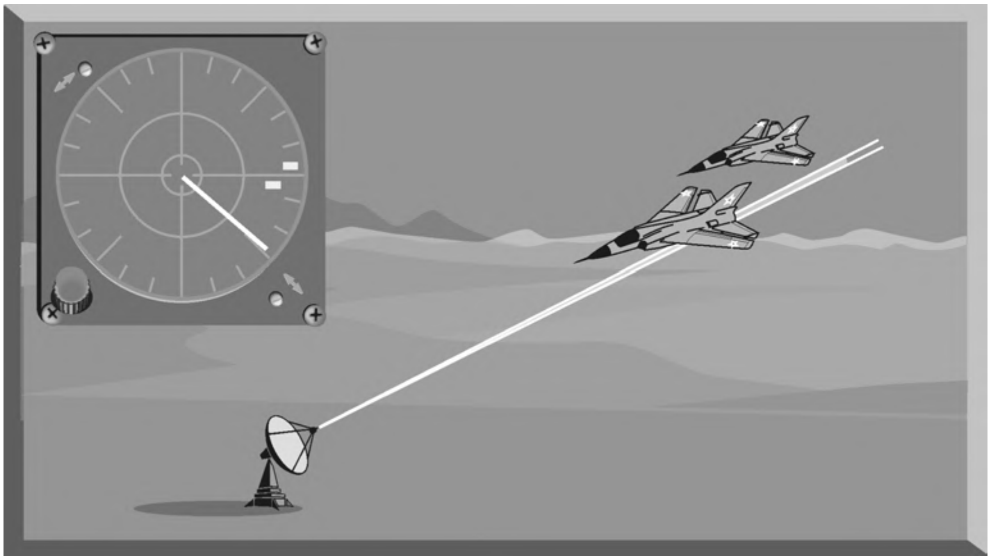

Figure 6-16. PPI Scope Display

c. In order to provide coverage for a large volume of airspace, the beamwidth associated with a circular scan radar is relatively wide. This wide beamwidth, coupled with the long pulse width and low PRF, gives the circular scan radar a large resolution cell, especially at long ranges (Figure 6-17). This limitation can be exploited to mask force size and composition. However, as range decreases, the dimensions of the resolution cell decrease, and a circular scan radar will begin to break out target formations.

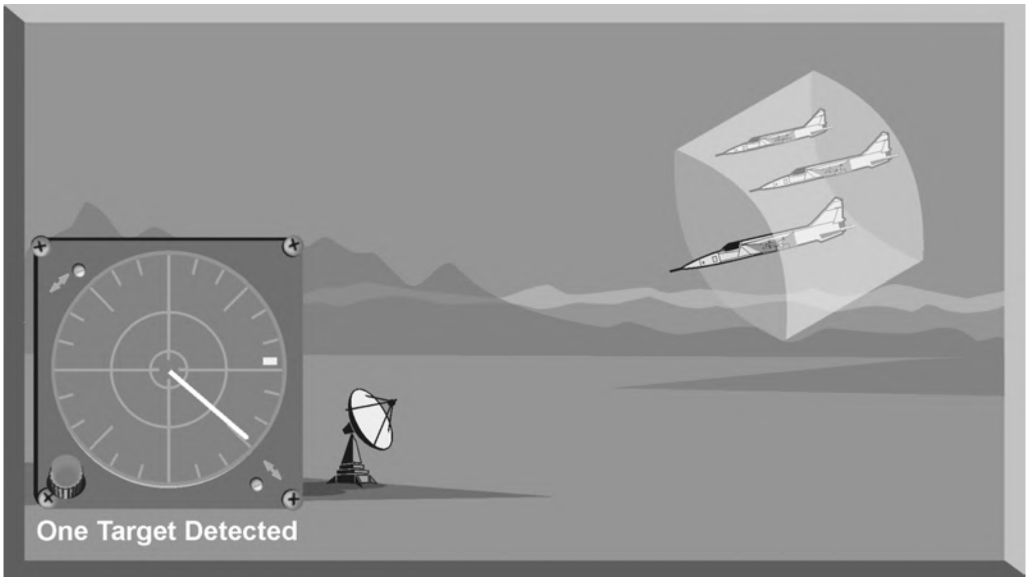

Figure 6-17. Resolution Cell

d. Circular scan radars provide range and azimuth information for both early warning and acquisition roles. Modified circular scan radars that can also provide elevation information may be used for ground control intercept (GCI) roles. Two modified circular scan radars that determine range, azimuth, and elevation are the V-beam and the stacked beam.

(1) The V-beam radar transmits two fan-shaped beams that are swept together (Figure 6-18). A vertical beam provides range and azimuth information. A second beam, rotated at some convenient angle, provides a measure of the altitude of the target.

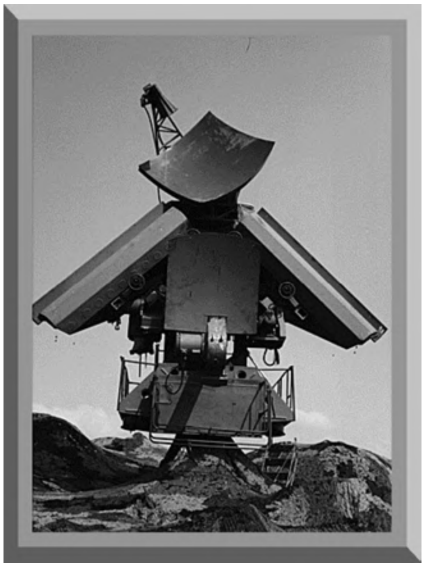

Figure 6-18. V-Beam Radar

(2) A stacked beam radar (Figure 6-19) employs a vertical stack of fixed elevation "pencil" beams which rotate 360°. Elevation information is obtained by noting which beam contains the target return. Range and azimuth information is determined in the same manner as in an early warning radar.

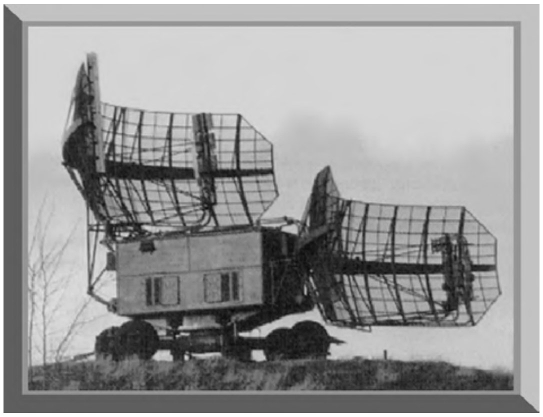

Figure 6-19. Stacked Beam Radar

8. LINEAR SCAN

Linear scan is a method used by some radar systems to sweep a narrow radar beam in a set pattern to cover a large volume of airspace. Linear scans can be oriented in a vertical direction for height finder radars or in a horizontal direction, or raster, for acquisition and target tracking radars. A unidirectional linear radar scans in a single direction then begins its sweep all over again (Figure 6-20). Generally, linear scans offer excellent single-axis coverage, and the narrow beam offers enhanced azimuth and elevation resolution.

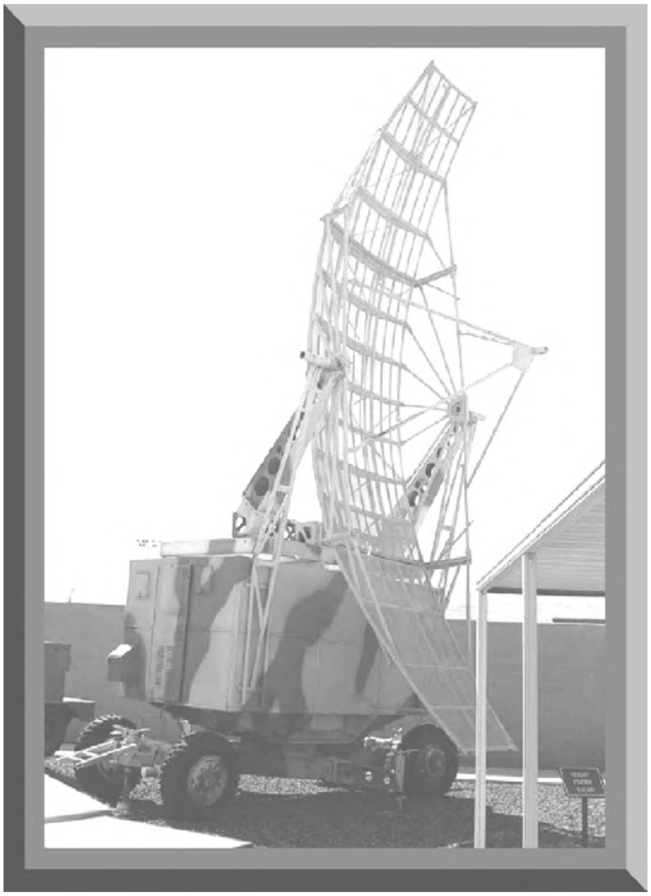

Figure 6-20. Unidirectional Linear Scan

9. UNIDIRECTIONAL SCAN

A helical scan is a unidirectional scan pattern that allows a "pencil" beam to search a 360° pattern. In Figure 6-21, the antenna sweeps a 360° sector in a clockwise direction. After each complete revolution, the antenna elevation is increased. This scan pattern is repeated for a specified number of revolutions, in this case, three, 360° sweeps. At the end of the scan pattern, the antenna elevation is reset to the initial elevation and the scan is repeated. A helical scan pattern is commonly used as a target acquisition mode for radar systems with narrow vertical and horizontal beamwidths.

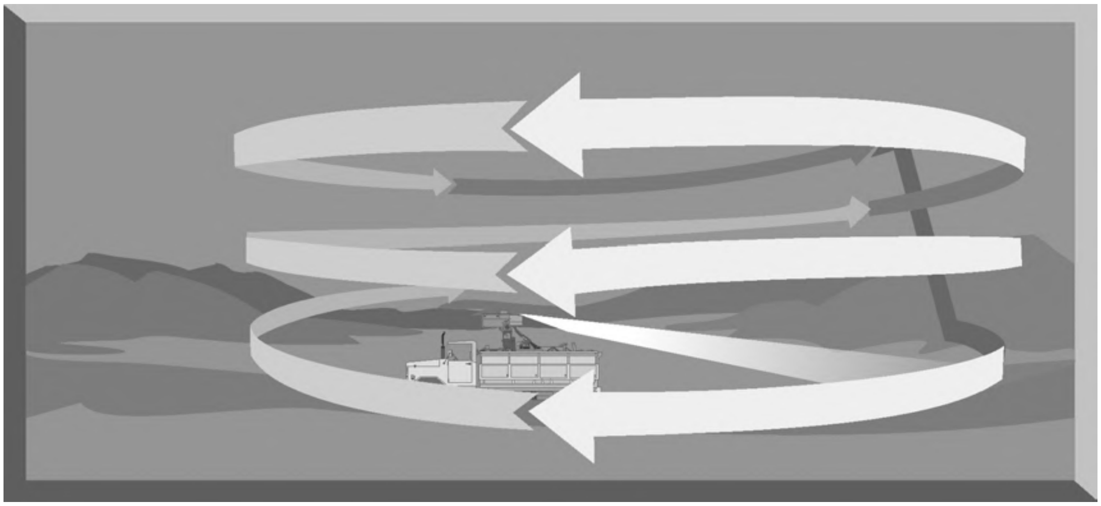

Figure 6-21. Helical Scan

10. BIDIRECTIONAL SCAN

A bidirectional linear scan, such as a raster scan, sweeps both horizontally and vertically (Figure 6-22). A raster scan uses a thin beam to cover a rectangular area by horizontally sweeping the area. The angle of elevation is incrementally stepped up or down with each horizontal sweep of the desired sector. After the sector has been covered, the angle of elevation is reset to the original value and the process is repeated. The number of raster bars is set by the number of horizontal sweeps in the basic raster pattern. Figure 6-22 shows a four-bar raster scan, which is normally associated with an airborne interceptor (AI) radar.

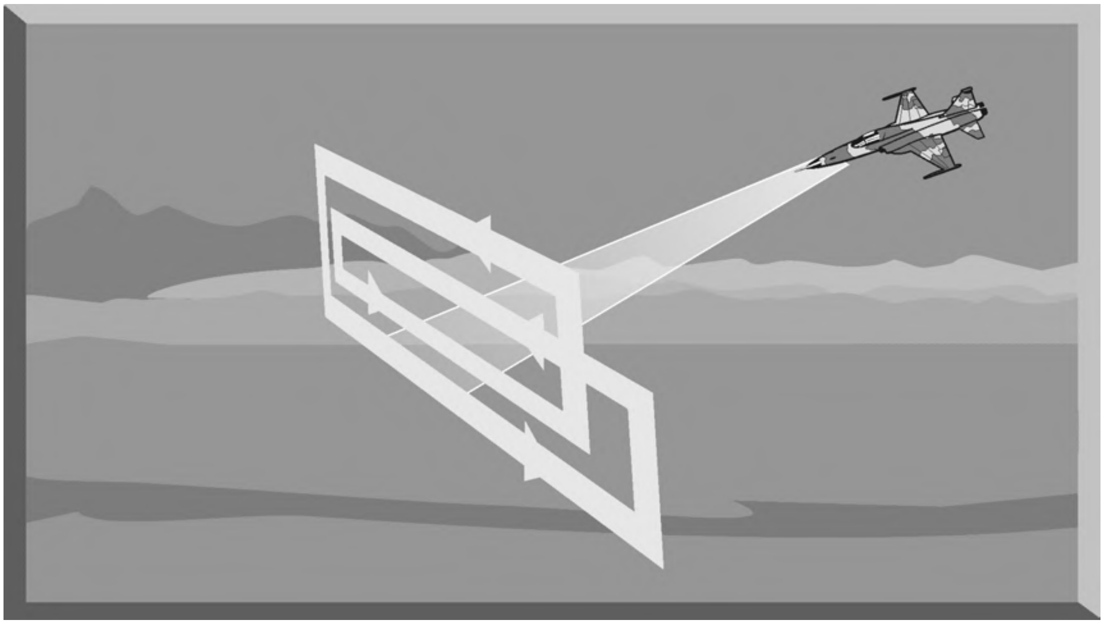

Figure 6-22. Raster Scan

11. CONICAL SCAN

A conical scan, or conscan, radar is generally used for precision target tracking. A conical scan radar (Figure 6-23) employs a pencil beam of radar energy that is continuously rotated around the target. This circular rotation of a pencil beam generates a cone-shaped scan pattern with the apex of the cone located at the antenna. Thus, the name conical scan.

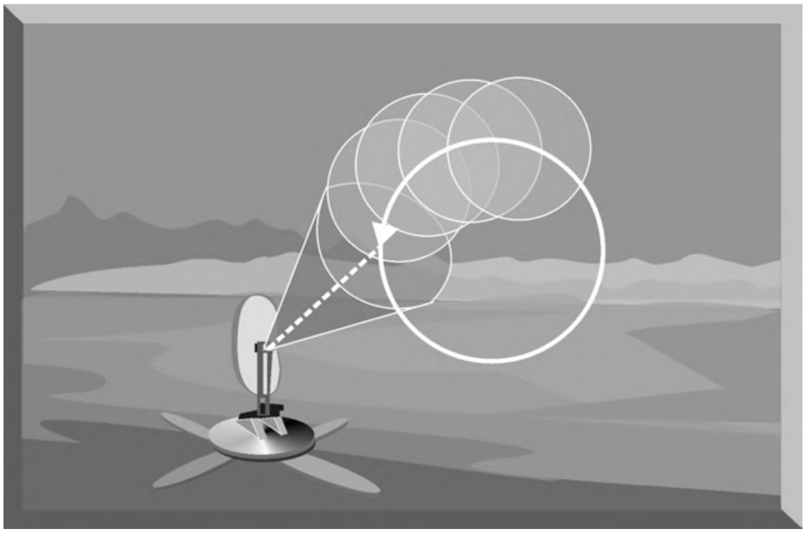

Figure 6-23. Conical Scan

a. As the pencil beam rotates, the circular scan patterns overlap in the center. This creates a central tracking area that has a much smaller effective beamwidth than the rotating pencil beam (Figure 6-24). This results in a very precise tracking solution.

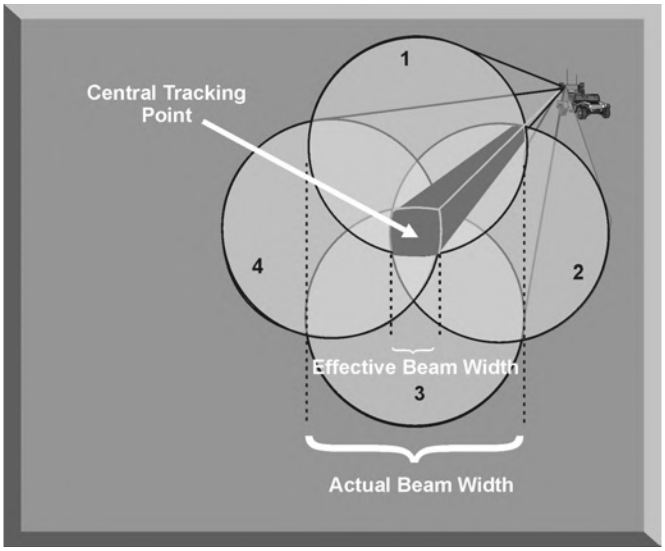

Figure 6-24. Conical Scan Tracking Area

b. Since conical scan radars are designed for precision target tracking, these radars normally operate at high frequencies, high PRF, narrow pulse widths, and narrow beamwidths. The rotation rate of the pencil beam can exceed 1,800 revolutions per minute. This means that both azimuth and elevation data can be updated about 30 times per second.

c. The combination of conical scan and raster scan is called a Palmer-raster scan (Figure 6-25). A Palmer-raster scan uses a thin beam, employing a conical scan searching pattern, for a specific sector of airspace. With each sweep of the sector, the angle of elevation is incrementally stepped up or down. After the vertical sector has been covered, the angle of elevation is set at the original elevation and the process is repeated. The number of bars is determined by the number of vertical search scans.

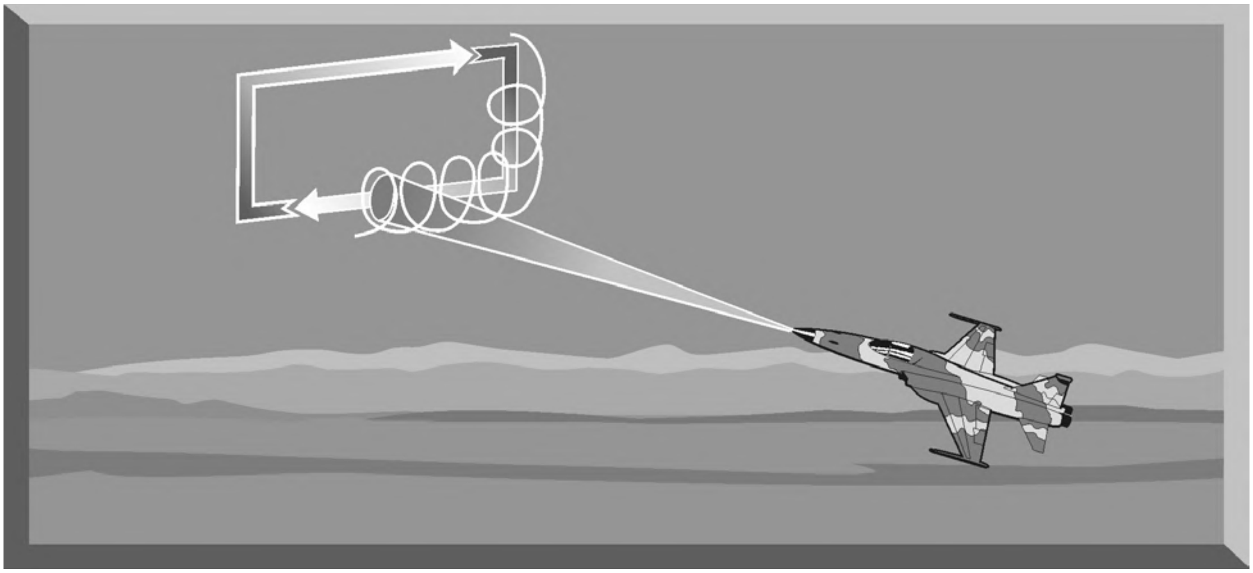

Figure 6-25. Palmer-Raster Scan

d. The combination of a conical scan and a circular scan is called a Palmer scan. Palmer scans incorporate a circular scanning antenna to search the entire horizon while simultaneously performing a conical scan. If the radar antenna is also performing a unidirectional altitude search in conjunction with this scan, it is employing a Palmer-helical scan (Figure 6-26).

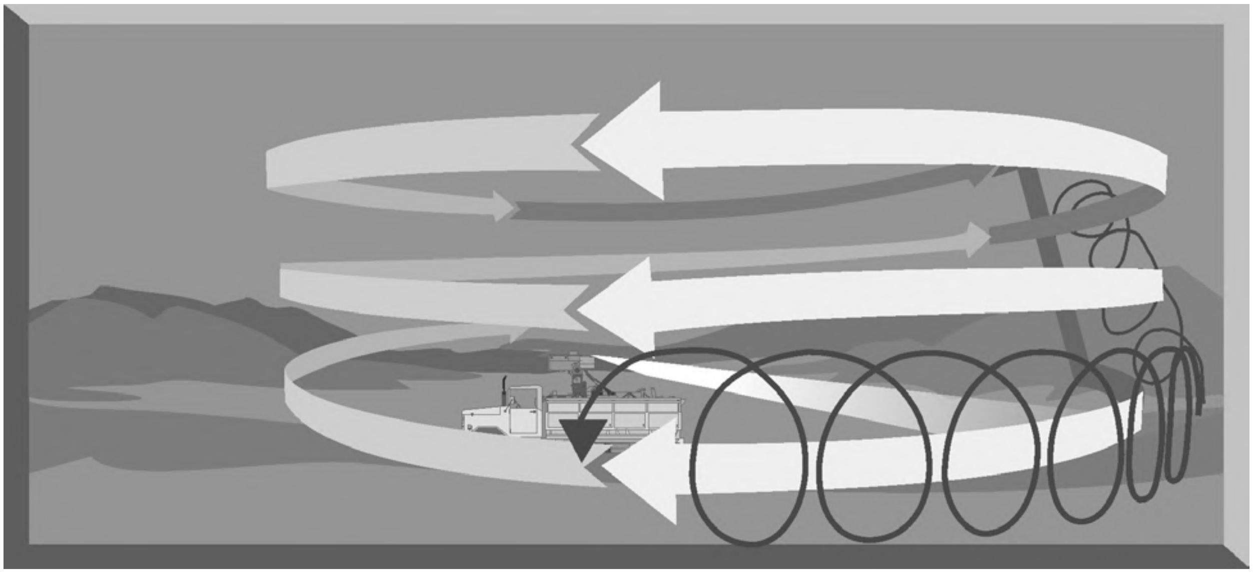

Figure 6-26. Palmer-Helical Scan

12. TRACK-WHILE-SCAN

A track-while-scan (TWS) system uses a technique that allows a radar to track one or more targets while scanning for others. Radar systems with a TWS capability must be able to generate two or more distinct radar beams.

a. A conventional TWS radar employs two antennas that work with each other to perform the scan function (Figure 6-27). Each antenna produces a separate unidirectional beam. Each beam is transmitted at a different frequency. The vertical antenna generates a beam employing a vertical sector scan similar to a height finder radar except the beamwidth is narrower and it scans at a higher rate. The horizontal antenna generates an identical beam employing a horizontal sector scan at a different frequency. The track function is accomplished in the area where the two beams pass through each other. A target that is within this center area is tracked, and positional information on range, elevation, and azimuth is updated each time the beams sweep through the area.

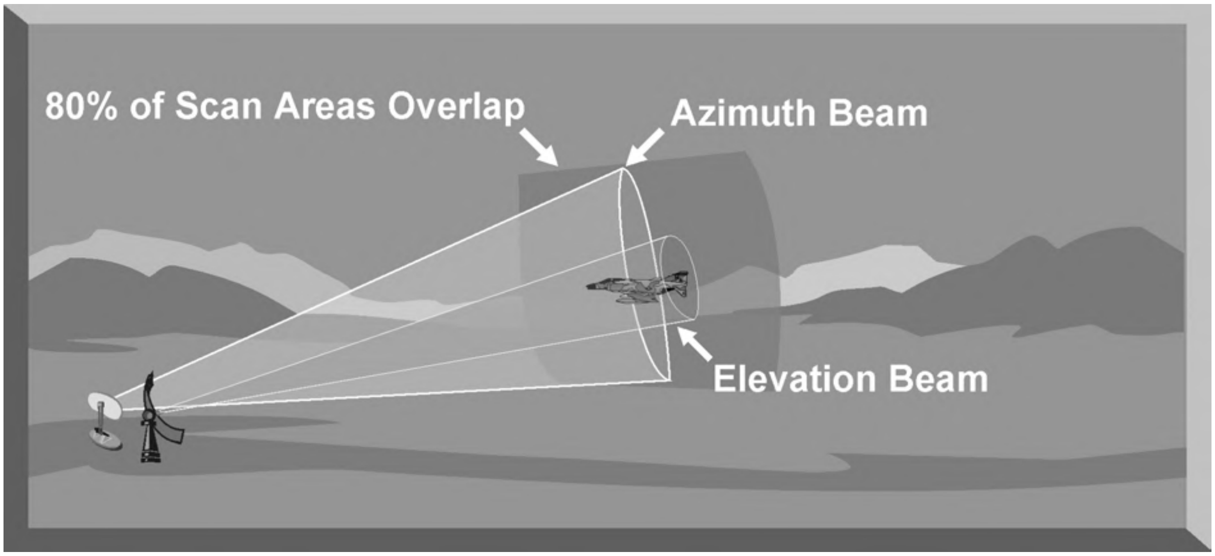

Figure 6-27. Conventional TWS Radar

b. The phased array radar is a product of the application of computer and digital technologies to the field of radar design. A phased array is a complex arrangement of many individual transmitting and receiving elements in a particular pattern (Figure 6-28). Common arrays include linear, planar, curved, and conformal, with linear being the most common. By using a computer to rapidly and independently control groups of these individual elements, a phased array antenna can, in effect, radiate more than one beam from the antenna. Multiple beams and computer processing of radar returns give the phased array radar the ability to perform the TWS function. The most common employment of the TWS capability of the phased array radar is in the air-to-air arena.

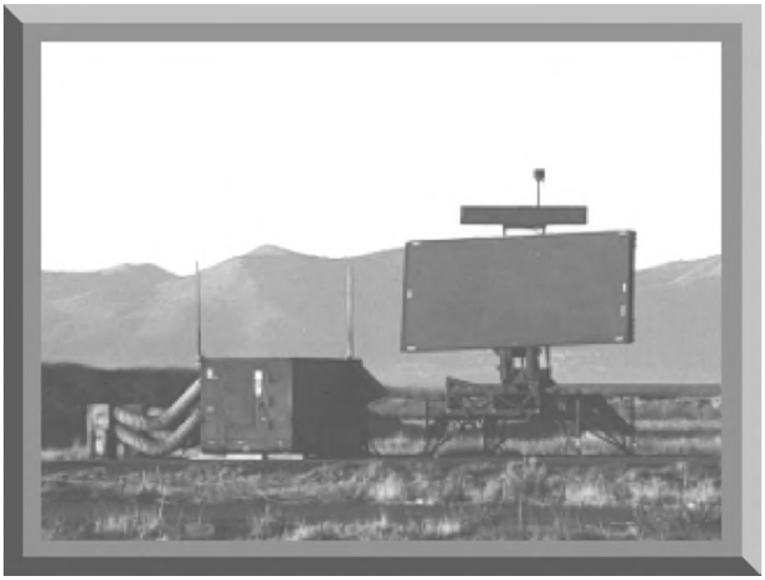

Figure 6-28. Phased Array Antenna

(1) The number of individual transmitting and receiving elements is limited by the size of the radar antenna. The number of targets a phased array radar can track is limited by the number of independent beams the antenna can generate. Many phased array radars, especially air-to-air radars, do not track and scan simultaneously, but rapidly switch between the two modes to overcome this limitation (Figure 6-29).

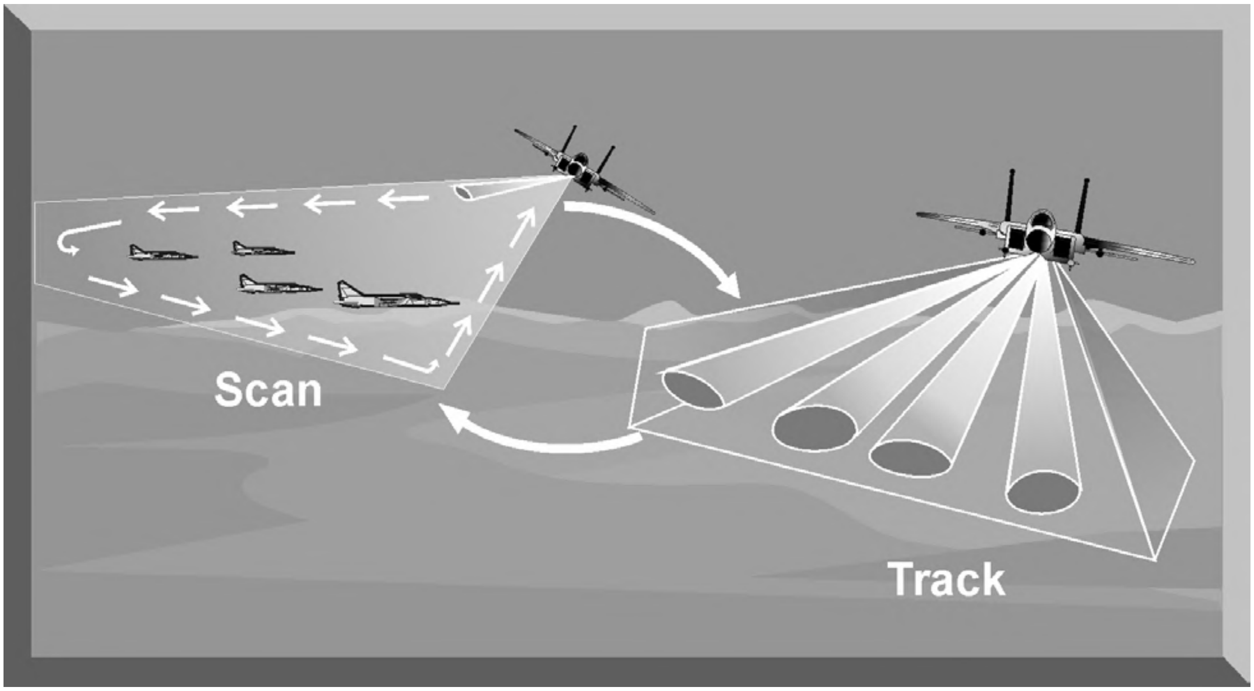

Figure 6-29. Phased Array TWS

(2) Modern TWS radars employ computer signal processing and complex computer algorithms to simplify the problem of target correlation (Figure 6-30). An air-to-air radar typically uses a raster scan to search a volume of airspace. In the search mode, the radar simply presents all targets detected in this airspace to the pilot on his radar display. In the TWS mode, the radar employs computer processing to figure out target correlation and update target information. This is done automatically, and the results are presented on the display.

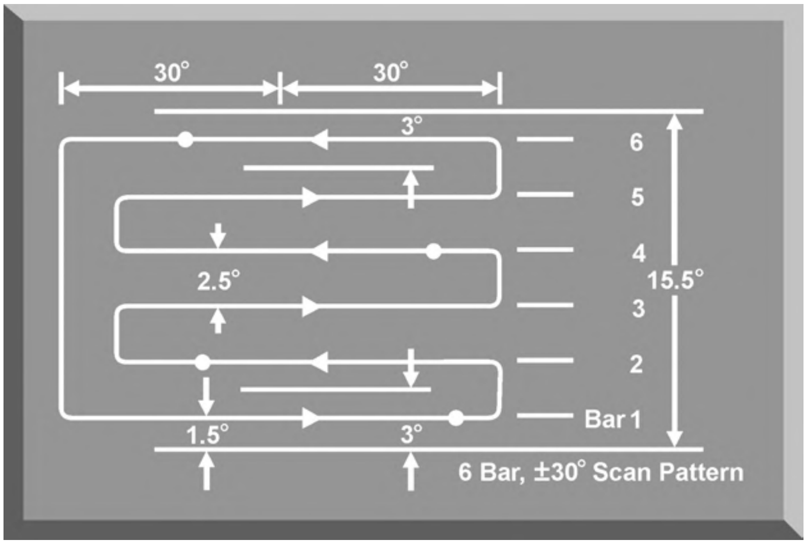

Figure 6-30. Phased Array TWS Radar Display

13. IMPACT OF TERRAIN ON RADAR SCANS

No matter what type of scan a radar system employs, terrain can limit radar line of sight (LOS) and target detection. The concepts of radar horizon, direct terrain masking, and indirect terrain masking are important factors in radar target detection and mission planning.

a. RF waves traveling in the atmosphere are bent, or refracted, and do not travel in a straight line. However, the degree of refraction depends on atmospheric conditions which vary significantly and are difficult to accurately quantify and predict. For these reasons, most radar computations are based on the assumption that RF waves travel in a straight line. The concept of the radar horizon is based on this assumption.

(1) The radar horizon shown in Figure 6-31 is the maximum range a radar system can detect a target due to the curvature of the earth. The distance (d) to the horizon for a radar antenna at a height (h) can be computed using Equation 6-6.

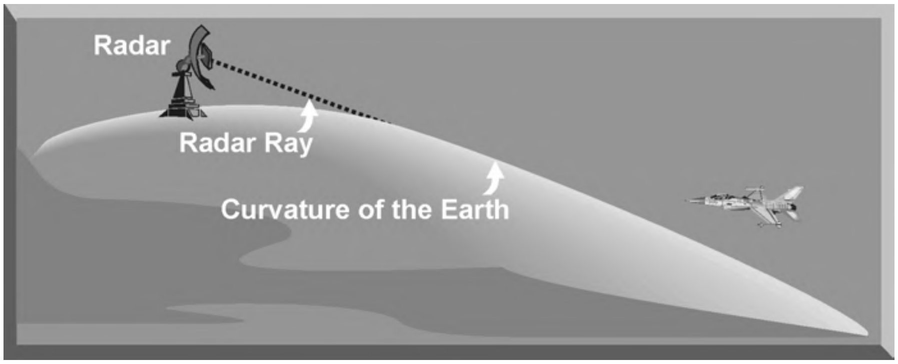

Figure 6-31. The Radar Horizon

$$\text{Distance to Horizon } (d) = \sqrt{2kah}$$

k = constant (4/3)
a = radius of earth
h = height of radar antenna

Equation 6-6. Basic Radar Horizon Equation

(2) In Equation 6-6, the constant, k, is commonly used to compensate for the assumption of straight line propagation of RF waves. Using the assumption that the height of the radar antenna is small compared to the radius of the earth, distance is measured in nautical miles (nm), and height is measured in feet, then Equation 6-6 reduces to Equation 6-7.

$$\text{Distance to Horizon } (d) = 1.06\sqrt{h}$$

Equation 6-7. Simplified Radar Horizon Equation

(3) Another application of Equation 6-7 is in calculating the range at which a radar antenna will achieve LOS with a low-altitude target. To compute this distance (D), Equation 6-8 can be used.

$$\text{Target Detection Distance } (D) = 1.06\left(\sqrt{h} + \sqrt{\text{target altitude}}\right)$$

Equation 6-8. Target LOS Distance

(4) To illustrate the use of Equations 6-7 and 6-8, consider the example of a radar antenna located at a height of 25 feet and a target aircraft flying at an altitude of 100 feet. From Equation 6-7, the radar horizon for this system would be 5.3 nm. From Equation 6-8, the radar antenna will have LOS with the target aircraft at 16 nm.

b. The previous discussions of the radar horizon assumes the radar is operating over water or level terrain. Radar operations over rough terrain can present other radar LOS limitations. As illustrated in Figure 6-32, prominent terrain features can limit radar detection. In this illustration, a mountain acts as a radar horizon and limits target LOS in one sector of the radar scan. This situation is called direct terrain masking. Placing prominent terrain features between the aircraft and threat radar systems effectively negates these systems and is an integral part of threat avoidance during combat mission planning.

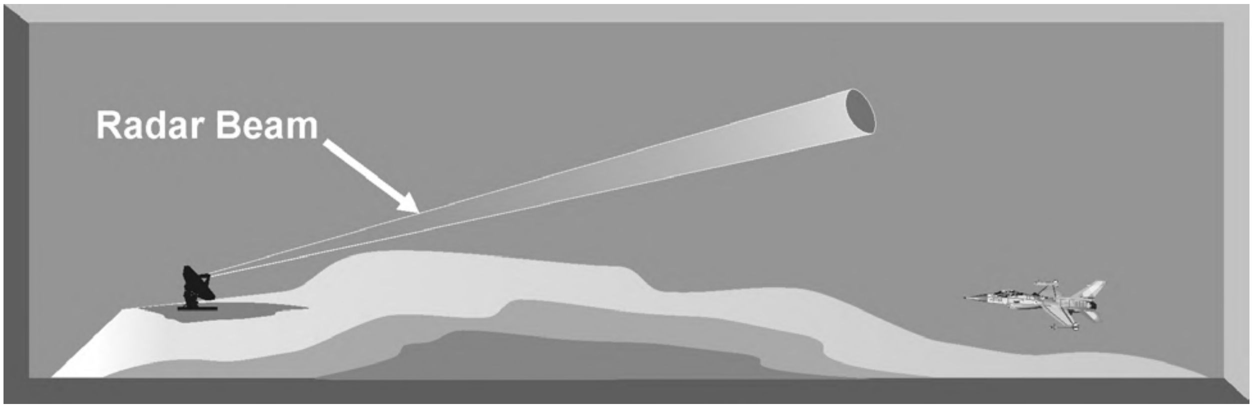

Figure 6-32. Direct Terrain Masking

c. Figure 6-33 depicts another impact of terrain on target detection/indirect terrain masking. When both the aircraft and a prominent terrain feature are illuminated by a radar beam, a pulse radar system may not be able to differentiate the target return from the return generated by the terrain. Indirect terrain masking is most effective when the aircraft is flying abeam the radar site. Pulse Doppler radars and radar systems employing moving target indicator (MTI) circuits may be able to negate the effectiveness of indirect terrain masking. However, indirect terrain masking is another important consideration for threat avoidance during combat mission planning.

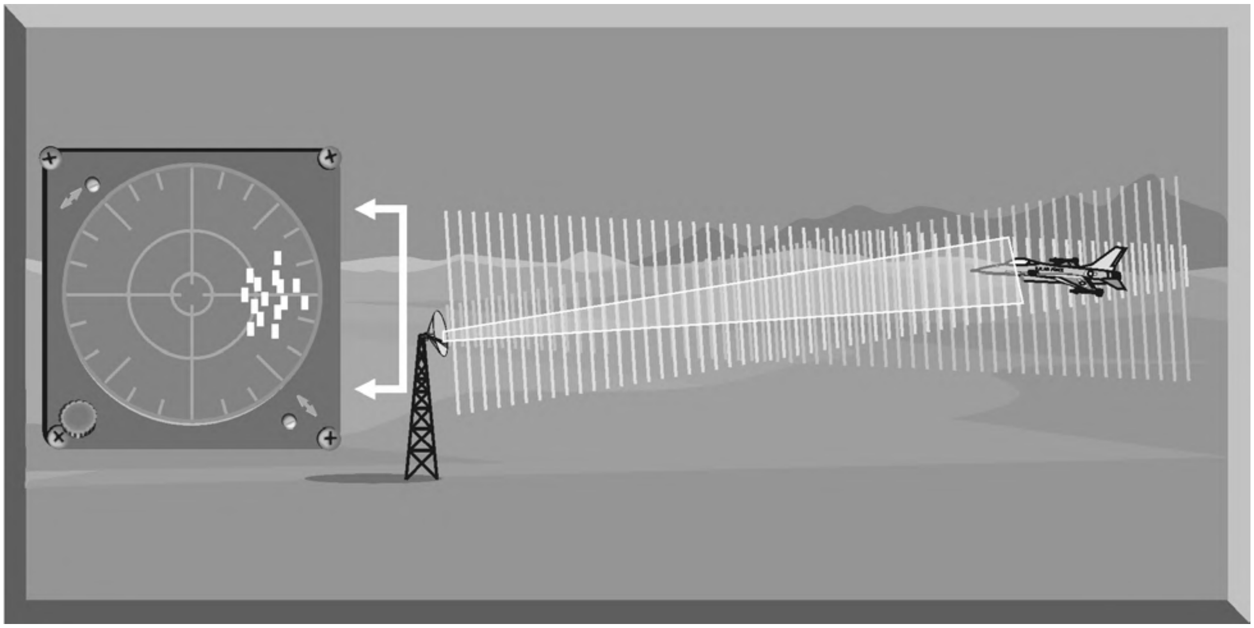

Figure 6-33. Indirect Terrain Masking

14. SUMMARY

This chapter has introduced basic antenna characteristics and how these characteristics influence their employment. The concepts of antenna gain and power density were also explained. The remainder of the chapter discussed the different types of radar scans. A radar's scan pattern is designed to enhance target detection and facilitate target tracking. The radar horizon, direct terrain masking, and indirect terrain masking are LOS limitations to all radar scans. The scan pattern chosen for a specific radar system determines the ability of that radar to provide accurate range, azimuth, elevation, or velocity of the detected target.

CHAPTER 7. TARGET TRACKING

1. INTRODUCTION

A target tracking radar (TTR) is designed to provide all the necessary information to guide a missile or aim a gun to destroy an aircraft. Once a target has been detected, either by a dedicated search radar or by using an acquisition mode, the TTR is designed to provide accurate target range, azimuth, elevation, or velocity information to a fire control computer.

a. A typical TTR has individual tracking loops to track a target in range, azimuth, elevation, or velocity. The antenna of the TTR is pointed at a single target, and the radar initiates acquisition and target track. TTRs normally employ automatic trackers to continuously measure target data. The range tracking loop employs an early gate/late gate range tracker to maintain automatic range tracking. The azimuth and elevation tracking loops generate error signals to position the antenna and maintain constant target illumination. The velocity tracking loop found on pulse Doppler and CW radars is used to reject clutter and generate accurate target radial velocity information. All this critical information is passed to a fire control computer for weapons employment.

b. The fire control computer is programmed with critical information on the capability of the weapon to be employed. For a missile, the fire control computer is programmed with the aerodynamic and range capabilities of the missile. For antiaircraft artillery (AAA), the fire control computer is programmed with the ballistics for the gun, rate of fire, and tracking rate. The fire control computer uses the precise target information from the TTR and the programmed weapon's parameters to compute a firing solution. Once a firing solution has been computed, the fire control computer either fires the weapon automatically or alerts the operator, who fires the weapon. For missile employment, the fire control computer may continue to provide missile guidance and fusing commands until missile impact or initiation of an active missile guidance mode. For AAA engagement, the fire control computer computes the required lead angle, aims the guns, and initiates firing.

c. To provide the required azimuth and elevation resolution, most TTRs use a high frequency to provide narrow antenna beamwidths for accurate target tracking. High frequency operation also allows the radar to transmit wide bandwidths. To provide the required range resolution, most TTRs employ narrow pulse widths and high pulse repetition frequencies (PRFs) to rapidly update target information. In this chapter, the target tracking techniques of conical scan, track-while-scan, lobe-on-receive-only, monopulse, Doppler radars, and pulse Doppler radars will be discussed.

2. RANGE TRACKING

In most TTR applications, the target is continuously tracked in range, azimuth, and elevation. Range tracking can be accomplished by an operator who watches an "A" scope presentation and manually positions a handwheel to maintain a marker over the desired target return. The setting of the handwheel is a measure of target range and is converted to a voltage used by the fire control computer. As target speeds and maneuvers increase, the operator may have extreme difficulty maintaining manual target range tracking. To avoid this situation, most TTRs employ an automatic range tracking loop. All pulse TTRs, which includes conical scan, track-while-scan, monopulse, and pulse Doppler radars, employ either a split gate or leading-edge automatic range tracking system. In a TTR, automatic range tracking serves two essential functions: (1) it provides the critical value of target range, and (2) it provides a target acceptance range gate that excludes clutter and interference from other returns. Since radar range is normally the first target discriminator used to initiate automatic target tracking, the second function is essential to the proper operation of the other tracking loops.

a. A range gate circuit is simply an electronic switch that is turned on for a period of time after a pulse has been transmitted. The time delay for switch activation corresponds to a specific range. Any target return that appears inside this range gate is automatically tracked. The most common type of automatic range tracking is accomplished by a split-gate tracker. Two range gates are generated as shown in Figure 7-1. The automatic range tracking loop attempts to keep the amount of energy from the target return in the early gate and late gate equal. The range tracking error is computed by subtracting the output of the late gate from the output of the early gate. The amount of the range tracking error signal is the difference between the center of the pulse and the center of the range gate. The sign of the error signal determines the direction in which the gates must be repositioned to continue to track the target.

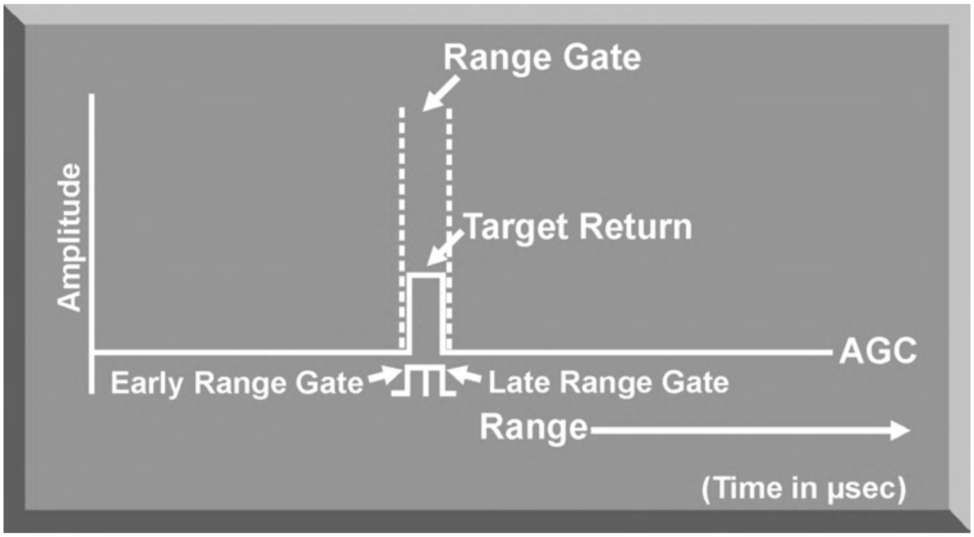

Figure 7-1. Split-Gate Range Tracker

b. Leading-edge range tracking is an electronic protection (EP) technique used to defeat range-gate-pull-off (RGPO) jamming. Figure 7-2 illustrates the application of leading-edge tracking. The leading-edge tracker obtains all range data from the leading edge of the target return. All RGPO cover pulse jamming tends to lag the target return by some increment of time (see (a) in Figure 7-2). By differentiating the entire return with respect to time, the target return can be separated from the jamming pulse (see (b) in Figure 7-2). Employing a split-gate tracker electronically positioned at the initial pan, or leading edge, of the returning pulse, the range tracking loop can track the target return and ignore any jamming signals. The range tracking loop then uses split-gate tracking logic to determine the magnitude and direction of range tracking errors and reposition the range gate.

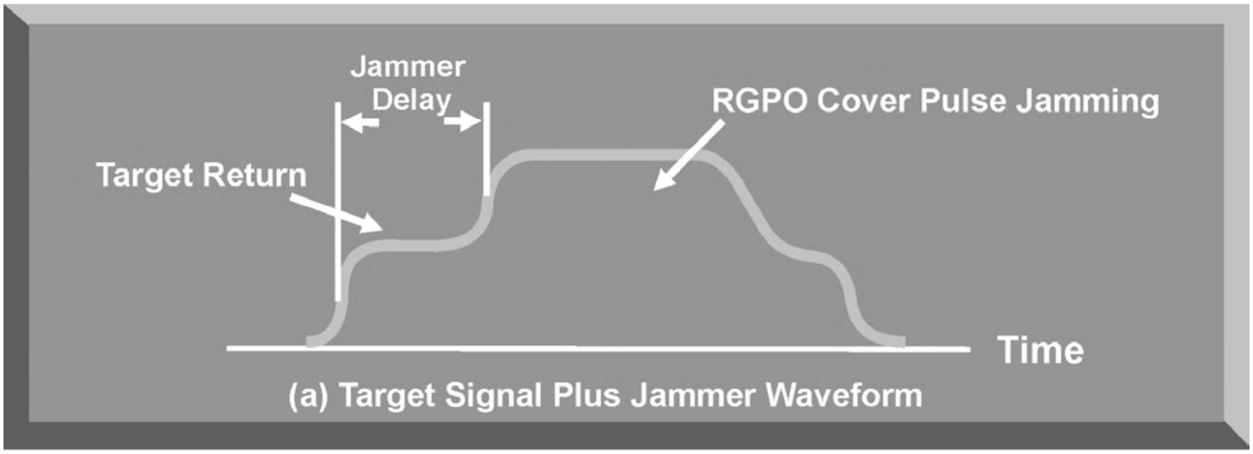

(a) Target Signal Plus Jammer Waveform

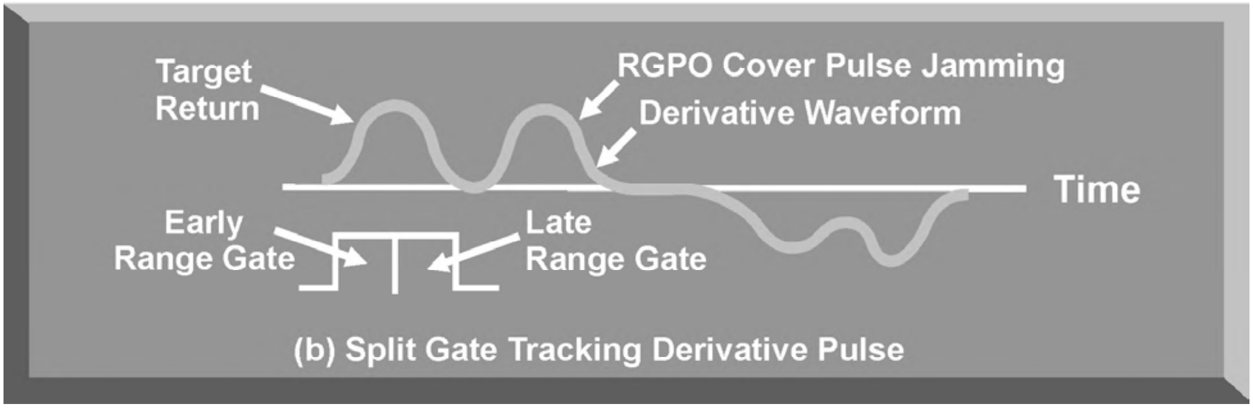

(b) Split Gate Tracking Derivative Pulse

Figure 7-2. Leading-Edge Range Tracker

c. The width of the tracking gate is an important radar design consideration. The range gate should be sufficiently narrow to effectively isolate the target from other returns at different ranges. It should be wide enough to allow sufficient energy from the target echo to be displayed. The width of the range tracking gate is normally equal to the pulse width of the radar.

d. Nearly all range tracking gates employ some form of automatic gain control (AGC). AGC is designed to limit target clutter and glint. It is also designed to avoid excessive false alarms.

3. CONICAL SCAN

A conical scan tracking system is a special form of sequential lobing. Sequential lobing implies that the radar antenna beam is sequentially moved between beam positions around the target to develop angle-error data. For a conical scan radar to generate azimuth and elevation tracking data, the beam must be switched between at least four beam positions as shown in Figure 7-3.

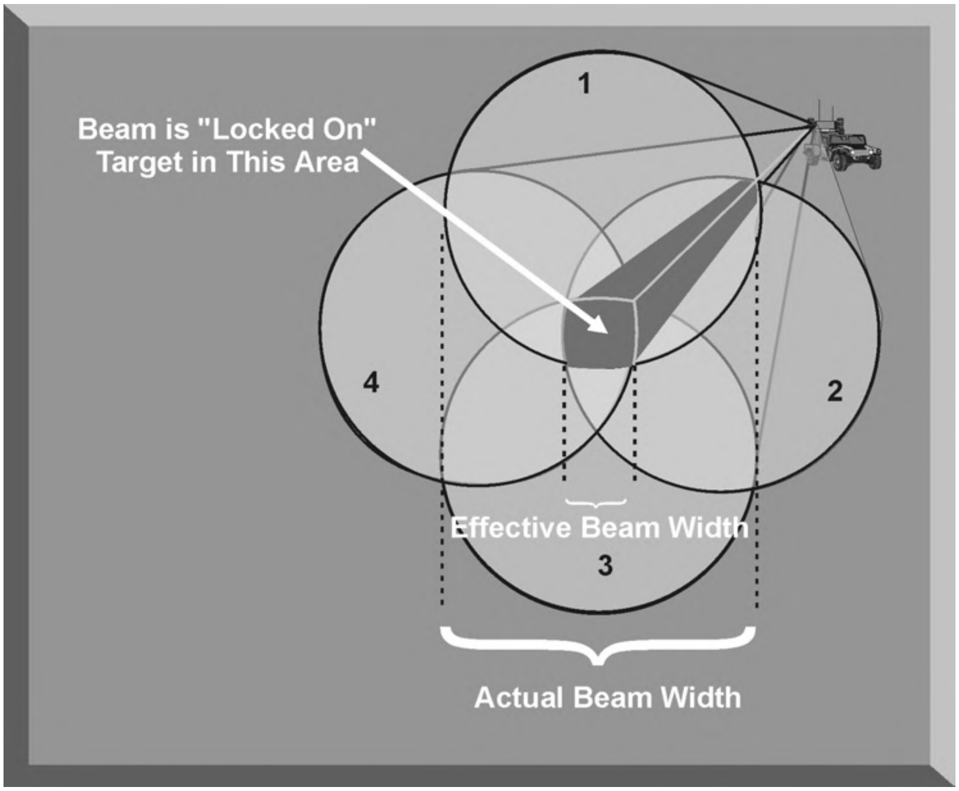

Figure 7-3. Conical Scan Positions

a. One of the simplest conical scan antennas is a parabola with an offset rear feed that rotates, or nutates, to maintain the plane of signal polarization. The radar beam is rotated at a fixed frequency around the target. The angle between the axis of rotation (normally the axis of the antenna) and the axis of the antenna beam is called the squint angle.

b. A conical scan radar first tracks the target aircraft in range. For azimuth and elevation tracking, the target return is modulated at a frequency equal to the rotation frequency of the radar beam. This results in a target signal output that resembles a sine wave (Figure 7-4). The azimuth and elevation tracking loops

drive servo motors to position the antenna to keep the energy level in each of the four positions equal. The amount of the modulated signal determines how far the target is off the antenna boresight while the phase of the modulation (positive or negative) determines the direction.

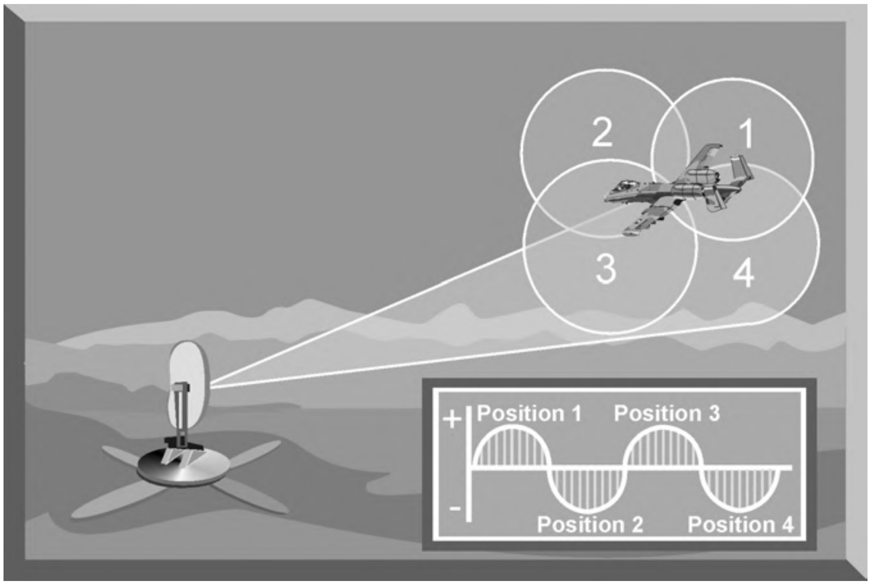
2
1
3
4
+
-
Position 1
Position 3
Position 2
Position 4

Figure 7-4. Conical Scan Modulation

c. In Figure 7-5, most of the target energy is in position 1, with a small amount of energy in position 4. The output from the elevation tracking loop is positive and drives the antenna servos upward. The output from the azimuth tracking loop moves the antenna to the right.

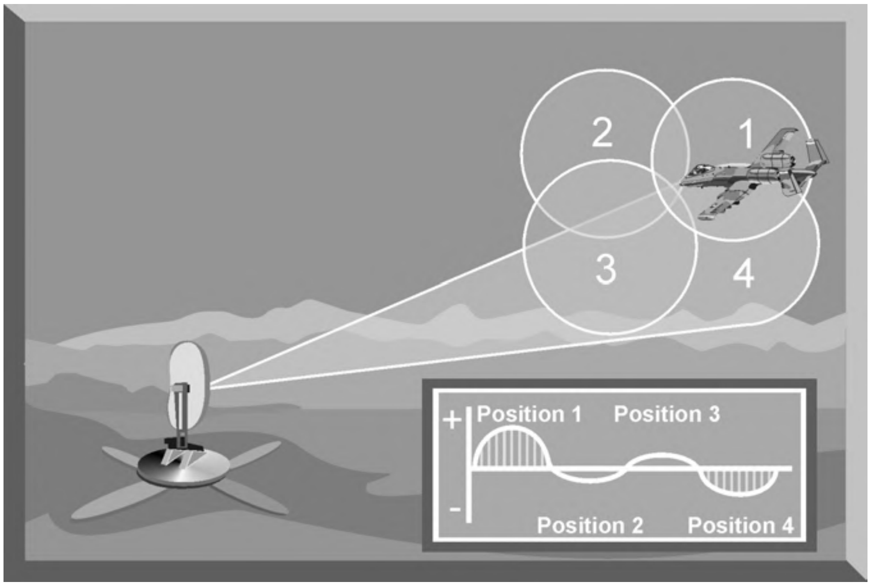
2
1
3
4
+
-
Position 1
Position 3
Position 2
Position 4

Figure 7-5. Conical Scan Tracking Errors

d. Once a balance of target energy in each scan position is achieved, the target is in the central tracking area (Figure 7-6). The azimuth and elevation tracking circuits continue to drive the antenna servos to maintain this energy balance which keeps the radar beam on the target.

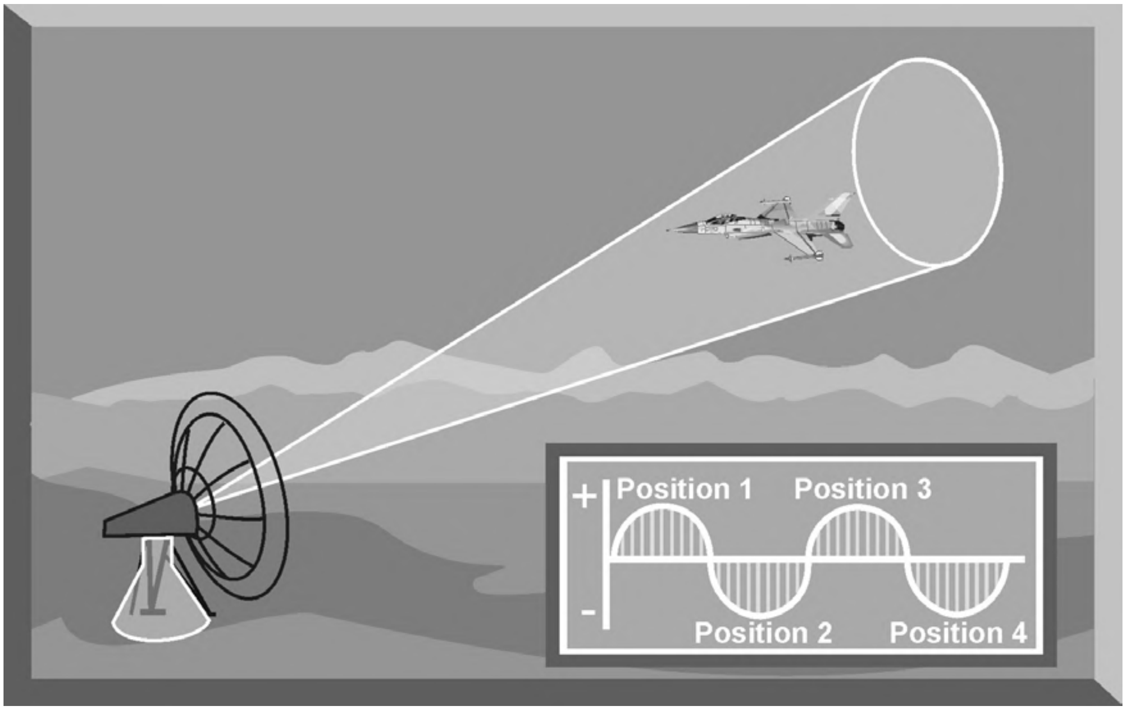

Figure 7-6. Conical Scan Tracking

e. The primary advantage of a conical scan radar is the small beamwidth which provides extremely accurate target tracking information. The primary disadvantages of conical scan include the following: (1) the narrow beamwidth makes target acquisition a problem. Even using a Palmer-helical scan, it may take considerable time to find and initiate track on a target; (2) conical scan radars are vulnerable to inverse gain modulation jamming based on the scanning frequency of the rotating beam; (3) a conical scan radar must analyze many radar return pulses to generate a tracking solution.

4. TRACK-WHILE-SCAN (TWS)

TWS is a combined search and tracking mode that sacrifices the continuous target observation capability of the dedicated tracker in return for the ability to monitor a finite sector of airspace. This is accomplished while maintaining tracks on multiple targets moving through the covered airspace. There are two types of radar systems capable of TWS operation: conventional and phased array.

a. Conventional track-while-scan threat radars use two separate antennas to generate two separate beams (Figure 7-7). These beams operate at two different

frequencies and are sectored so they overlap the same region of space. This overlap area provides a tracking area for a single target. One beam is sectored in the vertical plane to give range and elevation. The other beam is sectored in the horizontal plane to provide range and azimuth. Each beam scans its sector at a rate of 5 to 50 times per second. This provides a rapid update on target range, azimuth, and elevation.

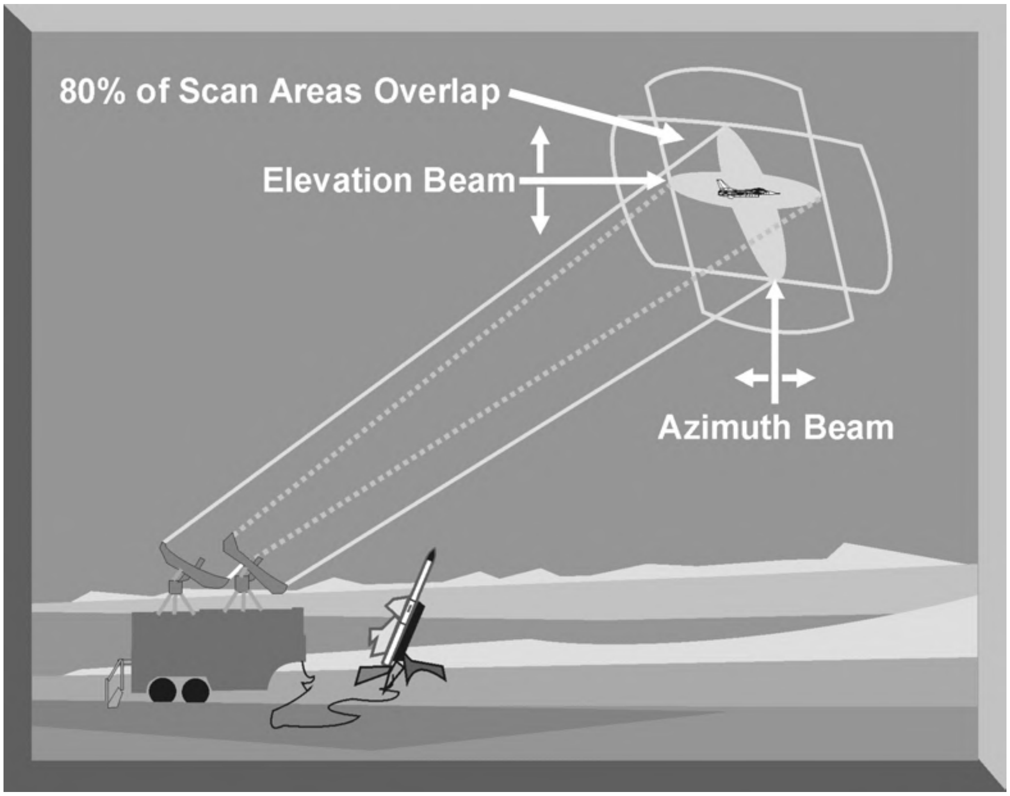

Figure 7-7. Conventional TWS Radar Beams

(1) The two TWS antennas generate their beams using an electro-mechanical principle. Each antenna provides inputs to its own display and provides angle and range information for all targets in the coverage of the radar. The display from the elevation beam is calibrated in range and elevation, while the display from the azimuth beam is calibrated in azimuth and range. Operators position a cursor over the returns on these displays using range as the primary parameter. Once a target has been designated for engagement, the radar automatically attempts to keep the tracking axis of the radar beams centered on the target.

(2) Once the target is designated by the operator, the range gate is enabled and tracks the target using a split-gate tracker. The azimuth and elevation tracking loops receive information only from targets inside the range gate. As the beams scan across the target, a burst of pulse returns is received that have an

amplitude envelope corresponding to the beam pattern. The azimuth beam pattern is shown in Figure 7-8.

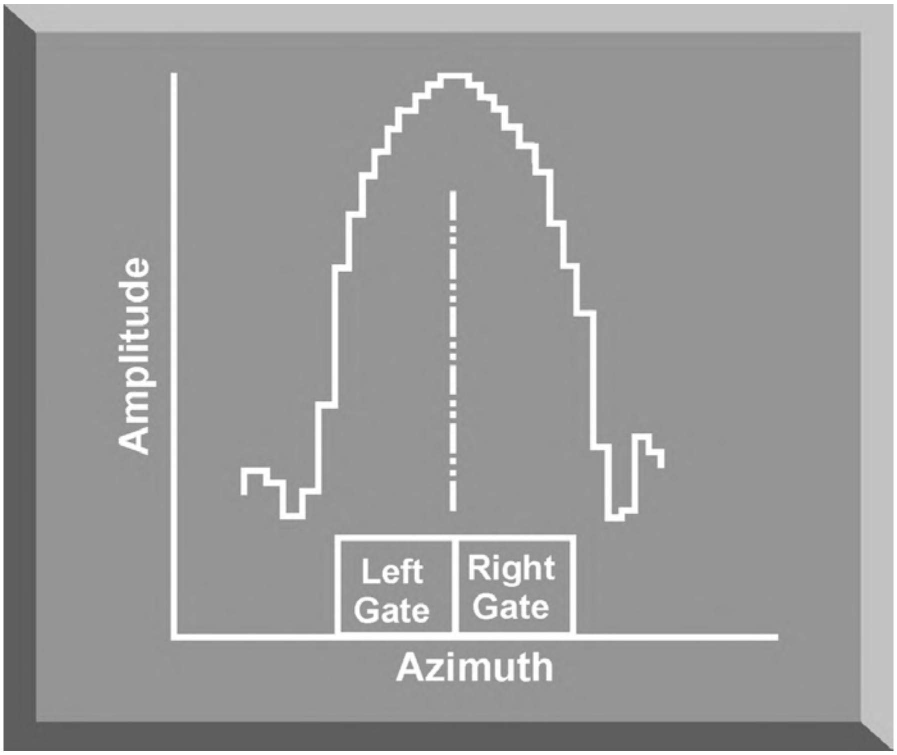

Figure 7-8. Conventional TWS Azimuth Tracking

(a) The azimuth tracker is typically a split-gate tracker, identical in concept to a split-gate range tracker. However, range delay time is replaced by azimuth scan time. The azimuth tracker uses a left gate and right gate. Each gate integrates its share of the target return to generate a voltage/time value. When the azimuth gate is centered on the target, the areas are equal and the error signal (right gate minus left gate) is zero. The azimuth tracking loop sends signals to the antenna servos to keep the target centered in the scan area.

(b) Elevation tracking is accomplished in the same manner by using an up gate and a down gate. The elevation tracking loop also sends signals to the antenna servos to keep the target centered in the scan area.

(3) Once the target is designated and the radar is automatically keeping the radar return centered in the tracking area, target range, azimuth, and elevation information is sent to a fire control computer. The radar continues to provide information on other targets in the scan area. The fire control computer indicates the firing solution has been achieved for the designated target, and a missile is launched. The radar tracks the target and the missile and provides in-flight corrections to the missile right up to the moment of missile impact. These corrections are based on both target and missile azimuth, range, and elevation

information. Information is passed to the missile from a dedicated antenna on the radar to special antennas on the missile. Commands from the radar to the missile are called uplink guidance commands. Information from the missile back to the radar and fire control computer is called downlink information.

(4) The advantages of a conventional TWS radar include the following: (1) TWS radars have the ability to maintain radar contact with all targets in the sector scan area while maintaining target track on a single target, and (2) the rapid sector scan rate provides a rapid update on target parameters. The primary disadvantages of a conventional TWS radar include: (1) a large resolution cell due to the wide azimuth and elevation beams, and (2) vulnerability to modulation jamming based on the scan rate of the independent beams.

b. Many modern radars employing a planar or phased array antenna system have a TWS mode. The radar does not really track and scan simultaneously, but rapidly switches between search and track (Figure 7-9).

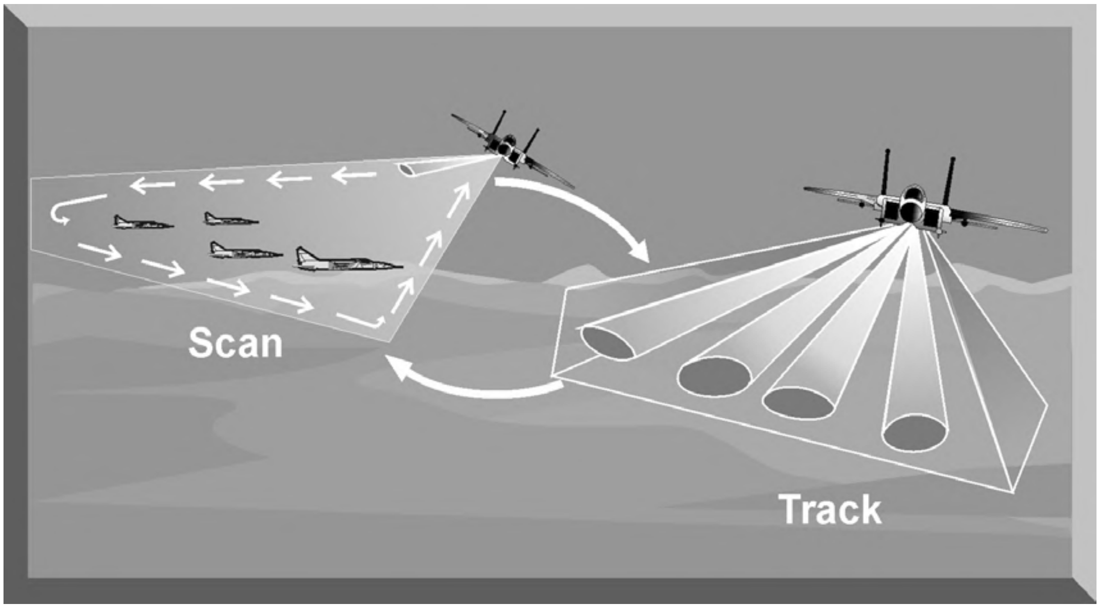

Figure 7-9. Phased Array TWS

(1) The most common air-to-air radar system uses a planar array antenna. In the scan mode, the radar antenna generates a pencil beam and uses a raster scan to detect targets in the search area. Targets detected are presented to the pilot on the aircraft's radar display (Figure 7-10).

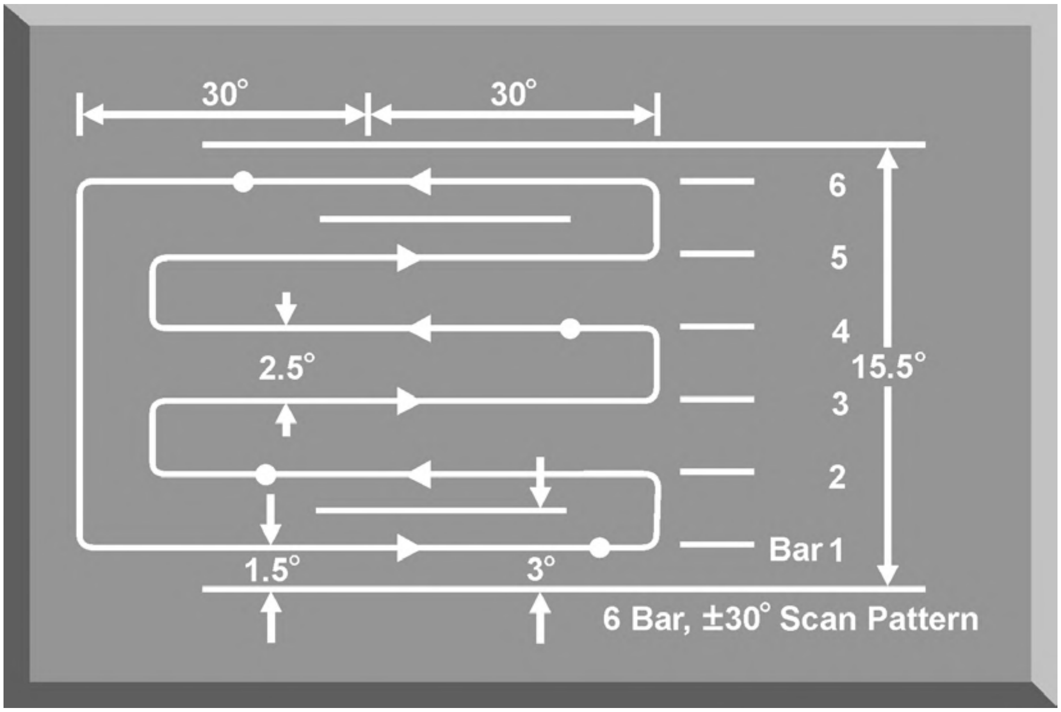
30°
30°
6
5
4
3
2
Bar 1
15.5°
2.5°
1.5°
3°
6 Bar, ±30° Scan Pattern

Figure 7-10. Air-to-Air TWS Display

(a) In the track mode, the antenna generates multiple beams to illuminate individual targets. The radar typically uses monopulse or pulse Doppler techniques to update target range, azimuth, elevation or velocity. These tracking techniques will be covered later in this chapter. The radar initiates a track file on each detected target that contains all current data on the target and an estimate of future target position.

(b) As the radar switches between track and scan modes, target parameters are updated in the tracking loop (Figure 7-11). The new target information is compared to the predicted information in the measurement data processing cell. If the two sets of data agree within certain limits, target position and information are updated. This process is called gating.

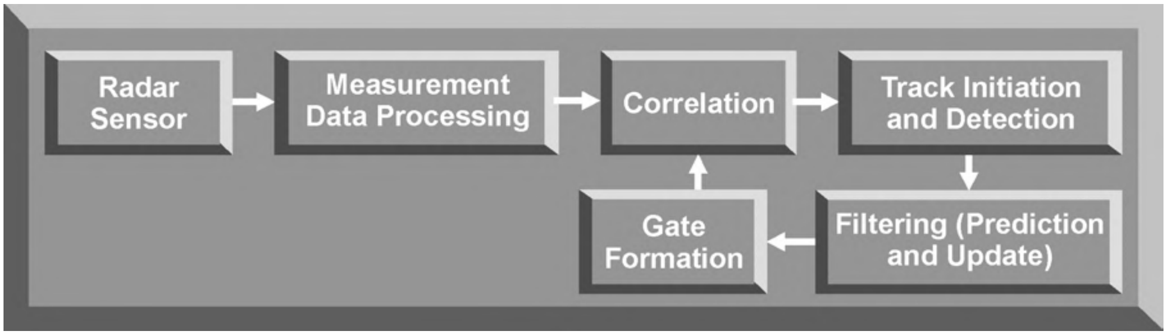
Radar Sensor
Measurement Data Processing
Correlation
Track Initiation and Detection
Filtering (Prediction and Update)
Gate Formation

Figure 7-11. Planar/Phased Array TWS Tracking Loop

(c) If the updated target information does not correspond to the predicted values, the information is sent to the correlation processor. The correlation processor attempts to resolve the conflict based on further refinement of target data. If the correlation processor cannot assign the target parameters to an existing track file, a new track file is generated and displayed.

(2) The obvious advantage of a planar/phased array TWS radar is that it can search a large volume of airspace while tracking individual targets. The number of targets that can be tracked is limited by the number of beams the radar can generate. Planar/phased array radars have increased peak and average power when compared to pulse radar systems. Since the radar beam of a planar/phased array radar is electronically controlled and can rapidly change beams and scans, it is resistant to many jamming techniques. The primary disadvantages of a planar/phased array TWS radar include its complexity, cost, and reliance on computer processing.

5. LOBE-ON-RECEIVE-ONLY (LORO)

LORO is a mode of radar operation developed as an EP feature for a track-while-scan radar. LORO can be employed by any radar that has the capability to passively track a target. In a LORO mode, the radar transmits a continuous signal from a set of illuminating antennas. This continuous signal hits the target, and the return echo is received by a different set of receive antennas (Figure 7-12). The receive antennas are passive and generate azimuth and elevation tracking signals by electronically scanning the reflected signal. The tracking signals are sent to the antenna servos to keep the illuminating antennas pointed at the target and centered in the receive antenna tracking area. The range tracking circuit uses the time delay between the transmission and reception of the illuminating antenna signals. A split-gate tracker is used to provide range tracking.

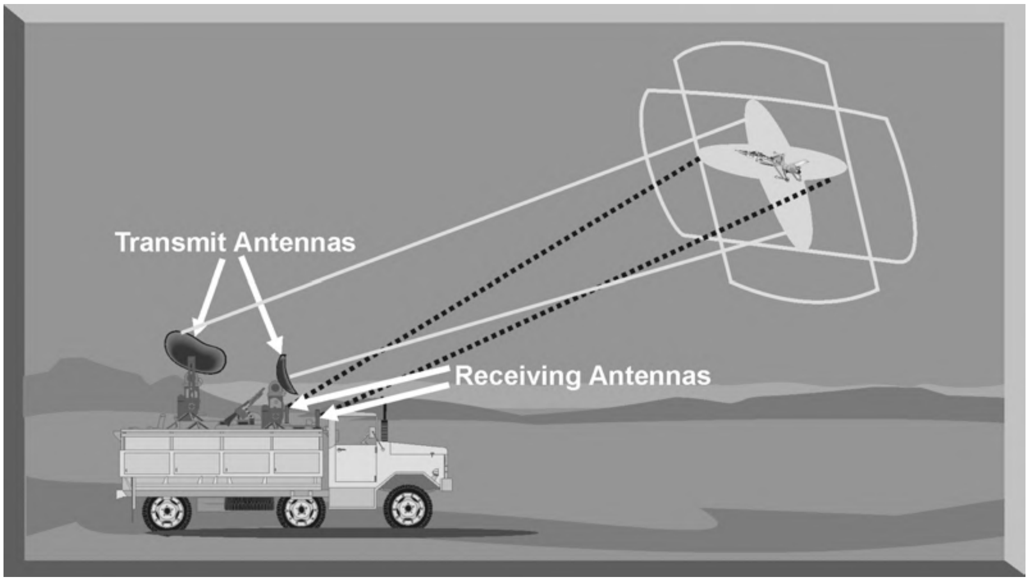

Figure 7-12. LORO Mode

a. The illuminating antennas used in the LORO mode have very narrow beamwidths and transmit at a high power level. This reduces the effectiveness of noise jamming techniques against a radar employing a LORO mode. In addition, the continuous signal from the illuminating antennas negate the effectiveness of most angle deception jamming techniques designed to defeat TWS radars. These specialized jamming techniques exploit the scan rate of TWS antennas. In the LORO mode, the illuminating antennas do not have a scan rate. The limited effectiveness of both noise and deception jamming techniques is the major advantage of the LORO mode.

b. The LORO mode also provides a track-on-jam (TOJ) capability to exploit noise jamming techniques. In a TOJ mode, the receive antennas passively track any detected noise jamming signals. The radar assumes that the most intense jamming signal is the target. The receive antennas process the strongest jamming signal as if it were a target echo from the transmit antenna signal. The receive antennas generate azimuth and elevation tracking signals to keep the jamming signal centered in the tracking area. The TOJ mode does not provide target range.

6. MONOPULSE RADAR

Monopulse radars are among the most complex radar systems. From a single pulse, a monopulse radar can derive all the data needed to update a target's position. It does this by comparing the relationship of two or more radar beams that are transmitted together from the same antenna but received separately. By comparing the phase or amplitude of the energy in these returned beams, target azimuth and elevation can be found. The speed that a monopulse radar updates the target's position, coupled with its azimuth/elevation accuracy and resistance to jamming, make this a popular choice amongst many newer TTRs.

a. The Magic T circuit allows monopulse radars to gather and process information from a single pulse that is transmitted and received using separate antennas. Figure 7-13 depicts a four-beam monopulse radar system. The Magic T is a sophisticated wave guide that can separate multiple signals by their phase relationships. This allows the radar tracking computer to compare the signal amplitude from the reflected pulses in several distinct ways. As the reflected energy enters the Magic T, it is separated by phase. The energy in the "H" arm will be in-phase and will exit from ports 1 and 2. The received energies entering the wave guide in the "E" arm exit at the number 1 port. This energy is exactly 180° out-of-phase with energy entering the H arm. This ensures there is no transfer of energy between the E and H plane arms. A typical monopulse radar would have eight Magic T's.

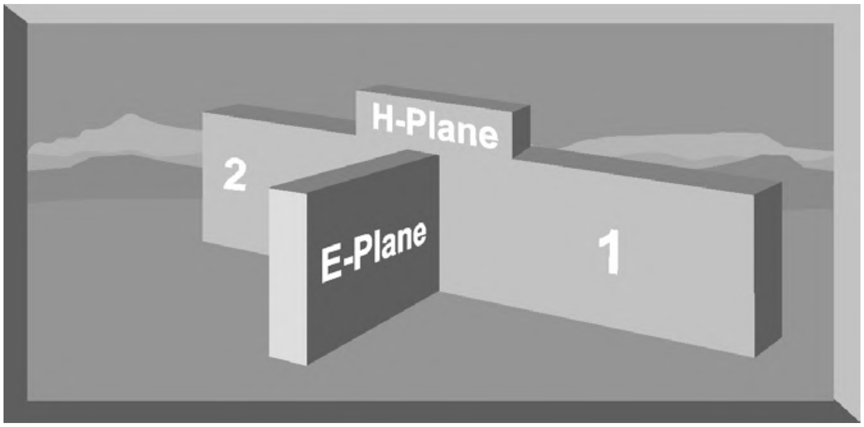

Figure 7-13. Monopulse Magic T

b. The output of a Magic T is the sum and difference of the two signals. These sum and difference values in amplitude or phase are used to generate azimuth and elevation error signals as well as to compute range. Monopulse radars may split the incoming signal as depicted in Figure 7-14. Upper antennas receive the A and B signals. Lower antennas receive the C and D signals. The various combinations of signals are processed and compared by simple addition and subtraction of the signal characteristics. From these steps, azimuth, range and elevation data are computed.

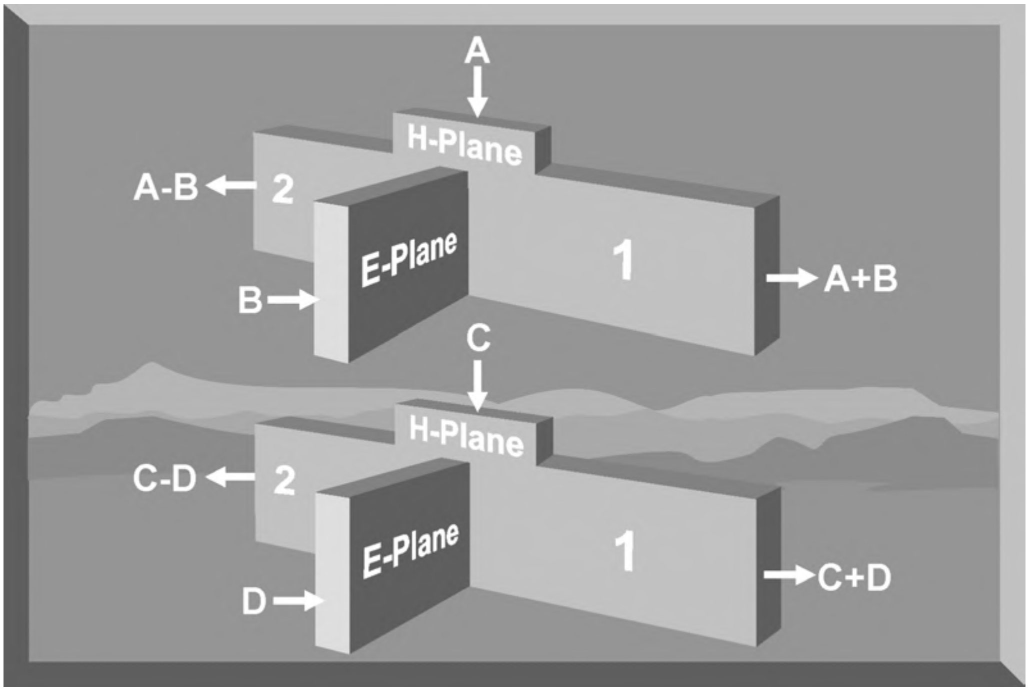

Figure 7-14. Magic T Output Signals

(1) The top equation in Figure 7-15 is used to compute target range. Target range is derived by adding the signal from the A scan to the signal from the B scan. This value is then added to the sum of the C and D scan signals. The output of these combinations is then passed to the range circuit which figures out the range of the target and displays it to the operator. Monopulse range tracking is accomplished by using either a leading-edge or split-gate tracking loop.

(2) Target elevation tracking error is derived using the middle equation from Figure 7-15. The signal from the A scan is added to the signal from the B scan. The signals from the C and D scans are also added. This time, however, the sum of A+B is subtracted from the sum of C+D. This value is then passed to the elevation circuit. Elevation signals are sent to the operator display and the servo mechanism, which corrects to the updated position of the target.

(3) The bottom equation from Figure 7-15 is used to compute the azimuth error. The signal from the A scan is added to the signal from the C scan. The B and D scans are also added together. The sums of these values are subtracted from each other. This difference equals the tracking error in azimuth. The radar system will then position the search beam to even the energy level between the two pairs of sums. When this occurs, the azimuth tracking error is zero.

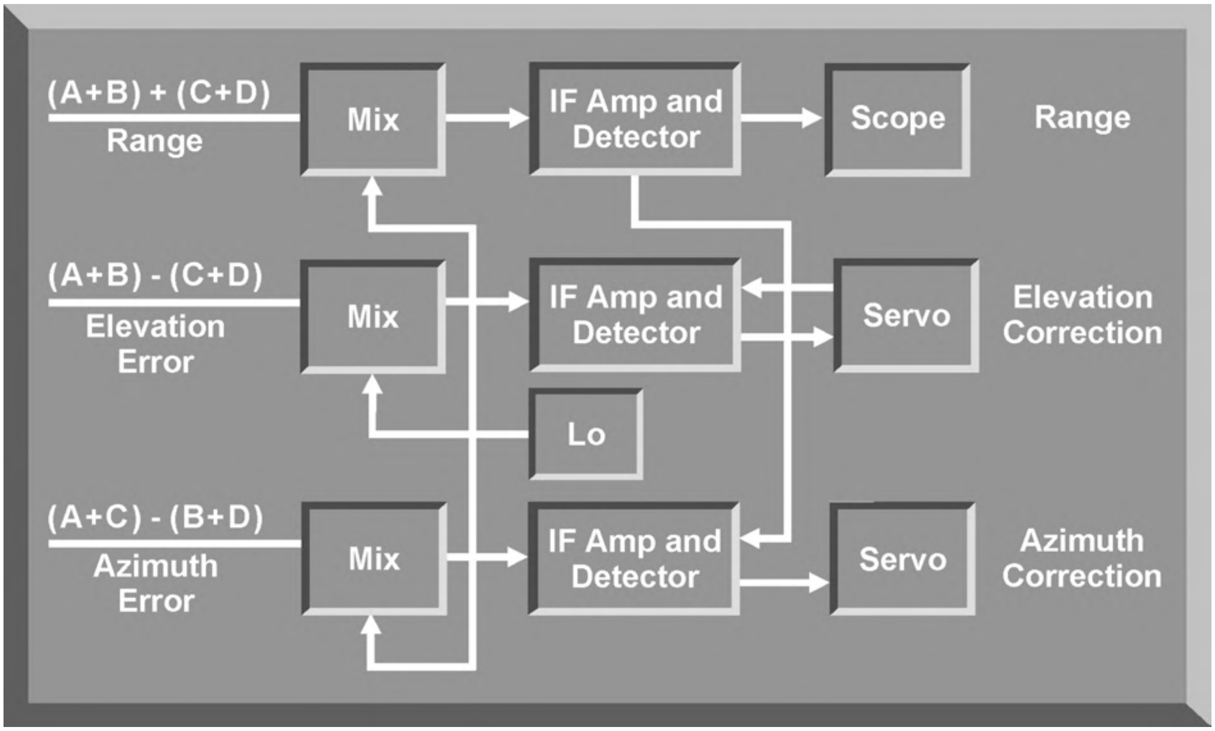
(A+B) + (C+D)
Range
Mix
IF Amp and Detector
Scope
Range
(A+B) - (C+D)
Elevation Error
Mix
IF Amp and Detector
Servo
Elevation Correction
Lo
(A+C) - (B+D)
Azimuth Error
Mix
IF Amp and Detector
Servo
Azimuth Correction

Figure 7-15. Monopulse Tracking Loops

c. A further illustration of the idea of signal combinations can be seen by referring to the F-16 in Figure 7-16. All the energy is received in the B scan area.

The A scan signal is added to the B scan signal. The signals from the C and D scans are also added. However, the sum of A+B is now subtracted from the sum of C+D. In this case, the values from the A scan and the C scan are zero. This total value of (A+B) - (C+D) is then passed to the elevation circuit.

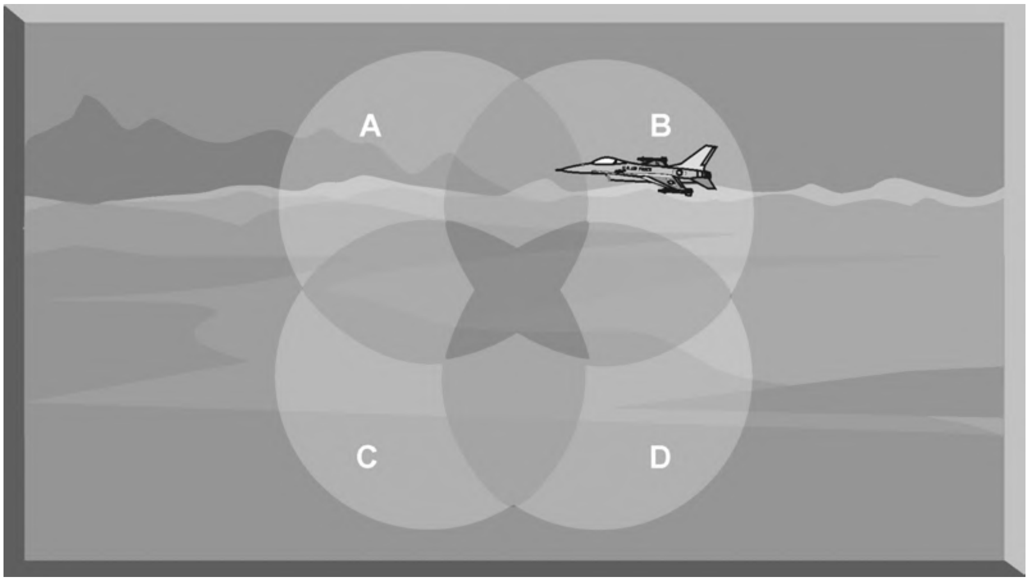

Figure 7-16. Monopulse Elevation Tracking Error

(1) The comparison in Figure 7-17 shows that all the energy is in the B scan. The radar will reposition the scan vertically to balance the energy between the B and D scans. When the energy level is balanced, the elevation error is zero.

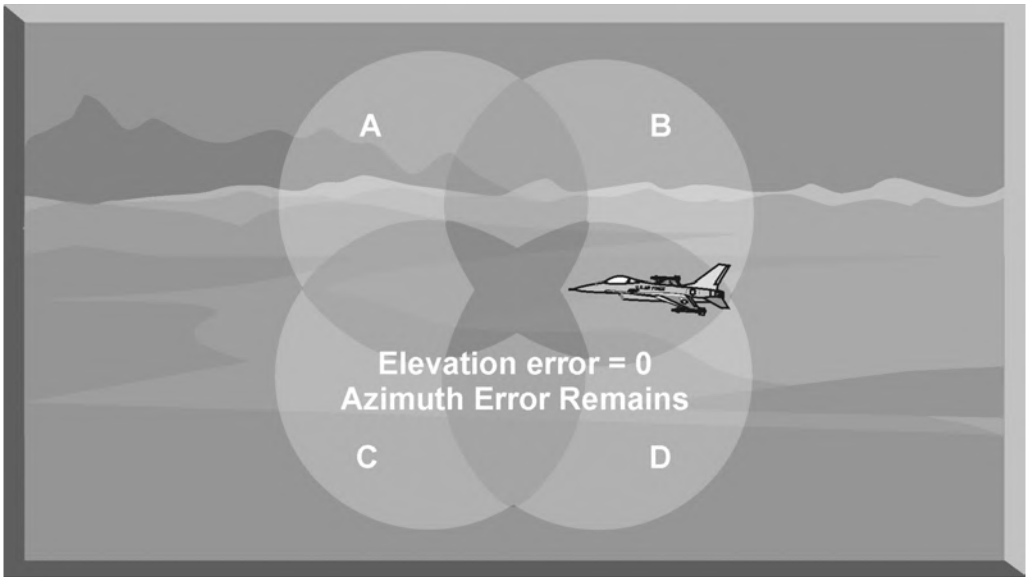

Figure 7-17. Monopulse Elevation Track

(2) Using the azimuth error equation from Figure 7-15, the azimuth tracking loop computes the azimuth error and repositions the antenna to equalize the received energy in all the beams. The monopulse radar has now established a tracking solution (Figure 7-18). All these computations are done instantaneously on a pulse-to-pulse basis.

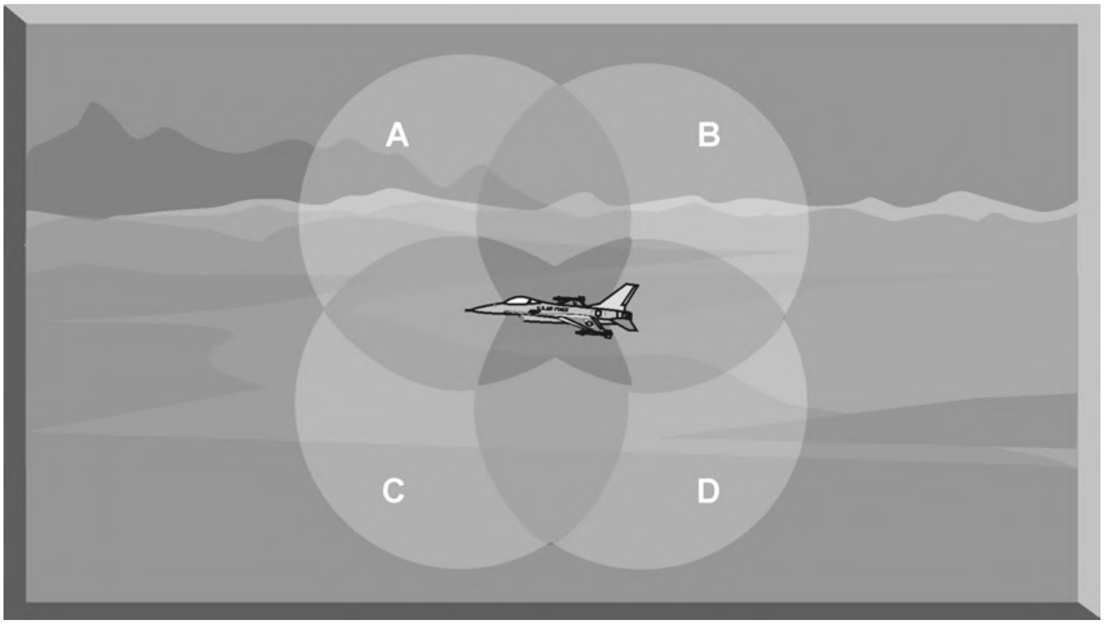

Figure 7-18. Monopulse Azimuth Track

7. CONTINUOUS WAVE (CW) RADARS

CW radar was one of the earliest forms of radar systems. Unlike pulse radar systems, CW radars emit a continuous beam of RF energy with no interruptions in the transmissions to detect returning echoes. A continuous radar transmission from the antenna requires that a classic CW radar have two antennas, one for transmission and one for reception (Figure 7-19).

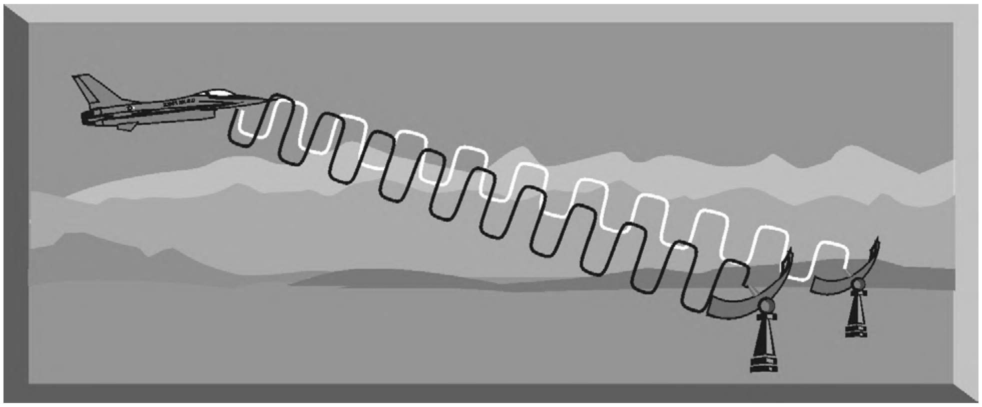

Figure 7-19. CW Radar

a. Since a continuous transmission results in a continuous echo signal, it is impossible to tell what part of the echo is associated with any particular part of the transmission. This makes range typical determination impossible. The azimuth and elevation tracking capability of a CW radar is based on the antenna position when the target is illuminated. The simple application of the Doppler principle provides a means for a CW radar to track a target in velocity and reject clutter. The Doppler principle deals with the fact that a radar return from a moving target will be shifted in frequency by an amount proportional to its radial velocity compared with the radar site. Using the difference in frequency from the transmitted signal to the received signal, a CW radar can separate the target return from clutter based on a change in frequency. This type of radar is called a CW Doppler radar.

b. The most serious disadvantage of a simple CW Doppler radar is that it does not provide any range information on the target. One method of obtaining range information from a CW radar uses frequency modulation (FM). The modulation can be sinusoidal, sawtooth, triangular, or any shape, as long as the rate of frequency change is known. The transmitter emits a continuous signal, but the frequency is changed in a known pattern. When the echo returns from the target, it is then compared to the frequency being transmitted. This frequency difference is directly proportional to the range of the target.

(1) Figure 7-20 shows how a FM CW radar measures range. Using a triangular wave for modulation, a plot of transmitted frequency over time would look like the solid line. It is important to note that this is not a depiction of the transmitted wave but a plot of how the frequency of the wave varies with time. For a target, without any relative motion, the frequency returning to the receiver is depicted by the dotted line. The target echo frequency lags the transmitted frequency by time (t). There is also a frequency difference between the transmitted and received signals. Range to the target may be computed by measuring this frequency difference and dividing by the rate of change of the transmitted frequency. The result is time. Dividing this time by 12.4 microseconds per radar mile yields range to the target. The frequency difference is constant for a target, without any relative motion, except for the brief intervals when the change in frequency goes from a positive to a negative slope. These “ditches” are negligible and can be disregarded when calculating range. The average amplitude is equal to target range.

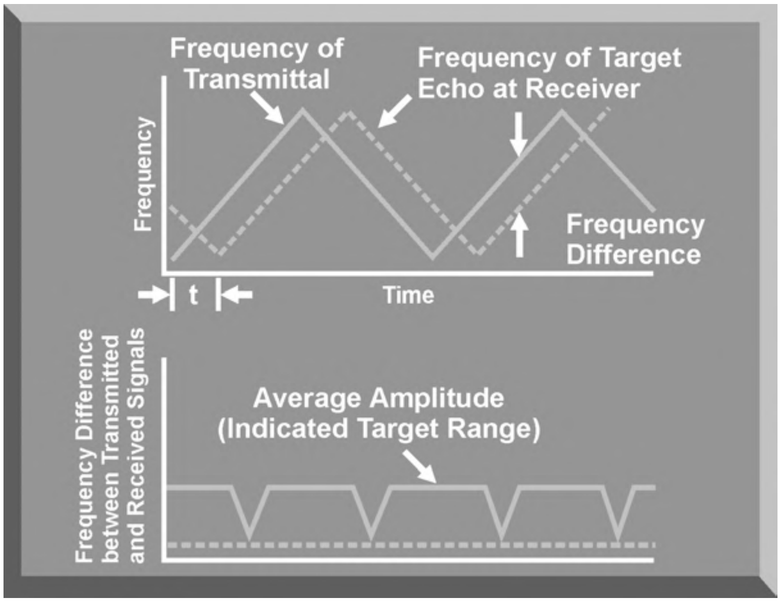
Frequency of Transmittal
Frequency of Target Echo at Receiver
Frequency
Frequency Difference
t
Time
Frequency Difference between Transmitted and Received Signals
Average Amplitude (Indicated Target Range)

Figure 7-20. FM CW Radar Range Determination (No Relative Motion)

(2) For a target moving toward the radar, the FM CW radar measures both target range and velocity. In Figure 7-21, the frequency of the return signal will be increased as depicted by the dotted line. Remember, this is a graph of frequency versus time, not a depiction of the radar wave. For a moving target, this results in a varying frequency difference. A plot of the frequency difference over time provides target range by averaging the difference, while target velocity is found by comparing the two frequency differences.

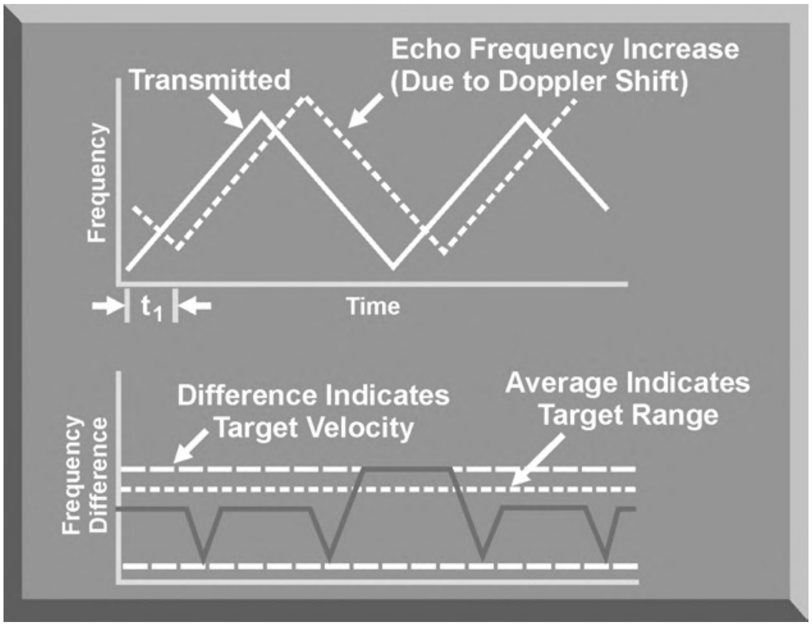
Transmitted
Echo Frequency Increase (Due to Doppler Shift)
Frequency
t_1
Time
Frequency Difference
Difference Indicates Target Velocity
Average Indicates Target Range

Figure 7-21. FM CW Radar Range Determination (Relative Motion)

c. The primary advantage of the FM CW Doppler radar is its ability to combine the clutter rejection features of a simple CW Doppler radar with the capability to detect range. The widest application of FM CW Doppler radars is in radar altimeters for aircraft. In addition, the HAWK missile system uses an FM CW Doppler acquisition radar and a CW target illuminator.

d. Another method used for clutter rejection in a pulse radar system is to employ special circuits, or Doppler processing, to identify and reject clutter. These special circuits are added to the receiver section of a pulse radar and are called moving target indicators (MTIs). There are two types of MTIs, non-coherent and coherent.

(1) The earliest form of MTI was called non-coherent MTI or “area” MTI. Non-coherent MTI radars do not process Doppler frequencies. The returns from one scan are subtracted from returns from the next scan. All targets that move at least one resolution cell in the time between scans are displayed. All stationary objects, including fixed clutter, are cancelled and not displayed. In this type of MTI, clutter cancellation is based on the size and movement of the return. Due to changes in the clutter cross section, instabilities in radar operations, variations such as rain or clouds, and noise from the transmitter, clutter cancellation is never complete. In another form of non-coherent MTI, the radar returns from moving targets are compared to the returns from fixed targets, and the fixed targets are cancelled. These non-coherent MTIs are simple, but they do not provide the clutter rejection available from coherent MTI radars.

(2) A coherent MTI uses the fact that Doppler shifts appear to a pulse radar as phase shifts on the received target pulse. Coherent MTI uses sophisticated circuitry, including stable local oscillators (STALOs) and coherent local oscillators (COHOs) to capture and process these phase shifts. Further processing of these phase shifts yields velocities for each return. Those velocities associated with stationary targets are rejected and only moving targets are displayed. Coherent MTIs have a major problem called “blind speeds.” Blind speeds occur for all target Doppler frequencies that are the exact PRF, or any multiple of the PRF, of the radar signal. When a target is moving at a velocity that produces this Doppler frequency, its return is cancelled along with fixed returns.

(3) There are three techniques to limit “blind speeds.” The first technique is to use a PRF stagger. By staggering the PRF, the blind speed associated with one PRF will be covered by the other PRF. The second method is called the delay line and canceler. This involves delaying each pulse so it can be compared to the next pulse before processing. This method enhances Doppler frequency comparison and rejects clutter more effectively. The third way is to use range gates and Doppler filters. A range gate is simply a switch that opens for a time corresponding to the time a radar return would arrive for a target at a specific range. For example, a range gate for all targets between 10 and 11 nautical miles would open 124 microseconds after the transmitted pulse (12.4 microseconds per mile) and close 12.4 microseconds later. A target return at this range would trip

this gate and be processed by Doppler filters to find velocity. Fixed targets would trip the range gate and be eliminated by the Doppler filters. This is a sophisticated technique that is also used by pulse Doppler radars.

8. PULSE DOPPLER RADAR

Pulse Doppler radars combine the advantages of both pulse and Doppler radar systems. Because the signal is pulsed, the radar can find range, azimuth, and elevation, similar to a conventional pulsed radar. A pulse Doppler radar can also compute overtake, or rate of closure, compared with the radar system on a pulse-to-pulse basis. A pulse Doppler radar operates much like an MTI, and the terms are sometimes used interchangeably. However, an MTI uses Doppler frequency shifts only to reject clutter while a pulse Doppler radar uses Doppler frequency shifts to reject clutter and to track targets in velocity. A pulse Doppler radar transmits a box or pulse of RF energy at the operating frequency of the radar (Figure 7-22). The frequency inside these boxes reacts the same way as the continuous waves of a CW radar, but since the RF waves are pulsed, range determination can be accomplished by measuring the time it takes for the reflected pulse to return from the target. Velocity determination and tracking are accomplished by capturing and quantifying the Doppler shift of the frequencies in each pulse.

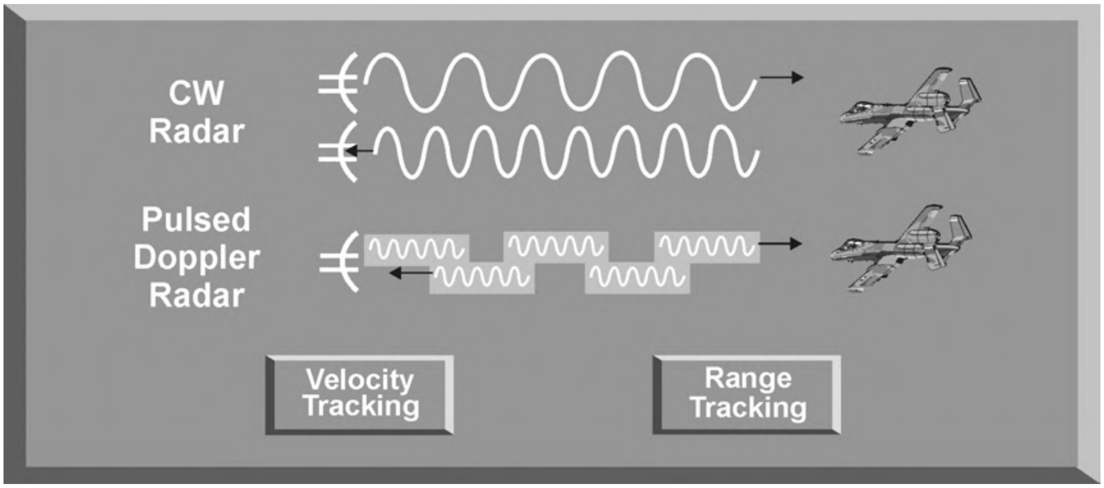

Figure 7-22. Comparison of CW Radar and Pulse Doppler Radar

a. A pulse Doppler radar tracks a single target in azimuth and elevation by employing either conical scan (sequential lobing) or monopulse tracking. Angle and elevation tracking employing these techniques is covered in Sections 3 and 6 of this chapter. The error signals generated by these techniques are sent to the antenna servos to keep the target return centered in the antenna beam.

b. Range tracking of a single target in a pulse Doppler radar is normally accomplished by a split-gate or leading-edge range tracking loop. Some pulse

Doppler radars employ an FM technique to provide range information during high PRF operations.

c. For velocity tracking, each range gate has a complete set of Doppler filter banks as depicted in Figure 7-23. Each pulse of RF energy is composed of many frequencies. To separate the returning target frequency from all other frequencies in the returning waveform, the pulse Doppler radar employs filters to cancel the unwanted frequencies. In addition, it cancels out all returns with no frequency shift, which equates to canceling all returns with no movement relative to the radar.

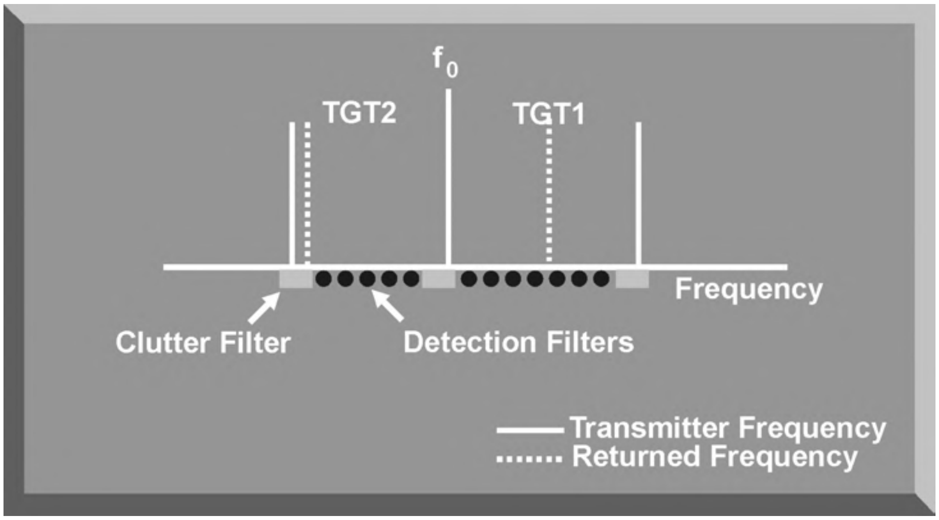

Figure 7-23. Range Ambiguity

d. The ability of a pulse Doppler radar to accomplish range, azimuth, and velocity tracking is dependent on the PRF of the radar. Table 7-1 summarizes these capabilities based on PRF. Low PRF tracking in velocity is extremely difficult due to the spacing of the spectral lines. Low PRF gives accurate range and azimuth as well as long, unambiguous range. Medium PRF tracking in range, azimuth, and velocity is easy for the radar to handle. High PRF provides excellent velocity resolution, but range ambiguities become a problem.

Table 7-1. Pulse Doppler Tracking Capabilities

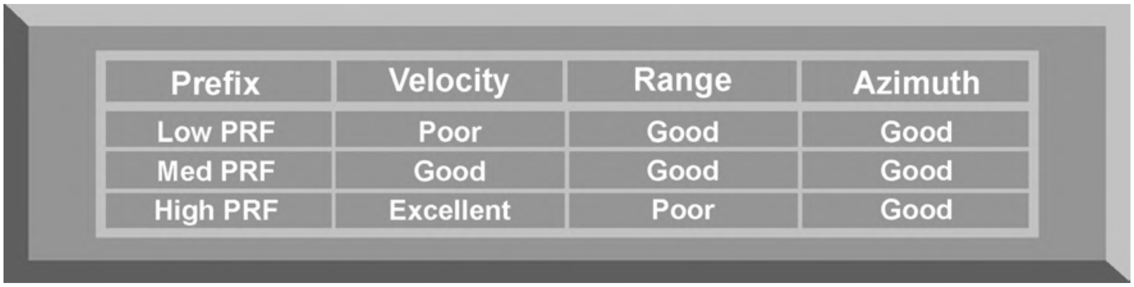

Prefix	Velocity	Range	Azimuth
Low PRF	Poor	Good	Good
Med PRF	Good	Good	Good
High PRF	Excellent	Poor	Good

e. Weaknesses of pulse Doppler radars include velocity blind speeds, range ambiguity, and range eclipsing.

(1) The primary operator exploitable weakness of a pulse Doppler radar takes advantage of the pulse Doppler radar's biggest strength. The pulse Doppler is designed to eliminate ground returns so that the attacker is able to track an aircraft that used to be able to hide in ground clutter. To remove the ground clutter and avoid tracking unwanted targets like cars on a road, a filter is designed in the radar to eliminate all targets with a low velocity relative to the radar. The key to breaking track of a pulse Doppler radar is to place the aircraft in a speed less than the speed gate relative to the radar (Figure 7-24), commonly referred to as the Doppler notch.

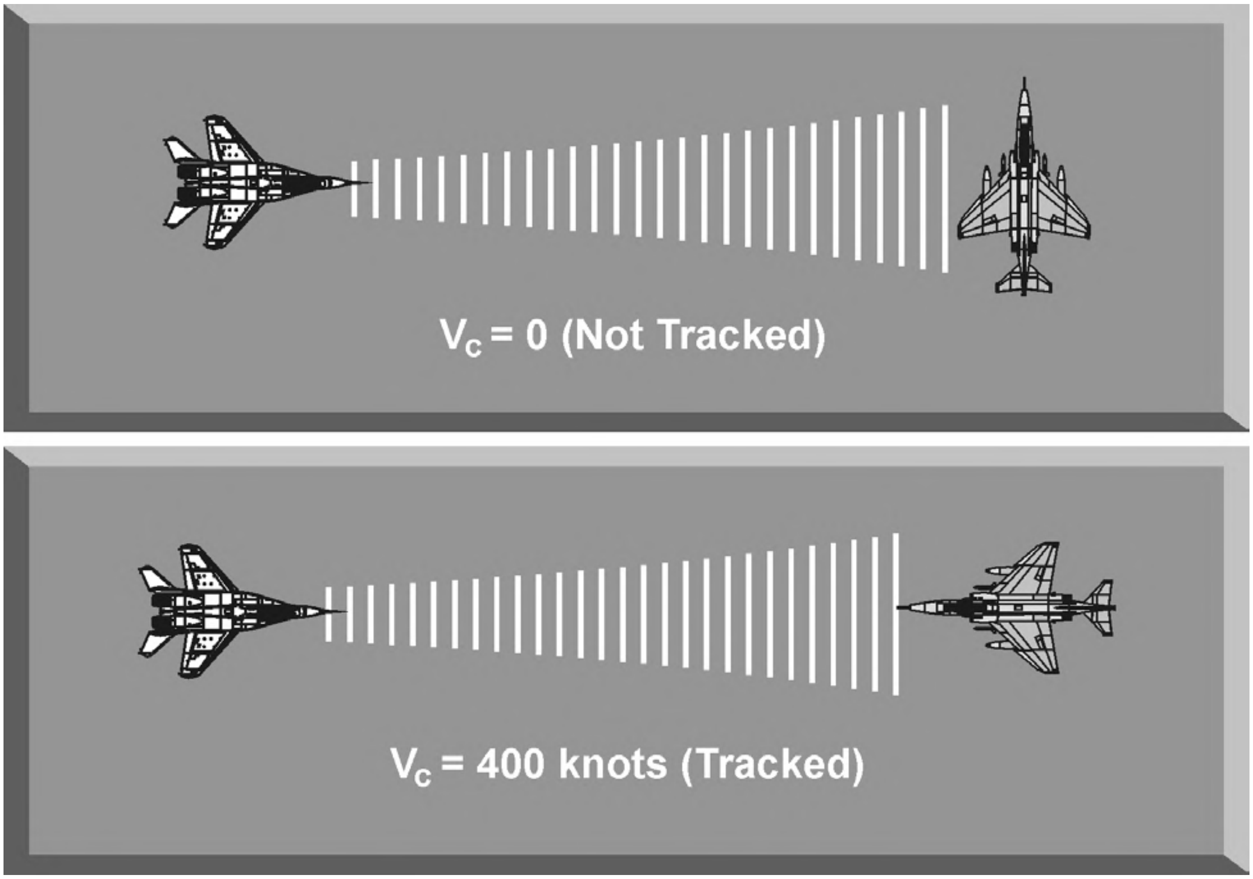

Figure 7-24. Pulse Doppler Speed Gate

(2) Range ambiguity occurs primarily with long-range targets when the return comes back to the radar after another pulse has already been transmitted (Figure 7-25). The radar will see this return and base its range calculations on the transmission time of the immediately preceding pulse, instead of the pulse that generated the return. The result will be an incorrect range calculation.

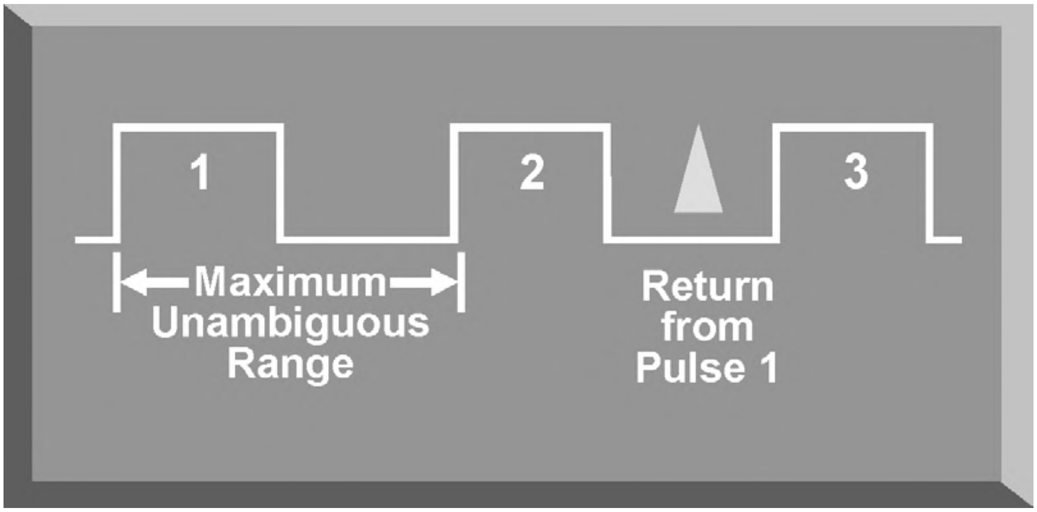

Figure 7-25. Range Ambiguity

(3) Range eclipsing occurs when a target return comes back to the radar antenna while a pulse is being transmitted (Figure 7-26). Since the radar cannot receive while transmitting, the return will not be displayed.

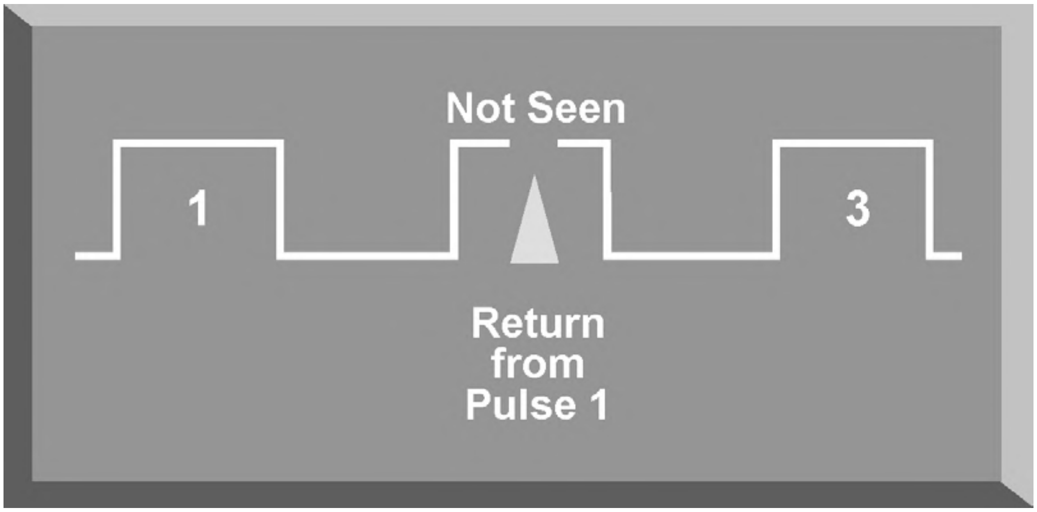

Figure 7-26. Range Eclipsing

(4) To solve the problems of range ambiguity and eclipsing, a pulse Doppler radar employs different PRFs and computer logic. In Figure 7-27, at PRF 1, the return from target 1 is eclipsed, and target 2 is ambiguous. By changing the PRF slightly, these range problems can be resolved. Notice that at PRF 2, neither target is eclipsed, and at PRF 3, target 2 is eclipsed. The computer logic needs to be extremely advanced to compensate for these problems. These simplistic examples show the complex problem of using multiple PRFs.

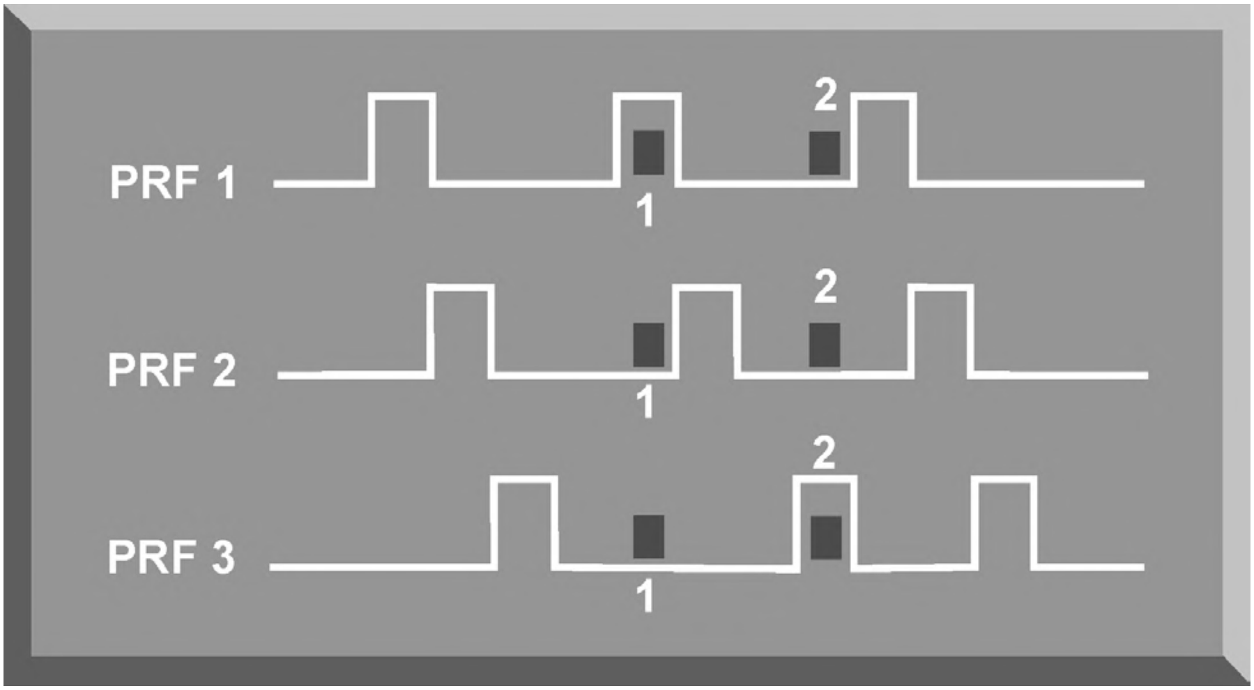

Figure 7-27. Resolving Range Ambiguity and Eclipsing

9. SUMMARY

This chapter has discussed the range, azimuth, elevation, and velocity tracking techniques employed by conical scan, TWS, LORO, monopulse, CW Doppler, and pulse Doppler radar systems. The method used by a specific radar system to track a target determines the type of jamming technique required to counter this system. A basic understanding of the target tracking techniques will enable you to understand the basic jamming techniques employed to defeat these threats.

CHAPTER 8. RADAR MISSILE GUIDANCE TECHNIQUES

1. INTRODUCTION

Once a target has been designated, acquired, and tracked by a radar system, the final stage in target engagement is to guide a missile or projectile to destroy the target. There are three basic requirements for successful missile guidance: (1) precise target tracking by a target tracking radar (TTR) to provide target parameters (range, azimuth, elevation, velocity, etc.), (2) a method to track the position of the missile compared with the target, and (3) a fire control computer to generate missile guidance commands based on target and missile position. The missile guidance techniques employed by modern air-to-air and surface-to-air missile (SAM) systems will be covered in this chapter. In addition, the target engagement techniques employed by antiaircraft artillery (AAA) systems will also be discussed. There are three distinct phases in any missile intercept: boost, mid-course, and terminal.

a. Nearly all missiles are unguided during the initial boost phase (Figure 8-1). During the boost phase, the missile electrical and hydraulic systems are activated and are coming up to operating parameters. The missile is gathering speed and normally will be in an unguided mode of flight.

Figure 8-1. Initial Boost Phase

b. During the mid-course phase, the missile is actively being guided to the target using some type of guidance signal (Figure 8-2). Guidance signals deflect

the control vanes of the missile to change its direction. These vanes change the roll, pitch, and yaw, in some combination, to control the missile flight path. Normally a gas grain generator powers a small hydraulic pump that deflects the control vanes in response to guidance signals. Each missile carries a limited supply of hydraulic fluid for maneuvering. The fluid is expended through vents with every control surface activation. The limited quantity of hydraulic fluid can be a significant factor during a long-range missile intercept.

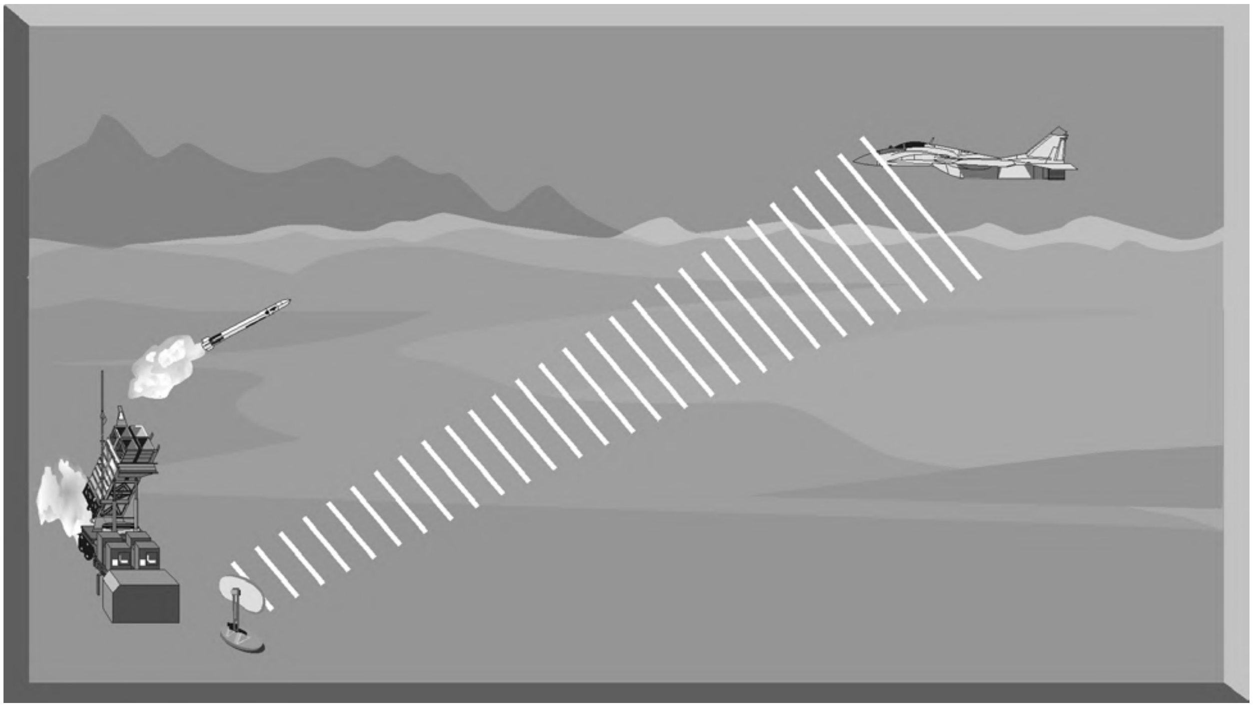

Figure 8-2. Mid-Course Guidance Phase

c. The final phase of an intercept is the terminal phase (Figure 8-3). During this phase, the missile attempts to pass close enough to the target to detonate the fuse while the target is within the lethal radius of the warhead. Modern missiles employ both a contact fuse and some type of proximity fuse. Proximity fuses range from command detonation for command-guided missiles, fractional Doppler gates for semi-active guided missiles, to active laser fuses for IR-guided missiles.

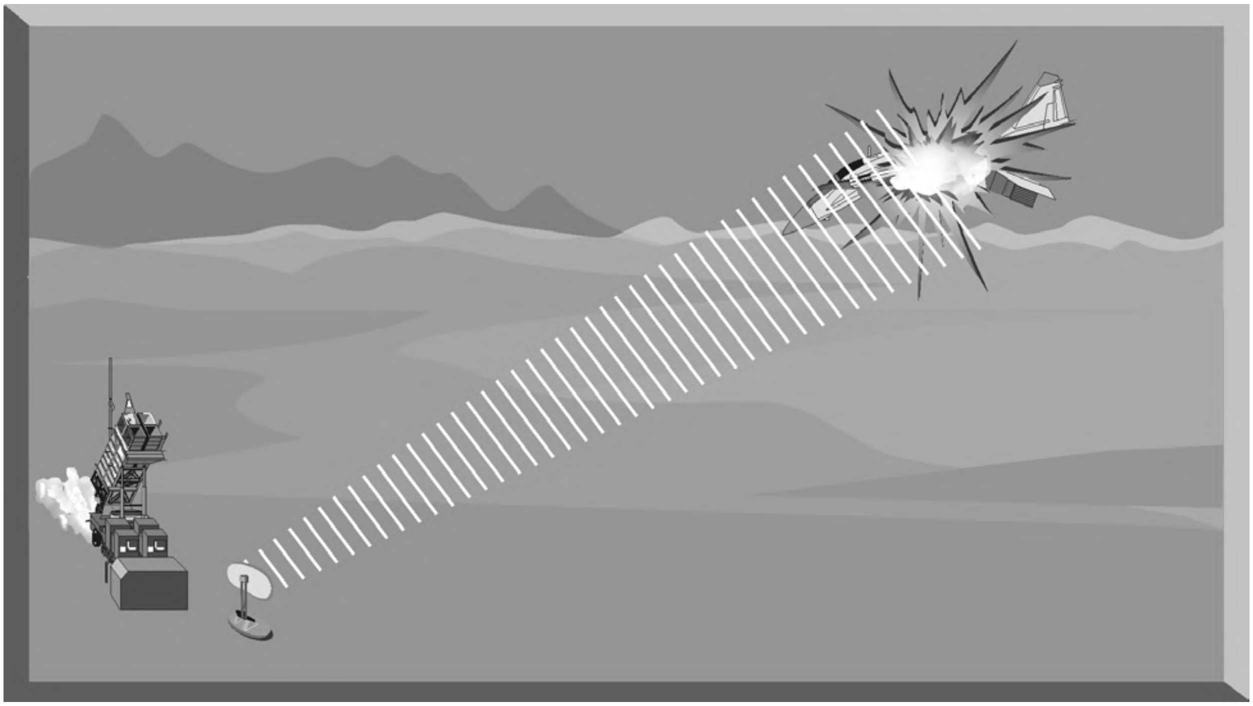

Figure 8-3. Terminal Guidance Phase

2. COMMAND GUIDANCE

Command guidance uses a fire control computer to constantly send course correction commands to the missile throughout its flight. These commands are a series of electrical missile guidance pulses called doublets or triplets. These pulses provide steering commands to the missile by varying the spacing between each guidance pulse. Each pulse, or pulse combination, relays some roll, pitch, and yaw command to the missile. These inputs are constantly corrected for the spatial relationship between the missile and the target's present position and rate of motion. Guidance commands are passed to the missile by specialized antennas on the TTR and an antenna installed on the missile, called a missile beacon. The beacon is a special radio receiver and transmitter that is attached to the rear of the missile. It acts like a transponder in that the TTR tracks and receives guidance commands. The guidance frequency may be widely separated from the target tracking radar frequency to minimize interference. This beacon is usually masked until missile booster separation. This results in the missile being launched unguided for the first 2-3 seconds. This type of delay is one of the reasons that all command-guided missile systems have a minimum launch range. Command guidance is used by the SA-2, SA-3, SA-4, and SA-8.

a. Command-guided missiles will generally fly a rectified (full or half) or three-point pursuit geometry during the mid-course portion of the intercept (Figure 8-4). However, a command-guided missile may transition to a pure pursuit geometry during the terminal phase of the intercept. Rectified geometry involves the prediction of where the target and the missile will be at some point in the future.

The target's direction and rate of movement is tracked and predicted. The missile is then launched, pulls lead on the target, and is guided to the point in the sky where the intercept is predicted to take place. This profile requires the constant update of both the target and missile positions.

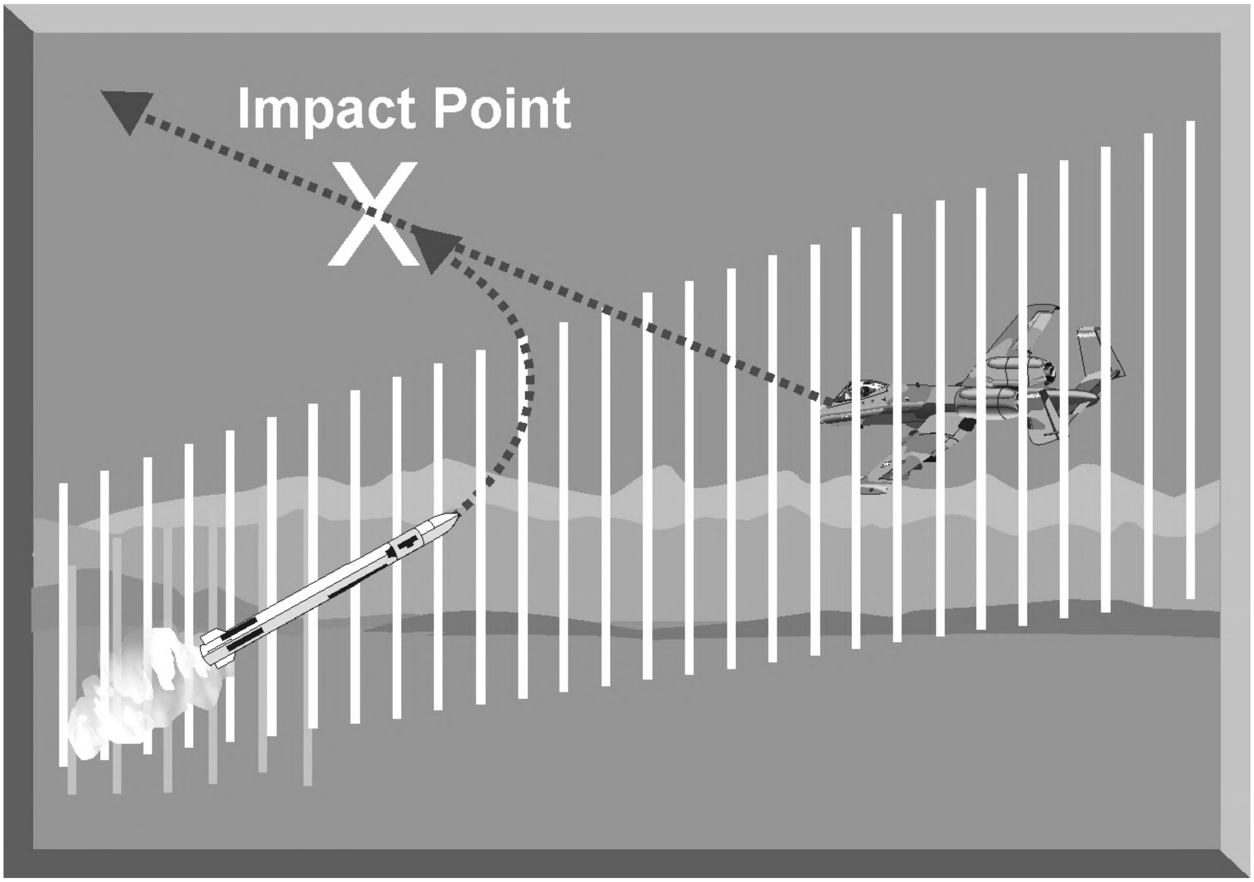

Figure 8-4. Rectified Flight Profile

b. Three-point pursuit geometry is often used when there is incomplete range tracking data on the target. In this case, it will be impossible to predict exactly where the target will be at some point in the future. In this profile, the target tracking radar constantly tracks the target. The missile location will be updated by the missile beacon. The fire control computer will direct the missile to fly directly down the tracking radar beam toward the target. In this geometry, the missile may start out on a direct intercept course and, depending on the target's direction and rate of movement, transition to a pure pursuit intercept. The three points in three-point missile geometry are depicted in Figure 8-5. Point one is the target tracking radar, point two is the missile itself, and point three is the target. By keeping all three points always in a line, the missile will intercept the target at some point, although the range of the target is unknown.

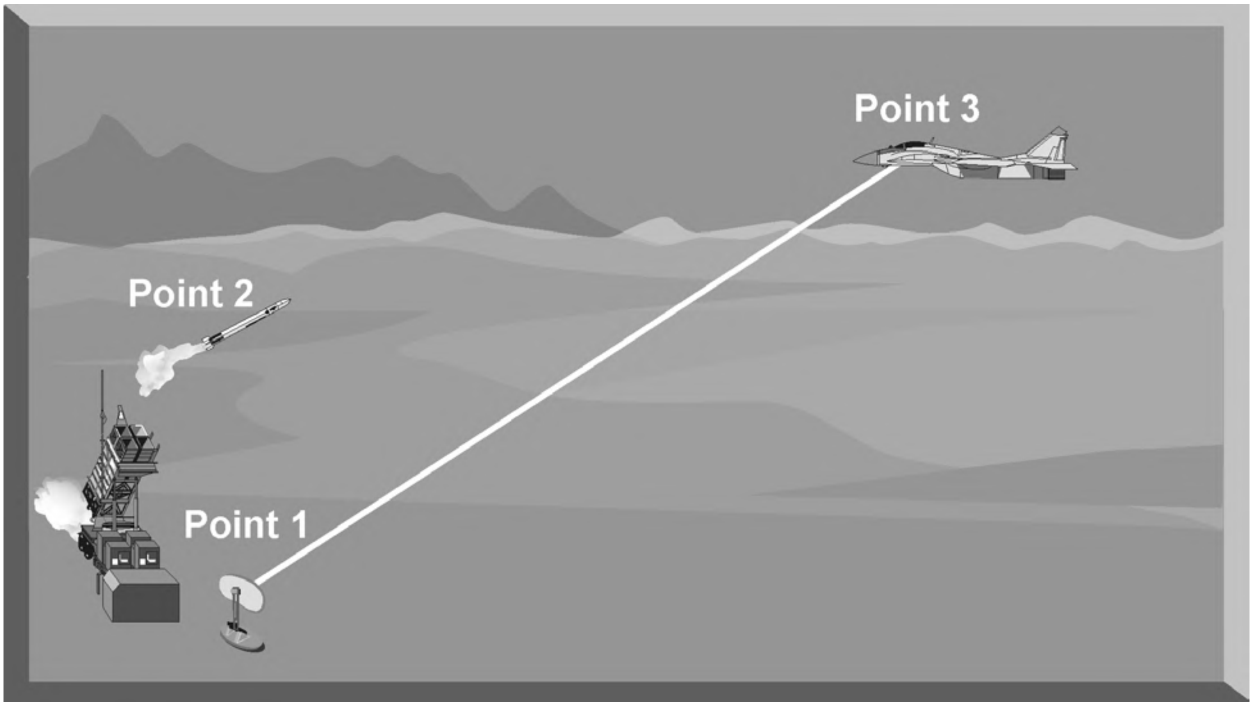

Figure 8-5. Command Guidance: Three-Point Pursuit

c. Command guidance techniques have many advantages. First, command-guided missiles can adjust their flight geometry throughout an intercept profile. Second, the missiles are uncomplicated since they do not carry onboard computers or target tracking equipment. The fire control computer associated with the TTR accomplishes all intercept calculations. Third, the primary intercept profile, a full- or half-rectified intercept, is the fastest and most fuel-efficient intercept. Fourth, command guidance is difficult to jam since the missile beacon antenna is at the rear of the missile and can be relatively high-powered. And finally, an intercept is possible even without accurate range information by using the three point intercept profile.

d. Command guidance has several disadvantages. First, the use of a missile beacon delays the capture of the missile by the tracking radar. This can cause a large dead zone which equates to a larger minimum engagement range. Second, the accuracy of the intercept geometry is only as good as the tracking information provided by the target tracking radar. Jamming, interference, or loss of signal will adversely affect the intercept accuracy. In addition, normal radar characteristics could produce sufficient errors to cause the missile to miss the target, especially at longer ranges. Third, with insufficient range information, the three-point intercept profile is very slow and could result in the missile running out of energy before it gets to the target. Fourth, command guidance is reactive. The fire control computer constantly updates the intercept geometry based on target maneuvering. This results in missile maneuvering lagging target maneuvers.

3. SEMI-ACTIVE GUIDANCE

Semi-active guidance is significantly different from command guidance, but only after launch. The first requirement is still for the target tracking radar to maintain a solid target track, with the tracking data being supplied to the fire control computer. The fire control computer then directs a target illumination antenna to point at the target and illuminates it with CW energy. The missile then passively homes on the reflected CW energy.

a. The missile used by a threat system that uses CW homing is vastly different from the missile being guided by a command guidance signal. The missile that homes on CW energy must be equipped with a seeker section composed of an antenna and an internal receiver. The seeker section processes and computes the necessary course corrections as it flies toward the target. It can do this by knowing the zero boresight line of the antenna within the missile (Figure 8-6). As the reflected CW energy is received by the seeker, there is normally some deviation from the zero reference position. The onboard computer then directs the control surfaces to change the flight path to reduce the reference errors in the antenna to zero, if possible. When the error between the antenna position and the boresight position is zero, the missile is pointed directly at the target.

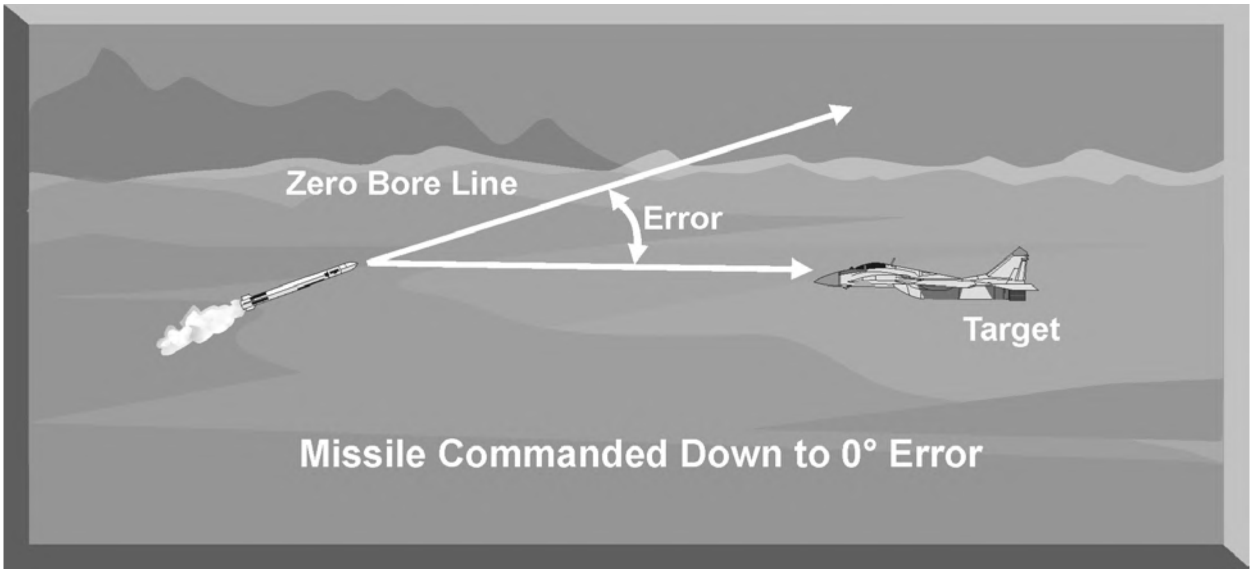

Figure 8-6. Semi-Active Guided Missile

b. Missile systems that use semi-active guidance normally use velocity as the primary target discriminator during the intercept. The missile seeker locks onto a reference Doppler signal provided by the fire control computer before launch. This Doppler signal establishes a tracking gate around the velocity of the target. After the missile is launched, it initially compares the reference Doppler to the target Doppler signal.

c. The mid-course phase for a semi-active missile is also different from that of a command-guided missile. A semi-active guided missile follows the reflected CW energy during the mid-course phase of the intercept and normally attempts to fly a lead pursuit profile to the target (Figure 8-7). If the target maneuvers, however, the missile may transition to a pure pursuit flight path. Unlike a command-guided missile, a semi-active guided missile does not use a missile beacon. The fire control computer does not need to know where the missile is to compute course corrections since all that is necessary is to illuminate the target with the CW illuminator. This also means that the missile can begin to track and guide when it is launched and locked on to the reference Doppler gate. Semi-active guidance is the primary mode of guidance for many surface-to-air missiles, and almost all radar-guided air-to-air missiles.

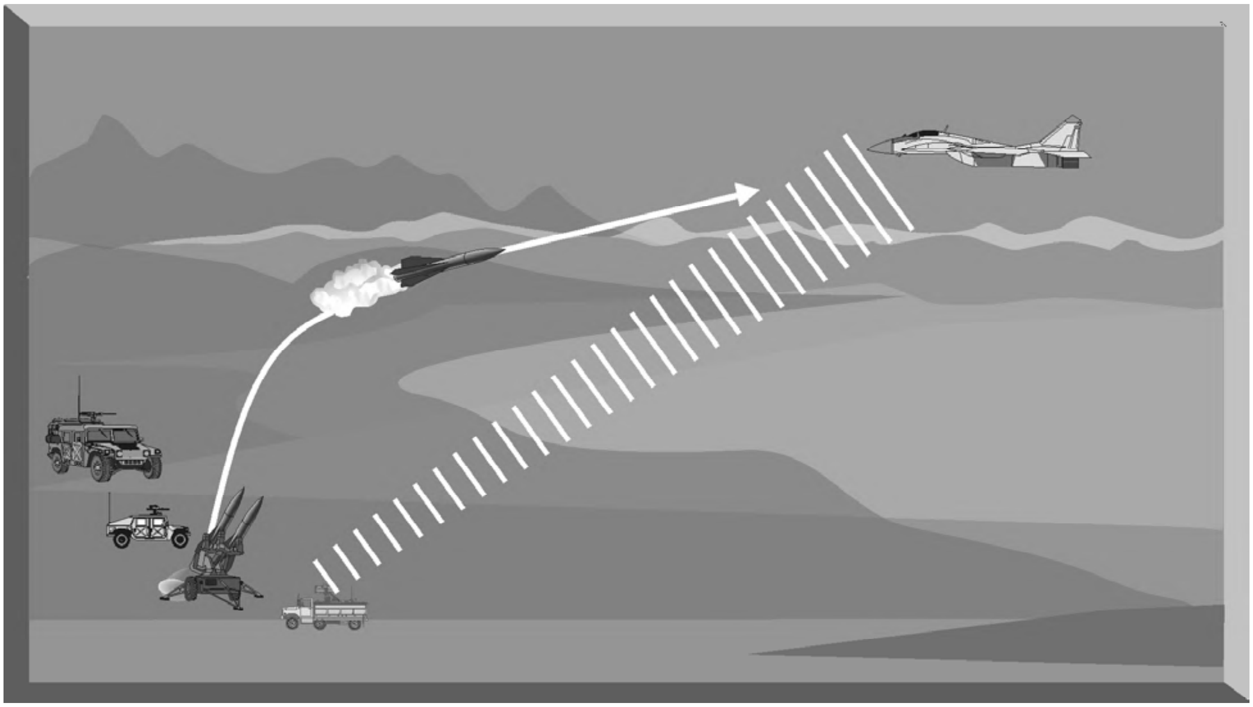

Figure 8-7. Semi-Active Guidance (Mid-Course)

d. As the missile enters the terminal phase of the intercept, there is no change in the guidance mode used by a CW homing missile. The missile may complete the terminal phase of the intercept geometry by going to a pure pursuit flight path, if necessary (Figure 8-8). The missile continues to home in on the reflected CW signal until it passes close enough for the fuse to function.

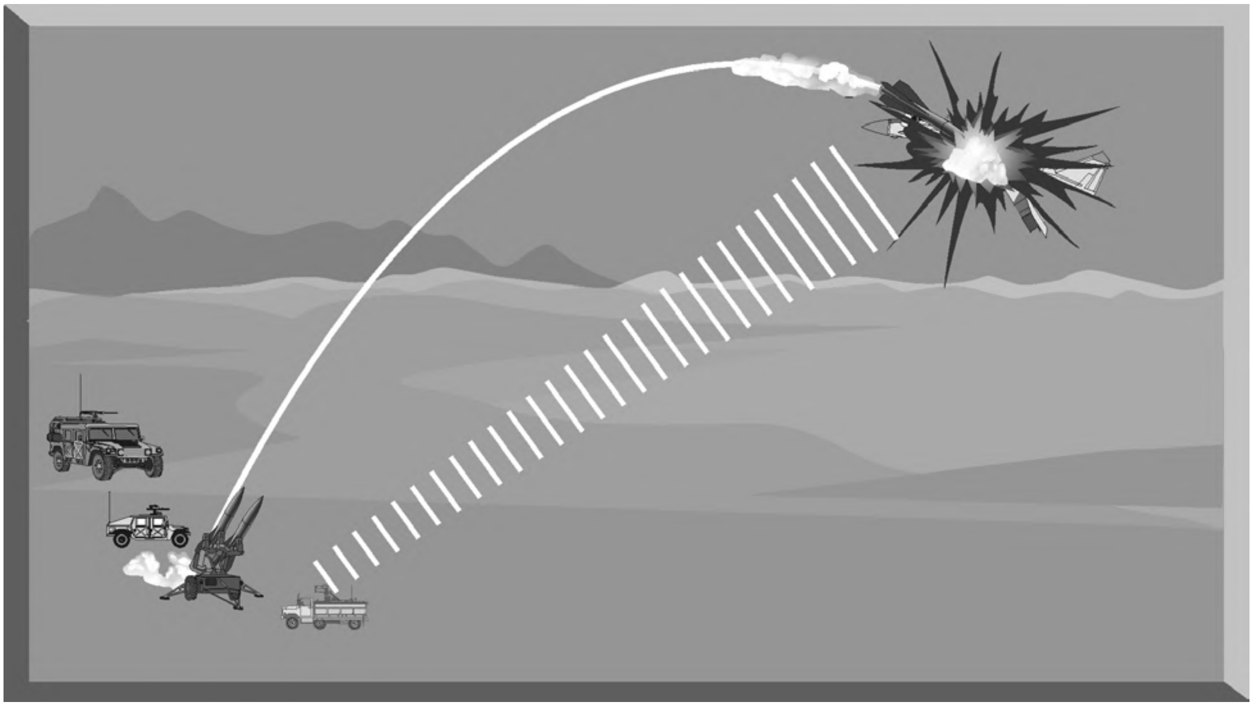

Figure 8-8. Semi-Active Guidance (Terminal Phase)

e. Semi-active missile guidance has many advantages. First, a semi-active guided missile is resistant to electronic jamming that may be used to deny range information. Second, a semi-active missile can be guided almost immediately after launch. This gives it a very small minimum range since it can maneuver almost as soon as it clears the launch rail. Third, it computes its own course corrections as necessary. This allows for a much quicker reaction to target maneuvers compared to a command-guided missile. Fourth, during a long-range intercept, a CW missile can be more accurate than a command-guided missile. This is accomplished by taking the inherent long-range radar tracking errors out of the equation. The target tracking radar only has to keep the target illuminated so that it can point the CW antenna at the target.

f. Although semi-active missile guidance is generally considered an excellent guidance technique, it does have some disadvantages. First, a semi-active guided missile normally requires reference Doppler values to be entered into the missile computer before launch. Without this reference, a semi-active missile cannot be launched (Figure 8-9). Second, a semi-active homing missile must maintain a lock onto the target Doppler. The use of chaff and beam maneuvers, which result in a near zero target Doppler, may cause a missile or radar to break lock. Third, if a break-lock occurs, a CW homing missile normally cannot regain target track and complete the intercept.

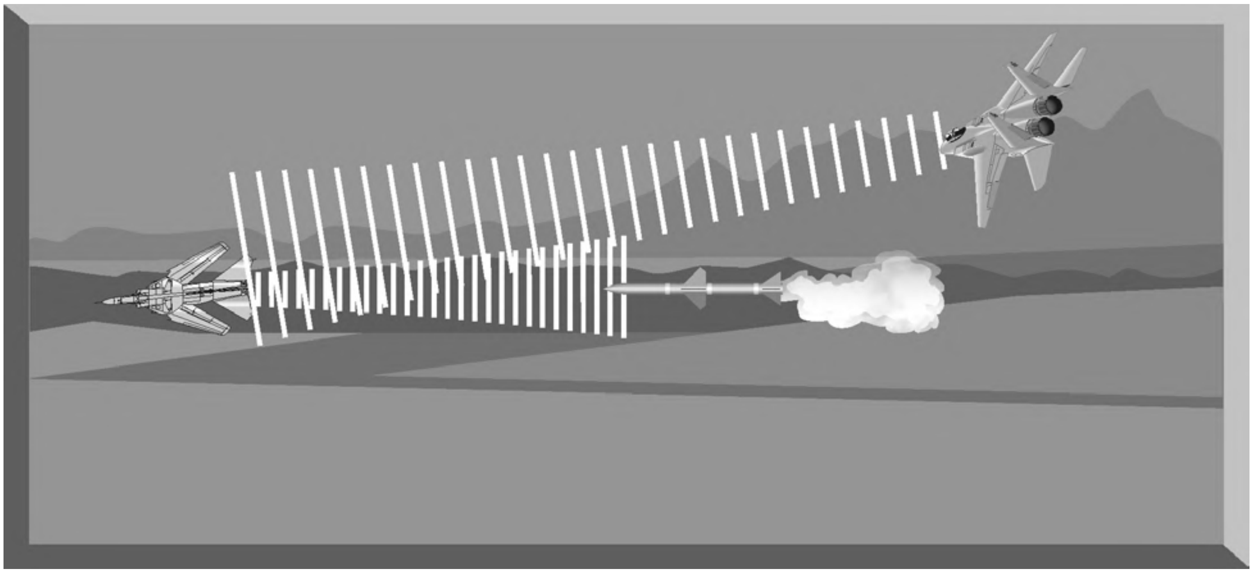

Figure 8-9. Semi-Active Guidance CW Reference Signal

4. ACTIVE GUIDANCE

Active guidance is an improvement that has been included in several new long-range missiles such as the AIM-54 Phoenix and the AIM-120 AMRAAM. This specialized guidance mode is only active during the terminal phase of flight. The mid-course phase usually employs semi-active or command guidance (Figure 8-10).

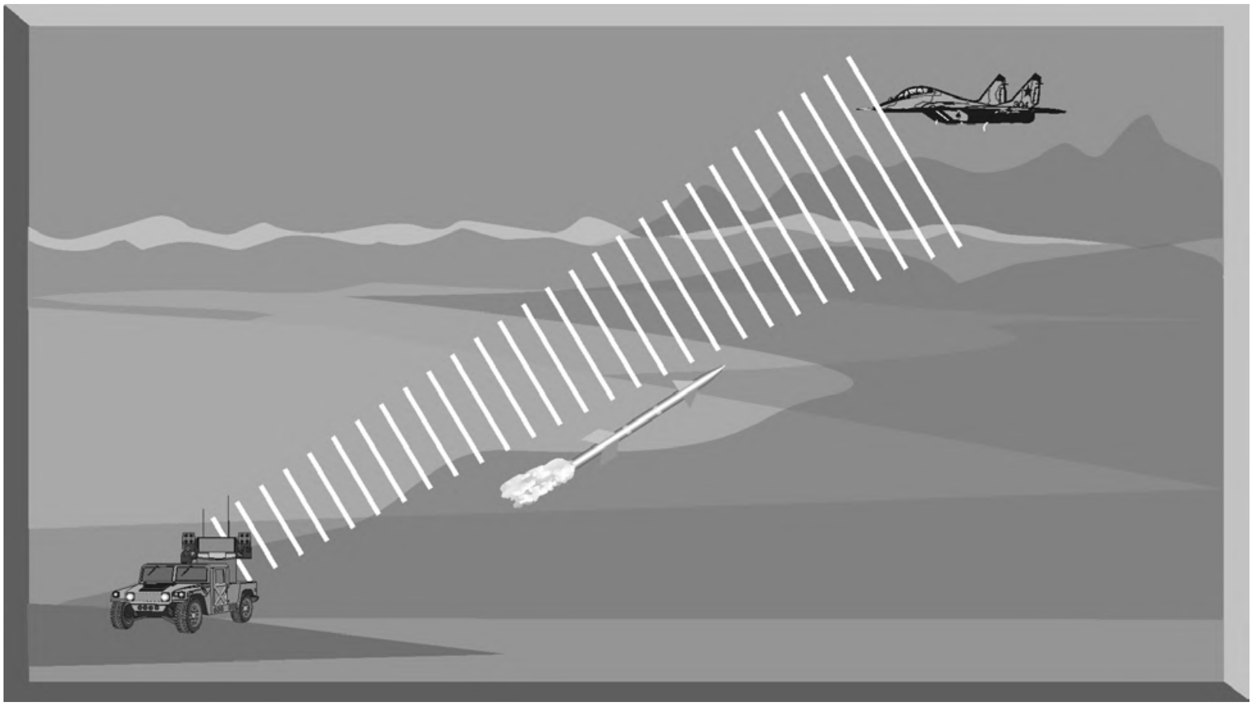

Figure 8-10. Active Guidance Mid-Course Intercept Phase

a. The range at which the missile goes "active" is dependent on the intercept geometry. High-aspect angle intercepts allow the activation of active guidance sooner than beam or tail-aspect intercepts. Missiles that employ active guidance carry a complete miniature radar system and fire control computer within the missile. As the missile nears the target, its internal radar system turns on and locks onto the target. The internal fire control computer directs control inputs to complete the intercept (Figure 8-11).

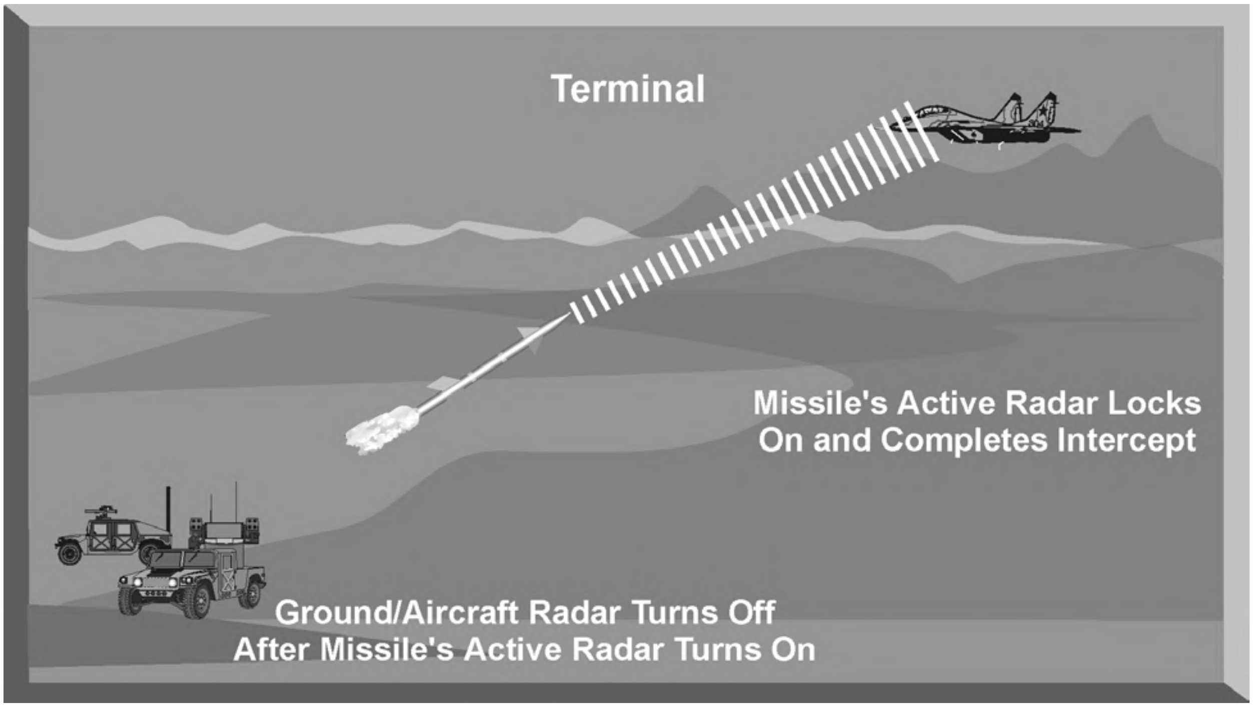
Terminal
Missile's Active Radar Locks On and Completes Intercept
Ground/Aircraft Radar Turns Off After Missile's Active Radar Turns On

Figure 8-11. Active Guidance Phase

b. Active-guided missiles have many advantages. First, active-guided missiles are very accurate at long ranges. This is because they do not rely on the target tracking radar once their internal radar takes over the intercept. Second, an active missile is extremely difficult to jam. It uses a narrow beam and its relative power is constantly increasing as it nears the target. Third, an active-guided missile is a fire-and-forget weapon. Command or semi-active missile guidance requires the target tracking radar to maintain lock-on until the intercept is completed. In an air-to-air engagement, this means the interceptor is predictable until the missile hits the target, and vulnerable to an enemy missile attack. An interceptor with an active missile, however, may launch the missile and, once it goes "active," can then turn around or maneuver defensively.

c. Active-guided missiles have a few disadvantages as well. First, the active homing missile is a complex missile integrating both command and active guidance modes. Second, the missile may still be susceptible to electronic jamming during the mid-course phase of flight. Remember, during the mid-course

phase, the missile relies on command or semi-active guidance. Jamming the target tracking radar may affect the missile's ability to "see" the target near the terminal phase.

5. SEEKER-AIDED GROUND GUIDANCE/TRACK-VIA-MISSILE GUIDANCE

In seeker-aided ground guidance (SAGG) and track-via-missile (TVM) guidance, the target is illuminated by the ground-based radar and the missile receives reflected energy from the target. Unlike conventional semi-active homing, the missile does not generate its own guidance commands. Instead, the missile transmits raw engagement data to the ground-based fire control system (FCS) in order to generate uplink guidance commands. TVM is similar to SAGG; however, additional processing is done on-board the missile prior to transmitting the engagement data to the ground-based FCS.

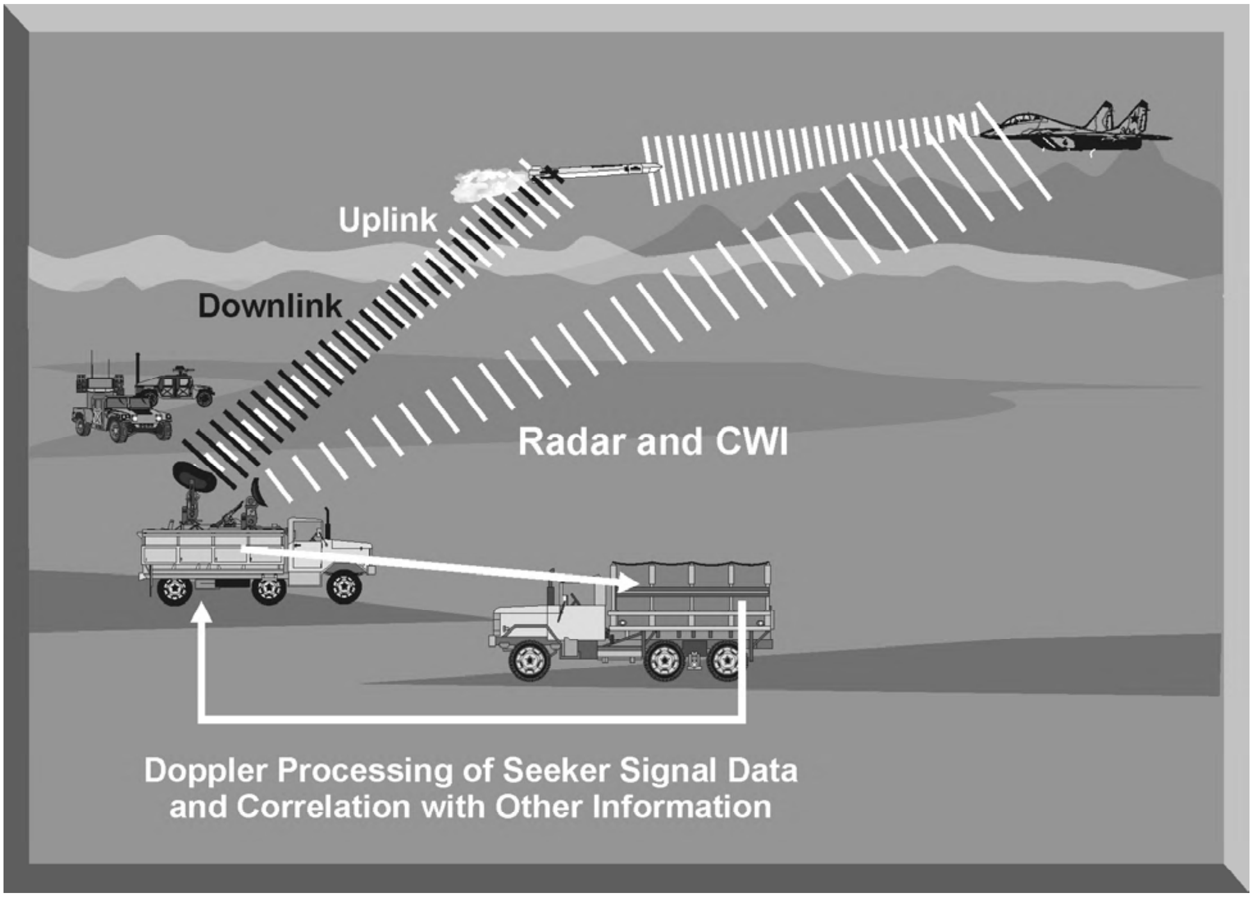
Uplink
Downlink
Radar and CWI
Doppler Processing of Seeker Signal Data and Correlation with Other Information

Figure 8-12. SAGG/Track-Via-Missile Guidance

a. Track-via-missile and seeker-aided ground guidance are two relatively new missile guidance techniques with similar advantages. First, they are extremely accurate at long ranges where the inherent radar tracking errors may be large

enough to cause a miss. Second, they can respond very quickly to any actions taken by the target since the missile seeker can track these changes and transmit the new position to the TTR fire control computer. Third, TVM and SAGG can be used with a large and capable fire control computer since most computations are accomplished by the TTR. Fourth, the integration of a phased array radar and the powerful TTR fire control computer allows the missile system to engage multiple targets. The Patriot missile battery, for example, can track and engage at least four targets simultaneously.

b. The major disadvantage of track-via-missile and seeker-aided ground guidance is that they are the most complex forms of missile guidance. They require the use of sophisticated computers to combine radar tracking data and data received from the missile. This required hardware is expensive and demands greater maintenance and logistical support. In addition, the missile itself needs to be large enough to store the appropriate hardware for computations and data transfer.

6. ANTIAIRCRAFT ARTILLERY (AAA)

The classic role of AAA is point defense. AAA systems provide close-in defense for high-value targets. AAA systems are deployed to defend cities, airfields, bridges, industrial centers, lines of communications, command and control centers, infantry/tank units, and SAM sites. There are two types of AAA systems: towed and mobile. Towed AAA is normally deployed in fixed sites around key targets. Mobile AAA systems are deployed to provide air defense for army units and to protect mobile SAM sites. The effectiveness of AAA systems, towed or mobile, depends on the ability of the system to predict an aircraft's future position to fire its unguided ballistic projectile to intercept the aircraft and destroy it. To accomplish this objective, AAA systems employ two primary tactics, aimed fire and sector/barrage fire.

a. Aimed AAA fire requires very accurate aircraft position information and an accurate prediction of future position. For aimed AAA fire, this information can be derived by using an optical sighting system on the gun or by employing a radar system coupled with a fire control computer. Smaller caliber AAA guns generally rely on optical target acquisition and firing (Figure 8-13). The high rate of fire, short range, and short projectile time of flight (TOF) simplifies the prediction and aiming problem for these systems. Smaller caliber AAA can also use tracer ammunition to help the gunner in correcting his optical firing solution. Larger caliber AAA systems, with slow rates of fire, long range, and long projectile TOF, generally use a TTR and a fire control computer to solve the problems associated with aimed fire (Figure 8-14).

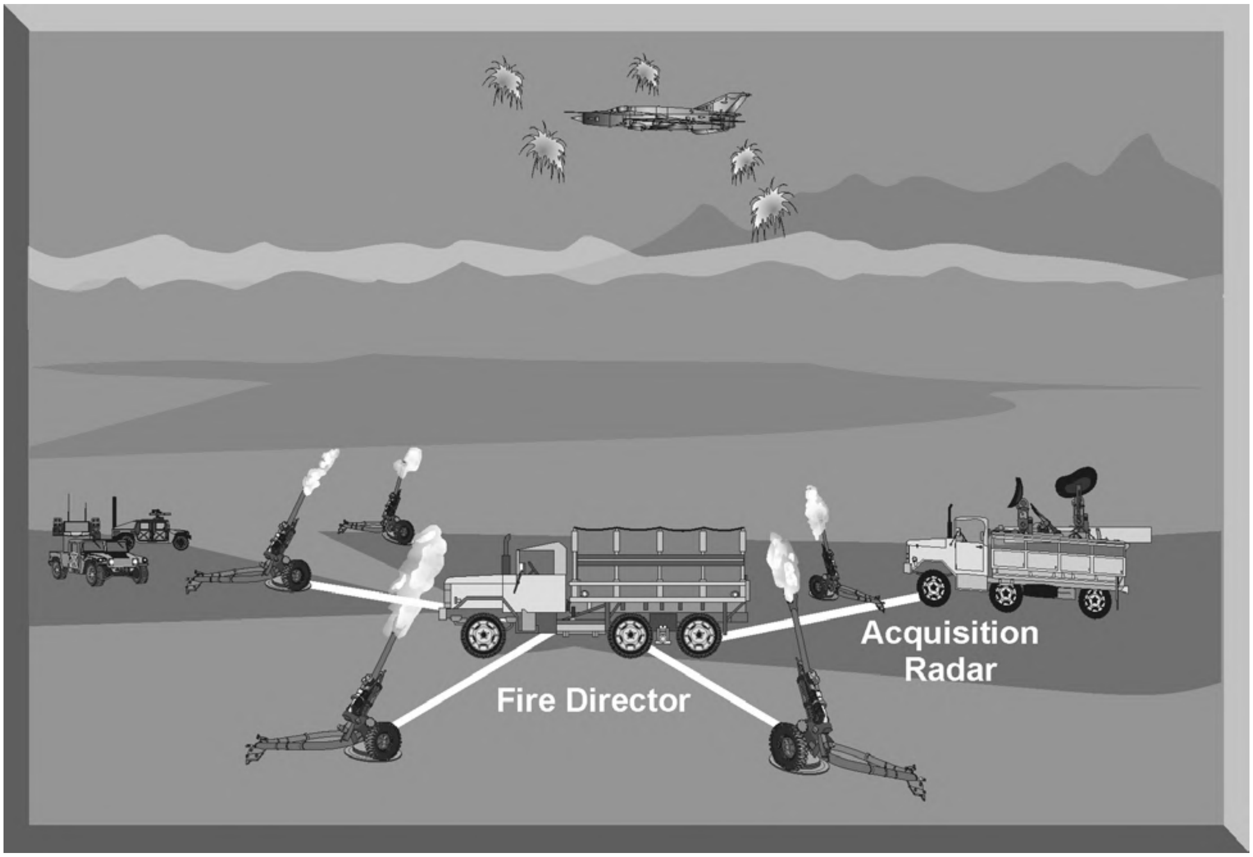

Figure 8-13. Optically Aimed AAA

(1) The typical engagement sequence for an aimed AAA engagement employing a TTR and fire control computer begins with initial target data from an acquisition radar. The guns and TTR are pointed toward the target. The TTR initiates search and lock-on to the target. The TTR associated with large caliber AAA is usually a conical scan radar to provide accurate target positioning information. Target information is fed into the fire control computer which calculates the aim point, points the guns, and initiates firing. The fire control computer uses the target kinematic data, gun ballistics, wind, air density, and projectile dispersion pattern to compute the required aim point. All these computations are based on the assumption that the target will continue on the same heading, at the same altitude, and at the same airspeed during the projectile TOF.

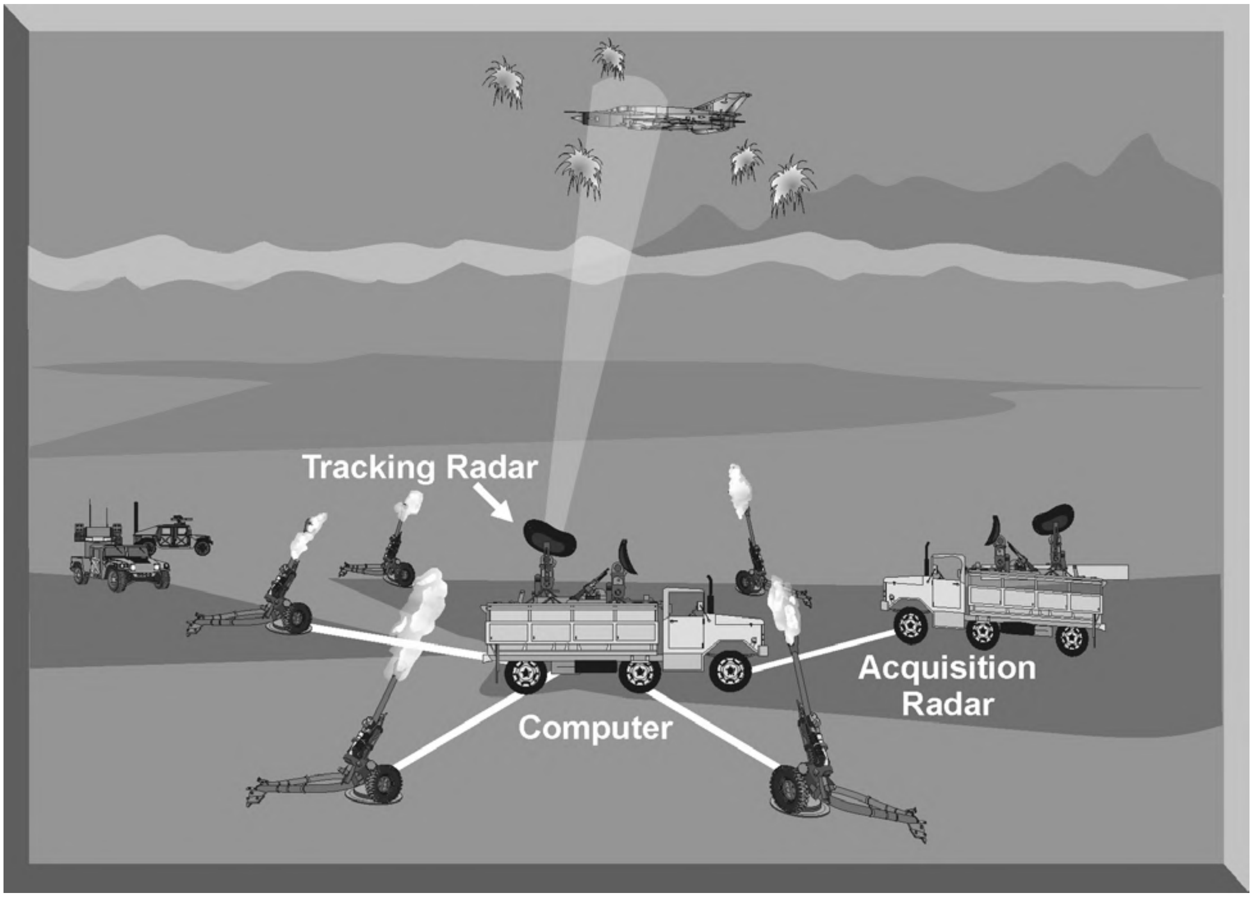

Figure 8-14. Radar-Directed AAA

(2) The typical engagement sequence for an aimed AAA engagement, employing optical target tracking begins with initial target information from an acquisition radar to the fire director. The fire director gives gross aiming commands to the individual guns. The gunners then visually search for the target and use the on-carriage gun sights to predict the required lead angle and initiate firing.

b. Sector or barrage fire tactics are employed when the aircraft cannot be accurately tracked (Figure 8-15). Acquisition information suggests an aircraft will traverse a volume of airspace or a specific sector. The fire director instructs the gunners to fire randomly into this sector in an effort to hit the aircraft with the barrage of AAA fire, or have the aircraft fly into a “curtain” of AAA fire. This tactic is especially effective for point defense for a fixed target. Attacking aircraft may have to fly a predictable flight path during weapons delivery. Sector/barrage fire can be directed to cover the expected attack directions and altitudes.

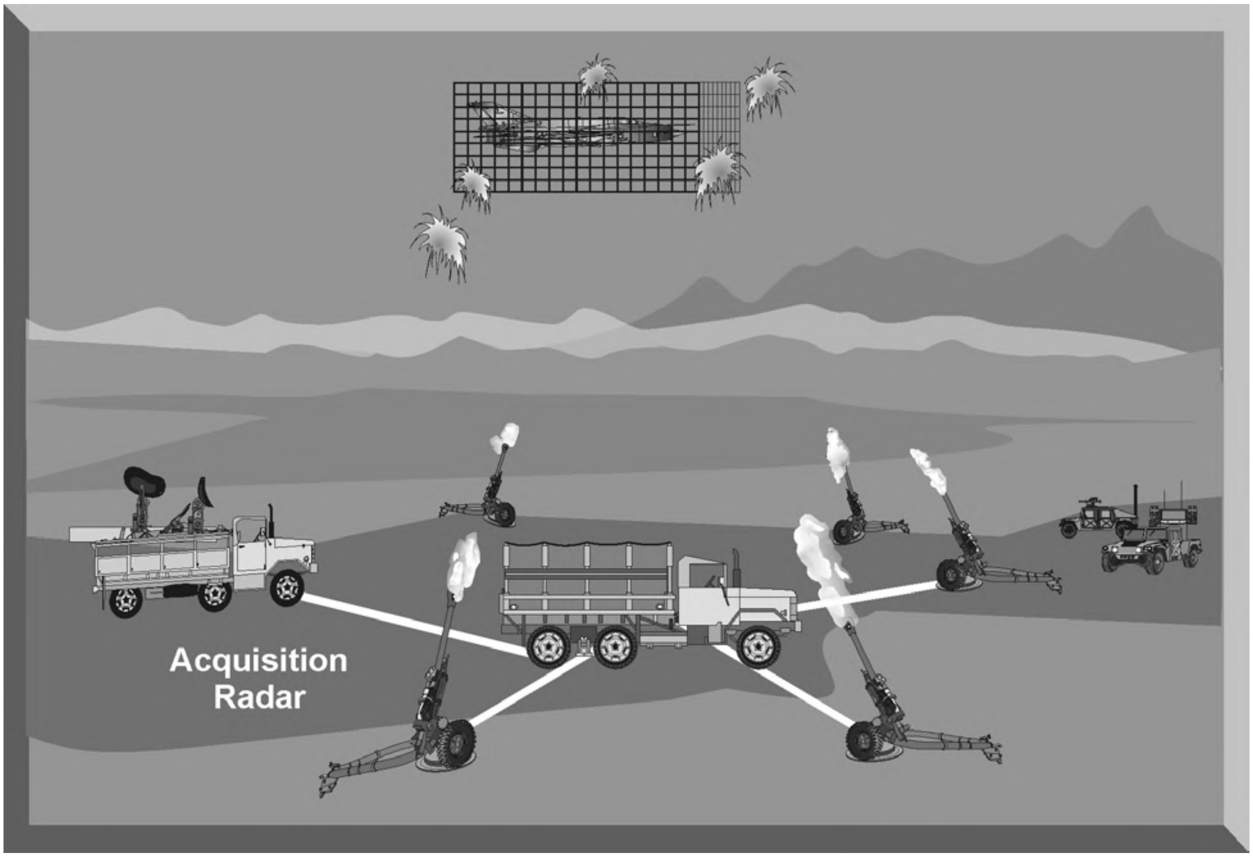

Figure 8-15. Sector/Barrage AAA

7. SUMMARY

This chapter has discussed the most common missile guidance techniques and AAA firing modes used by modern threat systems. A familiarity with the guidance technique employed by specific threat systems is the key to understanding the jamming techniques, chaff/flare employment settings, and tactical maneuvers designed to counter these systems.

CHAPTER 9. INTRODUCTION TO RADAR JAMMING

1. INTRODUCTION

Radar jamming is the intentional radiation or reradiation of radio frequency (RF) signals to interfere with the operation of a radar by saturating its receiver with false targets or false target information. Radar jamming is one principal component of electronic combat (EC). Specifically, it is the electronic attack (EA) component of electronic warfare (EW). Radar jamming is designed to counter the radar systems that play a vital role in support of an enemy integrated air defense system (IADS). The primary purpose of radar jamming is to create confusion and deny critical information to negate the effectiveness of enemy radar systems. This chapter will introduce the two types of radar jamming, the three radar jamming employment options, and discuss the fundamental principles that determine the effectiveness of radar jamming.

2. RADAR JAMMING TYPES

There are two types of radar jamming: noise and deception.

a. Noise jamming is produced by modulating a RF carrier wave with noise, or random amplitude changes, and transmitting that wave at the victim's radar frequency. It relies on high power levels to saturate the radar receiver and deny range and, occasionally, azimuth and elevation information to the victim radar (Figure 9-1). Noise jamming takes advantage of the extreme sensitivity of the radar receiver and the transmission pattern of the radar antenna to deny critical information to the victim radar.

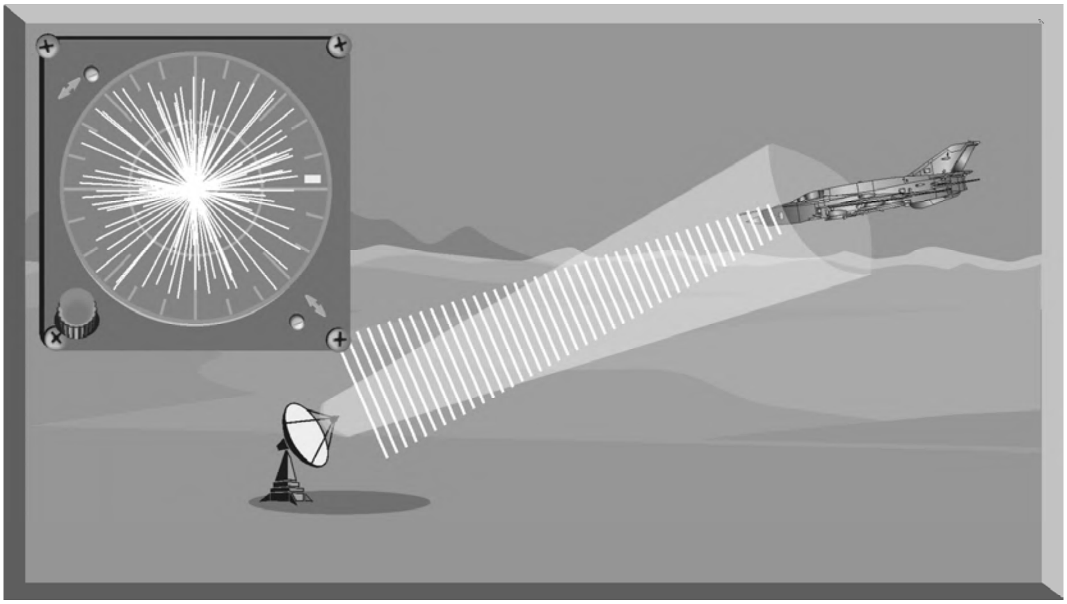

Figure 9-1. Noise Jamming

b. Deception jamming uses complex receiving and transmitting circuits to process and retransmit jamming pulses that appear as a real target to the victim radar. A deception jammer receives the signal from the victim radar and alters the signal to provide false range, azimuth, or velocity information. The altered signal is then retransmitted (Figure 9-2). The victim radar processes this signal, which disrupts the victim radar and confuses the radar operator. To be effective, deception jamming must match not only the victim radar's operating frequency, but all the other operating characteristics, including pulse repetition frequency (PRF), pulse repetition interval (PRI), pulse width, and scan rate.

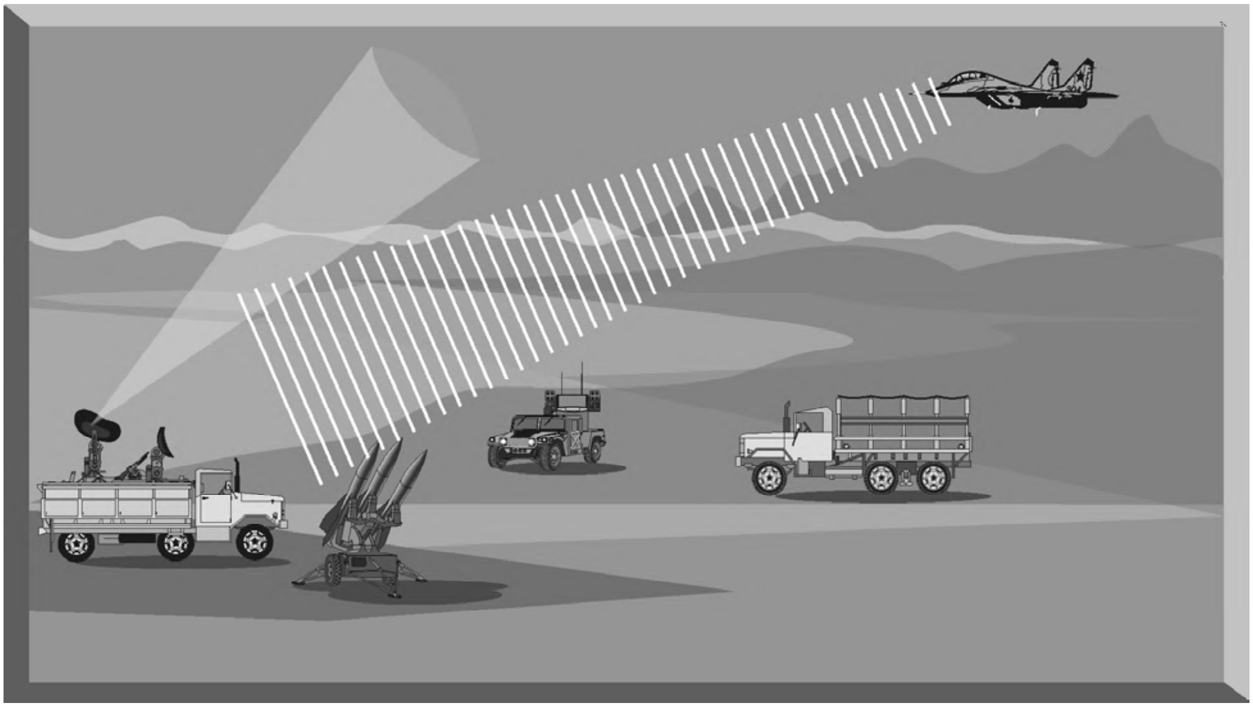

Figure 9-2. Deception Jamming

c. Both noise and deception jamming effectiveness are heavily dependent on another component of EW, specifically, electronic warfare support (ES). ES assets, either airborne or ground-based, provide the threat system specific radar parametric data and update this critical information based on observed threat system operations. This data provides the foundation for developing noise and deception jamming techniques. Intelligence and engineering assessment of this data are used to identify specific threat system weaknesses that can be exploited with the optimum noise, deception, or combination of jamming techniques. This information is then programmed into jamming systems to counter specific threats.

3. RADAR JAMMING EMPLOYMENT OPTIONS

There are currently two primary employment options for both noise and deception jamming techniques. These options are: (1) support jamming, and (2) self-protection jamming. Support jamming can be broken down further into stand-off jamming (SOJ), and escort jamming.

a. To counter early warning, ground control intercept (GCI), and acquisition radars associated with an enemy IADS, noise and deception jamming techniques are employed by specialized support jamming aircraft. The goal of support jamming is to create confusion and delays within the command and control structure of the IADS. Deny, delay or degrade the enemy's ability to engage friendly forces. Support jamming operations can be focused against a national level IADS through the use of a stand-off jamming (SOJ) profile (Figure 9-3) or against a target area threat array using an escort jamming profile.

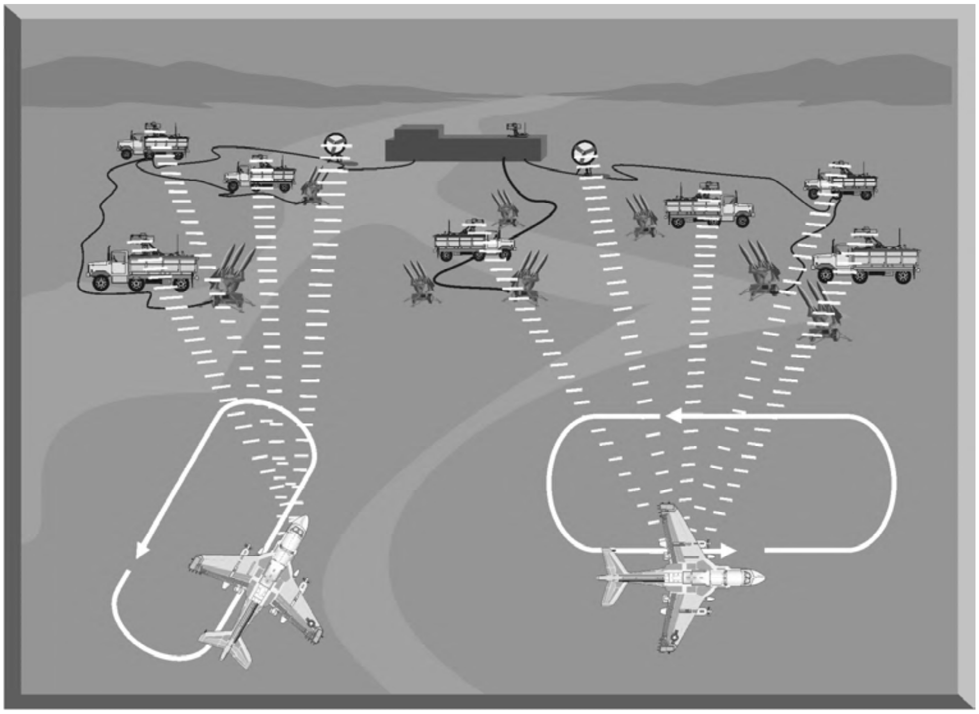

Figure 9-3. Stand-Off Jamming

(1) From an orbit area outside the surface-to-air missile (SAM) engagement zone, SOJ aircraft employ specialized jamming techniques to deny the enemy information about the attack package. SOJ aircraft employ specialized noise jamming techniques to generate jamming strobes on the victim radar display. This effectively denies range and azimuth information on aircraft ingressing and egressing the area covered by the noise jamming strobes. Intensity of the strobes is based on the power in the jamming. The area covered is based on the amount of jamming that can be injected into the main beam and sidelobes of the victim

radar. The effectiveness of SOJ noise jamming is determined by the power the jammer can generate relative to the power the victim radar can generate. This is called the jamming-to-signal (J/S) ratio.

(2) SOJ aircraft can also employ a deception technique to generate false targets to confuse the radar operator and mask the presence of real targets (Figure 9-4). In this specialized technique, the deception jammer must tune to the frequency, PRF, and scan rate of the victim radar. The jammer then transmits multiple jamming pulses that the victim radar receiver processes like real target returns. With enough power, the deception jammer can generate multiple false azimuth targets by injecting jamming pulses into the sidelobes of the victim radar. False moving targets and false range targets are generated by varying the time delay of the jamming pulses based on the PRF and scan rate of the victim radar.

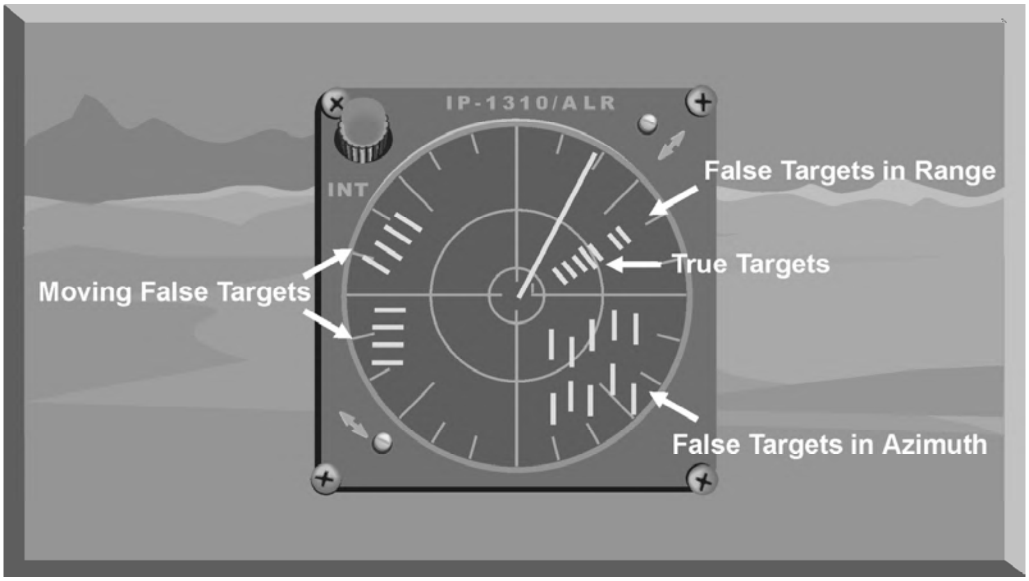

Figure 9-4. False Target Jamming

(3) Escort jamming is a specific tactic used by the EA-6B Prowler. The EA-6B is employed as an integral part of the attack package and is normally positioned behind and above the attack package (Figure 9-5).

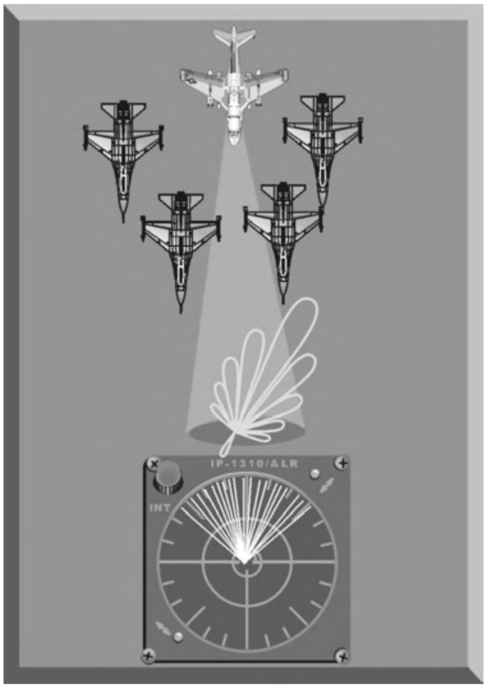

Figure 9-5. Escort Jamming

Using noise jamming, the EA-6B attempts to deny range and azimuth information to the victim radar by injecting high power signals into the main radar beam and sidelobes. To be effective, the EA-6B must be properly positioned in relation to the ingressing or egressing attack package (Figure 9-6).

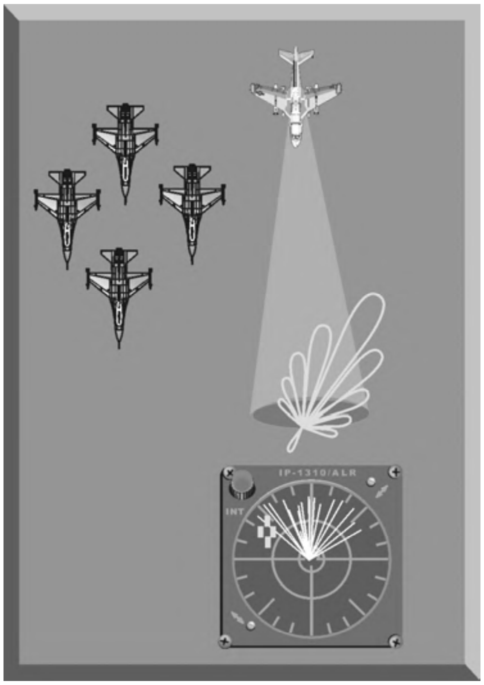

Figure 9-6. Escort Jamming Alignment

b. Self-protection radar jamming targets the radar systems that support jamming cannot negate. Self-protection jamming systems are part of a self-protection suite that includes a self-protection jamming pod, a chaff/flare dispenser, and on some aircraft, a towed decoy system. The overall purpose of these systems is individual aircraft survivability. These systems are designed to counter the individual SAM, AAA, and AI assets associated with the enemy IADS. They employ deception jamming techniques against the target tracking radars (TTRs) associated with these threats. They are designed to break the radar track or generate sufficient tracking errors to cause the missile or bullet to miss the aircraft.

(1) Self-protection radar jamming systems usually employ deception jamming techniques based on several factors. First, effective deception jamming techniques generally require less power than noise jamming techniques. Second, less power means less weight and space, which are very important considerations for modern tactical aircraft. Finally, deception jammers can be designed to jam multiple threats, which is a critical requirement for operations in a dense threat environment (Figure 9-7).

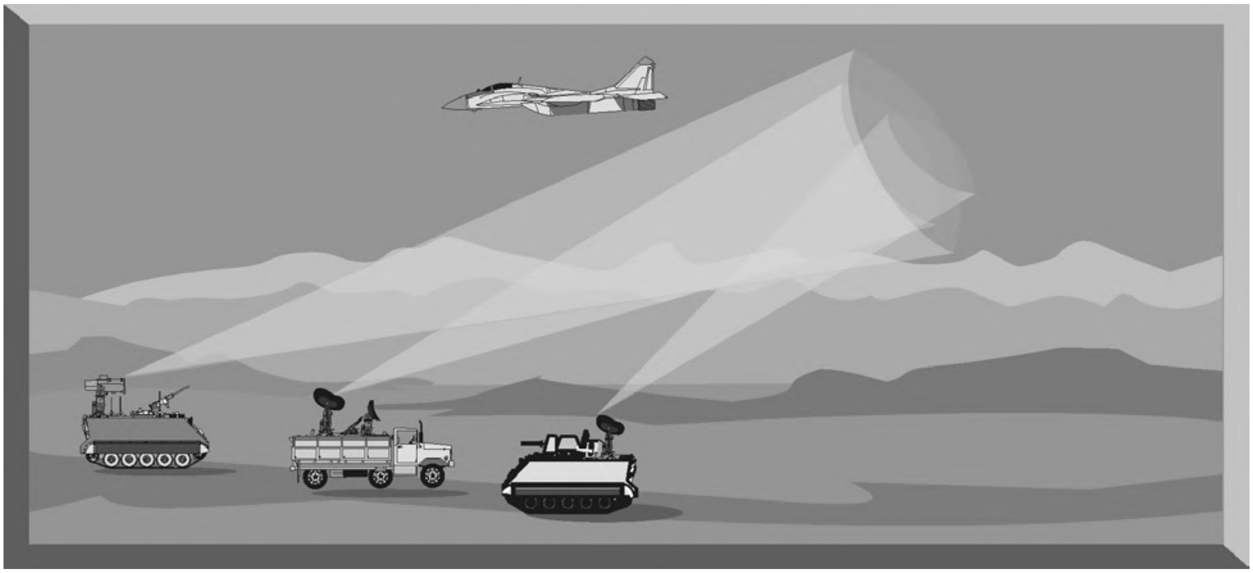

Figure 9-7. Self-Protection Jamming

(2) Despite the advantages of deception jamming techniques for self-protection jamming, there are some limitations that must be considered. First, deception jammers are complex electronic systems that must receive a victim radar's signal, memorize all its characteristics, modify the signal, and retransmit this modified signal at a high power level. Second, to be effective, deception jammers must be programmed with all the signal parameters (frequency, PRF, PRI, pulse width, scan rate, etc.) of the victim radar. Finally, because many deception techniques can be effective against specific threats, selecting optimum techniques to employ against these threats must be based on identified threat

system limitations. Identifying these specific threat systems limitations may be difficult.

4. FUNDAMENTALS OF RADAR JAMMING

There are some fundamental principles that apply to all types of jamming and to all jamming employment options. These principles are based on the characteristics of the jamming system and the characteristics of the victim radar. They include frequency matching, continuous interference, signal-to-noise ratio, jamming-to-signal ratio, and burnthrough range.

a. Based on the data provided by ES systems and intelligence evaluations, radar jamming systems must transmit signals at the frequency of the victim radar This applies to both noise and deception jamming. If a jamming signal does not match the transmitter frequency, the jamming signal is not received and displayed on the scope (Figure 9-8). When a jamming signal matches the transmitter frequency, the jamming signal is received and masks the target display (Figure 9-9).

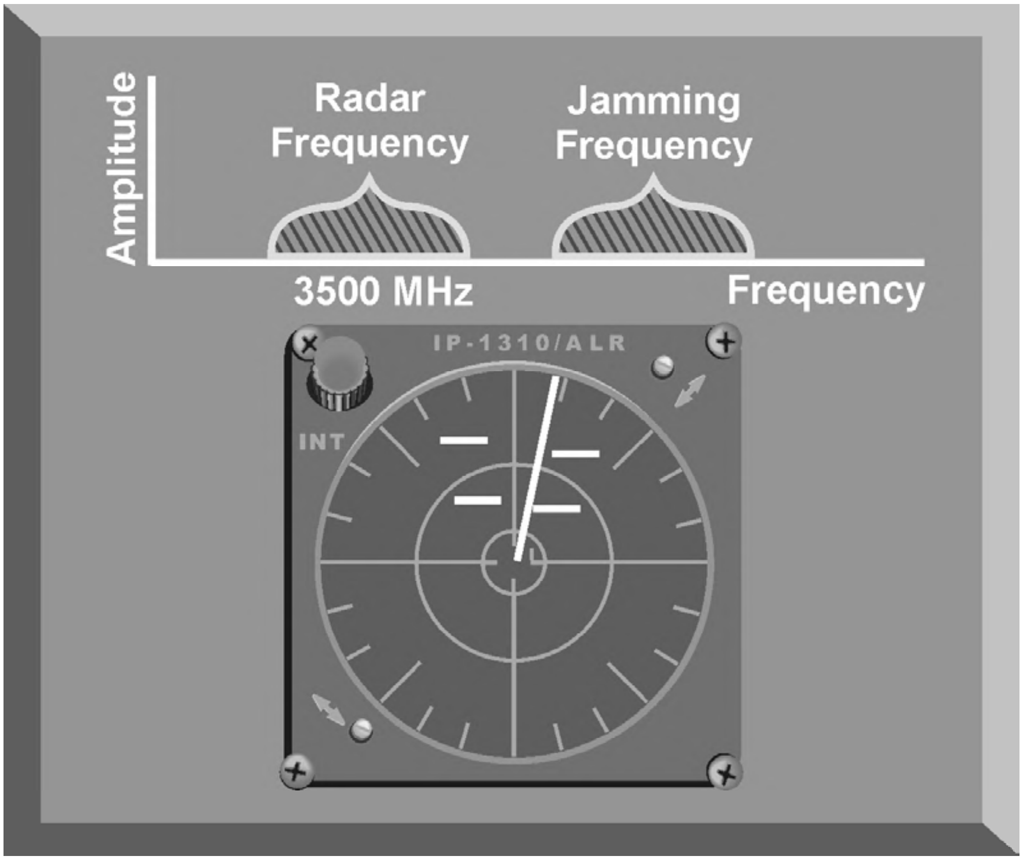

Figure 9-8. Jamming Frequency Error

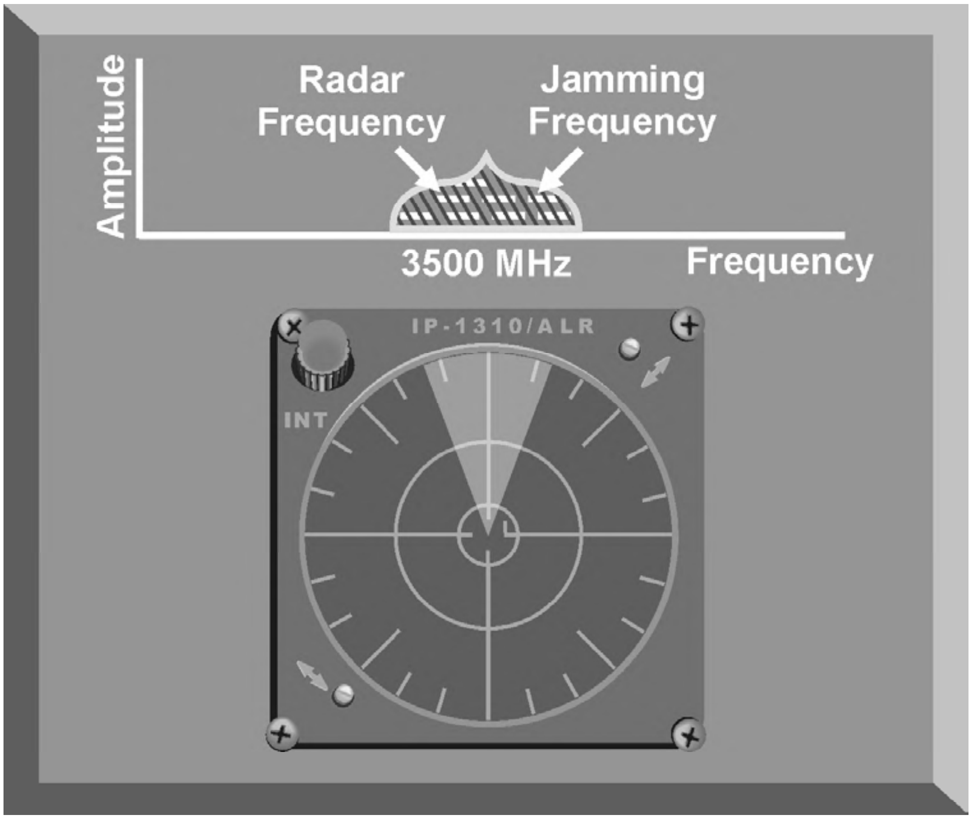

Figure 9-9. Correct Jamming Frequency Matching

b. For maximum effectiveness, a jamming transmitter should produce continuous interference. In much the same way intermittent static on a radio receiver does not completely block out a signal, intermittent jamming on a radar scope may not completely mask the target. An experienced radar operator or advanced automatic tracker can "read through" intermittent jamming and derive sufficient target information to negate jamming effectiveness. While true for noise jamming techniques, continuous interference also applies to deception techniques, especially when target reacquisition is considered.

c. The signal-to-noise (S/N) ratio is a measure of the ability of the victim radar to detect targets. It is also an indication of the vulnerability of the radar to certain jamming techniques, especially noise jamming.

(1) From the discussion of the basic radar equation in Chapter 5 (Equation 5-10), Equation 9-1 is the signal power density of a target return at the radar receiver. The signal power density of the target return is so weak that it requires very strong amplification before processing and display. Besides the signal power from the target, some level of thermal noise is also generated and amplified along with the target signal. For an "ideal" (no noise) amplifier, Equation 9-2 is used to compute the level of thermal noise generated by the amplifier.

$$\text{Signal Power Density} = \frac{P_T G \sigma A_e}{(4\pi)^2 R^4}$$

P_T = transmitted power
G = antenna gain
σ = target radar cross section (RCS)
A_e = antenna aperture area
R = range to the target

Equation 9-1. Signal Power Density

$$\text{Thermal Noise } (N) = KTBF$$

K = Boltzman's Constant
(1.38 x 10^{-23} watts/Hz degrees K)
T = standard temperature (290° K)
B = radar receiver equivalent bandwidth
F = radar receiver noise figure
(one for "ideal" receiver)

Equation 9-2. Thermal Noise

Note: The instantaneous bandwidth of a receiver is the frequency range over which the receiver can simultaneously amplify two or more signals to within a specified gain.

(2) The radar receiver amplifies both target signal and thermal noise. The output of the radar receiver will contain the target signal and the noise amplified across the bandwidth of the receiver. Separating the desired target signal from the undesired noise signal is one of the major problems confronting radar designers.

(3) Equation 9-3 is derived by dividing Equation 9-1 by Equation 9-2. Many factors in this equation fluctuate and must be estimated using statistical calculations. For example, target RCS fluctuates based on the changing angle of the antenna beam and corresponding changes in the reflected signal. Effective antenna aperture is also a statistical phenomenon based on the fluctuations in target RCS. The thermal noise generated by a receiver is also a fluctuating factor and must be treated statistically. This means that the S/N ratio is a statistical factor associated with a probability of target detection and a probability of a false alarm. A false alarm occurs when the radar operator or automatic tracking circuit designates a fluctuation in noise level as a target. The higher the S/N ratio, the

higher the probability of target detection with a corresponding reduction in the probability of a false alarm.

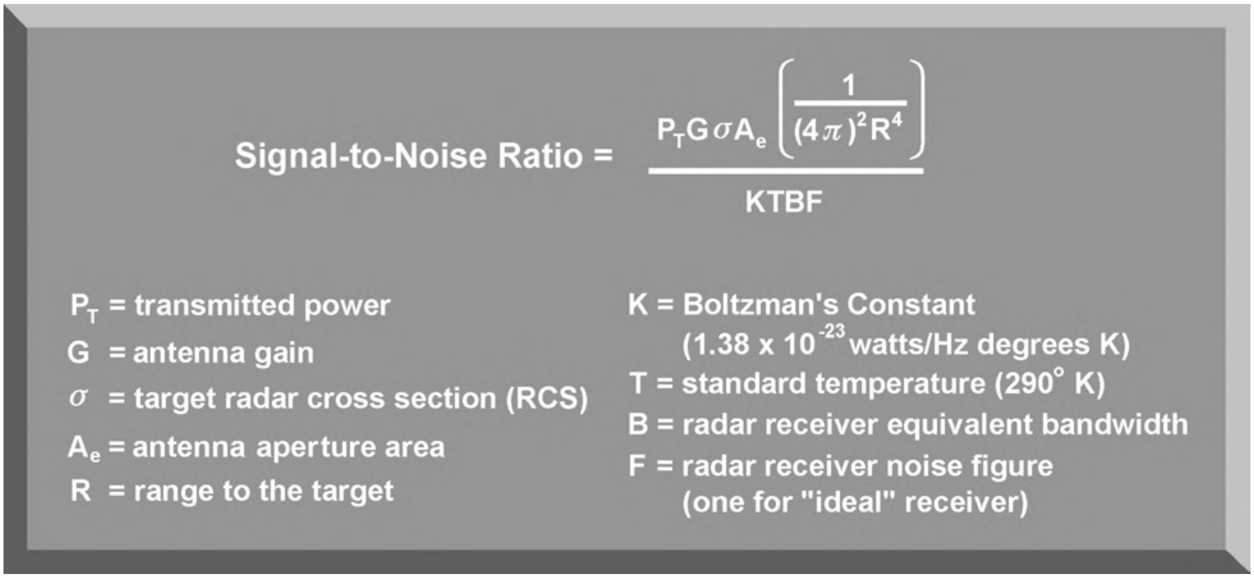

$$\text{Signal-to-Noise Ratio} = \frac{P_T G \sigma A_e \left(\frac{1}{(4\pi)^2 R^4}\right)}{KTBF}$$

P_T = transmitted power
G = antenna gain
σ = target radar cross section (RCS)
A_e = antenna aperture area
R = range to the target

K = Boltzman's Constant
(1.38 x 10^{-23} watts/Hz degrees K)
T = standard temperature (290° K)
B = radar receiver equivalent bandwidth
F = radar receiver noise figure
(one for "ideal" receiver)

Equation 9-3. Signal-to-Noise Ratio

(4) An analysis of Equation 9-3 suggests that any action that increases the power in the target signal (for example, increasing transmitted power, increasing antenna gain/aperture area, or decreasing target range) will improve the S/N ratio and improve the probability of target detection. It would also appear that decreasing the bandwidth of the radar receiver will increase the S/N ratio and enhance the probability of target detection. However, if the effective bandwidth of the receiver is reduced, this may eliminate a significant portion of the radar signal spectrum and decrease the probability of target detection.

(5) The S/N ratio is also an indication of the range at which a target will be detected. A plot of the receiver output of a typical radar is shown in Figure 9-10. The weak target signal at an extended range is just above the receiver noise level. The target at closer range is easily detected above the noise level. A radar operator or automatic target detector could mistake the very weak target return as a fluctuation in the receiver noise level. This could result in a missed detection. The lack of discrimination between noise and target returns because of a poor S/N ratio can also result in designating fluctuations in the noise level as actual target signals, known as false alarms.

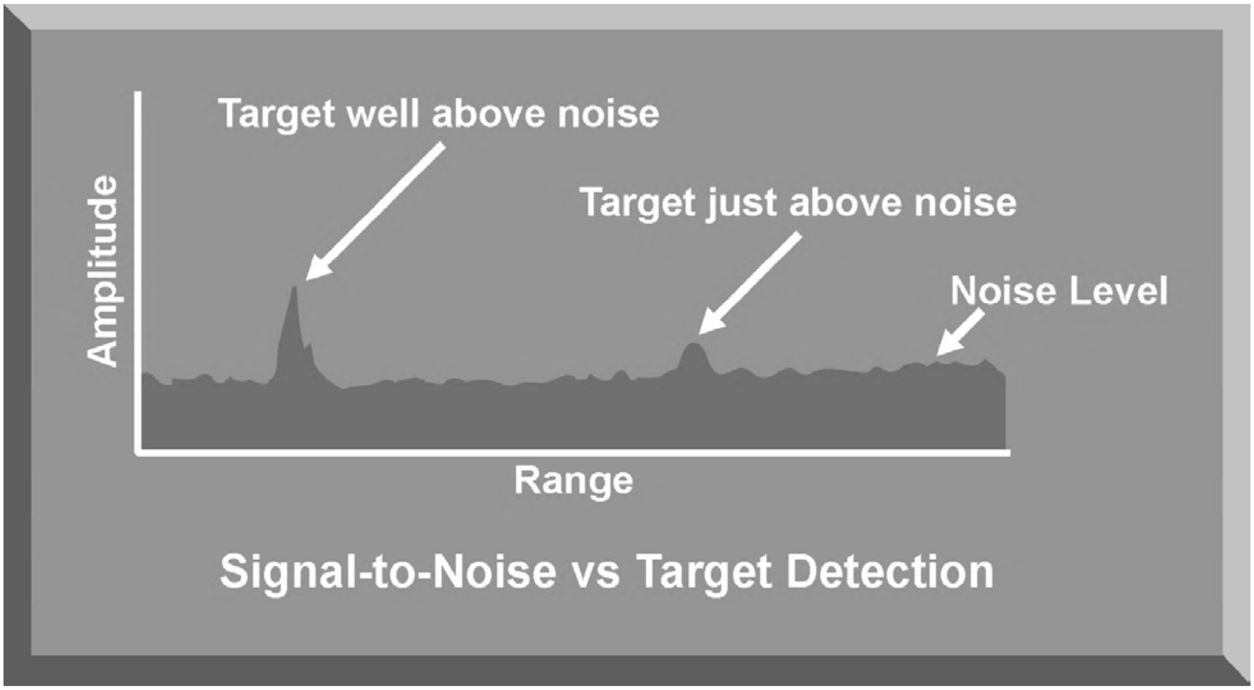

Figure 9-10. S/N Ratio and Target Detection

(6) To preclude, or minimize false alarms, the radar receiver may be equipped with electronic circuits to establish a false alarm threshold. If the signal strength of a radar return is below this threshold level, it will not be detected or displayed (Figure 9-11). This false alarm threshold also influences the probability of target detection. With the threshold set too high, many detected targets will not be displayed. Additionally, if the false alarm threshold is raised automatically in relation to the amplitude of the receiver noise, the radar receiver is more vulnerable to noise jamming.

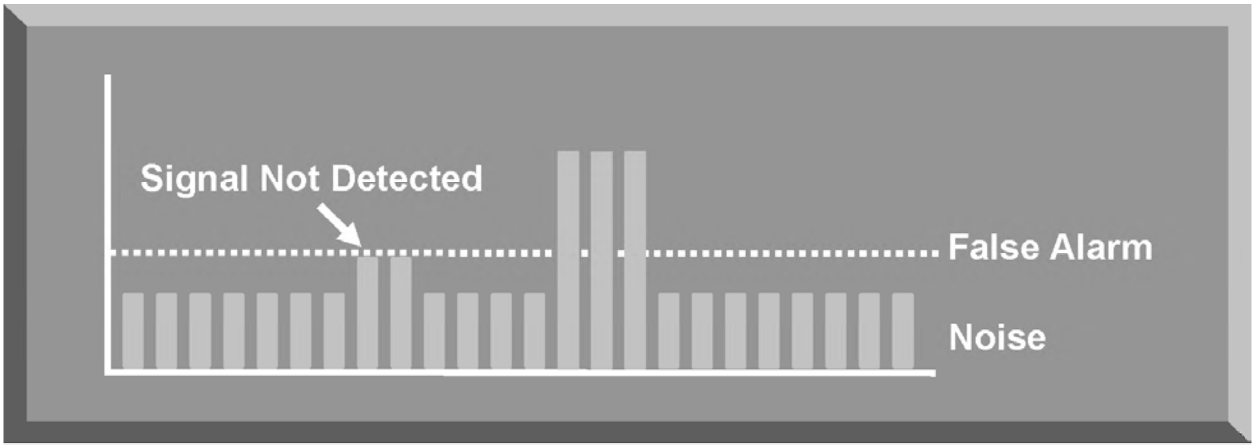

Figure 9-11. Receiver False Alarm Threshold

(7) For any target return to be detected by the radar, the S/N ratio must be greater than one. If the S/N ratio is less than one, the target will not be detected above the receiver noise level. The purpose of noise jamming is to raise the level of noise in the radar receiver to reduce the S/N ratio to less than one. This masks the presence of the true target return. If a false alarm threshold is used, noise jamming raises this threshold to further complicate target detection. Figure 9-12 depicts a S/N ratio greater than one. Figure 9-13 depicts a S/N ratio of less than one due to noise jamming.

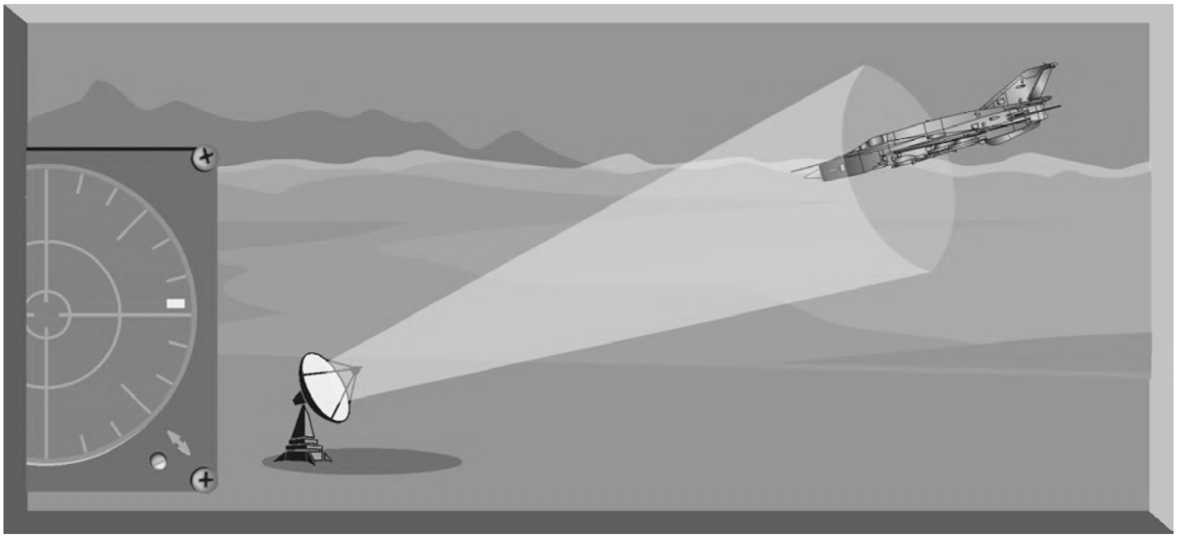

Figure 9-12. S/N Ratio Greater Than One

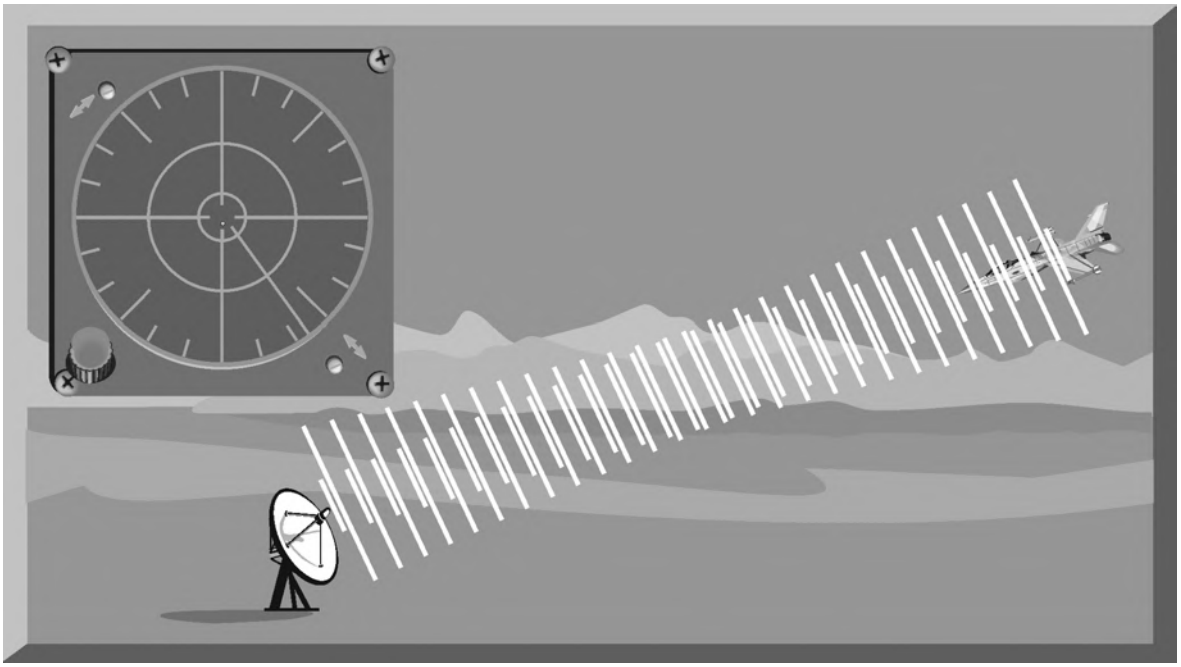

Figure 9-13. S/N Ratio Less Than One

d. The jamming-to-signal (J/S) ratio is a fundamental measure of jamming effectiveness. The J/S ratio compares the power in the jamming signal with the power in the radar return. Equation 9-4 is an expression of the J/S ratio. It is important to note that the J/S ratio should be measured at the output of the radar receiver. This will allow consideration of the receiver signal processing gain applied to the jamming signal.

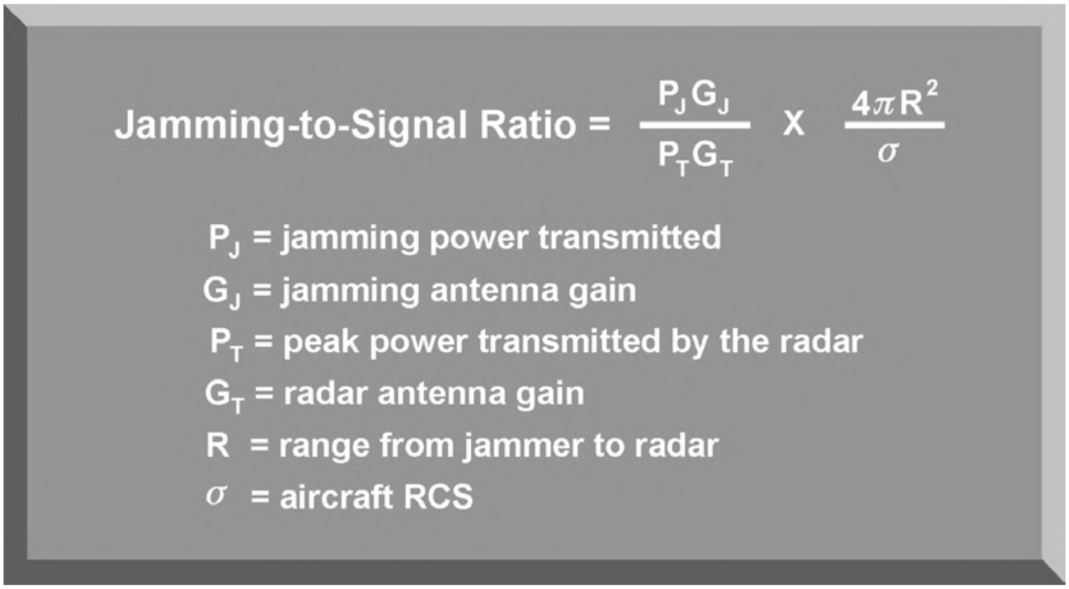

$$\text{Jamming-to-Signal Ratio} = \frac{P_J G_J}{P_T G_T} \times \frac{4\pi R^2}{\sigma}$$

P_J = jamming power transmitted
G_J = jamming antenna gain
P_T = peak power transmitted by the radar
G_T = radar antenna gain
R = range from jammer to radar
σ = aircraft RCS

Equation 9-4. Jamming-to-Signal Ratio

(1) The most critical factor in both the S/N and the J/S ratios is range. The S/N ratio is calculated based on R to the fourth power. This equates to a signal traveling from the radar to the target, and back to the radar receiver. The J/S ratio is calculated using R to the second power. This factor reflects the “one way” transmission of the jamming pulse from the jammer to the victim radar's receiver.

(2) For a jamming signal to be effective, the J/S ratio must be greater than one. In general, threat radars, especially ground-based radars, transmit much more power than does an airborne jamming system. However, this power must travel twice as far as the airborne jamming signal. At long ranges, a low power jamming system can generate a J/S ratio much greater than one. In Figure 9-14, the jamming pulse completely masks the target return. As the jamming system approaches the target, the distance the radar pulse travels decreases with a corresponding increase of power in the radar return. This reduces the J/S ratio to a value less than one and the radar “sees” the target. This is called the burnthrough range (Figure 9-15).

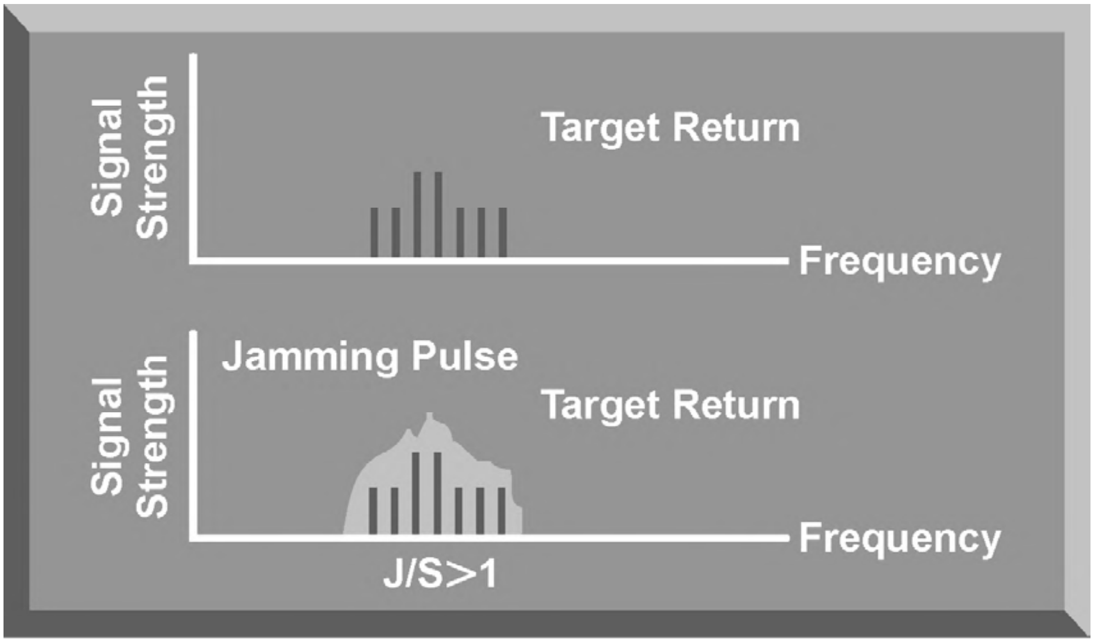
Signal Strength
Target Return
Frequency
Signal Strength
Jamming Pulse
Target Return
Frequency
J/S>1

Figure 9-14. J/S Ratio Greater Than One

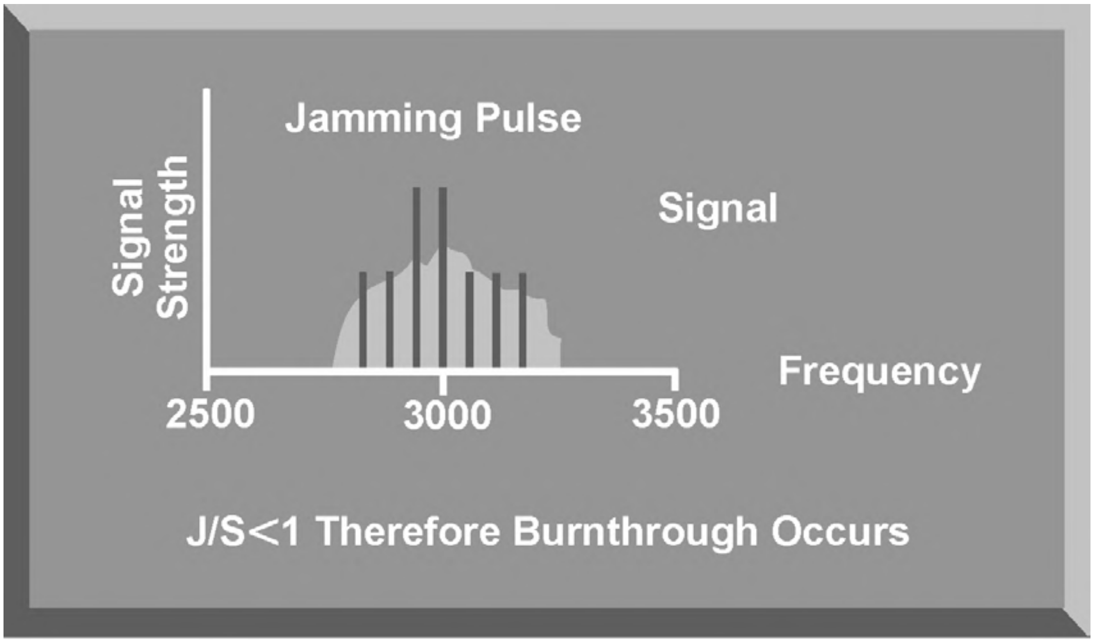
Jamming Pulse
Signal Strength
Signal
Frequency
2500
3000
3500
J/S<1 Therefore Burnthrough Occurs

Figure 9-15. J/S Ratio Less Than One

e. Burnthrough occurs when the power in the reflected target signal exceeds the power in the jamming signal. Even when an optimum and continuous jamming technique is transmitting on the exact frequency of the victim radar, the jamming starts to lose effectiveness as it nears the radar. For a particular radar

jamming technique, burnthrough range depends on the detection capability of the victim radar, expressed as the S/N ratio, and the capability of the aircraft's jamming system, expressed as the J/S ratio. The idea of burnthrough range explains why a jamming technique, especially noise jamming, loses its effectiveness as the aircraft approaches the radar. When plotting the jamming and signal power versus range (Figure 9-16), these two values intersect at the point where the J/S ratio is one. At closer ranges, the jamming pulse is no longer masking the aircraft, and the aircraft can be detected. Burnthrough range is the point where the radar can see through the jamming.

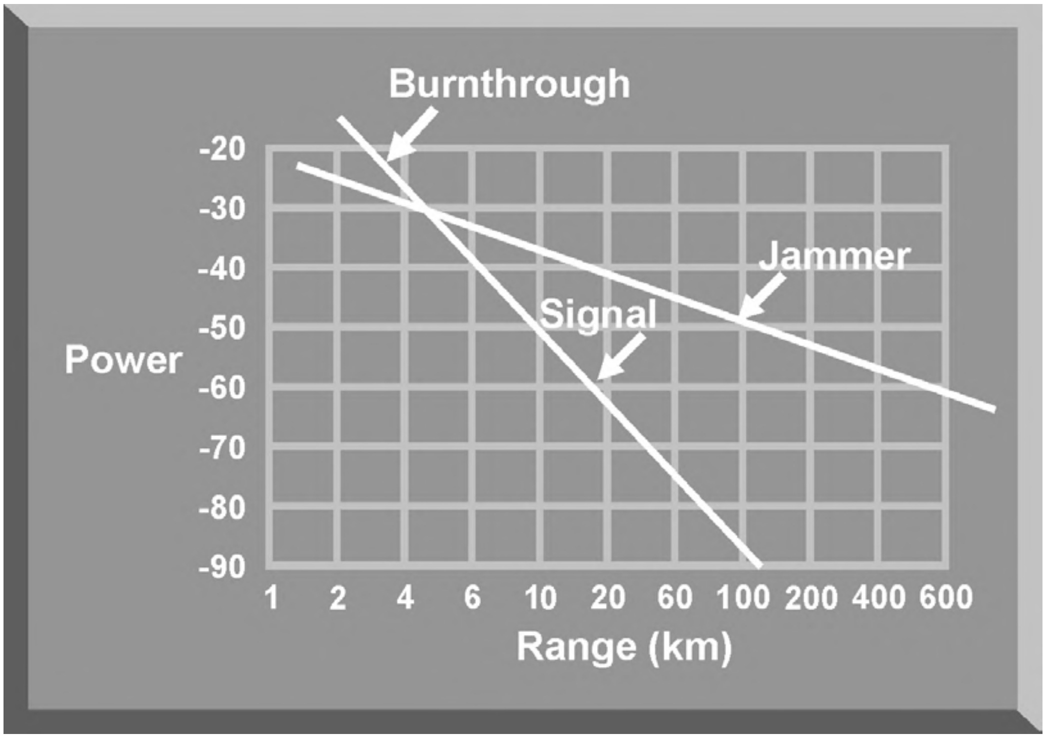

Figure 9-16. Burnthrough Range

5. SUMMARY

The purpose of radar jamming is to confuse or deny critical data to the radar systems that play a vital role in supporting the mission of an integrated air defense system. Two types of radar jamming, noise and deception, can be employed in a support-jamming role, or in a self-protection role for individual aircraft. The effectiveness of a jamming technique depends on the ability of the jamming system to generate a jamming signal that replicates the parameters of the victim radar, especially its frequency. The signal-to-noise ratio of the victim radar determines the vulnerability of the radar receiver to jamming while the jamming-to-signal ratio is an indication of the ability of the jamming system to effectively jam the victim radar. These basic radar jamming concepts are fundamental to understanding the impact of specific jamming techniques on radar systems.

CHAPTER 10. RADAR NOISE JAMMING

1. INTRODUCTION

A radar noise jamming system is designed to generate a disturbance in a radar receiver to delay or deny target detection. Since thermal noise is always present in the radar receiver, noise jamming attempts to mask the presence of targets by substantially adding to this noise level. Radar noise jamming can be employed by support jamming assets or as a self-protection jamming technique. Radar noise jamming usually employs high-power jamming signals tuned to the frequency of the victim radar. This chapter will discuss the factors that determine the effectiveness of radar noise jamming, radar noise jamming generation, and the most common noise jamming techniques. These noise jamming techniques include barrage, spot, swept spot, cover pulse, and modulated noise jamming.

2. RADAR NOISE JAMMING EFFECTIVENESS

The effectiveness of radar noise jamming depends on numerous factors. These factors include the jamming-to-signal (J/S) ratio, power density, the quality of the noise signal, and the polarization of the transmitted jamming signal (Figure 10-1).

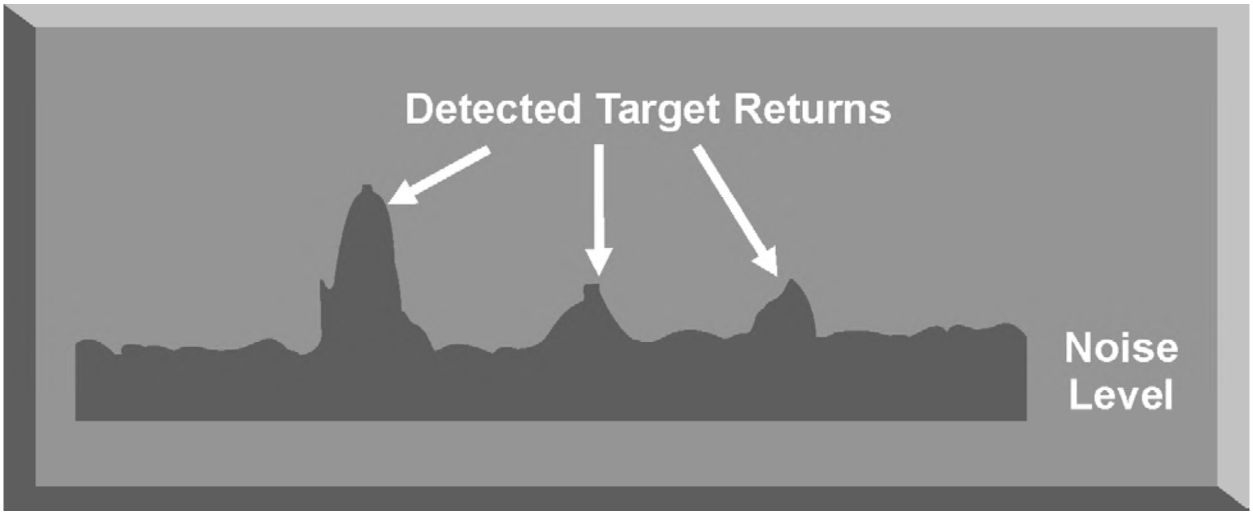

Figure 10-1. Noise Level in a Typical Radar Receiver Output

a. One of the most important factors that impacts the effectiveness of radar noise jamming is the J/S ratio (Figure 10-2). As discussed in Chapter 9, the power output of the noise jammer must be greater than the power in the target return, as measured at the output of the radar receiver. To achieve this level of jamming power, radar noise jammers usually generate high-power jamming signals. These high-power jamming signals can be introduced into the victim radar's main beam

to deny range information and into the victim radar's sidelobes to deny azimuth information.

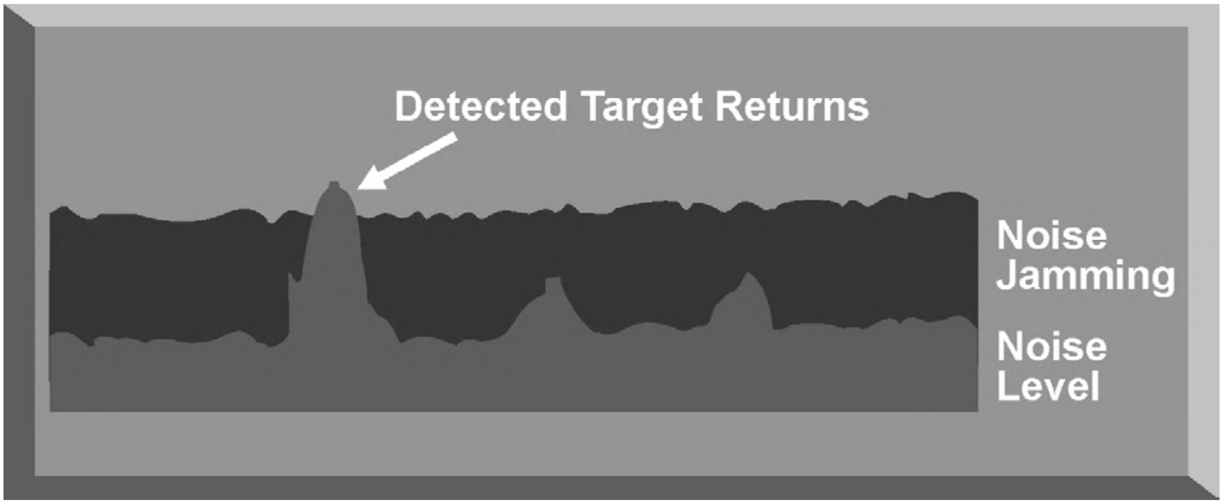
Detected Target Returns
Noise Jamming
Noise Level

Figure 10-2. Impact of Noise Jamming

b. Another factor which impacts the effectiveness of radar noise jamming is the power density. The power density of the noise jamming signal has a direct relation to the J/S ratio.

(1) If the noise jamming signal is centered on the frequency and bandwidth of the victim radar, the jamming signal has a high power density. The ability of a noise jammer to concentrate the jamming signal depends on the ability of the jammer to identify the exact frequency and bandwidth of the victim radar (Figure 10-3).

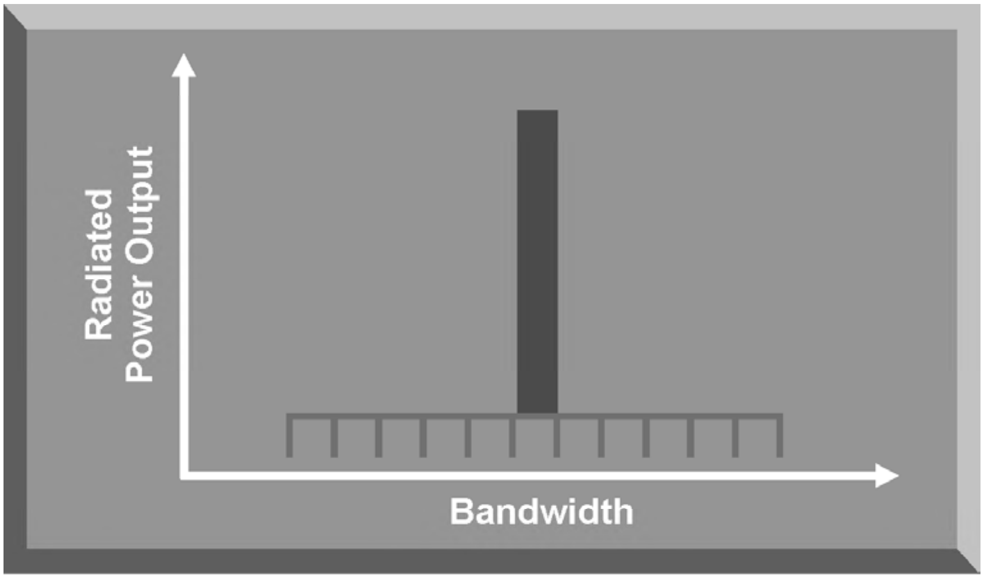
Radiated Power Output
Bandwidth

Figure 10-3. Power Density – Narrow Bandwidth

(2) If the generated noise jamming signal has to cover a wide bandwidth or frequency range, the power density at any one frequency is reduced (Figure 10-4). Radar systems that are frequency agile or that employ a wide bandwidth can reduce, or negate, the effectiveness of noise jamming by reducing the power density of the jamming signal.

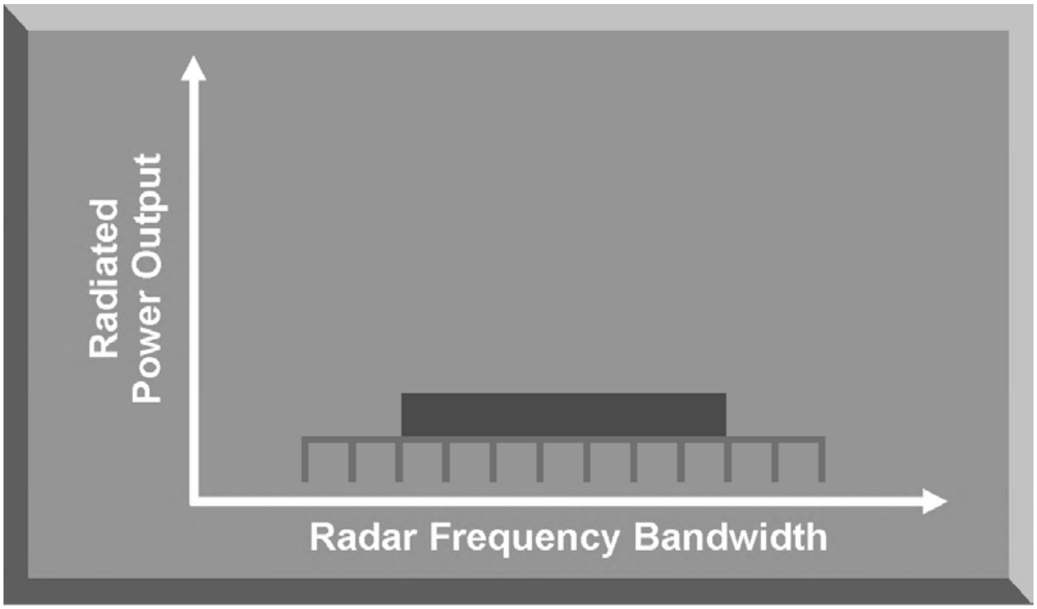

Figure 10-4. Power Density – Wide Bandwidth

c. The quality of the noise jamming also determines its effectiveness. To effectively jam a radar receiver with noise, the jamming signal must emulate the thermal noise generated by the receiver. This ensures that the radar operator or automatic detection circuit cannot distinguish between the noise jamming and normal thermal noise. Thermal noise is referred to as white noise and has a uniform spectrum. All of the frequencies in the bandwidth of the receiver have the same spectrum and an amplitude that varies based on Gaussian distribution. A Gaussian distribution is simply a bell-shaped distribution of amplitudes. In order to be effective, the jamming signal should exactly match the characteristics of the thermal noise signal of the victim radar receiver.

d. Polarization of the noise jamming signal is another significant factor that impacts its effectiveness. As discussed in Chapter 2, if the polarization of the jamming signal does not match the antenna polarization of the victim radar, there is a significant power loss in the jamming signal. Noise jamming systems designed to counter multiple threat radars, with various polarizations, generally use a transmitting antenna with a 45° slant or use circular polarization. Most threat systems are horizontally or vertically polarized. This results in a 50% reduction in effective radiated power (ERP) for most threat systems. A more serious power loss, nearly 100%, in ERP occurs when the jamming antenna is

orthogonally polarized with the victim antenna. The polarization of the noise jamming signal impacts the J/S ratio and the power density.

3. RADAR NOISE JAMMING GENERATION

Noise jamming is produced by modulating an RF carrier wave with random amplitude or frequency changes, called noise, and retransmitting that wave at the victim radar's frequency. Since noise from numerous sources is always present and displayed on a radar scope, noise jamming adds to the problem of target detection. Reflected radar pulses from target aircraft are extremely weak. To detect these pulses, a radar receiver must be very sensitive and be able to amplify the weak target returns. Noise jamming takes advantage of this radar characteristic to delay or deny target detection.

a. The simplest method of generating a high-power Gaussian noise jamming signal is to employ a highly amplified diode to generate a noise signal at the frequency of the victim radar. This signal is filtered and directly amplified to the maximum power that can be generated by the transmitter. This method is called direct noise amplification (DINA). The DINA method of noise generation has a serious limitation. The maximum power available from linear wideband power amplification is extremely limited. Employing any other form of power amplification would alter the Gaussian distribution of the jamming signal. This method of generating radar noise jamming was used extensively during WW II.

b. Modern noise jamming systems generate noise jamming signals by frequency modulating a carrier wave at the frequency of the victim radar. FM noise jammers employ a receiving antenna to intercept the victim's radar signal. The antenna passes the victim radar signal to the receiver for identification. The receiver also tunes the jamming signal generator to the correct frequency. The receiver uses an automatic frequency control (AFC) circuit to tune the voltage-controlled oscillator (VCO) to the frequency of the victim radar. A noise signal is generated by the jamming signal generator and added to the tuning voltage of the VCO to get an FM jamming signal. This signal is sent to a traveling wave tube (TWT) power transmitter. The TWT is normally operated in a saturated mode which produces a high-power jamming signal that covers a wider bandwidth than the victim radar. This reduces the power density of the signal, but the high power levels available from the TWT amplification of an FM signal compensate for this loss. The signal is sent to the transmitting antenna and directed toward the victim radar.

c. Figure 10-5 highlights an important feature of a modern radar noise jamming system: a look-through capability. A look-through mode allows the receiver to periodically sample the signal environment. The objective of the look-through mode is to allow the jammer to update victim radar parameters and change the jamming signal to respond to changes in the signal environment. This greatly enhances the effectiveness of noise jamming systems. One method used to provide a look-through capability is to isolate the transmit and receive

antennas to allow continuous operation of the receiver to update signal parameters. Another method is to switch off the jammer for a brief period to allow the receiver to sample the signal environment. Since this latter look-through method eliminates the jamming signal, the amount of time the jammer is switched off must be kept to a minimum.

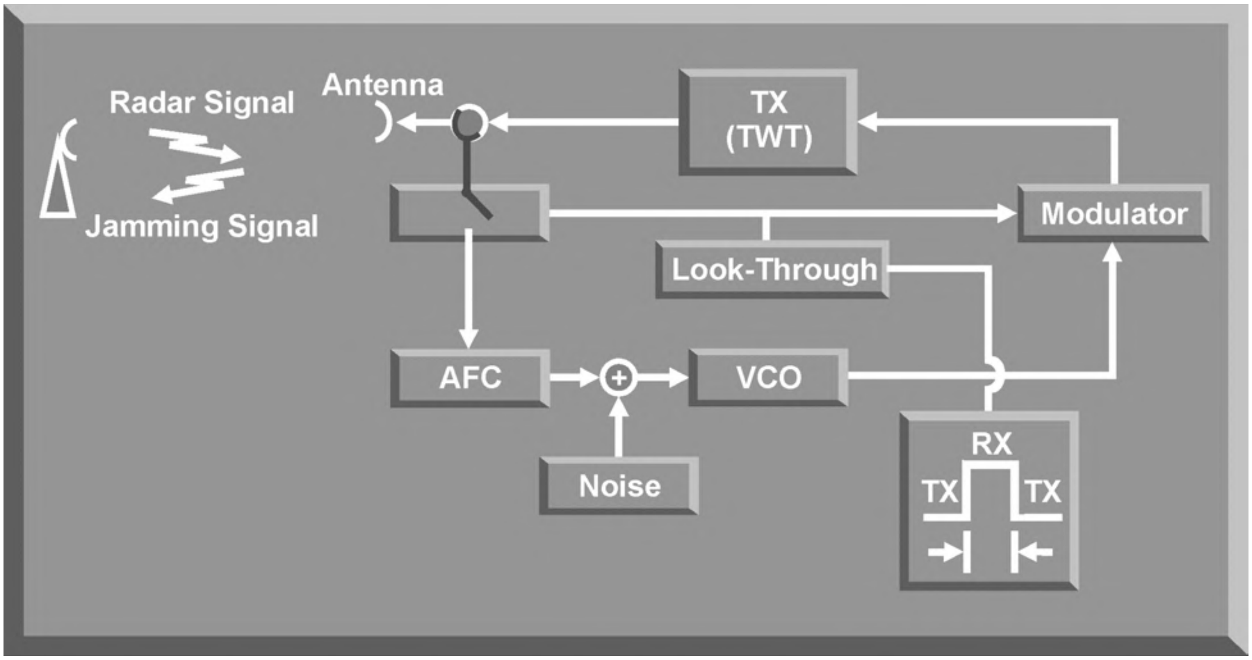

Figure 10-5. Frequency Modulated (FM) Noise Jamming System

4. BARRAGE JAMMING

An important aspect of jamming power is power density. Noise jamming depends on power density for its effectiveness. Power density is a function of the frequency range, or bandwidth, of the jamming signal. If a jammer covers a narrow frequency range, it can concentrate energy in a narrow band. If a jammer covers a wide frequency range, the energy is spread over that entire range. Since the jammer has fixed radiated power, this lowers the effective jamming power at a given frequency. Barrage jamming is a jamming technique where high power is sacrificed for the continuous coverage of several radar frequencies (Figure 10-6). The jamming signal is spread over a wide frequency range, which lowers the ERP at any one frequency. This type of jamming is useful against frequency-agile radars, against a radar system that uses multiple beams, or against multiple radar systems operating in a specific frequency range. By spreading the jamming over a wide frequency range, there is some level of jamming no matter what frequency the radar uses. Barrage jamming was used extensively during World War II. Advantages of barrage jamming are its simplicity and ability to cover a wide portion of the electromagnetic spectrum. The primary disadvantage is the low power density, especially when a high J/S ratio is needed against modern radars.

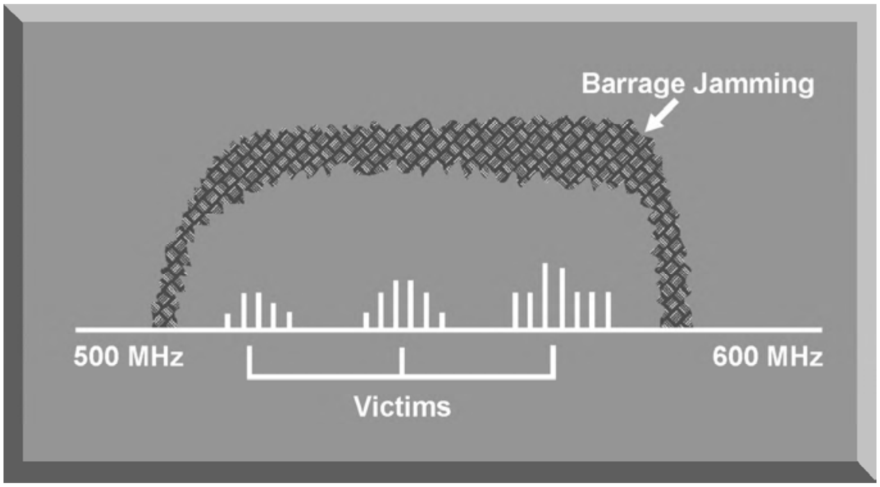

Figure 10-6. Barrage Jamming

5. SPOT JAMMING

One way to take advantage of the noise jammer's simplicity, but raise the jamming signal power, is to use a spot jammer. The earliest spot jammers were very narrow band jammers covering a bandwidth of 10 megahertz or less (Figure 10-7). This narrow band spot jammer was tuned to the anticipated frequency of the target radar. When it is necessary to jam a number of radars at different frequencies, more than one jammer is used. One problem that developed was of carrying the required number of spot jammers to counter a modern IADS. Also, radars that change their operating frequency, or are frequency-agile, defeat the spot jammer. Today, intercept panoramic receivers work with spot jammers to determine the frequency of the victim radar. A look-through capability is included in the system so that the target radar signal can be monitored to assess jamming effectiveness. The jamming signal can be adjusted for any changes in the operating frequency of the radar.

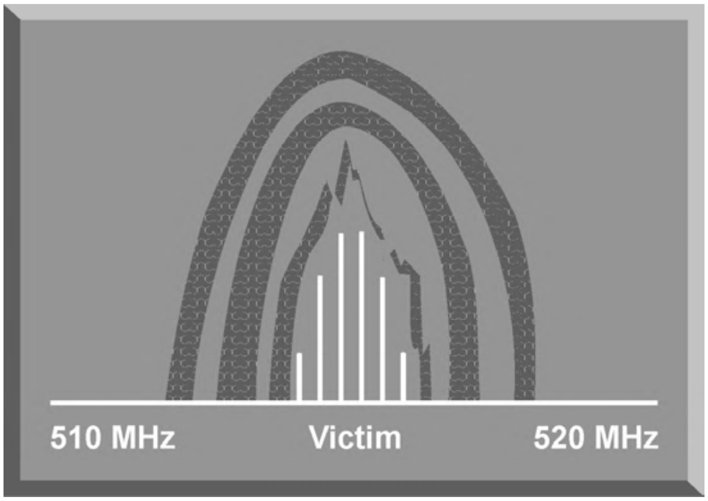

Figure 10-7. Spot Jamming

a. The primary advantage of spot jamming is its power density. Radar or communications receivers can be countered at longer ranges than when using a barrage jammer of equal output power (Figure 10-8).

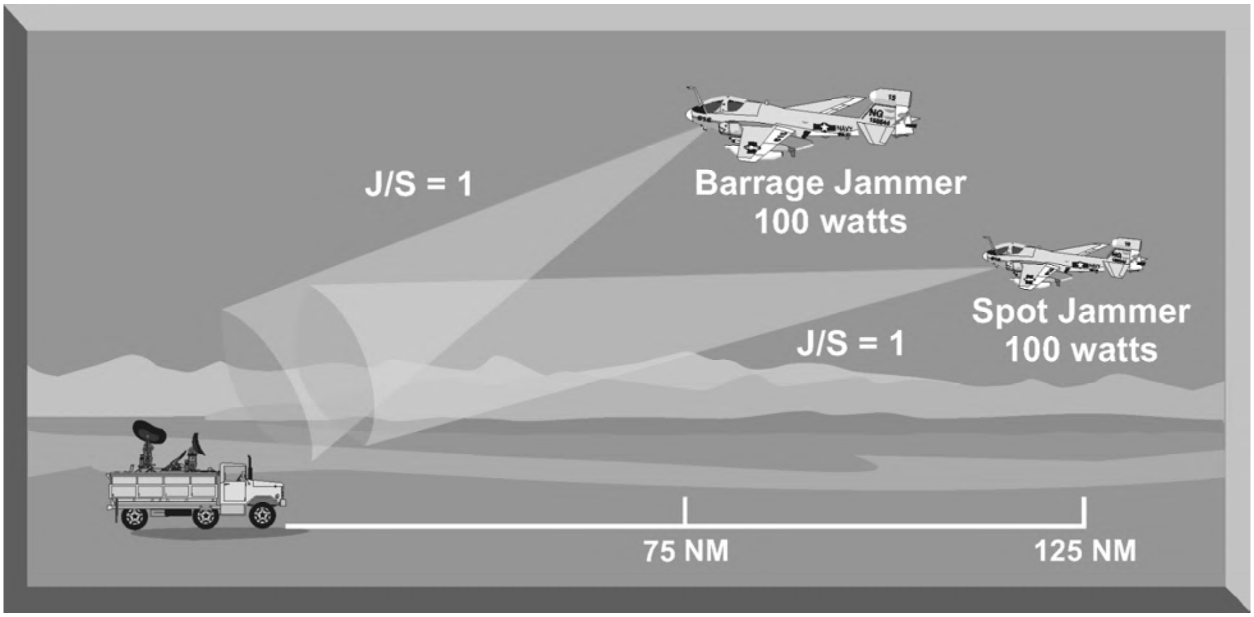

Figure 10-8. Spot Jamming Effectiveness

b. A disadvantage of the spot jammer is its coverage of a narrow band of the frequency spectrum. An operator or computer in the receiver must constantly monitor and tune the jamming signal to the target radar's frequency. The complexity of this process increases when jamming frequency-agile radars that can change frequencies with every pulse.

6. SWEPT-SPOT JAMMING

When high power density is required over a large bandwidth, one solution is to take spot jamming and sweep it across a wide frequency range (Figure 10-9). This preserves the high power density but allows the jamming to cover a large bandwidth. The jamming spot is swept across a broad frequency range at varying speeds. With this technique, a number of radar systems can be covered. Because of their high jamming power, swept-spot jammers are able to cover a number of radars operating in a broad frequency range. However, jamming is not continuous. Fast swept-spot jamming can approximate continuous jamming by causing a phenomenon known as "ringing." Fast sweeping spot noise is like a burst of energy which sets up vibrations within the receiver section. When these vibrations last until the next burst of energy is received, this is known as ringing. Three factors determine swept-spot jamming effectiveness. The first is the power in the spot. The next is the bandwidth, or frequency range, the spot covers. The last is the sweep rate.

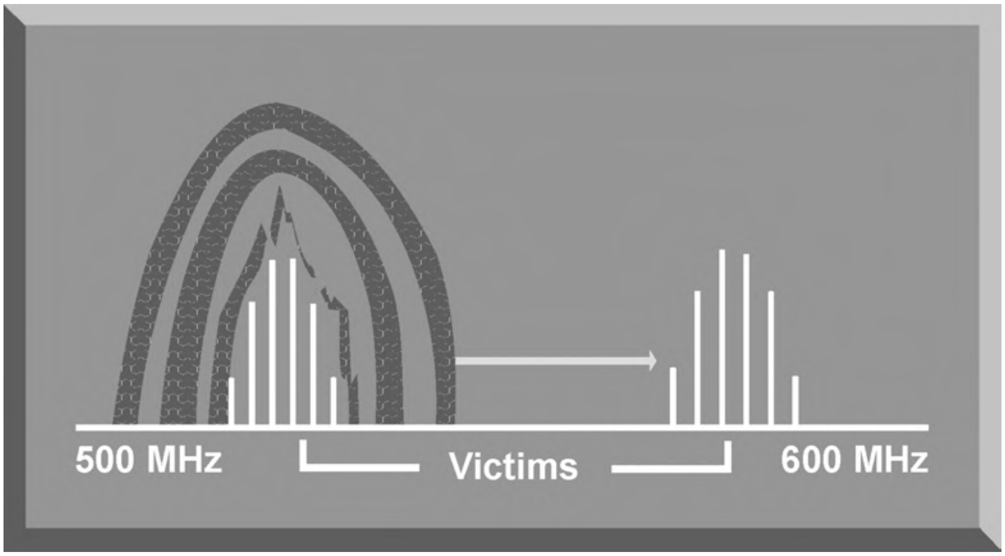

Figure 10-9. Swept-Spot Jamming

7. COVER PULSE JAMMING

Cover pulse jamming is a modification of swept-spot jamming. This is a "smart noise" technique that is responsive for a short period of time (Figure 10-10). A repeater jammer acts as a transponder. It receives several radar pulses and determines the PRF of the victim radar. It then uses this data to predict when the next radar pulse should arrive. Using an oscillator that is gated for a period of time based on predicted pulse arrival time, a noise-modulated signal is amplified and transmitted. This process works against a radar with a steady PRF, and allows a low-powered repeater to respond to a number of threats by time-sharing.

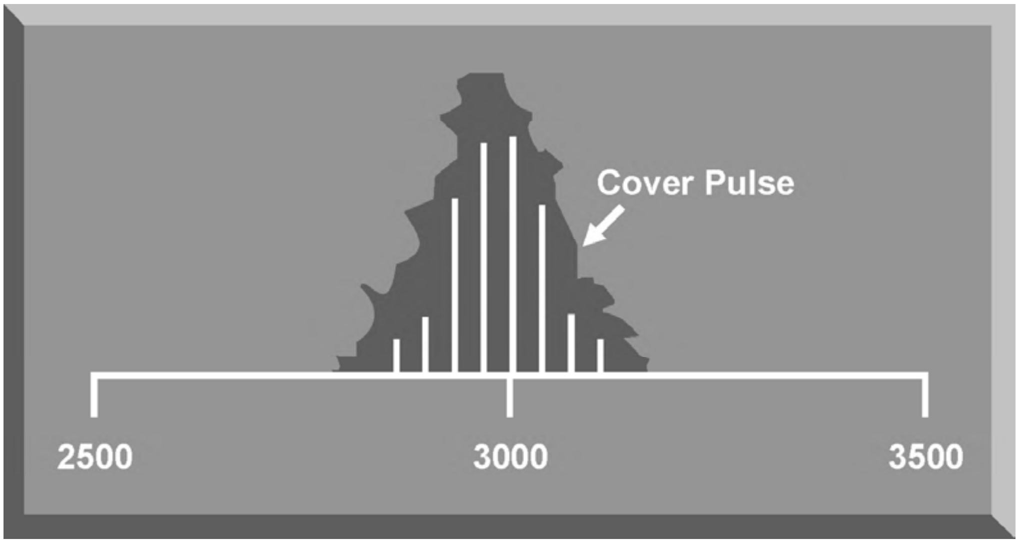

Figure 10-10. Cover Pulse Jamming

a. Cover pulse jamming is used to initiate a range gate pull-off (RGPO) deception jamming technique. The deception jammer transmits a noise jamming signal, or cover pulse, that is much stronger than the target return. The cover pulse raises the automatic gain inside the range gate, and the range tracking loop initiates tracking on the cover pulse. The deception jammer then increases the time delay in the jamming pulse and moves the range tracking gate away from the real target (Figure 10-11).

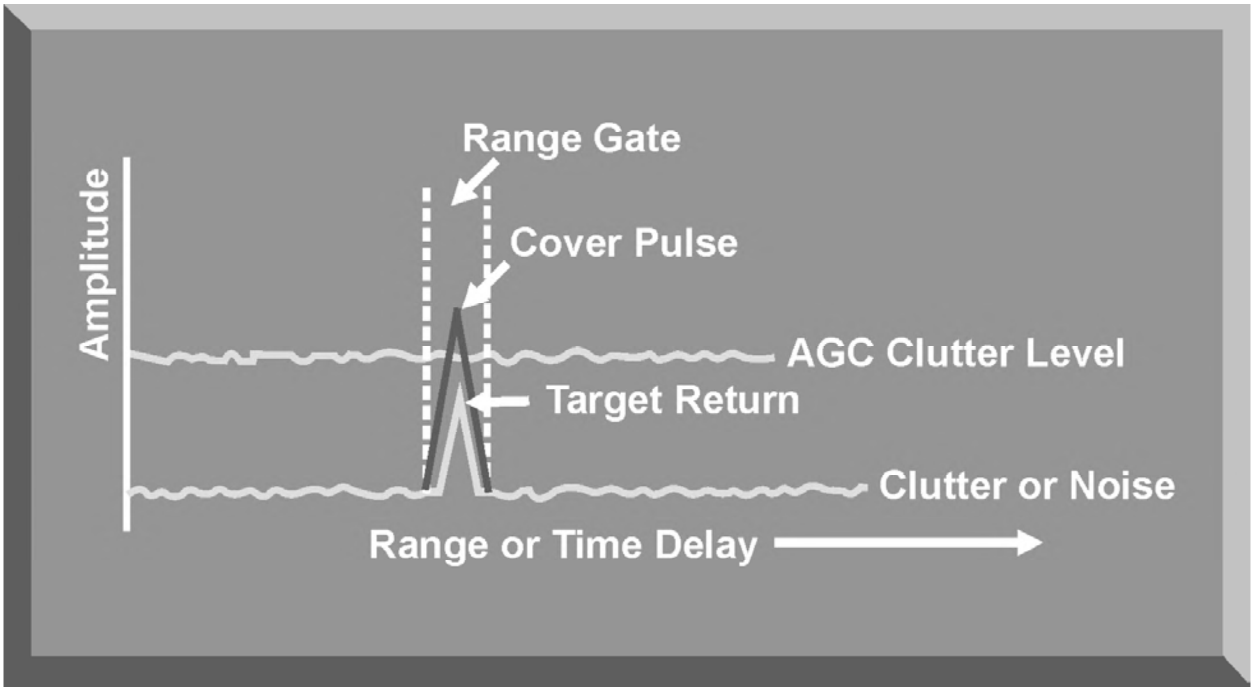

Figure 10-11. Range Gate Pull-Off Cover Pulse

b. A form of cover pulse jamming is also used to initiate a velocity gate pull-off (VGPO) technique against continuous wave and pulse Doppler radars. The cover pulse, in this case, is a strong jamming signal with the same frequency shift as the aircraft return. This cover pulse steals the velocity tracking gate and sets up the velocity tracking loop to steal the velocity tracking gate based on false target Doppler shifts (Figure 10-12).

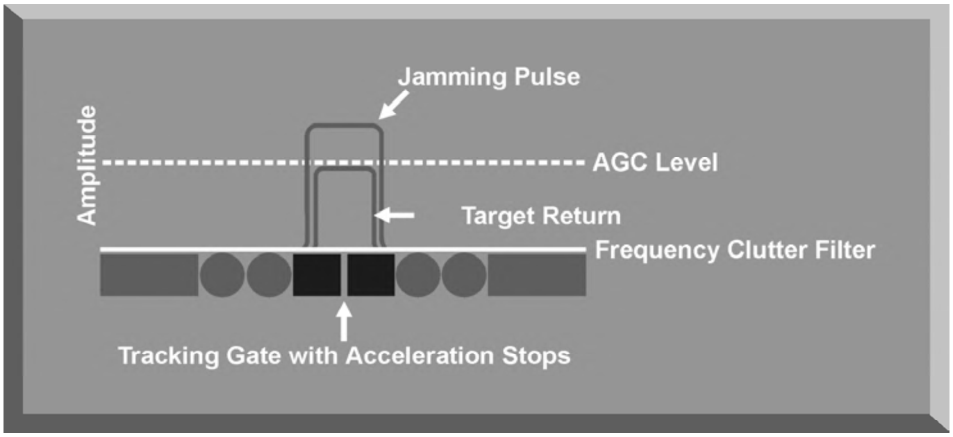
Jamming Pulse
Amplitude
AGC Level
Target Return
Frequency Clutter Filter
Tracking Gate with Acceleration Stops

Figure 10-12. Velocity Gate Pull-Off Cover Pulse

8. MODULATED NOISE JAMMING

Modulated jammers are special hybrid jammers which employ noise jamming that is either amplitude or frequency modulated. The purpose of this modulated noise is to defeat target tracking radars (TTRs) rather than deny range information. Modulated noise jamming has proven effective against conical scan and track-while-scan (TWS) TTRs.

a. Modulated jamming alters the noise jamming signal at a frequency that is related to the scan rate of the target radar. If modulated jamming is used against a conical scan radar, a sine wave signal is used (Figure 10-13). The frequency of the sine wave is slightly higher than the scan rate of the victim radar. The amplitude difference results in a constantly varying phase between the radar and the jamming signal. This phase differential produces false targets with a strong signal amplitude everywhere the signals reinforce each other. This causes the conical scan radar to track the false returns and lose the real target return. For this technique to work, the scan rate of the intended victim radar must be known.

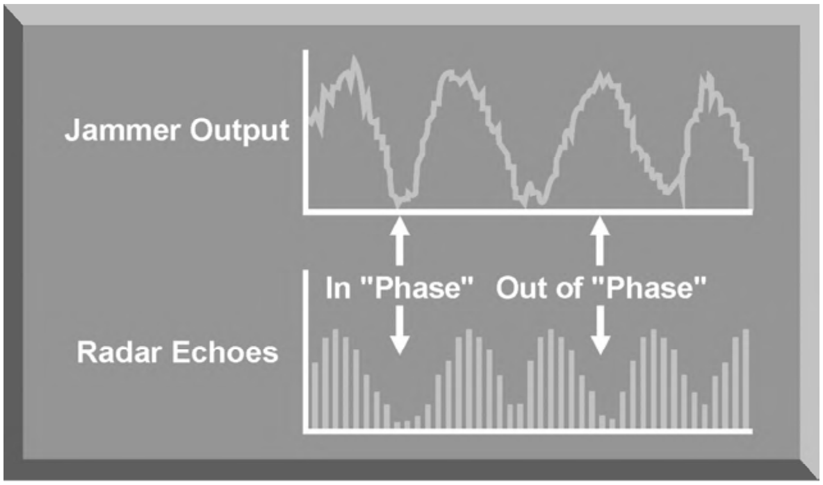
Jammer Output
In "Phase"
Out of "Phase"
Radar Echoes

Figure 10-13. Conical Scan Modulated Jamming

b. Against a TWS radar, a rectangular waveform is used to modulate the noise signal. The PRF of the modulation is set at some harmonic of the TWS rate. This synchronization results in a number of jamming strobes on the radar scope. Each jamming strobe is at a different azimuth or elevation depending on which radar beam is being jammed. The number of jamming strobes depends directly on the harmonic used to modulate the signal. In Figure 10-14, a modulating signal frequency that is four times the scan rate of the radar will produce four jamming strobes on the scope. If the jamming is slightly out of tune with the scan rate, the jamming strobes will appear to roll across the radar scope.

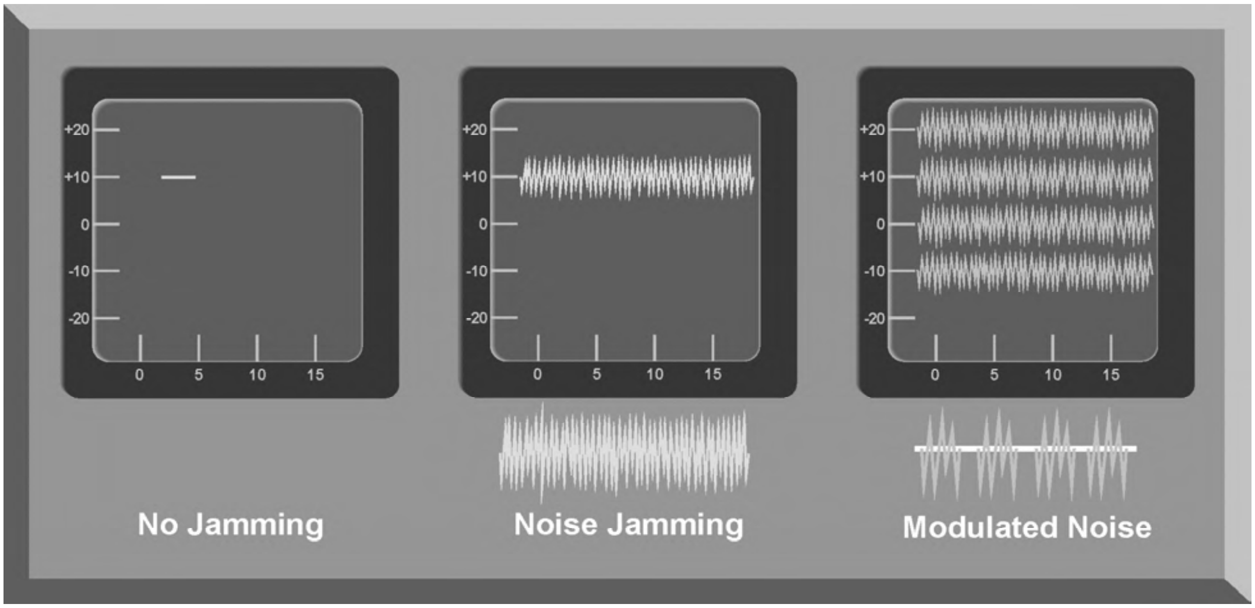

Figure 10-14. TWS Modulated Jamming

9. SUMMARY

Radar noise jamming is employed to deny target acquisition and target tracking data to a victim radar. This is accomplished by injecting amplitude or frequency modulated noise jamming signals into the victim radar's receiver. The radar noise jamming techniques discussed in this chapter included barrage, spot, swept-spot, cover pulse, and modulated jamming. The effectiveness of these noise jamming techniques depends on the power density of the jamming signal compared to the power in the radar return, or the J/S ratio. Radar noise jammers are generally simple, high-power systems which can be effectively employed in a support or self-protection role. Radar noise jamming can be employed in conjunction with deception jamming techniques to maximize the impact of jamming on victim radars.

CHAPTER 11. DECEPTION JAMMING

1. INTRODUCTION

Deception jamming systems are designed to inject false information into a victim radar to deny critical information on target azimuth, range, velocity, or a combination of these parameters. To be effective, a deception jammer receives the victim radar signal, modifies this signal, and retransmits this altered signal back to the victim radar. Because these systems retransmit, or repeat, a replica of the victim's radar signal, deception jammers are known as repeater jammers. The retransmitted signal must match all victim radar signal characteristics including frequency, pulse repetition frequency (PRF), pulse repetition interval (PRI), pulse width, and scan rate. However, the deception jammer does not have to replicate the power of the victim radar system.

a. A deception jammer requires significantly less power than a noise jamming system. The deception jammer gains this advantage by using a waveform that is identical to the waveform the radar's receiver is specifically designed to process. Therefore, the deception jammer can match its operating cycle to the operating cycle of the victim radar instead of using the 100% duty cycle required of a noise jammer. To be effective, a deception jammer's power requirements are dictated by the average power of a radar rather than the peak power required for a noise jammer. In addition, since the jammer waveform looks identical to the radar's waveform, it is processed like a real return. The jamming signal is amplified by the victim radar receiver, which increases its effectiveness. The reduced power required for effective deception jamming is particularly significant when designing and building self-protection jamming systems for tactical aircraft that penetrate a dense threat environment. Deception jamming systems can be smaller, lighter, and can jam more than one threat simultaneously. These characteristics give deception jammers a great advantage over noise jamming systems.

b. Although deception jammers require less power, they are much more complex than noise jammers (Figure 11-1). Memory is the most critical element of any deception jammer. The memory element must store the signal characteristics of the victim radar and pass these parameters to the control circuitry for processing. This must be done almost instantaneously for every signal that will be jammed. Any delay in the memory loop diminishes the effectiveness of the deception technique. Using digital RF memory (DRFM) reduces the time delay and enhances deception jammer effectiveness. Deception jamming employed in a self-protection role is designed to counter lethal radar systems. To be effective, deception jamming systems must be programmed with detailed and exact signal parameters for each lethal threat.

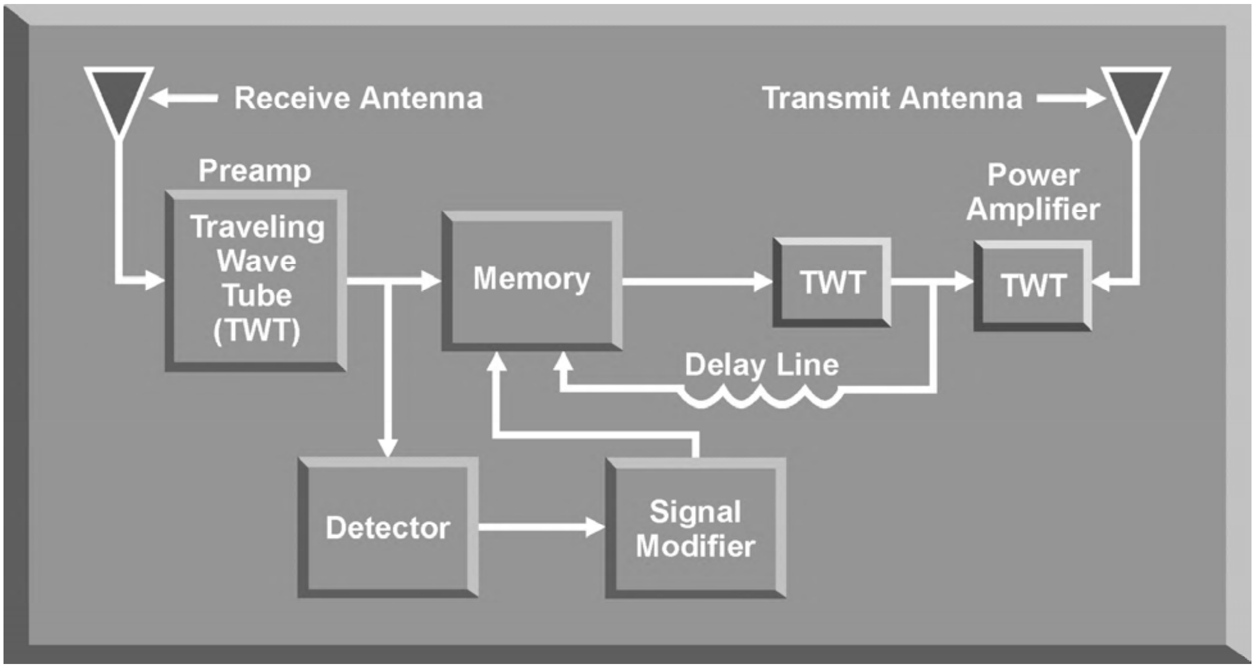

Figure 11-1. Deception Jamming System

c. The requirement for exact signal parameters increases the burden on electronic warfare support (ES) systems to provide and update threat information on operating frequency, PRF, PRI, power pulse width, scan rate, and other unique signal characteristics. An electronic intelligence (ELINT) architecture is required to collect, update, and provide changes to deception jamming systems. In addition, intelligence and engineering information on exactly how a specific threat system acquires, tracks and engages a target is essential in identifying system weaknesses. Once a weakness has been identified, an effective deception jamming technique can be developed and programmed into a deception jammer. For example, if a particular radar system relies primarily on Doppler tracking, a Doppler deception technique will greatly reduce its effectiveness. Threat system exploitation is the best source of detailed information on threat system capabilities and vulnerabilities. Effective deception jamming requires much more intelligence support than does noise jamming.

d. Most self-protection jamming techniques employ some form of deception against a target tracking radar (TTR). The purpose of a TTR is to continuously update target range, azimuth, and velocity. Target parameters are fed to a fire control computer that computes a future impact point for a weapon based on these parameters and the characteristics of the weapon being employed. The fire control computer is constantly updating this predicted impact point based on changes in target parameters. Deception jamming is designed to take advantage of any weaknesses in either target tracking or impact point calculation to maximize the miss distance of the weapon or to prevent automatic tracking. This chapter will discuss the most commonly employed deception jamming

techniques, including false target jamming, range deception jamming, angle deception jamming, velocity deception jamming, and monopulse jamming.

2. FALSE TARGET JAMMING

False target jamming is an effective jamming technique employed against acquisition, early warning, and ground control intercept (GCI) radars. The purpose of this type of jamming is to confuse the enemy radar operator by generating many false target returns on the victim radar scope. When false target deception jamming is successfully employed, the radar operator cannot distinguish between false targets and real targets (Figure 11-2).

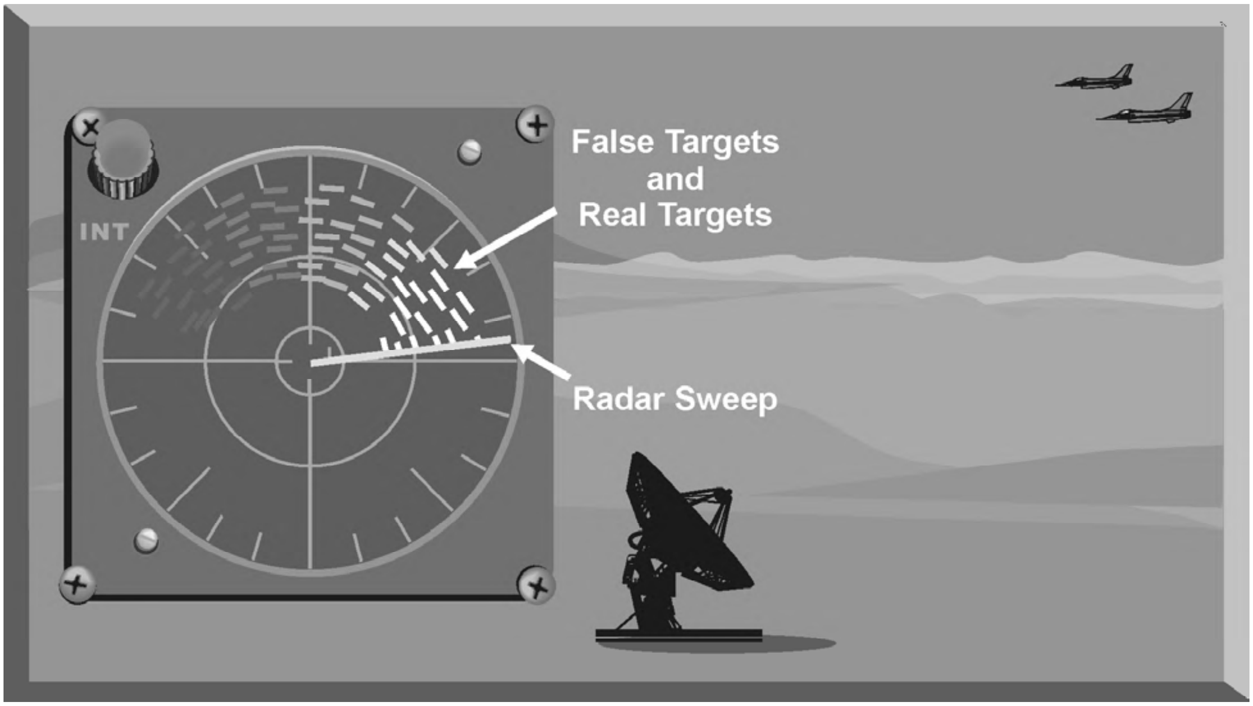

Figure 11-2. False Target Generation

a. To generate false targets, the deception jammer must tune to the frequency, PRF, and scan rate of the victim radar. The jamming pulse must appear on the radar scope exactly like a radar return from an aircraft. Multiple false targets greater in range than the jammer are generated by delaying the transmission of a jamming pulse until after the victim radar pulse has been received. False targets closer in range are generated by anticipating the arrival of a radar pulse and transmitting a jamming pulse before the victim radar pulse hits the aircraft. If the victim radar employs a jittered PRF, only targets greater in range can be generated.

b. To generate different azimuth false targets, the deception jammer synchronizes its transmitted pulse with the victim radar's sidelobes. Due to their

reduced power, when compared to the main beam, sidelobes are difficult to detect and analyze. The receiver in the deception jammer must be sensitive enough to detect these sidelobes and not be saturated by the power in the main radar beam. A false target deception jammer must inject a jamming pulse that looks like a target return into these sidelobes. To penetrate the radar sidelobes requires a lot of power. However, the power must be judiciously used. If a powerful jamming pulse is injected into the main beam, the false targets will be easy to detect. Most false target jammers vary the power in the jamming pulse inversely with the power in the received signal, on a pulse-by-pulse basis. This means the repeater jamming signal is at minimum power when the main beam of the victim radar is on the aircraft and at maximum power when the sidelobes are being jammed. To effectively generate false azimuth targets, the jammer must have a receiver with a wide dynamic range to detect both the main beam and the sidelobes. In addition, the jamming system must be able to generate high power that can be effectively controlled by the receiver.

c. To generate moving false targets, the deception jammer must synchronize with the main beam and the sidelobes in frequency, pulse width and PRF. Amplitude modulated jamming signals, with variable time delays, are transmitted into the sidelobes of the victim radar. The variable time delay provides a false target that changes range, either toward or away from the radar, depending on the time delay. The amplitude modulation provides false azimuth targets that appear to be moving.

d. The effectiveness of false target generation is based on the credibility of the generated false radar returns. If the victim radar can easily distinguish between false returns and target returns, the technique is a failure. The false returns must look identical to an aircraft return. The radar return on the victim radar scope should have the same intensity, depth, and width as a target return.

(1) Power determines the false target intensity when it is displayed on the victim radar scope. Varying jammer output power inversely with received power ensures that each false target has nearly the same intensity as a true target return. The depth, or thickness, of the false target depends on the pulse width of the victim radar. By matching the pulse width of the jamming pulse with the pulse width of the victim radar, the jammer can generate false targets with the same depth as a real target return.

(2) The width of the false target depends on the antenna pattern of the victim radar. This can pose a problem for false-target deception jammers. Because the jamming pulse is transmitted the entire time the radar beam is on the jammer, the width of a false target will tend to be greater than a real target return. Aircraft radar return varies with main beam cross-section. To correct this problem, most false target deception jammers use random modulation in the power of the transmitted pulses. This will vary the width of the false targets and make them look more like the variable returns of actual targets.

3. RANGE DECEPTION JAMMING

Although a specific TTR can track multiple targets and direct multiple weapons, the tracking circuit must select a single target return and track it while ignoring all other returns. Target selection is done by using gate bins. The range gate is used as the primary gate for target selection. A range gate is an electronic switch that is turned on for a period of microseconds based on a certain range or time delay after a pulse is transmitted (Figure 11-3).

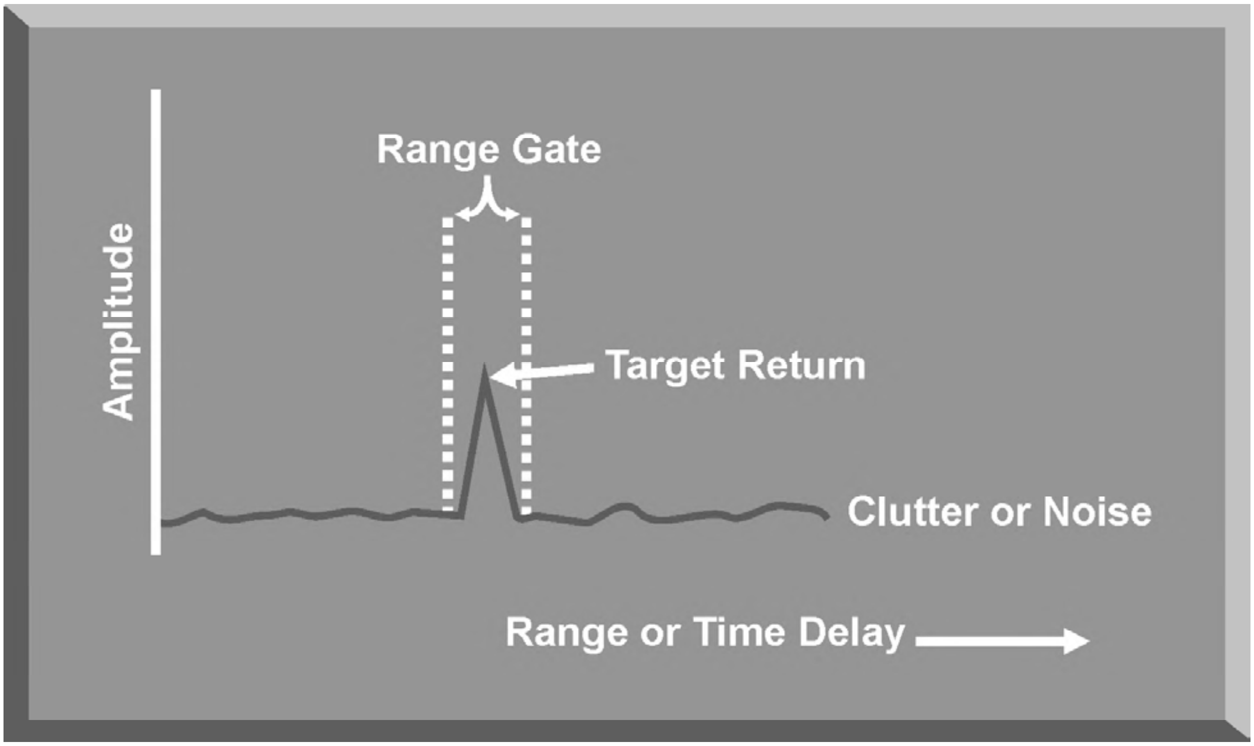

Figure 11-3. Range Gate Tracking

a. Range deception jamming exploits any inherent weakness in a TTR's automatic range gate tracking circuits. When a TTR's range gate locks on to an aircraft, the range deception jammer detects the radar signal. The range deception jammer then amplifies and retransmits a signal much stronger than the radar return. This retransmitted signal, called a cover pulse, is displayed in the range gate with the target signal (Figure 11-4).

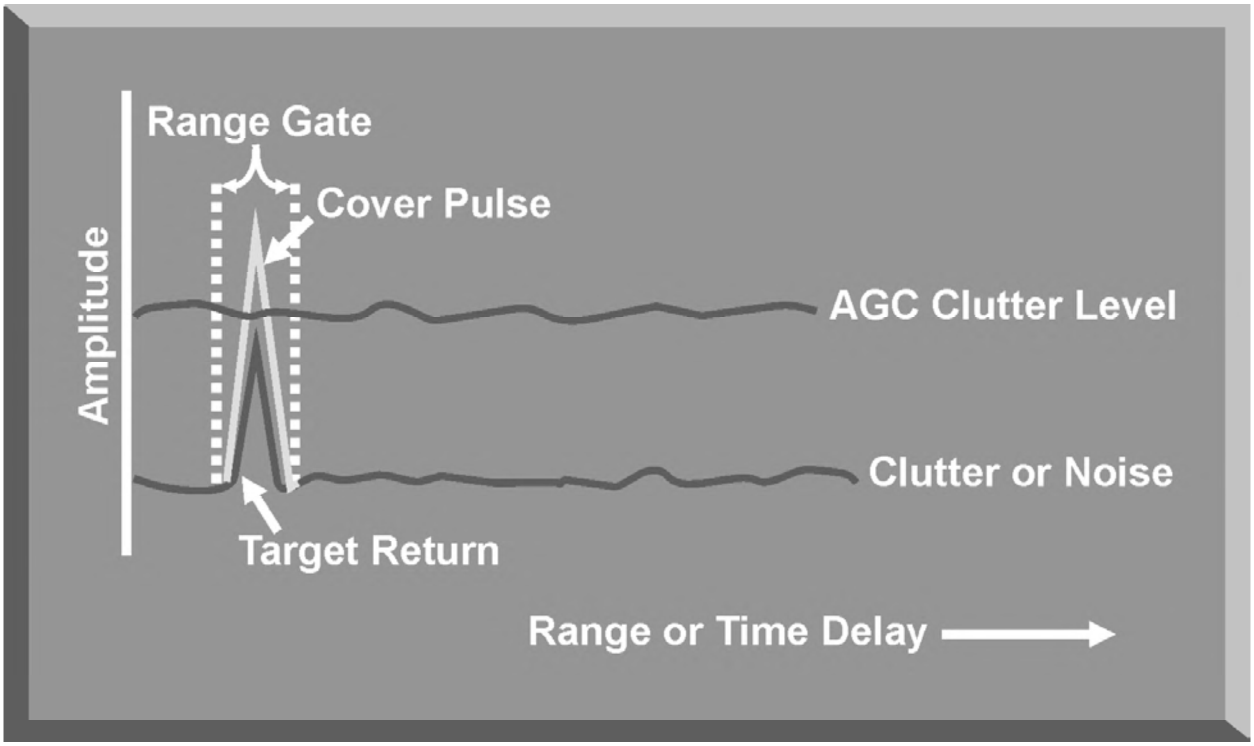

Figure 11-4. Range Gate Jamming Cover Pulse

b. The automatic gain control (AGC) circuit lowers the gain in the range tracking gate to control the amplitude of the cover pulse in the range gate. Reduced gain causes the real target return to be lost, and the range gate only tracks the jamming signal. This is known as range gate capture.

c. Once the range gate is captured by the cover pulse, a technique called range gate pull-off (RGPO) is employed (Figure 11-5). The deception jammer memorizes the radar signal and introduces a series of time delays before retransmitting. By increasing these time delays, the range gate will detect an increase in range and automatically move off to a false range. Once the range gate has moved well away from the real target, the range deception jammer shuts down, and the radar range gate is left with no target to track. The range gate breaks lock and the TTR must again go through the process of search, acquisition, and lock-on to re-engage the target.

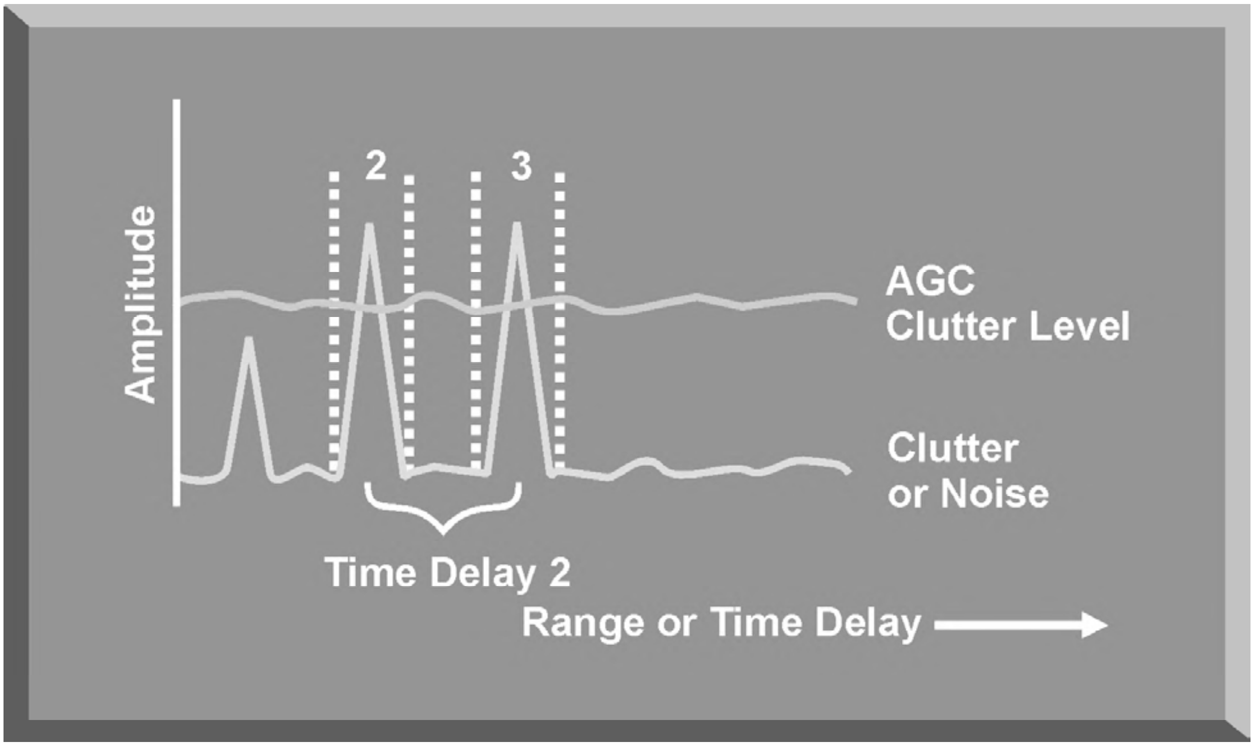

Figure 11-5. Range Gate Pull-Off

d. There are several advantages of range deception jamming, especially when used as a self-protection technique. It can generate sufficient errors to deny range information and is effective against most automatic range tracking systems. This technique does not require a large amount of power, just enough to cover the radar return of the aircraft. If the time delays are not exaggerated, an operator may not detect the loss of range lock-on until after a missile has been fired. The insidious nature of range deception jamming may generate enough miss distance to save the aircraft and pilot.

e. There are disadvantages to using range deception jamming. First, it can be defeated by a trained radar operator. If the operator detects a problem with the automatic range tracking circuit, the system can be switched to manual range tracking mode to defeat RGPO. Also, if the threat system is still able to track the aircraft's azimuth and elevation, range information may not be required to complete target engagement. To maximize range deception jamming effectiveness, it should be employed in conjunction with azimuth and elevation jamming. Finally, this type of range deception jamming is not effective against a leading-edge range tracking system. A leading-edge tracker will not see the delayed cover pulse. As the cover pulse moves off the target, AGC circuits reset the gain to continue tracking the real target. The only way to defeat a leading-edge range tracker is with a deceptive jammer that anticipates the next radar pulse and sends a jamming cover pulse before it reaches the aircraft. This jamming technique can also be defeated by randomly varying the radar PRF.

4. ANGLE DECEPTION JAMMING

Angle deception jamming is designed to exploit weaknesses in the angle tracking loop of the victim radar. The specific technique depends on the tracking method used to derive azimuth and elevation information. Inverse amplitude modulation jamming is the main angle deception technique used against TWS radars. For conical scan radars, scan rate modulation and inverse gain jamming are used. Swept square wave (SSW) jamming is used against LORO tracking radars. Monopulse angle deception jamming will be covered separately.

a. The azimuth and elevation tracking loop for a TWS radar is based on target signal amplitude modulation. The inverse amplitude modulation jammer generates a signal with modulation exactly opposite the expected return. To accomplish this, the angle deception jammer must receive the radar signals from the tracking beams. The jammer responds with a signal of the same frequency, PRF, and scan rate synchronized to the inverse of the radar antenna pattern (Figure 11-6). This induces an error in the angle tracking gate that, over a series of scans, causes the radar to lose target angle tracking.

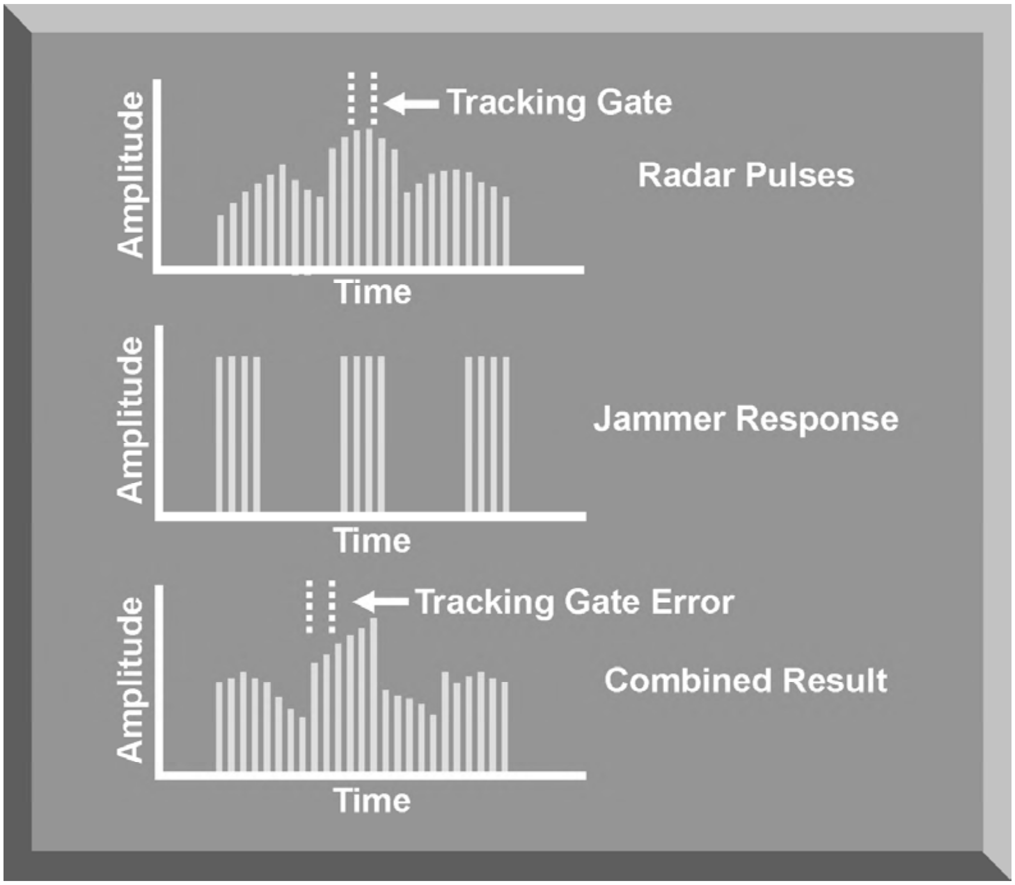

Figure 11-6. Inverse Amplitude Modulation Jamming

b. Inverse gain jamming is also effective against conical scan radars. Since conical scan radars use the phase of the target returns to generate error signals, an inverse gain deception jammer attempts to alter the phase by inducing fake signals into the antennas. In addition, by altering the amplitude of the signal, the jammer induces large errors into the tracking loop. To accomplish this, the jammer must determine the frequency, PRF, and scan rate of the victim radar. It then transmits signals that change the phase and amplitude of the target signal, resulting in a signal 180 degrees out of phase with the actual target (Figure 11-7). This 180-degree error rapidly drives the antenna off the target and causes break-lock.

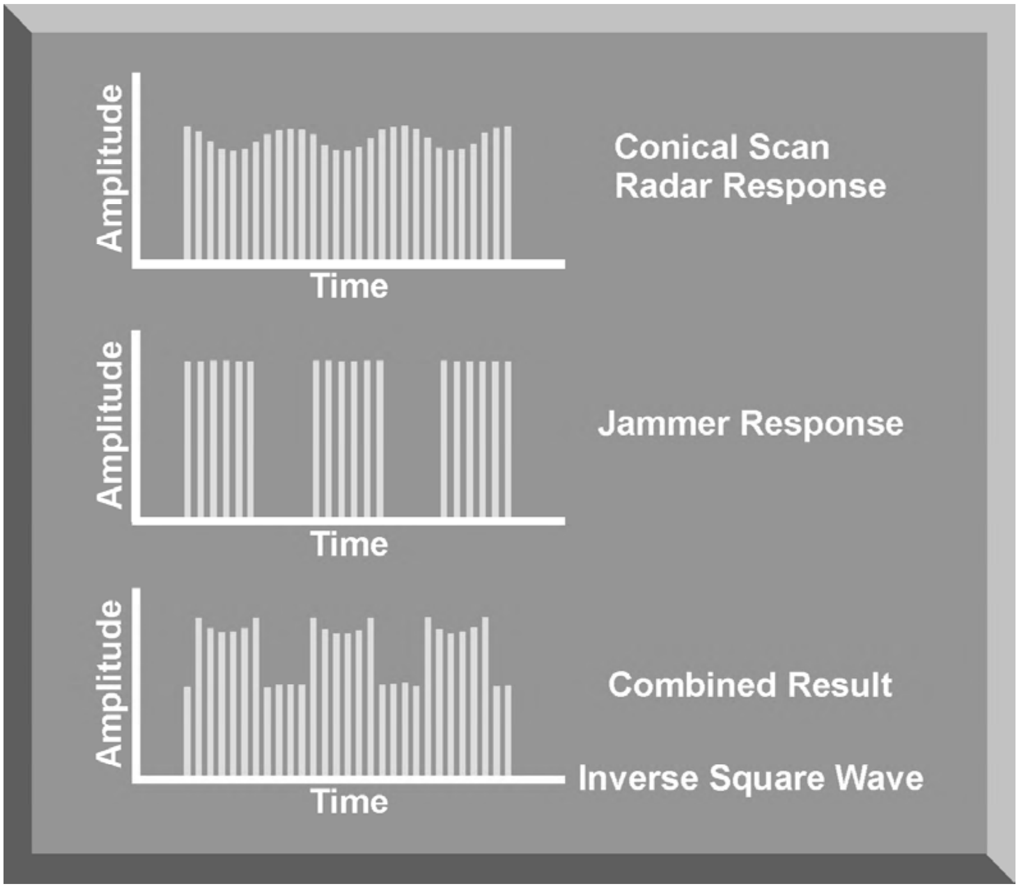

Figure 11-7. Inverse Gain Jamming

c. Scan rate modulation is also used against conical scan radars. This angle deception technique modulates the jamming pulse at or near the victim radar nutation frequency. As the modulation approaches the radar's nutation frequency, large error signals appear in the radar servo tracking loops, producing random gyrations in the antenna system, causing break-lock. This technique is most effective if the modulation jamming is slowly swept in frequency until it matches the nutation rate.

d. Both inverse scan and scan rate modulation jamming require very little power and have proven extremely effective against TWS and conical scan radars. To be effective, however, the angle deception jammer must find the precise scan rate of the victim radar. The jammer must concentrate on one signal at a time, limiting the number of threat systems that can be jammed simultaneously. In a dense threat environment, this can be a severe limitation.

e. The effectiveness of inverse gain and scan rate modulation jamming led radar designers to employ antennas that scan only during the receiving function of the radar system. Generally, this is accomplished by using two antennas. The transmitting antenna illuminates the target. Receiving antennas scan to produce the amplitude modulation of the reflected signal for effective angle tracking. This technique is called Lobe-On-Receive-Only (LORO). Since the transmitting antenna does not nutate, or scan, angle deception jammers cannot detect the modulation required to generate effective inverse gain modulation. Swept square wave (SSW) jamming is the angle deception technique developed to counter LORO angle tracking.

f. SSW jamming continuously varies the frequency of amplitude modulation on the jamming pulse over an expected range of nutation or scanning frequencies. This range is established by either electronic intelligence (ELINT) data on a particular system, or by exploitation. The dotted line in Figure 11-8 shows a threat's nutation or scan frequency. As the frequency of the modulated jamming pulse approaches the threat scan frequency, it induces errors in the angle tracking loop of the victim radar. The longer the SSW jamming stays near the scan frequency, the greater the induced errors. It is important that the sweep rate of the modulating jamming be slow enough to maximize its impact on the victim radar.

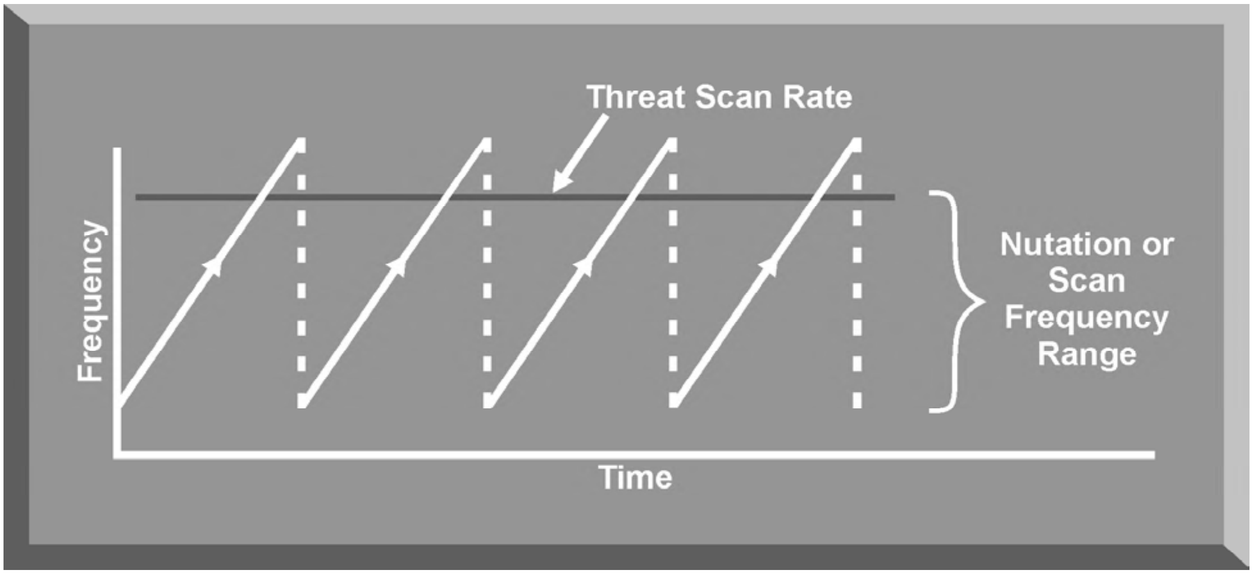

Figure 11-8. Swept Square Wave Jamming

5. VELOCITY DECEPTION JAMMING

Pulse Doppler and continuous wave (CW) radars track targets based on velocity or Doppler-shifted frequency (Figure 11-9). The objective of velocity deception jamming is to deny velocity tracking information and generate false velocity targets. The primary techniques include velocity gate pull-off (VGPO), Doppler noise, narrowband Doppler noise, and Doppler false targets.

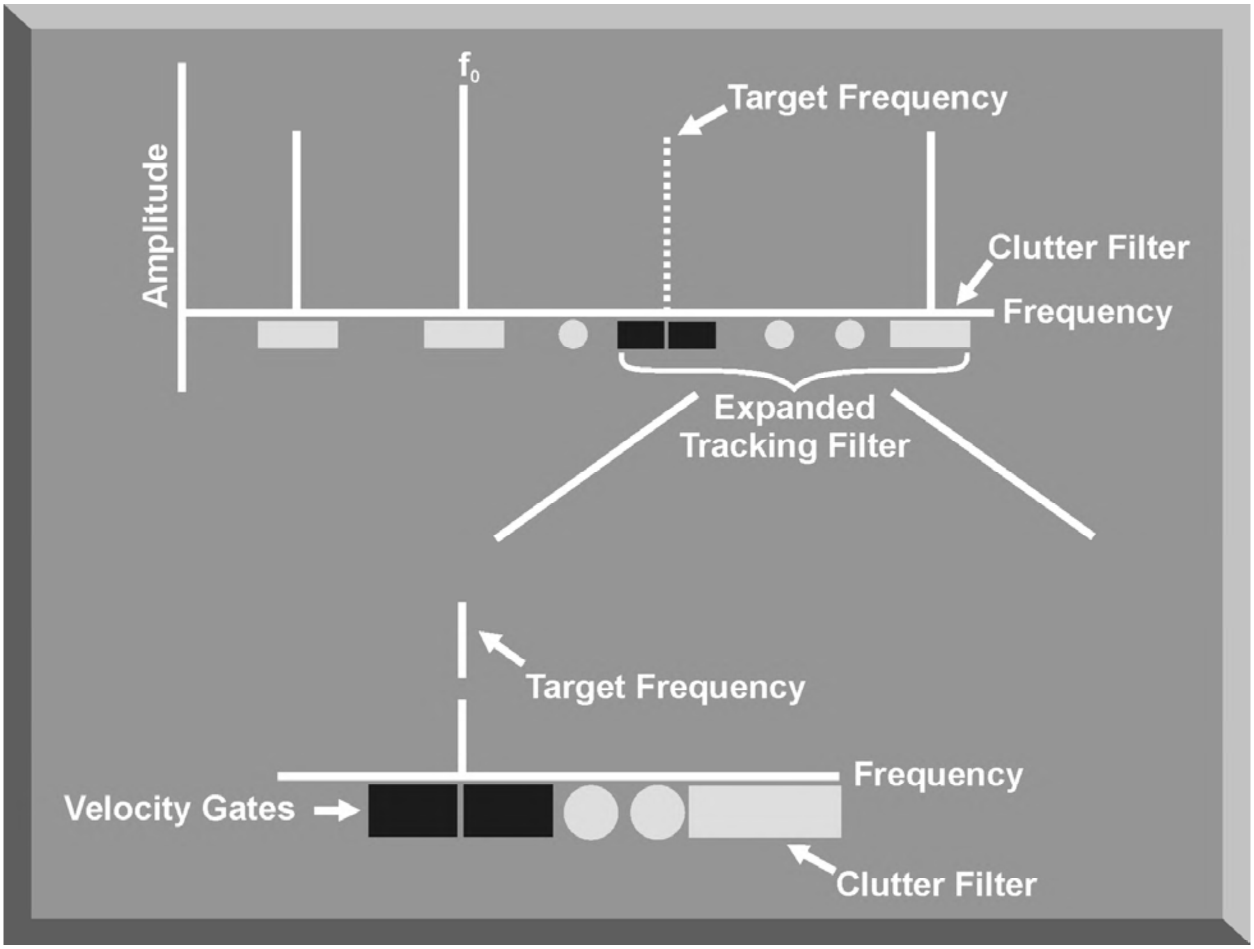
f_0
Target Frequency
Amplitude
Clutter Filter
Frequency
Expanded
Tracking Filter
Target Frequency
Frequency
Velocity Gates
Clutter Filter

Figure 11-9. Velocity Tracking Gate

a. Velocity gate pull-off counters pulse Doppler or CW radars by stealing the velocity gate of their automatic tracking loop. The objective of VGPO is to capture the Doppler velocity tracking gate by transmitting an intense false Doppler signal. Then the frequency of the false signal is changed to move the tracking gate away from the true target Doppler. This is analogous to the RGPO technique used against the range gate tracking loop.

(1) To accomplish an effective VGPO technique, the jammer receives the CW or pulse Doppler signal. It then retransmits a CW or pulse Doppler signal that is higher in power than the return from the aircraft, but at approximately the same

Doppler frequency (Figure 11-10). It is important that the frequency of this initial jamming pulse appears within the same velocity tracking filters as the target return or the victim radar will disregard it. The frequency band of the Doppler tracking filters is an important piece of intelligence information. The velocity tracking gates are quite narrow, roughly 50 to 250 MHz. Once the jamming pulse appears in the tracking gate, the automatic gain control circuit gains out the target return, and the jamming pulse has captured the velocity gate.

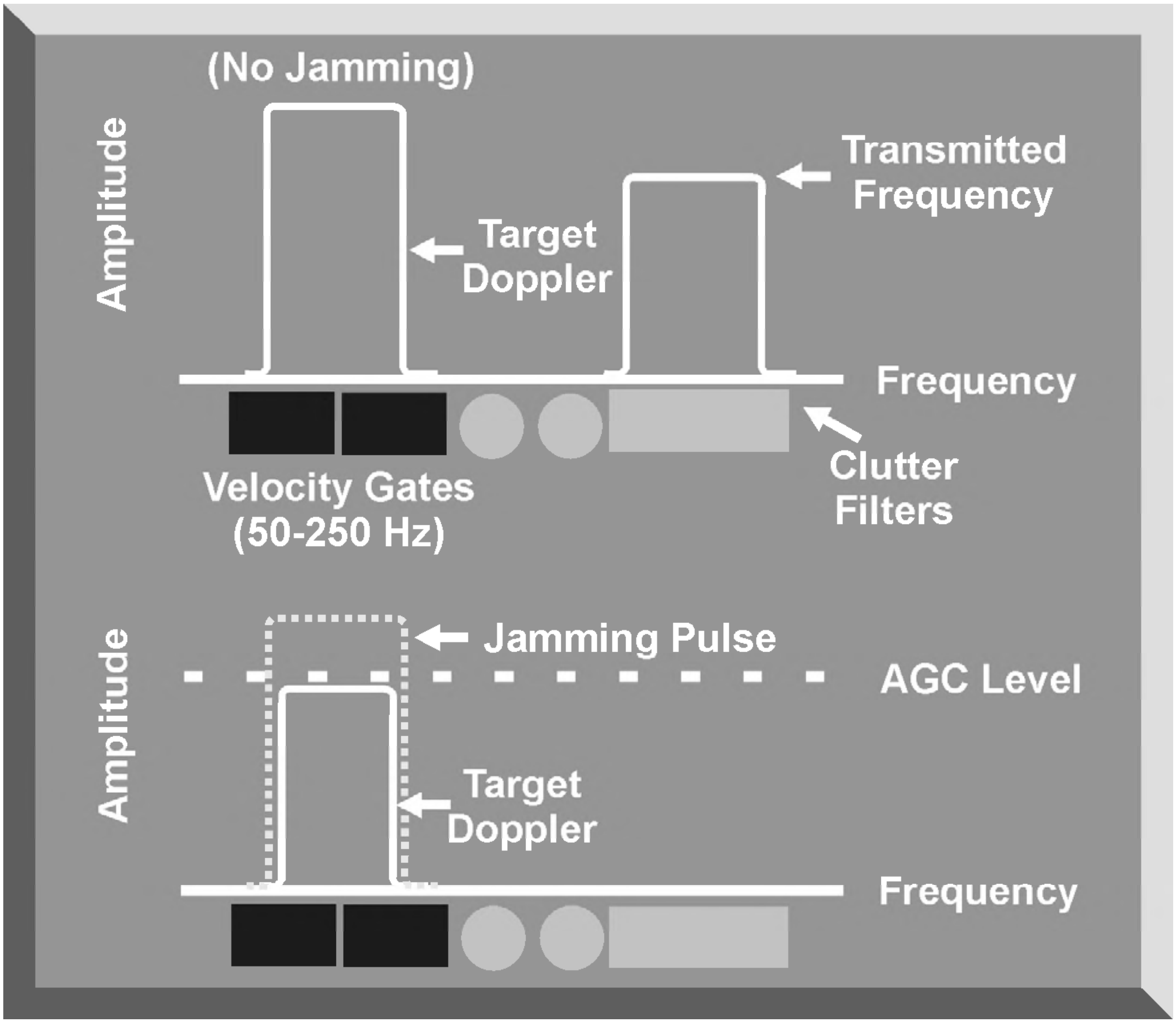
(No Jamming)
Amplitude
Transmitted Frequency
Target Doppler
Frequency
Clutter Filters
Velocity Gates (50-250 Hz)
Jamming Pulse
AGC Level
Amplitude
Target Doppler
Frequency

Figure 11-10. Velocity Gate Capture

(2) Once the jamming pulse has captured the tracking gate, the deception jammer slowly changes the Doppler frequency (Figure 11-11). This frequency shift is accomplished by several methods. The most common method uses frequency modulation (FM) within the jammer's traveling wave tube (TWT). By varying the TWT voltage, the Doppler frequency of the jamming pulse is changed

linearly, and the radar tracking gates follow the jamming pulse. By using FM, the jamming pulse can be moved in either a positive or negative direction, depending on the slope of the voltage. By slowly changing the frequency of the modulation, the jamming pulse pulls the tracking gates off the target. When the maximum offset has been achieved, nominally 5 to 50 kHz, the FM is "snapped back" to a minimum value, and the process is repeated to preclude target reacquisition.

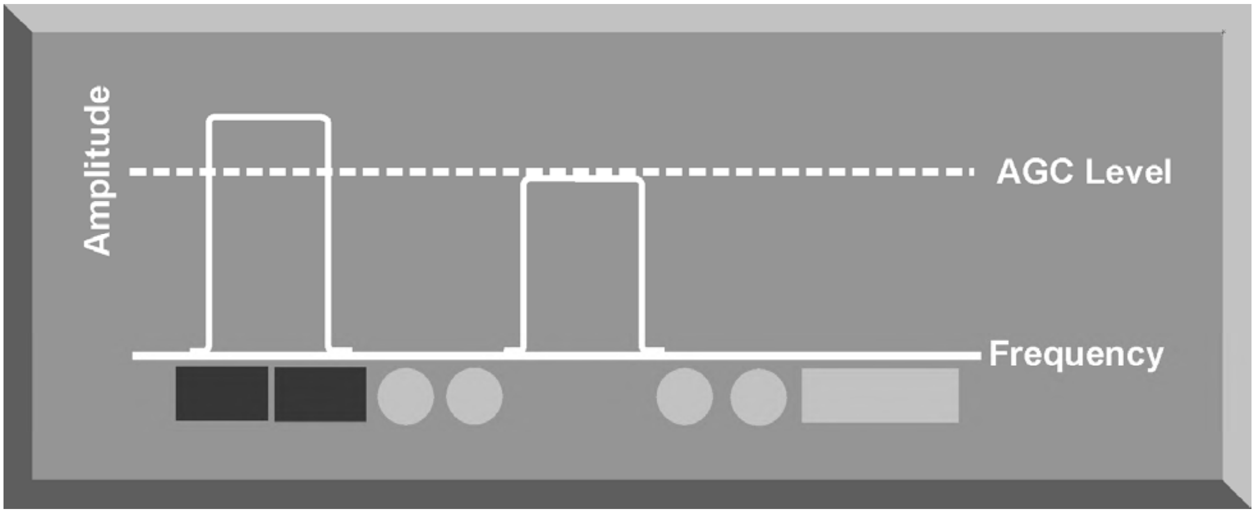

Figure 11-11. Velocity Gate Pull-Off

(3) The rate of change of frequency offset in a VGPO pulse is an extremely critical parameter. Many CW and pulse Doppler radars employ acceleration stops as part of the tracking gates. By differentiating the velocity outputs of the velocity tracking gates with respect to time, the velocity tracker computes target acceleration. Acceleration stops detect and reject unusually large changes in target acceleration. If the VGPO technique changes the frequency of the jamming pulse too rapidly, the tracking loop, with acceleration stops, will reject the jamming pulse and stay on the target. This means that an effective VGPO technique may take from one to ten seconds.

b. Doppler noise differs from most noise techniques in that it is a repeater technique. The jamming system must receive the pulse Doppler radar signal in order to generate an appropriate jamming pulse. Also, noise jamming output is done on a pulse-by-pulse basis and only lasts as long as the pulse duration, or pulse width, of the victim radar signal (Figure 11-12). The Doppler noise jammer receives each pulse and applies a random frequency shift, either positive or negative, to each pulse.

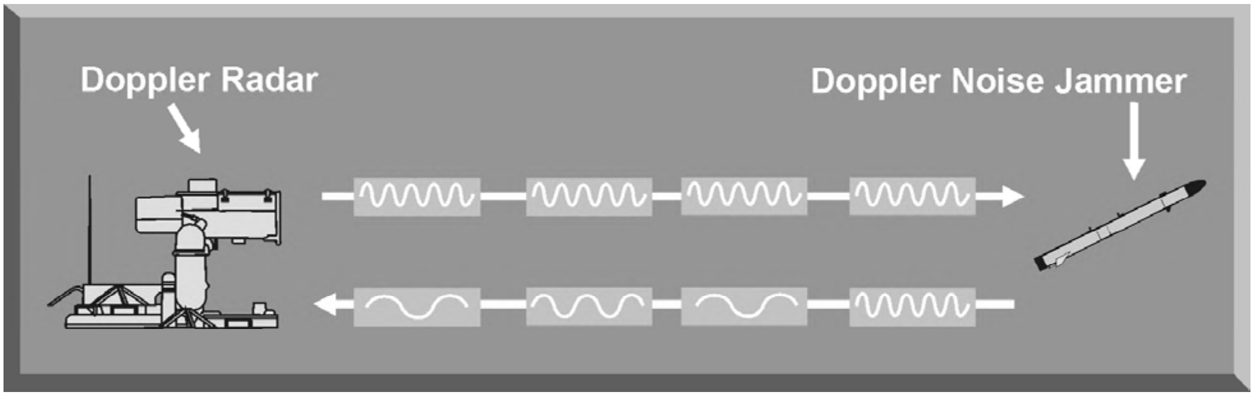
Doppler Radar
Doppler Noise Jammer

Figure 11-12. Doppler Noise Jamming

(1) When Doppler noise jamming pulses are processed by the signal processor, and the Doppler frequencies are sent to the velocity tracking gate, there are so many different velocities that the tracking gate cannot distinguish the target from the jamming. The random distribution of target velocities effectively masks the true target Doppler velocity. If the velocity tracking loop is not saturated, multiple false targets traveling at different speeds will be displayed.

(2) When a technique called Doppler noise blinking is employed, it interferes with the angle and velocity tracking within most semi-active radar missiles. Doppler noise blinking is accomplished by rapidly transmitting bursts of Doppler noise jamming (Figure 11-13).

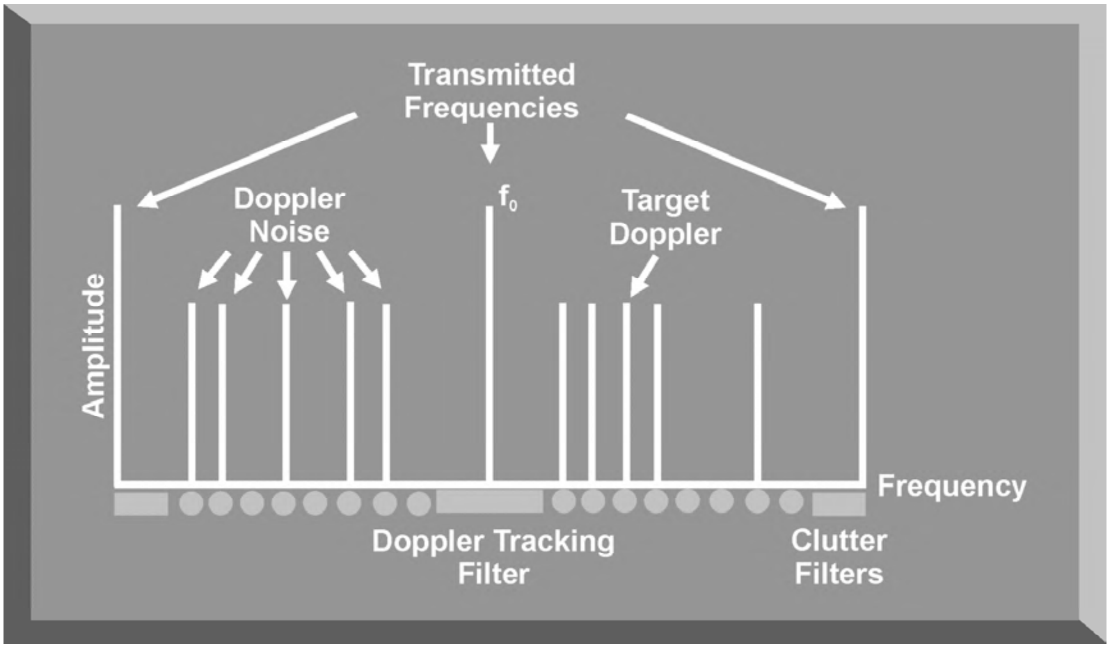
Transmitted Frequencies
f_0
Doppler Noise
Target Doppler
Amplitude
Frequency
Doppler Tracking Filter
Clutter Filters

Figure 11-13. Impact of Doppler Noise Jamming

(3) Doppler noise jamming is effective against most pulse Doppler radars and the semi-active missiles employed with these radars. One disadvantage, however, is that it is only effective against the velocity tracking loop. If range tracking is still available to the radar, Doppler noise may highlight the jamming aircraft. Another disadvantage is that Doppler noise requires a sophisticated jammer able to receive the victim radar pulse, generate random positive and negative frequency modulations on this pulse, and retransmit the jamming pulses at the PRF and pulse width of the victim radar. This requires an extremely fast signal processing capability and detailed intelligence information on the victim radar.

c. Narrowband Doppler noise is also a repeater technique. The jamming system receives the pulse Doppler radar signal and generates a noise jamming signal on a pulse-by-pulse basis (Figure 11-14). Narrowband Doppler noise requires detailed information on the frequency coverage of an individual velocity tracking filter, or velocity bin, employed by the victim radar. Once this frequency range is known, the jammer receives each pulse from the victim radar and transmits jamming pulses with a higher and lower frequency shift based on the real target Doppler. These frequency shifts are always within the frequency range of the velocity bin.

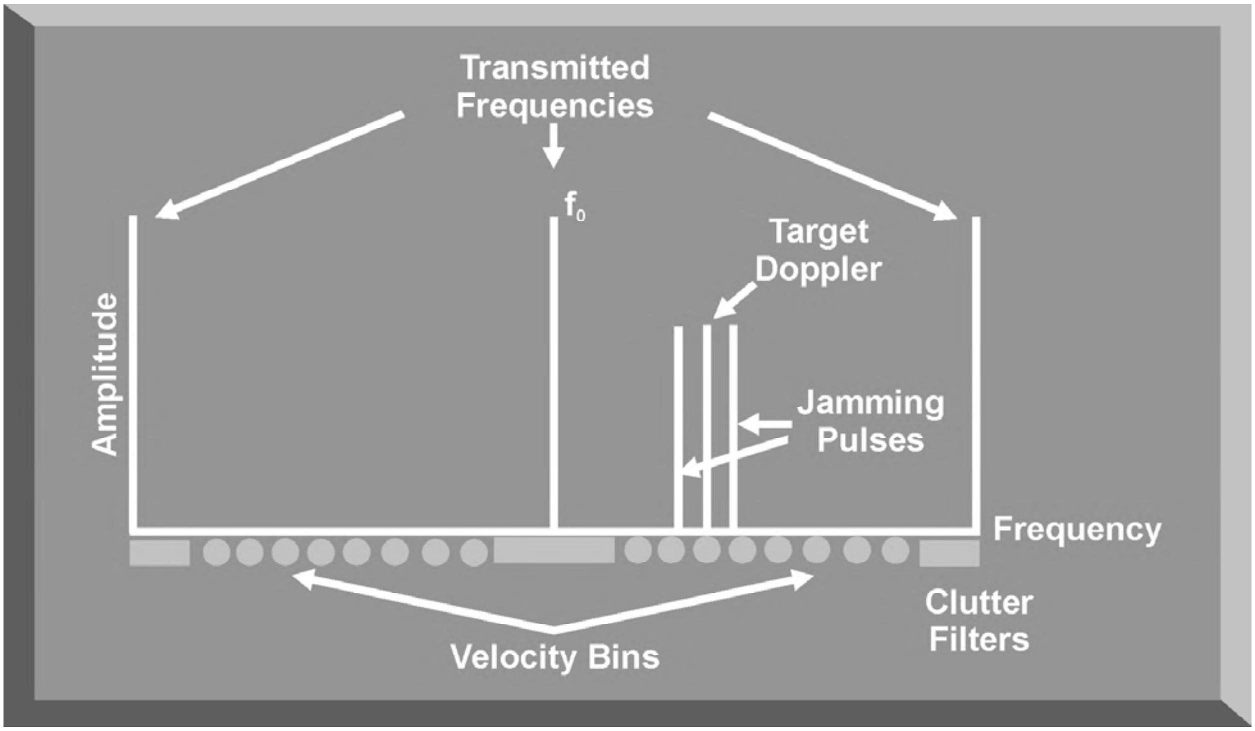

Figure 11-14. Narrowband Doppler Noise

(1) When these pulses are processed by the signal processor and the Doppler signals are sent to the velocity tracking gates, the particular bin that

contains the target Doppler also contains several other targets generated by the jammer. The victim radar signal processor attempts to distinguish the target Doppler from the jamming pulses. It raises the gain in the velocity tracking bins, thinking that the signal with the highest amplitude is the target. But, as the signal gain is increased, the target is "gained out" with the jamming signals and no target is displayed. This is called velocity bin masking and can completely deny target information to a pulse Doppler radar (Figure 11-15).

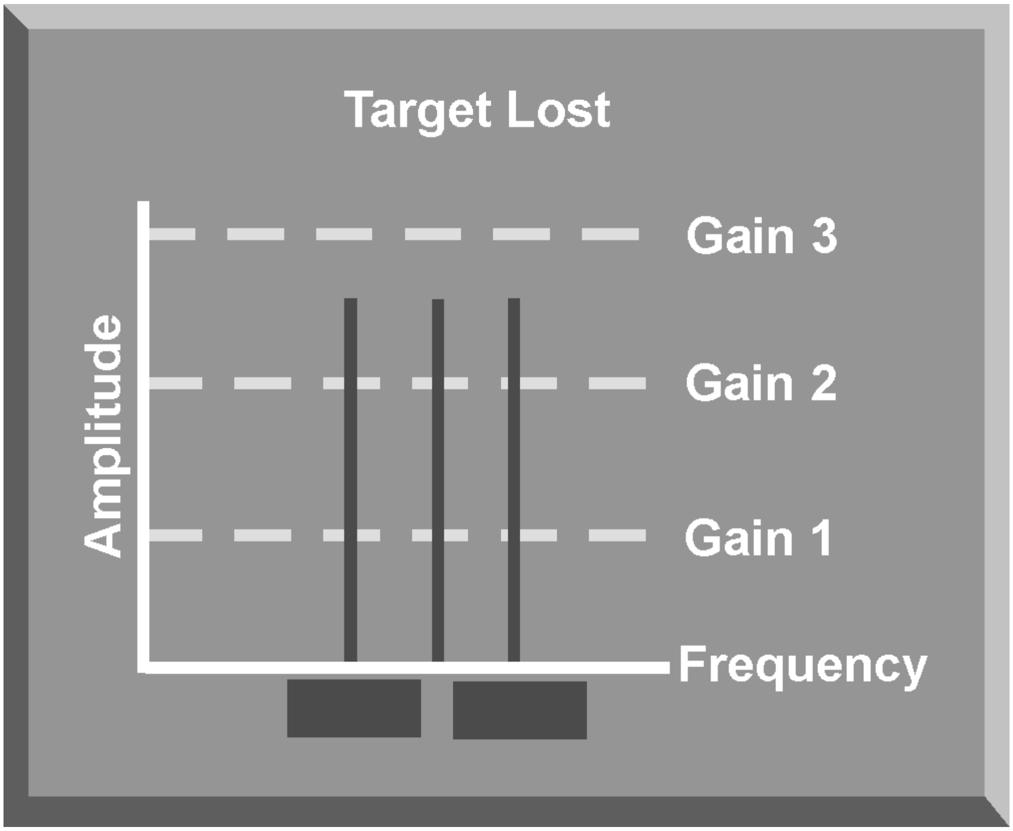

Figure 11-15. Velocity Bin Masking

(2) The advantage of narrowband Doppler noise is that it completely masks an aircraft's velocity from a pulse Doppler radar. The disadvantages include the following: When the victim radar can range-track an aircraft, narrowband Doppler noise highlights the aircraft's presence. To be effective, narrowband Doppler noise requires knowledge of the frequency range of the victim radar's velocity tracking bins, or filters. This detailed information may be available only through threat system exploitation. Finally, sophisticated signal processing and jamming systems are required to receive and transmit in the very narrow frequency band of the velocity bin.

d. Doppler false target jamming is normally used with narrowband Doppler noise or other deception techniques. Its purpose is to initially confuse the radar signal processor with multiple targets and then force the radar signal processor

to raise its gain levels in the velocity tracking loop. The Doppler false target jammer receives each pulse of the victim radar and applies a random frequency shift to a selected number of these pulses (Figure 11-16).

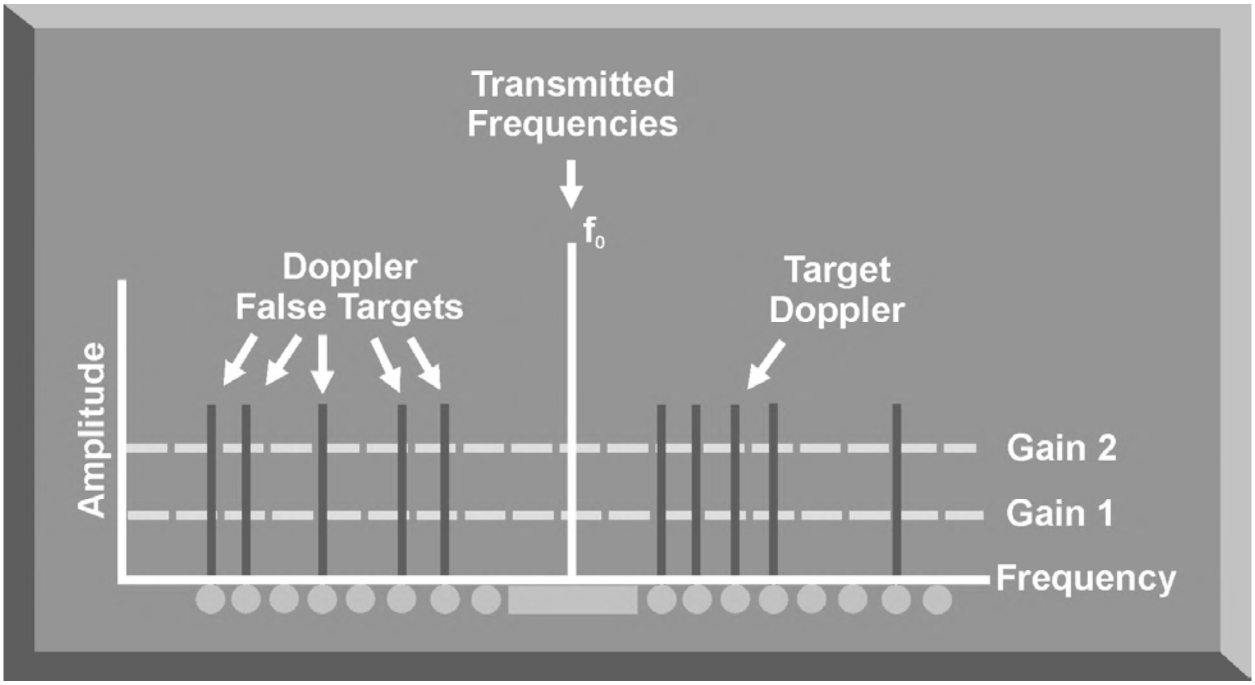

Figure 11-16. Impact of Doppler False Target Jamming

(1) The selected pulses are processed by the signal processor, and multiple Doppler frequencies are sent to the velocity tracking gate. In an attempt to distinguish the target from the jamming pulses, the signal processor increases the gain in each tracking filter, assuming the target Doppler has a higher amplitude than the jamming pulses. This increase in gain sets up the velocity tracking loop for a narrowband Doppler noise technique that will cause the real target to be lost among the generated false targets.

(2) The advantage of Doppler false target jamming is that it can initially confuse the radar signal processor and the radar operator as to the velocity of the real target. It also sets up the radar for narrowband Doppler noise technique and increases its effectiveness. The disadvantage is that the signal processor or the radar operator will eventually be able to distinguish the real target from the false targets based on its velocity. This jamming technique is much more effective when used in conjunction with other Doppler jamming techniques.

6. MONOPULSE DECEPTION JAMMING

The ability of monopulse tracking radars to obtain azimuth, range, and elevation information on a pulse-by-pulse basis make them extremely difficult to jam

(Figure 11-17). Amplitude modulation jamming used against conical scan or TWS radars, such as inverse scan and swept square wave, highlights a target, making monopulse tracking easier. Frequency modulation techniques, such as RGPO and VGPO, are equally ineffective. They serve as a beacon that aids the monopulse radar's target tracking ability. The monopulse radar may be able to track the jammer with more accuracy than tracking actual radar returns because target glint effects are absent from the jamming pulse. Monopulse angle jamming techniques can be divided into two main categories, system-specific and universal. Examples of system-specific jamming techniques include skirt frequency jamming, image jamming, and cross-polarization jamming. These techniques attempt to exploit weaknesses in the design and operation of specific monopulse radars. Cross-eye jamming, a universal technique, attempts to exploit all monopulse radar systems.

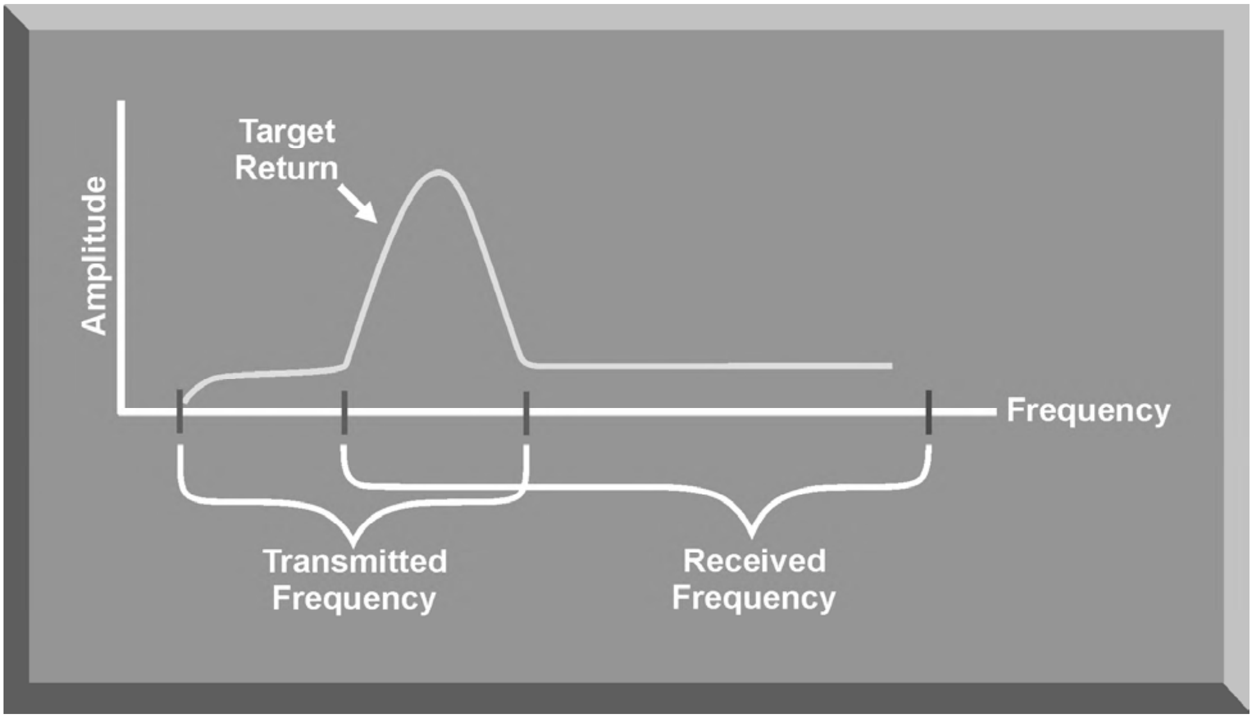

Figure 11-17. Monopulse Radar Receiver

a. Skirt frequency jamming, or filter skirt jamming, is designed to counter the monopulse receiver. Skirt frequency jamming is based on the fact that the intermediate frequency (IF) filter of the monopulse receiver must be correctly tuned to the transmitting frequency of the monopulse radar. It these two components are not exactly tuned, the target signal may be presented on the edge, or skirt, of the receiver IF filter. This offers an opportunity to inject a jamming signal into this skirt (Figure 11-18).

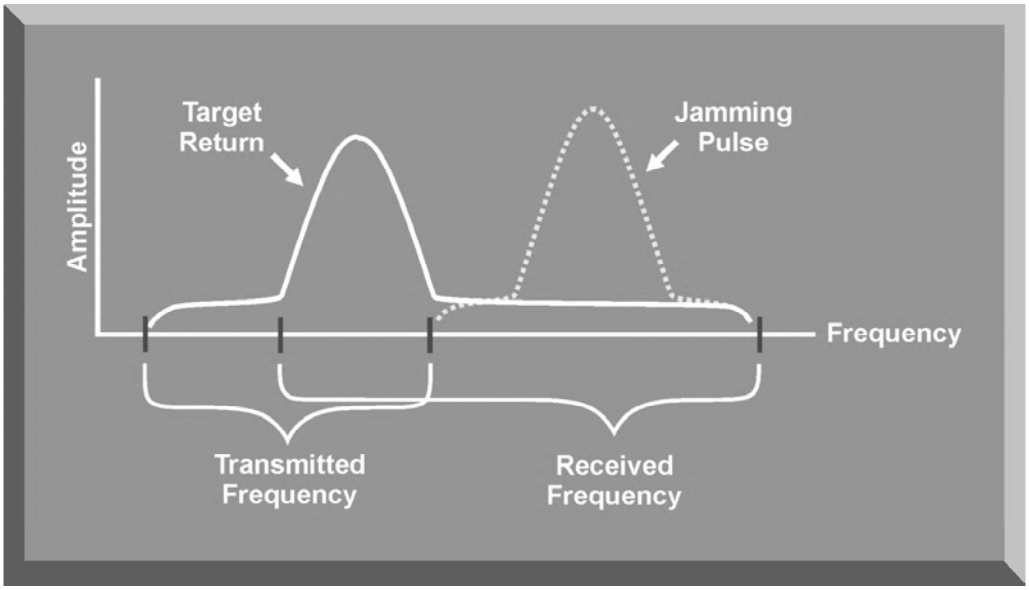
Amplitude
Target Return
Jamming Pulse
Frequency
Transmitted Frequency
Received Frequency

Figure 11-18. Filter Skirt Jamming Pulse

(1) Filter skirt jamming attempts to take advantage of this frequency imbalance by transmitting a jamming pulse tuned slightly off the radar transmitted frequency and in the middle of the receiver IF filter. This jamming pulse will generate a false error signal and drive the antenna away from the true target return.

(2) A well designed and maintained monopulse system does not have a frequency imbalance. The transmitter and IF filter frequencies will be identical. Jamming signals that are even slightly out of this narrow frequency range will not affect the monopulse tracking capability of the radar (Figure 11-19).

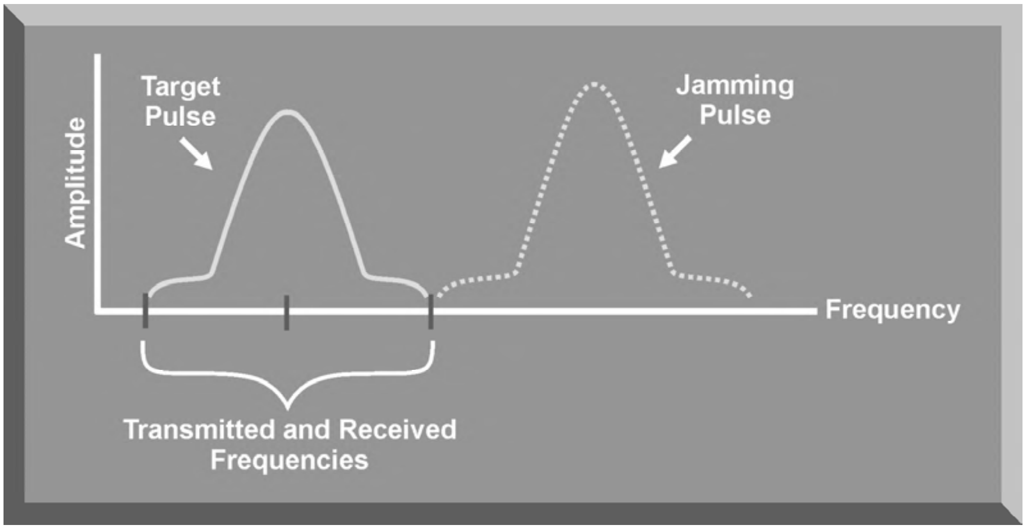
Amplitude
Target Pulse
Jamming Pulse
Frequency
Transmitted and Received Frequencies

Figure 11-19. Ineffective Filter Skirt Jamming

(3) Effective filter skirt jamming requires extensive knowledge of the internal operation of the IF filter. This information can normally be obtained only by system exploitation. Variances from radar to radar and frequency imbalance exists from one radar IF filter to another. This creates a high degree of uncertainty in the effectiveness of this technique.

b. Image jamming exploits another potential weakness in the monopulse receiver (Figure 11-20). Some monopulse receivers have a wide-open front end with no preselection before the mixer. If the jammer transmits a pulse at the intermediate, or image, frequency, but out of phase with this frequency, the phase of the target tracking signal will be reversed and the antenna will be driven away from the target (Figure 11-21). Effective image jamming requires detailed information on the operation of the monopulse receiver. Of particular importance is the image, or intermediate, frequency and whether the local oscillation frequency is above or below the transmitted frequency. This may require exploitation of the monopulse threat system. In addition, a well-designed monopulse system has preselection in the front end and will reject signals that are out of phase with the transmitted frequencies. This capability renders image jamming ineffective.

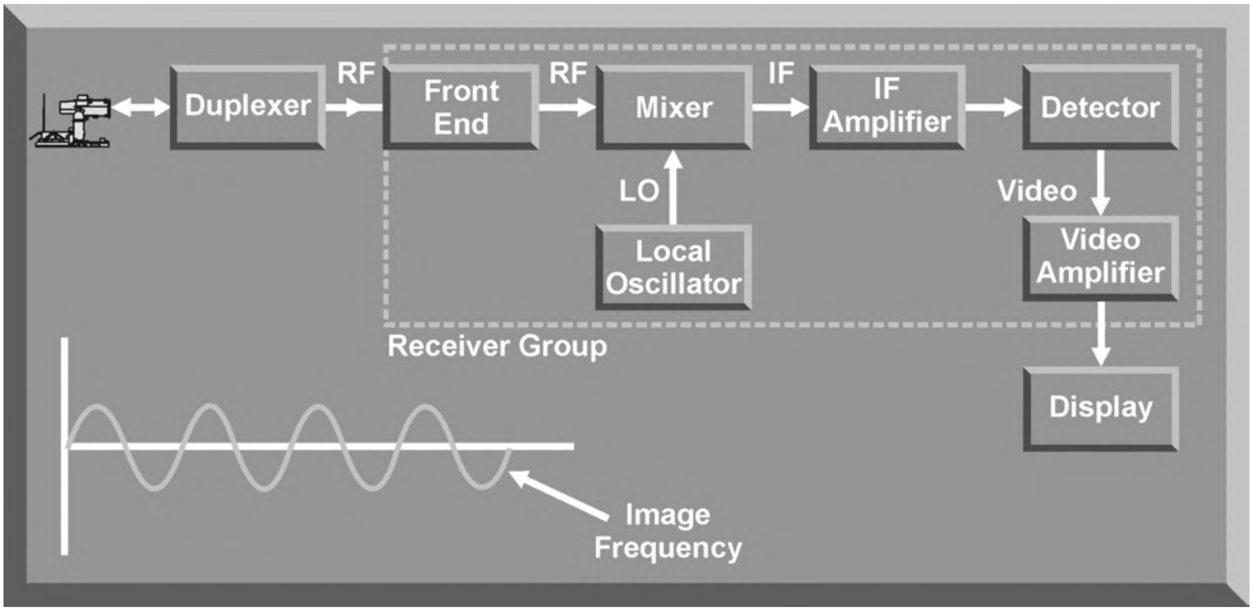

Figure 11-20. Monopulse Image Frequency

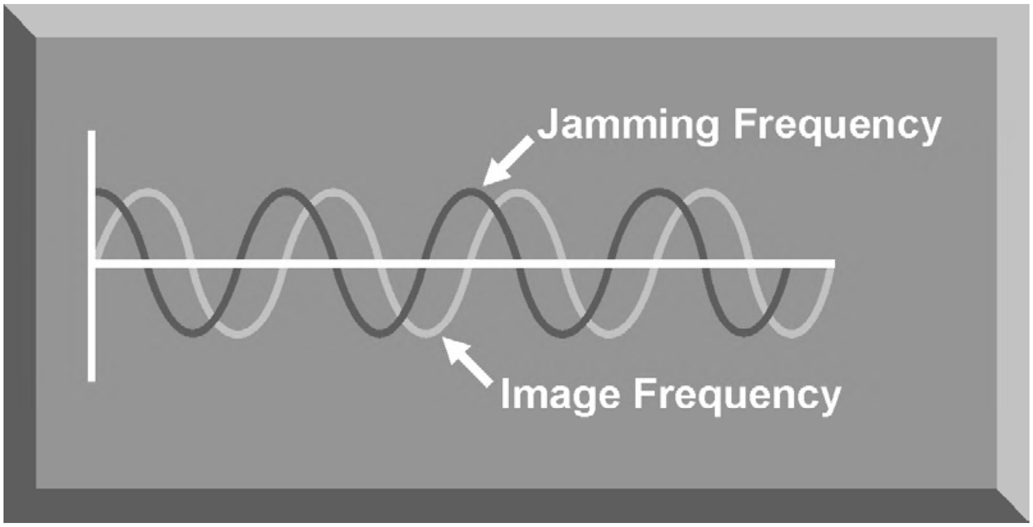
Jamming Frequency
Image Frequency

Figure 11-21. Monopulse Image Jamming

c. Cross-polarization jamming exploits the difference in the monopulse antenna pattern for a jamming pulse that is polarized orthogonal to the design polarization. The antenna pattern for a two-channel monopulse radar using sigma and delta beams shows the tracking point to be between the two beams (Figure 11-22). This is true if the radar is using its design polarization. However, the radar antenna also has a receiving pattern for a signal that is cross-polarized with the design frequency. For a cross-polarized signal, the tracking point is shifted one beamwidth to the right. This shift in the tracking point results in a target tracking signal that is 180° out of phase with the real signal. To be effective, a jamming signal polarized orthogonally to the design frequency of the radar would have to be 25 to 30 decibels, or about 1000 times, stronger than the radar signal.

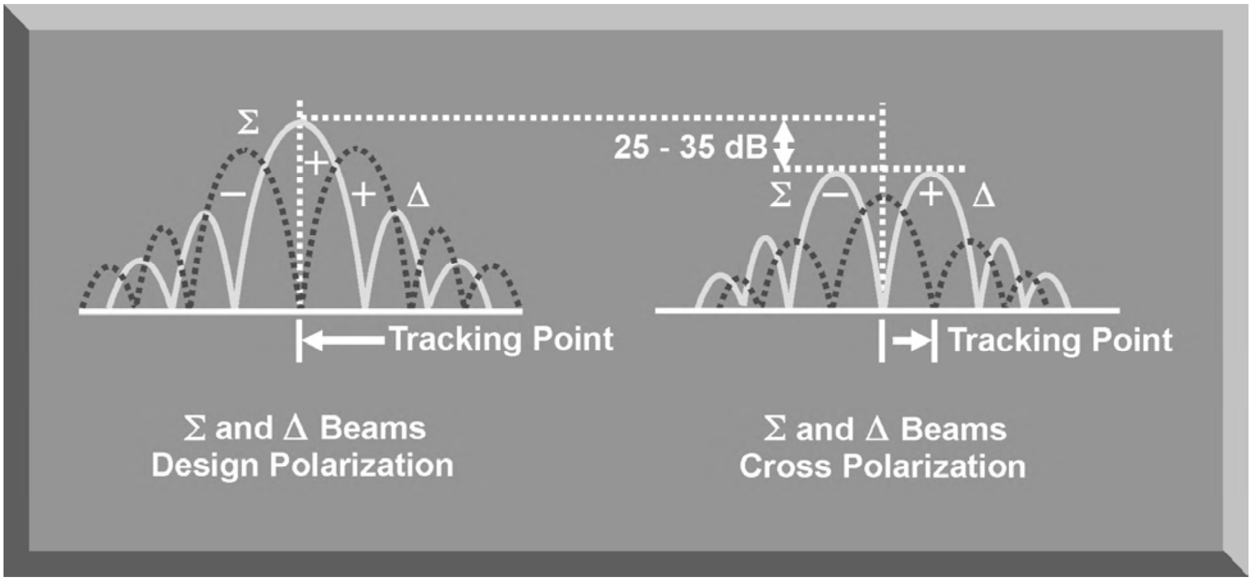
Σ
+
−
+
Δ
25 - 35 dB
Σ
−
+
Δ
Tracking Point
Tracking Point
Σ and Δ Beams
Design Polarization
Σ and Δ Beams
Cross Polarization

Figure 11-22. Cross-Polarization Antenna Pattern

(1) A cross-polarized jammer must receive and measure the polarization of the victim monopulse radar. The jammer then transmits a very high power jamming signal at the same frequency, but orthogonally polarized, to the victim radar. As a rule, the jamming signal must be 25 to 30 dBs stronger than the target return to exploit the tracking errors in the cross-polarized antenna pattern. Additionally, it must be as purely orthogonal to the design polarization as possible. Any jamming signal component that is not purely orthogonal will highlight the target and require more jamming power to cover the target return.

(2) A cross-polarized jammer must be able to generate a powerful jamming pulse that is polarized orthogonal to the victim radar. A cross-polarized jammer that generates the power and purity of polarization required to defeat monopulse angle tracking poses extreme technological challenges.

d. Cross-eye jamming is a complex technique that attempts to distort the wavefront of the beams in a monopulse radar and induce angle tracking errors. It exploits two basic assumptions of monopulse tracking logic in comparing target returns on a pulse-by-pulse basis. The first assumption is that a target return will always be a normal radar pulse echo. The second assumption is that any shift in amplitude or phase in a target return is due to the tracking antenna not pointing directly at a target. This condition generates an error signal and the antenna tries to null, but the amplitude or phase shifts.

(1) Cross-eye jamming attacks the two assumptions through a process of receiving and transmitting jamming pulses from different antennas separated as far apart as possible. In Figure 11-23, the phase front of a monopulse signal is received by the number 1 receive antenna, amplified by the repeater, and transmitted by the number 2 transmit antenna. The same phase front then hits receive antenna number 2, is shifted 180°, amplified by the repeater, and transmitted by the number 1 transmit antenna. These two out-of-phase signals must be matched in amplitude and must exceed the amplitude of the target return.

(2) When these jamming signals arrive at the victim radar, the tracking loop attempts to null out the amplitude and phase differences. With two widely spaced jamming sources at different phases, the antenna never achieves a null position or tracking solution. The distance between antenna pairs is an important parameter that determines the effectiveness of cross-eye jamming. The wider the spacing between antenna pairs, the more distortion in the victim's wave front near the true radar return. Most fighter aircraft do not provide sufficient spacing between the antennas to maximize effectiveness. Effectiveness is also lost when the aircraft is abeam or going away from the radar. To further complicate matters, when the radar is directly in front of the aircraft, the jamming pulses must have a power at least 20 dBs above the target return. Cross-eye jamming can also be defeated with a leading-edge tracker that rejects jamming signals arriving at the antenna behind the target return.

(3) Countering monopulse angle tracking is the greatest challenge for self-protection jamming systems. Skirt jamming and image jamming have had limited success. Cross-polarization and cross-eye jamming techniques require complex and sophisticated circuitry and much power.

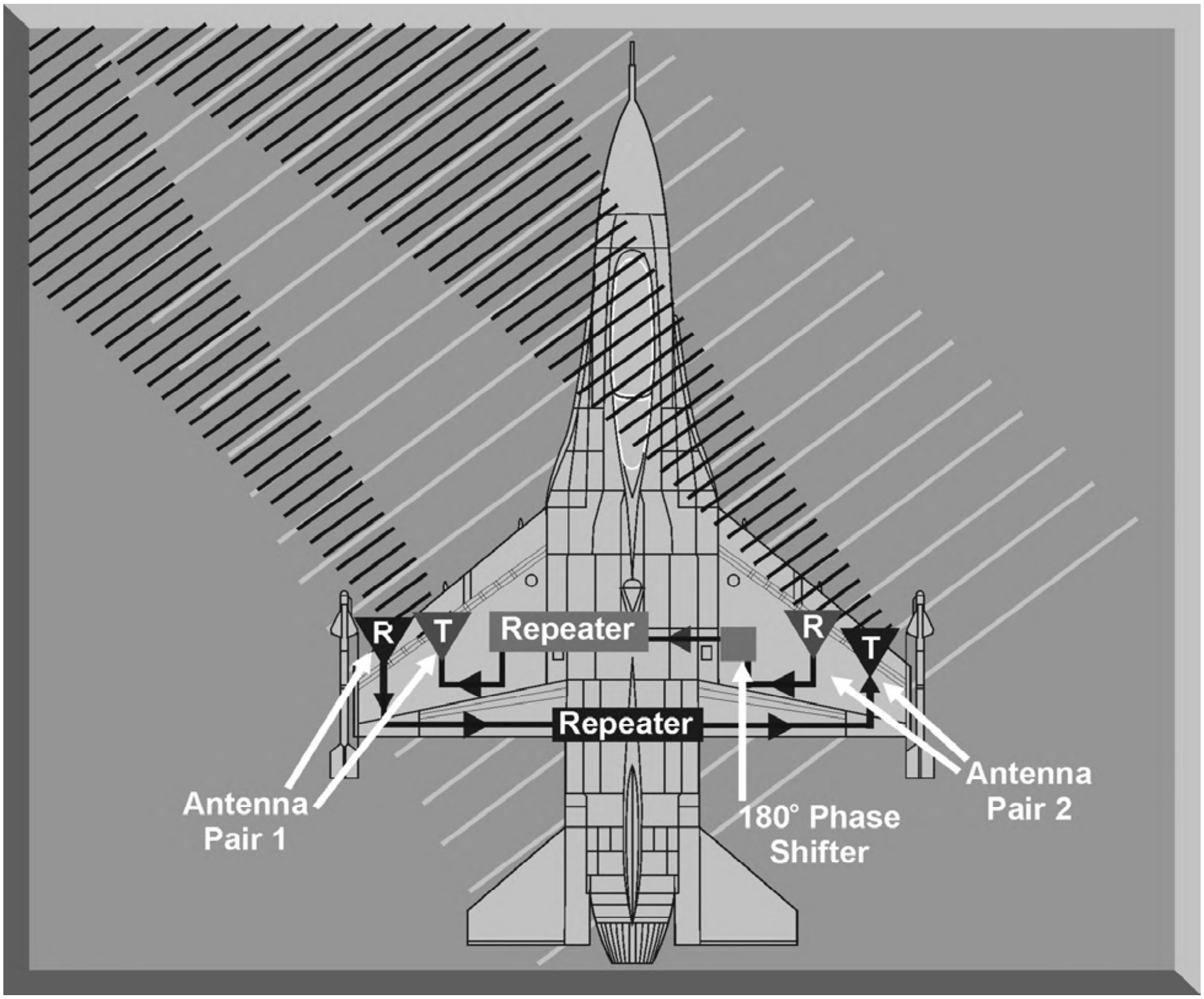

Figure 11-23. Cross-Eye Jamming

7. TERRAIN BOUNCE

Terrain bounce is a jamming technique used primarily at low altitude. It is used to counter semi-active, air-to-air missiles and monopulse tracking radars. The technique involves a repeater jammer that receives the radar or missile guidance signal. The jammer amplifies and directs this signal to illuminate the terrain directly in front of the aircraft. The missile or radar tracks the reflected energy from the spot on the ground instead of the aircraft (Figure 11-24).

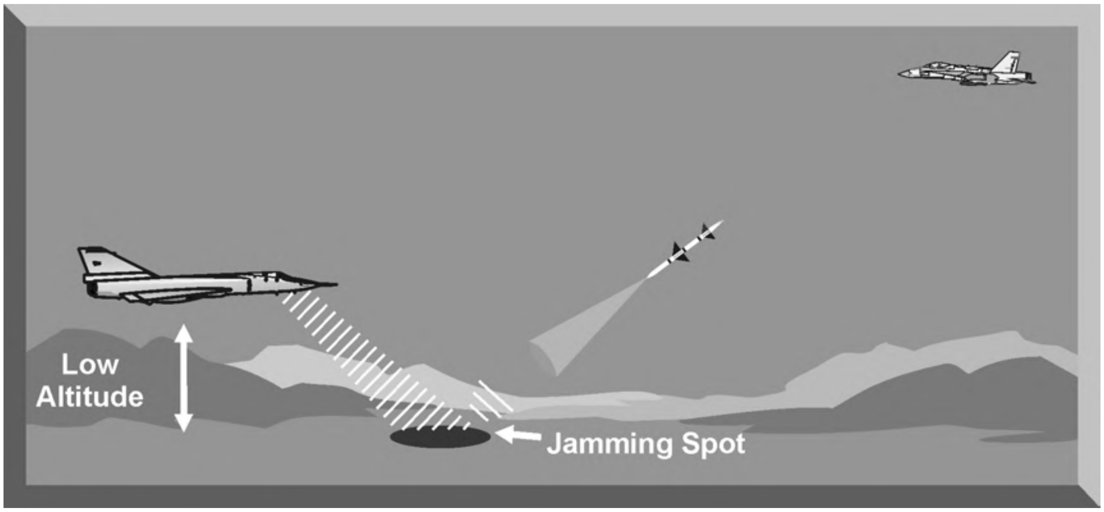

Figure 11-24. Terrain Bounce

a. To be effective, the terrain bounce jamming antennas should have a narrow elevation beamwidth and a broad azimuth beamwidth. This transmission pattern maximizes the energy directed toward the ground and minimizes the energy transmitted toward the missile or radar. To overcome signal losses associated with uncertain terrain propagation, the jamming system should also generate high jamming power. This ensures the energy reflected from the terrain is higher than the energy in the aircraft return. The terrain bounce jamming antennas should have very low sidelobes to preclude activation of any home-on-jam (HOJ) missile capability. For an air-to-air missile, the terrain bounce technique should be activated at long range. This will initially put the aircraft and the jamming spot in the same resolution cell. As the range decreases, the missile will be decoyed by the higher power in the jamming spot.

b. Some problems associated with terrain bounce jamming include the uncertainty of the signal scattering parameters of the various terrain features and the possible changes in signal polarization caused by terrain propagation. In addition, terrain bounce jamming can place maneuvering restrictions and maximum altitude limitations on the aircraft.

8. SUMMARY

There are several deception jamming techniques that can be employed to counter threat radar systems. The effectiveness of these techniques can be enhanced when they are employed in combination. For example, the effectiveness of an RGPO technique is enhanced when an angle deception technique is also employed. Determining the most effective deception technique, or combination of techniques, can present a challenge to intelligence and engineering analysts. However, when employed with maneuvers and chaff, deception techniques can mean the difference between success and failure on the modern battlefield.

CHAPTER 12. DECOYS

1. INTRODUCTION

A decoy is a device designed to look to an enemy radar more like an aircraft than the actual aircraft itself. Decoys do three primary missions: they saturate the enemy's integrated air defense system (IADS), coerce the enemy into exposing his forces prematurely, and defeat tracking by enemy radar. This chapter will discuss saturation decoys, towed decoys, and expendable active decoys. Chaff and flare systems will be discussed in separate chapters.

2. SATURATION DECOYS

A saturation decoy is usually an expendable vehicle designed to emulate a penetrating aircraft. Its mission is to deceive and saturate an enemy's IADS. Employing multiple saturation decoys can force an IADS to devote critical resources to engage these false targets. This depletes enemy assets available to engage penetrating aircraft. In addition, ground or air launched saturation decoys can be used to stimulate the IADS, to collect intelligence data, or to initiate attacks by suppression of enemy air defense (SEAD) assets. The three main characteristics of saturation decoys are their electronic signature, their flight program, and their mission type.

a. Saturation decoys must present an electronic signature, or radar return, that is indistinguishable from the aircraft they are protecting. Decoys can do this by either passive or active measures, or use a combination of both. A passive decoy is essentially a flying radar reflector. The size, shape, and materials used in the decoy are optimized to ensure that the proper amount of radar energy is returned to the enemy radars. Active decoys employ radar repeater systems to receive the enemy radar signal, amplify it and send back a radar return of the proper size to confuse the enemy. Reflecting or transmitting the proper size radar return is critical for both passive and active decoys. A return that is too large or too small will allow the enemy radar operator to differentiate between decoys and aircraft, causing the decoys to be ignored.

b. To continue deceiving an enemy IADS, a decoy must do more than provide the proper-sized radar return. Possessing flight characteristics similar to the aircraft it is protecting increases the probability that the decoy will effectively deceive an IADS for a sustained period of time. Modern decoys can either be powered with rockets, miniature engines, or simply glide for very long distances based upon the altitude and airspeed of the jet that releases them. Additionally, their flight paths can be pre-programmed into an onboard autopilot, allowing the decoy to fly an independent ground track, thus increasing their appearance as attack aircraft worth tracking.

c. Saturation decoys carry out two of the three decoy missions. Launched in significant numbers, they can saturate or overburden an IADS. Meanwhile, their realistic electronic image and preprogrammed flight paths entice the enemy to turn on radars and show his forces.

(1) Saturation decoys launched in coordination with an attacking strike package force the enemy to take time to process meaningless tracks and tie up critical assets. In this role, decoys primarily work against the early warning network of the enemy IADS by presenting the IADS with numerous targets to sort and track. Resources committed to tracking decoys may not be available to track actual aircraft. Additionally, if an enemy knows that decoys are present, he may not commit any resources against targets for fear they are just decoys.

(2) Time of radiation or “emission control” is a critical factor for acquisition and target tracking radars. To be effective and survive on the battlefield, ground threat radars radiate as little as possible; too much time radiating allows ELINT collectors to find their location and either direct aircraft to avoid them or call in an attack upon them. Therefore, when a decoy can get a radar to emit, the radar is now essentially compromised and can be avoided or attacked. Getting the enemy's radars to emit is called “stimulating the IADS,” which is generally a precursor to any threat suppression mission.

(3) An extremely successful example of using decoys to stimulate the IADS was carried out in the Bekaa Valley in 1982. The Israelis opened the conflict by launching saturation decoys to successfully simulate an attack. While the Syrians reloaded, Israeli fighters attacked, destroying 17 of 19 Syrian SA-6s in the beginning of the battle. With the ground threat neutralized, the Israeli Air Force went on to destroy 85 Syrian fighters in the pure air-to-air conflict that resulted.

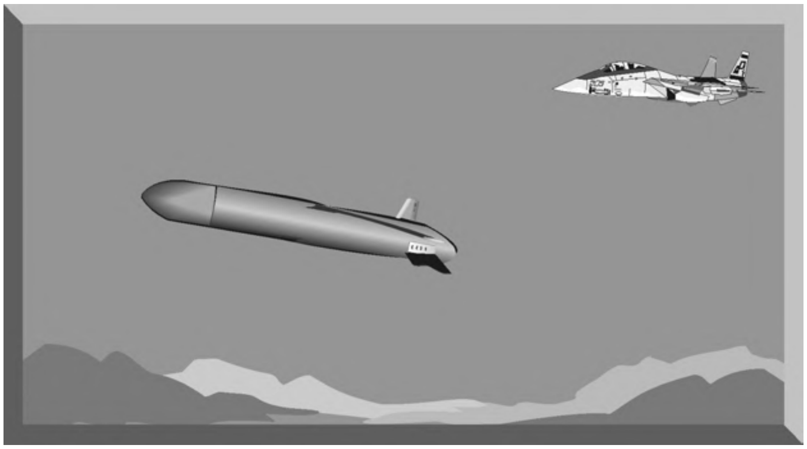

Figure 12-1. Tactical Air Launched Decoy

d. Two examples of saturation decoys are the Navy's Tactical Air Launched Decoy (TALD) in Figure 12-1 and the Air Force's proposed Miniature Air Launched

Decoy (MALD) in Figure 12-2. Both of these decoys can work actively or passively and both have pre-programmable flight paths. The TALD is an unpowered decoy normally launched from an F-14 Tomcat. The Air Force's MALD is a smaller jet-powered decoy also designed to be used by fighter aircraft. The MALD is 90 inches long, 6 inches in diameter, and has 25-inch wings that are foldable–essentially it is the size of an air-to-air missile. Because of its small size, the MALD can be carried into the target area before it is launched. Once launched it uses its speed, independent flight path, and electronically manipulated radar signature to make acquisition radars and target tracking radars mistake it for one of the attacking aircraft.

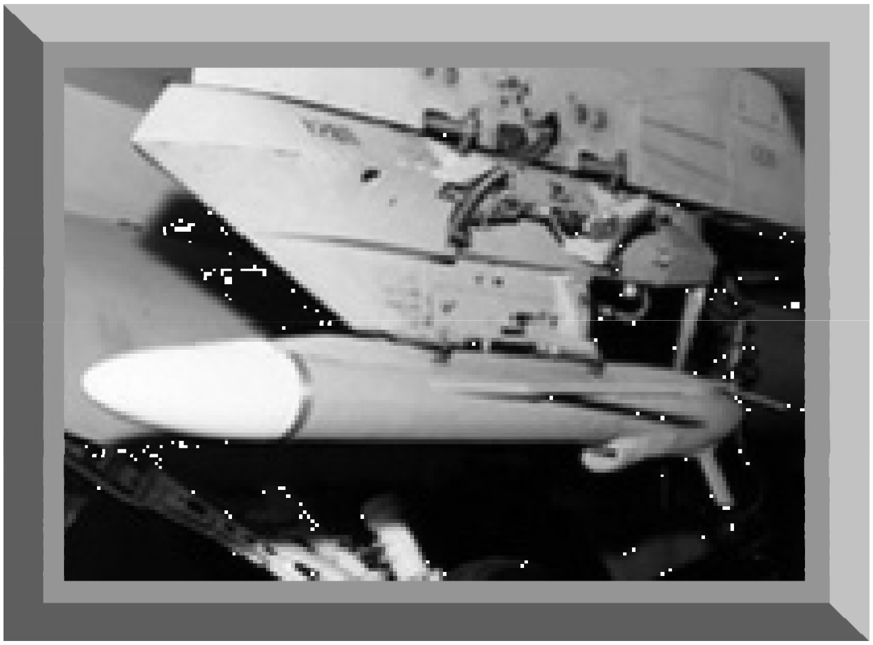

Figure 12-2. Miniature Air Launched Decoy (MALD)

3. TOWED DECOYS

A towed decoy is a small jammer that is physically attached to the aircraft (Figure 12-3). Unlike the saturation decoys that work against the IADS, the towed decoys are for individual aircraft survival. Towed decoys are designed to defeat enemy missiles in the final stages of an engagement; therefore, towed decoys, as well as other expendables, are known as endgame countermeasures. While towed decoys are primarily designed to provide sufficient miss distance between an attacking semi-active radar missile and the protected aircraft, they may also be effective against pulse Doppler radars and monopulse radars.

a. To be effective, the towed decoy must turn on within the threat radar's resolution cell after the radar is tracking the protected target. To successfully decoy the missile, the towed decoy must return radar signals with sufficient power to simulate a radar cross section (RCS) significantly larger than that of the

protected target. There are currently two generations of towed decoys on the market. Their primary difference lies in the connection each has with the aircraft towing it.

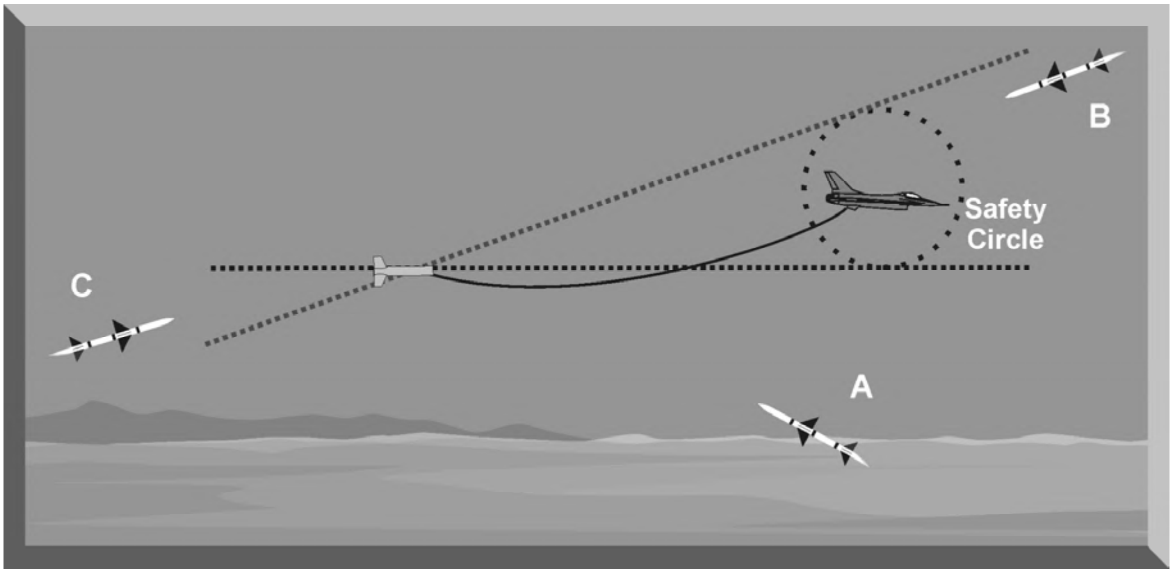

Figure 12-3. Towed Decoy

(1) The first generation of towed decoys contains a simple repeater jammer that enhances any signal it receives in the proper frequency range. The enhanced signal is stronger than the aircraft signal so the missile is lured towards the decoy. These decoys are stand-alone units that contain all the electronics, processors, receivers and transmitters within them. The only tie to the aircraft is for power and status. One of the big advantages of these simple repeater devices is that they do not require the exact frequency of the enemy radar systems to be effective, they will enhance any signal coming at them. An area of concern with the use of towed decoys is possible conflict between the onboard jamming system and the towed decoy. The onboard system could overpower the decoy, causing the attacking missile to ignore the decoy and track the aircraft.

(2) The second generation of decoys is tethered to the aircraft via fiber optic cable. Through this cable travels the different jamming modulations to be used by the decoy. These fiber optic towed decoys (FOTD) only contain the transmitters; the remaining items are in the jet or the pod. This system allows for more complex jamming through the decoy, including cooperative jamming between the aircraft and the decoy.

b. The separation required between the decoy and the aircraft is a primary consideration in developing a towed decoy system. The towed decoy should be positioned far enough behind the aircraft to preclude warhead fragments from missiles guiding on the decoy from also impacting the aircraft. Missile A in Figure

12-3 depicts a situation where the missile will detonate well outside of the aircraft's safety circle. From a pilot perspective, any restrictions on aircraft maneuvering imposed by a towed decoy are very important. The number of decoys that can be carried and the time required for decoy deployment are also important employment considerations.

c. Achieving 360° coverage is a primary limitation of a towed decoy system. When an aircraft equipped with a towed decoy is abeam a threat radar, the radar may be able to discriminate between the aircraft and the decoy. This is a function of the resolution cell of the radar. In addition, missiles approaching from a high-aspect angle, and above the aircraft (Figure 12-3 - Missile B), may fuse on the aircraft while guiding to the decoy. Missiles approaching from a low-aspect angle (Figure 12-3 - Missile C) may not fuse on the decoy and subsequently acquire and fuse on the aircraft. Finally, if the decoy is destroyed or lost, the time required to deploy a replacement decoy is critical, especially if the aircraft is engaged by multiple missiles.

d. An example of a fielded towed decoy system is the AN/ALE-50 (Figure 12-4). This first generation towed decoy system is found on Air Force F-16 and B-1 aircraft, and there is a version that is integrated into the ALQ-184 pod.

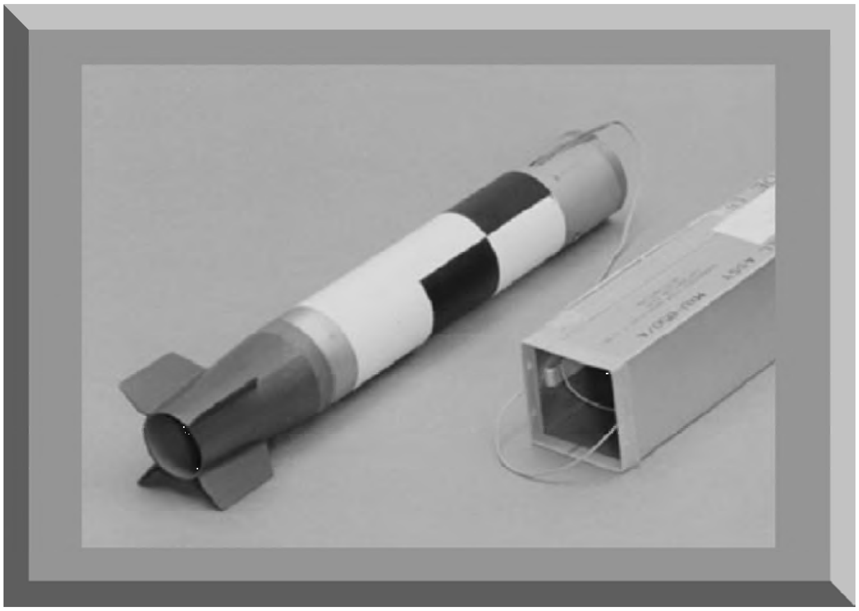

Figure 12-4. AN/ALE-50 System

(1) The system consists of a launch controller subsystem and towed decoys. The launch controller houses the decoy before it is launched, provides power to the decoy, and provides for the monitoring of the electronics. The decoy body is a factory sealed, self-contained unit with everything except for power. Power comes through the tether from the host aircraft; the decoy sends its operating status back through the tether to the aircraft.

(2) ALE-50 decoy use is cleared throughout the flight regime of the F-16 and B-1. Decoys can be deployed without being turned on, but once a decoy has been deployed, it cannot be reeled back in and must be severed before landing. Procedures are in place to reduce the chance that the onboard jamming system will negate the decoy.

(3) The ALE-50 towed decoy is a wideband RF repeater that provides self-protection EA by receiving, electronically amplifying, and retransmitting enemy radar RF signals. Upon receiving a threat radar signal, this simple repeater amplifies the signal and retransmits it. This provides the radar with two signals, one reflected from the aircraft and a stronger one from the decoy. With the signal from the decoy being the more attractive, the radar or missile guides towards the decoy. During combat operations over Kosovo, ALE-50 decoys were credited with saves for both F-16s and B-1s.

4. EXPENDABLE ACTIVE DECOYS

Expendable active decoys are designed to lure the tracking gates of an enemy's radar away from the aircraft. They are endgame countermeasures like towed decoys, but they differ in that expendable decoys free-fall or glide to the ground as opposed to being towed behind the aircraft.

a. Expendable decoys are small, active jamming systems designed to be expended by existing aircraft chaff and flare dispensers, such as the AN/ALE-40 or the AN/ALE-47. Expendable decoys can employ noise or deception jamming with noise jamming being the most common. Deception jamming techniques can be employed to enhance effectiveness against pulse Doppler radars. There are two challenges associated with expendable jammers: the amount of the time the jammer is effective and the packaging (Figure 12-5).

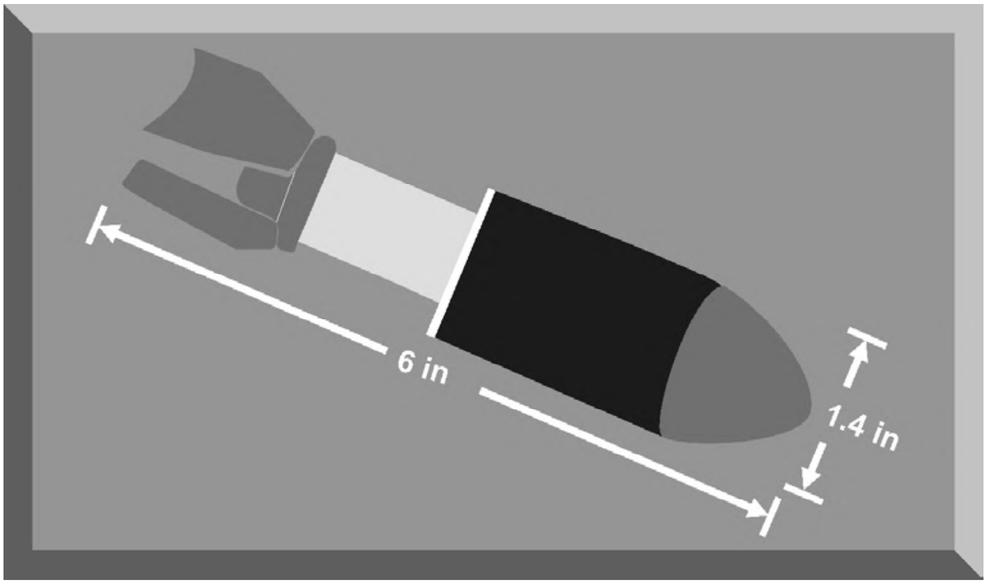

Figure 12-5. Generic Expendable

b. Expendable decoys are designed to provide protection for the dispensing aircraft for a specific period. The dispensing altitude and rate of fall determine this period of effective coverage. Expendable decoys can employ small parachutes of aerodynamic design to slow the rate of fall and increase the time of effective coverage. If the period of coverage is too short, multiple expendable decoys must be employed. This places a premium on timely employment and expendables management.

c. The primary components of an expendable decoy are the transmit and receive antennas, techniques generator, amplifier, and power supply. The transmit and receive antennas should be isolated and capable of high gain, wide bandwidth, and should use compatible polarization with the victim radar. The techniques generator must recognize the victim radar signal and generate the appropriate jamming response. The amplifier must be capable of generating a high power jamming signal over a wide frequency range. To meet these requirements, sophisticated computer and miniaturization techniques are used, and the components packaged to all fit in the aircraft dispenser. These factors impact the cost of expendable decoys and may limit the availability of these assets.

d. The Generic Expendable, RTE-1489, commonly called the GEN-X decoy is a fielded expendable active decoy. The decoy is sized to fit into a 1.4 x 1.4 x 5.8 inch cartridge and take advantage of new microwave/millimeter-wave integrated circuit (MMIC) technology. The GEN-X is programmable and features a broadband antenna and wide frequency coverage. After ejection, the decoy extends three small fins for stability. Its battery ignites to provide power, the receiver locks on to the threat radar signal, and a deception signal is generated and transmitted.

5. SUMMARY

Decoys simply provide the enemy with more targets to process. In the case of saturation decoys, this forces the enemy to commit resources against false targets, or show his defenses. For towed decoys and expendable active decoys, it makes the missile or tracking radar separate a real target from more electronically attractive decoys.

CHAPTER 13. CHAFF EMPLOYMENT

1. INTRODUCTION

Chaff was first used during World War II when the Royal Air Force, under the code name "WINDOW," dropped bales of metallic foil during a night bombing raid in July 1943 (Figure 13-1). The bales of foil were thrown from each bomber as it approached the target. The disruption of German AAA fire control and ground control intercept (GCI) radars rendered these systems almost totally ineffective. Based on this early success, chaff employment became a standard bomber tactic for the rest of the war.

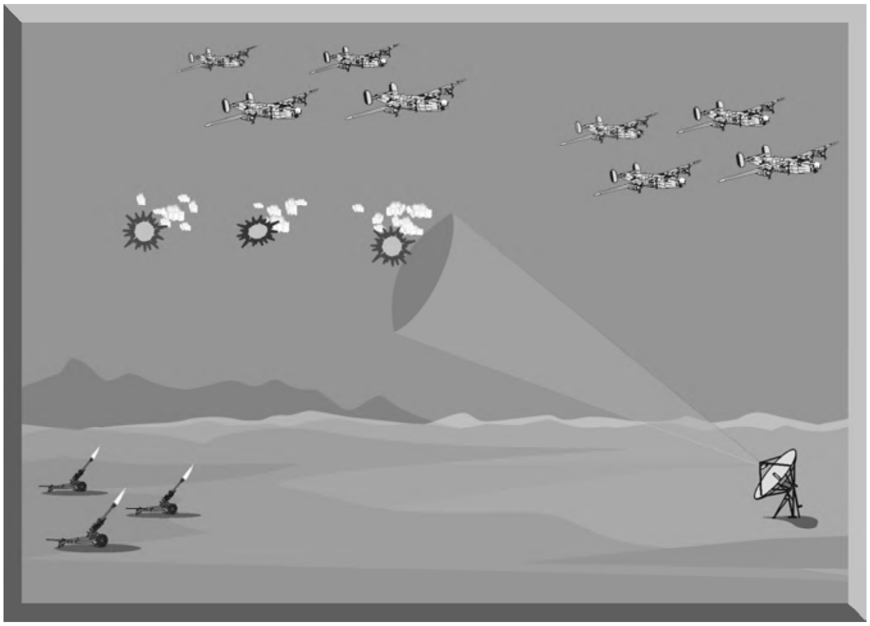

Figure 13-1. "WINDOW" – The First Operational Employment of Chaff

a. Chaff is one of the most widely used and effective expendable electronic attack (EA) devices. It is a form of volumetric radar clutter consisting of multiple metalized radar reflectors designed to interfere with and confuse radar operation. It is dispensed into the atmosphere to deny radar acquisition, generate false targets, and to deny or disrupt radar tracking. Chaff is designed to be dispensed from an aircraft and function for a limited period.

b. Even with the development and deployment of advanced radar threat systems, chaff continues to be an extremely effective EA device. Experience gained during the Vietnam conflict, the 1973 Yom Kippur War, and DESERT STORM clearly shows that chaff effectiveness against radar threats is still a factor with which the enemy must contend. This is especially true when chaff is employed with self-protection jamming and aircraft maneuvers.

c. Chaff screening and self-protection are the two basic chaff employment tactics. Chaff screening tactics, including area saturation and chaff corridor employment, are designed to confuse and deny acquisition information to the early warning, GCI, and acquisition radars supporting surface-to-air missile (SAM) systems. Self-protection tactics are designed to counter acquisition and target tracking radars (TTRs). When used with jamming and maneuvers, chaff can cause TTRs to break lock or generate survivable miss distances if a SAM is fired at the aircraft.

2. CHAFF CHARACTERISTICS

To understand how chaff affects radar systems, it is important to understand its characteristics. The most important chaff characteristics are radar cross section (RCS), frequency coverage, bloom rate, Doppler content, polarization, and persistence.

a. RCS is a measure of the net reradiated energy from a target to the illuminating radar. The RCS of an aircraft varies based on the size, shape, type of skin surface, configuration, and aspect to the illuminating radar. Figure 13-2 shows the effect of aspect on aircraft RCS. The RCS is greatest when the aircraft aspect is 90°, or abeam the radar. The lowest RCS occurs near the 30-70° and 110-150° of aspect. Since the aircraft RCS also varies based on frequency, the victim radar's frequency is a key factor. To be effective, chaff must be dispensed in large enough quantities to create an RCS greater than the aircraft RCS.

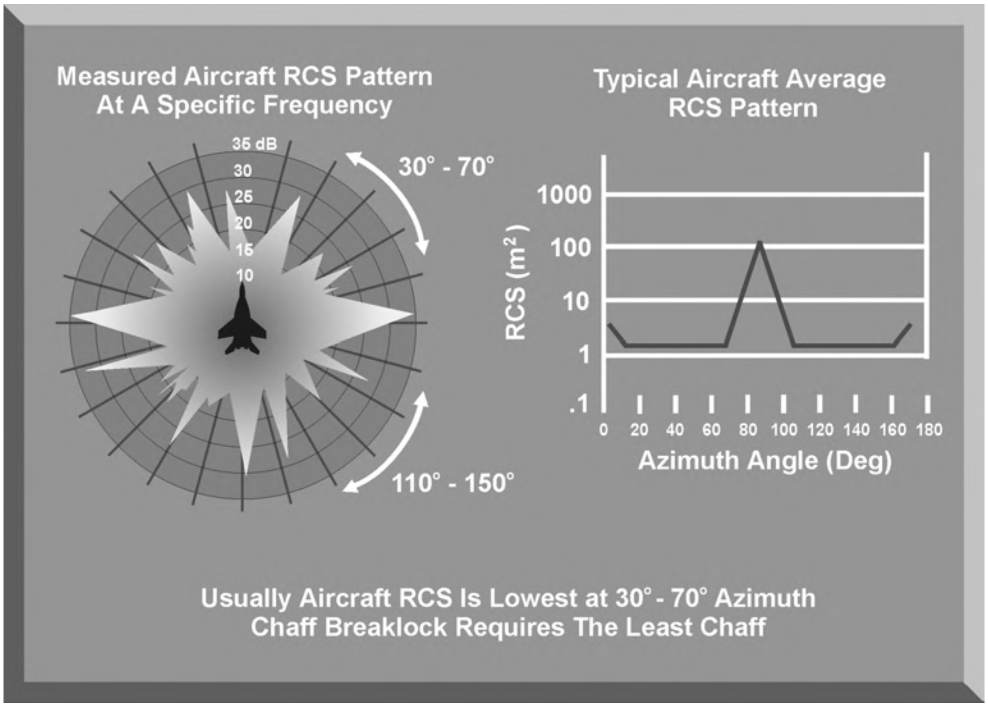

Figure 13-2. Aircraft Radar Cross Section (RCS)

(1) The RCS of a chaff bundle depends on the frequency of the victim radar and the dispensing aircraft's relative position, or aspect. Figure 13-3 shows the RCS of a single RR-170 chaff cartridge based on frequency. It shows that the largest RCS occurs at about 3 GHz. However, for the spectrum between 2-18 GHz, which includes most SAM TTRs, the RCS of the RR-170 cartridge is over 50 square meters. Since the typical fighter aircraft RCS varies between 1 and 10 square meters, depending upon frequency and aspect, the RR-170 chaff cartridge should provide a sufficient RCS to mask the aircraft RCS.

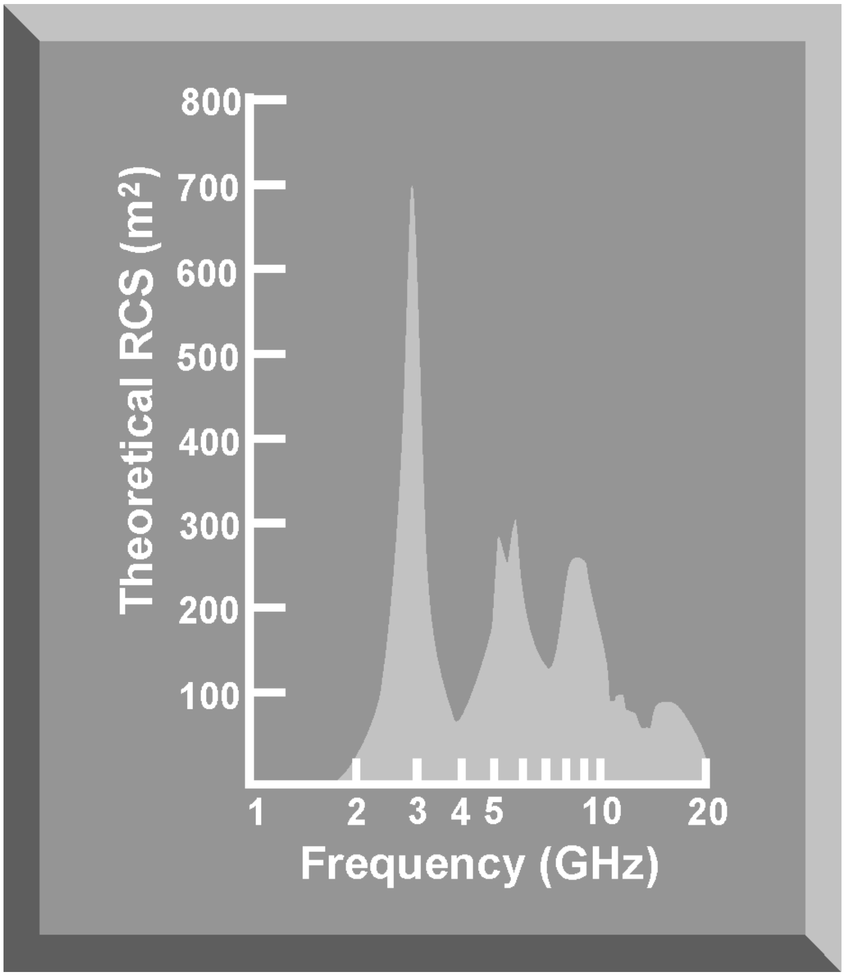

Figure 13-3. RR-170 Chaff Cartridge RCS

(2) The angular relationship, or aspect, between the aircraft and chaff bundle affects the chaff RCS presented to the victim radar. Chaff RCS is greatest when the chaff bundle and the aircraft are abeam the threat radar. It is smallest when the threat radar is off the nose or tail of the aircraft. Aspect is important when developing self-protection maneuvering and chaff dispensing tactics against threat radars (Figure 13-4).

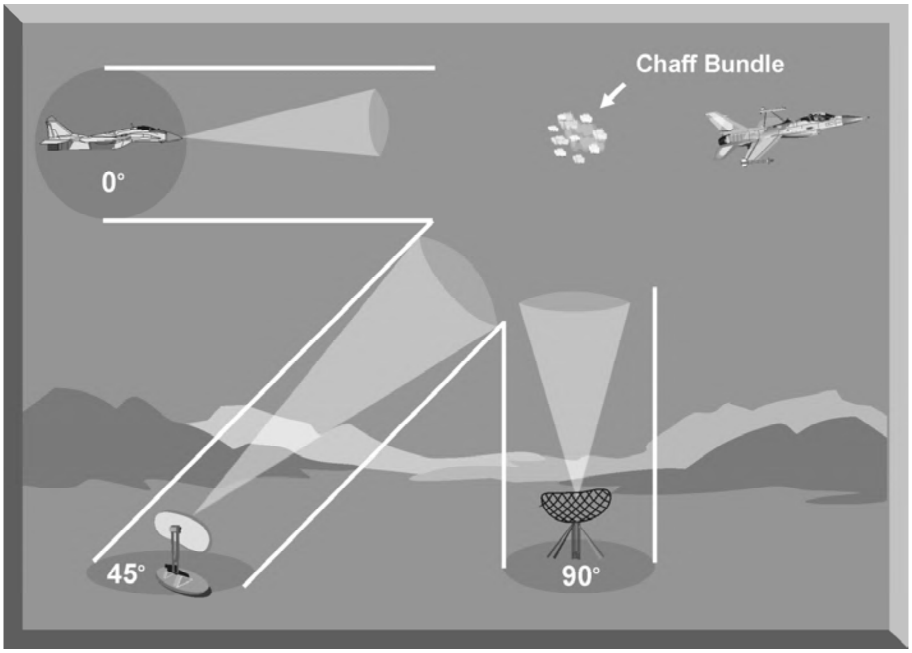
Chaff Bundle
0°
45°
90°

Figure 13-4. Threat Radar Aspect and Chaff RCS

(3) Dispensing multiple chaff bundles simultaneously does not necessarily increase chaff RCS. Multiple bundles increase the density of the chaff but do not directly enhance self-protection capabilities (Figure 13-5). This is an important consideration when developing chaff dispenser rates to counter threats.

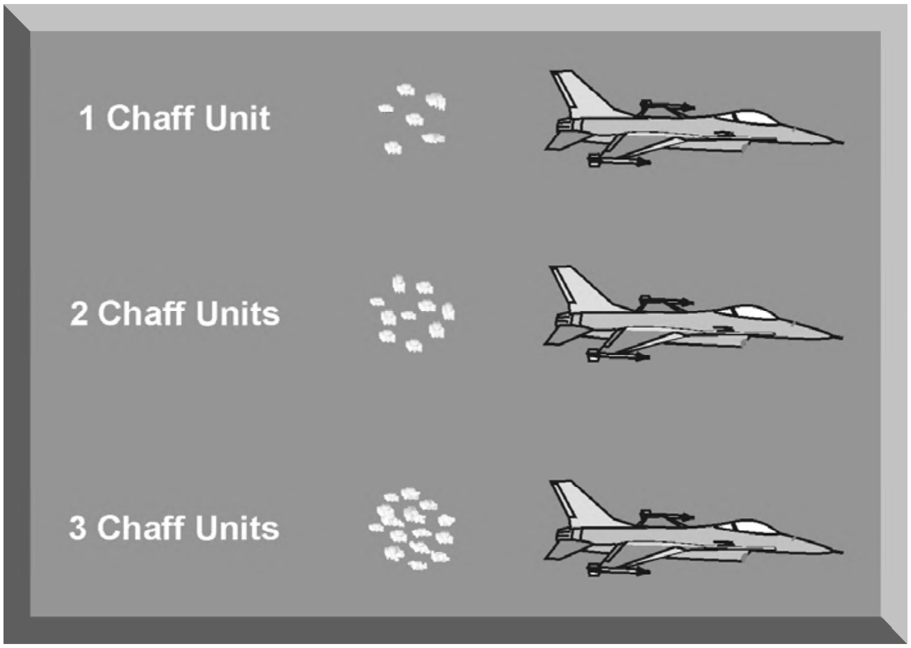
1 Chaff Unit
2 Chaff Units
3 Chaff Units

Figure 13-5. Impact of Multiple Chaff Cartridge Employment on Chaff RCS

b. Each strip of chaff is a dipole reflector that reradiates the electromagnetic energy received from an emitting radar and creates a radar echo. The optimum size is cut to about one-half the wavelength of the victim radar's RF. Since a single cut length is restricted in effectiveness to a narrow range of frequencies, different lengths are normally packaged together to provide coverage over a wide range of frequencies (Figure 13-6).

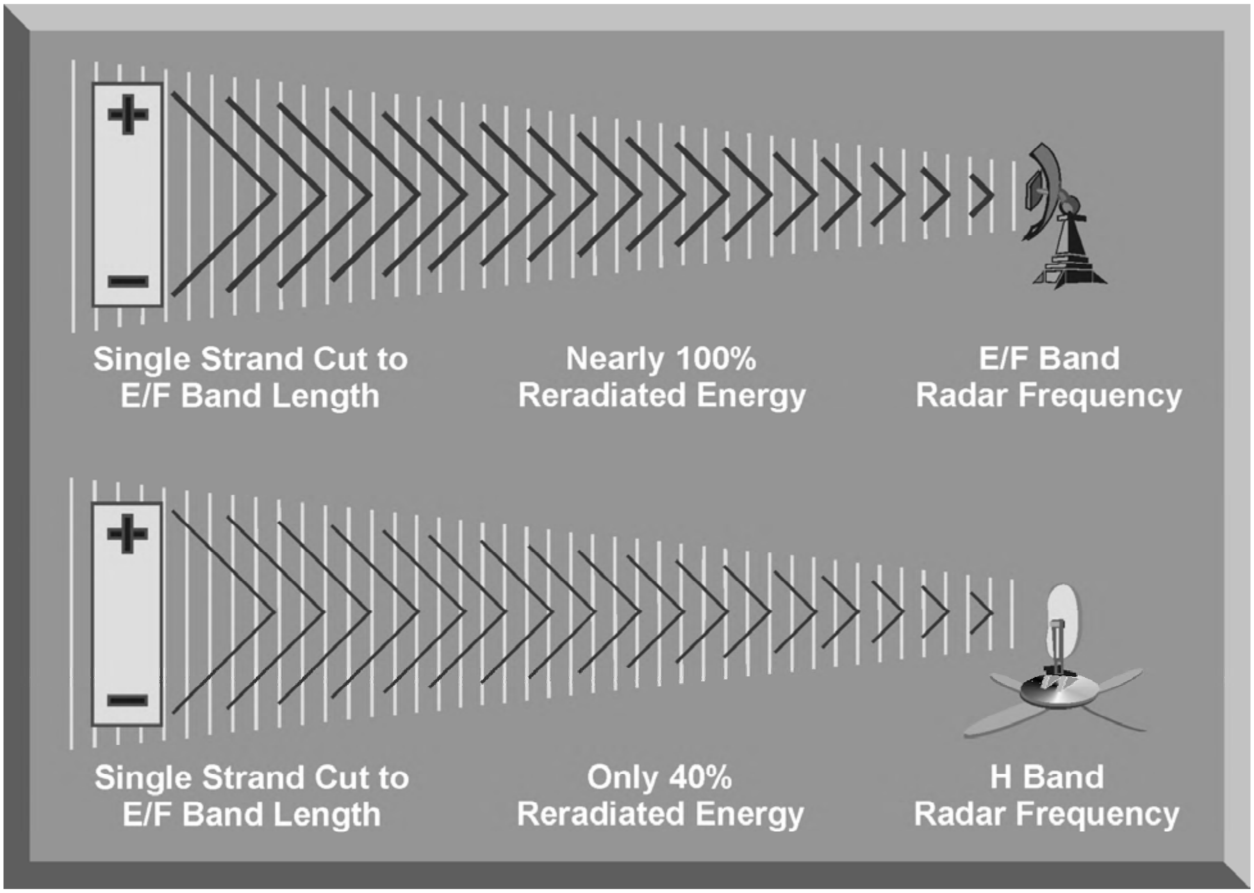
+
−
Single Strand Cut to E/F Band Length
Nearly 100% Reradiated Energy
E/F Band Radar Frequency
+
−
Single Strand Cut to E/F Band Length
Only 40% Reradiated Energy
H Band Radar Frequency

Figure 13-6. Chaff Length and Frequency Coverage

(1) Considerable research and development has reduced the size and increased the effectiveness of self-protection chaff. There are various chaff sizes, shapes, and materials. Most chaff carried on fighter aircraft are made of small aluminum strips, coated strips of nylon, or fiberglass. These strips are cut to various lengths and compressed into bundles that are small and light enough to allow the aircraft to carry and dispense multiple chaff bundles. These cuts of chaff are packaged into chaff cartridges and inserted into a dispenser on the aircraft. An explosive squib assembly ejects the cartridges from the dispenser and disperses the chaff (Figure 13-7).

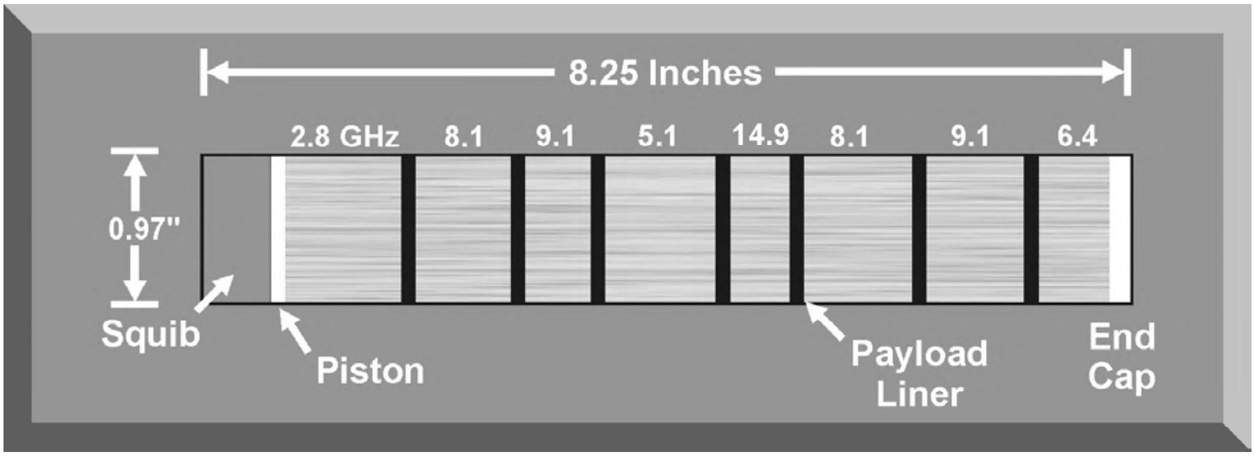

Figure 13-7. RR-170 Chaff Cartridge

(2) To provide as many dipoles as possible and present the maximum radar cross section, each chaff bundle has numerous chaff cuts to match a predetermined range of frequencies. Each chaff cartridge contains almost 3 million dipoles packaged in an eight inch by one-inch cartridge. The dipole frequencies cover the frequency range where most SAM TTRs and air-to-air radars operate (2 - 18 GHZ).

c. Bloom rate, the rate at which chaff will scatter, is also a very important characteristic of self-protection chaff. Self-protection chaff effectiveness is based on the relationship of bloom rate, chaff RCS, aircraft RCS, and the resolution cell of the threat radar system. The ability of chaff to effectively defeat a target tracking radar is directly related to the chaff dispense rate, which determines the chaff RCS, which should be larger than the aircraft's RCS. The chaff bundles must also bloom within the resolution cell of the radar.

(1) Chaff bloom rate is dependent on aerodynamic factors associated with the chaff type, the location of the dispenser on the aircraft, and the aircraft wake or turbulence. Heavy or dense chaff falls faster and blooms slower than lighter and less dense chaff. The location of the chaff dispenser on the aircraft affects the airflow in which the chaff will be dispensed. The ideal position for the dispenser is in the area where there is the most turbulence from the aircraft. Turbulence behind the aircraft is probably the most important factor affecting bloom rate. The more turbulent the airflow, the greater the bloom rate (Figure 13-8). Maneuvering the aircraft while dispensing chaff also enhances the chaff bloom rate.

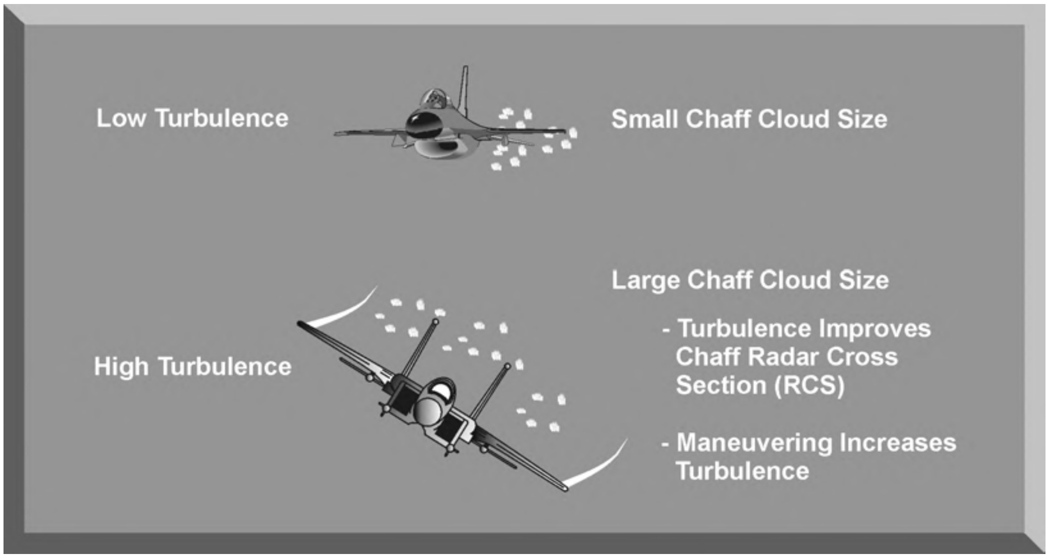
Low Turbulence
Small Chaff Cloud Size
Large Chaff Cloud Size
- Turbulence Improves Chaff Radar Cross Section (RCS)
High Turbulence
- Maneuvering Increases Turbulence

Figure 13-8. Impact of Turbulence and Chaff Bloom Rate

(2) To ensure that the victim radar is decoyed or that it transfers automatic tracking to the chaff, the chaff must bloom within the radar resolution cell. This resolution cell is a three-dimensional spheroid with dimensions based on the pulse width, horizontal beamwidth, vertical beamwidth, and the range of the aircraft (Figure 13-9). There are some rules of thumb that can be used when considering the bloom rate of chaff and the resolution cell of a particular radar. The shorter the pulse width of a radar, the faster the chaff has to bloom to be effective. The narrower the horizontal and vertical beamwidths, the faster the chaff has to bloom to be effective.

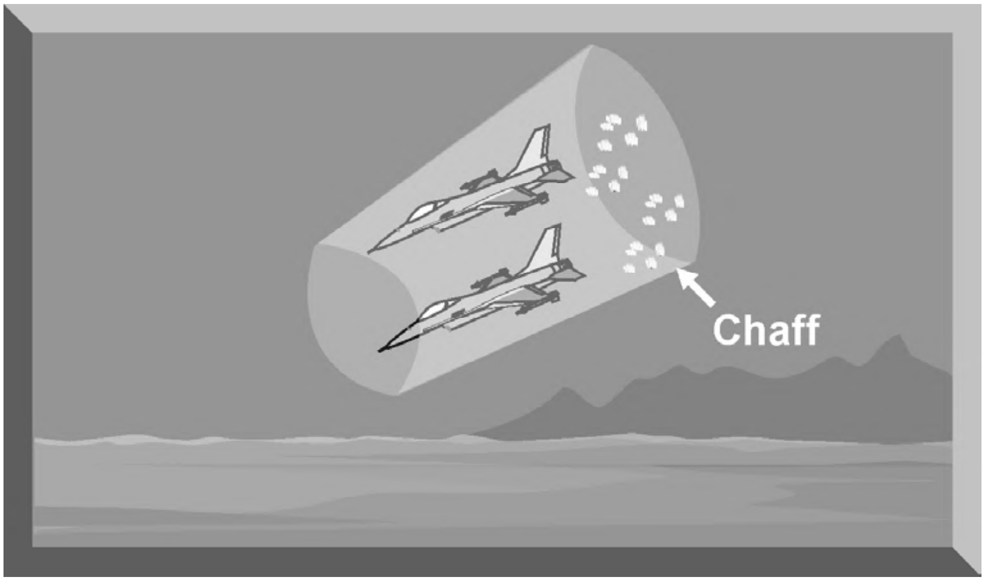

Figure 13-9. Chaff Bloom Rate and Radar Resolution Cell

d. Against Doppler radars, self-protection chaff is most effective when dispensed at or near the beam, relative to the threat radar. When chaff is dispensed in the airstream, the drag on an individual dipole is so great compared to its mass that it slows to the velocity of the surrounding air mass almost instantly. Since the relative velocity of the chaff, in relation to the radar, is zero, radar systems employing Doppler processing and tracking will not display the chaff. Doppler processing radars will continue to track the aircraft unless it also has a relative velocity of zero. This occurs when the aircraft is abeam the radar. Chaff corridor and area saturation tactics against Doppler tracking radars will have limited effectiveness.

e. Chaff persistence and polarization are two additional characteristics that are important employment considerations for area saturation or chaff corridor operations. These individual chaff element characteristics are directly related to the combined effects of aerodynamic, atmospheric, and gravitational influences.

(1) Chaff persistence is the length of time the chaff is at an effective altitude to screen ingressing aircraft during area saturation or chaff corridor operations. The time span depends on the fall rate of the chaff and varies according to the density of the dipoles. The prevailing atmospheric conditions, such as wind and temperature also affect chaff persistence. Generally, the longer cuts used for lower frequency radars fall faster than the shorter cuts used for higher frequency radars. Each type has its own rate of fall based on these conditions. The rate of fall is a critical mission planning consideration for determining the amount of time between chaff corridor or area saturation initiation and the arrival of the aircraft being screened. If the chaff is employed too early, it may not be at the correct altitude or may have dispersed to the point that it is not effective to screen ingressing aircraft.

(2) Each chaff strand is a polarized dipole with positive and negative ends. The orientation of these strands determine their polarity (Figure 13-10). Chaff cuts with the positive and negative ends oriented vertically are vertically polarized. Chaff cuts with the positive and negative ends oriented horizontally are horizontally polarized. Since chaff strands are initially buffeted by turbulence and airstream vortices, the dipole orientation and polarization, changes rapidly and randomly. Eventually, the strands separate into two groups; one descending horizontally, and one descending vertically. Since the vertically oriented strands tend to fall faster, the lower part of the chaff cloud tends to become more vertically polarized, while the upper portion is horizontally polarized. A threat radar that uses vertical polarization will receive minimal affects from the upper (horizontally polarized) portion of the chaff cloud. If the aircraft being screened are flying within this portion of the chaff cloud, they may be detected and engaged. This is another mission planning consideration for chaff area saturation or chaff corridor operations.

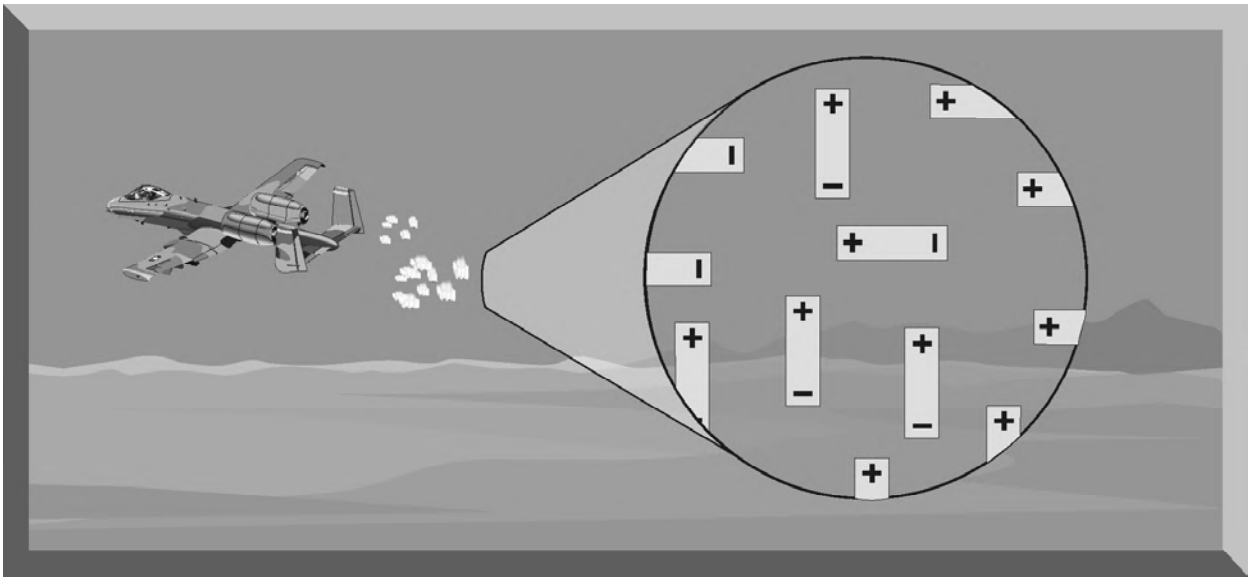

Figure 13-10. Impact of Chaff Polarization

3. CHAFF OPERATIONAL EMPLOYMENT

The two primary chaff employment tactics are force screening and self-protection. Force screening tactics include area saturation and corridor operations. Self-protection tactics include the reactive employment of chaff to negate a potentially lethal engagement. Different chaff dispensing techniques are used for each employment tactic and are important planning considerations for all chaff employment tactics. This section will discuss area saturation, corridor operations, and self-protection chaff employment.

a. The objective of area saturation operations is to present multiple false targets in a specific area in order to saturate radar systems and confuse the enemy integrated air defense system (IADS). Area saturation can be accomplished by fighter aircraft or drones equipped with chaff pods employing random chaff dispensing techniques. The chaff dispenser is set to release random bursts of chaff along the ingress and egress route of the attack package. Chaff pods may be supplemented with chaff bombs containing special fuses that provide false targets at varying altitudes. Attack aircraft can also contribute to area saturation by randomly dispensing self-protection chaff as they ingress and egress. However, this tactic can deplete the number of chaff bundles an attack aircraft may need to defeat a potentially lethal radar system encountered at a later time in the mission.

(1) The chaff cuts must provide frequency coverage for the threat radar systems. Also the RCS of each chaff burst should be large enough to present a realistic target to the victim radars. Multiple false targets created by chaff area saturation may confuse threat system operators and encourage them to expend missiles on false chaff targets.

(2) Saturation also masks the number of attacking aircraft (Figure 13-11). When used with false target deception jamming, area saturation can greatly enhance mission success. However, the technique is resource-intensive since aircraft employing chaff pods and chaff bombs cannot attack targets. These aircraft are vulnerable to attack and should be supported by standoff jamming. Area saturation tactics may have limited success against Doppler processing radars.

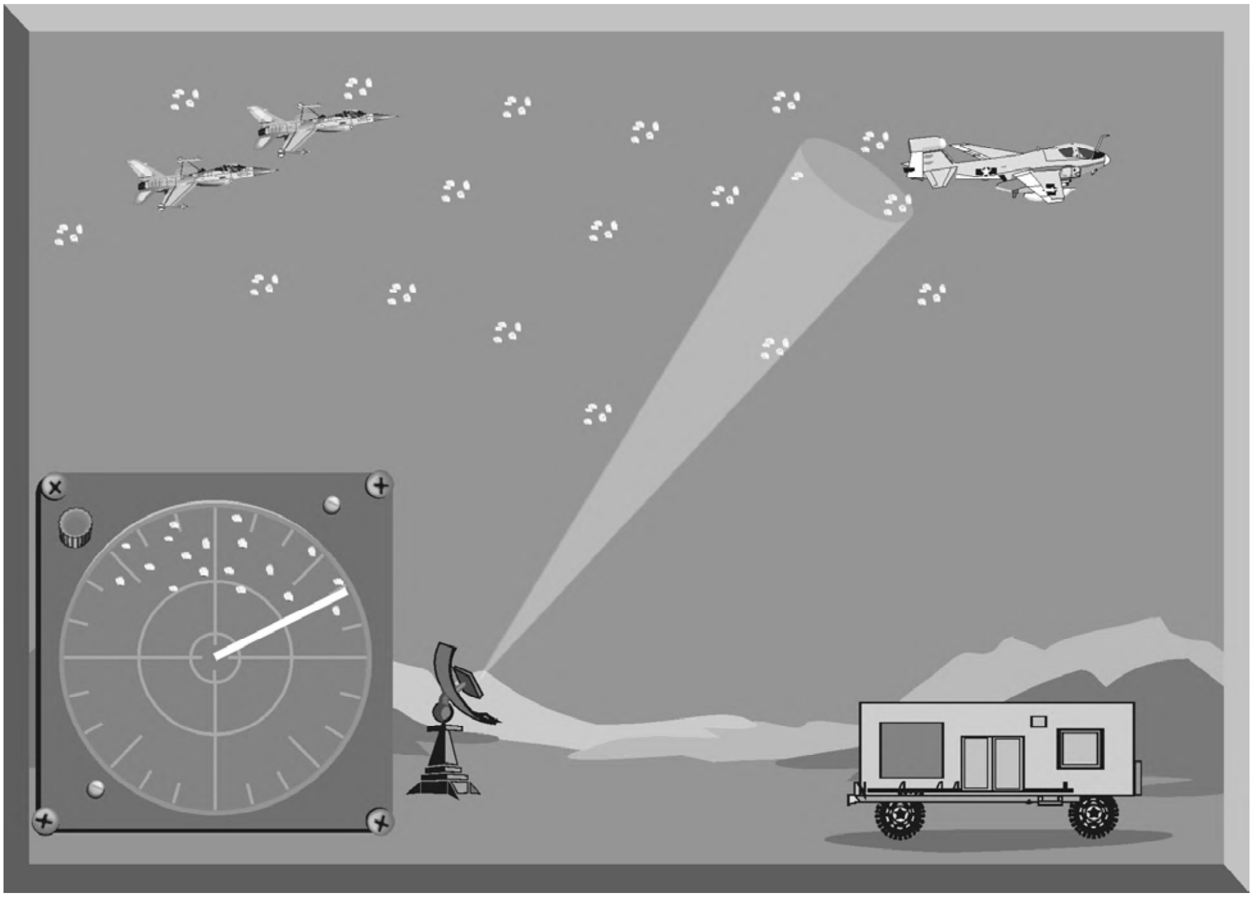

Figure 13-11. Area Saturation Tactics

b. The objective of chaff corridor operations is to screen the ingress and egress of an attack package by dispensing large quantities of chaff in a continuous "ribbon." Fighter aircraft, or drones equipped with chaff pods such as the ALE-38, employ a stream chaff dispensing technique to "lay" the chaff corridor. The pods are set to provide a continuous line of chaff dense enough to hide ingressing and egressing aircraft. The chaff cuts should provide frequency coverage for the radar systems that must be countered. Timing for the chaff aircraft in relation to the attack package must consider the fall rate and persistency of the chaff to ensure that the chaff corridor covers the required altitude for a time sufficient to allow the attack package to ingress and egress. An effective chaff corridor completely denies a radar's ability to distinguish between

the chaff and the attack aircraft. To do this, the radar cross section, or RCS, of the chaff within the resolution cell of the radar must exceed the RCS of the aircraft. This condition must be met throughout the length of the chaff corridor. When this condition is met, the chaff corridor will appear as a continuous return on the victim radar scope, and the attack package cannot be detected (Figure 13-12).

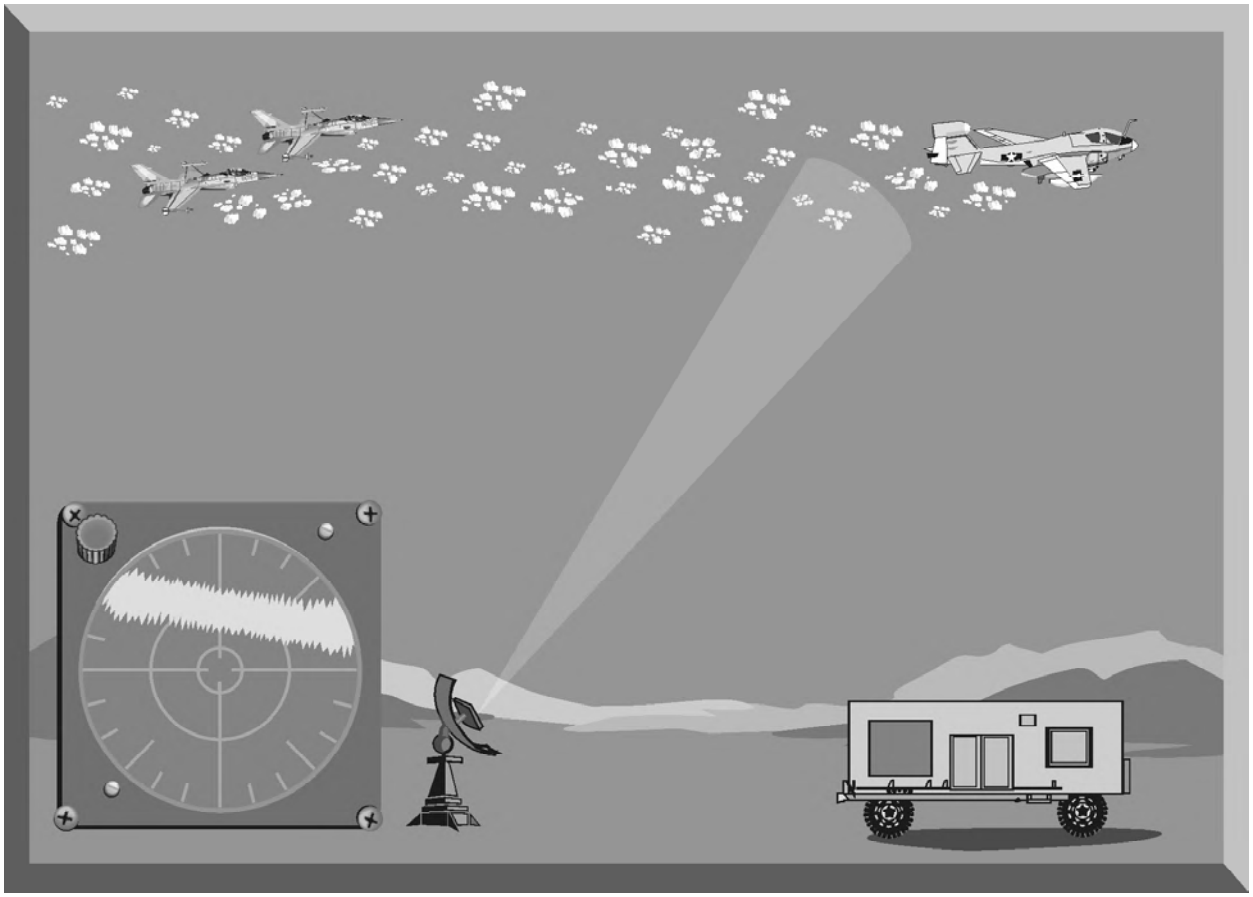

Figure 13-12. Chaff Corridor Tactics

(1) One advantage of a chaff corridor is that it can screen ingressing and egressing aircraft from pulse radar systems. However, chaff corridors are resource-intensive. Aircraft "laying" the corridor cannot strike critical targets. The chaff aircraft are also vulnerable to attack. Therefore, standoff jamming and self-protection jamming systems should be employed to provide some screening and protection for the chaff dispensing aircraft. Finally, chaff corridors may not be effective against radars with Doppler processing.

(2) To be effective, chaff corridor operations require detailed planning. Electronic combat (EC) planners must first determine that a chaff corridor is the most effective way to screen the attack force. This decision is based on the vulnerability of the attack aircraft to the anticipated threat radar systems and the availability of chaff assets. Once the decision is made to employ a chaff corridor,

planners must select the location, determine the length of the chaff corridor, select the ingress and egress altitudes, and establish the timing for the chaff aircraft and the attack package. Once the location of the chaff corridor is determined, planners must assess the threat radar systems that must be countered. The specific operating frequencies of the threat radars will determine the cuts of chaff that must be dispensed. The resolution cells of the threat radars will determine the density of chaff required. The length of the chaff corridor and the chaff density will determine the number of chaff aircraft required to seed the chaff corridor. The chaff fall rate and the atmospheric conditions impact the timing between the chaff aircraft and the attack package, and the altitude that the chaff dispensing aircraft must fly.

c. Self-protection chaff tactics are based on the use of chaff dispensers that use burst chaff dispensing techniques to defeat a TTR. Burst chaff dispensing, employed during the final phase of an engagement by air-to-air or surface-to-air weapons, can generate tracking errors or a radar break-lock. Burst chaff effectiveness is greatly enhanced when accompanied by jamming and evasive maneuvers (Figure 13-13).

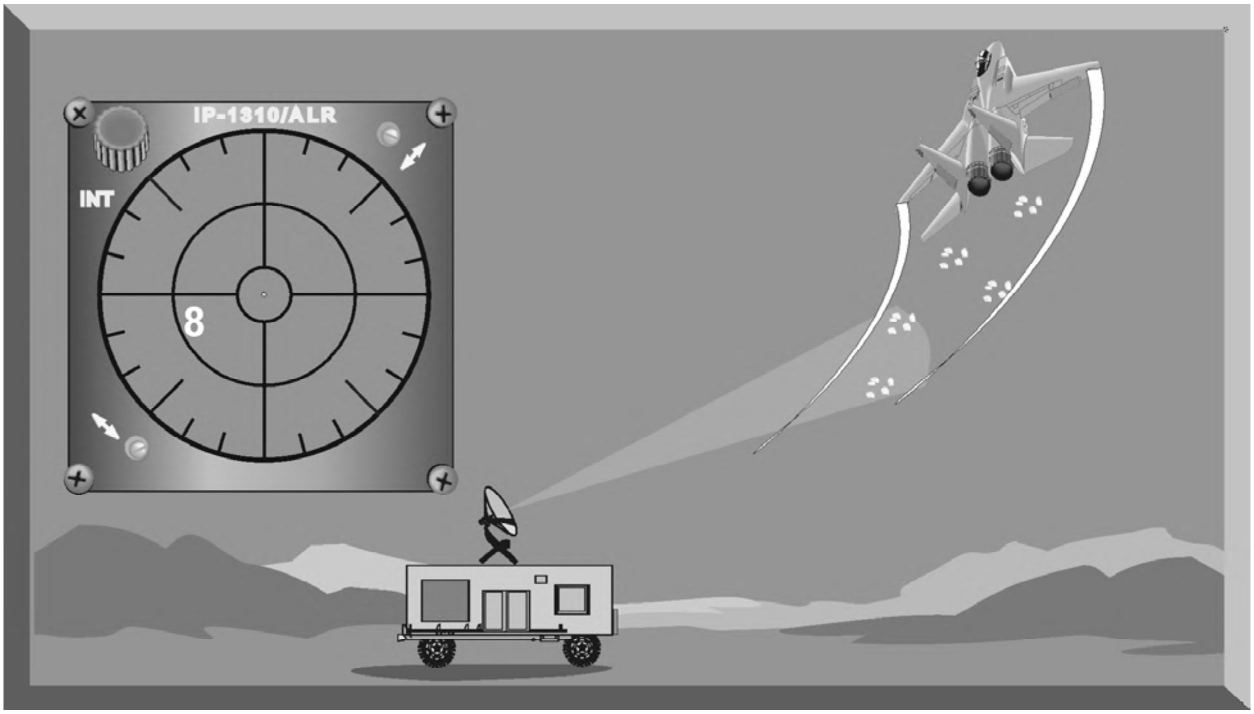

Figure 13-13. Self-Protection Chaff Tactics

(1) Self-protection chaff has proven effective against all pulse radar threat systems when employed with maneuvers and jamming. This is especially true for TTRs operating in an automatic tracking mode.

(a) Chaff employed against a track-while-scan (TWS) radar is designed to put multiple targets, with an RCS greater than the aircraft, in the resolution cell of the horizontal and vertical radar beams (Figure 13-14). Since the tracking loop tracks the largest return, the TWS radar will automatically switch to the chaff. After dispensing chaff, the pilot can maneuver vertically or horizontally to move the aircraft out of the resolution cell.

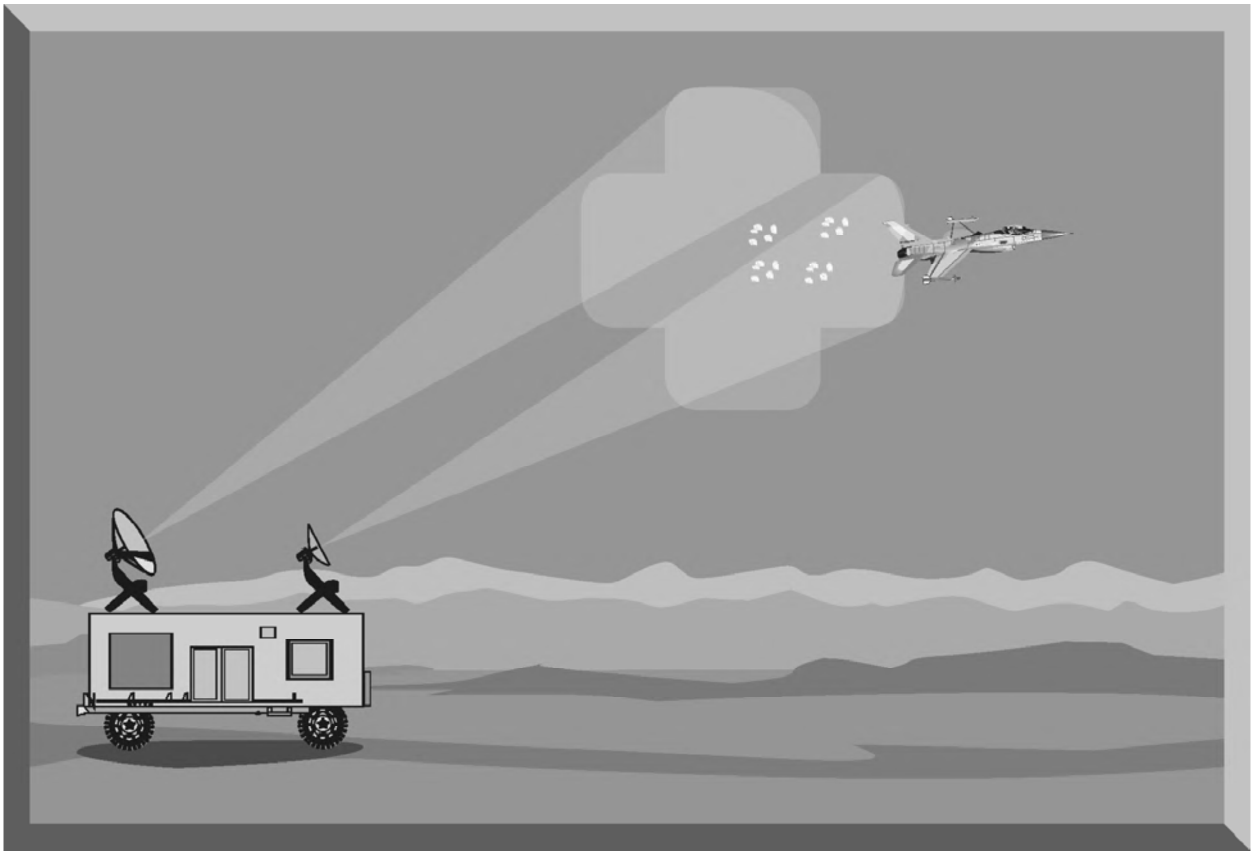

Figure 13-14. Self-Protection Chaff Effect on a TWS Radar

(b) Against a conical scan radar, chaff puts multiple, large RCS targets within the separate scans of the radar (Figure 13-15). These multiple targets generate error signals in the tracking loop and drive the separate scans off the aircraft return. As the conical scan radar tracking loop attempts to resolve these error signals, it will eventually lock on to the chaff. Maneuvering outside the overlapping scan area enhances chaff effectiveness and facilitates the transfer of radar lock-on to the chaff.

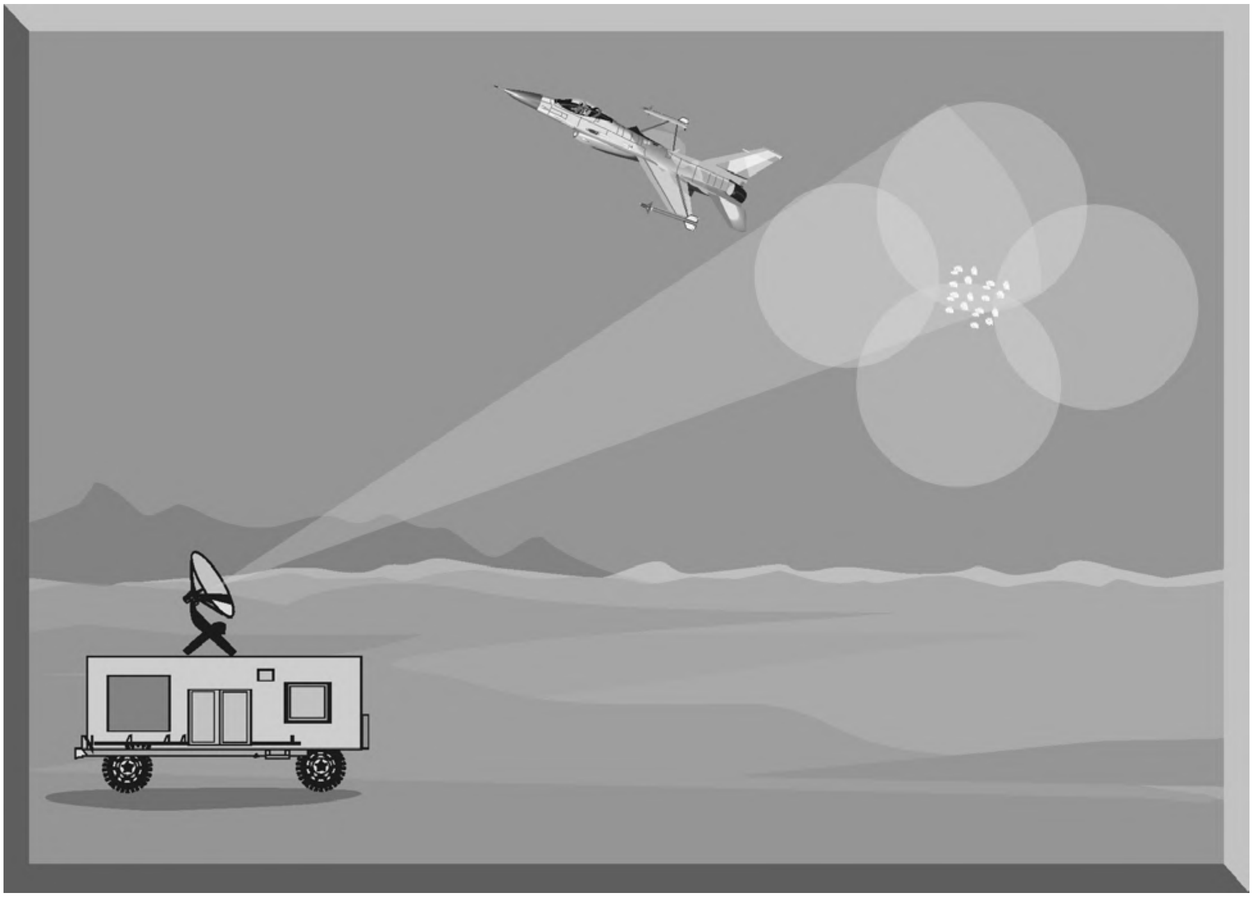

Figure 13-15. Self-Protection Chaff Impact on a Conical Scan Radar

(c) Chaff employed against a monopulse radar is designed to put multiple targets in at least two of the tracking beams (Figure 13-16). This generates errors in the azimuth, elevation, and range tracking circuits. Multiple chaff targets continue to generate azimuth and elevation errors that can eventually generate a break-lock condition, as the radar transfers lock-on to the chaff. Chaff is most effective against monopulse radars when employed on the beam in order to create the maximum angular tracking error.

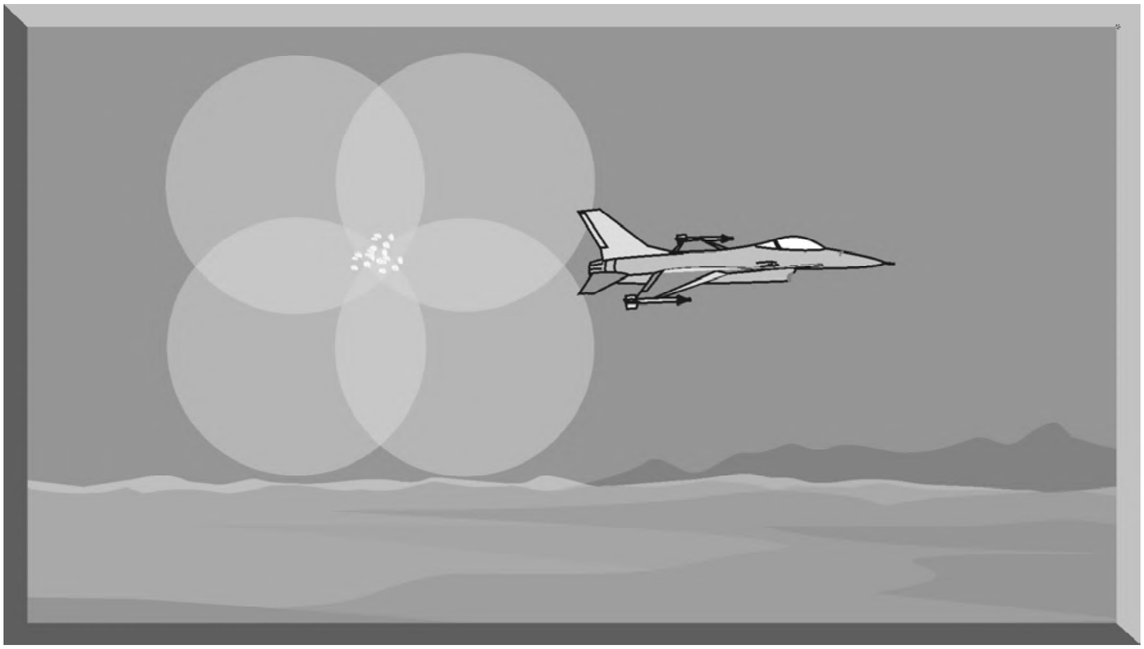

Figure 13-16. Self-Protection Chaff Impact on a Monopulse Radar

(d) Modern radars may employ some form of Doppler filtering to negate the effectiveness of chaff and other sources of clutter. Pulse Doppler and continuous wave radar systems track targets based on target velocity relative to the radar. Radars employing a moving target indicator (MTI) use relative target velocity to distinguish between targets and clutter. Chaff slows to near zero relative velocity almost immediately after dispensing. For self-protection chaff to be effective, the aircraft velocity relative to the radar site must also be near zero This occurs when the aircraft's aspect to the radar is 90°, or on the beam. By maneuvering to a beam aspect against a Doppler radar, the pilot is exploiting the “notch” where radar cannot discriminate targets based on Doppler frequency shift (Figure 13-17).

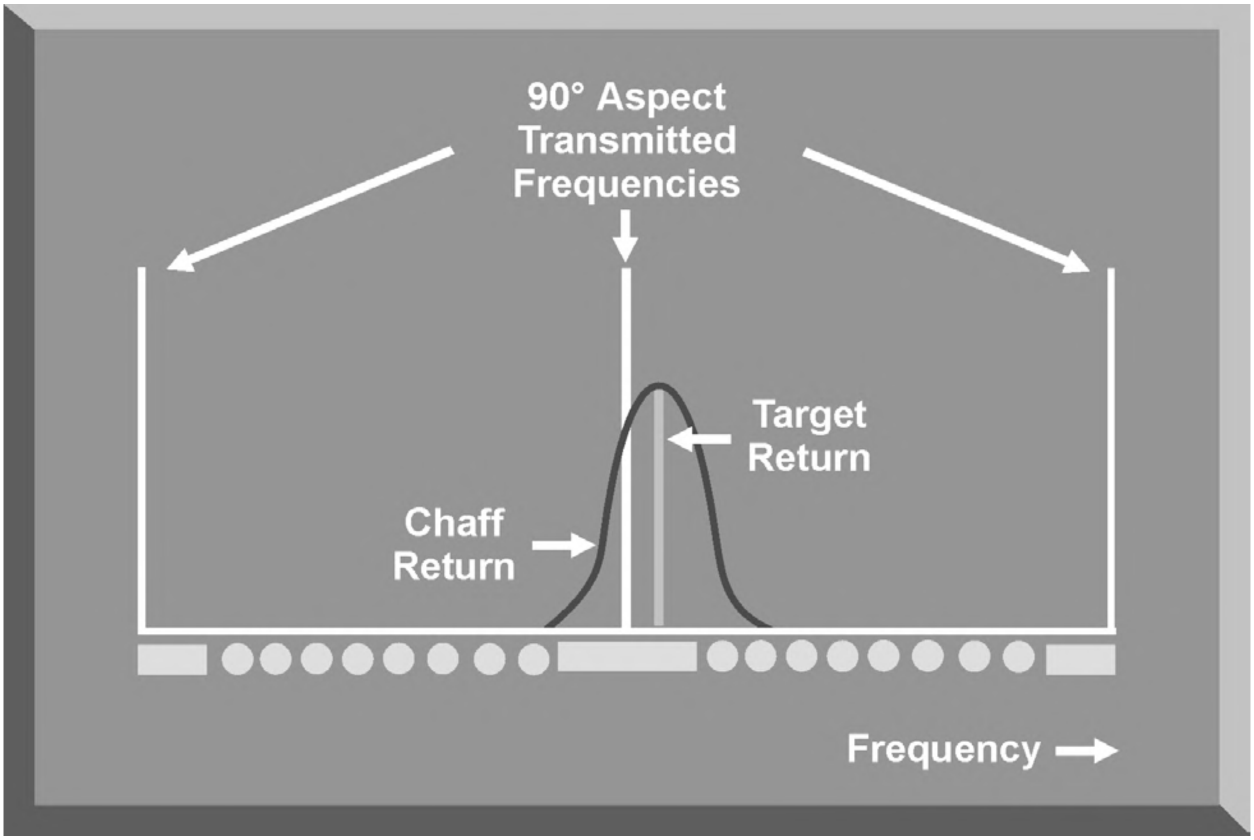
90° Aspect
Transmitted
Frequencies
Target
Return
Chaff
Return
Frequency →

Figure 13-17. Self-Protection Chaff Impact on a Doppler Radar

4. SUMMARY

Chaff is one of the oldest and most effective pulse radar countermeasures. The fundamental characteristics of chaff (RCS, frequency coverage, bloom rate, Doppler content, polarization, and persistence) determine the effectiveness of chaff employment. The primary chaff employment tactics of force screening and self-protection are designed to maximize the impact of chaff on threat radar systems. Self-protection chaff, together with jamming and maneuvers, is often the "last line of defense" against lethal radar threat systems.

CHAPTER 14. IR FUNDAMENTALS

1. INTRODUCTION

Since their introduction in the 1950s, infrared (IR) missiles have been an increasing threat from both ground-based and airborne systems. The range, reliability, and effectiveness of IR missiles have been continuously improved by advanced detector materials and computer technology. Since IR missiles are passive, they are relatively simple and inexpensive to produce. These characteristics have contributed to the proliferation of IR missiles in the combat arena. Nearly every aircraft flying in either the air-to-air or air-to-surface role now carries an all-aspect IR missile. Additionally, every infantry unit down to the platoon level is equipped with shoulder-fired IR missiles (Figure 14-1). This chapter will cover basic IR theory, IR missile detection, IR seekers, and conclude with a section on IR flare rejection.

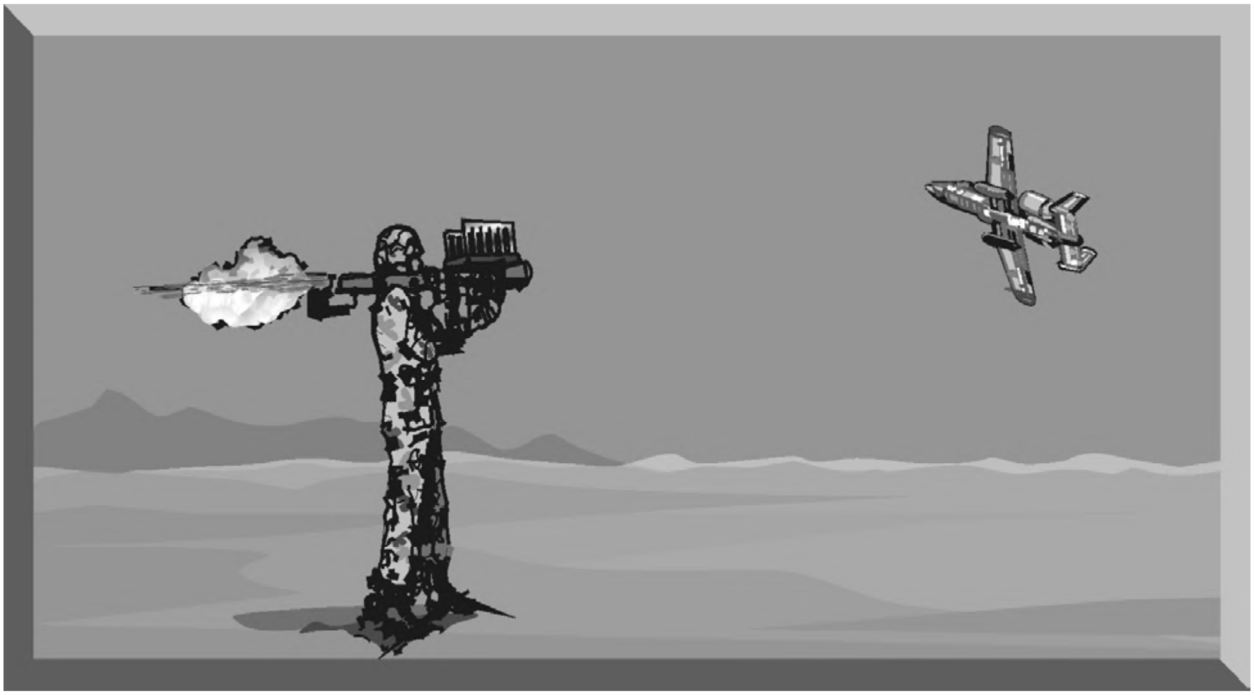

Figure 14-1. The IR Threat

2. BASIC IR THEORY

Because of its location in the frequency spectrum, IR radiation exhibits some of the characteristics and limitations of microwaves and visible light.

a. All warm objects emit IR energy. The object's temperature dictates the characteristics of this radiation. As the temperature of the material increases, the radiant intensity increases and shifts to shorter and shorter wavelengths or

higher and higher frequencies. The frequency band of IR radiation falls between the upper limit of microwaves and the lower limit of visible light. When discussing IR radiation, it is more convenient to refer to wavelength instead of frequency (Figure 14-2). The wavelength of the highest frequency IR is 0.72 x 10^{-6} meters. A unit of measure called the micron (µ), is one millionth, or 10^{-6}, of a meter and is used to designate IR wavelengths. IR energy falls in the electromagnetic spectrum between the wavelengths of 1000 and 0.72 microns, while visible light occupies the spectrum from 0.72 to 0.39 microns.

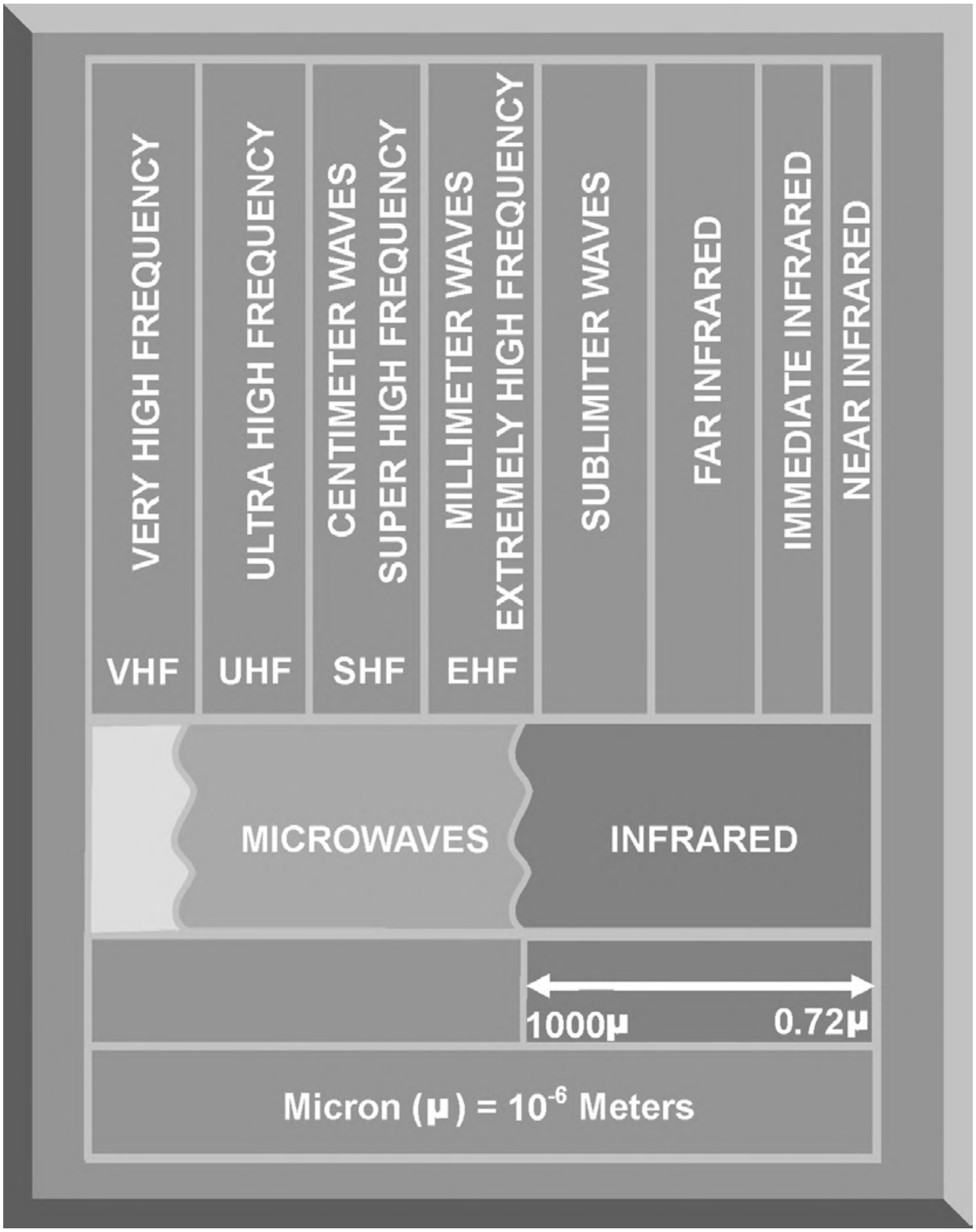
VERY HIGH FREQUENCY
VHF
ULTRA HIGH FREQUENCY
UHF
CENTIMETER WAVES
SUPER HIGH FREQUENCY
SHF
MILLIMETER WAVES
EXTREMELY HIGH FREQUENCY
EHF
SUBLIMITER WAVES
FAR INFRARED
IMMEDIATE INFRARED
NEAR INFRARED
MICROWAVES
INFRARED
1000µ
0.72µ
Micron (µ) = 10⁻⁶ Meters

Figure 14-2. IR Frequency Band

b. As the IR energy travels through the atmosphere, certain wavelengths are absorbed or attenuated. The greatest IR attenuator is atmospheric water vapor, which varies as the weather conditions vary, with negligible absorption at altitudes above 30,000 feet. Another significant attenuator is carbon dioxide. The percentage of carbon dioxide in the atmosphere is practically constant up to a height of 30 miles. Carbon dioxide absorption is predictable and occurs only in the IR region. Scattering is another form of atmospheric attenuation and is caused by dust particles and water droplets. Scattering is also largely dependent on weather conditions and cannot be predicted. Most of the scattering occurs at lower altitudes and at the shorter wavelengths. Other atmospheric elements cause little or no attenuation of IR energy. Figure 14-3 shows atmospheric IR transmission at sea level. There is a relatively large window of IR transmission in the region from one to five microns. This is the region where the aircraft engine heat signature is at its maximum intensity and where most IR missiles are designed to operate.

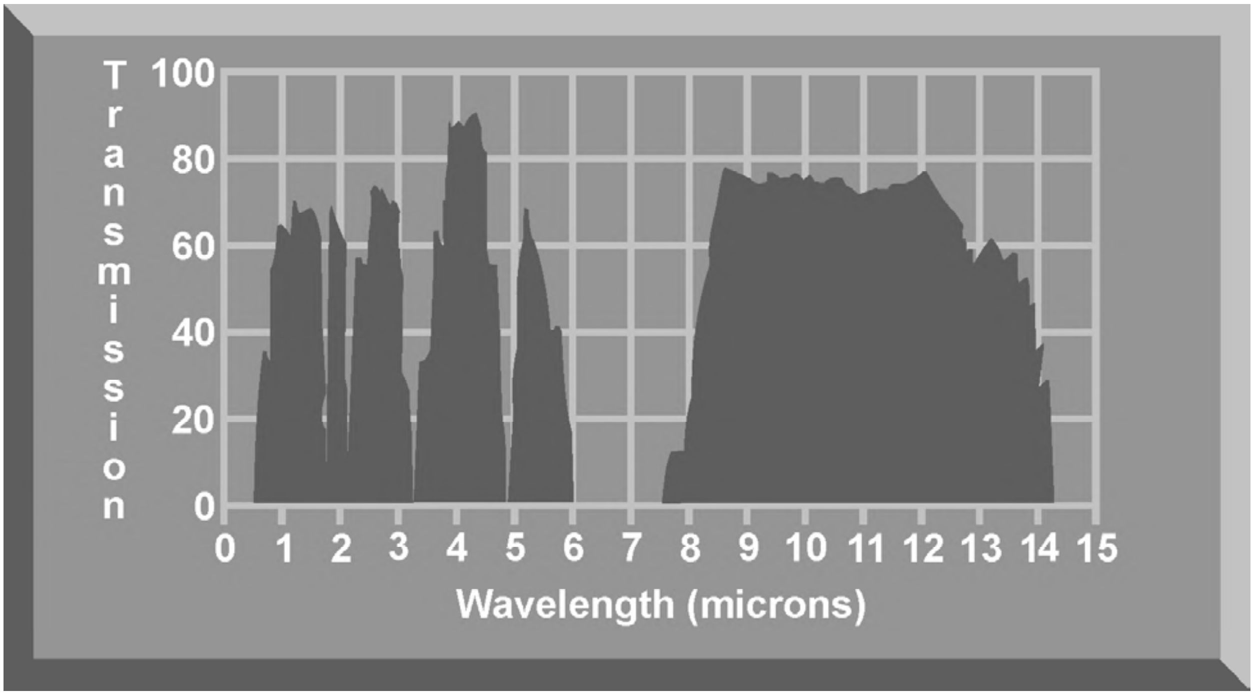

Figure 14-3. Atmospheric IR Transmission

c. IR energy exhibits some of the transmission characteristics of both RF energy and visible light. As with visible light, it can be optically focused by lenses and mirrors. This characteristic is used in the IR missile detector elements (Figure 14-4). As with RF energy, the intensity of IR radiation diminishes inversely with the square of the distance between the source and the receiver.

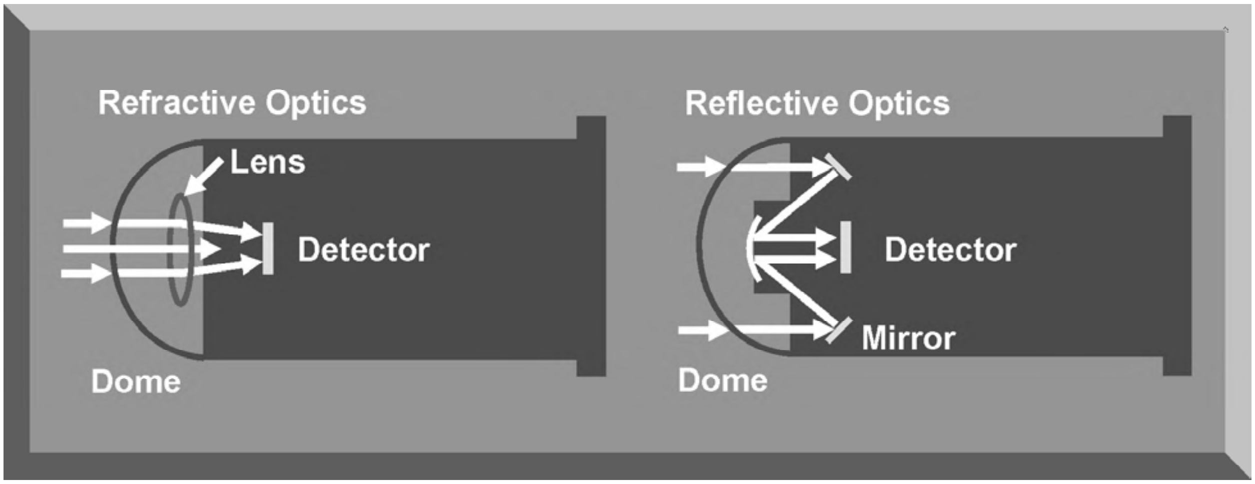
Refractive Optics
Lens
Detector
Dome
Reflective Optics
Detector
Mirror
Dome

Figure 14-4. IR Missile Seeker

d. If the intensity of the aircraft's IR signature, while in the military power setting, is plotted in relation to wavelength, it reaches a peak at approximately 3 microns (Figure 14-5). In afterburner, the aircraft's IR intensity reaches a peak at approximately 1.5 microns. Since IR missiles are designed to detect and track the aircraft's IR signature, most IR missiles operate in the region of 1 to 5 microns. To be effective, the IR intensity of a flare must also fall within this micron region.

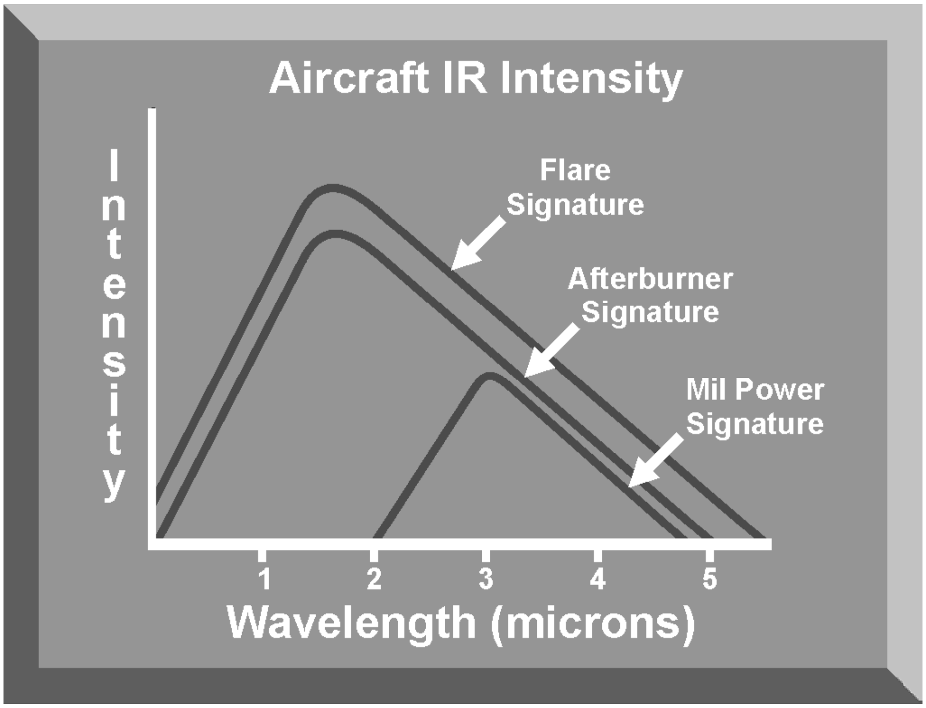

Figure 14-5. Aircraft IR Intensity

e. An aircraft's relative IR signature is based on aspect angle, airspeed, altitude, and afterburner status. The minimum relative IR intensity is at the nose and maximum at the tail. As airspeed increases, the relative IR intensity increases due to the heat generated by friction and increased engine temperatures. As altitude increases, the relative IR signature increases at all aspect angles, due to the reduction in atmospheric attenuation (Figure 14-6).

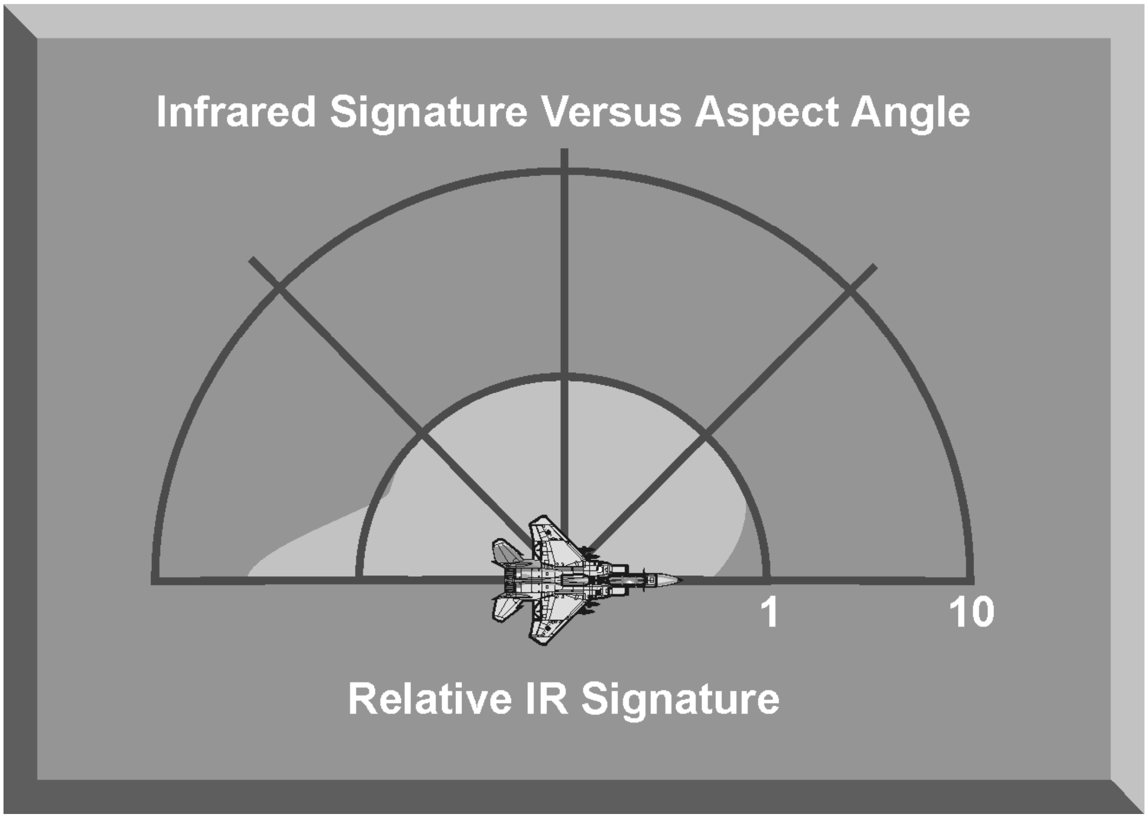

Figure 14-6. IR Signature Versus Aspect Angle

3. IR SIGNATURE SOURCES

IR guidance is based on the fact that every object with a temperature above zero degrees Kelvin emits IR radiation. The temperature of the object dictates the characteristics of this radiation. As the temperature of the object increases, the radiant intensity increases and shifts to higher and higher frequencies and correspondingly shorter and shorter wavelengths. The F-16 at Figure 14-7 demonstrates the different wavelengths found at different areas of the aircraft. Emissivity in Figure 14-7 is a relative measure of the IR energy emitted when a surface is directly viewed.

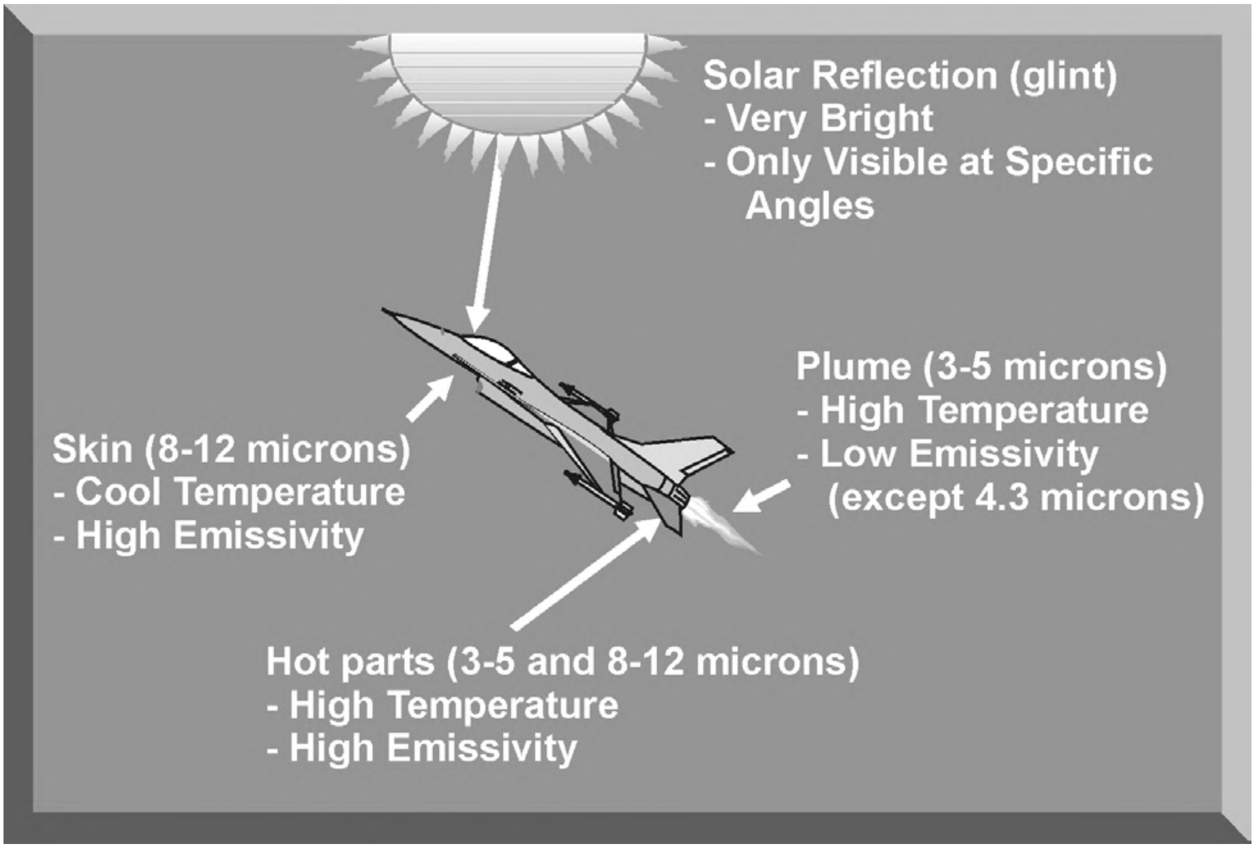
Solar Reflection (glint)
- Very Bright
- Only Visible at Specific Angles
Plume (3-5 microns)
- High Temperature
- Low Emissivity (except 4.3 microns)
Skin (8-12 microns)
- Cool Temperature
- High Emissivity
Hot parts (3-5 and 8-12 microns)
- High Temperature
- High Emissivity

Figure 14-7. Target Signature Sources

a. The main aircraft signature sources are the plume, the engine hot parts, and the skin. The skin is large in area, but usually fairly cool. Thus the best detection of the aircraft skin is usually in the long wave IR band (8-12 microns).

b. The engine hot parts offer excellent detection when one is looking at the right angle. They have high temperatures and high emissivity. However, depending on the viewing angle, they may have a low perceived area.

c. The plume has high temperatures and a high perceived area. This large perceived area allows near all-aspect detection. Unfortunately, it has a relatively low emissivity except near 4.2 microns. It is the strategy of the new breed of all-aspect missiles to detect the plume in the mid-wave IR, around 4.2 microns.

d. Figure 14-8 shows some of the common IR detector materials. Note that the cooled detectors are sensitive to longer wavelength (lower energy) photons, as discussed earlier.

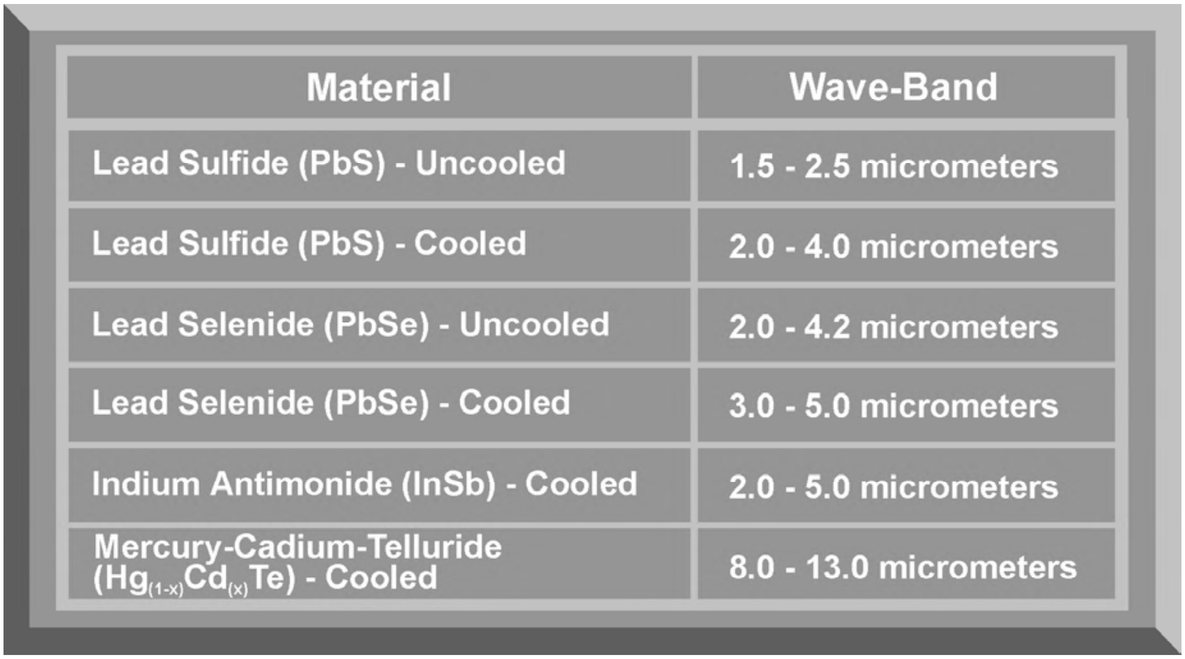

Material	Wave-Band
Lead Sulfide (PbS) - Uncooled	1.5 - 2.5 micrometers
Lead Sulfide (PbS) - Cooled	2.0 - 4.0 micrometers
Lead Selenide (PbSe) - Uncooled	2.0 - 4.2 micrometers
Lead Selenide (PbSe) - Cooled	3.0 - 5.0 micrometers
Indium Antimonide (InSb) - Cooled	2.0 - 5.0 micrometers
Mercury-Cadium-Telluride $(Hg_{(1-x)}Cd_{(x)}Te)$ - Cooled	8.0 - 13.0 micrometers

Figure 14-8. IR Detection Materials

4. IR SEEKER CHARACTERISTICS

Guidance units are designed to detect and home in on the IR radiation of the aircraft. The job of the seeker is to view the scene and output the estimated target position. Figure 14-9 shows the hardware that makes up the seeker of a generic IR missile system.

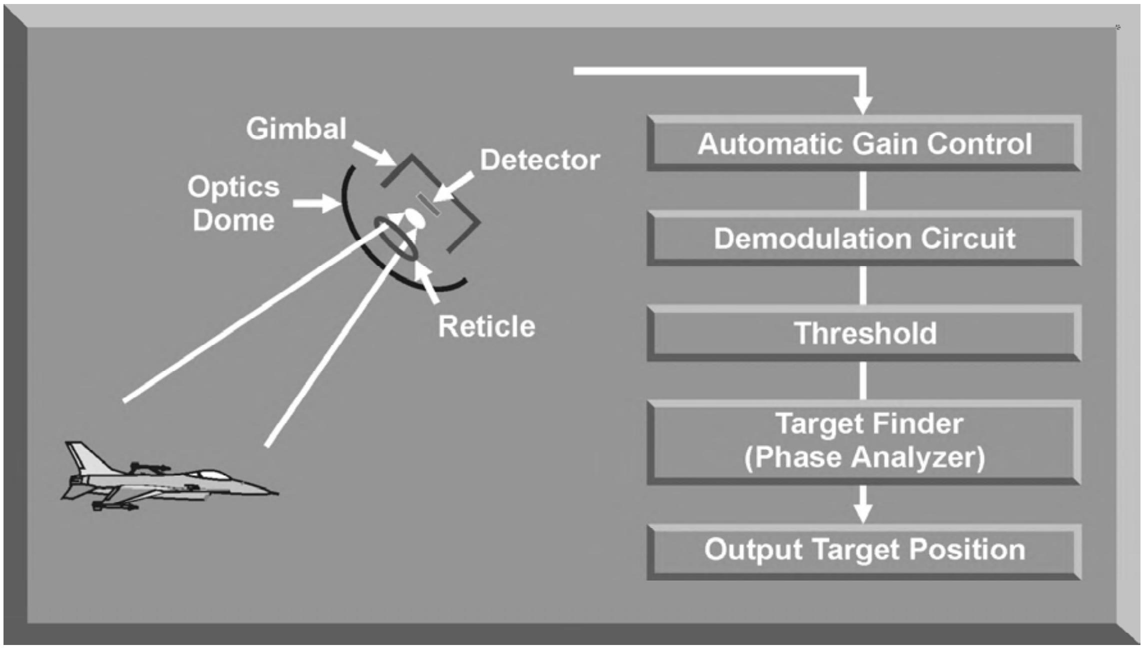
Gimbal
Optics Dome
Detector
Reticle
Automatic Gain Control
Demodulation Circuit
Threshold
Target Finder (Phase Analyzer)
Output Target Position

Figure 14-9. IR Missile Seeker

a. The dome acts as an IR transparent protective cover. Most IR missiles use the hemispherical dome because that shape does not alter the path of the incoming light rays. The optics focus the scene through the reticle onto the detector. The reticle pattern enables the detector output to "code" the position of the target.

b. The internal workings of the missile are similar to those found in a radar tracking system. There is an automatic gain control (AGC) to adjust the levels of the detector output so that it is not too large or too small. The demodulation circuit "decodes" the detector signal. The threshold circuit cancels signals that are below the specified threshold value, similar to the clutter rejection function found in radar systems. Finally, the phase analyzer reads the target position from the signal and sends the result to the autopilot. The autopilot then adjusts the missile's flight path to track the target.

c. IR systems employ filters and detectors to filter out unwanted IR radiation from the sun, sunlit clouds, smoke, the earth, and other background radiation sources (Figure 14-10). The detection unit is coupled with a guidance system to generate commands to the missile control vanes to keep the target centered in the field of view.

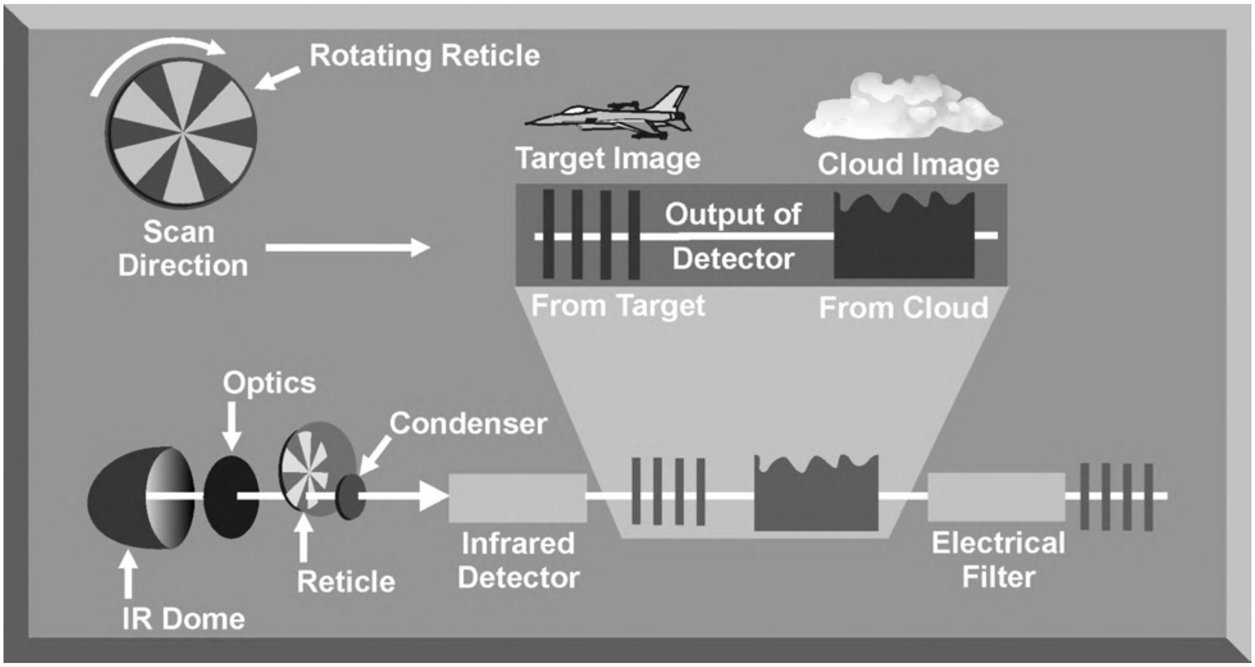

Figure 14-10. Spin Scan Reticle

d. A defining characteristic of an IR missile is its field of view (FOV) and field of regard (FOR) (Figure 14-11). The angular size of the image in degrees is called the field of view. To provide the greatest possibility of collecting IR radiation from the target, the receiver must have the greatest possible FOV. This in turn may

create problems. A large FOV increases the possibility that the receiver may not be able to distinguish the target from other sources of IR radiation. To avoid this problem, a relatively small FOV is scanned through a wider area. This wider area is referred to as the FOR. The gimbal enables the FOV to be scanned, so that an entire FOR can be searched.

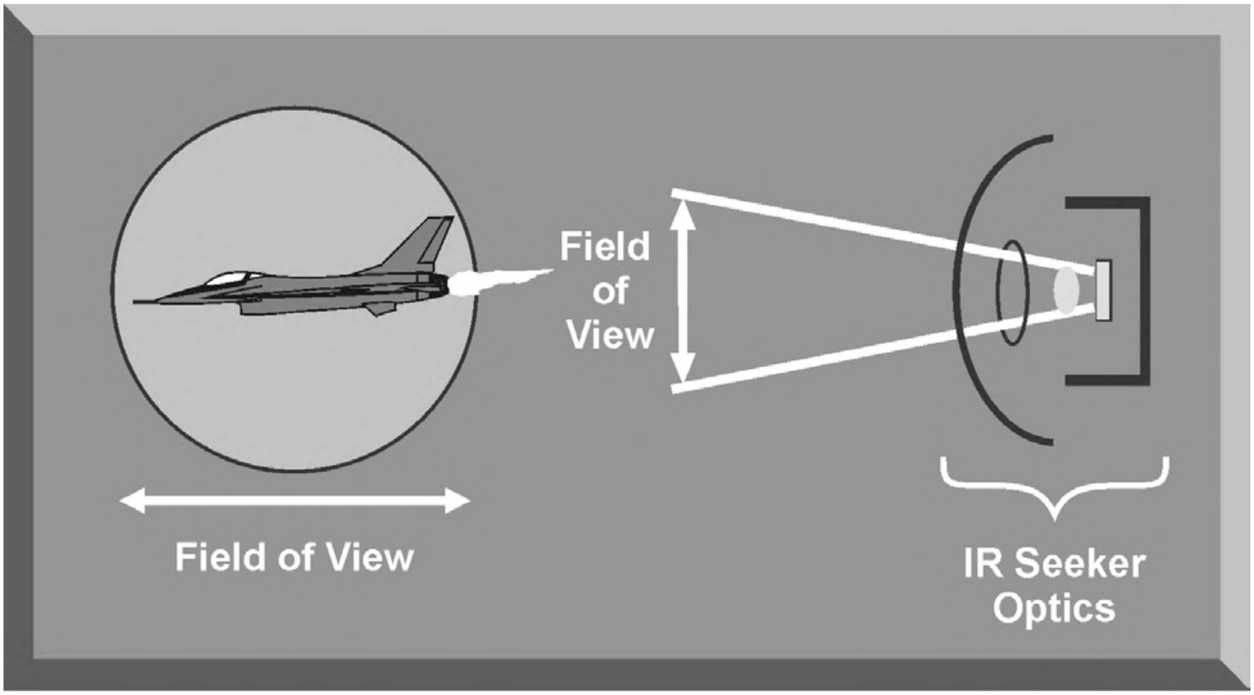

Figure 14-11. Missile Field of View

5. IR SEEKER TYPES

Through the years advances in seeker technology have resulted in significant changes to IR missile engagement tactics. This section will discuss spinning reticle, conical scan, cooled, and imaging IR seekers.

a. First generation IR missiles, like the SA-7, use a spinning reticle as the means to track the target. Due to their relatively low cost and ease of use, IR missiles of the first generation can still be encountered. The spinning reticle is inserted in the seeker just before the IR radiation reaches the detector. The reticle is a thin plate of optical material which has a transparent and opaque pattern on it. As the reticle is rotated, the IR energy is chopped at a rate determined by the reticle pattern. This system produces error signals when the target is not exactly centered in the field of view. Figure 14-12 is an example of a reticle pattern that can provide both azimuth and elevation information. If the target is located in the upper half of the pattern, the IR intensity on the detector is constant as the reticle rotates. As the pie-shaped half of the disc rotates over the target, the IR energy is pulsed and the amplitude of the pulses is an indication of relative elevation angle.

When the target moves to the right or left, the pulsing starts and stops at different times, indicating target azimuth. Center spun spin-scan seekers, also called center null reticles, are relatively insensitive when the target is in the center of the seeker scan where there is no tracking error. This is because the point target tends to bleed energy into all the spokes at once, eliminating the pulsed signal output of the detector. Once the target falls off the center of the reticle, the seeker generates an error signal that initiates guidance commands to recenter the target. This is the reason early IR missiles flew an undulating path toward the target.

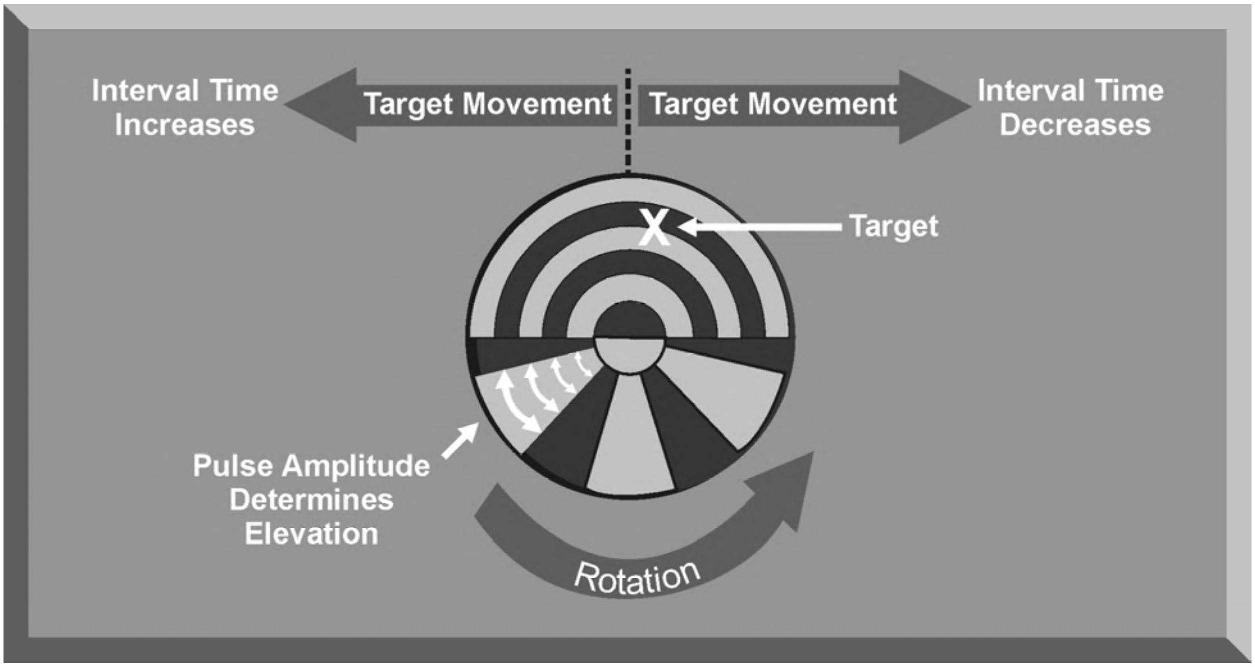

Figure 14-12. IR Missile Reticle

b. IR conical-scan (con-scan) seekers were developed to solve some of the problems with spin-scan seekers; notably the lack of error response when the target is near the center of the seeker field of view (FOV) (Figure 14-13). In a typical con-scan seeker, the reticle is fixed and does not spin. Instead, a secondary mirror is tilted and spun. This causes the target image to be scanned in a circular path around the outer edge of the reticle. When the target is centered in the seeker scan, the detector generates a pulsed output similar to that of the spin-scan seeker. However, as the target leaves the center, the output of the detector is a frequency modulated (FM) sine wave. The frequency of the modulation is directly proportional to the amount of target displacement from the center of the seeker scan. Con-scan optics are usually designed to spin the target very close to the edge of the reticle. This generates the greatest amount of FM modulation for a given target tracking error and gives the con-scan IR missile a more sensitive and tighter tracking loop. The center of the reticle is only used for acquisition.

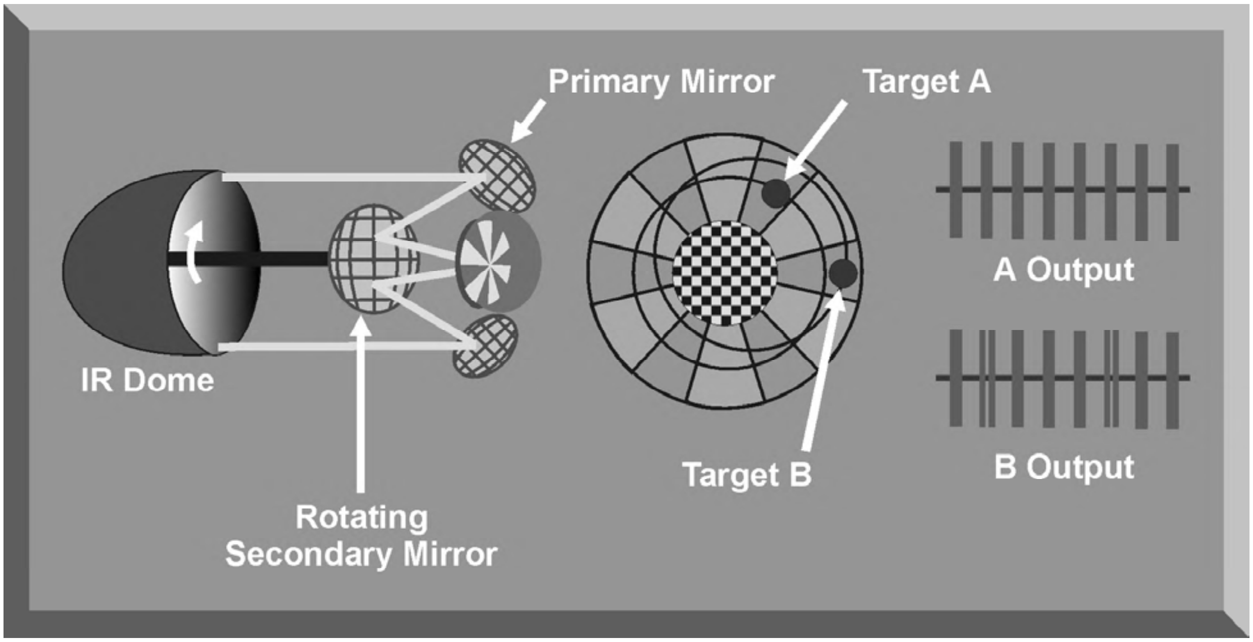

Figure 14-13. Conical Scan IR Seeker

c. A significant improvement to IR seekers resulted from the cooling of the detector with an inert gas such as argon. Older IR missiles, using uncooled lead-sulfide detectors, have a peak sensitivity in the 2 micron region. This limits these missiles, the SA-7 for example, to stern attacks because the missiles can only discriminate the IR signature of the engine turbine from background IR energy. By cooling the detector with an inert gas, like argon, the detectors of newer IR missiles can track longer wavelength IR radiation associated with airframe friction. Using newer detector materials like indium antimonide (InSb), require cooling to have increased target detection range and all-aspect tracking capability.

d. Imaging IR is the most recent advancement in IR seeker technology. The technology for these seekers is similar to that found in the AGM-65 Maverick missile. Imaging IR seekers are harder to decoy with flares than older seekers, and they are resistant to pulsed light jamming. Imaging detection involves creating an IR picture of the scene in one of two ways, scanning or staring. A scanning system uses one detector (or a mix of detectors and mirrors) which moves relative to the scene until the entire scene is scanned. This is an easy system to fabricate, but it can be noisy because the detector can't stay very long at each position, and it does not have a lot of time to measure the signal. A staring system uses many detectors, each of which detects a small portion of the scene. Each detector can "dwell" on its part of the scene for the entire frame time. However, such systems, also called focal plane arrays, are difficult to fabricate in a way such that each detector has the same sensitivity. One of the prime advantages with using imaging IR seekers is that they can be programmed

to track a particular IR shape or scene, significantly reducing the effectiveness of decoy flares.

6. IR MISSILE FLARE REJECTION

Flares are the primary countermeasure used to defeat the IR missile. Advanced IR missiles use different techniques to overcome the use of self-protection flares. There are two important characteristics of infrared (IR) missiles that influence the effectiveness of self-protection flares. The first is the ability of the IR missile seeker to discriminate between the IR signature of the aircraft and the IR signature of background interference, especially clouds. The second is the flare rejection capability built into the missile seeker and the missile guidance section.

a. The major source of IR background interference is sunlit clouds. Space filtering using a spinning reticle as described in the previous section is the method most widely used to suppress background interference and improve discrimination. IR energy from the target is gathered by the optical system and focused on the spinning reticle. The reticle chops the signal into a series of pulses that are focused onto the IR detector element. The signal output of a point source of IR energy, such as a target aircraft, will be a series of pulses. The signal output of a cloud, which covers several segments of the reticle, will be a single large pulse. An electrical filter eliminates the single large cloud pulse, and passes the multiple pulses generated by the target to the missile guidance section.

(1) IR missiles using spin-scan seekers have very little ability to reject flares. Flares provide the missile seeker with a hotter target than that of the aircraft, causing the spin-scan seeker to track the flare. The typical flare burns with a peak energy emission in the 2 micron range. Since the flare energy emission is greater than that of the target aircraft, the missile seeker transfers lock to the flare.

(2) Due to the scan pattern, con-scan seekers have some inherent resistance to flares. As described in the previous section, the reticle in a con-scan seeker is fixed and does not spin. Instead, a secondary mirror is tilted and spun. This causes the target image to be scanned in a circular path around the outer edge of the reticle. Because flares tend to drop away from the aircraft, they will drop off the con-scan reticle much faster than for a spin-scan reticle; therefore, flare resistance is built-in.

(3) Reducing the missile FOV is another method to help an IR missile discriminate the target IR signature from background IR (Figure 14-14). Limiting the FOV of the missile also makes the missile more resistant to flares. To decoy an IR missile seeker, the flare must create a heat source hotter than the aircraft and within the missile FOV. To decoy an IR missile with a narrow FOV, the flare must reach peak intensity almost immediately after ejection (Figure 14-15).

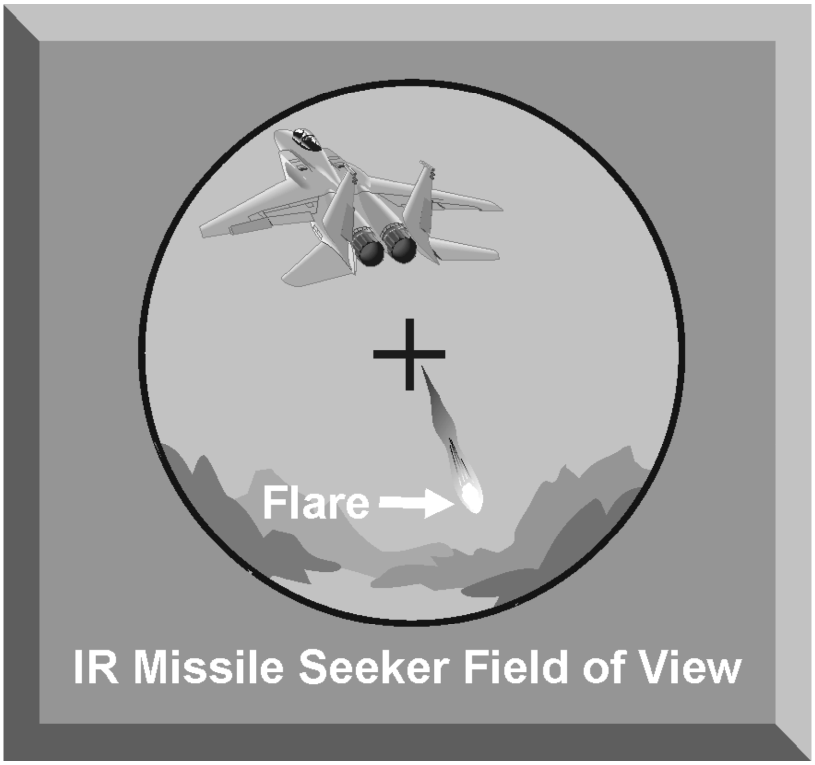
Flare
IR Missile Seeker Field of View

Figure 14-14. Impact of IR Missile Field of View

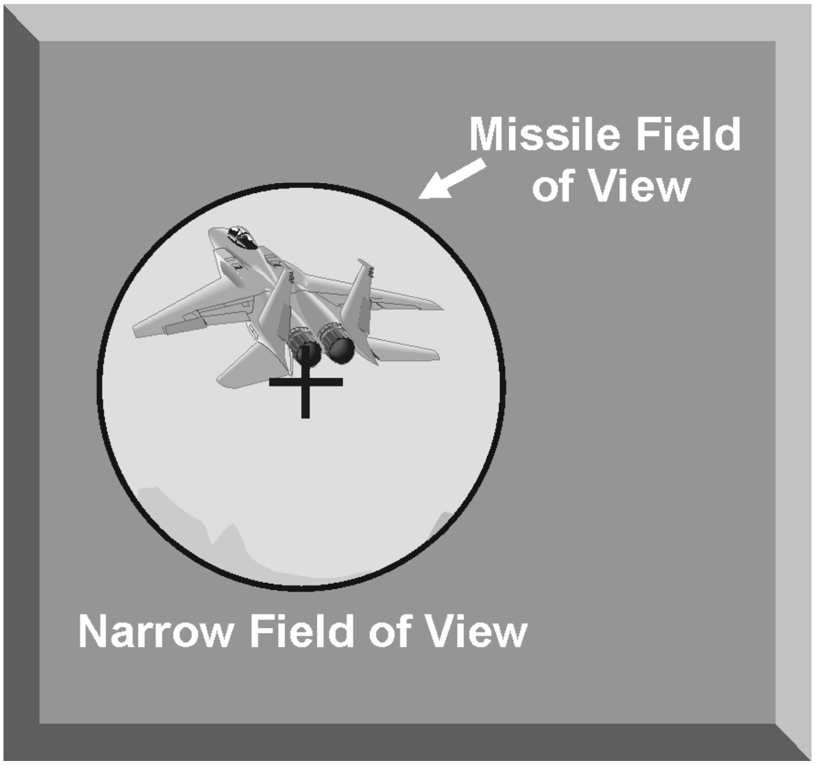
Missile Field of View
Narrow Field of View

Figure 14-15. IR Missile – Narrow Field of View

(4) The detector also influences the ability of an IR missile to discriminate target IR from background IR. Newer IR seekers can track longer wavelength IR radiation. The seeker's ability to track through a wider frequency range forces flares to decoy that wider frequency range to be successful. Additionally, the cooled seeker's ability to track from all angles could cause geometry problems, the missile may never see a flare ejected from the bottom rear of an aircraft.

b. IR missile designers have combined computer signal processing and modern IR missile seekers to detect the presence of flares and reject them as targets. This flare rejection capability, or IR counter-countermeasure (IRCCM), allows newer IR missiles to track the aircraft IR signature while rejecting multiple flares. Flare rejection is based on two computer functions. The first is called the "trigger," which detects the flare in the seeker field of view. The "trigger" function activates the "response" computer processing, the action the seeker takes to reject the flare. Both the "trigger" and "response" must operate to successfully reject flares. Each advanced IR missile employs different computer techniques for both the "trigger" and "response." A flare technique developed to counter one IR missile may not work against another IR missile that uses different flare rejection computer techniques.

(1) There are several "trigger" techniques that can be used by IR missiles to detect flares. Advanced IR missiles may employ one or more of these techniques to detect the presence of a flare in the seeker FOV. These "trigger" techniques include rise time, two-color, kinematic, and spatial.

(a) An IR missile using a rise time "trigger" monitors the IR energy level of the target. A sharp rise in the received IR energy within a specified time limit indicates a flare in the IR seeker FOV. When the missile detects this rapid rise in IR energy, the rise-time "trigger" triggers a flare "response." The "response" is switched off when the received IR energy drops to its original level. The thresholds for the rise in IR energy and the response time are set to preclude activating the rise time "trigger" when the aircraft selects afterburner. An IR missile using the rise time "trigger" can be decoyed by multiple flares with slow rise times.

(b) IR missiles using a two-color "trigger" to detect flares sample the energy level in two different wavelength bands. In Figure 14-16, a non-afterburning target would have more IR intensity in band B than in band A. The typical flare produces more IR intensity in band A than band B. A sudden increase in band A intensity compared to band B intensity indicates a flare in the seeker FOV. The two-color "trigger" would then trigger a flare "response." Advanced IR missiles using a two-color "trigger" can employ different detectors composed of different materials to monitor the intensity in two bands. A lead sulfide detector could be used for band A and indium antimonide for band B. IR missiles using a single detector can employ a reticle with different bandpass filters to monitor IR intensity in both bands. To track the target, IR missiles employing a two-color "trigger" may use either band, or use data from both

bands of a dual-channel tracker. An IR missile using the two-color "trigger" can be decoyed by multiple flares that provide equal IR intensity in each band.

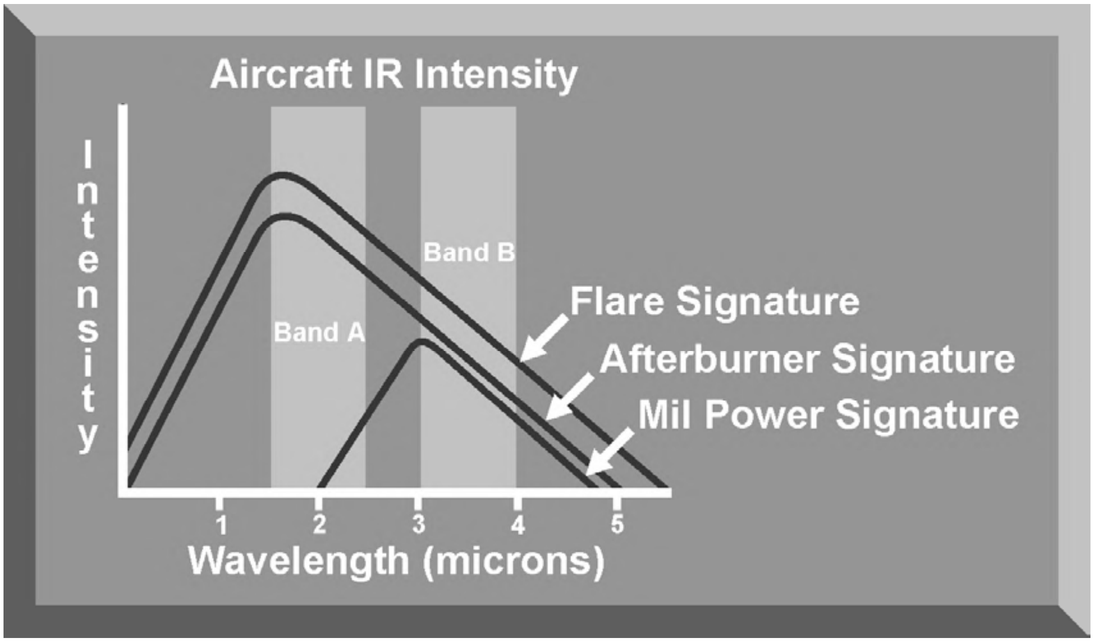

Figure 14-16. Two-Color IR "Trigger"

(c) A kinematic "trigger" takes advantage of the fact that flares separate very quickly from the dispensing aircraft due to aerodynamic drag. In a beam aspect engagement, an IR seeker that transfers track from the target to the flare will have a dramatic change in the line-of-sight rate due to the rapid separation of the flare from the aircraft. An IR missile employing a kinematic "trigger" detects this change and initiates the "response." IR missiles employing a kinematic "trigger" may have difficulty in flare detection in a head-on or stern engagement due to the small line-of-sight change between target and flare. Multiple flares dispensed at very short intervals will probably decoy an IR missile employing a kinematic "trigger".

(d) The spatial "trigger" operates like the kinematic "trigger" in that it uses the rapid separation of the flare from the aircraft to trigger the "response." An IR missile employing a spatial "trigger" uses the seeker to detect a flare. When the flare separates to the rear of the aircraft, the seeker will see the target on the edge of the FOV corresponding to the direction of target movement. The flare will be on the opposite side of the FOV. Once two hot objects on opposite sides of the FOV are distinguishable, the spatial "trigger" triggers the flare "response." IR missiles employing a spatial "trigger" can be decoyed by dispensing multiple flares at very short intervals.

(2) The seeker's "response" to the "trigger" is to reject the flare or limit its effect on target track. As long as the flare remains in the seeker FOV, the missile

is tracking the target in a degraded mode. Most IR seekers have a FOV of less than 2.5°. At long ranges, the flare will remain in the FOV for a relatively long period of time. At close range, the flare will be in the missile FOV for a relatively short period of time. Several different "response" techniques may be used, either alone or in combinations, to defeat a flare. These "response" techniques include simple memory, seeker push-ahead, seeker push-pull, and sector attenuation.

(a) When the simple memory "response" is initiated, the missile continues the maneuver it was performing just before the "trigger." This "response" assumes the flare will separate to the rear of the target. The missile rejects the seeker track data and maintains its motion relative to the target, waiting for the flare to leave the seeker FOV. The missile will continue to reject track data until the flare leaves the FOV or until the "trigger" times out. When the "trigger" times out, the "response" is discarded and the seeker operates in the normal track mode. If the "trigger" times out while a flare remains in the FOV, the seeker will usually transfer lock to the flare.

(b) The seeker push-ahead "response" causes the seeker gimbals to drive the seeker forward in the direction the target is moving (Figure 14-17). Pushing the seeker forward causes the flare to depart the FOV faster than with simple memory, minimizing the amount of time the missile is not tracking the target. The greater the amount of forward movement (called "bias"), the faster the flare will depart the FOV. If the amount of forward bias applied is too great, the seeker may be pushed forward of the target. This could cause both the target and the flare to depart the FOV and the missile would have to reacquire the target.

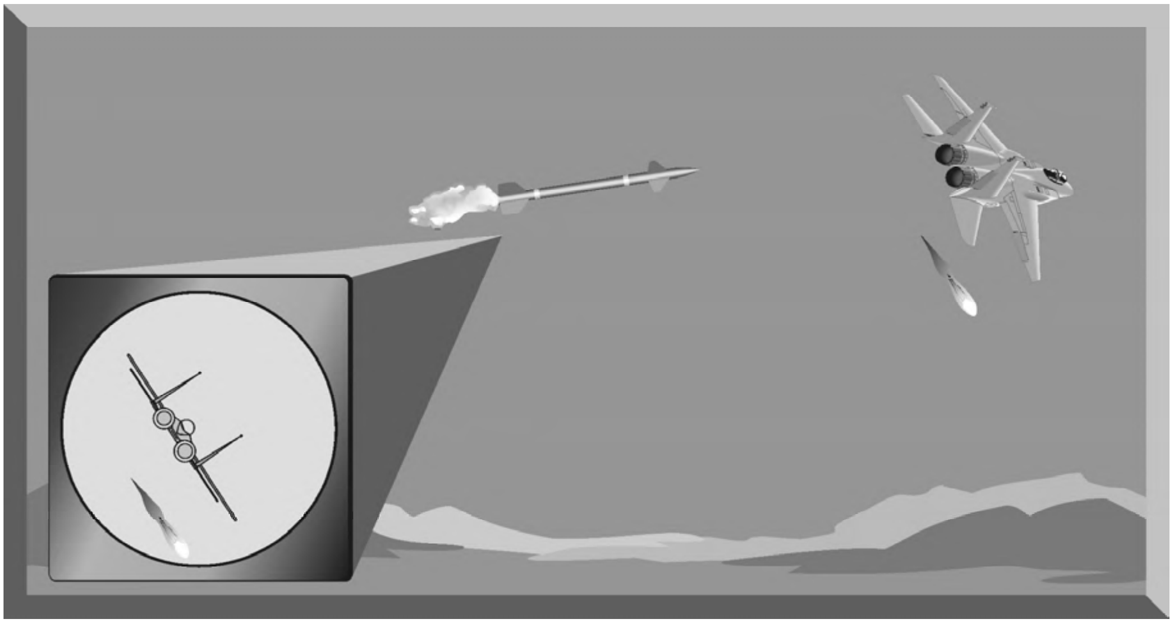

Figure 14-17. Seeker Push-Ahead "Response" Technique

(c) The seeker push-pull "response" assumes flares will have a higher intensity IR signature than the target. The "response" is initiated when the target and flare are on opposite sides of the seeker FOV (Figure 14-18). This corresponds to a spatial "trigger" condition. The received energy will rise and fall as the energy of the target and flare is scanned across the detector. When the flare energy is at a peak, the seeker gimbals drive the seeker away from the flare. When the lesser energy from the target is detected, the seeker's gimbals pull the seeker in the direction of the target. As a result, the seeker is moved away from the flare and toward the coolest IR source in the FOV, the target aircraft.

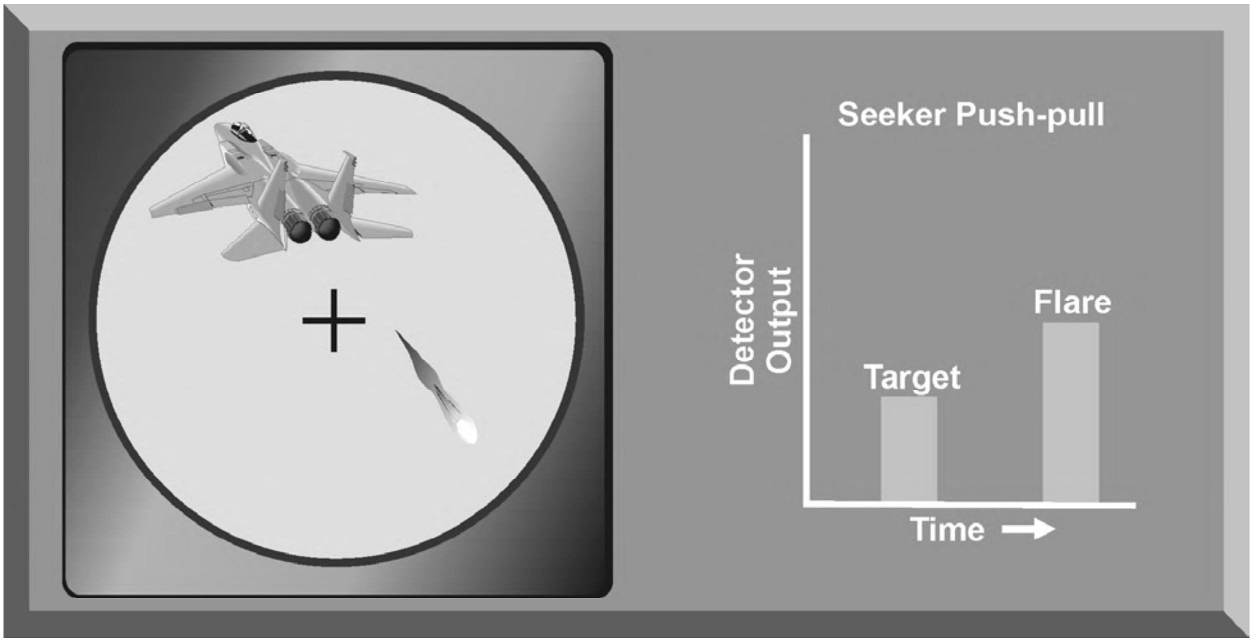

Figure 14-18. Seeker Push-Pull "Response" Technique

(d) The sector attenuation "response" is initiated by placing an attenuation filter across part of the seeker FOV (Figure 14-19). This filter reduces the seeker sensitivity in that part of the FOV. If the target being tracked is in the center of the FOV, then placing an attenuator in the quadrant below and to the rear of the target should reduce any energy received from a flare. If the attenuated flare energy is below that of the target energy, the seeker will continue to track the target.

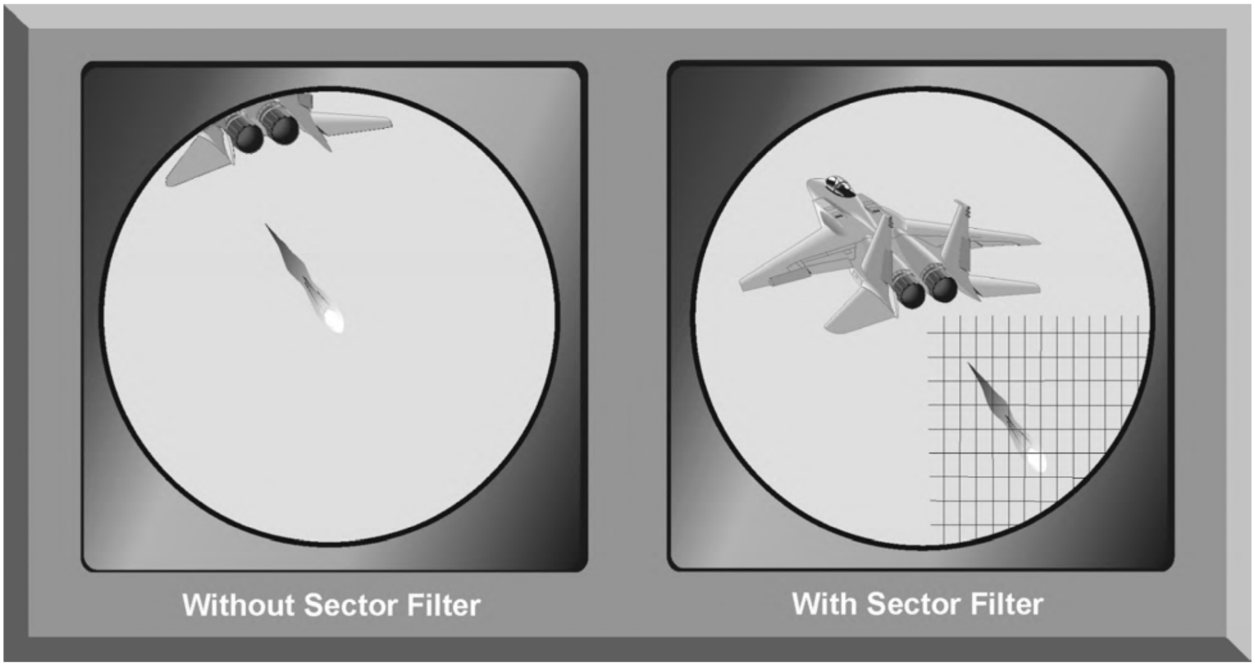

Figure 14-19. Sector Attenuation “Response” Technique

7. SUMMARY

This chapter provided some background on IR theory and IR detection, then went into some of the basics about the different types of IR missile seekers. Finally factors that impact the ability of an IR missile to reduce the effectiveness of self-protection flares were discussed. This ability depends on the discrimination capabilities of the missile seeker, the type of detector, the missile FOV, and the missile flare rejection capabilities. Modern IR missiles that employ sophisticated flare rejection techniques and advanced missile seeker technologies present a growing, and potentially lethal, threat.

CHAPTER 15. IR COUNTERMEASURES

1. INTRODUCTION

Defeating IR missiles used to be an afterthought placed well behind surviving the radar missile threat. Now with the proliferation of advanced IR missiles, defeating the IR threat is becoming more important and more difficult. The primary countermeasure used to defeat IR missiles is the self-protection flare. This chapter will discuss flare characteristics and employment considerations. One of the main difficulties in defeating an IR missile is knowing that one has been launched. This chapter will also discuss some of the methods currently available to detect an IR missile attack.

2. FLARE CHARACTERISTICS

Self-protection flares were developed to counter threat systems operating in the IR spectrum. Self-protection chaff and flare dispensers, such as the ALE-40, ALE-45, or the ALE-47, are designed to allow the pilot to dispense flare cartridges when engaged by an IR threat. These flare cartridges are pyrotechnic and pyrophoric devices designed to produce an IR source that is more attractive than the IR signature of the aircraft. To decoy an IR missile seeker, the flare must create a heat source more attractive than the aircraft, within the missile field of view. The most important flare characteristics that determine the ability of a flare to decoy an IR missile are IR wavelength matching, flare rise time, and flare burn time. The MJU-7 flare cartridge will be used as an example of a typical flare cartridge.

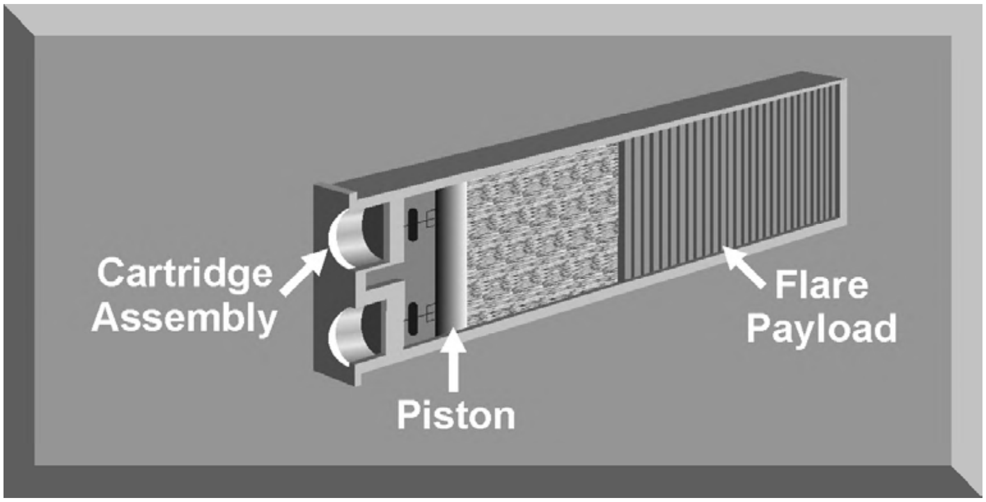

Figure 15-1. The MJU-7 Flare

a. The MJU-7 flare cartridge (Figure 15-1) is an example of a flare cartridge used in both the ALE-40 and the ALE-47 dispensers. The flare grain is composed of magnesium and tetraflouroethylene, or C_2F_4, which burns at 2000 to 2200°K. As the flare burns, it emits IR energy of different wavelengths from the luminous zone that emulates the aircraft IR signature. The burning flare also produces a large quantity of white smoke, which may highlight the position of the dispensing aircraft.

b. A flare must reach peak intensity shortly after ejection or it will not be effective in decoying the IR missile seeker. Flare rise time is the time required for the flare to reach peak intensity.

(1) To counter a short-range IR missile with a narrow field of view, a flare must reach peak intensity quickly. On the other hand, some advanced IR missiles now look for a rapid rise in IR energy as a trigger to know when a flare is trying to decoy it (Figure 15-2).

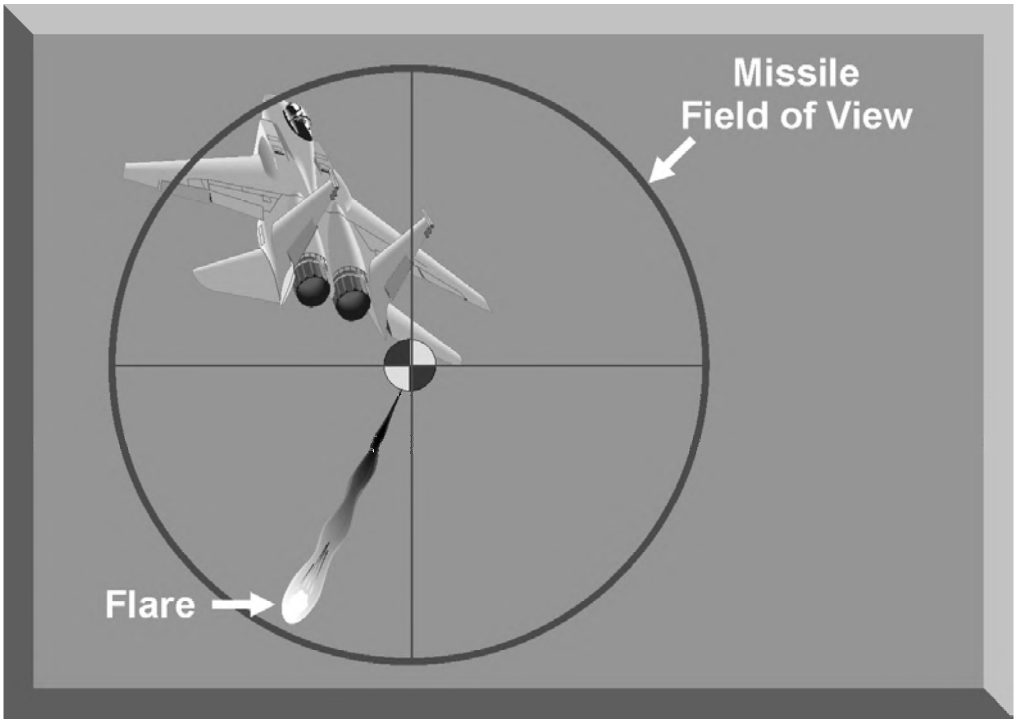

Figure 15-2. Impact of Flare Rise Time

(2) Flare rise time varies dramatically with altitude. The flare burns longer at high altitude, but it takes much longer to reach peak intensity. This increase in flare rise time at high altitude can impact the effectiveness of flares to decoy and defeat IR missiles particularly in the air-to-air combat environment.

c. Flare burn time is the time span that the flare burns and determines how far the IR seeker will be pulled off the target. The longer the burn time, the longer the

IR seeker is pulled off the target aircraft. The flare should burn long enough to ensure that the aircraft is no longer in the missile field of view. Flare burn time, like flare rise time, varies with altitude. A flare will burn longer, but at lower intensity, at higher altitudes. A longer flare burn time increases the probability the IR missile will be decoyed by a single flare.

3. ADVANCED FLARES

To counter the advances in IR missile seeker technology the Air Force and Navy formed the Advanced Strategic and Tactical Expendables program to develop and field advanced IR decoys. Two products to come out of this program are the kinematic flare, MJU-47, and the covert flare, MJU-50/51.

a. Kinematic Flares. A significant characteristic of conventional flares is that upon ejection they rapidly slow down and separate from the aircraft. As described in Chapter 14, modern IR missiles exploit this rapid separation between the aircraft and the flare. The rapid separation triggers a flare rejection response in the missile seeker causing it to ignore the flare and continue tracking the aircraft. The kinematic, or thrusted, flare delays the missile response by propelling itself in the direction of the aircraft and negating or delaying the missile's flare rejection trigger. The MJU-47 is the same size as the currently fielded MJU-10 flares. The flare's decoying pyrotechnics are vectored out the end of the flare to provide a means of propulsion (Figure 15-3).

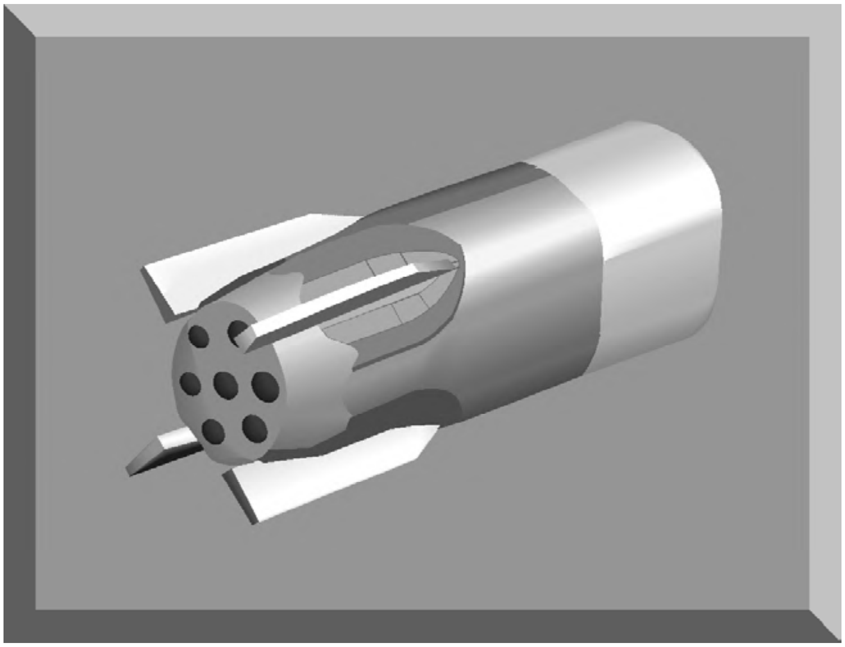

Figure 15-3. Kinematic Flare

b. Covert Flares. Though an effective tactic at defeating MANPADS, pre-emptively dropping flares in a target area carries a significant risk. The risk is that enemy air defense systems that may not have known the aircraft's location will surely see the flares, day or night, and any chance for surprise is lost. Figure 15-4 shows the visual signature of conventional flares. The danger of highlighting oneself becomes even more likely as aircraft become fitted with chaff and flare systems that automatically dispense expendables based upon inputs from the RWR or a missile attack warning system (MAWS). Automatic systems tend to error on the side of caution which means expendables will be dropped if there is any ambiguity. Visually covert flares, named the MJU-50 and MJU-51, remove this problem by not leaving a visual signature such as smoke or flame. These covert flares are made of material that oxidizes, pyrophoric instead of pyrotechnic, when released in the air. The MJU-50 is the size of the small M-206 flare while the MJU-51 is the size of the MJU-7 flare.

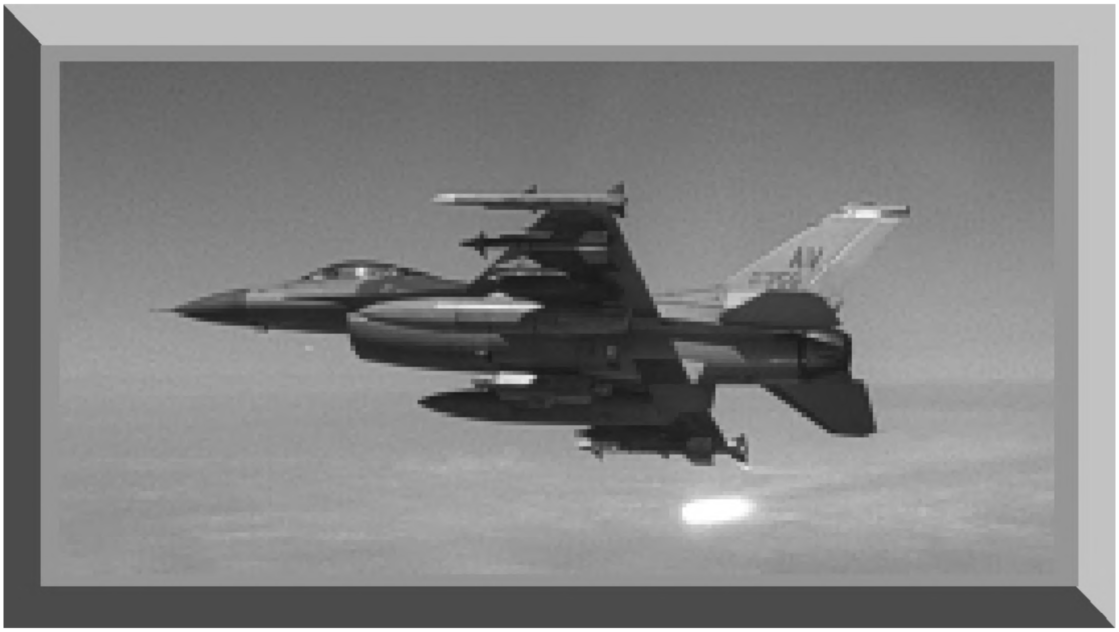

Figure 15-4. Flare Visual Signature

4. FLARE EMPLOYMENT

The purpose of employing a self-protection flare cartridge is to decoy the seeker head of an IR missile. This is accomplished by presenting the IR missile with a second heat source with an IR signature that exceeds the aircraft signature. The flare or IR source must appear in the field of view of the IR missile at the same time as the aircraft. As the flare separates from the aircraft, the IR missile seeker tracks the most intense IR signature, which ideally is the flare, and is decoyed away from the aircraft.

a. It is important to perform maneuvers along with flare employment to effectively defeat IR missiles. Maneuvers compound the IR missile's tracking problems and increase the distance between the aircraft and flare. Maneuvers that put the aircraft outside the IR missile field of view increase the ability of the flare to decoy the missile. When engaged by an IR threat, one tactic is to immediately dispense a flare and maneuver to put the missile on the beam. This increases the line-of-sight rate the missile guidance system must process and gives maximum separation between the aircraft and the flare (Figure 15-5).

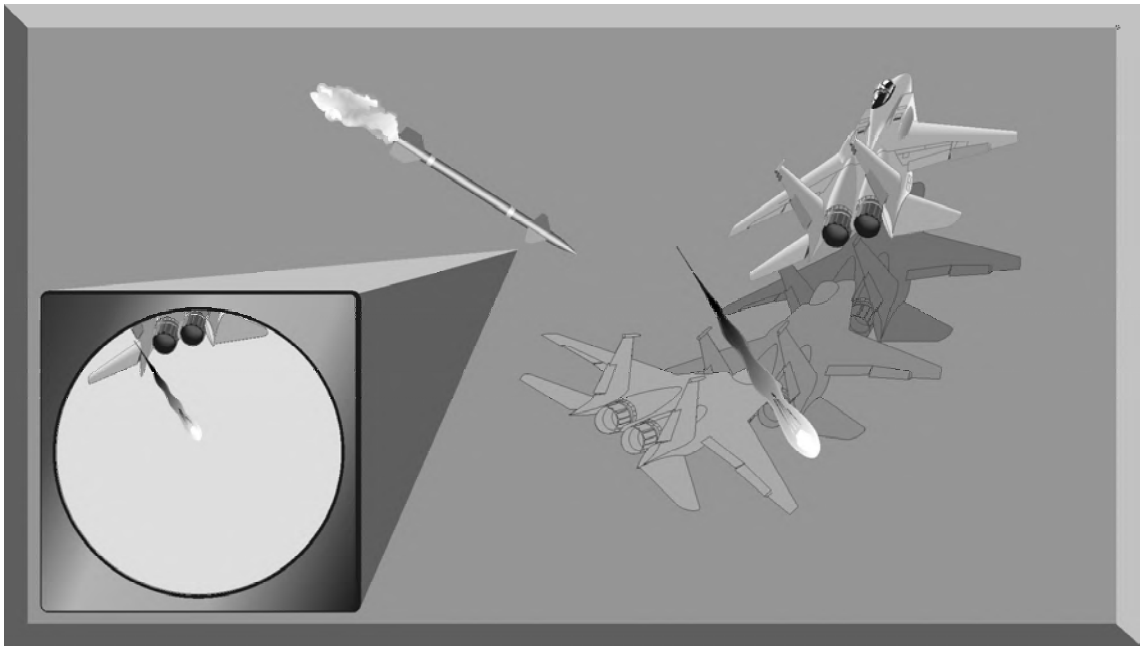

Figure 15-5. Initial Maneuver and Flare Employment

b. With the IR threat on the beam, the pilot has the best chance for achieving a "tally-ho" on the missile to determine range and keep the missile on the beam. Visually acquiring the missile increases the chances of surviving the encounter. Modern chaff and flare dispensers, such as the ALE-47 and ALQ-213, can be programmed to dispense flares in a sequence optimized to defeat specific IR missiles. Programs are selected by the pilot based upon the most likely threat. Classified tactics manuals provide aircraft specific maneuvers and flare dispensing programs to defeat the IR threat.

5. IRCM TACTICS

Besides maneuvers and flare employment, there are other IR countermeasures that can reduce the effectiveness of IR threat systems. The first is to reduce the intensity of the heat signature of the aircraft by reducing the power setting. The engines produce the largest IR signature and are the only source of IR radiation that the pilot can influence. By reducing the power setting, the IR signature is

reduced and the required flare intensity is thereby reduced. Use of afterburner should be minimized for exactly the same reason; the required flare intensity necessary to cover an aircraft in afterburner is difficult to obtain. The pilot should not reduce the power setting below the minimum required to maintain sufficient maneuvering airspeed.

a. Another effective IRCM is to use the sun. Maneuvering into the sun masks the aircraft's IR signature from most IR threat systems. When attacking a target defended by IR systems, an attack axis that places the sun behind the attacker may limit the effectiveness of these systems.

b. Clouds or smoke can also confuse IR threat systems (Figure 15-6). The water vapor making up clouds diffuses the IR energy making it difficult for an IR system to get a point to track, similar to how sun's light is diffused on a cloudy day. The particles found in smoke can have a similar effect on IR energy. The final IRCM is to reduce the IR signature of the aircraft by careful design of the engines and exhaust system. For example, the F-117 was specifically designed to provide the lowest possible IR signature.

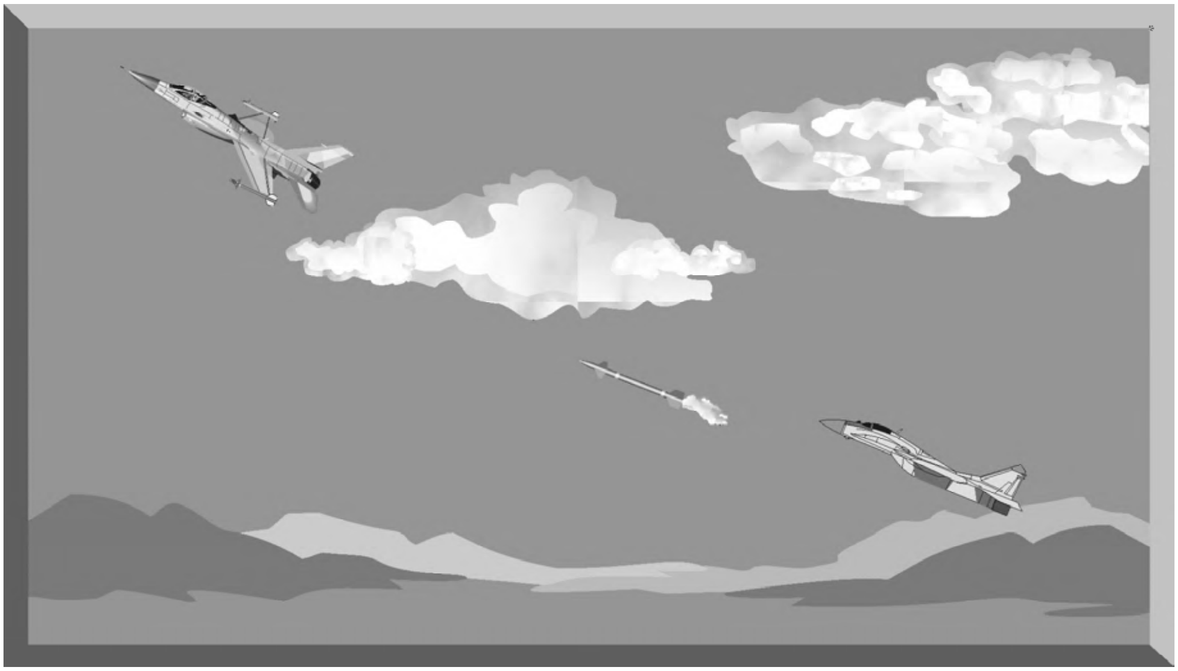

Figure 15-6. IRCM

6. MISSILE APPROACH WARNING SYSTEMS (MAWS)

One of the most important factors in defeating an IR missile attack is knowing that an attack is in progress. Since IR threat systems are generally passive, the radar warning receiver (RWR) will provide no attack warning unless the threat uses some detectable radar energy for acquisition prior to launching an IR

missile. MAWS are designed to provide the crew some warning of an ongoing missile attack. These systems have been widely deployed on larger aircraft. Advances in technology allowing them to be smaller and promising lower false alarm rates have increased the possibility that fighter size aircraft may be outfitted with this equipment.

a. The two primary requirements for MAWS are timeliness and reliability. An IR missile engagement is an extremely short event providing very little time for the targeted aircraft to respond with maneuvers and flares. Secondly, there must be a very low false alarm rate. In earlier systems the sun, flares, or the wingman's aircraft have been known to cause false alarms. Aside from being annoying, a false alarm can become a tactical problem if the MAWS is tied directly into the countermeasures dispenser set allowing for automatic dispensing of chaff and flares when the MAWS senses an attack. False alarms in this type of scenario would only serve to highlight a previously untargeted aircraft.

b. MAWS detect incoming missiles either actively or passively. The active MAWS use a pulse Doppler radar to detect and track the in-flight approach of an attacking missile. Pulse Doppler radar is used for this purpose because of its ability to use relative velocity to pull targets out of clutter. An incoming missile will have a high velocity relative to the surrounding background. Range and time to impact are computed automatically, updated continuously, and provided to the crew to assist in defensive maneuvers (Figure 15-7). The down side to active MAWS is the requirement to continuously radiate for long periods of time to ensure protection. This contradicts the common tactic of keeping radar transmissions to a minimum to avoid being tracked via passive detection measures. This situation has led to the latest generation of MAWS being mostly of the passive type.

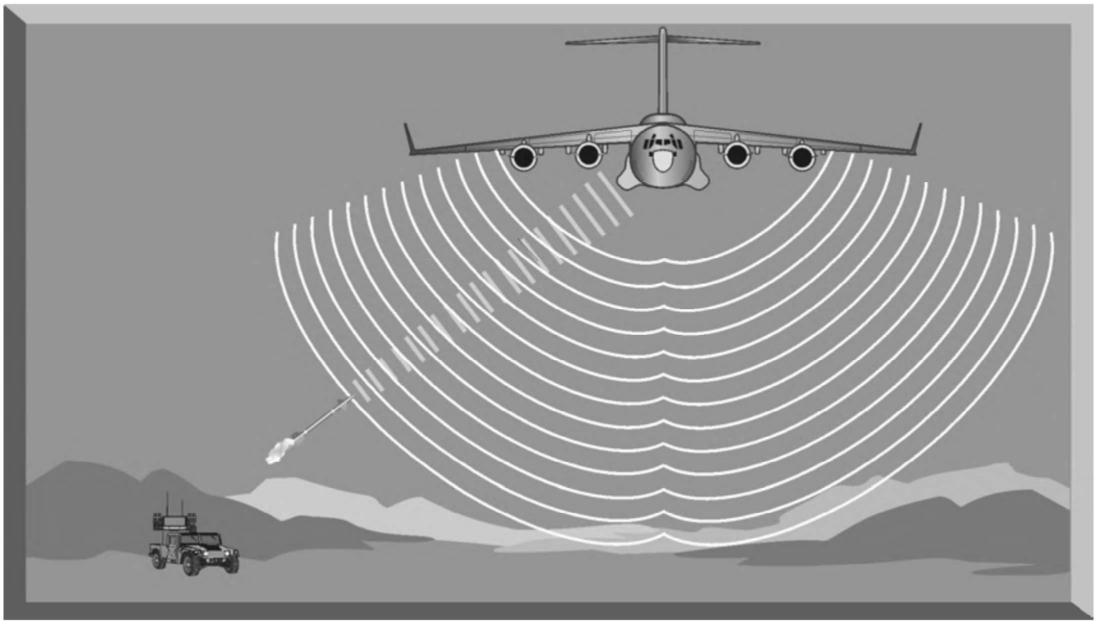

Figure 15-7. Missile Alert Warning System (MAWS)

c. The passive MAWS operate in either the ultraviolet, or the infrared frequency spectrum. These systems are tuned to look for the IR signature of a missile's rocket motor, then pass a warning and the position of the inbound threat to the pilot. To obtain all-aspect missile warning, multiple MAWS sensors must be positioned on the aircraft in a similar fashion to RWR antennas. All-aspect MAWS face the same challenges that RWR systems face: false multiple targets due to maneuvers and accurate threat position reporting, to name a few.

7. SUMMARY

The proliferation of IR threat systems has elevated the importance of IRCM to survival on the modern battlefield. The most effective IRCM is still the employment of flares in conjunction with maneuvers. The specific flare characteristics of IR spectrum coverage, rise time, and burn time are critical factors in determining flare effectiveness. Other IRCM tactics are designed to enhance the effectiveness of flares and take advantage of IR missile limitations. The effectiveness of all IRCM tactics depends on some type of attack warning. The new generation of MAWS should provide some measure of warning to the crews and in some cases automatically defeat the IR missile.

CHAPTER 16. RADAR ELECTRONIC PROTECTION (EP) TECHNIQUES

1. INTRODUCTION

Electronic warfare (EW) is defined as military action involving the use of electromagnetic and directed energy to control the electromagnetic spectrum or to attack the enemy. Nearly every military action, from command and control of an entire integrated air defense system (IADS) to precision guidance of an individual weapon, depends on effective use of the electromagnetic spectrum. Radar systems have become a vital element of nearly every military operation. Since these systems operate across the entire electromagnetic spectrum, much of the EW effort is concerned with countering radar systems. All of the jamming techniques discussed in Chapters 10 and 11 and the chaff employment options discussed in Chapter 13 are specifically designed to counter radar systems. These actions are classified as electronic attack (EA), which is a part of EW.

a. EW is somewhat like a chess game—a series of moves and countermoves within the electromagnetic spectrum. As we develop jamming techniques to counter radar systems, our adversaries develop counter-countermeasures to negate the effectiveness of these techniques. In response, we develop newer techniques and our adversaries respond with new modifications to their radar systems. This series of moves and countermoves can continue for decades. The development and application of radar counter-countermeasures are classified as electronic protection (EP), also a part of EW.

b. The continuing battle to control the electromagnetic spectrum for unrestricted radar employment has resulted in over 150 radar EP techniques. These techniques are designed to negate the effectiveness of electronic jamming and chaff on radar systems. These radar EP techniques can be incorporated into the design of a radar system or added to an existing radar system in response to a jamming technique. It is beyond the scope of this text to discuss all the radar EP techniques in use today. This chapter will discuss the most common EP techniques. They have been organized by function of the technique within the radar. These functions include radar receiver protection, jamming avoidance, jamming signal exploitation, overpowering the jamming signal, pulse duration discrimination, angle discrimination, bandwidth discrimination, Doppler discrimination, and time discrimination.

2. RADAR RECEIVER PROTECTION

The following are some of the most common radar counter-countermeasures designed to prevent receiver overload or saturation.

a. Sensitivity time control (STC) is used to counter close-in chaff or close-in clutter. Receiver gain is set at normal for long ranges and reduced for close-in ranges. One problem with using STC is that close-in targets may be missed if STC is improperly adjusted.

b. Automatic gain control (AGC) is used to counter chaff, clutter, and most types of transmitted jamming. AGC senses the signal level of a receiver's output and develops a back-bias, producing a constant output level. This technique has a slow response time compared to fast AGC and instantaneous AGC, both of which are employed instead of AGC. Also, it cannot maintain correct IF output levels for different intensity signals that are close in range because the bias voltage has a long buildup and decay time.

c. Fast automatic gain control (FAGC) is also employed against chaff, clutter, and most types of transmitted jamming. FAGC works by sensing the signal level of receiver output and develops a back-bias, tending to hold output constant. Response time is within milliseconds, permitting fast response and recovery as the antenna traverses the jammer's bearing. There are several precautions to note when using FAGC. First, targets may be suppressed and lost without the operator knowing that jamming is present. Second, a strong pulse or echo may cause ensuing weak targets to be lost. Lastly, FAGC has difficulty getting an accurate bearing on the jamming source.

d. Instantaneous automatic gain control (IAGC) is another technique to counter chaff, clutter, and most types of transmitted jamming. IAGC senses the signal level of each echo or jamming pulse and develops a back-bias that holds the stage output constant. Gain control response time is within milliseconds and extends the dynamic range of the receiver. There are several precautions to note when using IAGC. First, it is not effective against signals whose "in band" time is less than the IAGC response time. Also, with continuous duty cycle jammers, targets may be lost without the operator knowing that jamming is present. Finally, it is difficult to get an accurate bearing on the jamming source.

e. Automatic noise leveling (ANL) counters noise jamming and modulated or unmodulated constant wave jamming. ANL samples receiver noise content at the end of each PRF and sets the gain accordingly for the next pulse interval. Continuous jamming reduces gain to keep output the same as the original noise level. ANL also follows the scanning rate of the antenna so that receiver noise output is constant as the antenna rotates. When using ANL, targets may be suppressed and lost without the operator knowing that jamming is present. Also, receiver gain is unstable when pulses or swept jamming enter the sampling gate intermittently.

f. The logarithmic receiver (LOG) counters most types of transmitted jamming by amplifying and demodulating large dynamic-range signals in logarithmic amplifiers. This produces "amplitude compression" of the strong

signals. However, when using LOG, output is nearly constant so the operator cannot easily tell when jamming is present.

g. The logarithmic receiver with fast time constant (LOG-FTC) counters narrowband jamming, chaff, and clutter. This technique amplifies and demodulates large dynamic-range signals in logarithmic amplifiers, producing "amplitude compression" of the strong signals. Video is coupled through FTC circuits to eliminate rectified carrier and low frequency sideband products. There are problems with using LOG-FTC. First, the receiver output is nearly constant, so the operator cannot always tell when jamming is present. Second, LOG-FTC is not effective against wideband, or fast-swept, short pulse jamming. Lastly, LOG-FTC causes a broadening of displayed jam sector. as well as degrading bearing accuracy on the jam source.

h. Dicke-fix (DF) counters wideband and fast-swept jamming and is similar in employment to wideband limiting (WBL). DF amplifies without ringing, clips down all pulses to a common level, then amplifies the narrowband echo signal more than the wideband jamming. Noise level is held constant, independent of jamming intensity. There are precautions to note when using DF. Jamming that enters the wideband limiter can capture limiters, causing poor receiver sensitivity. Targets may be suppressed without the operator knowing that jamming is present. Also, resolution and target detection range are reduced, even in a non-jamming environment. Finally, DF is ineffective against extremely fast-swept spot jamming.

i. WBL is used to counter wideband jamming and fast-swept jamming. WBL amplifies without ringing, clips down all pulses to a common level, then amplifies the narrowband echo signal more than the wideband jamming. Noise level is held constant, independent of jamming intensity. However, jamming that enters the wideband limiter can capture limiters and cause poor receiver sensitivity. Resolution and target detection range is reduced, even in a clear environment. Targets may be suppressed without the operator knowing that jamming is present. Finally, WBL is ineffective against fast-swept spot jamming.

j. Adaptive video processing (AVP) counters chaff corridors, weather, sea clutter, and most types of transmitted jamming. AVP combines the adaptive threshold, beam-to-beam correlation, and wide-pulse blanking in frequency-scanning three-dimensional radars to avoid collapsing undesired returns on the PPI display. However, when using AVP, there is a decreased probability of detection in some multiple target situations. Also, targets may be suppressed without the operator knowing that jamming is present. Finally, AVP passes all point targets.

3. JAMMING SIGNAL AVOIDANCE

The following are some EP techniques used to avoid jamming signals.

a. Frequency agility (FA) counters narrowband jamming and some types of repeater and deception jamming. FA enables the radar to make rapid changes of transmitter and receiver operating frequency, sometimes on a pulse-to-pulse basis. Manual frequency changes may cause mutual interference with other radars and services.

b. Frequency diversity works against narrowband jamming and some types of repeaters and transponders. It is a multiple-radar coordination procedure in which radars are assigned operating frequencies that are separated to reduce mutual interference and their susceptibility to a single jammer. It is important to note that other radars may be operating at the same allocated operating frequency.

c. Polarization diversity is used to counter chaff, weather, and transmitted jamming. Polarization diversity attenuates jamming input to a radar receiver by using antenna polarization different from jammer polarization, and usually involves separate radars of different polarization. There are two precautions when using polarization diversity: (1) ground clutter worsens on vertical polarization, and (2) close coordination is necessary if separate radars are used; for example, one horizontally polarized search radar and one vertically polarized search radar.

d. Circular polarization (CP) works against chaff, weather, and transmitted jamming. CP attenuates jamming input to a radar receiver by using antenna polarization different from jammer polarization, and usually involves separate radars of different polarization. CP also improves target detection in rain clutter.

e. Conical-scan-on-receive-only (COSRO) is employed against inverse conical scan jamming to deny a jammer the ability to sense and upset scan angle tracking information. A constantly transmitted illumination beam is received and scanned to derive target angle information. However, the jammer can still degrade angle tracking if it can approximate the received signal scan rate.

f. Speedgate tracking is used against all types of transmitted jamming. The technique provides a very narrow bandpass having a center frequency related to Doppler shift. Only jamming within the restricted band is effective. It has the advantages of accurate target Doppler discrimination and good target tracking at low target levels. However, the speedgate can be stolen by gate stealers and some types of swept jamming.

g. Leading-edge track (LET) is used to counter an incoming target dropping chaff by allowing target tracking on the leading edge of the target. Trailing edge track (TET) is used to counter a receding target dropping chaff.

h. Track coast is used to counter chaff, clutter, multiple targets, range gate stealers, jam fades, and blinking jamming by placing tracking radar in a rate-aided coast condition. The system "estimates" target position to avoid interrupting the fire control solution. A lock-on or return to acquisition mode terminates the track coast condition. Track coast requires adequate storage of rate-aided information, and no true tracking information will be developed while track coast is operating.

i. Guard gates work against chaff, clutter, multiple targets, range gate stealers, jam fades, and blinking jamming. Guard gates provide automatic detection of a foreign signal and "estimates" target position to avoid interrupting fire control solutions. Like track coast, guard gates require an adequate store of rate-aided information with no true tracking information developed.

4. JAMMING SIGNAL EXPLOITATION

The following are some EP techniques that use the jamming signal for target acquisition and engagement.

a. Passive angle tracking (PAT) counters most types of transmitted jamming by allowing the radar to acquire and angle-track the source of jamming signals. There are some problems with this technique. Blinking jamming can cause severe instability, and the range of the jammer is unavailable until the target reaches burnthrough range.

b. Home-on-jamming (HOJ) counters most types of transmitted jamming by allowing the missile or radar to use the jamming signals, locate the source, and home on it. However, blinking jamming can cause severe instability, and the range of the jammer is unavailable until burnthrough.

c. Jamming signals produce recognizable sounds that help in their detection and identification. Aural recognition allows an operator to listen to the Doppler frequency associated with a moving target. It is used to counter most types of jamming.

d. The local oscillator off technique counters continuously transmitted jamming. No receiver output occurs unless a target echo signal and a jamming signal are present. Limitations of this technique include: targets only display in an area where jamming is present; and, if the antenna rotates away from the jammer, or if jamming is turned off, no targets are displayed on the radar scope.

e. The jamming strobe indicator counters any transmitted jamming with high-duty-cycle modulation. The indicator is a variable marker strobe on the radar display that moves in range proportional to jamming strength. The indicator traces an antenna lobe pattern on the display, showing the azimuth of the jamming source. There are some problems with the jamming strobe indicator. First, it interrupts normal video in some radars. Second, inverse or sidelobe

jamming can cause erroneous strobes. Finally, the jam strobe does not react to unmodulated CW or low-duty-cycle jamming for some radar systems.

f. The jamming indicator lamp, located on the operator console, is used on radars with automatic noise leveling (ANL) to counter continuous transmitted jamming. The lamp alerts the operator to the presence of jamming, and the ANL is manually shut off. This action allows the operator to determine jammer bearing.

g. Clean strobe generation (CSG) counters any transmitted jamming by using the sidelobe blanking circuits of a radar. An azimuth strobe appears when the jamming level in the main antenna exceeds the jamming level in the sidelobe auxiliary antennas. The operator is alerted to the presence and bearing of a jamming source, even with a constant false alarm radar (CFAR) receiver.

h. Jamming attenuation (JAM ATTEN) counters both clutter and any type of jamming. Receiver gain is reduced to avoid receiver saturation by inserting an attenuator pad that enables the operator to recognize presence, type, and bearing of a jamming source. When using JAM ATTEN, however, the reduced gain may cause loss of targets, even in non-jammed sectors. Also, any improvement in signal-to-jam ratio is not possible.

i. Receiver manual IF gain (MAN GAIN or IF GAIN) also counters clutter and jamming. Receiver gain is reduced to avoid jamming saturation by manually reducing stage gain, allowing the operator to identify jammer presence, type, and bearing. When using IF GAIN, the reduced gain may cause a loss of targets, even in non-jammed sectors. Also, any improvement in signal-to-jam ratio is not possible.

5. OVERPOWERING THE JAMMING SIGNAL

Following are some EP techniques a radar system can employ to overpower jamming and reduce the jamming-to-signal (J/S) ratio to less than one.

a. Burnthrough counters most types of transmitted jamming. Energy in the target pulse is raised by increasing the peak power, that is, the PRF or pulse width, or by increasing the time the radar illuminates the target by reducing the scan rate or scan angle. Some radars have modes in which the radar concentrates its power in narrow azimuth and elevation sectors about the suspected target position. However, burnthrough can degrade general radar performance by overloading the receiver if a large radar cross section target is detected. High power may impact radar operation in clutter or dense chaff environments.

b. Narrowband long pulse (NBLP or NLP) counters most types of transmitted jamming by using a high-energy long pulse. The signal uses a narrowband receiver for reception, and increases detection range for targets in jamming and

in the clear. Simultaneously, it reduces resolution, which causes poor radar performance in chaff and clutter.

6. PULSE DURATION DISCRIMINATION

The following are some radar counter-countermeasures that use pulse duration to discriminate between radar and jamming signals.

a. The fast time constant (FTC) is used to counter chaff, clutter, and narrowband jamming. A video circuit provides low frequency attenuation to reject carrier and low frequency modulation jamming. FTC passes normal radar pulse lengths with little attenuation, but causes some loss of receiver sensitivity.

b. Pulse width discrimination (PWD), clutter eliminate (CE), and wide pulse blanking (WPB) are designed to counter chaff, most types of jamming, EMI, and some types of deception jammers. A video coincidence gate, involving a delay line matched to the expected signal duration, senses if a return is the proper pulse width. PWD, CE, and WPB provide an enabling path for qualified signals. However, weak signals may be lost in the signal processing.

c. Pulse expansion-compression (PC) is used to counter most types of noise jamming and some types of deception jamming. An expanded pulse is coded for transmission. This expanded pulse is transmitted and decoded on return. Echo responses are then compressed in a decoding process. This expansion/compression is equivalent to NLP, which provides longer detection ranges, and wideband short pulse, which provides increased resolution. Using PC is not without problems. Unwanted residues may cause loss of weak targets. Additionally, range error proportional to the Doppler shift, or radial velocity, affects the accuracy of the PC.

7. ANGLE DISCRIMINATION

The following techniques use angle discrimination to distinguish between radar returns and jamming signals.

a. Sidelobe blanking (SLB) and sidelobe cancellation (SLC) are types of sidelobe suppression (SLS) used to counter sidelobe response to chaff, clutter, transmitted jamming, sidelobe jamming, and deception jamming. An auxiliary antenna approximates the pattern and gain of sidelobes of the main antenna and produces a signal for comparison with the signal received in the main antenna. If the signal in the auxiliary antenna is greater, the signal in the main antenna channel is blanked. This permits bearings to be obtained on a jamming source and rejects sidelobe jamming. SLB is useful only for determining the bearing to the jamming source.

b. Antenna manual positioning, antenna traverse and elevation angle offset, antenna jog, and antenna slow scan are EP techniques used to counter main

beam and sidelobe jamming and deception. These techniques are designed to increase the antenna scans across the jammed sector to increase the blip-scan ratio. These techniques increase the number of pulses integrated, as well as the operators' sorting capability.

8. BANDWIDTH DISCRIMINATION

The following are EP techniques that use bandwidth to distinguish between radar jamming and target returns.

a. Dicke-fix (DF) counters wideband and fast-swept jamming. A wideband limiter amplifies without ringing and clips all pulses down to a common level. Then an amplifier increases narrowband target echo signals more than the wideband jamming. There are some problems associated with DF. Any jamming entering the wideband limiter can "capture" the limiters and cause poor receiver sensitivity. Targets may be suppressed without the operator knowing that jamming is present. Resolution and target range is reduced, even in a clear environment. Finally, DF is ineffective, even harmful, when the jamming bandwidth is near the bandwidth of the desired echo signal.

b. Transmitter pulse lengthening (TPL) counters wideband and fast-swept jamming. TPL concentrates power into a narrow band about the carrier frequency by lengthening the transmitting pulse. While this allows use of a narrowband receiver, it impairs resolution, causing poor chaff and clutter performance.

c. Transmitter pulse shaping (TPS) counters wideband and fast-swept jamming. The sideband range is limited by shaping the transmitted pulse. This allows use of a narrowband receiver, but impairs resolution, causing poor performance in chaff and clutter.

d. Narrowband pulse limiting (NBLP or NLP) is a form of transmitter pulse lengthening that counters wideband jamming and fast-swept jamming. NBLP concentrates power into a narrow band in the carrier frequency by lengthening the transmitting pulse. This allows use of a narrowband receiver, but impairs resolution, causing poor chaff and clutter performance.

e. The fast time constant (FTC) is used to counter chaff, clutter, and narrowband jamming. A video circuit provides low frequency attenuation to reject carrier and low frequency modulation of jamming. FTC passes normal radar pulse lengths with little attenuation, but causes some loss of receiver sensitivity.

f. High video pass (HVP) is used to counter chaff, clutter, and narrowband jamming. It is similar to FTC. A video circuit provides low frequency attenuation to reject carrier and low frequency modulation jamming. HVP passes only the leading edge of the received pulses. HVP can cause some loss of receiver sensitivity.

g. The wideband short pulse (WSP) counters chaff, clutter, and narrowband jamming by transmitting a short pulse and using a wideband receiver for reception. Echo resolution and accuracy are improved, and performance against narrowband jamming is enhanced. However, the maximum detection range is decreased and system vulnerability to wideband jamming is increased.

h. Narrowband limiting (NBL) counters chaff, clutter and narrowband jamming. A narrowband filter positioned in the front of the amplifier section allows only the target signal bandwidth to enter the limiter, reducing wideband and off-frequency jamming. The limiter clips all signals and noise to a common level. This technique is useful only when followed by pulse compression or other "decode" techniques. NBL is not effective against wideband jammers capable of causing "ringing" of the NBL bandpass filters. Targets may be suppressed and lost without the operator knowing that jamming is present. Finally, target detection and resolution are poor.

9. DOPPLER DISCRIMINATION

EP techniques that use Doppler frequency discrimination between radar and jamming signals to negate jamming effectiveness include the following:

a. Moving target indication (MTI) is used to counter chaff and clutter. The phase of returned target echoes is compared on a pulse-to-pulse basis. Those with no phase change (no change in radial velocity) are cancelled using a delay-line canceler. Sensitivity using MTI is poor for weak targets, even in the clear. Also, it is blind to targets that have a Doppler frequency that is equal to a multiple of the radar PRF, unless PRF stagger is used. Finally, limited dynamic range does not allow full cancellation of strong clutter echoes.

b. Compensated coherent MTI, also known as compensated COHO MTI, counters chaff and clutter by comparing the phase of returned target echoes on a pulse-to-pulse basis. Those pulses with no phase change, that is, no change in radial velocity, are cancelled. Corrections to the coherent oscillator are applied to compensate for motion of the platform and radar antenna. Sensitivity is poor for weak targets, even in the clear, and it is blind to targets that have a Doppler frequency equal to, or a multiple of, the radar PRF, unless PRF stagger is used.

10. TIME DISCRIMINATION

The following EP techniques use time discrimination between radar and jamming signals to negate jamming effectiveness.

a. Video integration (VINT) and integrate-multiply (INT-MULT) counter any form of transmitted jamming not synchronous with radar PRF. The video continuously circulates through a delay line, delaying signals exactly one pulse recurrence time (PRT), then combines them with signals from the next PRT. Synchronous target signals add together to increase video output, but noise and

random pulses are suppressed. Unless MTI and FTC are employed, VINT and INT-MULT will enhance chaff, clutter, and jamming along with target returns. Also, feedback control, or loop gain, must be carefully adjusted for optimum results.

b. PRF stagger and jitter are EP techniques designed to counter quasi-synchronous jamming, EMI, MTI blind-speeds, "second trip" echoes, and repeater jammers simulating "closer-than-real" targets. The transmitter pulse interval is varied to break up synchronous patterns. Received signals must be "de-staggered" for use with MTI or integration. When using any of these techniques, video de-stagger balance must be accurate or "double video" occurs. These techniques are not effective against exact synchronous deception jamming.

c. Pulse-to-pulse correlation (PPC) is used to counter slow-swept blinking or pulsed jamming not synchronized with radar PRF, and some types of deception jammers. To be displayed, target video must exceed a threshold voltage for two successive pulses. The technique avoids displaying non-synchronous jamming and EMI, but is not effective against synchronous jamming, and may cause weak targets to be missed.

d. Beam-to-beam correlation (BBC) is used to counter slow-swept blinking or pulse jamming not synchronous with radar PRF and some types of repeaters and transponders. It is used in three-dimensional frequency scanning or frequency agile radars. To be displayed, the target return echo signal must exceed a threshold value in two adjacent antenna beams. BBC is not effective against synchronous jamming. It may also cause weak targets to be missed.

e. Single beam blanking (SBB) is used to counter slow-swept, blinking, pulsed jamming, and narrowband jamming. It is used in three-dimensional frequency scanning or frequency agile radars to avoid displaying vertical beams containing jamming. A blanking pulse is generated for vertical beams containing jamming so that they are not displayed. The technique is not used on RHI video. Also, it can cause a loss of targets at the jammed elevation angle on the PPI display.

11. SUMMARY

This chapter has discussed some of the most widely employed EP techniques designed to counter radar jamming. The capabilities of the individual radar operator were not discussed. However, the radar operator is as important as the EP techniques designed for the radar system. Many of the most effective EP techniques are designed to ease operator interpretation of the radar display. In the chess game of EW, the capabilities of individual radar operators can be as important as the sophisticated EP techniques in determining the final outcome.

CHAPTER 17. RADAR WARNING RECEIVER (RWR) BASIC OPERATIONS AND GEOLOCATION TECHNIQUES

1. INTRODUCTION

Radar surveillance and radar-directed weapons represent the biggest threat to aircraft survival on the modern battlefield. The first step in countering these threat systems is to provide the pilot or crew with timely information on the signal environment. The radar warning receiver (RWR) is designed to provide this vital information to the pilot. The RWR system is an example of an electronic warfare support (ES) system. The primary purpose of an RWR system is to provide a depiction of the electronic order of battle (EOB) that can have an immediate impact on aircraft survival. Though the RWR system is complex, the basic operations of the various components are straightforward. A step above RWR systems is threat geolocation. While an RWR provides the EOB for a single aircraft, threat geolocation systems can provide accurate threat location data for numerous aircraft over an entire region. Threat location data is used for aircraft threat avoidance and, more common today, the preemptive attacking of enemy radar sites. This chapter will discuss the functions of the various components of a RWR system including the antennas, receiver/amplifiers, signal processor, emitter identification (EID) tables, RWR scope, RWR audio, and limitations to RWR systems. This chapter will then go on to discuss three of the methods used to geolocate radar threat systems.

2. RWR ANTENNAS

Antennas are designed to receive radar pulses from threat radar systems. Factors that impact the operation of the RWR antennas include location, pattern, sensitivity, and polarization.

a. The physical location of the RWR antennas on the aircraft can affect its ability to detect a radar signal. Antennas are arranged to cover a predetermined area of horizontal and vertical space around the aircraft (Figure 17-1).

b. The antennas and their patterns play an essential part in displaying the spatial relationship of a threat radar to the aircraft. The antenna patterns are the areas, or “footprints,” that the antennas are specifically designed to cover. These footprints are directly affected by the relative position of the antennas to the threat systems. This is because the signal processor measures and compares signal strength from all the aircraft antennas to compute threat signal location relative to the aircraft. This relative location is then presented on the RWR scope display. Aircraft movement and maneuvering shifts these relative positions during flight and can distort the true threat position on the RWR scope. Precise position determination is not possible with most RWRs.

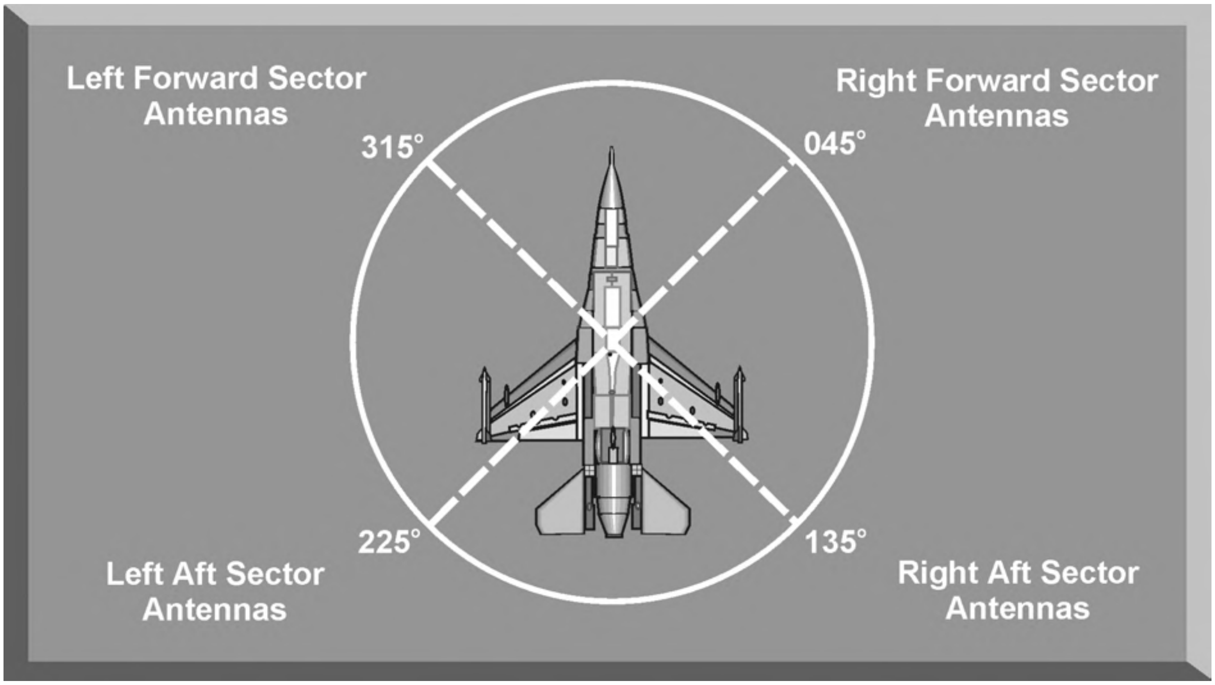
Left Forward Sector Antennas
Right Forward Sector Antennas
315°
045°
225°
135°
Left Aft Sector Antennas
Right Aft Sector Antennas

Figure 17-1. Radar Warning Receiver (RWR) Antennas

c. The sensitivity of an RWR antenna directly affects its ability to detect a radar signal. The more sensitive the antenna, the further it can detect a signal. The sensitivity of a system and its ability to intercept a radar signal is usually expressed in decibels relative to milliwatts or dBm units. A 10 dBm change in sensitivity can result in a 25 nm range difference in target detection. In general, sensitivity levels of -50 to -60 dBms are required to detect signals at long ranges (Figure 17-2).

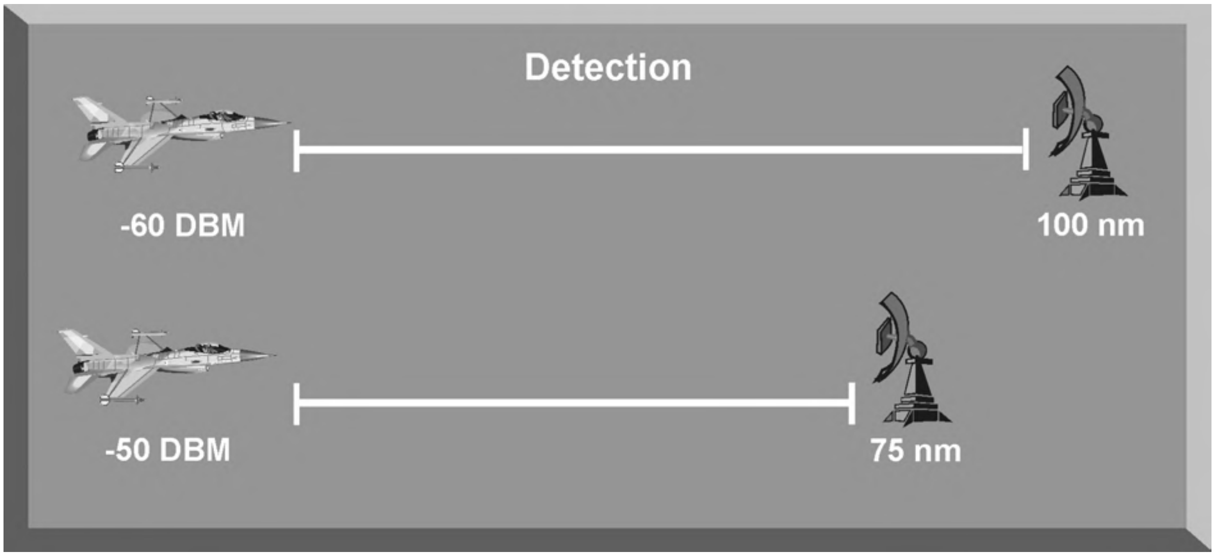
Detection
-60 DBM
100 nm
-50 DBM
75 nm

Figure 17-2. RWR Antenna Sensitivity

d. Another factor affecting antenna detection range is the polarization of the antennas. If the polarization of the RWR antenna and the threat system antenna are mismatched, or cross-polarized, initial detection of a threat signal could be delayed until the aircraft is within the lethal range of a threat system. In this situation, the aircraft could be engaged with minimal warning (Figure 17-3).

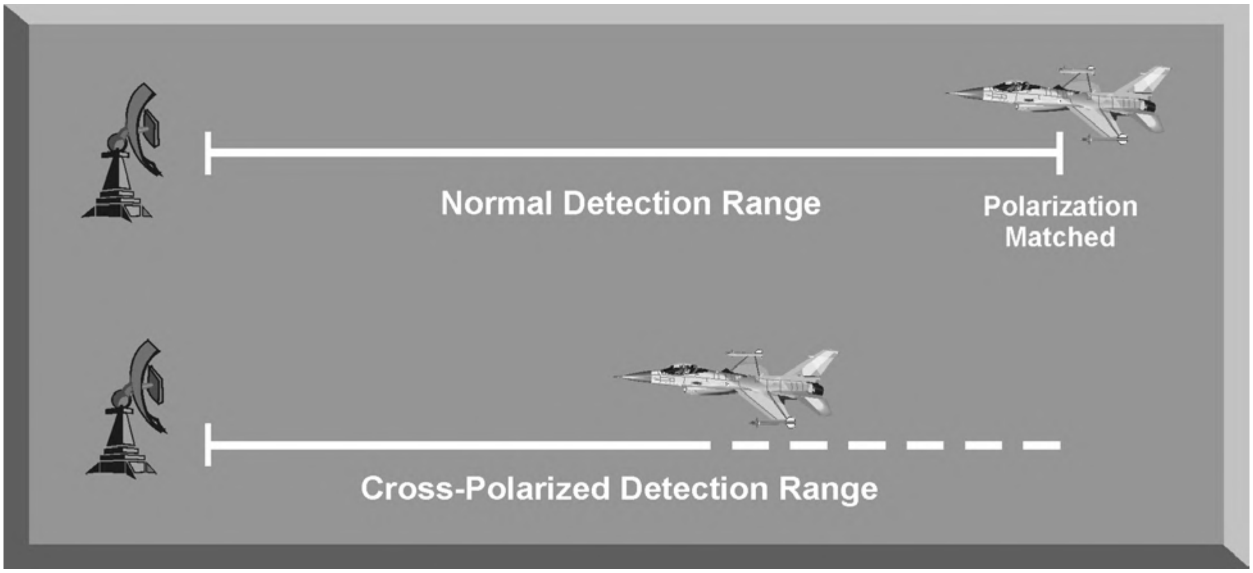

Figure 17-3. RWR Antenna Polarization

3. RWR RECEIVER/AMPLIFIERS

The RWR receiver/amplifier section processes the radar signals from the antennas. Most RWR systems use frequency bands to differentiate signals. Nominal band designations are summarized in Table 17-1. There are two types of receivers currently used in RWR systems: the crystal video receiver and the superheterodyne receiver.

Table 17-1. RWR Frequency Band Designators

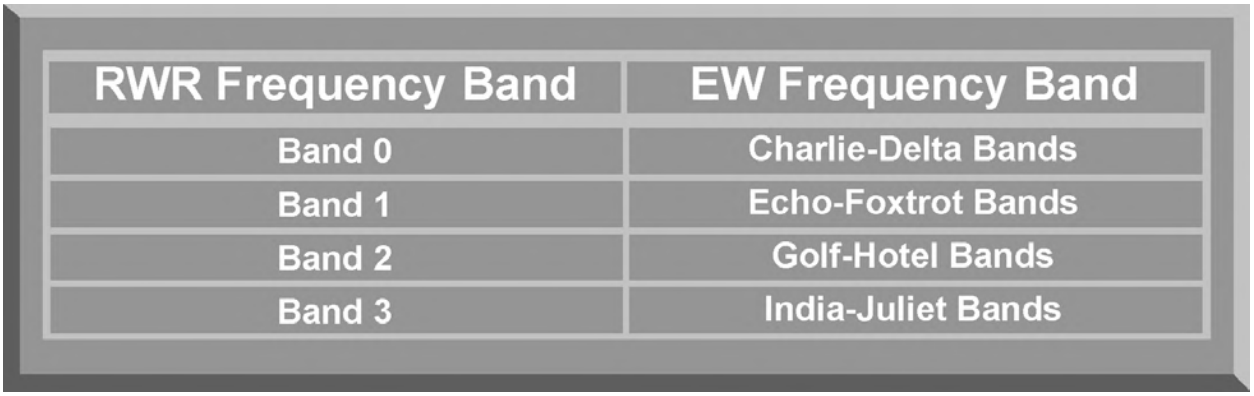

RWR Frequency Band	EW Frequency Band
Band 0	Charlie-Delta Bands
Band 1	Echo-Foxtrot Bands
Band 2	Golf-Hotel Bands
Band 3	India-Juliet Bands

a. A crystal video receiver (CVR) is the simplest type of microwave receiver. It is used primarily for detection of pulse radar signals in the 2 to 18 GHz band. A CVR used in a radar warning receiver incorporates crystal detectors for each designated frequency band. A pulse radar signal is detected by the antenna and passed to the multiplexer. The multiplexer divides the received radar signals by frequency band and sends the signals to the appropriate band channel. The RF amplifier boosts the radar signal and passes it to the crystal detector (Figure 17-4).

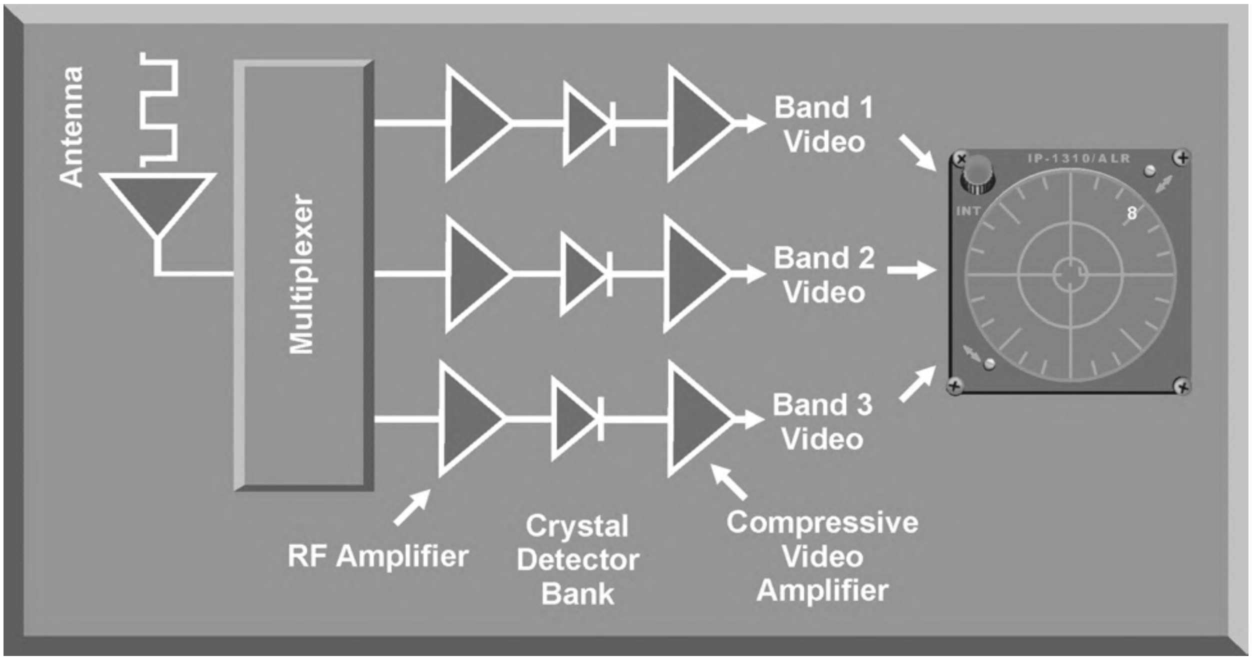

Figure 17-4. Crystal Video RWR Receiver

(1) The crystal detector is an RF diode, which converts the RF signal into a video signal. The voltage level of the output video signal is dependent only on the amplitude of the input signal and not on the frequency or phase. The sensitivity of a CVR is limited by the sensitivity of these crystal detectors. The sensitivity of crystal detectors currently available is generally adequate to detect main beam radiation from most threat radars. The video output of the crystal detectors is amplified by a high-grain compressive video amplifier and sent to the RWR scope for display.

(2) A CVR is extremely fast, sensitive, and covers a wide frequency range. These characteristics coupled with low cost and small size make CVRs ideal for use in radar warning receivers. The primary disadvantage of a CVR is that it is indiscriminate in reception and can be saturated in a dense signal environment. Multiple signals in the same band can cause amplitude distortion which can mask key threat signals.

b. A superheterodyne RWR receiver uses a pre-selector filter, mixer, and a local oscillator to translate the received signal to a lower intermediate frequency (IF). This lower IF allows the receiver to amplify and filter the received signal to provide greater sensitivity and frequency selectivity than a CVR (Figure 17-5).

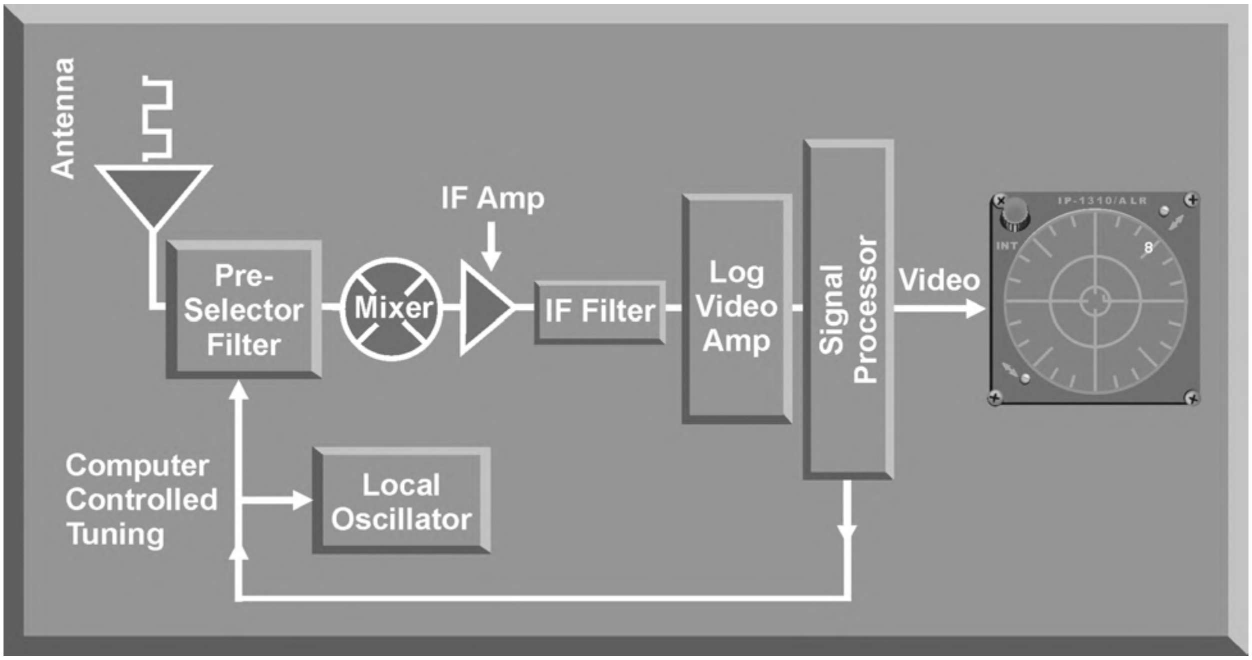

Figure 17-5. Superheterodyne RWR Receiver

(1) Superheterodyne RWR receivers use special scanning techniques controlled by the signal processor to tune the pre-selector filter and the local oscillator to rapidly scan selected threat system frequency bands. If the receiver detects activity in any of these bands, the scan stops to allow the processor to analyze the detected signals. The signal is combined with the local oscillator signal to lower the frequency to the IF. This signal is amplified, filtered, and amplified again before it reaches the signal processor. The signal processor classifies the threat and displays the proper threat symbol to the pilot. This entire process is accomplished in a matter of microseconds.

(2) The scanning superheterodyne receiver has important features that make it effective for RWR system application. It has excellent sensitivity and selectivity. It also has good frequency resolution. These features give the superheterodyne receiver a very low false alarm rate. The major disadvantage of the scanning superheterodyne receiver is its limited capability to receive signals from threat systems employing scanning antenna patterns. This limitation can be overcome with specific computer-controlled tuning to look for these threat signals.

4. SIGNAL PROCESSOR

The signal processor is the heart of the radar warning receiver. The signal processor is also known as the digital processor or analysis processor in different RWR systems. Its primary functions are to process numerous complex radar signals and identify, among the thousands of similar signals, those generated by lethal threat systems. The signal processor accomplishes this task continuously over the duration of the mission and displays the identified threat system to the aircrew almost instantly.

a. Signal processing begins when RF energy strikes the receive antennas on the aircraft. The received signals are then boosted in strength by intermediate amplifiers or antenna receivers. These amplified signals are then sent to the signal processor, or digital processor, where they are assigned a track file for reference to other signal characteristics. Data in these files is compared to those in the emitter identification (EID) table to process the signal for identification. Once identification is complete, a video and, if necessary, an audio signal, is sent to the cockpit display. The audio and video signals alert the aircrew to the electronic environment around the aircraft. This whole process takes less than a microsecond (Figure 17-6).

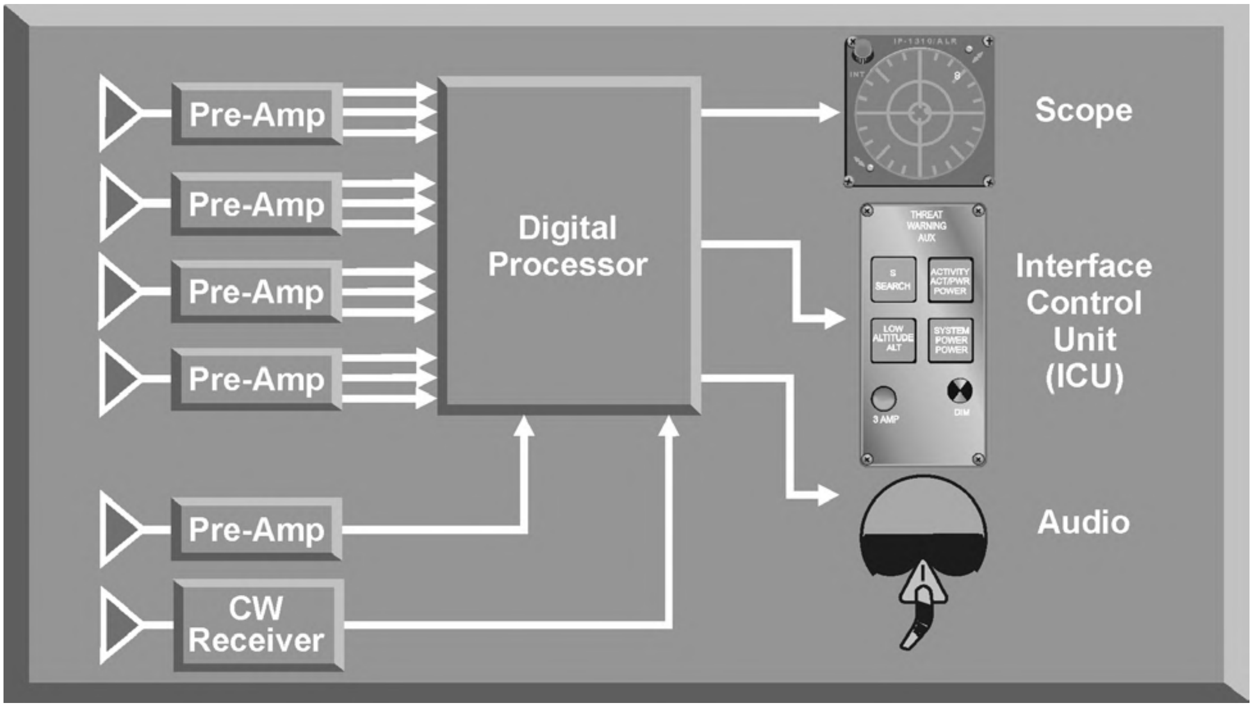
Pre-Amp
Pre-Amp
Pre-Amp
Pre-Amp
Pre-Amp
CW Receiver
Digital Processor
IP-1310/ALR
INT
8
Scope
THREAT WARNING AUX
S SEARCH
ACTIVITY ACT/PWR POWER
LOW ALTITUDE ALT
SYSTEM POWER POWER
3 AMP
DIM
Interface Control Unit (ICU)
Audio

Figure 17-6. RWR System

b. Since many signals may be present, the amplifier detectors boost the signal strength, and also tag each signal by certain characteristics such as its time of arrival, direction of arrival, and/or frequency. These signals, along with their respective tags, are sent to the signal processor for further processing and

identification. The signal processor then makes a track file for each signal it receives from the amplifier detector.

c. The signal processor classifies each received signal and corresponding track file by its unique radar signal characteristics. Identifying characteristics used by a signal processor can include radio frequency, pulse width, pulse repetition frequency, EP techniques, and more. Characteristics of one signal may be identical to characteristics from different signals, while certain other characteristics can be as unique as a human fingerprint. The signal processor uses these primary characteristics to identify specific signals. When the primary characteristics of two or more signals are similar, the signal processor uses additional signal characteristics to resolve any confusion between two or more signals.

d. The signal processor ranks the track files based on priorities determined from tests it conducts on the signal characteristics and the threat priorities contained in the EID tables. It then quickly processes signals belonging to lethal threats before it processes signals belonging to non-lethal threats. For example, three signals enter the processor together and separate track files are established for each signal (Figure 17-7). A test on the first characteristic discriminant, frequency, will delay the further processing of Signal 3, since no lethal threat systems operate at a frequency less than 2000 megahertz. A further test on the remaining prioritized signals may eliminate Signal 2 as a threat system signal, leaving more processor time for the identification of the threat system which generated Signal 1. These tests do not stop the processor from attempting to identify all received signals. The signal processor merely delays the identifying sequence until all high priority signals have been processed.

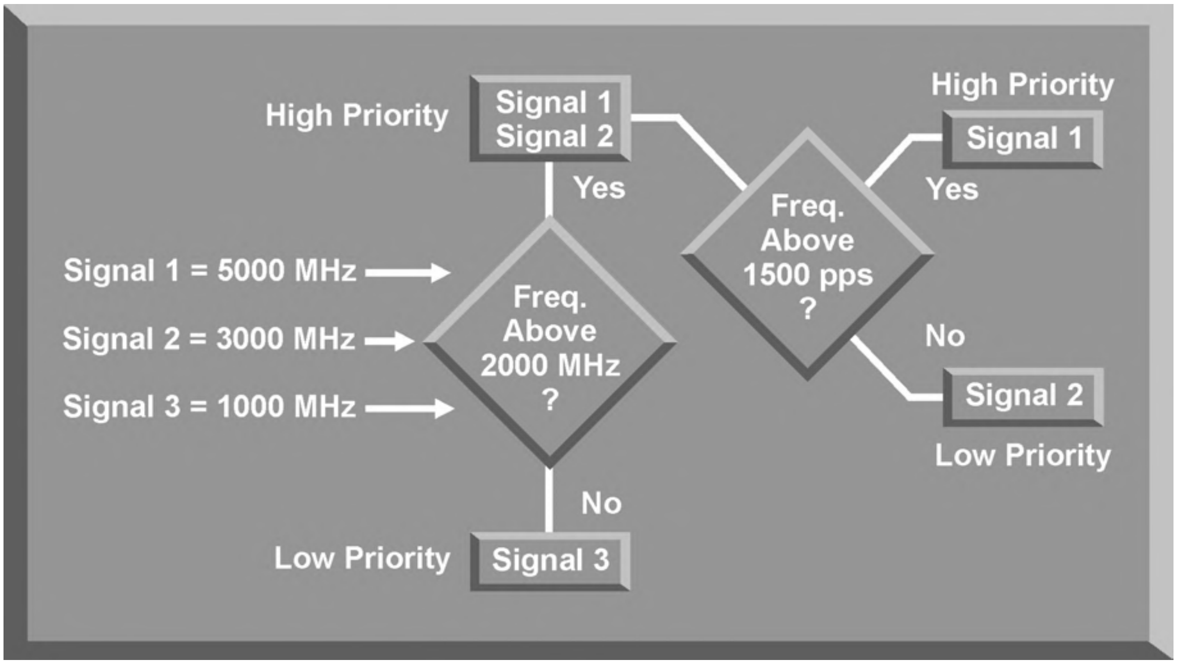

Figure 17-7. Signal Processor Signal Priority

5. EMITTER IDENTIFICATION (EID) TABLES

The signal characteristics in each track file are filled with processed data, and are constantly updated based on the time of arrival and location of the received signals. In addition, the track files are constantly compared to the EID table installed in the signal processor's computer memory. The EID table is a predefined table of radar characteristics associated with known radar systems (Figure 17-8). It is created from information gathered from electronic warfare support (ES) assets and intelligence sources. This table can be changed and updated as necessary to reflect the most current radar characteristics available for the anticipated threats in the planned theater of conflict. Each RWR system has unique procedures to reprogram the signal processor and update the EID tables. Emergency reprogramming actions, such as would be taken if a new threat appears that is not part of the current EID, are called a Pacer Ware.

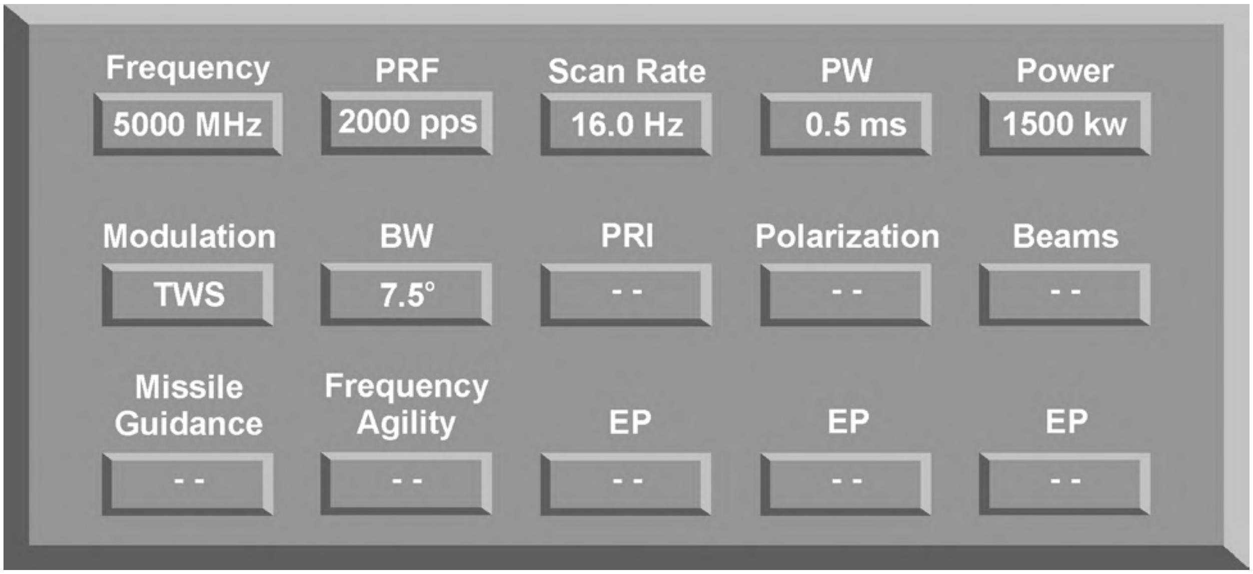

Frequency	PRF	Scan Rate	PW	Power
5000 MHz	2000 pps	16.0 Hz	0.5 ms	1500 kw
Modulation	BW	PRI	Polarization	Beams
TWS	7.5°	- -	- -	- -
Missile Guidance	Frequency Agility	EP	EP	EP
- -	- -	- -	- -	- -

Figure 17-8. Sample EID Table

6. RWR SCOPE DISPLAY

The signal processor continually compares signal characteristics in the track files with the data in the EID tables. Once the signal processor has determined that enough of the signal characteristics match the information in the EID tables, it generates and positions a video symbol on the RWR scope. The video symbol represents a specific threat, and each threat system has its own unique symbol. In addition, an audio tone is generated to alert the pilot. The signal processor also generates symbols and audio associated with specific threat system actions, including search, track, and missile launch. The position of the threat symbol on the RWR scope always represents the relative position of the threat in relation to the aircraft which is the center of the RWR scope. The signal processor compares the received signal strength in the different antennas to determine the proper location of the threat symbol. Figure 17-9 depicts a situation where the two

forward antennas receive equal signal strength therefore the signal processor places the symbol at the 12 o'clock position.

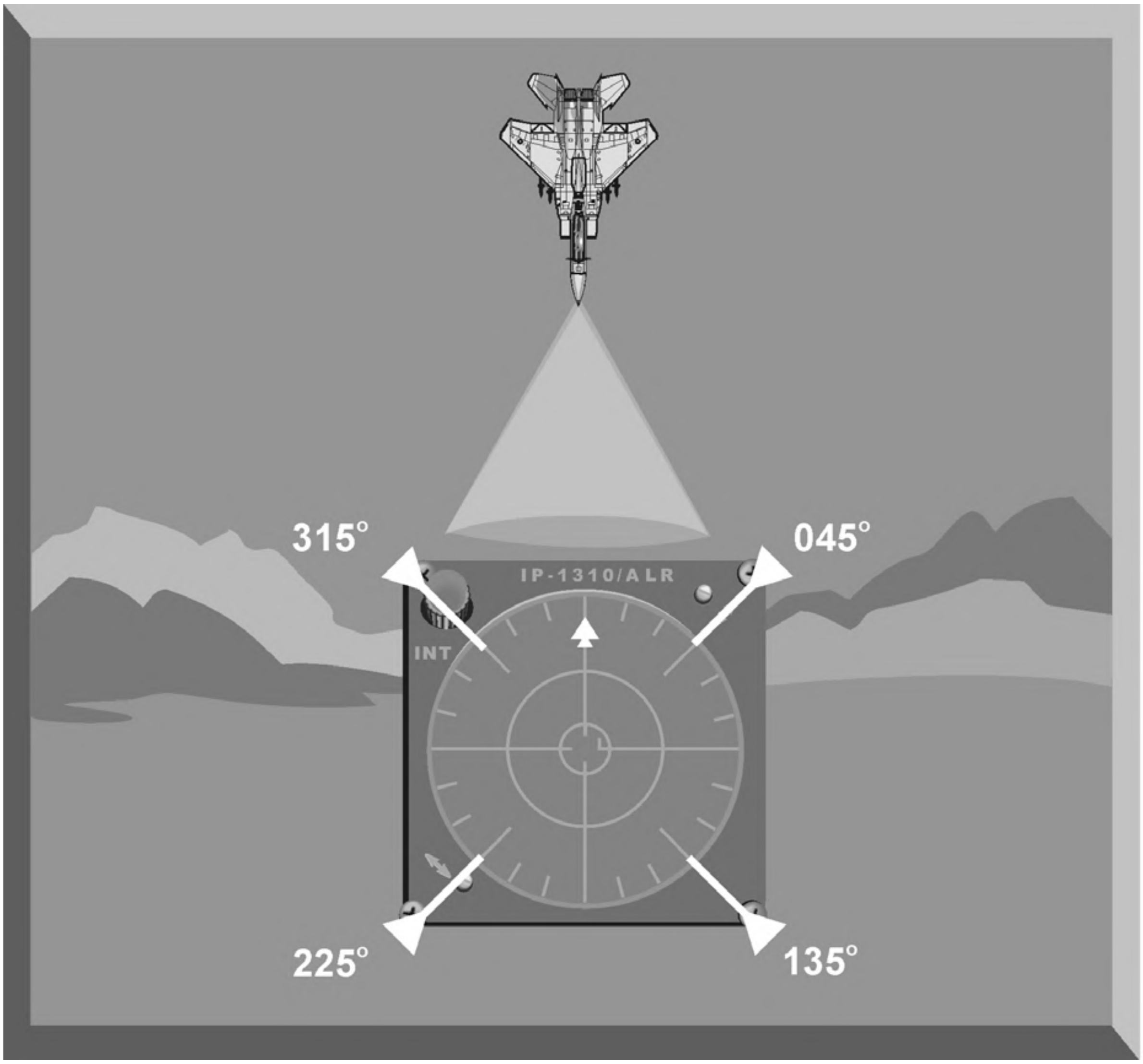

Figure 17-9. RWR Scope Azimuth Positioning

7. RWR AUDIO

In addition to generating threat symbols for each identified threat, the signal processor also generates threat audio. Threat audio first alerts the aircrew to the detection of a threat system. This RWR audio is generally referred to as "new guy" alert audio. The signal processor can also present constant audio from a selected threat. The aircrew controls this function through the interface control unit. The constant audio provided by an RWR system can be either "real" or synthetic. "Real" audio is normally based on the actual pulse repetition frequency (PRF) of the threat system radar whether the signal processor has identified it or not. Synthetic audio is based on the classification of the threat (SAM, AI, etc.) as

determined by the signal processor. The signal processor also generates a launch warning audio when the signal characteristics of the threat indicate a missile launch condition exists.

8. RWR INTERFACE CONTROL UNIT (ICU)

Every RWR system has some type of an ICU which provides the aircrew interface with the signal processor (Figure 17-10). The buttons on the ICU control specific functions of the signal processor. The ICU allows the aircrew to optimize the RWR system based on mission tactics. This optimization includes selecting appropriate priority lists based on ingress and egress tactics, controlling threat audio presentation, and determining the number and types of threats displayed. In addition, the ICU provides an additional visual indication of missile launch. All system test functions are controlled by the ICU to allow the aircrew to monitor the status of the RWR system.

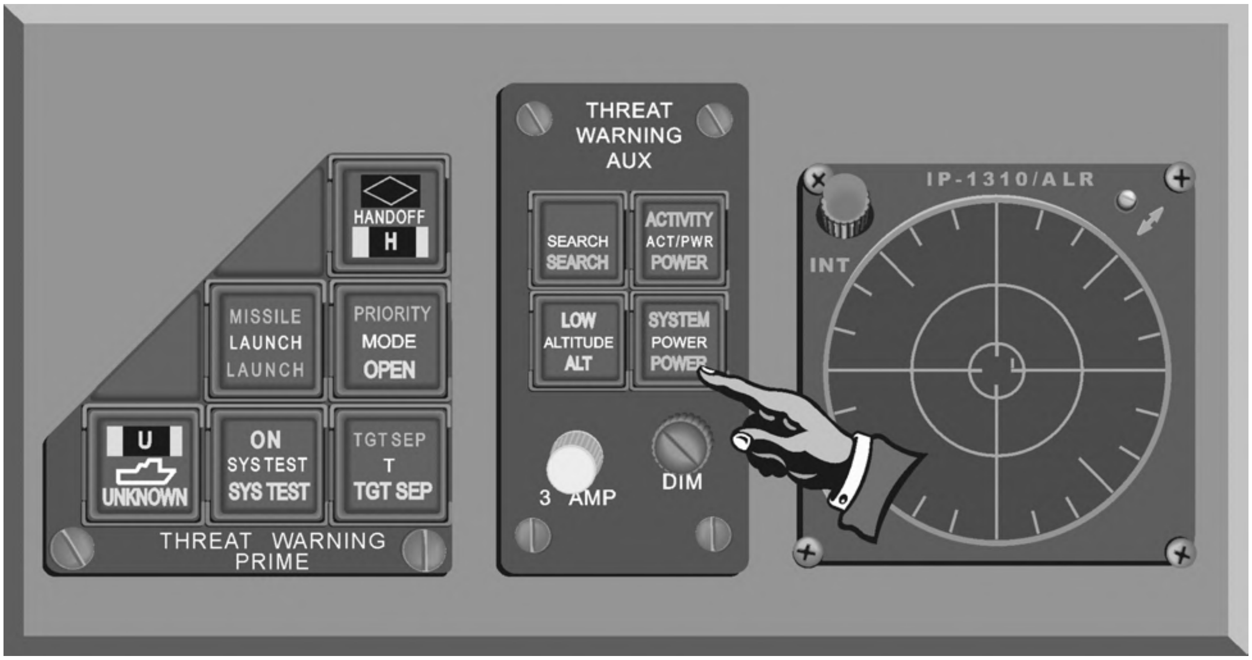
HANDOFF
H
MISSILE
LAUNCH
LAUNCH
PRIORITY
MODE
OPEN
U
UNKNOWN
ON
SYS TEST
SYS TEST
TGT SEP
T
TGT SEP
THREAT WARNING
PRIME
THREAT
WARNING
AUX
SEARCH
SEARCH
ACTIVITY
ACT/PWR
POWER
LOW
ALTITUDE
ALT
SYSTEM
POWER
POWER
3 AMP
DIM
IP-1310/ALR
INT

Figure 17-10. RWR Interface Control Unit

9. RWR LIMITATIONS

There are several limitations associated with all RWR systems. The most important limitations include ambiguities, impact of maneuvering, and electromagnetic interference (EMI).

a. The sheer number and diversity of radar systems associated with an enemy IADS greatly compound the problem of threat identification and warning for RWRs. Adding to this problem is the fact that many different threat systems use operating modes that are parametrically similar. When an RWR processes a

radar signal that has the same characteristics of a signal from a different system, an ambiguity may occur. An RWR ambiguity is defined as the display of more than one symbol for a specific threat signal. RWR ambiguities may occur from both enemy and friendly radar systems.

(1) Figure 17-11 depicts a number of friendly and threat signals that operate between 8000 and 10,000 megahertz. On any given combat mission, it is quite possible that the RWR will receive signals from one or more of these threat systems at the same time. If frequency is the only signal characteristic available for processing, the RWR will not be able to determine which system the signal represents. Since threat systems operating in this frequency range are potentially lethal systems, the signal processor will attempt to match the frequency with a threat system from the EID table.

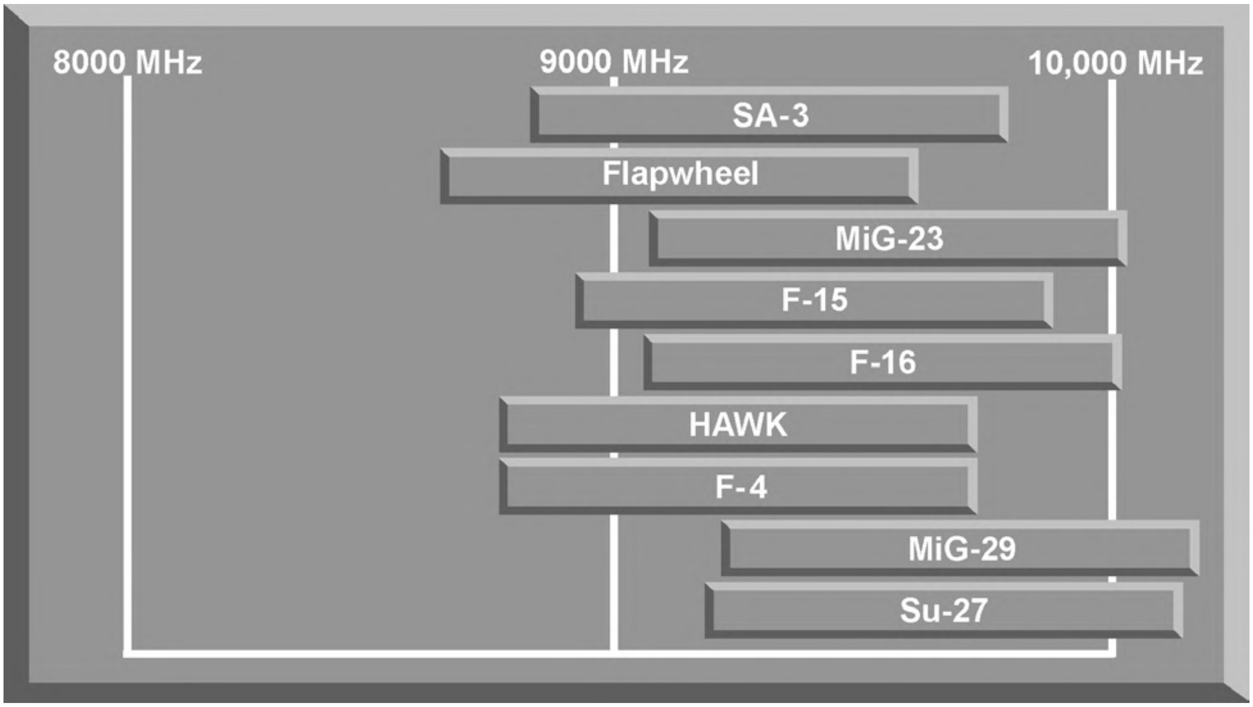

Figure 17-11. Signal Frequency Spectrum

(2) Matching partially processed signals to threat systems in the EID table may result in the wrong threat symbol being displayed, or numerous symbols being displayed on top of each other, making it difficult for the pilot to distinguish the exact threat. Incorrect threat symbology, or numerous combined symbols, are called RWR ambiguities.

b. RWRs are designed to provide accurate threat positioning information when the aircraft is flying straight and level. Most RWRs will also provide accurate threat positioning information when the aircraft is maneuvering up to certain limits of bank angle and turn rate. If aircraft maneuvering exceeds these

limits, RWR threat positioning data becomes unreliable. The two RWR limitations associated with aggressive maneuvering are inaccurate threat azimuth and multiple threat symbols.

(1) In Figure 17-12, the right forward and right aft RWR antennas detect a threat signal from a TTR. The left forward and left aft antennas are shielded by the aircraft and do not detect the signal. The signal processor determines the threat position, using the azimuth positioning algorithm, and displays the threat symbol at the 2 o'clock position.

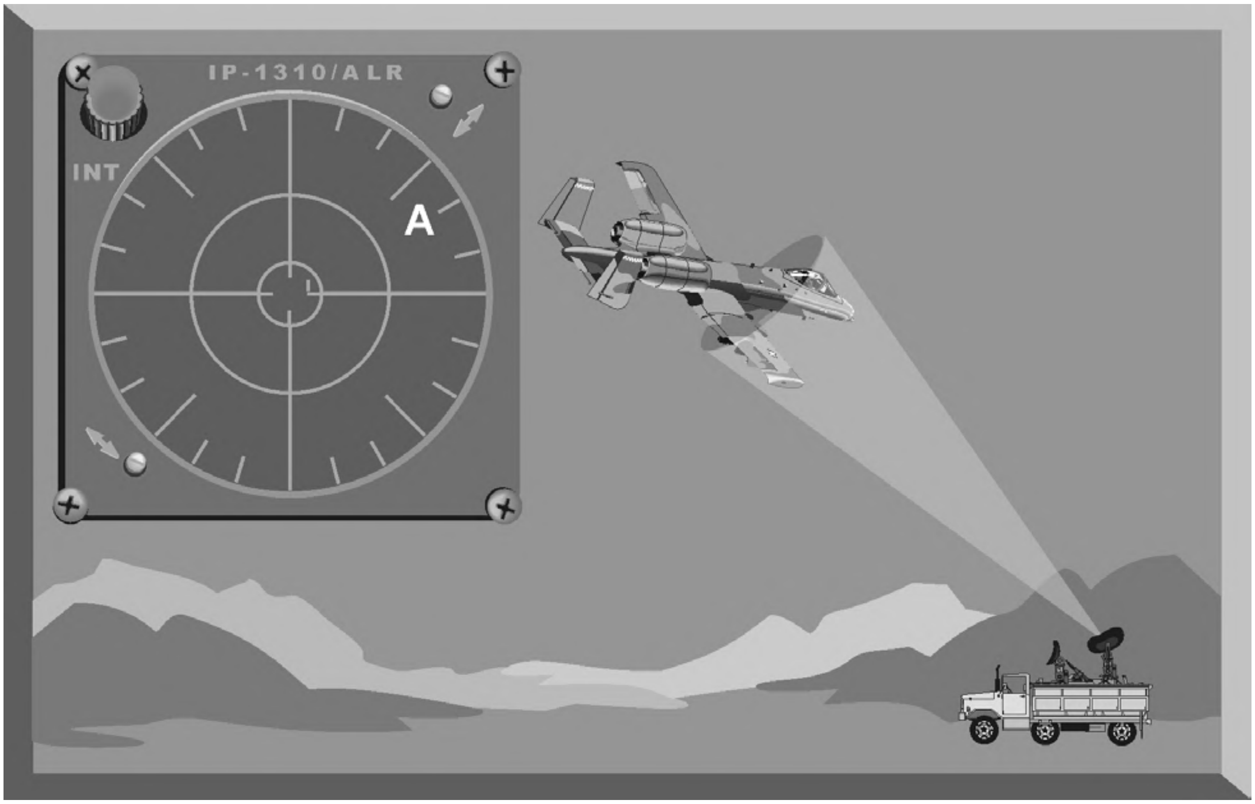

Figure 17-12. RWR Threat Azimuth Position Determination

(2) When the pilot maneuvers aggressively to put the threat symbol on the beam, he exceeds the RWR maneuvering limitations (Figure 17-13). Now all four RWR antennas detect the TTR signal. The signal strength in the right and left forward antennas is nearly equal. Based on this information, the signal processor displays the RWR symbol at the one o'clock position while the threat is actually at the three o'clock position.

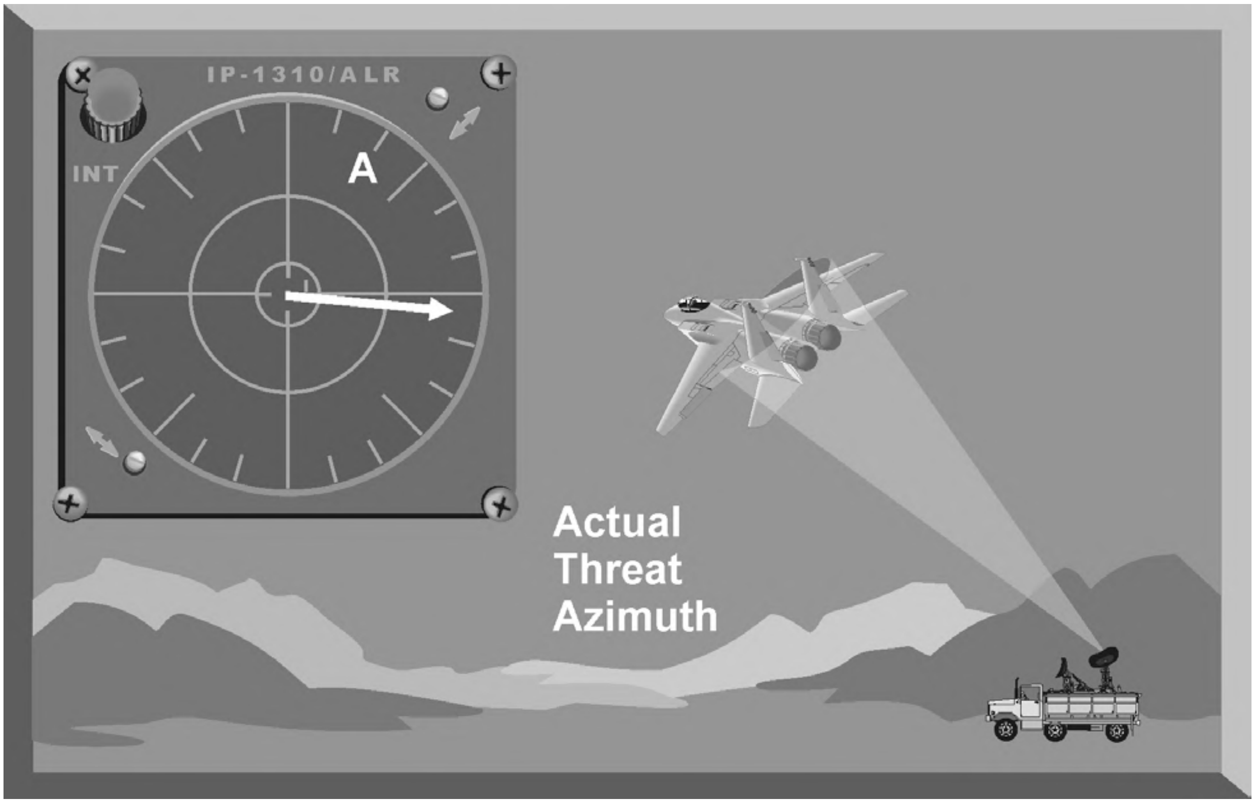

Figure 17-13. RWR Azimuth Error

(3) During aggressive maneuvering, the RWR signal processor may generate multiple symbols for a single threat emitter. In Figure 17-14, as the aircraft maneuvers, the relative azimuth of the threat position changes rapidly causing signal strength detected by each antenna to also change rapidly. The signal processor interprets these changing signals as new and different threat systems with the same signal characteristics. The number of "false" threat symbols displayed for a single threat is determined by the processing speed of the signal processor and a parameter called the symbol "age-out" time. Symbol age-out is the time, normally in seconds, that the RWR will continue to display a threat symbol after the signal processor has determined that the threat is no longer transmitting. The symbol age-out time is set for each threat in the EID tables.

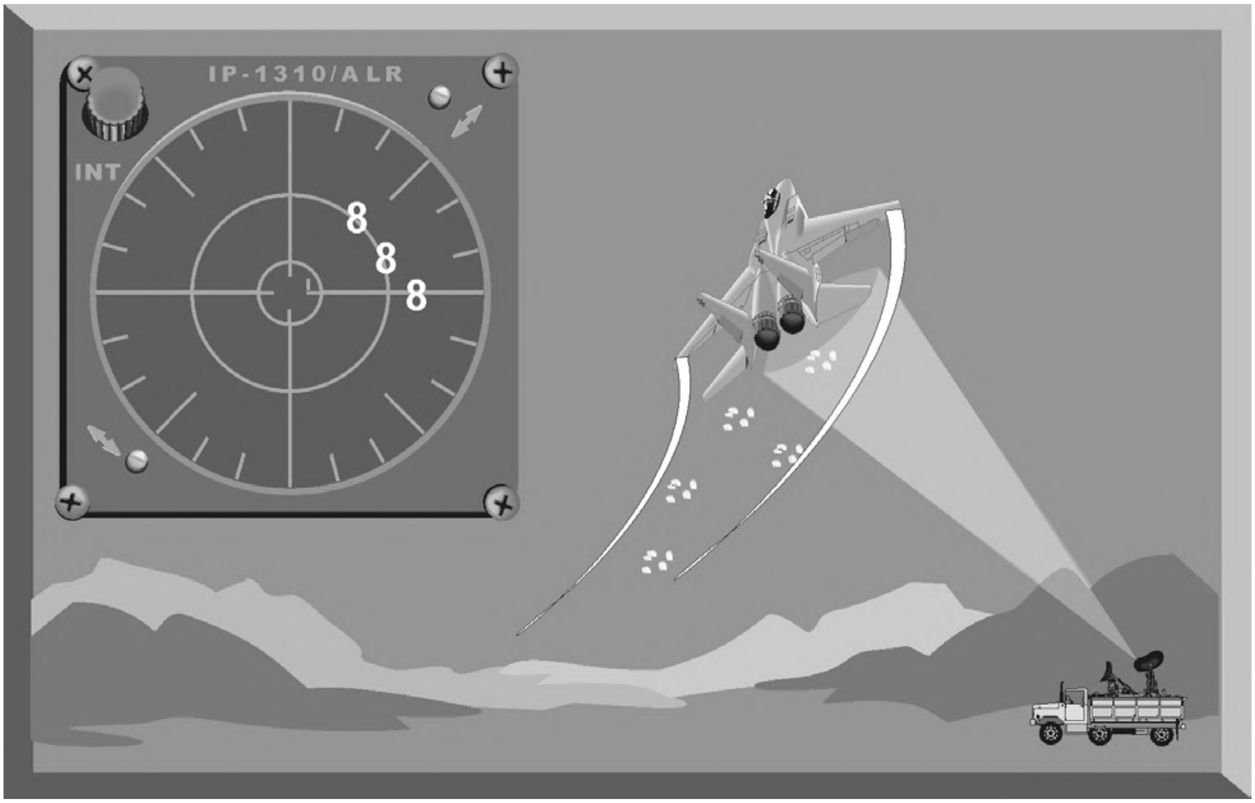

Figure 17-14. RWR Multiple Threat Symbols

c. Electromagnetic interference (EMI) is defined as any electromagnetic disturbance that interrupts, obstructs, or otherwise degrades or limits the effective performance of electronic systems. EMI can be induced intentionally, by way of jamming, or unintentionally as a result of spurious emissions and modulations. Certain RWR characteristics make them susceptible to EMI. Modern RWR systems are designed to receive and process signals in a wide frequency range, nominally 0.5 to 18 GHz, where most threats operate. This broad frequency coverage combined with the sensitive antennas make RWR systems susceptible to EMI. The primary source of EMI that impacts RWR operation is noise and deception jamming designed to counter enemy threat systems.

(1) High power noise jamming, such as that provided by a stand-off jamming aircraft, causes the RWR to raise the receiver threshold in the frequency band of the jamming. This effectively reduces the sensitivity of the RWR receiver and could delay the display of threat signals in that band. The reduction in sensitivity depends on the power that the jammer is transmitting, the beamwidth of the jamming beam, and the distance from the jammer to the aircraft. High power jamming may also generate multiple threat symbols on the RWR scope at the approximate azimuth to the jammer's position relative to the aircraft. Additionally, jamming from a wingman's self-protection system can generate multiple threat symbols and reduced sensitivity.

(2) A limitation of all RWR systems, related to EMI, is the problem of inaccurately identifying threat radar systems as friendly radar systems. This occurs when the parameters of a friendly radar are similar to the parameters of a threat radar system. The RWR will either display a threat radar symbol or an ambiguity associated with a threat radar. These RWR misidentifications are especially prevalent for AI radar systems.

(3) The impact of EMI on the operation of an RWR system depends on the signal environment. The pilot has little control over the number and diversity of friendly and enemy signals the RWR system must process. EMI is an unfortunate consequence of the reliance of modern military forces on operations in the electromagnetic spectrum. Aircrews should be keenly aware that EMI will impact RWR operation and be must be familiar with common RWR displays of EMI.

10. THREAT GEOLOCATION TECHNIQUES

The purpose of threat geolocation is to put a defined position, normally coordinates, on a threat radar. This information can be used to simply warn other aircraft about the threat or, if the coordinates are accurate enough, to allow for targeting and attack of the threat. Until recently, threat geolocation could only be performed by strategic assets and specialized tactical aircraft, specifically the F-4G Wild Weasel. Due to inherent time delays, data provided through strategic channels often did not apply to mobile threat systems. The mobile systems would relocate making the data obsolete. This section will discuss three techniques used to geolocate, also known as direction finding (DF), emitting radars that can be used by tactical assets to rapidly locate radar threat systems. The three are triangulation, interferometry, and time of arrival. All three techniques are heavily dependent upon the receiving aircraft's ability to accurately determine its present position, and the advent of GPS receivers has made this significantly easier.

a. Triangulation is the most basic form of DF available. It involves taking direction measurements from more than one source. The intersection of the azimuth measurements, called “lines of bearing”, is the likely location of the emitter (Figure 17-15). To be effective, the participating aircraft must have accurate data of their current positions when getting the lines of bearing.

(1) Triangulation can be carried out by multiple aircraft equipped with receiver equipment or by one aircraft over a period of time. The advantage of having multiple aircraft providing azimuth measurements is the increased angle-off and the speed of interception. In triangulation the best azimuth cuts are those that approach 90° angles. The speed of interception comes into play because threat emitters, knowing that DF operations are underway, attempt to transmit for as little time as possible. The disadvantage of multiple platforms is the communication required to ensure that all the platforms are measuring the same radar. In a dense radar environment this can be a difficult task.

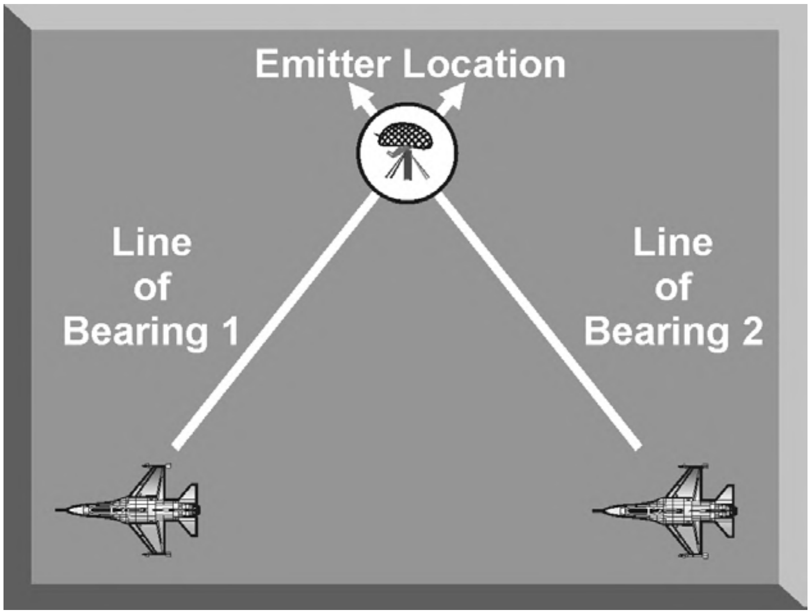

Figure 17-15. Triangulation

(2) Single aircraft triangulation eliminates the problem of signal coordination with other aircraft, but it also requires that the aircraft transit some distance to get multiple azimuth measurements. The accuracy of single ship DF operations is a function of the quality of the receiver equipment, distance away from the targeted radar, speed of the aircraft, and the amount of time that the targeted radar radiates.

b. The second technique is called interferometry. This technique is also known as phase interferometry, or phase difference of arrival. These systems operate by comparing the phase of a radar wave as it impacts two or more DF antennas; this phase difference is then used to compute an angle of arrival (AOA). For aircraft, the desire is to have these DF antennas on different parts of the same aircraft. Multiple AOA measurements are then used to provide the range and position of the threat.

(1) The key elements in an interferometer system are two antennas in fixed locations with matched receivers, a phase comparator, and a processor (Figure 17-16). An intermediate frequency output from each receiver is passed to a phase comparator, which measures the relative phase of the two signals. This relative phase position is passed to a processor, which calculates the AOA relative to the orientation of the two antennas (called the baseline). In most systems, the processor also accepts information about the orientation of the baseline (relative to true North or local horizontal) to determine the true azimuth or elevation angle to the emitter.

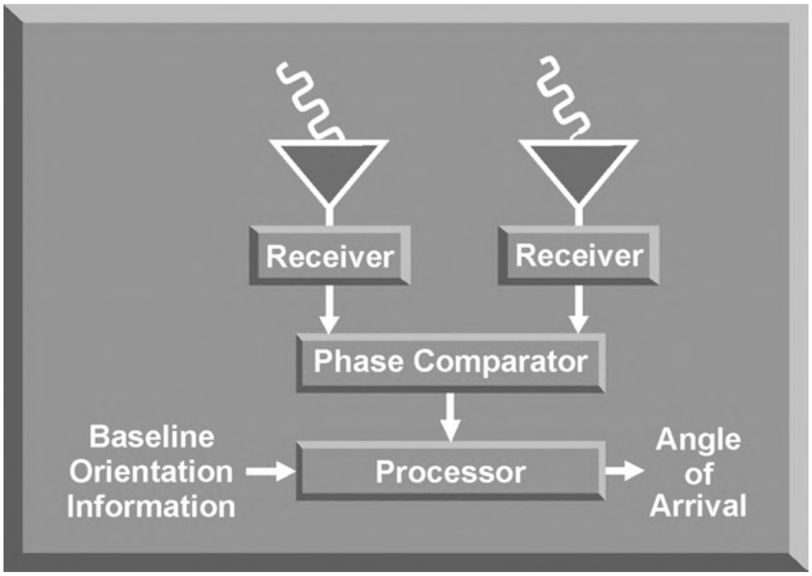
Receiver
Receiver
Phase Comparator
Baseline Orientation Information
Processor
Angle of Arrival

Figure 17-16. Interferometer Components

(2) To obtain rapid emitter locations some interferometer systems use multiple receive antennas in an array setup. This allows for simultaneous azimuth and elevation measurements to rapidly locate the emitter. An array allows for the mixing of long and short baselines in different patterns by selecting different pairs of antennas (Figure 17-17). The terms long baseline and short baseline are often used to designate the distance between the antenna elements in an interferometer system. Long baselines have the advantage of providing a quick and accurate location of the emitter, but they can suffer from ambiguities resulting from different wavefronts hitting the different antenna elements. In addition to the array depicted in Figure 17-17, a long baseline system could be created by using the existing RWR antennas on an aircraft, and supplementing these with a small short baseline system to compensate for the ambiguities.

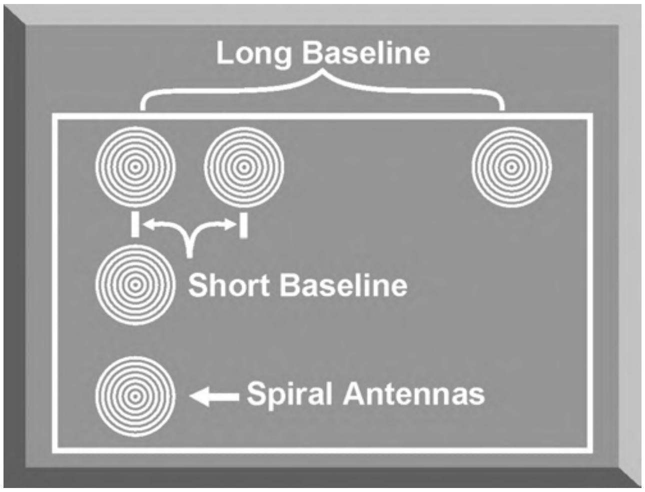
Long Baseline
Short Baseline
Spiral Antennas

Figure 17-17. Interferometer Array

c. Time of arrival (TOA) and time difference of arrival (TDOA) techniques are the final type of location techniques to be covered in this section. Both techniques are based around the fact that radar signals travel at approximately the speed of light. Both techniques, using the speed of light as a constant, solve for the distance that the emitter is away from the receiver using the equation distance equals rate multiplied by time. Because there is not any directional information, the equation represents the radius of a circle around the receiving antenna; multiple distance measurements taken from multiple receivers are then overlayed. The intersection of the circles is the position of the emitter. If only two receivers are used, a simple DF technique can solve the ambiguity of which intersection represents the emitter (Figure 17-18).

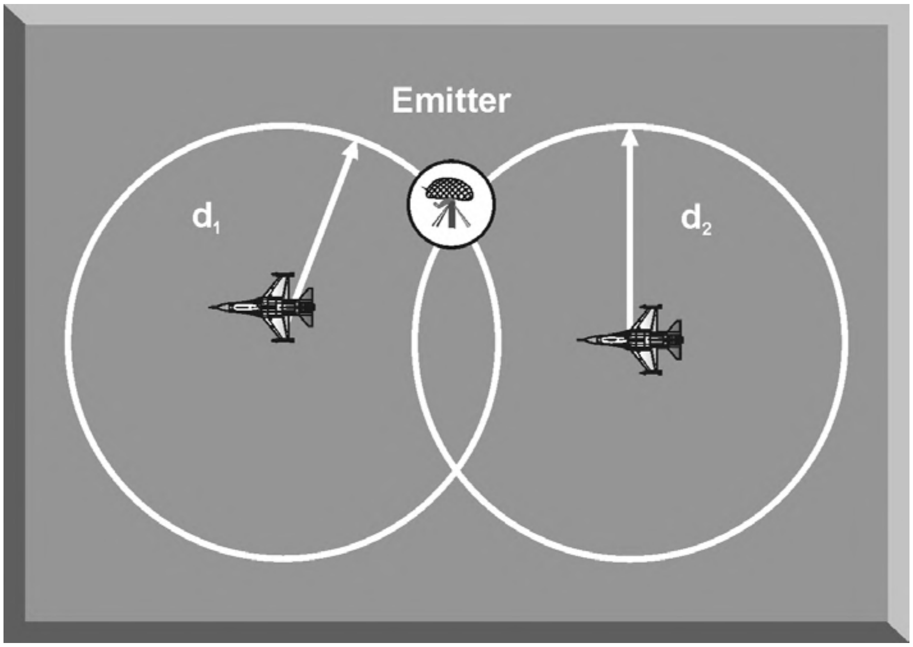

Figure 17-18. Time of Arrival Measurements

(1) TOA positioning is calculated by taking the time that a signal leaves a radar, measuring its arrival time at the receiver and mathematically solving for distance using the techniques described above. One of the primary challenges involved with TOA positioning is determining when the measured signal was transmitted, this requires either a very cooperative enemy or a radar signal with some type of exploitable time reference. Another challenge involves insuring that multiple receivers are timing the same signal. This is especially difficult in a threat intensive environment.

(2) TDOA is used when it is not possible to determine when a signal was transmitted. TDOA uses most of the same principles as TOA except that it must compensate for not knowing the time the signal was transmitted. It does this by using an extra receiver a known distance from the first receiver to generate a distance curve. The arrival of the signal is precisely measured at the two

receivers, the theory is that a signal that arrives at the two receivers at times t1 and t2 had to originate from a point on a curve defined by that difference in arrival time. For example, if the signal arrives at the two points at exactly the same time, then the transmitter must be equal distance from the two receivers. In this case the curve is actually a line of possible locations equal distance from the receivers and can easily be drawn. For situations where the signal arrives at different times at the receivers, a hyperbolic curve instead of a straight line denotes all the possible locations of the transmitter. Figure 17-19 shows an example of when the signal arrives at a different time, the constant value is the time difference multiplied by the speed of light yielding a distance. To solve for the emitter's location using TDOA another antenna receiver is required that is not in line with the first two receivers to generate an independent curve that will cross the first curve at the transmitter location.

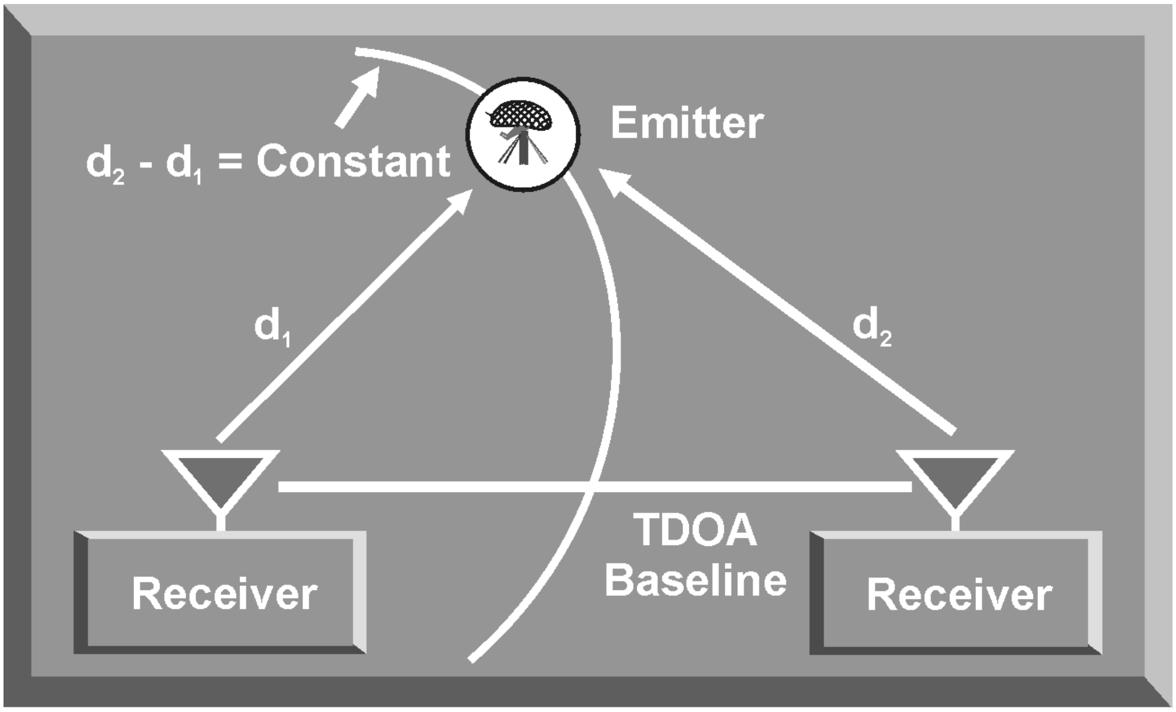

Figure 17-19. Time Difference of Arrival Measurements

(3) Timing techniques require very good timing accuracy in the receiving systems. If the receivers are far apart, such as different aircraft, then a separate time measurement is required at the source before sending to a processor. As with triangulation, one of the major challenges of multiple aircraft measurements is the coordination to ensure that the same signal is being looked at by all aircraft involved.

(4) Timing techniques are affected by the type of radar that they are trying to locate. Pure CW radars are not practical for a timing technique because there

is not a pulse to time. Pulsed signals, on the other hand, are much more susceptible to this type of location technique.

11. SUMMARY

An RWR system is designed to provide a picture of the electronic order of battle (EOB) operating in the vicinity of an aircraft. The signal processor is the heart of the RWR system and controls the function of all other system components. The signal processor, using inputs from the antennas and the receiver/amplifiers, compares the signal parameters with the parameters in the EID tables. The identified signals are displayed on the RWR scope with the appropriate audio. The ICU provides the aircrew interface with the signal processor to allow the aircrew to customize RWR operation for each combat mission. Modern RWR systems have some limitations that can effect aircrew survival. These limitations include ambiguous threat displays, maneuvering limitations, and EMI. Despite these limitations, an operational RWR system is one of the keys to survival on today's electronic battlefield. A product of advances on the electronic battlefield is rapid threat geolocation. New threat geolocation techniques are designed to provide advanced threat information to allow pilots to avoid, suppress, or destroy the mobile threat. The three most common techniques are basic triangulation using lines of bearing, interferometry using phase difference of arrival, and time difference of arrival using time differences to determine distance curves.

CHAPTER 18. SELF-PROTECTION JAMMING SYSTEM OPERATIONS

1. INTRODUCTION

Self-protection jamming systems are designed to counter surface-to-air missile (SAM), airborne interceptor (AI), and antiaircraft artillery (AAA) acquisition and target tracking radars. Self-protection jamming systems generate noise and deception jamming techniques to either deny threat system automatic tracking capability or generate sufficient tracking errors to prevent a successful engagement.

a. To counter the diverse array of threats and their associated frequencies in an integrated air defense system (IADS), a self-protection jamming system must be able to simultaneously jam multiple signals operating in a wide frequency range. The system must also be able to generate sufficient power to mask the radar return of the aircraft. Since many modern self-protection jamming systems are carried internally or externally on aircraft, their size, shape, and weight must be carefully controlled to minimize adverse effects on aircraft performance and handling. These design requirements may require a trade-off between additional power or capability and aircraft compatibility.

b. A self-protection jamming system must have receive antennas to receive radar signals, a system processor to identify received signals, a jamming techniques generator to produce an optimum jamming technique for the identified threat, and transmit antennas to transmit the required jamming techniques. These system components will be discussed in this chapter.

c. There are numerous internal and externally mounted self-protection jamming systems in use today. To simplify the discussion, the ALQ-184 pod will be used as an example of a modern, self-protection jamming system (Figure 18-1). The functions and operations of the ALQ-184 components are representative of currently deployed self-protection systems.

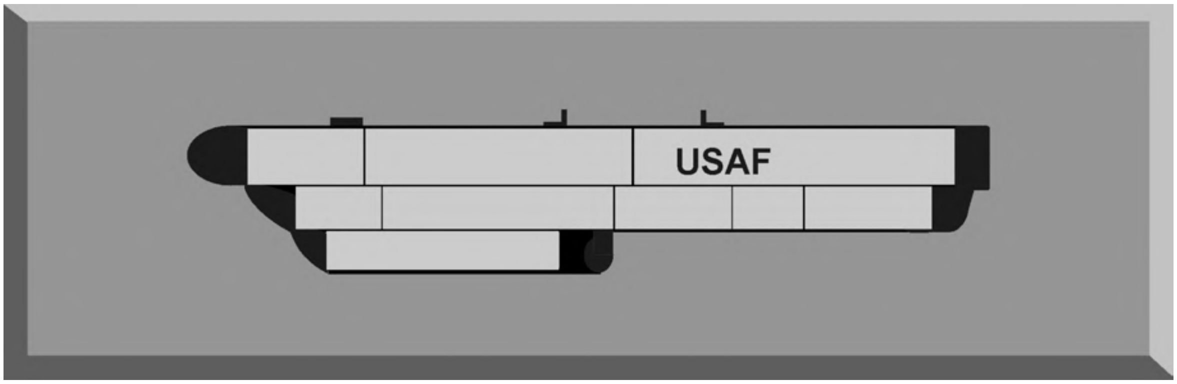

Figure 18-1. ALQ-184 Pod

2. RECEIVE ANTENNAS

The ALQ-184 pod has two sets of receive antenna assemblies located on the front and rear of the lower pod "gondola" (Figure 18-2). Each receive antenna assembly set consists of a low-band antenna and a mid/high-band antenna. The low-band antenna's signals are combined to form the input for the low-band receiver. The mid/high-band antennas provide signals to the separate mid/high-band portion of the pod.

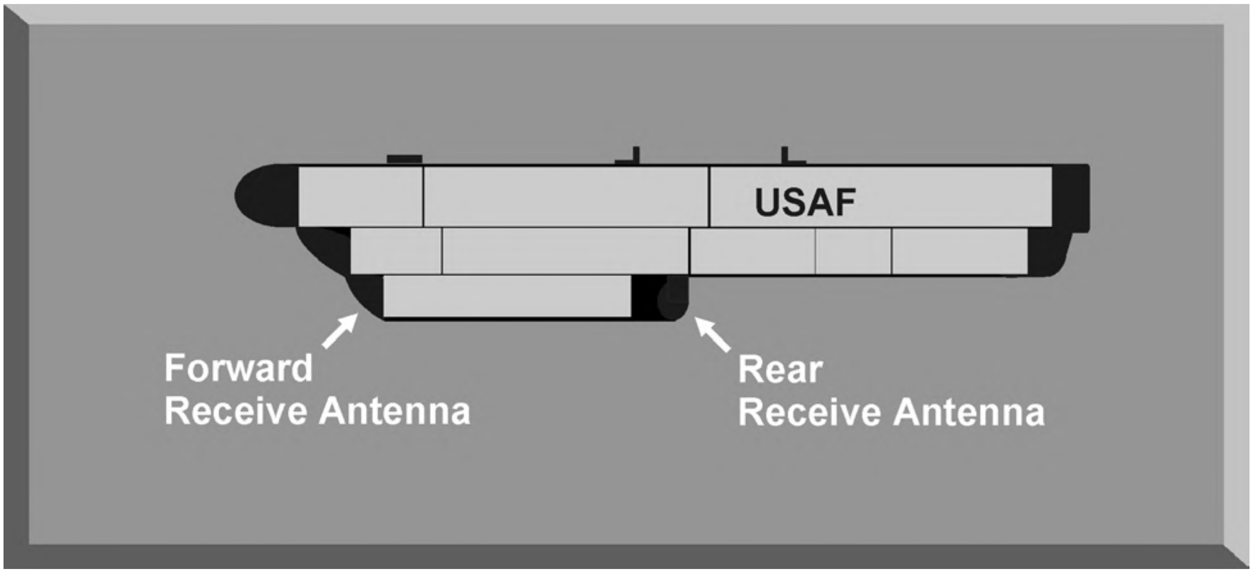

Figure 18-2. ALQ-184 Receive Antenna Location

a. The low-band receive antennas are circularly polarized spiral arrays. The mid/high-band receive antenna covers eight sub-bands via an eight element array that corresponds to the eight element transmit array. The mid/high-band antenna is also circularly polarized.

b. Each signal received via the receive antenna generates a corresponding transmit signal through a crystal video receiver. The angle of arrival (AOA) of each signal is determined by comparing the output of the crystal video receiver for each antenna.

3. RECEIVER SECTION

The forward and aft receive antennas detect radar signals and send them to the receiver section. The receiver section contains an AOA receiver and a multiplexer that separates frequencies. The AOA receiver determines the AOA of a signal based on the output signal level of each antenna element. The AOA data is passed to the system processor to control the angle of transmission of the jamming signal. The multiplexer separates all received threat signals into eight frequency sub-bands consisting of five mid-band and three high-band frequency

ranges. The multiplexer categorizes the received radar signals by sub-band and azimuth and then sends them to the system processor and the exciter.

4. SYSTEM PROCESSOR

The system processor is the "brain" of the ALQ-184 pod. It receives the threat signals from the receiver section that have already been categorized by sub-band and distinguished by AOA. Each received signal and its corresponding AOA is counted. When the signal count exceeds a pre-set threshold, the system processor validates the signal and identifies it using the emitter identification (EID) tables. Based on this threat identification, the system processor directs the techniques generator to initiate a jamming program through the exciter. At the same time, the system processor, through a signal switch control, directs the transmitter section to use either the forward or aft transmit antenna array, and specifies the transmit angle for the jamming program based on AOA.

a. The system processor also controls the low-band portion of the pod (Figure 18-3). The forward or aft low-band antenna receives and combines the threat signal and sends it to the multiplexer. The multiplexer channelizes this signal into one of two low sub-bands. The signal then passes to the modulator where the voltage controlled oscillators (VCOs) generate a jamming technique as directed by the system processor techniques generator. This ensures the jamming technique is at the proper frequency and modulation to provide maximum effect against the threat system. The generated jamming signal passes to a second multiplexer and then to a solid state amplifier for amplification. The amplified jamming signal is then sent to the two low-band traveling wave tubes (TWTs) and transmitted from the forward and aft low-band transmit antennas.

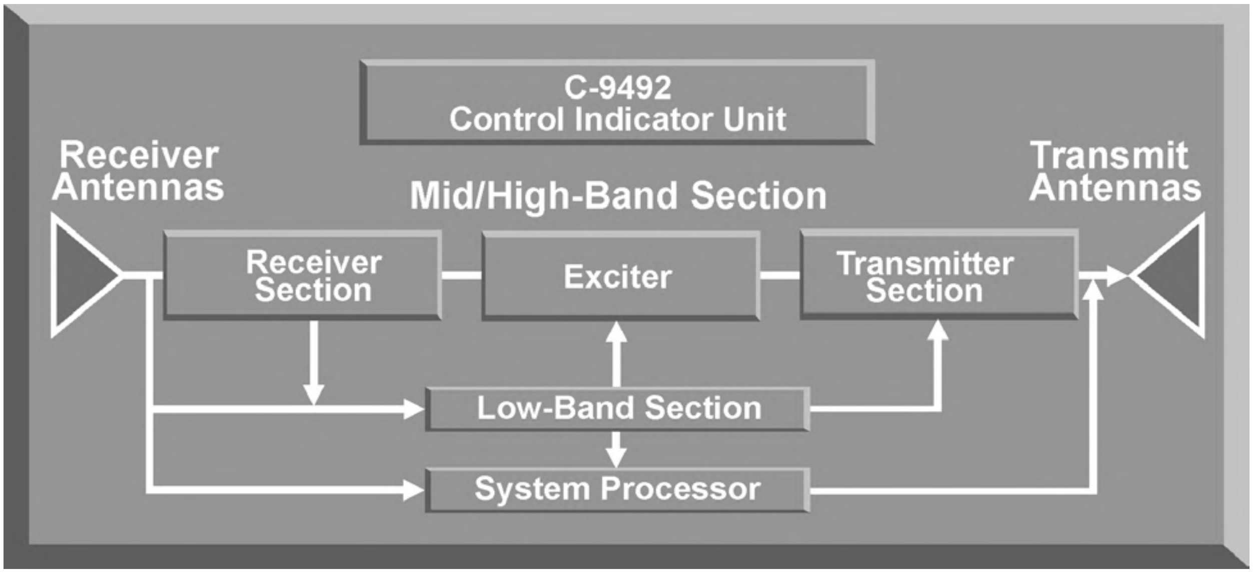

Figure 18-3. ALQ-184 Simplified Block Diagram

b. The system processor periodically performs a background built-in-test (B-BIT), to check the status and calibration of the ALQ-184. The system processor software is updated or reprogrammed as required from a memory loader verifier (MLV). The system processor uses two major software programs. The operational flight program (OFP) manages the receiver and transmitter functions, signal processing, and techniques generation. The mission data generator (MDG) contains the threat EID tables and the jamming techniques matrix for these threats.

5. JAMMING TECHNIQUES GENERATOR

The exciter section contains the VCOs and the keyed oscillators, which generate the jamming waveforms that are directed by the system processor. The exciter takes the threat signal from the receiver and modifies it with deception modulation or generates a noise program based on the techniques generator in the system processor. The selected jamming technique is then sent to the transmitter section.

6. TRANSMIT ANTENNAS

The transmitter section contains antenna switching circuits, a signal forming network, and TWTs. The antenna switching circuits are controlled by the system processor to ensure the transmitted jamming pulse is radiated at the proper azimuth. The signal-forming network controls the phase of each jamming signal. There are sixteen TWTs in the transmitter section, one for each front and aft antenna. The TWTs amplify the jamming signal for transmission from the appropriate transmit antenna.

a. The ALQ-184 has two sets of transmit antenna assemblies. One set is located on the front and one on the rear of the main pod assembly (Figure 18-4).

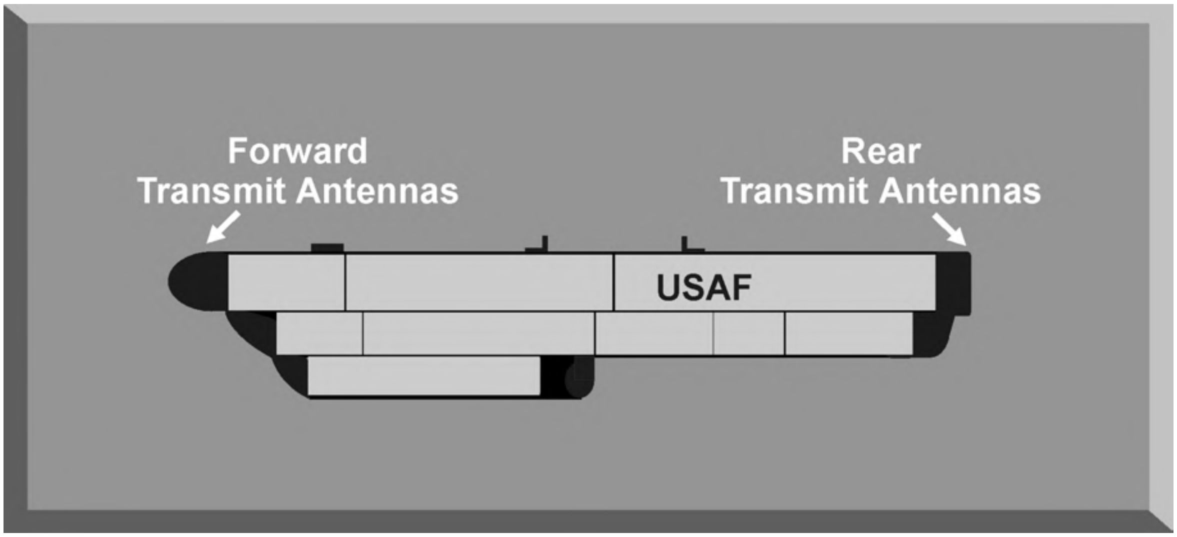

Figure 18-4. ALQ-184 Transmit Antenna Location

Each set has a low-band and mid/high-band antenna. The transmit antennas have opposite polarization from the receive antennas. This enables the pod to transmit and receive at the same time while preventing pod "ringing." Ringing occurs when the pod receives its own jamming signal and generates a jamming program to counter its own signal. This condition can highlight the aircraft to enemy radar and seriously degrade pod effectiveness.

b. The low-band transmit horn antennas are circularly polarized. The mid/high-band transmit antennas have the same eight-element array as the receive antennas. The transmit antennas generate a directional jamming signal for pulsed radar signals based on the AOA determined by the receive antennas. Detailed information on the antenna jamming pattern is contained in the ALQ-184 ECM Handbook.

7. C-9492 CONTROL INDICATOR UNIT

The C-9492 Control Indicator Unit (CIU) provides control of all functions of the ALQ-184. It allows the aircrew to control system power, select jamming techniques, and monitor system status (Figure 18-5).

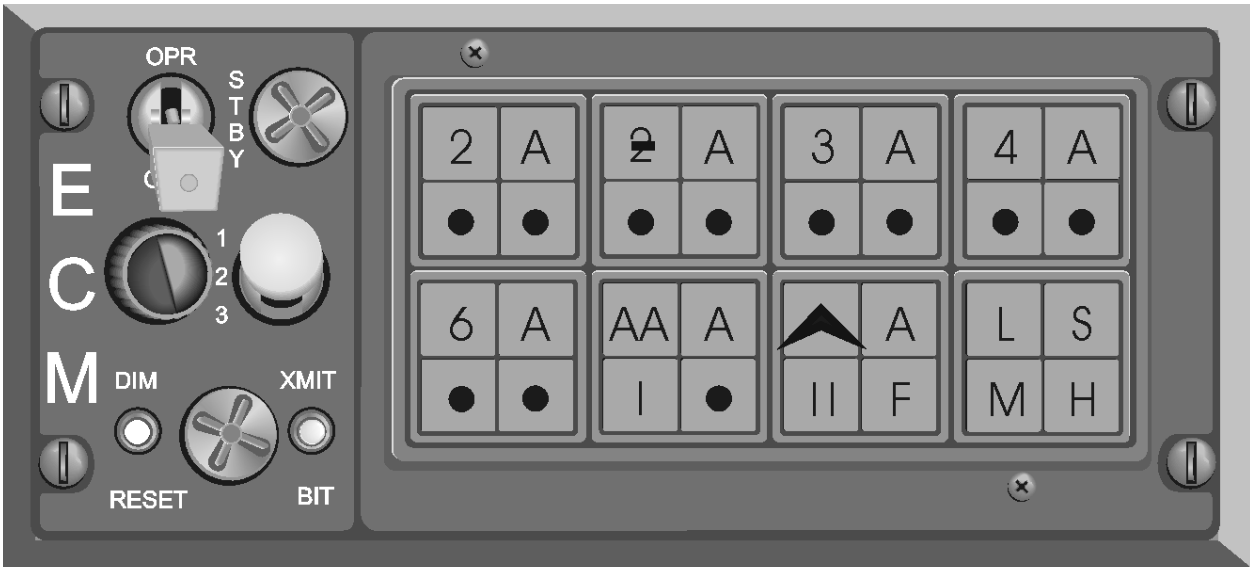

Figure 18-5. C-9492 Control Indicator Unit

8. JAMMING POD CONSIDERATIONS

Unfortunately, jamming pods do not make the aircraft completely invisible from enemy threat systems. When used in conjunction with the RWR, expendables, and threat countertactics they do drastically increase the probability of surviving a threat engagement. This section contains some items of consideration to ensure that your self-protection jamming is being properly utilized. These considerations include highlighting, burnthrough, and reprogramming.

a. If used improperly a self-protection pod can actually highlight the aircraft. This happens when a pod transmits a jamming program constantly or when another type of radar, not targeted for jamming but in the same frequency range, sees the jamming. Highlighting is normally associated with noise programs that transmit constantly. Most pods today avoid this by transmitting only when a threat requires it. Additionally, some aircraft have pod inhibit switches that allow the pod to remain in a standby condition until the pilot switches the pod to operate. What makes this system effective is that the switch is either on the stick or throttle allowing the pilot to rapidly go to operate without looking inside the cockpit (Figure 18-6).

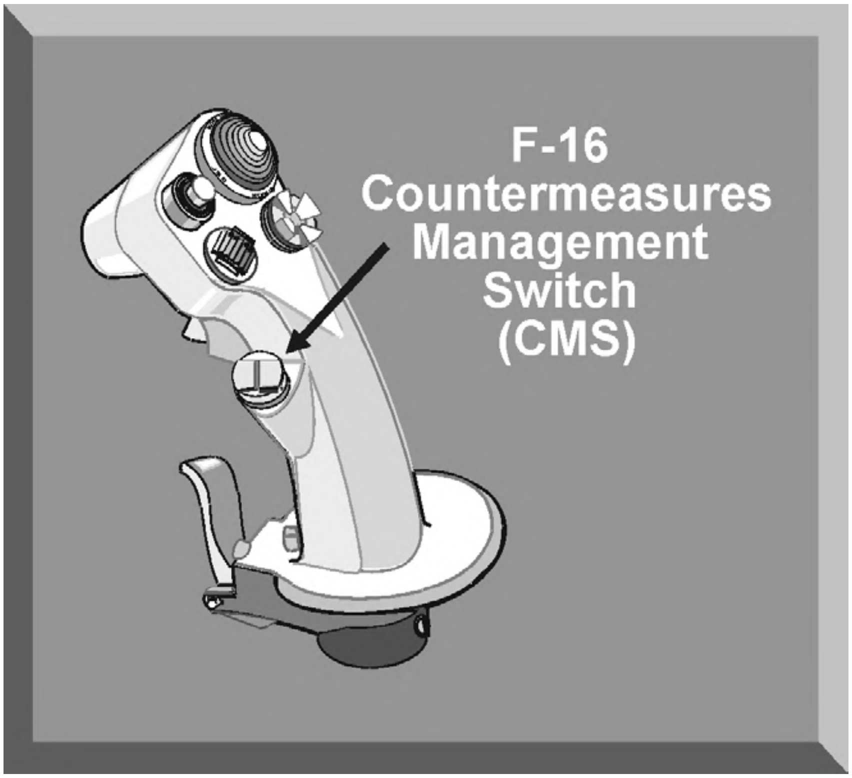

Figure 18-6. F-16 ECM CMS

b. Burnthrough is described as the range that the aircraft's radar return is stronger than its jamming signal. Basically, the jet has gotten so close to the threat radar that the threat radar is overpowering the jamming pod. The range that this occurs varies for different radars, but it should be planned for particularly when there are threats in the target area. Burnthrough is also a consideration for support jamming systems; there is a range at which the support jammer's signal will be overpowered and the protected aircraft can be seen.

c. Many self-protection jamming systems are what is called software-reprogrammable. This means that jamming programs can be rapidly changed to adjust for changes in enemy radar systems. These updates can be made to

systems while they are still on the jet, and can take as little as 15 minutes using an MLV. The process of changing electronic warfare systems is known as a Pacer Ware change. An exercise to check the ability to make changes is known as a Serene Byte exercise. The pilot of the aircraft carrying the self-protection system has one responsibility in this process: to ensure that his jamming system is carrying the most current software program, usually referred to as a software version. While Pacer Ware changes are not common during peacetime operations, they do happen quickly during combat operations often to fine tune the jamming systems for the particular theater of combat.

9. SUMMARY

The effective employment of self-protection jamming is one of the factors that can mean the difference between success and failure on the modern battlefield. When employed in concert with chaff and maneuvers, self-protection jamming can be extremely effective in negating potentially lethal attacks. Despite their limitations, self-protection jamming systems are an important part of the "bag of tricks" aircrews can employ to put bombs on target.

ANNEX A
GLOSSARY

Absorption - Dissipation of energy of electromagnetic waves, sound, and light waves into other forms of energy because of interaction with matter. Absorption characteristics of specific materials are used as blankets, coatings, or structural and surface materials for aircraft to reduce effective radar cross-sections.

Acoustic jamming - The deliberate radiation of mechanical or electro-acoustic signals with the objectives of obliterating or obscuring signals which the enemy is attempting to receive, and of deterring enemy weapon systems.

Acquire –

1. When applied to acquisition radars: the process of detecting the presence and location of a target in sufficient detail to permit identification.

2. When applied to tracking radars: the process of positioning a radar beam so that a target is in that beam to permit the effective employment of weapons.

Acquisition radar - A radar set that detects an approaching target and feeds approximate position data to a fire control or missile guidance radar that then takes over the function of tracking the target.

Active array radar - A phased array radar in which each radiating element contains a transmitter and receiver front end, as opposed to a single transmitter/receiver serving all phased array elements. Advantages attributed to active array radars include efficient use of prime power, no waveguide losses, very wide bandwidth, and extreme reliability due to the lack of single point of failure.

Active homing guidance - A system of homing guidance in which both the source for illuminating the target, and the receiver for detecting the energy reflected from the target as the result of illumination, are carried within the missile.

Aerosols - Solid particles dispersed in the atmosphere that have resonant size particles with a high index of refraction. The particles both scatter and absorb visual and laser-directed energy so as to lessen the effect of weapon systems directed by these techniques.

Airborne early warning and control - Air surveillance and control provided by airborne early warning vehicles that are equipped with search and height finding radar and communications equipment for controlling weapons.

Airborne interceptor (AI) - A manned aircraft used for identification and/or engagement of airborne objects. (An AI may or may not be equipped with radar to assist in the interception.)

Airborne Warning and Control System (AWACS) - An aircraft suitably equipped to provide control, surveillance, and communications capability for strategic defense and/or tactical air operations.

Air defense - All defensive measures designed to destroy attacking enemy aircraft or missiles in the earth's envelope of atmosphere, or to nullify or reduce the effectiveness of such attack.

Air surveillance radar (ASR) - A radar displaying range and azimuth that is normally employed in a terminal area as an aid to approach and departure control.

Air-to-air missile (AAM) - A missile launched from an airborne carrier at a target above the surface.

Air-to-surface missile (ASM) - A missile launched from an airborne carrier to impact on a surface target.

Alternating current (AC) - An electric current that reverses its direction at regularly recurring intervals, the frequency being determined by the frequency of the alternator supplying the current.

AM-CW jamming - Jamming in which a carrier wave is modulated at a constant recurring rate. The recurrence rate of amplitude modulation is variable, and reflects a noticeable change in the radar scope. Because of the spiraling or chaining effect produced, AM-CW jamming is easily identified.

Amplifier - An electronic circuit usually used to obtain amplification of voltage, current, or power.

Amplitude shift keying - A method of impressing a digital signal upon a carrier signal by causing the carrier amplitude to take different values corresponding to the different values of the digital signal.

Amplitude modulation (AM) - A method of impressing a message upon a carrier signal by causing the carrier amplitude to vary proportionally to the message waveform.

Angle jamming - A deception jamming technique used to deny azimuth and elevation information to a TTR by transmitting a jamming pulse similar to the radar pulse, but with modulation information out of phase with the returning target azimuth modulation information.

Angle tracking - Accomplished through the use of pulses to determine angular definition of a target.

Antenna - A device used for transmitting or receiving RF energy. The function of the antenna during transmission is to concentrate the radar energy from the transmitter into a shaped beam that points in the desired direction. During reception, or listening time, the function of the antenna is to collect the returning radar energy, contained in the echo signals, and deliver these signals to the receiver. Radar antennas are characterized by directive beams that are usually scanned in a recognizable pattern. The primary radar antenna types in use today fall into three categories: parabolic, Cassegrain, and phased array antennas.

Antenna cross talk - A measure of undesired power transfer through space from one antenna to another. Ratio of power received by one antenna to power transmitted by the other usually expressed in decibels.

Antenna gain - See Gain

Antenna polarization - See Polarization

Antenna sidelobes - See Sidelobes

Antiaircraft artillery (AAA) - Guns used to shoot unguided projectiles at airborne aircraft. Usually used in the air defense system.

Anti-clutter circuits (in radar) - Circuits that attenuate undesired reflections to permit detection of targets otherwise obscured by such reflections.

Antiradiation missile (ARM) - A missile that homes passively on a radiation source.

Area defense - The concept of locating defense units to intercept enemy attacks, remote from, and without reference to, individual vital installations, industrial complexes, or population centers.

Asynchronous pulsed jamming - An effective form of pulsed jamming where the jammer nearly matches the PRF of the radar then transmits multiples of the PRF. It is more effective if the jammer pulse width is greater than that of the radar. Asynchronous pulsed jamming is similar to synchronous jamming except that the target lines tend to curve inward or outward and appear fuzzy in the jammed sector of the radar scope.

Attenuation - The decrease in amplitude of a signal during its transmission from one point to another. Attenuation increases as distance increases. The higher the frequency of the propagating signal, the higher the rate of attenuation.

Automatic frequency control (AFC) - Circuits in a receiver that automatically correct the local oscillator frequency to prevent receiver drift in tuned frequency.

Automatic gain control (AGC) –

1. A feature involving special circuitry designed to maintain the output of a radio, radar, or television receiver essentially constant, or to prevent its exceeding certain limits, despite variations in the strength of the incoming signal. In a radio receiver, in particular, though something of a misnomer, also known as automatic volume control.

2. A self-acting compensating device that maintains the output of a transmission system constant with narrow limits, even in the face of wide variations in the attenuation of the system.

3. A radar circuit that prevents saturation of the radar receiver by long blocks of receiver signals, or by a carrier modulated at low frequency.

Automatic search jamming - An intercept receiver and jamming transmitting system that automatically searches for and jams enemy signals of specific radiation characteristics.

Automatic tracking - Tracking in which a system employs some mechanism, e.g., servo or computer, to automatically follow some characteristics of the signal.

Azimuth resolution - The ability of a radar to distinguish two targets in close azimuth proximity and distance.

Backlobe - The portion of the radiation pattern of an antenna that is oriented 180° in relation to the main beam. The antenna backlobe is a result of diffraction effects of the reflector and direct leakage through the reflector surface.

Backward wave oscillator (BWO) - A special traveling wave tube in which oscillatory currents are produced by using an oscillatory electromagnetic field to bunch the electrons as they flow from cathode to anode.

Ballistic missile - Any missile that does not rely upon aerodynamic surfaces to produce lift and consequently follows a ballistic trajectory when thrust is terminated.

Bandpass filter - A filter that allows a select range of frequencies to pass while attenuating all frequencies outside the range.

Bandwidth - The range of frequencies within which performance, with respect to some characteristics, falls with specific limits (i.e., the width of frequency of a barrage noise package).

Barrage noise jamming - Noise jamming spread in frequency to deny the use of multiple radar frequencies to effectively deny range information. Although this is attractive because it enables one jammer to simultaneously jam several radars of different frequencies, it does have the inherent problem that the wider the jamming spread, the less jamming power available per radar.

Beam rider - A missile guided by an electronic beam.

Beam-to-beam correlation (BBC) - Used by frequency scan radars to reject pulse jamming and jamming at swept frequencies. Correlation is made from two adjacent beams (pulses). The receiver rejects those targets (signals) that do not occur at the same place in two adjacent beams.

Beamwidth - The width of a radar beam measured between lines of half-power points on the polar pattern of the antenna. This width is measured at the 3 dB points.

Beat frequency oscillator (BFO) - Any oscillator whose output is intended to be mixed with another signal to produce a sum or difference beat frequency. Used particularly in reception of CW transmissions.

Bistatic radar - A radar where the transmitting and receiving antennas are separated by a considerable distance. Bistatic operation provides several advantages for its user. The covert positioning of the receivers poses problems for a potential attacking force since ELINT techniques locate the transmitter not the receiver. The proper placement of jamming assets is difficult, since the receiving sites are unknown. In addition, if a stand-off jammer is directed at the transmitter, its effectiveness in the direction of the covert receiver is diminished.

Jammers not in the same beam as the wanted targets will be attenuated by the receiver's sidelobe protection and these targets will be more readily detected.

Blanking –

1. The cutting off of the electronic beam in a cathode-ray tube when the picture is not being formed. The beam is blanked while the spot is returning to the starting position (normally right to left).

2. Process of making a channel or device noneffective for a desired interval.

Blinking - A jamming technique employed by two aircraft separated by a short distance and within the same azimuth resolution to appear as one target to a tracking radar. The two aircraft alternately spot jam, causing the radar system to oscillate from one plane to another, making an accurate solution of fire control problem impossible. However, keep in mind that too high a blinking frequency can cause the tracker to average the data while too low a frequency will cause a missile to home-in on one of the jammers.

Broad pulse jamming - Transmission of broad pulses for control system jamming when little is known about the command pulse group. For example, a broad pulse might cover a whole group of command pulses, thus jamming that command.

Burnthrough range - The ability of a radar to see through jamming. Usually described as the point when the radar's target return is stronger than the jamming signal.

Burst chaff - The formation of a reflective volume of chaff from an individual bundle.

Buzzer - Code name for airborne jamming.

Capture - Where the jammer takes control over the guidance signal by active jamming.

Capture effect - The tendency of a receiver to suppress the weaker of two signals within its bandpass.

Capture of AGC - Domination of the radar's automatic gain control (AGC) level by a strong transmitted jamming signal.

Carrier frequency - Frequency of an unmodulated radio wave emanated from a radio, radar, or other type of transmitter.

Carrier wave - Electromagnetic radiation used to "carry" information from one point to another.

Centroid homing - When applied to antiradiation missiles, the effect on a missile that has two or more radiation sources in its field of view causing it to home on the centroid of the power from the radiating sources.

Chaff - Ribbon-like pieces of metallic materials or metalized plastic that are dispensed by aircraft to mask or screen other aircraft or to cause a tracking radar to break lock. The foil materials are generally cut into small pieces for which the size is dependent upon the radar interrogation frequency (approximately 1/2 wavelength of the victim radar frequency). Being 1/2 wavelength long, chaff acts as a resonant dipole and reflects much of the energy back to the radar.

Chaff corridor - Operational technique of dropping large quantities of chaff for a continuous period of time. This results in a "ribbon" or "stream" of returns many miles long on radar scopes. The penetrating strike force can then use the resulting chaff corridor to mask its penetration.

Chirp - A pulse compression technique characterized by linear frequency modulation on pulse (LFMOP).

Chirp radar - Radar in which a swept-frequency signal is transmitted, received from a target, and then compressed in time to give a final narrow pulse called the chirp signal. It has high immunity to jamming and an inherent rejection of random noise signals.

Circularly polarized jamming - The techniques of radiating jamming energy in both planes of polarization. With this method, there is a 3-dB loss of effective power in either plane, but the enemy cannot cross-polarize his antenna to escape jamming.

Circular scan - The pattern generated by an antenna that is continuously rotating in one direction.

Clipped noise modulation - A clipping action is performed to increase the. bandwidth of the jamming signal. Results in more energy in the sidebands, correspondingly less energy in the carrier, and an increase in the ratio of average power to peak power.

Clutter - Unwanted signals, echoes, or images on the face of a scope that interferes with the observation of desired signals. Also called noise. This tends to mask the true target from detection or cause a tracking radar to break lock.

Clutter elimination - The clutter eliminator circuit discriminates against any target echo that exceeds three times the transmitted pulse width, and will not display it on the indicator. It is normally employed on the lower beams of a high frequency radar. This will allow targets above a preset signal strength to be presented, while the clutter (land) will be eliminated.

Clutter gating - A technique that provides switching between MTI and normal videos. It results in the normal video being displayed in regions with no clutter and the MTI video being switched in only for the clutter areas. Clutter gating is achieved automatically by the PW discrimination or the use of storage tubes. It also can be achieved by a manually operated range azimuth gate. The clutter gate vastly increases the effectiveness of noncoherent MTI against chaff.

Coherent MTI (in radar MTI) - A system in which the target echo is selected based on its Doppler frequency when compared to a local reference frequency maintained by a coherent oscillator.

Coherent noise jamming - Noise-like jamming signal that is repetitive and in synch with the PRI of the radar.

Coherent repeater jammer - A jammer that uses the phase information of the receiver radar signal in creating false targets thus transmitting a signal that is acceptable to the receiver processor of a victim coherent radar.

Command and control warfare (C2W) - The integrated use of operations security (OPSEC), military deception, psychological operations (PSYOP), electronic warfare (EW), and physical destruction, mutually supported by intelligence to deny information, influence, degrade, or destroy adversary C^2 capabilities while protecting friendly C^2 capabilities.

Command, control, communications, and computer systems (C4) - The process of, and means for, the exercise of authority and direction by a properly designated commander over assigned forces in the accomplishment of the commander's mission.

Command guidance - A guidance system in which intelligence transmitted to the missile from an offboard source causes the missile to traverse a directed flight path.

Communications intelligence (COMINT) - Intelligence derived from the interception of enemy communications signals.

Communications security (COMSEC) - The protection resulting from all measures designed to deny unauthorized persons information of value that might be

derived from the possession and study of telecommunications, or to mislead unauthorized persons in their interpretation of the results of such possession and study. COMSEC includes: 1. Cryptosecurity; 2. transmission security; 3. emission security; and 4. physical security of communications security material and information.

1. Cryptosecurity - The component of communications security that results from the provision of technically sound cryptosystems and their proper use.

2. Transmission security - The component of communications security from which all measures designed to protect transmissions from interception and exploitation by means other than cryptoanalysis.

3. Emission security - The component of communications security that results from all measures taken to deny unauthorized persons information of value that might be derived from intercept and analysis of compromising emanations from crypto equipment and telecommunications systems.

4. Physical security - The component of communications security that results from all physical measures necessary to safeguard classified equipment, material, and documents from access thereto or observation thereof by being within a friendly power.

Complex pulse - A pulse train in which there is more than one pulse width and/or more than one pulse repetition interval.

Conformal antenna - An antenna which conforms to a surface whose shape is determined by considerations other than electromagnetic, for example, an antenna shaped to aerodynamically fit the side of an aircraft.

Conical scan (CONSCAN) - A type of scanning in which the axis of the RF beam is tilted away from the axis of the reflector and rotated about it, thus generating a cone.

Constant false alarm rate (CFAR) receiver - A radar receiver with automatic detection circuits designed to produce a constant number of erroneous target detections independent of noise level at the receiver input. CFAR techniques are intended to prevent receiver saturation and overload, to present clear video information to the display, and a constant noise level to an automatic detector. A device that accomplishes these objectives may respond to the signal-to-noise ratio, for example, rather than the absolute signal level above a fixed threshold. CFAR does not usually permit the detection of a target if the target is weaker than the jamming, but it does attempt to remove the confusing jamming effects.

Continuous wave (CW) - A constant, single-frequency, unmodulated carrier wave that is transmitted and then reflected. A good system for determining velocity.

Correlation detection (modulation systems) - Detection based on the averaged product of the received signal and locally generated function possessing some known characteristics of the transmitted wave. Note the following caveats:

1. The averaged product can be formed, for example, by multiplying and integrating, or by using a matched filter whose impulse response, when reversed in time, is the locally generated function.

2. Strictly, the above definition applies to detection based on cross-correlation. The term correlation detection may also apply to detection involving autocorrelation, in which case the locally generated function is merely a delayed form of the received signal.

Countdown - A technique for forcing the radar AGC to continuously change value and oscillate. The jammer rapidly changes the duty cycle of the deception pulses.

Countdown blink - Self-screening AGC deception using a gated repeater or noise source; the period of which is short compared to the victim's AGC time constant.

Cover pulse - A jammer covers the radar return with an AM pulse usually much wider than the radar pulse. Since tracking circuits are looking for largest return, they will transfer to the cover pulse, thereby denying range information.

Cross-eye - A jamming technique used to produce angular errors in monopulse and other passive lobing radars. Jammer is a two-source interferometer that causes the phase front of the signal reaching the radar to be highly distorted. With such a technique, it is difficult for the radar to determine the points from which the transmissions are originating. Requires a high jam-to-signal ratio or the skin echo will show up in the pattern nulls.

Cross-gated CFAR - A CFAR technique employed to achieve the fast switching required for an optimum combination of normal and MTI modes. Here, the MTI video signals are used to “gate” on the normal video when the MTI indicates a target in clutter. CFAR action is achieved by the wideband as in the zero-crossing and Dicke fix CFARs.

Cross polarization - or “Cross Pole,” is a monopulse jamming technique where a cross-polarized signal is transmitted to give erroneous angle data to the radar. The component of the jamming signal with the same polarization as the radar must be very small.

Cross polarization jamming - A technique whereby the received signal is retransmitted cross-polarized and at a much higher power than the skin return-effective against some monopulse radars by generating erroneous angle information.

CW jamming - The transmission of constant-amplitude, constant-frequency, unmodulated jamming signals to change the signal-to-noise ratio of a radar receiver.

Data link - A communications link which permits automatic transmission of information in digital form.

Deception - Those measures designed to mislead the enemy by manipulation, distortion, or falsification of evidence to induce him to react in a manner prejudicial to his interests. (See Electronic Deception, or Manipulative Deception.)

Deception jamming - Any means of jamming consisting of false signals that have similar characteristics to the victim radar thereby deceiving the operator into erroneous conclusions.

Decibel (dB) - A dimensionless unit for expressing the ratio of two values, the number of decibels being 10 times the logarithm to the base 10 of a power ratio, or 20 times the logarithm to the base 10 of a voltage of current ratio. A power increase by 3 dB indicates a doubling of the original power.

(dBm) - Same as dBw except the reference level is one milliwatt instead of one watt.

(dBw) - Unit used to describe the ratio of the power at any point in a transmission system to a referenced level of one watt. The ratio expresses decibels above and below the reference level of one watt.

Defense suppression - A term applied to weapons systems that are intended to eliminate or degrade enemy detection, acquisition, or tracking equipment.

Delayed opening chaff - Chaff that blooms at a specific elapsed time after it is dispensed.

Detector balanced bias - Controlling circuit used in radar systems for anti-clutter purposes.

Dicke fix - A technique specifically designed to protect the receiver from ringing caused by noise, fast-sweep, or narrow-pulse jamming. The basic configuration

consists of a broadband limiting IF amplifier, followed by an IF amplifier of optimum bandwidth. The limit level is preset at approximately the peak amplitude of receiver noise. The bandwidth may vary from 10 to 20 MHz, depending on the jamming environment. The device provides excellent discrimination against fast-sweep jamming (10 - 500 MHz), usually something about 20 to 40 dB, without appreciable loss in sensitivity. However, strong CW jamming will seriously degrade the performance of the Dicke fix because the CW signal captures the radar signal in the limiter.

Dicke fix CFAR - Constant False Alarm Rate.

Dicke fix MT CFAR - An MTI CFAR technique similar to Dicke fix CFAR. Limiting and narrowbanding follow wideband amplification, phase detection, and cancellation so as not to impair the MTI performance.

Digital radio frequency memory (DRFM) - A computer-controlled digital device used in radar jamming systems. DRFM provides an extremely fast and accurate storage capability for victim radar signal parameters. Jamming systems employing DRFM can rapidly and accurately generate coherent jamming based on the memorized victim radar signal.

DINA - See Direct Amplified Noise.

Diode - An electron device having two electrodes, a cathode, and an anode.

Diplex - Two transmitters operating alternately on approximately the same RF and using a common antenna. The normal procedure is to pulse each transmitter at ½ the desired PRF, 180° out-of-phase. The advantage is that higher peak power per transmitter is possible because each transmitter is operating at ½ the normal duty cycle.

Dipole antenna - A straight, center-fed, ½ wavelength antenna. Horizontally polarized, it produces a figure-eight radiation pattern with maximum radiation at right angles to the plane of the antenna.

Direction finding (DF) - A procedure for obtaining bearings of radio frequency emitters with the use of a highly directional antenna and a display unit on an intercept receiver or ancillary equipment.

Direct amplified noise (DINA) - DINA without a carrier frequency is used for increasing (saturating) the radar receiver's noise level. This is called video stealing. The biggest danger from this type of jamming is that the radar operator may not realize he is being jammed when automatic gain control (AGC) or automatic noise leveling (ANL) are employed. This is a barrage type of jamming with a bandwidth normally more than 50 MHz. DINA appears as a bright, high intensity wedge when the manual gain is employed. Thus, all useful radar scope information is lost.

Doppler (effect) - Continuous wave (CW) Doppler radar modules are sensors that measure the shift in frequency created when an object moves. A transmitter emits energy at a specific frequency which, when reflected, can indicate both speed and direction of the target. When objects move closer to the Doppler source, they increase in shift (positive value), and when they move further away, they decrease in shift (negative value).

Doppler radar - A radar system that measures the velocity of a moving object by the apparent shift in carrier frequency of the returned signal as it approaches or recedes.

Downlink - The signal from a transponder beacon located on a surface-to-air missile (SAM) used to provide a traceable radar return for missile guidance.

Downlink jamming (DLJ) - Some command guidance missiles carry a beacon (downlink) which is used by the parent radar to track the missile. If this beacon reply can be hidden from the parent tracking radar, the missile guidance solution can be defeated. Hence, downlink (beacon) jamming is intended to screen the missile beacon signal from the parent radar's view.

Ducting - The bending of radar rays due to atmospheric conditions. Ducting can either extend radar coverage beyond normal line of sight or it can deny the radar picture above a duct. Ducting is also called Anomalous Propagation.

Dummy antenna - A device that has the necessary impedance characteristics of an antenna and the necessary power-handling capabilities, but does not radiate or receive radio waves. Note: In receiver practice, that portion of the impedance not included in the signal generator is often called a dummy antenna.

Dummy load (radio transmission) - A dissipative but essentially nonradiating substitute device having impedance characteristics simulating those of the substituted device. This allows power to be applied to the radar unit without radiating into free space.

Duplex - In radar, a condition of operation when two identical and interchangeable equipments are provided—one in an active state, and the other immediately available for operation.

Duplexer - A switching device used in radar to permit alternate use of the same antenna for both transmitting and receiving.

Duty cycle - The ratio of the time the transmitter is actually on versus the time it could be on in a given transmission cycle. Mathematically, it can be expressed as:

$$\text{Duty Cycle} = \frac{PD}{PRT} \quad \text{or} \quad \text{Duty Cycle} = PD \times PRF$$

Dynamic range –

1. The difference, in decibels, between the overload level and the minimum acceptable signal level in a system or transducer. Note: The minimum acceptable signal level of a system or transducer is ordinarily fixed by one or more of the following: noise level, low-level distortion, interference, or resolution level.

2. Ratio of the specified maximum signal level capability of a system or component to its noise or resolution level, usually expressed in decibels.

Early warning radar - A radar set or system used near the periphery of a defended area to provide early notification of hostile aircraft approaching the area.

EA pod - A jamming system that is designed to be carried externally on an aircraft.

Effective radiated power (ERP) - Input power to antenna time multiplied by the gain of the antenna, expressed in watts.

Electromagnetic interference (EMI) - Any electromagnetic disturbance that interrupts, obstructs, or otherwise degrades or limits the effective performance of electronic systems. EMI can be induced intentionally, by way of jamming, or unintentionally because of spurious emissions and modulations.

Electromagnetic pulse (EMP) - The generation and radiation in a transmission medium of a very narrow and very high-amplitude pulse of electromagnetic noise. The term is associated with the high-level pulse because of a nuclear detonation and with an intentionally generated narrow, high-amplitude pulse for EA applications. In nuclear detonations, the EMP signal consists of a continuous spectrum with most of its energy distributed throughout the low frequency band of 3 to 30 kHz.

Electromagnetic radiation - Radiation made up of oscillating electric and magnetic fields and propagated with the speed of light. Includes gamma radiation, x-rays, ultraviolet, visible and infrared radiation, plus radar and radio waves.

Electromagnetic spectrum - The total range of frequencies (or wavelengths) over which any form of electromagnetic radiation occurs.

Electromagnetic test environment (EMTE) - A range complex of radars, such as those located at Eglin AFB, FL, operating in different frequency bands and modes to provide a very flexible test facility for evaluating aircraft antenna patterns, reflectivity measurements, infrared, reconnaissance, airborne interceptors, and electronic warfare devices and techniques.

Electronic attack (EA) - The use of electromagnetic energy, directed energy, or antiradiation weapons to attack personnel, facilities, or equipment with the intent of degrading, neutralizing, or destroying enemy combat capability. Action taken to reduce the enemy’s effective use of the electromagnetic spectrum. EA is a division of electronic warfare (EW).

Electronic combat (EC) - Action taken in support of military operations against the enemy's electromagnetic capabilities. EC is task-oriented and includes electronic warfare (EW), command and control warfare (C2W), and suppression of enemy air defenses (SEAD).

Electronic protection (EP) - Active and passive means taken to protect personnel, facilities, and equipment from any effects of friendly or enemy employment of electronic warfare that degrade, neutralize or destroy friendly combat capability. EP is a division of electronic warfare (EW).

Electromagnetic deception - The deliberate radiation, reradiation, alteration, absorption, or reflection of electromagnetic radiations in a manner intended to mislead an enemy in the interpretation of, or use of, information received by his electronic systems. There are two categories of electronic deception:

1. **Manipulative deception** - The alteration or simulation of friendly electromagnetic radiation to accomplish deception.

2. **Imitative deception** - The introduction of radiations into enemy channels that imitate his own emissions.

Electronic intelligence (ELINT) - The intelligence information product of activities engaged in the collection and processing for subsequent intelligence purposes of foreign, noncommunications, electromagnetic radiations emanating from other than nuclear detonations or radioactive sources.

Electronic intelligence parameter limit list (EPL) - A compilation of identified signals with assigned ELINT notations.

Electronic jammers –

1. **Expendable** - A transmitter designed for special use such as being dropped behind enemy lines.

2. **Repeater** - A receiver-transmitter device that, when triggered by enemy radar impulses, returns synchronized false signals to the enemy equipment. The returned impulses are spaced and timed to produce false echoes or bearing errors in the enemy equipment. See Expendable and Repeater Jammers.

Electronic jamming - The deliberate radiation, reradiation, or reflection of electromagnetic energy with the object of impairing the use of electronic devices, equipment, or systems.

Electronic order of battle - A listing of all the electronic radiating equipment of a military force giving location, type function, and other pertinent data.

Electronic reconnaissance - Specific reconnaissance directed toward the collection of electromagnetic radiations. Examples:

COMINT	Communications Intelligence
ELINT	Electronic Intelligence
OPINT	Optical Intelligence
RINT	Radiated Intelligence
SIGINT	Signal Intelligence

Electronic warfare (EW) - Military action involving the use of electromagnetic energy and directed energy to control the electromagnetic spectrum. EW has three divisions: electronic attack (EA), electronic protection (EP), and electronic warfare support (ES).

Electronic warfare support (ES) - Actions taken to search for, intercept, identify, and locate sources of intentional radiated electromagnetic energy for the purpose of immediate threat recognition. Surveillance of the electromagnetic spectrum that directly supports an operational commander's electromagnetic information needs. ES is a division of EW.

Electro-optics (EO) - The interaction between optics and electronics leading to the transformation of electrical energy into light, or vice versa, with the use of an optical device.

Electro-optic counter-countermeasures (EOCCM) - Actions taken to ensure the effective friendly use of the electro-optic spectrum despite the enemy's use of countermeasures in that spectrum.

Emission control (EMCON) –

1. The management of electromagnetic radiations to counter an enemy's capability to detect, identify, or locate friendly emitters for exploitation by hostile action.

2. Controlling the radiation of an active system to minimize detection by enemy sensors.

Endgame - The period of military engagement 3-5 seconds before missile impact.

Endgame countermeasures (EGCM) - Actions taken to defeat a tracking missile. This includes expendables, decoys, and maneuvers.

Essential elements of information (EEI) - The critical items of information regarding the enemy and his environment needed by the commander by a particular time to relate with other available information and intelligence to be able to reach a logical decision.

Expendable jammer - A nonrecoverable jammer. Early expendables were limited to chaff and flare deployments; however, various radiating jamming systems exist that use noise or repeater techniques. These are dispensed by aircraft or other delivery systems and are designed to disrupt or deceive a victim radar for a short period of time.

Extremely high frequency (EHF) - Frequencies in the range of 30 to 300 GHz.

False target - Radiated bundle of electromagnetic energy that is displaced in time from the echo that creates a response in the receiver where no reflecting surface exists.

False target generator - Device for generating electromagnetic energy of the correct frequency of the receiver that is displaced in time from the reflected energy of the target.

Fast automatic gain control (FAGC) - An AGC scheme characterized by a response time that is long with respect to a PW and short with respect to the target. An ultra fast FAGC will reduce the CW capture effect on the Dicke fix.

Fast time constant circuit (in radar) –

1. A circuit with short time constant used to emphasize signals of short duration to produce discrimination against extended clutter, long-pulse jamming, or noise.

2. A resistance-capacitance differentiating network with a time constant about equal to the transmitted pulse width and employed in the video portion of the receiver to provide the discrimination against jamming with low modulating frequencies. FTC is effective against CW, long jamming pulses, swept jamming, and clutter. It is also fairly effective against chaff corridors. Based on time, the radar receiver will allow only pulses equal to its own pulse width to pass and be presented on the indicators as targets. Because of this, only the leading edges of long pulses will be displayed.

Fence –

1. Line or network of early warning radars.

2. Concentric steel fence erected around a ground radar transmitting antenna to serve as an artificial horizon and suppress ground clutter that would otherwise drown out weak signals returning at a low angle from the target.

Ferret - An aircraft, ship, or vehicle especially equipped for the detection, location, recording, and analyzing of electromagnetic radiations.

Field of view (FOV) - The maximum solid angle visible by an optical or electro-optic system.

Fire control radar - Specialized radar systems used to locate and track airborne and surface targets, compute an optimum weapons firing point, and control the firing and sometimes guidance of its weapons.

FM-by-noise modulation - A method of frequency modulating with effective jamming method against AM and fix-tuned FM receivers. Not very effective against continuously tunable PFM receivers; careful tuning can defeat a great portion of the jamming signal. For this reason, FM-by-noise is not considered optimum as a type of modulation for jamming FM receivers.

FM jamming - Technique consisting of a constant amplitude RF signal that is varied in frequency around a center frequency to produce a signal over a band of frequencies.

Forward edge of the battle area (FEBA) - The foremost limits of a series of areas in which ground combat units are deployed, excluding the areas in which the covering or screening forces are operating, designated to coordinate fire support, the positioning of forces, or the maneuver of units.

Frequency agility - A radar's ability to change frequency within its operating band, usually on a pulse-to-pulse basis. This is an EP technique employed to avoid spot jamming and to force the jammer to go into a less effective barrage mode.

Frequency band designations –

1. As decided upon by the Atlantic City Radio Convention of 1947, and later modified by Comite Consultatif International Radio (CCIR) Recommendation No. 142 in 1953:

Band	RF Range	Band	RF Range
VLF	0-30 kHz	VHF	30-300 MHz
LF	30-300 kHz	UHF	300-3000 MHz
MF	300-3000 kHz	SHF	3-30 GHz
HF	3-30 MHz	EHF	30-300 GHz

2. Electronic warfare uses the following designations:

A	0-250 MHz	H	6-8 GHz
B	250-500 MHz	I	8-10 GHz
C	500-1000 MHz	J	10-20 GHz
D	1000-2000 MHz	K	20-40 GHz
E	2-3 GHz	L	40-60 GHz
F	3-4 GHz	M	60-100 GHz
G	4-6 GHz		

Frequency bandpass - The number of hertz where maximum output is obtained between two limits usually defined and bounded by lower and upper half power (3 dB) points.

Frequency hopping - An anti-jamming technique used by a radar system. The carrier frequency of the pulsed transmissions are periodically or continuously shifted within limits on each successive pulse.

Frequency diversity - Method of transmission and/or reception using several frequencies simultaneously to minimize the effects of selective fading, deliberate jamming, or interference.

Frequency modulated CW (FM-CW) jamming - FM of a CW signal to produce a wide band of signals when the exact operating frequency of the victim is not known within the CW limits.

Frequency modulation (FM) - A method of impressing a message upon a carrier signal by causing the carrier frequency to vary proportionally to the message waveform.

Frequency shift keying (FSK) - A form of FM where the carrier is mined code. In multiple FSK, the carrier is shifted to more than two frequencies. FSK is used principally with teletype communications.

Frequency spectrum - The entire range of frequencies of electromagnetic radiation.

G - Acceleration due to gravity (32.2 ft/sec^2).

Gain (manual) - The receiver gain control allows the operator to vary the receiver sensitivity. It is not designed as an AJ feature; however, when properly employed it may greatly reduce the effects of jamming. The radar detection capability is also reduced by an equal amount.

Gain (transmission gain) - The increase in signal power in transmission from one point to another under static conditions. Note: Power gain is usually expressed in decibels.

Glint (in radar) –

1. The random component of target location error caused by variations in the phase front of the target signal (as contrasted with scintillation error). GLINT may affect angle, range of Doppler measurements, and may have peak values corresponding to locations beyond the true target extent in the measured coordinate.

2. Electronic countermeasures that use the scintillating, or flashing, effect of shuttered or rotating reflectors to degrade tracking or seeking functions of an enemy weapons system.

Ground controlled intercept (GCI) - Vectoring an interceptor aircraft to an airborne target by means of information relayed from a ground-based radar site that observes both the interceptor and target.

Guidance system (missile) - A system that evaluates flight information, correlates it with target data, determines the desired flight path of the missile, and communicates the necessary commands to the missile flight control system.

Guided missile - An unmanned vehicle moving above the surface of the earth, whose trajectory of flight path is capable of being altered by an external or internal mechanism.

Height finder - A radar used to detect the angular elevation, slant range and height of objects in the vertical sight plane. An air defense ground radar used specifically to accurately determine aircraft altitude for tracking and ground controlled intercepts.

Hertz (Hz) - The unit of frequency, equal to one cycle of variation per second. It supersedes the unit cycle per second (cps).

High frequency (HF) - Frequencies from 3000 - 30,000 kHz.

Home-on-jam (HOJ) - A missile mode of operation in which a jamming signal is used to develop steering information for the missile to home in on the jamming source.

Homing guidance - A system by which a missile steers itself toward a target by means of a self-contained mechanism which is activated by some distinguishing characteristics of the target.

Identification, friend or foe (IFF) - A system using radar transmission to which equipment carried by friendly forces automatically responds, for example, by emitting pulses, thereby distinguishing themselves from enemy forces. It is the primary method or determining the friendly or unfriendly character of aircraft and ships by other aircraft and ships, and by ground forces employing radar detection equipment and associated identification, friend or foe units.

Image frequency - An undesired input frequency capable of producing the selected frequency by the same process. NOTE: An image frequency is a frequency which differs from, but has a certain symmetrical relationship to, that which a superheterodyne receiver is tuned. Consequently, the image frequency can be mistakenly accepted and processed as a true frequency by the receiver.

Image jamming - Jamming at the image frequency of the radar receiver. Barrage jamming is made most effective by generating energy at both the normal operating and imaging frequency of the radar. Image jamming inverts the phase of the response and is thereby useful as an eagle deception technique. Not effective if the radar uses image rejection.

Imitative deception - The introduction of radiations into enemy channels which imitates their own emissions.

Imitative jamming - The jamming technique of transmitting a signal identical to the original guidance signal.

Infrared (IR) - That portion of the frequency spectrum lying between the upper end of the millimeter wave region and the lower (red) end of the visible spectrum. In wavelength, the IR lies between 0.78 and 300 microns; in frequency, it lies between one and 400 terahertz (THz).

Infrared counter-countermeasures (IRCCM) - Actions taken to effectively employ our own infrared radiation equipment and systems in spite of the enemy's actions to counter their use.

Infrared countermeasures (IRCM) –

1. Countermeasures used specifically against enemy threats operating in the infrared spectrum.

2. Actions taken to prevent or reduce the effectiveness of enemy equipment and tactics employing infrared radiation.

Instantaneous automatic gain control (IAGC) –

1. That portion of a system that automatically adjusts the gain of an amplifier with varying input pulse amplitudes. The adjustment is sufficiently fast to operate while a pulse is passing through the amplifier.

2. In radar, a quick-acting automatic gain control that responds to variations of mean clutter level, or jamming over different range or angular regions, avoiding receiver saturation (also known as back-bias). IAGC automatically adjusts the gain of the radar receiver so that strong signals do not block adjoining weak signals. IGAC is not quick enough to block short pulse jamming, extended clutter, or chaff.

Instantaneous frequency CFAR (IFCFAR) - A technique similar to the IF DICKE FIX CFAR, except for having a phase detector instead of a frequency discriminator. The primary use of this technique, combined with rapid random frequency changes, is to make chaff signals appear noise-like so they can be handled conventionally.

Instantaneous inverse gain - Pulse-by-pulse amplification, modulation, and reradiation of the victim's radar pulse to obscure angle data.

Intercept point - A computed point in space toward which an interceptor is vectored to complete an interception.

Interference (electronic) - An electrical or electromagnetic disturbance that causes undesirable responses on electronic equipment. Electrical interference refers specifically to interference caused by the operation of electrical apparatus that is not designed to radiate electromagnetic energy.

Interferometer - As pertains to radar, a receiving system which determines the angle of arrival of a wave by phase comparison of the signals received at separate antennas or separate points on the same antenna.

Interleaved jamming - Time sharing of a jammer's output around many threats usually to cover pulse jamming.

Intermediate frequency (IF) –

1. A fixed frequency to which all carrier waves are converted in a superheterodyne receiver.

2. A frequency to which a signaling wave is shifted locally as an intermediate step during transmission or reception.

3. A frequency resulting from the combination of the received signal and that of the local oscillator in a superheterodyne receiver.

Intermediate frequency jamming - Form of CW jamming that is accomplished by transmitting two CW signals separated by a frequency equal to the center frequency of the radar receiver IF amplifier.

Interrogator - A device used to transmit pulse-coded challenges to an IFF transponder and then evaluates the pulse-coded reply for identification purposes.

Intrapulse modulation repeater - A classified deception jamming technique.

Intrusion –

1. The entry of a nonfriendly aircraft or system into friendly air space.

2. The intentional interference in a communication system by which the intruder attempts to confuse, delay, or cause error by the selective introduction of additional data.

Inverse conscan - One method of confusing a radar operator or fire control radar system is to provide erroneous target bearings. This is accomplished by first sensing the radar antenna scan rate and then modulating repeater amplifier gain so that the weapons systems will fire at some bearing other than the true target bearing. The angle deception technique is used to break lock on CONSCAN radars.

Inverse gain - Amplification, inverse modulation, and reradiation of a radar's pulse train at the rotation rate of the radar scan. Deceives a conical scanning radar in angle.

Inverse-gain repeater jammer - A form of repeater in which the jammer creates false targets by varying the output power inverse with the strength of the received radar signal.

Jaff - Expression for the combination of electronic and chaff jamming. An ECM tactic involving the use of jammers to illuminate chaff corridors or chaff bursts to produce false targets.

Jam attenuator control - Used to prevent receiver saturation from any strong signals, including electronic jamming, chaff, or clutter. It also permits the determination of the bearing and elevation of jammers that would cause wide sectors of the scope to be obscured. The control reduces all signals equally, jamming as well as targets. This control should be used in the jammed sector only and not for an entire antenna revolution. Targets in the jammed sector would only be seen if they were stronger than the jamming. However, the jammed sector may be reduced enough in size to allow the operator to determine either bearing or elevation.

Jammer - A device used to deprive, limit, or degrade the use of communications or radar systems. Radio frequency jammers include noise, discrete frequency repeater, and deceptive equipment.

Jamming-to-signal (J/S) ratio - The relative power ratio of jamming to the radar return signal at the radar antenna. The inverse of the signal-to-jamming ratio.

Jam strobe - Also called JAVA (jamming amplitude versus azimuth). A circuit that generates a marker on the PPI to indicate signal strength as a function of bearing. It does this by sampling the jamming intensity once each repetition period. Besides showing the direction of the jammer, it also indicates the severity of main beam and sidelobe jamming.

Jet engine modulation (JEM) - Modulation present in the radar returns received from a jet aircraft, caused by the rotation of the fan or turbine blades of the aircraft's engines.

Jittered pulse repetition frequency (jittered PRF) - The PRF is rapidly varied at a random rate so that false targets appear to jitter or appear fuzzy on the scope. An alternative to jittered PRF is to change the PRF momentarily. This causes the false targets to change their position on the scope. It provides a discrimination against repeater-type jammers.

Klystron - A very stable microwave amplifier that provides high gain at good efficiency. This is accomplished by velocity modulating (accelerating a beam of electronics flowing from its cathode to its anode.

Laser target designation - The use of a laser to direct a light beam onto the target so that appropriate sensors can track or home on the reflected energy.

Leading-edge tracker - A tracking radar that obtains its data from the leading edge of the echo pulse from the target.

Light amplification by stimulated emission of radiation (LASER) - A process of generating coherent light. The process uses a natural molecular (and atomic) phenomenon whereby molecules absorb incident electromagnetic energy at specific frequencies. It then stores this energy for short but usable periods, then releases the stored energy as light at particular frequencies, and in an extremely narrow frequency band.

Lobe - One of the three-dimensional sections of the radiation pattern of a directional antenna bounded by 1-2 cones of nulls.

Lobe-on-receive-only (LORO) - Mode of operation consisting of transmitting on one antenna system and receiving the reflected energy on another system (TWS, conical, or monopulse).

Local oscillator off - A simple expedient of shutting off the local oscillator during barrage jamming. The barrage jammer will not be seen unless there is a beating signal such as a target; therefore, targets will appear. Since targets will not

appear in directions where no jamming arises, either an automatic azimuth switch or an additional receiver display is required.

Logarithmic fast time constant log (LOG-FTCL) - A device consisting of a logarithmic IF amplifier followed by an FTC circuit. The LOG-FTC combination is very effective in removing variations in the video output noise level caused by spot noise, wideband noise, and slow sweep noise modulated AM jamming.

Logarithmic receiver - A receiver whose response approximates the logarithm of the strength of the incoming signal. A special type of receiver having a large dynamic range of automatic gain control that gives considerable protection against receiver saturation by strong jamming on interference signals. Useful against weather, clutter, chaff, and spot jamming.

Look-down, shoot-down - Refers to an air interceptor (AI) equipped with a pulse Doppler radar, or a radar that has a moving target indicator (MTI) feature, that can detect and lock-on to a target within ground return clutter enabling the AI to track and shoot the target.

Look-through –

1. When jamming, a technique by which the jamming emission is interrupted irregularly for extremely short periods to allow monitoring of the victim signal during jamming operations.

2. When being jammed, the technique of observing or monitoring a desired signal during interruptions in the jamming signals.

Low frequency (LF) - Frequencies from 30 - 300 kHz.

Low power spread spectrum radar - A low power, high duty cycle radar whose spectrum is spread 100 MHz or more. Since this radar has a broad output spectrum and a high duty cycle, neither time nor frequency can be effectively used to resolve these signals. This leaves direction as the prime method of resolution. The spectrum of these radars is spread over the bandwidth by any of the pseudo random noise modulating techniques commonly used in communications. Techniques such as bi-phase modulation, quaternary phase modulation, chirp, random frequency jumping, etc., may be used to spread either a CW signal or a very high duty cycle signal. Such signals have a very good range resolution—approximately equal to the reciprocal of the bandwidth.

Magnetron - A radar microwave device whose operation is based on the motion of electrons (AC) under the influence of combined electric and magnetic fields.

Mainlobe - The lobe of a transmitting or receiving antenna centered on the directivity axis of the antenna.

Manipulative deception - The alteration or simulation of friendly electromagnetic radiations to accomplish the deception.

Meaconing, interference, jamming, and intrusion (MIJI) - An acronym of four component parts to determine the intent and technique of electromagnetic radiation.

Meaconing - a system of receiving radio beacon signals and rebroadcasting them on the same frequency to confuse navigation. Interference - unintentional electromagnetic radiation that may cause interference with electronic equipment.

Jamming - the deliberate radiation, reradiation, or reflection of electromagnetic energy with the intent of impairing the use of electronic devices, equipment, or systems being used by the enemy.

Intrusion - the intentional insertion of electromagnetic energy into transmission paths in any manner with the objective of deceiving operators or causing confusion.

Medium frequency (MF) - Frequencies from 300 to 3,000 kHz.

Micron - A unit of length equal to a micrometer (10^{-6} meters).

Microwave amplification by stimulated emission or radiation (MASER) - A low-noise, radio-frequency amplifier. The emission of energy stored in a molecular or atomic system by a microwave power supply is stimulated by the input signal.

Microwave communications - Line-of-sight communications, the frequency of which is higher than 300 MHz.

Millimeter waves - Frequencies (30 GHz to 300 GHz) in the millimeter portion of the electromagnetic spectrum.

Miss distance - The distance measured between the closest paths of a target and interceptor (i.e., aircraft and missile). One objective of self-protection jamming systems is to increase the miss distance to avoid destruction if missile launch cannot be prevented.

Missile approach warning system (MAWS) - A system used to detect and provide warning of approaching missiles. MAWS may be partitioned into active MAWS and passive MAWS.

1. **Active missile approach warning system (AMAWS)** - Generally employs pulse Doppler radar as its sensor. This radar is able to discern a moving target in stationary or slow-moving background clutter.

2. **Passive missile approach warning system (PMAWS)** - An ultraviolet (UV) or infrared-based detector system with the ability to detect and distinguish threat missiles from surrounding clutter and non-lethal missiles.

Modulated barrage jamming - A technique that varies the amplitude of the output power of a barrage jammer with sinusoidal or complex modulation.

Modulated CW jamming - A CW carrier waveform that has been modulated with another signal (such as noise, low, medium, or high frequencies), and is transmitted for the purpose of deception. May be AM, FM, pulse modulated (PM), etc.

Modulated PRF (MPRF) - The deliberate modulation of the interpulse spacing in a pulse train. See PRF Jitter and PRF Stagger.

Modulation - The variation of amplitude, frequency, or phase of an electromagnetic wave by impressing another wave on it.

Modulator - A device (such as an electron tube) for modulating a carrier wave or signal for the transmission of intelligence of some sort.

Monopinch - Anti-jam application of the monopulse technique where the error signal is used to provide discrimination against jamming signals.

Monopulse - A method of pulse generation that allows the simultaneous determining of azimuth, elevation and range, and/or speed from a single pulse.

Monopulse radar - A radar using a receiving antenna system having two or more partially overlapping lobes in the radiation patterns. Sum and difference channels in the receiver compare the amplitudes or the phases of the antenna outputs to determine the angle of arrival of the received signal relative to the antenna boresight. A well-designed monopulse tracking system will achieve a more accurate track under conventional jamming techniques than on the skin return. Certain monopulse trackers are susceptible to angular jamming techniques such as skirt and image jamming. Techniques such as “CROSS EYE” are designed to attack all monopulse tracking systems. Monopulse deception is a major area of advanced R&D with no clear “best technique” yet in sight.

Moving target indicator (MTI) - A radar presentation that shows only targets that are in motion. Signals from stationary targets are subtracted out of the return signal by the output of a suitable memory circuit.

MTI CFAR - A technique that provides CFAR capability in an MTI receiver. The cancellation of ground clutter is not impaired during radar jamming.

Multiband radar - Radar that simultaneously operates on more than one frequency band through a common antenna. This technique allows for many sophisticated forms of video processing and requires that a jammer must jam all channels simultaneously.

Multiplex - Simultaneous transmission of two or more signals on a common carrier wave. The three types of multiplex are called time division, frequency division, and phase division.

Multitarget generator jamming - A technique where a generator takes the radar's PRF, scan rate, and antenna lobe pattern and computes when to transmit false targets at the radar's frequency. The targets appear as true targets, but normally only about 20 percent of the targets are programmed to give a logical course and speed. The many targets saturate the tracking operator's capability by the sudden appearance of multiple targets, and/or many targets either stationary or on illogical courses and speeds, or maneuvers such as 90 degree turns at high speeds but with no displacement. Also, targets may appear in back and sidelobe positions.

Music - In air intercept, a term meaning electronic jamming.

Narrowband (NB-1) - Narrows the receiver frequency making it more selective. It limits the target signals and the jamming signal to a set level of amplitude and reduces the level of a jamming signal if the jammer is not tuned to the radar's exact frequency.

Narrow pulse jamming - Jamming where the jammer pulse width is less than the radar's pulse width. Technique may be required by interleaved jamming.

Noise –

1. Any unwanted disturbance within a dynamic electrical or mechanical system, such as undesired electromagnetic radiation, and any transmission channel or device.

2. Uncontrolled random disturbances that arise in a guided missile system because of various physical phenomena.

Noise jamming - Direct (straight) AM or FM noise on a carrier frequency that has a highly variable bandwidth for the purpose of increasing (saturating) the radar receiver's noise level.

Nuclear effects - The electromagnetic phenomena resulting from a nuclear explosion. The phenomena are listed as follows:

1. **Argus phenomena (trapped electrons)**. The trapping in the earth's magnetic field of electrons produced by a nuclear burst.

2. **Blackout (radio frequency interference)**. An effect that is the result of ionization produced by a nuclear explosion in or above the atmosphere. This ionization can cause interference (blackout) by attenuating, reflecting, cluttering, and scattering radar electromagnetic pulses in a high-intensity burst of electromagnetic radiation, predominantly in the radio frequency range of the spectrum.

3. **Optical phenomena**. Intense radiations covering all parts of the optical spectrum are produced by the interactions between the atmosphere and the nuclear radiation and fission products resulting from a nuclear detonation. The resulting auroras and airglows are created as an optical background that can affect reconnaissance, tracking, warning and homing systems, and personnel.

4. **Transient radiation effects on electronics**. Nuclear radiation impinging on electronic systems or components can substantially alter the operation and output of these systems. The word transient refers to the type of environment and not to the duration of the effect since the effect may be either transient or permanent.

Nulls - Points of no radiation in an antenna radiation pattern.

Obscuration - Effects produced by masking-type jammers. Denial of either range or angle is achieved by submerging data interference caused by noise or noise-like signals.

Off-frequency spot jamming - A type of spot jamming in which the jammer frequency is off the radar operating frequency but still within the radar receiver bandpass.

Optical countermeasures - Applications of electronic countermeasures in the visible light portion of the electromagnetic spectrum. Actions taken to prevent or reduce an enemy's effective use of the visible spectrum.

Oscillator - Electronic circuit or device capable of converting direct current (DC) into alternating current (AC) at a frequency determined by the inductive and the capacitive constants of the oscillator.

Over-the-horizon radar - A radar system that makes use of the ionosphere to extend its range of detection beyond line-of-sight. Over-the-horizon radars may be either forward scatter or backscatter systems.

Palmer scan - Conical scan superimposed on another type of scan pattern-usually a spiral pattern.

Passive angle tracking (PAT) - A target may be tracked “passively” if that target emits electromagnetic radiation; i.e., jamming radio, radar signal of sufficient duration that a DF bearing may be obtained. The emission from the target is DF-ed in azimuth and/or elevation. No range information will be available unless cross DF techniques are used by two or more passively tracking sites.

Passive detection and tracking - By combining azimuth data on jamming strobes from several stations, intersections are obtained which indicate the position of the jammers. The number of ghosts can be reduced by increasing the number of friendly stations and obtaining elevation angles of strobes when available.

Passive electronic countermeasures - Electronic countermeasures based on the reflection, absorption or modification of the enemy's electromagnetic energy. This distinction between active and passive countermeasures is not currently used, but is based on the presence or absence of an electronic transmitter.

Passive homing guidance - A system of homing guidance in which the receiver in the missile uses radiations only from the target.

Peak power - Maximum power output during transmission time. Expressed in watts or megawatts.

Penetration aids - Techniques and/or devices employed by aerospace systems to increase the probability of weapon system penetration of any enemy defense.

Examples are: Low altitude profiles, trajectory adjustments, reduced radar cross-sections of attack vehicles, improved vehicle hardness to effects of defense engagements, terrain avoidance radar, bomber defense missiles, decoys, chaff, electronic countermeasures, etc. Penetration aids are used by an offensive system to penetrate enemy defenses more effectively. Also called PENAIDS.

Phase modulation - A method of impressing a message upon a carrier signal by causing the carrier phase to vary proportionally to the waveform.

Phase shift keying - A method of impressing a digital signal upon a carrier signal by causing the carrier phase to take different values corresponding to the different values of the digital signal.

Phased array radar - Radar using many antenna elements that are fed out-of-phase to each other. The direction of the beam can be changed as rapidly as the phase relationships (usually less than 20 μsec). Thus, the antenna remains stationary while the beam is moved electronically. The use of many antenna elements allows for very rapid and high directivity of the beam(s) with a large peak and/or average power.

Point defense - The defense of specified geographical areas, cities, and vital installations. One distinguishing feature of point defense missiles is that their guidance information is received from radars located near the launching sites.

Polarization - The direction of an electrical field is considered the direction of polarization. When a half-wave dipole antenna is horizontally oriented, the emitted wave is horizontally polarized. A vertical polarized wave is emitted when the antenna is erected vertically.

Polarization diversity - The variation of polarization (such as horizontal, vertical, circular, or elliptical for radar use) either simultaneously or singularly.

Power management - Generally classified methods to optimize all related EW activities and facilities-usually in reference to a coordinate, optimized EW suite.

PPI scope - A radar display yielding range and azimuth information via an intensity modulated display and a circular sweep of a radial line. The radar is located at the center of the display.

PRF jitter - PRF rapidly varied at a random rate so that false targets appear to jitter, or appear fuzzy, on the scope. An alternative to jittered PRF is to momentarily change the PRF. This causes the false targets to change their position on the scope.

PRF stagger - The technique of switching PRF or PRI to different values on a pulse-to-pulse basis such that the various intervals follow a regular pattern. This is useful in compensating for "blind speeds" in pulsed MTI radars. Interpulse intervals differ but follow a regular pattern.

Pseudo noise - A modulation technique resulting in low signal selectability and decreased vulnerability to jamming.

Pulse compression - A scheme whereby a specifically modulated, medium power long pulse is stretched and transmitted. The returned pulse is compressed in the receiver demodulation process to obtain the advantage of narrow pulse operation. Long pulses provide the advantage of long range detection and short (compressed) pulses provide the advantage of better resolution and accuracy. This technique uses matched filter techniques for discriminating against signals that do not correspond to the transmitted code. It permits an increase in average transmitted power (without an increase in peak power) with no loss in range resolution. Pulse compression is a special form of pulse coding and correlation.

Pulse deception jamming - Any of the many false target techniques used to deceive a pulse radar, as opposed to obscurative noise techniques.

Pulse discriminator - Device that responds only to a pulse that has a particular characteristic, such as duration or period.

Pulse Doppler radar - A highly complex radar system that employs a very high pulse repetition frequency (usually 10,000 PPS or higher) to reduce "blind speeds" and measure the Doppler frequency shift to resolve target velocity. Pulse Doppler is applied principally to radar systems requiring the detection of moving targets in a ground clutter environment. It uses pulse modulation to achieve higher peak power, greater range, less susceptibility to unfriendly detection, and enhanced range resolution.

Pulse duration - The time in microseconds that the radar set is transmitting RF energy. Generally, the greater the pulse duration, the higher the average power, but the poorer the range resolution. Also known as pulse width. More technically, it is the time interval, measured at the half-amplitude points, from the leading edge to the trailing edge of a pulse.

Pulse group - In complex pulse trains, two or more pulses that are recognizably distinct from the others.

Pulse interference suppression and blanking (PISAB) - An EP automatic interference blanking device that will blank all video signals not synchronous with the radar PRF. PISAB does not require any trigger and operates on both normal and MTI modes. It is effective against random pulse signals.

Pulse jitter - Random variation of interpulse interval.

Pulse modulation (PM) - A special case of amplitude modulation in which the carrier wave is varied at a pulsed rate. This series of pulses is generally for transmitting data. The result is a short, powerful burst of electromagnetic radiation that can be used for measuring the distance from a radar set to a target.

Pulse modulated jamming - Use of jamming pulses of various widths and repetition rates.

Pulse position modulation - Modulation by variation of the interval that elapses between the pulse to be modulated in a group of pulses and a synchronizing pulse, usually the first pulse of the group.

Pulse repetition frequency (PRF) - The rate at which pulses or pulse groups are transmitted from a radar set. Generally, PRF is the number of pulses generated per second and is expressed in hertz (Hz).

Pulse repetition interval (PRI) - The interval of time between two transmitted radar pulses, usually expressed in microseconds. PRI is the inverse of PRF.

Pulse recurrence time (PRT) - Time elapsing between the start of one transmitted pulse and the start of the next. It is the reciprocal of PRF.

Pulse recurrence time (PRT) agility - Ability of the radar set to vary the number of pulses per set.

Pulse width (PW) - See Pulse Duration.

Pulse width discrimination (PWD) - An EP technique used to discriminate against received pulses that do not have the same duration as the radar transmitted pulse. PWD is used in eliminating the effects of pulse type interference when the interference pulses are not the same length as the real radar pulse. Since this circuit generates a blanking gate that shuts off the receiver whenever a pulse of improper length is sensed, loss of valid target data can result. PWD offers good discrimination against long-pulse jamming and jamming signals employing low frequency noise modulation. It affords little or no discrimination against short pulses and HF noise modulations.

Pulse width discriminator - A device that measures the pulse length of video signals and passes only those pulses whose time duration tails into some predetermined design tolerances. A pulse width discriminator will generally provide some gain against barrage jamming, similar to that of a matched filter in the video.

“Rabbits” - Interference from another radar on or near the frequency of the receiving radar. Shows on the indicators as interference at the rate of the relative PRF of the interfering radar.

Radar - From Radio Detection And Ranging. A device used to detect a distant target and determine and display its relative direction (azimuth) and its relative distance (range).

Radar absorbent material (RAM) - Material used as a radar camouflage device to reduce the echo area of an object.

Radar beacon - A receiver-transmitter combination that sends out a coded signal when triggered by the proper type of pulse enabling determination of range and bearing information by the interrogating station or aircraft.

Radar cross section - The equivalent area intercepted by a radiated signal and, if scattered uniformly in all directions, produces an echo at the radar receiver equal to that of the target. Typical radar cross sections of aircraft vary from one to over 1,000 square meters. The RCS of ships may exceed 10,000 square meters.

Radar definition - The accuracy with which a radar obtains target information such as range, azimuth, or elevation.

Radar homing - Homing on the source of a radar beam.

Radar homing and warning (RHAW) - Typically consists of an airborne, wideband video receiver designed to intercept, identify, and display the direction to pulse-type emitters.

Radar resolution - A measure of a radar's ability to separate targets that are close together in some aspect of range, azimuth, or elevation into individual returns.

Radar warning receiver (RWR) - A receiver onboard an aircraft that analyzes the hostile radar environment and determines radar threat by type, frequency, relative bearing, and relative distance. The threat is displayed to the aircrew by means of display lights, video symbols, and aural tones.

Radio frequency (RF) - Electromagnetic energy radiated at some frequency.

Radio frequency interference - An unintentional interfering signal capable of being propagated into electronic equipment, usually derived from sources outside the system.

Railing –

1. Pertains to radar pulse jamming at high recurrence rates (50 to 150 kHz). It results in an image on a radar indicator resembling fence railing.

2. The name given to that pattern produced on an "A" scope by CW modulated with a high frequency signal. Railings appear as a series of vertical lines resembling target echoes along the baseline.

Random modulation CW jamming - The use of a nonperiodic function to an AM, FM, or AM/FM CW carrier. The effects produced on a radar are similar to those produced by DINA.

Random noise - Electromagnetic energy having no particular modulation or pattern. May be generated by either natural atmospheric phenomena or by electromagnetic radiation devices.

Random pulse jamming - The technique by which a pulse transmission system is pulsed irregularly by random noise signals.

Range - The distance from one object to another.

Range gate - A gate voltage used to select radar echoes from a very short range interval.

Range gate capture - A jamming technique using a spoofer radar transmitter to produce a false target echo that can make a target tracking radar move off the real target and follow the false one.

Range gate pull-in (RGPI) - Same as range gate pull-off (see next text entry) except that the deceptive pulse is transmitted before the radar pulse is received. This is accomplished by digital storage of the pulse repetition period which must be extremely stable.

Range gate pull-off (RGPO) - A deception technique used against pulse tracking radars using range gates. The jammer initially repeats the skin echo with minimum time delay at a high power to capture the AGC circuitry. The delay is progressively increased, forcing the tracking gates to be pulled away ("walked off") from the target echo. Frequency memory loops (FML) or transponders provide the variable delay.

Range gate tracker - A radar system that tracks targets in range by measuring the elapsed time from the transmitted pulse to the echo return.

Range tracking - Pulse radars measure the time difference between radar pulse transmission and echo reception. The range gate is positioned at a range where the target is expected. The receiver is blanked off except during the period where the range gate is positioned. Range tracking may occur at the leading edge of the return pulse or between ON and OFF gates.

Recovery time (RT) - The time that a radar requires to "get ready" to receive after a pulse is sent out. Time is required to get ready because the high-powered pulse tends to fill the sensitive radar receiver with RF energy, which prevents target returns from being seen. A short "damping out" period occurs after the pulse width during which time the RF energy dissipates, allowing weak target echoes to once again be detected.

Rectifier - A device (such as a vacuum tube) for converting alternating current (AC) into direct current (DC).

Repeater - A receiver-transmitter combination that amplifies the received signal and retransmits it.

Repeater jammer - Equipment used to confuse or deceive the enemy by causing his equipment to present false information. This is done by a system that intercepts and reradiates a signal on the frequency of the enemy equipment. The reradiated signal is modified to present erroneous data on azimuth, range, number of targets, etc.

Repeater jamming - Interception and reradiation of a signal with the reradiation of a signal being modified to give erroneous azimuth, range, velocity, acceleration, or number of targets.

Resolution - The ability of a system to distinguish between two adjacent objects and to display them separately.

Resolution cell - The minimum volume in space in which a radar can discriminate between targets. It is determined numerically-for a conventional radar-by the width of the beam in angle, the pulse width in range, and the velocity gate width in speed.

Response noise jammer - A repeater jammer that transmits spot noise over the received radar frequency even in the case of radar frequency agility.

Ringing - The undesired oscillation or triggering of an electrical device by its own transmitter.

Rope - An element of chaff consisting of a long roll of metallic foil or wire that is designed for broad low frequency response. (See Chaff.)

Rope chaff - Chaff that contains one or more rope elements. (See Chaff.)

Rotating polarization - The result of a rotating feed. This should not be confused with circular polarization where the electric field rotates about the axis of propagation at the radar frequency.

SAM - Surface-to-air missile.

Saturating signal - In radar, a signal of an amplitude greater than the dynamic range of the receiving system.

Sawtooth modulated jamming - Electronic countermeasure technique when a high-level jamming signal is transmitted, thus causing large AGC voltages to be developed at the radar receiver that, in turn, causes the target return and receiver noise to completely disappear.

Scan - The process of directing a beam of RF energy successively over a given region, or the corresponding process in reception.

Scan interval - The time interval from the peak of one mainlobe in a scan pattern to the peak of the next mainlobe.

Scan period - The time period of basic scan types (except conical and lobe switching) or the period of the lowest repetitive cycle of complex scan combinations. The basic unit of measurement is degrees/mils per second or seconds per cycle.

Scan rate modulation - Modulation of the deception jamming signal with one or more frequencies that are related to the scan rate of the radar.

Scan type - The path made in space by a point on the radar beam, for example, circular, helical, conical, spiral, or sector.

Search –

1. A term applied to that phase of radar operation when the lobe, or beam of radiated energy, is directed in such a way to search for targets in the area.

2. A systematic examination of space to locate and identify targets of interest.

Sector scan - A scan in which the antenna sweeps back and forth through a selected angle.

Selective identification feature (SIF) - A capability which, when added to the basic IFF system, provides the means to transmit, receive, and display selected coded replies.

Self-protection jamming - Jamming to protect the vehicle upon which the jammer is deployed.

Semiactive radar homing - Semiactive homing guidance combines principles from both the beam rider and the active radar homing missile. Track on the target is established by the AI's radar; the missile is launched when the target comes within its effective range. During missile flight, the AI maintains track on the target. Radar returns from the target are received by the missile. Guidance commands are generated within the missile from the radar returns.

Sensitivity time control (STC) - Programmed variation of the gain (sensitivity) of a radar receiver as a function of time within each pulse repetition interval or observation time. STC prevents overloading of the receiver by strong echoes from targets or clutter at close ranges. STC reduces the gain of the radar receiver for detection of close-in targets. It is particularly effective in removing close-in clutter and strong nearby signals. STC also refers to a circuit that reduces the gain of the radar receiver immediately following the transmission of the radar pulse so that the receiver is not saturated by strong radar returns from nearby objects.

Serrodyne - A method of “pulling off” the velocity gate of a Doppler radar by using a voltage-controlled phase shifter, usually a traveling wave tube (TWT). This introduces a frequency shift from zero to some maximum value, pulling the Doppler tracking gate away from the skin echo. The phase shift is usually accomplished by modulating the TWT's helix voltages.

Sidelobe - Part of the beam from an antenna, other than the mainlobe. Sidelobe gain is usually less than mainlobe gain. Given that the mainlobe radiates most of the power at zero degrees azimuth, sidelobes inherently radiate significant power in the direction of +20°, 90°, and 150° relative to the mainlobe.

Sidelobe blanking (SLB) - A device that employs an auxiliary wide angle antenna and receiver to sense whether a received pulse originates in the sidelobe region of the main antenna and to gate it from the output signal if it does. This technique uses an omnidirectional antenna and compares relative signal strength between the omni and the radar antenna. The omnichannel (plus receiver) has slightly more gain than the sidelobes of the normal channel, but less gain than the main beam. Any signal that is greater in the omnichannel must have been received from a sidelobe and is blanked. This technique is effective in removing spoofed signals with a duty cycle up to 50%.

Sidelobe canceller (SLC) - A device that employs one or more auxiliary antennas and receivers to allow linear subtraction of interfering signals from the desired output if they are sensed to originate in the sidelobe of the main antenna. This technique employs the same antenna and receiver configurations of the SLB, except that a gain matching and canceling process takes place. Extraneous signals entering the sidelobes of the main antenna are canceled while the targets remain. This type of system exhibits cancellation roughly 20 dB against a single noise jammer. With multiple jammers at various azimuths, the performance of this device rapidly deteriorates.

Sidelobe jamming - Jamming through a sidelobe of the receiving antenna in an attempt to obliterate the desired signal received through the mainlobe of the receiving antenna at fixed points.

Sidelobe suppression - The suppression of that portion of the beam from a radar antenna other than the mainlobe.

Sidewinder - A solid-propellant, air-to-air missile with nonnuclear warhead and an infrared guidance system. Designated as the AIM-9 missile.

Signal intelligence (SIGINT) - Intelligence derived from the interception of enemy communications and noncommunication signals. A generic term that includes both COMINT and ELINT.

Signal-to-jamming ratio (S/J) - The ratio of the signal power to the jamming power or intentional interference at some point in the system. This ratio is often expressed in decibels.

Signal-to-noise ratio (S/N) - Ratio of the power of the signal to the power of the noise.

Signature - The set of parameters that describes the characteristics of a radar target or an RF emitter and distinguishes one emitter from another. Signature

parameters include the RF of the carrier, the modulation characteristics (typically the pulse modulation code), and the scan pattern.

Sine-wave modulated jamming - Jamming signal produced by modulating a CW signal with one or more sine waves.

Single beam blanking (SBB) - Used by phased array radars as an alternative method of Beam to Beam Correlation (BBC). It is effective, to some degree, against many multiple target generators and swept frequency jammers. Because of overlapping beamwidths, a target signal will appear in more than one beam as the beams are scanned past a true target. When jamming signals are transmitted along one beam, that beam is blanked by the radar receiver.

Skirt jamming - Jamming that places the signal somewhat off the radar center frequency but within the IF skirts. The technique will degrade the tracking accuracy of some monopulse radars; however, good radar design makes the technique ineffective.

Smart jamming - Selective jamming of threat radars with an optimized modulation signal at a correct time. Power management techniques are used to control smart jamming systems.

Smith modulation - Deceptive technique that operates on the servo loop of the victim's radar. Two RF carriers are transmitted with a few cycles difference.

Spiral scan - A pencil beam scan which moves around a central axis describing the surface of a cone in an ever-increasing cone size.

Spot jamming - Narrow frequency band jamming concentrated against radar at a particular frequency. The jamming bandwidth is comparable to the radar bandpass. Used to deny range and sometimes angle information.

Spread spectrum - Use of broader frequency bandwidths than normally required to transmit information, in order to gain advantages in interference rejection (anti-jam), message privacy/security, low probability of intercept (LPI), deny frequency resolution, multiplexing of more than one message in the same bandwidth, or high resolution range measurement.

Spurious radiation - Emissions from a radio transmitter at frequencies outside its assigned or intended emission frequency. Spurious emission includes harmonic emission, parasitic emission, and intermodulation products, but excludes emissions in the immediate vicinity of the necessary band that are a result of the modulation process for the transmission of information.

Spurious response –

1. Any response, other than the desired response, of an electric transducer or device.

2. A term used in electronic warfare to describe the undesirable signal images in the intercept receiver resulting from the mixing of the intercepted signal with harmonics of the local oscillators in the receiver.

Staggered PRF - A technique of using more than one PRF to reduce the blind speeds associated with MTI radars. PRF switching occurs on a pulse-to-pulse basis.

Stand-off jammer (SOJ) - A powerful jammer that remains at a safe range while providing jamming coverage (masking) for the attacking elements. The Navy EA-6B is an example of a SOJ aircraft.

Stand-off jamming (SOJ) - A jamming aircraft that orbits outside the maximum range of the SAM threat. As the attack package penetrates, the jamming aircraft directs jamming against all significant radars in the area.

Stream - Dispensing of chaff (solid/random interval/bursts).

Stream chaff - Operational technique of dropping large quantities of chaff for a continuous period of time. This results in a “ribbon” or “stream” of returns many miles in length on victim radar scopes. The penetrating attack package can then use the resulting chaff corridor to mask their penetration.

Super high frequency (SHF) - Frequencies from 3 to 30 GHz.

Support jamming - A tactic by which aircraft carrying electronic jamming equipment orbit at a safe distance from the enemy threat defenses or fly escort with the strike force for the primary purpose of screening them from the threat radars.

Suppression of enemy air defenses (SEAD) - That activity which neutralizes, destroys, or temporarily degrades enemy air defense systems in a specific area by using physical attack, deception, and/or electronic warfare.

Surface-to-air missile (SAM) - A missile launched from a surface launcher at a target above the surface.

Sweep jammer - Electronic jammer that sweeps a narrow band of electronic energy over a broad bandwidth.

Sweep lock-on jammer - A transmitter in which a narrowband jamming signal can be tuned over a broad frequency band and the signal locked on a particular frequency.

Swept audio - Jamming technique usually employed against conical scan-on-receive-only (COSRO) radars. The received pulses are amplified and retransmitted by the target and amplitude modulated at a frequency close to the suspected receiver antenna scan frequency.

Swept jamming - An EA technique of barrage jamming in which a CW carrier or noise source is swept over a selected bandwidth.

Swept-spot jamming - A jamming technique in which an oscillator is swept over a specific range of frequencies in the band of interest in order to be assured of exciting a receiver tuned to any frequency in that band.

Synchronized-pulse jamming - The technique of attempting to insert jamming pulses into a receiver each time the receiver gate opens.

Synchronous-pulsed jamming - A jamming technique that matches exactly the PRF of the victim's radar; then transmits multiples of the PRF. It is most effective if the jammer also matches the PW of the radar. Synchronous-pulsed jamming is easily recognized since the spacing between successive target lines is equal and each target line is the same in depth from the center outward. The width of the jammed sector is dependent upon the range of the jammer from the radar.

Synthetic aperture radar (SAR) - A high-resolution ground mapping technique in which advantage is taken of the forward motion of a coherent pulsed radar to synthesize the equivalent of a very long sidelooking array antenna from the radar returns received over a period of up to several seconds or more.

Target acquisition - The detection, identification, and location of a target in sufficient detail to permit the effective employment of weapons.

Terminal guidance –

1. The guidance applied to a guided missile between mid-course and arrival in the vicinity of the target.

2. Electronic, mechanical, visual, or other assistance given to aircraft pilots to facilitate arrival at, operation within or over, landing upon or departure from an air landing or air drop facility.

Terminal threat - The weapon systems, generally near a target, used to directly engage an aircraft in order to destroy it.

Terrain-avoidance radar - An airborne radar that provides a display of terrain ahead of low-flying aircraft to permit horizontal avoidance of obstacles.

Terrain-following radar (TFR) - An airborne radar that provides a display of terrain ahead of low-flying aircraft to permit manual control, or signals for automatic control to maintain constant altitude above the ground.

Thermal crossover - The natural phenomenon which normally occurs twice daily when temperature conditions are such that there is a loss of contrast between two adjacent objects on infrared imagery.

Threshold - The minimum value of a signal that can be detected by a system or sensor under consideration.

Time-of-arrival (TOA) - A method of locating a distant pulse emitter by measuring the difference in the time-of-arrival of its pulses at three separate locations. This method is also called Inverse LORAN.

Track –

1. A series of related contacts displayed on a plotting board.

2. To display or record the successive positions of a moving object.

3. To lock onto a point of radiation and obtain guidance from it.

4. To keep a gun properly aimed, or to continuously point a target-locating instrument at a moving target.

5. The actual path of an aircraft above, or a ship on, the surface of the earth. The course is the path that is planned; the track is the path that is taken.

Tracking - The continuous monitoring of range, velocity, or position of a target in space from a reference position. This is accomplished via radar and/or optical means.

Tracking radar - A radar that measures the range, azimuth, elevation, and/or velocity of the target and provides data that may be used by the fire control computer to determine the target path and predict its future position.

Track-on-jam - A method of passive target tracking using the jamming signal emitted by the target.

Track-while-scan (TWS) radar - Although it is not really a tracking radar in the true sense of the word, it does provide complete and accurate position information for missile guidance by using two separate beams produced by two separate antennas on two different frequencies. The system uses electronic computer techniques whereby raw data are used to track an assigned target, compute target velocity, and predict its future position while maintaining normal sector scan.

Transceiver - A combined radio transmitter and receiver in which some circuits other than those of the power supply are common to both transmitter and receiver, and not providing for simultaneous transmission and reception.

Transponder - A transmitter-receiver capable of accepting the electronic challenge of an interrogator and automatically transmitting an appropriate reply.

Traveling wave tube (TWT) - An electron tube in which a beam of electrons interacts continuously with a guided electromagnetic wave to produce amplification at microwave frequencies. A TWT capable of providing high amplification (60 dB) in frequencies over several octaves without adjustment. TWTs are classified by:

1. Power Output:

Low (less than one watt)
High (more than 10 watts)
Intermediate (one to 10 watts)

2. **Noise Characteristics:**

    Low (less than 20 dB)

3. **Operating Mode:**

    Pulse
    CW
    Dual Mode (pulse or CW)

4. **Method of Focusing:**

    PPM (periodic permanent magnet)
    Solenoid (single electromagnet)
    Coupled Cavity

5. **Terms associated with TWT operation:**

    a. **Gain Compression:** Change in amplification a device provides as it operates near saturation.

    b. **Overdrive:** TWT is operating beyond the point of saturation.

    c. **Saturation Gain:** Ratio of output power to input power when the device is being driven at maximum output.

    d. **Serrodyne:** Linear translation of the phase of a signal on the helix of a TWT by a linear sawtooth waveform that enables the TWT to operate as a single sideband frequency translator.

    e. **Small Signal Gain:** Ratio of output power to input power when the TWT is operating linearly.

Two-signal jamming (also called straddle jamming) - Jamming whereby two signals are transmitted on two RF frequencies slightly separated. Effective against certain types of radar where receiver bandwidth is narrow enough to defeat noise jamming.

Ultra high frequency (UHF) - Frequencies from 300 to 3,000 MHz.

Unit prefixes - Prefixes used to indicate scientific units:

Multiple	Prefix	Symbol
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k
10^{2}	hecto	h
10^{1}	deka	da
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	µ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

Unmodulated CW jamming - The transmission of a high power carrier frequency that causes an overload effect to occur in the radar receiver. Used against bandpass radars that have a limited tuning capability. Unmodulated CW jamming can be identified by a blackening of the scope background (no video present) in a wedge-shaped sector, or by a solid brightening of a wedge or sector, normally exceeding one bandwidth.

Uplink - The missile guidance signal that passes command guidance intelligence from the site to the missile.

Velocity gate pull-off (VGPO) - Method of capturing the velocity gate of a Doppler radar and moving it away from the skin echo. Similar to the RGPO, but used against Doppler radar systems. The target Doppler frequency, which is amplified and retransmitted, is shifted in frequency to provide an apparent rate change or Doppler shift.

Very high frequency (VHF) - Frequencies from 30 to 300 MHz.

Very low frequency (VLF) - Frequencies from 3 to 30 kHz.

Video frequency –

1. A band of frequencies extending from approximately 100 Hz to several MHz.

2. The frequency of the voltage resulting from television scanning. Range from zero to 4 MHz or more.

Warning receiver - A receiver with the primary function of warning the user that his unit is being illuminated by an electromagnetic signal of interest.

Wideband constant false alarm rate receiver (WB-1) - Used against individual or combinations of rapidly swept FM-CW, noise, or CW jamming. This mode has nonlinear limiting and gives poor resolution of overlapping targets.

Window –

1. Strips of frequency-cut metal foil, wire, or bars that may be dropped from aircraft or missiles, or expelled from shells or rockets as a radar countermeasure. A confusion reflector.

2. A passive radar deception or confusion device; usually consisting of some metallic structure, designed in size and shape to effectively reflect impinging signals, to simulate a true target.

3. British name for chaff.

Wobbulation - A periodic and usually slow variation of a parameter (such as frequency, amplitude, or period) about a central value. A low frequency modulation of a jamming carrier that appears on a radar display as a “wobbling” target. This effect is an undesired result of an angle deception or time-varying barrage technique.

Wooden round - An ordnance round (shell, missile, etc.) requiring no maintenance or preparation time prior to loading for firing.

Kenneth Vito Zichi

3018 Moyer Road
Williamston, MI 48895-9566

The ABC's of Propagation

As children, a lot of us thought of radio as a "magic" box. Just by turning a switch, voices and music from across the city, or state, would come drifting into our room. By the time most of us discover shortwave, we've outgrown the idea that supernatural forces were at work in our radios. Still, we remain awestruck at the idea that someone in Australia or South Africa could whisper into a microphone and we could hear it half a world away.

This decidedly non-supernatural method by which radio waves travel from point "A" to point "B" is called *propagation*. It is a concept basic to all radio communications. And despite popular myth, you do not need a Ph.D in physics to understand it.

Radio is a form of "electromagnetic radiation." That is a fancy way of saying that radio waves have both electrical and magnetic properties. If they are energetic enough, they can leave the confines of a conductor and travel freely through the air, a vacuum (like outer space), or other similar media. In this way, radio waves act much like light, which is another form of electromagnetic radiation, albeit higher in frequency than radio. In fact, light provides a very good analogy for explaining how radio waves propagate.

Types of Propagation

Like light, radio waves can propagate (travel) in several different ways. We've sketched out some of the most common types of propagation, how they work and on what band(s) you're likely to encounter them.

GROUND WAVE

The most common and reliable way for radio waves to travel is called "ground wave." This is a form of propagation that is common to *all* radio bands from VLF (Very Low Frequency) to radar. Simply put, when a radio wave travels directly from the transmitting antenna to the receiving antenna without bending, reflecting, or otherwise being diverted, it is traveling by ground wave.

Using the light analogy, imagine a lighthouse shining in the distance. So long as there is nothing between you and the light -- and you are not so far away that the curvature of the earth would place the lighthouse below the horizon -- you can see it. (See figure one, wave A).

You can increase the distance you can see the lighthouse by elevating either yourself (a receiver) or the light (a transmitter), but once these variables are fixed, there is a finite distance at which the light will be visible because of the earth's curvature.

Reception distance is therefore limited when using this form of propagation. However, this limitation is offset by the fact that groundwave reception is very reliable. Some examples of transmissions using ground wave reception are local AM and FM broadcasters, police communications and CB radio.

IONOSPHERIC ("Skip")

Another very common form of propagation is ionospheric propagation, which is also known as "skip" or "skywave." This common phenomenon of long distance communications on frequencies between .5 and 30 MHz is the result of the bending of radio waves in the earth's *ionosphere*.

The ionosphere is several regions of charged oxygen (and other gases) located some 60 and 200 miles above the ground. These layers have a number of unique properties, not the least is their ability to *refract* radio waves that enter it back toward the earth's surface.

To use the light analogy again, it is as if someone placed a huge mirror in the sky, and radio waves, which would ordinarily shoot off into space, are re-directed back to the earth's surface. (See figure one, wave B.)

Actually, this analogy is not quite correct since the ionosphere *refracts* radio waves and a mirror *reflects* light. At any rate, if a radio wave of the right frequency enters the ionosphere, it will be bent back to earth. The result is reception of the signal hundreds of miles away from the

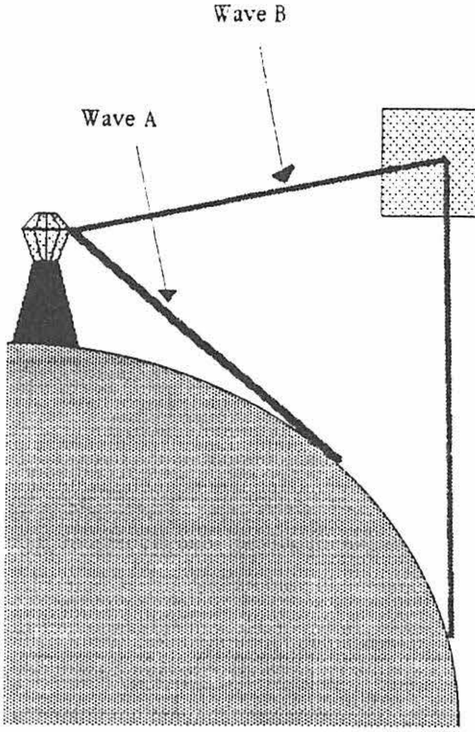

Figure One

Like the light from this lighthouse, radio waves travel only in straight lines. As a result of the curvature of the Earth, they can therefore only travel a finite distance (Wave "A").

Reception past the horizon can take place, but only if the radio waves are refracted or reflected off some surface, akin to someone in space holding amirror to allow you to see the lighthouse, as illustrated by wave "B."

Kenneth Vito Zichi

3018 Moyer Road
Williamston, MI 48895-9566

transmitter and well beyond the range possible by ground wave propagation.

Unfortunately, the ionosphere is also capable of absorbing or blocking reception of radio waves. Depending on the density of the ionosphere at a given time, the angle of the wave entering the ionosphere, and the frequency of the wave, long distance reception will be either enhanced or destroyed.

Too, the gap between where ground wave reception ends and the spot where the skywave signal hits the earth is also an area of non-reception. It is called the *skip zone* and the phenomenon is illustrated in figure two, with "S" denoting the skip zone.

Depending on the time of day and solar conditions (solar radiation is one of the major factors that cause gases in the upper atmosphere to ionize and form the ionosphere), different layers of ionized gas of varying intensity will be formed. Each layer affects radio waves slightly differently.

Lower frequency shortwave signals (eg. 2 through 9 MHz) will be absorbed if they pass through the ionosphere at low angles during the daylight hours. During this part of the day, there are more layers to the ionosphere and these layers are thicker. Thus, during the day, low frequencies do not propagate as well as higher frequencies (eg 11 through 25 MHz). Similarly, at night, the ionosphere is thinner and the low frequency waves are refracted without being absorbed. Then, the higher frequencies pass freely into space.

MULTI HOPS

Radio signals making the journey from transmitter to ionosphere to earth are said to have completed a "one hop" circuit. However, this is not necessarily the end of the trip. That same signal can be reflected from the earth, back up to the ionosphere once again, and down to earth, this time even further from the transmitter. If conditions are just right, a signal can go all the way around the world and double back on itself.

The "multi-hop" circuits do increase distances dramatically, but are less than single hops since they are subject to the uncertainties of the ionosphere more than once.

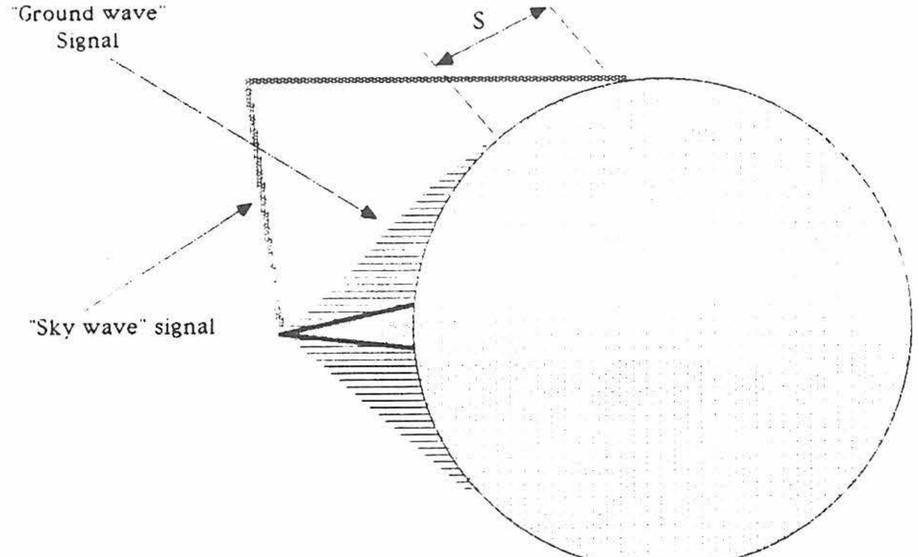

Figure Two

"S" = the "skip zone": the area between ground wave and sky wave propagation where reception of radio signals is difficult at best.

Drawing is not to scale! Actual distances for ground wave reception on the Earth would be about 300 km depending on effective height of transmitter and receiver antennae, not 1/4 of the planet's surface as indicated here!

EXOTIC STUFF

Lastly, there are numerous, less common ways that a radio wave can get from point A to point B.

One of these is called *sporadic E.* If it is sufficiently charged, the "E Layer" of the ionosphere is capable of refracting relatively high frequency radio waves. Occasionally, there will be "pockets" of highly charged particles within the E layer which allows for unusual reception.

This so called "sporadic E reception" is just that: sporadic. It is fairly common around local dusk and can boost reception of FM and TV signals well beyond their intended range. Sporadic E is also most common during the late spring and early summer in temperate latitudes but it is possible at any time of the year.

Moonbounce is a form of reception that's fascinating but rarely used now that man-made satellites populate the skies over planet earth. Pioneered by radio amateurs, it is a form of propagation in which signals are intentionally reflected off the moon's surface and back to earth.

Tropospheric Ducting is most common at relatively high frequencies and affects transmissions heard on FM, TV and scanners. It results when radio waves follow patterns of high and low pressure in the atmosphere (much the way weather patterns do). It allows for reception of signals slightly beyond the range you might otherwise expect.

Lastly, *Meteor Scatter* is probably one of the most unusual forms of propagation. Here, radio waves are bounced off the ionized trails of meteors. Like sporadic E reception, this results from a region of the atmosphere becomming supercharged, and thus refracting higher frequencies than normal back to earth. Reception of VHF signals using metero scatter can be very brief -- lasting in the order of seconds.

In short, there are a plethora of exotic reception techniques, but the ones outlined in the first part of this article are the ones you are most likely to run into, and should aid you in planning your listening to best enjoy the hobby.

■

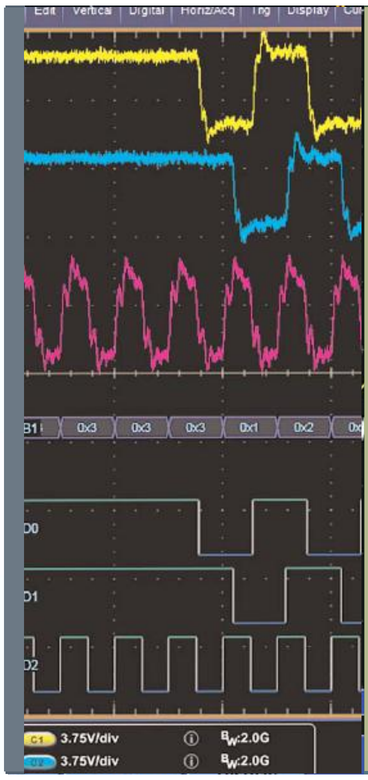
Edit
Vertical
Digital
Horiz/Acq
Trig
Display
B1
0x3
0x3
0x3
0x1
0x2
D0
D1
D2
C1 3.75V/div
C2 3.75V/div
BW:2.0G
BW:2.0G

XYZs of Oscilloscopes

Primer

Tektronix®

Table of Contents

Introduction 4

Signal Integrity . . . . 5 - 6

- The Significance of Signal Integrity 5
- Why is Signal Integrity a Problem? 5
- Viewing the Analog Origins of Digital Signals 6

The Oscilloscope . . . . 7 - 12

- Understanding Waveforms & Waveform Measurements . .7
- Types of Waves 8
  - Sine Waves 9
  - Square and Rectangular Waves 9
  - Sawtooth and Triangle Waves 9
  - Step and Pulse Shapes 9
  - Periodic and Non-periodic Signals 9
  - Synchronous and Asynchronous Signals 9
  - Complex Waves 10
- Waveform Measurements 11
  - Frequency and Period 11
  - Voltage 11
  - Amplitude 11
  - Phase 11
  - Waveform Measurements with Digital Oscilloscopes 12

Types of Oscilloscopes .13 - 17

- Digital Storage Oscilloscopes 13
- Digital Phosphor Oscilloscopes 15
- Mixed Signal Oscilloscopes 16
- Digital Sampling Oscilloscopes 17

The Systems and Controls of an Oscilloscope .18 - 31

- Vertical System and Controls 19
  - Position and Volts per Division 19
  - Input Coupling 19
  - Bandwidth Limit 19
  - Bandwidth Enhancement 20
- Horizontal System and Controls 20
  - Acquisition Controls 20
  - Acquisition Modes 20
  - Types of Acquisition Modes 21
  - Starting and Stopping the Acquisition System 21
  - Sampling 22
  - Sampling Controls 22
  - Real-time Sampling Method 22
  - Equivalent-time Sampling Method 24
  - Position and Seconds per Division 26
  - Time Base Selections 26
  - Zoom/Pan 26
  - Search 26
  - XY Mode 26
  - Z Axis 26
  - XYZ Mode 26
- Trigger System and Controls 27
  - Trigger Position 29
  - Trigger Level and Slope 29
  - Trigger Sources 29
  - Trigger Modes 30
  - Trigger Coupling 30
  - Trigger Holdoff 30
- Display System and Controls 31
- Other Oscilloscope Controls 31
  - Math and Measurement Operations 31
  - Digital Timing and State Acquisitions 31

The Complete Measurement System . . . . 32 - 34

Probes 32
Passive Probes 33
Active and Differential Probes 34
Logicl Probes 34
Specialty Probes 35
Probe Accessories 35

Performance Terms and Considerations . . . . . 35 - 42

Bandwidth 35
Rise Time 36
Sample Rate 37
Waveform Capture Rate 38
Record Length 38
Triggering Capabilities 39
Effective Bits 39
Frequency Response 39
Vertical Sensitivity 39
Sweep Speed 39
Gain Accuracy 39
Horizontal Accuracy (Time Base) 39
Vertical Resolution (Analog-to-Digital Converter) 39
Timing Resolution (MSO) 40
Connectivity 40
Expandability 40
Ease-of-use 42

Operating the Oscilloscope . . . . 43 - 45

Proper Grounding 43
Setting the Controls 43
Calibrating the Instrument 44
Connecting the Probes 44
Compensating the Probes 44

Oscilloscope Measurement Techniques . . . . . . 46 - 48

Voltage Measurements 46
Time and Frequency Measurements 47
Pulse Width and Rise Time Measurements 47
Phase Shift Measurements 48
Other Measurement Techniques 48

Written Exercises . . . . 49 - 54

Part I
  A. Vocabulary Exercises 49
  B. Application Exercises 50

Part II
  A. Vocabulary Exercises 51
  B. Application Exercises 52

Answer Key 54

Glossary . . . . 55 - 58

Introduction

Nature moves in the form of a sine wave, be it an ocean wave, earthquake, sonic boom, explosion, sound through air, or the natural frequency of a body in motion. Energy, vibrating particles and other invisible forces pervade our physical universe. Even light – part particle, part wave – has a fundamental frequency, which can be observed as color.

Sensors can convert these forces into electrical signals that you can observe and study with an oscilloscope. Oscilloscopes enable scientists, engineers, technicians, educators and others to "see" events that change over time.

Oscilloscopes are indispensable tools for anyone designing, manufacturing or repairing electronic equipment. In today's fast-paced world, engineers need the best tools available to solve their measurement challenges quickly and accurately. As the eyes of the engineer, oscilloscopes are the key to meeting today's demanding measurement challenges.

The usefulness of an oscilloscope is not limited to the world of electronics. With the proper sensor, an oscilloscope can measure all kinds of phenomena. A sensor is a device that creates an electrical signal in response to physical stimuli, such as sound, mechanical stress, pressure, light, or heat. A microphone is a sensor that converts sound into an electrical signal. Figure 1 shows an example of scientific data that can be gathered by an oscilloscope.

Oscilloscopes are used by everyone from physicists to repair technicians. An automotive engineer uses an oscilloscope to correlate analog data from sensors with serial data from the engine control unit. A medical researcher uses an oscillscope to measure brain waves. The possibilities are endless.

The concepts presented in this primer will provide you with a good starting point in understanding oscilloscope basics and operation.

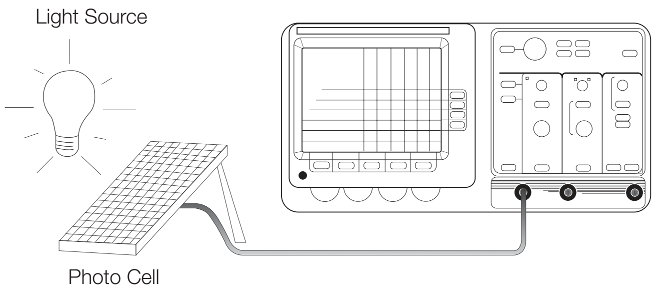

Figure 1. An example of scientific data gathered by an oscilloscope.

The glossary in the back of this primer will give you definitions of unfamiliar terms. The vocabulary and multiple-choice written exercises on oscilloscope theory and controls make this primer a useful classroom aid. No mathematical or electronics knowledge is necessary.

After reading this primer, you will be able to:

- Describe how oscilloscopes work
- Describe the differences between various oscilloscopes
- Describe electrical waveform types
- Understand basic oscilloscope controls
- Take simple measurements

The manual provided with your oscilloscope will give you more specific information about how to use the oscilloscope in your work. Some oscilloscope manufacturers also provide a multitude of application notes to help you optimize the oscilloscope for your application-specific measurements.

Should you need additional assistance, or have any comments or questions about the material in this primer, simply contact your Tektronix representative, or visit **www.tektronix.com**.

Signal Integrity

The Significance of Signal Integrity

The key to any good oscilloscope system is its ability to accurately reconstruct a waveform – referred to as signal integrity. An oscilloscope is analogous to a camera that captures signal images that we can then observe and interpret. Two key issues lie at the heart of signal integrity.

- When you take a picture, is it an accurate picture of what actually happened?
- Is the picture clear or fuzzy?
- How many of those accurate pictures can you take per second?

Taken together, the different systems and performance capabilities of an oscilloscope contribute to its ability to deliver the highest signal integrity possible. Probes also affect the signal integrity of a measurement system.

Signal integrity impacts many electronic design disciplines. But until a few years ago, it wasn't much of a problem for digital designers. They could rely on their logic designs to act like the Boolean circuits they were. Noisy, indeterminate signals were something that occurred in high-speed designs – something for RF designers to worry about. Digital systems switched slowly and signals stabilized predictably.

Processor clock rates have since multiplied by orders of magnitude. Computer applications such as 3D graphics, video and server I/O demand vast bandwidth. Much of today's telecommunications equipment is digitally based, and similarly requires massive bandwidth. So too does digital high-definition TV. The current crop of microprocessor devices handles data at rates up to 2, 3 and even 5 GS/s (gigasamples per second), while some DDR3 memory devices use clocks in excess of 2 GHz as well as data signals with 35 ps rise times.

Importantly, speed increases have trickled down to the common IC devices used in automobiles, consumer electronics, and machine controllers, to name just a few applications.

A processor running at a 20 MHz clock rate may well have signals with rise times similar to those of an 800 MHz processor. Designers have crossed a performance threshold that means, in effect, almost every design is a high-speed design.

Without some precautionary measures, high-speed problems can creep into otherwise conventional digital designs. If a circuit is experiencing intermittent failures, or if it encounters errors at voltage and temperature extremes, chances are there are some hidden signal integrity problems. These can affect time-to-market, product reliability, EMI compliance, and more. These high-speed problems can also impact the integrity of a serial data stream in a system, requiring some method of correlating specific patterns in the data with the observed characteristics of high-speed waveforms.

Why is Signal Integrity a Problem?

Let's look at some of the specific causes of signal degradation in today's digital designs. Why are these problems so much more prevalent today than in years past?

The answer is speed. In the "slow old days," maintaining acceptable digital signal integrity meant paying attention to details like clock distribution, signal path design, noise margins, loading effects, transmission line effects, bus termination, decoupling and power distribution. All of these rules still apply, but…

Bus cycle times are up to a thousand times faster than they were 20 years ago! Transactions that once took microseconds are now measured in nanoseconds. To achieve this improvement, edge speeds too have accelerated: they are up to 100 times faster than those of two decades ago.

This is all well and good; however, certain physical realities have kept circuit board technology from keeping up the pace. The propagation time of inter-chip buses has remained almost unchanged over the decades. Geometries have shrunk, certainly, but there is still a need to provide circuit board real estate for IC devices, connectors, passive components, and of course, the bus traces themselves. This real estate adds up to distance, and distance means time – the enemy of speed.

It's important to remember that the edge speed – rise time – of a digital signal can carry much higher frequency components than its repetition rate might imply. For this reason, some designers deliberately seek IC devices with relatively "slow" rise times.

The lumped circuit model has always been the basis of most calculations used to predict signal behavior in a circuit. But when edge speeds are more than four to six times faster than the signal path delay, the simple lumped model no longer applies.

Circuit board traces just six inches long become transmission lines when driven with signals exhibiting edge rates below four to six nanoseconds, irrespective of the cycle rate. In effect, new signal paths are created. These intangible connections aren't on the schematics, but nevertheless provide a means for signals to influence one another in unpredictable ways.

Sometimes even the errors introduced by the probe/instrument combination can provide a significant contribution to the signal being measured. However, by applying the "square root of the sum of the squares" formula to the measured value, it is possible to determine whether the device under test is approaching a rise/fall time failure. In addition, recent oscilloscope tools use special filtering techniques to de-embed the measurement system's effects on the signal, displaying edge times and other signal characteristics.

At the same time, the intended signal paths don't work the way they are supposed to. Ground planes and power planes, like the signal traces described above, become inductive and act like transmission lines; power supply decoupling is far less effective. EMI goes up as faster edge speeds produce shorter wavelengths relative to the bus length. Crosstalk increases.

In addition, fast edge speeds require generally higher currents to produce them. Higher currents tend to cause ground bounce, especially on wide buses in which many signals switch at once. Moreover, higher current increases the amount of radiated magnetic energy and with it, crosstalk.

Viewing the Analog Origins of Digital Signals

What do all these characteristics have in common? They are classic analog phenomena. To solve signal integrity problems, digital designers need to step into the analog domain. And to take that step, they need tools that can show them how digital and analog signals interact.

Digital errors often have their roots in analog signal integrity problems. To track down the cause of the digital fault, it's often necessary to turn to an oscilloscope, which can display waveform details, edges and noise; can detect and display transients; and can help you precisely measure timing relationships such as setup and hold times. Modern oscilloscopes can help to simplify the troubleshooting process by triggering on specific patterns in parallel or serial data streams and displaying the analog signal that corresponds in time with a specified event.

Understanding each of the systems within your oscilloscope and how to apply them will contribute to the effective application of the oscilloscope to tackle your specific measurement challenge.

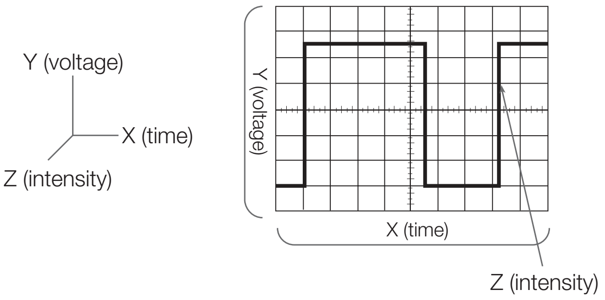
Y (voltage)
X (time)
Z (intensity)
Y (voltage)
X (time)
Z (intensity)

Figure 2. X, Y, and Z components of a displayed waveform.

The Oscilloscope

What is an oscilloscope and how does it work? This section answers these fundamental questions.

The oscilloscope is basically a graph-displaying device – it draws a graph of an electrical signal. In most applications, the graph shows how signals change over time: the vertical (Y) axis represents voltage and the horizontal (X) axis represents time. The intensity or brightness of the display is sometimes called the Z axis, as shown in Figure 2. In DPO oscilloscopes, the Z axis can be represented by color grading of the display, as seen in Figure 3.

This simple graph can tell you many things about a signal, such as:

- The time and voltage values of a signal
- The frequency of an oscillating signal
- The “moving parts” of a circuit represented by the signal
- The frequency with which a particular portion of the signal is occurring relative to other portions
- Whether or not a malfunctioning component is distorting the signal
- How much of a signal is direct current (DC) or alternating current (AC)
- How much of the signal is noise and whether the noise is changing with time

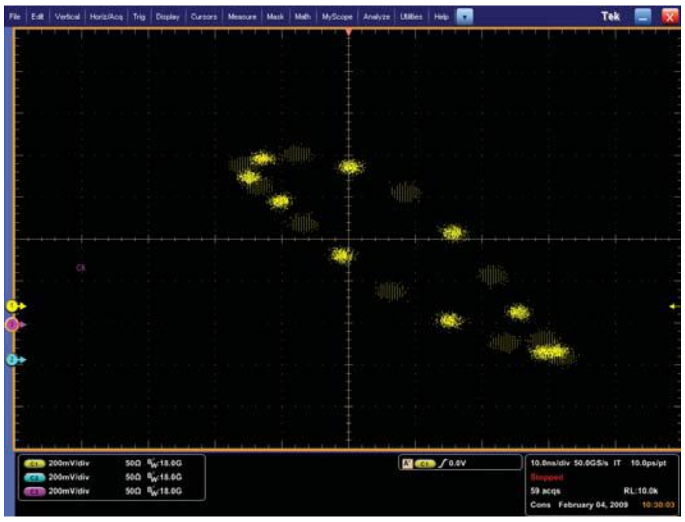
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
C1 200mV/div 50Ω BW:18.0G
C2 200mV/div 50Ω BW:18.0G
C3 200mV/div 50Ω BW:18.0G
A C1 0.0V
10.0ns/div 50.0GS/s IT 10.0ps/pt
Stopped
59 acqs RL:10.0k
Cons February 04, 2009 10:30:03

Figure 3. Two offset clock patterns with Z axis intensity grading.

Understanding Waveforms and Waveform Measurements

The generic term for a pattern that repeats over time is a wave – sound waves, brain waves, ocean waves, and voltage waves are all repetitive patterns. An oscilloscope measures voltage waves. Remember as mentioned earlier, that physical phenomena such as vibrations or temperature or electrical phenomena such as current or power can be converted to a voltage by a sensor. One cycle of a wave is the portion of the wave that repeats. A waveform is a graphic representation of a wave. A voltage waveform shows time on the horizontal axis and voltage on the vertical axis.

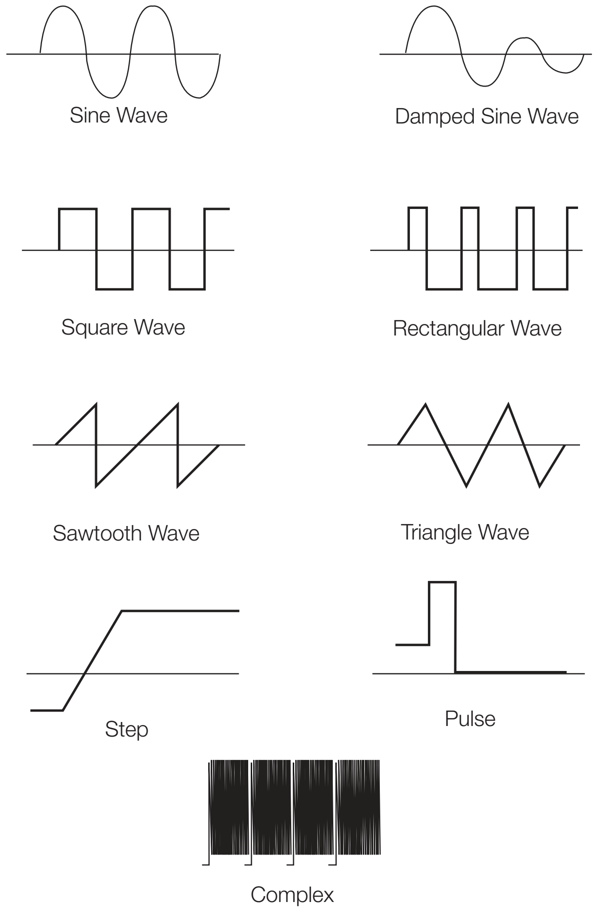
Sine Wave
Damped Sine Wave
Square Wave
Rectangular Wave
Sawtooth Wave
Triangle Wave
Step
Pulse
Complex

Figure 4. Common waveforms.

Waveform shapes reveal a great deal about a signal. Any time you see a change in the height of the waveform, you know the voltage has changed. Any time there is a flat horizontal line, you know that there is no change for that length of time. Straight, diagonal lines mean a linear change – rise or fall of voltage at a steady rate. Sharp angles on a waveform indicate sudden change. Figure 4 shows common waveforms and Figure 5 displays sources of common waveforms.

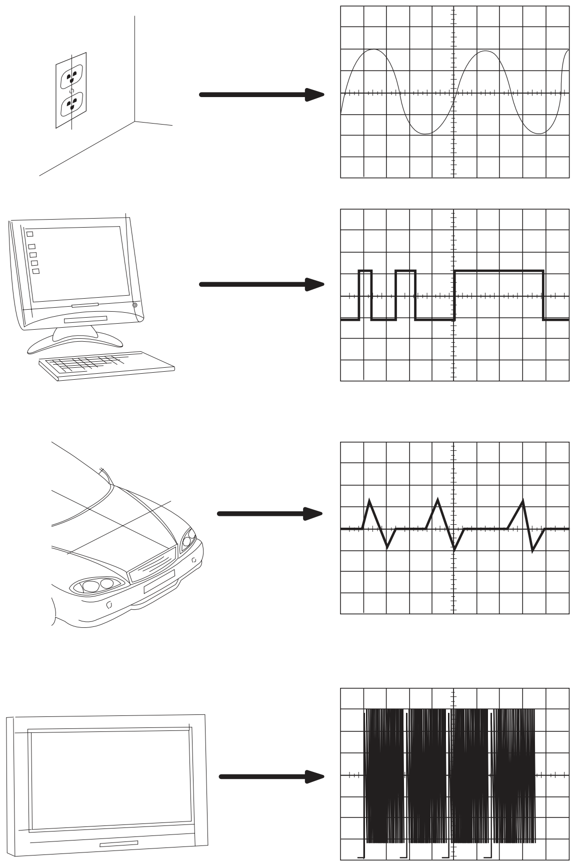

Figure 5. Sources of common waveforms.

Types of Waves

You can classify most waves into these types:

- Sine waves
- Square and rectangular waves
- Sawtooth and triangle waves
- Step and pulse shapes
- Periodic and non-periodic signals
- Synchronous and asynchronous signals
- Complex waves

Sine Waves

The sine wave is the fundamental wave shape for several reasons. It has harmonious mathematical properties – it is the same sine shape you may have studied in trigonometry class. The voltage in your wall outlet varies as a sine wave. Test signals produced by the oscillator circuit of a signal generator are often sine waves. Most AC power sources produce sine waves. (AC signifies alternating current, although the voltage alternates too. DC stands for direct current, which means a steady current and voltage, such as a battery produces.)

The damped sine wave is a special case you may see in a circuit that oscillates, but winds down over time.

Square and Rectangular Waves

The square wave is another common wave shape. Basically, a square wave is a voltage that turns on and off (or goes high and low) at regular intervals. It is a standard wave for testing amplifiers – good amplifiers increase the amplitude of a square wave with minimum distortion. Television, radio and computer circuitry often use square waves for timing signals.

The rectangular wave is like the square wave except that the high and low time intervals are not of equal length. It is particularly important when analyzing digital circuitry.

Sawtooth and Triangle Waves

Sawtooth and triangle waves result from circuits designed to control voltages linearly, such as the horizontal sweep of an analog oscilloscope or the raster scan of a television. The transitions between voltage levels of these waves change at a constant rate. These transitions are called ramps.

Step and Pulse Shapes

Signals such as steps and pulses that occur rarely, or non-periodically, are called single-shot or transient signals. A step indicates a sudden change in voltage, similar to the voltage change you would see if you turned on a power switch.

A pulse indicates sudden changes in voltage, similar to the voltage changes you would see if you turned a power switch on and then off again. A pulse might represent one bit of information traveling through a computer circuit or it might be a glitch, or defect, in a circuit. A collection of pulses traveling together creates a pulse train. Digital components in a computer communicate with each other using pulses. These pulses may be in the form of serial data stream or multiple signal lines may be used to represent a value in a parallel data bus. Pulses are also common in x-ray, radar, and communications equipment.

Periodic and Non-periodic Signals

Repetitive signals are referred to as periodic signals, while signals that constantly change are known as non-periodic signals. A still picture is analogous to a periodic signal, while a moving picture can be equated to a non-periodic signal.

Synchronous and Asynchronous Signals

When a timing relationship exists between two signals, those signals are referred to as synchronous. Clock, data and address signals inside a computer are an example of synchronous signals.

Asynchronous is a term used to describe those signals between which no timing relationship exists. Because no time correlation exists between the act of touching a key on a computer keyboard and the clock inside the computer, these are considered asynchronous.

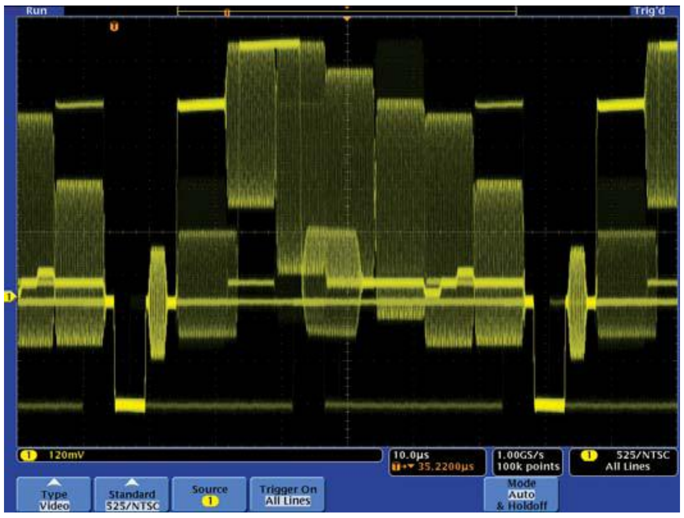
Run
Trig'd
1 120mV
10.0µs
35.2200µs
1.00GS/s
100k points
1 525/NTSC
All Lines
Type
Video
Standard
525/NTSC
Source
1
Trigger On
All Lines
Mode
Auto
& Holdoff

Figure 6. An NTSC composite video signal is an example of a complex wave.

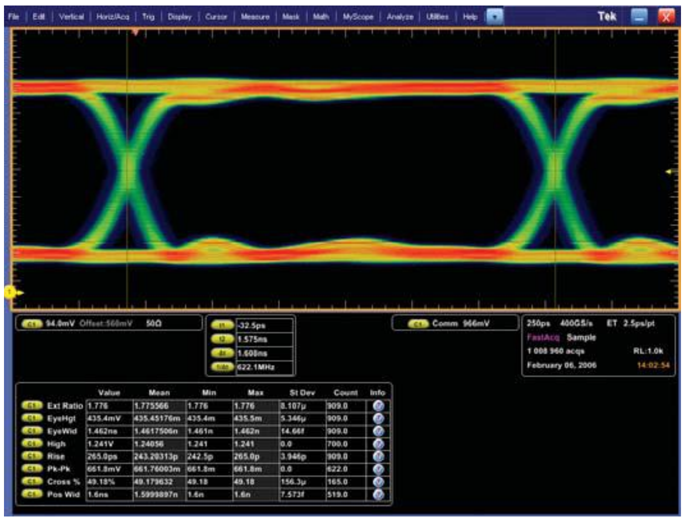
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursor
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
C1 94.0mV Offset:560mV 50Ω
t1 -32.5ps
t2 1.575ns
Δt 1.608ns
1/Δt 622.1MHz
C1 Comm 966mV
250ps 400GS/s ET 2.5ps/pt
FastAcq Sample
1 008 960 acqs RL:1.0k
February 06, 2006 14:02:54

	Value	Mean	Min	Max	St Dev	Count	Info
C1 Ext Ratio	1.776	1.775566	1.776	1.776	8.107µ	909.0	
C1 EyeHgt	435.4mV	435.45176m	435.4m	435.5m	5.346µ	909.0	
C1 EyeWid	1.462ns	1.4617506n	1.461n	1.462n	14.66f	909.0	
C1 High	1.241V	1.24056	1.241	1.241	0.0	700.0	
C1 Rise	265.0ps	243.20313p	242.5p	265.0p	3.946p	909.0	
C1 Pk-Pk	661.8mV	661.76003m	661.8m	661.8m	0.0	622.0	
C1 Cross %	49.18%	49.179632	49.18	49.18	156.3µ	165.0	
C1 Pos Wid	1.6ns	1.5999897n	1.6n	1.6n	7.573f	519.0	

Figure 7. 622 Mb/s serial data eye pattern.

Complex Waves

Some waveforms combine the characteristics of sines, squares, steps, and pulses to produce complex waveshapes. The signal information may be embedded in the form of amplitude, phase, and/or frequency variations. For example, although the signal in Figure 6 is an ordinary composite video signal, it is composed of many cycles of higher-frequency waveforms embedded in a lower-frequency envelope.

In this example, it is usually most important to understand the relative levels and timing relationships of the steps. To view this signal, you need an oscilloscope that captures the low-frequency envelope and blends in the higher-frequency waves in an intensity-graded fashion so that you can see their overall combination as an image that can be visually interpreted. Digital phosphor oscilloscopes are most suited to viewing complex waves, such as video signals, illustrated in Figure 6. Their displays provide the necessary frequency-of-occurrence information, or intensity grading, that is essential to understanding what the waveform is really doing.

Some oscilloscopes allow for displaying certain types of complex waveforms in special ways. For example, telecommunications data may be displayed as an eye pattern or a constellation diagram.

Telecommunications digital data signals can be displayed on an oscilloscope as a special type of waveform referred to as an eye pattern. The name comes from the similarity of the waveform to a series of eyes, as seen in Figure 7. Eye patterns are produced when digital data from a receiver is sampled and applied to the vertical input, while the data rate is used to trigger the horizontal sweep. The eye pattern displays one bit or unit interval of data with all possible edge transitions and states superimposed in one comprehensive view.

A constellation diagram is a representation of a signal modulated by a digital modulation scheme such as quadrature amplitude modulation or phase-shift keying.

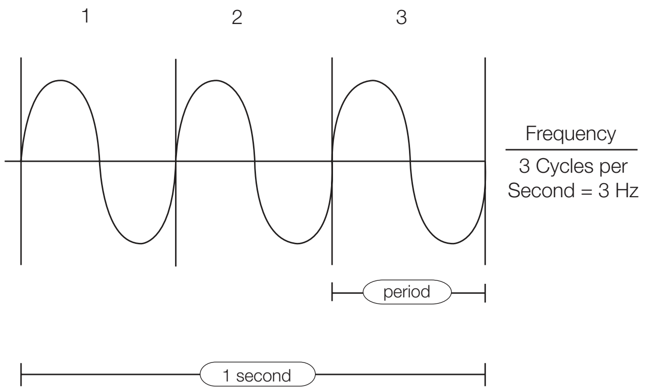

Figure 8. Frequency and period of a sine wave.

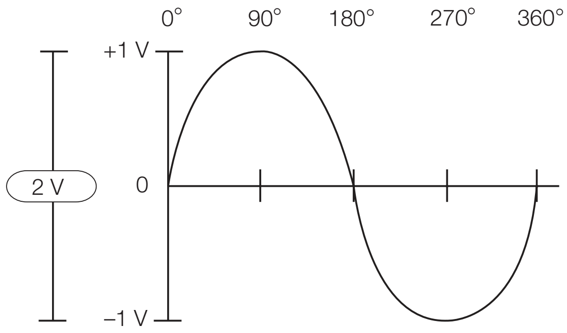

Figure 9. Amplitude and degrees of a sine wave.

Waveform Measurements

Many terms are used to describe the types of measurements that you make with your oscilloscope. This section describes some of the most common measurements and terms.

Frequency and Period

If a signal repeats, it has a frequency. The frequency is measured in Hertz (Hz) and equals the number of times the signal repeats itself in one second, referred to as cycles per second. A repetitive signal also has a period, which is the amount of time it takes the signal to complete one cycle. Period and frequency are reciprocals of each other, so that 1/period equals the frequency and 1/frequency equals the period. For example, the sine wave in Figure 8 has a frequency of 3 Hz and a period of 1/3 second.

Voltage

Voltage is the amount of electric potential, or signal strength, between two points in a circuit. Usually, one of these points is ground, or zero volts, but not always. You may want to measure the voltage from the maximum peak to the minimum peak of a waveform, referred to as the peak-to-peak voltage.

Amplitude

Amplitude refers to the amount of voltage between two points in a circuit. Amplitude commonly refers to the maximum voltage of a signal measured from ground, or zero volts. The waveform shown in Figure 9 has an amplitude of 1 V and a peak-to-peak voltage of 2 V.

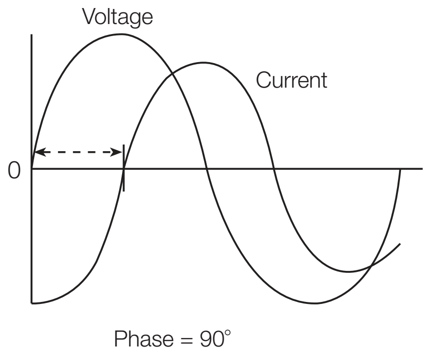

Figure 10. Phase shift.

Phase

Phase is best explained by looking at a sine wave. The voltage level of sine waves is based on circular motion. Given that a circle has 360°, one cycle of a sine wave has 360°, as shown in Figure 9. Using degrees, you can refer to the phase angle of a sine wave when you want to describe how much of the period has elapsed.

Phase shift describes the difference in timing between two otherwise similar signals. The waveform in Figure 10 labeled “current” is said to be 90° out of phase with the waveform labeled “voltage,” since the waves reach similar points in their cycles exactly 1/4 of a cycle apart (360°/4 = 90°). Phase shifts are common in electronics.

Waveform Measurements with Digital Oscilloscopes

Modern digital oscilloscopes have functions that make waveform measurements easier. They have front-panel buttons and/or screen-based menus from which you can select fully automated measurements. These include amplitude, period, rise/fall time, and many more. Many digital instruments also provide mean and RMS calculations, duty cycle, and other math operations. Automated measurements appear as on-screen alphanumeric readouts. Typically these readings are more accurate than is possible to obtain with direct graticule interpretation.

Examples of fully automated waveform measurements:

- Period
- Frequency
- Width +
- Width -
- Rise time
- Fall time
- Amplitude
- Extinction ratio
- Mean optical power
- Duty Cycle +
- Duty Cycle -
- Delay
- Phase
- Burst width
- Peak-to-peak
- Mean
- Cycle mean
- Cycle area
- High
- Low
- Minimum
- Maximum
- Overshoot +
- Overshoot -
- RMS
- Cycle RMS
- Jitter

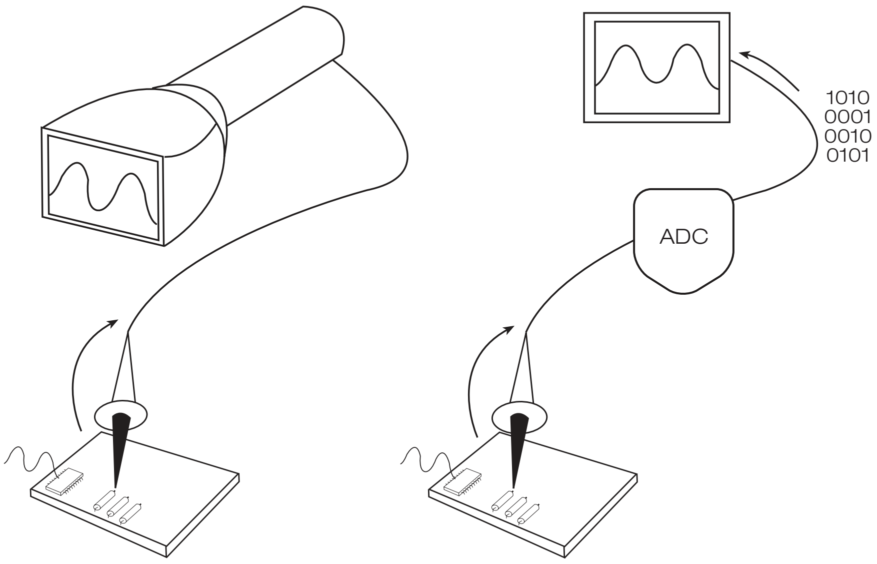

Analog Oscilloscopes
Trace Signals

Digital Oscilloscopes Samples
Signals and Construct Displays

Figure 11. Analog oscilloscopes trace signals, while digital oscilloscopes sample signals and construct displays.

The Types of Oscilloscopes

Electronic equipment can be classified into two categories: analog and digital. Analog equipment works with continuously variable voltages, while digital equipment works with discrete binary numbers that represent voltage samples. A conventional phonograph is an analog device, while a compact disc player is a digital device.

Oscilloscopes can be classified similarly – as analog and digital types. In contrast to an analog oscilloscope, a digital oscilloscope uses an analog-to-digital converter (ADC) to convert the measured voltage into digital information. It acquires the waveform as a series of samples, and stores these samples until it accumulates enough samples to describe a waveform. The digital oscilloscope then re-assembles the waveform for display on the screen, as seen in Figure 11.

Digital oscilloscopes can be classified into digital storage oscilloscopes (DSOs), digital phosphor oscilloscopes (DPOs), mixed signal oscilloscopes (MSOs), and digital sampling oscilloscopes.

The digital approach means that the oscilloscope can display any frequency within its range with stability, brightness, and clarity. For repetitive signals, the bandwidth of the digital oscilloscope is a function of the analog bandwidth of the front-end components of the oscilloscope, commonly referred to as the –3 dB point. For single-shot and transient events, such as pulses and steps, the bandwidth can be limited by the oscilloscope's sample rate. Please refer to the Sample Rate section under Performance Terms and Considerations for a more detailed discussion.

Digital Storage Oscilloscopes

A conventional digital oscilloscope is known as a digital storage oscilloscope (DSO). Its display typically relies on a raster-type screen rather than the luminous phosphor found in an older analog oscilloscope.

Digital storage oscilloscopes (DSOs) allow you to capture and view events that may happen only once – known as transients. Because the waveform information exists in digital form as a series of stored binary values, it can be analyzed, archived, printed, and otherwise processed, within the oscilloscope itself or by an external computer. The waveform need not be continuous; it can be displayed even when the signal disappears. Unlike analog oscilloscopes, digital storage oscilloscopes provide permanent signal storage and extensive waveform processing. However, DSOs typically have no real-time intensity grading; therefore, they cannot express varying levels of intensity in the live signal.

03W-8605-5.qxd 12/10/09 10:24 AM Page 14

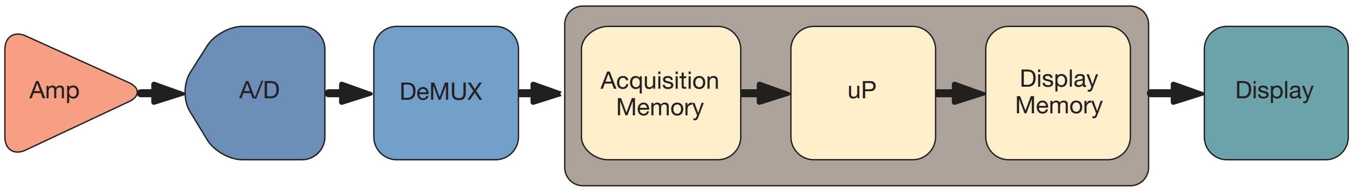

Figure 12. The serial-processing architecture of a digital storage oscilloscope (DSO).

Some of the subsystems that comprise DSOs are similar to those in analog oscilloscopes. However, DSOs contain additional data-processing subsystems that are used to collect and display data for the entire waveform. A DSO employs a serial-processing architecture to capture and display a signal on its screen, as shown in Figure 12. A description of this serial-processing architecture follows.

Serial-processing Architecture

Like an analog oscilloscope, a DSO's first (input) stage is a vertical amplifier. Vertical controls allow you to adjust the amplitude and position range at this stage. Next, the analog-to-digital converter (ADC) in the horizontal system samples the signal at discrete points in time and converts the signal's voltage at these points into digital values called sample points. This process is referred to as digitizing a signal.

The horizontal system's sample clock determines how often the ADC takes a sample. This rate is referred to as the sample rate and is expressed in samples per second (S/s).

The sample points from the ADC are stored in acquisition memory as waveform points. Several sample points may comprise one waveform point. Together, the waveform points comprise one waveform record. The number of waveform points used to create a waveform record is called the record length. The trigger system determines the start and stop points of the record.

The DSO's signal path includes a microprocessor through which the measured signal passes on its way to the display. This microprocessor processes the signal, coordinates display activities, manages the front panel controls, and more. The signal then passes through the display memory and is displayed on the oscilloscope screen.

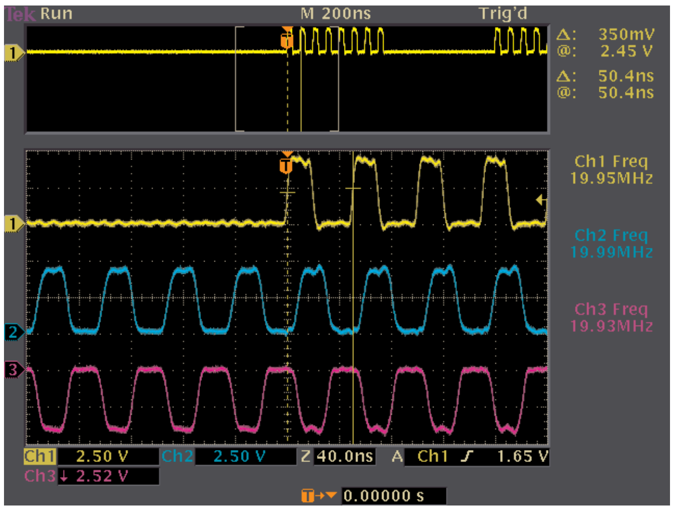

Figure 13. The digital storage oscilloscope delivers high-speed, single-shot acquisition across multiple channels, increasing the likelihood of capturing elusive glitches and transient events.

Depending on the capabilities of your oscilloscope, additional processing of the sample points may take place, which enhances the display. Pre-trigger may also be available, enabling you to see events before the trigger point. Most of today's digital oscilloscopes also provide a selection of automatic parametric measurements, simplifying the measurement process.

As shown in Figure 13, a DSO provides high performance in a single-shot, multi-channel instrument. DSOs are ideal for low-repetition-rate or single-shot, high-speed, multi-channel design applications. In the real world of digital design, an engineer usually examines four or more signals simultaneously, making the DSO a critical companion.

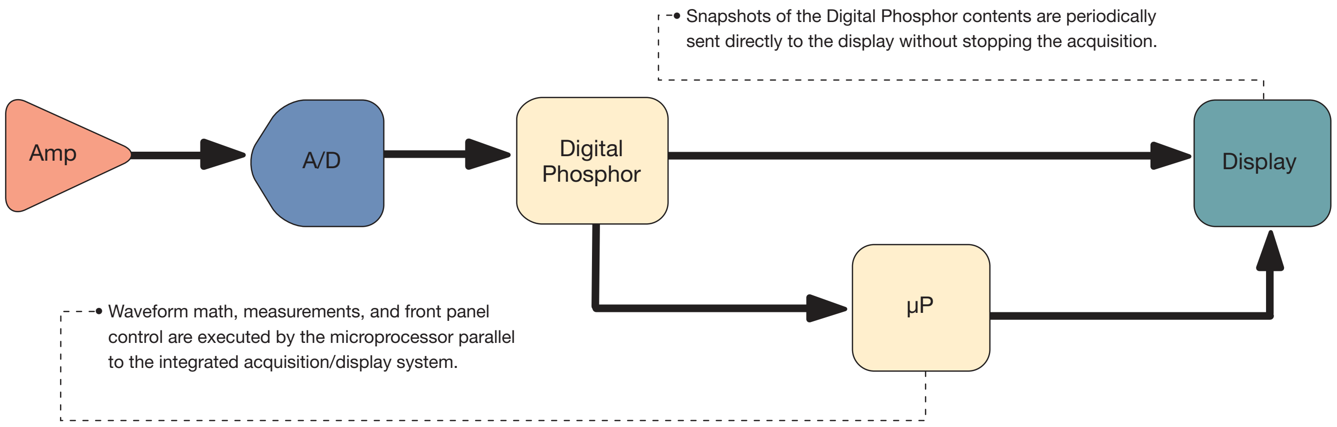
Snapshots of the Digital Phosphor contents are periodically sent directly to the display without stopping the acquisition.
Amp
A/D
Digital Phosphor
Display
μP
Waveform math, measurements, and front panel control are executed by the microprocessor parallel to the integrated acquisition/display system.

Figure 14. The parallel-processing architecture of a digital phosphor oscilloscope (DPO).

Digital Phosphor Oscilloscopes

The digital phosphor oscilloscope (DPO) offers a new approach to oscilloscope architecture. This architecture enables a DPO to deliver unique acquisition and display capabilities to accurately reconstruct a signal.

While a DSO uses a serial-processing architecture to capture, display and analyze signals, a DPO employs a parallel-processing architecture to perform these functions, as shown in Figure 14. The DPO architecture dedicates unique ASIC hardware to acquire waveform images, delivering high waveform capture rates that result in a higher level of signal visualization. This performance increases the probability of witnessing transient events that occur in digital systems, such as runt pulses, glitches and transition errors, and enables additional analysis capability. A description of this parallel-processing architecture follows.

Parallel-processing Architecture

A DPO's first (input) stage is similar to that of an analog oscilloscope – a vertical amplifier – and its second stage is similar to that of a DSO – an ADC. But, the DPO differs significantly from its predecessors following the analog-to-digital conversion.

For any oscilloscope – analog, DSO or DPO – there is always a holdoff time during which the instrument processes the most recently acquired data, resets the system, and waits for the next trigger event. During this time, the oscilloscope is blind to all signal activity. The probability of seeing an infrequent or low-repetition event decreases as the holdoff time increases.

It should be noted that it is impossible to determine the probability of capture by simply looking at the display update rate. If you rely solely on the update rate, it is easy to make the mistake of believing that the oscilloscope is capturing all pertinent information about the waveform when, in fact, it is not.

The digital storage oscilloscope processes captured waveforms serially. The speed of its microprocessor is a bottleneck in this process because it limits the waveform capture rate.

The DPO rasterizes the digitized waveform data into a digital phosphor database. Every 1/30th of a second – about as fast as the human eye can perceive it – a snapshot of the signal image that is stored in the database is pipelined directly to the display system. This direct rasterization of waveform data, and direct copy to display memory from the database, removes the data-processing bottleneck inherent in other architectures. The result is an enhanced "real-time" and lively display update. Signal details, intermittent events, and dynamic characteristics of the signal are captured in real-time. The DPO's microprocessor works in parallel with this integrated acquisition system for display management, measurement automation and instrument control, so that it does not affect the oscilloscope's acquisition speed.

A DPO faithfully emulates the best display attributes of an analog oscilloscope, displaying the signal in three dimensions: time, amplitude and the distribution of amplitude over time, all in real-time.

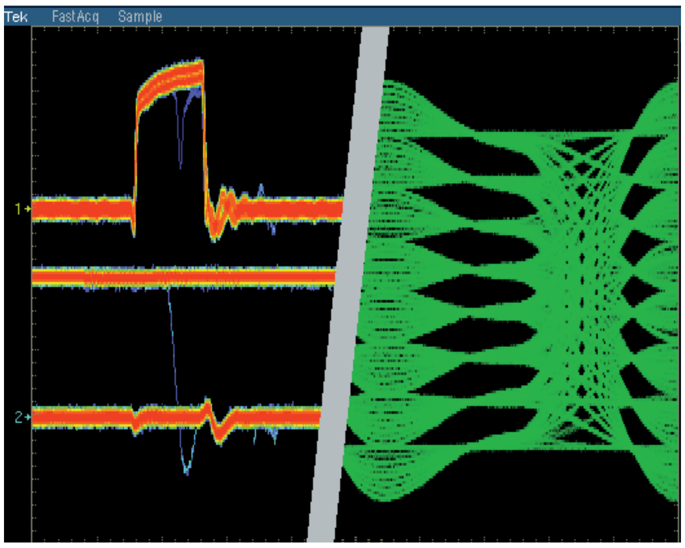

Figure 15. Some DPOs can acquire millions of waveform in just seconds, significantly increasing the probability of capturing intermittent and elusive events and revealing dynamic signal behavior.

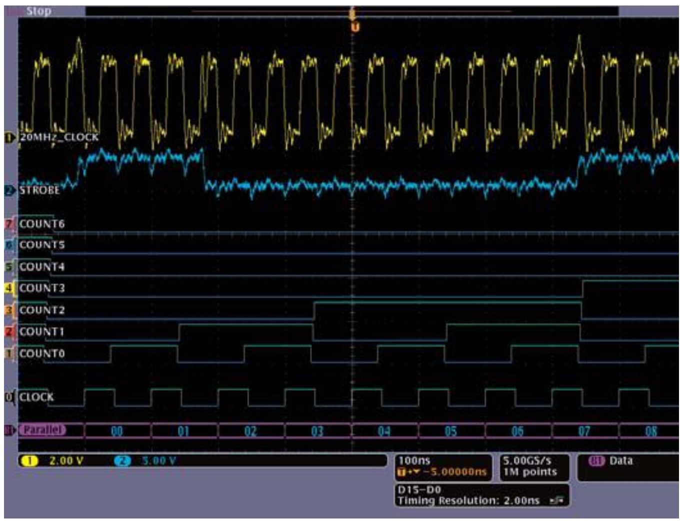

Figure 16. The MSO provides 16 integrated digital channels, enabling the ability to view and analyze time-correlated analog and digital signals.

Unlike an analog oscilloscope's reliance on chemical phosphor, a DPO uses a purely electronic digital phosphor that's actually a continuously updated database. This database has a separate "cell" of information for every single pixel in the oscilloscope's display. Each time a waveform is captured – in other words, every time the oscilloscope triggers – it is mapped into the digital phosphor database's cells. Each cell that represents a screen location and is touched by the waveform is reinforced with intensity information, while other cells are not. Thus, intensity information builds up in cells where the waveform passes most often.

When the digital phosphor database is fed to the oscilloscope's display, the display reveals intensified waveform areas, in proportion to the signal's frequency of occurrence at each point – much like the intensity grading characteristics of an analog oscilloscope. The DPO also allows the display of the varying frequency-of-occurrence information on the display as contrasting colors, unlike an analog oscilloscope. With a DPO, it is easy to see the difference between a waveform that occurs on almost every trigger and one that occurs, say, every 100th trigger.

Digital phosphor oscilloscopes (DPOs) break down the barrier between analog and digital oscilloscope technologies. They are equally suitable for viewing high and low frequencies, repetitive waveforms, transients, and signal variations in real-time. Only a DPO provides the Z (intensity) axis in real-time that is missing from conventional DSOs.

A DPO is ideal for those who need the best general-purpose design and troubleshooting tool for a wide range of applications, as seen in Figure 15. A DPO is exemplary for advanced analysis, communication mask testing, digital debug of intermittent signals, repetitive digital design and timing applications.

Mixed Signal Oscilloscopes

The mixed signal oscilloscope (MSO) combines the performance of a DPO with the basic functionality of a 16-channel logic analyzer, including parallel/serial bus protocol decoding and triggering. The MSO's digital channels view a digital signal as either a logic high or logic low, just like a digital circuit views the signal. This means as long as ringing, overshoot and ground bounce do not cause logic transitions, these analog characteristics are not of concern to the MSO. Just like a logic analyzer, a MSO uses a threshold voltage to determine if the signal is logic high or logic low.

The MSO is the tool of choice for quickly debugging digital circuits using its powerful digital triggering, high resolution acquisition capability, and analysis tools. The root cause of many digital problems is quicker to pinpoint by analyzing both the analog and digital representations of the signal, as shown in Figure 16, making a MSO ideal for verifying and debugging digital circuits.

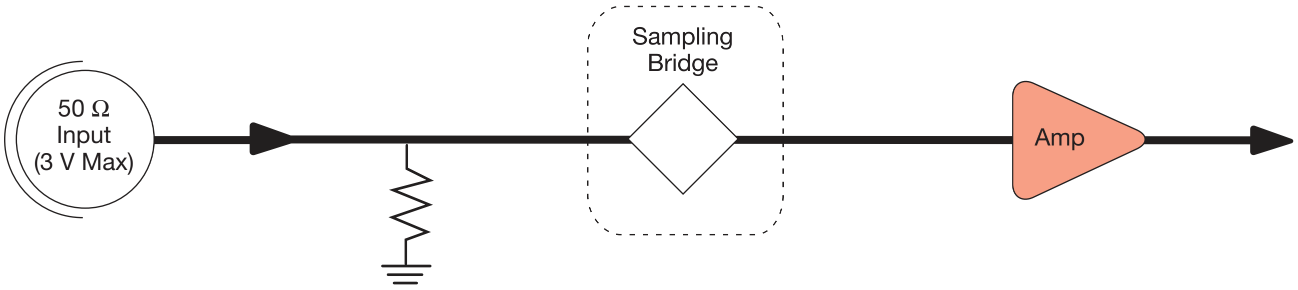
50 Ω
Input
(3 V Max)
Sampling
Bridge
Amp

Figure 17. The architecture of a digital sampling oscilloscope.

Digital Sampling Oscilloscopes

In contrast to the digital storage and digital phosphor oscilloscope architectures, the architecture of the digital sampling oscilloscope reverses the position of the attenuator/amplifier and the sampling bridge, as shown in Figure 17. The input signal is sampled before any attenuation or amplification is performed. A low bandwidth amplifier can then be utilized after the sampling bridge because the signal has already been converted to a lower frequency by the sampling gate, resulting in a much higher bandwidth instrument.

The tradeoff for this high bandwidth, however, is that the sampling oscilloscope's dynamic range is limited. Since there is no attenuator/amplifier in front of the sampling gate, there is no facility to scale the input. The sampling bridge must be able to handle the full dynamic range of the input at all times. Therefore, the dynamic range of most sampling oscilloscopes is limited to about 1 V peak-to-peak. Digital storage and digital phosphor oscilloscopes, on the other hand, can handle 50 to 100 volts.

In addition, protection diodes cannot be placed in front of the sampling bridge as this would limit the bandwidth. This reduces the safe input voltage for a sampling oscilloscope to about 3 V, as compared to 500 V available on other oscilloscopes.

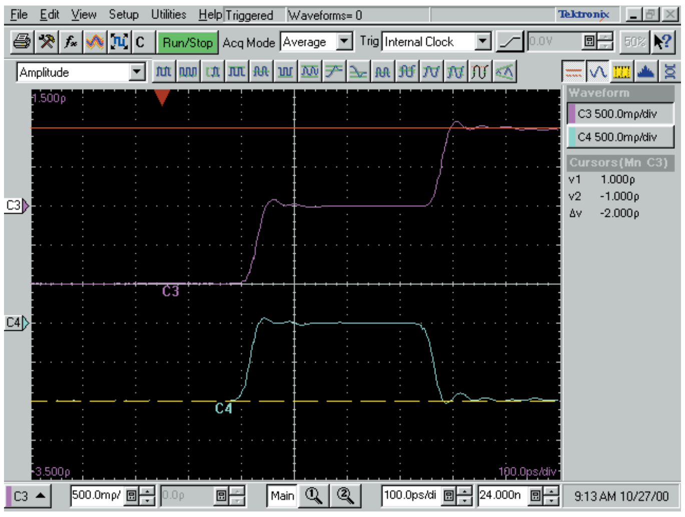
File
Edit
View
Setup
Utilities
Help
Triggered
Waveforms= 0
Tektronix
C
Run/Stop
0.0V
50%
Amplitude
1.500p
C3
C4
-3.500p
100.0ps/div
Waveform
C3 500.0mp/div
C4 500.0mp/div
Cursors(Mn C3)
C3
500.0mp/
0.0p
Main
100.0ps/di
24.000n
9:13 AM 10/27/00

Acq Mode Average
Trig Internal Clock
v1 1.000p
v2 -1.000p
Δv -2.000p

Figure 18. Time domain reflectometry (TDR) display from a digital sampling oscilloscope.

When measuring high-frequency signals, the DSO or DPO may not be able to collect enough samples in one sweep. A digital sampling oscilloscope is an ideal tool for accurately capturing signals whose frequency components are much higher than the oscilloscope's sample rate, as seen in Figure 18. This oscilloscope is capable of measuring signals of up to an order of magnitude faster than any other oscilloscope. It can achieve bandwidth and high-speed timing ten times higher than other oscilloscopes for repetitive signals. Sequential equivalent-time sampling oscilloscopes are available with bandwidths to 80 GHz.

The Systems and Controls of an Oscilloscope

This section briefly describes the basic systems and controls found on analog and digital oscilloscopes. Some controls differ between analog and digital oscilloscopes; your oscilloscope probably has additional controls not discussed here.

A basic oscilloscope consists of four different systems – the vertical system, horizontal system, trigger system, and display system. Understanding each of these systems will enable you to effectively apply the oscilloscope to tackle your specific measurement challenges. Recall that each system contributes to the oscilloscope's ability to accurately reconstruct a signal.

The front panel of an oscilloscope is divided into three main sections labeled vertical, horizontal, and trigger. Your oscilloscope may have other sections, depending on the model and type. See if you can locate these front-panel sections in Figure 19, and on your oscilloscope, as you read through this section.

When using an oscilloscope, you need to adjust three basic settings to accommodate an incoming signal:

- Vertical: The attenuation or amplification of the signal. Use the volts/div control to adjust the amplitude of the signal to the desired measurement range.
- Horizontal: The time base. Use the sec/div control to set the amount of time per division represented horizontally across the screen.
- Trigger: The triggering of the oscilloscope. Use the trigger level to stabilize a repeating signal, or to trigger on a single event.

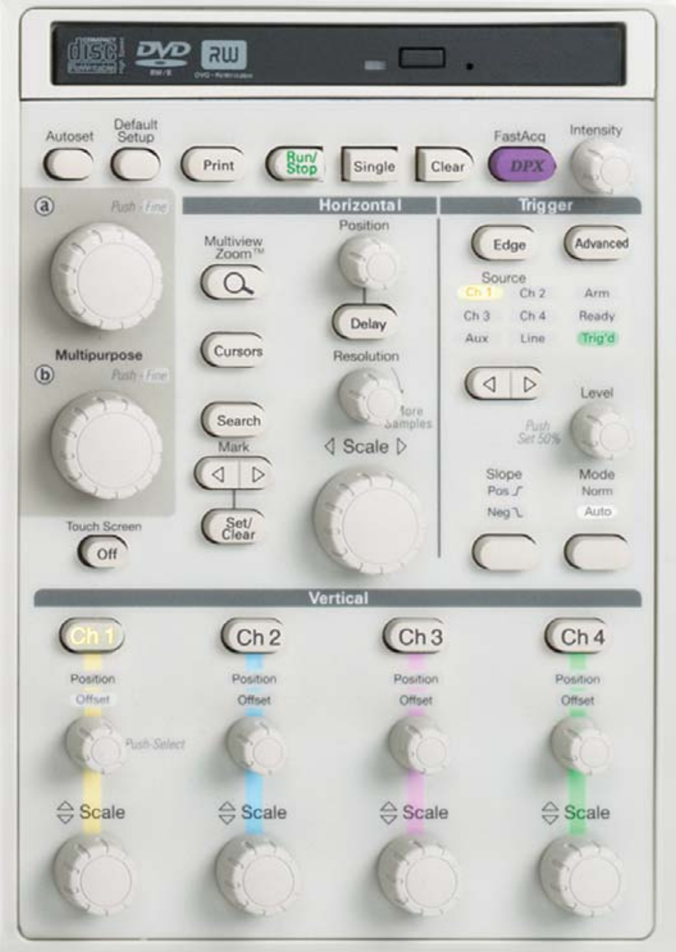
DVD RW
Autoset
Default Setup
Print
Run/Stop
Single
Clear
FastAcq
DPX
Intensity
Horizontal
Trigger
Position
Multiview Zoom
Delay
Resolution
Multipurpose
Push - Fine
Cursors
Search
Mark
Set/Clear
Touch Screen
Off
Scale
Edge
Advanced
Source
Ch 1
Ch 2
Arm
Ch 3
Ch 4
Ready
Aux
Line
Trig'd
Level
Push Set 50%
Slope
Pos
Neg
Mode
Norm
Auto
Vertical
Ch 1
Ch 2
Ch 3
Ch 4
Position
Offset
Push-Select
Scale

Figure 19. Front-panel control section of an oscilloscope.

Common vertical controls include:

- Termination
  - 1M ohm
  - 50 ohm
- Coupling
  - DC
  - AC
  - GND
- Bandwidth
  - Limit
  - Enhancement
- Position
- Offset
- Invert – On/Off
- Scale
  - Fixed steps
  - Variable

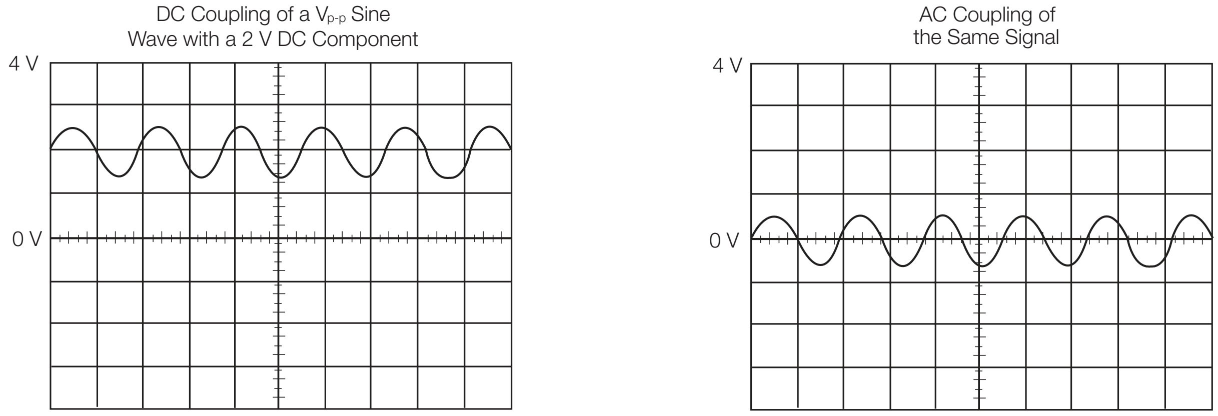
DC Coupling of a V_{p-p} Sine Wave with a 2 V DC Component

AC Coupling of the Same Signal

Figure 20. AC and DC input coupling.

Vertical System and Controls

Vertical controls can be used to position and scale the waveform vertically, set the input coupling, and adjust other signal conditioning.

Position and Volts per Division

The vertical position control allows you to move the waveform up and down exactly where you want it on the screen.

The volts-per-division setting (usually written as volts/div) is a scaling factor that varies the size of the waveform on the screen. If the volts/div setting is 5 volts, then each of the eight vertical divisions represents 5 volts and the entire screen can display 40 volts from bottom to top, assuming a graticule with eight major divisions. If the setting is 0.5 volts/div, the screen can display 4 volts from bottom to top, and so on. The maximum voltage you can display on the screen is the volts/div setting multiplied by the number of vertical divisions. Note that the probe you use, 1X or 10X, also influences the scale factor. You must divide the volts/div scale by the attenuation factor of the probe if the oscilloscope does not do it for you.

Often the volts/div scale has either a variable gain or a fine gain control for scaling a displayed signal to a certain number of divisions. Use this control to assist in taking rise time measurements.

Input Coupling

Coupling refers to the method used to connect an electrical signal from one circuit to another. In this case, the input coupling is the connection from your test circuit to the oscilloscope. The coupling can be set to DC, AC, or ground. DC coupling shows all of an input signal. AC coupling blocks the DC component of a signal so that you see the waveform centered around zero volts. Figure 20 illustrates this difference. The AC coupling setting is useful when the entire signal (alternating current + direct current) is too large for the volts/div setting.

The ground setting disconnects the input signal from the vertical system, which lets you see where zero volts is located on the screen. With grounded input coupling and auto trigger mode, you see a horizontal line on the screen that represents zero volts. Switching from DC to ground and back again is a handy way of measuring signal voltage levels with respect to ground.

Bandwidth Limit

Most oscilloscopes have a circuit that limits the bandwidth of the oscilloscope. By limiting the bandwidth, you reduce the noise that sometimes appears on the displayed waveform, resulting in a cleaner signal display. Note that, while eliminating noise, the bandwidth limit can also reduce or eliminate high-frequency signal content.

Bandwidth Enhancement

Some oscilloscopes may provide a DSP arbitrary equalization filter which can be used to improve the oscilloscope channel response. This filter extends the bandwidth, flattens the oscilloscope channel frequency response, improves phase linearity, and provides a better match between channels. It also decreases rise time and improves the time domain step response.

Horizontal System and Controls

An oscilloscope's horizontal system is most closely associated with its acquisition of an input signal – sample rate and record length are among the considerations here. Horizontal controls are used to position and scale the waveform horizontally.

Common horizontal controls include:

- Time Base
- XY
- Scale
- Trace Separation
- Record Length
- Resolution
- Sample Rate
- Trigger Position
- Zoom/Pan
- Search

Acquisition Controls

Digital oscilloscopes have settings that let you control how the acquisition system processes a signal. Look over the acquisition options on your digital oscilloscope while you read this description. Figure 21 shows you an example of an acquisition menu.

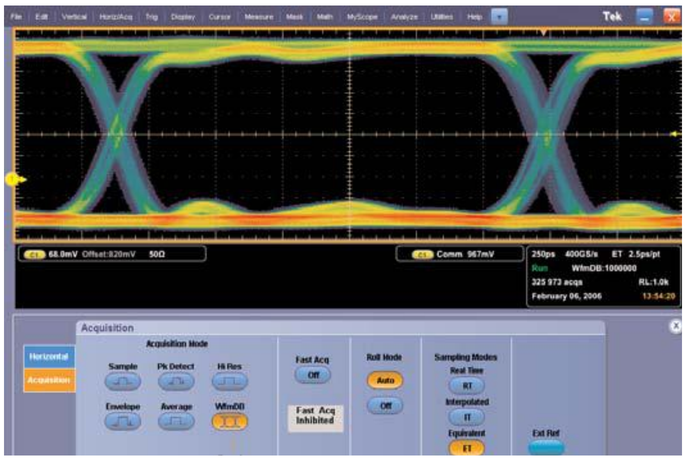
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursor
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
C1 68.0mV Offset:820mV 50Ω
C1 Comm 967mV
250ps 400GS/s ET 2.5ps/pt
Run WfmDB:1000000
325 973 acqs RL:1.0k
February 06, 2006 13:54:20
Acquisition
Horizontal
Acquisition
Acquisition Mode
Sample
Pk Detect
Hi Res
Envelope
Average
WfmDB
Fast Acq
Off
Fast Acq Inhibited
Roll Mode
Auto
Off
Sampling Modes
Real Time
RT
Interpolated
IT
Equivalent
ET
Ext Ref

Figure 21. Example of an acquisition menu.

Acquisition Modes

Acquisition modes control how waveform points are produced from sample points. Sample points are the digital values derived directly from the analog-to-digital converter (ADC). The sample interval refers to the time between these sample points. Waveform points are the digital values that are stored in memory and displayed to construct the waveform. The time value difference between waveform points is referred to as the waveform interval.

The sample interval and the waveform interval may, or may not, be the same. This fact leads to the existence of several different acquisition modes in which one waveform point is comprised of several sequentially acquired sample points.

Additionally, waveform points can be created from a composite of sample points taken from multiple acquisitions, which provides another set of acquisition modes. A description of the most commonly used acquisition modes follows.

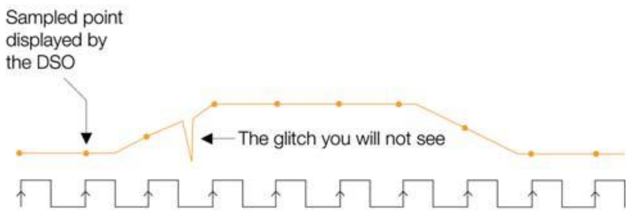

Figure 22. Sample rate varies with time base settings - the slower the time based setting, the slower the sample rate. Some digital oscilloscopes provide peak detect mode to capture fast transients at slow sweep speeds.

Types of Acquisition Modes

- Sample Mode: This is the simplest acquisition mode. The oscilloscope creates a waveform point by saving one sample point during each waveform interval.
- Peak Detect Mode: The oscilloscope saves the minimum and maximum value sample points taken during two waveform intervals and uses these samples as the two corresponding waveform points. Digital oscilloscopes with peak detect mode run the ADC at a fast sample rate, even at very slow time base settings (slow time base settings translate into long waveform intervals) and are able to capture fast signal changes that would occur between the waveform points if in sample mode, as shown in Figure 22. Peak detect mode is particularly useful for seeing narrow pulses spaced far apart in time, as demonstrated in Figure 23.
- Hi-Res Mode: Like peak detect, hi res mode is a way of getting more information in cases when the ADC can sample faster than the time base setting requires. In this case, multiple samples taken within one waveform interval are averaged together to produce one waveform point. The result is a decrease in noise and an improvement in resolution for low-speed signals. The advantage of Hi Res Mode over Average is that Hi-Res Mode can be used even on a single shot event.
- Envelope Mode: Envelope mode is similar to peak detect mode. However, in envelope mode, the minimum and maximum waveform points from multiple acquisitions are combined to form a waveform that shows min/max accumulation over time. Peak detect mode is usually used to acquire the records that are combined to form the envelope waveform.

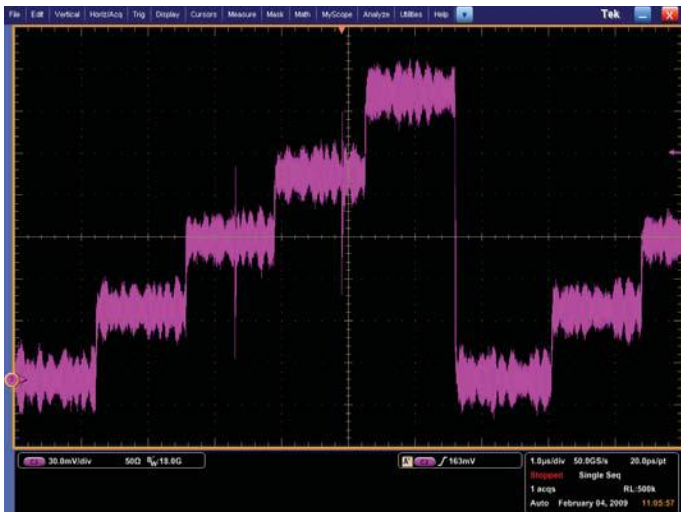

30.0mV/div 50Ω B_W:18.0G
163mV
1.0µs/div 50.0GS/s 20.0ps/pt
Stopped Single Seq
1 acqs RL:500k
Auto February 04, 2009 11:05:57

Figure 23. Peak detect mode enables the oscilloscope to capture extremely brief transient anomalies.

- Average Mode: In average mode, the oscilloscope saves one sample point during each waveform interval as in sample mode. However, waveform points from consecutive acquisitions are then averaged together to produce the final displayed waveform. Average mode reduces noise without loss of bandwidth, but requires a repeating signal.
- Waveform Database Mode: In waveform database mode, the oscillocope accumulates a waveform database that provides a three-dimensional array of amplitude, time, and counts.

Starting and Stopping the Acquisition System

One of the greatest advantages of digital oscilloscopes is their ability to store waveforms for later viewing. To this end, there are usually one or more buttons on the front panel that allow you to start and stop the acquisition system so you can analyze waveforms at your leisure. Additionally, you may want the oscilloscope to automatically stop acquiring after one acquisition is complete or after one set of records has been turned into an envelope or average waveform. This feature is commonly called single sweep or single sequence and its controls are usually found either with the other acquisition controls or with the trigger controls.

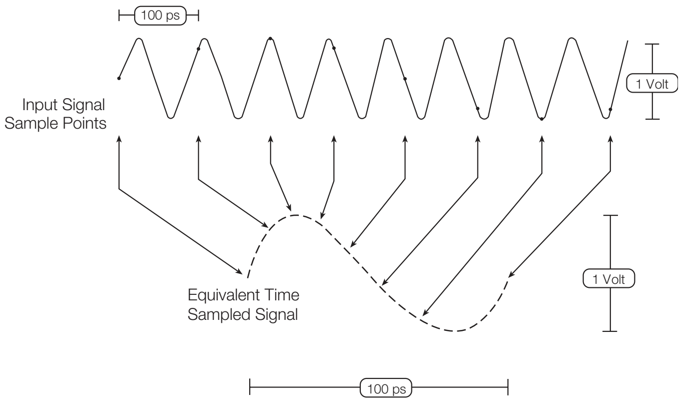

Figure 24. Basic sampling, showing sample points are connected by interpolation to produce a continuous waveform.

Sampling

Sampling is the process of converting a portion of an input signal into a number of discrete electrical values for the purpose of storage, processing and/or display. The magnitude of each sampled point is equal to the amplitude of the input signal at the instant in time in which the signal is sampled.

Sampling is like taking snapshots. Each snapshot corresponds to a specific point in time on the waveform. These snapshots can then be arranged in the appropriate order in time so as to reconstruct the input signal.

In a digital oscilloscope, an array of sampled points is reconstructed on a display with the measured amplitude on the vertical axis and time on the horizontal axis, as illustrated in Figure 24.

The input waveform in Figure 24 appears as a series of dots on the screen. If the dots are widely spaced and difficult to interpret as a waveform, the dots can be connected using a process called interpolation. Interpolation connects the dots with lines, or vectors. A number of interpolation methods are available that can be used to produce an accurate representation of a continuous input signal.

Sampling Controls

Some digital oscilloscopes provide you with a choice in sampling method – either real-time sampling or equivalent-time sampling. The acquisition controls available with these oscilloscopes will allow you to select a sample method to acquire signals. Note that this choice makes no difference for slow time base settings and only has an effect when the ADC cannot sample fast enough to fill the record with waveform points in one pass. Each sampling method has distinct advantages, depending on the kind of measurements being made.

Controls are typically available on modern oscilloscopes to give you the choice of three horizontal time base modes of operations. If you are simply doing signal exploration and want to interact with a lively signal, you will use the Automatic or interactive default mode that provides you with the liveliest display update rate. If you want a precise measurement and the highest real-time sample rate that will give you the most measurement accuracy, then the Constant Sample Rate mode is for you. It will maintain the highest sample rate and provide the best real-time resolution. The last mode is called the Manual mode because it ensures direct and independent control of the sample rate and record length.

Real-time Sampling Method

Real-time sampling is ideal for signals whose frequency range is less than half the oscilloscope's maximum sample rate. Here, the oscilloscope can acquire more than enough points in one "sweep" of the waveform to construct an accurate picture, as shown in Figure 25. Real-time sampling is the only way to capture fast, single-shot, transient signals with a digital oscilloscope.

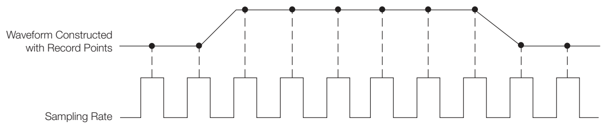
Waveform Constructed with Record Points
Sampling Rate

Figure 25. Real-time sampling method.

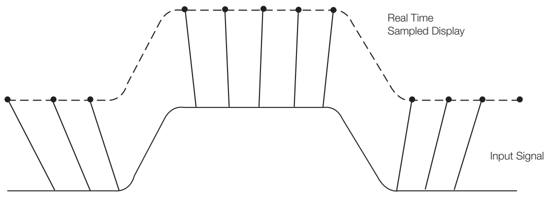
Real Time Sampled Display
Input Signal

Figure 26. In order to capture this 10 ns pulse in real-time, the sample rate must be high enough to accurately define the edges.

Real-time sampling presents the greatest challenge for digital oscilloscopes because of the sample rate needed to accurately digitize high-frequency transient events, as shown in Figure 26. These events occur only once, and must be sampled in the same time frame that they occur.

If the sample rate isn't fast enough, high-frequency components can "fold down" into a lower frequency, causing aliasing in the display, as demonstrated in Figure 27. In addition, real-time sampling is further complicated by the high-speed memory required to store the waveform once it is digitized. Please refer to the Sample Rate and Record Length sections under Performance Terms and Considerations for additional detail regarding the sample rate and record length needed to accurately characterize high-frequency components.

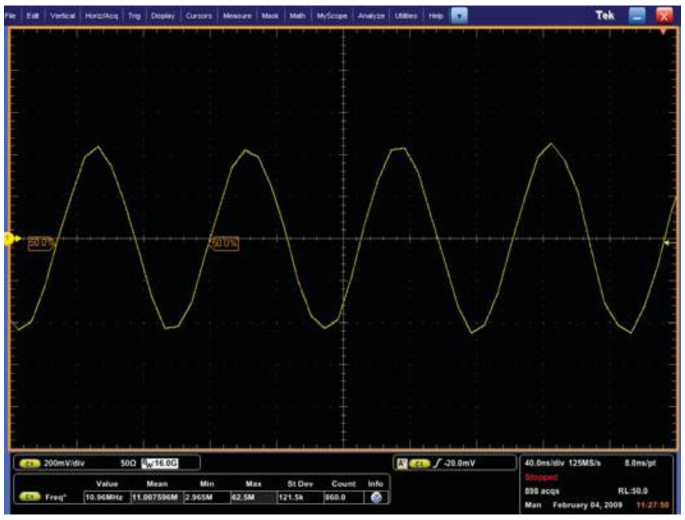
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
200mV/div
50Ω
16.0G
-20.0mV
40.0ns/div 125MS/s 8.0ns/pt
Stopped
898 acqs
RL:50.0
Mon February 04, 2009 11:27:50

	Value	Mean	Min	Max	St Dev	Count	Info
Freq*	10.96MHz	11.007596M	2.965M	62.5M	121.5k	860.0	

Figure 27. Undersampling of a 100 MHz sine wave introduces aliasing effects.

Real-time sampling with interpolation, dgital oscilloscopes take discrete samples of the signal that can be displayed. However, it can be difficult to visualize the signal represented as dots, especially because there can be only a few dots representing high-frequency portions of the signal. To aid in the visualization of signals, digital oscilloscopes typically have interpolation display modes.

In simple terms, interpolation "connects the dots" so that a signal that is sampled only a few times in each cycle can be accurately displayed. Using real-time sampling with interpolation, the oscilloscope collects a few sample points of the signal in a single pass in real-time mode and uses interpolation to fill in the gaps. Interpolation is a processing technique used to estimate what the waveform looks like based on a few points.

03W-8605-5.qxd 12/10/09 10:24 AM Page 24

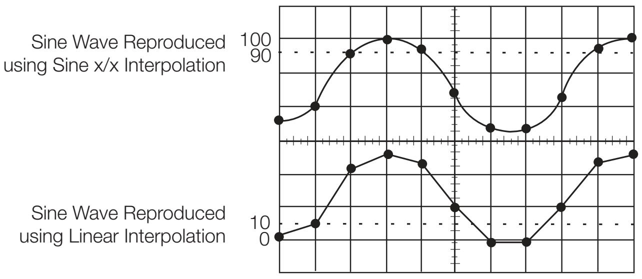

Figure 28. Linear and sin x/x interpolation.

Linear interpolation connects sample points with straight lines. This approach is limited to reconstructing straight-edged signals, as illustrated in Figure 28, which better lends itself to square waves.

The more versatile sin x/x interpolation connects sample points with curves, as shown in Figure 28. Sin x/x interpolation is a mathematical process in which points are calculated to fill in the time between the real samples. This form of interpolation lends itself to curved and irregular signal shapes, which are far more common in the real world than pure square waves and pulses. Consequently, sin x/x interpolation is the preferred method for applications where the sample rate is 3 to 5 times the system bandwidth.

Equivalent-time Sampling Method

When measuring high-frequency signals, the oscilloscope may not be able to collect enough samples in one sweep. Equivalent-time sampling can be used to accurately acquire signals whose frequency exceeds half the oscilloscope's sample rate, as illustrated in Figure 29. Equivalent-time digitizers (samplers) take advantage of the fact that most naturally occurring and man-made events are repetitive. Equivalent-time sampling constructs a picture of a repetitive signal by capturing a little bit of information from each repetition. The waveform slowly builds up like a string of lights, illuminating one-by-one. This allows the oscilloscope to accurately capture signals whose frequency components are much higher than the oscilloscope's sample rate.

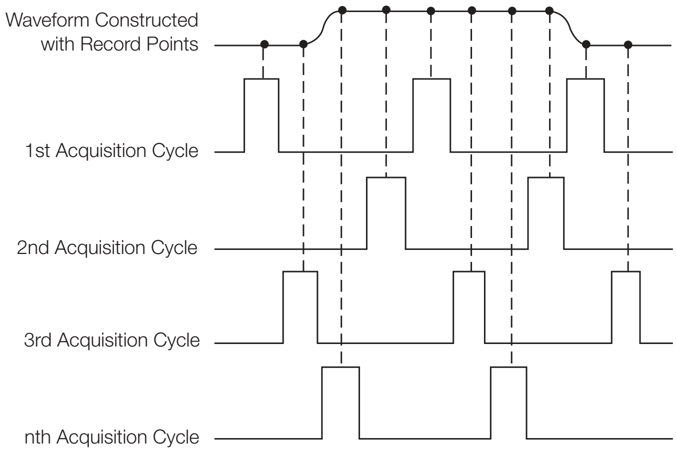

Figure 29. Some oscilloscopes use equivalent-time sampling to capture and display very fast, repetitive signals.

There are two types of equivalent-time sampling methods: random and sequential. Each has its advantages. Random equivalent-time sampling allows display of the input signal prior to the trigger point, without the use of a delay line. Sequential equivalent-time sampling provides much greater time resolution and accuracy. Both require that the input signal be repetitive.

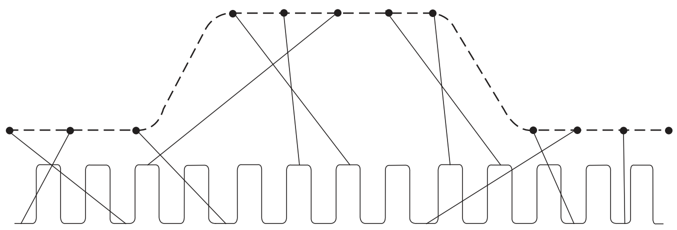

Figure 30. In random equivalent-time sampling, the sampling clock runs asynchronously with the input signal and the trigger.

Random Equivalent-time Sampling

Random equivalent-time digitizers (samplers) utilize an internal clock that runs asynchronously with respect to the input signal and the signal trigger, as illustrated in Figure 30. Samples are taken continuously, independent of the trigger position, and are displayed based on the time difference between the sample and the trigger. Although samples are taken sequentially in time, they are random with respect to the trigger – hence the name "random" equivalent-time sampling. Sample points appear randomly along the waveform when displayed on the oscilloscope screen.

The ability to acquire and display samples prior to the trigger point is the key advantage of this sampling technique, eliminating the need for external pretrigger signals or delay lines. Depending on the sample rate and the time window of the display, random sampling may also allow more than one sample to be acquired per triggered event. However, at faster sweep speeds, the acquisition window narrows until the digitizer cannot sample on every trigger. It is at these faster sweep speeds that very precise timing measurements are often made, and where the extraordinary time resolution of the sequential equivalent-time sampler is most beneficial. The bandwidth limit for random equivalent-time sampling is less than for sequential-time sampling.

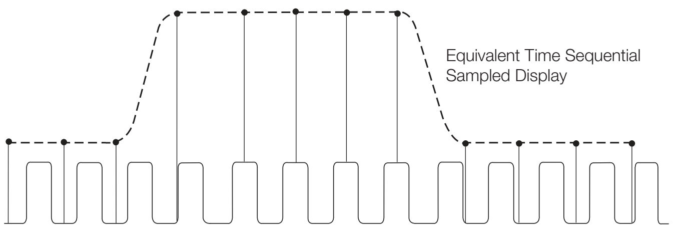

Figure 31. In sequential equivalent-time sampling, the single sample is taken for each recognized trigger after a time delay which is incremented after each cycle.

Sequential Equivalent-time Sampling

The sequential equivalent-time sampler acquires one sample per trigger, independent of the time/div setting, or sweep speed, as illustrated in Figure 31. When a trigger is detected, a sample is taken after a very short, but well-defined, delay. When the next trigger occurs, a small time increment – delta t – is added to this delay and the digitizer takes another sample. This process is repeated many times, with "delta t" added to each previous acquisition, until the time window is filled. Sample points appear from left to right in sequence along the waveform when displayed on the oscilloscope screen.

Technologically speaking, it is easier to generate a very short, very precise "delta t" than it is to accurately measure the vertical and horizontal positions of a sample relative to the trigger point, as required by random samplers. This precisely measured delay is what gives sequential samplers their unmatched time resolution. Since, with sequential sampling, the sample is taken after the trigger level is detected, the trigger point cannot be displayed without an analog delay line, which may, in turn, reduce the bandwidth of the instrument. If an external pretrigger can be supplied, bandwidth will not be affected.

Position and Seconds per Division

The horizontal position control moves the waveform left and right to exactly where you want it on the screen.

The seconds-per-division setting (usually written as sec/div) lets you select the rate at which the waveform is drawn across the screen (also known as the time base setting or sweep speed). This setting is a scale factor. If the setting is 1 ms, each horizontal division represents 1 ms and the total screen width represents 10 ms, or ten divisions. Changing the sec/div setting enables you to look at longer and shorter time intervals of the input signal.

As with the vertical volts/div scale, the horizontal sec/div scale may have variable timing, allowing you to set the horizontal time scale between the discrete settings.

Time Base Selections

Your oscilloscope has a time base, which is usually referred to as the main time base. Many oscilloscopes also have what is called a delayed time base – a time base with a sweep that can start (or be triggered to start) relative to a pre-determined time on the main time base sweep. Using a delayed time base sweep allows you to see events more clearly and to see events that are not visible solely with the main time base sweep.

The delayed time base requires the setting of a time delay and the possible use of delayed trigger modes and other settings not described in this primer. Refer to the manual supplied with your oscilloscope for information on how to use these features.

Zoom/Pan

Your oscilloscope may have special horizontal magnification settings that let you display a magnified section of the waveform on-screen. Some oscilloscopes add pan functions to the zoom capability. Knobs are used to adjust zoom factor or scale and the pan of the zoom box across the waveform.

Search

Some oscilloscopes offer search and mark capabilities, enabling you to quickly navigate through long acquisitions looking for user-defined events.

XY Mode

Most oscilloscopes have an XY mode that lets you display an input signal, rather than the time base, on the horizontal axis. This mode of operation opens up a whole new area of phase shift measurement techniques, explained in the Measurement Techniques section of this primer.

Z Axis

A digital phosphor oscilloscope (DPO) has a high display sample density and an innate ability to capture intensity information. With its intensity axis (Z axis), the DPO is able to provide a three-dimensional, real-time display similar to that of an analog oscilloscope. As you look at the waveform trace on a DPO, you can see brightened areas – the areas where a signal occurs most often. This display makes it easy to distinguish the basic signal shape from a transient that occurs only once in a while – the basic signal would appear much brighter. One application of the Z axis is to feed special timed signals into the separate Z input to create highlighted "marker" dots at known intervals in the waveform.

XYZ Mode with DPO and XYZ Record Display

Some DPOs can use the Z input to create an XY display with intensity grading. In this case, the DPO samples the instantaneous data value at the Z input and uses that value to qualify a specific part of the waveform. Once you have qualified samples, these samples can accumulate, resulting in an intensity-graded XYZ display. XYZ mode is especially useful for displaying the polar patterns commonly used in testing wireless communication devices – a constellation diagram, for example. Another method of displaying XYZ data is XYZ record display. In this mode the data from the acquisition memory is used rather than the DPO database.

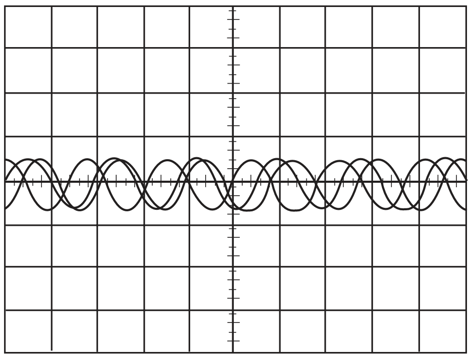

Figure 32. Untriggered display.

Trigger System and Controls

An oscilloscope's trigger function synchronizes the horizontal sweep at the correct point of the signal, essential for clear signal characterization. Trigger controls allow you to stabilize repetitive waveforms and capture single-shot waveforms.

The trigger makes repetitive waveforms appear static on the oscilloscope display by repeatedly displaying the same portion of the input signal. Imagine the jumble on the screen that would result if each sweep started at a different place on the signal, as illustrated in Figure 32.

Edge triggering, available in analog and digital oscilloscopes, is the basic and most common type. In addition to threshold triggering offered by both analog and digital oscilloscopes, many digital oscilloscopes offer numerous specialized trigger settings not offered by analog instruments. These triggers respond to specific conditions in the incoming signal, making it easy to detect, for example, a pulse that is narrower than it should be. Such a condition would be impossible to detect with a voltage threshold trigger alone.

Advanced trigger controls enable you to isolate specific events of interest to optimize the oscilloscope's sample rate and record length. Advanced triggering capabilities in some oscilloscopes give you highly selective control. You can trigger on pulses defined by amplitude (such as runt pulses), qualified by time (pulse width, glitch, slew rate, setup-and-hold, and time-out), and delineated by logic state or pattern (logic triggering).

Other advanced trigger functions include:

- **Pattern Lock Triggering:** Pattern lock triggering adds a new dimension to NRZ serial pattern triggering by enabling the oscilloscope to take synchronized acquisitions of a long serial test pattern with outstanding time base accuracy. Pattern lock triggering can be used to remove random jitter from long serial data patterns. Effects of specific bit transitions can be investigated, and averaging can be used with mask testing.
- **Serial Pattern Triggering:** Serial pattern triggering can be used to debug serial architectures. It provides a trigger on the serial pattern of an NRZ serial data stream with built-in clock recovery and correlates events across the physical and link layer. The instrument can recover the clock signal, identify transitions, and allow you to set the desired encoded words for the serial pattern trigger to capture.
- **A & B Triggering:** Some trigger systems offer multiple trigger types only on a single event (A event), with delayed trigger (B event) selection limited to edge type triggering and often do not provide a way to reset the trigger sequence if the B event doesn't occur. Modern oscilloscopes can provide the full suite of advanced trigger types on both A and B triggers, logic qualification to control when to look for these events, and reset triggering to begin the trigger sequence again after a specified time, state, or transition so that even events in the most complex signals can be captured.
- **Search & Mark Triggering:** Hardware triggers watch for one event type at a time, but Search can scan for multiple event types simultaneously. For example, scan for setup or hold time violations on multiple channels. Individual marks can be placed by Search indicating events that meet search criteria.
- **Trigger Correction:** Since the trigger and data acquisition systems share different paths there is some inherent time delay between the trigger position and the data acquired. This results in skew and trigger jitter. With a trigger correction system the instrument adjusts the trigger position and compensates for the difference of delay there is between the trigger path and the data acquisition path. This will eliminate virtually any trigger jitter at the trigger point. In this mode, the trigger point can be used as a measurement reference.

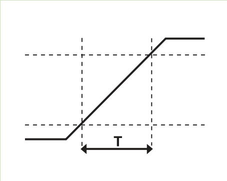

Slew Rate Triggering. High frequency signals with slew rates faster than expected or needed can radiate troublesome energy. Slew rate triggering surpasses conventional edge triggering by adding the element of time and allowing you to selectively trigger on fast or slow edges.

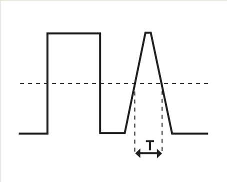

Glitch Triggering. Glitch triggering allows you to trigger on digital pulses when they are shorter or longer than a user-defined time limit. This trigger control enables you to examine the causes of even rare glitches and their effects on other signals

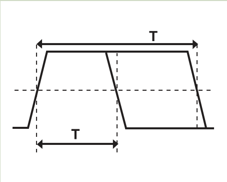

Pulse Width Triggering. Using pulse width triggering, you can monitor a signal indefinitely and trigger on the first occurrence of a pulse whose duration (pulse width) is outside the allowable limits.

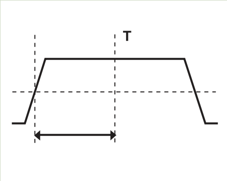

Time-out Triggering. Time-out triggering lets you trigger on an event without waiting for the trigger pulse to end, by triggering based on a specified time lapse.

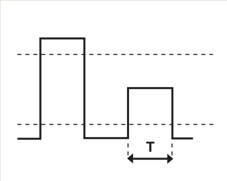

Runt Pulse Triggering. Runt triggering allows you to capture and examine pulses that cross one logic threshold, but not both.

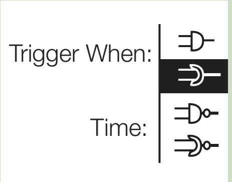

Logic Triggering. Logic triggering allows you to trigger on any logical combination of available input channels – especially useful in verifying the operation of digital logic.

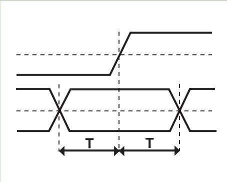

Setup-and-Hold Triggering. Only setup-and-hold triggering lets you deterministically trap a single violation of setup-and-hold time that would almost certainly be missed by using other trigger modes. This trigger mode makes it easy to capture specific signal quality and timing details when a synchronous data signal fails to meet setup-and-hold specifications.

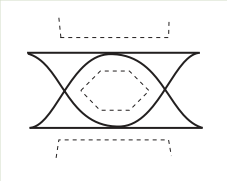

Communication Triggering. Optionally available on certain oscilloscope models, these trigger modes address the need to acquire a wide variety of Alternate-Mark Inversion (AMI), Code-Mark Inversion (CMI), and Non-Return to Zero (NRZ) communication signals.

Figure 33. Common trigger types.

- **Serial Triggering on Specific Standard Signals I²C, CAN, LIN, etc.)** - Some oscilloscopes provide the ability to trigger on specific signal types for standard serial data signals such as CAN, LIN, I²C, SPI, and others. The decode of these signal types is also available on many oscilloscopes today.
- **Parallel Bus Triggering** - Multiple parallel buses can be defined and displayed at one time to easily view decoded parallel bus data over time. By specifying which channels are the clock and data lines, you can create a parallel bus display on some oscilloscopes that automatically decodes bus content. Countless hours can be saved by using parallel bus triggers to simplify capture and analysis.

Optional trigger controls in some oscilloscopes are designed specifically to examine communications signals as well. Figure 33 highlights a few of these common trigger types in more detail. The intuitive user interface available in some oscilloscopes also allows rapid setup of trigger parameters with wide flexibility in the test setup to maximize your productivity.

Trigger Position

Horizontal trigger position control is only available on digital oscilloscopes. The trigger position control may be located in the horizontal control section of your oscilloscope. It actually represents the horizontal position of the trigger in the waveform record.

Varying the horizontal trigger position allows you to capture what a signal did before a trigger event, known as pre-trigger viewing. Thus, it determines the length of viewable signal both preceding and following a trigger point.

Digital oscilloscopes can provide pre-trigger viewing because they constantly process the input signal, whether or not a trigger has been received. A steady stream of data flows through the oscilloscope; the trigger merely tells the oscilloscope to save the present data in memory.
In contrast, analog oscilloscopes only display the signal – that is, write it on the CRT – after receiving the trigger. Thus, pre-trigger viewing is not available in analog oscilloscopes, with the exception of a small amount of pre-trigger provided by a delay line in the vertical system.

Pre-trigger viewing is a valuable troubleshooting aid. If a problem occurs intermittently, you can trigger on the problem, record the events that led up to it and, possibly, find the cause.

Trigger Level and Slope

The trigger level and slope controls provide the basic trigger point definition and determine how a waveform is displayed, as illustrated in Figure 34.

The trigger circuit acts as a comparator. You select the slope and voltage level on one input of the comparator. When the trigger signal on the other comparator input matches your settings, the oscilloscope generates a trigger.

The slope control determines whether the trigger point is on the rising or the falling edge of a signal. A rising edge is a positive slope and a falling edge is a negative slope. The level control determines where on the edge the trigger point occurs.

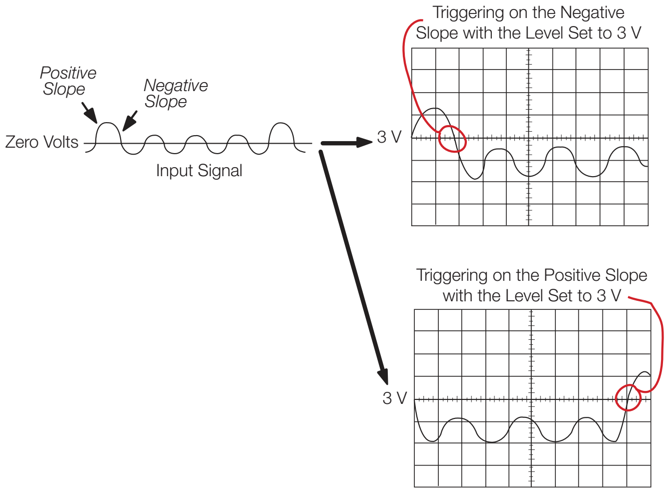
Triggering on the Negative Slope with the Level Set to 3 V
Positive Slope
Negative Slope
Zero Volts
Input Signal
3 V
Triggering on the Positive Slope with the Level Set to 3 V
3 V

Figure 34. Positive and negative slope triggering.

Trigger Sources

The oscilloscope does not necessarily need to trigger on the signal being displayed. Several sources can trigger the sweep:

- Any input channel
- An external source other than the signal applied to an input channel
- The power source signal
- A signal internally defined by the oscilloscope, from one or more input channels

Most of the time, you can leave the oscilloscope set to trigger on the channel displayed. Some oscilloscopes provide a trigger output that delivers the trigger signal to another instrument.

The oscilloscope can use an alternate trigger source, whether or not it is displayed, so you should be careful not to unwittingly trigger on channel 1 while displaying channel 2, for example.

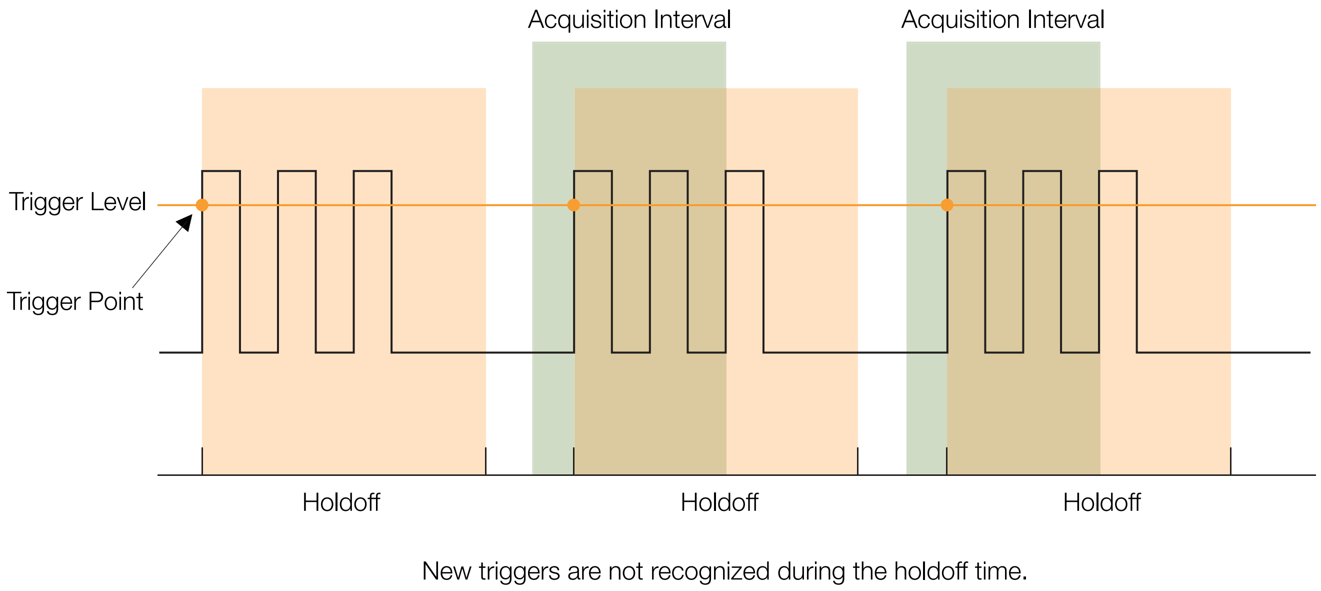

Figure 35. Trigger holdoff.

Trigger Modes

The trigger mode determines whether or not the oscilloscope draws a waveform based on a signal condition. Common trigger modes include normal and auto.

In normal mode the oscilloscope only sweeps if the input signal reaches the set trigger point; otherwise (on an analog oscilloscope) the screen is blank or (on a digital oscilloscope) frozen on the last acquired waveform. Normal mode can be disorienting since you may not see the signal at first if the level control is not adjusted correctly.

Auto mode causes the oscilloscope to sweep, even without a trigger. If no signal is present, a timer in the oscilloscope triggers the sweep. This ensures that the display will not disappear if the signal does not cause a trigger.

In practice, you will probably use both modes: normal mode because it lets you see just the signal of interest, even when triggers occur at a slow rate, and auto mode because it requires less adjustment.

Many oscilloscopes also include special modes for single sweeps, triggering on video signals, or automatically setting the trigger level.

Trigger Coupling

Just as you can select either AC or DC coupling for the vertical system, you can choose the kind of coupling for the trigger signal.

Besides AC and DC coupling, your oscilloscope may also have high frequency rejection, low frequency rejection, and noise rejection trigger coupling. These special settings are useful for eliminating noise from the trigger signal to prevent false triggering.

Trigger Holdoff

Sometimes getting an oscilloscope to trigger on the correct part of a signal requires great skill. Many oscilloscopes have special features to make this task easier.

Trigger holdoff is an adjustable period of time after a valid trigger during which the oscilloscope cannot trigger. This feature is useful when you are triggering on complex waveform shapes, so that the oscilloscope only triggers on an eligible trigger point. Figure 35 shows how using trigger holdoff helps create a usable display.

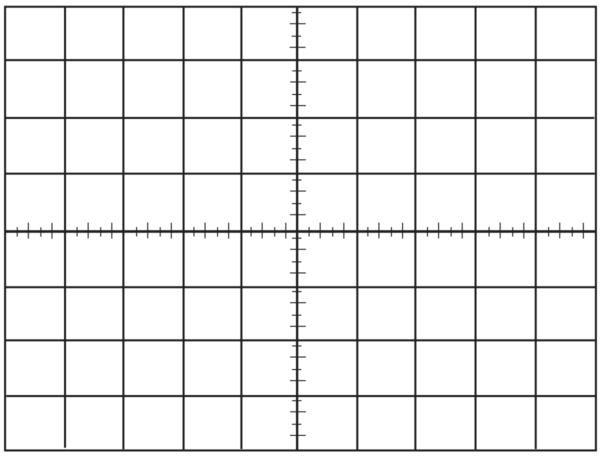

Figure 36. An oscilloscope graticule.

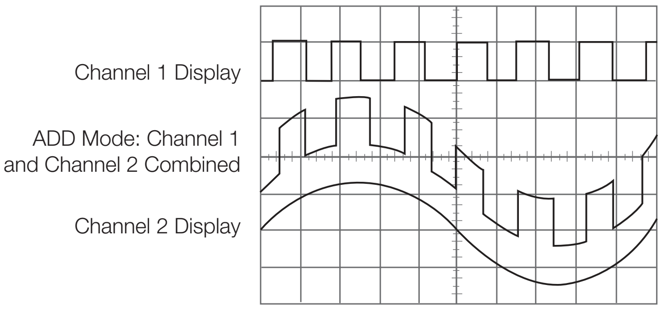

Figure 37. Adding channels.

Display System and Controls

An oscilloscope's front panel includes a display screen and the knobs, buttons, switches, and indicators used to control signal acquisition and display. As mentioned at the front of this section, front-panel controls are usually divided into vertical, horizontal and trigger sections. The front panel also includes input connectors.

Take a look at the oscilloscope display. Notice the grid markings on the screen – these markings create the graticule. Each vertical and horizontal line constitutes a major division. The graticule is usually laid out in an 8-by-10 or 10-by-10 division pattern. Labeling on the oscilloscope controls (such as volts/div and sec/div) always refers to major divisions. The tick marks on the center horizontal and vertical graticule lines, as shown in Figure 36, are called minor divisions. Many oscilloscopes display on the screen how many volts each vertical division represents and how many seconds each horizontal division represents.

Other Oscilloscope Controls

Math and Measurement Operations

Your oscilloscope may also have operations that allow you to add waveforms together, creating a new waveform display. Analog oscilloscopes combine the signals while digital oscilloscopes create new waveforms mathematically. Subtracting waveforms is another math operation. Subtraction with analog oscilloscopes is possible by using the channel invert function on one signal and then using the add operation. Digital oscilloscopes typically have a subtraction operation available. Figure 37 illustrates a third waveform created by combining two different signals.

Using the power of their internal processors, digital oscilloscopes offer many advanced math operations: multiplication, division, integration, Fast Fourier Transform, and more. This advanced signal processing capability can also perform functions such as the insertion of a filter block which can be used de-embed the characteristics of the fixture on the device under test or implement a filter block with desired frequency response such as a low pass filter. The processing block is flexible – not dedicated; it can perform as an arbitrary filter instead, for example for simulation of pre-emphasis/de-emphasis schemes.

Digital Timing and State Acquisitions

Digital channels provided by a mixed signal oscilloscope enable acquisition capabilities similar to those found on logic analyzers. There are two major digital acquisition techniques. The first technique is timing acquisition in which the MSO samples the digital signal at uniformly spaced times determined by the MSO's sample rate. At each sample point, the MSO stores the signal's logic state and creates a timing diagram of the signal. The second digital acquisition technique is state acquisition. State acquisition defines special times that the digital signal's logic state is valid and stable. This is common in synchronous and clocked digital circuits. A clock signal defines the time when the signal state is valid. For example, the input signal stable time is around the rising clock edge for a D-Flip-Flop with rising edge clocking. The output signal stable time is around the falling clock edge for a D-Flip-Flop with rising edge clocking. Since the clock period of a synchronous circuit may not be fixed, the time between state acquisitions may not be uniform as it is in a timing acquisition.

A mixed signal oscilloscope's digital channels acquire signals similar to how a logic analyzer acquires signals in timing acquisition mode. The MSO then decodes the timing acquisition into a clocked bus display, and event table which is similar to the logic analyzer's state acquisition display, providing you with important information during debug.

We have described the basic oscilloscope controls that a beginner needs to know about. Your oscilloscope may have other controls for various functions. Some of these may include:

- Automatic parametric measurements
- Measurement cursors
- Keypads for mathematical operations or data entry
- Printing capabilities
- Interfaces for connecting your oscilloscope to a computer or directly to the Internet

Look over the other options available to you and read your oscilloscope’s manual to find out more about these other controls.

The Complete Measurement System

Probes

Even the most advanced instrument can only be as precise as the data that goes into it. A probe functions in conjunction with an oscilloscope as part of the measurement system. Precision measurements start at the probe tip. The right probes matched to the oscilloscope and the device-under-test (DUT) not only allow the signal to be brought to the oscilloscope cleanly, they also amplify and preserve the signal for the greatest signal integrity and measurement accuracy.

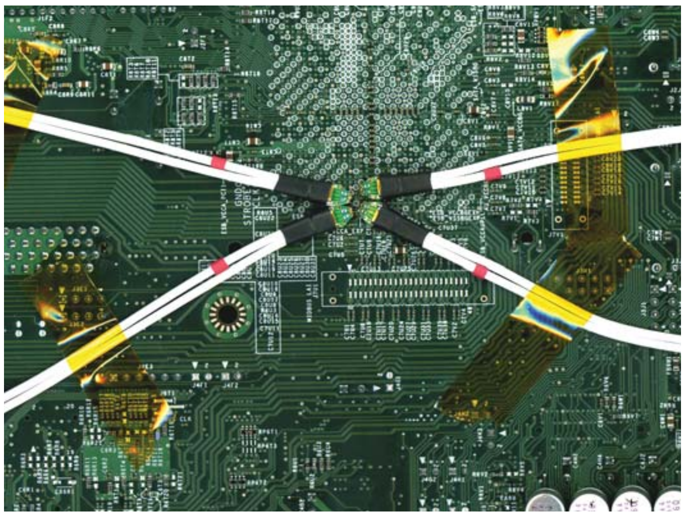

Figure 38. Dense devices and systems require small form factor probes.

To ensure accurate reconstruction of your signal, try to choose a probe that, when paired with your oscilloscope, exceeds the signal bandwidth by 5 times.

Probes actually become part of the circuit, introducing resistive, capacitive and inductive loading that inevitably alters the measurement. For the most accurate results, the goal is to select a probe with minimal loading. An ideal pairing of the probe with the oscilloscope will minimize this loading, and enable you to access all of the power, features and capabilities of your oscilloscope.

Another connectivity consideration is the probe’s form factor. Small form factor probes provide easier access to today’s densely packed circuitry, as shown in Figure 38.

A brief description of the types of probes follows. Please refer to the Tektronix ABCs of Probes primer for more information about this essential component of the overall measurement system.

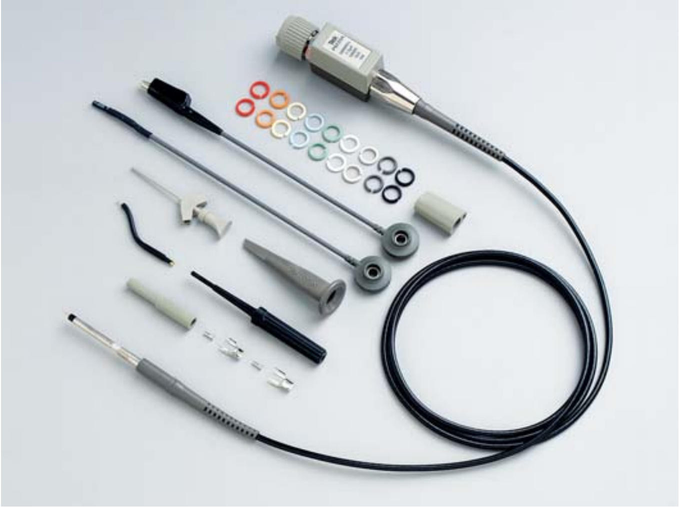

Figure 39. A typical passive probe with accessories.

Passive Probes

For measuring typical signal and voltage levels, passive probes provide ease-of-use and a wide range of measurement capabilities at an affordable price. The pairing of a passive voltage probe with a current probe will provide you with an ideal solution for measuring power.

Most passive probes have some attenuation factor, such as 10X, 100X, and so on. By convention, attenuation factors, such as for the 10X attenuator probe, have the X after the factor. In contrast, magnification factors like X10 have the X first.

The 10X (read as "ten times") attenuator probe reduces circuit loading in comparison to a 1X probe and is an excellent general-purpose passive probe. Circuit loading becomes more pronounced for higher frequency and/or higher impedance signal sources, so be sure to analyze these signal/probe loading interactions before selecting a probe. The 10X attenuator probe improves the accuracy of your measurements, but also reduces the signal's amplitude at the oscilloscope input by a factor of 10.

Because it attenuates the signal, the 10X attenuator probe makes it difficult to look at signals less than 10 millivolts peak-to-peak. The 1X probe is similar to the 10X attenuator probe but lacks the attenuation circuitry. Without this circuitry, more interference is introduced to the circuit being tested.

Use the 10X attenuator probe as your general-purpose probe, but keep the 1X probe accessible to measure slow-speed, low-amplitude signals. Some probes have a convenient feature for switching between 1X and 10X attenuation at the probe tip. If your probe has this feature, make sure you are using the correct setting before taking measurements.

Many oscilloscopes can automatically detect whether you are using a 1X or 10X probe and adjust their screen readouts accordingly. However with some oscilloscopes, you must set the type of probe you are using or read from the proper 1X or 10X marking on the volts/div control.

The 10X attenuator probe works by balancing the probe's electrical properties against the oscilloscope's electrical properties. Before using a 10X attenuator probe you need to adjust this balance for your particular oscilloscope. This adjustment is known as compensating the probe and is described in more detail in the Operating the Oscilloscope section of this primer.

Passive probes, such as the one shown in Figure 39, provide excellent general-purpose probing solutions. However, general-purpose passive probes cannot accurately measure signals with extremely fast rise times, and may excessively load sensitive circuits. The steady increase in signal clock rates and edge speeds demands higher speed probes with less loading effects. High-speed active and differential probes provide ideal solutions when measuring high-speed and/or differential signals.

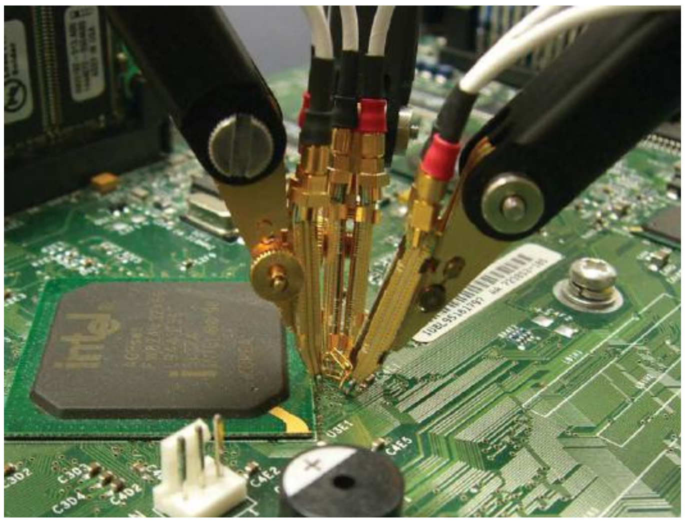

Figure 40. High-performance probes are critical when measuring the fast clocks and edges found in today's computer buses and data transmission lines.

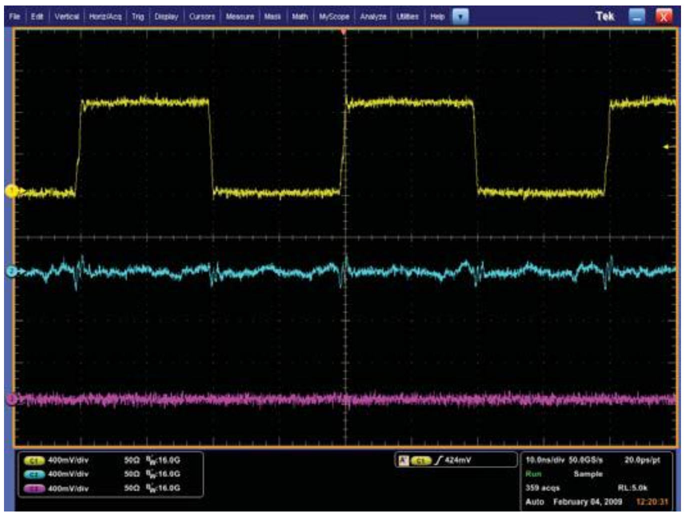

File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
C1 400mV/div 50Ω B_W:16.0G
C2 400mV/div 50Ω B_W:16.0G
C3 400mV/div 50Ω B_W:16.0G
A C1 ∫ 424mV
10.0ns/div 50.0GS/s 20.0ps/pt
Run Sample
359 acqs RL:5.0k
Auto February 04, 2009 12:20:31

Figure 41 Differential probes can separate common-mode noise from the signal content of interest in today's fast, low-voltage applications - especially important as digital signals continue to fall below typical noise thresholds found in integrated circuits.

Active and Differential Probes

Increasing signal speeds and lower-voltage logic families make accurate measurement results difficult to achieve. Signal fidelity and device loading are critical issues. A complete measurement solution at these high speeds includes high-speed, high-fidelity probing solutions to match the performance of the oscilloscope, as shown in Figure 40.

Active and differential probes use specially developed integrated circuits to preserve the signal during access and transmission to the oscilloscope, ensuring signal integrity. For measuring signals with fast rise times, a high-speed active or differential probe will provide more accurate results, as highlighted in Figure 41.

Newer probe types provide the advantage of being able to use one setup, and get three types of measurements without adjusting probe tip connections. These probes can make differential, single-ended and common mode measurements from the same probe setup.

Logic Probes

The logic probe shown in Figure 42 offers two eight-channel pods. Each channel ends with a probe tip featuring a recessed ground for simplified connection to the device-under-test. The coax on the first channel of each pod is colored blue making it easy to identify. The common ground uses an automotive style connector making it easy to create custom grounds for connecting to the device-under-test. When connecting to square pins, you can use an adapter that attaches to the probe head extending the probe ground flush with the probe tip so you can attach to a header. These probes offer outstanding electrical characteristics with minimal capacitive loading.

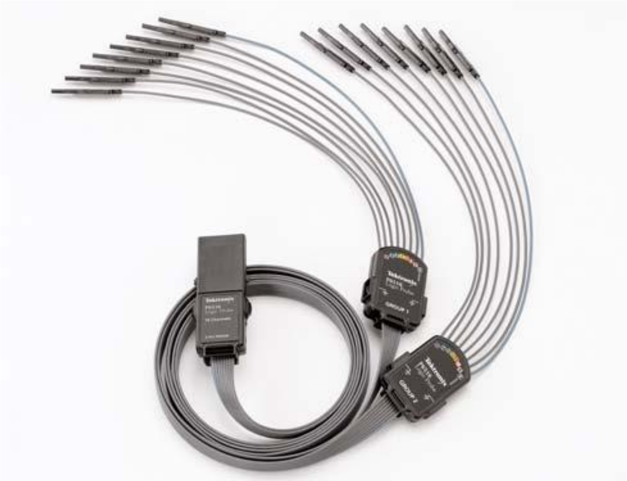

Figure 42. Logic probes for a mixed signal oscilloscope (MSO) simplify digital connectivity to your device.

Specialty Probes

In addition to the previously mentioned probe types, there are also a variety of other specialty probes and probing systems. These include current, high-voltage, and optical probes, just to name a few.

Probe Accessories

Many modern oscilloscopes provide special automated features built into the input and mating probe connectors. In the case of intelligent probe interfaces, the act of connecting the probe to the instrument notifies the oscilloscope about the probe's attenuation factor, which in turn scales the display so that the probe's attenuation is figured into the readout on the screen. Some probe interfaces also recognize the type of probe – that is, passive, active or current. The interface may act as a DC power source for probes. Active probes have their own amplifier and buffer circuitry that requires DC power.

Ground lead and probe tip accessories are also available to improve signal integrity when measuring high-speed signals. Ground lead adapters provide spacing flexibility between probe tip and ground lead connections to the DUT, while maintaining very short lead lengths from probe tip to DUT.

Please refer to the Tektronix ABCs of Probes primer for more information about probes and probe accessories.

Performance Terms and Considerations

As previously mentioned, an oscilloscope is analogous to a camera that captures signal images that we can observe and interpret. Shutter speed, lighting conditions, aperture and the ASA rating of the film all affect the camera's ability to capture an image clearly and accurately.

Like the basic systems of an oscilloscope, the performance considerations of an oscilloscope significantly affect its ability to achieve the required signal integrity.

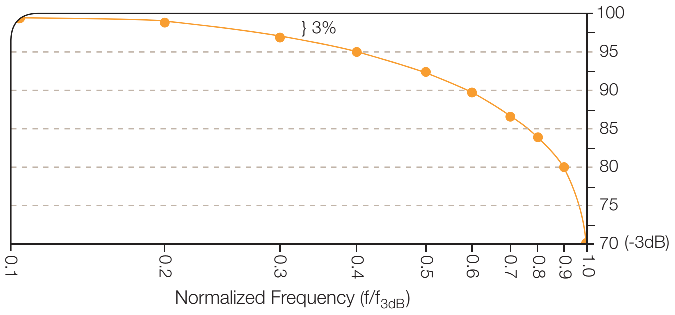

Figure 43. Oscilloscope bandwidth is the frequency at which a sinusoidal input signal is attenuated to 70.7% of the signal's true amplitude, known as the -3 dB point.

Learning a new skill often involves learning a new vocabulary. This idea holds true for learning how to use an oscilloscope. This section describes some useful measurement and oscilloscope performance terms. These terms are used to describe the criteria essential to choosing the right oscilloscope for your application. Understanding these terms will help you to evaluate and compare your oscilloscope with other models.

Bandwidth

Bandwidth determines an oscilloscope's fundamental ability to measure a signal. As signal frequency increases, the capability of the oscilloscope to accurately display the signal decreases. This specification indicates the frequency range that the oscilloscope can accurately measure.

Oscilloscope bandwidth is specified as the frequency at which a sinusoidal input signal is attenuated to 70.7% of the signal's true amplitude, known as the –3 dB point, a term based on a logarithmic scale, as seen in Figure 43.

Without adequate bandwidth, your oscilloscope will not be able to resolve high-frequency changes. Amplitude will be distorted. Edges will vanish. Details will be lost. Without adequate bandwidth, all the features, bells and whistles in your oscilloscope will mean nothing.

To determine the oscilloscope bandwidth needed to accurately characterize signal amplitude in your specific application, apply the "5 Times Rule."

$$\text{Oscilloscope Bandwidth} \geq \begin{matrix}\text{Highest Frequency}\\ \text{Component of Signal}\end{matrix} \; x \;\; 5$$

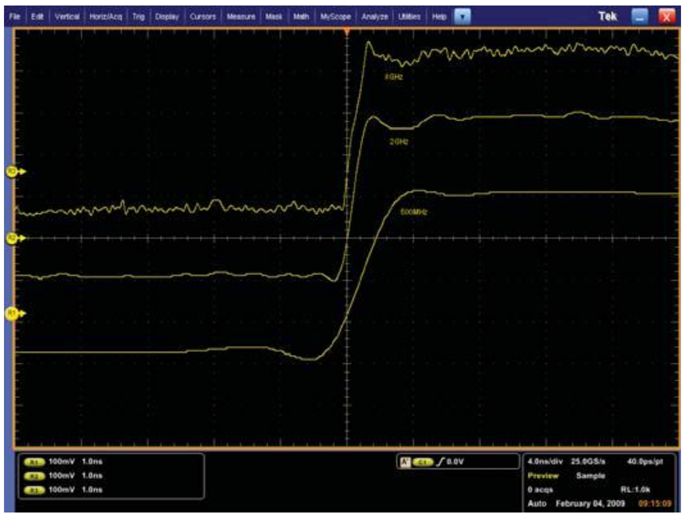

100mV 1.0ns
100mV 1.0ns
100mV 1.0ns

0.0V

4.0ns/div 25.0GS/s 40.0ps/pt
Preview Sample
0 acqs RL:1.0k
Auto February 04, 2009 09:15:09

Figure 44. The higher the bandwidth, the more accurate the reproduction of your signal, as illustrated with a signal captured at 250 MHz, 1 GHz and 4 GHz bandwidth levels.

An oscilloscope selected using the 5 Times Rule will give you less than ±2% error in your measurements – typically sufficient for today's applications. However, as signal speeds increase, it may not be possible to achieve this rule of thumb. Always keep in mind that higher bandwidth will likely provide more accurate reproduction of your signal, as demonstrated in Figure 44.

Some oscilloscopes provide a method of enhancing the bandwidth through digital signal processing. A DSP arbitrary equalization filter can be used to improve the oscilloscope channel response. This filter extends the bandwidth, flattens the oscilloscope's channel frequency response, improves phase linearity, and provides a better match between channels. It also decreases rise time and improves the time-domain step response.

Rise Time

In the digital world, rise time measurements are critical. Rise time may be a more appropriate performance consideration when you expect to measure digital signals, such as pulses and steps. As shown in Figure 45, your oscilloscope must have sufficient rise time to accurately capture the details of rapid transitions.

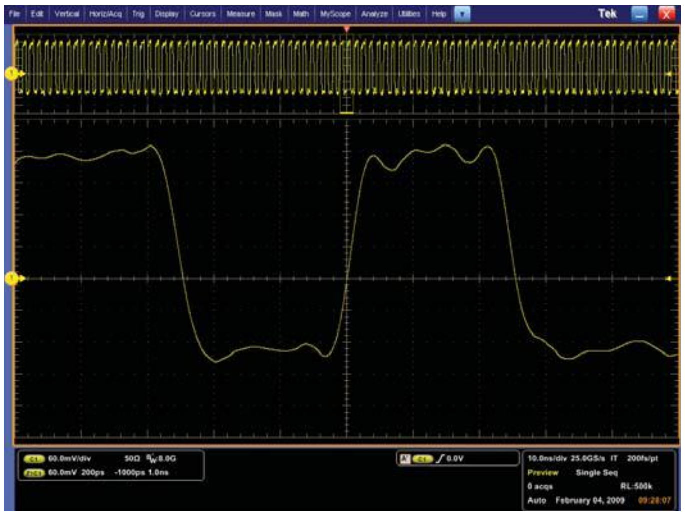

60.0mV/div 50Ω BW:8.0G
60.0mV 200ps -1000ps 1.0ns

0.0V

10.0ns/div 25.0GS/s IT 200fs/pt
Preview Single Seq
0 acqs RL:500k
Auto February 04, 2009 09:28:07

Figure 45. Rise time characterization of a high-speed digital signal.

Rise time describes the useful frequency range of an oscilloscope. To calculate the oscilloscope rise time required for your signal type, use the following equation:

$$\text{Oscilloscope Rise Time} \leq \text{Fastest Rise Time of Signal} \; x \; \frac{1}{5}$$

Note that this basis for oscilloscope rise time selection is similar to that for bandwidth. As in the case of bandwidth, achieving this rule of thumb may not always be possible given the extreme speeds of today's signals. Always remember that an oscilloscope with faster rise time will more accurately capture the critical details of fast transitions.

In some applications, you may know only the rise time of a signal. A constant allows you to relate the bandwidth and rise time of the oscilloscope, using the equation:

$$\text{Bandwidth} = \frac{k}{\text{Rise Time}}$$

where k is a value between 0.35 and 0.45, depending on the shape of the oscilloscope's frequency response curve and pulse rise time response. Oscilloscopes with a bandwidth of <1 GHz typically have a 0.35 value, while oscilloscopes with a bandwidth of > 1 GHz usually have a value between 0.40 and 0.45.

Logic Family	Typical Signal Rise Time	Calculated Signal Bandwidth
TTL	2 ns	175 MHz
CMOS	1.5 ns	230 MHz
GTL	1 ns	350 MHz
LVDS	400 ps	875 MHz
ECL	100 ps	3.5 GHz
GaAs	40 ps	8.75 GHz

Figure 46. Some logic families produce inherently faster rise times than others.

Some logic families produce inherently faster rise times than others, as illustrated in Figure 46.

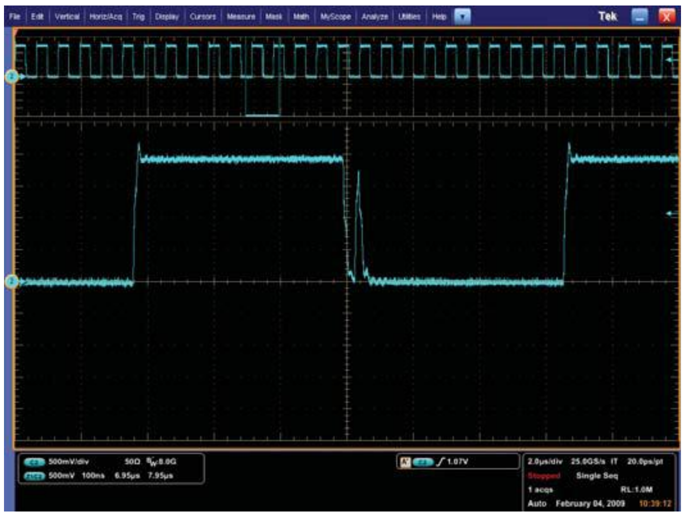

Figure 47. A higher sample rate provides greater signal resolution, ensuring that you'll see intermittent events.

Sample Rate

Sample rate – specified in samples per second (S/s) – refers to how frequently a digital oscilloscope takes a snapshot or sample of the signal, analogous to the frames on a movie camera. The faster an oscilloscope samples (i.e., the higher the sample rate), the greater the resolution and detail of the displayed waveform and the less likely that critical information or events will be lost, as shown in Figure 47. The minimum sample rate may also be important if you need to look at slowly changing signals over longer periods of time. Typically, the displayed sample rate changes with changes made to the horizontal scale control to maintain a constant number of waveform points in the displayed waveform record.

How do you calculate your sample rate requirements? The method differs based on the type of waveform you are measuring, and the method of signal reconstruction used by the oscilloscope.

In order to accurately reconstruct a signal and avoid aliasing, the Nyquist theorem states that the signal must be sampled at least twice as fast as its highest frequency component. This theorem, however, assumes an infinite record length and a continuous signal. Since no oscilloscope offers infinite record length and, by definition, glitches are not continuous, sampling at only twice the rate of highest frequency component is usually insufficient.

In reality, accurate reconstruction of a signal depends on both the sample rate and the interpolation method used to fill in the spaces between the samples. Some oscilloscopes let you select either sin (x)/x interpolation for measuring sinusoidal signals, or linear interpolation for square waves, pulses and other signal types.

For accurate reconstruction using sin(x)/x interpolation, your oscilloscope should have a sample rate at least 2.5 times the highest frequency component of your signal. Using linear interpolation, the sample rate should be at least 10 times the highest frequency signal component.

Some measurement systems with sample rates to 50 GS/s and bandwidths to 20 GHz have been optimized for capturing very fast, single-shot and transient events by oversampling up to 5 times the bandwidth.

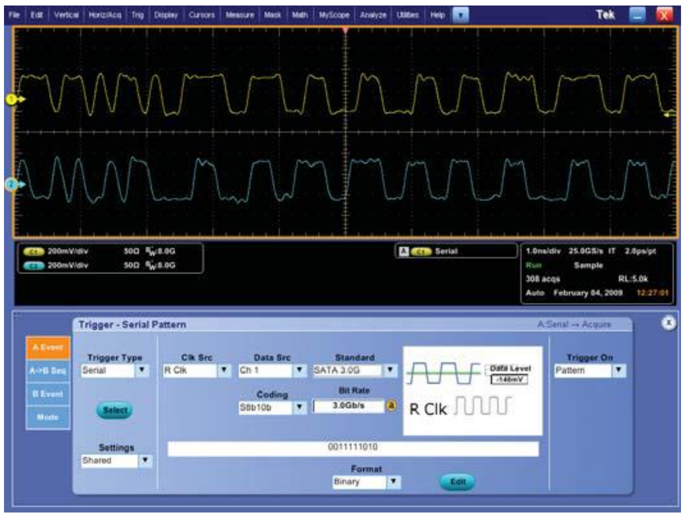
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
C1 200mV/div 50Ω 8.0G
C2 200mV/div 50Ω 8.0G
A C1 Serial
1.0ns/div 25.0GS/s IT 2.0ps/pt
Run Sample
308 acqs RL:5.0k
Auto February 04, 2008 12:27:01
A Event
A->B Seq
B Event
Mode
Select
Edit
R Clk

Trigger - Serial Pattern A:Serial → Acquire

Trigger Type: Serial
Clk Src: R Clk
Data Src: Ch 1
Standard: SATA 3.0G
Coding: S8b10b
Bit Rate: 3.0Gb/s
Data Level: -148mV
Trigger On: Pattern
Settings: Shared
0011111010
Format: Binary

Figure 48. A DPO provides an ideal solution for non-repetitive, high-speed, multi-channel digital design applications.

Waveform Capture Rate

All oscilloscopes blink. That is, they open their eyes a given number of times per second to capture the signal, and close their eyes in between. This is the waveform capture rate, expressed as waveforms per second (wfms/s). While the sample rate indicates how frequently the oscilloscope samples the input signal within one waveform, or cycle, the waveform capture rate refers to how quickly an oscilloscope acquires waveforms.

Waveform capture rates vary greatly, depending on the type and performance level of the oscilloscope. Oscilloscopes with high waveform capture rates provide significantly more visual insight into signal behavior, and dramatically increase the probability that the oscilloscope will quickly capture transient anomalies such as jitter, runt pulses, glitches and transition errors.

Digital storage oscilloscopes (DSOs) employ a serial-processing architecture to capture from 10 to 5,000 wfms/s. Some DSOs provide a special mode that bursts multiple captures into long memory, temporarily delivering higher waveform capture rates followed by long processing dead times that reduce the probability of capturing rare, intermittent events.

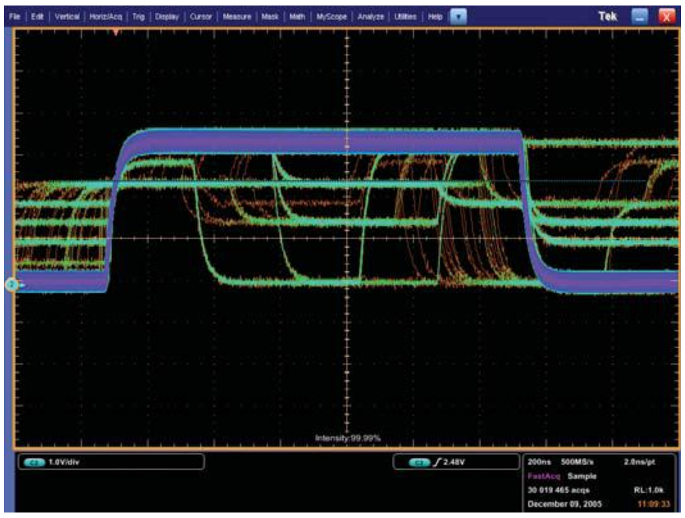
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursor
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
Intensity:99.99%
C1 1.0V/div
C1 2.48V
200ns 500MS/s 2.0ns/pt
FastAcq Sample
30 019 465 acqs RL:1.0k
December 09, 2005 11:09:33

Figure 49. A DPO enables a superior level of insight into signal behavior by delivering vastly greater waveform capture rates and three-dimensional display, making it the best general-purpose design and troubleshooting tool for a wide range of applications.

Most digital phosphor oscilloscopes (DPOs) employ a parallel-processing architecture to deliver vastly greater waveform capture rates. As seen in Figure 48, some DPOs can acquire millions of waveforms in just seconds, significantly increasing the probability of capturing intermittent and elusive events and allowing you to see the problems in your signal more quickly. Moreover, the DPO's ability to acquire and display three dimensions of signal behavior in real time – amplitude, time and distribution of amplitude over time – results in a superior level of insight into signal behavior, as shown in Figure 49.

Record Length

Record length, expressed as the number of points that comprise a complete waveform record, determines the amount of data that can be captured with each channel. Since an oscilloscope can store only a limited number of samples, the waveform duration (time) will be inversely proportional to the oscilloscope's sample rate.

$$\text{Time Interval} = \frac{\text{Record Length}}{\text{Sample Rate}}$$

Modern oscilloscopes allow you to select record length to optimize the level of detail needed for your application. If you are analyzing an extremely stable sinusoidal signal, you may need only a 500 point record length, but if you are isolating the causes of timing anomalies in a complex digital data stream, you may need a million points or more for a given record length, as demonstrated in Figure 50.

Triggering Capabilities

An oscilloscope's trigger function synchronizes the horizontal sweep at the correct point of the signal, essential for clear signal characterization. Trigger controls allow you to stabilize repetitive waveforms and capture single-shot waveforms.

Please refer to the Trigger section under Performance Terms and Considerations for more information regarding triggering capabilities.

Effective Bits

Effective bits represent a measure of a digital oscilloscope's ability to accurately reconstruct a sinewave signal's shape. This measurement compares the oscilloscope's actual error to that of a theoretical "ideal" digitizer. Because the actual errors include noise and distortion, the frequency and amplitude of the signal must be specified.

Frequency Response

Bandwidth alone is not enough to ensure that an oscilloscope can accurately capture a high frequency signal. The goal of oscilloscope design is a specific type of frequency response: Maximally Flat Envelope Delay (MFED). A frequency response of this type delivers excellent pulse fidelity with minimum overshoot and ringing. Since a digital oscilloscope is composed of real amplifiers, attenuators, ADCs, interconnects, and relays, MFED response is a goal that can only be approached. Pulse fidelity varies considerably with model and manufacturer.

Vertical Sensitivity

Vertical sensitivity indicates how much the vertical amplifier can amplify a weak signal – usually measured in millivolts (mV) per division. The smallest voltage detected by a general-purpose oscilloscope is typically about 1 mV per vertical screen division.

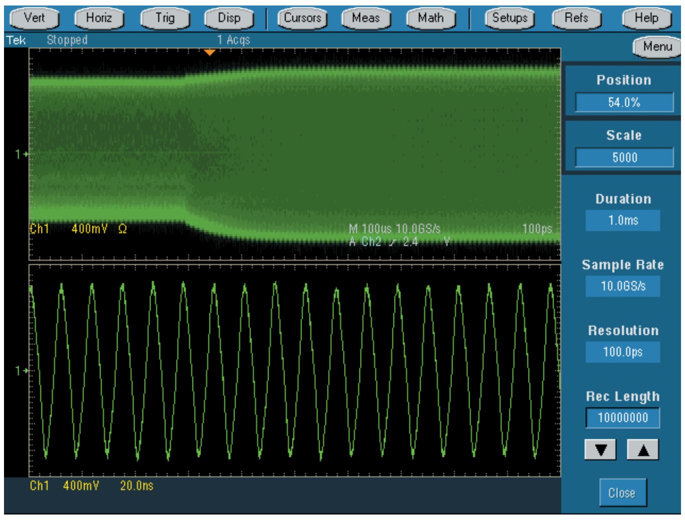

Position 54.0%
Scale 5000
Duration 1.0ms
Sample Rate 10.0GS/s
Resolution 100.0ps
Rec Length 10000000

Figure 50. Capturing the high frequency detail of this modulated 85 MHz carrier requires high resolution sampling (100 ps). Seeing the signal's complete modulation envelope requires a long time duration (1 ms). Using long record length (10 MB), the oscilloscope can display both.

Sweep Speed

Sweep speed indicates how fast the trace can sweep across the oscilloscope screen, enabling you to see fine details. The sweep speed of an oscilloscope is represented by time (seconds) per division.

Gain Accuracy

Gain accuracy indicates how accurately the vertical system attenuates or amplifies a signal, usually represented as a percentage error.

Horizontal Accuracy (Time Base)

Horizontal, or time base, accuracy indicates how accurately the horizontal system displays the timing of a signal, usually represented as a percentage error.

Vertical Resolution (Analog-to-Digital Converter)

Vertical resolution of the ADC, and therefore, the digital oscilloscope, indicates how precisely it can convert input voltages into digital values. Vertical resolution is measured in bits. Calculation techniques can improve the effective resolution, as exemplified with hi-res acquisition mode.

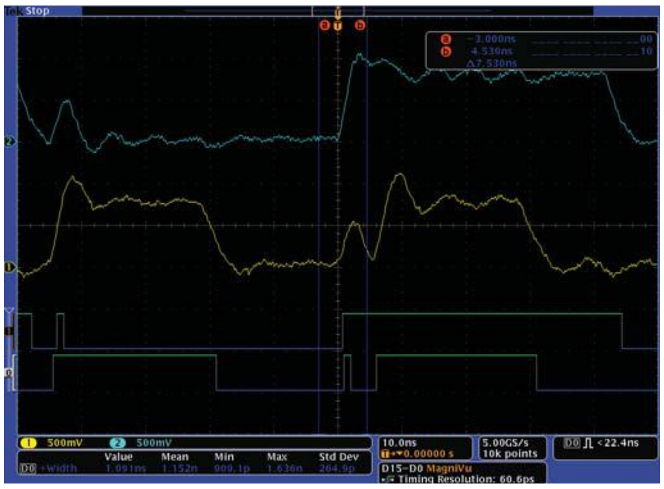

	Value	Mean	Min	Max	Std Dev
D0 +Width	1.091ns	1.152n	909.1p	1.636n	264.9p

Figure 51. The MSO provides 16 integrated digital channels, enabling the ability to view and analyze time-correlated analog and digital signals. A high speed timing acquisition provides more resolution to reveal narrow events such as glitches.

Timing Resolution (MSO)

An important MSO acquisition specification is the timing resolution used for capturing digital signals. Acquiring a signal with better timing resolution provides a more accurate timing measurement of when the signal changes. For example, a 500 MS/s acquisition rate has 2 ns timing resolution and the acquired signal edge uncertainty is 2 ns. A smaller timing resolution of 60.6 ps (16.5 GS/s) decreases the signal edge uncertainty to 60.6 ps and captures faster changing signals.

Some MSOs internally acquire digital signals with two types of acquisitions at the same time. The first acquisition is with standard timing resolution, and the second acquisition uses a high speed resolution. The standard resolution is used over a longer record length while the high speed timing acquisition offers more resolution around a narrow point of interest, as shown in Figure 51.

Connectivity

The need to analyze measurement results remains of utmost importance. The need to document and share information and measurement results easily and frequently has also grown in importance. The connectivity of an oscilloscope delivers advanced analysis capabilities and simplifies the documentation and sharing of results. As shown in Figure 52, standard interfaces (GPIB, RS-232, USB, Ethernet) and network communication modules enable some oscilloscopes to deliver a vast array of functionality and control.

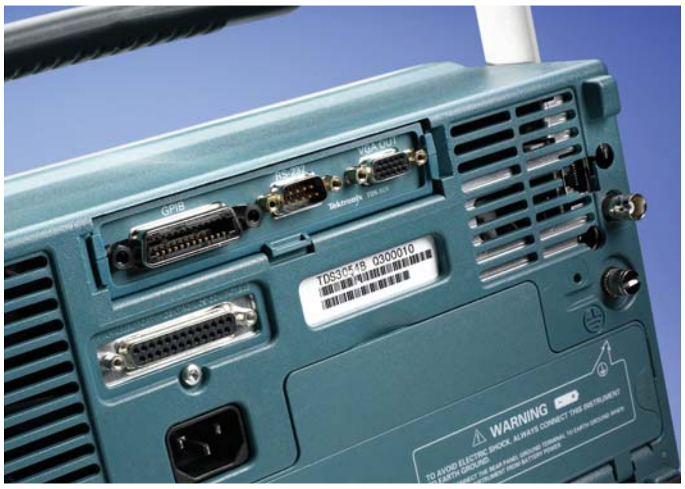

Figure 52. Today's oscilloscopes provide a wide array of communications interfaces, such as a standard Centronics port and optional Ethernet/RS-232, GPIB/RS-232, and VGA/RS-232 modules. There is even a USB port (not shown) on the front panel.

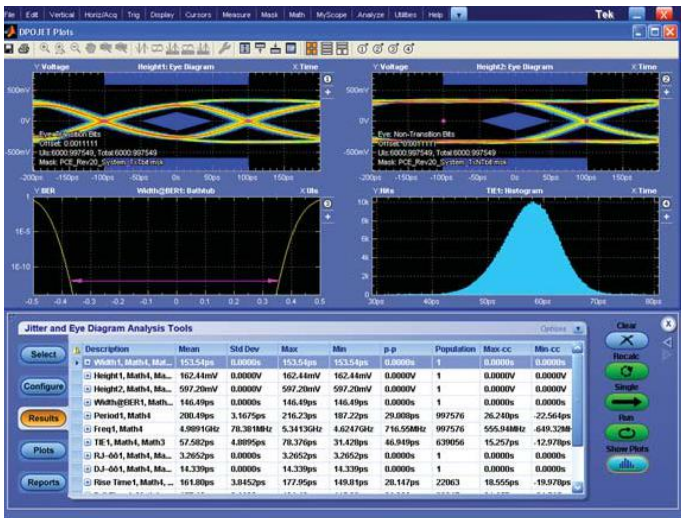

Jitter and Eye Diagram Analysis Tools

Description	Mean	Std Dev	Max	Min	p-p	Population	Max-cc	Min-cc
Width1, Math4, Mat...	153.54ps	0.0000s	153.54ps	153.54ps	0.0000s	1	0.0000s	0.0000s
Height1, Math4, Ma...	162.44mV	0.0000V	162.44mV	162.44mV	0.0000V	1	0.0000V	0.0000V
Height2, Math4, Ma...	597.20mV	0.0000V	597.20mV	597.20mV	0.0000V	1	0.0000V	0.0000V
Width@BER1, Math...	146.49ps	0.0000s	146.49ps	146.49ps	0.0000s	1	0.0000s	0.0000s
Period1, Math4	200.49ps	3.1675ps	216.23ps	187.22ps	29.008ps	997576	26.240ps	-22.564ps
Freq1, Math4	4.9891GHz	78.381MHz	5.3413GHz	4.6247GHz	716.55MHz	997576	555.94MHz	-649.32MH
TIE1, Math4, Math3	57.582ps	4.8895ps	78.376ps	31.428ps	46.949ps	639056	15.257ps	-12.978ps
RJ–δδ1, Math4, Ma...	3.2652ps	0.0000s	3.2652ps	3.2652ps	0.0000s	1	0.0000s	0.0000s
DJ–δδ1, Math4, Ma...	14.339ps	0.0000s	14.339ps	14.339ps	0.0000s	1	0.0000s	0.0000s
Rise Time1, Math4, ...	161.80ps	3.8452ps	177.95ps	149.81ps	28.147ps	22063	18.555ps	-19.978ps

Figure 53. Analysis software packages are specifically designed to meet jitter and eye measurement needs of today's high-speed digital designers.

Some advanced oscilloscopes also let you:

- Create, edit and share documents on the oscilloscope – all while working with the instrument in your particular environment
- Access network printing and file sharing resources
- Access the Windows® desktop
- Run third-party analysis and documentation software
- Link to networks
- Access the Internet
- Send and receive e-mail

Expandability

An oscilloscope should be able to accommodate your needs as they change. Some oscilloscopes allow you to:

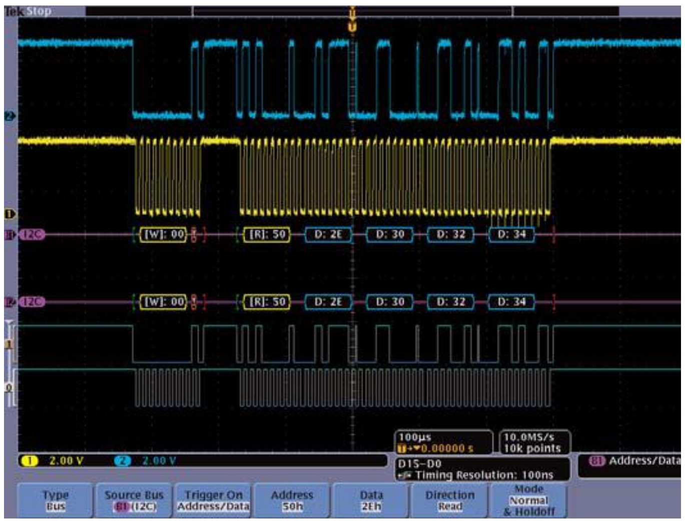
Stop
I2C
[W]: 00
[R]: 50
D: 2E
D: 30
D: 32
D: 34
2.00 V
2.00 V
100µs
10.0MS/s
10k points
Timing Resolution: 100ns
Address/Data
Type
Bus
Source Bus
Trigger On
Address/Data
Address
50h
Data
2Eh
Direction
Read
Mode
Normal
& Holdoff

Figure 54. Serial bus analysis is accelerated with automated trigger, decode, and search on serial packet context.

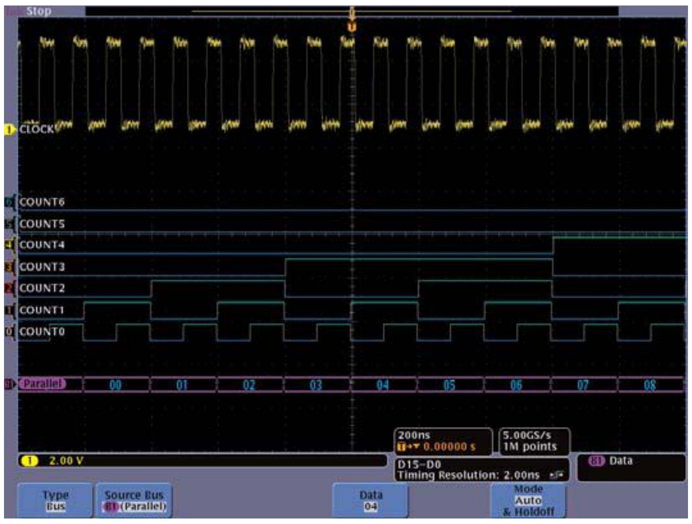
Stop
CLOCK
COUNT6
COUNT5
COUNT4
COUNT3
COUNT2
COUNT1
COUNT0
Parallel
00
01
02
03
04
05
06
07
08
2.00 V
200ns
5.00GS/s
1M points
Timing Resolution: 2.00ns
Data
Type
Bus
Source Bus
(Parallel)
Data
04
Mode
Auto
& Holdoff

Figure 55. Automatically trigger, decode, and search on clocked or unclocked parallel bus data.

- Add memory to channels to analyze longer record lengths
- Add application-specific measurement capabilities
- Complement the power of the oscilloscope with a full range of probes and modules
- Work with popular third-party analysis and productivity Windows-compatible software
- Add accessories, such as battery packs and rackmounts

Application modules and software may enable you to transform your oscilloscope into a highly specialized analysis tool capable of performing functions such as jitter and timing analysis, microprocessor memory system verification, communications standards testing, disk drive measurements, video measurements, power measurements and much more. Figures 53 - 58 highlight a few of these examples.

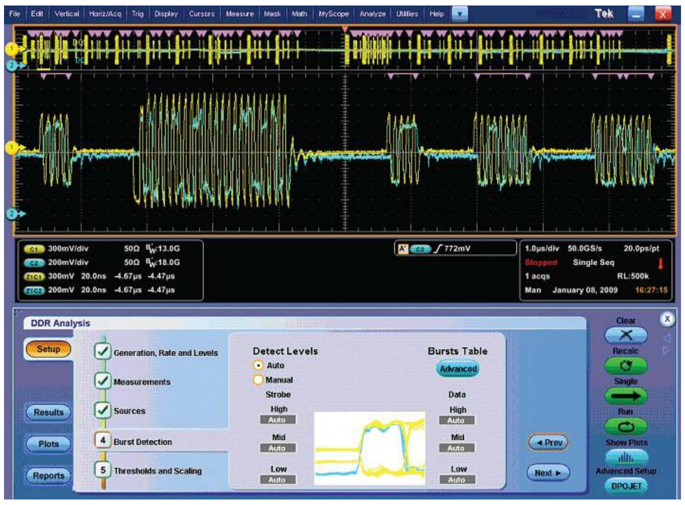
DDR Analysis
Setup
Results
Plots
Reports
Generation, Rate and Levels
Measurements
Sources
Burst Detection
Thresholds and Scaling
Detect Levels
Auto
Manual
Bursts Table
Advanced
Strobe
Data
High
Mid
Low
Prev
Next
Clear
Recalc
Single
Run
Show Plots
Advanced Setup
DPOJET

Figure 56. Advanced DDR analysis tools automate complex memory tasks like separating read/write bursts and performing JEDEC measurements.

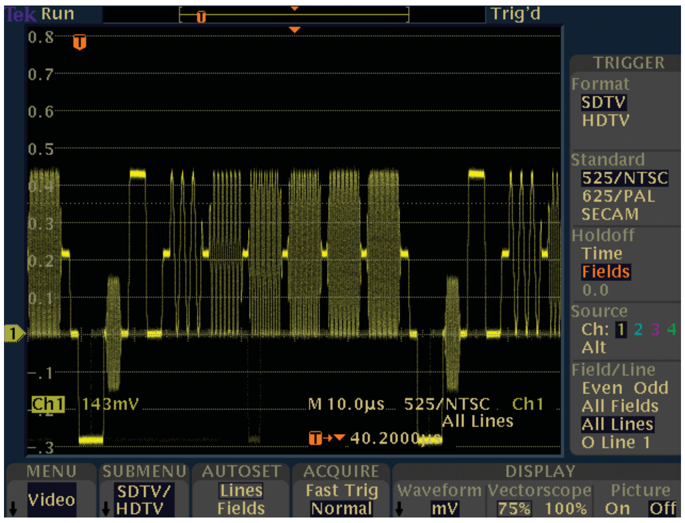
Run
Trig'd
TRIGGER
Format
SDTV
HDTV
Standard
525/NTSC
625/PAL
SECAM
Holdoff
Time
Fields
Source
Ch: 1 2 3 4
Alt
Field/Line
Even Odd
All Fields
All Lines
0 Line 1
Ch1 143mV
M 10.0µs
525/NTSC Ch1
All Lines
40.2000µs
MENU
Video
SUBMENU
SDTV/
HDTV
AUTOSET
Lines
Fields
ACQUIRE
Fast Trig
Normal
DISPLAY
Waveform
mV
Vectorscope
75% 100%
Picture
On Off

Figure 57. Video application modules make the oscilloscope a fast, tell-all tool for video troubleshooting.

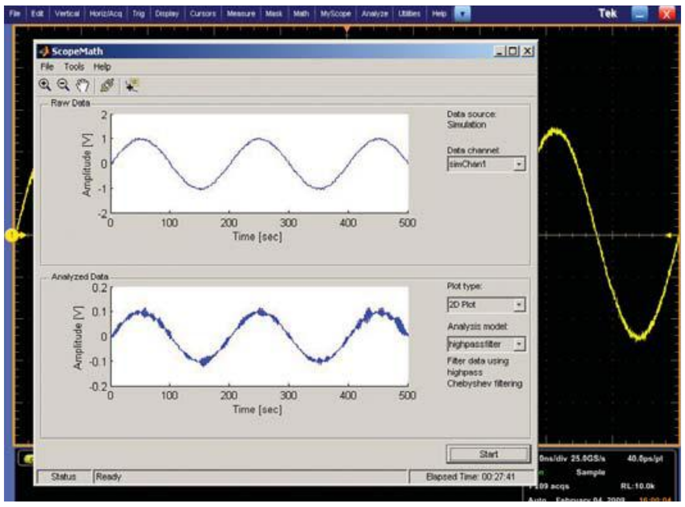
ScopeMath
File Tools Help
Raw Data
Amplitude [V]
Time [sec]
Data source:
Simulation
Data channel:
simChan1
Analyzed Data
Plot type:
2D Plot
Analysis model:
highpassfilter
Filter data using highpass Chebyshev filtering
Start
Status
Ready
Elapsed Time: 00:27:41

Figure 58. Advanced analysis and productivity software, such as MATLAB®, can be installed in Windows-based oscilloscopes to accomplish local signal analysis.

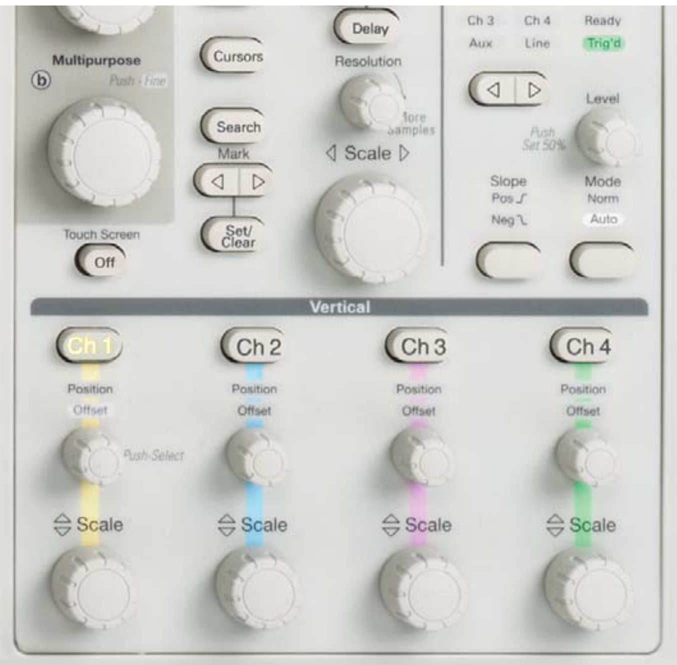
Multipurpose
Push - Fine
Cursors
Delay
Resolution
Search
Mark
Scale
Touch Screen
Off
Set/Clear
Ch 3
Ch 4
Ready
Aux
Line
Trig'd
Level
Push Set 50%
Slope
Pos
Neg
Mode
Norm
Auto
Vertical
Ch 1
Ch 2
Ch 3
Ch 4
Position
Offset
Push-Select
Scale

Figure 59. Traditional, analog-style knobs control position, scale, intensity, etc. – precisely as you would expect.

Ease-of-Use

Oscilloscopes should be easy to learn and easy to use, helping you work at peak efficiency and productivity. Allowing you to focus on your design, rather than the measurement tools. Just as there is no one typical car driver, there is no one typical oscilloscope user. Regardless of whether you prefer a traditional instrument interface, or that of a Windows® interface, it is important to have flexibility in your oscilloscope's operation.

Many oscilloscopes offer a balance between performance and simplicity by providing the user with many ways to operate the instrument. A front-panel layout, in Figure 59, provides dedicated vertical, horizontal and trigger controls. An icon-rich graphical user interface, as shown in Figure 60, helps you understand and intuitively use advanced capabilities. Touch-sensitive displays solve issues with cluttered benches and carts, while providing access to clear, on-screen buttons, as seen in Figure 61. Online help provides a convenient, built-in reference manual. Intuitive controls allow even occasional oscilloscope users to feel as comfortable driving the oscilloscope as they do driving a car, while giving full-time users easy access to the oscilloscope's most advanced features. In addition, many oscilloscopes are portable, such as the one shown in Figure 62, making the oscilloscope efficient in many different operating environments – in the lab or in the field.

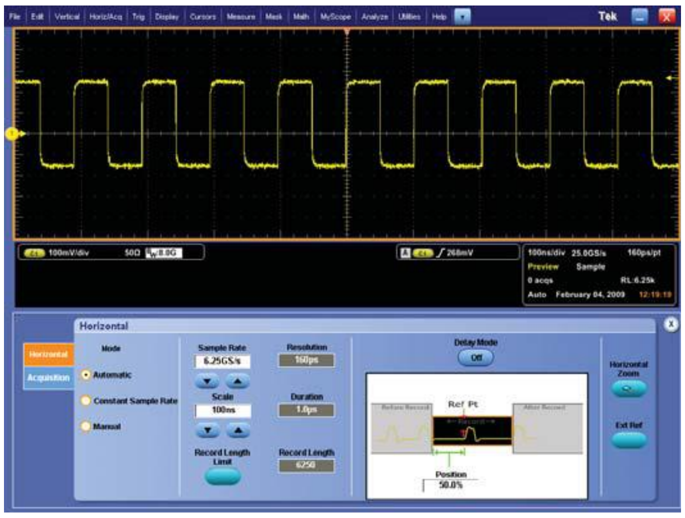
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
Horizontal
Acquisition
Mode
Automatic
Constant Sample Rate
Manual
Sample Rate
Scale
Record Length Limit
Resolution
Duration
Record Length
Delay Mode
Off
Ref Pt
Position
Horizontal Zoom
Ext Ref

Figure 60. Use graphical control windows to access even the most sophisiticated functions with confidence and ease.

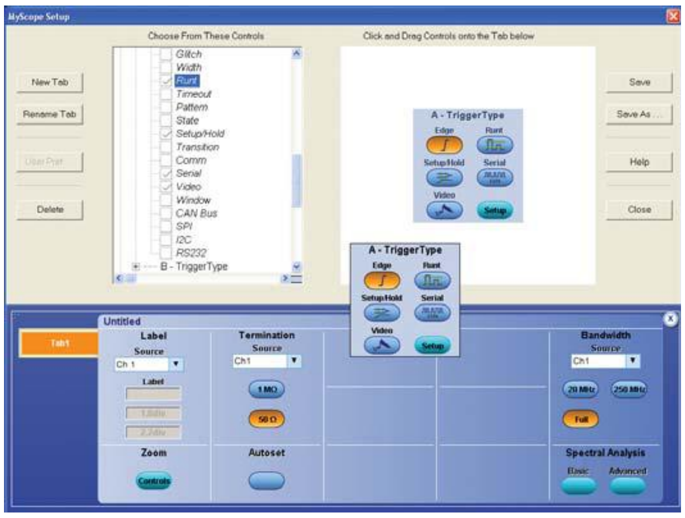
MyScope Setup
Choose From These Controls
Click and Drag Controls onto the Tab below
New Tab
Rename Tab
Delete
Save
Save As
Help
Close
Glitch
Width
Runt
Timeout
Pattern
State
Setup/Hold
Transition
Comm
Serial
Video
Window
CAN Bus
SPI
I2C
RS232
B - TriggerType
A - TriggerType
Edge
Runt
Setup/Hold
Serial
Video
Setup
Untitled
Tab1
Label
Source
Termination
Bandwidth
Zoom
Controls
Autoset
Spectral Analysis
Basic
Advanced
20 MHz
250 MHz
Full

Figure 61. Touch-sensitive display naturally solves issues with cluttered benches and carts, while providing access to clear, on-screen buttons.

Figure 62. The portability of many oscilloscopes makes the instrument efficient in many operating environments.

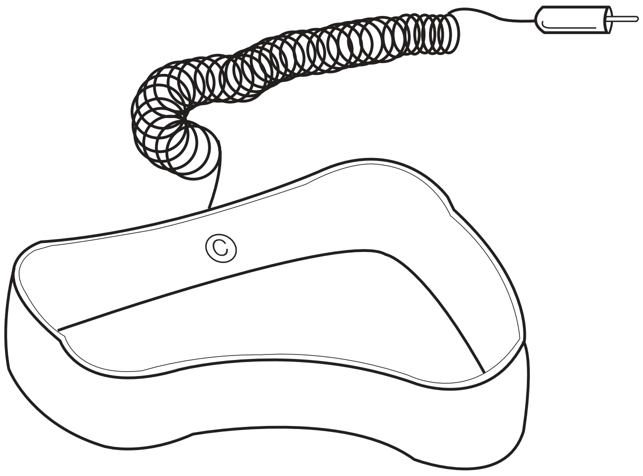

Figure 63. Typical wrist-type grounding strap.

Operating the Oscilloscope

This section briefly describes how to set up and start using an oscilloscope – specifically, how to properly ground the oscilloscope and yourself, set the oscilloscope controls, calibrate the oscilloscope, connect the probes, and compensate the probes.

Proper grounding is an important step when setting up to take measurements or work on a circuit. Proper grounding of the oscilloscope protects you from a hazardous shock and grounding yourself protects your circuits from damage.

Proper Grounding

To ground the oscilloscope means to connect it to an electrically neutral reference point, such as earth ground. Ground your oscilloscope by plugging its three-pronged power cord into an outlet grounded to earth ground.

Grounding the oscilloscope is necessary for safety. If a high voltage contacts the case of an ungrounded oscilloscope – any part of the case, including knobs that appear insulated – it can give you a shock. However, with a properly grounded oscilloscope, the current travels through the grounding path to earth ground rather than through you to earth ground.

Grounding is also necessary for taking accurate measurements with your oscilloscope. The oscilloscope needs to share the same ground as any circuits you are testing.

Some oscilloscopes do not require separate connection to earth ground. These oscilloscopes have insulated cases and controls, which keeps any possible shock hazard away from the user.

If you are working with integrated circuits (ICs), you also need to ground yourself. Integrated circuits have tiny conduction paths that can be damaged by static electricity that builds up on your body. You can ruin an expensive IC simply by walking across a carpet or taking off a sweater and then touching the leads of the IC. To solve this problem, wear a grounding strap, as shows in Figure 63. This strap safely sends static charges on your body to earth ground.

Setting the Controls

After plugging in the oscilloscope, take a look at the front panel. As described previously, the front panel is typically divided into three main sections labeled vertical, horizontal, and trigger. Your oscilloscope may have other sections, depending on the model and type.

Notice the input connectors on your oscilloscope – this is where you attach the probes. Most oscilloscopes have at least two input channels and each channel can display a waveform on the screen. Multiple channels are useful for comparing waveforms. As mentioned earlier, MSOs also have digital inputs as well.

Some oscilloscopes have AUTOSET and/or DEFAULT buttons that can set up the controls in one step to accommodate a signal. If your oscilloscope does not have this capability, it is helpful to set the controls to standard positions before taking measurements.

General instructions to manually set up the oscilloscope in standard positions are as follows:

- Set the oscilloscope to display channel 1
- Set the vertical volts/division scale and position controls to mid–range positions
- Turn off the variable volts/division
- Turn off all magnification settings
- Set the channel 1 input coupling to DC
- Set the trigger mode to auto
- Set the trigger source to channel 1
- Turn trigger holdoff to minimum or off
- Set the horizontal time/division and position controls to mid-range positions
- Adjust channel 1 volts/division such that the signal occupies as much of the 10 vertical divisions as possible without clipping or signal distortion

Calibrating the Instrument

In addition to proper oscilloscope setup, periodic instrument self-calibration is recommended for accurate measurements. Calibration is needed if the ambient temperature has changed more than 5° C (9° F) since the last self-calibration or once a week. In the oscilloscope menu this can sometimes be initiated as “Signal Path Compensation”. Refer to the manual that accompanied your oscilloscope for more detailed instructions.

Connecting the Probes

Now you are ready to connect a probe to your oscilloscope. A probe, if well-matched to the oscilloscope, will enable you to access all of the power and performance in the oscilloscope and will ensure the integrity of the signal you are measuring.

Measuring a signal requires two connections: the probe tip connection and the ground connection. Probes often come with a clip attachment for grounding the probe to the circuit under test. In practice, you attach the grounding clip to a known ground in the circuit, such as the metal chassis of a product you are repairing, and touch the probe tip to a test point in the circuit

Compensating the Probes

Passive attenuation voltage probes must be compensated to the oscilloscope. Before using a passive probe, you need to compensate it – to balance its electrical properties to a particular oscilloscope.

You should get into the habit of compensating the probe every time you set up your oscilloscope. A poorly adjusted probe can make your measurements less accurate. Figure 64 illustrates the effects on a 1 MHz test signal when using a probe that is not properly compensated.

Most oscilloscopes have a square wave reference signal available at a terminal on the front panel used to compensate the probe. General instructions to compensate the probe are as follows:

- Attach the probe to a vertical channel
- Connect the probe tip to the probe compensation, i.e. square wave reference signal
- Attach the ground clip of the probe to ground
- View the square wave reference signal
- Make the proper adjustments on the probe so that the corners of the square wave are square

Probe Undercompensated.

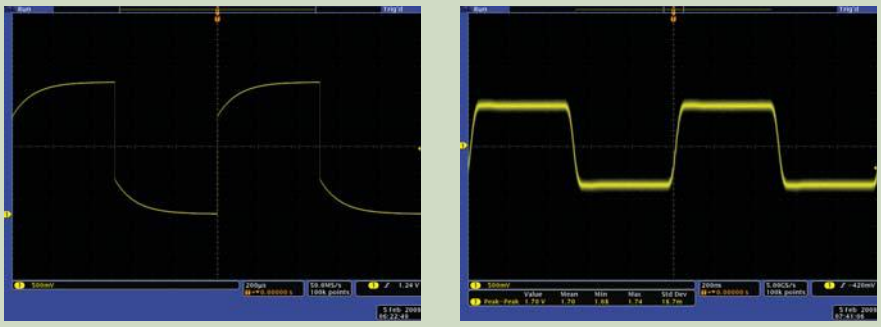

Probe Adjustment Signal

Note Reduced Amplitude

Probe Compensated Correctly.

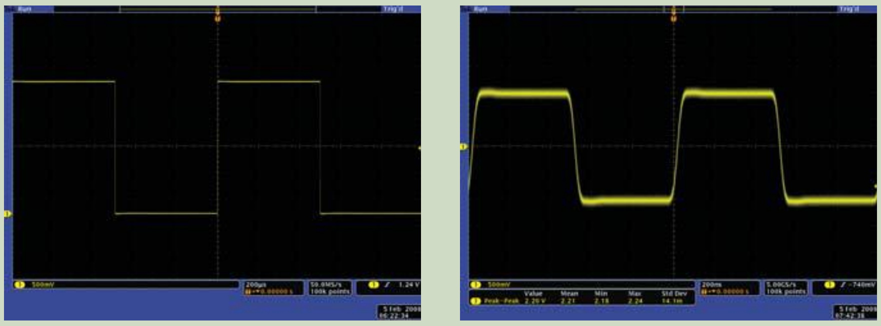

Probe Adjustment Signal

Note Proper Amplitude

Probe Overcompensated.

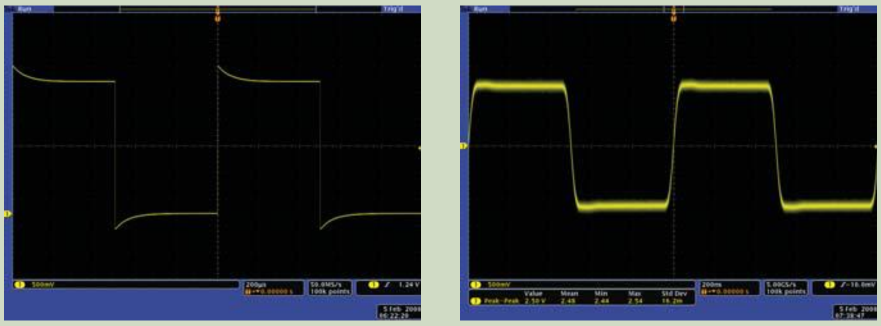

Probe Adjustment Signal

Note Increased Amplitude

Figure 64. The effects of improper probe compensation.

When you compensate the probe, always attach any accessory tips you will use and connect the probe to the vertical channel you plan to use. This will ensure that the oscilloscope has the same electrical properties as it does when you take measurements.

Oscilloscope Measurement Techniques

This section reviews basic measurement techniques. The two most basic measurements you can make are voltage and time measurements. Just about every other measurement is based on one of these two fundamental techniques.

This section discusses methods for taking measurements visually with the oscilloscope screen. This is a common technique with analog instruments, and also may be useful for "at-a-glance" interpretation of digital oscilloscope displays.

Note that most digital oscilloscopes include automated measurement tools that simplify and accelerate common analysis tasks, thus improving the reliability and confidence of your measurements. However, knowing how to make measurements manually as described here will help you understand and check the automatic measurements.

Voltage Measurements

Voltage is the amount of electric potential, expressed in volts, between two points in a circuit. Usually one of these points is ground (zero volts), but not always. Voltages can also be measured from peak-to-peak – from the maximum point of a signal to its minimum point. You must be careful to specify which voltage you mean.

The oscilloscope is primarily a voltage-measuring device. Once you have measured the voltage, other quantities are just a calculation away. For example, Ohm's law states that voltage between two points in a circuit equals the current times the resistance. From any two of these quantities you can calculate the third using the following formula:

$$\text{Voltage} = \text{Current } x \text{ Resistance}$$

$$\text{Current} = \frac{\text{Voltage}}{\text{Resistance}}$$

$$\text{Resistance} = \frac{\text{Voltage}}{\text{Current}}$$

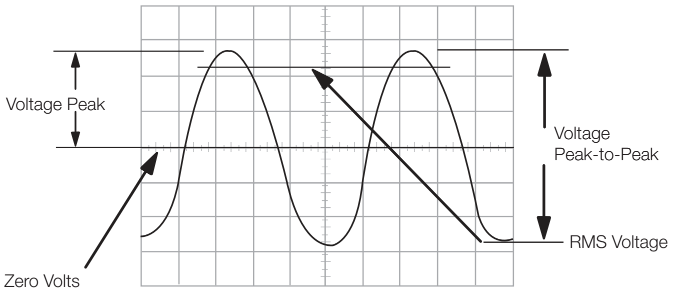

Figure 65. Voltage peak (V_p) and peak-to-peak voltage (V_{p-p})

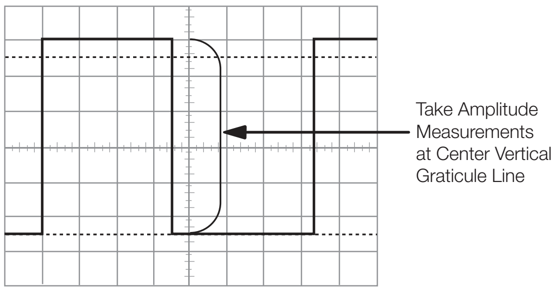
Take Amplitude Measurements at Center Vertical Graticule Line

Figure 66. Measure voltage on the center vertical graticule line.

Another handy formula is the power law, which states that the power of a DC signal equals the voltage times the current. Calculations are more complicated for AC signals, but the point here is that measuring the voltage is the first step toward calculating other quantities. Figure 65 shows the voltage of one peak (V_p) and the peak-to-peak voltage (V_{p-p}).

The most basic method of taking voltage measurements is to count the number of divisions a waveform spans on the oscilloscope's vertical scale. Adjusting the signal to cover most of the display vertically makes for the best voltage measurements, as shown in Figure 66. The more display area you use, the more accurately you can read the measurement.

Many oscilloscopes have cursors that let you make waveform measurements automatically, without having to count graticule marks. A cursor is simply a line that you can move across the display. Two horizontal cursor lines can be moved up and down to bracket a waveform's amplitude for voltage measurements, and two vertical lines move right and left for time measurements. A readout shows the voltage or time at their positions.

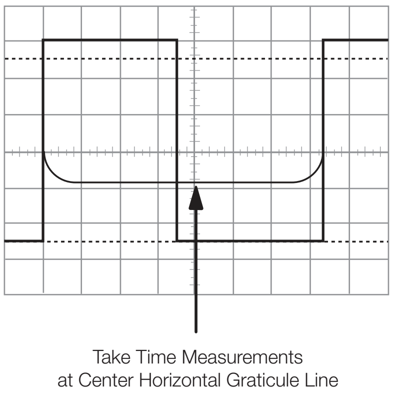

Figure 67. Measure time on the center horizontal graticule line.

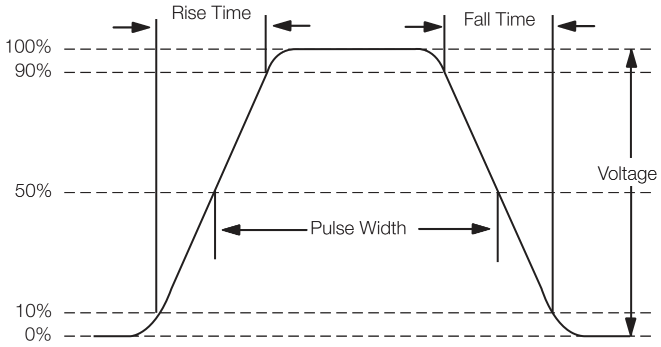

Figure 68. Rise time and pulse width measurement points.

Time and Frequency Measurements

You can make time measurements using the horizontal scale of the oscilloscope. Time measurements include measuring the period and pulse width of pulses. Frequency is the reciprocal of the period, so once you know the period, the frequency is one divided by the period. Like voltage measurements, time measurements are more accurate when you adjust the portion of the signal to be measured to cover a large area of the display, as illustrated in Figure 67.

Pulse Width and Rise Time Measurements

In many applications, the details of a pulse's shape are important. Pulses can become distorted and cause a digital circuit to malfunction, and the timing of pulses in a pulse train is often significant.

Standard pulse measurements are pulse rise time and pulse width. Rise time is the amount of time a pulse takes to go from a low to high voltage. By convention, the rise time is measured from 10% to 90% of the full voltage of the pulse. This eliminates any irregularities at the pulse's transition corners. Pulse width is the amount of time the pulse takes to go from low to high and back to low again. By convention, the pulse width is measured at 50% of full voltage. Figure 68 illustrates these measurement points.

Pulse measurements often require fine-tuning the triggering. To become an expert at capturing pulses, you should learn how to use trigger holdoff and how to set the digital oscilloscope to capture pretrigger data, as described in the Systems and Controls of an Oscilloscope section. Horizontal magnification is another useful feature for measuring pulses, since it allows you to see fine details of a fast pulse.

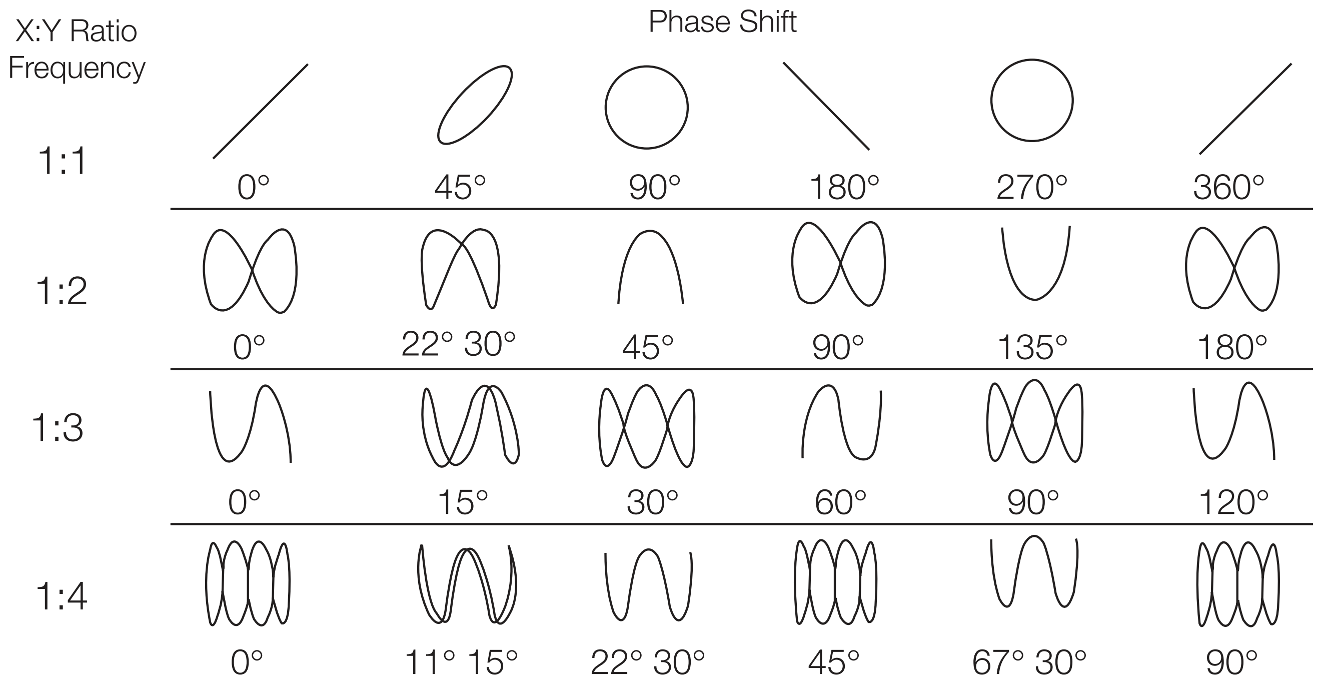

Figure 69. Lissajous patterns.

Phase Shift Measurements

One method for measuring phase shift – the difference in timing between two otherwise identical periodic signals – is to use XY mode. This measurement technique involves inputting one signal into the vertical system as usual and then another signal into the horizontal system – called an XY measurement because both the X and Y axis are tracing voltages. The waveform that results from this arrangement is called a Lissajous pattern (named for French physicist Jules Antoine Lissajous and pronounced LEE–sa–zhoo). From the shape of the Lissajous pattern, you can tell the phase difference between the two signals. You can also tell their frequency ratio. Figure 69 shows Lissajous patterns for various frequency ratios and phase shifts.

The XY measurement technique originated with analog oscilloscopes. DSOs may have difficulty creating real-time XY displays. Some DSOs create an XY image by accumulating triggered data points over time, then displaying two channels as an XY display.

DPOs, on the other hand, are able to acquire and display a genuine XY mode image in real-time, using a continuous stream of digitized data. DPOs can also display an XYZ image with intensified areas. Unlike XY displays on DSOs and DPOs, these displays on analog oscilloscopes are typically limited to a few megahertz of bandwidth.

Other Measurement Techniques

This section has covered basic measurement techniques. Other measurement techniques involve setting up the oscilloscope to test electrical components on an assembly line, capturing elusive transient signals, and many others. The measurement techniques you will use will depend on your application, but you have learned enough to get started. Practice using your oscilloscope and read more about it. Soon its operation will be second nature to you.

Written Exercises

This section contains written exercises that cover information in this primer. These are divided into two parts, Part I and Part II, with vocabulary and application exercises for each.

Check how well you have absorbed the information in these sections by verifying your responses against the answer key at the end of this section on page 54.

Part I covers information presented in these sections:

- The Oscilloscope
- Performance Terms and Considerations

Part II covers information presented in sections:

- The Systems and Controls of an Oscilloscope
- Operating the Oscilloscope
- Measurement Techniques

Part I A: Vocabulary Exercise

Write the letter of the definitions in the right column next to the correct words in the left column.

	Term		**Definition**
1. ___	Acquisition	**A**	The unit of electric potential difference.
2. ___	Analog	**B**	A performance measurement indicating the precision of an ADC, measured in bits.
3. ___	Bandwidth	**C**	Term used when referring to degree points of a signal's period.
4. ___	Digital Phosphor	**D**	The number of times a signal repeats in one second.
5. ___	Frequency	**E**	The amount of time it takes a wave to complete one cycle.
6. ___	Glitch	**F**	A stored digital value that represents the voltage of a signal at a specific point in time on the display.
7. ___	Period	**G**	A common waveform shape that has a rising edge, a width, and a falling edge.
8. ___	Phase	**H**	A performance measurement indicating the rising edge speed of a pulse.
9. ___	Pulse	**I**	Oscilloscope circuitry that controls the timing of the sweep.
10. ___	Waveform Point	**J**	An intermittent spike in a circuit.
11. ___	Rise Time	**K**	A signal measured by an oscilloscope that only occurs once.
12. ___	Sample Point	**L**	The oscilloscope's process of collecting sample points from the ADC, processing them, and storing them in memory.
13. ___	Digital Storage	**M**	Something that operates with continuously changing values.
14. ___	Time Base	**N**	Digital oscilloscope that captures 3 dimensions of signal information in real-time.
15. ___	Transient	**O**	Digital oscilloscope with serial processing.
16. ___	ADC Resolution	**P**	A sine wave frequency range, defined by the – 3 dB point.
17. ___	Volt	**Q**	The raw data from an ADC used to calculate and display waveform points.

Part I B: Application Exercise

Circle the best answers for each statement. Some statements have more than one right answer.

1. With an oscilloscope you can:

a. Calculate the frequency of a signal.

b. Find malfunctioning electrical components.

c. Analyze signal details.

d. All the above.

2. The difference between analog and digitizing oscilloscopes is:

a. Analog oscilloscopes do not have on-screen menus.

b. Analog oscilloscopes apply a measurement voltage directly to the display system, while digital oscilloscopes first convert the voltage into digital values.

c. Analog oscilloscopes measure analogs, whereas digitizing oscilloscopes measure digits.

d. Analog oscilloscopes do not have an acquisition system.

3. An oscilloscope's vertical section does the following:

a. Acquires sample points with an ADC.

b. Starts a horizontal sweep.

c. Lets you adjust the brightness of the display.

d. Attenuates or amplifies the input signal.

4. The time base control of the oscilloscope does the following:

a. Adjusts the vertical scale.

b. Shows you the current time of day.

c. Sets the amount of time represented by the horizontal width of the screen.

d. Sends a clock pulse to the probe.

5. On an oscilloscope display:

a. Voltage is on the vertical axis and time is on the horizontal axis.

b. A straight diagonal trace means voltage is changing at a steady rate.

c. A flat horizontal trace means voltage is constant.

d. All the above.

6. All repeating waves have the following properties:

a. A frequency measured in Hertz.

b. A period measured in seconds.

c. A bandwidth measured in Hertz.

d. All the above.

7. If you probe inside a computer with an oscilloscope, you are likely to find the following types of signals:

a. Pulse trains.

b. Ramp waves.

c. Sine waves.

d. All the above.

8. When evaluating the performance of an analog oscilloscope, some things you might consider are:

a. The bandwidth.

b. The vertical sensitivity.

c. The ADC resolution.

d. The sweep speed.

9. The difference between digital storage oscilloscopes (DSO) and digital phosphor oscilloscopes (DPO) is:

a. The DSO has a higher bandwidth.

b. The DPO captures three dimensions of waveform information in real-time.

c. The DSO has a color display.

d. The DSO captures more signal details.

Part II A: Vocabulary Exercise

Write the letter of the definitions in the right column next to the correct words in the left column.

	Term		**Definition**
1. ___	Averaging Mode	**A**	The unintentional interaction of the probe and oscilloscope with the circuit being tested which distorts a signal.
2. ___	Circuit Loading	**B**	A conductor that connects electrical currents to the Earth.
3. ___	Compensation	**C**	A sampling mode in which the digital oscilloscope collects as many samples as it can as the signal occurs, then constructs a display, using interpolation if necessary.
4. ___	Coupling	**D**	A sampling mode in which the digital oscilloscope constructs a picture of a repetitive signal by capturing a little bit of information from each repetition.
5. ___	Earth Ground	**E**	A device that converts a specific physical quantity such as sound, pressure, strain, or light intensity into an electrical signal.
6. ___	Equivalent-Time	**F**	A test device for injecting a signal into a circuit input.
7. ___	Graticule	**G**	A processing technique used by digital oscilloscopes to eliminate noise in a displayed signal.
8. ___	Interpolation	**H**	The method of connecting two circuits together.
9. ___	Real Time	**I**	A “connect-the-dots” processing technique to estimate what a fast waveform looks like based on only a few sampled points.
10. ___	Signal Generator	**J**	The grid lines on a screen for measuring oscilloscope traces.
11. ___	Single Sweep	**K**	A trigger mode that triggers the sweep once, must be reset to accept another trigger event.
12. ___	Sensor	**L**	A probe adjustment for 10X attenuator probes that balances the electrical properties of the probe with the electrical properties of the oscilloscope.

Part II B: Application Exercise

Circle the best answers for each statement. Some statements have more than one right answer.

1. To operate an oscilloscope safely, you should:

a. Ground the oscilloscope with the proper three-pronged power cord.

b. Learn to recognize potentially dangerous electrical components.

c. Avoid touching exposed connections in a circuit being tested even if the power is off.

d. All the above.

2. Grounding an oscilloscope is necessary:

a. For safety reasons.

b. To provide a reference point for making measurements.

c. To align the trace with the screen's horizontal axis.

d. All the above.

3. Circuit loading is caused by:

a. An input signal having too large a voltage.

b. The probe and oscilloscope interacting with the circuit being tested.

c. A 10X attenuator probe being uncompensated.

d. Putting too much weight on a circuit.

4. Compensating a probe is necessary to:

a. Balance the electrical properties of the 10X attenuator probe with the oscilloscope.

b. Prevent damaging the circuit being tested.

c. Improve the accuracy of your measurements.

d. All the above.

5. The trace rotation control is useful for:

a. Scaling waveforms on the screen.

b. Detecting sine wave signals.

c. Aligning the waveform trace with the screen's horizontal axis on an analog oscilloscope.

d. Measuring pulse width.

6. The volts per division control is used to:

a. Scale a waveform vertically.

b. Position a waveform vertically.

c. Attenuate or amplify an input signal.

d. Set the numbers of volts each division represents.

7. Setting the vertical input coupling to ground does the following:

a. Disconnects the input signal from the oscilloscope.

b. Causes a horizontal line to appear with auto trigger.

c. Lets you see where zero volts is on the screen.

d. All the above.

8. The trigger is necessary to:

a. Stabilize repeating waveforms on the screen.

b. Capture single-shot waveforms.

c. Mark a particular point of an acquisition.

d. All the above.

9. The difference between auto and normal trigger mode is:

a. In normal mode the oscilloscope only sweeps once and then stops.

b. In normal mode the oscilloscope only sweeps if the input signal reaches the trigger point; otherwise the screen is blank.

c. Auto mode makes the oscilloscope sweep continuously even without being triggered.

d. All the above.

10. The acquisition mode that best reduces noise in a repeating signal is:

a. Sample mode.

b. Peak detect mode.

c. Envelope mode.

d. Averaging mode.

11. The two most basic measurements you can make with an oscilloscope are:

a. Time and frequency measurements.

b. Time and voltage measurements.

c. Voltage and pulse width measurements.

d. Pulse width and phase shift measurements.

12. If the volts/division is set at 0.5, the largest signal that can fit on the screen (assuming an 8 x 10 division screen) is:

a. 62.5 millivolts peak-to-peak.

b. 8 volts peak-to-peak.

c. 4 volts peak-to-peak.

d. 0.5 volts peak-to-peak.

13. If the seconds/division is set at 0.1 ms, the amount of time represented by the width of the screen is:

a. 0.1 ms.

b. 1 ms.

c. 1 second.

d. 0.1 kHz.

14. By convention, pulse width is measured:

a. At 10% of the pulse's peak-to-peak (pk-pk) voltage.

b. At 50% of the pulse's peak-to-peak (pk-pk) voltage.

c. At 90% of the pulse's peak-to-peak (pk-pk) voltage.

d. At 10% and 90% of the pulse's peak-to-peak (pk-pk) voltage.

15. You attach a probe to your test circuit but the screen is blank. You should:

a. Check that the screen intensity is turned up.

b. Check that the oscilloscope is set to display the channel that the probe is connected to.

c. Set the trigger mode to auto since norm mode blanks the screen.

d. Set the vertical input coupling to AC and set the volts/division to its largest value since a large DC signal may go off the top or bottom of the screen.

e. Check that the probe isn't shorted and make sure it is properly grounded.

f. Check that the oscilloscope is set to trigger on the input channel you are using.

g. All of the above.

Answer Key

This section provides the answers to all written exercises in the previous section.

Part IA: Vocabulary Exercise Answers

1. L	5. D	9. G	13. O
2. M	6. J	10. F	14. I
3. P	7. E	11. H	15. K
4. N	8. C	12. Q	16. B
			17. A

Part IB: Application Exercise Answers

1. D	3. D	5. D	7. A
2. B,D	4. C	6. A,B	8. A,B,D
			9. B

Part IIA: Vocabulary Exercise Answers

1. G	4. H	7. J	10. F
2. A	5. B	8. I	11. K
3. L	6. D	9. C	12. E

Part IIB: Application Exercise Answers

1. D	5. C	9. B,C	13. B
2. A,B	6. A,C,D	10. D	14. B
3. B	7. D	11. B	15. G
4. A,C	8. D	12. C	

Glossary

A

Acquisition Mode – Modes that control how waveform points are produced from sample points. Some types include sample, peak detect, hi res, envelope, average, and waveform data base.

Alternating Current (AC) – A signal in which the current and voltage vary in a repeating pattern over time. Also used to indicate signal coupling type.

Amplification – An increase in signal amplitude during its transmission from one point to another.

Amplitude – The magnitude of a quantity or strength of a signal. In electronics, amplitude usually refers to either voltage or power.

Analog-to-Digital Converter (ADC) – A digital electronic component that converts an electrical signal into discrete binary values.

Analog Oscilloscope – An instrument that creates a waveform display by applying the input signal (conditioned and amplified) to the vertical axis of an electron beam moving across a cathode-ray tube (CRT) screen horizontally from left to right. A chemical phosphor coated on the CRT creates a glowing trace wherever the beam hits.

Analog Signal – A signal with continuously variable voltages.

Attenuation – A decrease in signal amplitude during its transmission from one point to another.

Averaging – A processing technique used by digital oscilloscopes to reduce noise in a displayed signal.

B

Bandwidth – A frequency range, usually limited by –3 dB.

C

Circuit Loading – The unintentional interaction of the probe and oscilloscope with the circuit being tested, distorting the signal.

Compensation – A probe adjustment for passive attenuation probes that balances the capacitance of the probe with the capacitance of the oscilloscope.

Coupling – The method of connecting two circuits together. Circuits connected with a wire are directly coupled (DC); circuits connected through a capacitor or transformer are indirectly (AC) coupled.

Cursor – An on–screen marker that you can align with a waveform to make more accurate measurements.

D

Delayed Time Base – A time base with a sweep that can start (or be triggered to start) relative to a pre-determined time on the main time base sweep. Allows you to see events more clearly and to see events that are not visible solely with the main time base sweep.

Digital Signal – A signal whose voltage samples are represented by discrete binary numbers.

Digital Oscilloscope – A type of oscilloscope that uses an analog-to-digital converter (ADC) to convert the measured voltage into digital information. Types include: digital storage, digital phosphor, mixed signal, and digital sampling oscilloscopes.

Digital Phosphor Oscilloscope (DPO) – A type of digital oscilloscope that closely models the display characteristics of an analog oscilloscope while providing traditional digital oscilloscope benefits (waveform storage, automated measurements, etc.) The DPO uses a parallel-processing architecture to pass the signal to the raster-type display, which provides intensity-graded viewing of signal characteristics in real time. The DPO displays signals in three dimensions: amplitude, time and the distribution of amplitude over time.

Digital Sampling Oscilloscope – A type of digital oscilloscope that employs equivalent-time sampling method to capture and display samples of a signal, ideal for accurately capturing signals whose frequency components are much higher than the oscilloscope's sample rate.

Digital Signal Processing – The application of algorithms to improve the accuracy of measured signals.

Digital Storage Oscilloscope (DSO) – A digital oscilloscope that acquires signals via digital sampling (using an analog-to-digital converter). It uses a serial-processing architecture to control acquisition, user interface, and the raster display.

Digitize – The process by which an analog-to-digital converter (ADC) in the horizontal system samples a signal at discrete points in time and converts the signal's voltage at these points into digital values called sample points.

Direct Current (DC) – A signal with a constant voltage and/or current. Also used to indicate signal coupling type.

Division – Measurement markings on the oscilloscope graticule indicating major and minor marks.

E

Earth Ground – A conductor that will connect electrical currents to the Earth.

Effective Bits – A measure of a digital oscilloscope's ability to accurately reconstruct a sine wave signal's shape. This measurement compares the oscilloscope's actual error to that of a theoretical "ideal" digitizer.

Envelope – The outline of a signal's highest and lowest points acquired over many displayed waveform repetitions.

Equivalent-time Sampling – A sampling mode in which the oscilloscope constructs a picture of a repetitive signal by capturing a little bit of information from each repetition. Two types of equivalent-time sampling: random and sequential.

F

Focus – The analog oscilloscope control that adjusts the cathode-ray tube (CRT) electron beam to control the sharpness of the display.

Frequency – The number of times a signal repeats in one second, measured in Hertz (cycles per second). The frequency equals 1/period.

Frequency Response – Frequency response curves of an oscilloscope define the accuracy in amplitude representation of the input signal in function of the signals frequency. In order to obtain maximum signal fidelity, it is important that the oscilloscope has a flat (stable) frequency response across the entire specified oscilloscopes bandwidth.

G

Gain Accuracy – An indication of how accurately the vertical system attenuates or amplifies a signal, usually represented as a percentage error.

Gigahertz (GHz) – 1,000,000,000 Hertz; a unit of frequency.

Glitch – An intermittent, high-speed error in a circuit.

Graticule – The grid lines on a display for measuring oscilloscope traces.

Ground –

1. A conducting connection by which an electric circuit or equipment is connected to the earth to establish and maintain a reference voltage level.
2. The voltage reference point in a circuit.

H

Hertz (Hz) – One cycle per second; the unit of frequency.

Horizontal Accuracy (Time Base) – An indication of how accurately the horizontal system displays the timing of a signal, usually represented as a percentage error.

Horizontal Sweep – The action of the horizontal system that causes a waveform to be drawn.

I

Intensity Grading – Frequency-of-occurrence information that is essential to understanding what the waveform is really doing.

Interpolation – A "connect-the-dots" processing technique to estimate what a fast waveform looks like based on only a few sampled points. Two types: linear and sin x/x.

K

Kilohertz (kHz) – 1,000 Hertz; a unit of frequency.

L

Loading – The unintentional interaction of the probe and oscilloscope with the circuit being tested which distorts a signal.

Logic Analyzer – An instrument used to make the logic states of many digital signals visible over time. It analyzes the digital data and can represent the data as real-time software execution, data flow values, state sequences, etc.

M

Megahertz (MHz) – 1,000,000 Hertz; a unit of frequency.

Megasamples per second (MS/s) – A sample rate unit equal to one million samples per second.

Microsecond (µs) – A unit of time equivalent to 0.000001 seconds.

Millisecond (ms) – A unit of time equivalent to 0.001 seconds.

Mixed Signal Oscilloscope (MSO) – A type of digital oscilloscope that combines the basic functionality of a 16-channel logic analyzer with the trusted performance of a 4-channel digital phosphor oscilloscope.

N

Nanosecond (ns) – A unit of time equivalent to 0.000000001 seconds.

Noise – An unwanted voltage or current in an electrical circuit.

O

Oscilloscope – An instrument used to make voltage changes visible over time. The word oscilloscope comes from “oscillate,” since oscilloscopes are often used to measure oscillating voltages.

P

Peak (V_p) – The maximum voltage level measured from a zero reference point.

Peak Detection – An acquisition mode available with digital oscilloscopes that enables you to observe signal details that may otherwise be missed, particularly useful for seeing narrow pulses spaced far apart in time.

Peak-to-peak (V_{p-p}) – The voltage measured from the maximum point of a signal to its minimum point.

Period – The amount of time it takes a wave to complete one cycle. The period equals 1/frequency.

Phase – The amount of time that passes from the beginning of a cycle to the beginning of the next cycle, measured in degrees.

Phase Shift – The difference in timing between two otherwise similar signals.

Pre-trigger Viewing – The ability of a digital oscilloscope to capture what a signal did before a trigger event. Determines the length of viewable signal both preceding and following a trigger point.

Probe – An oscilloscope input device, usually having a pointed metal tip for making electrical contact with a circuit element, a lead to connect to the circuit’s ground reference, and a flexible cable for transmitting the signal and ground to the oscilloscope.

Pulse – A common waveform shape that has a fast rising edge, a width, and a fast falling edge.

Pulse Train – A collection of pulses traveling together.

Pulse Width – The amount of time the pulse takes to go from low to high and back to low again, conventionally measured at 50% of full voltage.

R

Ramps – Transitions between voltage levels of sine waves that change at a constant rate.

Raster – A type of display.

Real-time Sampling – A sampling mode in which the oscilloscope collects as many samples as possible from one triggered acquisition. Ideal for signals whose frequency range is less than half the oscilloscope’s maximum sample rate.

Record Length – The number of waveform points used to create a record of a signal.

Rise Time – The time taken for the leading edge of a pulse to rise from its low to its high values, typically measured from 10% to 90%.

S

Sampling – The conversion of a portion of an input signal into a number of discrete electrical values for the purpose of storage, processing and/or display by an oscilloscope. Two types: real-time sampling and equivalent-time sampling.

Sample Point – The raw data from an ADC used to calculate waveform points.

Sample Rate – Refers to how frequently a digital oscilloscope takes a sample of the signal, specified in samples per second (S/s).

Sensor – A device that converts a specific physical quantity such as sound, pressure, strain, or light intensity into an electrical signal.

Signal Integrity – The accurate reconstruction of a signal, determined by the systems and performance considerations of an oscilloscope, in addition to the probe used to acquire the signal.

Signal Source – A test device used to inject a signal into a circuit input; the circuit’s output is then read by an oscilloscope. Also known as a signal generator.

Sine Wave – A common curved wave shape that is mathematically defined.

Single Shot – A signal measured by an oscilloscope that only occurs once (also called a transient event).

Single Sweep – A trigger mode to display one triggered screen of a signal and then stop.

Slope – On a graph or an oscilloscope display, the ratio of a vertical distance to a horizontal distance. A positive slope increases from left to right, while a negative slope decreases from left to right.

Square Wave – A common wave shape consisting of repeating square pulses.

Sweep – One horizontal pass of an anlog oscilloscope's electron beam from left to right across the CRT screen.

Sweep Speed – Same as the time base.

T

Time Base – Oscilloscope circuitry that controls the timing of the sweep. The time base is set by the seconds/division control.

Trace – The visible shapes drawn on a CRT by the movement of the electron beam.

Transient – A signal measured by an oscilloscope that only occurs once (also called a single–shot event).

Trigger – The circuit that references a horizontal sweep on an oscilloscope.

Trigger Holdoff – A control that allows you to adjust the period of time after a valid trigger during which the oscilloscope cannot trigger.

Trigger Level – The voltage level that a trigger source signal must reach before the trigger circuit initiates a sweep.

Trigger Mode – A mode that determines whether or not the oscilloscope draws a waveform if it does not detect a trigger. Common trigger modes include normal and auto.

Trigger Slope – The slope that a trigger source signal must reach before the trigger circuit initiates a sweep.

V

Vertical Resolution (Analog-to-Digital Converter) – An indication of how precisely an analog-to-digital converter (ADC) in a digital oscilloscope can convert input voltages into digital values, measured in bits. Calculation techniques, such as hi res acquisition mode, can improve the effective resolution.

Vertical Sensitivity – An indication of how much the vertical amplifier can amplify a weak signal – usually measured in millivolts (mV) per division.

Volt – The unit of electric potential difference.

Voltage – The difference in electric potential, expressed in volts, between two points.

W

Wave – The generic term for a pattern that repeats over time. Common types include: sine, square, rectangular, sawtooth, triangle, step, pulse, periodic, non-periodic, synchronous, asynchronous.

Waveform – A graphic representation of a voltage varying over time.

Waveform Capture Rate – Refers to how quickly an oscilloscope acquires waveforms, expressed as waveforms per second (wfms/s).

Waveform Point – A digital value that represents the voltage of a signal at a specific point in time. Waveform points are calculated from sample points and stored in memory.

Writing Speed – The ability of an analog oscilloscope to provide a visible trace of the movement of a signal from one point to another. This ability is restrictive for low-repetition signals that have fast-moving details, such as digital logic signals.

XY Mode – A measuremenbt technique that involves inputting one signal into the vertical system, as usual, and one into the horizontal system to trace voltages on both the X and Y axis.

Z

Z Axis – The display attribute on an oscilloscope that shows brightness variations as the trace is formed.

Contact Tektronix:

ASEAN / Australasia (65) 6356 3900
Austria 00800 2255 4835*
Balkans, Israel, South Africa and other ISE Countries +41 52 675 3777
Belgium 00800 2255 4835*
Brazil +55 (11) 3759 7600
Canada 1 (800) 833-9200
Central East Europe, Ukraine and the Baltics +41 52 675 3777
Central Europe & Greece +41 52 675 3777
Denmark +45 80 88 1401
Finland +41 52 675 3777
France 00800 2255 4835*
Germany 00800 2255 4835*
Hong Kong 400-820-5835
India (91) 80-42922600
Italy 00800 2255 4835*
Japan 81 (3) 6714-3010
Luxembourg +41 52 675 3777
Mexico, Central/South America & Caribbean 52 (55) 53 35 10 85
Middle East, Asia and North Africa +41 52 675 3777
The Netherlands 00800 2255 4835*
Norway 800 16098
People's Republic of China 400-820-5835
Poland +41 52 675 3777
Portugal 80 08 12370
Republic of Korea 82 (2) 6917-5000
Russia & CIS +7 (495) 7484900
South Africa +27 11 206 8360
Spain 00800 2255 4835*
Sweden 00800 2255 4835*
Switzerland 00800 2255 4835*
Taiwan 886 (2) 2722-9622
United Kingdom & Ireland 00800 2255 4835*
USA 1 (800) 833-9200

* European toll-free number. If not accessible, call: +41 52 675 3777

Contact List Updated 09 December 2009

For Further Information
Tektronix maintains a comprehensive, constantly expanding collection of application notes, technical briefs and other resources to help engineers working on the cutting edge of technology. Please visit **www.tektronix.com**

Copyright © 2009, Tektronix. All rights reserved. Tektronix products are covered by U.S. and foreign patents, issued and pending. Information in this publication supersedes that in all previously published material. Specification and price change privileges reserved. TEKTRONIX and TEK are registered trademarks of Tektronix, Inc. All other trade names referenced are the service marks, trademarks or registered trademarks of their respective companies.

12/09 JS 03W-8605-5
TEK1511

MORE ELECTRONIC GADGETS FOR THE EVIL GENIUS

40 NEW BUILD-IT-YOURSELF PROJECTS

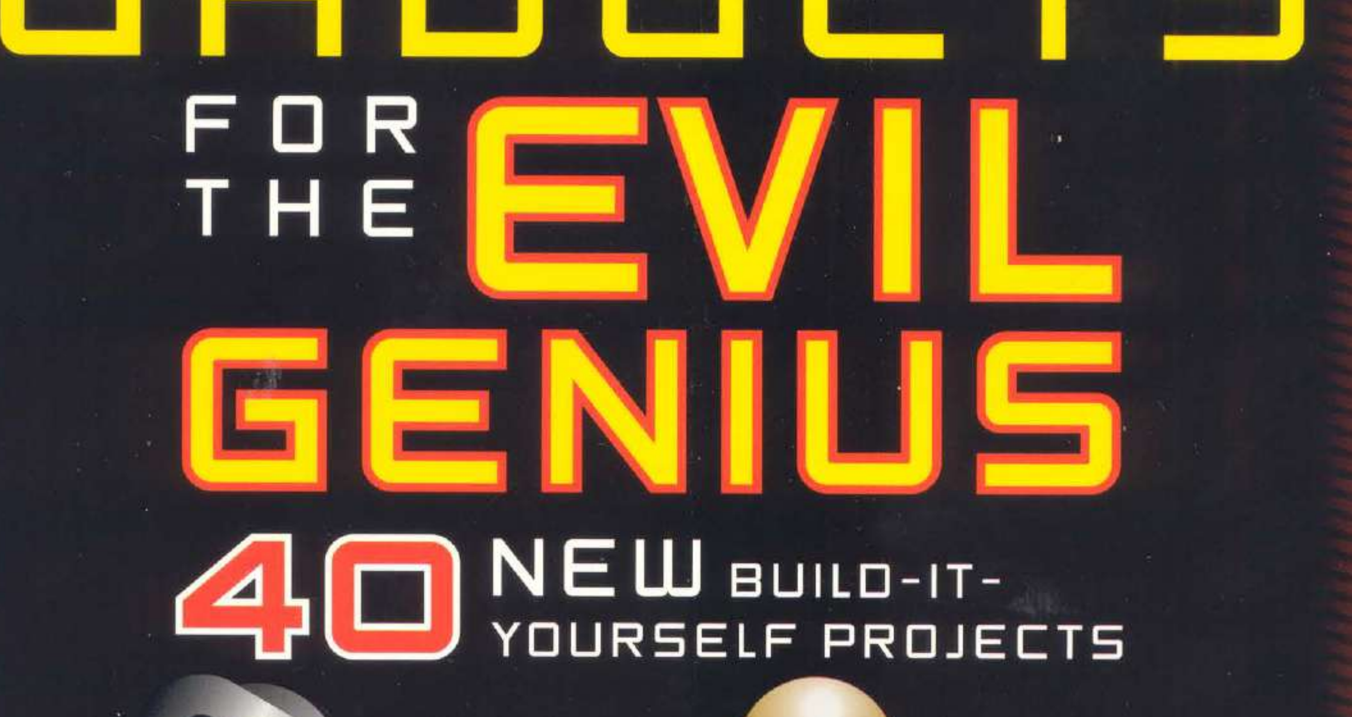

- Every project includes a list of materials, sources for obtaining parts, schematics, documentation, and clear instructions
- Projects are fun and inexpensive, but challenging enough for an electronics wiz
- Build your own lasers, surveillance devices, ultrasonic generators, display and detection devices

TAB ELECTRONICS

BOB IANNINI

Contents

Chapter One Battery-Powered Infrared Pulsed Laser 1

Theory of Operation 2

Circuit Theory of Operation 2

Current Monitor 5

Charging Circuits 5

Chassis Assembly 5

Assembly Testing 7

Final Assembly 8

Special Note 8

Chapter Two High-Speed Laser Pulse Detector 11

Circuit Description 12

System Assembly 13

Assembly Testing 15

Final Assembly 16

Special Note 17

Special Note on Photo Detectors 17

Chapter Three Ultra-Bright Green Laser Project 19

Assembly Steps 20

Chapter Four 115 VAC, 5- to 50-Watt Pulsed Infrared Laser 25

Theory of Operation 25

Circuit Theory of Operation 26

Notes on SCR 28

Notes on the Storage Capacitor 28

Notes on Layout Wiring 28

Notes on Circuit Monitoring 30

Assembly Steps 30

Assembly Testing 33

Final Assembly Steps 35

Chapter Five Laser Property Guard 37

System Installation Suggestions 39

Chapter Six 30-Milliwatt, 980-Nanometer Infrared Laser 43

Assembly 44

Collimator and Final Assembly 47

Chapter Seven High-Voltage, High-Frequency Driver Module 51

Electrical and Mechanical Specifications 51

Circuit Description 52

Construction 52

Board Assembly Steps 52

Testing Steps 56

Chapter Eight Negative Ion Machine with Reaction-Emitting Rotor 59

Benefits of Negative Ions 59

Circuit Description 60

Assembly Steps 61

Notes 62

Chapter Nine Kirlian Imaging Project 65

Assembly Steps 66

Testing Steps for the Circuit 66

Notes 67

Chapter Ten Plasma Etching and Burning Pen 69

Driver Circuit Description 69

Drive Circuit Assembly 70

Driver Board Testing 71

Mechanical Assembly 72

Sample Demonstration 72

Chapter Eleven How to Electrify Objects and Vehicles 75
Supplementary Electrification Data and Information 76
Assembly of Suggested Power Sources 81
Circuit Description of Pulsed Shocker 81
Construction Steps for Pulsed Shocker 81
Notes Pulsed Shocker 85
Construction of Continuous Shocker 85

Chapter Twelve Electromagnetic Pulse (EMP) Gun 87
Basic Theory Simplified 87
Circuit Description 89
Circuit Assembly 89
Notes 89

Chapter Thirteen Microwave Cannon 93

Chapter Fourteen Induction Heater 103
Circuit Operation 103
Project Assembly 104
Project Testing 108
Using the Heater Project 108

Chapter Fifteen 50-Kilovolt Laboratory DC Supply 113
Basic Description 113
Circuit Operation 114
Assembly of the Driver Section 114
Pretesting the Driver Section 116
Assembly of the Multiplier Section 121
Final Testing 125
Applications 129

Chapter Sixteen Magnetic High-Impact Cannon 133
Theory of Operation 133
Circuit Theory 134
Assembly Steps 136
Circuit Testing 136
Assembly of the Completed System 137
Testing and Operation of the Cannon 143

Chapter Seventeen Vacuum Tube Tesla Coil Project 151
Basic Description 153
Assembly Steps 154
Testing and Operating 167
Notes 167

Chapter Eighteen Universal Capacitance Discharge Ignition (CDI) Driver 169
Circuit Description 169
Assembly Steps 170
Test Steps 175
Notes on Operation 175

Chapter Nineteen Long-range Telephone Conversation Transmitter 177
Circuit Description 177
Assembly of the Circuit Board 179
Notes 180

Chapter Twenty Line-Powered Telephone Conversation Transmitter 183
Circuit Description 183
Assembly of the Circuit Board 185
Notes 186

Chapter Twenty-One Remote Wireless FM Repeater 187
Circuit Description 187
Assembly 188
Notes 189

Chapter Twenty-Two Tracking and Homing Transmitter 191
Circuit Description 191
Assembly of the Circuit Board 192
Notes 194

Chapter Twenty-Three Snooper Phone Room-Listening Device 195
Circuit Operation 195

Construction Steps 196

Final Assembly 197

Testing Steps 198

Notes 199

Chapter Twenty-Four Long-Range FM Voice Transmitter 201

Circuit Description 201

Assembly Steps 202

Notes 204

Chapter Twenty-Five FM Pocket Radio and TV Disrupter 205

Circuit Theory 206

Assembly Steps 206

Chapter Twenty-Six Ozone Generator for Water Treatment 209

Introduction to the Benefits of Ozone 209

Ozone for Water Applications 210

Ways to Generate Ozone 210

Chapter Twenty-Seven Therapeutic Magnetic Pulser 213

Circuit Description 214

Assembly 215

Chapter Twenty-Eight Noise Curtain Generator 219

Circuit Description 220

Assembly Steps 220

Testing Steps 221

Chapter Twenty-Nine Mind-Synchronizing Generator 223

Information on Mind Control 223

Operating a Mind Machine 225

Circuit Theory 226

Assemble the Board 227

Chapter Thirty Alternative Health Multiwave Machine 231

Theory of Operation 231

Construction of the Multiwave Coil and Antenna 233

Operation 234

Chapter Thirty-One Mind Mangler 237

Circuit Operation 237

Construction Steps 238

Testing 239

Chapter Thirty-Two 500-Milligram Ozone Air Purification System 241

The Story of Ozone 241

Ozone for Treating Air 242

Construction Steps 242

Special Notes 244

Chapter Thirty-Three Invisible Pain-field Generator 249

Circuit Description 250

Construction 251

Chapter Thirty-Four Canine Controller 257

Device Description 257

Circuit Description 258

Circuit Assembly 260

Testing the Circuit 263

Chapter Thirty-Five Ultrasonic Phaser Pain-Field Generator 265

Basic Device Description 265

Circuit Theory 266

Construction Steps 267

Testing the Assembly 268

Basic Operating Instructions 269

Information on the System 272

A Word of Caution 272

General Information on Ultrasonics 272

Application Supplement 272

Chapter Thirty-Six Magnetic Distortion Detector 275

Circuit Description 276

Assembly Steps 277

Pretesting the Sense Head 278

Construction of the Control Box 279
Control Box Protesting 279
Applications 281
Optional Ultra-Sensitive Vibration and Tremor Detector 283

Chapter Thirty-Seven Body Heat Detector 285
Description 285
Body Heat Detector Modes 286
Circuit Description 286
Assembly Instructions 288
Operating Instructions 290
Final Notes 291

Chapter Thirty-Eight Ion and Field Detector 293
Circuit Description 293
Construction Steps 294
Applications 298

Chapter Thirty-Nine Light Saber Recycling Stand 301
Circuit Description 301
Assembly 302
Parts Fabrication 302
Operation 303

Chapter Forty Andromeda Plasma Lamp 309
Circuit Description 309
Assembly Board Wiring 310

Index 317

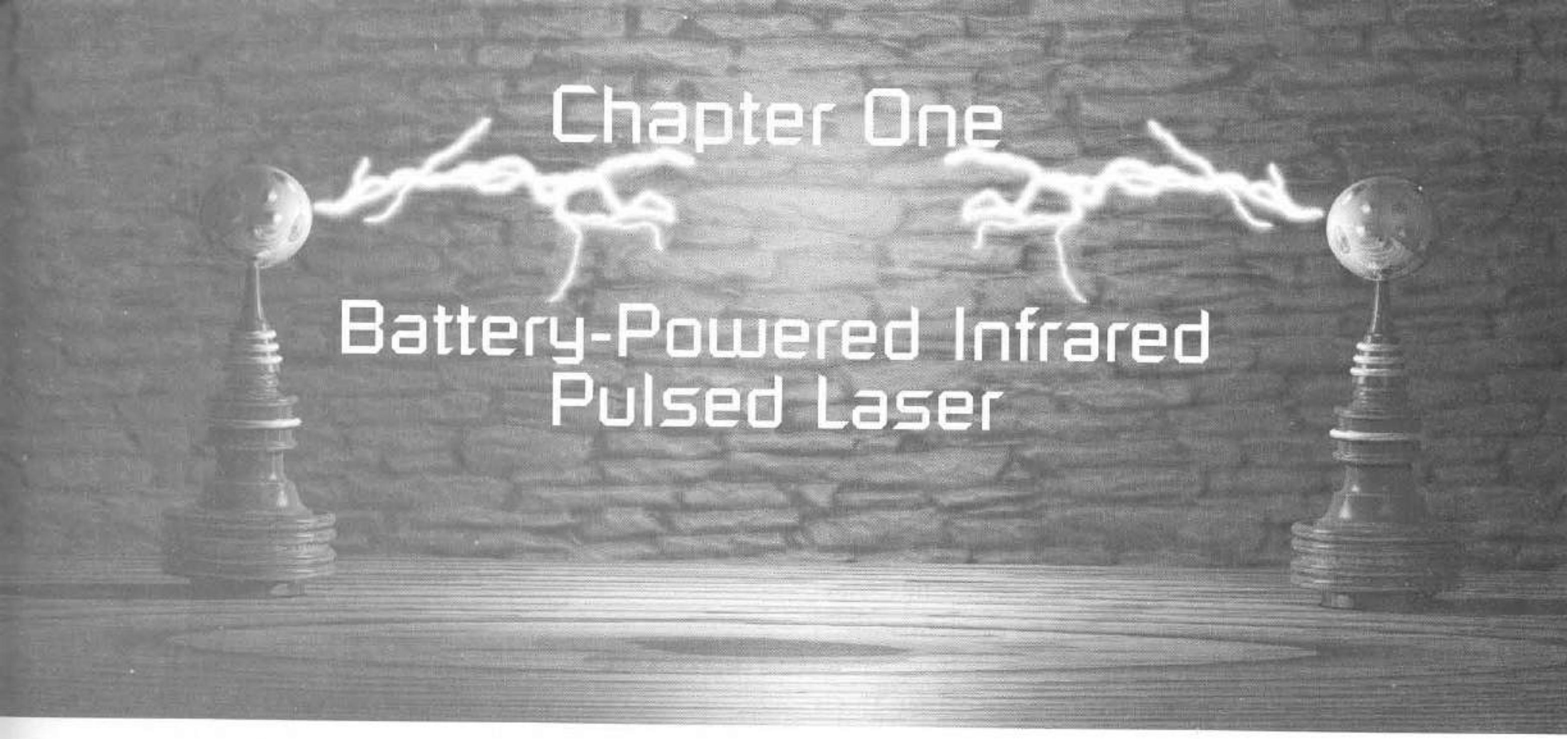

Chapter One

Battery-Powered Infrared Pulsed Laser

This project shows how to construct a low-powered, portable, solid-state, pulsed infrared laser. The system uses a gallium arsenide laser diode and provides pulse powers from 10 to 100 watts, depending on the diode used. The device operates from batteries and is completely self-contained. It may be built in a pistol, rifle, or a simple tubular configuration. The device is intended as a source of adjustable-frequency pulses from 10 to 1,000 repetitions per second of infrared energy at 900 nanometers

The system is shown in Figure 1-1 as being built on a perforated board assembly and combination copper ground plane that all fits into a tubular enclosure with a lens and holder for a collimator. The housing serves as the enclosure for the batteries and contains the control panel at its rear. It may be fitted with a handle that can hold an optional trigger switch when the device is designed in a gun configuration. A conventional sighting system is also easily adapted to the device.

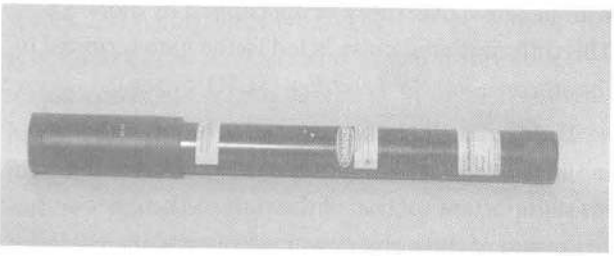

Figure 1-1 *Photograph of laser*

The laser is intended as a rifle-type simulated weapon, whose range can be several miles or as a long-range, laser-beam protection fence with a similar range of several miles. It is intended to be used with our high-speed laser pulse detector described in Chapter 2.

The laser, when assembled as shown, is a class 3B FDA certified device and requires the appropriate labeling and several included safety functions, as described in the assembly instructions. At no time should it be pointed at anyone without protective eyewear or at anything that could reflect the pulses. Never look into the unit when the power is on. The device is intended to be used for ranging, simulated weapons practice, intrusion detection, communications and signaling, and a variety of related scientific, optical experiments and uses.

This is an intermediate- to advanced-level project requiring electronic skills and basic electronic shop equipment. Expect to spend $100 to $150. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 1-1.

Theory of Operation

A laser diode is nothing more than a three-layer device consisting of a *pn* junction of *n*-type silicon, a *p* type of gallium arsenide, and a third *p* layer of doped gallium arsenide with aluminum. The *n*-type material contains electrons that readily migrate across the *pn* junction and fill the holes of the *p*-type material. Conversely, holes in the *p* type migrate to the *n* type and join with electrons. This migration causes a potential hill or barrier consisting of negative charges in the *p*-type material and positive charges in the *n*-type material that eventually ceases growing when a charge equilibrium exists. In order for current to flow in this device, it must be supplied at a voltage to overcome this potential barrier. This is the forward voltage drop across a common diode. If this voltage polarity is reversed, the potential barrier is simply increased, assuring no current flow. This is the reversed bias condition of a common diode.

A diode without an external voltage applied to it contains electrons that move and wander through the lattice structure at a low, lazy average velocity as a function of temperature. When an external current at a voltage exceeding the barrier potential is applied, these lazy electrons increase their velocity so that some of them, by colliding, acquire a discrete amount of energy and become unstable. They eventually emit the acquired energy in the form of a photon after returning to a lower-energy state. These photons of energy are random both in time and direction; hence, any radiation produced is incoherent, such as that of an LED.

The requirement for coherent radiation is that the discrete packets of radiation must be in the form of a lockstep phase and in a definite direction. This demands two essential requirements: first, sufficient electrons at the necessary excited energy levels, and second, an optical resonant cavity capable of trapping these energized electrons to stimulate more electrons and give them direction. The amount of energized electrons is determined by the forward diode current. A definite threshold condition exists where the device emits laser light rather than incoherent light, such as in an LED. This is why the device must be pulsed with high current. The radiation from these energized electrons is reflected back and forth between the square-cut edges of the crystal that form the reflecting surfaces due to the index of refraction of the material and air.

The electrons are initially energized in the region of the *pn* junction. When these energized electrons drift into the *p*-type transparent region, they spontaneously liberate other photons that travel back and forth in the optical cavity interacting with other electrons, commencing laser action. A portion of the radiation traveling back and forth between the reflecting surfaces of these mirrors escapes and constitutes the output of the device.

Circuit Theory of Operation

Figure 1-2 shows the inverter section increasing the 12 volts of the portable battery pack to 200 to 300 volts performed by the circuit, which consists of a switching transistor (Q1) and step-up transformer (T1). Q1 conducts until saturated for a time until the base reverse biases and can no longer sustain it at an on state and Q1 turns off. This causes the magnetic field in its collector winding (COL) to collapse, thus producing a stepped-up voltage in the secondary (SEC) of proper phase. The variable resistor (R3) controls the charging current to the capacitor (C2), the transistor turnoff time, and consequently system power.

The stepped-up square wave voltage on the secondary of T1 is rectified by diode (D1) and integrated onto the storage capacitor (C3). The trigger circuit determines the pulse rep rate of the laser and uses a timer (I1) whose pulse rate is determined by the timing resistor and capacitor (R9 and C6). R9 is adjustable for changing the laser pulse rate. The output trigger pulse is differentiated by capacitor (C4) with negative overshoot being clipped by diode D2. This differentiated pulse is fed to the gate terminal of the *silicon controlled rectifier* (SCR) switch.

The discharge circuit generates the current pulse in the laser diode (LD1) and consequently is the most important section of the pulser. The basic configuration of the pulse power supply is shown in the system schematic. The current pulse is generated by the charging storage capacitor C3 being switched through the SCR and laser diode LD1. The rise time

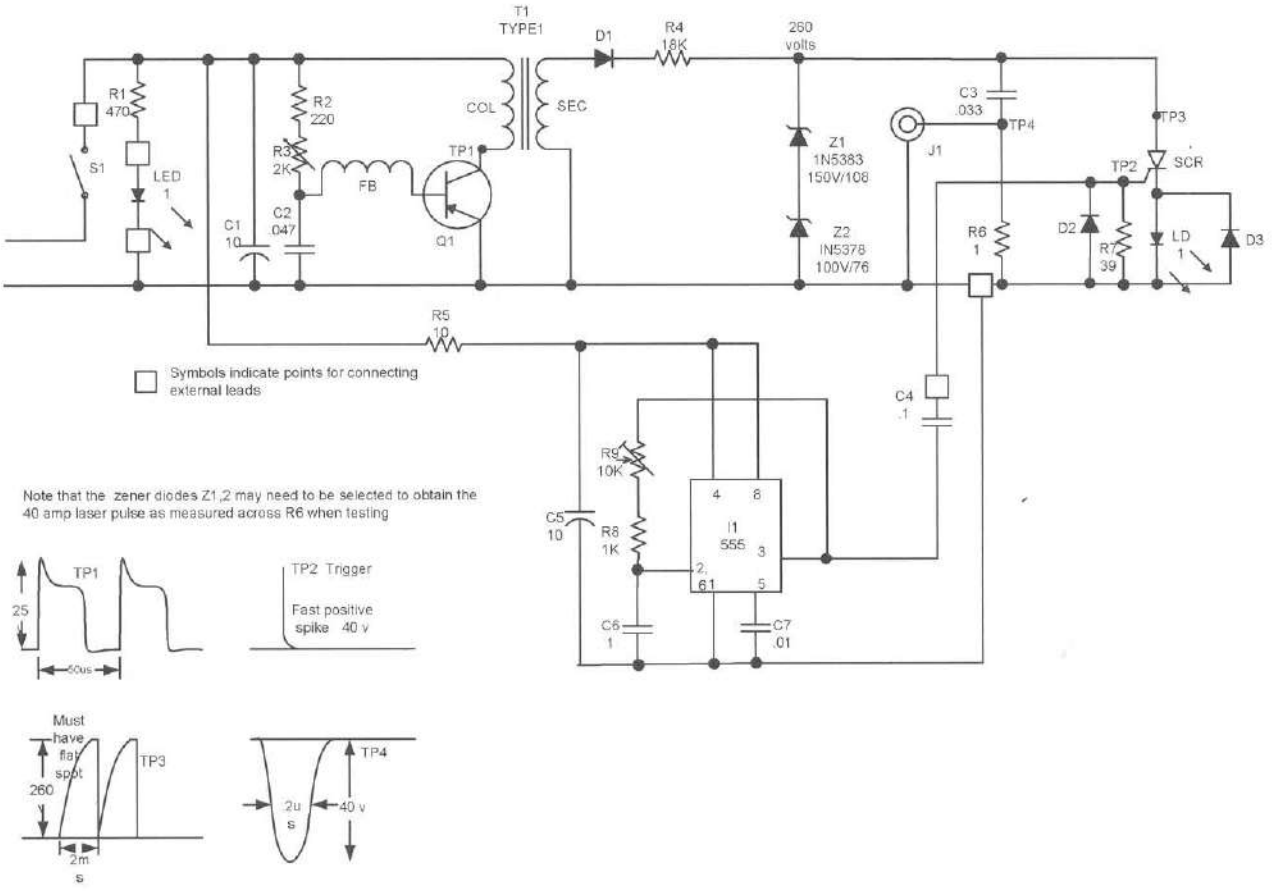
T1
TYPE1
D1
R4
18K
260
volts
R1
470
S1
LED
1
R2
220
R3
2K
COL
SEC
TP1
FB
Q1
C1
10
C2
.047
Z1
1N5383
150V/108
Z2
1N5378
100V/76
J1
C3
.033
TP4
TP3
TP2
SCR
R6
1
D2
R7
39
LD
1
D3
R5
10
Symbols indicate points for connecting external leads
C4
.1
R9
10K
R8
1K
C5
10
C6
1
C7
.01
I1
555
4
8
3
2,
61
5
Note that the zener diodes Z1,2 may need to be selected to obtain the 40 amp laser pulse as measured across R6 when testing
TP1
25
50us
TP2 Trigger
Fast positive spike 40 v
Must have flat spot
TP3
260
2m s
TP4
2u s
40 v

Figure 1-2 *Circuit schematic*

of the current pulse is usually determined by the SCR, while the fall time is determined by the capacitor value and the total resistance in the discharge circuit.

Figure 1-3 shows the typical anode voltage and current waveforms of the SCR during the current pulse through the diode laser. The peak current, pulse width, and voltage of the capacitor discharge circuit are related for various load and capacitance values. The peak laser current and charged capacitor voltage relationships are given for several different capacitor values and typical laser types. The voltage and current limits of the SCR are also shown. Short pulse widths provide less time for the SCR to turn on than longer pulse widths; therefore, the SCR impedance is higher and more voltage is required to generate the same current. Also shown are the current pulse waveforms for the three different values of the capacitance. The capacitor is charged to the same voltage in all three cases, that is, 400 volts.

In conventional SCR operation, the anode current, initiated by a gate pulse, rises to its maximum value in about 1 microsecond. During this time, the anode-to-cathode impedance drops from an open circuit to a fraction of an ohm. In injection laser pulsers, however, the duration of the anode-cathode pulse is much less than the time required for the SCR to turn on completely. Therefore, the anode-to-cathode impedance is at the level of 1 to 10 ohms throughout most of the conduction period.

The major disadvantage of the high SCR impedance is that it causes low circuit efficiency. For example, at a current of 40 amps, the maximum voltage would be across the SCR, while only 9 volts would be across the laser diode. These values represent very low circuit efficiency.

The efficiency of a laser array is greater due to its circuit impedance being more significant. Because the SCR is used unconventionally, many of the standard specifications such as peak current reverse

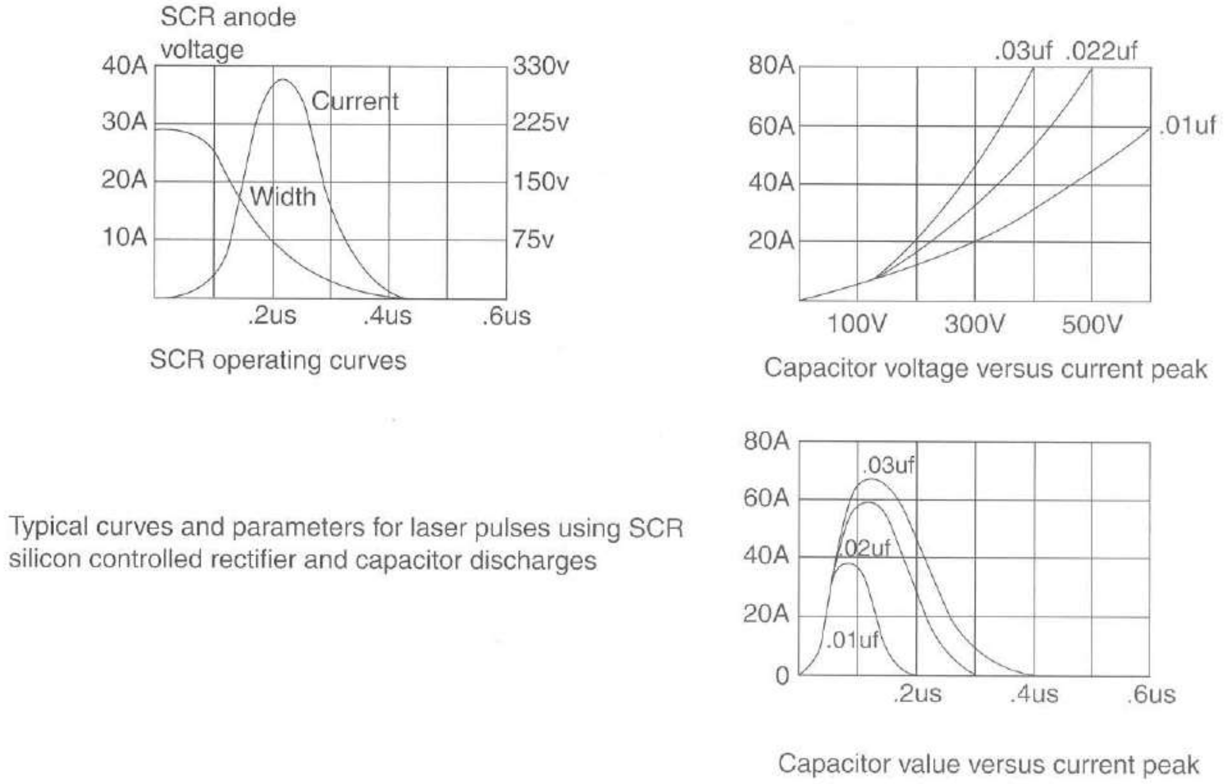

SCR operating curves

Capacitor voltage versus current peak

Typical curves and parameters for laser pulses using SCR silicon controlled rectifier and capacitor discharges

Capacitor value versus current peak

Figure 1-3 *SCR operating curves*

voltage, on-state forward voltage, and turn-off time are not applicable. In fact, it is difficult to select an SCR for a pulsing circuit on the basis of normally specified characteristics. The specifications important to laser pulser applications are the forward-blocking voltage and current rise time. A *use test* is the best and often the only practical method of determining the suitability of a particular SCR.

The voltage rating of the storage capacitor must be at least as high as the supply voltage. With the exception of ceramic types, most capacitors (metallized paper, mica, etc.) will perform well in this circuit. Ceramic capacitors have noticeably greater series resistance but are usable in slower speed pulsing circuits.

Lead lengths and circuit layout are very important to the performance of the discharge circuit. Lead inductance affects the rise time and peak value of the current, and it can also produce ringing and undershoot in the current waveform that can destroy the laser. A well-built discharge circuit might have a total lead length of only 1 inch and therefore an inductance of approximately 20 nanohenries. If the current rises to 75 amperes in 100 nanoseconds, the inductive voltage drop across this lead can be 15 volts. It is now obvious that if proper care is not taken in wiring the discharge circuits, high-inductive voltage drops will result.

A 1-ohm resistor in the discharge circuit will greatly reduce the current undershoot in single-diode lasers. Laser arrays usually have sufficient resistance to eliminate undershoot. The small resistance in the discharge circuit is also useful in monitoring the laser current, as described in the following section.

A clamping diode (D3) is added in parallel with the laser to reduce the current undershoot. Its polarity should be opposite that of the laser. Although the clamping diode is operated above its usual maximum current rating, the current undershoot caused by ringing is very short and the operating life of the diode is satisfactory.

Current Monitor

The current monitor in the discharge circuit provides a means of observing the laser current's waveform with an oscilloscope. A resistive-type monitor (R6) reduces circuit ringing and current undershoot, but the lead inductance of the resistor may cause a current reading that is higher than it actually is. A current transformer such as the Tektronix CT-2 can also be used to monitor the current and is not affected by lead inductance. Because the transformer does not respond to low-frequency signals, it should be used with fast waveforms that have a short pulse width and a fast fall time.

Charging Circuits

The second major section of the pulser is the charging circuit. This circuit charges the capacitor to the supply voltage during the time interval between the laser current pulses. It also isolates the supply voltage from the discharge circuit during the laser current pulse, thereby allowing the SCR to recover to the blocking state. Because the response times of the charging circuit are relatively long, lead lengths are not important and the circuit can be remotely located from the discharge circuit.

The simplest charging circuit is a resistor-cap combination. The resistor must limit the current to a value less than the SCR holding current, but it should be as low as practical, because this resistance also determines the charging time of the capacitor, C3.

Chassis Assembly

The following is a list of steps to construct the laser project.

1. Fabricate the copper chassis, as shown in Figure 1-4. This part is shown in Table 1-1 as being available.

2. Assemble the board as shown in Figure 1-5. Lay out and identify all the parts and pieces. Separate the components that go to the assembly board and fabricate the (PB1) perfboard from a 1.4 × 5-inch piece of .1 × .1 grid. Note the two holes for attachment to the copper chassis section from Figure 1-6.

3. Assemble the components as shown and observe the polarity of the capacitors and semiconductors.

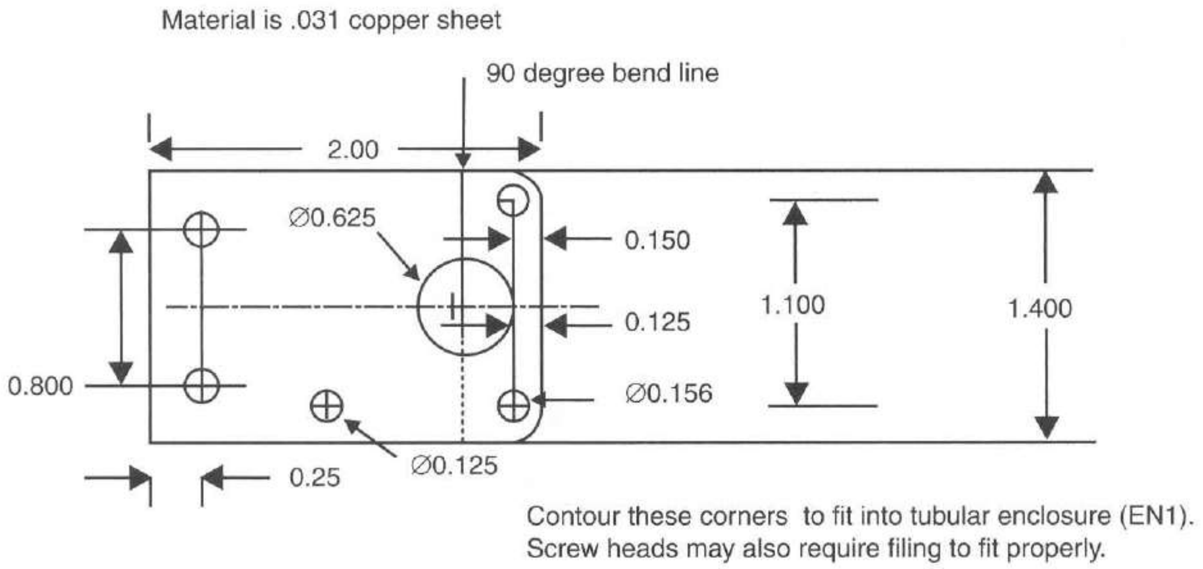
Material is .031 copper sheet
90 degree bend line
2.00
Ø0.625
0.150
0.125
1.100
1.400
0.800
Ø0.156
0.25
Ø0.125
Contour these corners to fit into tubular enclosure (EN1).
Screw heads may also require filing to fit properly.

Figure 1-4 *Fabrication of laser chassis*

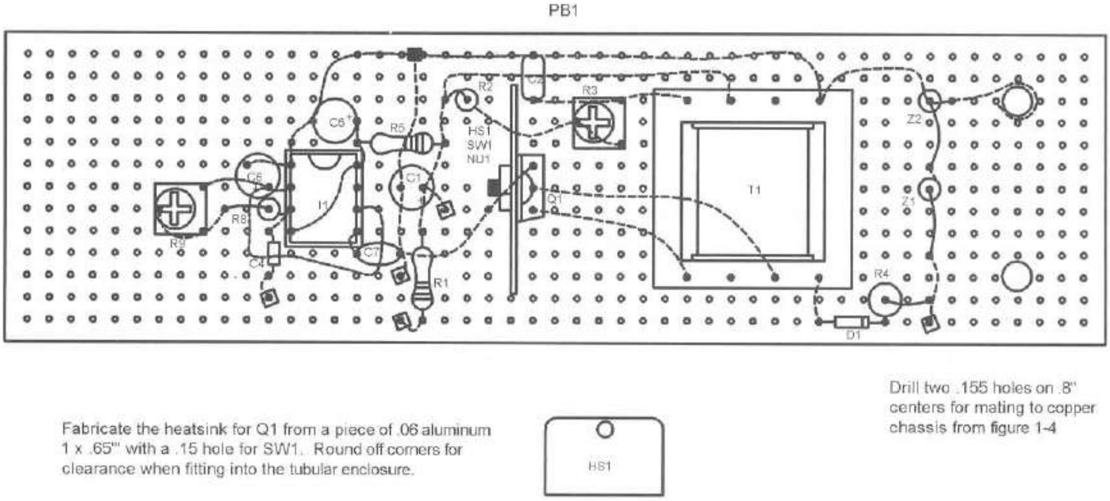
PB1
C5
R5
R2
C2
R3
HS1
SW1
NU1
T1
Z2
C6
I1
C1
Q1
R8
R9
Z1
C4
C7
R1
R4
D1
HS1

Drill two .155 holes on .8" centers for mating to copper chassis from figure 1-4

Fabricate the heatsink for Q1 from a piece of .06 aluminum 1 x .65" with a .15 hole for SW1. Round off corners for clearance when fitting into the tubular enclosure.

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

- Small dots are holes used for component insertion.

■ Black squares are solder junctions

▣ Squares with dots are holes that may require drilling for component mounting and points for external connecting wires indicating strain relief points. It is suggested to drill clearance holes for the leads and solder to points beneath board.

Figure 1-5 *Assembly board parts identification*

4. Connect the components using their actual leads wherever possible. Follow the layout and avoid bare-wire bridges.
5. Figure 1-6 shows the wiring of the copper chassis (COPCHA1). This type of wiring is preferred for high-frequency, sensitive, low-noise circuitry. It is used in this circuit due to the high-current pulses with fast rise times.

   Start by forming the leads and soldering to the copper plate. You will need a good hot soldering iron to do this correctly. The components themselves serve as anchor points and will be self-supporting when soldered to one another.
6. Figure 1-7 shows the interconnection of the copper chassis and the perforated assembly board via the two screws and nuts. Attach and connect the external leads and components as shown.

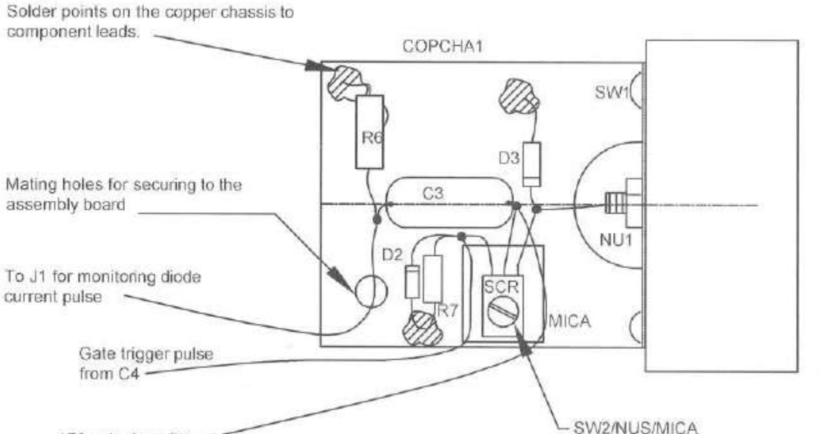
Solder points on the copper chassis to component leads.
COPCHA1
SW1
R6
D3
C3
Mating holes for securing to the assembly board
NU1
D2
To J1 for monitoring diode current pulse
SCR
R7
MICA
Gate trigger pulse from C4
170 volts from R4
SW2/NUS/MICA

Note that the SCR is insulated from the chassis by a mica or plastic washer and secured using a nylon screw.

Figure 1-6 *Assembly of laser chassis*

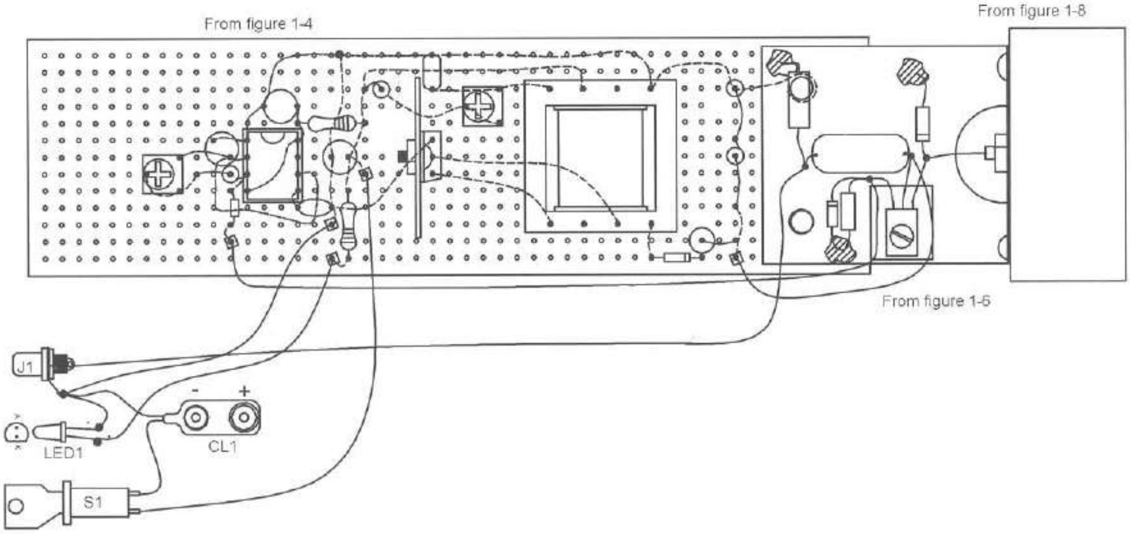

Figure 1-7 *Interconnecting the assembly board to the laser chassis*

7. Fabricate the LAHOLD1, as shown in Figure 1-8. Mount laser diode LD1 into the center hole and secure via a nut (NU1). Secure this assembly to the bent-down flange of the copper chassis. Use two ¼-inch screws as shown.
8. A collimator adapter is shown and is optional.

Assembly Testing

The following steps will require basic testing equipment.

1. It is assumed at this point that the assembly is complete, as shown in Figure 1-7. Verify the circuit and preset the two trimpots R3 and R9 to midrange.
2. Obtain a 12-volt bench power supply and connect the 12 volts to the battery clip CL1. Turn on the key switch and note the LED coming on. A current meter on the bench supply should read approximately 300 milliamps.
3. Connect a scope to TP1, as shown in Figure 1-2, and adjust R3 for the waveshape as shown.
4. Connect the scope to TP2 and note a spiky waveform that verifies the trigger pulse to the SCR.
5. Connect the scope to TP3 and note the waveform, which is the voltage across the SCR. Note the flat spot indicating that capacitor C3 has reached full charge before the SCR is triggered. The top of this waveshape is affected by the pulse repetition rate and should support 500 repetitions per second when set by R9.
6. Connect the scope to TP4 and set it for a negative-going fast pulse, indicating the current waveform through the LD1 laser diode. If you have followed the instructions and your assembly of the copper chassis is secure with short leads, you should see a picture-perfect waveform as shown.

   Note that the current pulse is 40 amps and corresponds to 1 amp for every volt read at TP4. If you find the pulse is less, you can increase the voltage ratings on one of the zener diodes. The laser diode specified requires 40 amps for full output. Any value over 40 amps will greatly shorten the useful life of this part.
7. Verify the laser output by using a night vision scope or a video camera. You should see an indication of the output. Screw in LEN1 so it is flush with the holder and adjust for the smallest footprint at over 50 feet. Our lab model adjusted to about two turns below being flush with the holder for the best results. The output should be a sharp, narrow bar shape with the diode specified.

Material can be preferably aluminum or fabricated from a PVC rod.

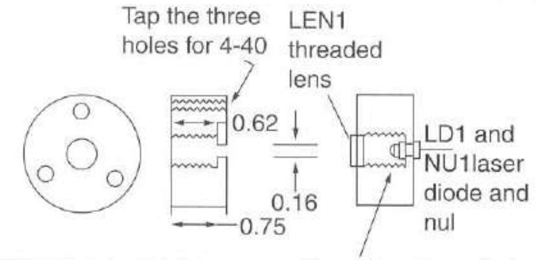

This section can be mated to section B when using a collimator with a C mount fitting.

Thread to .5" depth to mate with lens as supplied.

Optonal "C" mount collimator adapter fabricated from aluminum

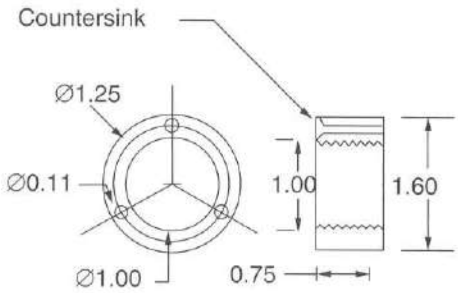

Section B adapter for fitting to the lens holder when using the optional collimator

Three mounting holes of .11 on a 1.25" diameter reference 120 degrees are mated to those in above

Thread for C lens at 1" diameter 32 tpi to a minimum of .375"

Figure 1-8 *Laser diode and lens holder*

Final Assembly

Create the EN1 enclosure from a 12 × 1.5-inch inner diameter plastic or aluminum tube (see Figure 1-9). Enclose the assembly and label it as shown in Figure 1-10. You may want to construct a mounting block or another assembly to mechanically connect the laser to a good, sturdy video tripod for future use.

Special Note

This laser is rated at 10 to 20 watts and is not to be confused with a continuous-output device rated at this power level. The advantage of this system is that the pulses are optically equivalent to 10 to 20 watts, and a suitable detection system will see them at this power level.

The equivalent energy output is related to the repetition rate times the peak power times the pulse duration and is approximately equal to joules. A joule is a watt second and is energy. If that energy is released in 1 microsecond as a pulse, the power of that pulse is now 1 megawatt.

A very suitable detector for this pulsed laser is our laser light detector described in Chapter 2, and a useful application for it as a laser property security fence is described in Chapter 5. Note that a laser pointer is 5 milliwatts, meaning the pulsed laser is 5,000 times more detectable.

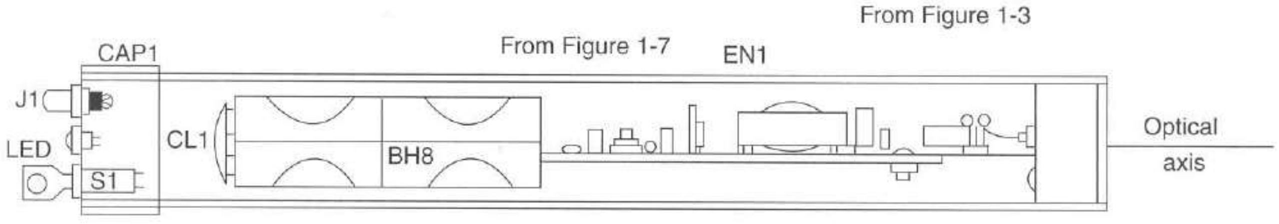

Figure 1-9 *X-ray view of enclosure*

CAUTION: Viewing of laser beam of direct reflections require the use of protective eye wear available through Information Unlimited at www.amazing1.com

Instruction

1. Turn key switch counterclockwise to verify unit is off. Note key is easily removed in the "off" position
2. Remove CAP1 and insert 8 AA batteries into holder.
3. Turn key clockwise and note LED lighting. If not, check the battery holder. Note key cannot be removed in the on position. Laser energy is infrared and can not be seen with the eye. You must use infrared detection devices such as our laser pulse detector or suitable night vision devices. Video cameras will often detect the laser.
4. Adjust collimator for desired effect if you are using.

Note that collimator will expand beam at near field but greatly reduce it at far field.

Compliance tests
1. Verify correct labels as shown
2. Key switch — nonremoval in off position
3. Beam indicator LED indicates beam emission

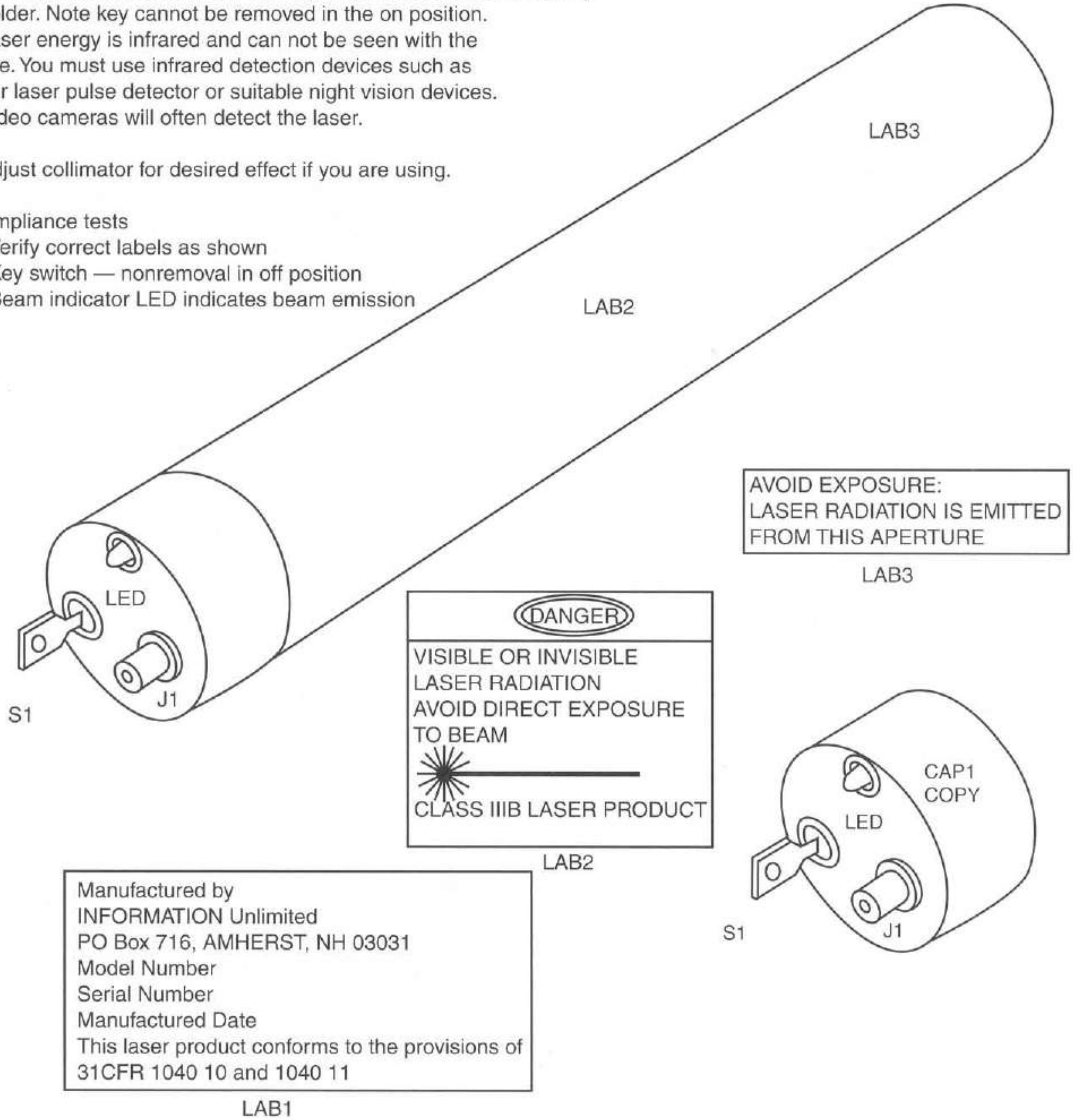

AVOID EXPOSURE:
LASER RADIATION IS EMITTED
FROM THIS APERTURE

VISIBLE OR INVISIBLE
LASER RADIATION
AVOID DIRECT EXPOSURE
TO BEAM
CLASS IIIB LASER PRODUCT

Manufactured by
INFORMATION Unlimited
PO Box 716, AMHERST, NH 03031
Model Number
Serial Number
Manufactured Date
This laser product conforms to the provisions of
31CFR 1040 10 and 1040 11

Figure 1-10 *Final assembly view and labels*

Table 1-1 Battery-Operated Infrared Laser Parts List

Ref. #	*Description*	*DB Part #*
R1	470 1/4-watt resistor (yel-pur-br)	
R2	220 1/4-watt resistor (red-red-br)	
R3	2K horizontal trimmer resistor	
R4	18K, 3-watt metal oxide resistor	
R5	10-ohm, 3/4-watt resistor (br-blk-blk)	
R6	1-ohm, 1/2-watt carbon resistor (br-blk-gold)	
R7	39 1/4-watt resistors (or-wh-blk)	
R8	1K 1/4-watt resistor (br-blk-red)	
R9	10K horizontal trimmer resistor	
C1,5	Two 10-microfarad, 50-volt vertical electrolytic capacitors	
C2	.047-microfarad, 50-volt plastic capacitor 473	
C3	.033-microfarad, 250-volt polypropylene capacitor	
C4	.1-microfarad, 50-volt plastic capacitor	
C6	1-microfarad, 50-volt vertical electrolytic capacitor	
C7	.01-microfarad/50-volt plastic capacitor (103)	
D1,3	Two 1N4937 1-kilovolt, 1-amp diode	
D2	IN914 silicon diode	
LED1	Bright green *light-emitting diode* (LED)	
LD1	40-amp, 10- to 20-watt, 904-nanometer TO18 pack laser diode	DB# LD650
SCR1	C106M 600-volt, 4-amp *silicon-controlled rectifier* (SCR)	DB# C106M
Z1	100/76-volt, 5-watt zener diode 1N5378	
Z2	150/108-volt, 5-watt zener diode 1N5383	
Q1	MJE3055T NPN power transistor in TO220 case	
I1	555 *dual inline package* (DIP) timer	
T1	400-volt square-wave-switching transformer	DB#TYPE1PC
PB1	1.4 × 5-inch, .1 × .1 grid perforated board	
S1	Key switch with non-removable key in on position	DB# KEYSWSM
J1	RCA phono jack	
COPCHAS1	Copper chassis, created in Figure 1-4	DB# COPCHAS1
LAHOLD	Laser and lens holder, created in Figure 1-8	DB# LAHOLD1
LENS13	Threaded lens	DB# LENS13
EN1	1 1/2-inch inner diameter × 1/32-inch wall × 12-inch plastic or aluminum housing tube	
CAP1,2	1 5/8-inch plastic caps, reworked as shown in Figure 1-9	
BH8	Eight AA cell holder	
SW1	Five 6-32 × 1/4-inch screws	
NU1	Four 6-32 hex nuts	
SW2	4-40 × 3/8-inch nylon screw	
MICA	Mica insulating washer	
NU2	4-40 hex nut	
HSINK	Heatsink tab for Q1	

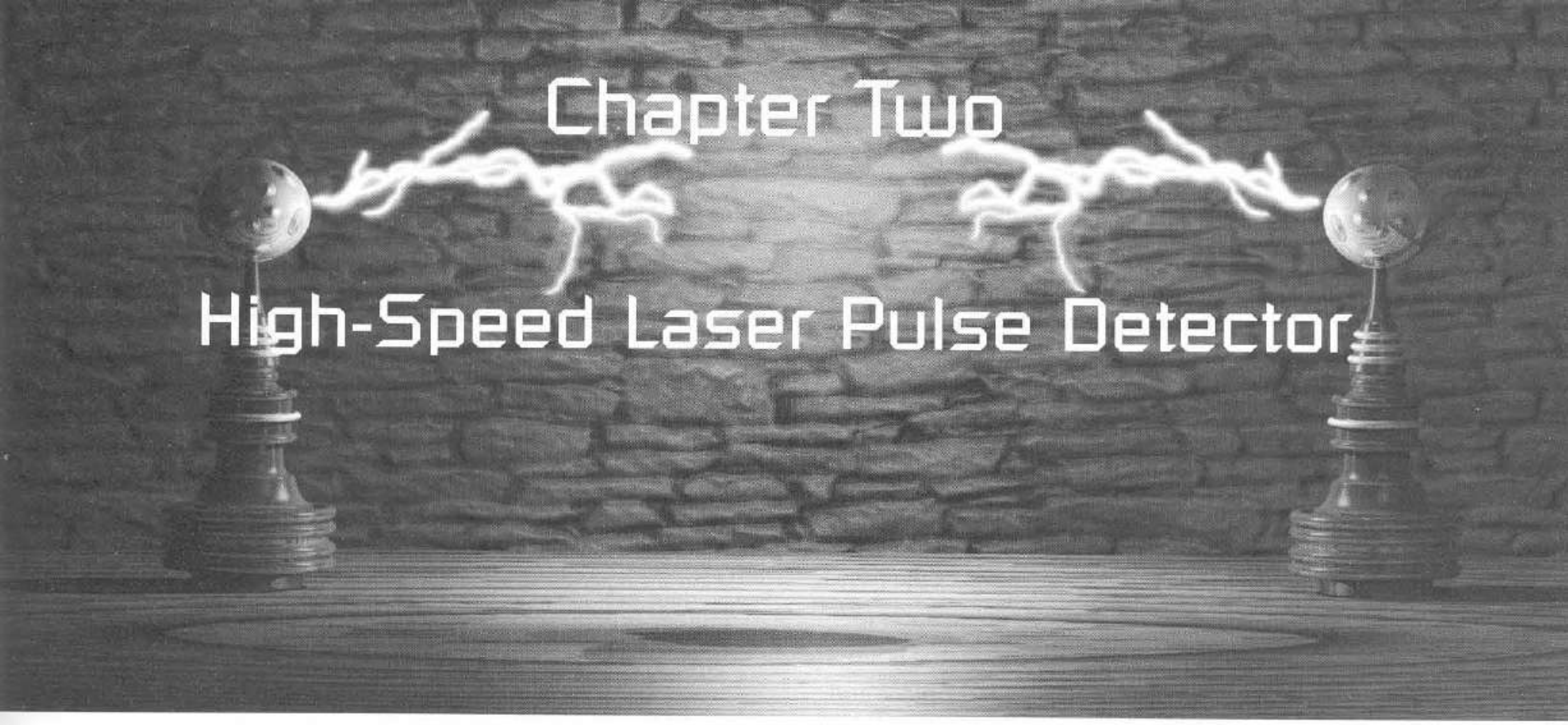

Chapter Two

High-Speed Laser Pulse Detector

This project shows how to build a specialized device capable of detecting fast, high-speed optical pulses. A suitable laser generating these pulses is described in Chapters 1 and 4 and can have a range, limited only by the curvature of the earth. Its uniqueness is its ability to resolve these pulses with rise times in the nanosecond range. This capability makes it possible to transmit large amounts of information within a small period of time.

The heart of this device is the actual front-end detector, in this case a PIN diode device. PIN comes from *positive negative* (PN) junction with intrinsic layer (I). The characteristics of this detector are low capacitance, high resistance, and low leakage when operating in the reverse bias mode. It may also be used as a photo cell when operated in the forward mode.

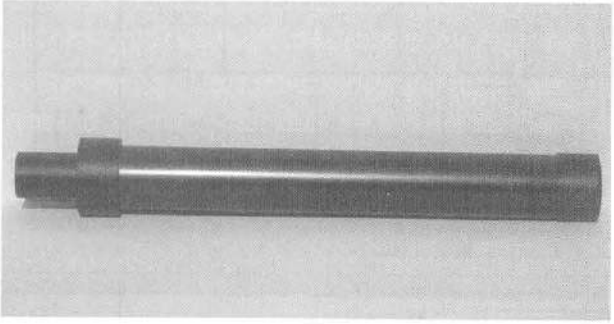

Figure 2-1 *The high-speed laser pulse detector*

The first part of the circuit is built on a piece of thin copper and uses discrete components. The builder may choose to build on a fully copper-clad circuit board if available, and a facsimile output can be set up via a phono jack or a BNC connector. A signal-conditioning circuit is shown built on a normal perforated circuit board and is designed to actuate a relay in the absence of the laser pulses. This mode of operation allows its use as a long-range laser perimeter or invisible fence intrusion detector, as described in Chapter 5.

Optics are shown for those wanting to operate at long ranges. Machined fittings are shown to mate with the 1.5-inch inner-diameter plastic enclosure used for the housing.

This is an intermediate- to advanced-level project requiring certain electronic skills, and one should expect to spend $35 to $70. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 2-1.

Circuit Description

The circuit in Figure 2-2 utilizes a pin photodiode (LPIN) reverse biased by resistor R3. The reverse-biased photodiode produces a voltage across bias resistor R2. The response of the PIN photo detector is determined by the resistance of R2 and is relatively low due to the nanosecond rise time of the laser pulses detected. Capacitor C3 couples the positive-going detected laser pulses to the gate of a preamplifying *field effect transistor* (FET) Q1 where the negative-going pulse is amplified and fed into a second FET (Q2) for further amplification. NPN transistor Q3 again amplifies the positive-going pulse for Q2 to a negative pulse. This amplified signal is a reasonably good facsimile of the actual laser pulse and can be monitored or further processed via phono jack J1.

The negative-going pulses are also fed to a Schmidt discriminator that is in an untriggered state, providing that transistor Q4 continues to see these pulses. When the pulses are interrupted by an object in the laser path, Q4 turns on due to a positive bias set by trimpot resistor R15 and turns transistor Q5 off, regenerating the trigger state of the Schmidt discriminator. R15 allows the adjustment of a trigger threshold if required. The triggering time of the circuit also helps prevent false triggering such as flying birds, windblown debris, and insects.

The output of the Schmidt is DC coupled into the base of the emitter-follower transistor Q6 where the signal is inverted by transistor Q7. The negative-going level at the collector of Q7 now triggers timer I1 on for a preset time that controls an alarm or another function resulting from a qualified laser beam interruption. A relay (RE1) is shown in the circuit and can be used as normal open or normal closed contacts for control of external circuit functions. This is shown in the phasor pain field generator in Chapter 35.

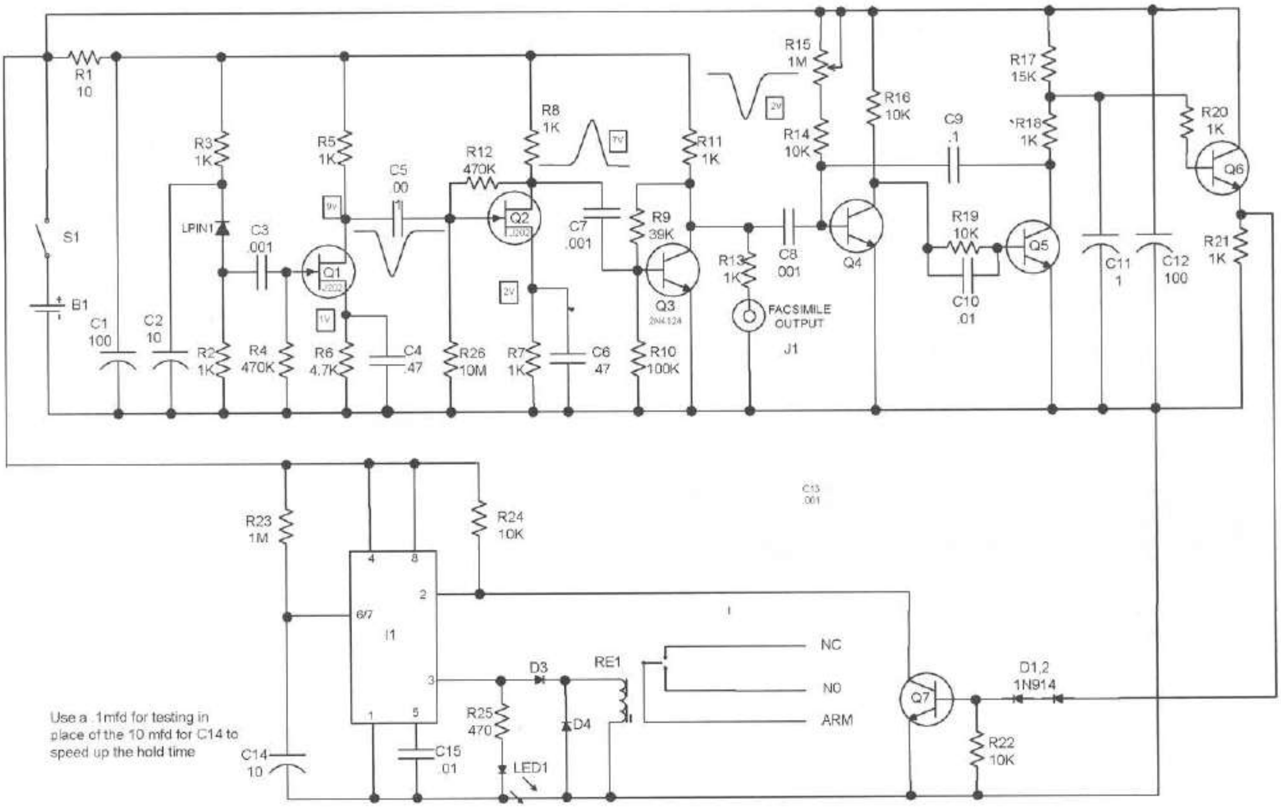
R1
10
S1
B1
C1
100
C2
10
R3
1K
LPIN1
R2
1K
C3
.001
R4
470K
Q1
J202
R5
1K
9V
1V
R6
4.7K
C4
.47
C5
.00
1
R12
470K
R26
10M
Q2
J202
R8
1K
7V
2V
R7
1K
C6
.47
C7
.001
R9
39K
Q3
2N4124
R10
100K
R11
1K
R15
1M
2V
R14
10K
R13
1K
C8
.001
FACSIMILE OUTPUT
J1
Q4
R16
10K
C9
.1
R19
10K
C10
.01
Q5
R17
15K
R18
1K
C11
1
C12
100
R20
1K
Q6
R21
1K
C13
.001
R23
1M
R24
10K
4
8
2
6/7
I1
3
1
5
C14
10
C15
.01
R25
470
D3
D4
LED1
RE1
NC
NO
ARM
Q7
D1,2
1N914
R22
10K

Use a .1mfd for testing in place of the 10 mfd for C14 to speed up the hold time

Figure 2-2 *Circuit schematic*

System Assembly

1. Assemble the board, as shown in Figure 2-3. Lay out and identify all the parts and pieces. Separate the components that go to the assembly board and create the PB1 perfboard from a 1½ × 5½-inch piece of .1 × .1 grid. Note the two holes for attachment to the copper chassis section.
2. Assemble the components as shown and observe polarity of the capacitors and semiconductors.
3. Connect the components using their actual leads wherever possible. Follow the layout and avoid bare-wire bridges. Note that Figure 2-3 shows pieces of #20 bus lead used as VC+ and COMMON rails. This approach is easier for wiring as it gives a direct route for the component leads.
4. Create the copper chassis as shown in Figure 2-4. This part is listed as a fabricated piece in Table 2-1.
5. Figure 2-5 shows the wiring of the COPCHA1 copper chassis. This type of wiring is preferred for high-frequency and sensitive, low-noise circuitry. It is used in this circuit due to the fast-rise-time, low-level-voltage pulses.

   Start by forming the leads and soldering to the copper plate. You will need a good, hot soldering iron to do this correctly. The components themselves serve as anchor points and will be self-supporting when soldered to one another.

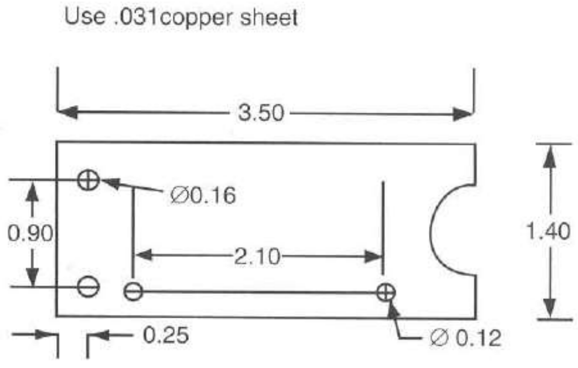

Figure 2-4 *Fabrication of the laser chassis*

6. Figure 2-6 shows the connection of the copper chassis and the perforated assembly board via the two screws and nuts. Attach and connect the external leads and components as shown.

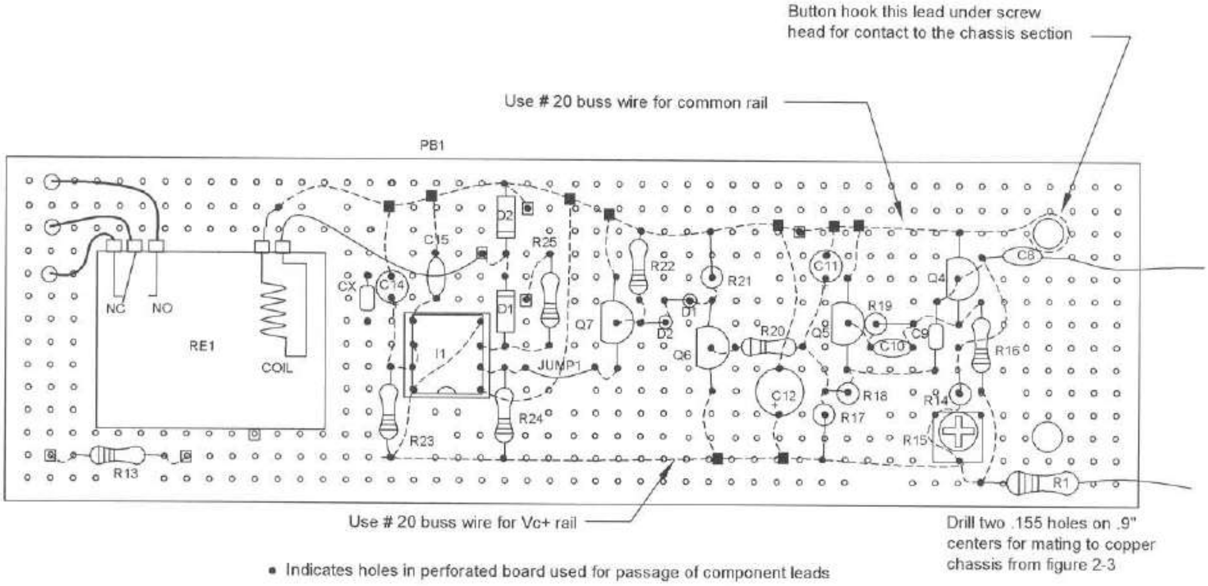

Button hook this lead under screw head for contact to the chassis section

Use # 20 buss wire for common rail

Use # 20 buss wire for Vc+ rail

Drill two .155 holes on .9" centers for mating to copper chassis from figure 2-3

• Indicates holes in perforated board used for passage of component leads

▣ Indicates holes in perforated board used for external leads connection points

■ Indicates component soldering points on (+) plus and (-) negative rails

Use a piece of #20 buss wire for rail leads indicated with heavy dashed lines

Figure 2-3 *Assembly board parts identification*

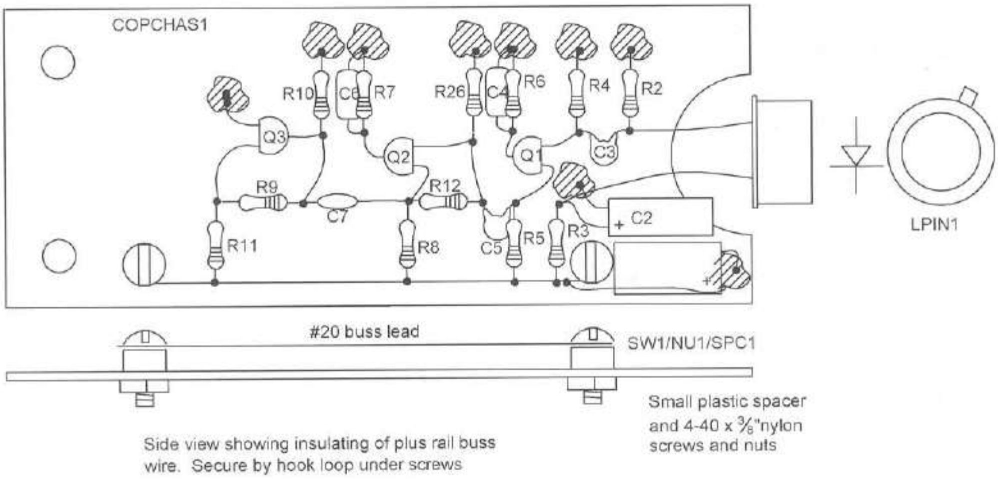
COPCHAS1
R10
C6
R7
R26
C4
R6
R4
R2
Q3
Q2
Q1
C3
R9
C7
R12
+ C2
LPIN1
R11
R8
C5
R5
R3
+
#20 buss lead
SW1/NU1/SPC1

Small plastic spacer and 4-40 x ⅜"nylon screws and nuts

Side view showing insulating of plus rail buss wire. Secure by hook loop under screws

Figure 2-5 *Assembly of the laser chassis*

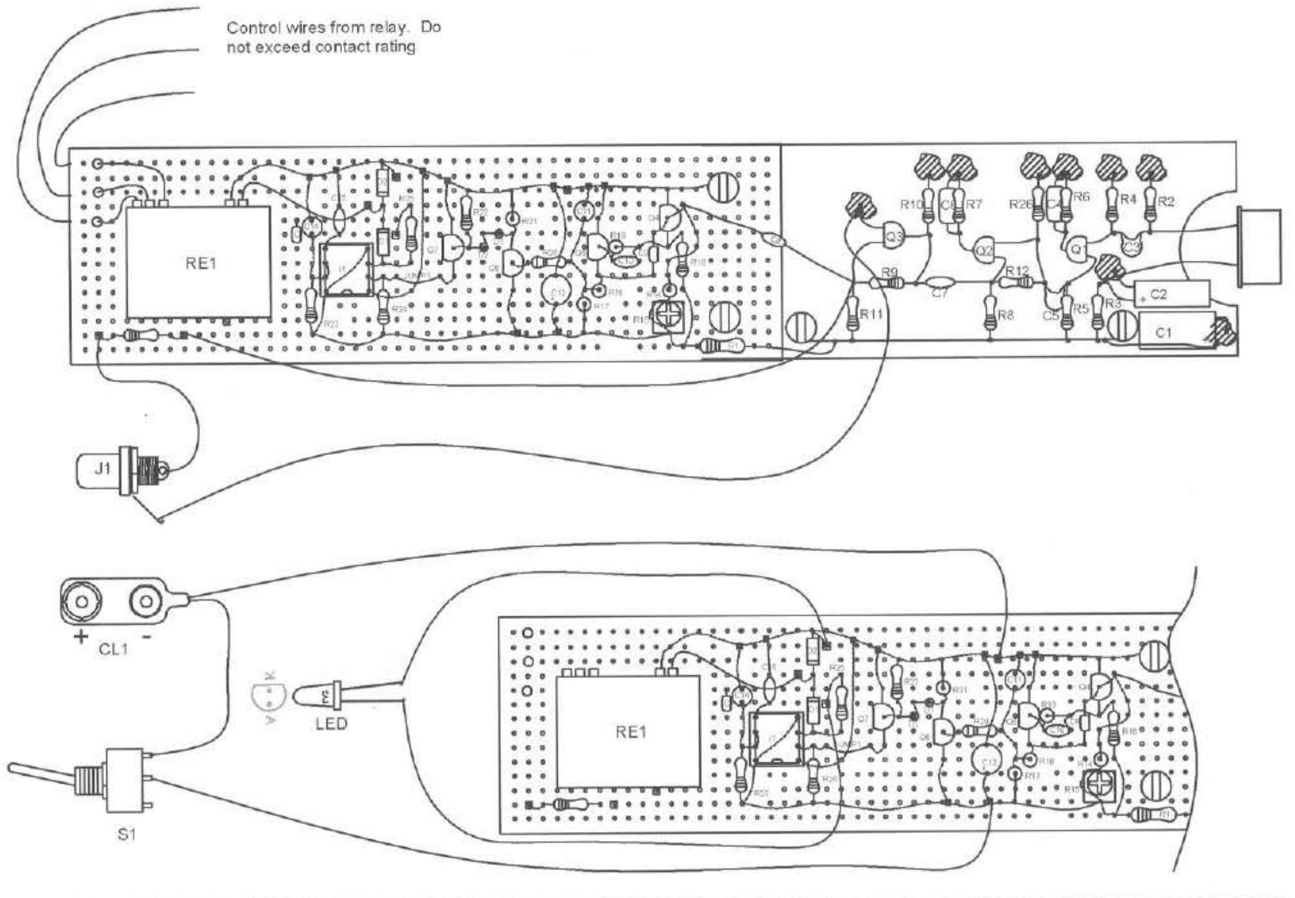
RE1
R10
C6
R7
R26
C4
R6
R4
R2
Q3
Q2
Q1
C3
R9
C7
R12
C2
R11
R8
C5
R5
R3
C1
J1
+
CL1
-
LED
S1
RE1

Control wires from relay. Do not exceed contact rating

Figure 2-6 *Connection of board to chassis*

Assembly Testing

Testing of the laser receiver involves the use of a compatible laser transmitter to provide the necessary high-speed pulses. The laser described in Chapter 1 is recommended.

1. It is assumed at this point that the assembly is complete, as shown in Figure 2-6. Verify the circuit and preset the R15 trimpot to midrange. This test is performed *without* the circuitry installed and without a lens. You should temporarily replace C14 with a .1-microfarad value.
2. Obtain a 12-volt bench power supply and connect the 12 volts to the battery clip CL1. Turn on the S1 switch and note the LED coming on and flashing erratically. A current meter on the bench supply should read approximately 10 milliamps.
3. Measure the DC voltage levels, as indicated in Figure 2-2, noted in the squares. These values should be within 10 to 15 percent of those indicated.

   At this point, you will need an optical laser pulse transmitter, such as that described in Chapter 1. Any suitable laser or optical transmitter will work with a fast pulse rise of 1 usec or so. The detector, as shown, uses a silicon PIN diode with a spectral sensitivity peak at 960 nanometers. This is very close to the spectral output of the lasers described in Chapters 1 and 4.
4. Position the laser pulse generator source so that it is pointing in the direction of the laser receiver being tested. You will need to connect a scope at the drain of Q1 and then attempt to position the receiver so it is receiving the pulses and is secure in this position. This may be tricky and require patience. When the laser is perfectly aligned with the receiver, the signal will block the Q1 and produce an overloaded waveshape. Carefully adjust the receiver off axis until the scope indicates the negative-going facsimile pulse as shown. Continue through the circuit and observe the waveshapes as shown, again adjusting the receiver off axis to obtain the required waveshape.

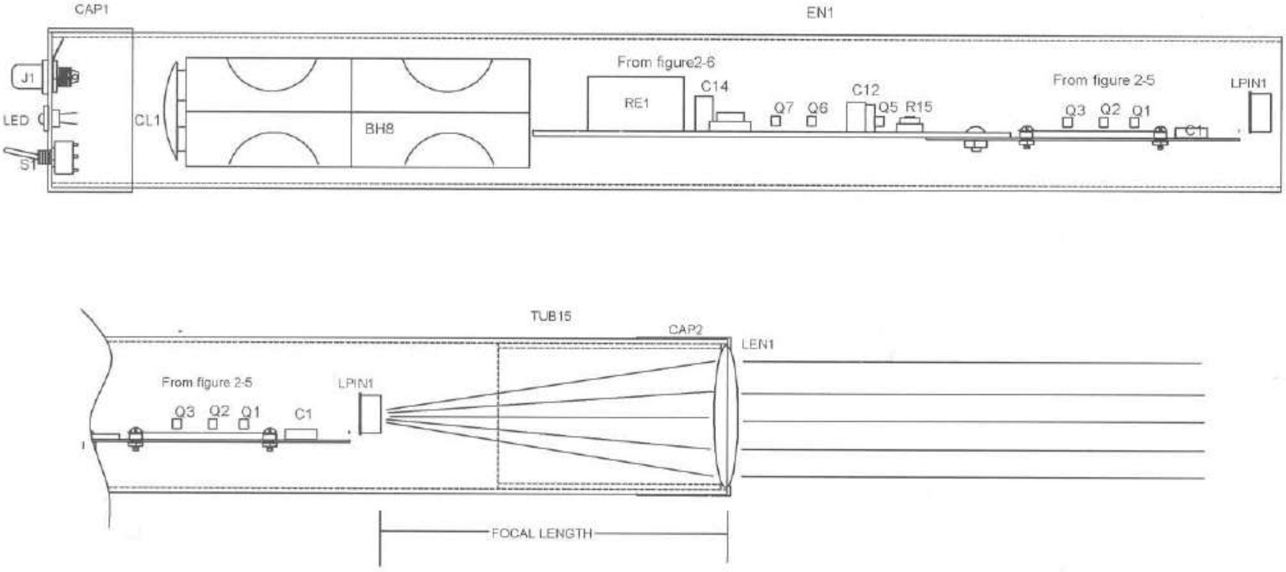

The lens system shown uses a 38 mm diameter 100 mm focal length. A 2" x 1.5" diameter tube (TUB15) is sleeved into the main enclosure EN1 and is positioned to abut to the lens on it's shoulders. A plastic cap (CAP2) retains the lens against TUB15. Acenter hole of over 1" is removed for the light aperture.

Figure 2-7 *X-ray view of enclosure*

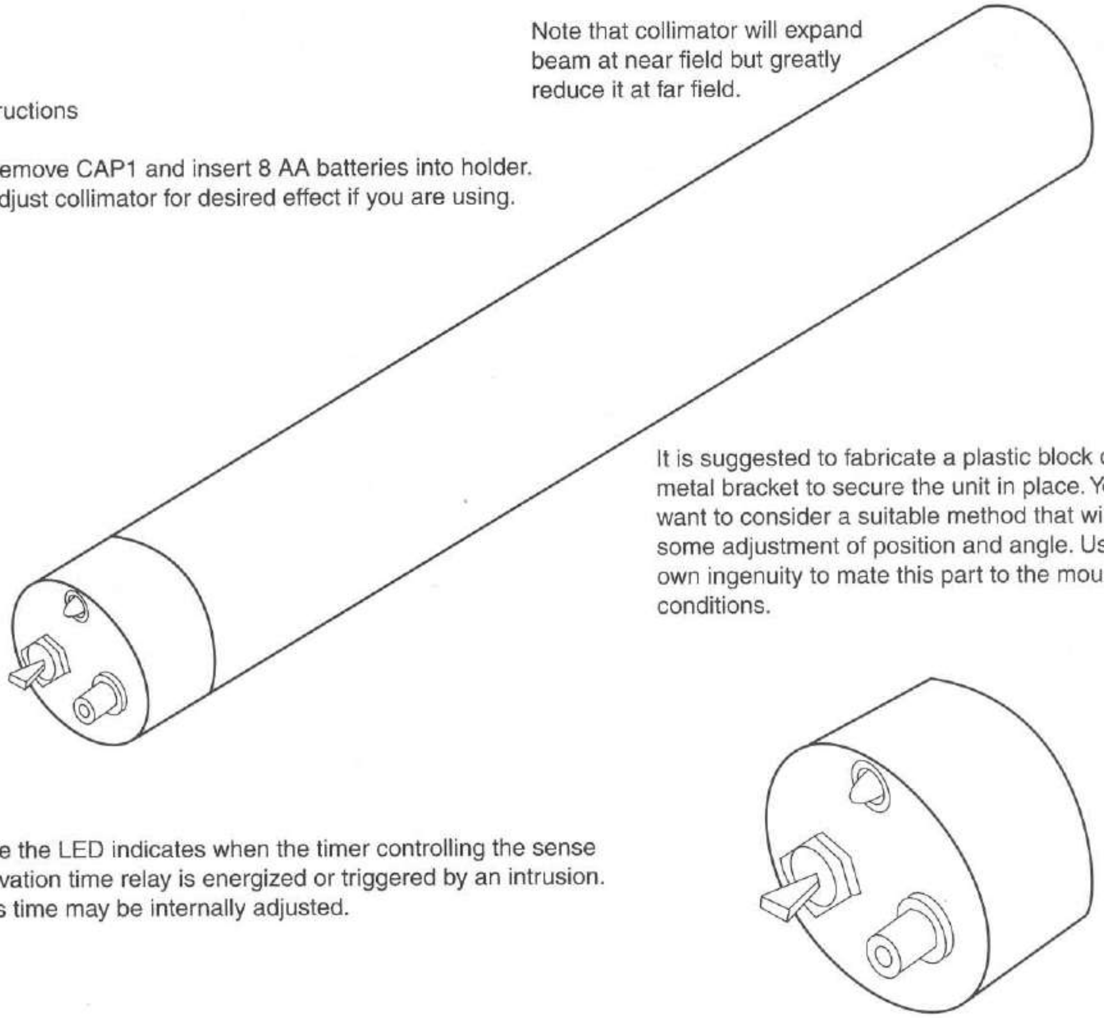

Note that collimator will expand beam at near field but greatly reduce it at far field.

Instructions

1. Remove CAP1 and insert 8 AA batteries into holder.
2. Adjust collimator for desired effect if you are using.

It is suggested to fabricate a plastic block or sturdy metal bracket to secure the unit in place. You may want to consider a suitable method that will allow some adjustment of position and angle. Use your own ingenuity to mate this part to the mounting conditions.

Note the LED indicates when the timer controlling the sense activation time relay is energized or triggered by an intrusion. This time may be internally adjusted.

Figure 2-8 *Final assembly view*

Note that the LED will be off if the circuit is working and will come on if you block the light with a solid object.

5. Verify that the discriminator works and you can set the sensitivity trip level via R15. Verify the timer trips when the light beam is broken, activating the relay for a preset time as determined by the value of C14. This feature now allows external circuitry to be activated when the laser beam is broken, powering alarms or deterrents. The relay as shown provides a normally closed and open circuit that can be selected by the user. Do not switch loads that exceed the relay contact rating.

Final Assembly

Fabricate the EN1 enclosure from a 15 × 1.5-inch ID plastic or aluminum tube. Enclose the assembly as shown in Figure 2-7. You may want to create a mounting block or another assembly to mechanically connect the laser to a good, sturdy video tripod for future use

The system as shown is enclosed in a long tube that includes the batteries. You may separate the batteries and relay control circuits from the actual detection section consisting of the copper chassis and optics. This approach will be a two-part system that is connected via a cable or cord.

Special Note

This laser pulse detector receiver, when used with a pulsed laser like that described in Chapter 1, can have a control range exceeding that allowed by the curvature of the earth when used with suitable optics. Alignment is about as critical as sighting a high-powered rifle and requires extreme mechanical stability once set. The alignment of a long-range system requires patience, perseverance, and a little black magic.

Chapter 5 describes how to use both the laser pulser from Chapter 1 and the above assembly in a multiple-reflection property protection device with an ultrasonic shock deterrent.

Special Note on Photo Detectors

Two photo detectors are referenced in Table 2-1. LPIN is the low-cost piece that will usually provide the necessary operating parameters for most hobbyist laser projects. LPINX is more expensive and has a much lower dark current rating. It is used for lower noise and more sensitive circuitry.

Table 2-1 Laser Pulse Detector Parts List

Ref. #	*Description*	*DB Part #*
R1	10 ohms, 1/4 watt (br-blk-blk)	
R2, 3, 5, 7, 8, 11, 13, 18, 20, 21	Ten 1K, 1/4-watt resistors (br-blk-red)	
R4, 12	Two 470K, 1/4-watt resistors (yel-pur-yel)	
R6	4.7K, 1/4-watt resistor (yel-blk-red)	
R9	39K, 1/4-watt resistor (or-wh-or)	
R10	100K, 1/4-watt resistor (br-blk-yel)	
R14, 16, 19, 22, 24	Five 10K, 1/4-watt resistors (br-blk-or)	
R15	1M horizontal trimmer resistor	
R17	15K, 1/4-watt resistor (br-blk-or)	
R23	1M, 1/4-watt resistor (br-blk-grn)	
R25	470-ohm, 1/4-watt resistor (yel-pur-br)	
R25	10M, 1/4-watt resistor (br-blk-blue)	
C1, 12	Two 100-microfarad, 50-volt vertical electrolytic capacitors	
C2, 14	Two 10-microfarad, 50-volt vertical electrolytic capacitors	
C3, 5, 7, 8	Four .001-microfarad, 100-volt disc capacitors	
C4, 6	Two .47-microfarad, 50-volt plastic capacitors	
C9, 13	Two .1-microfarad, 50-volt plastic capacitors	
C10, 15	Two .01-microfarad, 50-volt plastic capacitors (103)	
C11	1-microfarad, 50-volt vertical electrolytic capacitor	
Q1, 2	Two J202 N-channel FET transistors	
Q3	2N4124 *negative positive negative* (NPN) high-frequency transistor	
Q4,5,6,7	Four PN2222 NPN general-purpose transistors	
D1,2	Two IN914 silicon diodes	
D3,4	Two IN4007 1-kilovolt, 1-amp diodes	
I1	555 *dual inline package* (DIP) timer integrated circuit	
LED1	High-brightness *light-emitting diode* (LED)	
*LPIN1	Silicon PIN photodiode dark current less than .10 nano-amp (see text)	DB# PIN
*LPIN1X	Silicon photo diode dark current less than .02 nano-amp (see text)	DB# PINX
RE1	12-volt coil, 400-ohms, 120 VAC 15-amp contacts	DB# RE12115
COPCHAS1	1.4 × 3.5-inch .031 copper plate (see Figure 2-4)	
SW1/NU1	Four 4-40 × 3/8 nylon screws and metal hex nuts	
SPC1	Two 1/4-inch plastic spacers	
PB1	5½″ × 1½″ .1 × .1 grid perforated circuit board	
J1	Chassis mount RCA phono jack	
S1	Small *single pole, single throw* (SPST) toggle switch	
CL1	Battery clip	
BH8	Holder for 8 AA cell battery	
EN1	1 1/2-inch inner diameter × 1/32-inch wall plastic or aluminum tube	
CAP1, 2	Two 1 5/8″ plastic caps reworked as shown in Figure 2-8	
LEN1	38 mm × 100 mm focal length DCX glass lens	DB# 38100
RET1	Reworked 1 1/2-inch plastic cap for retaining lens	
TUB15	2-inch length of 1.5-inch outer diameter plastic or aluminum tube	

Chapter Three

Ultra-Bright Green Laser Project

This ultra-bright green laser project, as shown in Figure 3-1, is intended for the laser experimenter who desires to add a simple optical system that provides far-field focusing, beam collimation, or beam expanding.

Construction parts such as tubing, handles, brackets, and hardware are all readily available from most local hardware stores (see Table 3-1). Our suggested lab approach is to locate some telescopic tubing of metal or plastic, or a combination of both, and fabricate the laser and lens holders from a suitable sized plastic rod or equivalent. The center holes of the holders must be as true as possible or the optical system's integrity will be greatly degraded.

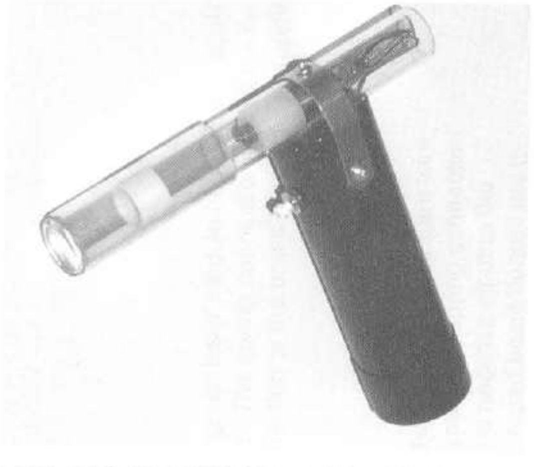

Figure 3-1 *Photo of green laser project*

The grip or handle is created from a 1-inch-thick piece of finished pine that has a section carefully removed for the battery holder and push-button switch. Passage holes for the power leads are drilled as shown. You may also use a suitable sized plastic box. A clear piece of Lexan (polycarbonate) sheet is formed with a flange at a 90-degree bend and is the surface for mounting the switch. This cover is screwed to the wooden handle and makes an attractive design if done with care and precision. The handle can be constructed with a contoured shape to match the curvature of the tube, and you may also add rings, fins, and other décor to make a space-age optical ray gun.

The operation of the laser allows usage without any optics, providing similar performance to a high-quality laser pointer. The addition of the lens system allows far-field focusing, close range expansion, and far-field collimation where the beam spot impact will be reduced by the collimating power of the system. As an example, the bare foot beam diameter (or operation without external optics) at 300 meters will be approximately a 50-centimeter diameter spot. With an ×10 collimator, the diameter can be reduced to 5 centimeters.

This is a class iiia beginner's laser project that uses a working laser module. Class iiia is a classification for lasers under 5 milliwatts of output power

Never point it at vehicles, aircraft, or directly at people. Expect to spend $40 to $50 for this very rewarding laser project.

Assembly Steps

1. Rework the laser module (LM1) as shown in Figure 3-2.
2. Wire the laser module with pushbutton switch (SW1) and battery clip (CL1) as shown in Figure 3-3.
3. Make final assembly as shown in Figure 3-4.
4. Connect batteries and activate SW1. Note a green impact point. Adjust lenses as shown for desired beam profile.

The green laser module has a built in push button switch that may be bypassed by adding a connection lead as shown. This added lead is the negative input. The spring connection is also the negative input but now requires activation of the built in switch. The positive input can be a mechanical connection to the brass housing by wrapping a bare lead around the slot. You may also carefully solder a lead as shown for this positive input point.

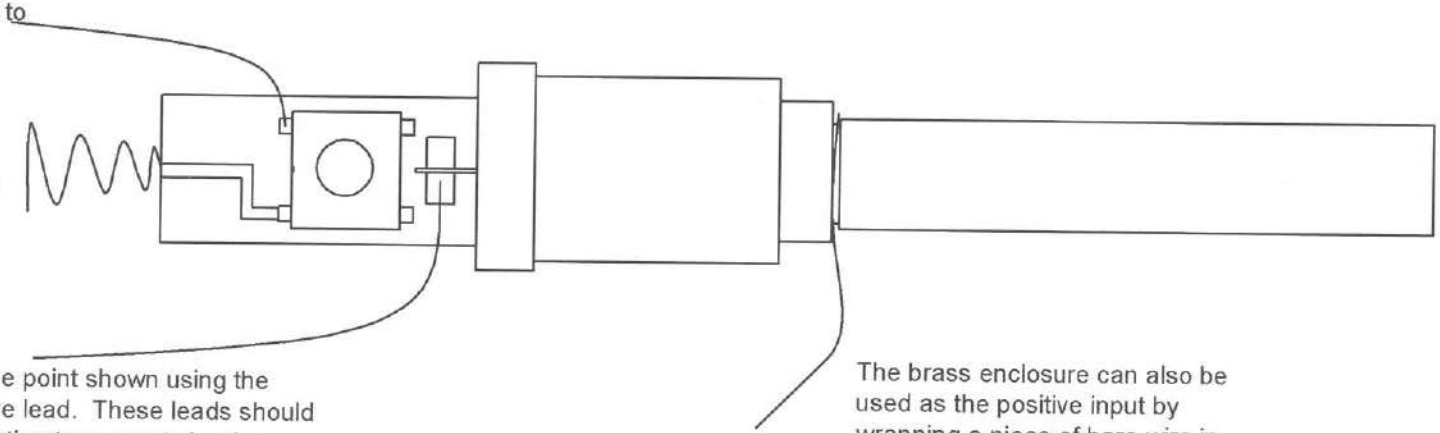

Solder a piece of #24 bare wire to bypass the spring connection for the negative input to the switch pad being careful not to damage the circuit board

The spring is the normal negative input point but will require activation of the built in push button switch

You may solder a lead to the point shown using the same method for the positive lead. These leads should be sleeved with some insulation to prevent shorting. You may also want to strain relieve these leads using silicon rubber etc.

The brass enclosure can also be used as the positive input by wrapping a piece of bare wire in the narrow slit as shown.

Figure 3-2 *Laser module rework*

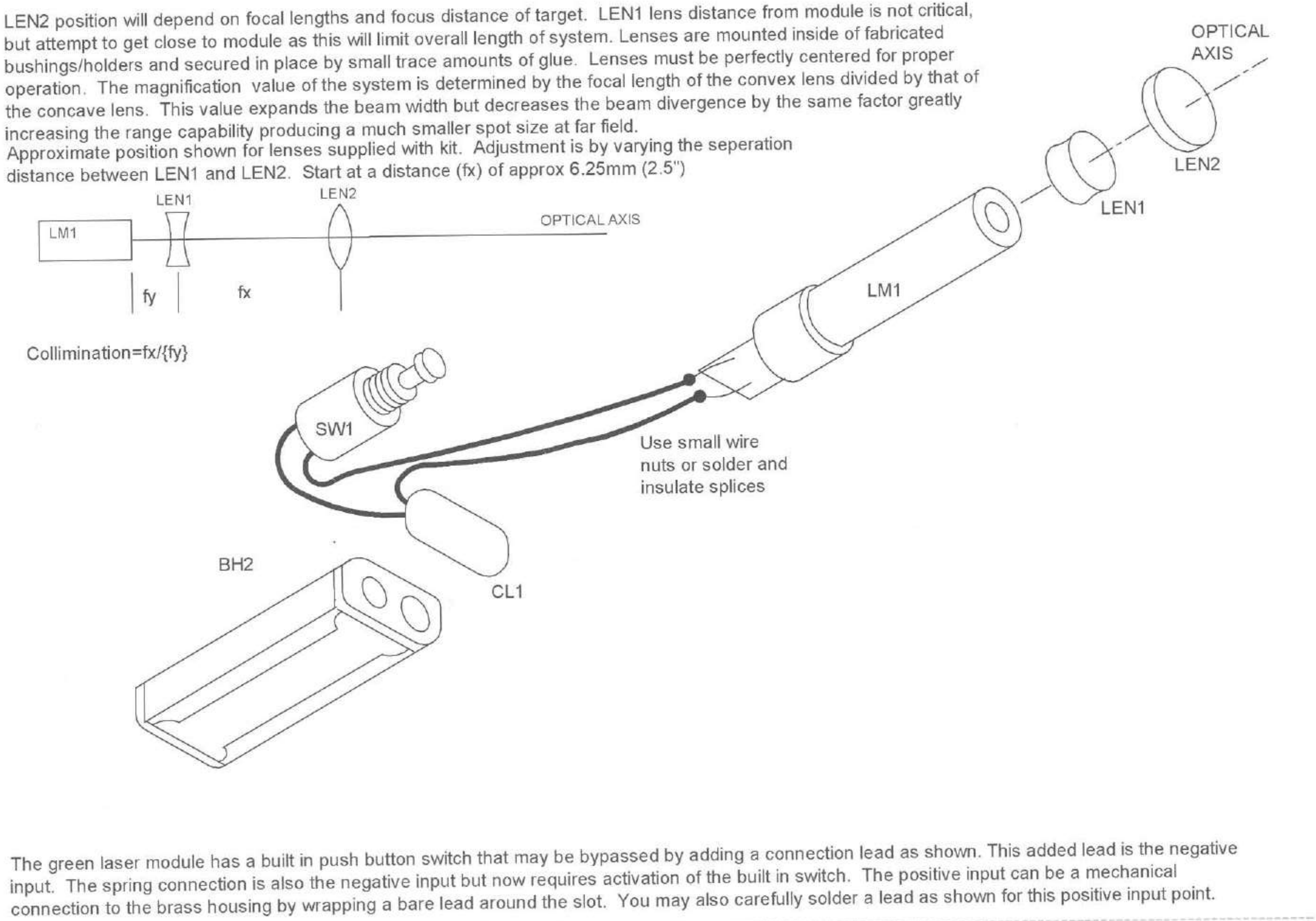

LEN2 position will depend on focal lengths and focus distance of target. LEN1 lens distance from module is not critical, but attempt to get close to module as this will limit overall length of system. Lenses are mounted inside of fabricated bushings/holders and secured in place by small trace amounts of glue. Lenses must be perfectly centered for proper operation. The magnification value of the system is determined by the focal length of the convex lens divided by that of the concave lens. This value expands the beam width but decreases the beam divergence by the same factor greatly increasing the range capability producing a much smaller spot size at far field.

Approximate position shown for lenses supplied with kit. Adjustment is by varying the seperation distance between LEN1 and LEN2. Start at a distance (fx) of approx 6.25mm (2.5")

Collimination=fx/{fy}

Use small wire nuts or solder and insulate splices

The green laser module has a built in push button switch that may be bypassed by adding a connection lead as shown. This added lead is the negative input. The spring connection is also the negative input but now requires activation of the built in switch. The positive input can be a mechanical connection to the brass housing by wrapping a bare lead around the slot. You may also carefully solder a lead as shown for this positive input point.

Figure 3-3 *Ultra bright green isometric assembly*

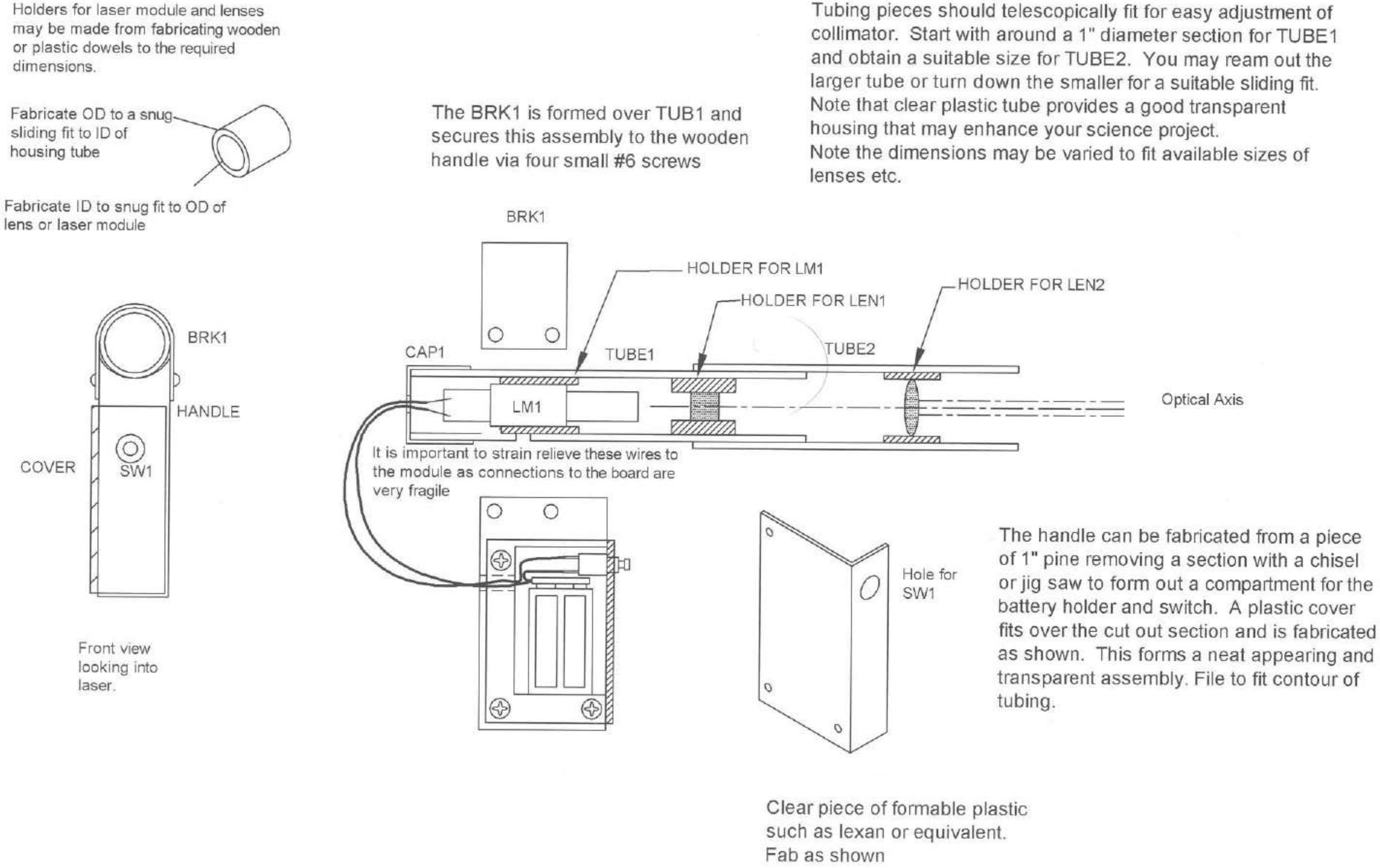

Holders for laser module and lenses may be made from fabricating wooden or plastic dowels to the required dimensions.

Fabricate OD to a snug sliding fit to ID of housing tube

Fabricate ID to snug fit to OD of lens or laser module

The BRK1 is formed over TUB1 and secures this assembly to the wooden handle via four small #6 screws

Tubing pieces should telescopically fit for easy adjustment of collimator. Start with around a 1" diameter section for TUBE1 and obtain a suitable size for TUBE2. You may ream out the larger tube or turn down the smaller for a suitable sliding fit. Note that clear plastic tube provides a good transparent housing that may enhance your science project.
Note the dimensions may be varied to fit available sizes of lenses etc.

It is important to strain relieve these wires to the module as connections to the board are very fragile

The handle can be fabricated from a piece of 1" pine removing a section with a chisel or jig saw to form out a compartment for the battery holder and switch. A plastic cover fits over the cut out section and is fabricated as shown. This forms a neat appearing and transparent assembly. File to fit contour of tubing.

Clear piece of formable plastic such as lexan or equivalent.
Fab as shown

Front view looking into laser.

Figure 3-4 *Ultra bright green optical laser gun suggested final design*

Table 3-1 Ultra-Bright Green Laser Parts List

Ref. #	*Description*	*DB Part #*
SW1	Pushbutton NO switch	
BH2AA	Two-cell AA battery holder	
CL1	Battery snap clip	
LM1	5-milliwatt green laser module	DB# LM-532-P5
LEN1	Concave lens .6 × -.75-inch focal length	DB# LE15
LENS	Convex lens .9 × 3.5-inch focal length	DB# LE2475
LAB1	Class 3a danger laser label	
LAB2	Certification label	
LAB3	Aperture label	
The mechanical parts you get yourself; they are not critical and those shown represent our finalized lab approach.		
TUBE1	1-inch OD thin-walled plastic tubing	
TUBE2	Tube for telescope over TUBE1	
HANDLE	2 × 4 × 1-inch piece of soft pine	
COVER	Piece of 1/16-inch clear plastic sheet	
BRK1	1/32 × 1/2-inch aluminum bracket	
LENS HOLDERS	Wood or plastic dowels	
SCREWS	Small wood	

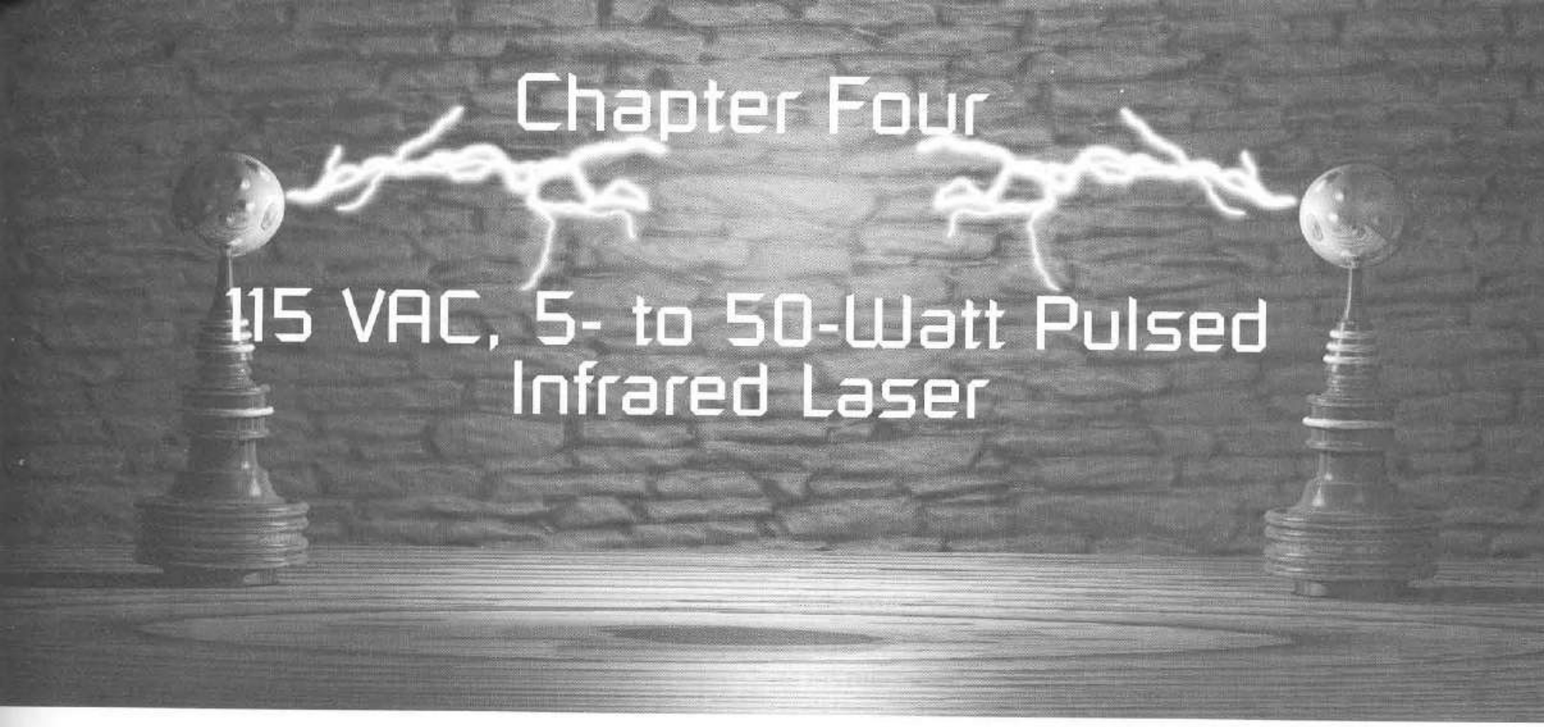

Chapter Four

115 VAC, 5- to 50-Watt Pulsed Infrared Laser

This project shows how to construct a laboratory-use, solid-state, pulsed infrared laser (see Figure 4-1). The system utilizes a gallium arsenide laser diode providing pulse powers of 10 to 100 watts depending on the diode used. The device operates from the 115 VAC line via an isolation transformer and is constructed in two parts. The laser is an excellent source of variable-frequency pulses adjustable from 10 to 2,500 repetitions per second of infrared energy.

The laser, when assembled as shown, is a class iiib device and requires the appropriate labeling and several included safety functions, as described in the assembly instructions. At no time should it be pointed at anyone without protective eyewear or at anything that could reflect these pulses. Also, never look into the unit when the power is on. It is intended to be used for ranging, simulated weapons practice, intrusion detection, communications and signaling, and a variety of related scientific, optical experiments and uses.

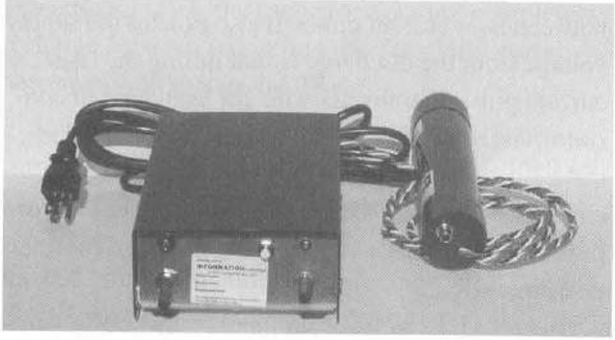

Figure 4-1 *IR laser*

This is an intermediate- to advanced-level project requiring electronic skills and basic electronic shop equipment. Expect to spend $100 to $150. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 4-1.

Theory of Operation

A laser diode is nothing more than a three-layer device consisting of a *pn* junction of *n*-type silicon, a *p*-type of gallium arsenide, and a third *p* layer of doped gallium arsenide with aluminum. The *n*-type material contains electrons that readily migrate across the *pn* junction and fill the holes of the *p*-type material. Conversely, holes in the *p*-type migrate to the *n*-type and join with electrons. This migration causes a potential hill or barrier consisting of negative charges in the *p*-type material and positive charges in the *n*-type material that eventually cease growing when a charge equilibrium exists. In order for current to flow in this device, it must be supplied at a voltage to overcome this potential barrier. This is the forward voltage

Table 4-1

INFO #	*Pulse width*	*Package*	*Diodes*	*Peak current*	*Emitting area*	*Peak power*	*Duty factor*	*Beam symmetry*	*Spectral width*
LD660	200 ns	TO18	1	40 amps	9 × 1	12 watts	.1%	15 × 20	3.5 nm
LD780	200 ns	TO5	2	40 amps	21 × 1	20 watts	.1%	15 × 20	3.5 nm
LD1630	200 ns	TO5	2	40 amps	21 × 1	30 watts	.1%	15 × 20	3.5 nm

drop across a common diode. If this voltage polarity is reversed, the potential barrier is simply increased, assuring no current flow. This is the reversed bias condition of a common diode.

A diode without an external voltage applied to it contains electrons that move and wander through the lattice structure at a low, lazy average velocity as a function of temperature. When an external current at a voltage exceeding the barrier potential is applied, these lazy electrons now increase their velocity to where some, by colliding, acquire a discrete amount of energy and become unstable, eventually emitting this acquired energy in the form of a photon after returning to a lower energy state. These photons of energy are random both in time and direction; hence, any radiation produced is incoherent, such as that of a *light-emitting diode* (LED).

The requirements for coherent radiation are that the discrete packets of radiation must be in the form of a lockstep phase and in a definite direction. This demands two essential requirements; first, sufficient electrons at the necessary excited energy levels and, second, an optical resonant cavity capable of trapping these energized electrons for stimulating more and giving them direction. The amount of energized electrons is determined by the forward diode current. A definite threshold condition exists where the device emits laser light rather than incoherent light, such as in an LED. This is why the device must be pulsed with a high current. The radiation from these energized electrons is reflected back and forth between the square-cut edges of the crystal that form the reflecting surfaces due to the index of refraction of the material and air.

The electrons are initially energized in the region of the *pn* junction. When these energized electrons drift into the *p*-type transparent region, they spontaneously liberate other photons that travel back and forth in the optical cavity interacting with other electrons commencing laser action. A portion of the radiation traveling back and forth between the reflecting surfaces of these mirrors escapes and constitutes the output of the device.

Circuit Theory of Operation

AC power, as shown in Figure 4-2, is obtained via polarized plug CO1 through fuse FH1. Proper grounding of the green cord lead is a necessity to prevent an unnecessary shock hazard along with usage of grounded test and measuring equipment. S1 is a key switch type where the key can be removed only in the off position. Transformer T1 provides a one-to-one ratio and isolation from the power line, while step-down transformer T2 provides the low voltages necessary for the control circuits. The power indicator lamp consists of neon lamp NE1 and associated current-limiting resistor R1.

Diodes D1 and D2, along with capacitors C1 and C2, comprise a voltage doubler. The voltage across C1 and C2 is 1.4 × 230 or approximately 340 volts.

A major section of this laser pulser is the charging circuit. This circuit charges the pulse discharge capacitor (C10) to the supply voltage during the time interval between laser current pulses. It also isolates the supply voltage from the discharge circuit during the laser current pulse, thereby allowing the switching *silicon-controlled rectifier* (SCR) to recover to the blocking state. Because the response times of the charging circuit are relatively long, lead lengths are not important, and the circuit can be remotely located from the discharge circuit.

The simplest charging circuit is a resistor-capacitor combination. In this simple case, the resistor must limit the current to a value less than the SCR holding

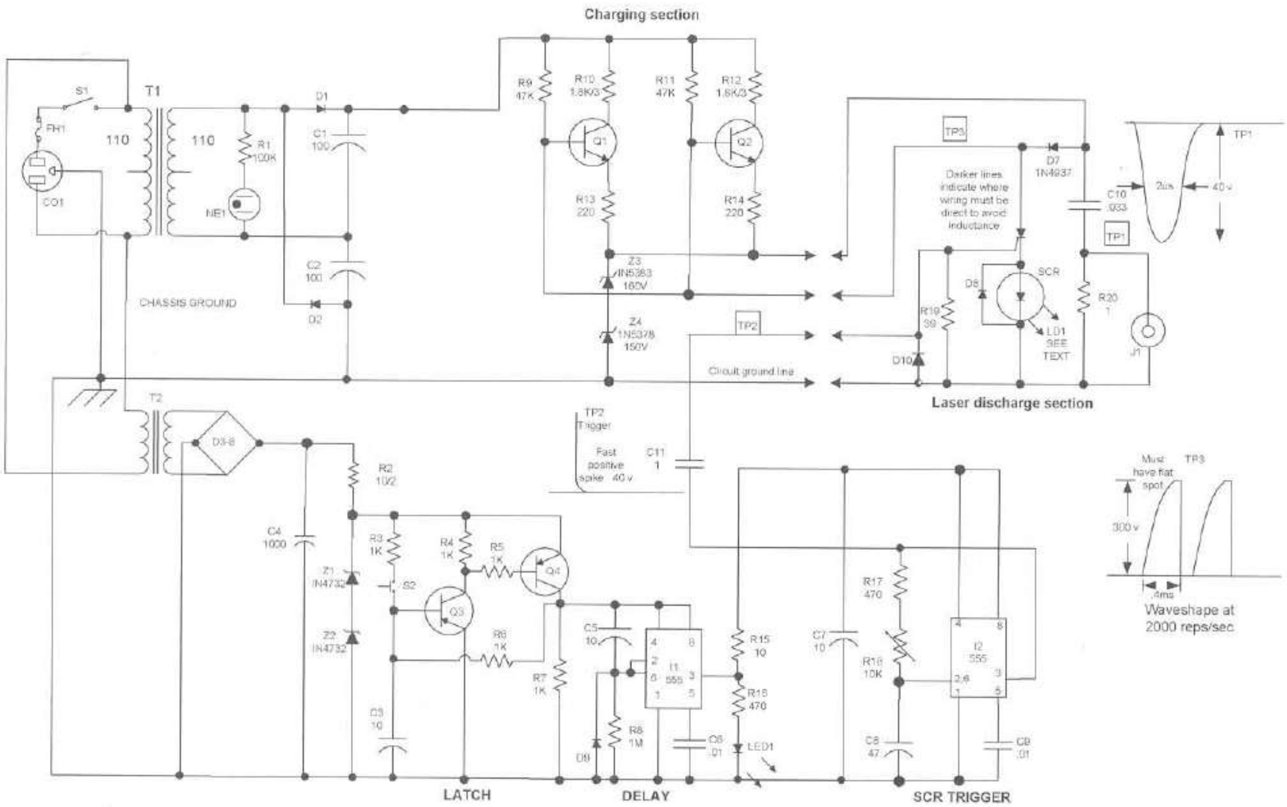
Charging section
S1
T1
FH1
110
110
CO1
R1
100K
NE1
D1
C1
100
C2
100
D2
CHASSIS GROUND
R9
47K
R10
1.8K/3
R11
47K
R12
1.8K/3
Q1
Q2
R13
220
R14
220
Z3
1N5383
150V
Z4
1N5378
150V
TP2
Circuit ground line
TP3
D7
1N4937
Darker lines indicate where wiring must be direct to avoid inductance
C10
.033
TP1
SCR
D8
R19
39
R20
1
D10
LD1
SEE TEXT
J1
Laser discharge section
TP1
2us
40 v
T2
D3-8
R2
10/2
TP2
Trigger
Fast positive spike 40 v
C11
1
C4
1000
Z1
IN4732
Z2
IN4732
R3
1K
S2
R4
1K
R5
1K
Q3
Q4
R6
1K
R7
1K
C3
10
C5
10
I1
555
R8
1M
D9
C6
.01
R15
10
R16
470
LED1
C7
10
R17
470
R18
10K
I2
555
C8
47
C9
.01
LATCH
DELAY
SCR TRIGGER
Must have flat spot
TP3
300 v
.4ms
Waveshape at 2000 reps/sec

Figure 4-2 *Circuit schematic*

current but should be as low as practical because this resistance also determines the charging time of the capacitor C10, thus determining the available laser pulse repetition rate at full energy. For example, a resistance of 40 kilo-ohms limits the current to 10 milliamps from a 400-volt supply. This current value is just at the holding level of an average SCR. A time of almost 4 milliseconds is required to charge C10, a 0.033-microfarad capacitor, to the supply voltage in three time constants through a 40-kilo-ohm resistor. Therefore, the *pulse repetition rate* (PRR) of the pulsing circuit is limited to less than 250 Hz. If the PRR exceeds this value, the capacitor does not completely recharge between pulses and the peak laser current decreases with increasing PRR.

Varying the supply voltage may control the peak current in the discharge circuit, provided the PRR is low enough to allow the capacitor to fully recharge between current pulses. Both the supply voltage and the PRR determine the peak laser current. A considerable risk exists in increasing the supply voltage to compensate for an insufficient recharge time. If the PRR is decreased while the supply voltage is high, the capacitor again recharges completely, and the laser pulse current increases to a value that may damage or destroy the laser diode.

The needs for a variable voltage supply and the low PRR limit are the major disadvantages of the resistive-type charging network. Therefore, the limitations of the simple resistor drivers are that a resistor large enough to keep the current below the holding current of the SCR also limits the pulse repetition rate.

The frequency capability of the pulse power supply can be greatly improved with the charging circuits, as described. Capacitor C10 is charged to the supply voltage when the SCR is not conducting. Diode D7 is in the off state, because no current flows through it during this period. At the onset of a pulse trigger at the gate, the impedance of the SCR drops rapidly and capacitor C10 discharges toward ground potential. During this period, current is surging through diode D7, causing zero biasing of the current-control transistors (Q1, Q2) by shorting the bias resistors (R9, R11). The SCR now turns off when the current drops below its holding current. The voltage across capacitor

C10 then charges back to the supply voltage. During the capacitor charge cycle, diode D7 passes no current, and Q1 and Q2 are forward biased into the saturation region. Obviously, the charging time of C10 is now shorter due to being charged through the smaller-value resistors (R10, R12). The emitter resistors R13 and R14 balance the current through Q1 and Q2.

Zener diodes (Z3, Z4) are selected to regulate and provide the proper required current pulse for the laser diode in operation. The zener diodes provide a maximum charge voltage of 310 volts that produces 40 amps through laser diode LD1 for up to 2,500 laser pulses per second.

Other laser diodes may require a selection of other voltage-rated zener diodes. The builder may wish to provide a multiposition switch to select these various combinations of zener diodes externally via a front panel control.

The trigger circuit consists of an astable timer (I2) that derives its power through the output of timer I1, which provides a delay time required for FDA laser compliance. The trigger and delay circuitry, as well as the timers, are powered by 12 VDC from transformer T2 and rectifier diodes D3 through D6. Capacitor C4 filters the rectified wave and steps up the voltage to 15 VDC. Resistor R2 drops the 15 volts to the zener voltage of zener diodes Z1 and Z2 of regulated 12 VDC.

Operation of the laser requires activating pushbutton switch S2, which latches transistor Q4 upon supplying power to timer I1. The timer must time out to 10 seconds, as determined by resistor R8 and capacitor C5. The emission indicator LED1 comes on simultaneously with power to the trigger timer I2, which pulses the laser diode LD1. The laser pulse rate is determined by external pot R18.

Notes on SCR

In the conventional operation of an SCR, the anode current, initiated by a gate pulse, rises to its maximum value in about 1 microsecond. During this time, the anode-to-cathode impedance drops from open circuit to a fraction of an ohm. In injection laser pulsers, however, the duration of the anode-cathode pulse is much less than the time required for the SCR to turn on completely. Therefore, the anode-to-cathode impedance is at the level of 1 to 10 ohms throughout most of the conduction period. The major disadvantage of the high SCR impedance is that it causes low circuit efficiency. For example, at a current of 40 amps, the maximum voltage would be across the SCR, while only 9 volts would be across the laser diode, which represents very low circuit efficiency. The efficiency of a laser array is greater due to its circuit impedance being more significant.

Because the SCR is used unconventionally, many of the standard specifications such as peak current reverse voltage, on-state forward voltage, and turn-off time are not applicable. In fact, it is difficult to select an SCR for a pulsing circuit on the basis of normally specified characteristics. The specifications important to laser pulser applications are forward-blocking voltage and current rise time. A usage test is the best and many times the only practical method of determining the suitability of a particular SCR.

Notes on the Storage Capacitor

The voltage rating of the storage capacitor must be at least as high as the supply voltage. A high-quality metallized capacitor must be used for this part as peak currents are in the tens of amps.

Notes on Layout Wiring

Lead lengths and circuit layout are very important to the performance of the discharge circuit. Lead inductance affects the rise time and peak value of the current and can also produce ringing and undershoots in the current waveform that can destroy the laser. A well-built discharge circuit might have a total lead length of only 1 inch and therefore an inductance of approximately 20 nanohenries. If the current rises to 75 amperes in 100 nanoseconds, the inductive voltage drop will be $e = L\, di/dt$. If proper care is not taken in wiring the discharge circuits, high-inductance voltage drops will result.

A 1-ohm resistor (R20) in the discharge circuit will greatly reduce the current undershoot in single-diode lasers. Laser arrays usually have sufficient resistance

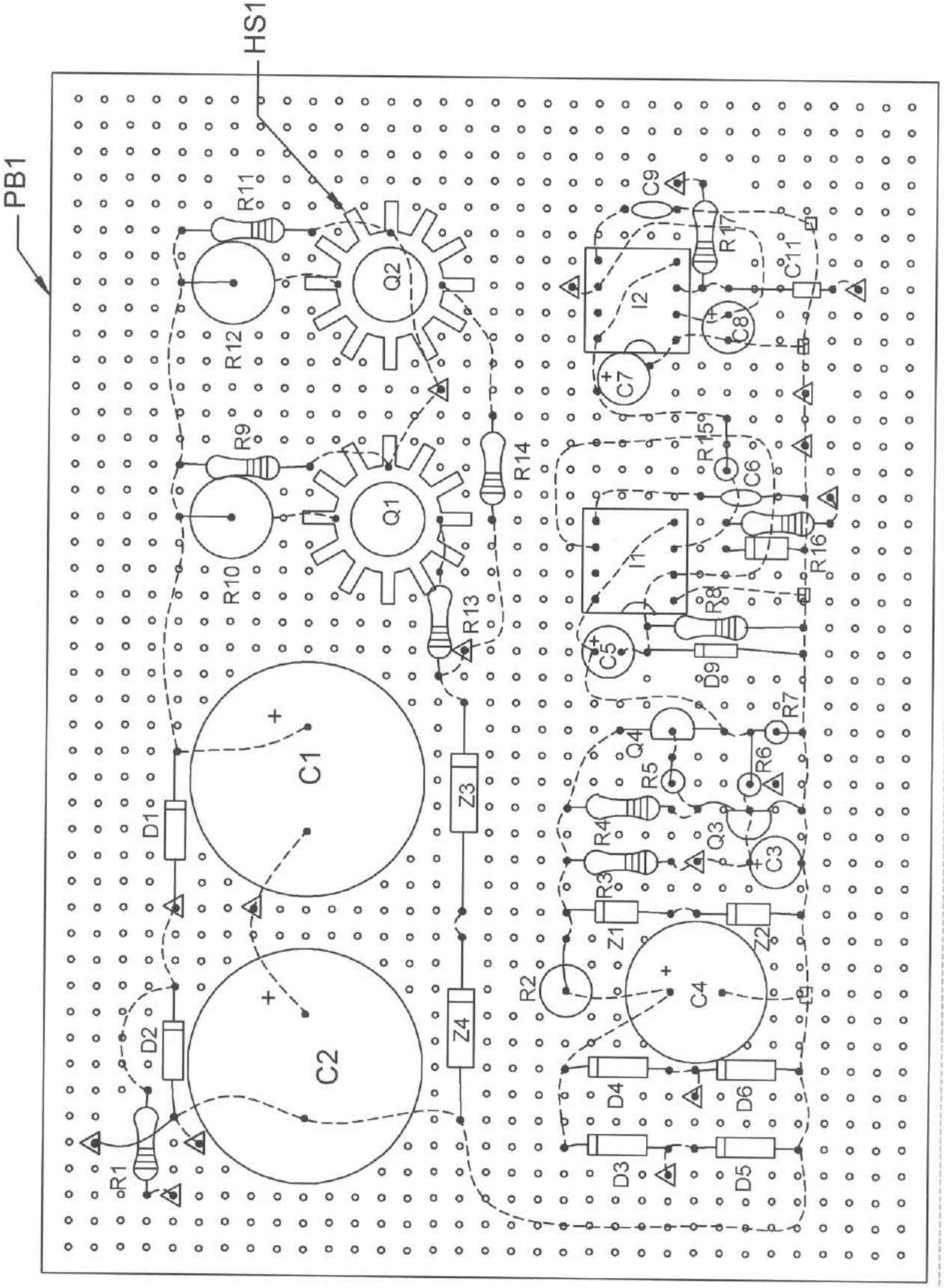
PB1
HS1
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
C1
C2
C3
C4
C5
C6
C7
C8
C9
C11
Q1
Q2
Q3
Q4
D1
D2
D3
D4
D5
D6
D9
Z1
Z2
Z3
Z4
I1
I2
+

Figure 4-3 *Board assembly*

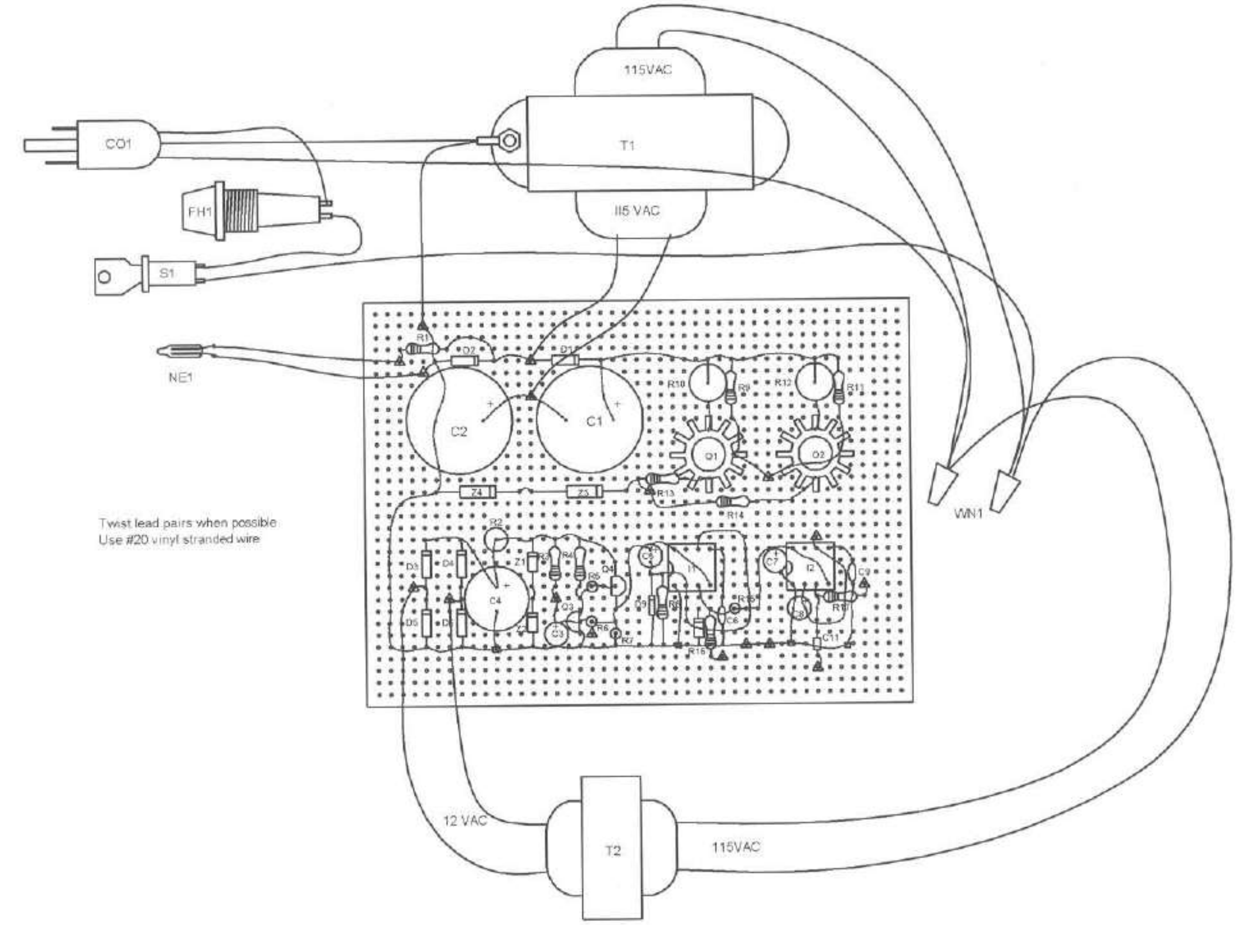

Figure 4-4 *Board input power wiring.*

to eliminate an undershoot. The small resistance in the discharge circuit is also useful in monitoring the laser current, as described in the following section.

A clamping diode (D8) is added in parallel with the laser to reduce the current undershoot. Its polarity should be opposite that of the laser. Although the clamping diode is operated above its usual maximum current rating, the current undershoot caused by ringing is very short and the operating life of the diode is satisfactory.

Notes on Current Monitoring

The current monitor in the discharge circuit provides a means of observing the laser current waveform with an oscilloscope. A resistive-type monitor (R20) reduces circuit ringing and current undershoot, but the lead inductance of the resistor may cause a higher than actual current reading. A current transformer such as the Tektronix CT-2 can also be used to monitor the current and is not affected by lead inductance. Because the transformer does not respond to low-frequency signals, it should be used with fast, short-pulse-width, fast fall-time waveforms.

Assembly Steps

Assembly of the laser system is straight forward and will require basic tools and test equipment.

1. Lay out and identify all parts and pieces. Verify them with the parts list, and separate the resistors as they have a color code to determine their value. Colors are noted on the parts list.
2. Cut a piece of .1-inch grid perforated board to a size of 5.8 × 3.5 inches, as shown in Figure 4-3.

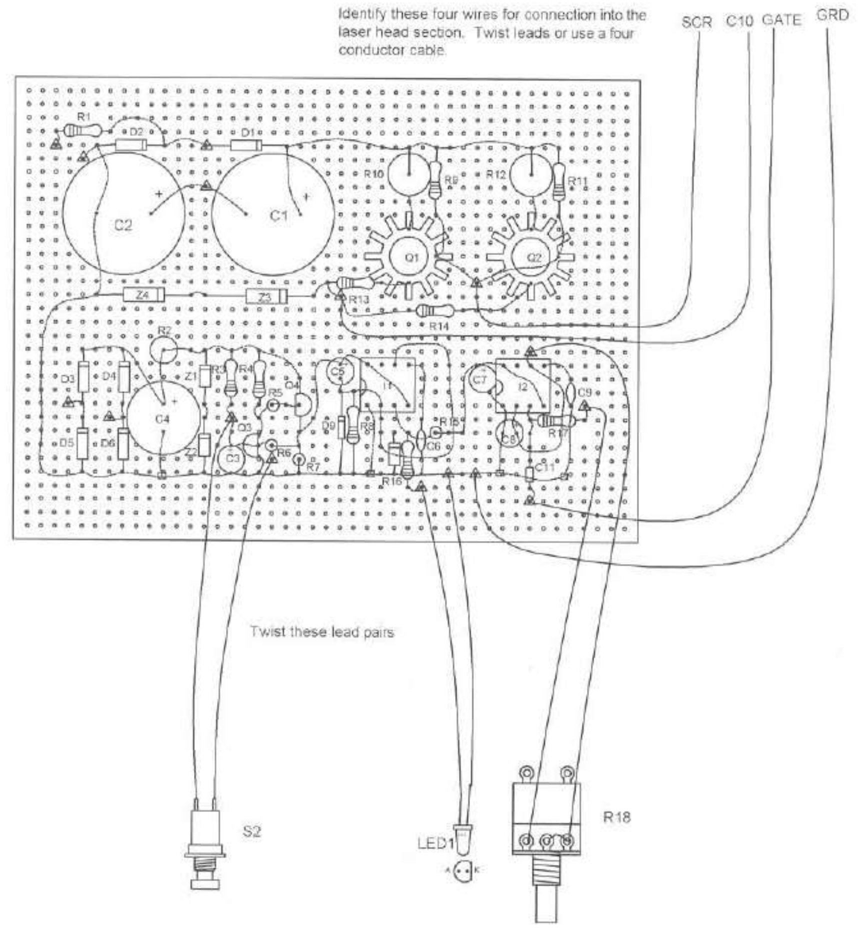
Identify these four wires for connection into the laser head section. Twist leads or use a four conductor cable.
SCR
C10
GATE
GRD
R1
D2
D1
C2
C1
R10
R9
R12
R11
Q1
Q2
Z4
Z3
R13
R14
R2
D3
D4
Z1
R3
R4
Q4
R5
C5
I1
C7
I2
C9
C4
D5
D6
Z2
C3
Q3
R6
R7
R8
C6
R15
R16
C8
R17
C11
Twist these lead pairs
S2
LED1
A
K
R18

Figure 4-5 *Board output and control wiring*

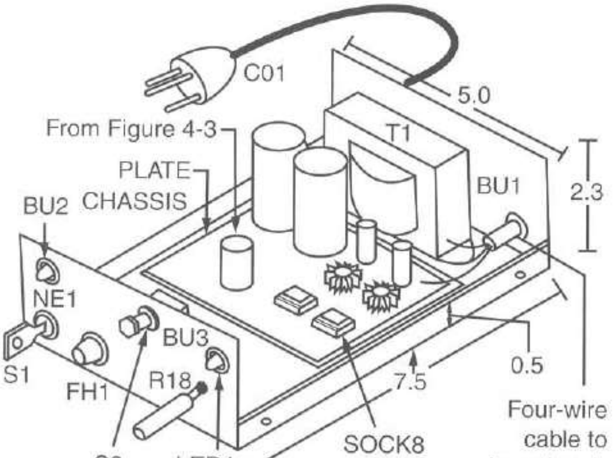
C01
5.0
From Figure 4-3
T1
PLATE
CHASSIS
BU2
BU1
2.3
NE1
BU3
S1
FH1
R18
7.5
0.5
Four-wire cable to laser head section
SOCK8
S2
LED1

Fabricate the front panel holes as required for the components used.Trial fit before drilling.

Figure 4-6 *Fabrication and chassis assembly*

3. If you are building from a perforated board, insert the components, starting in the lower left-hand corner as shown. Pay attention to the polarity of the capacitors that have polarity signs and all semiconductors. Use eight-pin sockets for I1 and I2.

   Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of the assembly board. The heavy dashed lines indicate the use of thicker #20 bus wire, as this is a high-current discharge path.

4. Attach the external leads as shown in Figure 4-4 to the external components. Note the chassis grounding to the green lead of the power cord (CO1) and the use of wire nuts (WN1) for the junction of the transformer's primary inputs.

5. Attach the external leads as shown in Figure 4-5 to the external components. Note the four leads that go to the laser head assembly. These

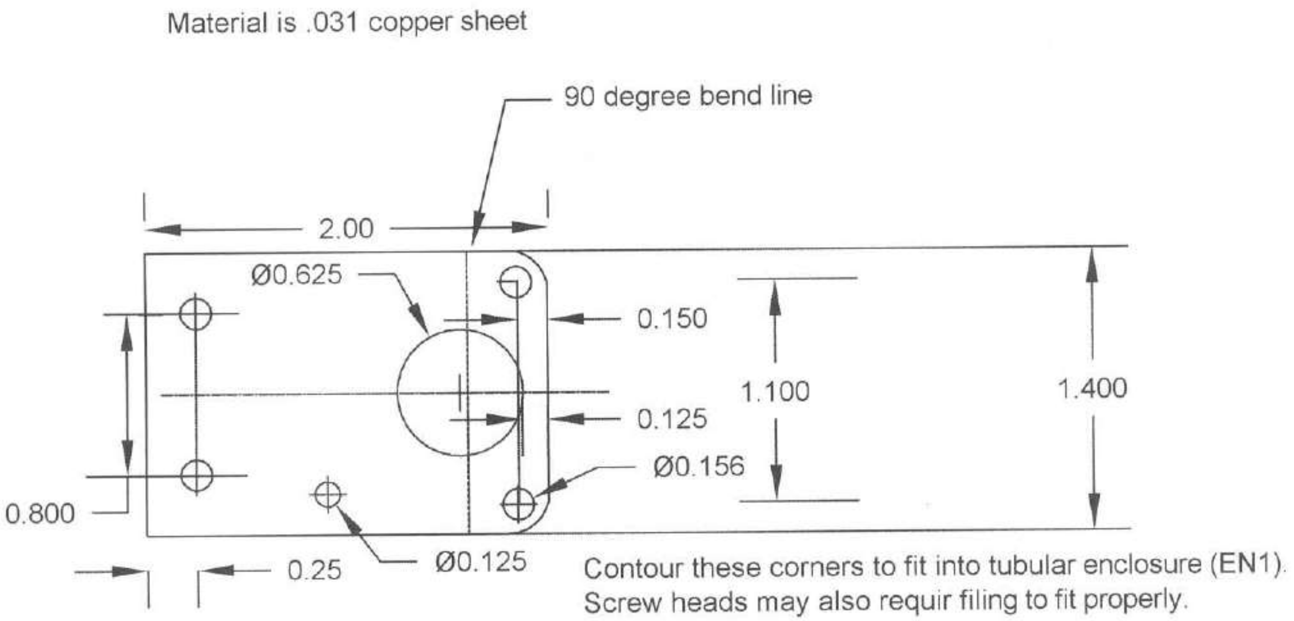
Material is .031 copper sheet
90 degree bend line
2.00
Ø0.625
0.150
1.100
1.400
0.125
Ø0.156
0.800
0.25
Ø0.125
Contour these corners to fit into tubular enclosure (EN1).
Screw heads may also requir filing to fit properly.

Figure 4-7 *Assembly of laser chassis*

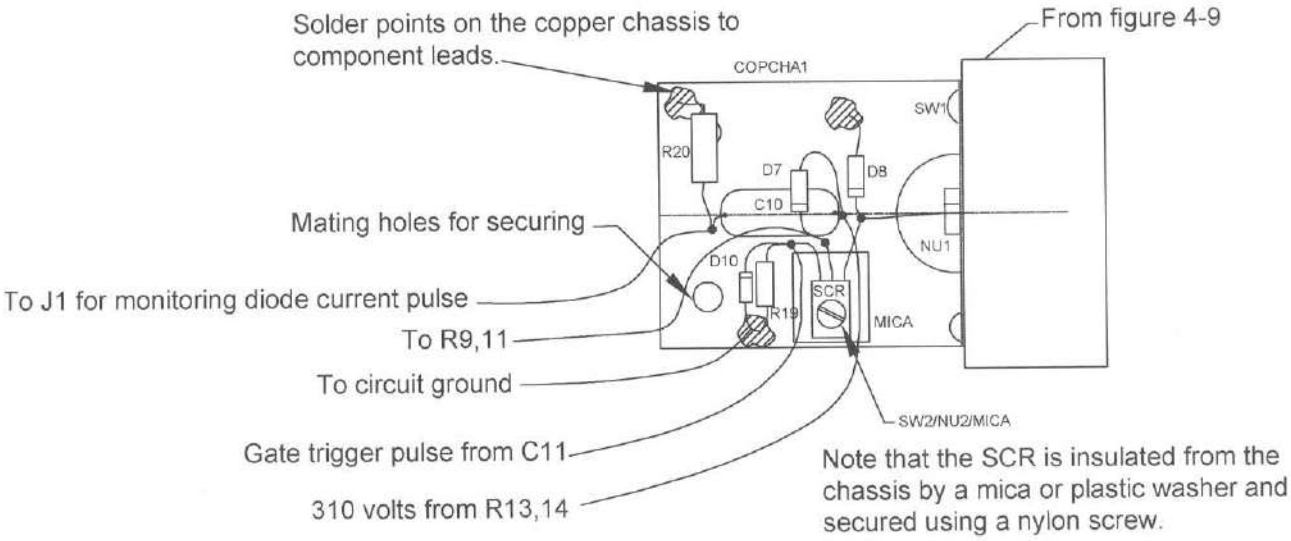
Solder points on the copper chassis to component leads.
From figure 4-9
COPCHA1
SW1
R20
D7
D8
C10
NU1
Mating holes for securing
D10
To J1 for monitoring diode current pulse
SCR
R19
MICA
To R9,11
To circuit ground
SW2/NU2/MICA
Gate trigger pulse from C11
310 volts from R13,14
Note that the SCR is insulated from the chassis by a mica or plastic washer and secured using a nylon screw.

Figure 4-8 *Fabricate the laser holder*

are suggested to be 24 to 36 inches in length as well as twisted and identified. You may use a four-conductor cable.

6. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.

7. Fabricate the chassis, as shown in Figure 4-6. Then position and drill the front and rear panel holes to fit the necessary external components. Use appropriate bushings for the power cord and wires to the laser head. Also, place an insulating piece (PLATE) under the assembly board to prevent contact with the metal chassis.

8. Fabricate the laser chassis, as shown in Figure 4-7, from a .031 piece of copper sheet metal.

9. Assemble the laser chassis, as shown in Figure 4-8. This section takes advantage of the ground plane effect of using the copper for

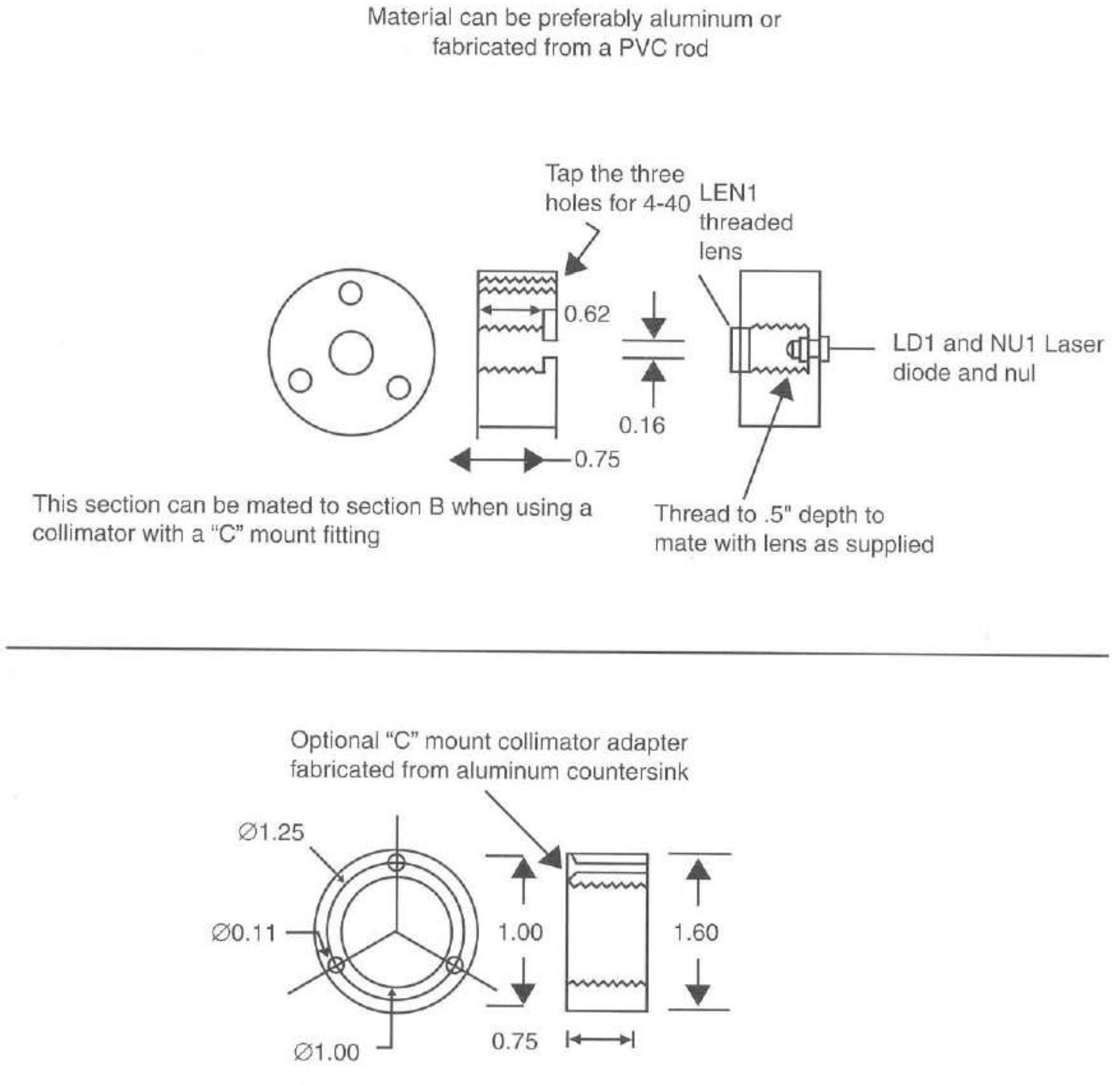

Material can be preferably aluminum or fabricated from a PVC rod

This section can be mated to section B when using a collimator with a "C" mount fitting

Thread to .5" depth to mate with lens as supplied

Optional "C" mount collimator adapter fabricated from aluminum countersink

Section B adapter for fitting to the lens holder when using the optional collimator.

Three mounting holes of .11 on a 1.25" diameter 120 degres are mated to those in above.

Thread for "C" lens at 1" diameter 32 tpi to a minimum of .375".

Figure 4-9 *Laser diode and lens holder*

direct solder points. It is important to maintain a good waveshape because the switching currents are 40 amps and any undershoot is undesirable across the laser diode. Do not install the laser diode at this point.

10. Create the laser holder (LAHOLD), as shown in Figure 4-9. This assembly is attached to the bent-up lip of the chassis section by two small screws.

Assembly Testing

Testing will involve the use of an oscilloscope with a 100 MHz bandwidth and infrared detection equipment such as a night vision scope or camera.

1. It is assumed at this point that the electronic assembly is complete. Verify the circuit and preset the two trimpots, R3 and R9, to midrange.

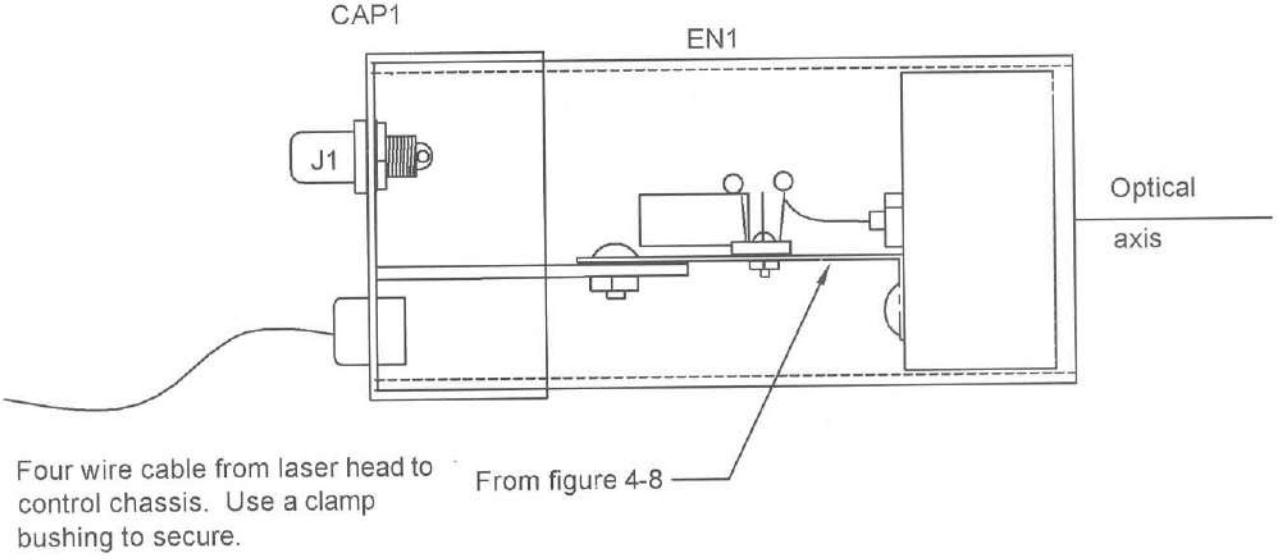
CAP1
EN1
J1
Optical
axis
Four wire cable from laser head to control chassis. Use a clamp bushing to secure.
From figure 4-8

Figure 4-10 *Final head assembly*

2. Insert a .5-amp fuse into the fuse holder. Verify that the key switch is off and R18 is fully counterclockwise. Plug in a ballasted, 25-watt, 115-volt, 60 Hz power source. This is nothing more than the cord with a 25-watt lamp in series with the black lead. It protects the circuitry from the catastrophic currents of the power line should a major fault occur. Obtain a 1N4007 power diode and temporarily wire it in place of the laser diode LA1. The diode simulates the actual laser diodes and avoids costly replacement should a gross error exist. The actual laser diode is not to be wired in at this point. Turn on the key switch and note the NE1 coming on. The ballast lamp should not be lit.

13. Connect the scope to TP2 and note a spiky waveform that verifies the trigger pulse to the SCR, as shown in Figure 4-2.

14. Connect the scope to TP3 and note the waveform, which is the voltage across the SCR. Notice the flat spot indicating that capacitor C10 has reached a full charge before the SCR is triggered. The top of this waveshape is affected by the pulse repetition rate and should support 2,500 repetitions per second when set by R18 (see Figure 4-2).

15. Connect the scope to TP1 and set it up for a negative-going fast pulse, indicating the current waveform through the LD1 laser diode. If you have followed the instructions and your assembly of the copper chassis is secure with short leads, you should see a picture-perfect waveform, as shown.

    Note that the current pulse is 40 amps and corresponds to 1 amp for every volt read at TP1. If you find the pulse is less, you can increase the voltage ratings on one of the zener diodes, Z3 or Z4. The laser diode specified requires 40 amps for full output. Any value over 40 amps will greatly shorten the useful life of this part. The range of the pulse rep rate should be 200 to 2,500 pulses per second. The pulse amplitude should only vary several percent throughout this range. You will note that the amplitude will slightly decrease as the rep rate is increased.

16. At this point, you should be wearing your safety glasses. Remove the test diode D5 and connect the laser diode. Place an infrared indicator several inches from the diode and note an orange glow when the unit is turned on. Then recheck the pulse shape and amplitude.

17. Verify the laser output by using a night vision scope or a video camera. Screw in LENS13, so it is flush with the holder. Adjust it to find the smallest footprint at over 50 feet. Our lab model adjusted to about two turns below being flush with the holder for the best results. The output should be a sharp, narrow bar shape with this diode specified.

Please note that laser diodes are very easily destroyed by current pulses that are too large and overshoot. Pay attention to the correct current pulse shown in Figure 4-2.

The recommended laser diodes are available from Information Unlimited and are shown in Table 4-1.

Final Assembly Steps

Final assembly of head section as shown in Figure 4-10. For final assembly place the label as shown and verifying the compliance tests in Figure 4-11

Compliance Tests

1. Verify correct labels as shown.
2. Key switch — nonremoval in off position.
3. Beam indicator LED indicates beam emission.

Instruction

1. Turn key switch counterclockwise to verify unit is off. Note key is easily removed in the off position.
2. Plug unit into 115 vac.
3. Turn key clockwise and note NE1 lighting. Note key cannot be removed in the on position Laser energy is infrared and cannot be seen with the eye. You must use infrared detection devices such as our laser pulse detector or suitable night-vision devices. Video cameras will often detect the laser.
4. Adjust collimator for desired effect if you are using.

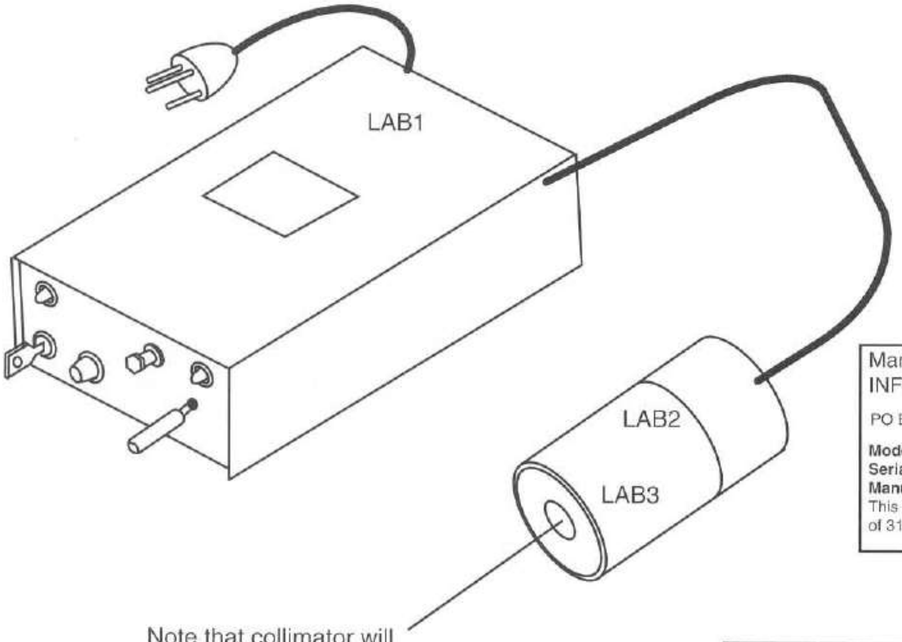

Note that collimator will expand beam at near field but greatly reduce it at far field.

Manufactured By
INFORMATION unlimited
PO Box 716, AMHERST, NH 03031
Model NUMBER
Serial Number
Manufactured Date
This laser product conforms to the provisions of 31CFR1040 10 and 1040 11

LAB1

AVOID EXPOSURE
LASER RADIATION IS EMITTED
FROM THIS APERTURE

LAB3

DANGER

VISIBLE OR INVISIBLE
LASER RADIATION
AVOID DIRECT EXPOSURE
TO BEAM

CLASS IIIB LASER PRODUCT

LAB2

Figure 4-11 *Isometric of final view*

Table 4-2 Parts List for Pulsed Laser

Ref #	*Description*	*DB Part #*
R1	100K, ¼ –watt resistor (br-blk-yel)	
R2	10-ohm, 1-watt resistor (br-blk-blk)	
R3, 4, 5, 6, 7	Two 1K ¼-watt resistors (br-blk-red)	
R8	1M, ¼-watt resistor (br-blk-gr)	
R9, 11	47K, ¼-watt resistor (yel-pur-or)	
R10, 12	1.8K, 3-watt mox resistor	
R13, 14	220, 1/2-watt resistor (red-red-br)	
R15	10-ohm, ¼-watt resistor (br-blk-blk)	
R16, 17	470-ohm, ¼-watt resistor (yel-pur-br)	
R18	10K, 10 mm potentiometer	
R19	39-ohm, ¼-watt resistor (or-wh-blk)	
R20	1-ohm, ½-watt carbon resistor (br-blk-gold)	
C1, 2	Two 330 mfd, 16-volt vertical electrolytic capacitor	
C4	1,000 mfd, 25-volt vertical electrolytic capacitor	
C3, 5, 7	10 mfd, 25-volt vertical electrolytic capacitor	
C6, 9	.01 mfd, 50-volt disc capacitor	
C10	.033 mfd, 400-volt polypropylene capacitor	
C11	.1 mfd 50-volt plastic capacitor	
D1, 2	1N4007 1KV, 1-amp rectifier diodes	
D3–6	Four 1N4001 50-volt 1-amp rectifier	
D7, 8	1N4937 fast recovery 1KV, 1-amp diode	
D9, 10	Two 1N914 signal diodes	
Z1, 2	1N4735 6.2-volt zener diode	
Z3	1N5383 150-volt, 5-watt zener diode	
Z4	1N5384 160-volt, 5-watt zener diode	
Q1, 2	Two 2N3439 NPN high-voltage transistors	
Q3	PN2907 PNP general-purpose transistor	
Q4	PN2222 NPN general-purpose transistor	
I1, 2	Two timer dip integrated circuits LM555	
LED1	Green LED	
NE1	Neon indicator lamp	
SCR1	12-amp, 800-volt SCR S8012R	DB# S8012R
LD1	Laser diode 40-amp, 10–20-watt, 904 nm TO18 package	DB# LD650
LAHOLD	Laser and lens holder fabricated as shown in Figure 4-9	
LENS13	Threaded lens	
T1	115/115 volt, 15-volt-amp isolation transformer	DB# TR115/115
T2	115/12 volt, 1-amp transformer	DB# 12DC/.1
J1	RCA phono jack	
CO1	Three-wire #18 power cord	
S1	Compliant key switch	DB# KEYSWSM
S2	Pushbutton no switch	
FH1	Panel-mount fuse holder	
FS1	1-amp fuse	
WN1	Two small Hi3 wire nuts	
PB1	4.8 × 3.5-inch .1 × .1 perforated circuit vector board	
SOCK8	Two 8-pin integrated circuit sockets	

Chapter Five

Laser Property Guard

This project, as shown in Figure 5-1, allows the user to protect a perimeter of over 1,000 meters. The system utilizes a laser transmitter, as described in Chapter 1, and a laser receiver, described in Chapter 2. The system can easily control any deterrent to an invader, such as flood lighting, alarms, man-trap systems, or our phaser pain field generator, as described in Chapter 35.

The operation of the laser property guard is simple. The laser beam is directed through the area to be protected, using front-surface mirrors located on sturdy and secure mounts and at the points necessary to obtain the required coverage (see Figure 5-2). The receiver is normally quiescent when receiving this beam. Once the beam is broken by an intrusion, the receiver now processes the interruption as a fault and powers the desired deterrent or alarm.

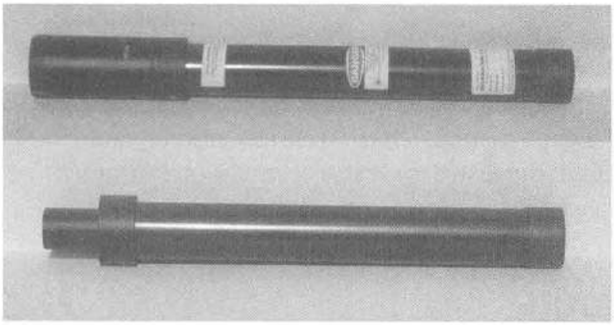

Figure 5-1 *Laser property guard protector components*

This project requires assembly of the laser devices, as described in Chapters 1 and 2. The setup requires basic mechanical skills and installation of the optical reflecting mirrors, and the hookup requires basic electrical skills. Expect to spend $25 to $50 after the assembly of the necessary components. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com) and listed in Table 5-1.

The laser property guard system, as shown in Figure 5-3, utilizes a pulsed infrared laser to provide a perimeter path of programmed light pulses that, when interrupted, will sound an alarm, create an alert, or trigger a deterrent. The trigger fault signal actually triggers a timer or a latching circuit, which drives a switch for the external control of other functions. The timer mode allows the adjustment of the alarm period and automatically resets. The latch mode requires a manual reset.

Even though the laser pulses are rated in tens of watts, the overall energy over a period of 1 second is in the milliwatts. Installing the system at short ranges of only one to two reflection points is reasonably simple. Longer ranges requiring multiple reflections can be challenging.

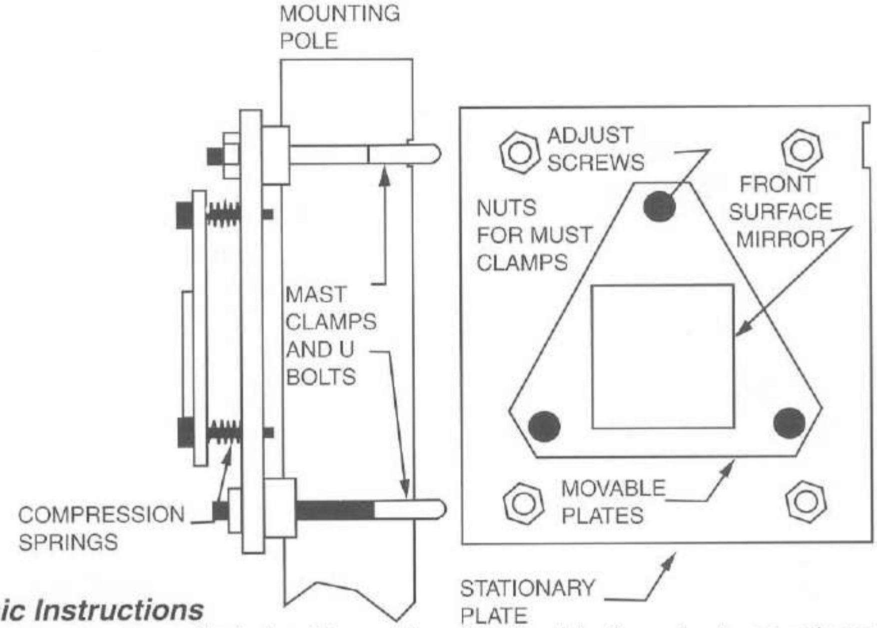

Basic Instructions

Your laser protector guard is designed to provide an intrusion detection perimeter around your home or target area. It uses a Class iiia pulsed diode laser to minimize but not eliminate any optical hazard. A low-liability sonic shock field generator such as that described in Chapter 35 is suggested as a deterrent. Caution: Check local laws for proper posting of this equipment on your property. Remember if you injure a criminal, even if he is robbing you, it can result in a stiff penalty in certain states for violating his rights.

System Setup

1. Obtain the following equipment listed in the latest Information Unlimited catalog.
   - ☐ Class iiia pulsed diode laser as described in Chapter 1.
   - ☐ Optical light detector and controller with built-in alarm as descrbed in Chapter 2.
   - ☐ Low-liability sonic shock pain field generator as described in Chapter 35.

2. Examine the area you want to protect. Attempt to position beam travel over level terrain to eliminate possible easy-to-sneak-under points.

3. Establish corner points of property where mirrors are to be positioned. Note that a clear view of adjacent mirrors, laser and receiver must be maintained in all weather conditions for reliable operation.

4. Determine the corner where you want to position the actual laser and laser detector at. Note: There should be access to these devices for powering and control.

5. Construct mirror mounts using your own ingenuity. Allow for adjustment and stability. Rough align mirror to approximate position.

6. Turn on laser and position to hit center of first mirror. Secure laser in place.

Adjust first mirror to hit center of second and repeat for remaining mirrors all the way back to the optical receiver. Secure in place as you go.

Special note: Large areas may require an optical collimator at the laser output to reduce beam width at longer ranges. The collimator can be a rifle scope, telescope, binocular, etc., positioned in axial alignment with the laser. Beam divergence will be reduced by the magnification factor of the system used, however the actual beam will be expanded by same factor. The net result is a smaller cross-section at longer ranges.

Figure 5-2 *Low-cost, adjustable mirror mount*

System Installation Suggestions

There are many ways to align an optical system of this type. The use of a visible laser module as listed in Table 5-1 can simplify the initial setup and positioning of the mirrors. Those fortunate to have night vision equipment will find this a considerable advantage.

- Determine the area you want to protect. Flat land is the most desirable, as gullies and hills are difficult to cover without using more mirrors.
- You must also consider the position of the laser transmitter and receiver, noting the power and connection into the alarm or the deterrent section. Weather exposure and convenience must also be taken into consideration, and mounting must be stable and somewhat adjustable.
- Determine the path that the light must travel to provide acceptable coverage. Always attempt to use as few mirrors as possible and consider the laser transmitter's and receiver's positions. If you are lucky, you may have some objects to mount the mirrors to. They must be stable and allow placement at the required height for your application.

  The mirrors used for this project must be the front-surface-coated type if multiple reflections are required. Regular household mirrors can be used up to two reflection points, but they will cause multiple reflections. A low, adjustable-cost mirror mount is shown in Figure 5-2. You may in certain cases use a good-quality automotive mirror, as they are easily adjustable.
- Evaluate the wildlife in your area and determine the best laser height that will provide minimal false alarms from the animal population. Obviously, deer will always be a problem, as their height may be close to that of the target intruder.
- Determine the deterrent you wish to use against the intruder. You can activate bright lights, sonic pain generators (described in Chapter 35), alarms, or just alerts so you can take further action. For more serious applications, you can activate shocking devices, such as those described in Chapter 11, nonlethal kinetic devices, or explosive charges for shock and surprise. Very serious deterrents can consist of deadly force, such as high-powered kinetic devices, shotgun traps, high-powered shocking devices, and buried explosives for lethality.

Caution: As man trapping is illegal in the United States, it may also be so in your country. Always check your laws before ever implementing a deterrent that can cause injury.

- The quality of the mirrors and their mounts cannot be overstressed when designing a high-integrity system. Here we discuss a low-cost system that may suffice in smaller installations but may be limited for use in larger, more reliable systems.

The initial alignment requires a visible, low-power laser to align the mirrors. The laser may have to be collimated for larger multiple reflection systems.

Optics on the laser light transmitter and receiver, as shown, may be sufficient for small systems. They do not have to be high-cost precision devices, as the only objective is to establish a properly aligned, position-stable beam spot.

Once the preliminary alignment is accomplished, you can put the laser transmitter in place and, with a night vision scope, recheck the alignment. The laser receiver, as constructed here, has an indicating *light-emitting diode* (LED). This function can be eliminated if required. The actual control of external functions is via a set of normally open or closed relay contacts that can handle 115 VAC at 10 amps AC.

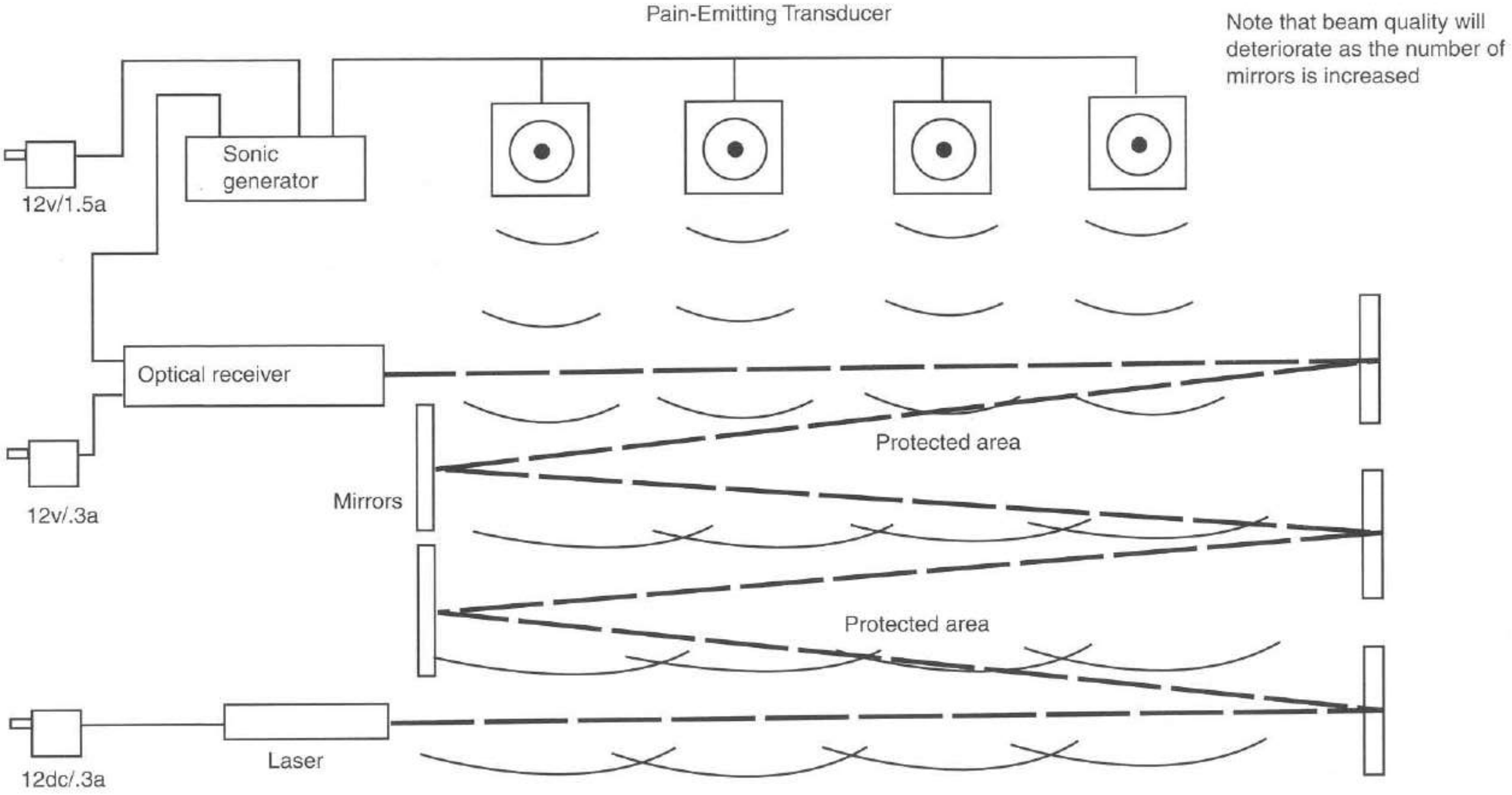
Pain-Emitting Transducer
Note that beam quality will deteriorate as the number of mirrors is increased
Sonic generator
12v/1.5a
Optical receiver
12v/.3a
Mirrors
Protected area
Protected area
Laser
12dc/.3a

Laser beam is reflected back and forth in the protected area via strategically placed mirrors. Once the beam is broken by an intruder the optical receiver detects and sends a level to the sonic generator where the transducers are activated producing an intolerable sonic pain field discouraging the intrusion.

Even though there is no lasting effect from exposure to the transducers, it is always best to check with local laws before implementing this system.

Figure 5-3 *Example of system layout*

Table 5-1 Laser Property Guard Parts List

Ref. #	*Description*	*DB Part #*
MIRROR	3 × 4-inch front surface mirrors	DB# MIR FRSUR
LRG4K	Pulsed infrared laser kit, as in Chapter 1	DB# LRG4K
LRG40	Assembled pulsed infrared laser, as in Chapter 1	DB# LRG40
LSD4K	Laser detector kit, as in Chapter 2	DB# LSD4K
LSD40	Assembled laser detector, as in Chapter 2	DB# LSD40
PPF4K	Pain field generator kit, as in Chapter 35	DB# PPF4K
PPF40	Assembled pain field generator, as in Chapter 35	DB# PPF40
SHOCK1K	Kit of shockers, as in Chapter 11	DB# PFSHK1K
SHOCK10	Assembled shockers, as in Chapter 11	DB# PFSHK10
LM650-3	Visible red laser module for alignment aid	DB# LM650- Description P3System

Chapter Six

30-Milliwatt, 980-Nanometer Infrared Laser

The laser project shown in Figure 6-1 is a continuous infrared device providing a continuous 30-milliwatt output at 980 nanometers. The laser, when built as shown, should comply with a class 3 laser-compliant system. The completed laser can be very optically dangerous, and eye protection must be worn to prevent direct or scattered reflections to the eyes.

The project can be used for many optical applications, including illuminating a surface, such as a window, for laser listening. Laser listening is were a laser beam illuminates a window and a sensitive receiver picks up the light reflections and processes them into audio signals similar to the picked-up voices that vibrate the illuminated window glass. A system such as this is described on www.amazing1.com. It should be noted that using a class 3b laser for this application can be an optical hazard to anyone in the target area.

Converting illumination using night vision cameras and spectrally compatible viewers is another potential application, again being fully aware of a possible optical hazard. Other applications are invisible target identification, voice and data communication, and optical signaling and experimenting.

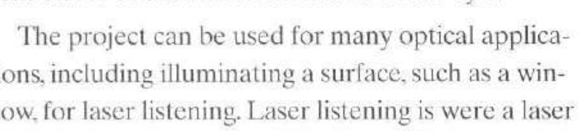

Figure 6-1 *The infrared laser module*

This is an intermediate-level project requiring basic electronic skills. Safety glasses are strongly advised, as the laser can cause eye damage if viewed directly or via direct reflections. Expect to spend $75 to $150. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 6-1.

For those wanting to use this project as an illuminating laser for window bounce listeners, it is suggested that you refer to Chapter 13 in *Electronic Gadgets for the Evil Genius*, published by McGraw-Hill, ISBN 0-07-142609-4.

Circuit Description

Figure 6-2 shows the laser module LM1 and associated support circuitry. The system is powered by batteries B1 and B2 and is controlled by an internal key switch (S1) or external enabling control leads. A light-emitting diode (LED1) and a current-limiting resistor (R1) indicate the presence of laser emission.

The system will require S1 to be a key switch with a nonremovable key in the activated position if the assembly is integrated as a fully functional laser system. The power line or bench operation will require a

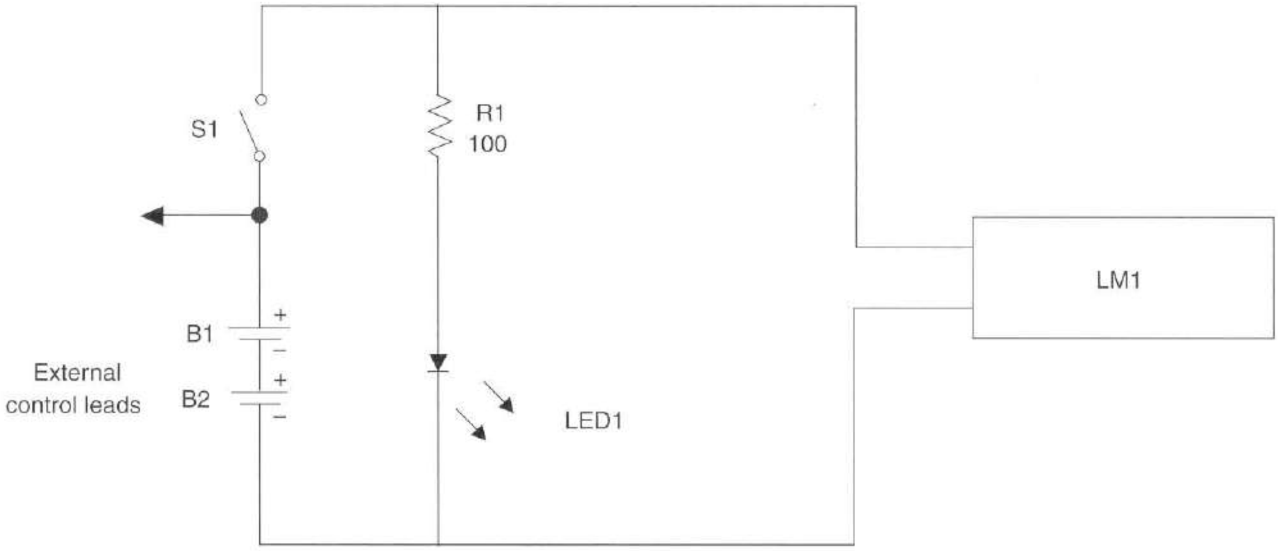

Laser current should be 50 to 60 milliamps at 3 volts input.

Warning: If you overvolt or reverse polarity on your laser module, you will ruin it!!

Assembly shows a battery-operated compliant laser system. You may wire direct to a 3-volt source that now contains the key switch and a built-in time delay necessary for compliance and safety.

Figure 6-2 *Circuit schematic*

key switch for external enabling and associated control circuitry that may require an emission delay and automatic resetting should a loss of power occur.

Assembly

1. Rework the laser module, as shown in Figure 6-3. This step requires care as it is easy to damage the module if one is not very careful.
2. Connect the soldered leads to a source of 3 volts, paying attention to the correct polarity. If you have a current meter, you can measure 40 milliamps.

   Output can be detected using infrared photosensitive paper or a night vision scope. Certain video cameras may also detect the laser output.

   Danger! Do not view directly or indirectly without safety glasses.
3. Fabricate the laser holder (HOLD1), as shown in Figure 6-4.
4. Fabricate the laser module tube (TUB1) from a piece of 3-inch × 1-inch OD × 1/16-inch wall plastic tubing, as shown in Figure 6-5.

   Fabricate the two spacing washers (WASH1, 2) from 1/16-inch plastic. Note the inner hole is 1 inch to allow for the sleeving onto TUB1 as shown. The outer diameter fits into the inner diameter of the main enclosure (EN1). It is these two spacer washers that position and secure TUB1 into the center of EN1. Space the two washers as shown.
5. Insert the laser module LM1 as shown and wire it up to the circuitry as shown. It is a good idea to leave enough lead length to allow one to pull out the battery holder for battery replacement by removing the rear cap (CAP1). Use a piece of foam to secure the battery holder in place.
6. Verify the proper circuit operation so that when the key switch is activated, the LED emits, indicating laser emission.

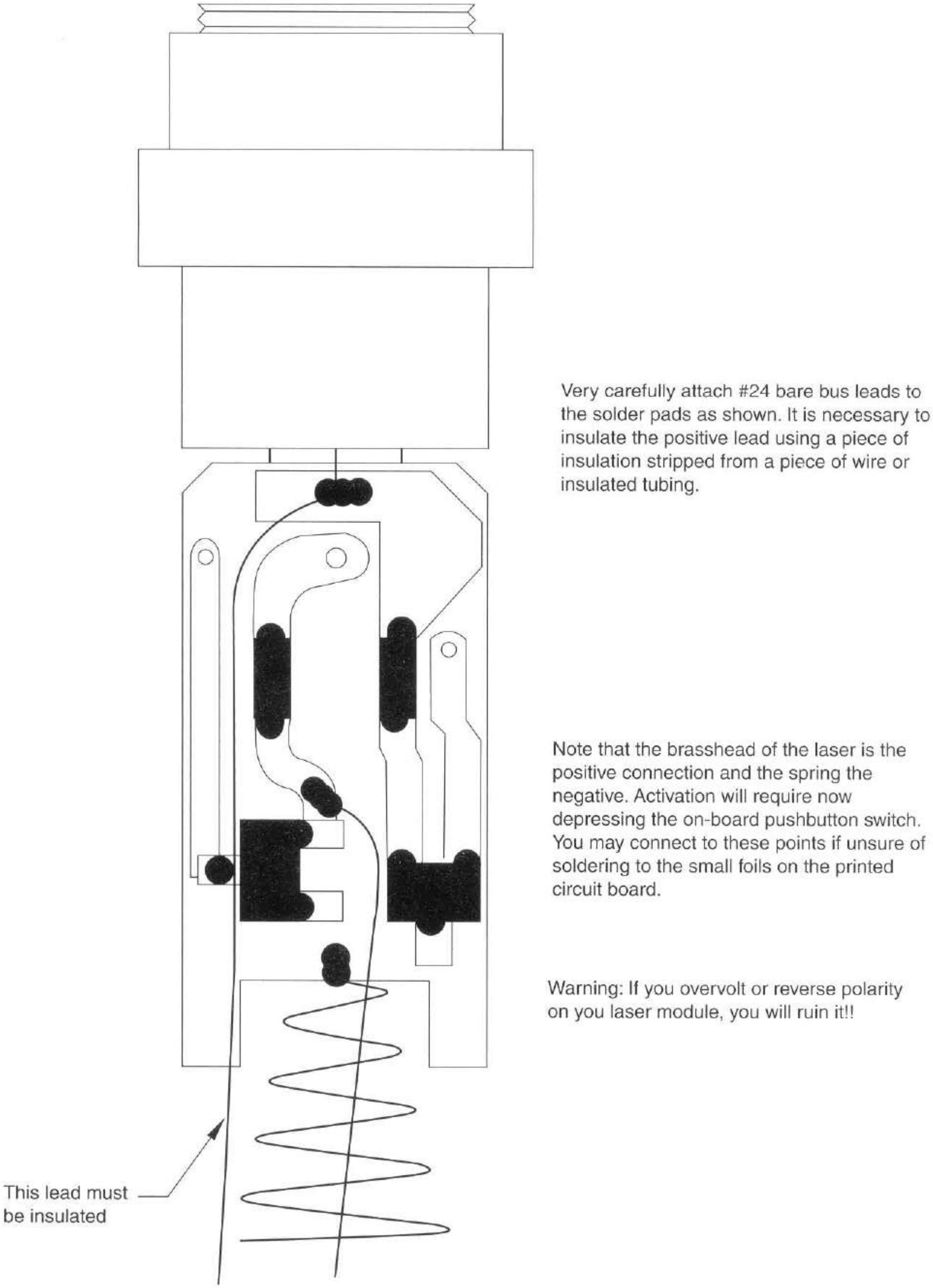

Very carefully attach #24 bare bus leads to the solder pads as shown. It is necessary to insulate the positive lead using a piece of insulation stripped from a piece of wire or insulated tubing.

Note that the brasshead of the laser is the positive connection and the spring the negative. Activation will require now depressing the on-board pushbutton switch. You may connect to these points if unsure of soldering to the small foils on the printed circuit board.

Warning: If you overvolt or reverse polarity on you laser module, you will ruin it!!

This lead must be insulated

Figure 6-3 *Laser module rework*

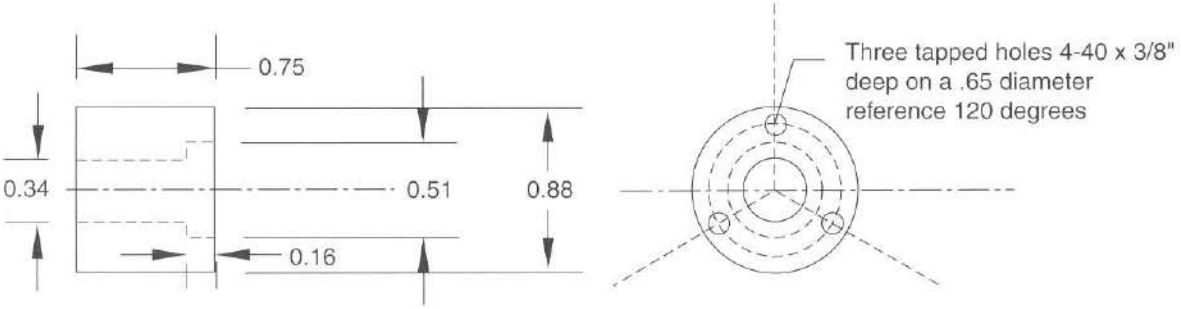
0.75
0.34
0.51
0.88
0.16
Three tapped holes 4-40 x 3/8"
deep on a .65 diameter
reference 120 degrees

HOLD1 Holder material: PVC

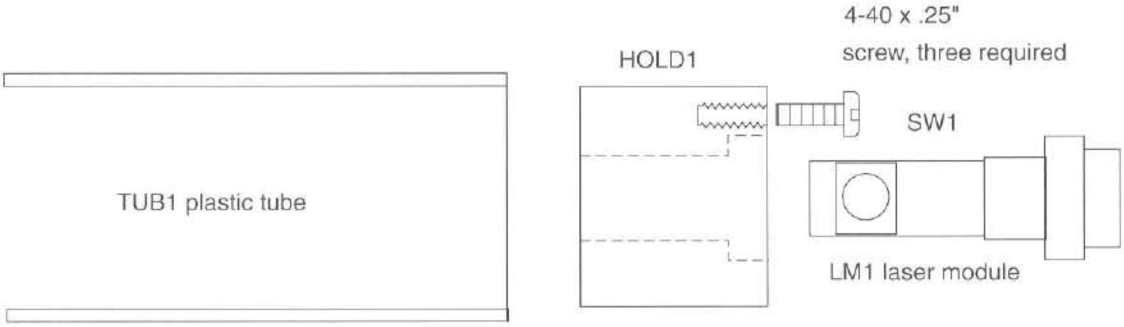
4-40 x .25"
screw, three required
HOLD1
SW1
TUB1 plastic tube
LM1 laser module

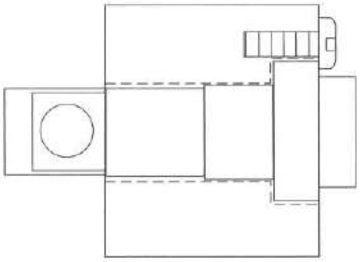

X-ray view showing assembled LM1 laser module inserted into HOLD1 holder with three retaining screws SW1

Figure 6-4 *Creation of laser module holder and retaining washer*

This is our approach to assembling this laser project with external optics. The builder may use his own ingenuity noting that the objective is optical stability and safety.

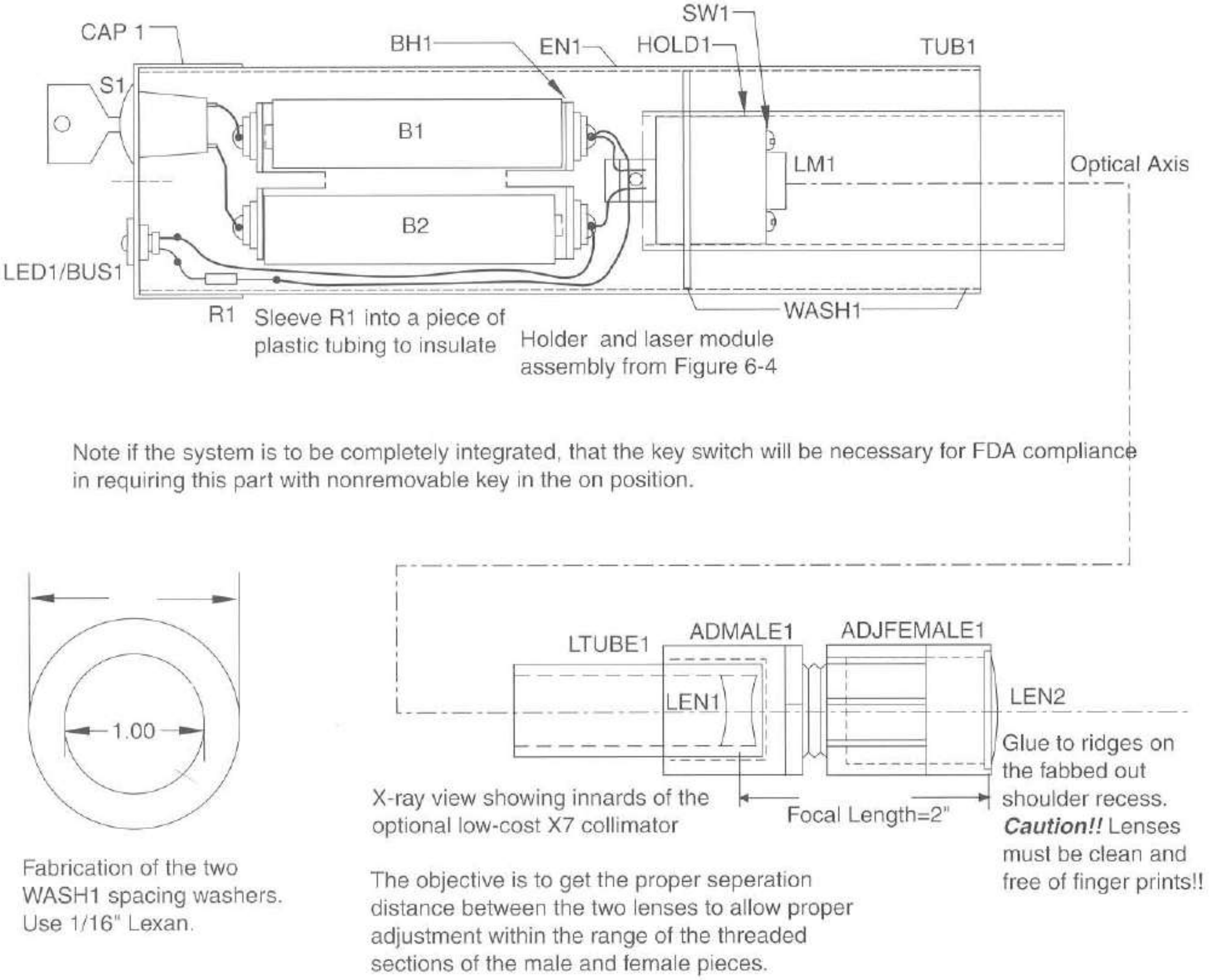

R1 Sleeve R1 into a piece of plastic tubing to insulate

Holder and laser module assembly from Figure 6-4

Note if the system is to be completely integrated, that the key switch will be necessary for FDA compliance in requiring this part with nonremovable key in the on position.

Glue to ridges on the fabbed out shoulder recess. ***Caution!!*** Lenses must be clean and free of finger prints!!

X-ray view showing innards of the optional low-cost X7 collimator

Fabrication of the two WASH1 spacing washers. Use 1/16" Lexan.

The objective is to get the proper seperation distance between the two lenses to allow proper adjustment within the range of the threaded sections of the male and female pieces.

Figure 6-5 *X-ray of complete assembly*

Collimator and Final Assembly

1. Glue the larger lens LEN2 into ADFEMALE1 as shown, using silicon rubber glue or an equivalent. Clean the lens before gluing.
2. Insert LEN1 into the laser output end of the LTUB1 assembly as shown. Do not glue at this time.
3. Fully insert the LTUB1 into the 1-inch tube TUB1. Shim it with tape for a secure but removable fit.
4. Finish the assembling and label it as shown in Figure 6-6.
5. Point the laser at a target around 100 meters away and adjust the collimator to the smallest spot. This should be done in low light, preferably at night, using a night vision device and only looking at the beam impact point. Note

Instructions

1. Turn key switch full ccw. Note switch has several positions–only two are used in this version.
2. Remove CAP1 and insert 2 AA batteries into holder.
3. Turn on key two positions cw and note LED lighting. If not check the battery holder.
4. Adjust coliminator for desired effect.

Note that coliminator will expand the beam at near field but greatly reduce it at far field. Laser module may require beam adjustment over time using the special tool.

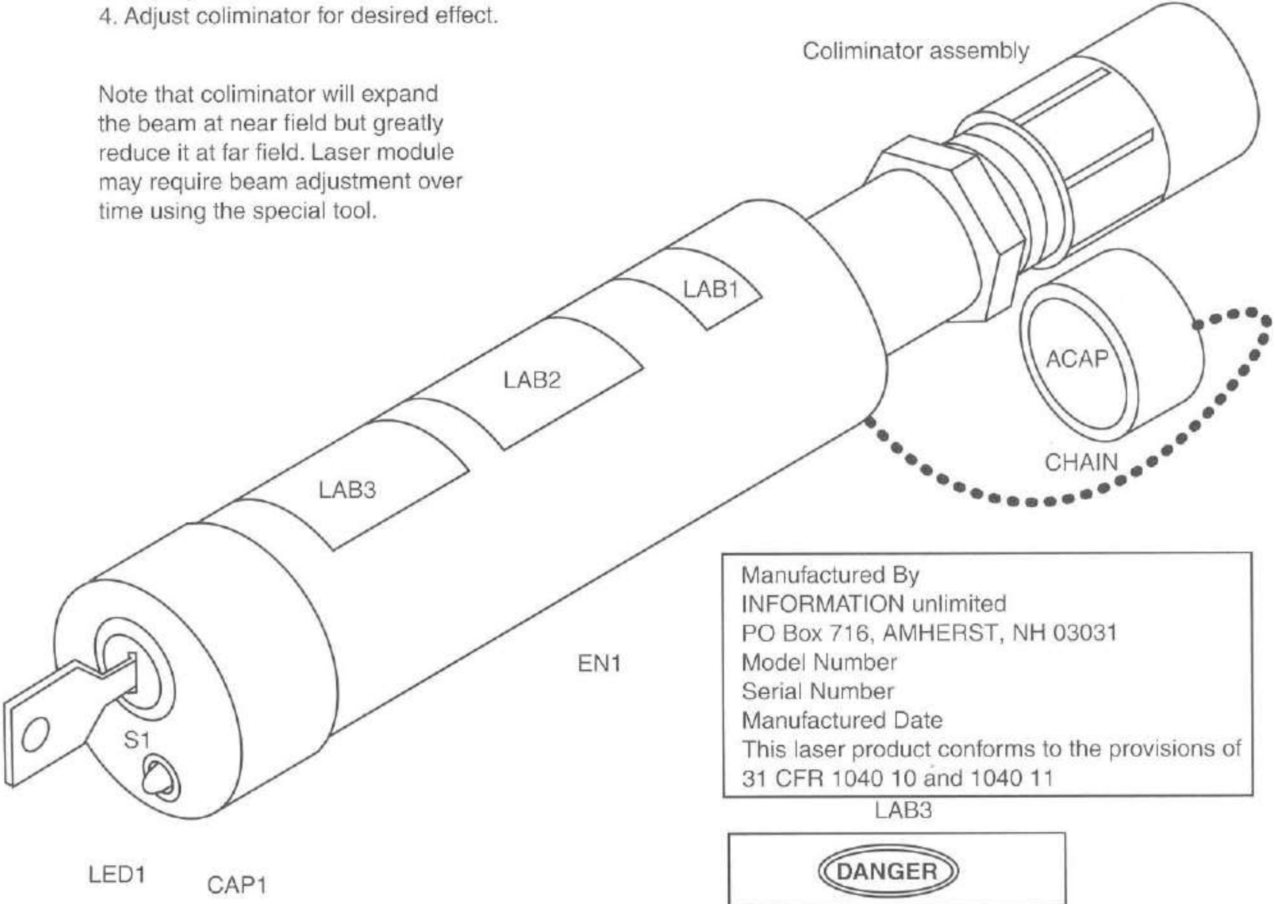

Manufactured By
INFORMATION unlimited
PO Box 716, AMHERST, NH 03031
Model Number
Serial Number
Manufactured Date
This laser product conforms to the provisions of
31 CFR 1040 10 and 1040 11

LAB3

DANGER

VISIBL AND/OR INVISIBLE
LASER RADIATION
AVOID DIRECT EXPOSURE
TO BEAM

30 Milliwatts @ 980 nm
CLASS 3B LASER PRODUCT

LAB2

AVOID EXPOSURE

INVISIBLE LASER RADIATION IS
EMITTED FROM THIS APERTURE

LAB1

Compliance test

1. Verify correct labels as shown
2. Key switch–nonremoval in off position
3. Beam indicator LED indicates after delay
4. Aperture cap included

WARNING: Do not under any circumstances point this device at people, vehicles, or especially aircraft. YOU WILL BE PROSECUTED IF CAUGHT!!!

Figure 6-6 *Final assembly view*

that the adjustment should be about 1/2 to 3/4 turns from being fully tight for a mechanically stable operation. If it is too far out, the assembly will be loose. You may have to use a wire brush on the threads and lube with dry teflon for smooth action. You may compensate the distance between lenses by changing the length of TUBE1. You may now glue LEN1 in place as noted.

The collimator will expand the beam width but decrease the divergence by the same ratio. This greatly reduces the far-field beam diameter.

Note that the laser module is at a wavelength of 980 nanometers and is invisible to the eye. Also, note that most of the mechanical parts used in this project can be obtained through most hardware stores.

Table 6-1 Infrared Laser Parts List

Ref. #	*Description*	*DB Part #*
R1	220-ohms, ¼-watt resistor (red-red-br)	
LED1	High-brightness green LED	
LM1	30-milliwatt, 980-nanometer infrared laser module	DB# LM980-25
S1	Key switch with nonremovable key in "on" position	
BH1	Dual, side-by-side two-AA battery holder	
WR24	12-inch #24 bus wire	
WR22	24-inch #22 vinyl hookup wire	
HOLD1	Fabricated holder with washer for laser module	DB# HOLD1
SWI/NU1	Two 4-40, 1.25-inch machine screws and nuts	
TUB1	3 × 1-inch OD × .875-inch plastic or metal tube	
WASH1	Two fabricated plastic spacing washers, 1.5 OD with 1-inch center hole	
EN1	6-inch × 1.625 OD × 1.5-inch ID plastic or metal tube	
CAP1	1⅝-inch plastic slip on cap	
Parts for low-cost collimator		
LEN1	6 mm × (-1) double-concave negative glass lens	DB# LENS13
LEN2	24 mm × 75 mm plano/convex glass lens	DB# LE2475
*LTUB1	625-inch ID × 2-inch schedule 40 PVC tube	
*ADMALE1	½-inch PVC schedule 40 slip fit to male thread, GENOVA 30405	
*ADFEMALE1	½-inch PVC schedule 40 slip fit to female threads, GENOVA 30305	
*ACAP1	Plastic cap and chain for aperture cap	
LAB1	Aperture label	
LAB2	Classification label	
LAB3	Certification label	

*Available in most hardware stores

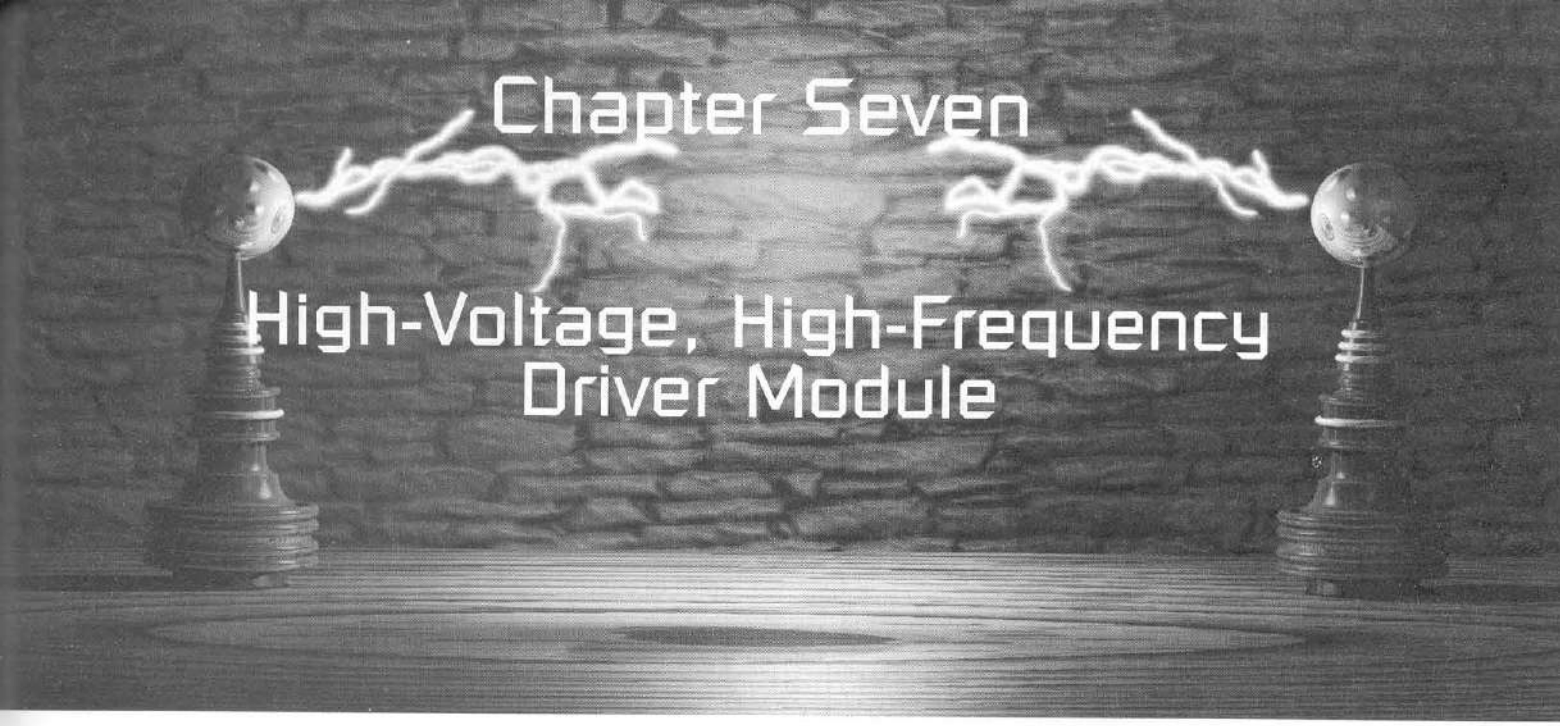

Chapter Seven

High-Voltage, High-Frequency Driver Module

Figure 7-1 shows that a universal, high-voltage, modular power supply project will provide you with many options. This "sweetheart" of a circuit has features that make it a priceless tool for the high-voltage researcher and experimenter.

The circuit operates on 11 to 15 volts DC, drawing 3 amps under a full load, allowing a portable battery or 115 volts AC via a converter. The output voltage is a 60 KHz, high-frequency current that is fully short-circuit protected. The high-frequency output also makes it possible to use low-energy voltage multiplier stacks for high-voltage DC sources, and it can also serve as an excellent plasma driver when used directly. The output current is fully adjustable via a control pot.

The unit is excellent for powering neon and all types of gas-filled vessels using one or two electrodes, or it can power objects simply by proximity. It easily retrofits to our voltage multipliers' modules that provide DC voltages up to 100 kilovolts and currents of up to .3 milliamperes. The current-limiting and control features make this combination an excellent choice for charging capacitors for low loss charging, utilizing 12-volt portable or 115 vac line operation. Also, the device is an excellent choice for powering large and small antigravity craft, ozone air purification, and other applications requiring a high-voltage current-controlled source. The module is shown built on both a rugged *printed circuit board* (PCB) and a vector perforated box. It is mounted in a plastic channel.

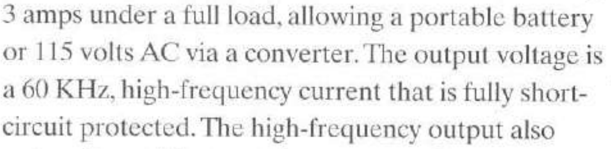

Figure 7-1 *Completed assembly module*

The module is used for several projects in this book. In order not to waste a lot of space, we show it as a common subassembly for the following chapters:

- Chapter 12, "Electro-magnetic (EMP) Gun"
- Chapter 32, "100-Milligram Ozone Machine"

Electrical and Mechanical Specifications

- Open-circuit voltage, 7,500 peak at 60 KHz
- Short-circuit current, 10 milliamperes short-circuit protected
- Input 11 to 15 volts DC at 3 amps fully loaded

- Adjustable current by duty-cycle-controlled pulse
- Compact size, $7 \times 2\frac{1}{8} \times 1\frac{1}{8}$ inches, weighing less than 5 ounces
- Easily retrofits to our voltage multipliers

This is an intermediate-level project requiring basic electronic skills and should cost between \$25 and \$50. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 7-1.

High-voltage DC output is obtained from this module using a Cockcroft-Walton voltage multiplier with multiple stages of multiplication as required. Note that this method of obtaining high voltages was used in the first atom smasher that ushered in the nuclear age.

The multiplier section requires a high-voltage AC source for input supplied by the circuit transformer (T1) producing 6 to 8 kilovolts at approximately 60 KHz. You will note that this transformer is of a unique design, being owned by our company Information Unlimited. The part is very small, versatile, and lightweight for the power produced.

Circuit Description

The primary of T1 is current driven through inductor L1 and switched at the desired frequency by FET switch Q1. Capacitor C6 is resonated with the primary of T1 and zero voltage switches when the frequency is properly adjusted. (This mode of operation is very similar to class E operation.) The timing of the drive pulses to Q1 is therefore critical to obtain optimum operation (see Figure 7-2).

The drive pulses are generated by a 555 timer circuit (I1) connected as an astable multivibrator with a rep rate determined by the setting of trimpot R1 and the fixed-value timing capacitor C2.

The timer circuit is now turned on and off by a second timer, I2. This timer operates at a fixed frequency of 100 Hz but has an adjustable duty cycle (a ratio of on to off time) determined by the setting of control pot R10. I1 is now gated on and off with this controlled pulse now providing an adjustment of output power.

When the unit is interfaced with a DC voltage multiplier, an over-voltage protection spark gap is placed across the output and is easily set to break down at a preset voltage level for circuit protection. Even though the output is short-circuit protected against continuous overload, constant hard discharging of the output can cause damage and must be limited. A pulse-current-limiting resistor (R17) helps to protect the unit from these catastrophic current spikes.

Power input is controlled by switch S1, which is part of control pot R10. Actual power can be a small battery- capable of supplying up to 2 amps or a 12-volt, 2 to 3-amp converter for 115-volt operation.

Construction

The circuit is shown using the more challenging *perforated circuit board* often required for a science fair project. A *printed circuit board* is also individually available, requiring that you identify only the particular part and insert it into the respective holes as noted. The PCB is plainly marked with the part identification, and soldering is now very simple as you solder the component leads to the conductive metal traces on the underside of the board.

The perforated board approach is more challenging, as now the component leads must be routed and used as the conductive metal traces. We suggest that the builder closely follow the figures in this section and mark the actual holes with a pen before inserting the parts. Start from a corner, using it as a reference, and proceed from left to right. Note that the perforated board is the preferred approach for science projects, as the system looks more homemade.

Board Assembly Steps

If you are a beginner it is suggested to obtain our *GCAT1 General Construction Practices and Techniques*. This informative literature explains basic practices that are necessary in proper construction of electromechanical kits and is listed in Table 7-1.

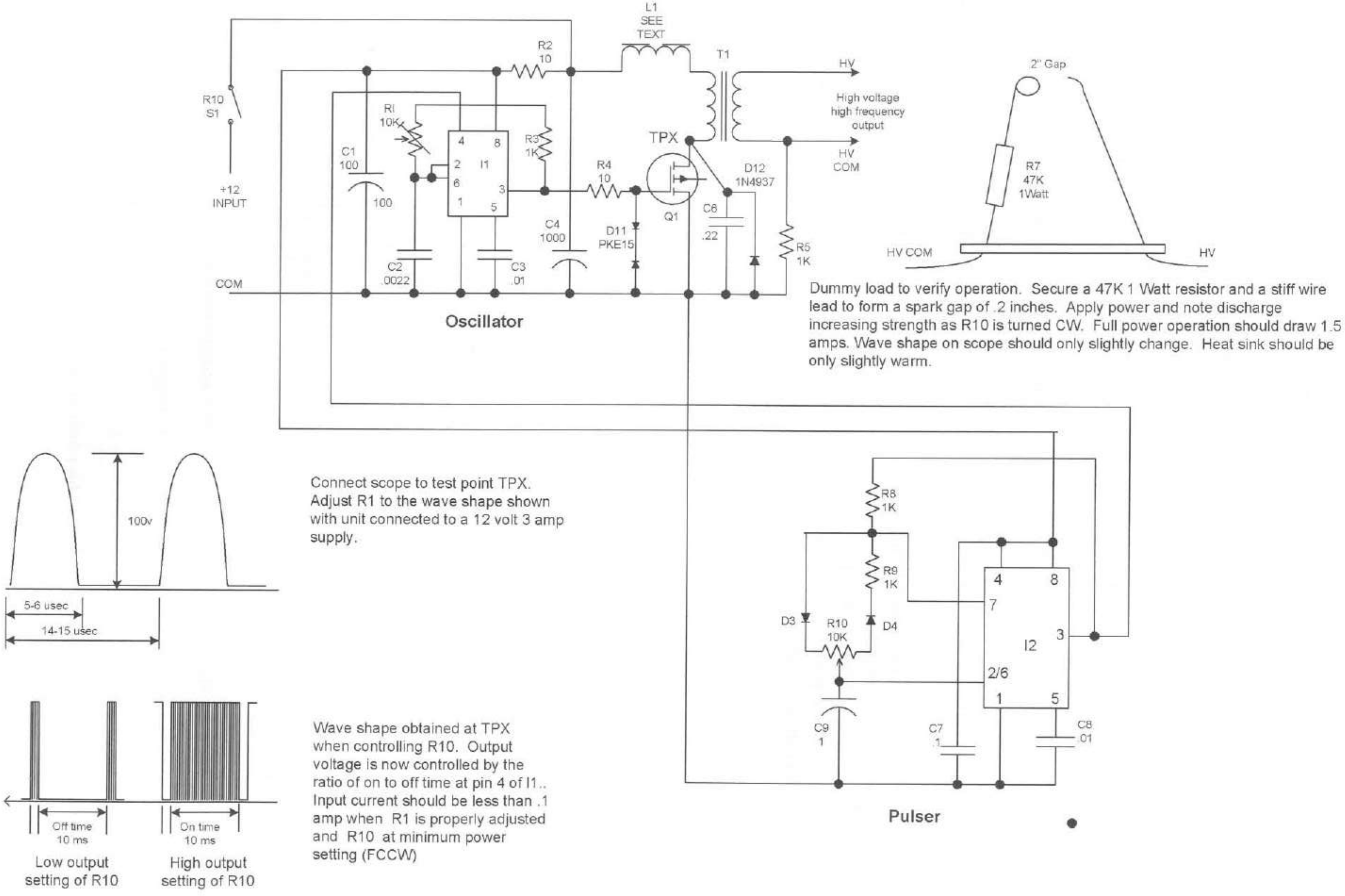
L1
SEE
TEXT
T1
R2
10
R10
S1
+12
INPUT
COM
C1
100
100
RI
10K
C2
.0022
4
8
2
I1
6
3
1
5
C3
.01
R3
1K
R4
10
C4
1000
D11
PKE15
TPX
Q1
C6
.22
D12
1N4937
R5
1K
HV
High voltage
high frequency
output
HV
COM
Oscillator
2" Gap
R7
47K
1Watt
HV COM
HV
R8
1K
R9
1K
D3
R10
10K
D4
C9
1
C7
.1
C8
.01
I2
4
8
7
3
2/6
1
5
Pulser
100v
5-6 usec
14-15 usec
Off time
10 ms
On time
10 ms
Low output
setting of R10
High output
setting of R10

Dummy load to verify operation. Secure a 47K 1 Watt resistor and a stiff wire lead to form a spark gap of .2 inches. Apply power and note discharge increasing strength as R10 is turned CW. Full power operation should draw 1.5 amps. Wave shape on scope should only slightly change. Heat sink should be only slightly warm.

Connect scope to test point TPX. Adjust R1 to the wave shape shown with unit connected to a 12 volt 3 amp supply.

Wave shape obtained at TPX when controlling R10. Output voltage is now controlled by the ratio of on to off time at pin 4 of I1.. Input current should be less than .1 amp when R1 is properly adjusted and R10 at minimum power setting (FCCW)

Figure 7-2 *Circuit schematic*

1. Lay out and identify all the parts and pieces. Verify them with the parts list, and separate the resistors, as they have a color code to determine their value. Colors are noted on the parts list.
2. Cut a piece of .1-inch grid perforated board to a size of 5 × 3 inches. Locate and drill the holes, as shown in Figure 7-3. An optional PCB is individually available (refer to Table 7-1).
3. Fabricate a metal heatsink for Q1 from a piece of .063 aluminum, at 1.5 × .75 inches, as shown in the inset of Figure 7-4.
4. Assemble L1 as shown in the Figure 7-4 inset. This part is individually available and listed in Table 7-1
5. If you are building from a perforated board, it is suggested that you insert components starting at the lower left-hand corner, as in Figure 7-5. Pay attention to the polarity of the capacitors with polarity signs and all semiconductors. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing, as these indicate connection runs on the underside of the assembly board. The heavy dashed lines indicate the use of thicker #20 bus wire, as this is a high-current discharge path.

   Figures 7-6 and 7-7 show the available PCB, which is also listed as an available part in Table 7-1.
6. Attach external leads as shown in Figure 7-5.
7. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.

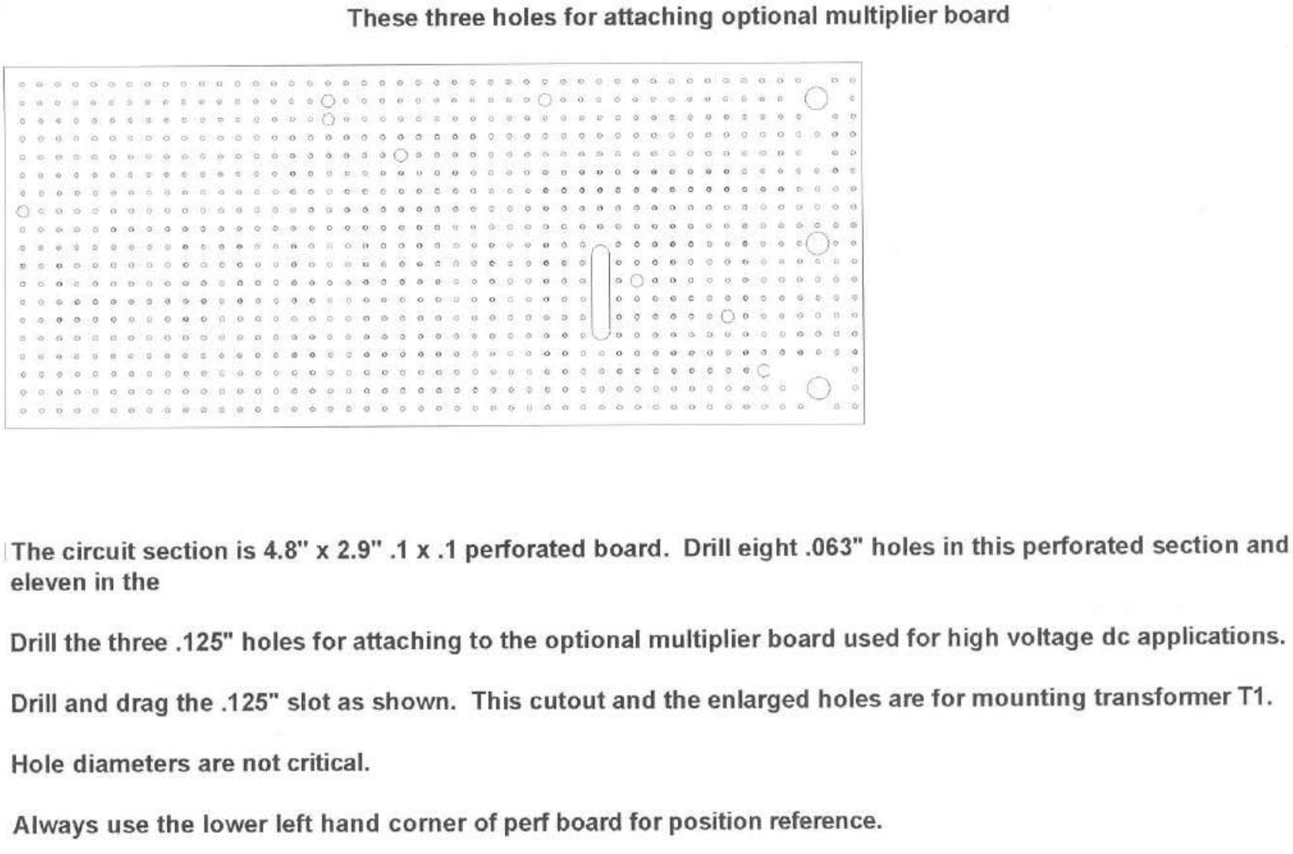
These three holes for attaching optional multiplier board
The circuit section is 4.8" x 2.9" .1 x .1 perforated board. Drill eight .063" holes in this perforated section and eleven in the
Drill the three .125" holes for attaching to the optional multiplier board used for high voltage dc applications.
Drill and drag the .125" slot as shown. This cutout and the enlarged holes are for mounting transformer T1.
Hole diameters are not critical.
Always use the lower left hand corner of perf board for position reference.

Figure 7-3 *Perforated board*

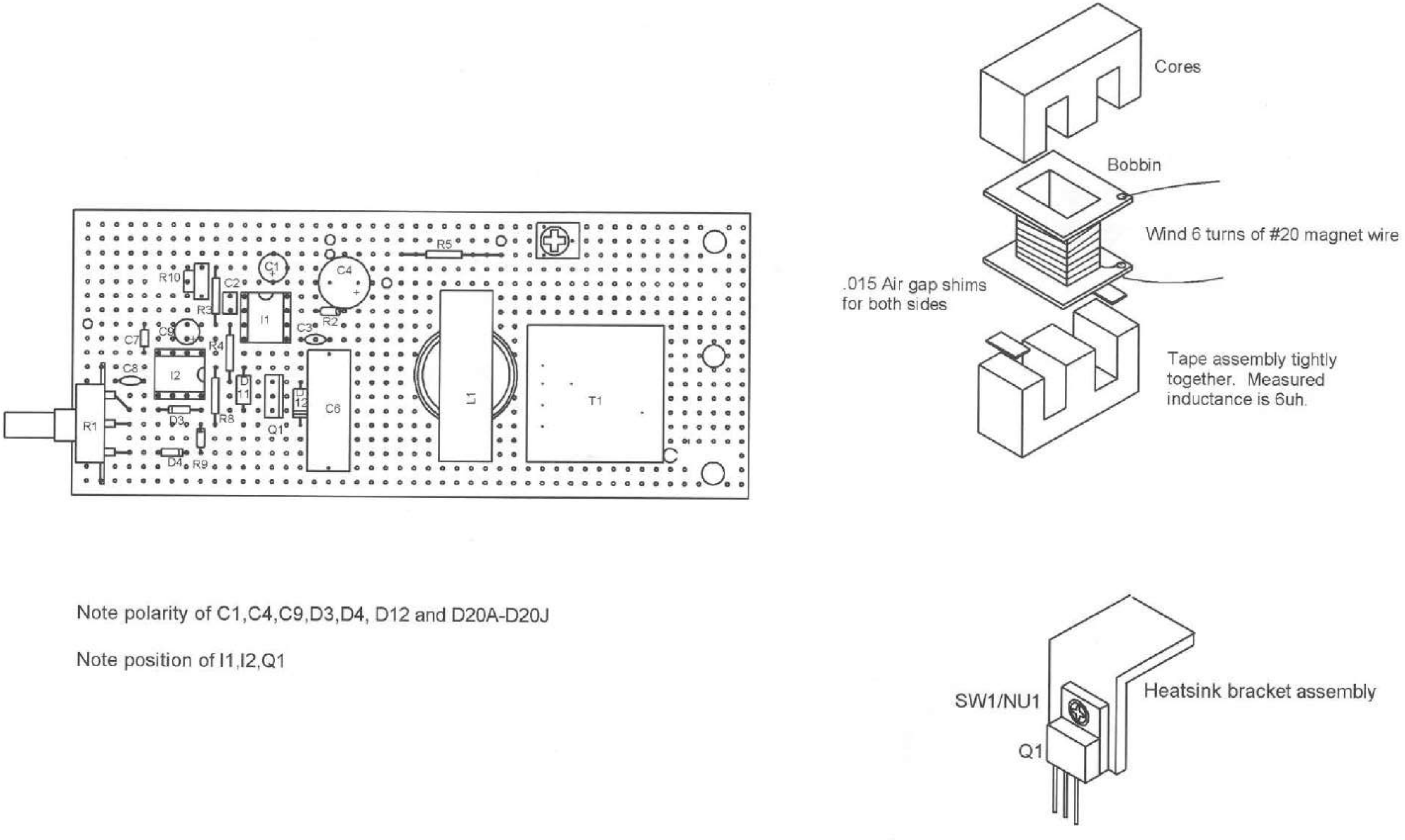
R1
C8
C7
I2
D3
D4
R9
C9
R10
R3
R4
R8
C2
C1
I1
D11
Q1
C3
R2
C4
D12
C6
R5
L1
T1
Cores
Bobbin
Wind 6 turns of #20 magnet wire
.015 Air gap shims
for both sides
SW1/NU1
Q1
Heatsink bracket assembly

Tape assembly tightly together. Measured inductance is 6uh.

Note polarity of C1,C4,C9,D3,D4, D12 and D20A-D20J

Note position of I1,I2,Q1

HSINK Bracket fabricated as per step 3 from 1/16" aluminium piece. Note hole for attaching tab of Q1

Figure 7-4 *Parts assembly and fabrication*

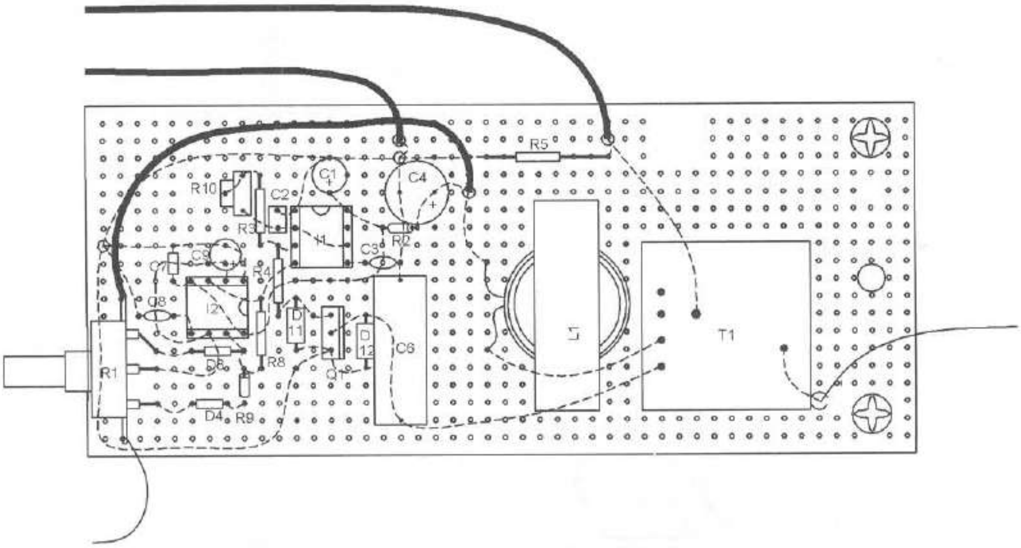

- - - - - Thick dashed lines are direct connection runs beneath board of #20 buss wire (WR20BUSS) and are extended for the spark switch electrodes.

Thinner dashed lines are #24 buss wire (WR24BUSS) and component leads wherever possible.

◁ Triangles are direct connection point junctions.

— Solid black lines are external leads for input and output lines. Use red (WR20R) for +12 input. Use green (WR20G) for lifter - connection. Use black (WR20B) for com -12 input

Use 1/8-3/16 wide smooth globular solder joints for connections to C20A-J and D20A-J, R7 and HV output points. Thisis contrary to normal soldering but is necessary to prevent corona leakage.

Figure 7-5 *Wiring connections and external leads*

8. Fabricate a channel from a piece of 1/16-inch plastic material. Place it in the assembly and secure it at the corners using silicon rubber adhesive. You may also enclose it in a suitable plastic box.

   This step may not be required when used for another project referenced in this book.

Testing Steps

This step will require basic electronic laboratory equipment including a 60 MHz oscilloscope.

1. Preset the trimpot R1 to midrange and R10 to full counterclockwise. Short out the output leads using a short clip lead.
2. Obtain a 47K, 1-watt resistor and construct the spark gap "dummy load," as shown in Figure 7-2.
3. Obtain a 12-volt DC, 3-amp power converter* or a 12-volt, 4-amp rechargeable battery.*

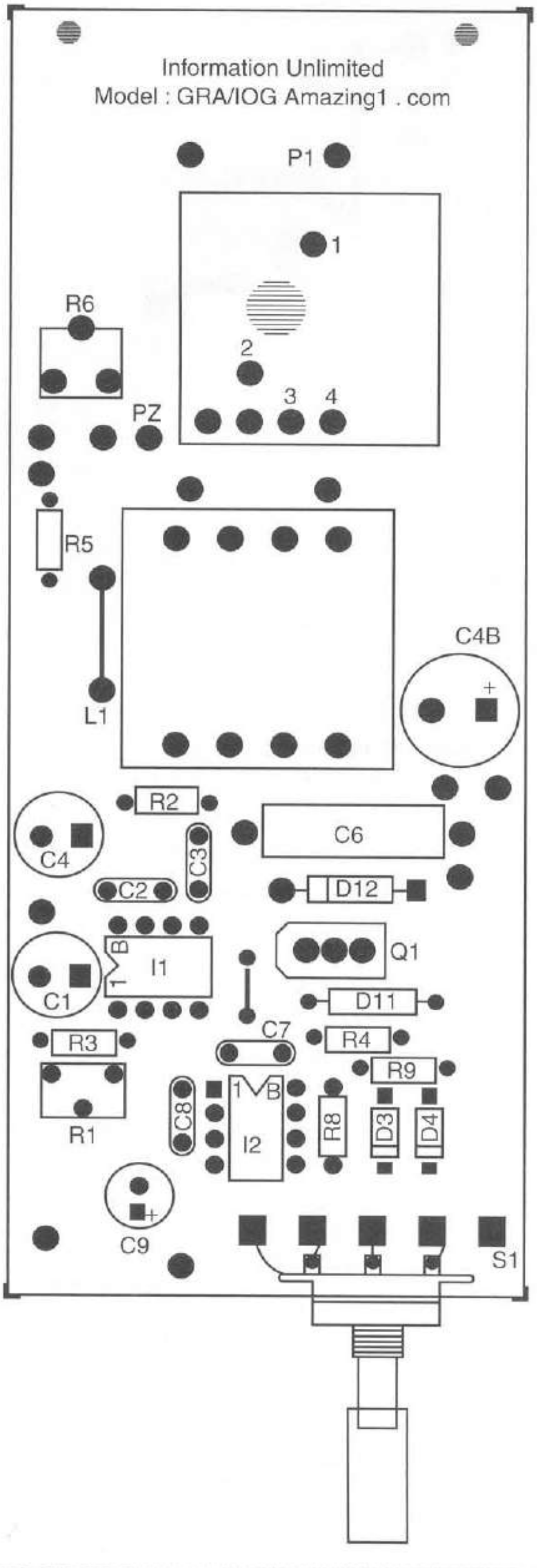

Figure 7-6 *PCB part identification*

4. Connect the input to the power converter and connect the scope, set to read 100 volts and with a sweep time of 5 usecs, to the drain of Q1.

5. Apply power and quickly adjust R1 to the waveshape shown in the inset of Figure 7-2. Note an input current dip of 100 milliamps or less. Check the heatsink of Q1, noting only warm to touch or slightly above 110 degree F.

   Note our laboratory-built units usually tune in with the pot set to 11 o'clock.

6. Remove the clip lead short in step 9 and connect the dummy load. Apply power and rotate R10 clockwise, noting the input current smoothly increasing to 1.5 amps and the spark increasing in energy. This control varies the ratio of off to on time and nicely controls the system current. Note that this system can easily provide 30 watts of usable power.

Table 7-1 High-Voltage, High-Frequency Driver Module Parts List

Ref. #	*Description*	*DB Part #*
R1	10K vertical trimpot	
R2, R4	Two 10-ohm, 1/4-watt resistor (br-blk-blk)	
R3, 5, 8, 9	Four 1K, 1/4-watt resistor (br-blk-red)	
R7	47K, 1-watt resistor (yel-pur-or)	
R10/S1	10K, 17 millimeter pot and switch	
C1	100-microfarad, 25-volt vertical electroradial leads	
C2	.0022-microfarad, 50-volt green plastic cap (222)	
C3,8	Two .01-microfarad, 50-volt disks (103)	
C4	1,000-microfarad, 25-volt vertical electrolytic capacitor	
C6	.22-microfarad, 250-volt metallized polypropylene	
C9	1-microfarad, 25-volt vertical electro cap	
C7	.1-microfarad, 50-volt cap, INFO#VG22	
D3, 4	Two IN914 silicon diodes	
D11	One PKE15, 15-volt transient suppressor	

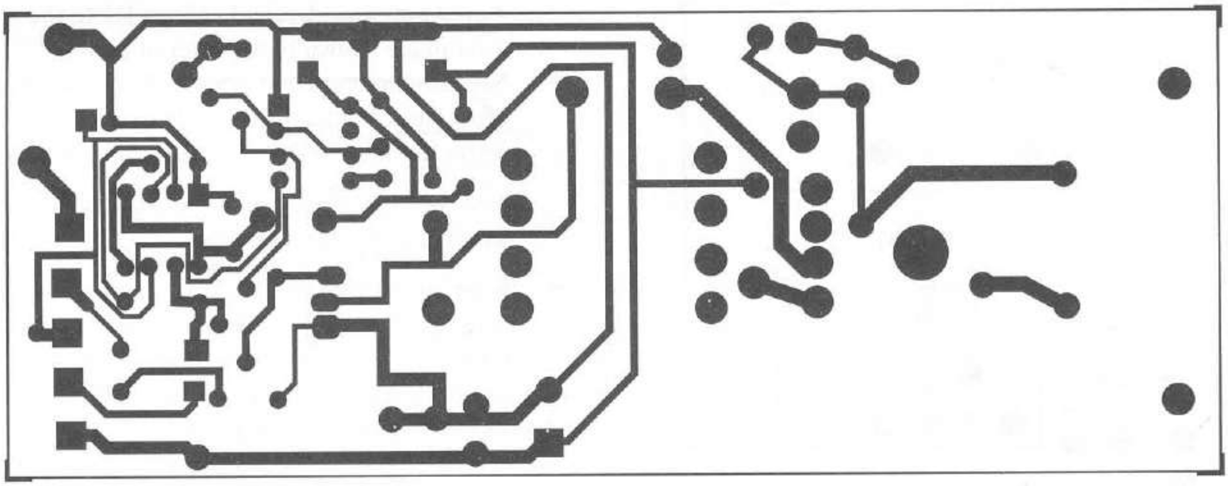

Figure 7-7 *PCB foil traces*

D12	1N4937 fast-switching 1-kilovolt diode	
Q1	IRF540 *metal-oxide-semiconductor field effect transistor* (MOSFET) transistor TO220	
I1, 2	Two 555 DIP timers	
T1	7-kilovolt, 10-milliamp mini-switching transformer	DB# IU28K089
L1	6 uH inductors (see text on assembly)	DB# IU6UH
PBOARD	5 × 2.8 × .1-inch grid perforated board; build to size per Figure 7-2	
PCGRA	Optional PCB	DB# PCGRA
WR20R	12-inch #20 vinyl red wire for positive input	
WR20B	12-inch #20 vinyl black wire for negative input	
WR20G	12-inch #20 vinyl green wire for output ground to craft return	
WR20KV	Four 20 kilovolt silicon high-voltage wires for output	
WR20BUSS	18 inches of #20 bus wire for spark gap and heavy leads (see Figure 7-5)	
WR24BUSS	12 inches of #24 bus wire for light leads (see Figure 1-5)	
HSINK	1.5 × 1-inch .063 AL plate created as per Figure 7-4	
SW1	One 6-32½ Philips screw	
NUT	Four 6-32 keep nuts	
12DC/7	115 VAC to 12 VDC, 3-amp converter	DB# 12DC/7
BAT12	12-volt, 4-amp, hour rechargeable battery	DB# BAT12
BC12K	Battery charger kit for above BAT12	DB# BC12K
GCAT1	*GENERAL CONSTRUCTION PRACTICES*	DB# GCAT1

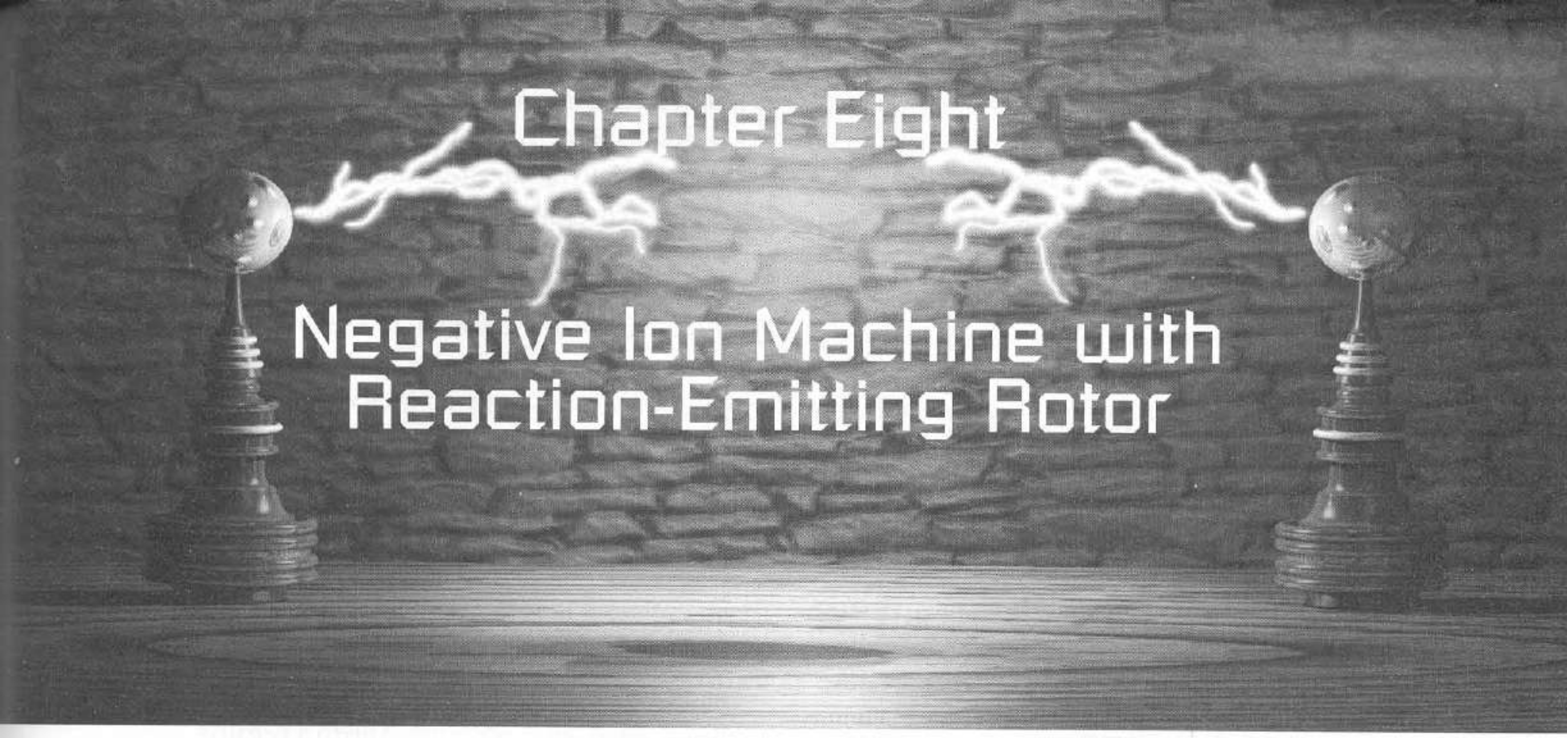

Chapter Eight

Negative Ion Machine with Reaction-Emitting Rotor

Figure 8-1 shows a new concept in negative ion generation. A spinning rotor powered by the emitted ions themselves generates copious amounts of beneficial ions. Large volumes of ions can now be produced, as the area circumscribed by the rotor greatly reduces the local charge flux. This feature allows many times more ions without ozone-producing voltages.

Negative ions are highly beneficial and cause a sense of well-being, as well as rid the air of pollen and other breathable irritants. This is an excellent device for people suffering from asthma and other bronchial problems. It is also a nice conversation piece with a desktop unit's spinning rotor silently reaching a relatively high rotational velocity. The device operates from 115 VAC/12 VDC or a battery. The project utilizes a 25-kilovolts negative output power module requiring 12 volts DC at 500 milliamps along with our proprietary rotating emitter.

This is a beginner- to intermediate-level project requiring basic electronic skills, and one should expect to spend $25 to $50. All the parts are readily available, with specialized parts obtainable from Information Unlimited (www.amazing1.com), and they are listed in Table 8-1.

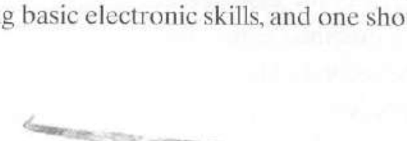

Figure 8-1 *The ion machine*

Precautions: Do not use this device in explosive or flammable atmospheres, as spark discharges may cause dangerous ignitions.

Benefits of Negative Ions

In the last several decades, a medical controversy has evolved pertaining to the beneficial effects of these minute electrical particles. As with any device that appears to affect people in a beneficial sense, there are those who sensationalize and exaggerate these claims as a cure for all aliments and ills. Such people manufacture and market these devices under these false pretenses and consequently give the products an adverse name. The Food and Drug Administration now steps in on these claims, and the product, along with its beneficial facets, goes down the tubes.

People are affected by the property of these particles, which increase the activity rate of cilia, whose function is to keep the trachea clean from foreign objects, thus enhancing oxygen intake and increasing the flow of mucus. This property neutralizes the effects of cigarette smoking, which slows down this activity of the cilia. Hay fever and bronchial asthma victims are greatly relieved by these particles, and burn and surgery patients are relieved of pain and heal faster.

Tiredness, lethargy, and general worn-out feelings are replaced by a sense of well-being and renewed energy. Negative ions destroy bacteria and purify the air with a country-air freshness. They also cheer people up by decreasing the serotonin content of the blood. As can be seen in countless articles and technical writings, negative ions are a benefit to humans and their environment.

Negative ions occur naturally from static electricity, the wind, waterfalls, crashing surf, cosmic radiation, radioactivity, and ultraviolet radiation. Positive ions are also produced from some of these phenomenon and both of them usually cancel each other out as a natural statistical occurrence. However, many manmade objects and devices have a tendency to neutralize the negative ions, thus leaving an abundance of positive ions, which create sluggishness and most of the opposite physiological effects of its negative counterpart.

One method for producing negative ions is by obtaining a radioactive source rich in beta radiation (negative electrons). Alpha and gamma emissions from this source produce positive ions that are neutralized electrically. The resulting negative ions are directed by electrostatic forces to the output exit of the device and are further dispersed by the action of a fan (this method has recently come under attack by the Bureau of Radiological Health and Welfare for the use of tritium or other radioactive salts). This approach appears to be the more hazardous of the two according to the Consumer Product Safety people. The second method is to produce a potential level of negative high-voltage electricity without ozone and allow the creation of negatively charged particles by the high voltage.

Your negative ion generator produces ions via a high-potential, low-current source of DC power. The high potential causes negative charges to be produced at the sharp emission points of the rotor blade. The reactionary force of the emitted particles now causes the rotor to spin, allowing these negative ions to escape into the air stream due to the high-charge density at the sharp, pointed rotor tips. (The point of a pin will have a much higher charge density than a larger diameter spheroid for the same potential.) Ozone is kept at a minimum by keeping the voltage relatively low while allowing a sufficient charge density lets the negative particles escape into the air.

Circuit Description

The circuit (MOD25KV) consists of a high-frequency, high-voltage oscillator being fed into a multistage Cockcroft-Walton multiplier. This energy is converted into a potential of up to a negative 25,000 volts. The high-frequency stage consists of a transistor connected as a simple oscillator where its collector drives the primary winding of the resonant transformer (T1).

The high-frequency output of T1 is fed into a voltage multiplier stage consisting of a diode string (D1 to D5) and a capacitor string (C1 to C5). A 12-volt DC wall adapter (T2) powers the unit. A separate ground is required for optimum performance and involves connecting the negative 12 volts to earth ground. This is done by connection to the AC receptacle plate or ground pin. This ground provides a virtual ground for the ions.

A charge concentration occurring at the ends of the rotor blade's points now produces ions, and reactionary forces move the rotor at a high speed. Ions are produced at a high rate as the effective emitter area is equal to that of the rotating rotor. Also, charge neutralization is minimal using this method and a high ion current is generated.

The output of the unit can be approximated by the fact that a Coulomb charge equals the current multiplied by the seconds. There are also 6.25 × E18 negative chargers per Coulomb and the maximum current to the emitters is 400 microamps. Therefore, the unit can produce 6.25 × E18 × 200 × E-6 = 2.5 × 10E15

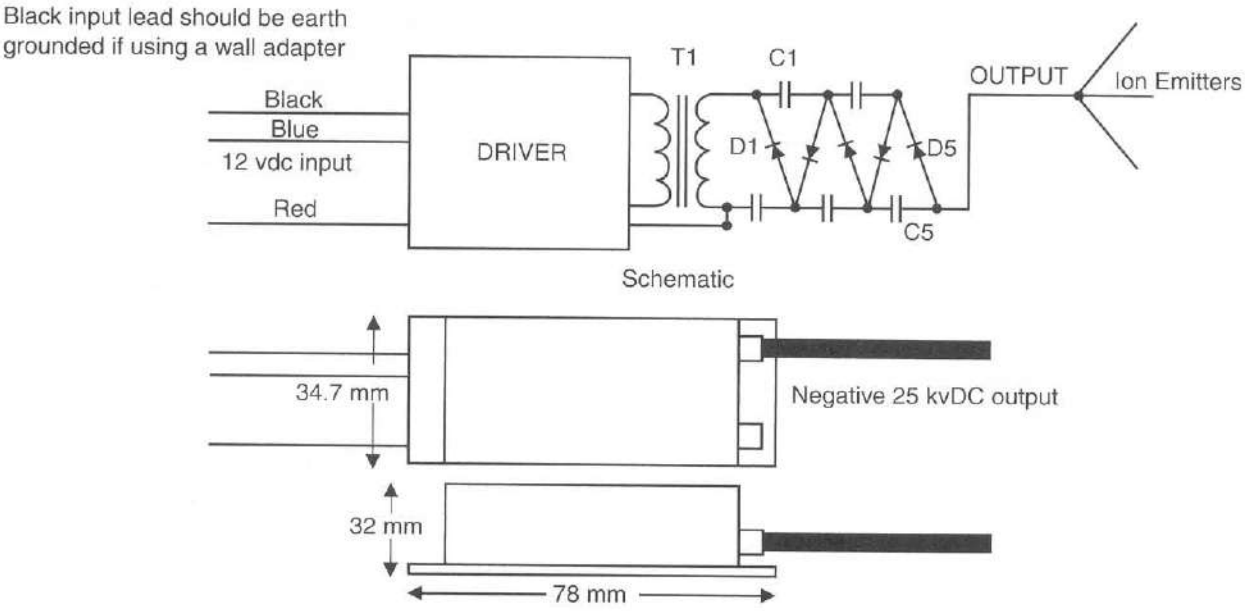
Black input lead should be earth grounded if using a wall adapter

Function	Value	Notes	Other
Input Voltage	12 vdc		
Input Current (No load)			
Input Current (load)	400 ma		
Input Watt (No load)			
Input Watt (load)			
Output Volts	–20 kvDC		
Output Configuration	Single Output		
Internal Operating Frequency	35 KHz		

Use a 12 vdc/.5 amp DC wall adapter or other suitable power supply.
Always check entire system for any excessive heating.

Figure 8-2 *Circuit schematic*

negative charges per second. This is usually sufficient for a large room.

Assembly Steps

1. Lay out and identify all parts and pieces. Verify them with the parts list.
2. Carefully fabricate the rotor blade as shown in Figure 8-3 and note that it must be balanced for optimum rotational velocity and ion emission.
3. Position a common plastic pushpin (PIN1) and hot glue or epoxy to the module as shown. Glue is more forgiving if should you ever wish to replace the pin.
4. Carefully strip about $^{3}/_{4}$ of an inch of insulation from the output lead. Tin and solder it to the bottom of the actual pin. This connection point must allow complete clearance of the entire bearing section of the rotor assembly when in place.
5. Wire up the wall adapter (T2) and ground the lead using (WN1) wire nuts, as shown in Figure 8-4.

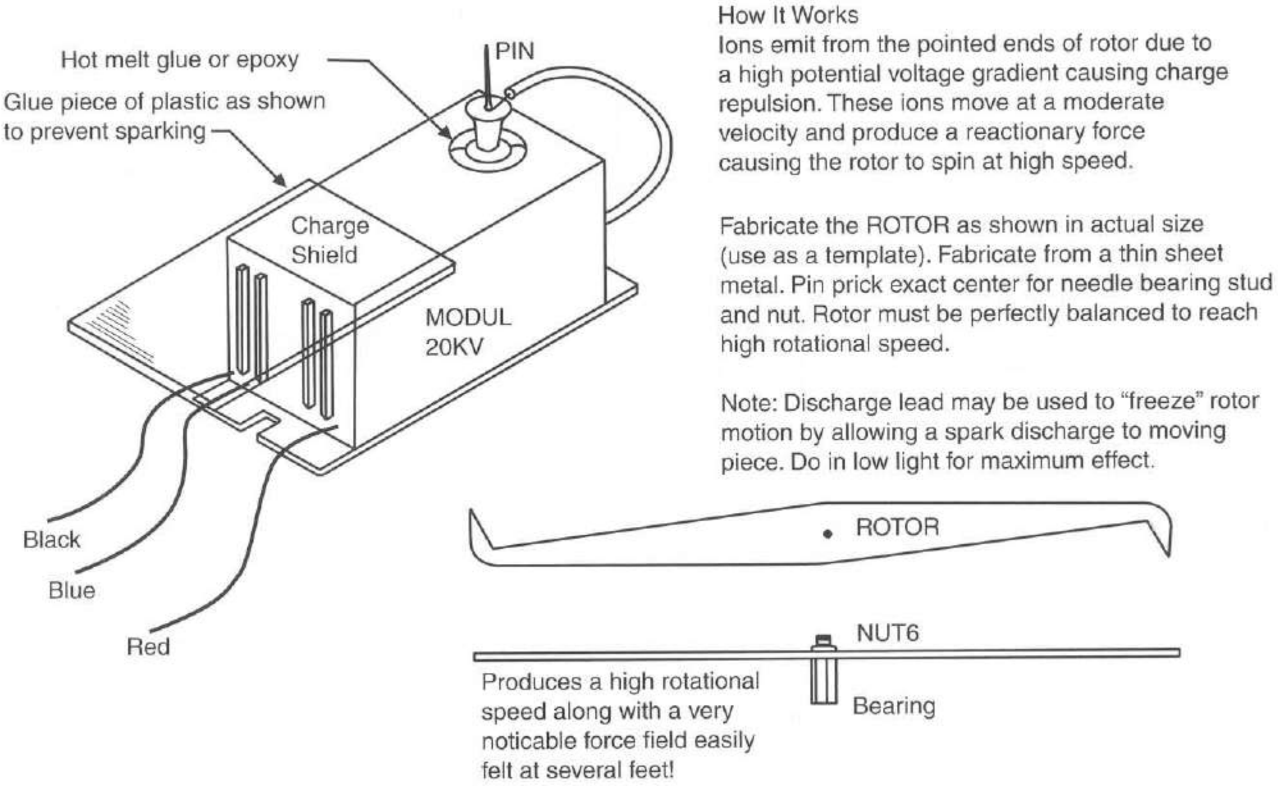

How It Works
Ions emit from the pointed ends of rotor due to a high potential voltage gradient causing charge repulsion. These ions move at a moderate velocity and produce a reactionary force causing the rotor to spin at high speed.

Fabricate the ROTOR as shown in actual size (use as a template). Fabricate from a thin sheet metal. Pin prick exact center for needle bearing stud and nut. Rotor must be perfectly balanced to reach high rotational speed.

Note: Discharge lead may be used to "freeze" rotor motion by allowing a spark discharge to moving piece. Do in low light for maximum effect.

Produces a high rotational speed along with a very noticable force field easily felt at several feet!

Figure 8-3 *Isometric view of the overall unit*

6. Glue on the CS1 charge shield plate to prevent premature sparking to the metal heatsink of the module. Also, stick on the rubber feet.
7. The unit is perfectly usable. However, you may want to enclose it in some sort of plastic box. Always use good ventilation and make sure the box is large enough or about 1½ times the diameter of the rotor.
8. Find a suitable location, such as a nonmetallic table, to place the unit on. Place the rotor blade on the pin and verify it is unobstructed. Then connect the grounding lead to the receptacle mounting screw and plug in the T2 adapter. The rotor should begin rotating and quickly reach a high speed.

Notes

Ions are now being emitted into the air and can easily be detected by holding a fluorescent tube or a small neon indicating lamp (LA1) near the unit, noting a flashing. This must be done in the dark. You can also place a piece of paper near the base of the unit and note the force field generated that tries to pull the paper.

Figure 8-4 *Side view showing power connections*

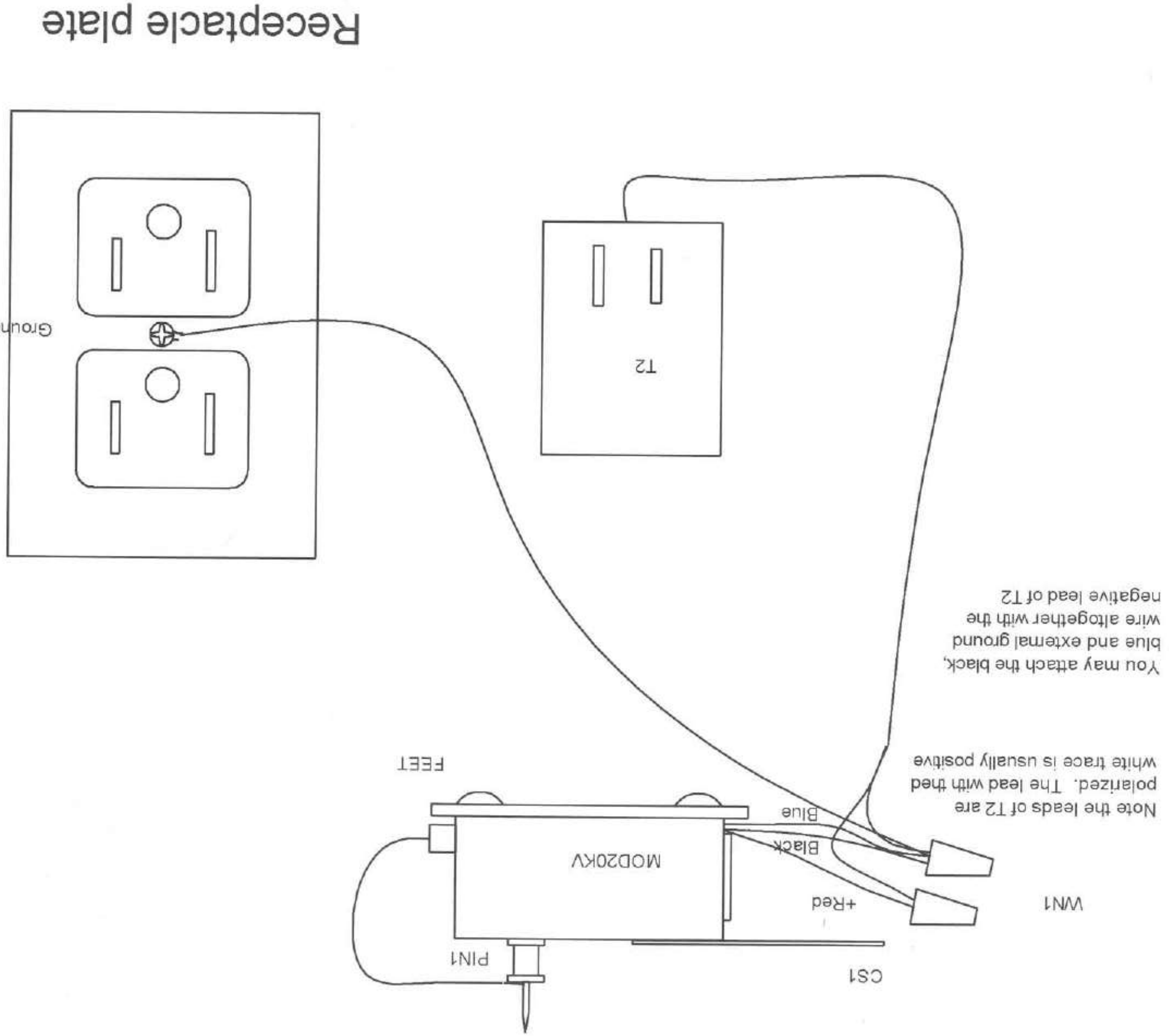
Receptacle plate
Ground
T2
FEET
MOD20KV
PIN1
CS1
+Red
Black
Blue
WN1

Note the leads of T2 are polarized. The lead with thed white trace is usually positive

You may attach the black, blue and external ground wire altogether with the negative lead of T2

Table 8-1 Negative Ion Machine Parts List

Ref. #	*Description*	*DB Part #*
MODULE	Negative 20-kilovolt module reworked	DB# MOD20KV
T2	12-volt DC, 500-milliamp wall adapter	DB# 12DC/.5
ROTOR	Metal fabricated rotor	DB# ROTOR
PIN	Common plastic pushpin	
BEARING	#6-32 threaded standoff	
NUT6	6-32 hex nut	
CS1	2¼ × 1½-inch-thin, 1/16 to 1/8-inch-thick plastic	
WN1	Small wire nuts	
FEET	Four stick-on rubber feet	

Chapter Nine

Kirlian Imaging Project

Kirlian photography uses *cold electron emission* to expose a photographic film. Cold electron emission is the result of electrons accelerating in an electric field. These accelerated electrons ionize the air and recombine producing ultraviolet light that is mostly invisible to the naked eye. This phenomenon is commonly known as *corona*. Many living objects are claimed to emit this radiation that will vary according to the health and condition of the object being exposed. This in itself is now subject to scientific research in material analysis. A more "psychic application" can be seen in the practice of radiation being referred to as an aura and being dependent on other more metaphysical functions. This project describes a simple method of producing Kirlian images of small objects and is an introduction to those who want to explore this interesting field. Kirlian imaging ranges from the psychic to materials analysis. Often it is used for experiments in which it is claimed that objects all possess a unique aura signature. This depends on the subject's mental makeup and condition at the time. Basically, it is a form of corona discharge that occurs as a precursor to a dielectric breakdown by most conductive objects when placed in a high-frequency, high-voltage field. Corona is a form of electrical leakage occurring without actual air breakdown such as a spark discharge. A corona discharge possesses ultraviolet emissions that easily develop film and thermographic paper, producing an image print.

This is an intermediate-level project requiring basic electronic skills and one can expect to spend $25 to $50. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com). The parts are listed in Table 9-1.

The system is simple and built from readily available parts (see Figure 9-1). The builder may deviate from what is shown as long as the basics are adhered to. The system requires a high-voltage, high-frequency module, as shown in Chapter 10 where it will be used to supply the burning plasma for the plasma pen project.

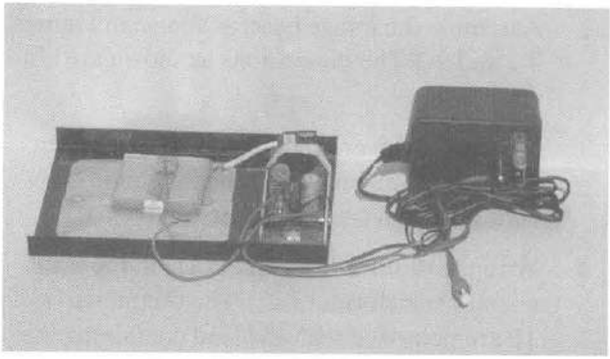

Figure 9-1 *A complete unit*

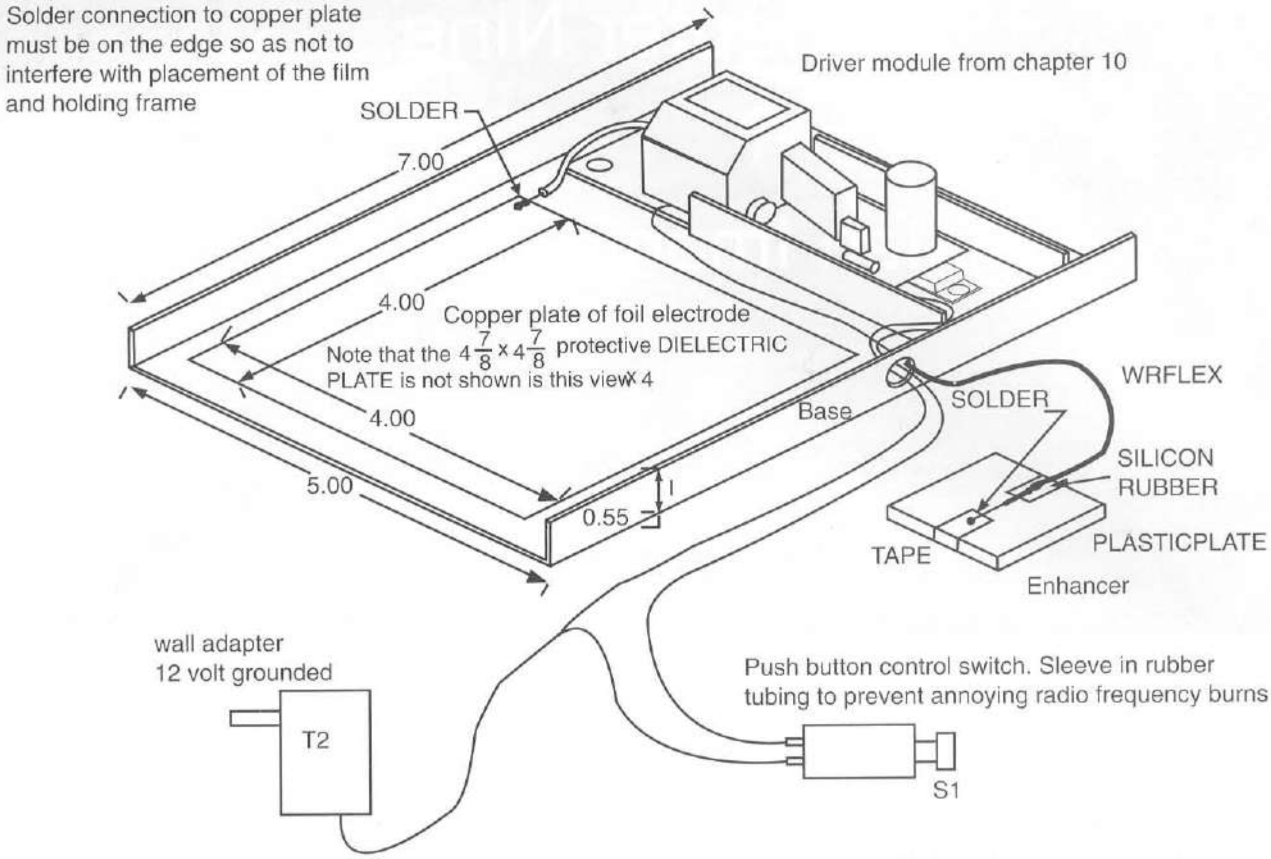
Solder connection to copper plate must be on the edge so as not to interfere with placement of the film and holding frame
SOLDER
7.00
Driver module from chapter 10
4.00
Copper plate of foil electrode
Note that the $4\frac{7}{8} \times 4\frac{7}{8}$ protective DIELECTRIC PLATE is not shown is this view X 4
4.00
5.00
0.55
Base
SOLDER
WRFLEX
SILICON RUBBER
PLASTICPLATE
TAPE
Enhancer
wall adapter
12 volt grounded
T2
Push button control switch. Sleeve in rubber tubing to prevent annoying radio frequency burns
S1

Figure 9-2 *Isometric of the image base and charge plate*

Assembly Steps

If you are a beginner, it is suggested to obtain our *GCAT1 General Construction Practices and Techniques*. This informative literature explains basic practices that are necessary in proper construction of electromechanical kits and is listed in Table 9-1.

1. Assemble the driver module, as shown in Figures 10-2 and 10-3 in Chapter 10. Verify proper operation before using.
2. Assemble the image base, as shown in Figures 9-2 and 9-3. The dimensions as shown are not critical.
3. Mount the driver module and solder the output lead as shown. Route the input and high-voltage return lead to the ENHANCER.
4. Wire the pushbutton switch (S1) to the wall adapter transformer (T2). The output wires of T2 are polarized with the lead containing the white trace, usually the positive. It is a good idea to cover the entire body of S1 except the plastic actuator; any contact to metal will result in small, very annoying burns. The wall adapter recommended has a third pin being an isolated ground for the high-frequency return.

Testing Steps for the Circuit

These steps usually will not require any special testing equipment.

1. Power up the unit and activate S1. Place your finger on the insulated plate and note a bluish corona discharge occurring under your finger. You may feel some heat or a tingling sensation. Try this for only up to 20 seconds.
2. Obtain some thermographic fax paper and place it on the dielectric plate. Place your finger on the paper and energize the circuit for 10 to 20 seconds. You will note an image on the paper.

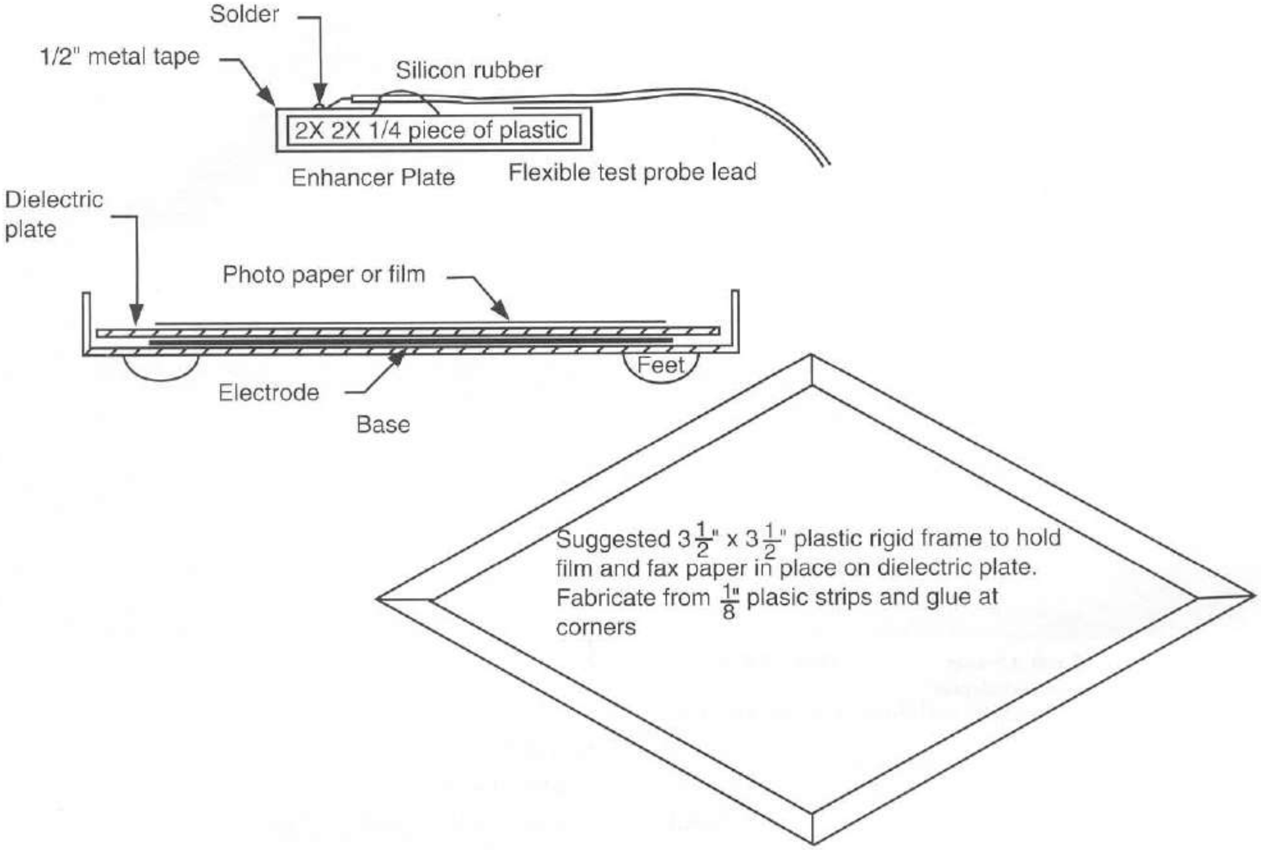
Solder
1/2" metal tape
Silicon rubber
2X 2X 1/4 piece of plastic
Enhancer Plate
Flexible test probe lead
Dielectric plate
Photo paper or film
Feet
Electrode
Base
Suggested $3\frac{1}{2}$" x $3\frac{1}{2}$" plastic rigid frame to hold film and fax paper in place on dielectric plate. Fabricate from $\frac{1}{8}$" plasic strips and glue at corners

Figure 9-3 *Image base and charge plate*

Try it again with a metal object such as a coin on the dielectric plate and place the enhancer on the object. Push the button for 5 to 10 seconds and note an image forming.

3. You can upgrade your image to an actual photograph by using actual camera film. The old Polaroid 600 or other similar instant film will work. You must work entirely in the dark to avoid film exposure to any ambient light. The objective is to place the film on the dielectric plate along with the object to be photographed and the enhancer placed on top.

Notes

You will note that a few variables here can affect performance: the output voltage of the driver module and the dielectric plate's thickness and dielectric properties. The dielectric plate is important because it must form a high enough capacity between the object and the contacts to allow an adequate corona current to flow to create an image. The thinner the plastic dielectric plate, the more corona current. However, you must not allow a breakdown to occur by using material that is too thin, as this can produce a painful burn.

Many different sources of information, including some excellent books, are available on the applications, methods, and uses of Kirlian photography. You are encouraged to continue to investigate this unusual electronic phenomenon.

Table 9-1 Kirlian Imaging Parts List

Ref. #	*Description*	*DB Part #*
DRIVER	Driver module from Chapter 10	DB# MINIMAX7
BASE	6 × 7-inch plastic piece, formed as shown in Figure 9-2	
COPPER ELECTRODE	4 × 4-inch, thin piece of copper	
DIELECTRIC PLATE	5 × 5 × .02-inch piece of polypropylene sheet	DB# POLYP20
PLASTIC PLATE	2 × 2 × 1/4-inch plastic plate	
TAPE	Four inches of 1/2-inch solderable metal sticky tape	
FEET	Four sticky rubber feet	
T2	12-volt, 1.5-amp groundedadapter with third pin	DB# 12DC/1.5G
FRAME PIECES	Four strips of 1/8-inch plastic to fabricate the frame as shown in Figure 9-3	
WRFLEX	12 inches of flexible test lead wire	
S1	Push-button switch reworked as shown to prevent burns	
GCAT1	General construction practices	DB# GCAT1

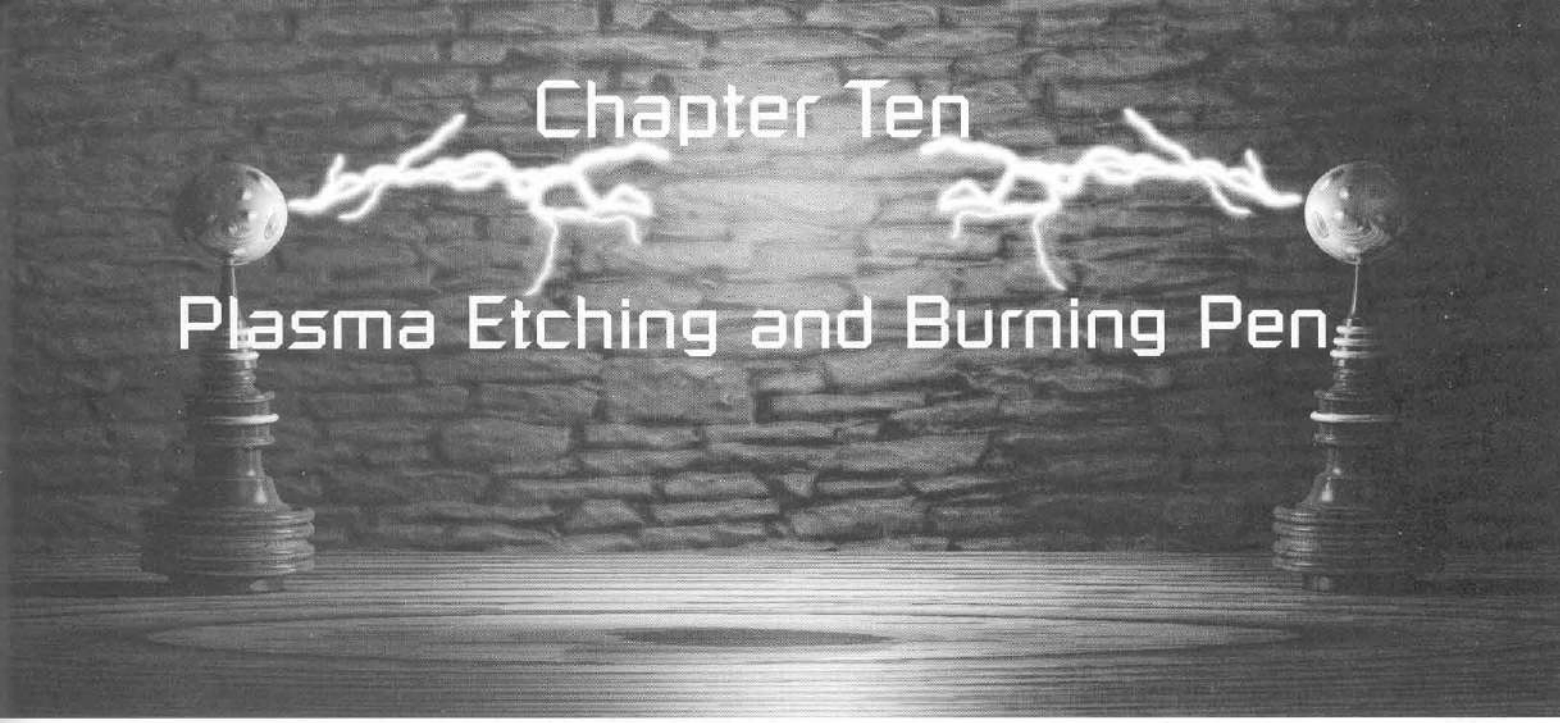

Chapter Ten

Plasma Etching and Burning Pen

A plasma pen uses high frequency, high voltage to burn or etch permanent markings into many different materials (see Figure 10-1). With a little practice, the user can create many intriguing designs of fine, lace-like etches. Names and other text can easily be etched into almost any organic material. Wood, paper, plastics, pumpkins, gourds, eggs, and leaves can all be marked with fine and fancy designs or inscriptions. The pen's other applications include Kirlian photography, aura enhancing, and magically lighting household fluorescent and neon tubes, all without any connecting wires, providing a great magic act. You can also use this for plasma experimentation or even removing warts.

This is an intermediate-level project requiring basic electronic skills. Expect to spend $25 to $35. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and all of them are listed in Table 10-1.

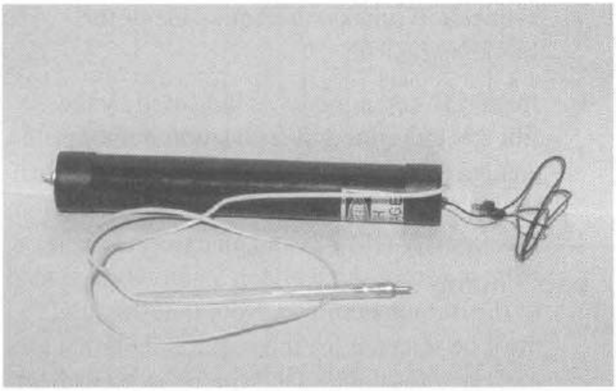

Figure 10-1 *Photo of assembly*

The system is shown in two parts: the power driver and the pen or etching section. These are connected together via a suitable, flexible, high-voltage wire. The power driver consists of the electronics module and the internal batteries for portable use. A jack is provided for external use from a 12 VDC wall adapter or another suitable source. This section is housed in a tubular enclosure with plastic end caps as shown.

Driver Circuit Description

The circuit is a frequency, high-voltage oscillator that consists of transistor Q1 connected as a simple oscillator where its collector drives the primary winding PR1 of the resonant transformer T1. Feedback is obtained via a second winding (FB) and fed to the base of Q1 through a current-limiting resistor (R1). Resistor R2 biases Q1 into conduction and initiates the oscillation (see Figure 10-2).

Capacitor C3 speeds up the turn-off time of Q1, while resistor R3 and capacitor C5 provide a filter to prevent oscillation at the self-resonant frequency of T1. Resonating capacitor C4 resonates the transformer

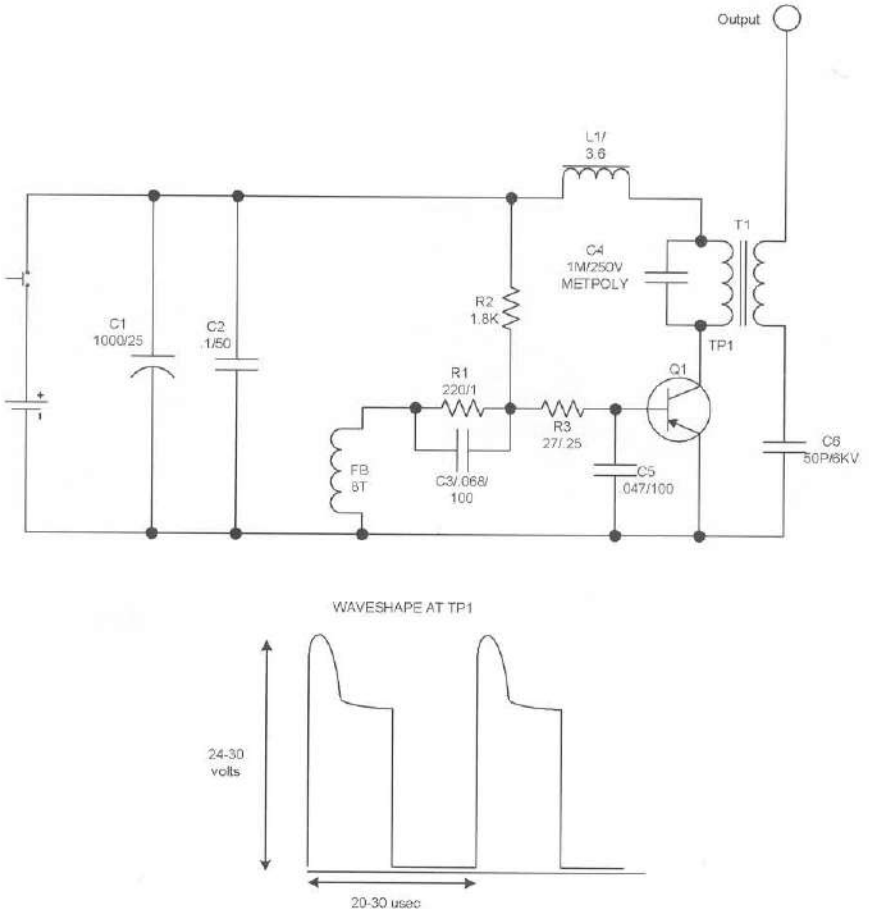

Figure 10-2 *Circuit schematic*

to around 35 KHz. The current-limiting inductor L1 limits the short-circuit current to a noncatastrophic value, and capacitors C1 and C2 bypass any signal to ground.

Driver Circuit Assembly

1. Lay out and identify all the parts and pieces. Verify them with the parts list, and separate the resistors as they have a color code to determine their value. Colors are noted on the parts list.
2. Obtain the available *printed circuit board* (PCB), as shown in Figure 10-3, or fabricate a piece of perforated circuit board to the PC as laid out on the PCB. Note the size of the PCB is 3½ × 1⅜ inches and contains the silk screening that shows the positioning of the mounted parts.
3. If you are building from a perforated circuit board, it is suggested that you insert the components starting in the lower left-hand corner. Pay attention to the polarity of the capacitors that have polarity signs and all semiconductors. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs and follow the foil traces on the drawing as these indicate the connection runs on the underside of the assembly board.
4. Insert the components as indicated by the silk-screen printed identification numbers and compare this with the bill of materials. Attach three 6-inch, #22-20 leads as shown for the input power (P1, P2) and an external high-voltage return. Also attach a high-voltage lead to the output as shown. Note that this lead must be selected for the required length. Use a silicon, 20-kilovolt DC wire or an equivalent.

Note the ground lead that connects to the metal frame under the screw head as shown.

5. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
6. Cut the metal bracket (BRK1) from a piece of .062 sheet aluminum and make the final assembly, as shown in Figure 10-4. Note that the metal tab of Q1 must be insulated from the bracket as shown on the mounting scheme. This piece is also the heatsink for Q1.

Driver Board Testing

This step will require basic electronic laboratory equipment. A 60 MHz oscilloscope is suggested.

1. Obtain a 12-volt DC source, preferably with a volt and ammeter. Leave the output leads as open circuit and apply power. Note a current draw of less than 250 milliamps. If you have a scope, observe the waveshape at the collector of Q1, as shown in Figure 10-2.
2. Contact the high-voltage output lead to the bracket and note that you can draw a 1/2-inch arc with the input current going to 1 amp. This completes this module.

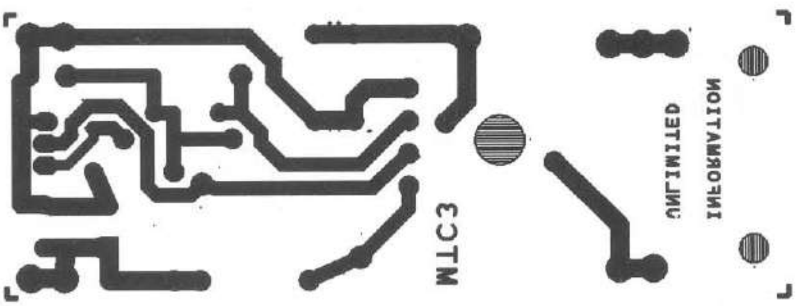
MTC3
INFORMATION UNLIMITED

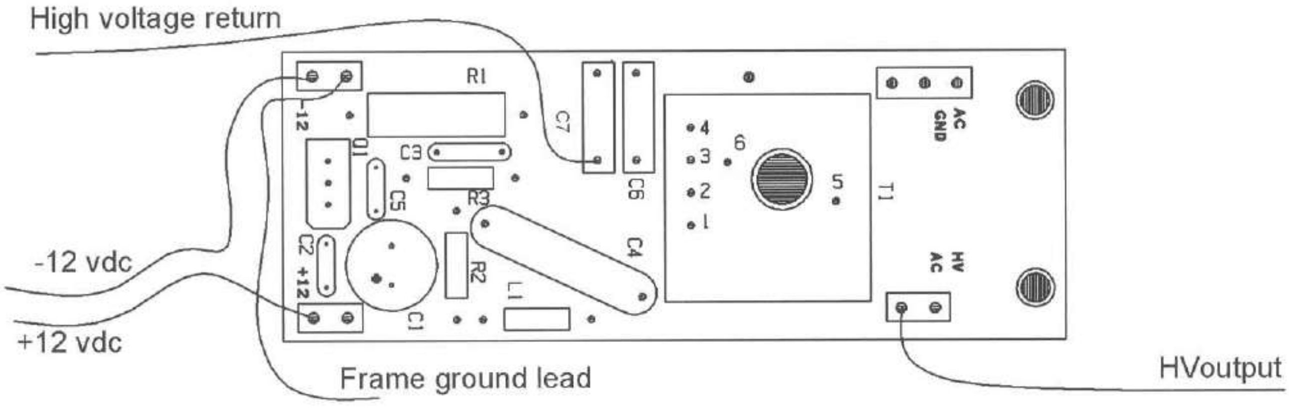
High voltage return
R1
-12
Q1
C3
C7
C6
C5
R3
C2
+12
R2
L1
C1
C4
4
3
2
1
6
5
T1
AC GND
HV AC
-12 vdc
+12 vdc
Frame ground lead
HVoutput

Figure 10-3 *Assembly board*

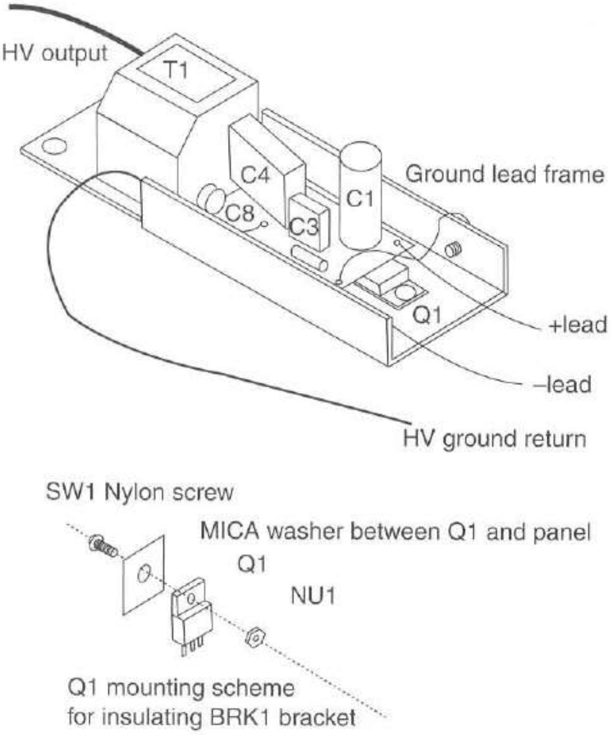
HV output
T1
C4
C8
C3
C1
Ground lead frame
Q1
+lead
–lead
HV ground return
SW1 Nylon screw
MICA washer between Q1 and panel
Q1
NU1
Q1 mounting scheme
for insulating BRK1 bracket

Figure 10-4 *Isometric of assembly*

Mechanical Assembly

You will need some basic cutting tools for the following steps.

1. Fabricate TUB2, as shown in Figure 10-5, and then make the caps CA1 and CA2 for passage of the output leads and for switch S1, jack J1, and indicator lamp LA1.
2. Insert the components and wire them up as shown. Note the current-limiting resistor R10. The exiting leads should be tied in a knot to relieve any strain.
3. Assemble the pen section as shown and solder the lead to the brass POINT. Shim the lead and POINT, and mechanically secure them in place.

Sample Demonstration

1. Obtain a 4 × 6-inch square piece of 1/4-inch masonite plywood, as shown in Figure 10-6. Moisten it with a sponge and place it on top of an approximately equal-size metal plate attached to the ground lead.
2. Contact the surface with the plasma pen and note the burning action. Experiment for the best results with other materials.

Note that your pen will light fluorescent lamps in your house just from its proximity to them, and it can also be used for Kirlian photography.

Switch S1 selects either internal batteries or external 12 volts, jack J1 is for the external 12 volts. LED indicates when the system is powered up. Resistor R1 limits the LED current. These components are mounted on the rear cap CAP2.

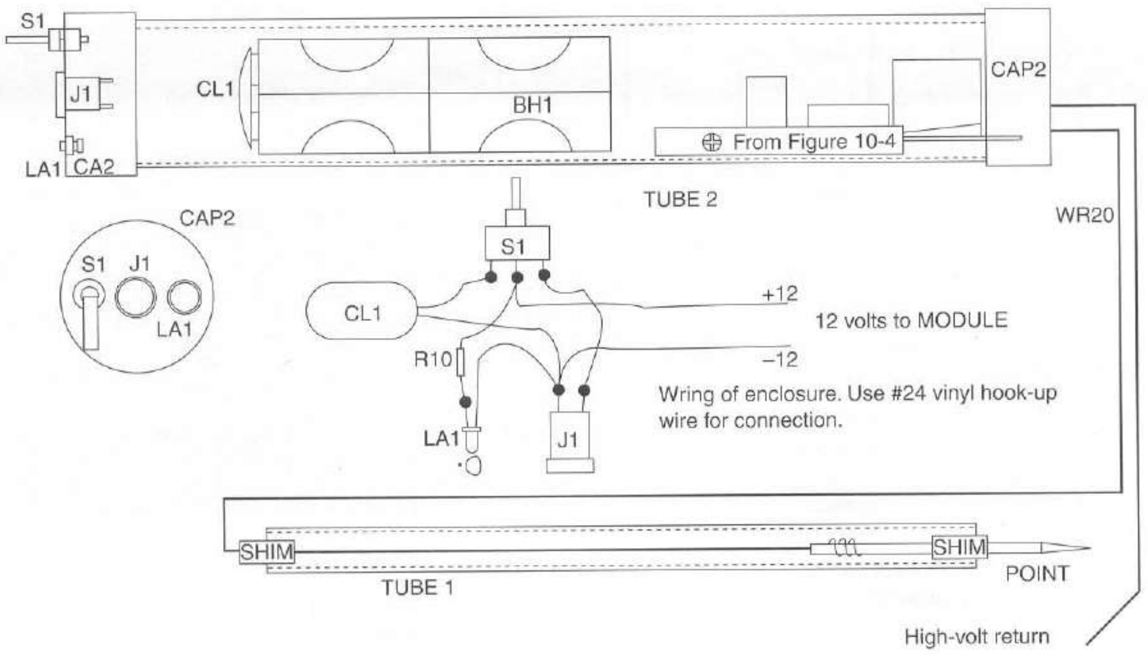
S1
J1
CL1
BH1
LA1
CA2
CAP2
From Figure 10-4
TUBE 2
WR20
CAP2
S1
J1
LA1
S1
CL1
+12
12 volts to MODULE
−12
R10
LA1
J1
SHIM
SHIM
TUBE 1
POINT

Wring of enclosure. Use #24 vinyl hook-up wire for connection.

High-volt return

The POINT can be a 1/5 to 1/4 brass rod that can be soldered to. The SHIMS can be rubber tubing that strain relieve and position the feed lead and burning point. TUBE1 can be a 3/8" diameter plastic tube.

Figure 10-5 *Final enclosure and wiring*

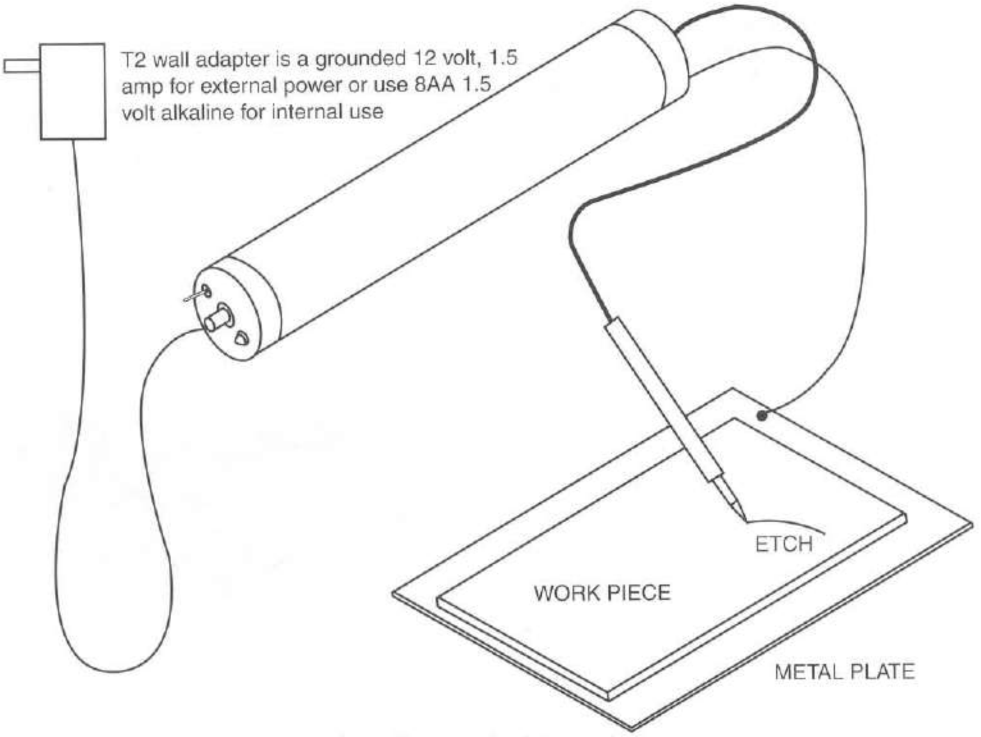

T2 wall adapter is a grounded 12 volt, 1.5 amp for external power or use 8AA 1.5 volt alkaline for internal use

ETCH
WORK PIECE
METAL PLATE

For optimum performance, use a piece of mansonite, thin wood, etc.
Place on top of metal plate connected to the high-voltage ground lead.
You may moisten the beginning area to start the etch.

Figure 10-6 *Sample demonstration*

Table 10-1 Plasma Etching and Burning Pen Parts List

Ref. #	*Description*	*DB Part #*
R1	220 ohms, 1 watt (red-red-br)	
R2	1.8K, 1/4-watt (br-gray-red)	
R3	27 ohms, 1/4-watt (red-pur-blk)	
C1	1,000-microfarad, 25-volt vertical electrolytic capacitor	
C2	.1-microfarad/50-volt small, plastic, capacitor	
C3	.068-microfarad, 50-volt plastic capacitor (683)	
C4	1-microfarad, 250-volt metallized foil large, blue, plastic capacitor	
C5	.047-microfarad, 100-volt plastic capacitor (473)	
C6, 7	Two 25-picofarad, 6-kilovolt ceramic capacitors or a single 50-picofarad, 6-kilovolt unit	
Q1	MJE3055T TO220 power tab NPN transistor	
T1	Special high-frequency transformer	DB# 28K089
WR20	24-inch #22-20 vinyl stranded hookup wire	
PCMTC	MTC3 printed circuit board or use a 3 1/2 × 1 1/2 × .1-inch grid perf board	DB# PCMTC
BRKT1	Create as shown in Figure 10-4	
TUBE2	12 × 2-inch ID schedule, 40 PVC tubing	
TUBE1	3/8 × 1/4-inch ID length of plastic tubing	
CA1, 2	Two mating flat-end caps for TUBE1	
POINT	1/8 × 3-inch brass rod filed to shape	
LA1	*Light-emitting diode* (LED), any color	
R10	220 1/4-watt resistor (red-red-br)	
CL1	Battery snap clip	
BH8	Eight-AA-cell battery holder	
S1	*Single pole, double throw* (SPDT) small toggle switch	
J1	Chassis-mounting DC power jack	
WR24	48-inch, #24 vinyl hookup wire	
WR20	24-inch, 20-kilovolt, silicon, high-voltage wire	
NEONX21	Assembled high-voltage module (see Figure 10-3)	DB# NEON21

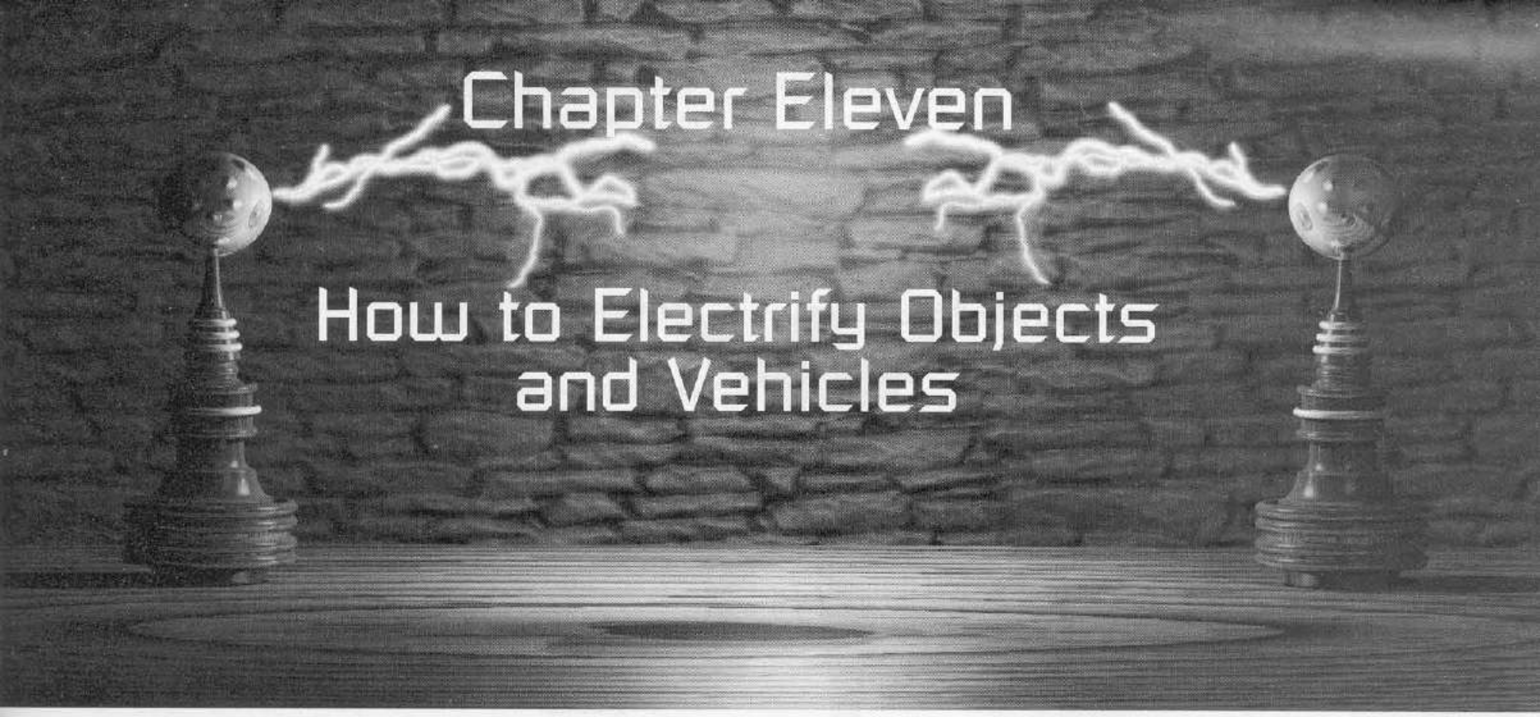

Chapter Eleven

How to Electrify Objects and Vehicles

Figure 11-1 shows two board-level devices capable of producing electrical shocks. The larger circuit generates low-energy, 20,000-volt pulses similar to an automotive ignition system. Although these pulses are nonlethal, they can still be very painful when contacted. The smaller circuit generates a steady 5-milliamp, low current at 2,000 volts. By itself, it is not lethal but again can produce a painful shock. If the output is allowed to charge a capacitor to a sufficient energy level, then the rules change and a potentially lethal situation exists. This level usually is any value over 25 joules as computed from W (joules) = half of C (capacitance in farads) × V^2 (in volts) that the capacitor is allowed to charge too.

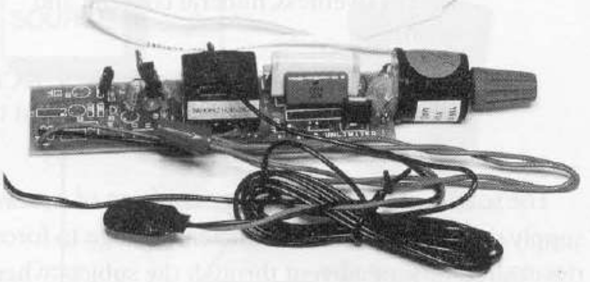

Figure 11-1 *Photograph of devices*

Caution: The use of electrification to prevent human intrusion is illegal. This application is legally referred to as "man trapping." Consult legal advice before using this if human contact is intended!

Warning: Never expose any person wearing a pacemaker or having a known physical condition to any form of electrical shock.

Data shows how to use several shocking systems with different types of outputs and applications. Even though the energy levels may be well below dangerous levels, secondary reactions can cause severe injury. Use caution if intended for human contact, and never connect the DC output version to a capacitor as a lethal charge may exist.

The following information is intended to serve as a guide to those contemplating using our equipment or other devices capable of producing a mild, severe, or possibly lethal electric shock. Any device of this nature, if used against people, may be considered a potential weapon. Remember that what may be a mild shock to one person could be harmful to another. Secondary effects, resulting from involuntary action, must also be considered.

The devices sold are intended for laboratory experiments or for use against animals. They should only be used for intimidation purposes to prevent an encounter. Human contact must be totally justified, as serious consequences can result from the illegality of their use.

Even though our electronic devices provide an average current below what is considered lethal to the average person, circumstances such as a person's heart condition or other physical parameters can influence both the immediate effect and after contact reaction.

Table 11-1 shows certain effects from various electrical currents. Remember that the indicated current must have an appropriate voltage behind it to flow through a given resistance, such as the human body. When working with sources that have storage capacitors (especially dangerous sources), one must take into account the high peak current that can result for a period of time and result in possible electrocution if this duration of time is sufficient.

Table 11-1 Electrical current scale

Voltage	*Degree of pain*	*Examples*
.01 to 1 ma	Tingling to annoying	Static electric shocks
1 to 5 ma	Annoying to painful	Spark plugs, TV picture tubes, trip current of GFI
5 to 20 ma	Painful to very painful	Oil burner ignition, bug killers, stun guns
20 to 50 ma	Very painful to possibly lethal	Neon sign transformer, old tube radios
50 to 100 ma	Possibly lethal to lethal	Low-powered transmitters, capacitor charges for lasers
100 to 500 ma	Lethal to deadly	Medium-powered transmitters, laboratory power supplies
500 ma to 1amp (+)	Deadly, with usually no second chance	Electric chair, 220-volt house current under certain circumstances

Supplementary Electrification Data and Information

The following information explains how to place an electrical charge on an object to protect against animal damage, intrusion, or theft, and the construction of several energy sources is referenced. These must be treated with caution and never be used on human beings without proper discretion due to potential injury or electrocution. Accidental contact must also be avoided. Many laws unfortunately are made to protect the criminal, and therefore any attempt made to protect oneself or property using these devices could result in a suit by the criminal if he or she can prove injury.

The methods for electrifying objects and areas will be shown as several examples that the user may adapt to his or her own needs. To successfully electrify, an object requires the following conditions to exist. Note that the following pertains to flowing electric current and not to static or super-high voltages. Figure 11-2 provides the basic explanation.

- The object of contact must be of a conductive nature such as metal, conductive paint, or wet ground.
- The subject (the intruder or animal) must supply the link to complete the circuit; that is, his or her physical extremities must come in contact with the object, along with a second common surface such as the ground, water, or another object anticipating simultaneous contact. No current flows unless contact is made to complete the circuit by the subject physically touching the object.
- The electric current flows from the +lead of SOURCE to the −lead. In doing so, it flows from the object to the point of contact by the subject to the ground and back to SOURCE via the −lead. This produces a current flow through the subject, developing an electrical shock. The severity of this shock is dependent on:
    - How well the ground will conduct this current (wetness, mineral content, and so on)
    - The electrical parameters of the SOURCE, that is, the voltage and available current to support this voltage

The source of electrical energy is a form of a power supply made to generate a sufficient voltage to force the desired flow of current through the subject when contact is made. This can take on many different forms depending on the applications and the desired results. If a system was built to kill a human being, a power supply with a current supply capability of $^1/_2$ to 20 amps would be necessary to support a voltage high

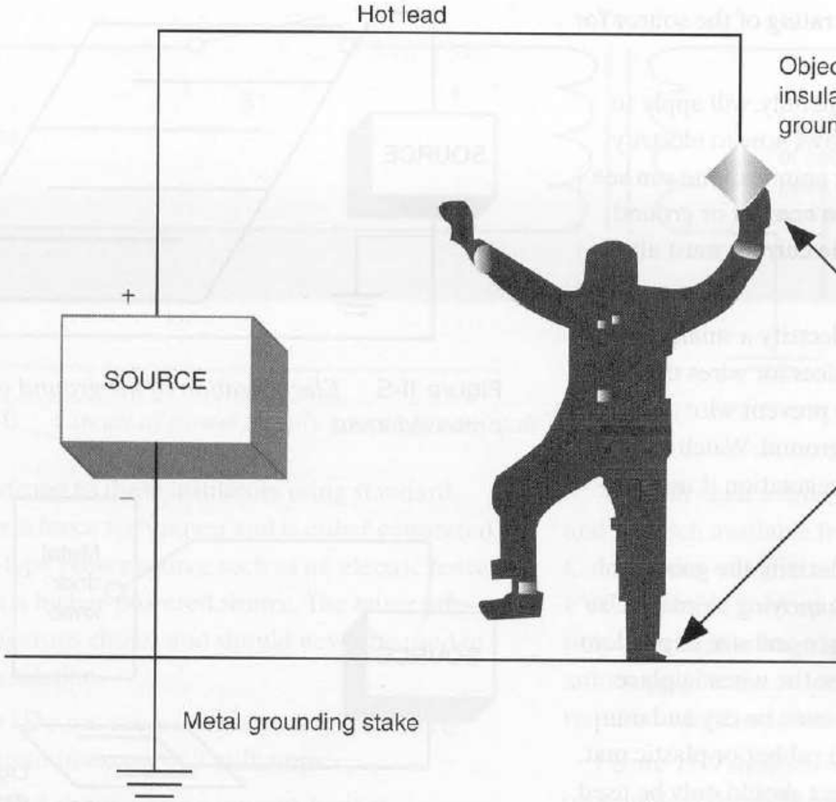
Hot lead
Object must be insulated from ground
+
SOURCE
Contact points
−
Metal grounding stake

Figure 11-2 *Electrification principle*

enough to cause this current to flow, usually several thousand volts. (These parameters are similar to those used in electric chairs.) Currents of more than 1/10 of an amp are considered lethal to the average human under certain circumstances, 1/100 of an amp produces intense pain, and 1/1,000 of an amp is felt with no difficulty and is very annoying.

Special pulse-type energy sources, similar to capacitor discharge ignitions, can be used relatively safely as peak currents are high, yet the average is low. A source of this type is the pulsed model described in this project. This source is capable of generating the high-peak current pulses and maintaining a substantial voltage on the body of a vehicle in spite of leakages due to steel-belted tires and so on. The source is also relatively safe to exposure but will produce moderate shocks. With these parameters, one

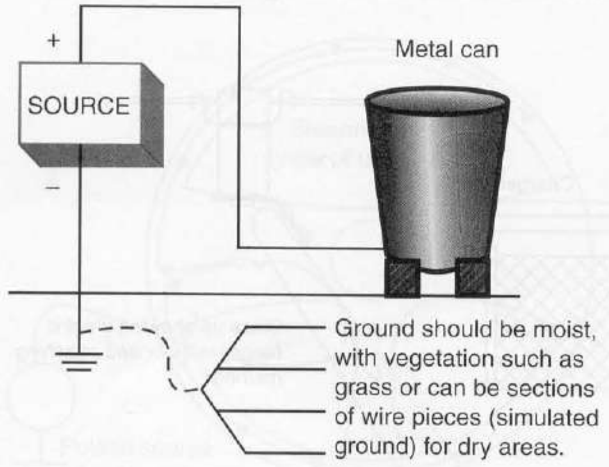
+
SOURCE
−
Metal can
Ground should be moist, with vegetation such as grass or can be sections of wire pieces (simulated ground) for dry areas.

Figure 11-3 *Electrification of a garbage can or similar object*

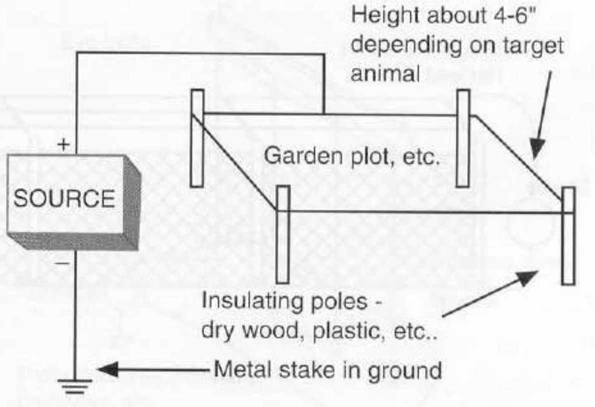
Height about 4-6" depending on target animal
Garden plot, etc.
+
SOURCE
Insulating poles - dry wood, plastic, etc..
Metal stake in ground

Figure 11-4 *Electrification of a small area*

can get an idea of the size and rating of the source for a particular application.

The following examples, hopefully, will apply to most situations. Figure 11-3 shows how to electrify garbage cans against annoying animals. One can see that the condition of a common contact or ground system is important because the current must also flow through this medium.

Figure 11-4 shows how to electrify a small garden area using plastic stakes with slots for wires that are pushed into the ground. Try to prevent wire droop and possible contact with the ground. Watch for fire hazards and proximity to dry vegetation if using a high, average current source.

Figure 11-5 shows how to electrify the ground of an area and protect it against annoying animals. Use #20 wire as shown with the space and size dependent on the animal and area. Anchor the wires in place with small stones. The ground must be dry and nonconductive, or use an insulated rubber or plastic mat. Please be aware that this project should only be used with a low-energy pulsed system or a hand-cranked magneto. Do not create this system if you have a heart condition.

Figure 11-6 shows a prank, but this is only to be done using a low-energy pulsed system or a hand-cranked magneto. The victim stands on a rug to open a door and zap! You turn the power on.

Figure 11-7 shows the electrification of a fence. This method is for educational use only, as serious injury can result either from the electric shock or the physical reaction. The intended use would be for military or other high-security situations.

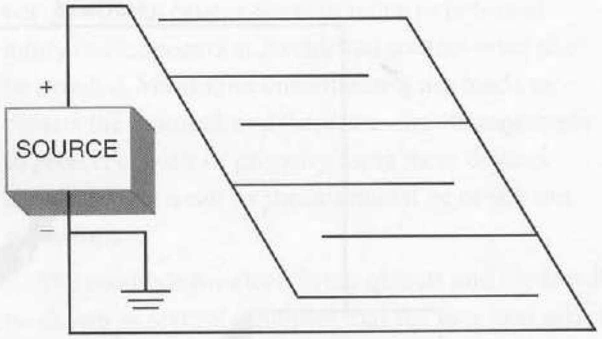

Figure 11-5 *Electrification of the ground or protected areas*

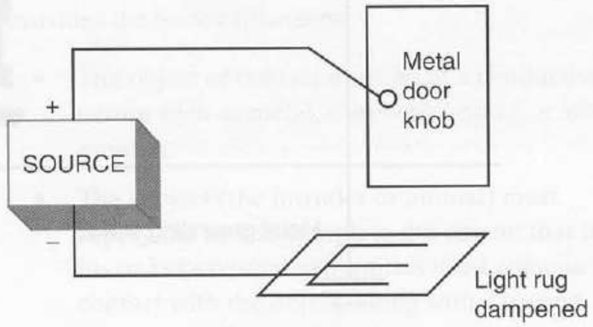

Figure 11-6 *Door knob prank*

The object here is to electrify the top-most wire of a chain-link fence, as anyone attempting to crawl over will no doubt come in contact with the top wire. This is a very effective method and can cause severe injury and even death from electrocution or physical injury from falling. The best method is to simply replace the original top-most wire with a standard, electrical farm fence insulator, as shown in Figure 11-7. The wire is

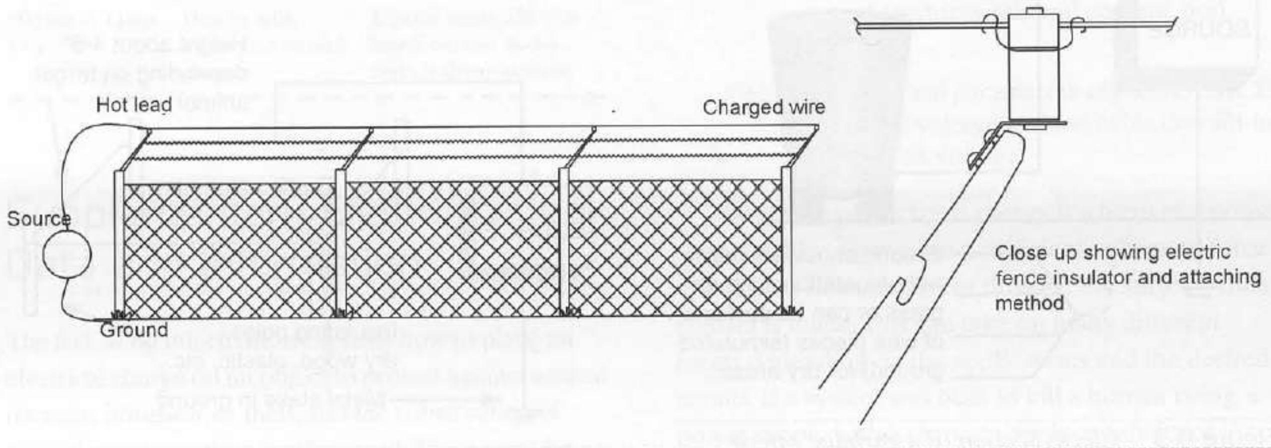
Hot lead
Charged wire
Source
Ground
Close up showing electric fence insulator and attaching method

Figure 11-7 *Electrification of a chain-link or similar fence*

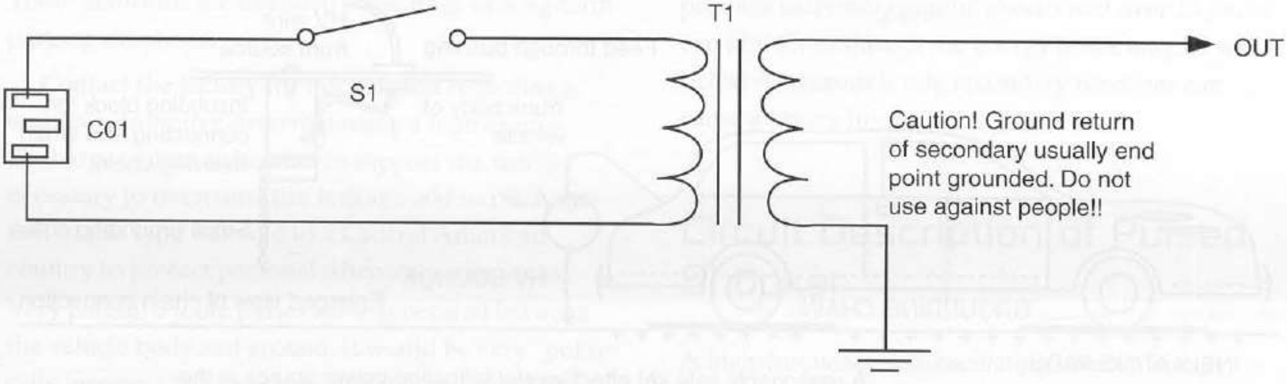

Caution! Ground return of secondary usually end point grounded. Do not use against people!!

Figure 11-8 *Circuit of power supply source for animals only*

then just strung to these insulators using standard, electric farm fence techniques and is either connected to a farm-type power source such as an electric fence charger or a higher-powered source. The latter presents a dangerous choice and should never be used in a normal situation.

Danger! Do not use a transformer with a short-circuit current in excess of 9 milliamps.

Figure 11-8 shows using a current-limited transformer powered by 115 VAC. The following are suggested power sources for the circuit shown in Figure 11-8.

> 4KV/009: T1 −4 KV, 9 ma, 115 VAC, current-limited floating secondary Database #4 KV/.03 A
>
> 6KV/02: T1 6.5 KV, 20 ma, 115 VAC, current-limited floating secondary Database #6 KV/.02 A

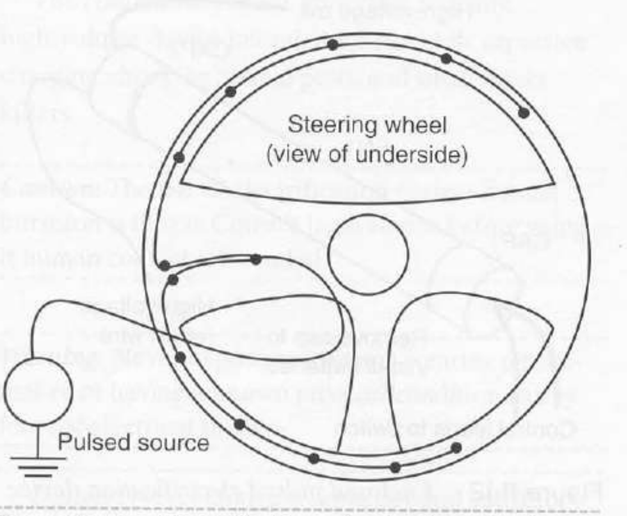

Figure 11-9 *Electrification of a steering wheel or similar object*

You will need a three-wire grounded power cord and a switch available from most hardware stores. Connect the −connection to common using standard PVC or plastic hookup wire. Connect the +connection via a high-voltage ignition wire available from an auto supply house. A weatherproof system is required when installed outside.

Figure 11-9 shows how to electrify a steering wheel or a similar object. The project is intended as a theft or unauthorized use deterrent. Use a low-energy, pulsed, high-voltage source.

The subject sits in the seat and activates the pressure switch to the source. When he grabs the steering wheel, he experiences a mild pulse-type shock between his hand and rear end. The voltage return is via metal springs in the seat. To create this project, please follow these steps:

1. Tape some thin, bare wire on the underside of the steering wheel. Secure it with small pieces of tape, wax, or silicon rubber.

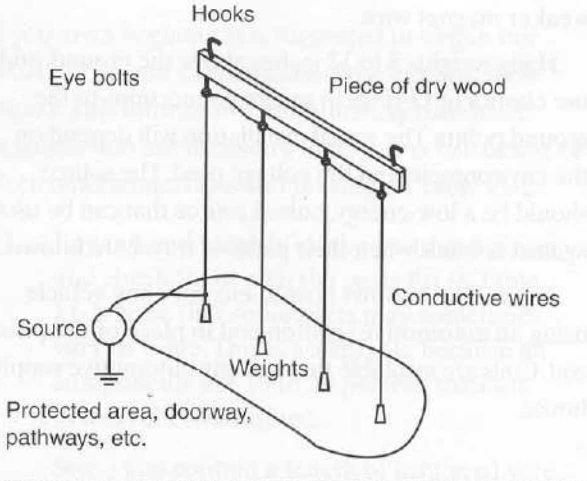

Figure 11-10 *Space or field electrification*

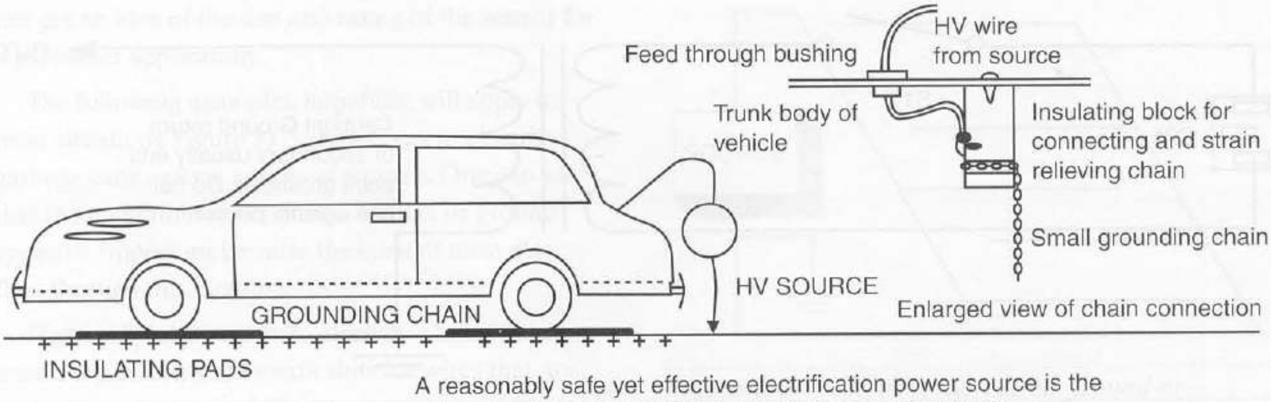
Feed through bushing
HV wire from source
Trunk body of vehicle
Insulating block for connecting and strain relieving chain
Small grounding chain
HV SOURCE
Enlarged view of chain connection
GROUNDING CHAIN
INSULATING PADS

A reasonably safe yet effective electrification power source is the pulsed version shown, with a standard automotive ignition coil replacing the small trigger coil. This approach lengthens the pulse to overcome the electrical capacity of the vehicle body and provides a better overall effect.

Figure 11-11 *Electrification of a vehicle or similar object*

2. Run a concealed connecting wire up the steering column to the bare wire. This wire should be insulated and connect to the output of the source.
3. Conceal an activation switch as desired and don't forget to de-energize it before entering the car. You may want to rig up a pressure-sensitive switch on the seat so when the thief gets in the unit it activates before he starts the car. Important caution: The unit must be activated without the car being in motion or serious accidents may result.
4. The use of our pulsed low-power source is suggested.

Figure 11-10 shows how to protect an area, doorway, or paths using space or field electrification. Hanging supports use ultra-thin steel wire for strength or weaker magnet wire.

Hang weights 8 to 12 inches above the ground, and use elastics or O-rings to secure connections to the ground points. The actual installation will depend on the environment and the voltage used. The source should be a low-energy, pulsed source that can be used against animals when their paths of travel are known.

Figure 11-11 shows how to electrify any vehicle using an automotive ignition coil in place of the pulse coil. Coils are available through any automotive supply house.

A remote switch must be mounted so that it is accessible from the outside of the vehicle. This may be actuated via an insulating object such as a plastic comb or something similar. An override switch may be installed to actuate this system while one is inside the vehicle in order to discourage certain encounters involving contact while driving or stopped at a light.

The suitable method for connecting to earth ground as the common terminal is accomplished via a grounding chain. This is shown strain relieved via a plastic insulating block. The user is asked to use his or her own ingenuity to create other methods of implementation.

Insulated platforms for wheels should be used if the tires have steel-belted radials or if roads are salty.

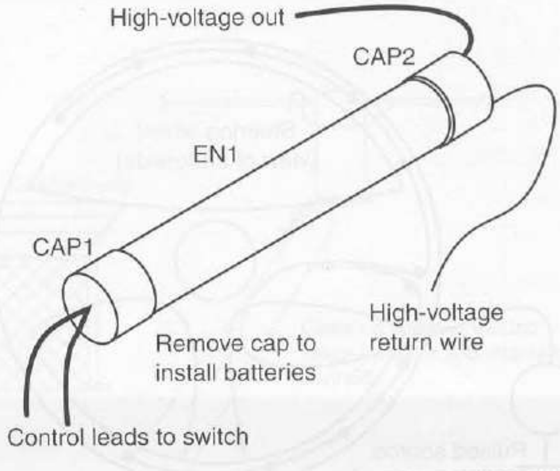
High-voltage out
CAP2
EN1
CAP1
High-voltage return wire
Remove cap to install batteries
Control leads to switch

Figure 11-12 *Enclosed pulsed electrification device*

These platforms are intended for storage or long-term parking situations.

Contact the factory for information regarding a serious automotive deterrent using a high-energy discharge system to be able to support the voltage necessary to overcome tire leakage and so on. A system of this type was sold to a Central American country to protect personal when in hostile areas. Very painful 5 joule pulses were generated between the vehicle body and ground. It would be very "politically incorrect" if it were used in this country. This system produces high-current pulses of several tens of amperes for short durations. This is necessary as most tires offer considerable leakage at these high voltages, and a steady-state supply would no doubt burn off the tires and be very lethal to the perpetrator.

Assembly of Suggested Power Sources

The *pulsed version* shown in Figure 11-12 can also be used as a high-voltage trigger for triggering spark gaps, flash tubes, chemical igniters, Kirklian photography, and for electrification in animal control. The unit is housed in a tubular enclosure and powered by built-in batteries or by an external 12 VDC wall adapter. The device produces 20,000-volt pulses for electrifying objects, fences, and so on. Dual-mode operation produces low-energy pulses of .02 joules at a repetition rate of 30 per second or medium-energy pulses of .2 joules at a repetition rate of 4 per second.

The *continuous version* generates a steady, high-voltage device intended for research, capacitor charging, shocking animal pests, and small insect killers.

Caution: The use of electrification against human intrusion is illegal. Consult legal advice before using if human contact is intended.

Warning: Never expose any person wearing a pacemaker or having a known physical condition to any form of electrical shock.

Danger! Never charge a capacitor and use it for shocking humans. Energy storage over 5 joules can produce extremely painful shocks and over 25 joules can kill! Even though the energy levels may be well below dangerous levels, secondary reactions can cause a severe injury.

Circuit Description of Pulsed Shocker

A high-frequency, self-oscillating inverter circuit comprised of switching transistor Q1 and stepup transformer T1 produces a high-voltage high frequency at the secondary winding. This high-voltage AC is rectified by diode D1 and charges up storage capacitor C3 or C4 through isolation resistor R3. When this voltage charges up to the breakdown potential of SIDAC (SID1), the energy stored in the capacitors is "dumped" into the primary of the high-voltage pulse transformer T2, producing a high-voltage pulse at the output terminals (see Figure 11-13).

The oscillator circuit utilizes a winding on T1 to produce the necessary positive feedback to the base of Q1 to sustain oscillation. Resistor R1 initiates Q1 to turn on while resistor R2 and C2 control the base current and operating point.

A charging circuit consisting of the current-limiting resistor R4 and rectifier diode D2 allows external charging of battery B1 when *nickle cadmium* (NiCad) or other rechargeable batteries are used.

Construction Steps for Pulsed Shocker

If you are a beginner it is suggested to obtain our *GCAT1 General Construction Practices and Techniques*. This informative literature explains basic practices that are necessary in proper construction of electromechanical kits and is listed on Table 11-2.

1. Lay out and identify all the parts and pieces, and check them with the parts list in Table 11-2. Note that some parts may sometimes vary in value. This is acceptable because all components are 10 to 20 percent tolerant unless otherwise noted.

   Some kits contain a length of insulated wire.

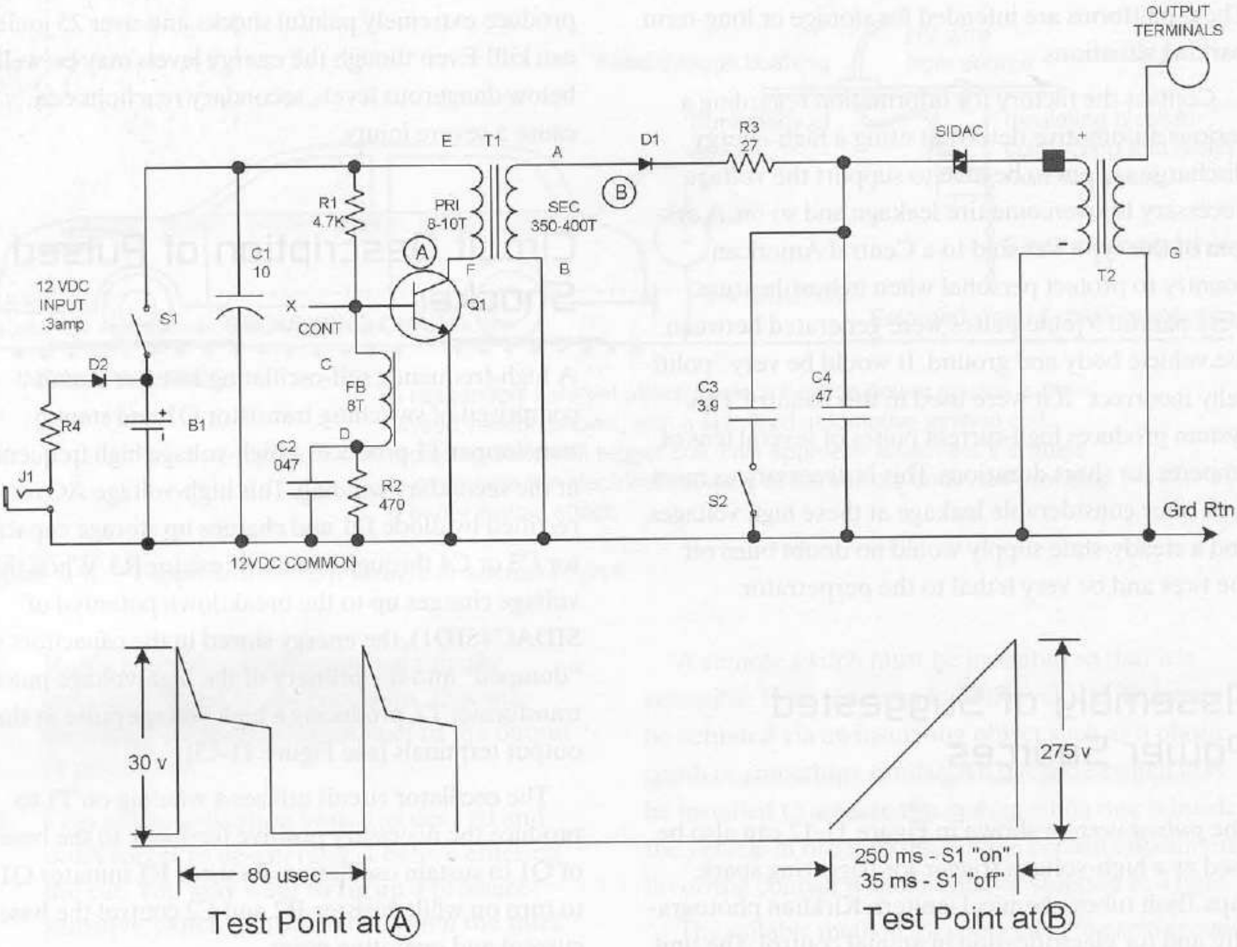
OUTPUT TERMINALS
E
T1
A
D1
R3
27
SIDAC
+
B
R1
4.7K
PRI
8-10T
SEC
350-400T
B
F
A
Q1
C1
10
12 VDC
INPUT
.3amp
S1
X
CONT
C
FB
8T
D
D2
R4
B1
C2
.047
R2
470
J1
12VDC COMMON
C3
3.9
C4
47
S2
T2
G
Grd Rtn
30 v
80 usec
Test Point at A
275 v
250 ms - S1 "on"
35 ms - S1 "off"
Test Point at B

Figure 11-13 *Circuit schematic*

This must be cut, stripped, and tinned according to where it is used.

2. Assemble the board as shown in Figure 11-14. Insert the components into the board holes as shown, proceeding from left to right. Note the polarity of the components.

Note: Certain leads of the actual components will be used for connecting points and circuit runs. Do not cut or trim them at this time. It is best to temporarily fold the leads over to secure the individual parts from falling out of the board holes for now.

---

3. Verify the wiring, proper components, and polarity to the schematics. Check for cold solder, excess solder, and bare-wire bridge shorts. Now the unit is ready to test.
4. Test the unit as per the following:

A. Position the bare end of the ground return lead (GRD RTN) to allow a ½- to ¾-inch air gap between the output pins of T2.

B. Connect 12 volts from a bench power supply or use eight AA NiCads at 1.25 volts each or eight AA alkalines at 1.5 volts for 12 volts.

C. Note with S2 open a fast pulsing action will produce a thin, bluish discharge. This can cause a very mild electrical shock and could be safely used within reason as a prank.

D. Note a thick, slow pulsing discharge when S2 is closed. This can produce a painful shock and is intended for use against animals.

E. The current draw with the unit properly operating should be approximately 250 milliamps.

F. Check the power tabs of both Q1 and SIDAC. These should be cool to warm to the touch.

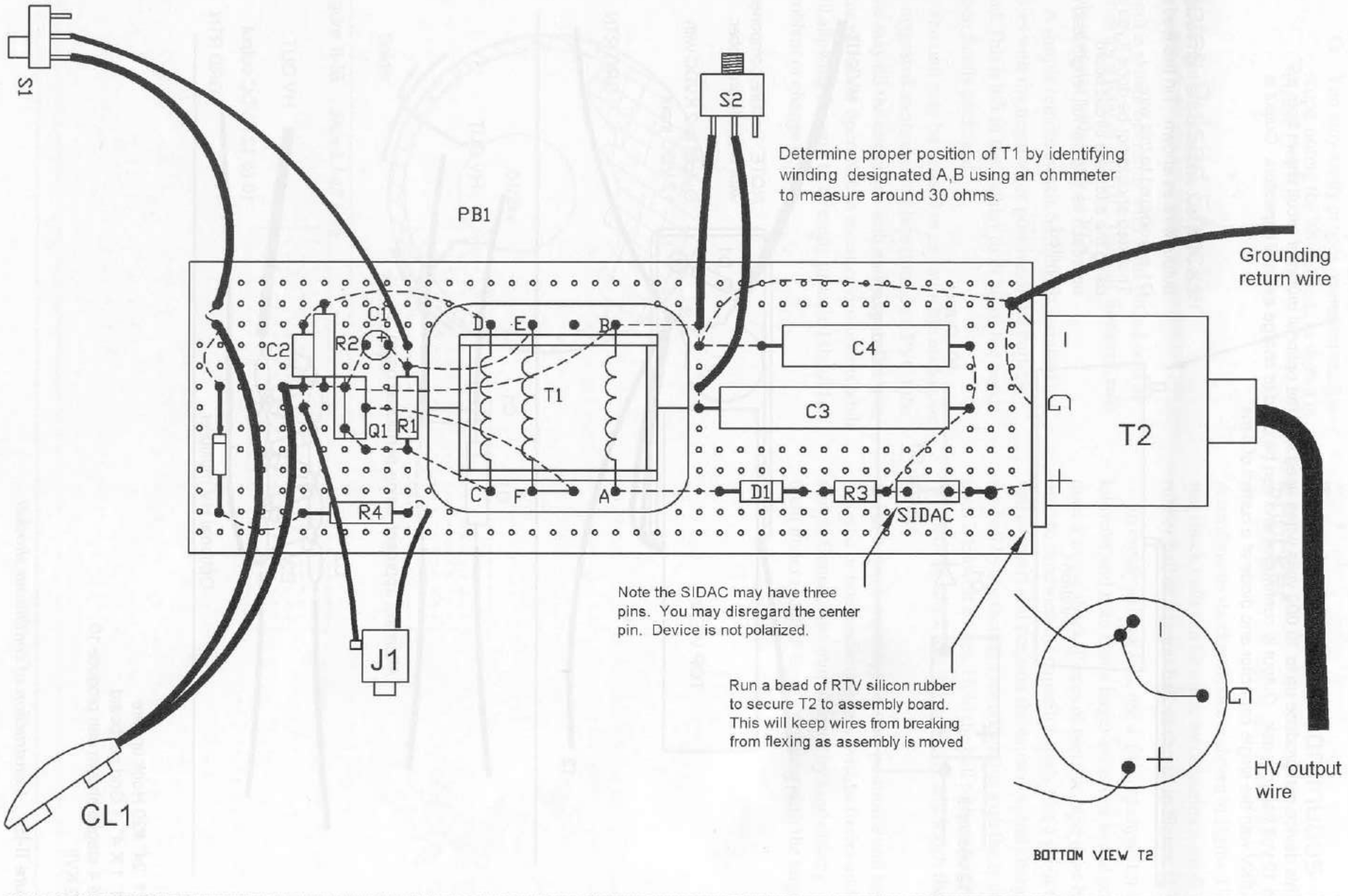
S1
S2
PB1
C1
C2
R2
Q1
R1
R4
D
E
B
C
F
A
T1
C4
C3
D1
R3
SIDAC
T2
G
-
+
J1
CL1
Grounding return wire
HV output wire
BOTTOM VIEW T2

Determine proper position of T1 by identifying winding designated A,B using an ohmmeter to measure around 30 ohms.

Note the SIDAC may have three pins. You may disregard the center pin. Device is not polarized.

Run a bead of RTV silicon rubber to secure T2 to assembly board. This will keep wires from breaking from flexing as assembly is moved

Figure 11-14 *Assembly board*

This device can produce up to 10,000 volts when used with the optional multiplier circuit shown from a 9 to 12 volt battery pack. Output is continuous and can be used to charge external capacitors. Output is 1.5KV with the single capacitor and diode at a current of 5 ma.

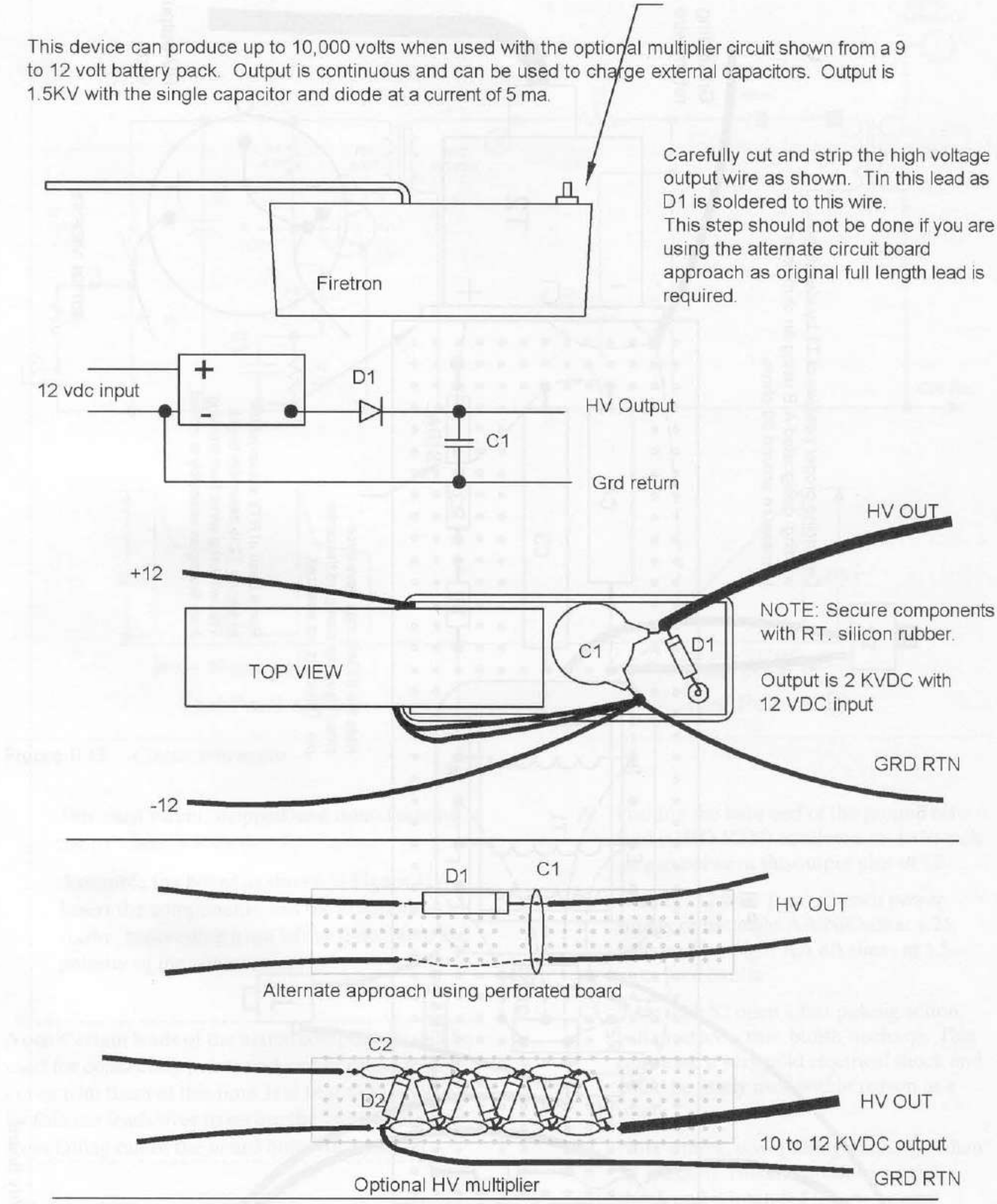

Carefully cut and strip the high voltage output wire as shown. Tin this lead as D1 is soldered to this wire.
This step should not be done if you are using the alternate circuit board approach as original full length lead is required.

NOTE: Secure components with RT. silicon rubber.

Output is 2 KVDC with 12 VDC input

Note 4 stages shown can produce 10 to12 KV!!

Firetron
12 vdc input
+
-
D1
HV Output
C1
Grd return
HV OUT
+12
TOP VIEW
C1
D1
-12
GRD RTN
D1
C1
HV OUT
Alternate approach using perforated board
C2
D2
HV OUT
10 to 12 KVDC output
GRD RTN
Optional HV multiplier

WR1 24" #20 Hook up wire
PB1 1 X 4" .1 Grid perfboard

Figure 11-15 *Construction of continuous shocker*

G. You may verify proper operation using a scope, noting the waveshapes as shown in Figure 11-13.

Notes Pulsed Shocker

Input is shown operating with eight NiCad cells in series for a total of 10 VDC. The unit, however, reliably operates within 8 to 14 VDC.

A simple remote switch, S1, may be installed in series with the negative or positive of the battery lead. This is left to the builder to fit his or her needs when finally packaged.

The unit may be housed in any suitable enclosure. A suggested method is to place it inside a PVC tube You may fill the enclosure with melted paraffin wax. This will allow operation in moist environments while still allowing the unit to be easily unsealed should a problem or change occur

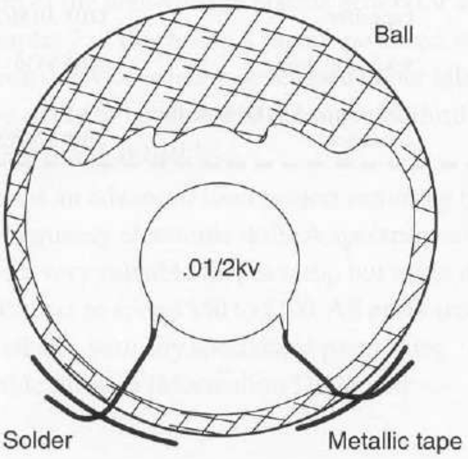

Figure 11-16 *Shock ball joke*

Construction of Continuous Shocker

A continuous shocker device is shown in Figure 11-15, and shock balls can be made by obtaining a small rubber ball or a tennis ball, as shown in Figure 11-16.

To create a shock ball, use a .01-amp capacitor at 2 kilovolts and note that a larger capacitor would produce a worse shock. Pieces of metallic tape must be used to sandwich the capacitor leads. Cut a slit in the ball to insert and position the capacitor, and charge the ball by contacting a metallic strip with the continuous 2-kilovolt source. Hold the ball without touching the metallic strips, toss it to a buddy, and watch the results.

For a shock wand application, position a unit inside a plastic or nonconductive tube with batteries and a switch. Connect the output leads by sandwiching them under metallic tape for contact with the target subject.

Table 11-2 Parts List for Shockers

Ref. #	*Description*	*DB Part #*
R1	4.7K, ¼-watt carbon film resistor (yel, pur, red)	
R2	470 ohm, ¼-watt carbon film resistor (yel, pur, br)	
R3	27-ohm, ¼-watt carbon film resistor (red, pur, blk)	
R4	100-ohm, ¼-watt carbon film resistor (br, blk, br)	
C1	10-microfarad, 25-volt electrolytic capacitor vertical mount	
C2	.047-microfarad, 50-volt polyester capacitor	
C3	3.9 to 4-microfarad, 350-volt polyester capacitor	
C4	.47-microfarad, 250-volt polyester capacitor	
Q1	MJE3055 NPN transistor TO220	
D1, 2	Two IN4007 1-kilovolt rectifier diode	
SIDAC	300-volt Sidactor switch	SIDAC
T1	400-volt switching square-wave transformer	DB# TYPE1PC
T2	25-kilovolt pulse transformer	DB# CD25B
S1, 2,	Two *single pole, single throw* (SPST) 3-amp toggle switch or equivalent	
PB1	5 × 1.5 × .1-inch grid perforated circuit board	
PCLITE	Optional *printed circuit board* (PCB)	DB# PCLITE
WR20B	36-inch #20 vinyl stranded hookup wire, black	
WR20R	36-inch #20 vinyl stranded hookup wire, red	
WRHV20	12-inch 20-kilovolt wire	
CL1	Battery clip	
CAP1, 2	Two 1⅝-inch plastic caps	
EN1	12 inches of 1⅝ × 1½ × 12 plastic tube	
BH1	Eight-AA-cell battery holder	
J1	3.5-millimeter mono jack	
12DC/.3	Optional 12 VDC, .3-amp wall adapter	DB# 12DC/.3
Parts for the continuous version		
FIRETRON	Firetron 2,000-volt module	DB# FIRETRON
C1	.01-microfarad, 2,000-volt disc capacitor	DB# .01M/2KV
D1	6-kilovolt diode	DB# VG6
C2	270-picofard, 3,000-volt disc capacitor	DB# 270P/3KV

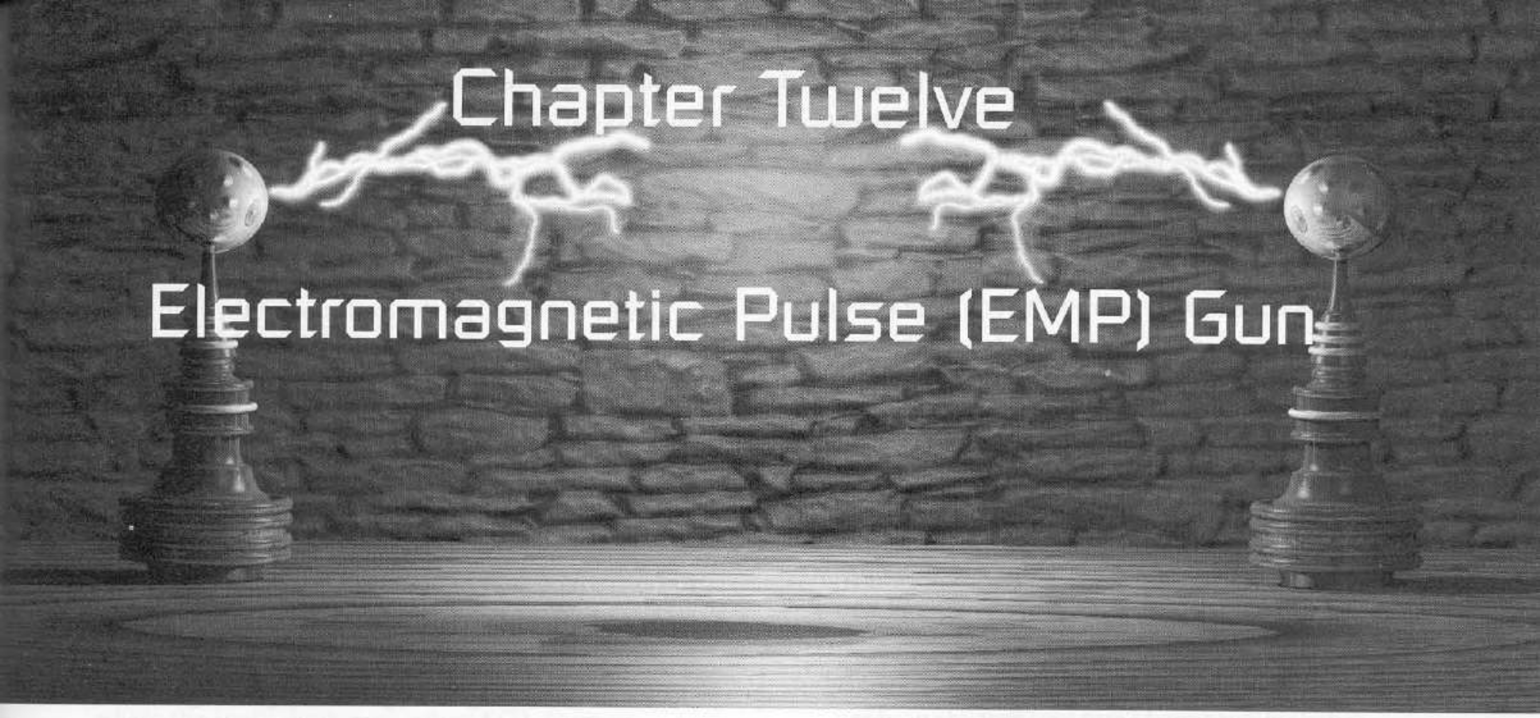

Chapter Twelve

Electromagnetic Pulse (EMP) Gun

Figure 12-1 shows the construction of a low-power pulse gun that will provide kilowatts of peak power at frequencies up to 100 MHz with harmonics. The unit shown is battery powered and uses a modified version of the high-voltage plasma generator shown in Chapter 7 of this book. A higher powered, more functional device is being developed in our labs and will be ready to be copied at the time the third book of this series is available.

This is an advanced level project requiring basic high-frequency electronic skills. A spectrum analyzer can be a very valuable tool in setup but is not necessary. Expect to spend $50 to $100. All parts are readily available, with any specialized parts being available through Information Unlimited (www.amazing1.com) as listed in the parts list at the end of this chapter.

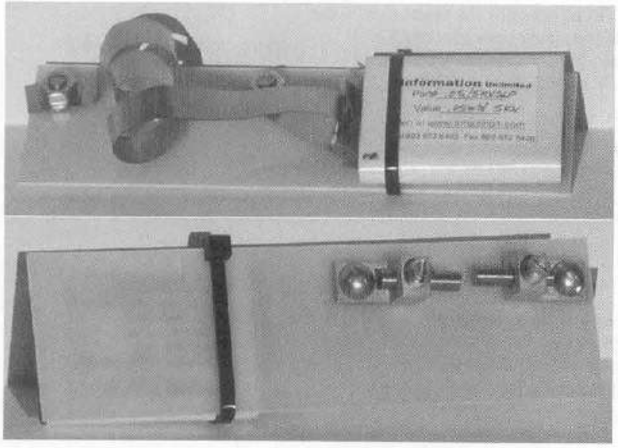

Figure 12-1 *Electromagnetic pulse gun*

Basic Theory Simplified

The ability of a signal to disrupt sensitive circuitry requires several properties. Most microprocessors consist of *field effect transistors* (FET) operating at very low voltages. Once these voltages are exceeded, catastrophic failure becomes imminent. Forgiveness to an overvoltage fault is practically nonexistent due to a microthin metal oxide between controlling elements. Any overvoltage generated across these elements consequently produces permanent damage or, in some less severe cases, causes deprogramming. To generate these damaging voltages from an external source requires a wave that can produce a standing wave of energy across circuit board traces, components, and other key points. The external signal energy therefore must be high enough for the circuitry geometry to be a significant part of energy at the wavelength. Therefore microwaves having fast rise times (high Fourier equivalents) and burst duration help to maximize the effect. The energy required is great, as it must be sufficient to enhance damage. A good measure would be the quotient of energy/wavelength. A high-power pulse of microwaves can be generated in several ways.

Explosive flux compression-driven virtual cathode oscillators and their generic relatives can produce gigawatts of peak power from only hundreds of kilojoules. This is where a seed current is pulsed into an inductor and at its peak is compressed by a shaped explosive charge thereby trapping the flux and creating a source of high energy. The coil must be compressed both along its axis and along its radius using a high-detonation velocity explosive such as the cyclotrimethyltrinitramine, its derivatives PETN, or some other equivalent energetic explosive. This trapped flux now produces an energy gain that is conditioned into the final peak-powered pulse of microwave power (HEPM). Flux compression like that of a nuclear initiation requires precise timing of the explosive chargers. For flux compression, Krytron switches or similar can be used instead of the more radiation-hardened Sprytrons that are used in nuclear initiations where ionizing radiation is produced from the inherent fissionable materials.

A virtual cathode oscillator can also easily be energized from a small Marx impulse generating 200 to 400 kilovolts. The fast current rise and high peak power can produce a powerful burst of microwaves.

Other methods include exploding wires where energy is allowed to flow into a LCR circuit and then is rapidly disrupted by the explosion of the feed wire as it vaporizes at near the peak injection current. A very fast and energetic pulse is produced, which is capable of generating an *electromagnetic pulse* (EMP).

Microwave pulses are excellent candidates for damaging sensitive electronic circuitry. But much lower frequencies are better for disrupting power grids and other similarly sized systems, as now the lengths of the conductive elements are more conducive to generating the high standing voltage waves. Obviously more energy is now required, as breakers, switches, and transformers require more energetic pulses

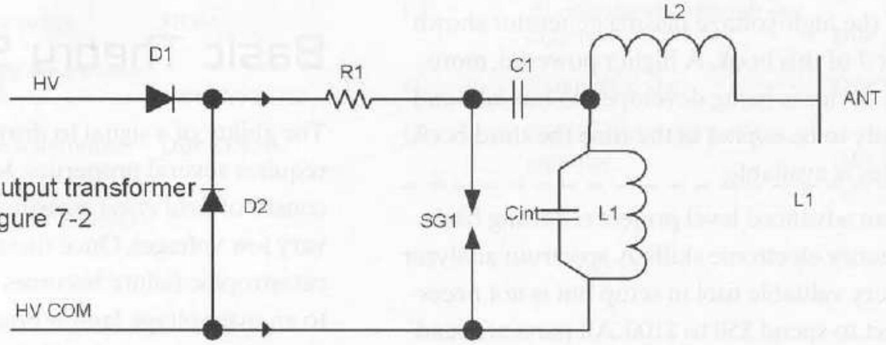

The ideal peak power of this circuit can be approximated by evaluating the product of the charge voltage across C1 x the peak current as determined by Ipk = Epk x the square root of the capacitance of C1/the total inductance of the discharge circuit. This expression implies no resistance (R) in the discharge circuit that now is only in an ideal situation. In all actuality the resistance in the real world will now be a factor in where e to the -(pi/square root of L/C) / R must be a factor in the above ideal condition. This now implies a damped waveform.

Diodes D1,2 are 10 kv 10 ma fast recovery
Resistor R1 is three 47k 1 watt resistors in series and isolate the diodes from the dv/dt of the discharge.
L2 inductor tunes out the capacitive reactance of the antenna at the desired resonant frequency

Capacitor C1 is a "slapper" capacitor. These are used to produce a very fast rise time necessary to detonate initiators necessary for initiating high explosive. Its claim to fame is the high peak current discharge current. The capacitor is constructed as strip lines with the connection leads exiting a common end.

Figure 12-2 *Circuit schematic*

Circuit Description

The circuit shown in Figure 12-2 shows a simple method of obtaining a high-power pulse with Fourier equivalents above 100 MHz. Even though the power and frequency are relatively low, close range effects are possible on many target circuits.

The project uses a high-frequency plasma source, which is converted to a direct current charging source and is short circuited functionally by the use of lossless reactive ballasting. This means that the capacitor can be charged without the use of an energy-robbing resistor, as only the real part of the complex current is seen by the battery supply.

The modified source now supplies a current charge to the reservoir capacitor (C1) to a value where the spark discharges across SG1. Current through L1 rapidly rises and rings along with the circuit and lumped capacitance (Cint). Spark gap SG1 must turn off and allow the energy to circulate in the discharge in order to generate a resonant peak of power that is now coupled into the system emitter. Experimentation of the gap settings is necessary to obtain an optimum effect.

Circuit Assembly

1. Fabricate a piece of Lexan plastic or G10 circuit board, as shown in Figure 12-3, to a 7 × 2 inch plate for PL1. Locate the two $^1/_4$-inch spark-gap-holder (SGH1, 2) screws holes as shown. The holes are spaced $2^3/_4$ inches from one another.
2. Form a three-turn 1-inch diameter coil (L1) from $^1/_2$-inch sheet copper or #14 solid copper wire, as shown in Figure 12-4. Note the leads as attached to the capacitor (C1) and the end gap holder.
3. Connect an output port at the junction of C1 and L1 as shown. This point is now considered the output and can be connected to the radiating emitter.
4. Connect it to the converted plasma driver in the manner shown in Chapter 7 to the generator (see the schematic Figure 12-2). Note the added diodes D1 and D2 for converting the output to direct current
5. You will note a coil (L2) connected in series with the output lead. This inductor tunes out the capacitive reactance of the lead and the end capacitance. Experiment using a radio wave or absorption wave meter to determine the resonant frequency of L1/C1. Select a value for L2 to provide maximum radiation from a distance.
6. Experiment on various electronic devices and observe the effects at various distances.

Notes

A low-cost spectrum analyzer would be a great aid in setting up this system. Please note this is a low-power device intended as an introductory project for those desiring to experiment with shock pulse and EMP research. Several more functional and sophisticated devices will be featured in the next book in this series.

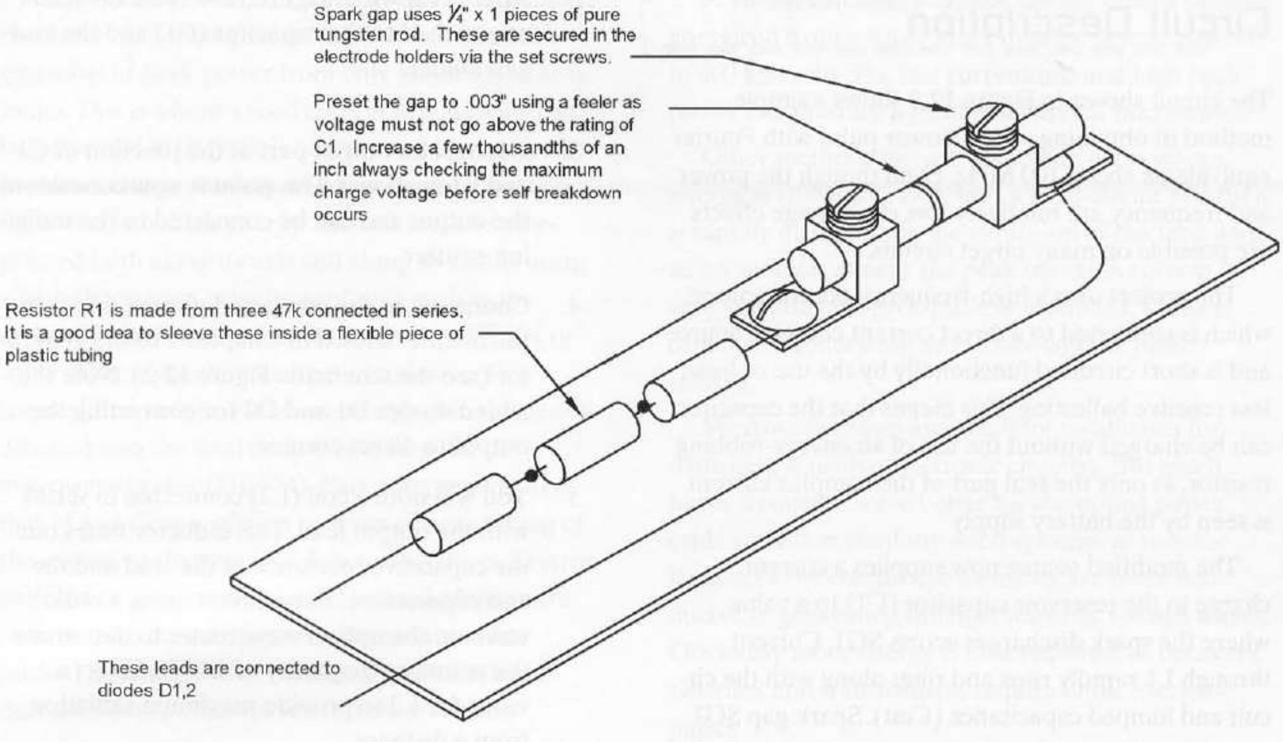

Spark gap uses ¼" x 1 pieces of pure tungsten rod. These are secured in the electrode holders via the set screws.

Preset the gap to .003" using a feeler as voltage must not go above the rating of C1. Increase a few thousandths of an inch always checking the maximum charge voltage before self breakdown occurs

Resistor R1 is made from three 47k connected in series. It is a good idea to sleeve these inside a flexible piece of plastic tubing

These leads are connected to diodes D1,2

Figure 12-3 *Generator board bottom view*

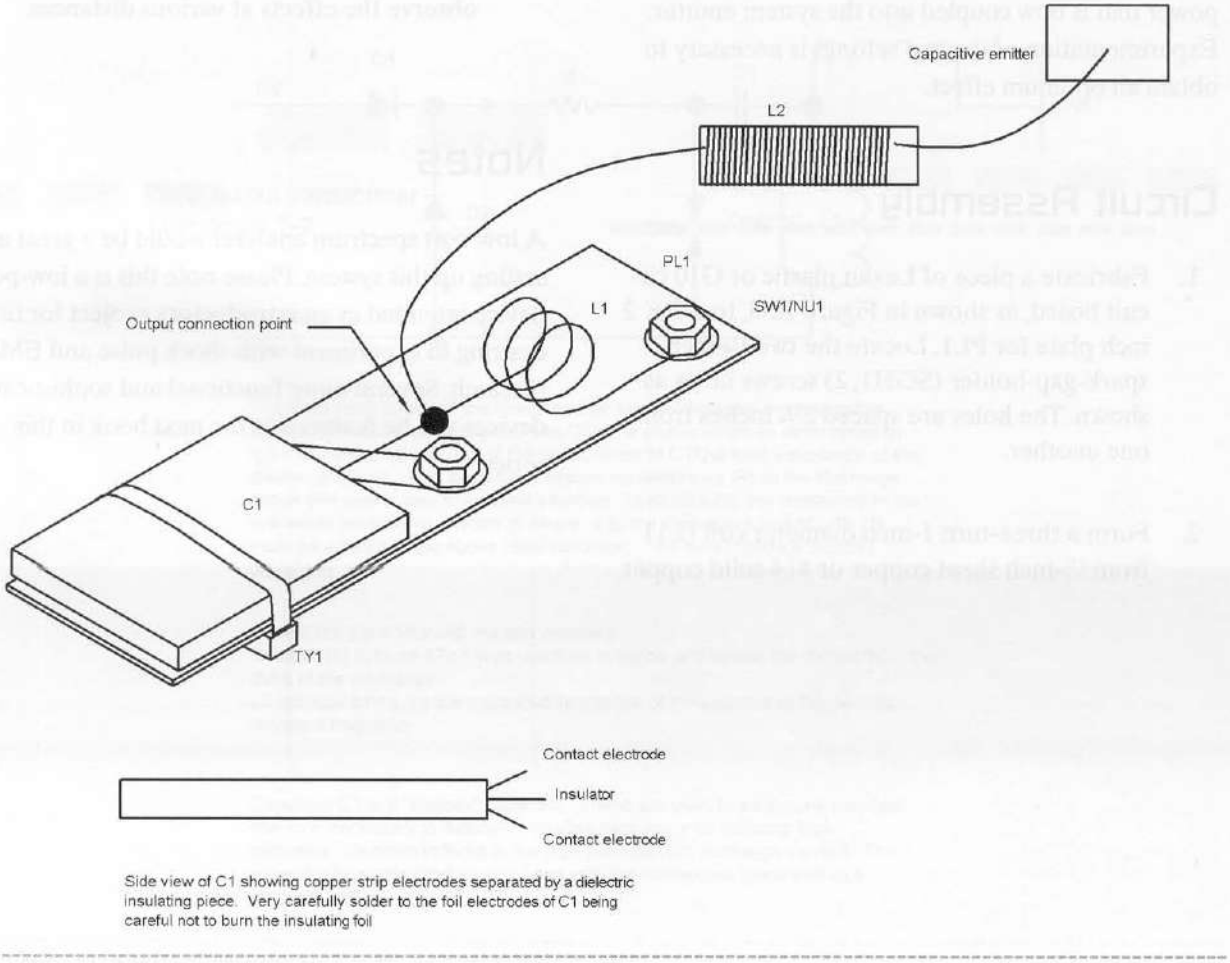

Capacitive emitter
L2
PL1
L1
SW1/NU1
Output connection point
C1
TY1
Contact electrode
Insulator
Contact electrode

Side view of C1 showing copper strip electrodes separated by a dielectric insulating piece. Very carefully solder to the foil electrodes of C1 being careful not to burn the insulating foil

Figure 12-4 *Generator board bottom view*

Table 12-1 EMP Gun Parts List

Ref. #	*Description*	*DB part #*
R1	Three 47K, 1-watt resistors (yel-pur-or) connected in series	
D1, D2	Two 16 KV, 10 ma fast-recovery high-voltage rectifiers	DB# VG16
C1	.05 mfd, 5 KV stripline capacitor	DB#.05M/5KV
L1	Inductor 3-turn, 1-inch diameter use #12 copper wire	
L2	Inductor wind, as directed in text	
ELECTRODES	Two ½ × 1 inch pure tungsten electrodes	DB# TUNG141B
PL1	7 × 2 inch × .063 Lexan or G10 plastic	
TYE1	10-inch nylon tie wraps	
SW1/NU1	Two ¼ × 1 inch brass screw and nuts	
GRAVDRIV1K	Plasma driver kit	DB# GRAVDRIV1K
GRAVDRIV10	Plasma driver assembly	DB# GRAVDRIV10

Chapter Thirteen

Microwave Cannon

This project shows how to build a powerful microwave transmitter intended to be a useful laboratory resource or as a device for disrupting bothersome stereos, boom boxes, and other annoyances perpetrated by inconsiderate hip-hop miscreants. This chapter was prepared by Dr. Barney Vincellete, who holds several doctorate degrees in the electrical sciences. It is a very useful tutorial involving the design of a microwave system, going into basic mathematical details for those who are interested. The system can be built from the parts of a microwave oven.

Warning: The project uses dangerously high voltages and can be a radiation hazard at close range if not used with protective shielding. This is an advanced project intended for experienced builders and must not be attempted by anyone unfamiliar with high voltages or high-power radio frequency circuitry.

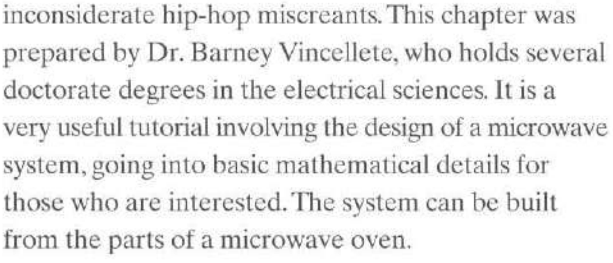

Figure 13-1 *Microwave cannon*

The microwave cannon consists of three parts: the parabolic reflector horn, the electric field focusing lens, and the magnetron. The horn focuses the magnetic oscillations that circumscribe the magnetron antenna into a narrow beam in order to concentrate the radiation into a two-dimensional space. It also prepares the electric field oscillations for the lens to concentrate the radiation in the third dimension. Because the electric field and magnetic field are orthogonal, calculations upon them and the modifications that we will perform upon them can be done on one without altering the other.

In designing the horn, we will assign the Cartesian coordinate system with the z axis along the longitudinal axis of the horn; that is, the z axis will occupy the center of the shaft of radiation that will be fired into your cultural tyrants' stereos. The Poynting vector will be on the z axis, and the x axis will be directed horizontally to the left and right of the z axis. The x axis will, during the calculations for the shape of the horn, be given an absolute value; that is, it will be endowed with a positive value whether it increases to the left or right of the z axis. This is because its values will be the square root of x squared, and it will be symmetrical to the left and right of the z axis. The origin will be located at the vertex of the parabola that will define the sides of the horn.

The y axis will be in the vertical direction to complete the basic vectors in our discourse and calculations. It will pass through the z axis at the mouth of the horn, rather than through the vertex, as the x axis does. It too will have only positive values because, due to its symmetry about the z axis, there is no need to consider negative values of y (see Figure 13-2).

The dimensions of the sides of the horn will be performed first. In building the wooden mold upon which the sheet copper will be formed, an adjustment will be made to correct for the flare that can be seen in Figure 13-3.

The horn will be made of 20-inch-wide sheet copper that can be purchased from a roofing supply house. Because the dimensions of the parabolic sides shown in Figure 13-2 are fixed by the top and bottom pieces of copper that will be soldered to the copper side pieces, the top and bottom pieces will be cut to a parabolic shape with an extra 1/4-inch margin on the sides that will be bent into a ridge that will be soldered to the side piece. The actual parabola will have a maximum width of 19.5 inches, or the maximum value of x will be 9.75 inches. Also, for structural rigidity, a half-inch lip will be bent along the front edge of the top and bottom pieces.

The parabola will contain the magnetron antenna aligned parallel to the y axis and positioned at the focal point of the parabola. The focal point will be one-eighth of the wavelength of the magnetron frequency, which, when the wavelength is computed by dividing the speed of light by the frequency or 2.45 GHz, gives us a wavelength of 2.45 inches and a focal length of 0.6 inches. The parabola is described by the equation x squared = 4 pz, where *p* equals the focal length. When *p* equals .6 inches, *r* equals 2.4 z. From this, we construct Table 13-1.

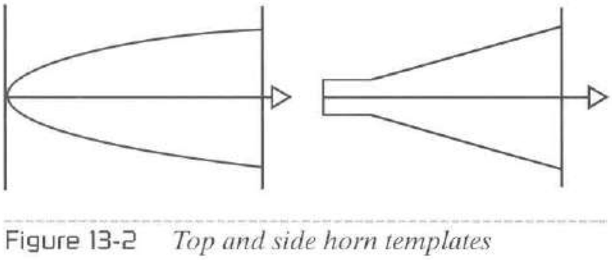

Figure 13-2 *Top and side horn templates*

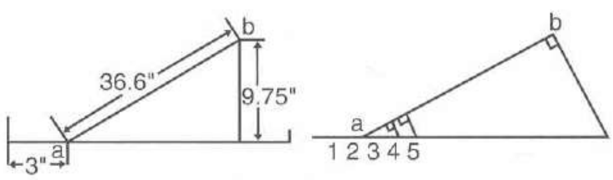

Figure 13-3 *Flair angle template*

Table 13-1 Parabolic Dimensions (measurements in inches)

z	*x*	*z*	*x*	*z*	*x*	*z*	*x*	*z*	*x*
.125	.55	3	2.68	8	4.38	16	6.2	29	8.34
.25	.77	4	3.1	9	4.65	18	6.57	32	8.76
.5	1.1	5	3.46	10	4.9	20	6.93	35	9.17
1	1.55	6	3.79	12	5.37	23	7.43	38	9.55
2	2.19	7	4.1	14	5.8	26	7.9	39.6	9.75

When regarding the side view of the horn (Figure 13-2), it can be seen that plotting the parabola, which is done by cutting the wooden sheets used for making a mold of the horn, will require an adjustment so that the sides of the parabola are not distorted where z is greater than 3 inches. The dimensions of z, if projected from the z axis to the top and bottom of the horn where z is greater than 3 inches, will have to be enlarged by dividing each inch by the direction cosine of the angle between the z axis and the top (also the bottom) of the horn. This can be done by drawing upon a sheet of 2 by 4 foot particle board, as shown in Figures 13-3.

After you have marked the board, as shown in Figure 13-3, you will begin plotting dots measured above and below the z axis at distances of x from Table 13-1 measured with a carpenter's square. You will connect these dots with a ruler, and when you cut out the parabola, the saber saw that you use will make a curved path that will very closely approximate the parabolic shape needed. After you cut out the parabola, you will make another cut along the locus of points where z equals 3 inches.

You will trace these pieces and cut duplicates from another piece of particle board, as shown in Figure 13-4. With nails, screws, glue, and two trapezoidal-shaped pieces of particle board, you will put together the form that you will use to build the copper horn (see Figure 13-5).

Stand the wooden form on its end, where z is 39.6 inches against the floor, and make a template from poster paper that has been taped together to form large enough pieces. Make a template that will wrap about the sides, and trim it to fit so that it matches the sides. Make another template for the top (also the bottom) piece the same way, as shown in Figure 13-6.

You will need 15 feet of copper roofing sheet metal to cut the pieces that you will solder together. Trace and cut the long piece from the template you made from the sides and rear of the wooden form. Add a half-inch to the length of each side so that you can bend a lip that will make the front of the sides rigid. This lip will also provide a place where you can attach the sides of a wooden frame that will make the horn more stable. Then trace the top and bottom pieces against the copper and draw a 1/4-inch additional margin surrounding the parabola you just traced. Also add a half-inch to the front that you can bend into a lip to make the top and bottom more rigid. Cut out two of these pieces, one for the top and one for the bottom.

With a pair of pliers, bend a 90-degree lip from the 1/4-inch margin you traced. You can cut notches in this lip to make this bending easier. This lip will be soldered to the piece that will form the sides of the horn. Bend the half-inch lip at the front edges of the pieces that you will solder together, and clean the insides of the lip that you will be soldering to the sides of the horn using steel wool or sand paper. Clean the exterior 1/4-inch sides of the horn to which you will be soldering the top piece. Apply a thin coat of soldering paste and clamp the sides and top securely into place on the wooden form you built. Hammer the side lip against the sides of the horn for a tight fit. You do not need a sharp crease where the horn begins to flare out where z is 3 inches; a modest bend will do. You are now ready to begin soldering the seam, as shown in Figure 13-7.

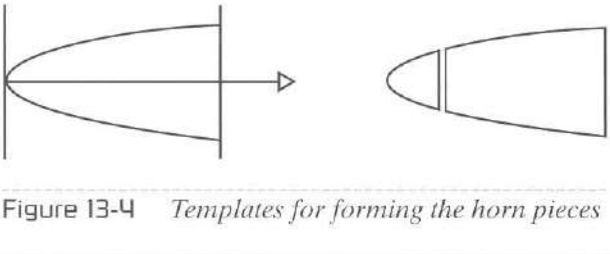

Figure 13-4 *Templates for forming the horn pieces*

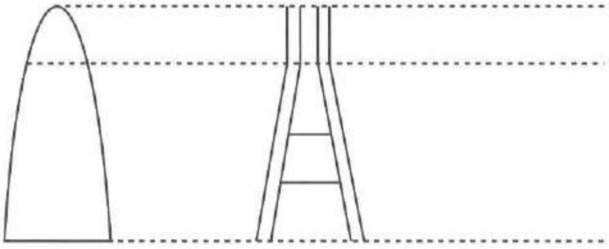

Figure 13-5 *Poster paper template*

Begin with a propane torch and plumber's solder. Apply the flame to a spot and, as soon as it is hot enough to draw in some solder, let just enough solder draw into the seam to fill the gap. Use more hammer taps to close any open spots where the lip is not tightly against the side piece. As soon as you have an inch or two soldered, remove the flame and solder another place a few inches away. Do not use enough solder to cause it to form drips or beads on the inside of the horn. Continue until the entire seam is soldered. After this is done, remove the clamps and sheet copper structure you soldered together. If you see any lumps of solder on the interior seam, you can melt it away by applying a flame.

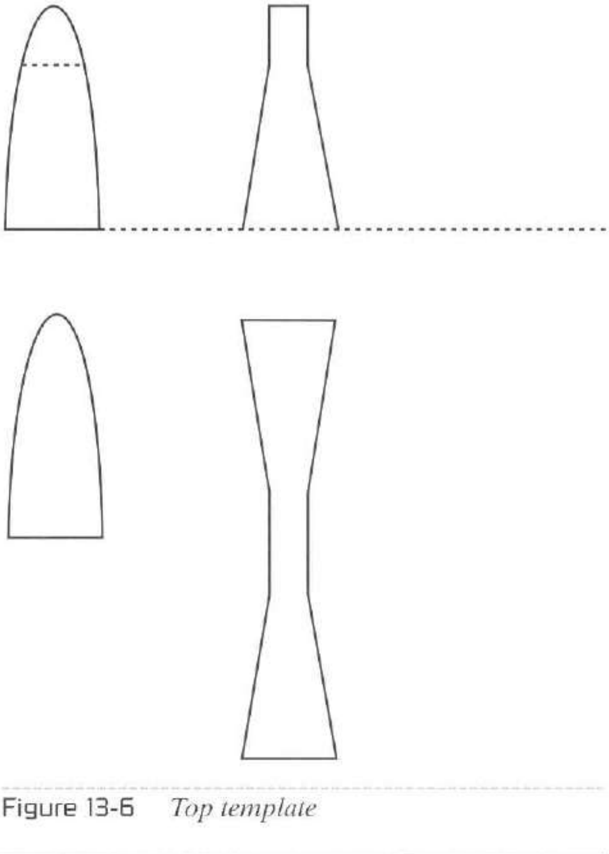

Figure 13-6 *Top template*

Now you must locate the point on the center axis at the top of the horn. This can be done by measuring the center of the front of the top section of the horn, measuring 2 feet along each side of the horn's interior, drafting a line between these points that are 2 feet back from the front, and bisecting it. You bisect the width of the front to find one point. You then measure 2 feet back on each side and draw a line between the ends. You bisect the line for the second point. This will provide two points to extend back to the vertex of the parabola's interior.

Now measure, as punctiliously as you can, 0.6 inches from the vertex. Drill a pilot hole and cut a circle that is $1^{3}/_{16}$ inches in diameter. With a hammer and a piece of pipe, tap on the interior of this hole so that its edges flare outward to make a good seal with the brass wool gasket on the magnetron you will be using (see Figure 13-8).

At this point, solder the bottom piece into place and complete the construction of the copper horn. A wooden frame must be built to make the front edges of the horn more rigid, to provide a structure for mounting the lens, and to provide a place for the front of an aluminum rail where a telescopic rifle sight will be mounted.

Here you have great artistic liberty in how you will build this cabinetry. Half-inch plywood is the most practical material to use, and cedar has the most pulchritudinous finish to stain, varnish, and paint. You can paint or carve artwork upon this frame, and you can also upholster the rear or sides with velvet and sunken buttons. In the most unlikely event the cannon becomes an exhibit in court, these characteristics will humor the jury, many of whom will be as hostile as you against the music the neighbors were forcing into your home. If it becomes newsworthy enough to attract a television crew, it will be a splendid show for the millions of people whom you will inspire to follow your example and join the revolution against cultural tyranny.

The rear of the frame will protect the operator from the deadly 4,500-volt wires between the power supply and the magnetron. It will also provide a place to mount the rear of the aluminum rail for the telescopic rifle sight.

Let us digress to a disquisition on the flair angle formed by the sides of the cannon. In order to form the best possible cancellation of standing waves, due to the impedance vicissitudes where the horn begins to flair at the mouth of the horn, the following equation must be emulated within approximately 10 percent. The dimensions of the horn must match this within a percent.

$$0.25 = [a/(2\lambda)]\tan(\theta/2)$$

where a is the dimension of the mouth in the horn in the y direction: 20 inches. λ is the wavelength in

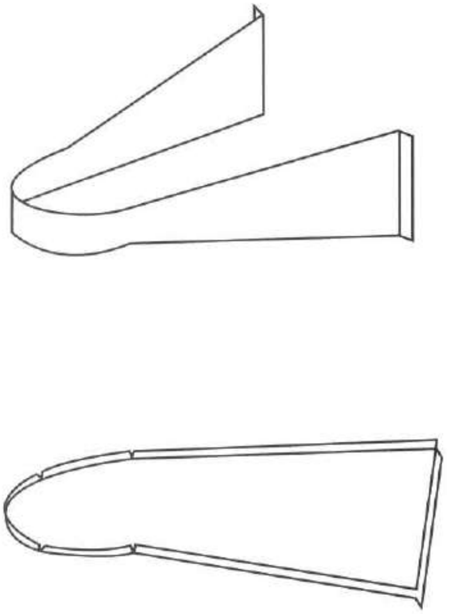

Figure 13-7 *Copper side with top and bottom pieces*

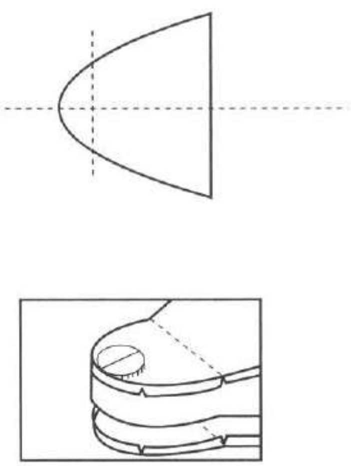

Figure 13-8 *Position of aperture hole for magnetron probe*

inches. θ is the angle between the z axis and the top (or bottom) side of the horn.

Let us now examine the efficiency of the parabolic sides of the horn for their focus of the magnetic field. The antenna is circumscribed by a magnetic field that propagates from a line source at the center of the antenna. The radiation either reflects off the sides of the parabola formed by the horn or it radiates directly from the antenna out the mouth of the horn. The latter is not useful because it dissipates rapidly with distance from the cannon. The former, the reflected radiation, is projected parallel to the zy plane, and because it originates from a line source parallel to the y axis, no diffraction occurs off the sides of the horn's mouth. The fraction of the radiation that is reflected is 92.2 percent of the total from the antenna. This is far superior to the typical 40 percent efficiency given by radar dishes that use large sizes to achieve gain.

The focusing of the electric field can now be performed by placing a lens in front of the cannon's mouth, and a description of the electric field must be provided. On the surface of a conductor there can be no voltage parallel to the surface of that conductor, because any voltage would be short-circuited out. This makes the components of the gradient of any voltage parallel to that surface zero. Since the electric field is the negative gradient of the voltage, the electric field near a conducting surface must perforce be perpendicular to that surface. As the electric field propagates forward, it forms a cylindrical-shaped wavefront in the flared part of the horn. At the mouth of the horn, this cylinder has a radius of 41 inches. A lens will convert this cylindrical wavefront to a flat wavefront, and that will complete the focusing of the radiation.

The lens will be constructed of metal plates that will be aligned parallel to the y and z axis 4 inches apart. Between these plates, any wave vector, $k = 2\mathrm{n}/\pi$, can be the vector sum of two components, kx and kz. In the x direction, the sine wave of the wave vector will be zero every 4 inches, because at the plates the electric field in the y direction must be zero. Taking recourse to the Pythagorean theorem, the wavelength in the z direction is dilated by the inverse of the refraction index, that is, the refraction index, $n = [1-(\lambda/8 \text{ inches})\mathrm{E}2]\mathrm{E}-.5 = 0.8$ ($\lambda = 4.8$ inches). See Figure 13-9.

The cylinder formed by the cylindrical-shaped wavefront plus the refraction index times the distance parallel to the z axis between the lens elements to a flat plane in front of the lens is to be constant so that the electric field will be in phase on the flat front of the lens. The y axis is shifted to the most concave depth of the lens elements, and the z axis is directed rearward. The algebra for calculating the concave curve to cut into the rear of the lens plates is rudimentary. From the following equation, the dimensions of the lens are shown in Figure 13-10.

$$[(47'' - z)\mathrm{E}2 + y]\mathrm{E}.5 + 0.8\,z = 47'' \text{ from which } z = \{18.8 - [(18.8)\mathrm{E}2 - (1.44y)\mathrm{E}2]\,\mathrm{E}.5\}\,\{1/.72\}$$

The lens plates mounted against the far right and far left sides of the lens frames should be cut from leftover sheet copper and glued to wooden sides, also cut to the shape of the lens element. The other lens elements should be made from 0.025-inch sheet aluminum. They should be 21 inches long in order to mount them to the wood lens frame you will build. Small wooden blocks can hold the ends of lens plates in place. Pieces of wood with grooves cut into them will also work.

The efficiency of the lens can also be calculated. This efficiency is due to the reflection of radiation that takes place when radiation passes from a medium of one refraction index into a medium of a different refraction index where the space impedance is changed. Two reflections occur, one when the radiation enters the rear of the lens and a second reflection when the radiation exits the front of the lens. The second reflection is in opposite phase with the first reflection, so it is 2π radians out of phase with the first reflection. The appropriate Fresnel equation for calculating the power reflected when the electric field

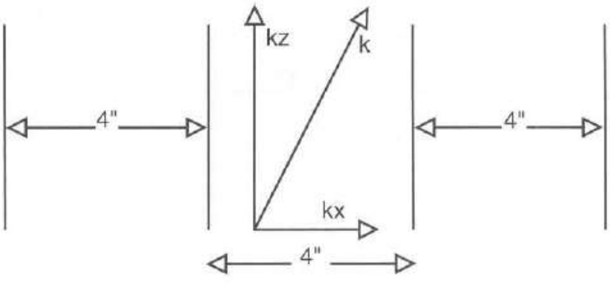

Figure 13-9 *Focus plate geometry*

is parallel to the plane containing the angle of incidence, θi, and the angle of transmission, θt, is as follows:

$$R = \{[-\cos\theta\mathrm{i} + (\mathrm{n1/n2})\cos\theta\mathrm{t}] \div [\cos\theta\mathrm{i} + (\mathrm{n1/n2})\cos\theta\mathrm{t}]\}\mathrm{E2}$$

θt is arctan dz/dy, as can be seen in Figure 13-11. θi is θt − arctan $[y/(47 - z)]$.

In the second reflection, θi and θt are both zero. n1 is 1 and nt is 0.8. The second reflection that must be subtracted from the first one is cos $[2\pi$ radians $2(z + .25'')\lambda z]T$, where T is the fraction of the power transmitted into the lens and λz is 4.8 inches/0.8. Since the total reflected power is almost everywhere less than a percent, we can make the approximation that T equals 1. If the reflected power is calculated every inch for y, the average power reflected is approximately half a percent. Thus, the gain of the horn is

$$4\pi\mathrm{A}/\Lambda\mathrm{E2} = 218 \text{ or } 23.4 \text{ db}$$

With the ideal horn and lens system having been designed so that the common man can build it without having to know all the electromagnetic laws of physics that make it work, we direct our attention to the choice of a microwave source. The cardinal rule of electronic warfare is that the radiation must convey noise in order to disrupt the signals in the circuits you will be attacking. This requires modulation that will drown out enough of the unwanted music and interfere with the digital signals in disc players to make them useless. Half-wave rectification of the current through the radiation generator is subjectively the most disrupting. The lower frequencies promoted by stereo manufacturers and music producers seem to be designed to most effectively penetrate the walls of apartments and homes for the purpose of destroying the privacy and cultural autonomy of all but the wealthiest who can afford the enormous cost of solitude.

The 60-cycle voltage from a household wall socket offers a nearly ideal source of modulation to create this disruption, because in a microwave oven it is half-wave rectified by a voltage-doubling circuit. This provides a splendid spectrum of Fourier frequencies that concentrate upon the lower frequencies but distribute themselves upward through the audio range, and it gives the utmost modulation of amplitude. It also reduces the duty cycle of the magnetron while preserving the maximum power in each pulse it shoots into the neighbor's stereo for the best range.

The first inclination is to choose a magnetron that will have the highest output power. The best choice would be a radar magnetron, which might offer 30 to 100 kilowatts in pulses that are only a microsecond or two in duration. However, tests using radar against a compact disc player yield disappointing results. The Fourier series expansion on a series of microsecond-long pulses modulates less than 1 percent in the audio band. The series of ticking sounds that it introduces into a stereo amplifier is scarcely noticeable, and not enough pulses occur to put sufficient errors in the digital circuits of the compact disc player to do any good. Also, audio circuits are not as sensitive to the

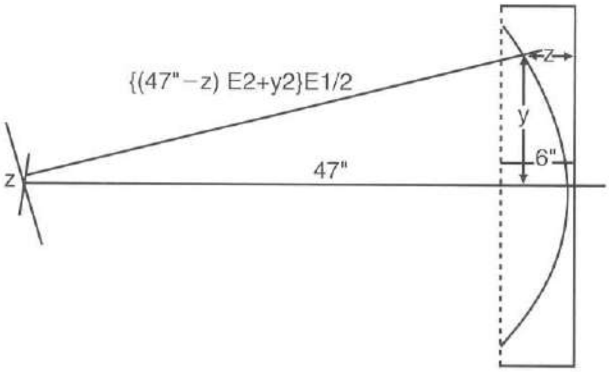

Figure 13-10 *Wavefront and electric lens geometry*

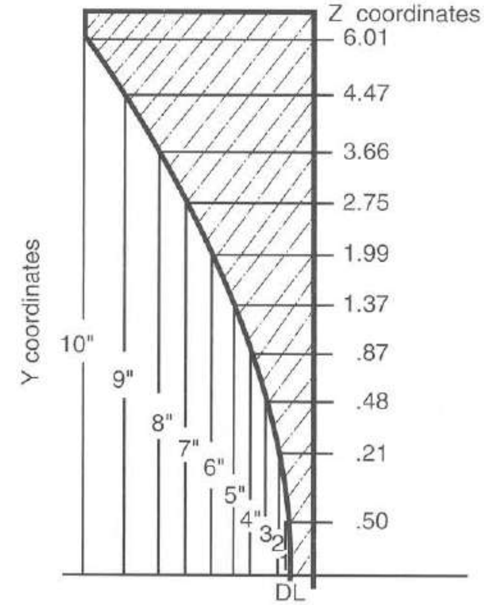

Figure 13-11 *Calculated lens dimension*

higher frequencies of radar as they are to microwave oven frequencies.

We might also be inclined to consider the higher-powered magnetrons used in industrial heaters and plasma generators, but this can be problematic. First, higher power is achieved more by increasing the duty cycle than by increasing the peak output power we desire. The 3,000-watt Panasonic 2M265 magnetron has a 7-kilowatt peak output power, but it runs on pulse-modulated direct current for a longer duty cycle than the half-wave voltage doubler of a microwave oven. Its operation is more critical; the cathode heater must have its voltage adjusted to the power output, and the cathode is more brittle in these more powerful magnetrons. It also costs as much as $1,000 and the power supply is even more expensive. The 6,000-watt Toshiba E3328 magnetron runs on filtered direct current and it has even more expensive computer controls that adjust an electromagnet to control the current. The peak output power is less than that of the 3,000-watt magnetron.

Another problem occurs with these magnetrons as well. They are so unforgiving of such things as a large metal object near the cannon in which they might be installed that reflections could mismatch the impedance enough to shorten the service life. The manufacturers recommend that they have a branch off of a waveguide to a dummy load that will absorb any energy due to such an impedance mismatch in order to protect the magnetron. This makes the parabolic horn impossible to use because it is not practical to feed with a waveguide. A pyramidal horn would be needed and it only has two-thirds of the gain of the parabolic horn. Thus, the 6,000-watt magnetron would be as good as a 4,000-watt one in a parabolic horn, and the 3,000-watt magnetron would be as good as two-thirds of 7,000 watts in a parabolic horn. Several vastly cheaper household oven magnetrons can easily outclass the more expensive and difficult-to-operate choices that appear more attractive until they are examined with greater perspicacity.

Another difficulty with purchasing industrial magnetrons also exists. Companies will not sell them to private individuals out of fear that they might be used for negative or hateful enterprises, such as in machines designed to sabotage overly loud stereos. Although almost any magnetron and power supply from any microwave oven will work, the best oven magnetron for our purposes is the 2M121A, which has its number conflated with the number 53 or 57. These last two numbers make no difference in the choice of editions of the 2M121A you procure. They are used in the following ovens: the Panasonic 1030, the Brother MF5000 and MF7000, and the Merrichef models 136M, 165M, and 206M. It has a peak output power of 5,800 watts, better than the 2,000-watt magnetrons that cost more than twice as much. It can be ordered from the Expert Appliance site (www.expertappliance.com) under the part number Z9-Panasonic NEI 030-412 if you type in Panasonic NEI 030 for the type of oven for which you are purchasing parts. The transformer number is 210 Panasonic NEI 030-5414. Global Microwave Parts has voltage-doubling capacitors that are 0.85 microfarads rated at 2,500 volts AC and diodes. You should use one rated at half an ampere. Magnetrons and parts are available and listed in Table 13-2. You will need copper and aluminum sheeting, along with mat-erials available from hardware or building suppliers.

The magnetron is attached to the antenna, as shown in Figure 13-12. Follow the instructions closely and observe that the brass washer is flush with the copper conductive surface.

The power supply is a simple voltage-doubler circuit, shown in Figure 13-13. It must be understood that the voltages in the wires will be in the thousands and can produce a spark that can jump through the air and through an insulator, conveying enough current that it would probably cause electrocution. Accordingly, the wires should be positioned where they will not be approached or touched. The magnetron and horn must be securely grounded or it will be 4,500 volts above ground potential. The frame that you must build around it will keep people's hands away from these wires. Also, after the cannon is shut off, four of the capacitors can hold a charge that would inflict a shock unless it is discharged, either through a bleeder resistor that may be built into the capacitor or it can be discharged with a safety discharge probe.

The magnetron must also have a fan to cool it during operation. The 2M121A requires at least 85 cubic feet per minute of air directed through its fins for best operation.

When your microwave cannon is completed, testing it is little more than a matter of switching on the power and aiming it at a fluorescent tube or a small 120-volt lightbulb. The lightbulb should be no more than 25 watts and should glow slightly when it is within a few feet of the mouth of the cannon. Fluorescent lamp tubes should light up as far as 10 feet away.

You should not put any body parts directly in front of the cannon at close range because, after a few seconds to a few minutes, you could burn yourself, but a distance of over 50 feet away would provide safety. However, prolonged overheating of internal organ tissue is remotely possible and the cornea of the eye can be damaged by overheating. Also, testicles can overheat, and at close enough range for the *root mean square* (RMS) microwave field to exceed 4,000 watts per square meter, a 5-minute uninterrupted exposure is supposed to build up enough internal steam pressure to cause the testicles to explode. However, the possibility of having a hot spot this intense a few inches in front of the cannon is marginal.

When testing or using your microwave cannon, you can do so with utmost safety to yourself by constructing a gown out of metal screen from a hardware store. The gown should include a hood that completely covers the head and should extend over the torso down to just below the vaudevillian parts of the body. The sex organs should be stuffed into a tin can, the mouth of which can be lined with fake fur to make it more comfortable. (Obviously, the ladies need not take recourse to this expedient.) This is more than enough to prevent any harm when performing experiments that last half an hour or more and is probably hundreds of times the protection you will need.

It can be advantageous to test the cannon in your backyard. Sooner or later the neighbors will see it, and if they see a mysterious-looking machine that lights up fluorescent tubes 10 or more feet away and they see you dressed in a wire screen gown complete with a tin can, it will frighten the emunctory indiscretions out of them. Further, when they complain to the authorities they will sound much the same way mental patients sound as they describe exotic macrunes, stereos malfunctioning, strange costumes complete with tin cans, and the like. It is even slightly possible

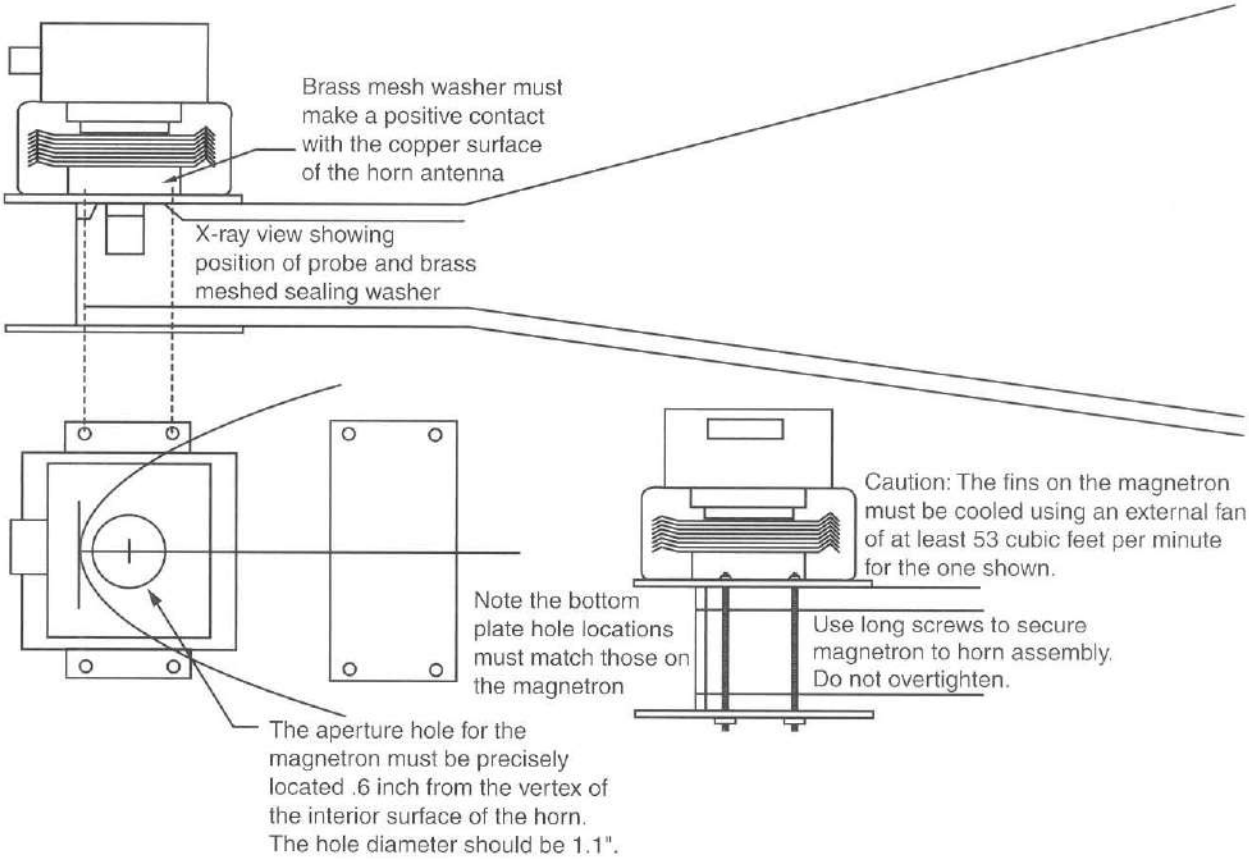
Brass mesh washer must make a positive contact with the copper surface of the horn antenna
X-ray view showing position of probe and brass meshed sealing washer
Caution: The fins on the magnetron must be cooled using an external fan of at least 53 cubic feet per minute for the one shown.
Note the bottom plate hole locations must match those on the magnetron
Use long screws to secure magnetron to horn assembly. Do not overtighten.
The aperture hole for the magnetron must be precisely located .6 inch from the vertex of the interior surface of the horn. The hole diameter should be 1.1".

Figure 13-12 *Attaching the magnetron*

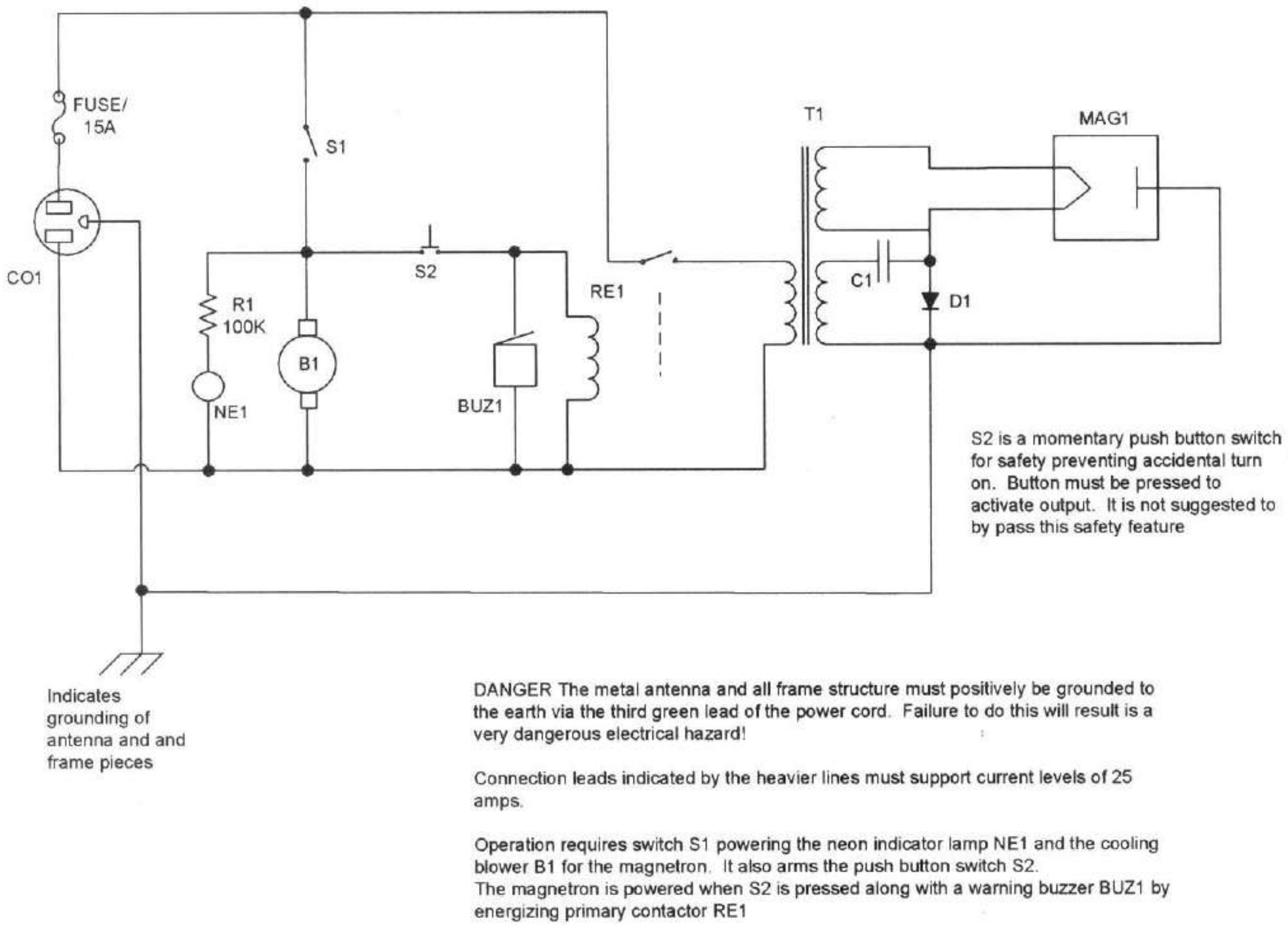

S2 is a momentary push button switch for safety preventing accidental turn on. Button must be pressed to activate output. It is not suggested to by pass this safety feature

Indicates grounding of antenna and and frame pieces

DANGER The metal antenna and all frame structure must positively be grounded to the earth via the third green lead of the power cord. Failure to do this will result is a very dangerous electrical hazard!

Connection leads indicated by the heavier lines must support current levels of 25 amps.

Operation requires switch S1 powering the neon indicator lamp NE1 and the cooling blower B1 for the magnetron. It also arms the push button switch S2.
The magnetron is powered when S2 is pressed along with a warning buzzer BUZ1 by energizing primary contactor RE1

Figure 13-13 *Microwave cannon schematic*

that they might be invited to undergo observation at the local state psychiatric hospital, where, if you are really ambitious, you could locate their room, aim your microwave cannon, and make their stereo buzz whenever the psychiatrist's back is turned and cause a diagnosis that will keep them out of the neighborhood for a longer time. But in the remote case you get caught doing this, it will be almost impossible to prosecute you because no state legislature has the time to write and process special laws against shooting microwaves into peoples' stereos. The television publicity possible if the state attempts to prosecute you will be a disaster for the system that supports cultural tyranny because it will inspire countless television viewers to join in the revolution to make the world a better place by building more microwave cannons and shooting down more obnoxious stereos. And even if they can win a battle in court, they will ultimately lose the war that they started against human dignity.

Table 13-2 Microwave Cannon Parts List

Ref. #	*Description*	*DB Part #*
R1	100K 1/4-watt resistor (br-blk-yel)	
CO1	Heavy-duty, 15-amp, three-wire power cord	
FUSE/15	15-amp fuse and holder	
B1	85 cubic feet/minute or better cooling fan for magnetron fins	
S1	*Single pole, single throw* (SPST), 5-amp, 115 VAC toggle switch	
S2	Momentary push-button switch (see note on Figure 13-13)	
NE1	Neon indicator lamp with leads or equivalent	
BUZ1	120 VAC high output safety buzzer	
RE1	10- to 20-amp contacter or relay	
T1	Magnetron power transformer NEI030-5414	DB# MAGTRAN
C1	.85-microfarad, 2500 VAC voltage-doubler capacitor	DB# MAGCAP
D1	5 kilovolt, 1/2-amp diode	DB# MAGDIODE
MAG1	2M121A magnetron	DB# MAGNETRON

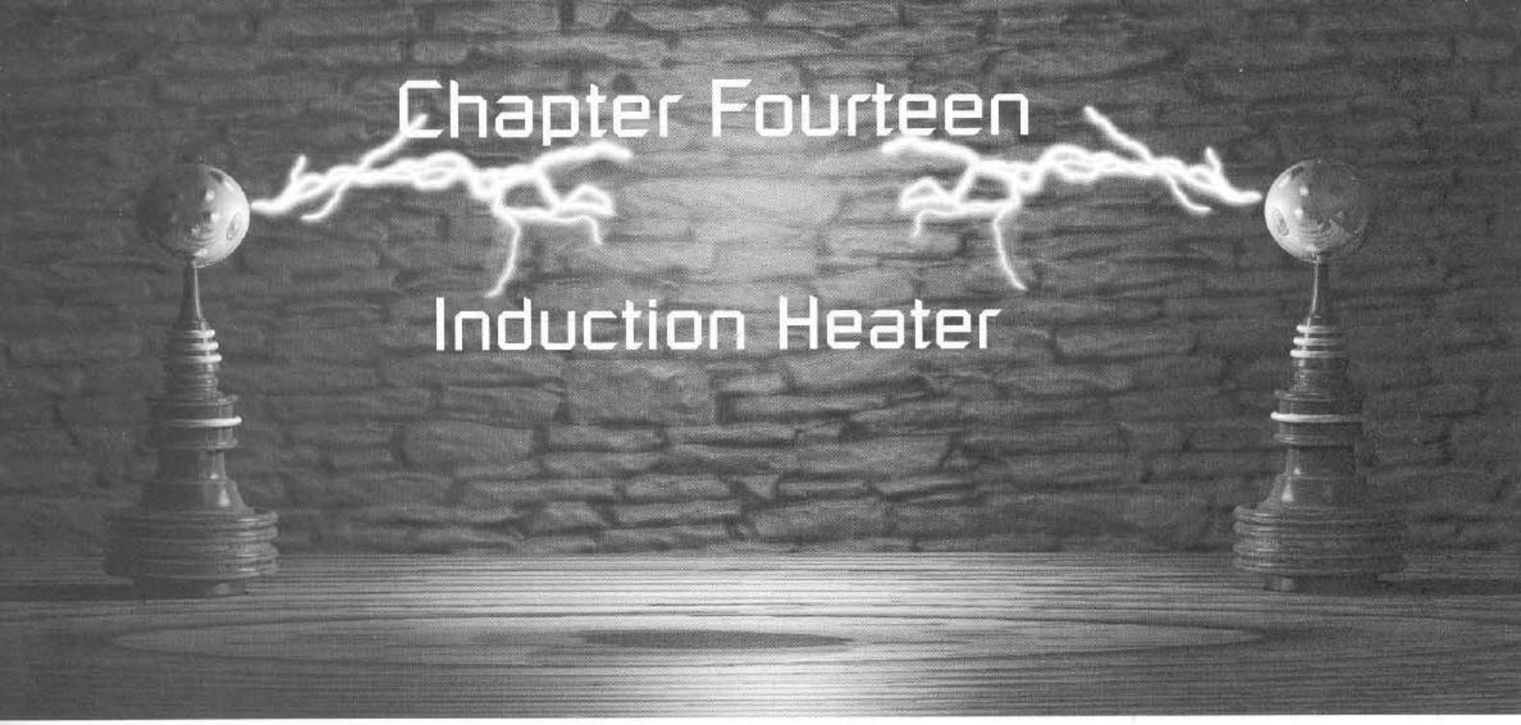

Chapter Fourteen

Induction Heater

This device shown in Figure 14-1 shows an induction heater capable of heating metals to over 1,500 degrees Fahrenheit. It can be used in the processing of gas discharge tubes often used for plasma, neon displays, laser discharge tubes, and spectrum tubes. These tubes usually require electrodes that must be heated to drive out the impurities that are now sucked out by a vacuum system. The induction heater heats these electrodes directly through the glass.

The completed system consists of a donut-shaped heating coil head connected via an umbilical cord, which is in turn connected to the power-conditioning and control box. The system is powered directly from 115 VAC and requires caution when building and using. The head is a coil of wire that surrounds the target area to be heated. As shown, it is an air-cooled system that requires a waiting time for processing between operations because the coil will get hot. The common circuit ground is above ground and requires an isolation transformer or an ungrounded scope to view the waveforms from a grounded scope.

This is an advanced-level project requiring electronic skills and the use of an oscilloscope. Expect to spend $35 to $75. All the parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 14-1.

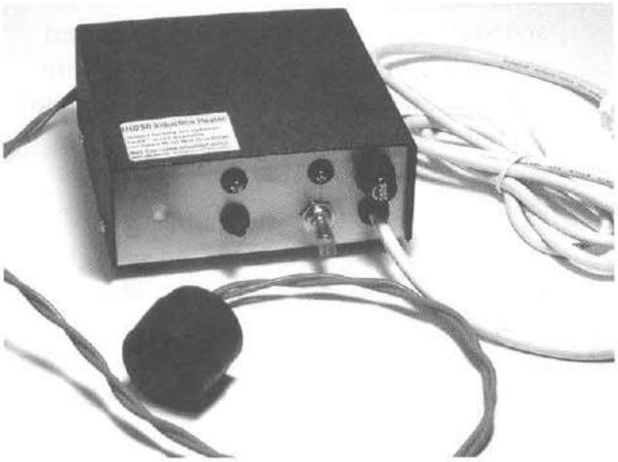

Figure 14-1 *Induction heater*

Circuit Operation

The power cord CO1 supplies 115/220 VAC from a wall receptacle where the third, green wire, grounding lead must be securely attached to the metal chassis (CHASSIS) for safety (see Figure 14-2).

Switch SW1 energizes the primary circuits, and fuse FUS1 protects the main circuit from any catastrophic faults. Indicator lamp NEON1 requires the current-limiting resistor R12 and lights up when SW1 is turned on. The peak charging current to capacitors C10 and C11 is limited by in-rush resistor Rx. The rectifiers D1, D2, D3, and D4 are shown in a full-bridge configuration when used for 220 VAC input. For 115 VAC operation, only D1 and D3 are used along with the JUMP jumper in a voltage-doubler configuration. The output voltage across C10 and C11 is 300 volts DC and is referred to as the positive and negative rails.

Dropping resistors R2 and R3 provide voltage to the oscillator driver IC1 with zener diode Z1 regulating the voltage at 15 volts. Switch SW2 is normally closed, shorting out this voltage across Z1. Capacitor C2 provides the high instant of current for the output pulses and must be physically close to IC1. Operation requires SW2 being pressed, allowing 15 volts to be applied to the oscillator driver generating the drive pulses.

These drive pulses turn the main MOSFET switches, Q1 and Q2, off and on. The switching frequency is determined by the time constant of power control pot R10 and timing capacitor C3. A limit resistor, R11, prevents driving the circuit beyond the intended limits.

The switching circuit is in a half-bridge configuration where the MOSFET connected to the positive end (+ rail) must be driven with its source pin referenced at 150 volts above the common end (− rail). This is accomplished by biasing a bootstrap capacitor (C4) through an ultrafast diode (D5), providing the correct DC level to fully control Q1. Resistors R6 and R7 eliminate the high-frequency parasitic oscillation that occurs as a result of rapidly switching the capacitive load of the MOSFET gates. Capacitor C40 limits the current flow to a second neon indicator light, NEON2, which is connected in series and is energized by the output frequency when the heating cycle is activated by SW2.

A network consisting of capacitor C7 and resistor R8 slows down the transition time of the switched pulses across Q1 and Q2. The resultant time constant limits the rate of voltage rise, dv/dt, that could cause a premature turn-on at the wrong switch, creating a catastrophic fault mode.

Capacitors C5 and C6 provide a voltage midpoint and produce the necessary storage energy to maintain the voltage level of the individual pulses. The output is taken at the junction of Q1/Q2 and C5/C6 and is a square wave fed to the induction heating coil L1. A current rise through L1 now commences as a function of Et/L where E is the voltage of the drive pulse, t is the applied time, and L is the inductance of L1. This changing current now induces eddy currents in the target piece to be heated as a function of $E = L\,di/dt$. These circulating currents quickly heat up the target piece.

Project Assembly

1. Assemble the heating head coil, as shown in Figure 14-3. This coil must be tightly and evenly wound for optimal performance. It may be difficult for the beginner and is available in three sizes, as indicated in Table 14-1.
2. Assemble the circuit board, as shown in Figure 14-4. Note that you should allow a half-inch for the Q1 and Q2 pins. This is necessary to allow these parts to be placed flat against the chassis when mounting as they exit at the midsection of the part.

   It is also strongly suggested that you use an eight-pin IC socket (SO8) because IC1 is very sensitive to static electricity and possible errors during the assembly.

   Note that if you are building from a perforated or vector circuit board, it is suggested that you use the indicated traces for the wire runs and insert components starting in the lower left-hand corner. Pay attention to the polarity of capacitors with polarity signs and all semiconductors.

   Route the leads of the components as shown and solder as you go, cutting away the unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. The heavy foil runs should use the thicker #20 bus wire because these are the high-current discharge paths.
3. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.

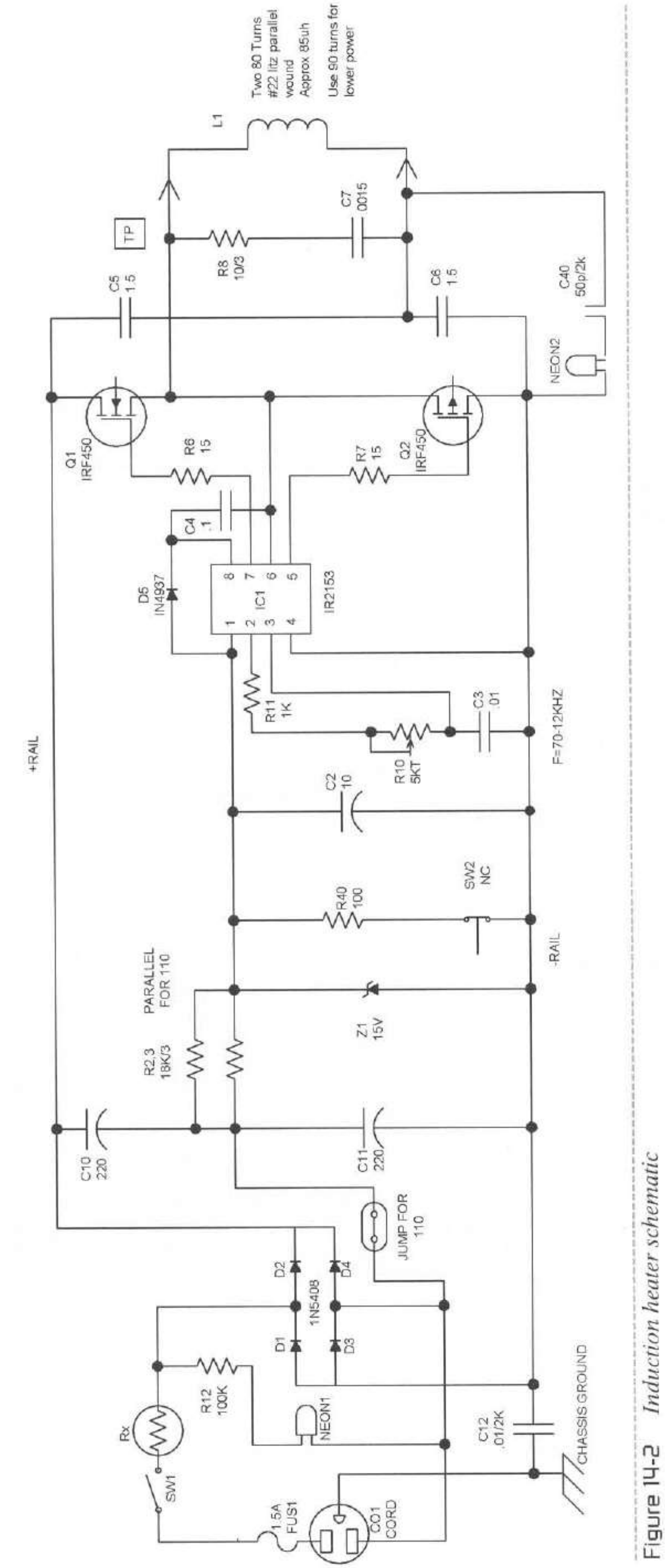

L1
Two 80 Turns #22 litz parallel wound Approx 85uh
Use 90 turns for lower power
TP
C5 1.5
R8 10/3
C7 .0015
C6 1.5
C40 50p/2k
NEON2
Q1 IRF450
Q2 IRF450
R6 15
R7 15
C4 .1
D5 1N4937
IC1
IR2153
1 2 3 4 8 7 6 5
R11 1K
C3 .01
R10 5KT
F=70-12KHZ
+RAIL
C2 10
R40 100
SW2 NC
-RAIL
PARALLEL FOR 110
R2,3 18K/3
Z1 15V
C10 220
C11 220
JUMP FOR 110
D2 D4 D1 D3
1N5408
Rx
R12 100K
NEON1
SW1
1.5A FUS1
CO1 CORD
C12 .01/2K
CHASSIS GROUND

Figure 14-2 Induction heater schematic

4. Connect the external components as shown in Figure 14-5. Use 6-inch lengths of the wire size as noted.

5. Create the chassis, as shown in Figure 14-6, from a piece of .065-inch 5052 bendable aluminum. Note to trial-fit and measure the diameters of the components before drilling the holes. Pay attention to the holes for mounting Q1 and Q2 because these must be reasonably accurate to align with the layout of the PCB. Also, there must be at least 1/8 inch of clearance between the board and the chassis for sandwiching in the PLASTPL plastic plate. This prevents grounding out of the board wiring traces to the metal chassis.

6. The final assembly is shown in Figure 14-7. Note the mounting scheme for Q1 and Q2, as they must be insulated from the chassis using the thermo pads, THERM1. These pads allow heat transfers and provide electrical insulation from the metal chassis.

   Drill the hole for grounding lug LUG1 and solder the green grounding lead from the power cord and circuit board's earth ground.

The heating head coil assembly can be built to accommodate a diameter from a 1/2" to a 1" object. The assembly with the material shown is intended for using enclosed electrodes in glass tubes such as those used in neon tubes, laser tubes, and spectrum tubes. High-temperature materials may be used if you are anticipating using for other applictions.

The coil should be wound with high-frequency wire to limit heating losses. Use #22 litz wire consisting of multiple strands of #36 magnet wire.

1. Fabricate the winding bobbin from 1.25" lengths of 1/16" wall GT fiber glass tubing of your diameter choice of 1/2, 5/8, 3/4, 7/8 or 1".
2. Wind an even 1" layer of the #22 litz wire. Glue the start wire with a fast-curing glue. Tape the finished winding using a single layer of high-temperature glass tape.
3. Continue winding layers and repeat step 2 until there are 80 turns. You may truncate the successive layers by several turns at each end to keep from slipping off or use a quick-curing glue.
4. Attach the input leads and attach by gluing and taping into place.

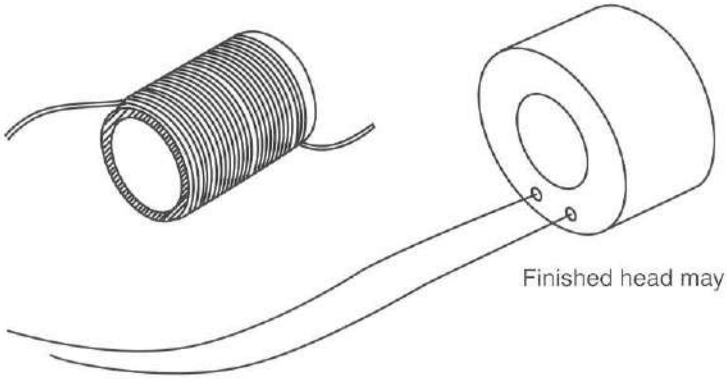

Finished head may be potted in epoxy or sealed in plastic caps

The inductances of the various heads will vary as a function of the diameter. Ideal current through the head is 20 amps peak. You should therefore adjust the frequency of R10 to maintain this value when changing the head sizes as the following:

1/2" head	50 uh	75 KHz
5/8" head	75 uh	50 KHz
3/4" head	100 uh	40 KHz
7/8" head	125 uh	35 KHz
1" head	150 uh	30 KHz

These heads are available assembled for those not wanting to attempt winding their own. They are listed in Table 14-1.

Figure 14-3 *Heater head coil assembly*

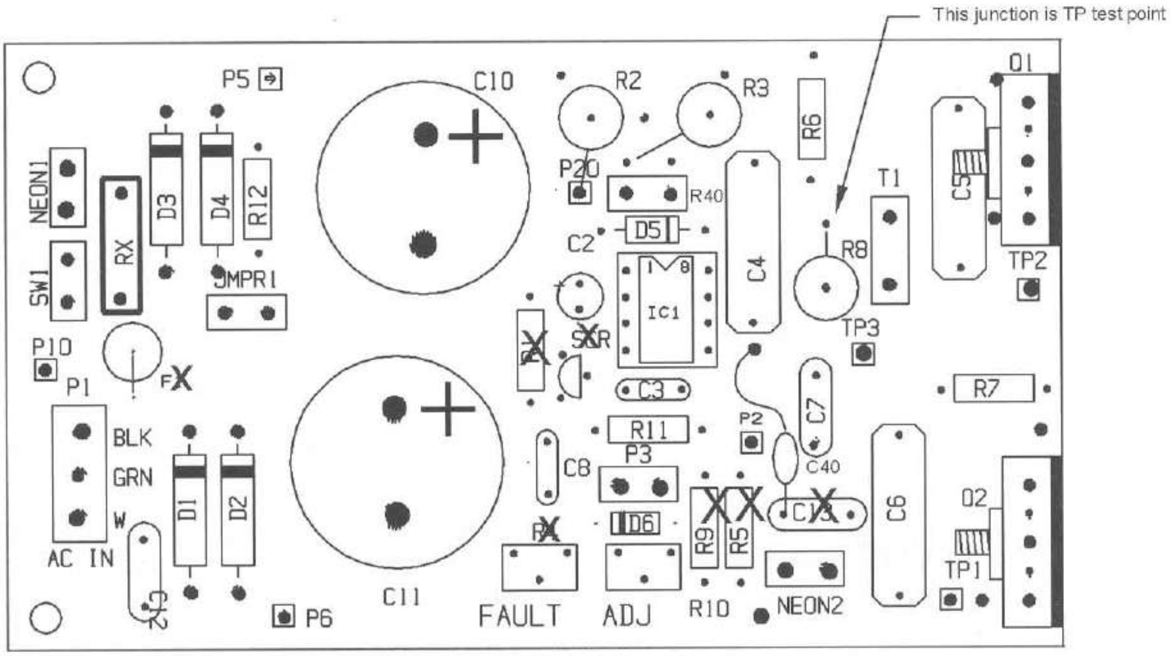
This junction is TP test point
P5
C10
R2
R3
R6
Q1
NEON1
D3
D4
R12
RX
P20
R40
T1
C5
D5
C2
R8
SW1
JMPR1
IC1
C4
TP2
TP3
P10
P1
F
SCR
C3
R7
BLK
GRN
W
AC IN
D1
D2
R11
P2
C7
P3
C8
C40
C9
C6
Q2
R9
R5
D6
TP1
P6
C11
FAULT
ADJ
R10
NEON2

1. Black dots are connections points to external leads as shown figure 14-5

2. The PC board as shown is used in several projects and does not use all the components as indicated. Those indicated by an X are omitted for this project. R40 and C40 are added to the pads as shown and are not identified on the PC board printing

3. This circuit uses a 8 pin socket for IC1 as it is prone to failing if there are circuit faults

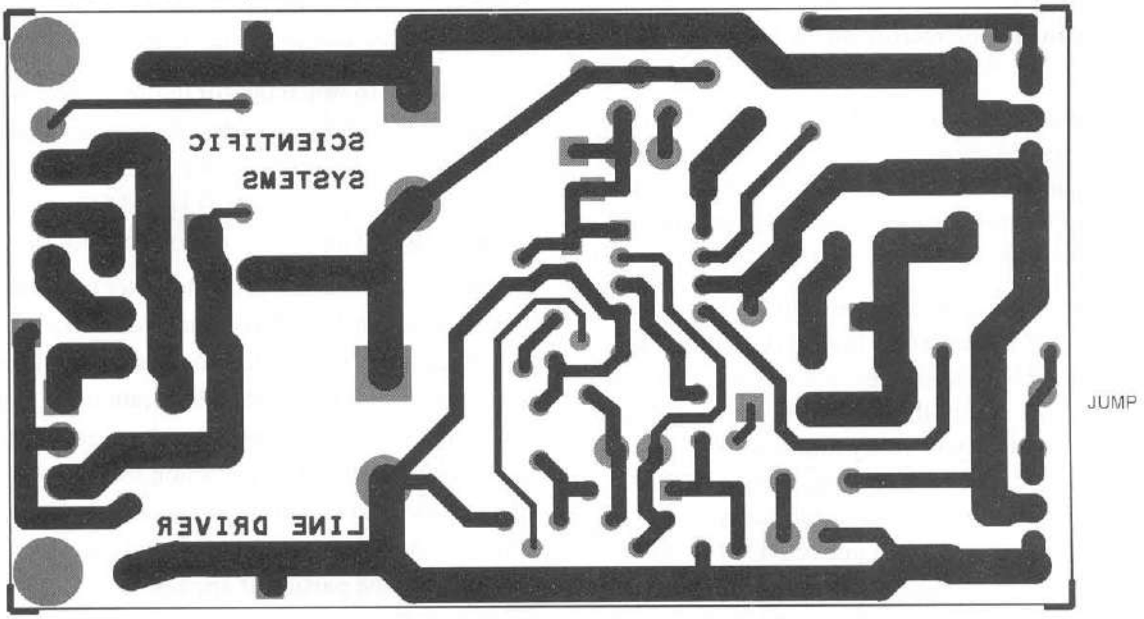
SCIENTIFIC
SYSTEMS
LINE DRIVER
JUMP

Assembly of the board assembly is shown using a printed circuit board available through www.amazing1.com Builders may use a vector board and use the component leads to duplicate the foil traces as shown.

On all IND40 assembly boards this jump lead is to be the first part on the board. Once the unit is final assembled with the heating head connected in place the jump now can now be disconnected. This step prevents static electric damage to the semiconductors.

Figure 14-4 *Assembly of the circuit board*

Note that we have shown the use of plastic bushings (BU38) for the neon indicator lamps and output leads, and also BU58CL clamp bushing for the line cord.

7. Assemble the test jig as shown in Figure 14-8. This is a very useful fixture for any 115 VAC project up to 300 watts.
8. Now fabricate the cover to fit the chassis section. Use bendable plastic or a thin, perforated aluminum sheet. Secure it to the chassis via four screws (SW1) on the folded-up lip in this section.

Project Testing

Caution: The circuit commonly noted as −RAIL is electrically above earthground by over 150 volts. It is strongly recommended that you obtain a 200-watt isolation transformer before proceeding and testing this or other similar direct line powered circuits. Experienced experimenters working on dry, wooden floors may choose to unground their test equipment. This can be a dangerous shock hazard!

1. Connect the ungrounded scope to the TP test point, as noted in the Figure 14-2 schematic or in Figure 14-4. The lead of R8 is shown connected to the junction of the drain and the source of Q1 and Q2 respectively.
2. Verify that the main power switch S1 is off and insert a 2-amp fuse into FH1. Then adjust trimpot R10 to midrange and connect CO1 to the female end of the cord on the test jig. It is important that you verify that the test jig is not shorting the ballast lamp. Plug the jig into the output of the isolation transformer. Quickly turn on S1 and notice NEON1 lighting up and the ballast lamp not lighting. If it lights, you must troubleshoot and find your error before proceeding.

Press and hold PB1, noting a picture-perfect square wave on the scope with the lamp only dimly lighting. Quickly adjust R10 for a 20-usec pulse period. This corresponds to 50 kHz. NEON2 will indicate when PB1 is activated. Experienced builders may wish to monitor the current to the head using a current probe attachment to the scope. This wave shape should be another picture-perfect sawtooth of 20 amps maximum.

Using the Heater Project

Obtain a gas discharge tube with an electrode to fit into the head size you chose in assembly.

At this point, it is a good idea to have a meter capable of measuring the AC line current, as this value is very dependent on the object that you place into the head and should not exceed 2 amps. You can drop this value by increasing the frequency, which is done by adjusting R10. When heating electrodes, you will note the ampmeter dropping in level when the electrode is red hot. This is due to the change in the magnetic characteristics of the heated material.

Something to watch out for in this project is overheating the head coil. You may implement a cooling fan stand when using the device for multiple tasks. Also, the chassis panel where Q1 and Q2 are mounted should not be uncomfortably hot to touch.

Those experienced can create other heads for different heating tasks and you can even make a "quick" heat solder pot. Just make sure you keep the switching current below the maximum of 20 amps. The inductance of the head should not go below 80 uh. Alternate head construction must always take into consideration heating fatigue and the eventual failure of the head wires, bobbin, and construction materials for the particular application.

Notes:
1. The parts on the printed circuit board with an X are not used for this project.
2. C40 and R40 are added components not indicated on the board
3. Lead lengths can be shortened in final assembly

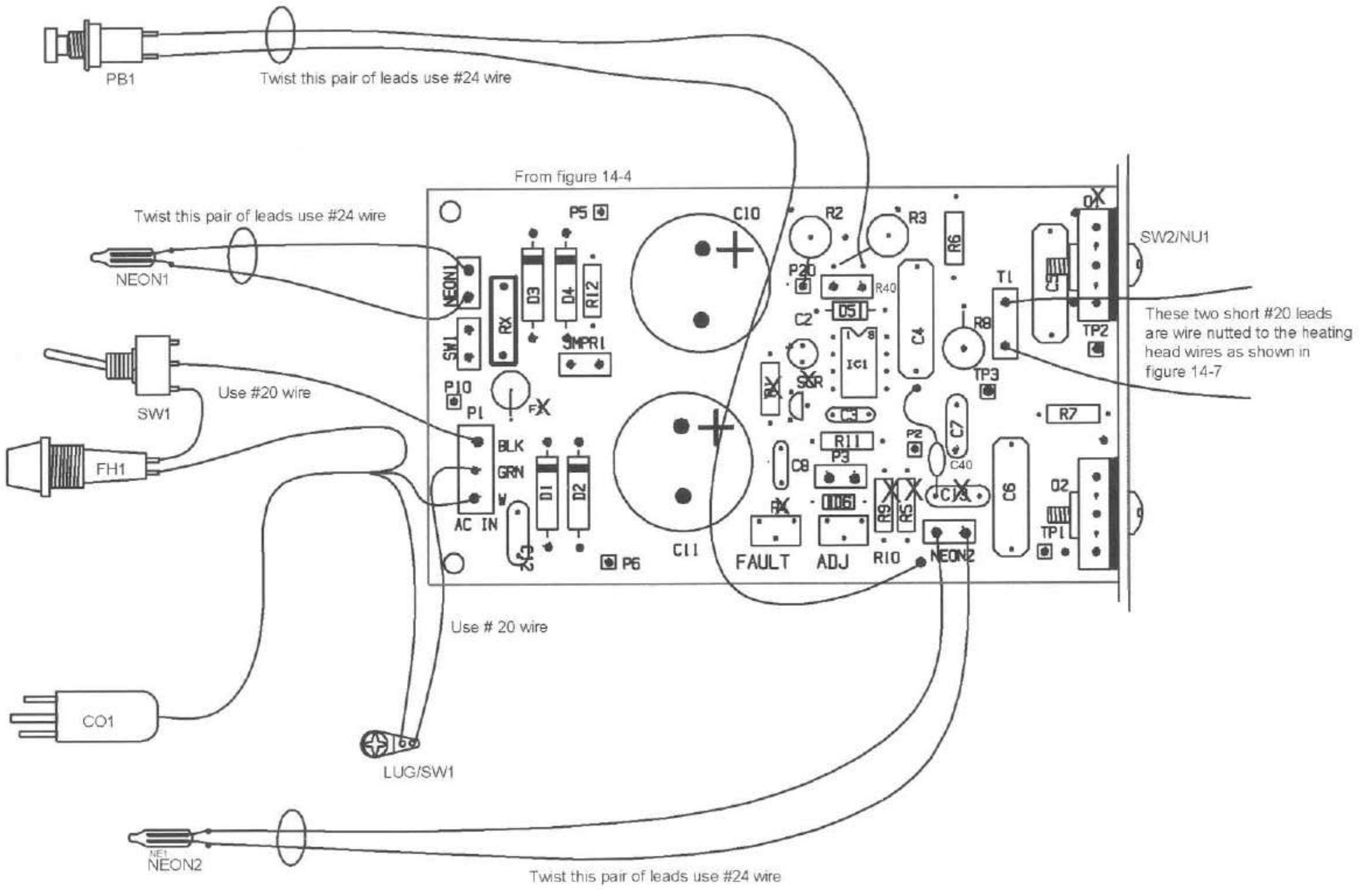
PB1
Twist this pair of leads use #24 wire
From figure 14-4
Twist this pair of leads use #24 wire
NEON1
SW1
Use #20 wire
FH1
CO1
LUG/SW1
Use # 20 wire
NE1
NEON2
Twist this pair of leads use #24 wire
SW2/NU1
These two short #20 leads are wire nutted to the heating head wires as shown in figure 14-7
P5
C10
R2
R3
R6
T1
TP2
P20
R40
D5
C2
IC1
C4
R8
TP3
SCR
C3
R7
R11
P2
C7
C40
P3
C8
C6
Q2
TP1
C11
FAULT
ADJ
R10
NEON2
R9
R5
C13
P6
NEON1
SW1
RX
D3
D4
R12
JMPR1
P10
P1
FX
BLK
GRN
W
AC IN
D1
D2
C12

Figure 14-5 *External wiring*

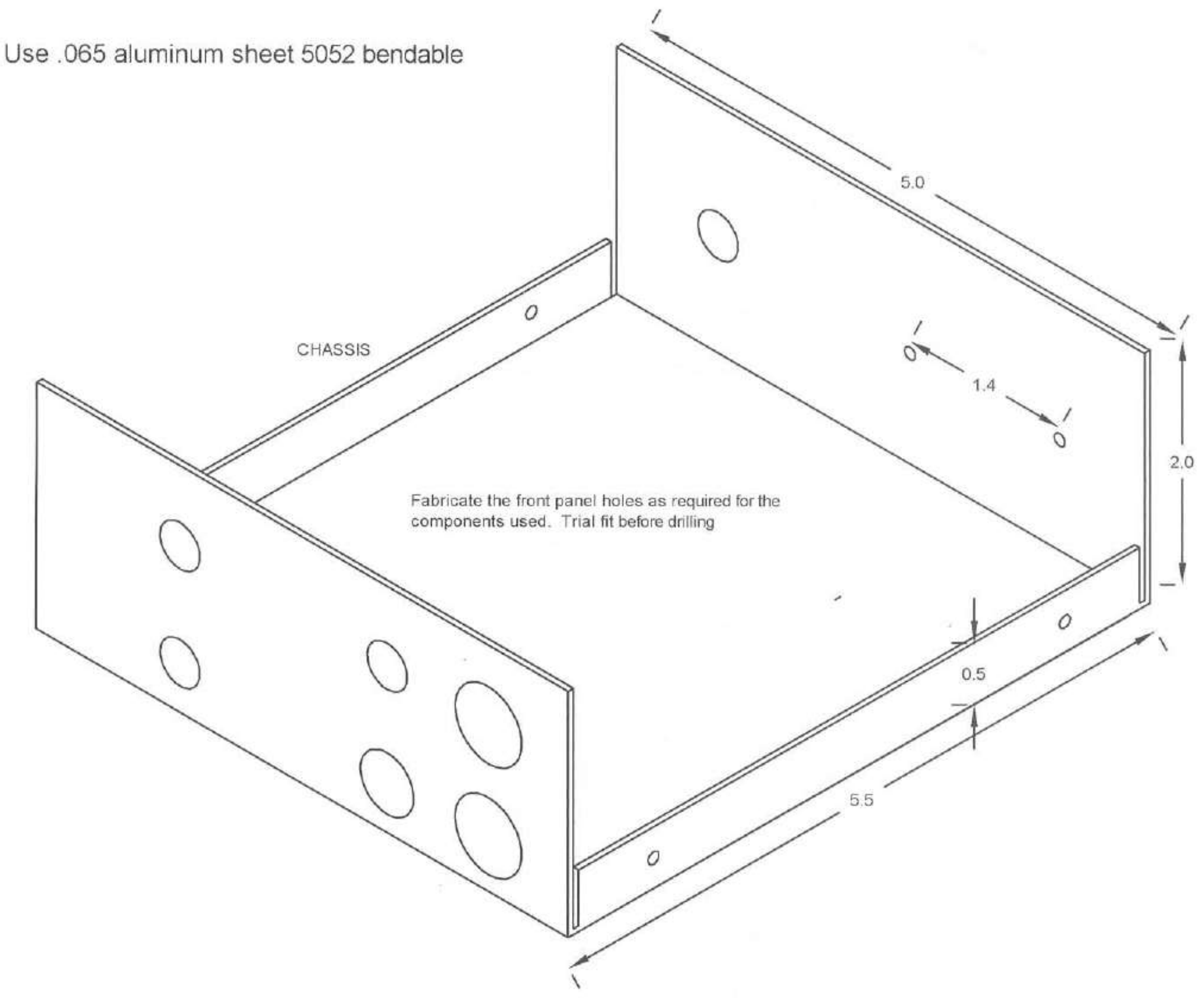
Use .065 aluminum sheet 5052 bendable
5.0
CHASSIS
1.4
2.0
Fabricate the front panel holes as required for the components used. Trial fit before drilling
0.5
5.5

Figure 14-6 *Chassis fabrication*

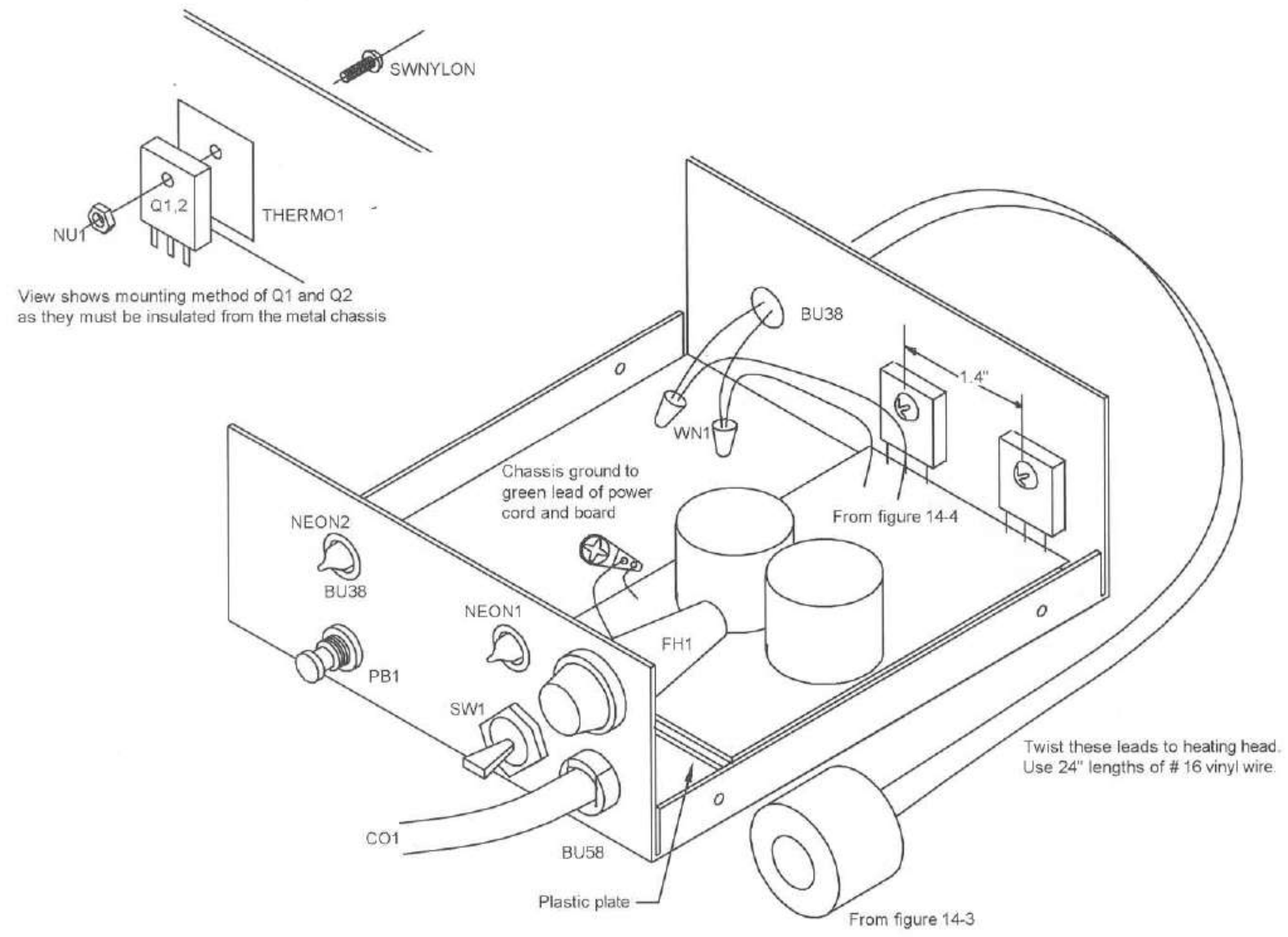
SWNYLON
Q1,2
THERMO1
NU1
View shows mounting method of Q1 and Q2
as they must be insulated from the metal chassis
BU38
1.4"
WN1
Chassis ground to
green lead of power
cord and board
From figure 14-4
NEON2
BU38
NEON1
FH1
PB1
SW1
Twist these leads to heating head.
Use 24" lengths of # 16 vinyl wire.
CO1
BU58
Plastic plate
From figure 14-3

Figure 14-7 *Isometric of final assembly*

This test jig as shown prevents catastrophic damage to AC powered circuitry in the event of a fault when first assembled. Basically it uses a 60-watt lamp as a ballast, connected in series with the hot wire of the AC line. The lamp will usually allow ample current to flow for operational circuit verification yet limit short circuit current to that of the lamp should the circuit have a catastrophic fault. It is suggested that the serious experimenter construct this test jig and enclose it in a suitable container with receptacle for plugging in any AC circuit under test.

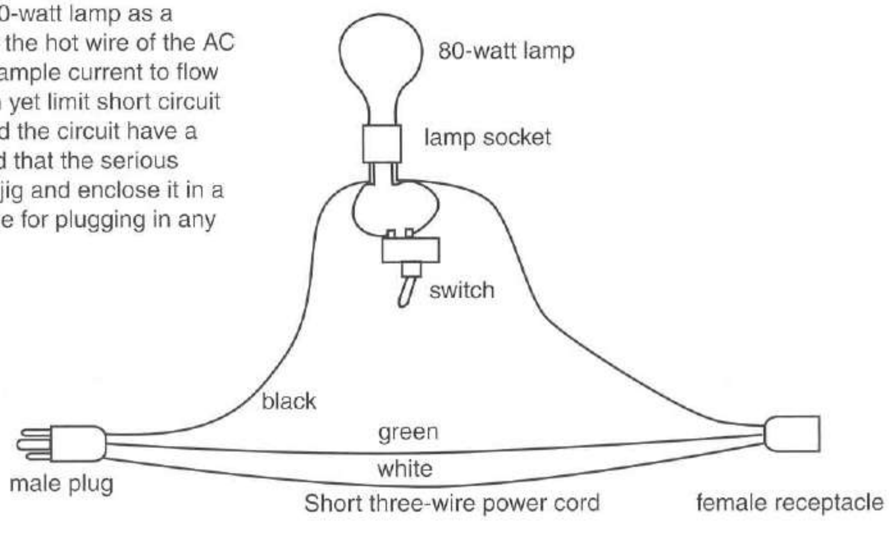
80-watt lamp
lamp socket
switch
black
green
white
male plug
Short three-wire power cord
female receptacle

Figure 14-8 *Useful test jig for line-drive circuits*

Table 14-1 Induction Heater Parts List

Ref. #	Description	DB Part #
R2, 3	Two 18K 3-watt resistor	
R4	100-ohm, 1/4-watt resistor (br-blk-br)	
R6, 7	Two 15 1/4-watt resistor (br-grn-blk)	
R8	10-ohm, 3-watts metal oxide resistor	
R10	5K vertical trimmer	
R11	1 K 1/4-watt resistor (br-blk-red)	
R12	100 K 1/4-watt resistor (br-blk-yel) or 10-ohm 3-watt metal oxide resistor	
Rx	5-amp in-rush limiter (KC006L-ND)	
C2	10-microfarad, 50-volt vertical electrolytic capacitor	
C3	.01- microfarad, 50-volts plastic capacitor (103)	
C4	.1-microfarad, 600-volt plastic capacitor	
C5, 6	Two 1.5-microfarad, 400-volt metal polyester capacitors	
C7	.0015-microfarad, 600-volt polypropylene capacitor	
C10, 11	Two 200- to 800-microfarad, 200-volt vertical electrolytic capacitors	
C12	.01-microfarad, 2-kilovolt disc capacitor	
C40	50-picofarad, 6-kilovolt ceramic capacitor	
D1, 2, 3, 4	Four 1,000-volt, 3-amp rectifiers (1N5408)	
D5	1-kilovolt, 1-amp fast diode (1N4937)	
Z1	15-volt, .5-watt zener diode (1N5245)	
Q1, 2	IRFP450/460 *metal-oxide-semiconductor field effect transistors* (MOSFET)	
IC1	IR2153 half-bridge driver	
NEON1, 2	Two neon indicator bulbs with leads	
FH1	Fuse holder panel mount	
FUS1	2-amp slow blow 3AG fuse	
PB1	Normally closed push-button switch	
SW1	*Single pole, single throw* (SPST) toggle switch	
CO1	Three-wire line cord	
LUG1	#6 solder lug	
SO8	Eight-pin *integrated circuit* (IC) socket	
THERM1	Two thermo mounting pads for Q1 and Q2	
PCLINE	*Printed circuit board* (PCB)	DB# PCLINE
WN1	Two small wire nuts	
CHASSIS	Chassis fabricated from .065-inch aluminum sheet per Figure 14-6	
COVER	Cover fabricated from plastic sheet to fit chassis (see Figure 14-6)	
BU38	Three 3/8-inch plastic bushings	
BU58CL	5/8-inch clamp bushing for cord CO1	
SW1	Five #6 × 3/8-inch sheet metal screws for cover and LUG1	
SW2/NU1	Two 6-32 × 1/2-inch nylon screws and nuts	
WR24	24 inches of #24 vinyl stranded hookup wire	
WR20	24 inches of #20 vinyl stranded hookup wire	
WR16	48 inches of #16 vinyl stranded hookup wire	
PLASTIC	2 × 4-inch plastic insulating plate for under PC board	
HEAD12	Head with 1/2-inch clearance hole	DB# HEAD12
HEAD58	Head with 5/8-inch clearance hole	DB# HEAD58
HEAD34	Head with 3/4-inch clearance hole	DB# HEAD34
HEAD78	Head with 7/8-inch clearance hole	DB# HEAD78
HEAD10	Head with 1-inch clearance hole and head parts for those wanting to attempt assembly	DB# HEAD10
WR22LITZ	#22 LITZ of multiple strands of #36 enamel wire	
BOBBIN	1.25 inches of 5/8-inch GT tubing	

Chapter Fifteen

50-Kilovolt Laboratory DC Supply

This project, as shown in Figure 15-1, provides the high-voltage experimenter or researcher a DC source of 5 to 60 kilovolts at 5 milliamps. The system is short-circuit protected and current limited by inductive ballasting. This circuitry is excellent for high-energy capacitor charging intended for high-power applications. The charging current is fully adjustable and easily converts to a voltage source by the addition of load resistors. It is an excellent device for large gravity and lifter research, as well as normal, high-voltage laboratory functions.

This is an advanced-level project requiring electronic skills and high-voltage experience. Expect to spend $100 to $250. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 15-1.

Basic Description

This project is shown constructed in two sections. The first section is the power conditioning and control circuitry. This is where the 115 VAC level is converted to a variable-frequency, high-voltage current source between 15 and 30 kHz. The second section is the high-voltage multiplier circuitry interconnected via a short umbilical cable from the power control section.

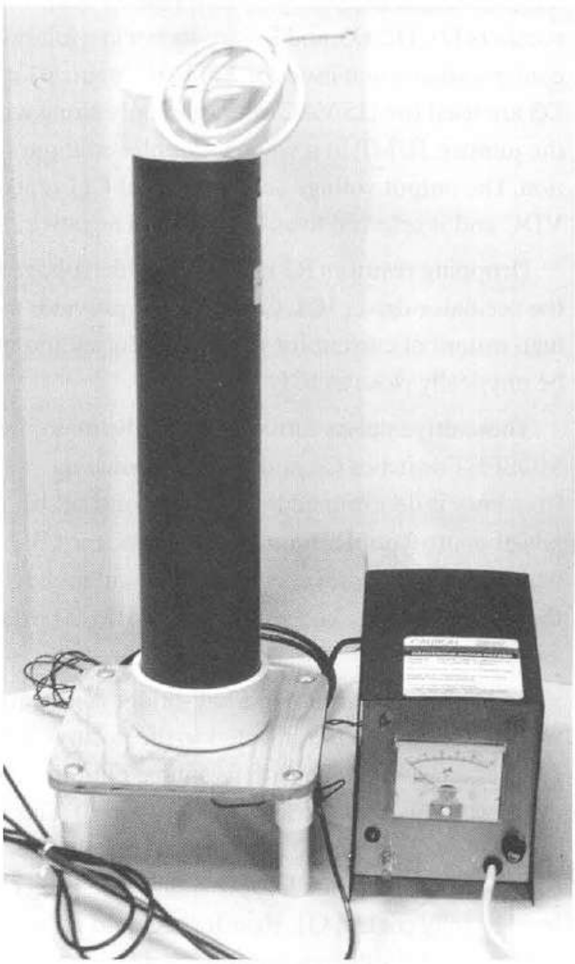

Figure 15-1 *50-Kilovolt power supply*

The multiplier section is a Cockcroft-Walton half-wave circuit consisting of 10 capacitors and diodes connected in the ladder configuration, along with a peak-current, high-voltage-limiting resistor.

Circuit Operation

The power cord (CO1) supplies 115/220 VAC from a wall receptacle where the third, green, grounding lead wire must be securely attached to the metal chassis (CHASSIS) for safety (see Figure 15-2).

Switch S1 energizes the primary circuits and fuse FUS1 protects the main circuit from any catastrophic faults. The indicator lamp NEON1 requires current-limiting resistor R12 and lights up when SW1 is turned on. The peak charging current to capacitors C10 and C11 is limited by in-rush resistor Rth. The rectifiers D1, D2, D3, and D4 are shown in a full-bridge configuration when used for 220 VAC input. D1 and D3 are used for 115 VAC operation only along with the jumper, JUMP, in a voltage-doubler configuration. The output voltage across C10 and C11 is 300 VDC and is referred to as the plus and negative rails.

Dropping resistors R2 and R3 provide voltage to the oscillator driver IC1. Capacitor C2 provides the high instant of current for the output pulses and must be physically close to IC1.

These drive pulses turn off and on the main MOSFET switches Q1 and Q2. The switching frequency is determined by the time constant of power control pot R21 and timing capacitor C3. A limit resistor R11 prevents driving the circuit beyond the intended limits. You will note that R21 is part of switch S1.

The switching circuit is in a half-bridge configuration where the MOSFET connected to the positive end (+ rail) must be driven with its source pin referenced at 150 volts above the common end (− rail). This is accomplished by biasing a bootstrap capacitor C4 through ultrafast diode D5, providing the correct DC level to fully control Q1. Resistors R6 and R7 eliminate the high-frequency parasitic oscillation that occurs as a result of rapidly switching the capacitive load of the MOSFET gates. The transition time of the switched pulses across Q1 and Q2 is slowed down by a network consisting of capacitor C7 and resistor R8. The resultant time constant limits the rate of voltage rise dv/dt rate, which could cause premature turn-on at the wrong switch, creating a catastrophic fault mode.

Capacitors C5 and C6 provide a voltage midpoint and produce the necessary storage energy to maintain the voltage level of the individual pulses. The output is taken at the junction of Q1/Q2 and C5/C6 and is a square wave fed to the primary of transformer T1, inducing the high-voltage output in the secondary winding.

A safety shutdown circuit, consisting of a spark breakover switch, fires when the voltage exceeds a certain level and is fed to isolation transformer T2. The output is now rectified by diode D6 and integrated onto capacitor C8. When this voltage exceeds a preset level controlled by trimpot R4, a trigger level now turns on the silicon switch, SCR, and crowbars the voltage to the drive chip IC1, shutting it down and disabling the high voltage until reset by power removal.

The 15,000-volt, high-frequency output is fed to the voltage multiplier stack, consisting of capacitors C14 through C21 and diodes D7 through D14. The resultant 50,000-volt DC output is in series with peak-current-limiting resistors (R17 through R20).

Assembly of the Driver Section

1. Assemble the circuit board as shown in Figure 15-3. Note to allow a length of a half-inch for the pins of Q1 and Q2. This is necessary to allow these parts to be placed flat against the chassis when mounting.

   It is also strongly suggested that you use an eight-pin integrated circuit socket (SO8) because IC1 is very sensitive to static electricity and possible errors during assembly.

   Note that if you are building from a perforated or vector circuit board, it is suggested that you use the indicated traces for the wire runs and insert components starting in the lower left-hand corner. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors.

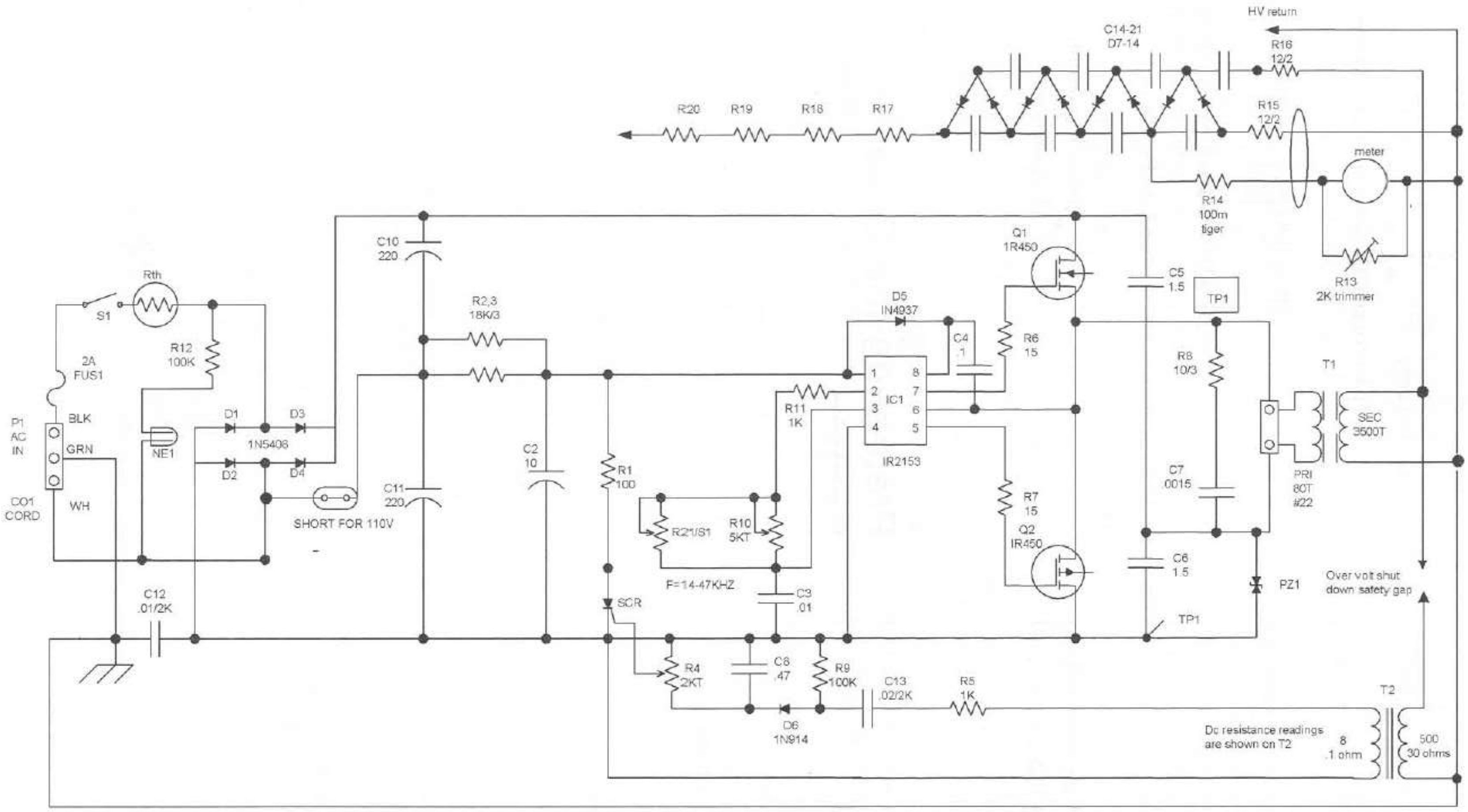
HV return
C14-21
D7-14
R16
12/2
R15
12/2
R20
R19
R18
R17
meter
R14
100m
tiger
R13
2K trimmer
Q1
1R450
C5
1.5
TP1
R8
10/3
T1
SEC
3500T
PRI
80T
#22
C7
.0015
C6
1.5
PZ1
TP1
Over volt shut
down safety gap
T2
500
30 ohms
8
.1 ohm
Dc resistance readings
are shown on T2
D5
1N4937
C4
.1
R6
15
IC1
IR2153
1
2
3
4
5
6
7
8
R7
15
Q2
IR450
R11
1K
R21/S1
R10
5KT
F=14-47KHZ
C3
.01
C10
220
R2,3
18K/3
C2
10
C11
220
R1
100
SCR
R4
2KT
C8
.47
R9
100K
C13
.02/2K
R5
1K
D6
1N914
Rth
S1
2A
FUS1
R12
100K
P1
AC
IN
BLK
GRN
WH
CO1
CORD
NE1
D1
D3
1N5406
D2
D4
SHORT FOR 110V
C12
.01/2K

Figure 15-2 *Circuit schematic*

TRANSFORMER

PRIMARY	80	#22 LITZ	9MH	4.75MH
XFMR	TURNS	WIRE	Lsc	Loc
SECONDARY	3500	#39 MAG	3HY	9H
AIR GAP=	2 MILS/SIDE			

Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. The heavy foil runs should use the thicker
#20 bus wire as these are the high-current discharge paths.

2. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
3. Assemble the transformer as shown in Figure 15-4, or you may purchase this part as noted on the parts list.
4. Connect the external components as shown in Figure 15-5.
5. Fabricate the chassis as shown in Figure 15-6 from a piece of .035-inch 5052 bendable aluminum. Note to trial-fit and measure the diameters of the components before drilling the holes. Pay attention to the holes for mounting Q1 and Q2, as these must be reasonably accurate to align with the layout of the PCB. Also, there must be at least 1/8 inch of clearance between the board and chassis for sandwiching in the PLASTPL plastic plate. This prevents grounding out of the board's wiring traces.
6. Fabricate the plastic base from a 9 × 6-inch piece of 1/16-inch Lexan sheet, as shown in Figure 15-7.
7. Fabricate the parts as shown in Figure 15-8 for the over-voltage shutdown switch.
8. Attach the metal chassis to the plastic base via four screws and nuts, SW12/NU1 (see Figure 15-9).
9. Attach the shutdown switch assembly from Figure 15-8 using sheet metal screws SW1.
10. Mount the assembled PCB, as in Figure 15-3. Note that Q1 and Q2 are mounted using the insulating THERMO1 pads with the nylon screws. Mount the prewired front panel components, as in Figure 15-5.
11. Mount transformer T1 via an 8-inch tie wrap. Note the cores of T1 are wrapped with a single turn of #18 bus wire and this is connected to the ground lug electrically grounding the core.
12. Mount transformer T2 using a piece of double-sided foam tape or glue it in place.
13. Mount the remaining hardware, wiring lugs, and so on.
14. Proceed to wire everything as shown in Figure 15-10. Use twist pairs where noted or tie-wrap to neaten its appearance (see Figure 15-11).
15. Double-check the wiring for any obvious errors. You should have a finished assembly resembling Figure 15-12. The unit is now ready for the basic pretest.

Pretesting the Driver Section

This step will require basic electronic laboratory equipment including a 60 MHz oscilloscope

1. Assembly the test jig as shown in Chapter 14 Figure 14-8. This fixture is a very useful for any 115 VAC project up to 300 watts.

Caution: The circuit commonly noted as −RAIL is electrically above ground by over 150 volts. It is strongly recommended that you obtain a 200-watt isolation transformer before proceeding and that you test this or similar direct-line-powered AC circuits. Experienced people working on dry, wooden floors may choose to unground their test equipment. This still can be a dangerous shock hazard!

2. Connect the scope to the test points as noted in the Figure 15-2 schematic or as shown in Figure 15-4. The lead of R8 is shown connected to the junction of the drain and the source of Q1 and Q2 respectively.
3. Short out the output leads.

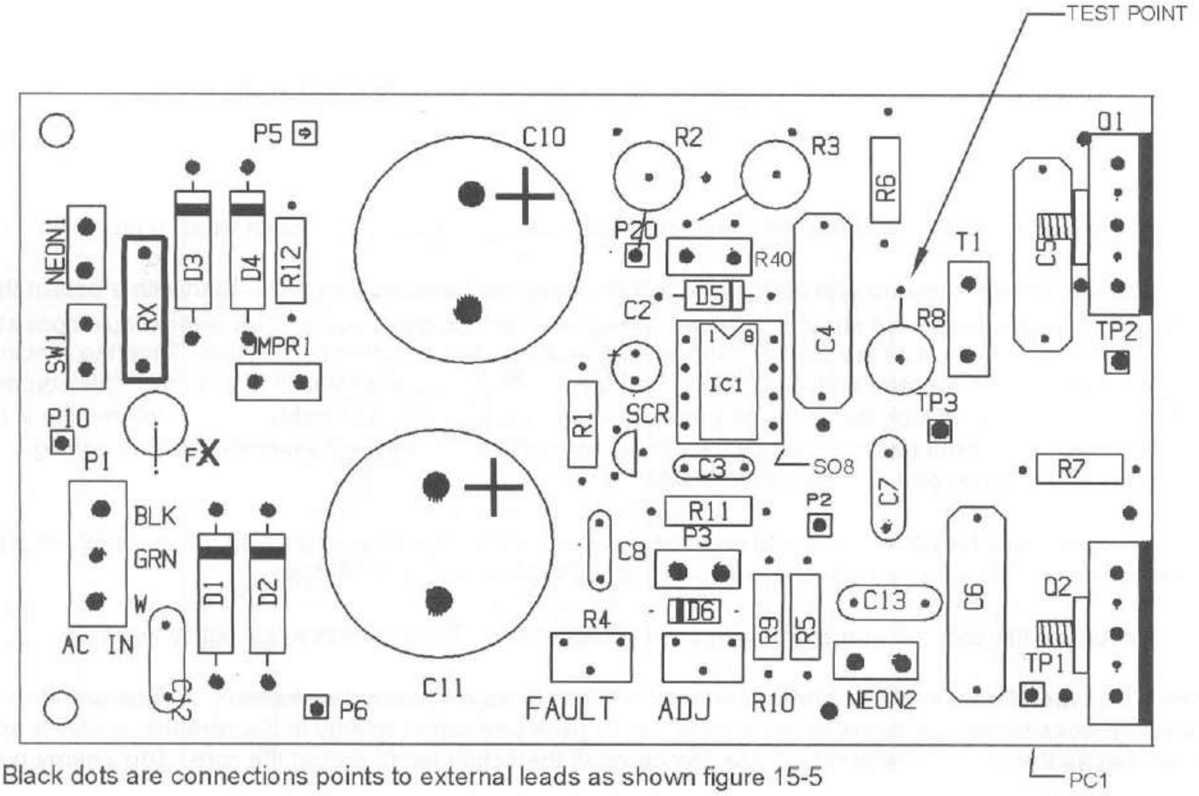
TEST POINT
P5
C10
R2
R3
R6
Q1
NEON1
D3
D4
R12
P20
R40
T1
C5
SW1
RX
JMPR1
C2
D5
IC1
C4
R8
TP2
P10
P1
FX
R1
SCR
SO8
TP3
C3
R7
BLK
GRN
W
AC IN
D1
D2
C8
P3
R11
P2
C7
C6
Q2
R4
D6
R9
R5
C13
TP1
C11
FAULT
ADJ
R10
NEON2
P6
PC1

1. Black dots are connections points to external leads as shown figure 15-5

2. The PC board as shown is used in several projects and may not use all the components as indicated.

3. This circuit uses a 8 pin socket for IC1 as it is prone to failing if there are circuit faults

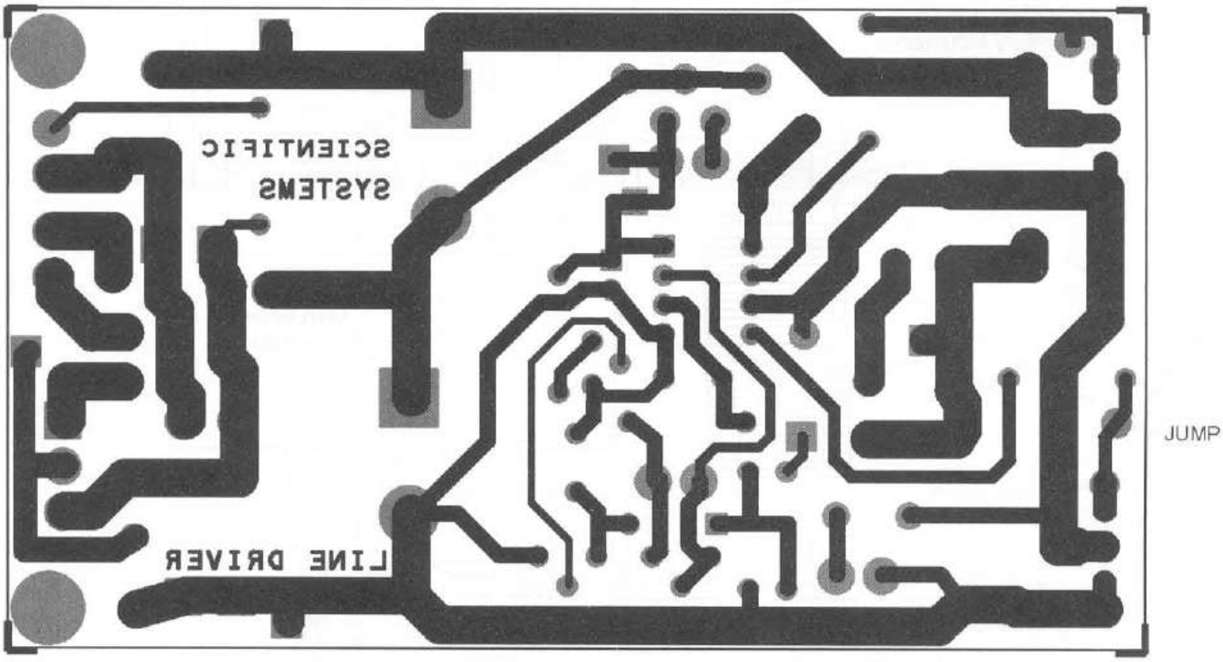
SCIENTIFIC
SYSTEMS
LINE DRIVER
JUMP

Assembly of the board is shown using a printed circuit board available through www.amazing1.com Builders may use a vector board and use the component leads to duplicate the foil traces as shown.

Figure 15-3 *Assembly of the circuit board*

The special ferrite transformer used in this project is available ready to use and is listed on table 15-1. The transformer is designed with a certain amount of leakage reactance between the primary and secondary winding to allow limited short circuit current. This approach allows charging capacitors without the use of lossy resistors. While this is an advantage in circuit protection is has the disadvantage of requiring transistor switches (FETS) that now must switch the reactive leakage current in addition to the real load current. We show our circuit using high current switches that alleviate this problem.

The design specifications are shown for those advanced builders desiring to attempt winding their own.

1. The secondary is wound with 3500 turns of #39 heavy insulated magnet wire. Start with a bobbin that will fit over the cores and wind about a ¾" layer evenly and without cross overs. The layer should consist of about 200 turns. Note to exit the start of your winding as this is the common connection. Tape the winding using 2 mil polypropylene tape and wind the second layer until you have a total of 3500 turns. This equates out to 18 layers!! (good luck doing this by hand) You now must pot the assembly and this will require a cup to hold the windings and out gassing using a vacuum pump on both the winding assembly and the potting material. We use a two part silicon rubber GE#627.

2. The primary is far easier to do and requires another bobbin now only winding 80 turns of #22 Litz high frequency wire. This will require two layers and is taped together using mylar tape.

3. Assemble the core set and space with 5 mil (.005)" shims. Tape the assembly tightly together.

Note: This type of transformer is often incorrectly referred to as a "flyback transformer". A flyback unit physically looks similar but this is where it ends. A fly back unit stores energy in the core that is mainly across the air gap as this is a very high reluctance compared to the actual ferrite part of the core. This energy builds up as long as the transistor is on and then "flies back" when the transistor switches off.

Measured parameters:

Open circuit primary inductance..........5.5 mH
short circuit primary inductance.......... 1 mH
Open circuit secondary inductance.......10 H
Short circuit secondary inductance....... 2 H

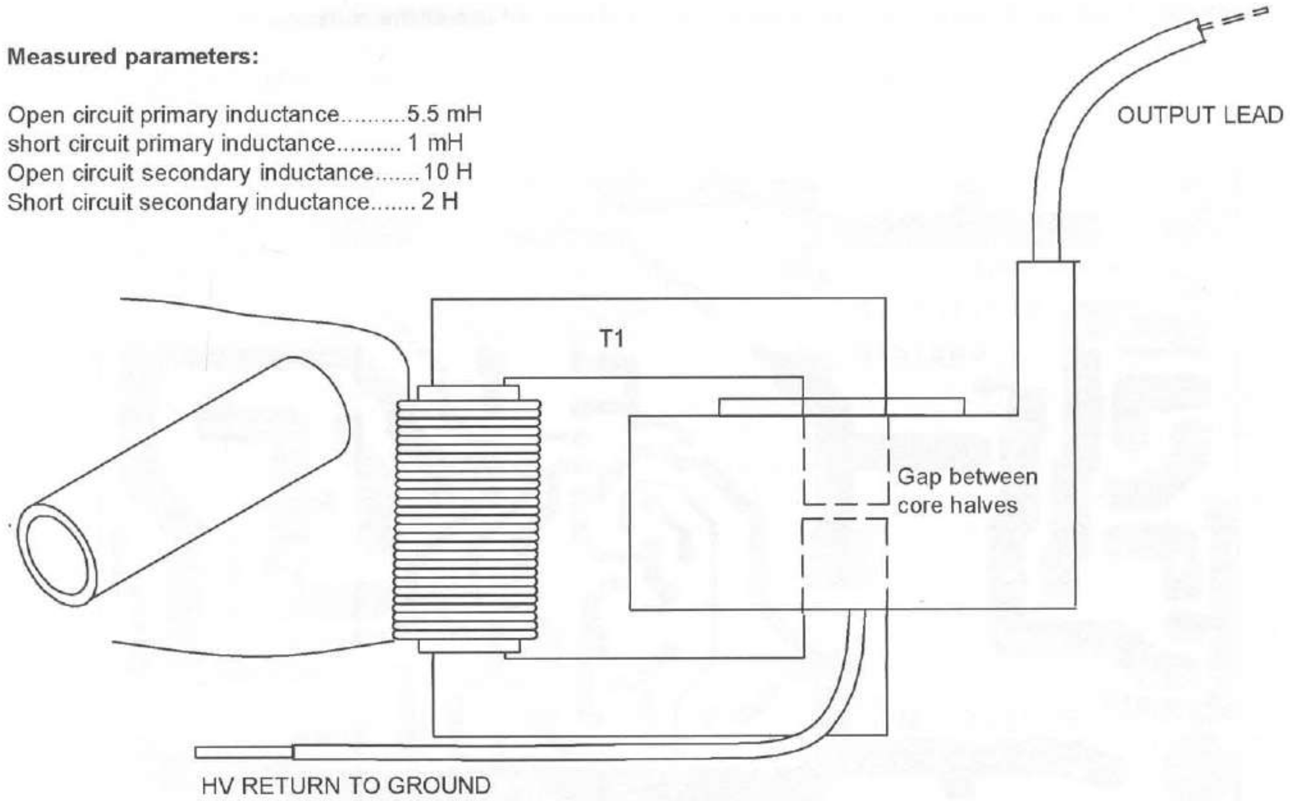

*#18 solid magnet wire can be used however high frequency LITZ wire will give a slight improvement. You can make this wire by obtaining 6 pieces of #26 magnet wire and twist together as a single wire.

Figure 15-4 *Assembly of transformer*

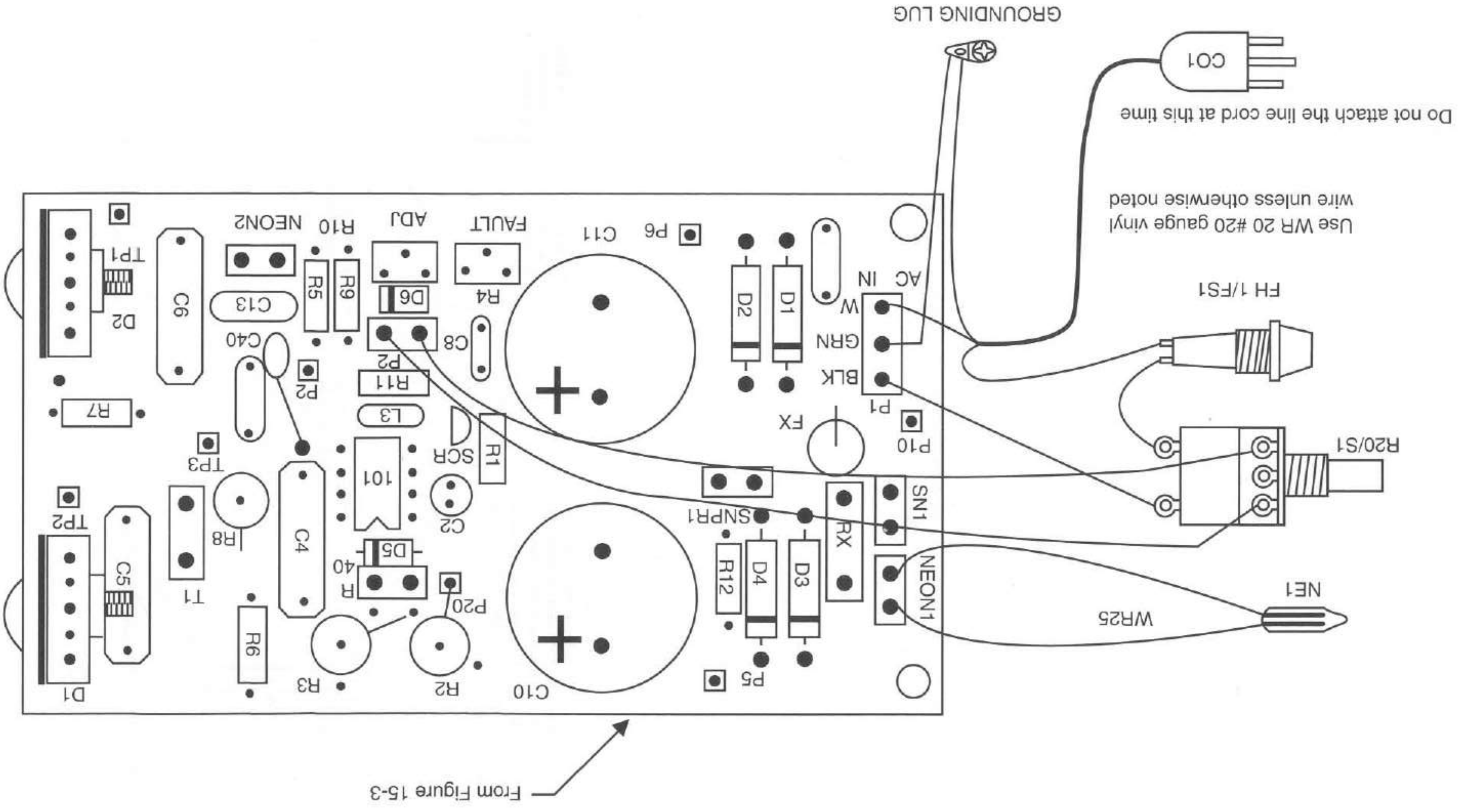
From Figure 15-3
D1
TP2
C5
R7
D2
TP1
T1
C6
R8
TP3
R6
C40
C13
NEON2
R3
C4
P2
R5
R10
R9
R
40
D5
101
L3
R11
P2
D6
ADJ
R2
P20
C2
SCR
C8
R4
FAULT
R1
C10
C11
P6
SNPR1
R12
P5
D4
D2
D3
D1
FX
RX
P1
BLK
GRN
W
AC IN
NEON1
SN1
P10
GROUNDING LUG
WR25
NE1
R20/S1
FH 1/FS1
CO1

Use WR 20 #20 gauge vinyl wire unless otherwise noted

Do not attach the line cord at this time

Figure 15-5 *External wiring*

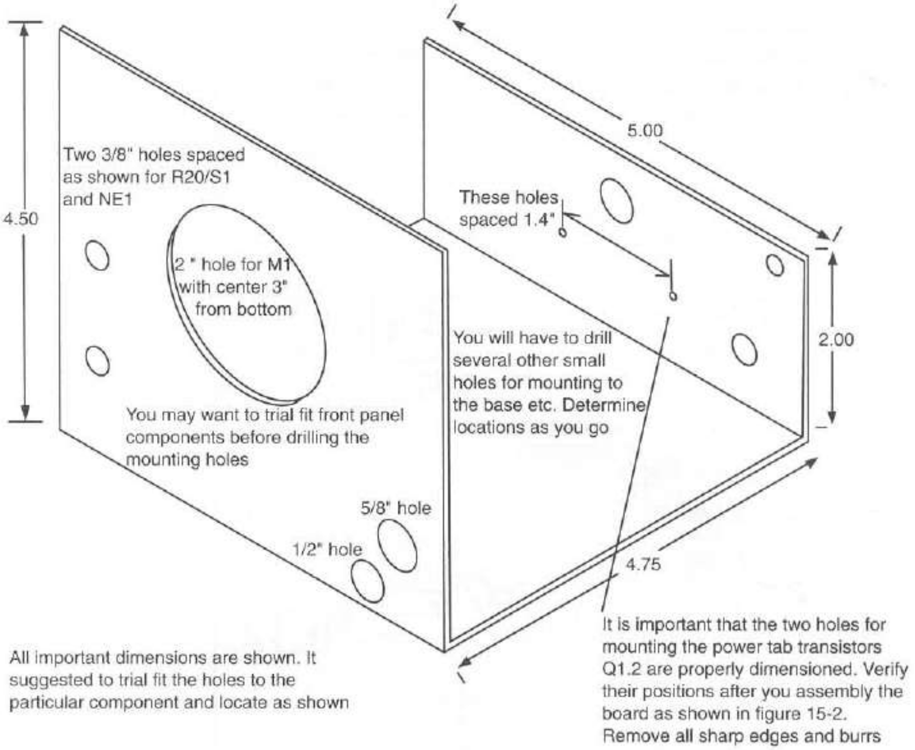
4.50
5.00
2.00
4.75
2 " hole for M1 with center 3" from bottom
These holes spaced 1.4"
5/8" hole
1/2" hole

Two 3/8" holes spaced as shown for R20/S1 and NE1

You may want to trial fit front panel components before drilling the mounting holes

You will have to drill several other small holes for mounting to the base etc. Determine locations as you go

All important dimensions are shown. It suggested to trial fit the holes to the particular component and locate as shown

It is important that the two holes for mounting the power tab transistors Q1.2 are properly dimensioned. Verify their positions after you assembly the board as shown in figure 15-2. Remove all sharp edges and burrs

Figure 15-6 *Chassis fabrication*

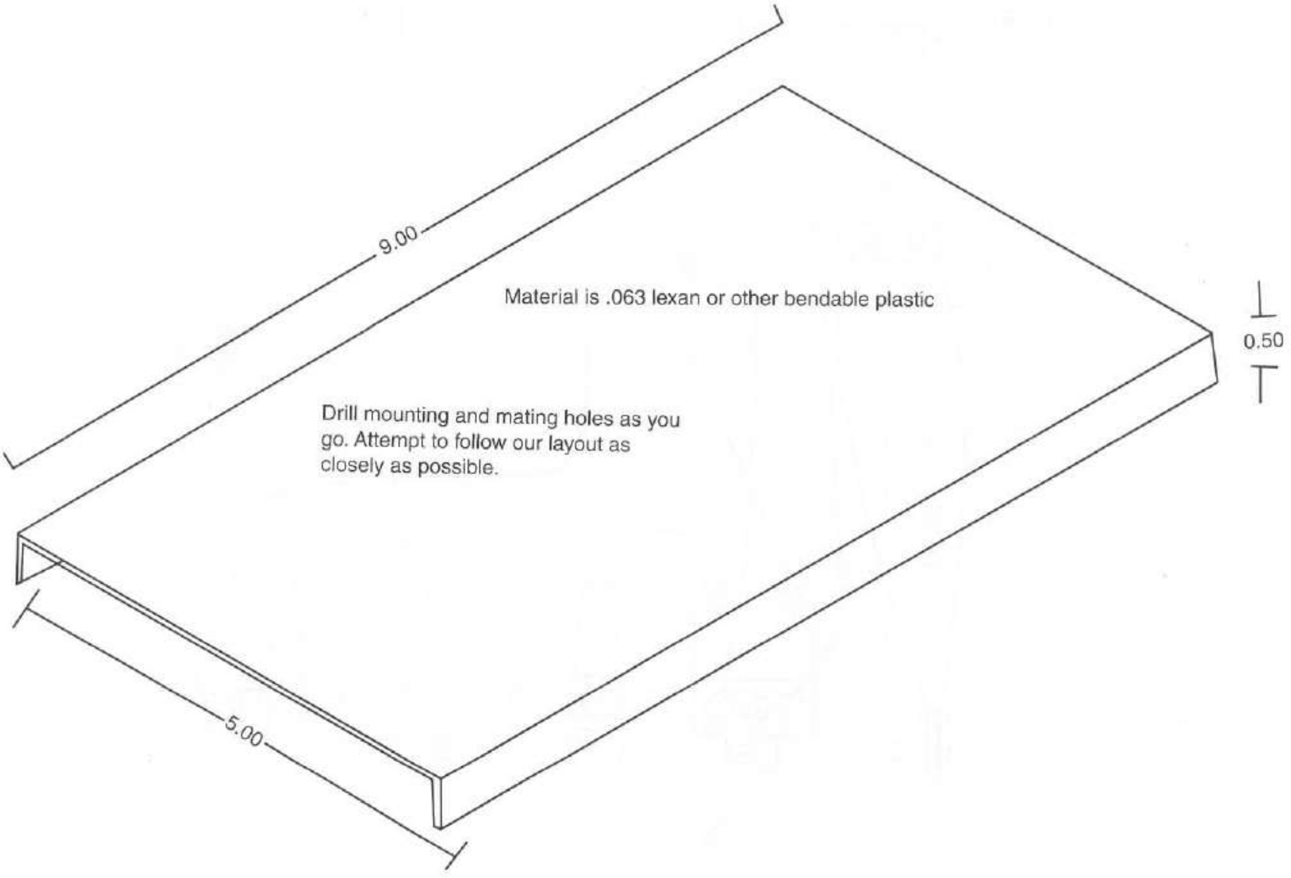

Material is .063 lexan or other bendable plastic

Drill mounting and mating holes as you go. Attempt to follow our layout as closely as possible.

Figure 15-7 *Plastic base fabrication*

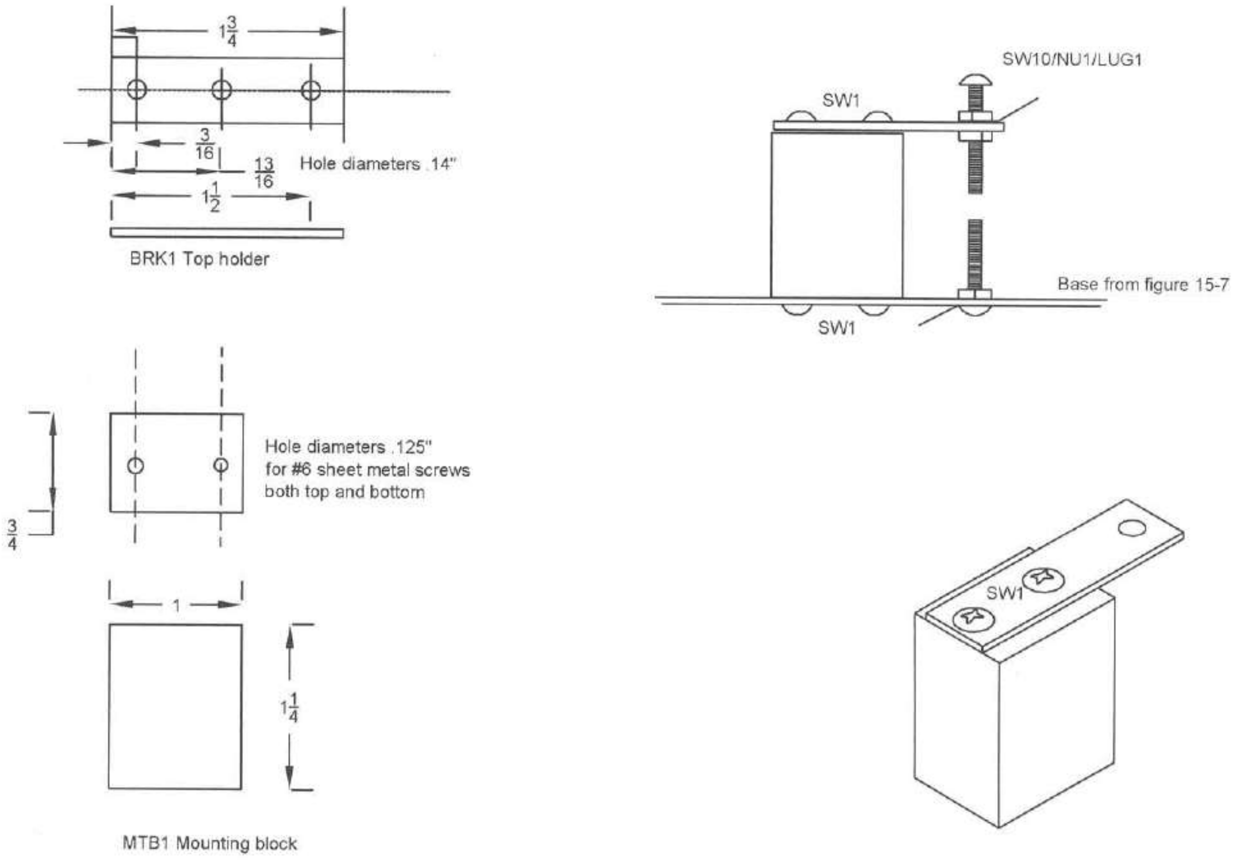

Figure 15-8 *Creating the shutdown switch*

4. Verify the main power switch S1 is off and insert a 2-amp fuse into FH1. Connect CO1 to the female end of the cord on the test jig and adjust both trimpots to midrange. Verify the test jig is not shorting the ballast lamp and plug the jig into the output of the isolation transformer. Quickly turn on S1 and note NEON1 lighting and the ballast lamp not lighting. If it brightly lights up, you must troubleshoot and find your error before proceeding. Otherwise, note a picture-perfect square wave on the scope with the lamp only dimly lighting. Adjust R21 full clockwise and adjust trimpot R10 for an 80-usec pulse period. This corresponds to 12.5 kHz. Experienced builders may wish to monitor the current to T1 using a current probe attachment to the scope. This waveshape should be another picture-perfect saw tooth of 8 amps maximum. This is a reactive current and is not reflected as an input current due to the phase angle.
5. Turn the unit off and unshort the output leads. Adjust the spark switch screws 1/8 of an inch.
6. Turn on R21/S1 and very quickly rotate clockwise, noting the unit shutting down as apparent by the absence of the waveshape at the test point. If this does not occur instantly, you will need to readjust the switch gap or trimpot R4 clockwise until the shutdown activates. If it does not, you must troubleshoot the circuit. This completes the driver section of this system.

Assembly of the Multiplier Section

Assembly of this section requires solder joints to be large, smooth, and globular. This is contrary to normal soldering practices but is necessary to reduce hv leakage.

1. Fabricate the MULTIBOARD board and two BRK2 brackets, as shown in Figure 15-13.
2. Assemble the diodes and components as shown in Figure 15-14, but do not solder at

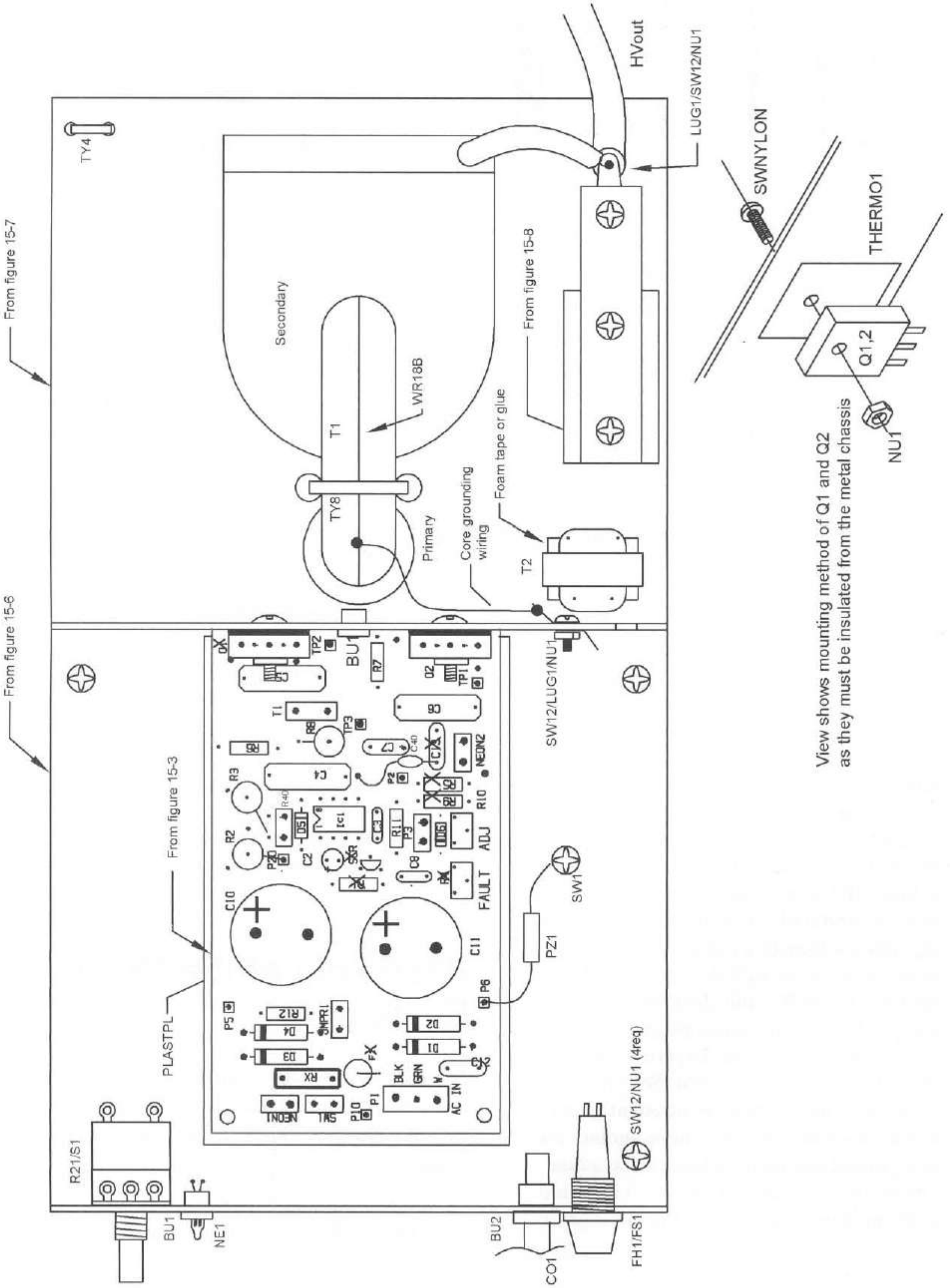
From figure 15-7
From figure 15-6
From figure 15-3
From figure 15-8
PLASTPL
R21/S1
BU1
NE1
BU2
CO1
FH1/FS1
SW12/NU1 (4req)
SW1
PZ1
P6
C10
C11
FAULT
ADJ
R10
NEON1
NEON2
AC IN
BLK
GRN
W
RX
D1
D2
D3
D4
R12
SW12/LUG1/NU1
BU1
T2
Core grounding wiring
Foam tape or glue
Primary
TY8
T1
WR18B
Secondary
TY4
LUG1/SW12/NU1
HVout
SWNYLON
THERMO1
Q1,2
NU1

View shows mounting method of Q1 and Q2 as they must be insulated from the metal chassis

Figure 15-9 *Parts layout to driver*

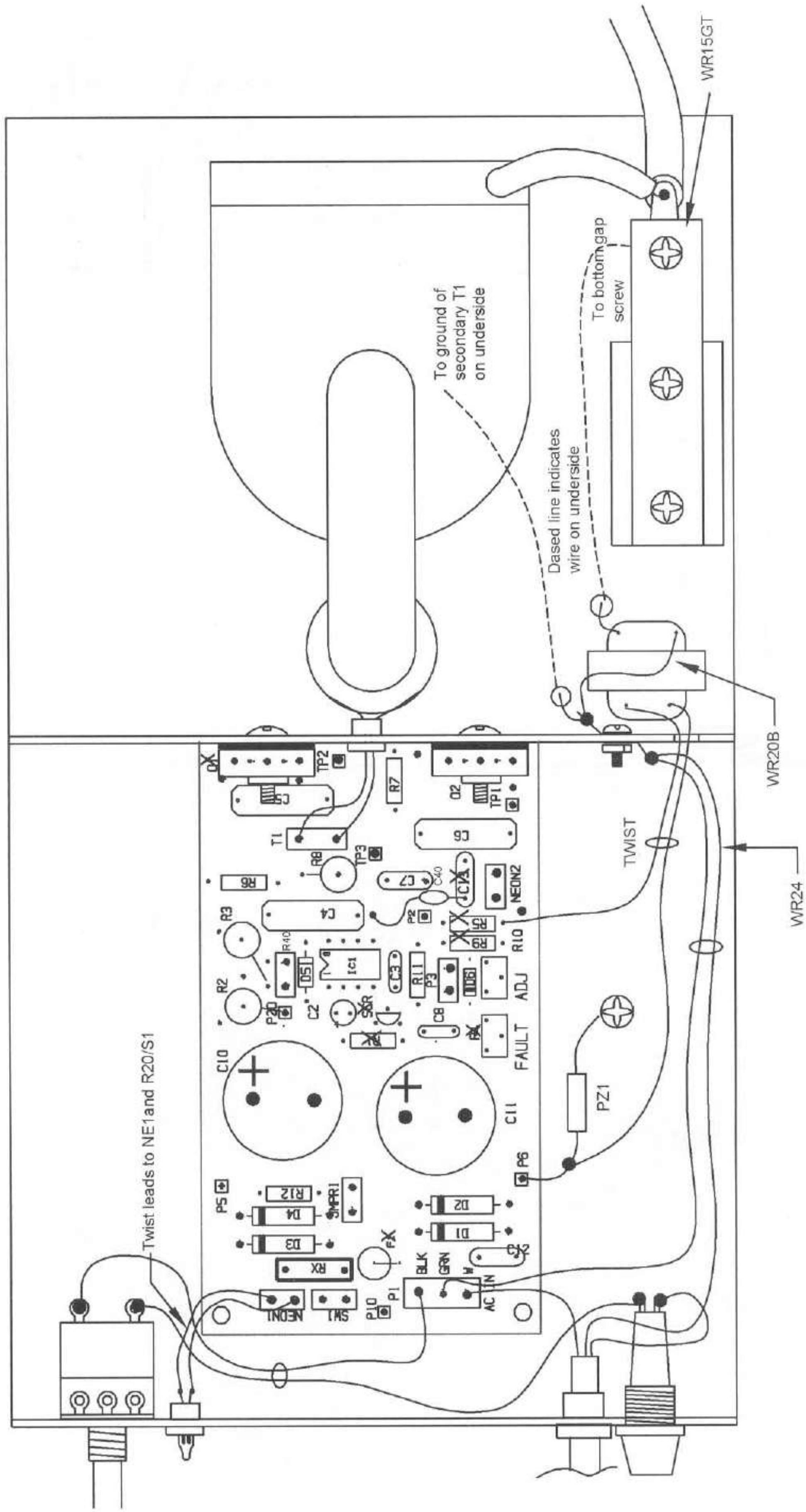
WR15GT
To bottom gap screw
To ground of secondary T1 on underside
Dased line indicates wire on underside
WR20B
TWIST
WR24
PZ1
Twist leads to NE1and R20/S1
Q1
TP2
C5
R7
Q2
TP1
T1
R8
TP3
C6
R6
C7
C40
C14
NEON2
C4
P2
R5
R9
R10
R3
R40
D51
IC1
C3
R11
P3
D6
ADJ
R2
P20
C2
C8
FAULT
C10
C11
P6
P5
R12
D4
D3
D2
D1
FX
RX
C12
BLK
GRN
W
AC LIN
P1
P10
SW1
NEON1

Figure 15-10 *Primary wiring aid*

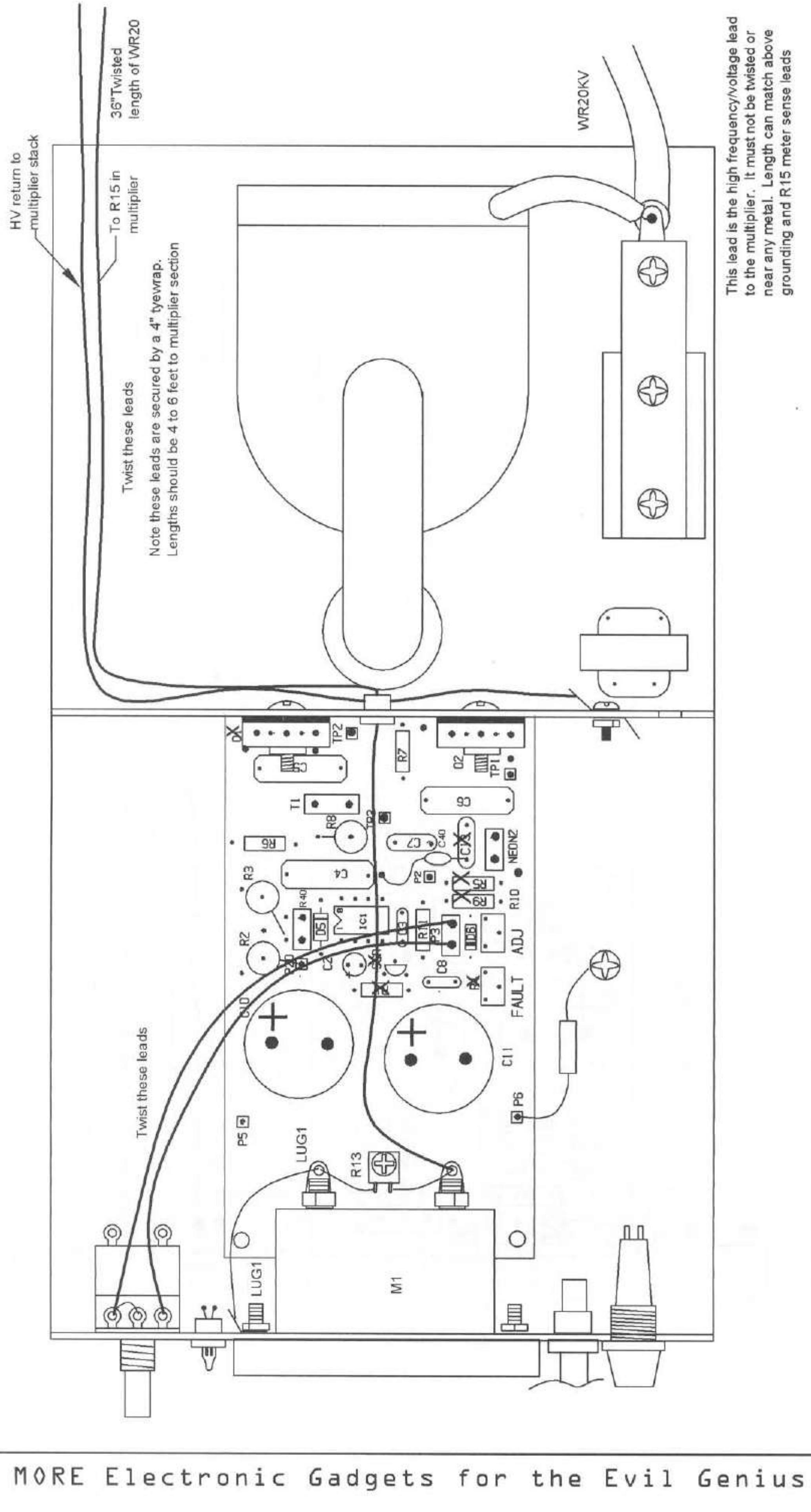
HV return to multiplier stack
To R15 in multiplier
36" Twisted length of WR20
WR20KV
M1
LUG1
LUG1
R13
P5
P6
C10
C11
FAULT
ADJ
NEON2
TP1
TP2
TP3
D2
T1
R2
R3
R5
R6
R7
R8
R9
R10
R11
R40
C2
C4
C5
C6
C7
C8
C40
IC1
D51
P2
P3

Twist these leads

Twist these leads

Note these leads are secured by a 4" tyewrap.
Lengths should be 4 to 6 feet to multiplier section

This lead is the high frequency/voltage lead to the multiplier. It must not be twisted or near any metal. Length can match above grounding and R15 meter sense leads

Figure 15-11 *Secondary wiring aid*

this time. Assemble the resistor tube for R17-20 as shown.

3. Mount the capacitors C14 through C21 and the remaining components as shown in Figure 15-15. Insert the BAF1 baffle plate to insulate and separate the left and right rows.

   Proceed to solder all the joints, noting that the capacitor and diode connections should be smooth, round balls of solder 1/8 to 3/16 of an inch wide. This may be tricky with all the leads but should be done to avoid leakage, especially the last four capacitors.

4. Sleeve resistor R14 into a plastic sleeve, as shown in Figure 15-16.
5. The final assembly is shown in Figure 15-17. Note you have a choice of an output lead or a spherical terminal.

Final Testing

This step requires basic electronic laboratory equipment including a 60 MHz oscilloscope along with a high-voltage meter with a 40 kilovolt dc rating. You will need a high-voltage load resistor to verify output power capabilities.

1. Connect the driver and multiplier sections together and separate them by at least 3 feet. Keep the high-voltage wire from the driver away from other conductive objects. Follow Figure 15-11 for the driver and Figure 15-17 for the multiplier interconnecting wiring.
2. Obtain a load resistor of around 10 megohms of at least 100 watts. This resistor must be able to support at least 50,000 volts. Our laboratory load resistor consists of 180 47k, 1-watt resistors connected in series and then inserted into

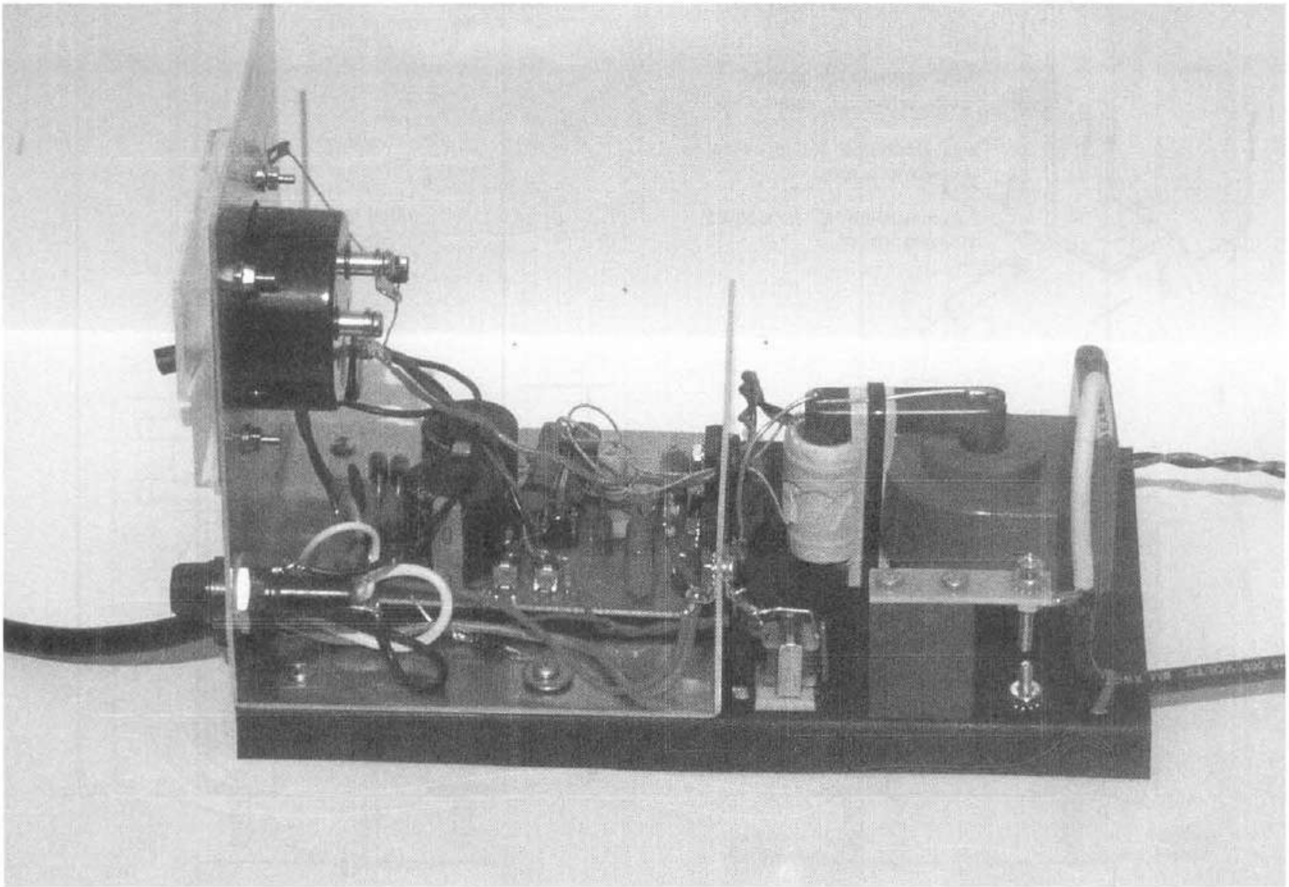

Figure 15-12 *Driver section*

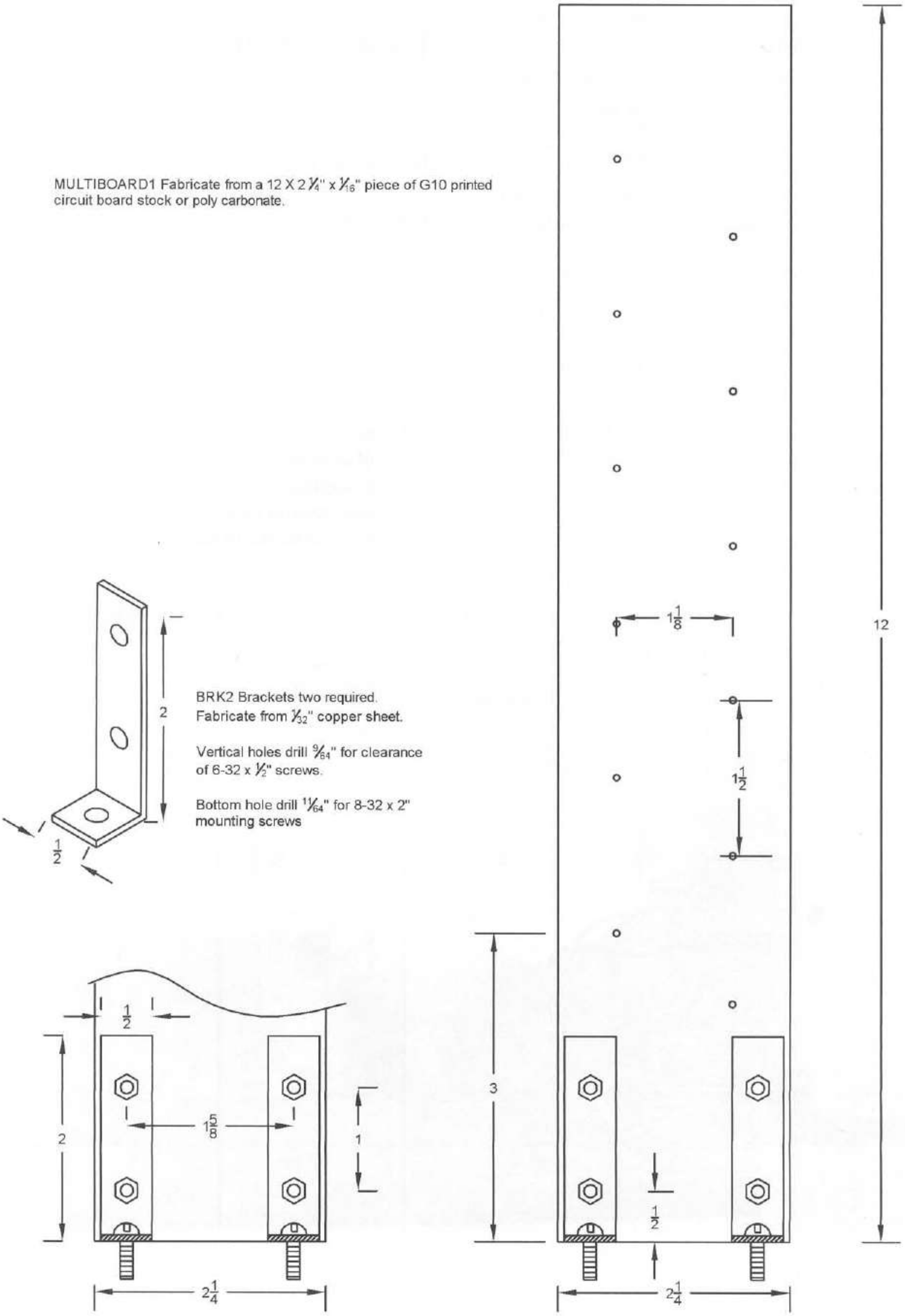

MULTIBOARD1 Fabricate from a 12 X 2 ¼" x 1/16" piece of G10 printed circuit board stock or poly carbonate.

BRK2 Brackets two required. Fabricate from 1/32" copper sheet.

Vertical holes drill 9/64" for clearance of 6-32 x ½" screws.

Bottom hole drill 11/64" for 8-32 x 2" mounting screws

Figure 15-13 *Creating the multiplier board*

some polyethylene tubing, along with the terminating high-voltage leads. The tubing is then coiled, maintaining adequate clearance of the terminating leads. This is a very tedious assembly but works well for powers up to 300 watts at 100 kilovolts. Connect the load to the output and common leads.

3. Assuming that the driver is properly functioning, proceed to apply power and slowly turn the power control pot R21 to full clockwise, noting a line current draw of about 2 amps. Very carefully measure around 35 kilovolts at the output using a 40-kilovolt probe such as an HV44 or similar. The load resistors should start to get warm.

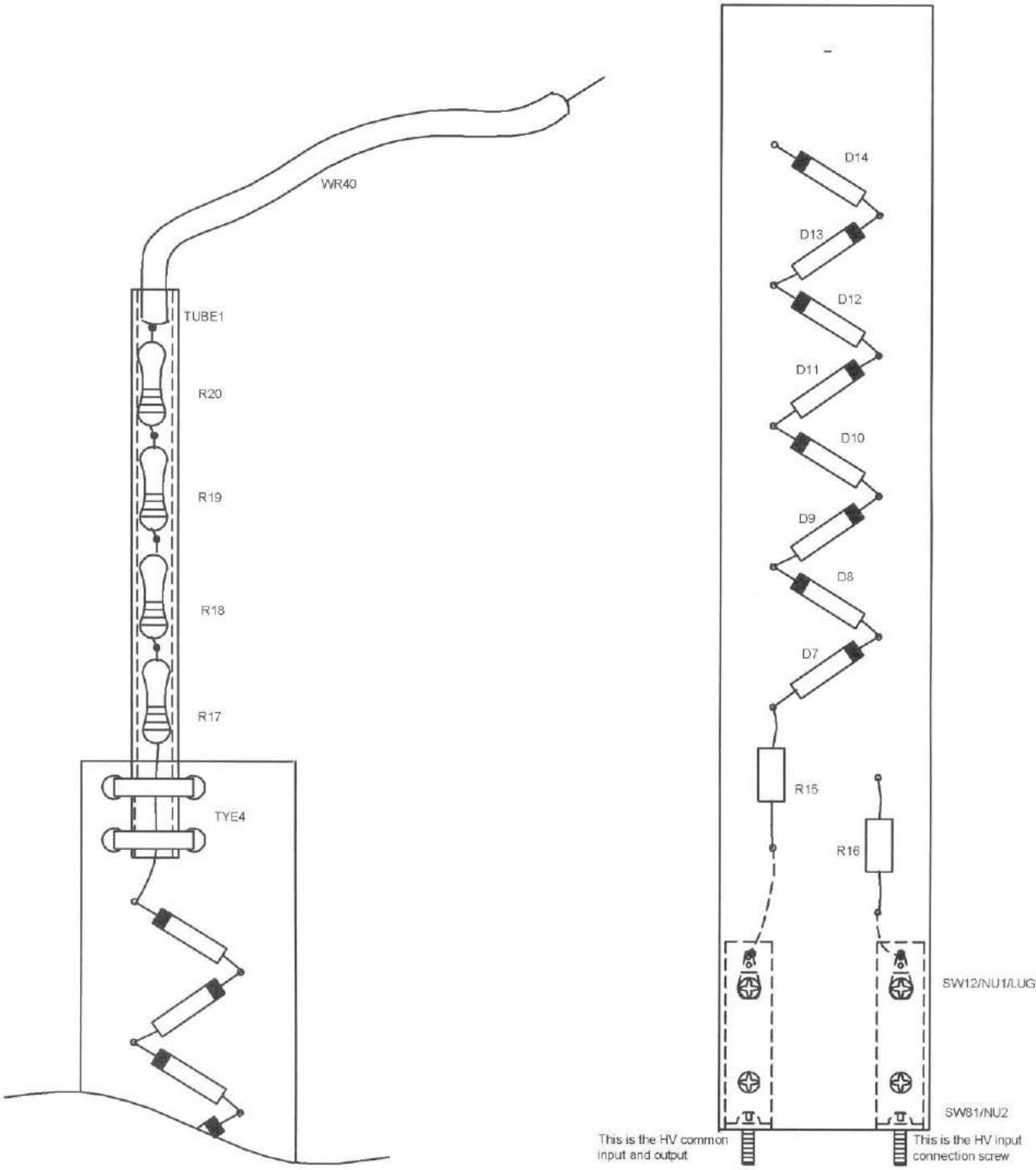

This is the HV common input and output

This is the HV input connection screw

Figure 15-14 *Diode layout and wiring*

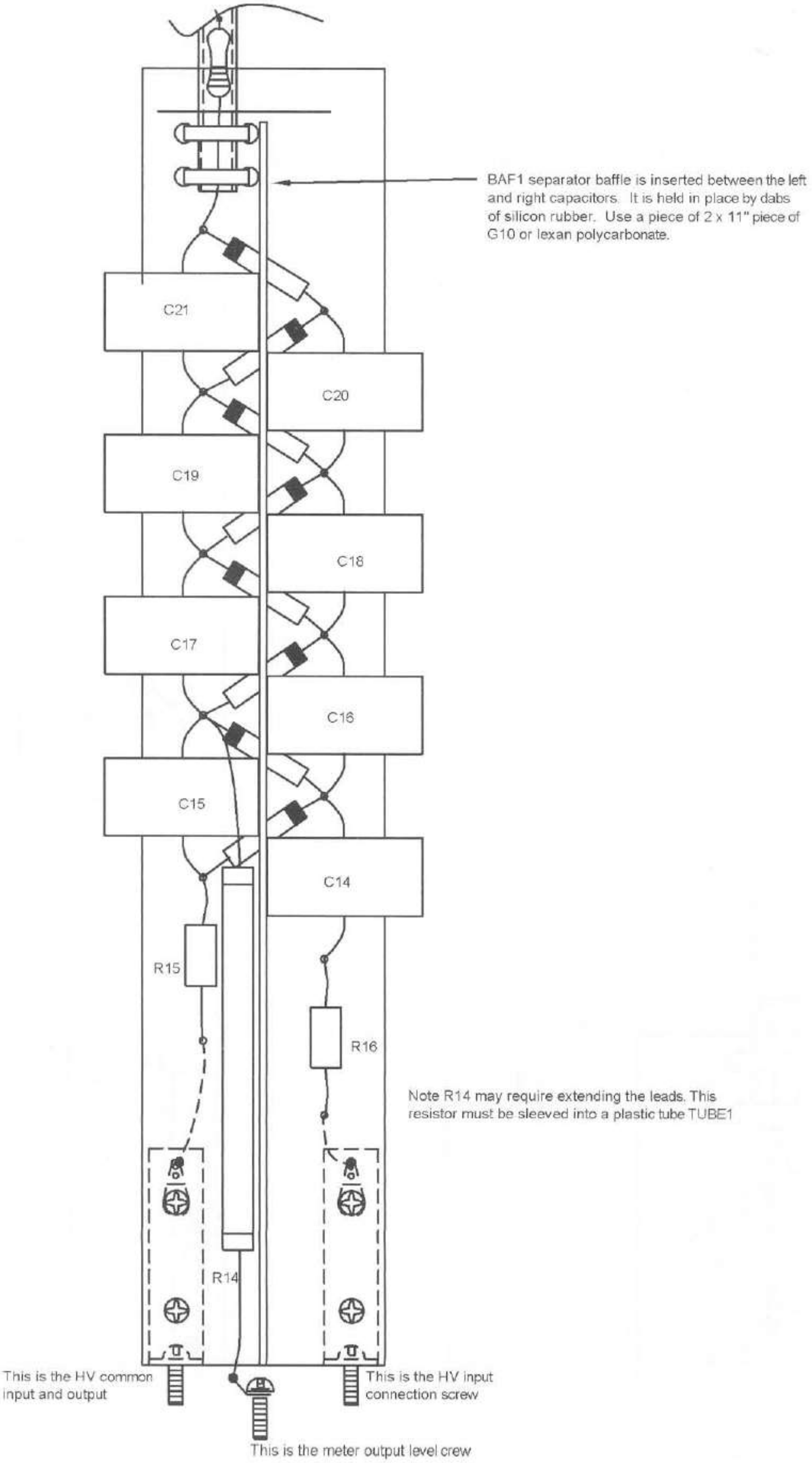

BAF1 separator baffle is inserted between the left and right capacitors. It is held in place by dabs of silicon rubber. Use a piece of 2 x 11" piece of G10 or lexan polycarbonate.
C21
C20
C19
C18
C17
C16
C15
C14
R15
R16
R14
Note R14 may require extending the leads. This resistor must be sleeved into a plastic tube TUBE1
This is the HV common input and output
This is the HV input connection screw
This is the meter output level crew

Figure 15-15 *Mount capacitors and remaining components as shown here.*

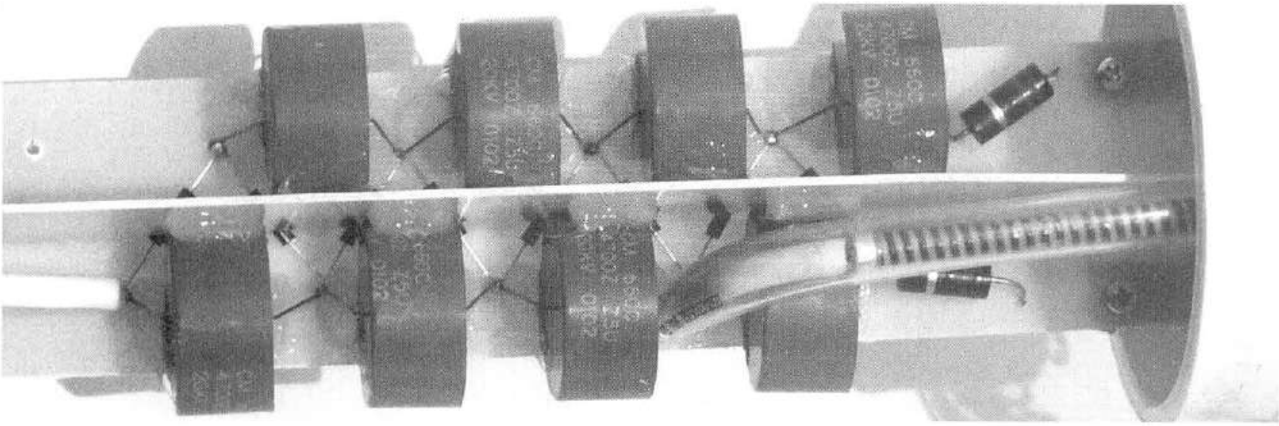

Figure 15-16 *Multiplier internal photo*

4. Allow to it run for several minutes and note that Q1 and Q2 are only slightly warm to the touch.
5. Reverify the operation of the shutdown circuit to prevent damage to the circuit components in the event of over-volting.

Applications

This unit is excellent to use for powering large antigravity lifters because it is not damaged by short-circuiting the output. It can also be used for charging capacitors up to 50,000 volts or as a voltage source with a load resistor of 20 megohms. The charging current is 4 to 5 milliamps and charges as a current source.

DANGER! Improper contact or use of high-energy capacitors can result in death by electrocution or injury due to explosive discharges.

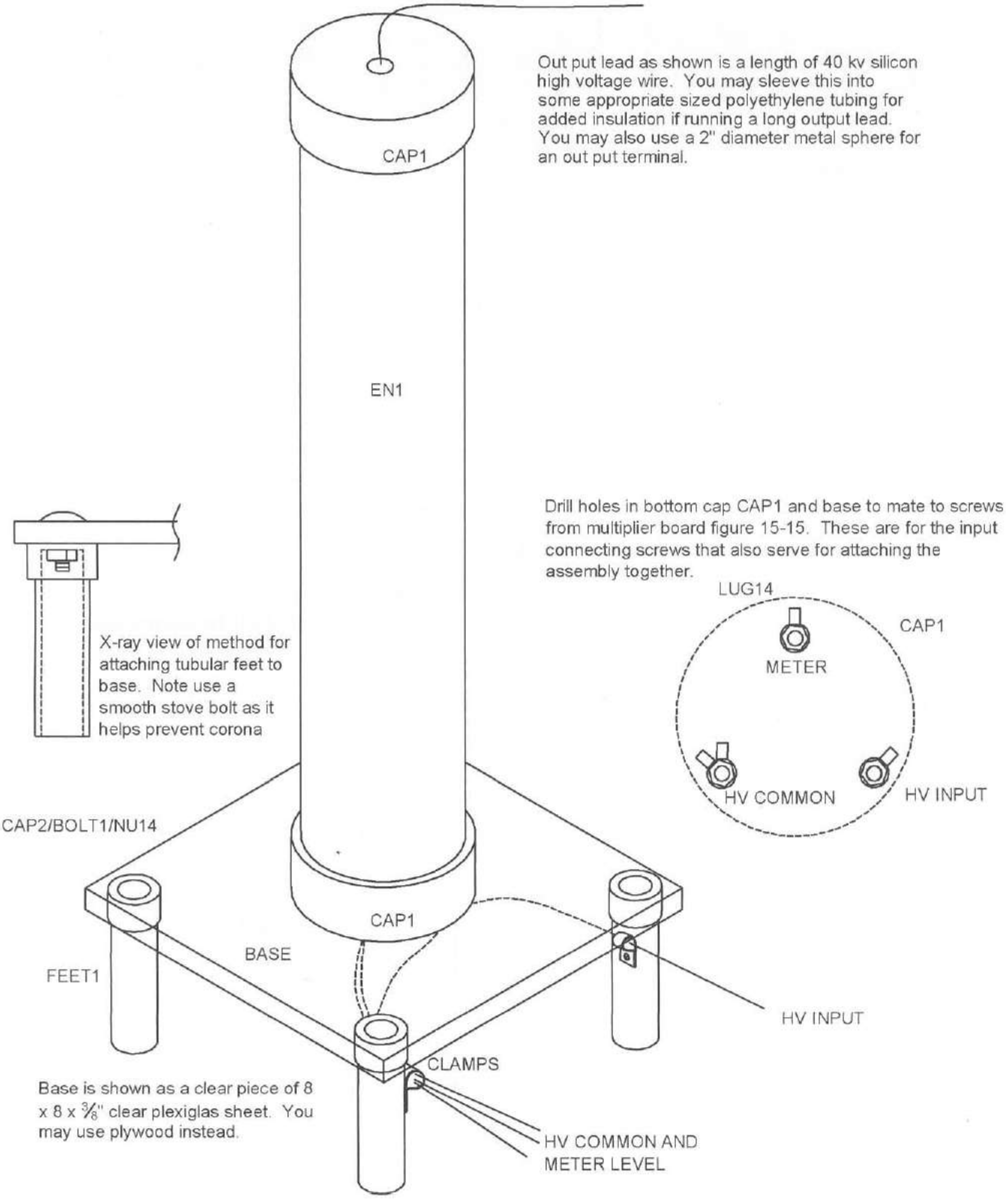

Out put lead as shown is a length of 40 kv silicon high voltage wire. You may sleeve this into some appropriate sized polyethylene tubing for added insulation if running a long output lead. You may also use a 2" diameter metal sphere for an out put terminal.

Drill holes in bottom cap CAP1 and base to mate to screws from multiplier board figure 15-15. These are for the input connecting screws that also serve for attaching the assembly together.

X-ray view of method for attaching tubular feet to base. Note use a smooth stove bolt as it helps prevent corona

Base is shown as a clear piece of 8 x 8 x ⅜" clear plexiglas sheet. You may use plywood instead.

CAP1
EN1
LUG14
CAP1
METER
HV COMMON
HV INPUT
CAP2/BOLT1/NU14
CAP1
BASE
FEET1
HV INPUT
CLAMPS
HV COMMON AND
METER LEVEL

Figure 15-17 *Final assembly*

Table 15-1 50-Kilovolt Power Supply

Ref. #	*Description*	*DB Part #*
R1	100-ohm, 1/4-watt resistor (br-blk-br)	
R2, 3	Two 18K, 3-watt resistors	
R4, 13	Two 2K vertical trimmers	
R6, 7	Two 15 1/4-watt resistors (br-grn-blk)	
R8	10-ohm, 3-watt metal oxide resistor	
R10	5K vertical trimmer	
R5, 11	Two 1K 1/4-watt resistors (br-blk-red)	
R9, 12	Two 100K 1/4-watt resistors (br-blk-yel)	(KC006-LND)
Rth	limiter 5-amp in-rush	
R15, 16	Two 12-ohm, 2-watt carbon resistors (br-red-blk)	
R14	100 meg, 20-kilovolt tiger resistor	
R17–20	4 100K, 2-watt ceramic resistor	DB# 100KCER
R21/S1	10K pot and 115 VAC switch	
C2	10-microfarad, 50-volt vertical electrolytic capacitor	
C3	.01-microfarad, 50-volt plastic capacitor (103)	
C4	.1-microfarad, 600-volt plastic capacitor	
C5, 6	Two 1.5-microfarad, 400-volt metal polyester capacitors	
C7	.0015-microfarad, 600-volt polypropylene capacitor	
C8	.47-microfarad, 50-volt plastic capacitor	
C10, 11	Two 200- to 800-microfarad, 200-volt vertical electrolytic capacitors	
C12, 13	Two .01-microfarad, 2-kilovolt disc capacitor	
C14–21	Eight .0047-microfarad, 20-kilovolt ceramic capacitors	DB#.0047/20 KV
PZ1	420-volt suppressor	
D1, D2, D3, D4	Four 1,000-volt, 3-amp rectifier (1N5408)	
D5	1-kilovolt, 1-amp fast diode (1N4937)	
D6	1N914 small-signal diode	
D7–14	Eight 30-kilovolt, 5-milliamp fast-recovery diodes	DB# 3VG30
SCR	EC103 A, 12 uamp sensitive gate *silicon-controlled rectifier* (SCR)	
Q1, 2	IRFP450/460 *metal-oxide-semiconductor field effect transistor* (MOSFET)	
IC1	IR2153 half-bridge driver	
T1	High-volt transformer assembled as shown in Figure 15-4	DB# FLYPVM
T2	8/500-ohm audio transformer	
SO8	Eight-pin DIP socket	
NE1	Neon indicator bulb with leads	
FH1	Fuse holder panel mount	
FUS1	3-amp slow blow 3AG fuse	
METER	50 uamp, 3-panel meter	DB# METER50L
CO1	Three-wire line cord	
LUG1	Eight #6 solder lugs	
LUG14	Three 1/4-20 lugs	
THERM1	Two thermo mounting pads for Q1 and Q2	
PCLINE	*Printed circuit board* (PCB)	DB# PCLINE
BU38	Three 3/8-inch plastic bushings	
BU58CL	5/8-inch clamp bushing for cord CO1	
WR15GTO	4 feet of 15-kilovolt GTO flexible wire	
WR40KV	12 inches of 40-kilovolt high-voltage silicon wire	DB# WR40KV
WR20	10 feet of #20 vinyl hookup wire	
WR24	12 inches of # 24 vinyl hookup wire	
WR20B	12 inches of # 20 bus wire	
WR18B	24 inches of #18 bus wire	

Table 15-1 *continued*

Fabrication

Ref. #	*Description*	*DB Part #*
CHASSIS	Chassis created from .065-inch aluminum as per Figure 15-6	
COVER	Cover created from plastic sheet to fit chassis	
BASEPLATE	Base plate created from .065-inch Lexan per Figure 15-7	
BRK1	Bracket created from .035-inch copper per Figure 15-8	
BRK2	Two brackets created from .035-inch copper per Figure 15-13	
MTB1	Mounting block created from .75-inch PVC plastic	
PLASTPL	4.5 × 2.5 piece of insulating plastic, shown in Figure 15-9	
MULTIBOARD1	Multiplier board created from .065-inch Lexan per Figure 15-13	
BASE	Base created from 8 × 8 × 3/8-inch Plexiglas per Figure 15-17	
BAF1	Baffle separator created from .065-inch Lexan per Figure 15-15, available from a hardware store	
SW1	Five #6 1/4-inch sheet metal screws for cover and MTB1	
SWNYLON	Two 6-32 1/2-inch nylon screws	
SW12	Ten 6-32 1/2-inch screws	
SW10	Two 6-32 × 1-inch screws	
SW81	Three 8-32 × 2-inch screws	
BOLT1	Four 1/4-20 × 1 1/2-inch stove bolts	
NU1	Sixteen 6-32 keep nuts	
NU2	Six 8-32 keep nuts	
NU14	Four 14-20 hex nuts	
CAP1	Two 3-inch PVC flat caps (GENOVA# 70153)	
CAP2	Four 1/2-inch PVC flat caps (GENOVA# 30155)	
FEET1	Four pieces of 1/2 × 3-inch PVC tubing	
EN1	3 × 15-inch PVC tubing	
TUBE1	10 inches of 3/8 ID × 1/2 OD flexible plastic tubing	

Chapter Sixteen

Magnetic High-Impact Cannon

This very advanced project (shown in Figure 16-1) is intended for the experienced experimenter and researcher in the field of high-voltage and high-energy circuitry. The magnetic cannon accelerates a projectile at high velocities with considerable kinetic energy; this is accomplished strictly from a magnetic pulse. Not only is this project a dangerous electrical device, it is also kinetically hazardous, as the accelerated projectile can cause serious injury and even death. The same consideration and respect given to a firearm must be given to this project. A video demonstration of this device blasting large holes through a wall can be seen at www.amazing1.com.

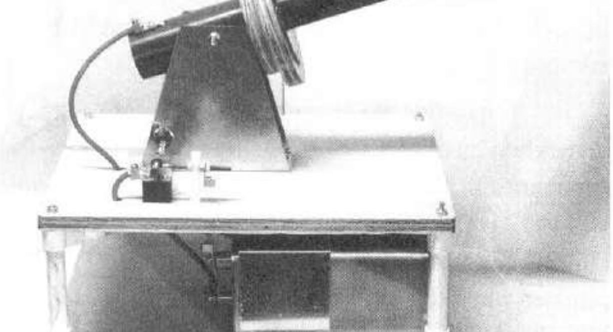

Figure 16-1 *The magnetic cannon*

Under the correct supervision, the magnetic cannon can provide a high-action science project demonstrating several important electrical laws that involve electromagnetic reactions. Lenz's law and the Lorentz JXB forces are clearly utilized in this device. The system as shown can provide the experienced hobbyist hours of fun and entertainment by experimenting with its effect (its impact) on various objects.

This is an advanced level project requiring electronic skills and high-voltage experience. Expect to spend $300 to $500 unless you have access to the surplus market. All parts are readily available and any specialized parts are available through Information Unlimited (www.amazing1.com) and are listed in the parts list at the end of the chapter.

Theory of Operation

A nonmagnetic conductor such as aluminum is placed in a time-variant magnetic field. Induced currents in the aluminum produce currents that, in turn, produce opposing magnetic fields thereby causing a moment of acceleration of the aluminum piece. In this project the aluminum piece is in the form of a large flat washer that is the projectile. The aluminum projectile has a hole in its center with a mandrel guide to keep it traveling in a straight line.

The heart of this project is the accelerating coil. It is what couples the energy (the electrical current) to the aluminum ring (the projectile). Optimum efficiency is dependent on the coupling to achieve maximum projectile kinetic energy. This requires a minimal proximity of the projectile to the coil along with minimal air gaps (such as the spacing between the wires of the coil and the geometry of the coil as related to the dimensional format of the projectile). Also the inductive value of the coil is related to the storage capacity to provide the current rise over a time period that is dependent on the physical parameters of the projectile. The associated second-order differential equations that mathematically determine these parameters are beyond the scope of this material. Although the mathematical purist will find a deviation from maximum efficiency by using different materials, practically speaking, there are materials of different sizes and materials that offer a cost-effective compromise. An example is the use of square magnetic wire in place of conventional round stock wire. Using square magnetic wire for the coil, when properly wound, will provide more kinetic energy to the projectile due to reduced air space (reluctance); however it is quite difficult to wind and usually requires purchasing a significant amount at a healthy price.

The shape and timing parameters of the coil-magnetizing pulse must be related to the projectile to achieve optimum efficiency. A current rise that is too fast will cause slippage (magnetic cavitation). Note that one definite disadvantage of this method of acceleration is that the projectile is influenced over a very short distance. Velocity, now being the square root of the acceleration times the distance, is limited. To achieve high resultant velocities requires a very high moment of acceleration. Projectiles using conventional explosives current detonators and boosters could be prone to sympathetic initiation by these high g accelerating forces at the time of launching.

Circuit Theory

This project as shown in Figure 16-2 is constructed operating from 115 vac household current. It also can be built to operate from 12 volts or built-in batteries.

A high-voltage current-limited 60 Hz transformer (T30) steps up the 115 vac to 6,500 vac and is rectified by high-voltage diodes (D31 through D34). DC current now charges the energy storage capacitors (C30) through the isolation resistor (R30) to a programmable value, as selected by the operator. It is this stored energy that is discharged into the accelerator coil (L1) as it is switched by the spark gap (GAP1). Once switched into the coil, the now rapidly rising current wave induces a current into the aluminum ring projectile (PROJ1). It is this induced current that now generates a very high magnetic moment and repels the initial field in the accelerator coil, causing a moment of intense accelerative forces. Those not familiar with this concept often ask why an aluminum ring? The answer is that a magnetic material would now be attracted and thus would neutralize the repulsion.

The initializing of the circuit commences by turning on the key switch (S2). This switch is intended to keep unauthorized personnel from powering up the system. The key switch controls 12 volts of DC power necessary to energize the relay (RE1), with normally open contacts controlling power to the high-voltage transformer (T30).

The controlling system, as shown, consists of momentary pushbutton switches that start the charging action (S4) and can stop this action via switch (S5). Triggering the momentary pushbutton switch (S3) supplies power to the trigger module (TRIG10) and firing trigger gap (GAP2), thereby initiating the main gap GAP2 and switching the energy from the storage capacitors C30 into the accelerator coil L1. Charging voltage to C30 is controlled by the potentiometer R14. Once set, the voltage will maintain its preset level until triggered or readjusted. Meter M1 indicates the charge voltage and is calibrated by the trimpot (R16).

The transformer (T1) supplies 12 volts of AC that is rectified by diodes (D1 through D4), filtered by capacitor C4, and regulated by zener diodes (Z1 and Z2) to 12 volts dc in order to power the control circuits. The indicator *light-emitting diode* (LED1) illuminates when the key switch is energized. The second indicator LED (LED2) illuminates when charging of C30 is taking place. A buzzer (BUZ1) sounds whenever there is a charge voltage on C30. This is a safety

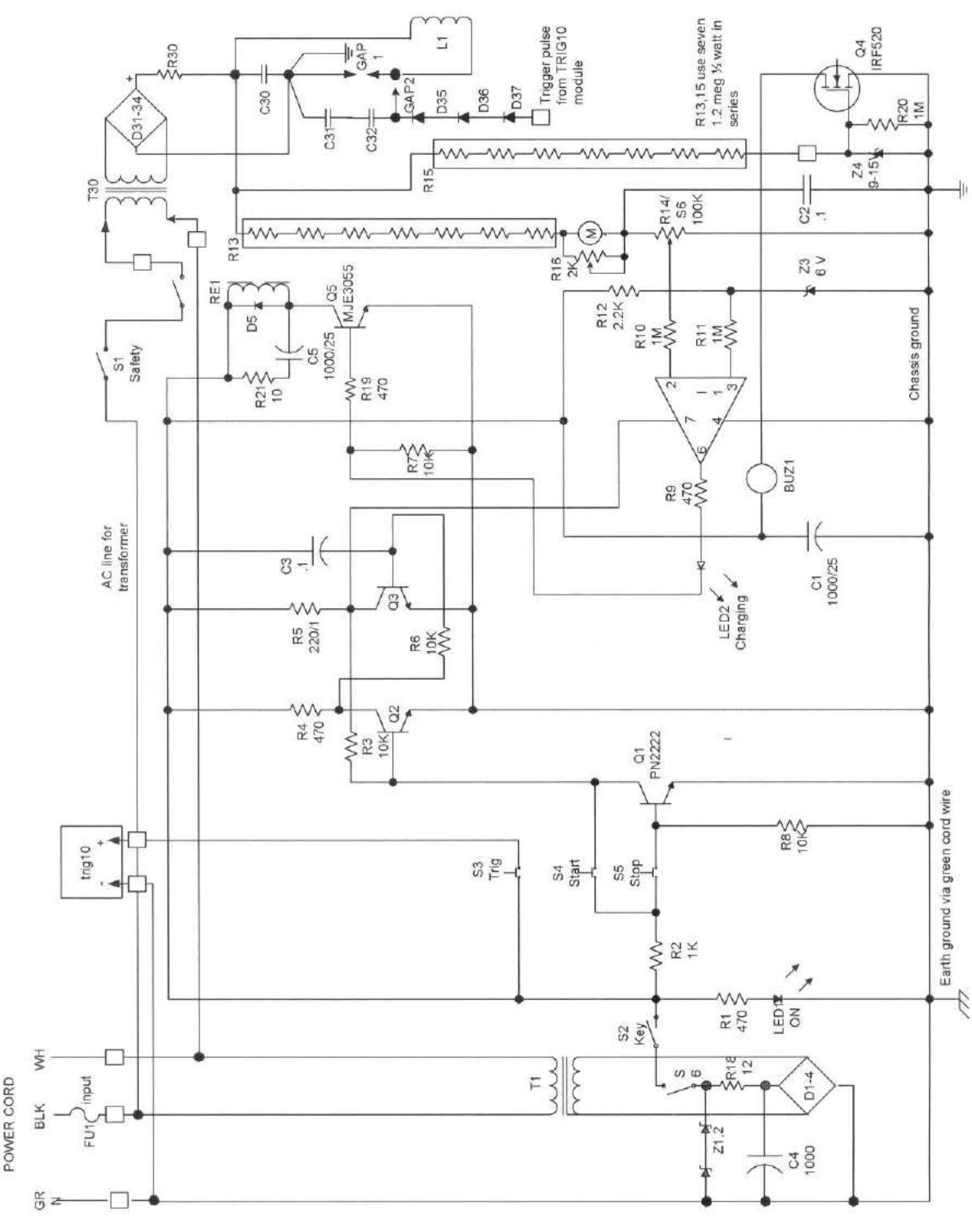
POWER CORD
GR
BLK
WH
FU1
input
T1
S2 Key
S 6
R18 12
D1-4
Z1,2
C4 1000
R1 470
LED1 ON
R2 1K
S4 Start
S5 Stop
S3 Trig
trig10
Earth ground via green cord wire
R8 10K
Q1 PN2222
R3 10K
R4 470
Q2
R5 220/1
Q3
R6 10K
C3 .1
AC line for transformer
LED2 Charging
C1 1000/25
R9 470
BUZ1
Chassis ground
R7 10K
R19 470
R21 10
C5 1000/25
D5
Q5 MJE3055
RE1
S1 Safety
R10 1M
R11 1M
R12 2.2K
Z3 6 V
R16 2K
R13
M
R14/ S6 100K
C2 .1
T30
D31-34
R30
C30
C31
C32
GAP 1
GAP2
D35
D36
D37
Trigger pulse from TRIG10 module
L1
R15
R13,15 use seven 1.2 meg ½ watt in series
Z4 9-15V
R20 1M
Q4 IRF520

Figure 16-2 *Schematic*

device to warn the operator that a voltage exists on the storage capacitors. Sense voltage is obtained for the meter and charge voltage control circuit via string resistor R13. The safety level voltage is obtained through the resistor string (R15).

Assembly Steps

1. Cut a piece of .1 × .1 grid perforated circuit board (PB1) 6 × 4½ inch. You will have to drill holes for transformer T1 and relay RE1.
2. Insert the components as shown in Figure 16-3, starting from the lower left-hand corner. Count the perforated holes as a guide. Note the polarity marks on capacitors and diodes. It is suggested that you use a socket (SO8) for the LM741 integrated circuit (I1).
3. Wire the components as shown in Figure 16-3 using the leads of the actual components as the connection runs. These are indicated by the dashed lines. Always avoid bare-wire bridges and globby solder joints. Check for cold or loose solder joints. Note the symbols indicating wires to external components and to solder junctions beneath board.
4. Assemble the resistor dividers (R13 and R15) as shown in Figure 16-4 from seven 1.2 meg, ½-watt resistors that are all connected in series and sleeved into a ¼-inch ID flexible plastic tubing. Place these along with the connecting leads. The leads connected to R30 must be rated for 10 kilovolts. Connect the external components using #20 vinyl-jacketed, stranded wire. Attach and solder shunt resistor R16 across meter M1. Sleeve in the connection.
5. Connect the three high-voltage diodes (D35, D36, and D37) in series as shown in Figure 16-5. Make sure there are no sharp edges on solder joints and then sleeve into a plastic tube with a short piece of high-voltage wire lead.

   Connect the remaining components, noting that some may have to be unsoldered when routing through the panel, as shown. Verify proper lengths of connecting leads with other figures in the project.
6. Fabricate the chassis from a sheet of .063 aluminum as shown in Figure 16-6. The front panel holes are shown to approximate layout and size. These should be verified with all parts for hole size and location.
7. Fabricate a 6 × 7½ inch sheet of plastic (PLATE1) to insulate the assembly board connections as shown in Figure 16-7. Assemble T30 to the chassis using screws and nuts. Include solder lugs for grounding the power cord and assembly board.
8. Fabricate a 2¼ × 7 inch piece of plastic (PLATE2) for mounting the four rectifiers (D31 through D34) and resistor R30. Use two-sided sticky tape or silicon rubber to secure. Use a long screw or piece of threaded rod with an insulating washer on top side of resistor.
9. Assemble components to the front panel, as shown in Figure 16-8, and then finalize the wiring to complete assembly. Use pieces of 10 kilovolt-rated wire for those that are shown as heavy traces. Other leads are made from #20 vinyl-jacketed, stranded wire. The grounding leads shown are from the power cord, assembly board, and D33/D34 to the lug on T30. Output leads are shown as the TRIGGER, COMMON GROUND, and HV OUTPUT. Observe all notes on this figure.

Circuit Testing

10. Obtain a ballasted 115 vac power source. You can make this simply by placing a 60-watt lightbulb in series with the hot side of the power line, which is usually designated by the black lead. Do not eliminate the green earth ground connection.
11. Verify all switches are off and then insert a 3-amp fuse into FU1. Temporarily connect the high-voltage output to the chassis ground. Output is current limited by the leakage inductance of T30.

12. Plug into the ballasted 115 vac source. Turn on S6 and the key switch. Note LED1 lighting. Check control of both these switches as they are connected in series.
13. Push S3 and note a high-voltage spark at the trigger output lead.
14. Push S4 and note the relay latching and 60-watt lamp starting to light. The unit may chatter in this mode. LED2 should light in coincidence with the relay.

This concludes the basic test of this circuit but does not verify the programming voltage function, meter reading and calibration, or the safety buzzer.

Assembly of the Completed System

15. Fabricate the upper and lower deck pieces, as shown in Figure 16-9. Fabricate four pieces of 5½-inch PVC tubing for the pillar spacers located at each corner.

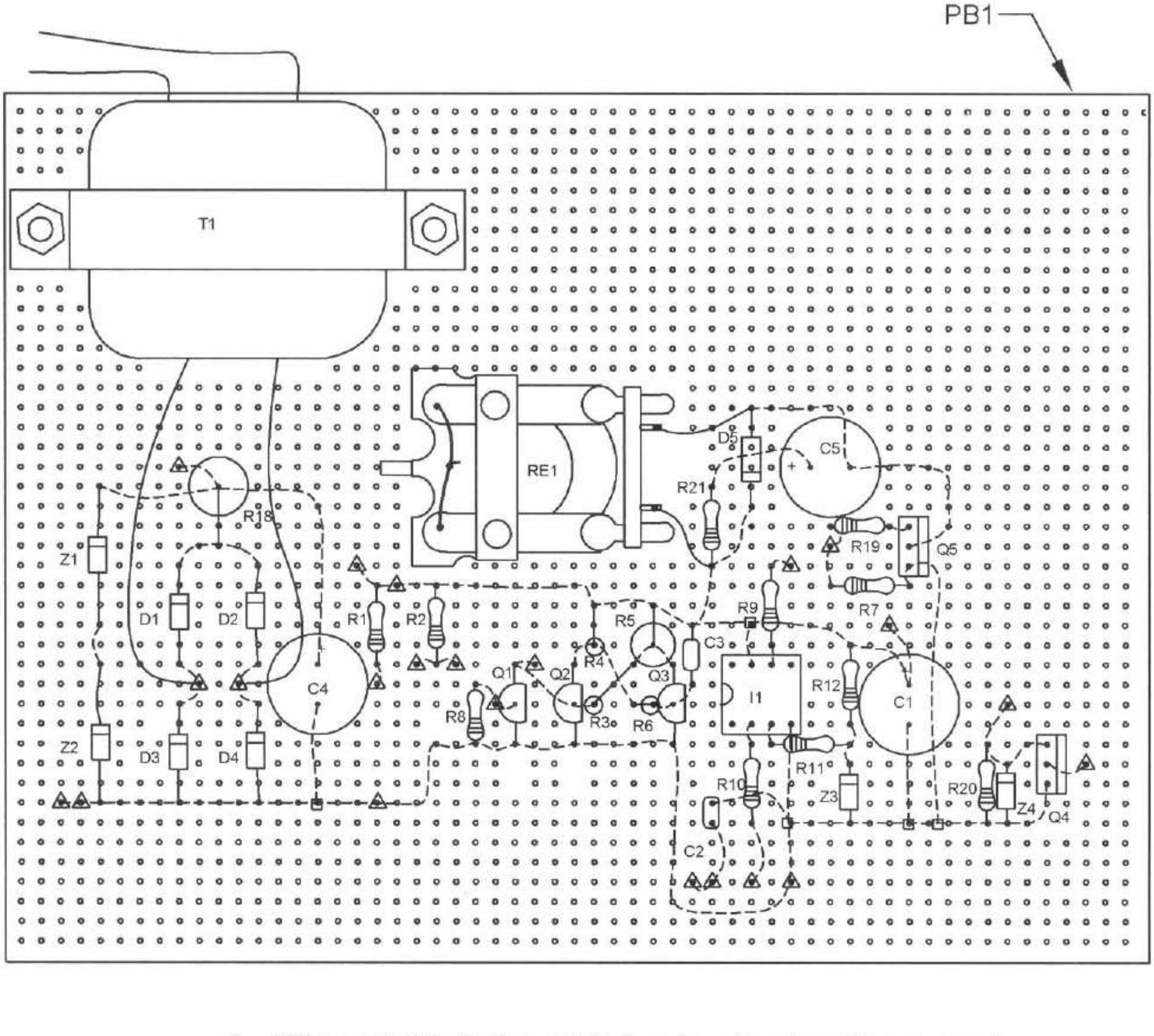

▲ This symbol indicates points for wires to external components

□ This symbol indicates solder junction of leads

~ This symbol indicates wiring on underside of board

Figure 16-3 *Assembly board parts layout*

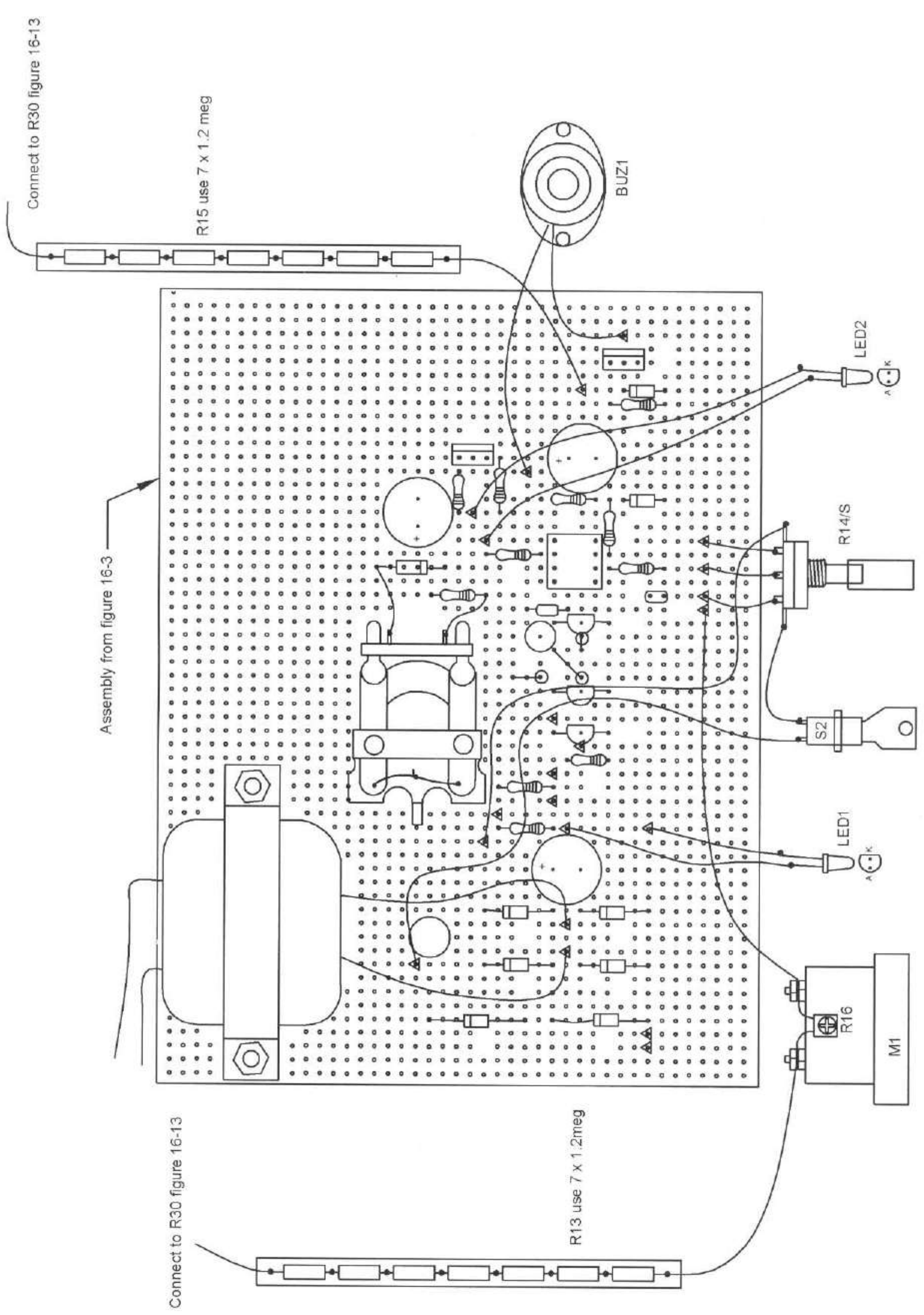
Connect to R30 figure 16-13
R15 use 7 x 1.2 meg
BUZ1
Assembly from figure 16-3
LED2
A K
R14/S
S2
LED1
A K
R16
M1
R13 use 7 x 1.2meg
Connect to R30 figure 16-13

Figure 16-4 *Assembly board external wiring first level*

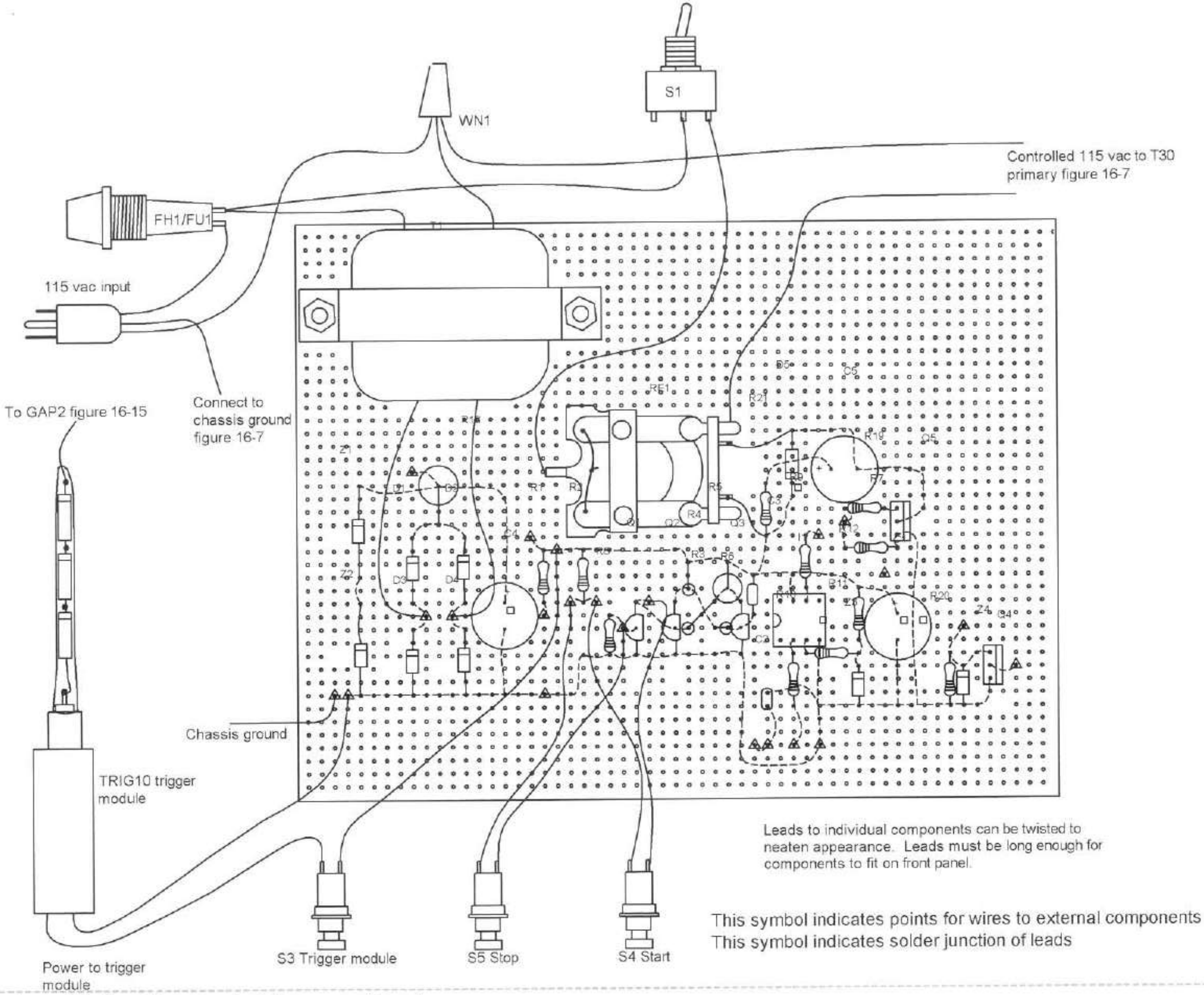
S1
WN1
Controlled 115 vac to T30
primary figure 16-7
FH1/FU1
115 vac input
Connect to
chassis ground
figure 16-7
To GAP2 figure 16-15
T1
RE1
R21
D5
C5
R19
Q5
R7
R8
R12
R13
R11
R20
Z4
Q4
R1
R2
R3
R4
R5
R6
Q1
Q2
Q3
C1
C3
C4
Z1
Z2
D1
D3
D4
R18
Chassis ground
TRIG10 trigger
module
Power to trigger
module
S3 Trigger module
S5 Stop
S4 Start
Leads to individual components can be twisted to neaten appearance. Leads must be long enough for components to fit on front panel.
This symbol indicates points for wires to external components
This symbol indicates solder junction of leads

Figure 16-5 *Assembly board external wiring second level*

Fabricate from .063 aluminum sheet.

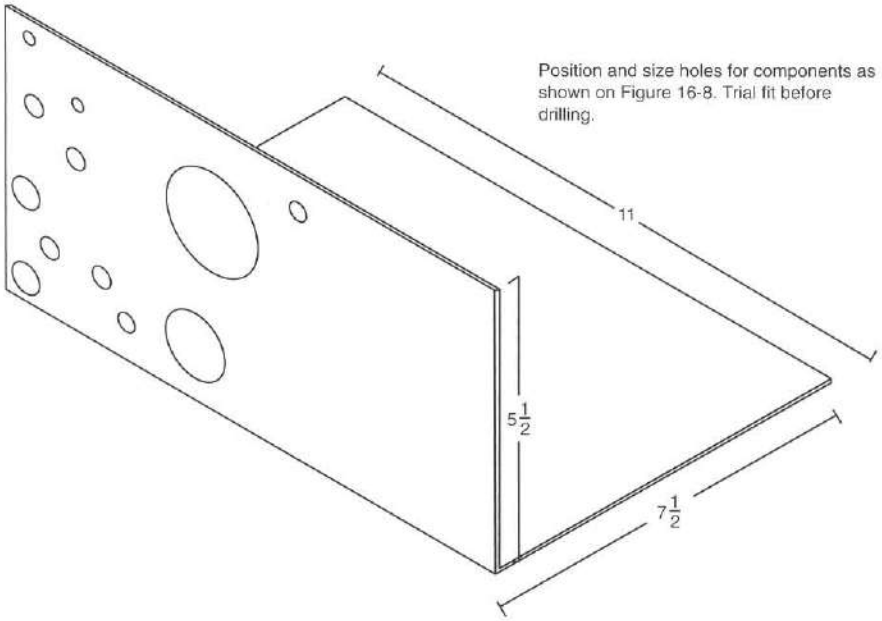

Position and size holes for components as shown on Figure 16-8. Trial fit before drilling.

Figure 16-6 *Fabrication of chassis*

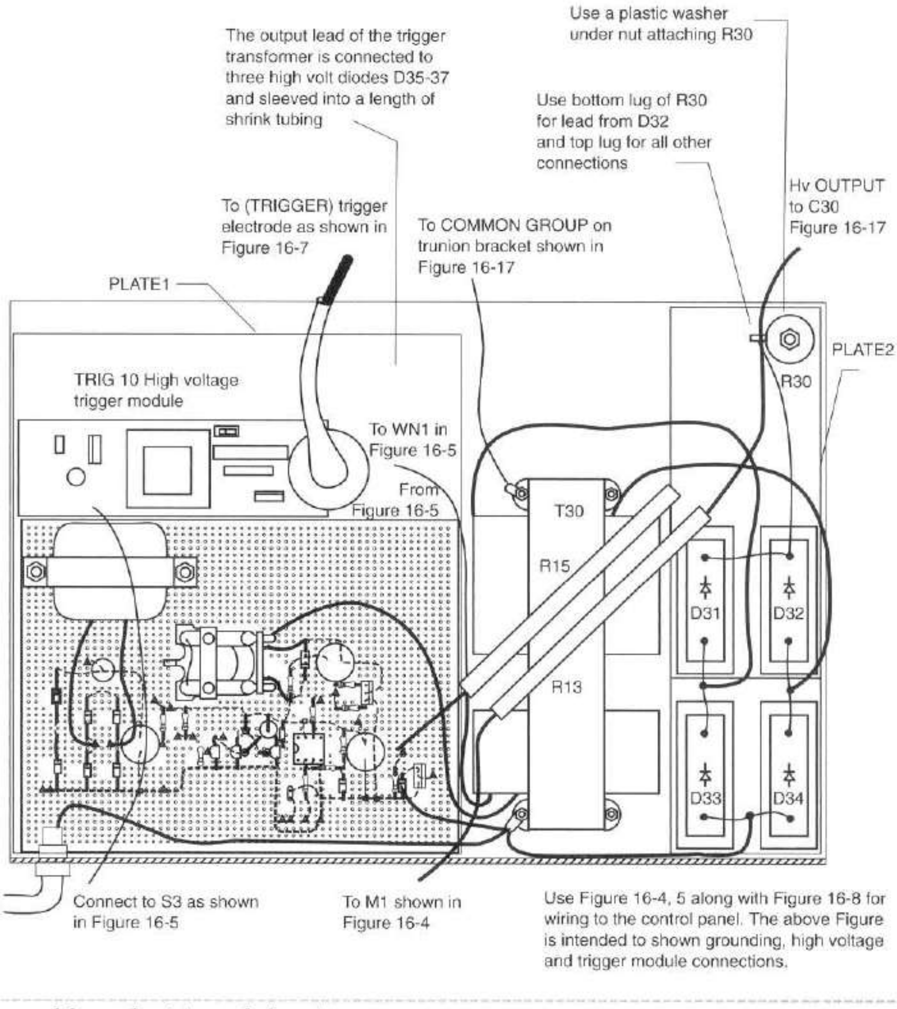

The output lead of the trigger transformer is connected to three high volt diodes D35-37 and sleeved into a length of shrink tubing

Use a plastic washer under nut attaching R30

Use bottom lug of R30 for lead from D32 and top lug for all other connections

To (TRIGGER) trigger electrode as shown in Figure 16-7

To COMMON GROUP on trunion bracket shown in Figure 16-17

Hv OUTPUT to C30 Figure 16-17

To WN1 in Figure 16-5

From Figure 16-5

Connect to S3 as shown in Figure 16-5

To M1 shown in Figure 16-4

Use Figure 16-4, 5 along with Figure 16-8 for wiring to the control panel. The above Figure is intended to shown grounding, high voltage and trigger module connections.

Figure 16-7 *Assembly and wiring of chassis*

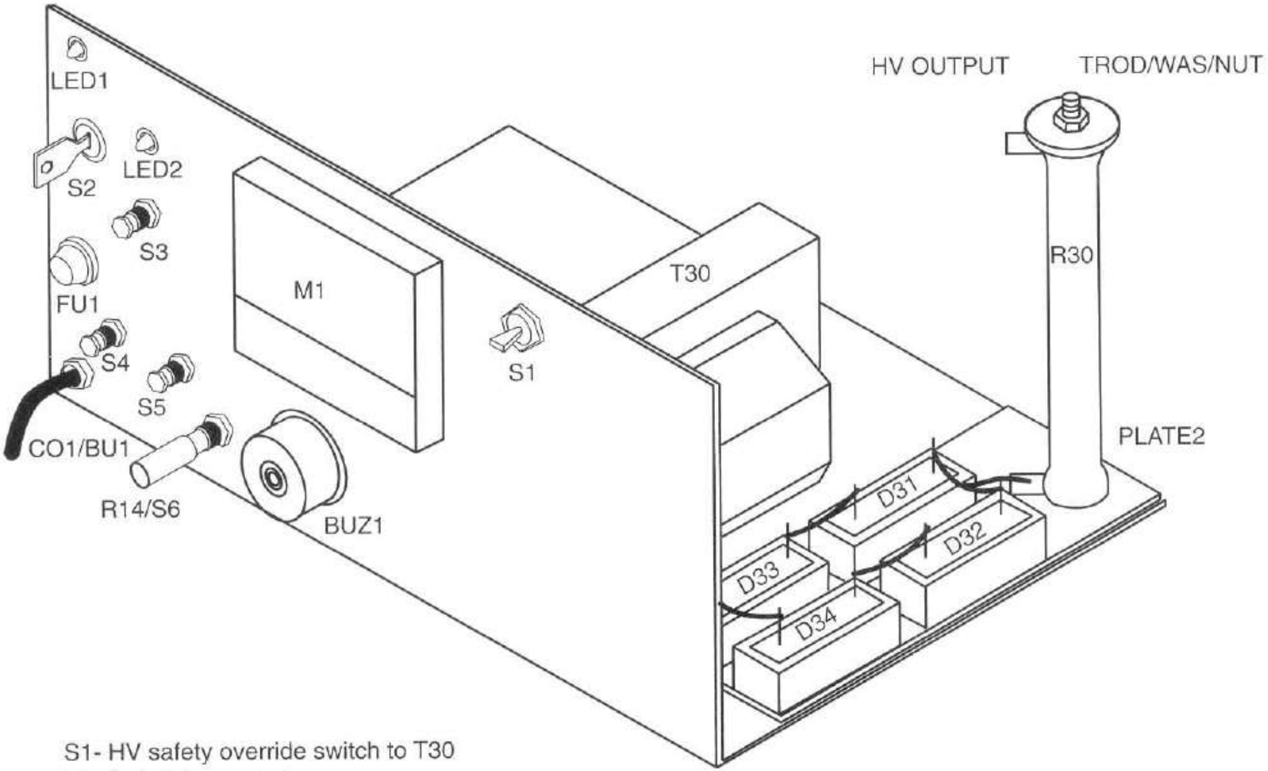

S1- HV safety override switch to T30
S2- Safety key switch
S3- Trigger "fire" switch
S4- Charge start switch
S5- Charge abort switch
S6- Low control voltage switch
LED1- Power "on" indicator
LED2- Charging cycle "on" indicator
R14- Programmed charge voltage level control
FU1- 3 amp fuse holder
BUZ1- High voltage presence audible warning
M1- Charge voltage meter.. 5 kilovolt full scale
CO1/BU1- Clamp bushing for power cord CO1

Note that it is a good idea to label the controls

Figure 16-8 *Front-panel controls and wiring*

16. Fabricate the capacitor bracket, as shown in Figure 16-10. This piece secures the three energy storage capacitors C10A, C10B, and C10C on the bottom deck.

17. Fabricate the trunion bracket, as shown in Figure 16-11. This piece secures the actual flyway and accelerating coil and therefore must be able to withstand the kinetic recoil from the accelerating reaction of the projectile. This bracket is the central electrical grounding point of the discharge circuit and must also be actively grounded to earth ground via the power cord green wire. There is also a $^3/_8$-inch hole for the electrode holding collar (COLLAR1) for the ground electrode (TUNG38) of the main spark switch (GAP1). This part must be centered and carefully soldered in place using a propane torch.

18. Fabricate the front and rear disk section, as shown in Figure 16-12. You may also fabricate the projectile (PROJ1) from a piece of aluminum, as shown.

19. Wind the coils, as shown in Figure 16-13. It is a good idea to study Figures 16-12, 16-13, and 16-14 before attempting this step. Make sure to leave connection leads that are a minimum of 10 inches long.

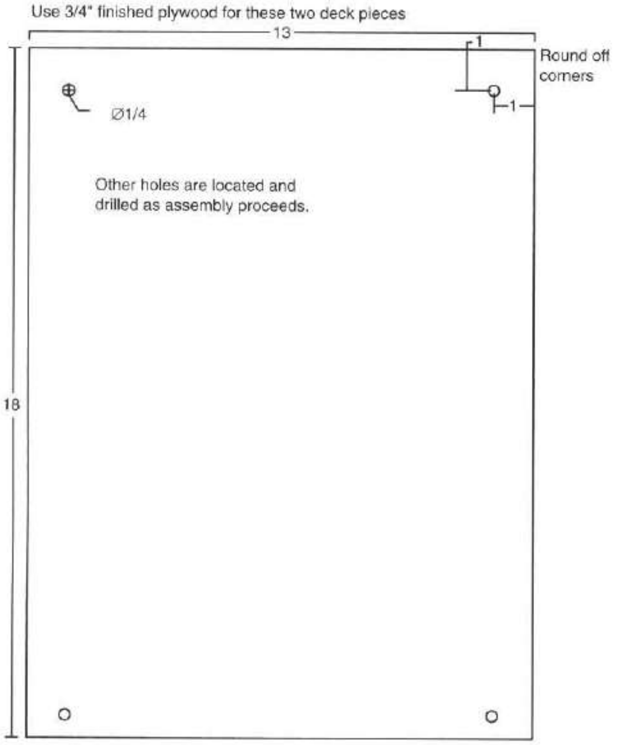
Use 3/4" finished plywood for these two deck pieces

Round off corners

Other holes are located and drilled as assembly proceeds.

Figure 16-9 *Fabrication of deck sections*

20. Assemble the flyway and breech, as shown in Figure 16-14. Study the notes on this figure.

21. Fabricate the two plastic blocks (PVCBLK and TFBLK), as shown in Figure 16-15, to the suggested dimensions. Note that the heights of these pieces must allow close alignment of the electrodes as dictated by the position of the soldered collar on the trunion bracket. C31 and C32 are wired in place and must be spread apart to prevent sparking. The blocks are secured to the top deck using sheet metal screws. Use two screws for the PVC block. Do not allow these screws to penetrate more than 1/2 inch into the plastic material. Note the clearance hole in the deck for the high-voltage lead from the trigger module, shown in Figure 16-7.

22. Assemble the three energy-discharge capacitors (shown in Figure 16-16) along with the charger and control module.

23. Make the final assembly steps and wire, as shown in Figure 16-17. Note the direct and

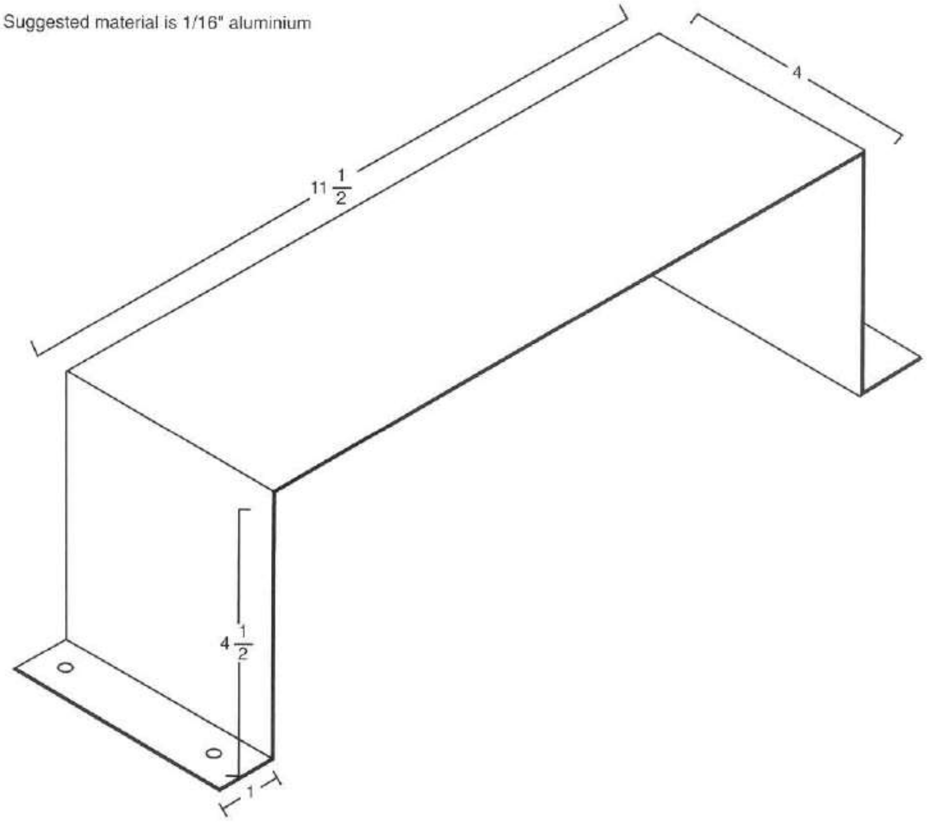
Suggested material is 1/16" aluminium

Figure 16-10 *Fabrication of the capacitor bracket*

Suggested material is 1/16" galvanized steel

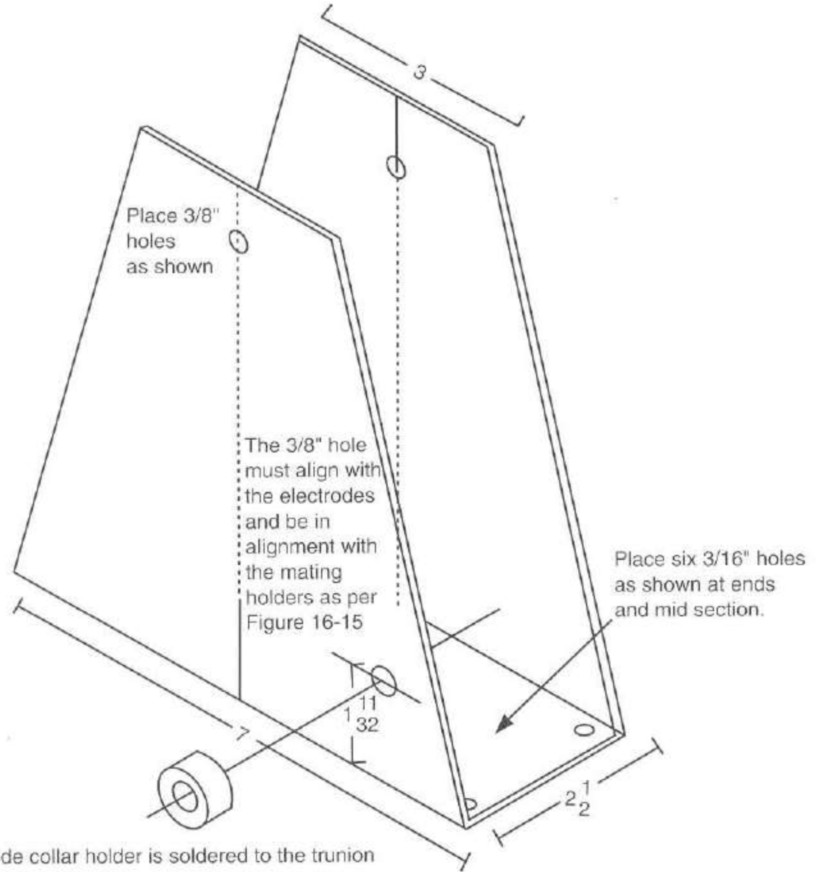
3
Place 3/8" holes as shown
The 3/8" hole must align with the electrodes and be in alignment with the mating holders as per Figure 16-15
Place six 3/16" holes as shown at ends and mid section.
1 11/32
7
2 1/2

The electrode collar holder is soldered to the trunion bracket using a propane torch and carefully centering with clearance hole.

Figure 16-11 *Fabrication of the trunion bracket*

heavy leads for the discharge path. You may want to sleeve some plastic tubing over these leads, as they are rated only for 600 volts.

Note: The trunion bracket is the common grounding point for the system. It is very important that the earth ground green lead of the power cord be firmly attached at this point to ensure system operating safety.

24. Verify and use the diagram shown in Figure 16-18 as a wiring aid and reconfirm grounding and general circuitry and integrity.

Testing and Operation of the Cannon

25. Locate a test site with a backstop capable of stopping the projectile. Obtain a 5,000 volt DC reading meter to set M1. It is assumed that the circuitry is correctly assembled and the basic electrical pretest as described in steps 10 through 14 was successively performed. Preset the main gap to 1/16 inch and the trigger gap to 1/8 inch.

26. Preset all switches shown back in Figure 16-8 to the off position. Familiarize yourself with these controls

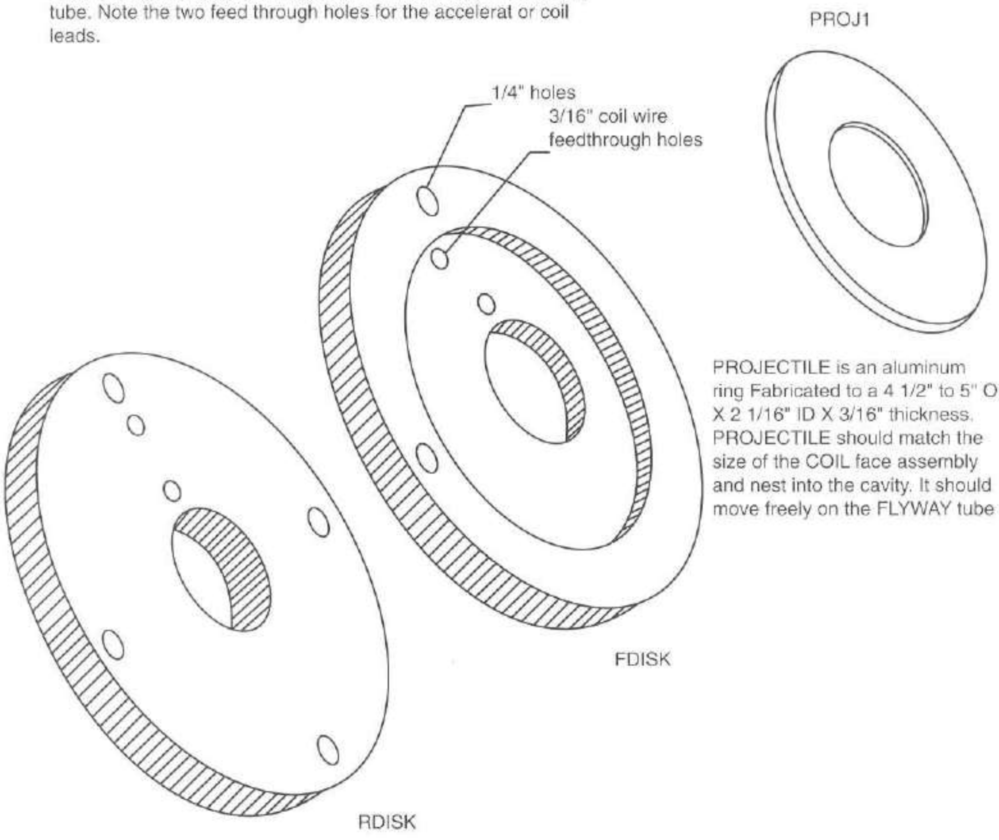

Suggested material is 3/4" plywood or machined PVC

The accelerator coil construction is very important to the overall performance and efficiency of the system. These pieces as shown are fabricated from a high grade birch plywood. You may use polycarbonate for strength or PVC for cost. The objective is to provide a cavity for the flat pan cake shaped windings.
The parts shown are 7 inch overall diameter of 3/4" thickness. Both pieces have all holes as shown and must be accurately aligned with one another. The front piece has the 5/16" deep cavity milled out for nesting of the coil windings. Leave a 1" shoulder allowing a 5 inch winding diameter. Both center holes should allow a reasonably tight sliding action for the 1 7/8" PVC fly way tube. Note the two feed through holes for the accelerat or coil leads.

PROJECTILE is an aluminum ring Fabricated to a 4 1/2" to 5" OD X 2 1/16" ID X 3/16" thickness. PROJECTILE should match the size of the COIL face assembly and nest into the cavity. It should move freely on the FLYWAY tube

Figure 16-12 *Fabrication of the accelerator coil holder*

Danger: The next steps involve exposure to dangerous high voltages. Failure to take any of the following described steps will require a complete shutdown and power removal, along with a shorting of C30 before performing any trouble shooting.

27. Connect the test meter from chassis ground to the output end of resistor R30. Plug the unit into 115 vac and turn on S6. Turn on the key switch S2 and note LED1 lighting. Place the projectile over the accelerator coil.
28. Depress pushbutton switch S3 and note a spark occurring in both gaps.
29. Verify S1 safety switch is off. Depress pushbutton switch S4 and note the relay RE1 energizing and LED2 lighting. Depress pushbutton switch S5 and note RE1 de-energizing and LED2 turning off. Verify these on and off functions of S4 and S5 several times before the next steps.

    S5 is important as it must stop the charging action in the event that you choose to quickly abort the cycle.
30. Turn on toggle switch S1. Restart the system by again depressing R4, noting a voltage reading on both the panel meter M1 and the external test meter. This voltage will build up to a level as set by the charge-voltage level control R14 and then shut down, only to restart the charge cycle when the voltage charge level

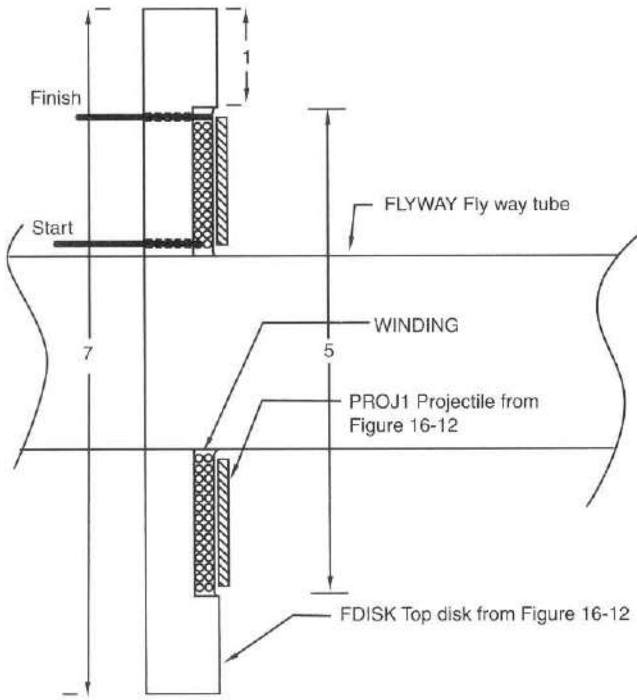
1
Finish
Start
FLYWAY Fly way tube
WINDING
7
5
PROJ1 Projectile from Figure 16-12
FDISK Top disk from Figure 16-12

1. Fabricate the FLYWAY from a 18" piece of 2" PVC tubing
2. Wind a coil of 30 turns of #12 in fifteen layers of 2 turns each. Finished coil must be in a tight pancake configuration and have 10 inch connection leads. *You may have to make up a temporary bobbin jig for this step.* Hold together with tape etc.
3. Place COIL section into cavity routing the leads as shown. COIL section should nest into cavity and be positioned flat.
4. Seal all points that can leak as epoxy must be carefully poured over the COIL forming a 1/16" layer on top.

Figure 16-13 *Side view showing winding*

Suggested material is ¾" plywood or machined PVC

You will note that the front disk housing the accelerator coil is allowed to move over the flyway tube and against the springs to absorb some of the recoil. The rear disk abuts to the shoulder of the breech tube that is non movable along with the flyway tube.

This feature in important if you are intending to repetitively fire high energy shots..

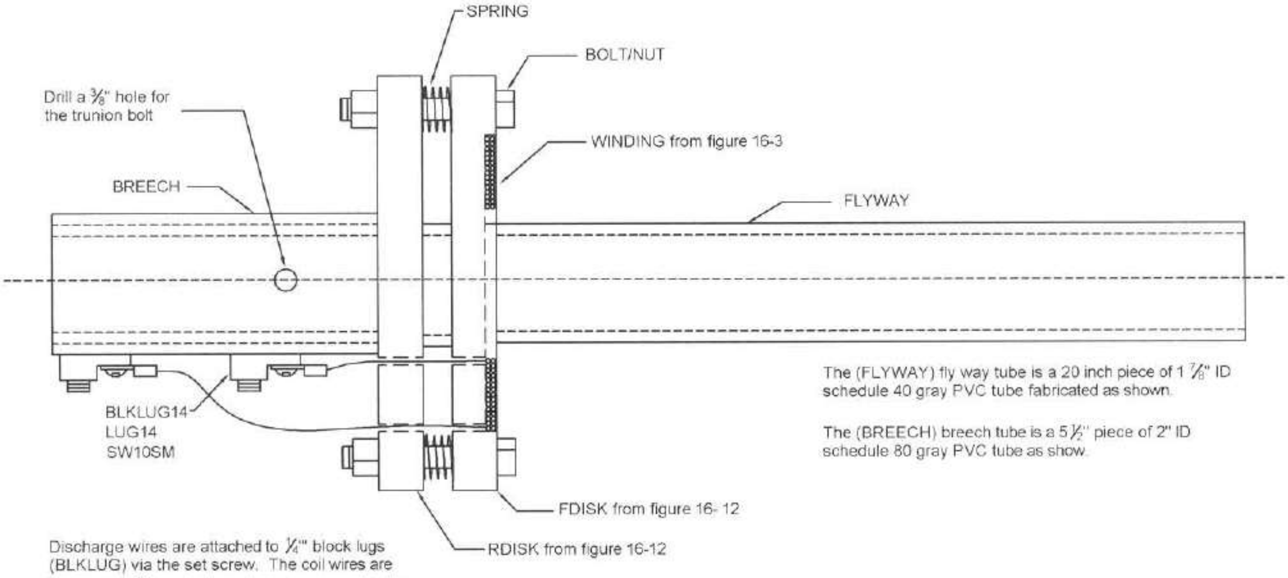
SPRING
BOLT/NUT
Drill a ⅜" hole for the trunion bolt
WINDING from figure 16-3
BREECH
FLYWAY
BLKLUG14
LUG14
SW10SM
FDISK from figure 16- 12
RDISK from figure 16-12

The (FLYWAY) fly way tube is a 20 inch piece of 1 ⅞" ID schedule 40 gray PVC tube fabricated as shown.

The (BREECH) breech tube is a 5½" piece of 2" ID schedule 80 gray PVC tube as show.

Discharge wires are attached to ¼" block lugs (BLKLUG) via the set screw. The coil wires are soldered to ¼" spade lugs that are secured to the block lugs via #10 x ¾" sheet metal screws, screwed into both the FLYWAY and BREECH tube.

The bare leads from the coil should be individually sleeved into plastic tubing to prevent catastrophic contact

Figure 16-14 *Assembly of the breech and flyway*

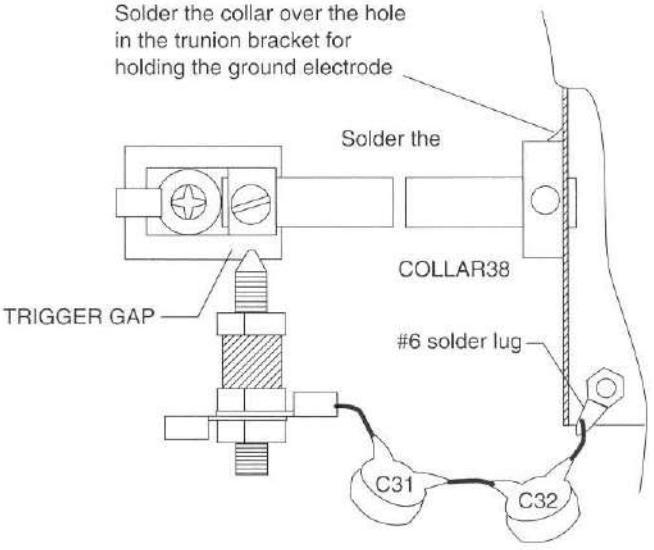
Solder the collar over the hole
in the trunion bracket for
holding the ground electrode
Solder the
COLLAR38
TRIGGER GAP
#6 solder lug
C31
C32

Solder gap switch assembly top view

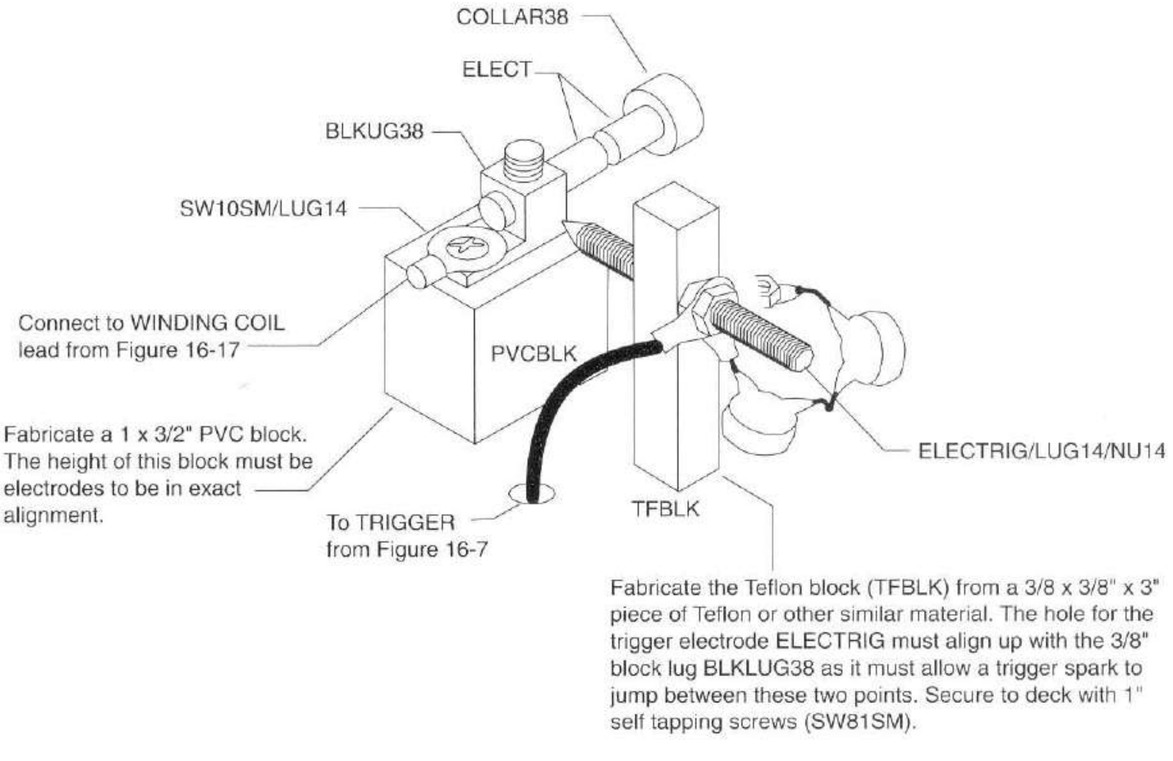
COLLAR38
ELECT
BLKUG38
SW10SM/LUG14
Connect to WINDING COIL
lead from Figure 16-17
PVCBLK
Fabricate a 1 x 3/2" PVC block.
The height of this block must be
electrodes to be in exact
alignment.
To TRIGGER
from Figure 16-7
TFBLK
ELECTRIG/LUG14/NU14
Fabricate the Teflon block (TFBLK) from a 3/8 x 3/8" x 3"
piece of Teflon or other similar material. The hole for the
trigger electrode ELECTRIG must align up with the 3/8"
block lug BLKLUG38 as it must allow a trigger spark to
jump between these two points. Secure to deck with 1"
self tapping screws (SW81SM).

Spark gap switch is o view

Figure 16-15 *Details of the spark and trigger switches*

falls below the set level. Obviously there is an amount of hysteresis that prevents off-and-on chatter. With R14 set to full counterclockwise (CCW), this voltage should reset at less than 1,000 volts. The buzzer BUZ1 should be producing an annoying sound, thus alerting you to a charge presence on the capacitors.

31. Slightly turn R14 clockwise and allow it to cycle, repeating up to 1,500 volts as read on the external test meter. Depress the trigger

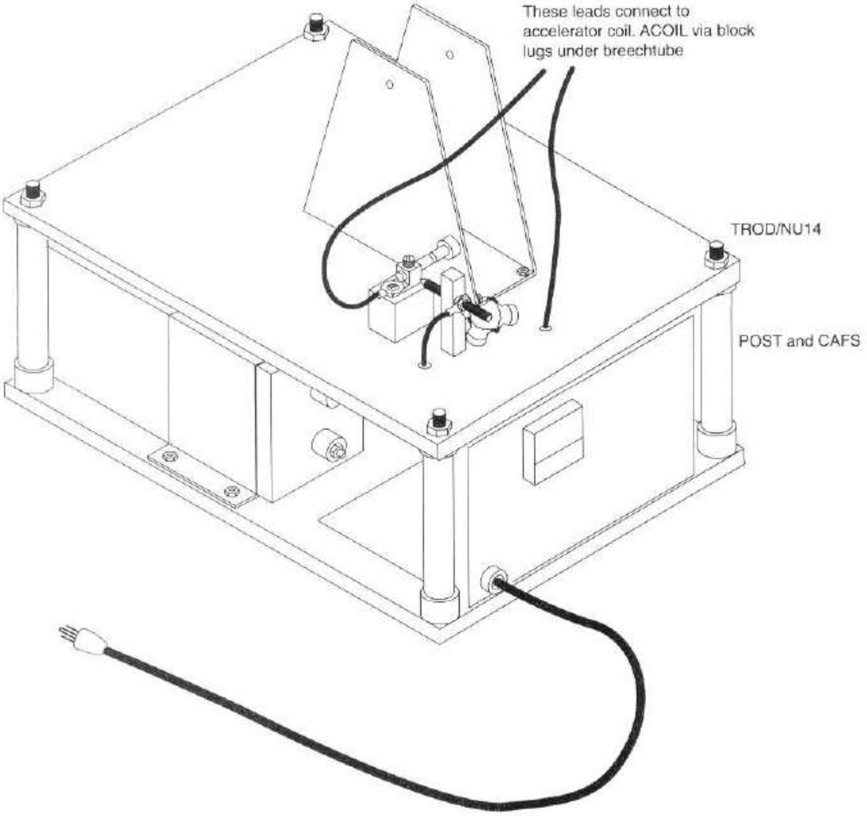
These leads connect to accelerator coil. ACOIL via block lugs under breechtube
TROD/NU14
POST and CAFS

Figure 16-16 *Final assembly isometric view*

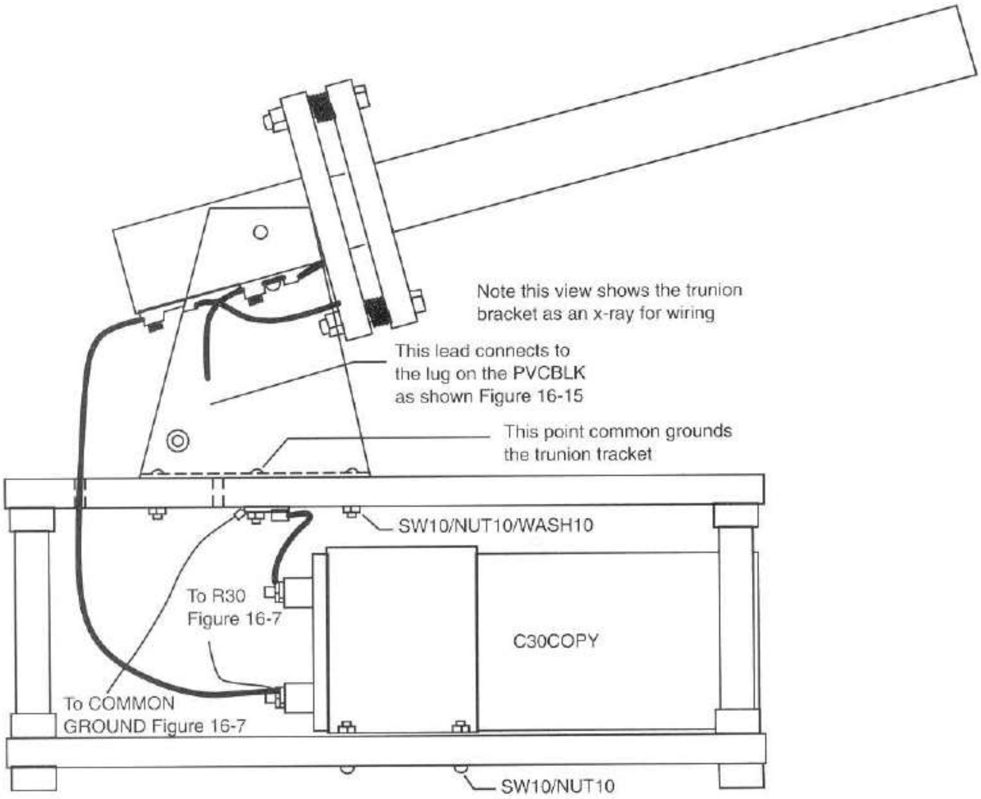
Note this view shows the trunion bracket as an x-ray for wiring
This lead connects to the lug on the PVCBLK as shown Figure 16-15
This point common grounds the trunion tracket
SW10/NUT10/WASH10
To R30 Figure 16-7
C30COPY
To COMMON GROUND Figure 16-7
SW10/NUT10

Figure 16-17 *Final assembly side view*

fire pushbutton S3 and note the projectile shooting into the backstop. Immediately press the stop switch S5 unless you want it to charge up for the next shot.

Note: The system will automatically recharge to the 1,500-volt present level unless the stop switch is activated.

---

32. Repeat the cycle, increasing the charge voltage up to 2,500 volts, and then adjust the trimpot R16 on the meter M1 to agree with the test meter.
33. Repeat this step allowing the charge voltage to go up to 5,000 volts, noting the impact of the projectile.
34. Verify the unit is not breaking apart as a result of recoil. Rethink the recoil springs if a problem exists.

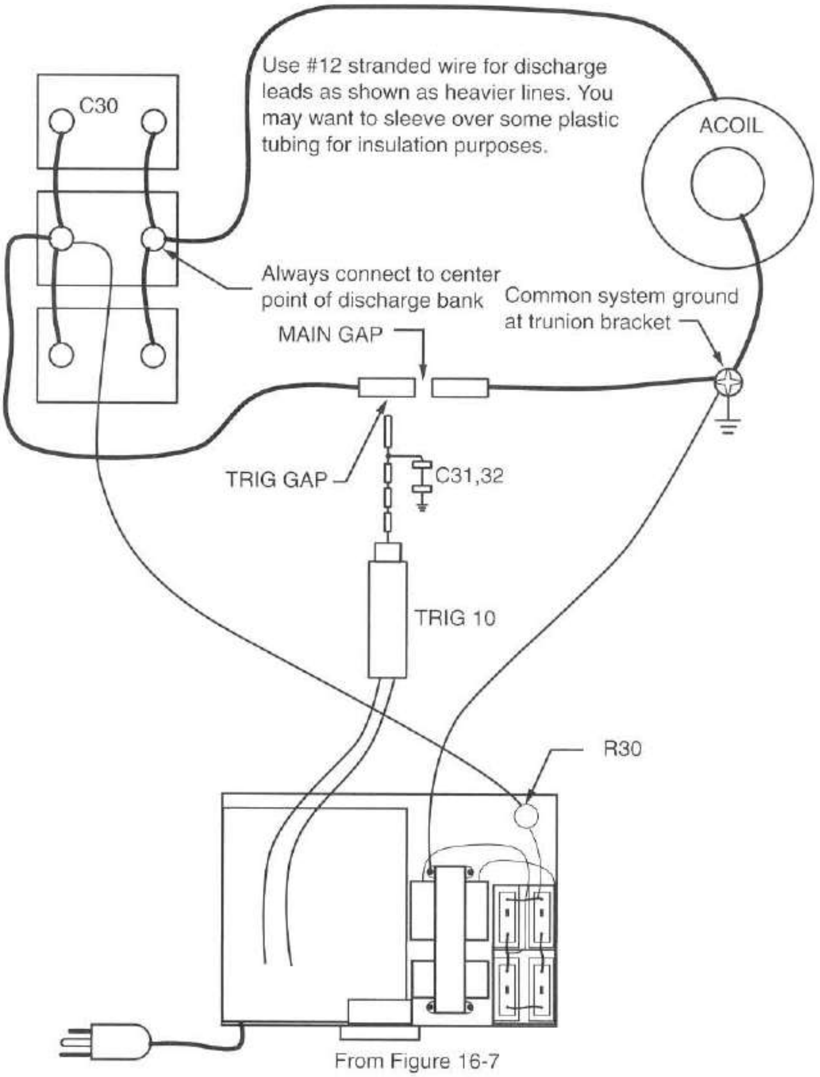

C30
Use #12 stranded wire for discharge leads as shown as heavier lines. You may want to sleeve over some plastic tubing for insulation purposes.
ACOIL
Always connect to center point of discharge bank
Common system ground at trunion bracket
MAIN GAP
TRIG GAP
C31,32
TRIG 10
R30
From Figure 16-7

Figure 16-18 *Overall system wiring diagram*

Table 16.1 Magnetic Cannon Part List

Ref. #	*Description*	*DB Part #*
R1, 4, 9, 19	Four 470, ¼-watt resistors (yel-pur-br)	
R2	1K, ¼-watt resistor (br-blk-red)	
R3, 6, 7, 8	Four 10 kilohm, ¼-watt resistors (br-blk-or)	
R5	220, ¼-watt resistor (red-red-br)	
R10, 11	Two 10M, ¼-watt resistors (br-blk-bl)	
R12	2.2K, ¼-watt resistor (red-red-red)	
R13, 15	Fourteen 1.2M, ¼-watt resistors (br-red-gr) (see schematic in Figure 16-2)	
R14/S6	100K potentiometer, 17 mm with low-volt switch	
R16	2K trimpot	
R20	1M, ¼-watt resistor (br-blk-gr)	
R21	10-ohm, ¼-watt resistors (br-blk-blk)	
R18	12-ohm, 3-watt resistor	
R30	16K, 50-watt power resistor	DB# 16K50W
C1, 4, 5	Three 1,000 mfd, 25-volt vertical electrolytic capacitors	
C2, 3	Two .1 mfd, 50-volt plastic capacitors	
*C30A	Three 32 mfd, 4,500-volt oil-filled cans in parallel for 96 mfd/4.5 kv	DB# 32m/4500
*C30B	Four 24 mfd, 5,200-volt oil-filled cans in parallel for 96 mfd/5 kv	DB# 24m/5200
I1	741 op-amp DIP IC	
Q1, 2, 3	Three PN2222 GP NPN transistors	
Q4	IRF520 Mosfet	
Q5	MJE3055 NPN power transistor	
D1-4	Four 1N4001 50-volt 1-amp diodes	
D5	1N4007 1,000-volt, 1-amp diode	
D31-34	Four 8-kilovolt, 100 ma standard recovery diodes	DB# H407
D35-37	Three 30 kilovolt, 10 ma diodes	DB# VG30
Z1, 2, 3	Three 1N4733 6-volt zener diodes	
Z4	1N4745 15-volt zener diode	
LED1	Red LED indicator	
LED2	Green LED indicator	
S1	SPST toggle switch	
S2	Key switch	
S3, 4, 5	Three pushbutton NO switches	
RE1	12-volt, 5-amp relay	
BUZ1	12-volt buzzer	DB# BUZROUND
Q1	100-volt IRF540 MOSFET	
T1	12-volt, 100 ma transformer	DB# 12/.1
T2	6,500-volt, 20 ma current-limited transformer	DB# 6kv/20
*L10	Accelerator coil assembly, as shown in Figures 16-12 and 16-13	DB# ACCOIL
*GAP1	Main gap assembly, as shown in Figure 16-15	DB# GAPMAIN
*GAP2	Trigger gap assembly, as shown in Figure 16-15	DB# GAPTRIG
CHASSIS1	Metal chassis fabricated as shown in Figure 16-6	
PLATE1	6 × 7½ × 1/16 inch plastic insulating plate	
PLATE2	2¼ × 7½ × 1/16 inch plastic insulating plate	
TAPE1	12 inches of 1-inch wide, thin sticky tape	
PLTUBE	18 inches of ¼-inch ID vinyl flexible tubing for R13, R15, and trigger diodes	
CO1	6 feet of 3-wire #18 power cord	
HVWIRE	12 inches 20 kv silicon high-voltage wire	DB# WIRE20KV
BUSSWIRE	24 inches #20 bare busswire	

WR20	10 feet #20 vinyl stranded hook-up wire	
WN1	Medium wire nut	
BU2, 3	Two bushings for LEDS	
BU1	Power cord clamp bushing	
SW6	Six 6-32 × 3/4-inch screws	
SW1	6-32 × 1/4-inch screw for mounting relay	
NU6	Six #6 hex nuts	
LUG6	#6 solder lug	
TROD	5-inch length of 1/4-20 threaded rod, 2 nuts, and plastic washers	
TRIG10	Assembled trigger pulse module	DB#TRIG10
TRIG1K	Kit of trigger pulse module	DB#TRIG1K

Parts List to Final System

DECKS	Two fabricated pieces of 7/8-inch birch plywood 13 × 17 inches
POSTS	Four fabricated pieces of 5 3/4 × 7/8 inch PVC tubing
PLATE1	Fabricate a piece of 1/16-inch plastic sheet 6 × 7 inches
PLATE2	Fabricate a piece of 1/16-inch plastic sheet 2 1/4 × 7 inches
BRKCAP1	Fabricate the capacitor bracket seen in Figure 16-10
BRKTRUN1	Fabricate the trunion bracket seen in Figure 16-11
RDISK	Fabricate the rear disk seen in Figure 16-12
FDISK	Fabricate the front disk seen in Figure 16-12
PROJ1	Fabricate the projectile seen in Figure 16-12
BLKELECT	Fabricate the electrode holding block seen in Figure 16-15
BLKTRIG	Fabricate the trigger electrode holding block seen in Figure 16-15
FLYWAY	Fabricate the flyway tube seen in Figure 16-14
BREECH	Fabricate the breech tube seen in Figure 16-14
CAPS	Twelve 7/8-inch PVC caps for posts and feet seen in Figure 16-1
SPRINGS	Four 1/2-inch diameter × 3/4-inch stiff compression springs
COLLAR	3/8-inch shaft collar
BLKLUG25	Three 1/4-inch block lugs for breech connections
BLKLUG38	3/8-inch block lug for electrode
LUG14	Five 1/4-inch wide solder lugs
ELECTRGAP	Two 3/8 × 2-inch tungsten or tool steel electrodes
ELECTRTRIG	1/4-20 × 3-inch threaded rod with point
TROD	Four 1/4-20 × 8-inch threaded rod
BOLT/NUT	Five 3 1/2 × 3/8-inch bolt and nut
NUT14	Eleven 1/4 × 20 nuts for trigger gap and posts
SW10M	Five #10 × 1/2-inch sheet metal screws
SW10	Ten #10-32 × 1.5-inch machine screw
NU10	Ten #10-32 nuts
WASH10	Twenty #10 wide shoulder washers
WIRE12	4 feet #12 stranded, vinyl-covered wire
WIRE12	30 feet #12 heavy-coated, magnetic wire for winding ACOIL

Chapter Seventeen

Vacuum Tube Tesla Coil Project

This advanced project, as shown in Figure 17-1, is similar to a coil made by John Weisner some years ago. The geometry of the secondary not being the most efficient is a good compromise for wireless energy transmission and also provides a reasonable output display. The project is intended for the experienced experimenter and researcher in the field of high-voltage and high-frequency circuitry. Not only is this project a dangerous electrical device, but it can also be a bothersome source of radio interference, especially in the standard AM broadcast band and should be used in a shielded enclosure or a Faraday cage if operated on a continuous basis. A gas discharge lamp such as a standard household fluorescent lamp will glow up to 10 feet away when this device is used. Annoying burns are experienced when contact is made with conductive objects up to similar distances.

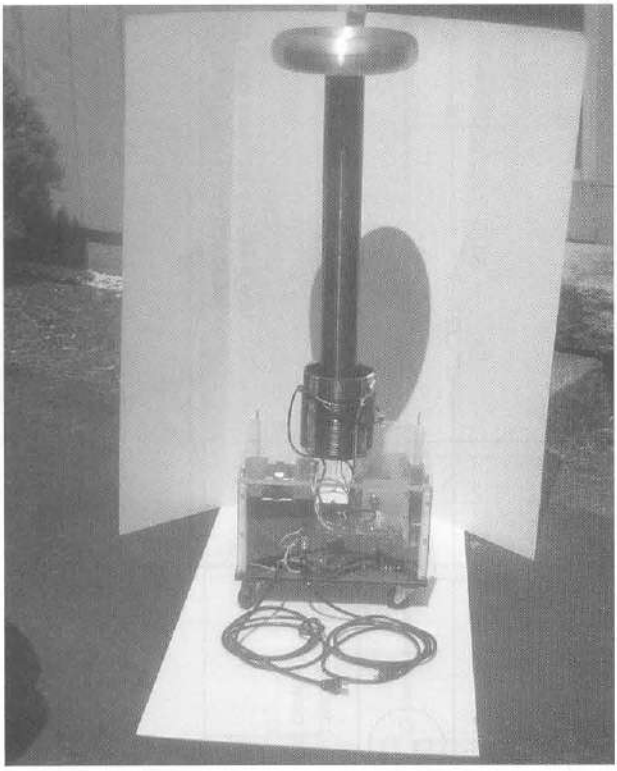

Figure 17-1 *The vacuum tube Tesla coil*

We also have built coils using 304TH, 450TH, and parallel 2500THs with a full 5,000 watts of plate dissipation. This coil produced spectacular flaming arcs up to 8 feet long with 8 kilowatts of input. A video showing the operation is on our web site. The coil shown here will easily produce 1 to 2 feet of flaming arcs.

This system can provide the experienced hobbyist with hours of fun and entertainment experimenting with terminals and settings to generate wireless energy as well as output arcs and sparks. This is an advanced-level and potentially expensive project requiring electronic skills and high-voltage experience. Expect to spend $300 to $750. All the parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 17-1.

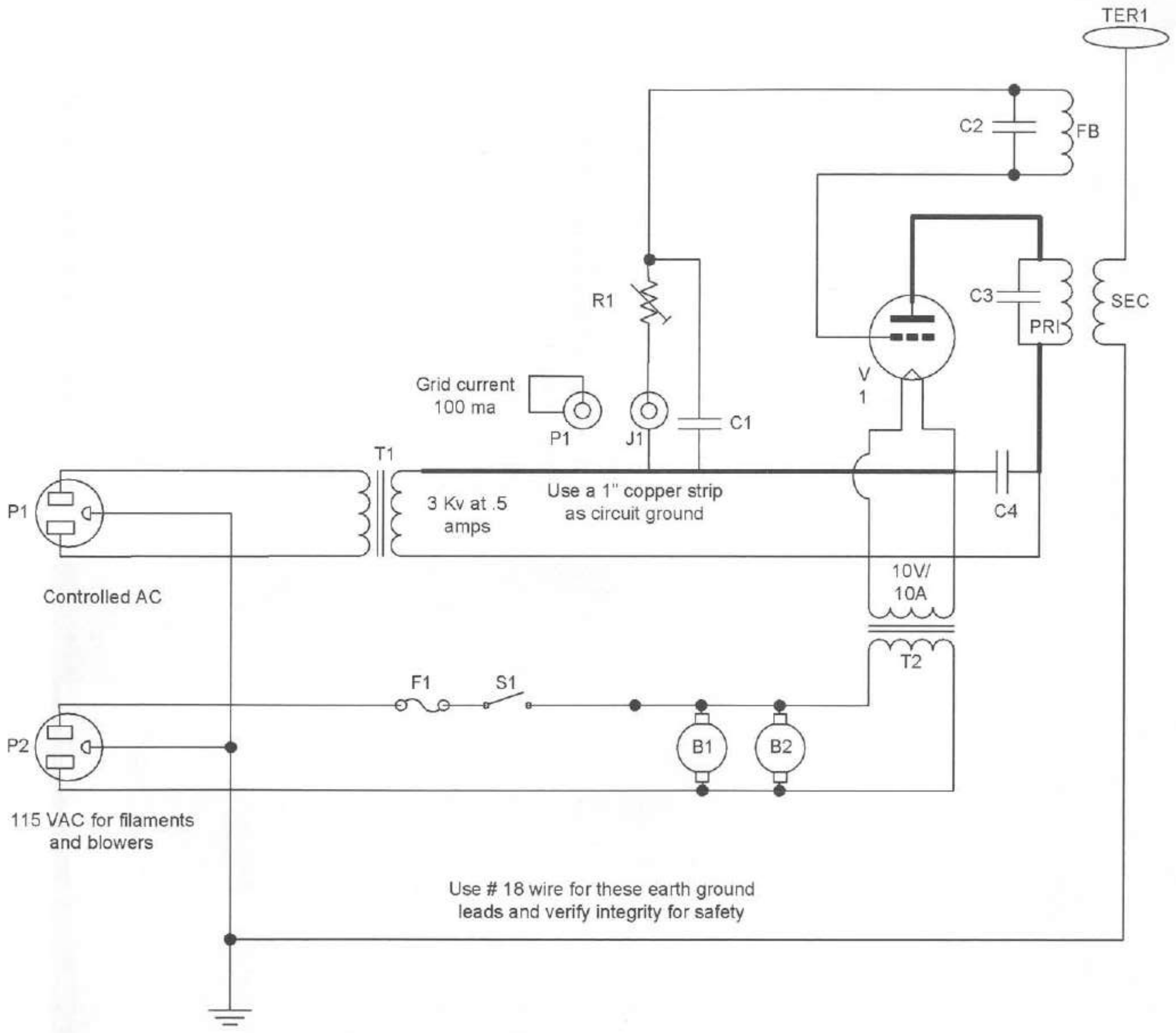
TER1
C2
FB
R1
C3
SEC
PRI
V
1
Grid current
100 ma
P1
J1
C1
T1
P1
3 Kv at .5
amps
Use a 1" copper strip
as circuit ground
C4
Controlled AC
10V/
10A
T2
F1
S1
P2
B1
B2
115 VAC for filaments
and blowers
Use # 18 wire for these earth ground
leads and verify integrity for safety

Feed back winding is 10 to 20 turns of #18 stranded vinyl wire. Open circuit inductance is approximately 75 micro-henries

Primary windings are 20 to 22 turns of #8 insulated wire. Open circuit inductance 55 micro-henries

Secondary winding is a 33" length of tightly wound #28 magnet wire for about 2000 turns. Free standing resonant frequency is approximately 200 kHz.

Resistor R1 is 5000 ohms at 100 watts

Figure 17-2 *Circuit schematic*

Basic Description

These spectacular display devices produce arcs and sparks quite unlike the undamped spark gap-driven Tesla coils. Operation does not require a noisy spark gap that produces copious amounts of *radio frequency interference* (RFI) but operates efficiently at the quarter-wave frequency of the secondary coil. Vacuum tubes, while large and requiring filament power, do offer robust and forgiving operation and are intended as a good starter approach to solid-state Tesla coils.

This approach is far less temperamental than the solid-state *metal-oxide-semiconductor field effect transistors* (MOSFET) or *insulated gate bipolar transistor* (IGBT) drivers. However, once these FET devices are debugged and tweaked, one can obtain some excellent results. Our lab FET coil generates over 4-foot sparks using only 3 kilowatts. The output frequency tracks within 5 percent of the resonant point using one of our proprietary circuits that will be on our site in the late spring of 2006 with the complete system scheduled for publishing in the third edition of this book series.

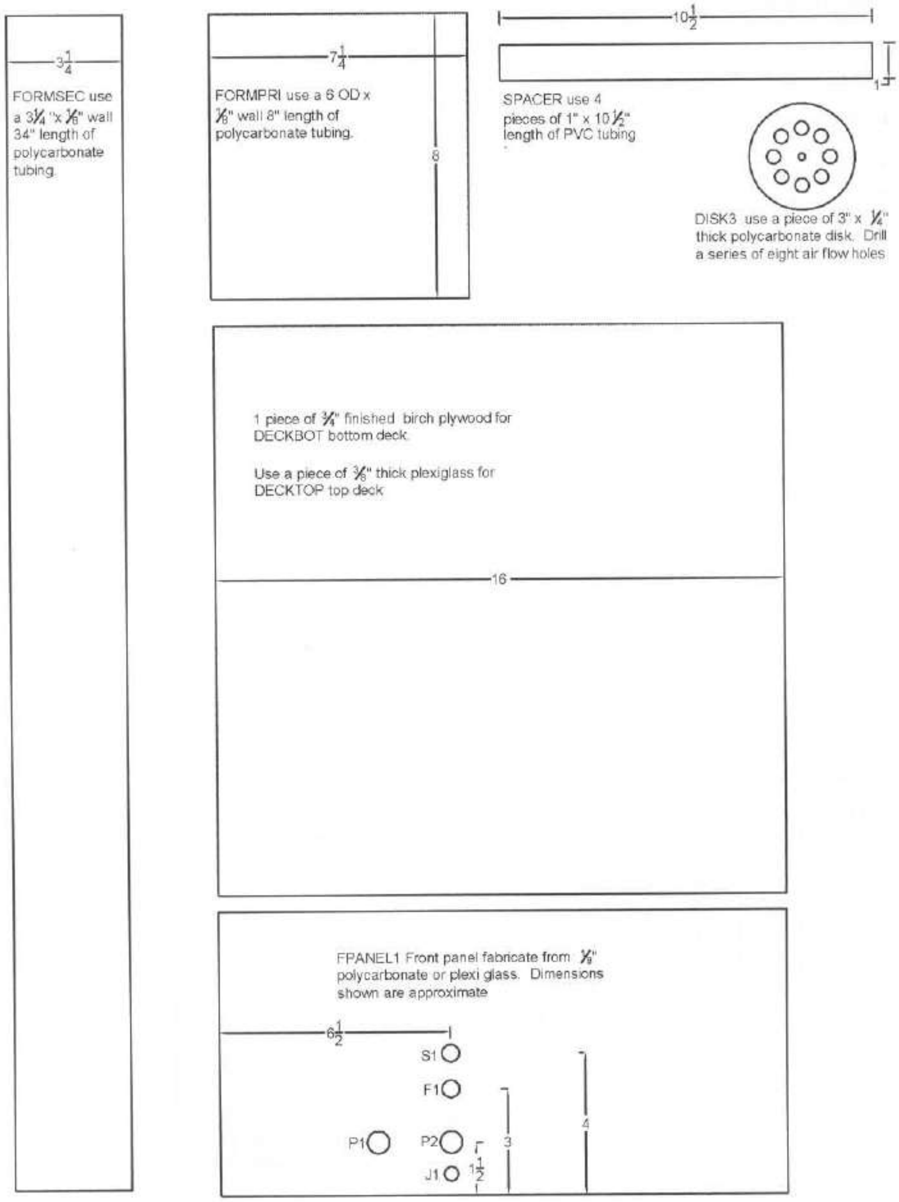
3¼
FORMSEC use a 3¼ "x ⅛" wall 34" length of polycarbonate tubing.
7¼
8
FORMPRI use a 6 OD x ⅛" wall 8" length of polycarbonate tubing.
10½
1
SPACER use 4 pieces of 1" x 10½" length of PVC tubing
DISK3 use a piece of 3" x ¼" thick polycarbonate disk. Drill a series of eight air flow holes
1 piece of ¾" finished birch plywood for DECKBOT bottom deck.
Use a piece of ⅜" thick plexiglass for DECKTOP top deck
16
FPANEL1 Front panel fabricate from ⅛" polycarbonate or plexi glass. Dimensions shown are approximate
6½
S1
F1
P1
P2
J1
1½
3
4

Figure 17-3 *Creating the major parts*

The circuit as shown on the schematic in Figure 17-2 of the vacuum tube device is nothing more than a medium-powered Hartley radio frequency oscillator tuned to the resonant frequency of the secondary coil. The circuit uses a medium-powered, readily available 833A triode transmitting tube that inherently has a high grid to plate capacitance necessary for self-oscillation. The plate load is a resonant transformer with both the primary and secondary tuned to the approximate identical frequencies. The output potential is now a function of the secondary-to-primary coil inductances times the peak primary plate voltage. This system generates voltages at frequencies sufficient to cause air breakdown, thus producing the sparks and arc jumping into open air.

The output of the oscillator is relatively closely coupled to the secondary coil designed for high Q performance and self-resonant to the quarter-wave of the oscillator frequency. The voltage distribution is now that of a quarter section with a current node at the base and voltage node at the top. The input power to the coil is raw, unrectified AC at 3,000 volts *root mean square* (rms) with a current of .5 amps. The peak voltage is over 4,000 volts and is supplied by a conventional plate transformer being fed by a voltage-adjustable variac, which allows you to adjust the output voltage of the system. A pulse signal is also shown that controls the grid of the tube, allowing a wide range of spark texture variation by changing the duty cycle and frequency.

Assembly Steps

1. Fabricate the secondary coil form (FORMSEC) from a $3\frac{1}{4} \times \frac{1}{8}$-inch wall polycarbonate tubing, as shown in Figure 17-3. It is not suggested that you use PVC tubing due to moisture retention and breakdown.

   Fabricate the primary coil form (FORMPR1)as shown in Figure 17-3 from a length of $7\frac{1}{2} \times \frac{1}{4}$-inch wall plastic tubing. Polycarbonate tubing is preferred.

   Fabricate two 3-inch diameter discs (DISK3) as shown in Figure 17-3 from $\frac{1}{4}$-inch polycarbonate. These pieces are for the ends of the secondary coil and are secured via three #6 × $\frac{3}{8}$-inch brass wood screws. Drill air passage holes as shown for the blower (B2).

   Fabricate four 10.5-inch spacers (SPACER) as shown in Figure 17-3 from lengths of 1-inch OD PVC tubing.

   Fabricate the 16 × 16-inch square bottom deck (DECKBOT) as shown in Figure 17-3 from a piece of $\frac{3}{4}$-inch finished plywood and paint it black.

   Fabricate the 16 × 16-inch deck top (DECKTOP) as shown in Figure 17-3 from a $\frac{3}{8}$-inch-thick Plexiglas or you may use plywood.

2. Fabricate the parts as shown in Figure 17-4. Note to use $\frac{1}{16}$-inch G10 or another high-temperature plastic for the actual tube

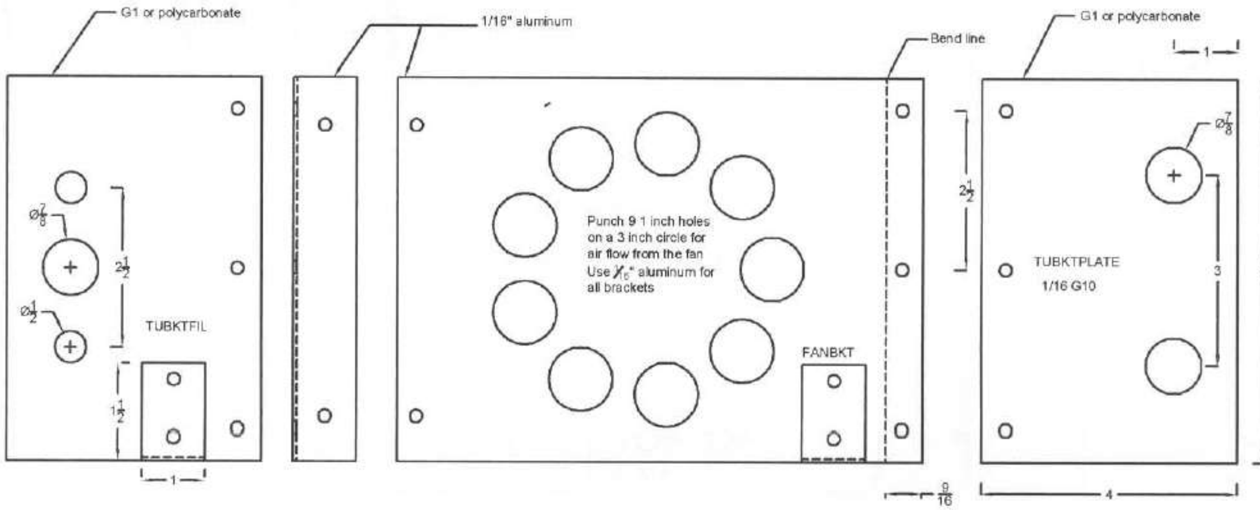

Figure 17-4 *Assembly of the tube and mounting bracket*

holders (TUBKTFIL and TUBKTPLATE). Use $^{1}/_{16}$-inch aluminum for the other parts. Note the mating holes for assembling them together, as shown in Figure 17-4.

3. Assemble the fan and tube-mounting bracket from the parts in Figure 17-5. Use thirteen 6-32 × $^{1}/_{2}$-inch screws and nuts. Foot brackets secure the assembly to the wooden deck via four #8 × $^{1}/_{2}$-inch wood screws. TUBKTPLATE must be disassembled to insert the tube V1 in place. The blower fan B1 is centered to maximize air flow across the plate section of V1. It is mounted using four 8-32 × $^{1}/_{2}$-inch screws and nuts. The entire assembly

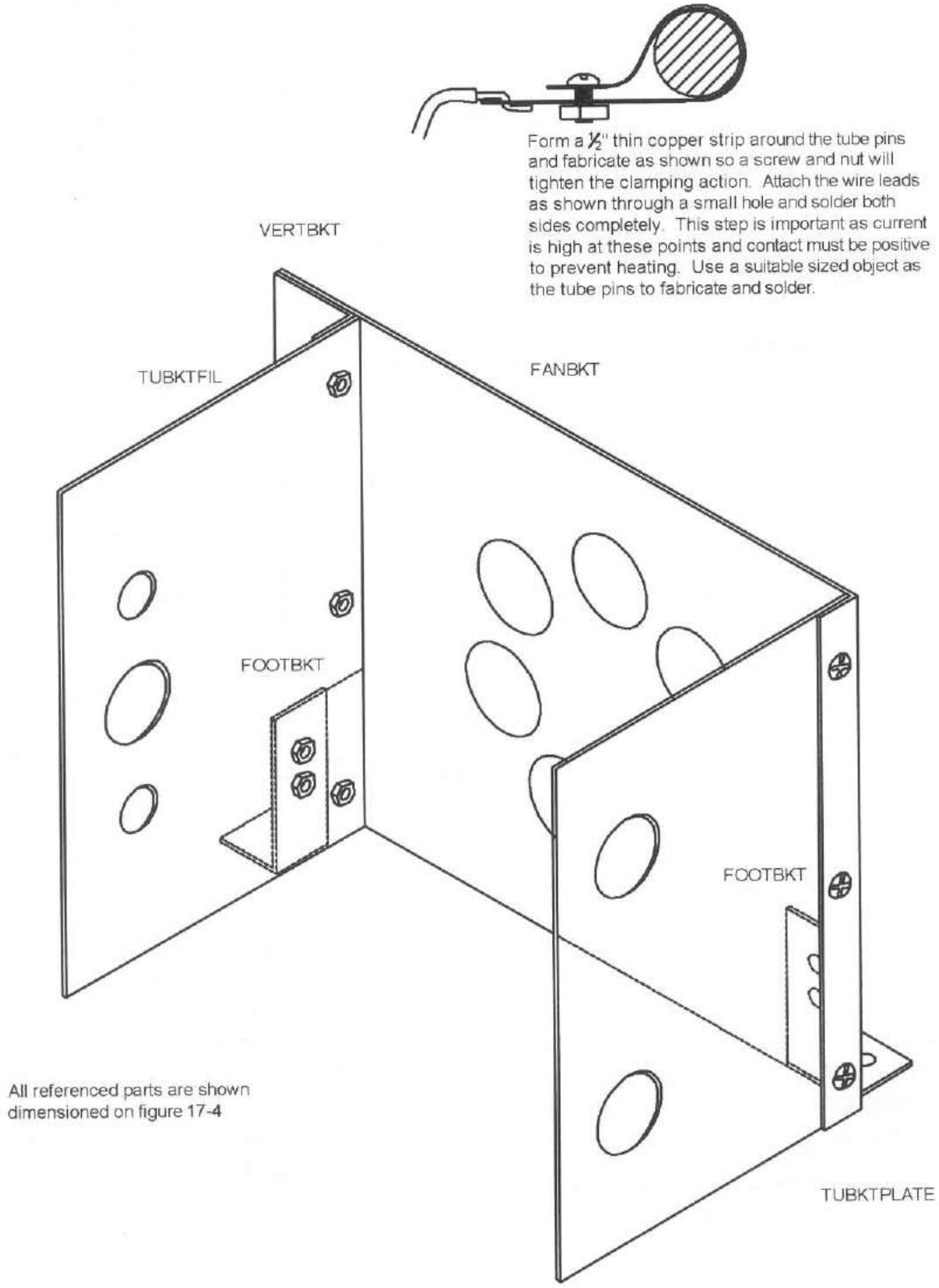

Form a ½" thin copper strip around the tube pins and fabricate as shown so a screw and nut will tighten the clamping action. Attach the wire leads as shown through a small hole and solder both sides completely. This step is important as current is high at these points and contact must be positive to prevent heating. Use a suitable sized object as the tube pins to fabricate and solder.

All referenced parts are shown dimensioned on figure 17-4

Figure 17-5 *Assembly of the tube-holding bracket*

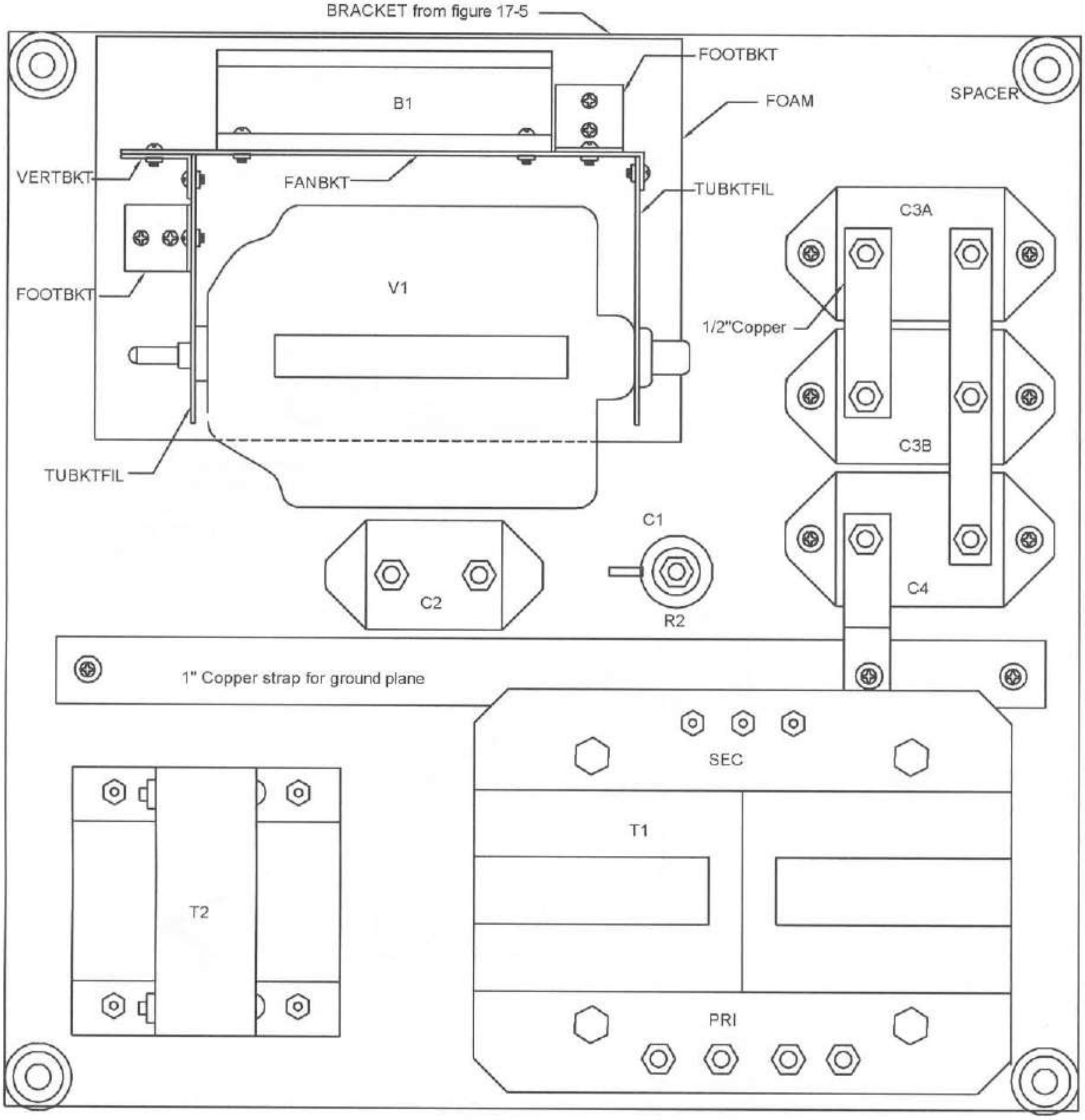
BRACKET from figure 17-5
FOOTBKT
B1
FOAM
SPACER
VERTBKT
FANBKT
TUBKTFIL
C3A
FOOTBKT
V1
1/2"Copper
C3B
TUBKTFIL
C1
C2
R2
C4
1" Copper strap for ground plane
SEC
T1
T2
PRI

Figure 17-6 *Bottom deck assembly*

must set on a 1/4-inch foam rubber pad to avoid any vibration. Fabricate the tube pin connections from 1/2-inch strips of thin copper in the form of clamping connectors. Use screws to tighten the clamps securely around the tube pins. Crimp the connecting wires in place and carefully solder the connecting leads completely. This is important as the tube pins must pass a high current.

4. Lay out the bottom deck DECKBOT and trial-fit the components, as shown in Figure 17-6. (You may want to consider castors at the corners for system mobility.) Cut a 14-inch length of a 1-inch-wide copper strip for the main ground plane. Secure it via two wood screws at the end points. Secure the plate transformer (T1) using four 1/4-20 × 1 1/2-inch bolts with flat washers on the bottom side of the wooden deck. Mount the three mica capacitors (C3A, C3B, and C4) using 6-32 × 1 1/2-inch screws, nuts, and washers. Cut 1/2-inch lengths of copper straps and attach them as shown. Note the grounding piece on C4 must be shaped and formed to have a flat surface abutting the ground plane strap and fitting as shown.

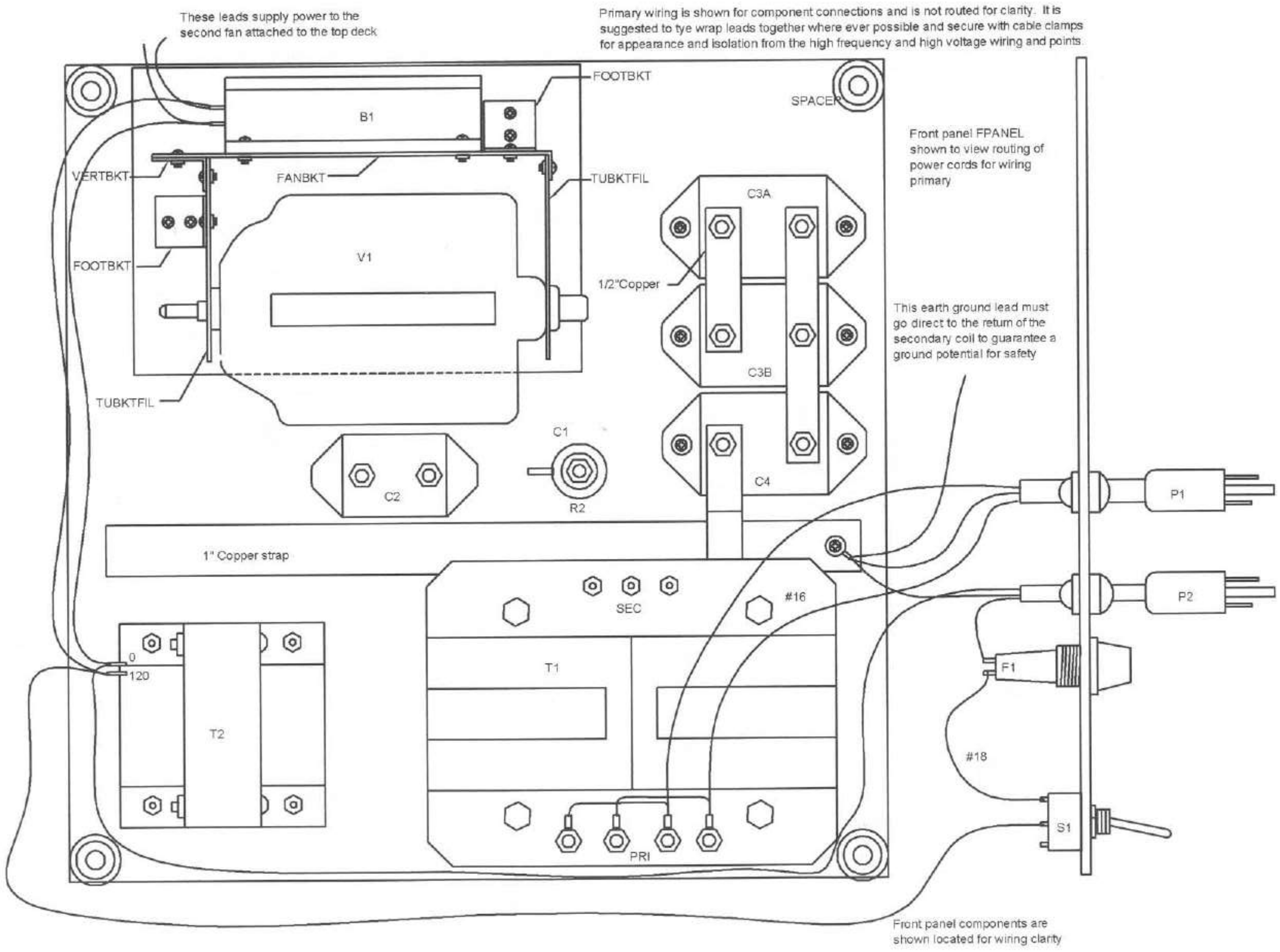

These leads supply power to the second fan attached to the top deck

Primary wiring is shown for component connections and is not routed for clarity. It is suggested to tye wrap leads together where ever possible and secure with cable clamps for appearance and isolation from the high frequency and high voltage wiring and points.

Front panel FPANEL shown to view routing of power cords for wiring primary

This earth ground lead must go direct to the return of the secondary coil to guarantee a ground potential for safety

Front panel components are shown located for wiring clarity

FOOTBKT
SPACER
B1
VERTBKT
FANBKT
TUBKTFIL
FOOTBKT
V1
TUBKTFIL
C3A
1/2"Copper
C3B
C4
C1
C2
R2
1" Copper strap
#16
SEC
P1
P2
F1
0
120
T2
T1
#18
S1
PRI

Figure 17-7 *Bottom deck primary wiring*

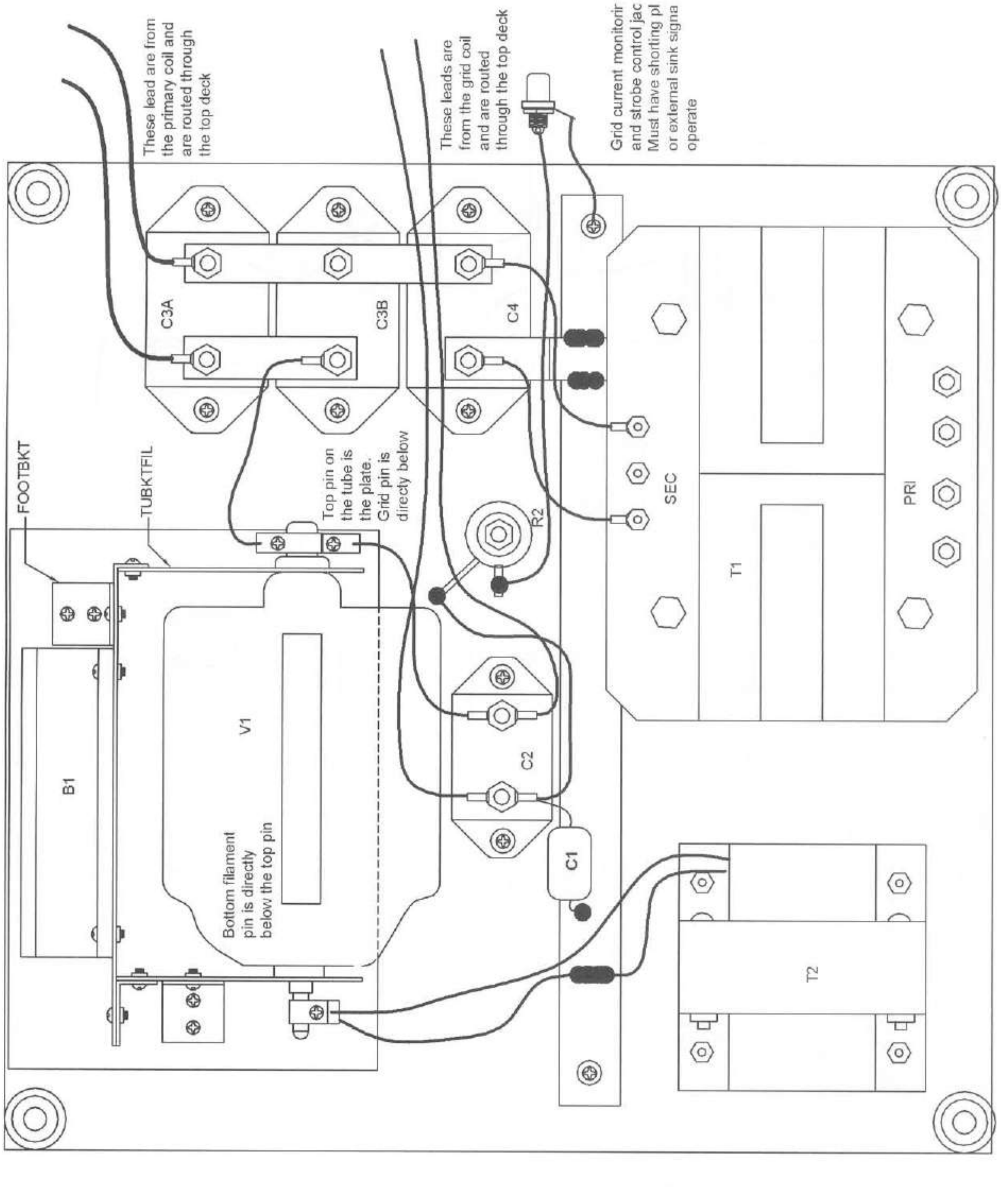
These lead are from the primary coil and are routed through the top deck
These leads are from the grid coil and are routed through the top deck
Grid current monitorir and strobe control jac Must have shorting pl or external sink signa operate
C3A
C3B
C4
FOOTBKT
TUBKTFIL
Top pin on the tube is the plate. Grid pin is directly below
R2
SEC
PRI
T1
B1
V1
C2
C1
T2
Bottom filament pin is directly below the top pin

Figure 17-8 *Bottom deck high-voltage wiring*

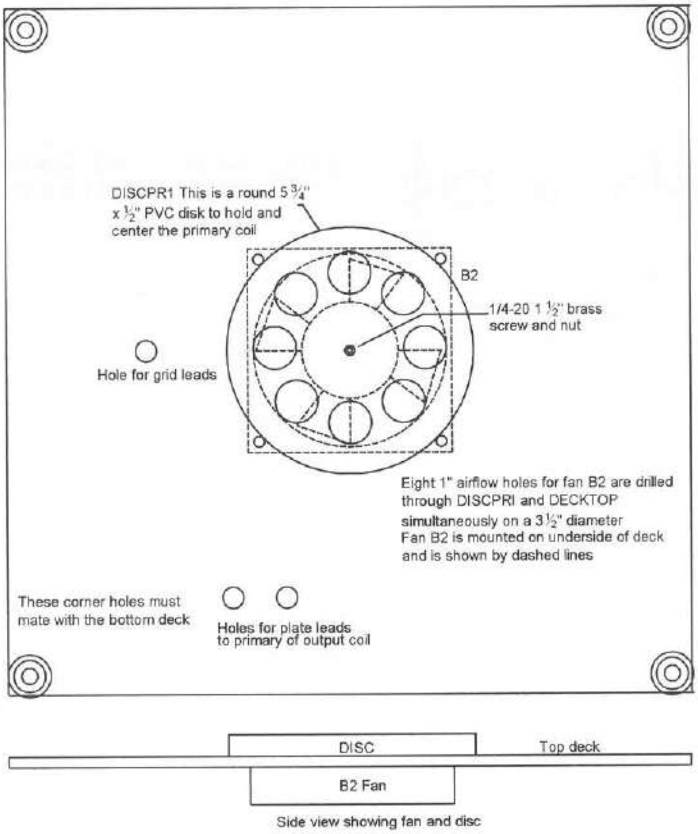
DISCPR1 This is a round 5 3/4" x 1/2" PVC disk to hold and center the primary coil
B2
1/4-20 1 1/2" brass screw and nut
Hole for grid leads
Eight 1" airflow holes for fan B2 are drilled through DISCPRI and DECKTOP simultaneously on a 3 1/2" diameter
Fan B2 is mounted on underside of deck and is shown by dashed lines
These corner holes must mate with the bottom deck
Holes for plate leads to primary of output coil
DISC
Top deck
B2 Fan
Side view showing fan and disc

Figure 17-9 *Top deck fabrication and assembly*

Do not tighten the nuts at this time. Mount mica capacitor C2 using similar hardware and grid resistor R2 vertically using a long 1/4-inch bolt or a threaded rod. Do not overtighten. Secure the filament transformer T2 using four 8-32 × 1 1/2-inch screws and nuts. Secure the tube and fan bracket via four #8 × 1/2-inch wood screws through the two foot brackets FOOBKT. Also note the foam rubber mount FOAM. Drill holes for the 1/4-inch −20 threaded rods for the deck spacer.

5. Thread the power cords (P1, P2) through the clamp bushings (BU1, BU2) as shown in Figure 17-7 on the front-panel section (FPANEL). The actual location on the panel section is shown in Figure 17-3. Figure 17-7 is shown to clarify the wiring. Power cord P1 must be a heavy-duty, #16, three-wire cord, with P2 being a #18, three-wire cord. Connect the two green ground leads to a crimp lug along with a 24-inch piece of #18 stranded wire that eventually connects to the center ground lug of the secondary coil. Solder this lug and verify the integrity as this is mandatory for electrical safety. Proceed to wire as shown using #18 stranded wires. Note that the two leads eventually connect to the second blower (B2) mounted on the top deck. Use crimp lugs and solder wherever shown. Route the wires, tie wrap, and clamp into place for appearance and isolation from other components.

6. Wire the filament clips to the 10-volt output leads of T2 as shown in Figure 17-8. These clips are shown in Figure 17-5. You will note that insulation is removed from one of these leads, and the uncovered section is soldered to the 1-inch copper strip. Solder in the bypass resistor (C1) to the copper strip and to a lug with the other leads to R2. High-voltage wire should be used from transformer T2 to capacitor C4. Proceed to complete the wiring using crimp lugs and soldering. Two pairs of wires are shown from the primary and grid coil. These are not actually connected at this time as they are part of the top deck assembly (DECKTOP). Jack (J1) is attached to the front panel shown located in Figure 17-3.

7. Fabricate the DECKTOP, as shown in Figure 17-9. This piece was originally shown as being

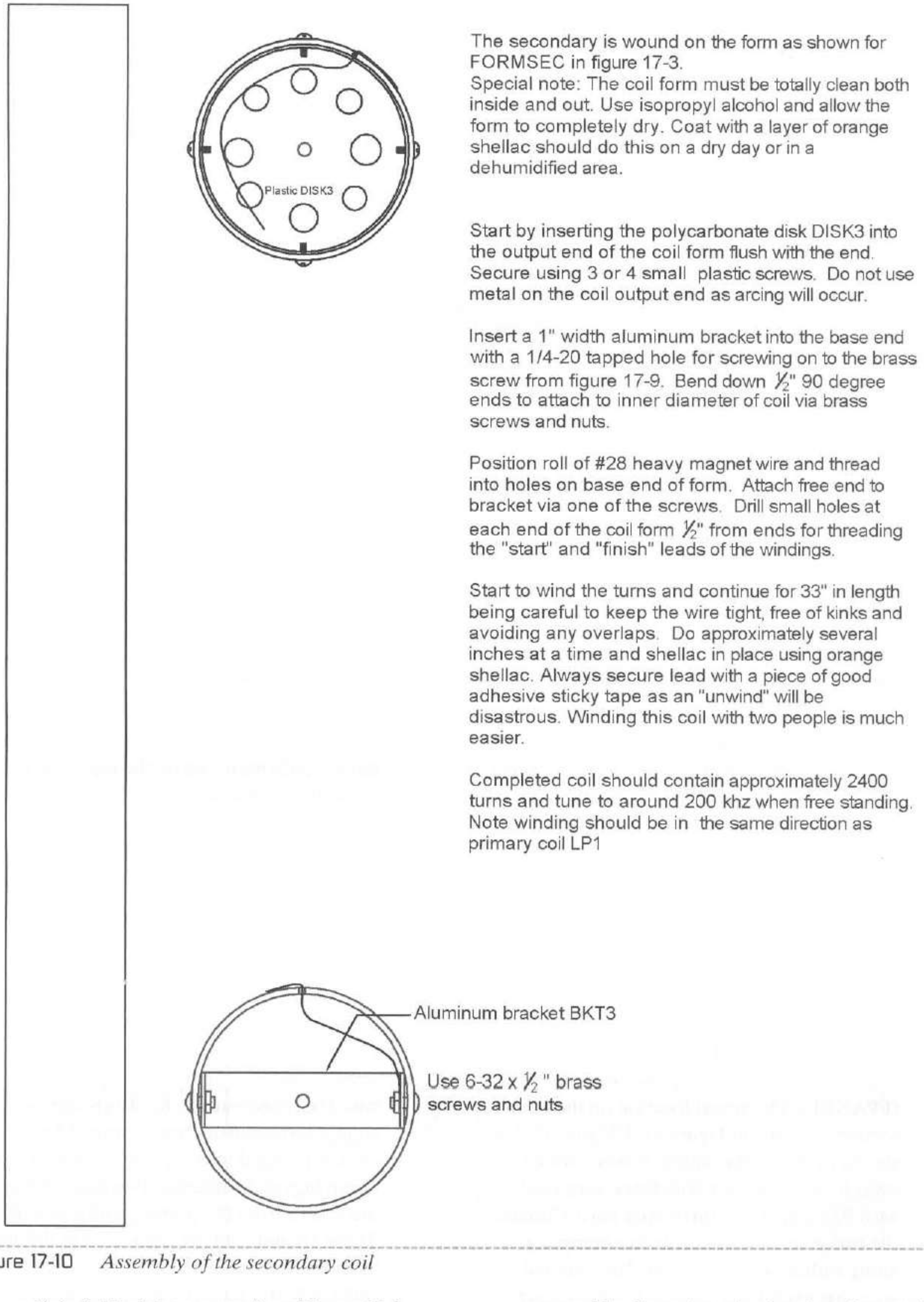

The secondary is wound on the form as shown for FORMSEC in figure 17-3.
Special note: The coil form must be totally clean both inside and out. Use isopropyl alcohol and allow the form to completely dry. Coat with a layer of orange shellac should do this on a dry day or in a dehumidified area.

Start by inserting the polycarbonate disk DISK3 into the output end of the coil form flush with the end. Secure using 3 or 4 small plastic screws. Do not use metal on the coil output end as arcing will occur.

Insert a 1" width aluminum bracket into the base end with a 1/4-20 tapped hole for screwing on to the brass screw from figure 17-9. Bend down ½" 90 degree ends to attach to inner diameter of coil via brass screws and nuts.

Position roll of #28 heavy magnet wire and thread into holes on base end of form. Attach free end to bracket via one of the screws. Drill small holes at each end of the coil form ½" from ends for threading the "start" and "finish" leads of the windings.

Start to wind the turns and continue for 33" in length being careful to keep the wire tight, free of kinks and avoiding any overlaps. Do approximately several inches at a time and shellac in place using orange shellac. Always secure lead with a piece of good adhesive sticky tape as an "unwind" will be disastrous. Winding this coil with two people is much easier.

Completed coil should contain approximately 2400 turns and tune to around 200 khz when free standing. Note winding should be in the same direction as primary coil LP1

Figure 17-10 *Assembly of the secondary coil*

a $\frac{3}{8}$-inch Plexiglas square from Figure 17-3. Fabricate a 5$\frac{3}{4}$-inch PVC disc (DISCPRI) from $\frac{1}{2}$-inch-thick sheet. This piece is for securing and centering the primary coil assembly, shown later in the plans. Secure it in place on the deck via a $\frac{1}{4}$-20 × 1-inch brass screw and nut. This will be the grounding point for the bottom of the secondary coil and

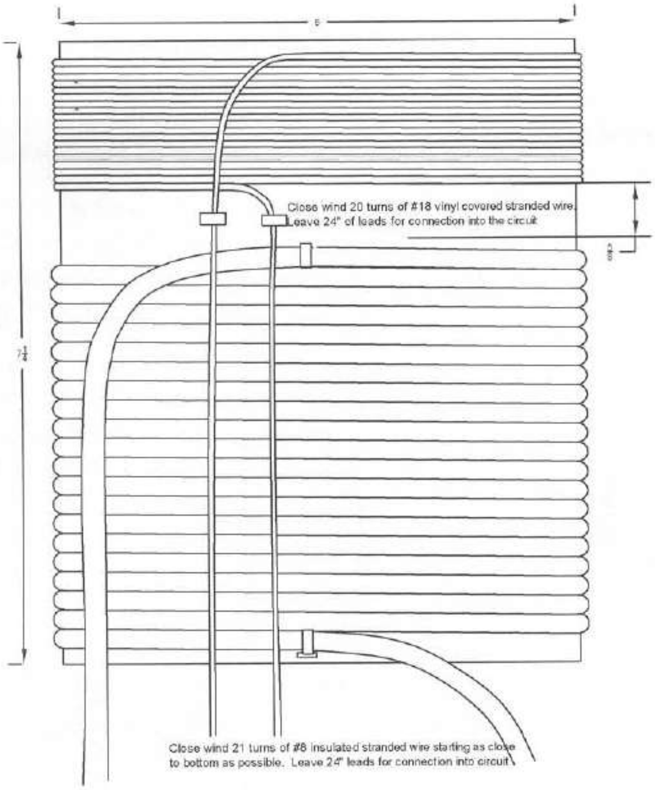
6
7¾
Close wind 20 turns of #18 vinyl covered stranded wire.
Leave 24" of leads for connection into the circuit
Close wind 21 turns of #8 insulated stranded wire starting as close
to bottom as possible. Leave 24" leads for connection into circuit

Figure 17-11 *Assembly of the primary coil*

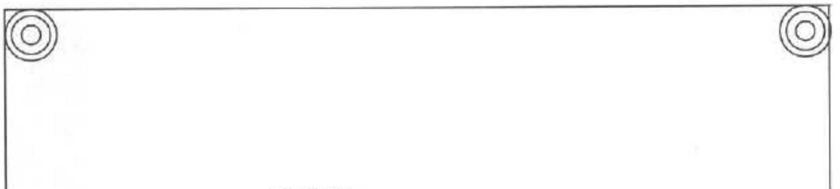
DISCPR1
FROM FIG 17-11
FROM FIG 17-10

Figure 17-12 *Final assembly of top deck*

goes to earth ground, as shown in Figure 17-7. Proceed to drill the eight 1-inch air flow holes.

At this time, the builder should be aware that if his or her coil is to be used on a continuous basis, the second blower mounted on the underside of the top deck is necessary to keep the bottom of the coils cool. They will heat up over time due to the resonant rise of current associated with the base of a quarter-wave resonator. If your coil is to be used for a short-time operation of several minutes, the part involving the drilling of the air flow holes and the second fan may be omitted.

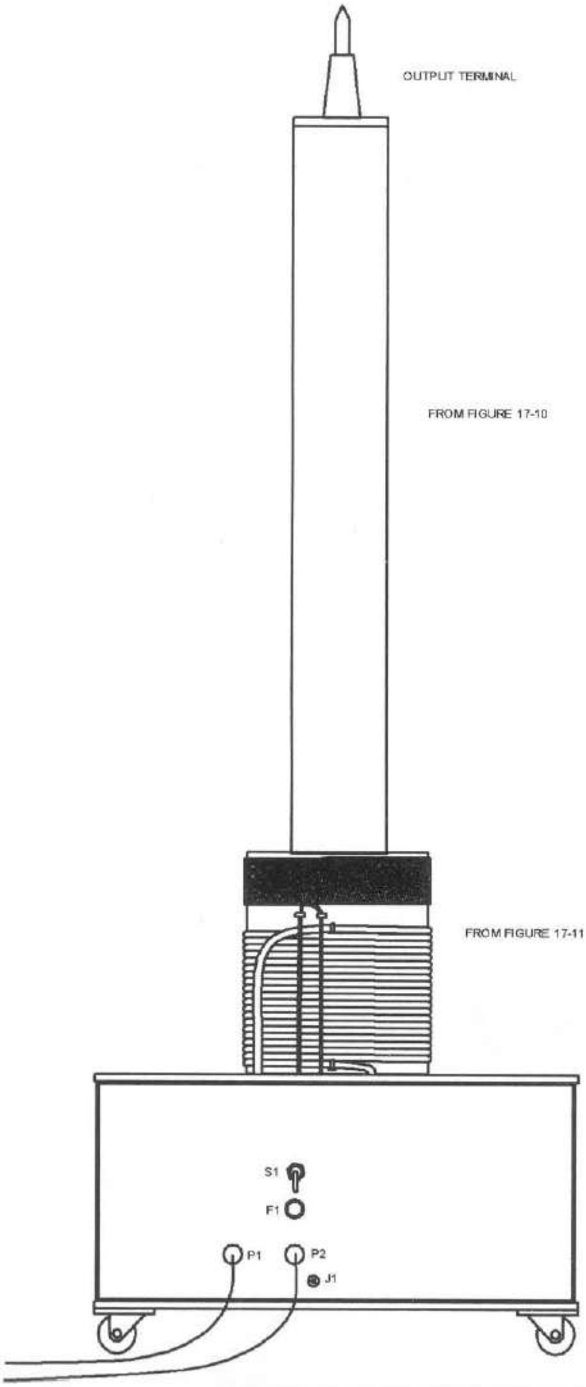

Figure 17-13 *Side view of final assembly*

8. Assemble and wind the secondary coil as shown in Figure 17-10. Proper winding of the coil is mandatory for achieving optimal operation. Kinks and overlaps will seriously affect output.
9. Assemble the primary coil assembly, as shown in Figure 17-11. Start at the bottom and tightly wind the turns. Use tie wraps to secure the start and finish of this heavier 21 turns of the #8 stranded wire. You may want to place dabs of silicon rubber to secure them in place. Start

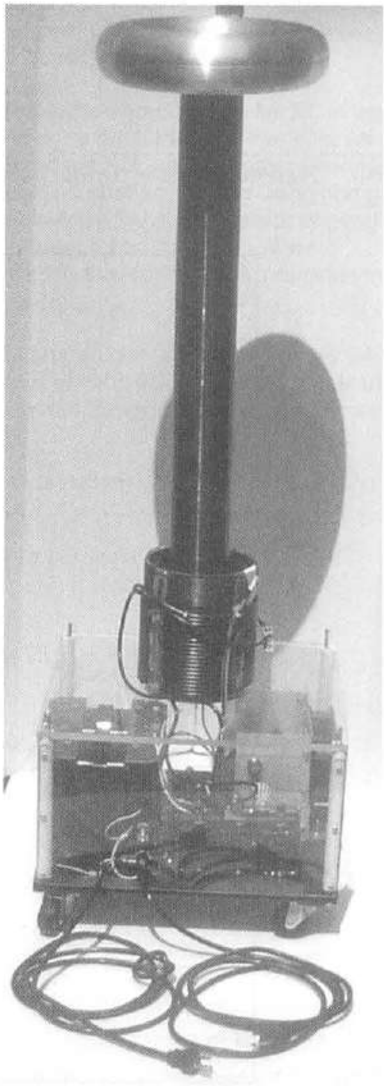

Figure 17-14 *The final assembly*

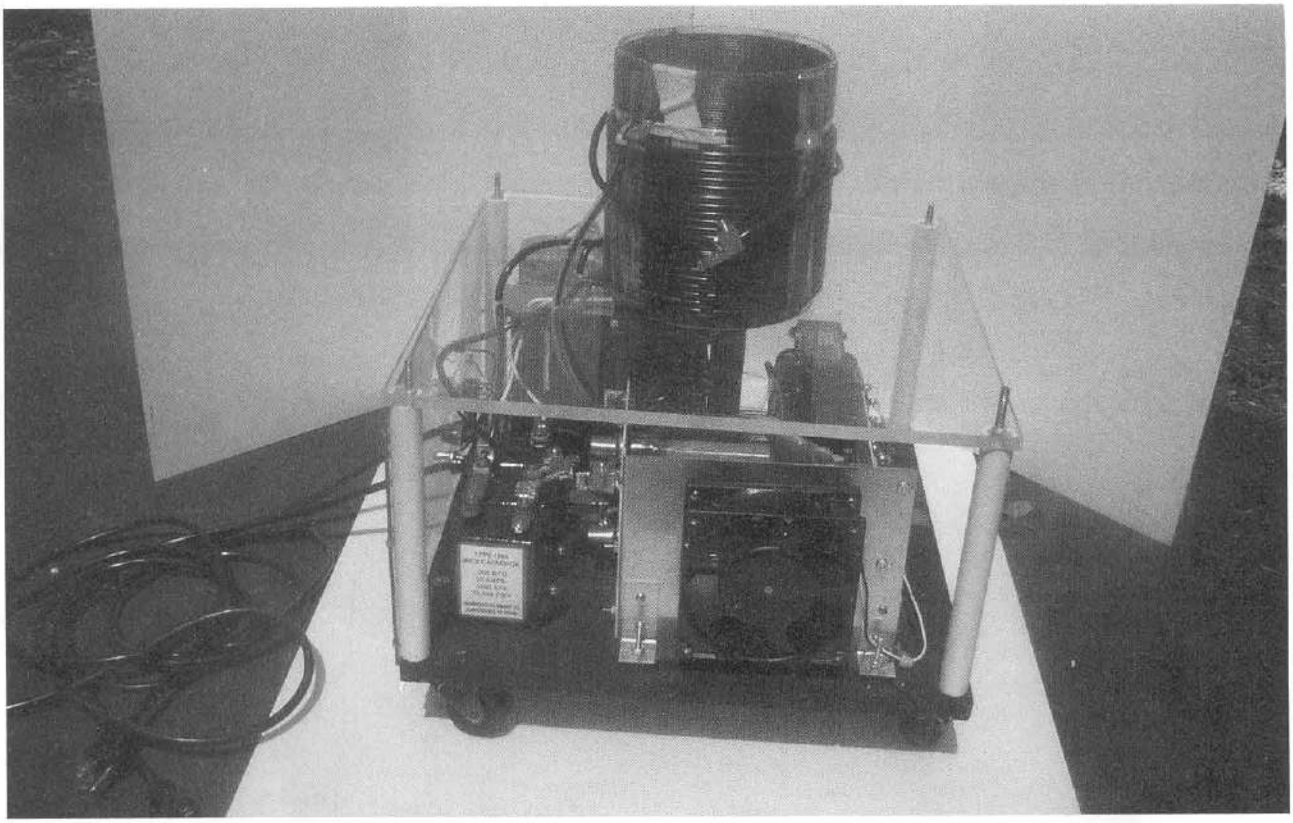

Figure 17-15 *The right-side view*

from the top and wind 20 turns of #18 vinyl-covered stranded wire and tie wrap it in place. Note that there must be at least ½-inch between the closest proximities of these two coils as a flashover (or spark break) could damage the vacuum tube. Leave at least 2 feet of extra leads for connections, as shown in Figure 17-8.

10. Finally, assemble the top deck, as shown in Figure 17-12. The primary coil section slips over the disc DISCPRI, as shown in Figure 17-8. The tapped hole in the bracket of the secondary coil screws onto the center brass screw shown in Figure 17-8. It is this screw that grounds and secures the secondary coil.

    Route the leads from the coils through their respective holes for connection into the circuit. You will note back in Figure 17-8 that the heavy plate leads are shown connected to C3A via heavy crimp lugs and are also soldered. The same applies for the smaller grid leads that connect across C2. These leads may require switching during the test steps. Both sets of these leads carry high-frequency energy with the heavier leads carrying high-radio-frequency voltages and currents. These heavier leads must be as short and direct as possible and not be in close proximity to other objects.

    Figure 17-13 shows the completed graphic of the coil assembly with an insulator and a pointed discharge pin. Any suitable feed through the insulator around 2 to 3 inches in height can be used. You may also use a piece of polycarbonate or Teflon rod for this part. Other plastics may carbonize and burn.

    The following figures are actual photos showing different views of the assembly apart and intact. They are intended as assembly aids.

Figure 17-14 shows a 12 × 3-inch toroid as the output terminal. You will note the two power cords exiting through the front panel.

Figure 17-15 shows the right side with a view of the blower B1 and tube-mounting bracket.

Figure 17-16 shows the left side with a view of the filament and plate transformers.

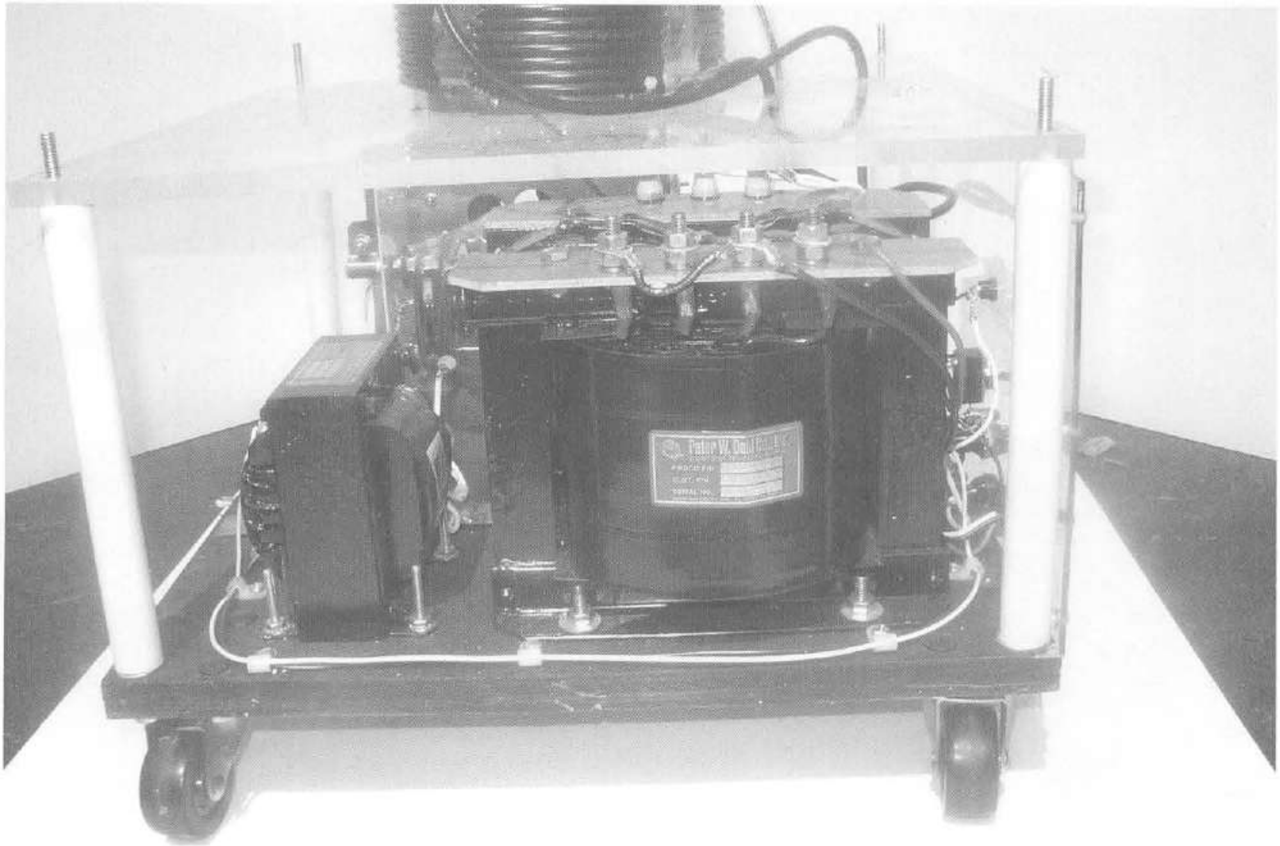

Figure 17-16 *The left-side view*

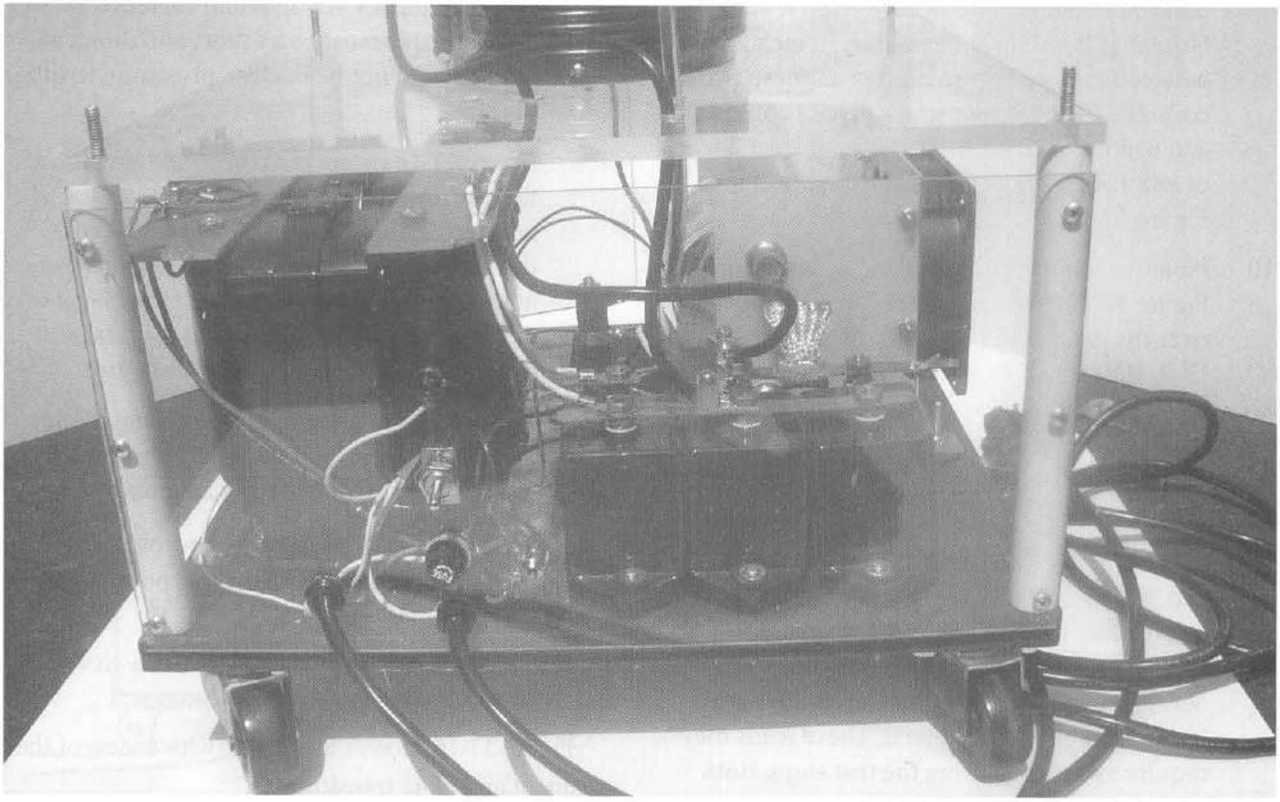

Figure 17-17 *The front view*

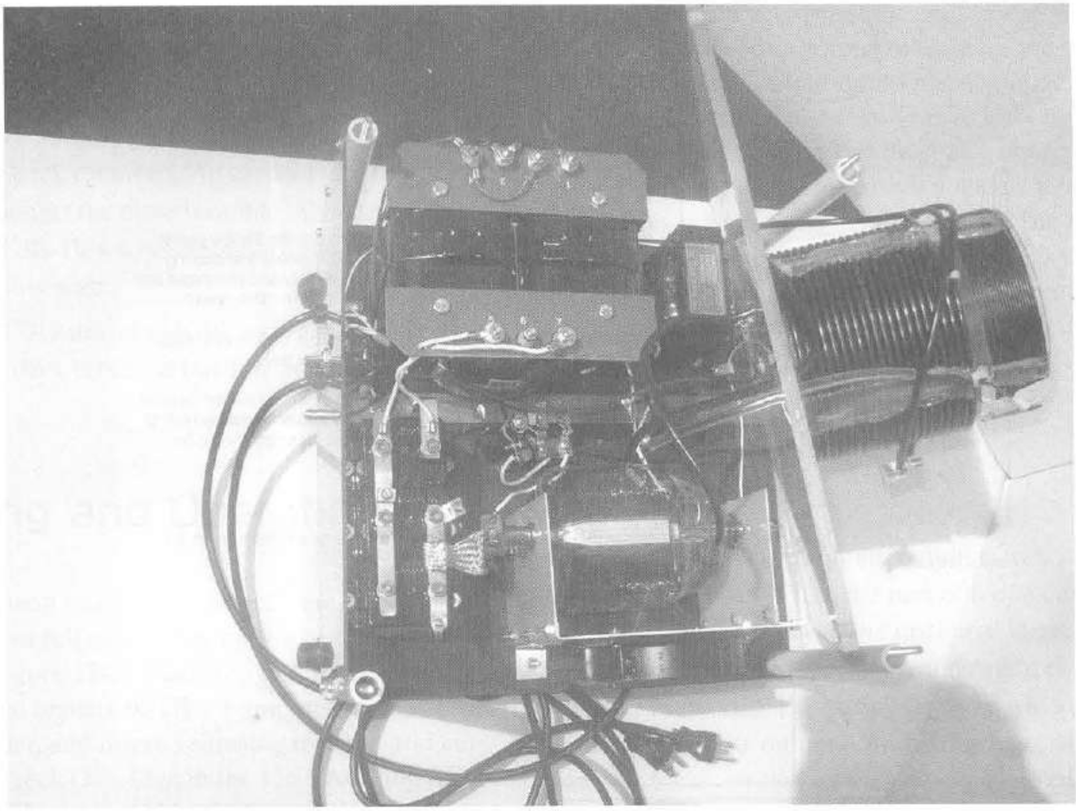

Figure 17-18 *The top view*

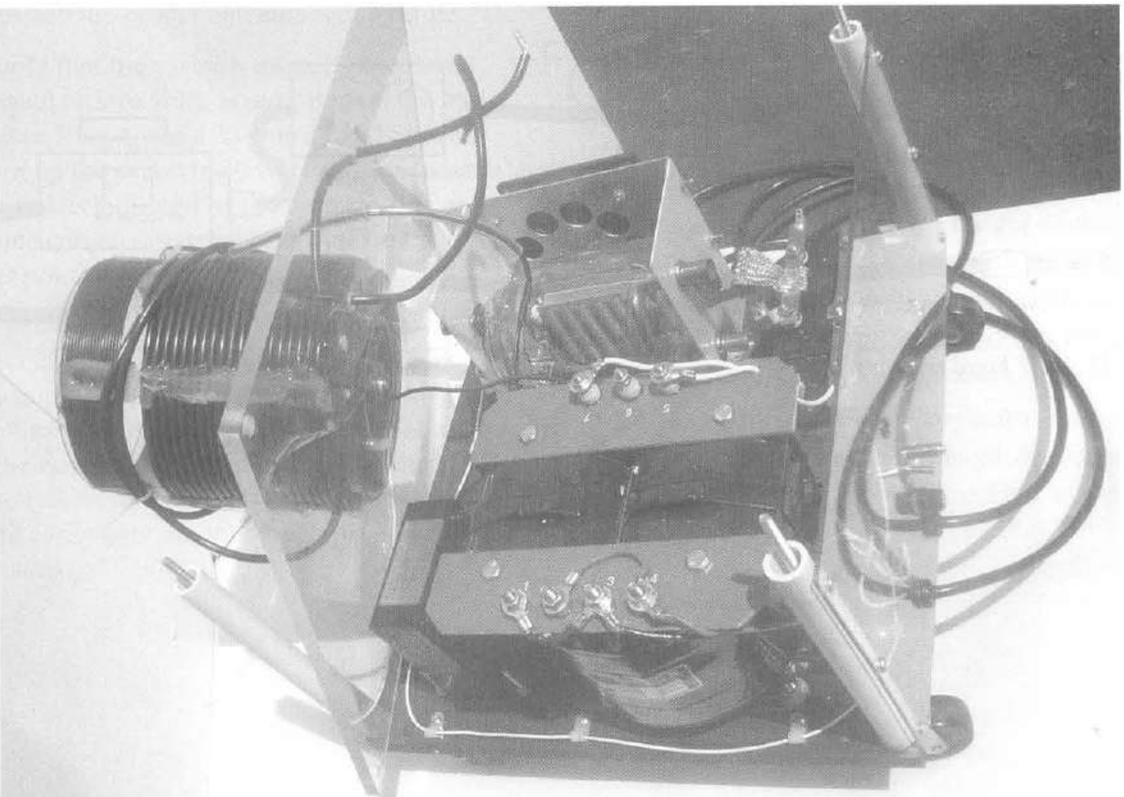

Figure 17-19 *Another top view*

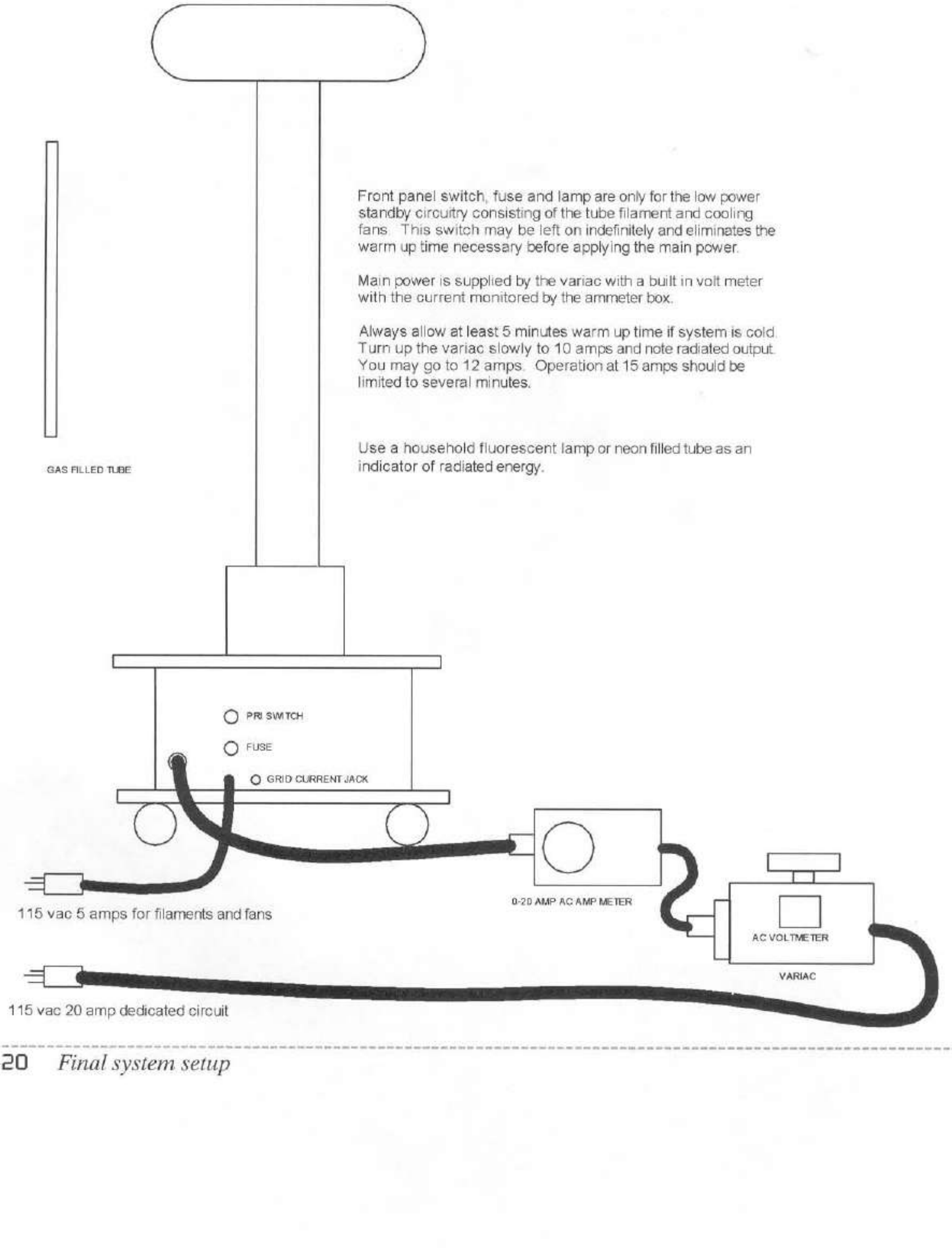
GAS FILLED TUBE
PRI SWITCH
FUSE
GRID CURRENT JACK
0-20 AMP AC AMP METER
AC VOLTMETER
VARIAC
115 vac 5 amps for filaments and fans
115 vac 20 amp dedicated circuit

Front panel switch, fuse and lamp are only for the low power standby circuitry consisting of the tube filament and cooling fans. This switch may be left on indefinitely and eliminates the warm up time necessary before applying the main power.

Main power is supplied by the variac with a built in volt meter with the current monitored by the ammeter box.

Always allow at least 5 minutes warm up time if system is cold. Turn up the variac slowly to 10 amps and note radiated output. You may go to 12 amps. Operation at 15 amps should be limited to several minutes.

Use a household fluorescent lamp or neon filled tube as an indicator of radiated energy.

Figure 17-20 *Final system setup*

Figure 17-17 shows the front view through the clear plastic panel. Note the three mica capacitors C3A, C3B, and C4.

Figure 17-18 shows the top view with the clear plastic top deck removed. We used a piece of shielded strap to connect the plate terminal of the tube to C3A and C3B. This is not necessary and may be heavy flexible wire.

Figure 17-19 also shows the top view with the clear plastic top deck removed but at a different angle.

Testing and Operating

1. You will need a dedicated 20-amp line to obtain full output. Set up the system as shown in Figure 17-20. Place the unit away from metal objects, insert a 7-amp fuse in the holder, and insert a shorting into the grid current jack (J1). Plug in the 115 VAC plug for the filament and blower fans, and wait for the tube to glow and the fans to blow. Always allow at least 3 minutes for the tube filament to reach the proper emission temperature.
2. Verify that the variac is properly wired and turned to zero volts, as indicated on the voltmeter. Plug it into a 20-amp circuit. Slowly turn up the variac to 30 volts and verify some output as indicated by the bulb glowing or some corona discharge occurring on the output pin.

   If you do not observe some output, it will be necessary to completely shut the system down by unplugging both plugs, reversing the grid coil wires that attach to C2, and repeating the previous step. This is referred to as phasing the system. You should get substantial output and easily light the tube when properly phased.
3. Once you verify that the coil is properly phased, you may start to increase the voltage, noting the output discharge, the input current, and the applied AC voltage as indicated on the variac meter. You must also observe that the plate of the tube does not start to glow, as this is an indication of mistuning. The tube may have a very soft glow at full 120 VAC where the output should be 15 inches or more. The input current will approach 20 amps.

Notes

Tuning the coil for maximum output is tricky and may take some effort on the part of the builder. The tuning values will vary as the discharge increases due to the virtual capacitance of the ionization cloud at the top of the coil. The output terminal will also greatly effect the output tuning. Try adding or subtracting a turn or two on the primary plate coil, as this will change the resonant drive signal. You will need more inductance as the output increases, and the trick is to select the optimum value at the desired input power.

A 12-inch toroid or another large capacitive terminal will cause larger circulating currents within the coil and greatly enhance the wireless transmission of energy, as can be demonstrated by observing the distance that is possible for lighting the gas discharge tube without a wire. This will eliminate the air breakdown discharge when operating in this mode as the terminal voltage drops unless you increase the input power to overcome this loading effect.

The geometry of the secondary coil as shown is not the optimum for maximum length discharges. A squatter size would be more efficient but would radiate less.

Table 17-1 Vacuum Tube Tesla Coil Parts List

Ref. #	Description	DB Part #
R2	5K, 100-watt wire-wound resistor	
C1	.004-microfarad, 1,000-volt polypropylene capacitor	DB# .0039/400
C2	.003-microfarad, 3-kilovolt mica capacitor	DB# .003/3 KVM
C3A, B, C4	Three .01-microfarad, 10-kilovolt mica capacitors	DB# .01/10 KVM
V1	833A Triode vacuum tube	DB# 833A
T1	3,000-volt, .5-amp plate transformer	DB# 3KV/.5A
T2	10-volt, 10-amp filament transformer	DB# 10V/10A
B1, 2	5 × 5-inch muffin fan and 5 × 5-inch high-output muffin fan	DB # FAN1
S1	*Single pole, single throw* (SPST) 5-amp toggle switch	
F1/FS1	Panel-mount fuse holder and 5-amp fuse	
P1	8-foot #16 three-wire power cord	
P2	8-foot #18 three-wire power cord	
AMPMETER	20-amp AC panel mount ammeter	
VARIAC	18-amp variac with built-in meter	DB# VARIAC18A
TUBE	15-inch neon tube	DB# NE15
TER1	12 × 3-inch toroidal terminal	DB# TO12

Chapter Eighteen

Universal Capacitance Discharge Ignition (CDI) Driver

Figure 18-1 shows how to construct a very useful driver circuit that can be used to pulse most induction and ignition coils in the forward direction. High-quality automotive ignition coils can produce 50- to 75,000-volt pulses capable of jumping several inches. The system allows control of the pulse repetition rate up to 100 reps/second and individual pulse energy up to .3 joules. A maximum power output of over 30 watts is possible. Our laboratory model has generated discharges of up to 10 inches from some high-performance ignition coils.

The advantage of a *capacitance discharge ignition* (CD) system is its ability to produce high peak currents and voltages simultaneously. This property is well understood and appreciated in the automotive industry where fouled spark plugs are prevented by the brisance of the discharge. Electric fences burn away vegetation growth from contacting the wire by using high current pulses. Stun guns use these high peak current pulses to thwart the neuromuscular system. All this peak power with very low resultant energy allows many useful applications where battery-operated systems can be efficiently utilized.

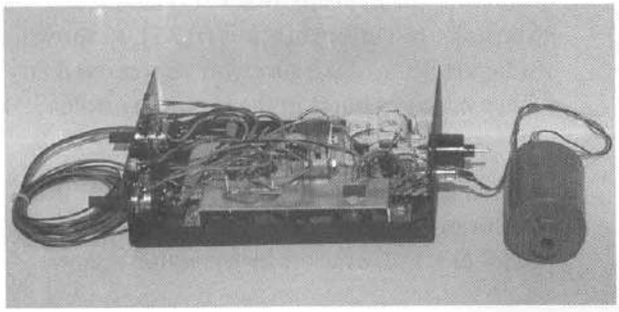

Figure 18-1 *The capacitor discharge driver*

This is an intermediate-level project requiring basic electronic skills and one should expect to spend $25 to $50. All the parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 18-1.

Circuit Description

Figure 18-2 shows the complete circuit schematic of the adjustable energy and repetition rate capacitor discharge driver. The adjustable voltage regulator I2 controls the operating voltage to the system via control pot R9/S1 from 6 to 12 volts. Input to I2 is a 12-volt DC wall adapter transformer with a current rating of 1 to 2 amps being controlled by switch S1. This adjustable voltage is fed to the switching transistors Q1 and Q2 via the center tap of stepup transformer T1. The switching action of these transistors is controlled by a complementary pulse driver circuit I1 with its frequency controlled by trimpot R3. This circuit generates a high-frequency signal that alternates between the two primary windings, in turn inducing a higher voltage in the secondary winding.

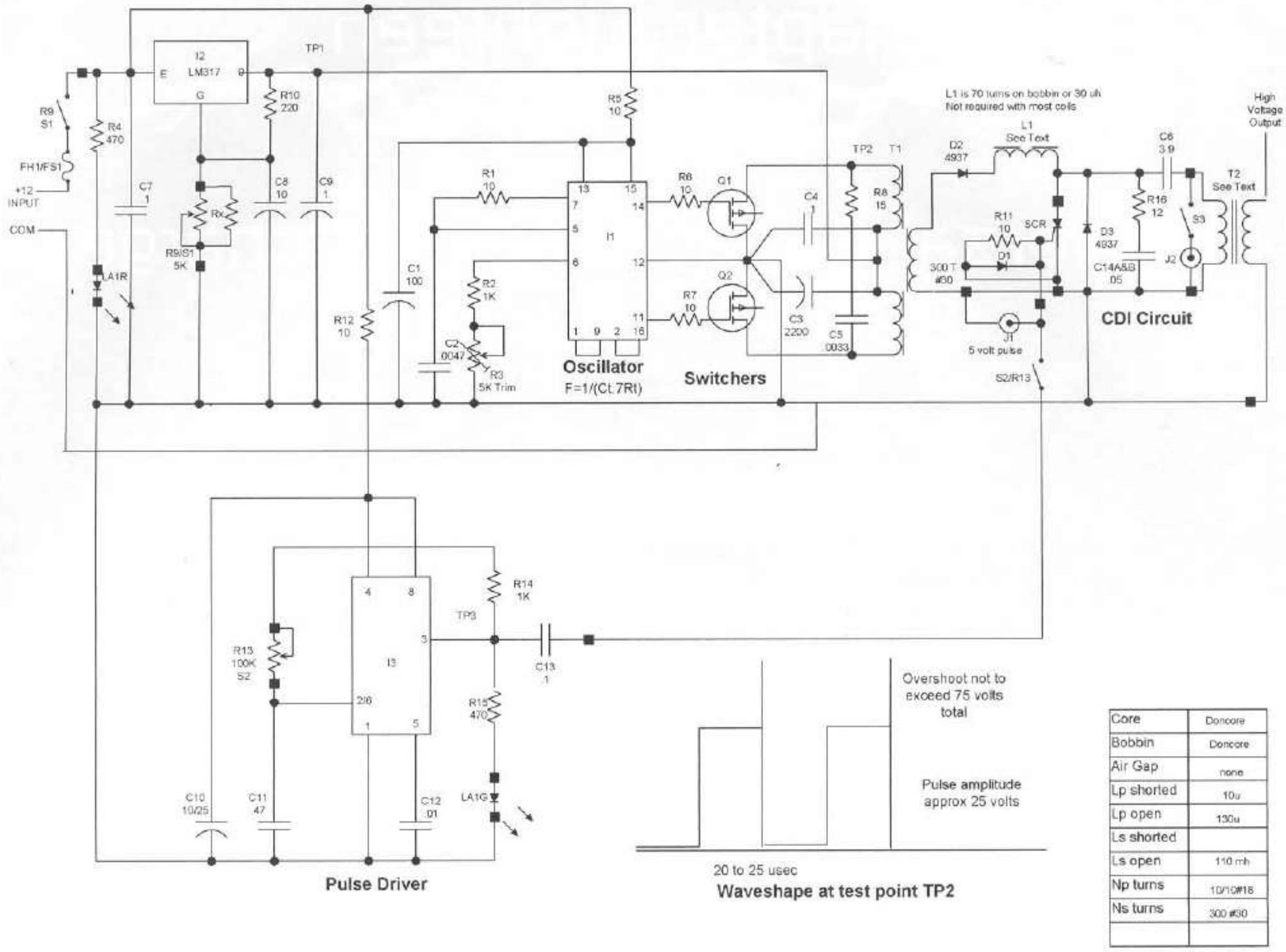
I2
LM317
E
G
TP1
R10
220
R9
S1
R4
470
FH1/FS1
+12
INPUT
COM
C7
1
R9/S1
5K
Rx
C8
10
C9
1
LA1R
R5
10
R1
10
C1
100
R12
10
R2
1K
C2
.0047
R3
5K Trim
I1
Oscillator
F=1/(Ct.7Rt)
R6
10
Q1
R7
10
Q2
C4
1
C3
2200
C5
.0033
R8
15
TP2
T1
Switchers
L1 is 70 turns on bobbin or 30 uh
Not required with most coils
L1
See Text
D2
4937
R11
10
SCR
D1
300 T
#30
J1
5 volt pulse
S2/R13
D3
4937
C14A&B
.05
C6
3.9
R16
12
S3
J2
T2
See Text
High
Voltage
Output
CDI Circuit
I3
R13
100K
.52
R14
1K
TP3
C13
.1
R15
470
LA1G
C10
10/25
C11
.47
C12
.01
Pulse Driver
Overshoot not to
exceed 75 volts
total
Pulse amplitude
approx 25 volts
20 to 25 usec
Waveshape at test point TP2

Figure 18-2 *Circuit schematic*

Core	Doncore
Bobbin	Doncore
Air Gap	none
Lp shorted	10u
Lp open	130u
Ls shorted	
Ls open	110 mh
Np turns	10/10#18
Ns turns	300 #30

T1 is a transformer with a preset, built-in leakage reactance that controls the short-circuit-charging current to discharge capacitor C6 through rectifying diode D2. Once C6 is charged, it can be discharged by the silicon controlled rectifier switch (SCR) through the primary high-voltage pulse transformer T2. The SCR is controlled by timer chip I3, connected as a variable pulse generator via pulse rate adjust pot R13/S2. It is these pulses that trigger the SCR to switch energy from C6 into T2, generating the high-voltage output pulses. You may also externally switch the SCR via a pulse signal at input jack J1. R13/S2 must be off for this external function. The pulse transformer is shown as a low-power unit in the parts list.

The real intention of the circuit is the ability to pulse-drive many different ignition coils at different energy levels as selected by R9 and pulse repetition rates as selected by R13. We have obtained sparks up to 4 inches with conventional automotive coils.

Assembly Steps

1. Assemble the switching transformer as shown in Figure 18-3.
2. Fabricate the heatsink bracket (HS1) and transistor-mounting bracket (BR1), as shown in Figure 18-4. Make sure you've removed any sharp edges or burrs in the mounting holes that could break through the insulating pads.
3. Attach Q1 and Q2 as shown, noting that the mounting tabs are insulated from the actual metal heatsink bracket by the nylon screws.

1. Bifilar (parallel) wind two different color #18 magnet wire for 10 turns on the primary bobbin. Tape in place leaving 8" leads. Note that different colors will help identifying the lead ends.
   #18 solid magnet wire can be used however high frequency LITZ wire will give a slight improvement. You can make this wire by obtaining 6 pieces of #26 magnet wire and twist together as a single wire.
2. Wind 300 turns of #30 magnet wire in the 10 segments of the secondary bobbin. Wind 30 turns per segment and attach the start and finish leads to the connection lugs
3. Place the wound coils onto the cores as shown and tape tightly into place.
4. You may verify the inductance as follows secondary open:
   A to B&C are around 130 micro henries.
   D to B&C are around 130 micro henries
   Secondary winding 110 millihenrys

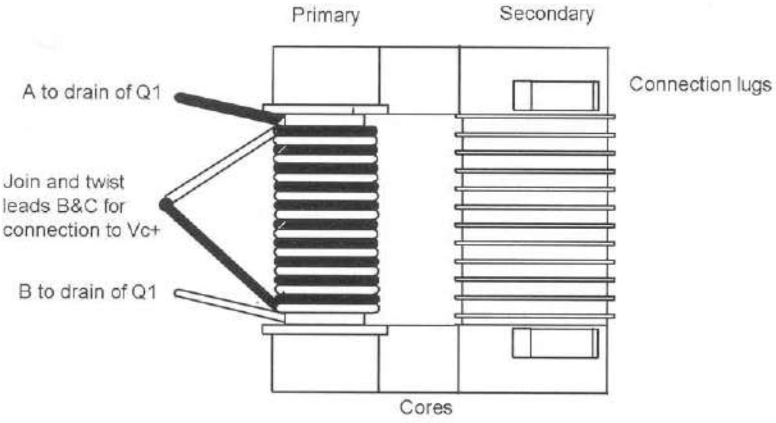
Primary
Secondary
A to drain of Q1
Connection lugs
Join and twist leads B&C for connection to Vc+
B to drain of Q1
Cores

*#18 solid magnet wire can be used however high frequency LITZ wire will give a slight improvement. You can make this wire by obtaining 6 pieces of #26 magnet wire and twist together as a single wire.

Figure 18-3 *T1 Transformer construction*

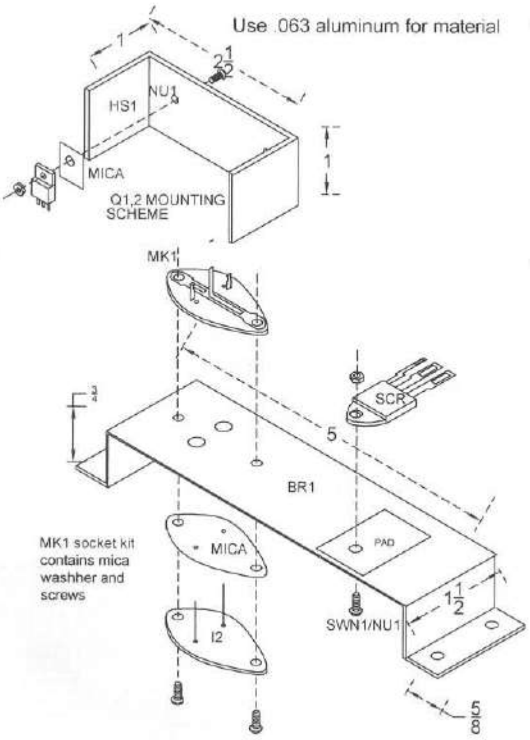
Use .063 aluminum for material
1
2 1/2
HS1
NU1
MICA
Q1,2 MOUNTING SCHEME
1
MK1
3/4
SCR
5
BR1
PAD
MK1 socket kit contains mica washher and screws
MICA
I2
SWN1/NU1
1 1/2
5/8

Figure 18-4 *Assembly of brackets*

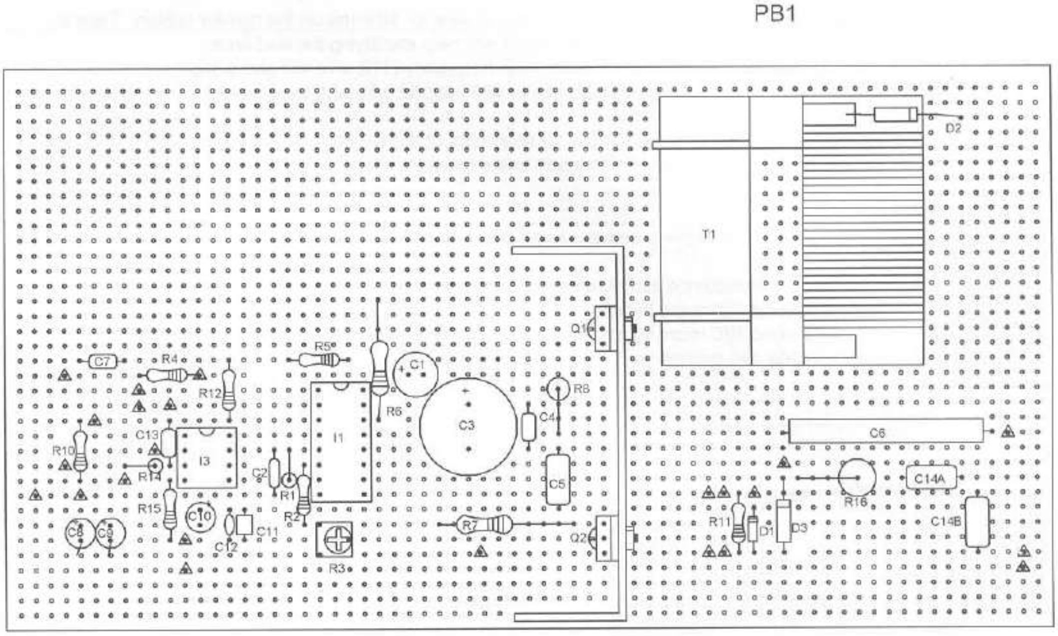
PB1
D2
T1
Q1
C7
R4
R5
C1
R12
R8
R6
C3
C4
I1
C13
I3
R10
C2
R14
C5
R1
R15
C8
C9
C10
C11
R2
C12
R3
R7
Q2
C6
R16
C14A
R11
D1
D3
C14B

▲ This symbol indicates points for wires to external components

Figure 18-5 *Assembly board and parts location*

Attach I2 using the socket assembly kit (MK1) and attach the SCR using a nylon screw and an insulating pad (PAD).

4. Cut a piece of .1 × .1 grid perforated circuit board (PB1) at 7 × 3.8 inches.
5. Insert the components as shown in Figure 18-5, starting from the lower left-hand corner. Count the perforated holes for a guide. Note the polarity marks on C1, C3, C8, C9, and C10 and diodes D1, D2, and D3. It is suggested that you use sockets SO16 for I1 and SO8 for I3.
6. Wire the components as shown in Figure 18-6 using the leads of the actual components as the connection runs, which are indicated by the dashed lines. Always avoid bare-wire bridges and messy solder joints. Check for cold or loose solder joints and secure T1 using double-sided sticky tape or plastic tie wraps.

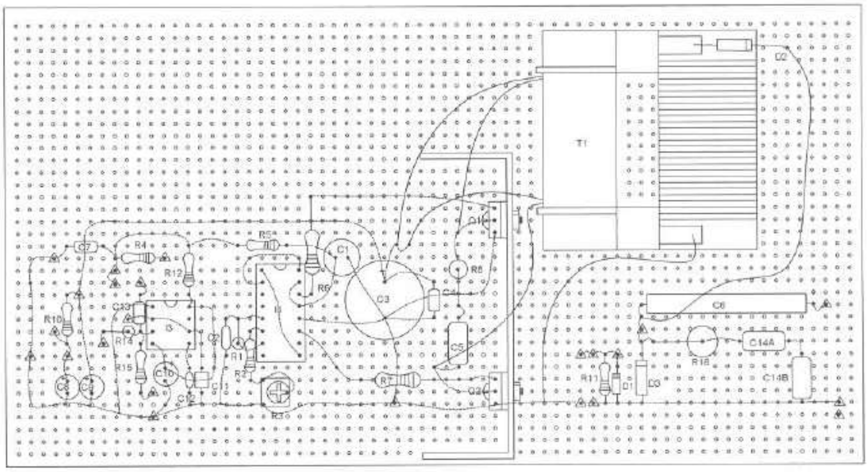
D2
T1
Q1
C7
R4
R5
C1
R12
R8
R6
C3
C4
I1
C13
I3
R10
C2
R14
C5
R1
R15
C8
C9
C10
C11
R2
C12
R3
R7
Q2
C6
R16
C14A
R11
D1
D3
C14B

Figure 18-6 *Internal onboard wiring*

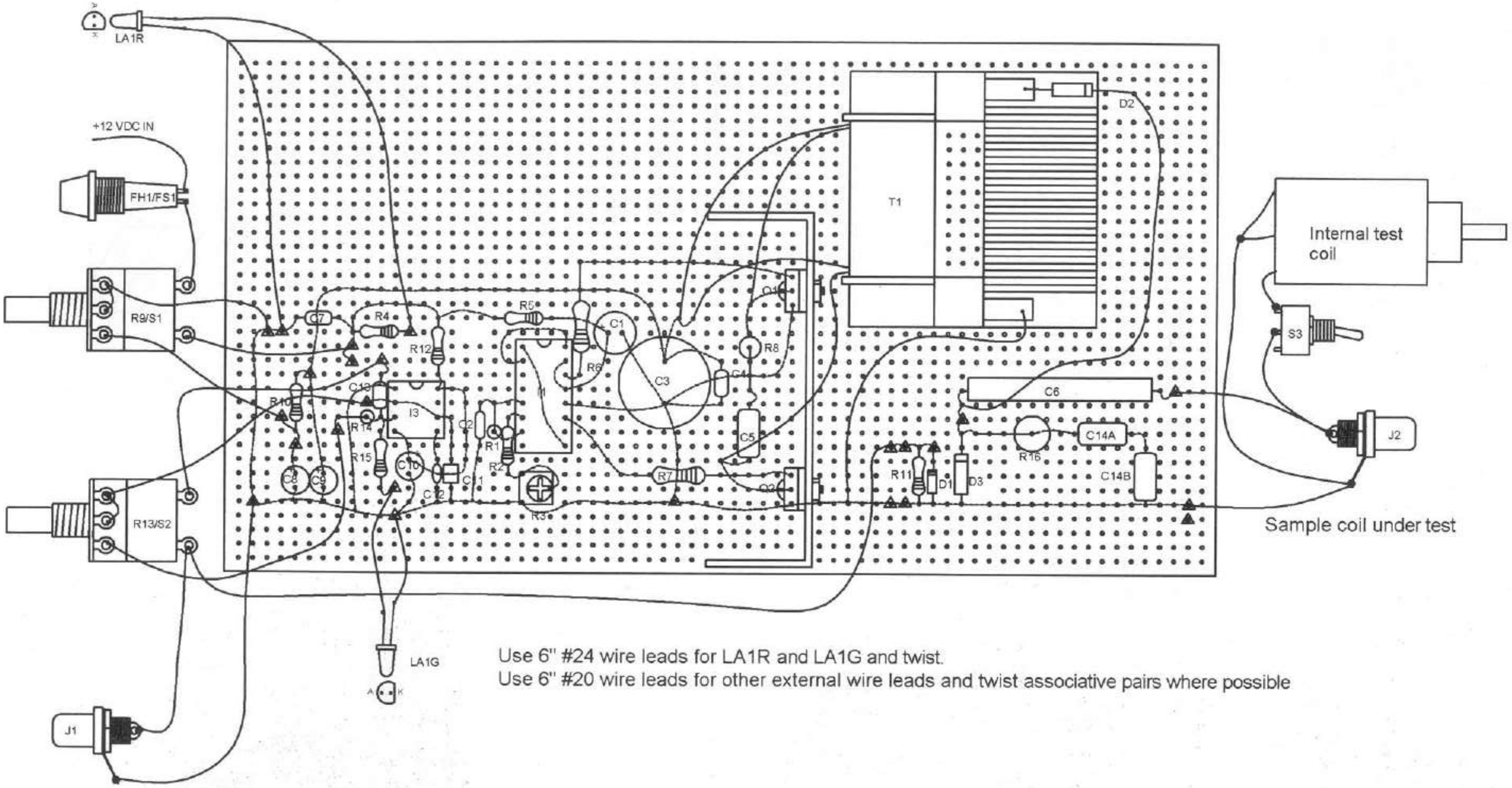
LA1R
A
K
+12 VDC IN
FH1/FS1
R9/S1
R13/S2
J1
LA1G
A
K
C7
R4
R5
R12
R10
C13
I3
R14
R15
C8
C9
C10
C12
C11
C2
R1
R2
R3
I1
R6
C1
C3
C4
R8
C5
R7
Q1
Q2
T1
D2
C6
R16
C14A
C14B
R11
D1
D3
Internal test coil
S3
J2
Sample coil under test

Use 6" #24 wire leads for LA1R and LA1G and twist.
Use 6" #20 wire leads for other external wire leads and twist associative pairs where possible

Figure 18-7 *External wiring to controls and input*

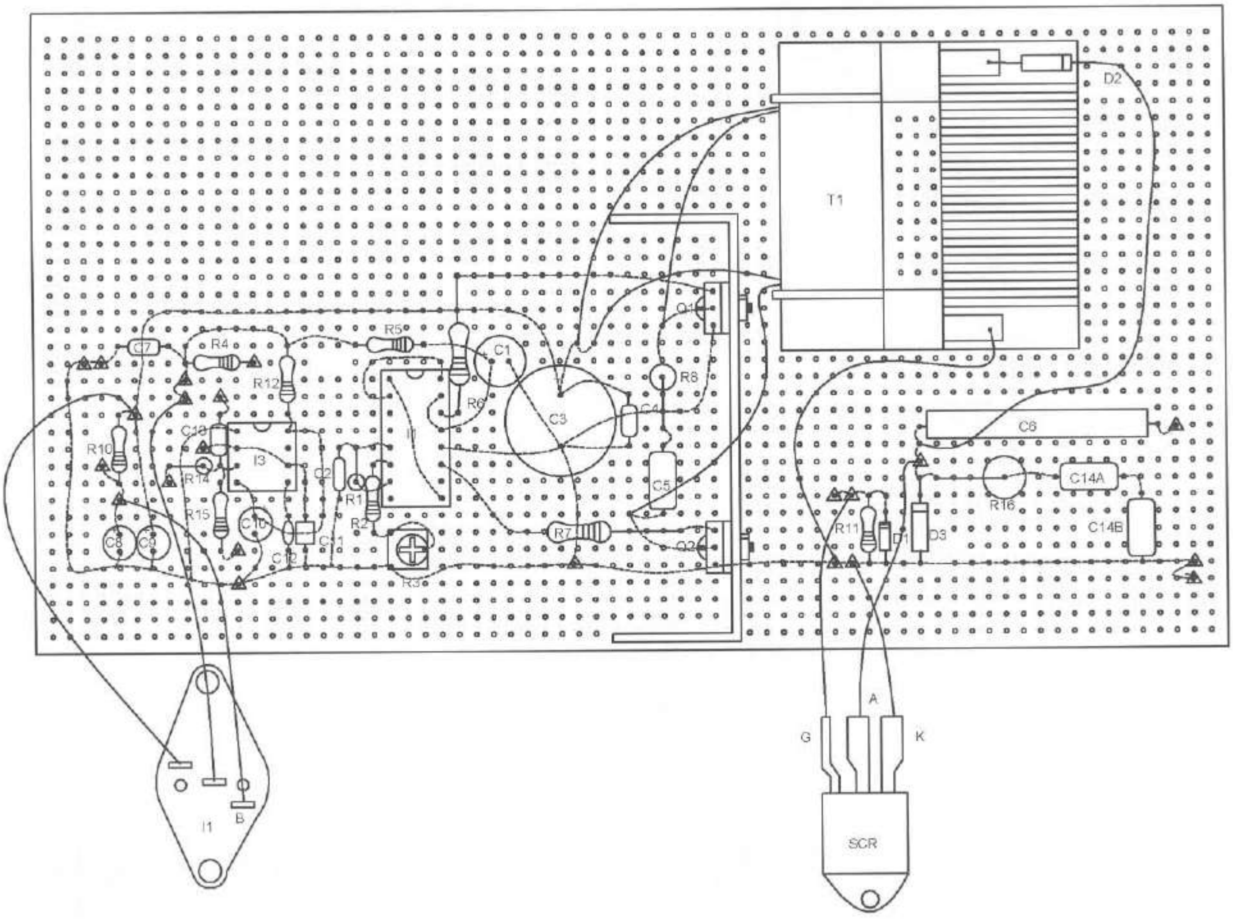
D2
T1
Q1
R5
R4
C7
C1
R12
R8
R6
C3
C4
C6
C13
I3
I2
R10
R14
C2
R1
C5
C14A
R16
C14B
R11
D1
D3
R15
C10
C8
C9
C12
C11
R2
R3
R7
Q2
A
G
K
I1
B
SCR

Figure 18-8 *External wiring to transistor modules*

The external coil of choice is connected to J3 using short lengths of #18 wire. It is a good idea to rig a ground return for the output spark to jump to. Always use one of the primary leads of the external coil as the ground return point. Other points used for ground may cause voltage loops that can damage the unit. We use a dedicated gap with connection leads containing clips. The gap is adjustable from ½ to 3".

S3 disables the internal test coil.

J1 is intended for input from an external trigger and requires R13 to be fully off.

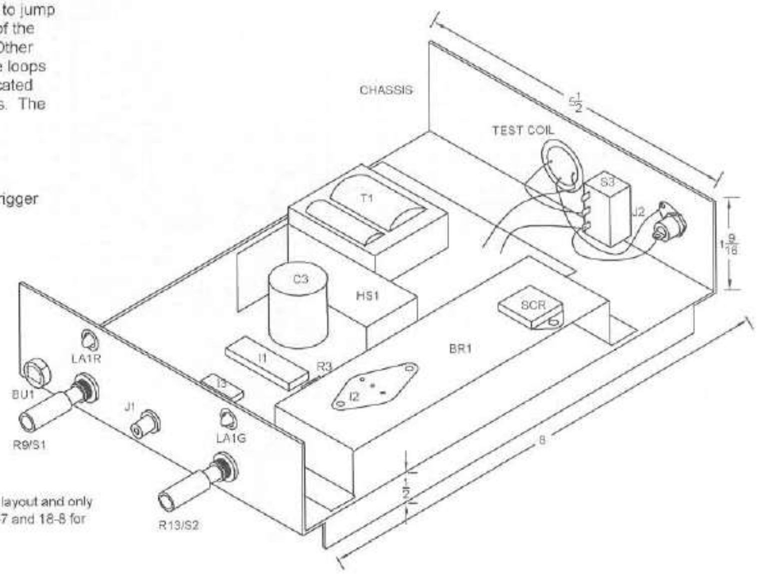
CHASSIS
5½
TEST COIL
S3
J2
1 9/16
T1
C3
HS1
SCR
BR1
I1
R3
I3
I2
LA1R
BU1
J1
R9/S1
LA1G
R13/S2
8
¾

Note this view is intended to show the general layout and only indicates certain components. Use figures 18-7 and 18-8 for wiring of the specific parts as shown

Figure 18-9 *Final assembly*

7. Wire the external components as shown in Figure 18-7.

8. Wire the transistors from the assembled bracket, as shown in Figure 18-8 and detailed in Figure 18-4.

Fabricate the mounting chassis at the overall dimensions shown in Figure 18-9. Trial-fit the remaining components, proceeding to drill and locate with the proper sized holes. Attempt to follow the figure and note the position of the BR1 bracket assembly and front panel controls. Route and twist all lead pair to neaten appearance. You may want to splice in some output leads to replace the connections to the test pulse transformer T2. These should be #18 wire and as short as physically possible to the external ignition coil. Also create a suitable cover if required and ventilate it with holes for air cooling.

Test Steps

This step will require basic electronic laboratory equipment, including a 60 MHz oscilloscope.

1. Turn off both controls R9 and R13. Set the R3 trimpot at midrange.
2. Connect a scope to TP2 and a voltmeter to TP1.
3. Apply 12 VDC to the input leads and turn on R9/S1. Adjust the meter for 5 volts and adjust the period of the waveshape at TP2 to a 25 us period. The input current should be less than .2 amps on a 12-volt power supply.
4. Note that the LED indicators are on with LA1G flashing at the pulse repetition rate determined by the R13 setting. This occurs even with R13/S2 fully off.
5. Turn on R13 and adjust it for 10 volts. Note the test coil sparking over its outer surface. Rotate R13 and notice the pulse rate increasing. Only operate the test coil to test circuit performance.

Notes on Operation

The unit is intended to drive larger ignition and induction coils. Jack J2 is intended as the output to the external coil under operation.

Note that the test coil must be disabled for external operation and may be disconnected by unsoldering or using a suitable high-current switch that may be mounted on the rear panel adjacent to J2.

When using unknown coils, it is advised that you monitor the DC voltage across the SCR. It should never exceed 400 volts, as the storage capacitor will be overrated. The SCR can handle up to 800 volts.

A change occurs in the charging voltage as the pulse repetition rate is changed. It decreases as the repetition rate increases. You can compensate for this by adjusting R9, but you must be careful to turn it down before lowering the repetition rate. Always start with the voltage control R9 at the lowest setting and increase as necessary. Additionally, jack J1 is intended for external triggering and must be used with R13/S2 in the off position.

Table 18-1 Parts List for CDI Driver

Ref. #	*Description*	*DB Part #*
R1, 5, 6, 7, 11, 12	Six 10-ohm 1/4-watt resistors (br-blk-blk)	
R2, 14	Two 1K, 1/4-watt resistors (br-blk-red)	
R3	5K horizontal trimmer	
R4, 15	Two 470 ohm, 1/4-watt resistors (yel-pur- br)	
R8	15-ohm, 3-watt metal oxide resistor	
R9/S1, R13/S2	Two 10K pots and switches	
R10	220-ohm 1/2-watt resistor (red-red-br)	
R16	12-ohm, 2-watt metal oxide resistor	
Rx	Select a 1/4-watt resistor to shunt R9 to improve range	
C1	100-microfarad, 25-volt vertical electrolytic capacitor	
C2	4,700-picofarad, 50-volt polyester capacitor	
C3	2,200-microfarad, 25-volt vertical electrolytic capacitor	
C4, 7, 13	Three .1-microfarad, 50-volt polyester capacitors	
C5	.0033-microfarad, 250-volt polypropylene capacitor	
C6	3- to 5-microfarad, 350-volt polyester capacitor	DB# 3.9M/350
C9	1-microfarad, 25-volt vertical electrolytic capacitor	
C8, 10	Two 10-microfarad, 25-volt vertical electrolytic capacitors	
C11	.47-microfarad, 50-volt polyester capacitor	
C12	.01-microfarad, 50-volt disc capacitor	
C14	.05-microfarad, 400-volt metalized polypropylene	
I1	*Integrated circuit* (IC) driver and pulse-width modulator (LM3525)	
I2	1.5-amp adjustable voltage regulator (LM317)	
I3	Timer *dual inline package* (DIP) IC (LM555)	
Q1, 2	Two 100-volt *metal-oxide-semiconductor field effect transistors* (MOSFETS) (IRF540)	
SCR1	70-amp, 700-volt *silicon-controlled rectifier* (SCR)	DB# S8070W
D1	IN914 small-signal diode	
D2, 3	Two IN4937 fast-recovery, 1,000-volt, 1-amp diodes	
LA1R	Red *light-emitting diode* (LED)	
LA1G	Green LED	
T1	Inverter high-voltage transformer, as shown in Figure 18-3	DB# TRANCDI
T2	20-kilovolt pulse transformer (see text)	DB# CD25B
FH1	Panel-mount fuse holder	
FS1	3-amp fuse	
J1, 2	RCA phono jack	
S3	*Single pole, single throw* (SPST) 5-amp or more toggle switch	
PB1	7 × 3.8 × .1-inch perforated circuit vector board	
SOCK8	Eight-pin IC socket	
SOCK16	16-pin IC socket	
MK1	TO3 transistor mounting kit	
PAD	Thermo pad insulator	
CHASSIS	#20-gauge aluminum chassis, as shown in Figure 18-9	
HS1	Heatsink, as shown in Figure 18-4	
BR1	Transistor-mounting bracket, as shown in Figure 18-4	
MICA	Two Mica insulating pads for Q1 and Q2	
SWN1	Two 6-32 × 1/2-inch nylon screws	
SW1	6-32 × 1/2-inch screws	
NU1	Seven 6-32 nuts	
WIRE20B	12 inches of #20 bus wire	
WR20	Six feet of #20 vinyl hookup wire	
BU1	Small clamp bushing	
BU2, 3	Two 3/8-inch plastic bushings for LED	

Chapter Nineteen

Long-range Telephone Conversation Transmitter

Figure 19-1 shows a device that allows the user to monitor a phone's incoming and outgoing calls without any connecting wires. The system is intended to transmit both sides of a telephone conversation to any FM radio or headset-type receiver, such as a Walkman radio. It allows the user to perform outside activities such as mowing the lawn or working outside the house while completely monitoring the telephone over the radio.

The unique feature of this device is that activation occurs only when the phone is off the hook. This eliminates transmitting a continuous signal, thus preventing potential interference and unnecessarily wearing down the batteries. It also eliminates dead-tape listening time, as now the tape only records voices when the telephone is used. You may also listen to a radio station while monitoring the telephone. When the phone is used, the radio station is replaced by the transmission of the unit.

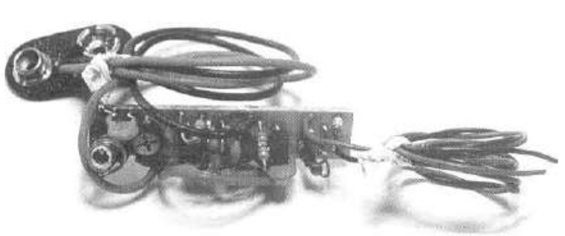

Figure 19-1 *Wireless circuit*

The device is shown assembled on a perforated circuit board and may be housed in a small plastic enclosure. Potting (or encapsulating with epoxy) and sealing is at the discretion of the builder.

Circuit Description

This chapter shows how to build a mini-powered FM transmitter that transmits both sides of a telephone conversation to a nearby FM radio. The circuit consists of a *radio-frequency* (RF) oscillator section combined with a telephone-activated switch section that controls the oscillator power (see Table 19-1). Transistor Q1 forms a relatively stable RF oscillator whose frequency is determined by the values of coil L1 and tuning capacitor C4. The C4 setting determines the desired operating frequency and is in the standard FM broadcast band with tuned circuit design, favoring the high end up to 110 MHz. Capacitor C2 supplies the necessary feedback voltage developed across resistor R3 in the emitter circuit of Q1, sustaining an oscillating condition. Resistors R1 and R2 provide the necessary bias of the base emitter junction for proper operation, while capacitor C1 bypasses any RF to ground fed to the base circuit. Capacitor C3 provides an RF return path for the tank circuit of L1 and C4 while blocking the DC supply voltage fed to the collector of Q1 (see Figure 19-2).

You will note that the junction of the base bias resistors R1 and R2 is a feed point consisting of capacitor C5 and resistor R4. This point is where the speech voltage from the telephone is applied. Because of the nature of the oscillator frequency being subject to change by varying the base bias condition, a varying AC voltage superimposed at this point causes a corresponding frequency shift (FM) along with an amplitude-modulated (AM) condition. It is this property that allows the circuit to be FM modulated by the speech occurring on the telephone lines. The signal is clearly detected by any FM receiver when properly tuned.

Be aware that before the phone receiver is lifted off the hook, a DC voltage of approximately 50 volts along with some AC hum is measured across the green and red wires of the phone line. This 50 volts is used to keep transistors Q2 and Q3 "off" by preventing forward biasing of the Q2 junction. The zener diode Z1 reverse voltage is exceeded as long as the voltage on the telephone lines remains above the zener voltage. When the phone is off the hook, the voltage is now below the zener voltage, and resistor R5 now biases Q2 "on," which turns Q3 on and clamps the common return of the oscillator to the negative of the battery, commencing operation by turning on.

Note that Q2 and Q3 make a simple DC switch that produces a minimal battery current during the on-hook conditions. When the receiver is lifted, commencing normal conversation, this DC voltage drops to less than 10 volts, and an existing AC voltage corresponding to the speech conversation now modulates the oscillator as described. You will also note Q2 and Q3 being a simple DC switch that only allows a minimal battery current during the on-hook condition.

Capacitor C5 is necessary to block the DC component of the phone line when an on-hook condition exists and allows little attenuation of the varying AC speech voltage. Resistor R4 is necessary for attenuation of this speech voltage that could cause overmodulation and unnecessary serious sidebands if not properly selected.

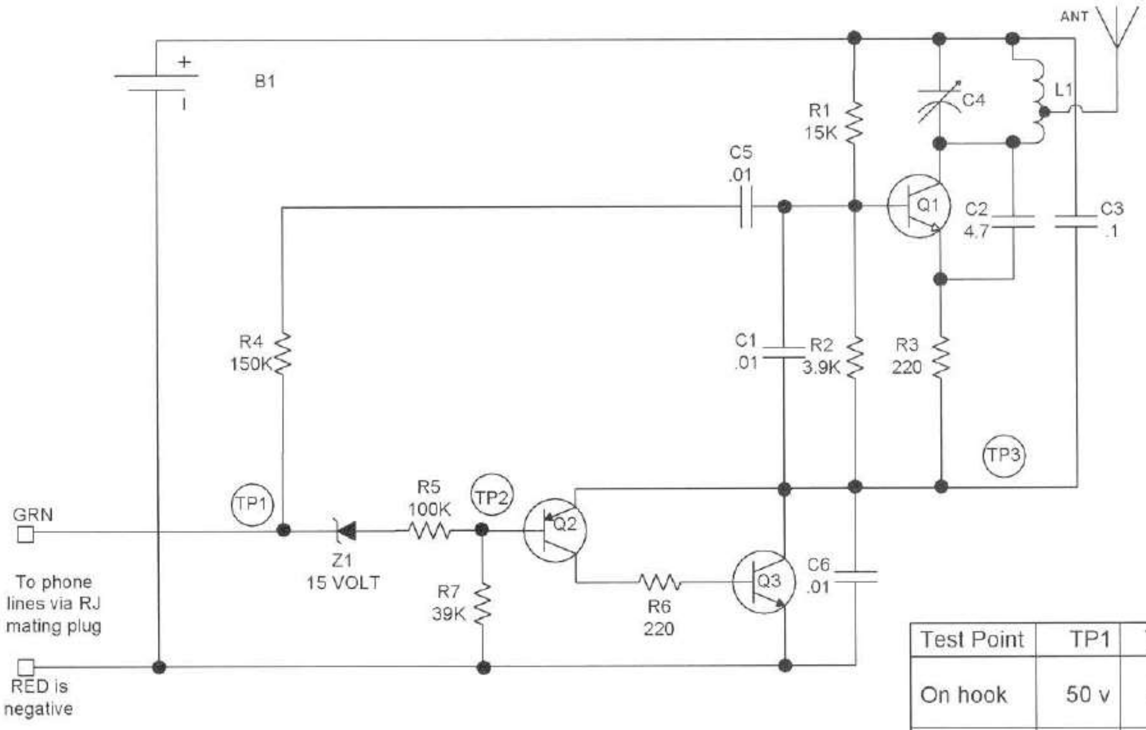

Test Point	TP1	TP2	TP3
On hook	50 v	10 v	9 v
Off hook	8 v	.7 v	.1 v

Measured from negative battery lead

Figure 19-2 *Circuit schematic*

Assembly of the Circuit Board

Assemble the circuit as shown in Figure 19-3.

1. Fabricate a piece of .1-inch grid perforated board to a size of 2 × 1 inches.
2. Form L1 by tightly wrapping eight turns of #16 bus wire on a #8 wood screw. This produces an eight-turn coil with an inner diameter of approximately .135 inches and is about .625 inches in length. Insert it in the proper hole and solder as shown.
3. Insert trimmer capacitor C4 into the holes as shown. You may put this component on either side of the PC board. This choice is up to you and should be determined by the final packaging scheme you require. Three holes are available for the part, with two of them electrically the same point. This is necessary as some trimmers may have three pins. Make sure that the common pins are connected to the same electrical points with the odd pin to the other point.
4. If you are building from a perforated board, it is suggested that you insert components starting in the lower left-hand corner. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate connection runs on the underside of the assembly board.
5. Attach the external leads of battery clip CL1 and the RJ cable and plug them into the phone connection jack.
6. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
7. Tune an FM receiver to a fairly strong station at the high end of the band (108 MHz or higher). Turn up the volume and position it at a distance of 25 to 50 feet.
8. If a multimeter or 50-milliamp meter is available, connect it in series with the battery lead.

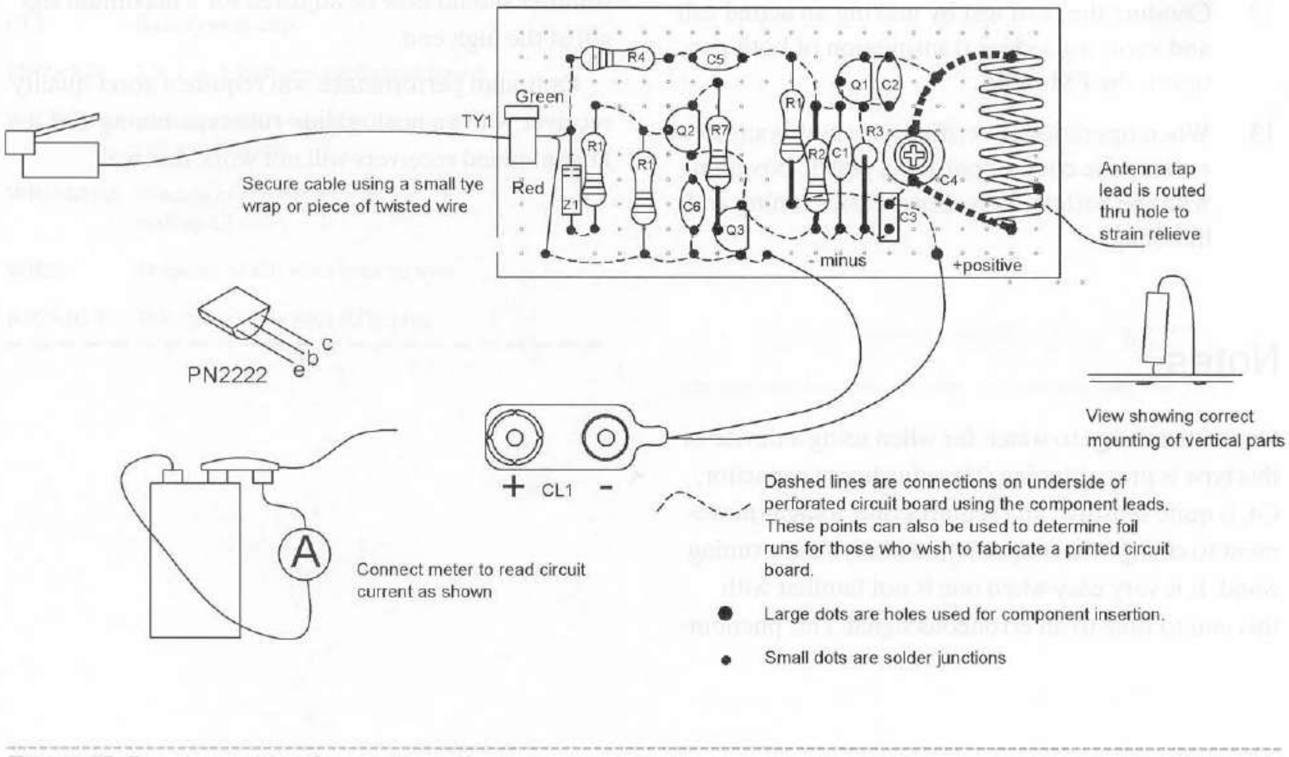

Secure cable using a small tye wrap or piece of twisted wire

Antenna tap lead is routed thru hole to strain relieve

View showing correct mounting of vertical parts

Connect meter to read circuit current as shown

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

● Large dots are holes used for component insertion.

• Small dots are solder junctions

Figure 19-3 *Assembly of circuit board*

This can be done by removing one of the clip fasteners and connecting the meter to the free contacts, as shown in Figure 19-3. The meter should read 5 to 7 milliamps. Use a short piece of bare wire to touch the coil L1 a turn at a time, starting from the C3 end. Note that as you progress turn by turn away from C3 that the indicated meter current will drop or change.

9. Connect the operating unit to the phone line with the phone on hook and notice the meter reading dropping to almost nothing. You may have to reverse the connections to the phone lines. Check the following voltage points indicated on the schematic to verify proper circuit operation, as shown in Figure 19-2.
10. Take the phone off the hook and tune C4 with a tuning wand until you hear the telephone tone. It may be difficult at first to spot the signal as this adjustment is very touchy. It is important to know that several spots in the adjustment may be erroneous and will be weak and unstable. The correct signal will be strong, stable, and clear, and it can be verified by distance-of-transmission testing.
11. Once the desired setting of C4 is found, the position should be marked on the adjustment capacitor with the frequency noted.
12. Conduct the final test by making an actual call and verifying a clear transmission of both parties to the FM radio.
13. When operation is verified, you may want to enclose the circuit board in a plastic box along with the battery. This allows easier tuning and handling.

Notes

One of the things to watch for when using a device of this type is proper tuning. The adjustment capacitor, C4, is quite sensitive and requires only a slight movement to change the frequency, so always use a tuning wand. It is very easy when one is not familiar with this unit to tune to an erroneous signal. This phenomenon is likely to occur when the unit is close to the monitoring receiver. As stated previously, an erroneous signal will be weak, distorted, and unstable (often it is mistaken for the main signal and the unit is blamed for poor performance), and the main signal will be strong, stable, and undistorted, if modulated.

Several experiments in tuning the unit should be done before attempting to use it for the desired application. Also, whenever possible, the unit should be used around 108 to 109 MHz, which is at the boundary of the aircraft band and upper FM broadcast band. When the approximate desired spot is found, final touchup tuning should be done at the receiver end for clarity. In most areas, these upper frequencies are clear and allow uninterrupted use in contrast to the lower ones in the FM band where a slight change in frequency from a clear spot results in the unit being drowned out by a strong broadcast station. Do not go above 108 MHz if near an airport or air traffic lane.

Most available FM radios can easily be detuned slightly to shift the dial readings down to where the dial reading of 108 MHz is actually 109 MHz. This is accomplished by carefully adjusting the "oscillator," the padding trimmer located on the main tuning capacitor, and "walking" a known station down the necessary megahertz or two. An antenna-peaking trimmer should now be adjusted for a maximum signal at the high end.

Optimum performance will require a good-quality receiver with an analog slide-rule-type tuning dial. Digital tuned receivers will not work that well.

Table 19-1 Long-Range Phone Conversation Transmitter Parts List

Ref. #	*Description*	*DB Part #*
R1	15K, 1/4-watt resistor (br-gr-or)	
R2	3.9K, 1/4-watt resistor (or-wh-red)	
R3, 6	Two 220 ohm, 1/4-watt resistor (red-red-br)	
R4	150K, 1/4-watt resistor (br-gr-yel)	
R5	100K, 1/4-watt resistor (br-blk-yel)	
R7	39K, 1/4-watt resistor (or-wh-or)	
C1, 5, 6	Three .01-microfarad, 50-volt disc capacitors	
C2	4- to 6-picofarad small silver mica disk (4.7)	DB# 4.7P
C3	.1-microfarad, 50-volt disc capacitor	
C4	6- to 35-picofarad trimmer capacitor	DB# 635P
Z1	15-volt zener diode (1N5245)	
Q1, 3	Two PN2222 NPN general-purpose transistors	
Q2	PN2907 PNP general-purpose transistor	
L1	Coil wound as described in assembly steps	DB# COIL8T
CL1	Battery snap clip	
PBOARD	2 × 1 × .1-inch grid perforated board	
WR24BUSS	12 inches of #24 bus wire for wiring and antenna tap	
WR16BUSS	6 inches of #16 bus wire for making L1 coil	
WR20	24 inches of #20 vinyl hookup wire	
RJCABLE	Telephone cable with RJ11 plug	

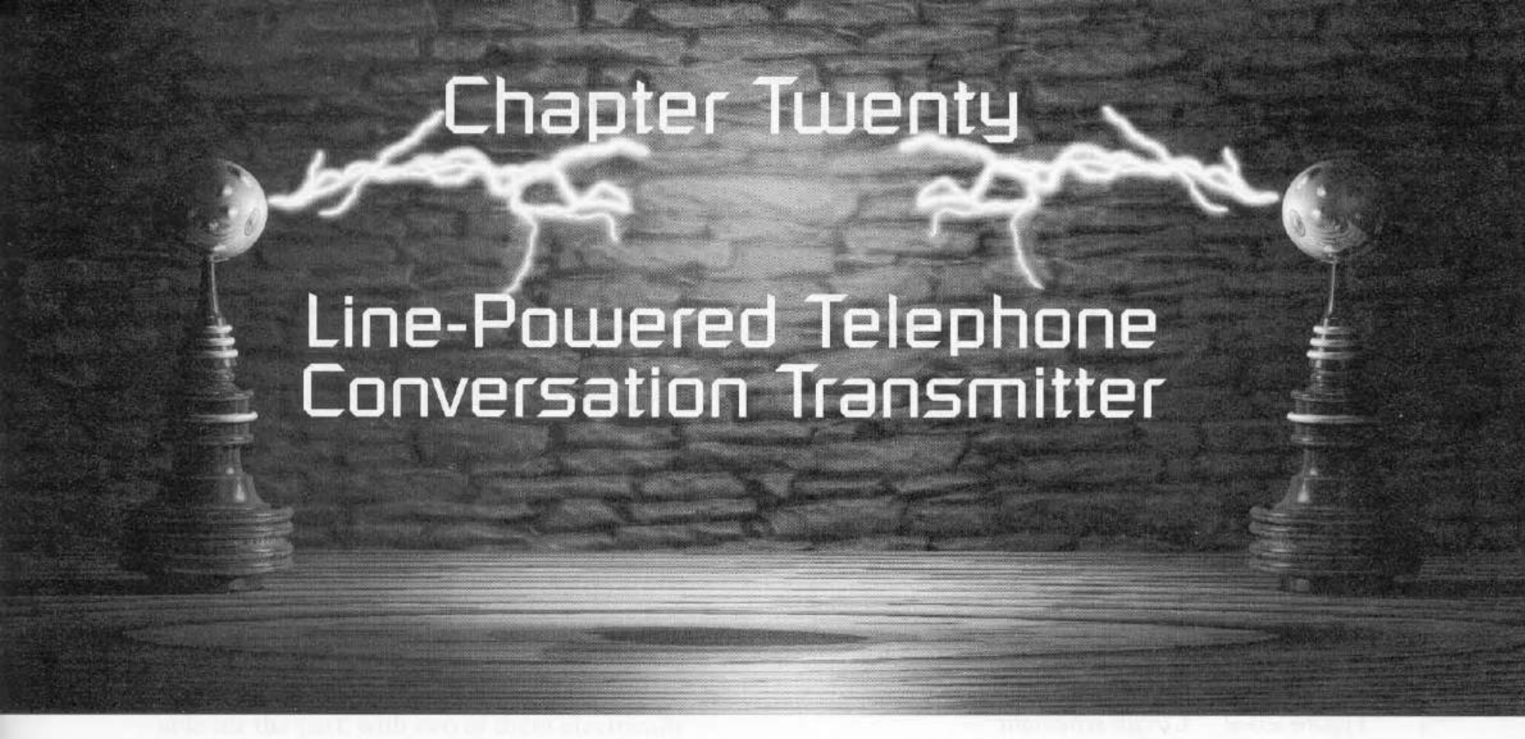

Chapter Twenty

Line-Powered Telephone Conversation Transmitter

This project shown in Figure 20-1 is similar to that described in Chapter 19, but this one can be permanently implanted in a telephone and can operate indefinitely, as it does not require batteries. The circuit as shown derives its power from the phone lines once the receiver is lifted off of the hook. Operating range is limited to several hundred feet, far less than the battery-powered version.

The unique features of this device is that it does not require a battery, and activation occurs only when the phone is off the hook. This eliminates transmitting a continuous signal, thus preventing potential interference and eliminates dead-tape listening time. You may also listen to a radio station while monitoring the telephone. When the phone is used, the radio station is replaced by the transmission of the unit. The device is shown assembled on a perforated circuit board and may be housed in a small plastic enclosure. Potting and sealing is at the discretion of the builder.

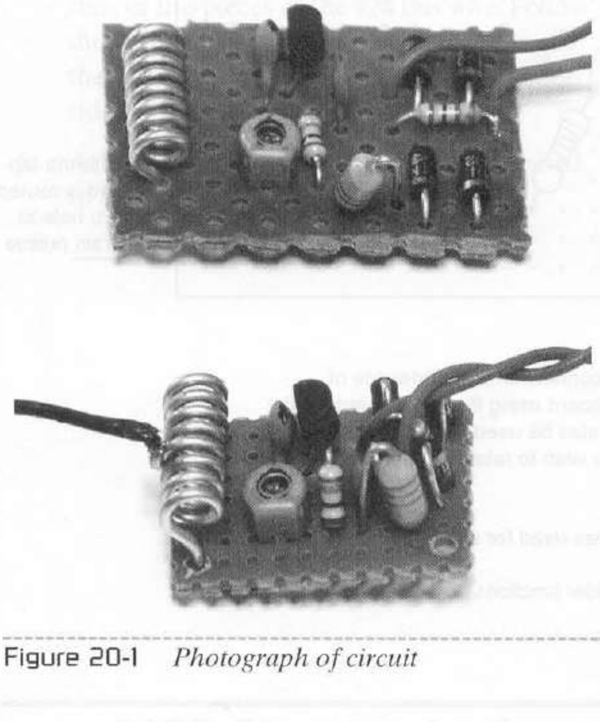

Figure 20-1 *Photograph of circuit*

Circuit Description

This chapter shows how to build a mini-powered FM transmitter that transmits both sides of a telephone conversation to a nearby FM radio. The circuit consists of a *radio-frequency* (RF) oscillator section combined with a telephone power conditioner that powers the oscillator (see Table 20-1). Transistor Q1 forms a relatively stable RF oscillator whose frequency is determined by the value of coil LI and tuning capacitor C3. The setting of C3 determines the desired operating frequency and is in the standard FM broadcast band with tuned circuit design favoring the high end up to 110 MHz. Capacitor C2 supplies the necessary feedback voltage necessary to sustain oscillation. Resistor R2 provides the necessary bias of the base emitter junction for proper operation, while

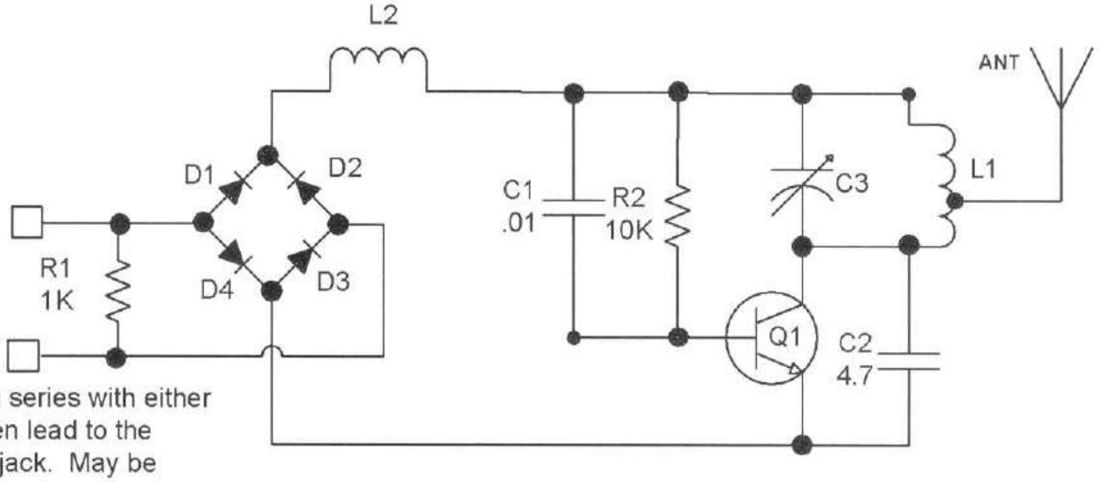

Connect in series with either red or green lead to the telephone jack. May be installed anywhere line is accessible. Connection will mean breaking one of the leads and connecting in place.

Figure 20-2 *Circuit schematic*

capacitor C1 bypasses any RF to ground. Choke L2 blocks the signal energy from going back to the phone lines yet allows DC power to Q1 (see Figure 20-2).

Connecting the circuit requires breaking one of the lines and connecting the input leads in series with the telephone. It is this method that leaches voltage from the series load resistor R1 when the telephone is in use, as current now flows across R1. The bridge diodes consisting of D1 through D4 polarize the audio voltages necessary to modulate the oscillator.

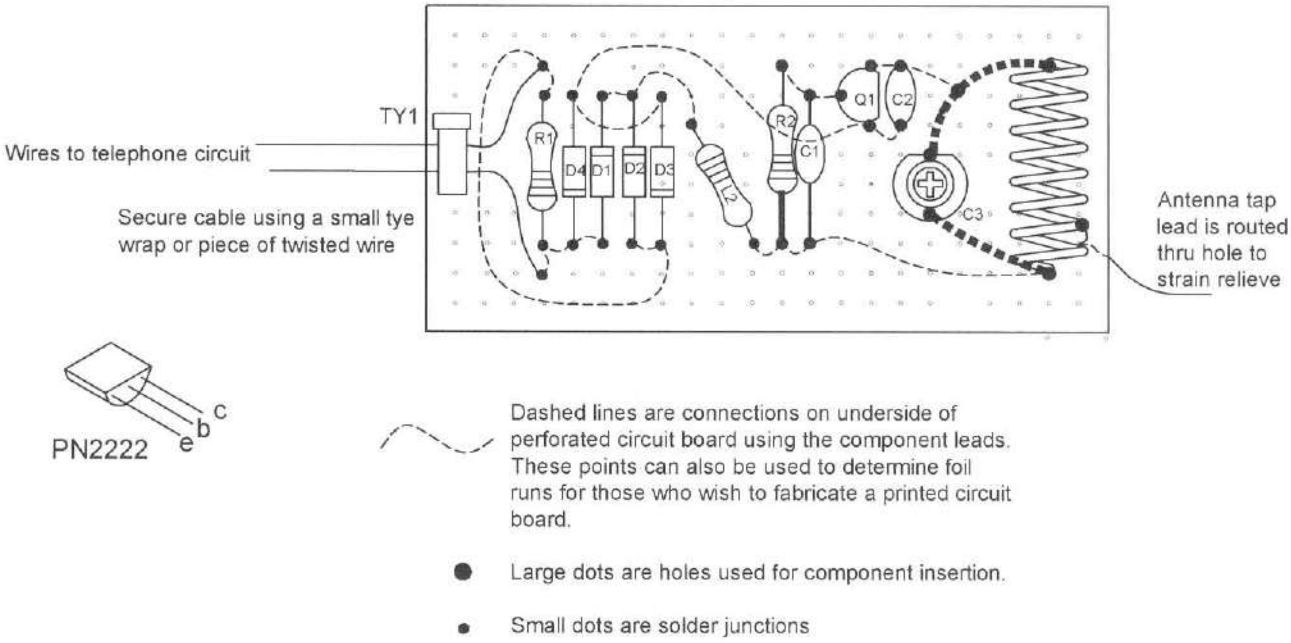

Secure cable using a small tye wrap or piece of twisted wire

Antenna tap lead is routed thru hole to strain relieve

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

Large dots are holes used for component insertion.

Small dots are solder junctions

Figure 20-3 *Assembly of circuit board*

Assembly of the Circuit Board

Assemble the circuit as shown in Figure 20-3:

1. Cut a piece of .1-inch grid perforated board to 2 × 1 inches.
2. Form L1 by tightly wrapping eight turns of #16 bus wire on a #8 wood screw. This produces an eight-turn coil with an inner diameter of approximately .135 inches and is about .625 inches in length. Insert it in the proper hole and solder as shown.
3. Insert trimmer capacitor C4 into the holes as shown. You may put this component on either side of the PC board. The choice is up to you and should be determined by the final packaging scheme you require. Three holes are available for the part, with two of them electrically the same point. This is necessary, as some trimmers may have three pins. Make sure that the common pins are connected to the same electrical points, with the odd pin to the other point.
4. If you are building from a perforated board, it is suggested that you insert the components starting in the lower left-hand corner. Pay attention to the polarity of the capacitors that have polarity signs and all the semiconductors. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate connection runs on the underside of the assembly board.
5. Attach the external leads that are connected to the telephone lines.
6. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
7. Tune an FM receiver to a fairly strong station at the high end of the band (108 MHz or higher). Turn up the volume and position it 25 to 50 feet away from the transmitter.
8. Connect the operating unit to the phone line with the phone on the hook. Remove the phone and adjust C3 until you hear the dial tone. Tuning the trimmer must be done with an insulating tuning wand. It is best to tune the unit by placing it on a nonconductive surface and securing it with a sponge or a piece of rubber.

   It may be difficult at first to spot the signal as this adjustment is very touchy. Also note that several spots in the adjustment may be erroneous and will be weak and unstable. The correct signal will be strong, stable, and clear and can be verified through distance-of-transmission testing.
9. Once the desired setting of C4 is found, it should be marked, with the frequency noted.
10. Conduct a final test by making an actual call and verifying the clear transmission of the FM radio to both parties.
11. When operation is verified, it may be a good idea to enclose the circuit board in a plastic box. This allows easier tuning and handling.

Notes

Please see notes in Chapter 19.

Table 20-1 Line-Powered Telephone Conversation Transmitter Parts List

Ref. #	*Description*	*DB Part #*
R1	1K 1/4-watt resistor (br-blk-red)	
R2	10K 1/4-watt resistor (br-blk-or)	
C1	.01-microfarad, 50-volt disc capacitor	
C2	4- to 6-picofarad small silver mica disk (4.7)	DB# 4.7P
C3	6- to 35-picofarad trimmer capacitor	DB# 635P
D1–4	100-volt diodes IN4002	
Q1	PN2222 NPN general-purpose transistor	
L1	Coil wound as described in assembly steps	DB# COIL8T
L2	3.6 microhenry inductor	DB# 3.6U
PBOARD	2 × 1 × .1-inch grid perforated board	
WR24BUSS	12 inches of #24 bus wire for wiring and antenna tap	
WR16BUSS	6 inches of #16 bus wire for making L1 coil	
WR20	24 inches of #20 vinyl hookup wire	

Chapter Twenty-One

Remote Wireless FM Repeater

Figure 21-1 shows a wireless FM transmitter designed to transmit the output of a tape recorder, VCR, DVD, or any device having composite audio and video outputs to a nearby radio or TV tuned to the unit's frequency. Assembly is built on a piece of circuit board and may be placed in any suitable plastic container along with a 9-volt battery.

Circuit Description

The parts for this project are listed in Table 21-1. Transistor Q1 forms a relatively stable *radio-frequency* (RF) oscillator whose frequency is determined by the value of coil L1 and tuning capacitor C4. The C4 setting determines the desired operating frequency and is in the standard FM broadcast band with tuned circuit design favoring the high end up to 110 MHz. Capacitor C2 supplies the necessary feedback voltage developed across emitter resistor R3 in the emitter circuit of Q1, sustaining an oscillating condition. Resistor R1 and R2 provide the necessary bias of the base emitter junction for proper operation, while capacitor C1 bypasses any RF to ground fed to the base circuit. Capacitor C9 provides an RF return path for the tank circuit of L1 and C4 while blocking the DC supply voltage fed to the collector of Q1 (see Figure 21-2). R4 is connected in series with DC-blocking capacitor C5 to the input cable and plug.

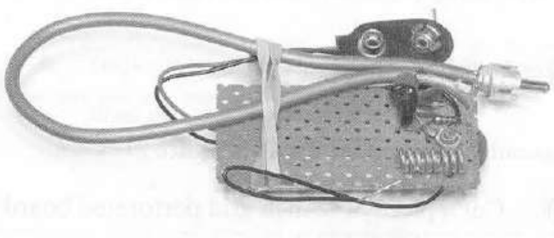

Figure 21-1 *The completed device*

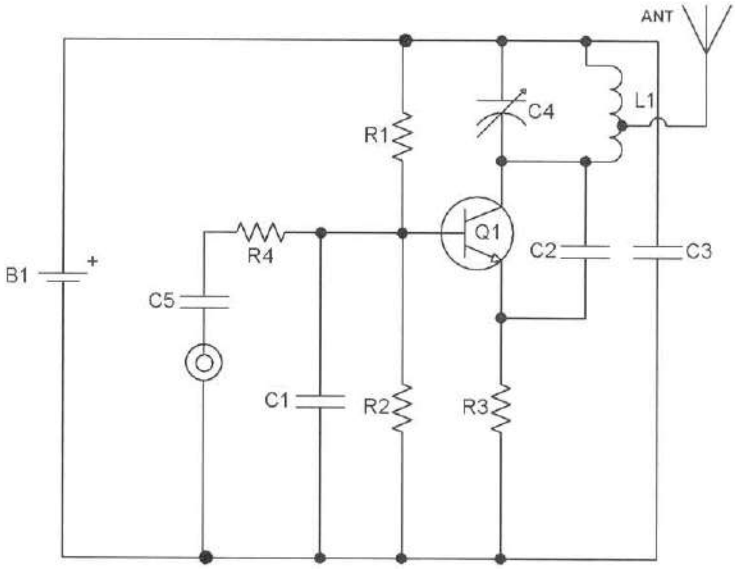

Figure 21-2 *Circuit schematic*

Assembly

Assemble the circuit as shown Figure 21-3.

1. Cut a piece of .1-inch grid perforated board to a size of 2 × 1 inches.
2. Form L1 by tightly wrapping eight turns of #16 bus wire on a #8 wood screw. This produces an eight-turn coil with an inner diameter of approximately .135 inches and a length of .625 inches. Insert it in the proper hole and solder as shown.
3. Insert trimmer capacitor C4 into the holes as shown. You may put this component on either side of the PC board. This choice is up to you and should be determined by the final packaging scheme you require. Three holes are available for the part with two of them electrically the same point. This is necessary, as some trimmers may have three pins. Make sure that the common pins are connected to the same electrical points with the odd pin to the other point.
4. If you are building from a perforated board, it is suggested that you insert components starting in the lower left-hand corner. Pay attention to the polarity of the capacitors that have polarity signs and all the semiconductors. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of the assembly board.
5. Attach the external leads of the battery clip (CL1) and shielded cable (WR10).
6. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
7. Tune an FM receiver to a fairly strong station at the high end of the band (108 MHz or higher). Turn up the volume and position the receiver about 25 to 50 feet from the transmitter circuit.
8. If a multimeter or 50-milliamp meter is available, connect it in series with a battery lead. This can be done by removing one of the clip fasteners and connecting the meter to one of the free contacts, as shown in Figure 21-2. The meter should read 5 to 7 milliamps. Pick up a short piece of bare wire and touch coil L1. Touch it a turn at a time, starting from the C3

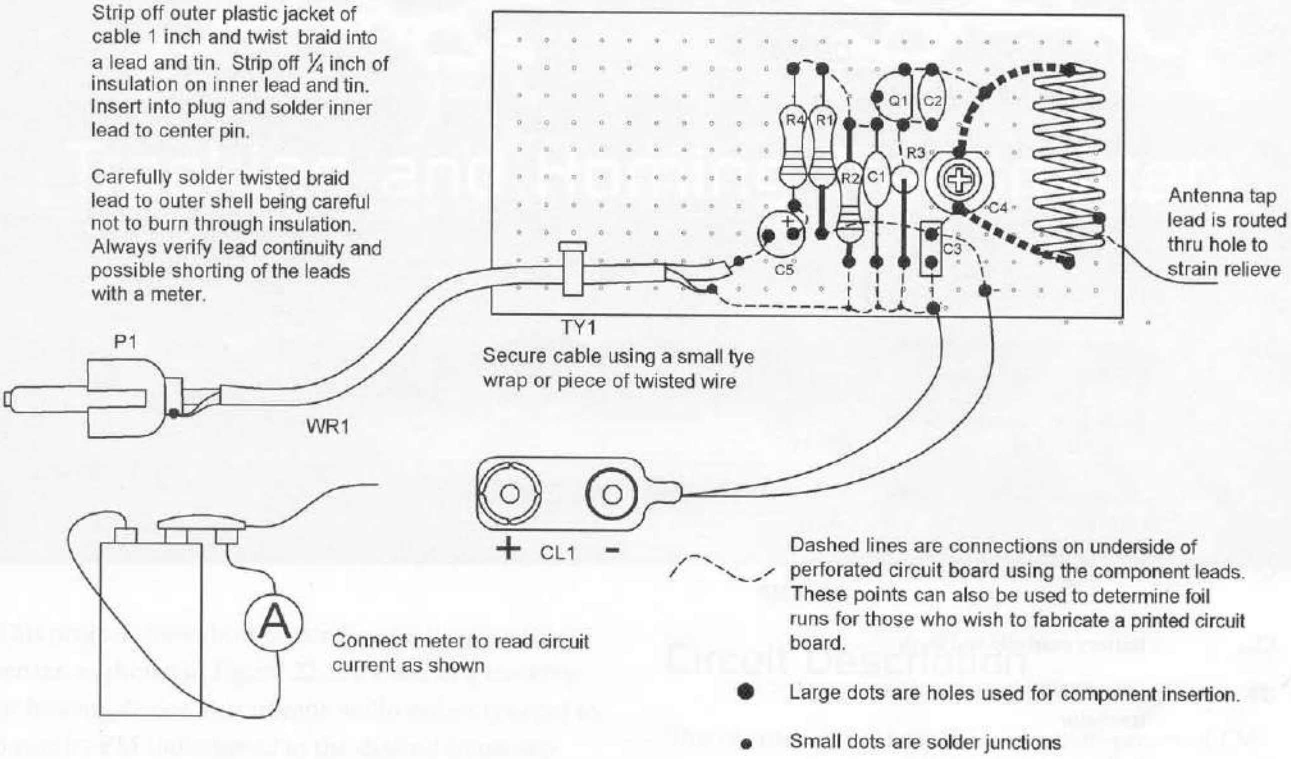

Strip off outer plastic jacket of cable 1 inch and twist braid into a lead and tin. Strip off ¼ inch of insulation on inner lead and tin. Insert into plug and solder inner lead to center pin.

Carefully solder twisted braid lead to outer shell being careful not to burn through insulation. Always verify lead continuity and possible shorting of the leads with a meter.

Antenna tap lead is routed thru hole to strain relieve

Secure cable using a small tye wrap or piece of twisted wire

Connect meter to read circuit current as shown

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

● Large dots are holes used for component insertion.

• Small dots are solder junctions

Figure 21-3 *Assembly of board*

end. Note that as you progress turn by turn away from C3 that the indicated meter current will drop or change.

9. Connect the cable into an audio source such as a recorder or another radio. Slowly rotate C4 with an insulated tuning wand until the station being received by the radio is at approximately 108 MHz. It may be difficult at first to spot the signal, as this adjustment is very touchy. Also note that several spots in the adjustment may be erroneous and will be weak and unstable. The correct signal will be strong, stable, and clear and can be verified by distance-of-transmission testing.
10. Once the desired setting of C4 *is* found, it should be marked with the frequency noted.
11. When the operation is verified, it may be desired to pot the assembly as shown. This allows easier tuning and handling.

Notes

Please refer to the notes in Chapter 19 on operation of wireless circuits.

Table 21-1 Remote Wireless FM Repeater Parts List

Ref. #	*Description*	*DB Part #*
R1	15K, 1/4-watt resistor (br-gr-or)	
R2	3.9K, 1/4-watt resistor (or-wh-red)	
R3	220-ohm, 1/4-watt resistor (red-red-br)	
R4	1K, 1/4-watt resistor (br-blk-red)	
C1, 5	Two .01-microfarad, 50-volt disc capacitors	
C2	Small 4- to 6-picofarad silver mica disk (4.7)	DB #4.7P
C3	.1-microfarad, 50-volt plastic capacitor	
C4	6- to 35-picofarad trimmer capacitor	DB #635P
CL1	Battery snap clip and leads	
Q1	PN2222 NPN general-purpose transistor	
L1	Coil wound as described in assembly steps	DB #Coil8T
PBOARD	2 × 1 × .1-inch grid perforated board	
WR24BUSS	12 inches of #24 bus wire for wiring and antenna tap	
WR16BUSS	6 inches of #16 bus wire for making L1 coil	
WR1	12 inches of shielded cable	
P1	RCA phono plug	

Chapter Twenty-Two

Tracking and Homing Transmitter

This project shows how to construct a beeping transmitter, as shown in Figure 22-1, for use as a tracking or homing device. It transmits audio pulses (beeps) to a nearby FM radio tuned to the desired frequency. The unit is intended to be used in transmitter hunt games or as a homing beacon for hunters or hikers. It is suggested that you mark a bearing heading on a small but good-quality FM radio and use it as a low-cost receiver for determining the direction of the signal pickup. The range of the device, when used as a homing device with a quarter-wave vertical placed on top of a metal roof, such as a vehicle, can be in excess of 3 miles when using a quality analog FM receiver.

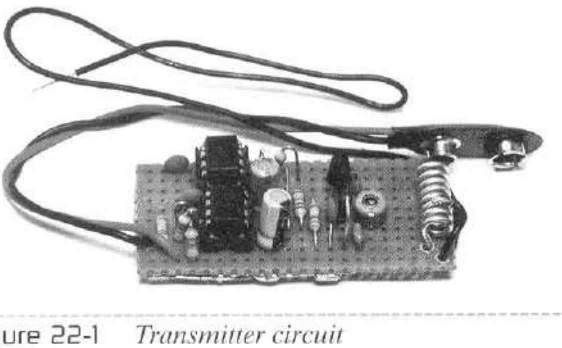

Figure 22-1 *Transmitter circuit*

Circuit Description

This chapter shows how to build a mini-powered FM transmitter that transmits audio pulses for tracking and homing applications (see Figure 22-2). The circuit consists of a *radio-frequency* (RF) oscillator section interfaced with a tone modulator and a pulse-switching section that controls the oscillator power. Transistor Q1 forms a relatively stable RF oscillator whose frequency is determined by the value of coil LI and tuning capacitor C4. The C4 setting determines the desired operating frequency and is in the standard FM broadcast band with the tuned circuit design favoring the high end up to 110 MHz. Capacitor C2 supplies the necessary feedback voltage developed across resistor R3 in the emitter circuit of Q1, sustaining an oscillating condition. Resistors R1 and R2 provide the necessary bias of the base emitter junction for proper operation, while capacitor C1 bypasses any RF to ground fed to the base circuit. Capacitor C3 provides an RF return path for the tank circuit of LI and C4 while blocking the DC supply.

You will note that the junction of the base bias resistors R1 and R2 is a feed point consisting of resistor R4. This point is where the beeping tone from the tone generator chip I2 is applied. Because of the nature of the oscillator frequency being subject to change by varying the base bias condition, a pulsed voltage superimposed at this point causes a corresponding frequency shift (FM) along with an amplitude modulated shift (AM) condition. It is this property that allows the circuit to be FM modulated, producing the tone frequency. The signal is clearly detected on any FM receiver when the receiver is properly tuned. The tone frequency is the pulse repetition rate of I1 wired as an astable pulse generator. This repetition rate is determined by timing resistor R5 and capacitor C6.

The ratio of beeping on to off time is controlled by chip I1, wired as a free-running astable device with duty cycle control. This function is determined by the ratio of resistors R7 and R8 and timing capacitor C8. The output of I1 controls the bias voltage of oscillator Q1 through resistor R1, turning the oscillator on and off (see Figure 22-2).

Assembly of the Circuit Board

Assemble the circuit as shown in Figure 22-3.

1. Cut a piece of .1-inch grid perforated board to a size of 2 × 1 inches.
2. Form L1 by tightly wrapping eight turns of #16 bus wire on a #8 wood screw. This produces an eight-turn coil with an inner diameter of .135 inches and a length of .625 inches. Insert it in the proper hole and solder as shown.
3. Insert trimmer capacitor C4 into the holes as shown. You may put this component on either side of the PC board. The choice is up to you and should be determined by the final packaging scheme you require. Three holes are available for the part, with two of them electrically the same point. This is necessary, as some trimmers may have three pins. Make sure that the common pins are connected to the same electrical points with the odd pin to the other point.
4. If you are building from a perforated board, it is suggested that you insert components, starting in the lower left-hand corner. Pay atten-

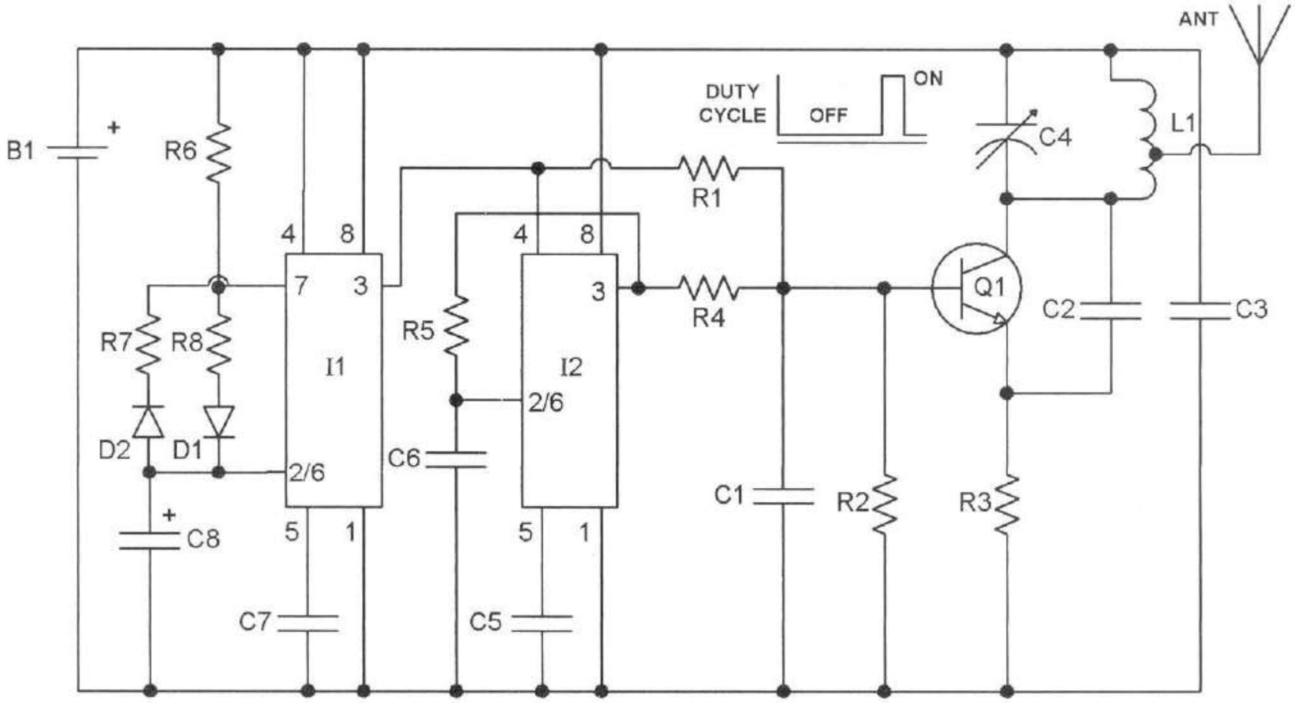

Figure 22-2 *Circuit schematic*

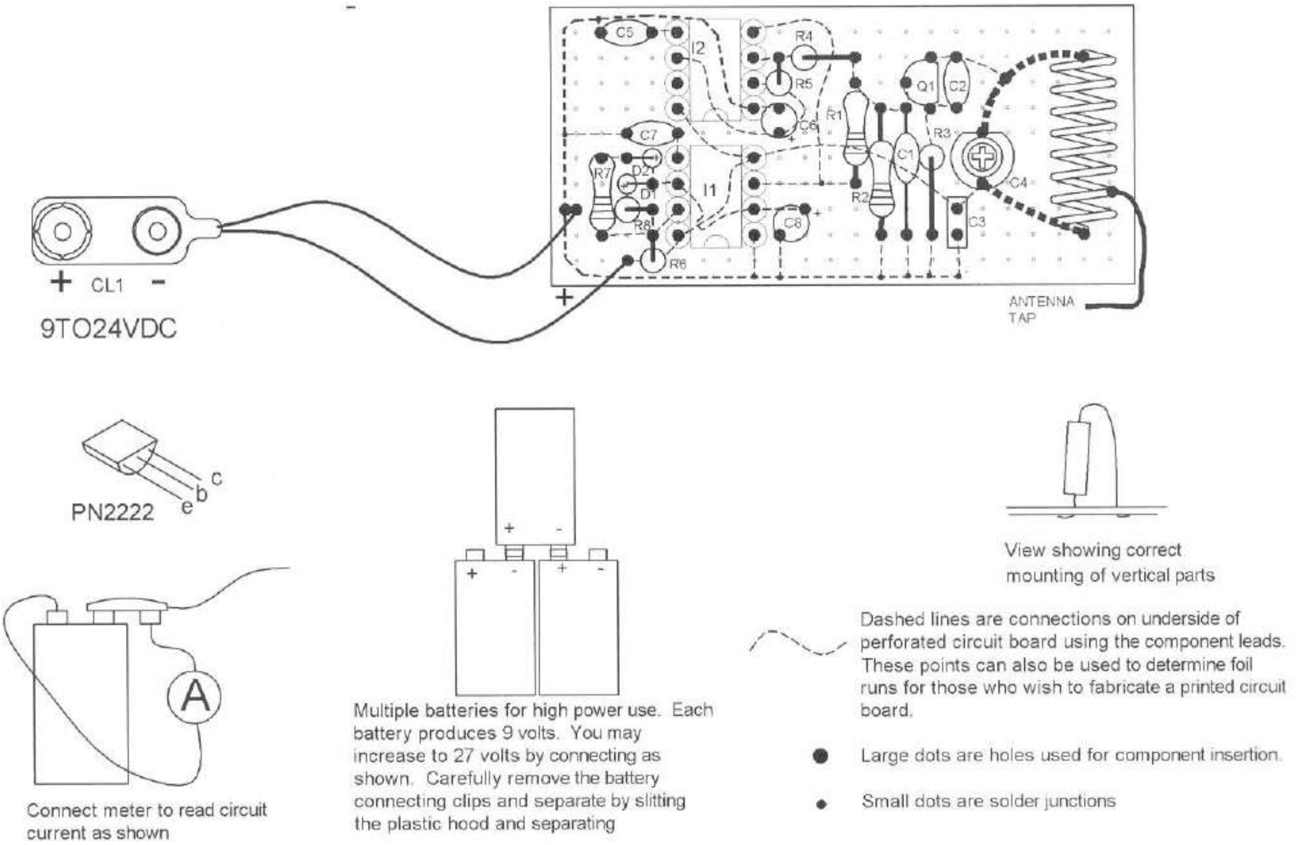

View showing correct mounting of vertical parts

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

Large dots are holes used for component insertion.

Small dots are solder junctions

Multiple batteries for high power use. Each battery produces 9 volts. You may increase to 27 volts by connecting as shown. Carefully remove the battery connecting clips and separate by slitting the plastic hood and separating

Connect meter to read circuit current as shown

Figure 22-3 *Assembly of the circuit board*

tion to the polarity of the capacitors with polarity signs and all the semiconductors. Route the leads of the components as shown and solder as you go, cutting away the unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of the assembly board.

5. Attach the external leads of battery clip CL1.
6. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
7. Tune an FM receiver to a fairly strong station at the high end of the band (108 MHz or higher). Turn up the volume and position the receiver 25 to 50 feet from the transmitter circuit.
8. If a multimeter or 50-milliamp meter is available, connect it in a series with a battery lead. This can be done by removing one of the clip fasteners and connect the meter to the free contacts, as shown in Figure 22-2. The meter should read a pulsing of 5 to 7 milliamps. Use a short piece of bare wire to touch the coil L1 a turn at a time, starting from the C3 end. Note that as you progress turn by turn away from C3 that the indicated meter current will drop or change.
9. Tune C4 with an insulated tuning tool until you hear the beeping tone. It may be difficult at first to spot the signal, as this adjustment is very touchy. Also note that several spots in the adjustment may be erroneous and will be weak and unstable. The correct signal will be strong, stable, and clear and can be verified by distance-of-transmission testing.

10. Once the desired setting of C4 is found, it should be marked with alignment marks, with the frequency noted.

11. Finally, test the unit using a good-quality FM analog receiver. Note to place an arrow on the radio once you determine the proper positioning to obtain the maximum signal strength. Do this step from a distance of at least several hundred feet.

12. When operation is verified, it may be desired to enclose the circuit board into a plastic box along with the battery. This allows easier tuning and handling.

Notes

Please see the notes in Chapter 19 on operating.

Table 22-1 Tracking and Homing Transmitter Parts List

Ref. #	*Description*	*DB Part #*
R1	Two 10K, 1/4-watt resistor (br-gr-or)	
R2	3.9K, 1/4-watt resistor (or-wh-red)	
R3	220-ohm, 1/4-watt resistor (red-red-br)	
R5	1K, 1/4-watt resistor (br-blk-red)	
R6	39K, 1/4-watt resistor (or-wh-or)	
R7	100K, 1/4-watt resistor (br-blk-yel)	
R8	22K, 1/4-watt resistor (red-red-or)	
C1, 5, 7	Three .01-microfarad, 50-volt disc capacitors	
C2	Small, 4- to 6-picofarad, silver mica disk (4.7)	DB# 4.7P
C3	.1-microfarad, 50-volt plastic capacitor	
C4	6- to 35-picofarad trimmer capacitor	DB# 635P
C6	.47-microfarad, 25-volt vertical electrolytic capacitor	
C8	4.7-microfarad, 25-volt vertical electrolytic capacitor	
D1, 2	Two 1N914 small-signal diodes	
I1, 2	Two 555 timers in a *dual inline package* (DIP)	
CL1	Battery snap clip and leads	
Q1	PN2222 NPN general-purpose transistor	
L1	Coil wound as described in assembly steps	DB# COIL8T
PBOARD	2 × 1 × .1-inch grid perforated board	
WR24BUSS	12 inches of #24 bus wire for wiring and antenna tap	
WR16BUSS	6 inches of #16 bus wire for making L1 coil	

Chapter Twenty-Three

Snooper Phone Room-Listening Device

Figure 23-1 shows a simple electronic device intended to be an intrusion detector and listening device for checking a home or office while away or on vacation. It may also be used to trigger other electrical devices. The circuitry does not require a tone encoder, so it does not provide selected activation by a particular caller. Thus, it limits the device from being used for illegal purposes and allows the sale of the kit or fully assembled unit without the requirement of special authorization.

This is an intermediate-level project requiring basic electronic skills. Expect to spend $25 to $40. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Table 23-1.

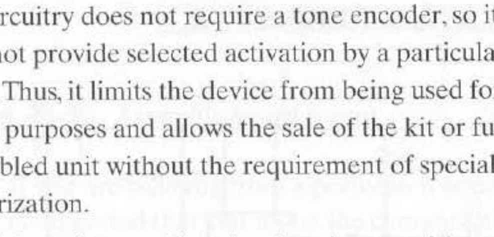

Figure 23-1 *The snooper phone*

Unlike the controversial Infinity transmitter, our system does not utilize decoder and encoder circuitry. Once connected to the phone lines, it will open a microphone to anyone who happens to dial the number. This feature obviously reduces the potential use of the device for illegal interception of oral communication. Even though the phone does not ring, it would not be long before the potential victim of an illegal surveillance would be aware of the setup. It is therefore quite obvious that the applications are strictly limited to use as a security device that checks for unauthorized intrusion or checks household appliances by placing a pickup device in certain favorable locations to overhear these desired sounds and noises. I am sure that anyone with this capability would breathe a sigh of relief to be able to dial his or her home or office phone while being away and hear the familiar sounds of appliances and other systems properly performing their duties and functions.

Circuit Operation

Your circuit consists of a high-gain amplifier (I1) fed into the telephone lines via transformer T1. The circuit is initiated by the action of a voltage transient pulse occurring across the phone line at the instant of a telephone circuit connection being made. This transient is a voltage change of about 48 to 5 VDC and usually occurs before the ring signal. It is this change that immediately forward-biases diode D1 and triggers pin 2 of timer I2, whose output pin 3 goes positive, turning on transistors Q1 and Q2. Timer I2 now remains in this state for a period depending on the values of timing resistor R12 and the selected value of the capacity determined by the activated time of switch S2. This four-position switch allows time selections of 10, 30, 100, and 1,000 seconds by the selection of the respective capacitors C12 through C15 (see Figure 23-2).

You will note that when Q2 is turned on by timer I2 that a simulated off-hook condition exists by the switching action connecting the 500-ohm winding of the transformer directly across the phone lines. Simultaneously, Q1 clamps the ground of the I1 amplifier to the negative return of battery Bl, enabling this amplifier section. System power is controlled by switch S1 being part of the pickup sensitivity control R1.

Electret microphone M1 picks up the sounds fed to operational amplifier I1 with the output coupled by capacitor C6 to the 1,500-ohm winding of Tl, and R1 controls the pickup sensitivity of the system. The system described will operate when any incoming call is received without the phone ever ringing.

Construction Steps

If you are a beginner, it is suggested to obtain our *GCAT1 General Construction Practices and Techniques*. This informative literature explains basic practices that are necessary in proper construction of electromechanical kits and is listed in Table 23-1.

1. Lay out and identify all the parts and pieces. Verify them with the parts list, and separate the resistors as they have a color code to determine their value. Colors are noted on the parts list.
2. Cut a piece of .1-inch grid perforated board to a size of 4.2 × 1.5. Locate and drill the holes as shown in Figure 23-3.

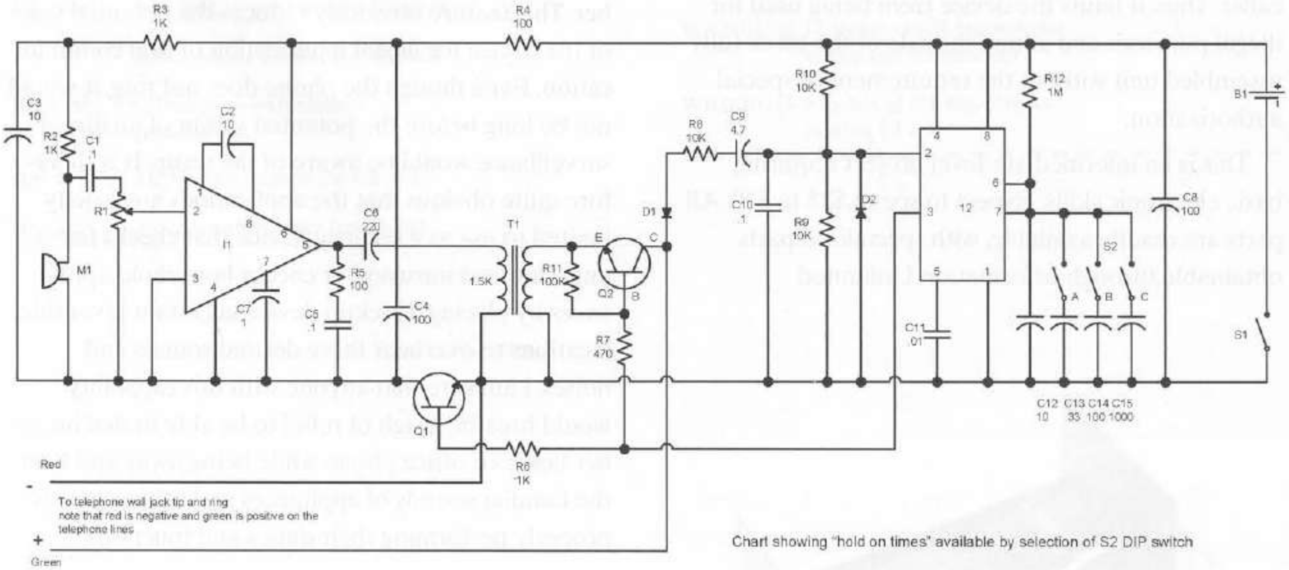

To telephone wall jack tip and ring
note that red is negative and green is positive on the telephone lines

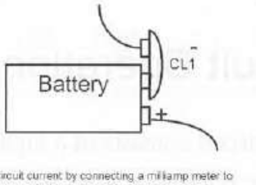

Measure circuit current by connecting a milliamp meter to the unconnected battery and battery clip contact

Chart showing "hold on times" available by selection of S2 DIP switch

S2A	S2B	S2C	*TIME
off	off	off	10 sec
on	off	off	43 sec
on	on	off	145 sec
on	on	on	1145 sec

*times may vary from that shown due to component variances

Figure 23-2 *Circuit schematic*

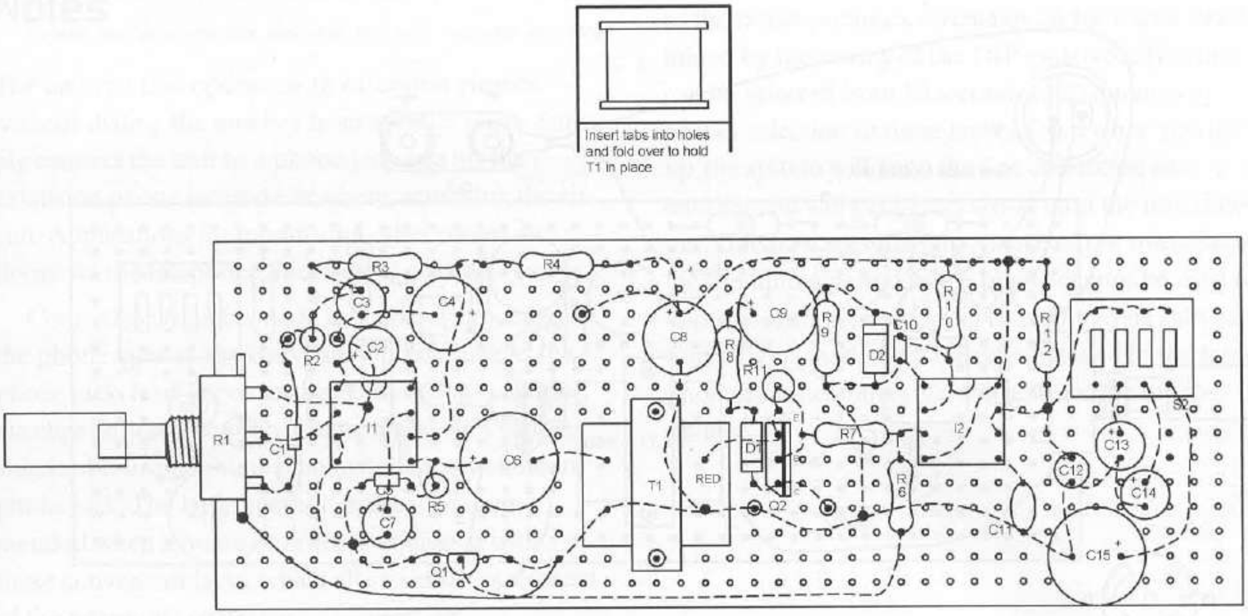
Insert tabs into holes and fold over to hold T1 in place

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

• Small dots are holes used for component insertion.

● Large dots are solder junctions

⊙ Circles with dots are holes that may require drilling for component mounting and points for external connecting wires indicating strain relief points. It is suggested to drill clearance holes for the leads and solder to points beneath board.

Figure 23-3 *Assembly board layout*

3. If you are building from a perforated board, it is suggested that you insert the components starting in the lower left-hand corner (see Figure 23-3). Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors.

   Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate connection runs on the underside of the assembly board.

4. Attach the external leads as shown in Figure 23-4. You may relieve any strain they might have by drilling holes to allow the insertion of the entire wire. Note that the short pieces of bus wire are soldered to R1 for connection to the circuit board.

5. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.

Final Assembly

Create a suitable housing from a piece of aluminum or purchase a ready-made aluminum box approximately 5 × 3 × 2 inches. Drill the holes as required for the microphone grommet GR1, gain control R1, and strain relief bushing for the phone line cable assembly RJ11. The metal box should be grounded to the circuit common line to reduce hum pickup. Note that a piece of Velcro sticky tape can be used to secure the battery or use your own ingenuity.

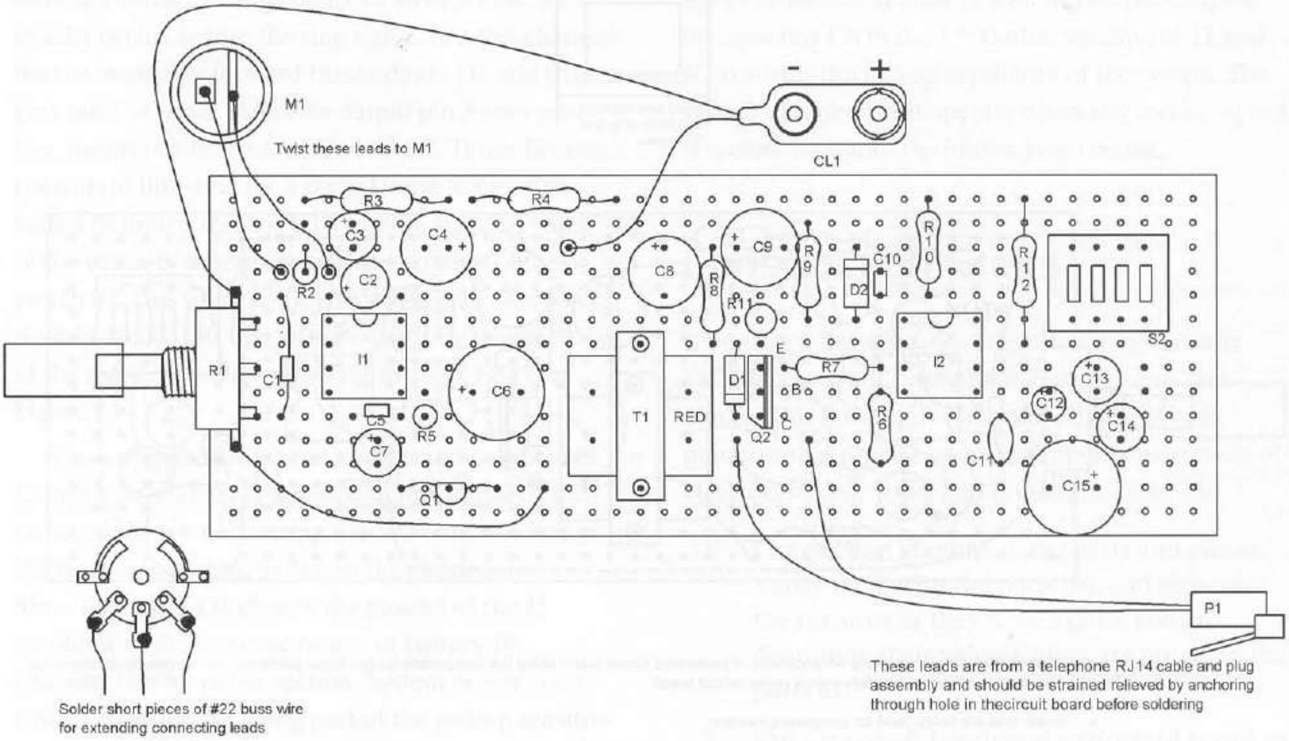

These leads are from a telephone RJ14 cable and plug assembly and should be strained relieved by anchoring through hole in thecircuit board before soldering

Solder short pieces of #22 buss wire for extending connecting leads

Figure 23-4 *External connection and parts identification*

Testing Steps

This step will require basic electronic laboratory equipment including a 60 MHz oscilloscope if trouble shooting the circuit is necessary.

1. Turn the unit off by rotating R1 full counter-clockwise, noting the click of the switch. Verify that all the DIP switch levers of S2 are down (off). Connect the battery as shown in series with a current meter and note zero current flow. Turn the unit on and note a current of 5 milliamps, which may momentarily jump to 30 milliamps as the unit times out.
2. Locate a phone jack so that you can monitor the DC voltage. Connect a meter across the tip and ring leads (green and red), and note the 40 to 50 volts with the phone on the hook. Also note the green lead is positive and the red is negative. This is the opposite of conventional wiring. Lift the phone and notice the 40 to 50 volts dropping down to 8 to 12 volts. This step verifies a properly working phone line.
3. Turn the unit off and connect it to the phone line. Rotate the R1 so that it just clicks on, and note the voltage remaining as or timing out to read the on-hook 40 to 50 volts. Preset R2 to midway.
4. Call the target phone using a cell or another telephone. The unit should connect at the beginning of the ring signal before the phone has a chance to ring. You should clearly hear sounds until the unit times out, at which time the device disconnects and restores the phone to normal operation. Adjust R1 for desired pickup sensitivity. The timeout time should be only about 10 seconds for the presetting of S2, as instructed in step 1. You may lengthen these times as shown in Figure 23-2.

Notes

The unit can also operate with extension phones without dialing the number from another place. Simply connect the unit to a phone jack and lift the extension phone located elsewhere, activating the circuit. Applications for monitoring other rooms and floors via the telephone lines now is possible.

Connections can be made in simple ways across the phone lines at any convenient point, such as the phone jack, lead-in box, or inside the telephone itself via clips or lugs. They can also be made via a simple, standard four-pin phone plug mated to any standard phone jack. The latter method is usually recommended when a home or office is equipped with these convenient jacks, which allow simple placement of the telephone or device.

When the phone activates, allowing you to listen to the target premises, it remains on for a time determined by the setting of the DIP switch S2. This time can be selected from 10 seconds to 20 minutes by proper selection of these levers. Even when you hang up, the system will keep the line connected so another call will get a busy signal until the unit times out. One now sees the importance of not overselecting the timing period as the phone cannot be used for either incoming or outgoing calls. Outgoing calls can normally be made with the unit installed to the target phone, but incoming calls will activate the microphone.

Table 23-1 Snooper Phone Room-Listening Device Parts List

Ref. #	*Description*	*DB Part #*
R1/S1	10K, 17-millimeter pot and switch	
R2, 3, 6	Four 1K, ¼-watt resistors (br-blk-red)	
R4, 5	One 100-ohm, ¼-watt resistor (br-blk-br)	
R7	470 ¼-watt resistor (yel-pur-br)	
R8, 9, 10	Three 10K, ¼-watt resistors (br-blk-or)	
R11	100K, ¼-watt resistor (br-blk-yel)	
R12	1M ¼-watt resistor (br-blk-gr)	
C1, 5, 10	Three .1-microfarad, 50-volt caps	
C2, 3, 12	Three 10-microfarad, 25-volt vertical electro radial leads	
C4,8,14	Three 100-microfarad, 25-volt vertical electro radial leads	
C6	.22-microfarad, 250-volt metalized polypropylene	
C7	.1-microfarad, 50-volt cap	
C9	1-microfarad, 25-volt vertical electro cap	
C13	33-microfarad, 25-volt vertical electro radial leads	
C11	.01-microfarad, 50-volt disk (103)	
C15	1,000-microfarad, 25-volt vertical electrolytic capacitor	
C6	.22-microfarad, 250-volt metallized polypropylene	
C9	1-microfarad, 25-volt vertical electro cap	
C7	.1-microfarad, 50-volt cap (INFO#VG22)	
D1, 2	Two 1N4007 1-kilovolt diodes	
I1	LM386 operational amplifier *dual inline package* (DIP) *integrated circuit* (IC)	
I2	555 timer DIP IC	
Q1	PN2222 NPN small-signal transistor	
Q2	D40D5 high-voltage transistor	
T1	1.5K, 500-ohm audio interstage transformer	
M1	Electret microphone	
S2	4PST DIP switch	
PBOARD	4.2 × 1.5 × .1-inch grid perforated board, sized per chapter	
GR1	½-inch rubber grommet	
BUSCLP	¼-inch Heyco clamp bushing	
CL1	9-volt battery clip with leads	
RJ11	RJ11 telephone plug and cable assembly	
WR20R	12 inches of #20 vinyl red wire for positive input	
WR20B	12 inches of #20 vinyl black wire for negative input	
ENC1	5 × 3 × 2-inch aluminum box and cover	

Chapter Twenty-Four

Long-Range FM Voice Transmitter

This project shows how to construct an electronic device that can transmit low-level sounds over a considerable distance (see Figure 24-1). It is intended for use as a security system when entry sounds are broadcasted over a standard FM radio, providing an early warning of an intrusion. It may be used to monitor children, swimming pools, sick and invalid people, and animals. The device can also be used for nature listening, which is made possible by the wide frequency response of the audio section, allowing good-quality reproducibility of rare bird calls, for example. An excellent application, when one is living in hostile areas, is to install the unit in one's home. This allows monitoring of the premises before possibly being surprised by a burglar.

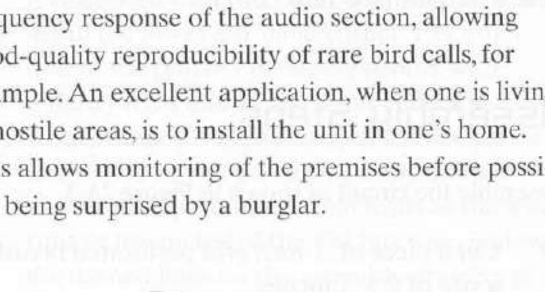

Figure 24-1 *Long-range FM voice transmitter*

The assembled transmitter is small enough to be placed at the foci of a sonic reflector, creating a highly directional wireless mike for targeting particular areas of interest. (The Audubon Society purchased 25 of these systems for monitoring bird and animal calls in mosquito-infested swamp areas.)

Circuit Description

This circuit shows how to build a super-sensitive, mini-powered FM transmitter consisting of a *radio frequency* (RF) oscillator section interfaced with a high-sensitivity, wide pass-band audio amplifier and capacitance mike with a built-in *field effect transistor* (FET) that modulates the base of the RF oscillator transistor (see Figure 24-2). Transistor Q1 forms a relatively stable RF oscillator whose frequency is determined by the value of coil L1 and tuning capacitor C4. The setting of C4 determines the desired operating frequency and is in the standard FM broadcast band with the tuned circuit design favoring the high end up to 110 MHz. Capacitor C2 supplies the necessary feedback voltage developed across resistor R3 in the emitter circuit of Q1, sustaining an oscillating condition. Resistors R1 and R2 provide the necessary bias of the base emitter junction for proper operation, while capacitor C1 bypasses any RF to ground fed to the base circuit. Capacitor C3 provides an RF return path for the tank circuit of L1 and C4 while

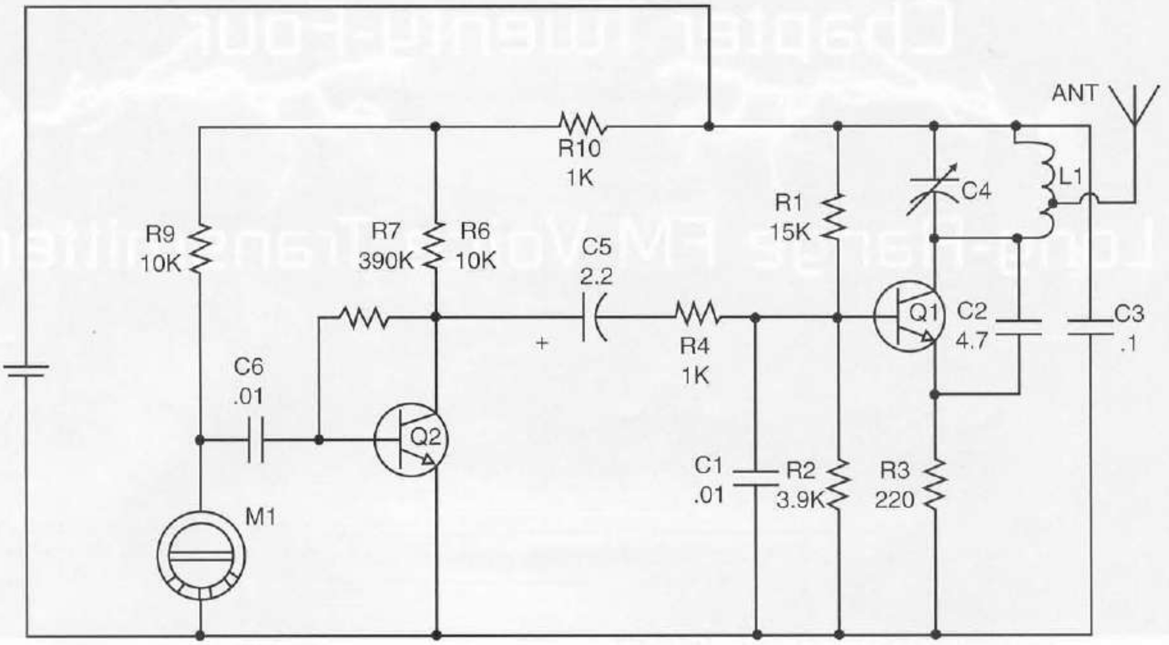

Note that the microphone M1 has polarity. You can identify the ground pad by the connection tabs to the outer shell.

Figure 24-2 *Circuit schematic*

blocking the DC supply voltage fed to the collector of Q1.

The audio section utilizes a high-sensitivity capacitance mike (M1) and built-in FET transistor (field effect) and will clearly pick up all low-level sounds in the speech audio spectrum. The speech voltage developed across resistor R9 by M1 is capacitively coupled by C6 to the base of audio amplifier transistor Q2. You will note the base of oscillator transistor Q1 is biased by resistors R1 and R2. A signal voltage now developed across resistor R6 is capacity coupled through nonpolarized capacitor C5 to the base of Q1 through resistor R4. The gain of Q2 is controlled by the ratio of R7 and R6. The operating point is set to allow full excursion of the collector to the amplified signal. The amplified speech signal now causes FM and AM modulation of the oscillator circuit by slight shifting of the operating point of the base section. Resistor R10 decouples the oscillator and audio circuits and is necessary to prevent feedback and other undesirable effects.

When properly assembled, the circuit should produce crystal-clear quality when the receiver is properly tuned to the unit. Note that a shunt capacitor may be connected across the base lead of Q1 to reduce sensitivity. The circuit with the components listed operates best in the upper FM band. This is a clear spot without interference from FM radio stations. However, satisfactory performance is obtained above 110 MHz for limited range use. Use caution as this is the aircraft band and should not be used if near an airport (see Table 24-1).

Assembly Steps

Assemble the circuit as shown in Figure 24-3.

1. Cut a piece of .1-inch grid perforated board to a size of 2 × 1 inches.
2. Form L1 by tightly wrapping eight turns of #16 bus wire on a #8 wood screw. This produces an eight-turn coil with an inner diameter of approximately .135 inches and a length of about .625 inches. Insert it in the proper hole and solder as shown.
3. Insert trimmer capacitor C4 into the holes as shown. You may put this component on either side of the PC board. This choice is up to you and should be determined by the final packaging scheme you require. Three holes are available for the part with two of them electrically the same point. This is necessary as some

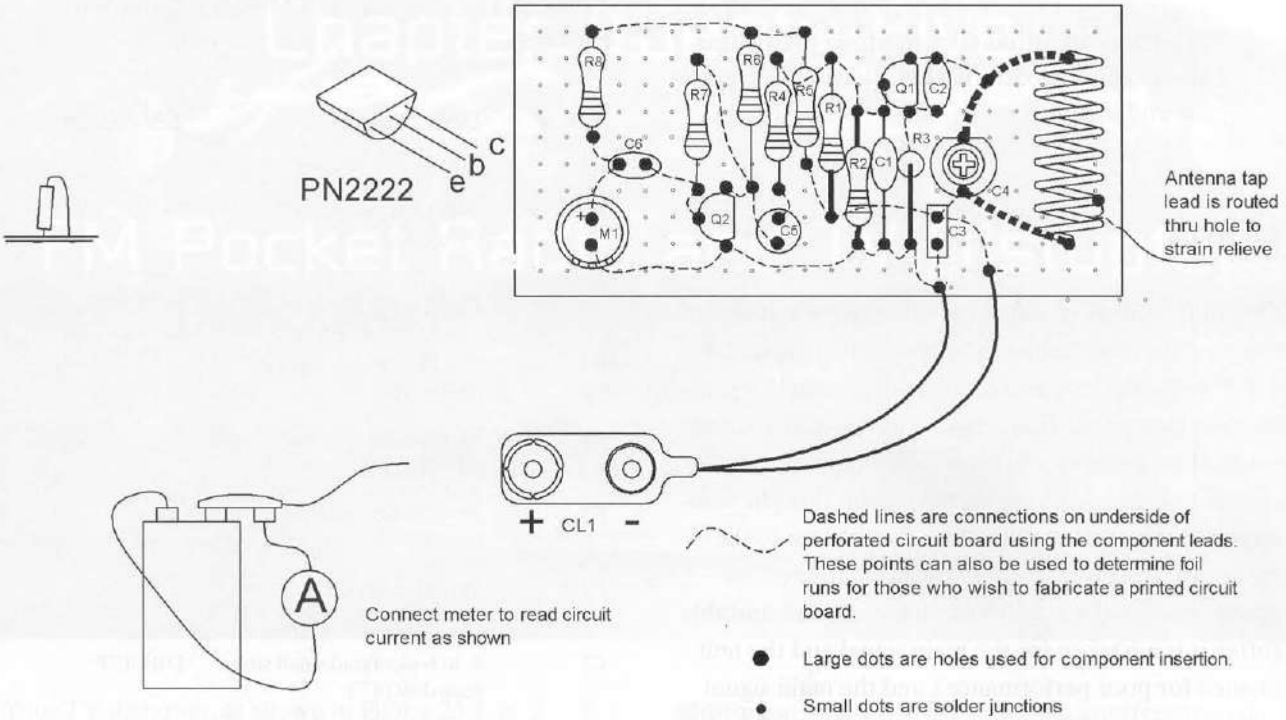

Connect meter to read circuit current as shown

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

● Large dots are holes used for component insertion.

• Small dots are solder junctions

Figure 24-3 *Assembly board*

trimmers may have three pins. Make sure that the common pins are connected to the same electrical points with the odd pin to the other point.

4. If you are building from a perforated board, it is suggested that you insert components starting in the lower left-hand corner. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of the assembly board.
5. Attach the external leads of the battery clip (CL1).
6. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
7. Tune an FM receiver to a fairly strong station at the high end of the band (108 MHz or higher). Turn up the volume and position the receiver 25 to 50 feet from the circuit.
8. If a multimeter or 50-milliamp meter is available, connect it in series with a battery lead. This can be done by removing one of the fasteners of the clip and connecting the meter to the free contacts, as shown in Figure 24-3. The meter should read 5 to 7 milliamps. Use a short piece of bare wire to touch the coil L1 a turn at a time, starting from the C3 end. Note that as you progress turn by turn away from C3 that the indicated meter current will drop or change.
9. If the device performs as follows with the battery connected, slowly rotate C4 with an insulated tuning wand until the station being received by the radio at approximately 108 MHz breaks out in audio feedback or is blocked out. It may be difficult at first to spot the signal as this adjustment is very touchy. Also note that several spots in the adjustment may be erroneous and will be weak and unstable. The correct signal will be strong, stable, and clear and can be verified by distance-of-transmission testing.
10. Once the desired setting of C4 is found, it should be marked with the frequency noted.

11. When operation is verified, it may be desired to enclose the circuit board into a plastic box along with the battery. This allows easier tuning and handling.

Notes

One of the things to watch for when using a device of this type is proper tuning. The adjustment capacitor, C4, is quite sensitive and requires only a slight movement to change the frequency, so always use a tuning wand. It is very easy when one is not familiar with this unit to tune to an erroneous signal. This phenomenon is likely to occur when the unit is close to the monitoring receiver. As previously mentioned, an erroneous signal will be weak, distorted, and unstable (often it is mistaken for the main signal and the unit blamed for poor performance), and the main signal will be strong, stable, and undistorted, if modulated. Several experiments in tuning the unit should be done before attempting to use it for the desired application.

Also, whenever possible, the unit should be used around 108 to 109 MHz, which is at the boundary of the aircraft band and upper FM broadcast band. When the approximate desired spot is found, final touchup tuning should be done at the receiver end for clarity. In most areas, these "upper" frequencies are clear and allow uninterrupted use in contrast to the lower ones in the FM band where a slight change in frequency from a clear spot results in the unit being drowned out by a strong broadcast station. Do not go above 108 MHz if near an airport or air traffic lane.

Most available FM radios can easily be detuned slightly to shift the dial readings down where 108 is 109. This is accomplished by carefully adjusting the "oscillator," the padding trimmer located on the main tuning capacitor, and "walking" a known station down the necessary megahertz or two. An antenna-peaking trimmer should now be adjusted for maximum signal at the high end.

Optimum performance will require a good-quality receiver with an analog slide-rule-type tuning dial. Digital tuned receivers will not work that well.

Table 24-1 Long-Range FM Voice Transmitter Part List

Ref. #	*Description*	*DB Part #*
R1	15K, 1/4-watt resistor (br-gr-or)	
R2	3.9K, 1/4-watt resistor (or-wh-red)	
R3	220-ohm, 1/4-watt resistor (red-red-br)	
R4, 5	Two 1K, 1/4-watt resistors (br-blk-red)	
R6, 9	Two 10K, 1/4-watt resistors (br-blk-or)	
R7	390K, 1/4-watt resistor (or-wh-yel)	
C1, 6	Two .01-microfarad, 50-volt disc capacitors	
C2	4- to 6-picofarad small silver mica disk (4.7)	DB# 4.7P
C3	.1-microfarad, 50-volt plastic capacitor	
C4	6- to 35-picofarad trimmer capacitor	DB# 635P
C5	2.2-microfarad, 25-volt vertical nonpolarized electrolytic capacitor	
CL1	Battery snap clip and leads	
Q1, 2	Two PN2222 NPN general-purpose transistors	
L1	Coil wound as described in assembly steps	DB# COIL8T
PBOARD	2 × 1 × .1-inch grid perforated board	
WR24BUSS	12 inches of #24 bus wire for wiring and antenna tap	
WR16BUSS	6 inches of #16 bus wire for making L1 coil	
M1	Special FET-bypassed microphone	DB# FETMIKE
PCXFM1		DB# PCXFM

Chapter Twenty-Five

FM Pocket Radio and TV Disrupter

Your TV disrupter, as shown in Figure 25-1, is designed and built on a piece of perforated circuit board that can fit inside a small plastic case. It operates using two standard, 9-volt transistor batteries in series. A unique pulsing circuit causes carrier deviation at 1,000 Hz, which creates a highly effective interfering device. The unit is tuned to a certain channel or frequency via a miniature tuning capacitor and then is controlled by a slight action of hand movement for further control.

This device is intended as a barroom joke and must not be abused. No attempts should be made to add an antenna because the range can be excessive and become a highly illegal device. When used as a harmless gag, such as in a bar, you can drive the bartender crazy because just as he is about to make an adjustment to a TV to get rid of the interference, you simply detune the device to make an adjustment to get rid of the interference and take the disrupter off the channel. When he thinks the problem is solved and continues his other duties, you hit it again. You can only begin to imagine the enjoyment (as a gag) of creating such havoc and being completely unnoticed.

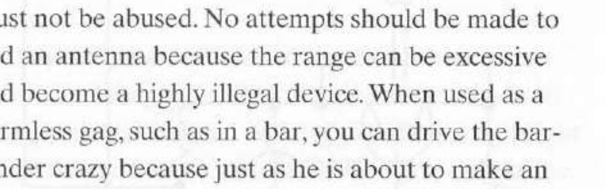

Figure 25-1 *Assembly board*

Also, any FM radio station can be interrupted with a 1,000 Hz tone that is many times louder than the normal received station. Obviously, a fun-loving person can use his or her imagination and come up with many harmless uses for the device.

It must be stated that using a device of this nature could be considered illegal, and discretion should be used at all times because a good gag can become a hateful nuisance. Also note that when cable TV is used, it may be necessary to get as close to the set as possible for maximum effect.

The following plans show how a miniature, simple, inexpensive electronic circuit can disrupt TV and radio communications in a very limited area. The unit can also be used to blank out those tiresome commercials.

Circuit Theory

The I1 timer generates positive pulses that occur across R4. The frequency of those pulses is a function of R1 and C1 and consequently establishes the pulse repetition rate of the device. This frequency modulates the base of Q1, which is a variable oscillator whose frequency is determined by C6 (see Figure 25-2). See Table 25-1 for the parts list.

Assembly Steps

Lay out and identify the parts with the bill of materials. Insert the components as shown in Figure 25-3, starting from one end of the perforated circuit board, and follow the locations using the holes as guides. Use the leads of the actual components as the connection runs, which are indicated by the dashed lines. It is a good idea to trial-fit the parts before actual soldering. Always avoid bare wire bridges and globby solder joints, and check for cold or loose solder joints. Always pay attention to the capacitors with polarity signs and all the semiconductors.

To adjust for frequency, set the TV to channel 2 and turn C6 for maximum capacitance (for a preliminary frequency, when adjusting, it is advisable to use a lower voltage due to the possibility of interfering with other than TV or radio stations such as police or aircraft bands). Adjust L1 by compressing turns until the interference is noted on channel 2. Turn to other channels, turn on C4, and note the complete coverage. Channels above 6 are covered by the harmonic output of the device.

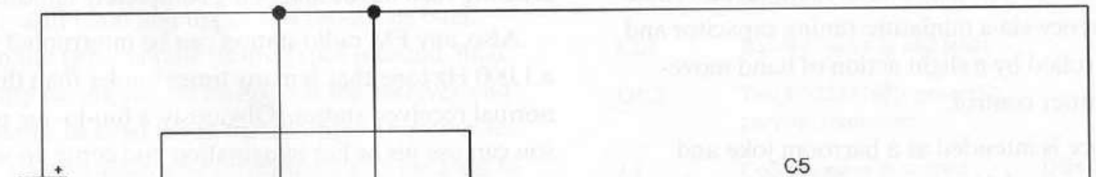

Figure 25-2 *Circuit schematic*

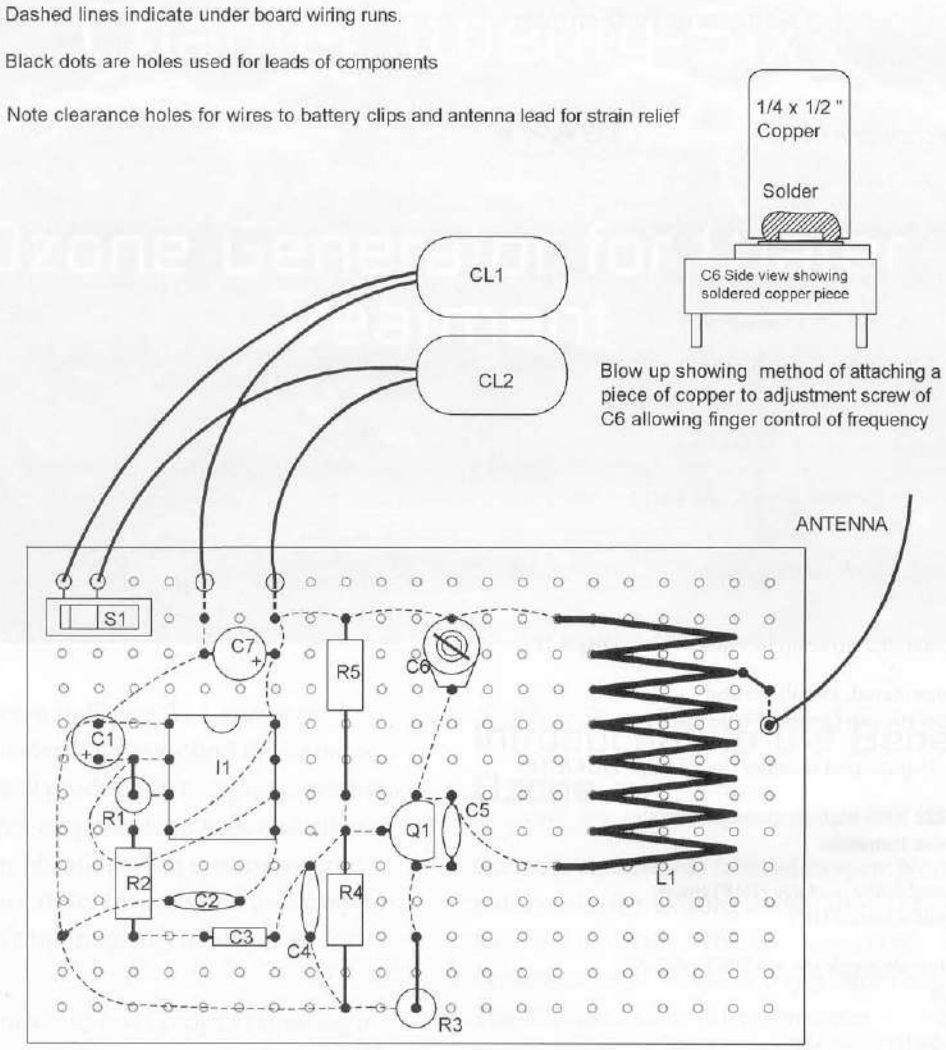
Dashed lines indicate under board wiring runs.
Black dots are holes used for leads of components
Note clearance holes for wires to battery clips and antenna lead for strain relief
1/4 x 1/2 "
Copper
Solder
C6 Side view showing soldered copper piece
CL1
CL2
Blow up showing method of attaching a piece of copper to adjustment screw of C6 allowing finger control of frequency
ANTENNA
S1
C7
+
R5
C6
C1
+
I1
R1
C5
Q1
R2
R4
C2
C3
C4
L1
R3

Figure 25-3 *Assembly board*

Several points of interest should be mentioned here, one being the ability to tune the unit on or off of a particular channel with just hand capacity (slight movement near L1); that is, the ability to detune the circuit with your hand creates a virtual capacitance. The other is the ability to also cover the FM radio band. This is accomplished by decompressing the coil (spreading the turns apart) and eventually adjusting L1 and C4 for the necessary coverage. It should be noted that all channels can be covered along with the FM band when properly adjusted. It may be necessary to remove a turn or two from L1 to a range on the FM radio stations.

Cable and satellite TV will often require the unit to be very close to cause an effect. There is no control over this, as the unit is designed to interfere with the actual frequency of the station being received, as was the case before cable and satellite services.

Table 25-1 FM Pocket Radio and TV Disrupter Parts List

Ref. #	*Description*	*DB Part #*
R1, 2	Two 1K, ¼-watt resistors (br-blk-red)	
R3	220-ohm, ¼-watt resistor (red-red-br)	
R3	100-ohm, ¼-watt resistor (br-blk-br)	
R4	3.9K, ¼-watt resistor (or-wh-red)	
R5	15K, ¼-watt resistor (br-gr-or)	
C1	1-microfarad, 35-volt vertical electrolytic cap (green or blue can)	
C2, 4	Two .01-microfarad, 25-volt disk caps (103)	
C3	.1-meter, 50-volt ceramic cap (blue tablet .1)	
C5	6-picofarad, zero temp (4.7 disk)	DB# 4.7P
C7	.47-microfarad, 35-volt vertical electrolytic cap (green or blue can)	
C6	6- to 35-picofarad trimmer tuning Cap	DB# 635P
Q1	PN2222 NPN high-frequency, general-purpose transistor	
I1	555 *dual inline package* (DIP) timer *integrated circuit* (IC)	
S2	*Single pole, single throw* (SPST) slider switch	
CL1	Two-battery snap clip	
WR1	9 inches of #16 buss wire; wind seven turns on a pencil-sized object	
WR2	12 inches of #24 vinyl hookup wire	
PB1	2.5 × 1.5 × .1-inch grid perforated circuit board	
L1	Coil as shown	DB# COIL8T

Chapter Twenty-Six

Ozone Generator for Water Treatment

This project, as shown in Figure 26-1, can very effectively purify water for use around the home or farm. The system will produce up to 5 grams per hour sufficient for swimming pools, laundries, and other medium-volume applications. The system covered here is shown in five different sizes and uses easy-to-construct modules requiring only a flow of air for operation.

This is an intermediate-level project requiring a basic electrical hookup. Expect to spend \$50 to \$250 depending on the size of the system chosen and the availability of the required air source. All parts are readily available, with specialized parts obtainable through Information Unlimited at www.amazing1.com, and they are listed in the parts list at the end of the chapter.

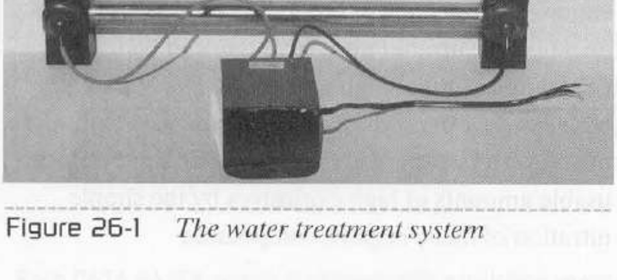

Figure 26-1 *The water treatment system*

Introduction to the Benefits of Ozone

Ozone is an unstable form of oxygen. Normal oxygen is diatomic (0_2), existing as two atoms of oxygen that form the molecule. Whereas ozone is tri-atomic (0_3), existing as three atoms of oxygen for the molecule. The tri-atomic form of oxygen is very unstable, wanting to lose the third oxygen atom and combine with whatever atom it can (oxidization). This property makes it the most active oxidizer known, with the exception of the very hazardous fluorine gas.

Ozone at normal pressures is colorless and produces a pleasing, fresh-air odor often like that after an old-fashioned thunderstorm. Under pressure, it becomes a bluish gas.

Ozone is also a very powerful bactericide. It is not affected by pH, as is chlorine, thus making it an excellent candidate for pools, spas, laundries, and general water treatment applications. It is many times more soluble in water, further enhancing its purifying effect. Ozone will combine with diatomic nitrogen N^2, forming nitrous oxide, 2NO. This oxide quickly combines with water, forming nitric acid, HNO_3. This is often a very undesirable effect when used with straight air. Pure oxygen greatly minimizes this effect and is often required in many applications. However, those requiring a supply of concentrated nitric acid for nitration may wish to consider ozone and air with a condensing apparatus to obtain this useful acid. A condensing apparatus takes vapor from the ozone gas along with water and nitrogen to form a liquid by contacting them with cooler surfaces. The liquid (now nitric acid) flows down into a collection container by gravity. In the same way the old timers made moonshine, this uses a simple and basic chemical reaction where $2O^3 + 2N^2 + 2H_2O$ yields $2H(NO^3) + H_2O + 2NO$.

Ozone for Water Applications

It is estimated that 20 percent of all ground water is contaminated by pesticides, benzene, and phenol derivatives along with other undesirable organic substances. Ozone can oxidize many of these compounds, along with deactivating many viruses and harmful bacteria. Ozone will also oxidize certain inorganic compounds, such as iron and manganese, making them more easily removable by filtration.

Chlorine and bromine are often the choice of disinfectant for swimming pools and spas. The effect of these halogens is often dictated by pH, temperature, and agitation. Extreme heat and agitation can produce chloroform, a very toxic carcinogenic. The *Environmental Protection Agency* (EPA) is already taking a look at these chemicals for this use.

Ozone-treated water will destroy fungus, mold, and many pathogens found in water when used for washing fruits and vegetables in packing lines. When discharged, ozone causes little change to the beneficial bacteria in sewage treatment facilities.

Additionally, freshly caught fish will last longer when washed with ozone-treated water, and ozone's oxidizing action can eliminate odors from stored cheese. Egg storage time is increased, and wine can be aged faster. The removal of odors produced by the bleaching of beeswax, starch, flour, straw, bones, and feathers are all aided by ozone treatment. Also, the grease and wax on cotton and wool fibers can be decomposed by ozone, and gray mold on the surface of fruits and vegetables can be controlled by ozone.

Ways to Generate Ozone

Ozone can be produced by an electrical discharge or by a high-frequency electromagnetic wave. A high frequency requires the wave to be in the ultraviolet spectrum where Planck's energy formula, $W = hc/v$, starts to become effective. This is where energy in a wave packet in ergs is equal to Planck's constant times the speed of light in centimeters divided by the wavelength in centimeters. It is this energy that causes the stable O_2 to break up and recombine with other O_2 to form unstable O_3. Germicidal lamps operating at 253.7 nanometers can produce ozone.

The method presented here utilizes the ozone-producing properties of an electrical discharge. We have all, at one time, smelled the by-products of ozone. After a thunderstorm, it can be detected, as well as on certain days where electrical activity is spawning a storm. A sparking electric discharge such as brushes on a motor will create ozone also.

Our method uses a metal tube with a conductor running down the center. The conductor and the tube are insulated by one another and must support the high frequency and high voltage necessary to create a corona without breaking down into an arc. The ozone produced in the tubular cavity must be made to flow using moving air for cooling and replenishing air to be converted.

This method, while producing usable ozone, has several disadvantages. First, normal air contains nitrogen that likes to combine with the ozone to produce nitrous oxides. Second, air contains water in the form of moisture. Without getting into basic chemistry, we know that water plus nitrous oxide forms nitric acid, which is corrosive and undesirable for air purification applications. However, a very effective way of making this acid is to allow the oxides to combine with water vapor or steam and condense in a cooling tube, producing concentrated nitric acid that can be used in the manufacture of high explosives. The 5 gram per hour system can produce enough nitric acid that, when mixed with battery acid (sulfuric acid), can produce usable amounts of high explosives by the simple nitration of many organic compounds.

Select the system you require from Figure 26-2. Note that each is complete with the mating power supply.

Wire up the system as directed in Figure 26-3. Obtain a suitable air supply and connect up the hoses as shown. Use ozone-resistant material for connections to the cell.

Turn on the air supply and note a bubbling from the output hose immersed under the water. Apply power to the high voltage and note a distinct smell coming from the hose when pulled out of the water. Reinsert and rig the hose end to stay under the water. The unit is now ozonating the water.

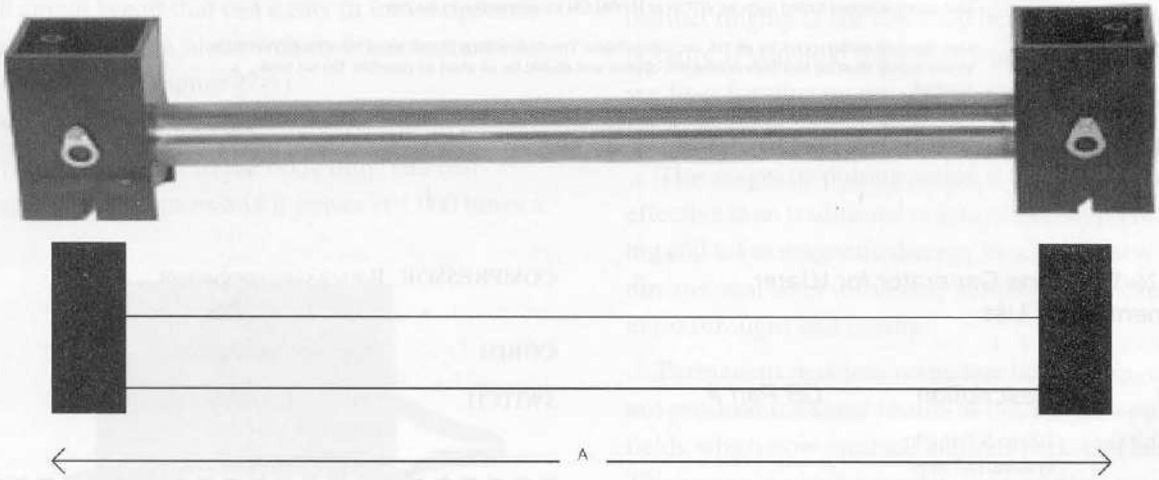

DATA BASE#	LENGTH A	OZONE OUTPUT/HR	AIR/O3 FLOW/MIN	INPUT V	INPUT I
OZONE300	182mm	100-500 mg/hr	.25-.55 CFM*	115 Vac	.2 Amps
OZONE800	282mm	500-1000 mg/hr	.35-.7 CFM*	115 Vac	.6 Amps
OZONE2000	382mm	1000-2500 mg/hr	1-1.5 CFM*	115 Vac	.9 Amps
OZONE5000	482mm	2500-5000 mg/hr	1.5-3 CFM*	115 Vac	I.2 Amps

* Multiply CFM by 28.3 to get Liters per minute

Each DATA BASE# system is complete with a matching power supply and ready to connect up as shown on figure 26-3

Figure 26-2 *System selection specifications*

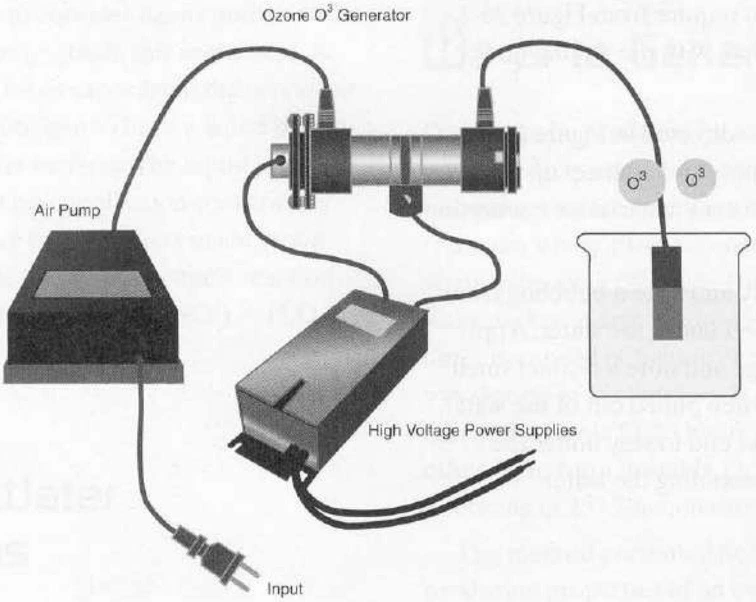

The high voltage power supplies used in these systems operate from normal 115 Vac 60 Hz. They are listed with the appropriate ozone cell on figure 26-2

The air pumps used can be that appropriate for the volume of the particular syste. A flow meter should be used when first setting up the system to verify proper air flow. Pure oxygen works the best and is available in cylinders with regulator valves and optional flow meters. Large aquarium air pumps may be used with the smaller systems while rotary vane air compressors are suggested for the larger systems. Note to multiply cubic feet per minute (CFM) by the factor 28.3 when converting to liters per minute.

Use ozone resistant tubing such as VITOR or HYPALON for connection to the cell

Use standard wiring codes for all 115 vac connections. The high voltage output wires from the high voltage power supply must be free from conductive objects and should be as short as possible. Do not twist together.

Figure 26-3 *Ozone system diagram*

Table 26-1 Ozone Generator for Water Treatment Parts List

Ref. #	*Description*	*DB Part #*
OZONE300	100 to 500 mg/hr system/182 mm/ 115 VAC @ .2 amps	DB# OZONE500
OZONE800	500 to 1,000 mg/hr system/282 mm/ 115 VAC @ .6 amps	DB# OZONE800
OZONE2000	1,000 to 2,000 mg/hr system/382 mm/ 115 VAC @ .9 amps	DB# OZONE2000
OZONE5000	2,500 to 5,000 mg/ hr system/482 mm/ 115 VAC @ .2 amps	DB# OZONE5000
HOSE	Connecting hoses, use ozone-resistant VITOR or HYPALON	
COMPRESSOR	Rotary vane compressor at required airflow	
CORD1	Three-wire power cord	
SWITCH	Two switches to control air compressor and high-voltage supply	

Chapter Twenty-Seven

Therapeutic Magnetic Pulser

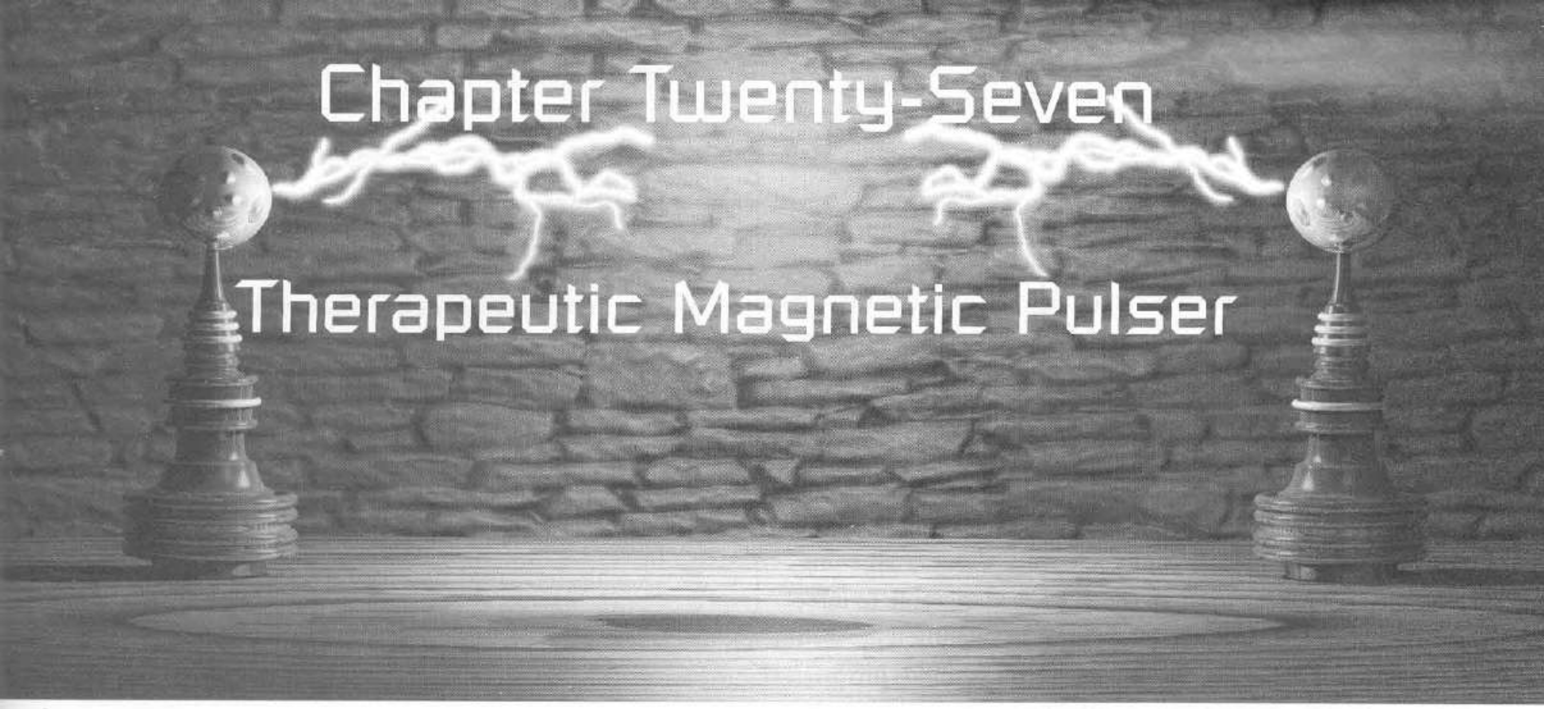

This project shows how to build a battery-powered magnetic pulser that can provide relief from aches and pains, as well as improve upon many different health concerns and conditions. The unit is built on a small circuit board that can easily fit into a cigarette pack-size enclosure, along with its 9-volt battery or wall adapter (see Figure 27-1).

This magnetic pulser is designed so the magnetic north pole is applied to the body only. The coil strength is 10,000 gauss and it pulses at 1,000 times a second. A small switch can be used to turn the device off or you may simply unsnap the battery clip.

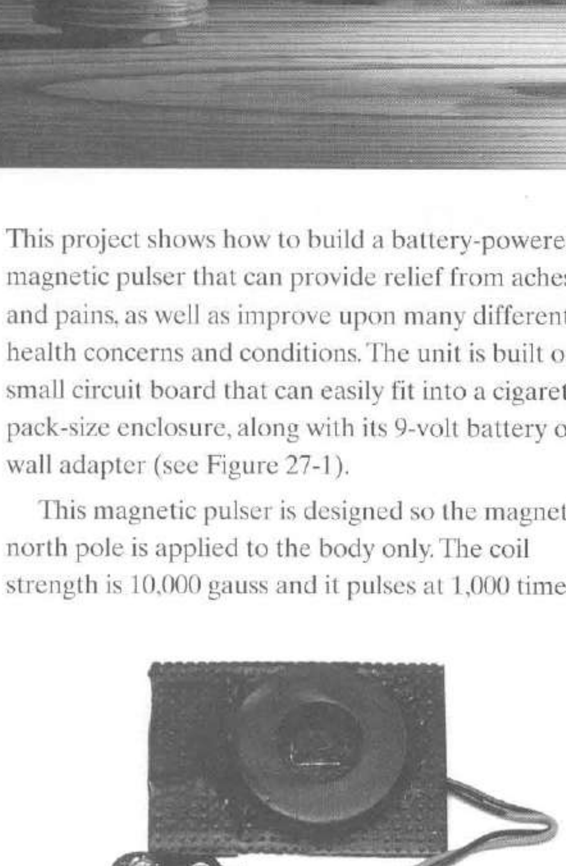

Figure 27-1 *The thermomagnetic pulser*

When an aluminum beer or pop can is held between your ear and this pulser, you hear a strong, distinct ringing as the can itself becomes the opposing polarity. If you hold the coil in front of a TV, you can see lines forming on the screen. On some TVs, the lines are visible when it's held as far away as 3 inches.

This magnetic pulsing action is more powerful and effective than traditional magnetic therapy. The pulsing coil takes magnetic therapy to a whole new dimensional level with many yet-to-explore exciting breakthroughs and results.

Permanent magnets, no matter how strong, will not produce the same results as this device's pulsed fields, which now produce induced back-emf currents. This requires a high-intensity, time-varying, magnetic impulse, not just a steady-state magnetic field.

Permanent magnets have their uses, but they work in different ways with different results. Equally important is the fact that we have "electric-powered" brains: All our thoughts and perceptions consist of complex networks of electrical signals and electromagnetic fields that pulse and sweep throughout the brain. So it then makes sense that harmonic electrical revitalization of the brain can harmonically influence your mental state and positively alter mental effectiveness.

Externally applied magnetic pulses to the lymphatic system, spleen, kidney, and liver help neutralize germinating, latent, and incubating parasites of all types, helping to block reinfection. This speeds up the elimination of disease, restores the immune system, and supports detoxification. The movement of the lymphatic system is essential in purifying, detoxifying, and regenerating the body, supporting the immune system, and maintaining health. Normally, the lymph is pumped by the movement of our body's electromagnetic field with vigorous exercise and physical activity. However, a clogged, sluggish, or weak lymphatic system prevents the body from circulating vital fluids and eliminating toxic wastes, thus weakening the immune system. It makes us vulnerable to infections and diseases. In order to be healthy, it is essential to keep the energy balanced and fluids moving so that the body's natural intelligence may operate at its full healing capacity. In addition, each cell must be enlivened with its own unique energetic frequencies and harmonic energy state and be harmonically connected to the life-force energy throughout the rest of the body.

Pulsed electromagnetic fields influence cell behavior by inducing electrical changes around and within the cell. Improved blood supply increases the oxygen pressure, activating and regenerating cells. Improved calcium transport increases the absorption of calcium in bones and improves the quality of cartilage in joints, decreasing pain dramatically. Acute and even chronic pain may disappear completely.

This device uses complex energy pulses of magnetic waves to stimulate certain body functions. It accelerates the production of vitally important hormones often providing miraculous effects. Magnetic pulsing aids in human growth hormone production and neotransmitter production. The results are remarkable, increasing vitality, sexuality, and accelerated healing. It also helps learning and the reduction of memory loss with reports of increased psychic ability for some people. The magnetic pulser also revitalizes and rejuvenates. As we get older, hormone production drops off considerably and is a reason why we age.

Users are experiencing faster healing of injuries, including bone fractures, carpal tunnel, and arthritis. The pulsed magnetic field stimulates blood flow and cellular respiration in the area applied. Reports have also documented that people's migraine headaches have ceased after magnetic therapy has been applied.

This pulser can be worn, held, or wrapped to any part of the body or held in place with a tension bandage. You can use your pulser while driving and on the go, rather than having to stop and apply it when sitting or lying down, as is necessary when using a 110-volt, standard magnetic pulser. The product is currently being sold only for experimental research and testing.

This is a simple, basic-level project requiring minimal electronic skills. Expect to spend $15 to $20. All parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in the parts list at the end of the chapter.

Circuit Description

Figure 27-2 shows a 555 timer (I1) connected as a free-running, astable pulse generator. The output pulses are nonsymmetrical with a ratio of low to high time of 10 to 1. Discharge resistor R3 and forward-conducting diode D2 control the low time, whereas charge resistor R2 and forward-conducting diode D1 control the high time. Timing capacitor C2 is common to both states of the pulse being alternately charged and discharged.

Magnetic energy is produced when the timer I1 turns on transistor Q1. The current now has the "on" time of the pulse to build up and then collapse in the flux coil L1 when Q1 is turned off. The stored energy is returned to the circuit by diode D3. The result is a steady train of magnetic pulses at 1,000 pulses per second.

The switch S1 allows the unit to be turned on and off and also allows switching in another range, dropping the pulse repetition to 100 per second. It is interesting to note that claims using specific pulse repetition rates of 20, 72, 95, 100, 125, 146, 440, 600, 660, and 727 have proven very beneficial. Do not attempt to extend the frequency without first determining the current rise in the coil. Current may be excessive and will require more turns as the frequency is lowered.

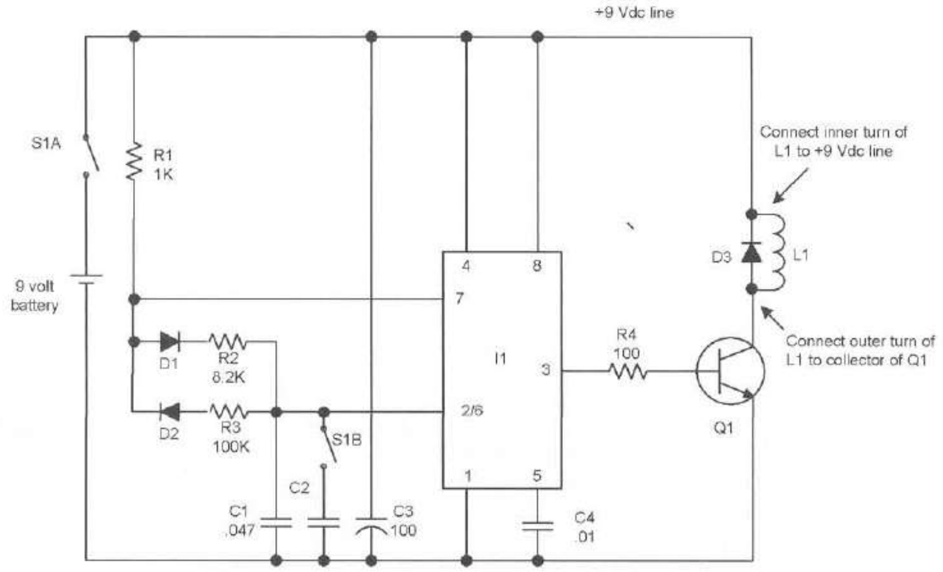

Select C2 for optional repetition rate

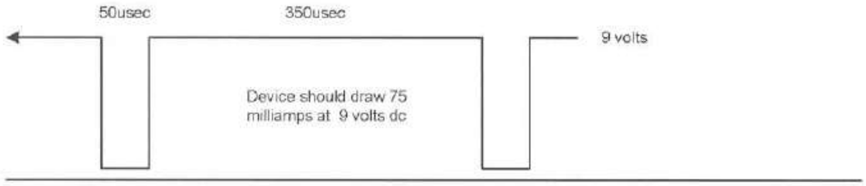

Waveform at collector of Q1

Figure 27-2 *Circuit schematic*

Assembly

1. Assemble the flux coil as shown in Figure 27-3 by first drilling a small hole adjacent to the inner diameter so that the "start" lead can exit the winding. Carefully wind 100 turns of #24 magnet wire in even windings. This is not necessary but makes for a neat-looking construction. Note to wind the wire so that its direction is the same as the figure with the leads routed as directed. This will guarantee the correct polarity as shown.
2. Cut a piece of .1-inch grid perforated board to a size of 2.6 × 1.8 inches, as shown in Figure 27-4. Locate and drill the holes as shown for the leads of battery clip CL1.
3. When building from a perforated board, it is suggested that you insert the components starting in the lower left-hand corner as shown. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors.

   Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of the assembly board.
4. Attach the external leads from the battery clip and flux coil. Note the polarity of the leads on this part as indicated in Figures 27-2 and 27-4.
5. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
6. You are now ready to test the unit. Connect it to a 9-volt battery and note a current draw of 60 to 80 milliamps. Place the coil of the unit

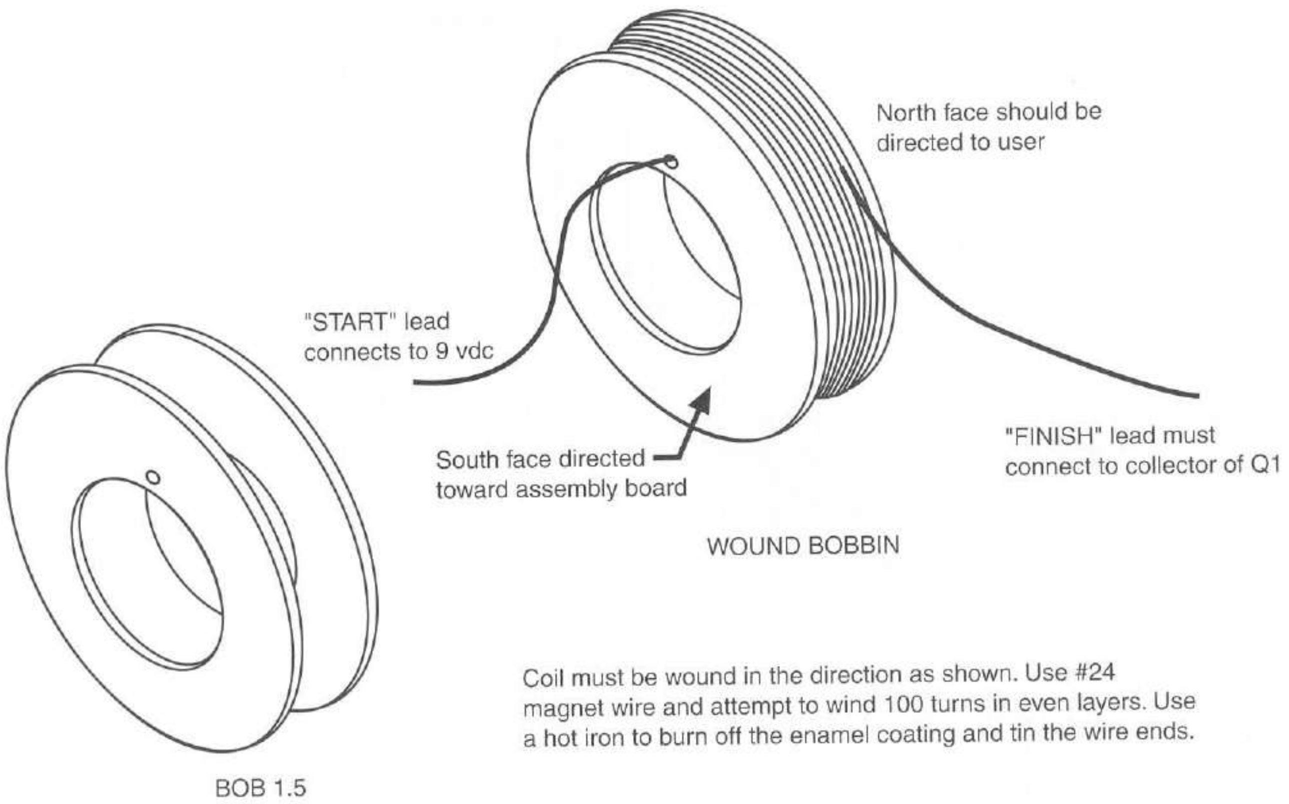
North face should be directed to user
"START" lead connects to 9 vdc
"FINISH" lead must connect to collector of Q1
South face directed toward assembly board
WOUND BOBBIN
BOB 1.5

Coil must be wound in the direction as shown. Use #24 magnet wire and attempt to wind 100 turns in even layers. Use a hot iron to burn off the enamel coating and tin the wire ends.

Figure 27-3 *Flux coil assembly*

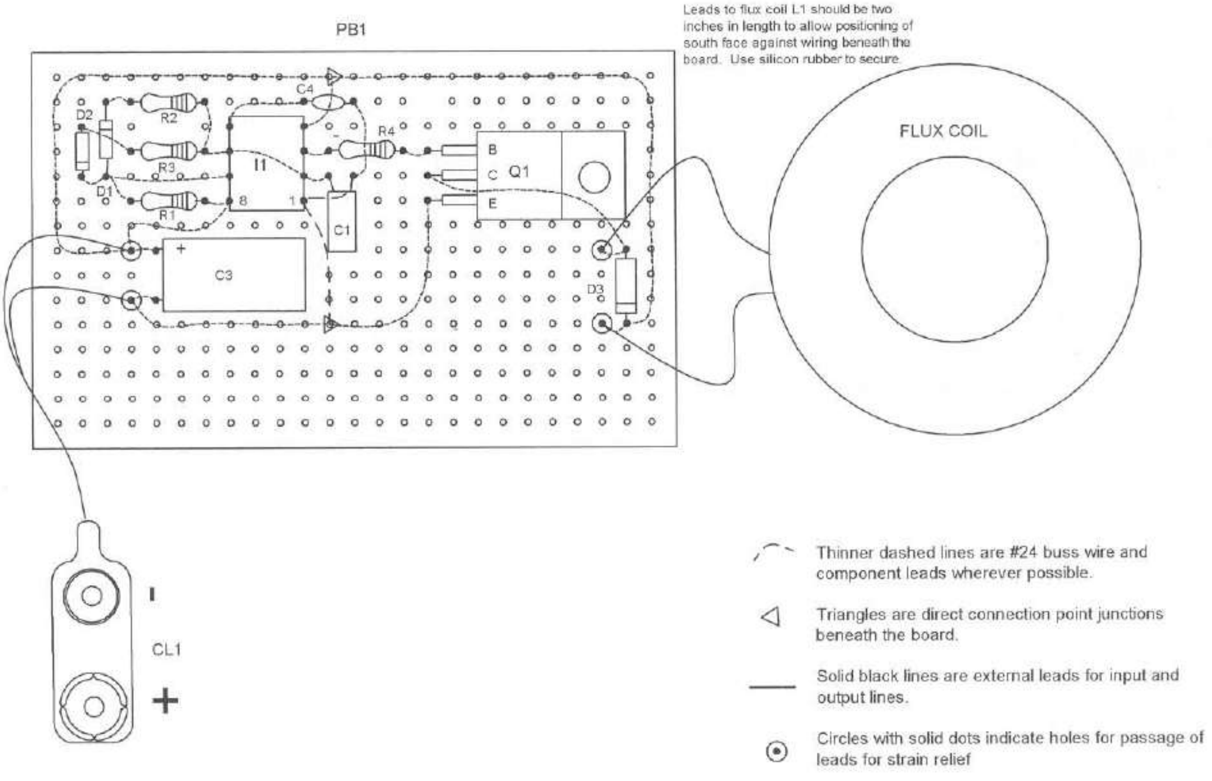
PB1
D2
R2
C4
R4
B
C
Q1
E
R3
I1
D1
8
1
R1
C1
+
C3
D3
FLUX COIL
I
CL1
+

Leads to flux coil L1 should be two inches in length to allow positioning of south face against wiring beneath the board. Use silicon rubber to secure.

Thinner dashed lines are #24 buss wire and component leads wherever possible.

Triangles are direct connection point junctions beneath the board.

Solid black lines are external leads for input and output lines.

Circles with solid dots indicate holes for passage of leads for strain relief

Figure 27-4 *Board assembly*

near a television screen and note the lines appearing. You may also put a metal can onto the coil head and place your ear to the can, noting a 1 to 2 KHz tone. If you have a scope, note the waveshape shown in Figure 27-2.

7. To use the unit, simply place the north face of the coil to the target area and allow exposure for several hours.

Table 27-1 Therapeutic Magnetic Pulser Parts List

Ref. #	*Description*	*DB Part #*
R1	1K, ¼-watt resistor (br-blk-red)	
R2	8.2K, ¼-watt resistor (gray-red-red)	
R3	100K, ¼-watt resistor (br-blk-yel)	
R4	100-ohm, ¼-watt resistor (br-blk-br)	
C1	.047-microfarad, 50-volt plastic capacitor	
C2	1-microfarad, 25-volt vertical electrolytic capacitor	
C3	100-microfarad, 25-volt vertical electrolytic capacitor	
C4	.01-microfarad, 50-volt disk capacitor (103)	
D1, 2	Two IN914 silicon diodes	
D3	1-kilovolt, 1-amp diode 1N4007	
Q1	TIP31 NPN TO-220 power transistor	
I1	555 dual inline package (DIP) timer	
BOB1.5	1.5 × .3 × .75-inch ID nylon bobbin (see Figure 27-3)	DB# BOB1.5
L1	Flux coil (see Figure 27-3)	DB# FLUXCOIL
PBOARD	2.6 × 1.8 × .1-inch grid perforated board as shown in Figure 27-4	
CL1	9-volt battery snap clip	

Chapter Twenty-Eight

Noise Curtain Generator

This unique circuit shown in Figure 28-1 is designed to produce a relaxing sound like that of a breaking surf. It is technically known as pink noise, being defined as a random distributor of equal sounds in the audible spectrum, favoring the higher-frequency end.

Pink noise has a spectral intensity that is inversely proportional to frequency for a given range. Therefore equal power is dissipated into a fixed resistance in any octave bandwidth in that range. White noise is random noise, which is typically thermal and shot noise, has a constant energy per unit bandwidth, and is independent of the central frequency at the band.

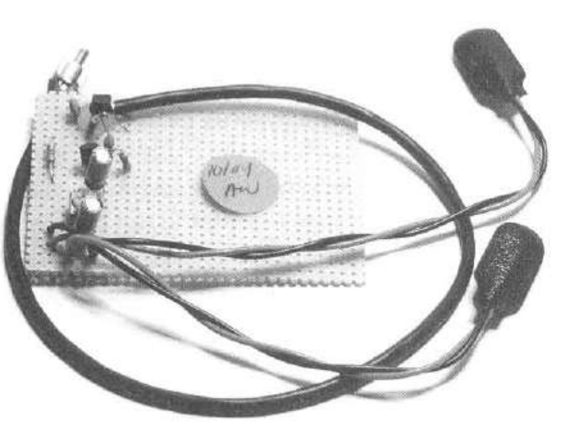

Figure 28-1 *The board-level generator*

This device is intended for use as a relaxing source of sound and also as a mask for certain types of surveillance equipment. Crashing surf is an excellent example of naturally produced pink noise that has a great relaxing effect. Due to the nearly equal amplitude of all the frequencies generated, the electronic equipment of an acoustical nature becomes easily overloaded and saturated, whereby normal voice sounds of varying amplitude and of limited bandwidth are much more finite and are unable to be processed by the equipment. In other words, the noise jams the microphone and audio preamplifier circuitry, thus rendering it unable to detect normal voices. (It should be noted that in certain instances a running shower will sometimes help simulate pink noise.)

The signal from the device is intended to be fed into the input of any amplifier, radio, or tape deck. It is easily connected to the center arm of the volume control if an input jack is not provided in a regular radio. (You will note the phono plug attached.) Also, the two batteries used are left on permanently, due to the fact that so little power is used.

The optional "speaker amplifier" reference schematic in Figure 28-2 is built on a piece of perfboard using standard audio-frequency wiring techniques. When used as a totally self-contained system, the electronics may be housed along with all batteries into a case or enclosure determined by the builder. This approach provides a convenient portable unit.

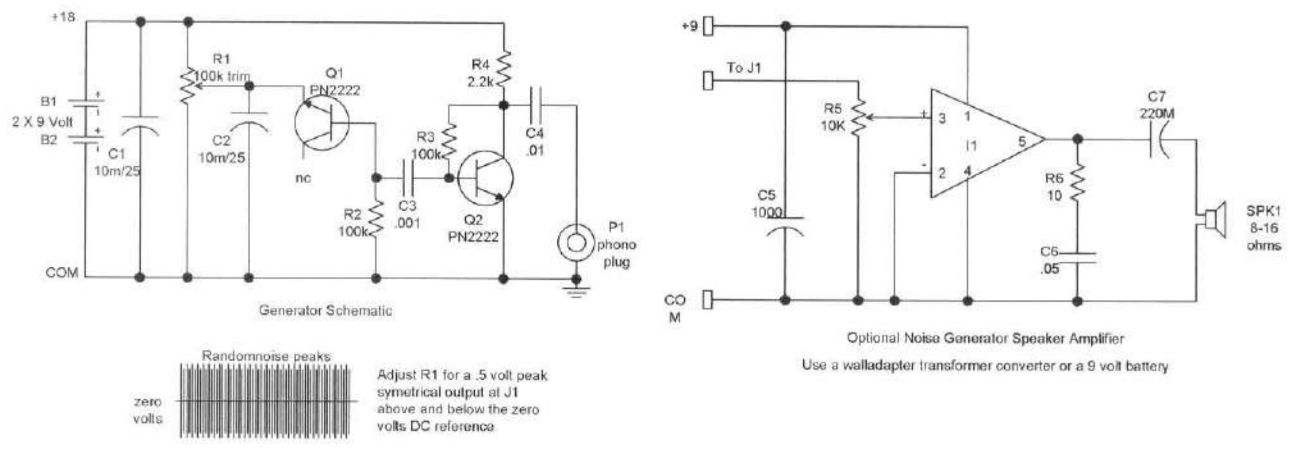
+18
B1
2 X 9 Volt
B2
C1
10m/25
R1
100k trim
C2
10m/25
Q1
PN2222
nc
R2
100k
C3
.001
R3
100k
R4
2.2k
C4
.01
Q2
PN2222
P1
phono
plug
COM
Generator Schematic
Randomnoise peaks
zero
volts
Adjust R1 for a .5 volt peak symetrical output at J1 above and below the zero volts DC reference
+9
To J1
R5
10K
C5
1000
I1
3
1
2
4
5
R6
10
C6
.05
C7
220M
SPK1
8-16
ohms
CO
M
Optional Noise Generator Speaker Amplifier
Use a walladapter transformer converter or a 9 volt battery

Figure 28-2 *Circuit schematic*

Circuit Description

Pink noise is a form of white noise that is the Gaussian distribution of all possible frequencies, the difference being that pink noise is more weighted to the audio spectrum. This form of noise has some very interesting properties, one being the ability to cause relaxation and a sense of well-being. Another property is that it provides a background that will completely mask an annoying device, rendering it unable to affect normal conversation. The device described is designed to work with any sound system with an audio input or is easily adapted to drive a loudspeaker by connecting it to a normal radio or building an optional speaker amp, as shown in the accompanying plans, making a complete system (see the parts list at the end of the chapter).

The circuit works in the following manner: A base emitter junction of transistor Q1 is reversed biased through a current-limiting resistor into breakdown (avalanche) condition. The random shot noise created is fed to the common emitter amplifier transistor, Q2, and to a filter, which in turn provides a low-level signal output. The unit module board and the batteries are placed in a plastic case with pieces of foam rubber or plastic. The outlet leads are fed out per the builder's requirements. The batteries are permanently installed and left on because the operating current is so low that there is no reason for a switch. Note that Jl is *not* required if using the optional speaker and amplifier circuit.

Assembly Steps

If you are a beginner it is suggested you obtain our *GCAT1 General Construction Practices and Techniques*. This informative manual explains basic practices that are necessary in the proper construction of electromechanical kits and is listed in Table 28-1.

1. Lay out and identify all the parts and pieces. Verify them with the parts list, and separate the resistors as they have a color code to determine their value. Colors are noted on the parts list.
2. Cut a piece of .1-inch grid perforated board to a size of 3 × 2.4 inches. Locate and drill the holes as shown in Figure 28-3 for the leads of the battery clips.
3. If you are building from a perforated board, it is suggested that you insert the components starting in the lower left-hand corner as shown in Figure 28-3. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate connection runs on the underside of the assembly board.

4. Attach the external leads as shown in Figure 28-4. Notice the special note on the shielded cable.
5. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.

Testing Steps

Insert the two battery clips and plug the output cable into the microphone or auxiliary jack of the existing system. Note the rushing sound and adjust the trimpot R1 to the desired effect. If you have a scope, make the adjustment as shown in Figure 28-2. If you build the speaker amplifier section, simply connect it as shown for a self-contained system suitable for nightstand use.

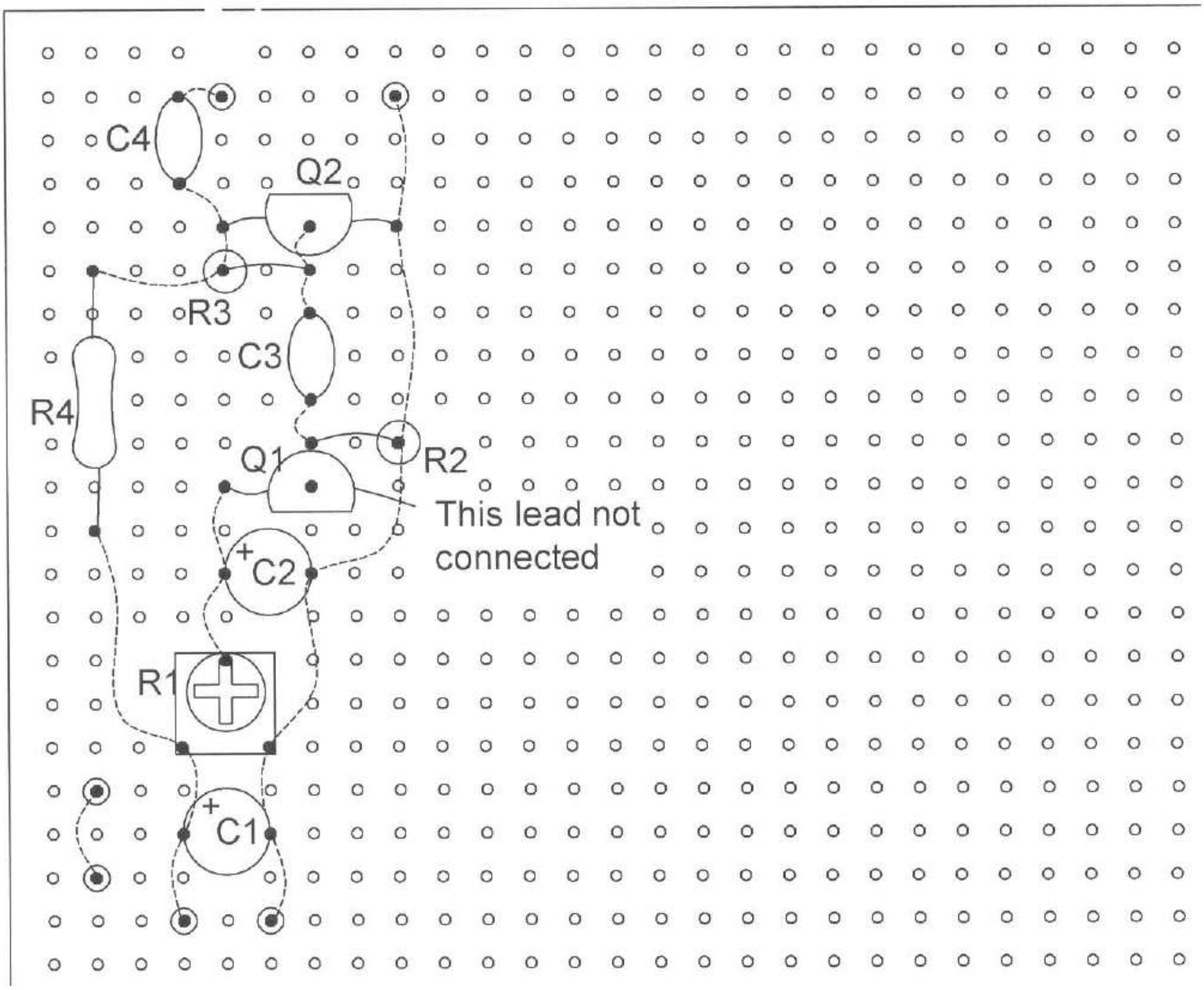

Figure 28-3 *Assembly board layout*

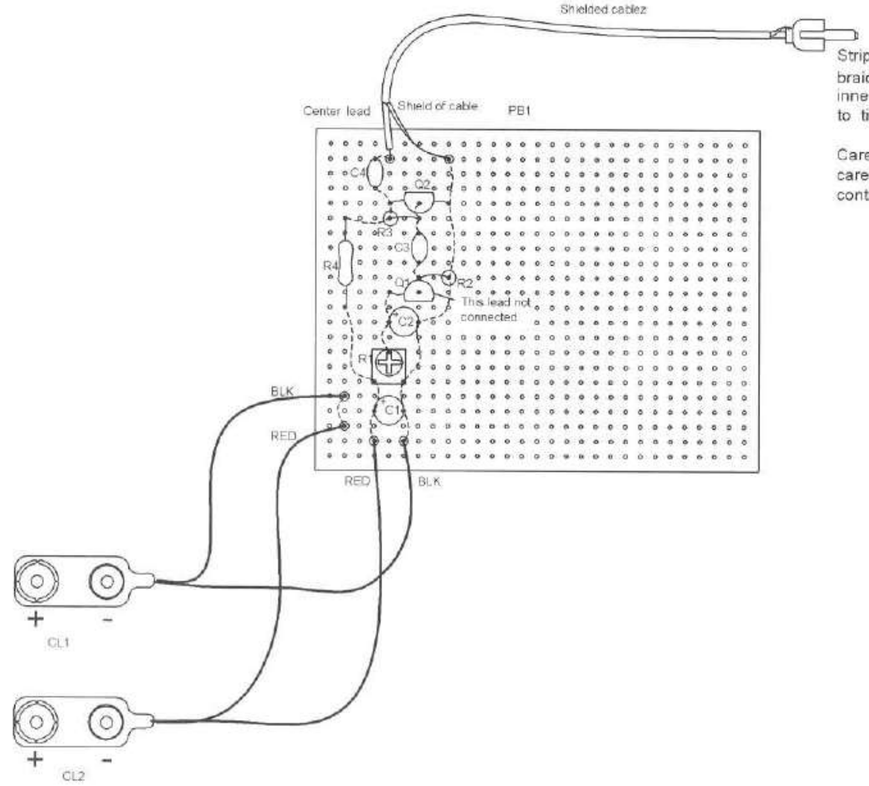

Strip off outer plastic jacket of cable 1 inch and twist braid into a lead and tin. Strip off ¼ inch of insulation on inner lead and tin. Insert into plug and solder inner lead to tip.

Carefully solder twisted braid lead to outer shell being careful not to burn through insulation. Always verify lead continuity and possible shorting of the leads with a meter.

Figure 28-4 *Assembly showing external leads*

Table 28-1 Noise Curtain Generator Parts List

Ref. #	*Description*	*DB Part #*
R1	100K trimpot vertical mount	
R2, 3	Two 100K, ¼-watt resistors (br-blk-yel)	
R4	2.2K, ¼-watt resistor (red-red-red)	
R5	10K trimpot vertical mount	
R6	10-ohm, ¼-watt resistor (br-blk-blk)	
C1, 2	10-microfarad, 50-volt vertical electrolytic capacitor	
C3	.001-microfarad, 50-volt disc capacitor	
C4	.01-microfarad, 50-volt disk capacitor	
C5	1,000-microfarad, 25-volt vertical electrolytic capacitor	
C6	.047-microfarad, 50-volt plastic capacitor	
C7	220-microfarad, 25-volt vertical electrolytic capacitor	
Q1, 2	Two PN2222 NPN general-purpose transistors	
I1	LM386 dual inline package (DIP) operational amplifier	
PB1	2.25 × 3 × .1 × .1-inch grid perforated circuit board	
SPK1	Small 8- to 16-ohm speaker	
CL1, 2	Two 9-volt battery clips and leads	
P1	RCA phono plug	
SHC1	12-inch shielded microphone cable	

Chapter Twenty-Nine

Mind-Synchronizing Generator

Caution: This machine can trigger epileptic seizures and should be avoided at all costs by anyone who suspects even slightly that they are epileptic.

Your mind synchronizer is designed to be a completely self-contained unit with built-in batteries (see Figure 29-1). The system produces a variable-rate, flashing, monochromatic, directional, "shaped" light source (a light source with a variable ratio of off to on times), and a complementary synchronized audio pulser with tone-shaping control. The unit may be used so that a group can be exposed to light and sound pulses. For individual use, headphones can be used, along with special assembled light glasses. These are shown made from an existing pair of sunglasses. Optimum performance requires a quiet and reasonably low light condition, and distractions must be at a minimal. The mind-synchronizer can produce strange and bizarre hallucinations, as well as provide a sense of relaxation and well-being.

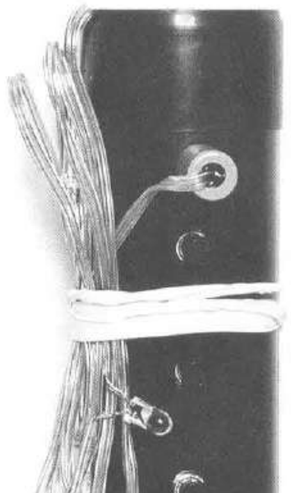

Figure 29-1 *Photograph of unit*

Information on Mind Control

Today's work environment demands that we have the ability to excel, to compete, and to resolve problems. Hence, the most important tool we possess is our mind. To optimize our potential, we now offer our low-cost mind-synchronizing unit.

After 25 years of research, Dr. Axel Bruck of Berlin, Germany, created a machine that can enhance, develop, and perfect the qualities within you in order to attain the goal of mental fitness. It has been found that external stimulation evokes a mental reaction in people that is either active or passive. The mind synchronizer can determine automatically the kind of person you are. An evaluation is performed via the means of a hand sensor using a technique called biofeedback, which is sensitive to the flow of blood. This knowledge is very important for providing optimal stimulation to the users.

Having determined the kind of person you are, the mind synchronizer now uses "programmed" light pulses and "shaped" sound to help you reach a required mental condition, be it relaxation or efficiency. It has been known for hundreds of years that an external rhythm of light and sound has an effect on your mental condition. The mind synchronizer is a modern version of a drum and a flickering fireplace.

Through extensive scientific measurements, experiments, and results, the mind synchronizer can improve your performance level by using the right stimuli:

- Deep relaxation to the point of sleeping
- Relaxation
- Equilibrium
- Activation of muscles doing physical movements
- Efficiency

Whether in sports, professions, or personal lives, the development of personality and identity is vital to personal success and satisfaction. The mind synchronizer trains and perfects your mental and psychic abilities so that you have a stable way of thinking with a healthy outlook. The mind synchronizer teaches you the technique of concentration, stress handling, focusing, visualization, relaxation, and self-awareness

When using the mind synchronizer, you learn to set a goal and concentrate on this goal completely. You eliminate all distractions and thereby maintain your concentration over a long period of time. By doing so, you improve your power of concentration. In the long run, you will think clearly, make better decisions, and deal more effectively with life itself.

The program of relaxation combats stress, tension, and direct pressure. Almost everything we do causes stress. Stress means that the body is in a chaotic condition. The opposite of stress is relaxation. The right kind of relaxation is an optimization of mind, concentration, learning ability, performance, and health.

In reality, a relaxed person saves more than half a billion heartbeats over a nervous or hectic person, according to Professor Prinzinger's theory of the Quiet Pulse. Professor Prinzinger performed much research into the various relaxed states of the body and the relationship of relaxation to heart rate and blood pressure. The body consumes less energy, organs are burdened less, and life expectancy increases.

The program of efficiency is a newly developed psycho-active technique: the indication of unconscious eye movement and the employment of rhythm displacement. Certain stimulating rhythms such as that of a strobing light can produce nausea when strobing at 20 repetitions per minute. Other stimulating rhythms of lower repetitions per minute can produce relaxation and the feeling of well being. In such a case, eye movement usually follows at the lower repetition rates. The combination of visual effects (flashing *light-emitting diodes* [LEDs]) with audio input (via a headset) produces high-efficiency performance. The mind synchronizer establishes physical and mental readiness to attain this goal of high performance.

The program of activation involves more optic, visual patterns than acoustic ones to provide "imagination journeys." To experience associations with your own history and your own personality, you will be able to discover your emotional condition and personality.

The program of equilibrium allows for meditation with your eyes open. You would observe attentively, simply, and exactly each individual light source and its changes. The mind synchronizer will bring you the healthy balance you desire through light and sound.

Mind machines are available from a wide variety of sources today. Often referred to as hemispheric synchronizers or cortical-frequency entertainment devices, they all basically perform one basic function: They lead the brain to synchronize both hemispheres and then lead those synchronized hemispheric frequencies to a specific target frequency. In order to understand why anyone would want to do such a thing, let's examine what frequencies the brain produces and what people experience when the brain is at these frequencies.

Generally, brain waves are broken down into four different levels. They are referred to as beta, alpha, theta, and delta. With respect to frequency, they are as follows:

Beta: 13 cycles per second or more

Alpha: 8 to 12 cycles per second

Theta: 5 to 7 cycles per second

Delta: 1 to 4 cycles per second

Profound changes take place in a person's conscious experience at each of these levels. The beta frequency is the "normal" waking frequency, the drive-in-the-city, go-to-work, drink-coffee frequency, and it can range as high as 40 cycles per second. It is *not* a relaxing frequency.

The alpha frequency is a realm of relaxed awareness. It has been shown in a concrete fashion that this frequency is associated with an increased capacity for learning, comprehension, and retention. It is in the alpha state that the best learning occurs.

Theta is a hypnotic state often associated with meditation, a zen-like state of waking dreaminess. It can be a profoundly relaxed state where internal associations and thoughts can form ideas with clarity that could never have emerged in the normal, high-noise environment of the beta state. This is also the range that many native shamans have been found to enter when engaging in healing trance work. It is a state of profound tranquility.

The delta state is much simpler to define; you're asleep. When your brain waves are in the delta range, you're in a sound and restful sleep.

A number of studies have been done that indicate that a flashing light has a profound impact on the entire brain, more so than a pulsing sound or feeling. When the brain is exposed for a time (as short as 2 to 5 minutes) to a flashing light at a speed close to the brain's existing operating frequency, the brain will begin to synchronize with the stimulation. Moreover, both hemispheres, which are often operating at different frequencies, will harmonize. This has also been shown clinically and statistically to be an aid to mental functioning. Many books are available on the market on this subject for anyone who would like more detailed information on these studies.

Operating a Mind Machine

Now with the previous information in mind, the operation of the mind machine in all of its forms is basically the same. Start the machine off at a frequency that is roughly equivalent to the operating frequency of your brain at the time. (Use your own feelings as a gauge. If you're frazzled, start at maximum frequency; if you're pretty mellow, start at about half-frequency.) Allow yourself about 3 to 4 minutes at this frequency, and then slowly begin bringing the frequency down, very slowly. You should take about 7 to 10 minutes to bring the control all the way down. When you reach a minimum, leave it there for about 5 to 7 minutes.

After you've gotten to the bottom range, you can either lead yourself back to normal, in which case you will feel what you may consider a "good normal." If you so choose, you can exit directly from the lowest setting. If you do, have taken your time getting to it, and stayed there for several minutes, you will notice that it may be very hard to begin to move. Your limbs may feel very heavy, as if your body were asleep. If you manage to get to your feet, you may notice that things won't feel as solid; they'll seem to have a rubbery quality to them. These sensations can be quite pleasant as long as you're prepared and expect them (and don't have to drive or operate heavy machinery afterward!).

The visual effects are the most stunning and can only be likened to the experience of ingesting certain mind-altering substances. As the frequencies change, you will notice colors. Yes, two little flashing red lights in front of your closed eyes will create the illusion of fantastic, swirling colors and geometric designs ranging from ultimate simplicity to incredible complexity. One effect will be observed as myriads of little shapes that form and begin to slowly spiral around your field of vision, first clockwise, then counterclockwise, then back again. You will see dreamy, smooth, cloudlike fields of frosty blue-white, deep blues, deep reds, electric greens, yellows, whites, and purples, all in balls, stripes, stars, and triangles, dancing and moving like the most intense kaleidoscope you can imagine. The imagery is absolutely fascinating and can transfix you with the dazzling display. You will be absolutely amazed and impressed with the fact that it is your brain that produces this light show.

This unit is designed to take you from a medium-level beta down to a low theta. You can use these states as you desire. You may wish to stop the frequency decrease early and remain in the alpha state before an exam, prior to entering a class, or when trying to learn something important. You can continue on down and enter the theta level and create healing imagery in your mind to assist your body's own natural healing processes.

Most of all, play with your unit and enjoy the show. Built right, it will provide many years' worth of reliable service. Through usage, you can become familiar with the settings and be able to use the machine in a way that's exactly right for you. Always leave yourself plenty of time to use the machine and at least a half-hour after to "normalize." Use the unit to help you drift off to sleep; you can actually lead your brain down into relaxation and turn the unit off after you've gone into a deep theta, and if you're ready to sleep away, it's only a short slide from theta into the peaceful slumber of delta.

Circuit Theory

Integrated circuit I1 is an astable oscillator with a pulse rate being controlled from approximately 3 to 40 reps per second by external pot R2. The output of U1 now triggers integrated circuit I2 being connected as a monostable circuit that controls these pulses' ratios of on to off time. This duty cycle control is via external pot R5. Light-emitting diodes (LED1 and LED2) are connected to the output pin 3 of I2 through brightness control pot R7. These are the diodes you install in the sunglasses or they can be a single ultrabright LED for group use. The result is a variable flash rate from 3 to 40 pulses per second with a variable duty cycle rate (ratio of on to off time).

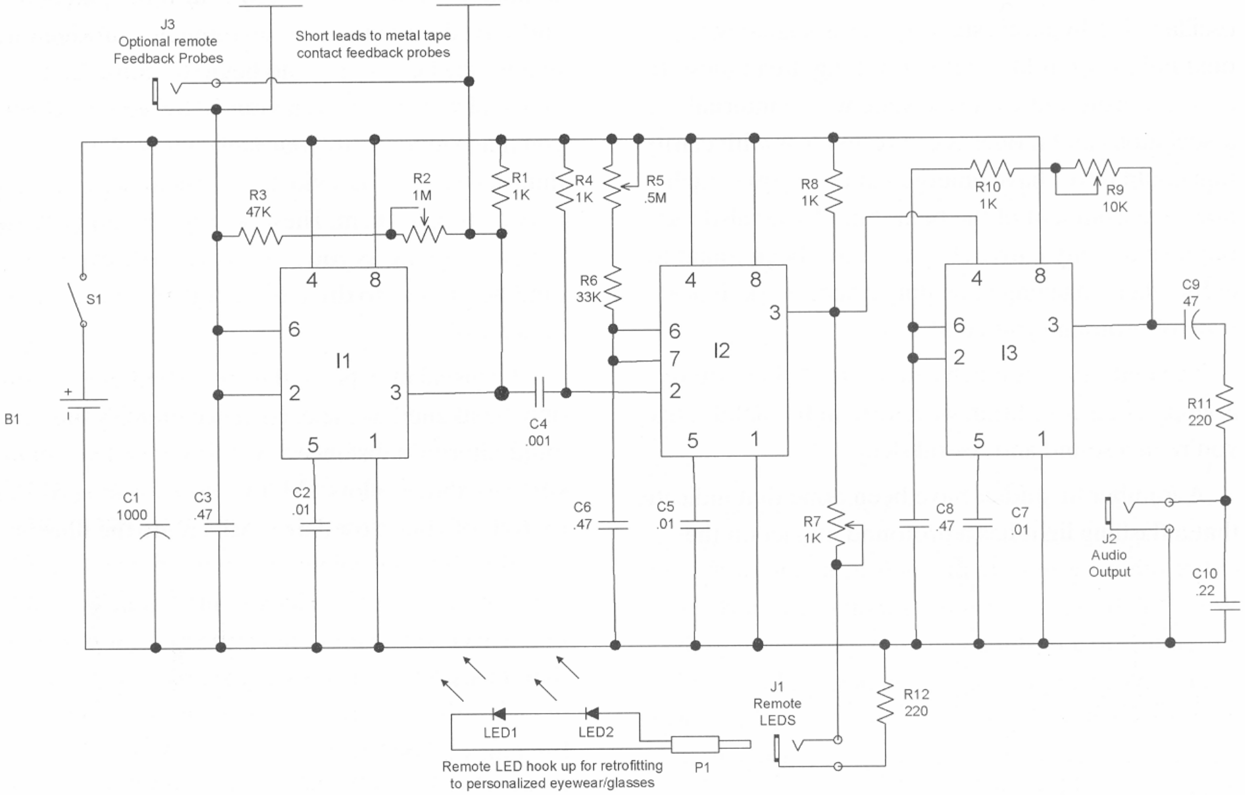
J3
Optional remote
Feedback Probes
Short leads to metal tape
contact feedback probes
R3
47K
R2
1M
R1
1K
R4
1K
R5
.5M
R8
1K
R10
1K
R9
10K
S1
B1
I1
R6
33K
I2
I3
C4
.001
C9
47
R11
220
C1
1000
C3
.47
C2
.01
C6
.47
C5
.01
R7
1K
C8
.47
C7
.01
J2
Audio
Output
C10
.22
J1
Remote
LEDS
R12
220
LED1
LED2
Remote LED hook up for retrofitting
to personalized eyewear/glasses
P1

Figure 29-2 *Circuit schematic*

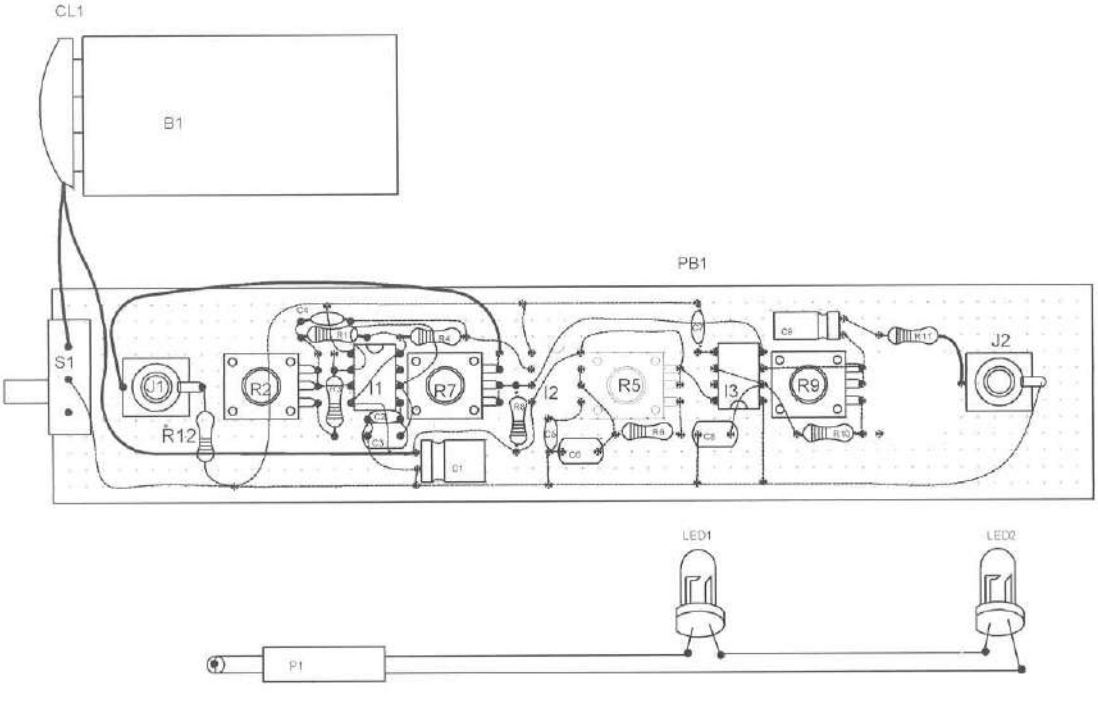

Emitter assembly for retrofitting to personalized eyewear glasses

Figure 29-3 *Assembly board*

Note it is best to experiment with and record the duty cycle settings and their effects (see Figure 29-2).

The circuit also includes integrated circuit I3 connected as a variable-rate, sonically "shaped" audio oscillator. The output is via jack J2 driving a set of 32-ohm headphones.

The feedback circuit consists of two external pieces of metallic tape providing points of contact for the user's hand. Body resistance now partially controls the pulse rate of I1. An uptight situation provides lower contact resistance and consequently raises the pulse rate. As one relaxes, this resistance becomes higher and lowers the pulse rate. An external jack (J3) provides a connection for other bodily contact probes to the feedback system. These contact probes are left to the discretion of the user. This jack is omitted when using the external tape probes. Switch S1 controls power to the circuit and should be off when not in use to conserve the internal batteries.

Assemble the Board

If you are a beginner, it is suggested to obtain our *GCAT1 General Construction Practices and Techniques*. This informative manual explains basic practices that are necessary in proper construction of electromechanical kits and is listed in Table 29-1.

1. Lay out and identify all the parts and pieces. Verify them with the parts list, and separate the resistors as they have a color code to determine their value. Colors are noted on the parts list at the end of the chapter.
2. Cut a piece of .1-inch grid perforated board to a size of 6¼ × 1¼ inches, as shown in Figure 29-3. An optional PCB is individually available.

   The circuit is shown with the more challenging perforated circuit board often required for a science fair project. A PCB is also available and requires that you only identify the particular part and insert it into the respective holes as noted. The PCB is plainly marked with the part identification. Soldering is very simple as you solder the component leads to the conductive metal traces on the underside of the board.

   The perforated board approach is more challenging as now the component leads must be routed and used as the conductive metal traces. We suggest that the builder closely follow the drawings on this section and mark the actual holes with a pen before inserting the

Obtain a pair of sunglasses preferably the ones with side shields. Measure the distance between your eye centers. This is usually about 2½". Drill the two holes just a bit smaller than the LEDS to provide a secure fit. Press into the holes and glue if necessary to further secure.

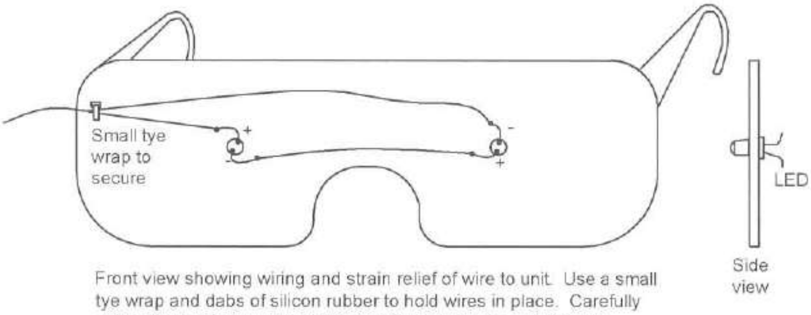

Front view showing wiring and strain relief of wire to unit. Use a small tye wrap and dabs of silicon rubber to hold wires in place. Carefully solder the leads from the LEDS to extension leads and observe polarity as shown on the emitter assembly on figure 29-5

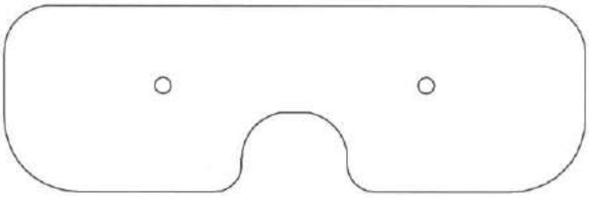

View towards the eyes. Secure the leads to the frames of the assembly to keep out of the way. You may also use pieces of duct or electrical tape to hold all in place.

Figure 29-4 *Assembly of glasses*

parts. Start from a corner, using it as a reference, and proceed from left to right. Note that the perforated board is the preferred approach for science projects as the system looks more homemade.

3. If you are building from a perforated board, it is suggested that you insert the components, starting in the lower left-hand corner, as shown in Figure 29-3. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors. Resistor RI2 is connected from the pin of 11 to the PC board. Note the control pots R2, R5, R7, and R9 and jacks J1, J2, and J3 should be positioned to closely match the holes, as shown in Figure 29-5. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of assembly board.
4. Attach the external leads to the battery holder as shown.
5. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
6. Assemble an LED to cable P1 as shown. Also, retrofit to the eyewear as per Figure 29-4.
7. Fabricate the enclosure tube (TUB1) as shown in Figure 29-5 and insert the assembly board and battery. Verify that the controls mounted on the assembly board will match the drilled holes with the dimensions as shown. Note the dimensions are shown for use with the optional PCB and may be changed. The unit is secured attaching the adjustment pots via nuts.
8. The final assembly is shown in Figure 29-6 and should be operated as instructed.

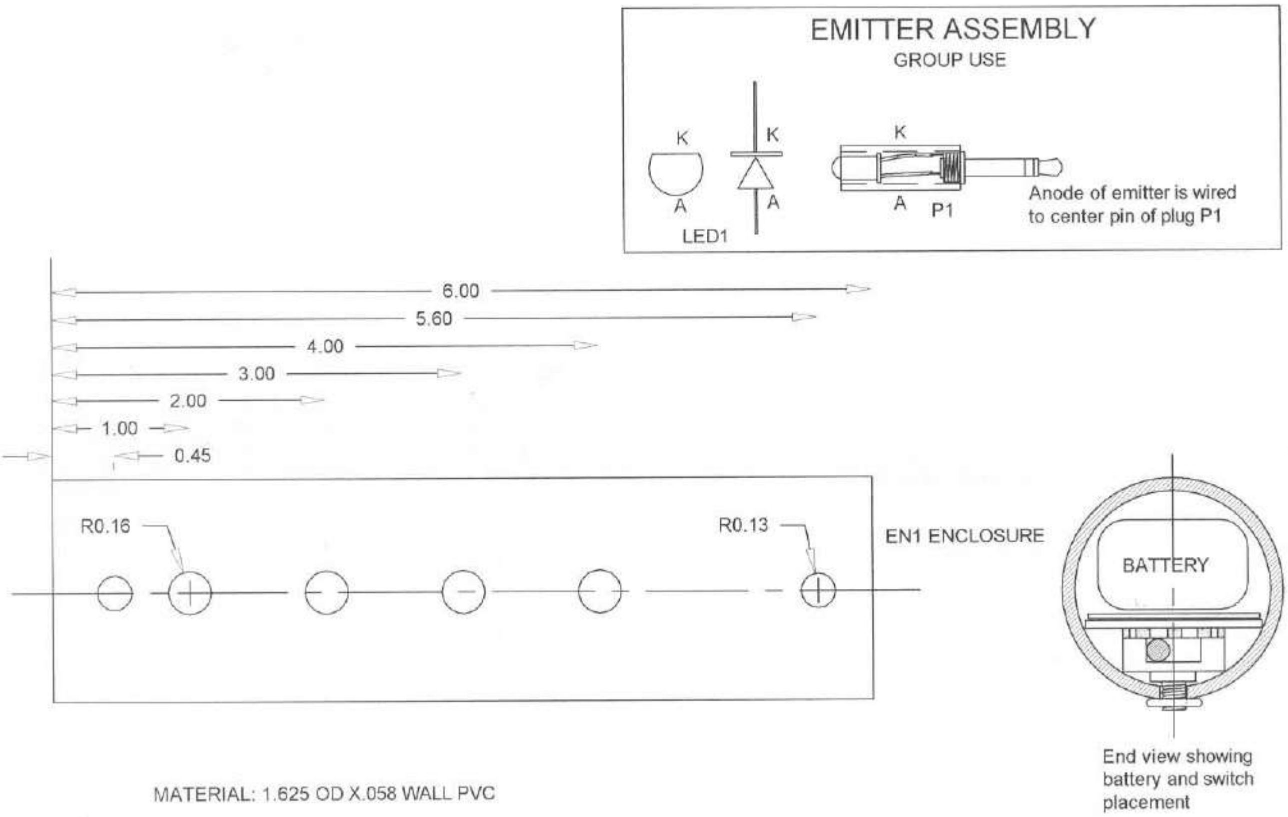
EMITTER ASSEMBLY
GROUP USE
K
K
K
A
A
A
P1
LED1
Anode of emitter is wired to center pin of plug P1
6.00
5.60
4.00
3.00
2.00
1.00
0.45
R0.16
R0.13
EN1 ENCLOSURE
BATTERY
End view showing battery and switch placement
MATERIAL: 1.625 OD X.058 WALL PVC

Figure 29-5 *Fabrication of enclosure*

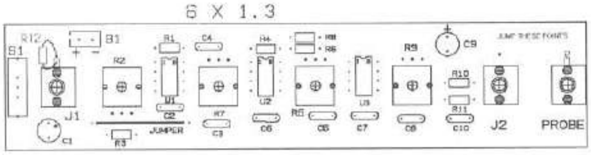
6 X 1.3
R12
B1
R1
C4
R4
R8
R6
R9
C9
JUMP THESE POINTS
R2
R10
R11
J1
C2
JUMPER
R7
C3
C6
R5
U1
U2
U3
C5
C7
C8
C10
J2
PROBE
C1
R3

Figure 29-6 *Final assembly on PC Board*

Table 29-1 Mind-Synchronizing Generator Parts List

Ref. #	*Description*	*DB Part #*
R1, 4, 8, 10	Four 1K, ¼-watt resistors (br-blk-red)	
R2, 5	.5-1 meg potentiometer resistor	
R3	47K, ¼-watt resistors (yel-pur-or)	
R6	33K resistor (or-w-or)	
R7	1K pot resistor	
R9	10K pot resistor	
RI1, 12	Two ¼-watt 220- to 330-ohm, resistors (red-red-br)	
CI	1,000-microfarad, 25-volt vertical electrolytic capacitor	
C2, 5, 7	Four .01-microfarad, 50-volt disc capacitors	
C4	.001 mfd, 50-volt disk capacitor	
C3, 6, 8	Three .47-microfarad, 50-volt mylar capacitors	
C9	47-microfarad, vertical electrolytic capacitor	
C10	.22-microfarad, 50-volt mylar capacitor	
Il, 2, 3	Three 555 dual inline package (DIP) timers	
*LED1, 2	Two bright-red LEDs or use single, optional 15,000 mcd assembled as shown	
S1	Single pole, single throw (SPST) PC switch	
CLI	9-volt battery snap	
11, 3	Two 3.5 mm mono jacks	

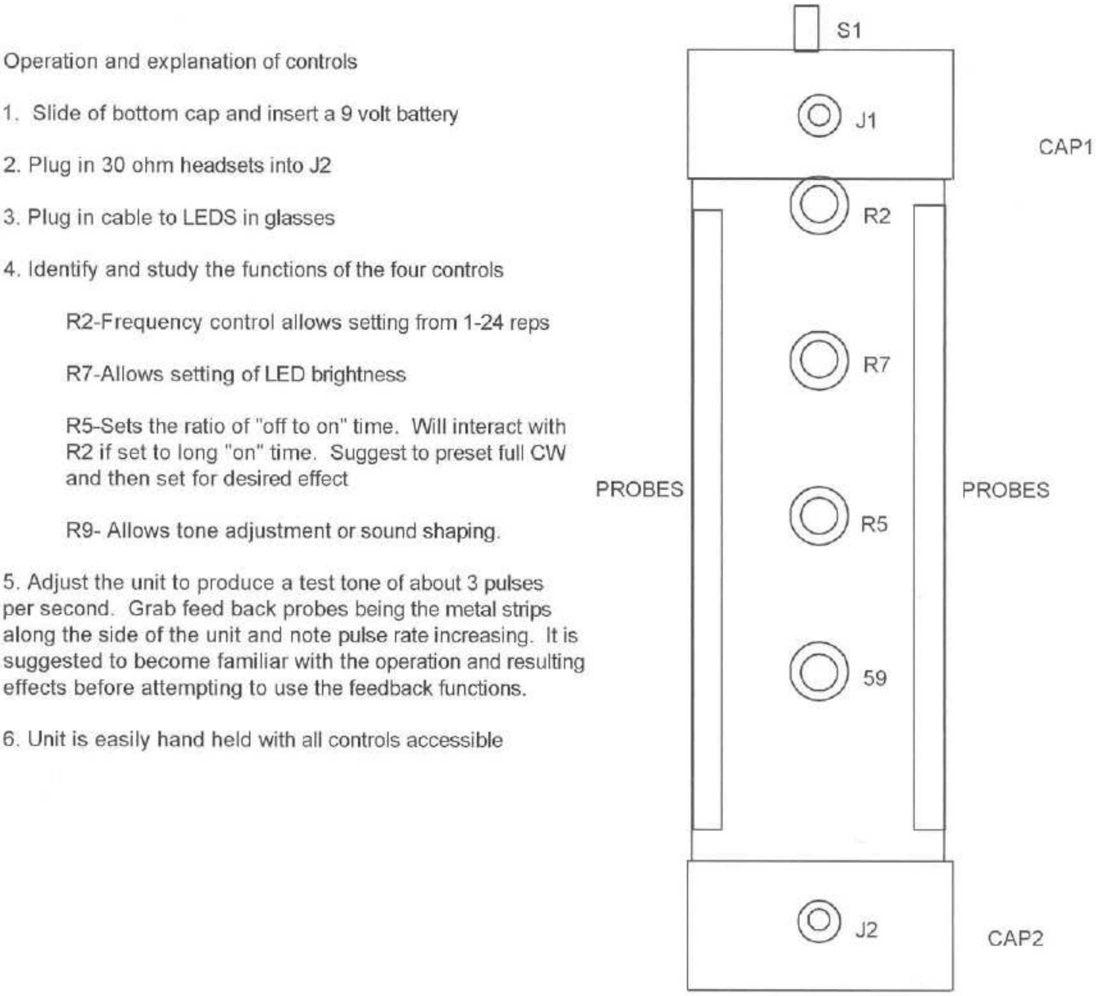

Operation and explanation of controls

1. Slide of bottom cap and insert a 9 volt battery
2. Plug in 30 ohm headsets into J2
3. Plug in cable to LEDS in glasses
4. Identify and study the functions of the four controls

    R2-Frequency control allows setting from 1-24 reps

    R7-Allows setting of LED brightness

    R5-Sets the ratio of "off to on" time. Will interact with R2 if set to long "on" time. Suggest to preset full CW and then set for desired effect

    R9- Allows tone adjustment or sound shaping.

5. Adjust the unit to produce a test tone of about 3 pulses per second. Grab feed back probes being the metal strips along the side of the unit and note pulse rate increasing. It is suggested to become familiar with the operation and resulting effects before attempting to use the feedback functions.
6. Unit is easily hand held with all controls accessible

Figure 29-7 *Operation and Explanation of Controls*

J2	3.5 mm stereo jack
WR1	12 inches of #24 hookup wire
WR2	24 inches of #24 bus wire
WR3	48 inches of two-cord speaker wire
PB1	.1 × 6¼ × 1¼-inch grid perforated board
PCBOARD	Optional printed circuit board (PCB) with printing DB# PBMIND
EN1	Enclosure, sized at 1⅝ OD × 1⁷⁄₁₆ ID × 7 inches
CA1, 2	Two 1⅝ inch plastic caps #A 1⅝
P1	3.5 mm plug for LED cable to eyewear
HEADSETS	Stereo or mono, 32-ohm, cushioned headphones
TAPE	4 × ½-inch metallic silicon tape229

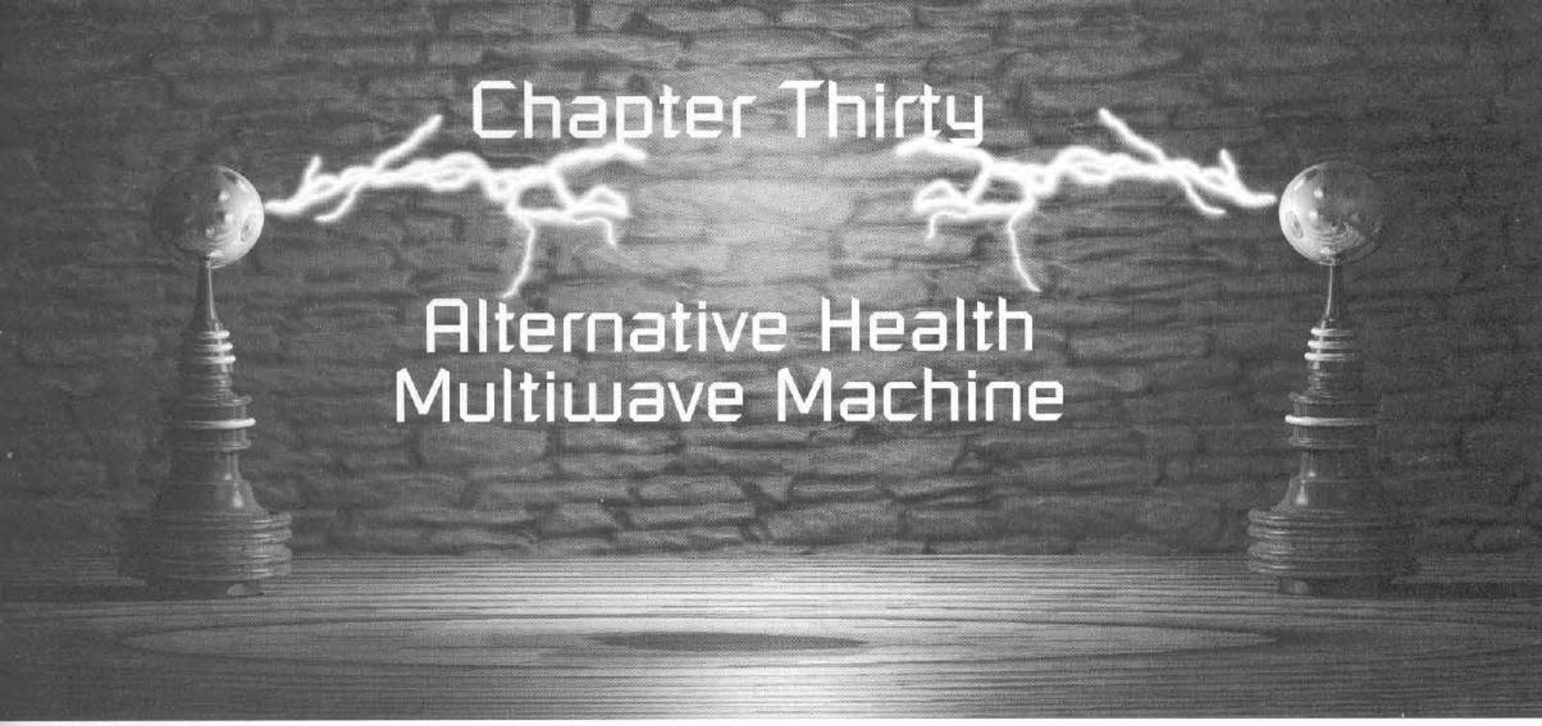

Chapter Thirty

Alternative Health Multiwave Machine

Figure 30-1 shows how to construct a Lakhovsky multiwave alternative health system. Dr. Lakhovsky was a Russian researcher who experimented with the effects of high-frequency electromagnetic radiation on various parts of the living organism. This controversial system is claimed to cure many health ills, including cancer. Its claim to fame is that all cells in a living body possess an intrinsic, vibrating resonant frequency that can be energized by an external means. Diseased cells appear to have a weaker and different frequency that can be brought back into step with the adjacent nondiseased cells. Antiaging appears to occur when cells are exposed to this full-spectrum electromagnetic vibratory energy. It is of special interest that this method of cellular regeneration has generated considerable interest in several research fields as a possible cure for cancer.

This project shows how to construct the multiwave radiating coil and antenna from readily available parts and pieces. System operation will require a working BTC30 12-inch-spark Tesla coil, as described in Chapter 14 of *Electronic Gadgets for the Evil Genius*. This Tesla coil is also available through plans, a kit, or as a complete and ready-to-use coil on the Information Unlimited Web site at amazing1.com.

It is suggested that the builder obtain the Lakhovsky handbook as listed in the parts list at the end of this chapter and study the methods of application and testimonials on the use of this controversial health machine before using it on one's self or others.

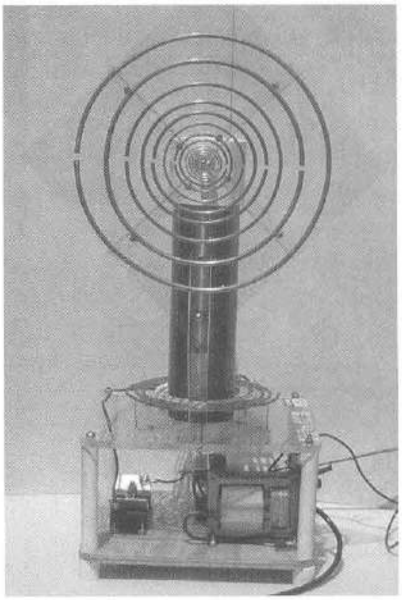

Figure 30-1 *The multiwave antenna*

Theory of Operation

Figure 30-2 shows a block diagram of the radiating coil and Tesla driver. The antenna coil contains a series of concentrically wound coil rings mounted in a flat plane. These coils all possess an inherent resonant

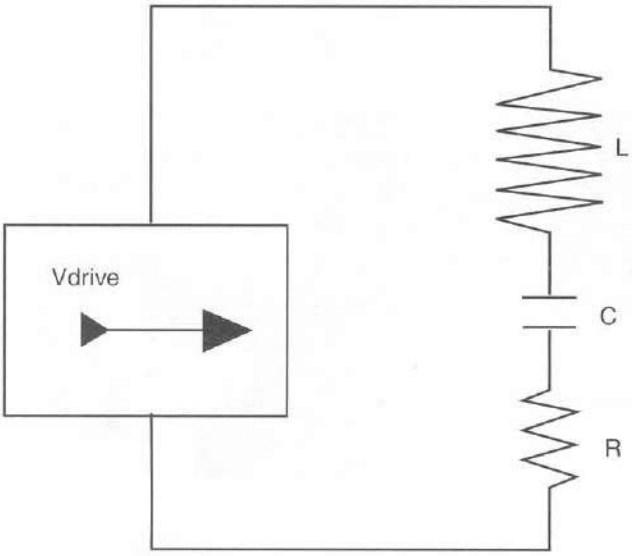

L1 to Ln represents the inductance of all of the coils.
C1 to Cn reprasents the capacity of all of the coils.
R1 to Rn is the radiation resistance of each section.
Frequency are the sum of the resonant sections ranging up to the gigahertz.

L is the sum of L1 to Ln
C is the sum of C1 to Cn
R is the sum of R1 to Rn and is mainly the radiation resistance.

The undamaged current wave from the output of the Tesla driver induces voltages at many frequencies into the multicoils of the antenna.

Figure 30-2 *Diagram of system*

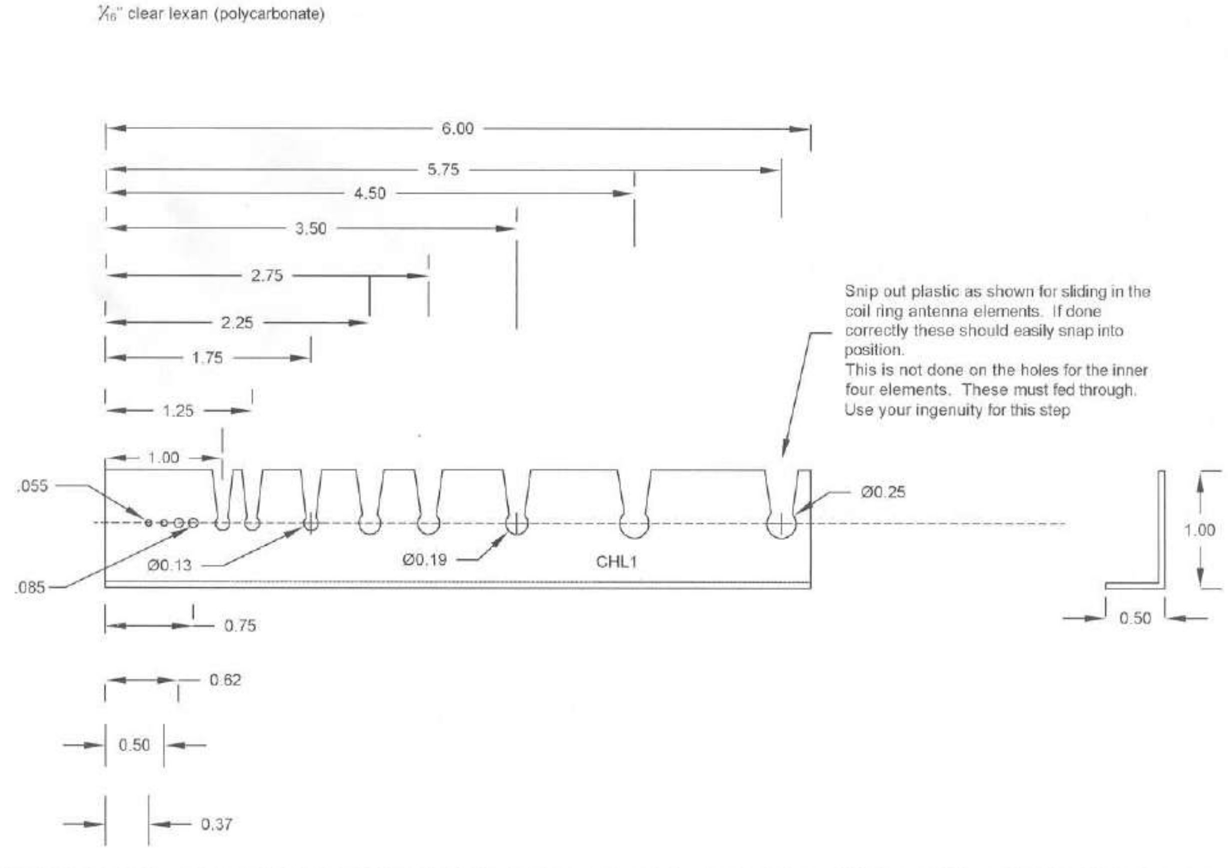
1/16" clear lexan (polycarbonate)
6.00
5.75
4.50
3.50
2.75
2.25
1.75
1.25
1.00
.055
.085
Ø0.13
Ø0.19
CHL1
Ø0.25
1.00
0.50
0.75
0.62
0.50
0.37

Snip out plastic as shown for sliding in the coil ring antenna elements. If done correctly these should easily snap into position.
This is not done on the holes for the inner four elements. These must fed through. Use your ingenuity for this step

Figure 30-3 *Fabricating the four coil holders*

frequency as a function of their physical parameters. Harmonics are also produced and with all the coils oscillating they create a near-continuous spectrum of radio-frequency energy. A spark-driven Tesla coil is required for optimum results, as it generates many fast steps of current rich in harmonic content that drives the antenna coil.

Construction of the Multiwave Coil and Antenna

1. Create the four coil holders (CHOLD1) as shown in Figure 30-3. Use 1/16-inch Lexan polycarbonate.
2. Create the 12 × 12-inch main plate (MP1) from a piece of 1/16-inch Lexan sheet, as shown in Figure 30-4.
3. Mount the coil holders as shown and carefully drill the mounting holes. Use nylon or brass screws and nuts for mounting the holders to the plate. It is very important to have these holders as exact as possible or the wound coils will not fit properly in their predrilled slots.
4. Form the circular coil rings from the appropriate diameter copper tube and snap the tubing into place. Again attempt to form circles as perfect as possible for proper operation and esthetics. The inner coil rings are formed pieces of bus wire and are threaded through their respective holes. Note a 1/8-inch gap at the coil ends.
5. Attach a PVC cap (CAP1) using a small metal screw, solder lug, and nut. This screw is the high-voltage feed point to the antenna and is connected to a piece of wire soldered to the innermost coil ring (see Figure 30-5).
6. The antenna may be used with any suitable high-voltage, high-frequency generator. It is shown in Figure 30-6 with our BTC30 250-kilovolt Tesla coil, as described in Chapter 14 of *Electronic Gadgets*. It is available as a kit or completed system as noted in Table 30-1.

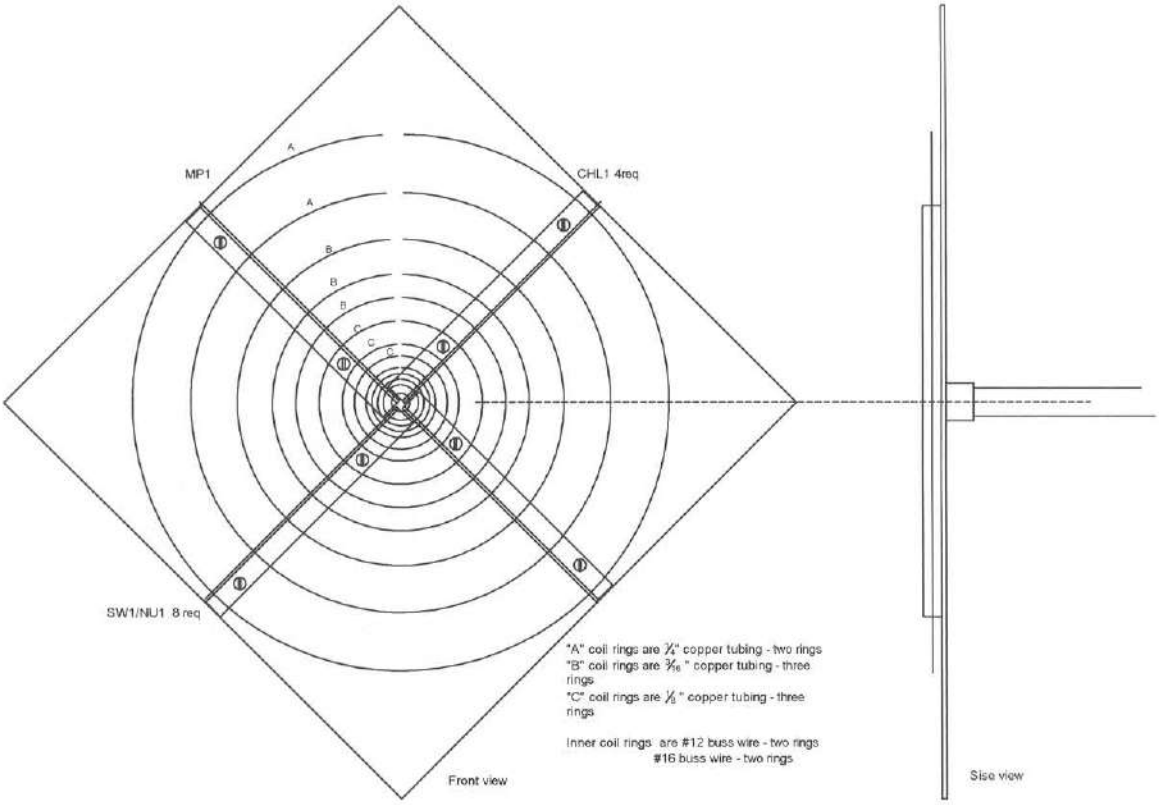
MP1
A
A
B
B
B
C
C
C
CHL1 4req
SW1/NU1 8req
"A" coil rings are 1/4" copper tubing - two rings
"B" coil rings are 3/16" copper tubing - three rings
"C" coil rings are 1/8" copper tubing - three rings
Inner coil rings are #12 buss wire - two rings
#16 buss wire - two rings
Front view
Sise view

Figure 30-4 *Creating the base*

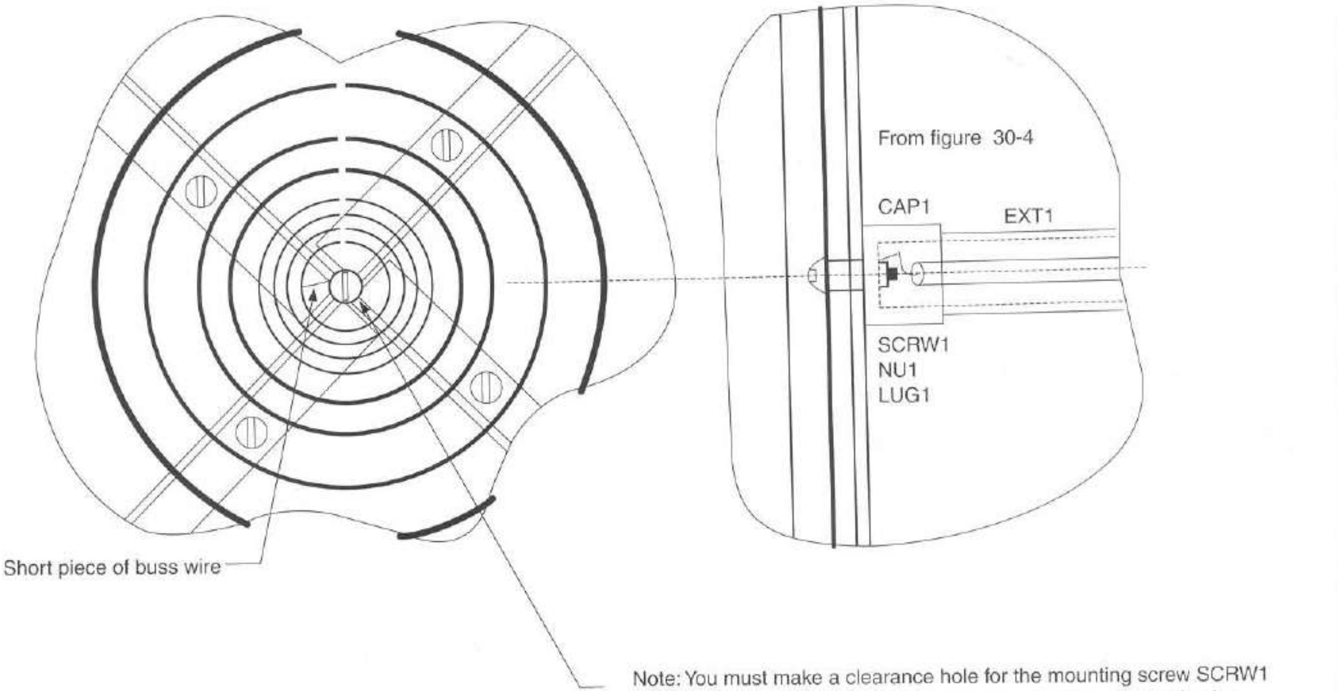

Note: You must make a clearance hole for the mounting screw SCRW1 exactly in the center. It may be necessary to heat up a suitable size diameter metal rod and carefully melt a clearance section out of the four coil holders for the head of this screw. You must now connect a short piece of buss wire from this screw to the inner most ring and solder in place.

Figure 30-5 *Mounting and feeding details*

7. If you intend to use our coil, you will need to obtain a 3-inch slip cap (SLP1) and drill a hole for the extension arm tube (EXT1). You will also need the plastic stabilizing bracket (BRK1) required to keep the coil assembly upright.
8. The connection between the output of the Tesla coil and the antenna is made via the wire sleeved through the extension tube that connects to the mounting screw and exits through the rear to make contact with the Tesla coil output lead.

Operation

The following instructions reference the use of our 250-kilovolt Tesla coil as listed in Table 30-1 or as described in Chapter 14 of *Electronic Gadgets*.

1. Preset the spark gap on the Tesla coil to no more than 1/16 of an inch. Connect the tap lead for maximum inductance.
2. Turn on the switch on the Tesla coil and note all the antenna coil rings adjacently sparking and arcing between one another. Note that the outer ring may be grounded to the chassis to increase output activity.

At this point, it is recommended that you obtain the referenced book (LSK1), the *Lakhovsky Multiple Wave Oscillator Handbook,* as referenced in the parts list and familiarize yourself with the potentials, treatments, and experiments possible with this system.

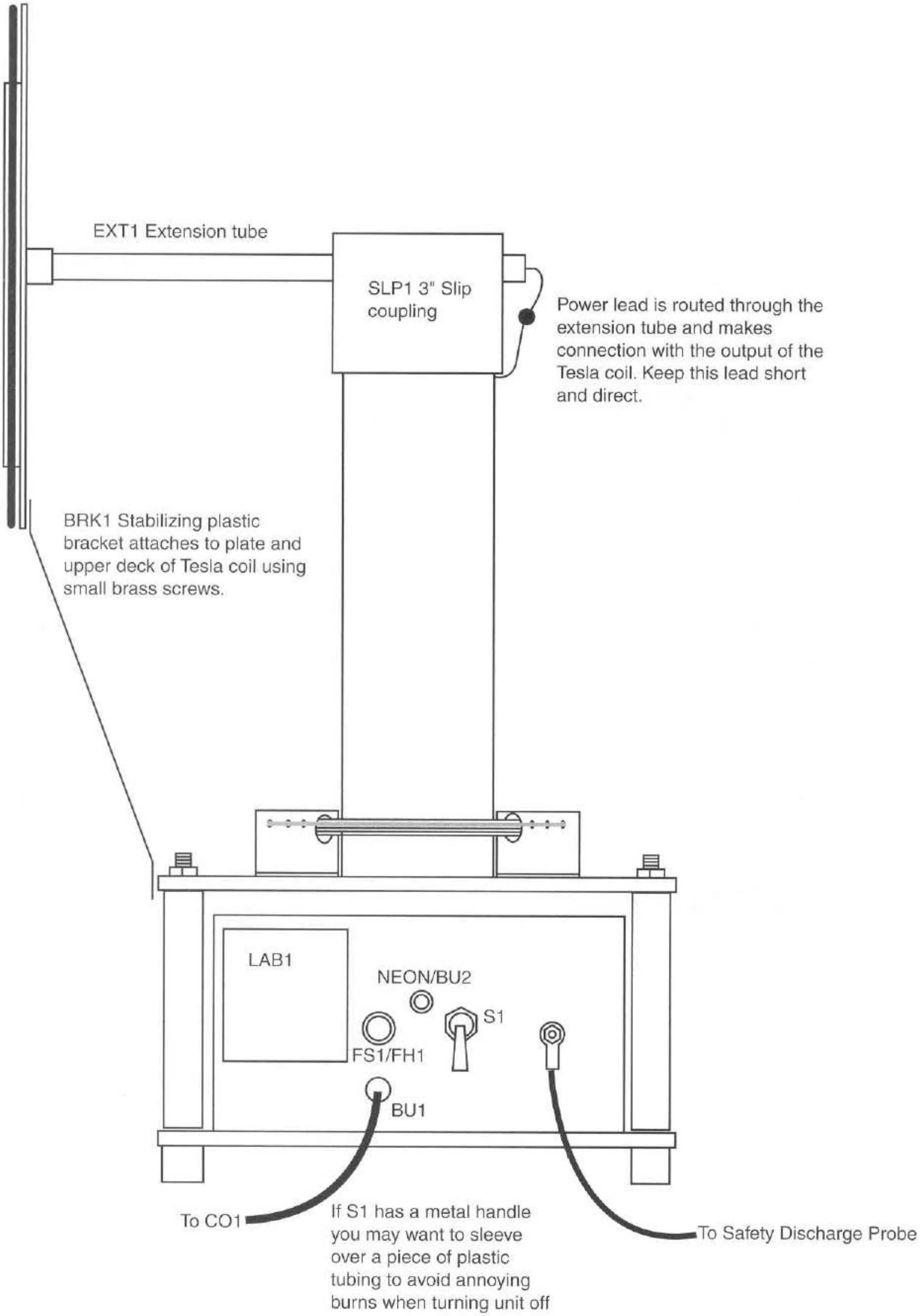
EXT1 Extension tube
SLP1 3" Slip coupling
LAB1
NEON/BU2
S1
FS1/FH1
BU1
To CO1
To Safety Discharge Probe

Power lead is routed through the extension tube and makes connection with the output of the Tesla coil. Keep this lead short and direct.

BRK1 Stabilizing plastic bracket attaches to plate and upper deck of Tesla coil using small brass screws.

If S1 has a metal handle you may want to sleeve over a piece of plastic tubing to avoid annoying burns when turning unit off

Figure 30-6 *Completed system*

Table 30-1 Multiwave Machine Parts List

Ref. #	*Description*	*DB Part #*
CHL1	Four coil-ring holders shown in Figure 30-3	DB# CHL1LAK
MP1	Mounting plate shown in Figure 30-4	
ARING	Two ¼-inch copper, soft drawn tubings	
BRING	Three 3⁄16-inch copper, soft drawn tubings	
CRING	Three ⅛-inch copper, soft drawn tubings	
INNERRING	Two #12 bus wire circular rings	
INNERRING SMALL	Two #16 bus wire circular rings	
SW1/NU1	106-32 × ½-inch brass or nylon screws and nuts	
SCRW1/NU1	6-32 × ¾-inch brass screw and nut	
LUG1	#6 solder lug	
WR20	24 inches of #20 bus wire	
BRK1	Stabilizing bracket shown in Figure 30-6	
CAP1	½-inch PVC slip cap	
SLP1	3 inches of PVC slip coupling	
EXT1	18 inches of ½-inch PVC tubing	
BTC3K	250-kilovolt Tesla coil generator kit and plans	DB# BTC3K
BTC30	250-kilovolt Tesla coil generator ready to use	DB# BTC30
LSK1	Book on Lakhovsky, 144 pages	DB# LSK1

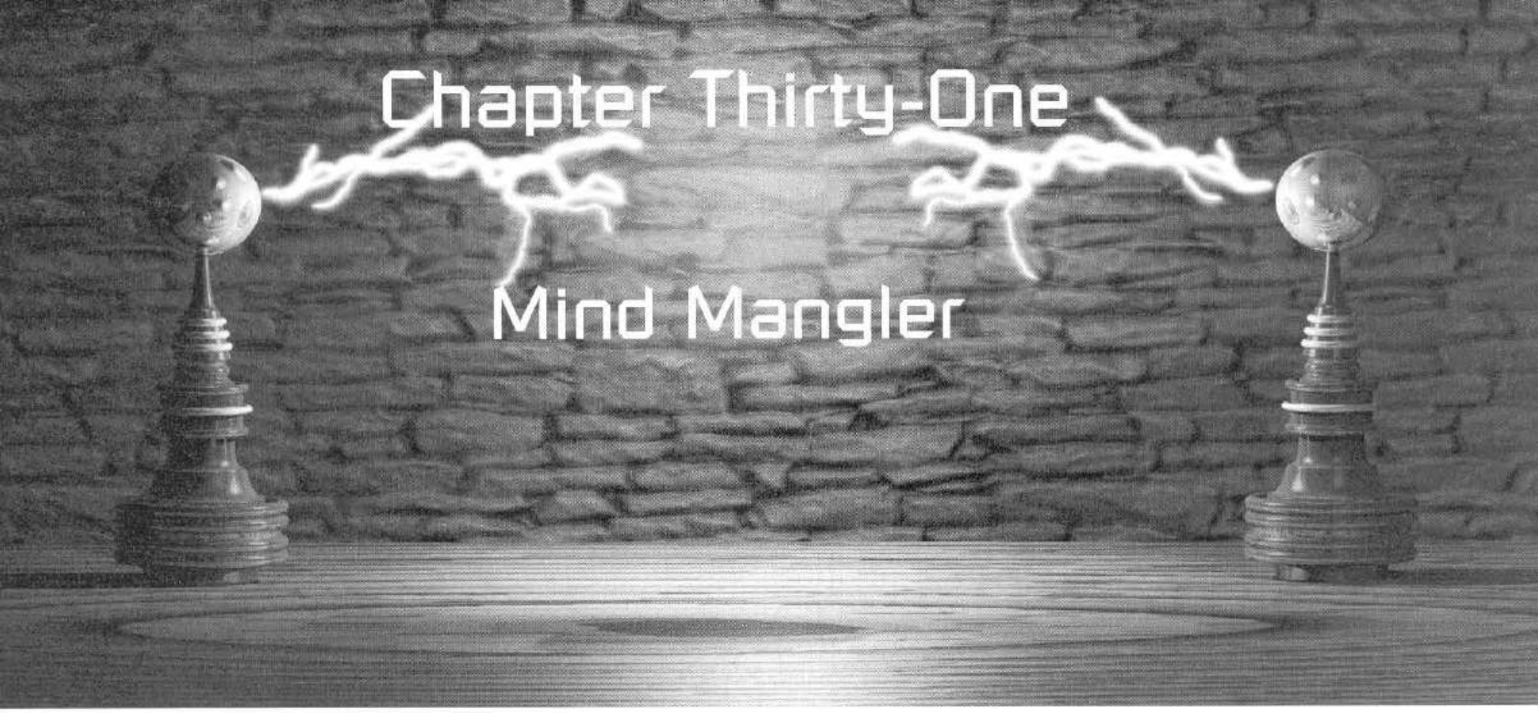

Chapter Thirty-One

Mind Mangler

This board-level device, as shown in Figure 31-1, when properly assembled, can be a great prank. When placed in one's bedroom under normal light condition, it does nothing. As soon as the lights are turned off, it comes to life, producing pulses of high-frequency sound much like an insect. These control pulses can be timed to occur about every minute. When the unsuspecting victim attempts to locate it by turning on the lights, it ceases operation. Even using a flashlight will disable it. The long time between pulses also makes it very difficult to detect. Properly hidden, it can take a long time to locate and may turn a prank into a nasty situation.

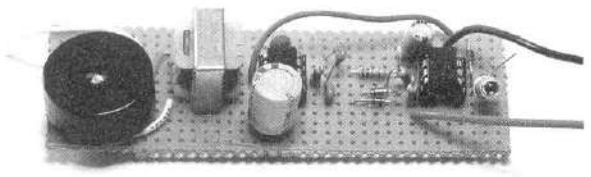

Figure 31-1 *Board-level view of the mind mangler*

Circuit Operation

Figure 31-2 shows a timer, I1, as a free-running, astable pulse generator producing a symmetrical 10-second, off and on pulse controlling the second timer, I2. Timer I1 has a phototransistor Q1 connected in series with the reservoir capacitor C1. As long as there is light on Q1, C1 can never discharge due to the conductance of Q1. Therefore, pin 3 is at a low and I2 is disabled. When the lights go out, C1 now is allowed to discharge and pin 3 goes high, turning on I2 and producing the high-frequency sound to the output transducer TD1. The high-audio frequency is determined by resistor R3 and C3, and can be changed by altering these values. The output signal from I2 is stepped up by transformer T1 (see the parts list at the end of the chapter).

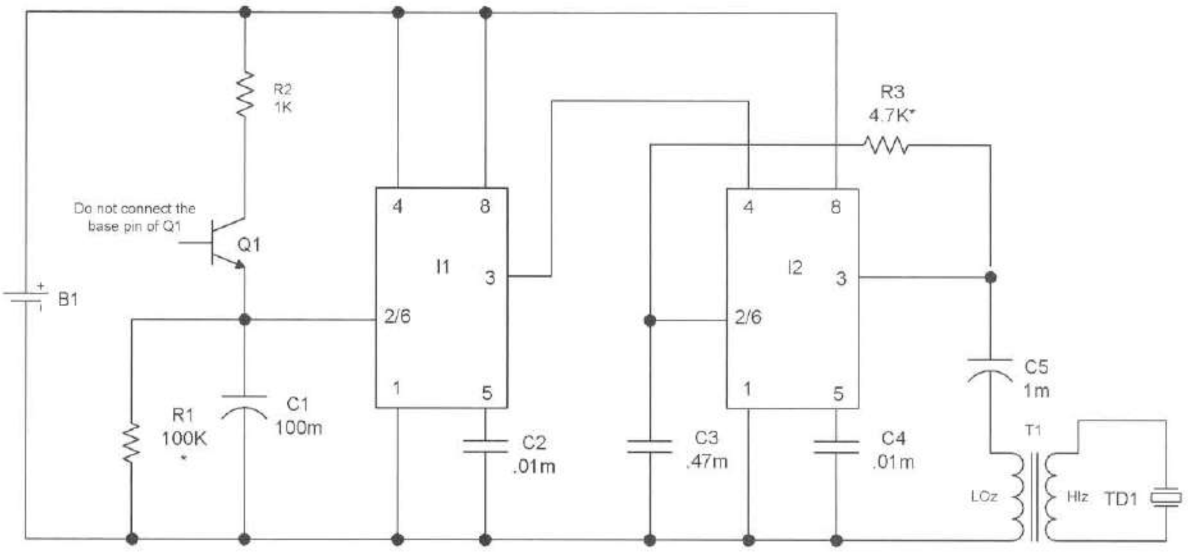
R2
1K
Do not connect the base pin of Q1
Q1
B1
R1
100K
*
C1
100m
4
8
I1
3
2/6
1
5
C2
.01m
R3
4.7K*
4
8
I2
3
2/6
1
5
C3
.47m
C4
.01m
C5
1m
T1
LOz
HIz
TD1

*The value of R1 May be changed to adjust for dark and light conditions. You may want to replace with a 500 K trimpot

Figure 31-2 *Circuit schematic*

Construction Steps

1. Identify all the parts and pieces and verify them with the bill of materials.
2. Insert the components starting from one end of the perforated circuit board and follow the locations shown in Figure 31-3 using the individual holes as guides. Use the leads of the actual components as the connection runs, which are indicated by the dashed lines. It is a good idea to trial-fit the larger parts before actually starting to solder. Always avoid bare wire bridges, globby solder joints, and potential solder shorts. Also check for cold or loose solder joints.

Additionally, pay attention to the polarity of the capacitors with polarity signs and all the semiconductors. The transformer position is determined using an ohmmeter.

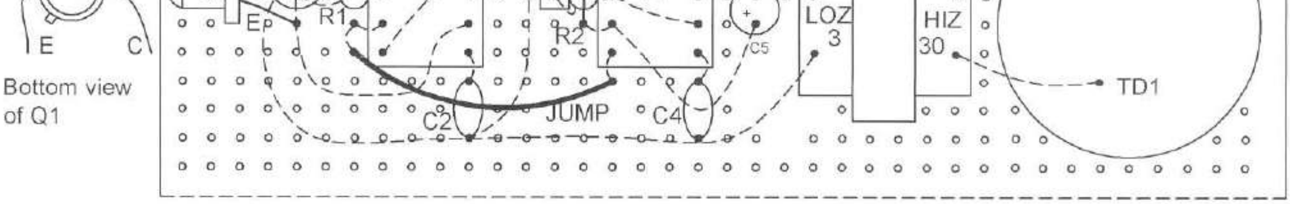
CL1 BATTERY CLIP
+
NC
E
C
Bottom view of Q1
Q1
Ep
C1
R2
C1
R1
I1
C3
R2
I2
C5
LOZ
3
T1
HIZ
30
TD1
C2
JUMP
C4

You can determine the low impedance side of T1 with an ohmeter. It will always read as the lower of the two

Figure 31-3 *Assembly board*

Testing

Verify the wiring accuracy and connection. Insert a 9-volt battery into the clip, cover the lens of Q1 or turn off the lights, and note the unit emitting a low-level, raspy, clicking tone like a sick cricket. Uncover the lens of Q1 and note the circuit shutting down. There will be a delay that is dependent on the level of light and dark. You may vary the response by changing the value of R1. Decreasing its value will require more light to shut down and vice versa.

Table 31-1 Mind Mangler Part List

Ref. #	*Description*	*DB Part #*
R1	100K, ¼-watt resistor (br-blk-yel)	
R2	1K, ¼-watt resistor (br-blk-yel)	
R3	4.7K, ¼-watt resistor (yel-pur-red)	
C1	100-microfarad, 25-volt vertical electrolytic capacitor	
C2, 4	Two .01-microfarad, 50-volt plastic capacitors (103)	
C3	.47-microfarad, 50-volt vertical electrolytic capacitor	
C5	1-microfarad, 50-volt vertical electrolytic capacitor	
I1, 2	Two 555 timer dual inline packages (DIP)	
Q1	L14P phototransistor	
CL1	9-volt battery clip	
TD1	Small piezo transducer	
T1	Small audio transformer, 8 to 1K	
PB1	4 × 1.5 × .1-inch grid perforated circuit board	

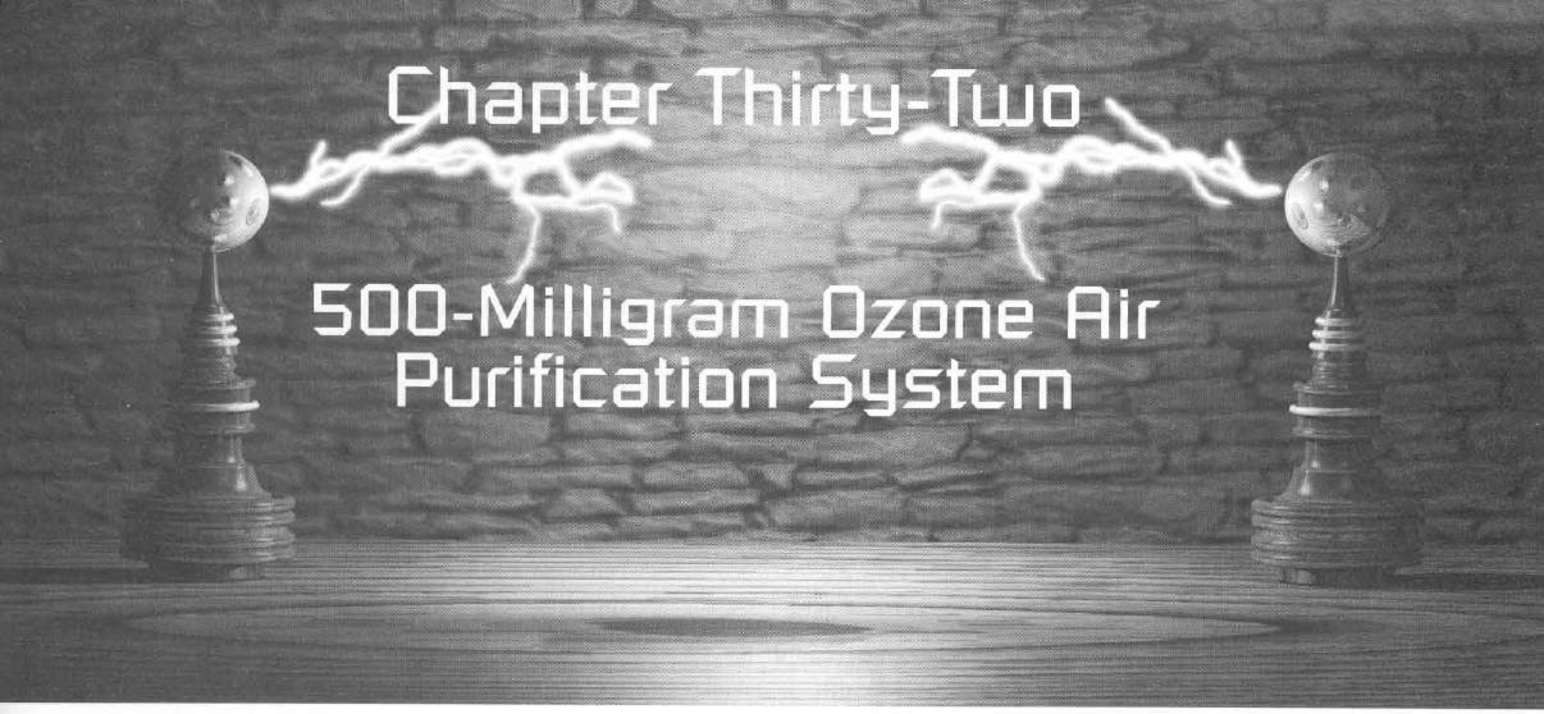

Chapter Thirty-Two

500-Milligram Ozone Air Purification System

Figure 32-1 shows a useful home device that generates ozone for eliminating odors, killing bacteria and mold, and disinfecting unsanitary areas such as pet boxes and bathrooms.

This project shows how you can build a system that could cost over $300 for less than $50. Its simple and basic design uses two 12-volt fans to blow air over a unique ozone-producing cell that is powered by a variable-output, high-voltage, high-frequency power supply. The power supply is featured in Chapter 7 of this book and is used in several other projects.

This is a beginner- to intermediate-level project requiring minimal basic electronic skills. Expect to spend $40 to $60. All the parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in Tables 32-1 and 7-1 at the end of this chapter and Chapter 7.

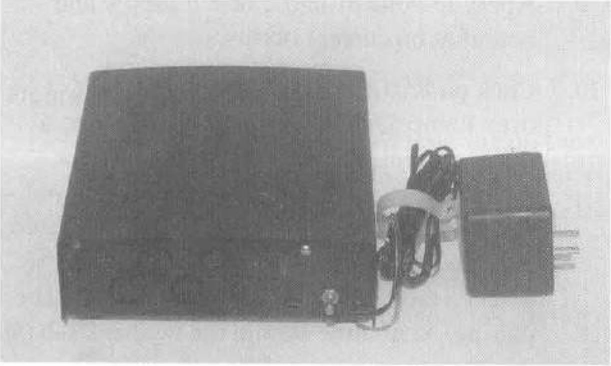

Figure 32-1 *Ozone purification system*

The Story of Ozone

Ozone is an unstable form of oxygen. Normal oxygen is diatomic (0_2), existing as two atoms of oxygen making up the molecule. Ozone is triatomic (0_3), existing as three atoms of oxygen for the molecule. The triatomic form of oxygen is very unstable, wanting to lose the third oxygen atom and combine with whatever atom it can (oxidization). This property makes it the most active oxidizer known, with the exception of the very hazardous fluorine gas. Ozone at normal pressures is colorless and odorless, yet it produces a pleasing, fresh-air odor as a result of the nitrous oxides it produces. The odor is usually noticeable after an old-fashioned thunderstorm. Under pressure, it becomes a bluish gas.

Ozone is also a very powerful bactericide. It is not affected by pH, as is chlorine, thus making it an excellent candidate for pools, spas, laundries, and general water treatment applications. It is many times more soluble in water, further enhancing its purifying effect. Ozone will combine with diatomic nitrogen, N_2, forming nitrous oxide, 2NO. This oxide quickly combines with water, forming nitric acid, HNO_3. This is often a very undesirable effect when used with straight air. Pure oxygen greatly minimizes this effect and is often required in many applications.

However, those requiring a supply of concentrated nitric acid for the nitration of many carbohydrates, such as cellulose, glycerine, hexamine, phenol, and toluene, may wish to consider combining ozone and air with a condensing apparatus to obtain this useful acid.

Ozone for Treating Air

Air pollution is one of the most serious environmental issues that we face. Visible and detectable smoke, dust, mildew, mold, and toxic odors, bacteria, pollen, and static electricity, along with the more elusive and invisible chemicals, have become a serious health threat. As buildings and homes are constructed with tighter air sealants for energy conservation, the problem will become even more severe. Often the effects from these problems include burning eyes, headaches, dizziness, depression, allergies, and general lethargy, all of which are usually attributed to colds and viruses. Air filters, such as the overpriced Sharper Image unit currently being advertised, are only partially effective, with some producing undesirable positive ions.

Ozone purifies the air from these undesirable pollutants by oxidization, as most pollutants readily combine with this highly reactive tri-atomic form of oxygen and break down into water and other nontoxic compounds. Ozone is produced artificially by electricity and is many times more antiseptically effective than oxygen.

Construction Steps

1. The system is shown using the high-voltage, high-frequency driver shown in Chapter 7. It is built on a printed circuit board or a more difficult perforated board. The PCB version is the easiest, as construction involves placing the correct components into the correct holes and soldering. The more difficult approach is also shown and uses a perforated circuit board. This is far more challenging and is intended for the experienced assembler. Figure 32-2 shows the board with the necessary changes for use with this circuit.
2. Create the base section, as shown in Figure 32-3. You can use sheet aluminum or bendable plastic sheets. Design a mating cover with a slight overhang over the front and rear panels. Use similar material and mount using #6 sheet metal screws along the lip of the base.
3. Fabricate the fan and cell mounting bracket as shown in Figure 32-4. Use clear 1/16-inch polycarbonate plastic and initially layout the piece, marking off the bend and cutting lines. You may fabricate separate pieces for the cell brackets and fans if you do not have access to the necessary equipment. Note dimensions and cuts must result in a single piece, as shown.
4. Cut out a piece of window screen large enough to cover the input air holes and secure it from the inside of the base enclosure using screws and silicon rubber cement.
5. Mechanically assemble the components, as shown in Figure 32-5. Verify all proper fittings and the clearance of components, especially the high-voltage points on the cell. Finally, wire everything as shown and verify the accuracy and any potential errors. Note: Figure 32-6 shows the wiring using the perforated board approach. Do not connect the 12 DC/1.5 adapter wall transformer at this time.
6. Obtain a 12-volt, 3-amp bench power supply with a voltage and current meter. Connect it to where the adapter's plus and minus wires go.
7. Connect a scope to the test points, as shown in Figure 32-5.
8. Preset the trimpot R1 to midrange and control pot R10 to full counterclockwise and click off.
9. Apply 12 volts from the bench supply and note that no current occurs.
10. Click on R10 and note a pulsing current of over 1 amp. Quickly adjust the trimpot to a dip in current, as noted on the meter. Our model was tuned correctly when the trimpot was set to one to two o'clock. The fans should both be in full operation and you should detect a faint, bluish glow coinciding with the pulsing. At no time should the heatsink tab on the power transistor Q1 be hot.

It will be necessary to construct the printed circuit assembly board as shown in chapter 7 with the following below list of revisions.

Wind L1 using 10 turns of #22 magnet wire. Air gap core with .007 (7mils) shims. Measured inductance should be between 15 -16 micro-henries.

Trimpot R6 is not used

Replace R5 with a wire jump

Replace C2 from a .0022 to a .0047 microfarad capacitor

Replace C6 from a .22 to a .47 microfarad 250 volt capacitor

Capacitor C4 is not used. insert a short piece of buss wire for a test point ground

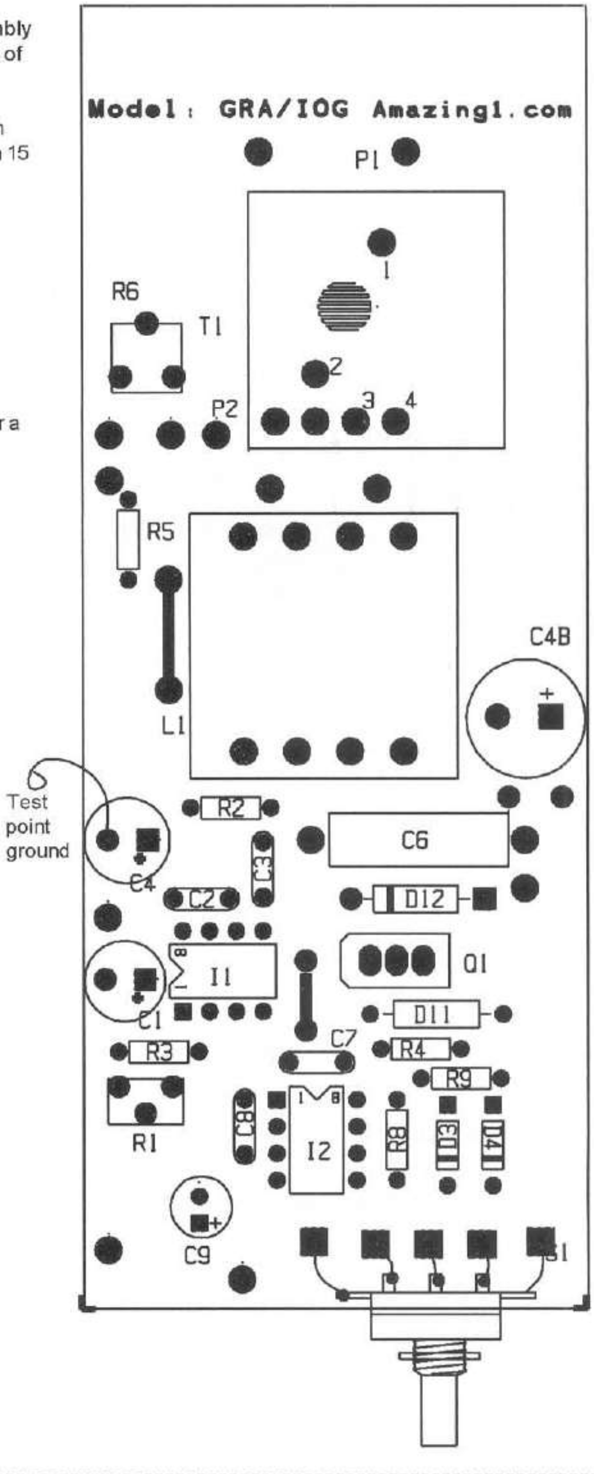
Model : GRA/IOG Amazing1.com
P1
R6
T1
1
2
3
4
P2
R5
L1
C4B
+
Test point ground
C4
R2
C3
C6
C2
D12
Q1
8
I1
C1
D11
R3
C7
R4
R9
1
8
R1
C8
I2
R8
D13
D14
+
C9
S1

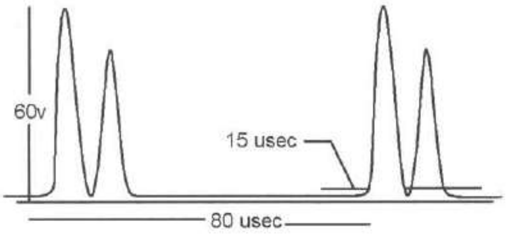

Correct wave shape measured at the drain of Q1 and test ground point

Figure 32-2 *PCB rework*

11. Turn R10 full clockwise and note the on time of the pulse increasing with the current but not exceeding 1.5 amps at these pulse peaks. Allow it to operate for 30 minutes and note ozone being produced, a soft, bluish, even glow on the cell, and the heatsink not getting too hot to touch.
12. Once this is verified, you can hook in the 12 VDC/1.5-amp adapter transformer. Allow it to operate and verify that the adapter does not get hot.
13. Attach the cover and place the unit in the target area.

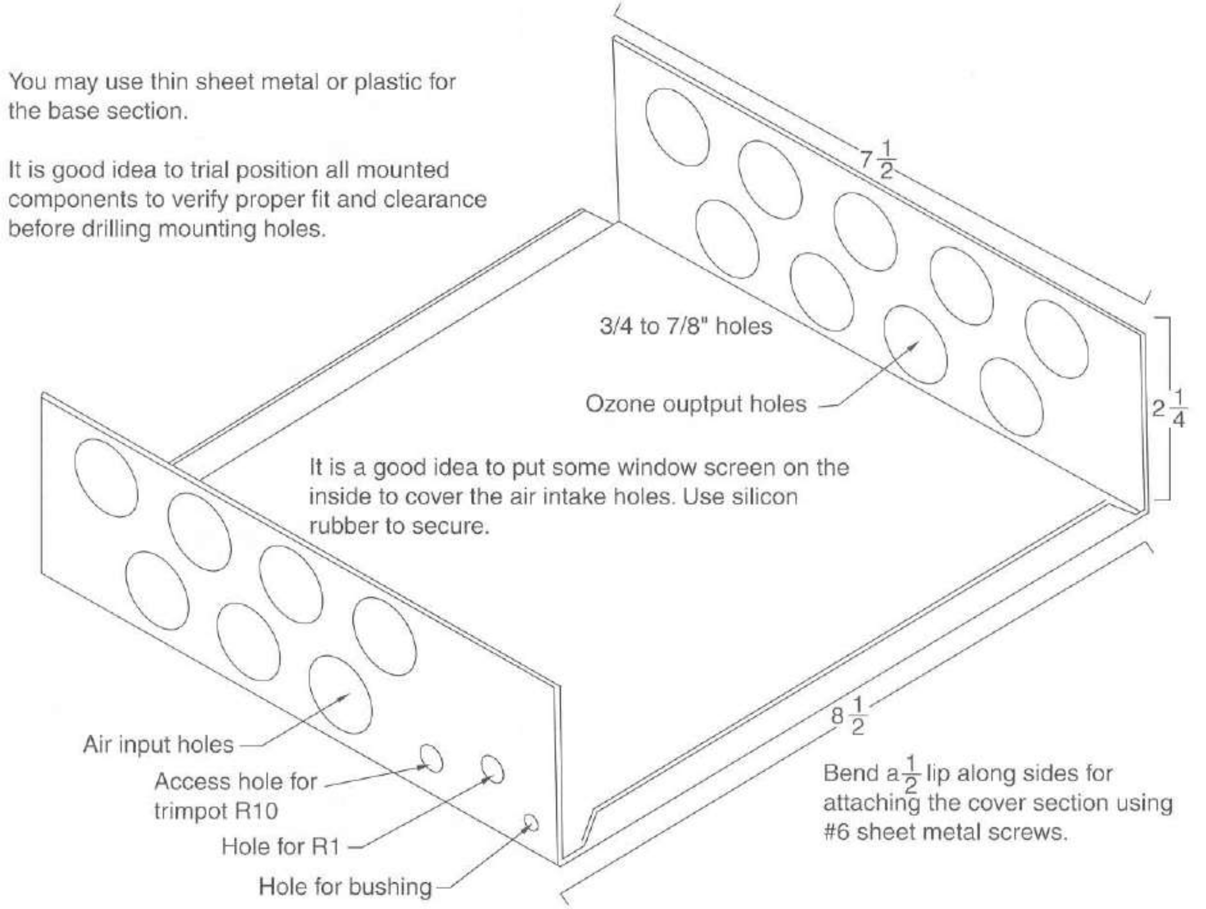
You may use thin sheet metal or plastic for the base section.
It is good idea to trial position all mounted components to verify proper fit and clearance before drilling mounting holes.
7 1/2
3/4 to 7/8" holes
Ozone ouptput holes
2 1/4
It is a good idea to put some window screen on the inside to cover the air intake holes. Use silicon rubber to secure.
8 1/2
Bend a 1/2 lip along sides for attaching the cover section using #6 sheet metal screws.
Air input holes
Access hole for trimpot R10
Hole for R1
Hole for bushing

Figure 32-3 *Base section creation*

14. Place the unit with the output directed toward the center of the area.
15. Plug in the wall adapter and rotate the control switch to on. Note the fan rotating and freely turning without obstructions. It should produce a good air flow.
16. Adjust the control a full clockwise turn and note a smell of fresh air emitting from output ports. You will observe a purplish glow around the cell when viewed in darkness.
17. Allow it to run for several hours and adjust the control to where the fresh air smell is just detectable.
18. Check both the unit and the wall adapter for excessive heating. They should only be warm to the touch.
19. This unit may need cleaning from dust. Use compressed dry air or a soft cleaning brush.

Special Notes

You should never operate the unit when the odor becomes pungent or domineering. Ozone is a colorless, odorless gas composed of unstable diatomic oxygen. The smell of fresh air is not the ozone, but the result of secondary reactions with other chemicals.

The unit can be adapted to 12 VDC operation for vehicle or battery use. Simply remove the adapter and splice in some 12-volt connections. Use caution when observing the polarity. You may request a factory modification for this step.

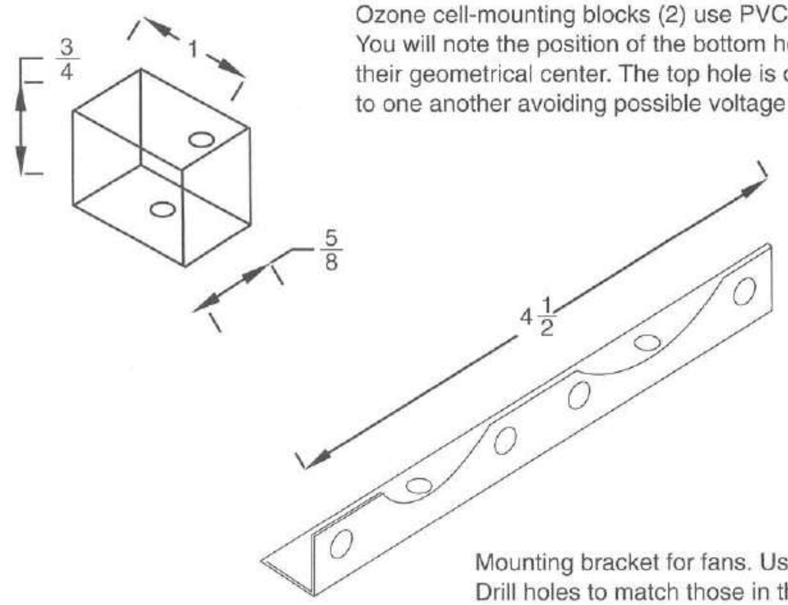

Ozone cell-mounting blocks (2) use PVC or another suitable insulting plastic. You will note the position of the bottom hole mounting the blocks of the base is at their geometrical center. The top hole is offset to this location so as not to be adjacent to one another avoiding possible voltage breakdown.

Mounting bracket for fans. Use $\frac{1}{32}$ aluminum or plastic. Drill holes to match those in the fan assemblies.

Figure 32-4 *Fabricated parts*

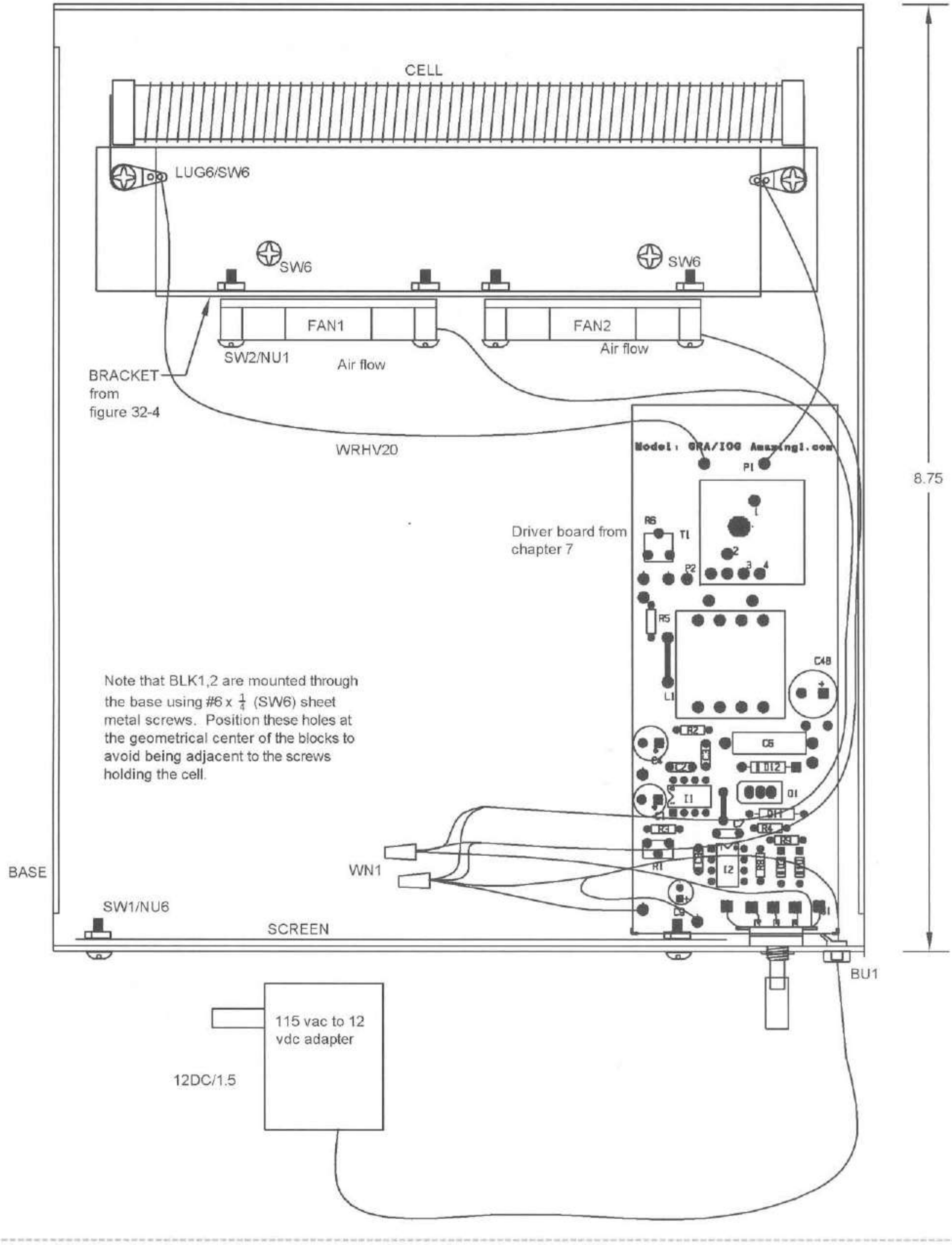
CELL
LUG6/SW6
SW6
SW6
FAN1
FAN2
SW2/NU1
Air flow
Air flow
BRACKET from figure 32-4
WRHV20
Model: GRA/IOG Amazing1.com
P1
Driver board from chapter 7
R6
T1
P2
R5
L1
C4B
R2
C6
C4
I1
I2
Q1
R4
R3
R1
C9
8.75
BASE
SW1/NU6
SCREEN
WN1
BU1
115 vac to 12 vdc adapter
12DC/1.5

Note that BLK1,2 are mounted through the base using #6 x $\frac{1}{4}$ (SW6) sheet metal screws. Position these holes at the geometrical center of the blocks to avoid being adjacent to the screws holding the cell.

Figure 32-5 *Mechanical layout showing wiring when using the PCB*

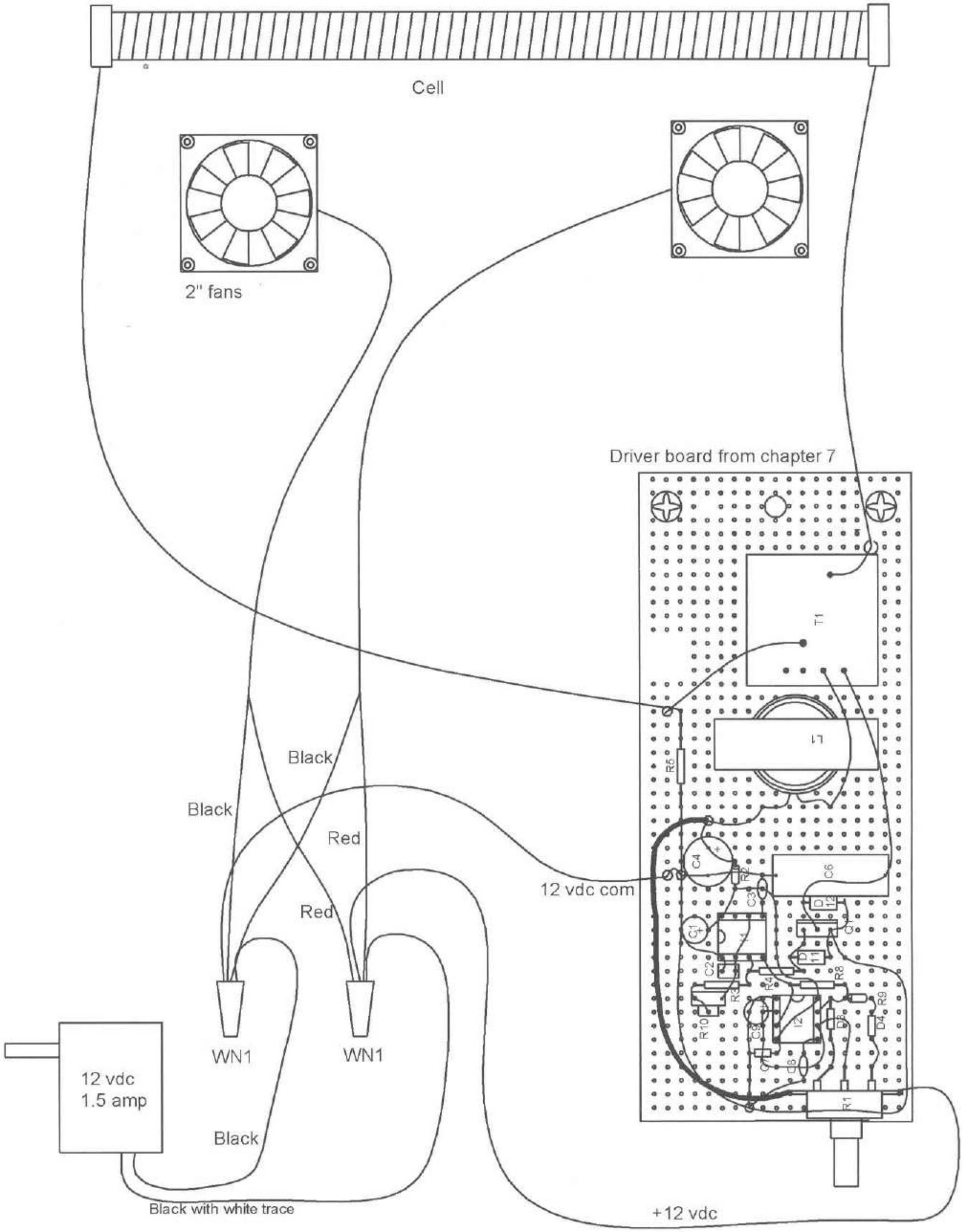
Cell
2" fans
Driver board from chapter 7
T1
L1
R6
Black
Black
Red
Red
12 vdc com
C4
R2
C3
C6
D12
Q1
C1
I1
D11
C2
R3
R4
R8
R10
C9
I2
D3
R9
D4
C7
C8
R1
WN1
WN1
12 vdc
1.5 amp
Black
Black with white trace
+12 vdc

Figure 32-6 *Final wiring using a perforated board*

Table 32-1 500-Milligram Ozone Machine Parts List

Ref. #	*Description*	*DB Part #*
DRIVER	Modified high-voltage assembly from Chapter 7	DB# GRADRIV10
BASE	Base section fabricated per Figure 32-3	
COVER	Fabricated to fit BASE section	
BLOCKS	Two blocks fabricated per Figure 32-4	
BRACKET	Bracket fabricated per Figure 32-4	
FAN1, 2	Two small, 2 × 2-inch, 12-volt fans	DB# FAN2
CELL500	500-milligram ozone corona cell	DB# CELL500
SCREEN	2 × 5-inch piece of window screen	
12DC/1.5	12-volt, 1.5-amp wall adapter transformer	DB# 12 DC/1.5
WRHV20	6-inch, 20-kilovolt, high-voltage wire	
WN1,2	Two small wire nuts	
BU1	Small Heyco bushing #2P-4	
SW6	10 #6 × ¼-inch sheet metal screws for cover and blocks	
SW2/NU2	Four #4-40 × ¾-inch machine screws and nuts for fans	
SW1/NU1	Two #6-32 × ½-inch machine screws	
LUG6	Two #6 solder lugs	
WR22	12 inches of #22 vinyl hookup wire	
FEET	Four stick-on rubber feet for base section	

Chapter Thirty-Three

Invisible Pain-Field Generator

The invisible pain-field generator device shown Figure 33-1 is a handheld, battery-operated sonic, shock-wave generator that produces a *sound pressure level* (SPL) of up to 125 decibels (db) at 30 centimeters. The shock wave frequencies are user presettable at 25, 16, and 12 KHz. A sweep function is included in the circuitry where the selected center frequency varies between two set limits at an adjustable rate, providing a complex sonic signal that further enhances the effect.

Applications of the device can range from the routing out of agricultural pests in silage bins, chicken houses, grain bins, or wherever rats are a problem. The optimum frequency for this particular application is 16 KHz or just above that of human hearing. Many farmers use these devices for determining the effectiveness of pesticides, noting the reduction in infestation levels. Other applications include spooking birds and animals from unwanted areas and discouraging dogs and deer from decimating shrubs and ornamentals.

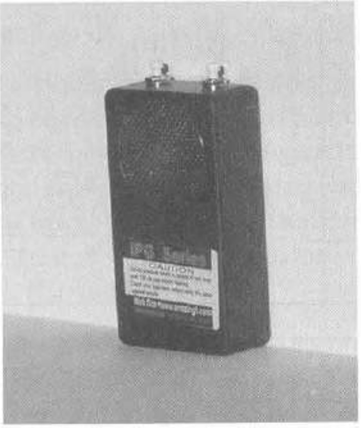

Figure 33-1 *Pain-field generator*

Adjustment of the frequency down to 12 KHz (within the range of human hearing) will produce extremely painful and annoying affects. The unit is excellent for disbursing crowds of potentially unruly people. A simple test is to point the device at an unsuspecting subject and momentarily push the button. You will notice a very positive response. Unfortunately maximum affect seems to favor younger women; older men seem to be less sensitive.

The unit is very directional when used outside, but it looses this property when used inside due to reflections from walls, ceilings, and furniture.

The unit is shown assembled into a small plastic enclosure that includes the electronics emitter transducer and the 9-volt battery. A removable plastic cover allows access in order to change the battery. Controls include a pushbutton emission control and sweep-activation pushbutton switches located on the top of the enclosure. The size of the main enclosure is 4 × 2 × 1 inches. The weight is only about 6 ounces with batteries installed

The unit can be used as a research tool for producing an effect on certain animals for their control and experimentation. The device has been successfully used for controlling certain dogs or other vicious animals by joggers or other outdoors enthusiasts. The unit has also been used in flushing out rats.

Caution: Caution must be used, as the effect on most people causes pain, headache, nausea, and extreme irritability. (Younger women are especially affected.) Do not, under any circumstances, point the unit at a person's ears or head at close range, as severe discomfort and possible ear damage may result. This also applies to dogs and other animals.

Include a caution label on the device to avoid exposure over 105 db for any continuous period of time. When using the device around people, you must be careful, as unjustified harassment is illegal and can result in prosecution.

This is an intermediate-level project requiring basic electronic skills. Expect to spend $25 to $50. All parts are readily available, and any specialized parts are available through Information Unlimited (www.amazing1.com) and are listed in Table 33-1 at the end of the chapter.

Circuit Description

A timer (IC2) is connected as an astable free-running multivibrator whose frequency is internally controlled by trimpot (R9). Resistor R8 selects the range limit of R9. The square-wave output of IC2 is via pin 3 and is directly coupled to power amplifier Q2. The drain of Q2 is DC biased through audio-frequency blocking chokes (L1A and L1B), providing a high impedance to the AC component of the signal.

DIP switches (S1,2,3) preselect the internal frequency for the required application and are shown in Figure 33-2. These switches select resonating inductors L2 and L3 connected in series with the output transducer TD1. The resonant action between the inherent capacity of TD1 and these selected inductances now produces a sinusoidal-shaped wave

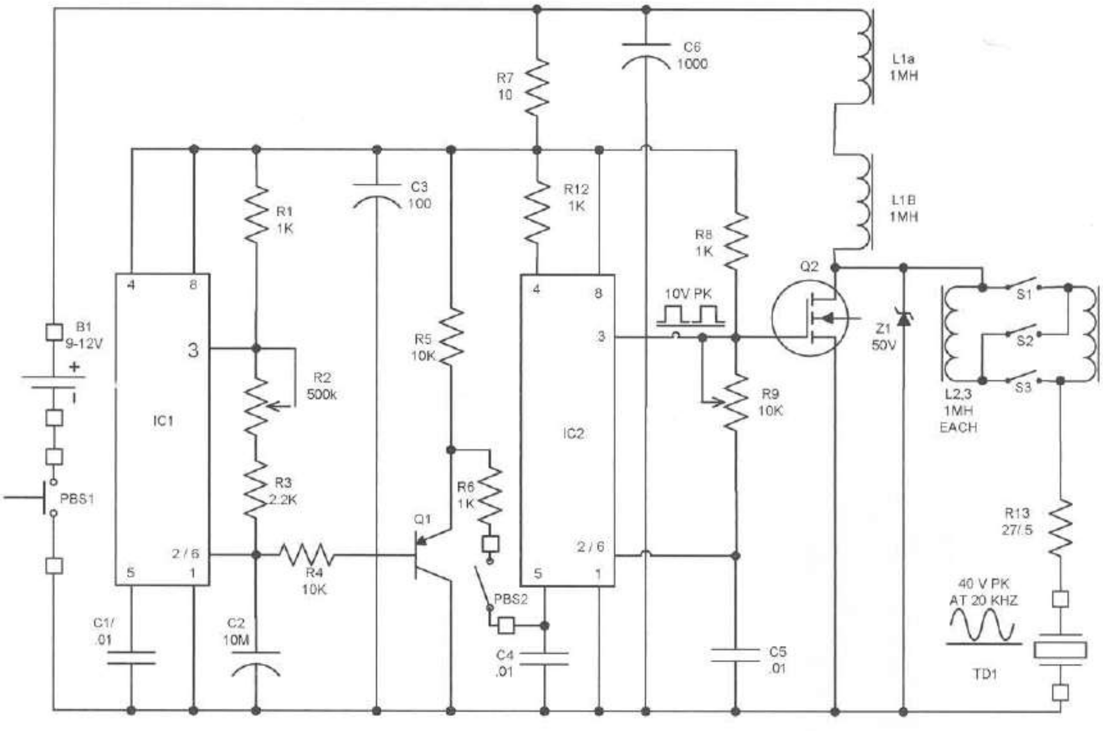

☐ These symbols represent external connection to off board components

ma	freq	1	2	3	usec
225	25khz	on	off	on	tune max 40-50 u
175	16khz	on	off	off	tune max 60-70 u
150	12khz	off	on	off	tune max 80-90u

Figure 33-2 *Pain-field generator circuit schematic*

peaking around the upper limit of the tuning range. This *signal* waveform now has a peak-to-peak voltage several times that of the *original* square wave. Transducer TD1 now can take advantage of these peak voltages to produce the high sound-pressure levels necessary without exceeding the high RMS ratings of an equivalent voltage-level square wave. Zener diode Z1 clips any excessive overshoots across Q2.

Timer IC1 is similarly connected as an astable running multivibrator and is used to produce the sweeping voltage necessary for modulating the frequency of IC2. This sweep repetition rate is controlled by trimpot R2. Resistor R3 limits the range of this repetition time. C2 sets the sweep time range. Output for IC1 is via pins 6 and 2, where the signal ramp function voltage is resistively coupled to inverter transistor Q1 via resistor R6. The output of Q1 is fed to pin 5 of IC2 and provides the output modulation voltage necessary to vary the frequency as required. Note that the modulation signal is enabled by pushbutton PBS2.

Power to the system is via battery B1 and pushbutton PBSl. Capacitor C6 guarantees an AC return path for the output signal. Power to the driver circuits IC1 and IC2 is thru a decoupling network consisting of resistor R7 and capacitor C3.

Construction

1. Layout and identify all parts and pieces. Verify with parts list. Separate resistors, as they have a color code to determine their value. Colors are noted on the parts list.
2. If you are building from a perforated board, use Figure 33-3 as a parts placement guide and insert components starting in the lower left-hand corner as shown. Pay attention to polarity of capacitors and all semiconductors. It is suggested you use sockets for integrated circuits IC1 and IC2.

   Route leads of components, as shown, and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 buss wire. Follow dashed lines on assembly drawing, as these indicate connection runs on underside of assembly board. The heavy dashed lines indicate use of thicker #20 buss wire, as this is a high-current discharge path.

   Please note that this circuit is very cramped, and it is suggested you obtain the optional printed circuit board shown in the parts list.
3. Attach external leads to components as shown in Figure 33-4, noting the individual lengths and twisted pairs.
4. Double-check accuracy of wiring and quality of solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
5. Connect an ohm meter between CLI contacts and pushbutton PBSl. Note a reading of several thousand ohms. This may vary but should not indicate a short circuit. Preset trimpots R2 and R7 to midrange.
6. Connect a 9-volt battery and note a current of 200 to 250 milliamps when R9 is set to midrange. The meter can be connected across PBSl for this step.
7. Simultaneously press both PBS2 and PBSI and note a piercing sweeping tone coming from the transducer TD1. Turn R9 fully ccw and note the decrease in signal tone.
8. Frequency range should be approximately 10 to 25 KHz, with a sweep of approximately 1 to 20 times per second selectable by trimpot R2.
9. Check the waveshapes as shown at pin 3 of the output timer IC2. This is determined by the following formula: $F = 1/(1.57 \times R9 \times C5)$.
10. Check the waveshapes as shown across transducer TD1. This waveshape is approximately 40 to 50 volts, peak to peak, and approaches a sine wave. This is the approximate resonant frequency point of the selected values of series inductances within the inherent capacitance of the transducer. A voltage-peaking effect will be noted, as the frequency is varied by trimpot R9. As the frequency is varied off of resonance, this wave will severely distort but will not damage the circuitry. Note the chart in Figure 33-2 showing the selection

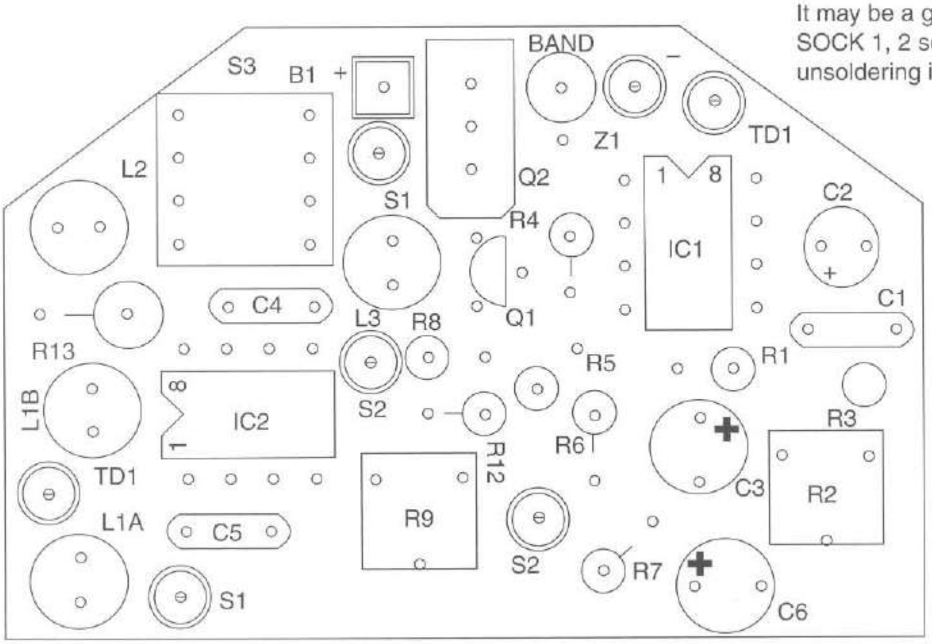
S3
B1
+
BAND
–
TD1
Z1
L2
Q2
S1
R4
1
8
IC1
C2
+
C4
C1
L3
R8
Q1
R13
R5
R1
8
L1B
IC2
1
S2
R12
R6
R3
+
TD1
C3
R2
L1A
C5
R9
S2
R7
+
C6
S1

It may be a good idea to use SOCK 1, 2 sockets for IC1, 2 to avoid unsoldering if you make an error.

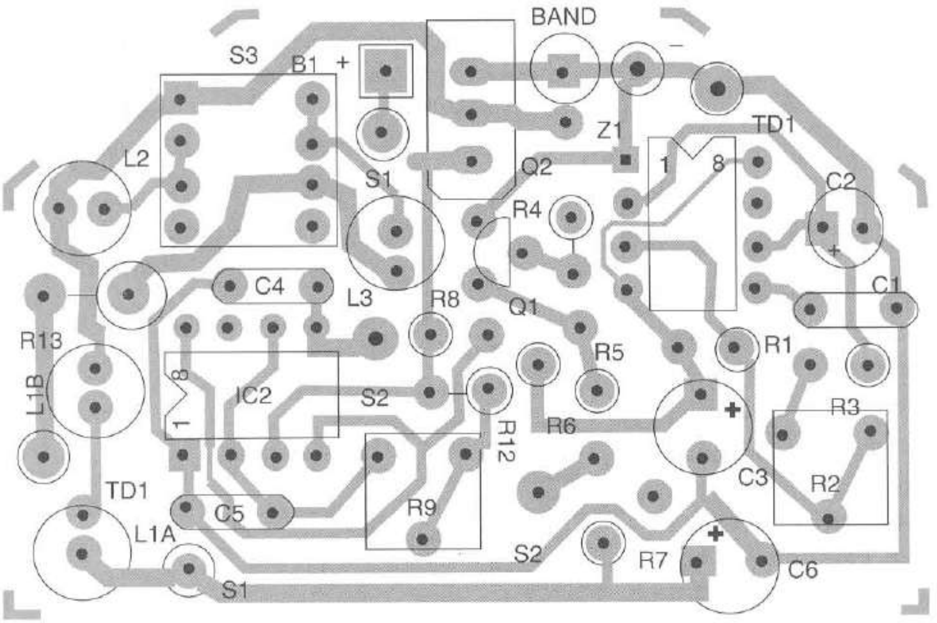
BAND
S3
B1
+
–
Z1
TD1
L2
1
8
Q2
S1
C2
R4
+
C4
C1
L3
R8
Q1
R13
R5
R1
8
L1B
IC2
S2
+
R3
1
R12
R6
TD1
C3
R2
C5
R9
L1A
+
S2
R7
C6
S1

Figure 33-3 *Printed circuit board parts and wiring*

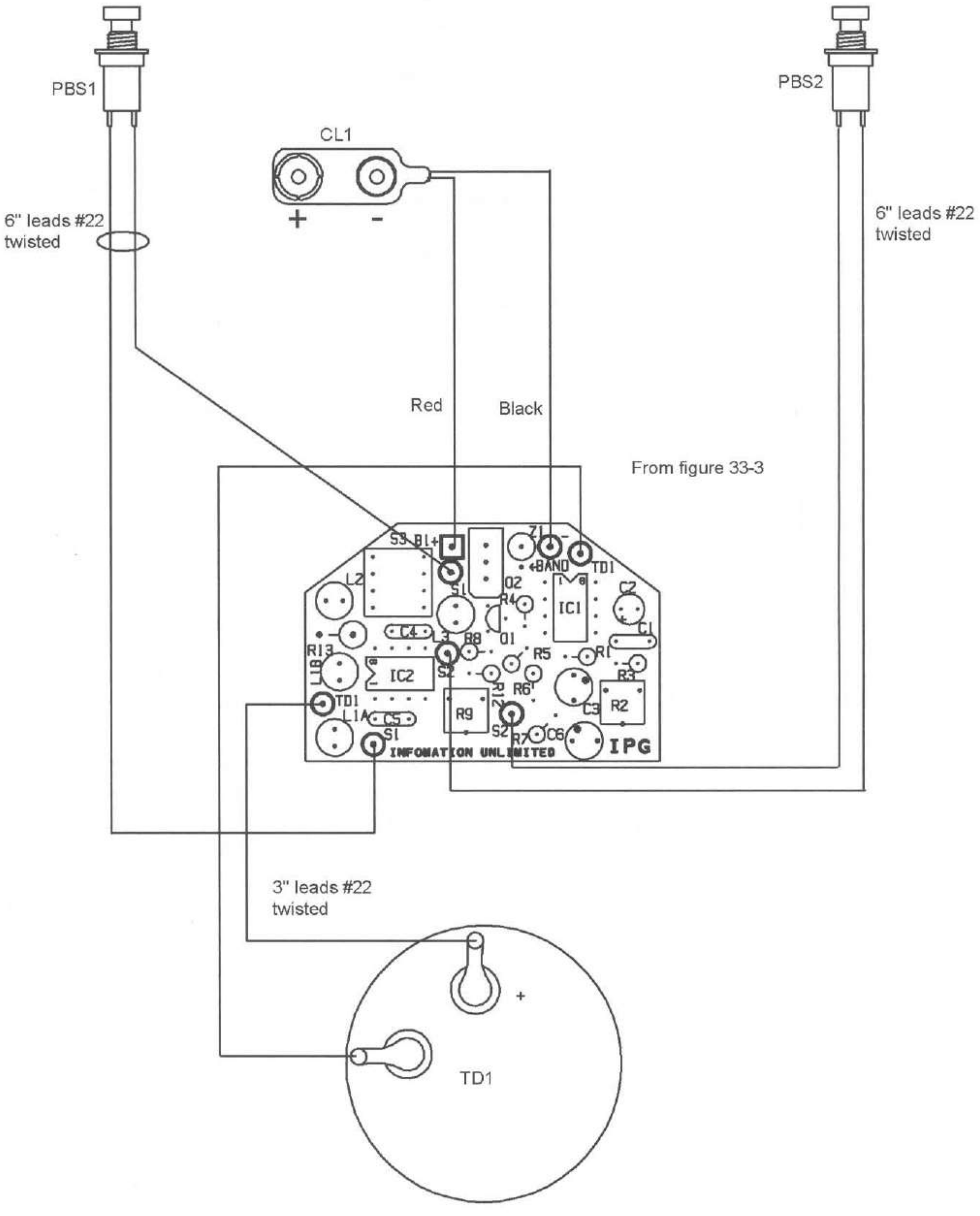
PBS1
CL1
PBS2
+
-
6" leads #22 twisted
6" leads #22 twisted
Red
Black
From figure 33-3
S3
B1+
Z1
-
BAND
TD1
L2
S1
Q2
R4
IC1
C2
C4
L3
R8
Q1
R5
R1
C1
R13
L1B
IC2
S2
R12
R6
R3
C3
R2
TD1
L1A
C5
R9
S1
S2
R7
C6
IPG
INFOMATION UNLIMITED
3" leads #22 twisted
+
TD1

Figure 33-4 *External wiring*

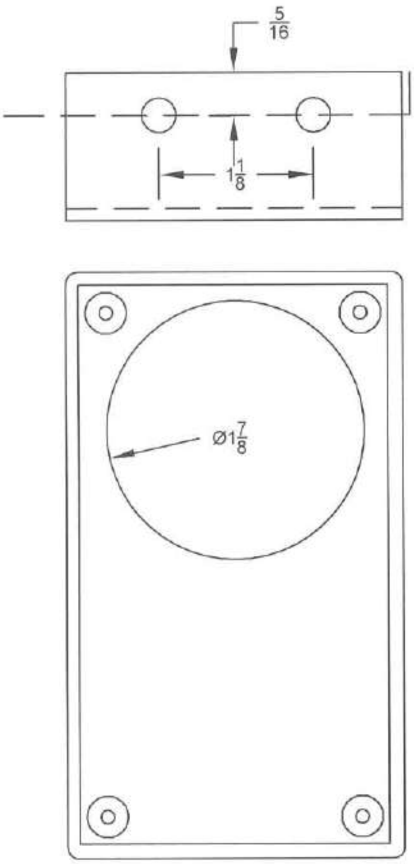

Operating instructions:

Remove rear cover and insert a 9 volt alkaline battery. Note that proper positioning is required for cover to properly fit without bowing.

Default settings of the two trimpots is midrange. To change for obtaining the maximum effect on the target you may adjust trimpots using a small plastic tuning tool or screw driver.

Simply direct unit transducer opening towards the target and push the control button. You should notice an immediate effect. The range will depend on the acoustical sensitivity of the target subject.

Use the sweep button to possible enhance the effect on the target.

This unit is designed for generating intermittent bursts of ultrasonic and high frequency acoustical energy.
Do not direct at people.
Use out side as walls and ceilings will cause the signal to loose directional characteristics and effect the user.
Use in 2 to 5 second bursts.

Figure 33-5 *Fabrication of plastic case*

of the inductances for the choice of operating frequencies that provide the maximum output.

Use a scope for these measurements. These steps are not necessary if the unit appears to function as described.

11. A sound pressure measurement of approximately 125 db was measured at 10 KHz at a distance of 16 inches on our model. Voltage across TD1 was measured at 40 volts peak.

This completes the electronic testing of the unit.

12. Fabricate the EN1 enclosure, as shown in Figure 33-5. Note $1^{7}/_{8}$-inch hole for the aperture of TDI. This hole should be made with a large punch or hole saw using a drill press. The hole may also be cut out with a sharp knife. Drill $^{1}/_{4}$-inch holes for PBSl and PBS2, as shown.
13. Cut a piece of $2^{1}/_{8} \times 2^{1}/_{8}$-inch window screen and secure it along with transducer TDI. Use RTV or suitable adhesive.
14. Make final assembly, as shown Figure 33-6, and secure assembly board with two-sided foam tape.
15. Attach the battery and position as shown. Note: The battery must be up against the enclosure with all leads properly routed or the rear cover will not fit properly.
16. See Figure 33-5 for operating instructions.

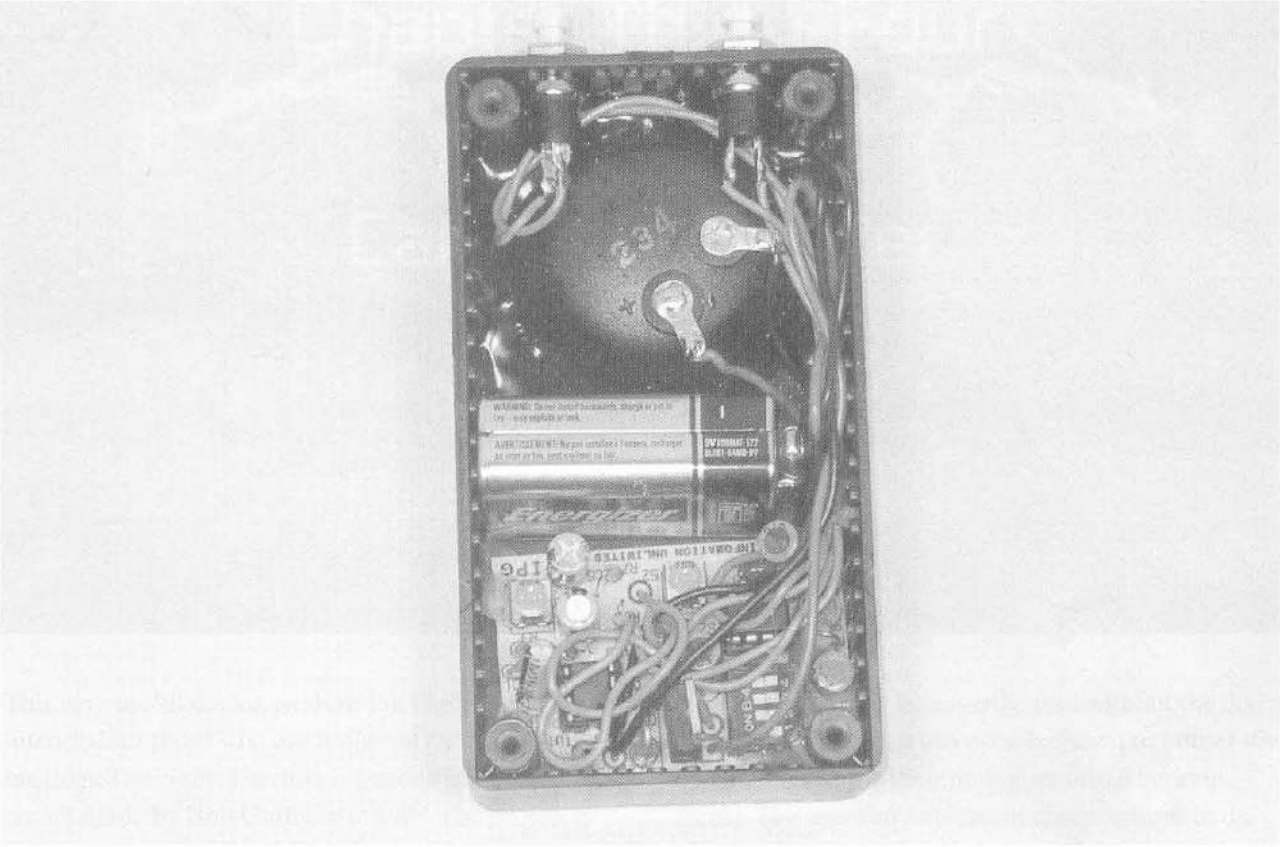

Figure 33-6 *Final assembly into EN1 enclosure*

Table 33-1 Invisible Pain-Field Generator Parts List

Ref#	*Description*	*DB Part #*
R1,6,8,12	Four 1 K, ¼-watt resistors (br-blk-red)	
R2	500 K vertical trimpot	
R3	2.2 K, ¼-watt resistor (red-red-red)	
R4,5	Two 10 K, ¼-watt resistors (br-blk-or)	
R7	10-ohm, ¼-watt (br-blk-blk)	
R9	10 K vertical trimpot	
R13	27-ohm, ½-watt resistor (red-pur-blk)	
C1,4	Two .01 mfd, 50-volt disc capacitors	
C2	10 mfd, 50-volt vertical electrolytic capacitor	
C5	.01 mfd, 50-volt plastic capacitor	
C3	100 mfd, 50-volt vertical electrolytic capacitor	
C6	1,000 mfd, 25-volt vertical electrolytic capacitor	
IC1,2	Two 555 timer DIP integrated circuits	
Q1	PN2907 PNP TO92 transistor	
SOCK1,2	Two 8-pin DIP sockets for above IC1,2 (Not shown)	
Q2	IRF540 Mosfet TO220	
L1AB,2,3	Four 1-millihenry inductors	DB# 1MH
Z1	50-volt, 1-watt Zener diode	
PBS1,2	Two pushbutton switches	
S1,2,3	3 or 4 section DIP switch	
CL1	Battery snap clip	
TD1	Special polarized transducer	DB# 1020A
PCB1	Printed circuit board	DB# PCIPG9
CASE	4 × 2 ⅜ × 1⅛-inch plastic enclosure	
SCREEN	2.5 × 2.5 inch piece of window screen	

Chapter Thirty-Four

Canine Controller

This very useful device, as shown in Figure 34-1, is intended for those who are bothered by nearby barking dogs. The control section is a modification of a circuit made by Bob Gaffigan in 1993. The project is intended to detect the dog's bark, which triggers a high-output, pain-field sonic generator, as described in Chapter 35, producing very uncomfortable sounds to the animal's sensitive hearing.

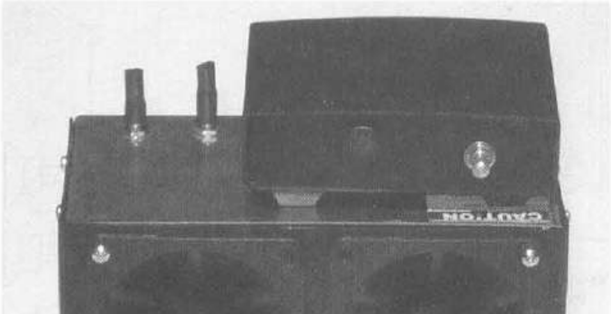

Figure 34-1 *The canine controller system with our sonic blaster*

The system may be covertly used without the dog owners knowing or it can even be tuned to bother the owners every time their mutt goes into a barking frenzy. The user can implement many options to discourage these constant barkers or give some of these inconsiderate owners a taste of their own medicine.

This is an intermediate-level project requiring basic electronic skills. Expect to spend $25 to $50. All the parts are readily available, with the specialized parts obtainable through Information Unlimited (www.amazing1.com), and they are listed in the parts list in Table 34-1 at the end of the chapter.

Device Description

A directional device picks up the dog's bark and processes it so that it turns on the sonic burst pain-field generator pointed in the offending animal's direction. The animal now experiences a very uncomfortable sound, much like a person would find the scratching of chalk on a blackboard. Eventually, the animal associates this sound with his barking, thus conditioning him to stop. The device contains adjustments that actually count the number of yelps so that it can be set not to trigger, thereby discouraging the animal from a normal barking trend that may be the warning of an intrusion or other important event.

The unit is easily mounted in any convenient location and operates from internal batteries, an external 12 volts, or 115 VAC via a wall adapter converter. The detection system contains adjustments for bark-level sensitivity, the number of barks detected before triggering, and the length of the triggered on time of the sonic burst. The pain-field generation section, as described in Chapter 35, has controls for the frequencies of the burst, sweep rate, sweep on/off, and main power control. The canine controller section is designed to easily connect to the existing jacks on the sonic generator section via a three-conductor cable. It may be mechanically attached as a single unit for a compact integrated system. The pickup microphone may even be placed at the focal point of a parabolic dish, providing very discrete selection of the target. As in any device such as this, many variables will enter into the equation that defines the overall performance. The suggested effective range will vary but can be up to 100 feet.

Special Note: In some circumstances, the owner is more to blame than the actual animal. As previously stated, the unit may be tuned to a frequency that the owner him- or herself finds very uncomfortable.

Circuit Description

Figure 34-2 shows the block diagram where the mic picks up the bark and amplifies it. A filter favors the assumed frequencies of the bark and feeds them to a bark counter. This circuit contains a selector switch to preset the necessary number of barks to trigger the bark period timer. The time of the bark period is also selectable by a switch. The bark period timer gates on the MOSFET switch transistor turns on the sonic generator, which drives the four transducers that now produce the directional, uncomfortable sound pain-field shockwaves.

Figure 34-3 shows the complete schematic of the controller section circuit, and the sonic pain-field generator section schematic is shown in Chapter 35. The bark is detected by microphone M1 and amplified by operational amplifier IC2A. The DC follower IC2D sets the midpoint bias, and resistor R2 sets the gain of this stage. The output level at this stage is controlled by potentiometer R5 and fed to amplifier/filter IC2C with a passband set by resistors R8 and R9 and capacitors C8 and C9. Schmidt discriminators

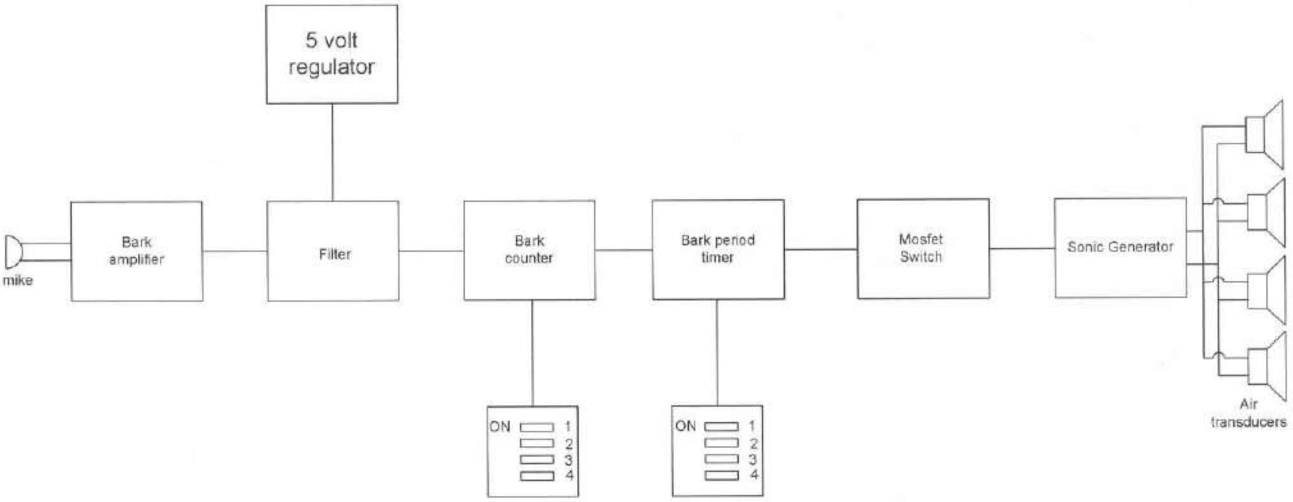

S1 determines the relative number of barks that must be detected before triggering the timer for turning on the sonic generator.
Pos1.........8
Pos2.........7
Pos3.........5
Pos4.........3

S2 determines the on time of the sonic generator once it triggers from the bark signal.
Pos1.......1sec
Pos2.......2sec
Pos3.......10sec
Pos4.......33sec
Combinations of the switch positions will be additive. Note the switch must always be on in one position.

Figure 34-2 *Circuit block diagram*

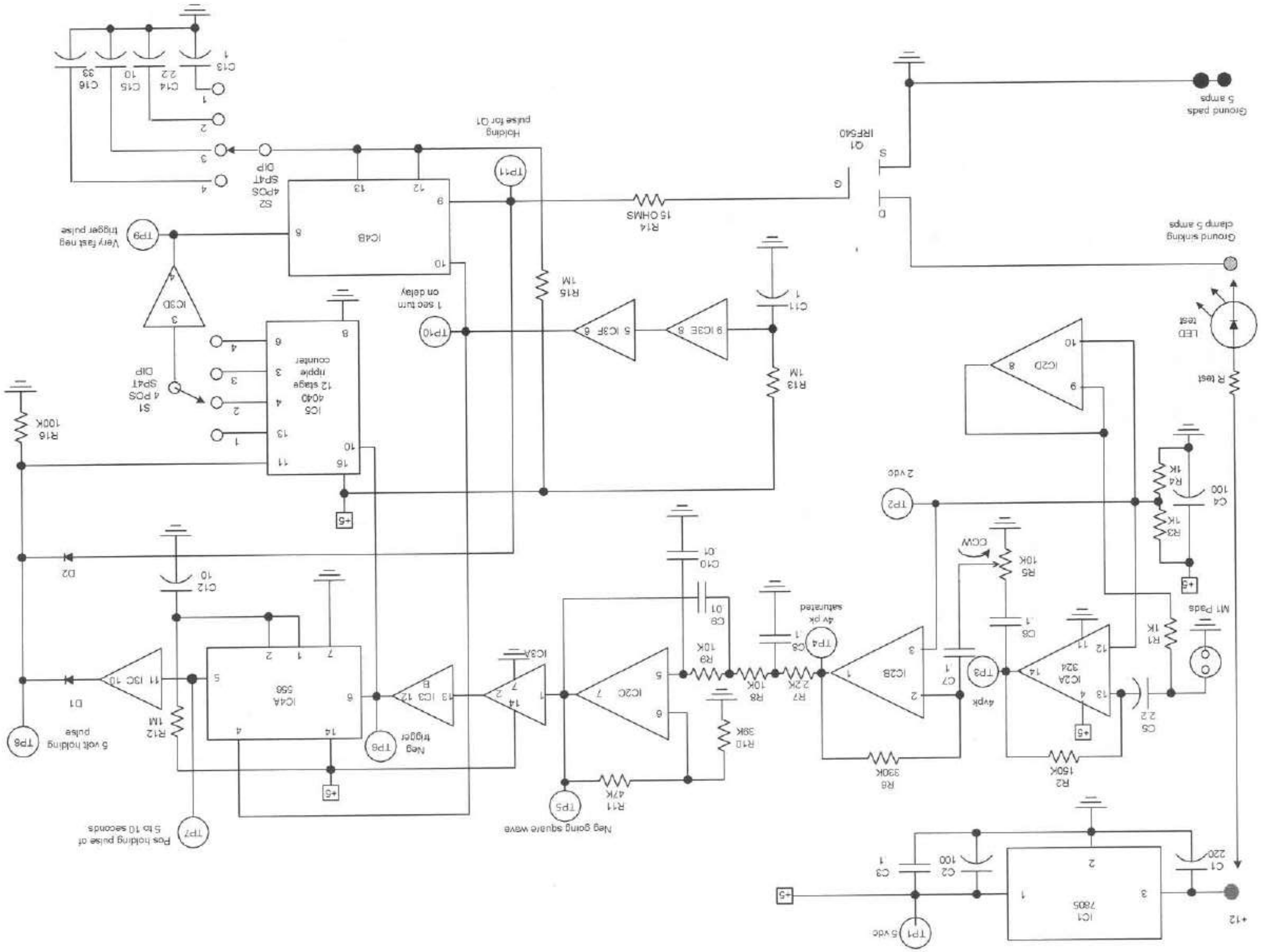
+12
C1
220
IC1
7805
3
2
1
C2
100
C3
1
TP1
5 vdc
+5
M1 Pads
C5
2.2
R1
1K
R2
150K
IC2A
324
4
13
12
11
14
TP3
4vpk
C6
.1
R5
10K
CCW
C7
.1
R6
330K
IC2B
2
3
1
TP2
2 vdc
R3
1K
R4
1K
C4
100
IC2D
9
10
8
R test
LED
test
Ground sinking
clamp 5 amps
Ground pads
5 amps
TP4
4v pk
saturated
R7
2.2K
C8
.1
R8
10K
R9
10K
C9
.01
C10
.01
R10
39K
IC2C
6
5
7
R11
47K
TP5
Neg going square wave
IC3A
14
7
2
IC3
13
12
B
Neg
trigger
TP6
IC4A
556
4
14
6
7
1
2
5
C12
10
R12
1M
TP7
Pos holding pulse of
5 to 10 seconds
11 I3C 10
D1
5 volt holding
pulse
TP8
D2
R16
100K
IC5
4040
12 stage
ripple
counter
11
13
4
3
6
16
10
8
S1
4 POS
SP4T
DIP
IC3D
TP9
Very fast neg
trigger pulse
IC4B
10
9
12
13
8
TP10
1 sec turn
on delay
TP11
Holding
pulse for Q1
R15
1M
IC3F
5
6
IC3E
9
8
R13
1M
C11
1
R14
15 OHMS
Q1
IRF540
G
D
S
S2
4POS
SP4T
DIP
C13
1
C14
2.2
C15
10
C16
33

Figure 34-3 *Circuit schematic*

IC3A and IC3B trigger the signal, provide a defined level of the correct polarity to initiate timer IC4A, and enable ripple counter IC5 to count the barks within a time period determined by IC4A. If the number of barks exceeds a limit set by bark count switch S1, timer IC4B is initiated, turning on MOSFET switch Q1 for a predetermined amount of time selected by switch S2, enabling the sonic pain-field deterrent signal to be sent toward the target and then resetting. You will note that control of the sonic pain-field generator is made by sinking the negative return lead into the ground.

Circuit Assembly

1. Assemble the circuit board as shown in Figure 34-4. Note that Figure 34-5 shows the foil traces of the printed circuit. It is not recommended that you assemble this circuit on a perforated vector board unless you are an advanced assembler.

   Note that if you are building from a perforated or vector circuit board, it is suggested that you use the indicated traces for the wire runs and insert components starting at the lower left-hand corner. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors. It is a good idea to use sockets for all the integrated circuits.

   Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #26 bus wire. The heavy foil runs should use the thicker #24 bus wire, as these are the high-current discharge paths.

2. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire

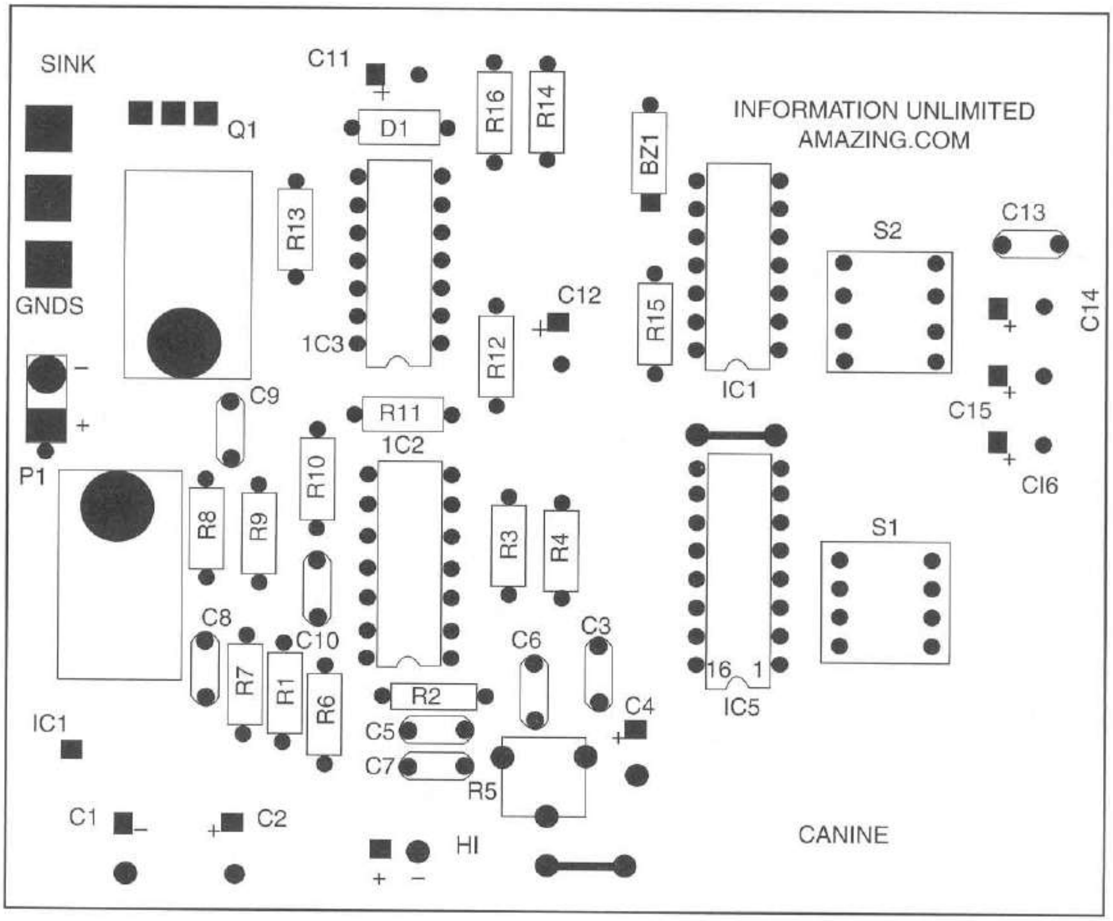
SINK
Q1
C11
D1
R16
R14
BZ1
INFORMATION UNLIMITED
AMAZING.COM
R13
C13
S2
GNDS
C12
R15
C14
1C3
R12
IC1
P1
C9
R11
C15
1C2
R10
CI6
R8
R9
R3
R4
S1
C8
C10
C6
C3
R7
R1
R6
16
1
R2
C4
IC5
IC1
C5
C7
R5
C1
C2
HI
CANINE
+
-

Figure 34-4 *Assembly of the circuit board*

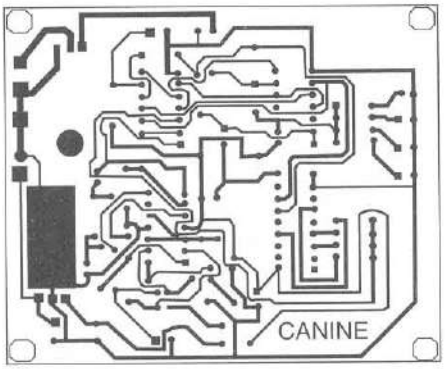

Figure 34-5 *Assembly circuit board traces*

bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.

3. Connect the external components as shown in Figure 34-6. Use shielded microphone cable if you are not installing it on the actual board.
4. Fabricate the chassis as shown in Figure 34-7 from a piece of .035-inch 5052 bendable aluminum or plastic. If you use metal, you will need a piece of plastic material under the assembly board to prevent shorting the foil traces.
5. Create a mating cover and finally assemble everything as shown in Figure 34-8. Note the hole for the microphone with a screen. Wire the P1 plug and cable for interconnecting the controller and the sonic blaster using desired lengths per your requirements. Leave the other end with leads that will eventually connect to the sonic blaster.

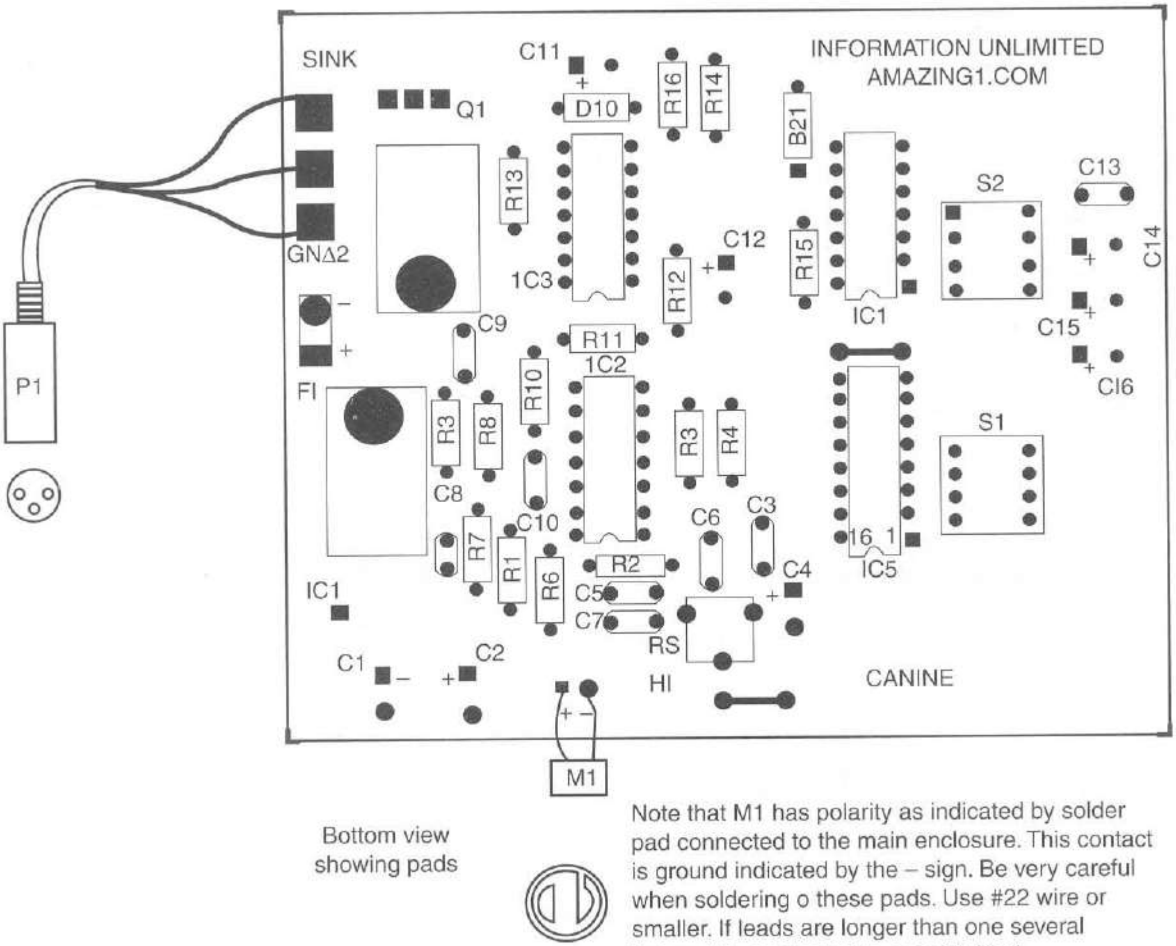
SINK
Q1
C11
D10
R16
R14
INFORMATION UNLIMITED
AMAZING1.COM
B21
R13
1C3
C12
R12
R15
IC1
S2
C13
C14
C15
CI6
GNΔ2
C9
R11
1C2
FI
P1
R10
R3
R8
R3
R4
S1
C8
C6
C3
R7
R1
C10
R6
R2
C4
16
1
IC5
IC1
C5
C7
RS
C1
C2
HI
CANINE
M1
Bottom view
showing pads

Note that M1 has polarity as indicated by solder pad connected to the main enclosure. This contact is ground indicated by the – sign. Be very careful when soldering o these pads. Use #22 wire or smaller. If leads are longer than one several inches it is suggested to use shielded microphone cable.

Figure 34-6 *External wiring*

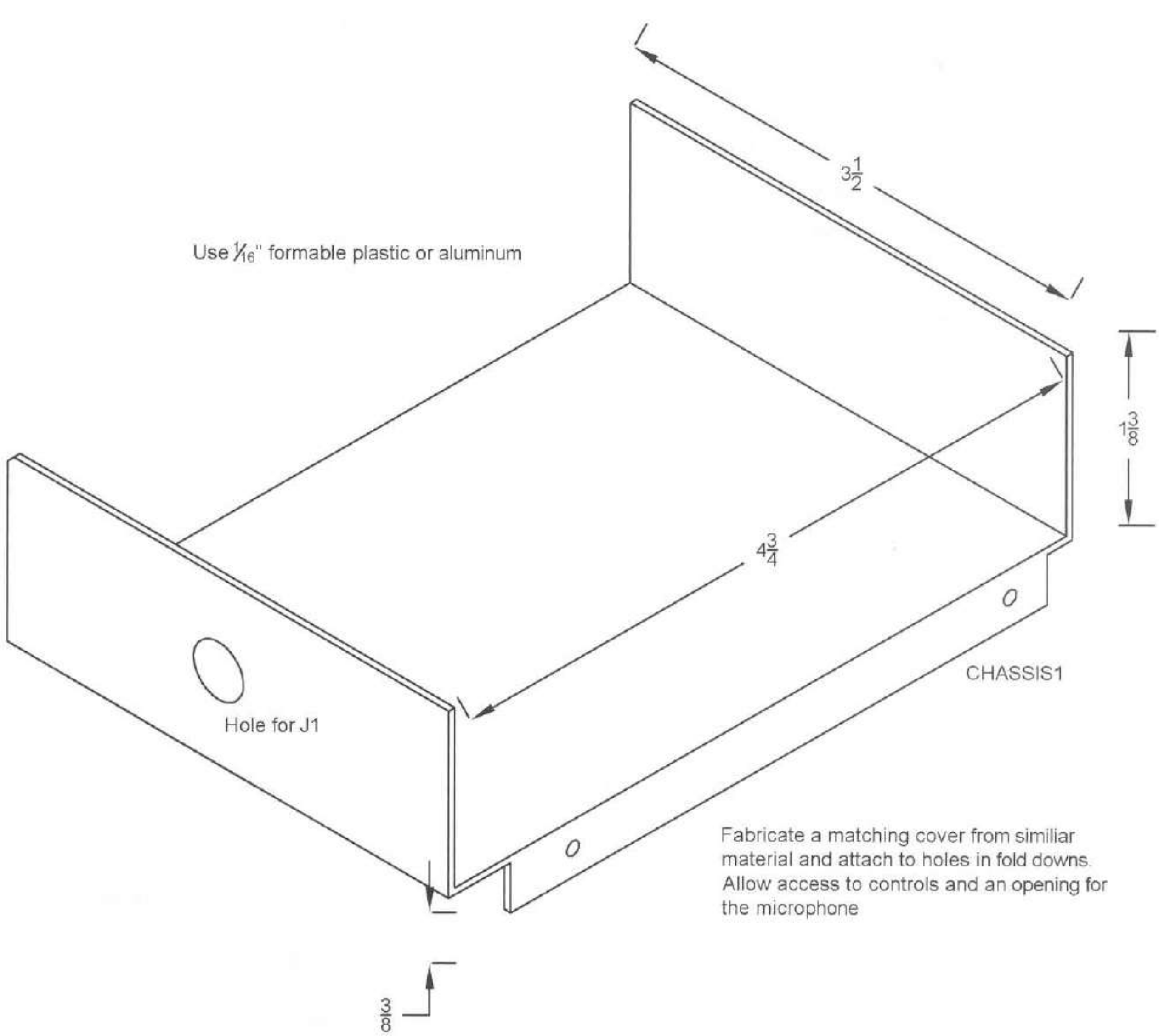
3½
Use 1/16" formable plastic or aluminum
1⅜
4¾
CHASSIS1
Hole for J1
Fabricate a matching cover from similiar material and attach to holes in fold downs. Allow access to controls and an opening for the microphone
⅜

Figure 34-7 *Chassis fabrication*

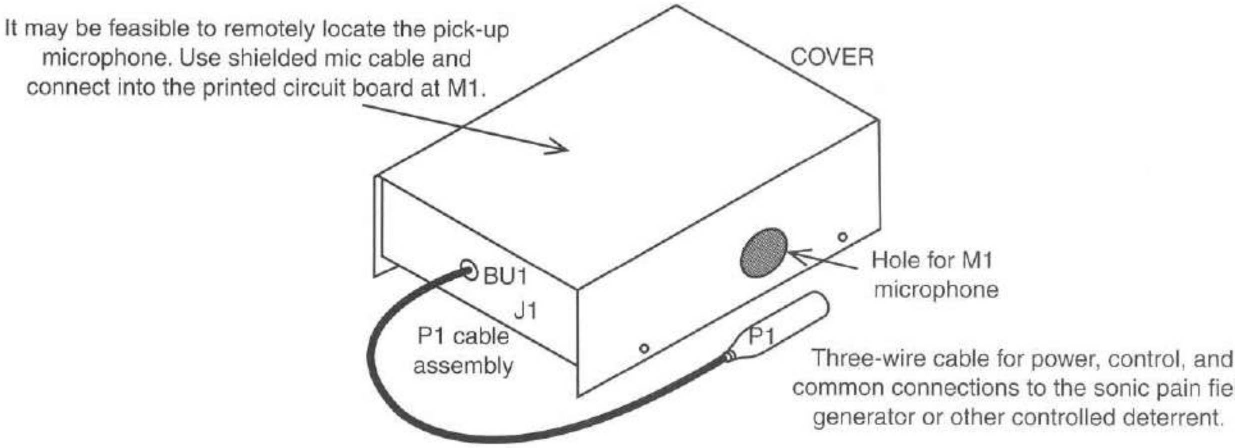
It may be feasible to remotely locate the pick-up microphone. Use shielded mic cable and connect into the printed circuit board at M1.
COVER
BU1
J1
P1 cable assembly
Hole for M1 microphone
P1
Three-wire cable for power, control, and common connections to the sonic pain field generator or other controlled deterrent.

Figure 34-8 *Isometric of final assembly*

Testing the Circuit

1. Assemble the test jig, referring back to Figure 34-3. If the circuit is correct, you may go right to the connection of the test jig as shown and perform the following steps:
   a. Preset S1 to position 4 (the lowest number of barks for activation).
   b. Preset S2 to position 1 (the shortest time for the sonic generator).
   c. Preset R5 to midrange.
2. Apply a 12 VDC input and note a low-current draw of less than 10 milliamps, which is the same value as the quiescent operating current. C cell batteries will work approximately 30 days before requiring replacement.
3. Simulate a bark and note the test LED momentarily coming on. If this occurs, you may want to test the various positions of S1 and S2. Also verify the sensitivity control, R5.
4. If the circuit does not work, you may need a scope to test the various test points as shown in Figure 34-3. They should have the following values:

   TP1: +5 volts DC

   TP2: 2 to 3 volts DC

   TP3: 4-peak-volt audio level

   TP4: Saturated audio signal

   TP5: Negative-going pulse

   TP6: Negative-going square wave

   TP7: Positive 10-second holding pulse

   TP8: Positive holding pulse

   TP9: Very fast negative trigger pulse

   TP10: 1-second turn-on delay

   TP11: Holding pulse for Q1

10. Once the controller operation is verified, you may connect the system as shown in Figure 34-9. You will need a properly working sonic pain-field generator, as described in Chapter 35.

It is suggested that you experiment with the system before actually installing it. Once familiar with it, you may position the system for the best effect. You also have the option of adjusting the sonic generator when the dog starts barking so that it can also be very annoying to the animal owners when their mutt goes into an uncontrolled barking rage. Use your own judgment as you can get tagged for harassment!

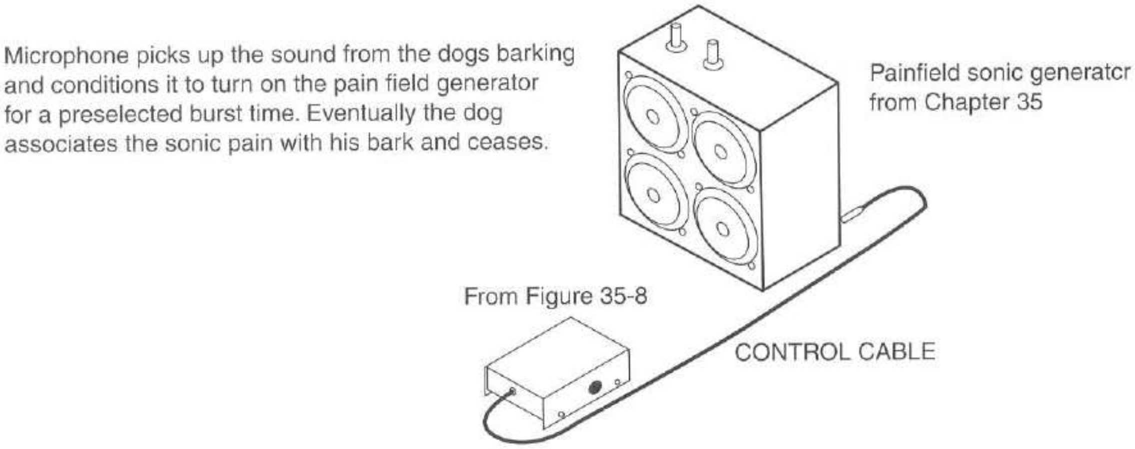
Microphone picks up the sound from the dogs barking and conditions it to turn on the pain field generator for a preselected burst time. Eventually the dog associates the sonic pain with his bark and ceases.
Painfield sonic generatcr from Chapter 35
From Figure 35-8
CONTROL CABLE

Figure 34-9 *Systems view*

Table 34-1 Canine Controller Parts List

Ref. #	*Description*	*DB Part #*
R1, 3, 4	Three 1K, ¼-watt resistors (br-blk-red)	
R2	150K, ¼-watt resistor (br-gr-yel)	
R5	10K pot and 12-volt switch, 17 mm	
R6	330K, ¼-watt resistor (or-or-yel)	
R7	2.2K, ¼-watt resistor (red-red-red)	
R8, 9, 10	Three 10 kilo-ohm, ¼-watt resistors (br-blk-or)	
R11	47 kilo-ohm, ¼-watt resistor (yel-pur-or)	
R12, 13, 15	Three 1M, ¼-watt resistors (br-blk-gr)	
R14	15-ohm, ¼-watt resistor (br-gr-blk)	
R16	100K, ¼-watt resistor (br-blk-yel)	
C1	220-microfarad, 25-volt vertical electrolytic capacitor	
C2, 4	Two 100-microfarad, 25-volt vertical electrolytic capacitors	
C3, 6, 7, 8	Four .1-microfarad, 50-volt plastic capacitors	
C5, 14	Two 2.2-microfarad, 50-volt vertical electrolytic capacitors	
C9, 10	Two .01-microfarad, 50-volt plastic capacitors	
C11, 13	Two 1-microfarad, 50-volt vertical electrolytic capacitors	
C12, 15	Two 10-microfarad, 50-volt vertical electrolytic capacitors	
C16	33-microfarad, 50-volt vertical electrolytic capacitor	
D1, 2	Two IN914 silicon diodes	
IC1	7805 5-volt regulator TO220	
IC2A, B, C, D	LM324 quad amp in dual inline package (DIP)	
IC3A, B, C, D, E, F	40106 hex Schmidt in DIP	
IC4	556 dual timer in DIP	
IC5	4040 complementary metal oxide semiconductor (CMOS) PLL phase lock loop in DIP	
Q1	100-volt IRF540 metal-oxide-semiconductor field effect transistor	MOSFET
M1	FET mic element	DB# FETMIK
S1, 2	Two four-position single-throw DIP switches	
PCCANINE	Printed circuit board (PCB)	DB# PCCANINE
CHASSIS1	Metal or plastic chassis as shown in Figure 34-7	
COV1	Cover fabricated as shown in Figure 34-8	
J1/P1	Three-pin chassis mount jack and mating plug	
WR3C	3-inch three-conductor cable for interconnecting	
PPF4K	Pain-field generator kit as described in Chapter 35	DB# PPF4K
PPF40	Pain-field generator assembled as described in Chapter 35	DB# PPF40

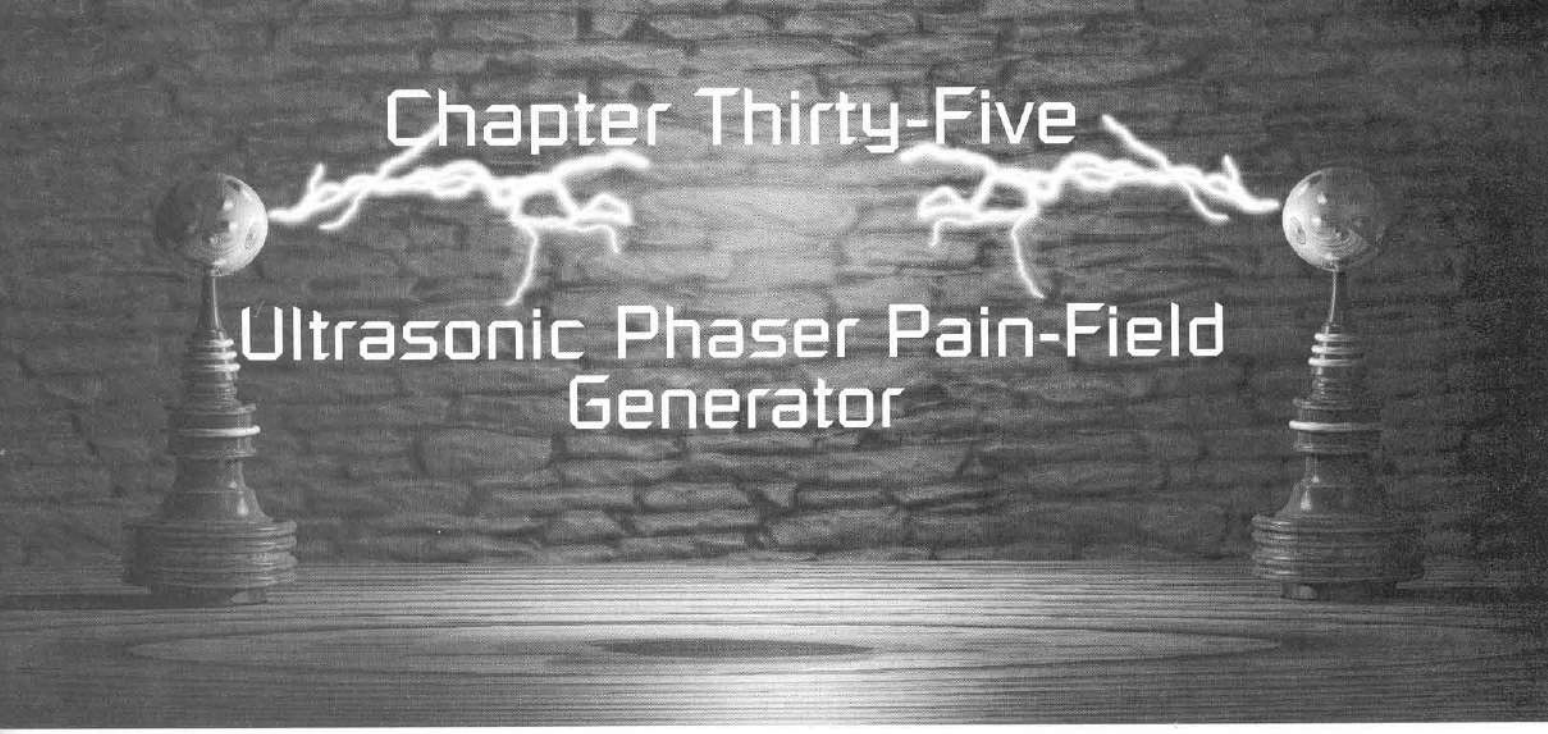

Chapter Thirty-Five

Ultrasonic Phaser Pain-Field Generator

This project, as shown in Figure 35-1, shows how to construct a moderately high powered sonic generator that can be used for tasks that range from animal control to discouraging personal encounters. It can be used as part of our laser property protection guard described in Chapter 5 or as the deterrent for use with the canine controller, as described in Chapter 34. The unit can generate a variable rate of complex waves from 5 to 25 kHz well into the ultrasonic range. These waves can be very painful or disorientating to animals and people, depending on where the controls of the unit are set.

This is an excellent device for use in animal control as well as a low-liability deterrent in anti-intrusion alarms and detection systems. This is an intermediate-level project requiring basic electronic skills. Expect to spend $50 to $75. All parts are readily available, with specialized parts obtainable through Information Unlimited at www.amazing1.com, and they are listed in the parts list at the end of the chapter.

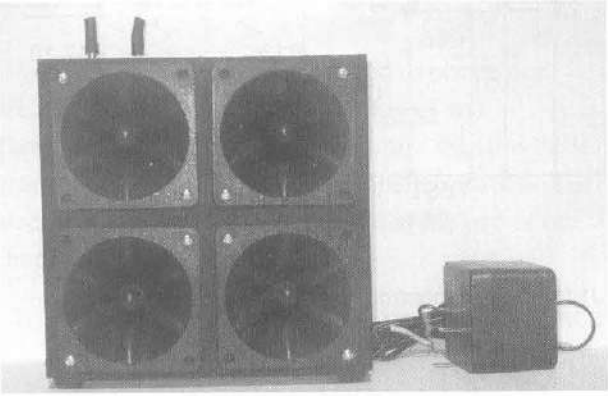

Figure 35-1 *The phaser pain-field generator*

Basic Device Description

This chapter demonstrates how to create a variable-sweep-frequency, ultrasonic, pain-field generator capable of producing the equivalent of 400 watts of resultant power such as that obtainable from a conventional dynamic transducer system. This is possible due to recently developed piezoelectric ceramic devices. The high efficiency requires very little driving power, consequently resulting in a lightweight, portable, handheld device that is battery driven and capable of producing high sound pressure. It is a directional device and can be set up in a target area such as a field or garden to discourage any sound-sensitive animals. It is also capable of being powered by a car's 12-volt system and can be mounted on the hood or roof of the vehicle. Transducers are mounted in an array for a concentration of energy in one direction.

It should be understood that certain people subjected to different degrees of exposure are affected more than others, some to a point where they may vomit or experience severe headaches and cranial pains. Some people will experience severe pain in the ear, teeth, or lower head. Statistically, women and younger children are many times more sensitive to this device than the average male adult. With this in mind, the user must exercise consideration when testing and using the device for animal control, as many people will not be aware of the source of this pain and attribute it to a headache or other physical ailment. Also, certain people are affected mentally to a point of actually losing their tempers completely or becoming extremely quick-tempered. Some will experience a state of extreme anxiety when overly exposed. Therefore, consideration must be used at all times when testing or using this or similar devices.

It should also be noted that using the transducers in an array configuration may be damaging to hearing at close range. The array approach produces high sound-pressure density occurring on or near the output axis.

Circuit Theory

Figure 35-2 shows a timer, IC2, connected as an astable free-running multivibrator whose frequency is externally controlled by pot R9. Resistor R10 selects the range limit of R9, and capacitor C5 determines the frequency range of the device along with the previous resistors.

The square wave output of IC2 is via pin 3 and is directly coupled to power amplifier Q2. The drain of Q2 is DC biased through choke L1.

The square wave output signal is then fed into transducer TD1 in series with parallel combination of resonating coils L2 and L3. The resonant action between the inherent capacity of TD1, the C7 tuning capacitor, and the inductance now produces a sinusoidal-shaped wave peaking around 25 kHz or the

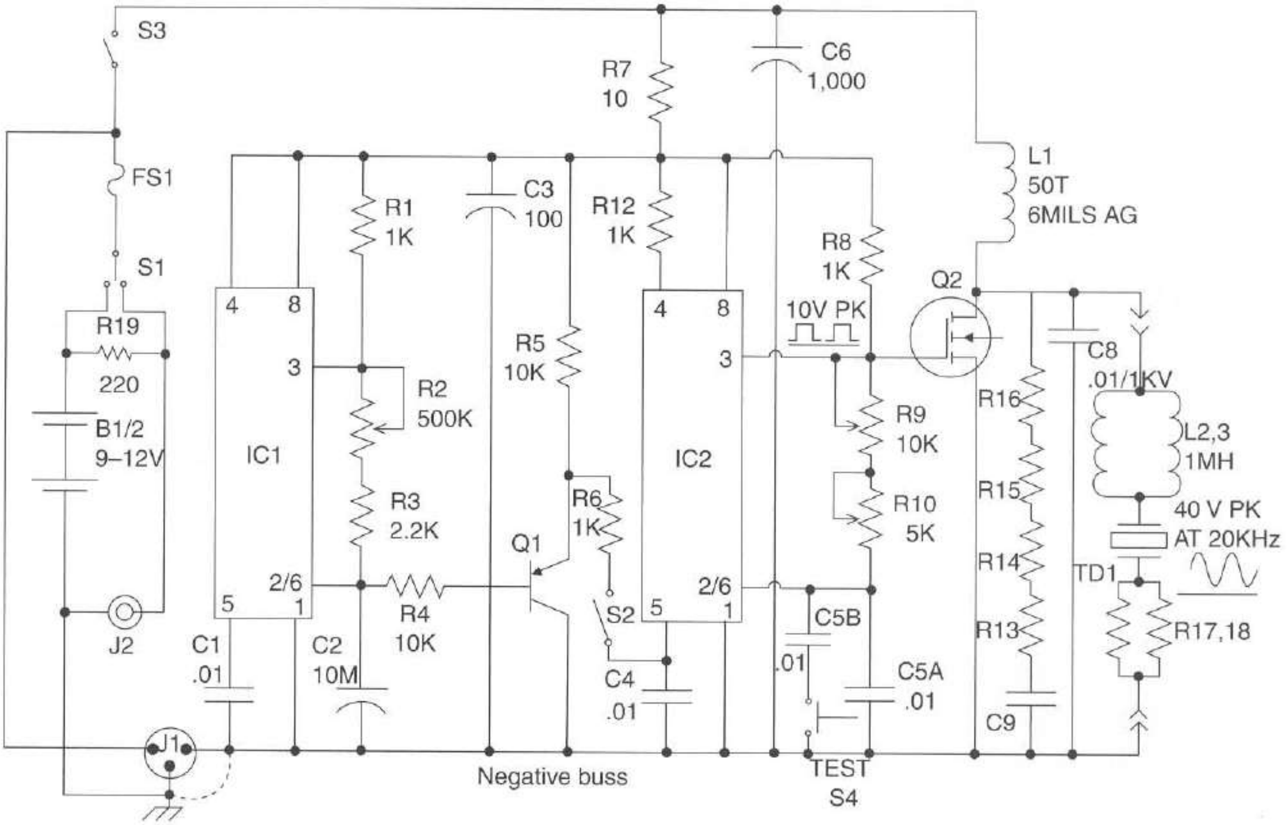
S3
FS1
S1
R19
220
B1/2
9–12V
J2
J1
4
8
3
IC1
2/6
5
1
C1
.01
C2
10M
R1
1K
R2
500K
R3
2.2K
R4
10K
C3
100
R5
10K
Q1
R6
1K
S2
R7
10
R12
1K
C4
.01
C6
1,000
IC2
R8
1K
10V PK
R9
10K
R10
5K
C5B
.01
C5A
.01
TEST
S4
Negative buss
Q2
L1
50T
6MILS AG
C8
.01/1KV
L2,3
1MH
40 V PK
AT 20KHz
TD1
R16
R15
R14
R13
C9
R17,18

Dashed line indicates bypass connection when not using J1 for external control. You may also short these pins using a mating plug with pins connected together.

Note four transducers and associated networks.

Figure 35-2 *Schematic of the phaser pain-field generator*

upper limit of the tuning range. This signal waveform now has a peak-to-peak voltage several times that of the original square wave. Transducer TD1 now can take advantage of these peak voltages to produce the high sound-pressure levels necessary without exceeding the high *root means square* (rms) ratings of an equivalent voltage-level square wave.

Timer IC1 is similarly connected as an astable, running multivibrator and is used to produce the sweeping voltage necessary for modulating the frequency of IC2. This sweep repetition rate controlled by pot R2 and resistor R3 limits the range of this repetition time. Resistor R1 selects the duty cycle of the pulse, while capacitor C2 sets the sweep time range. The output for IC1 is via pins 6 and 2 where the signal ramp function voltage is resistively coupled to inverter transistor Q1 via resistor R4. The output of Q1 is fed to pin 5 of IC2 and provides the output modulation voltage necessary to vary the frequency as required. Note that the modulation signal is easily disabled via R2/S2.

Capacitor C6 guarantees an AC return path for the output signal. Power to driver circuits IC1 and IC2 is through a decoupling network consisting of resistor R7 and capacitor C3.

Power to the system is via internal battery B1 or an external 12 VDC such as from a vehicle. Three-way switch S1 selects between these two power sources or shuts the unit off in the center position. Resistor R19 provides a trickle charge to the internal battery when using an external source. Jack J1 allows a connection that is similar to our canine controller, turning on the unit when it detects a barking dog.

Construction Steps

If you are a beginner it is suggested to obtain our *GCAT1 General Construction Practices and Techniques*. This informative literature explains basic practices that are necessary in proper construction of electromechanical kits and is listed at the end of the chapter.

1. Lay out and identify all the parts and pieces, and check them with the parts list. Note that certain parts may sometimes vary in value. This is acceptable as all components are 10 to 20 percent tolerant unless otherwise noted.
2. Fabricate the heatsink bracket (HS1) as shown in Figure 35-3 from a .75 × 2 × .065-inch aluminum piece. Bend it 90 degrees at its midsection and drill a hole for the SW1/NU1 screw and nut. Attach it to Q2 as shown.
3. Assemble the L1 choke coil as shown in Figure 35-4 by wrapping 50 turns of #24 magnet wire on the nylon bobbin as evenly as possible. Leave 2 inches of leads for a connection to the circuitry. Assemble the E core as shown and shim each side with pieces of yellow cardboard strips of 3 mils each for a total of 6 mils. If you have an LCR bridge, measure 1 millihenry.
4. Assemble the PCB as shown in Figure 35-5. Note the two wire jumps and component polarity. Wire in inductor L1 and secure it to the board with *room temperature vulcanizing* (RTV) silicon rubber or another suitable adhesive.

   If you are building from a perforated or vector circuit board, it is suggested that you use the indicated traces for the wire runs and insert the components starting in the lower left-hand corner. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors.

   Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire

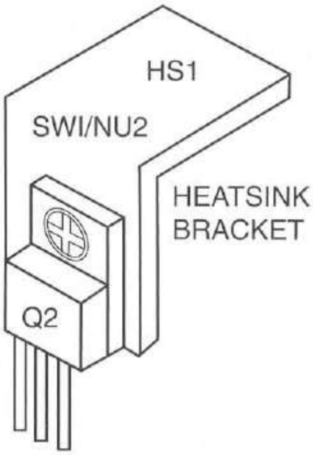

Figure 35-3 *Fabrication of heatsink*

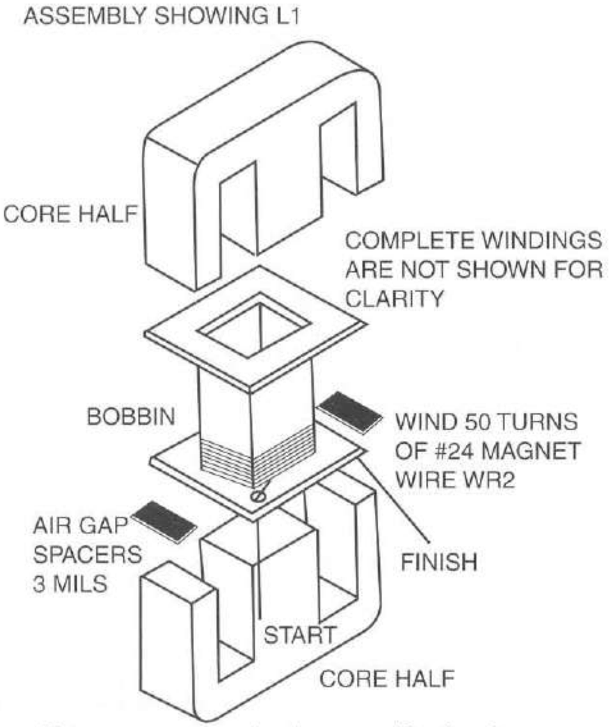
ASSEMBLY SHOWING L1
CORE HALF
COMPLETE WINDINGS ARE NOT SHOWN FOR CLARITY
BOBBIN
WIND 50 TURNS OF #24 MAGNET WIRE WR2
AIR GAP SPACERS 3 MILS
FINISH
START
CORE HALF

Note you can use a business card for the air gap shims as thickness will be close enough.

Use tape to secure core halves tightly together.

Figure 35-4 *Assembly of L1 inductor*

runs or use pieces of the #24 bus wire. The heavy foil runs should use the thicker #20 bus wire as these are the high-current discharge paths.

It is suggested that you use an 8-pin integrated circuit socket (SO8) for IC1 and IC2.

5. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.

6. Fabricate the chassis and cover, as shown in Figure 35-6 from 1/16-inch Lexan plastic or aluminum. Use a large punch or circle saw to cut out the four holes for the transducers. Drill the remaining holes as you assemble it, verifying the proper clearances and sizes with components.

7. Complete the final assembly shown in Figure 35-7 and mount the assembly board, battery holders, switch S1, jacks J1 and J2, and fuse holder FS1. It is a good idea to trial-fit all the parts in this section before actually drilling or making any mounting holes. Note a piece of plastic (SPACER) that insulates the bottom solder connections of the PCB from shorting out to the metal chassis base. Also note resistor R19 across S1. Secure the battery holders using pieces of two-sided foam sticky tape (TAPE). Wire using #20 vinyl wire as shown.

8. Mount the transducers as shown in Figure 35-8 using SW2/NU2 6-32 × 1/2-inch screws and nuts.

   Twist the leads of the two inductors L2 and L3 together, effectively paralleling these components. Repeat this for resistors R17 and R18. You should end up with four sets of these parallel components. These components should be self-supporting by being connected to stiff pieces of bus wire. Take caution to observe any potential shorts. Solder the two 18-inch connecting leads and connect them to point P6 on the assembly board.

9. The final assembly is shown in Figure 35-9. Do not attach the cover until you have tested the unit.

Testing the Assembly

10. Turn the pots counterclockwise, set the switch to mid position (off), and install eight fresh batteries into the battery holders. If you have access to a bench supply of 12 VDC at 2 amps, it may be connected to J2 via a mating plug, eliminating the need for the batteries in the preliminary test. Short out the chassis ground and negative bus pins of J1. This jack is for connection to the canine controller module or other remote control devices.

Note: A variable bench supply capable of 12 VDC at 2 amps with a voltmeter and current meter can be a great convenience for the remaining steps and the testing of other similar circuits.

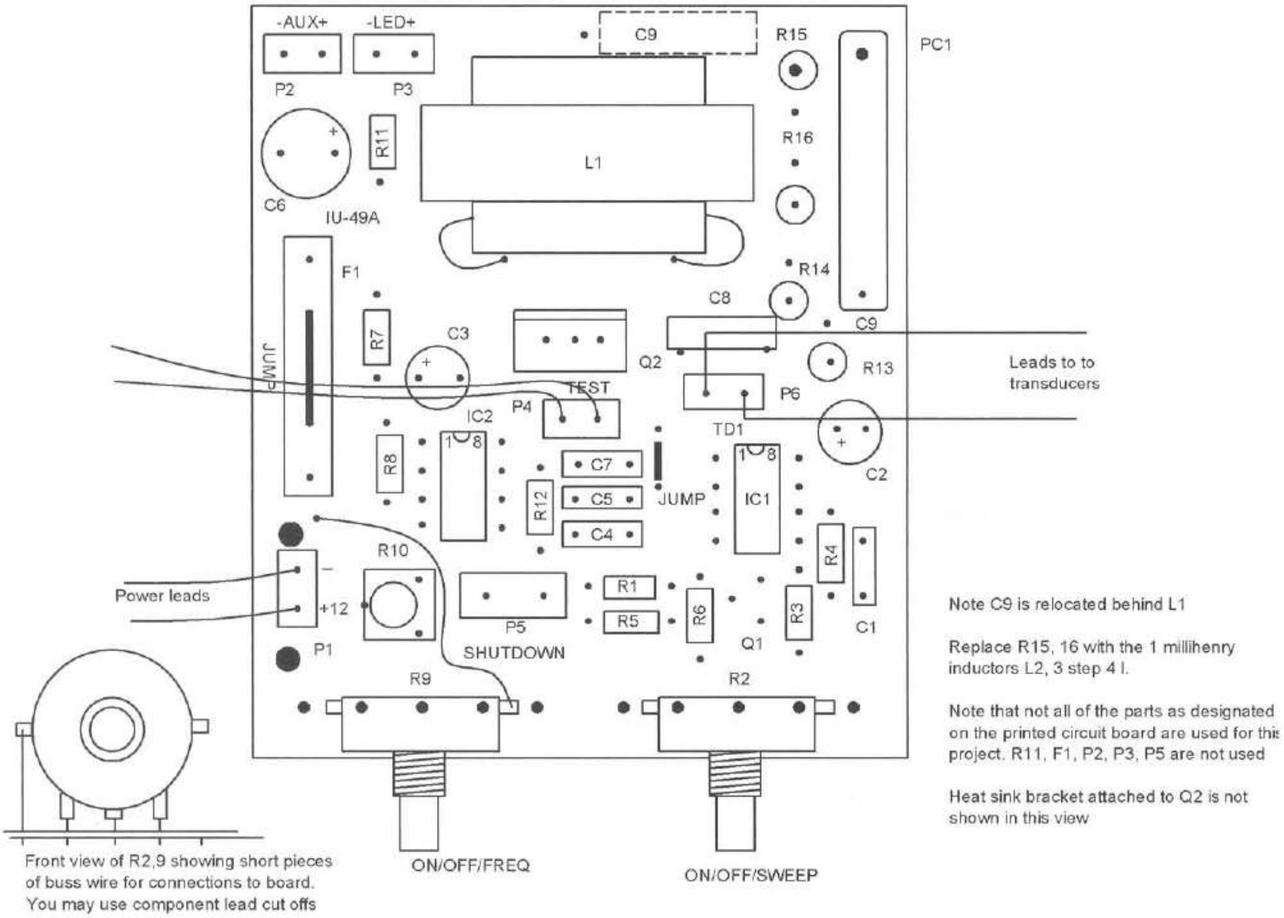

Note C9 is relocated behind L1

Replace R15, 16 with the 1 millihenry inductors L2, 3 step 4 I.

Note that not all of the parts as designated on the printed circuit board are used for this project. R11, F1, P2, P3, P5 are not used

Heat sink bracket attached to Q2 is not shown in this view

Front view of R2,9 showing short pieces of buss wire for connections to board. You may use component lead cut offs

Figure 35-5 *Assembly of PCB*

11. Press push button S1 and note a loud, piercing wave coming from TD1. Measure a current of 500 to 600 milliamps in series with a battery or bench supply. Note that the switch position when up should be power supplied by internal batteries, and the down position being supplied by the external 12 VDC supply. This function may be reversed but should be noted, as the external position will automatically trickle charge the internal batteries.

12. Rotate R9 and note the frequency increasing to above the audible range; measure a current of 400 to 800 milliamps. You may preset the limit of the high range by the R9 setting. The normal factory setting is 25 kHz maximum with R10 full clockwise. Note the wave shapes shown in Figure 35-2 for those who have a scope.

13. Turn on sweep control R2/S2 and note the frequency being modulated by a changing rate as this control is adjusted. Use caution as certain sweep rates may cause epileptic seizures and other undesirable effects. Sweep rates between 7 to 20 per sec should be used with caution.

Basic Operating Instructions

14. Connect a 12-volt source capable of supplying 2 amps of current to designated leads or use internal *nickel cadmium* (NiCad) batteries that are type C cells. These will provide approximately 3½ hours continuous operation at slightly reduced output.

15. The battery-charging function occurs when the system is connected to an external power source and will charge the NiCad batteries at a 100-milliamp rate, taking about 14 hours to fully charge.

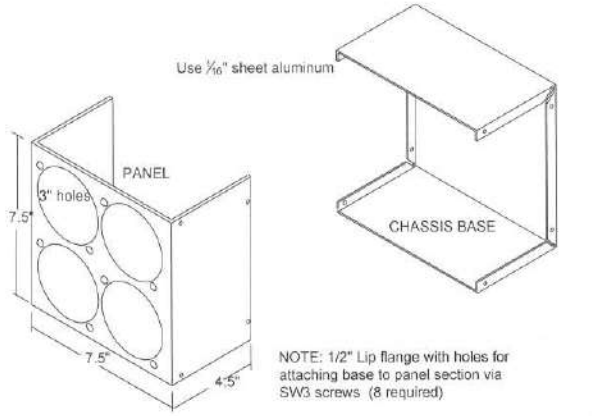
Use 1/16" sheet aluminum
PANEL
3" holes
7.5"
7.5"
4.5"
CHASSIS BASE
NOTE: 1/2" Lip flange with holes for attaching base to panel section via SW3 screws (8 required)

Figure 35-6 *Chassis fabrication*

16. If continued use is anticipated via the external 12 volts, it is a good idea to remove the internal batteries' connections as overcharge may occur that will ruin the batteries or actually cause them to explode. The recommended method of setting the controls is the following:
    a. Determine the frequency limits per the application. These types of devices have two basic applications. When used as an anti-intrusion device to discourage unauthorized entrance, adjustments are made for maximum human annoyance, usually with frequencies from 10 to 15 kHz.

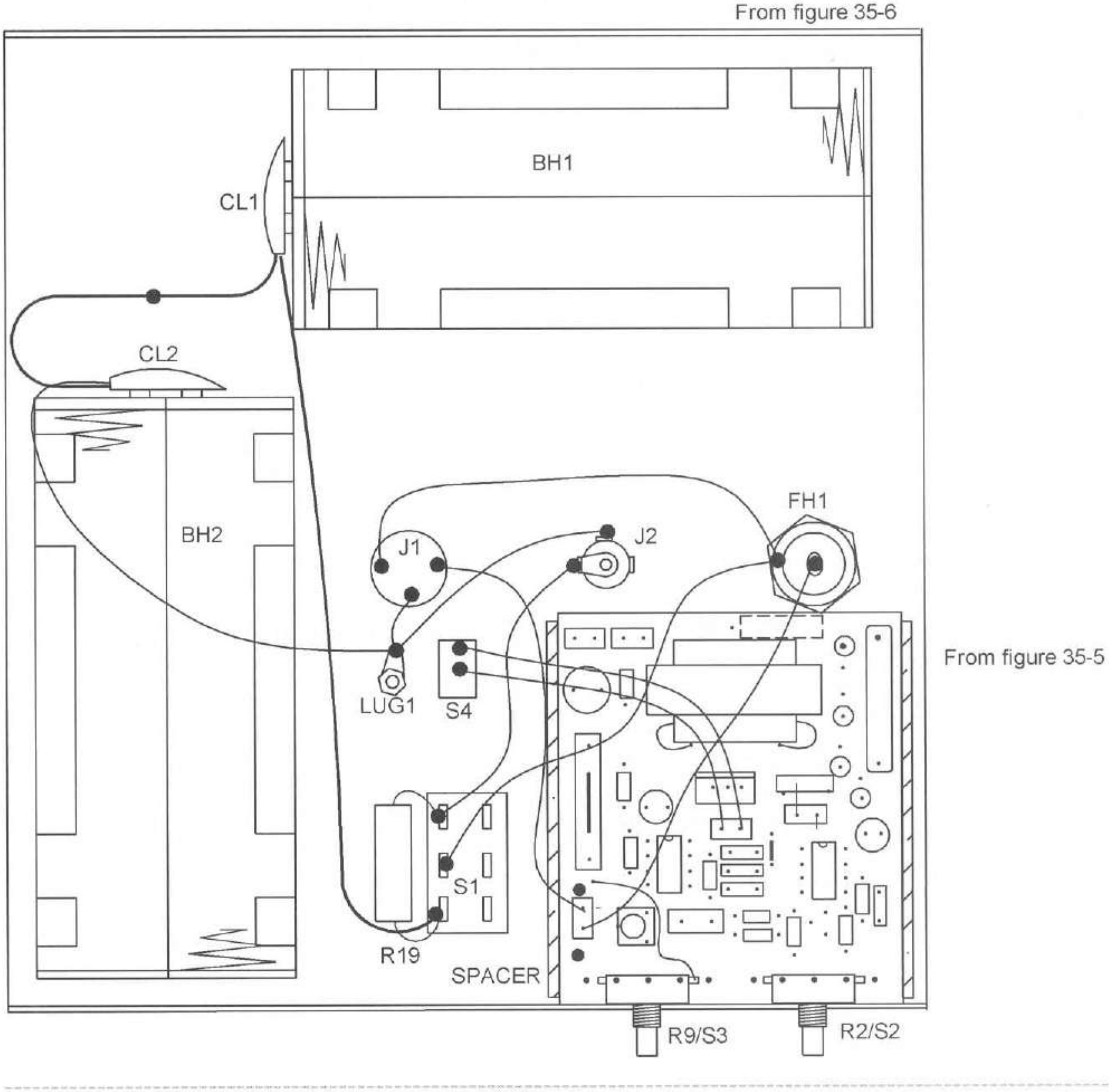
From figure 35-6
BH1
CL1
CL2
BH2
FH1
J1
J2
LUG1
S4
From figure 35-5
S1
R19
SPACER
R9/S3
R2/S2

Figure 35-7 *Chassis wiring*

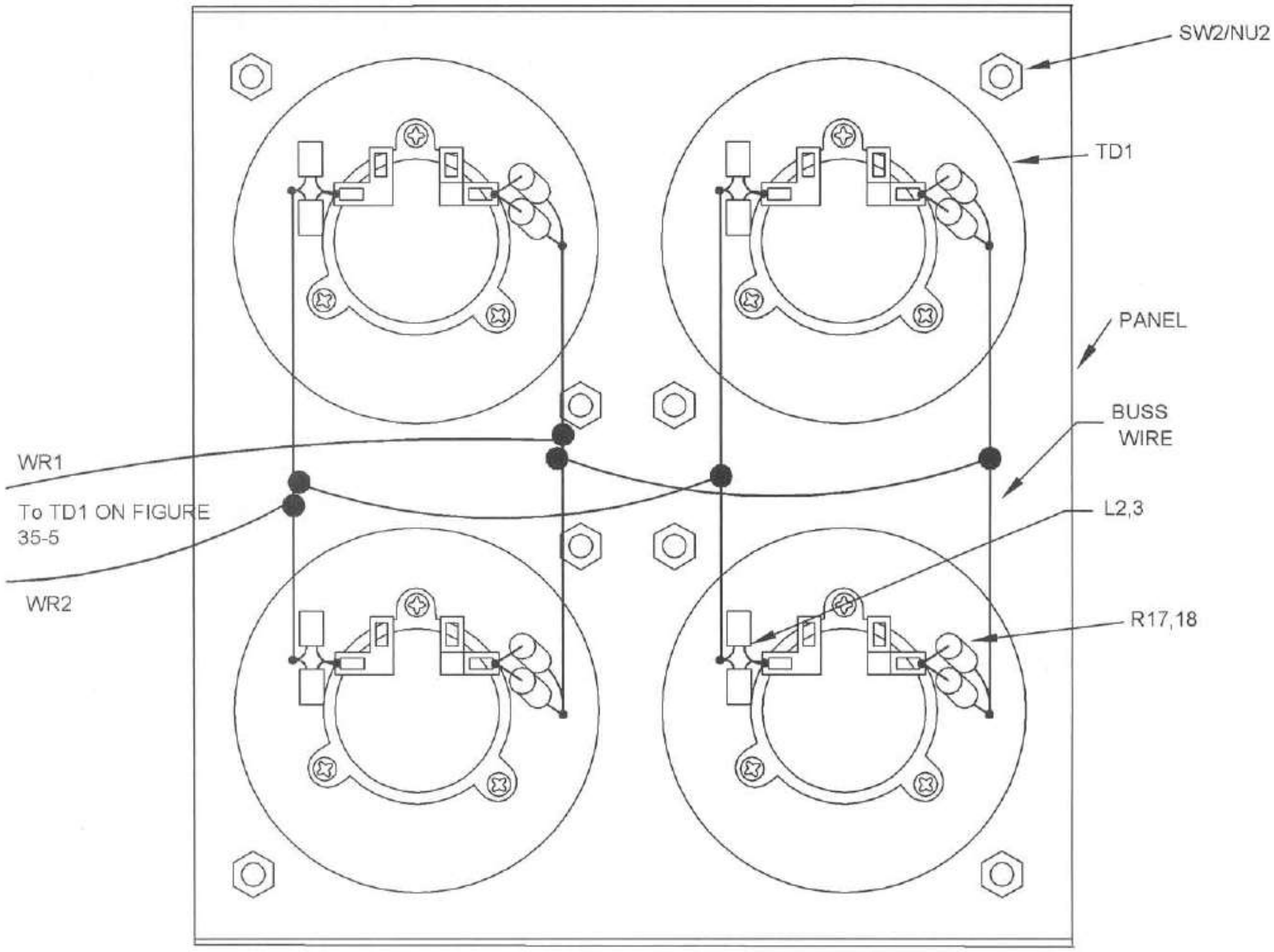

Figure 35-8 *Transducer panel wiring*

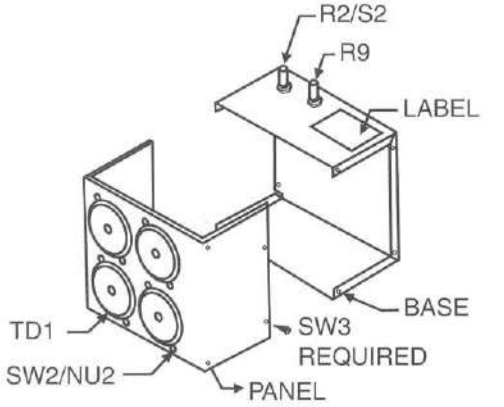

NOTE: 1/2" Lip flange with holes for attaching base to panel section via SW3 screws (8 required)

Figure 35-9 *Final assembly*

When used as a rodent device, we have found that the lowest tolerable frequency to humans usually has the greatest effect on the common species of rats. This is not always true but serves as a starting point in initially setting the adjustments. A frequency setting of 15 to 20 kHz for nonhuman living areas and 20 to 25 kHz for areas where people are present usually suffices for good rat control. Note that at no time should the unit be set higher than 25 kHz.

Dog control is usually around 18 kHz, but experimenting with the target animal may be necessary for optimum effect. Stubborn dog owners fall into the category of anti-intrusion if they will not cooperate.

b. Turn the SWP control off. Adjust the TONE control so that the output is detectable by the human ear or just above the point of annoyance.

c. Set the TEST switch to an audible verification. Note that experimentation is required for optimizing any effect on a given set of target applications.

Information on the System

Your phaser system is capable of operating in two modes. Mode one is at a frequency that is known to produce paranoia, nausea, disorientation, and many other physiological effects. Mode two allows using the system as an audible alarm to frighten off intruders or warn the user of an intrusion. Both modes may be used in combination and are easily rear-panel-controlled by the user.

A Word of Caution

Ultrasonics is a grey area in many respects when an application involves the control of animals or even a deterrent to unauthorized intrusion. It is always best to consult with local municipal and state laws before using this device to protect home or property. Remember many state laws lean more towards the right of the criminal rather than the victim.

General Information on Ultrasonics

Numerous requests have been made for information on the effect of these devices on people. First, let us make it clear that no device such as this should purposely be used on humans, and we discourage this use due to the possibility of acoustically sensitive people being highly irritated.

None of these devices has the ability to stop a person with the same effect as a gun, club, or more conventional weapon. They will, however, produce an extremely uncomfortable, irritating, sometimes painful effect in most people. Everyone will experience this effect to some degree. Unfortunately, younger women are much more affected than older men due to being more acoustically sensitive. The range of the device depends on many variables and is normally somewhere between 10 and 100 feet depending on the acoustical sensitivity of the target.

One possible use of the device (which deserves careful consideration) is the installation of all transducers in an area to protect against unauthorized intrusion. This would produce an irritating and painful feeling for the intruder, along with a condition of paranoia of not knowing what to expect next.

Application Supplement

The following describes an acoustical, ultrasonic device for dispersing potentially unruly crowds or gatherings. The reason for using the higher-frequency energy is that it is easily directed without being diffracted in all directions. Therefore, the use of directional-type transducers producing a front-to-back relative signal ratio of 20 to 30 dbs is easily achieved. The higher-frequency sound can produce nausea, headaches, and depression without the subject necessarily being able to determine the source. This effect no doubt presents an advantage to law enforcement when using these devices.

It is a known fact that the development of any type of demonstration in "psychological crowd control" follows the natural laws of growth and decay, in that a positive influx of people encourages a crowd growth, whereas a negative outflux creates a diminishing situation. One must realize that differences exist among most people regarding sound level and frequency sensitivity. When sampling a group or crowd as to the overall sensitivity of these devices, it is usually found that certain ones may not be affected to the same degree as others. However, taking into consideration that those who find the effect intolerable does generate the negative outflux that usually terminates the gathering.

This method, to this date, produces no permanent, lasting effects or symptoms and therefore becomes a more humane means of dispersing potentially unruly crowds. It should also be mentioned that younger people appear to be many times more sensitive to this effect. This usually goes hand in hand in most demonstrations, since younger people are more apt to become unruly.

This section encompasses many of the methods for applying this device via a permanent installation where transducers can be placed and adjusted to produce the necessary sound-pressure levels at distances to achieve the desired effects. This method could prove effective as an intrusion deterrent for a certain protected area.

Another method could be a handheld directional device similar to a megaphone where the amount of transducers used would be determined by the affects desired. This device would take into consideration the highly directional characteristics of this energy, being able to be directed at problem areas as a handheld portable system.

Another method for effective control would be clusters or arrays of transducers with these systems being mounted to vehicles via conventional roof mounts or on the hoods of these vehicles. Operation could take further advantage of the fair to good soundproofing found in most automobiles and vehicles could approach problem areas, taking advantage of close-proximity situations.

It has been found that certain frequencies also can be made to produce intense irritability in certain types of people. This method could justify the eventual use of more severe methods of restraint, assuming that the target subjects would eventually become more aggressive or violent, thus warranting this action. This is only to be considered as a potential extreme use of the device.

The objective here is to describe a similar method whereby the sound-pressure levels and frequencies can be obtained economically and efficiently, using state of the art methods and easily available parts, yet they can still be portable and easily handled for complete operational flexibility. Construction is based around the use of the piezoelectric tweeter transducers. These lightweight devices are easily driven and produce a conversion efficiency of over six times their electromagnetic counterpart. Being lightweight, economical, and easy to use when driving, they are an excellent candidate for this type of system. Higher driving impedance versus frequency is the inverse of the electromagnetic type.

The system, as described, contains several modes of operation. The first mode is a steady state of a certain frequency determined by the operator. This mode is termed *manual frequency control* and is adjusted by its appropriate control. The next mode is *low speed sweep*, which consists of the unit starting at a low frequency and automatically increasing to its higher limit where it again repeats itself. The remaining modes are controlled rates of this sweep, determined by the appropriate control referred to as *sweep rate control*. It is these controls that can be remotely operated from the inside of a vehicle when high-powered, outside arrays in clusters are used. As was mentioned before, certain frequencies combined with certain sweep rates can cause different degrees of effectiveness. It is not the objective of this information to analyze the potential types of effect and behavior versus control setting, but to describe the working system as a generator of these controlled sounds.

Table 35-1 Pain-Field Generator Parts List

Ref. #	*Description*	*DB Part #*
R1, 6, 8, 12	Four 1K, ¼-watt resistors (br-blk-red)	
R2/S2	500K, 1 meg pot/switch, 17 mm	
R3	2.2K, ¼-watt resistor (red-red-red)	
R4, 5	Two 10K, ¼-watt resistors (br-blk or)	
R7	10-ohm, ¼-watt resistor (br-blk-blk)	
R9/S3	10K pot/switch, 17 mm	
R10	5K horizontal trimpot	
R13–16	Four 4.7-ohm, 3-watt resistors (yel-pur-sil-gold)	
R17, 18, 19	Nine 120-ohm, 5-watt resistors (br-red-br); use two per transducer and one across S1	
C1, 4	Two .01-microfarad, 50-volt disk capacitors (103)	
C3	100-microfarad, 25-volt vertical electrolytic capacitor	
C2	10-microfarad, 25-volt vertical electrolytic capacitor	
C5A, B	Two .01-microfarad, 50-volt polyester capacitor	
C6	1,000-microfarad, 25-volt vertical electrolytic capacitor	
C8	.01-microfarad, 2-kilovolt disk capacitor	
C9	.22-microfarad, 250-volt polypropylene capacitor	
L2, 3	Eight 1 mh inductors (6080), 2 used per transducer, in place of R15 and R16 as marked on PC board	DB# 1MH
L1	Inductor: two Hitachi 30.48 E cores and mating bobbin assembled and wound L1 inductor	DB# PPPL1L1
Q1	PN2907 PNP GP transistor	
Q2	IRF530 or 540 N-channel metal-oxide-semiconductor field effect transistor (MOSFET)	
IC1, 2	Two 555 dual inline package (DIP) timers	
HS1	Heatsink bracket, created as shown	
SW1	6-32 × ¼-inch screw	
SW2	Eight 6-32 × ½-inch screw	
SW3	Eight #8 × ¼-inch sheet metal screw	
NU2	Nine 6-32 hex nuts	
FH1/FS1	Fuse holder and 2-amp fuse	
J1	Three-pin din chassis mount jack	
J2	DC 2.5 mm jack panel mounting	
CL1, 2	Two battery 9-volt clips	
BH1, 2	Two four-C-Cell holders	
S1	Single pole, single throw (SPDT) power switch	
S4	Small SPST toggle switch	
PC1	Printed circuit board (PCB) or use perforated vector board	DB# PCSONIC
TD1	Four polarized 130 db piezo transducer	DB# MOTRAN
WR1/2	36 inches of #20 vinyl hookup wire, red and black	
CHASSISBASE	Create as shown in Figure 35-6	
PANEL	Create as shown in Figure 35-6	
TAPE	8 inches of two-sided 1-inch foam tape for securing battery holders to base	
SPACER	3 × 3-inch piece of plastic sheet to insulate assembly board from metal BASE	
LABULTRA	Label, ultrasonic warning	

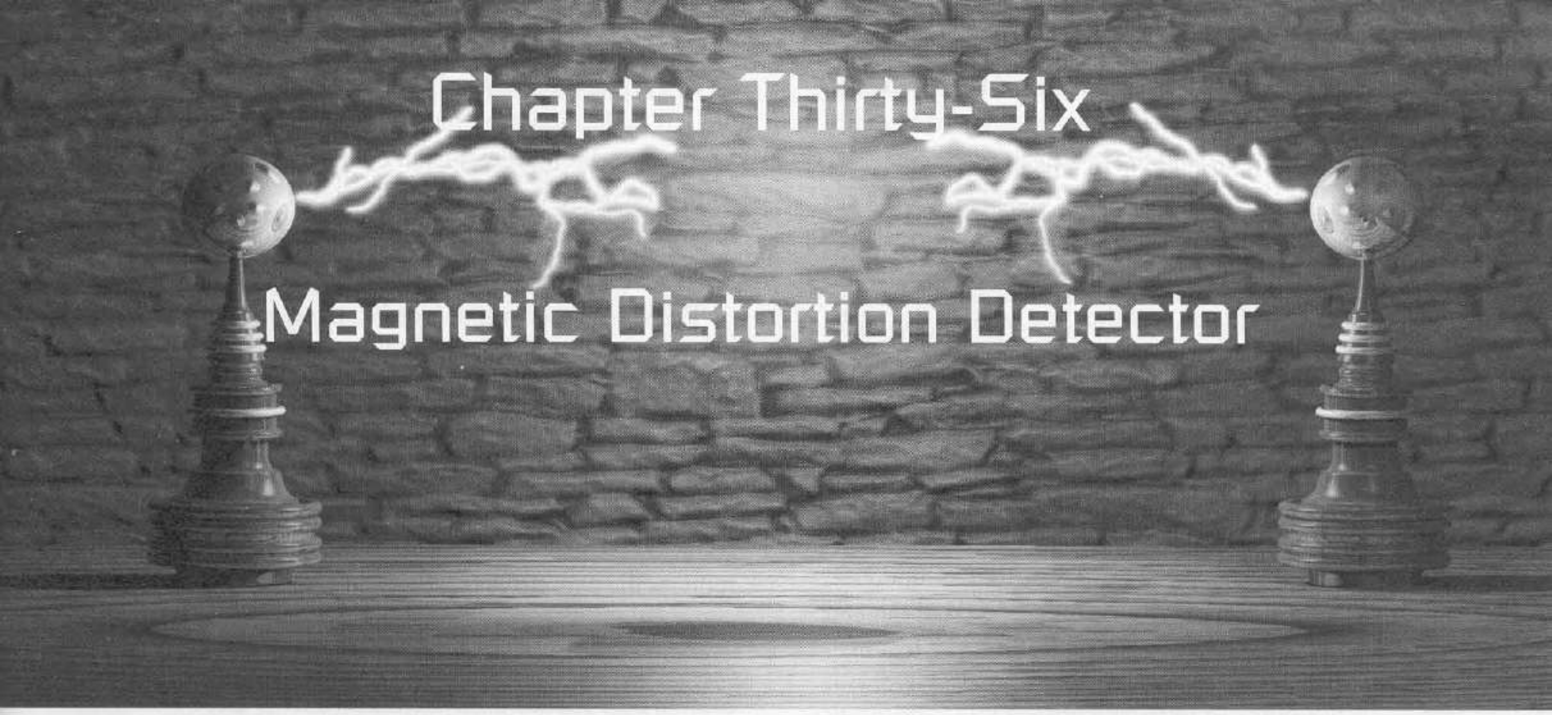

Chapter Thirty-Six

Magnetic Distortion Detector

The following project, as shown in Figure 36-1, describes an excellent, low-cost device that has amazing sensitivity in its ability to detect magnetic disturbances down into the microgauss. A microgauss is one millionth of a gauss. The earth's magnetic field is 500 milligauss where a milligauss is one thousandth of a gauss. The unit easily detects sunspot activity and auroras, and it can even be used as a vibration sensor that can detect a person's finger tapping on a rock. It can also detect moving aircraft and vehicles, as well as other changing electric and magnetic phenomena.

Actual detection is accomplished via an analog meter that responds to the relative strength of the changing magnetic fields. An audible alarm also allows the presetting of an activation level where it will alert the user of certain activity. An output port is included for connecting to a chart recorder.

The magnetic fields produced by many devices, both manmade and natural, are detectable by this system. Interesting results are produced when this unit is placed near anything electrical or magnetic such as a car, motor, or TV. The sensitivity is impressive when one sees the results of a 1-amp pulse of electrical current through a wire located a meter away from the unit. (This corresponds to a milligauss. Earth's magnetic field is .5 gauss). When the system is properly installed, it will detect passing automobiles, aircraft, or any moving, ferrous source that will cause the terrestrial magnetic field to distort. The unit is an excellent science fair project and can supply the amateur scientist with hours of knowledgeable entertainment in listening and detecting magnetic fields and their abnormalities.

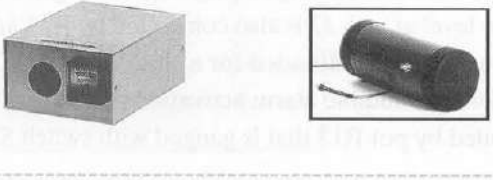

Figure 36-1 *The magnetic distortion detector*

Construction is shown to be as flexible as possible. It is built in two sections: the sensing head and the control box, which are interconnected via a length of shielded cable. The maximum usable sensitivity requires the sensing head to be placed away from power lines and other potential disturbances. The sensing head, which must be securely mounted to prevent interference from mechanical movement due to wind and vibration, can cause erroneous signals as it shifts position through the earth's magnetic field. The unit can also be made as a sensitive magnetic probe by installing the sensing head and the control box together, bearing in mind that the usable sensitivity will be decreased due to erroneous movement near magnetic materials when moving its position. The detection of field changes produces a moderately loud alarm along with a relative indication from an analog meter located on the panel. An optional jack is used for an X/Y chart recorder, and the system is fully battery powered.

This is an intermediate to advanced-level project requiring basic electronic skills. Expect to spend $45 to $65. All parts are readily available, with specialized parts obtainable through Information Unlimited at www.amazing1.com, and they are listed in the parts list at the end of the chapter.

Circuit Description

Figure 36-2 shows a coil (L1) consisting of many thousands of turns is wound on a 2-inch bobbin and placed over a soft iron core of about 1 to 2 feet in length. Any magnetic changes induce small voltages to occur at the output of this coil. This output signal is fed into a nano-amp, sensitive DC amplifier (I1), providing a current gain of 500 times the original. The iron core of L1 serves to concentrate the magnetic flux lines by offering a lower reluctance path than that of air. The input of I1 is protected by overvoltage and transients via clipping diodes D1 and D2. The output of I1 is forced to near zero via a multiturn trimpot (R4) for balancing and compensating any offset.

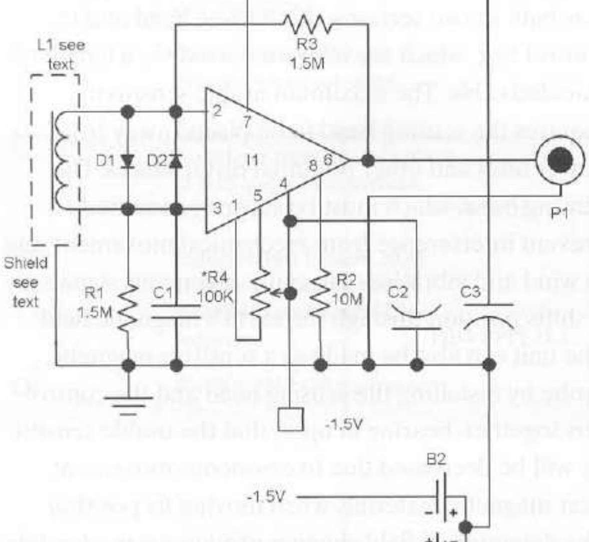

Figure 36-2 *Sense head schematic*

Power for I1 consists of batteries B1 and B2, which are left in the circuit permanently as current drain is very low and replacement can be on a yearly basis. It should be noted that if the sensing head is to be located in a hard-to-get area the builder might use D-size cells, as these will last that much longer.

The sense head output is fed to the control box via any reasonable length of shielded cable, WR10. As mentioned before, the sensing head should be located away from any potential magnetic interfering devices, such as anything electrical and large, ferrous objects.

Aluminum foil can be wrapped around the sense coil and is connected to the circuit common line. This reduces the capacitive pickup of unwanted signals.

Again, the sensing head and control box may be assembled together and used as a probe. This configuration may reduce the system's sensitivity due to erroneous movements and interaction of the circuits such as magnetic speakers and meters. This method is not encouraged due to these erroneous readings.

Figure 36-3 shows the control box providing another amplifier, I2, with a gain of 200 times the original amount, which gives the system an overall gain of 100,000. Full-scale meter deflection of the 50 ua meter (M1) occurs when a current of .5 nano-amp is obtained from L1. It is obvious that with the many thousands of turns on L1 that a detectable signal can occur with very minute flux changes, such as the result of changes in the microgausses. It should be noted that the earth's magnetic field is only .5 gauss.

I1 is connected as a conventional DC operational amplifier with its output DC coupled to I2 via a shielded cable. This cable is necessary to separate the sense head and the control box by at least 6 feet. This is necessary to prevent magnetic coupling from the speaker SPK1. Sensitivity control pot R5 controls the system gain and is usually set at maximum.

You will note that meter deflection pot R11 also is ganged with switch S1, which controls the +6-volt power to the control box section. R11 controls the deflection of meter M1 with diode D3 and capacitor C6, serving to smooth out and average the signal. The voltage level at jack J2 is also controlled by R11 and is an auxiliary jack intended for a chart or other recorder. The audible alarm activation point is controlled by pot R12 that is ganged with switch S2.

Switch S2 controls the −6 volt power to the control box.

The alarm is nothing more than a reflex-type oscillator consisting of transistors Q2 and Q3 with its feedback paths consisting of resistor R15 and capacitor C8. Trimpot resistor R16 is set for a reliable, clean-sounding signal.

Assembly Steps

1. Figure 36-4 shows coil L1 wound with thousands of turns of fine magnet wire. This may be difficult to do and it is suggested that the builder locate a high-voltage oil burner ignition, neon sign, bug killer, or other similar transformer and carefully remove the high-voltage secondary winding. The parameters of these transformers are usually around 5,000 volts at 10 to 20 milliamps. These are available from junkyards and other places. Many are rejected due to radio and TV interference and yet have intact secondary winding suitable for this project.

   The objective is to obtain one with as many turns as possible; however, most will work quite well. A quick test is to measure the DC resistance of between 5 and 20K across the coil. A relay coil of high impedance such as 5 to 10K may also work. The coil must have connecting leads and these should be secured with *room temperature vulcanizing* (RTV) adhesive, wax, or something similar to prevent breaking that may render the coil useless. A soft iron core of 1 to 2 feet is inserted through the coil and serves to offer a lower reluctance to the magnetic flux lines. It is these lines that, as they change in the core, induce a voltage due to magnetic induction. Note to shield the coil from electrical and capacitive pickup by wrapping it in aluminum foil and connecting the foil to the circuit ground. A special custom-made coil is available from Information Unlimited and is noted in the parts list.

2. Lay out and identify all the parts and pieces, and check them with the parts list. Note that some parts may sometimes vary in value. This is acceptable as all components are 10 to 20 percent tolerant unless otherwise noted.

3. Assemble the board by inserting components into the board holes as shown in Figure 36-5.

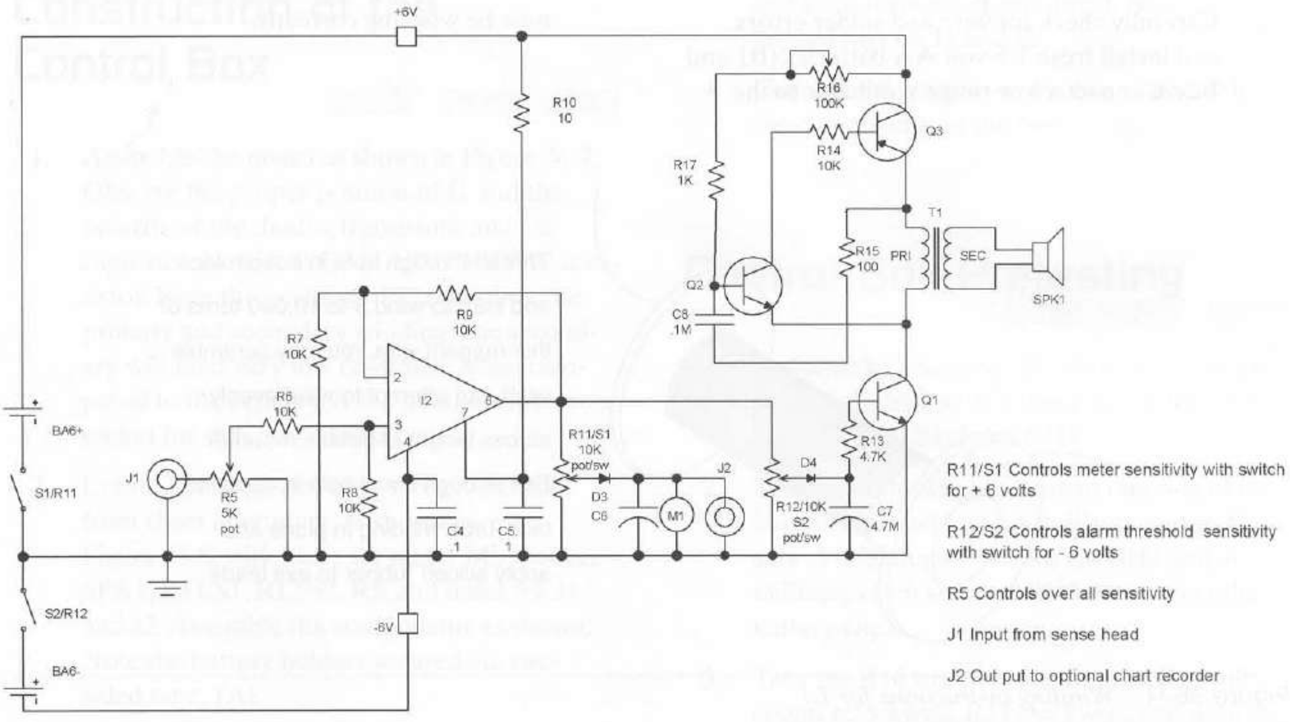

R11/S1 Controls meter sensitivity with switch for + 6 volts

R12/S2 Controls alarm threshold sensitivity with switch for - 6 volts

R5 Controls over all sensitivity

J1 Input from sense head

J2 Out put to optional chart recorder

Figure 36-3 *Control box schematic*

Proceed from left to right, attempting to obtain the layout. Note the polarity of the components as indicated. Certain leads of the actual components are to be used for connecting points and circuit runs. Do not cut or trim them at this time. It is best to temporarily fold the leads over to secure the individual parts from falling out of the board holes for now. Solder as required and clip off excess leads. Some kits contain a length of insulated wire. This must be cut, stripped, and tinned according to where it is used.

4. Wire in the 6-foot length of shielded cable, being careful not to burn through the insulation and short out the cable.
5. Attach the battery clip HOLD1 using silicon rubber.
6. Attach the L1 coil to the PB1 perforated board as shown in Figure 36-5 and secure it with RTV adhesive. You will have to drill a hole for the iron core. The core should fit snuggly and can be glued in place in the final assembly.

Pretesting the Sense Head

1. Carefully check for wire and solder errors, and install fresh 1.5-volt AA batteries (B1 and B2). Connect a low-range voltmeter to the leads of the output cable and adjust the 25-turn pot (R4) for as close to a near-zero reading as possible. Finally, adjust the lowest range of the meter, preferably the 50 microamp range. This should be done outside away from power lines.
2. With the meter still connected, observe a reading when a piece of iron is moved near sensing coil L1 and moved about. If the piece of iron contains any magnetism, the effect will be enhanced.
3. Wrap the sense coil with several layers of aluminum foil and ground the foil to the negative circuit. Diodes D1 and D2 will require shading against light sources as the junction behaves like photodiodes at these low signal levels.
4. Create enclosure EN1, as shown in Figure 36-6, from an 8 × 3½-inch outer-diameter piece of plastic tubing.

   Note the small, ⅜-inch access holes for adjusting R4 are drilled when all is in place. Create the brackets as required for mounting the sense head.
5. Install the assembly into ENl with caps CPl and CP2 as shown. Recheck it by connecting the meter across P1. Readjust R4 through the access hole if needed. The sense head should now be working correctly.

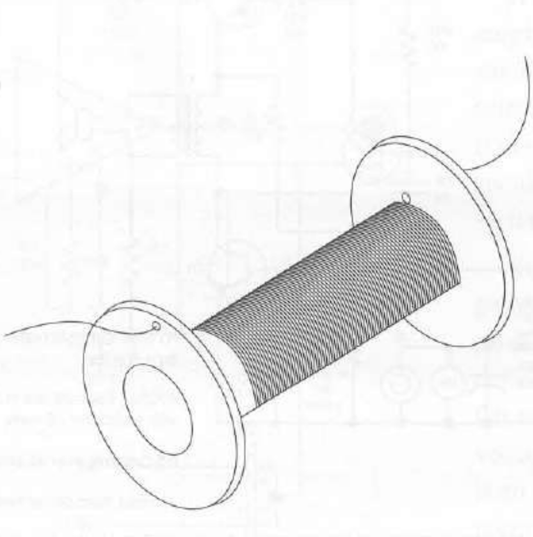

Thread through hole in bobbin face and start to wind 5 to 10,000 turns of thin magent wire. You may scramble wind, but attempt to wind evenly across length of bobbin mandrel.

Exit through small hole in opposite face. Tape winding in place and apply silicon rubber to exit leads.

Figure 36-4 *Winding instructions for L1*

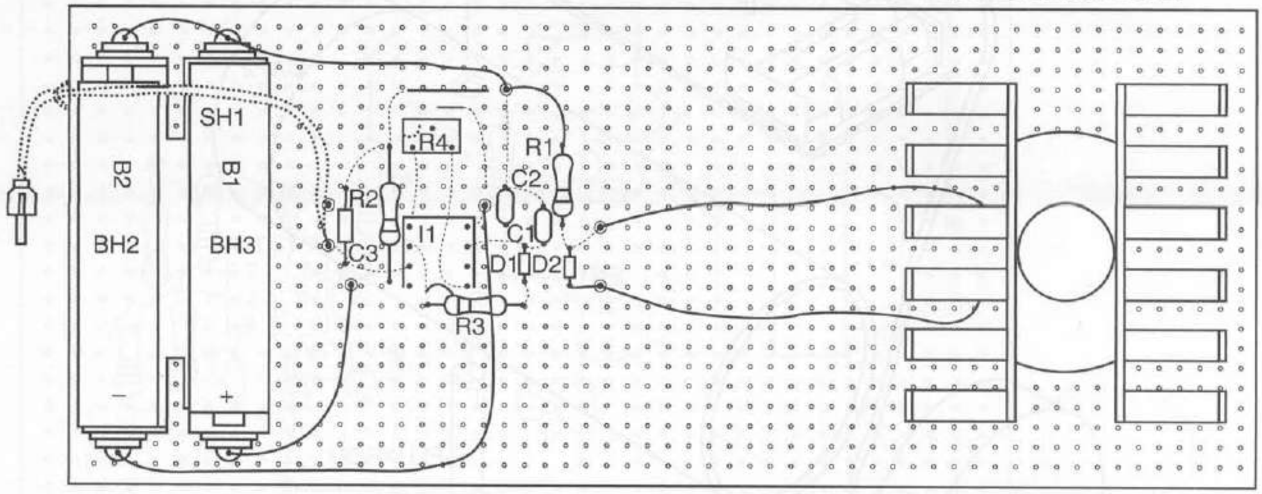

Strip off outer plastic jacket of cable 1 inch and twist braid into a lead and tin. Strip off 1/4 inch of insulation on inner lead and tin. Insert into plug and solder inner lead to lin tip.

Carefully solder twisted braid lead to outer shell being careful not to burn through insulation. Always verify lead continuity and possible shorting of the leads with a meter.

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

- Small dots are holes used for component insertion.

⊙ Circles with dots are holes that may require drilling for component mounting and points for external connecting wires, indicating strain relief points. It is suggested to drill clearance holes for the leads and solder to points beneath board.

Note: The sense head as shown in this sketch is wound from a transformer bobbin. Any plastic bobbin 2"x1" will usually suffice. See Figure 36-4.

Figure 36-5 *Assembly of the sense board section*

Construction of the Control Box

1. Assemble the board as shown in Figure 36-7. Observe the proper position of I1 and the polarity of the diodes, transistors, and C6 capacitor. Note that Q3 is a PNP PN2907 transistor. Note the position of Tl regarding the primary and secondary winding. The secondary will read very low resistance when compared to the primary. It is a good idea to use a socket for the integrated circuit.
2. Create the enclosure base section (BASE) from sheet aluminum, as shown in Figure 36-8, with holes for meter Ml, speaker SPK1, R11/S1, R12/S2, R5, and holes for J1 and J2. Assemble the components as shown. Note the battery holders secured via two-sided tape, TAl.
3. Attach the interconnecting leads, as shown in Figure 36-9. Twist the leads whenever possible and keep them short and direct. Note the ground lugs on J1 and J2.
4. Connect the leads from the board assembly to the components in the box.

Control Box Pretesting

1. Position the board, but do not secure it at this point. Turn S1 and S2 off and insert four AA batteries into BH2 and BH3.
2. Turn on both switches, unsnap one side of the battery clip, and insert a milliamp meter. Measure .5 milliamps (−6 volts) at BH2 and .8 milliamps (+6 volts) at BH3. Reconnect the battery clips.
3. Turn the R16 trimpot to midrange, R11 full counterclockwise, R12 clockwise, and note the meter remaining at zero and the alarm

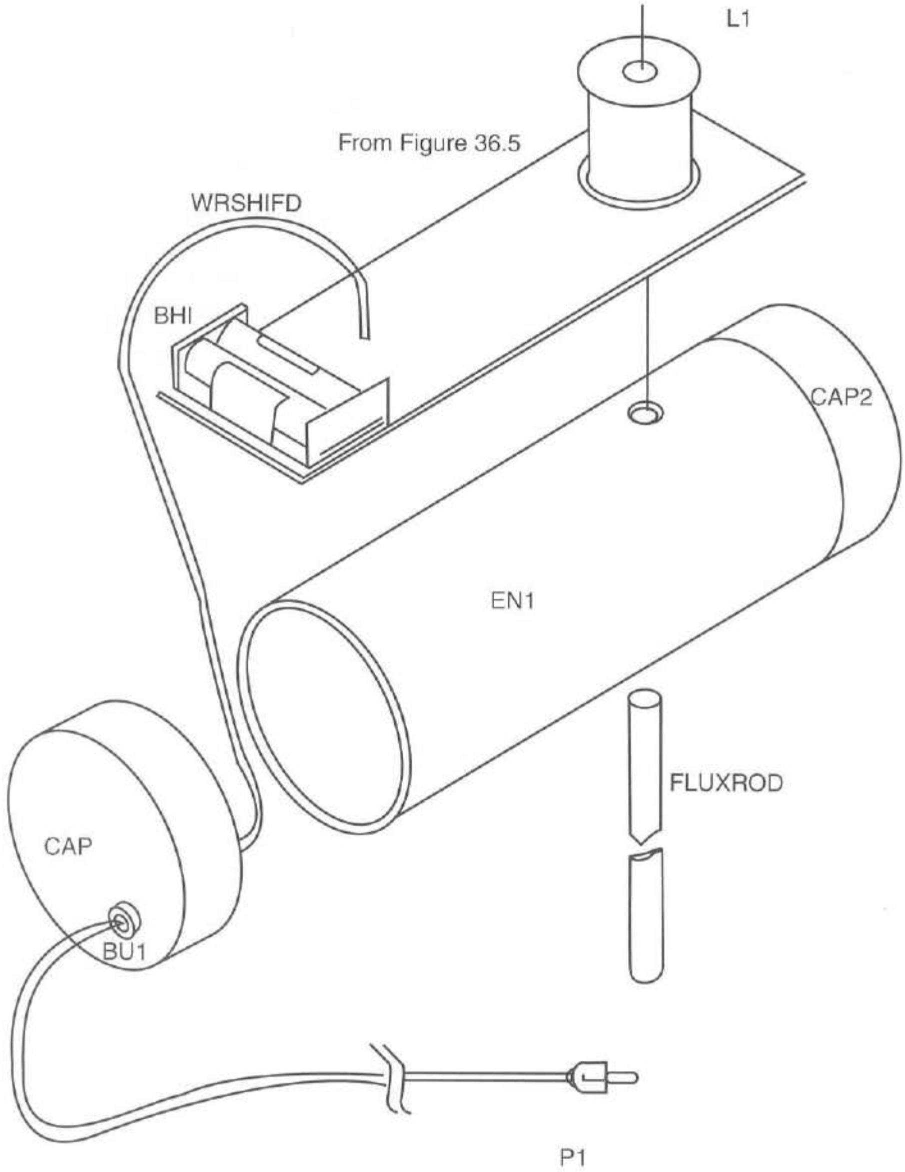

Figure 36-6 *Sense head enclosure assembly*

remaining silent. Now turn R11 clockwise and note the meter deflecting to full scale.

The following steps will require the sense head located away from electrical wiring and other potential sources of magnetic interference. The distance should be at least 5 feet from the control box to avoid feedback from the meter and speaker magnets.

4. Turn both R11 and R12 full clockwise. Plug in P1 from the sense head J2. Note the meter may pin, or jump off scale, and the alarm may sound. Have a friend slowly rotate R4 on the sense head to attempt to bring the meter into range. You may have to desensitize the circuit by turning R11 counterclockwise until the meter reads midrange and turn R12 to deactivate the alarm as it can be annoying during testing. Note that R5 is an overall sensitivity control and should be turned fully on in most applications for maximum sensitivity. The meter and alarm can be ranged by their respective controls, R11 and R12.

When properly set, the reading will constantly fluctuate and is the result of sunspots or other phenomena distorting the earth's magnetic field.

Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

• Small dots are holes used for component insertion.

⊙ Circles with dots are holes that may require drilling for component mounting and points for external connecting wires indicating strain relief points. It is suggested to drill clearance holes for the leads and solder to points beneath board.

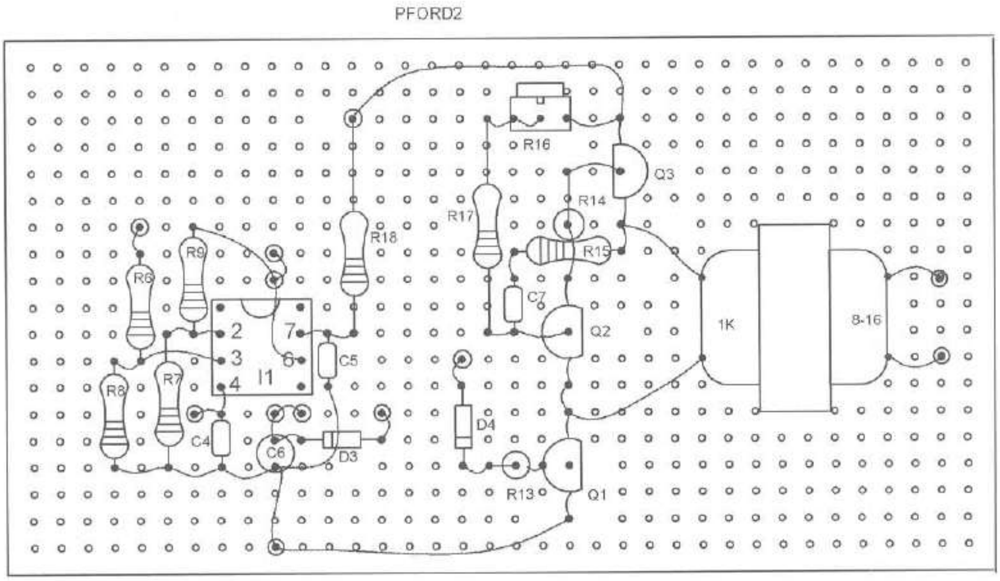

Figure 36-7 *Assembly control board*

If you choose a magnetically noisy location, you will have an added signal as the result of power lines, motors, and other magnetic disturbance. Note that if R11 cannot bring the meter into range without overly desensitizing it, it may be necessary to internally readjust R4 for proper ranging.

5. Preset R14 at midrange and adjust R12 until the alarm sounds. Readjust R16 for a clean, desirable-sounding tone if necessary.

   Appreciating the sensitivity of this unit requires operation in a magnetically quiet location where all the controls can be advanced fully. Under the conditions of a low magnetic background, the unit can detect automobiles, manmade devices, and natural magnetic disturbances. Also, simply moving a piece of iron near the sensing head should produce an indication. (Note low-level-detection sensitivity may be difficult during the peak of the sunspot cycle due to the high background signal fluctuations.)

   Jack J2 may be used to drive an XY recorder when observing sunspots or other time-related disturbances.

Note: No steps were taken to filter out 60 HZ disturbances, as it appears that many uses of this device involve the detection of power lines and other related equipment.

Applications

The uses of this device are obviously only appreciated by those who can understand its sensitivity to a minute change in magnetic flux. The unit can be made to detect the motion of large steel bodies as they distort the existing earth's magnetic field. Many applications and experiments are therefore possible with the device. Any high-impedance headphones

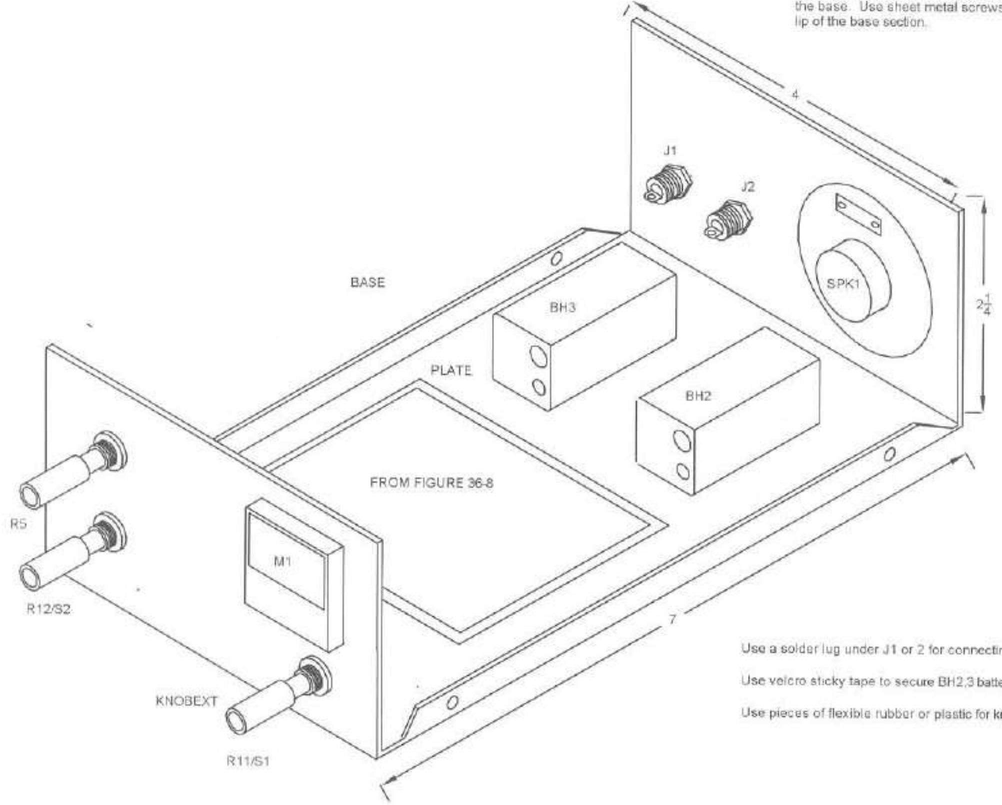
4
J1
J2
SPK1
2¼
BASE
BH3
PLATE
BH2
FROM FIGURE 36-8
R5
M1
R12/S2
7
KNOBEXT
R11/S1

Fabricate a cove to fit the base section from any suitable material. Allow a over hang over the front and rear panels of the base. Use sheet metal screws to secure to the folded up lip of the base section.

Use a solder lug under J1 or 2 for connecting to base ground.

Use velcro sticky tape to secure BH2,3 battery packs

Use pieces of flexible rubber or plastic for knob extensions

Figure 36-8 *Control box base section fabrication and layout*

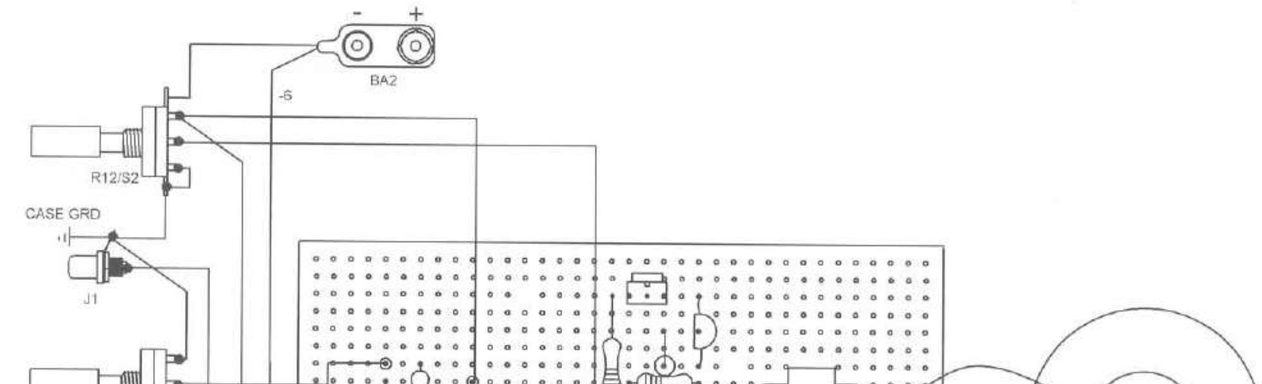
-
+
BA2
-6
R12/S2
CASE GRD
J1
R5
CASE GRD
J2
2
8
4
5
CASE GRD
R11/S1
-
+
BA1
+6
+
-
M1 METER
SPK1 SPEAKER

Figure 36-9 *Final assembly control box*

can be plugged into J2 where the experimenter can actually "listen in" and hear the field changing.

One verifying test that is easily done to determine the system's ability to detect the earth's magnetic field is the following. Rotate the sensing head with short, quick jerks in a clockwise direction. Note the meter pinning, and as rotation is continued a point will be found where the meter will respond less. This nulling point should occur with the fluxrod pointed in a north/south direction. If rotation is continued, it will be found that the short, quick jerks must occur in a counterclockwise direction to obtain a positive meter movement. This is similar to an electrical generator using the earth as its field magnet and the sensing head core as the moving armature.

Optional Ultra-Sensitive Vibration and Tremor Detector

You may easily convert this device to a vibration detector by simply modifying the sense head so it can be mounted to a rigid object such as a rock. The trick now is to place a magnet adjacent to the end of the sense coil, L1. The magnet should be held in place by elastic or other means whereby it is mechanically isolated relative to the sense coil. It should be very close to the sense coil for optimum sensitivity.

The theory is simple. Positioning a powerful magnet near a coil with many turns produces a signal. It is now the relative motion between this magnet and the coil that produces a voltage by the familiar equation $E = Nd0/dt$, where E is volts, N is the coil turns, 0 theta is the flux of the magnet, and t is the rate of relative motion in seconds. The combination of the high-resultant signal volts and the amplification sensitivity of the circuit provides amazing results.

Many variations are possible with the common objective being to obtain the highest reading for a given tremor. This unit will easily detect earthquakes occurring anywhere on the planet; however, the builder must experiment with motion dampening and other schemes to differentiate between manmade vibrations in order to isolate these signals. This is an excellent, low-cost project for any interested individual to learn by experimentation.

Table 36-1 Magnetic Distortion Detector Parts List

Ref. #	*Description*	*DB Part #*
R1, 3	Two 1.5M, ¼-watt resistors (br-gr-gr)	
R2	Two 10M, ¼-watt resistors (br-blk-bl)	
R4	100-kilo-ohm, 10-turn vertical trimmer	
R5	5K linear pot	
R6, 7, 8, 9, 14	Five 10K, ¼-watt resistor (br-blk-or)	
R10	10-ohm, ¼-watt resistor (br-blk-blk)	
R11/S1, 12/S2	Two 10K linear pots and switches	
R13	4.7K, ¼-watt resistor (yel-pur-red)	
R15	100-ohm ¼-watt resistor (br-blk-br)	
R16	100-kilo-ohm vertical trimmer	
R17	1-kilo-ohm, ¼-watt resistor (br-blk-red)	
C1, 2, 3, 4, 5, 8	Six .1-microfarad, 50-volt plastic capacitors	
C6	1-microfarad, 50-volt vertical electrolytic capacitor	
C7	4.7-microfarad, 50-volt vertical electrolytic capacitor	
I1	LM4250 nano-amp operational amplifier in a dual inline package (DIP)	
I2	LM741 operational differential amplifier in DIP	
D1, 2, 3, 4	Four IN914 small-signal diodes	
Q1, 2	Two PN2222 NPN GP silicon transistors	
Q3	PN2907 PNP GP silicon transistor	
TI	1K to 8-ohm, small audio transformer	
M1	50 uamp panel meter	DB# METER50S
SPK1	2-inch, 8- to 16-ohm speaker	
J1, 2	Two RCA panel-mount phono jacks	
P1, 2	Two RCA mating phono plugs	
CL1, 2	Two 9-volt battery clips	
SH1	3 feet of single-conductor, shielded mike cable	
WIRE24	4 feet of #24 vinyl stranded hookup wire	
BH1	Two-AA-battery holder	
BH2, 3	Two 4-AA-battery holders	
BU1	Small, plastic feed-through bushing	
L1SENSE	Sense coil per Figure 36-4	DB# SENSE
PBORD1	6¼ × 2½-inch, .1 × .1-inch grid perforated vector board	
PBORD2	4½ × 2-inch, .1 × .1-inch grid perforated vector board	
EN1	7½ × 3-inch schedule 40 PVC tubing	
CAP1	Two 3½-inch black plastic caps	
TAP1	4 inches of 4 × 1-inch strips of Velcro sticky tape	
KNOBEXT	4 inches of 3/16-inch inner-diameter plastic or rubber tubing for shafts	
FLUXROD	6 inches of ¼ × 20-inch steel-threaded rod and two mating nuts	
BASE	Created as shown in Figure 36-8 from .063 aluminum	
COVER	Cut to fit BASE section	
PLATE	5 × 2½-inch piece of insulating plastic to mount under board	

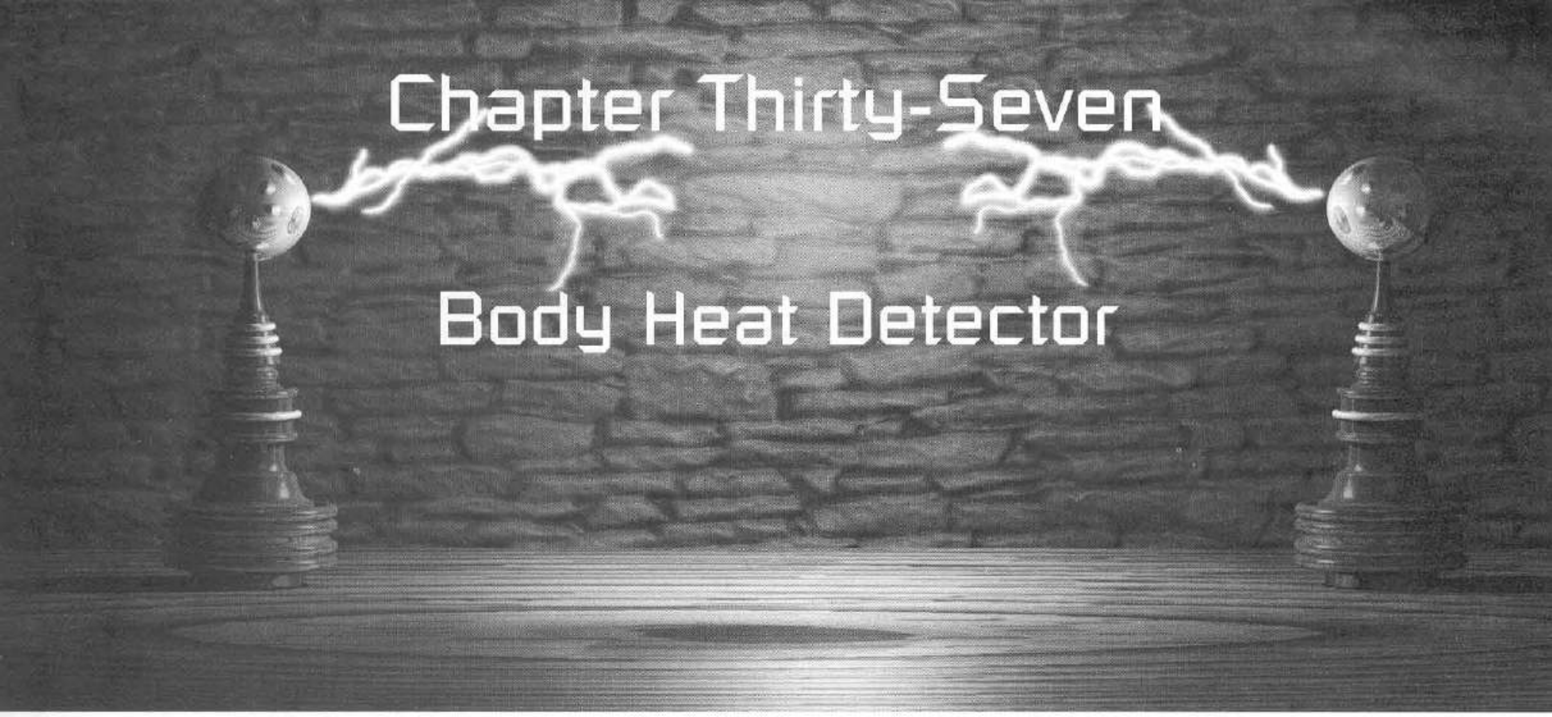

Chapter Thirty-Seven

Body Heat Detector

This project, as shown in Figure 37-1, shows how to construct a useful device that can detect human or animal motion up to several hundred meters. It provides a narrow field of view and can be used for paths and walkways. Farmers use these devices to detect when a predator comes into range, taking the necessary action when the alarm sounds. It may also be used as a medium-sensitive thermal detector for lab and research use. The device is an excellent science project demonstrating medium-sensitivity detection of relatively low temperatures.

This is an intermediate-level project requiring basic electronic skills. Expect to spend $30 to $45. All the parts are readily available, with specialized parts obtainable through Information Unlimited at www.amazing1.com. These are listed in the parts list at the end of the chapter.

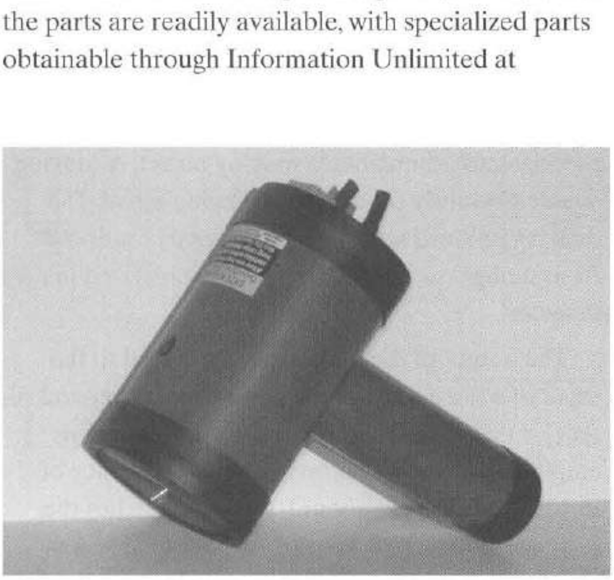

Figure 37-1 *Body heat detector*

Description

Your body heat detector is a device capable of sensing an object at a temperature increase slightly over ambient. It is this property that allows it to detect a living body at a considerable distance. The sensitivity of the device corresponds to a black body wavelength of 10 microns, corresponding to a temperature of 100° Fahrenheit. Black-body radiation defines color temperature in wavelengths below what is visible to the naked eye. All objects above absolute zero radiate energy as a function of their temperature to the fourth power. This radiation decreases in wavelength as the radiating body becomes hotter. This radiation will consist of the visible color spectrum for a given range of temperatures. Hotter bodies will show visible color such as blue being on the hotter side of red. A body hot enough to glow red is around 900 degrees F. Black bodies radiate in the infrared (invisible) and lower as the generating temperatures are near ambient being above 70 degrees F. All objects above absolute zero radiate energy as a function of the fourth power of temperature, and body heat energy is in the infrared region at around 100° degrees F.

The applications for this device are numerous and as an example you could a create motion detection system for home security. These are normally short-range, wide-angle detection systems and are limited in performance. Our system would be mounted similar to an optical scope on a rifle or a tripod. A living target now can easily be ferreted out in darkness or under moderate cover due to heat reflections. A predator, as it approaches its target victim, is easily detected, alerting the farmer or rancher to take the next step of action. Garden pests, as they emerge from their burrow, and downed game can easily be detected by the device.

The ability to detect hot spots is an extremely useful function. An example of sensitivity can easily be demonstrated by the following. Obtain a piece of cardboard or plastic approximately 6 × 12 inches. Place your hand on one end of this piece for several seconds to produce a hot spot of your hand print. Then place the piece in front of the unit, noting a detected signal increase on the surface previously touched by your hand due to the heat stored. (Don't touch the other end of the piece). Note that as time continues, the hot spot dissipates into the object thereby reducing the temperature differential and the signal pickup on the cardboard piece becomes evenly distributed. The differences in temperature detection capability is in the millidegrees. You can now realize the many useful applications of this unit.

Body Heat Detector Modes

Your body heat detector operates in two modes:

- **MODE A:** This mode is used to detect moving bodies such as animals, people, and so on. It is intended for sensing intrusions along pathways and into areas where a narrow field of view is necessary at long ranges (100 meters or so). This mode is the most stable of the two and rejects steady background sources of thermal energy. Detected bodies, however, must be moving.
- **MODE B:** Operation in this mode allows the detection of stationary hot spots and is most useful with a "cold background" such as at night. It is useful for ferreting out thermal sources such as hidden persons and animals. Hot object spots and areas are easily detected, but unfortunately this mode is so sensitive that often signals appear to be erroneous and mask the target object.

Both modes of operation require a camera tripod for reliable and simple operation. Very often a panning action can be used to reject or detect the desired targets.

Circuit Description

The heart of the detector is a pyroelectric detector (PYR1) capable of detecting very minute temperature changes. This device is responsive to temperatures of 100° Fahrenheit equivalent to approximately 10 microns. Unfortunately, its response speed is relatively slow, peaking at about .1 Hz. This is shown in Figure 37-2.

The pyroelectronic effect utilizes dual crystals of lithium tantalum oxide (LiTa 02). These crystals are polarized in an electric field below the Curie point. Electrodes attached to the crystal become electrically charged and eventually neutralize due to the surroundings. When the neutralized crystal is exposed to an infrared signal, its temperature changes, producing a charge on the electrodes and hence a detected signal. This signal is amplified by a built-in, low-noise *field effect transistor* (FET) and becomes the output. It is important to note that only a changing infrared signal can produce output; therefore, a chopper must be used for stationary sources. A chopper is nothing more than a mechanical wheel that has vanes to periodically interrupt the heat radiation signal to the pyrodetector, simulating a moving object. A moving source obviously produces a changing signal. The dual crystals will also cancel out signals produced from sunlight as they are oppositely polarized in the detector.

The output of the detector is connected to the input of a low-frequency, dual-stage amplifier and filter (U1) with a gain of approximately 2,500. This amplifier is designed to respond to a frequency of .1 to 10 Hz, peaking at about 1 Hz. This matches the response of the pyrodetector and the signal commonly generated by a living, moving body in the field of detection.

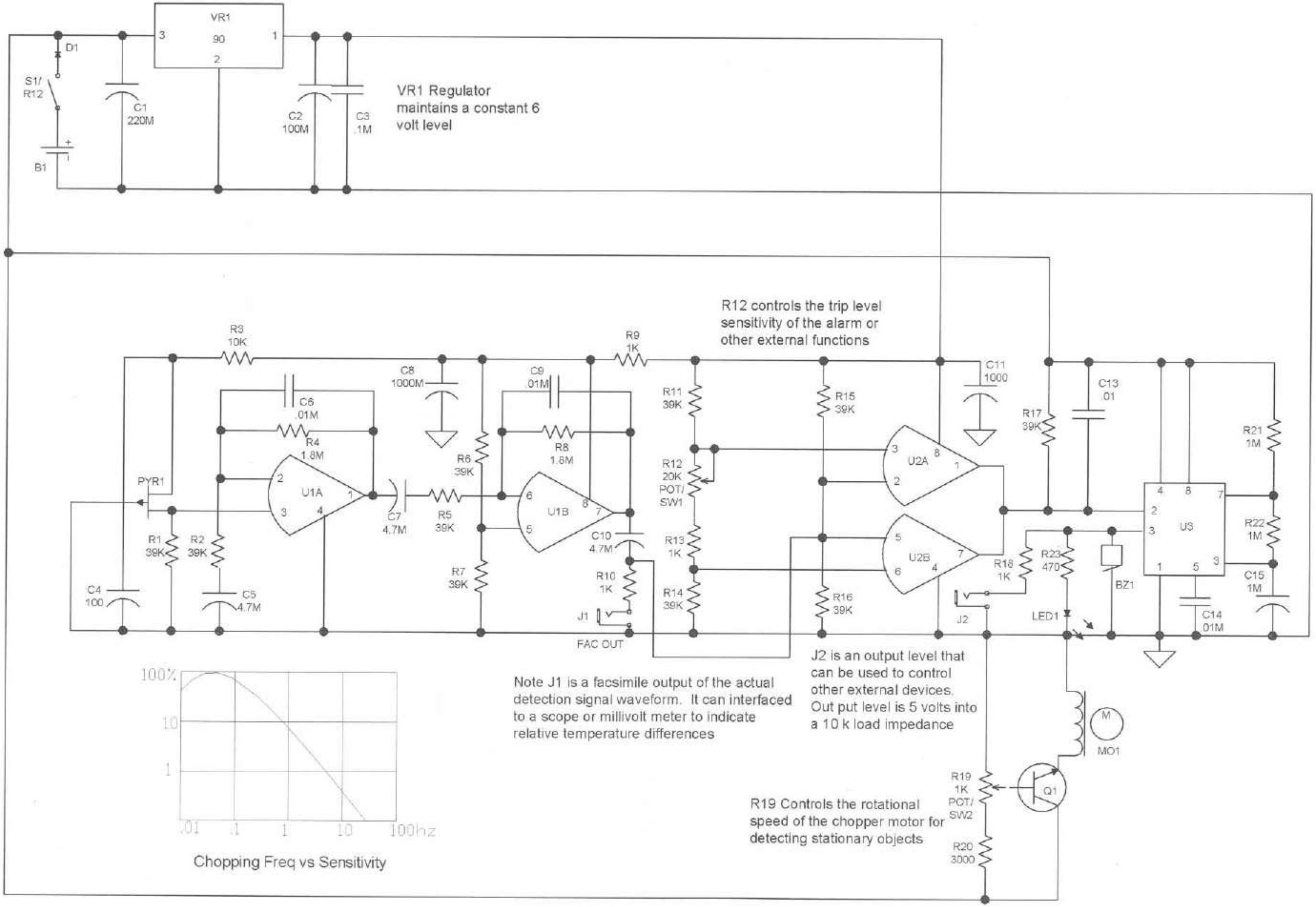
VR1
90
3
1
2
D1
S1/
R12
B1
C1
220M
C2
100M
C3
.1M
R3
10K
C6
.01M
R4
1.8M
U1A
PYR1
R1
39K
R2
39K
C4
100
C5
4.7M
C7
4.7M
R5
39K
C8
1000M
R6
39K
R7
39K
C9
.01M
R8
1.8M
U1B
C10
4.7M
R10
1K
J1
FAC OUT
R9
1K
R11
39K
R12
20K
POT/
SW1
R13
1K
R14
39K
R15
39K
R16
39K
U2A
U2B
C11
1000
J2
R17
39K
C13
.01
R18
1K
R23
470
LED1
BZ1
U3
R21
1M
R22
1M
C15
1M
C14
.01M
MO1
M
Q1
R19
1K
POT/
SW2
R20
3000
100%
10
1
.01
.1
1
10
100hz

VR1 Regulator maintains a constant 6 volt level

R12 controls the trip level sensitivity of the alarm or other external functions

Note J1 is a facsimile output of the actual detection signal waveform. It can interfaced to a scope or millivolt meter to indicate relative temperature differences

J2 is an output level that can be used to control other external devices. Out put level is 5 volts into a 10 k load impedance

R19 Controls the rotational speed of the chopper motor for detecting stationary objects

Chopping Freq vs Sensitivity

Figure 37-2 *Circuit schematic*

The changing AC output of the dual-signal amplifier (U1A and U1B) is capacitive coupled to output jack J1. It is this signal that is measured by an external AC millivolt meter and provides an analog indication of signal strength. This signal may be further processed for ultra-low-sensitive detection or maybe demodulated where the resulting DC level can control a VCO for audible detection via a changing tone. This AC signal is also fed to the dual plus and minus ("window") comparator (U2A and U2B). A threshold control (R12) presets the trip voltage or sensitivity of activation. The output of these two comparators is "anded" together and fed to the trigger input of the timer (U3). The output of U3 now activates audible buzzer BZ1 and visual indicator LED1. The buzzer on time is controlled by timing resistors R21 and R22 and capacitor C15.

An optional motor speed control for the chopper consists of Q1 and speed control pot R19. Jack J2 provides a 5-volt level through 1K resistor R18 and is used to trigger other systems.

Assembly Instructions

1. Lay out and identify all the parts and pieces, and verify them within the parts list at the end of the chapter.
2. Examine the PC board, as shown in Figure 37-3 and 37-4.
3. Proceed to assemble the device by inserting and soldering all the fixed resistors.
4. Insert and solder the capacitors, carefully noting the polarity on all the electrolytics: C1, C2, C4, C5, C7, C8, C10, C11, and C15. The other capacitors are not polarized.
5. Insert and solder in the semiconductors (D1 and VR1) and Q1. Note the polarity.
6. Identify and note the position marks on the integrated circuits U1, U2, and U3. Insert and carefully solder them, avoiding solder shorts between the pins. It is suggested to always use sockets for these components.
7. Identify the pots R12 and R9 and solder them in place. Note the switch section, S1, of R12. These connections are made using short pieces of cut leads to the pads directly beneath. Note these pots must be properly installed to allow final alignment with the PANEL plate.
8. Insert the wire jumps using short pieces of the cut leads of components that were used priorly.

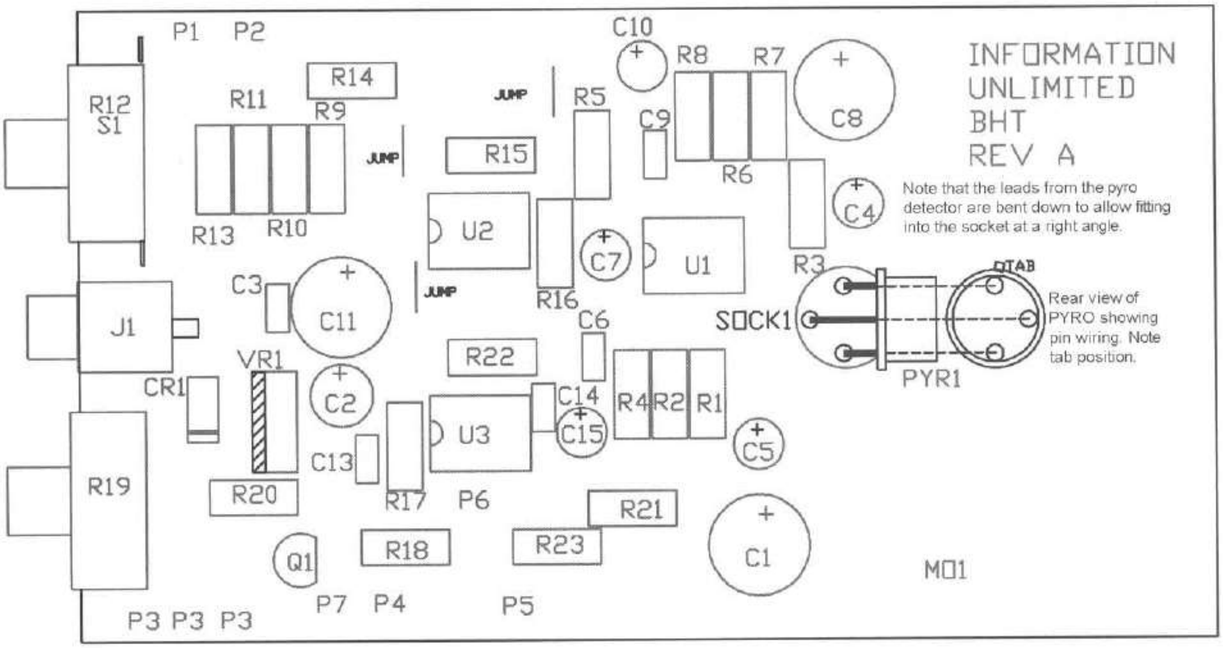

Note that the leads from the pyro detector are bent down to allow fitting into the socket at a right angle.

Rear view of PYRO showing pin wiring. Note tab position.

Figure 37-3 *PCB wiring and assembly*

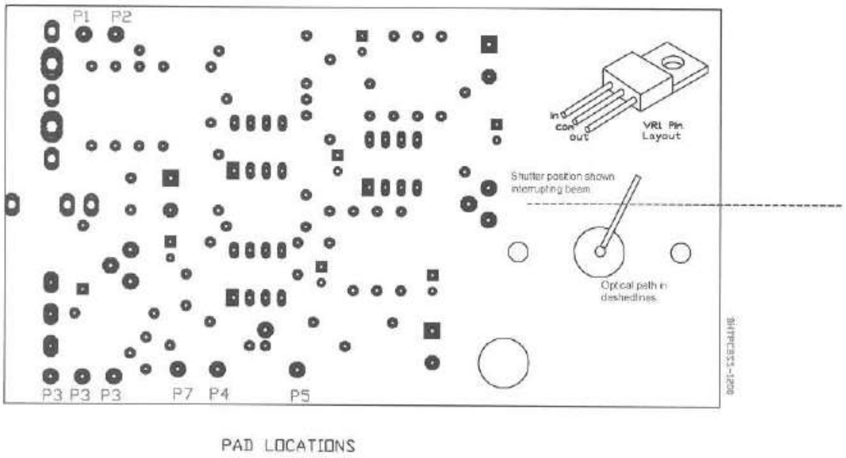
P1 P2
in
com
out
VR1 Pin Layout
Shutter position shown interrupting beam
Optical path in dashedlines
P3 P3 P3
P7 P4
P5
PAD LOCATIONS

Figure 37-4 *Pad and connections*

9. Insert J1 and note the rear connection lead requiring a short piece of bare wire to connect to the board pad directly beneath.

10. Insert the SOCK1 pyrodetector socket as shown. Carefully form the three leads of the pyrodetector at 90°. Do not sharply bend it. Note the proper leads, and position the tab as shown. This part must be at the optical center of the system for proper operation.

11. Connect the BZ1 buzzer using its leads, as shown in Figure 37-5.

12. Connect output control jack J2 and indicator light-emitting diode (LED1) using a 3-inch length of #22 vinyl wire. Note the cathode lead of LED1 directly connecting to the ground lug of J2. Note these parts eventually will be mounted to the panel.

13. Connect the leads from the chopper motor MO1 and battery clip CL1 as shown.

14. Create the panel, as shown in Figure 37-6. Attach it to the PC board via the two pots and the jack. The panel and PC board should be flush to one another. Attach the remaining J2 and LED1 to the panel. Note to use a bushing for the LED.

15. Create the shutter flap (FLAP), as directed in Figure 37-6.

16. Recheck all wiring for shorts, accuracy, and correctness of parts. Also check for solder joints. You are now ready to bench test the assembly.

17. Connect a 9-volt supply to the battery clip.

18. Rotate R12 to on, and note the shutter flap rotating with R19 full counterclockwise. It should slow down as R12 is turned in a clockwise direction. This rotation should be variable in the range of 10 *revolutions per second* (rps), down to stalling at less than 1 rps. The motor will stall if there is any obstruction.

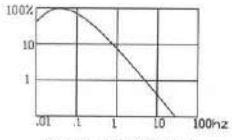

Chopping Freq vs Sensitivity

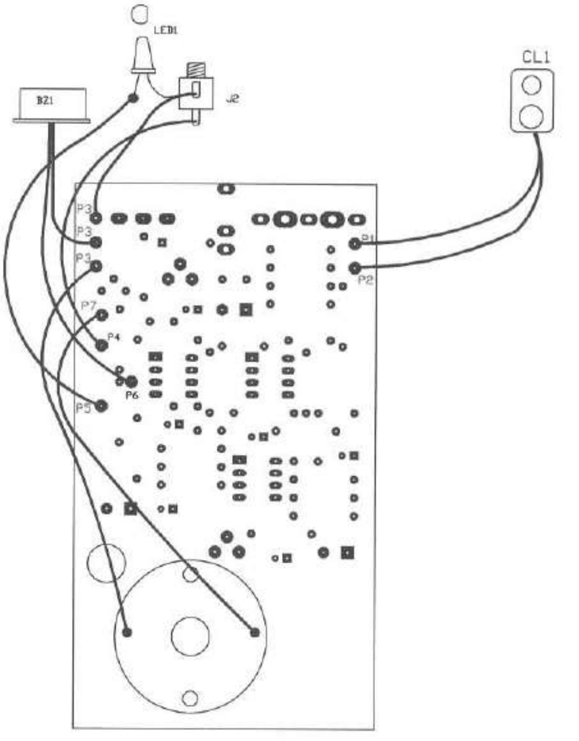

Figure 37-5 *Final wiring to external components*

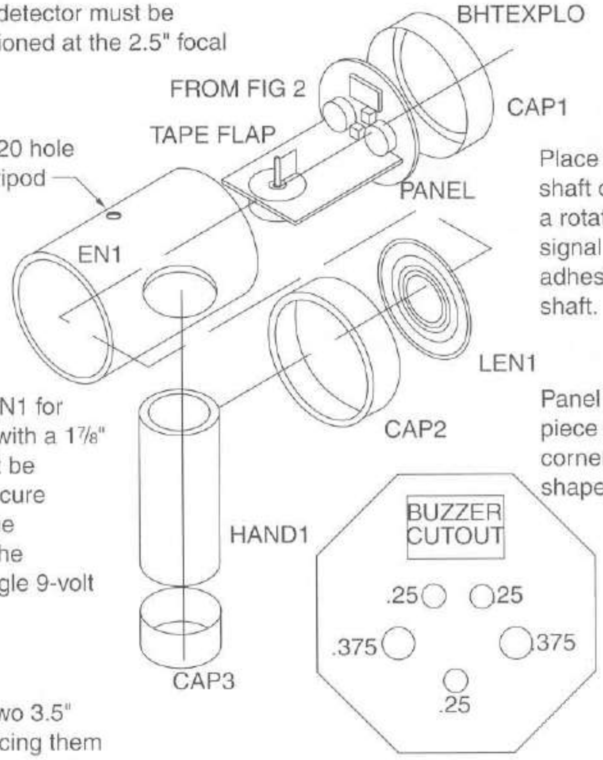

Note that the pyro detector must be centered and positioned at the 2.5" focal length of the lens.

Drill and tap a 1/4-20 hole for mounting to a tripod screw.

Place a small "flap" on to the shaft of the chopper motor to act as a rotating shutter for breaking up the signal path. Use a small piece of adhesive metal tape folded onto the shaft.

Note that hole in EN1 for handle is best cut with a 1⅞" circle saw. Fit must be tight to properly secure handle in place. The handle serves as the housing for the single 9-volt or 8 AA cells.

Panel is cut from a 3 1/4 x 3 1/4" piece of .035 aluminum. Cut corners to approach a circular shape.

Cut out center of two 3.5" plastic caps by placing them on the enclosure tube and cutting out the center section with an xacto knife using the inner wall as a guide.

Front Panel Fab

Note that holes must be accurately positioned for proper alignment to R12, R19, and J1 on the printed circuit board.

Figure 37-6 *Final housing setup*

Make sure the flap is clear and not touching the bearing bottom.

19. Connect a scope to J1 and note a changing level as your hand or another warm object is waved in view of the detector. Note the signal responding to a stationary source with the shutter flap or to some form of motion with the shutter flap stopped. Check the sensitivity control pot R12 and then verify a DC output level at J2, occurring when a body is detected. This function is for control of external alarms and alerts.
20. Check that the buzzer and light emitter activate when detecting a thermal source. Use your solder iron as a heat source and experiment with different sources.
21. Finally, assemble the device as shown in Figure 37-7 and proceed to use it as outlined in the instruction section.

Operating Instructions

1. Turn all controls fully counterclockwise to off.
2. Remove the battery compartment cover and slide out the eight-cell holder. Insert fresh batteries and reinstall.
3. Rotate the chopper control fully clockwise. View the shutter through the access hole and note how it is positioned in the figure.
4. Slowly turn the sensitivity control clockwise until the alarm sounds. If the unit does not respond, try moving your hand in front of it, noting its response. Turn the control back, just to the point where the unit does not respond.
5. Turn the chopper control until the unit starts to respond. You are now measuring the thermal background as indicated by the periodic sounding of the alarm as the shutter interrupts the heat path.

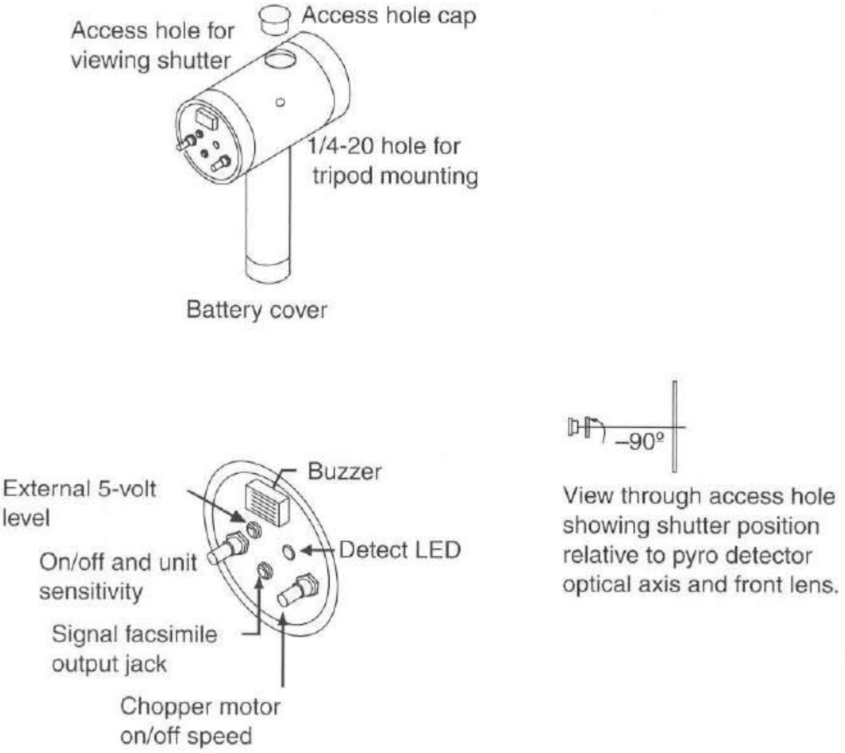
Access hole for viewing shutter
Access hole cap
1/4-20 hole for tripod mounting
Battery cover
–90°
External 5-volt level
Buzzer
On/off and unit sensitivity
Detect LED
Signal facsimile output jack
Chopper motor on/off speed

View through access hole showing shutter position relative to pyro detector optical axis and front lens.

Figure 37-7 *Final assembly view*

6. Connect a standard 3.5-millimeter plug into the 5-volt level jack and measure a positive 5-volt level every time the alarm responds. This output is through a 1K resistor and is intended for triggering other alarms or external functions.
7. Connect a standard 3.5-millimeter plug into the signal facsimile jack and connect it to a scope input set on the AC millivolt scale. Note the changing voltage level indicating the slightest amount of differential thermal energy. Check this level with and without the chopper and try it with various thermal sources, noting the voltage levels. This output port is intended for the lab technician desiring to explore low-level thermal sources. It requires a scope or a sensitive AC millivolt meter capable of measuring 10 to 100 millivolts.

The previous steps are performed without considering the thermal stabilizing of the unit or its immediate surroundings. Maximum sensitivity will not be obtainable under these conditions.

8. Select a target area and mount the unit on a tripod, using the 1/4-20 mounting hole located on the side of the unit. This scheme will place the handle in the horizontal plane (off to the side) but should not pose a problem. The access hole will be opposite the handle.
9. Point the unit at the target area and allow it several hours to thermally stabilize. Do not touch the enclosure and try not to get too close to the front of the system, as this can destabilize the system for low-level detection.
10. Point and pan the unit at some known source to verify the anticipated range of operation. Experiment with the chopper speed or no shutter to obtain the optimum results for your application.
11. Handheld operation will vary considerably due to the many variables in the thermal background. We can only suggest that you get accustomed by trial and error to determine your application.

Final Notes

The access hole for the chopper shutter is necessary in order to prevent the blockage of the thermal path when the chopper motor is off. There is no guarantee that the shutter will not randomly stop at the wrong place, blocking the detector and preventing normal operation. In such cases, the shutter must be manually moved with a pencil or some object.

The unit must be temperature stable to optimize maximum sensitivity. This usually takes about an hour if the unit has been recently handled or placed near radiant sources.

For actual body detection, the best performance will occur on a cold night. This does not mean that the unit will not work on a warm night, but that it might be less sensitive.

The detection of stationary objects may vary with chopper motor speed and background temperature. The maximum response of the pyrodetector is around .1 chops per second. This is very slow for many applications and we suggest increasing it to 1 to 3 per second.

The field of view may be narrow due to the Fresnel lens used, but you may change to other suitable lenses of polyethylene. Also, a panning action allows zeroing in on certain target subjects.

Table 37-1 Body Heat Detector Parts List

Ref.#	*Description*	*DB Part #*
R1, 2, 5, 6, 7, 11, 14, 16, 17	Ten 39K, ¼-watt resistors (or-wh-or)	
R3	10K, ¼-watt resistor (br-blk-or)	
R4, 8	Two 1.8-megohm ¼-watt resistors (br-gray-gr)	
R9, 10, 13, 18	Four 1K, ¼-watt resistors (br-blk-red)	
R20	4.7K, ¼-watt resistor (yel-pur-red)	
R21, 22	Two 1-megohm, ¼-watt resistor (br-blk-gr)	
R12/S1	20K pot and 12 VDC switch	
R19	500 to 1-kilohm pot	
R23	470-ohm, ¼-watt resistor (yel-pur-br)	
C1	220-microfarad, 25-volt vertical electrolytic capacitor	
C2, 4	Two 100-microfarad, 25-volt vertical electrolytic capacitors	
C3	.1-microfarad, 50-volt plastic capacitor	
C5, 7, 10	Three 4.7-microfarad, 25-volt vertical electrolytic capacitors	
C6, 9, 13, 14	Four .01-microfarad, 50-volt plastic capacitors (103)	
C8, 11	Two 1,000-microfarad, 25-volt vertical electrolytic capacitors	
C15	1-microfarad, 25-volt vertical electrolytic capacitor	
PC1	Printed circuit board (PCB)	DB# PCBHT
SOCK1	Three-pin TO5 socket	
Q1	PN2222 NPN transistor	
U1	LM358 dual-operational-amplifier integrated circuit	
U2	LM393 dual-comparator integrated circuit	
U3	LM555 timer integrated circuit	
VR1	7805 5-volt regulator TO220 package	
MO1	Low-voltage chopper motor	DB# CHOPMOT
PYR1	Pyroelectric detector	DB# PYR1
BZ1	Piezoelectric buzzer	
LED1	High-brightness light emitter	
J1, 2	Two 3.5-millimeter panel-mount jacks	
CL1	Battery clip and 9-inch leads	
PANEL	3.25 × 3.25-inch aluminum as shown	
EN1	6 × 3.5-inch OD schedule 40 PVC tubing as shown	
HANDLE	6 × 1.875-inch schedule 40 PVC tubing as shown	
CAP1, 2	Two 3 ½-inch plastic caps, as shown in Figure 37-6	
CAP3	2-inch plastic cap for retaining batteries in handle section	
LEN1	3-inch Fresnel polyethylene lens	DB# FRELEN3
BU1	Lens holder for LED1 or use a ⅜-inch bushing	
BHOLD1	8-AA-cell, 4 × 2 battery holder with snaps	
WR20B	24 inches #22 black vinyl hookup wire	

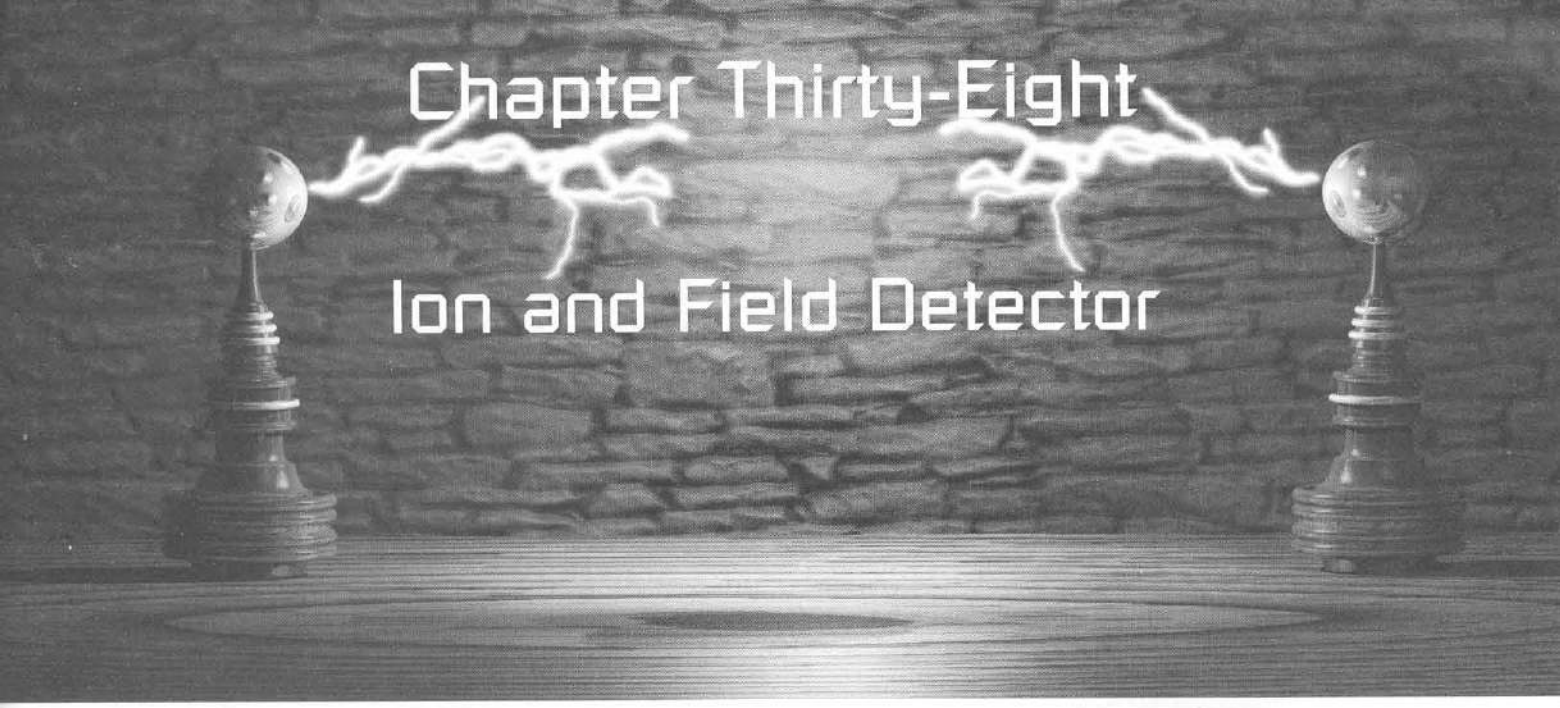

Chapter Thirty-Eight

Ion and Field Detector

This chapter focuses on a device capable of detecting and indicating the relative amount of free ions in the air (see Figure 38-1). It can determine the output of ion generators, high-voltage leakage points, static electrical conditions electric field gradients, and any other application where the presence of ions or a measurement of their relative flux density is required. It can also be used as a very sensitive lightning indicator, showing dangerous electric field levels.

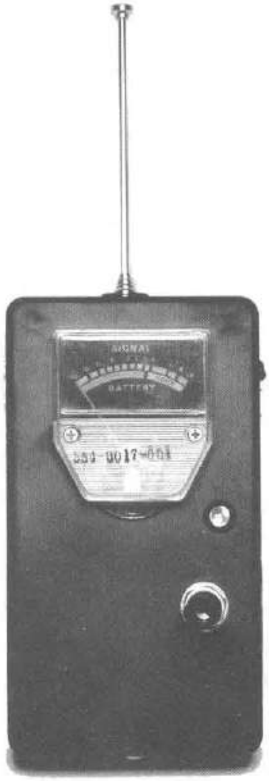

Figure 38-1 *Ion and field detector*

The unit is handheld and enclosed in a plastic enclosure about the size of a king-size pack of cigarettes. A sensitivity control with an on/off switch, a high flux indicator lamp, and a meter are located on the front panel. A telescoping antenna exits a hole in the top of the unit and is used as a moving probe or a stationary pickup point for these charge fields. Grounding the unit is a strip of metallic foil on the outside of the enclosure that contacts the user's hand or may be directly connected to earth ground via a wire. Please note that the performance of this unit can be seriously affected by high humidity conditions.

Circuit Description

As shown in Figure 38-2, ions are accumulated on the collector probe ANT1 and cause a minute, negative base current to flow in transistor Q1 connected as a PNP Darlington pair with transistor Q2. Capacitor C1 and resistor R1 form a time constant to eliminate any rapid fluctuations. The response of the time constant can be changed for various applications and is discussed in Figure 38-2. Diodes D1 and D2 prevent excessive voltage from destroying Q1 by clamping any transients to ground. The collector of Q2 is DC

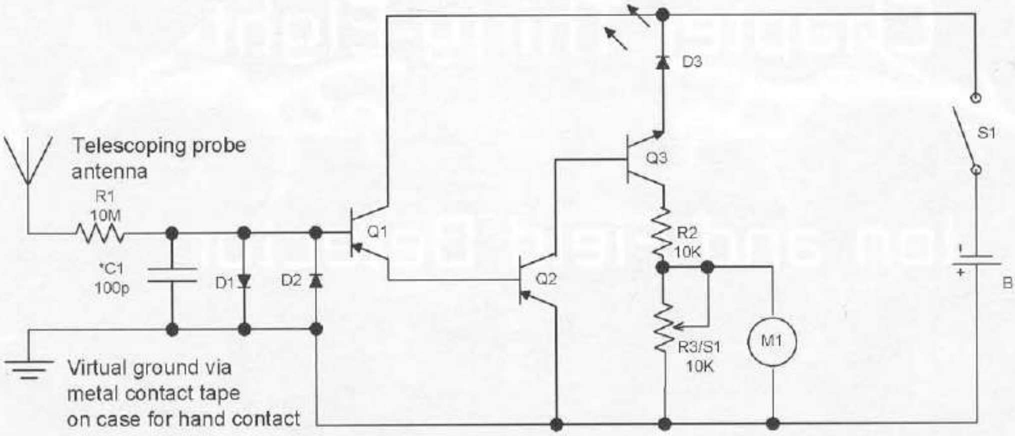
Telescoping probe antenna
R1
10M
*C1
100p
D1
D2
Q1
Q2
Q3
D3
S1
B1
R2
10K
R3/S1
10K
M1
Virtual ground via metal contact tape on case for hand contact

C1 is used to slow down the response and can be eliminated when detecting fast discharge fields such as lightning etc

Unit is shown wired for positive ion detection. You may change to negative by replacing the PN2222 transistors with PN2907 and the opposite. For a quick indication you may grab the antenna and use the body of the unit to detect the negative ions

Please note that the antenna probe on you unit is a telescoping antenna properly secured and electrically isolated. It is important to remember that any type of leakage around the input of QI can reduce the sensitivity. You may wish to coat the circuitry with a good quality varnish etc. Make sure unit is dry and clean before sealing.

Figure 38-2 *Circuit schematic*

coupled to the base of NPN transistor Q3 in series with current-limiting resistor R2 and meter sensitivity pot R3. Current flowing through Q3 now causes a voltage to develop across R3, driving the indicator meter M1. *Light-emitting diode* (LED) D3 is connected in series with the emitter of Q3 and serves as a visual indicator of strong charge fields. It should be noted that the meter serves only as a relative indicator of ion flux and is not totally that linear.

A threshold voltage, as a result of the base emitter, drops off Q1 and Q2 is only noticeable at almost non-detectable levels. It should be noted that in order for the unit to operate properly, some sort of ground is usually required. Metallic tape is used as shown and provides contact to the operator's hand, providing a partial ground. If the unit is placed remotely, it should be earth grounded to a water pipe, for example.

Measuring positive ions simply involves holding the antenna, using the body of the unit as the probe, or reversing the polarity of the transistors. A *double pole, double throw* (DPDT) switch may be optionally installed for polarity reversal, allowing reversal at the flick of a switch. Be aware that a ceramic or quality switch must be used to avoid leakage in high humidity conditions.

Construction Steps

1. Lay out and identify all the parts and pieces. Verify them with the parts list, and separate the resistors as they have a color code to determine their value. Colors are noted in the parts list at the end of the chapter.
2. Cut a piece of .1-inch grid perforated board to 2 × 1 inches. Locate and drill the holes as shown in Figure 38-3. If you are building from a perforated board, it is suggested that you insert the components starting in the lower left-hand corner as shown. Pay attention to the polarity of all the semiconductors. Route

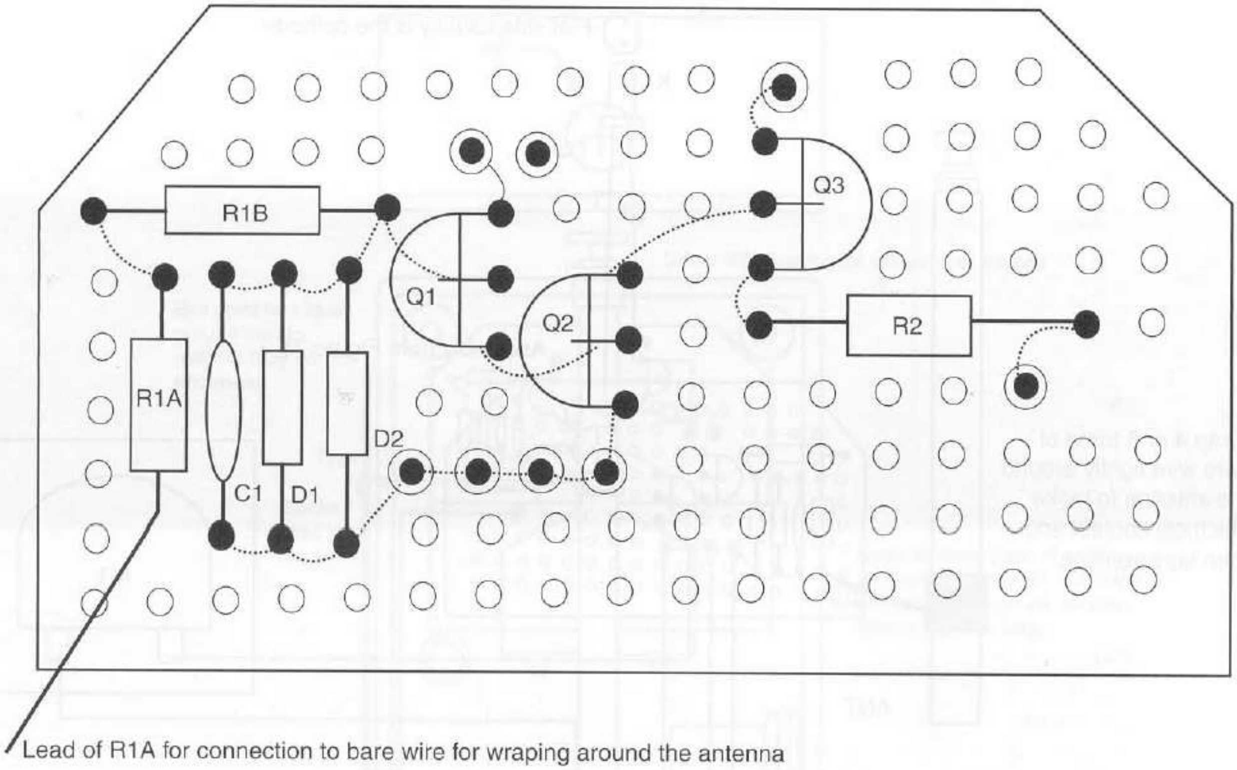

Lead of R1A for connection to bare wire for wraping around the antenna

● This symbol indicates holes used for leads of components.

◉ This symbol indicates holes used for connection points of external leads.

Dashed line indicates wiring performed on underside of board. Use the leads of the components whenever possible for these wire runs.

Board size is 1.9 x 1 inch and is .1 x .1 centers.

Figure 38-3 *Assembly board wiring*

the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of the assembly board.

3. Attach the external leads and components as shown in Figure 38-4.
4. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
5. Create the plastic case, as shown in Figure 38-5. Note the 1-inch meter hole centered and located approximately $1^1/_2$ inch from the top as shown. Holes for the switches are centered and located 1 inch below the meter line. Location is not critical, but make sure everything is clear and fits together.
6. Place a strip of foil tape as shown, noting a hole for the connection to the common circuit

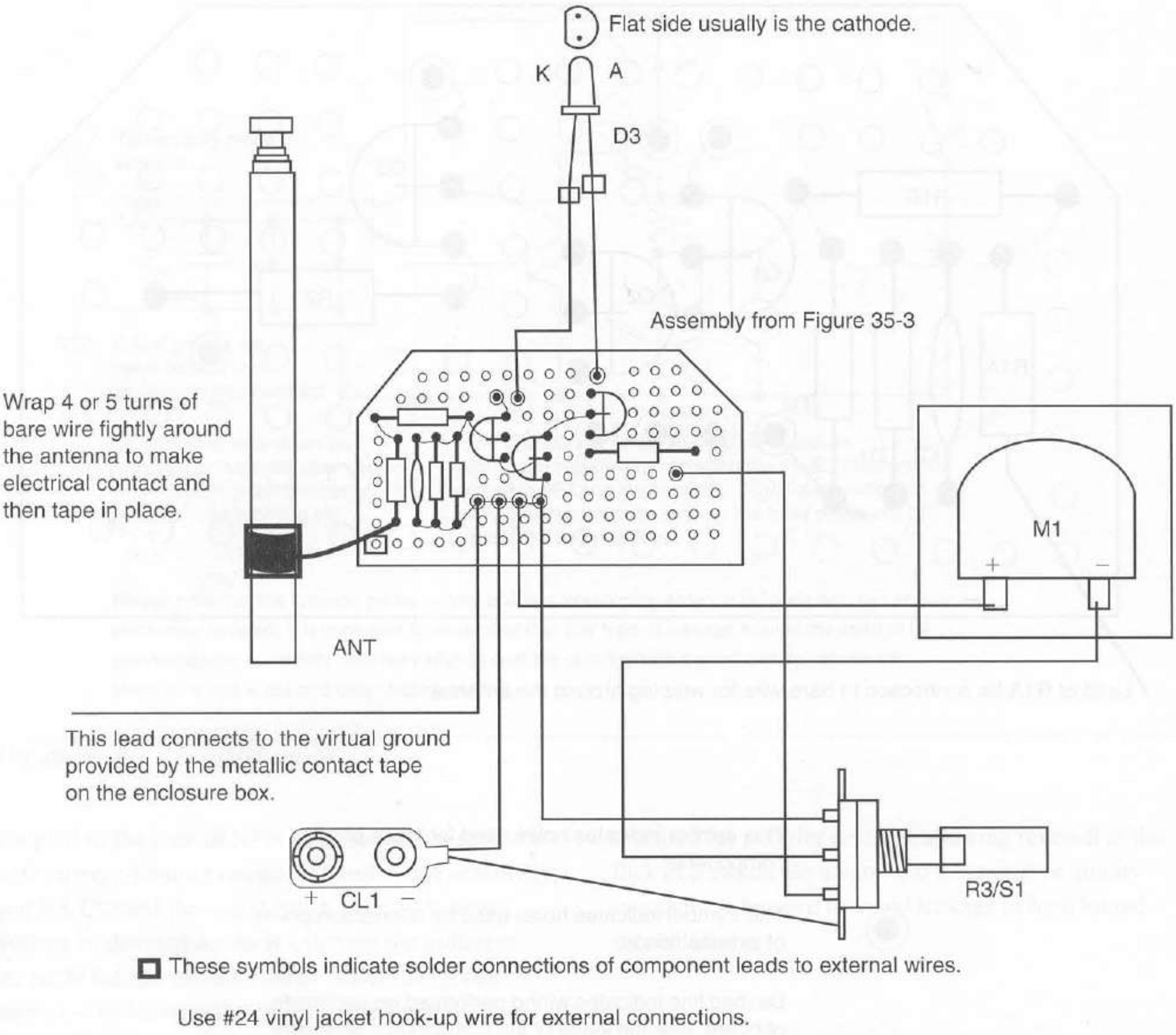
Flat side usually is the cathode.
K
A
D3
Assembly from Figure 35-3
Wrap 4 or 5 turns of bare wire tightly around the antenna to make electrical contact and then tape in place.
M1
+
−
ANT
This lead connects to the virtual ground provided by the metallic contact tape on the enclosure box.
+ CL1 −
R3/S1
These symbols indicate solder connections of component leads to external wires.
Use #24 vinyl jacket hook-up wire for external connections.

Figure 38-4 *Final wiring and assembly*

point. Secure the assembly board with double-sided tape after circuit operation has been verified.

7. Install battery B1 as shown and turn on the control. Rotate R3 fully clockwise and note the meter showing a slight indication. This is due to transistor and circuit leakage and should not be considered an indication of the ions. If meter sensitivity responds in the opposite way, you may have to reverse the wires across the outer terminals of R3 or the meter leads.

8. Extend the antenna about 6 inches and adjust R3 fully clockwise. Obtain a plastic comb and run it through your hair or rub a plastic object with a dry cloth. Place it near the probe and note the meter indicating a strong charge and the diode lighting. This effect may be reduced if the air is damp or moist.

9. Make the final assembly, as shown in Figure 38-6.

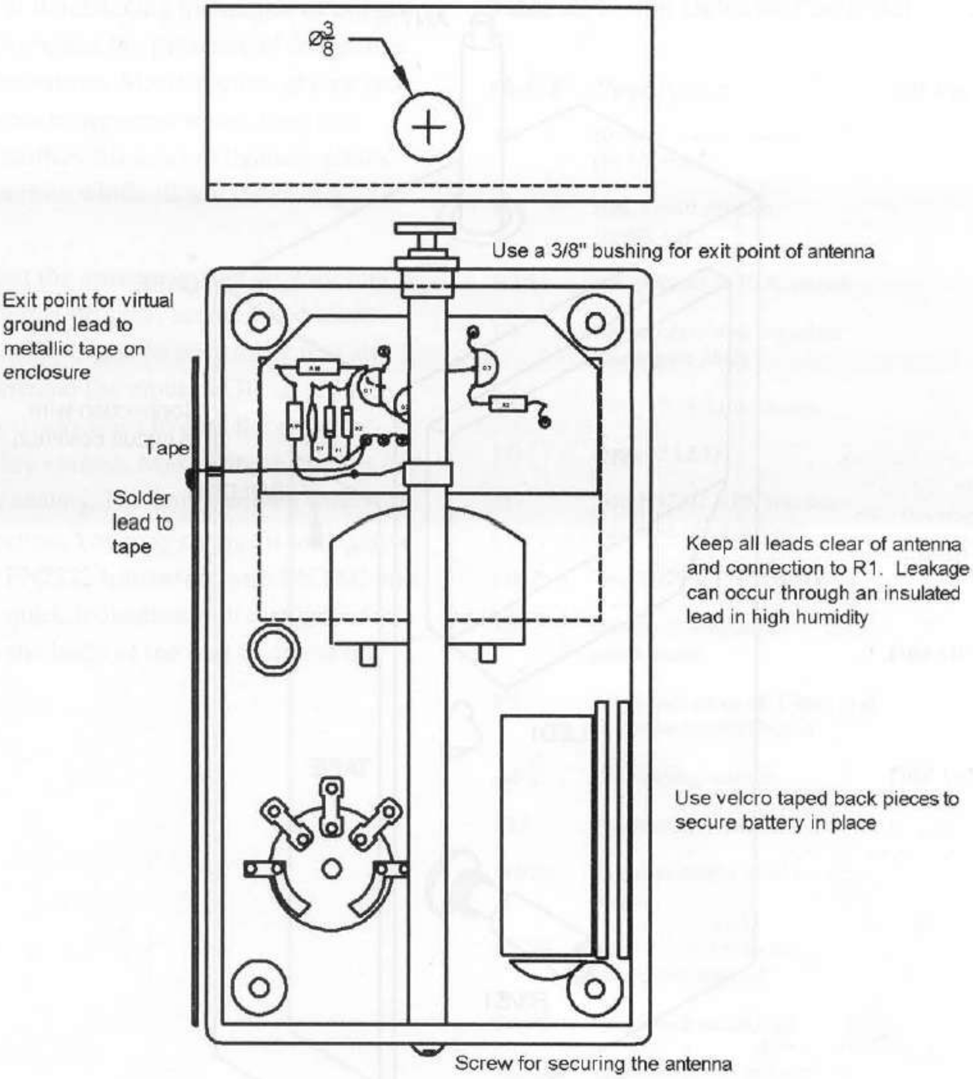

Use a 3/8" bushing for exit point of antenna

Exit point for virtual ground lead to metallic tape on enclosure

Keep all leads clear of antenna and connection to R1. Leakage can occur through an insulated lead in high humidity

Use velcro taped back pieces to secure battery in place

Screw for securing the antenna

Rear view with cover removed. Wiring as not shown for clarity. Use figure 35-4 for wiring and this figure for component location. Keep all lead neat and short.

Figure 38-5 *Final assembly*

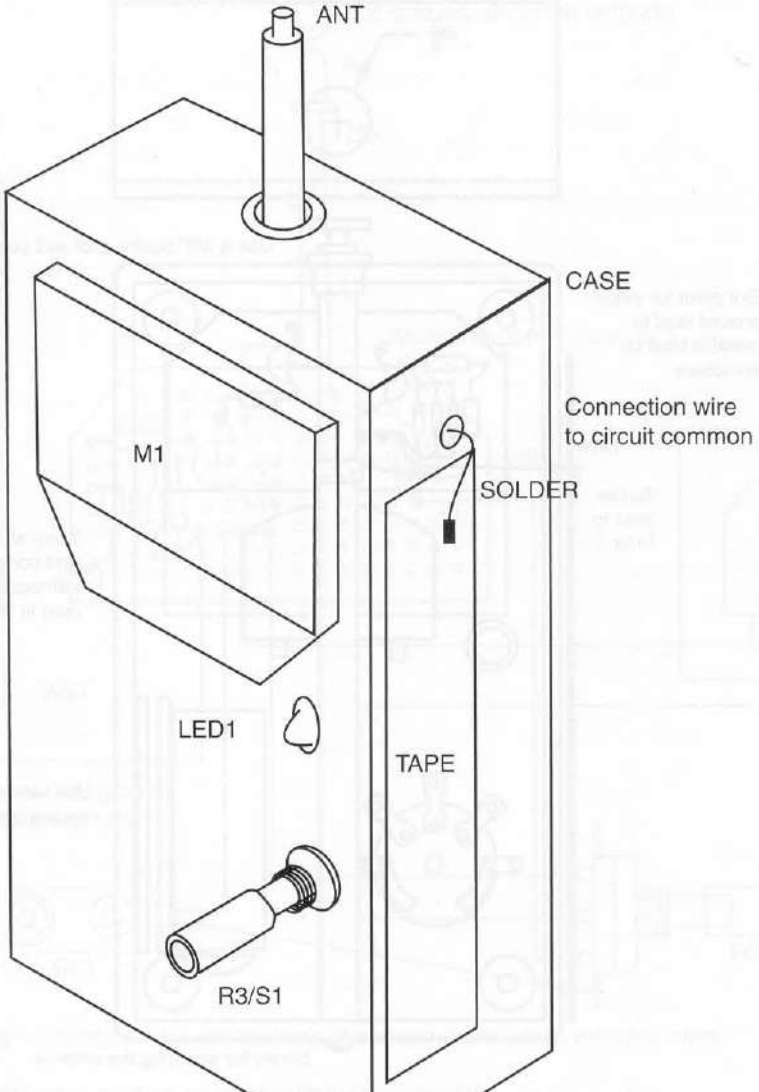

Figure 38-6 *Completed unit*

Applications

Your ion detector is a very, very sensitive ion-measuring device. It can be used for relative measurement but is not designed for absolute measurement.

For quick indications of the presence of a negative ion field, the unit may be handheld and actually used to determine where the source is. The sensitivity of this device can be realized by the simple experiment of running a plastic comb through one's hair and laying it near the probe, as described above.

Those who are familiar with the metallic leaf electroscope will soon realize the advantage of portability and sensitivity. When used for indicating or testing the relative strength of ion sources, the unit should be hard wire grounded for the best results. Adjustments to an ion source to determine the ion output may be made, noting the meter reading and then readjusting R3 to bring the meter reading to scale.

A very interesting phenomenon will be noted when using this device for detecting residual ion fields, the shielding of ions, field direction, static charges, resultant polarity, and the intensity of static charges, as well as a host of others. The unit is an invaluable tool for determining the output of ion generators, any purifiers, and the presence of dangerous static electricity situations. Many sources of charged particles soon become apparent when using the device. People's clothes, fluorescent lighting, plastic containers, and certain winds all will indicate a charge.

Please note that the antenna probe on your unit is a telescoping antenna properly secured and electrically isolated. It is important to remember that any type of leakage around the input of Q1 can reduce the sensitivity. You may want to coat the circuitry with a good-quality varnish. Make sure the unit is dry and clean before sealing. The unit is shown wired for positive ion detection. You may change it to negative by replacing the PN2222 transistors with PN2907 and vice versa. For a quick indication, you may grab the antenna and use the body of the unit to detect the negative ions.

Table 38-1 Ion Detector Parts List

Ref. #	*Description*	*DB Part #*
R1	10 meg, ¼-watt resistor (br-blk-blue)	
R2	10K, 1-watt resistor (br-blk-or)	
R3/S1	10K pot and 12 VDC switch	
C1	100-picofard disk capacitor (see Figure 38-2)	
D1, 2	Two 1N914 signal diodes	
D3	Colored LED	
O3	Two PN2222 NPN transistor (see note on reversing)	
Q1, 2	Two PN2907 PNP transistors	
M1	100-micro-amp, small panel meter	DB# METER50S
PB1	1 × 2-inch piece of .1-inch grid perforated circuit board	
ANT	Telescoping antenna	DB# ANT1
CL1	9-volt battery clip	
WR24	24 inches of #24 vinyl hookup wire	
CASE	4⅜ × 2⁷⁄₁₆ × 1¼-inch plastic box with lid	
TAPE	4 × ½-inch metal tape	
WR24B	12 inches of #24 bus wire	

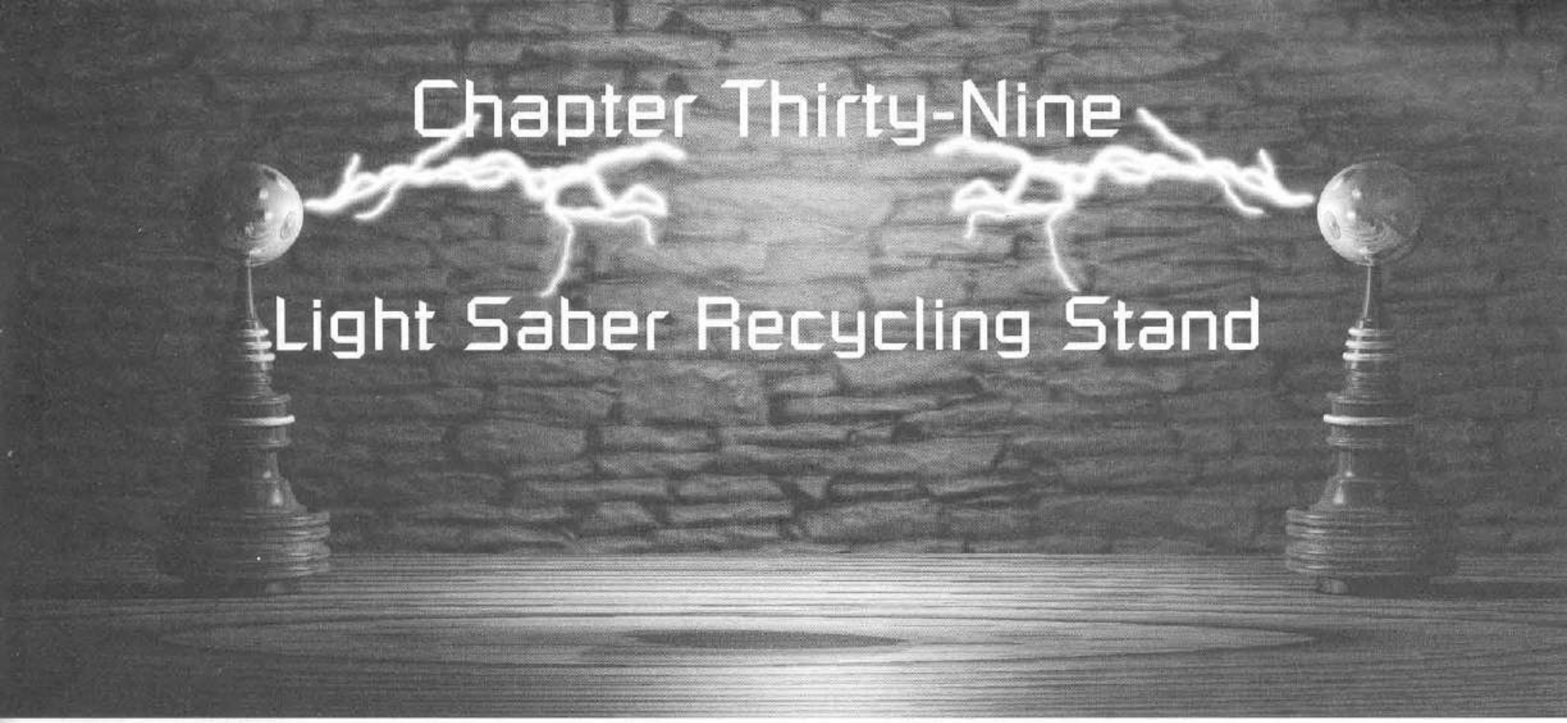

Chapter Thirty-Nine

Light Saber Recycling Stand

This project is intended for those who have built or purchased a light saber from Information Unlimited or who have built one from our first book as directed in Chapter 21 in this series of three.

The saber stand recycles your saber where the plasma display travels up the tube, appearing to evaporate into open space (see Figure 39-1). It then retracts into the hilt and continues repeating until power is removed. The effect is very relaxing and provides an excellent display piece or prop for a bar, a point-of-purchase display, an attention getter, a conversation piece, or just a nightlight for a bedroom.

Operation is powered from a 12-volt wall adapter and is completely safe for children. This is an intermediate-level project requiring basic electronic skills. Expect to spend $20 to $30. All the parts are readily available, with specialized parts obtainable through Information Unlimited (www.amazing1.com), and are listed in the parts list at the end of the chapter.

Circuit Description

Figure 39-2 shows a 12-volt wall adapter (T1) being fed into a voltage regulator chip (I2) where a regulated 9 volts is applied to the remaining circuitry. Capacitors C1 and C2 smooth out the voltage waveform. An astable pulse generator using a timer (I1) provides an uneven "on to off" time determined by the ratio of resistors R2 and R3. The output pulse repetition rate is a function of timing capacitor C3 with

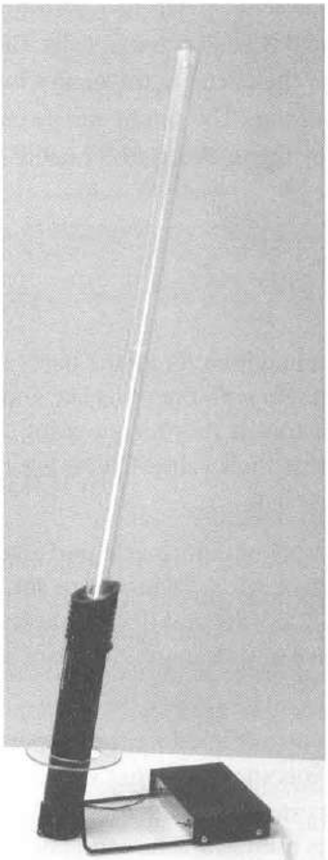

Figure 39-1 *The saber recycling stand*

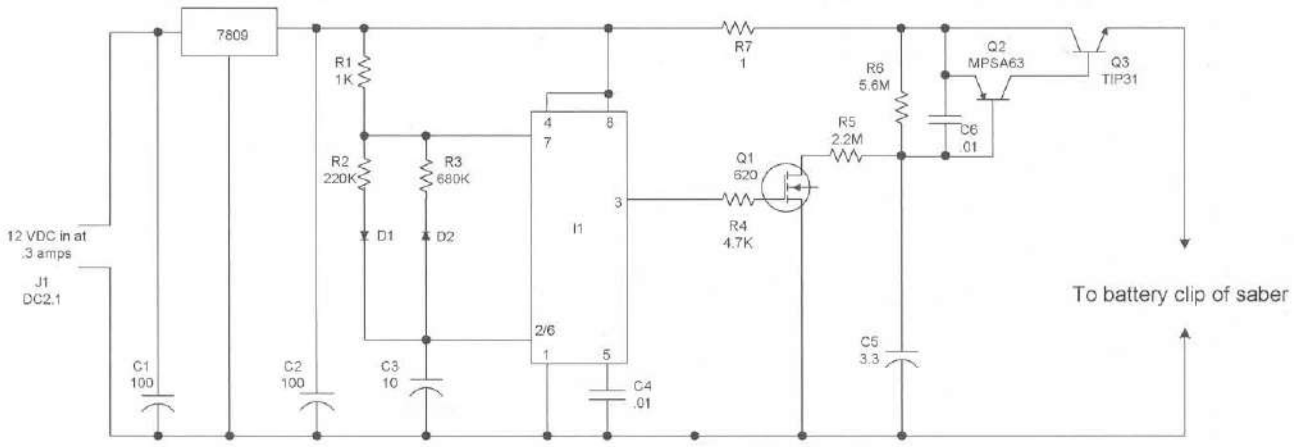

Figure 39-2 *Circuit schematic*

output at pin 3 of I1. The pulse output serves as a gate for MOSFET switch Q1, creating a charging and discharging of low-leakage capacitor C5. A quasi-linear voltage ramp now appears across C5, controlling the current through a high beta Darlington transistor (Q2) in turn controlling the main-current pass transistor Q3. It is this ramping up and down of current that now powers the saber electronics, allowing the plasma to flow up and down, all automatically. The cycling action is set by the circuit to travel up a half-second and down 3 seconds. This timing can be changed to some extent by the selection of R2 and R3.

Assembly

1. Lay out and identify all the parts and pieces. Verify them with the parts list, and separate the resistors as they have a color code to determine their value. Colors are noted on the parts list.
2. Cut a piece of .1-inch grid perforated board to a size of 3.1 × 1.8 inches, as shown in Figure 39-3. Locate and drill the holes as shown to relieve any strain on the input and output leads.
3. When you are building from a perforated board, it is suggested that you insert the components starting in the lower left-hand corner, which is used as a reference, and proceed to the opposite corner.

   Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors.

   Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of the assembly board.
4. Attach the external leads for battery clip CL1 and the leads to input power jack J1. *Important!* The battery clip CL1 must be wired in reverse by connecting the red lead to the circuit ground and the black lead to the emitter of Q3.
5. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.

Parts Fabrication

1. Create the chassis shown in Figure 39-4. This is attached to the BASE section using four #6 sheet metal screws.
2. Create the base section from a piece of plastic or aluminum, as shown in Figure 39-5. Note

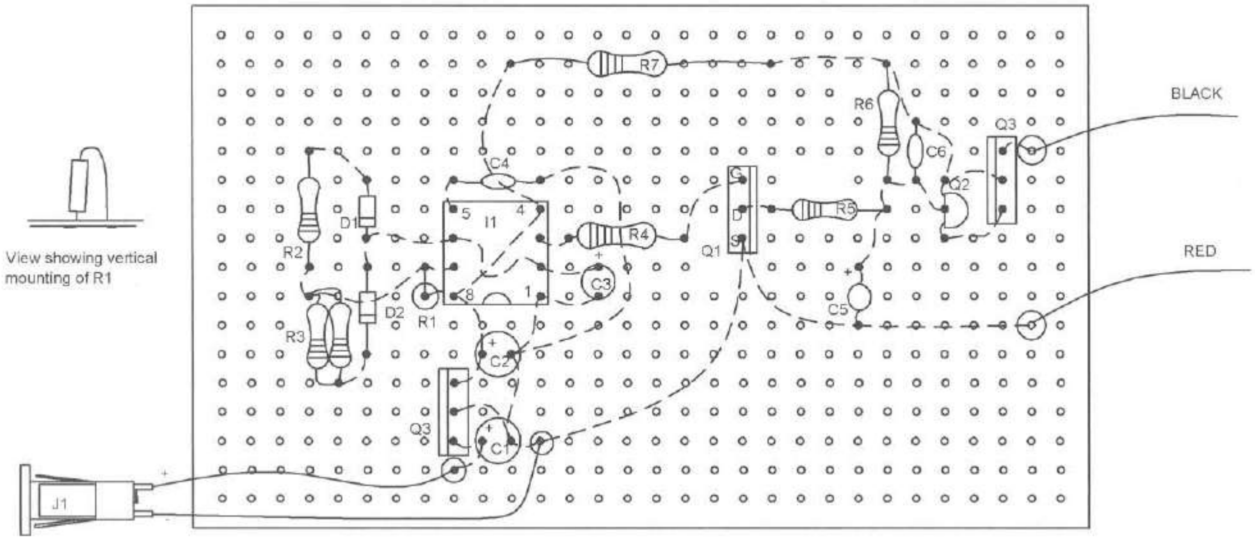

Notes

The looped dashed lind connecting pins 2 and 6 of I1 must have a sleeve of insulated tubing to prevent shorting the lead between pins 4 and 8.

〜 Dashed lines are connections on underside of perforated circuit board using the component leads. These points can also be used to determine foil runs for those who wish to fabricate a printed circuit board.

• Small dots are holes used for component insertion.

● Large dots are solder junctions

⊙ Circles with dots are holes that may require drilling for component mounting and points for external connecting wires indicating strain relief points. It is suggested to drill clearance holes for the leads and solder to points beneath board.

Figure 39-3 *Board assembly*

the four holes for mounting the chassis section, two holes for the holder section, and two holes for the tie wraps.

3. Create the holder from a piece of $2^{3}/_{8}$-inch, schedule 40 gray PVC tubing. Note the slight angle cut for aesthetics being congruent with the light saber when in the stand. This piece is mounted to the base using two #6 × $^{3}/_{8}$-inch sheet metal screws.

4. Form the frame as shown in Figure 39-6 from $^{3}/_{16}$-inch cold roll steel rod or an equivalent. The frame must be formed to provide a stable assembly with the saber in place in the stand. The longer 36-inch blade unit is the most critical, requiring the most attention. Attempt to follow the dimensions as shown and determine the base angle as a final adjustment with the saber in place.

   The frame is secured in place by inserting the horizontal members through the corner holes in the chassis section. The front section of the frame is secured in place by two small tie wraps, as shown in Figure 39-5.

10. Finish the assembly, as shown in Figures 39-7 and 39-8.

Operation

Remove the cap at the rear of the saber hilt and remove both batteries. Take out one of the battery clips and snap it into the mating clip of the saber stand. Position the saber in the stand and verify its mechanical stability. Plug in the wall adapter and turn on the switch. The plasma should immediately go to the end and then drop down into the hilt. Recycling should commence and continue until power is removed.

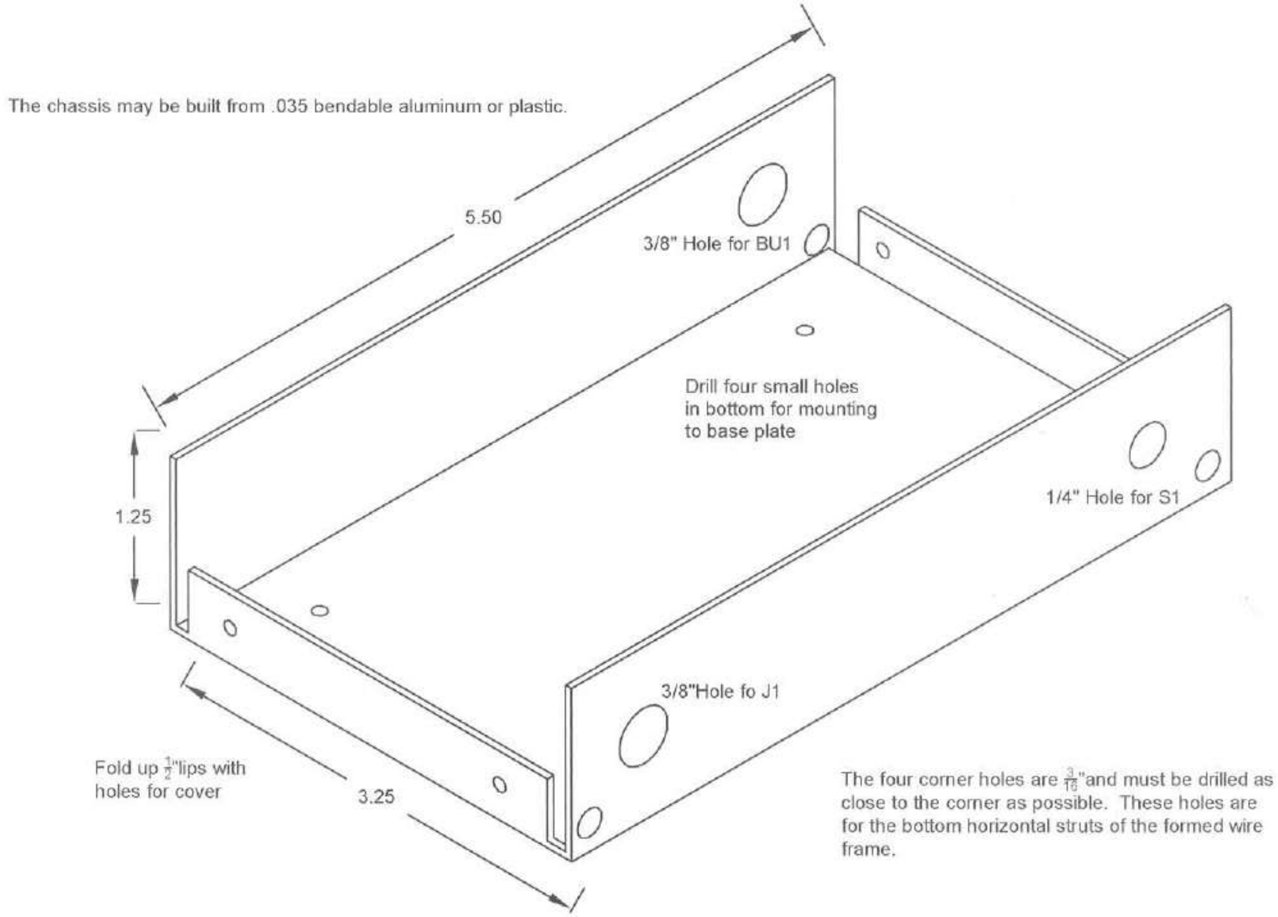
The chassis may be built from .035 bendable aluminum or plastic.
5.50
3/8" Hole for BU1
Drill four small holes
in bottom for mounting
to base plate
1.25
1/4" Hole for S1
3/8"Hole fo J1
Fold up $\frac{1}{2}$"lips with
holes for cover
3.25
The four corner holes are $\frac{3}{16}$"and must be drilled as
close to the corner as possible. These holes are
for the bottom horizontal struts of the formed wire
frame.

Figure 39-4 *Chassis fabrication*

Light sabers are available from Information Unlimited at www.amazing1.com in 24- or 36-inch lengths in colors of phasor green, photon blue, starfire purple, neon red, or solar yellow.

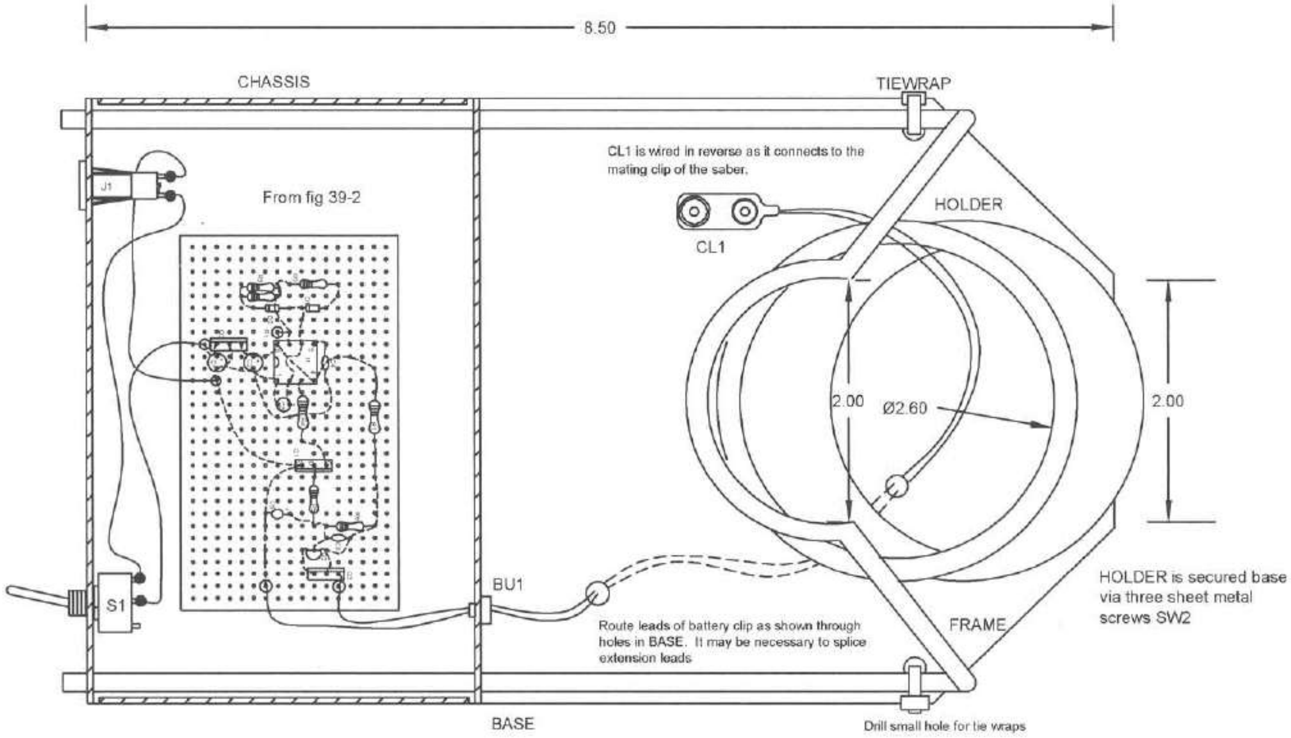
8.50
CHASSIS
TIEWRAP
J1
From fig 39-2
CL1
HOLDER
2.00
Ø2.60
2.00
S1
BU1
FRAME
BASE

CL1 is wired in reverse as it connects to the mating clip of the saber.

HOLDER is secured base via three sheet metal screws SW2

Route leads of battery clip as shown through holes in BASE. It may be necessary to splice extension leads

Drill small hole for tie wraps

Figure 39-5 *Base fabrication*

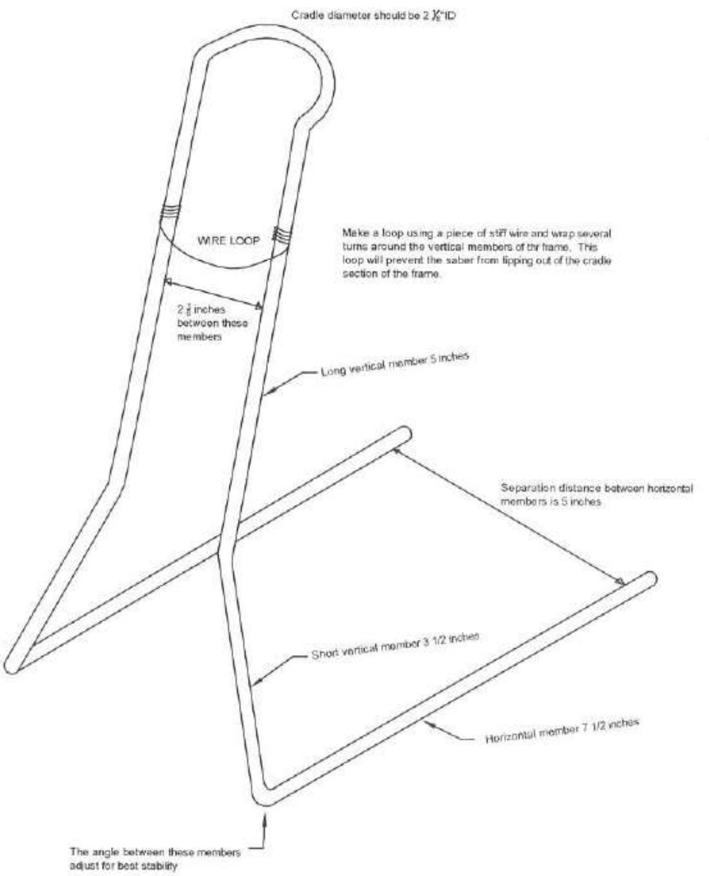
Fabricate from $\frac{3}{16}$" cold roll steel or brass
Cradle diameter should be 2 $\frac{1}{8}$" ID
WIRE LOOP
Make a loop using a piece of stiff wire and wrap several turns around the vertical members of thr frame. This loop will prevent the saber from tipping out of the cradle section of the frame.
2 $\frac{1}{8}$ inches between these members
Long vertical member 5 inches
Separation distance between horizontal members is 5 inches
Short vertical member 3 1/2 inches
Horizontal member 7 1/2 inches
The angle between these members adjust for best stability

Figure 39-6 *Frame section*

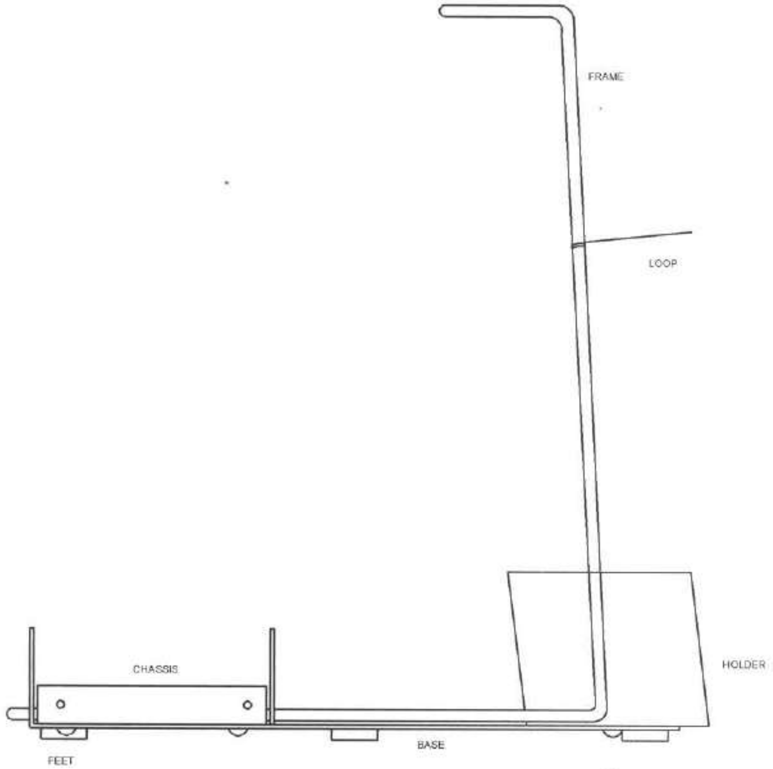
FRAME
LOOP
HOLDER
CHASSIS
BASE
FEET

Figure 39-7 *Side view of assembly*

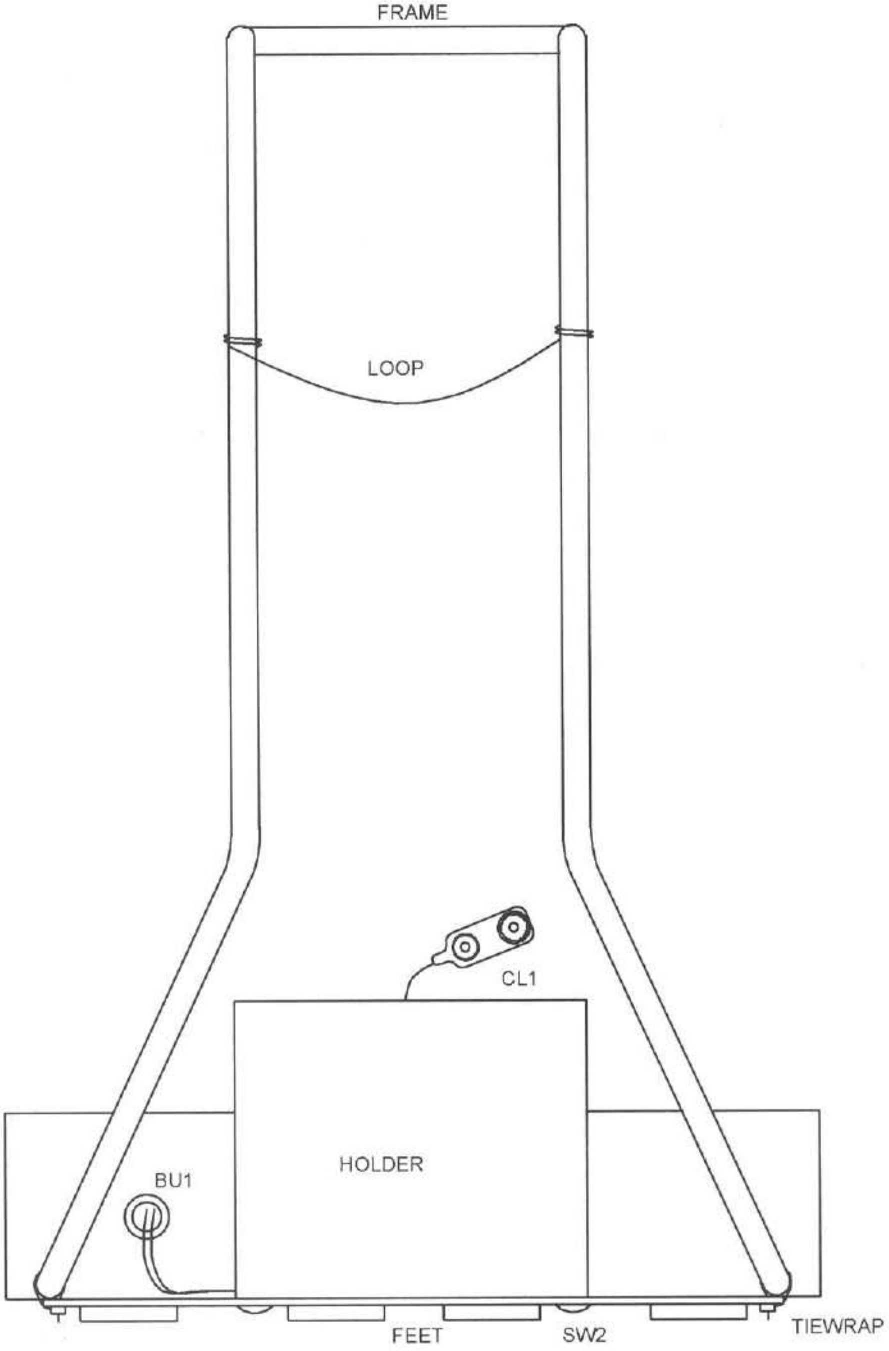
FRAME
LOOP
CL1
HOLDER
BU1
FEET
SW2
TIEWRAP

Figure 39-8 *Front view of assembly*

Table 39-1 Light Saber Stand Parts List

Ref. #	Description	DB Part #
R1	1K, ¼-watt resistor (br-blk-red)	
R2	220K, ¼-watt resistor (red-red-yel)	
R3	680K, ¼-watt resistor (blue-gray-yel)	
R4	4.7K, ¼-watt resistor (yel-pur-red)	
R5	2.2 m, ¼-watt resistor (red-red-gr)	
R6	5.6 m, ¼-watt resistor (gr-blue-gr)	
R7	1-ohm, ¼-watt resistor (br-blk-gold)	
C1, 2	Two 100-microfarad, 25-volt vertical electrolytic capacitors	
C3	10-microfarad, 25-volt vertical electrolytic capacitor	
C4, 6	Two .01-microfarad, 50-volt disc capacitors	
C5	3.3-microfarad, 35-volt tantalum capacitor	
D1, 2	Two IN914 silicon diodes	
I1	555 dual inline package (DIP) timer integrated circuit	
I2	7809 T0220 9-volt regulator	
Q1	IRF620 metal-oxide-semiconductor field effect transistor (MOSFET)	
Q2	MPSA63 PNP Darlington transistor	
Q3	TIP31 T0220 NPN transistor	
J1	Snap-in DC panel, 2.5 mm jack	
S1	Single pole, single throw (SPST) small toggle switch	
PB1	3.5 × 2 .1 × .1-inch grid perforated board; trim to size	
CL1	Battery clip with 12-inch leads or extend shorter leads	
TIEWRAP	Two small tie wraps	
CHASSIS	Create as shown in Figure 39-4 from .035-inch aluminum	
BASE	Create as shown in Figure 39-5 from .035-inch plastic	
HOLDER	Create from a length of 2⅜-inch schedule 40 gray PVC	
FRAME	Create as shown in Figure 39-6	
FEET	Six large, rubber stick-on feet	
SW2	Six #6 × ⅜-inch sheet metal screws	
BU1	Two ⅜-inch plastic bushings	
LIGHT SABER	36- or 24-inch saber in blue, green, red, or purple	

Chapter Forty

Andromeda Plasma Lamp

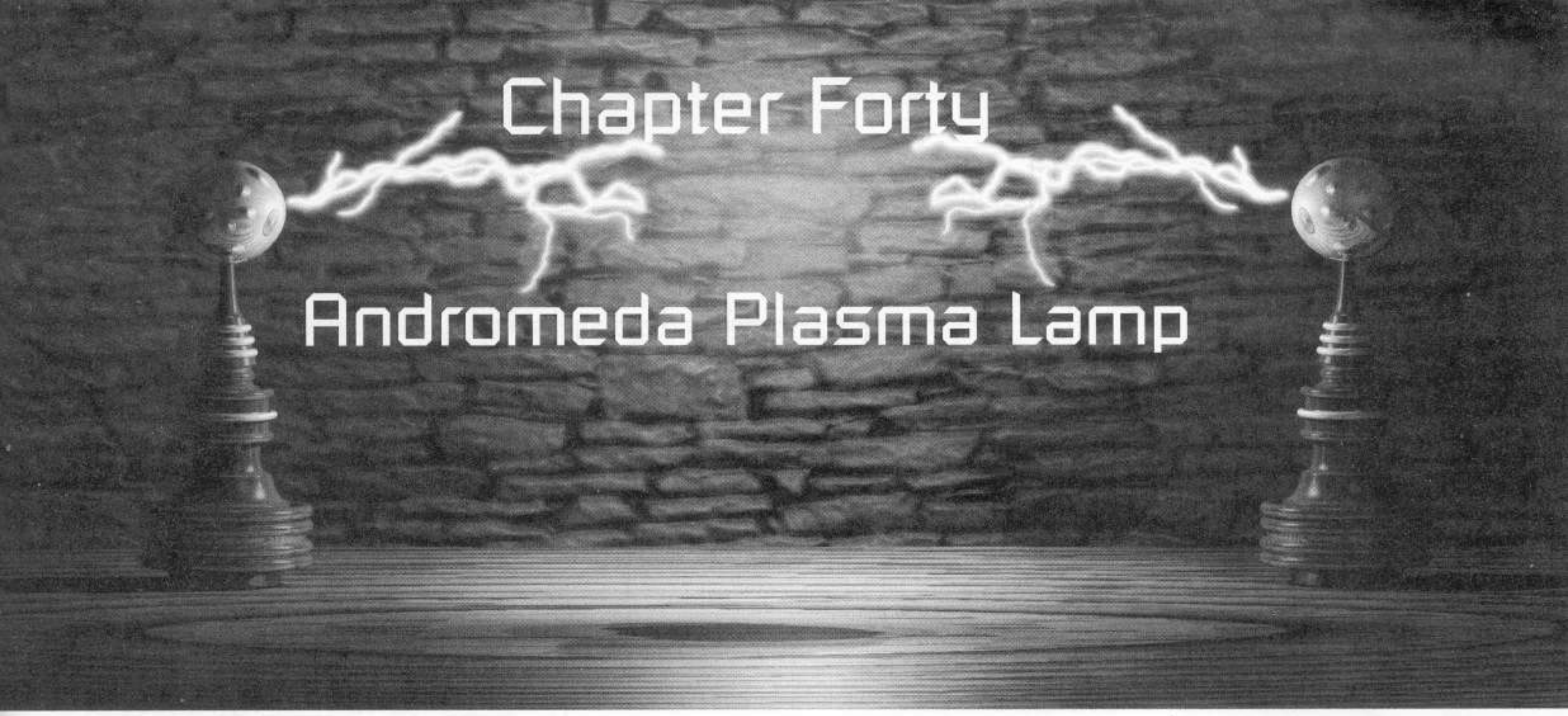

This project shows how to build a unique type of outside lamp suitable for illuminating driveways, pools, gardens, walkways, docks, or any place where colorful or functioning illumination is desired. The lamps are patented by the author and are currently in preproduction.

The lamp's features are high efficiency as five individual lamps can be powered by a single, conventional, 12-volt wall adapter as the lamps require only 3 watts each. The lamps provide a high amount of useful illumination and must not be compared to the low-light, solar-powered accent lamps.

Another feature of the lamp is that it never needs replacing and should last a lifetime. Other lamps such as incandescent ones are always burning out filaments, whereas conventional fluorescents darken and fail to start due to metallic sputtering.

The third feature is that the lamps are available in pure spectral colors of green, blue, purple, yellow, red, and cool and warm white. These are vivid colors and not the result of filters or paint. Their color is the result of the true spectral emission.

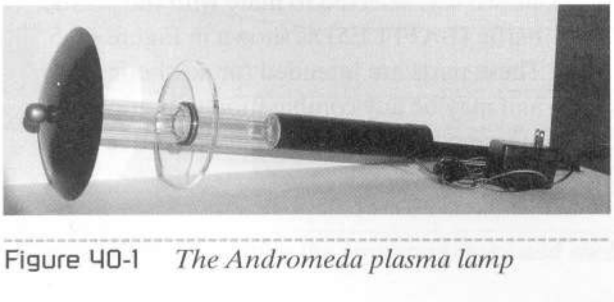

Figure 40-1 *The Andromeda plasma lamp*

This is an intermediate-level project requiring basic electronic skills. Expect to spend $25 to $35. All the parts are readily available, with specialized parts obtainable through Information Unlimited at www.amazing1.com, which are listed in the parts list at the end of the chapter.

Circuit Description

Transistor Q1 forms a self-oscillating circuit where current rises in the collector, COL, winding transformer T1 and inducing a voltage in the feedback (FB). This also causes a winding that keeps Q1 on until T1 saturates and the feedback voltage can no longer be induced. The collector current rapidly falls and forces the base to go negative further, turning Q1 off. When the collector current falls to zero, Q1 is turned on by bias-set resistors R1 and R2, which must charge up timing capacitor C2. The setting of trimpot R1 determines the power output by controlling the number of cycles per second, and capacitor C1 bypasses any residual signal voltage to ground (see Figure 40-2).

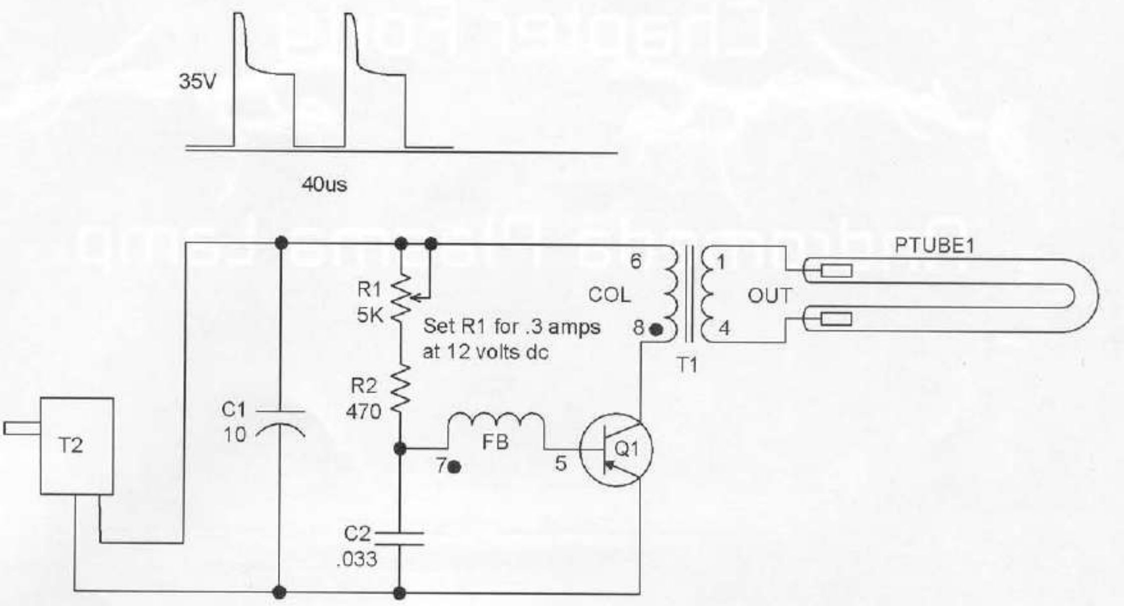

Figure 40-2 *Circuit schematic*

Assembly Board Wiring

1. Cut a piece of .1 × .1 grid vector circuit board to 1 × 4.5 inches, as shown in Figure 40-3.
2. Whenever you are using a perforated board, it is suggested that you insert components starting in the lower left-hand corner. Pay attention to the polarity of the capacitors with polarity signs and all the semiconductors. Route the leads of the components as shown and solder as you go, cutting away unused wires. Attempt to use certain leads as the wire runs or use pieces of the #24 bus wire. Follow the dashed lines on the assembly drawing as these indicate the connection runs on the underside of the assembly board.
3. Double-check the accuracy of the wiring and the quality of the solder joints. Avoid wire bridges, shorts, and close proximity to other circuit components. If a wire bridge is necessary, sleeve some insulation onto the lead to avoid any potential shorts.
4. Attach leads from the plasma tube PT1, being careful not to stress the glass enclosure. Note the small tie wraps (TY1) securing the tube to the perforated board. Again, do not cause any stress to the glass. Make sure that when tightening the wraps they do not stress the tube. When soldering, you must not allow the glass or metal seal to get too hot. Leave at least a half-inch lead between the tube and the connection. Use flux and make the solder joints quickly.
5. Connect the input power leads to a 12-volt DC source and note the lamp lighting. Quickly adjust trim pot R1 to 300 milliamps of current. The lamp should be very bright and the heatsink tab on Q1 should only be warm. Allow it to run for several hours and check adapter transformer T2 for excessive heating.
6. Create the EN1 enclosure tube and the STK1 mounting stake, as shown in Figure 40-4. Note the two mating holes for attaching the stake and the 5/16-inch hole for the rubber tubular bushing for routing the input power leads.
7. Obtain a top canopy (CANOPY7). The one shown was selected to mate with the plastic baffle (BAFFLE5) as shown in Figure 40-5. These parts are intended for aesthetics only and may be any combination that appears pleasing to the builder. A good source is a

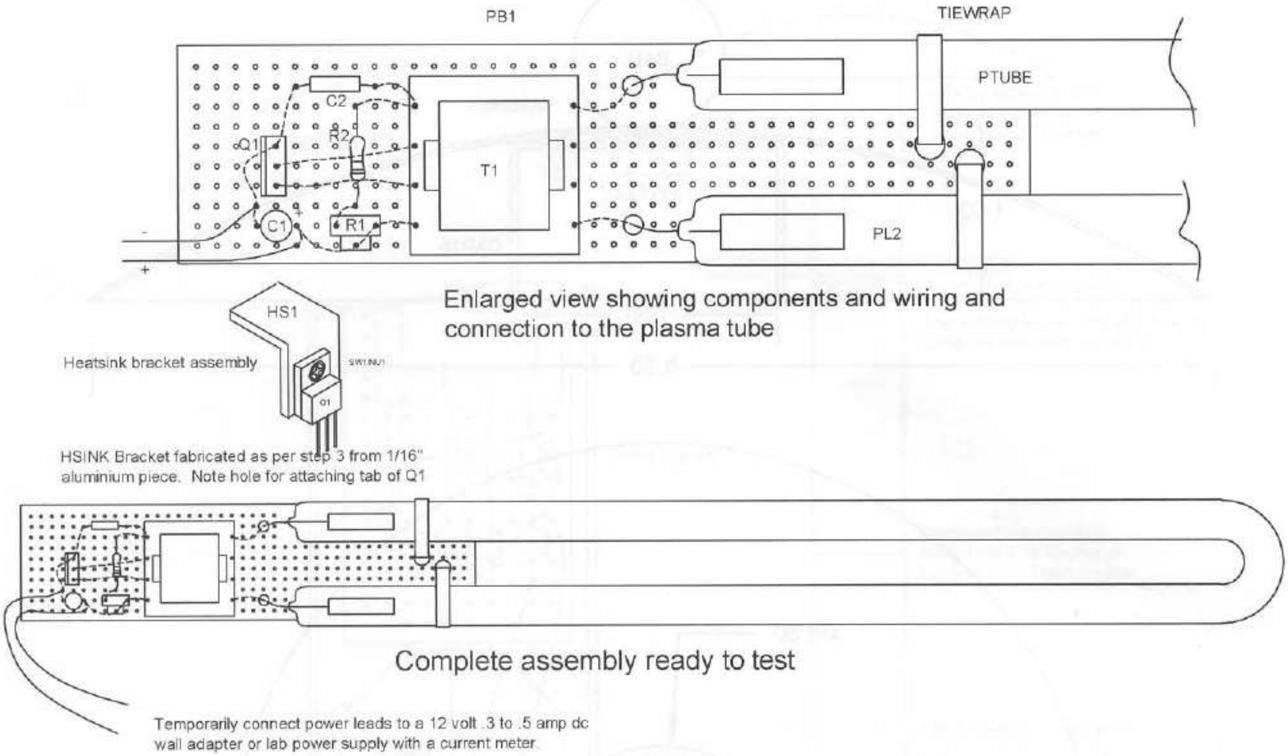
PB1
TIEWRAP
PTUBE
C2
R2
Q1
T1
C1
R1
PL2
-
+
HS1
Enlarged view showing components and wiring and connection to the plasma tube
Heatsink bracket assembly
SW1NU1
Q1
HSINK Bracket fabricated as per step 3 from 1/16" aluminium piece. Note hole for attaching tab of Q1
Complete assembly ready to test
Temporarily connect power leads to a 12 volt .3 to .5 amp dc wall adapter or lab power supply with a current meter.

Figure 40-3 *Assembly board*

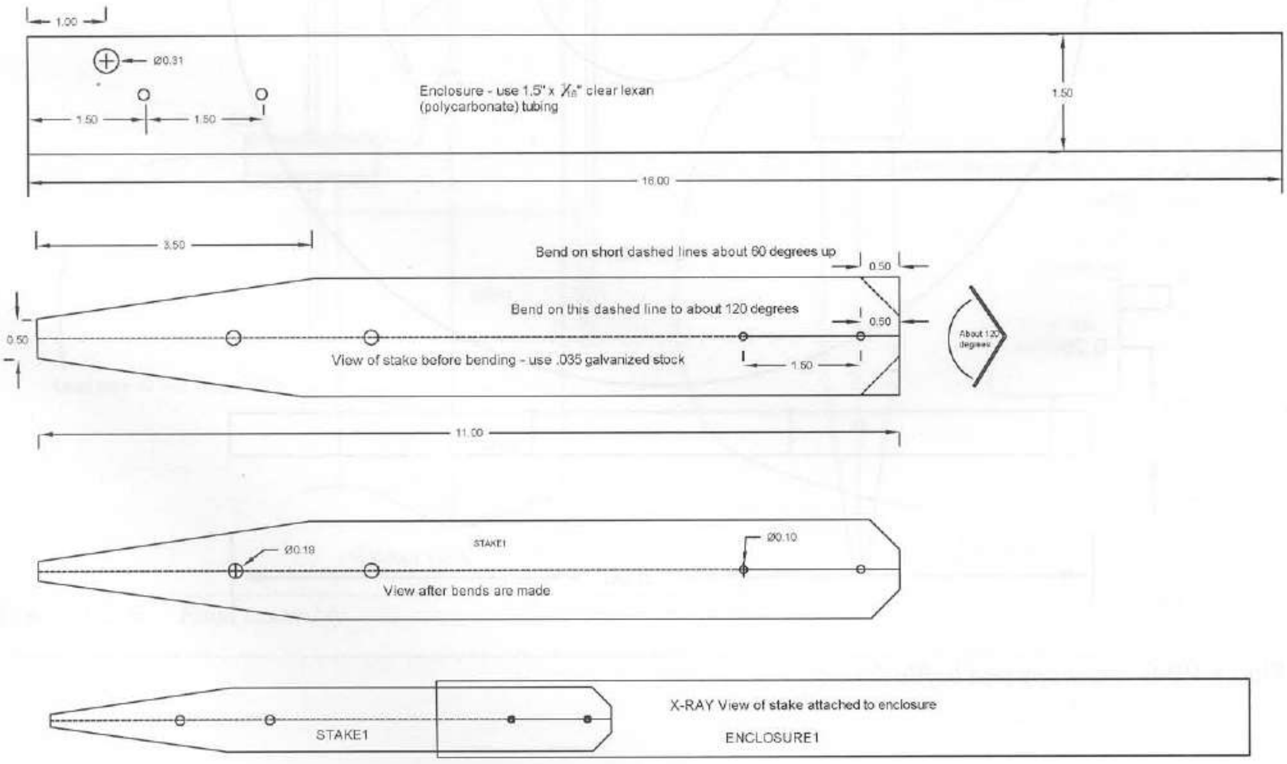
1.00
Ø0.31
1.50
1.50
Enclosure - use 1.5" x 1/16" clear lexan (polycarbonate) tubing
1.50
16.00
3.50
Bend on short dashed lines about 60 degrees up
0.50
Bend on this dashed line to about 120 degrees
0.50
0.50
View of stake before bending - use .035 galvanized stock
1.50
About 120 degrees
11.00
Ø0.19
STAKE1
Ø0.10
View after bends are made
X-RAY View of stake attached to enclosure
STAKE1
ENCLOSURE1

Figure 40-4 *Creating the mounting stake and enclosure*

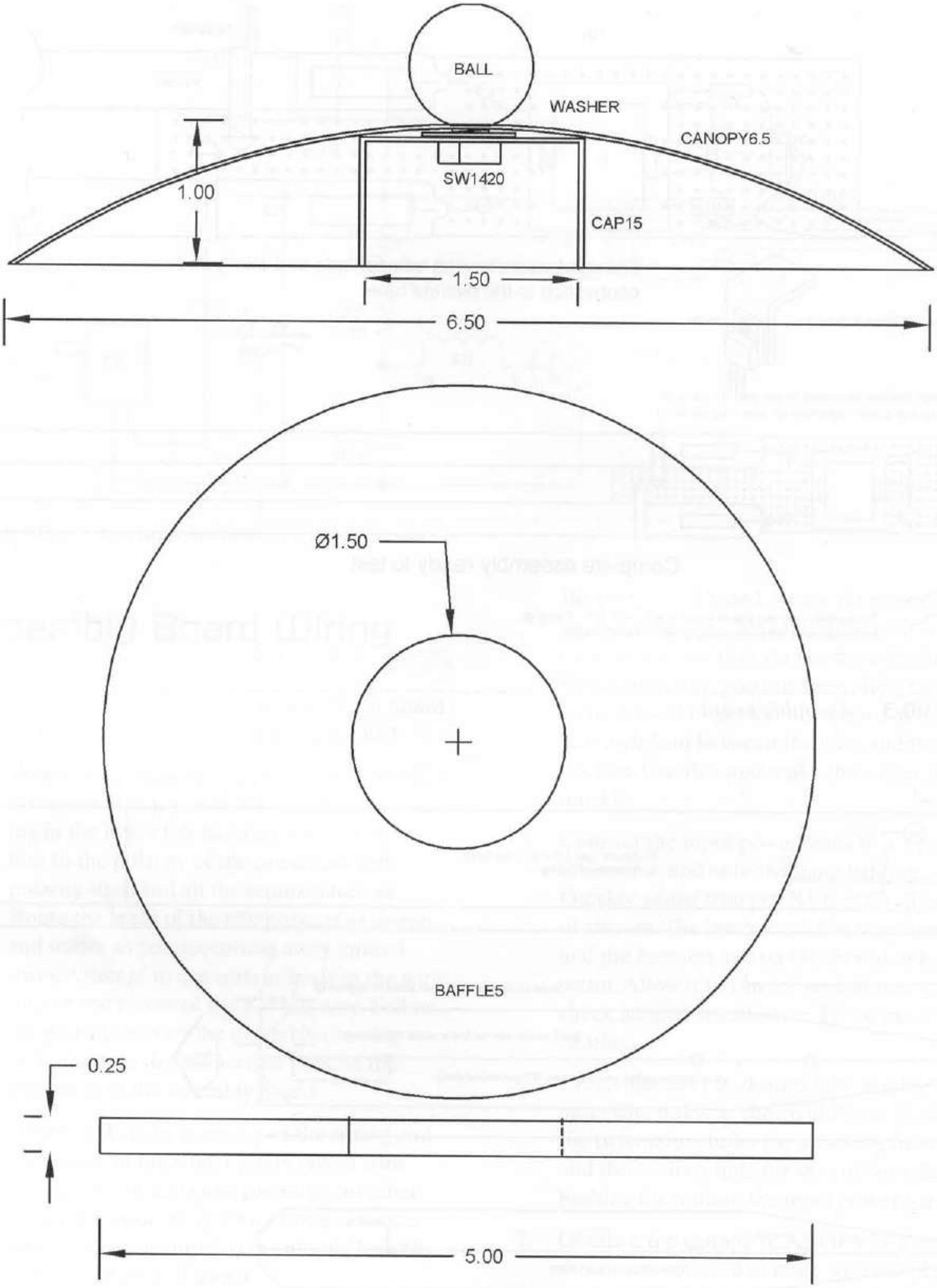
BALL
WASHER
CANOPY6.5
SW1420
1.00
CAP15
1.50
6.50
Ø1.50
BAFFLE5
0.25
5.00

Figure 40-5 *Canopy and baffle design*

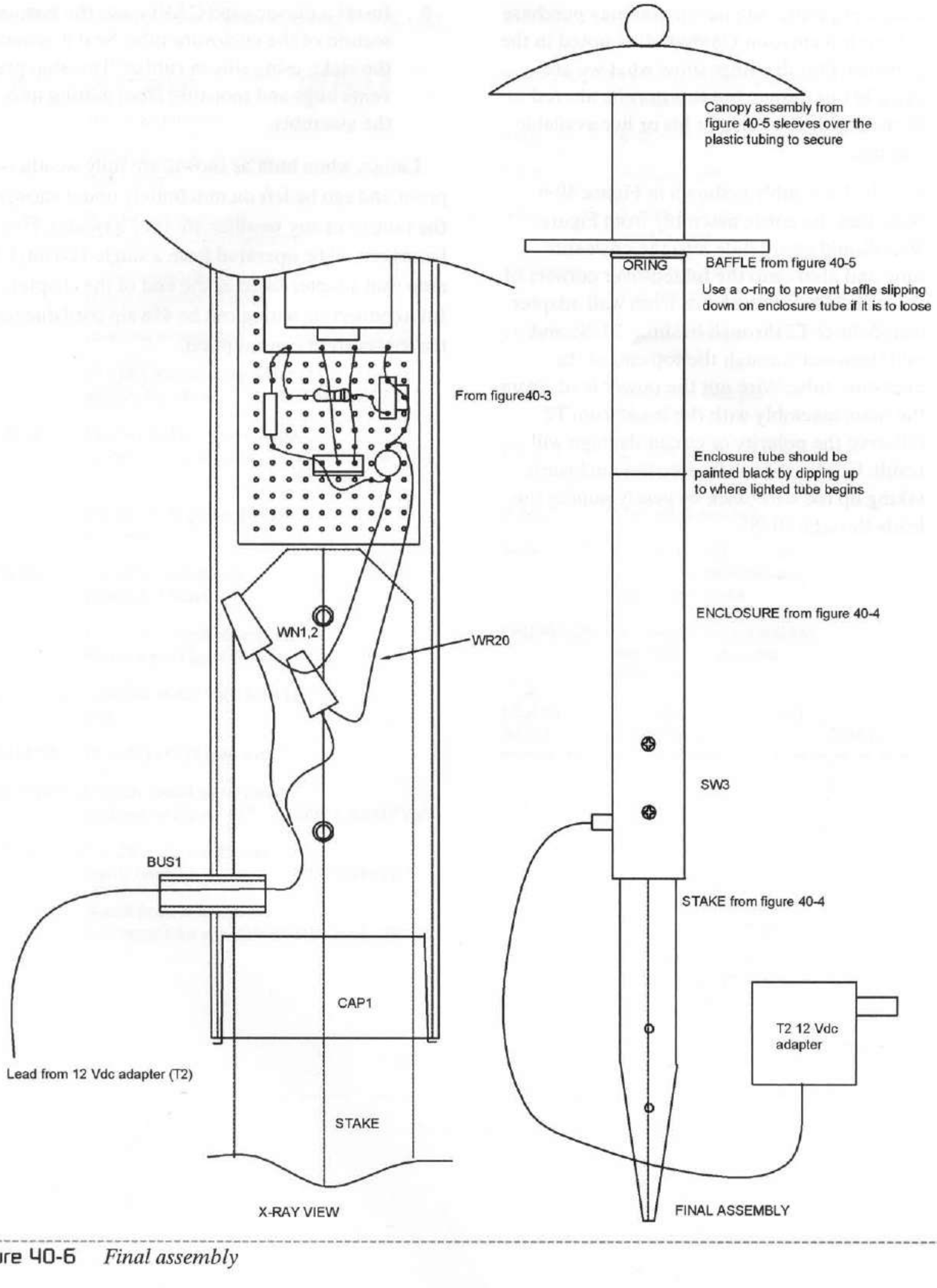
Canopy assembly from figure 40-5 sleeves over the plastic tubing to secure
ORING
BAFFLE from figure 40-5
Use a o-ring to prevent baffle slipping down on enclosure tube if it is to loose
From figure40-3
Enclosure tube should be painted black by dipping up to where lighted tube begins
ENCLOSURE from figure 40-4
WN1,2
WR20
SW3
BUS1
STAKE from figure 40-4
CAP1
T2 12 Vdc adapter
Lead from 12 Vdc adapter (T2)
STAKE
X-RAY VIEW
FINAL ASSEMBLY

Figure 40-6 *Final assembly*

lamp or lighting supplier or you may purchase it from Information Unlimited, as noted in the parts list. Our drawings show what we are using in our design, but this may be altered to fit the builder's choice or his or her available sources.

8. The final assembly is shown in Figure 40-6. Note that the entire assembly from Figure 40-3 should easily slide into the enclosure tube and abut onto the folded-over corners of the stake. Thread the leads from wall adapter transformer T2 through bushing BUS1 and pull them out through the top end of the enclosure tube. Wire nut the power leads from the main assembly with the leads from T2. Observe the polarity or circuit damage will result. Gently slide them into the enclosure, taking up the wire slack by gently pulling the leads through BUS1.
9. Insert a plastic cap (CAP1) into the bottom section of the enclosure tube. Seal it around the stake using silicon rubber. This step prevents bugs and moisture from getting into the assembly.

Lamps, when built as shown, are fully weatherproof and can be left on indefinitely under snow, in the rain, or in any weather, short of a twister. Five lamps can all be operated from a single 12-volt, 1.5-amp wall adapter listed at the end of the chapter. Interconnecting wiring can be #18 zip cord due to the low-current consumption.

Table 40-1 Andromedia Plasma Lamp Parts List

Ref. #	*Description*	*DB Part #*
R1	5K vertical trimmer	
R2	470-ohm, ¼-watt resistor (yel-pur-br)	
C1	10-microfarad, 50-volt vertical electrolytic capacitor	
C2	.033-microfarad, 50-volt plastic capacitor (332)	
T1	Inverting transformer	DB# ANQP2824
T2	12 VDC, 300-milliamp wall adapter transformer	DB# 12 DC/.3
PTUBE-C	Cool white/blue/green/purple/red/yellow plasma tube	DB# PTUBE-COLOR
Q1	PN3055 NPN power T0-220 transistor	
PBOARD	1 × 4.5 × .1-inch grid perforated board	
HS1	1.5 × 1 × 1/16-inch aluminum heatsink per Figure 40-3	
WR20	12 inches of #20 vinyl hookup wire	
WR24BUSS	12 inches of #24 buss wire	
CANOPY6.5	6.5-inch, round metal canopy, as shown in Figure 40-5	DB# CANOPY65
BAFFLE	5 × .25-inch, round, clear plastic baffle	DB# BAFFLE5
BALL	⅞-inch brass ball with a hole tapped for a ¼-20 bolt	DB# BALL78
STAKE1	Metal stake created as shown in Figure 40-4	
ENCLOSURE	Plastic tube created as shown in Figure 40-4	
SW1	6-32 ¼ Phillips screw	
NUT	6-32 keep nut	
SW3	Two #6 × ¼-inch sheet metal screws	
SW1420X1	¼-20 × 1-inch hex bolt for attaching the ball (see Figure 40-5)	
TIEWRAP	Two tie wraps used per Figure 40-3	
CAP1	Plastic cap used per Figure 40-6	
CAP1.5	1.5-inch plastic cap shown in Figure 40-6	
WN1, 2	Two small wire nuts	
BUS1	5/16 OD × 3/16 ID × 1-inch length of rubber tubing (see Figure 40-6)	
ORING1.25	1¼-inch O-ring for holding the baffle in place per Figure 40-6	
5 LAMP ADAPTER	12 VDC, 1.5-amp adapter for five lamps	DB# 12/1 .5

Burn things out, mess things up—that's how you learn.

Make: Electronics

Learn by Discovery

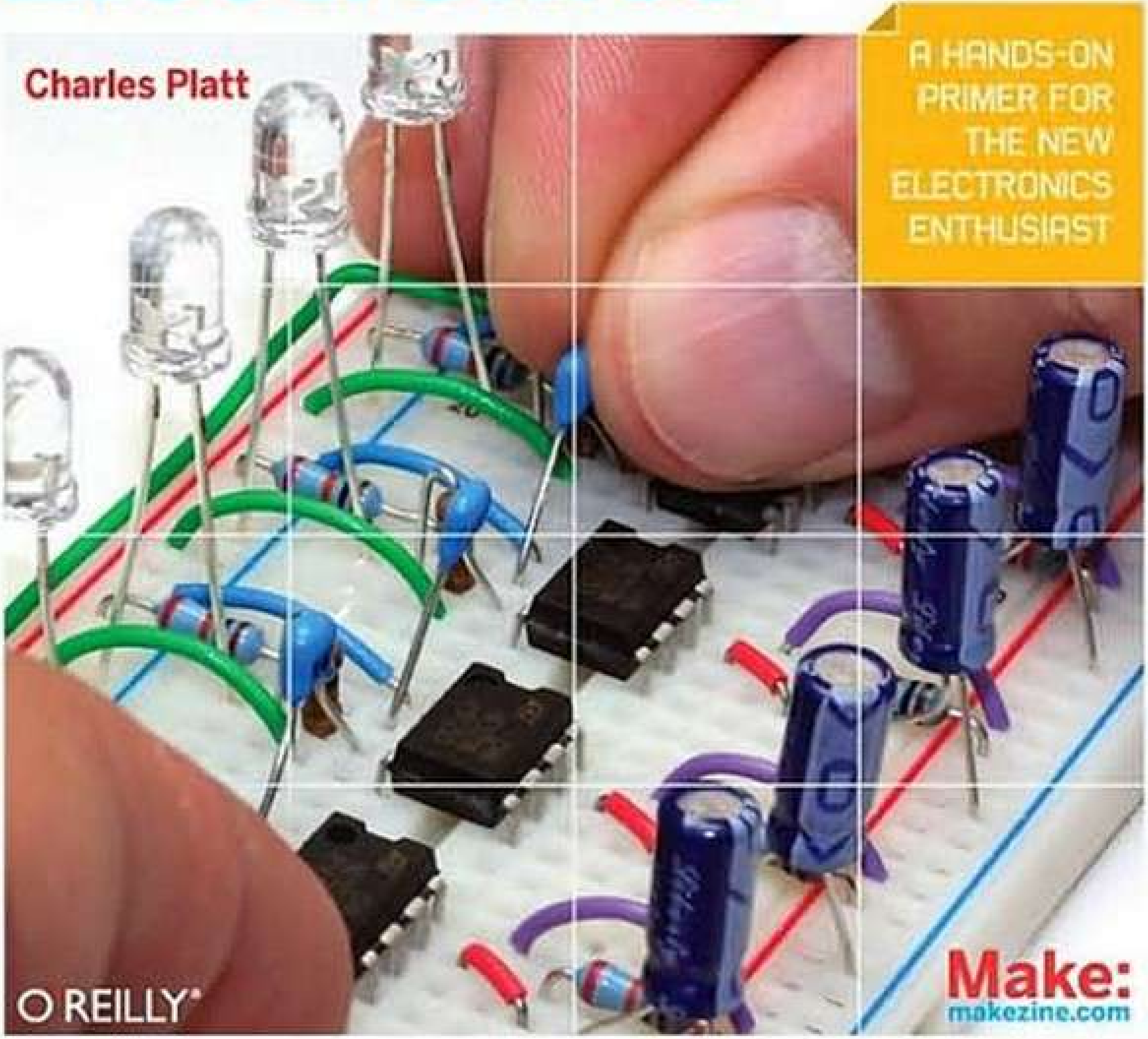

Charles Platt

A HANDS-ON PRIMER FOR THE NEW ELECTRONICS ENTHUSIAST

O'REILLY®

Make:
makezine.com

Contents

Preface ix

1. Experiencing Electricity 1

Shopping List: Experiments 1 Through 5 1

Experiment 1: Taste the Power! 5

Experiment 2: Let's Abuse a Battery! 9

Experiment 3: Your First Circuit 13

Experiment 4: Varying the Voltage 18

Experiment 5: Let's Make a Battery 32

2. Switching Basics and More 39

Shopping List: Experiments 6 Through 11 39

Experiment 6: Very Simple Switching 43

Experiment 7: Relay-Driven LEDs 55

Experiment 8: A Relay Oscillator 60

Experiment 9: Time and Capacitors 68

Experiment 10: Transistor Switching 73

Experiment 11: A Modular Project 82

3. Getting Somewhat More Serious 95

Shopping List: Experiments 12 Through 15 95

Experiment 12: Joining Two Wires Together 104

Experiment 13: Broil an LED 114

Experiment 14: A Pulsing Glow 117

Experiment 15: Intrusion Alarm Revisited 127

4. Chips, Ahoy! 147
Shopping List: Experiments 16 Through 24 147
Experiment 16: Emitting a Pulse 153
Experiment 17: Set Your Tone 162
Experiment 18: Reaction Timer 170
Experiment 19: Learning Logic 181
Experiment 20: A Powerful Combination 197
Experiment 21: Race to Place 205
Experiment 22: Flipping and Bouncing 211
Experiment 23: Nice Dice 214
Experiment 24: Intrusion Alarm Completed 223

5. What Next? 227
Shopping List: Experiments 25 Through 36 228
Customizing Your Work Area 228
Reference Sources 233
Experiment 25: Magnetism 236
Experiment 26: Tabletop Power Generation 239
Experiment 27: Loudspeaker Destruction 242
Experiment 28: Making a Coil React 246
Experiment 29: Filtering Frequencies 249
Experiment 30: Fuzz 257
Experiment 31: One Radio, No Solder, No Power 262
Experiment 32: A Little Robot Cart 268
Experiment 33: Moving in Steps 284
Experiment 34: Hardware Meets Software 293
Experiment 35: Checking the Real World 306
Experiment 36: The Lock, Revisited 311
In Closing 317

Appendix. Online Retail Sources and Manufacturers 319

Index 325

Preface

How to Have Fun with This Book

Everyone uses electronic devices, but most of us don't really know what goes on inside them.

Of course, you may feel that you don't need to know. If you can drive a car without understanding the workings of an internal combustion engine, presumably you can use an iPod without knowing anything about integrated circuits. However, understanding some basics about electricity and electronics can be worthwhile for three reasons:

- By learning how technology works, you become better able to control your world instead of being controlled by it. When you run into problems, you can solve them instead of feeling frustrated by them.
- Learning about electronics can be fun—so long as you approach the process in the right way. The tools are relatively cheap, you can do all the work on a tabletop, and it doesn't consume a lot of time (unless you want it to).
- Knowledge of electronics can enhance your value as an employee or perhaps even lead to a whole new career.

Learning by Discovery

Most introductory guides begin with definitions and facts, and gradually get to the point where you can follow instructions to build a simple circuit.

This book works the other way around. I want you to start putting components together right away. After you see what happens, you'll figure out what's going on. I believe this process of *learning by discovery* creates a more powerful and lasting experience.

Stay Within the Limits!

Although I believe that everything suggested in this book is safe, I'm assuming that you will stay within the limits that I suggest. Please always follow the instructions and pay attention to the warnings, denoted by the icon you see here. If you go beyond the limits, you will expose yourself to unnecessary risks.

Learning by discovery occurs in serious research, when scientists notice an unusual phenomenon that cannot be explained by current theory, and they start to investigate it in an effort to explain it. This may ultimately lead to a better understanding of the world.

We're going to be doing the same thing, although obviously on a much less ambitious level.

Along the way, you will make some mistakes. This is good. Mistakes are the best of all learning processes. I want you to burn things out and mess things up, because this is how you learn the limits of components and materials. Since we'll be using low voltages, there'll be no chance of electrocution, and so long as you limit the flow of current in the ways I'll describe, there will be no risk of burning your fingers or starting fires.

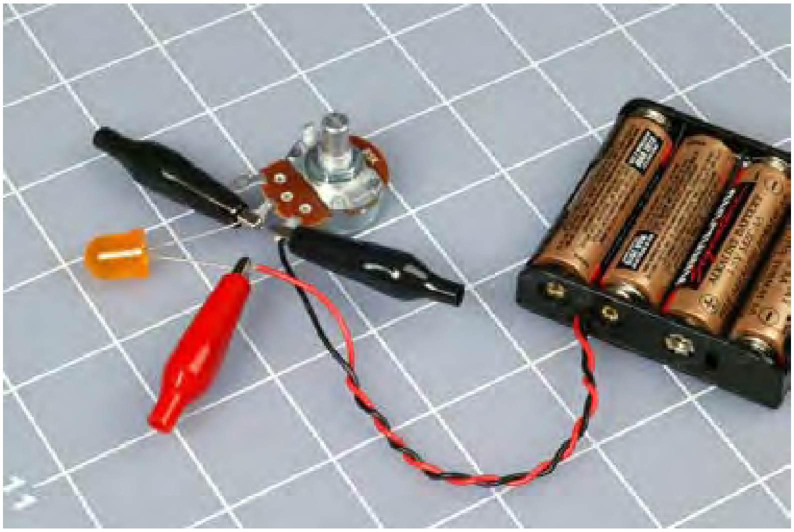

Figure P-1. *Learning by discovery allows you to start building simple circuits right away, using a handful of cheap components, a few batteries, and some alligator clips.*

How Hard Will It Be?

I assume that you're beginning with no prior knowledge of electronics. So, the first few experiments will be ultra-simple, and you won't even use solder or prototyping boards to build a circuit. You'll be holding wires together with alligator clips.

Very quickly, though, you'll be experimenting with transistors, and by the end of Chapter 2, you will have a working circuit that has useful applications.

I don't believe that hobby electronics has to be difficult to understand. Of course, if you want to study electronics more formally and do your own circuit design, this can be challenging. But in this book, the tools and supplies will be inexpensive, the objectives will be clearly defined, and the only math you'll need will be addition, subtraction, multiplication, division, and the ability to move decimal points from one position to another.

Moving Through This Book

Basically there are two ways to present information in a book of this kind: in tutorials and in reference sections. I'm going to use both of these methods. You'll find the tutorials in sections headed as follows:

- Shopping Lists
- Using Tools
- Experiments

You'll find reference sections under the following headings:

- Fundamentals
- Theory
- Background

How you use the sections is up to you. You can skip many of the reference sections and come back to them later. But if you skip many of the tutorials, this book won't be of much use to you. Learning by discovery means that you absolutely, positively have to do some hands-on work, and this in turn means that you have to buy some basic components and play with them. You will gain very little by merely imagining that you are doing this.

It's easy and inexpensive to buy what you need. In almost any urban or suburban area in the United States, chances are you live near a store that sells electronic components and some basic tools to work with them. I am referring, of course, to RadioShack franchises. Some Shacks have more components than others, but almost all of them have the basics that you'll need.

You can also visit auto supply stores such as AutoZone and Pep Boys for basics such as hookup wire, fuses, and switches, while stores such as Ace Hardware, Home Depot, and Lowe's will sell you tools.

If you prefer to buy via mail order, you can easily find everything you need by searching online. In each section of the book, I'll include the URLs of the most popular supply sources, and you'll find a complete list of URLs in the appendix.

Fundamentals

Mail-ordering components and tools

Here are the primary mail-order sources that I use myself online:

http://www.radioshack.com
RadioShack, a.k.a. The Shack. For tools and components. Not always the cheapest, but the site is easy and convenient, and some of the tools are exactly what you need.

http://www.mouser.com
Mouser Electronics.

http://www.digikey.com
Digi-Key Corporation.

http://www.newark.com

Newark.

Mouser, Digi-Key, and Newark are all good sources for components, usually requiring no minimum quantities.

http://www.allelectronics.com

All Electronics Corporation. A narrower range of components, but specifically aimed at the hobbyist, with kits available.

http://www.ebay.com

You can find surplus parts and bargains here, but you may have to try several eBay Stores to get what you want. Those based in Hong Kong are often very cheap, and I've found that they are reliable.

http://www.mcmaster.com

McMaster-Carr. Especially useful for high-quality tools.

Lowe's and Home Depot also allow you to shop online.

McMASTER-CARR
MOUSER ELECTRONICS
Soldering Irons & Guns
RobotShop
Robotics at your service!
SERVO
Digi-Key
ALL ELECTRONICS
3 1/2 DIGIT LCD MULTIMETER W/ BACK
Mr. Clock Radio
LITTER-ROBOT
No more litter box cleaning!
Contact the Experts

Figure P-2. *You'll find no shortage of parts, tools, kits, and gadgets online.*

Companion Kits

Maker Shed (*www.makershed.com*) offers a number of *Make: Electronics* companion kits, both toolkits and bundles of the various components used in the book's experiments. This is a simple, convenient, and cost-effective way of getting all the tools and materials you need to do the projects in this book.

Comments and Questions

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, larger versions of the book's figures, and any additional information. You can access this page at:

http://oreilly.com/catalog/9780596153748

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com

For more information about our books, conferences, Resource Centers, and the O'Reilly Network, see our website at:

http://oreilly.com

Safari® Books Online

Safari Books Online is an on-demand digital library that lets you easily search over 7,500 technology and creative reference books and videos to find the answers you need quickly.

With a subscription, you can read any page and watch any video from our library online. Read books on your cell phone and mobile devices. Access new titles before they are available for print, and get exclusive access to manuscripts in development and post feedback for the authors. Copy and paste code samples, organize your favorites, download chapters, bookmark key sections, create notes, print out pages, and benefit from tons of other time-saving features.

O'Reilly Media has uploaded this book to the Safari Books Online service. To have full digital access to this book and others on similar topics from O'Reilly and other publishers, sign up for free at *http://my.safaribooksonline.com*.

Experiencing Electricity

1

I want you to get a taste for electricity—literally!—in the first experiment. This first chapter of the book will show you:

- How to understand and measure electricity and resistance
- How to handle and connect components without overloading, damaging, or destroying them

Even if you have some prior knowledge of electronics, you should try these experiments before you venture on to the rest of the book.

IN THIS CHAPTER

Shopping List: Experiments 1 Through 5
Experiment 1: Taste the Power!
Experiment 2: Let's Abuse a Battery!
Experiment 3: Your First Circuit
Experiment 4: Varying the Voltage
Experiment 5: Let's Make a Battery

Shopping List: Experiments 1 Through 5

If you want to limit your number of shopping trips or online purchases, look ahead in the book for additional shopping lists, and combine them to make one bulk purchase.

In this first chapter, I will give you part numbers and sources for every tool and component that we'll be using. Subsequently, I won't expect you to need such specific information, because you will have gained experience searching for items on your own.

Maker Shed (www.makershed.com) *has put together a series of* Make: Electronics *companion kits. These include all of the tools and components used in book's experiments. This is a quick, simple, and cost-effective way of getting everything you need to complete the projects in this book.*

Tools

Small pliers

RadioShack Kronus 4.5-inch, part number 64-2953 or Xcelite 4-inch mini long-nose pliers, model L4G.

Or similar. See Figures 1-1 through 1-3. Look for these tools in hardware stores and the sources listed in the preface. The brand is unimportant. After you use them for a while, you'll develop your own preferences. In particular, you have to decide whether you like spring-loaded handles. If you decide you don't, you'll need a second pair of pliers to pull the springs out of the first.

Wire cutters

RadioShack Kronus 4.5-inch, part number 64-2951, or Stanley 7-inch model 84-108.

Or similar. Use them for cutting copper wire, not harder metals (Figure 1-4).

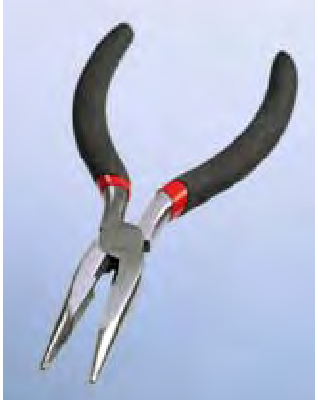

Figure 1-1. *Generic long-nosed pliers are your most fundamental tool for gripping, bending, and picking things up after you drop them.*

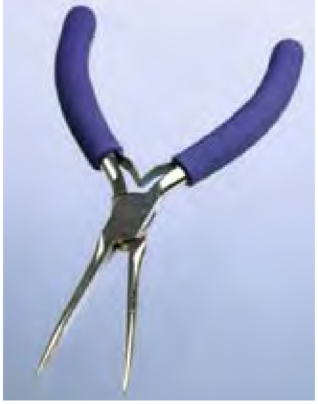

Figure 1-2. *Longer-nosed pliers: these are useful for reaching into tiny spaces.*

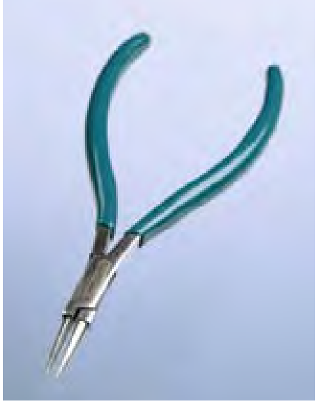

Figure 1-3. *Sharp-pointed pliers are designed for making jewelry, but are also useful for grabbing tiny components.*

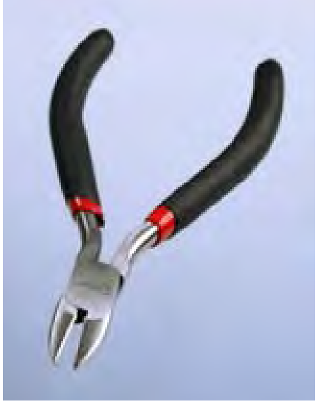

Figure 1-4. *Wire cutters, sometimes known as side cutters, are essential.*

Multimeter

Extech model EX410 or BK Precision model 2704-B or Amprobe model 5XP-A.

Or similar. Because electricity is invisible, we need a tool to visualize the pressure and flow, and a meter is the only way. A cheap meter will be sufficient for your initial experiments. If you buy online, try to check customer reviews, because reliability may be a problem for cheap meters. You can shop around for retailers offering the best price. Don't forget to search on eBay.

The meter must be digital—don't get the old-fashioned analog kind with a needle that moves across a set of printed scales. This book assumes that you are looking at a digital display.

I suggest that you do not buy an autoranging meter. "Autoranging" sounds useful—for example, when you want to check a 9-volt battery, the meter figures out for itself that you are not trying to measure hundreds of volts, nor fractions of a volt. The trouble is that this can trick you into making errors. What if the battery is almost dead? Then you may be measuring a fraction of a volt without realizing it. The only indication will be an easily overlooked "m" for "millivolts" beside the large numerals of the meter display.

On a manual-ranging meter, you select the range, and if the source that you are measuring is outside of that range, the meter tells you that you made an error. I prefer this. I also get impatient with the time it takes for the autoranging feature to figure out the appropriate range each time I make a measurement. But it's a matter of personal preference. See Figures 1-5 through 1-7 for some examples of multimeters.

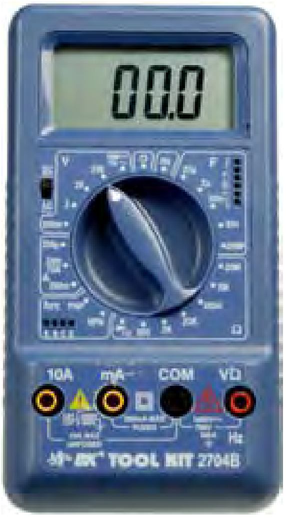

Figure 1-5. *You can see by the wear and tear that this is my own favorite meter. It has all the necessary basic features and can also measure capacitance (the F section, for Farads). It can also check transistors. You have to choose the ranges manually.*

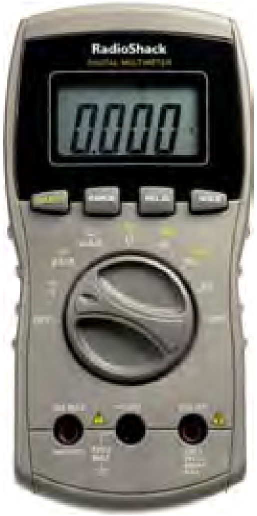

Figure 1-6. *Mid-priced RadioShack meter, which has the basic features; however, the dual purpose for each dial position, selected with the SELECT button, may be confusing. This is an autoranging meter.*

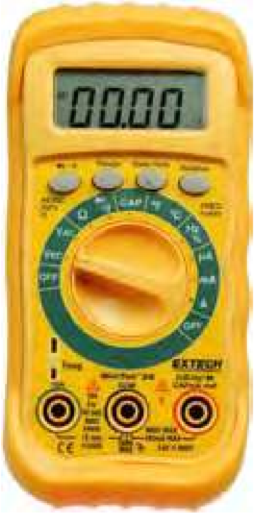

Figure 1-7. *An autoranging meter from Extech offers basic functions, plus a temperature probe, which may be useful to check whether components such as power supplies are running unduly hot.*

Supplies

Batteries

9-volt battery. Quantity: 1.

AA batteries, 1.5 volts each. Quantity: 6.

The batteries should be disposable alkaline, the cheapest available, because we may destroy some of them. You should *absolutely not* use rechargeable batteries in Experiments 1 and 2.

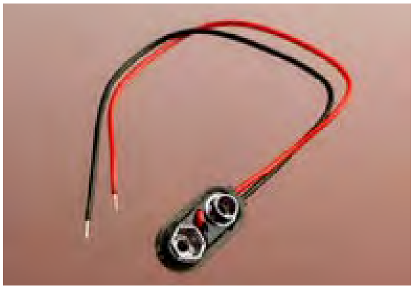

Figure 1-8. *Snap connector for a 9-volt battery.*

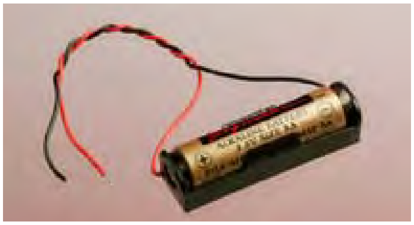

Figure 1-9. *Single AA-sized battery carrier with wires.*

Battery holders and connectors

Snap connector for 9-volt battery, with wires attached (Figure 1-8). Quantity: 1. RadioShack part number 270-325 or similar. Any snap connector that has wires attached will do.

Battery holder for single AA cell, with wires attached (Figure 1-9). Quantity: 1. RadioShack part number 270-401 or Mouser.com catalog number 12BH410-GR, or similar; any single-battery holder that has thin wires attached will do.

Battery holder for four AA cells, with wires attached (Figure 1-10). Quantity: 1. All Electronics catalog number BH-342 or RadioShack part 270-391 or similar.

Alligator clips

Vinyl-insulated. Quantity: at least 6. All Electronics catalog number ALG-28 or RadioShack part number 270-1545 or similar (Figure 1-11).

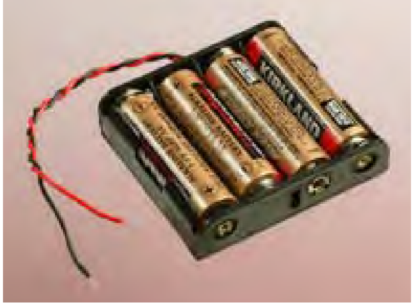

Figure 1-10. *Battery carrier for four AA cells, to be installed in series, delivering 6 volts.*

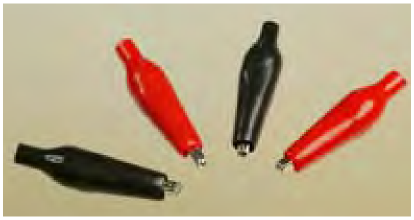

Figure 1-11. *Alligator clips inside vinyl sheaths, which reduce the chance of accidental short circuits.*

Components

You may not know what some of these items are, or what they do. Just look for the part numbers and descriptions, and match them with the photographs shown here. Very quickly, in the learning by discovery process, all will be revealed.

Fuses

Automotive-style, mini-blade type, 3 amps. Quantity: 3. RadioShack part number 270-1089, or Bussmann part ATM-3, available from automotive parts suppliers such as AutoZone (Figure 1-12).

Or similar. A blade-type fuse is easier to grip with alligator clips than a round cartridge fuse.

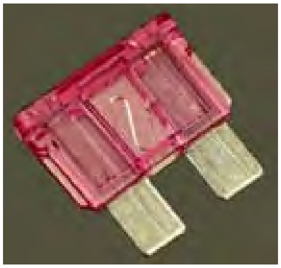

Figure 1-12. *A 3-amp fuse intended primarily for automotive use, shown here larger than actual size.*

Potentiometers

Panel-mount, single-turn, 2K linear, 0.1 watt minimum. Quantity: 2. Alpha part RV170F-10-15R1-B23 or BI Technologies part P160KNPD-2QC25B2K, from Mouser.com or other component suppliers (Figure 1-13).

Or similar. The "watt" rating tells you how much power this component can handle. You don't need more than 0.5 watts.

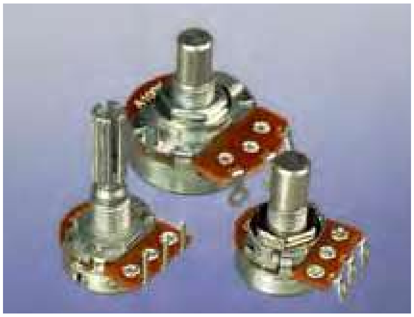

Figure 1-13. *Potentiometers come in many shapes and sizes, with different lengths of shafts intended for different types of knobs. For our purposes, any style will do, but the larger-sized ones are easier to play with.*

Resistors

Assortment 1/4-watt minimum, various values but must include 470 ohms, 1K, and 2K or 2.2K. Quantity: at least 100. RadioShack part number 271-312.

Or search eBay for "resistor assorted."

Light-emitting diodes (LEDs)

Any size or color (Figures 1-14 and 1-15). Quantity: 10. RadioShack part number 276-1622 or All Spectrum Electronics part K/LED1 from Mouser.com.

Or similar. Just about any LEDs will do for these first experiments.

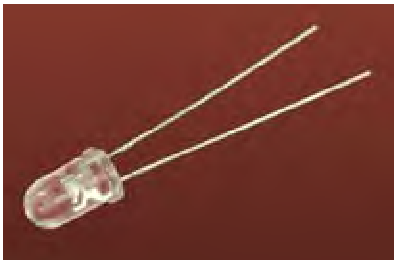

Figure 1-14. *Typical 5-mm diameter light-emitting diode (LED).*

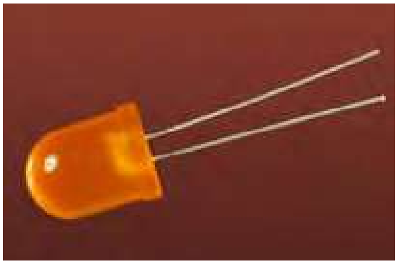

Figure 1-15. *Jumbo-sized LED (1 cm diameter) is not necessarily brighter or more expensive. For most of the experiments in this book, buy whatever LEDs you like the look of.*

Experiment 1: Taste the Power!

Can you taste electricity? Maybe not, but it feels as if you can.

You will need:

- 9-volt battery
- Snap connector for battery terminals
- Multimeter

No More Than 9 Volts

A 9-volt battery won't hurt you. But do not try this experiment with a higher-voltage battery or a larger battery that can deliver more current. Also, if you have metal braces on your teeth, be very careful not to touch them with the battery.

Procedure

Moisten your tongue and touch the tip of it to the metal terminals of a 9-volt battery. The sudden sharp tingle that you feel is caused by electricity flowing from one terminal of the battery (Figure 1-16), through the moisture on and in your tongue, to the other terminal. Because the skin of your tongue is very thin (it's actually a mucus membrane) and the nerves are close to the surface, you can feel the electricity very easily.

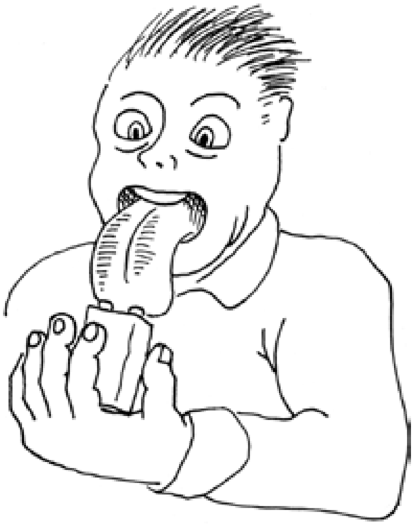

Figure 1-16. *Step 1 in the process of learning by discovery: the 9-volt tongue test.*

Now stick out your tongue, dry the tip of it very thoroughly with a tissue, and repeat the experiment without allowing your tongue to become moist again. You should feel less of a tingle.

What's happening here? We're going to need a meter to find out.

Tools

Setting up your meter

Check the instructions that came with the meter to find out whether you have to install a battery in it, or whether a battery is preinstalled.

Most meters have removable wires, known as *leads* (pronounced "leeds"). Most meters also have three sockets on the front, the leftmost one usually being reserved to measure high electrical currents (flows of electricity). We can ignore that one for now.

The leads will probably be black and red. The black wire plugs into a socket labeled "COM" or "Common." Plug the red one into the socket labeled "V" or "volts." See Figures 1-17 through 1-20.

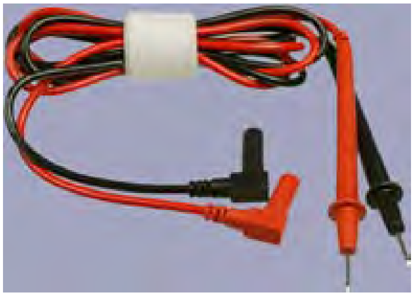

Figure 1-17. *The black lead plugs into the Common (COM) socket, and the red lead plugs into the red socket that's almost always on the righthand side of a multimeter.*

The other ends of the leads terminate in metal spikes known as *probes*, which you will be touching to components when you want to make electrical measurements. The probes detect electricity; they don't emit it in significant quantities. Therefore, they cannot hurt you unless you poke yourself with their sharp ends.

If your meter doesn't do autoranging, each position on the dial will have a number beside it. This number means "no higher than." For instance if you want to check a 6-volt battery, and one position on the voltage section of the dial is numbered 2 and the next position is numbered 20, position 2 means "no higher than 2 volts." You have to go to the next position, which means "no higher than 20 volts."

If you make a mistake and try to measure something inappropriate, the meter will show you an error message such as "E" or "L." Turn the dial and try again.

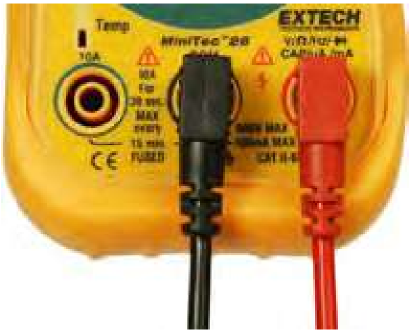

Figure 1-18

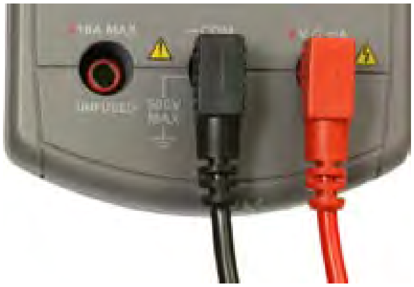

Figure 1-19

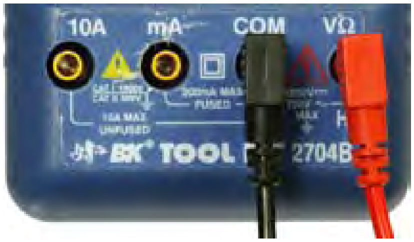

Figure 1-20. *To measure resistance and voltage, plug the black lead into the Common socket and the red lead into the Volts socket. Almost all meters have a separate socket where you must plug the red lead when you measure large currents in amps, but we'll be dealing with this later.*

FUNDAMENTALS

Ohms

We measure distance in miles or kilometers, weight in pounds or kilograms, temperature in Fahrenheit or Centigrade—and electrical resistance in ohms. The ohm is an international unit.

The Greek omega symbol (Ω) is used to indicate ohms, as shown in Figures 1-21 and 1-22. Letter K (or alternatively, KΩ) means a kilohm, which is 1,000 ohms. Letter M (or MΩ) means a megohm, which is 1,000,000 ohms.

Number of ohms	Usually expressed as	Abbreviated as
1,000 ohms	1 kilohm	1KΩ or 1K
10,000 ohms	10 kilohms	10KΩ or 10K
100,000 ohms	100 kilohms	100KΩ or 100K
1,000,000 ohms	1 megohm	1MΩ or 1M
10,000,000 ohms	10 megohms	10MΩ or 10M

A material that has very high resistance to electricity is known as an *insulator*. Most plastics, including the colored sheaths around wires, are insulators.

A material with very low resistance is a *conductor*. Metals such as copper, aluminum, silver, and gold are excellent conductors.

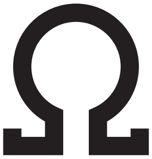

Figure 1-21. *The omega symbol is used internationally to indicate resistance on ohms.*

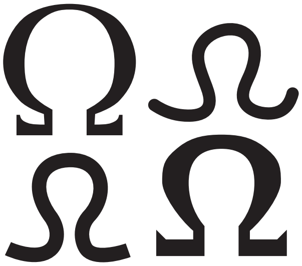

Figure 1-22. *You'll find it printed or written in a wide variety of styles.*

Procedure

We're going to use the meter to discover the electrical resistance of your tongue. First, set your meter to measure resistance. If it has autoranging, look to see whether it is displaying a K, meaning kilohms, or M, meaning megohms. If you have to set the range manually, begin with no less than 100,000 ohms (100K). See Figures 1-23 through 1-25.

Touch the probes to your tongue, about an inch apart. Note the reading, which should be around 50K. Now put aside the probes, stick out your tongue, and use a tissue to dry it very carefully and thoroughly. Without allowing your tongue to become moist again, repeat the test, and the reading should be higher. Finally, press the probes against the skin of your hand or arm: you may get no reading at all, until you moisten your skin.

When your skin is moist (for instance, if you perspire), its electrical resistance decreases. This principle is used in lie detectors, because someone who knowingly tells a lie, under conditions of stress, tends to perspire.

A 9-volt battery contains chemicals that liberate electrons (particles of electricity), which want to flow from one terminal to the other as a result of a chemical reaction inside it. Think of the cells inside a battery as being like two water tanks—one of them full, the other empty. If they are connected with a pipe, water flows between them until their levels are equal. Figure 1-26 may help you visualize this. Similarly, when you open up an electrical pathway between the two sides of a battery, electrons flow between them, even if the pathway consists only of the moisture on your tongue.

Electrons flow more easily through some substances (such as a moist tongue) than others (such as a dry tongue).

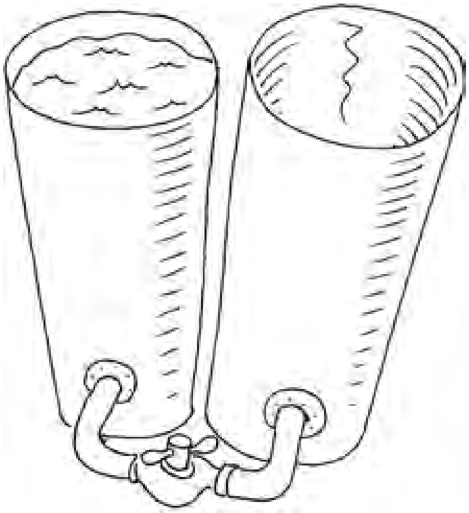

Figure 1-26. *Think of the cells in a battery as being like two cylinders: one full of water, the other empty. Open a connection between the cylinders, and the water will flow until the levels are equal on both sides. The less resistance in the connection, the faster the flow will be.*

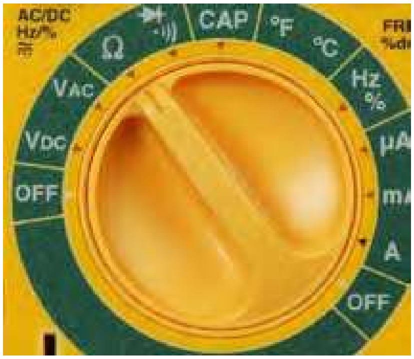

Figure 1-23

Figure 1-24

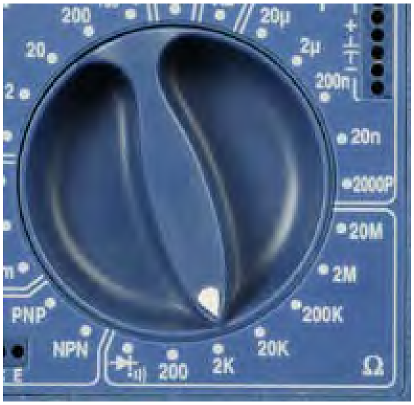

Figure 1-25. *To measure ohms, turn the dial to the ohm (omega) symbol. On an autoranging meter, you can then press the Range button repeatedly to display different ranges of resistance, or simply touch the probes to a resistance and wait for the meter to choose a range automatically. A manual meter requires you to select the range with the dial (you should set it to 100K or higher, to measure skin resistance). If you don't get a meaningful reading, try a different range.*

BACKGROUND

The man who discovered resistance

Georg Simon Ohm, pictured in Figure 1-27, was born in Bavaria in 1787 and worked in obscurity for much of his life, studying the nature of electricity using metal wire that he had to make for himself (you couldn't truck on down to Home Depot for a spool of hookup wire back in the early 1800s).

Despite his limited resources and inadequate mathematical abilities, Ohm was able to demonstrate in 1827 that the electrical resistance of a conductor such as copper varied in proportion with its area of cross-section, and the current flowing through it is proportional to the voltage applied to it, as long as temperature is held constant. Fourteen years later, the Royal Society in London finally recognized the significance of his contribution and awarded him the Copley Medal. Today, his discovery is known as Ohm's Law.

Figure 1-27. *Georg Simon Ohm, after being honored for his pioneering work, most of which he pursued in relative obscurity.*

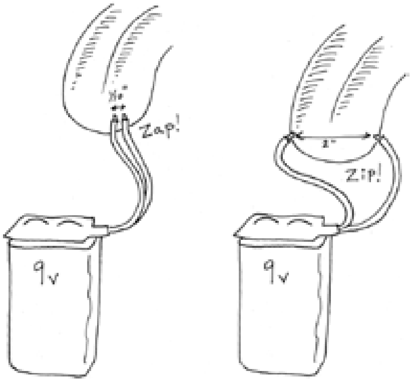

Figure 1-28. *Modifying the tongue test to show that a shorter distance, with lower resistance, allows greater flow of electricity, and a bigger zap.*

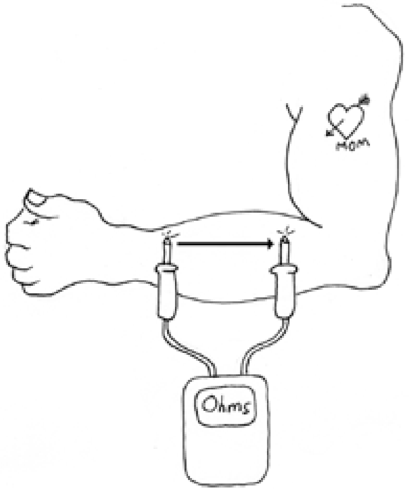

Figure 1-29. *Moisten your skin before trying to measure its resistance. You should find that the resistance goes up as you move the meter probes farther apart. The resistance is proportional to the distance.*

Further Investigation

Attach the snap-on terminal cap (shown earlier in Figure 1-8) to the 9-volt battery. Take the two wires that are attached to the cap and hold them so that the bare ends are just a few millimeters apart. Touch them to your tongue. Now separate the ends of the wires by a couple of inches, and touch them to your tongue again. (See Figure 1-28.) Notice any difference?

Use your meter to measure the electrical resistance of your tongue, this time varying the distance between the two probes. When electricity travels through a shorter distance, it encounters less total resistance. As a result, the current (the flow of electricity per second) increases. You can try a similar experiment on your arm, as shown in Figure 1-29.

Use your meter to test the electrical resistance of water. Dissolve some salt in the water, and test it again. Now try measuring the resistance of distilled water (in a clean glass).

The world around you is full of materials that conduct electricity with varying amounts of resistance.

Cleanup and Recycling

Your battery should not have been damaged or significantly discharged by this experiment. You'll be able to use it again.

Remember to switch off your meter before putting it away.

Experiment 2: Let's Abuse a Battery!

To get a better feeling for electrical power, you're going to do what most books tell you not to do. You're going to short out a battery. A short circuit is a direct connection between the two sides of a power source.

Short Circuits

Short circuits can be dangerous. Do not short out a power outlet in your home: there'll be a loud bang, a bright flash, and the wire or tool that you use will be partially melted, while flying particles of melted metal can burn you or blind you.

If you short out a car battery, the flow of current is so huge that the battery might even explode, drenching you in acid (Figure 1-30).

Lithium batteries are also dangerous. Never short-circuit a lithium battery: it can catch fire and burn you (Figure 1-31).

Use only an alkaline battery in this experiment, and only a single AA cell (Figure 1-32). You should also wear safety glasses in case you happen to have a defective battery.

You will need:

- 1.5-volt AA battery
- Single-battery carrier
- 3-amp fuse
- Safety glasses (regular eyeglasses or sunglasses will do)

Procedure

Use an alkaline battery. Do not use any kind of rechargeable battery.

Put the battery into a battery holder that's designed for a single battery and has two thin insulated wires emerging from it, as shown in Figure 1-32. Do not use any other kind of battery holder.

Use an alligator clip to connect the bare ends of the wires, as shown in Figure 1-32. There will be no spark, because you are using only 1.5 volts. Wait one minute, and you'll find that the wires are getting hot. Wait another minute, and the battery, too, will be hot.

Figure 1-30. *Anyone who has dropped an adjustable wrench across the bare terminals of a car battery will tell you that short circuits can be dramatic at a "mere" 12 volts, if the battery is big enough.*

Figure 1-31. *The low internal resistance of lithium batteries (which are often used in laptop computers) allows high currents to flow, with unexpected results. Never fool around with lithium batteries!*

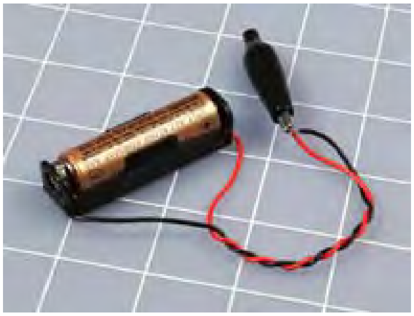

Figure 1-32. *Shorting out an alkaline battery can be safe if you follow the directions precisely. Even so, the battery is liable to become too hot to touch comfortably. Don't try this with any type of rechargeable battery.*

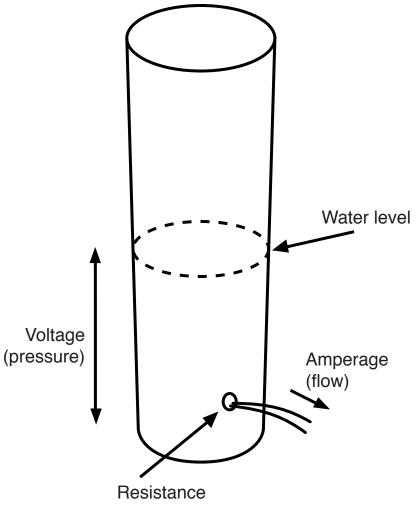

Figure 1-33. *Think of voltage as pressure, and amperes as flow.*

The heat is caused by electricity flowing through the wires and through the electrolyte (the conductive fluid) inside the battery. If you've ever used a hand pump to force air into a bicycle tire, you know that the pump gets warm. Electricity behaves in much the same way. You can imagine the electricity being composed of particles (electrons) that make the wire hot as they push through it. This isn't a perfect analogy, but it's close enough for our purposes.

Chemical reactions inside the battery create electrical pressure. The correct name for this pressure is *voltage*, which is measured in volts and is named after Alessandro Volta, an electrical pioneer.

Going back to the water analogy: the height of the water in a tank is proportionate to the pressure of the water, and comparable to voltage. Figure 1-33 may help you to visualize this.

But volts are only half of the story. When electrons flow through a wire, the flow is known as *amperage*, named after yet another electrical pioneer, André-Marie Ampère. The flow is also generally known as current. It's the current—the amperage—that generates the heat.

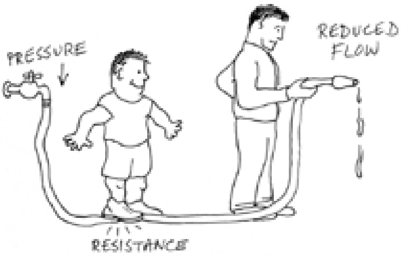

Figure 1-34. *Larger resistance results in smaller flow—but if you increase the pressure, it may overcome the resistance and increase the flow.*

BACKGROUND

Why didn't your tongue get hot?

When you touched the 9-volt battery to your tongue, you felt a tingle, but no perceptible heat. When you shorted out a battery, you generated a noticeable amount of heat, even though you used a lower voltage. How can we explain this?

The electrical resistance of your tongue is very high, which reduces the flow of electrons. The resistance of a wire is very low, so if there's only a wire connecting the two terminals of the battery, more current will pass through it, creating more heat. If all other factors remain constant:

- Lower resistance allows more current to flow (Figure 1-34).
- The heat generated by electricity is proportional to the amount of electricity (the current) that flows.

Here are some other basic concepts:

- The flow of electricity per second is measured in amperes, or amps.
- The pressure of electricity causes the flow, measured in volts.
- The resistance to the flow is measured in ohms.
- A higher resistance restricts the current.
- A higher voltage overcomes resistance and increases the current.

If you're wondering exactly how much current flows between the terminals of a battery when you short it out, that's a difficult question to answer. If you try to use your multimeter to measure it, you're liable to blow the fuse inside the meter. Still, you can use your very own 3-amp fuse, which we can sacrifice because it didn't cost very much.

First inspect the fuse very carefully, using a magnifying glass if you have one. You should see a tiny S-shape in the transparent window at the center of the fuse. That S is a thin section of metal that melts easily.

Remove the battery that you short-circuited. It is no longer useful for anything, and should be recycled if possible. Put a fresh battery into the battery carrier, connect the fuse as shown in Figure 1-35, and take another look. You should see a break in the center of the S shape, where the metal melted almost instantly. Figure 1-36 shows the fuse before you connected it, and Figure 1-37 depicts a blown fuse. This is how a fuse works: it melts to protect the rest of the circuit. That tiny break inside the fuse stops any more current from flowing.

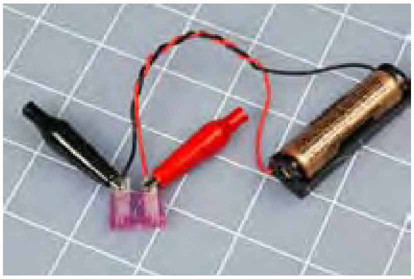

Figure 1-35. *When you attach both wires to the fuse, the little S-shaped element inside will melt almost instantly.*

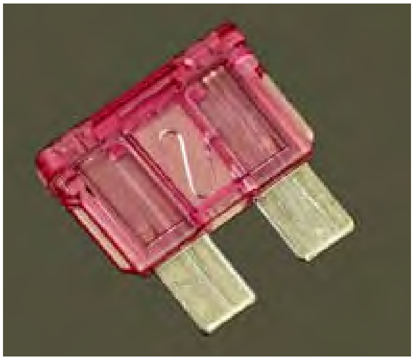

Figure 1-36. *A 3-amp fuse, before its element was melted by a single 1.5-volt battery.*

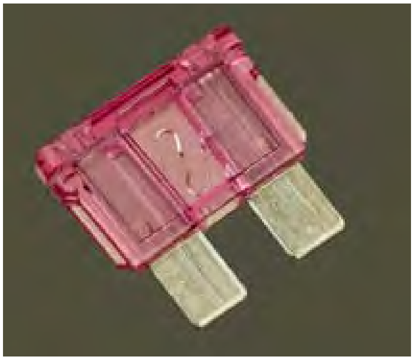

Figure 1-37. *The same fuse after being melted by electric current.*

FUNDAMENTALS

Volt basics

Electrical pressure is measured in volts. The volt is an international unit. A millivolt is 1/1,000 of a volt.

Number of volts	Usually expressed as	Abbreviated as
0.001 volts	1 millivolt	1 mV
0.01 volts	10 millivolts	10 mV
0.1 volts	100 millivolts	100 mV
1 volt	1,000 millivolts	1V

Ampere basics

We measure electrical flow in amperes, or amps. The ampere is an international unit, often referred to as an "amp." A milliamp is 1/1,000 of an ampere.

Number of amperes	Usually expressed as	Abbreviated as
0.001 amps	1 milliamp	1 mA
0.01 amps	10 milliamps	10 mA
0.1 amps	100 milliamps	100 mA
1 amp	1,000 milliamps	1A

BACKGROUND

Inventor of the battery

Alessandro Volta (Figure 1-38) was born in Italy in 1745, long before science was broken up into specialties. After studying chemistry (he discovered methane in 1776), he became a professor of physics and became interested in the so-called galvanic response, whereby a frog's leg will twitch in response to a jolt of static electricity.

Using a wine glass full of salt water, Volta demonstrated that the chemical reaction between two electrodes, one made of copper, the other of zinc, will generate a steady electric current. In 1800, he refined his apparatus by stacking plates of copper and zinc, separated by cardboard soaked in salt and water. This "voltaic pile" was the first electric battery.

Figure 1-38. *Alessandro Volta discovered that chemical reactions can create electricity.*

FUNDAMENTALS

Direct and alternating current

The flow of current that you get from a battery is known as *direct current*, or DC. Like the flow of water from a faucet, it is a steady stream, in one direction.

The flow of current that you get from the "hot" wire in a power outlet in your home is very different. It changes from positive to negative 50 times each second (in Great Britain and some other nations, 60 times per second). This is known as *alternating current*, or AC, which is more like the pulsatile flow you get from a power washer.

Alternating current is essential for some purposes, such as cranking up voltage so that electricity can be distributed over long distances. AC is also useful in motors and domestic appliances. The parts of an American power outlet are shown in Figure 1-39. A few other nations, such as Japan, also use American-style outlets.

For most of this book I'm going to be talking about DC, for two reasons: first, most simple electronic circuits are powered with DC, and second, the way it behaves is much easier to understand.

I won't bother to mention repeatedly that I'm dealing with DC. Just assume that everything is DC unless otherwise noted.

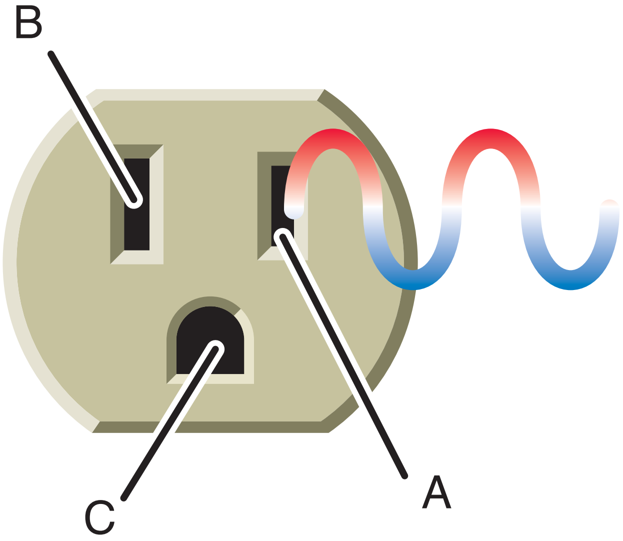

Figure 1-39. *This style of power outlet is found in North America, South America, Japan, and some other nations. European outlets look different, but the principle remains the same. Socket A is the "live" side of the outlet, supplying voltage that alternates between positive and negative, relative to socket B, which is called the "neutral" side. If an appliance develops a fault such as an internal loose wire, it should protect you by sinking the voltage through socket C, the ground.*

Cleanup and Recycling

The first AA battery that you shorted out is probably damaged beyond repair. You should dispose of it. Putting batteries in the trash is not a great idea, because they contain heavy metals that should be kept out of the ecosystem. Your state or town may include batteries in a local recycling scheme. (California requires that almost all batteries be recycled.) You'll have to check your local regulations for details.

The blown fuse is of no further use, and can be thrown away.

The second battery, which was protected by the fuse, should still be OK. The battery holder also can be reused later.

Experiment 3: Your First Circuit

Now it's time to make electricity do something that's at least slightly useful. For this purpose, you'll use components known as resistors, and a light-emitting diode, or LED.

You will need:

- 1.5-volt AA batteries. Quantity: 4.
- Four-battery holder. Quantity: 1.
- Resistors: 470Ω, 1K, and either 2K or 2.2K (the 2.2K value happens to be more common than 2K, but either will do in this experiment). Quantity: 1 of each resistor.
- An LED, any type. Quantity: 1.
- Alligator clips. Quantity: 3.

Setup

It's time to get acquainted with the most fundamental component we'll be using in electronic circuits: the humble resistor. As its name implies, it resists the flow of electricity. As you might expect, the value is measured in ohms.

If you bought a bargain-basement assortment package of resistors, you may find nothing that tells you their values. That's OK; we can find out easily enough. In fact, even if they are clearly labeled, I want you to check their values yourself. You can do it in two ways:

- Use your multimeter. This is excellent practice in learning to interpret the numbers that it displays.
- Learn the color codes that are printed on most resistors. See the following section, "Fundamentals: Decoding resistors," for instructions.

After you check them, it's a good idea to sort them into labeled compartments in a little plastic parts box. Personally, I like the boxes sold at the Michaels chain of crafts stores, but you can find them from many sources.

BACKGROUND

Father of electromagnetism

Born in 1775 in France, André-Marie Ampère (Figure 1-40) was a mathematical prodigy who became a science teacher, despite being largely self-educated in his father's library. His best-known work was to derive a theory of electromagnetism in 1820, describing the way that an electric current generates a magnetic field. He also built the first instrument to measure the flow of electricity (now known as a *galvanometer*), and discovered the element fluorine.

Figure 1-40. *Andre-Marie Ampere found that an electric current running through a wire creates a magnetic field around it. He used this principle to make the first reliable measurements of what came to be known as amperage.*

FUNDAMENTALS

Decoding resistors

Some resistors have their value clearly stated on them in microscopic print that you can read with a magnifying glass. Most, however, are color-coded with stripes. The code works like this: first, ignore the color of the body of the resistor. Second, look for a silver or gold stripe. If you find it, turn the resistor so that the stripe is on the righthand side. Silver means that the value of the resistor is accurate within 10%, while gold means that the value is accurate within 5%. If you don't find a silver or gold stripe, turn the resistor so that the stripes are clustered at the left end. You should now find yourself looking at three colored stripes on the left. Some resistors have more stripes, but we'll deal with those in a moment. See Figures 1-41 and 1-42.

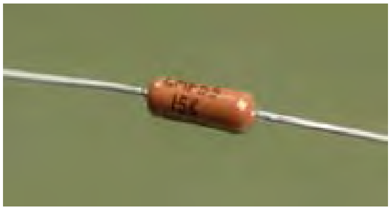

Figure 1-41. *Some modern resistors have their values printed on them, although you may need a magnifier to read them. This 15K resistor is less than half an inch long.*

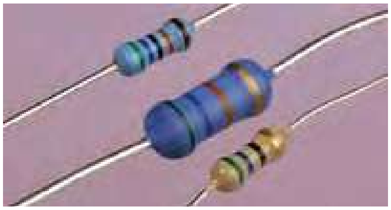

Figure 1-42. *From top to bottom, these resistor values are 56,000 ohms (56K), 5,600 ohms (5.6K), and 560 ohms. The size tells you how much power the resistor can handle; it has nothing to do with the resistance. The smaller components are rated at 1/4 watt; the larger one in the center can handle 1 watt of power.*

Starting from the left, the first and second stripes are coded according to this table:

Black	0
Brown	1
Red	2
Orange	3
Yellow	4
Green	5
Blue	6
Violet	7
Gray	8
White	9

The third stripe has a different meaning: It tells you how many zeros to add, like this:

Black	-	No zeros
Brown	0	1 zero
Red	00	2 zeros
Orange	000	3 zeros
Yellow	0000	4 zeros
Green	00000	5 zeros
Blue	000000	6 zeros
Violet	0000000	7 zeros
Gray	00000000	8 zeros
White	000000000	9 zeros

FUNDAMENTALS

Decoding resistors (continued)

Note that the color-coding is consistent, so that green, for instance, means either a value of 5 (for the first two stripes) or 5 zeros (for the third stripe). Also, the sequence of colors is the same as their sequence in a rainbow.

So, a resistor colored brown-red-green would have a value of 1-2 and five zeros, making 1,200,000 ohms, or 1.2MΩ. A resistor colored orange-orange-orange would have a value of 3-3 and three zeros, making 33,000 ohms, or 33KΩ. A resistor colored brown-black-red would have a value of 1-0 and two additional zeros, or 1KΩ. Figure 1-43 shows some other examples.

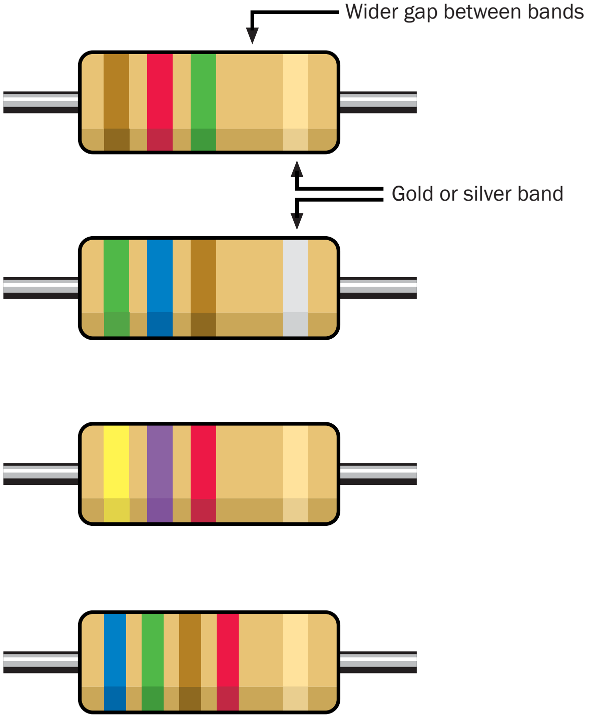

Figure 1-43. *To read the value of a resistor, first turn it so that the silver or gold stripe is on the right, or the other stripes are clustered on the left. From top to bottom: The first resistor has a value of 1-2 and five zeros, or 1,200,000, which is 1.2MΩ. The second is 5-6 and one zero, or 560Ω. The third is 4-7 and two zeros, or 4,700, which is 4.7KΩ. The last is 6-5-1 and two zeros, or 65,100Ω, which is 65.1KΩ.*

If you run across a resistor with four stripes instead of three, the first *three* stripes are digits and the *fourth* stripe is the number of zeros. The third numeric stripe allows the resistor to be calibrated to a finer tolerance.

Confusing? Absolutely. That's why it's easier to use your meter to check the values. Just be aware that the meter reading may be slightly different from the claimed value of the resistor. This can happen because your meter isn't absolutely accurate, or because the resistor is not absolutely accurate, or both. As long as you're within 5% of the claimed value, it doesn't matter for our purposes.

Lighting an LED

Now take a look at one of your LEDs. An old-fashioned lightbulb wastes a lot of power by converting it into heat. LEDs are much smarter: they convert almost all their power into light, and they last almost indefinitely—as long as you treat them right!

An LED is quite fussy about the amount of power it gets, and the way it gets it. Always follow these rules:

- The *longer* wire protruding from the LED must receive a *more positive* voltage than the shorter wire.
- The voltage difference between the long wire and the short wire must not exceed the limit stated by the manufacturer.
- The current passing through the LED must not exceed the limit stated by the manufacturer.

What happens if you break these rules? Well, we're going to find out!

Make sure you are using fresh batteries. You can check by setting your multimeter to measure volts DC, and touching the probes to the terminals of each battery. You should find that each of them generates a pressure of at least 1.5 volts. If they read slightly higher than this, it's normal. A battery starts out above its rated voltage, and delivers progressively less as you use it. Batteries also lose some voltage while they are sitting on the shelf doing nothing.

Load your battery holder (taking care that the batteries are the right way around, with the negative ends pressing against the springs in the carrier). Use your meter to check the voltage on the wires coming out of the battery carrier. You should have at least 6 volts.

Now select a 2KΩ resistor. Remember, "2KΩ" means "2,000 ohms." If it has colored stripes, they should be red-black-red, meaning 2-0 and two more zeros. Because 2.2K resistors are more common than 2K resistors, you can substitute one of them if necessary. It will be colored red-red-red.

Wire it into the circuit as shown in Figures 1-44 and 1-45, making the connections with alligator clips. You should see the LED glow very dimly.

Now swap out your 2K resistor and substitute a 1K resistor, which will have brown-black-red stripes, meaning 1-0 and two more zeros. The LED should glow more brightly.

Swap out the 1K resistor and substitute a 470Ω resistor, which will have yellow-violet-brown stripes, meaning 4-7 and one more zero. The LED should be brighter still.

This may seem very elementary, but it makes an important point. The resistor blocks a percentage of the voltage in the circuit. Think of it as being like a kink or constriction in a flexible hose. A higher-value resistor blocks more voltage, leaving less for the LED.

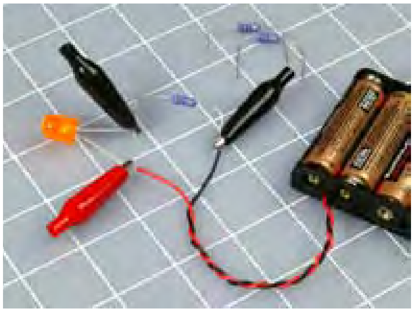

Figure 1-44. *The setup for Experiment 3, showing resistors of 470Ω, 1KΩ, and 2KΩ. Apply alligator clips where shown, to make a secure contact, and try each of the resistors one at a time at the same point in the circuit, while watching the LED.*

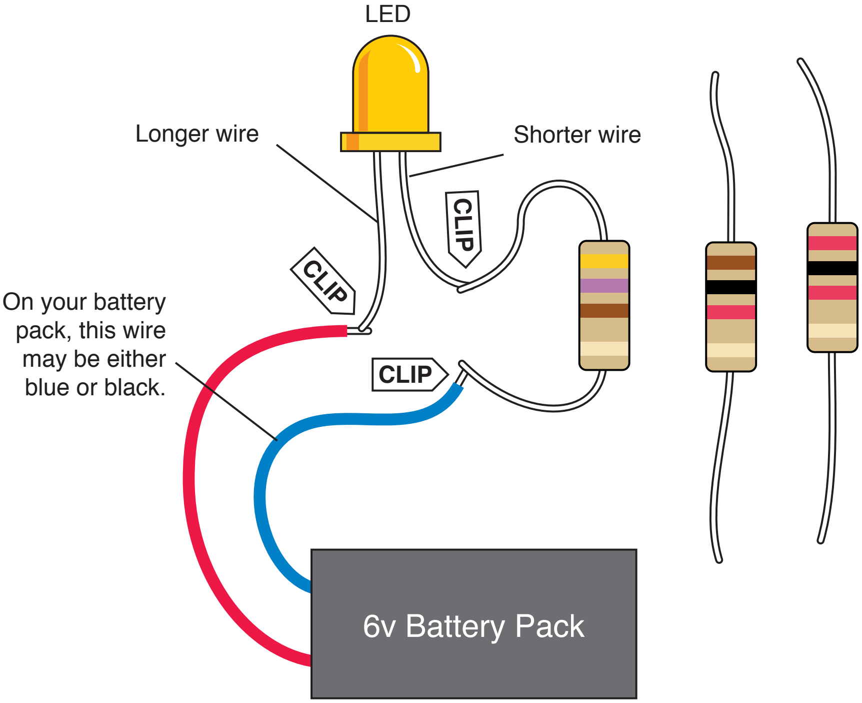
LED
Longer wire
Shorter wire
CLIP
CLIP
CLIP
On your battery pack, this wire may be either blue or black.
6v Battery Pack

Figure 1-45. *Here's how it actually looks, using a large LED. If you start with the highest value resistor, the LED will glow very dimly as you complete the circuit. The resistor drops most of the voltage, leaving the LED with insufficient current to make it shine brightly.*

Cleanup and Recycling

We'll use the batteries and the LED in the next experiment. The resistors can be reused in the future.

Experiment 4: Varying the Voltage

Potentiometers come in various shapes and sizes, but they all do the same thing: they allow you to vary voltage and current by varying resistance. This experiment will enable you to learn more about voltage, amperage, and the relationship between them. You'll also learn how to read a manufacturer's data sheet.

You will need the same batteries, battery carrier, alligator clips, and LED from the last experiment, plus:

- Potentiometer, 2KΩ linear. Quantity: 2. (See Figure 1-46.) Full-sized potentiometers that look like this are becoming less common, as miniature versions are taking their place. I'd like you to use a large one, though, because it's so much easier to work with.
- One extra LED.
- Multimeter.

Look Inside Your Potentiometer

The first thing I want you to do is find out how a potentiometer works. This means you'll have to open it, which is why your shopping list required you to buy two of them, in case you can't put the first one back together again.

Most potentiometers are held together with little metal tabs. You should be able to grab hold of the tabs with your wire cutters or pliers, and bend them up and outward. If you do this, the potentiometer should open up as shown in Figures 1-47 and 1-48.

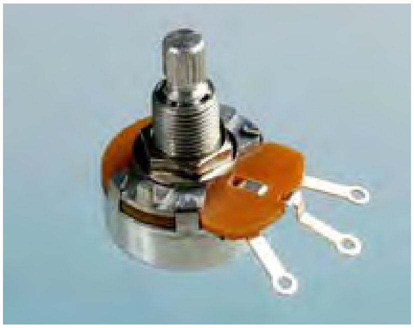

Figure 1-46

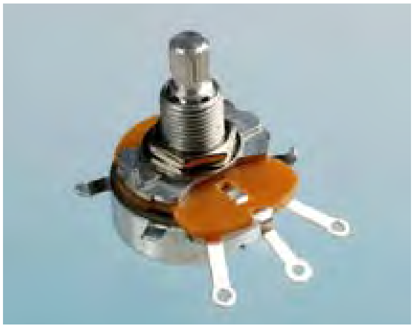

Figure 1-47

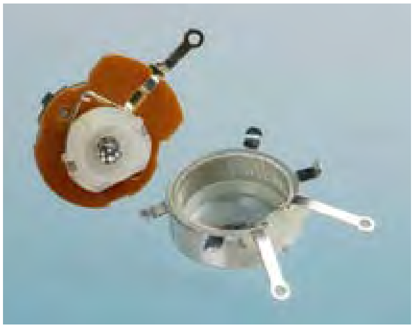

Figure 1-48. *To open the potentiometer, first pry up the four little metal tabs around the edge (you can see one sticking out at the left and another one sticking out at the right in Figure 1-47). Inside is a coil of wire around a flat plastic band, and a pair of springy contacts (the wiper), which conduct electricity to or from any point in the coil when you turn the shaft.*

Depending whether you have a really cheap potentiometer or a slightly more high-class version, you may find a circular track of conductive plastic or a loop of coiled wire. Either way, the principle is the same. The wire or the plastic possesses some resistance (a total of 2K in this instance), and as you turn the shaft of the potentiometer, a wiper rubs against the resistance, giving you a shortcut to any point from the center terminal.

You can try to put it back together, but if it doesn't work, use your backup potentiometer instead.

To test your potentiometer, set your meter to measure resistance (ohms) and touch the probes while turning the potentiometer shaft to and fro, as shown in Figure 1-49.

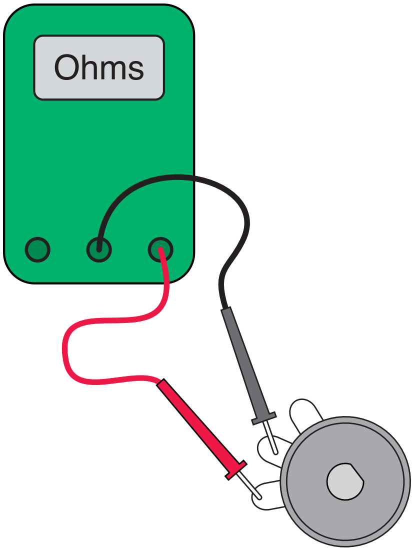

Figure 1-49. *Measure the resistance between these two terminals of the potentiometer while you turn its shaft to and fro.*

Dimming Your LED

Begin with the potentiometer turned all the way counterclockwise, otherwise you'll burn out the LED before we even get started. (A very, very small number of potentiometers increase and decrease resistance in the opposite way to which I'm describing here, but as long as your potentiometer looks like the one in Figure 1-48 after you open it up, my description should be accurate.)

Now connect everything as shown in Figures 1-50 and 1-51, taking care that you don't allow the metal parts of any of the alligator clips to touch each other. Now turn up the potentiometer *very* slowly. You'll notice the LED glowing brighter, and brighter, and brighter—until, oops, it goes dark. You see how easy it is to destroy modern electronics? Throw away that LED. It will never glow again. Substitute a new LED, and we'll be more careful this time.

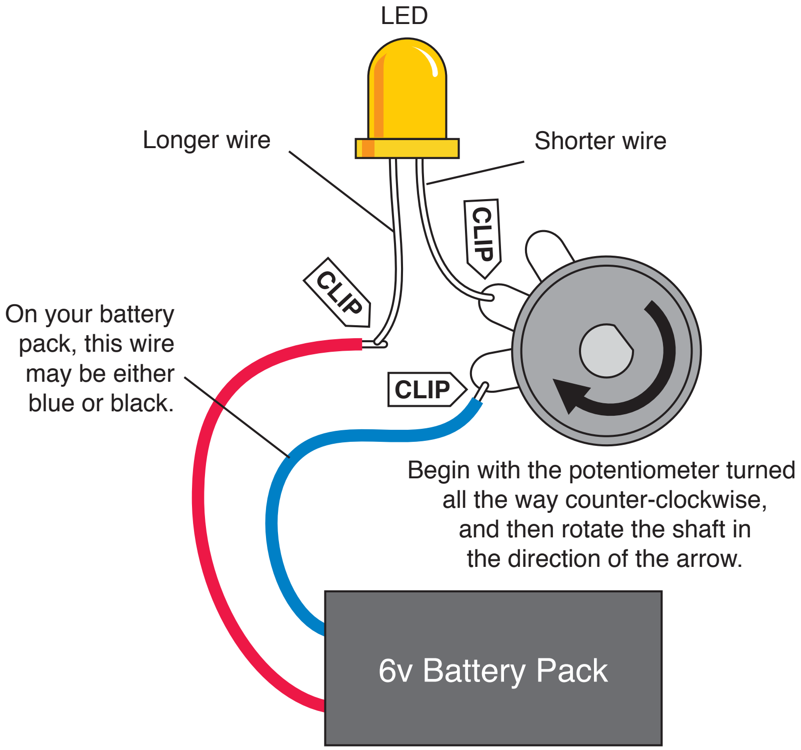
LED
Longer wire
Shorter wire
CLIP
CLIP
CLIP
On your battery pack, this wire may be either blue or black.
Begin with the potentiometer turned all the way counter-clockwise, and then rotate the shaft in the direction of the arrow.
6v Battery Pack

Figure 1-50. *The setup for Experiment 4. Rotating the shaft of the 2K potentiometer varies its resistance from 0 to 2,000Ω. This resistance protects the LED from the full 6 volts of the battery.*

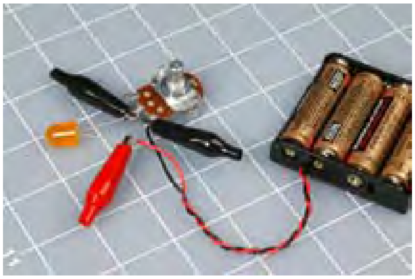

Figure 1-51. *The LED in this photo is dark because I turned the potentiometer up just a little bit too far.*

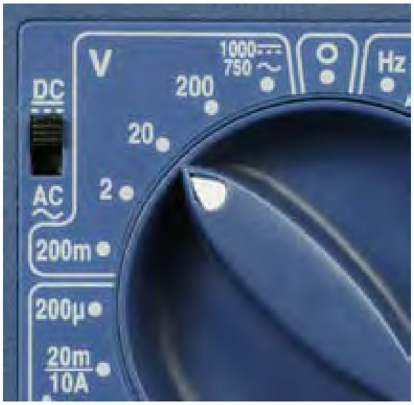

Figure 1-52

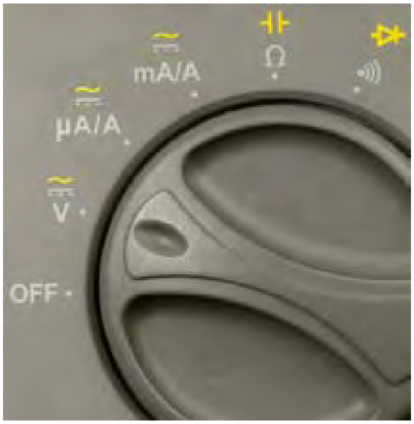

Figure 1-53

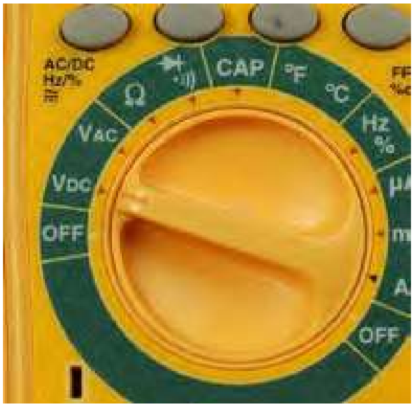

Figure 1-54. *Each meter has a different way to measure volts DC. The manually adjusted meter (top) requires you to move a slider switch to "DC" and then choose the highest voltage you want to measure: In this case, the selected voltage is 20 (because 2 would be too low). Using the autoranging RadioShack meter, you set it to "V" and the meter will figure out which range to use.*

While the batteries are connected to the circuit, set your meter to measure *volts DC* as shown in Figures 1-52 through 1-54. Now touch the probes either side of the LED. Try to hold the probes in place while you turn the potentiometer up a little, and down a little. You should see the voltage pressure around the LED changing accordingly. We call this the *potential difference* between the two wires of the LED.

If you were using a miniature old-fashioned lightbulb instead of an LED, you'd see the potential difference varying much more, because a lightbulb behaves like a "pure" resistor, whereas an LED self-adjusts to some extent, modifying its resistance as the voltage pressure changes.

Now touch the probes to the two terminals of the potentiometer that we're using, so that you can measure the potential difference between them. The potentiometer and the LED share the total available voltage, so when the potential difference (the voltage drop) around the potentiometer goes up, the potential difference around the LED goes down, and vice versa. See Figures 1-55 through 1-57. A few things to keep in mind:

- If you add the voltage drops across the devices in the circuit, the total is the same as the voltage supplied by the batteries.
- You measure voltage relatively, between two points in a circuit.
- Apply your meter like a stethoscope, without disturbing or breaking the connections in the circuit.

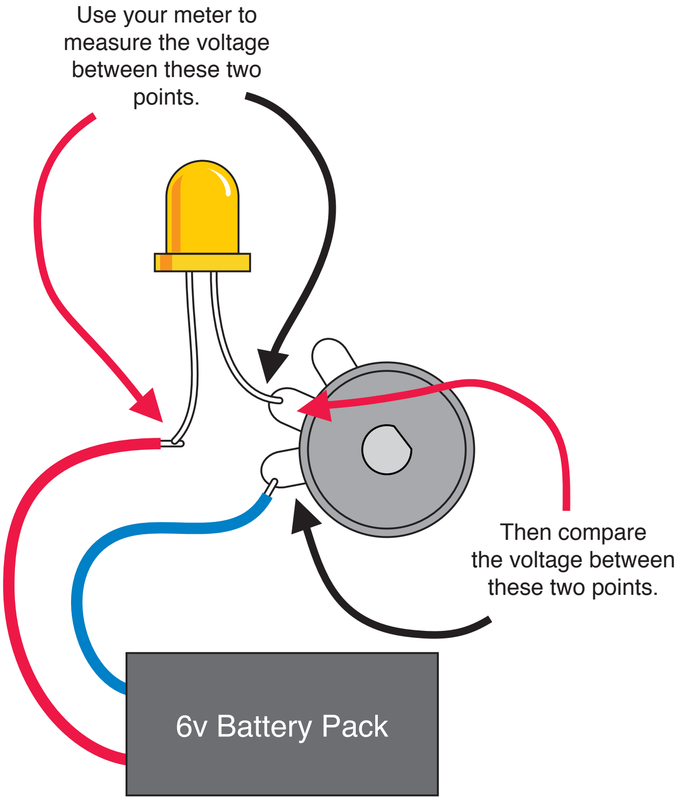
Use your meter to measure the voltage between these two points.

Then compare the voltage between these two points.

Figure 1-55. *How to measure voltage in a simple circuit.*

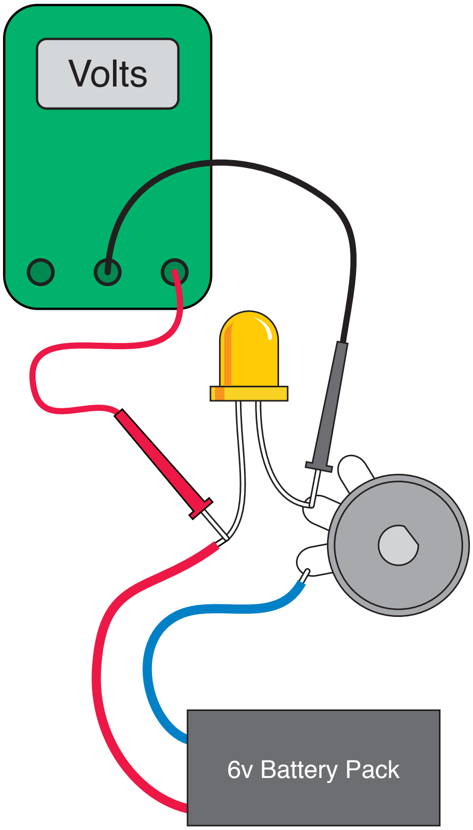

Figure 1-56. *The meter shows how much voltage the LED takes.*

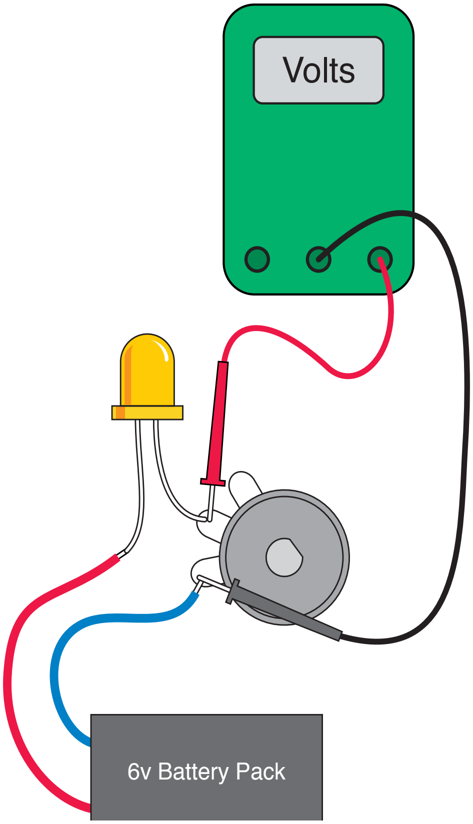

Figure 1-57. *The meter shows how much voltage the potentiometer takes.*

Checking the Flow

Now I want you to make a different measurement. I want you to measure the flow, or current, in the circuit, using your meter set to mA (milliamps). Remember, to measure current:

- You can only measure current when it passes *through* the meter.
- You have to insert your meter into the circuit.
- Too much current will blow the fuse inside your meter.

Make sure you set your meter to measure mA, not volts, before you try this. Some meters require you to move one of your leads to a different socket on the meter, to measure mA. See Figures 1-58 through 1-61.

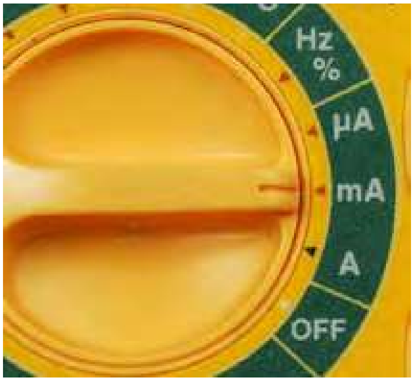

Figure 1-58. *Any meter will blow its internal fuse if you try to make it measure too high an amperage. In our circuit, this is not a risk as long as you keep the potentiometer in the middle of its range. Choose "mA" for milliamps and remember that the meter displays numbers that mean thousandths of an amp.*

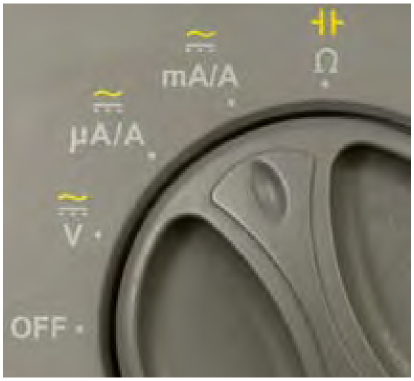

Figure 1-59

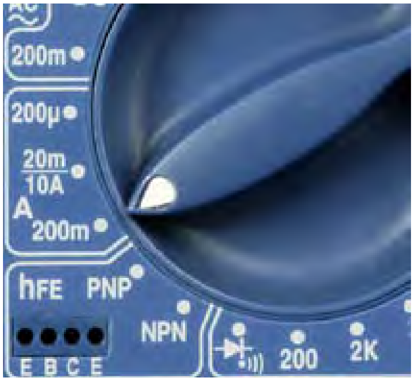

Figure 1-60

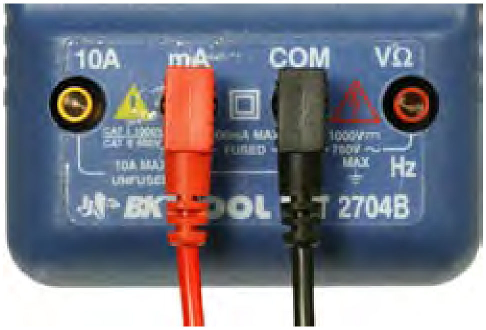

Figure 1-61. *A manual meter such as the one here may require you to shift the red lead to a different socket, to measure milliamps. Most modern meters don't require this until you are measuring higher currents.*

Insert your meter into the circuit, as shown in Figure 1-62. Don't turn the potentiometer more than halfway up. The resistance in the potentiometer will protect your meter, as well as the LED. If the meter gets too much current, you'll find yourself replacing its internal fuse.

As you adjust the potentiometer up and down a little, you should find that the varying resistance in the circuit changes the flow of current—the amperage. This is why the LED burned out in the previous experiment: too much current made it hot, and the heat melts it inside, just like the fuse in the previous experiment. *A higher resistance limits the flow of current, or amperage*.

Now insert the meter in another part of the circuit, as shown in Figure 1-63. As you turn the potentiometer up and down, you should get exactly the same results as with the configuration in Figure 1-64. This is because the current is the same at all points in a similar circuit. *It has to be, because the flow of electrons has no place else to go*.

It's time now to nail this down with some numbers. Here's one last thing to try. Set aside the LED and substitute a 1KΩ resistor, as shown in Figure 1-64. The total resistance in the circuit is now 1KΩ plus whatever the resistance the potentiometer provides, depending how you set it. (The meter also has some resistance, but it's so low, we can ignore it.)

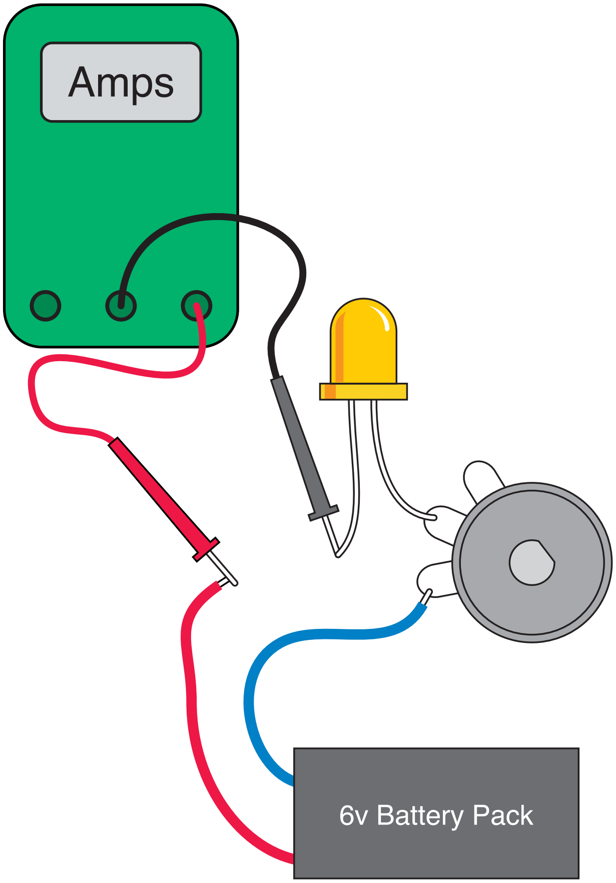

Figure 1-62. *To measure amps, as illustrated here and in Figure 1-63, the current has to pass through the meter. When you increase the resistance, you restrict the current flow, and the lower flow makes the LED glow less brightly.*

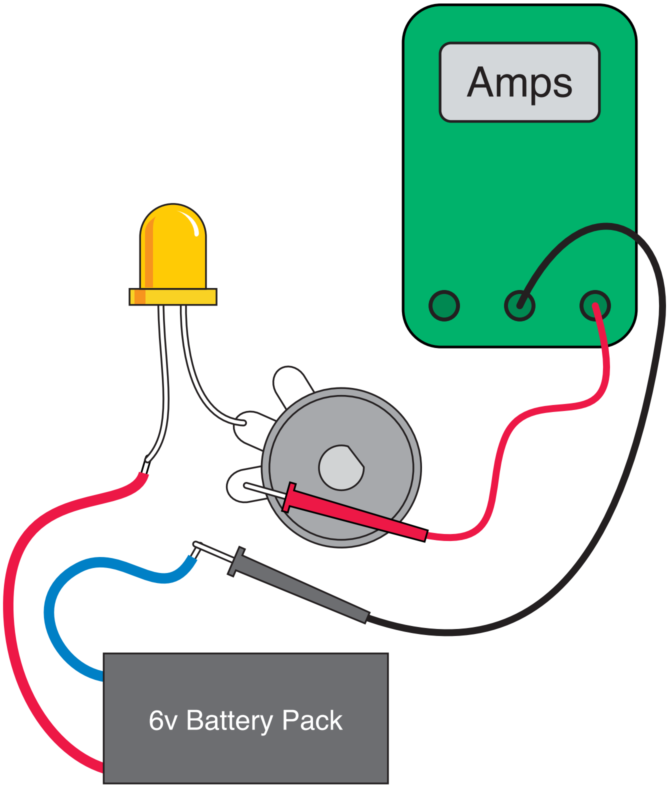

Figure 1-63

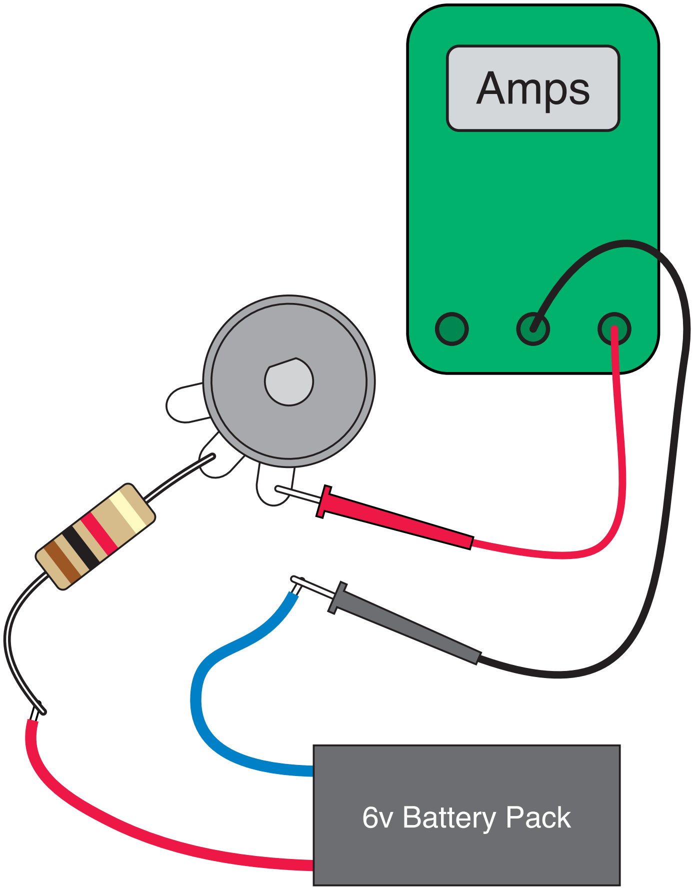

Figure 1-84. *If you substitute a resistor instead of the LED, you can confirm that the current flowing through the circuit varies with the total resistance in the circuit, if the voltage stays the same.*

Turn the potentiometer all the way counterclockwise, and you have a total of 3K resistance in the circuit. Your meter should show about 2 mA flowing. Now turn the potentiometer halfway, and you have about 2K total resistance. You should see about 3 mA flowing. Turn the potentiometer all the way clockwise, so there's a total of 1K, and you should see 6 mA flowing. You may notice that if we multiply the resistance by the amperage, we get 6 each time—which just happens to be the voltage being applied to the circuit. See the following table.

Total resistance	**Current**	**Voltage**
(KΩ)	(mA)	(Volts)
3	2	6
2	3	6
1	6	6

In fact, we could say:

voltage = kilohms × milliamps

But wait a minute: 1K is 1,000 ohms, and 1mA is 1/1,000 of an amp. Therefore, our formula should really look like this:

voltage = (ohms × 1,000) × (amps/1,000)

The two factors of 1,000 cancel out, so we get this:

volts = ohms × amps

This is known as Ohm's Law. See the section, "Fundamentals: Ohm's Law," on the following page.

FUNDAMENTALS

Series and parallel

Before we go any further, you should know how resistance in a circuit increases when you put resistors in series or in parallel. Figures 1-65 through 1-67 illustrate this. Remember:

- Resistors in series are oriented so that one follows the other.
- Resistors in parallel are oriented side by side.

When you put two equal-valued resistors in series, you double the total resistance, because electricity has to pass through two barriers in succession.

When you put two equal-valued resistors in parallel, you divide the total resistance by two, because you're giving the electricity two paths which it can take, instead of one.

In reality we don't normally need to put resistors in parallel, but we often put other types of components in parallel. Lightbulbs in your house, for instance, are all wired that way. So, it's useful to understand that resistance in a circuit goes down if you keep adding components in parallel.

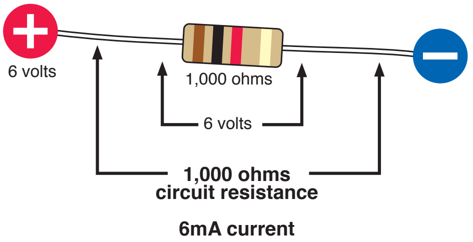

6 volts
1,000 ohms
6 volts
1,000 ohms
circuit resistance
6mA current

Figure 1-65. One resistor takes the entire voltage, and according to Ohm's Law, it draws v/R = 6/1,000 = 0.006 amps = 6mA of current.

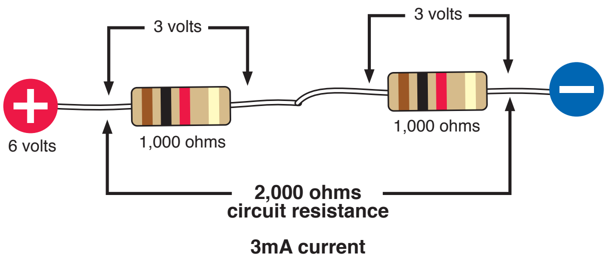

3 volts
3 volts
6 volts
1,000 ohms
1,000 ohms
2,000 ohms
circuit resistance
3mA current

Figure 1-66. When two resistors are in series, the electricity has to pass through one to reach the other, and therefore each of them takes half the voltage. Total resistance is now 2,000 ohms, and according to Ohm's Law, the circuit draws v/R = 6/2,000 = 0.003 amps = 3mA of current.

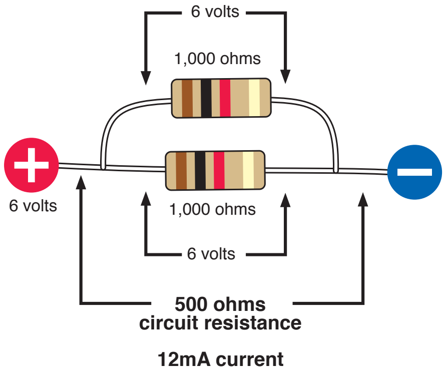

6 volts
1,000 ohms
6 volts
1,000 ohms
6 volts
500 ohms
circuit resistance
12mA current

Figure 1-67. When two resistors are in parallel, each is exposed to the full voltage, so each of them takes 6 volts. The electricity can now flow through both at once, so the total resistance of the circuit is half as much as before. According to Ohm's Law, the circuit draws v/R = 6/500 = 0.012 amps = 12mA of current.

FUNDAMENTALS

Ohm's Law

For reasons I'll explain in a moment, amps are normally abbreviated with the letter I. V stands for volts and R stands for resistance in ohms (because the omega symbol, Ω, is not easily generated from most keyboards). Using these symbols, you can write Ohm's Law in three different ways:

$V = I \times R$
$I = V/R$
$R = V/I$

Remember, V is a *difference* in voltage between two points in a simple circuit, R is the resistance in ohms *between* the same two points, and I is the current in amps flowing *through* the circuit between the two points.

Letter I is used because originally current was measured by its *inductance*, meaning the ability to induce magnetic effects. It would be much less confusing to use A for amps, but unfortunately it's too late for that to happen.

Using Ohm's Law

Ohm's Law is extremely useful. For example, it helps us to figure out whether a component can be used safely in a circuit. Instead of stressing the component until we burn it out, we can predict whether it will work.

For instance, the first time you turned the potentiometer, you didn't really know how far you could go until the LED burned out. Wouldn't it be useful to know precisely what resistance to put in series with an LED, to protect it adequately while providing as much light as possible?

How to Read a Data Sheet

Like most information, the answer to this question is available online.

Here's how you find a manufacturer's data sheet (Figure 1-68). First, find the component that you're interested in from a mail-order source. Next, Google the part number and manufacturer's name. Usually the data sheet will pop up as the first hit. A source such as *Mouser.com* makes it even easier by giving you a direct link to manufacturers' data sheets for many products.

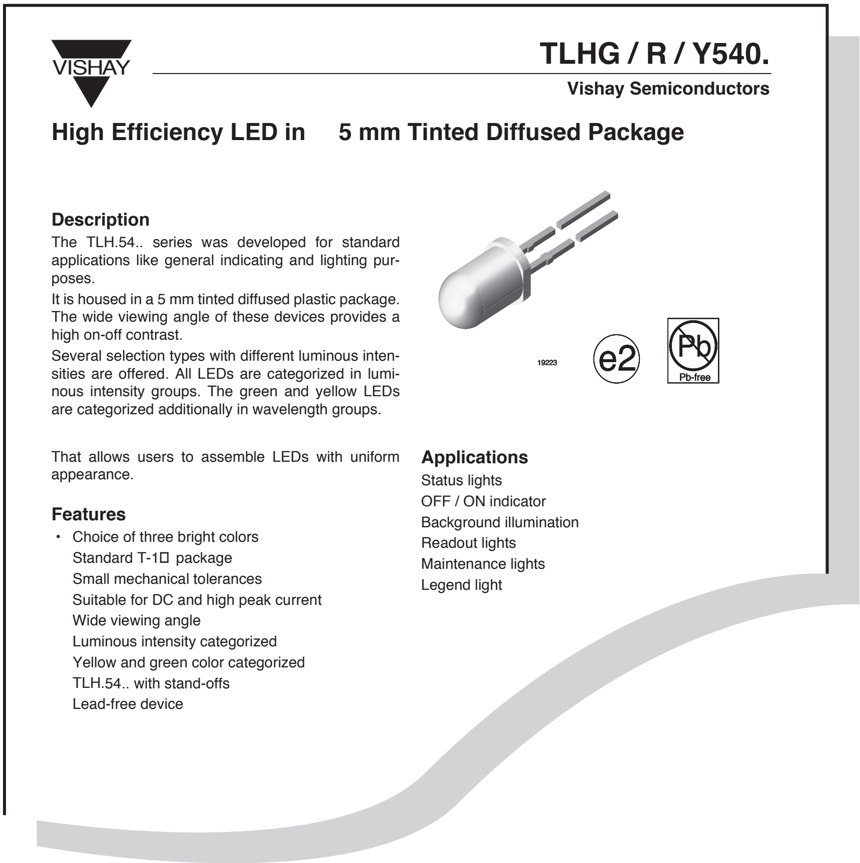

VISHAY

TLHG / R / Y540.

Vishay Semiconductors

High Efficiency LED in 5 mm Tinted Diffused Package

Description

The TLH.54.. series was developed for standard applications like general indicating and lighting purposes.

It is housed in a 5 mm tinted diffused plastic package. The wide viewing angle of these devices provides a high on-off contrast.

Several selection types with different luminous intensities are offered. All LEDs are categorized in luminous intensity groups. The green and yellow LEDs are categorized additionally in wavelength groups.

That allows users to assemble LEDs with uniform appearance.

Features

- Choice of three bright colors
  Standard T-1□ package
  Small mechanical tolerances
  Suitable for DC and high peak current
  Wide viewing angle
  Luminous intensity categorized
  Yellow and green color categorized
  TLH.54.. with stand-offs
  Lead-free device

Applications

Status lights
OFF / ON indicator
Background illumination
Readout lights
Maintenance lights
Legend light

Figure 1-68. *The beginning of a typical data sheet, which includes all relevant specifications for the product, freely available online.*

BACKGROUND

How much voltage does a wire consume?

Normally, we can ignore the resistance in electric wires, such as the little leads of wire that stick out of resistors, because it's trivial. However, if you try to force large amounts of current through long lengths of thin wire, the resistance of the wire can become important.

How important? Once again, we can use Ohm's Law to find out.

Suppose that a very long piece of wire has a resistance of 0.2Ω. And we want to run 15 amps through it. How much voltage will the wire steal from the circuit, because of its resistance?

Once again, you begin by writing down what you know:

$R = 0.2$

$I = 15$

We want to know V, the potential difference, for the wire, so we use the version of Ohm's Law that places V on the left side:

$V = I \times R$

Now plug in the values:

$V = 15 \times 0.2 = 3$ volts

Three volts is not a big deal if you have a high-voltage power supply, but if you are using a 12-volt car battery, this length of wire will take one-quarter of the available voltage.

Now you know why the wiring in automobiles is relatively thick—to reduce its resistance well below 0.2Ω. See Figure 1-69.

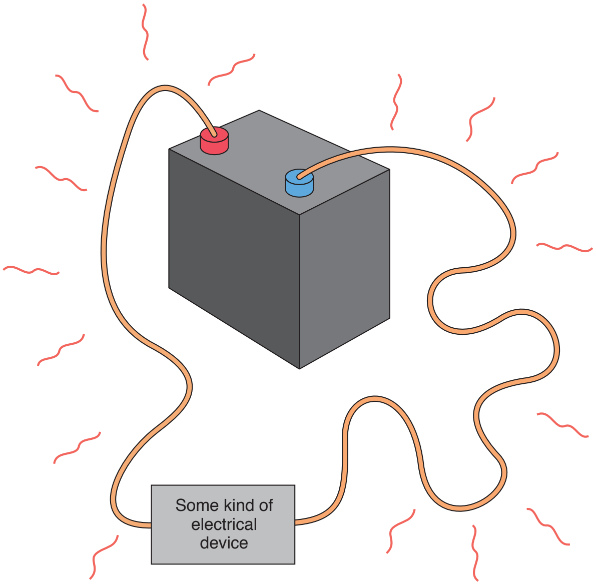
Some kind of electrical device

Figure 1-69. *When a 12-volt car battery runs some kind of electrical device through a long piece of thin wire, the resistance of the wire steals some of the voltage and dissipates it as heat.*

BACKGROUND

The origins of wattage

James Watt (Figure 1-70) is known as the inventor of the steam engine. Born in 1736 in Scotland, he set up a small workshop in the University of Glasgow, where he struggled to perfect an efficient design for using steam to move a piston in a cylinder. Financial problems and the primitive state of the art of metal working delayed practical applications until 1776.

Despite difficulties in obtaining patents (which could only be granted by an act of parliament in those times), Watt and his business partner eventually made a lot of money from his innovations. Although he predated the pioneers in electricity, in 1889 (70 years after his death), his name was assigned to the basic unit of electric power that can be defined by multiplying amperes by volts. See the Fundamentals section, "Watt Basics," on page 31.

Figure 1-70. *James Watt's development of steam power enabled the industrial revolution. After his death, he was honored by having his name applied to the basic unit of power in electricity.*

Here's an example. Suppose I want a red LED, such as the Vishay part TLHR5400, which has become such a common item that I can buy them individually for 9 cents apiece. I click the link to the data sheet maintained by the manufacturer, Vishay Semiconductor. Almost immediately I have a PDF page on my screen. This data sheet is for TLHR, TLHG, and TLHY types of LED, which are red, green, and yellow respectively, as suggested by the R, G, and Y in the product codes. I scroll down and look at the "Optical and Electrical Characteristics" section. It tells me that under conditions of drawing a current of 20 mA, the LED will enjoy a "Typ," meaning, typical, "forward voltage" of 2 volts. The "Max," meaning maximum, is 3 volts.

Let's look at one other data sheet, as not all of them are written the same way. I'll choose a different LED, the Kingbright part WP7113SGC. Click on the link to the manufacturer's site, and I find on the second page of the data sheet a typical forward voltage of 2.2, maximum 2.5, and a maximum forward current of 25 mA. I also find some additional information: a maximum reverse voltage of 5 and maximum reverse current of 10 uA (that's microamps, which are 1,000 times smaller than milliamps). This tells us that you should avoid applying excessive voltage to the LED the wrong way around. If you exceed the reverse voltage, you risk burning out the LED. Always observe polarity!

Kingbright also warns us how much heat the LED can stand: 260° C (500° F) for a few seconds. This is useful information, as we'll be putting aside our alligator clips and using hot molten solder to connect electrical parts in the near future. Because we have already destroyed a battery, a fuse, and an LED in just four experiments, maybe you won't be surprised when I tell you that we will destroy at least a couple more components as we test their limits with a soldering iron.

Anyway, now we know what an LED wants, we can figure out how to supply it. If you have any difficulties dealing with decimals, check the Fundamentals section "Decimals," on the next page, before continuing.

How Big a Resistor Does an LED Need?

Suppose that we're use the Vishay LED. Remember its requirements from the data sheet? Maximum of 3 volts, and a safe current of 20mA.

I'm going to limit it to 2.5 volts, to be on the safe side. We have 6 volts of battery power. Subtract 2.5 from 6 and we get 3.5. So we need a resistor that will take 3.5 volts from the circuit, leaving 2.5 for the LED.

The current flow is the same at all places in a simple circuit. If we want a maximum of 20mA to flow through the LED, the same amount of current will be flowing through the resistor.

Now we can write down what we know about the resistor in the circuit. Note that we have to convert all units to volts, amps, and ohms, so that 20mA should be written as 0.02 amps:

$V = 3.5$ (the potential drop across the resistor)
$I = 0.02$ (the current flowing through the resistor)

We want to know R, the resistance. So, we use the version of Ohm's Law that puts R on the left side:

$R= V/I$

Now plug in the values:

$R = 3.5/0.02$

Run this through your pocket calculator if you find decimals confusing. The answer is:

$R = 175\Omega$

It so happens that 175Ω isn't a standard value. You may have to settle for 180 or 220Ω, but that's close enough.

Evidently the 470Ω resistor that you used in Experiment 3 was a very conservative choice. I suggested it because I said originally that you could use any LED at all. I figured that no matter which one you picked, it should be safe with 470Ω to protect it.

Cleanup and Recycling

The dead LED can be thrown away. Everything else is reusable.

FUNDAMENTALS

Decimals

Legendary British politician Sir Winston Churchill is famous for complaining about "those damned dots." He was referring to decimal points. Because Churchill was Chancellor of the Exchequer at the time, and thus in charge of all government expenditures, his difficulty with decimals was a bit of a problem. Still, he muddled through in time-honored British fashion, and so can you.

You can also use a pocket calculator—or follow two basic rules.

Doing multiplication: combine the zeros

Suppose you need to multiply 0.04 by 0.005:

1. Count the total number of zeros following both of the decimal points. In this case, three zeros.
2. Multiply the numbers which follow the zeros. In this case, $4 \times 5 = 20$.
3. Write down the result as 0 followed by a decimal point, followed by the number of zeros, followed by the multiplication result. Like this: 0.00020, which is the same as 0.0002.

Doing division: cancel the zeros

Suppose you need to divide 0.006 by 0.0002:

1. Shift the decimal points to the right, in both the numbers, by the same number of steps, until both the numbers are greater than 1. In this case, shift the point four steps in each number, so you get 60 divided by 2.
2. Do the division. The result in this case is 30.

THEORY

Doing the math on your tongue

I'm going to go back to the question I asked in the previous experiment: why didn't your tongue get hot?

Now that you know Ohm's Law, you can figure out the answer in numbers. Let's suppose the battery delivered its rated 9 volts, and your tongue had a resistance of 50K, which is 50,000 ohms. Write down what you know:

$V = 9$
$R = 50{,}000$

We want to know the current, I, so we use the version of Ohm's Law that puts this on the left:

$I = V/R$

Plug in the numbers:

$I = 9/50{,}000 = 0.00018$ amps

Move the decimal point three places to convert to milliamps:

$I = 0.18$ mA

That's a tiny current that will not produce much heat at 9 volts.

What about when you shorted out the battery? How much current made the wires get hot? Well, suppose the wires had a resistance of 0.1 ohms (probably it's less, but I'll start with 0.1 as a guess). Write down what we know:

$V = 1.5$
$R = 0.1$

Once again we're trying to find I, the current, so we use:

$I = V/R$

Plug in the numbers:

$I = 1.5/0.1 = 15$ amps

That's 100,000 times the current that may have passed through your tongue, which would have generated much more heat, even though the voltage was lower.

Could that tiny little battery really pump out 15 amps? Remember that the battery got hot, as well as the wire. This tells us that the electrons may have met some resistance inside the battery, as well as in the wire. (Otherwise, where else did the heat come from?) Normally we can forget about the internal resistance of a battery, because it's so low. But at high currents, it becomes a factor.

I was reluctant to short-circuit the battery through a meter, to try to measure the current. My meter will fry if the current is greater than 10A. However I did try putting other fuses into the circuit, to see whether they would blow. When I tried a 10A fuse, it did not melt. Therefore, for the brand of battery I used, I'm fairly sure that the current in the short circuit was under 10A, but I know it was over 3A, because the 3A fuse blew right away.

The internal resistance of the 1.5-volt battery prevented the current in the short circuit from getting too high. This is why I cautioned against using a larger battery (especially a car battery). Larger batteries have a much lower internal resistance, allowing dangerously high currents which generate explosive amounts of heat. A car battery is designed to deliver literally hundreds of amps when it turns a starter motor. That's quite enough current to melt wires and cause nasty burns. In fact, you can weld metal using a car battery.

Lithium batteries also have low internal resistance, making them very dangerous when they're shorted out. High current can be just as dangerous as high voltage.

FUNDAMENTALS

Watt basics

So far I haven't mentioned a unit that everyone is familiar with: watts.

A watt is a unit of work. Engineers have their own definition of work—they say that work is done when a person, an animal, or a machine pushes something to overcome mechanical resistance. Examples would be a steam engine pulling a train on a level track (overcoming friction and air resistance) or a person walking upstairs (overcoming the force of gravity).

When electrons push their way through a circuit, they are overcoming a kind of resistance, and so they are doing work, which can be measured in watts. The definition is easy:

watts = volts × amps

Or, using the symbols customarily assigned, these three formulas all mean the same thing:

$W = V \times I$

$V = W/I$

$I = W/V$

Watts can be preceded with an "m," for "milli," just like volts:

Number of watts	Usually expressed as	Abbreviated as
0.001 watts	1 milliwatt	1mW
0.01 watts	10 milliwatts	10 mW
0.1 watts	100 milliwatts	100 mW
1 watt	1,000 milliwatts	1W

Because power stations, solar installations, and wind farms deal with much larger numbers, you may also see references to kilowatts (using letter K) and megawatts (with a capital M, not to be confused with the lowercase m used to define milliwatts):

Number of watts	Usually expressed as	Abbreviated as
1,000 watts	1 kilowatt	1 KW
1,000,000 watts	1 megawatt	1 MW

Lightbulbs are calibrated in watts. So are stereo systems. The watt is named after James Watt, inventor of the steam engine. Incidentally, watts can be converted to horsepower, and vice versa.

THEORY

Power assessments

I mentioned earlier than resistors are commonly rated as being capable of dealing with 1/4 watt, 1/2 watt, 1 watt, and so on. I suggested that you should buy resistors of 1/4 watt or higher. How did I know this?

Go back to the LED circuit. Remember we wanted the resistor to drop the voltage by 3.5 volts, at a current of 20 mA. How many watts of power would this impose on the resistor?

Write down what you know:

$V = 3.5$ (the voltage drop imposed by the resistor)

$I = 20mA = 0.02$ amps (the current flowing through the resistor)

We want to know W, so we use this version of the formula:

$W = V \times I$

Plug in the values:

$W = 3.5 \times 0.02 = 0.07$ watts (the power being dissipated by the resistor)

Because 1/4 watt is 0.25 watts, obviously a 1/4 watt resistor will have about four times the necessary capacity. In fact you could have used a 1/8 watt resistor, but in future experiments we may need resistors that can handle 1/4 watt, and there's no penalty for using a resistor that is rated for more watts than will actually pass through it.

Experiment 5: Let's Make a Battery

Long ago, before web surfing, file sharing, or cell phones, kids were so horribly deprived that they tried to amuse themselves with kitchen-table experiments such as making a primitive battery by pushing a nail and a penny into a lemon. Hard to believe, perhaps, but true!

This is seriously old-school—but I want you to try it anyway, because anyone who wants to get a feel for electricity should see how easy it is to extract it from everyday objects around us. Plus, if you use enough lemons, you just *might* generate enough voltage to power an LED.

The basic components of a battery are two metal electrodes immersed in an electrolyte. I won't define these terms here (they're explained in the following section "Theory: The nature of electricity"). Right now all you need to know is that lemon juice will be your electrolyte, and copper and zinc will be your electrodes. A penny provides the necessary copper, as long as it is fairly new and shiny. Pennies aren't solid copper anymore, but they are still copper-plated, which is good enough.

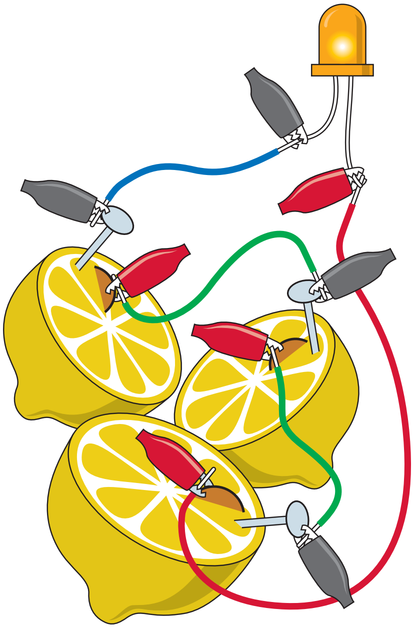

Figure 1-71. *A three-lemon battery. Don't be too disappointed if the LED fails to light up. The lemons have a high electrical resistance, so they can't deliver much current, especially through the relatively small surface area of the nails and the pennies. However, the lemon battery does generate voltage that you can measure with your meter.*

To find some metallic zinc, you will have to make a trip to a hardware store, where you should ask for roofing nails. The nails are zinc-plated to prevent them from rusting. Small metal brackets or mending plates also are usually zinc-plated. They should have a slightly dull, silvery look. If they have a mirror-bright finish, they're more likely to be nickel-plated.

Cut a lemon in half, set your multimeter so that it can measure up to 2 volts DC, and hold one probe against a penny while you hold the other probe against a roofing nail (or other zinc-plated object). Now force the penny and the nail into the exposed juicy interior of the lemon, as close to each other as possible, but not actually touching. You should find that your meter detects between 0.8 volts and 1 volt.

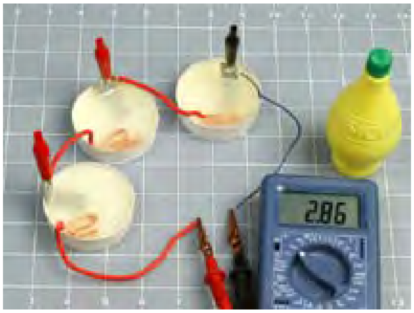

Figure 1-72. *Bottled lemon juice seems to work just as well as fresh lemon juice. I cut the bottoms off three paper cups, inserted a galvanized bracket into each, and used heavyweight stranded copper wire to make the positive electrodes*

You can experiment with different items and liquids to see which works best. Immersing your nail and penny in lemon juice that you have squeezed into a shot glass or egg cup may enhance the efficiency of your battery, although you'll have a harder time holding everything in place. Grapefruit juice and vinegar will work as substitutes for lemon juice.

To drive a typical LED, you need more than 1 volt. How to generate the extra electrical pressure? By putting batteries in series, of course. In other words, more lemons! (Or more shot glasses or egg cups.) You'll also need lengths of wire to connect multiple electrodes, and this may entail skipping ahead to Chapter 2, where I describe how to strip insulation from hookup wire. Figures 1-71 and 1-72 show the configuration.

If you set things up carefully, making sure than none of the electrodes are touching, you may be able to illuminate your LED with two or three lemon-juice batteries in series. (Some LEDs are more sensitive to very low currents than others. Later in the book I'll be talking about very-low-current LEDs. If you want your lemon-juice battery to have the best chance of working, you can search online for low-current LEDs and buy a couple.)

THEORY

The nature of electricity

To understand electricity, you have to start with some basic information about atoms. Each atom consists of a nucleus at the center, containing protons, which have a positive charge. The nucleus is surrounded by electrons, which carry a negative charge.

Breaking up the nucleus of an atom requires a lot of energy, and can also liberate a lot of energy—as happens in a nuclear explosion. But persuading a couple of electrons to leave an atom (or join an atom) takes very little energy. For instance, when zinc reacts chemically with an acid, it can liberate electrons. This is what happens at the zinc electrode of the chemical battery in Experiment 5.

The reaction soon stops, as electrons accumulate on the zinc electrode. They feel a mutual force of repulsion, yet they have nowhere to go. You can imagine them like a crowd of hostile people, each one wanting the others to leave, and refusing to allow new ones to join them, as shown in Figure 1-73.

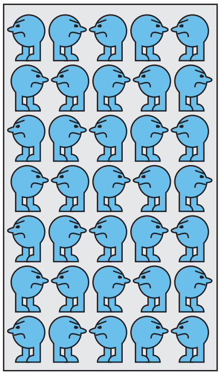

Figure 1-73. *Electrons on an electrode have a bad attitude known as mutual repulsion.*

Now consider what happens when a wire connects the zinc electrode, which has a surplus of electronics, to another electrode, made from a different material, that has a shortage of electrons. The electrons can pass through the wire very easily by jumping from one atom to the next, so they escape from the zinc electrode and run through the wire, propelled by their great desire to get away from each other. See Figure 1-74. This mutual force of propulsion is what creates an electrical current.

Now that the population of electrons on the zinc electrode has been reduced, the zinc-acid reaction can continue, replacing the missing electrodes with new ones—which promptly imitate their predecessors and try to get away from each other by running away down the wire. The process continues until the zinc-acid reaction grinds to a halt, usually because it creates a layer of a compound such as zinc oxide, which won't react with acid and prevents the acid from reacting with the zinc underneath. (This is why your zinc electrode may have looked sooty when you pulled it out of the acidic electrolyte.)

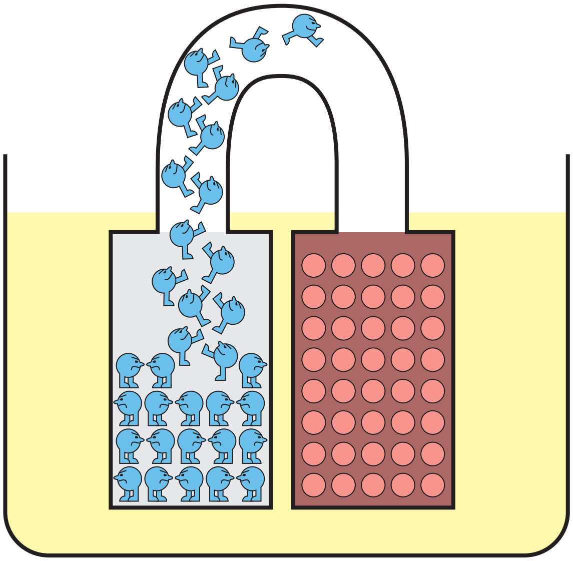

Figure 1-74. *As soon as we open up a pathway from a zinc electrode crowded with electrons to a copper electrode, which contains "holes" for the electrons, their mutual repulsion makes them try to escape from each other to their new home as quickly as possible.*

This description applies to a "primary battery," meaning one that is ready to generate electricity as soon as a connection between its terminals allows electrons to transfer from one electrode to the other. The amount of current that a primary battery can generate is determined by the speed at which chemical reactions inside the battery can liberate electrons. When the raw metal in the electrodes has all been used up in chemical reactions, the battery can't generate any more electricity and is dead. It cannot easily be recharged, because the chemical reactions are not easily reversible, and the electrodes may have oxidized.

In a rechargeable battery, also known as a secondary battery, a smarter choice of electrodes and electrolyte does allow the chemical reactions to be reversed.

How much current is being generated in your lemon battery? Set your meter to measure milliamps, and connect it between the nail and the penny. I measured about 2mA, but got 10mA when I used some #10 stranded copper wire instead of a penny and a large mending plate instead of a roofing nail, immersed in a cup of grapefruit juice. When a larger surface area of metal makes better contact with the electrolyte, you get a greater flow of current. (Don't ever connect your meter to measure amps directly between the terminals of a real battery. The current will be too high, and can blow the fuse inside your meter.)

What's the internal resistance of your lemon? Put aside the copper and zinc electrodes and insert your nickel-plated meter probes into the juice. I got a reading of around 30K when both probes were in the same segment of the lemon, but 40K or higher if the probes were in different segments. Is the resistance lower when you test liquid in a cup?

Here are a couple more questions that you may wish to investigate. For how long will your lemon battery generate electricity? And why do you think your zinc-plated electrode becomes discolored after it has been used for a while?

Electricity is generated in a battery by an exchange of ions, or free electrons, between metals. If you want to know more about this, check the section "Theory: The nature of electricity" on the previous page.

Cleanup and Recycling

The hardware that you immersed in lemons or lemon juice may be discolored, but it is reusable. Whether you eat the lemons is up to you.

BACKGROUND

Positive and negative

If electricity is a flow of electrons, which have a negative charge, why do people talk as if electricity flows from the positive terminal to the negative terminal of a battery?

The answer lies in a fundamental embarrassment in the history of research into electricity. For various reasons, when Benjamin Franklin was trying to understand the nature of electric current by studying phenomena such as lightning during thunderstorms, he believed he observed a flow of "electrical fluid" from positive to negative. He proposed this concept in 1747.

In fact, Franklin had made an unfortunate error that remained uncorrected until after physicist J. J. Thomson announced his discovery of the electron in 1897, 150 years later. Electricity actually flows from an area of greater negative charge, to some other location that is "less negative"—that is, "more positive." In other words, electricity is a flow of negatively charged particles. In a battery, they originate from the negative terminal and flow to the positive terminal.

You might think that when this fact was established, everyone should have discarded Franklin's idea of a flow from positive to negative. But when an electron moves through a wire, you can still think of an equal positive charge flowing in the opposite direction. When the electron leaves home, it takes a small negative charge with it; therefore, its home becomes a bit more positive. When the electron arrives at its destination, its negative charge makes the destination a bit less positive. This is pretty much what would happen if an imaginary positive particle traveled in the opposite direction. Moreover, all of the mathematics describing electrical behavior are still valid if you apply them to the imaginary flow of positive charges.

As a matter of tradition and convenience we still retain Ben Franklin's erroneous concept of flow from positive to negative, because it really makes no difference. In the symbols that represent components such as diodes and transistors, you will actually find arrows reminding you which way these components should be placed—and the arrows all point from positive to negative, even though that's not the way things really work at all! Ben Franklin would have been surprised to learn that although most lightning strikes occur when a negative charge in clouds discharges to neutralize a positive charge on the ground, some forms of lightning are actually a flow of electrons from the negatively charged surface of the earth, up to a positive charge in the clouds. That's right: someone who is "struck by lightning" may be hurt by *emitting* electrons rather than by receiving them, as shown in Figure 1-75.

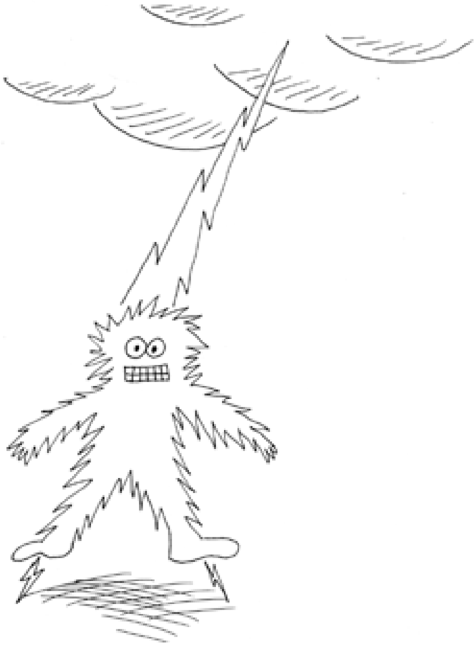

Figure 1-75. *In some weather conditions, the flow of electrons during a lightning strike can be from the ground, through your feet, out of the top of your head, and up to the clouds. Benjamin Franklin would have been surprised.*

THEORY

Basic measurements

Electrical potential is measured by adding up the charges on individual electrons. The basic unit is the *coulomb*, equal to the total charge on about 6,250,000,000,000,000,000 electrons.

If you know how many electrons pass through a piece of wire each second, this establishes the flow of electricity, which can be expressed in amperes. In fact 1 ampere can be defined as 1 coulomb per second. Thus:

1 ampere = 1 coulomb/second
= about 6.25 quintillion electrons/second

There's no way to "see" the number of electrons running through a conductor (Figure 1-76), but there are indirect ways of getting at this information. For instance, when an electron goes running through a wire, it creates a wave of electromagnetic force around it. This force can be measured, and we can calculate the amperage from that. The electric meter installed at your home by the utility company functions on this principle.

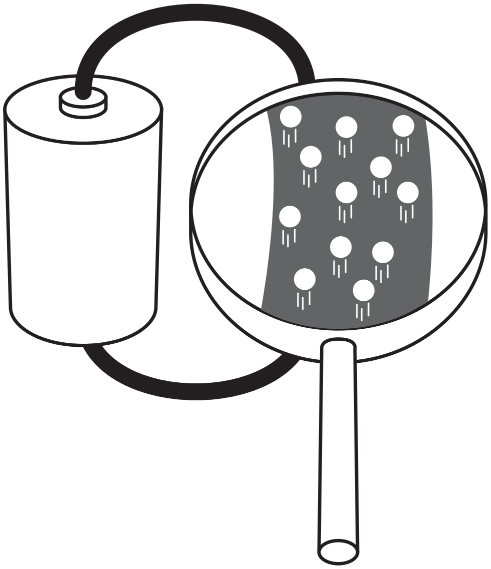

Figure 1-76. *If you could look inside an electric wire with a sufficiently powerful magnifying device, and the wire happened to be carrying 1 ampere of electron flow at the time, you might hope to see about 6.25 quintillion electrons speeding past each second.*

If electrons are just moving freely, they aren't doing any work. If you had a loop of wire of zero resistance, and you kick-started a flow of electrons somehow, they could just go buzzing around forever. (This is what happens inside a superconductor—almost.)

Under everyday conditions, even a copper wire has some resistance. The force that we need to push electrons through it is known as "voltage," and creates a flow that can create heat, as you saw when you shorted out a battery. (If the wire that you used had zero resistance, the electricity running through it would not have created any heat.) We can use the heat directly, as in an electric stove, or we can use the electrical energy in other ways—to run a motor, for instance. Either way, we are taking energy out of the electrons, to do some work.

One volt can be defined as the amount of pressure that you need to create a flow of 1 ampere, which does 1 watt of work. As previously defined, 1 watt = 1 volt × 1 ampere, but the definition actually originated the other way around:

1 volt = 1 watt/1 ampere

It's more meaningful this way, because a watt can be defined in nonelectrical terms. Just in case you're interested, we can work backward through the units of the metric system like this:

1 watt = 1 joule/second
1 joule = a force of 1 newton acting through 1 meter
1 newton = the force required to accelerate 1 kilogram by 1 meter per second, each second

On this basis, the electrical units can all be anchored with observations of mass, time, and the charge on electrons.

Practically Speaking

For practical purposes, an intuitive understanding of electricity can be more useful than the theory. Personally I like the water analogies that have been used for decades in guides to electricity. Figure 1-77 shows a tall tank half full of water, with a hole punched in it near the bottom. Think of the tank as being like a battery. The height of the water is comparable to voltage. The volume of flow through the hole, per second, is comparable to amperage. The smallness of the hole is comparable to resistance. See Figure 1-79 on the next page.

Where's the wattage in this picture? Suppose we place a little water wheel where it is hit by the flow from the hole. We can attach some machinery to the water wheel. Now the flow is doing some work. (Remember, wattage is a measurement of work.)

Maybe this looks as if we're getting something for nothing, extracting work from the water wheel without putting any energy back into the system. But remember, the water level in the tank is falling. As soon as I include some helpers hauling the waste water back up to the top of the tank (in Figure 1-78), you see that we have to put work in to get work out.

Similarly, a battery may seem to be giving power out without taking anything in, but the chemical reactions inside it are changing pure metals into metallic compounds, and the power we get out of a battery is enabled by this change of state. If it's a rechargeable battery, we have to push power back into it to reverse the chemical reactions.

Going back to the tank of water, suppose we can't get enough power out of it to turn the wheel. One answer could be to add more water. The height of the water will create more force. This would be the same as putting two batteries end to end, positive to negative, in series, to double the voltage. See Figure 1-80. As long as the resistance in the circuit remains the same, greater voltage will create more amperage, because amperage = voltage/resistance.

What if we want to run two wheels instead of one? We can punch a second hole in the tank, and the force (voltage) will be the same at each of them. However, the water level in the tank will drop twice as fast. Really, we'd do better to build a second tank, and here again the analogy with a battery is good. If you wire two batteries side by side, in parallel, you get the same voltage, but for twice as long. The two batteries may also be able to deliver more current than if you just used one. See Figure 1-81.

Summing up:

- Two batteries in series deliver twice the voltage.
- Two batteries in parallel can deliver twice the current.

All right, that's more than enough theory for now. In the next chapter, we'll continue with some experiments that will build on the foundations of knowledge about electricity, to take us gradually toward gadgets that can be fun and useful.

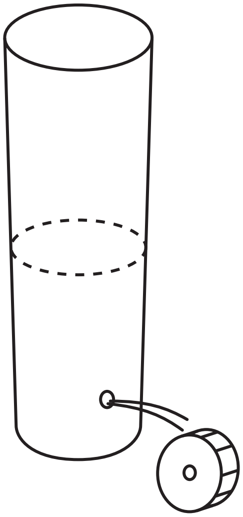

Figure 1-77. *If you want to get work out of a system...*

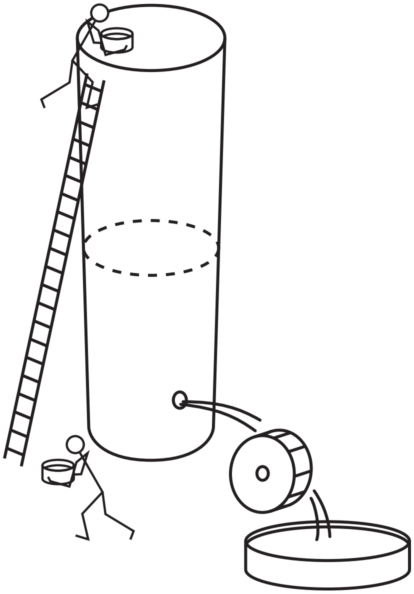

Figure 1-78. *. . . somehow or other you have to put work back into it.*

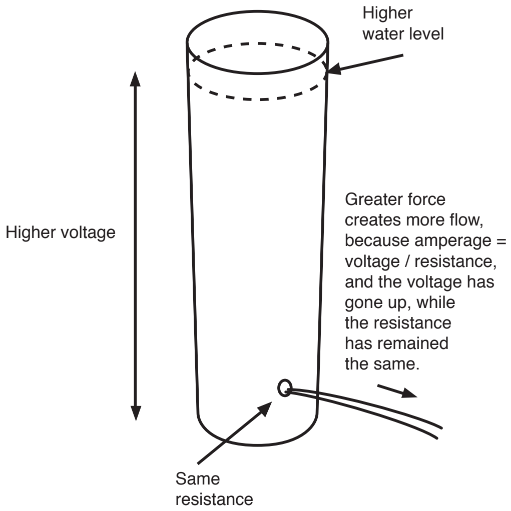
Higher water level
Higher voltage
Greater force creates more flow, because amperage = voltage / resistance, and the voltage has gone up, while the resistance has remained the same.
Same resistance

Figure 1-79. *Greater force generates more flow, as long as the resistance remains the same.*

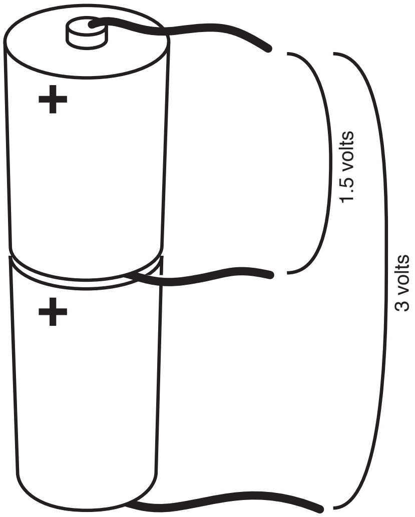
1.5 volts
3 volts

Figure 1-80. *When you place two equal batteries in series, you double the voltage.*

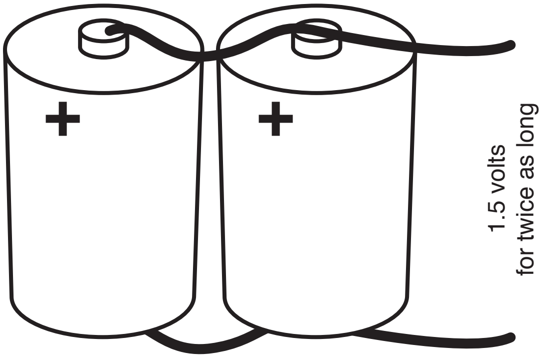
1.5 volts
for twice as long

Figure 1-81. *Two equal batteries that are wired in parallel will deliver the same voltage for twice as long as one.*

Switching Basics and More

2

IN THIS CHAPTER

Shopping List: Experiments 6 Through 11
Experiment 6: Very Simple Switching
Experiment 7: Relay-Driven LEDs
Experiment 8: A Relay Oscillator
Experiment 9: Time and Capacitors
Experiment 10: Transistor Switching
Experiment 11: A Modular Project

The concept of switching is fundamental in electronics, and I'm not just talking about power switches. By "switching," I mean using one flow of electricity to switch, or control, another. This is such an important principle that no digital device can exist without it.

Today, switching is mostly done with semiconductors. Before I deal with them, I'll back up and illustrate the concept by introducing you to relays, which are easier to understand, because you can see what's going on inside them. And before I get to relays, I'll deal with everyday on/off switches, which may seem very simple—but we have to nail down the basics.

Also in this chapter, I'll deal with capacitance, because capacitance and resistance are fundamental to electronic circuits. By the end of the chapter, you should have a basic grounding in electronics and be able to build the noise-making section of a simple intrusion alarm. This will be your first circuit that does something genuinely useful!

Shopping List: Experiments 6 Through 11

As in the previous shopping list, you should visit the various online suppliers for availability and pricing of components and devices. Manufacturers seldom sell small numbers of parts directly. Check the appendix for a complete list of URLs for all the companies mentioned here.

Devices

- Power supply/universal AC adapter, 3 to 12 volts at 1A (1,000 mA). See Figure 2-1. Part number 273-029 from RadioShack, part PH-62092 by Philips, or similar.
- Breadboard suitable for integrated circuits. Quantity: 1. See Figures 2-2 and 2-3. Part 276-002 from RadioShack, model 383-X1000 made by PSP, part 923252-I by 3M, or similar. A breadboard that has screw terminals mounted beside it will be a little easier to use but more expensive than one that does not have terminals.

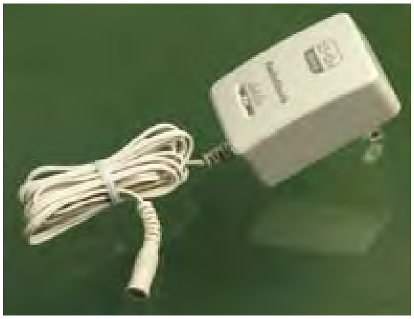

Figure 2-1. *This AC adapter delivers DC from 3 to 12 volts, and is ideal for electronics projects.*

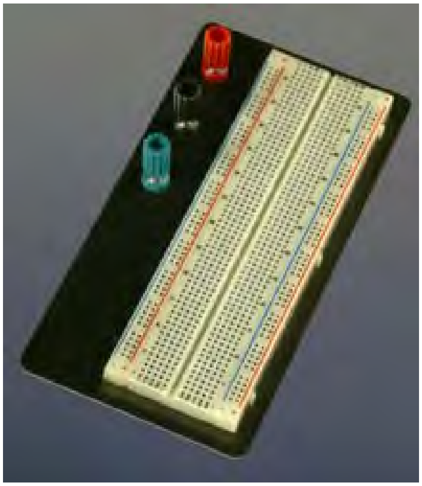

Figure 2-2. *This "breadboard" for quickly constructing electronic circuits has a metal base, and screw terminals for attaching wires from a power supply.*

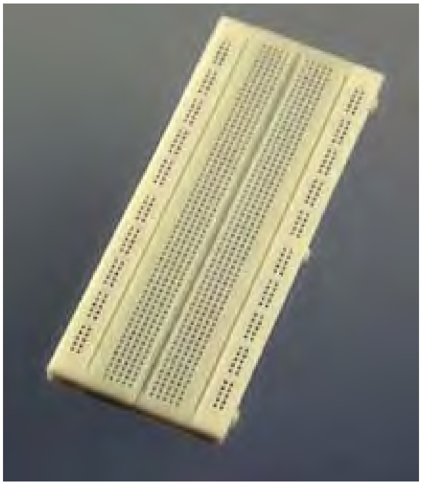

Figure 2-3. *A breadboard without screw terminals is almost as convenient, and is cheaper.*

Tools

Wire strippers

Ideal model 45-121 wire strippers for 16- to 26-gauge wire, or similar. See Figure 2-4. (The "gauge" of the wire tells you how thick it is. A higher gauge means a thinner wire. In this book, we will mainly be using thin wire of 20- to 24-gauge.)

You may also consider the Kronus Automatic Wire Strippers, part 64-2981 from RadioShack, or GB Automatic Wire Strippers, part SE-92 from Amazon.com. See Figure 2-5.

The Kronus and GB wire strippers are functionally identical. The advantage of their design is that it enables you to strip insulation from a wire with one hand. But they do not work well on really thin wire.

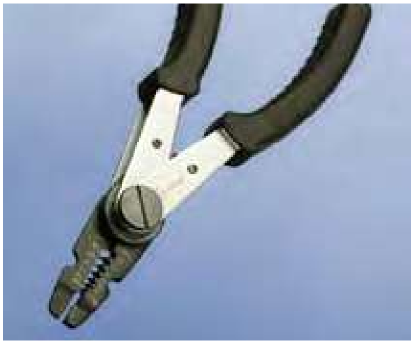

Figure 2-4. *To use these wire strippers, insert a piece of insulated wire in the appropriate-sized hole between the jaws, grip the handles, and pull a section of insulation away. See page 45.*

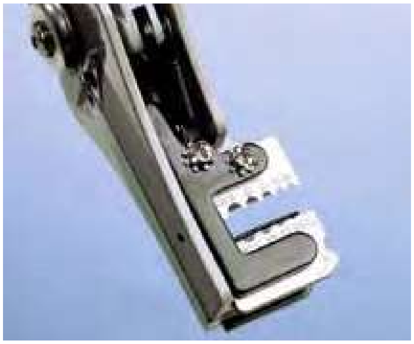

Figure 2-5. *These automatic wire strippers enable one-handed operation, but are not suitable for very small wire diameters. See page 44.*

Supplies

Hookup wire

Solid-conductor, 22-gauge, minimum 25 feet of each color. See Figure 2-6. Part 278-1222 from RadioShack, catalog item 9948T17 from McMaster-Carr, or check eBay for deals.

It's easy to buy the wrong kind of wire. You need solid-core wire, which has a single conductor inside the plastic insulation, not stranded, which has multiple, thinner conductors. See Figures 2-7 and 2-8. You're going to be pushing wires into little holes in a "breadboard," and stranded wire won't let you do this. You will also have problems if you buy wire thicker than 22-gauge. Remember: the lower the gauge number, the thicker the wire.

For a little extra money, you can buy an assortment of precut sections of wire, with ends stripped and ready for use. Try catalog item JW-140 (jumper wire assortment) from All Electronics or search eBay for "breadboard wire." See Figure 2-9.

Patch cords

Patch cords are not strictly necessary but very convenient. You don't want audio or video patch cords, which have a plug on each end; you want wires with alligator clips on each end, also sometimes referred to as "test leads." Try catalog item 461-1176-ND from Digi-Key or catalog item MTL-10 from All Electronics. See Figure 2-10.

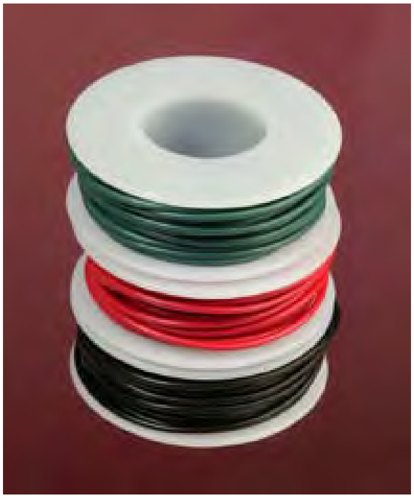

Figure 2-6. *Using hookup wire with different colors of plastic insulation will help you to distinguish one wire from another in your circuits.*

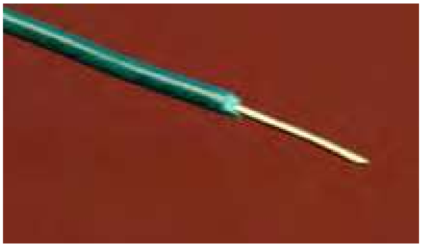

Figure 2-7. *Solid-conductor wire of 22 or 24 gauge is suitable for most of the experiments in this chapter.*

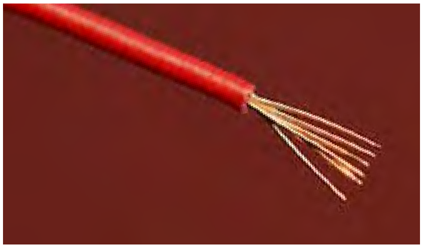

Figure 2-8. *Stranded is more flexible, but cannot be used easily with breadboards.*

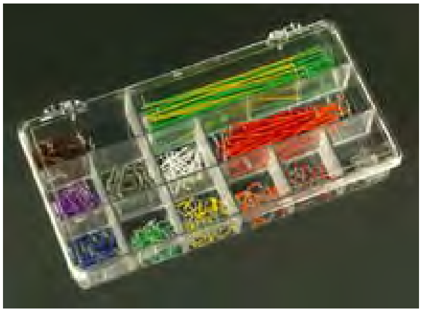

Figure 2-9. *Precut wires with stripped ends can save a lot of time and trouble—if you don't mind paying a little extra.*

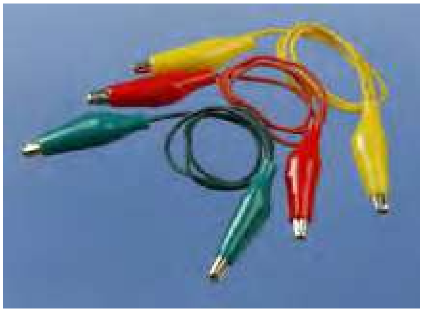

Figure 2-10. *Patch cords, sometimes known as test leads, consist of wires preattached to alligator clips. This is another of those little luxuries that reduces the hassle factor in hobby electronics.*

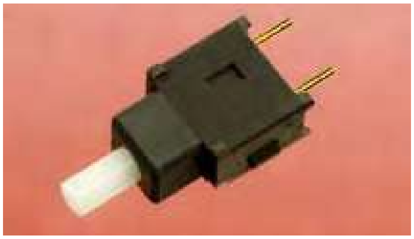

Figure 2-11. *The terminals protruding from this tiny pushbutton are spaced 0.2 inches apart, making it ideal for the "breadboard" that you'll be using.*

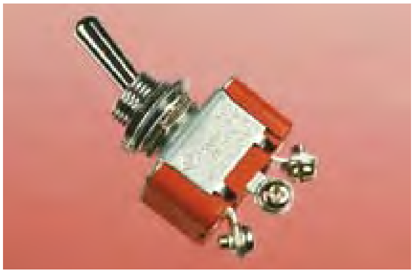

Figure 2-12. *This relatively large toggle switch made by NKK has screw terminals, which will reduce the inconvenience of attaching it to hookup wire.*

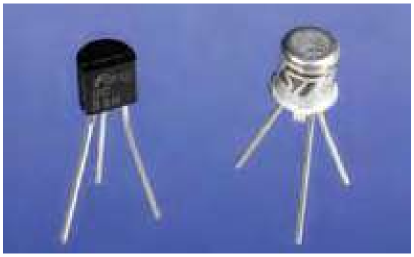

Figure 2-13. *Transistors are commonly sold either in little metal cans or sealed into little lumps of plastic. For our purposes, the packaging makes no difference.*

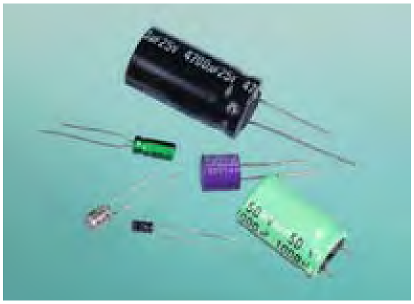

Figure 2-14. *An assortment of electrolytic capacitors.*

Components

Pushbutton

Momentary-on, SPST, sometimes referred to as OFF-(ON) or (ON)-OFF type. Must be PCB- or PC-mount, meaning is extremely small with thin spiky contacts on the bottom. Quantity: 1. See Figure 2-11.

Examples are part number AB11AP by NKK, part MPA103B04 by Alcoswitch, or part EP11SD1CBE by C&K. If you have a choice, buy the cheapest, as we're going to switch very low current.

Switches

Toggle switch, single-pole, double-throw (SPDT), sometimes referred to as ON-ON type. Quantity: 2. See Figure 2-12.

Model S302T-RO by NKK is ideal; it has screw terminals that will eliminate the need for alligator clips. Other options are catalog item MTS-4PC from All Electronics or part 275-603 from RadioShack.

We won't be switching large currents or high voltages, so the exact type of switch is unimportant. However, the terminals on larger-size switches are spaced wider apart, which makes them easier to deal with.

Relays

DPDT, nonlatching, 12v DC. Quantity: 2.

It's important to get the right kind of relay—one whose configuration matches the pictures I'll be using. Look for parts FTR-F1CA012V or FTR-F1CD012V by Fujitsu, G2RL-24-DC12 by Omron, or OMI-SH-212D by Tyco. Avoid substitutions.

Potentiometer

1 megohm linear potentiometer, Part number 271-211 from RadioShack, part number 24N-1M-15R-R from Jameco, or similar.

Transistors

NPN transistor, general-purpose, such as 2N2222 by STMicroelectronics, part PN2222 by Fairchild, or part 2N2222 from RadioShack. Quantity: 4. See Figure 2-13.

2N6027 programmable unijunction transistor manufactured by On Semiconductor or Motorola. Quantity: 4.

Capacitors

Electrolytic capacitors, assorted. Must be rated for a minimum of 25 volts and include at least one capacitor of 1,000 µF (microfarads). If you search on eBay, make sure you find *electrolytic* capacitors. If they're rated for higher voltages, that's OK, although they will be physically larger than you need. See Figure 2-14.

Ceramic capacitors, assorted. Make sure you get at least one rated at 0.0047 µF (which can also be written as 47 nF). See Figure 2-15.

Resistors

If you bought only a minimal selection for experiments 1 through 5, now's the time to buy a larger assortment, so that you won't be stuck needing the one value that you don't have. 1/4-watt minimum.

Loudspeaker

Any 8Ω, 1-inch miniature loudspeaker such as part 273-092 from RadioShack. See Figure 2-16.

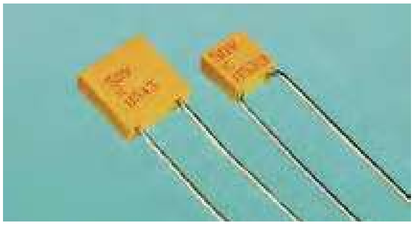

Figure 2-15. *Ceramic capacitors mostly look like this, although many of them are round or bead-shaped instead of square. The packaging shape is unimportant to us.*

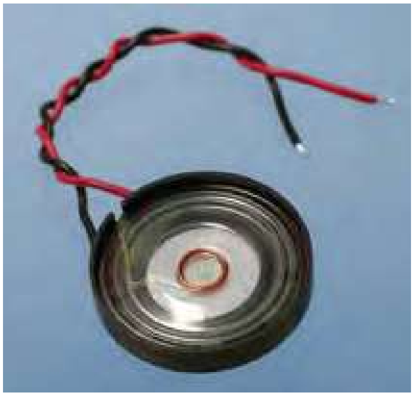

Figure 2-16. *This miniature loudspeaker, just over 1 inch in diameter, is useful for verifying audio output direct from transistor circuits.*

Experiment 6: Very Simple Switching

You will need:

- AA batteries. Quantity: 4.
- Battery carrier for 4 AA batteries. Quantity: 1.
- LED. Quantity: 1.
- Toggle switches, SPDT. Quantity: 2. See Figure 2-12.
- 220Ω or similar value resistor, 1/4-watt minimum. Quantity: 1.
- Alligator clips. Quantity: 8.
- Wire or patch cords. See Figure 2-10, shown previously.
- Wire cutters and wire strippers if you don't use patch cords. See Figure 2-4, shown previously.

In Experiment 3, you illuminated an LED by attaching a battery, and switched it off by removing the battery. For greater convenience our circuits should have proper switches to control power, and while I'm dealing with the general topic of switches, I'm going to explore all the varieties, using a circuit to suggest some possibilities.

Assemble the parts as shown in Figures 2-17 and 2-18. The long lead on the LED must connect with the resistor, because that is the more positive side of the circuit.

You'll notice that you have to include a couple lengths of wire. I suggest green wire to remind you that these sections are not connected directly to positive or to negative power. But you can use any color you like. You can also substitute patch cords, if you have them. However, learning to strip insulation from pieces of wire is a necessary skill, so let's deal with that now.

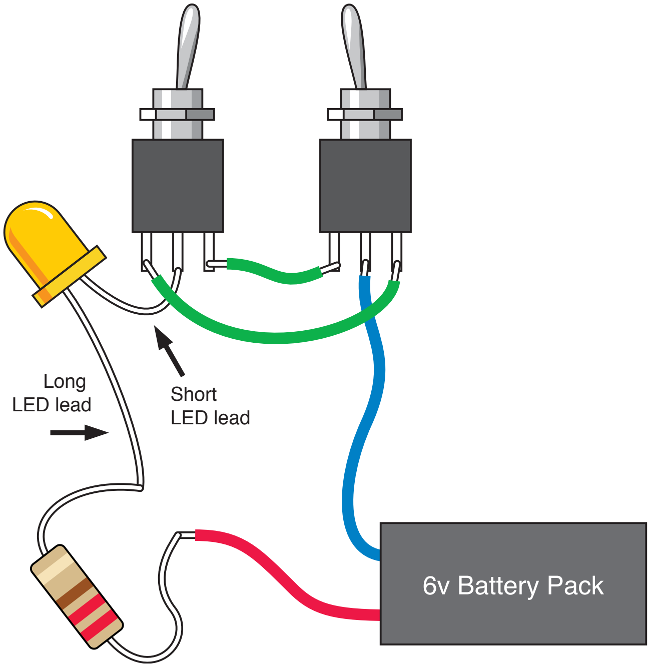

Figure 2-17. *If the LED is on, flipping either of the switches will turn it off. If the LED is off, either of the switches will turn it on. Use alligator clips to attach the wires to each other, and to the switches if your switches don't have screw terminals. Be careful that the clips don't touch each other.*

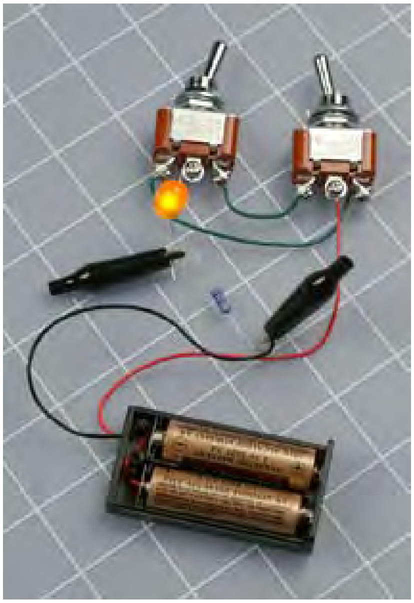

Figure 2-18. *Full-size toggle switches with screw terminals make it easy to hook up this simple circuit.*

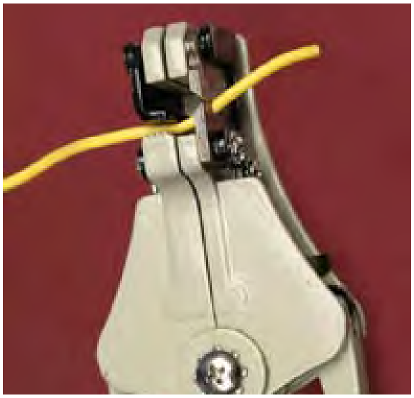

Figure 2-19. *Using automatic wire strippers, when you squeeze the handles the jaw on the left clamps the wire, the sharp grooves on the right bite into the insulation. Squeeze harder and the jaws pull away from each other, stripping the insulation from the wire.*

Tools

If automatic wire strippers (Figure 2-19) don't grip skinny 22-gauge wire very effectively, try the Ideal brand of wire strippers shown back in Figure 2-4, or use plain and simple wire cutters as shown in Figure 2-20. When using wire cutters, you hold the wire in one hand and apply the tool in your other hand, squeezing the handles with moderate pressure—just enough to bite into the insulation, but not so much that you chop the wire. Pull the wire down while you pull the cutters up, and with a little practice you can rip the insulation off to expose the end of the wire.

Macho hardware nerds may use their teeth to strip insulation from wires. When I was younger, I used to do this. I have two slightly chipped teeth to prove it. Really, it's better to use the right tool for the job.

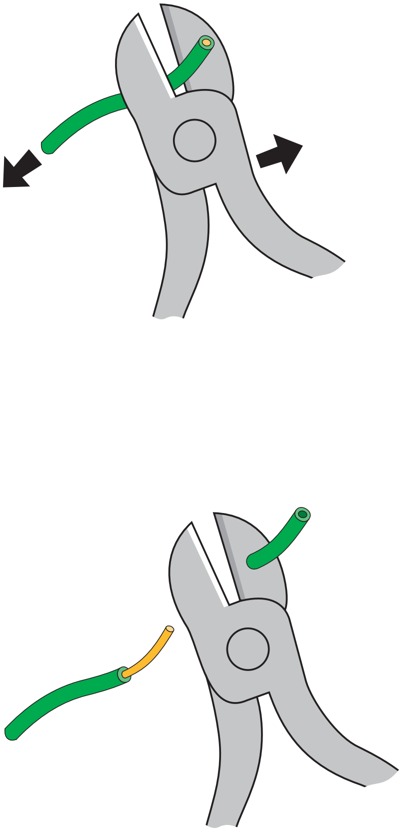

Figure 2-20. *To remove insulation from the end of a thin piece of wire, you can also use wire cutters. This takes a little practice.*

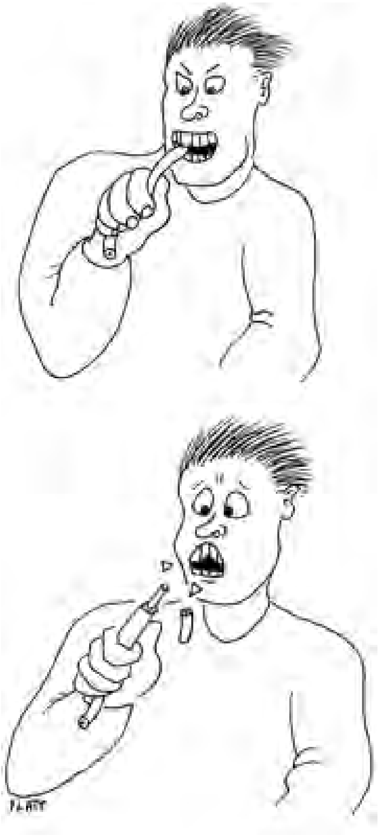

Figure 2-21. *Those who tend to misplace tools, and feel too impatient to search for them, may feel tempted to use their teeth to strip insulation from wire. This may not be such a good idea.*

Connection Problems

Depending on the size of toggle switches that you are using, you may have trouble fitting in all the alligator clips to hold the wires together. Miniature toggle switches, which are more common than the full-sized ones these days, can be especially troublesome (see Figure 2-22). Be patient: fairly soon we'll be using a breadboard, which will eliminate alligator clips almost completely.

Testing

Make sure that you connect the LED with its long wire toward the positive source of power (the resistor, in this case). Now flip either of the toggle switches. If the LED was on, it will go off, and if it was off, it will go on. Flip the other toggle switch, and it will have the same effect. If the LED does not go on at all, you've probably connected it the wrong way around. Another possibility is that two of your alligator clips may have shorted out the battery.

Assuming your two switches do work as I described them, what's going on here? It's time to nail down some basic facts.

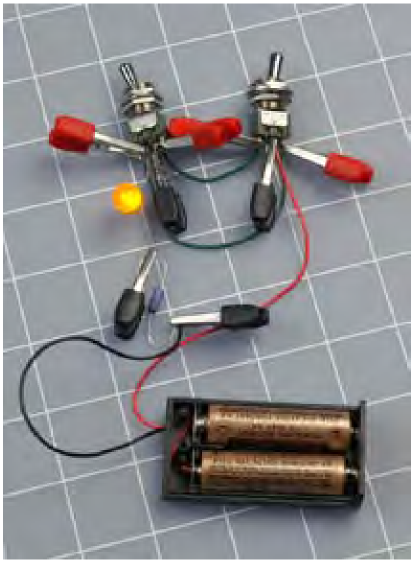

Figure 2-22. *Miniature toggle switches can be used—ideally, with miniature alligator clips—but watch out for short circuits.*

FUNDAMENTALS

All about switches

When you flip the type of toggle switch that you used in Experiment 6, it connects the center terminal with one of the outer terminals. Flip the switch back, and it connects the center terminal with the other outer terminal, as shown in Figure 2-23.

The center terminal is called the pole of the switch. Because you can flip, or throw, this switch to make two possible connections, it is called a *double-throw switch*. As mentioned earlier, a single-pole, double-throw switch is abbreviated *SPDT*.

Some switches are on/off, meaning that if you throw them in one direction they make a contact, but in the other direction, they make no contact at all. Most of the light switches in your house are like this. They are known as *single-throw switches*. A single-pole, single-throw switch is abbreviated *SPST*.

Some switches have two entirely separate poles, so you can make two separate connections simultaneously when you flip the switch. These are called *double-pole switches*. Check the photographs in Figures 2-24 through 2-26 of old-fashioned "knife" switches (which are still used to teach electronics to kids in school) and you'll see the simplest representation of single and double poles, and single and double throws. Various toggle switches that have contacts sealed inside them are shown in Figure 2-27.

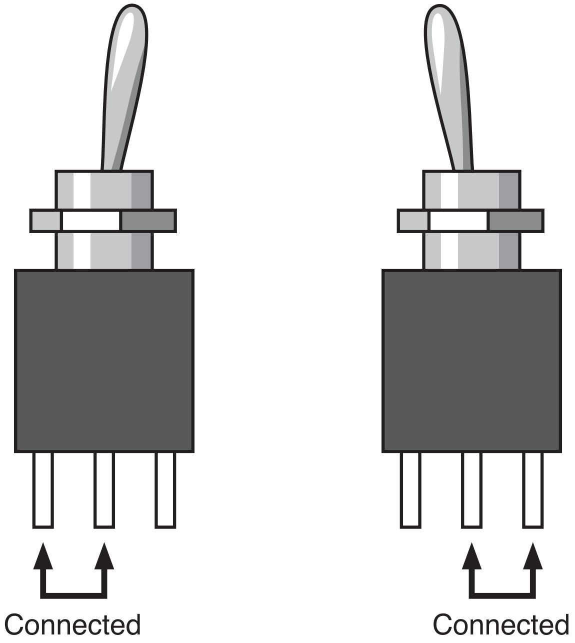

Figure 2-23. *The center terminal is the pole of the switch. When you flip the toggle, the pole changes its connection.*

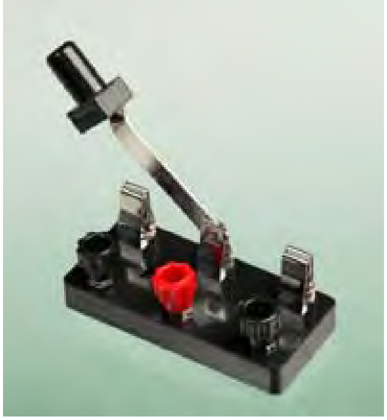

Figure 2-24. *This primitive-looking single-pole, double-throw switch does exactly the same thing as the toggle switches in Figures 2-23 and 2-27.*

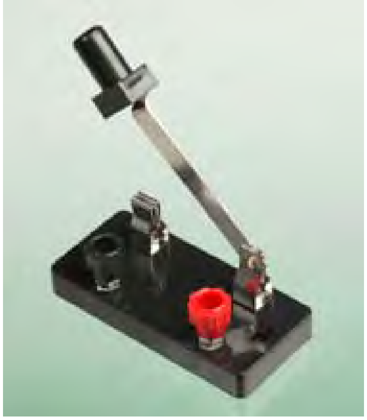

Figure 2-25. *A single-pole, single-throw switch makes only one connection with one pole. Its two states are simply open and closed, on and off.*

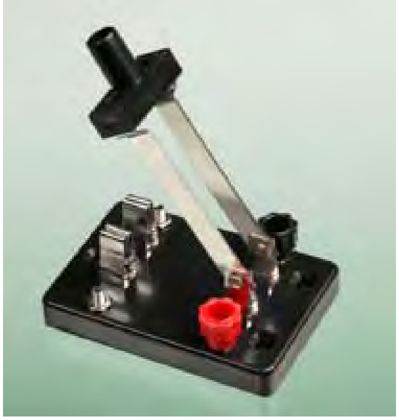

Figure 2-26. *A double-pole, single-throw switch makes two separate on/off connections.*

FUNDAMENTALS

All about switches (continued)

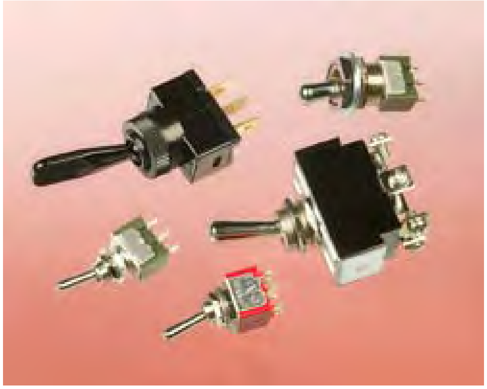

Figure 2-27. *These are all toggle switches. Generally, the larger the switch, the more current it can handle.*

To make things more interesting, you can also buy switches that have three or four poles. (Some rotary switches have even more, but we won't be using them.) Also, some double-throw switches have an additional "center off" position.

Putting all this together, I made a table of possible types of switches (Figure 2-28). When you're reading a parts catalog, you can check this table to remind yourself what the abbreviations mean.

	Single Pole	Double Pole	3-Pole	4-Pole
Single Throw	SPST ON-OFF	DPST ON-OFF	3PST ON-OFF	4PST ON-OFF
Double Throw	SPDT ON-ON	DPDT ON-ON	3PDT ON-ON	4PDT ON-ON
Double Throw with Center Off	SPDT ON-OFF-ON	DPDT ON-OFF-ON	3PDT ON-OFF-ON	4PDT ON-OFF-ON

Figure 2-28. *This table summarizes all the various options for toggle switches and pushbuttons.*

Now, what about pushbuttons? When you press a door bell, you're making an electrical contact, so this is a type of switch—and indeed the correct term for it is a momentary switch, because it makes only a momentary contact. Any spring-loaded switch or button that wants to jump back to its original position is known as a momentary switch. We indicate this by putting its momentary state in parentheses. Here are some examples:

- OFF-(ON): Because the ON state is in parentheses, it's the momentary state. Therefore, this is a single-pole switch that makes contact only when you push it, and flips back to make no contact when you let it go. It is also known as a "normally open" momentary switch, abbreviated "NO."
- ON-(OFF): The opposite kind of momentary single-pole switch. It's normally ON, but when you push it, you break the connection. So, the OFF state is momentary. It is known as a "normally closed" momentary switch, abbreviated "NC."
- (ON)-OFF-(ON): This switch has a center-off position. When you push it either way, it makes a momentary contact, and returns to the center when you let it go.

Other variations are possible, such as ON-OFF-(ON) or ON-(ON). As long as you remember that parentheses indicate the momentary state, you should be able to figure out what these switches are.

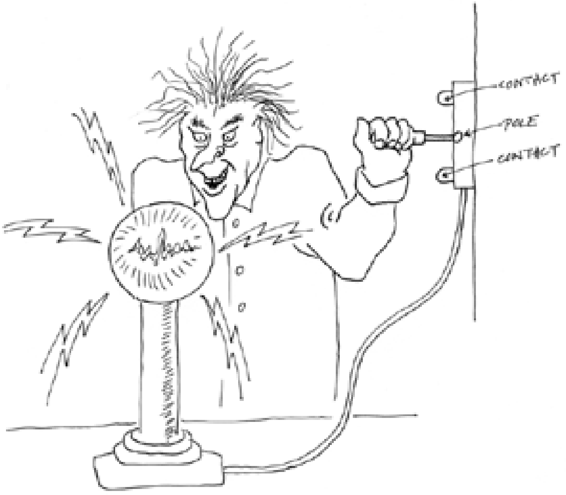

Figure 2-29. *This evil mad scientist is ready to apply power to his experiment. For this purpose, he is using a single-pole, double-throw knife switch, conveniently mounted on the wall of his basement laboratory.*

FUNDAMENTALS

All about switches (continued)

Sparking

When you make and break an electrical connection, it tends to create a spark. Sparking is bad for switch contacts. It eats them until the switch doesn't make a reliable connection anymore. For this reason, you must use a switch that is appropriate for the voltage and amperage that you are dealing with. Electronic circuits generally are low-current, and low-voltage, so you can use almost any switch, but if you are switching a motor, it will tend to suck an initial surge of current that is at least double the rating of the motor when it is running constantly. You should probably use a 4-amp switch to turn a 2-amp motor on and off.

Checking a switch

You can use your meter to check a switch. Doing this helps you find out which contacts are connected when you turn a switch one way or the other. It's also useful if you have a pushbutton and you can't remember whether it's the type that is normally open (you press it to make a connection) or normally closed (you press it to break the connection). Set your meter to measure ohms, and touch the probes to the switch terminals while you work the switch.

This is a hassle, though, because you have to wait while the meter makes an accurate measurement. When you just want to know whether there is a connection, your meter has a "continuity tester" setting. It beeps if it finds a connection, and stays silent if it doesn't. See Figures 2-30 through 2-32 for examples of meters set to test continuity. Figure 2-33 offers an example of a toggle switch being tested for continuity.

Use the continuity-testing feature on your meter only on circuits or components that have no power in them at the time.

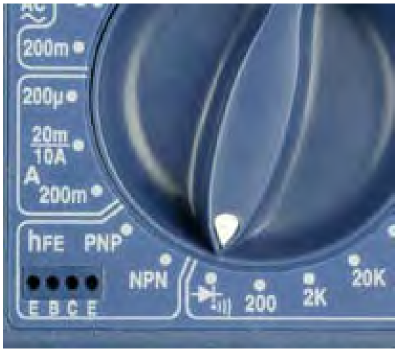

Figure 2-30

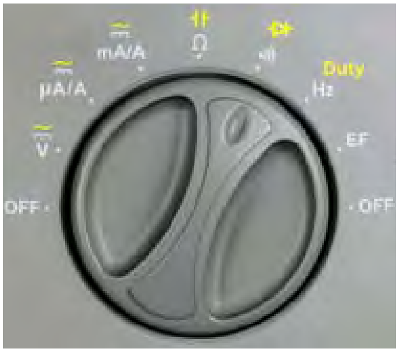

Figure 2-31

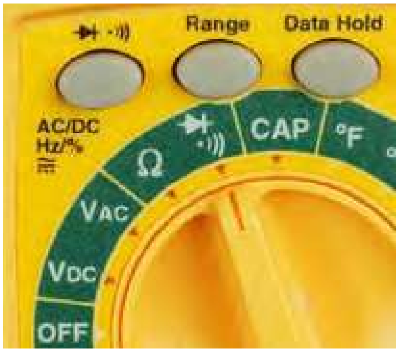

Figure 2-32. *To check a circuit for continuity, turn the dial of your meter to the symbol shown. Only use this feature when there is no power in the component or the circuit that you are testing.*

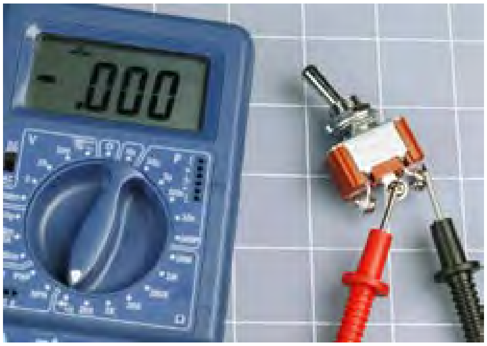

Figure 2-33. *When the switch connects two of its terminals, the meter shows zero resistance between them and will beep if you have set it to verify continuity.*

BACKGROUND

Early switching systems

Switches seem to be such a fundamental feature of our world, and their concept is so simple that it's easy to forget that they went through a gradual process of development and refinement. Primitive knife switches were quite adequate for pioneers of electricity who simply wanted to connect and disconnect electricity to some apparatus in a laboratory, but a more sophisticated approach was needed when telephone systems began to proliferate. Typically, an operator at a "switchboard" needed a way to connect any pair of 10,000 lines on the board. How could it be done?

In 1878, Charles E. Scribner (Figure 2-34) developed the "jack-knife switch," so called because the part of it that the operator held looked like the handle of a jackknife. Protruding from it was a plug, and when the plug was pushed into a socket, it made contact inside the socket. The socket, in fact, was the switch.

Figure 2-34. *Charles E. Scribner invented the "jack-knife switch" to satisfy the switching needs of telephone systems in the late 1800s. Today's audio jacks still work on the same basis.*[*]

Audio connectors on guitars and amplifiers still work on the same principle, and when we speak of them as being "jacks," the term dates back to Scribner's invention. Switch contacts still exist inside a jack socket.

Today, of course, telephone switchboards have become as rare as telephone operators. First they were replaced with relays—electrically operated switches, which I'll talk about later in this chapter. And then the relays were superceded by transistors, which made everything happen without any moving parts. Before the end of this chapter, you'll be switching current using transistors.

---

[*] The photo on which this drawing is based first appeared in *The History of the Telephone* by Herbert Newton Casson in 1910 (Chicago: A. C. McClurg & Co.).

Introducing Schematics

In Figure 2-35, I’ve redrawn the circuit from Experiment 6 in a simplified style known as a “schematic.” From this point onward, I will be illustrating circuits with schematics, because they make circuits easier to understand. You just need to know a few symbols to interpret them.

Larger versions of all schematics and breadboard photos are available online at this book's website: http://oreilly.com/catalog/9780596153748.

Compare the schematic here with the drawing of the circuit in Figure 2-17. They both show exactly the same thing: Components, and connections between them. The gray rectangles are the switches, the zigzag thing is the resistor, and the symbol with two diagonal arrows is the LED.

The schematic LED symbol includes two arrows indicating that it emits light, because there are some kinds of diodes, which we’ll get to later, that don’t. The triangle inside the diode symbol always points from positive to negative.

Trace the path that electricity can take through the circuit and imagine the switches turning one way or the other. You should see clearly now why either switch will reverse the state of the LED from on to off or off to on.

This same circuit is used in houses where you have a switch at the bottom of a flight of stairs, and another one at the top, both controlling the same lightbulb. The wires in a house are much longer, and they snake around behind the walls, but because their connections are still the same, they could be represented with the same basic schematic. See Figure 2-36.

A schematic doesn’t tell you exactly where to put the components. It just tells you how to join them together. One problem: Different people use slightly different schematic symbols to mean the same thing. Check the upcoming section, “Fundamentals: Basic schematic symbols,” for the details.

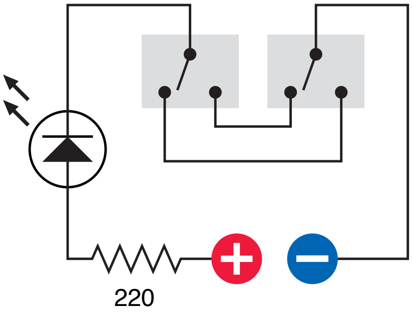

Figure 2-35. *This schematic shows the same circuit as in Figure 2-17 and makes it easier to see how the switches function.*

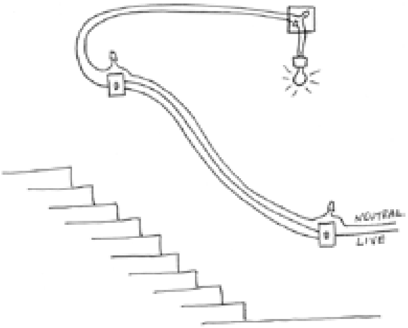

Figure 2-36. *The two-switch circuit shown in Figures 2-17 and 2-35 is often found in house wiring, especially where switches are located at the top and bottom of a flight of stairs. This sketch shows what you might find inside the walls. Wires are joined with “wire nuts” inside boxes that are hidden from everyday view.*

FUNDAMENTALS

Basic schematic symbols

Schematic symbols are like words in a language: they have mutated over the years into a confusing range of variations. A simple on/off (single-pole, single-throw, or SPST) switch, for instance, can be represented by any of the symbols shown in Figure 2-37. They all mean exactly the same thing.

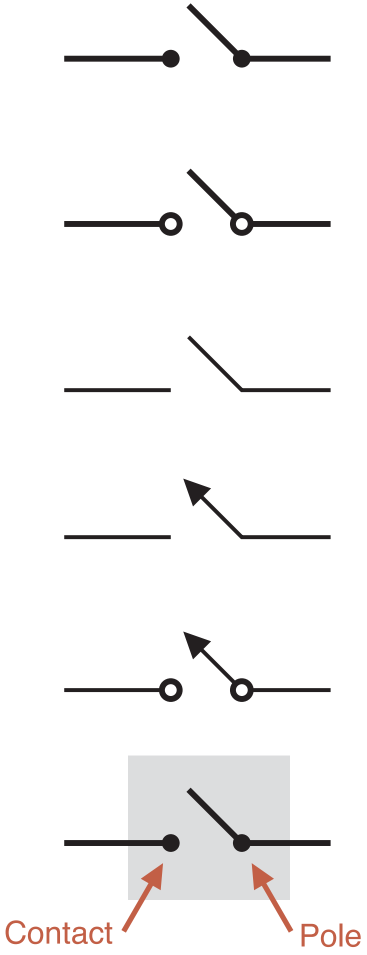

Figure 2-37. *Variations on a theme: Just some of the different styles used to depict a single-pole, single-throw switch in schematic diagrams. The bottom version is the style used in this book.*

Figure 2-38 shows double-pole, double-throw switches. A dotted line indicates a mechanical connection inside the switch, so that when you flip it, you affect both the poles simultaneously. Remember, the poles are electrically isolated from each other.

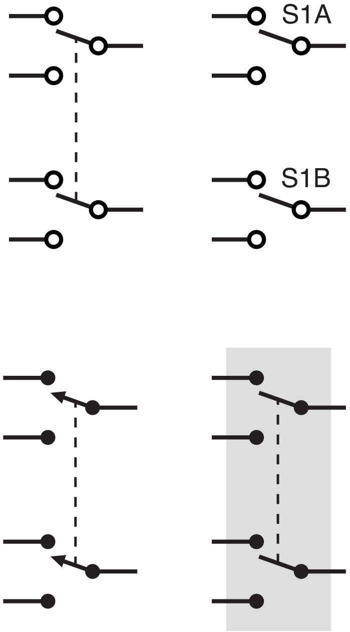

Figure 2-38. *More variations: some different styles for depicting a double-pole, double-throw switch. The style at bottom-right is used in this book.*

Once in a while, you may find a schematic in which switches seem to be scattered around, but the way they are identified (such as S1A, S1B, S1C, and so on) tells you that this is really all one switch with multiple poles.

In the schematics in this book, I'll place a gray rectangle behind each switch. This gray rectangle is not a standard symbol; you won't find it in other books. I'm just including it to remind you that the parts inside are all contained in one package.

A very important stylistic variation in schematics is the way they show whether wires make a connection with each other. Old schematics used to show a little semicircular bump in a wire if it crossed another wire without making a connection. Because modern circuit-drawing software doesn't create this style of schematic, it is no longer often used. The modern style, which you are likely to find if you browse through schematics online, can be summarized like this:

- A dot joining two wires indicates an electrical connection.
- No dot indicates no connection.

FUNDAMENTALS

Basic schematic symbols (continued)

The problem is that this is not very intuitive, especially when you're just beginning to use schematics. When you see two wires crossing, you can easily imagine that they are making a connection, even though there's no dot at the intersection. Therefore, in the interests of clarity, I've chosen to use the old "semicircular bump" style of schematic in this book (see Figure 2-39). It can be summarized like this:

- A dot joining two wires indicates an electrical connection.
- A bump in a wire that crosses another wire indicates no connection.

In this book, you won't find wires crossing each other without either a dot or a bump.

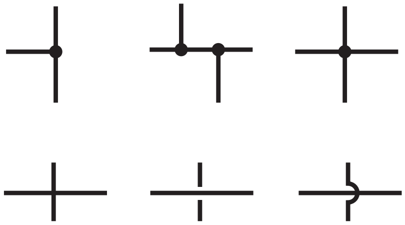

Figure 2-39. *In wiring schematics, a dot always indicates an electrical connection. However, the cross-shaped intersection of wires at top-right is considered bad style because if the dot is accidentally omitted or poorly printed, the intersection can be mistaken for the type shown at bottom left, in which the wires do not make a connection. All three of the configurations in the bottom row indicate no connection, the first example being the most common style, the center example being least common, and the third being the most old-fashioned—although for reasons of clarity, it is used in this book.*

In a battery-powered circuit, you may find a battery symbol, but more often you will find a little note indicating where positive voltage enters the system, while negative is indicated by a "ground" symbol. In fact there may be ground symbols all over the place. You have to remember that when you build a circuit, all the wires leading to grounds must actually be connected together, to the negative side of the voltage supply.

The idea of the ground symbol dates back to the time when electronic gadgets were mounted on a metal chassis, which was connected to the negative side of the power supply. The ground symbol really meant "connect to the chassis." Some variants in the ground symbol are shown in Figure 2-40.

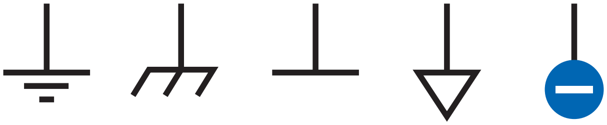

Figure 2-40. *All of these symbols are used to mean the same thing: connect the wire to "ground" or "chassis" or the negative side of the power source. The far-right symbol is used in this book.*

In this book, we have color throughout, so I'll show a red positive and blue negative to clarify where the power is connected, and I won't use ground symbols. Once again, my purpose is to minimize the risk of misunderstandings, because I know how frustrating it is to build a circuit that doesn't work.

A big inconsistency in schematics is the way in which they show resistors. The traditional zigzag symbol has been abandoned in Europe. Instead they use a rectangle with a number inside indicating the number of ohms. See Figure 2-41. The Europeans also changed the way in which decimal points are represented: they omit them as much as possible, because in badly printed schematics, the little dots tend to get lost (or can be confused with dust and dirt). So, a 4.7KΩ resistor will be listed as 4K7, and a 1.2MΩ resistor will be 1M2. I like this notation, so I'm going to be using it myself, but I'll be keeping the zigzag resistor symbol, which remains widely used in the United States.

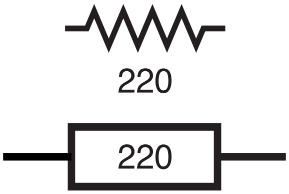

Figure 2-41. *Two styles for depicting a 220Ω resistor. The upper version is traditional, and still used in the United States. The lower version is European.*

FUNDAMENTALS

Basic schematic symbols (continued)

Potentiometers suffer from the same inconsistent style between the United States and Europe, but either way, you'll find an arrow showing where the wiper (usually, the center terminal) touches the resistance. See Figure 2-42. And sometimes LEDs are shown inside circles, and sometimes not. I prefer circles, myself. See Figure 2-46.

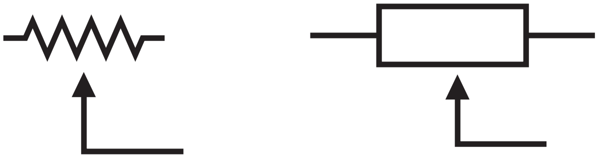

Figure 2-42. *Potentiometer symbols: the left is traditional and used in the United States, the right is European. In both cases the arrow indicates the wiper (usually the center terminal).*

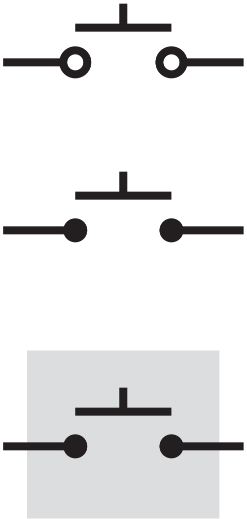

Figure 2-43. *Three ways of indicating a pushbutton switch.*

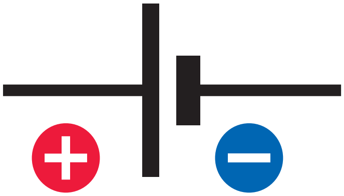

Figure 2-44. *The battery symbol is usually shown without + and – symbols. I've added them for clarity.*

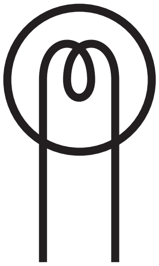

Figure 2-45. *Symbol for an incandescent lightbulb.*

I'll explore other symbol variants later in the book. Meanwhile, the most important things to remember are:

- The positions of components in a schematic are not important.
- The styles of symbols used in a schematic are not important.
- The connections between the components are extremely important.

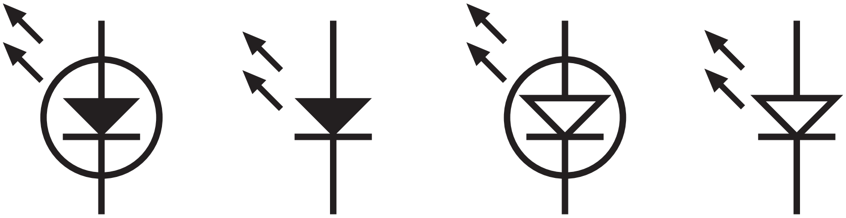

Figure 2-46. *Sometimes an LED is shown with a circle around it; sometimes not. In this book, I will include the circle. The arrows indicate emitted light.*

FUNDAMENTALS

Basic schematic symbols (continued)

For example, the three LED circuits that I have included in Figure 2-47 show components in different positions, using different symbols, but all three circuits function exactly the same way, because their connections are the same. In fact, they all depict the circuit that you built in Experiment 4, shown in Figure 1-50.

Often the symbols in a schematic are placed so that the circuit is most intuitively easy to understand, regardless of how you may build it with actual components. Compare the example in Figure 2-48, showing the two DPDT switches, with the version shown back in Figure 2-35. The previous one looked more like your bench-top version of it, but Figure 2-48 shows the flow of electricity more clearly.

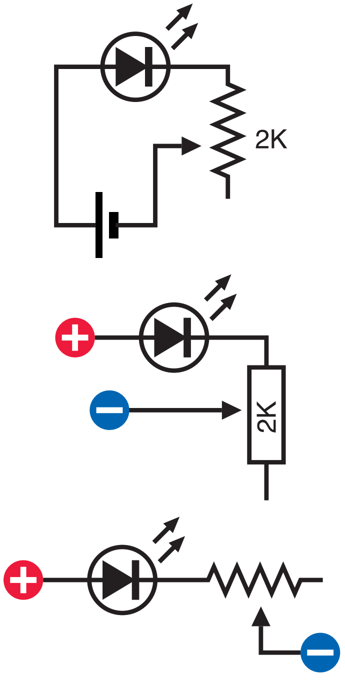

Figure 2-47. *These three schematics all depict the same basic circuit. It's the circuit that you built with the potentiometer in Experiment 4.*

In many schematics, the positive side of the power supply is shown at the top of the diagram, and negative or ground at the bottom. Many people also tend to draw a schematic with an input (such as an audio input, in an amplifier circuit) at the left side, and the output at the right. So, "positive voltage" flows from top to bottom while a signal tends to pass from left to right.

When I was planning this book, initially I drew the schematics to conform with this top-to-bottom, left-to-right convention, but as I started building and testing the circuits, I changed my mind. We use a device known as a "breadboard" to create circuits, and its internal connections require us to lay out components very differently from a typical schematic. When you're starting to learn electronics, it's very confusing to try to rearrange components from a schematic in the configuration that you need for a breadboard.

Therefore, throughout this book, you'll find that I have drawn the schematics to imitate the way you'll wire them on a breadboard. I believe the advantages of doing things this way outweigh the disadvantages of being a little different from the schematic styles that are used elsewhere.

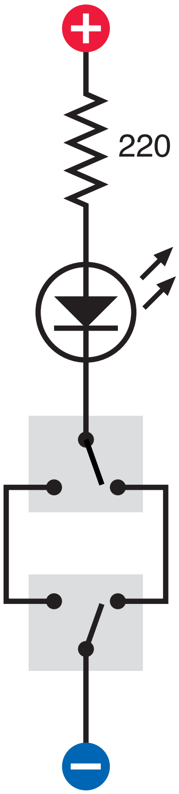

Figure 2-48. *This schematic is just another, clearer, simpler way of showing the circuit that appeared in Figure 2-35.*

Experiment 7: Relay-Driven LEDs

You will need:

- AC adapter, wire cutters and strippers.
- DPDT relay. Quantity: 2.
- LEDs. Quantity: 2.
- Resistor, 680Ω approx. Quantity: 1.
- Pushbutton, SPST. Quantity: 1.
- Hookup wire, 22 gauge, or patch cords.
- Alligator clips. Quantity: 8.
- Utility knife.

The next step in our exploration of switching is to use a remote-controlled switch. By "remote-controlled," I mean one to which you can send a signal to turn it on or off. This kind of switch is known as a *relay*, because it relays an instruction from one part of a circuit to another. Often a relay is controlled by a low voltage or small current, and switches a larger voltage or higher current.

This setup can be cost-effective. When you start your car, for instance, a relatively small, cheap switch sends a small signal down a relatively long, thin, inexpensive piece of wire, to a relay that is near the starter motor. The relay activates the motor through a shorter, thicker, more expensive piece of wire, capable of carrying as much as 100 amps.

Similarly, when you raise the lid on an a top-loading washing machine during its spin cycle, you close a lightweight switch that sends a small signal down a thin wire to a relay. The relay handles the bigger task of switching off the large motor spinning the drum full of wet clothes.

Before you begin this experiment, you need to upgrade your power supply. We're not going to use batteries anymore, because most relays require more than 6 volts, and in any case you should have a power supply that can give you a variety of voltages on demand. The simplest way to achieve this is by using an AC adapter.

First you'll set up the AC adapter. After you have it running, you'll use it to power the relay. Initially the relay will just switch between two LEDs, but then you'll modify the circuit to make the LEDs flash automatically. Finally you'll rebuild the circuit on a breadboard, and say goodbye to alligator clips, for most of the time at least.

Preparing Your AC Adapter

An AC adapter plugs into the wall and converts the high-voltage AC supply in your home into a safe, low DC voltage for electronic devices. Any charger that you use with your cell phone, or iPod, or laptop computer is a special-purpose AC adapter that delivers only one voltage via a specific type of plug. I've asked you to buy a general-purpose adapter that delivers many different voltages, and we're going to begin by getting rid of its plug.

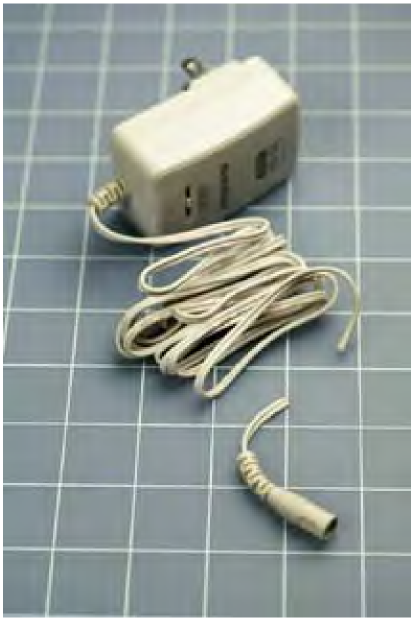

Figure 2-49. *Preparing an AC adapter. First, cut off the little low-voltage plug and throw it away.*

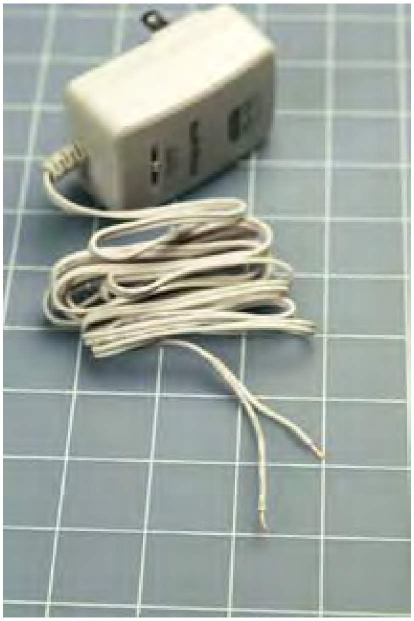

Figure 2-50. *Second, strip the wires, making one shorter than the other to reduce the risk of them touching. Color one of the adapter wires red with a marker, to identify it as the positive one.*

1. It's important to make sure that your AC adapter is not plugged into the wall!
2. Chop off the little plug at the end of its wire. See Figure 2-49.
3. Use a box cutter or utility knife or scissors to make a half-inch cut between the two conductors, and then pull the conductors apart a couple of inches.
4. Use wire cutters to trim one of the conductors shorter than the other, so that after you strip away a little of the insulation, the exposed copper wires cannot easily touch each other. This is a precaution against short-circuiting your AC adapter and burning it out.
5. Strip the two conductors using your wire strippers. Twist the copper strands between finger and thumb so that there are no loose strands sticking out. See Figure 2-50.
6. Make sure that the two wires are not touching each other, and plug your AC adapter into a wall outlet. Set your meter to DC volts and apply the meter probes to the wires from the adapter. If the voltage is preceded with a minus sign, you have the probes the wrong way around. Reverse them and the minus sign should go away. This tells you which wire is positive.
7. Mark the positive wire from the adapter. If the wire has white insulation, you can mark it with a red marker. If it has black insulation, you can tag it with a label. The positive wire will remain positive regardless of which way up you plug the AC adapter into a wall outlet.

The Relay

The type of relay that I want you to use has little spiky legs on the bottom, in a standardized layout. If you buy some other kind of relay, you will have to figure out for yourself which pins are connected to the coil inside, which pins go to the poles of the switch inside it, and which go to the normally closed and normally open contacts. You can check the manufacturer's data sheet for this purpose, but I strongly suggest you use one of the relays mentioned in the shopping list, so that you can follow the instructions here more easily.

I asked you to buy two relays so that you can use one for investigational purposes—meaning that you can break it open and take a look inside. If you do this very, very carefully, the relay should still be usable afterward. If not, well, you still have a spare.

The easiest way to open the relay is with a box cutter or utility knife. Figures 2-52, 2-53, and 2-54 show the technique. Shave the edges of the plastic shell containing the relay, beveling them until you see just a hair-thin opening. Don't go any farther; the parts inside are very, very close to its housing. Now pop the top off. You can use needle-nosed pliers to nibble the rest of the shell away. Read the following section, "Fundamentals: Inside a relay," and then apply power to the relay to see how it works.

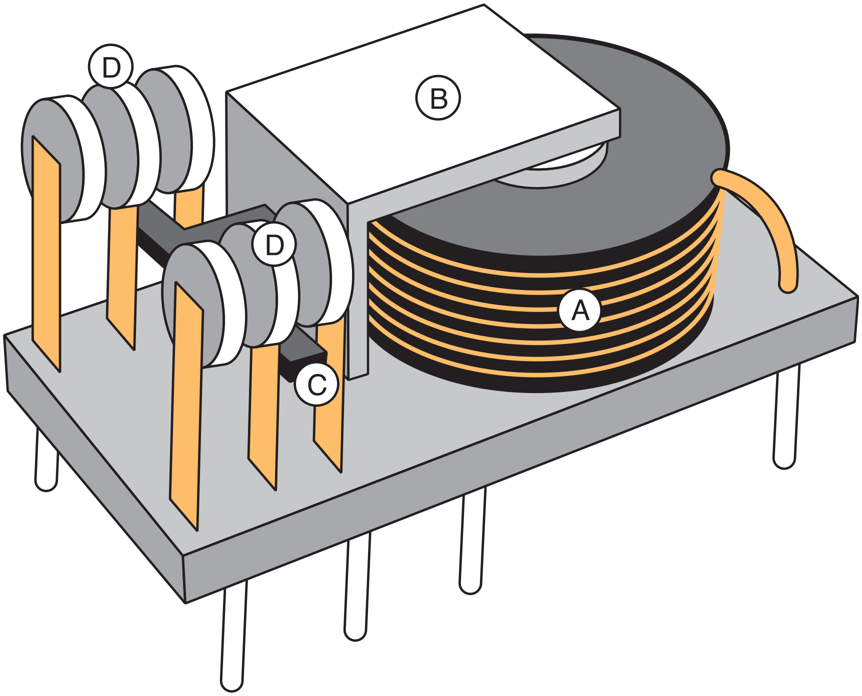

Figure 2-51. *This is one way that the parts inside a relay can be arranged. The coil, A, generates a magnetic attraction pulling lever B downward. A plastic extension, C, pushes outward against flexible metal strips and moves the poles of the relay, D, between the contacts.*

Figure 2-55

Figure 2-56. *Patience is essential when carving the edges of a relay package in order to open it. Faster methods such as a tomahawk or a flamethrower will satisfy the emotional needs of those with a short attention span, but results may be unpredictable.*

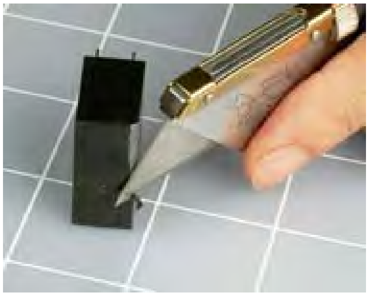

Figure 2-52. *To look inside a sealed relay, shave the top edges of the plastic package with a utility knife til you open a thin crack.*

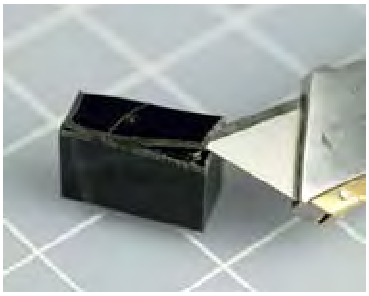

Figure 2-53. *Insert the blade of your knife to pry open the top, then repeat the procedure for the sides.*

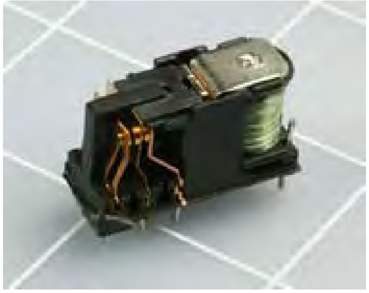

Figure 2-54. *If you are really, really careful, the relay should still work after you open it.*

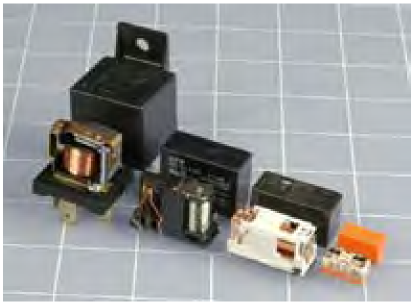

Figure 2-57. *Four assorted 12-volt relays, shown with and without their packages. The automotive relay (far left) is the simplest and easiest to understand, because it is designed without much concern for the size of the package. Smaller relays are more ingeniously designed, more complex, and more difficult to figure out. Usually, but not always, a smaller relay is designed to switch less current than a larger one.*

FUNDAMENTALS

Inside a relay

A relay contains a coil of wire wrapped around an iron core. When electricity runs through the coil, the iron core exerts a magnetic force, which pulls a lever, which pushes or pulls a springy strip of metal, closing two contacts. So as long as electricity runs through the coil, the relay is "energized" and its contacts remain closed.

When the power stops passing through the coil, the relay lets go and the springy strip of metal snaps back into its original position, opening the contacts. (The exception to this rule is a latching relay, which requires a second pulse through a separate coil to flip it back to its original position; but we won't be using latching relays until later in the book.)

Relays are categorized like switches. Thus, you have SPST relays, DPST, SPDT, and so on.

Compare the schematics in Figure 2-58 with the schematics of switches in Figure 2-38. The main difference is that the relay has a coil that activates the switch. The switch is shown in its "relaxed" mode, when no power flows through the coil.

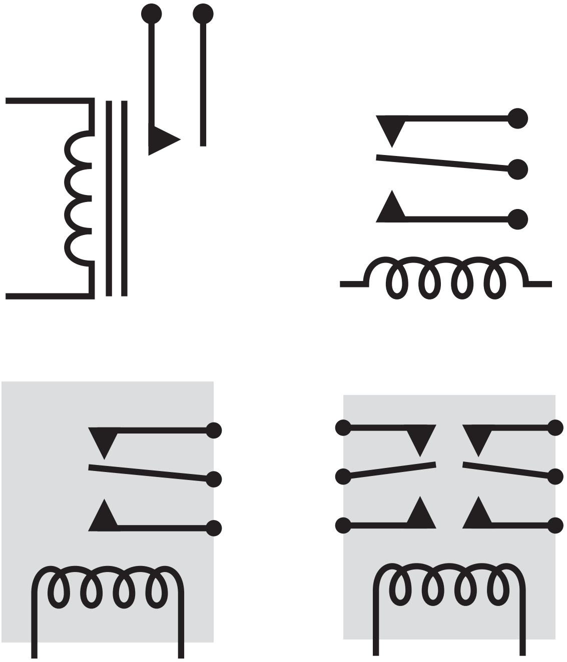

Figure 2-58. *Various ways to show a relay in a schematic. Top left: SPST. Top right and bottom left: SPDT. Bottom right: DPDT. The styles at bottom-left and bottom-right will be used in this book.*

The contacts are shown as little triangles. When there are two poles instead of one, the coil activates both switches simultaneously.

Most relays are nonpolarized, meaning that you can run electricity through the coil in either direction, and the relay doesn't care. You should check the data sheet to make sure, though. Some relay coils work on AC voltage, but almost all low-voltage relays use direct current—a steady flow of electricity, such as you would get from a battery. We'll be using DC relays in this book.

Relays suffer from the same limitations as switches: their contacts will be eroded by sparking if you try to switch too much voltage. It's not worth saving a few dollars by using a relay that is rated for less current or voltage than your application requires. The relay will fail you when you need it most, and may be inconvenient to replace.

Because there are so many different types of relays, read the specifications carefully before you buy one. Look for these basics:

Coil voltage

The voltage that the relay is supposed to receive when you energize it.

Set voltage

The minimum voltage that the relay needs to close its switch. This will be a bit less than the ideal coil voltage.

Operating current

The power consumption of the coil, usually in milliamps, when the relay is energized. Sometimes the power is expressed in milliwatts.

Switching capacity

The maximum amount of current that you can switch with contacts inside the relay. Usually this is for a "resistive load," meaning a passive device such as light bulb. When you use a relay to switch on a motor, the motor takes a big initial surge of current before it gets up to speed. In this case, you should choose a relay rated for double the current that the motor draws when it is running.

Procedure

Turn the relay with its legs in the air and attach wires and LEDs as shown in Figure 2-59, with a 680Ω resistor (a 1K resistor will be OK if you don't have the correct value). Also attach a pushbutton switch. (Your pushbutton switch may look different from the one shown, but as long as it is a SPST pushbutton with two contacts at the bottom, it will work the same way.) When you press the pushbutton, the relay will make the first LED go out and the second LED light up. When you release the pushbutton, the first LED lights up and the second one goes out.

How It Works

Check the schematic in Figure 2-60 and compare it with Figure 2-59. Also see Figure 2-62, which shows how the pins outside the relay make connections inside the relay when its coil is energized, and when it is not energized.

This is a DPDT relay, but we are only using one pole and ignoring the other. Why not buy a SPDT relay? Because I want the pins to be spaced the way they are when you will upgrade this circuit by transferring it onto a breadboard, which will happen very shortly.

On the schematic, I have shown the switch inside the relay in its relaxed state. When the coil is energized, the switch flips upward, which seems counterintuitive, but just happens to be the way that this particular relay is made.

When you're sure you understand how the circuit works, it's time to move on to the next step: making a small modification to get the relay to switch itself on and off, as we'll do in Experiment 8.

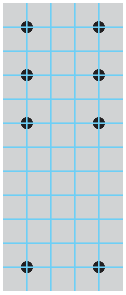

Figure 2-61. *The layout of the pins of the relay, superimposed on a grid of 1/10-inch squares. This is the type of relay that you will need in Experiment 8.*

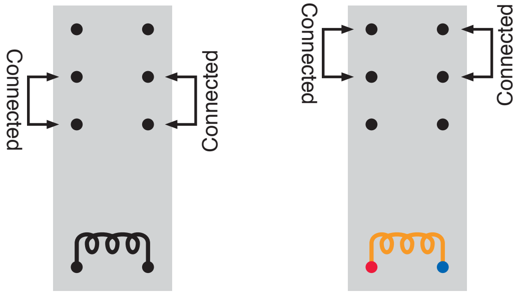

Figure 2-62. *How the relay connects the pins, when it is not energized (left) and when it is energized (right).*

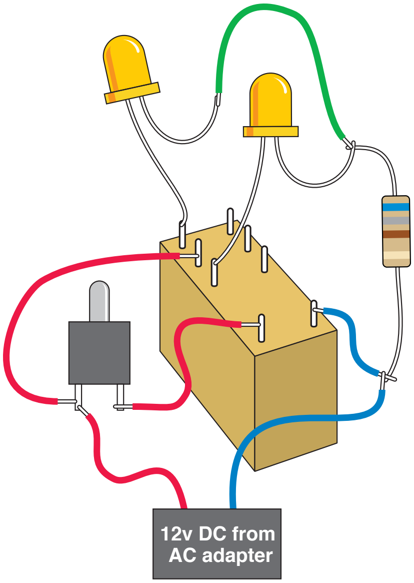
12v DC from AC adapter

Figure 2-59. *As before, you can use patch cords, if you have them, instead of some of the wired connections shown here.*

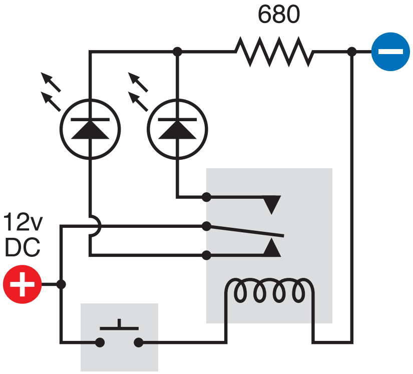

Figure 2-60. *Same circuit, shown in schematic form.*

Larger versions of all schematics and breadboard photos are available online at this book's website: http://oreilly.com/catalog/9780596153748.

Experiment 8: A Relay Oscillator

You will need:

- AC adapter, breadboard, wire, wire cutters and strippers.
- DPDT relay. Quantity: 1.
- LEDs. Quantity: 2.
- Pushbutton, SPST. Quantity: 1.
- Alligator clips. Quantity: 8.
- Resistor, approximately 680Ω. Quantity: 1.
- Capacitor, electrolytic, 1,000 µF. Quantity: 1.

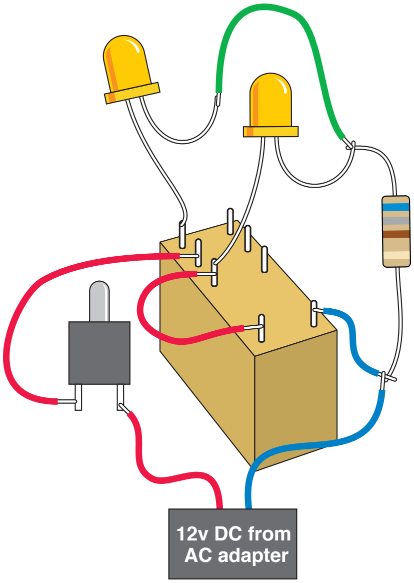

Figure 2-63. *A small revision to the previous circuit causes the relay to start oscillating when power is applied.*

Look at the revised drawing in Figure 2-63 and the revised schematic in Figure 2-64 and compare them with the previous ones. Originally, there was a direct connection from the pushbutton to the coil. In the new version, the power gets to the coil by going, first, through the contacts of the relay.

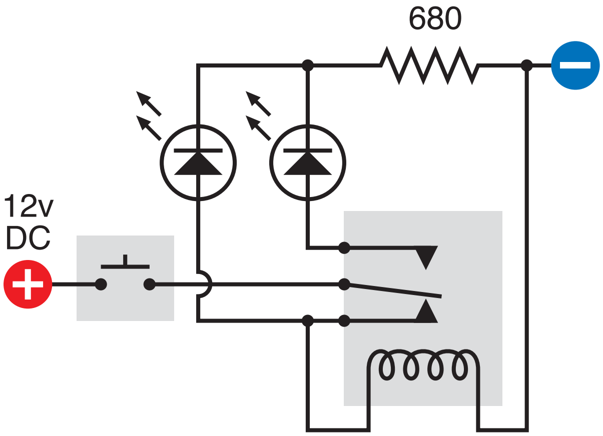

Figure 2-64. *The oscillator circuit shown in schematic form.*

Now, when you press the button, the contacts in their relaxed state feed power to the coil as well as to the lefthand LED. But as soon as the coil is energized, it opens the contacts. This interrupts the power to the coil—so the relay relaxes, and the contacts close again. They feed another pulse of power to the coil, which opens the contacts again, and the cycle repeats endlessly.

Because we're using a very small relay, it switches on and off extremely fast. In fact, it oscillates perhaps 50 times per second (too fast for the LEDs to show what's really happening). Make sure your circuit looks like the one in the diagram, and then press the pushbutton very briefly. You should hear the relay make a buzzing sound. If you have impaired hearing, touch the relay lightly with your finger, and you should feel the relay vibrating.

When you force a relay to oscillate like this, it's liable to burn itself out or destroy its contacts. That's why I asked you to press the pushbutton briefly. To make the circuit more practical, we need something to slow the relay down and prevent it from self-destructing. That necessary item is a capacitor.

Adding Capacitance

Add a 1,000 µF electrolytic capacitor in parallel with the coil of the relay as shown in the diagram in Figure 2-65 and the schematic in Figure 2-66. Check Figure 2-14 if you're not sure what a capacitor looks like. The 1,000 µF value will be printed on the side of it, and I'll explain what this means a little later.

Make sure the capacitor's *short* wire is connected to the *negative* side of the circuit; otherwise, it won't work. In addition to the short wire, you should find a minus sign on the body of the capacitor, which is there to remind you which side is negative. Electrolytic capacitors are fussy about this.

When you press the button now, the relay should click slowly instead of buzzing. What's happening here?

A capacitor is like a tiny rechargeable battery. It's so small that it charges in a fraction of a second, before the relay has time to open its lower pair of contacts. Then, when the contacts are open, the capacitor acts like a battery, providing power to the relay. It keeps the coil of the relay energized for about one second. After the capacitor exhausts its power reserve, the relay relaxes and the process repeats.

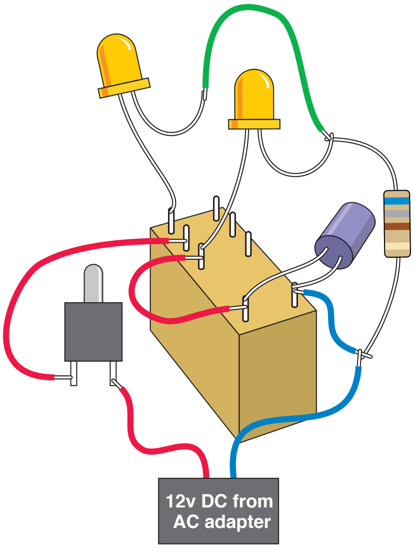
12v DC from AC adapter

Figure 2-65. *Adding a capacitor makes the relay oscillate more slowly.*

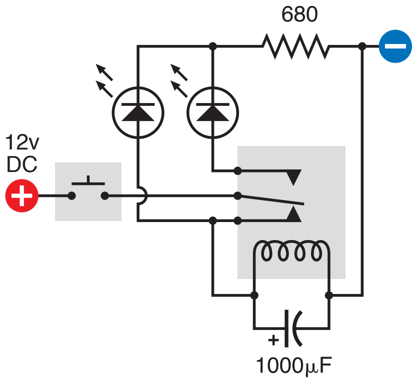

Figure 2-66. *The capacitor appears at the bottom of this schematic diagram.*

FUNDAMENTALS

Farad basics

The Farad is an international unit to measure capacitance. Modern circuits usually require small capacitors. Consequently it is common to find capacitors measured in microfarads (one-millionth of a farad) and even picofarads (one-trillionth of a farad). Nanofarads are also used, more often in Europe than in the United States. See the following conversion table.

0.001 nanofarad	1 picofarad	1 pF
0.01 nanofarad	10 picofarads	10 pF
0.1 nanofarad	100 picofarads	100 pF
1 nanofarad	1,000 picofarads	1,000 pF
0.001 microfarad	1 nanofarad	1 nF
0.01 microfarad	10 nanofarads	10 nF
0.1 microfarad	100 nanofarads	100 nF
1 microfarad	1,000 nanofarads	1,000 nF
0.000001 Farad	1 microfarad	1 µF
0.00001 Farad	10 microfarads	10 µF
0.0001 Farad	100 microfarads	100 µF
0.001 Farad	1,000 microfarads	1,000 µF

(You may encounter capacitances greater than 1,000 microfarads, but they are uncommon.)

Getting Zapped by Capacitors

If a large capacitor is charged with a high voltage, it can retain that voltage for a long time. Because the circuits in this book use low voltages, you don't have to be concerned about that danger here, but if you are reckless enough to open an old TV set and start digging around inside (which I do not recommend), you may have a nasty surprise. An undischarged capacitor can kill you as easily as if you stick your finger into an electrical outlet. Never touch a large capacitor unless you really know what you're doing.

FUNDAMENTALS

Capacitor basics

DC current does not flow through a capacitor, but voltage can accumulate very quickly inside it, and remains after the power supply is disconnected. Figures 2-67 and 2-68 may help to give you an idea of what happens inside a capacitor when it is fully charged.

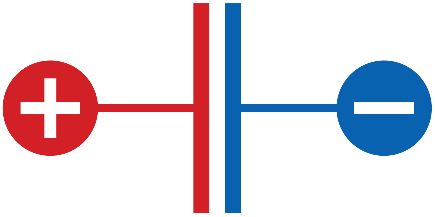

Figure 2-67. *When DC voltage reaches a capacitor, no current flows, but the capacitor charges itself like a little battery. The positive and negative charges are equal and opposite.*

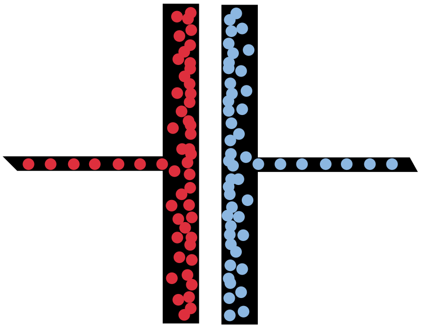

Figure 2-68. *You can imagine positive "charge particles" accumulating on one side of the capacitor and attracting negative "charge particles" to the opposite side.*

In most modern electrolytic capacitors, the plates have been reduced to two strips of very thin, flexible, metallic film, often wrapped around each other, separated by an equally thin insulator. Disc ceramic capacitors typically consist of just a single disc of nonconductive material with metal painted on both sides and leads soldered on.

The two most common varieties of capacitors are ceramic (capable of storing a relatively small charge) and electrolytic (which can be much larger). Ceramics are often disc-shaped and yellow in color; electrolytics are often shaped like miniature tin cans and may be just about any color. Refer back to Figures 2-13 and 2-14 for some examples.

FUNDAMENTALS

Capacitor basics (continued)

Ceramic capacitors have no polarity, meaning that you can apply negative voltage to either side of them. Electrolytics do have polarity, and won't work unless you connect them the right way around.

The schematic symbol for a capacitor has two significant variants: with two straight lines (symbolizing the plates inside a capacitor), or with one straight line and one curved line, as shown in Figure 2-69. When you see a curved line, that side of the capacitor should be more negative than the other. The schematic symbol may also include a + sign. Unfortunately, some people don't bother to draw a curved plate on a polarized capacitor, yet others draw a curved plate even on a nonpolarized capacitor.

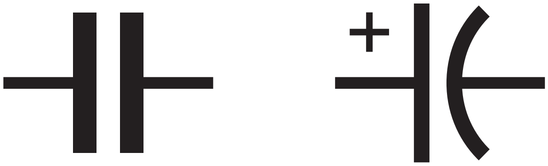

Figure 2-69. *The generic schematic for a capacitor is on the left. The version on the right indicates a polarized capacitor which requires its left plate to be "more positive" than its right plate. The plus sign is often omitted.*

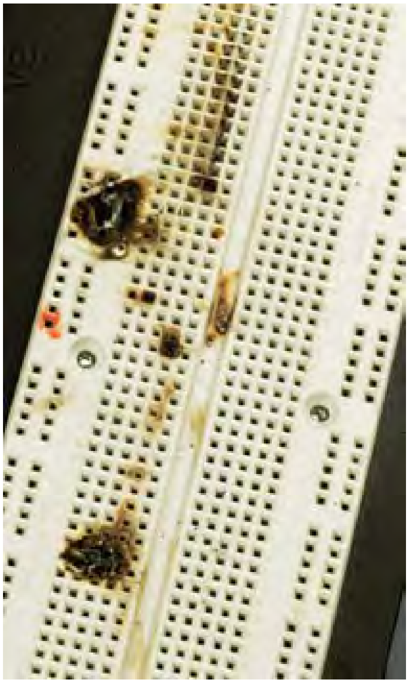

Figure 2-70. *A tantalum capacitor was plugged into this breadboard, accidentally connected the wrong way around to a power source capable of delivering a lot of current. After a minute or so of this abuse, the capacitor rebelled by popping open and scattering small flaming pieces, which burned their way into the plastic of the breadboard. Lesson learned: observe polarity!*

Capacitor Polarity

You must connect an electrolytic capacitor so that its longer wire is more positive than its shorter wire. The shell of the capacitor is usually marked with a negative sign near the shorter wire.

Some capacitors may behave badly if you don't observe their polarity. One time I connected a tantalum capacitor to a circuit, using a power supply able to deliver a lot of current, and was staring at the circuit and wondering why it wasn't working when the capacitor burst open and scattered little flaming fragments of itself in a 3-inch radius. I had forgotten that tantalum capacitors can be fussy about positive and negative connections. Figure 2-70 shows the aftermath.

BACKGROUND

Michael Faraday and capacitors

The earliest capacitors consisted of two metal plates with a very small gap between them. The principle of the thing was simple:

- If one plate was connected to a positive source, the positive charges attracted negative charges onto the other plate.
- If one plate was connected to a negative source, the negative charges attracted positive charges onto the other plate.

Figures 2-67 and 2-68, shown previously, convey the basic idea.

The electrical storage capacity of a capacitor is known as its *capacitance*, and is measured in farads, named after Michael Faraday (Figure 2-71), another of the pantheon of electrical pioneers. He was an English chemist and physicist who lived from 1791 to 1867.

Although Faraday was relatively uneducated and had little knowledge of mathematics, he had an opportunity to read a wide variety of books while working for seven years as a bookbinder's apprentice, and thus was able to educate himself. Also, he lived at a time when relatively simple experiments could reveal fundamental properties of electricity. Thus he made major discoveries including electromagnetic induction, which led to the development of electric motors. He also discovered that magnetism could affect rays of light.

His work earned him numerous honors, and his picture was printed on English 20-pound bank notes from 1991 through 2001.

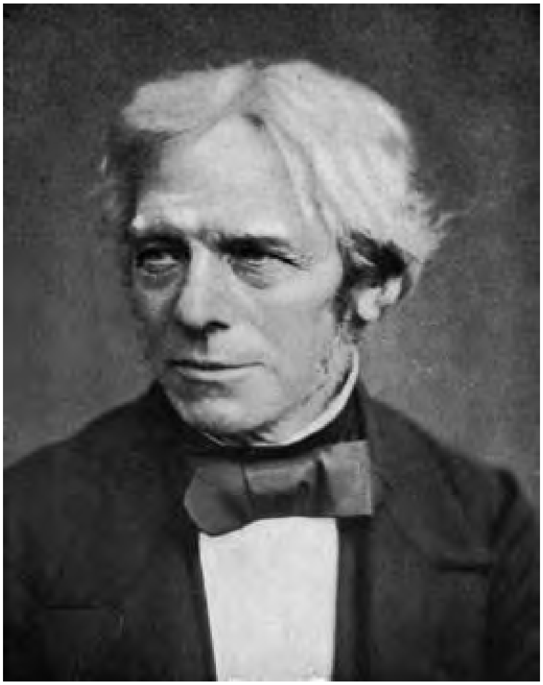

Figure 2-71. *Michael Faraday*

Breadboarding the Circuit

I promised to free you in time from the frustrations of alligator clips, and that time has come. Please turn your attention to the block of plastic with lots of little holes in it that I asked you to buy. For reasons that I do not know, this is called a *breadboard*. When you plug components into the holes, hidden metal strips inside the breadboard connect the components for you, allowing you to set up a circuit, test it, and modify it very easily. Afterward you can pull the components off the breadboard and put them away for future experiments.

Without a doubt, breadboarding is the most convenient way to test something before you decide whether you want to keep it.

Almost all breadboards are designed to be compatible with integrated circuit chips (which we will be using in Chapter 4 of this book). The chip straddles an empty channel in the center of the breadboard with rows of little holes either side—usually five holes per row. You insert other components into these holes.

In addition, the breadboard should have columns of holes running down each side. These are used to distribute positive and negative power.

Take a look at Figures 2-72 and 2-73, which show the upper part of a typical breadboard seen from above, and the same breadboard seen as if with X-ray vision, showing the metal strips that are embedded behind the holes.

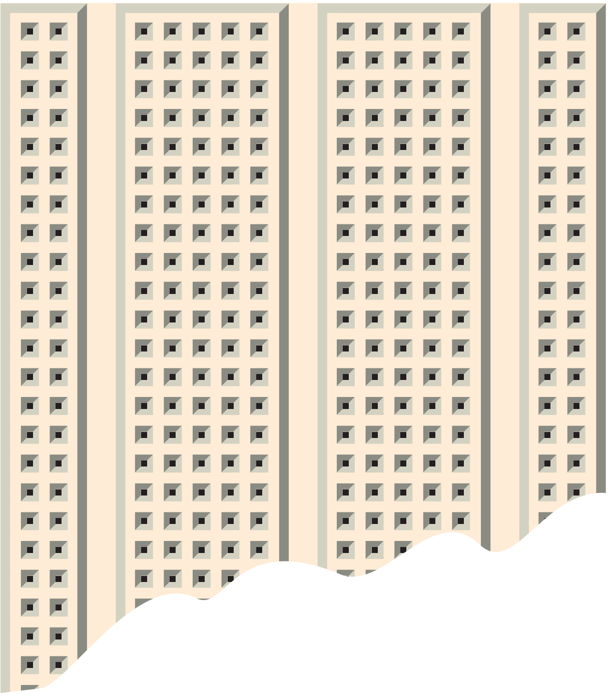

Figure 2-72. *A typical breadboard. You can plug components into the holes to test a circuit very quickly.*

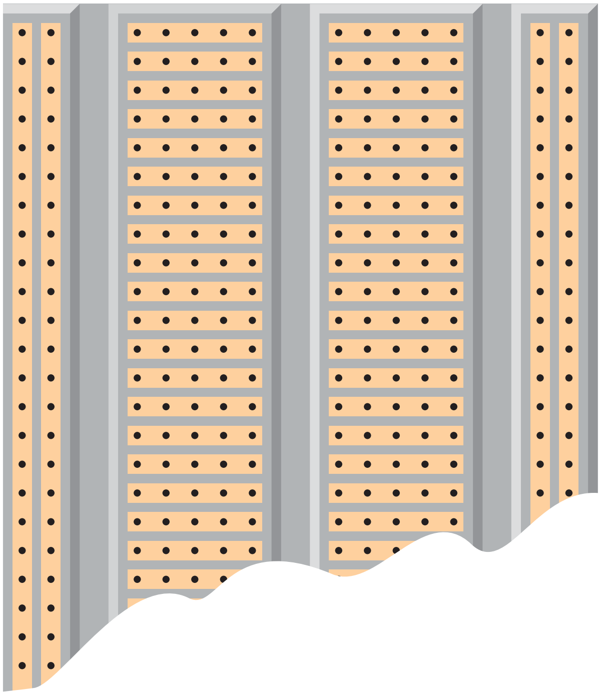

Figure 2-73. *This X-ray-vision view of the breadboard reveals the copper strips that are embedded in it. The strips conduct electricity from one component to another.*

Important note: some breadboards divide each vertical column of holes, on the left and the right, into two separate upper and lower sections. Use your meter's continuity testing feature to find out if your breadboard conducts power along its full length, and add jumper wires to link the upper and lower half of the breadboard if necessary.

Figure 2-74 shows how you can use the breadboard to replicate your oscillating relay circuit. To make this work, you need to apply the positive and negative power from your AC adapter. Because the wire from your AC adapter is almost certainly stranded, you'll have difficulty pushing it into the little holes. A way around this is to set up a couple of pieces of bare 22-gauge wire, and use them as terminals to which you clip the wire from the adapter, as in Figure 2-75. (Yes, you still need just a couple of alligator clips for this purpose.) Alternatively, you can use a breadboard with power terminals built into it, which is more convenient.

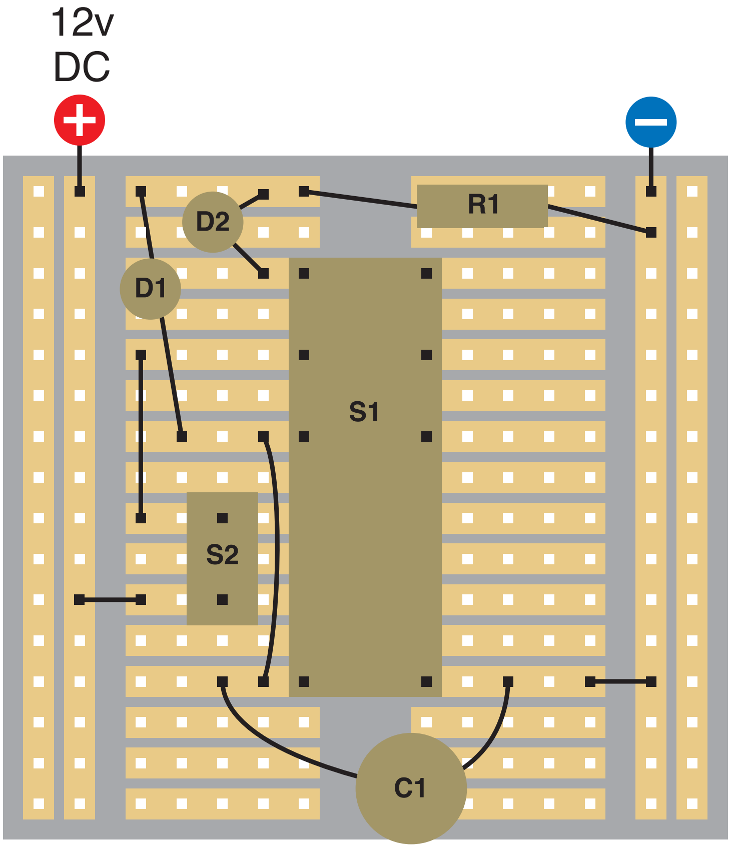

Figure 2-74. *If you place the components on your breadboard in the positions shown, they will create the same circuit that you built from wire and alligator clips in Experiment 8. Component values:*

D1, D2: Light-emitting diodes
S1: DPDT relay
S2: SPST momentary switch
C1: Electrolytic capacitor, 1,000 μF
R1: Resistor, 680Ω minimum

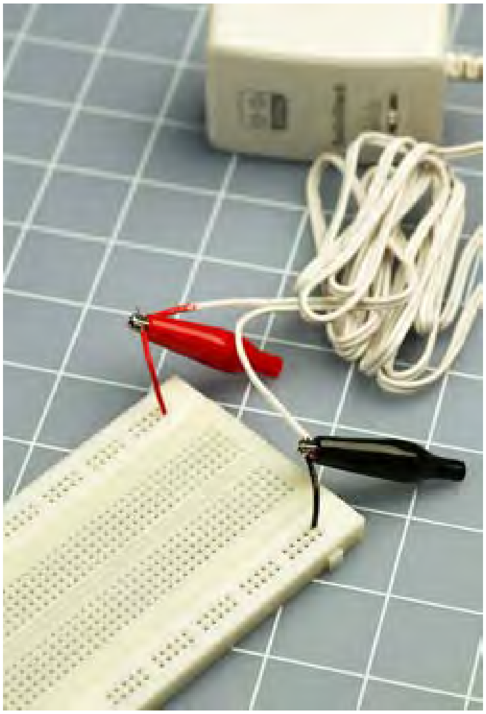

Figure 2-75. *If your breadboard doesn't have screw terminals, insert two short pieces of solid-core wire with stripped ends and then attach the stranded wires from the adapter using alligator clips.*

You'll need some more 22-gauge wire, or some precut hookup wire, to supply the power to your components, which are plugged into the breadboard as shown in Figures 2-76 and 2-77. If you get all the connections right, the circuit should function the same way as before.

The geometry of the metal connecting strips in the breadboard often forces you to connect components in a roundabout way. The pushbutton, for instance, supplies power to the pole of the relay but cannot be connected directly opposite, because there isn't room for it.

Remember that the strips inside the breadboard that don't have any wires or components plugged into them are irrelevant; they don't do anything.

I'll include some suggested breadboard layouts for circuits as you continue through this book, but eventually you'll have to start figuring out breadboard layouts for yourself, as this is an essential part of hobby electronics.

Larger versions of all schematics and breadboard photos are available online at this book's website: http://oreilly.com/catalog/9780596153748.

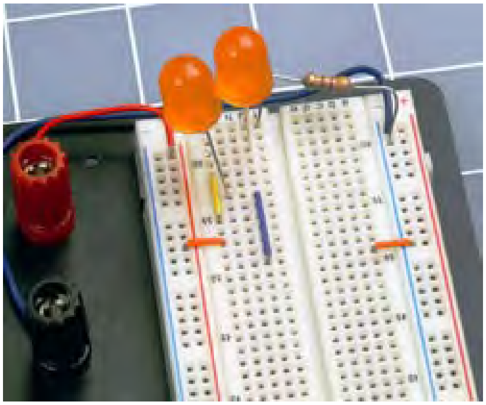

Figure 2-76. *Two oversized LEDs, one resistor, and the necessary jumper wires have been added to the breadboard.*

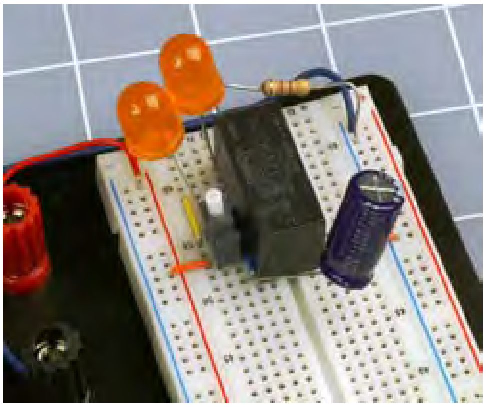

Figure 2-77. *Now the pushbutton, relay, and capacitor have been added to complete the circuit shown in the diagram and the schematic. When the pushbutton is pressed, the relay oscillates and the LEDs flash.*

Experiment 9: Time and Capacitors

You will need:

- AC adapter, breadboard, wire, wire cutters, and strippers.
- Multimeter.
- Pushbutton, SPST. Quantity: 1.
- Resistors and electrolytic capacitors, assorted.

In Experiment 8, when you put a capacitor in parallel with the coil of the relay, the capacitor charged almost instantly before discharging itself through the relay coil. If you add a resistor in series with a capacitor, the capacitor will take longer to charge. By making a capacitor take longer to charge, you can measure time, which is a very important concept.

Clean the components off your breadboard and use it to set up the very simple circuit shown in Figure 2-78, where C1 is a 1,000 μF capacitor, R1 is a 100K resistor, R2 is a 100Ω resistor, and S1 is the pushbutton that you used previously. Set your meter to measure volts DC, place the probes around the capacitor, and hold down the pushbutton. You should see the meter counting upward as the voltage accumulates on the capacitor. (This is easier with a meter that doesn't have autoranging, because you won't have to wait while the meter figures out which range to apply.) Resistor R1 slows the charging time for the capacitor.

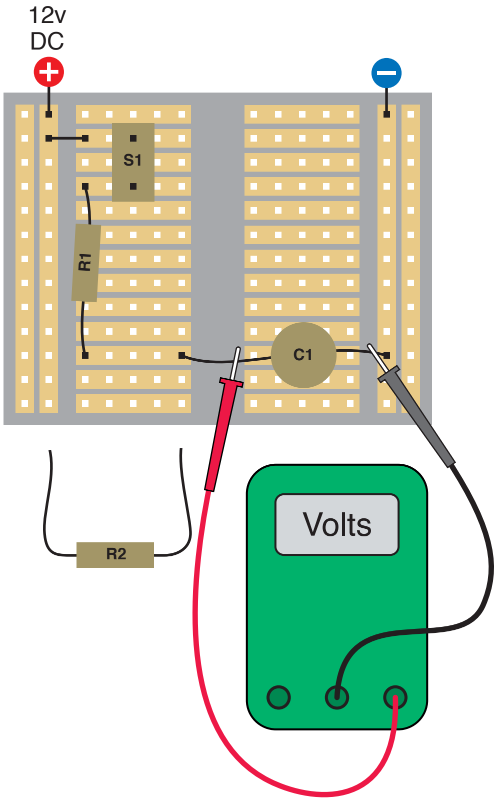

Figure 2-78. *Watch the voltage building up on the capacitor while you hold down the pushbutton. Substitute other values for R1, discharge the capacitor by touching R2 across it, and then repeat your measurement process.*

S1: Momentary pushbutton, OFF (ON)
R1: 100K initially
R2: 100Ω
C1: 1,000 μF

Release the pushbutton, set aside your meter, and discharge the capacitor by touching R2 across it for a second or two. Now substitute a 50K resistor for R1, and repeat the measurement. The meter should count upward almost twice as fast as before.

Voltage, Resistance, and Capacitance

Think of the resistor as a faucet, and the capacitor as a balloon that you are trying to fill with water. When you screw down the faucet until only a trickle comes through, the balloon will take longer to fill. But a slow flow of water will still fill the balloon completely if you wait long enough, and (assuming the balloon doesn't burst) the process ends when the pressure inside the balloon is equal to the water pressure in the pipe supplying the faucet. See Figure 2-79.

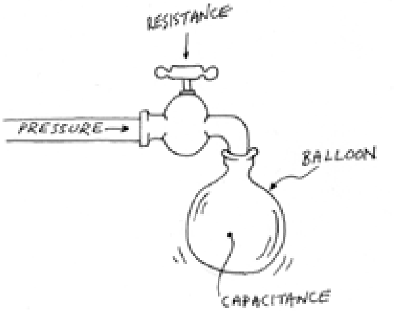

Figure 2-79. *When the faucet is closed half-way, the balloon will take longer to fill, but will still contain as much water and as much pressure in the end.*

Similarly, in your circuit, if you wait long enough, eventually the voltage across the capacitor should reach the same value as the voltage of the power supply. In a 12-volt circuit, the capacitor should eventually acquire 12 volts (although "eventually" may take longer than you think).

This may seem confusing, because earlier you learned that when you apply voltage at one end of a resistor, you get less voltage coming out than you have going in. Why should a resistor deliver the full voltage when it is paired with a capacitor?

Forget the capacitor for a moment, and remember how you tested just two 1K resistors. In that situation, each resistor contained half the total resistance of the circuit, so each resistor dropped half the voltage. If you held the negative probe of your meter against the negative side of your power supply and touched the positive probe to the center point between the two resistors, you would measure 6 volts. Figure 2-80 illustrates this.

Now, suppose you remove one of the 1K resistors and substitute a 9K resistor. The total resistance in the circuit is now 10K, and therefore the 9K resistor drops 90% of the 12 volts. That's 10.8 volts. You should try this and check it with your meter. (You are unlikely to find a 9K resistor, because this is not a standard value. Substitute the nearest value you can find.)

Now suppose you remove the 9K resistor and substitute a 99K resistor. Its voltage drop will be 99% of the available voltage, or 11.88 volts. You can see where this is heading: the larger the resistor, the larger its contribution to voltage drop.

However, I noted previously that a capacitor blocks DC voltage completely. It can accumulate an electrical charge, but no current passes through it. Therefore, a capacitor behaves like a resistor that has infinite resistance to DC current.

(Actually the insulation inside the capacitor allows a little bit of "leakage," but a perfect capacitor would have infinite resistance.)

The value of any resistor that you put in series with the capacitor is trivial by comparison. No matter how high the value of the resistor is, the capacitor still provides much more resistance in the circuit. This means that the capacitor steals almost the complete voltage drop in the circuit, and the voltage difference between one end of the resistor and the other will be zero (assuming that we ignore little imperfections in the components). Figure 2-80 may help to clarify this concept.

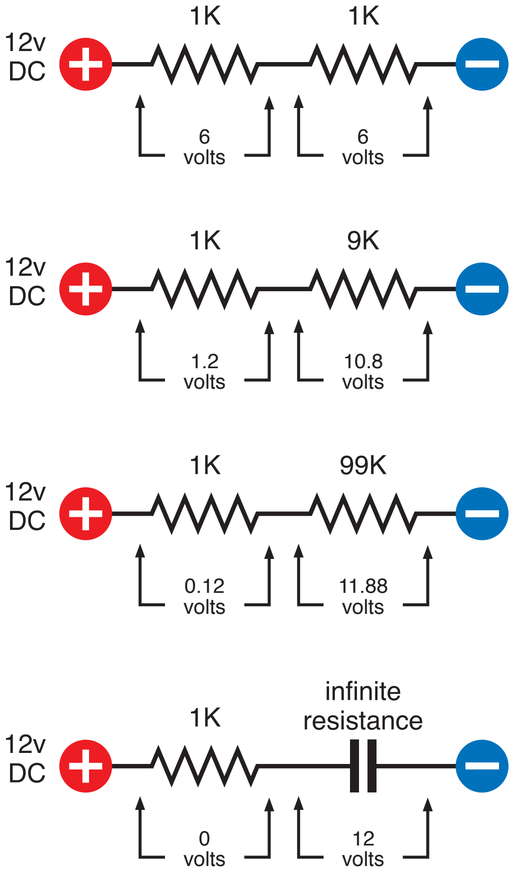

Figure 2-80. *When two resistances are in series, the larger one drops the voltage more than the smaller one. If the larger resistance becomes infinite (as in the case of a capacitor), the smaller one no longer has any measurable contribution to the voltage drop, and the voltage is almost exactly the same at both ends.*

You should try this using real resistors and capacitors—although if you do, you will run into a little problem. When you use your meter in its "DC volts" mode, it diverts a little of the current in the circuit—just a tiny taste—in the process of measuring it. The meter steals such a small amount, it doesn't affect the reading significantly when you are checking voltage across a resistor. The internal resistance of the meter is higher than the values of most resistors. However, remember that the internal resistance of a capacitor is almost infinite. Now the internal resistance of the meter becomes significant. Because you can never have an ideal meter, any more than you can have an ideal capacitor or resistor, your meter will always interfere with the circuit slightly, and you will get only an approximate indication.

If you try to measure the voltage on a capacitor that has been charged but is now not connected to anything else, you'll see the number slowly falling, as the capacitor discharges itself through the meter.

THEORY

The time constant

You may be wondering if there's a way to predict exactly how much time it takes for various capacitors to charge, when they are paired with various resistors. Is there a formula to calculate this?

Of course, the answer is yes, but the way we measure it is a bit tricky, because a capacitor doesn't charge at a constant rate. It accumulates the first volt very quickly, the second volt not quite as quickly, the third volt even less quickly—and so on. You can imagine the electrons accumulating on the plate of a capacitor like people walking into an auditorium and looking for a place to sit. The fewer seats that are left, the longer people take to find them.

The way we describe this is with something called a "time constant." The definition is very simple:

$$TC = R \times C$$

where TC is the time constant, and a capacitor of C farads is being charged through a resistor of R ohms.

Going back to the circuit you just tested, try using it again, this time with a 1K resistor and the 1,000 µF capacitor. We have to change those numbers to farads and ohms before we can put them in the formula. Well, 1,000 µF is 0.001 farads, and 1K is 1,000 ohms, so the formula looks like this:

$$TC = 1{,}000 \times 0.001$$

In other words, TC = 1—a lesson that could not be much easier to remember:

> ***A 1K resistor in series with a 1,000 µF capacitor has a time constant of 1.***

Does this mean that the capacitor will be fully charged in 1 second? No, it's not that simple. TC, the time constant, is the time it takes for a capacitor to acquire *63%* of the voltage being supplied to it, if it starts with zero volts.

(Why 63%? The answer to that question is too complicated for this book, and you'll have to read about time constants elsewhere if you want to know more. Be prepared for differential equations.) Here's a formal definition for future reference:

> TC, the time constant, is the time it takes for a capacitor to acquire 63% of the difference between its current charge and the voltage being applied to it. When TC=1, the capacitor acquires 63% of its full charge in 1 second. When TC=2, the capacitor acquires 63% of its full charge in 2 seconds. And so on.

What happens if you continue to apply the voltage? History repeats itself. The capacitor accumulates *another* 63% of the *remaining* difference between its current charge, and the voltage being applied to it.

Imagine someone eating a cake. In his first bite he's ravenously hungry, and eats 63% of the cake in one second. In his second bite, not wanting to seem too greedy, he takes just another 63% of the cake that is left—and because he's not feeling so hungry anymore, he requires the same time to eat it as he took to eat the first bite. In his third bite, he takes 63% of what still remains, and still takes the same amount of time. And so on. He is behaving like a capacitor eating electricity (Figure 2-81).

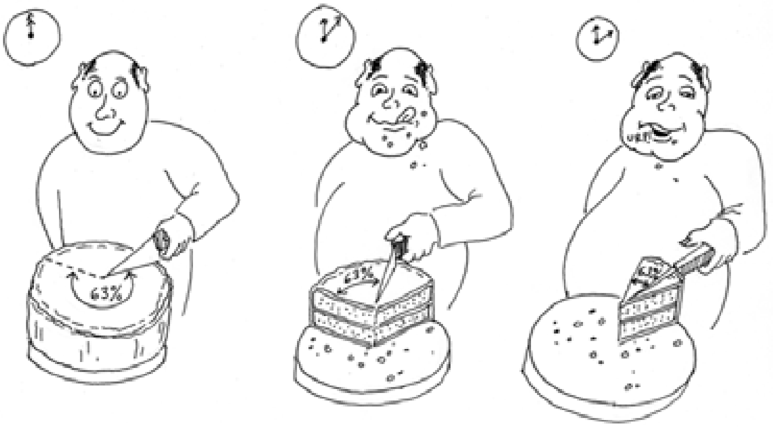

Figure 2-81. *If our gourmet always eats just 63% of the cake still on the plate, he "charges up" his stomach in the same way that a capacitor charges itself. No matter how long he keeps at it, his stomach is never completely filled.*

THEORY

The time constant (continued)

The cake eater will always have a few crumbs to eat, because he never takes 100% of the remainder. Likewise, the capacitor will never acquire a full charge. In a perfect world of perfect components, this process would continue for an infinite time.

In the real world, we say rather arbitrarily:

> After 5 × TC the capacitor will be so nearly fully charged, we won't care about the difference.

In the table is a calculation (rounded to two decimal places) showing the charge accumulating on a capacitor in a 12-volt circuit where the time constant is 1 second.

Here's how to understand the table. V1 is the current charge on the capacitor. Subtract this from the supply voltage (12 volts) to find the difference. Call the result V2. Now take 63% of V2, and add this to the current charge (V1) and call the result V4. This is the new charge that the capacitor will have after 1 second, so we copy it down to the next line in the table, and it becomes the new value for V1.

Now we repeat the same process all over again. Figure 2-82 shows this in graphical form. Note that after 5 seconds, the capacitor has acquired 11.92 volts, which is 99% of the power supply voltage. This should be close enough to satisfy anyone's real-world requirements.

If you try to verify these numbers by measuring the voltage across the capacitor as it charges, remember that because your meter steals a little current, there will be a small discrepancy that will increase as time passes. For practical purposes, the system works well enough.

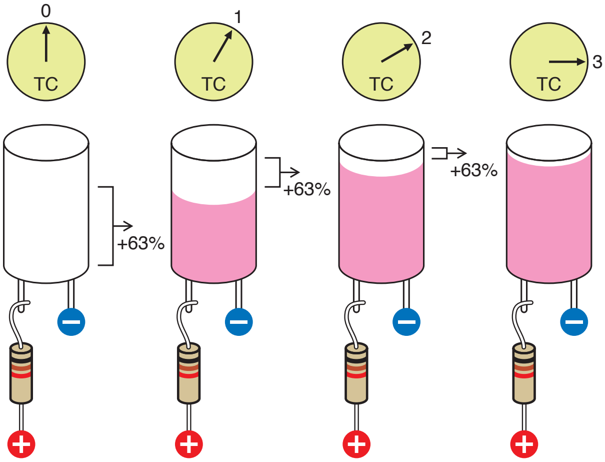

Figure 2-82. *A capacitor starts with 0 volts. After 1 time constant it adds 63% of the available voltage. After another time constant, it adds another 63% of the remaining voltage difference, and so on.*

Time in secs	V1: Charge on capacitor	V2: 12 – V1	V3: 63% of V2	V4: V1 + V3
0	0.00	12.00	7.56	7.56
1	7.56	4.44	2.80	10.36
2	10.36	1.64	1.03	11.39
3	11.39	0.61	0.38	11.77
4	11.77	0.23	0.15	11.92
5	11.92			

Experiment 10: Transistor Switching

You will need:

- AC adapter, breadboard, wire, and meter.
- LED. Quantity: 1.
- Resistors, various.
- Pushbutton, SPST. Quantity: 1.
- Transistor, 2N2222 or similar. Quantity: 1.

A transistor can switch a flow of electricity, just like a relay. But it's much more sensitive and versatile, as this first ultra-simple experiment will show.

We'll start with the 2N2222 transistor, which is the most widely used semiconductor of all time (it was introduced by Motorola in 1962 and has been in production ever since).

First, you should get acquainted with the transistor. Because Motorola's patents on the 2N2222 ran out long ago, any company can manufacture their own version of it. Some versions are packaged in a little piece of black plastic; others are enclosed in a little metal "can." (See Figure 2-83.) Either way, it contains a piece of silicon divided into three sections known as the collector, the base, and the emitter. I'll describe their function in more detail in a moment, but initially you just need to know that in this type of transistor, the collector receives current, the base controls it, and the emitter sends it out.

Use your breadboard to set up the circuit shown in Figure 2-85. Be careful to get the transistor the right way around! (See Figure 2-84.) For the three brands I have mentioned in the shopping list, the flat side should face right, if the transistor is packaged in black plastic, or the little tab should face toward the lower left, if the transistor is packaged in metal.

Figure 2-83. *A typical transistor is packaged either in a little metal can or a molded piece of black plastic. The manufacturer's data sheet tells you the identities of the three wire leads, relative to the flat side of a black plastic transistor or the tab that sticks out of a metal-can transistor.*

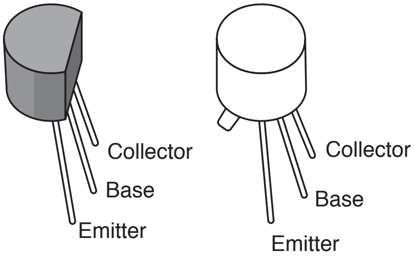

Figure 2-84. *The 2N2222 transistor may be packaged in either of these formats. Left: RadioShack or Fairchild. Right: STMicroelectronics (note the little tab sticking out at the lower-left side). If you use a different brand, you'll have to check the manufacturer's data sheet. Insert the transistor in your breadboard with the flat side facing right, as seen from above, or the tab pointing down and to the left, seen from above.*

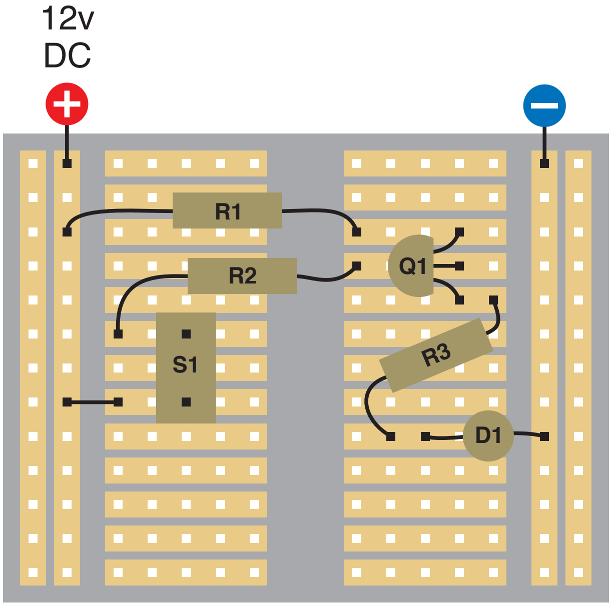

Figure 2-85. *The transistor blocks voltage that reaches it through R1. But when pushbutton S1 is pressed, this tells the transistor to allow current to pass through it. Note that transistors are always identified with letter Q in wiring diagrams and schematics.*

S1: Pushbutton, momentary, OFF (ON)
R1: 180Ω
R2: 10K
R3: 680Ω
Q1: 2N2222 or similar
D1: LED

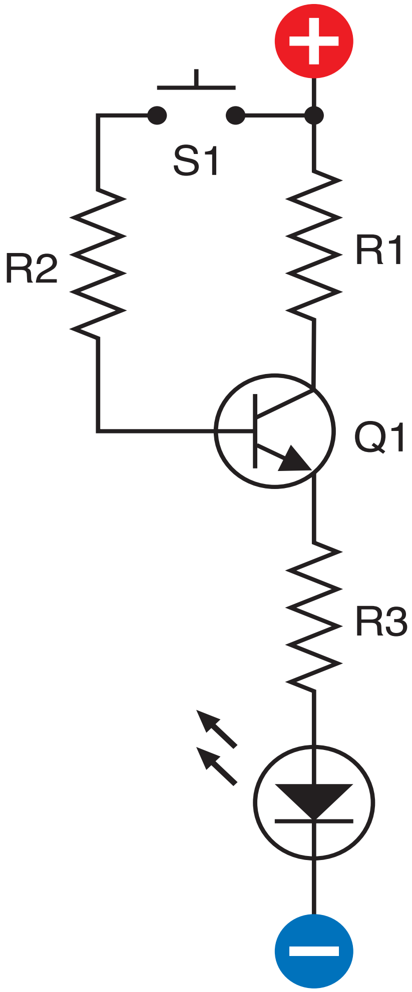

Figure 2-86. *This shows the same circuit as the breadboard diagram in Figure 2-85.*

Initially, the LED should be dark. Now press the pushbutton and the LED should glow brightly. Electricity is following two paths here. Look at the schematic in Figure 2-86, which shows the same circuit more clearly. I've shown positive at the top and negative at the bottom (the way most schematics do it) because it helps to clarify the function of this particular circuit. If you view the schematic from the side, the similarity with the breadboard layout is easier to see.

Through R1, voltage reaches the top pin (the collector) of the transistor. The transistor only lets a tiny trickle of it pass through, so the LED stays dark. When you press the button, voltage is also applied along a separate path, through R2 to the middle pin (the base) of the transistor. This tells the transistor to open its solid-state switch and allow current to flow out through its third pin (the emitter), and through R3, to the LED.

You can use your meter in volts DC mode to check the voltage at points in the circuit. Keep the negative probe from the meter touching the negative voltage source while you touch the positive probe to the top pin of the transistor, the middle pin, and the bottom pin. When you press the button, you should see the voltage change.

Fingertip Switching

Now here's something more remarkable. Remove R2 and the pushbutton, and insert two short pieces of of wire as shown in Figure 2-87. The upper piece of wire connects with the positive voltage supply; the lower piece connects with the middle pin of the transistor (its base). Now touch the tip of your finger to the two wires. Once again, the LED should glow, although not as brightly as before. Lick the tip of your finger, try again, and the LED should glow more brightly.

Never Use Two Hands

The fingertip switching demo is safe if the electricity passes just through your finger. You won't even feel it, because it's 12 volts DC from a power supply of 1 amp or less. But it's not a good idea to put the finger of one hand on one wire, and the finger of your other hand on the other wire. This would allow the electricity to pass through your body. Although the chance of hurting yourself this way is extremely small, you should never allow electricity to run through you from one hand to the other. Also, when touching the wires, don't allow them to penetrate your skin.

Your finger is conducting positive voltage to the base of the transistor. Even though your skin has a high resistance, the transistor still responds. It isn't just switching the LED on and off; it is *amplifying the current* applied to its base. This is an essential concept: *a transistor amplifies any changes in current that you apply to its base*.

Check Figure 2-88 to see more clearly what's happening.

If you studied the section "Background: Positive and negative" in Chapter 1, you learned that there is really no such thing as positive voltage. All we really have is negative voltage (created by the pressure of free electrons) and an absence of negative voltage (where there are fewer free electrons). But because the idea of a flow of electricity from positive to negative was so widely believed before the electron was discovered, and because the inner workings of a transistor involve "holes" which are an absence of electrons and can be thought of as positive, we can still pretend that electricity flows from positive to negative. See the following section, "Essentials: All about NPN and PNP transistors," for more details.

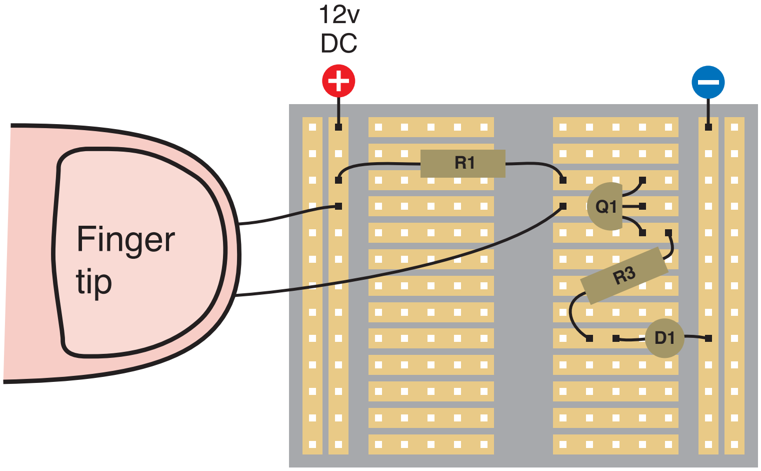

Figure 2-87

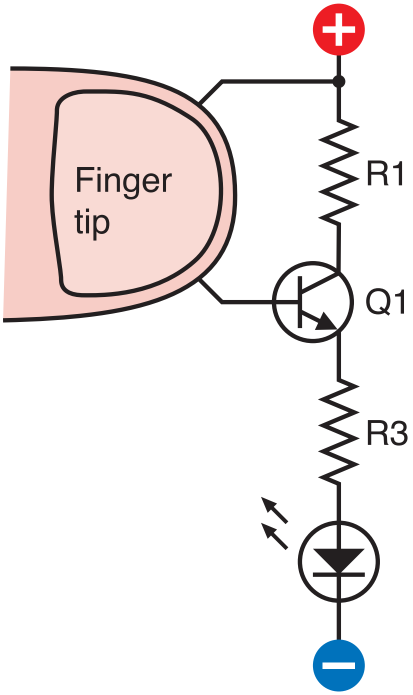

Figure 2-88. *These two diagrams show the same components as before, with a fingertip substituted for R2. Although only a trickle of voltage now reaches the base of the transistor, it's enough to make the transistor respond.*

ESSENTIALS

All about NPN and PNP transistors

A transistor is a semiconductor, meaning that sometimes it conducts electricity, and sometimes it doesn't. Its internal resistance varies, depending on the power that you apply to its base.

NPN and PNP transistors are bipolar semiconductors. They contain two slightly different variants of silicon, and conduct using both polarities of carriers—holes and electrons.

The NPN type is a sandwich with P-type silicon in the middle, and the PNP type is a sandwich with N-type silicon in the middle. If you want to know more about this terminology, and the behavior of electrons when they try to cross an NP junction or a PN junction, you'll have to read a separate source on this subject. It's too technical for this book. All you need to remember is:

- All bipolar transistors have three connections: Collector, Base, and Emitter, abbreviated as C, B, and E on the manufacturer's data sheet, which will identify the pins for you.
- NPN transistors are activated by *positive* voltage on the base relative to the emitter.
- PNP transistors are activated by *negative* voltage on the base relative to the emitter.

In their passive state, both types block the flow of electricity between the collector and emitter, just like an SPST relay in which the contacts are normally open. (Actually a transistor allows a tiny bit of current known as "leakage.")

You can think of a bipolar transistor as if it contains a little button inside, as shown in Figures 2-89 and 2-90. When the button is pressed, it allows a large current to flow. To press the button, you inject a much smaller current into the base by applying a small voltage to the base. In an NPN transistor, the control voltage is positive. In a PNP transistor, the control voltage is negative.

NPN transistor basics

- To start the flow of current from collector to emitter, apply a relatively positive voltage to the base.
- In the schematic symbol, the arrow points from base to emitter and shows the direction of positive current.
- The base must be at least 0.6 volts "more positive" than the emitter, to start the flow.
- The collector must be "more positive" than the emitter.

PNP transistor basics

- To start the flow of current from emitter to collector, apply a relatively negative voltage to the base.
- In the schematic symbol, the arrow points from emitter to base and shows the direction of positive current.
- The base must be at least 0.6 volts "more negative" than the emitter, to start the flow.
- The emitter must be "more positive" than the collector.

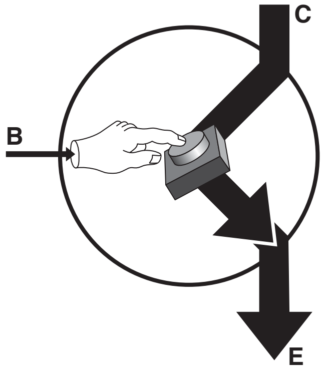

Figure 2-89. *You can think of a bipolar transistor as if it contains a button that can connect the collector and the emitter. In an NPN transistor, a small positive potential presses the button.*

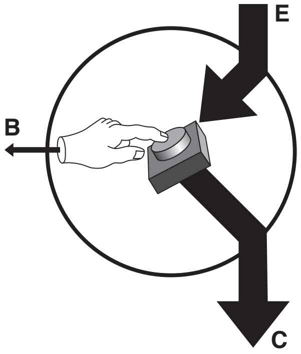

Figure 2-90. *In a PNP transistor, a small negative potential has the same effect. The arrows point in the direction of "positive current flow."*

ESSENTIALS

All about NPN and PNP transistors (continued)

All-transistor basics

- Never apply a power supply directly across a transistor. You can burn it out with too much current.
- Protect a transistor with a resistor, in the same way you would protect an LED.
- Avoid reversing the connection of a transistor between positive and negative voltages.
- Sometimes an NPN transistor is more convenient in a circuit; sometimes a PNP happens to fit more easily. They both function as switches and amplifiers, the only difference being that you apply a relatively positive voltage to the base of an NPN transistor, and a relatively negative voltage to the base of a PNP transistor.
- PNP transistors are used relatively seldom, mainly because they were more difficult to manufacture in the early days of semiconductors. People got into the habit of designing circuits around NPN transistors.
- Remember that bipolar transistors amplify current, not voltage. A small fluctuation of current through the base enables a large change in current between emitter and collector.
- Schematics sometimes show transistors with circles around them, and sometimes don't. In this book, I'll use circles to draw attention to them. See Figures 2-91 and 2-92.
- Schematics may show the emitter at the top and the collector at the bottom, or vice versa. The base may be on the left, or on the right, depending on what was most convenient for the person drawing the schematic. Be careful to look carefully at the arrow in the transistor to see which way up it is, and whether it is NPN or PNP. You can damage a transistor by connecting it incorrectly.
- Transistors come in various different sizes and configurations. In many of them, there is no way to tell which wires connect to the emitter, the collector, or the base, and some transistors have no part numbers on them. Before you throw away the packaging that came with a transistor, check to see whether it identifies the terminals.
- If you forget which wire is which, some multimeters have a function that will identify emitter, collector, and base for you. Check your multimeter instruction booklet for more details.

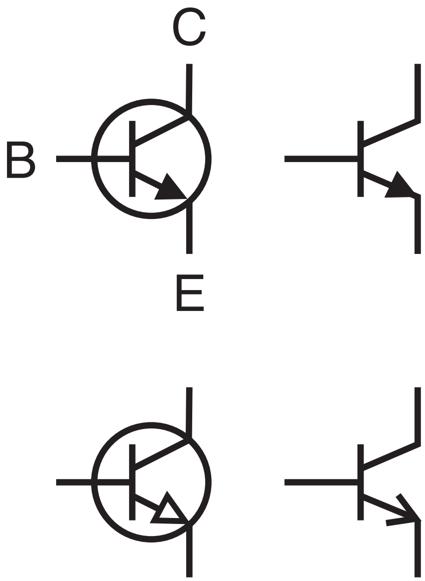

Figure 2-91. *The symbol for an NPN transistor always has an arrow pointing from its base to its emitter. Some people include a circle around the transistor; others don't bother. The style of the arrow may vary. But the meaning is always the same. The top-left version is the one I use in this book.*

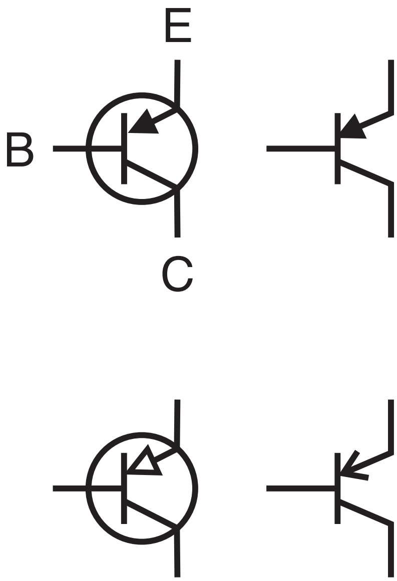

Figure 2-92. *The symbol for a PNP transistor always has an arrow pointing from its emitter to its base. Some people include a circle around the transistor; others don't bother. The style of the arrow may vary. But the meaning is always the same. The top-left version is the one I use in this book.*

BACKGROUND

Transistor origins

Though some historians trace the origins of the transistor back to the invention of diodes (which allow electricity to flow in one direction while preventing reversal of the flow), there's no dispute that the first working transistor was developed at Bell Laboratories in 1948 by John Bardeen, William Shockley, and Walter Brattain (Figure 2-93).

Shockley was the leader of the team, who had the foresight to see how potentially important a solid-state switch could be. Bardeen was the theorist, and Brattain actually made it work. This was a hugely productive collaboration—until it succeeded. At that point, Shockley started maneuvering to have the transistor patented exclusively under his own name. When he notified his collaborators, they were—naturally—unhappy about this idea.

A widely circulated publicity photograph didn't help, in that it showed Shockley sitting at the center in front of a microscope, as if he had done the hands-on work, while the other two stood behind him, implying that they had played a lesser role. In fact Shockley, as the supervisor, was seldom present in the laboratory where the real work was done.

The productive collaboration quickly disintegrated. Brattain asked to be transferred to a different lab at AT&T. Bardeen moved to the University of Illinois to pursue theoretical physics. Shockley eventually left Bell Labs and founded Shockley Semiconductor in what was later to become Silicon Valley, but his ambitions outstripped the capabilities of the technology of his time. His company never manufactured a profitable product.

Eight of Shockley's coworkers in his company eventually betrayed him by quitting and establishing their own business, Fairchild Semiconductor, which became hugely successful as a manufacturer of transistors and, later, integrated circuit chips.

Figure 2-93. *Photographs provided by the Nobel Foundation show, left to right, John Bardeen, William Shockley, and Walter Brattain. For their collaboration in development of the world's first working transistor in 1948, they shared a Nobel prize in 1956.*

ESSENTIALS

Transistors and relays

One limitation of NPN and PNP transistors is that they are naturally "off" until you turn them "on." They behave like a normally open pushbutton, which conducts electricity only for as long as you hold it down. They don't normally behave like a normal on switch, which stays on until you apply a signal to turn it off.

A relay offers more switching options. It can be normally open, normally closed, or it can contain a double-throw switch, which gives you a choice of two "on" positions. It can also contain a double-pole switch, which makes (or breaks) two entirely separate connections when you energize it. Single-transistor devices cannot provide the double-throw or double-pole features, although you can design more complex circuits that emulate this behavior.

Here's a list of transistor and relay characteristics.

	Transistor	Relay
Long-term reliability	Excellent	Limited
Configurable for DP and DT switching	No	Yes
Ability to switch large currents	Limited	Good
Able to switch alternating current (AC)	Usually not	Yes
Can be triggered by alternating current (AC)	Usually not	Optional
Suitability for miniaturization	Excellent	Very limited
Sensitive to heat	High	Moderate
Ability to switch at high speed	Excellent	Limited
Price advantage for low-voltage low-current	Yes	No
Price advantage for high-voltage high-current	No	Yes
Current leakage when "off"	Yes	No

The choice between relays or transistors will depend on each particular application.

THEORY

See the current

If you want to get a more precise understanding of how a transistor works, you should try this little test. It shows the precise behavior and limits of the 2N2222 transistor that you used in the previous experiment.

I've said that in an NPN transistor, the collector should always be more positive than the emitter and that the base should have a potential somewhere between those two voltages. Figure 2-94 shows this rather vague relationship. Now I want to substitute some numbers for these general statements.

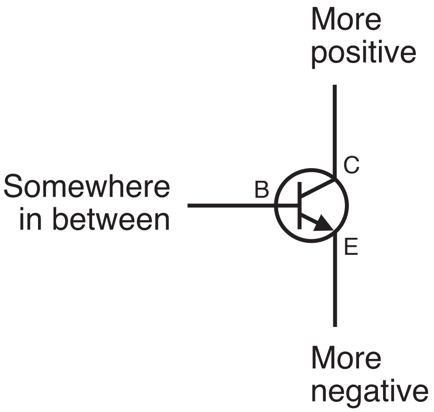

Figure 2-94. *The proper functioning of an NPN transistor requires you to maintain these voltage relationships.*

Take a look at the schematic in Figure 2-95, and check the component values. Notice that the total resistance above the transistor, from R1 + R2, is the same as the total resistance below it, from R3 + R4. Therefore the potential on the base of the transistor should be halfway between the two extremes—until you use potentiometer P1 to adjust the voltage of the base of the transistor up and down.

The two 180Ω resistors, R1 and R3, protect the transistor from passing excessive current. The two 10K resistors, R2 and R4, protect the base when the potentiometer is turned all the way up or all the way down.

I would like you to see what the transistor is doing by measuring the amperage flowing into the base at the position marked A1, and the total amperage flowing out through the emitter at the position marked A2. To do this, it would be really helpful if you had two meters. As that may be impractical, the breadboard diagrams in Figures 2-96 and 2-97 show how you can swap one meter between the two locations.

Remember that to measure milliamps, you have to pass electricity through the meter. This means that the meter must be inserted into the circuit, and whenever you remove the meter, you have to remake the connection where the meter was. The breadboard diagram shows how you can do this. Fortunately, it's very easy to remove and replace wires in a breadboard. Where wires are connected to the potentiometer, you may need to revert to using alligator clips.

Begin with the potentiometer turned about halfway through its range. Measure at A1 and A2. Turn the potentiometer up a bit, and measure current at the two locations again. Following is a table showing some actual readings I obtained at those two locations, using two digital meters simultaneously.

Milliamps passing through location A1	Milliamps passing through location A2
0.01	1.9
0.02	4.9
0.03	7.1
0.04	9.9
0.05	12.9
0.06	15.5
0.07	17.9
0.08	19.8
0.09	22.1
0.10	24.9
0.11	26.0
0.12	28.3

There's a very obvious relationship. The current emerging from the emitter of the transistor, through location A2, is about 24 times the current passing through location A1, into the base. The ratio of current coming out from the emitter of an NPN transistor to current going into the base is known as the *beta value* for a transistor. The beta value expresses the transistor's amplifying power.

It's a very constant ratio, until you push it a little too far. Above 0.12 mA, this particular transistor becomes "saturated," meaning that its internal resistance cannot go any lower.

THEORY

See the current (continued)

In my little experiment, I found that the maximum current at A2 was 33mA. A simple calculation using Ohm's Law showed me that this meant the transistor's internal resistance was near zero. This is why you should protect a transistor with some additional resistance in the circuit. If you don't, its low internal resistance would allow a huge current flow that would immediately burn it out.

What about the other end of its range? When it passes only 1.9 mA, the transistor has an internal resistance of around 6,000Ω. The conclusion is that depending how much current you apply to this transistor, its internal resistance varies between zero and 6,000Ω, approximately.

So much for the theory. Now what can we do with a transistor that's fun, or useful, or both? We can do Experiment 11!

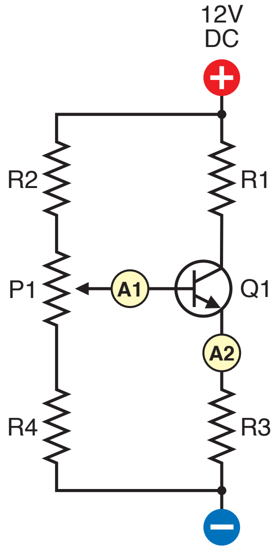

Figure 2-95. *This is basically the same as the previous circuit, with a potentiometer added and the LED removed. Component values:*

R1: 180Ω
R2: 10K
R3: 180Ω
R4: 10K
P1: 1M linear potentiometer
Q1: 2N2222 transistor

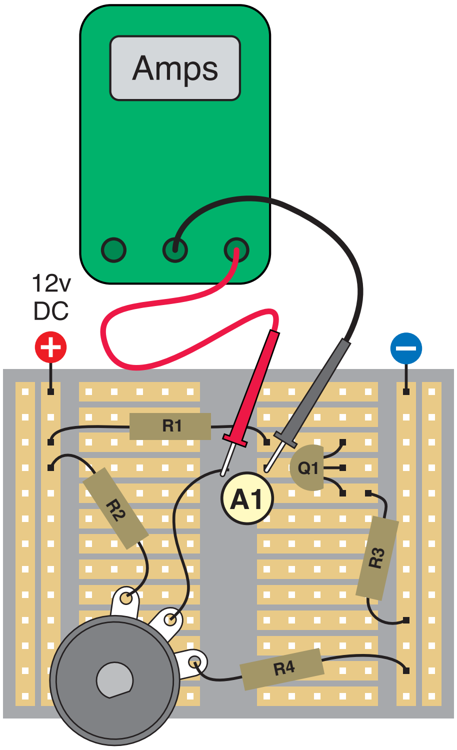

Figure 2-96. *The meter is measuring current flowing from the potentiometer into the base of the transistor at position A1 (see Figure 2-95).*

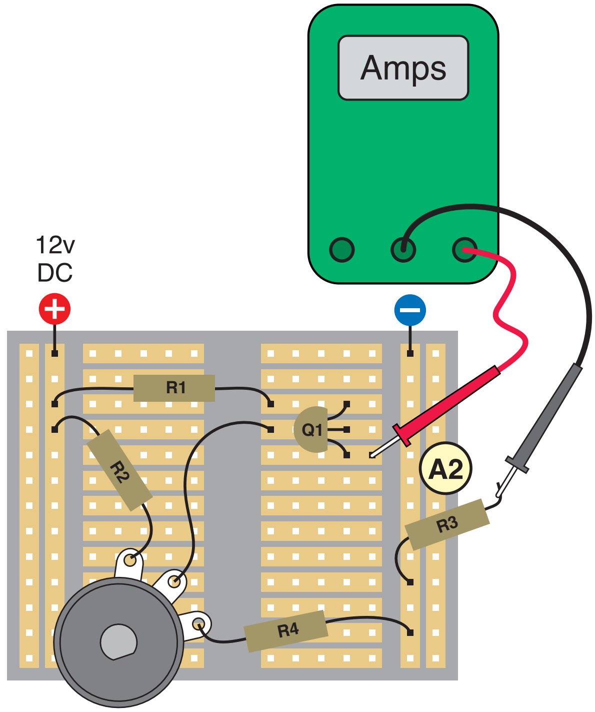

Figure 2-97. *One end of resistor R3 has been unplugged from the breadboard so that the meter now measures current flowing out through the emitter of the transistor, into R3, at position A2.*

Experiment 11: A Modular Project

You will need:

- AC adapter, breadboard, wire, and meter.
- LED. Quantity: 1.
- Resistors, various.
- Capacitors, various.
- Transistor, 2N2222 or similar. Quantity: 2.
- 2N6027 programmable unijunction transistor (PUT). Quantity: 2.
- Miniature 8Ω loudspeaker. Quantity: 1.

So far, I've described small circuits that perform very simple functions. Now it's time to show how modules can be combined to create a device that does a bit more.

The end product of this experiment will be a circuit that makes a noise like a small siren, which could be used in an intrusion alarm. You may or may not be interested in owning an alarm, but the four-step process of developing it is important, because it shows how individual clusters of components can be persuaded to communicate with each other.

I'll begin by showing how to use a transistor to make a solid-state version of the oscillating circuit that you built with a relay in Experiment 8. The relay, you may remember, was wired in such a way that the coil received power through the contacts of the relay. As soon as the coil was energized, it opened the contacts, thus cutting off its own power. As soon as the contacts relaxed they restored the power, and the process repeated itself.

There's no way to do this with a single bipolar transistor. You actually need two of them, switching each other on and off, and the way that this works is quite hard to understand. An easier option is to use a different thing known as a programmable unijunction transistor, or PUT.

Unijunction transistors were developed during the 1950s, but fell into disuse when simple silicon chips acquired the ability to perform the same kinds of functions, more accurately and more cheaply. However, the so-called programmable unijunction transistor is still widely available, often used in applications such as lamp dimmers and motor controllers. Because its primary use is in generating a stream of pulses, it's ideal for our purposes.

If you put together the components shown in Figure 2-98, the LED should start flashing as soon as you apply power.

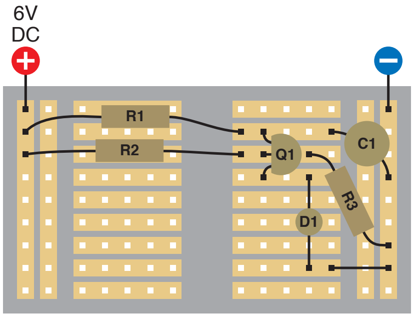

Figure 2-98. *Assemble these components, apply power, and the LED should start flashing.*

R1: 470K
R2: 15K
R3: 27K
C1: 2.2 μF electrolytic capacitor
D1: LED
Q1: 2N6027 programmable unijunction transistor

Note that this circuit will work on 6 volts. You won't damage anything if you run it with 12 volts, but as we continue adding pieces to it, you'll find that it actually performs better at 6 volts than at 12. If you read the next section, "Essentials: All about programmable unijunction transistors," you'll find out how the circuit works.

ESSENTIALS

All about programmable unijunction transistors

The schematic symbol for a programmable unijunction transistor, or PUT, looks very different from the symbol for a bipolar transistor, and its parts are named differently, too. Nevertheless, it does have a similar function as a solid-state switch. The symbol and the names of the three connections are shown in Figure 2-99.

Note that this is a rare case (maybe the only one in the whole of electronics!) in which you won't run into confusing variations of the basic schematic symbol. A PUT always seems to look the way I've drawn it here. Personally I think it would be clearer if we added a circle around it, but no one seems to do that, so I won't, either.

The 2N6027 is probably the most common PUT, and seems to be standardized in its packaging and pin-outs. I've only seen it in a plastic module rather than a little tin can. Figure 2-100 shows the functions of the leads if your 2N6027 is manufactured by Motorola or On Semiconductor. If you have one from another source, you should check the data sheet.

Note that the flat side of the plastic module faces the opposite way around compared with the 2N2222 bipolar transistor, when the two devices are functioning similarly.

The PUT blocks current until its internal resistance drops to allow flow from the "anode" to the "cathode." In this way, it seems very similar to an NPN transistor, but there's a big difference in the circumstances that cause the PUT to lower its resistance. The voltage at the anode determines when the PUT allows current to flow.

Suppose you start with, say, 1 volt at the anode. Slowly, you increase this voltage. The transistor blocks it until the anode gets close to 6 volts. Suddenly this pressure breaks down the resistance and current surges from the anode to the cathode. If the voltage goes back down again, the transistor reverts to its original state and blocks the flow.

I've included another version of the "finger on the button" drawing to convey this concept. The voltage on the anode is itself responsible for pushing the button that opens the pathway to the cathode. See Figure 2-101.

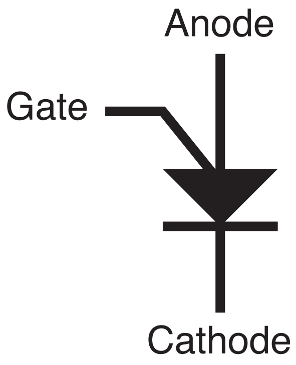

Figure 2-99. *The schematic symbol for a PUT.*

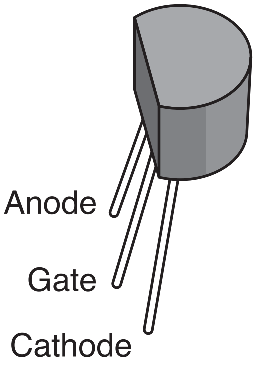

Figure 2-100. *In PUTs manufactured by On Semiconductor and Motorola, the leads have these functions.*

ESSENTIALS

All about programmable unijunction transistors (continued)

This may cause you to wonder what the function of the gate is. You can think of it as "assisting" the finger on the button. In fact, the gate is the "programmable" part of a PUT. By choosing a voltage for the gate, you establish the threshold point when current starts to flow.

Here's a simple take-home summary:

- The anode has to be more positive than the cathode, and the gate should be between those two extremes.
- If anode voltage increases above a threshold point, current bursts through and flows from the anode to the cathode.
- If anode voltage drops back down below the threshold, the transistor stops the flow.
- The voltage you apply to the gate determines how high the threshold is.
- The gate voltage is adjusted with two resistors, shown as R1 and R2 in the simple schematic in Figure 2-102. Typically, each resistor is around 20K. The PUT is protected from full positive voltage by R3, which can have a high value, 100K or greater, because very little current is needed to bias the transistor.
- You add your input signal in the form of positive voltage at the anode. When it exceeds the threshold, it flows out of the cathode and can work some kind of output device.

The only remaining question is how we make a PUT oscillate, to create a stream of on/off pulses. The answer is the capacitor that you included in the circuit that you breadboarded at the beginning of Experiment 11.

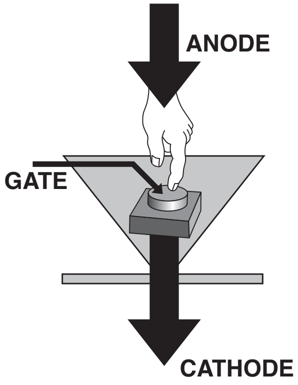

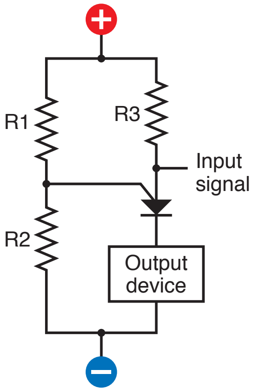

Figure 2-101. *When voltage at the anode of a PUT crosses a threshold (determined by a preset voltage at the gate), current breaks through and surges from the anode to the cathode. In this sense, the anode voltage acts as if it presses a button itself to open a connection inside the PUT, with some assistance from control voltage at the gate.*

Figure 2-102. *This simple schematic shows how a PUT is used. R1 and R2 determine the voltage at the gate, which sets the threshold point for the input at the anode. Above the threshold, current flows from anode to cathode.*

Step 1: Slow-Speed Oscillation

Figure 2-103 is a schematic version of the previous PUT breadboard circuit shown in Figure 2-98, drawn so that the layout looks as much like the breadboard as possible.

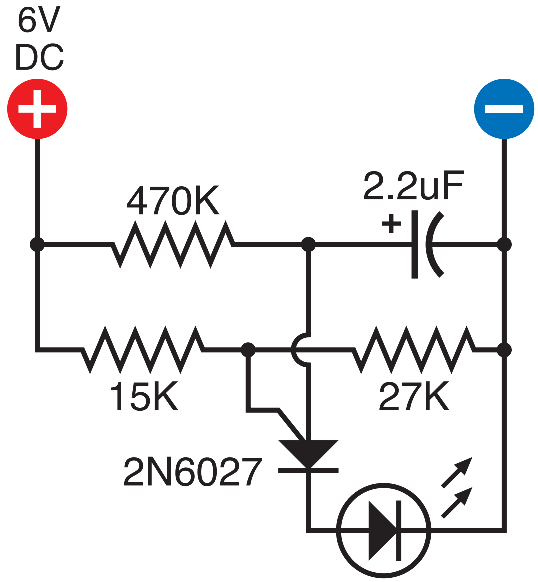

Figure 2-103. *This makes it easier to see what's happening in the breadboard version.*

The 15K resistor and 27K resistor establish the voltage at the gate. The 470K resistor supplies the anode of the PUT, but the PUT begins in its "off" condition, blocking the voltage. So the voltage starts to charge the 2.2 μF capacitor.

You may remember that a resistor slows the rate at which a capacitor accumulates voltage. The bigger the resistor and/or the larger the capacitor, the longer the capacitor takes to reach a full charge. In this circuit, the capacitor takes about half a second to get close to 6 volts.

But notice that the PUT is connected directly with the capacitor. Therefore, whatever voltage accumulates on the capacitor is also experienced by the PUT. As the voltage gradually increases, finally it reaches the threshold, which flips the PUT into its "on" state. The capacitor immediately discharges itself through the PUT, through the LED (which flashes), and from there to the negative side of the power supply.

The surge depletes the capacitor. The voltage drops back down, and the PUT returns to its original state. Now the capacitor has to recharge itself all over again, until the whole process repeats itself.

If you substitute a 22 μF capacitor, the charge/discharge cycle should take about 10 times as long, which will give you time to measure it. Set your meter to measure volts DC and place its probes on either side of the capacitor. You can actually watch the charge increasing until it reaches the threshold, at which point the capacitor discharges and the voltage drops back down again.

So now we have an oscillator. What's next?

Step 2: Beyond the Persistence of Vision

If you substitute a much smaller capacitor, it will charge much more quickly, and the LED will flash faster. Suppose you use a capacitor of 0.0047 µF (which can also be expressed as 47 nanofarads, or 47 nF). This seems like an odd number, but it's a standard value for a capacitor. This will reduce the capacitance by a factor of more than 500, and therefore the LED should flash about 500 times as fast, which should be about 1,000 times per second. The human eye cannot detect such rapid pulses. The human ear, however, can hear frequencies up to 10,000 per second and beyond. If we substitute a miniature loudspeaker for the LED, we should be able to hear the oscillations.

Figure 2-104 shows how I'd like you to make this happen. Please leave your original, slow-flashing circuit untouched, and make a duplicate of it farther down the breadboard, changing a couple of component values as indicated. In the schematic in Figure 2-105, the new part of the circuit is in solid black, while the previous section is in gray.

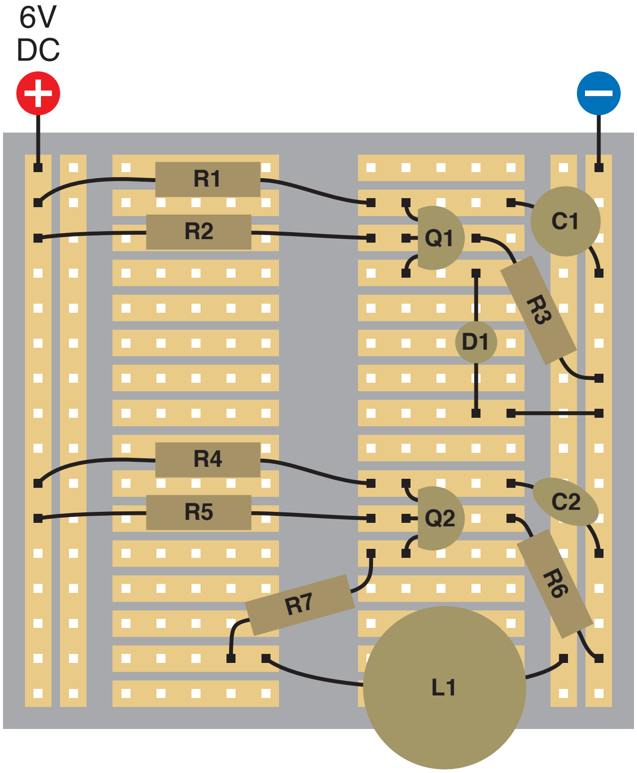

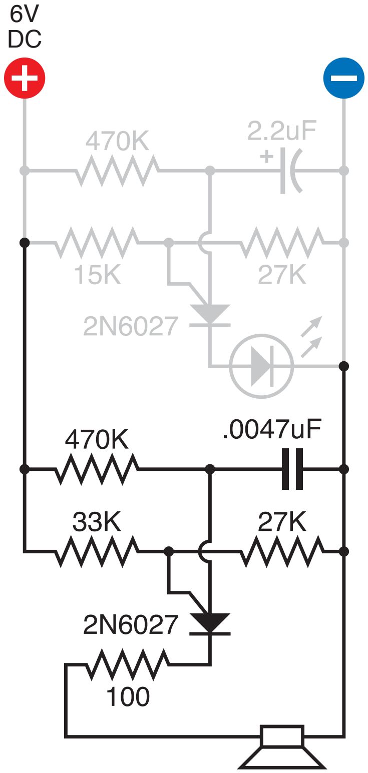

Figure 2-104. *The extra components which have been added at the lower half of the breadboard have the same functions as the components at the top, but some values are slightly different:*

R4: 470K
R5: 33K
R6: 27K
R7: 100Ω
C2: 0.0047 µF
Q2: 2N6027
L1: 8Ω 1-inch loudspeaker

Figure 2-105. *The previous section that you built is shown in gray. Just add the new section in black.*

I want you to keep the slow-flashing circuit separately, untouched, because I have an idea to make use of it a little later. You can leave the LED blinking.

The loudspeaker should be wired in series with a 100Ω resistor to limit the current that flows out of the PUT. The loudspeaker doesn't have any polarity, even though it is fitted with a red wire and a black wire. You can connect it either way around.

Initially, you may be disappointed, because the circuit will not seem to be doing anything. However, if you place your ear very, very close to the loudspeaker, and if you wired the circuit correctly, you should hear a faint buzz, like a mosquito. Obviously, this isn't loud enough to serve any practical purpose. We need to make it louder. In other words, we need to amplify it.

Maybe you remember that the 2N2222, which you played with previously, can function as an amplifier. So let's try using that.

Step 3: Amplification

Disconnect the loudspeaker and its 100Ω series resistor. Then add the 2N2222, which is linked with the output from the PUT via a 1K resistor to protect it from excessive current. See Figure 2-107.

The emitter of the 2N2222 is connected to ground, and the collector is supplied through the loudspeaker and its 100Ω series resistor. This way, small fluctuations in the output from the PUT are sensed by the base of the 2N2222 which converts them into bigger fluctuations between the collector and the emitter, which draw current through the loudspeaker. Check the schematic in Figure 2-108.

Now the sound should be louder than an insect buzz, but still not really loud enough to be useful. What to do?

Well—how about if we add another 2N2222? Bipolar transistors can be placed in series, so that the output from the first one goes to the base of the second one. The 24:1 amplification of the first one is multiplied by another 24:1, giving a total amplification of more than 500:1.

There are limits to this technique. The 2N2222 can only conduct so much current before getting overloaded, and excess amplification can cause distortion. But when I built this circuit, I used a meter to verify that we're still within the design limits of a 2N2222, and for this project, I don't care whether the sound is slightly distorted.

BACKGROUND

Mounting a loudspeaker

The diaphragm or cone of a loudspeaker is designed to radiate sound, but as it oscillates to and fro, it emits sound from its back side as well as its front side. Because the sounds are opposite in phase, they tend to cancel each other out.

The perceived output from a loudspeaker can increase dramatically if you add a horn around it in the form of a tube to separate the output from the front and back of the speaker. For a miniature 1-inch loudspeaker, you can bend and tape a file card around it. See Figure 2-106.

Better still, mount it in a box so that the box absorbs the sound from the rear of the loudspeaker. For purposes of these simple experiments, I won't bother to go into the details of vented enclosures and bass-reflex designs.

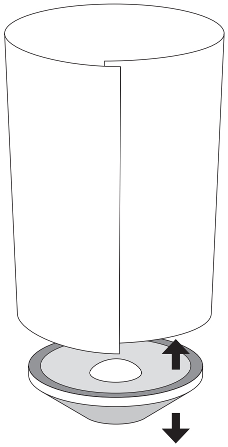

Figure 2-106. *A loudspeaker emits sound from its bottom surface as well as its top surface. To increase the perceived audio volume, use a cardboard tube to separate the two sound sources, or mount the speaker in a small box.*

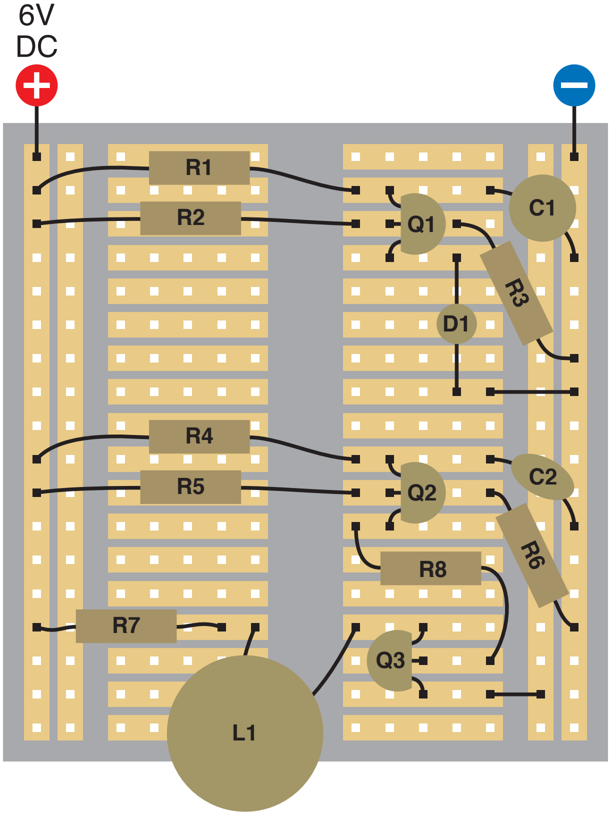

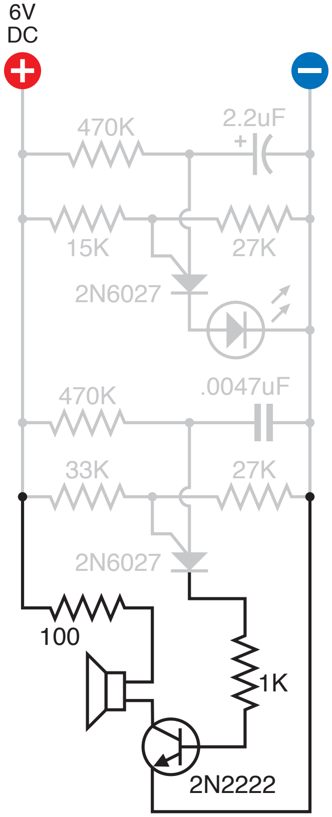

Figure 2-107. *By adding a 2N2222 general-purpose transistor, we amplify the signal from Q2:*

R8: 1K
Q3: 2N2222

Other components are the same as in the previous step in constructing this circuit.

Figure 2-108

Add the second 2N2222 as shown in Figure 2-109. In Figure 2-110, once again the previously wired section is in gray.

If the accumulation of electrical components is beginning to seem confusing, remember that each cluster of parts has a separate defined function. We can draw a block diagram to illustrate this, as in Figure 2-112.

Using the second 2N2222, you should find that the output is more clearly audible, at least within the limits of your tiny 1-inch loudspeaker. Cup your hands around it to direct the sound, and you'll find that the volume seems to increase. You can also try using a 3-inch loudspeaker, which will create a generally better audio output while still remaining within the limits of the little 2N2222 transistor. See Figure 2-106, shown previously, and Figure 2-111.

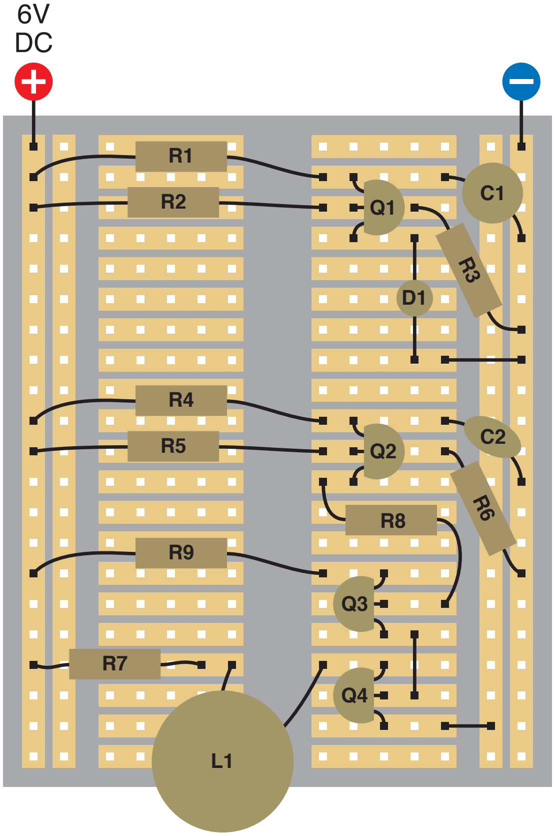
6V
DC
R1
R2
Q1
C1
R3
D1
R4
R5
Q2
C2
R6
R8
R9
Q3
R7
Q4
L1

Figure 2-109. *Q4 is another 2N2222 transistor that further amplifies the signal. It receives power through R9: 2.2K.*

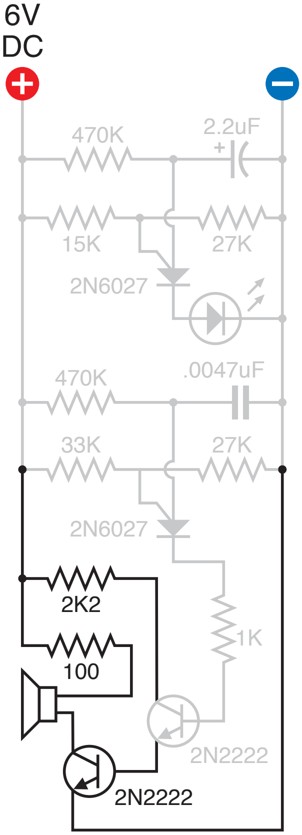
6V
DC
470K
2.2uF
15K
27K
2N6027
470K
.0047uF
33K
27K
2N6027
2K2
1K
100
2N2222
2N2222

Figure 2-110. *This schematic is comparable with the component layout in Figure 2-109.*

Figure 2-111. *The 2N2222 transistor is quite capable of driving a 3-inch loudspeaker, which will create much better sound than a 1-inch speaker.*

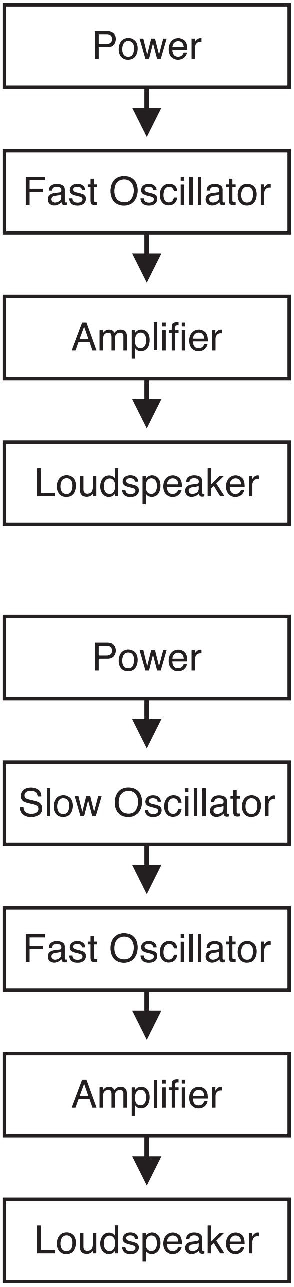

Figure 2-112. *Top: The basic functions of the noisemaking oscillator circuit shown as a block diagram. Bottom: The same functions with a slow oscillator added to control the fast oscillator.*

Step 4: Pulsed Output

If you wanted to use this audio signal as some kind of an alarm, a steady droning noise is not very satisfactory. A pulsing output would be a much better attention-getter.

Well, the first section of the circuit that you assembled created a pulsing signal about twice per second. You used it to flash an LED. Maybe we can get rid of the LED and feed the output from the first section to the second section. The lower block diagram in Figure 2-112 explains this concept.

Can it really be that simple? Well, yes and no. The trick is to make the output from the first section compatible with the input to the second section. If you simply connect a wire from the cathode of the first PUT to the anode of the second PUT, that's not going to work, because the second PUT is already oscillating nicely between low and high voltage, about 1,000 times each second. Add more voltage, and you will disrupt the balance that enables oscillation.

However, remember that the voltage on the gate of a PUT affects its threshold for conducting electricity. Maybe if we connect the output from Q1 to the gate of Q2, we'll be able to adjust that threshold automatically. The voltage still has to be in a range that the PUT finds acceptable, though. We can try various resistors to see which one works well.

This sounds like trial and error—and that's exactly what it is. Doing the math to predict the behavior of a circuit like this is far too complicated—for me, anyway. I just looked at the manufacturer's data sheet, saw the range of resistor values that the PUT would tolerate, and chose one that seemed as if it should work.

If you remove the LED and substitute R10 as shown in the breadboard diagram in Figure 2-113, you'll find that the fluctuating output from Q1 makes Q2 emit a two-tone signal. This is more interesting, but still not what I want. I'm thinking that if I make the pulses out of Q1 less abrupt, the result could be better, and the way to smooth a pulsing output is to hook up another capacitor that will charge at the beginning of each pulse and then release its charge at the end of each pulse. This is the function of C3 in Figure 2-114, and it completes the circuit so that it makes a whooping sound almost like a "real" alarm.

If you don't get any audio output, check your wiring very carefully. It's easy to make a wrong connection on the breadboard, especially between the three legs of each transistor. Use your meter, set to DC volts, to check that each section of the circuit has a positive voltage relative to the negative side of the power supply.

Figure 2-115 shows how your circuit should actually look on the breadboard.

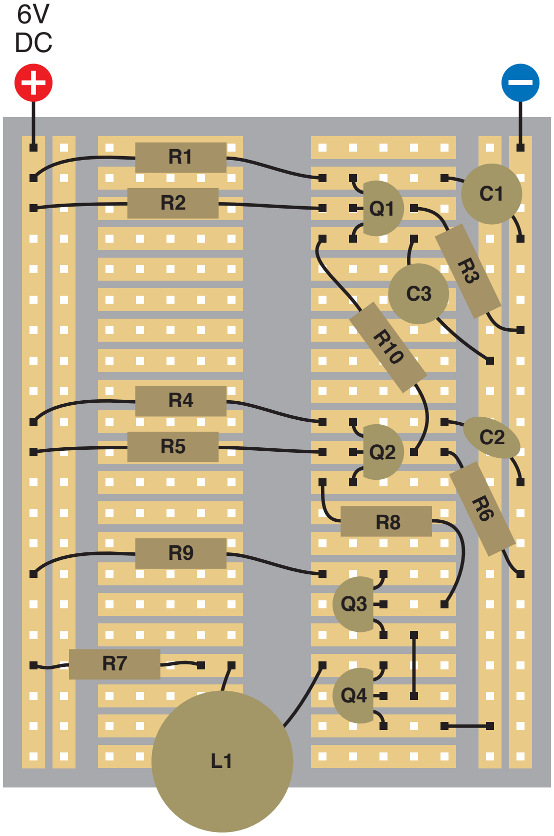
6V
DC
R1
R2
Q1
C1
R3
C3
R10
R4
R5
Q2
C2
R6
R8
R9
Q3
R7
Q4
L1

Figure 2-113. *R10 connects the slow-running oscillator at the top of the breadboard to the gate of Q2, the PUT in the middle of the breadboard. This modulates the audio oscillator, with addition of a smoothing capacitor.*

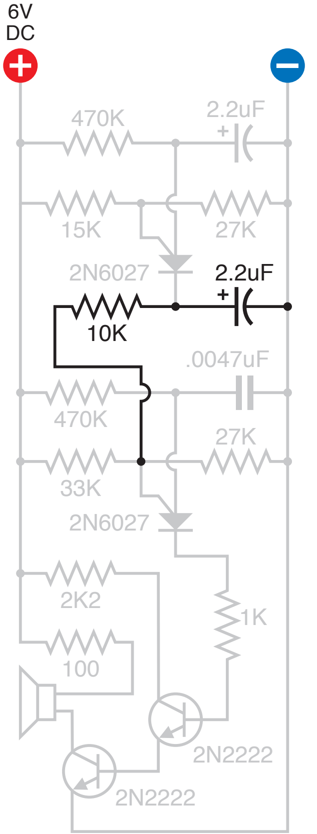
6V
DC
470K
2.2uF
+
15K
27K
2N6027
2.2uF
+
10K
.0047uF
470K
27K
33K
2N6027
2K2
1K
100
2N2222
2N2222

Figure 2-114. *This schematic shows the same circuit as in Figure 2-113:*

R10: 10K
C3: 2.2 μF

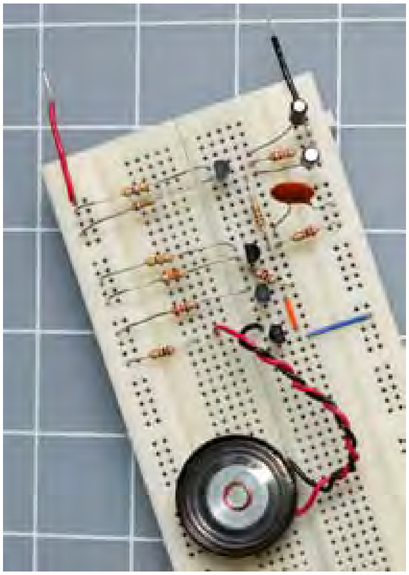

Figure 2-115. *This photograph shows the complete alarm-audio circuit on a breadboard.*

Tweaking it

There's still a lot of room for creativity here:

- Adjust the frequency of the sound: Use a smaller or larger capacitor instead of C2 (half or twice the current value). Use a smaller or larger value for R5.
- Adjust the pulsing feature: Use a smaller or larger capacitor instead of C1 (half or twice the current value). Use a smaller or larger value for R2.
- General performance adjustments: try a larger value for R1. Try smaller or larger values for C3.
- Try running the circuit at 7.5 volts, 10 volts, and 12 volts.

The circuits in this book are suggested as only a starting point. You should always try to tweak them to make them your own. As long as you follow the general rule of protecting transistors and LEDs with resistors, and respecting their requirements for positive and negative voltage, you're unlikely to burn them out. Of course, accidents will happen—I myself tend to be careless, and fried a couple of LEDs while working on this circuit, just because I connected them the wrong way around.

Step 5: Enhancements

A noisemaking circuit is just the output of an alarm. You would need several enhancements to make it useful:

1. Some kind of an intrusion sensor. Maybe magnetic switches for windows and doors?
2. A way to start the sound if any one of the sensors is triggered. The way this is usually done is to run a very small but constant current through all of the switches in series. If any one switch opens, or if the wire itself is broken, this interrupts the current, which starts the alarm. You could make this happen with a double-throw relay, keeping the relay energized all the time until the circuit is broken, at which point, the relay relaxes, opening one pair of contacts and closing the other pair, which can send power to the noisemaker.

   The trouble is that a relay draws significant power while it's energized, and it also tends to get hot. I want my alarm system to draw very little current while it's in "ready" mode, so that it can be powered by a battery. Alarm systems should never depend entirely on AC house current.

   If we don't use a relay, can we use a transistor to switch on the rest of the circuit when the power is interrupted? Absolutely; in fact, one transistor will do it.
3. But how do we arm the alarm in the first place? Really, we need a three-step procedure. First, check a little light that comes on when all the doors and windows are closed. Second, press a button that starts a 30-second countdown, giving you time to leave, if that's what you want to do. And third, after 30 seconds, the alarm arms itself.

4. If the alarm is triggered, what then? If someone forces open a window, should the alarm stop sounding as soon as the window is closed again? No, the alarm should lock itself on, until you turn it off.
5. How do you turn it off? Some kind of secret-code keypad would be good.
6. But to avoid driving everyone crazy if the alarm is triggered when you're not there, it should eventually stop itself, perhaps after about 10 minutes. At that point it should remain quiet but should light an LED to tell you what happened. You can then press a reset button to switch off the LED.

Implementing the Wish List

I've compiled a wish list that seems likely to make the project at least five times as complicated as it is already. Well, that's what tends to happen when you go beyond little demo circuits and try to design something that will be useful in everyday life. Suddenly you find yourself having to accommodate all kinds of circumstances and situations.

Actually, I can and will show you how to take care of all the enhancements on the wish list, but I'm thinking that they will require us to get a little more serious about electronics projects in general first. If you're going to build something ambitious, you'll want to make it more permanent, and probably more compact, than a breadboard with components pushed into it.

You will need to know how to connect everything permanently with solder, on a piece of perforated board that you can install in a neat little project box with switches and lights on the outside.

I'm going to deal with soldering and packaging in the next chapter. After that, we'll get back to the alarm project.

Getting Somewhat More Serious

3

I don't know how far you'll want to delve into electronics, but I do know that I've shown you about as much as I can with just a handful of components, some wires, a breadboard, and a few tools. To continue, you'll need:

- Some more tools and supplies
- Basic soldering skills
- Additional knowledge about:
    - Integrated circuits
    - Digital electronics
    - Microcontrollers
    - Motors

The tools are not particularly exotic or expensive, and the soldering skills are easily acquired. Learning to join wires with solder is far easier than mastering high-level crafts such as jewelry making or welding.

As for additional areas of knowledge about electronics, they are no more challenging than those that I have covered already.

By the end of this section, you should be able to transplant components from a breadboard onto perforated board, where you will solder everything together, and then mount the board in a little box with switches and warning lights on the front, for everyday use.

IN THIS CHAPTER

Shopping List: Experiments 12 Through 15
Experiment 12: Joining Two Wires Together
Experiment 13: Broil an LED
Experiment 14: A Pulsing Glow
Experiment 15: Intrusion Alarm Revisited

Shopping List: Experiments 12 Through 15

Tools

Each of the following tools is rated Essential, Recommended, or Optional. The Essential tools will take you through this chapter of the book. If you supplement them with the Recommended tools, they should be sufficient to get you to the end of the book. The Optional tools will make your work easier, but whether they're worth the money is for you to decide. Remember that URLs for manufacturers and sources of supply are all listed in the appendix.

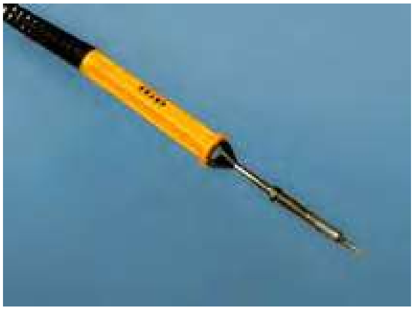

Figure 3-1. *The low wattage of this pencil-style soldering iron enables you to use it safely on sensitive components, and the sharp tip helps to apply heat selectively.*

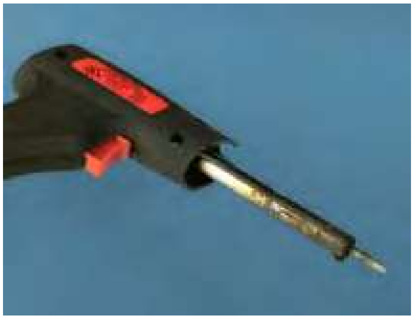

Figure 3-2. *This higher-wattage soldering iron delivers the additional heat necessary for thicker wire or larger components. The discoloration quickly occurs as a result of everyday use and has no effect on the capability of the iron, as long as the tip of it is clean.*

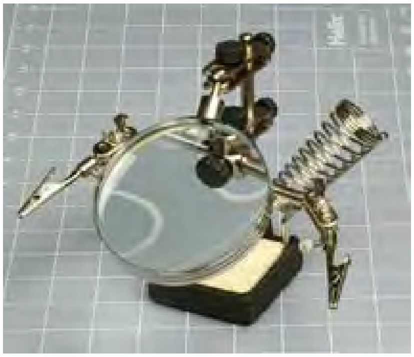

Figure 3-3. *The helping hand is fitted with two alligator clips to hold your work. The metal spiral is a safe place to holster a hot soldering iron, and you use the sponge to wipe its tip.*

I am assuming that you already have some commonly used workshop basics, such as an electric drill.

Essential: Pencil-type 15-watt soldering iron

Examples are RadioShack part 64-2051, McMaster-Carr catalog item 7016A34, or Xytronic model 252. See Figure 3-1. Soldering irons rated at 15 watts are less common than those that deliver 25 watts or more. Still, the 15-watt size is desirable for the small-scale work you'll be doing, and greatly reduces the risk of damaging components by inflicting excess heat.

When comparing prices, remember that a plated tip, which costs a little more, will last longer, will be easier to keep clean, and will conduct heat more reliably than a plain copper tip. If the manufacturer's specification doesn't mention a plated tip, the soldering iron probably doesn't have one.

Essential: General-duty soldering iron, 30 to 40 watts

Although most of the projects in this book entail small, heat-sensitive components and thin wire, at some point you're likely to want to make a solder joint with larger components and/or thicker wire. A 15-watt soldering iron will be unable to deliver enough heat. You should consider having a larger soldering iron in reserve, especially because they are relatively inexpensive.

Personally, I like the Weller Therma-Boost, shown in Figure 3-2, because it has an extra button that delivers more heat on demand. This is useful when you want the iron to get hot quickly, or if you are trying to solder very thick wire, which absorbs a lot of heat.

If you can't find or don't like the Weller, almost any 30-watt or 40-watt soldering iron will do. Check eBay or your local hardware store.

Essential: Helping hand

The so-called "helping hand" (or "third hand") has two alligator clips that hold components or pieces of wire precisely in position while you join them with solder. Some versions of the "helping hand" also feature a magnifying lens, a wire spiral in which you can rest your soldering iron, and a little sponge that you use to clean the tip of your iron when it becomes dirty. These additional features are desirable. Helping hands are available from all electronics hobby sources. Examples are the catalog item HH55 from Elenco or model 64-2991 from RadioShack. See Figure 3-3.

Essential: Magnifying lens

No matter how good your eyes are, a small, handheld, powerful magnifying lens is essential when you are checking solder joints on perforated board. The three-lens set in Figure 3-4 is designed to be held close to your eye, and is more powerful than the large lens on a "helping hand." The folding lens in Figure 3-5 stands on your workbench for hands-free operation. Both are available from RadioShack and similar items are stocked by art supply stores and hobby shops. Plastic lenses are quite acceptable if you treat them carefully.

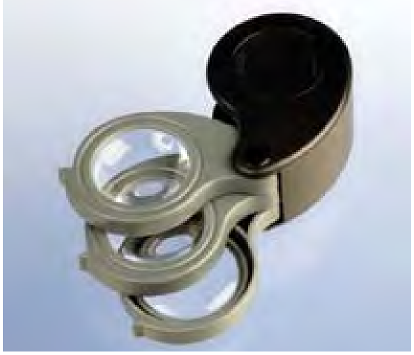

Figure 3-4. *As long as you treat it carefully, a cheap set of plastic magnifying lenses is perfectly acceptable. Handheld magnification is essential for inspecting the solder joints that you make on perforated board.*

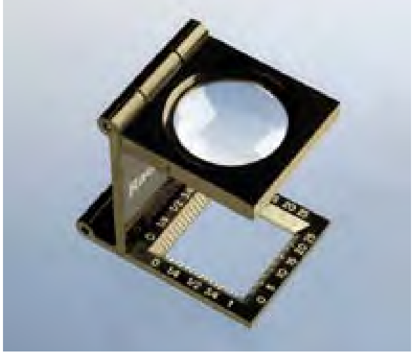

Figure 3-5. *This kind of folding magnifier can stand on your desktop and is useful for checking part numbers on tiny components.*

Essential: Clip-on meter test leads

The probes that came with your multimeter require you to hold them in contact while you make a reading. This requires both hands, preventing you from doing anything else at the same time.

When you use a pair of "minigrabber" probes with little spring-loaded clips at the end, you can attach the Common (negative) lead from your meter to the negative side of your circuit and leave it there, while you touch or attach the positive probe elsewhere.

The Pomona model 6244-48-0 (shown in Figure 3-6) from Meter Superstore and some other suppliers is what you need. If you have trouble finding it or you object to the cost, you may consider making your own by buying a couple of "banana plugs" (such as RadioShack part 274-721) that will fit the sockets on your meter, and then use 16-gauge or thicker stranded wire to connect the plugs with IC test clips, such as Kobiconn 13IC331 or RadioShack "mini hook clips," part number 270-372C. See Figures 3-7 and 3-8.

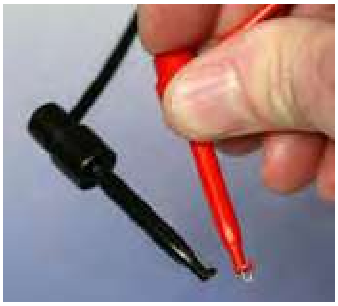

Figure 3-6. *These "minigrabber" add-ons for meter leads make it much easier to measure voltage or current. Push the spring-loaded button, and a little copper hook slides out. Attach it to a wire, release the button, and you have your hands free for other tasks. It's a mystery that meters are not supplied with these grabbers as standard equipment.*

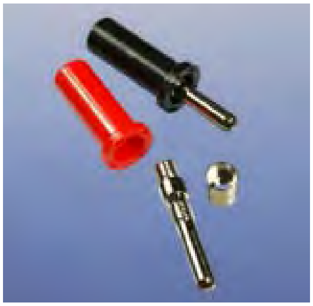

Figure 3-7. *To make your own minigrabber meter leads, first attach a banana plug to a wire by sliding the wire through the cap, into the plug, and out through a hole in the side.*

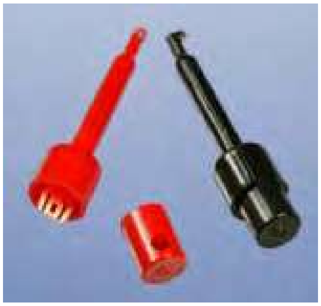

Figure 3-8. *Then screw a collar over the protruding piece of wire, and screw on the cap. The other end of the wire is soldered to a probe.*

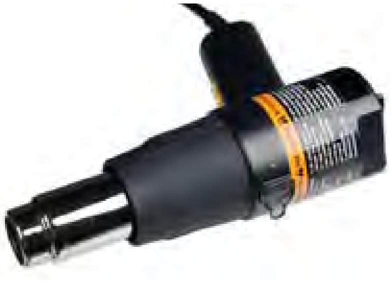

Figure 3-9. *Like an overpowered hair dryer, the heat gun is used with heat-shrink tubing to create a snug, insulated sheath around bare wire.*

Essential: Heat gun

After you join two wires with solder, you often need to insulate them. Electrical tape, sometimes called insulating tape, is messy and tends to come unstuck. You'll be using heat-shrink tube, which forms a safe, permanent sheath around a bare-metal joint. To make the tube shrink, use a heat gun, which is like a very powerful hair dryer. They're available from any hardware supply source, and I suggest you buy the cheapest one you can find. See Figure 3-9.

Essential: Solder pump

This little gadget sucks up hot, melted solder when you are trying to remove a solder joint that you made in the wrong place. Available from All Electronics (catalog item SSR-1) or RadioShack 64-2086. See Figure 3-10.

Essential: Desoldering wick

Also known as desoldering braid. See Figure 3-11. You use this to soak up solder, in conjunction with the Solder Pump. Available from All Electronics (catalog item SWK) or RadioShack (part 64-2090).

Essential: Miniature screwdriver set

Dinky little electronic parts often have dinky little screws in them, and if you try to use the wrong size of screwdriver, you'll tend to mash the heads of the screws. I like the Stanley precision set, part number 66-052, shown in Figure 3-12. But any set will do as long as it has both small Phillips and straight-blade screwdrivers.

Recommended: Soldering stand

Like a holster for a gun, you rest your soldering iron in this stand when the iron is hot but not on use. Examples are catalog item 50B-205 from All Electronics, RadioShack model 64-2078, or check eBay. See Figure 3-13. This item may be built into the helping hand, but you need an extra one for your second soldering iron.

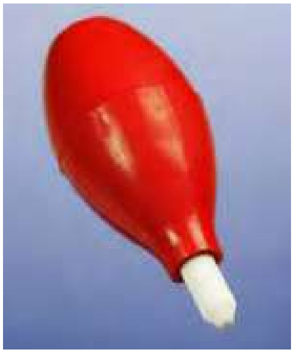

Figure 3-10. *To remove a solder joint, you can heat the solder until it's liquid, then suck it up into this squeezable rubber bulb.*

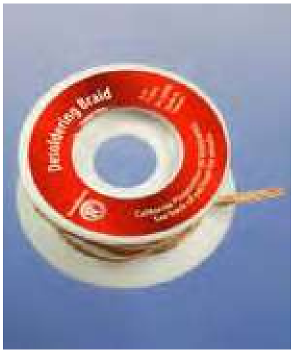

Figure 3-11. *An additional option for removing liquid solder is to soak it up in this copper braid.*

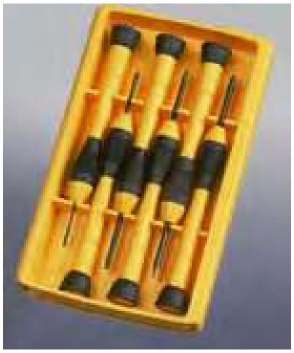

Figure 3-12. *A set of small screwdrivers is essential.*

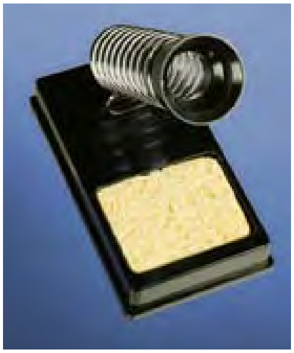

Figure 3-13. *A safe and simple additional stand for a hot soldering iron.*

Recommended: Miniature hand saw

I assume that you will want to mount a finished electronics project in a decent-looking enclosure. Consequently, you are likely to need tools to cut, shape, and trim thin plastic. For example, you may want to cut a square hole so that you can mount a square power switch in it.

Power tools are not suitable for this kind of delicate work. A miniature handsaw (a.k.a. a "hobby saw") is ideal for trimming things to fit. X-Acto makes a range of tiny saw blades. I suggest the #15 blade, plus the handle that it fits in, shown in Figure 3-14. Available online from Tower Hobbies, Hobbylinc, ArtCity, and many other arts/crafts sources. Also look for the larger X-Acto saw blade, #234 or #239, which you can use for cutting perforated board.

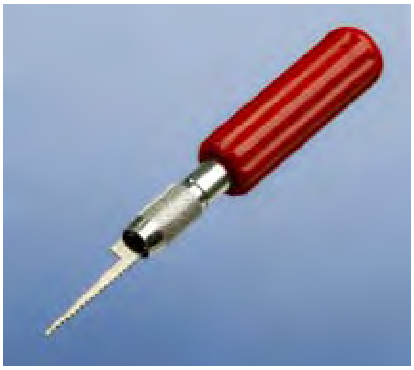

Figure 3-14. *X-Acto makes a range of small saw blades that are ideal for cutting square holes to mount components in plastic panels.*

Recommended: Miniature vise

A miniature vise can do things that the helping hand cannot. I use mine when I'm sawing small pieces of plastic and as a dead weight to anchor a piece of perforated board while I'm working on it. See Figure 3-15.

Look for a cast-iron vise that is listed as being 1 inch or slightly larger, available from Megahobby, eBay, and other arts/crafts sources. Also consider the PanaVise, which has a tilting head to allow you to turn your work to any angle.

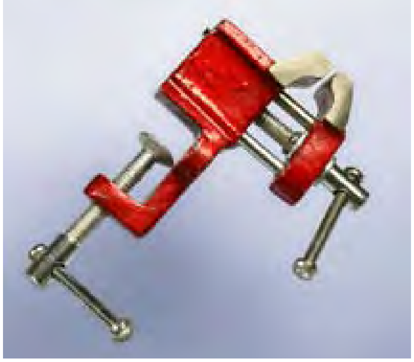

Figure 3-15. *This one-inch vise is available from the McMaster-Carr catalog.*

Recommended: Deburring tool

A deburring tool instantly smoothes and bevels any rough edge (when you have sawn or drilled a piece of plastic, for instance) and also can enlarge holes slightly. This may be necessary because some components are manufactured to metric sizes, which don't fit in the holes that you drill with American bits. Your small local hardware store may not stock deburring tools, but they are very inexpensively available from Sears, McMaster-Carr, KVM Tools, or Amazon. See Figure 3-16.

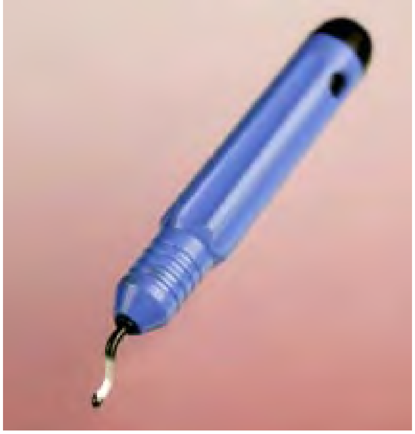

Figure 3-16. *This cunning little blade, safety-tipped with a round bump on the end, removes rough edges from saw cuts with a single stroke, and can enlarge holes that are almost big enough—but not quite.*

Optional: Hand-cranked countersink

You need a countersink to bevel the edges of screw holes to accept flat-headed screws. If you use a countersink bit in an electric drill, it won't give you precise control when you're working with thin, soft plastic.

Handheld countersinks that you grasp and turn like a screwdriver are easy to find, but McMaster-Carr (catalog item 28775A61) is the only source I've found for a hand-cranked tool that is much quicker to use. It comes with a set of bits, as shown in Figure 3-17.

Optional: Pick and hook set

Made by Stanley, part number 82-115, available from Amazon and hardware stores. You can find imported imitations for a few dollars less. See Figure 3-18.

Optional: Calipers

These may seem like a luxury, but are useful for measuring the external diameter of a round object (such as the screw thread on a switch or a potentiometer) or the internal diameter of a hole (into which you may want a switch or potentiometer to fit).

I like Mitutoyo calipers, and the low-end model 505-611 (shown in Figure 3-19) does everything I need. You can find cheaper brands, but economizing on precision measuring tools may not be a wise policy in the long term. The manufacturer's site will show you all their available models, after which you can Google "Mitutoyo" to find retail sources.

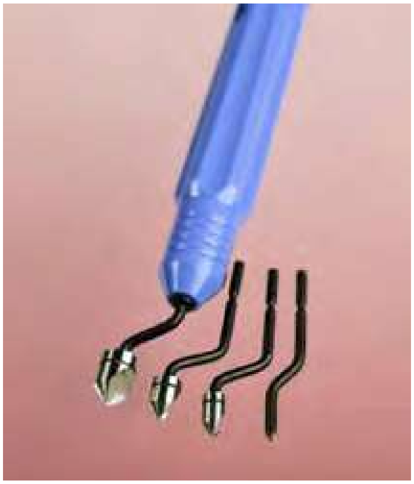

Figure 3-17. *You spin this countersink tool like a hand crank to add just the right amount of bevel to a hole, so that it will accommodate a flat-head screw.*

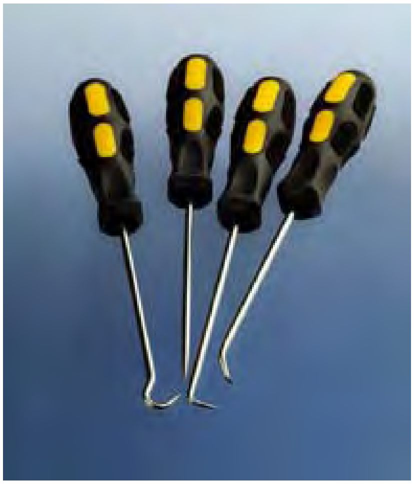

Figure 3-18. *This pick-and-hook set is useful in many unexpected ways.*

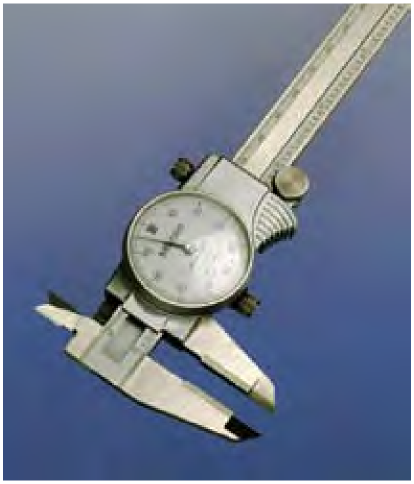

Figure 3-19. *Calipers can be digital (which automatically convert from millimeters to 1/64 inch to 1/1,000 inch), or analog like these (so you never need to worry about a dead battery).*

Supplies

Solder

This is the stuff that you will melt to join components together on a permanent (we hope) basis. You need some very thin solder, size 0.022 inches, for very small components, and thicker solder, 0.05 inches, for heavier items. Avoid buying solder that is intended for plumbers, or for craft purposes such as creating jewelry. A range of solder thicknesses is shown in Figure 3-20. You want to make sure to get lead-free solder.

Electronics solder has a nonacidic rosin core that is appropriate for electronic components. Rolls of solder are available from all hobby-electronics sources including All Electronics, RadioShack, and Jameco, or search for "electronic solder" on Amazon.

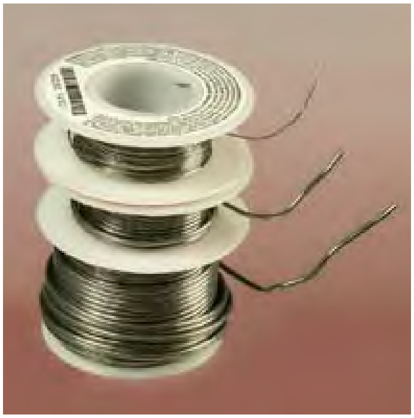

Figure 3-20. *Spools of solder in various thicknesses.*

Wire

You'll need some stranded wire to make flexible external connections with the circuit that you'll be building. Look for 22-gauge stranded hookup wire, in red, black, and green, 10 feet (minimum) of each.

If you want to install the intrusion alarm after completing that project in Experiment 15, you'll need white-insulated two-conductor wire of the type sold for doorbells or furnace controls. This is available by the foot from Lowe's, Home Depot, Ace Hardware, and similar stores. You'll decide how much to buy after you measure the distances between the magnetic sensor switches that you decide to install.

Heat-shrink tube

For use in conjunction with your heat gun, described previously. You'll need a range of sizes in any colors of your choice. See Figure 3-21. Check RadioShack part 278-1627, other electronics suppliers, or your local hardware store. Prices will vary widely. You can buy the cheapest.

Copper alligator clips

These absorb heat when you are soldering delicate components. The Mueller BU-30C is a full-size solid copper alligator clip for maximum heat absorption. RadioShack sells smaller clips (part number 270-373, shown in Figure 3-22) that are suitable for tiny components.

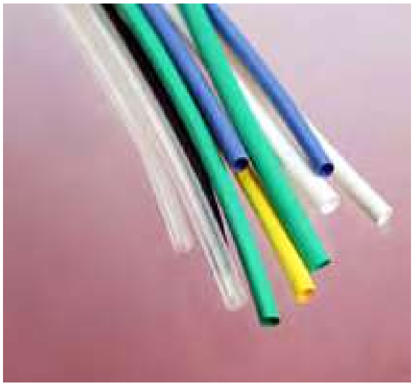

Figure 3-21. *Slide heat-shrink tubing over a bare joint and apply heat from a heat gun to make a tight insulating seal around the joint.*

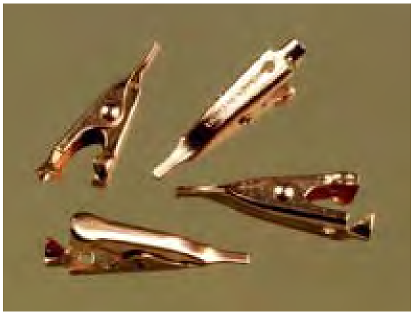

Figure 3-22. *These small clips absorb heat to protect components when you're soldering them.*

Perforated board

When you're ready to move your circuit from a breadboard to a more permanent location, you'll want to solder it to a piece of perforated board, often known as "prototyping board" but also called "perfboard."

You need the type that has copper strips etched onto the back, in exactly the same "breadboard layout" as the conductors hidden inside a breadboard, so that you can retain the same layout of your components when you solder them into place. Examples are RadioShack part 276-150 (shown in Figure 3-23) for small projects and part 276-170 (in Figure 3-24) for larger projects, such as Experiment 15.

For very small projects in which you will connect components using their wires alone, you need perfboard that isn't etched with copper strips connecting the holes. I like the Twin Industries 7100 range (available from Mouser.com) or Vectorboard from Newark Electronics, shown in Figure 3-25. You use a saw to cut out as small a piece as you need. Cheaper options are RadioShack part 276-147 (shown in Figure 3-26), or PC-1 from All Electronics. These have little copper circles around each hole that are not necessary for our purposes, but not a problem, either.

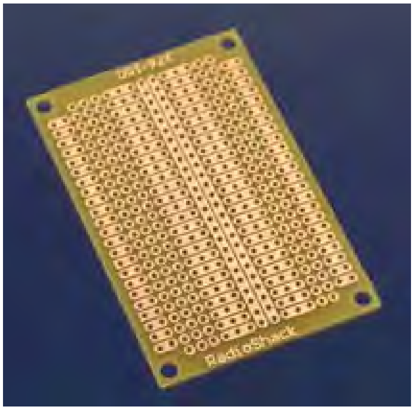

Figure 3-23. *This perforated board has a pattern of copper traces similar to the pattern inside a breadboard, so that you can lay out the components with minimal risk of wiring errors, when you're ready to create a permanently soldered version of your project.*

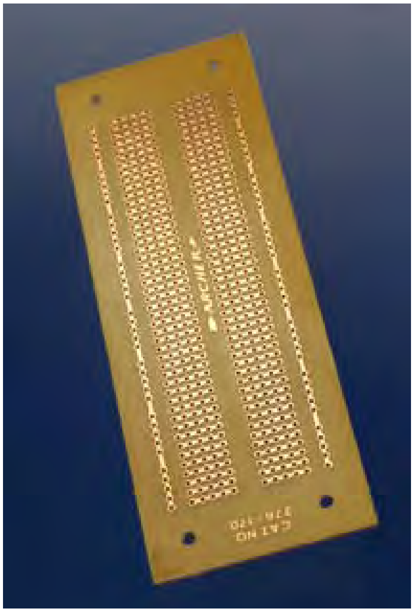

Figure 3-24. *A larger example of perforated board with breadboard geometry.*

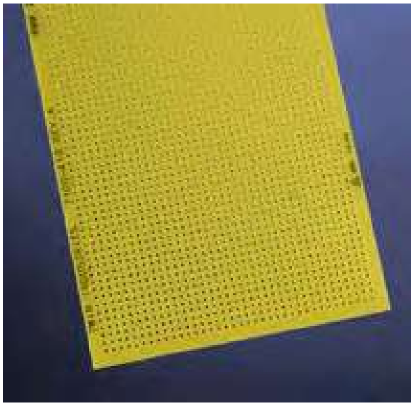

Figure 3-25. *Plain perforated board (with no copper traces) can be used for mounting components when you want to do point-to-point wiring.*

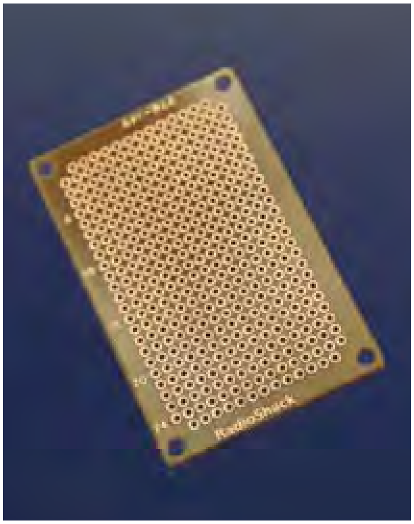

Figure 3-26. *A small piece of perforated board with individual copper solder pads to assist you in mounting components.*

Plywood

When you use a soldering iron, hot drops of solder tend to fall onto your table or workbench. The solder solidifies almost instantly, can be difficult to remove, and will leave a scar. Consider using a 2-foot square of half-inch plywood to provide disposable protection. You can buy it precut at Home Depot or Lowe's.

Machine screws

To mount components behind a panel, you need small machine screws (or "bolts"). They look nice if they have flat heads that fit flush against the panel. I suggest stainless-steel machine screws, #4 size, in 1/2-, 5/8-, 3/4-, and 1-inch lengths, 100 of each, plus 400 washers and 400 #4 locknuts of the type that have nylon inserts, so that they won't work loose. Check McMaster-Carr for a large and reasonably priced selection.

Project boxes

A project box is just a small box (usually plastic) with a removable lid. You mount your switches, potentiometers, and LEDs in holes that you drill through the box, and you attach your circuit on a perforated board that goes inside the box. Search All Electronics for "project box" or RadioShack for "project enclosure."

You need a box measuring approximately 6 inches long, 3 inches wide, and 2 inches high, such as RadioShack part 270-1805. Anything similar will do. I suggest you buy a couple other sizes as well, as they will be useful in the future.

Components

Power plugs, sockets, and binding posts

After you finish a project and put it in a box, you'll need a convenient way to supply it with power. Buy yourself a pair of insulated binding posts, such as RadioShack part 274-661, shown in Figure 3-27. Also obtain a panel-mounted power jack, size N, such as RadioShack part 274-1583, and DC power plug, size N, such as RadioShack 274-1573. The plug-and-socket pair is pictured in Figure 3-28.

Finally, you will need interconnects that are sized to fit a perforated board that is drilled at intervals of 1/10 inch. Sometimes known as "single inline sockets and headers," but also known as "boardmount sockets and pin-strip headers," they come in strips of 36 or more, and you can snip off as many as you need. Examples are Mill-Max part numbers 800-10-064-10-001000 and 801-93-050-10-001000, or 3M part numbers 929974-01-36-RK and 929834-01-36-RK. You can buy them from the usual electronics suppliers. Figure 3-29 shows headers before and after being snapped into small sections. Make sure that the interconnects have a terminal spacing of 0.1 inch.

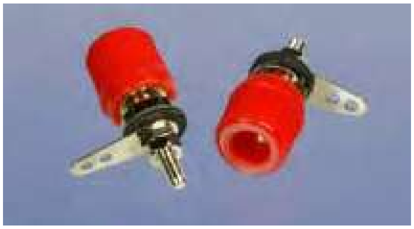

Figure 3-27. *These terminals, also known as binding posts, enable a solderless connection with wires that have stripped ends. Also available in black.*

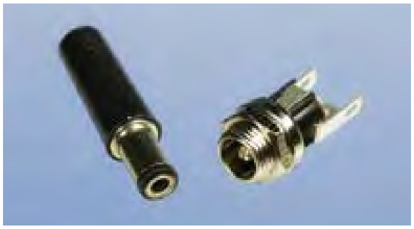

Figure 3-28. *The socket on the right can be mounted in a project box to receive power from the plug on the left.*

Battery

After you complete Experiment 15 at the end of this section of the book, if you want to use the project on a practical basis, you'll need a 12-volt battery. Search online for "12v battery" and you'll find many sealed, rechargeable lead-acid batteries that are designed for alarm systems, some measuring as small as 1×2×3-inch and costing under $10. You need a charger with it, which will probably cost you about $10.

Switches and relays

You will need the same DPDT relay and the same SPDT toggle switch that were mentioned in Chapter 2 shopping list.

For Experiment 15, you'll need magnetic switches that you can apply to doors or windows, such as the Directed model 8601, available from dozens of sources online.

Also you will need a DPDT pushbutton switch, ON-(ON) type, with solder terminals. Examples are model MPG206R04 by Tyco or model MB2061SS1W01-RO by NKK (with optional cap). Or search eBay for "DPDT pushbutton."

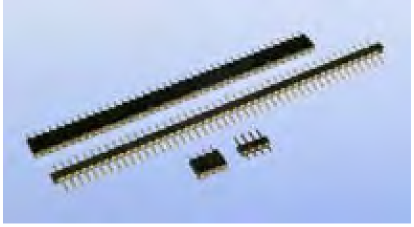

Figure 3-29. *Single inline sockets (top) and headers (middle) allow you to make very compact plug-and-socket connections to a PC board. They can be sawn, cut, or snapped into smaller sections (bottom). The terminals are 0.1 inch apart.*

Diodes

Buy at least half-a-dozen red 5 mm LEDs rated for approximately 2 volts, such as the Optek part number OVLFR3C7, Lumex part number SSL-LX-5093IT, or Avago part HLMP-D155. Buy half-a-dozen similar green LEDs at the same time.

In addition, you'll need a signal diode, type 1N4001 (any brand will do). Figure 3-30 shows an example, highly magnified. They're cheap, and likely to be useful in the future, so buy 10 of them.

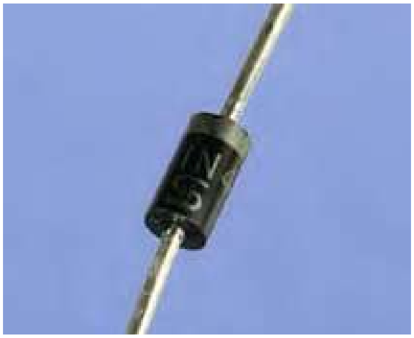

Figure 3-30. *This 1N4001 diode is about 1/4 inch long and can handle up to 50 volts.*

Loudspeaker

To complete the project in Experiment 15, you'll need a loudspeaker small enough to fit inside your project box but louder than the 1-inch speaker that you used previously. It should be 2 inches or 2.5 inches (50 to 60 mm) in diameter. If you can find a 100Ω speaker, it will give you more output, but an 8Ω speaker will be acceptable.

Soldering Irons Get Hot!

Please take these basic precautions:

Use a proper stand (such as the one incorporated in your helping hands) to hold your soldering iron. Don't leave it lying on a workbench.

If you have infants or pets, remember that they may play with, grab, or snag the wire to your soldering iron. They could injure themselves (or you).

Be careful never to rest the hot tip of the iron on the power cord that supplies electricity to the iron. It can melt the plastic in seconds and cause a dramatic short circuit.

If you drop a soldering iron, don't be a hero and try to catch it. Most likely you will grab the hot part, which hurts. (I speak from experience.) When you burn your hand, you will instinctively let go of the iron, so you may as well let it drop freely without the intermediate step of burning yourself while it's on its way to the floor. Naturally, you should pick it up quickly after it does hit the floor, but by then you will have gained the necessary time in which to make a sensible decision to grab it by the cool end.

Always bear in mind that others in your home are more at risk of hurting themselves on a soldering iron than you are, because they won't know that it's hot. Most soldering irons have no warning lights to tell you that they're plugged in. As a general rule, always assume that a soldering iron is hot, even if it's unplugged. It may retain sufficient heat to burn you for longer than you expect.

Experiment 12: Joining Two Wires Together

Your adventure into soldering begins with the prosaic task of joining one wire to another, but will lead quickly to creating a full electronic circuit on perforated board. So let's get started!

You will need:

- 30-watt or 40-watt soldering iron
- 15-watt pencil-type soldering iron
- Thin solder (0.022 inches or similar)
- Medium solder (0.05 inches or similar)
- Wire strippers and cutters
- "Helping hand" gadget to hold your work
- Shrink-wrap tubing, assorted
- Heat gun
- Something to protect your work area from drops of solder

Your First Solder Joint

We'll start with your general-duty soldering iron—the one rated for 30 or 40 watts. Plug it in, leave it safely in its holder, and find something else to do for five minutes. If you try to use a soldering iron without giving it time to get fully hot, you will *not* make good joints.

Strip the insulation from the ends of two pieces of 22-gauge solid wire and clamp them in your helping hand so that they cross each other and touch each other, as shown in Figure 3-31.

To make sure that the iron is ready, try to melt the end of a thin piece of solder on the tip of the iron. The solder should melt instantly. If it melts slowly, the iron isn't hot enough yet.

Now follow these steps (shown in Figures 3-32 through 3-36):

1. Make sure the tip of the soldering iron is clean (wipe it on the moistened sponge in the base of your helping hand if necessary), then touch it against the intersection of the wires steadily for three seconds to heat them. If you have hard tap water, use distilled water to wet the sponge to avoid a buildup of mineral deposits on the tip of your soldering iron.
2. While maintaining the iron in this position, feed a little solder onto the intersection of the wires, also touching the tip of the soldering iron. Thus, the two wires, the solder, and the tip of the iron should all come together at one point. The solder should spread over the wires within another two seconds.
3. Remove the iron and the solder. Blow on the joint to cool it. Within 10 seconds, it should be cool enough to touch.
4. Unclamp the wires and try to tug them apart. Tug hard! If they defeat your best attempts to separate them, the wires are electrically joined and should stay joined. If you didn't make a good joint, you will be able to separate the wires relatively easily, probably because you didn't apply enough heat or enough solder to connect them.

The reason I asked you to begin by using the higher-powered soldering iron is that it delivers more heat, which makes it easier to use.

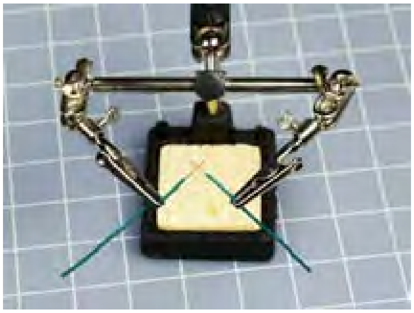
Figure 3-31. *A helping hand work aid is shown here holding two wires with their stripped ends touching. The magnifying glass has been hinged out of the way.*

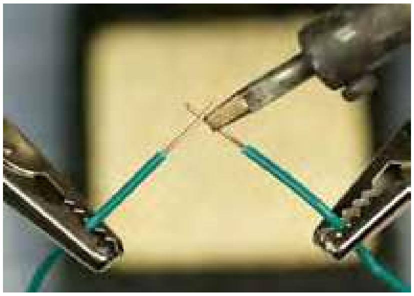
Figure 3-32

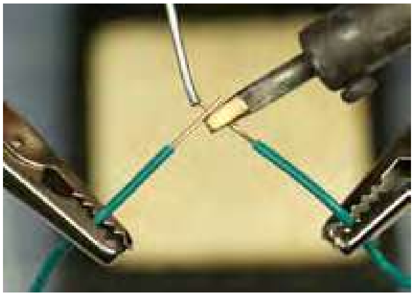
Figure 3-33

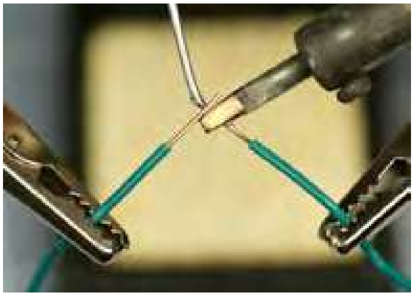
Figure 3-34

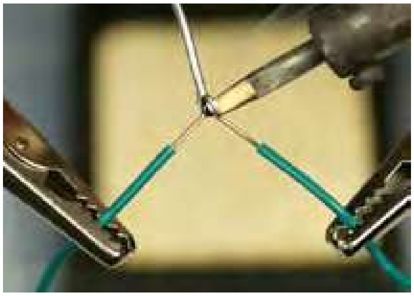
Figure 3-35. *This and the preceding three figures illustrate four steps to making a solder joint: apply heat to the wires, bring in the solder while maintaining the heat, wait for the solder to start to melt, and wait a moment longer for it to form a completely molten bead. The whole process should take between 4 and 6 seconds.*

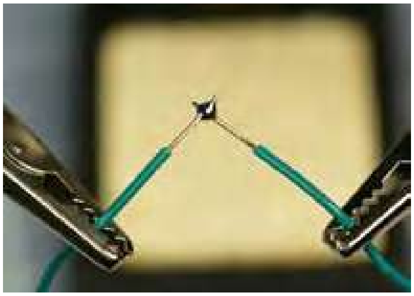
Figure 3-36. *The completed joint should be shiny, uniform, and rounded in shape.*

BACKGROUND

Soldering myths

Myth #1: Soldering is very difficult.

Millions of people have learned how to do it, and statistically, you are unlikely to be less coordinated than all of them. I have a lifelong problem with a tremor in my hands that makes it difficult for me to hold small things steadily. I also get impatient with repetitive detail work. If I can solder components, almost anyone should be able to.

Myth #2: Soldering involves poisonous chemicals.

Modern solder contains no lead. You should avoid inhaling the fumes for prolonged periods, but that also applies to everyday products such as bleach and paint. If soldering was a significant health hazard, we should have seen a high death rate among electronics hobbyists decades ago.

Myth #3: Soldering is hazardous.

A soldering iron is less hazardous than the kind of iron that you might use to iron a shirt, because it delivers less heat. In fact, in my experience, soldering is safer than most activities in a typical home or basement workshop. That doesn't mean you can be careless!

Soldering alternatives

As recently as the 1950s, connections inside electronic appliances such as radio sets were still being hand-soldered by workers on production lines. But the growth of telephone exchanges created a need for a faster way to make large numbers of rapid, reliable point-to-point wiring connections, and "wire wrap" became a viable alternative.

In a wire-wrapped electronics project, components are mounted on a circuit board that has long, gold-plated, sharp-cornered square pins sticking out of the rear. Special silver-plated wire is used, with an inch of insulation stripped from its ends. A manual or power-driven wire-wrap tool twirls the end of a wire around one of the pins, applying sufficient tension to "cold-weld" the soft silver plating of the wire to the pin. The wrapping process exerts sufficient pressure to make a very reliable joint, especially as 7 to 9 turns of wire are applied, each turn touching all four corners of the pin.

During the 1970s and 1980s, this system was adopted by hobbyists who built their own home computers. A wire-wrapped circuit board from a hand-built computer is shown in Figure 3-37. The technique was used by NASA to wire the computer in the Apollo spacecraft that went to the moon, but today, wire-wrapping has few commercial applications.

The widespread industrial use of "through-hole" components, such as the chips on early desktop computers, encouraged development of wave soldering, in which a wave or waterfall of molten solder is applied to the underside of a preheated circuit board where chips have been inserted. A masking technique prevents the solder from sticking where it isn't wanted.

Today, surface-mount components (which are significantly smaller than their through-hole counterparts) are glued to a circuit board with a solder paste, and the entire assembly is then heated, melting the paste to create a permanent connection.

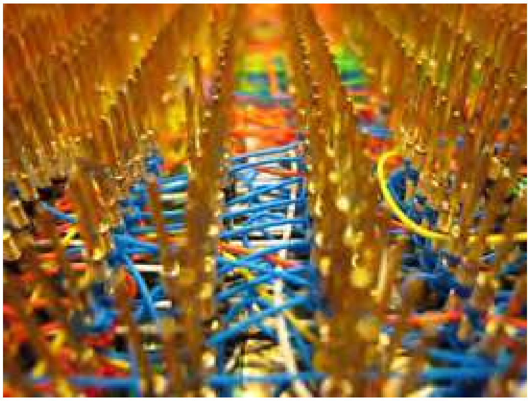

Figure 3-37. *This picture shows some of the wire-wrapping in Steve Chamberlin's custom-built, retro 8-bit CPU and computer. "Back in the day," connecting such a network of wires with solder joints would have been unduly time-consuming and prone to faults. Photo credit: Steve Chamberlin.*

TOOLS

Eight most common soldering errors

1. Not enough heat.

   The joint looks OK, but because you didn't apply quite enough heat, the solder didn't melt sufficiently to realign its internal molecular structure. It remained granular instead of becoming a solid, uniform blob, and you end up with a "dry joint," also known as a "cold joint," which will come apart when you pull the wires away from each other. Reheat the joint thoroughly and apply new solder.

   A leading cause of underheated solder is the temptation to use the soldering iron to carry solder to the joint. This results in the cold wires reducing the temperature of the solder. What you should do is touch the soldering iron to heat the wires first, and then apply the solder. This way, the wires are hot and help to melt the solder, which wants to stick to them.

   Because this is such a universal problem, I'll repeat myself: *Never melt solder on the tip of the iron and then use it to carry the solder to the joint.*

   You don't want to put hot solder on cold wires. You want to put cold solder on hot wires.

2. Too much heat.

   This may not hurt the joint, but can damage everything around it. Vinyl insulation will melt, exposing the wire and raising the risk of short circuits. You can easily damage semiconductors, and may even melt the internal plastic components of switches and connectors.

   Damaged components must be desoldered and replaced, which will take time and tends to be a big hassle (see "Tools: Desoldering" on page 109 for advice).

3. Not enough solder.

   A thin connection between two conductors may not be strong enough. When joining two wires, always check the underside of the joint to see whether the solder penetrated completely.

4. Moving the joint before the solder solidifies.

   You may create a fracture that you won't necessarily see. It may not stop your circuit from working, but at some point in the future, as a result of vibration or thermal stresses, the fracture can separate just enough to break electrical contact. Tracking it down will then be a chore. If you clamp components before you join them, or use perforated board to hold the components steady, you can avoid this problem.

5. Dirt or grease.

   Electrical solder contains rosin that cleans the metal that you're working with, but contaminants can still prevent solder from sticking. If any component looks dirty, clean it with fine sandpaper before joining it.

6. Carbon on the tip of your soldering iron.

   The iron gradually accumulates flecks of black carbon during use, and they can act as a barrier to heat transfer. Wipe the tip of the iron on the little sponge mounted in the base of your soldering iron stand or your helping hand.

7. Inappropriate materials.

   Electronic solder is designed for electronic components. It will not work with aluminum, stainless steel, or various other metals. You may be able to make it stick to chrome-plated items, but only with difficulty.

8. Failure to test the joint.

   Don't just assume that it's OK. Always test it, by applying manual force if you can (see Figures 3-38 and 3-39 for the ideal protocol) or, if you can't get a grip on the joint, slip a screwdriver blade under it and flex it just a little, or use small pliers to try to pull it apart. Don't be concerned about ruining your work. If your joint doesn't survive rough treatment, it wasn't a good joint.

Of the eight errors, dry/cold joints are by far the worst, because they are easy to make and can look OK.

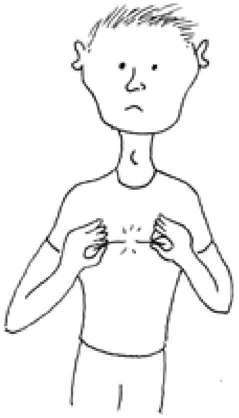

Figure 3-38. *Test result of a bad solder joint.*

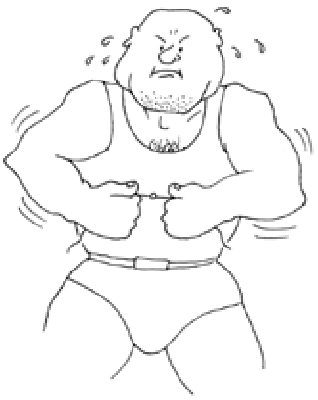

Figure 3-39. *Test result of a good solder joint.*

Your Second Solder Joint

Time now to try your pencil-style soldering iron. Once again, you must leave it plugged in for a good five minutes to make sure it's hot enough. In the meantime, don't forget to unplug your other soldering iron, and put it somewhere safe while it cools.

This time I'd like you to align the wires parallel with each other. Joining them this way is a little more difficult than joining them when they cross each other, but it's a necessary skill. Otherwise, you won't be able to slide heat-shrink tubing over the finished joint to insulate it.

Figures 3-40 through 3-44 show a successful joint of this type. The two wires do not have to make perfect contact with each other; the solder will fill any small gaps. But the wires must be hot enough for the solder to flow, and this can take an extra few seconds when you use the low-wattage pencil-style iron.

Be sure to feed the solder in as shown in the pictures. Remember: don't try to carry the solder to the joint on the tip of the iron. Heat the wires first, and then touch the solder to the wires and the tip of the iron, while keeping it in contact with the wires. Wait until the solder liquifies, and you will see it running eagerly into the joint. If this doesn't happen, be more patient and apply the heat for a little longer.

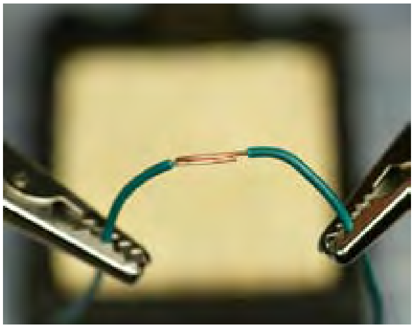
Figure 3-40

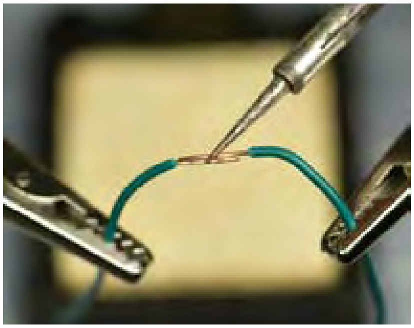
Figure 3-41

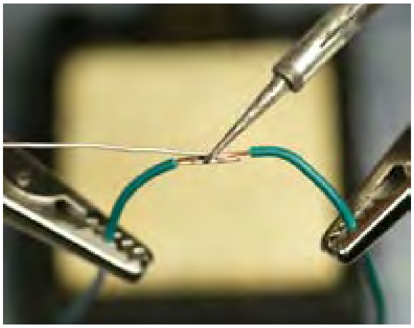
Figure 3-42

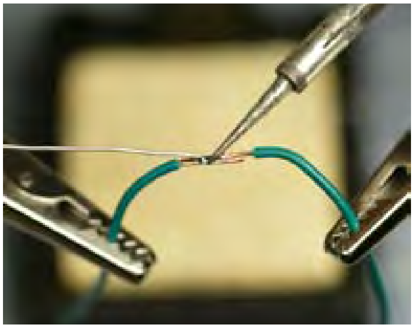
Figure 3-43. *This and the preceding three figures show how joining two wires that are parallel is more difficult, and the low-wattage, pencil-type soldering iron will require longer to heat them sufficiently for a good joint. Thinner solder can be used.*

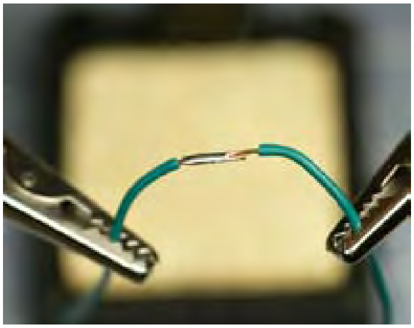
Figure 3-44. *The finished joint has enough solder for strength, but not so much solder that it will prevent heat-shrink tubing from sliding over it.*

THEORY

Soldering theory

The better you understand the process of soldering, the easier it should be for you to make good solder joints.

The tip of the soldering iron is hot, and you want to transfer that heat into the joint that you are trying to make. In this situation, you can think of the heat as being like a fluid. The larger the connection is between the soldering iron and the joint, the greater the quantity of heat, per second, that can flow through it.

For this reason, you should adjust the angle of the soldering iron so that it makes the widest possible contact. If it touches the wires only at a tiny point, you'll limit the amount of heat flow. Figures 3-45 and 3-46 illustrate this concept. Once the solder starts to melt, it broadens the area of contact, which helps to transfer more heat, so the process accelerates naturally. Initiating it is the tricky part.

The other aspect of heat flow that you should consider is that it can suck heat away from the places where you want it, and deliver it to places where you don't want it. If you're trying to solder a very heavy piece of copper wire, the joint may never get hot enough to melt the solder, because the heavy wire conducts heat away from the joint. You may find that even a 40-watt iron isn't powerful enough to overcome this problem, and if you are doing heavy work, you may need a more powerful iron.

As a general rule, if you can't complete a solder joint in 10 seconds, you aren't applying enough heat.

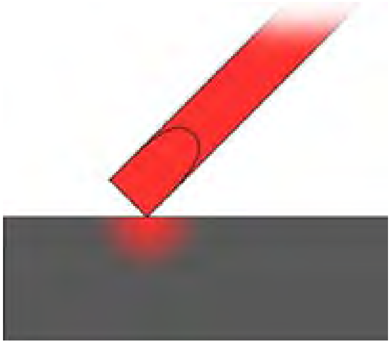

Figure 3-45. *With only a small surface area of contact between the iron and the working surface, an insufficient amount of heat is transferred.*

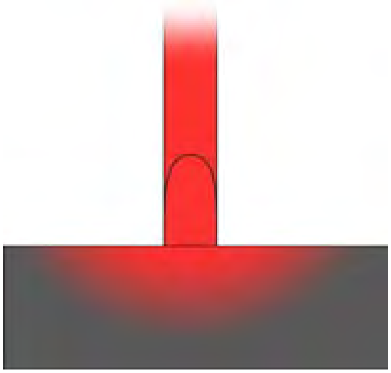

Figure 3-46. *A larger area of contact between the soldering iron and its target will greatly increase the heat transfer.*

TOOLS

Desoldering

Desoldering is much, much harder than soldering. Two simple tools are available:

- *Suction pump*. First, you apply the soldering iron to make the solder liquid. Then you use this simple gadget to try to suck up as much of the liquid as possible. Usually it won't remove enough metal to allow you to pull the joint apart, and you will have to try the next tool. Refer back to Figure 3-10.
- *Desoldering wick or braid*. Desoldering wick, also known as braid, is designed to soak up the solder from a joint, but again, it won't clean the joint entirely, and you will be in the awkward position of trying to use both hands to pull components apart while simultaneously applying heat to stop the solder from solidifying. Refer back to Figure 3-11.

I don't have much advice about desoldering. It's a frustrating experience (at least, I think so) and can damage components irrevocably.

Heat Guns Get Hot, Too!

Notice the chromed steel tube at the business end of your heat gun. Steel costs more than plastic, so the manufacturer must have put it there for a good reason—and the reason is that the air flowing through it becomes so hot that it would melt a plastic tube.

The metal tube stays hot enough to burn you for several minutes after you've used it. And, as in the case of soldering irons, other people (and pets) are vulnerable, because they won't necessarily know that the heat gun is hot. Most of all, make sure that no one in your home ever makes the mistake of using a heat gun as a hair dryer (Figure 3-47).

This tool is just a little more hazardous than it appears.

Adding Insulation

After you've succeeded in making a good inline solder connection between two wires, it's time for the easy part. Choose some heat-shrink tubing that is just big enough to slide over the joint with a little bit of room to spare.

Figure 3-47. *Other members of your family should understand that although a heat gun looks like a hair dryer, appearances may be deceptive.*

Slide the tubing along until the joint is centered under it, hold it in front of your heat gun, and switch on the gun (keeping your fingers away from the blast of superheated air). Turn the wire so that you heat both sides. The tubing should shrink tight around the joint within half a minute. If you overheat the tubing, it may shrink so much that it splits, at which point you must remove it and start over. As soon as the tubing is tight around the wire, your job is done, and there's no point in making it any hotter. Figures 3-48 through 3-50 show the desired result. I used white tubing because it shows up well in photographs. Different colors of heat-shrink tubing all perform the same way.

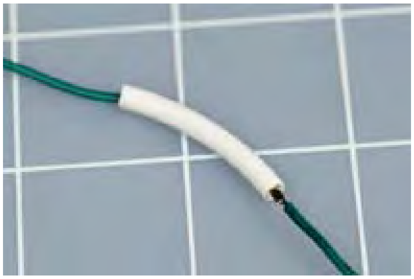

Figure 3-48. *Slip the tubing over your wire joint.*

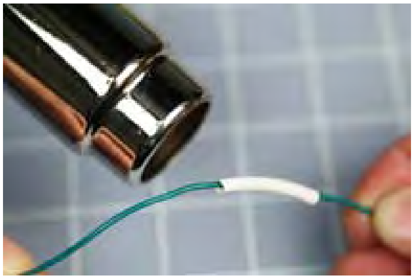

Figure 3-49. *Apply heat to the tubing.*

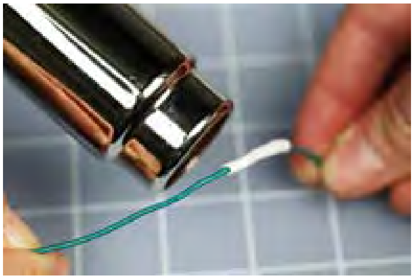

Figure 3-50. *Leave the heat on the tubing until it shrinks to firmly cover the joint.*

I suggest you next practice your soldering skills on a couple of practical projects. In the first one, you can add color-coded, solid-core wires to your AC adapter, and in the second one, you can shorten the power cord for a laptop power supply. You can use your larger soldering iron for both of these tasks, because neither of them involves any heat-sensitive components.

Modifying an AC Adapter

In the previous chapter, I mentioned the irritation of being unable to push the wires from your AC adapter into the holes of your breadboard. So, let's fix this right now:

1. Cut two pieces of solid-conductor 22-gauge wire—one of them red, the other black or blue. Each should be about 2 inches long. Strip a quarter-inch of insulation from both ends of each piece of wire.
2. Trim the wire from your AC adapter. You need to expose some fresh, clean copper to maximize your chance of getting the solder to stick.

   I suggest that you make one conductor longer than the other to minimize the chance of the bare ends touching and creating a short circuit. Use your meter, set to DC volts, if you have any doubt about which conductor is positive.

Solder the wires and add heat-shrink tubing as you did in the practice session. The result should look like Figure 3-51.

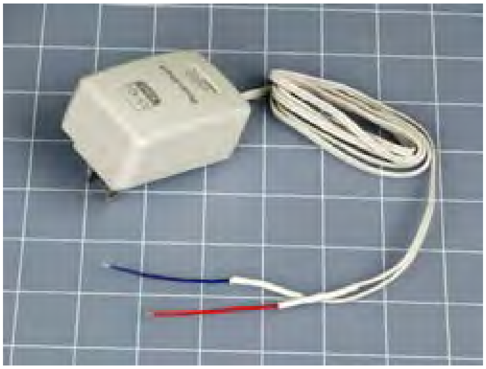

Figure 3-51. *Solid-core color-coded wires, soldered onto the wires from an AC adapter, provide a convenient way to feed power to a breadboard. Note that the wires are of differing lengths to reduce the risk of them touching each other.*

Choose the Right Tubing

If you use heat-shrink tubing on 110v AC cord, as is being done in this experiment, make sure you use tubing that's been rated for 110v use.

Shortening a Power Cord

When I travel, I like to minimize everything. It always annoys me that the power cord for the power supply of my laptop is 4 feet long. The thinner wire that connects the power supply to the computer is also 4 feet long, and I just don't need that much wire.

After searching exhaustively I couldn't find any laptop power cables shorter than 3 feet, so I decided to shorten one myself. If you feel no need to do this, you should try the following procedure on an old extension cord, just as an exercise. You do need to go through these steps to acquire some practice in soldering heavier, stranded wire and using heat-shrink tubing:

1. Use your wire cutters to chop the wire, and then a utility knife to split the two conductors, with one shorter than the other. When splicing a power cord or similar cable containing two or more conductors, it's good to avoid having the joints opposite each other. They fit more snugly if they are offset, and there's less risk of a short circuit if a joint fails.

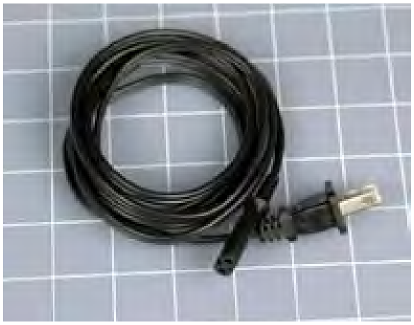
Figure 3-52

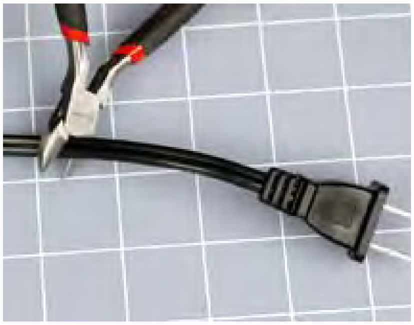
Figure 3-53

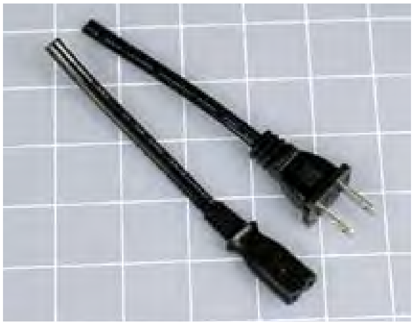
Figure 3-54

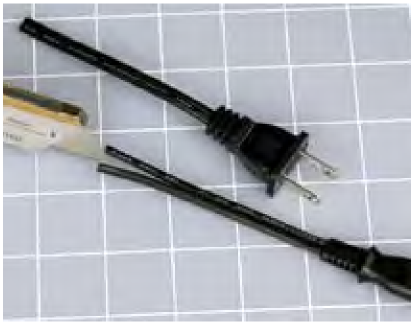
Figure 3-55

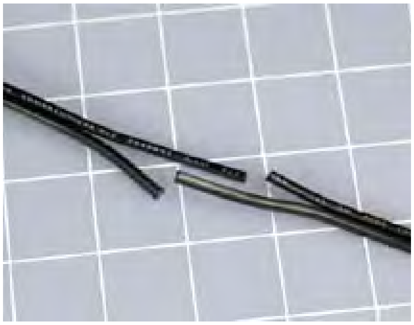
Figure 3-56

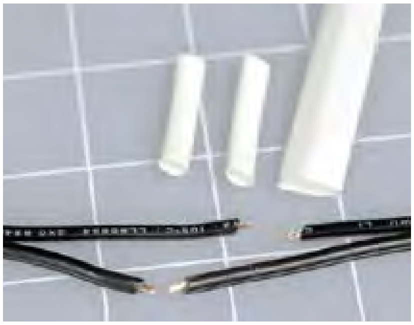
Figure 3-57

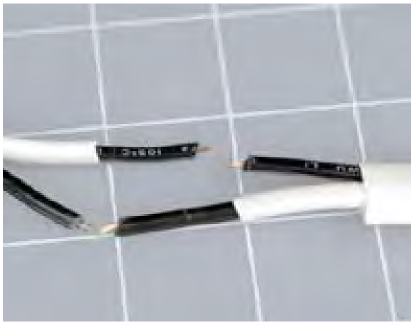
Figure 3-58. *Figures 3-52 through 3-58 illustrate the sequence of steps to prepare for making a shortened power cord for a laptop computer power supply.*

2. Strip off a minimal amount of insulation. One-eighth of an inch (3 mm) is sufficient. The automatic wire strippers that I mentioned in the shopping list in Chapter 1 are especially convenient, but regular wire strippers will do the job.
3. Cut two pieces of heat-shrink tubing, each 1 inch long, big enough to slide over the separate conductors in your cable. Cut a separate 2-inch piece of larger tubing that will slide over the entire joint when it's done. The steps described so far are illustrated in Figures 3-52 through 3-58.
4. Now for the most difficult part: activating your human memory. You have to remember to slide the tubing onto the wire *before* you make your solder joint, because the plugs on the ends of the wires will prevent you from adding any heat-shrink tubing later. If you're as impatient as I am, it's very difficult to remember to do this every time.

5. Use your helping hand to align the first joint. Push the two pieces of wire together so that the strands intermingle, and then squeeze them tight between finger and thumb, so that there are no little bits sticking out. A stray strand of wire can puncture heat-shrink tubing when the tubing is hot and soft and is shrinking around the joint.

6. The wire that you're joining is much heavier than the 22-gauge wire that you worked with previously, so it will suck up more heat, and you must touch the soldering iron to it for a longer time. Make sure that the solder flows all the way into the joint, and check the underside after the joint is cool. Most likely you'll find some bare copper strands there. The joint should become a nice solid, rounded, shiny blob. Keep the heat-shrink tubing as far away from the joint as possible while you're using the soldering iron, so that heat from the iron doesn't shrink the tubing prematurely, preventing you from sliding it over the joint later.

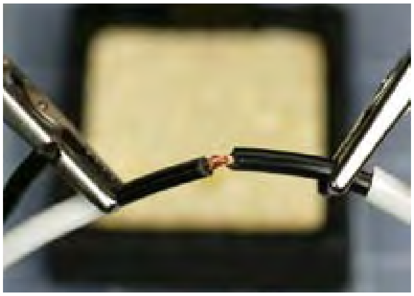
Figure 3-59

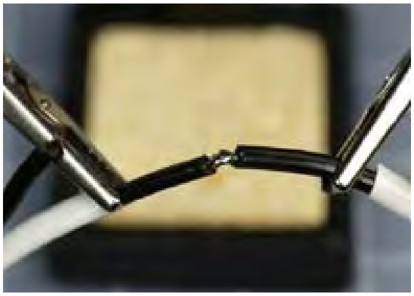
Figure 3-60

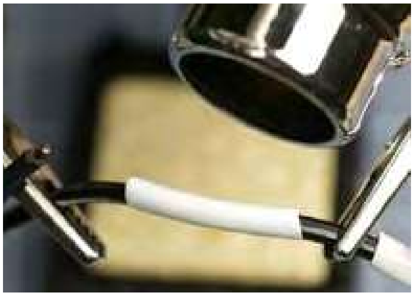
Figure 3-61

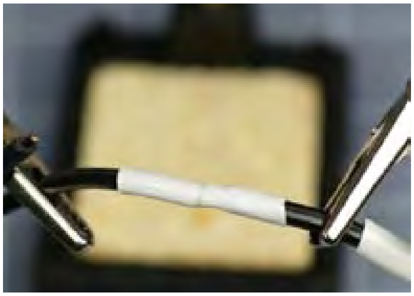
Figure 3-62

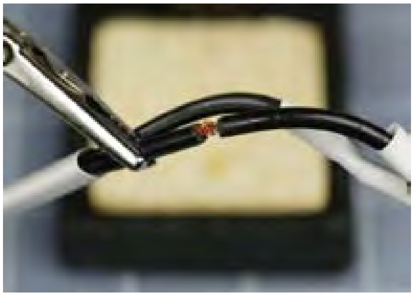
Figure 3-63

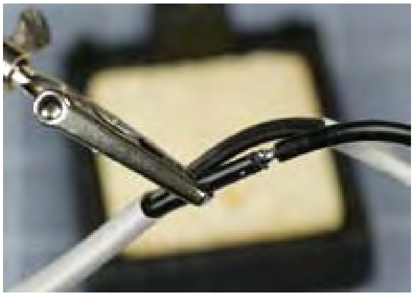
Figure 3-64

7. When the joint has cooled, slide the heat-shrink tubing over it, and apply the heat gun. Now repeat the process with the other conductor. Finally, slide the larger piece of tubing over the joint. You did remember to put the large tubing onto the wire at the beginning, didn't you?

Figures 3-59 through 3-65 show the steps all the way through to the end.

If you have completed the soldering exercises so far, you now have sufficient basic skills to solder your first electronic circuit. But first, I want you to verify the vulnerability of components to heat.

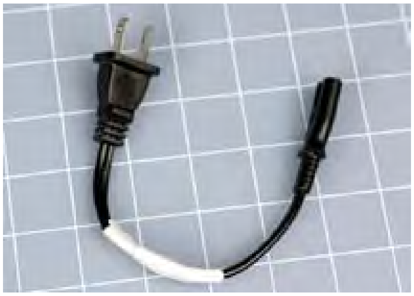
Figure 3-65. *Completion of the shortened power cord for a laptop power supply.*

Experiment 13: Broil an LED

In Chapter 1, you saw how an LED can be damaged if too much current flows through it. The electricity caused heat, which melted the LED. Unsurprisingly, you can just as easily melt it by applying too much heat to one of its leads with a soldering iron. The question is: how much heat is too much? Let's find out.

You will need:

- 30-watt or 40-watt soldering iron
- 15-watt pencil-type soldering iron
- A couple of LEDs (that are expendable)
- 680Ω resistor
- Wire cutters and sharp-nosed pliers
- "Helping hand" gadget to hold your work

I don't want you to use alligator clips to join the LED to a power supply, because the alligator clip will divert and absorb some of the heat from your soldering iron. Instead, please use some sharp-nosed pliers to bend each of the leads from an LED into little hooks, and do the same thing with the wires on a 680Ω load resistor. Finally bend the new wires on your AC adapter so that they, too, are tiny hooks. Now you can put the hooks together like links in a chain, as shown in Figure 3-66.

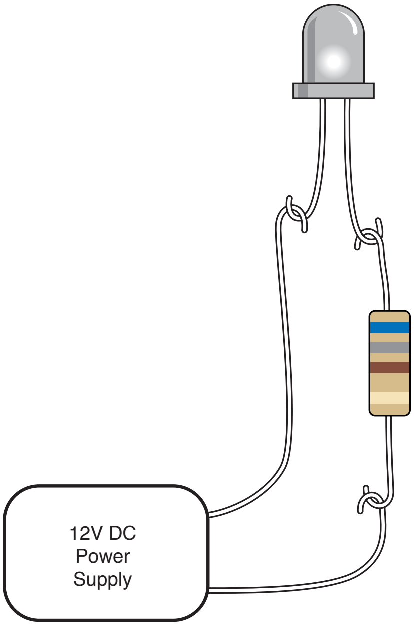

Figure 3-66. *By literally hooking together the leads from a resistor and a white-light LED, we minimize pathways for heat to escape during the subsequent test.*

Grip the plastic body of the LED in your helping hand. Plastic is not a good thermal conductor, so the helping hand shouldn't siphon too much heat away from our target. The resistor can dangle from one of the leads on the LED, and the wire from the AC adapter can hang from that, a little farther down. Gravity should be sufficient to make this work. Set your AC adapter to deliver 12 volts as before, plug it in, and your LED should be shining brightly. I used a white LED in this experiment, because it's easier to photograph.

Make sure your two soldering irons are really hot. They should have been plugged in for at least five minutes. Now take the pencil-style iron and hold its tip firmly against one of the leads on your glowing LED, while you check the time with a watch. Figure 3-67 shows the setup.

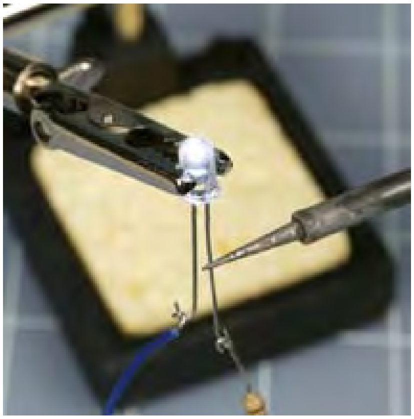

Figure 3-67. *Applying heat with a 15-watt soldering iron. A typical LED should withstand this treatment for two or three minutes, but if you substitute a 30-watt soldering iron, the LED is likely to burn out in under 15 seconds.*

I'm betting that you can sustain this contact for a full three minutes without burning out the LED. This is why you use a 15-watt soldering iron for delicate electronics work—it doesn't endanger the components.

Allow your LED wire to cool, and then apply your more powerful soldering iron to the same piece of wire as before. Again, make sure it is completely hot, and I think you'll find that the LED will go dark after as little as 10 seconds (note, some LEDs can survive higher temperatures than others). This is why you *don't* use a 30-watt soldering iron for delicate electronics work.

The large iron doesn't necessarily reach a higher temperature than the small one. It just has a larger heat capacity. In other words, a greater quantity of heat can flow out of it, at a faster rate.

Throw away your burned-out LED. Substitute a new one, connected as before, but add a full-size copper alligator clip to one of the leads up near the body of the LED, as shown in Figure 3-68. Press the tip of your 30-watt or 40-watt soldering iron against the lead just *below* the alligator clip. This time, you should be able to hold the powerful soldering iron in place for a full two minutes without burning out the LED.

Imagine the heat flowing out through the tip of your soldering iron, into the wire that leads to the LED—except that the heat meets the alligator clip along the way, as shown in Figure 3-69. The clip is like an empty vessel waiting to be filled. It offers much less resistance to heat than the remainder of the wire leading to the LED, so the heat prefers to flow into the copper clip, leaving the LED unharmed. At the end of your experiment, if you touch the clip, you'll find that it's hot, while the LED remains relatively cooler.

The alligator clip is known as a heat sink, and it should be made of copper, because copper is one of the best conductors of heat.

Because the 15-watt soldering iron failed to harm the LED, you may conclude that the 15-watt iron is completely safe, eliminating all need for a heat sink. Well, this may be true. The problem is, you don't really know whether some semiconductors may be more heat-sensitive than LEDs. Because the consequences of burning out a component are so exasperating, I suggest you should play it safe and use a heat sink in these circumstances:

- If you apply 15-watt iron extremely close to a semiconductor for 20 seconds or more.
- If you apply a 30-watt iron near resistors or capacitors for 10 seconds or more. (Never use it near semiconductors.)
- If you apply a 30-watt iron near anything meltable for 20 seconds or more. Meltable items include insulation on wires, plastic connectors, and plastic components inside switches.

Rules for Heat Sinking

1. Full-size copper alligator clips do work better.
2. Clamp the alligator clip as close as possible to the component and as far as possible from the joint. (You don't want to suck too much heat away from the joint.)
3. Make sure there is a metal-to-metal connection between the alligator clip and the wire to promote good heat transfer.

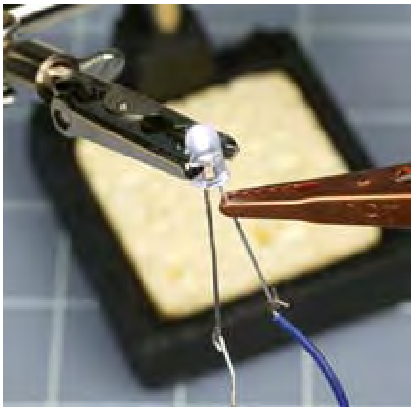

Figure 3-68. *When a copper alligator clip is used as a heat sink, you should be able to apply a 30-watt soldering iron (below the clip) without damaging the LED.*

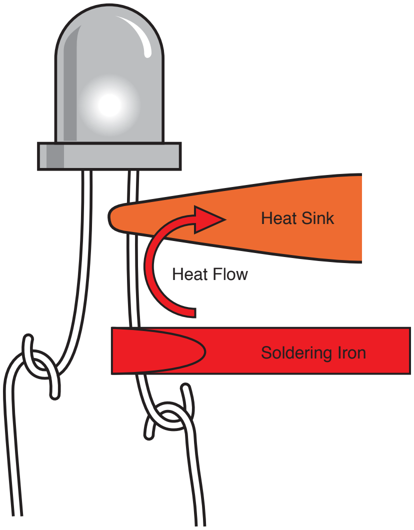

Figure 3-69. *The heat sink intercepts the heat, sucks it up, and protects the LED from damage.*

FUNDAMENTALS

All about perforated board

For the remainder of this book, you'll be using perforated board whenever you want to create permanent, soldered circuits. There are three ways to do this:

1. Point-to-point wiring. You use perforated board that has no connections behind the holes. Either the board has no copper traces on it at all, as in Figure 3-70, or you will find a little circular copper circle around each hole, as in Figure 3-71. These circles are not connected with each other and are used only to stabilize the components that you assemble.

   Point-to-point wiring allows you to place the components in a convenient, compact layout that can be very similar to a schematic. Under the board you bend the wires to link the components, and solder them together, adding extra lengths of wire if necessary. The advantage of this system is that it can be extremely compact. The disadvantage is that the layout can be confusing, leading to errors.

2. Breadboard-style wiring. Use perforated board that is printed with copper traces in exactly the same pattern as the conductors inside a breadboard. Once your circuit works on the breadboard, you move the components over to the perfboard one by one, maintaining their exact same positions relative to each other. You solder the "legs" of the components to the copper traces, which complete the circuit. Then you trim off the surplus wire. The advantage of this procedure is that it's quick, requires very little planning, and minimizes the possibility for errors. The disadvantage is that it tends to waste space. A cheap example is shown in Figure 3-72.

3. You can etch your own circuit board with customized copper traces that link your components in a point-to-point layout. This is the most professional way to complete a project, but it requires more time, trouble, and equipment than is practical in this book.

   Point-to-point wiring is like working with alligator clips, on a much smaller scale. The first soldered project will use this procedure.

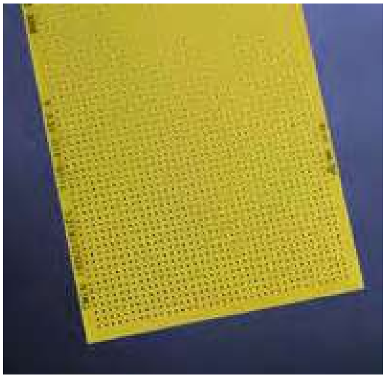

Figure 3-70

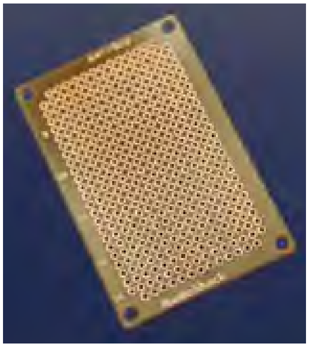

Figure 3-71. *Either this type of perforated board or the type in Figure 3-70 can be used for point-to-point wiring in Experiment 14.*

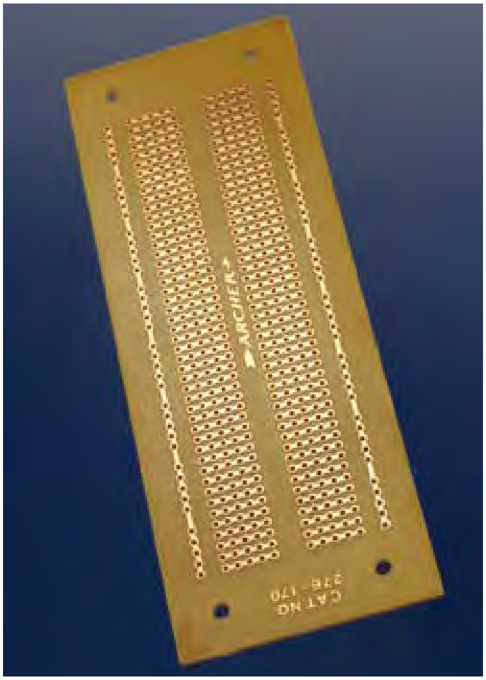

Figure 3-72. *Perforated board etched with copper in variants of a breadboard layout. This example is appropriate for Experiment 15.*

Experiment 14: A Pulsing Glow

You will need:

- Breadboard
- 15-watt pencil-type soldering iron
- Thin solder (0.022 inches or similar)
- Wire strippers and cutters
- Plain perforated board (no copper etching necessary)
- Small vise or clamp to hold your perforated board
- Resistors, various
- Capacitors, electrolytic, 100 µF and 220 µF, one of each
- Red LED, 5 mm, rated for 2 volts approximately
- 2N6027 programmable unijunction transistor

Your first circuit using a PUT was a slow-speed oscillator that made an LED flash about twice each second. The flashes looked very "electronic," by which I mean that the LED blinked on and off without a gradual transition between each state. I'm wondering if we can modify this circuit to make the LED pulse in a more gentle, interesting way, like the warning light on an Apple MacBook when it's in "sleep" mode. I'm thinking that something of this sort might be wearable as an ornament, if it's small enough and elegant enough.

I'm also thinking that this first soldering project will serve three purposes. It will test and refine your skill at joining wires together, will teach you point-to-point wiring with perfboard, and will give you some additional insight into the way that capacitors can be used to adjust timing.

Look back at the original schematic in Experiment 11, on page 82. Refresh your memory about the way it worked. The capacitor charges through a resistor until it has enough voltage to overcome the internal resistance in the PUT. Then the capacitor discharges through the PUT and flashes the LED.

If you drew a graph of the light coming out of the LED, it would be a thin, square-shaped pulse, as shown in Figure 3-73. How can we fill it out to make it more like the curve in Figure 3-74, so that the LED fades gently on and off, like a heartbeat?

One thing is obvious: the LED is going to be emitting a greater total amount of light in each cycle. Therefore, it's going to need more power. This means that C1, in Figure 3-75, must be a larger capacitor.

When we have a larger capacitor, it takes longer to charge. To keep the flashes reasonably frequent, we'll need a lower-value resistor for R1 to charge the capacitor quickly enough. In addition, reducing the values of R2 and R3 will program the PUT to allow a longer pulse.

Most important, I want to discharge the capacitor through a resistor to make the onset of the pulse gradual instead of sudden. Remember, when you have a resistor in series with a capacitor, the capacitor not only charges more slowly, but discharges more slowly.

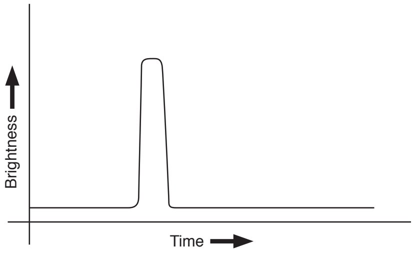

Figure 3-73

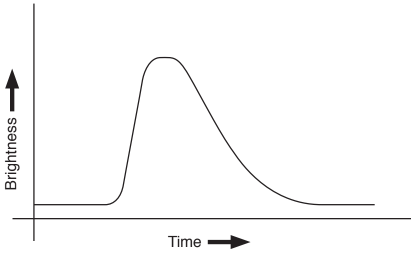

Figure 3-74. *The original PUT oscillator circuit in Experiment 11 made the LED emit sharp, short flashes. The upper graph shows what we might find if we measured light output over time. The second graph shows a gentler onset to each flash, followed by a slow fade-out. Capacitors can be used to create this effect.*

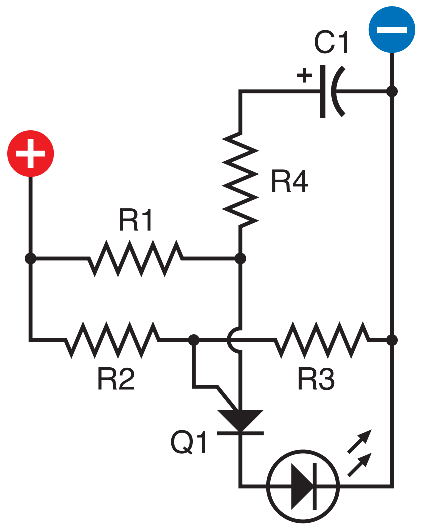

Figure 3-75. *The first step toward creating a gentler flashing effect is to use a larger capacitor for C1 and discharge it through a resistor, R4. Lower-value resistors are necessary to charge the capacitor rapidly enough.*

R1: 33K
R2: 1K
R3: 1K
R4: 1K
C1: 100 µF electrolytic
Q1: 2N6027

Figure 3-75 shows these features. Compare it with Figure 2-103 on page 85. R1 is now 33K instead of 470K. R2 and R3 are reduced to 1K. R4 also is 1K, to so that the capacitor takes longer to discharge through it. And C1 is now 100 µF instead of 2.2 µF.

Assemble this circuit on a breadboard, and compare the results when you include R4 or bypass it with a plain piece of jumper wire. It softens the pulse a bit, but we can work on it some more. On the output side of the PUT, we can add another capacitor. This will charge itself when the pulse comes out of the PUT, and then discharge itself gradually through another resistor, so that the light from the LED dies away more slowly.

Figure 3-76 shows the setup. C2 is large—220 µF—so it sucks up the pulse that comes out of the PUT, and then gradually releases it through 330Ω resistor R5 and the LED. You'll see that the LED behaves differently now, fading out inside of blinking off. But the resistances that I've added have dimmed the LED, and to brighten it, you should increase the power supply from 6 volts to 9 volts.

Remember that a capacitor imposes a smoothing effect only if one side of it is grounded to the negative side of the power supply. The presence of the negative charge on that side of the capacitor attracts the positive pulse to the other side.

I like the look of this heartbeat effect. I can imagine a piece of wearable electronic jewelry that pulses in this sensual way, very different from the hard-edged, sharp-on-and-off of a simple oscillator circuit. The only question is whether we can squeeze the components into a package that is small enough to wear.

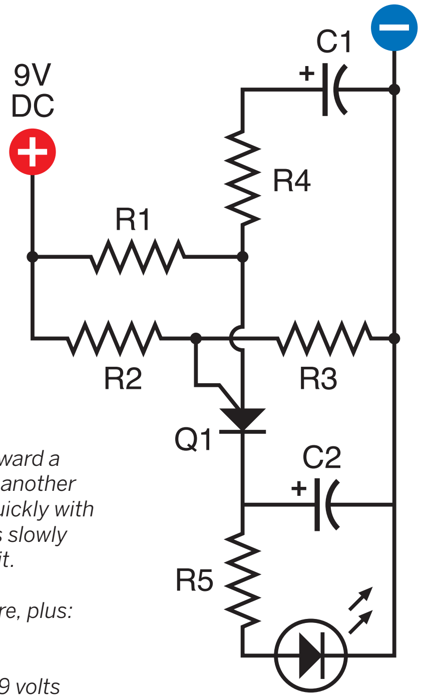

Figure 3-76. *The second step toward a gentler flashing effect is to add another capacitor, C2, which charges quickly with each pulse and then discharges slowly through R5 and the LED below it.*

Same components as before, plus:
R5: 330Ω
C2: 220 µF electrolytic
Power supply increased to 9 volts

Figure 3-77. *On a dark night in a rural area, the heartbeat flasher may be attractive in unexpected ways.*

Resizing the Circuit

The first step is to look at the physical components and imagine how to fit them into a small space. Figure 3-78 shows a 3D view of a compact arrangement. Check this carefully, tracing all the paths through the circuit, and you'll see that it's the same as the schematic. The trouble is that if we solder the components together like this, they won't have much strength. All the little wires can bend easily, and there's no easy way to mount the circuit in something or on something.

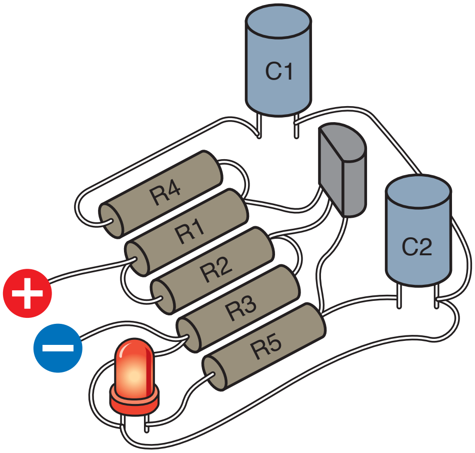

Figure 3-78. *This layout of components replicates their connections in the schematic diagram while squeezing them into a minimal amount of space.*

The answer is to put it on a substrate, which is one of those terms that people in the electronics field like to use, perhaps because it sounds more technical than "perfboard." But perforated board is what we need, and Figure 3-79 shows the components transferred onto a piece of board measuring just 1 inch by 0.8 inch.

The center version of this diagram uses dotted lines to show how the components will be connected with each other underneath the board. Mostly the leads that stick out from underneath the components will be long enough to make these connections.

Finally, the bottom version of the perfboard diagram shows the perfboard flipped left-to-right (notice the L and the R have been transposed to remind you, and I've used a darker color to indicate the underside of the board). Orange circles indicate where solder joints will be needed.

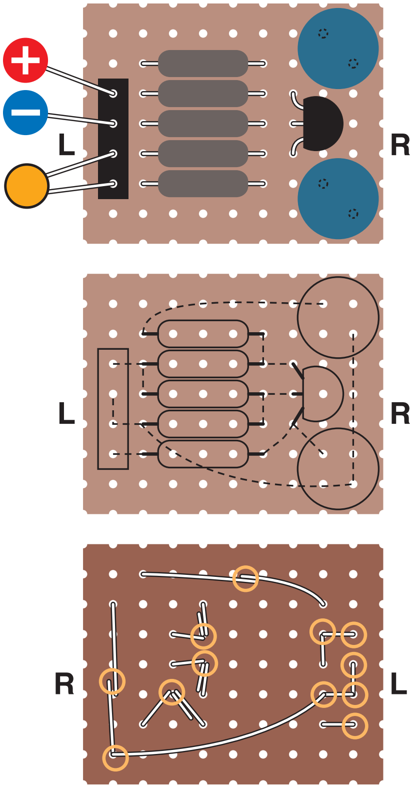

Figure 3-79. *Perforated board can be used to support the layout of components. Their leads are soldered together under the board to create the circuit. The middle diagram shows the wires under the board as dashed lines. The bottom diagram shows the board from underneath, flipped left to right. Orange circles indicate where solder joints will be necessary.*

The LED should be unpluggable, because we may want to run it at some distance from the circuit. Likewise the power source should be unpluggable. Fortunately we can buy miniature connectors that fit right into the perforated board. You may have to go to large online retail suppliers such as Mouser.com for these. Some manufacturers call them "single inline sockets and headers," while others call them "boardmount sockets and pinstrip headers." Refer back to Figure 3-29 and check the shopping list for more details.

This is a very compact design that will require careful work with your pencil-style soldering iron. Because a piece of perforated board as small as this will tend to skitter around, I suggest that you apply your miniature vise to one end to anchor it with some weight while still allowing you to turn it easily.

When I'm working on this kind of project, I like to place it (with the vise attached) on a soft piece of polyurethane foam—the kind of slab that is normally used to make a chair cushion. The foam protects the components from damage when the board is upside-down, and again helps to prevent the work from sliding around unpredictably.

Step by Step

Here's the specific procedure for building this circuit:

1. Cut the small piece of perfboard out of a sheet that has no copper traces on it. You can cut the section using your miniature hobby saw, or you may be able to snap the board along its lines of holes, if you're careful. Alternatively, use a small ready-cut piece of perfboard with copper circles on it that are not connected to one another. You'll ignore the copper circles in this project. (In the next experiment, you'll deal with the additional challenge of making connections between components and copper traces on perforated board.)
2. Gather all the components and carefully insert them through holes in the board, counting the holes to make sure everything is in the right place. Flip the board over and bend the wires from the components to anchor them to the board and create connections as shown. If any of the wires isn't long enough, you'll have to supplement it with an extra piece of 22-gauge wire from your supply. You can remove all the insulation, as we'll be mounting the perfboard on a piece of insulating plastic.
3. Trim the wires approximately with your wire cutters.
4. Make the joints with your pencil soldering iron. Note that in this circuit, you are just joining wires to each other. The components are so close together that they'll prevent each other from wiggling around too much. If you are using board with copper pads (as I did), and some solder connects with them, that's OK—as long as it doesn't creep across to the neighboring component and create a short circuit.
5. Check each joint using a close-up magnifying glass, and wiggle it with pointed-nosed pliers. If there isn't enough solder for a really secure joint, reheat it and add more. If solder has created a connection that shouldn't be there, use a utility knife to make two parallel cuts in the solder, and scrape away the little section between them.

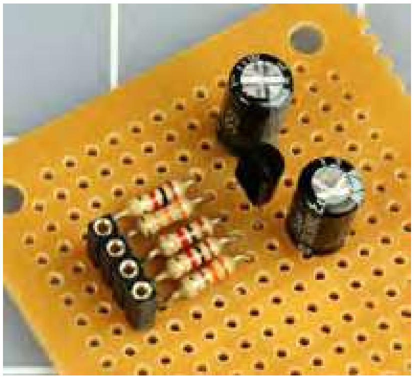

Figure 3-80. *Components mounted on a piece of perforated board.*

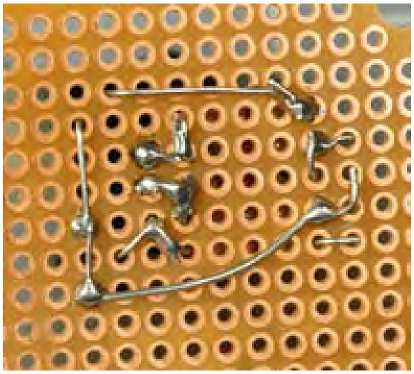

Figure 3-81. *The assembly seen from below. The copper circles around the holes are not necessary for this project. Some of them have picked up some solder, but this is irrelevant as long as no unintentional short circuits are created.*

Generally, I insert three or four components, trim the wires approximately, solder them, trim their wires finally, then pause to check the joints and the placement. If I solder too many components in succession, there's a greater risk of missing a bad joint, and if I make an error in placing a component, undoing it will be much more problematic if I have already added a whole lot more components around it.

Figures 3-80 and 3-81 show the version of this project that I constructed, before I trimmed the board to the minimum size.

Flying Wire Segments

The jaws of your wire cutters exert a powerful force that peaks and then is suddenly released when they cut through wire. This force can be translated into sudden motion of the snipped wire segment. Some wires are relatively soft, and don't pose a risk, but harder wires can fly in unpredictable directions at high speed, and may hit you in the eye. The leads of transistors are especially hazardous in this respect.

I think it's a good idea to wear safety glasses when trimming wires.

Finishing the Job

I always use bright illumination. This is not a luxury; it is a necessity. Buy a cheap desk lamp if you don't already have one. I use a daylight-spectrum fluorescent desk lamp, because it helps me identify the colored bands on resistors more reliably. Note that this type of fluorescent lamp emits quite a lot of ultraviolet light, which is not good for the lens in your eye. Avoid looking closely and directly at the tube in the lamp, and if you wear glasses, they will provide additional protection.

No matter how good your close-up vision is, you need to examine each joint with that close-up magnifier. You'll be surprised how imperfect some of them are. Hold the magnifier as close as possible to your eye, then pick up the thing that you want to examine and bring it closer until it comes into focus.

Finally, you should end up with a working circuit. You can insert the wires from your power supply into two of the tiny power sockets, and plug a red LED into the remaining two sockets. Remember that the two center sockets are negative, and the two outer sockets are positive, because it was easier to wire the circuit this way. You should color-code them to avoid mistakes.

So now you have a tiny circuit that pulses like a heartbeat. Or does it? If you have difficulty making it work, retrace every connection and compare it with the schematic. If you don't find an error, apply power to the circuit, attach the black lead from your meter to the negative side, and then go around the circuit with the red lead, checking the presence of voltage. Every part of this circuit should show at least some voltage while it's working. If you find a dead connection, you may have made a bad solder joint, or missed one entirely.

When you're done, now what? Well, now you can stop being an electronics hobbyist and become a crafts hobbyist. You can try to figure out a way to make this thing wearable.

First you have to consider the power supply. Because of the components that I used, we really need 9 volts to make this work well. How are you going to make this 9-volt circuit wearable, with a bulky 9-volt battery?

I can think of three answers:

1. You can put the battery in a pocket, and mount the flasher on the outside of the pocket, with a thin wire penetrating the fabric. Note that the tiny power connector on the perforated board will accept two 22-gauge wires if they are solid core, or if they are stranded (like the wires from a 9-volt battery connector) but have been thinly coated with solder.
2. You could mount the battery inside the crown of a baseball cap, with the flasher on the front.
3. You can put together three 3-volt button batteries in a stack, held in some kind of plastic clip. If you try this option, it may not be a good idea to try to solder wire to a battery. You will heat the liquid stuff inside the battery, which may not be good for it, and may not be good for you if the liquid starts boiling and the battery bursts open. Also, solder doesn't stick easily to the metallic finish on most battery terminals.

Most LEDs create a sharply defined beam of light, which you may want to diffuse to make it look nicer. One way to do this is to use a piece of transparent acrylic plastic, at least 1/4 inch thick, as shown in Figure 3-82. Sandpaper the front of the acrylic, ideally using an orbital sander that won't make an obvious pattern. Sanding will make the acrylic translucent rather than transparent.

Drill a hole slightly larger than the LED in the back of the acrylic. Don't drill all the way through the plastic. Remove all fragments and dust from the hole by blasting some compressed air into it, or by washing it if you don't have an air compressor. After the cavity is completely dry, get some transparent silicone caulking or mix some clear five-minute epoxy and put a drop in the bottom of the hole. Then insert the LED, pushing it in so that it forces the epoxy to ooze around it, making a tight seal. See Figure 3-82.

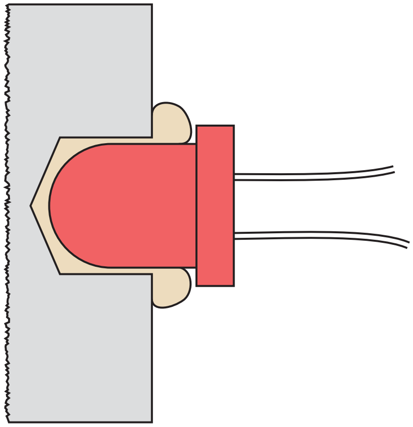

Figure 3-82. *This cross-sectional view shows a sheet of transparent acrylic in which a hole has been drilled part of the way from the back toward the front. Because a drill bit creates a hole with a conical shape at the bottom, and because the LED has rounded contours, transparent epoxy or silicone caulking can be injected into the hole before mounting the LED.*

Try illuminating the LED, and sand the acrylic some more if necessary. Finally, you can decide whether to mount the circuit on the back of the acrylic, or whether you want to run a wire to it elsewhere.

Because the LED will flash at about the speed of a human heart while the person is resting, it may look as if it's measuring your pulse, especially if you mount it on the center of your chest or in a strap around your wrist. If you enjoy hoaxing people, you can suggest that you're in such amazingly good shape, your pulse rate remains constant even when you're taking strenuous exercises.

To make a good-looking enclosure for the circuit, I can think of options ranging from embedding the whole thing in clear epoxy to finding a Victorian-style locket. I'll leave you to consider alternatives, because this is a book about electronics rather than handicrafts.

However, I will address one final issue: how long will this gadget continue flashing?

If you check the following section "Essentials: Battery life," you'll find that a regular alkaline 9-volt battery should keep the LED flashing for about 50 hours.

ESSENTIALS

Battery life

Any time you finish a circuit that you intend to run from a battery, you'll want to calculate the likely battery life. This is easily done, because manufacturers rate their batteries according to the "ampere hours" they can deliver. Keep the following in mind:

- The abbreviation for amp-hours is Ah, sometimes printed as AH. Milliampere-hours are abbreviated mAh.
- The rating of a battery in amp-hours is equal to the current, in amps, multiplied by the number of hours that the battery can deliver it.

Thus, in theory 1 amp-hour can mean 1 amp for 1 hour, or 0.1 amp for 10 hours, or 0.01 amp for 100 hours—and so on. In reality, it's not as simple as this, because the chemicals inside a battery become depleted more quickly when you draw a heavy current, especially if the battery gets hot. You have to stay within limits that are appropriate to the size of the battery.

For instance, if a small battery is rated for 0.5 amp-hours, you can't expect to draw 30 amperes from it for 1 minute. But you should be able to get 0.005 amps (i.e., 5 milliamps) for 100 hours without any trouble. Remember, though, that the voltage delivered by a battery will be greater than its rated voltage when the battery is fresh, and will diminish below its rated voltage while the battery is delivering power.

According to some test data that I trust (I think they are a little more realistic than the estimates supplied by battery manufacturers), here are some numbers for typical batteries:

- Typical 9 volt alkaline battery: 0.3 amp-hours, while delivering 100 mA.
- Typical AA size, 1.5-volt alkaline battery: 2.2 amp-hours, while delivering 100 mA.
- Rechargeable nickel-metal hydride battery: about twice the endurance of a comparably sized alkaline battery.
- Lithium battery: maybe three times the endurance of an alkaline battery.

BACKGROUND

Maddened by measurement

Throughout most of this book, I've mostly used measurements in inches, although sometimes I've digressed into the metric system, as when referring to "5-mm LEDs." This isn't inconsistency on my part; it reflects the conflicted state of the electronics industry, where you'll find inches and millimeters both in daily use, often in the very same data sheet.

The United States is the only major nation still using the old system of units that originated in England. (The other two holdouts are Liberia and Myanmar, according to the CIA's *World Factbook*.) Still, the United States has led many advances in electronics, especially the development of silicon chips, which have contacts spaced 1/10 inch apart. These standards became firmly established, and show no sign of disappearing.

To complicate matters further, even in the United States, you can encounter two incompatible systems for expressing fractions of an inch. Drill bits, for instance, are measured in multiples of 1/64 inch, while metal thicknesses may be measured in decimals such as 0.06 inch (which is approximately 1/16 inch).

The metric system is not necessarily more rational than the U.S. system. Originally, when the metric system was formally introduced in 1875, the meter was defined as being 1/10,000,000 of the distance between the North Pole and the equator, along a line passing through Paris—a quixotic, Francocentric conceit. Since then, the meter has been redefined three times, in a series of efforts to achieve greater accuracy in scientific applications.

As for the usefulness of a 10-based system, moving a decimal point is certainly simpler than doing calculations in 64ths of an inch, but the only reason we count in tens is because we happen to have evolved with that number of digits on our hands. A 12-based system would really be more convenient, as numbers would be evenly divisible by 2 and 3.

As we're stuck with the whimsical aspects of length measurement, I've created the charts in Figures 3-83 and 3-84 to assist you in going from one system to another. From these you will see that when you need to drill a hole for a 5 mm LED, a 3/16-inch drill bit is about right. (In fact, it results in a better, tighter fit than if you drill an actual 5 mm hole.)

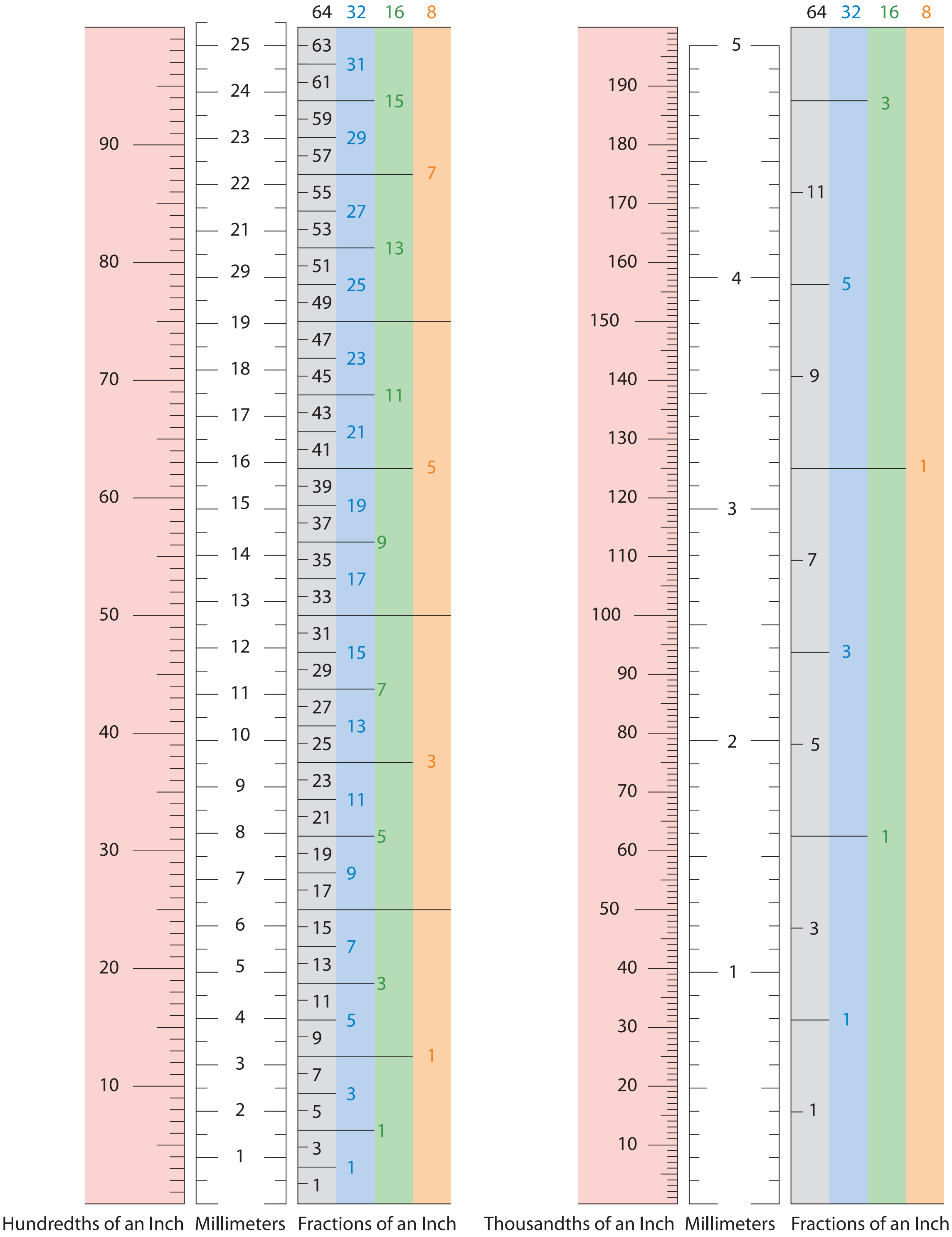

Figure 3-83. *Because units of measurement are not standardized in electronics, conversion is often necessary. The chart on the right is a 5x magnification of the bottom section of the chart on the left.*

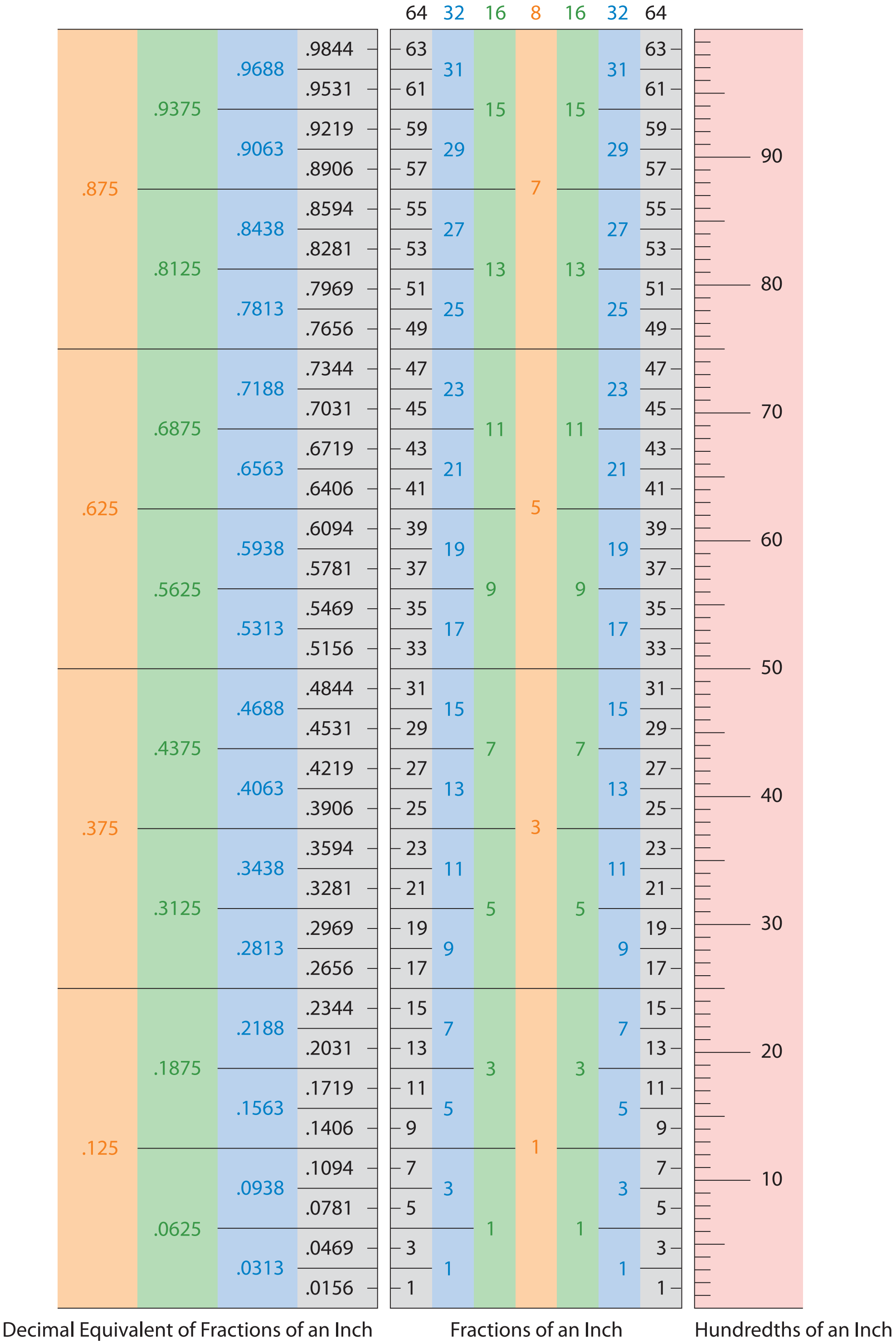

				64	32	16	8	16	32	64
.875	.9375	.9688	.9844	63	31	15	7	15	31	63
			.9531	61						61
		.9063	.9219	59	29				29	59
			.8906	57						57
	.8125	.8438	.8594	55	27	13		13	27	55
			.8281	53						53
		.7813	.7969	51	25				25	51
			.7656	49						49
.625	.6875	.7188	.7344	47	23	11	5	11	23	47
			.7031	45						45
		.6563	.6719	43	21				21	43
			.6406	41						41
	.5625	.5938	.6094	39	19	9		9	19	39
			.5781	37						37
		.5313	.5469	35	17				17	35
			.5156	33						33
.375	.4375	.4688	.4844	31	15	7	3	7	15	31
			.4531	29						29
		.4063	.4219	27	13				13	27
			.3906	25						25
	.3125	.3438	.3594	23	11	5		5	11	23
			.3281	21						21
		.2813	.2969	19	9				9	19
			.2656	17						17
.125	.1875	.2188	.2344	15	7	3	1	3	7	15
			.2031	13						13
		.1563	.1719	11	5				5	11
			.1406	9						9
	.0625	.0938	.1094	7	3	1		1	3	7
			.0781	5						5
		.0313	.0469	3	1				1	3
			.0156	1						1

Figure 3-84. *This chart allows conversion between hundredths of an inch, conventional U. S. fractions of an inch, and fractions expressed in thousandths of an inch.*

Experiment 15: Intrusion Alarm Revisited

Time now to add some of the enhancements to the intrusion alarm that I discussed at the end of Experiment 11. I'm going to show you how the alarm can be triggered if you install various detectors on windows and doors in your home. I'll also show how the alarm can be wired so that it locks itself on and continues to make noise even after a door or window is reclosed.

This experiment will demonstrate the procedure for transferring a project from a breadboard to a piece of perforated board that has copper connections laid out identically to the ones inside the breadboard, as shown earlier in Figure 3-72. And you'll mount the finished circuit in a project box with switches and connectors on the front.

When all is said and done, you'll be ready for wholesale circuit building. The explanations in the rest of this book will get gradually briefer, and the pace will increase.

You will need:

- 15-watt pencil-type soldering iron
- Thin solder (0.022 inches or similar)
- Wire strippers and cutters
- Perforated board etched with copper in a breadboard layout
- Small vise or clamp to hold your perforated board
- The same components that you used in Experiment 11, plus:
    - 2N2222 NPN transistor. Quantity: 1.
    - DPDT relay. Quantity: 1.
    - SPDT toggle switch. Quantity: 1.
    - 1N4001 diode. Quantity: 1.
    - Red and green 5mm LEDs. Quantity: 1 each.
    - Project box, 6 × 3 × 2 inches.
    - Power jack, type N, and matching power socket, type N.
    - Binding posts.
    - Stranded 22-gauge wire, three different colors.
    - Magnetic sensor switches, sufficient for your home.
    - Alarm network wiring, sufficient for your home.

Magnetic Sensor Switches

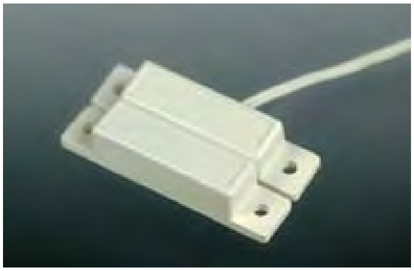

Figure 3-85. *In this simple alarm sensor switch, the lower module contains a magnet, which opens and closes a reed switch sealed into the upper module.*

A typical alarm sensor switch consists of two modules: the magnetic module and the switch module, as shown in Figures 3-85 and 3-86. The magnetic module contains a permanent magnet, and nothing else. The switch module contains a "reed switch," which makes or breaks a connection (like a contact inside a relay) under the influence of the magnet. When you bring the magnetic module close to the switch module, you may faintly hear the reed switch click as it flips from one state to the other.

Like all switches, reed switches can be normally open or normally closed. For this project, you want the kind of switch that is normally open, and closes when the magnetic module is close to it.

Attach the magnetic module to the moving part of a door or window, and attach the switch module to the window frame or door frame. When the window or door is closed, the magnetic module is almost touching the switch module. The magnet keeps the switch closed until the door or window is opened, at which point the switch opens.

The only question is: how do we use this component to trigger our alarm? As long as a small current flows through all our magnetic sensor switches, the alarm should be off, but if the flow of current stops, the alarm should switch on.

We could use a relay that is "always on" while the alarm is armed. When the circuit is interrupted, the relay relaxes and its other pair of contacts closes, which could power up the alarm noisemaker.

But I don't like this idea. Relays take significant power, and they can get hot. Most of them are not designed to be kept "always-on." I'd prefer to handle the task using a transistor.

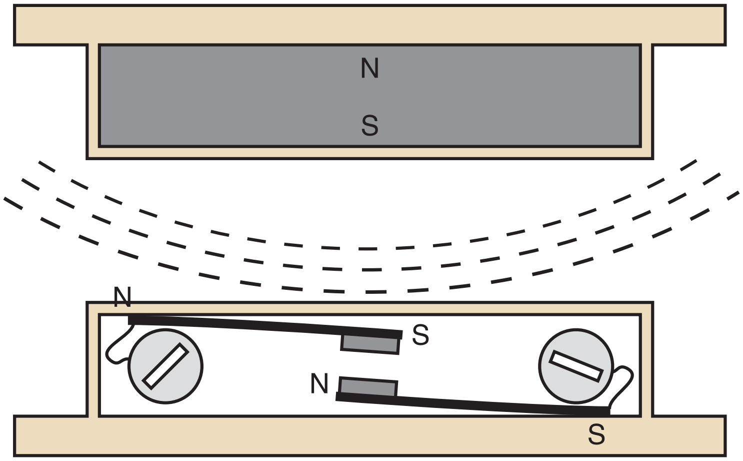

Figure 3-86. *This cutaway diagram shows a reed switch (bottom) and the magnet that activates it (top), inside an alarm sensor. The switch contains two flexible magnetized strips, the upper one with its south pole adjacent to an electrical contact, the lower one with its north pole adjacent to an electrical contact. When the south pole of the magnet approaches the switch, the magnetic force (shown as dashed lines) repels the south contact and attracts the north contact, causing them to snap together. Two screws on the outside of the casing are connected with the strips inside.*

A Break-to-Make Transistor Circuit

First, recall how an NPN transistor works. When the base is not sufficiently positive, the transistor blocks current between its collector and emitter, but when the base is relatively positive, the transistor passes current.

Take a look at the schematic in Figure 3-87, which is built around our old friend the 2N2222 NPN transistor. When the switch is closed, it connects the base of the transistor to the negative side of the power supply through a 1K resistor. At the same time, the base is connected with the positive side of the power supply through a 10K resistor. Because of the difference in resistances and the relatively high turn-on voltage for the LED, the base is forced below its turn-on threshold, and as a result, the transistor will not pass much current. The LED will glow dimly at best.

Now what happens when the switch is opened? The base of the transistor loses its negative power supply and has only its positive power supply. It becomes much more positive, above the turn-on threshold for the transistor, which tells the transistor to lower its resistance and pass more current. The LED now glows brightly. Thus, when the switch is turned off and breaks the connection, the LED is turned on.

This seems to be what we want. Imagine a whole series of switches instead of just one switch, as shown in Figure 3-88. The circuit will still work the same way, even if the switches are scattered all over your home, because the resistance in the wires connecting the switches will be trivial compared with the resistance of the 1K resistor.

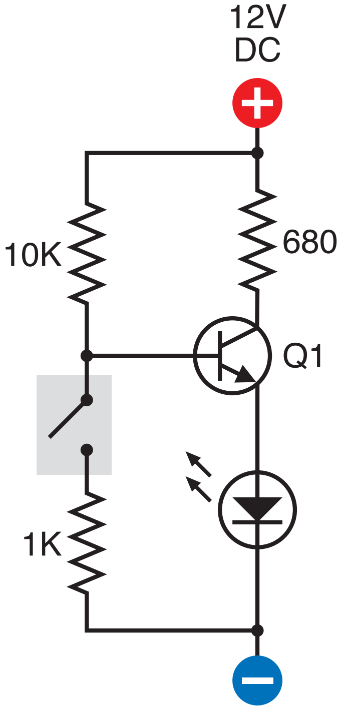

Figure 3-87. *In this demonstration circuit, when the switch is opened, it interrupts negative voltage to the base of the transistor, causing the transistor to lower its resistance, allowing current to reach the LED. Thus, when the switch is turned off, it turns on the LED.*

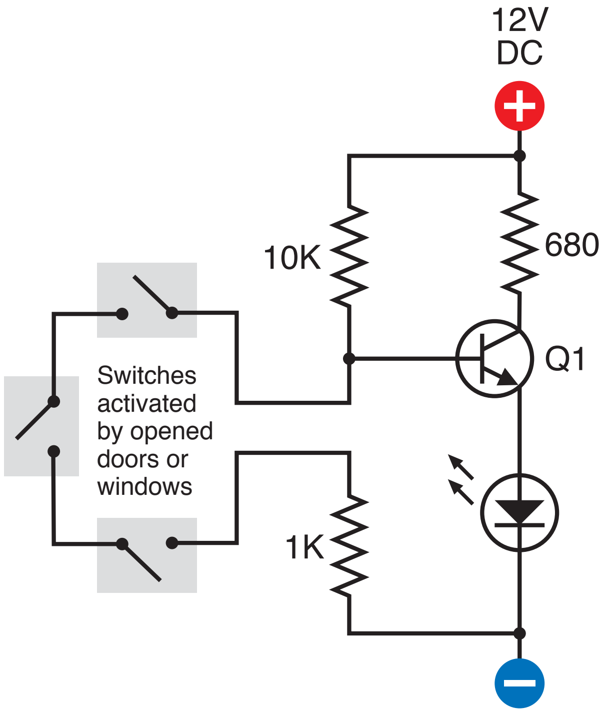

Figure 3-88. *A network of switches, wired in series, can be substituted for the single switch in Figure 3-87. Now any one switch will break continuity and trigger the transistor.*

I have shown the switches open, because that's the way the schematic for a switch is drawn, but imagine them all closed. The base of the transistor will now be supplied through the long piece of wire connecting all the closed switches, and the LED will stay dark. Now if just one switch is opened, or if anyone tampers with the wire linking them, the base of the transistor loses its connection to negative power, at which point the transistor conducts power and the LED lights up.

While all the switches remain closed, the circuit is drawing very little current—probably about 1.1 mA. So you could run it from a typical 12-volt alarm battery.

Now suppose we swap out the LED and put a relay in there instead, as shown in Figure 3-89. I don't mind using a relay in this location, because the relay will not be "always on." It will normally be off, and will draw power only when the alarm is triggered.

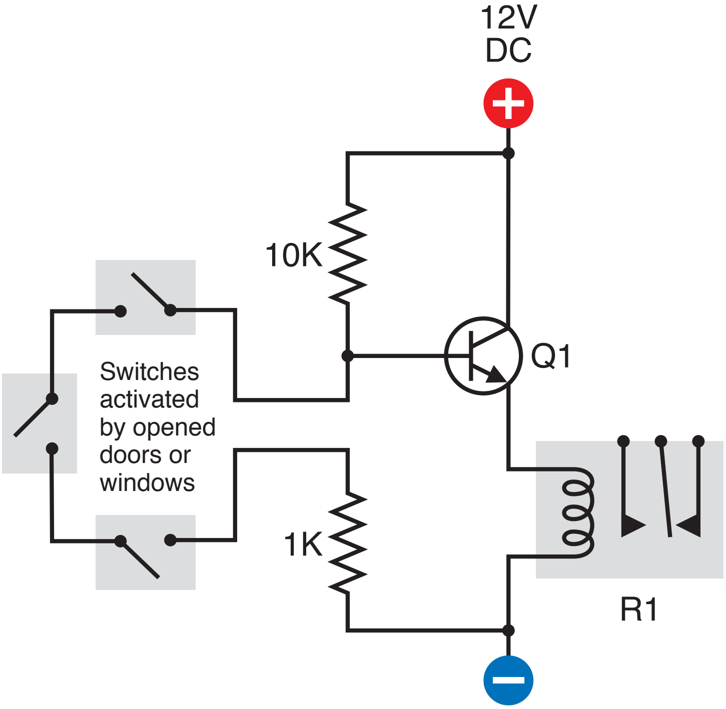

Figure 3-89. *If the LED and 680Ω resistor shown in are removed, and a relay takes their place, the relay will be activated when any switch in the sensor network is opened.*

Try one of the 12-volt relays that you used previously. You should find that when you open the switch, the relay is energized. When you close the switch, the relay goes back to sleep. Note that I eliminated the 680Ω resistor from the circuit, because the relay doesn't need any protection from the 12-volt power supply.

Self-Locking Relay

There's only one remaining problem: we want the alarm to continue making noise even after someone who has opened a door or window closes it again quickly. In other words, when the relay is activated, it must lock itself on.

One way to do this would be by using a latching relay. The only problem is that we would then need another piece of circuitry to unlatch it. I prefer to show you how you can make any relay keep itself switched on after it has received just one jolt of power. This idea will be useful to you later in the book as well.

The secret is to supply power to the relay coil through the two contacts inside the relay that are normally open. (Note that this is exactly opposite to the relay oscillator, which supplied power to its coil through the contacts that were normally closed. That setup caused the relay to switch itself off almost as soon as it switched itself on. This setup causes the relay to keep itself switched on, as soon as it has been activated.)

In Figure 3-90, the four schematics illustrate this. You can imagine them as being like frames in a movie, photographed microseconds apart. In the first picture, the switch is open, the relay is not energized, and nothing is happening. In the second, the switch has been closed to energize the coil. In the third, the coil has pulled the contact inside the relay, so that power now reaches the coil via two paths. In the fourth, the switch has been opened, but the relay is still powering its own coil through its contacts. It will remain locked in this state until the power is disconnected.

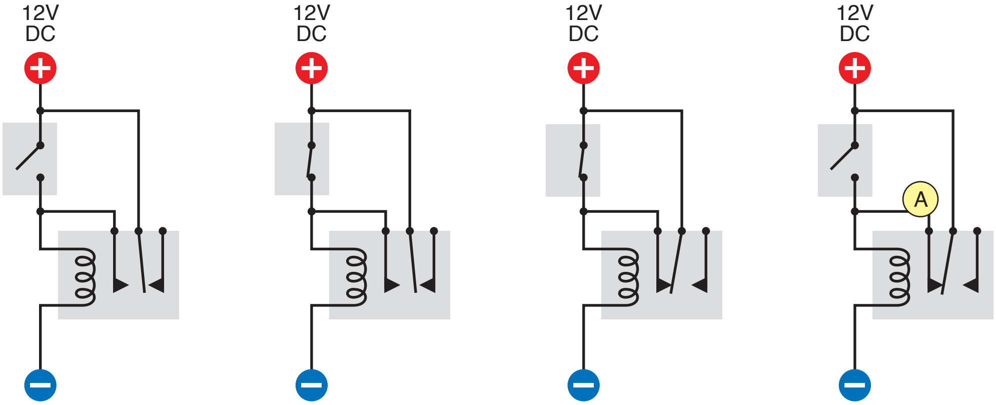

Figure 3-90. *This sequence of schematics shows the events that occur when a relay is energized. Initially, the switch is open. Then the switch is closed, activating the relay. The relay then powers itself through its own internal contacts. The relay remains energized even after the switch is opened again. Power switched by the relay can be taken from the circuit at point A.*

All we need to do, to make use of this idea, is to substitute the transistor for the on/off switch, and tap into the circuit at point A, running a wire from there to the noisemaking module.

Figure 3-91 shows how that would work. When the transistor is activated by any of the network of sensor switches, as previously explained, the transistor conducts power to the relay. The relay locks itself on, and the transistor becomes irrelevant.

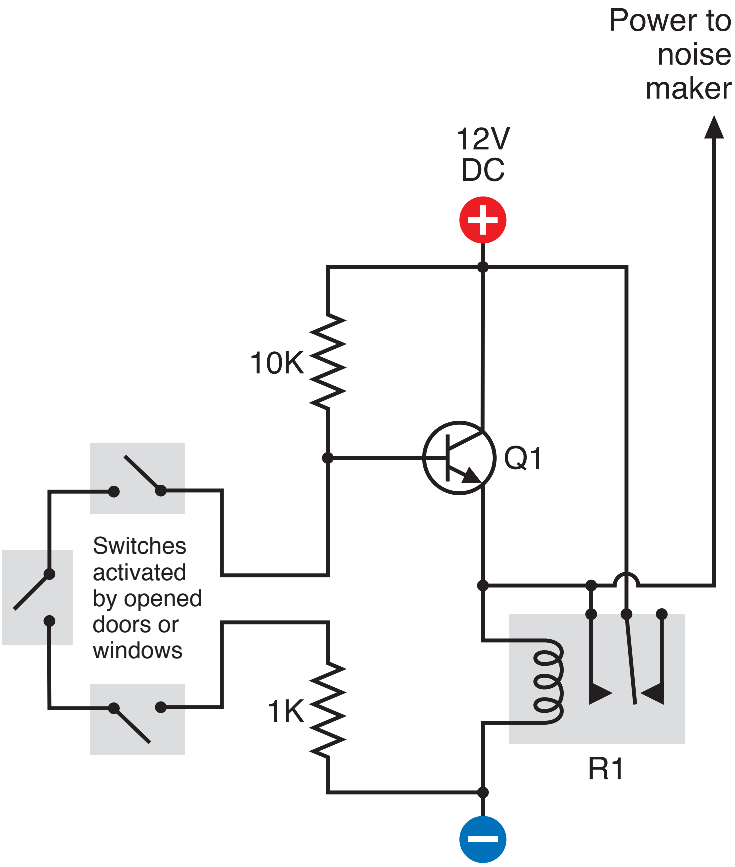
Power to
noise
maker
12V
DC
+
10K
Q1
Switches
activated
by opened
doors or
windows
1K
R1
−

Figure 3-91. *The self-locking relay depicted in has been incorporated in the alarm circuit, so that if any switch in the network is opened, the relay will continue to power the noise maker even if the switch is closed again.*

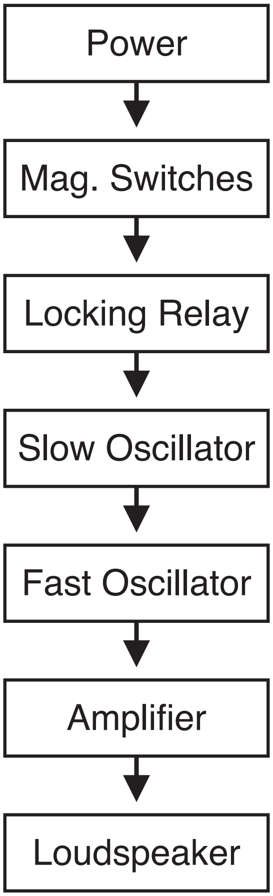
Power
Mag. Switches
Locking Relay
Slow Oscillator
Fast Oscillator
Amplifier
Loudspeaker

Figure 3-92. *This block diagram previously shown in Figure 2-112 on page 90 has been updated to include the magnetic-switch network and locking-relay control system.*

Because I've been adding pieces to the original alarm noisemaker circuit, I've updated the block diagram from Figure 2-112 to show that we can still break this down into modules with simple functions. The revised diagram is shown in Figure 3-92.

Blocking Bad Voltage

One little problem remains: in the new version of the circuit, if the transistor goes off while the relay is still on, current from the relay can flow back up the wire to the emitter of the transistor, where it will try to flow backward through the transistor to the base, which is "more negative," as it is linked through all the magnetic switches and the 1K resistor to the negative side of the power supply.

Applying power backward through a transistor is not a nice thing to do. Therefore the final schematic in this series shows one more component, which you have not seen before: a diode, labeled D1. See Figure 3-93. The diode looks like the heart of an LED, and indeed, that's pretty much what it is, although some diodes are much more robust. It allows electricity to flow in only one direction, from positive to negative, as shown by its arrow symbol. If current tries to flow in the opposite direction, the diode blocks it. The only price you pay for this service is that the diode imposes a small voltage drop on electricity flowing in the "OK" direction.

So now, positive flow can pass from the transistor, through the diode, to the relay coil, to get things started. The relay then supplies itself with power, but the diode prevents the positive voltage from getting back into the transistor the wrong way.

Perhaps a more elegant solution to the problem is to connect the NO leg of the relay via a 10k resistor to the base connection. When the relay is not energized, the NO leg is inert and simply behaves as a parasitic capacitance on the node. When the relay becomes energized, the NO leg shunts +12V through the common terminal via a 10k resistor into the base of the transistor. In this circuit configuration, the transistor is never exposed to a potentially harmful voltage and you are not depending on leakage currents of non-ideal elements to protect devices.

However, I needed an opportunity to introduce you to the concept of diodes. You can check the following section "Essentials: All about diodes" to learn more.

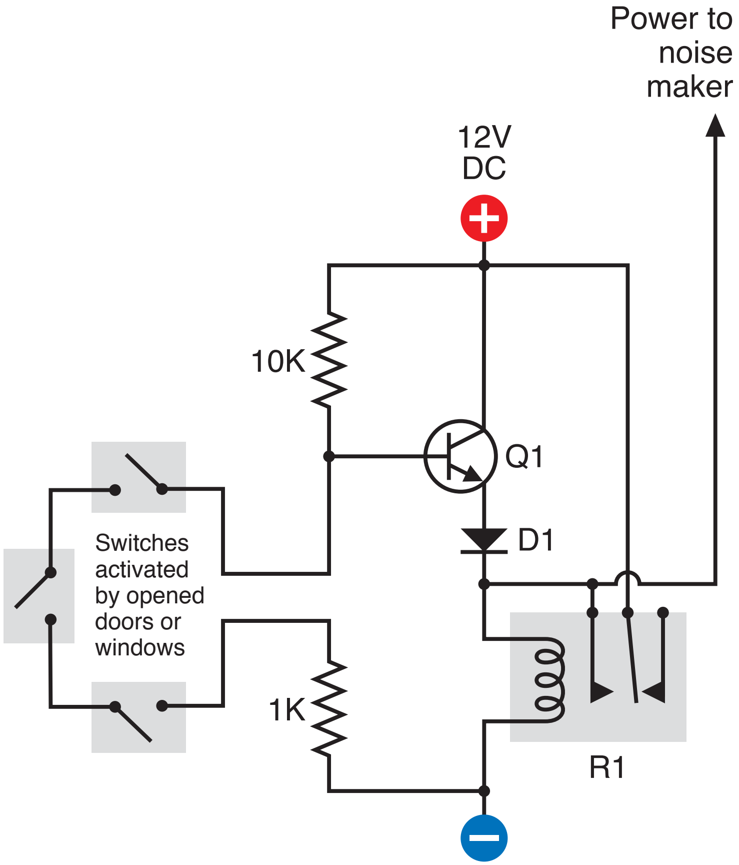

Figure 3-93. *Diode D1 has been added to protect the emitter of Q1 from positive voltage when the relay is energized.*

ESSENTIALS

All about diodes

A diode is a very early type of semiconductor. It allows electricity to flow in one direction, but blocks it in the opposite direction. (A light-emitting diode is a much more recent invention.) Like an LED, a diode can be damaged by reversing the voltage and applying excessive power, but most diodes generally have a much greater tolerance for this than LEDs. The end of the diode that blocks positive voltage is always marked, usually with a circular band, while the other end remains unmarked. Diodes are especially useful in logic circuits, and can also convert alternating current (AC) into direct current (DC).

A Zener diode is a special type that we won't be using in this book. It blocks current completely in one direction, and also blocks it in the other direction until a threshold voltage is reached—much like a PUT.

Signal diodes are available for various different voltages and wattages. The 1N4001 diode that I recommend for the alarm activation circuit is capable of handling a much greater load at a much higher voltage, but I used it because it has a low internal resistance. I wanted the diode to impose a minimal voltage drop, so that the relay would receive as much voltage as possible.

It's good practice to use diodes at less than their rated capacity. Like any semiconductor, they can overheat and burn out if they are subjected to mistreatment.

The schematic symbol for a diode has only one significant variant: sometimes the triangle is outlined instead of filled solid black (see Figure 3-94).

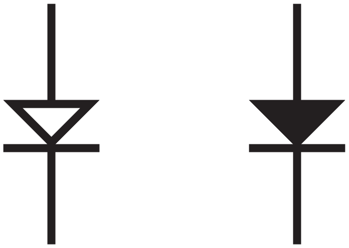

Figure 3-94. *Either of these schematic symbols may be used to represent a diode, but the one on the right is more common than the one on the left.*

Completing the Breadboard Alarm Circuit

It's time now to breadboard the control circuit for your alarm noisemaker. Figure 3-95 shows how this can be done. I am assuming that you still have the noisemaker, which functions as before. I'm assuming that you still have its relevant components mounted on the top half of the breadboard. To save space, I'm just going to show the additional components mounted on the bottom half of the same breadboard.

It's important to remember that you are not supplying power directly to the left and right "rails" on the breadboard anymore; you are supplying power to the relay-transistor section, and when the relay closes its contacts, the *relay* supplies power to the rails. These then feed the power up to the top half of the breadboard. So disconnect your power supply from the breadboard rails and reconnect it as shown in Figure 3-95.

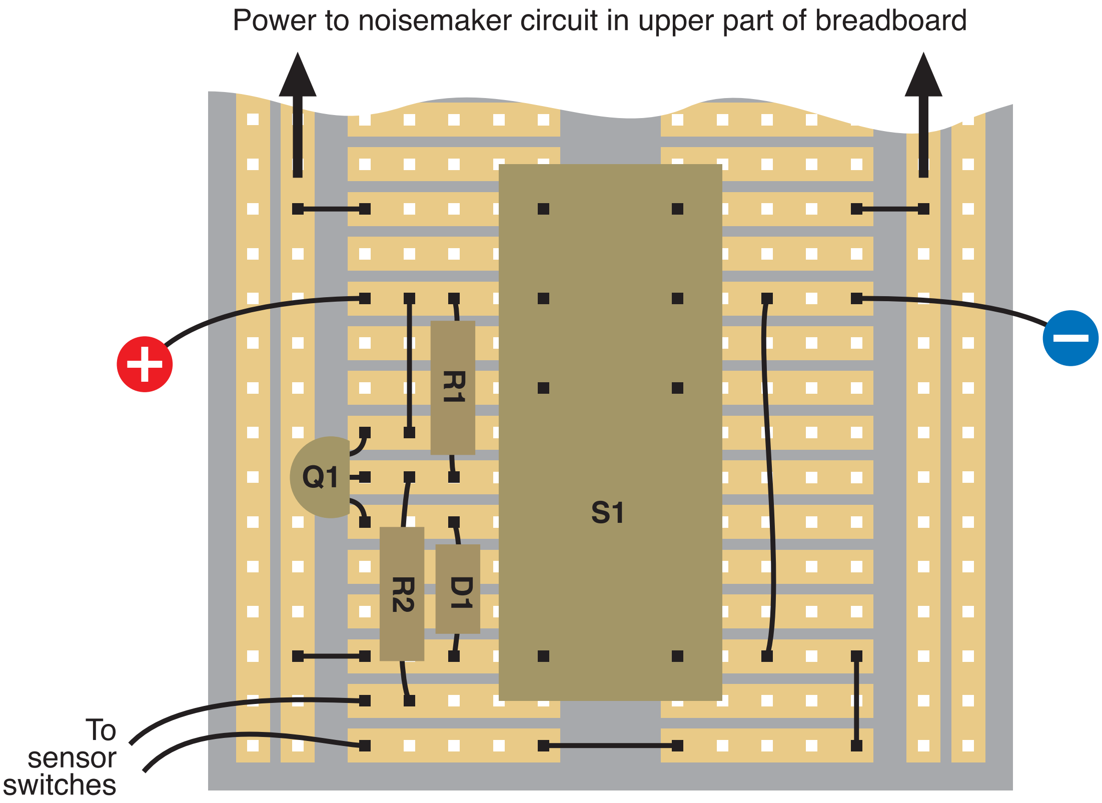

Figure 3-95. *The schematic that was developed in the previous pages can be emulated with components on a breadboard, as shown here. S1 is a DPDT relay. Wires to the sensor switch network and to the power supply must be added where shown.*

Because it's a double-pole relay, I am using it to switch negative as well as positive. This means that when the relay contacts are open, the noisemaking section of the circuit is completely isolated from the rest of the world.

The breadboarded relay circuit is exactly the same as the schematic in Figure 3-93. The components have just been rearranged and squeezed together so that they will fit alongside the relay. Two wires at the lower-left corner go to the network of magnetic sensor switches that will trip the alarm; for testing purposes, you can just hold the stripped ends of these two wires together to simulate all the switches being closed, and separate the wires to simulate a switch opening.

Two more wires bring power to the breadboard on either side of the relay. This is where you should connect your power supply during testing. The output from the relay, through its top pair of contacts, is connected with the rails of the breadboard by a little jumper wire at top left, and another at top right. Don't forget to include them! One more little wire at the lower-left corner (easily overlooked) connects the lefthand side rail to the lefthand coil terminal of the relay, so that when the relay is powering the noisemaker circuit, it powers itself as well.

When you mount the diode, remember that the end of it that is marked with a band around it is the end that blocks positive current. In this circuit, that's the lower end of the diode.

Try it to make sure that it works. Short the sensor wires together and then apply power. The alarm should remain silent. You can use your meter to check that no voltage exists between the side rails. Now separate the sensor wires, and the relay should click, supplying power to the side rails, which activates the noisemaker. Even if you bring the sensor wires back together, the relay should remain locked on. The only way to unlock it is to disconnect the power supply.

When the circuit is active, the transistor followed by the diode drops the voltage slightly, but the 12-volt relay should still work.

In my test circuit, trying three different relays, they drew between 27 and 40 milliamps at 9.6 volts. Some current still leaked through the transistor when it was in its "off" mode, but only a couple of milliamps at 0.5 volts. This low voltage was far below the threshold required to trip the relay.

Ready for Perfboarding

If the circuit works, the next step is to immortalize it on perforated board. Use the type of board that has a breadboard contact pattern etched on it in copper, as shown in Figure 3-72 on page 116. Check the following section, "Essentials: Perfboard procedure," for guidance on the best way to make this particular kind of solder joint—and the subsequent section for the most common problems.

ESSENTIALS

Perfboard soldering procedure

Carefully note the position of a component on your breadboard, and then move it to the same relative position on the perfboard, poking its wires through the little holes.

Turn the perfboard upside down, make sure that it's stable, and examine the hole where the wire is poking through, as shown in Figure 3-96. A copper trace surrounds this hole and links it with others. Your task is to melt solder so that it sticks to the copper and also to the wire, forming a solid, reliable connection between the two of them.

Take your pencil-style soldering iron in one hand and some solder in your other hand. Hold the tip of the iron against the wire and the copper, and feed some thin solder to their intersection. After two to four seconds, the solder should start flowing.

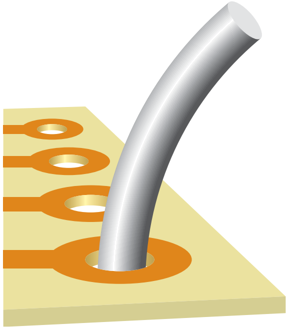

Figure 3-96

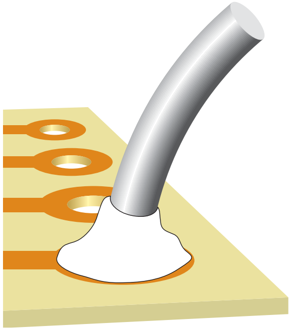

Figure 3-97

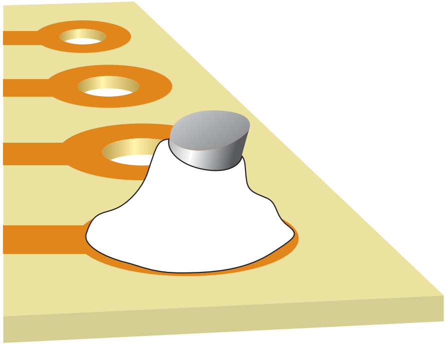

Figure 3-98. *To establish a connection between a section of wire and a copper trace on perforated board, the wire is pushed through the hole, and solder (shown in pure white for illustrative purposes) completes the connection. The wire can then be snipped short.*

Allow enough solder to form a rounded bump sealing the wire and the copper, as shown in Figure 3-97. Wait for the solder to harden thoroughly, and then grab the wire with pointed-nosed pliers and wiggle it to make sure you have a strong connection. If all is well, snip the protruding wire with your cutters. See Figure 3-98.

Because solder joints are difficult to photograph, I'm using drawings to show the wire before and after making a reasonably good joint, which is shown in pure white, outlined with a black line.

Actual soldered perfboard is shown in the photographs in Figures 3-99 and 3-100.

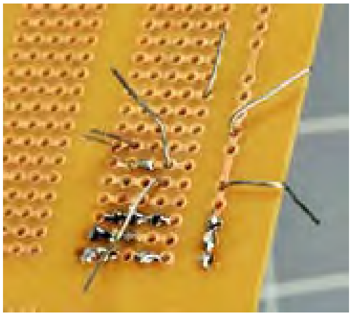

Figure 3-99. *This photograph was taken during the process of transferring components from breadboard to perforated board. Two or three components at a time are inserted from the other side of the board, and their leads are bent over to prevent them from falling out.*

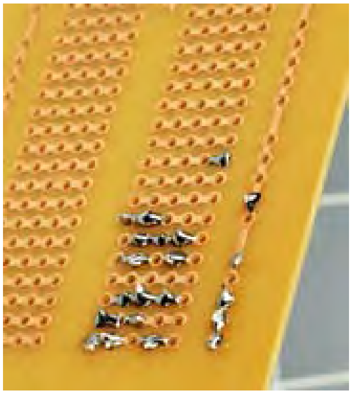

Figure 3-100. *After soldering, the leads are snipped short and the joints are inspected under a magnifying glass. Another two or three components can now be inserted, and the process can be repeated.*

TOOLS

Four most common perfboarding errors

1. Too much solder

Before you know it, solder creeps across the board, touches the next copper trace, and sticks to it, as depicted in Figure 3-101. When this happens, you have to wait for it to cool, and then cut it with a utility knife. You can also try to remove it with a rubber bulb and solder wick, but some of it will tend to remain.

Even a microscopic trace of solder is enough to create a short circuit. Check the wiring under a magnifying glass while turning the perfboard so that the light strikes it from different angles (or use your solder wick to suck it away).

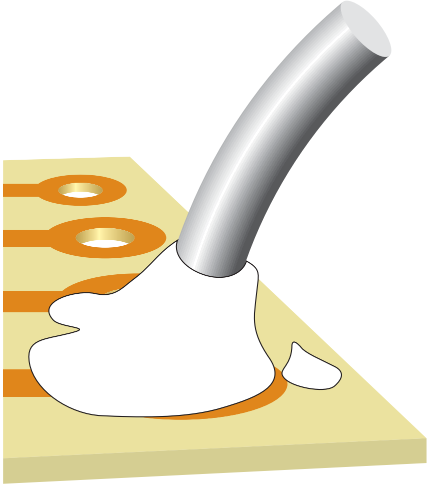

Figure 3-101. *If too much solder is used, it makes a mess and can create an unwanted connection with another conductor.*

2. Not enough solder

If the joint is thin, the wire can break free from the solder as it cools. Even a microscopic crack is sufficient to stop the circuit from working. In extreme cases, the solder sticks to the wire, and sticks to the copper trace around the wire, yet doesn't make a solid bridge connecting the two, leaving the wire encircled by solder yet untouched by it, as shown in Figure 3-102. You may find this undetectable unless you observe it with magnification.

You can add more solder to any joint that may have insufficient solder, but be sure to reheat the joint thoroughly.

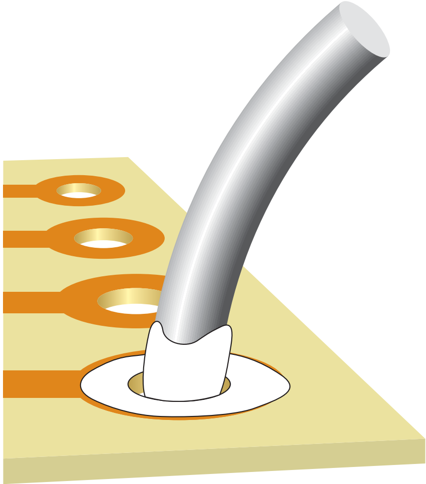

Figure 3-102. *Too little solder (or insufficient heat) can allow a soldered wire to remain separate from the soldered copper on the perforated board. Even a hair-thin gap is sufficient to prevent an electrical connection.*

3. Components incorrectly placed

It's very easy to put a component one hole away from the position where it should be. It's also easy to forget to make a connection.

I suggest that you print a copy of the schematic, and each time you make a connection on the perforated board, you eliminate that wire on your hardcopy, using a highlighter.

4. Debris

When you're trimming wires, the little fragments that you cut don't disappear. They start to clutter your work area, and one of them can easily get trapped under your perforated board, creating an electrical connection where you don't want it.

This is another reason for working with something soft, such as polyurethane foam, under your project. It tends to absorb or hold little pieces of debris, reducing the risk of you picking them up in your wiring.

Clean the underside of your board with an old (dry) toothbrush before you apply power to it, and keep your work area as neat as possible. The more meticulous you are, the fewer problems you'll have later.

Once again, be sure to check every joint with a magnifying glass.

Transferring components from the breadboard to the perforated board should be fairly simple, as long as you don't try to move too many at once. Follow the suggestions described in previous section "Essentials: Perfboard procedure," and pause frequently to check your connections. Impatience is almost always the cause of errors in this kind of work.

Figure 3-103 shows the noisemaker section of the circuit on perfboard, with the components positioned to minimize wasted space. Figure 3-104 shows the perfboard with the relay and its associated components added. The two black wires will go to the loudspeaker, the black-and-red pair of wires will bring power to the board, and the green wires will go to the magnetic sensor switches. Each wire penetrates the board, and its stripped end is soldered to the copper beneath.

Test it now, in the same way that you tested the same circuit on the breadboard. If it doesn't work, check the following section, "Essentials: Real-world fault tracing." If it does work, you're ready to trim the board and mount it in a project box.

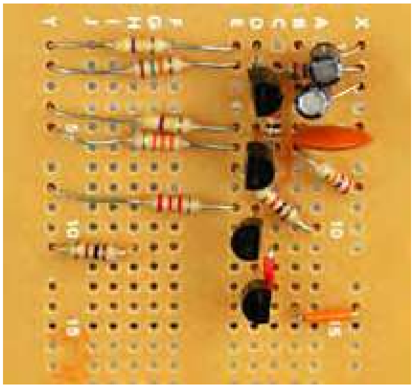

Figure 3-103. *The noisemaker circuit has been transplanted from breadboard to perforated board, with no additions or changes.*

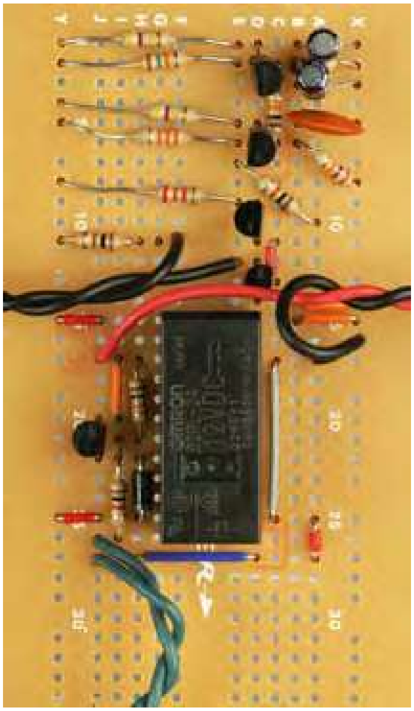

Figure 3-104. *The relay-transistor control circuit has been added. Wires to external devices have been stripped and poked into the perforated board, where they are soldered in place. The green wires connect with the sensor network, the black wires go to the loudspeaker, and the red-and-black wires supply power.*

ESSENTIALS

Real-world fault tracing

Here's a real-life description of the procedure for tracing a fault.

After I assembled the perfboard version of the combined noisemaker and relay circuit, I checked my work, applied power—and although the relay clicked, no sound came out of the loudspeaker. Of course, everything had worked fine on the breadboard.

First I looked at component placement, because this is the easiest thing to verify. I found no errors. Then I flexed the board gently while applying power—and the loudspeaker made a brief "beep." Any time this happens, you can be virtually certain that a solder joint has a tiny crack in it.

The next step was to anchor the black lead of my meter to the negative side of the power supply, and then switch on the power and go through the circuit point by point, from top to bottom, checking the voltage at each point with the red lead of the meter. In a simple circuit like this, every part should show at least some voltage.

But when I got to the second 2N2222 transistor, which powers the loudspeaker, its output was completely dead. Either I had melted the transistor while soldering it (unlikely), or there was a bad joint. I checked the perfboard beneath the transistor with the magnifying glass, and found that solder had flowed around one of the leads of the transistor without actually sticking to it. The gap must have been less than one-thousandth of an inch, but still, that was enough. Probably, the problem had been caused by dirt or grease.

This is the kind of patient inquiry you need to follow when a circuit doesn't work. Check whether your components are placed correctly, check your power supply, check the power on the board, check the voltage at each stage, and if you are persistent, you'll find the fault.

Switches and Inputs for the Alarm

Now you need to make the system easy to use. The block diagram in Figure 3-105 shows one additional box near the top of the sequence: User Controls. These will consist of switches, LEDs, and connections to the outside world. To plan this part of the job, first I have to summarize the way in which our alarm system works at this point in its evolution.

A full-featured home alarm system normally has two modes: in-home and away-from-home:

- Using the in-home mode, you switch on the alarm while you are at home so that it will alert you if an intruder opens a door or window.
- Using the away-from-home mode, typically you enter a code number, after which you have 30 seconds to leave and close the door behind you. When you return, you trigger the alarm by opening the door, but now you have 30 seconds to go to the control panel and enter your code number again to stop the alarm from sounding.

So far, the alarm system that you've been building has only an in-home mode. Still, many people find this function useful and reassuring. Later in the book I'll suggest a way in which you can modify it to incorporate an away-from-home mode, but for now, making it practical for in-home use is enough of a challenge.

Consider how it should be used on an everyday basis. It should have an on/off switch, naturally. When it's on, any of the magnetic sensor switches should trigger the alarm. But what if you switch it on without realizing that you've left a window open? At that time it won't be appropriate for the alarm to sound. What you really need is a circuit-test feature, to tell you if all the doors and windows are closed. Then you can switch on the alarm.

I think a pushbutton would be useful to test the alarm circuit. When you press it, a green LED should light up to indicate that the circuit is good. After you see the green light, you let go of the pushbutton and turn on the power switch, which illuminates a red LED, to remind you that the alarm is now armed and ready.

One additional feature would be useful: an alarm noisemaker test feature, so that you can be sure that the system is capable of sounding its alert when required to do so.

The circuit shown in Figure 3-106 incorporates all of these features. S1 is a SPDT switch; S2 is a DPDT momentary pushbutton of ON-(ON) type. The schematic shows it in its "relaxed" mode, when the button is not being pressed.

D1 is a red LED, D2 is a green LED, J1 is a power input jack (to be connected with an external 12-volt supply), and R1 is a 680Ω resistor to protect the LEDs. Note that J1 follows the usual practice of supplying positive voltage in its center contact, and negative in the circular shell around the center.

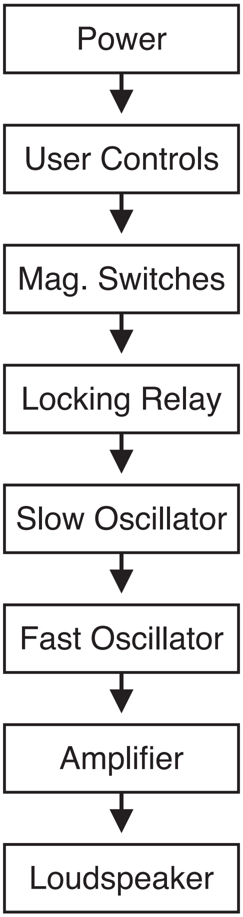

Figure 3-105. *The final block diagram for this phase of the project shows where user controls fit in the series of functions.*

When S1 is in its Off position, it still supplies positive power through its upper contact to S2, the pushbutton. When the pushbutton is pressed, so that it goes into its "Test" position, the pole of S2 connects with the power and sends it out through the sensor switches on doors and windows. The wires to these switches will be attached via a couple of binding posts, shown here as two circles. If the sensor switches are all closed, power returns through the second binding post, passes through the lower set of contacts in S2, and lights D2, the green LED. Because S1 is not supplying power to the alarm circuit board, the alarm does not sound at this time.

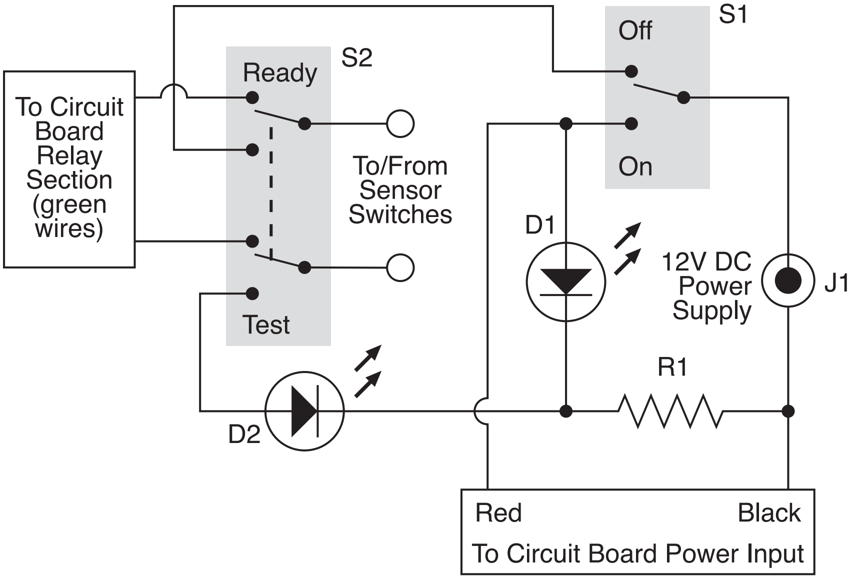

Figure 3-108. *This schematic suggests a convenient way to add an on/off switch, a continuity testing feature, and a noisemaker testing feature to the alarm.*

Now if S1 is turned to its On position, it sends power to the components on the perforated board. The transistor circuit sends power out along the green wires to S2, and as long as the button is not being pressed, the power goes out through the switch network and back through S2, to the transistor, keeping its base relatively negative. So the alarm remains silent. But as soon as a sensor switch is opened, the circuit is broken and the alarm sounds. The only way to stop it will be by switching S1 into its Off position.

Finally, if you press S2, the pushbutton, while S1 is in its On mode, you interrupt the network of switches and activate the noisemaker. In this way, S2 does dual duty: when S1 is off, pushbutton S2 tests the sensor switches for continuity. When S1 is on, S2 tests the noisemaker to make sure it creates a noise. I think this is the simplest possible way to implement these features.

Installing the Switches

If you bought a project box from RadioShack, it may have come with two optional top panels: one made of metal, the other made of plastic. I'll assume that you'll use the plastic one, as you'll have more trouble drilling holes in metal. The type of plastic used by RadioShack is ABS, which is very easy to shape with the tools that I have recommended.

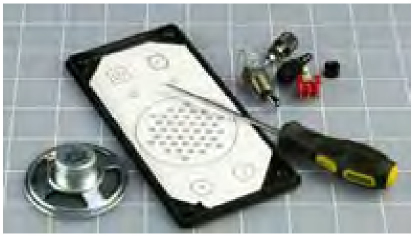

Figure 3-107. *A printed layout for the switches, LEDs, and other components has been taped to the underside of the lid of the project box. An awl is pressed through the paper to mark the center of each hole to be drilled in the lid.*

You have to choose a layout for the switches and other components that will share the top panel of the project box. I like a layout to be neat, so I take the trouble to draw it using illustration software, but a full-size pencil sketch is almost as good. Just make sure there's room for the components to fit together, and try to place them similarly to the schematic, to minimize the risk of confusion.

Tape your sketch to the inside of the top panel, as shown in Figure 3-107, and then use a sharp pointed tool, such as a pick, to press through and mark the plastic at the center of each hole. The indentations will help to center your bit when you drill the holes. Remember that you'll need to make multiple holes to vent the sound from the loudspeaker, which will be beneath the top panel of the box. The result is shown in Figure 3-108.

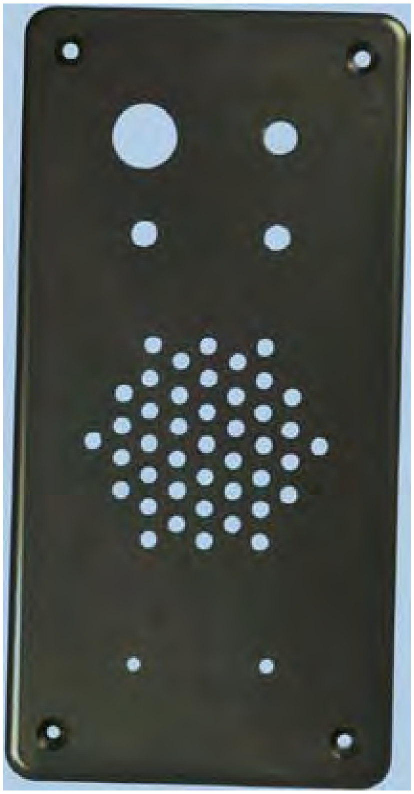

Figure 3-108. *The exterior of the panel after drilling. A small handheld cordless drill can create a neat result if the holes were marked carefully.*

I placed all the components on the top panel, with the exception of the power input jack, which I positioned at one end of the box. Naturally, each hole has to be sized to fit its component, and if you have calipers, they'll be very useful for taking measurements and selecting the right drill bit. Otherwise, make your best guess, too small being better than too large. A deburring tool is ideal for slightly enlarging a hole so that a component fits snugly. This may be necessary if you drill 3/16-inch holes for your 5-inch mm LEDs. Fractionally enlarge each hole, and the LEDs should push in very snugly.

If your loudspeaker lacks mounting holes, you'll have to glue it in place. I used five-minute epoxy to do this. Be careful not to use too much. You don't want any of the glue to touch the speaker cone.

Drilling large holes in the thin, soft plastic of a project box can be a problem. The drill bit tends to dig in and create a mess. You can approach this problem in one of three ways:

1. Use a Forstner drill bit if you have one. It creates a very clean hole.
2. Drill a series of holes of increasing size.
3. Drill a smaller hole than you need, and enlarge it with a deburring tool.

Regardless of which approach you use, you'll need to clamp or hold the top panel of the project box with its outside surface face-down on a piece of scrap wood. Then drill from the inside, so that your bit will pass through the plastic and into the wood.

Finally, mount the components in the panel, as shown in Figure 3-109, and turn your attention to the underneath part of the box.

The circuit board will sit on the bottom, held in place with four #4-size machine screws (bolts) with washers and nylon-insert locknuts. You need to use locknuts to eliminate the risk of a nut working loose and falling among components where it can cause a short circuit.

You'll have to cut the perfboard to fit, taking care not to damage any of the components on it. Also check the underside the board for loose fragments of copper traces after you finish cutting.

Drill bolt holes in the board, if necessary, taking care again not to damage any components. Then mark through the holes to the plastic bottom of the box, and drill the box. Countersink the holes (i.e., bevel the edges of a hole so that a flat-headed screw will fit into it flush with the surrounding surface), push the little bolts up from underneath, and install the circuit board. Be extremely careful not to attach the circuit board too tightly to the project box. This can impose bending stresses, which may break a joint or a copper trace on the board.

I like to include a soft piece of plastic under the board to absorb any stresses. Because you're using locknuts, which will not loosen, there's no need to make them especially tight.

Test the circuit again after mounting the circuit board, just in case.

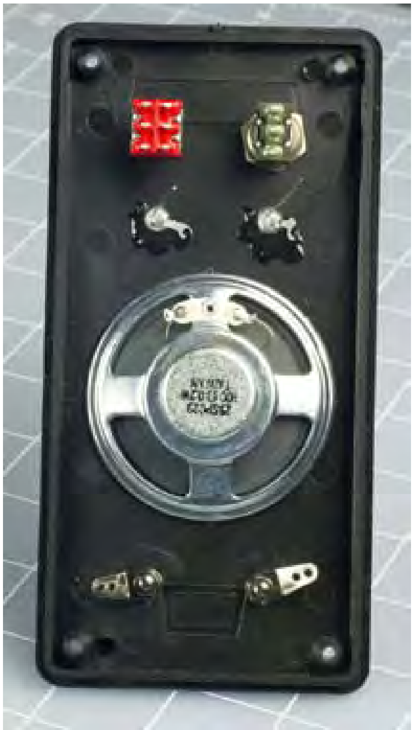

Figure 3-109. *Components have been added to the control panel of the project box (seen from the underside). The loudspeaker has been glued in place. Spare glue was dabbed onto the LEDs, just in case. The SPDT on/off switch is at the top right, the DPDT pushbutton is at top left, and the binding posts, which will connect with the network of magnetic sensor switches, are at the bottom.*

Soldering the Switches

Figure 3-110 shows how the physical switches should be wired together. Remember that S1 is a toggle switch and S2 is a DPDT pushbutton. Your first step is to decide which way up they should be. Use your meter to find out which terminals are connected when the switch is flipped, and when the button is pressed. You'll probably want the switch to be on when the toggle is flipped upward. Be especially careful with the orientation of the pushbutton, because if you wire it upside-down, it will constantly have the alarm in "test" mode, which is not what you want.

Remember, the center terminal of any double-throw switch is almost always the pole of the switch, connecting with the terminals immediately above it and below it.

Stranded wire is appropriate to connect the circuit board with the components in the top panel, because the strands flex easily and impose less stress on solder joints. Twisting each pair of wires together helps to minimize the mess.

Remember to install the LEDs with their short, negative wires connected with the resistor. This will entail some wire-to-wire soldering. You may want to protect some of these bare leads and joints with thin heat-shrink tubing, to minimize the risk of short circuits when you push all the parts into the box.

When you connect wires or components with the lugs on the switches, your pencil-style soldering iron probably won't deliver enough heat to make good joints. You can use your higher-powered soldering iron in these locations, but you absolutely must apply a good heat sink to protect the LEDs when you attach them, and don't allow the iron to remain in contact with anything for more than 10 seconds. It will quickly melt insulation, and may even damage the internal parts of the switches.

In projects that are more complex than this one, it would be good practice to link the top panel with the circuit board more neatly. Multicolored ribbon cable is ideal for this purpose, with plug-and-socket connectors that attach to the board. For this introductory project, I didn't bother. The wires just straggle around, as shown in Figure 3-111.

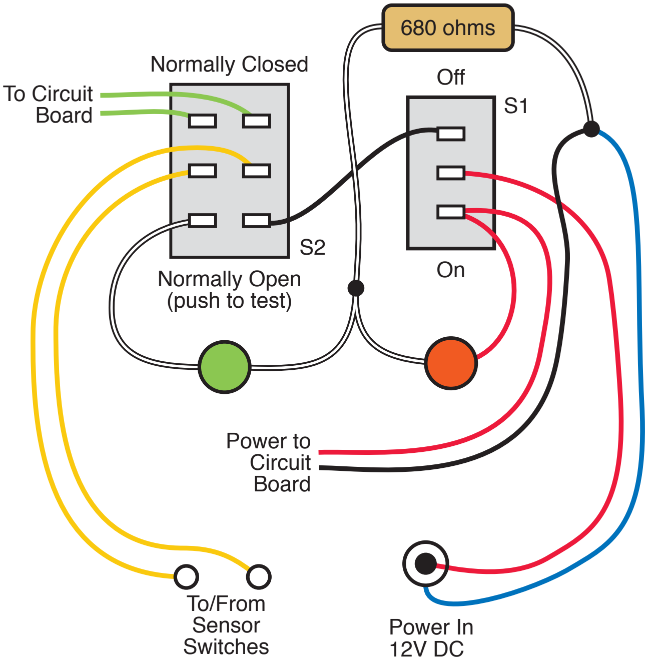

Figure 3-110. *The components can be wired together like this to replicate the circuit shown in The red and green circles are LEDs. Small, solid black circles indicate wire-to-wire solder joints.*

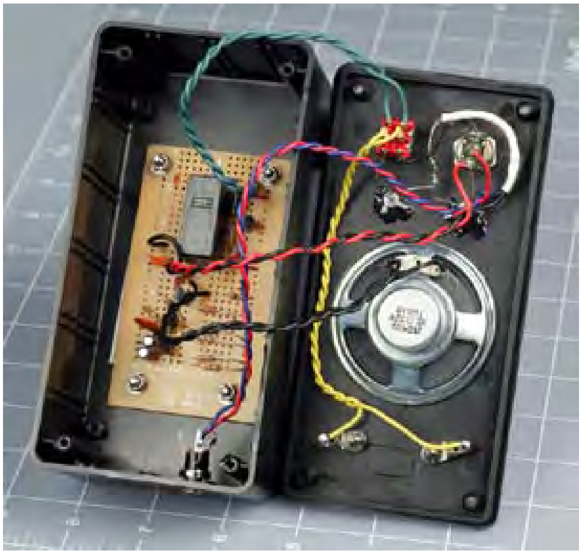

Figure 3-111. *The circuit board has been installed in the base of the project box, and the power input jack has been screwed into the end of the box. Twisted wire-pairs have been connected on a point-to-point basis, without much concern over neatness, as this is a relatively small project. The white insulation at the top-right corner of the front panel is heat-shrink tube that encloses a solder joint and the 680Ω load resistor. Soldering wires to the pushbutton switch requires care and precision, as the contacts are closely spaced.*

Final Test

When you've completed the circuit, test it! If you don't have your network of magnetic sensor switches set up yet, you can just use a piece of wire to connect the two binding posts. Make sure that S1 is in its Off position, then solder the appropriate plug to your 12-volt power source, and plug it into the power jack. When you press the button, the green LED should light up to show continuity between the two binding posts. Now disconnect the wire between the binding posts, press the button again, and the green LED should remain dark.

Reconnect the binding posts, flip S1 to its On position, and the red LED should light up. Press the button, and the alarm should start. Reset it by turning S1 off and then on again; then disconnect the wire between the binding posts. Again, the alarm should start, and it should continue even if you reconnect the wire.

If everything works the way it should, it's time to screw the top of the box in place, pushing the wires inside. Because you're using a large box, you should have no risk of metal parts touching each other accidentally, but still, proceed carefully.

Alarm Installation

Before you install your magnetic sensor switches, you should test each one by moving the magnetic module near the switch module and then away from it, while you use your meter to test continuity between the switch terminals. The switch should close when it's next to the magnet, and open when the magnet is removed.

Now draw a sketch of how you'll wire your switches together. Always remember that they have to be in series, not in parallel! Figure 3-112 shows the concept in theory. The two terminals are the binding posts on top of your control box (which is shown in green), and the dark red rectangles are the magnetic sensor switches on windows and doors. Because the wire for this kind of installation usually has two conductors, you can lay it as I've indicated but cut and solder it to create branches. The solder joints are shown as orange dots. Note how current flows through all the switches in series before it gets back to the control box.

Figure 3-113 shows the same network as you might actually install it in a situation where you have two windows and a door. The blue rectangles are the magnetic modules that activate the switch modules.

You'll need a large quantity of wire, obviously. The type of white, stranded wire that is sold for doorbells or furnace thermostats is good. Typically, it is 20-gauge or larger.

After you install all the switches, clip your meter leads to the wires that would normally attach to the alarm box. Set your meter to test continuity, and open each window or door, one at a time, to check whether you're breaking the continuity. If everything is OK, attach the alarm wires to the binding posts on your project box.

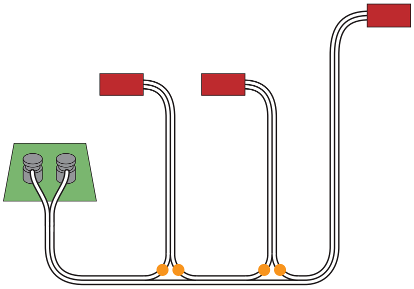

Figure 3-112. *Dual-conductor, white insulated wire can be used to connect the terminals on the alarm control box with magnetic sensors (shown in dark red). Because the sensors must be in series, the wire is cut and joined at positions marked with orange dots.*

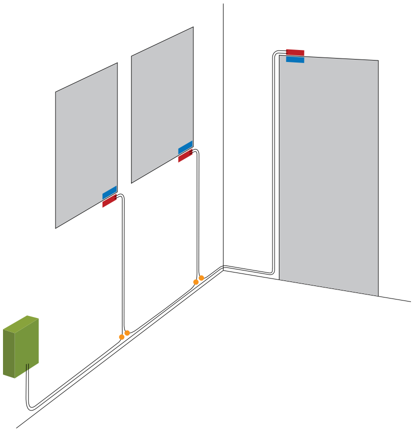

Figure 3-113. *In an installation involving two windows and a door, the magnetic components of the sensors (blue rectangles) could be placed as shown, while the switches (dark red) are located alongside them.*

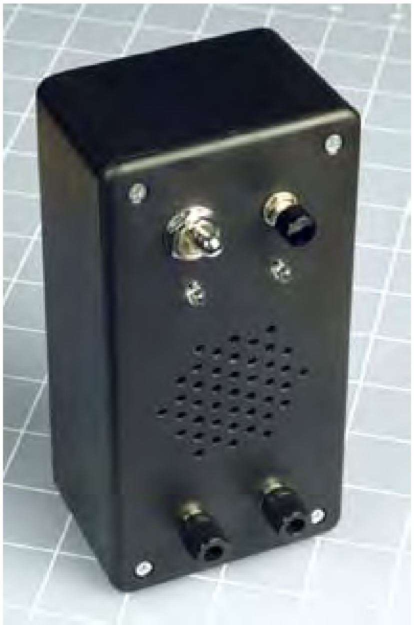

Figure 3-114. *The intrusion alarm completed and in its project box.*

Now deal with the power supply. Use your AC adapter, set to 12 volts, hooked up to your type N power plug, or attach the power plug to a 12-volt alarm battery.

If you use a battery, be especially careful that the wire leading to the center terminal of your power plug is positive! A 12-volt battery can deliver substantial current, which can fry your components if you connect it the wrong way around. It would be a shame to destroy your entire project at the very last step.

The only remaining task is to label the switch, button, power socket, and binding posts on the alarm box. You know that the switch turns the power on and off, and the button tests the circuit and the noisemaker, but no one else knows, and you might want to allow a guest to use your alarm while you're away. For that matter, months or years from now, you may forget some details. Will you remember that the power source for this unit should be 12 volts?

Labeling really is a good idea. But as you can see in Figure 3-114, I haven't quite gotten around to it for the box that I built.

Conclusion

The alarm project has taken you through the basic steps that you will usually follow any time you develop something:

1. Draw a schematic and make sure that you understand it.
2. Modify it to fit the pattern of conductors on a breadboard.
3. Install components on the breadboard and test the basic functions.
4. Modify or enhance the circuit, and retest.
5. Transfer to perforated board, test, and trace faults if necessary.
6. Add switches, buttons, power jack, and plugs or sockets to connect the circuit with the outside world.
7. Mount everything in a box (and add labeling).

While going through this sequence, I hope you've learned the basics of electricity, along with some simple electrical theory, and fundamentals about electronic components. This knowledge should enable you to move on to the much more powerful realm of integrated circuit—which I'll cover in Chapter 4.

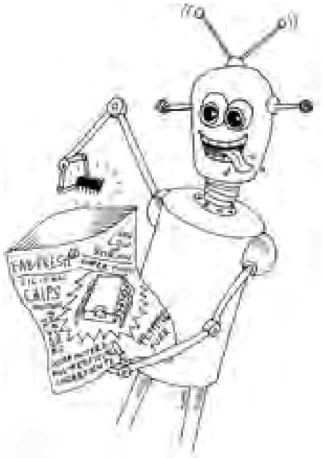

4

Chips, Ahoy!

IN THIS CHAPTER

Shopping List: Experiments 16 Through 24
Experiment 16: Emitting a Pulse
Experiment 17: Set Your Tone
Experiment 18: Reaction Timer
Experiment 19: Learning Logic
Experiment 20: A Powerful Combination
Experiment 21: Race to Place
Experiment 22: Flipping and Bouncing
Experiment 23: Nice Dice
Experiment 24: Intrusion Alarm Completed

Before I get into the fascinating topic of integrated circuit (IC) chips, I have to make a confession: some of the things I asked you to do in Chapter 3 could have been done a bit more simply. Does this mean you have been wasting your time? No, I firmly believe that by building circuits with old-fashioned components—capacitors, resistors, and transistors—you acquire the best possible understanding of the principles of electronics. Still, you are going to find that integrated circuit chips, containing dozens, hundreds, or even thousands of transistor junctions, will enable some shortcuts.

Shopping List: Experiments 16 Through 24

Tools

The only new tool that I recommend using in conjunction with chips is a logic probe. This tells you whether a single pin on a chip has a high or low voltage, which can be helpful in figuring out what your circuit is doing. The probe has a memory function so that it will light its LED, and keep it lit, in response to a pulse that may have been too quick for the eye to see.

Search online and buy the cheapest logic probe you can find. I don't have any specific brand recommendations. The one shown in Figure 4-1 is fairly typical.

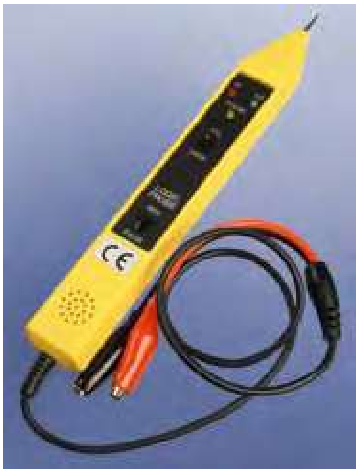

Figure 4-1. *A logic probe detects the high or low voltage on each pin of a chip, and reveals pulses that may occur too quickly for you to perceive them with the unaided eye.*

Supplies

Integrated circuit chips

If you buy everything on this shopping list, and you bought basic parts such as resistors and capacitors that were listed previously, you should have everything you need for all the projects in this chapter.

As chips are quite cheap (currently around 50 cents apiece), I suggest you buy extras. This way, if you damage one, you'll have some in reserve. You'll also have a stock for future projects.

Please read the next section, "Fundamentals: Choosing chips," before you begin chip-shopping. Chips should be easily obtainable from all the major electronics retail suppliers, and sometimes are found on eBay shops. Look in the appendix for a complete list of URLs.

FUNDAMENTALS

Choosing chips

Figure 4-2 shows what is often referred to as an integrated circuit (IC). The circuit is actually etched onto a tiny wafer or "chip" of silicon, embedded in a black plastic body, which is properly referred to as the "package." Tiny wires inside the package link the circuit with the two rows of pins on either side. Throughout this book, I will use the word "chip" to refer to the whole object, including its pins, as this is the most common usage.

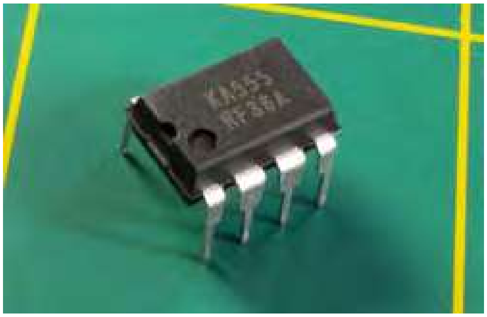

Figure 4-2. *An integrated circuit chip in Plastic Dual-Inline Pin package, abbreviated PDIP, or, more often, DIP.*

The pins are mounted at intervals of 1/10 inch in two rows spaced 3/10 inch apart. This format is known as a Plastic Dual Inline Package, abbreviated PDIP, or, more often, just DIP. The chip in the photograph has four pins in each row; others may have many more. The first thing you need to know, when shopping for chips, is that you'll only be using the DIP package. This book will not be featuring the more modern type, known as "surface-mount," because they're much smaller, more difficult to handle, and require special tools that are relatively expensive. Figure 4-3 shows a size comparison between a 14-pin DIP package and a 14-pin surface-mount package. Many surface-mount chips are even smaller than the one shown.

Just about every chip has a part number printed on it. In Figure 4-2, the part number is KA555. In Figure 4-3, the DIP chip's part number is M74HC00B1, and the surface-mount chip is a 74LVC07AD. You can ignore the second line of numbers and/or letters on each chip, as they are not part of the part number.

Notice in Figure 4-3 that even though the chips look quite different from each other, they both have "74" in their part numbers. This is because both of them are members of the "7400" family of logic chips, which originally had part numbers from 7400 and upward (7400, 7401, 7402, 7403, and so on). Often they are now referred to as "74xx" chips, where "xx" includes all the members of the family. I'll be using this family a lot, so you need to know how to buy them. I'll give you some advice on that without going into details yet about what the chips actually do.

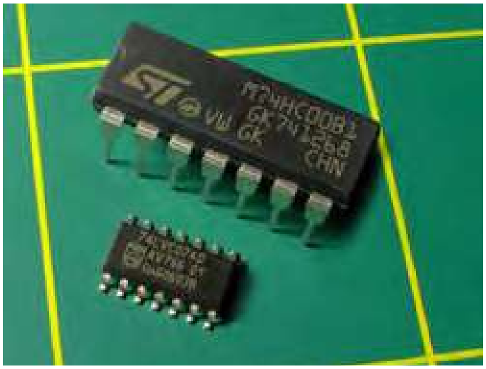

Figure 4-3. *The DIP chip, at the rear, has pins spaced 1/10 inch apart, suitable for insertion in a breadboard or perforated board. It can be soldered without special tools. The small-outline integrated circuit (SOIC) surface-mount chip (foreground) has solder tabs spaced at 1/20 inch. Other surface-mount chips have pins spaced at 1/40 inch or even less (these dimensions are often expressed in millimeters). Surface-mount chips are designed primarily for automated assembly and are difficult to work with manually. In this photo, the yellow lines are 1 inch apart to give you an idea of the scale.*

Take a look at Figure 4-4, which shows how to interpret a typical part number in a 74xx family member. The initial letters identify the manufacturer (which you can ignore, as it really makes no difference for our purposes). Skip the letters until you get to the "74." After that, you find two more letters, which are important. The 74xx family has evolved through many generations, and the letter(s) inserted after the "74" tell you which generation you're dealing with. Some generations have included:

- 74L
- 74LS
- 74C
- 74HC
- 74AHC

And there are more. Generally speaking, subsequent generations tend to be faster or more versatile than previous generations. In this book, for reasons I'll explain later, we are mostly using the HC generation.

FUNDAMENTALS

Choosing chips (continued)

After the letters identifying the generation, you'll find two (sometimes more) numerals. These identify the specific function of the chip. You can ignore any remaining letters and numerals. Looking back at Figure 4-3, the DIP chip part number, M74HC00B1 tells you that it is a chip in the 74xx family, HC generation, with its function identified by numerals 00. The surface-mount chip number, 74LVC07AD, tells you that it is in the 74xx family, LVC generation, with function identified by numerals 07. For convenience we could refer to the first chip as a "74HC00" and the second chip as a "74HC07" because, regardless of their different manufacturers and package sizes, the fundamental behavior of the circuit inside remains the same.

The purpose of this long explanation is to enable you to interpret catalog listings when you go chip shopping. You can search for "74HC00" and the online vendors are usually smart enough to show you appropriate chips from multiple manufacturers, even though there are letters preceding and following the term that you're searched for.

Suppose a circuit requires a 74HC04 chip. If you search for "74HC04" on the website of a parts supplier, you may find versions such as the CD74HC04M96 by Texas Instruments, the 74HC04N by NXP Semiconductors, or MM74HC04N by Fairchild Semiconductor. Because they all have "74HC04" in the middle, any of them will work.

Just be careful that you buy the larger DIP-style package, not the surface-mount package. If the part number has an "N" on the end, you can be sure that it's a DIP package. If there is no "N" on the end, it may or may not be a DIP package, and you will have to check a photo or additional description to make sure. If the part number begins with SS, SO, or TSS, it's absolutely definitely surface-mount, and you don't want it. Many catalogs show photographs of the chips to assist you in buying the right package style.

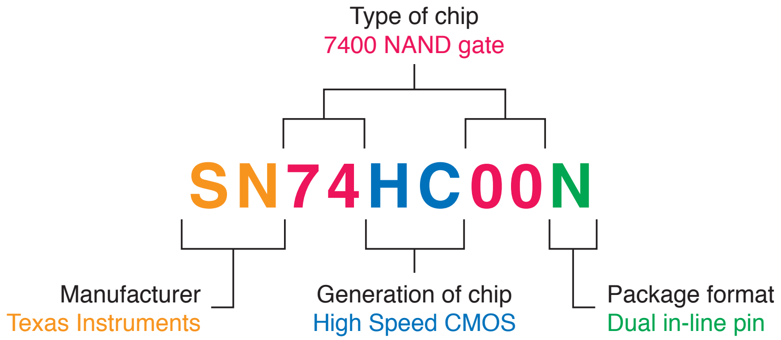

Figure 4-4. *Look for the chip family (74xx, in this case) with the correct generation (HC, in this case) embedded in the number. Make sure you buy the DIP version, not the surface-mount version. The manufacturer is unimportant.*

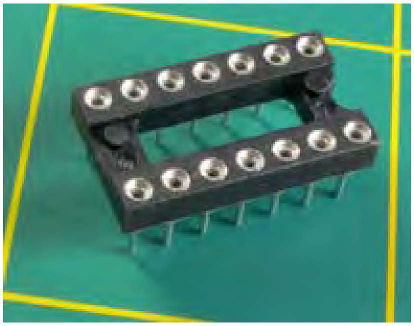

Figure 4-5. *When you're soldering a circuit onto perforated board, sockets eliminate the risk of overheating integrated circuit chips and reduce the risk of zapping them with static electricity, and enable easy replacement.*

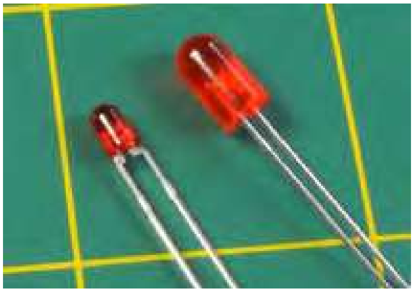

Figure 4-6. *An HC series logic chip is rated to deliver only 4mA at each pin. This is inadequate to drive a typical 5 mm LED (right), which is rated for 20mA forward current. Miniature, low-current LEDs (left) will use as little as 1mA in series with a suitable resistor, and are ideal for test circuits in which you want to see the output with a minimum of hassle.*

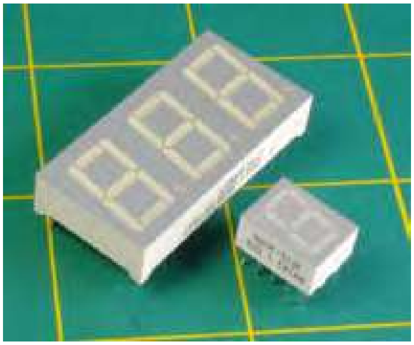

Figure 4-7. *Seven-segment displays are the simplest way to show a numeric output and can be driven directly by some CMOS chips. For finished projects, they are typically mounted behind transparent red acrylic plastic panels.*

Here's your chip list:

- 555 Timer. STMicroelectronics SA555N, Fairchild NE555D, RadioShack TLC555, or similar. Do not get the "CMOS" version of this chip, or any fancy versions such as those of high precision. Buy the cheapest you can find. Quantity: 10. The chip in Figure 4-2 is a 555 timer.
- Logic chips types 74HC00, 74HC02, 74HC04, 74HC08, 74HC32, and 74HC86. Actual part numbers could be M74HC00B1, M74HC02B1, M74HC04B1, and so on, by STMicroelectronics, or SN74HC00N, SN74HC02N, SN74HC04N, and so on, by Texas Instruments. Any other manufacturers are acceptable.
- Remember, each part number should have "HC" in the middle of it, and you want the DIP or PDIP package, not surface-mount. Quantity of each: At least 4.
- 4026 Decade Counter (a chip that counts in tens). Texas Instruments CD-4026BE or similar. Quantity: 4 (you'll need 3, but because this is a CMOS chip sensitive to static electricity, you should have at least one in reserve). Any chip with "4026" in its part number should be OK.
- 74LS92 counter chip, 74LS06 open-collector inverter chip, and 74LS27 triple-input NOR chip. Quantity: 2 of each. Note the "LS" in these part numbers! There will be one experiment in which I want you to use the LS generation instead of the HC generation.

IC sockets

I suggest that you avoid soldering chips directly onto perforated board. If you damage them, they're difficult to remove. Buy some DIP sockets, solder the sockets onto the board, and then plug the chips into the sockets. You can use the cheapest sockets you can find (you don't need gold-plated contacts for our purposes). You will need 8-pin, 14-pin, and 16-pin DIP sockets, such as parts 276-1995, 276-1999, and 276-1998 from RadioShack. See Figure 4-5. Quantity of each: 5 minimum.

Low-power LEDs

The logic chips that you'll be using are not designed to deliver significant power. You'll need to add transistors to amplify their output if you want to drive bright LEDs or relays. Because adding transistors is a hassle, I suggest an alternative: Special low-power LEDs that will draw as little as 1mA, such as the Everlight model T-100 Low Current Red, part number HLMPK150. Figure 4-6 shows a size comparison with a regular 5mm LED. Quantity: 10 (at a minimum).

LED numeric displays

In at least one of our projects, you'll want to illuminate some seven-segment LED numerals. You'll need either three individual numerals, or one package containing three numerals, such as the Kingbright High Efficiency Red Diffused, part number BC56-11EWA, which will be specifically referred to in schematics in this book. If you buy a different seven-segment display, it must be an LED with a "common cathode." (Don't buy liquid-crystal LCD numerals; they require different electronics to drive them.) If you have a choice of power consumption, buy whichever product consumes the least current. See Figure 4-7.

Latching relays

You're going to need a 5-volt latching relay that has two coils, instead of one. The first coil flips the relay one way; the second coil flips it back. The relay consumes no additional power while remaining passively in each state. I suggest the Panasonic DS2E-SL2-DC5V relay. If you buy a different relay, it must be dual-coil latching to run off 5 volts DC, switching at least 1 amp, in a "2 form C" package, to fit your breadboard.

Potentiometers

You'll need 5K, 10K, and 100K linear potentiometers (one of each). Also, a 10K trimmer potentiometer (which you may find described just as a "trimmer"). The manufacturer is unimportant.

Voltage regulators

Because many logic chips require precisely 5 volts DC, you need a voltage regulator to deliver this. The LM7805 does the job. Here again, the chip number will be preceded or followed with an abbreviation identifying the manufacturer and package style, as in the LM7805CT from Fairchild. Any manufacturer will do, but the package should look like the one in Figure 4-8, and if you have a choice, buy a regulator that can deliver at least 1 amp.

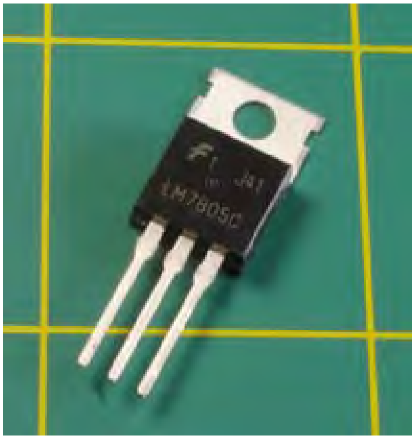

Figure 4-8. *Many integrated circuit chips require a controlled power supply of 5 volts, which can be delivered by this regulator when you apply 7.5 to 9 volts to it. The lefthand pin is for positive input, the center pin is a common ground, and the righthand pin is the 5V output. For currents exceeding 250mA, you should bolt the regulator to a metal heat sink using the hole at the top.*

Tactile switches

These are SPST pushbuttons (momentary switches), usually with four legs. Look for the ALPS part number SKHHAKA010 or any similar item that has pins to fit your breadboard or perforated board. See Figure 4-9.

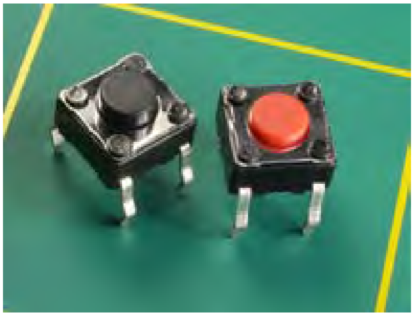

Figure 4-9. *A tactile switch delivers tactile feedback through your fingertip when you press it. They are almost always SPST pushbuttons designed for mounting in circuit boards with standard 1/10-inch hole spacing.*

12-key numeric keypad

Velleman "12 keys keyboard with common output" (no part number, but has been available through All Electronics under catalog code KP-12). Quantity: 1.

This type of keypad has the same layout as an old-fashioned touchtone phone. It should have at least 13 pins or contacts, 12 of which connect with individual pushbuttons, the thirteenth connecting with the other side of all the pushbuttons. In other words, the last pin is "common" to all of them, and this type of keypad is often described as having a "common output." The type of keypad that you don't want is "matrix-encoded," with fewer than 13 contacts, requiring additional external circuitry. See Figures 4-10 and 4-11. If you can't find the Velleman keypad that I suggest, look carefully at keypad descriptions and photographs to make sure that the one you buy is not matrix-encoded and has a common terminal.

Alternatively, you may substitute 12 cheap SPST NO pushbuttons and mount them in a small project box.

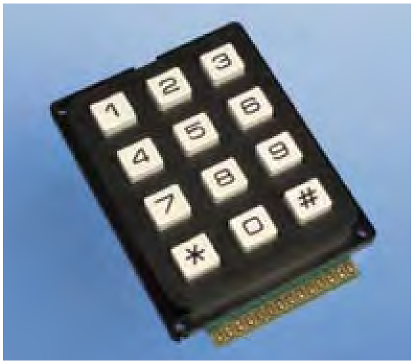

Figure 4-10. *When shopping for a numeric keypad, it should have 12 keys in "touch-tone phone" layout, and should have at least 13 contacts for input/output. The contacts are visible here along the front edge.*

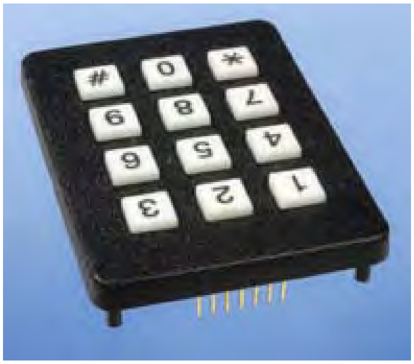

Figure 4-11. *This keypad has insufficient pins and will not work in the circuit in this book.*

BACKGROUND

How chips came to be

The concept of integrating solid-state components into one little package originated with British radar scientist Geoffrey W. A. Dummer, who talked about it for years before he attempted, unsuccessfully, to build one in 1956. The first true integrated circuit wasn't fabricated until 1958 by Jack Kilby, working at Texas Instruments. Kilby's version used germanium, as this element was already in use as a semiconductor. (You'll encounter a germanium diode when I deal with crystal radios in the next chapter of this book.) But Robert Noyce, pictured in Figure 4-12, had a better idea.

Born in 1927 in Iowa, in the 1950s Noyce moved to California, where he found a job working for William Shockley. This was shortly after Shockley had set up a business based around the transistor, which he had coinvented at Bell Labs.

Noyce was one of eight employees who became frustrated with Shockley's management and left to establish Fairchild Semiconductor. While he was the general manager of Fairchild, Noyce invented a silicon-based integrated circuit that avoided the manufacturing problems associated with germanium. He is generally credited as the man who made integrated circuits possible.

Early applications were for military use, as Minuteman missiles required small, light components in their guidance systems. These applications consumed almost all chips produced from 1960 through 1963, during which time the unit price fell from around $1,000 to $25 each, in 1963 dollars.

In the late 1960s, medium-scale integration chips emerged, each containing hundreds of transistors. Large-scale integration enabled tens of thousands of transistors on one chip by the mid-1970s, and today's chips can contain as many as several billion transistors.

Robert Noyce eventually cofounded Intel with Gordon Moore, but died unexpectedly of a heart attack in 1990. You can learn more about the fascinating early history of chip design and fabrication at *http://www.siliconvalleyhistorical.org*.

Figure 4-12. *This picture of Robert Noyce, late in his career, is from the Wikimedia Commons.*

Experiment 16: Emitting a Pulse

I'm going to introduce you to the most successful chip ever made: the 555 timer. As you can find numerous guides to it online, you might question the need to discuss it here, but I have three reasons for doing so:

1. It's unavoidable. You simply have to know about this chip. Some sources estimate that more than 1 billion are still being manufactured annually. It will be used in one way or another in most of the remaining circuits in this book.
2. It provides a perfect introduction to integrated circuits, because it's robust, versatile, and illustrates two functions that we'll be dealing with later: comparators and a flip-flop.
3. After reading all the guides to the 555 that I could find, beginning with the original Fairchild Semiconductor data sheet and making my way through various hobby texts, I concluded that its inner workings are seldom explained very clearly. I want to give you a graphic understanding of what's happening inside it, because if you don't have this, you won't be in a good position to use the chip creatively.

You will need:

- 9-volt power supply.
- Breadboard, jumper wires, and multimeter.
- 5K linear potentiometer. Quantity: 1.
- 555 timer chip. Quantity: 1.
- Assorted resistors and capacitors.
- SPST tactile switches. Quantity: 2.
- LED (any type). Quantity: 1.

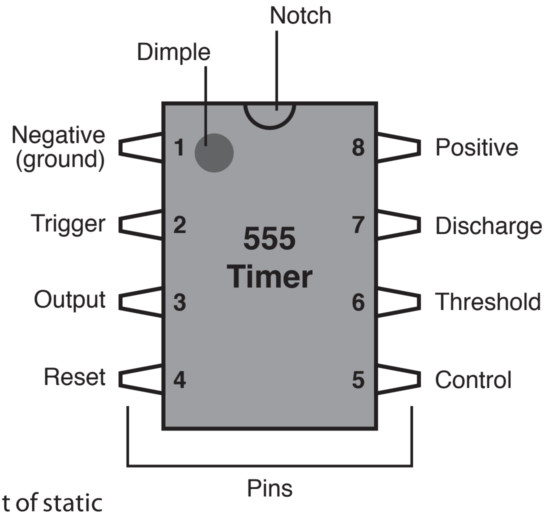

Figure 4-13. *The 555 timer chip, seen from above. Pins on chips are always numbered counterclockwise, from the top-left corner, with a notch in the body of the chip uppermost, or a circular indentation at top-left, to remind you which end is up.*

Procedure

The 555 chip is very robust, but still, in theory, you can zap it with a jolt of static electricity and kill it. Therefore, to be on the safe side, you should ground yourself before handling it. See the "Grounding yourself" warning on page 172 for details. Although this warning primarily refers to the type of chips known as CMOS, which are especially vulnerable, grounding yourself is always a sensible precaution.

Look for a small circular indentation, called the *dimple*, molded into the body of the chip, and turn the chip so that the indentation is at the top-left corner with the pins pointing down. Alternatively, if your chip is of the type with a notch at one end, turn the chip so that the notch is at the top.

The pins on chips are always numbered counterclockwise, starting from the top-left pin (next to the dimple). See Figure 4-13, which also shows the names of the pins on the 555 timer, although you don't need to know most of them just yet.

Insert the chip in your breadboard so that its pins straddle the channel down the center. Now you can easily feed voltages to the pins on either side, and read signals out of them. See Figure 4-14 for a precise guide to placement, in the first project. The timer is identified as "IC1," because "IC" is the customary abbreviation for "Integrated Circuit."

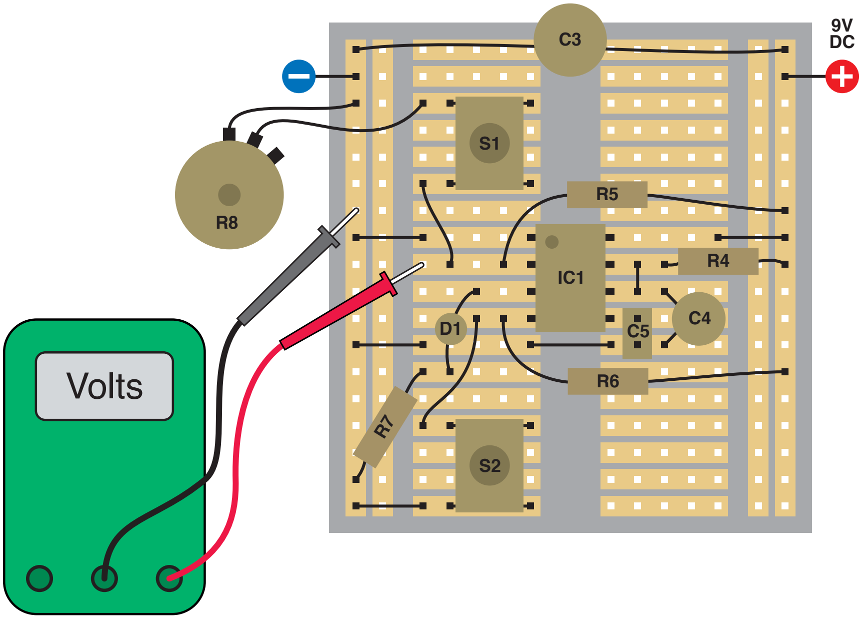

Figure 4-14. *This circuit allows you to explore the behavior of the 555 timer chip. Use your meter to monitor the voltage on pin 2 as shown. There are no resistors labeled R1, R2, or R3 and no capacitors labeled C1 or C2, because they'll be added in a later schematic. Component values in this schematic:*

R4: 100K
R5: 2K2
R6: 10K
R7: 1K
R8: 5K linear potentiometer
C3: 100 µF electrolytic
C4: 47 µF electrolytic
C5: 0.1 µF ceramic
IC1: 555 timer
S1, S2: SPST tactile switches (pushbuttons)
D1: Generic LED

R5 holds the trigger (pin 2) positive until S1 is pressed, which lowers the voltage depending on the setting of potentiometer R8. When the trigger voltage falls below 1/3 of the power supply, the chip's output (pin 3) goes high for a period determined by the values of R4 and C4. S2 resets (zeros) the timer, by reducing the voltage to pin 4, the Reset. C3 smoothes the power supply, and C5 isolates pin 5, the control, so that it won't interfere with the functioning of this test circuit. (We'll use the control pin in a future experiment.)

All integrated circuit chips require a power supply. The 555 is powered with negative voltage applied to pin 1 and positive to pin 8. If you reverse the voltage accidentally, this can permanently damage the chip, so place your jumper wires carefully.

Set your power supply to deliver 9 volts. It will be convenient for this experiment if you supply positive down the righthand side and negative down the lefthand side of the breadboard, as suggested in Figure 4-14. C3 is a large capacitor, at least 100 µF, which is placed across the power supply to smooth it out and provide a local store of charge to fuel fast-switching circuits, as well as to guard against other transient dips in voltage. Although the 555 isn't especially fast-switching, other chips are, and you should get into the habit of protecting them.

Begin with the potentiometer turned all the way counterclockwise to maximize the resistance between the two terminals that we're using, and when you apply the probe from your meter to pin 2, you should measure about 6 volts when you press S1.

Now rotate the potentiometer clockwise and press S1 again. If the LED doesn't light up, keep turning the potentiometer and pressing and releasing the button. When you've turned the potentiometer about two-thirds of the way, you should see the LED light up for just over 5 seconds when you press and release the button. Here are some facts that you should check for yourself:

- The LED will keep glowing after you release the button.
- You can press the button for any length of time (less than the timer's cycle time) and the LED always emits the same length of pulse.
- The timer is triggered by a fall in voltage on pin 2. You can verify this with your meter.
- The LED is either fully on or fully off. You can't see a faint glow when it's off, and the transition from off to on and on to off is very clean and precise.

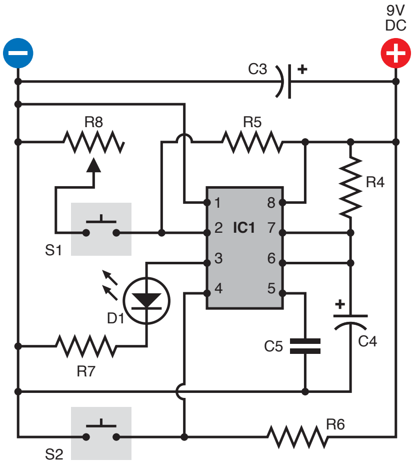

Figure 4-15. *A schematic view of the circuit shown in Figure 4-14. Throughout this chapter, the schematics will be laid out to emulate the most likely placement of components on a breadboard. This is not always the simplest layout, but will be easiest for you to build. Refer to Figure 4-14 for the values of the components.*

Check Figure 4-16 to see how the components should look on your breadboard, and then look at the schematic in Figure 4-15 to understand what's happening. I will be adding more components later, which I will be labeling R1, R2, C1, and C2 to be consistent with data sheets that you may see for the 555 timer. Therefore, in this initial circuit the resistors are labeled R4 and up, and capacitors C3 and up.

When S1 (the tactile switch) is open, pin 2 of the 555 timer receives positive power through R5, which is 2K2. Because the input resistance of the timer is very high, the voltage on pin 2 is almost the full 9 volts.

When you press the button, it connects negative voltage through R8, the 5K potentiometer to pin 2. Thus, R8 and R5 form a voltage divider with pin 2 in the middle. You may remember this concept from when you were testing transistors. The voltage between the resistances will change, depending on the values of the resistances.

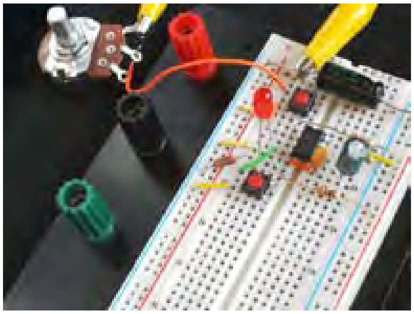

Figure 4-16. *This is how the components look when installed on the breadboard. The alligator clips are attached to a patch cord that links the 100 µF capacitor to the potentiometer. The power supply input is not shown.*

If R8 is turned up about halfway, it is approximately equal to R5, so the midpoint, connected to pin 2, has about half the 9-volt power supply. But when you turn the potentiometer so that its resistance falls farther, the negative voltage outweighs the positive voltage, so the voltage on pin 2 gradually drops.

If you have clips on your meter leads, you can hook them onto the nearest jumper wires and then watch the meter while you turn the potentiometer up and down and press the button.

The graphs in Figure 4-17 illustrate what is happening. The upper graph shows the voltage applied to pin 2 by random button-presses, with the potentiometer turned to various values. The lower graph shows that the 555 is triggered if, and only if, the voltage on pin 2 actively drops from above 3 volts to below 3 volts. What's so special about 3 volts? It's one-third of our 9-volt power supply.

Here's the take-home message:

- The output of the 555 (pin 3) emits a *positive* pulse when the trigger (pin 2) drops *below* one-third of the supply voltage.
- The 555 delivers the *same duration* of positive pulse every time (so long as you don't supply a prolonged low voltage on pin 2).
- A *larger* value for R4 or for C4 will *lengthen* the pulse.
- When the output (pin 3) is high, the voltage is almost equal to the supply voltage. When the output goes low, it's almost zero.

The 555 converts the imperfect world around it into a precise and dependable output. It doesn't switch on and off absolutely instantly, but is fast enough to *appear* instant.

Now here's another thing to try. Trigger the timer so that the LED lights up. While it is illuminated, press S2, the second button, which grounds pin 4, the reset. The LED should go out immediately.

When the reset voltage is pulled *low*, the output goes *low*, regardless of what voltage you apply to the trigger.

There's one other thing I want you to notice before we start using the timer for more interesting purposes. I included R5 and R6 so that when you first switch on the timer, it is not emitting a pulse—but is ready to do so. These resistors apply a positive voltage to the trigger and the reset pin, to make sure that the 555 timer is ready to run when you first apply power to it.

As long as the *trigger* voltage is *high*, the timer *will not* emit a pulse. (It emits a pulse when the trigger voltage drops.)

As long as the *reset* voltage is *high*, the timer is *able* to emit a pulse. (It shuts down when the reset voltage drops.)

R5 and R6 are known as "pull-up resistors" because they pull the voltage up. You can easily overwhelm them by adding a direct connection to the negative side of the power supply. A typical pull-up resistor for the 555 timer is 10K. With a 9-volt power supply, it only passes 0.9mA (by Ohm's Law).

Finally, you may be wondering about the purpose of C5, attached to pin 5. This pin is known the "control" pin, which means that if you apply a voltage to it, you can control the sensitivity of the timer. I'll get to this in more detail a little later. Because we are not using this function right now, it's good practice to put a capacitor on pin 5 to protect it from voltage fluctuations and prevent it from interfering with normal functioning.

Make sure you become familiar with the basic functioning of the 555 timer before you continue.

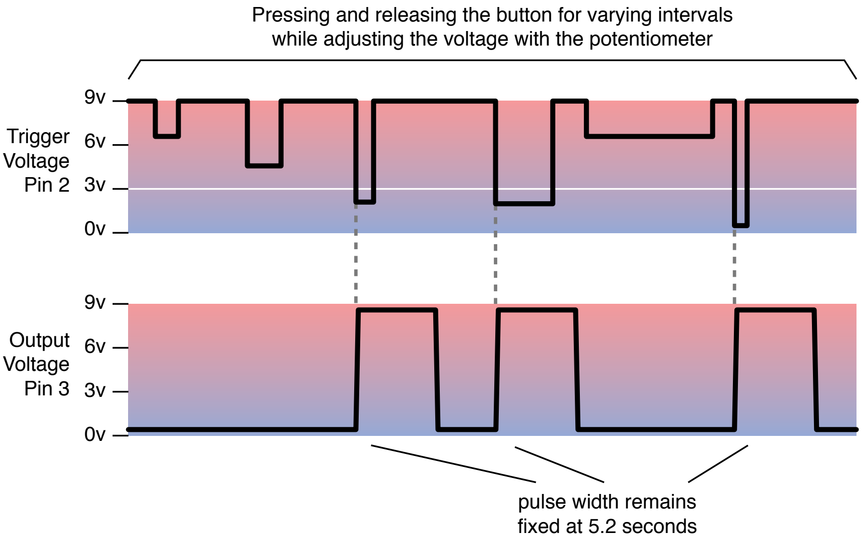

Figure 4-17. *The top graph shows voltage on the trigger (pin 2) when the pushbutton is pressed, for different intervals, at different settings of the potentiometer. The lower graph shows the output (pin 3), which rises until it is almost equal to the power supply, when the voltage on pin 2 drops below 1/3 the full supply voltage.*

FUNDAMENTALS

The following table shows 555 pulse duration in monostable mode:

- Duration is in seconds, rounded to two figures.
- The horizontal scale shows common resistor values between pin 7 and positive supply voltage.
- The vertical scale shows common capacitor values between pin 6 and negative supply voltage.

To calculate a different pulse duration, multiply resistance × capacitance × 0.0011 where resistance is in kilohms, capacitance is in microfarads, and duration is in seconds.

	1K	2K2	4K7	10K	22K	47K	100K	220K	470K	1M
47 µF	0.05	0.11	0.24	0.52	1.1	2.4	5.2	11	24	52
22 µF	0.02	0.05	0.11	0.24	0.53	1.1	2.4	5.3	11	24
10 µF	0.01	0.02	0.05	0.11	0.24	0.52	1.1	2.4	5.2	11
4.7 µF		0.01	0.02	0.05	0.11	0.24	0.52	1.1	2.4	5.2
2.2 µF			0.01	0.02	0.05	0.11	0.24	0.53	1.1	2.4
1.0 µF				0.01	0.02	0.05	0.11	0.24	0.52	1.1
0.47 µF					0.01	0.02	0.05	0.11	0.24	0.52
0.22 µF						0.01	0.02	0.05	0.11	0.24
0.1 µF							0.01	0.02	0.05	0.11
0.047 µF								0.01	0.02	0.05
0.022 µF									0.01	0.02
0.01 µF										0.01

THEORY

Inside the 555 timer: monostable mode

The plastic body of the 555 timer contains a wafer of silicon on which are etched dozens of transistor junctions in a pattern that is far too complex to be explained here. However, I can summarize their function by dividing them into groups, as shown in Figure 4-18. An external resistor and two external capacitors are also shown, labeled the same way as in Figure 4-15.

The negative and positive symbols inside the chip are power sources which actually come from pins 1 and 8, respectively. I omitted the internal connections to those pins for the sake of clarity.

The two yellow triangles are "comparators." Each comparator compares two inputs (at the base of the triangle) and delivers an output (from the apex of the triangle) depending on whether the inputs are similar or different. We'll be using comparators for other purposes later in this book.

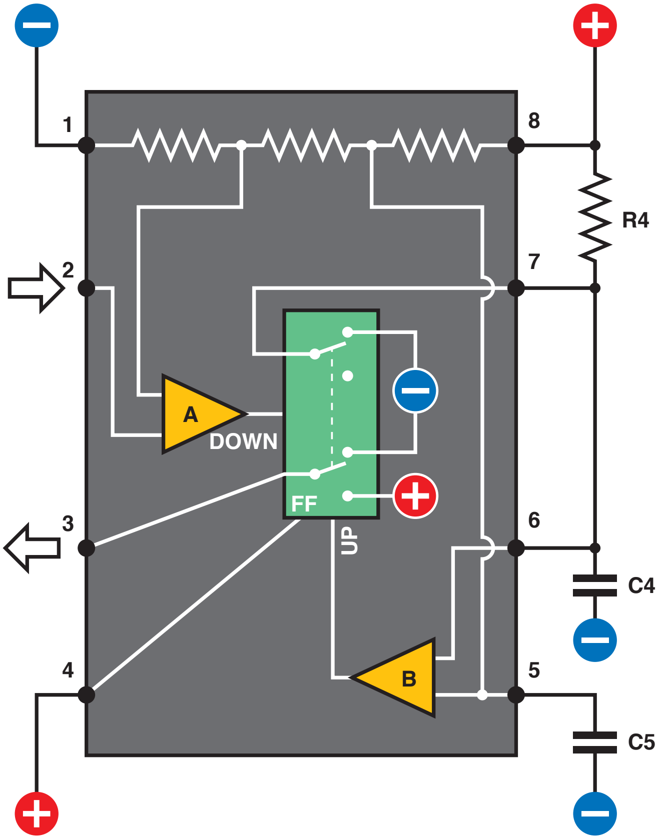

Figure 4-18. *Inside the 555 timer. White lines indicate connections inside the chip. A and B are comparators. FF is a flip-flop which can rest in one state or the other, like a double-throw switch. A drop in voltage on pin 2 is detected by comparator A, which triggers the flip-flop into its "down" position and sends a positive pulse out of pin 3. When C4 charges to 2/3 of supply voltage, this is detected by comparator B, which resets the flip-flop to its "up" position. This discharges C4 through pin 7.*

THEORY

Inside the 555 timer: monostable mode (continued)

The green rectangle, identified as "FF," is a "flip-flop." I have depicted it as a DPDT switch, because that's how it functions here, although of course it is really solid-state.

Initially when you power up the chip, the flip-flop is in its "up" position which delivers low voltage through the output, pin 3. If the flip-flop receives a signal from comparator A, it flips to its "down" state, and flops there. When it receives a signal from comparator B, it flips back to its "up" state, and flops there. The "UP" and "DOWN" labels on the comparators will remind you what each one does when it is activated.

Flip-flops are a fundamental concept in digital electronics. Computers couldn't function without them.

Notice the external wire that connects pin 7 with capacitor C4. As long as the flip-flop is "up," it sinks the positive voltage coming through R4 and prevents the capacitor from charging positively.

If the voltage on pin 2 drops to 1/3 of the supply, comparator A notices this, and flips the flip-flop. This sends a positive pulse out of pin 3, and also disconnects the negative power through pin 7. So now C4 can start charging through R4. While this is happening, the positive output from the timer continues.

As the voltage increases on the capacitor, comparator B monitors it through pin 6, known as the threshold. When the capacitor accumulates 2/3 of the supply voltage, comparator B sends a pulse to the flip-flop, flipping it back into its original state. This discharges the capacitor through pin 7, appropriately known as the discharge pin. Also, the flip-flop ends the positive output through pin 3 and replaces it with a negative voltage. This way, the 555 returns to its original state.

I'll sum up this sequence of events very simply:

1. Initially, the flip-flop grounds the capacitor and grounds the output (pin 3).
2. A drop in voltage on pin 2 to 1/3 the supply voltage or less makes the output (pin 3) positive and allows capacitor C4 to start charging through R4.
3. When the capacitor reaches 2/3 of supply voltage, the chip discharges the capacitor, and the output at pin 3 goes low again.

In this mode, the 555 timer is "monostable," meaning that it just gives one pulse, and you have to trigger it again to get another.

You adjust the length of each pulse by changing the values of R4 and C4. How do you know which values to choose? Check the table on page 157, which gives an approximate idea and also includes a formula so that you can calculate values of your own.

I didn't bother to include pulses shorter than 0.01 second in the table, because a single pulse of this length is usually not very useful. Also I rounded the numbers in the table to 2 significant figures, because capacitor values are seldom more accurate than that.

BACKGROUND

How the timer was born

Back in 1970, when barely a half-dozen corporate seedlings had taken root in the fertile ground of Silicon Valley, a company named Signetics bought an idea from an engineer named Hans Camenzind. It wasn't a huge breakthrough concept—just 23 transistors and a bunch of resistors that would function as a programmable timer. The timer would be versatile, stable, and simple, but these virtues paled in comparison to its primary selling point. Using the emerging technology of integrated circuits, Signetics could reproduce the whole thing on a silicon chip.

This entailed some trial and error. Camenzind worked alone, building the whole thing initially on a large scale, using off-the-shelf transistors, resistors, and diodes on a breadboard. It worked, so then he started substituting slightly different values for the various components to see whether the circuit would tolerate variations during production and other factors such as changes in temperature when the chip was in use. He made at least 10 different versions of the circuit. It took months.

Next came the crafts work. Camenzind sat at a drafting table and used a specially mounted X-Acto knife to scribe his circuit into a large sheet of plastic. Signetics then reduced this image photographically by a ratio of about 300:1. They etched it into tiny wafers, and embedded each of them in a half-inch rectangle of black plastic with the product number printed on top. Thus, the 555 timer was born.

It turned out to be the most successful chip in history, both in the number of units sold (tens of billions and counting) and the longevity of its design (unchanged in almost 40 years). The 555 has been used in everything from toys to spacecraft. It can make lights flash, activate alarm systems, put spaces between beeps, and create the beeps themselves.

Today, chips are designed by large teams and tested by simulating their behavior using computer software. Thus, chips inside a computer enable the design of more chips. The heyday of solo designers such as Hans Camenzind is long gone, but his genius lives inside every 555 timer that emerges from a fabrication facility. (If you'd like to know more about chip history, see *http://www.semiconductormuseum.com/Museum_Index.htm*.)

Figure 4-13. *Hans Camenzind, inventor and developer of the 555 timer chip for Signetics.*

FUNDAMENTALS

Why the 555 is useful

In its monostable mode (which is what you just saw), the 555 will emit a single pulse of fixed (but programmable) length. Can you imagine some applications? Think in terms of the pulse from the 555 controlling some other component. A motion sensor on an outdoor light, perhaps. When an infra-red detector "sees" something moving, the light comes on for a specific period—which can be controlled by a 555.

Another application could be a toaster. When someone lowers a slice of bread, a switch will close that triggers the toasting cycle. To change the length of the cycle, you could use a potentiometer instead of R4 and attach it to the external lever that determines how dark you want your toast. At the end of the toasting cycle, the output from the 555 would pass through a power transistor, to activate a solenoid (which is like a relay, except that it has no switch contacts) to release the toast.

Intermittent windshield wipers could be controlled by a 555 timer—and on earlier models of cars, they actually were.

And what about the burglar alarm that was described at the end of Chapter 3? One of the features that I listed, which has not been implemented yet, is that it should shut itself off after a fixed interval. We can use the change of output from a 555 timer to do that.

The experiment that you just performed seemed trivial, but really it implies all kinds of possibilities.

555 timer limits

1. The timer can run from a stable voltage source ranging from 5 to 15 volts.
2. Most manufacturers recommend a range from 1K to 1M for the resistor attached to pin 7.
3. The capacitor value can go as high as you like, if you want to time really long intervals, but the accuracy of the timer will diminish.
4. The output can deliver as much as 100mA at 9 volts. This is sufficient for a small relay or miniature loudspeaker, as you'll see in the next experiment.

Beware of Pin-Shuffling!

In all of the schematics in this book, I'll show chips as you'd see them from above, with pin 1 at top left. Other schematics that you may see, on websites or in other books, may do things differently. For convenience in drawing circuits, people shuffle the pin numbers on a chip so that pin 1 isn't necessarily shown adjacent to pin 2.

Look at the schematic in Figure 4-20 and compare it with the one in Figure 4-15. The connections are the same, but the one in Figure 4-20 groups pins to reduce the apparent complexity of the wiring.

"Pin shuffling" is common because circuit-drawing software tends to do it, and on larger chips, it is necessary for functional clarity of the schematic (i.e., logical groupings of pin names versus physical groupings on memory chips, for example). When you're first learning to use chips, I think it's easier to understand a schematic that shows the pins in their actual positions. So that's the practice I will be using here.

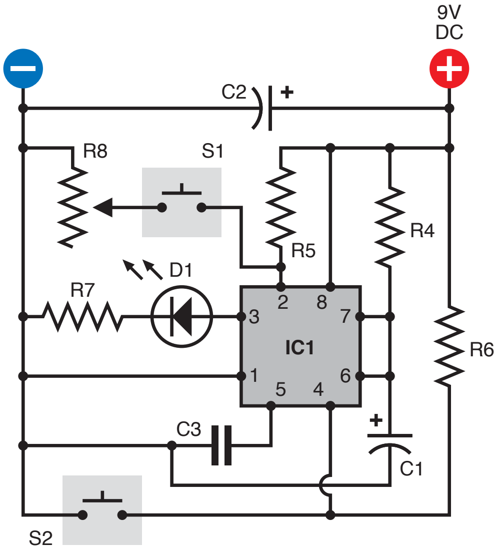

Figure 4-20. *Many people draw schematics in which the pin numbers on a chip are shuffled around to make the schematic smaller or simpler. This is not helpful when you try to build the circuit. The schematic here is for the same circuit as in Figure 4-15. This version would be harder to recreate on a breadboard.*

Experiment 17: Set Your Tone

I'm going to show you two other ways in which the 555 timer can be used.

You will need the same items as in Experiment 16, plus:

- Additional 555 timer chip. Quantity: 1.
- Miniature loudspeaker. Quantity: 1.
- 100K linear potentiometer. Quantity: 1.

Procedure

Leave the components from Experiment 16 where they are on the breadboard, and add the next section below them, as shown in Figures 4-21 and 4-22. Resistor R2 is inserted between pins 6 and 7, instead of the jumper wire that shorted the pins together in the previous circuit, and there's no external input to pin 2 anymore. Instead, pin 2 is connected via a jumper wire to pin 6. The easiest way to do this is by running the wire across the top of the chip.

I have omitted the smoothing capacitor from the schematic in Figure 4-22, because I'm assuming that you're running this circuit on the same breadboard as the first, where the previous smoothing capacitor is still active.

A loudspeaker in series with a 100Ω resistor (R3) has been substituted for the LED to show the output from the chip. Pin 4, the reset, is disabled by connecting it to the positive voltage supply, as I'm not expecting to use the reset function in this circuit.

Now what happens when you apply power? Immediately, you should hear noise through the loudspeaker. If you don't hear anything, you almost certainly made a wiring error.

Notice that you don't have to trigger the chip with a pushbutton anymore. The reason is that when C1 charges and discharges, its fluctuating voltage is connected via a jumper wire across the top of the chip to pin 2, the trigger. In this way, the 555 timer now triggers itself. I'll describe this in more detail in the next section "Theory: Inside the 555 timer: astable mode," if you want to see exactly what is going on.

In this mode, the chip is "astable," meaning that it is not stable, because it flips to and fro endlessly, sending a stream of pulses for as long as the power is connected. The pulses are so rapid that the loudspeaker reproduces them as noise.

In fact, with the component values that I specified for R1, R2, and C1, the 555 chip is emitting about 1,500 pulses per second. In other words, it creates a 1.5 KHz tone.

Check the table on page 166 to see how different values for R2 and C1 can create different pulse frequencies with the chip in this astable mode. Note that the table assumes a fixed value of 1K for R1!

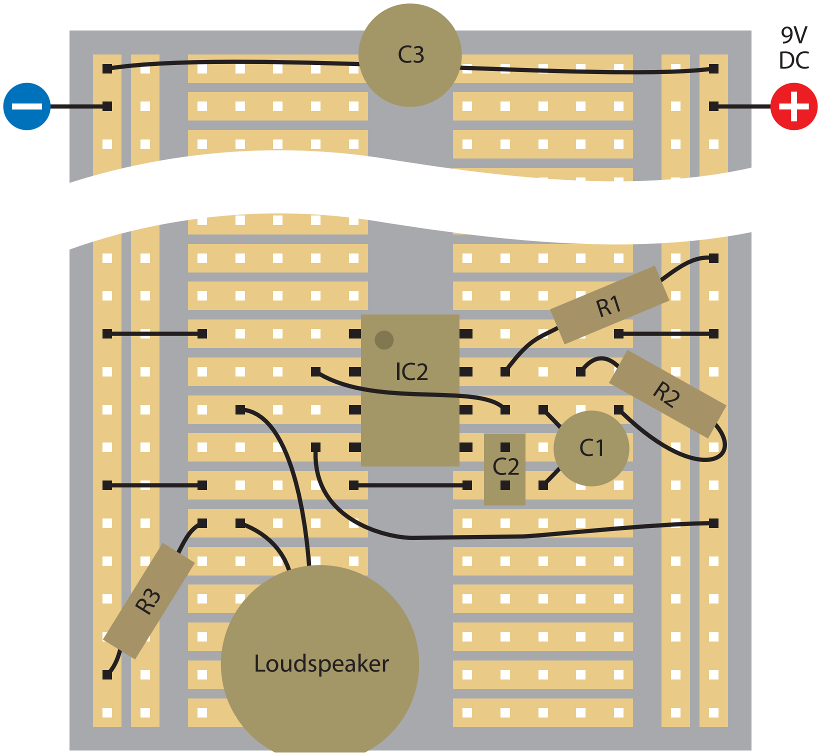
C3
9V
DC
R1
IC2
R2
C1
C2
R3
Loudspeaker

Figure 4-21. *These components should be added on the same breadboard below the components shown in Figure 4-14. Use the following values to test the 555 timer in its astable mode:*

R1: 1K
R2: 10K
R3: 100Ω
C1: 0.047 µF ceramic or electrolytic
C2: 0.1 µF ceramic
IC2: 555 timer

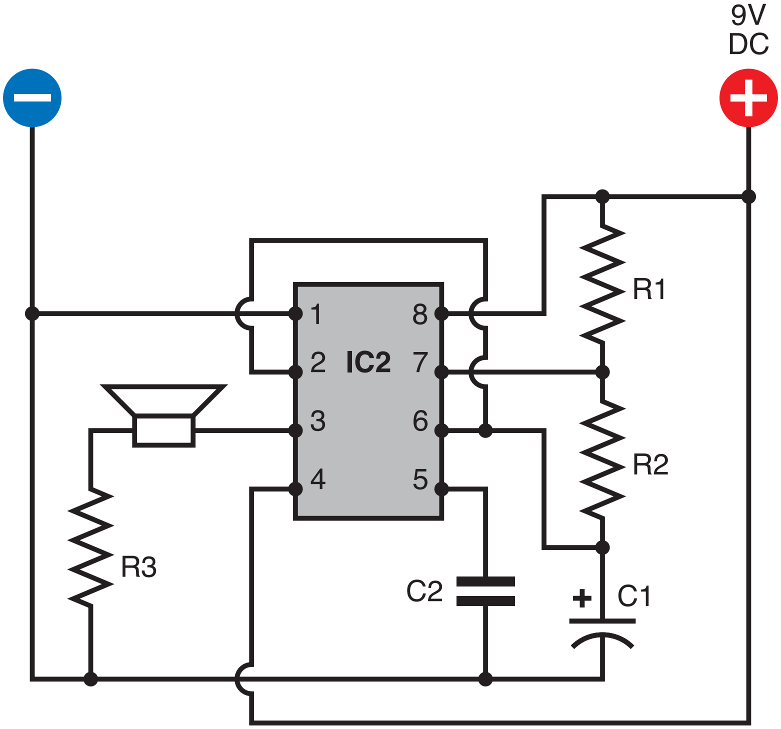
9V
DC
1
2
IC2
3
4
8
7
6
5
R1
R2
R3
C2
+
C1

Figure 4-22. *This is the schematic version of the circuit shown in Figure 4-21. The component values are the same.*

THEORY

Inside the 555 timer: astable mode

Here's what is happening now, illustrated in Figure 4-23. Initially, the flip-flop grounds C1 as before. But now the low voltage on the capacitor is connected from pin 7 to pin 2 through an external wire. The low voltage tells the chip to trigger itself. The flip-flop obediently flips to its "on" position and sends a positive pulse to the loudspeaker, while removing the negative voltage from pin 7.

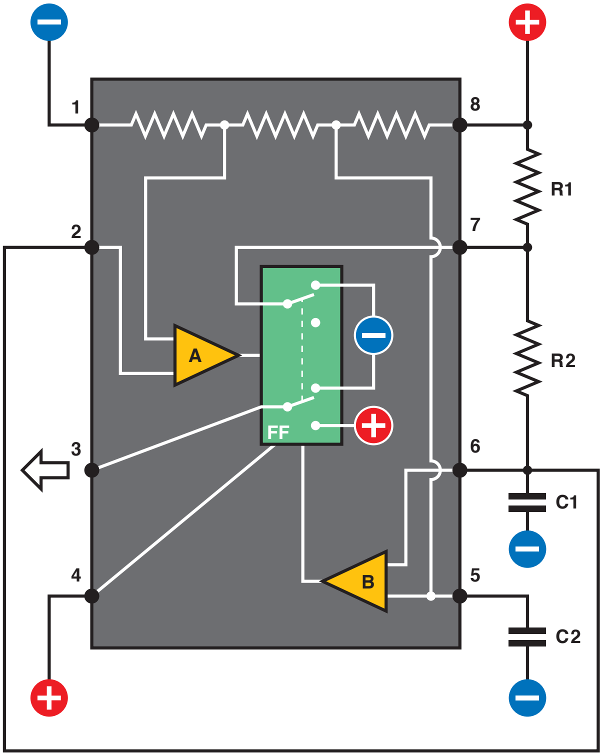

Figure 4-23. *When the 555 timer is used in astable mode, resistor R2 is placed between pin 6 and pin 7, and pin 6 is connected via an external wire to pin 2, so that the timer triggers itself.*

Now C1 starts charging, as it did when the timer was in monostable mode, except that it is being charged through R1 + R2 in series. Because the resistors have low values, and C1 is also small, C1 charges quickly. When it reaches 2/3 full voltage, comparator B takes action as before, discharging the capacitor and ending the output pulse from pin 3.

The capacitor takes longer to discharge than before, because R2 has been inserted between it and pin 7, the discharge pin. While the capacitor is discharging, its voltage diminishes, and is still linked to pin 2. When the voltage drops to 1/3 of full power or less, comparator A kicks in and sends another pulse to the flip-flop, starting the process all over again.

Summing up:

1. In astable mode, as soon as power is connected, the flip-flop pulls down the voltage on pin 2, triggering comparator A, which flips the flip-flop to its "down" position.
2. Pin 3, the output, goes high. The capacitor charges through R1 and R2 in series.
3. When the capacitor reaches 2/3 of supply voltage, the flip-flop goes "up" and the output at pin 3 goes low. The capacitor starts to discharge through R2.
4. When the charge on the capacitor diminishes to 1/3 of full voltage, the pull-down on pin 2 flips the flip-flop again and the cycle repeats.

Unequal on/off cycles

When the timer is running in astable mode, C1 charges through R1 and R2 in series. But when C1 discharges, it dumps its voltage through R2 only. This means that the capacitor charges more slowly than it discharges. While it is charging, the output on pin 3 is high; while it is discharging, the output on pin 3 is low. Consequently the "on" cycle is always longer than the "off" cycle. Figure 4-24 shows this as a simple graph.

If you want the on and off cycles to be equal, or if you want to adjust the on and off cycles independently (for example, because you want to send a very brief pulse to another chip, followed by a longer gap until the next pulse), all you need to do is add a diode, as shown in Figure 4-25.

Now when C1 charges, the electricity flows through R1 as before but takes a shortcut around R2, through diode D1. When C1 discharges, the diode blocks the flow of electricity in that direction, and so the discharge goes back through R2.

THEORY

Inside the 555 timer: astable mode (continued)

R1 now controls the charge time on its own, while R2 controls the discharge time. The formula for calculating the frequency is now:

Frequency = 1440 / ((R1 + R2) × C1)

If you set R1 = R2, you should get almost equal on/off cycles ("almost" because the diode itself imposes a small internal voltage drop of about 0.6V). The exact value depends primarily on the manufacturing process used to make the diode.

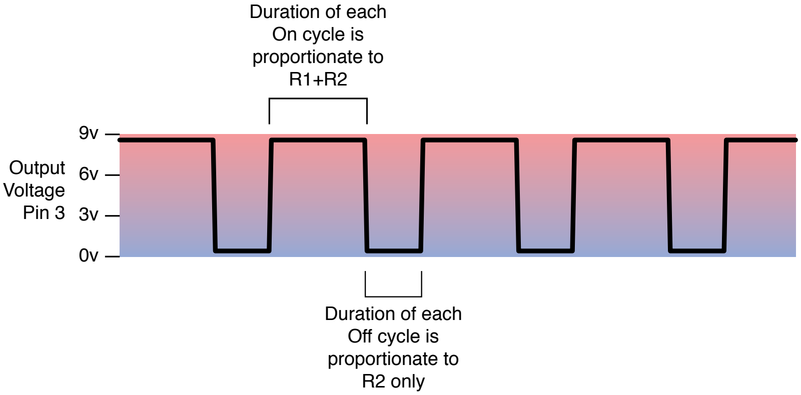

Figure 4-24. *In its usual astable configuration, the timer charges a capacitor through R1+R2 and discharges the capacitor through R2 only. Therefore its output on cycles are longer than its output off cycles.*

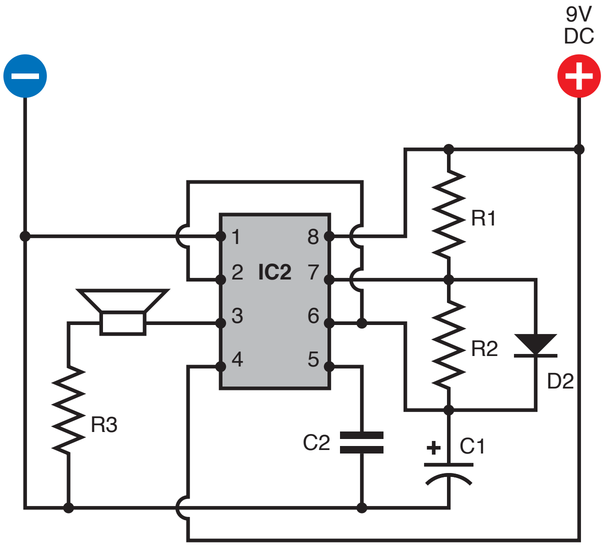

Figure 4-25. *This is a modification of the schematic shown in Figure 4-22. By adding a diode to a 555 timer running in astable mode, we eliminate R2 from the charging cycle of capacitor C1. Now we can adjust the output on cycle with the value of R1, and the output off cycle with the value of R2, so that the on and off durations are independent of each other.*

FUNDAMENTALS

The following table shows 555 timer frequency in astable mode:

- Frequency is in pulses per second, rounded to two figures.
- The horizontal scale shows common resistor values for R2.
- The vertical scale shows common capacitor values for C1. Resistor R1 is assumed to be 1K.
- Resistor R1 is assumed to be 1K.

To calculate a different frequency: double R2, add the product to R1, multiply the sum by C1, and divide the result into 1440. Like this:

Frequency = 1440 / ((R1 + 2R2) × C1) cycles per second

In this formula, R1 and R2 are in kilohms, C1 is in microfarads, and the frequency is in hertz (cycles per second). Note that the frequency is measured from the start of one pulse to the start of the next. The duration of each pulse is not the same as the length of time between each pulse. This issue is discussed in the previous section.

47 µF	10	5.7	3.0	1.5	0.7	0.3	0.2	0.1		
22 µF	22	12	6.3	3.1	1.5	0.7	0.3	0.2	0.1	
10 µF	48	27	14	6.9	3.2	1.5	0.7	0.3	0.2	0.1
4.7 µF	100	57	30	15	6.8	3.2	1.5	0.7	0.3	0.2
2.2 µF	220	120	63	31	15	6.9	3.3	1.5	0.7	0.3
1.0 µF	480	270	140	69	32	15	7.2	3.3	1.5	0.7
0.47 µF	1,000	570	300	150	68	32	15	7	3.3	1.5
0.22 µF	2,200	1,200	630	310	150	69	33	15	7	3.3
0.1 µF	4,800	2,700	1,400	690	320	150	72	33	15	7.2
0.047 µF	10,000	5,700	3,000	1,500	680	320	150	70	33	15
0.022 µF	22,000	12,000	6,300	3,100	1,500	690	330	150	70	33
0.01 µF	48,000	27,000	14,000	6,900	3,200	1,500	720	330	150	72
	1K	2K2	4K7	10K	22K	47K	100K	220K	470K	1M

Astable Modifications

In the circuits shown in Figures 4-22 or 4-25, if you substitute a 100K potentiometer for R2, you can adjust the frequency up and down by turning the shaft.

Another option is to "tune" the timer by using pin 5, the control, as shown in the Figure 4-26. Disconnect the capacitor that was attached to that pin and substitute the series of resistors shown. R9 and R11 are both 1K resistors, either side of R10, which is a 100K potentiometer. They ensure that pin 5 always has at least 1K between it and the positive and negative sides of the power supply. Connecting it directly to the power supply won't damage the timer, but will prevent it from generating audible tones. As you turn the potentiometer to and fro, the frequency will vary over a wide range. If you want to generate a very specific frequency, a trimmer potentiometer can be used instead.

A primary advantage of using pin 5 to adjust frequency is that you can control it remotely. Take the output from pin 3 of another 555 timer running slowly in astable mode, and pipe it through a 2K2 resistor to pin 5. Now you get a two-tone siren effect, as one timer controls the other. If, in addition, you add a 100 µF capacitor between pin 5 and ground, the charging and discharging of the capacitor will make the tone slide up and down instead of switching abruptly. I'll describe this in more detail shortly. This leads me to the whole topic of one chip controlling another chip, which will be our last variation on this experiment.

Chaining Chips

Generally speaking, chips are designed so that they can talk to each other. The 555 couldn't be easier in this respect:

- Pin 3, the output, from one 555 can be connected directly to pin 2, the trigger, of a second 555.
- Alternatively, the output can be sufficient to provide power to pin 8 of a second 555.
- The output is appropriate to control or power other types of chips too.

Figure 4-27 shows these options.

When the output from the first 555 goes high, it is about 70 to 80% of its supply voltage. In other words, when you're using a 9V supply, the high output voltage is at least 6 volts. This is still above the minimum of 5V that the second chip needs to trigger its comparator, so there's no problem.

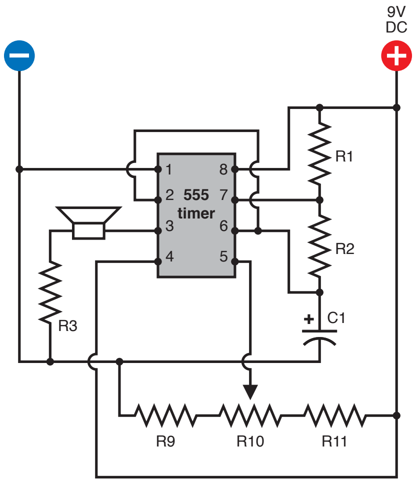
9V
DC
1
2
3
4
8
7
6
5
555
timer
R1
R2
C1
R3
R9
R10
R11

Figure 4-28. *The control (pin 5) is seldom used but can be useful. Varying the voltage on it will adjust the speed of the timer. This circuit enables you to test the behavior of it. Component values:*

R1: 1K
R2: 10K
R3: 100 ohms
R9, R11: 1K
R10: 100K linear potentiometer
C1: 0.047 µF

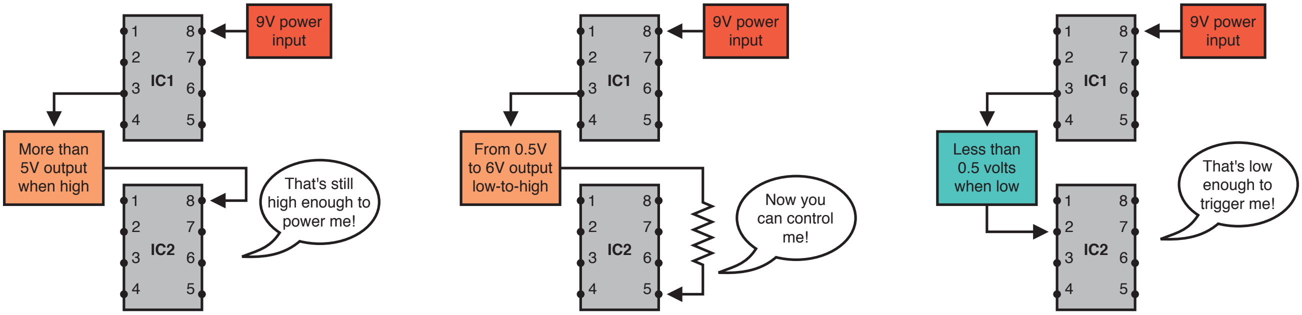
9V power input
IC1
IC2
More than 5V output when high
That's still high enough to power me!
9V power input
IC1
IC2
From 0.5V to 6V output low-to-high
Now you can control me!
9V power input
IC1
IC2
Less than 0.5 volts when low
That's low enough to trigger me!

Figure 4-27. *Three ways to chain 555 timers together. The output of IC1 can power a second timer, or adjust its control voltage, or activate its trigger pin.*

You can chain together the two 555 timers that you already have on your breadboard. Figure 4-28 shows how to connect the two circuits that were shown previously in Figures 4-15 and 4-22. Run a wire from pin 3 (the output) of the first chip to pin 8 (the positive power supply) of the second chip, and disconnect the existing wire connecting pin 8 to your power supply. The new wire is shown in red. Now when you press the button to activate the first chip, its output powers the second chip.

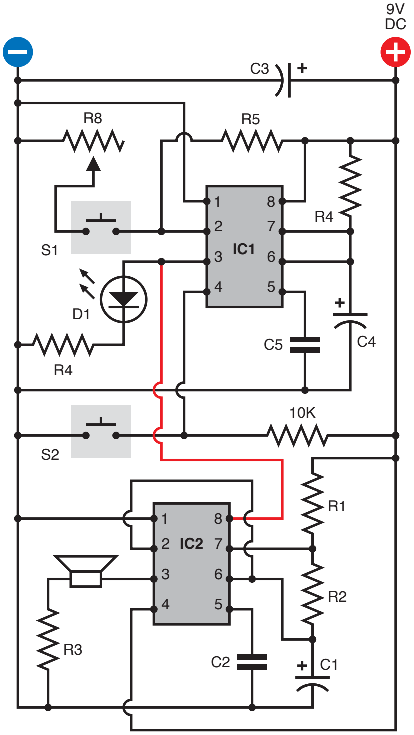

Figure 4-28. *You can combine the two circuits shown in Figures 4-15 and 4-22 simply by disconnecting the wire that provides power to pin 8 of the second timer, and running a substitute wire (shown in red).*

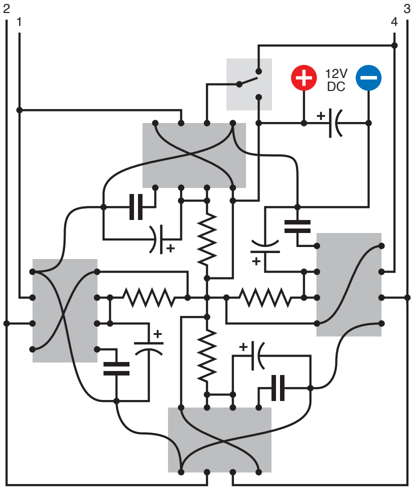

Figure 4-29. *Four 555 timers, chained together in a circle, can flash a series of four sets of LEDs in sequence, like Christmas lights or a movie marquee.*

You can also use the output from one chip to trigger another (i.e., you can connect pin 3 from the first chip to pin 2 of the second). When the output from the first chip is low, it's less than half a volt. This is well below the threshold that the second chip requires to be activated. Why would you want to do this? Well, you might want to have both timers running in monostable mode, so that the end of a high pulse from the first one triggers the start of a high pulse in the second one. In fact, you could chain together as many timers as you like in this way, with the last one feeding back and triggering the first one, and they could flash a series of LEDs in sequence, like Christmas lights. Figure 4-29 shows how four timers could be linked this way, in a configuration that would occupy minimal space (and would be wired point-to-point on perforated board, not on breadboard-format board). Each of the outputs numbered 1 through 4 would have about enough power to run maybe 10 LEDs, if you used relatively high load resistors to limit their current.

Incidentally, you can reduce the chip count (the number of chips) by using two 556 timers instead of four 555 timers. The 556 contains a pair of 555 timers in one package. But because you have to make the same number of external connections (other than the power supply), I haven't bothered to use this variant.

You can even get a 558 timer that contains four 555 circuits, all preset to function in astable mode. I decided not to use this chip, because its output behaves differently from a normal 555 timer. But you can buy a 558 timer and play with it if you wish. It is ideal for doing the "chain of four timers" that I suggested previously. The data sheet even suggests this.

Lastly, going back to the idea of modifying the frequency of a 555 timer in astable mode, you can chain two timers, as shown in Figure 4-30. The red wire shows the connection from the output of the first timer to the control pin of the second. The first timer has now been rewired in astable mode, so that it creates an oscillating on/off output around four times per second. This output flashes the LED (to give you a visual check of what's going on) and feeds through R7 to the control pin of the second timer.

But C2 is a large capacitor, which takes time to charge through R7. While this happens, the voltage detected by pin 5 slowly rises, so that the tone generated by IC2 gradually rises in pitch. Then IC1 reaches the end of its on cycle and switches itself off, at which point C2 discharges and the pitch of the sound generated by IC2 falls again.

You can tweak this circuit to create all kinds of sounds, much more controllably then when you were using PUT transistors to do the same kind of thing. Here are some options to try:

- Double or halve the value of C2.
- Omit C2 completely, and experiment with the value of R7.
- Substitute a 10K potentiometer for R7.
- Change C4 to increase or decrease the cycle time of IC1.

- Halve the value of R5 while doubling the value of C4, so that the cycle time of IC1 stays about the same, but the On time becomes significantly longer than the Off time.
- Change the supply voltage in the circuit from 9 volts to 6 volts or 12 volts.

Remember, you can't damage a 555 timer by making changes of this kind. Just make sure that the negative side of your power supply goes to pin 1 and the positive side to pin 8.

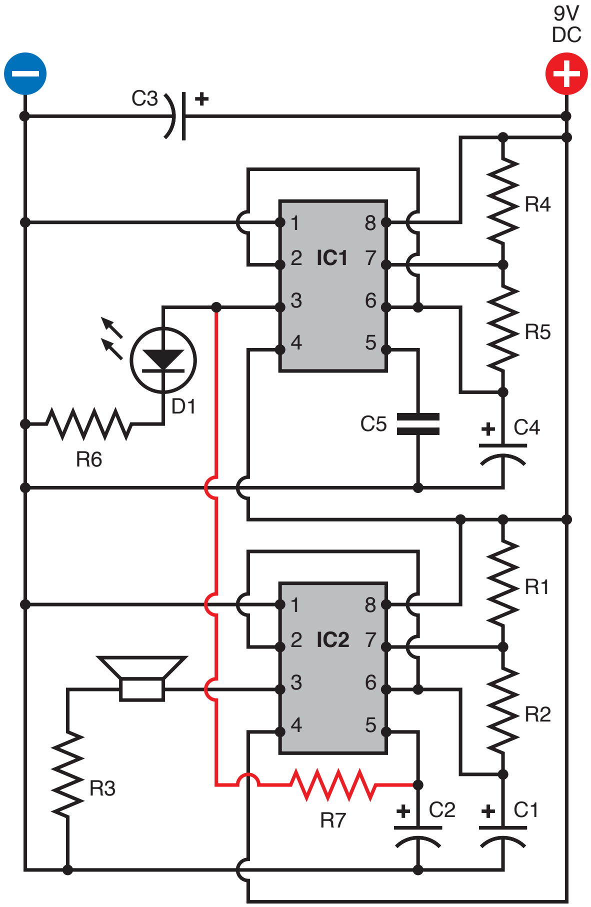

Figure 4-30. *When both timers are astable, but IC1 runs much more slowly than IC2, the output from IC1 can be used to modulate the tone generated by IC2. Note that as this is a substantial modification to the previous schematics, several components have been relabeled. To avoid errors, you may need to remove the old circuit from your breadboard and build this version from scratch. Try these values initially:*

R1, R4, R6, R7: 1K
R2, R5: 10K
R3: 100 ohms
C1: 0.047 µF
C2, C3: 100 µF
C4: 68 µF
C5: 0.1 µF

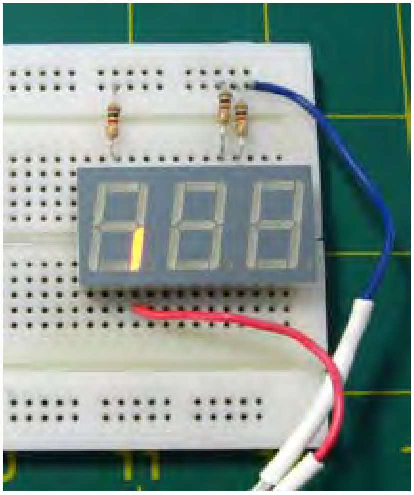

Figure 4-31. *After putting a 1K resistor between the common cathode of the display and the negative supply voltage, you can use the positive supply voltage to illuminate each segment in turn.*

Experiment 18: Reaction Timer

Because the 555 can easily run at thousands of cycles per second, we can use it to measure human reactions. You can compete with friends to see who has the fastest response—and note how your response changes depending on your mood, the time of day, or how much sleep you got last night.

Before going any further, I have to warn you that this circuit will have more connections than others you've tackled so far. It's not conceptually difficult, but requires a lot of wiring, and will only just fit on a breadboard that has 63 rows of holes. Still, we can build it in a series of phases, which should help you to detect any wiring errors as you go.

You will need:

- 4026 chip. Quantity: 4 (really you need only 3, but get another one in case you damage the others).
- 555 timers. Quantity: 3.
- Tactile switches (SPST momentary switches). Quantity: 3.
- Three numeric LEDs, or one 3-digit LED display (see the shopping list at the beginning of this chapter). Quantity: 1.
- Breadboard, resistors, capacitors, and meter, as usual.

Step 1: Display

You can use three separate LED numerals for this project, but I suggest that you buy the Kingbright BC56-11EWA on the shopping list at the beginning of this chapter. It contains three numerals in one big package.

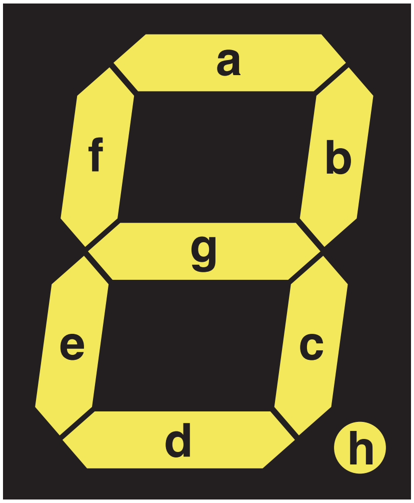

Figure 4-32. *The most basic and common digital numeral consists of seven LED segments identified by letters, as shown here, plus an optional decimal point.*

You should be able to plug it into your breadboard, straddling the center channel. Put it all the way down at the bottom of the breadboard, as shown in Figure 4-31. Don't put any other components on the breadboard yet.

Now set your power supply to 9 volts, and apply the *negative* side of it to the row of holes running up the breadboard on the *righthand* side. Insert a 1K resistor between that negative supply and each of pins 18, 19, and 26 of the Kingbright display, which are the "common cathode," meaning the negative connection shared by each set of LED segments in the display. (The pin numbers of the chip are shown in Figure 4-33. If you're using another model of display, you'll have to consult a data sheet to find which pin(s) are designed to receive negative voltage.)

Switch on the power supply and touch the free end of the positive wire to each row of holes serving the display on its left and right sides. You should see each segment light up, as shown in Figure 4-31.

Each numeral from 0 to 9 is represented by a group of these segments. The segments are always identified with lowercase letters *a* through *g*, as shown in Figure 4-32. In addition, there is often a decimal point, and although we won't be using it, I've identified it with the letter *h*.

Check Figure 4-33 showing the Kingbright display, and you'll see I have annotated each pin with its function. You can step down the display with the positive wire from your power supply, making sure that each pin lights an appropriate segment.

Incidentally, this display has two pins, numbered 3 and 26, both labeled to receive negative voltage for the first of the digits. Why two pins instead of one? I don't know. You need to use only one, and as this is a passive chip, it doesn't matter if you leave the unused one unconnected. Just take care not to apply positive voltage to it, which would create a short circuit.

A numeric display has no power or intelligence of its own. It's just a bunch of light-emitting diodes. It's not much use, really, until we can figure out a way to illuminate the LEDs in appropriate groups—which will be the next step.

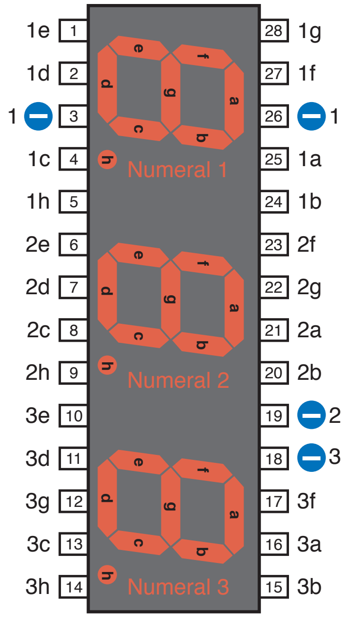

Figure 4-33. *This Kingbright unit incorporates three seven-segment numeric displays in one package, and can be driven by three chained 4026 decade counters. The pin numbers are shown close to the chip. Segments a through g of numeral 1 are identified as 1a through 1g. Segments a through g of numeral 2 are identified as 2a through 2g. Segments a through g of numeral 3 are identified as 3a through 3g.*

Step 2: Counting

Fortunately, we have a chip known as the 4026, which receives pulses, counts them, and creates an output designed to work with a seven-segment display so that it shows numbers 0–9. The only problem is that this is a rather old-fashioned CMOS chip (meaning, Complementary Metal Oxide Semiconductor) and is thus sensitive to static electricity. Check the caution on page 172 before continuing.

Switch off your power supply and connect its wires to the top of the breadboard, noting that for this experiment, we're going to need positive and negative power on both sides. See Figure 4-34 for details. If your breadboard does not already have the columns of holes color-coded, I suggest you use Sharpie markers to identify them, to avoid polarity errors that can fry your components.

The 4026 counter chip is barely powerful enough to drive the LEDs in our display when powered by 9 volts. Make sure you have the chip the right way up, and insert it into the breadboard immediately above your three-digit display, leaving just one row of holes between them empty.

The schematic in Figure 4-35 shows how the pins of the 4026 chip should be connected. The arrows tell you which pins on the display should be connected with pins on the counter.

Figure 4-36 shows the "pinouts" (i.e., the functions of each pin) of a 4026 counter chip. You should compare this with the schematic in Figure 4-35.

Include a tactile switch between the positive supply and pin 1 of the 4026 counter, with a 10K resistor to keep the input to the 4026 counter negative until the button is pressed. Make sure all your positives and negatives are correct, and turn on the power. You should find that when you tap the tactile switch lightly, the counter advances the numeric display from 0 through 9 and then begins all over again from 0. You may also find that the chip sometimes misinterprets your button-presses, and counts two or even three digits at a time. I'll deal with this problem a little further on.

The LED segments will not be glowing very brightly, because the 1K series resistors deprive of them of the power they would really like to receive. Those resistors are necessary to avoid overloading the outputs from the counter.

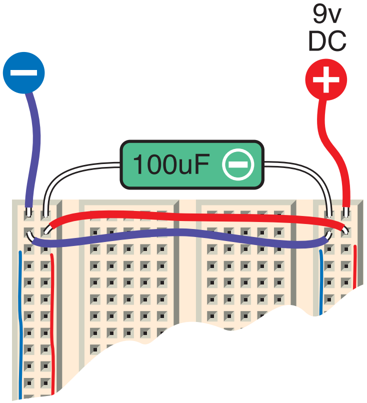

Figure 4-34. *When building circuits around chips, it's convenient to have a positive and negative power supply down each side of your breadboard. For the reaction timer circuit, a 9V supply with a 100 µF smoothing capacitor can be set up like this. If your breadboard doesn't color-code the columns of holes on the left and right sides, I suggest you do that yourself with a permanent marker.*

Grounding Yourself

To avoid the frustration that occurs when you power up a circuit and nothing happens, be sure to take these precautions when you use the older generation of CMOS chips (which often have part numbers from 4000 upward, such as 4002, 4020, and so on):

Chips are often shipped with their legs embedded in black foam. This is electrically conductive foam, and you should keep the chips embedded in it until you are ready to use them.

If the chips are supplied to you in plastic tubes, you can take them out and poke their legs into pieces of conductive foam or, if you don't have any, use aluminum foil. The idea is to avoid one pin on a chip acquiring an electric potential that is much higher than another pin.

While handling CMOS components, grounding yourself is important. I find that in dry weather, I accumulate a static charge merely by walking across a plastic floor-protecting mat in socks that contain some synthetic fibers. You can buy a wrist strap to keep yourself grounded, or simply touch a large metal object, such as a file cabinet, before you touch your circuit board. I am in the habit of working with my socked foot touching a file cabinet, which takes care of the problem.

Never solder a CMOS chip while there is power applied to it.

Grounding the tip of your soldering iron is a good idea.

Better still, don't solder CMOS chips at all. When you're ready to immortalize a project by moving it from a breadboard into perforated board, solder a socket into your perforated board, then push the chip into the socket. If there's a problem in the future, you can unplug the chip and plug in another.

Use a grounded, conductive surface on your workbench. The cheapest way to do this is to unroll some aluminum foil and ground it (with an alligator clip and a length of wire) to a radiator, a water pipe, or a large steel object. I like to use an area of conductive foam to cover my workbench—the same type of foam that is used for packaging chips. However, this foam is quite expensive.

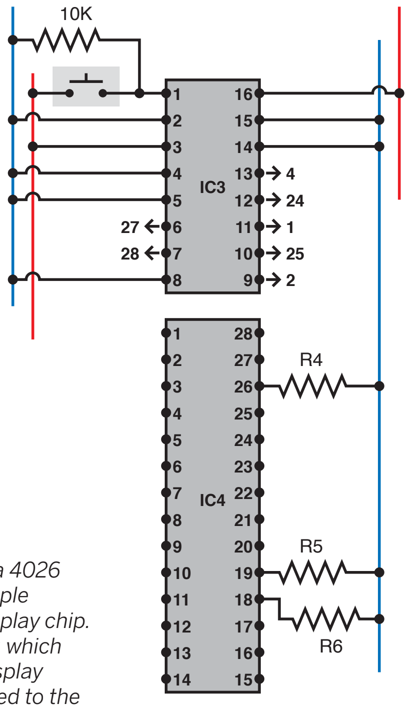
10K
IC3
1
2
3
4
5
6
7
8
16
15
14
13
12
11
10
9
27
28
4
24
1
25
2
IC4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
28
27
26
25
24
23
22
21
20
19
18
17
16
15
R4
R5
R6

Figure 4-35. *IC3 is a 4026 counter. IC4 is a triple seven-segment display chip. The arrows tell you which pins on the LED display should be connected to the pins on the counter.*

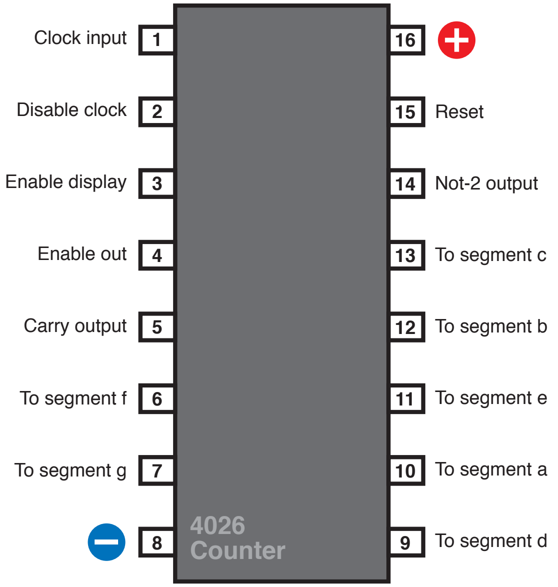
Clock input
Disable clock
Enable display
Enable out
Carry output
To segment f
To segment g
1
2
3
4
5
6
7
8
16
15
14
13
12
11
10
9
Reset
Not-2 output
To segment c
To segment b
To segment e
To segment a
To segment d
4026
Counter

Figure 4-36. *The 4026 decade counter is a CMOS chip that accepts clock pulses on pin 1, maintains a running total from 0 to 9, and outputs this total via pins designed to interface with a seven-segment LED numeric display.*

FUNDAMENTALS

Counters and seven-segment displays

Most counters accept a stream of pulses and distribute them to a series of pins in sequence. The 4026 decade counter is unusual in that it applies power to its output pins in a pattern that is just right to illuminate the segments of a 7-segment numeric display.

Some counters create positive outputs (they "source" current) while others create negative outputs (they "sink" current). Some seven-segment displays require positive input to light up the numbers. These are known as "common cathode" displays. Others require negative input and are known as "common anode" displays. The 4026 delivers positive outputs and requires a common cathode display.

Check the data sheet for any counter chip to find out how much power it requires, and how much it can deliver. CMOS chips are becoming dated, but they are very useful to hobbyists, because they will tolerate a wide range of supply voltage—from 5 to 15 volts in the case of the 4026. Other types of chips are much more limited.

Most counters can source or sink only a few milliamps of output power. When the 4026 is running on a 9-volt power supply, it can source about 4mA of power from each pin. This is barely enough to drive a seven-segment display.

You can insert a series resistor between each output pin of the counter and each input pin of the numeric display, but a simpler, quicker option is to use just one series resistor for each numeral, between the negative-power pin and ground. The experiment that I'm describing uses this shortcut. Its disadvantage is that digits that require only a couple of segments (such as numeral 1) will appear brighter than those that use many segments (such as numeral 8).

If you want your display to look bright and professional, you really need a transistor to drive each segment of each numeral. An alternative is to use a chip containing multiple "op amps" to amplify the current.

When a decade counter reaches 9 and rolls over to 0, it emits a pulse from its "carry" pin. This can drive another counter that will keep track of tens. The carry pin on that counter can be chained to a third counter that keeps track of hundreds, and so on. In addition to decade counters, there are hexadecimal counters (which count in 16s), octal counters (in 8s), and so on.

Why would you need to count in anything other than tens? Consider that the four numerals on a digital clock each count differently. The rightmost digit rolls over when it reaches 10. The next digit to the left counts in sixes. The first hours digit counts to 10, gives a carry signal, counts to 2, and gives another carry signal. The leftmost hours digit is either blank or 1, when displaying time in 12-hour format. Naturally there are counters specifically designed to do all this.

Counters have control pins such as "clock disable," which tells the counter to ignore its input pulses and freeze the display, "enable display," which enables the output from the chip, and "reset," which resets the count to zero.

The 4026 requires a positive input to activate each control pin. When the pins are grounded, their features are suppressed.

To make the 4026 count and display its running total you must ground the "clock disable" and "reset" pins (to suppress their function) and apply positive voltage to the "enable display" pin (to activate the output). See Figure 4-36 to see these pins identified.

Assuming that you succeed in getting your counter to drive the numeric display, you're ready to add two more counters, which will control the remaining two numerals. The first counter will count in ones, the second in tens, and the third in hundreds.

In Figure 4-37, I've continued to use arrows and numbers to tell you which pins of the counters should be connected to which pins of the numeric display. Otherwise, the schematic would be a confusing tangle of wires crossing each other.

At this point, you can give up in dismay at the number of connections—but really, using a breadboard, it shouldn't take you more than half an hour to complete this phase of the project. I suggest you give it a try, because there's

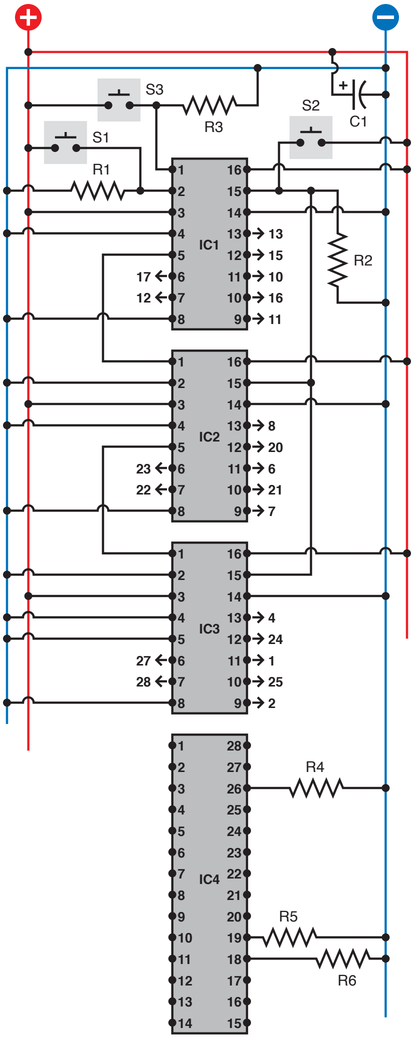

Figure 4-37. This test circuit, laid out as you would be likely to place it on a breadboard, allows you to trigger a counter manually to verify that the display increments from 000 upward to 999.

Component values:

All resistors are 1K.
S1, S2, S3: SPST tactile switches, normally open
IC1, IC2, IC3: 4026 decade counter chips
IC4: Kingbright 3-digit common-cathode display
C1: 100 µF (minimum) smoothing capacitor

Wire the output pins on IC1, IC2, and IC3 to the pins on IC4, according to the numbers preceded by arrows. The actual wires have been omitted for clarity. Check for the pinouts of IC4.

something magical about seeing a display count from 000 through 999 "all by itself," and I chose this project because it also has a lot of instructional value.

S1 is attached to the "clock disable" pin of IC1, so that when you hold down this button, it should stop that counter from counting. Because IC1 controls IC2, and IC2 controls IC3, if you freeze IC1, the other two will have to wait for it to resume. Therefore you won't need to make use of their "clock disable" features.

S2 is connected to the "reset" pins of all three counters, so that when you hold down this button, it should set them all to zero.

S3 sends positive pulses manually to the "clock input" pin of the first counter.

S1, S2, and S3 are all wired in parallel with 1K resistors connected to the negative side of the power supply. The idea is that when the buttons are not being pressed, the "pull-down" resistors keep the pins near ground (zero) voltage. When you press one of the buttons, it connects positive voltage directly to the chip, and easily overwhelms the negative voltage. This way, the pins remain either in a definitely positive or definitely negative state. If you disconnect one of these pull-down resistors you are likely to see the numeric display "flutter" erratically. (The numeric display chip has some unconnected pins, but this won't cause any problem, because it is a passive chip that is just a collection of LED segments.)

Always connect input pins of a CMOS chip so that they are either positive or negative. See the "No Floating Pins" warning on the next page.

I suggest that you connect all the wires shown in the schematic first. Then cut lengths of 22-gauge wire to join the remaining pins of the sockets from IC1, IC2, and IC3 to IC4.

Switch on the power and press S2. You'll see three zeros in your numeric display.

Each time you press S3, the count should advance by 1. If you press S2, the count should reset to three zeros. If you hold down S1 while you press S3 repeatedly, the counters should remain frozen, ignoring the pulses from S3.

FUNDAMENTALS

Switch bounce

When you hit S3, I think you'll find that the count sometimes increases by more than 1. This does not mean that there's something wrong with your circuit or your components; you are just observing a phenomenon known as "switch bounce."

On a microscopic level, the contacts inside a pushbutton switch do not close smoothly, firmly, and decisively. They vibrate for a few microseconds before settling; the counter chip detects this vibration as a series of pulses, not just one.

Various circuits are available to "debounce" a switch. The simplest option is to put a small capacitor in parallel with the switch, to absorb the fluctuations; but this is less than ideal. I'll come back to the topic of debouncing later in the book. Switch bounce is not a concern in this circuit, because we're about to get rid of S3 and substitute a 555 timer that generates nice clean bounceless pulses.

Pulse Generation

A 555 timer is ideal for driving a counter chip. You've already seen how to wire a 555 to create a stream of pulses that made noise through a loudspeaker. I'm reproducing the same circuit in Figure 4-38 in simplified form, using the positive and negative supply configuration in the current project. Also I'm showing the connection between pins 2 and 6 in the way that you're most likely to make it, via a wire that loops over the top of the chip.

For the current experiment, I'm suggesting initial component values that will generate only four pulses per second. Any faster than that, and you won't be able to verify that your counters are counting properly.

Install IC5 and its associated components on your breadboard immediately above IC1. Don't leave any gap between the chips. Disconnect S3 and R3 and connect a wire directly between pin 3 (output) of IC5 and pin 1 (clock) of IC1, the topmost counter. Power up again, and you should see the digits advancing rapidly in a smooth, regular fashion. Press S1, and while you hold it, the count should freeze. Release S1 and the count will resume. Press S2 and the counter should reset, even if you are pressing S1 at the same time.

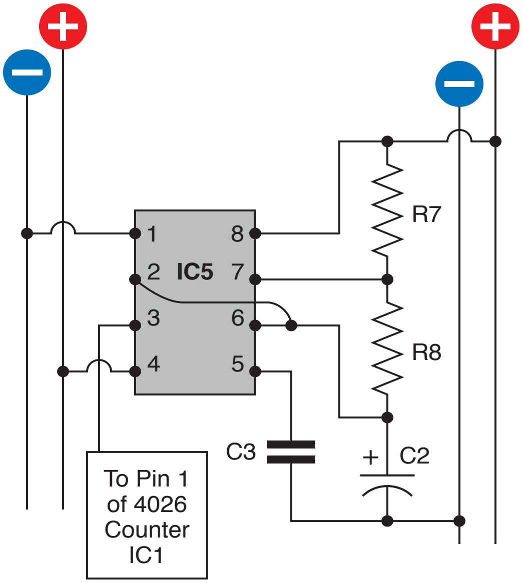

Figure 4-38. *A basic astable circuit to drive the decade counter in the previous schematic. Output is approximately 4 pulses per second.*

R7: 1K
R8: 2K2
C2: 68 μF
C3: 0.1 μF
IC5: 555 timer

No Floating Pins!

A CMOS chip is hypersensitive. Any pin that is not wired either to the supply voltage or to ground is said to be "floating" and may act like an antenna, sensitive to the smallest fluctuations in the world around it.

The 4026 counter chip has a pin labeled "clock disable." The manufacturer's data sheet helpfully tells you that if you give this pin a positive voltage, the chip stops counting and freezes its display. As you don't want to do that, you may just ignore that pin and leave it unconnected, at least while you test the chip. This is a very bad idea!

What the data sheet doesn't bother to tell you (presumably because "everyone knows" such things) is that if you want the clock to run normally, the clock-disable feature itself has to be disabled, by wiring it to negative (ground) voltage. If you leave the pin floating (and I speak from experience), the chip will behave erratically and uselessly.

All input pins must be either positively or negatively wired, unless otherwise specified.

Refinements

Now it's time to remember that what we really want this circuit to do is test a person's reflexes. When the user starts it, we want an initial delay, followed by a signal—probably an LED that comes on. The user responds to the signal by pressing a button as quickly as possible. During the time it takes for the person to respond, the counter will count milliseconds. When the person presses the button, the counter will stop. The display then remains frozen indefinitely, displaying the number of pulses that were counted before the person was able to react.

How to arrange this? I think we need a flip-flop. When the flip-flop gets a signal, it starts the counter running—and keeps it running. When the flip-flop gets another signal (from the user pressing a button), it stops the counter running, and keeps it stopped.

How do we build this flip-flop? Believe it or not, we can use yet another 555 timer, in a new manner known as bistable mode.

FUNDAMENTALS

The bistable 555 timer

Figure 4-39 shows the internal layout of a 555 timer, as before, but the external components on the righthand side have been eliminated. Instead, I'm applying a constant negative voltage to pin 6. Can you see the consequences? Suppose you apply a negative pulse to the trigger (pin 2). Normally when you do this and the 555 starts running, it generates a positive output while charging a capacitor attached to pin 6. When the capacitor reaches 2/3 of the full supply voltage, this tells the 555 to ends its positive output, and it flips back to negative.

Well, if there's no capacitor, there's nothing to stop the timer. Its positive output will just continue indefinitely. However, pin 4 (the reset) can still override everything, so if you apply negative voltage to pin 4, it flips the output to negative. After that, the output will stay negative indefinitely, as it usually does, until you trigger the timer by dropping the voltage to pin 2 again. This will flip the timer back to generating its positive output.

Here's a quick summary of the bistable configuration:

- A negative pulse to pin 2 turns the output positive.
- A negative pulse to pin 4 turns the output negative.
- The timer is stable in each of these states. Its run-time has become infinite.

It's OK to leave pins 5 and 7 of the timer unconnected, because we're pushing it into extreme states where any random signals from those pins will be ignored.

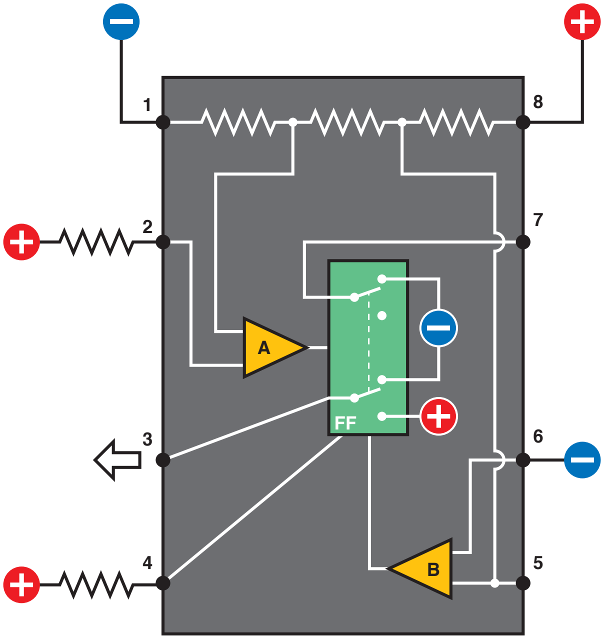

Figure 4-39. *In the bistable configuration, pin 6 of a 555 timer is perpetually negative, so the timer cycle never ends, unless you force it to do so by applying a negative pulse to pin 4 (the reset).*

In bistable mode, the 555 has turned into one big flip-flop. To avoid any uncertainty, we keep pins 2 and 4 normally positive via pull-up resistors, but negative pulses on those pins can overwhelm them when we want to flip the 555 into its opposite state. The schematic for running a 555 timer in bistable mode, controlled by two pushbuttons, is shown in Figure 4-40. You can add this above your existing circuit. Because you're going to attach the output from IC6 to pin 2 of IC1, the topmost counter, you can disconnect S1 and R1 from that pin. See Figure 4-41.

Now, power up the circuit again. You should find that it counts in the same way as before, but when you press S4, it freezes. This is because your bistable 555 timer is sending its positive output to the "clock disable" pin on the counter. The counter is still receiving a stream of pulses from the astable 555 timer, but as long as pin 2 is positive on the counter, the counter simply ignores the pulses.

Now press S5, which flips your bistable 555 back to delivering a negative output, at which point the count resumes.

We're getting close to a final working circuit here. We can reset the count to zero (with S3), start the count (with S5), and wait for the user to stop the count (with S4). The only thing missing is a way to start the count unexpectedly.

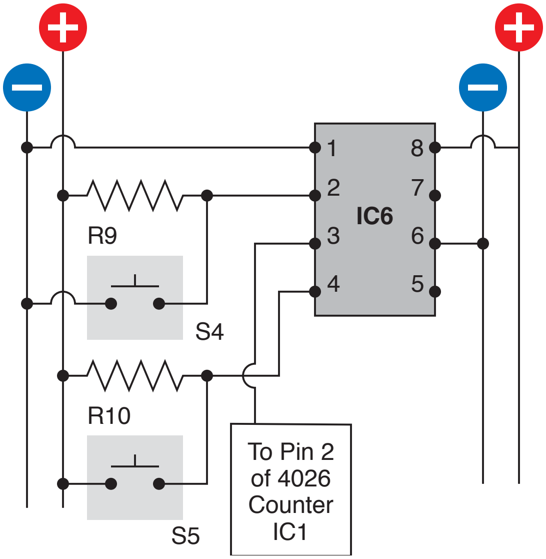

Figure 4-40. *Adding a bistable 555 timer to the reflex tester will stop the counter with a touch of a button, and keep it stopped.*

R9, R10: 1K
IC6: 555 timer

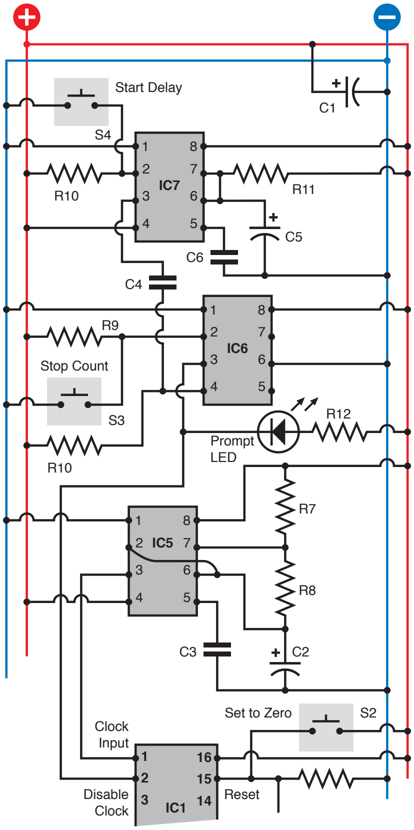

Figure 4-41. *The completed control section of the circuit, to be added above these timers.*

R7, R9, R10, R12: 1K
R8: 2K2
R11: 330K
C1: 100 µF
C2: 68 µF
C3, C4, C6: 0.1 µF
C5: 10 µF
S1, S2, S3: tactile switches
IC5, IC6, IC7: 555 timers

The Delay

Suppose we set up yet another 555 in monostable mode. Trigger its pin 2 with a negative pulse, and the timer delivers a positive output that lasts for, say, 4 seconds. At the end of that time, its output goes back to being negative. Maybe we can hook that positive-to-negative transition to pin 4 of IC6. We can use this instead of switch S5, which you were pressing previously to start the count.

Check the new schematic in Figure 4-41 which adds another 555 timer, IC7 above IC6. When the output from IC7 goes from positive to negative, it will trigger the reset of IC6, flipping its output negative, which allows the count to begin. So IC7 has taken the place of the start switch, S4. You can get rid of S4, but keep the pull-up resistor, R9, so that the reset of IC6 remains positive the rest of the time.

This arrangement works because I have used a capacitor, C4, to connect the output of IC7 to the reset of IC6. The capacitor communicates the sudden change from positive to negative, but the rest of the time it blocks the steady voltage from IC7 so that it won't interfere with IC6.

The final schematic in Figure 4-41 shows the three 555 timers all linked together, as you should insert them above the topmost counter, IC1. I also added an LED to signal the user. Figure 4-42 is a photograph of my working model of the circuit.

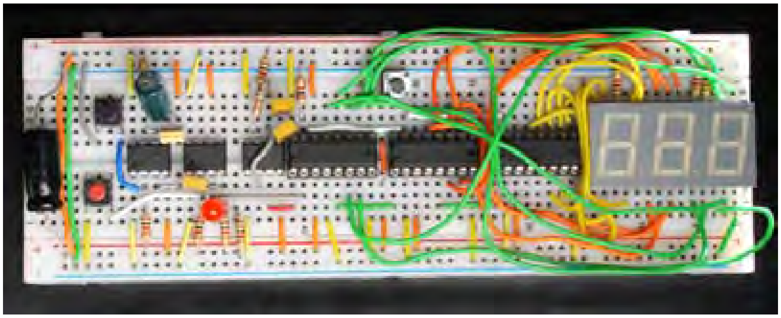

Figure 4-42. *The complete reaction-timer circuit barely fits on a 63-row breadboard.*

Because this circuit is complicated, I'll summarize the sequence of events when it's working. Refer to Figure 4-41 while following these steps:

1. User presses Start Delay button S4, which triggers IC7.
2. IC7 output goes high for a few seconds while C5 charges.
3. IC7 output drops back low.
4. IC7 communicates a pulse of low voltage through C4 to IC6, pin 4.
5. IC 6 output flips to low and flops there.
6. Low output from IC6 sinks current through LED and lights it.
7. Low output from IC6 also goes to pin 2 of IC1.

8. Low voltage on pin 2 of IC1 allows IC1 to start counting.
9. User presses S3, the "stop" button.
10. S3 connects pin 2 of IC6 to ground.
11. IC6 output flips to high and flops there.
12. High output from IC6 turns off the LED.
13. High output from IC6 also goes to pin 2 of IC1.
14. High voltage on pin 2 of IC1 stops it from counting.
15. After assessing the result, user presses S2.
16. S2 applies positive voltage to pin 15 of IC1, IC2, IC3.
17. Positive voltage resets counters to zero.
18. The user can now try again.
19. Meanwhile, IC5 is running continuously throughout.

In case you find a block diagram easier to understand, I've included that, too, in Figure 4-43.

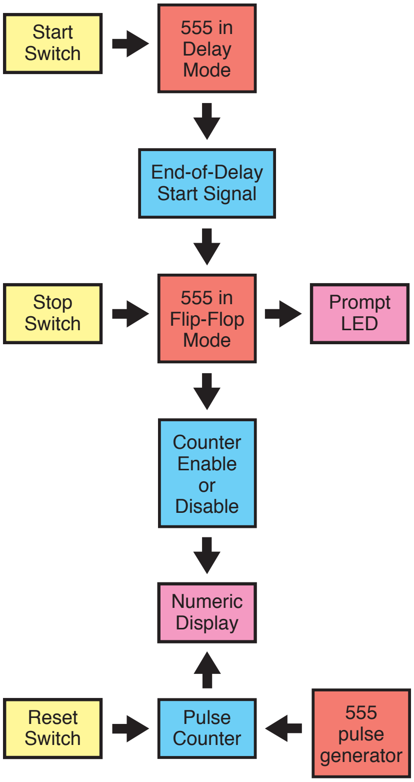

Figure 4-43. *The functions of the reflex tester, summarized as a block diagram.*

Using the Reflex Tester

At this point, you should be able to test the circuit fully. When you first switch it on, it will start counting, which is slightly annoying, but easily fixed. Press S3 to stop the count. Press S2 to reset to zero.

Now press S4. Nothing seems to happen—but that's the whole idea. The delay cycle has begun in stealth mode. After a few seconds, the delay cycle ends, and the LED lights up. Simultaneously, the count begins. As quickly as possible, the user presses S3 to stop the count. The numerals freeze, showing how much time elapsed.

There's only one problem—the system has not yet been calibrated. It is still running in slow-motion mode. You need to change the resistor and capacitor attached to IC5 to make it generate 1,000 pulses per second instead of just three or four.

Substitute a 10K trimmer potentiometer for R8 and a 1 μF capacitor for C2. This combination will generate about 690 pulses per second when the trimmer is presenting maximum resistance. When you turn the trimmer down to decrease its resistance, somewhere around its halfway mark the timer will be running at 1,000 pulses per second.

How will you know exactly where this point is? Ideally, you would attach an oscilloscope probe to the output from IC5. But, most likely, you don't have an oscilloscope, so here are a couple other suggestions.

First remove the 1 µF capacitor at C2 and substitute a 10 µF capacitor. Because you are multiplying the capacitance by 10, you will reduce the speed by 10. The leftmost digit in your display should now count in seconds, reaching 9 and rolling over to 0 every 10 seconds. You can adjust your trimmer potentiometer while timing the display with a stopwatch. When you have it right, remove the 10 µF capacitor and replace the 1 µF capacitor at C2.

The only problem is, the values of capacitors may be off by as much as 10%. If you want to fine-tune your reflex timer, you can proceed as follows.

Disconnect the wire from pin 5 of IC3, and substitute an LED with a 1K series resistor between pin 5 and ground. Pin 5 is the "carry" pin, which will emit a positive pulse whenever IC3 counts up to 9 and rolls over to start at 0 again. Because IC3 is counting tenths of a second, you want its carry output to occur once per second.

Now run the circuit for a full minute, using your stopwatch to see if the flashing LED drifts gradually faster or slower than once per second. If you have a camcorder that has a time display in its viewfinder, you can use that to observe the LED.

If the LED flashes too briefly to be easily visible, you can run a wire from pin 5 to another 555 timer that is set up in monostable mode to create an output lasting for around 1/10 of a second. The output from that timer can drive an LED.

Enhancements

It goes without saying that anytime you finish a project, you see some opportunities to improve it. Here are some suggestions:

1. No counting at power-up. It would be nice if the circuit begins in its "ready" state, rather than already counting. To achieve this you need to send a negative pulse to pin 2 of IC6, and maybe a positive pulse to pin 15 of IC1. Maybe an extra 555 timer could do this. I'm going to leave you to experiment with it.
2. Audible feedback when pressing the Start button. Currently, there's no confirmation that the Start button has done anything. All you need to do is buy a piezoelectric beeper and wire it between the righthand side of the Start button and the positive side of the power supply.
3. A random delay interval before the count begins. Making electronic components behave randomly is very difficult, but one way to do it would be to require the user to hold his finger on a couple of metal contacts. The skin resistance of the finger would substitute for R11. Because the finger pressure would not be exactly the same each time, the delay would vary. You'd have to adjust the value of C5.

Summing Up

This project demonstrated how a counter chip can be controlled, how counter chips can be chained together, and three different functions for 555 timers. It also showed you how chips can communicate with each other, and introduced you to the business of calibrating a circuit after you've finished building it.

Naturally, if you want to get some practical use from the circuit, you should build it into an enclosure with heavier-duty pushbuttons—especially the button that stops the count. You'll find that when people's reflexes are being tested, they are liable to hit the stop button quite hard.

Because this was a major project, I'll follow it up here with some quicker, easier ones as we move into the fascinating world of another kind of integrated circuit: logic chips.

Experiment 19: Learning Logic

You will need:

- Assorted resistors and capacitors.
- 74HC00 quad 2-input NAND chip, 74HC08 quad 2-input AND chip, and LM7805 voltage regulator. Quantity: 1 of each.
- Signal diode, 1N4148 or similar. Quantity: 1.
- Low-current LED. Quantity: 1.
- SPST tactile switches. Quantity: 2.

You're going to be entering the realm of pure digital electronics, using "logic gates" that are fundamental in every electronic computing device. When you deal with them individually, they're extremely easy to understand. When you start chaining them together, they can seem intimidatingly complex. So let's start with them one at a time.

Logic gates are much fussier than the 555 timer or the 4026 counter that you used previously. They demand an absolutely precise 5 volts DC, with no fluctuations or "spikes" in the flow of current. Fortunately, this is easy to achieve: just set up your breadboard with an LM7805 voltage regulator, as shown in the schematic in Figure 4-44 and the photograph in Figure 4-45. The regulator receives 9 volts from your usual voltage supply, and reduces it to 5 volts, with the help of a couple of capacitors. You apply the 9 volts to the regulator, and distribute the 5 volts down the sides of your breadboard instead of the unregulated voltage that you used previously. Use your meter to verify the voltage, and make sure you have the polarity clearly marked.

After installing your regulator, take a couple of tactile switches, two 10K resistors, a low-current LED, and a 1K resistor, and set them around a 74HC00 logic chip as shown in Figure 4-46. You may notice that many of the pins of the chip are shorted together and connected to the negative side of the power supply. I'll explain that in a moment.

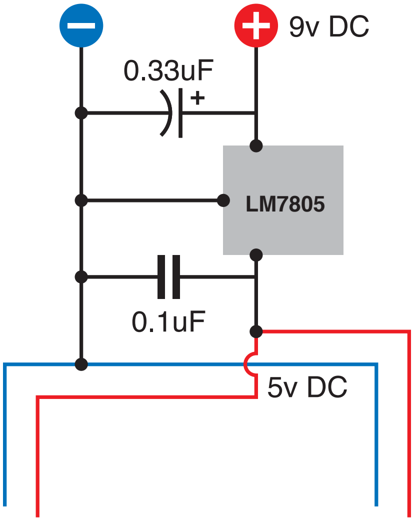

Figure 4-44. *This simple circuit is essential to provide a regulated 5V DC supply for logic chips.*

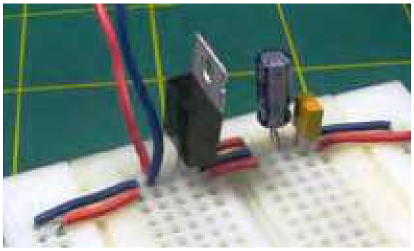

Figure 4-45. *The voltage regulator and its two capacitors can fit snugly at the top of a breadboard. Remember to apply the 9V input voltage at the left pin of the regulator, and distribute the 5V output down the sides of the breadboard.*

FUNDAMENTALS

Voltage regulators

The simplest versions of these little semiconductors accept a higher DC voltage on one pin and deliver a lower DC voltage on another pin, with a third pin (usually in the middle) serving as a common negative, or ground. You should also attach a couple of capacitors to smooth the current, as shown in Figure 4-46.

Typically you can put a 7.5-volt or 9-volt supply on the "input" side of a 5-volt regular, and draw a precise 5 volts from the "output" side. If you're wondering where the extra voltage goes, the answer is, the regulator turns the electricity into heat. For this reason, small regulators (such as the one in Figure 4-8) often have a metal back with a hole in the top. Its purpose is to radiate heat, which it will do more effectively if you bolt it to a piece of aluminum, since aluminum conducts heat very effectively. The aluminum is known as a heat sink, and you can buy fancy ones that have multiple cooling fins.

For our purposes, we won't be drawing enough current to require a heat sink.

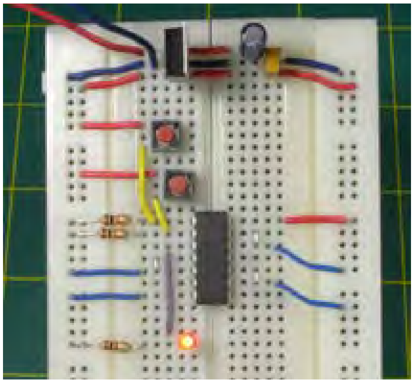

Figure 4-47. *This breadboard layout is exactly equivalent to the schematic in Figure 4-46.*

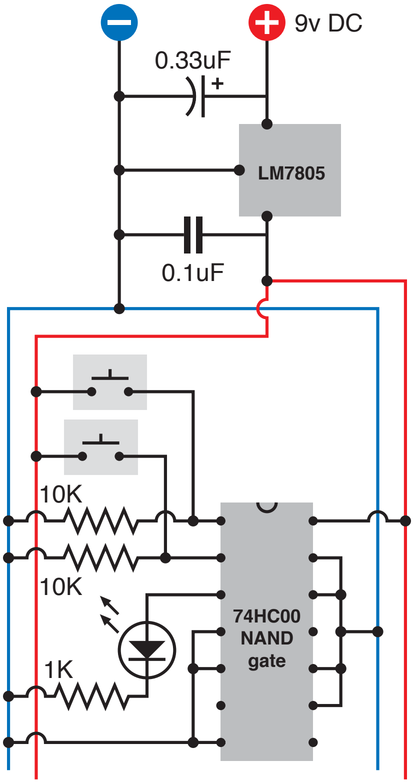

Figure 4-46. *By observing the LED when you press either, both, or neither of the buttons, you can easily figure out the logical function of the NAND gate.*

When you connect power, the LED should light up. Press one of the tactile switches, and the LED remains illuminated. Press the other tactile switch, and again the LED stays on. Now press both switches, and the light should go out.

Pins 1 and 2 are logic inputs for the 74HC00 chip. Initially they were held at negative voltage, being connected to the negative side of the power supply through 10K resistors. But each pushbutton overrides its pull-down resistor and forces the input pin to go positive.

The logic output from the chip, as you saw, is normally positive—but *not* if the first input *and* the second input are positive. Because the chip does a "Not AND" operation, it's known as a NAND logic gate. You can see the breadboard layout in Figure 4-47. Figure 4-48 is a simplified version of the circuit. The U-shaped thing with a circle at the bottom is the logic symbol for a NAND gate. No power supply is shown for it, but in fact all logic chips require a power supply, which enables them to put out more current than they take in. Anytime you see a symbol for a logic chip, try to remember that it has to have power to function.

The 74HC00 actually contains four NAND gates, each with two logical inputs and one output. They are arrayed as shown in Figure 4-49. Because only one gate was needed for the simple test, the input pins of the unused gates were shorted to the negative side of the power supply.

Pin 14 supplies positive power for the chip; pin 7 is its ground pin. Almost all the 7400 family of logic chips use the same pins for positive and negative power, so you can swap them easily.

In fact, let's do that right now. First, disconnect the power. Carefully pull out the 74HC00 and put it away with its legs embedded in conductive foam. Substitute a 74HC08 chip, which is an AND chip. Make sure you have it the right way up, with its notch at the top. Reconnect the power and use the pushbuttons as you did before. This time, you should find that the LED comes on if the first input AND the second input are both positive, but it remains dark otherwise. Thus, the AND chip functions exactly opposite to the NAND chip. Its pinouts are shown in Figure 4-50.

You may be wondering why these things are useful. Soon you'll see that we can put logic gates together to do things such as create an electronic combination lock, or a pair of electronic dice, or a computerized version of a TV quiz show where users compete to answer a question. And if you were really insanely ambitious, you could build an entire computer out of logic gates.

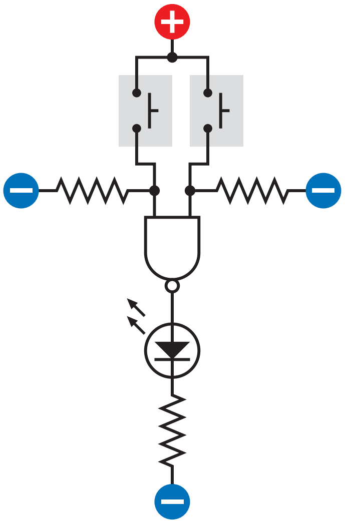

Figure 4-48. *The structure and function of the NAND gate is easier to visualize with this simplified schematic that omits the power supply for the chip and doesn't attempt to place the wires to fit a breadboard layout.*

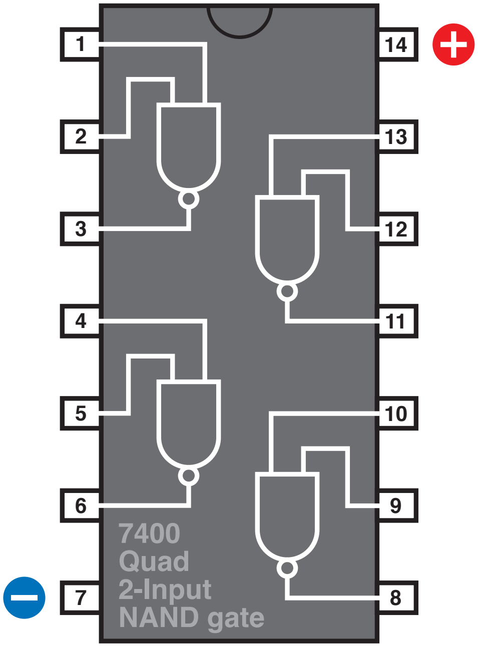

Figure 4-49. *The pinouts of the logic gates in a 74HC00chip.*

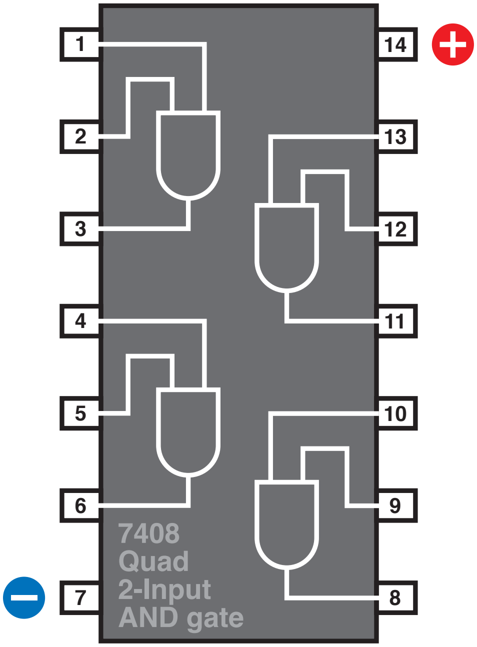

Figure 4-50. *The pinouts of the logic gates in a 74HC08chip.*

BACKGROUND

From Boole to Shannon

George Boole was a British mathematician, born in 1815, who did something that few people are ever lucky enough or smart enough to do: he invented an entirely new branch of mathematics.

Interestingly, it was not based on numbers. Boole had a relentlessly logical mind, and he wanted to reduce the world to a series of true-or-false statements which could overlap in interesting ways. For instance, suppose there is a couple named Ann and Bob who have so little money, they only own one hat. Clearly, if you happen to run into Ann and Bob walking down the street, there are four possibilities: neither of them may be wearing a hat, Ann may be wearing it, or Bob may be wearing it, but they cannot both be wearing it.

The diagram in Figure 4-51 illustrates this. All the states are possible except the one where the circles overlap. (This is known as a Venn diagram. I leave it to you to search for this term if it interests you and you'd like to learn more.) Boole took this concept much further, and showed how to create and simplify extremely complex arrays of logic.

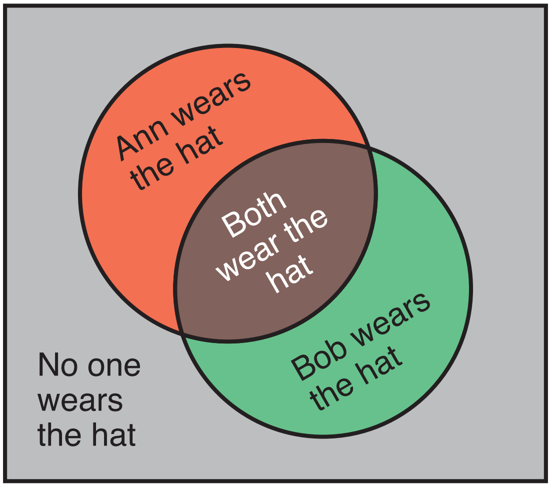

Figure 4-51. *This slightly frivolous Venn diagram illustrates the various possibilities for two people, Ann and Bob, who own only one hat.*

Another way to summarize the hat-wearing situation is to make the "truth table" shown in Figure 4-52. The rightmost column shows whether each combination of propositions can be true. Now check the table in Figure 4-53. It's the same table but uses different labels, which describe the pattern you have seen while using the NAND gate.

Boole published his treatise on logic in 1854, long before it could be applied to electrical or electronic devices. In fact, during his lifetime, his work seemed to have no practical applications at all. But a man named Claude Shannon encountered Boolean logic while studying at MIT in the 1930s, and in 1938 he published a paper describing how Boolean analysis could be applied to circuits using relays. This had immediate practical applications, as telephone networks were growing rapidly, creating complicated switching problems.

BACKGROUND

From Boole to Shannon (continued)

Ann is wearing the hat	Bob is wearing the hat	This combination can be
NO	NO	TRUE
NO	YES	TRUE
YES	NO	TRUE
YES	YES	FALSE

Figure 4-52. *The hat-wearing possibilities can be expressed in a "truth table."*

Input A	Input B	Output
OFF	OFF	ON
OFF	ON	ON
ON	OFF	ON
ON	ON	OFF

Figure 4-53. *The truth-table from can be relabeled to describe the inputs and outputs of a NAND gate.*

A very simple telephone problem could be expressed like this. Suppose two customers in a rural area share one telephone line. If one of them wants to use the line, or the other wants to use it, or neither of them wants to use it, there's no problem. But they cannot both use it at once. You may notice that this is exactly the same as the hat-wearing situation for Ann and Bob.

We can easily draw a circuit using two normally closed relays that creates the desired outcome (see Figure 4-54), but if you imagine a telephone exchange serving many thousands of customers, the situation becomes very complicated indeed. In fact, in Shannon's time, no logical process existed to find the best solution and verify that it used fewer components than some other solution.

Shannon saw that Boolean analysis could be used for this purpose. Also, if you used an "on" condition to represent numeral 1 and an "off" condition to represent numeral 0, you could build a system of relays that could count. And if it could count, it could do arithmetic.

When vacuum tubes were substituted for relays, the first practical digital computers were built. Transistors took the place of vacuum tubes, and integrated circuit chips replaced transistors, leading to the desktop computers that we now take for granted today. But deep down, at the lowest levels of these incredibly complex devices, they still use the laws of logic discovered by George Boole. Today, when you use a search engine online, if you use the words AND and OR to refine your search, you're actually using Boolean operators.

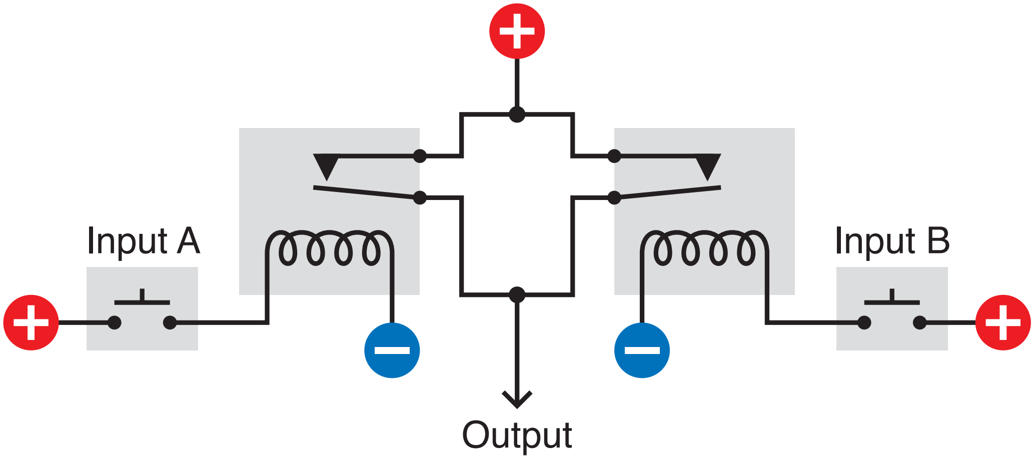

Figure 4-54. *This relay circuit could illustrate the desired logic for two telephone customers wanting to share one line, and its behavior is almost identical to that of the NAND schematic shown in Figure 4-48.*

Figure 4-55. *Ann and Bob attempt to overcome the limitations of Boolean logic.*

ESSENTIALS

Logic gate basics

The NAND gate is the most fundamental building block of digital computers, because (for reasons which I don't have space to explain here) it enables digital addition. If you want to explore more try searching online for topics such as "binary arithmetic" and "half-adder."

Generally, there are seven types of logic gates:

- AND
- NAND
- OR
- NOR
- XOR
- XNOR
- NOT

Of the six two-input gates, the XNOR is hardly ever used. The NOT gate has a single input, and simply gives a negative output when the input is positive or a positive output when the input is negative. The NOT is more often referred to as an "Inverter." The symbols for all seven gates are shown in Figure 4-56.

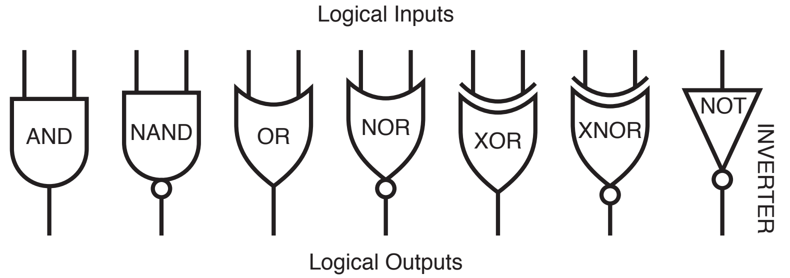

Figure 4-56. *American symbols for the six types of two-input logic gates, and the single-input inverter.*

I've shown the American symbols. Other symbols have been adopted in Europe, but the traditional symbols shown here are the ones that you will usually find, even being used by Europeans. I also show the truth tables, in Figure 4-57, illustrating the logical output (high or low) for each pair of inputs of each type of gate.

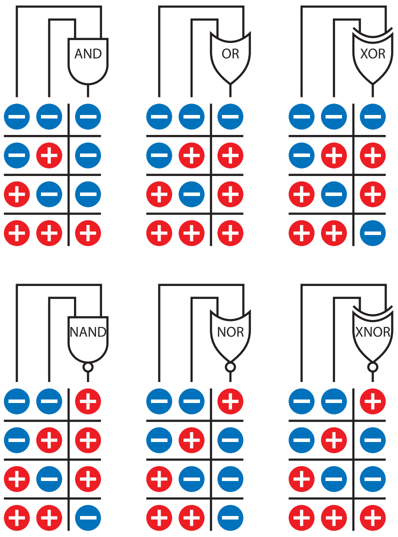

		AND	OR	XOR	NAND	NOR	XNOR
−	−	−	−	−	+	+	+
−	+	−	+	+	+	−	−
+	−	−	+	+	+	−	−
+	+	+	+	−	−	−	+

Figure 4-57. *Inputs and corresponding outputs for the six types of logic gates (note that the XNOR gate is seldom used). The minus signs indicate low voltage, close to ground potential. The plus signs indicate higher voltage, close to the positive potential of the power supply in the circuit. The exact voltages will vary depending on other components that may be actively connected.*

ESSENTIALS

Logic gate basics (continued)

If you have difficulty visualizing logic gates, a mechanical comparison may help. You can think of them as being like sliding plates with holes in them, in a bubblegum machine. Two people, A and B, can push the plates. These people are the two inputs, which are considered positive when they are doing something. (Negative logic systems also exist, but are uncommon, so I'm only going to talk about positive systems here.)

The flow of bubblegum represents a flow of positive current. The full set of possibilities is shown in Figures 4-58 through 4-63.

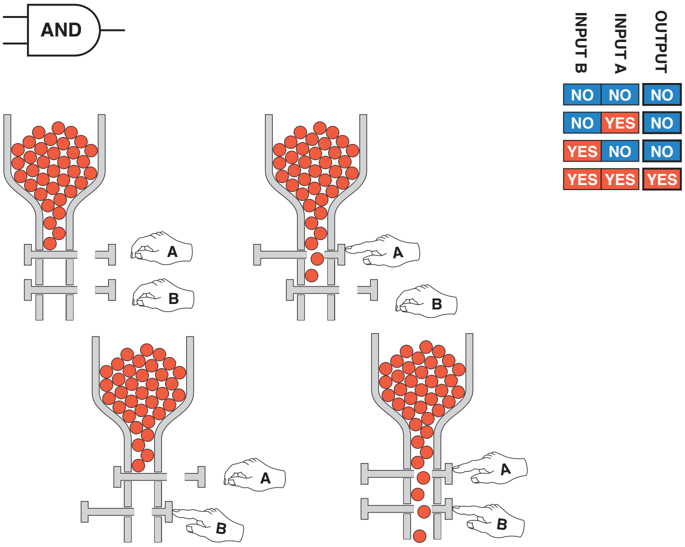

INPUT B	INPUT A	OUTPUT
NO	NO	NO
NO	YES	NO
YES	NO	NO
YES	YES	YES

Figure 4-58

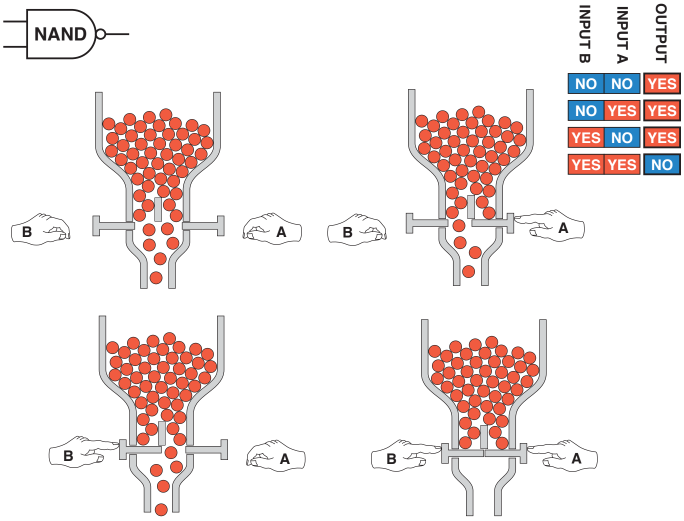

INPUT B	INPUT A	OUTPUT
NO	NO	YES
NO	YES	YES
YES	NO	YES
YES	YES	NO

Figure 4-59

ESSENTIALS

Logic gate basics (continued)

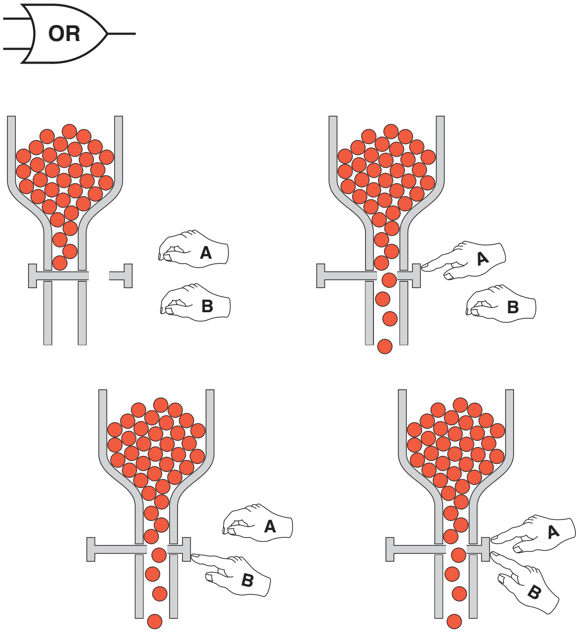
OR
A
B
A
B
A
B
A
B

INPUT B	INPUT A	OUTPUT
NO	NO	NO
NO	YES	YES
YES	NO	YES
YES	YES	YES

Figure 4-80

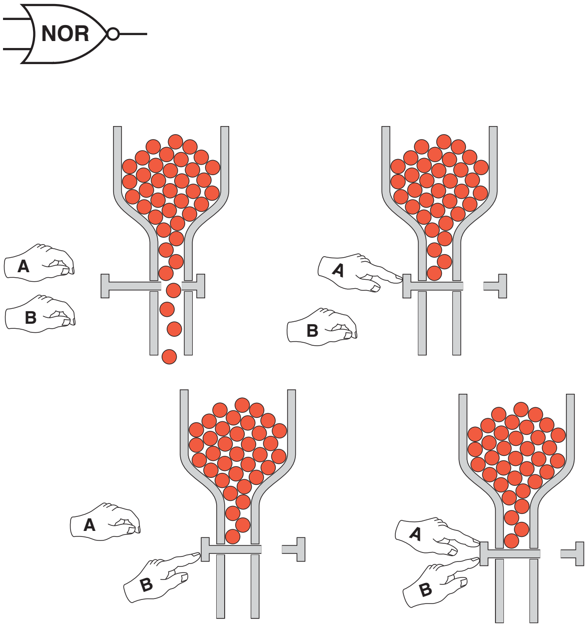
NOR
A
B
A
B
A
B
A
B

INPUT B	INPUT A	OUTPUT
NO	NO	YES
NO	YES	NO
YES	NO	NO
YES	YES	NO

Figure 4-81

ESSENTIALS

Logic gate basics (continued)

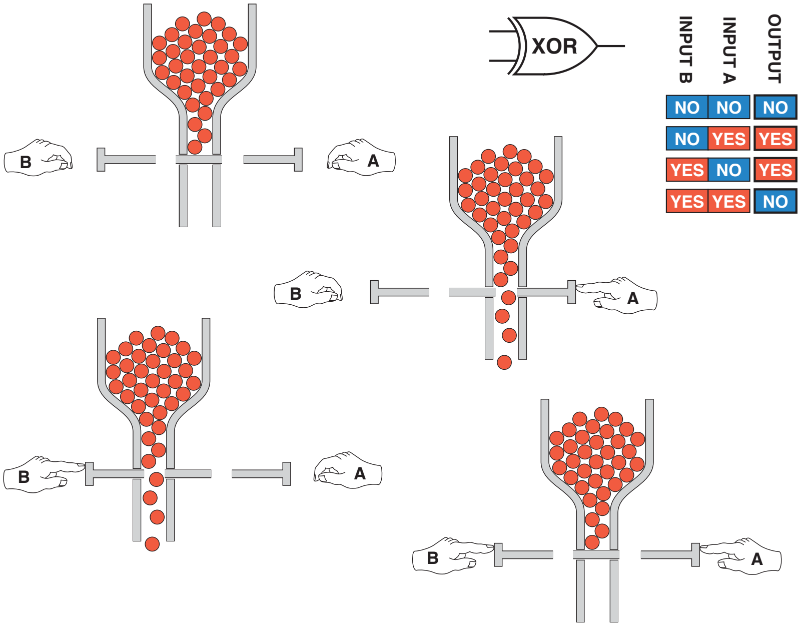

INPUT B	INPUT A	OUTPUT
NO	NO	NO
NO	YES	YES
YES	NO	YES
YES	YES	NO

Figure 4-82

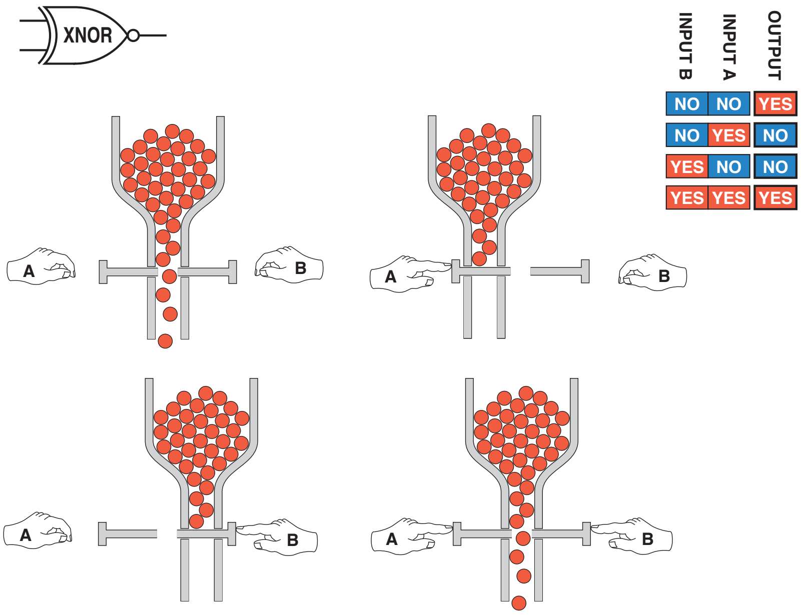

INPUT B	INPUT A	OUTPUT
NO	NO	YES
NO	YES	NO
YES	NO	NO
YES	YES	YES

Figure 4-83

BACKGROUND

The confusing world of TTL and CMOS

Back in the 1960s, the first logic gates were built with Transistor-Transistor Logic, abbreviated TTL, meaning that tiny bipolar transistors were etched into a single wafer of silicon. Soon, these were followed by Complementary Metal Oxide Semiconductors, abbreviated CMOS. Each of these chips was a collection of metal-oxide field-effect transistors (known as MOSFETs). The 4026 chip that you used earlier is an old CMOS.

You may remember that bipolar transistors amplify current. TTL circuits are similar: they are sensitive to current, rather than voltage. Thus they require a significant flow of electricity, to function. But CMOS chips are like the programmable unijunction transistor that I featured previously. They are voltage-sensitive, enabling them to draw hardly any current while they are waiting for a signal, or pausing after emitting a signal.

The two families named TTL and CMOS still exist today. The table in Figure 4-64 summarizes their basic advantages and disadvantages. The CMOS series, with part numbers from 4000 upward, were easily damaged by static electricity but were valuable because of their meager power consumption. The TTL series, with part numbers from 7400 upward, used much more power but were less sensitive and very fast. So, if you wanted to build a computer, you used the TTL family, but if you wanted to build a little gizmo that would run for weeks on a small battery, you used the CMOS family.

From this point on everything became extremely confusing, because CMOS manufacturers wanted to grab market share by emulating the advantages of TTL chips. Newer generations of CMOS chips even changed their part numbers to begin with "74" to emphasize their compatibility, and the functions of pins on CMOS chips were swapped around to match the functions of pins on TTL chips. Consequently, the pinouts of CMOS and TTL chips are usually now identical, but the meaning of "high" and "low" states changed in each new generation, and the maximum supply voltages for CMOS chips were revised downward. Note I have included question marks beside two categories in the CMOS column, as modern CMOS chips have overcome those disadvantages—at least to some extent.

Here's a quick summary, which will be useful to you if you look at a circuit that you find online, and you wonder about the chips that have been specified.

Where you see a letter "x," it means that various numbers may appear in that location. Thus "74xx" includes the 7400 NAND gate, the 7402 NOR gate, the 74150 16-bit data selector, and so on. A combination of letters preceding the "74" identifies the chip manufacturer, while letters following the part number may identify the style of package, may indicate whether it contains heavy metals that are environmentally toxic, and other details.

	TTL	CMOS
Part number series	7400	4000 (Later adopted 7400 numbering)
Vulnerable to static electricity	Less	More?
Speed	Faster	Slower?
Power Consumption	Higher	Very Low
Power Supply Range	Narrow 5v	Wider 3v-6v
Input Impedance	Low	Very High

Figure 4-64. *The basic differences between the two families of logic chips. In successive generations, these differences have gradually diminished.*

TTL family:

74xx
The old original generation, now obsolete.

74Sxx
Higher speed "Schottky" series, now obsolete.

74LSxx
Lower power Schottky series, still used occasionally.

74ALSxx
Advanced low-power Schottky.

74Fxx
Faster than the ALS series.

BACKGROUND

The confusing world of TTL and CMOS (continued)

CMOS family:

40xx
The old original generation, now obsolete.

40xxB
The 4000B series was improved but still susceptible to damage from static electricity. Many hobby circuits still use these chips because they will run from relatively high voltages and can power LEDs and even small relays directly.

74HCxx
Higher-speed CMOS, with part numbers matching the TTL family, and pinouts matching the TTL family, but input and output voltages not quite the same as the TTL family. I've used this generation extensively in this book, because it's widely available, and the circuits here have no need for greater speed or power.

74HCTxx
Like the HC series but matching the TTL voltages.

74ACxx
Advanced version of HC series. Faster, with higher output capacity.

74ACTxx
Like the AC series but with the same pin functions and voltages as TTL.

74AHCxx
Advanced higher-speed CMOS.

74AHCTxx
Like the AHC series but with the same pin functions and voltages as TTL.

74LVxx
Lower voltage (3.3v) versions, including LV, LVC, LVT, and ALVC series.

As you can see, these days you have to interpret part numbers very carefully. But which family and generation of chips should you use? Well, that depends! Following are some guidelines.

What you don't need:

1. Speed differences are irrelevant from our point of view, as we're not going to be building circuits running at millions of cycles per second.
2. Price differences are so small as to be inconsequential.
3. Lower-voltage (LV) CMOS chips are not very interesting for our small experimental circuits.
4. Try to avoid mixing different families, and different generations of the same family, in the same circuit. They may not be compatible.
5. Some modern chip varieties may be only available in the surface-mount package format. Because they're so much more difficult to deal with, and their only major advantage is miniaturization, I don't recommend them.
6. In the TTL family, the LS and ALS series cannot handle as much output current as the S series and the F series. You don't need them.

What you should use:

1. The old 74LSxx series of TTL chips was so popular, you'll still find schematics that specify these chips. You should still be able to buy them from sources online, but if not, you can substitute the 74HCTxx chips, which are designed to function identically.
2. The old 4000B series of CMOS chips are still used by hobbyists because their willingness to tolerate high voltages is convenient. While TTL or TTL-compatible chips require a carefully regulated 5 volts, the 4000B chips would handle 15 volts—and also delivered enough power to energize LEDs or even very small relays. Some hobbyists also have a nostalgic affection for the 74Cxx series of chips, which had the same pin connections as the TTL chips but could still tolerate higher voltages and higher output current. The trouble is, some of the 74Cxx chips are almost extinct, and while the 4000B chips are still available, they are considered almost obsolete.

BACKGROUND

The confusing world of TTL and CMOS (continued)

Bottom line: I suggest you use the 4000B chips only if you want to replicate an old circuit, or if a modern equivalent is unavailable (which is why I specified the 4026B chip for the reaction timer—I could not find a modern equivalent that will drive seven-segment numeric displays directly, and I didn't want you to have to deal with more parts than necessary).

If you check online suppliers such as *Mouser.com* you'll find that the HC family is by far the most popular right now. They are all available in through-hole format (to fit your breadboard and perforated board). They have the high input impedance of CMOS (which is useful) and they have the same pin identities as the old 74LSxx series.

Abbreviations

When looking at data sheets, you are likely to encounter some or all of these abbreviations:

- VOH min: Minimum output voltage in high state
- VOL max: Maximum output voltage in low state
- VIH min: Minimum input voltage to be recognized as high
- VIL max: Maximum input voltage recognized as low

BACKGROUND

Logic gate origins

The 7400 family of integrated circuits was introduced by Texas Instruments, beginning with the 7400 NAND gate in 1962. Other companies had sold logic chips previously, but the 7400 series came to dominate the market. The Apollo lunar missions used a computer built with 7400 chips, and they were a mainstay of minicomputers during the 1970s.

RCA introduced its 4000 series of logic chips in 1968, built around CMOS transistors; Texas Instruments had chosen TTL. The CMOS chips used less power, thus generating much less heat and enabling flexible circuit design, as each chip could power many others. CMOS was also tolerant of wide voltage ranges (from 3 to 15 volts) but prohibited switching speeds faster than around 1MHz. TTL was 10 times faster.

Design tweaks gradually eradicated the speed penalty for CMOS, and TTL chips have become relatively rare. Still, some people retain a special nostalgic loyalty to "the logic gates that went to the moon." A hardcore enthusiast named Bill Buzbee has built an entire web server from TTL-type 7400 chips, currently online at *http://magic-1.org*. Figure 4-65 shows just one of the handmade circuit boards that Bill assembled to run his computer.

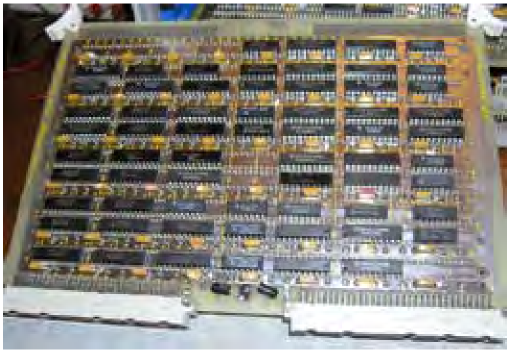

Figure 4-65. *Hobbyist Bill Buzbee built himself a web server entirely from 7400 series logic chips, the oldest of which was fabricated back in 1969. The web server can be found online at http://magic-1.org, displaying pictures of itself and details of its construction. The picture here that Bill took shows just one of the circuit boards of this remarkable project.*

FUNDAMENTALS

Common part numbers

Each 14-pin chip can contain four 2-input gates, three 3-input gates, two 4-input gates, one 8-input gate, or six single-input inverters, as shown in the following table.

	2 input	3 input	4 input	8 input
AND	7408	7411	7421	
NAND	7400	7410	7420	7430
OR	7432			744078*
NOR	7402	7427		744078*
XOR	7486			
XNOR	747266			
Inverter	(1 input) 7404			

*The 744078 has an OR output and a NOR output on the same chip.

Figures 4-66 through 4-74 show the internal connections of the logic chips that you are most likely to use. Note that the 7402 NOR gate has its logical inputs and outputs arranged differently from all the other chips.

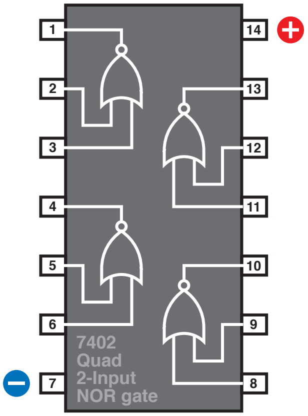
7402
Quad
2-Input
NOR gate

Figure 4-66. *Figures 4-66 through 4-74 show pinouts for some of the most widely used logic chips. Note that the inputs of the 7402 are reversed compared with the other chips.*

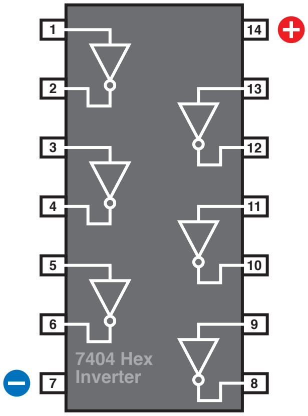
7404 Hex
Inverter

Figure 4-67

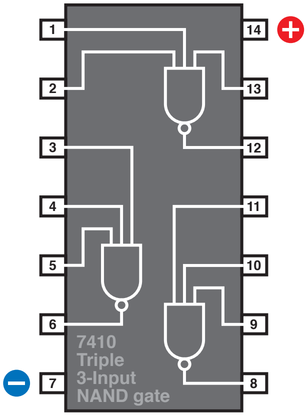
7410
Triple
3-Input
NAND gate

Figure 4-68

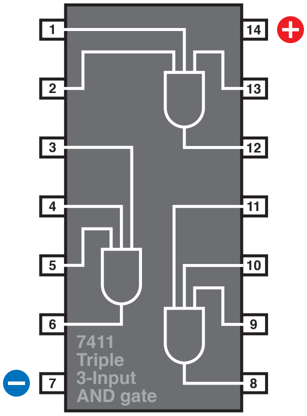
7411
Triple
3-Input
AND gate

Figure 4-69

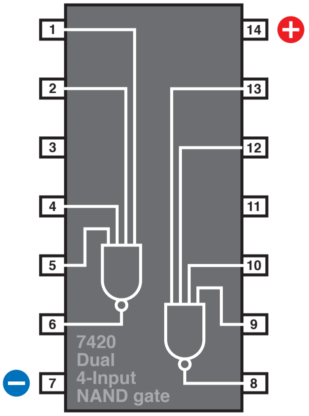
7420
Dual
4-Input
NAND gate

Figure 4-70

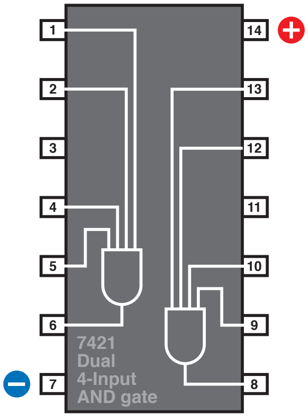
7421
Dual
4-Input
AND gate

Figure 4-71

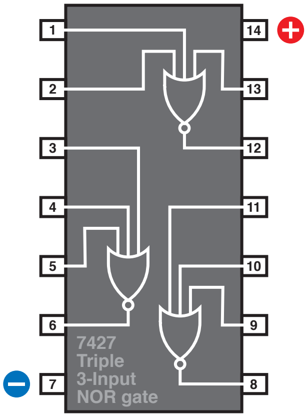
7427
Triple
3-Input
NOR gate

Figure 4-72

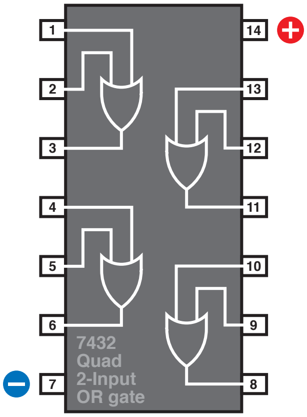
7432
Quad
2-Input
OR gate

Figure 4-73

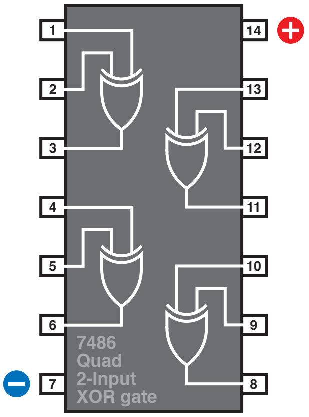
7486
Quad
2-Input
XOR gate

Figure 4-74

FUNDAMENTALS

Rules for connecting logic gates

Permitted:

- You can connect the input of a gate directly to your regulated power supply, either positive side or negative side.
- You can connect the output from one gate directly to the input of another gate.
- The output from one gate can power the inputs of many other gates (this is known as "fanout"). The exact ratio depends on the chip, but you can always power at least ten inputs with one logic output. The output from a logic chip can drive the trigger (pin 3) of a 555 timer. The output from the timer can then deliver 100mA, easily enough for half-a-dozen LEDs or a small relay.
- Low input doesn't have to be zero. A 74HCxx logic gate will recognize any voltage up to 1 volt as "low."
- High input doesn't have to be 5 volts. A 74HCxx logic gate will recognize any voltage above 3.5 volts as "high."

See Figures 4-75 and 4-76 for a comparison of permitted voltages on the input and output side of 74HCxx and 74LSxx chips.

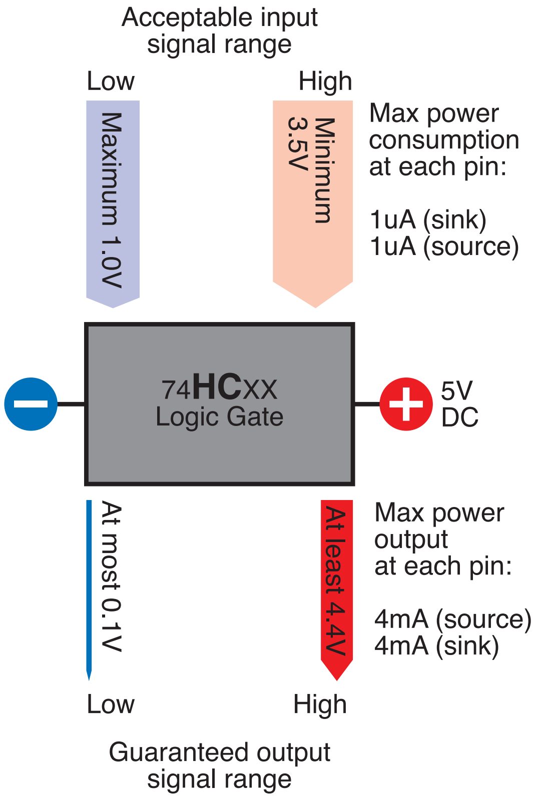
Acceptable input signal range
Low
High
Maximum 1.0V
Minimum 3.5V
Max power consumption at each pin:
1uA (sink)
1uA (source)
74HCXX
Logic Gate
5V DC
At most 0.1V
At least 4.4V
Max power output at each pin:
4mA (source)
4mA (sink)
Low
High
Guaranteed output signal range

Figure 4-75. *Each family of logic chips, and each generation in each family, has different standards for input and output minimum and maximum voltages. This diagram shows the standards used by the HC generation of the CMOS family, which was chosen for most of the projects in this book. Note that the current required for input is minimal compared with the current available for output. The power supply to the chip makes up the difference.*

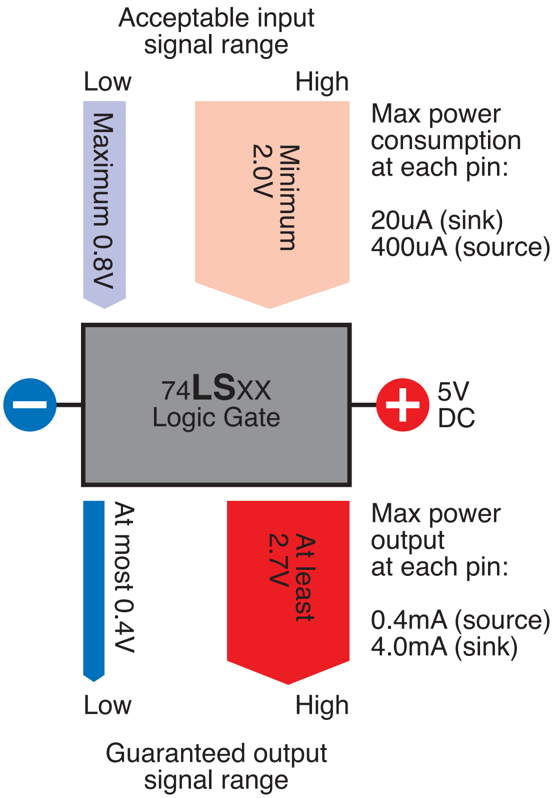
Acceptable input signal range
Low
High
Maximum 0.8V
Minimum 2.0V
Max power consumption at each pin:
20uA (sink)
400uA (source)
74LSXX
Logic Gate
5V DC
At most 0.4V
At least 2.7V
Max power output at each pin:
0.4mA (source)
4.0mA (sink)
Low
High
Guaranteed output signal range

Figure 4-76. *Because the LS generation of the TTL family has such different tolerances for input voltages and different standards for output voltages, the LS generation of TTL chips should not be mixed in the same circuit as the HC generation of CMOS chips, unless pull-up resistors are used to bring the LS chips into conformance with standards expected by the HC chips. See Experiment 21 for a case study in using LS chips.*

FUNDAMENTALS

Rules for connecting logic gates (continued)

Not permitted:

- No floating-input pins! On CMOS chips such as the HC family, you must always connect all input pins with a known voltage, even if they supply a gate on the chip that you're not using. When you use a SPST switch to control an input, remember that in its "off" position, it leaves the input unconnected. Use a pull-up or pull-down resistor to prevent this situation. See Figure 4-77.
- Don't use an unregulated power supply, or more than 5 volts, to power 74HCxx or 74LSxx logic gates.
- Be careful when using the output from a logic gate to power even a low-current LED. Check how many milliamps are being drawn. Also be careful when "sharing" the output from a logic gate with the input of another gate, at the same time that it is driving an LED. The LED may pull down the output voltage, to a point where the other gate won't recognize it. Always check currents and voltages when modifying a circuit or designing a new one.
- Never apply a significant voltage or current to the output pin of a logic gate. In other words, don't force an input into an output.
- Never link the outputs from two or more logic gates. If they must share a common output wire, use diodes to protect them from each other. See Figure 4-78.

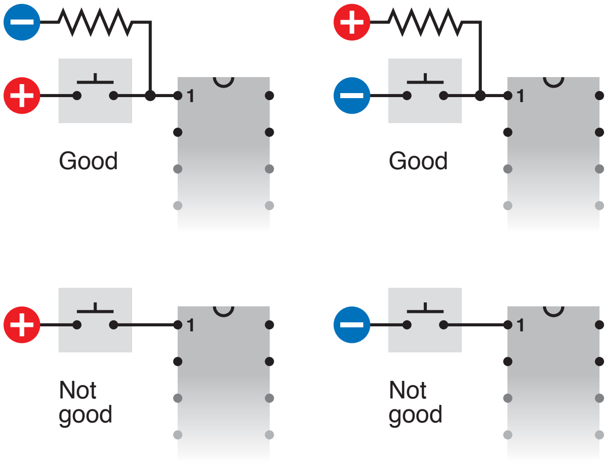

Figure 4-77. *Because a CMOS chip is so sensitive to input fluctuations, a logical input should never be left "floating," or unattached to a defined voltage source. This means that any single-throw switch or pushbutton should be used with a pull-up or pull-down resistor, so that when the contacts are open, the input is still defined.*

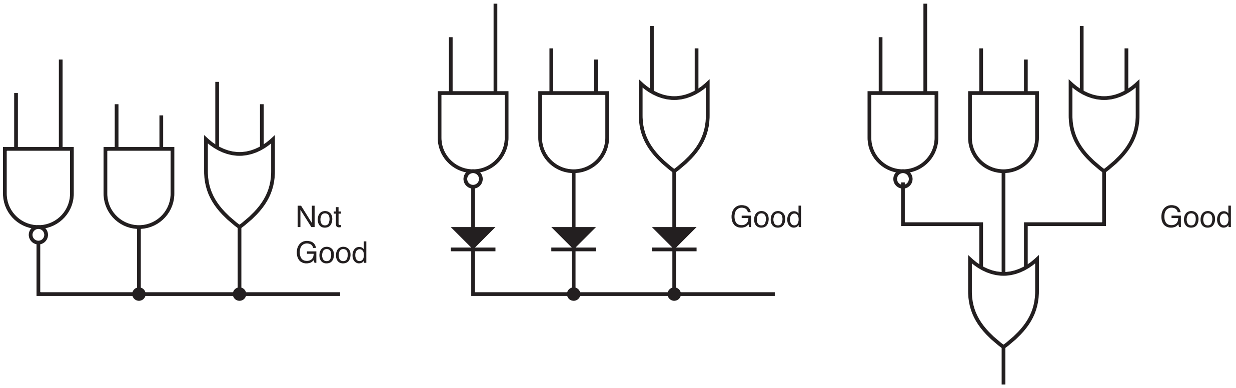

Figure 4-78. *The output from one logic gate must not be allowed to feed back into the output from another logic gate. Diodes can be used to isolate them, or they can be linked via another gate.*

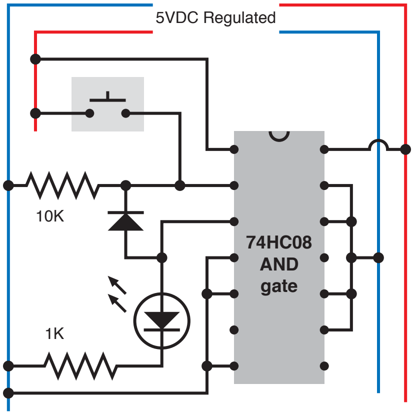

Figure 4-79. *Using a diode, the logical output from a gate can be allowed to feed back to one of its inputs, so that the gate latches after receiving a brief logical input pulse.*

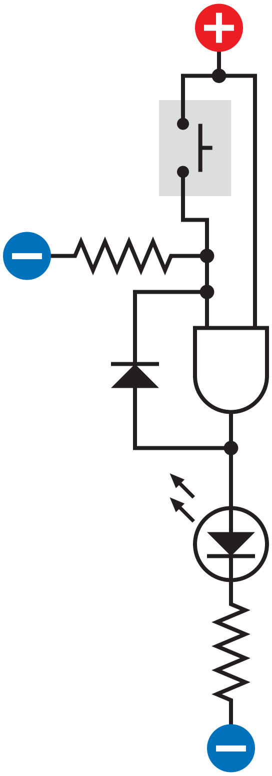

Figure 4-80. *The breadboard-format schematic in is simplified here to show more clearly the way in which a gate can latch itself after receiving an input pulse.*

In the 74HCxx logic family, each input of a logic gate consumes just a microamp, while the output can source 4 milliamps. This seems paradoxical: how can the chip give out more than it takes in? The answer is that it also consumes power from the power supply attached to pins 7 and 14. That's where the additional electricity comes from.

Because the logical output from a chip can be greater than the logical input, we can put the chip in a state where it keeps itself "switched on" in a way which is similar to the way the relay in the alarm project was wired to lock itself on. The simplest way to do this in a logic chip is by feeding some of the output back to one of the inputs.

Figure 4-79 shows an AND gate with one of its inputs wired to positive and its other input held low by a pull-down resistor, with a pushbutton that can make the input high. A signal diode connects the output of the chip back to the pushbutton-controlled input. Remember that the diode has a mark on it indicating the end which should be connected to the *negative* side of the power supply, which in this case will be the end of the 10K resistor.

The schematic in Figure 4-79 shows how the circuit should look in breadboard format. Figure 4-80 shows it in a simpler format.

From this point on, I won't bother to show the power regulator and the capacitors associated with it. Just remember to include them every time you see the power supply labeled as "5V DC Regulated."

When you switch on the power, the LED is dark, as before. The AND gate needs a positive voltage on both of its logical inputs, to create a positive output, but it now has positive voltage only on one of its inputs, while the other input is pulled down by the 10K resistor. Now touch the pushbutton, and the LED comes on. Let go of the pushbutton, and the LED stays on, because the positive output from the AND gate circulates back through the diode and is high enough to overcome the negative voltage coming through the pull-down resistor.

The output from the AND gate is powering its own input, so the LED will stay on until we disconnect it. This arrangement is a simple kind of "latch," and can be very useful when we want an output that continues after the user presses and releases a button.

You can't just connect the output from the gate to one of its inputs using an ordinary piece of wire, because this would allow positive voltage from the tactile switch to flow around and interfere with the output signal. Remember, you must never apply voltage to the output pin of a logic gate. The diode prevents this from happening.

If you've grasped the basics of logic gates, you're ready now to continue to our first real project, which will use all the information that I've set out so far.

Experiment 20: A Powerful Combination

Suppose you want to prevent other people from using your computer. I can think of two ways to do this: using software, or using hardware. The software would be some kind of startup program that intercepts the normal boot sequence and requests a password. You could certainly do it that way, but I think it would be more fun (and more relevant to this book) to do it with hardware. What I'm imagining is a numeric keypad requiring the user to enter a secret combination before the computer can be switched on.

You will need:

- Numeric keypad. As specified in the shopping list at the beginning of this chapter, it must have a "common terminal" or "common output." The schematic in Figure 4-82 shows what I mean. Inside the keypad, one conductor (which I have colored red to distinguish it from the others) connects with one side of every pushbutton. This conductor is "common" to all of them. It emerges from the keypad on an edge connector or set of pins at the bottom, which I've colored yellow.

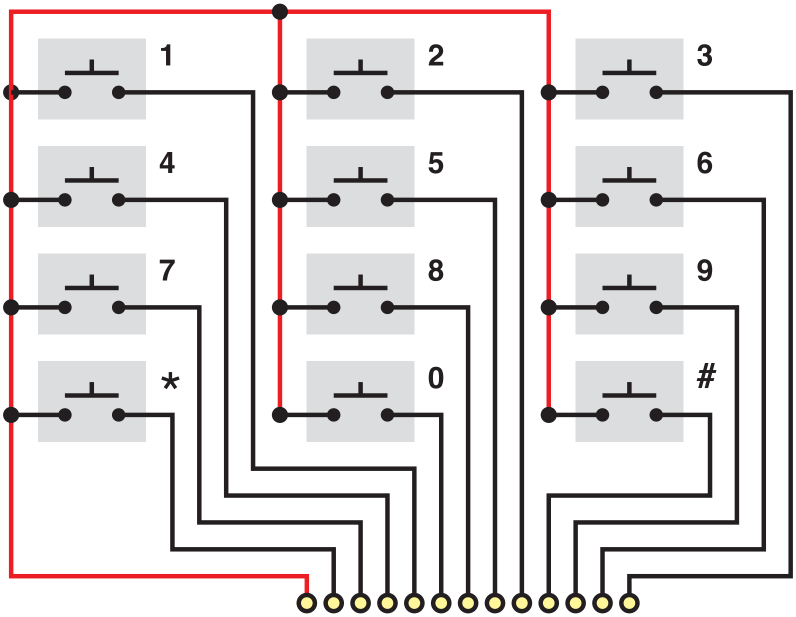

Figure 4-82. *A keypad of the type required for Experiment 20 incorporates a common terminal connected to one side of each of the 12 pushbuttons. The wire from the common terminal is shown red, here, to make it more easily identifiable.*

- Keypads that use "matrix encoding" won't work with the circuit that I'm going to describe. If the Velleman keypad, which I recommend, is unavailable, and you can't find another like it, you can use 12 separate SPST pushbuttons. Of course, that will cost a little more.
- 74HC08 logic chip containing four AND gates. Quantity: 1.
- 74HC04 logic chip containing six inverters. Quantity: 1.
- 555 timer chip. Quantity: 1.

The Warranty Issue

If you follow this project all the way to its conclusion, you'll open your desktop computer, cut a wire, and saw a hole in the cabinet. Without a doubt, this will void your warranty. If this makes you nervous, here are three options:

1. Breadboard the circuit for fun, and leave it at that.

2. Use the numeric keypad on some other device.

3. Use it on an old computer.

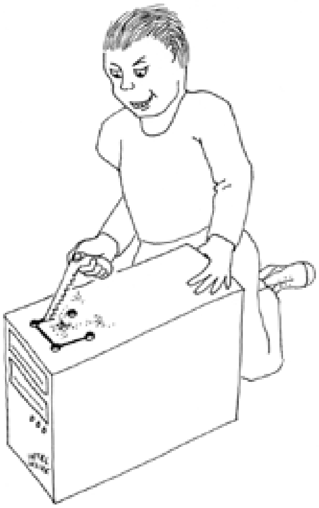

Figure 4-81. *Caution: This just might void your warranty.*

- Latching relay, 5 volt, DPST or DPDT, "2 form C" package, Panasonic DS2E-SL2-DC5V or similar. Must have two separate coils (one to latch, one to unlatch) with separate inputs. Quantity: 1.
- LEDs, 5mm generic, your choice of colors. Quantity: 3.
- Ribbon cable, with six conductors minimum, if you want to do a really neat job. You can use a cable of the type sold for hard drives, and split off the six conductors that you need, or shop around on eBay.
- Tools to open your computer, drill four holes, and make saw cuts between the holes, to create a rectangular opening for the keypad (if you want to take this project to its conclusion). Also, four small bolts to attach the keypad to the computer cabinet after you create the opening for it.

The Schematic

This time I'd like you to study the schematic before building anything. Let's start with the simplified version, shown in Figure 4-83.

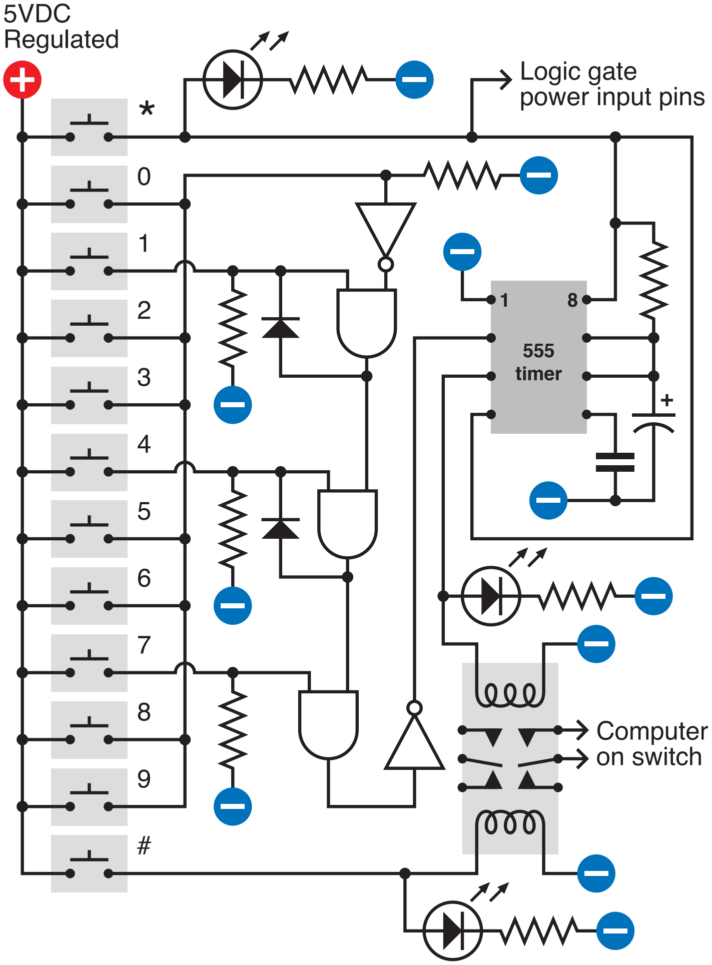

Figure 4-83. *A simplified schematic showing the basic structure of the combination lock circuit.*

I want this to be a battery-powered circuit, so that you don't have to run a separate power supply to it or (worse) try to tap into your computer's 5-volt bus. Battery power means that the circuit has to be "off" most of the time, to prevent the battery from running down. Because the keypad has two spare buttons (the asterisk and the pound sign), I'm going to use the asterisk as the "power on" button. When you press it, the LED at the top of the schematic lights up to confirm that everything's working, and the button sends power to the two logic chips and the 555 timer. You have to hold down the asterisk button while you punch in a three-digit code to unlock the computer.

Arbitrarily, I've chosen 1-4-7 as the three-digit code. Let's track what happens when you enter this sequence. (Naturally, if you build the circuit, you can wire it to choose any three digits you prefer.)

Pressing the 1 button sends positive power to one logical input of the first AND gate. The other logical input of this gate is also positive, because an inverter is supplying it, and the input of the inverter is being held negative by a pull-down resistor. When an inverter has a negative input, it gives a positive output, so pressing the 1 button activates the AND gate, and makes its output positive. The AND gate locks itself on, as its output cycles back to its switched input via a diode. So the gate output remains high even after you let go of the 1 button.

The output from the first AND gate also supplies one logical input of the second AND gate. When you press the 4 button, you send positive voltage to the other logical input of this AND gate, so its output goes high, and it locks itself on, just as the first gate did.

The second AND gate feeds the third AND gate, so when you press the 7 button, the third AND gate changes its output from low to high. This passes through an inverter, so the output from the inverter goes from high to low. This in turn goes to the trigger of a 555 timer wired in monostable mode.

When the trigger of a 555 timer goes from high to low, the timer emits a positive pulse through its output, pin 3. This runs down to the upper coil of the latching relay, and also flashes an LED to confirm that the code has been accepted and the relay has been activated.

Two of the contacts in the relay are wired into the power-up button of your computer. A little later in this description I'll explain why this should be safe with any modern computer.

Because we're using a latching relay, it flips into its "on" state and remains there, even when the power pulse from the timer ends. So now you can let go of the asterisk button to disconnect the battery power to your combination lock, and press the power-up button that switches on your computer.

At the end of your work session, you shut down your computer as usual, then press the pound button on your keypad, which flips the relay into its other position, reactivating the combination lock.

Incorrect Inputs

What happens if you enter the wrong code? If you press any button other than 1, 4, or 7, it sends positive voltage to the inverter near the top of the schematic. The positive voltage overwhelms the negative voltage being applied to the inverter through a pull-down resistor, and causes the inverter to output a negative voltage, which it applies to one of the logical inputs of the first AND gate. If the AND gate was locked on, the negative input will switch it off. If it was supplying the second AND gate, it'll switch that one off too.

Thus, any error when entering the first, second, or third digit of the secret code will reset the AND gates, forcing you to begin the sequence all over again.

What if you enter 1, 4, and 7 out of their correct sequence? The circuit won't respond. The third AND gate needs a high input supplied by the second AND gate, and the second AND gate needs a high input supplied by the first AND gate. So you have to activate the AND gates in the correct sequence.

Questions

Why did I use a 555 timer to deliver the pulse to the relay? Because the logical output from an AND gate cannot deliver sufficient power. I could have passed it through a transistor, but I liked the idea of a pulse of a fixed length to flip the relay and illuminate an LED for about 1 second, regardless of how briefly the user presses the 7 button.

Why do I need three LEDs? Because when you're punching buttons to unlock your computer, you need to know what's going on. The Power On LED reassures you that your battery isn't dead. The Relay Active LED tells you that the system is now unlocked, in case you are unable to hear the relay click. The System Relocked LED reassures you that you have secured your computer.

Because all the LEDs are driven either directly from the 5-volt supply or from the output of the 555 timer, they don't have to be low-current LEDs and can be used with 330Ω series resistors, so they'll be nice and bright.

How do you connect the keypad with the circuit? That's where your ribbon cable comes in. You carefully strip insulation from each of the conductors, and solder them to the contact strip or edge connector on your keypad. Push the conductors on the other end of the cable into your breadboard (when you're test-building the circuit) or solder them into perforated board (when you're building it permanently). Find a convenient spot inside your computer case where you can attach the perforated board, with double-sided adhesive or small bolts or whatever is convenient. Include a 9-volt battery carrier, and don't forget your power regulator to step the voltage down to 5 volts.

Breadboarding

No doubt you have realized by now that breadboards are very convenient as a quick way to push in some components and create connections, but the layout of their conductors forces you to put components in unintuitive configurations. Still, if you carefully compare the breadboard schematic in Figure 4-83 with the simplified schematic in Figure 4-84, you'll find that the connections are the same.

To help it make sense, I've shown the logic gates that exist inside the chips. I've also colored the power supply wires, as before, to reduce the risk of confusion. The positive side of the supply goes only to the common terminal on your keypad, and you have to press the asterisk key to send the power back down the ribbon cable, to supply the chips.

If you build the circuit and you can't understand why everything's dead, it's most likely because you forgot to hold down the asterisk button.

Note that the "wrong" numbers on the keypad are all shorted together. This will create some inconvenience if you want to change the combination in the future. I'll suggest a different option in the "enhancements" section that follows. For now, ideally, you should run a wire from every contact on your keypad, down to your circuit on its breadboard, and short the "wrong" keypad numbers together with jumper wires on the breadboard.

Also note that if you use a meter to test the inputs to the AND gates, and you touch your finger against the meter probe while doing so, this can be sufficient to trigger the sensitive CMOS inputs and give a false positive.

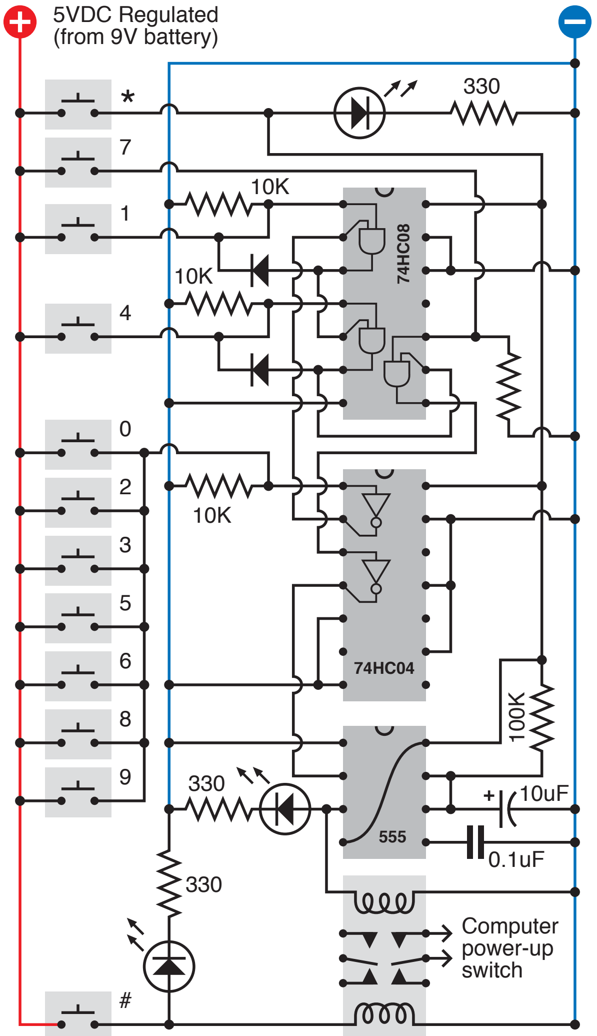

Figure 4-84. *The combination lock schematic redrawn to show how the components can be laid out on a breadboard.*

One Little Detail: The Computer Interface

Old computers used to have a big switch at the back, attached to the heavy metal box inside the computer, that transformed house current to regulated voltages that the computer needs. Most modern computers are not designed this way; you leave the computer plugged in, and you touch a little button on the box (if it's a Windows machine) or the keyboard (if it's a Mac), which sends a low-voltage pulse to the motherboard.

This is ideal from our point of view, because we don't have to mess with high voltages. Don't even think of opening that metal box with the fan mounted in it, containing the computer power supply. Just look for the wire (usually containing two conductors, on a Windows machine) that runs from the "power up" button to the motherboard.

To check that you found the right one, *make sure that your computer is unplugged*, ground yourself (because computers contain CMOS chips that are sensitive to static electricity) and very carefully snip just one of the two conductors in the wire. Now plug in your computer and try to use the "power up" button. If nothing happens, you've probably cut the right wire. Even if you cut the wrong wire, it still prevented your computer from booting, which is what you want, so you can use it anyway. Remember, we are not going to introduce any voltage to this wire. We're just going to use the relay as a switch to reconnect the conductor that you cut. You should have no problem if you maintain a cool and calm demeanor, and look for that single wire that starts everything. Check online for the maintenance manual for your computer if you're really concerned about making an error.

After you find the wire and cut just one of its conductors, unplug your computer again, and keep it unplugged during the next steps.

Find where the wire attaches to the motherboard. Usually there's a small unpluggable connector. First, mark it so that you know how to plug it back in the right way around, and then disconnect it while you follow the next couple of steps.

Strip insulation from the two ends of the wire that you cut, and solder an additional piece of two-conductor wire, as shown in Figure 4-85, with heat-shrink tube to protect the solder joints. (This is very important!)

Run your new piece of wire to the latching relay, making sure you attach it to the pair of contacts which close, inside the relay, when it is energized by the unlocking operation. You don't want to make the mistake of unlocking your computer when you think you're locking it, and vice versa.

Reconnect the connector that you disconnected from your motherboard, plug in your computer, and try to power it up. If nothing happens, this is probably good! Now enter the secret combination on your keypad (while holding down the asterisk button to provide battery power) and listen for the click as the relay latches. Now try the "power up" button again, and everything should work.

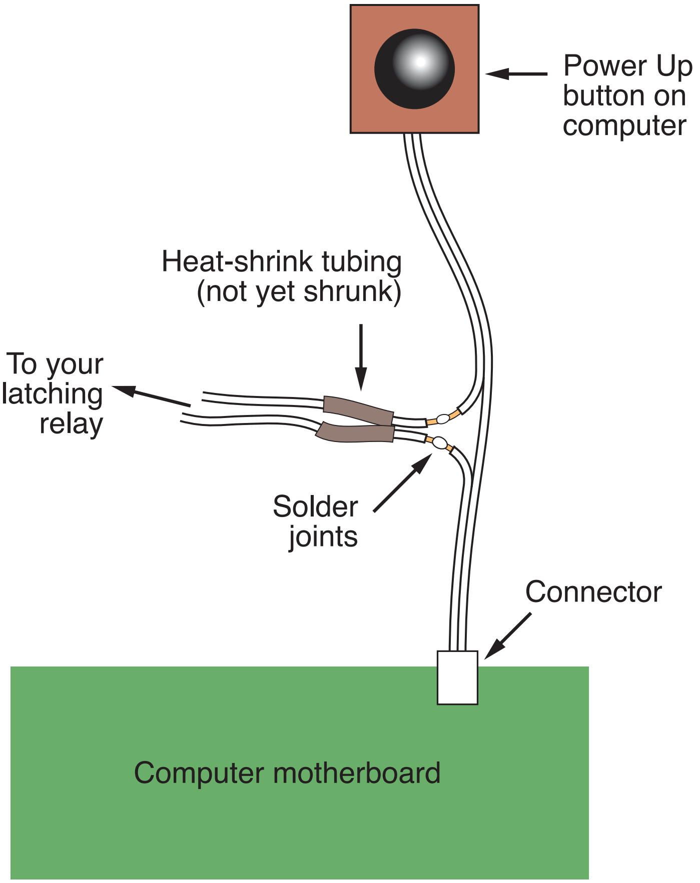
Power Up
button on
computer
Heat-shrink tubing
(not yet shrunk)
To your
latching
relay
Solder
joints
Connector
Computer motherboard

Figure 4-85. *The combination lock project can be interfaced with a typical desktop computer by cutting one conductor in the wire from the "power up" pushbutton, soldering an extension, and covering the joints with heat-shrink tube.*

Enhancements

At the end of any project, there's always more you can do.

To make this setup more secure, you could remove the usual screws that secure the case of the computer, and replace them with tamper-proof screws. Check any online source for "tamper-proof screws," such as *http://www.mcmaster.com*. Naturally, you will also need the special tool that fits the screws, so that you can install them (or remove them, if your security system malfunctions for any reason).

Another enhancement could be an additional 555 timer that is activated by the asterisk button, and delivers power to the other chips for, say, a limited period of 30 seconds, allowing you that much time to unlock the system. This would eliminate the need to hold down the asterisk button while you enter the unlocking code. A 555 timer can supply power to all the other chips, because they don't use very much. I omitted this feature for the sake of simplicity.

Yet another enhancement, if you are security-crazed, is to go for a four-button code. After all, the 74HC08 chip still has one unused AND gate. You could insert that into the chain of the existing AND gates and wire it to another keypad button of your choice.

Still another enhancement would be a way to change the code without unsoldering and resoldering wires. You can use the miniature sockets that I suggested in the heartbeat flasher project. These should enable you to swap around the ends of your wires from the keypad.

And for those who are absolutely, positively, totally paranoid, you could fix things so that entering a wrong code flips a second high-amperage relay which supplies a massive power overload, melting your CPU and sending a big pulse through a magnetic coil clamped to your hard drive, instantly turning the data to garbage (Figure 4-86). Really, if you want to protect information, messing up the hardware has major advantages compared with trying to erase data using software. It's faster, difficult to stop, and tends to be permanent. So, when the Record Industry Association of America comes to your home and asks to switch on your computer so that they can search for illegal file sharing, just accidentally give them an incorrect unlocking code, sit back, and wait for the pungent smell of melting insulation.

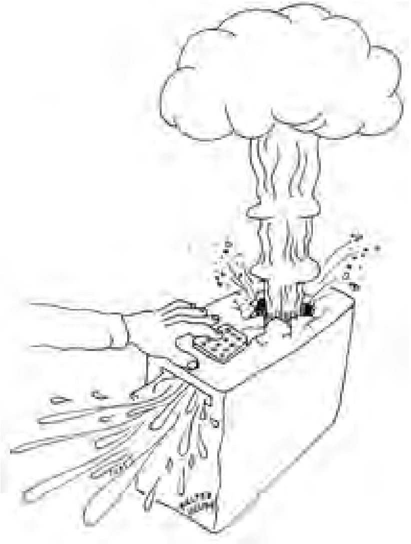

Figure 4-86. *For those who are absolutely, positively, totally paranoid: a meltdown/self-destruct system controlled by a secret key combination provides enhanced protection against data theft or intrusions by RIAA investigators asking annoying questions about file sharing.*

If you pursue this option, I *definitely* take no responsibility for the consequences.

On a more realistic level, no system is totally secure. The value of a hardware locking device is that if someone does defeat it (for instance, by figuring out how to unscrew your tamper-proof screws, or simply ripping your keypad out of the computer case with metal shears), at least you'll know that something happened—especially if you put little dabs of paint over the screws to reveal whether they've been messed with. By comparison, if you use password-protection software and someone defeats it, you may never know that your system has been compromised.

Experiment 21: Race to Place

The next project is going to get us deeper into the concept of feedback, where the output is piped back to affect the input—in this case, blocking it. It's a small project, but quite subtle, and the concepts will be useful to you in the future.

You will need:

- 74HC32 chip containing four OR gates. Quantity: 1.
- 555 timers. Quantity: 2.
- SPDT switch. Quantity: 1.
- SPST tactile switches. Quantity: 2.
- Various resistors.
- 5V supply using power regulator as before.

The Goal

On quiz shows such as *Jeopardy*, contestants race to answer each question. The first person who hits his answer button automatically locks out the other contestants, so that their buttons become inactive. How can we make a circuit that will do the same thing?

If you search online, you'll find several hobby sites where other people have suggested circuits to work this way, but they lack some features that I think are necessary. The approach I'm going to use here is both simpler and more elaborate. It's simpler because it has a very low chip count, but it's more elaborate in that it incorporates "quizmaster control" to make a more realistic game.

I'll suggest some initial ideas for a two-player version. After I develop that idea, I'll show how it could be expanded to four or even more players.

A Conceptual Experiment

I want to show how this kind of project grows from an idea to the finished version. By going through the steps of developing a circuit, I'm hoping I may inspire you to develop ideas of your own in the future, which is much more valuable than just replicating someone else's work. So join me in a conceptual experiment, thinking our way from a problem to a solution.

First consider the basic concept: two people have two buttons, and whoever goes first locks out the other person. I always find it helps me to visualize this kind of thing if I draw a sketch, so that's where I'll begin. In Figure 4-87, the signal from each button passes through a component that I'll call a "button blocker," activated by the other person's button. I'm not exactly sure what the button blocker will be or how it will work, yet.

Now that I'm looking at it, I see a problem here. If I want to expand this to three players, it will get complicated, because each player must activate the "button blockers" of *two* opponents. Figure 4-88 shows this. And if I have four players, it's going to get even more complicated.

Anytime I see this kind of complexity, I think there has to be a better way.

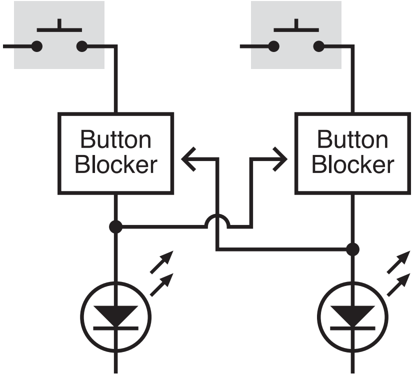

Figure 4-87. *The basic concept of the quiz project is that the output from one button should feed back to intercept the output of another button. At this point, the way in which the "button blocker" circuit works has not been figured out.*

Also, there's another problem. After a player lets his finger off the button, the other players' buttons will be unblocked again. I need a latch to hold the signal from the first player's button and continue to block the other players.

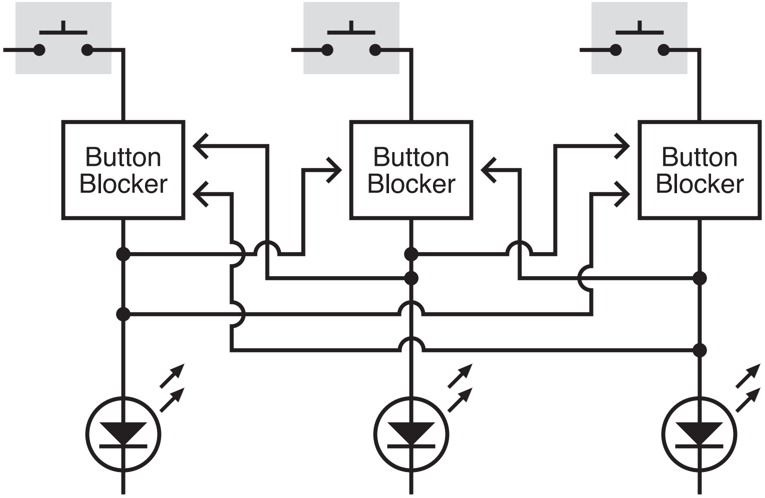

Figure 4-88. *The quiz concept becomes more complicated when an additional pushbutton is added. Now each button must block two other buttons. If a fourth button is added, the circuit will become unmanageably complex. There has to be a better way.*

This now sounds even more complicated. But wait a minute, if I have a latch which allows the winning player to take his finger off his button, I don't care if any of the buttons are being pressed anymore—including the button of the winning player. As soon as his signal is latched, *all* the buttons can be blocked. This makes things much simpler. I can summarize it as a sequence of events:

1. First player presses his button.
2. The signal is latched.
3. The latched signal feeds back and blocks all the buttons.

The new sketch in Figure 4-89 shows this. Now the configuration is modular, and can be expanded to almost any number of players, just by adding more modules.

There's something important missing, though: a reset switch, to put the system back to its starting mode after the players have had time to press their buttons and see who won. Also, I need a way to prevent players from pressing their buttons too soon, before the quizmaster has finished asking the question. Maybe I can combine this function in just one switch, which will be under the quizmaster's control. In its Reset position, the switch can reset the system and remove power to the buttons. In its Play position, the switch stops holding the system in reset mode, and provides power to the buttons. Figure 4-90 shows this. I've gone back to showing just two players, to minimize the clutter of lines and boxes, but the concept is still easily expandable.

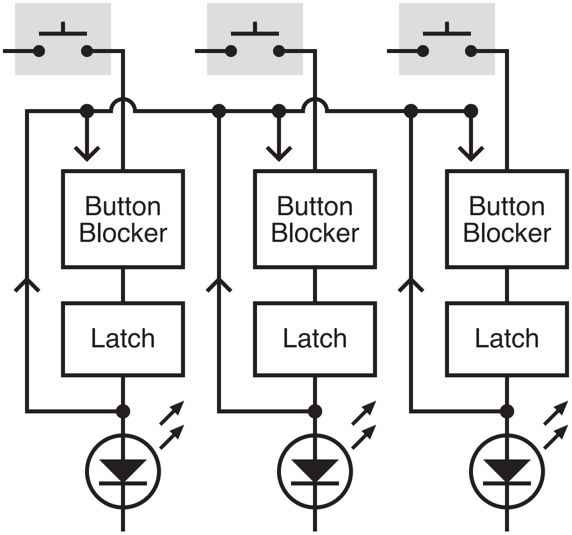

Figure 4-89. *If a latch is added below each button, it can retain one input and then block all inputs from all buttons. This simplifies the concept.*

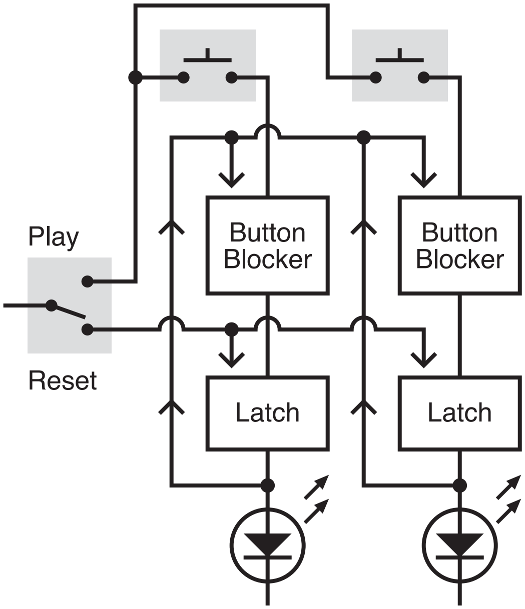

Figure 4-90. *A quizmaster switch will be needed to activate the buttons initially and then reset the circuit after a winning input has been recorded.*

Now I have to deal with a logic problem in the diagram. The way I've drawn it, after the output from the lefthand latch goes up to the "button blockers," it can also run down the wire to the other half of the circuit (against the direction to the arrows), because everything is joined together. In other words, if the lefthand LED lights up, the righthand LED will light up, too. How can I stop this from happening?

Well, I could put diodes in the "up" wires to block current from running down them. But I have a more elegant idea: I'll add an OR gate, because the inputs to an OR gate are separated from each other electrically. Figure 4-91 shows this.

Usually an OR gate has only two logical inputs. Will this prevent me from adding more players? No, because you can actually buy an OR that has eight inputs. If any one of them is high, the output is high. For fewer than eight players, I can short the unused inputs to ground, and ignore them.

Looking again at Figure 4-91, I'm getting a clearer idea of what the thing I've called a "button blocker" should actually be. I think it should be another logic gate. It should say, "If there's only one input, from a button, I'll let it through. But if there is a second input from the OR gate, I won't let it through."

That sounds like a NAND gate, but before I start choosing chips, I have to decide what the latch will be. I can buy an off-the-shelf flip-flop, which flips "on" if it gets one signal and "off" if it gets another, but the trouble is, chips containing flip-flops tend to have more features than I need for a simple circuit like this. Therefore I'm going to use 555 timers again, in flip-flop mode. They require very few connections, work very simply, and can deliver a good amount of current. The only problem with them is that they require a negative input at the trigger pin to create a positive output. But I think I can work with that.

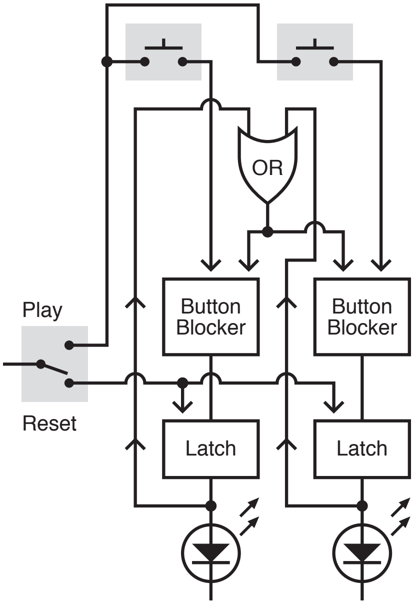

Figure 4-91. *To prevent the output from one latch feeding back around the circuit to the output from another latch, the outputs can be combined in an OR gate.*

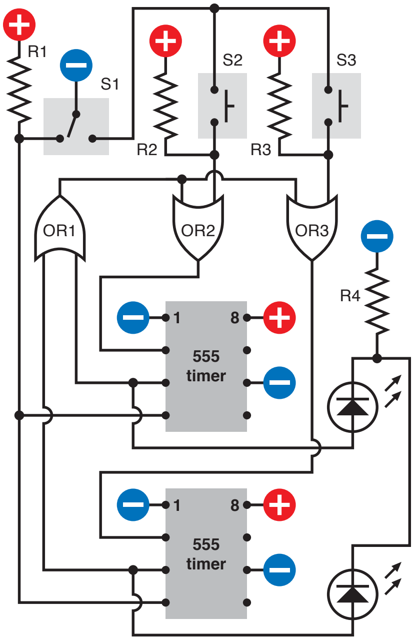

Figure 4-92. *Now that the basic concept of the quiz circuit has been roughed out, specific components can be inserted, with compatible inputs and outputs.*

So now, finally, here's a simplified schematic, in Figure 4-92. I like to show the pins of the 555 timers in their correct positions, so I had to move the components around a little to minimize wire crossovers, but you can see that logically, it's the same basic idea.

Before you try to build it, just run through the theory of it, because that's the final step, to make sure there are no mistakes. The important thing to bear in mind is that because the 555 needs a negative input on its trigger pin to create its output, when any of the players presses a button, the button has to create a negative "flow" through the circuit. This is a bit counterintuitive, so I'm including a three-step visualization in Figure 4-93, showing how it will work.

In Step 1, the quizmaster has asked a question and flipped his switch to the right, to supply (negative) power to the players' buttons. So long as no one presses a button, the pull-up resistors supply positive voltage to OR2 and OR3. An OR gate has a positive output if it has any positive input, so OR2 and OR3 keep the trigger inputs of the 555 timers positive. Their outputs remain low, and nothing is happening yet.

In Step 2, the lefthand player has pressed his button. Now OR2 has two negative inputs, so its output has gone low. But IC1 hasn't reacted yet.

In Step 3, just a microsecond later, IC1 has sensed the low voltage on its trigger, so its output from pin 3 has gone high, lighting the LED. Remember, this 555 timer is in flip-flop mode, so it locks itself into this state immediately. Meanwhile its high output also feeds back to OR1. Because OR1 is an OR gate, just one high input is enough to make a high output, so it feeds this back to OR2 and OR3. And now that they have high inputs, their outputs also go high, and will stay high, regardless of any future button-presses.

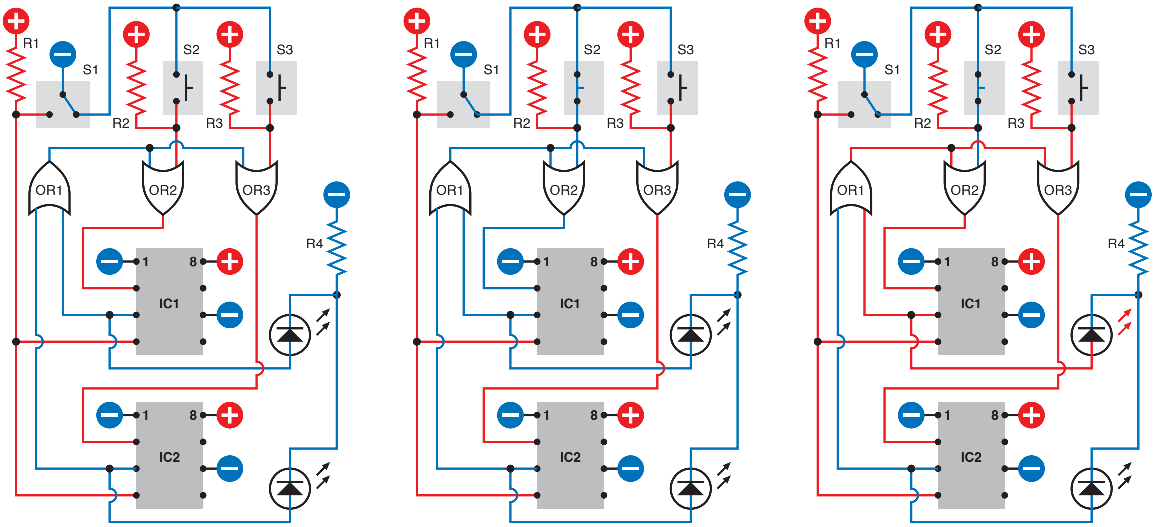

Figure 4-93. *These three schematics show the prevalence of higher and lower voltages (red and blue lines) through the quiz circuit when a pushbutton is pressed.*

Because OR2 and OR3 now have high inputs and outputs, IC1 and IC2 cannot be triggered. But IC1 is still locked into its "on" state, keeping the LED illuminated.

The only way to change IC1 is if the quizmaster flips his switch back to the left. That applies negative power to the reset pins of both the timers. Consequently their outputs go low, the LED goes out, and the circuit goes back into the same state as where it started. Having reset it, the quizmaster can ask another question, but the players' buttons are not activated until the quizmaster flips the switch back to the right again.

There's only one situation that I haven't addressed: what if both players press their buttons absolutely simultaneously? In the world of digital electronics, this is highly unlikely. Even a difference of a microsecond should be enough time for the circuit to react and block the second button. But if somehow both buttons are pressed at the same instant, both of the timers should react, and both of the LEDs will light up, showing that there has been a tie.

In case you feel a little uncertain about the way in which a two-player circuit can be upgraded to handle extra players, I've included a simplified three-player schematic in Figure 4-94.

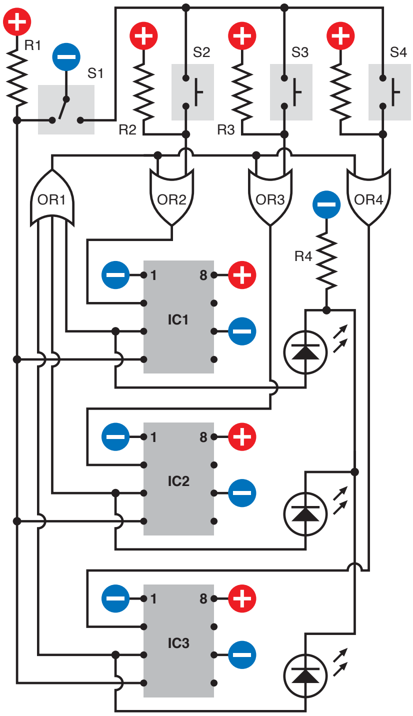

Figure 4-94. *The two-player schematic in can be easily upgraded to a three-player version, as shown here, provided the first OR gate can handle three inputs.*

Breadboarding It

Now it's time to create a schematic that's as close to the breadboard layout as possible, so that you can build this thing easily. The schematic is shown in Figure 4-95 and the actual components on a breadboard are in Figure 4-96. Because the only logic gates that I've used are OR gates, and there are only three of them, I just need one logic chip: the 74HC32, which contains four 2-input OR gates. (I've grounded the inputs to the fourth). The two OR gates on the left side of the chip have the same functions as OR2 and OR3 in my simplified schematic, and the OR gate at the bottom-right side of the chip works as OR1, receiving input from pin 3 of each 555 timer. If you have all the components, you should be able to put this together and test it quite quickly.

You may notice that I've made one modification of the previous schematic. A 0.01 µF capacitor has been added between pin 2 of each 555 timer (the Input) and negative ground. Why? Because when I tested the circuit without the capacitors, sometimes I found that one or both of the 555 timers would be triggered simply by flipping S1, the quizmaster switch, without anyone pressing a button.

At first this puzzled me. How were the timers getting triggered, without anyone doing anything? Maybe they were responding to "bounce" in the quizmaster switch. Sure enough, the small capacitors solved the problem. They may also slow the response of the 555 timers fractionally, but not enough to interfere with slow human reflexes.

As for the buttons, it doesn't matter if they "bounce," because each timer locks itself on at the very first impulse and ignores any hesitations that follow.

You can experiment building the circuit, disconnecting the 0.01 µF capacitors, and flipping S1 to and fro a dozen times. If you have a high-quality switch, you may not experience any problem. If you have a lower-quality switch, you may see a number of "false positives." I'm going to explain more about "bounce," and how to get rid of it, in the next experiment.

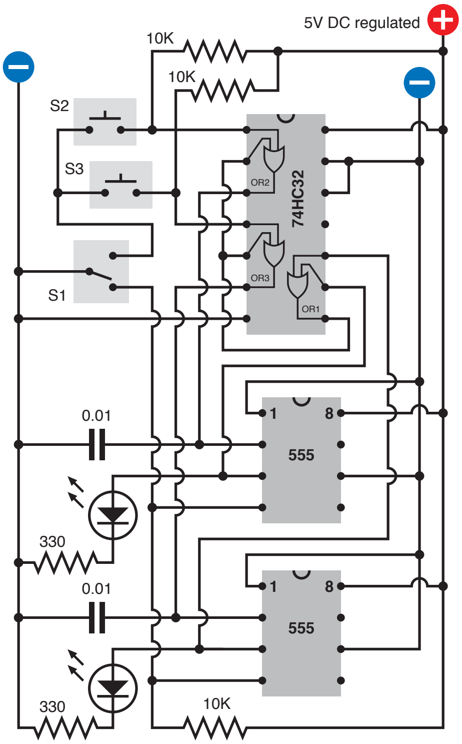

Figure 4-95. *Applying the simplified schematic to a breadboard inevitably entails a wiring layout that is less intuitively obvious and appears more complex. The connections are the same, though.*

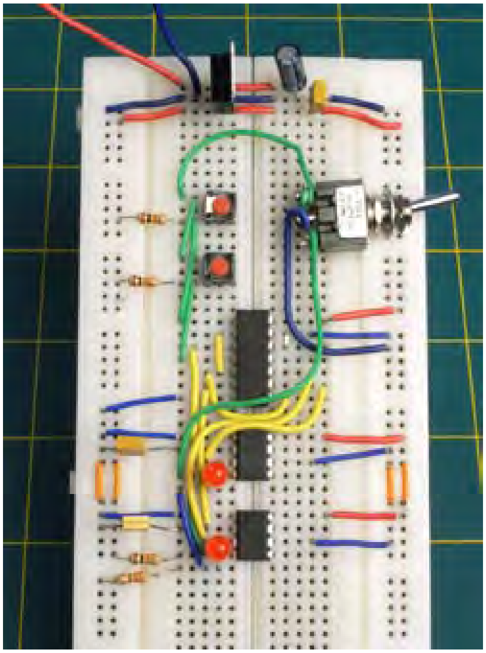

Figure 4-96. *The quiz schematic applied to a breadboard, to test the concept prior to full-scale implementation.*

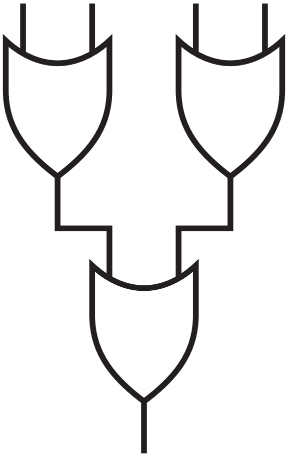

Figure 4-97. *Although a four-input OR gate is not manufactured, its functionality can be achieved easily by linking three 2-input OR gates together.*

Enhancements

After you breadboard the circuit, if you proceed to build a permanent version, I suggest that you expand it so that at least four players can participate. This will require an OR gate capable of receiving four inputs. The 74HC4078 is the obvious choice, as it allows up to eight. Just connect any unused inputs to negative ground.

Alternatively, if you already have a couple of 74HC32 chips and you don't want to bother ordering a 74HC4078, you can gang together three of the gates inside a single 74HC32 so that they function like a four-input OR. Look at the simple logic diagram in Figure 4-97 showing three ORs, and remember that the output from each OR will go high if at least one input is high.

And while you're thinking about this, see if you can figure out the inputs and output of three ANDs in the same configuration.

For a four-player game, you'll also need two additional 555 timers, of course, and two more LEDs, and two more pushbuttons.

As for creating a schematic for the four-player game—I'm going to leave that to you. Begin by sketching a simplified version, just showing the logic symbols. Then convert that to a breadboard layout. And here's a suggestion: pencil, paper, and an eraser can still be quicker, initially, than circuit-design software or graphic-design software, in my opinion.

Experiment 22: Flipping and Bouncing

I mentioned in the previous experiment that "bounce" from the buttons in the circuit wouldn't be a problem, because the buttons were activating 555 timers that were wired in bistable, flip-flop mode. As soon as the timer receives the very first pulse, it flips into its new state and flops there, ignoring any additional noise in the circuit. So can we debounce a switch or a button using a flip-flop? And as some 74HCxx chips are available containing flip-flops, can we use them?

The answers are yes, and yes, although it's not quite as simple as it sounds.

You will need:

- 74HC02 logic chip containing 4 NOR gates. 74HC00 logic chip containing 4 NAND gates. Quantity: 1 of each.
- SPDT switch. Quantity: 1.
- LEDs, low-current. Quantity: 2.
- 10K resistors and 1K resistors. Quantity: 2 of each.

Assemble the components on your breadboard, following the schematic shown in Figure 4-98. When you apply power (through your regulated 5-volt supply), one of the LEDs should be lit.

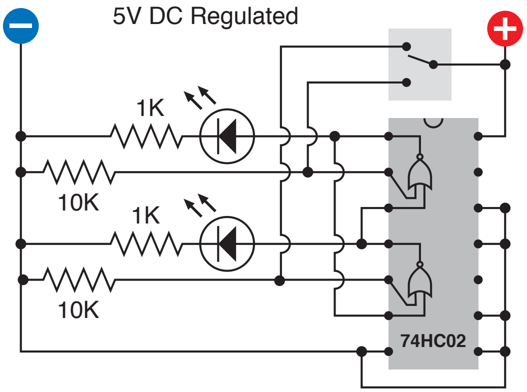

Figure 4-98. *A simple circuit to test the behavior of two NOR gates wired as a simple flip-flop that retains its state after an input pulse ceases.*

Now I want you to do something odd. Please disconnect the SPDT switch by taking hold of the wire that connects the positive power supply to the pole of the switch, and pulling the end of the wire out of the breadboard. When you do this, you may be surprised to find that the LED remains lit.

Push the wire back into the breadboard, flip the switch, and the first LED should go out, while the other LED should become lit. Once again, pull out the wire, and once again, the LED should remain lit.

Here's the take-home message:

- A flip-flop requires only an initial pulse.
- After that, it ignores its input.

How It Works

Two NOR gates or two NAND gates can function as a flip-flop:

- Use NOR gates when you have a positive input from a double-throw switch.
- Use NAND gates when you have a negative input from a double-throw switch.

Either way, you have to use a double-throw switch.

I've mentioned the double-throw switch three times (actually, four times if you count this sentence!) because for some strange reason, most introductory books fail to emphasize this point. When I first started learning electronics, I went crazy trying to understand how two NORs or two NANDs could debounce a simple SPST pushbutton—until finally I realized that they can't. The reason is that when you power up the circuit, the NOR gates (or NAND gates) need to be told in which state they should begin. They need an initial orientation, which comes from the switch being in one state or the other. So it has to be a double-throw switch. (Now I've mentioned it five times.)

I'm using another simplified multiple-step schematic, Figure 4-99, to show the changes that occur as the switch flips to and fro with two NOR gates. To refresh your memory, I've also included a truth table showing the logical outputs from NOR gates for each combination of inputs.

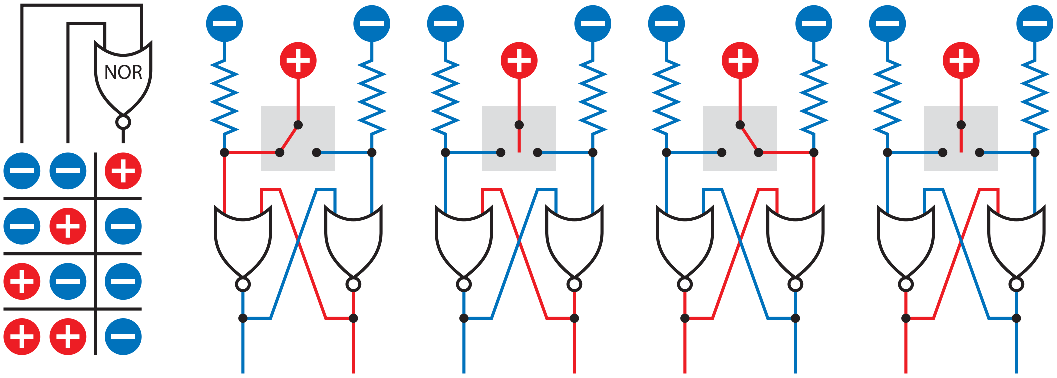

−	−	+
−	+	−
+	−	−
+	+	−

Figure 4-99. *Using two NOR gates in conjunction with a positive input through a SPDT switch, this sequence of four diagrams shows how a flip-flop circuit responds.*

Suppose that the switch is turned to the left. It sends positive current to the lefthand side of the circuit, overwhelming the negative supply from the pull-down resistor, so we can be sure that the NOR gate on the left has one positive logical input. Because any positive logical input will make the NOR give a negative output (as shown in the truth table), the negative output crosses over to the righthand NOR, so that it now has two negative inputs, which make it give a positive output. This crosses back to the lefthand NOR gate. So, in this configuration everything is stable.

Now comes the clever part. Suppose that you move the switch so that it doesn't touch either of its contacts. (Or suppose that the switch contacts are bouncing, and failing to make a good contact. Or suppose you disconnect the switch entirely.) Without a positive supply from the switch, the lefthand input of the left NOR gate goes from positive to negative, as a result of the pull-down resistor. But the righthand input of this gate is still positive, and one positive is all it takes to make the NOR maintain its negative output, so nothing changes. In other words, the circuit has "flopped" in this state.

Now if the switch turns fully to the right and supplies positive power to the righthand pin of the right NOR gate, quick as a flash, that NOR recognizes that it now has a positive logical input, so it changes its logical output to negative. That goes across to the other NOR gate, which now has two negative inputs, so its output goes positive, and runs back to the right NOR.

In this way, the output states of the two NOR gates change places. They flip, and then flop there, even if the switch breaks contact or is disconnected again. The second set of drawings in Figure 4-100 shows exactly the same logic, using a negatively powered switch and two NAND gates. You can use your 74HC00 chip, specified in the parts list for this experiment, to test this yourself.

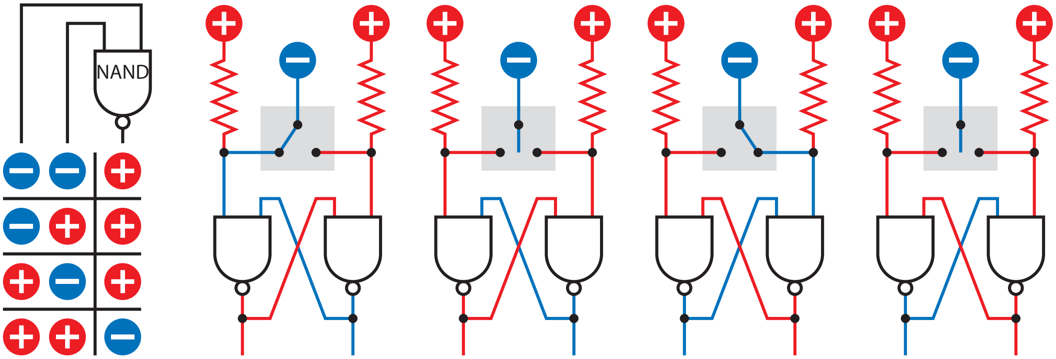

Figure 4-100. *The schematic from Figure 4-99 can be rewired with NAND gates and a negative switched input.*

Both of these configurations are examples of a jam-type flip-flop, so called because the switch forces it to respond immediately, and jams it into that state. You can use this circuit anytime you need to debounce a switch (as long as it's a double-throw switch).

A more sophisticated version is a clocked flip-flop, which requires you to set the state of each input first and then supply a clock pulse to make the flip-flop respond. The pulse has to be clean and precise, which means that if you supply it from a switch, the switch must be debounced—probably by using another jam-type flip-flop! Considerations of this type have made me reluctant to use clocked flip-flops in this book. They add a layer of complexity, which I prefer to avoid in an introductory text.

What if you want to debounce a single-throw button or switch? Well, you have a problem! One solution is to buy a special-purpose chip such as the 4490 "bounce eliminator," which contains digital delay circuitry. A specific part number is the MC14490PG from On Semiconductor. This contains six circuits for six separate inputs, each with an internal pull-up resistor. It's relatively expensive, however—more than 10 times the price of a 74HC02 containing NOR gates. Really, it may be simpler to use double-throw switches that are easily debounced as described previously.

Experiment 23: Nice Dice

This is the one experiment where I want you to use the 74LSxx generation of the TTL logic family, instead of the 74HCxx family of CMOS. Two reasons: first, I need to use the 7492 counter, which is unavailable in the HC family. And second, you should know the basic facts about the LS series of TTL chips, as they still crop up in circuits that you'll find in electronics books and online.

In addition, you'll learn about "open collector" TTL chips such as the 74LS06 inverter, which can be a convenient substitute for transistors when you want to deliver as much as 40mA of current.

The idea of this circuit is simple enough: run a 555 timer sending very fast pulses to a counter that counts in sixes, driving LEDs that are placed to imitate the spots on a die. (Note that the word "die" is the singular of "dice.") The counter runs so fast, the die-spots become a blur. When the user presses a button, the counter stops arbitrarily, displaying an unpredictable spot pattern.

Dice simulations have been around for many, many years, and you can even buy kits online. But this one will do something more: it will also demonstrate the principles of binary code.

So, if you're ready for the triple threat of TTL chips, open collectors, and binary, let's begin.

You will need:

- 74LS92 counter such as SN74LS92N by Texas Instruments. Quantity: 1 if you want to create one die, 2 to make two dice.
- 74LS27 three-input NOR gate such as SN74LS27N by Texas Instruments. Quantity: 1.
- 555 timers. Quantity: 1 if you want to make one die, 2 to make two dice.
- Signal diodes, 1N4148 or similar. Quantity: 4, or 8 to make two dice.

Seeing Binary

The counter that we dealt with before was unusual, in that its outputs were designed to drive seven-segment numerals. A more common type has outputs that count in binary code.

The 74LS92 pinouts are shown in Figure 4-101. Plug the chip into your breadboard and make connections as shown in Figure 4-102. Initially, the 555 timer will drive the counter in slow-motion, at around 1 step per second. Figure 4-103 shows the actual components on a breadboard.

Note that the counter has unusual power inputs, on pins 5 and 10 instead of at the corners. Also four of its pins are completely unused, and do not connect with anything inside the chip. Therefore, you don't need to attach any wire to them on the outside.

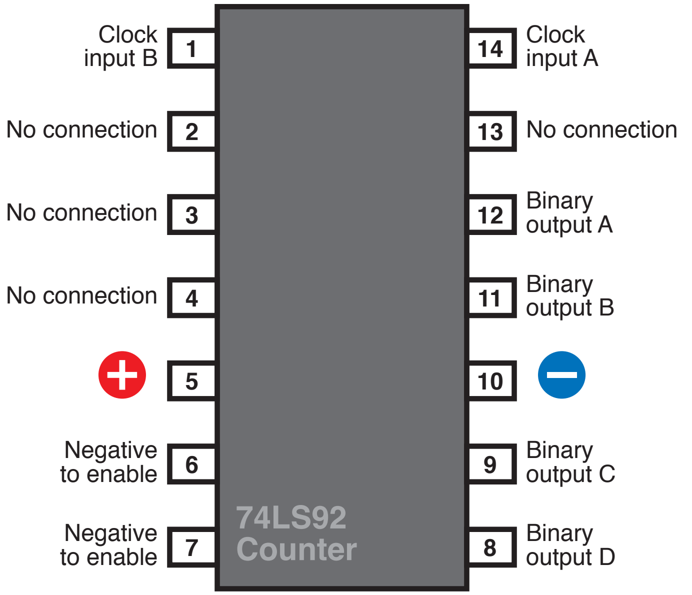
Clock input B
1
No connection
2
No connection
3
No connection
4
+
5
Negative to enable
6
Negative to enable
7
74LS92 Counter
14
Clock input A
13
No connection
12
Binary output A
11
Binary output B
10
−
9
Binary output C
8
Binary output D

Figure 4-101. *The unusual pin assignments include four that have no connection of any kind inside the chip, and can be left unattached.*

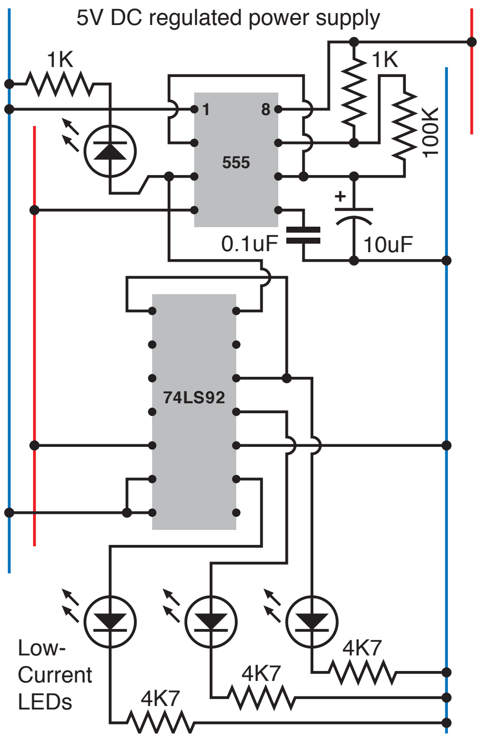
5V DC regulated power supply
1K
1K
100K
1
8
555
+
0.1uF
10uF
74LS92
Low-Current LEDs
4K7
4K7
4K7

Figure 4-102. *This simple circuit uses a 555 timer running slowly to control the 74LS92 binary counter and display the succession of high states from its outputs.*

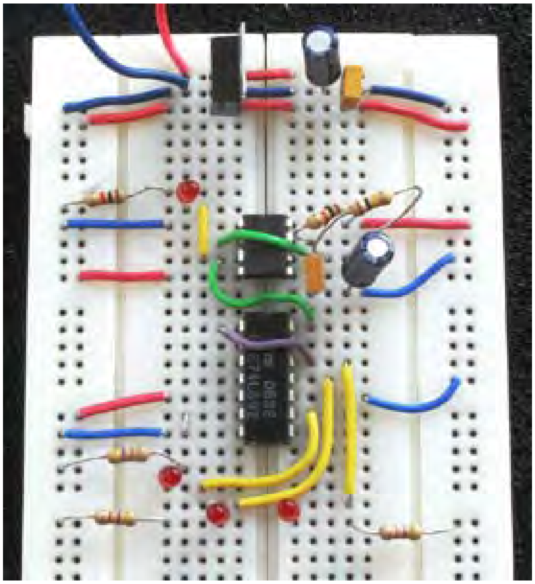

Figure 4-103. *The breadboard version of the schematic in Figure 4-102 to display the outputs from a 74LS92 counter.*

Now we come to the first new and difficult fact about the 74LSxx generation of TTL chips that makes them less desirable, for our purposes, than the 74HCxx generation of CMOS chips that I have recommended in previous projects. The modern and civilized HC chips will source 4mA or sink 4mA at each logical output, but the older LS generation is fussier. It will sink around 8mA into each output pin from a positive source, but when its output is high, it hardly gives you anything at all. This is a very basic principle:

- Outputs from TTL logic chips are designed to sink current.
- They are not designed to source significant current.

In fact, the 74LS92 is rated to deliver less than half a milliamp. This is quite acceptable when you're just connecting it with another logic chip, but if you want to drive an external device, it doesn't provide much to work with.

The proper solution is to say to the chip, "All right, we'll do it your way," and set things up with a positive source that flows through a load resistor to the LED that you want to use, and from there into the output from the chip. This is the "better" option shown in Figure 4-104.

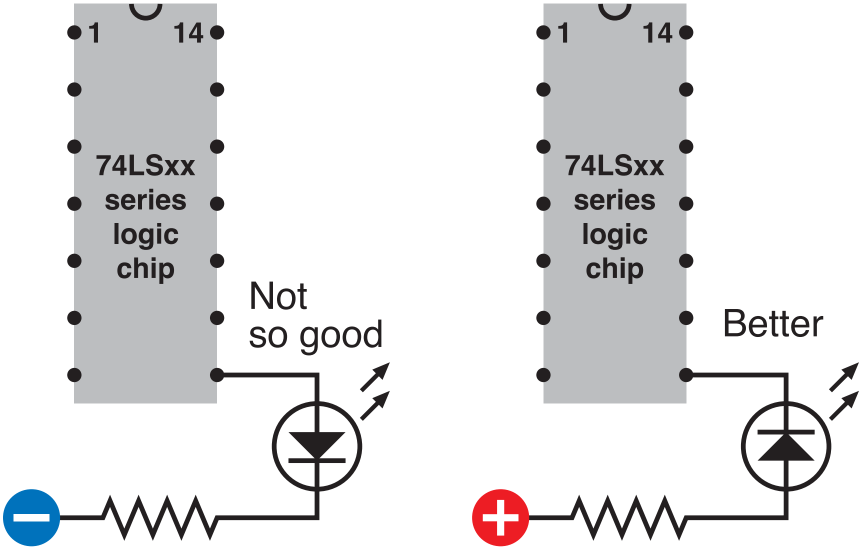

Figure 4-104. *Most TTL chips, including those in the LS generation, are unable to source much current from their logical output pins (left) and should usually be wired to sink current from a positive source (right).*

The only problem is that now the LED lights up when the counter's output is low. But the counter is designed to display its output in high pulses. So your LED is now off when it should be on, and on when it should be off.

You can fix this by passing the signal through an inverter, but already I'm getting impatient with this inconvenience. My way around the problem, at least for demo purposes, is to use the "Not so good" option in Figure 4-104 and make it work by connecting a very-low-current LED with a large 4K7 load resistor. This will enable us to "see" the output from the counter without asking it to give more than its rated limit, and if you want to create a more visibly powerful display for a finished version of the dice circuit, I'll deal with that later. According to my meter, the 4K7 resistor holds the current between 0.3mA and 0.4mA, which is the counter's rated maximum.

Set up your initial version of the circuit as shown in Figures 4-102 and 4-103. Be careful when you wire the positive and negative power supply to the counter chip, with its nonstandard pin assignments.

The 555 will run in astable mode, at about 1 pulse per second. This becomes the clock signal for the counter. The first three binary outputs from the counter then drive the three LEDs.

The counter advances when the input signal goes from *high* to *low*. So when the LED beside the 555 timer goes out, that's when the counter advances.

If you stare at the pattern generated by the outputs for long enough, you may be able to see the logic to it, bearing in mind that its zero state is when they are all off, and it counts up through five more steps before it repeats. The diagram in Figure 4-105 shows this sequence. If you want to know why the pattern works this way, check the following section, "Theory: Binary arithmetic."

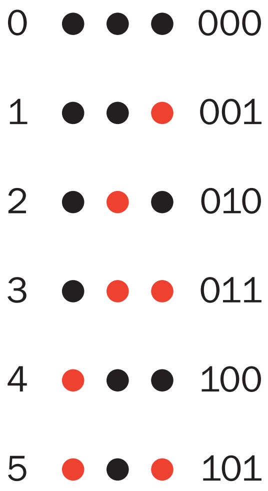

Figure 4-105. *The three output pins of the 74LS92 counter have high states shown by the red circles as the counter steps from 000 to 101 in binary notation.*

THEORY

Binary arithmetic

The rule for binary counting is just a variation of the rule that we normally use for everyday counting, probably without thinking much about it. In a 10-based system, we count from 0 to 9, then carry 1 over to the next position on the left, and go from 0 to 9 again in the right-most position. We repeat this procedure until we get to 99, then carry a 1 over to a new position to make 100, and continue counting.

In binary we do the same thing, except that we restrict ourselves to digits 0 and 1 only. So begin with 0 in the rightmost position, and count up to 1. As 1 is our limit, to continue counting we carry 1 over to the next place on the left, and start again from 0 in the right-most position. Count up to 1, then add 1 to the next place on the left—but, it already has a 1 in it, so it can't count any higher. So, carry 1 from there one space further, to the next place beside that—and so on.

If a glowing LED represents a 1, and a dark LED represents a 0, the diagram in Figure 4-105 shows how the 74LS92 counts up from 0 to (decimal) 6 or (binary) 101 in its inimitable fashion. I've also included a diagram in Figure 4-106 showing how a counter with four binary outputs would display decimal numbers from 0 through 15, again using the LEDs to represent 1s and 0s.

Here's a question for you: how many LEDs would you need to represent the decimal number 1024 in binary? And how many for 1023?

Obviously binary code is ideally suited to a machine full of logic components that either have a high or a low state. So it is that all digital computers use binary arithmetic (which they convert to decimal, just to please us).

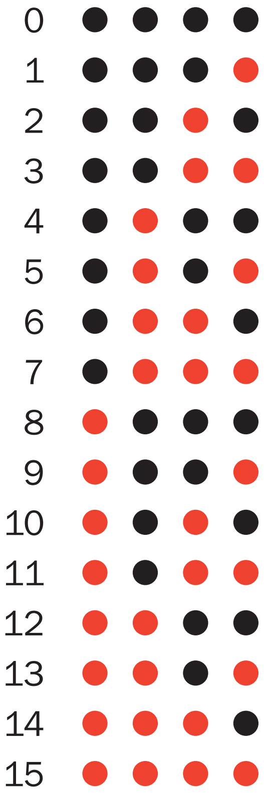

Figure 4-106. *A hexadecimal (16-based) binary counter would generate this succession of high states from its four output pins as it counts from 0 through 15 in decimal notation.*

Getting back to our project: I want to take the three binary outputs and make them create patterns like the spots on a die. How can I do this? Quite easily, as it turns out.

I'm assuming that I'll use seven LEDs to simulate the patterns of spots on a die. These patterns can be broken down into groups, which I have assigned to the three outputs from the counter in Figure 4-108. The first output (farthest to the right) can drive an LED representing the dot at the center of the die face. The second (middle) output can drive two more diagonal LEDs. The third output must switch on all four corner LEDs.

This will work for patterns 1 through 5, but won't display the die pattern for a 6. Suppose I tap into all three outputs from the counter with a three-input NOR gate. It has an output that goes high only when all three of its inputs are low, so it will only give a high output when the counter is beginning with all-low outputs. I can take advantage of this to make a 6 pattern.

Note that it's bad practice to mix the LS generation of TTL chips with the HC generation of CMOS chips, as their input and output ranges are different; so, the NOR chip has to be a 74LS27, not a 74HC27.

We're ready now for a simple schematic. In Figure 4-107 I've colored some of the wires just to make it easier for you to distinguish them. The colors have no other significance.

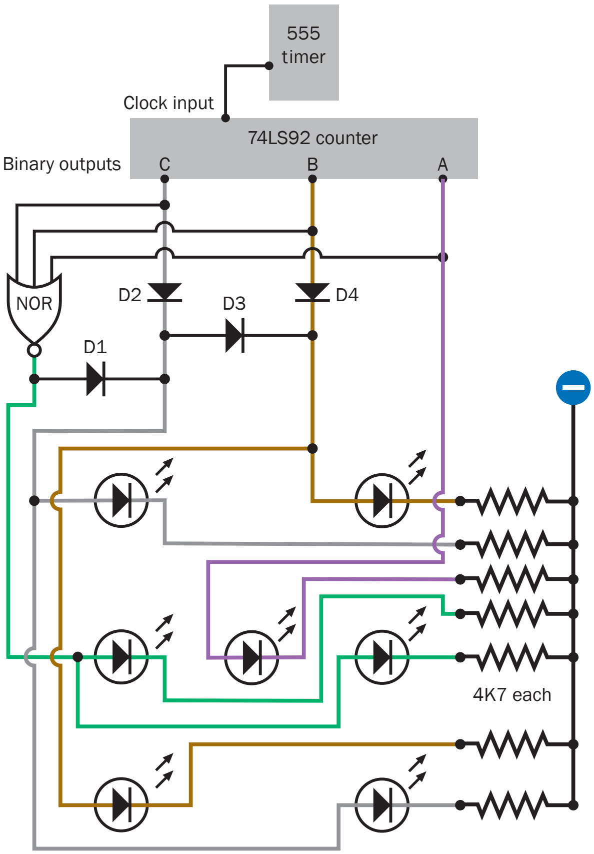
555
timer
Clock input
74LS92 counter
Binary outputs
C
B
A
NOR
D1
D2
D3
D4
4K7 each

Figure 4-107. A simplified schematic shows how outputs from the 74LS92 counter can be combined, with signal diodes and a single three-input NOR gate, to generate the spot patterns on a die. The wire colors have no special meaning and are used just to distinguish them from each other.

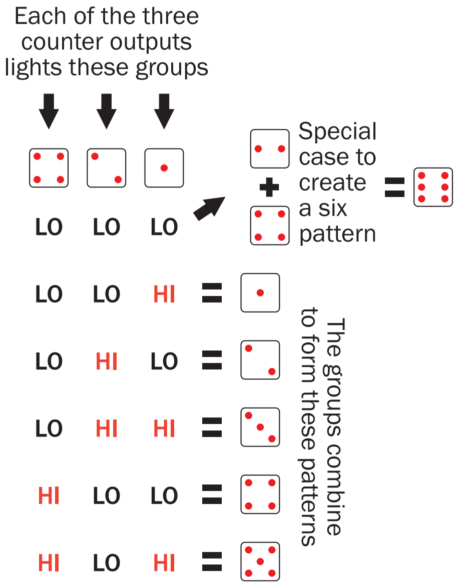
Each of the three
counter outputs
lights these groups
LO LO LO
Special
case to
create
a six
pattern
LO LO HI
LO HI LO
LO HI HI
HI LO LO
HI LO HI
The groups combine
to form these patterns

Figure 4-108. Binary outputs from the 74LS92 counter can be used to power LEDs arrayed in groups to simulate the pattern of spots on a die.

Each of the LEDs is grounded through a separate 4K7 load resistor. Unfortunately, this means that when they are displaying the pattern for a 6, all of them are running in parallel from the output of the NOR gate, which overloads it. As long as you don't leave the display in this mode for very long periods, it shouldn't cause a problem. You could compensate by increasing the load resistors, or by running pairs of the LEDs through one resistor, but this will make them so dim that they'll be difficult to see, as they're so close to their lower limit for current already.

Notice how I have added four signal diodes, D1 through D4. When Output C goes high, it has to illuminate all four corner LEDs, and so its power goes into the brown wire as well as the gray wire. But we must never allow one output to feed back into another, so D4 is needed to protect Output B when Output C is high.

Because there is now a connection between B and C, we need D2 to protect Output C when Output B is high. And because Output B must only feed two of the corner LEDs, we also need D3 to stop it from illuminating the other two. And, we have to protect the output from the NOR gate when either Output C or Output B is high. This requires D1.

Figure 4-109 shows everything that I've described so far assembled in breadboard format, while Figure 4-110 shows the test version that I built. Note that the unused logical inputs on the 74LS27 chip are shorted together and connected to the *positive* side of the power supply. Here's the rule:

- When using CMOS chips (such as the HC series), connect unused logical inputs to the negative side of the power supply.
- When using TTL chips (such as the LS series), connect unused logical inputs to the positive side of the power supply.

I assume that you have had enough fun watching the LEDs count slowly, so I've changed the capacitor and resistor values for the 555 to increase its speed from approximately 1 pulse per second to about 50,000 pulses per second. The counter could run much faster than this, but I just want it to cycle fast enough so that when the user presses and releases a button, the count will stop at an unforeseeable number.

The button starts and stops the 555 timer by applying and releasing power to the timing circuit only. This is the equivalent of shaking and then throwing the die.

While the counter is running fast, the LEDs are flashing so fast that all of them will seem to be on at once. At the same time, the circuit charges a new 68 µF capacitor, which I have added between the pushbutton and ground. When you release the button, this capacitor discharges itself through the 1K timing resistor. As the charge dissipates, the timing capacitor will take longer and longer to charge, and discharge, and the frequency of the 555 will gradually diminish. Consequently the LED display will also flash slower, like the reel on a Las Vegas slot machine gradually coming to a stop. This increases the tension as players can see the die display counting to the number that they're hoping for—and maybe going one step beyond it.

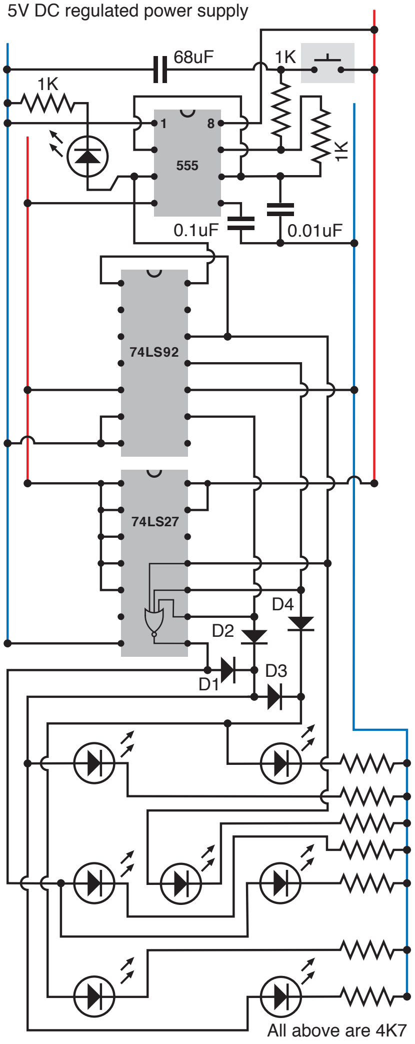

Figure 4-109. *With some extra components, the schematics from Figures 4-102 and 4-107 can be combined to make the working dice simulation.*

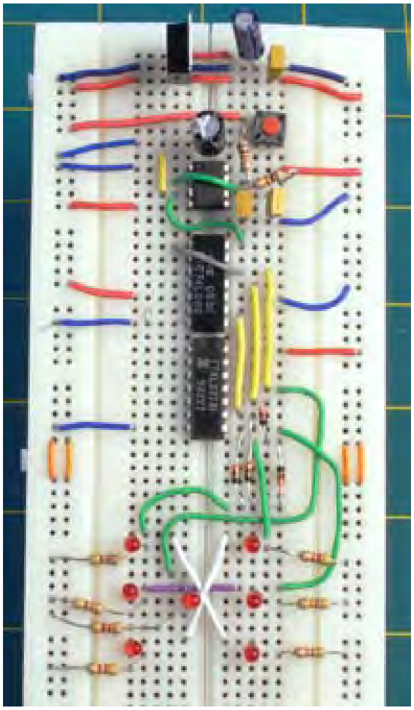

Figure 4-110. *The electronic dice schematic applied to a breadboard, with a pushbutton at the top to start and stop the counter, and 7 LEDs at the bottom to display the output.*

Note that to maximize this effect, the button has to be held down for a full second or more, so that the 68 μF capacitor becomes fully charged before the button is released.

So, this circuit now fulfills the original goal. But can it be better? Of course it can.

Enhancements

The main thing I want to improve is the brightness of the LEDs. I could add a transistor to amplify the current to each one, but there's a simpler alternative: a TTL "open collector" inverter.

I want to use an inverter because in the world of TTL, as I mentioned earlier, we can sink much more power into the output pin of a chip than we can source from it. So, I'm going to turn each LED the other way around and connect their load resistors to the positive side of the power supply. This way, they'll sink their power into the outputs of the inverter.

And the great advantage of an "open collector" version of the inverter chip is that it is designed to sink much more current than a normal TTL logic chip. It is rated for 40mA per pin. The only disadvantage is that it cannot source any current at all; instead of its output going high, it just behaves like an open switch. But that's OK for this circuit.

So the next and final schematic, in Figure 4-111, includes the 74LS06 inverter, which has also been added to the breadboarded version shown in Figure 4-112. I suggest that you put aside the little low-current LEDs and substitute some normal-size ones. Using Kingbright "standard" WP15031D 5mm LEDs, I find that each draws almost exactly 20mA with a voltage drop of about 2V with a 120 ohm series resistor. Because each output pin from the 74LS06 inverter powers no more than two LEDs at a time, this is exactly within its specification. I suggest that if you build this circuit, you check the consumption of your particular choice of LEDs and adjust the resistors if necessary.

Remember: to measure the voltage drop across an LED, simply touch the probes of your meter across it while it is illuminated. To measure the current, disconnect one side of the LED and insert the meter, in milliamp mode, between the leg of the LED and the contact that it normally makes in the circuit.

For a really dramatic display, you can get some 1 cm diameter LEDs (Figure 4-113). Check the specification, and you should find that many of these don't use more power than the usual 5 mm type. But whatever kind you use, don't forget to turn them around so that their negative sides face toward the inverter, and their positive sides face the resistors, which are connected to the positive side of the power supply.

One last detail: I had to add two 10K resistors to this version of the circuit. Can you see why? Diodes D1 through D4 are designed to transmit positive voltage through to the inverter when appropriate, but they prevent the inputs of the inverter from "seeing" the negative side of the power supply when the counter outputs are low. These inverter inputs require pull-down resistors to prevent them from "floating" and producing erroneous results.

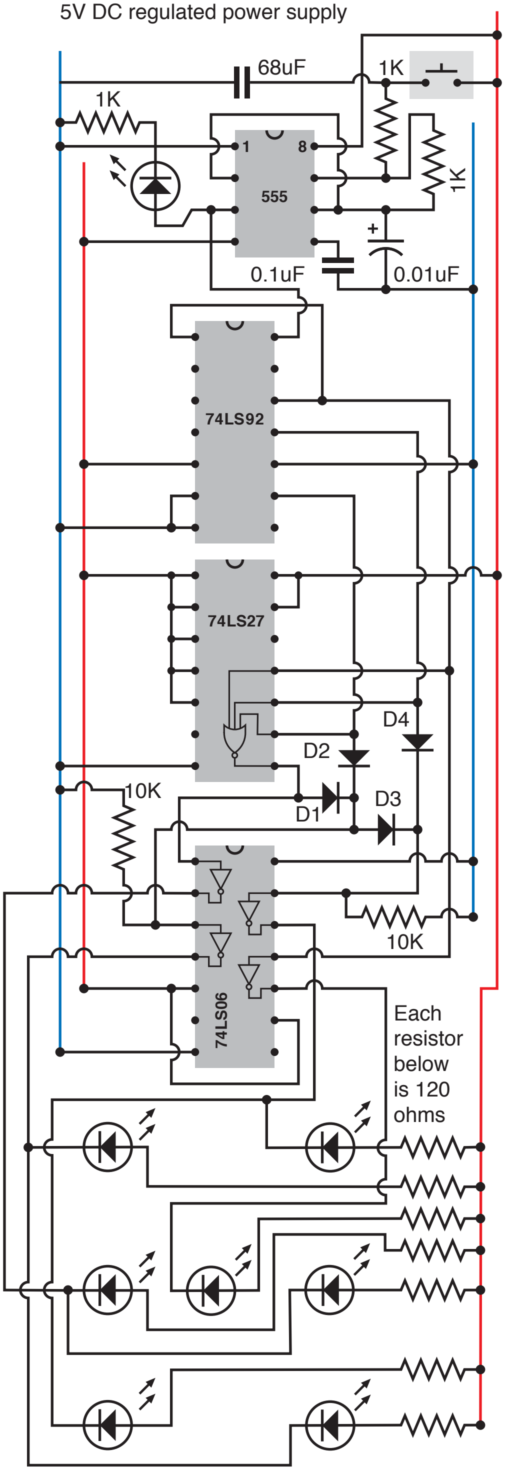
5V DC regulated power supply
68uF
1K
1K
1
8
555
1K
+
0.1uF
0.01uF
74LS92
74LS27
D4
D2
10K
D1
D3
10K
74LS06
Each resistor below is 120 ohms

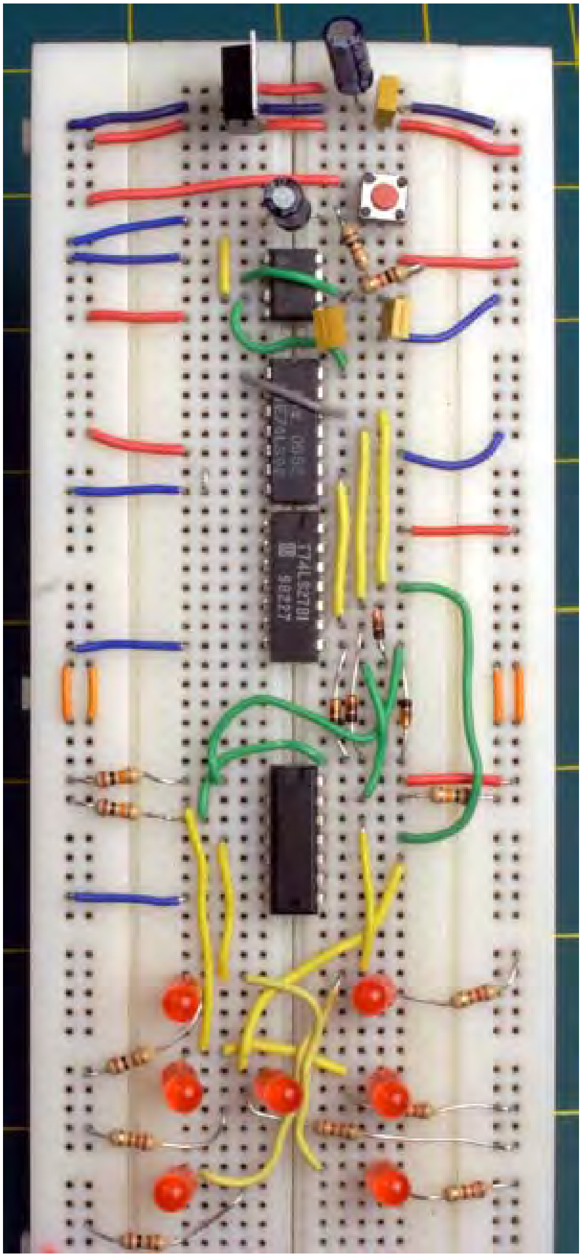

Figure 4-111. *If open-collector inverters are added to the dice schematic, it can drive full-size LEDs with up to 40mA, as long as the LEDs are turned around to sink current into the TTL output stage instead of trying to source current from it.*

Figure 4-112. *The completed circuit using an open-collector inverter to drive full-size LEDs.*

The final enhancements are up to you. Most obviously, you can add a second die, as many games require two dice. The 74LS27 chip still has a couple of spare NOR gates in it, one of which you can make use of, but you will need an additional 555 timer, running at a significantly different speed to ensure randomness, and it will have to drive a second counter.

After you get your dice up and running, you may want to test them for randomness. Because the pulses from a 555 timer are of equal length, every number has an equal chance of coming up; but the longer you hold down the Start button, the better your odds are of interrupting the counting process at a truly random moment. Anyone using your electronic dice should be told that "shaking" them for a full second is mandatory.

Of course, I could have simulated dice more easily by writing a few lines of software to generate random numbers on a screen, but even a fancy screen image cannot have the same appeal as a well-made piece of hardware. Figure 4-113 shows white 1 cm LEDs mounted in a sanded polycarbonate enclosure for dramatic effect.

Most of all, I derived satisfaction from using simple, dedicated chips that demonstrate the binary arithmetic that is fundamental in every computer.

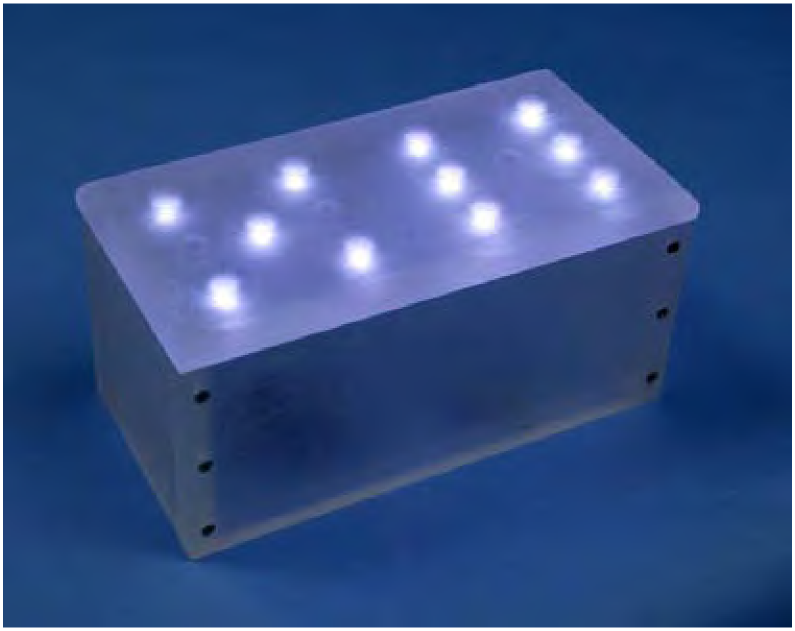

Figure 4-113. *The open-collector inverter chip in the dice circuit is sufficiently powerful to drive 1-cm white LEDs that draw about 20mA each, using a potential of 2V. In this finished version, the LEDs were embedded in cavities drilled from the underside of half-inch polycarbonate, which has been treated with an orbital sander to create a translucent finish.*

Experiment 24: Intrusion Alarm Completed

Now let me suggest how you can apply the knowledge from this chapter of the book to upgrade the burglar alarm project that was last modified in Experiment 15. You'll probably need to check Chapters 2 and 3 to refamiliarize yourself with some features of the alarm.

Upgrade 1: Delayed Activation

The biggest flaw in the alarm was that as soon as it was activated, it would immediately respond to any signal from the door and window sensors. It needed a feature to delay activation to give you a chance to exit from the building before the alarm armed itself. A 555 timer can provide this functionality, probably in conjunction with a relay. The power to the alarm should pass through the contacts of the relay, which are normally closed. When you press a button on the timer, it sends a positive pulse to the relay lasting for around 30 seconds, holding the relay open for that period. You could mount the timer in its own little box with a button on it, which you press when you're ready to leave the building. The 12-volt power supply to the burglar alarm passes through the box containing the delay circuit. For 30 seconds, the 555 interrupts power to the alarm, and then restores it, ready for action.

Upgrade 2: Keypad Deactivation

This is now really simple. You can substitute a latching relay instead of the switch, S1, on the alarm box (shown in Figure 3-110), and use a keypad to set and reset the relay in exactly the same way as in the combination lock in Experiment 20. You'll have to run an additional three wires from the relay, out of the alarm box, to the keypad (one supplying power to the "on" relay coil, another supplying power to the "off" coil, the third being a common ground). You can either use a 9-volt battery to power the electronics in association with the keypad, or run an additional fourth wire from the alarm box, to carry positive power to the logic chips, bearing in mind that you have to insert a voltage regulator at some point, to drop the 12 volts that the alarm uses to the 5 volts that the logic gates require. As the gates consume so little power, the 12-to-5 drop should be OK for the voltage regulator; it won't have to dissipate too much heat.

With this additional feature, you can use the alarm like this:

- Press the pound key on the keypad to flip the latching relay into its "on" mode, so that it passes power to the alarm box, which is now armed.
- If you want to leave the house, push the button on the delay unit to give you 30 seconds in which to do so.
- If the alarm is triggered, enter your secret code on the keypad to deactivate it by flipping the latching relay to its "off" position and cutting power to the alarm box.

These modifications are so simple that I think the block diagram in Figure 4-114 should be all you need. I don't think I need to give you any schematics. The only change you have to make to the existing alarm is to substitute the latching relay for the on/off switch.

But, there is still one obvious necessary enhancement needed: how can you get back into the house without instantly triggering the alarm?

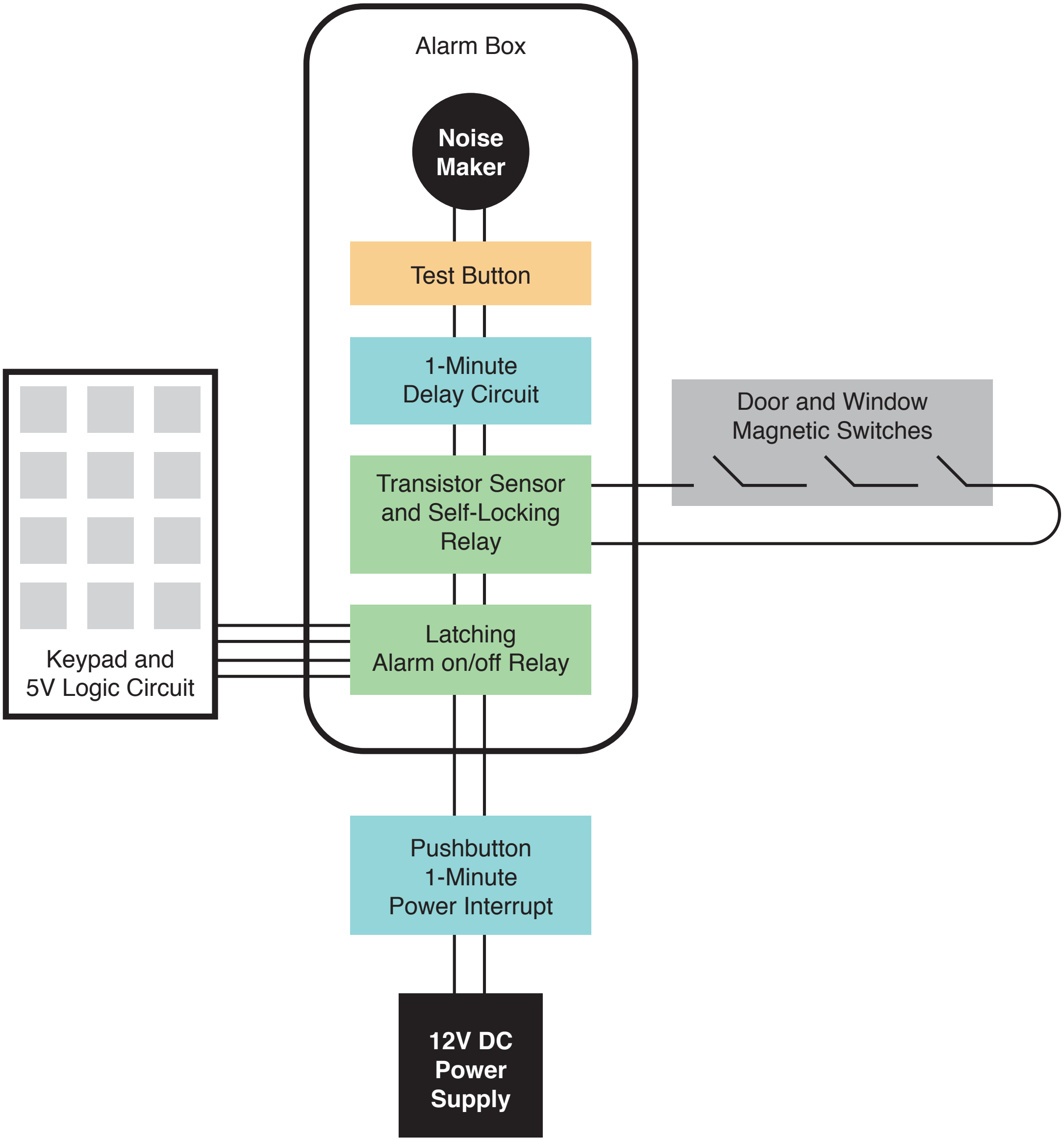
Alarm Box
Noise Maker
Test Button
1-Minute Delay Circuit
Door and Window Magnetic Switches
Transistor Sensor and Self-Locking Relay
Keypad and 5V Logic Circuit
Latching Alarm on/off Relay
Pushbutton 1-Minute Power Interrupt
12V DC Power Supply

Figure 4-114. *This block diagram shows the relative placement of the old and new components. The pushbutton power interrupt (which allows you to leave the house before the alarm switches itself back on) goes between the power supply and everything else. The latching relay substitutes for the DPDT switch on the previous version of the alarm. The transistor and self-locking relay, connected with the door and window magnetic switches, remain unchanged. The new delay circuit is inserted between the self-locking relay and the noisemaker. The test button is wired with the latching relay in the same way that it was wired previously with the DPDT switch.*

Upgrade 3: Delay Before Deactivation

Typically, alarms include another delay feature. When you open a door on your way into the building and it triggers the alarm, you have 30 seconds to deactivate it, before it starts making a noise.

How can we implement this delay feature? If I try to use another 555 timer to generate a pulse to inhibit the noise, that won't work, because the output from either the transistor or the relay can continue indefinitely. The relay locks itself on, and the transistor can continue passing voltage for as long as someone leaves a door open. If either of these signals activates a timer in monostable mode, the pulse from the timer will never end, and it will suppress the alarm indefinitely.

I think what I have to do is use a resistor and a capacitor to create a delay. I'll power them through the existing relay, so that I can be sure that they'll receive the full voltage of the power supply, after beginning from zero. Gradually the capacitor will acquire voltage—but I can't connect this directly to the noisemaker, because the noisemaker will gradually get louder as the voltage increases.

I have to insert a device that will be triggered to give full voltage when the input rises past a certain point. To do this, I'll use a 555 timer that's wired in bistable mode. This kind of modification is generally known as a "kludge," because it's not elegant, uses too many components, and does not use them appropriately. What I really need is a comparator, but I don't have space to get into that topic. So, using the knowledge that you have so far, the schematic in Figure 4-115 shows how a delay can be added to the alarm—not elegantly, but reliably.

The only problem is that if you power up a 555 timer in bistable mode, there's a 50-50 chance that the timer starts itself with its output high or low. So I need to pull the voltage low on the reset pin (to start the timer with its output inhibited) and gradually let it become positive (to permit the output). At the same time, I want to start with the voltage high on the trigger pin and gradually lower it, until it falls below 1/3 of the power supply and triggers the output.

So there are two timing circuits. The one for the reset pin works faster than the one on the trigger pin, so that at the point when the timer is triggered, the reset won't stop it.

The schematic shows component values that will do this. The 10 µF capacitor starts low but is charged through the 10K resistor in a couple of seconds. The timer is now ready to be triggered. But the 68 µF capacitor starts high (being connected with the positive side of the power supply) and takes a full minute to be pulled down to 1/3 of supply voltage through the 1M resistor. At that point, its voltage is low enough to trigger the 555. The timer output goes high and supplies the noisemaker.

You should be able to insert this little delay module in your alarm box, between the output from the relay and the input to the noisemaker without too much trouble. And if you want to adjust the delay, just use a higher or lower value resistor than 1M.

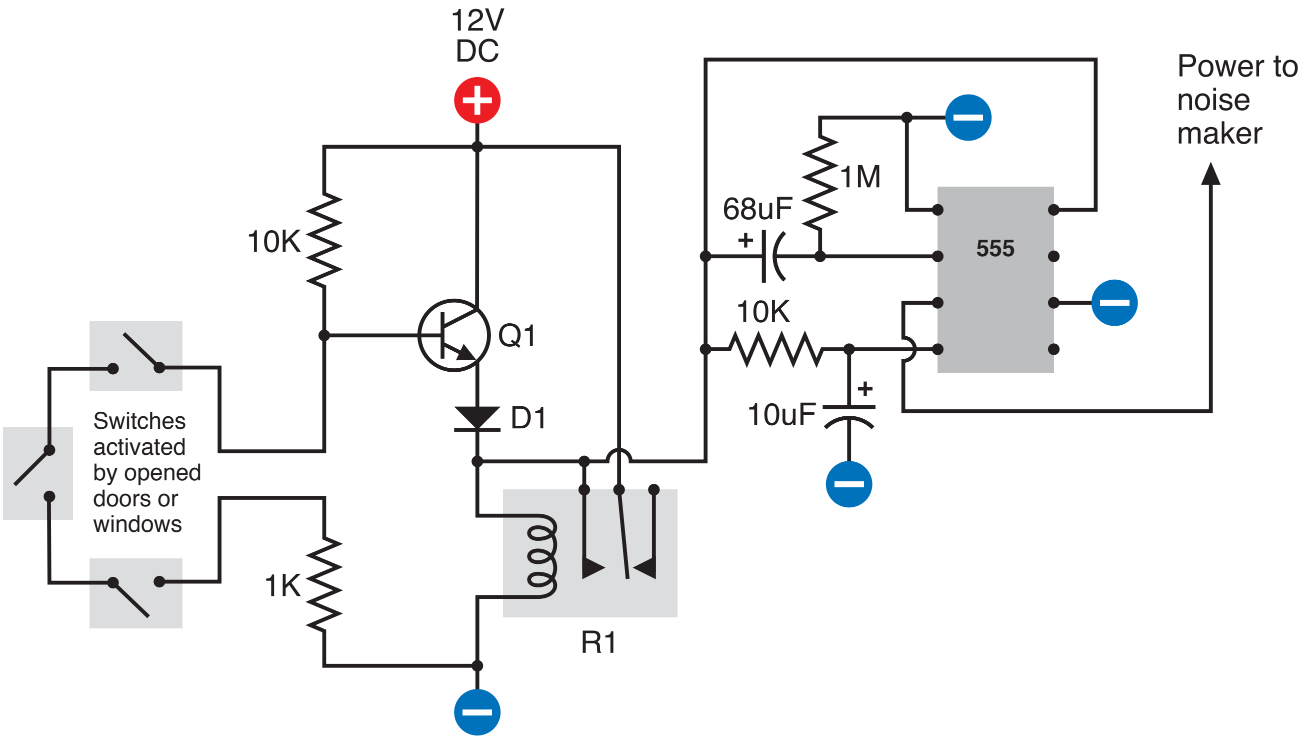

Figure 4-115. *This addition to the original alarm circuit imposes a one-minute delay before the alarm sounds. The 555 timer (wired in bistable mode) receives power through relay R1. The lower timing circuit initially applies negative voltage to the reset, ensuring that the 555 powers up with its output suppressed. This voltage quickly rises. Meanwhile the upper timing circuit applies a voltage to the trigger that gradually diminishes as the 68 µF capacitor equalizes its charge through the 1M resistor. When the voltage diminishes to 1/3 of the supply, the timer's output goes high and starts the noisemaker. If the power to the circuit is interrupted at any time before this, the relay relaxes, the capacitors gradually discharge, and the alarm does not sound.*

The Wrap-Up

If you add these three enhancements, your alarm will have all the features on my original wish list. Of course, if I were designing it from scratch, with all the information that has been added in this chapter of the book, it could be more elegant. But the modifications have not entailed making destructive changes to our original project, and all the design goals have been met.

5

What Next?

At this point, we can branch out in numerous directions. Here are some possibilities:

Audio electronics
This is a field in itself, including hobby projects, such as simple amplifiers and "stomp boxes," to modify guitar sound.

Radio-frequency devices
Anything that receives or transmits radio waves, from an ultra-simple AM radio to remote controllers.

Motors
The field of robotics has encouraged the growth of many online sites selling stepper motors, gear motors, synchronous motors, servo motors, and more.

Programmable microcontrollers
These are tiny computers on a single chip. You write a little program on your desktop computer, which will tell the chip to follow a series of procedures, such as receiving input from a sensor, waiting for a fixed period, and sending output to a motor. Then you download your program onto the chip, which stores it in nonvolatile memory. Popular controllers include the PICAXE, BASIC Stamp, Arduino, and many more. The cheapest ones retail for a mere $5 each.

Obviously, I don't have space to develop all of these topics fully, so what I'm going to do is introduce you to them by describing just one or two projects in each category. You can decide which interests you the most, and then proceed beyond this book by reading other guides that specialize in that interest.

I'm also going to make some suggestions about setting up a productive work area, reading relevant books, catalogs, and other printed sources, and generally proceeding further into hobby electronics.

IN THIS CHAPTER

Shopping List: Experiments 25 Through 36
Customizing Your Work Area
Reference Sources
Experiment 25: Magnetism
Experiment 26: Tabletop Power Generation
Experiment 27: Loudspeaker Destruction
Experiment 28: Making a Coil React
Experiment 29: Filtering Frequencies
Experiment 30: Fuzz
Experiment 31: One Radio, No Solder, No Power
Experiment 32: A Little Robot Cart
Experiment 33: Moving in Steps
Experiment 34: Hardware Meets Software
Experiment 35: Checking the Real World
Experiment 36: The Lock, Revisited
In Closing

Shopping List: Experiments 25 Through 36

Tools

You won't need any new tools for this section of the book.

Supplies and Components

As we have progressed to the point where you may want to pick and choose which projects you attempt, I will list the supplies and components at the beginning of each experiment.

Customizing Your Work Area

At this point, if you're getting hooked on the fun of creating hardware but haven't allocated a permanent corner to your new hobby, I have some suggestions. Having tried many different options over the years, my main piece of advice is this: don't build a workbench!

Many hobby electronics books want you to go shopping for 2×4s and plywood, as if a workbench has to be custom-fabricated to satisfy strict criteria about size and shape. I find this puzzling. To me, the exact size and shape of a bench is not very important. I think the most important issue is storage.

I want tools and parts to be easily accessible, whether they're tiny transistors or big spools of wire. I certainly don't want to go digging around on shelves that require me to get up and walk across the room.

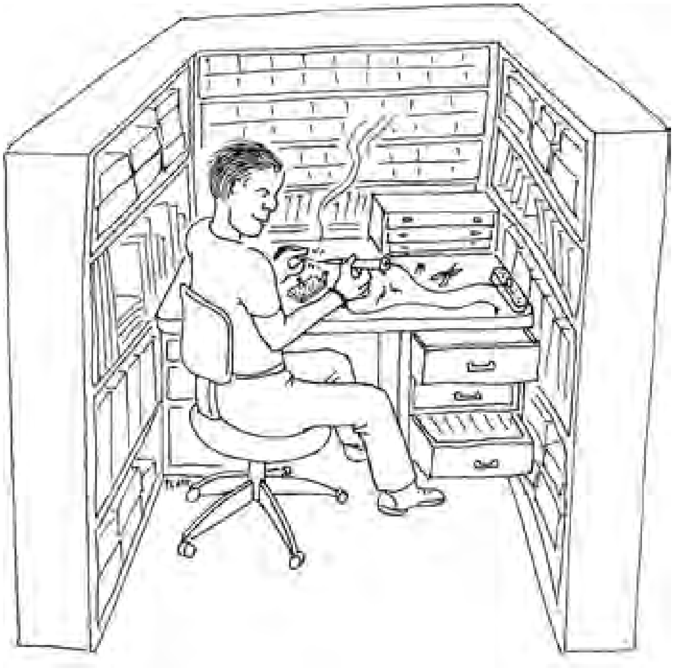

Figure 5-1. *The ideal work area: surrounded by storage. Never again will you need to get out of your chair!*

This leads me to two conclusions:

1. You need storage above the workbench.
2. You need storage below the workbench.

Many DIY workbench projects allow little or no storage underneath. Or, they suggest open shelves, which will be vulnerable to dust. My minimum configuration would be a pair of two-drawer file cabinets with a slab of 3/4-inch plywood or a Formica-clad kitchen countertop placed across them. File cabinets are ideal for storing all kinds of objects, not just files.

Of all the workbenches I've used, the one I liked best was an old-fashioned steel office desk—the kind of monster that dates back to the 1950s. They're difficult to move (because of their weight) and don't look beautiful, but you can buy them cheaply from used office furniture dealers, they're generous in size, they withstand abuse, and they last forever. The drawers are deep and usually slide in and out smoothly, like good file-cabinet drawers. Best of all, the desk has so much steel in it that you can use it to ground yourself before touching components that are sensitive to static electricity. If you use an anti-static wrist strap, you can simply attach it to a sheet-metal screw that you drive into one corner of the desk.

What will you put in the deep drawers of your desk or file cabinets? Some paperwork may be useful, perhaps including the following documents:

- Product data sheets
- Parts catalogs
- Sketches and plans that you draw yourself

The remaining capacity of each drawer can be filled with plastic storage boxes. The boxes can contain tools that you don't use so often (such as a heat gun or a high-capacity soldering iron), and larger-sized components (such as loudspeakers, AC adapters, project boxes, and circuit boards). You should look for storage boxes that measure around 11 inches long, 8 inches wide, and 5 inches deep, with straight sides. Boxes that you can buy at Wal-Mart will be cheaper, but they often have tapering sides (which are not space-efficient).

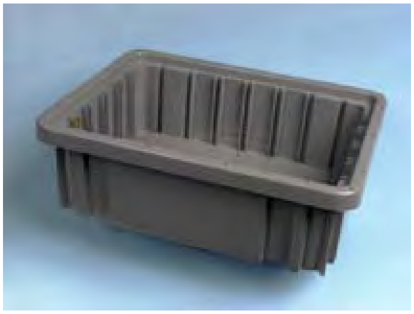

Figure 5-2. *Akro-Grid boxes contain grooves allowing them to be partitioned into numerous compartments for convenient parts storage.*

The boxes that I like best are Akro-Grids, made by Akro-Mils (see Figures 5-2 and 5-3). These are very rugged, straight-sided, with optional transparent snap-on lids. You can download the full Akro-Mills catalog from *http://www.akro-mils.com* and then search online for retail suppliers. You'll find that Akro-Mils also sells an incredible variety of parts bins, but I don't like open bins because their contents are vulnerable to dust and dirt.

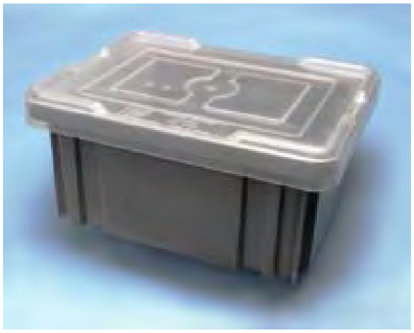

Figure 5-3. *Lids are sold separately for Akro-Grid boxes to keep the contents dust-free. The height of the box in Figure 5-2 allows three to be stacked in a typical file-cabinet drawer. The box shown here allows two to be stacked.*

For medium-size components, such as potentiometers, power connectors, control knobs, and toggle switches, I like storage containers measuring about 11 inches long, 8 inches wide, and 2 inches deep , divided into four to six sections. You can buy these from Michaels (the craft store), but I prefer to shop online for the Plano brand, as they seem more durably constructed. The Plano products that are most suitable for medium-size electronic parts are classified as fishing-tackle boxes, and you'll see them at *http://www.planomolding.com/tackle/products.asp*.

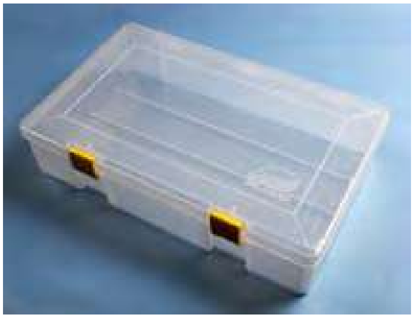

Figure 5-4. *This Plano brand box is undivided, making it useful for storing spools of wire or medium-size tools. When stacked upright on its long edge, three will fit precisely in a file-cabinet drawer.*

For undivided, flat-format storage boxes, the Prolatch 23600-00 is ideally sized to fit a file-cabinet drawer, and the latches are sufficiently secure that you could stack a series of them on their long edges. See Figure 5-4.

Plano also sells some really nicely designed toolboxes, one of which you can place on your desktop. It will have small drawers for easy access to screwdrivers, pliers, and other basics. Because you need a work area that's only about three feet square for most electronics projects, surrendering some desk space to a toolbox is not a big sacrifice.

If you have a steel desk with relatively shallow drawers, one of them can be allocated for printed catalogs. Don't underrate the usefulness of hard copy, just because you can buy everything online. The Mouser catalog, for instance, has an index, which is better in some ways than their online search feature, and the catalog is divided into helpful categories. Many times I've found useful parts that I never knew existed, just by browsing, which is much quicker than flipping through PDF pages online, even with a broadband connection. Currently, Mouser is still quite generous about sending out their catalogs, which contain over 2,000 pages. McMaster-Carr will also send you a catalog, but only after you've ordered from them, and only once a year.

Now, the big question: how to store all those dinky little parts, such as resistors, capacitors, and chips? I've tried various solutions to this problem. The most obvious is to buy a case of small drawers, each of which is removable, so you can place it on your desk while you access its contents. But I don't like this system, for two reasons. First, for very small components, you need to subdivide the drawers, and the dividers are never secure. And second, the removability of the drawers creates the risk of accidentally emptying the contents on the floor. Maybe you're too careful to allow this to happen, but I'm not!

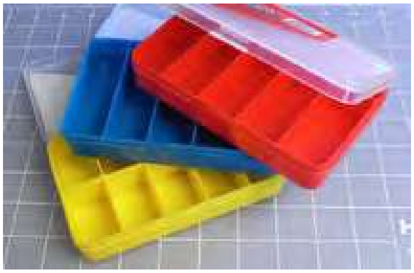

Figure 5-5. *Darice Mini-Storage boxes are ideal for components such as resistors, capacitors, and semiconductors. The boxes can be stacked stably or stored on shelves, with their ends labeled. The brand sticker is easily removed after being warmed with a heat gun.*

My personal preference is to use Darice Mini-Storage boxes, shown in Figure 5-5. You can find these at Michaels in small quantities, or buy them more economically in bulk online from suppliers such as *http://www.craftamerica.com*. The blue boxes are subdivided into five compartments that are exactly the right size and shape for resistors. The yellow boxes are subdivided into ten compartments, which are ideal for semiconductors. The purple boxes aren't divided at all, and the red boxes have a mix of divisions.

The dividers are molded into the boxes, so you don't have the annoyance associated with removable dividers that slip out of position, allowing components to mix together. The box lids fit tightly, so that even if you drop one of the boxes, it probably won't open. The lids have metal hinges, and a ridge around the edge that makes the boxes securely stackable.

I keep my little storage boxes on a set of shelves above the desk, with a gap of 3 inches between one shelf and the next, allowing two boxes to be stacked on each shelf. If I want to work with a particular subset of boxes, I shift them onto the desktop and stack them there.

Labeling

No matter which way you choose to store your parts, labeling them is essential. Any ink-jet printer will produce neat-looking labels, and if you use peelable (nonpermanent) labels, you'll be able to reorganize your parts in the future, as always seems to become necessary. I use color-coded labels for my collection of resistors, so that I can compare the stripes on a resistor with the code on the label, and see immediately if the resistor has been put in the wrong place. See Figure 5-6.

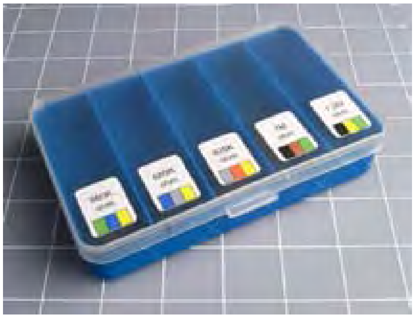

Figure 5-6. *To check that resistors are not placed in the wrong compartments, print the color code on each label.*

Even more important: you need to place a second (non-adhesive) label inside each compartment with the components. This label tells you the manufacturer's part number and the source, so that reordering is easy. I buy a lot of items from Mouser, and whenever I open their little plastic bags of parts, I snip out the section of the bag that has the identifying label on it, and slide it into the compartment of my parts box before I put the parts on top of it. This saves frustration later.

If I were *really* well organized, I would also keep a database on my computer listing everything that I buy, including the date, the source, the type of component, and the quantity. But I'm not that well organized.

On the Bench

Some items are so essential that they should sit on the bench or desktop on a permanent basis. These include your soldering iron(s), helping hands with magnifier, desk lamp, breadboard, power strip, and power supply. For a desk lamp, I prefer the type that has a daylight-spectrum fluorescent bulb, because it spreads a uniform light and helps me to distinguish colors of adjacent stripes on resistors.

The power supply is a matter of personal preference. If you're serious about electronics, you can buy a unit that delivers properly smoothed current at a variety of properly regulated and calibrated voltages. Your little wall-plug unit from RadioShack cannot do any of these things, and its output may vary depending on how heavily you load it. Still, as you've seen, it is sufficient for basic experiments, and when you're working with logic chips, you need to mount a 5-volt regulator on your breadboard anyway. Overall, I consider a good power supply optional.

Another optional item is an oscilloscope. This will show you, graphically, the electrical fluctuations inside your wires and components, and by applying probes at different points, you can track down errors in your circuit. It's a neat gadget to own, but it will cost a few hundred dollars, and for our tasks so far, it has not been necessary. If you plan to get seriously into audio circuits, an oscilloscope becomes far more important, because you'll want to see the shapes of the waveforms that you generate.

You can try to economize on an oscilloscope by buying a unit that plugs into the USB port of your computer and uses your computer monitor to display the signal. I have tried one of these, and was not entirely happy with the results. It worked, but did not seem accurate or reliable for low-frequency signals. Maybe I was unlucky; I decided not to try any other brands.

The surface of your desk or workbench will undoubtedly become scarred by random scuffs, cut marks, and drops of molten solder. I use a piece of half-inch plywood, two feet square, to protect my primary work area, and I clamp a miniature vise to its edge. To reduce the risk of static electricity when working with sensitive components, I cover the plywood with a square of conductive foam. This is not cheap, but offers advantages in addition to protecting chips from being zapped. Instead of scattering stray components, I can stick them into the foam, like plants growing in a garden. And like a garden, I can divide it into sections, with resistors on one side, capacitors on the other, and chips straight ahead.

Inevitably, during your work you'll create a mess. Little pieces of bent wire, stray screws, fasteners, and fragments of stripped insulation tend to accumulate, and can be a liability. If metal parts or fragments get into a project that you're building, they can cause short circuits. So you need a trash container. But it has to be easy to use. I use a full-size garbage pail, because it's so big that I can't miss it when I throw something toward it, and I can never forget that it's there.

Last, but most essential: a computer. Now that all data sheets are available online, and all components can be ordered online, and many sample circuits are placed online by hobbyists and educators, I don't think anyone can work efficiently without quick Internet access. To avoid wasting space, I suggest you use a small, cheap laptop that has a minimal footprint. A possible workbench configuration, using a steel desk, is shown in Figure 5-7.

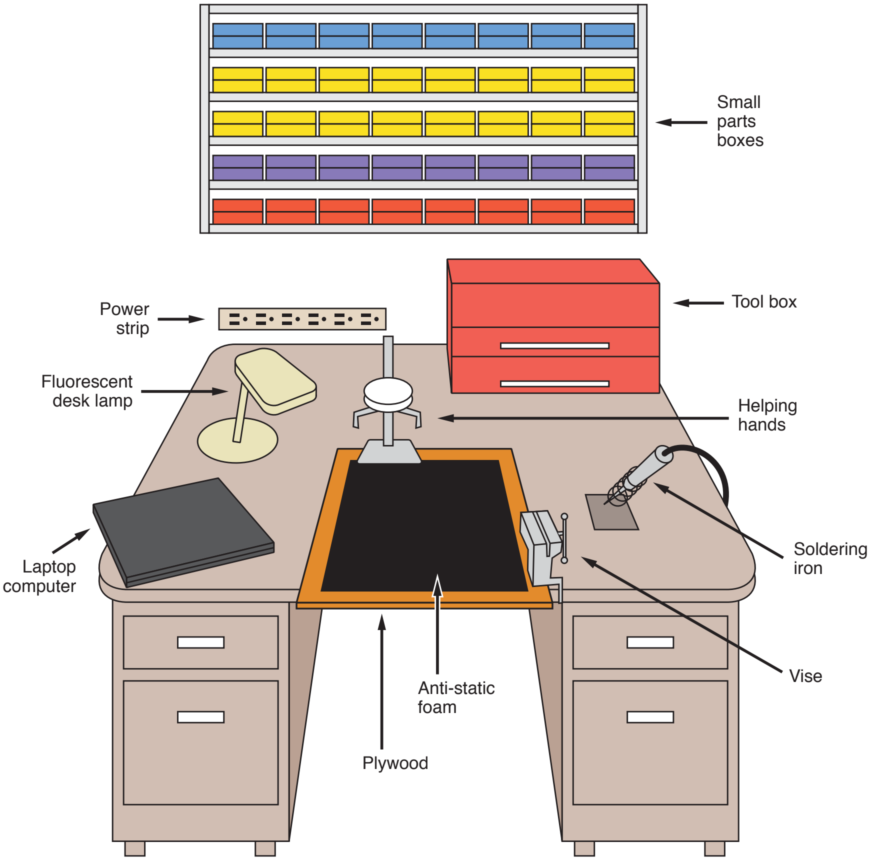

Figure 5-7. *An old steel office desk can be as good as, if not better than, a conventional workbench when building small electronics projects. It provides a large work area and ample storage, and has sufficient mass for you to ground yourself when dealing with components that are sensitive to static electricity.*

Reference Sources

Online

My favorite educational and reference site is Doctronics (*http://www.doctronics.co.uk*). I like the way they draw their schematics, and I like the way they include many illustrations of circuits on breadboards (which most sites don't bother to do). They also sell kits, if you're willing to pay and wait for shipping from the UK. Part of a page from the doctronics website is reproduced in Figure 5-8.

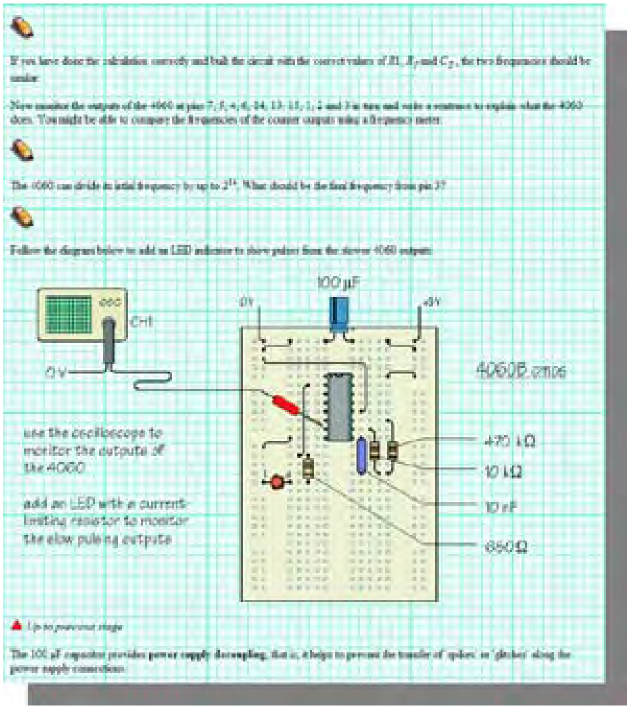
100 µF
0V
+9V
OSC
CH1
0V
4060B cmos
use the oscilloscope to monitor the outputs of the 4060
add an LED with a current-limiting resistor to monitor the slow pulsing outputs
470 kΩ
10 kΩ
10 nF
680Ω

Figure 5-8. *A sample page from http://www.doctronics.co.uk shows their detailed instructional approach. This is a valuable free online resource.*

My next favorite hobby site is also British-based: the Electronics Club (*http://www.kpsec.freeuk.com*). It's not as comprehensive as Doctronics, but very friendly and easy to understand.

For a more theory-based approach, try *http://www.electronics-tutorials.ws*. This will go a little farther than the theory sections I've included here.

For an idiosyncratic selection of electronics topics, try Don Lancaster's Guru's Lair (*http://www.tinaja.com*). Lancaster wrote *The TTL Cookbook* more than 30 years ago, which opened up electronics to at least two generations of hobbyists and experimenters. He knows what he's talking about, and isn't afraid of getting into some fairly ambitious areas such as writing his own PostScript drivers and creating his own serial-port connections. You'll find a lot of ideas there.

Books

Yes, you do need books. As you're already reading this one, I won't recommend other beginners' guides. Instead, in keeping with the orientation of this chapter, I'll suggest some titles that will take you farther in various directions, and can be used for reference. I own all of these myself, and find them valuable:

Practical Electronics for Inventors, by Paul Scherz (McGraw-Hill, Second Edition, 2007)

> This is a massive, comprehensive book, well worth the $40 cover price. Despite its title, you won't need to invent anything to find it useful. It's my primary reference source, covering a wide range of concepts, from the basic properties of resistors and capacitors all the way to some fairly high-end math. If you buy only one book (in addition to this one, of course!), this would be my recommendation.

Getting Started with Arduino, by Massimo Banzi (Make: Books, 2009)

> If you enjoy the simplicity and convenience of the PICAXE programmable microcontroller that I describe later in this chapter, you'll find that the Arduino can do a lot more. *Getting Started* is the simplest introduction around, and will help to familiarize you with the Processing language used in Arduino (similar to the C language, not much like the version of BASIC used by the PICAXE).

Making Things Talk, by Tom Igoe (Make: Books, 2007)

> This ambitious and comprehensive volume shows how to make the most of the Arduino's ability to communicate with its environment, even getting it to access sites on the Internet.

TTL Cookbook, by Don Lancaster (Howard W. Sams & Co, 1974)

> The 1974 copyright date is not a misprint! You may be able to find some later editions, but whichever one you buy, it will be secondhand and possibly expensive, as this title now has collectible value. Lancaster wrote his guide before the 7400 series of chips was emulated on a pin-for-pin basis by CMOS versions, but it's still a good reference, because the concepts and part numbers haven't changed, and his writing is so accurate and concise.

CMOS Sourcebook, by Newton C. Braga (Sams Technical Publishing, 2001)

> This book is entirely devoted to the 4000 series of CMOS chips, not the 74HC00 series that I've dealt with primarily here. The 4000 series is older and must be handled more carefully, because it's more vulnerable to static electricity than the generations that came later. Still, the chips remain widely available, and their great advantage is their willingness to tolerate a wide voltage range, typically from 5 to 15 volts. This means you can set up a 12-volt circuit that drives a 555 timer, and use output from the timer to go straight into CMOS chips (for example). The book is well organized in three sections: CMOS basics, functional diagrams (showing pinouts for all the main chips), and simple circuits showing how to make the chips perform basic functions.

The Encyclopedia of Electronic Circuits, by Rudolf F. Graf (Tab Books, 1985)

A totally miscellaneous collection of schematics, with minimal explanations. This is a useful book to have around if you have an idea and want to see how someone else approached the problem. Examples are often more valuable than general explanations, and this book is a massive compendium of examples. Many additional volumes in the series have been published, but start with this one, and you may find it has everything you need.

The Circuit Designer's Companion, by Tim Williams (Newnes, Second Edition, 2005)

Much useful information about making things work in practical applications, but the style is dry and fairly technical. May be useful if you're interested in moving your electronics projects into the real world.

The Art of Electronics, by Paul Horowitz and Winfield Hill (Cambridge University Press, Second Edition, 1989)

The fact that this book has been through 20 printings tells you two things: (1) Many people regard it as a fundamental resource; (2) Secondhand copies should be widely available, which is an important consideration, as the list price is over $100. It's written by two academics, and has a more technical approach than *Practical Electronics for Inventors*, but I find it useful when I'm looking for backup information.

Getting Started in Electronics, by Forrest M. Mims III (Master Publishing, Fourth Edition, 2007)

Although the original dates back to 1983, this is still a fun book to have. I think I have covered many of its topics here, but you may benefit by reading explanations and advice from a completely different source, and it goes a little farther than I have into some electrical theory, on an easy-to-understand basis, with cute drawings. Be warned that it's a brief book with eclectic coverage. Don't expect it to have all the answers.

Making Things
Getting Started with Arduino
Make:

Figure 5-9. *These books from MAKE provide guidance if you want to go beyond basic microcontrollers into the more exotic realms of the Arduino chip.*

TTL Cookbook
HOROWITZ AND HILL
THE ART OF ELECTRONICS
PRACTICAL ELECTRONICS FOR INVENTORS

Figure 5-10. *A sun-damaged copy of the Don Lancaster's classic TTL Cookbook, a 2,000-page catalog from the Mouser Electronics supply company, and two comprehensive reference books that can provide years of additional guidance in all areas of electronics.*

Experiment 25: Magnetism

This experiment should be a part of any school science class, but even if you remember doing it, I suggest that you do it again, because setting it up takes only a matter of moments, and it's going to be our entry point to a whole new topic: the relationship between electricity and magnetism. Quickly this will lead us into audio reproduction and radio, and I'll describe the fundamentals of self-inductance, which is the third and final basic property of passive components (resistance and capacitance being the other two). I left self-inductance until last because it's not very relevant to the experiments that you've done so far. But as soon as we start dealing with analog signals that fluctuate, it becomes essential.

You will need:

- Large screwdriver.
- 22-gauge wire (or thinner). Quantity: 6 feet.
- AA battery.

Procedure

This couldn't be simpler. Wind the wire around the shaft of the screwdriver, near its tip. The turns should be neat and tight and closely spaced, and you'll need to make 100 of them, within a distance of no more than 2 inches. To fit them into this space, you'll have to make turns on top of previous turns. If the final turn tends to unwind itself (which will happen if you're using stranded wire), secure it with a piece of tape. See Figure 5-11.

Now apply a battery, as shown in Figure 5-12. At first sight, this looks like a very bad idea, because you're going to short out your battery just as you did in Experiment 2. But by passing the current through a wire that's coiled instead of straight, we'll get some work out of it before the battery expires.

Put a small paper clip near the screwdriver blade, on a soft, smooth surface that will not present much friction. A tissue works well. Because many screwdrivers are already magnetic, you may find that the paper clip is naturally attracted to the tip of the blade. If this happens, move the clip just outside the range of attraction. Now apply the 1.5 volts to the circuit, and the clip should jump to the tip of the screwdriver. Congratulations: you just made an electromagnet.

FUNDAMENTALS

A two-way relationship

Every electric motor that was ever made uses some aspect of the relationship between electricity and magnetism. It's absolutely fundamental in the world around us. Remember that electricity can create magnetism:

When electricity flows through a wire, it creates a magnetic force around the wire.

The principle works in reverse: magnetism can create electricity.

When a wire moves through a magnetic field, it creates a flow of electricity in the wire.

This second principle is used in power generation. A diesel engine, or a water-powered turbine, or a windmill, or some other source of energy either turns coils of wire through a powerful magnetic field, or turns magnets amid some massive coils of wire. Electricity is induced in the coils. In the next experiment, you'll see a dramatic mini-demo of this effect.

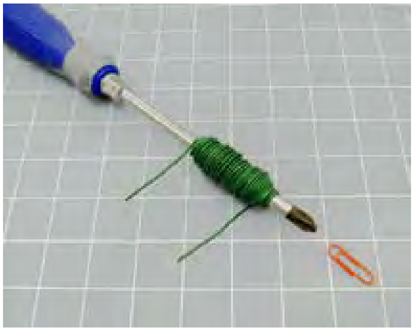

Figure 5-11. *Anyone who somehow missed this most basic childhood demo of electromagnetism should try it just for the fun of proving that a single AA battery can move a paper clip.*

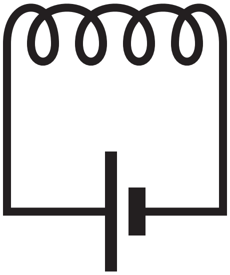

Figure 5-12. *A schematic can't get much simpler than this.*

THEORY

Inductance

When electricity flows through a wire, it creates a magnetic field around the wire. Because the electricity "induces" this effect, it is known as *inductance*. The effect is illustrated in Figure 5-13.

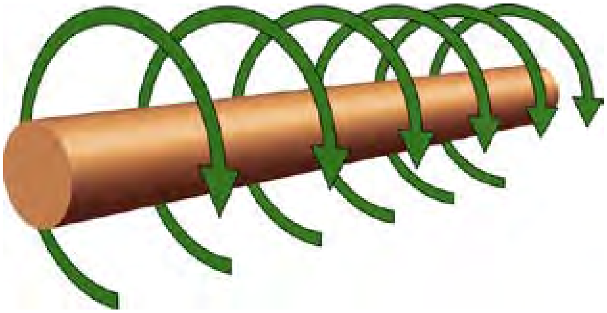

Figure 5-13. *When the flow of electricity is from left to right along this conductor, it induces a magnetic force shown by the green arrows.*

The field around a straight wire is very weak, but if we bend the wire into a circle, the magnetic force starts to accumulate, pointing through the center of the circle, as shown in Figure 5-14. If we add more circles, to form a coil, the force accumulates even more. And if we put a magnetic object (such as a screwdriver) in the center of the coil, the effectiveness increases further.

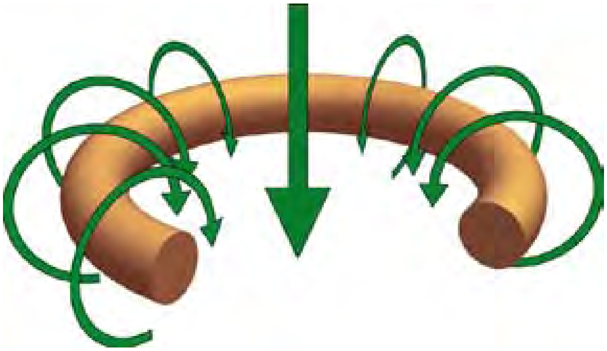

Figure 5-14. *When the conductor is bent to form a circle, the cumulative magnetic force acts through the center of the circle, as shown by the large arrow.*

Here's an approximated formula showing the relationship between the diameter of the coil, the width of the coil from end to end, the number of turns, and its inductance. The letter L is the symbol for inductance, even though the unit is the Henry, named after an American electrical pioneer named Joseph Henry:

L (in microHenrys) =
[(D × D) × (N x N)] / [(18 × D) + (40 × W)]
(Approximately)

In this formula, D is the diameter of the coil, N is the number of turns, and W is the width of the coil from end to end. See Figure 5-15. Here are three simple conclusions from this formula:

- Inductance increases with the diameter of the coil.
- Inductance increases with the square of the number of turns. (In other words, three times as many turns create nine times the inductance.)
- If the number of turns remains the same, inductance is lower if you wind the coil so that it's slender and long, but is higher if you wind it so that it's fat and short.

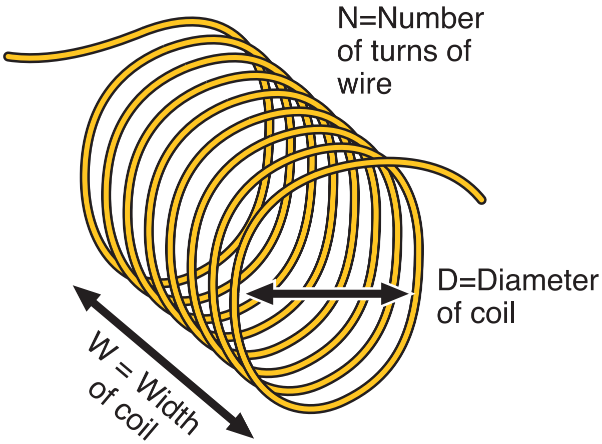

Figure 5-15. *The inductance of a coil increases with its diameter and with the square of its number of turns. If all other parameters remain the same, reducing the width (the distance from end to end) by packing the turns more tightly will increase the inductance.*

FUNDAMENTALS

Coil schematics and basics

Check the schematic symbols for coils in Figure 5-16. Note that if a coil has an iron core, this is shown with an extra couple of lines (sometimes only one line). If it has a ferrite core, the line is sometimes shown with dashes.

An iron core will add to the inductance of a coil, because it increases the magnetic effect.

A coil in isolation does not generally have any polarity. You can connect it either way around, but the magnetic force will be reversed accordingly (coils that interact with stuff—such as in transformers and solenoids—do have polarity).

Perhaps the most widespread application of coils is in transformers, where alternating current in one coil induces alternating current in another, often sharing the same iron core. If the primary (input) coil has half as many turns as the secondary (output) coil, the voltage will be doubled, at half the current—assuming hypothetically that the transformer is 100% efficient.

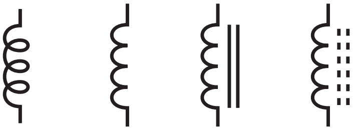

Figure 5-16. *Schematic diagrams represent coils. At far right is the older style. The third and fourth symbols indicate that the coil is wound around a solid or powdered magnetic core, respectively.*

BACKGROUND

Joseph Henry

Born in 1797, Joseph Henry was the first to develop and demonstrate powerful electromagnets. He also originated the concept of "self-inductance," meaning the "electrical inertia" that is a property of a coil of wire.

Henry started out as the son of a day laborer in Albany, New York. He worked in a general store before being apprenticed to a watchmaker, and was interested in becoming an actor. Friends persuaded him to enroll at the Albany Academy, where he turned out to have an aptitude for science. In 1826, he was appointed Professor of Mathematics and Natural Philosophy at the Academy, even though he was not a college graduate and described himself as being "principally self-educated." Michael Faraday was doing similar work in England, but Henry was unaware of it.

Henry was appointed to Princeton in 1832, where he received $1,000 per year and a free house. When Morse attempted to patent the telegraph, Henry testified that he was already aware of its concept, and indeed had rigged a system on similar principles to signal his wife, at home, when he was working in his laboratory at the Philosophical Hall.

Henry taught chemistry, astronomy, and architecture, in addition to physical science, and because science was not divided into strict specialties as it is now, he investigated phenomena such as phosphorescence, sound, capillary action, and ballistics. In 1846, he headed the newly founded Smithsonian Institution as its secretary.

Figure 5-17. *Joseph Henry was an American experimenter who pioneered the investigation of electromagnetism. This photograph is archived in Wikimedia Commons.*

Experiment 26: Tabletop Power Generation

If you have just three components, you can see magnetism generating electricity right in front of you, right now.

You will need:

- Cylindrical neodymium magnet, 3/4-inch diameter, axially magnetized. Quantity: 1. (Obtainable online at sites such as *http://www.kjmagnetics.com*.)
- Spool of hookup wire, 26-gauge, 100 feet. Quantity: 1.
- Spool of magnet wire, quarter-pound, 26-gauge, about 350 feet. Quantity: 1. (Search online for sources for "magnet wire.")
- Generic LED. Quantity: 1.
- 100 µF electrolytic capacitor. Quantity: 1.
- Signal diode, 2N4001 or similar. Quantity: 1.
- Jumper wires with alligator clips on the ends. Quantity: 2.

Procedure

You may be able to make this experiment work with the spool of hookup wire, depending on the size of the spool relative to the size of your magnet, but as the results are more likely to be better with the magnet wire, I'll assume that you're using that—initially, at least. The advantage of the magnet wire is that its very thin insulation allows the coils to be closely packed, increasing their inductance.

First peek into the hollow center of the spool to see if the inner end of the wire has been left accessible, as is visible in Figures 5-18 and 5-19. If it hasn't, you have to unwind the wire onto any large-diameter cylindrical object, then rewind it back onto the spool, this time taking care to leave the inner end sticking out.

Scrape the transparent insulation off each end of the magnet wire with a utility knife or sandpaper, until bare copper is revealed. To check, attach your meter, set to measure ohms, to the free ends of the wire. If you make a good contact, you should measure a resistance of 30 ohms or less.

Place the spool on a nonmagnetic, nonconductive surface such as a wooden, plastic, or glass-topped table. Attach the LED between the ends of the wire using jumper wires. The polarity is not important. Now take a cylindrical neodymium magnet of the type shown in Figure 5-20 and push it quickly down into the hollow core, then pull it quickly back out. See Figure 5-21. You should see the LED blink, either on the down stroke or the up stroke.

The same thing may or may not happen if you use 100 feet of 26-gauge hookup wire. Ideally, your cylindrical magnet should fit fairly closely in the hollow center of the spool. If there's a big air gap, this will greatly reduce the effect of the magnet. Note that if you use a weaker, old-fashioned iron magnet instead of a neodymium magnet, you may get no result at all.

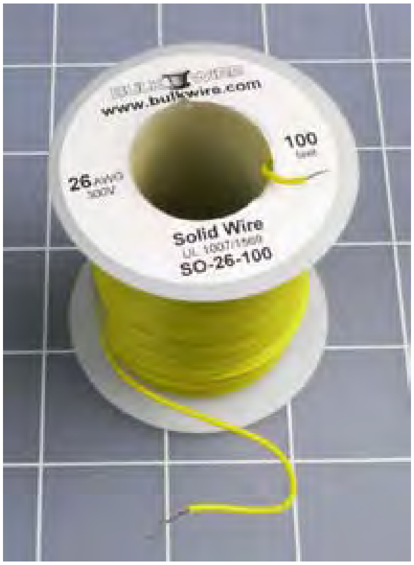

Figure 5-18. *An everyday 100-foot spool of hookup wire is capable of demonstrating the inductive power of a coil.*

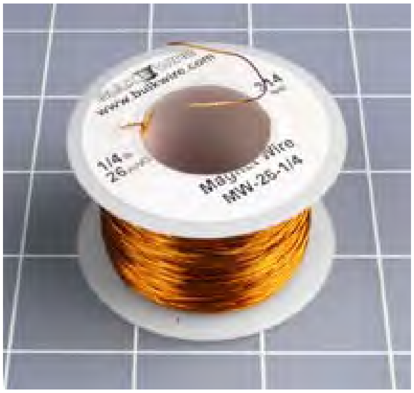

Figure 5-19. *Magnet wire has thinner insulation than hookup wire, allowing the turns to be more densely packed, inducing a more powerful magnetic field.*

Blood Blisters and Dead Media

Neodymium magnets can be hazardous. They're brittle and can shatter if they slam against a piece of magnetic metal (or another magnet). For this reason, many manufacturers advise you to wear eye protection.

Because a magnet pulls with increasing force as the distance between it and another object gets smaller, it closes the final gap very suddenly and powerfully. You can easily pinch your skin and get blood blisters.

If there's an object made of iron or steel anywhere near a neodymium magnet, the magnet will find it and grab it, with results that may be unpleasant, especially if the object has sharp edges and your hands are in the vicinity. When using a magnet, create a clear area on a nonmagnetic surface, and watch out for magnetic objects underneath the surface. My magnet sensed a steel screw embedded in the underside of a kitchen countertop, and slammed itself into contact with the countertop unexpectedly.

Be aware that magnets create magnets. When a magnetic field passes across an iron or steel object, the object picks up some magnetism of its own. Be careful not to magnetize your watch!

Don't use magnets anywhere near a computer, a disk drive, credit cards with magnetic stripes, cassettes of any type, and other media. Also keep magnets well away from TV screens and video monitors (especially cathode-ray tubes). Last but not least, powerful magnets can interfere with the normal operation of cardiac pacemakers.

Figure 5-20. *Three neodymium magnets, 1/4-, 1/2-, and 3/4-inch in diameter. I would have preferred to photograph them standing half-an-inch apart, but they refused to permit it.*

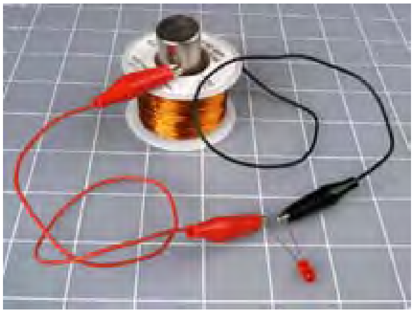

Figure 5-21. *By moving a magnet vigorously up and down through the center of a coil, you generate enough power to make the LED flash brightly.*

Here's another thing to try. Disconnect the LED and connect a 100 µF electrolytic capacitor in series with signal diode, as shown in Figure 5-23. Attach your meter, measuring volts, across the capacitor. If your meter has a manual setting for its range, set it to 20V DC. Make sure the positive (unmarked) side of the diode is attached to the negative (marked) side of the capacitor, so that positive voltage will pass through the capacitor and then through the diode.

Now move the magnet vigorously up and down in the coil. The meter should show that the capacitor is accumulating charge, up to about 10 volts. When you stop moving the magnet, the voltage reading will gradually decline, mostly because the capacitor discharges itself through the internal resistance of your meter.

This experiment is more important than it looks. Bear in mind that when you push the magnet into the coil, it induces current in one direction, and when you pull it back out again, it induces current in the opposite direction. You are actually generating alternating current.

The diode only allows current to flow one way through the circuit. It blocks the opposite flow, which is how the capacitor accumulates its charge. If you jump to the conclusion that diodes can be used to change alternating current to direct current, you're absolutely correct. We say that the diode is "rectifying" the AC power.

Experiment 24 showed that voltage can create a magnet. Experiment 25 has shown that a magnet can create voltage. We're now ready to apply these concepts to the detection and reproduction of sound.

Figure 5-22. *Because inductance increases with the diameter of a coil and with the square of the number of turns, your power output from moving a magnet through the coil can increase dramatically with scale. Those wishing to live "off the grid" may consider this steam-powered configuration, suitable for powering a three-bedroom home.*

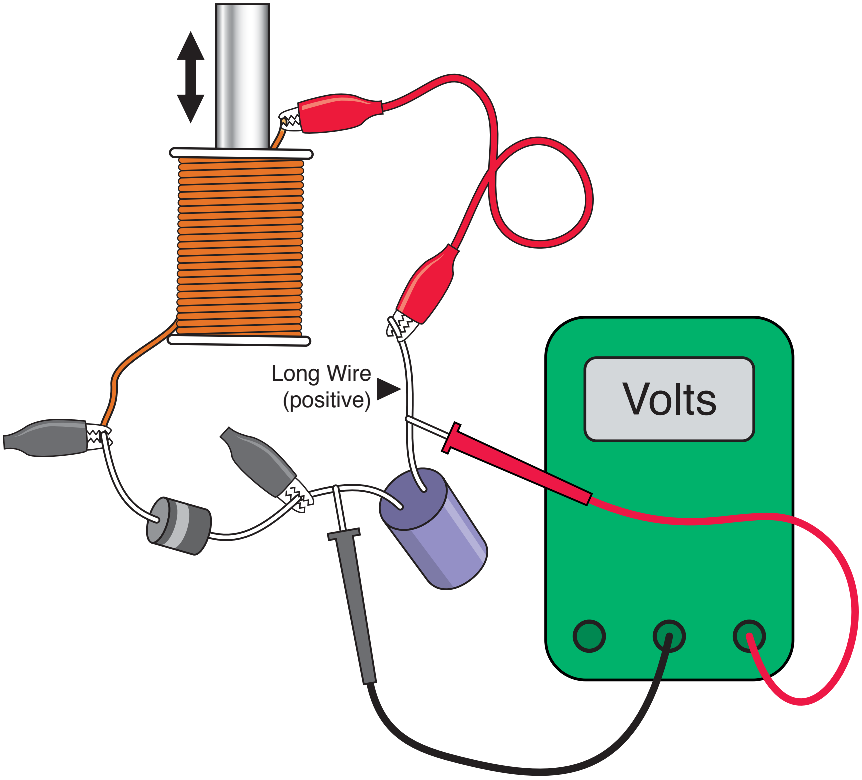

Figure 5-23. *Using a diode in series with a capacitor, you can charge the capacitor with the pulses of current that you generate by moving the magnet through the center of the coil. This demo illustrates the principle of rectifying alternating current.*

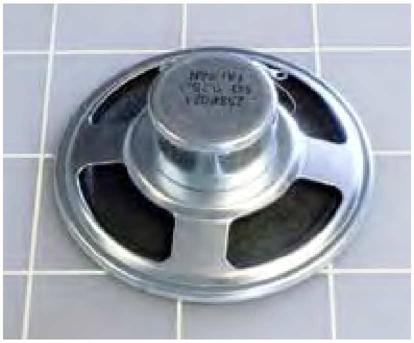

Figure 5-24. *A 2-inch loudspeaker can be instructively destroyed with a utility knife or X-Acto blade.*

Experiment 27: Loudspeaker Destruction

I'd like you to sacrifice a 2-inch loudspeaker, even though it means wasting the $5 or so that it probably costs. Actually, I don't consider this a waste, because if you want to learn how a component works, there's no substitute for actually seeing inside it. You might also have such a speaker already, part of a piece of cast-off personal electronics or toy you have in your basement.

You will need:

- Cheapest possible 2-inch loudspeaker. Quantity: 1. Figure 5-24 shows a typical example.

Procedure

Turn the loudspeaker face-up (as shown in Figure 5-25) and cut around the edge of its cone with a sharp utility knife or X-Acto blade. Then cut around the circular center and remove the ring of black paper that you've created. The result should look like Figure 5-26: you should see the flexible neck of the loudspeaker, which is usually made from a yellow weave. If you cut around its edge, you should be able to pull up the hidden paper cylinder, which has the copper coil of the loudspeaker wound around it. In Figure 5-27, I've turned it over so that it is easily visible. The two ends of this copper coil normally receive power through two terminals at the back of the speaker. When it sits in the groove visible between the inner magnet and the outer magnet, the coil reacts to voltage fluctuations by exerting an up-and-down force in reaction to the magnetic field. This vibrates the cone of the loudspeaker and creates sound waves.

Large loudspeakers in your stereo system work exactly the same way. They just have bigger magnets and coils that can handle more power (typically, as much as 100 watts).

Whenever I open up a small component like this, I'm impressed by the precision and delicacy of its parts, and the way it can be mass-produced for such a low cost. I imagine how astonished the pioneers of electrical theory (such as Faraday and Henry) would be, if they could see the components that we take for granted today. Henry spent days and weeks winding coils by hand to create electromagnets that were far less efficient than this cheap little loudspeaker.

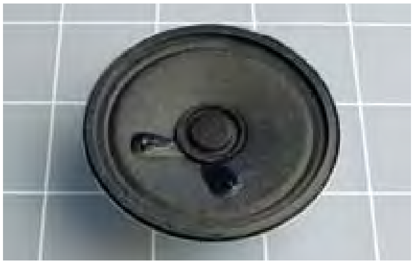

Figure 5-25. *Loudspeaker ready for creative destruction.*

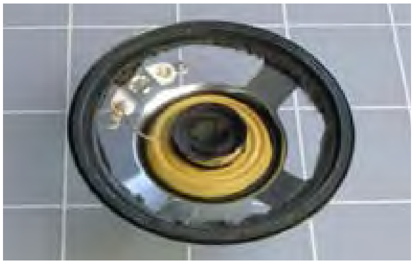

Figure 5-26. *The cone has been removed.*

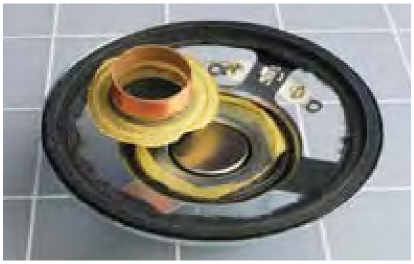

Figure 5-27. *The neck of the cone has been pulled out. Note the coil of copper wire, which fits precisely in the groove between two magnets in the base of the speaker.*

BACKGROUND

Origins of loudspeakers

Loudspeakers utilize the fact that if you run a varying electrical current through a coil situated in a magnetic field, the coil will move in response to the current. This idea was introduced in 1874 by Ernst Siemens, a prolific German inventor. (He also built the world's first electrically powered elevator in 1880.) Today, Siemens AG is one of the largest electronics companies in the world.

When Alexander Graham Bell patented the telephone in 1876, he used Siemen's concept to create audible frequencies in the earpiece. From that point on, sound-reproduction devices gradually increased in quality and power, until Chester Rice and Edward Kellogg at General Electric published a paper in 1925 establishing basic principles that are still used in loudspeaker design today.

At *http://www.radiolaguy.com/Showcase/Gallery-HornSpkr.htm* you'll find photographs of very beautiful early loudspeakers, which used a horn design to maximize efficiency. As sound amplifiers became more powerful, speaker efficiency became less important compared with quality reproduction and low manufacturing costs. Today's loudspeakers convert only about 1% of electrical energy into acoustical energy.

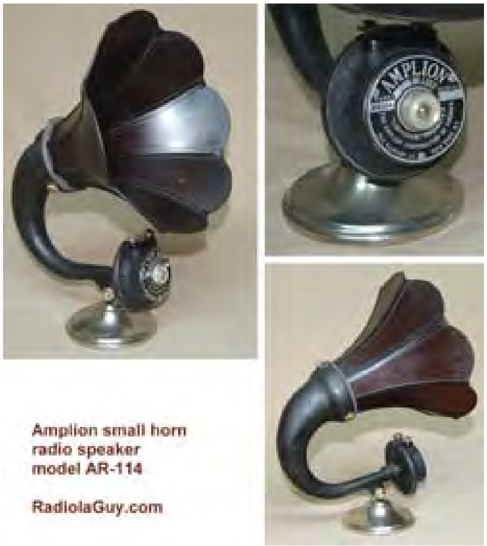

Figure 5-28. *This beautiful Amplion AR-114x illustrates the efforts of early designers to maximize efficiency in an era when the power of audio amplifiers was very limited. Photos by "Sonny, the RadiolaGuy." Many early speakers are illustrated at www.radiolaguy.com. Some are for sale.*

THEORY

Sound, electricity, and sound

Time now to establish a clear idea of how sound is transformed into electricity and back into sound again.

Suppose someone bangs a gong with a stick. The flat metal face of the gong vibrates in and out, creating sound waves. A sound wave is a peak of higher air pressure, followed by a trough of lower air pressure.

The wavelength of the sound is the distance (usually ranging from meters to millimeters) between one peak of pressure and the next peak.

The frequency of the sound is the number of waves per second, usually expressed as hertz.

Suppose we put a very sensitive little membrane of thin plastic in the path of the pressure waves. The plastic will flutter in response to the waves, like a leaf fluttering in the wind. Suppose we attach a tiny coil of very thin wire to the back of the membrane so that it moves with the membrane, and let's position a stationary magnet inside the coil of wire. This configuration is like a tiny, ultra-sensitive loudspeaker, except that instead of electricity producing sound, it is configured so that sound produces electricity. Sound pressure waves make the membrane move to and fro along the axis of the magnet, and the magnetic field creates a fluctuating voltage in the wire.

This is known as a *moving-coil* microphone. There are other ways to build a microphone, but this is the configuration that is easiest to understand. Of course, the voltage that it generates is very small, but we can amplify it using a transistor, or a series of transistors. Then we can feed the output through the coil around the neck of a loudspeaker, and the loudspeaker will recreate the pressure waves in the air. Figures 5-29 through 5-32 illustrate this sequence.

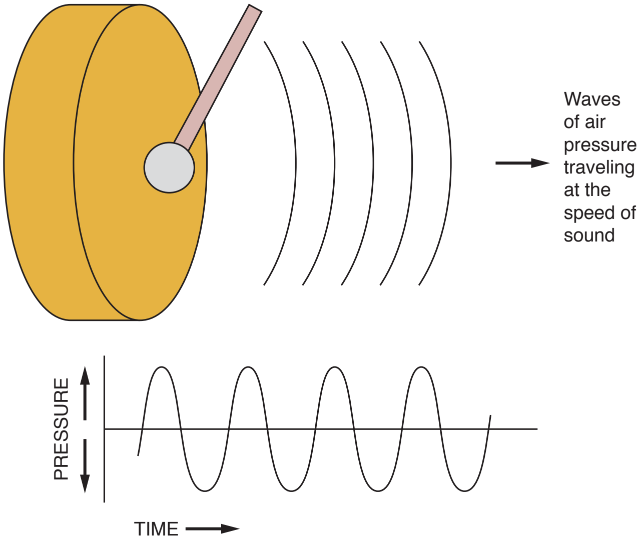

Figure 5-29. *Step 1 in the process of converting sound to electricity, and back again. When the hammer hits the gong, the face of the gong vibrates, creating pressure waves that travel through the air.*

THEORY

Sound, electricity, and sound (continued)

Somewhere along the way, we may want to record the sound and then replay it. But the principle remains the same. The hard part is designing the microphone, the amplifier, and the loudspeaker so that they reproduce the waveforms *accurately* at each step. It's a significant challenge, which is why accurate sound reproduction can be elusive.

Time now to think about what happens inside the wire when it generates a magnetic field. Obviously, some of the power in the wire is being transformed into magnetic force. But just what exactly is going on?

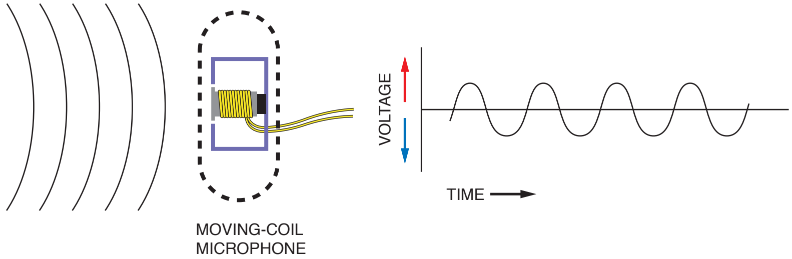

Figure 5-30. *Step 2: the pressure waves penetrate the perforated shell of a microphone and cause a diaphragm to vibrate in sympathy. The diaphragm has a coil attached to it. When the coil vibrates to and fro, a magnet at its center induces alternating current.*

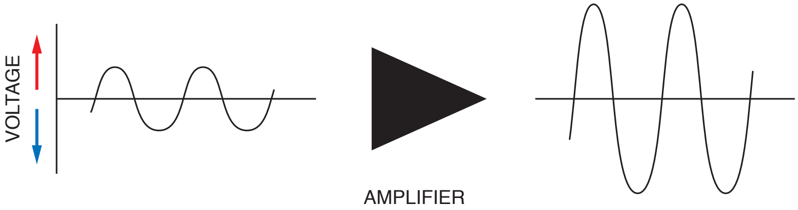

Figure 5-31. *Step 3: the tiny signals from the microphone pass through an amplifier, which enlarges their amplitude while retaining their frequency and the shape of their waveform.*

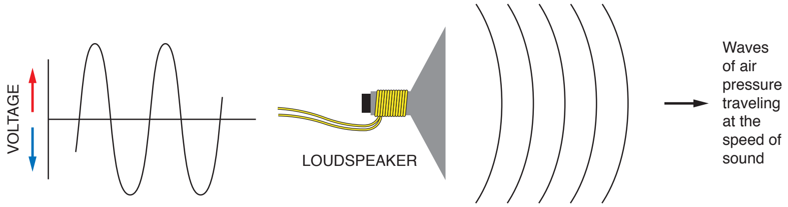

Figure 5-32. *Step 4: the amplified electrical signal is passed through a coil around the neck of a loudspeaker cone. The magnetic field induced by the current causes the cone to vibrate, reproducing the original sound.*

Experiment 28: Making a Coil React

A capacitor will absorb some DC current until it is fully charged, at which point it blocks the flow. There's another phenomenon that I haven't mentioned so far, which is the exact opposite of capacitance. It's known as *self-inductance*, and you find it in any coil of wire. Initially it blocks DC current (it reacts against it), but then its opposition gradually disappears. Here are a few definitions:

Resistance

Constrains current flow and drops voltage.

Capacitance

Allows current to flow initially and then blocks it. This behavior is properly known as capacitive reactance.

Self-Inductance

Blocks the flow of current initially and then allows it. This is also often referred to as *inductive reactance*. In fact, you may find the term "reactance" used as if it means the same thing, but since self-inductance is the correct term, I'll be using it here.

In this experiment, you'll see self-inductance in action.

You will need:

- LEDs, low-current type. Quantity: 2.
- Spool of hookup wire, 26-gauge, 100 feet. Quantity: 1.
- Resistor, 220Ω, rated 1/4 watt or higher. Quantity: 1.
- Capacitor, electrolytic, 2,000 µF or larger. Quantity: 1.
- SPST tactile switch. Quantity: 1.

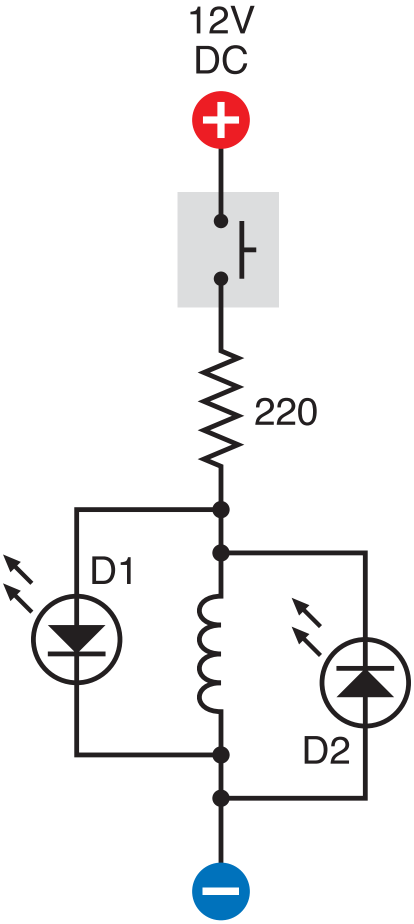

Figure 5-33. *In this demonstration of self-inductance, D1 and D2 are light-emitting diodes. When the switch is closed, D1 flashes briefly because the coil obstructs the initial flow of electricity. When the switch is opened, D2 flashes as the collapsing magnetic field induced by the coil releases another short burst of current.*

Procedure

Take a look at the schematic in Figure 5-33. At first it may not make much sense. The curly symbol is a coil of wire—nothing more than that. So apparently the voltage will pass through the 220Ω resistor, and then through the coil, ignoring the two LEDs, because the coil obviously has a much lower resistance than either of them (and one of them is upside-down anyway).

Is that what will happen? Let's find out. The coil can be a spool of 100 feet of 26-gauge (or smaller) hookup wire, although the spool of magnet wire listed in Experiment 25 will work better, if you have that. Once again, you will need access to both ends of the wire, and if the inner end is inaccessible, you'll need to rewind the coil, leaving the end sticking out.

Now that you have a coil, you can hook it up on your breadboard as shown in Figure 5-34, where the green circle is a tactile switch and the two circular red objects are LEDs. Make sure that you use low-current LEDs (otherwise, you may not see anything) and make sure that one of them is negative-side-up, positive-side-down and the other is positive-side-up, negative-side-down. Also, the 220Ω resistor should be rated at 1/4 watt or higher, if possible (see the following caution).

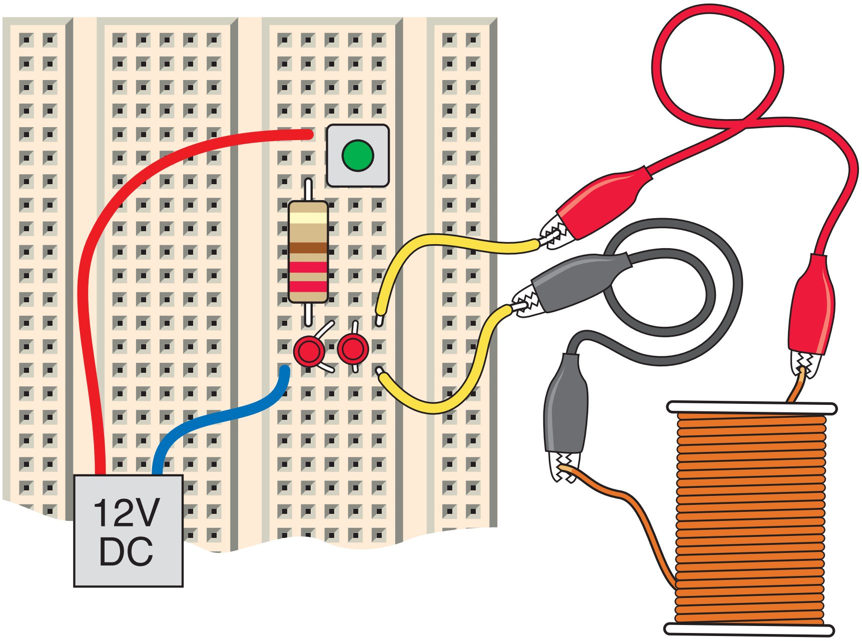

Figure 5-34. *The breadboarded version of the schematic in Figure 5-33 shows a simple way to set it up for a quick demo. The green button is a tactile switch. The two red LEDs should be placed so that the polarity of one is opposite to the polarity of the other.*

Hot Resistors

You'll be passing about 50mA through the 220Ω resistor, while the current is flowing. At 12 volts, this works out at 0.6 watts. If you use a 1/8-watt resistor, you will be overloading it, and it will get quite hot and may burn out. If you use a 1/4-watt resistor, it will still get hot, but is unlikely to burn out, as long as you don't press the button for more than a second or two.

Don't run the circuit without the coil of wire; you'll be trying to pass more than 50mA through the LEDs.

When you press the button, one LED should flash briefly. When you release the button, the other LED should flash.

What's happening here? The coil possesses self-inductance, which means that it reacts against any sudden change in the flow of electricity. First it fights it, and during that brief moment, it blocks most of the current. Consequently, the current looks for an alternative path and flows through D1, the lefthand LED in the schematic. (D2 doesn't respond, because it can pass current only in the opposite direction.)

Meanwhile, the voltage pressure overcomes the coil's self-inductance. When the self-inductance disappears, the resistance of the coil is no more than 10 ohms—so now the electricity flows mostly through the coil, and because the LED receives so little, it goes dark.

When you disconnect the power, the coil reacts again. It fights *any* sudden changes. After the flow of electricity stops, the coil stubbornly sustains it for a moment, because as the magnetic field collapses, it is turned back into electricity. This residual flow of current depletes itself through D2, the LED on the right.

In other words, the coil stores some energy in its magnetic field. This is similar to the way a capacitor stores energy between two metal plates, except that the coil blocks the current initially and then lets it build up, whereas the capacitor sucks up current initially, and then blocks it.

The more turns of wire you have in your coil, the more self-inductance the coil will have, causing your LEDs to flash more brightly.

Here's one last variation on this experiment to test your understanding of electrical fundamentals. Remove the 220Ω resistor, and substitute a 1K resistor (to protect your LED from sustained current). Remove the coil, and substitute a very large capacitor—ideally, about 4,700 µF. (Be careful to get its polarity the right way around.) What will you see when you press the button? Note that you will have to hold it down for a couple of seconds to get a result. And what will you see when you release the button? Remember: the behavior of capacitance is opposite to the behavior of self-inductance.

THEORY

Alternating current concepts

Here's a simple thought experiment. Suppose you set up a 555 timer to send a stream of pulses through a coil. This is a primitive form of alternating current.

We might imagine that the self-inductance of the coil will interfere with the stream of pulses, depending how long each pulse is, and how much inductance the coil has. If the pulses are too short, the self-inductance of the coil will tend to block them. Maybe if we can time the pulses exactly right, they'll synchronize with the time constant of the coil. In this way, we can "tune" a coil to allow a "frequency" to pass through it.

What happens if we substitute a capacitor? If the pulses are too long, compared with the time constant of the capacitor, it will tend to block them, because it will have enough time to become fully charged. But if the pulses are shorter, the capacitor can charge and discharge in rhythm with the pulses, and will seem to allow them through.

I don't have space in this book to get deeply into alternating current. It's a vast and complicated field where electricity behaves in strange and wonderful ways, and the mathematics that describe it can become quite challenging, involving differential equations and imaginary numbers. However, we can easily demonstrate the audio filtering effects of a loudspeaker and a coil.

Experiment 29: Filtering Frequencies

In this experiment, you'll see how self-inductance and capacitance can be used to filter audio frequencies. You're going to build a crossover network: a simple circuit that sends low frequencies to one place and high frequencies to another.

You will need:

- Loudspeaker, 8Ω, 5 inches in diameter. Quantity: 1. Figure 5-35 shows a typical example.
- Audio amplifier, STMicroelectronics TEA2025B or similar. Quantity: 1. See Figure 5-36.

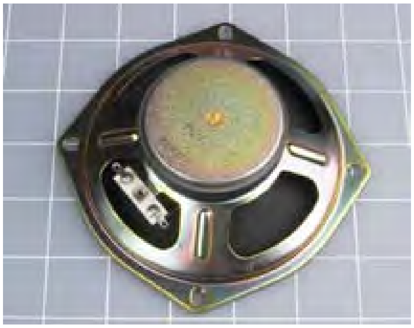

Figure 5-35. *To hear the effects of audio filters using coils and capacitors, you'll need a loudspeaker capable of reproducing lower frequencies. This 5-inch model is the minimum required.*

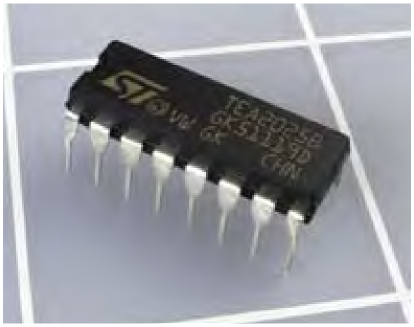

Figure 5-36. *This single chip contains a stereo amplifier capable of delivering a total of 5 watts into an 8Ω speaker when the two channels are combined.*

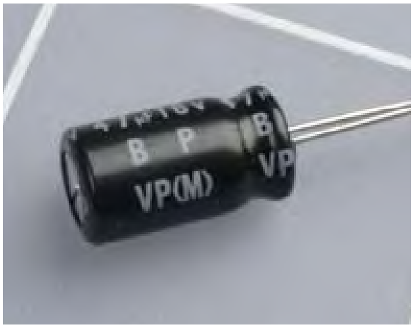

Figure 5-37. *A nonpolarized electrolytic capacitor, also known as a bipolar capacitor, looks just like an electrolytic capacitor, except that it will have "NP" or "BP" printed on it.*

- Nonpolarized electrolytic capacitors (also known as bipolar). 47 µF. Quantity: 2. A sample is shown in Figure 5-37. They should have "NP" or "BP" printed on them to indicate "nonpolarized" or "bipolar."
- Nonpolarized electrolytic capacitors (also known as bipolar). 100 µF. Quantity: 5. (Because you'll be working with audio signals that alternate between positive and negative, you can't use the usual polarized electrolytic capacitors. If you want to avoid the trouble and expense of ordering nonpolarized capacitors, you can substitute two regular electrolytics in series, facing in opposite directions, with their negative sides joined in the middle. Just remember that when you put capacitors in series, their total capacitance is half that of each individual component. Therefore, you would need two 220 µF electrolytics in series to create 110 µF of capacitance. See Figure 5-38.)
- Potentiometer, with audio taper if possible. 100K. Quantity: 1.
- Coil, for crossover network. Quantity: 1. You can search a source such as eBay for keywords "crossover" and "coil," but if you can't find one at a reasonable price, you can make do with a spool of 100 feet of 20-gauge hookup wire.
- Plastic shoebox. Quantity: 1.

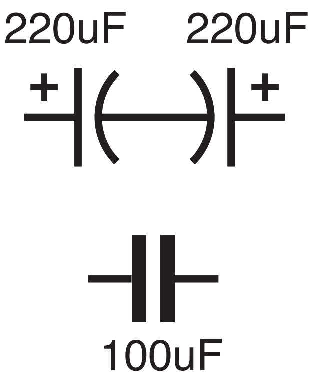

Figure 5-38. *You can make a nonpolarized electrolytic capacitor by putting two regular electrolytics in series. (In fact, that's what you'd find if you opened a real nonpolarized capacitor.) The symbol at the bottom is roughly equivalent to the pair of symbols at the top; bear in mind that two capacitors in series have a total capacitance that is half that of each of them.*

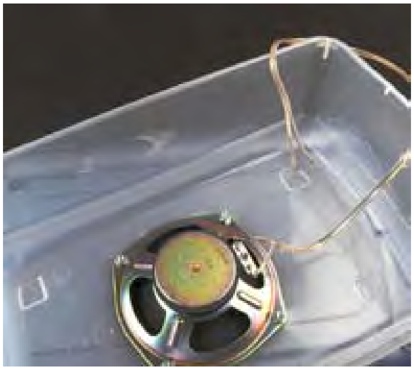

Figure 5-39. *A resonant enclosure is necessary if you want to hear some bass (lower frequencies) from your speaker. A cheap plastic shoebox is sufficient for demo purposes.*

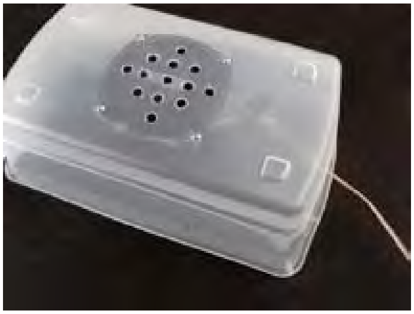

Figure 5-40. *Drill some half-inch holes in the bottom of the box, then bolt the speaker in place, running a wire out through a hole in one end. Snap on the lid, and you're ready for not-quite-high-fidelity audio.*

Procedure

The purpose of the audio amplifier chip is to provide enough power to get a decent amount of sound out of your loudspeaker. The purpose of using a 5-inch speaker is to enable you to hear lower-frequency sounds than the baby speakers that we have used previously. Bass notes have long wavelengths that small speakers are not able to generate effectively.

Maybe you remember from building the intrusion alarm that a speaker makes much more noise if you prevent the sound waves from the back of the cone from cancelling the sound waves from the front of the cone. The obvious way to achieve this is by enclosing the speaker in a box. I suggest a plastic box, because they're cheap, and we don't care too much about sound quality as long as we can hear at least some of the low frequencies. Figure 5-39 shows the speaker bolted into the bottom of a plastic box, and Figure 5-40 shows the box turned upside-down after snapping its lid into place.

Normally, a speaker should be mounted in a cabinet of heavy, thick material that has a very low resonant frequency—below the limits of human hearing. To minimize the resonance of the shoebox, you can put some soft, heavy fabric inside it before you snap the lid on. A hand towel or some socks should be sufficient to absorb some of the vibration.

Adding an Amplifier

Back in the 1950s, you needed vacuum tubes, transformers, and other power-hungry heavyweight components to build an audio amplifier. Today, you can buy a chip for about $1 that will do the job, if you add a few capacitors around it, and a volume control. The TEA2025B that I'm recommending is intended for use in cheap portable cassette players and CD players, and can work in stereo or mono mode, from a power supply ranging from 3 to 9 volts. With 9 volts and the two sides of the chip bridged together to drive one 8Ω speaker, it can generate 5 watts of audio power. That doesn't sound much compared with a typical home theater system rated at 100 watts per channel, but because loudness is a logarithmic scale, 5 watts will be quite enough to irritate any family members in the same room—and possibly even in other rooms.

If you can't find the TEA2025B chip, you can use any alternative listed as an audio amplifier. Try to find one that is designed to drive an 8Ω speaker with up to 5 watts in mono mode. Check the manufacturer's data sheet to see where you attach capacitors around it. Note carefully whether some of the capacitors have no polarity marked, even though they have fairly high values, such as 100 µF. These capacitors must function regardless of which way the alternating current is flowing, and I've marked them "NP" in my schematic in Figure 5-41, meaning "nonpolarized." (You may find them identified as "bipolar" or "BP" in parts catalogs.) As noted in the shopping list, you can put two 220 µF capacitors in series, negative-to-negative, to get the same effect as a single 100 µF nonpolarized capacitor.

For this project, it's essential to include the regular 100 µF electrolytic smoothing capacitor across the power supply. Otherwise, the amplifier will pick up and—yes, amplify—small voltage spikes in the circuit.

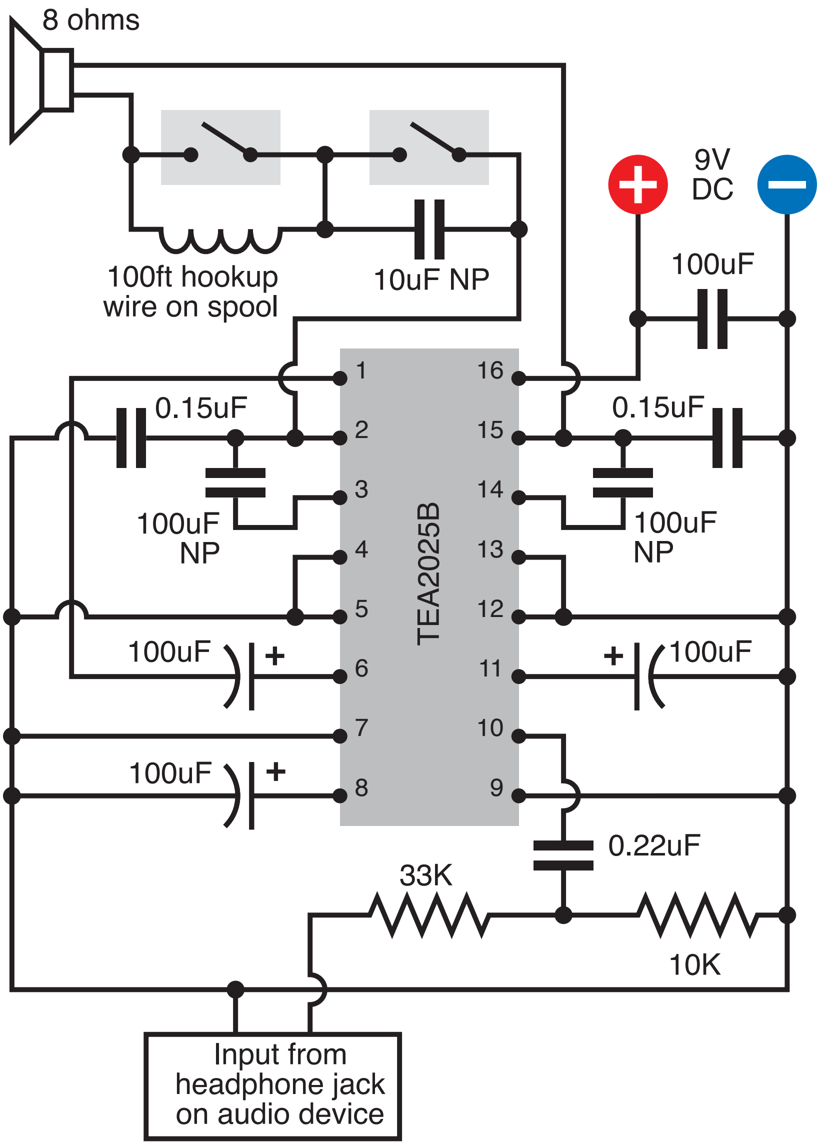
8 ohms
9V DC
100ft hookup wire on spool
10uF NP
100uF
0.15uF
0.15uF
100uF NP
100uF NP
TEA2025B
1
2
3
4
5
6
7
8
16
15
14
13
12
11
10
9
100uF
100uF
100uF
0.22uF
33K
10K
Input from headphone jack on audio device

Figure 5-41. *The audio amplifier chip should be wired with capacitors around it as shown, "NP" denoting the ones that are not polarized. The acronym "BP," meaning bipolar, is also often used to mean the same thing. The output from pins 2 and 15 of the chip can be passed through a coil or a 10 µF capacitor to demonstrate audio filtering.*

The input shown in the schematic can receive a signal from a typical media player, such as a portable MP3 player, CD player, or cassette player. To connect its headphone jack to the breadboard, you can use an adapter that converts it to a pair of RCA-type audio jacks, and then stick a wire into one of them as shown in Figure 5-42. The wire will connect to the 33K resistor on the breadboard circuit. The chromed neck of the RCA jack (which is sometimes gold-plated, or at least gold-colored) *must* be connected with the negative side of your power supply on the breadboard; otherwise, you won't hear anything. You can ignore the second output on the adapter, because we're working in mono, here, not stereo.

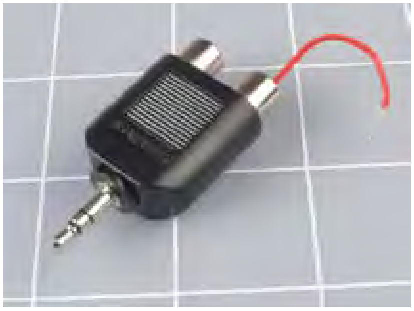

Figure 5-42. *To sample the output from the headphone socket of a music player, you can use this adapter and push the stripped end of a piece of hookup wire into one of the sockets. Then use alligator clips on a jumper wire to connect the audio to your breadboarded circuit. Don't forget to use an additional jumper wire to connect the outside of the socket to the negative side of the power supply on the breadboard. Because we're only using one speaker, the amplifier is connected to only one of the stereo outputs from your player. The other is ignored.*

The 33K resistor is necessary to protect the amplifier from being overdriven. If you don't get enough volume using your music player, decrease the 33K value. If the music is too loud and distorted, increase the value. You can also try omitting or increasing the 10K resistor next to it, which is included in an effort to reduce background hum noise.

I've shown two switches at the top of the schematic: one to bypass a coil, the other to bypass a capacitor. You can use alligator clips instead, as long as you can easily compare the sound when each of the components is inserted into the circuit.

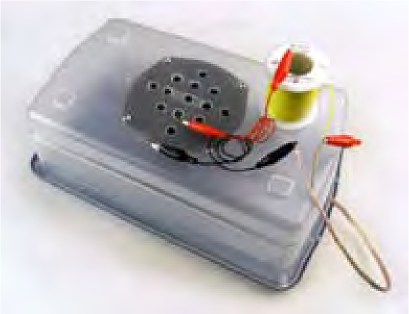

Figure 5-43. *The red and black alligator clips, lying on top of the shoebox, should connect with the output from your amplifier chip. The red jumper wire passes the signal through a coil of hookup wire on its way to the speaker. Note the change in sound when you short out the coil.*

Figure 5-43 shows a coil consisting of a spool of hookup wire being used. The red and black alligator clips resting loose on top of the shoebox will go to the output from the chip (on pins 2 and 15). There is no polarity; it doesn't matter which clip goes to which pin.

Begin by turning the volume control on your music source *all the way down* before you apply power. Don't be surprised if you hear humming or crackling noises when you activate the amplifier; it will pick up any stray voltages, because in this simple experiment, I haven't suggested that you should shield the input, and the amplifier circuit can pick up noise, as its wires can act like antennas.

Note that you may also get additional unwanted sound if you use the amplifier on a conductive desktop surface. Remove any aluminum foil or conductive foam for this project.

Make sure that your player is playing music, and slowly turn up its volume control until you hear it. If you don't hear anything, you'll have to check for circuit errors.

Now comes the interesting part. Insert the 100-foot spool of hookup wire between one output from the amplifier, and one input of the speaker (it doesn't matter which one), or if you used switches, open the switch that bypasses the coil. You should find that the music loses all its high-end response. By comparison, if you disconnect the coil and substitute a 10 µF capacitor, you should find that the music sounds "tinny," meaning that it loses all its low range, leaving only the high frequencies.

You've just tested two very simple filters. Here's what they are doing:

- The coil is a low-pass filter. It passes low frequencies but blocks high frequencies, because brief audio cycles don't have time to overcome the coil's self-inductance. A bigger coil eliminates a wider range of frequencies.
- The capacitor is a high-pass filter. It passes high frequencies and blocks low frequencies because longer audio cycles can fill the capacitance, at which point the capacitor stops passing current. A smaller capacitor eliminates a wider ranger of frequencies.

You can go a lot farther into filter design, using complex combinations of coils and capacitors to block frequencies at any point in the audible spectrum. Search online for audio filter schematics—you'll find hundreds of them.

Crossover Networks

In a traditional audio system, each speaker cabinet contains two drivers—one of them a small speaker called a *tweeter*, which reproduces high frequencies, the other a large speaker known as a *woofer*, which reproduces low frequencies. (Modern systems often remove the woofer and place it in a separate box of its own that can be positioned almost anywhere, because the human ear has difficulty sensing the direction of low-frequency sounds.)

The schematic that you just looked at and may have constructed is known as a "crossover network," and truly hardcore audiophiles have been known make their own (especially for use in car systems) to go with speakers of their choice in cabinets that they design and build themselves.

If you want to make a crossover network, you should use high-quality polyester capacitors (which have no polarity, last longer than electrolytics, and are better made) and a coil that has the right number of turns of wire and is the right size to cut high frequencies at the appropriate point. Figure 5-44 shows a polyester capacitor.

Figure 5-45 shows an audio crossover coil that I bought on eBay for $6. I was curious to find out what was inside it, so I bought two of them, and took one apart.

First I peeled away the black vinyl tape that enclosed the coil. Inside was some typical magnet wire—copper wire thinly coated with shellac or semitransparent plastic, as shown in Figure 5-46. I unwound the wire and counted the number of turns. Then I measured the length of the wire, and finally used a micrometer to measure the diameter of the wire, after which I checked online to find a conversion from the diameter in mils (1/1,000 of an inch) to American wire gauge.

As for the spool, it was plain plastic with an air core—no iron or ferrite rod in the center. Figure 5-47 shows the spool and the wire.

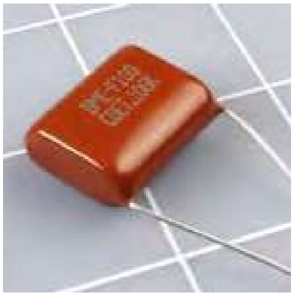

Figure 5-44. *Some nonelectrolytic capacitors have no polarity, such as this high-quality polyester film capacitor. However, they tend to be much more expensive, and are hard to find in values higher than 10 μF.*

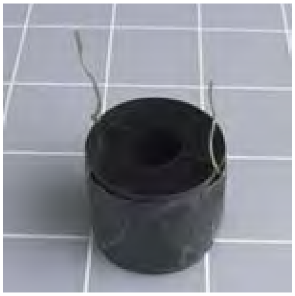

Figure 5-45. *What exotic components may we find inside this high-end audio component that's used with a subwoofer to block high frequencies?*

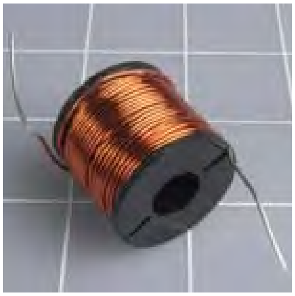

Figure 5-46. *The black tape is removed, revealing a coil of magnet wire.*

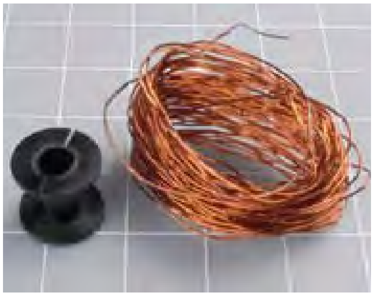

Figure 5-47. *The audio crossover coil consists of a plastic spool and some wire. Nothing more.*

So here's the specification for this particular coil in an audio crossover network. Forty feet of 20-gauge copper magnet wire, wrapped in 200 turns around a spool of 1/16–inch-thick plastic with a hub measuring 7/8 inch in length between the flanges and 1/2-inch external diameter. Total retail cost of materials if purchased separately: probably about $1, assuming you can find or make a spool of the appropriate size.

Conclusion: there's a lot of mystique attached to audio components. They are frequently overpriced, and you can make your own coil if you start with these parameters and adjust them to suit yourself.

Suppose you want to put some thumping bass speakers into your car. Could you build your own filter so that they only reproduce the low frequencies? Absolutely—you just need to wind a coil, adding more turns until it cuts as much of the high frequencies as you choose. Just make sure the wire is heavy enough so that it won't overheat when you push 100 or more audio watts through it.

Here's another project to think about: a color organ. You can tap into the output from your stereo and use filters to divide audio frequencies into three sections, each of which drives a separate set of colored LEDs. The red LEDs will flash in response to bass tones, yellow LEDs in response to the mid-range, and green LEDs in response to high frequencies (or whatever colors you prefer). You can put signal diodes in series with the LEDs to rectify the alternating current, and series resistors to limit the voltage across the LEDs to, say, 2.5 volts (when the music volume is turned all the way up). You'll use your meter to check the current passing through each resistor, and multiply that number by the voltage drop across the resistor, to find the wattage that it's handling, to make sure the resistor is capable of dissipating that much power without burning out.

Audio is a field offering all kinds of possibilities if you enjoy designing and building your own electronics.

THEORY

Waveforms

If you blow across the top of a bottle, the mellow sound that you hear is caused by the air vibrating inside the bottle, and if you could see the pressure waves, they would have a distinctive profile.

If you could slow down time and draw a graph of the alternating voltage in any power outlet in your house, it would have the same profile.

If you could measure the speed of a pendulum swinging slowly to and fro in a vacuum, and draw a graph of the speed relative to time, once again it would have the same profile.

That profile is a *sine wave*, so called because you can derive it from basic trigonometry. In a right-angled triangle, the sine of an angle is found by dividing the length of the side opposite the angle by the length of the hypoteneuse (the sloping side of the triangle).

To make this simpler, imagine a ball on a string rotating around a center point, as shown in Figure 5-48. Ignore the force of gravity, the resistance of air, and other annoying variables. Just measure the vertical height of the ball and divide it by the length of the string, at regular instants of time, as the ball moves around the circular path at a constant speed. Plot the result as a graph, and there's your sine wave, shown in Figure 5-49. Note that when the ball circles below its horizontal starting line, we consider its distance negative, so the sine wave becomes negative, too.

Why should this particular curve turn up in so many places and so many ways in nature? There are reasons for this rooted in physics, but I'll leave you to dig into that topic if it interests you. Getting back to the subject of audio reproduction, what matters is this:

- Any sound can be broken down into a mixture of sine waves of varying frequency and amplitude.

Or, conversely:

- If you put together the right mix of audio sine waves, you can create *any sound at all*.

Suppose that there are two sounds playing simultaneously. Figure 5-50 shows one sound as a red curve, and the other as pale blue. When the two sounds travel either as pressure waves through air or as alternating electric currents through a wire, the amplitudes of the waves are added together to make the more complex curve, which is shown in black. Now try to imagine dozens or even hundreds of different frequencies being added together, and you have an idea of the complex waveform of a piece of music.

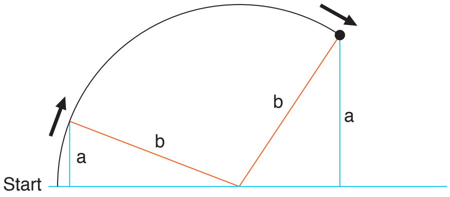

Figure 5-48. *If a weight on the end of a string (length b, in the diagram) follows a circular path at a steady speed, the distance of the weight from a horizontal center line (length a, in the diagram) can be plotted as a graph relative to time. The graph will be a sine wave, so called because in basic trigonometry, the ratio of a/b is the sine of the angle between line b and the horizontal baseline, measured at the center of rotation. Sinewaves occur naturally in the world around us, especially in audio reproduction and alternating current.*

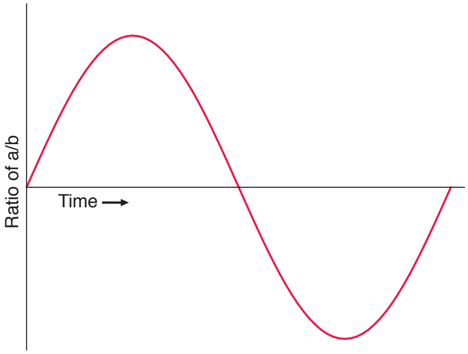

Figure 5-49. *This is what a "pure" sinewave looks like.*

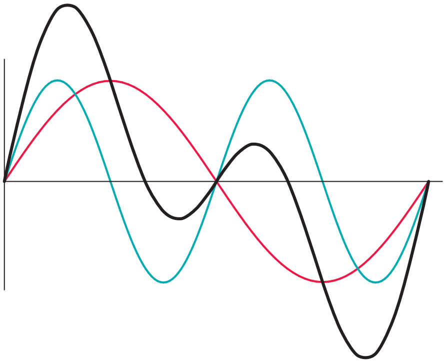

Figure 5-50. *When two sinewaves are generated at the same time (for instance, by two musicians, each playing a flute), the combined sound creates a compound curve. The blue sinewave is twice the frequency of the red sinewave. The compound curve (black line) is the sum of the distances of the sinewaves from the baseline of the graph.*

THEORY

Waveforms (continued)

You can create your own waveform as an input for your audio amplifier with the basic astable 555 timer circuit shown in Figure 5-51. You have to be careful, though, not to overload the amplifier input. Note the 680K series resistor on the output pin of the timer. Also note the 500Ω potentiometer.

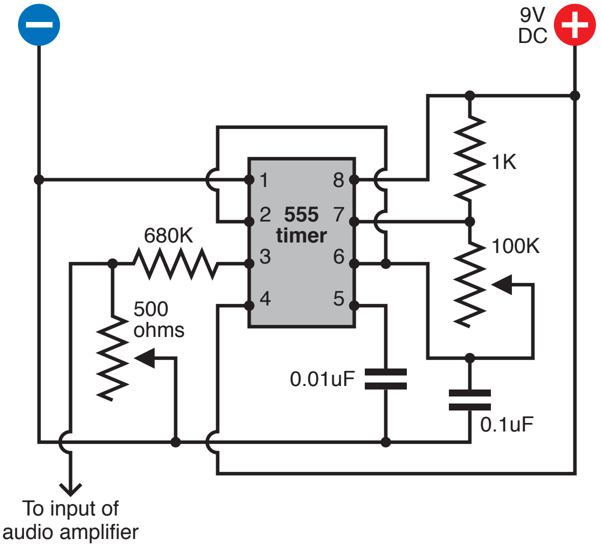

Figure 5-51. *A 555 timer is wired in astable mode using the component values shown here to generate a wide range of audible frequencies when the 100K potentiometer is adjusted. After the output is reduced in power, it can feed into the amplifier chip that was used previously.*

Disconnect your music player and connect the output from the 555 circuit to the input point (the 33K resistor) in the amplifier circuit shown earlier in Figure 5-41. You don't have to worry about a separate connection on the negative side as long as the 555 timer shares the same breadboard and the negative side of its power supply.

Make sure that the 500Ω potentiometer is turned all the way to short the output from the timer to the negative side of the power supply. This functions as your volume control. Also make sure the 100K potentiometer is in the middle of its range. Switch on the power and slowly turn up the 500Ω potentiometer until you hear a tone.

Now adjust the 100K potentiometer to create a low-pitched note. You'll find that it doesn't have a "pure" sound. There are some buzzing overtones. This is because the 555 timer is generating square waves such as those shown in Figure 5-52, not sine waves, and a square wave is actually a sum of many different sine waves, some of which have a high frequency. Your ear hears these harmonics, even though they are not obvious when you look at a square-shaped waveform.

Route one of the connections to your loudspeaker through your spool of hookup wire, and now you should hear a much purer tone, as the buzzing high frequencies are blocked by the self-inductance of the coil. Remove the coil and substitute the 10 μF capacitor, and now you hear more buzzing and less bass.

You've just taken a small step toward sound synthesis. If this subject interests you, you can go online and search for oscillator circuits. For a thorough understanding of the relationship between waveforms and the sounds you hear, you'll really need an oscilloscope, which will show you the shape of each waveform that you generate and modify.

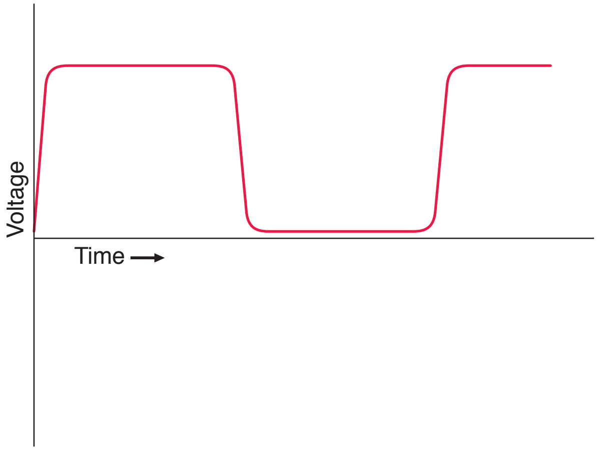

Figure 5-52. *The output from a 555 timer is either "on" or "off," with a very fast transition between those two states. The result is an almost perfect square wave. Theoretically, this can be disassembled into a complex set of sine waves that have many different frequencies. The human ear hears the high frequencies as harsh overtones.*

Experiment 30: Fuzz

Let's try one more variation on the circuit in Experiment 28. This will demonstrate another fundamental audio attribute: distortion.

You will need:

- One more 100K potentiometer.
- Generic NPN transistors: 2N2222 or similar. Quantity: 2.
- Various resistors and capacitors.

BACKGROUND

Clipping

In the early days of "hi-fi" sound, engineers labored mightily to perfect the process of sound reproduction. They wanted the waveform at the output end of the amplifier to look identical with the waveform at the input end, the only difference being that it should be bigger, so that it would be powerful enough to drive loudspeakers. Even a very slight distortion of the waveform was unacceptable.

Little did they realize that their beautifully designed tube amplifiers would be abused by a new generation of rock guitarists whose intention was to create as much distortion as possible.

The most common form of waveform abuse is technically known as "clipping." If you push a vacuum tube or a transistor to amplify a sine wave beyond the component's capabilities, it "clips" the top and bottom of the curve. This makes it look more like a square wave, and as I explained in the section on waveforms, a square wave has a harsh, buzzing quality. For rock guitarists trying to add an edge to their music, the harshness is actually a desirable feature.

Figure 5-53. *This Vox Wow-Fuzz pedal was one of the early stomp boxes, which deliberately induced the kind of distortion that audio engineers had been trying to get rid of for decades.*

The first gadget to offer this on a commercial basis was known as a "fuzz box," which deliberately clipped the input signal. An early fuzz box is shown in Figure 5-53. The clipping of a sine wave is shown in Figure 5-54.

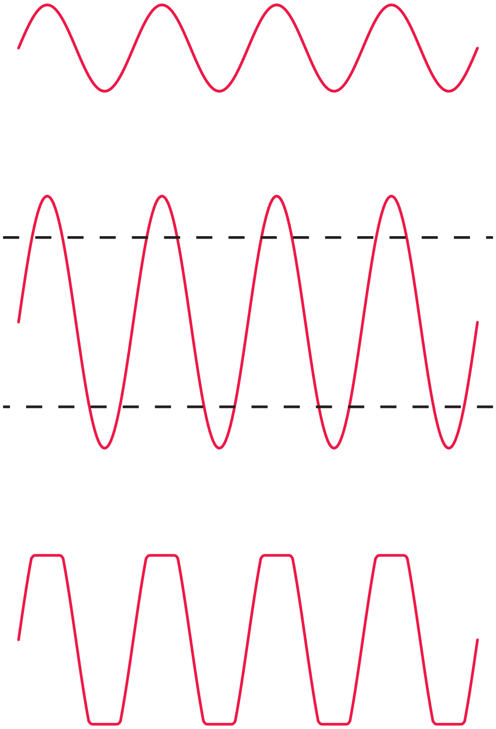

Figure 5-54. *When a sinewave (top) is passed through an amplifier which is turned up beyond the limit of its components (shown as dashed lines, center), the amplifier chops the wave (bottom) in a process known as "clipping." The result is close to a square wave and is the basic principle of a fuzz box commonly used to create a harsh guitar sound.*

Schematic

The output from the 555 timer is a square wave, so it already sounds quite "fuzzy," but we can make it more intense to demonstrate the clipping principle. I've redrawn the whole circuit in Figure 5-55, as several components have changed. The principal alteration is the addition of two NPN transistors.

If you assemble this circuit on your breadboard, note carefully that the 33K and 10K resistors at the bottom of the amplifier chip have been removed, and there's now just an 820Ω resistor in that location. The bottom of the adjacent 0.22 μF capacitor is still the input point for the amplifier, and if you follow the connection around to the middle of the schematic, you'll find it leading to a 100K potentiometer. This is your "fuzz adjuster."

The two NPN transistors are arranged so that the one on the left receives output from the 555 timer. This signal controls the flow of electricity through the transistor from a 33K resistor. This flow, in turn, controls the base of the righthand transistor, and the flow of current through that is what ultimately controls the amplifier.

When you power up the circuit, use the 100K potentiometer attached to the 555 timer to adjust the frequency (as before) and crank the "fuzz adjuster" potentiometer to hear how it adds increasing "bite" to the sound until ultimately it becomes pure noise.

The two transistors act as amplifiers. Of course, we didn't need them for that purpose—the input level for the amplifier chip was already more than adequate. The purpose of the lefthand transistor is simply to overload the righthand transistor, to create the "fuzz" effect. And when you turn up the output from the transistors with the "fuzz adjuster," eventually they overload the input of the amplifier chip, creating even more distortion.

If you want to tweak the output, try substituting different values for the 1K resistor and the 1 μF capacitor that are positioned between the emitter of the righthand transistor and the negative side of the power supply. A larger resistor should overload the transistor less. Different capacitor values should make the sound more or less harsh.

You can find literally thousands of schematics online for gadgets to modify guitar sound. The circuit I've included here is one of the most primitive. If you want something more versatile, you should search for "stomp box schematics" and see what you can find.

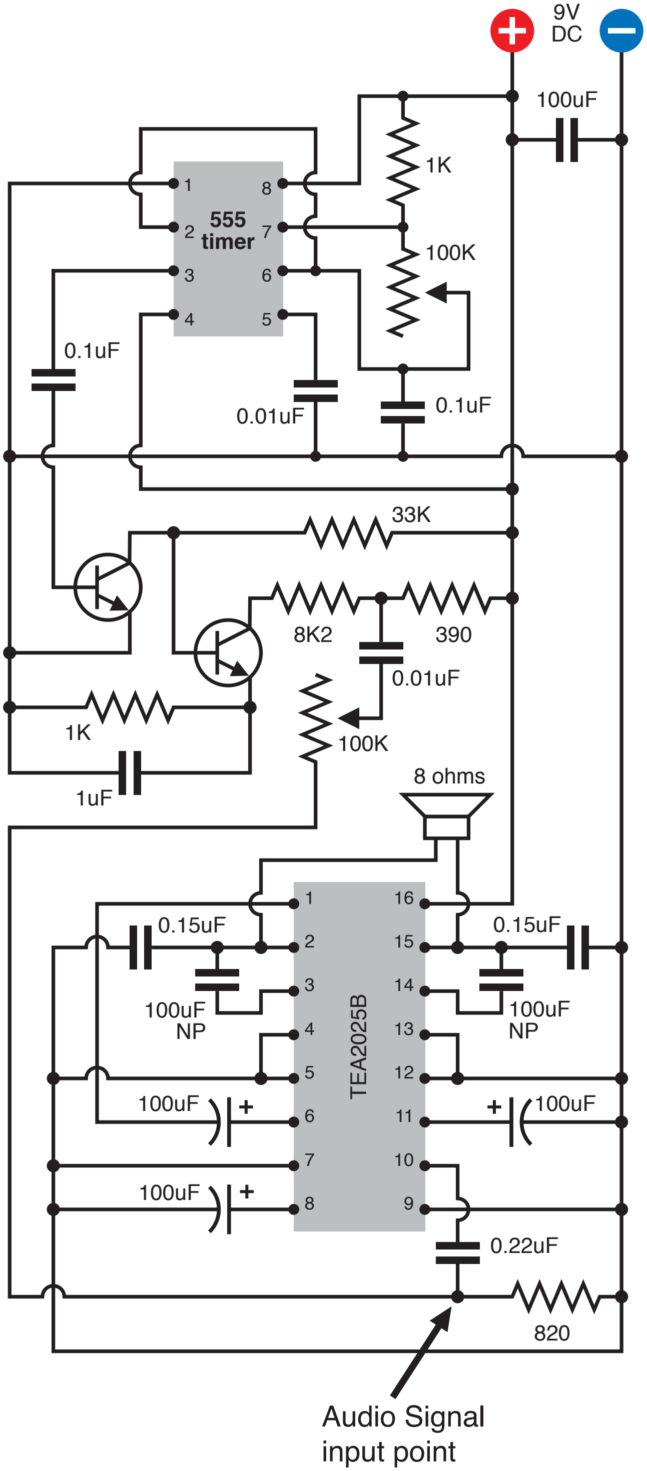
9V
DC
100uF
1K
555
timer
1
2
3
4
8
7
6
5
100K
0.1uF
0.01uF
0.1uF
33K
8K2
390
0.01uF
1K
1uF
100K
8 ohms
0.15uF
0.15uF
100uF
NP
100uF
NP
TEA2025B
1
2
3
4
5
6
7
8
16
15
14
13
12
11
10
9
100uF
+
100uF
+
100uF
+
0.22uF
820
Audio Signal
input point

Figure 5-55. *For a quick demo of clipping, insert a couple of transistors between the output from the 555 timer and the input of the amplifier chip. One transistor overdrives the other, so that when you adjust the potentiometer at the center of the circuit, you hear an increasingly harsh, distorted sound.*

BACKGROUND

Stomp-box origins

The Ventures recorded the first single to use a fuzz box, titled "The 2,000 Pound Bee," in 1962. Truly one of the most awful instrumentals ever made, it used distortion merely as a gimmick and must have discouraged other musicians from taking the concept seriously.

Ray Davies of the Kinks was the first to embody distortion as an integral part of his music. Davies did it initially by plugging the output from one amp into the input of another, supposedly during the recording of his hit "You Really Got Me." This overloaded the input and created clipping—the basic fuzz concept. From there it was a short step to Keith Richards using a Gibson Maestro Fuzz-Tone when the Rolling Stones recorded "(I Can't Get No) Satisfaction" in 1965.

Today, you can find thousands of advocates promoting as many different mythologies about "ideal" distortion. In Figure 5-56, I've included a schematic from Flavio Dellepiane, a circuit designer in Italy who gives away his work (with a little help from Google AdSense) at *http://www.redcircuits.com*. Dellepiane is self-taught, having gained much of his knowledge from electronics magazines such as the British *Wireless World*. In his fuzz circuit, he uses a very high-gain amplifier consisting of three field-effect transistors (FETs), which closely imitate the rounded square-wave typical of an overdriven tube amp.

Dellepiane offers dozens more schematics on his site, developed and tested with a dual-trace oscilloscope, low-distortion sinewave oscillator (so that he can give audio devices a "clean" input), distortion meter, and precision audio volt meter. This last item, and the oscillator, were built from his own designs, and he gives away their schematics, too. Thus his site provides one-stop shopping for home-audio electronics hobbyists in search of a self-administered education.

Before fuzz, there was tremolo. A lot of people confuse this with vibrato, so let's clarify that distinction right now:

- Vibrato applied to a note makes the frequency waver up and down, as if a guitarist is bending the strings.
- Tremolo applied to a note makes its volume fluctuate, as if someone is turning the volume control of a guitar up and down very quickly.

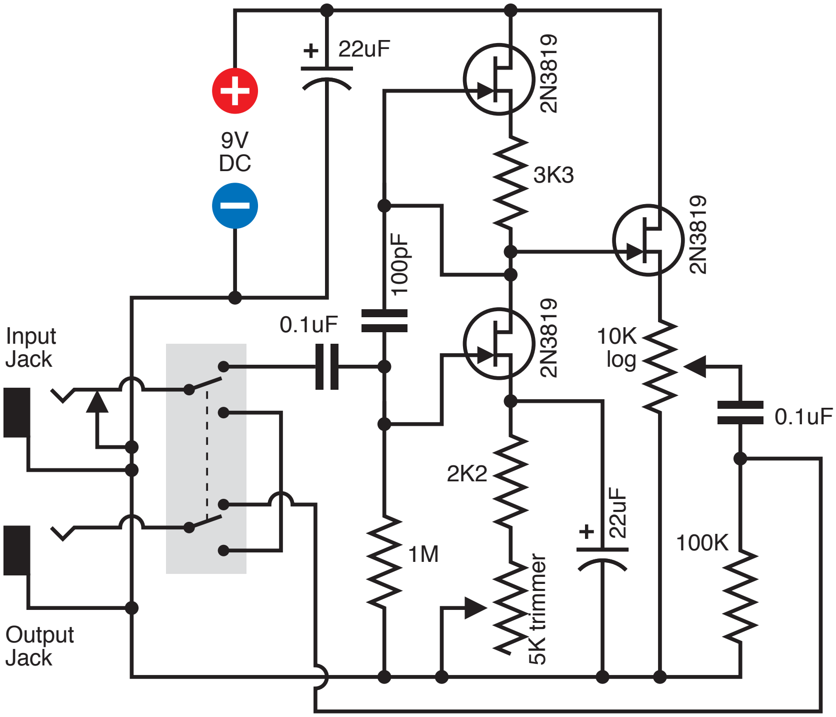

Figure 5-56. *This circuit designed by Flavio Dellepiane uses three transistors to simulate the kind of distortion that used to be created by overloading the input of a tube amplifier.*

BACKGROUND

Stomp-box origins (continued)

Harry DeArmond sold the first tremolo box, which he named the Trem-Trol. It looked like an antique portable radio, with two dials on the front and a carrying handle on top. Perhaps in an effort to cut costs, DeArmond didn't use any electronic components. His steam-punkish Trem-Trol contained a motor fitted with a tapered shaft, with a rubber wheel pressing against it. The speed of the wheel varied when you turned a knob to reposition the wheel up and down the shaft. The wheel, in turn, cranked a little capsule of "hydro-fluid," in which two wires were immersed, carrying the audio signal. As the capsule rocked to and fro, the fluid sloshed from side to side, and the resistance between the electrodes fluctuated. This modulated the audio output.

Today, Trem-Trols are an antique collectible. When industrial designer Dan Formosa acquired one, he put pictures online at *http://www.danformosa.com/dearmond.html*. And Johann Burkard has posted an MP3 of his DeArmond Trem-Trol so you can actually hear it: *http://johannburkard.de/blog/music/effects/DeArmond-Tremolo-Control-clip.html*.

The idea of a mechanical source for electronic sound mods didn't end there. The original Hammond organs derived their unique, rich sound from a set of toothed wheels turned by a motor. Each wheel created a fluctuating inductance in a sensor like the record head from a cassette player.

It's easy to think of other possibilities for motor-driven stomp boxes. Going back to tremolo: imagine a transparent disc masked with black paint, except for a circular stripe that tapers at each end. While the disc rotates, if you shine a bright LED through the transparent stripe toward a light-dependent resistor, you would have the basis for a tremolo device. You could even create never-before-heard tremolo effects by swapping discs with different stripe patterns. Figures 5-57 and 5-58 show the kind of thing I have in mind. For a real fabrication challenge, how about an automatic disc changer?

Today's guitarists can choose from a smorgasbord of effects, all of which can be home-built using plans available online. For reference, try these special-interest books:

- *Analog Man's Guide to Vintage Effects* by Tom Hughes (For Musicians Only Publishing, 2004). This is a guide to every vintage stomp box and pedal you can imagine.
- *How to Modify Effect Pedals for Guitar and Bass* by Brian Wampler (Custom Books Publishing, 2007). This is an extremely detailed guide for beginners with little or no prior knowledge. Currently it is available only by download, from sites such as *http://www.openlibrary.org*, but you may be able to find the previous printed edition from secondhand sellers, if you search for the title and the author.

Of course, you can always take a shortcut by laying down a couple hundred dollars for an off-the-shelf item such as a Boss ME-20, which uses digital processing to emulate distortion, metal, fuzz, chorus, phaser, flanger, tremolo, delay, reverb, and several more, all in one convenient multi-pedal package. Purists, of course, will claim that it "doesn't sound the same," but maybe that's not the point. Some of us simply can't get no satisfaction until we build our own stomp box and then tweak it, in search of a sound that doesn't come off-the-shelf and is wholly our own.

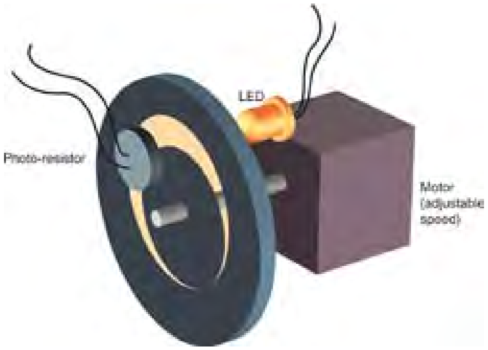

Figure 5-57. *Although electromechanical audio devices are obsolete now, some unexplored possibilities still exist. This design could create various tremolo effects, if anyone had the patience to build it.*

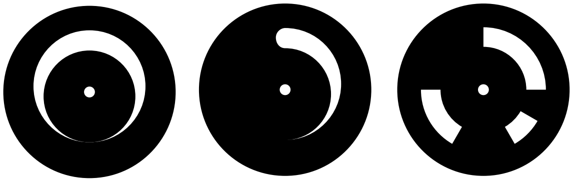

Figure 5-58. *Different stripe patterns could be used in conjunction with the imaginary electromechanical device in Figure 5-57 to create various tremolo effects.*

Experiment 31: One Radio, No Solder, No Power

Time now to go back one more time to inductance and capacitance, and demonstrate an application which also makes use of the way that waveforms can be added to each other. I want to show you how a simple circuit with no power supply at all can receive AM radio signals and make them audible. This is often known as a crystal radio, because the circuit includes a germanium diode, which has a crystal inside it. The idea dates back to the dawn of radio, but if you've never tried it, you've missed an experience that is truly magical.

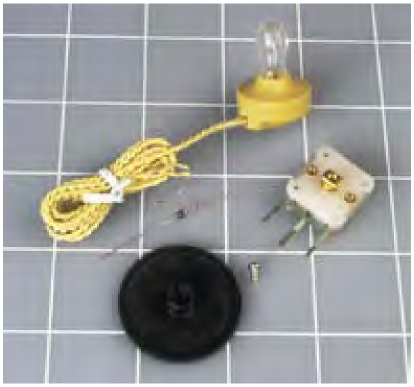

Figure 5-59. *Just add wire and a coil, and this is all you need to receive AM radio signals. The black disc becomes the tuning dial, after it is screwed onto the variable capacitor (right). This is actually an optional extra. A germanium diode (center) rectifies the radio signal. The high-impedance earphone (top) creates a barely audible sound.*

You will need:

- Rigid cylindrical object, such as a vitamin bottle. Quantity: 1.
- 22-gauge hookup wire, solid-core. Quantity: 60 feet.
- 16-gauge wire, stranded. Quantity: 100 feet.
- Polypropylene rope ("poly rope") or nylon rope. Quantity: 10 feet.
- Germanium diode. Quantity: 1.
- High-impedance headphone. Quantity: 1.

The diode and headphone can be ordered from *http://www.scitoyscatalog.com*. You cannot use a modern headphone of the type you wear with an MP3 player.

Some of these items are shown in Figure 5-59.

First, you need to make a coil. It should be about 3 inches in diameter, and you can wind it around any empty glass or plastic container of that size, so long as it's rigid. A soda bottle or water bottle isn't suitable, because the cumulative squeezing force of the turns of wire can deform the bottle so that it isn't circular anymore.

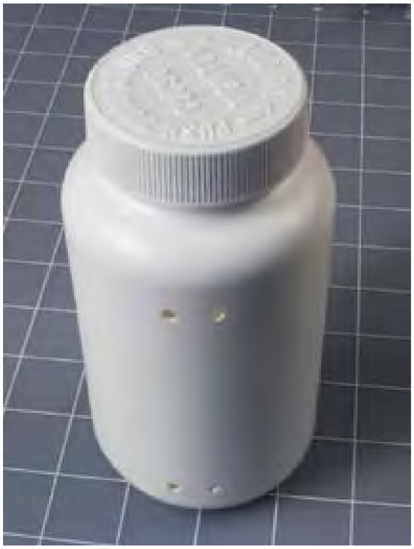

Figure 5-60. *A large, 3-inch diameter empty vitamin bottle makes an ideal core for a crystal radio coil. The drilled holes will anchor wire wrapped around the bottle.*

I chose a vitamin bottle that just happened to be exactly the right size. To remove the label, I softened its adhesive with a heat gun (lightly, to avoid melting the bottle) and then just peeled it off. The adhesive left a residue, which I removed with Xylol (also known as Xylene). This is a handy solvent to have around, as it can remove "permanent" marker stains as well as sticky residues, but you should always use latex or nitrile gloves to avoid getting it on your skin, and minimize your exposure the fumes. Because Xylol will dissolve some plastics, clearly it's not good for your lungs.

After you prepare a clean, rigid bottle, drill two pairs of holes in it, as shown in Figure 5-60. You'll use them to anchor the ends of the coil.

Now you need about 60 feet of 22-gauge solid-core wire. If you use magnet wire, its thin insulation will allow the turns of the coil to be more closely spaced, and the coil may be slightly more efficient. But everyday vinyl-insulated wire will do the job, and is much easier to work with.

Begin by stripping the first 6 inches of insulation from the end of the wire. Now measure 50 inches along the insulated remainder and apply your wire strippers at that point, just enough to cut the insulation without cutting the wire. Use your two thumb nails to pull the insulation apart, revealing about a half-inch of bare wire, as shown in Figure 5-61. Bend it at the center point and twist it into a loop, as shown in Figure 5-62.

You just made a "tap," meaning a point where you will be able to tap into the coil after you wind it. You'll need another 11 of these taps, all of them spaced 50 inches apart. (If the diameter of the bottle that you'll be using as the core of your coil is not 3 inches, multiply its diameter by 16 to get the approximate desired spacing of taps.)

After you have made 12 taps, cut the wire and strip 6 inches off that end. Now bend the end into a U shape about a half-inch in diameter, so that you can hook it through the pair of holes that you drilled at one end of the bottle. Pull the wire through, then loop it around again to make a secure anchor point.

Now wind the rest of the wire around the bottle, pulling it tightly so that the coils stay close together. When you get to the end of the wire, thread it through the remaining pair of holes to anchor it as shown in Figure 5-63. The completed coil is shown in Figure 5-64.

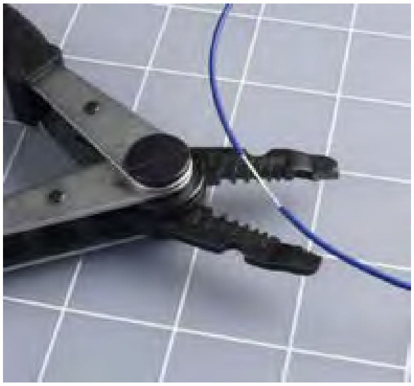

Figure 5-61. *Wire strippers expose the solid conductor at intervals along a 22-gauge wire.*

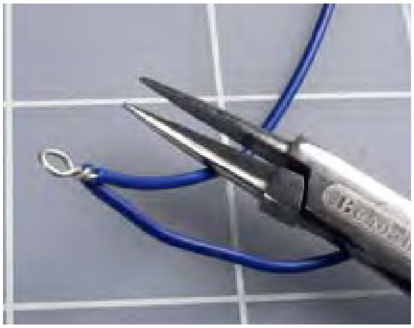

Figure 5-62. *Each exposed section of wire is twisted into a loop using sharp-nosed pliers.*

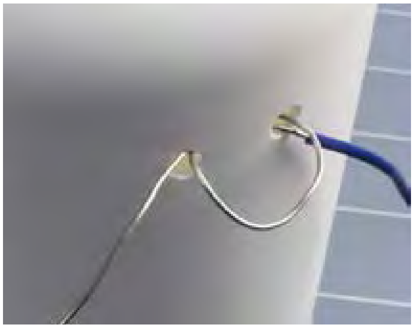

Figure 5-63. *The stripped end of the wire is secured through the holes drilled in the bottle.*

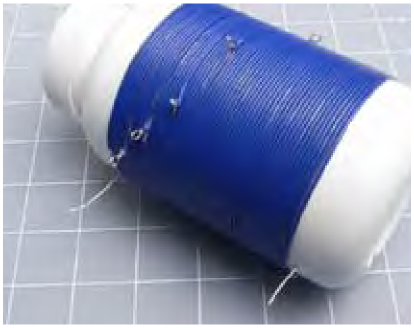

Figure 5-64. *The completed coil, wrapped tightly around the bottle.*

Your next step is to set up an antenna. If you live in a house with a yard outside, this is easy: just open a window, toss out a reel of 16-gauge wire while holding the free end, then go outside and string up your antenna by using polypropylene rope ("poly rope") or nylon rope, available from any hardware store, to hang the wire from any available trees, gutters, or poles. The total length of the wire should be about 100 feet. Where it comes in through the window, suspend it on another length of poly rope. The idea is to keep your antenna wire as far away from the ground or from any grounded objects as possible.

If you live in an apartment where you don't have access to a yard outside, you can try stringing your antenna around the room, hanging the wire from more pieces of poly rope. The antenna should still be about 100 feet long, but obviously it won't be in a straight line.

Hook the free end of your antenna to one end of your coil. At this point, you also need to add a germanium diode, which functions like a silicon-based diode but is better suited to the tiny voltages and currents that we'll be dealing with. The other end of the diode attaches to one of the wires leading to a high-impedance

High Voltage!

The world around us is full of electricity. Normally we're unaware of it, but a thunderstorm is a sudden reminder that there's a huge electrical potential between the ground below and the clouds above.

If you put up an outdoor antenna, never use it if there is any chance of a lightning strike. This can be extremely dangerous. Disconnect the indoor end of your antenna, drag it outside, and push the end of the wire into the ground to make it safe.

earphone. *A normal modern earphone or headphone will not work in this circuit.* The return wire from the headphone is connected to a jumper wire, the other end of which can be clipped to any of the taps in your coil.

One last modification, and you'll be ready to tune in. You have to ground the jumper wire. By this I mean connect it to something that literally goes into the ground. A cold-water pipe is the most commonly mentioned option, but (duh!) this will work only if the pipe is made of metal. Because a lot of plumbing these days is plastic, check under the sink to see if you have copper pipes before you try using a faucet for your ground.

Another option is to attach the wire to the screw in the cover plate of an electrical outlet, as the electrical system in your house is ultimately grounded. But the sure-fire way to get a good ground connection is to go outside and hammer a 4-foot copper-clad grounding stake into reasonably moist earth. Any wholesale electrical supply house should be able to sell you a stake. They're commonly used to ground welding equipment.

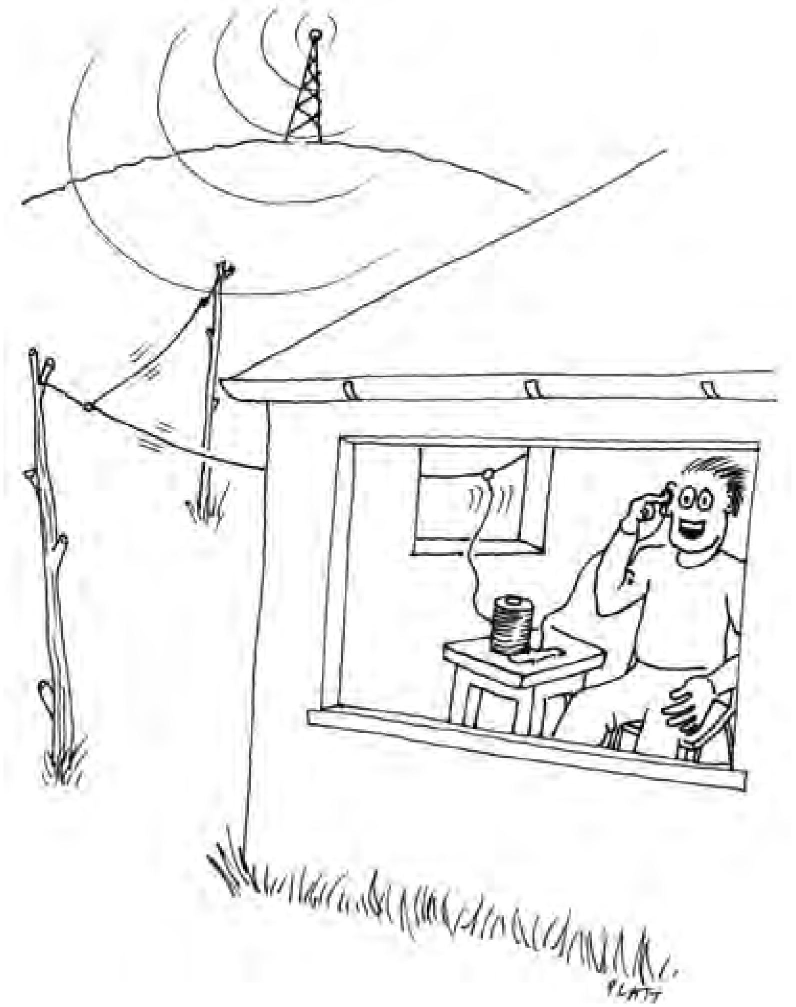

Figure 5-85. *The simple pleasure of picking up a radio signal with ultra-simple components and no additional power.*

Figures 5-66 and 5-67 show the completed radio.

If you've managed to follow these instructions (one way or another), it's time to tune your radio to the nearest station. Move the alligator clip at the end of your patch cord from one tap to another on your coil. Depending on where you live, you may pick up just one station, or several, some of them playing simultaneously.

It may seem that you're getting something for nothing here, as the earphone is making noise without any source of power. Really, though, there is a source of power: the transmitter located at the radio station. A large amplifier pumps power into the broadcasting tower, modulating a fixed frequency. When the combination of your coil and antenna resonates with that frequency, you're sucking in just enough voltage and current to energize a high-impedance headphone.

The reason you had to make a good ground connection is that the radio station broadcasts its signal at a voltage relative to ground. The earth completes the circuit between you and the transmitter. For more information on this and other concepts relating to radio, see the upcoming section "Theory: How radio works."

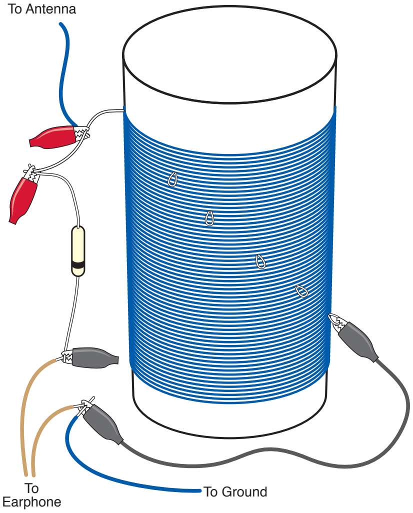

Figure 5-66. *A signal from the antenna can pass through the coil to ground. If the jumper wire is attached to an appropriate tap on the coil, it resonates with the radio signal, just powerfully enough to energize the earphone which is wired in series with a diode.*

Enhancements

The higher your antenna is, the better it should work. In my location, this is a major problem, as I live in a desert area without any trees. Still, just stringing the wire out of my window and tethering it (with rope) to the front bumper of my car enabled me to pick up a faint radio signal.

To improve the selectivity of your radio, you can add a variable capacitor, as shown in the following section. This allows you to "tune" the resonance of your circuit more precisely. Variable capacitors are uncommon today, but you can find one at the same specialty source that I recommended for the earphone and the germanium diode: the Scitoys Catalog (*http://www.scitoyscatalog.com)*.

This source is affiliated with a smart man named Simon Quellan Field, whose site suggests many fun projects that you can pursue at home. One of his clever ideas is to remove the germanium diode from your radio circuit and substitute a low-power LED in series with a 1.5-volt battery. This didn't work for me, because I live 40 miles from the nearest AM broadcaster. If you're closer to a transmitter, you may be able to see the LED varying in intensity as the broadcast power runs through it.

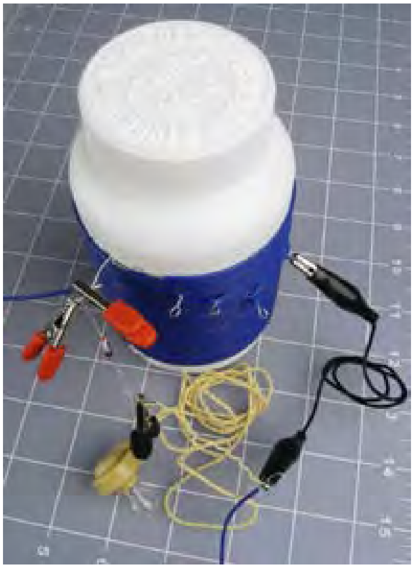

Figure 5-67. *The real-life version of Figure 5-66.*

THEORY

How radio works

When electrical frequencies are very high, the radiation they create has enough energy to travel for miles. This is the principle of radio transmission: A high-frequency voltage is applied to a broadcasting antenna, relative to the ground.

When I say "ground" in this instance, I literally mean the planet beneath us. If you set up a receiving antenna, it can pick up a faint trace of the transmission relative to the ground—as if the earth is one huge conductor. Actually the earth is so large and contains so many electrons, it can function as a common sink, like a gigantic version of the file cabinet that I suggested you should touch to get rid of static electricity in your body before touching a CMOS logic chip.

To make a radio transmitter, I could use a 555 timer chip running at, say, 850 kHz (850,000 cycles per second), and pass this stream of pulses through an amplifier to a transmission tower; if you had some way to block out all the other electromagnetic activity in the air, you could detect my signal and reamplify it.

This was more or less what Marconi (shown in Figure 5-68) was doing in 1901, after he had purchased rights to Edison's wireless telegraphy patent, although he had to use a primitive spark gap, rather than a 555 timer, to create the oscillations. His transmissions were of limited use, because they had only two states: on or off. You could send Morse code messages, and that was all.

Figure 5-68. *Marconi, the great pioneer of radio (photograph from Wikimedia Commons).*

Five years later, the first true audio signal was transmitted by imposing lower audio frequencies on the high-frequency carrier wave. In other words, the audio signal was "added" to the carrier frequency, so that the power of the carrier varied with the peaks and valleys of the audio.

At the receiving end, a very simple combination of a capacitor and a coil detected the carrier frequency out of all the other noise in the electromagnetic spectrum. The values of the capacitor and the coil were chosen so that their circuit would "resonate" at the same frequency as the carrier wave. Figures 5-69 and 5-70 illustrate these concepts.

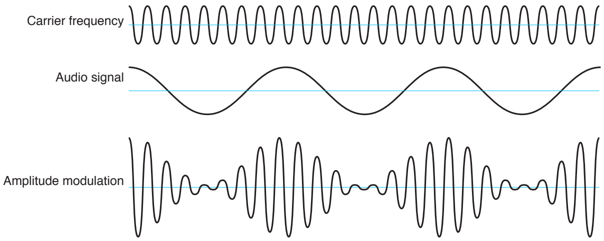

Figure 5-69. *When an audio signal (middle) is combined electronically with a high carrer frequency (top), the result looks something like the compound signal at the bottom. In actuality, the carrier frequency would be much higher compared with the audio frequency, by a radio of perhaps 1,000:1.*

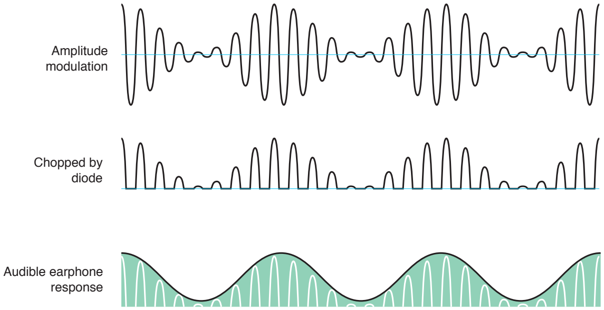

Figure 5-70. *When the compound signal is passed through a diode, only the upper half remains. An earphone cannot react fast enough to reproduce the high carrier frequency, so it "rides" the peaks and thus reproduces the audio frequency.*

The schematic in Figure 5-71 shows the simple circuit that you built by wrapping a coil around an empty vitamin bottle. When a positive pulse was received by the antenna, it resonated with the antenna and the coil, provided that the antenna was long enough and the coil was tapped at the appropriate number of turns.

THEORY

How radio works (continued)

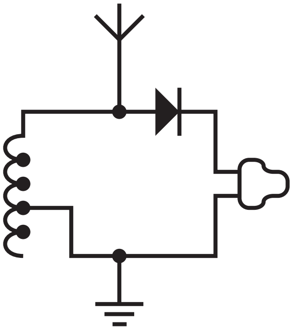

Figure 5-71. *An antenna at the top of the schematic picks up faint electromagnetic radiation from a distant transmitter. The coil at the left side is tapped at intervals so that its resonance can be adjusted to match the carrier frequency of the radio signal. Other frequencies are grounded (at the bottom of the schematic). The diode passes the "top half" of the signal to the earphone at the right, which is incapable of responding fast enough to reproduce the carrier frequency, and thus filters it out, leaving only the audio frequencies that were superimposed on it.*

By adding a capacitor, you can tune the circuit. Now an incoming pulse from the transmitter is initially blocked by the self-inductance of the coil, while it charges the capacitor. If an equally negative pulse is received after an interval that is properly synchronized with the values of the coil and the capacitor, it coincides with the capacitor discharging and the coil conducting. In this way, the right frequency of carrier wave makes the circuit resonate in sympathy. At the same time, audio-frequency fluctuations in the strength of the signal are translated into fluctuations in voltage in the circuit.

What happens to other frequencies pulled in by the antenna? The lower ones pass through the coil to ground; the higher ones pass through the capacitor to ground. They are just "thrown away."

The righthand half of the circuit samples the signal by passing it through a germanium diode and an earphone. The power from the transmitter is just sufficient to vibrate the diaphragm in the earphone, after the diode has subtracted the negative half of the signal.

Look back at the diagram of the amplitude-modulated signal. You'll see that it fluctuates up and down so rapidly, the earphone cannot possibly keep up with the positive-negative variations—hence the need for the diode. It will remain hesitating at the midpoint between the highs and lows, producing no sound at all. The diode solves this problem by subtracting the lower half of the "audio envelope," leaving just the positive spikes of voltage. Although these are still very small and rapid, they are now all pushing the diaphragm of the earphone in the same direction, so that it averages them out, approximately reconstructing the original sound wave.

Figure 5-72 shows how the circuit can be enhanced with a variable capacitor, to tune it without needing to tap the coil at intervals.

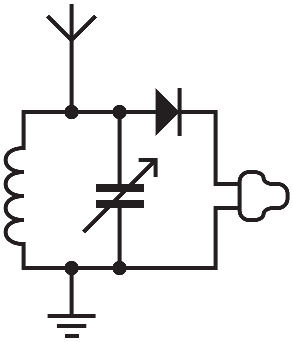

Figure 5-72. *By adding a capacitor to the circuit, its resonance can be tuned more precisely. The diagonal arrow indicates that a variable capacitor is used.*

The radio can pull in the stations on the AM (amplitude-modulated) waveband that happen to be most powerful in your area. The waveband ranges from 300 kHz to 3 MHz. If you find yourself interested in radio, your next step could be to build a powered radio using a couple of transistors. Alternatively you could build your own (legal) low-power AM transmitter. There's an ultra-simple kit available from *http://www.scitoys.com* consisting of just two principal components: a crystal oscillator, and a transformer, shown in Figure 5-73. That's all it takes.

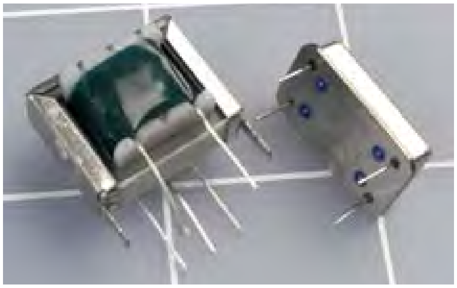

Figure 5-73. *An AM radio transmitter can be made from just two components: a transformer (left) and a crystal oscillator (right), available from http://www.scitoys.com.*

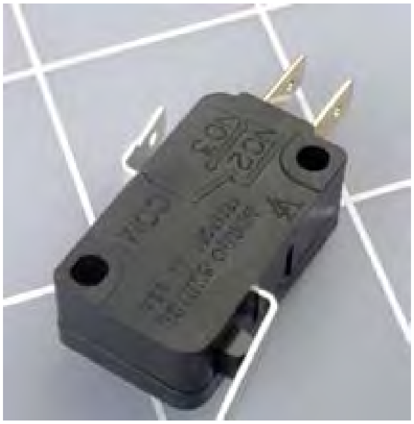

Figure 5-74. *A microswitch has a small button (at the front, righthand side in this picture) that is often actuated by a pivoted metal lever. The switch can respond to a very light pressure, but can handle relatively high currents.*

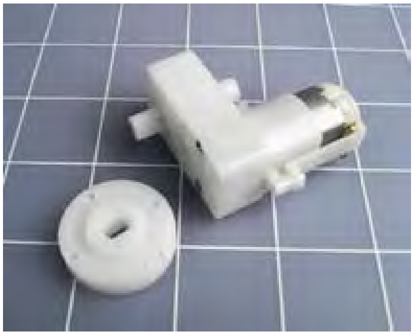

Figure 5-75. *For the Little Robot Cart, I found this 5-volt motor, which is supplied with a disc that fits its output shaft. The combination cost less than $10.*

Experiment 32: A Little Robot Cart

Robotics is another application of electronics that deserves a book in itself—or several books. So, once again, I'm going to give you an introduction followed by some points that you can follow if you want to go further. As always, I will start with the simplest possible device, which in the world of robotics is a cart that finds its way around your living room.

You will need:

- SPST or SPDT microswitches requiring minimal pressure to activate them. A force between 0.02 and 0.1 newtons would be ideal. Quantity: 2. See Figure 5-74.
- DC gear-motor, rated for any voltage between 5 and 12, drawing a maximum of 100mA in its free-running state, output shaft turning between 30 and 60 RPM. Quantity: 1. A motor is shown in Figure 5-75.
- Disc or arm that fits securely onto your motor shaft. Quantity: 1.
- 555 timer. Quantity: 1.
- DPDT nonlatching relay rated for the same voltage as your motor. Quantity: 1.
- 1/4-inch plywood or plastic, one piece about 2 feet square.
- #4 sheet-metal screws, 5/8 inch or 3/4 inch long. Quantity: 2 dozen.
- #6 bolts, 3/4 inch long, with nylon-insert lock nuts. Quantity: 2 dozen.
- 1/4-inch bolts, 1 inch long, with nuts, to mount the wheels. Quantity: 4.

I'm not specifying one particular motor, because if I did, it might not be available by the time you read this. Motors aren't like logic chips, which have retained their basic function throughout various improvements over a period of several decades. Motors come and go, and many that you may run across will be surplus parts that will never been seen again. Search online for "gear-motor" or "gearhead motor" and find one as close as possible to the specification that I have provided. The mechanical power output of the motor shouldn't be important, because we won't be requiring it to do much work.

The important consideration when you buy your motor is that you should also obtain something that fits onto its output shaft. Typically, this will be a disk or arm that can be screwed into place. To this you can then add a larger wheel of your own, which you can cut with a hole saw or make from the screw-on lid of a jar, or anything else circular that you may find in the house.

A larger wheel will make your cart move faster than a smaller wheel, but will reduce its torque, thus limiting its power to overcome obstacles.

This brings me to my next topic: fabrication. Although this is an electronics book, motors are electromechanical devices, and you have to be able to install them in some kind of a machine to get any interesting results. You can use plywood to complete the two little robotics projects here (ideally, the kind of thin, high-quality plywood sold by hobby stores) but I recommend something that looks better and is easier to work with: ABS plastic. Before you start on the robotic cart, you may want to check the section "Fundamentals: All about ABS."

FUNDAMENTALS

All about ABS

Unless you think the steampunk movement isn't going back far enough, you probably don't want your autonomous robot cart to resemble a relic from before the 1800s. Therefore, wood may not be the best construction material. Metal can look nice, but is not easy to work with. For quick results that have a twentieth-century look (maybe even a 21st-century look), plastic is the obvious choice, and I feel that ABS is the best type of plastic to use, because it provides such quick, easy results. ABS stands for "acrylonitrile butadiene styrene." Lego® blocks are made of ABS. Car-stereo installers and model-railroad buffs use it. You can use it, too. You can saw it, drill it, sand it, whittle it, and drive screws into it, and it won't warp, split, or splinter. It's washable, doesn't need to be painted, and will last almost forever.

Delrin is another type of plastic, but tends to cost more and is a little tougher to drill and cut. It's a matter of personal preference. ABS machines fairly well, but when you drill it, for example, it can "catch" on the bit and the piece will spin with the bit due to the way that plastic chips off with the bit. Delrin is self-lubricating and has better melting properties under the heat of machining, so it drills and cuts much more cleanly and easily than ABS.

Where to find ABS

Pieces of ABS a couple of feet square are available from online sources such as *http://hobbylinc.com* or estreetplastics (an eBay store), but you'll save money if you can truck on down to your nearest plastic supply house and buy it like plywood, in sheets measuring 4 by 8 feet. To discover whether you have a nearby plastic supply house, search for "plastic supply" in your yellow pages or Google Local.

Piedmont Regal Plastics has many supply centers around the nation, but you'll have to collect it yourself, and they may not be willing to cut small pieces. You can check online at *http://www.piedmontplastics.com* for their locations.

Stock colors of ABS include black, white, and "natural," which is beige. Sheets usually are textured on one side, which is the side that should face outward, as it is more scratch-resistant than the smooth side.

Because you won't be adding paint or other finishes, you'll have to be careful not to scuff the plastic or scratch it while working. Clean your bench thoroughly before you begin, taking special care to remove any metal particles, which tend to become embedded in the plastic. Use wooden shims in the jaws of your vise, and avoid resting the plastic accidentally on any sharp tools or screws. Working with ABS requires a clean environment and a very gentle touch.

Cut with Care

You can saw ABS, but if you use a table saw, the plastic will tend to melt and stick to the blade. These smears will get warm and sticky when you feed the next piece of plastic into the saw, and the result will be extremely unpleasant. The whirling blade will grab the plastic and hurl it at you powerfully enough to break bones. This is known as "kickback" and is a very serious risk when sawing plastic.

If you have extensive experience using a table saw, you are actually more vulnerable, because the reflexes and cautions you have developed while dealing with wood will not be adequate for working with plastic. Please take this warning seriously!

Your first and most obvious precaution is to use a plastic-cutting blade, which has a larger number of thicker teeth to absorb the heat. The blade I have used is a Freud 80T, but there are others. If you use a blade that is not suitable, you will see it starting to accumulate sticky smears. This is the only warning you will get. Clean that blade with a solvent such as acetone, and never use it for ABS again.

Regardless of other precautions, always wear gloves and eye protection when using a table saw, and stand to one side when feeding materials into it. Personally, after one episode of kickback that I thought had broken my arm, I prefer not to use a table saw on plastic at all.

For long, straight cuts, the alternatives include:

- *Panel saw (big and expensive, but safe and accurate).*
- *Miniature handheld circular saw with a blade around 4 inches in diameter, guided with a straight edge clamped to the sheet.*
- *Hand saw. This is my old-school preference. My favorite is a Japanese pull-to-cut saw, which makes very clean cuts: the Vaughan Extra-Fine Cross-Cut Bear Saw, 9-1/2 inches, 17 tpi (teeth per inch). If you use one of these, be careful to keep your free hand out of the way, as the saw can easily jump out of the cut. Because it is designed to cut hard materials such as wood, it has no difficulty cutting soft flesh. Gloves are strongly recommended.*

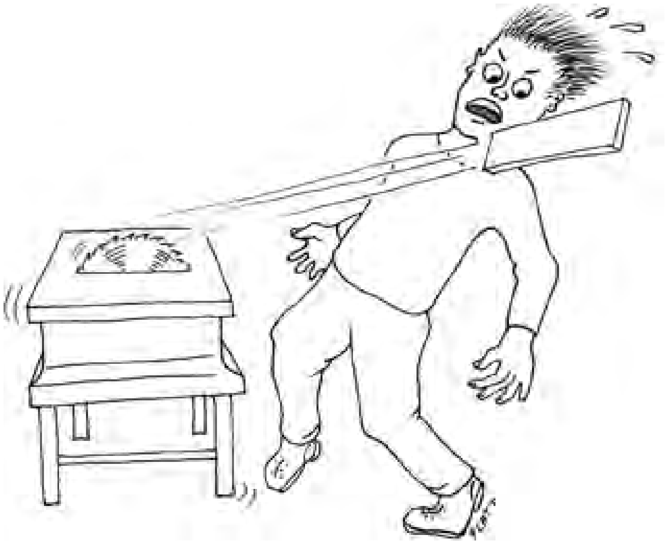

Figure 5-78. *The perils of kickback. Plastic easily sticks to the blade of a table saw, which will hurl it at you unexpectedly. Use other tools to cut plastic.*

Curving Cuts

Curving cuts involve relatively little danger, although eye protection and gloves are still advisable. My preferred tools:

- Band saw with a 3/8- or 1/4-inch blade designed for thin wood or plywood.
- Jigsaw. I have a special liking for the DeWalt XRP using Bosch blades that are designed for hardwood or plastic. This will cut complex curves in ABS as easily as scissors cutting paper.

No matter what type of saw you use, you'll have to clean ragged bits of plastic off the cut afterward, and the absolutely necessary item for this purpose is a deburring tool, available from *http://www.mcmaster.com* and most other online hardware sources. A belt sander or disc sander is ideal for rounding corners, and a metal file can be used to remove bumps from edges that are supposed to be straight.

Figures 5-77 through 5-80 show various cutting tools. Figure 5-81 shows a deburring tool, and Figure 5-82 shows a disc sander.

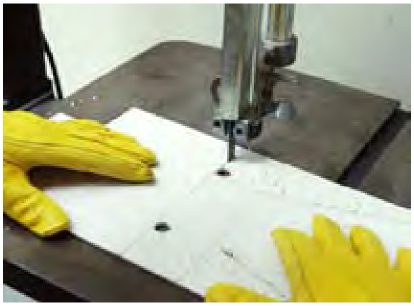

Figure 5-77. *A band saw is an ideal tool for cutting complex shapes out of ABS plastic. You can often find them secondhand for under $200.*

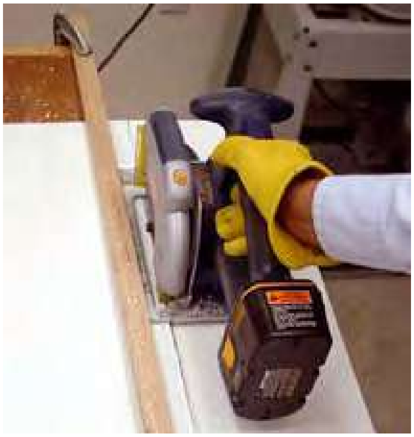

Figure 5-78. *A handheld circular saw, running along a straight-edge, is much safer than a table saw for cutting plastic, and can produce comparable results.*

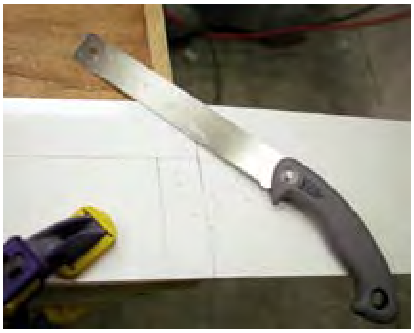

Figure 5-79. *This Japanese-style saw cuts when you pull it, rather than when you push it. After some practice, you can use it to make very accurate cuts. Because ABS is so soft, minimal muscle-power is required.*

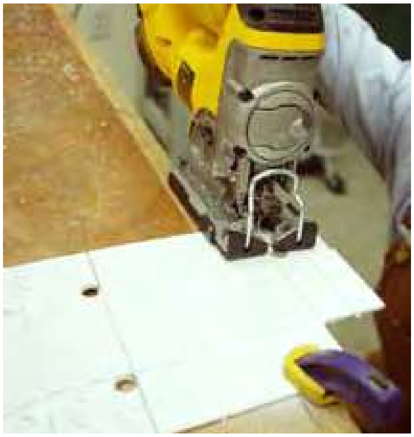

Figure 5-80. *This DeWalt jigsaw can run at very slow speeds, enabling precise and careful work with plastic.*

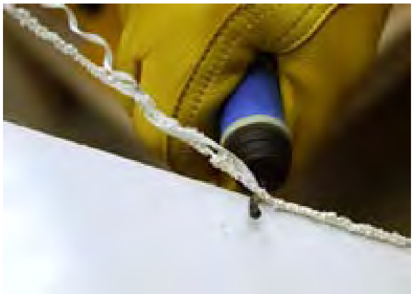

Figure 5-81. *A deburring tool will clean and bevel the sawn edge of a piece of plastic in just a couple of quick strokes.*

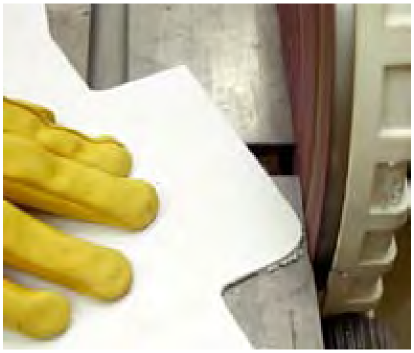

Figure 5-82. *A belt sander or disc sander is the ideal tool for rounding corners when working with ABS plastic.*

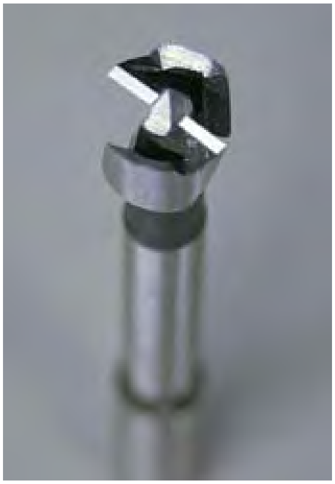

Figure 5-83. *A Forstner drill bit makes clean, precise holes; a large regular drill bit will chew up ABS plastic and make a mess.*

Making Plans

I like to use drawing software to create plans, and I try to print them at actual scale. I tape them to the smooth side of a piece of white or natural-color ABS, then use an awl to prick through the plan into the soft surface beneath. I remove the paper and connect the awl marks by drawing onto the plastic using a pencil or a fine-point water-soluble pen. Its lines can be wiped away later with a damp cloth. Don't use a permanent marker, as the solvents that you will need to clean it may dissolve the plastic.

Because ABS will tend to open a fissure when you bend it at any inside corner where you don't have a smooth radius, you need to drill holes at these locations, as shown in the cart plans in Figure 5-92 on page 275.

A regular half-inch drill bit is too aggressive; it will tend to jam itself into the plastic within one turn of the drill. Use Forstner bits (shown in Figures 5-83 and 5-84) to cut nice smooth circles.

Note that the heat from bending will tend to make any marks on the plastic permanent.

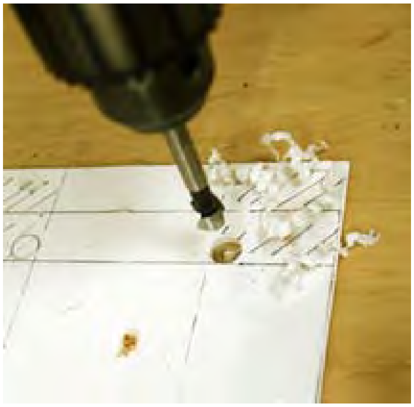

Figure 5-84. *By drilling holes at any location where two bends intersect, you reduce the risk of the plastic fissuring.*

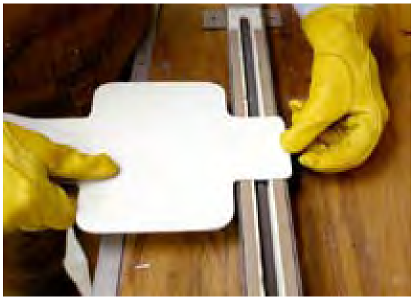

Figure 5-85. *Making clean, precise bends in ABS is simply a matter of resting the plastic over a bender that consists of an electric heating element.*

Bending It

A big advantage of plastic over wood is that you can make complex shapes by bending them, instead of cutting separate pieces and joining them with nails, screws, or glue. Unfortunately, bending does require an appropriate bender: an electric heating element mounted in a long, thin metal enclosure that you place on your workbench. The bender I use is made by FTM, a company that offers all kinds of neat gadgets for working with plastic. Their cheapest bender, shown in Figure 5-85, is just over $200 with a 2-foot element. You can get a 4-foot model for about $50 more. Check them out at *http://thefabricatorssource.com*.

Avoid Burns While Bending

A plastic bender will inflict serious burns if you happen to rest your hand on it accidentally, and because it has no warning light, you can easily forget that you have left it plugged in. Use gloves!

To bend plastic, lay it over the hot element of a plastic bender for a brief time (25 to 30 seconds for 1/8-inch ABS, 40 to 45 seconds for 3/16-inch, and up to a minute for 1/4-inch). If you overheat the plastic, you'll smell it, and when you turn it over you'll find it looks like brown melted cheese. Naturally you should learn to intervene before the plastic reaches that point.

ABS is ready to bend when it yields to gentle pressure. Take it off the bender and bend it *away* from the side that you heated. If you bend it toward the hot side, the softened plastic will bunch up inside the bend, which doesn't look nice.

You can work with it for about half a minute, and when you have it the way you want it, spray or sponge water onto it to make it set quickly. Alternatively, if you need more time, you can reheat it. The amount of force necessary to bend the sheet increases in proportion with the length of the bend, so a long bend can be difficult, and I usually insert it into a loose vise, push it a bit, move it along to the next spot, and push it again.

Because plastic bending is very similar to making shapes in origami, it's a good idea to model your projects in paper before you commit yourself to ABS.

If you decide that you don't want to spend money on a bender, don't abandon plastic just yet—you can use screws to assemble separate sections with greater ease and convenience than if you were working with wood.

Making 90-Degree Joints

Driving screws into the edge of a piece of plywood will almost always separate its layers, but ABS has no layers (or grain, either), and never splits or shatters. This means that you can easily join two pieces at 90° using small screws (#4 size, 5/8-inch long).

Figures 5-86 through 5-90 show the procedure for joining 1/8-inch (or thicker) ABS to 1/4-inch ABS, which I regard as the minimum thickness when you're inserting screws into its edge:

1. Mark a guideline on the thinner piece of plastic, 1/8 inch from its edge. For #4 screws, drill holes using a 7/64-inch bit. If you're using flat-headed screws, countersink the holes very gently.

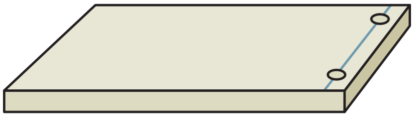

Figure 5-86

2. Hold or clamp the pieces in place and poke a pen or pencil through the holes to mark the edge of the 1/4-inch plastic beneath.
3. Remove the thin plastic, clamp the 1/4-inch plastic in a vise, and drill guide holes for the screws at each mark, centered within the thickness of the plastic. Because ABS does not compress like wood, the holes must be larger than you may expect; otherwise, the plastic will swell around the screw. A 3/32-inch bit is just right for a #4 screw.
4. Assemble the parts. Be careful not to overtighten the screws; it's easy to strip the threads that they cut in the soft plastic.

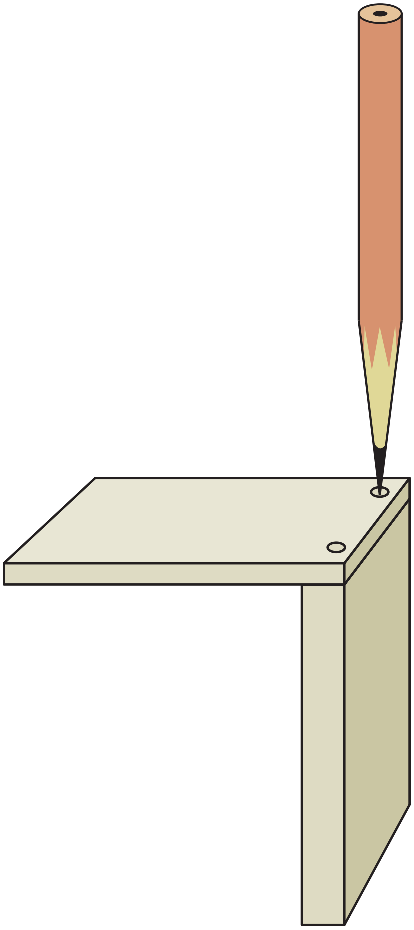

Figure 5-87

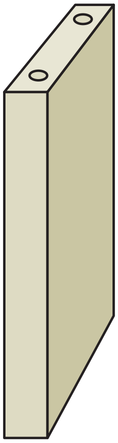

Figure 5-88

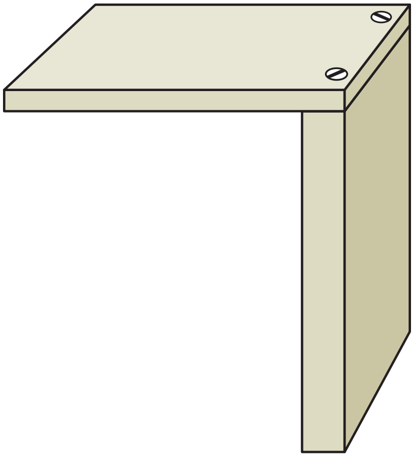

Figure 5-89. *Figures 5-86 through 5-89 illustrate four steps to join two pieces of ABS using #4 sheet-metal screws. Cut 7/64-inch holes on a line 1/8 inch from the edge of the first piece, then mark through the holes to the edge of the second piece. Drill 3/32-inch holes that are precisely centered in the edge, then screw the pieces together.*

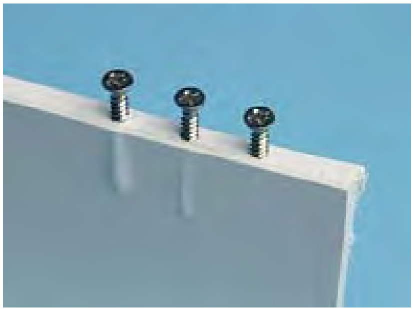

Figure 5-90. *Three #4 screws driven into the edge of ABS, using a 1/16-inch guide hole, a 5/64-inch guide hole, and a 3/32-inch guide hole. respectively Because the first two guides holes were too small, the plastic swelled around the screw (but did not break).*

Framing Your Cart

For reasons that will soon be apparent, I've chosen an unusual diamond-shaped configuration of wheels. In the rendering shown in Figure 5-91, the front wheel (at the far end of the cart) applies power, the rear wheel (at the near end of the cart) steers the cart when backing up, and the side wheels prevent it from falling over.

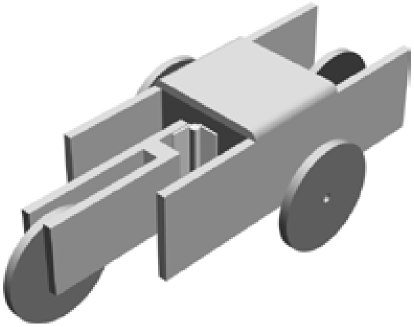

Figure 5-91. *If you have 3D rendering software, it can be a great way to test the feasibility of a construction project before you start cutting materials and trying to fit pieces together. This rendering was a proof-of-concept for the Little Robot Cart.*

Depending on the type of motor that you buy, you'll have to improvise a way to mount it in the front section of the cart. Don't be afraid to use kludges such as cable ties, duct tape, or even rubber bands to attach the motor to the frame. We're making a rough prototype, here, not a thing of beauty (although if you decide you like the cart, you can always rebuild it beautifully later).

The plan in Figure 5-92 shows the pieces that you will need. Part A is the body of the cart. If you're going to bend it from ABS, you should drill half-inch holes, with a forstner bit, at the four inside corners, so that these corners have rounded edges. If you simply saw the plastic to make sharp 90° corners, the plastic may develop fissures at the corners when you bend it. If you don't have a plastic bender and don't feel inclined to buy one, you can make Part A from three separate rectangles and then screw them together.

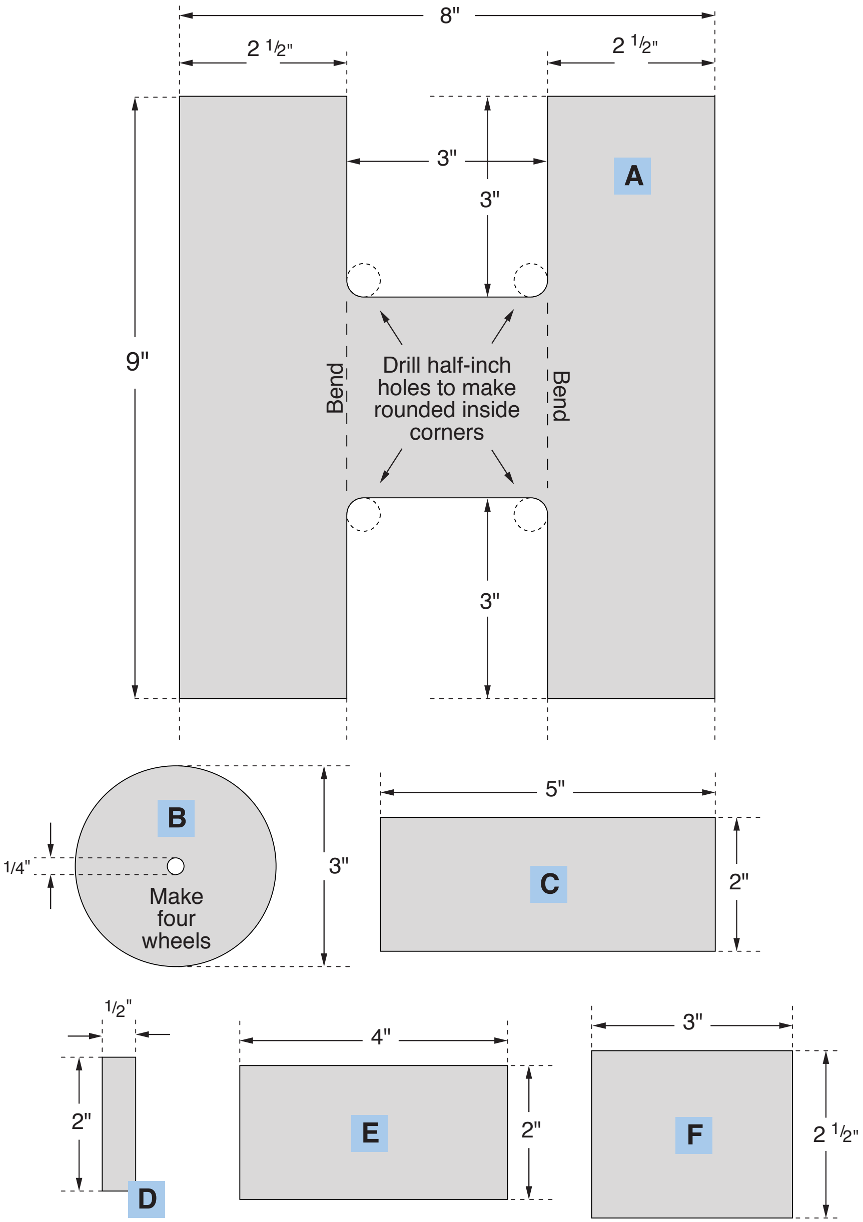
8"
2 1/2"
2 1/2"
3"
3"
A
9"
Bend
Drill half-inch holes to make rounded inside corners
Bend
3"
B
1/4"
Make four wheels
3"
5"
C
2"
1/2"
2"
D
4"
E
2"
3"
F
2 1/2"

Figure 5-92. *These sections of 1/4-inch plastic can be assembled to create the simple cart described in Experiment 31.*

Part B is a wheel, of which you will need four. I cut them using a 3-inch hole saw. The front wheel is screwed to whatever disc or arm you obtained to mate with the shaft of your motor. See Figure 5-93.

Parts C, D, and E assemble to form a yoke in which the rear wheel is mounted. I used a 2-inch hinge to pivot the yoke. The hinge is mounted on Part F, which is a partition located midway in the frame of the cart. The photographs in Figures

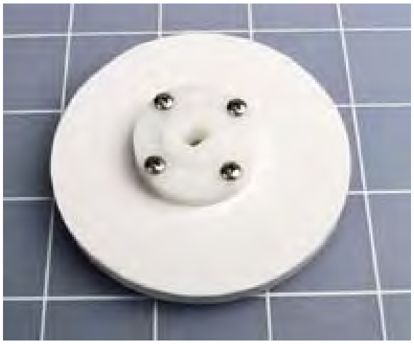

Figure 5-93. *A 3-inch wheel is screwed to the disc that mates with the drive shaft of the motor.*

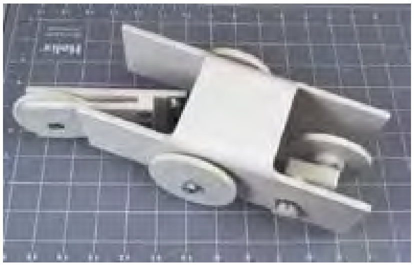

Figure 5-94. *The assembled body of the cart, before adding control electronics. The wheel at the righthand end will pull the cart from left to right. The hinged trailing wheel will allow the cart to move in a relatively straight line when it moves forward, but will tend to turn it when it backs up.*

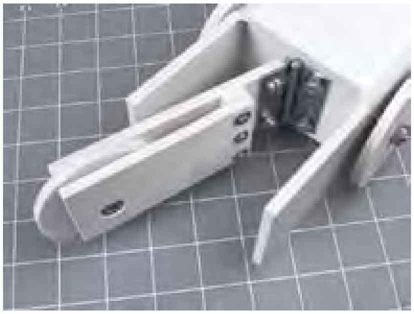

Figure 5-95. *A closeup of the hinged trailing wheel, which rotates freely and can flip from side to side with minimal friction.*

5-94 and 5-95 will help to make this clear. Initially, when you install Part F, use only two screws, one each side, so that you can adjust its angle a little. This will be necessary to optimize the contact of the wheels with the floor.

The side wheels and rear wheel must spin freely, but on the other hand, they shouldn't wobble. I simply tightened the nuts on the bolts that serve as axles for the wheels, until there was maybe half a millimeter of clearance. I added a drop of Loctite to stop the nuts from getting loose.

The plans don't show precisely where to drill holes for the axle bolts, because the location will depend on the size of your wheels. You can figure this out as you go along. Just make sure that the side wheels aren't mounted too low. We don't want them to lift the front wheel or the rear wheel off the floor. If the side wheels are a fraction higher off the ground than the front and rear wheels, that's good.

If you have tile or wood floors, your cart may acquire better traction if you wrap a thick rubber band around each disc that you use for the drive wheel and the steering wheel.

The most important aspect of the construction is to place microswitches where they'll be triggered when the cart runs into something. I placed mine at the front corners, as shown in Figures 5-96 and 5-97. And that brings me to the electronics.

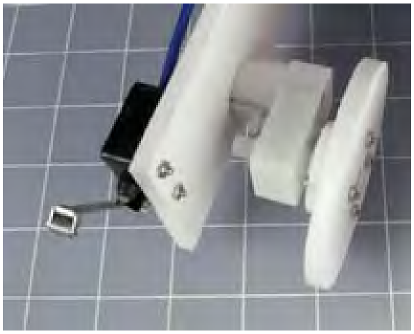

Figure 5-96

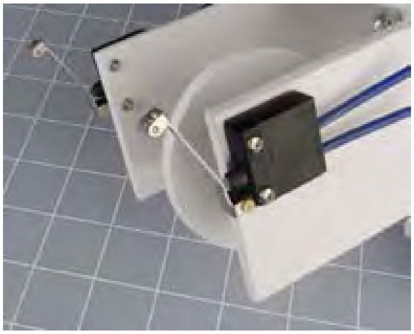

Figure 5-97. *Two microswitches with metal arms are mounted on each side of the cart, where they will sense any obstacle.*

The Circuit

The schematic is very, very simple, with only four principal components: two microswitches that sense obstacles in front of the cart, one relay, and one 555 timer. You will also need a small power switch, a battery or battery pack, and a resistor, and capacitors to go with the timer. A trimmer potentiometer will allow you to adjust the "on" time of the 555 timer, which will determine how long the cart takes to back up. See Figure 5-98.

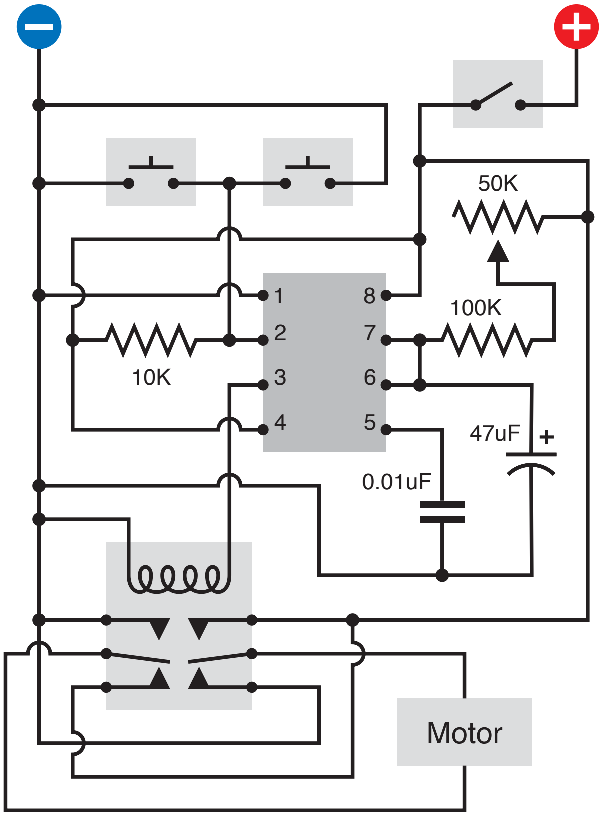

Figure 5-98. *This ultrasimple schematic is all the cart needs to enable it to back up when it hits an obstacle.*

The motor I chose requires 5 volts, so I had to use a voltage regulator with a 9-volt battery. If your motor uses 6 volts, you can wire four AA batteries to it directly. If you have a 12-volt motor, you can use two 9-volt batteries in series, supplying power through a 12-volt voltage regulator.

Assemble the components, mount them on the cart, and switch it on, and it should move forward slowly in a more-or-less straight line. If it moves backward, reverse your connection to the terminals on the motor.

When the cart bumps into something, either of the microswitches will connect negative voltage to the input pin of the 555 timer. This triggers the timer, which runs in monostable mode, generating a single pulse lasting about 5 seconds, which closes the relay, which is wired so that it reverses the voltage to the motor.

When the voltage is reversed to a simple DC motor, it runs backward. So the cart backs up. Because the rear wheel is mounted in a yoke that pivots, the yoke will tend to flip one way or the other, causing the cart to describe an arc as it moves backward. At the end of the timer cycle, the relay relaxes and the cart starts moving forward again. In forward mode, the rear wheel just follows along without applying any steering force, so the cart tends to follow a straight line—until it hits another obstacle, at which point it backs up, and tries another path.

FUNDAMENTALS

All about limit switches

The most obvious enhancement for your cart would be a better steering mechanism. You could use another motor to take care of this, with a pair of limit switches. Because limit switches are a basic, important idea in conjunction with motors, I'll explain them in detail.

Figure 5-99 shows three successive views of a motor with an arm attached to it, which can press either a lower pushbutton or an upper pushbutton. Both of the pushbuttons are normally closed, but will open when pressed by the motor arm. These buttons are the limit switches. Typically you would use microswitches for this purpose, just like the ones that I suggested as barrier-sensors at the front of the cart.

In addition, there's a DPDT relay that is activated by a simple on/off switch at the righthand side. On the cart, the 555 timer takes the place of the on/off switch, by feeding power to the relay.

Suppose that the motor begins with the arm pointing downward, as shown in the top view in Figure 5-99, and the motor is wired so that when it receives negative voltage at its lower terminal and positive at its upper terminal, it rotates counter-clockwise. This is what happens when the on/off switch closes and sends power to the DPDT relay. Positive voltage from the relay contacts cannot pass through the upper diode, but can pass through the upper limit switch, which is closed. Negative voltage cannot pass through the lower limit switch, because it's open, but can pass through the lower diode. So, the motor starts to turn counterclockwise. During the midpoint of its arc, it receives power through both of the limit switches.

Finally, the motor arm reaches the upper switch, and opens it. This prevents positive voltage from reaching the motor through that switch, and the positive voltage is also blocked by the upper diode. So, at this time, the motor stops.

Now suppose that the on/off switch is opened, as in the top view in Figure 5-100. The relay loses its power, so its contacts relax. The voltage to the motor is now reversed. Negative voltage passes through the upper diode, while positive voltage reaches the motor through the lower limit switch. The motor starts running clockwise, until its arm hits the lower switch, opening it and cutting off power to the motor.

Limit switches are necessary, because if you continue to apply voltage to a simple DC motor that is unable to turn, the motor sucks more current, gets hot, and may burn out.

You can easily see how this kind of system could be used to control the cart's steering. Even though the motor has only two positions, these are sufficient to make the cart turn when going backward, and proceed straight ahead when going forward.

To reduce power consumption, the DPDT relay could be replaced with a two-coil latching relay. The circuit would then have to be revised so that the relay is flipped to and fro by a pulse to each of its coils.

FUNDAMENTALS

All about limit switches (continued)

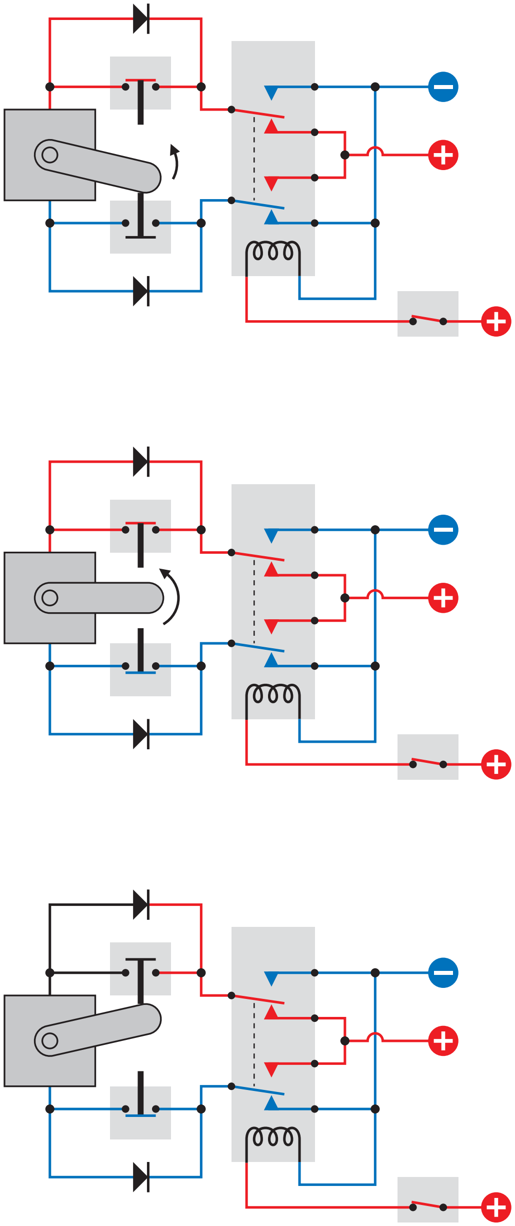

Figure 5-99. *The three diagrams, from top to bottom, show three snapshots of a motor controlled by a DPDT relay and two limit switches. When the on/off switch at bottom-right sends power to the relay, the lower relay contacts cause the motor to run counterclockwise until it stops itself as its arm opens the upper limit switch.*

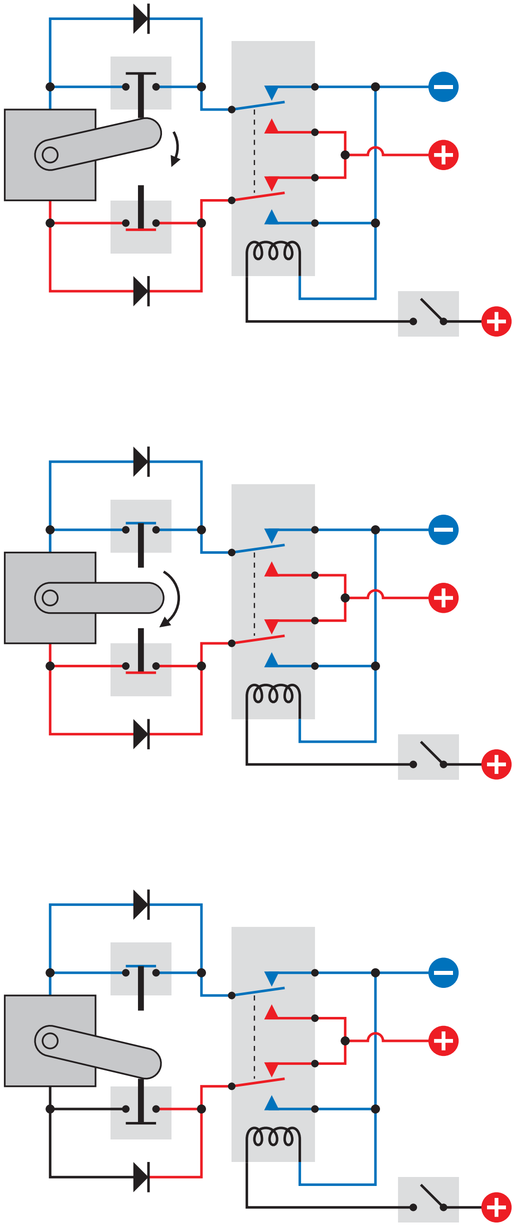

Figure 5-100. *When the on/off switch at bottom-right opens, the relay connects its upper contacts. This causes the motor to run clockwise until its arm opens the lower limit switch. Limit switches avoid the overheating and possible damage that are likely when power is delivered to a motor that is prevented from turning.*

FUNDAMENTALS

All about motors

Brushed DC motor

This is the oldest, simplest design for an electric motor, shown in very simplified form in Figure 5-101. Coils are attached to a shaft where they can interact with stationary magnets around them. The magnetic attraction turns the shaft a little, at which point the next coil on the shaft is energized to turn the shaft a little more, and then the next coil—and so on. To make this happen, electricity has to be fed into the coils by "brushes," often consisting of soft carbon pads that conduct power to a hub, known as a commutator, divided into sections, each of which is connected to a separate coil.

This basic design has several advantages if we want to build a small motorized gadget, such as a miniature robot or even a model airplane:

- Widely available
- Low cost
- Simple
- Reliable
- Will run in reverse when voltage reverses

In addition, brushed motors are often sold with reduction gearing built in. Such units are known as *gearhead motors* or *gear motors*. They free you from the need to use your own gears or belts to adjust the output speed yourself. You simply choose the motor that fits your specification.

DC stepper motor

This requires a controller, consisting of some electronics to tell the motor to rotate its shaft in small, discrete steps. The advantages of a stepper motor are:

- Precise positioning of the shaft
- Precise speed adjustment

Stepper motors are ideal for devices such as computer printers, where the paper has to roll up by a precise distance and the print head has to move laterally by an equally precise distance, but they are also useful in robots. If the motor is small enough to draw less than 200mA and will run on 12 volts or less, you can control it with pulses from a 555 timer. I'll describe stepper motors in more detail in Experiment 33.

Servo motor

This is generally used in conjunction with a programmable microcontroller, which sends instructors to rotate the motor shaft to a specific position and then hold it there. I'll mention servo motors when I introduce you to microcontrollers, but we won't be dealing with them in detail.

Other types of motors exist, including brushless DC motors (which require a different type of controller and are found in computer disk drives and CD players), and AC motors (including synchronous motors, which synchronize their rotation with the frequency of AC voltage, and were used extensively in clocks, before clocks mostly became digital).

In this book, I'll be talking mostly about brushed DC motors and DC stepper motors.

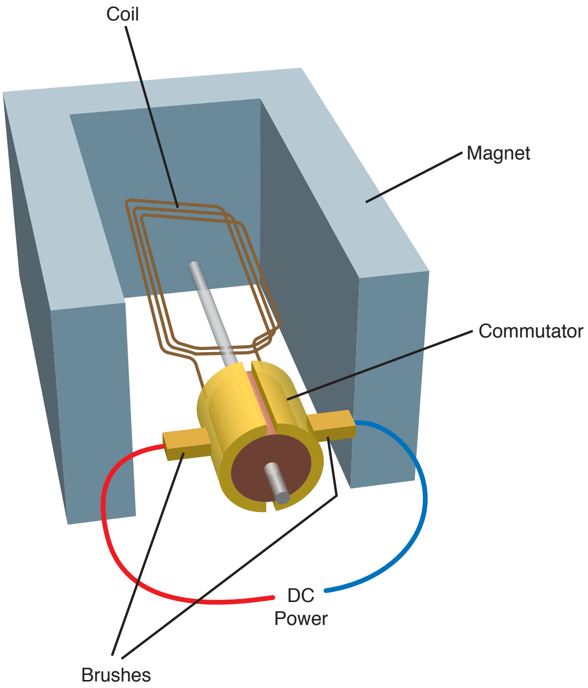

Figure 5-101. *The basic principle of a simple DC motor. The commutator passes electricity through a coil, creating a magnetic field that interacts with a magnet around the motor. The coil turns, and the commutator turns with it, until the electric field through the coil is reversed. This causes the process to repeat. In reality, a motor is likely to have a commutator formed from multiple segments, connected with multiple coils. The principle, however, remains the same.*

Take-home messages from this experiment include the following:

- You can buy simple DC motors with reduction gearing built in, providing your choice of RPM. Literally hundreds of websites will sell you small motors for robotics projects.
- When you reverse the voltage to a DC motor, the motor runs in reverse.
- A DPDT relay can be wired so that when it closes its contacts, it reverses a power supply to a motor.
- You can use two limit switches and a pair of diodes to stop a motor at two positions. In each of its stopping positions, the motor consumes no power and you won't have the risk of it burning out.

What other projects can you imagine using this simple set of techniques?

Mechanical Power

In the United States, the turning force, or torque, of a motor is usually measured in pound-feet or ounce-inches. In Europe, the metric system is used to measure torque in dynes.

A pound-foot is easy to understand. Imagine a lever pivoted at one end, as shown in Figure 5-102. If the lever is one foot long, and you hang a one-pound weight at the end of it, the turning force is one pound-foot.

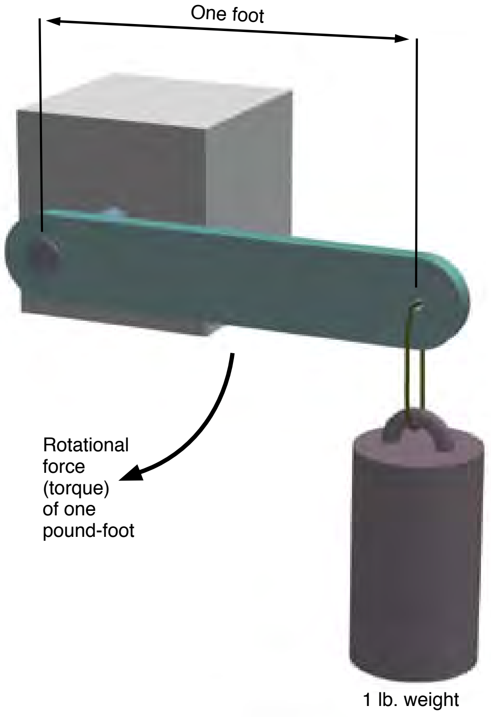

Figure 5-102. *The rotational force created by a motor is known as "torque," and in the United States it is measured in pound-feet (or ounce-inches, for small motors). In the metric system, torque is measured in dynes. Note that the torque created by a motor will vary according to the speed at which the motor is running.*

FUNDAMENTALS

Wire gauges

If you're going to power larger motors, or other components that take more current than LEDs or small relays, you really need to know about wire gauges. In particular, what's the relationship between wire thickness and AWG (American Wire Gauge)? And what gauge of wire should you use for any given current?

You can find numerous charts and tables if you go online, but many of these sources contradict each other, especially on the topic of how much current is safe to run through each gauge of wire.

After making several comparisons (and testing some wire samples myself), I've compiled the table in Figure 5-103, which I recommend as a compromise. Note the following:

- This table applies to solid-core copper wire.
- For stranded wire, or copper that has been tinned (giving it a silver appearance), the number of ohms per foot will *increase*, the number of feet per ohm will *decrease*, and the maximum amperage will *decrease*, probably by around 20%.

The maximum amperage assumes that the wire is insulated, preventing it from radiating heat as effectively as a bare conductor. I am also assuming that the wire is likely to be at least partially enclosed, inside a box or cabinet. At the amperages listed for each gauge of wire, you should expect the wire to become noticeably warm, and personally I would tend to use thicker wire instead of the maximums indicated in the table.

Most tables of this type only tell you the resistance of each gauge of wire in ohms per 1,000 feet. I have included that number but have also expressed the function the other way around, as the number of feet per ohm, as this doesn't require you to do so much arithmetic with decimals.

AWG		Diameter in inches	Ohms per 1,000 ft	Feet per ohm	Maximum amps (insulated)
0000	●	0.46	0.049	20,400	225
000	●	0.41	0.062	16,200	200
00	●	0.365	0.078	12,800	175
0	●	0.325	0.098	10,200	150
1	●	0.289	0.124	8,070	125
2	●	0.258	0.156	6,400	100
3	●	0.229	0.197	5,080	90
4	●	0.204	0.249	4,020	80
5	●	0.182	0.313	3,190	70
6	●	0.162	0.395	2,510	60
7	●	0.144	0.498	2,010	51
8	●	0.128	0.628	1,590	44
10	●	0.102	0.999	1,000	32
12	●	0.081	1.59	630	23
14	●	0.064	2.53	396	17
16	●	0.051	4.02	249	13
18	●	0.04	6.39	157	10
20	●	0.032	10.2	99	8
22	●	0.025	16.1	62	5
24	●	0.02	25.7	39	2.5
26	●	0.016	40.1	25	1.5
28	●	0.013	64.9	15	1.0
30	●	0.010	103.2	10	0.5

Figure 5-103. *American wire gauges (AWG) and their properties.*

THEORY

Calculating voltage drop

Another fact that you often need to know is how much of a voltage drop a particular length of wire will introduce in a circuit. If you want to get maximum power from a motor, you don't want to lose too much voltage in the wires that go to and from the motor.

Voltage drop is tricky, because it depends not only on the wire, but also on how heavily the circuit is loaded. Suppose that you are using 100 feet of 22-gauge wire, which has a resistance of about 1.5 ohms. If you attach it to a 12-volt battery and drive an LED and a series resistor offering a total effective resistance of about 1,200 ohms, the resistance of the wire is trivial by comparison. According to Ohm's Law:

amps = volts / ohms

so the current through the circuit is only about 10mA.

Again, by Ohm's Law:

volts = ohms × amps

so the wire with resistance of 1.5 ohms imposes a voltage drop of 1.5 × 0.01 = 0.015 volts.

Now suppose you're running a motor. The coils in the motor create impedance, rather than resistance, but still if we measure how much current is going through the circuit, we can establish its effective resistance. Suppose the current is 1 amp. Repeating the second calculation:

volts = ohms × amps

So the voltage drop in the wire is now 1.5 × 1 = 1.5 volts! This is illustrated in Figure 5-104.

Bearing these factors in mind, I have compiled a table for you. I've rounded the numbers to just two digits, as variations in the wire that you use make any pretense of greater accuracy unrealistic.

To use this table, you need to know how much current is passing through your circuit. You can calculate it (by adding up all the resistances and dividing it into the voltage that you are applying) or you can simply measure the current with a meter. Just make sure that your units are consistent (all in ohms, amps, and volts, or milliohms, milliamps, and millivolts).

In the table, I have arbitrarily assumed a length of 10 feet of wire. Naturally you will have to make allowances for the actual length of wire in your circuit. The shorter the wire, the less the loss will be. A circuit with only 5 feet of wire, and the same amperage and voltage, will suffer half of the percentage loss shown in the table. A circuit with 15 feet of wire, and the same amperage and voltage, will suffer 1.5 times the percentage loss. So, to use the table:

1. Divide your length of wire by 10. (Make sure that you measure the length in feet.)
2. Use the result to multiply the number in the table.

The table also arbitrarily assumes that you have a 12-volt supply. Again, you will have to make allowances if you are using a different voltage. So, to use the table:

1. Divide 12 by the actual voltage of your power supply.
2. Use the result to multiply the number in the table.

I can summarize those two steps like this:

Percent voltage lost = P × (12 / V) × (L / 10)

where P is the number from the table, V is your power-supply voltage, and L is the length of your wire.

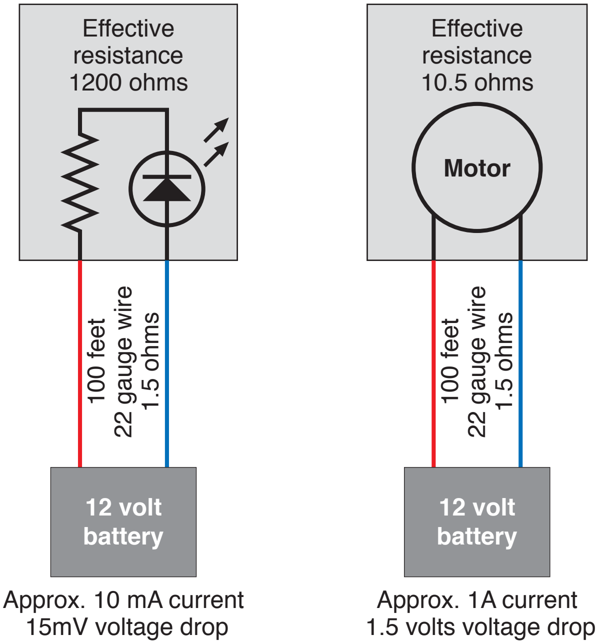
Effective resistance 1200 ohms
100 feet 22 gauge wire 1.5 ohms
12 volt battery
Approx. 10 mA current 15mV voltage drop
Effective resistance 10.5 ohms
Motor
100 feet 22 gauge wire 1.5 ohms
12 volt battery
Approx. 1A current 1.5 volts voltage drop

Figure 5-104. *The voltage drop imposed by wiring will depend on the current and the resistance in the circuit. The drop will be greatest when the resistance of the circuit is low and the amperage is high.*

THEORY

Calculating voltage drop (continued)

This table shows the percent voltage lost in a circuit with 10-foot wire at 12 volts.

Wire Gauge	Amperes									
	1	2	3	4	5	6	7	8	9	10
10	0.08	0.17	0.25	0.33	0.42	0.50	0.58	0.67	0.75	0.83
12	0.13	0.27	0.40	0.53	0.66	0.80	0.93	1.1	1.2	1.3
14	0.21	0.42	0.63	0.84	1.1	1.3	1.4	1.5	1.9	2.1
16	0.33	0.67	1.0	1.3	1.7	2.0	2.3	2.7	3.0	3.4
18	0.53	1.1	1.6	2.1	2.7	3.2	3.7	4.3	4.8	5.3
20	0.85	1.7	2.6	3.4	4.3	5.1	6.0	6.8	7.7	8.5
22	1.3	2.7	4.0	5.4	6.7	8.1	9.4	11	12	13
24	2.1	4.3	6.4	8.6	11	13	15	17	19	21
26	3.4	6.8	10	14	17	20	24	27	31	34
28	5.4	11	16	22	27	32	38	43	49	54
30	8.6	17	26	34	43	52	60	69	77	86

Remember, though, that the wire resistance will be higher if you are using stranded copper wire or tinned copper wire, and this will increase the percentage of voltage lost.

Experiment 33: Moving in Steps

Time now to build something more sophisticated: a cart that orients itself toward a light source. I'm going to tell you all you need to get started on this project, but this time I won't go all the way to the end in exhaustive detail. I want you to get into the habit of figuring out the details, improving on plans, and eventually inventing things for yourself.

You will need:

- 555 timers. Quantity: 8.
- Trimmer potentiometer, 2K linear. Quantity: 2.
- LEDs. Quantity: 4. If you get tired of using series resistors to protect LEDs in a 12-volt circuit, consider buying 12-volt LEDs such as Chicago Miniature 606-4302H1-12V, which contain their own resistors built in. However, the schematic in Figure 5-108 assumes that you will use regular 2V or 2.5V LEDs.
- Stepper motor: Unipolar, four-phase, 12-volt. Parallax 27964 or similar, consuming 100mA maximum. Quantity: 2.
- Photoresistors, ideally 500 to 3,000Ω range. Quantity: 2.

- ULN2001A or ULN2003A Darlington arrays by STMicroelectronics. Quantity: 2.
- CMOS octal or decade counter. Quantity: 2.
- Various resistors and capacitors.

Exploring Your Motor

I've specified a unipolar, four-phase, 12-volt motor because this is a very common type. A typical sample is shown in Figure 5-105. If you can't easily find the one that I've listed, you should feel safe in buying any other that has the same generic description. "Unipolar" means that you don't have to switch the power supply from positive to negative and back to positive again, to run the motor. Four-phase means that the pulses that run the motor must be applied in sequence to four separate wires. Because you will be running your motor directly from 555 timers, the lower its power consumption, the better.

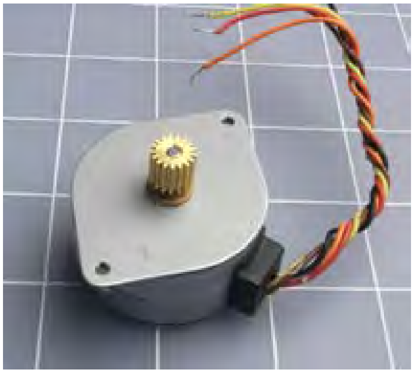

Figure 5-105. *A typical stepper motor. The shaft rotates in steps when negative pulses are applied to four of the wires in sequence, the fifth wire being common-positive.*

First, though, we can apply voltage to the motor without using any other components at all. Most likely it will have five wires already attached, with the ends stripped and tinned, so that you can easily insert them into holes in a breadboard, as shown in Figure 5-106. Check the data sheet for your motor; you should find that four of the wires are used to energize the motor and turn it in steps, while the fifth is the common connection. In many cases, the common connection should be hooked to the positive side of your power supply, while you apply negative voltage to the other four wires in sequence, one step at a time.

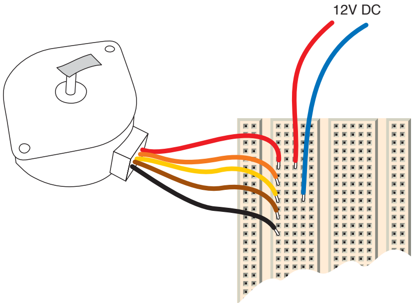

Figure 5-106. *The simplest test of a stepper motor is to apply voltage manually to each of its four control wires, while a piece of duct tape, attached to the output shaft, makes it easy to see how the motor responds.*

The data sheet will tell you in what sequence to apply power to the wires. You can figure this out by trial and error if necessary. One thing to bear in mind: a stepper motor is very tolerant. As long as you apply the correct voltage to it, you can't burn it out.

To see exactly what the motor is doing, stick a piece of duct tape to the end of the shaft. Then apply voltage to wires, one at a time, by moving your negative power connection from one to the next. You should see the shaft turning in little steps.

Inside the motor are coils and magnets, but they function differently from those in a DC motor. You can begin by imagining the configuration as being like the diagram in Figure 5-107. Each time you apply voltage to a different coil, the black quadrant of the shaft turns to face that coil. In reality, of course, the motor turns less than 90° from one coil to the next, but this simplified model is a good way to get a rough idea of what's happening. For a more precise explanation, see the upcoming section "Theory: Inside a stepper motor."

Bear in mind that as long as any of the wires of the motor are connected, it is constantly drawing power, even while sitting and doing nothing. Unlike a regular DC motor, a stepper motor is designed to do nothing for much of the time. When you apply voltage to a different wire, it steps to that position and then resumes doing nothing.

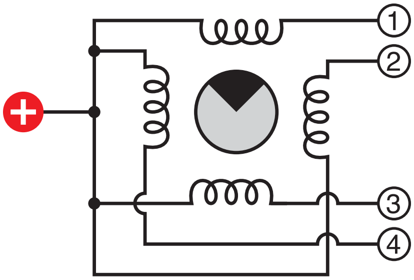

Figure 5-107. *This greatly simplified diagram helps in visualizing the way in which a stepper motor works. In reality, almost all motors rotate by less than 90° in response to each pulse.*

The coil inside the motor is holding the shaft in position, and the power that the motor draws will be dissipated as heat. It's quite normal for the motor to get warm while you're using it. The trouble is, if you use a battery to power it, and you forget that you have it connected, the battery will not hold its charge for long.

Quick Demo

Now that you've proved that your motor is functional, how can you actually run it? You need to send a pulse to each of the four wires in turn, in a rapidly repeating sequence. If you can also adjust the speed of the pulses, so much the better. I'm thinking that for a quick and simple demo, you can handle the challenge simply by using four 555 timers, all of them in monostable mode, with each one triggering the next.

The schematic in Figure 5-108 shows what I have in mind. It looks more complex than it really is. Each timer has the same pattern of components around it, so after you create the first module, you just make three copies of it.

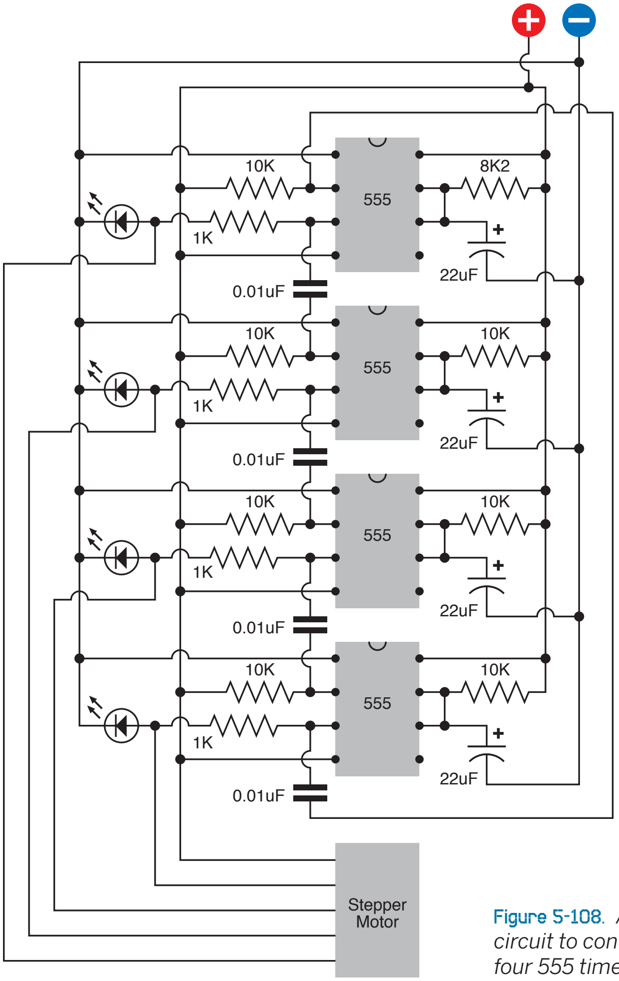

Figure 5-108. *A very quick and simple circuit to control a stepper motor uses four 555 timers, each in monostable mode, triggering each other in a repeating sequence.*

I've used a 10K resistor to pull up the input to each 555, so that the timers are naturally in their quiescent state. A 0.01 µF capacitor links the output from one timer to the input of the next so that they are electrically isolated from each other, and the capacitor just conveys a "spike" of voltage when one timer finishes its "on" cycle, and its output goes low, which triggers the next.

On the righthand side, I've used 10K resistors and 22 µF capacitors to generate a cycle of about a quarter of a second—except that the topmost timer has a 8K2 timing resistor. The reason for this is that when power is first applied, the timers will all be waiting for each other to begin, and timers 2 and 4 or 1 and 3 may fire together. By giving one timer a shorter cycle than the others, I minimize this problem.

The LEDs are included just to give you some visual verification of what's happening. Without them, if you make a wiring error, the motor may turn to and fro erratically, and you won't know why. Initially you can run your circuit with only the LEDs connected, just to make sure it works. Figure 5-109 shows the breadboarded circuit before the motor is plugged in. Then add the motor by plugging its wires into the breadboard, where you'll make connection with the outputs (pins 3) of the timers. See Figure 5-110.

Apply power, and you should see the motor turning in steps, in sequence with the LEDs. If the LED sequence isn't stable:

1. Connect a wire directly from the input (pin 2) of the topmost timer to the positive side of the voltage supply, and wait for the timers to calm down.
2. Restart the sequence by disconnecting the free end of this wire, or (if necessary) touch the free end of it briefly to the negative side of the supply, to trigger the first timer.

One thing you may have noticed, if you're paying very close attention: the common terminal of the motor is connected to positive. Therefore, when each timer flashes positive, that positive signal isn't actually powering the motor. The *low* outputs from the three timers that are *not* firing at any given moment are sinking current from the motor. It seems quite happy with this arrangement. You'll need some theory to understand why.

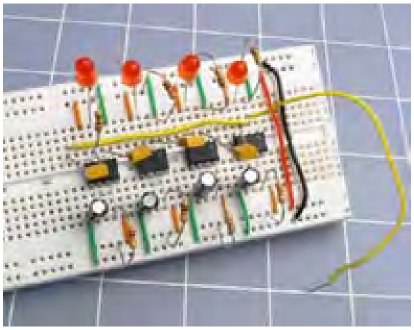

Figure 5-109. *To test the control circuit for errors, four LEDs show the outputs from the four 555 timers. The loose yellow wire at the righthand side connects to pin 2 of the first timer. Touch the free end of this wire to the positive side of the power supply to reset the timers, and then, if necessary, make a brief negative connection with the free end of the wire to restart their sequence.*

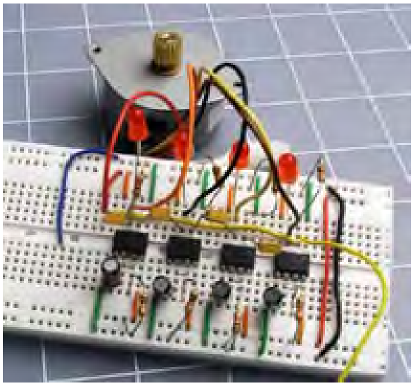

Figure 5-110. *After the circuit has been tested, the motor can be added by hooking its control wires to the outputs of the four 555 timers.*

THEORY

Inside a stepper motor

If you check the Wikipedia entry for stepper motors, you may see a very nice 3D rendering showing a toothed rotor and four coils arrayed around it. Maybe stepper motors used to be manufactured like this once upon a time, but not anymore.

Imagine two horizontal rows of coils. In the space between them is a series of little magnets, like a freight train, that can move left or right, as shown in Figures 5-111 and 5-112. Each coil has two windings, in opposite directions, so that current through one winding will create an upward magnetic force while current through the other will create a downward force. Each row of windings is connected in parallel, so that they switch on and off simultaneously.

In Step 1, the negative connection energizes the upper windings of the upper coils, which creates an upward magnetic force. I've shown this force using blue-green arrows so that you won't mistake it for a flow of electricity. It so happens that this force attracts the north poles of the magnets and repels the south poles, so if the magnets begin in the position shown in Step 1, they will want to move one step to the right.

This brings them to the position shown in Step 2. Now the upper windings of the lower coils are energized, and again, this produces an upward force, which again attracts the north poles and repels the south poles.

This advances the magnets to their location in Step 3. Now the lower windings of the upper coils are energized, producing a downward force. This repels the north poles of the magnets and attracts their south poles. So the magnets keep moving.

They reach the position shown in Step 4. The lower windings of the lower coils are energized, producing a downward force which continues to attract the south poles while repelling the north poles. So the magnets move a final step to the right—which leaves them in the same orientation shown in Step 1. And the process can repeat all over again.

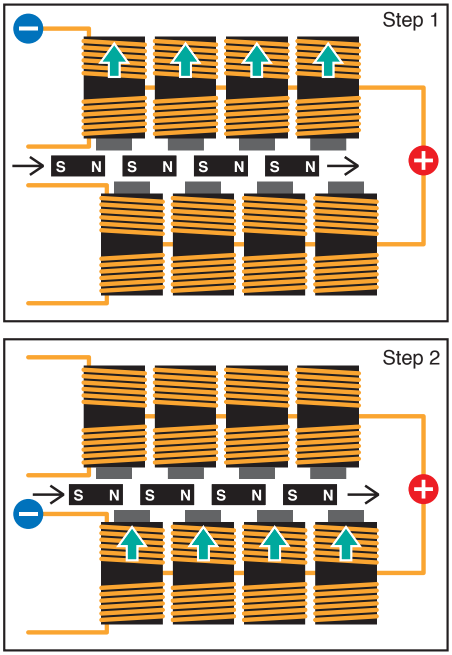

Figure 5-111. *This sequence shows the first two steps as the rotor of a stepper motor (shown as a series of north-south magnets) moves in response to pulses through electromagnets.*

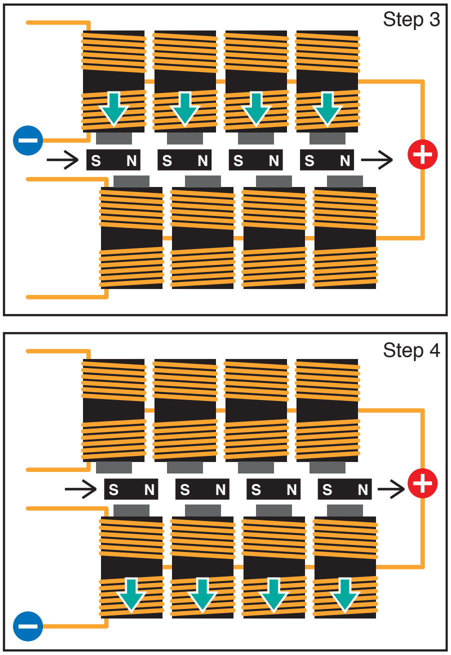

Figure 5-112. *After taking another two steps, the motor will be back where it started at Step 1 in Figure 5-111.*

THEORY

Inside a stepper motor (continued)

In reality, the magnets are not separate from each other. The edge of a rotor is magnetized in zones that alternate between south and north polarity. And instead of multiple coils, there are just four windings that go around all the magnetic cores. But the principle is exactly the same. The 3D rendering gives a general idea, and the photograph shows what I found when I cracked open a typical stepper motor.

Now bear in mind that when this device is driven by a set of 555 timers, we don't just connect negative to one wire at a time on the left, leaving the others floating. In reality, at any given moment, three of the timers have a negative output and the fourth has a positive output. The last diagram in Figure 5-112 shows this situation.

Suppose the top wire is positive while the other three are negative, as shown in Figure 5-113. The positive output does nothing, because it is balanced by the positive power on the other end of the coils. The two negatives attached to the bottom set of coils create equal and opposite forces that cancel each other out (while wasting some power). So the net result is the same as in Step 3.

In fact, you should find that you can disconnect the common wire completely while using the stepper motor with 555 timers, and the motor will still turn, because one of the timers is providing positive power while the others are negative. In fact, you'll be running them more efficiently this way.

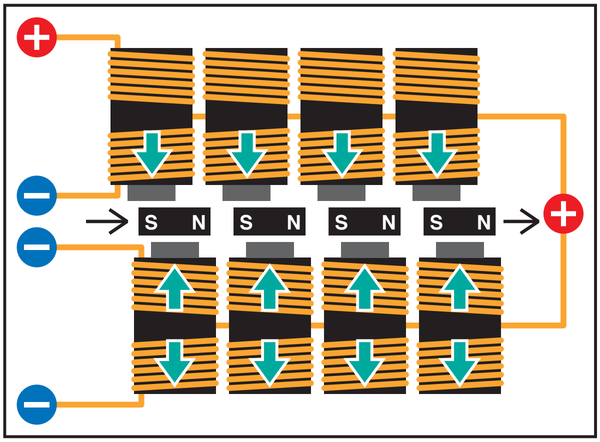

Figure 5-113. *When the motor is driven by four 555 timers, they are activating it by sinking positive voltage from it. The interior workings of the motor look something like this. It's not the most efficient way to do the job.*

Figures 5-114 and 5-115 may help to give you a clearer idea of what the motor actually looks like inside.

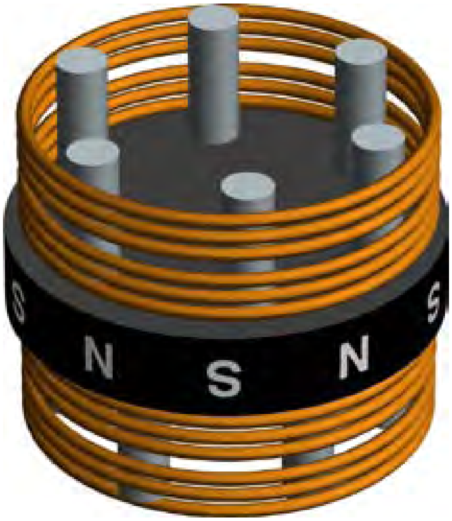

Figure 5-114. *This 3D rendering gives a better idea of what a typical stepper motor looks like inside. The copper coils and gray cylinders are stationary, while the black disc rotates between them.*

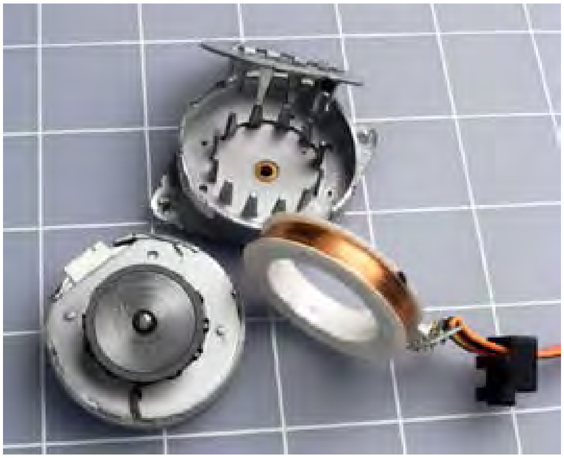

Figure 5-115. *When a stepper motor is broken open, this is what you're likely to find. On the left, the rotor of the motor, which has a magnetized band around its circumference, is still attached to the lower half of the casing. On the right, the upper half of the casing has been opened, and the coil has been removed (actually the winding you can see consists of two coils, wound in opposite directions). The spikes are the magnetic cores that exert force on the rotor.*

Speed Control

If you are a truly exceptionally observant, you may have noticed that I left pin 5 of each of the timers unconnected in the schematic for driving the stepper motor in Figure 5-108. Normally, pin 5 should be grounded through a capacitor to prevent it from picking up stray voltages which can affect the accuracy of the chip.

I left the pins unconnected because I had a plan for them. In fact, changing the timing of the chip is exactly what we want to do now, as a way to change the speed of the stepper motor.

If you tie pin 5 of all four timers together, as shown in Figure 5-116, and put a 2K trimmer potentiometer (shown in Figure 5-117) between them and the negative side of the power supply, you'll find that as you turn the trimmer to reduce its resistance, the timers start to run faster. Figure 5-118 shows the breadboard layout. Eventually, when the resistance goes below around 150 ohms, everything stops. The LEDs go dark, because you've reduced the voltage on pin 5 below the threshold level that the 555 timer finds acceptable.

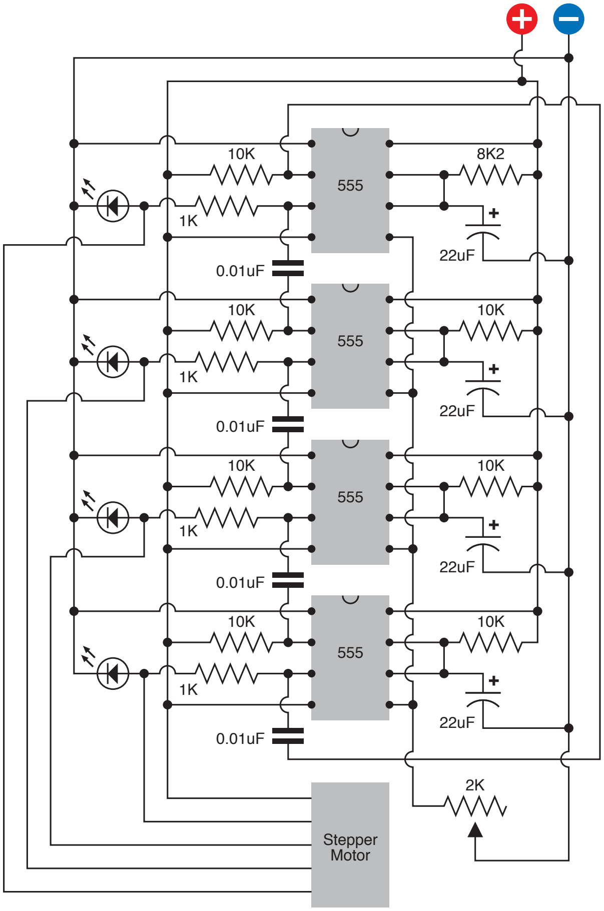

Figure 5-116. *To adjust the speed of the sequence of 555 timers, their control pins (pin 5 on each timer) are linked together and attached to a trimmer potentiometer that adjusts the resistance between the pins and the negative side of the power supply.*

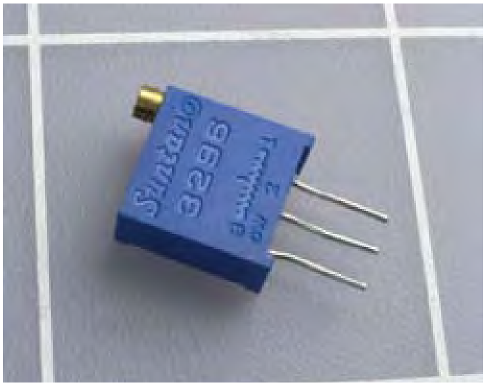

Figure 5-117. *Close-up of a trimmer potentiometer with pins spaced at 1/10 inch for insertion in a breadboard or perforated board. The brass screw, at top-left, turns a worm gear inside the unit, allowing precise adjustment of internal resistance.*

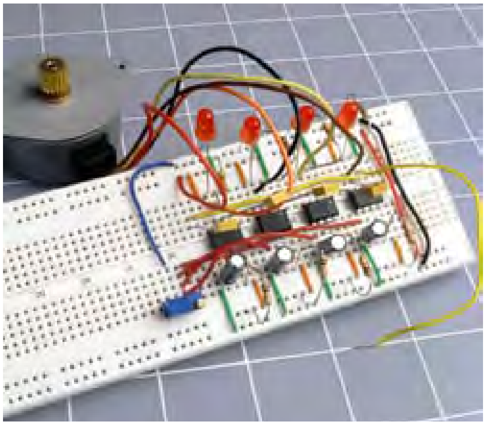

Figure 5-118. *The trimmer potentiometer has been added to the circuit, allowing motor speed control.*

Initially I suggested a step time of 1/4 second just so that you could see what was happening. When you're actually using this circuit, you'll never need it to run as slowly as that. So you can increase the entire range of speeds. Remove the 22 µF timing capacitors and substitute, say, 4.7 µF capacitors, or smaller. Now when you adjust the potentiometer, you'll get a useful range of speed.

Adding Autonomy

Currently, the circuit simply does what you tell it to do. The next step is to make it autonomous—in other words, give it the illusion of making up its own mind. I'm thinking that instead of a trimmer potentiometer, we could substitute a photocell, properly known as a photoresistor. Typically, the resistance of a cadmium sulfide photo resistor is highest in the dark, and lowest when light shines on it.

One problem with photoresistors is that they're not as widely available as many other types of electronic components. If you search Mouser.com, for instance, you'll find virtually nothing. Partly this is because the online search function at Mouser is the weakest feature of the site, and partly it's because Mouser is not oriented toward hobbyists. What you need to do is conduct a "product search." Go to *http://www.google.com/products*, enter the search terms "CdS" and "photocell," and you'll find a bunch of cheap cadmium sulfide components from places you may never have heard of.

Because photoresistors seem to come and go as erratically as DC motors, I am not offering any part numbers. You can buy any product that has an appropriate minimum resistance (in bright light) and maximum resistance (in the dark). If you find a component that ranges from 500 to 3,000Ω, that would be a good choice. If the only ones you can find have a higher minimum than 500Ω, you could consider putting a couple of them in parallel.

Setting Up Your Light Seeking Robot

Why would you want to control the speed of a stepper motor by using a photo resistor? Because the original objective was to build a robot that is attracted to light.

The idea is simple enough: use two stepper motors, each controlling the speed of one wheel of the cart. Use two photoresistors, each controlling the speed of the opposite stepper motor. When the righthand photoresistor picks up more light, its resistance lowers, causing the lefthand set of timers to run faster, which will make the lefthand wheel run faster. Thus, the cart will turn toward the light. Figure 5-119 illustrates the concept.

Before you start wiring more 555 timers, though, you might consider doing the job with a more appropriate component. The ULN2001A and ULN2003A are chips containing Darlington amplifiers specifically designed to deliver current to inductive loads such as solenoids, relays, and (you guessed it) motors. Each chip has seven inputs that require very little current, and seven outputs that can deliver 500mA each. The inputs are TTL and CMOS compatible (the 2001A has a wider tolerance for voltages than the 2003A) and each channel of

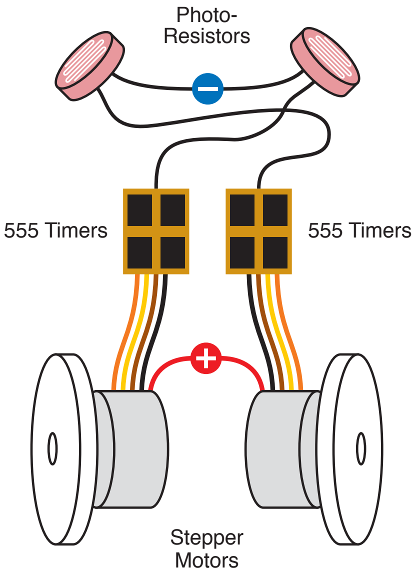

Figure 5-119. *If two photoresistors control the speed of two 555-timer arrays, the difference in speed between one wheel and the other can turn the cart toward a light source.*

the chip functions as an inverter, so that when the input goes high, the output goes low and sinks current. This is of course just what we need for our stepper motor that has a common positive connection.

The ULN2001A is only an amplification device, so you have to precede it with a counter that runs from 1 to 4 and then repeats. You can stick with your 555 timers, as you've already assembled them, or substitute almost any CMOS octal or decade counter that sends its output pulses to a series of pins. Just use the output from the fifth pin as the "carry" output to restart the counting sequence. I suggest a CMOS counter simply because it will run on 12 volts, so you can use the same power supply that suits your stepper motors.

If you switch to CMOS counters, you will still need a pair of 555 timers sending pulses to the counters. The timers will be free-running in astable mode, and your photoresistors will control their speed. Figure 5-120 shows the configuration.

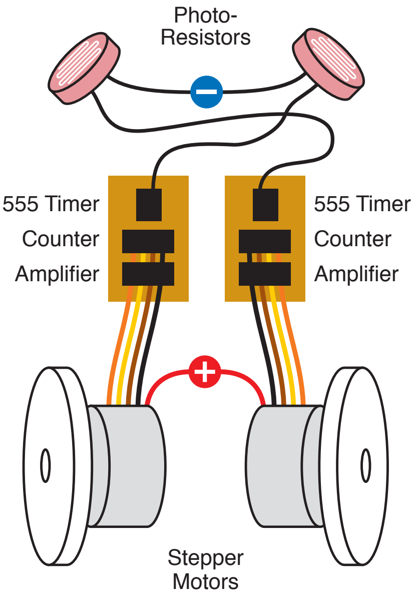

Figure 5-120. *A more efficient way to drive the motors is to use just one timer to set the speed of each, with a counter and amplifier (such as a Darlington array chip) sending the pulses down the wires. The principle is still the same, though.*

One last item: you'll need a 12-volt battery. You can of course put eight AA cells together, but I think you should consider a rechargeable pack from a source such as *http://www.all-battery.com*, which has a section entirely devoted to "robot batteries."

If you put it all together, you should find that when you place your robot cart in a very dimly lit room, it will turn toward the beam from a bright, well-focused flashlight. To get reliable results, you may have to recess each of the photoresistors in little tubes, so that they receive much more light when they face your flashlight than when they face away from it. Figure 5-121 is a 3D rendering of the concept.

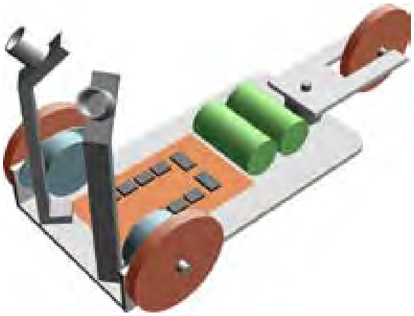

Figure 5-121. *This 3D rendering shows a possible configuration of the light-seeking cart, with two photoresistors enclosed in small tubes to restrict their response to light.*

Another idea is to rewire your cart so that it actually runs away from the light. Can you imagine how this might be done?

Just one more thought: if you use infrared photoresistors, you can control your cart with beams from infrared LEDs, in normal room lighting. If you and a couple of friends all have infrared transmitters, you can get your cart to run from one of you to the next, like an obedient dog.

This takes us about as far as I'm going to go into robotics. I urge you to check out the sites online if you want to pursue the topic further. You can also buy a wide variety of robot kits, although of course I feel that it's more fun to invent or develop things for yourself.

All that's left now is to perform one last introduction: to a device that should make your life much easier, even though the device is much more complicated than anything we have dealt with so far.

Experiment 34: Hardware Meets Software

Throughout this book, in accordance with the goal of *learning by discovery*, I have asked you to do an experiment first, after which I've suggested the general principles and ideas that we can learn from it. I now have to change that policy, because the next experiment involves so much setup that it's only fair to tell you what to expect before you begin the preparations.

We are about to enter the realm of controller chips, often known as MCUs, which is an acronym for micro controller unit. An MCU contains some flash memory, which stores a program that you write yourself. The flash memory is like the memory in a portable media player, or the memory card that you use in a digital camera. It needs no electricity to power it. In addition, the chip has a processor which carries out the instructions in your program. It has RAM to store the values of variables on a temporary basis, and ROM, which tells it how to perform tasks such as sensing a varying voltage input and converting it into digital form for internal use. It also contains an accurate oscillator, so that it can keep track of time. Put it all together, and it's a tiny computer that you can buy for under $5.

Let's suppose that you have a greenhouse where the temperature must never fall below freezing. You set up a temperature sensor, and you have two different heaters. You want to switch on the first heater if the temperature falls below 38° Fahrenheit. But if, for some reason, that heater is broken, you want to switch on the second, backup heater when the temperature goes below 36° Fahrenheit.

Programming a MCU to take care of this can be very simple indeed. You could even add extra features, such as a second temperature sensor, just in case the first one fails, and you could tell the chip to use whichever sensor gives a lower reading.

Another application for a MCU would be in a fairly elaborate security system. The chip could monitor the status of various intrusion sensors, and can take various preprogrammed steps, depending on the sensors' status. You could include delay intervals, too.

Many MCUs have additional useful features built in, such as the ability to control servo motors that turn to a specific angle in response to a stream of pulses. Servos are widely used in radio-controlled model boats, airplanes, and hobby robotics.

Perhaps you are now wondering why, if MCUs can do all this, haven't we been using them all along? Why did I spend so much time describing the development of an alarm system using discrete components, if one chip could have done everything?

There are three answers:

1. MCUs cannot do everything. They need other components to help them interact with the world, such as transistors, relays, sensors, and amplifiers. You need to know how those things work, so that you can make intelligent use of them.
2. MCUs can introduce their own kinds of problems and errors, associated with using software in addition to hardware. I'll have more to say about this later.
3. MCUs have limits and restrictions, most obviously their requirement for a 5-volt regulated power supply, and their inability to source or sink much current from each pin. They also demand that you learn a programming language (which differs from one brand of MCU to the next). And to get the program into the chip, you have to be able to plug it into a computer and do a download, which is not always convenient.

In this experiment, you'll learn how to write a program for a small and simple MCU, and you'll transfer the program into it and see how it works.

BACKGROUND

Origins of programmable chips

In factories and laboratories, many procedures are repetitive. A flow sensor may have to control a heating element. A motion sensor may have to adjust the speed of a motor. Microcontrollers are perfect for this kind of routine task.

A company named General Instrument introduced an early line of MCUs in 1976, and called them PICs, meaning Programmable Intelligent Computer—or Programmable Interface Controller, depending which source you believe. General Instrument sold the brand to another company named Microchip Technology, which owns it today.

"PIC" is trademarked, but is sometimes used as if it's a generic term, like Scotch tape. In this book, I've chosen a range of controllers based on the PIC architecture. They are licensed by a British company named Revolution Education Ltd., which calls its range of chips the PICAXE, for no apparent logical reason other than that it sounds cool.

I like these microcontrollers because they were developed originally as an educational tool and because they are very easy to use. They're cheap, and some of them are quite powerful. Despite their odd name, I think they're the best way to get acquainted with the core concepts of MCUs.

After you play with the PICAXE, if you want to go farther into MCUs, I suggest the BASIC Stamp (which uses a very similar language, but with additional powerful commands) and the very popular Arduino (which is a more recent design, packed with powerful features, but requires you to learn a variant of C language to program it). I'll say more about these chips later.

If you search for "picaxe" on Wikipedia, you'll find an excellent introduction to all the various features. In fact, I think it's a clearer overview than you'll get from the PICAXE website.

Supplies

Figure 5-122 shows some of the chips in the PICAXE range. I'll be telling you how to use the smallest—the 08M—which costs less than $5 and is cheaper than any other MCU that I've found. It has only 256 bytes of memory to store a program (not gigabytes, megabytes, or kilobytes, just 256 bytes!), but you'll be surprised how many possibilities this amount allows for. Figure 5-123 shows a closeup of the 08M with its legs safely embedded in a piece of conductive foam.

In the United States, there are three distributors of this chip:

- *http://www.advancedmicrocircuits.com*
- *http://www.phanderson.com/picaxe*
- *http://www.sparkfun.com*

I like P. H. Anderson for its grass-roots hobbyist approach, and they have good prices if you want to buy multiple chips. But SparkFun Electronics offers other associated products that you may find interesting.

All the distributors will want to sell you "starter kits," such as the one in Figure 5-124, perhaps because the PICAXE itself is so cheap that it doesn't offer much of a profit margin. Still, for our purposes, you should buy the chip as a standalone item. And buy two of them, just in case you damage one (for example, by connecting voltage to it incorrectly).

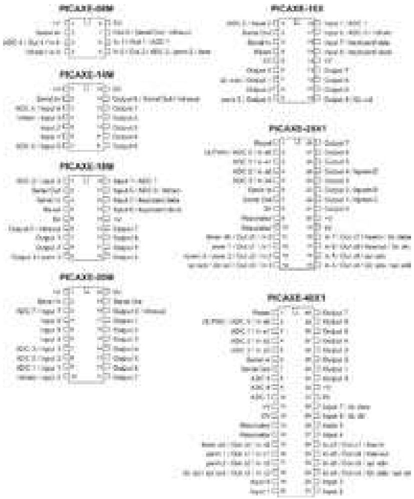

Figure 5-122. *A page from the PICAXE catalog lists only some of the chips that are available. What began as an educational aid has become a useful prototyping tool.*

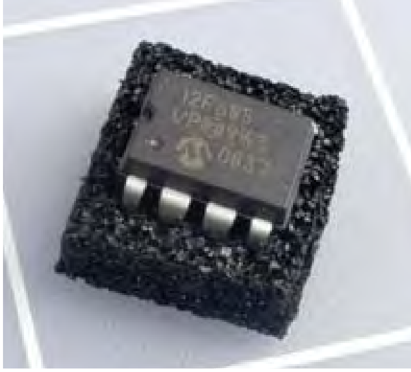

Figure 5-123. *When supplied by one of its American distributors, a PICAXE 08M arrives embedded in a little square of conductive foam. The chip is the same size as a 555 timer but has the power of a tiny computer.*

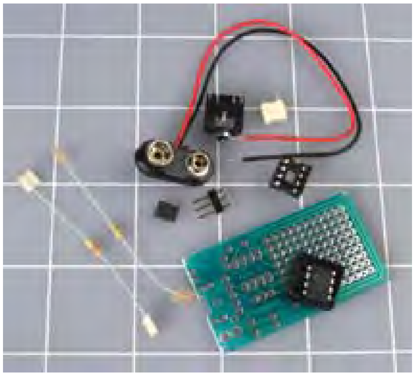

Figure 5-124. *A typical PICAXE kit includes a printed circuit board, which you may not really need, and some other not-entirely-essential items. But the 3.5-mm stereo jack socket (top, center) is absolutely necessary.*

To download your programming instructions into the chip, first you'll type the instructions on a computer, and then you'll feed them through a cable into the PICAXE memory. So you'll need to buy a cable, and you'll need software to help you to write the program.

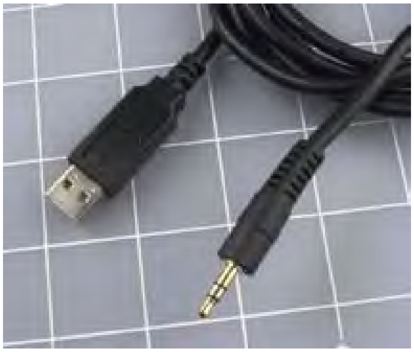

Figure 5-125. *The USB download cable made for use with the PICAXE terminates in a 3.5-mm audio plug. This should not be inserted in any audio device. It establishes a serial connection with a computer, allowing program code to be downloaded into the chip.*

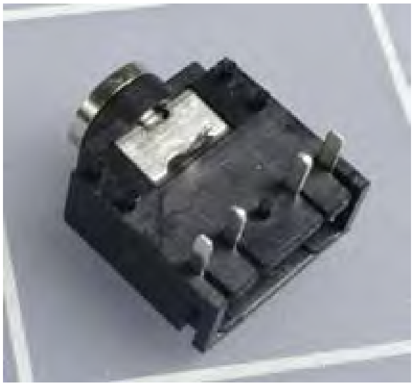

Figure 5-126. *Closeup of the 3.5-mm stereo socket that is used with the USB download cable.*

You can use the PICAXE with a serial cable, but I don't recommend it. The old RS-232 serial communications standard on PCs is pretty much obsolete, and PICAXE has recognized this by offering a USB cable (which contains a serial converter inside its plug). The USB cable is a little more expensive, but is also simpler and compatible with Apple computers. From any of the U.S. distributors, buy USB cable part AXE027, also sold as part PGM-08312 by *http://www.sparkfun.com* (quantity: 1). The cable is shown in Figure 5-125.

To write your software and send it down the wire to the chip, the PICAXE Programming Editor is the tool of choice. It comes in only a Windows version. For those who prefer Mac OS or Linux, you can get a free download of another piece of software known as AXEpad, which has fewer features, but will do the job. All the downloadable software is freely available from *http://www.rev-ed.co.uk/picaxe/software.htm*.

Finally, you need a 3.5-mm stereo audio socket with solder connections. The reason for this is that the manufacturers of the PICAXE have used a stereo audio plug on the free end of their USB cable, and you have to be able to plug it into something. The PICAXE breadboard adapter, SparkFun stock number DEV-08331, contains the necessary stereo socket in addition to a few other little items. Quantity: 1. See Figure 5-126.

Oddly enough, the USB cable is the most expensive item on the list, because of the electronics hidden inside it.

Software Installation and Setup

Now you have to go through a setup procedure. There is no way around this. Here is what you will be doing:

1. Install a driver so that your computer will recognize the special USB cable.
2. Install the Programming Editor software (or AXEpad for Mac/Linux) so that your computer will help you to write programs and then download them into the chip.
3. Mount the PICAXE on your breadboard and add the socket to receive downloads.

These steps are explained in the following sections.

The USB driver

Fair warning: If you go to the PICAXE website and try to use its search function, it probably won't find what you want. Search for "USB Driver," for instance, and it will pretend it has never heard of such a thing.

The PICAXE home page also has irritating drop-down menus that tend to disappear just when you're about to click on them, but at the time of writing, you can bypass these issues by going straight to the Software Downloads section at *http://www.rev-ed.co.uk/picaxe/software.htm*.

Scroll down past all the software until you get to Additional Resources. Look for the AXE027 PICAXE USB Download Cable. At first glance, it looks as if they want to sell you a cable, but in fact this is the list of drivers. Double-click the one appropriate to your computer, and choose a destination on your computer for the download—a place where you will find it easily, such as your desktop.

Be careful not to download the driver for the USB010 USB-Serial adapter by mistake. The USB-Serial adapter is something else entirely.

The download will leave you with a zipped file folder. You will have to unzip it. On Windows XP, right-click the folder and choose "Extract all." View the extracted files and you will find a PDF installation guide. Linux and Mac users can find instructions currently stashed at *http://www.rev-ed.co.uk/docs/AXE027.pdf*.

When installing the driver on a Windows platform, here are a few tips to minimize your exasperation level:

1. Remember, the special USB cable contains some electronics. It is not just a cable, but a device designed for interacting with a PICAXE chip. *Don't try to use it for anything else!*
2. You have to plug the cable into a USB port *before* you install the driver, because your computer will need to verify that the driver matches the cable.
3. You must not attach the PICAXE to the other end of the cable until *after* you have installed the driver.
4. Every USB port on your computer has a separate identity. Whichever one you choose when you first plug in the cable, you should use that port *every time in the future*. Otherwise, you will have to repeat the process of telling your computer what the cable is.
5. Bearing in mind Tip #4, you should avoid using the cable in a standalone USB hub.
6. The cable is fooling the PICAXE into thinking that it's talking to a serial port on your computer. Those "communication" ports are known as COM1, COM2, COM3, or COM4. When you install the driver, the installer will choose one of those COM ports for you, and later you will have to know which one it is. The PDF guide should help you through this procedure. Unfortunately, you cannot skip it.

The Programming Editor software

If you have come this far, you're ready for the next big step, which is much easier. You need the PICAXE Programming Editor, available for free on the Software Downloads web page where you found the USB driver. (If you are using a Mac or Linux, you will need AXEpad, which is on the same web page.)

Downloading and installing the Programming Editor should be simple and painless. Once you have done that, you should find that it has placed a shortcut on your desktop. Double-click it, go to View→Options, and in the window that opens (shown in Figure 5-127), click the Serial Port tab. You should see a dialog box like the one in Figure 5-128. Now make sure that the Programming Editor is looking at the same COM port that was chosen by the USB driver. Otherwise, the Programming Editor won't know where to find your PICAXE chip.

In the Programming Editor, go to View→Options and click the Mode tab, then click the button to select the 08M chip.

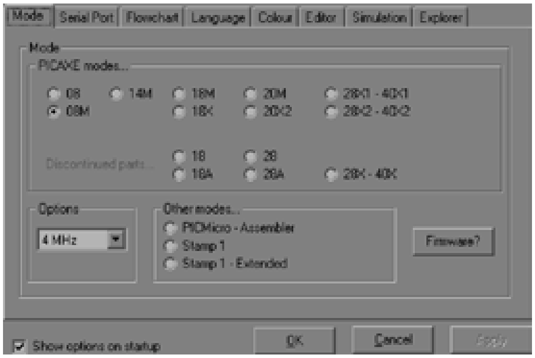
Mode
Serial Port
Flowchart
Language
Colour
Editor
Simulation
Explorer
Mode
PICAXE modes...
○ 08
○ 14M
○ 18M
○ 20M
○ 28X1 - 40X1
● 08M
○ 18X
○ 20X2
○ 28X2 - 40X2
Discontinued parts...
○ 18
○ 28
○ 18A
○ 28A
○ 28X - 40X
Options
4 MHz
Other modes...
○ PICMicro - Assembler
○ Stamp 1
○ Stamp 1 - Extended
Firmware?
☑ Show options on startup
OK
Cancel
Apply

Figure 5-127. *This screenshot shows the options window of the PICAXE Program Editor, which you must use to select the type of chip that you intend to program (in our case, the 08M).*

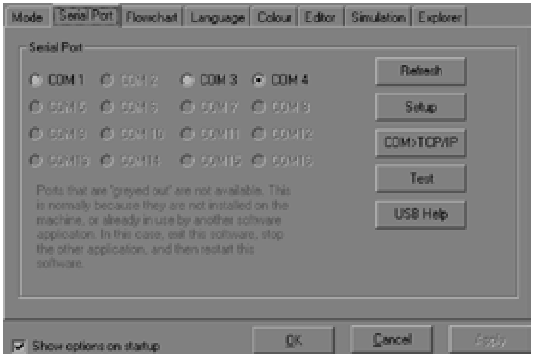
Mode
Serial Port
Flowchart
Language
Colour
Editor
Simulation
Explorer
Serial Port
○ COM 1
○ COM 2
○ COM 3
● COM 4
○ COM 5
○ COM 6
○ COM 7
○ COM 8
○ COM 9
○ COM 10
○ COM11
○ COM12
○ COM13
○ COM14
○ COM15
○ COM16
Refresh
Setup
COM>TCP/IP
Test
USB Help
☑ Show options on startup
OK
Cancel
Apply

Ports that are 'greyed out' are not available. This is normally because they are not installed on the machine, or already in use by another software application. In this case, exit this software, stop the other application, and then restart this software.

Figure 5-128. *Another screenshot of the options window shows the second essential choice that you must make: selection of the COM port that the installer chose on your computer.*

Are we having fun yet? Obviously not, but you're through with software hassles for the time being. The last step before you're ready to use the PICAXE is to mount it, and its socket, on your breadboard.

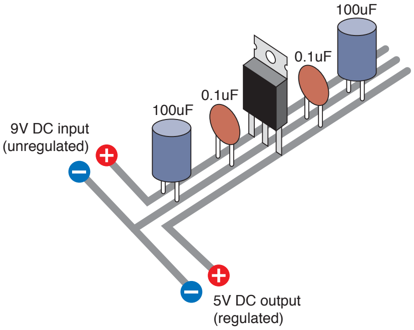
100uF
0.1uF
0.1uF
100uF
9V DC input
(unregulated)
5V DC output
(regulated)

Figure 5-129. *PICAXE documentation specifies a 100 μF and 0.1 μF capacitor on the input side of a 5-volt regulator, and a similar pair of capacitors on its output side. On a breadboard, they can be arrayed like this.*

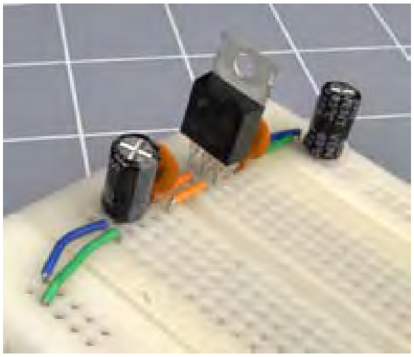

Figure 5-130. *The actual components for power regulation, applied to a breadboard, delivering 5 volts (positive and negative) down each side.*

Setting up the hardware

The PICAXE 08M looks like a 555 timer. (Other chips in the PICAXE range have more pins and more features.) It requires a properly regulated 5 volts, just like the logic chips you dealt with previously. In fact, the PICAXE people are rather emphatic about protecting it from voltage spikes. They want you to use two capacitors (one 100 μF, one 0.1 μF) on either side of an LM7805 regulator. This seems like overkill, but the PICAXE is more inconvenient to replace than a 555 timer. You certainly can't run down to RadioShack to buy one. So let's do what the manufacturer says, just in case, and set up a breadboard as shown in Figures 5-129 and 5-130.

Now for the chip itself. Note that the pins for positive and negative power are *exactly opposite* to those for the 555 timer, so be careful!

Set up your breadboard following the schematic shown in Figure 5-131. Note that I am showing the stereo socket on its underside, because I think that's how you'll have to use it with the breadboard. If you try to stick its pins into the holes in the board, they will fit, but when you insert the plug into the socket, the thickness of the plug will tend to raise the socket up so that it loses contact. I really think the way to go is to solder wires to the pins on the socket and push the wires into the breadboard. See Figure 5-133.

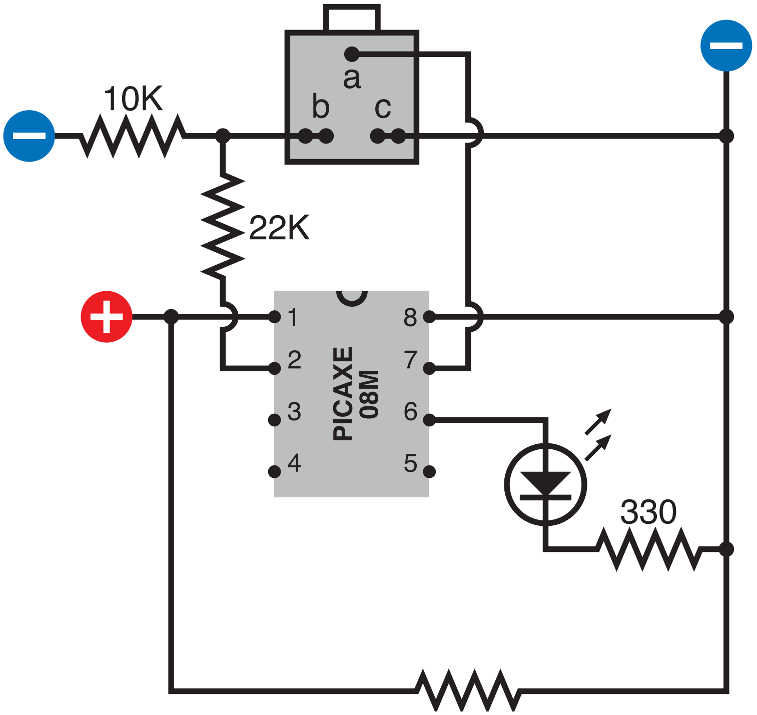

Figure 5-131. *The schematic of a test circuit for the PICAXE 08M shows the underside of the stereo socket, the essential 10K and 22K resistors on the input pin, and an LED to show an output from the chip.*

Be aware that the PICAXE manual shows things differently (although I have retained their labeling convention for the parts of the socket and the parts of the plug, identified as a, b, and c).

One little detail about the socket that is commonly supplied for use with the PICAXE: typically it has two *pairs* of contacts for the connections labeled b and c in the manual, and in my diagram. When you solder a connection, your solder joint should include *both* of the contacts in each pair, as shown in Figure 5-132.

Remember that the PICAXE must have 5 volts DC, and remember that your voltage regular will deliver this voltage reliably only if you give it extra voltage on its input side. If you provide it with 9 volts, that will provide a good amount of headroom.

The 22K and 10K resistors are essential for using the chip; see the following warning note for an explanation. My schematic also includes an LED and a 330Ω resistor, but they are needed only for the test that we'll be making momentarily.

Pin 2 Pull-Down

Always include the 22K resistor and the 10K resistor in the configuration shown in Figure 5-131. These resistors apply correct voltage to the serial connection, and when you're using the PICAXE on its own, they pull down the voltage on pin 2.

If pin 2 is left unconnected (floating), it may pick up random voltages, which the chip can misinterpret as a new program or other instructions, with unpredictable and undesirable results.

The 22K and 10K resistors should be regarded as permanent items accompanying your PICAXE regardless of whether you have it attached to your computer.

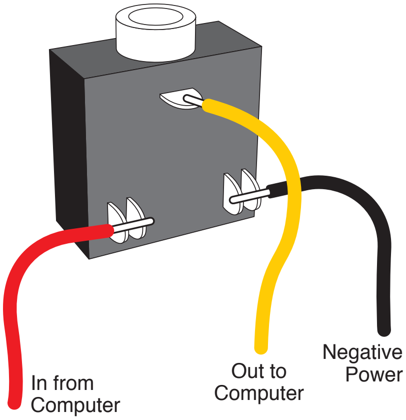

Figure 5-132. *Correct wiring of the socket is essential. When soldering wires to the lower terminals, make sure that you attach the wires to both of the terminals in each pair.*

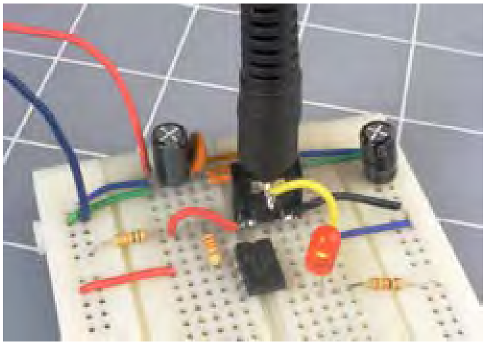

Figure 5-133. *The breadboard version of the test schematic, with the plug of the USB download cable inserted in the socket on the board. The PICAXE chip can now receive a downloaded program, and will immediately start to execute it.*

Verifying the Connection

Follow these steps carefully every time you want to program or reprogram your PICAXE chip:

1. Insert the USB plug of your PICAXE cable into the same USB port that you used before.
2. Start the Programming Editor (or AXEpad if you are using a Mac OS or Linux).
3. In the Programming Editor, select View→Options to verify that the editor is using the right COM port and is expecting the 08M PICAXE chip.
4. Plug the stereo plug on the free end of the USB cable into the stereo socket that is now wired into your breadboard. See Figures 5-133 and 5-134.
5. Check your wiring, and then connect your power supply to the breadboard.
6. Click the button labeled "program" in the Program Editor window to tell the software to look for the PICAXE.

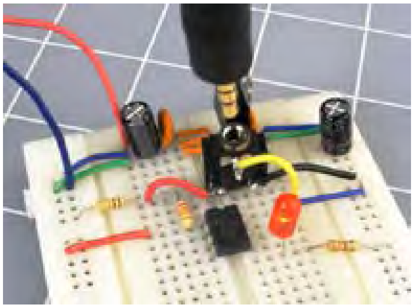

Figure 5-134. *After the program has been downloaded, the plug can be removed, and the program will continue to run, causing the LED to continue flashing.*

What If It Doesn't Work?

The first thing to do is pull out the plug of the USB cable from the PICAXE breadboard, leaving the other end of the cable attached to your computer. Set your multimeter to measure DC volts, and attach its probes to sections b and c of the plug. See Figure 5-135. Now click the "program" button again, and your meter should show 5 volts briefly coming out of your computer to the plug on the end of the cable.

If you detect the voltage, the software is installed and working properly. In that case, there's a problem on your breadboard, either in the chip or in the wiring around it.

If you cannot detect any voltage, the software probably wasn't installed properly, or is looking for the wrong serial port. Try uninstalling it and reinstalling it.

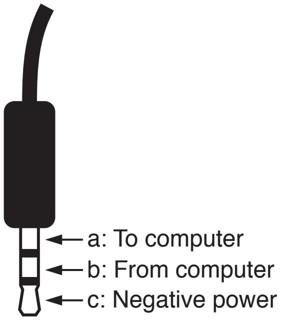
a: To computer
b: From computer
c: Negative power

Figure 5-135. *The stereo plug on the end of the USB download cable can be used for fault tracing. A multimeter set to measure DC volts can be attached to sections b and c of the plug to establish whether the Programming Editor is sending data through the serial connection.*

Your First Program

Finally you're ready to create your first program. Type the following code into the Programming Editor window:

```
main:
    high 1
    pause 1000
    low 1
    pause 1000
    goto main
```

Be sure to include the colon after the word "main" on the first line. See Figure 5-136 for a screenshot. The indents are created by pressing the Tab key. Their only purpose is to make program listings more legible. The software ignores them.

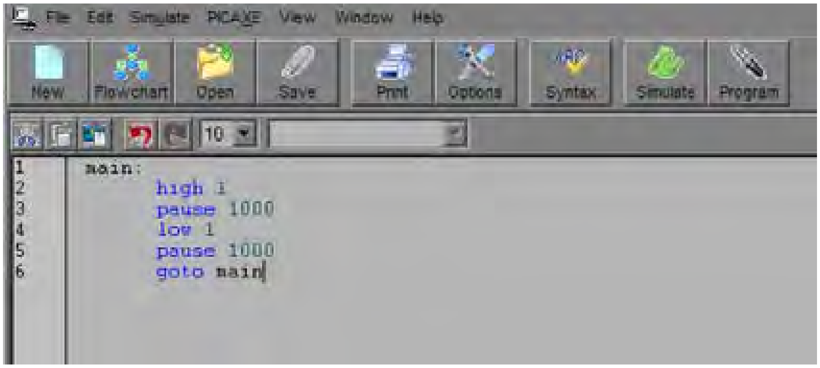
File
Edit
Simulate
PICAXE
View
Window
Help
New
Flowchart
Open
Save
Print
Options
Syntax
Simulate
Program
10

```
main:
        high 1
        pause 1000
        low 1
        pause 1000
        goto main
```

Figure 5-136. *This screenshot shows the first test program as it should be displayed by the Programming Editor (on a Windows computer).*

Click the Program button in the Programming Editor to download this program into the chip. As soon as the download is complete, the chip should start flashing the LED, lighting it for 1 second and then switching it off for 1 second. Figure 5-137 shows the steps that you should have followed to program the chip.

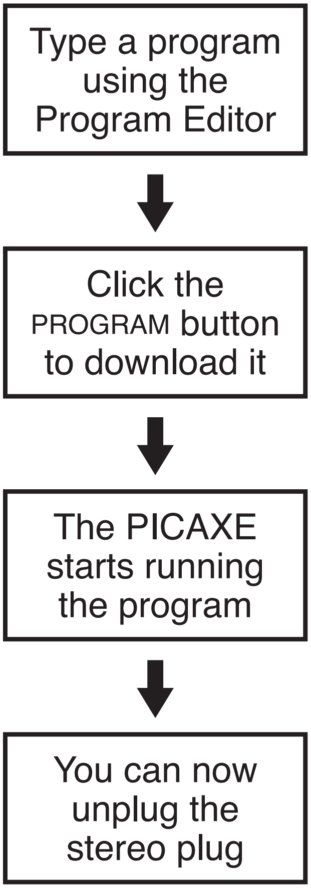
Type a program using the Program Editor
Click the PROGRAM button to download it
The PICAXE starts running the program
You can now unplug the stereo plug

Figure 5-137. *Four steps to create and run a program on the PICAXE controller chip.*

Now for the interesting part: disconnect the USB cable from the breadboard. The chip should continue flashing.

Disconnect the power supply from the breadboard and wait a minute or two for the capacitors to lose their charge. Reconnect the power, and the chip will start flashing again.

The program that you downloaded to the chip will remain in the memory inside the chip and will begin running every time power is applied to the chip.

The program doesn't use the usual pin numbers on the chip. It uses what I will call "logic pin numbers." Figure 5-138 shows how they are numbered. Figure 5-139 shows their multiple functions. I have put Logic Pin 0 in parentheses because its main purpose is to send data to the computer through the USB cable. It can do double duty as a digital output, but you have to disconnect it from the USB cable first. It's easy to forget to do this. It's a hassle that I prefer to avoid.

Decoding the Code

Let's take a look at the little program that you typed in. The first line identifies a section of the program. This program only has one section, and we're calling it "main." Any word with a colon after it is the name of a section of a program:

```
main:
```

The second line tells the chip to send a high output from Logic Pin 1:

```
high 1
```

The third line of the program tells the chip to wait for 1,000 milliseconds. This of course is the same as one second:

```
pause 1000
```

The fourth line tells the chip to change Logic Pin 1 back to its low state:

```
low 1
```

The fifth line tells the chip to wait for another 1,000 milliseconds:

```
pause 1000
```

The last line tells the chip to go back to the beginning of the "main" section:

```
goto main
```

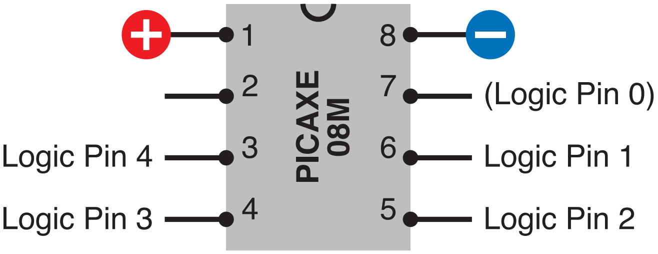
1
2
3
4
8
7
6
5
PICAXE
08M
Logic Pin 4
Logic Pin 3
(Logic Pin 0)
Logic Pin 1
Logic Pin 2

Figure 5-138. *The conventional pin numbers of the PICAXE chip are incompatible with the numbering system that is used in the PICAXE programming language. To minimize confusion, this guide refers to "Logic Pins" when using the numbering system that is required for programming the chip.*

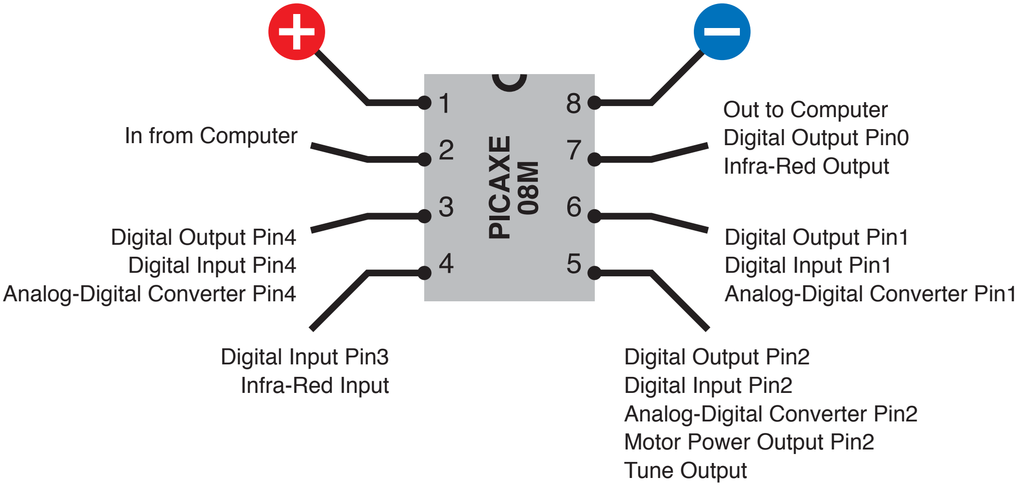
1
2
3
4
8
7
6
5
PICAXE
08M
In from Computer
Digital Output Pin4
Digital Input Pin4
Analog-Digital Converter Pin4
Digital Input Pin3
Infra-Red Input
Out to Computer
Digital Output Pin0
Infra-Red Output
Digital Output Pin1
Digital Input Pin1
Analog-Digital Converter Pin1
Digital Output Pin2
Digital Input Pin2
Analog-Digital Converter Pin2
Motor Power Output Pin2
Tune Output

Figure 5-139. *Many of the pins on the PICAXE 08M have multiple functions, which can be selected by appropriate program instructions.*

Editing

What if you want to change the program? No problem! Use the Programming Editor to change one of the lines in the program. Substitute 100 instead of 1000 milliseconds, for instance. (The `pause` command can be followed by any number up to 65535.) In your program, don't use a thousands comma in any of the numbers that you specify.

Plug the USB cable into the breadboard again, hit the Program button on the screen, and the new version of the program will be automatically downloaded to the chip, overwriting the old version.

What if you want to save the program for future use? Just go to the File menu in the Programming Editor and save the program onto your computer's hard drive. Because the PICAXE uses a variant of the BASIC computer language, it adds a *.bas* filename extension.

Simulation

If you make a simple typing error, the Programming Editor will find it and stop you from downloading your program. It will leave you to figure out how to fix the line that contains the error.

Even if all the statements in your program are correctly typed, it's a good idea to run a simulation of what they'll do, before you download them. This is easily done: click the "simulate" button on the menu bar of the Programming Editor. A new window will open, displaying a diagrammatic view of the PICAXE chip and showing you the states of its pins. (Note that if you use very short `pause` commands, the simulation won't run fast enough to display the time accurately.) A simulation screenshot is shown in Figure 5-140.

The >> button at the bottom-right corner of the simulation window will open up a list of all the variables in your program. So far, it doesn't have any variables, but it soon will. All the zeros on the righthand side are binary numbers, which you can ignore for now.

You'll need to check the second part of the PICAXE documentation, which contains all the programming statements and their correct syntax. At the time of writing, this is stored at http://www.rev-ed.co.uk/docs/picaxe_manual2.pdf.

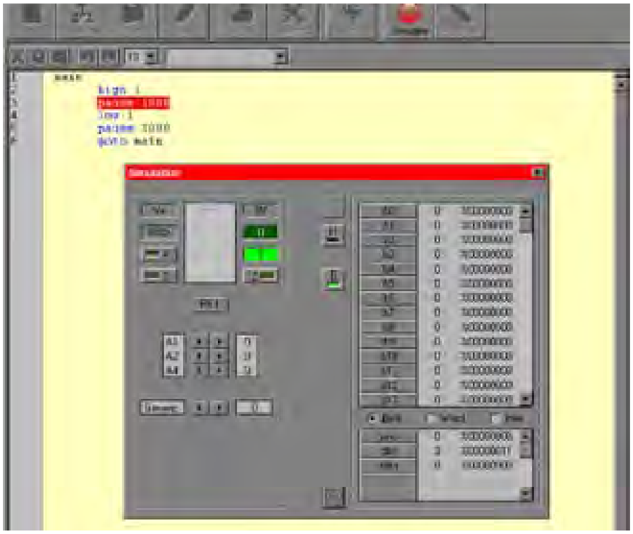

Figure 5-140. *This screenshot shows the simulation window that can be opened in the Program Editor to test program code before it is downloaded to the chip. The values of variables are shown in the section on the right. The pin states are shown on the left.*

Loops

Here's another thing I'd like you to try. Rewrite your program code as shown here and download it onto the PICAXE:

```
main:
    for b0 = 1 to 5
        high 1
        pause 200
        low 1
        pause 200
    next
    wait 2
    goto main
```

Note that `b0` is letter b followed by a zero, not letter b followed by letter O. The extra indents once again are added to make the listing easier to understand. The four lines beginning "high 1" and ending "pause 200" will be executed repeatedly. It's helpful to see them as a block.

Watch the light and see what happens. It should flash five times quickly, then wait for two seconds, and then repeat. You just added a *loop* to your program. You can use a loop if you want something to happen more than once.

b0 is known as a *variable*. Think of it as being like a little "memory box" with its name, b0, on a label on the outside. Figure 5-141 illustrates this concept. This particular memory box can contain any number from 0 through 255. The loop begins by telling the computer to put number 1 in the box, then process the remaining statements, until the word "next" sends the processor back to the first line, at which point it adds 1 to the contents of b0. If the value of b0 is 5 or less, the loop repeats. If the value is 6, the loop has run five times, so it's over, and the PICAXE skips down to the "wait 2" statement after "next." See Figure 5-142 for an annotated version of the program listing.

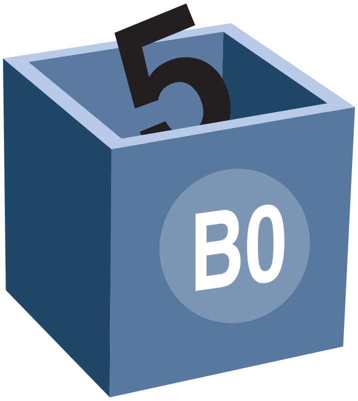

Figure 5-141. *To understand how a program works, visualize a variable as being like a "memory box" with its name on the outside and a number stored on the inside.*

"Wait" is a PICAXE command that is measured in whole seconds, so "wait 2" waits for 2 seconds. Then "goto main" begins the procedure all over again.

If your flashing-light demo worked out as planned, it's time to take the next step and make the chip do something more useful.

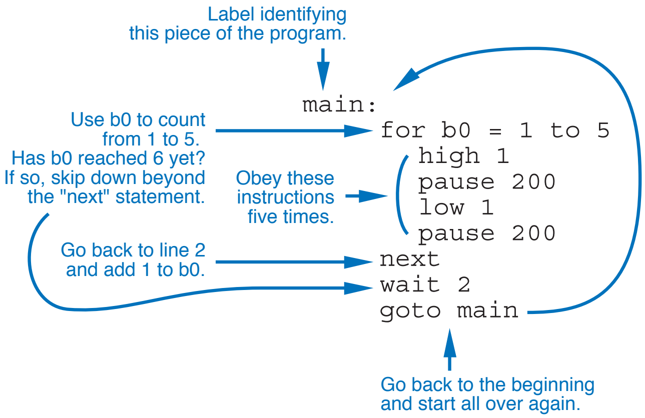
Label identifying this piece of the program.
Use b0 to count from 1 to 5. Has b0 reached 6 yet? If so, skip down beyond the "next" statement.
Obey these instructions five times.
Go back to line 2 and add 1 to b0.
Go back to the beginning and start all over again.

```
main:
for b0 = 1 to 5
  high 1
  pause 200
  low 1
  pause 200
next
wait 2
goto main
```

Figure 5-142. *The blue annotations explain what the program, on the right, is telling the PICAXE to do.*

FUNDAMENTALS

Basic PICAXE parameters

Here are some of the most useful parameters of the PICAXE:

- The PICAXE requires 5 volts DC, regulated.
- The inputs and outputs of the PICAXE are compatible with 5-volt logic chips. You can attach them directly.
- Each PICAXE pin can sink or source up to 20mA. The whole chip can deliver up to 90mA. This means that you can run LEDs directly from the pins, or a piezo noisemaker (which draws very little current), or a transistor.
- You can use a chip such as the ULN2001A Darlington array (mentioned in the previous experiment) to amplify the output from the PICAXE and drive something such as a relay or a motor.
- The chip executes each line of your program in about 0.1 milliseconds.
- The 08M chip has enough flash memory for about 80 lines of program code. Other PICAXE chips have more memory.
- The PICAXE provides 14 variables named b0 through b13. The "b" stands for "byte," as each variable occupies a single byte. Each can hold a value ranging from 0 through 255.
- No negative or fractional values are allowed in variables.
- You also have 7 double-byte variables, named w0 through w6. The "w" stands for "word." Each can hold a value ranging from 0 through 65535.
- The "b" variables share the same memory space as the "w" variables. Thus:
  - b0 and b1 use the same bytes as w0.
  - b2 and b3 use the same bytes as w1.
  - b3 and b4 use the same bytes as w2.
  - b5 and b6 use the same bytes as w3.
  - b7 and b8 use the same bytes as w4.
  - b9 and b10 use the same bytes as w5.
  - b11 and b12 use the same bytes as w6.
  - b13 and b14 use the same bytes as w7.

  Therefore, if you use w0 as a variable, do not use b0 or b1. If you use b6 as a variable, do not use w3, and so on.
- Variable values are stored in RAM, and disappear when the power is switched off.
- The program is stored in nonvolatile memory, and remains intact when the power is off.
- The manufacturer's specification claims that the nonvolatile memory is rewritable up to about 100,000 times.
- If you want to attach a switch or pushbutton to a pin and use it as an input, you should add a 10K pull-down resistor between the pin and the negative side of the power supply to hold the pin in a low state when the switch is open. Figure 5-143 shows how pull-down resistors should be used in conjunction with a SPST switch or a pushbutton.
- On the 08M chip, if you apply a varying resistance between Logic Pins 1, 2, or 4, and the negative side of the power supply, the chip can measure it and "decide" what to do. This is the "Analog-Digital Conversion" feature—which leads to our next experiment.

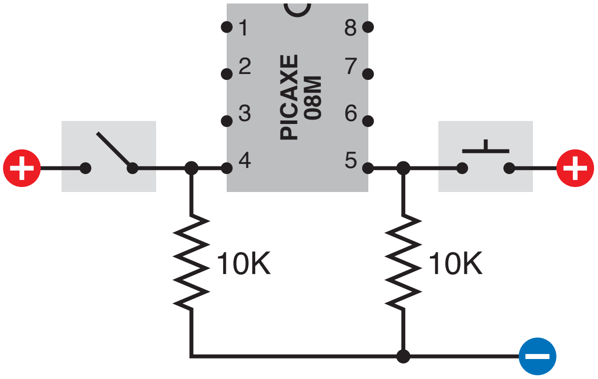

Figure 5-143. *The PICAXE can respond to the state of a switch or button attached to any of its input-capable pins. A 10K resistor must be used to pull down the state of the pin when the contact of the switch or button is open. Otherwise, you may get unpredictable results.*

Experiment 35: Checking the Real World

Often we want a microcontroller to measure something and respond in an appropriate way. For instance, it can measure a low temperature and sound an alarm, as I suggested in the example that I gave earlier.

The PICAXE has three analog-to-digital converters (ADCs) built in, accessible via logic pins 1, 2, and 4, as shown in Figure 5-139. The best way to use them is by applying a potential somewhere between 0 and 5 volts. In this experiment, I'll show you how to calibrate the response of the chip.

You will need:

- Trimmer potentiometer, 2K. Quantity: 1.
- PICAXE 08M chip and associated USB cable and socket. Quantity: 1 of each.

Procedure

Take the same trimmer potentiometer that you used in Experiment 32 and wire its center terminal to Logic Pin 2 of the PICAXE (which is hardware pin 5). The other two terminals of the 2K trimmer go to positive and to negative, respectively. So depending how you set the trimmer, the pin of the PICAXE is directly connected to positive (at one end of the scale), or directly connected to negative (at the other end of the scale), or somewhere in between. See Figure 5-144 for the revised schematic, and Figure 5-145 for a photograph of the breadboarded circuit.

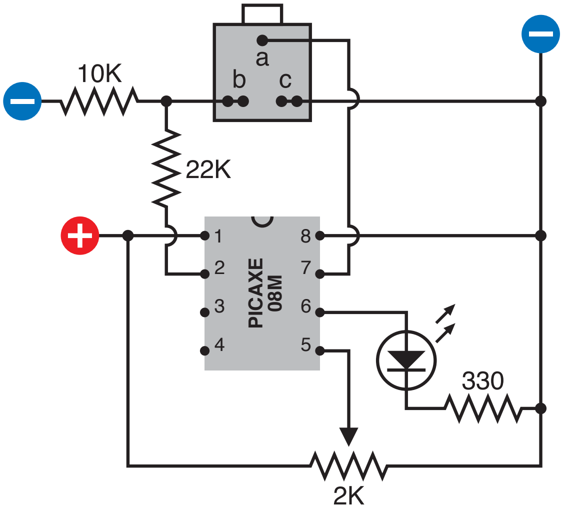

Figure 5-144. *This schematic, drawn in a layout suitable for breadboarding, shows how a 2K potentiometer can be used to apply a varying voltage to one of the pins of the PICAXE that is capable of converting an analog signal to a digital value.*

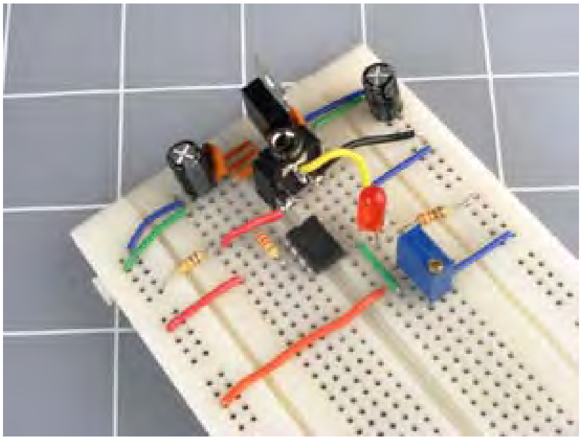

Figure 5-145. *The trimmer potentiometer added to the previously breadboarded circuit.*

Now we need a program to tell the chip what to do. Using the Programming Editor, start a new document. The code should look like this:

```
main:
    readadc 2,b0
    debug b0
    goto main
```

The command "readadc 2,b0" means "read the analog input on Logic Pin 2, convert from analog to digital, and store the result in b0."

The command "debug b0" tells the chip to go into program debugging mode, in which it uses its USB cable to tell the Programming Editor the values of all the variables while the program is running. The variables are displayed in a debugging window.

Download the program, and as the program starts to execute, the debugging window should open. Start adjusting the trimmer while looking at the value of b0, and you'll see b0 change its value.

You can make a table and draw a graph showing the relationship between the resistance between Logic Pin 2 and ground, and the value of b0. Just pull the trimmer off the breadboard, measure its resistance with a meter, then increase its resistance by, say, 200Ω, put it back into the breadboard, and look at the value of b0 again.

This is laborious, but calibrating equipment is always laborious—and in any case, I decided to do it for you. The graph is shown in Figure 5-146. You can also see the raw data numbers in the following table. I was pleased to find that the PICAXE gives a very precise, linear response to the input voltage. In other words, the graph is a straight line.

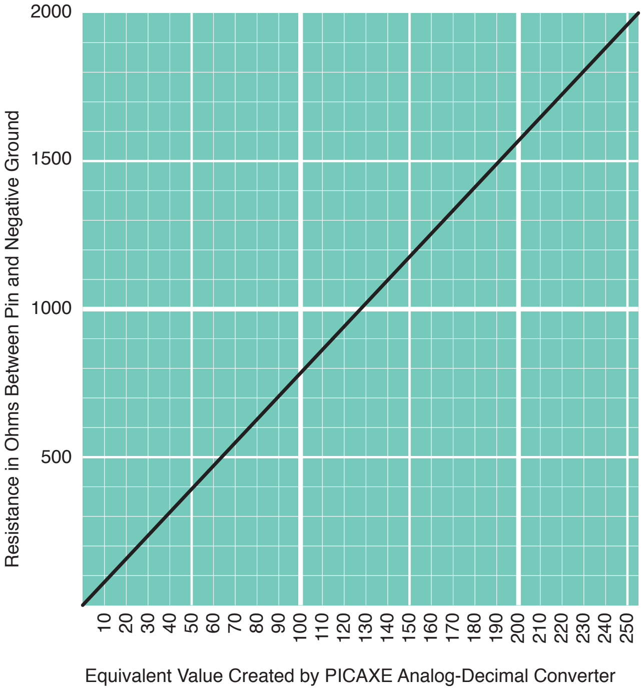

Figure 5-146. *When an ADC input pin is hooked up to a 2K potentiometer, which is connected across the same voltage that powers the chip, you should find that the resistance between the input pin and the negative side of the power supply generates the series of digital values shown on the graph. Note that the potentiometer must have a 2K value, and the power supply is assumed to be precisely 5 volts.*

This table shows measurements made with PICAXE 08M controller.

Resistance (in ohms) between the ADC pin and the negative supply	Equivalent digital value
2000	255
1900	243
1800	230
1700	218
1600	205
1500	192
1400	179
1300	166
1200	154
1100	141
1000	128
900	115
800	102
700	90
600	77
500	64
400	51
300	38
200	26
100	13
0	0

Now we can modify the program to make it do something with the information that it's taking in:

```
main:
    readadc 2,b0
    let w1 = 5 * b0
    high 1
    pause w1
    low 1
    pause w1
    goto main
```

Notice what's happening here. First we get a value in b0, and then on the next line, we do some arithmetic with it. The asterisk means "multiply." So the statement says, "Take whatever value is in b0, multiply by 5, and transfer it to another variable, w1." We have to use a w variable, because when we multiply the value of b0 by 5, we may get a number that is bigger than 255—too big to fit into a byte variable.

Finally, we take variable w1 and use it with a "pause" statement instead of a fixed number value. We're saying to the PICAXE, "pause for whatever number of microseconds you get by checking the value of w1."

So the software checks a variable resistance, turns it into a number, and applies that number to adjust the flashing speed of the LED.

Think back to the need of the cart powered by stepper motors. It was supposed to check two photoresistors, and adjust the speed of each motor accordingly. Well, this PICAXE program is a step in that direction. It can measure voltage on a pin and change the output frequency on another pin. If you had two PICAXE chips, you could wire each of them to a photoresistor and a motor. Then you could adjust the behavior of your cart by editing the second line in the program, where it converts the value of b0 to the value of w1 which will be used in the "pause" command to determine the number of pulses per second. Instead of multiplying by 5 you could multiply by 7 or whatever number gives you the result you need. This leads to an important conclusion: *a big advantage of a programmable chip is that you can make adjustments in software*.

Because the PICAXE 08M actually has more than one ADC input, and has three pins that can be used for output, you might wonder whether you could use just the one chip to control both motors in response to inputs from two sensors. The problem is that the three output pins on the 08M also function as the three ADC input pins. You'd do better to buy one of the more advanced PICAXE chips, such as the 18M, which has more pins to choose from. It uses the same basic set of programming instructions, and doesn't cost much more money.

Also, you should read the PICAXE documentation and look up the "pwmout" command, which is short for "pulse-width modulation output," but you can think of as meaning "power motor output." This is specifically intended to run stepper motors. It establishes an output frequency of pulses that will continue while the chip obeys other instructions in its program.

FUNDAMENTALS

Extra features

A complete guide to the 08M would fill a book of its own, and of course such books already exist (just search the books section of Amazon.com for keyword "picaxe"). But I'll finish my introduction to the controller by listing some of its extra capabilities, leaving you to look them up and explore them. Then I'm going to suggest one last experiment.

Interrupts

The PICAXE 08M allows you to set one "interrupt." This feature tells the chip to make a mental note that if a particular event occurs—such as a switch applying voltage to one pin—it should stop doing whatever else it was doing, and respond to the interruption.

Infrared

One pin on the PICAXE 08M can be used to receive infrared signals from a TV-style remote that you can buy from the same suppliers that sell the PICAXE itself. With an infrared sensor attached to the chip, you can issue commands remotely. If you want to build a remote-controlled robot, the chip is specifically designed with this in mind.

Servo motors

Every PICAXE chip has at least one pin that can send a stream of pulses to control a typical servo motor. On the 08M chip, it's Logic Pin 2. The width of each pulse tells the motor how far to rotate from its center position before stopping. A 555 timer can send this stream, but the PICAXE makes it easier. You can search online for more information about servo motors, which are especially useful for applications such as steering model vehicles, adjusting the flaps on model airplanes, and actuating robots.

Music

The PICAXE has an onboard tone generator that can be programmed with a "tune" command to play tunes that you write using a simple code.

Alphanumeric input/output

The "kbin" programming command is available in the PICAXE models 20X2, 28X1 and 28X2, and 40X1 and 40X2. You can plug a standard computer keyboard into the chip, and it will read the keypresses. You can also attach alphanumeric displays, but these procedures are nontrivial. For instance, when you're trying to figure out which key someone has pressed on a keyboard, your program has to contain a list of the special hexadecimal codes that the keyboard creates.

Pseudorandom number generation

All PICAXE models can generate pseudorandom numbers using a built-in algorithm. If you initialize the number generator by asking the user to press a button, and you measure the arbitrary time that this takes, you can seed the pseudorandom number generator with the result, and the pseudorandom number generator will have a less repeatable sequence.

Visit *http://www.rev-ed.co.uk/docs/picaxe_manual1.pdf* to learn more.

Experiment 36: The Lock, Revisited

The combination lock that I described in Experiment 20 is especially appropriate for a microcontroller, because it requires a series of operations that resemble a computer program. I'm going to show how this project can be redesigned using a PICAXE 08M, and then leave it to you to consider how some of the other projects in this book could be converted.

You will need:

- The same type of keypad and relay recommended in Experiment 20.
- A transistor or Darlington array to amplify the output from the PICAXE so that it can drive the relay.

Getting the User Input

Any of the input pins on the PICAXE can sense a switch closing. The trouble is that we only have three pins capable of doing this, and even the most advanced PICAXE chip has fewer than 10 such pins. So how can we attach a 10-key keypad to the 08M?

I have a suggestion: attach various resistors to the keypad, so that each key applies a different voltage to one of the ADC pins. Then use the ADC feature to convert the voltage to a number, and use a table of possible numbers to figure out which key is being pressed. This may not be the most elegant solution, but it works!

The keypad can be wired as shown in Figure 5-147. The asterisk key is still being used to supply power, as in the original experiment, while the pound key resets the relay at the end of your computing session, as before.

Current flows through a series of resistors, beginning with one that has a value of 500Ω. Because this is not a standard value, you will either have to make it by combining other resistors in series, or by presetting a trimmer potentiometer. After that, each button is separated from the next button by a 100Ω resistor. Finally, at the end of the chain, a 600Ω resistor separates the last button from the negative side of the power supply. Again, this is not a standard value, and you may have to use a trimmer.

Add up all the resistances and you have 2K, which is the range that the PICAXE wants us to use. When you press a button, you tap into the chain of resistances. Button 9 puts 600Ω between the PICAXE ADC pin and ground. Button 6 is 700Ω, button 3 is 800Ω, and so on. (You may prefer to lay out the buttons so that the resistance progresses in a more logical fashion. That's up to you. I chose to lay them out in the way that would be easiest to visualize on a keypad.)

Now look back at the ADC values that I supplied in the table on page 308. These are the values that you should get when you press various keypad buttons—but you cannot count on them being absolutely precise, because they may vary if your resistor values are not quite accurate, or if your power supply isn't exactly 5 volts. It's not safe to say, for example, that the PICAXE will deliver an ADC conversion value of precisely 77 when the resistance is 600Ω. It's safer to say that the value will be between 71 and 83. If we specify a range as shown in the following table, we have a much better chance of interpreting each button correctly.

Button number		Range
9	600	71–83
6	700	84–96
3	800	97–108
0	900	109–121
8	1000	122–134
5	1100	135–147
2	1200	148–160
7	1300	161–172
4	1400	173–185
1	1500	191–198

Suppose you attach the common pin of your keypad to ADC Logic Pin 2 of the PICAXE. You can now use the Program Editor to write a program that looks like this:

```
getkey:
    readadc 2,b0
    let b1 = 9
    if b0 < 84 then finish
    let b1 = 6
    if b0 < 97 then finish
    let b1 = 3
    if b0 < 109 then finish
    let b1 = 0
    if b0 < 122 then finish
    let b1 = 8
    if b0 < 135 then finish
    let b1 = 5
    if b0 < 148 then finish
    let b1 = 2
    if b0 < 161 then finish
    let b1 = 7
    if b0 < 173 then finish
    let b1 = 4
    if b0 < 186 then finish
    let b1 = 2

finish:
    return
```

What does the word "return" mean at the end? I'll get to that in a second. I want to explain the rest of the routine first.

b0 receives the value supplied by the analog-digital converter when it looks at the keypad. After storing the number in b0, the routine has to figure out which keypad key it matches. The key identity (0 through 9) will be stored in another variable, b1.

The program starts by assigning value 9 to b1. Then it checks to see whether b0 < 84. This means "if b0 is less than 84." If it is, then the routine tells the PICAXE to "finish," which means "jump to the finish: label." But if b0 is *not* less than 84, by default the PICAXE continues on to the next line, which makes a second attempt at guessing which key has been pressed. It assigns number 6 to b1. Now there's another if-then test—and so on. This process of reassigning values to b1 stops only when it gets to the point where b0 is greater than a number in the table.

If you're familiar with other dialects of BASIC, this may seem a bit laborious to you. You may wonder why we can't use a statement such as this:

```
    if b0 > 70 and b0 < 84 then b1 = 9
```

The answer is that PICAXE BASIC isn't sufficiently sophisticated to allow this. An if-then statement has to result in a jump to another section of the program. That's the only permitted outcome.

If you don't have any prior programming experience, the routine may still seem laborious to you, and perhaps a bit puzzling, too. This is understandable, because you're getting a crash course in software design without any formal preparation. Still, the PICAXE Programming Editor can be a big help, because it has its simulation feature. Before you can use this, though, you have to precede the routine that I just supplied with a control routine that you must type above it. The screenshot in Figure 5-148 shows you how it should look.

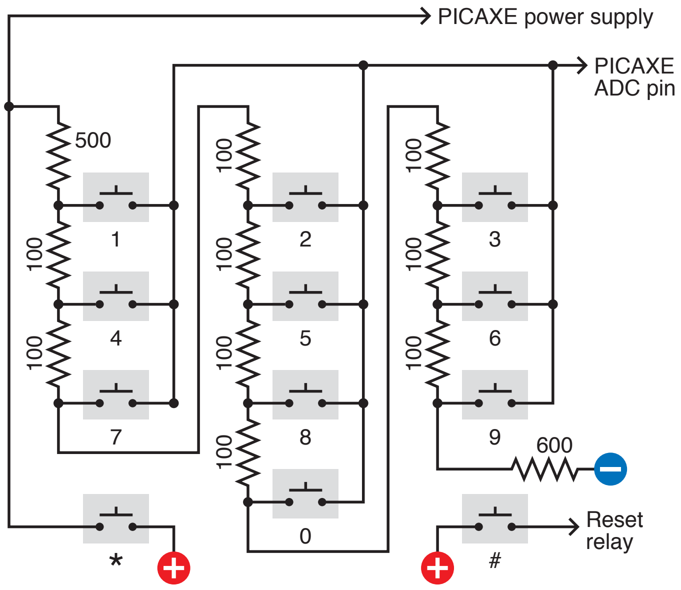

Figure 5-147. *A quick and simple way of attaching a keypad to provide numeric input to the PICAXE uses a chain of resistors totalling 2,000Ω. When a button is pressed, it connects the ADC input pin to a point in the chain. The resistance detected by the input pin can then be converted by the program in the chip to determine which key has been pressed.*

I have chosen an arbitrary combination of 7-4-1 for our combination lock. Using these numbers, the main section of the program looks like this:

```
main:
    low 1
    gosub getkey
    if b1<>7 then main
    gosub getkey
    if b1<>4 then main
    gosub getkey
    if b1<>1 then main
    high 1
    end
```

I should explain that the <> pair of symbols mean "is not equal to." So the fourth line of the program means, "if b1 is not equal to 7."

The value of b1 is supposed to be 7 if the user is putting in the correct combination. So if it's not 7, the user has entered the wrong value, and the if-then statement sends the PICAXE back to the beginning. In fact anytime the user inputs a number that is not in the correct 7-4-1 sequence, the program sends the PICAXE back to the beginning. This is the way the pure-hardware version of this experiment was set up.

But what is this word "gosub"? It means "go to a subroutine." A subroutine is any sequence of program statements that ends with the instruction to "return." So "gosub getkey" tells the PICAXE to mark its current place in the program while it skips to the `getkey:` section of code, which it obeys, until it finds the word "return," which returns it to the place from where it came.

The PICAXE continues in this fashion until it reaches the word "end." I had to insert the word "end" because otherwise the PICAXE will continue executing the program and will fall into the subroutine. "End" stops it from doing so. Figure 5-148 shows a screenshot of the complete listing.

So—is that all? Yes, that's it. If you enter the code into the Programming Editor exactly as I have supplied it, you should be able to run it in simulation mode, and in the simulation window, click the right-arrow beside Logical Pin A2 to increase its value in steps. Each time you pass one of the values in the `getkey:` subroutine, you should see the value for variable b1 change in the display.

This is really all you need to perform the functions of the combination lock. When the PICAXE runs this program, it waits for the correct combination. If it receives the combination, it sends the output from logical pin 1 high; otherwise, logical pin 1 stays low.

The only additional item you need is a transistor or CMOS gate between logical pin 1 and the relay that unlocks the computer, because the PICAXE cannot deliver enough current to operate the relay by itself.

Putting this procedure into a controller chip not only simplifies the circuit, but offers another advantage: you can change the combination simply by rewriting the program and downloading the new version into the chip.

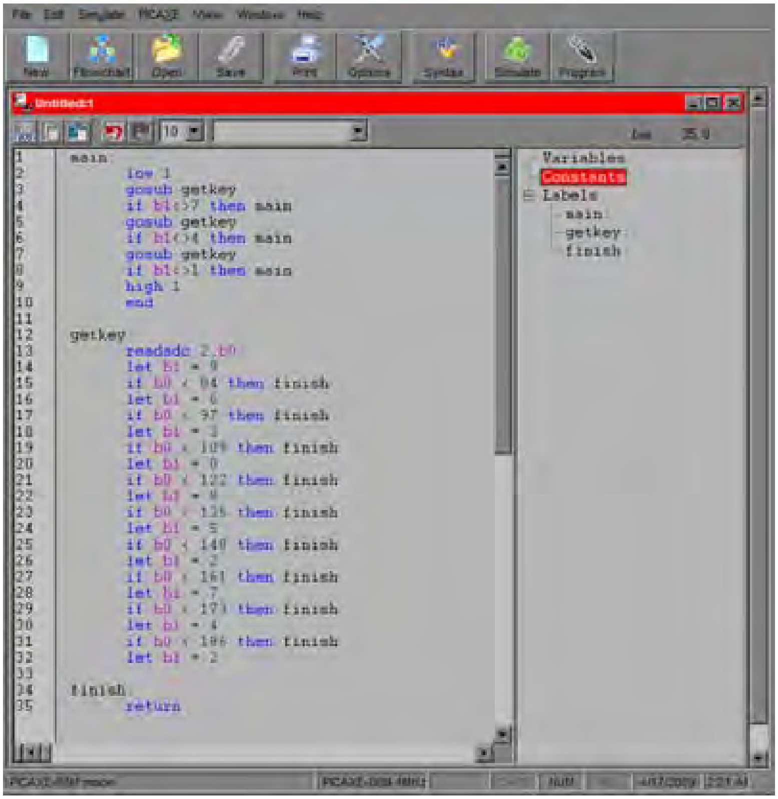

```
main:
        low 1
        gosub getkey
        if b1<>7 then main
        gosub getkey
        if b1<>4 then main
        gosub getkey
        if b1<>1 then main
        high 1
        end

getkey:
        readadc 2,b0
        let b1 = 9
        if b0 < 84 then finish
        let b1 = 6
        if b0 < 97 then finish
        let b1 = 3
        if b0 < 109 then finish
        let b1 = 0
        if b0 < 122 then finish
        let b1 = 8
        if b0 < 135 then finish
        let b1 = 5
        if b0 < 148 then finish
        let b1 = 2
        if b0 < 161 then finish
        let b1 = 7
        if b0 < 173 then finish
        let b1 = 4
        if b0 < 186 then finish
        let b1 = 2

finish:
        return
```

Figure 5-148. *This screenshot shows the complete listing of a program to read a sequence of three keypresses in conjunction with a combination lock. If the sequence is correct, the PICAXE sends a high output from one of its pins. If the sequence is incorrect, the program loops back to the beginning.*

FUNDAMENTALS

Limitations of MCUs

The PICAXE does have some disadvantages. Its voltage requirements alone restrict you from using it with the kind of freedom of a 555 timer.

Also, although I can get an instant result by plugging a 555 timer into a breadboard and adding a couple of resistors and a couple of capacitors, the PICAXE requires me to add a download socket, hook it up to my computer, write a program in the Programming Editor, and download the program.

Some people don't like writing software, or they have difficulty thinking in the relentlessly left-brain way that computer programming requires. They may prefer the hands-on process of assembling hardware.

Other people may have the opposite preference. This of course is a matter of taste, but one thing we know beyond all doubt is that computer programs often contain errors that may not reveal themselves until weeks or months later. The PICAXE, for instance, doesn't protect you if a number is assigned to a variable that exceeds the limit for that type of variable. Suppose b1=200 and b2=60 and your program tells the PICAXE:

```
let b3 = b1 + b2
```

The result should be 260, but byte-size variables can only count up to 255. What happens? You will find that b3 acquires a value of 4, without any warning or explanation. This is known as an "overflow error," which can be very difficult to predict, because it happens at runtime, when external factors are in control. The code looks perfectly good; the Programming Editor doesn't find any syntax errors; the simulation behaves properly. But in the real world, days or even months later, an unexpected set of circumstances results in an input that causes the overflow, and because the code is residing inside the chip at this point, you may have a hard time figuring out what on earth went wrong.

Software has its problems. Hardware has its advantages.

FUNDAMENTALS

Unexplored territory

If you've taken the time to complete most of the projects in this book with your own hands, you have gained a very rapid introduction to the most fundamental areas of electronics.

What have you missed along the way? Here are some topics that remain wide open for you to explore. Naturally you should search online if they interest you.

The informal, learning-by-discovery approach that I have used in this book tends to be light on theory. I've avoided most of the math that you'd be expected to learn in a more rigorous course on the subject. If you have mathematical aptitude, you can use it to gain a much deeper insight into the way in which circuits work.

I didn't deal much with computer architecture, either. We didn't go very far into binary code, and you didn't build a half-adder, which is a great way to learn how computers function on the most fundamental level. Perhaps you should think about assembling one.

I avoided going deeply into the fascinating and mysterious properties of alternating current. Here again, some math is involved, but just the behavior of current at high frequencies is an interesting topic in itself.

For reasons already stated, I avoided surface-mount components—but you can still go into this area yourself for a relatively small investment, if you like the idea of creating fascinatingly tiny devices. This may be the future of hobby electronics, so if you stick with it, you'll probably end up in the world of surface-mount.

Vacuum tubes were not mentioned, because at this point, they are mainly of historical interest. But there's something very special and beautiful about tubes, especially if you can enclose them in fancy cabinetwork. In the hands of a skilled craftsperson, tube amplifiers and radios become art objects.

I didn't show you how to etch your own printed circuit boards. This is a task that appeals to only certain people, and the preparation for it requires you to make very neat drawings or use computer software for that purpose. If you happen to have those resources, you might want to do your own etching. It could be a first step toward mass-producing your own devices.

I didn't cover static electricity at all. High-voltage sparks don't have any practical applications, and they entail some safety issues—but they are stunningly impressive, and you can easily obtain the necessary information to build the equipment. Maybe you should try.

Other Controllers

If you want something more powerful, a BASIC Stamp is the logical next step after a PICAXE. the BASIC Stamp is so called because it originally looked like a postage stamp. The BASIC Stamp has a larger vocabulary of commands and a bigger range of add-on devices (including displays with graphical capability, and a little keyboard that is specifically designed for use with the controller). The BASIC Stamp is shown in Figure 5-149.

On the downside, you'll find that everything associated with the BASIC Stamp is a bit more expensive than in the PICAXE world, and the download procedure isn't quite as simple.

One of the more recent developments in the world of MCUs is the Arduino, which is both sophisticated and powerful. It does require programming in the C language. This language is a little more difficult to understand, and has only the vaguest similarity to the syntax that is used in the PICAXE and BASIC Stamp. On the other hand, because C dominates the larger world of computing, learning it might not be such a bad idea—and the Arduino offers some truly amazing capabilities. Because it is so popular, there are also many software tools, documentation, user forums, and many enthusatic hobbyists to help you. Two other Make: Books titles that I mentioned previously, *Getting Started with Arduino* and *Making Things Talk*, provide a great introduction.

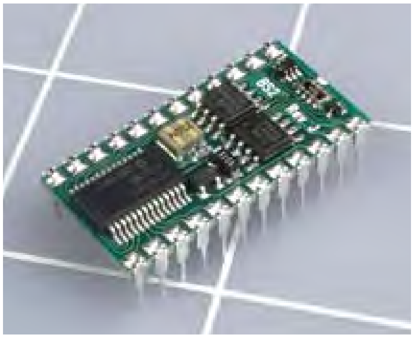

Figure 5-149. *The BASIC Stamp controller consists of surface-mounted components on a platform that has pins spaced at 1/10-inch intervals, for insertion in a breadboard or perforated board. This component uses a version of BASIC that is similar to the programming language of the PICAXE, but has many more extensions. The BASIC Stamp is available for use with a wide range of peripheral devices, including many alphanumeric dot-matrix displays.*

In Closing

I believe that the purpose of an introductory book is to give you a taste of a wide range of possibilities, leaving you to decide for yourself what you want to explore next. Electronics is ideal for those of us who like to do things ourselves, because almost any application—from robotics, to radio-controlled aircraft, to telecommunications, to computing hardware—allows opportunities that we can explore at home, with limited resources.

As you delve deeper into the areas of electronics that interest you most, I trust you'll have a satisfying learning experience. But most of all, I hope you have lots of fun along the way.

Online Retail Sources and Manufacturers

A

This appendix contains URLs for companies mentioned as retail sources or manufacturers, along with the commonly used name of the source and the company name.

Colloquially used name	Actual corporate or company name	URL
3M	Minnesota Mining and Manufacturing Co.	*http://solutions.3m.com/en_US/*
Ace Hardware	Ace Hardware Corporation	*http://www.acehardware.com*
Advanced Micro Circuits	Advanced Micro Circuits Corp	*http://www.advancedmicrocircuits.com*
Akro-Mils	Myers Industries, Inc.	*http://www.akro-mils.com*
Alcoswitch	Division of Tyco Electronics Corporation	*http://www.tycoelectronics.com/catalog/menu/en/18025*
All Electronics	All Electronics Corporation	*http://www.allelectronics.com*
All Spectrum Electronics	All Spectrum Electronics	*http://www.allspectrum.com*
All-Battery.com	Tenergy Corporation	*http://www.all-battery.com*
Alpha potentiometers	Alpha Products Inc.	*http://www.alphapotentiometers.net*
ALPS pushbutton	ALPS Electric Co., Ltd.	*http://www.alps.com*
Amazon	Amazon.com, Inc.	*http://www.amazon.com*
Amprobe	Amprobe Test Tools	*http://www.amprobe.com*
Arduino	No corporate identity	*http://www.arduino.cc*
ArtCity	ArtCity	*http://www.artcity.com*
AutoZone	AutoZone, Inc.	*http://www.autozone.com*
Avago	Avago Technologies	*http://www.avagotech.com*
BASIC Stamp	Brand owned by Parallax, Inc.	*http://www.parallax.com*
BI Technologies	BI Technologies Corporation	*http://www.bitechnologies.com*
BK Precision	B&K Precision Corp.	*http://www.bkprecision.com*
Bussmann fuses	Cooper Bussman, Inc.	*http://www.cooperbussmann.com*
C&K switch	CoActive Technologies, Inc.	*http://www.ck-components.com*
Chicago lighting	CML Innovative Technologies	*http://www.cml-it.com*
CraftAmerica	Cardinal Enterprises	*http://www.craftamerica.com*
Darice	Darice Inc.	*http://www.darice.com*

Colloquially used name	**Actual corporate or company name**	**URL**
DeWalt	DeWalt Industrial Tool Company	*http://www.dewalt.com*
Digi-Key	Digi-Key Corporation	*http://www.digikey.com*
Directed switches	Directed Electronics Inc.	*http://www.directed.com*
Doctronics	Doctronics Educational Publishing	*http://www.doctronics.co.uk*
eBay	eBay Inc.	*http://www.ebay.com*
Elenco	Elenco Electronics Inc.	*http://www.elenco.com*
Everlight	Everlight Electronic Co. Ltd.	*http://www.everlight.com*
Extech	Extech Instruments Corporation	*http://www.extech.com*
Fairchild	Fairchild Semiconductor Incorporated	*http://www.fairchildsemi.com*
FTM	FTM Incorporated	*http://thefabricatorssource.com*
Fujitsu	Fujitsu America, Inc.	*http://www.fujitsu.com/us/*
GB wire strippers	Gardner Bender Inc.	*http://www.gardnerbender.com*
Hobbylinc	Hobbylinc Hobbies	*http://www.hobbylinc.com*
Home Depot	Homer TLC, Inc.	*http://www.homedepot.com*
Ideal wire strippers	Ideal Industries Inc.	*http://www.idealindustries.com*
Jameco	Jameco Electronics	*http://www.jameco.com*
K&J Magnetics	K&J Magnetics Inc.	*http://www.kjmagnetics.com*
Kingbright	Kingbright Corporation	*http://www.kingbrightusa.com*
Kobiconn	No web page found; use mouser.com	
KVM Tools	KVM Tools Inc.	*http://www.kvmtools.com*
Lowe's hardware	LF, LLC	*http://www.lowes.com*
Lumex	Lumex Inc.	*http://www.lumex.com*
McMaster-Carr	McMaster-Carr Supply Company	*http://www.mcmaster.com*
Megahobby	Megahobby.com	*http://www.megahobby.com*
Meter Superstore	Division of SRS Market Solutions Inc.	*http://www.metersuperstore.com*
Michaels craft stores	Michaels Stores, Inc.	*http://www.michaelscrafts.com*
Mill-Max	Mill-Max Manufacturing Corp.	*http://www.mill-max.com*
Mitutoyo	Mitutoyo America Corporation	*http://www.mitutoyo.com*
Motorola	Motorola, Inc.	*http://www.motorola.com/us*
Mouser electronics	Mouser Electronics, Inc.	*http://www.mouser.com*
Mueller alligator clip	Mueller Electric Company	*http://www.muellerelectric.com*
Newark	Subsidiary of Premier Farnell plc	*http://www.newark.com*
NKK switches	Nihon Kaiheiki Industry Co. Ltd.	*http://www.nkkswitches.com*
NXP semiconductors	NXP Semiconductors	*http://www.nxp.com*
Omron	Omron Corporation	*http://www.omron.com*
On Semiconductor	Semiconductor Components Industries, LLC	*http://www.onsemi.com*
Optek	Subsidiary of TT Electronics plc	*http://www.optekinc.com*
Panasonic	Panasonic Electric Works Corporation	*http://pewa.panasonic.com*

Colloquially used name	Actual corporate or company name	URL
PanaVise	Panavise Products, Inc.	*http://www.panavise.com*
Parallax	Parallax, Inc.	*http://www.parallax.com*
Pep Boys	Pep Boys-Manny, Moe and Jack	*http://www.pepboys.com*
Philips	Koninklijke Philips Electronics N.V.	*http://www.usa.philips.com*
PICAXE	Revolution Education Ltd.	*http://www.rev-ed.co.uk*
Piedmont Plastics	Piedmont Plastics, Inc.	*http://www.piedmontplastics.com*
Plano storages boxes	Plano Molding Company	*http://www.planomolding.com*
Pomona test equipment	Pomona Electronics Inc.	*http://www.pomonaelectronics.com*
RadioShack	RadioShack Corporation	*http://www.radioshack.com*
Sears	Sears Brands, LLC	*http://www.sears.com*
SparkFun Electronics	Sparkfun Electronics	*http://www.sparkfun.com*
Stanley tools	The Stanley Works	*http://www.stanleytools.com*
STMicroelectronics	STMicroelectronics Group	*http://www.st.com*
Texas Instruments	Texas Instruments Incorporated	*http://www.ti.com*
Tower Hobbies	Tower Hobbies	*http://www.towerhobbies.com*
Twin Industries	Twin Industries	*http://www.twinind.com*
Tyco	Tyco Electronics Corporation	*http://www.tycoelectronics.com*
Vaughan	Vaughan & Bushnell Mfg.	*http://hammernet.com/vaughan/*
Velleman keyboards	Velleman nv	*http://www.velleman.eu*
Vishay	Vishay Intertechnology Inc.	*http://www.vishay.com*
Wal-Mart	Wal-Mart Stores, Inc.	*http://www.walmart.com*
Weller	Division of Cooper Industries, LLC	*http://www.cooperhandtools.com/brands/weller/*
X-Acto	Division of Elmer's Products, Inc.	*http://www.xacto.com*
Xcelite	Division of Cooper Industries, LLC	*http://www.cooperhandtools.com/brands/xcelite/*
Xytronic	Xytronic Industries Ltd.	*http://www.xytronic-usa.com*

Secrets of Building Electrostatic LIGHTNING BOLT GENERATORS

including high voltage
test equipment,
experiments, motors
and more!

by Walt Noon

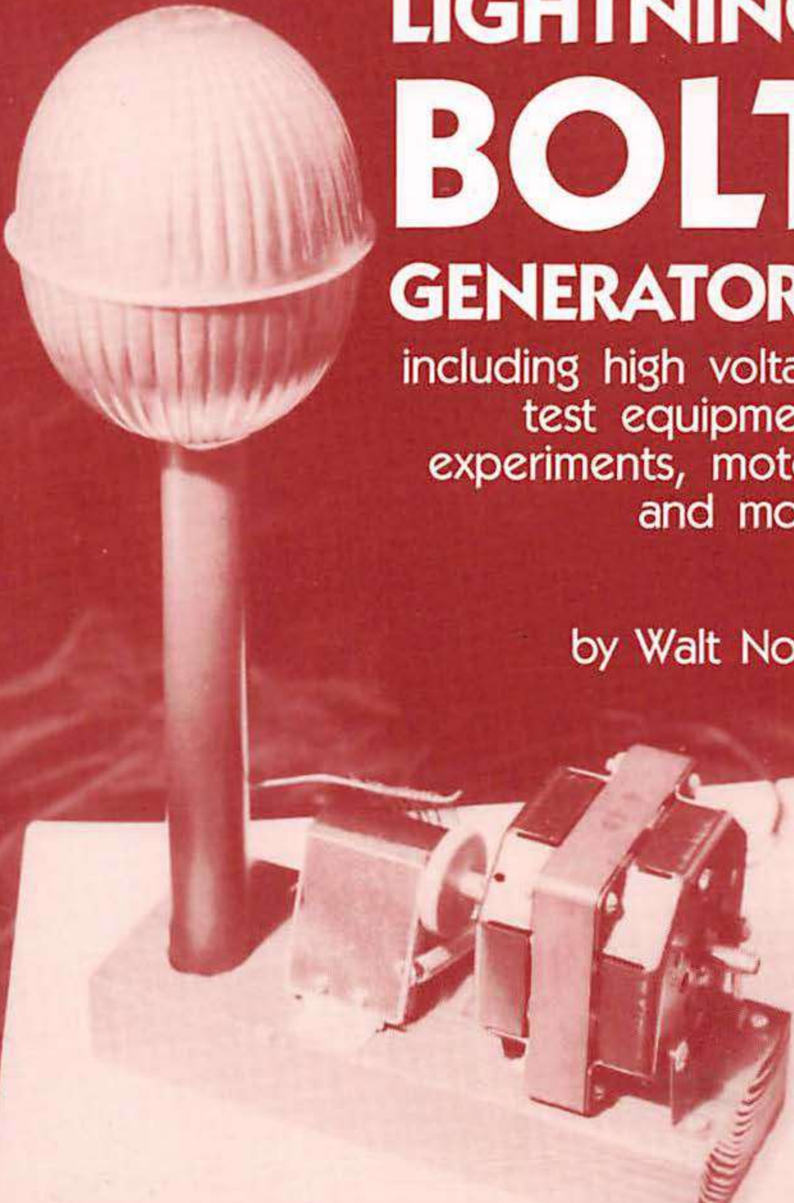

Lindsay Publications Inc

Secrets of Building
Electrostatic

LIGHTNING
BOLT
GENERATORS

including
high voltage test equipment, experiments, motors and more!

by Walt Noon

Lindsay Publications Inc

Copyright ©1992 by Lindsay Publications Inc

All rights reserved.

Lindsay Publications Inc
PO Box 12
Bradley IL 60915-0012

ISBN 1-55918-090-0

2 3 4 5 6 7 8 9 0

To my Mom and Dad who
always took the time to answer
my questions

Original CAD drawings
and photographs by
Matthew J. Roddy

Cartoon drawings by
Mike Noon

Special thanks to
Kevin E. Curtin
Matthew J. Roddy
Mike Noon

WARNING

The author of this book is not a professional engineer nor has he had formal training in the design or operation of high voltage devices. The author is an amateur but has been successful in building and operating the devices discussed herein.

The methods that he used and describes are presented merely as guidelines for other amateurs in developing similar devices. These devices can be dangerous, possibly even lethal, and dangers have been pointed out wherever possible. Since the author is not a professional in this field, there may well be other dangers involved in the building and operation of the devices described.

The author hereby disclaims any liability for injury to persons or property that may result while using this project book, and the author does not intend by this publication to explain all dangers known or unknown that may exist in the building and operation of this project.

Lindsay Publications Inc has not built this project nor does it endorse the methods described. Lindsay Publications Inc assumes no liability for injury to person or property that may result form used of this information.

Contents

Electrophorus 4
Rotostatic Generator 9
"Cat-O-Static" Generator 17
Motor Control 18
External Van de Graaff Generator 23
Classic Van de Graaff Generator..............................31
Extremely High Power Van de Graaff Generator 39
Inductive Electrostatic Generator 42
High Voltage Measurements 46
FET Electroscope 47
Mechanical Electroscope 50
Neon Lamp Blank 51
Spark Gap 53
Leyden Jar...58
Water Capacitor...61
Plate Capacitor..62
Experiments 67
Ion Blower 76
Franklin Motor...78
Corona Motor..80
Atmospheric Electricity 81
Bibliography..86
Electroplating Terminals 87

Every text I've seen on the subject of electrostatics seems to begin with a history of the subject. A. D. Moore, one writer whose works I particularly enjoy, said that if cavemen had cats that rubbed up against them then they must certainly have been aware of the presence of static electricity. There are theories now going around the scientific community that claim lightning triggered the formation of the first chains of DNA. I'll take that one step further and guess that lightning began the evolutionary process that created our species.

Static electricity is one of the most fascinating and possibly most overlooked fields of modern electronics. We all take static for granted. Before I began experimenting with static electricity it would have been very difficult to convince me that it was easy to get 3500 volts from an old plastic bottle or 10,000 volts and more from a rubber band! Even now as I watch one of these tiny generators in action it is difficult to imagine that it is producing these voltages.

As you learn about electrostatics you will find many strange and wonderful experiments to perform. For example, it is possible to mysteriously levitate objects, blow a candle flame to one side with an invisible wind, and to light neon bulbs without wires.

There are motors that can be built to run on static electricity! Well-made electrostatic motors are so sensitive that they have been powered by the earth's natural electrical charge alone! Connect your motor to an antenna, and off it goes!

In addition to the many curiosities static electricity makes possible, there is also the potential for serious research with many astounding possibilities.

Consider the work of Professor John G. Trump, formerly of the Massachusetts Institute of Technology. Professor Trump points out that electrostatic forces are the most powerful forces known in nature – stronger by far than gravity or magnetism. Professor Trump demonstrates the practical application of electrostatic generators with the following illustration:

Consider two conductive plates, 100 square inches in area, facing each other and separated by an insulator. If the plates are each charged oppositely to 300 volts per centimeter the force of attraction between them will be one 2,000th of a pound. Increase this charge to 30,000 volts per centimeter and the attraction escalates to half a pound. Now increase this voltage to the order of three million volts per centimeter and the force of attraction jumps to 5,700 pounds! As Trump has said "Force of this order has more than a passing interest for power engineers." Trump and his colleagues had been pursuing the creation of electrostatic generators they hoped would one day rival modern electromagnetic generators.

Just how much electrostatic energy is theoretically possible? Consider an aluminum cube only one centimeter square. Imagine that all the electrons in that cube were removed and kept one meter away. The cube has only positive charges, and the electrons one meter away only negative charges. Since positive and negative charges attract, what would be the force of attraction generated between them? The answer turns out to be an astounding thirty-two million million million million pounds! Professor A.D. Moore, an expert in electrostatics, points out that this force would be equal to the weight of a steel cube 76 miles high!

Of course, as much fun and as inspiring as it is to ponder the theoretical world of electrostatics, back in the real world we find that electrons don't leave home so easily. The most useful electrostatic inventions to date make use of the 'little' forces of electrostatics. But keep that cube in mind and let everyone wonder why you have that curious smirk on your face whenever you run your generator.

What voltages are possible for the experimenter? The four easily assembled generators shown in the following text will give you an idea. Each one intentionally operates on a slightly different principal of gathering an electrostatic charge. If you build them, or just simply become familiar with them, you will develop a understanding of the most popular ways of generating static electricity.

I am confident that if you become acquainted with these machines you will have no difficulty at all in constructing many others of your own design. The reason that I am so confident is that I designed and constructed all of these machines to teach myself about electrostatics! Each machine represents the best of my thoughts at the time I constructed it.

You will soon find that it is easily possible to construct a static generator capable of producing 50 to 100 thousand volts! If you are a craftsman, 200 thousand volts and more can be obtained by optimizing the designs of the higher power generators such as the Van De Graaff shown later in the text.

Electrostatic generators are also remarkably inexpensive. With a little foraging through the surplus bins, no generator in this text cost me more than $30 to build! I know of no other way to approach these voltages for such a low cost.

A Little History of the First Practical Generators

The first electrostatic generators produced their voltage with friction. Towards the end of the 17th century Otto von Guericke, an amateur physicist from Magdeburg, developed the first practical electrical machine – an electrostatic generator. What follows are his own instructions for building one:

"Secure one of the glass globes which are called phials, about the size of a youngster's head; fill it with sulfur, ground in a mortar and melted by the application of flame. After it freezes, break the phial, take out the sulfur globe and keep it in a dry place, not a moist one. Perforate it with a hole so that it can spin upon an iron axle. Thus the globe is prepared.

"To demonstrate the power developed by this globe, place it with its axis on two supports in the machine – a hand's breadth above the baseboard – and spread under it various sorts of fragments such as bits of leaves, gold dust, silver filings, snips of paper, hairs, shavings, etc. Apply a dry hand to the globe so that it is stroked or grazed two or three times or more. Now it attracts the fragments and, as it turns on its axis, carries them around with it.

"When a feather is in contact with the globe, and afterwards in the air, it puffs itself out and displays a sort of vivacity ... and if someone places a lighted candle on the table and brings the feather to within a hand's breadth of the flame, the feather regularly darts back suddenly to the globe and, as it were, seeks sanctuary there."

Otto Von Guericke goes on to describe experiments by which he produced both light and sound from his amazing globe!

The instructions for von Guericke's globe are simple enough for any experimenter today to follow. The only note that I would add is that in casting my own globe I found that large quantities of sulfur would become much smaller pools when melted, and that if the globe were not filled on the first pouring it would not appear nearly as attractive when finished. So, remember to begin with a large amount of sulfur.

THE ELECTROPHORUS

Many generations passed and many ingenious machines were developed without great change using materials and processes similar to von Guericke's globe. The most important historically was Alessandro Volta's *electro perpetuo* known now as the electrophorus. With it we can dramatically demonstrate that electrostatic charges are much like magnetic forces in that like charges repel one another while opposite charges attract.

An electrophorus is very easily assembled. It consists simply of a rectangular block of insulating material such Lucite or, better yet, polyethylene and a metal disk which is attached to an insulating handle (Figure 1).

When the Lucite is rubbed with a woolen cloth, electrons are removed from the cloth and held on the Lucite. This happens because the cloth will hold its electrons less tightly than the Lucite. An insulating material, in this case Lucite, by its nature, will not allow the electrons to move around freely. The charges are therefore 'static' or held in place by the insulating material. The Lucite will now have a negative charge, and the wool will pick up a positive charge, respectively.

Next, the metal disk with the handle is set on top of the Lucite. The negative charges on the Lucite are attracted to positive charges on the metal disk, so, the positive charges in the metal disk move to the side of the metal plate facing the Lucite. The negative charges are similarly repelled to the back of the metal disk. Now, your hand or, better yet, a ground wire are touched briefly to the top (handle side) of the metal disk and then removed. In this way the negative charge is transferred to your hand (or ground) leaving only a positive charge. The metal plate will now have a positive charge, and the plastic negative.

If you have achieved a large enough charge, the Lucite can be picked up by the metal disk!

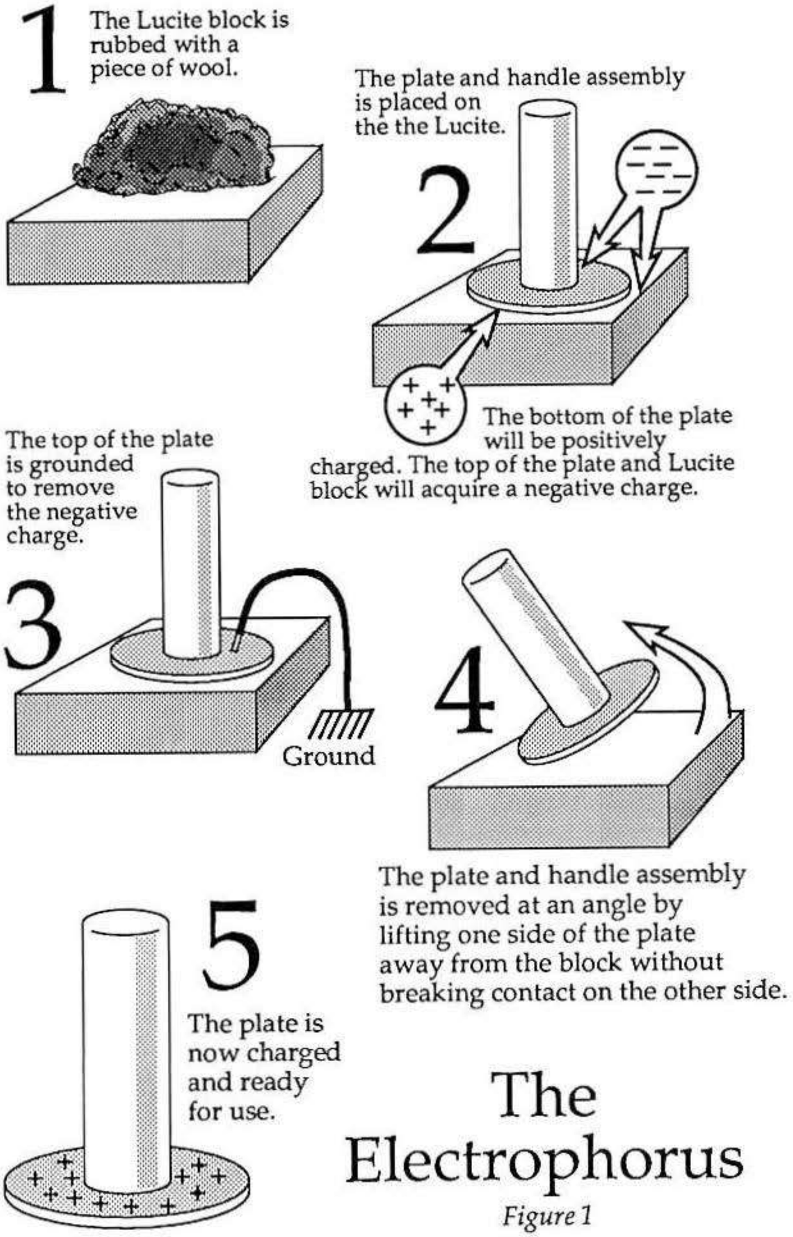

The Lucite block is rubbed with a piece of wool.

The plate and handle assembly is placed on the the Lucite.

The bottom of the plate will be positively charged. The top of the plate and Lucite block will acquire a negative charge.

The top of the plate is grounded to remove the negative charge.

The plate and handle assembly is removed at an angle by lifting one side of the plate away from the block without breaking contact on the other side.

The plate is now charged and ready for use.

The Electrophorus

Figure 1

Slowly remove the metal plate by the handle, drawing it away from the Lucite at an angle. The metal plate now holds the charge you seek. Having charged your electrophorus, you can perform the experiments with Von Guericke's globe mentioned above including attracting hair, styrofoam and other small objects.

Because the charged surface of the electrophorus is flat, a startling demonstration can be performed wherein the electrophorus disk is actually floated above the Lucite by electrostatic forces. Performing this trick takes some practice. Make a second metal plate and handle. Remember that if the plate is to float above the other it must be very light weight. (Figure 2)

Now, we have the challenge of charging both the disk and the Lucite with a like charge so that they will repel each other. Charge your first disk with a positive charge as described above, and then bring it close to, but not touching, the second disk. This will charge the second disk by attracting a negative charge to the front of the disk. Next, ground the handle side of the second disk to bleed off the positive charge. The second disk is now negatively charged. The Lucite is also negatively charged, so, provided everything has been accomplished and the weight is not too great, the second disk will now float above the Lucite block!

Electrophoruses act a great deal like capacitors discussed later on and can be used for many surprising experiments including charging a Leyden jar and flickering a neon bulb in a darkened room!

Special Notes on Friction Generators

Before we move on to the more powerful friction generators, there are a number of things you should be aware of that will help you in both troubleshooting and designing generators.

My first bit of advice is to avoid humidity. It makes air more conductive. Moisture in the air can have a devastating effect on the charges stored by electrostatic devices.

I was given a first rate example of this shortly after building my first two generators. Both machines, only days old in their construction, were whirling away in my shop. I was doing "careful scientific study" – in other words, I was zapping everything in sight to see what would happen. Suddenly, and most mysteriously, both generators

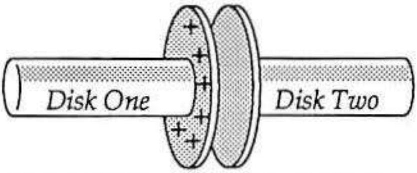
Disk One
Disk Two

Charge disk one on the Lucite block.
Then bring it close to a second disk.

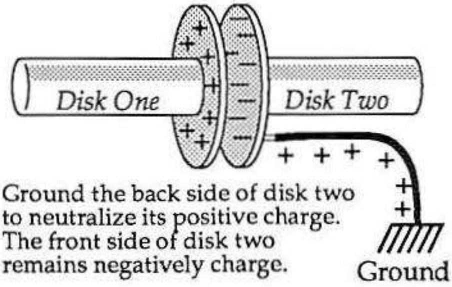
Disk One
Disk Two
Ground

Ground the back side of disk two
to neutralize its positive charge.
The front side of disk two
remains negatively charge.

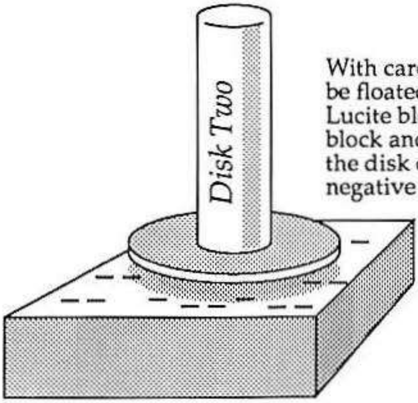

With care disk two can
be floated above the
Lucite block. Both the
block and bottom of
the disk carry repelling
negative charges.

Floating the Electrophorus

Figure 2

stopped producing a charge. I couldn't imagine why both, each being very different in construction, might fail at exactly the same time. Then the culprit made itself known! Just outside the door the automatic lawn sprinklers had come on and extremely humid air was coming through my door!

Another good example of this problem is demonstrated by walking across a padded carpet. When the relative humidity is 20% or less, charges of up to 35,000 volts can be obtained by shuffling one's feet along the carpet. However, raise the humidity to 60% or more, and the maximum charge will plummet to a mere 1,500 volts! If you live in a humid region, you may not be able to get the same performance from a generator that someone living in the desert might. Hopefully, humidity will only be a seasonal problem for you. Dry winter days are best.

There are fascinating static generators which are more or less immune to humidity. These rely on generating a charge by induction rather than by friction. These machines will also be discussed in detail later on.

Be careful in selecting paint for your generator. I have found that many spray paints contain ingredients that will conduct static electricity. For the most part I have had few problems with this, although in a darkened room I have seen the telltale blue corona glow of escaping electrons coming from painted plastic surfaces on a generator. To correct the problem I simply coated the nearby metal surfaces that were losing their charges to the paint with high resistance "corona dope" sold in T.V. and better electronic stores.

Avoid using wood to build generators. Although we think of wood as an insulator, in reality it becomes an excellent conductor of very high voltages. In the machines that follow, you'll see polyvinyl chloride (pvc) pipe used extensively. If you can get acrylic or phenolic tubing, use it. It offers much better dielectric strength.

Dust can bleed away a charge. Dust on a high potential terminal can reduce your generator's efficiency by a whopping 40%. Because your high potential terminal carries a charge, dust will be attracted to it in much the same way that dust accumulates on the front of a T.V. screen. I have actually seen a tiny particle of dust prevent an eighth inch spark from arcing. Keeping your terminal clean is important.

On the other hand, the attraction of dust to your generator can be quite entertaining. One way that I enjoy cleaning my generator is to ground myself and then pass my hand over the high potential terminal. The dust which has been sitting there has acquired a high voltage charge and jumps to your hand which it sees as ground. The effectiveness of this method for really getting your generator clean is doubtful since after a few seconds the dust in your hand looses its charge and heads back for the generator, but it's lots of fun.

Once during a rare Southern California thunderstorm I observed a particularly confused particle of dust. Apparently the storm had charged the air, and my generator had charged a large particle of fluff. The fluff couldn't decide where it wanted to be and would repeatedly fly around my bench only to return to the generator and then again off into the air. It was an astounding sight.

SUPER FRICTION GENERATOR – ROTOSTATIC

Now that you understand the basics of how a charge is picked up by friction, doesn't it seem that it would be possible to design a mechanical means to make and store a charge automatically and continuously? One solution to the problem is what I call a "rotostatic" generator, so named for an old magazine article that called a similar device a "rotostat." I said earlier that it was possible to get 3500 volts from an old plastic bottle, and I wasn't kidding. All you have to do is build the "rotostat".

These generators are extremely simple to build and may be hand cranked or operated by a small electric motor. They consist simply of a plastic drum – an old plastic jar works great – and a rubbing block – wool, nylon, asbestos, leather. When the bottle is rotated against the rubbing block a charge is developed on the plastic and on the block.

We need to attach some metallic brushes to the side of the plastic drum opposite the block to pick up the charges as they are generated. The brushes are attached to a metal sphere where the charges can be stored. A complete schematic of a rotostatic generator is shown in figure 3.

There is an infinite variety of possible configurations for such a generator, but keep in mind in designing yours that for static voltages even wood is a good conductor. It is most important to remember to keep your high potential components well insulated from ground.

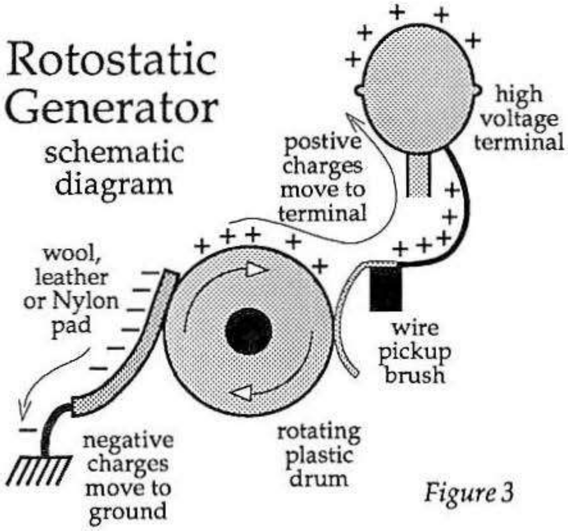
Rotostatic
Generator
schematic
diagram
postive
charges
move to
terminal
high
voltage
terminal
wool,
leather
or Nylon
pad
wire
pickup
brush
negative
charges
move to
ground
rotating
plastic
drum

Figure 3

If your generator is well made it will generate enough power to light a neon bulb (Ne2 type), levitate styrofoam, and arc a spark perhaps as long as 1/16 to 1/8 inch – all of this while still being perfectly safe to touch. When touched, your generator will give your finger only a very light "pop" much like a very small static discharge from a doorknob. As an interesting side note, if you can feel a static charge at all, it will be at least 3,500 volts!

Though static machines deliver these high voltages, their rapid, low current discharges make them unusually safe. Even generators producing 100,000 volts or more rarely present a danger, though I for one would not intentionally take a shock from one. A later chapter on capacitors (energy storage devices) will give you an equation for finding out when a charge has become dangerous.

Building a Film Case Rotostatic Generator

In my rotostatic generator design I have attempted to maximize the output and storage capacity of this type of machine. I am sure that there are still many improvements that can be made. This device works extremely well, is reliable, and very easy to construct. A schematic and a photograph are shown in figure 4.

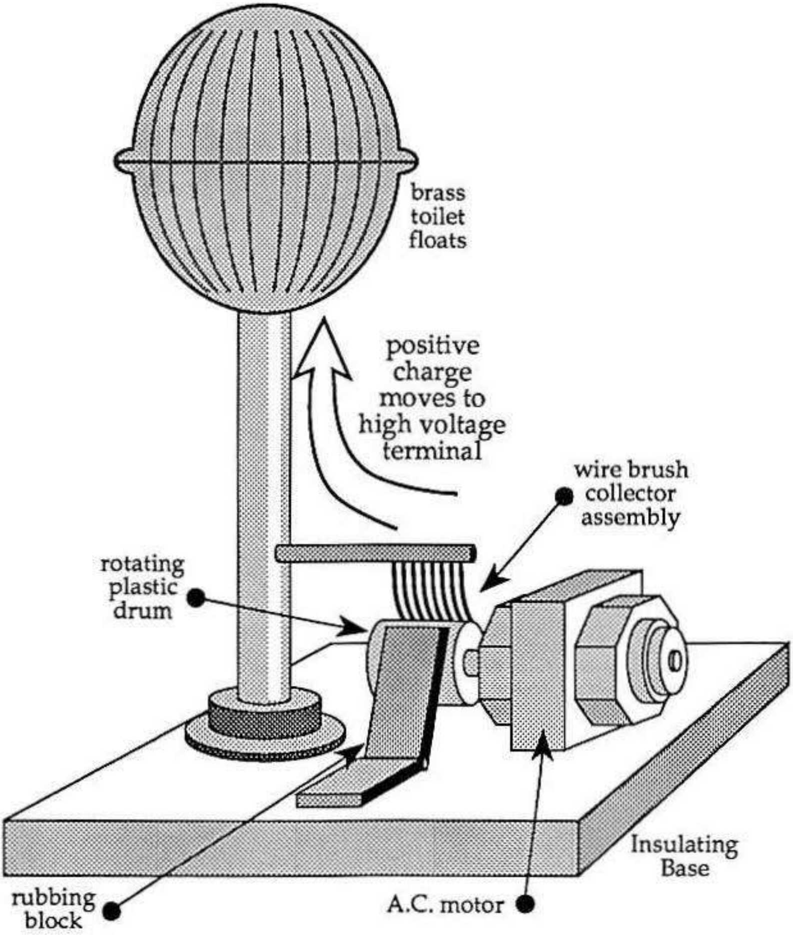
brass toilet floats
positive charge moves to high voltage terminal
wire brush collector assembly
rotating plastic drum
Insulating Base
rubbing block
A.C. motor

Rotostatic Generator

Figure 4

I used a surplus AC tape recorder motor to drive the unit. I like to use tape recorder motors in many projects because they are quiet and often have threaded axles that are very easy to mount components onto. In this case, I wanted to be able to switch plastic drums easily so I could experiment with a wide variety of materials. If you are unable to find such a motor, something similar can be used. A breakdown of

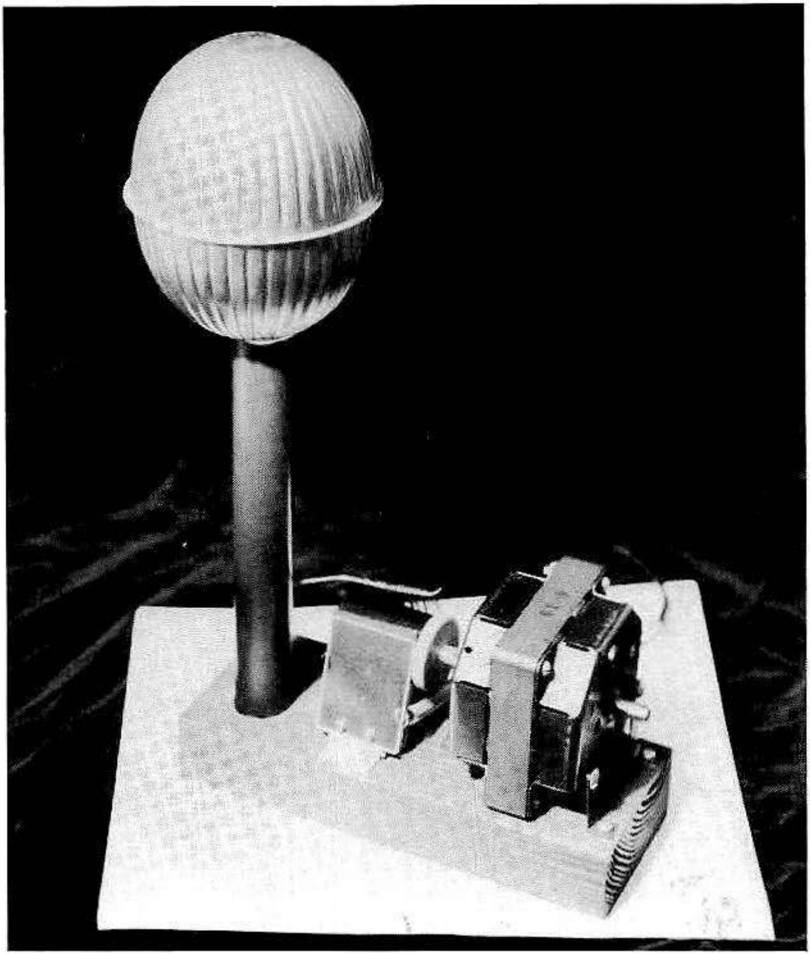

Fig. 5 – The complete Rotostatic Generator uses only a small plastic drum and a few common parts, and yet will light a hand-held 90 volt neon lamp!

AC and DC motors and their suggested hookups are shown in detail in the next chapter.

Assembly is straightforward and simple. The motor is mounted securely with wood screws to the base – in this case a section of 2x4. Make sure this connection is secure since in experimenting with different drums you will undoubtedly have some which will wobble due to imbalances in the drum or in the mounting. Also, remember to select a motor with plenty of power. Tiny D.C. toy motors can work, but they have little torque and can stall, and that can be frustrating.

The drum can be mounted to the motor in a variety of ways. If the motor axle is hollow as the tape recorder axles are, you can simply use a screw to secure the bottle cap directly to the motor axle. If the axle is solid, you will probably want to extend the axle all the way through the drum.

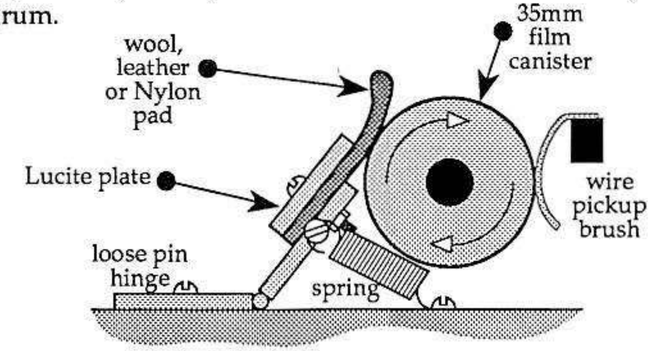

Wiper Assembly
for Roto Static Generator

Figure 6

To mate the motor axle to the axle running through your drum, glue progressively larger pieces of plastic tubing to the smaller axle until the tubing size matches that of the larger axle, then slip the tubing over both axles. Tubing used as fuel line on model airplanes, aquarium tubing, compressor hose, and the insulation stripped from wire can be used to make excellent adapters. Heat shrink tubing can also be used, however, after it has been shrunk it tends to become somewhat brittle.

As long as the pieces of tubing fit tightly over one another, any glue can be used to hold them together. I like cyanoacrylate (Super Glue) because it seems to hold rubber tubing nearly as well as it does skin. If you have used "Super Glue" for any length of time will know what I mean by this.

This method of connecting a motor to an axle also has the beneficial effect of acting as a "universal" joint. That is, your motor can actually power your axle from a variety of angles since the tubing will flex without difficulty. This is useful when using a variety of motors to drive a variety of unusual drums.

The drum must be very well centered on the motor axle, but fortunately, this is surprisingly easy to do. Most plastic pill bottles, film cases etc. already have a little nub right on the center of the lid. I assume this is some natural artifact of the process by which they are formed. Just hunt around a bit and use a bottle with this built-in center mark, and you'll have no difficulty here.

Personally, I like to mount the lid from the bottle to the axle so that the bottle can be easily removed and replaced.

The rubbing block shown in figure 6 turned out to be a very handy design. By looping a piece of wool against the Lucite as shown, the block makes excellent contact with the drum. A variety of rubbing materials can be tried by simply slipping them in between the block and the drum. I think that leather or nylon probably provided the best results when used with the film case. To make these experiments easier, use "loose pin" hinges for mounting the rubbing block.

Rather than use steel brushes to collect the charge from the drum, I came up with what I call "riders." These are formed by wrapping a short length of wire around the copper wire conductor coming from the globe. These wires should be loose enough to dangle down onto the drum and "ride" along its surface. Riders have many advantages. They are very tolerant of wobble in the drum. You may be able to change drum types and sizes without adjusting them in any way. You can improve the efficiency of "riders" by coating unusued surfaces with corona dope.

The PVC tube and collecting ball are easily mounted by drilling a tight hole in the 2x4 to accommodate the PVC tubing, by applying a little Silicone glue to the base, and by tapping it into place with a hammer. A large copper wire is run through the tubing as shown and is connected to the collecting ball with a brass screw. The ball, itself, is then mounted with a bead of Silicone glue.

Testing and Operation

Turn on the static generator's motor and allow it to run for 20 to 30 seconds. It takes some time to build up a charge. Next, hold one lead of a Ne2 neon bulb, available at electronics stores for about 49 cents, and bring the other lead near the brass bulb. When the free lead is nearly touching the bulb, the lamp should flicker or flash. Though the shock from this tiny generator can barely be felt, keep in mind that it takes nearly 90 volts to light the bulb!

Next spread some small, dry, pieces of styrofoam on the top of the ball. By bringing your hand close to the styrofoam you can actually cause them to move mysteriously toward your hands.

Touching this generator is perfectly safe. Under good conditions you should feel and hear a light static "pop" when you touch the fully charged top terminal. If you have trouble feeling the "pop" with your fingers try using your nose. I don't know why, but bringing your nose near the ball brings on a very subtle and strange electric sensation.

With this generator you can do many of the experiments mentioned later including charging up capacitors and energizing spark gaps. So experiment and enjoy.

REMEMBER!

If you are using yourself as ground, as described in the above experiments, you might find that generator performance seems to decrease after several experiments have been performed. What has actually happened is that by repeatedly touching the ball, you have charged yourself up to the same degree of charge as that of the ball, giving the electricity no desire to jump to you! This is simply remedied. Just touch a cold water pipe or other ground from time to time while you experiment.

One minor problem that can pop up is motor speed. Your motor may run too fast. We'll discuss a motor speed control a little later on. Don't worry if the riders hop around a bit. This is usually OK and won't cause problems.

Although a plastic film canister appears in the photos, jars with semi-transparent finishes perform. PVC might also make an exellent drum. Experiment!

A Useful List of Materials for Friction Generators

What follows is a useful list of materials that have a predisposition for taking on a negative or positive charge. By choosing materials that are far apart on this list you can increase the efficiency of a generator.

This list is excerpted from *Nature's Electricity* TAB books No. 2769 and is called "The TriboElectric Series."

POSITIVE CHARGE

Air
Human hands
Asbestos
Glass
Mica
Human hair
Nylon
Wool
Fur
Lead
Silk
Aluminum
Paper

NEUTRAL

Cotton
Steel
Wood
Amber
Sealing Wax
Hard rubber
Nickel, Copper
Brass, Silver
Gold, Platinum
Sulfur
Acetate
Polyester
Styrofoam
Orlon
Saran
Polyurethane
Polyethylene
Polypropylene
Vinyl (PVC)
Silicon
Teflon

NEGATIVE CHARGE

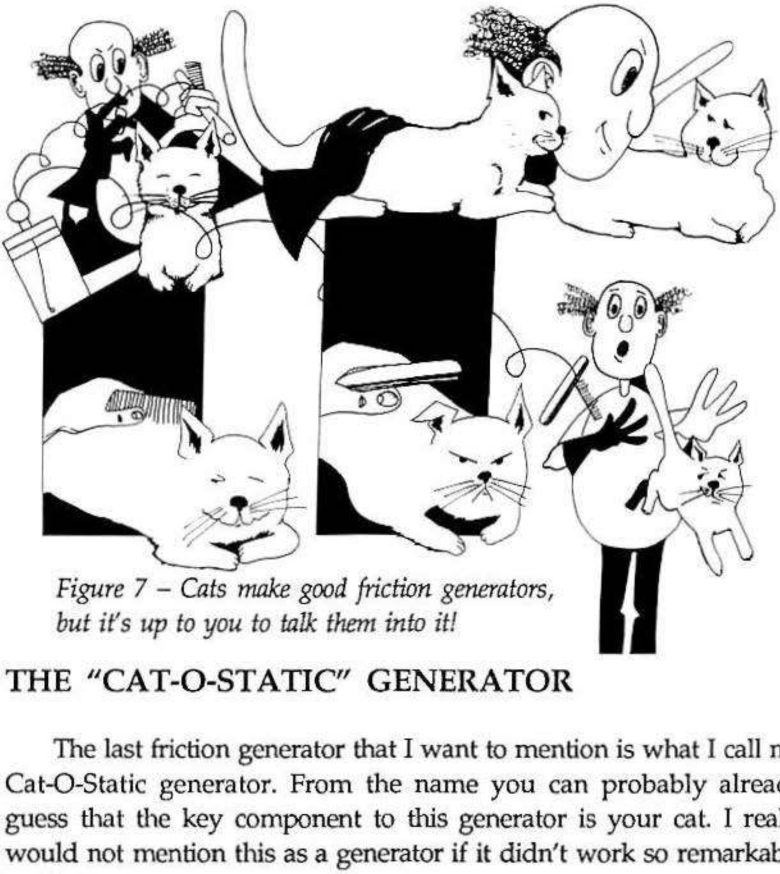

Figure 7 – Cats make good friction generators, but it's up to you to talk them into it!

THE "CAT-O-STATIC" GENERATOR

The last friction generator that I want to mention is what I call my Cat-O-Static generator. From the name you can probably already guess that the key component to this generator is your cat. I really would not mention this as a generator if it didn't work so remarkably well!

For the most part cats enjoy being electrostatic generators, the trick is to make them think they are being petted, and not used as a science experiment. (The cat commandments state somewhere: "Thou shalt allow thyself to be petted only when it suits your purposes and not those of another.")

First, you will need a rubbing material. The very best material I found for this purpose was the artificial leather used in cheap motorcycle gloves. It's actually a type of vinyl. You can secure a pair of these gloves for 5 or 10 dollars at any motorcycle shop. Some are called "man-made leather."

The next thing you will need is a metal flea comb and a length of PVC tubing. You will also need a Leyden jar to store the charge built up by your Cat-o-stat. Leyden jar construction is described in detail later on. The metal flea comb is held at the end of the PVC tubing by a slit cut in the tubing and some glue. A wire is attached from the flea comb to the top of the Leyden jar.

Put a glove on one hand and hold the PVC tubing with the other. Coax you cat into your lap, and tell him wonderful things about himself. This never fails to distract them. Pet him with the glove, and follow the glove with the metal comb to pick up the charge.

With my cat's fur and my particular gloves I am able to build up a charge so large that it will penetrate my vinyl gloves with a strong prickling! However, keep in mind that this type of generator will usually run off and hide before a large Leyden jar can be fully charged. Of course, if you have several cats, use them all to the charge the jar!

AC vs DC Motors and Controls

As we move on to the higher power generators, a little discussion is in order with regards to AC (alternating current) and DC (direct current) motors, and some good ways to control them.

In the generators I've built and discuss here, I have used both AC and DC motors. AC motors have advantages in that usually no external transformers or circuitry are required. They develop plenty of torque – so much so in fact, that you rarely have to worry about stalling the motor. Most AC motors have no brushes or other components to short or wear. They will generally run for years and years, often until the bearings themselves wear out! In general, I prefer using AC motors over DC motors whenever possible.

The major shortcoming of an AC motor is low rpm. The high performance Van-De-Graaff-type generators describe below work best at around 5,000 rpm. If the motor runs faster than this, electrons can literally be "blown" off of the carrying belt, and performance suffers. At lower rpm, leakage can impede the build up of high voltage. Since AC motors do not usually develop high rpm, DC motors should be considered for these high performance generators.

Unlike AC motors, DC motors are portable when battery powered and achieve high rpm easily. Keep in mind that a DC motor rated at

6500 rpm may not achieve that speed when driving your device. The rpm rating is usually the maximum no load speed of the motor.

Although not always necessary, you will probably find that you will want a motor with speed control as part of your DC static generator design. A number of dependable methods are shown in figures 8 through 12.

Figure 8 shows a piece of high resistance wire soldered or tied between two copper posts. One side of the wire is tied to a 12v (or similar) power supply. A wire coming from your motor is attached to an alligator clip which can be moved along the wire to increase or decrease the motor speed. This method works extremely well. However, be aware that the wire can become red hot when supplying current to the motor! Make sure to house it safely.

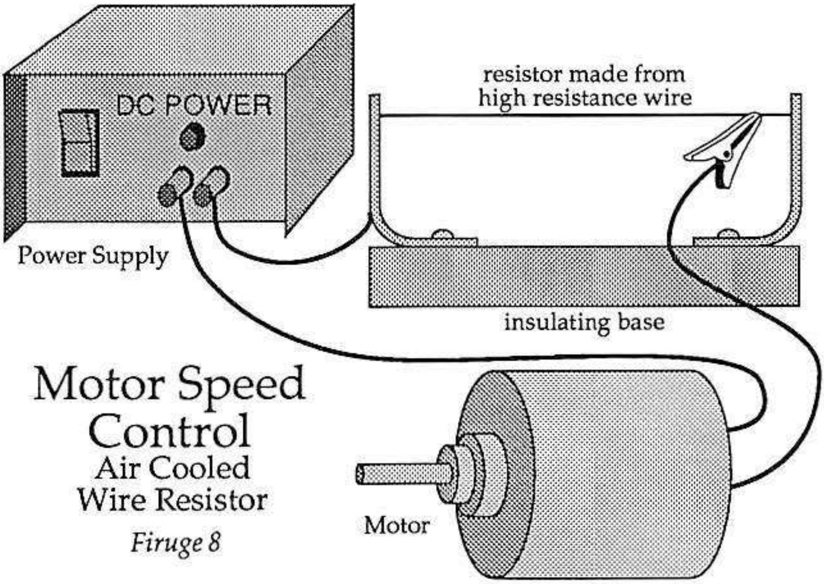

Motor Speed Control
Air Cooled
Wire Resistor
Firuge 8

High resistance wire is easily obtained from a variety of sources. You can use the NiChrome wire that comprises the heating coils in an old toaster, hair dryer, or similar heating device. You can also purchase NiChrome wire from any well-stocked hardware store. Keep in mind that a fairly long piece may be required. Another variation that can be tried is to use a high wattage rheostat, or even the speed control handle from a "slot car" set.

Figure 9 shows the use of power resistors to slow a DC motor.

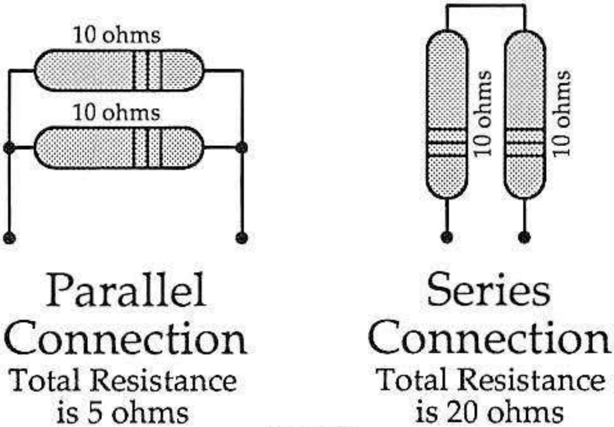

Figure 9

Power resistors are inexpensive, and if you buy a handful of different values you can connect them in parallel or series to get a wide variety of different values. Keep in mind that resistors have two ratings, the power handling capacity measured in Watts, and the resistance measured in Ohms.

An Ohm is a measure of the opposition to current flow. A resistor of 10 Ohms will oppose current flowing through it twice as much as a resistor of five Ohms.

Wattage is simply the amount of power your resistor can handle without being destroyed by overheating. In general, I'd say the higher the wattage rating the better. The Winter's Ring generator shown next required a 25 watt, 10 Ohm resistor, but had to be water cooled to increase its power handling capacity above the 25 watt air cooled rating.

Resistors oppose current flow and in doing so, convert electrical energy into heat. It's normal for them to become very warm. If your resistor gets too hot, its resistance can increase well above its stated value. A change in resistance can cause your motor to run at uneven speeds. You will either need to buy a resistor of a larger wattage, or use some means to cool it such as water (shown and discussed later). As a rule-of-thumb, don't use a resistor of less than 10 watts unless you have a very small or efficient motor.

Figure 10

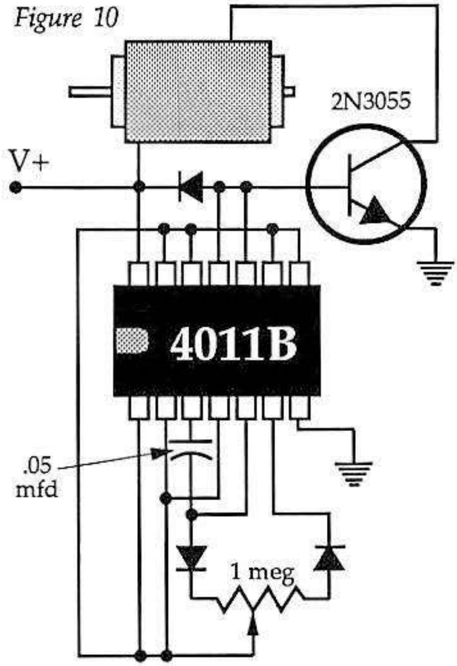

Motor Speed Control
Solid State

Figure 10 shows a very accurate and stable means of controlling a DC motor. An integrated circuit and power transistor are used to send a stream of pulses to the motor. As the pulse rate increases, the motor speed also increases. By simply turning the potentiometer, you can adjust motor speed at any time. For mid-sized DC motors a very large heat sink or water cooling is required for the transistor.

I have found that DC motors operating at high speeds can behave strangely. Slight variances in the motor's speed can set up an oscillation within the belt being driven. Because one side of the belt is being "pulled" and the other

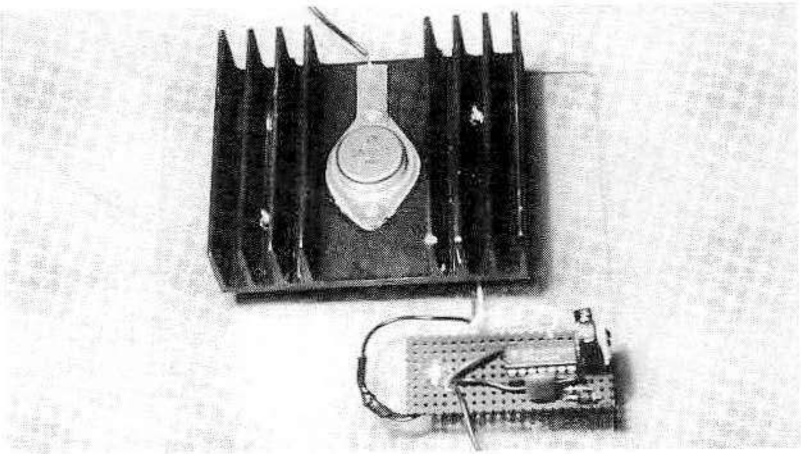

Figure 11 – Solid State Motor Control

is being "pushed", uneven tensions react with the belt's flexibility to produce a "flying" belt that makes only intermittent contact with the rollers or pulleys. This can cause a belt to fly off, bend brushes, or just make a lot of noise. Also, if ball bearings are not used in the rollers, an oscillation can occur among the bushings of the upper roller, the axle, and the belt. All of these problems can be easily solved by using a versatile speed control.

Component Cooling

You may to have to improvise methods of keeping electrical components cool. If you are using a fairly large, high RPM, DC motor which will need sizeable amounts of electrical power, components in the speed control are going to get hot. It may seem a little unorthodox, but water makes a fantastic heat sink for electrical components! Power resistors and even power transistors can be run well beyond ordinary heat sink tolerances when submerged in water. (Figure 12)

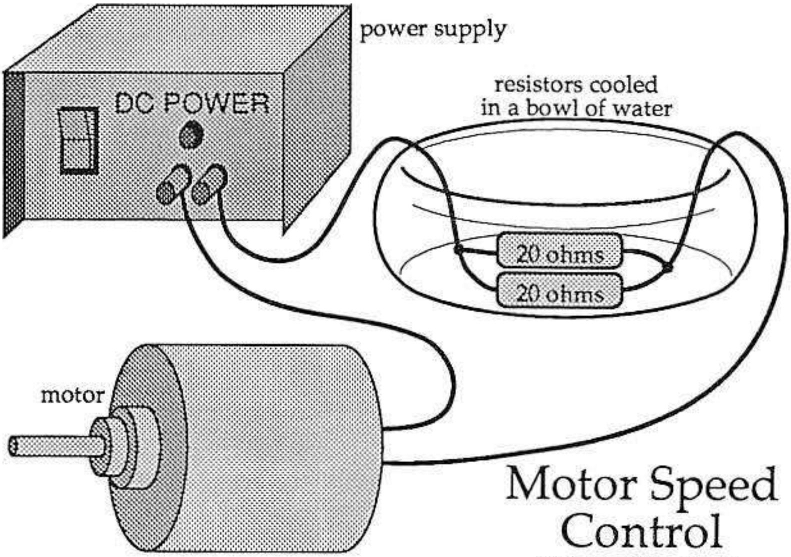

Motor Speed Control
Water Cooled Resistors

Figure 12

Distilled water is a very poor conductor of electricity and will not short out most components. Tap water is usually OK for resistors, but minerals and other gunk will cloud the water after a short time.

As long as the circuit is submerged in water, you know that the temperature of the water must be below 212 degrees Fahrenheit (100 Celsius) at normal atmospheric pressures. Above this temperature the water will boil away. Many electronic components work quite well up to 212 degrees. Just remember to add a little water now and then to compensate for evaporation.

AN EXTERNAL VAN DE GRAAFF USING "WINTER'S RING"

If you built the last generator described, you know that I wasn't exaggerating when I said you could get 3500 volts (or more) from an old film canister. Although interesting, this simple rotostatic generator won't keep you satisfied for long. There are higher voltages to conquer! So let's move on to the next type of generator and try for 10,000 volts (and possibly several times that amount) from a rubber band.

I call the following generator an "external" generator because all of the components of the generator have been mounted externally in such a way as to make construction and experimental modifications fast and easy. Since all of the workings are exposed, this generator is great for the experimenter who wants to quickly try different materials and configurations in developing his own generator design, or for someone who wants to whip together a simple generator to demonstrate.

The high voltage terminal of this generator consists of a metal ring, or toroid, of exceptional efficiency. I call it a "Winter's Ring" although such rings have been traditionally made of wood. They're rarely used anymore. As with other parts of this generator, the Winter's Ring may be quickly replaced by a more traditional spherical-shaped terminal or many others you might wish to try. Why not keep several on hand for demonstration?

Despite the fact that, as we've discussed, electrostatic principles have been well know for thousands of years, it was not until as recently as 1928 that a truly practical machine for generating high static voltages was perfected. Robert J. Van de Graaff, a young Rhodes scholar working at Princeton, designed his first machine to generate the high potential needed to power a particle accelerator to bombard atomic nuclei as part of his experiments in nuclear research.

Although it can be difficult, it is not impossible for the home experimenter to make his own atom smasher! A Van de Graaff generator such as described here is an integral part of such an apparatus. If you are interested in taking on a project of epic proportions, find a copy *"The Amateur Scientist"* by C. L. Stong which is now out of print and very difficult to find, but worth the effort. Or write for similar information from Information Unlimited, P. O. Box 716, Amherst NH 03031.

The Van de Graaff is so simple and obvious in design that it amazes me that is wasn't developed sooner. I have always thought that this is a good example of just how inventions don't have to be extremely complex to have a significant impact on society. How many other simple devices are there waiting to be designed and built? It could be you that dramatically changes the world!

The easiest way to understand the Van de Graaff is to refer to the schematic diagram in figure 13. A Van de Graaff generator consists of a belt, two pulleys, a motor, brushes, a high and a low potential terminals.

In newer machines the bottom pulley is plastic while older machines used an insulator covered by wool or similar rubbing material. The top pulley must be of a conductive material such as metal or even wood which is not a good insulator at high voltages.

When the machine is in operation, a small motor attached to the bottom pulley drives the belt that is lightly stretched between the pulleys. Friction occurs between the surfaces of the belt and the plastic wheel, putting an electrostatic charge on the belt. The charge is carried by the belt toward the upper terminal. A grounded wire brush makes contact with the lower portion of the belt and bleeds off some of the negative charge. The brush at the top of the belt accepts the positive charge remaining and distributes it on the high potential terminal. (Note: It is possible to make a Van de Graaff generator that develops a negative charge on the high potential terminal.)

As the belt moves, charge builds up on the high potential terminal. Theoretically, this charge could increase forever. However, leakage into the air from the upper terminal limits the maximum potential to

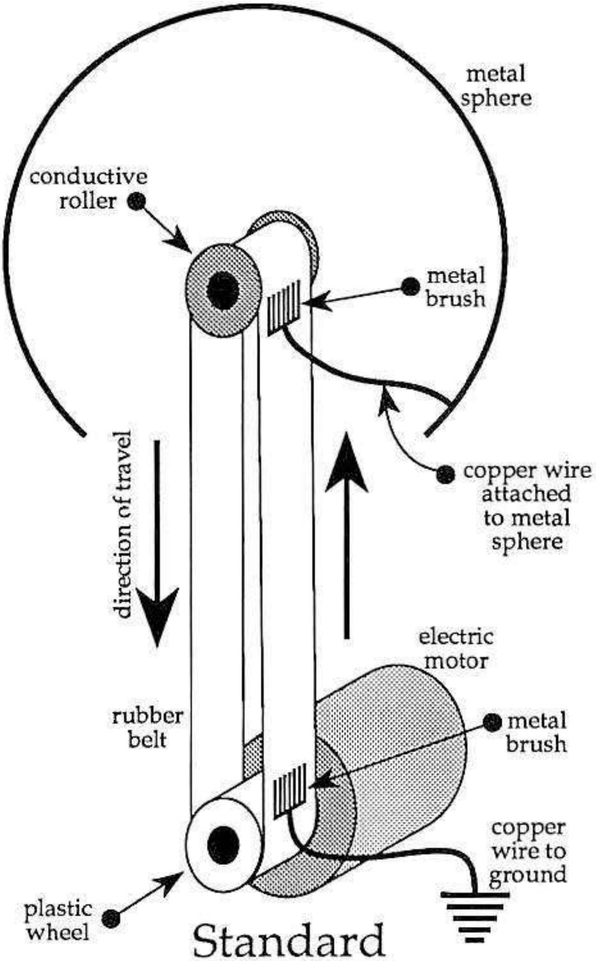
metal sphere
conductive roller
metal brush
direction of travel
copper wire attached to metal sphere
electric motor
rubber belt
metal brush
copper wire to ground
plastic wheel

Standard
Van de Graaff Generator
Schematic Diagram
Figure 13

about 50 to 100 thousand volts typically.

A simple way to calculate the theoretical maximum for a "perfect" spherical top potential is to figure a voltage 70,000 times the smallest radius of the curvature of the high potential terminal. For example–most spheres are sold with reference to their diameter. For a 12 inch diameter sphere, we divide 12 by 2 to get the radius, or 6 inches. We multiply 6 by 70,000 to get a theoretical voltage of 420,000 volts.

Because your terminal will never be a perfect sphere, and because there will be leakage into the atmosphere, theoretical voltages won't be reached. You might find these ball-park calculations interesting, nonetheless.

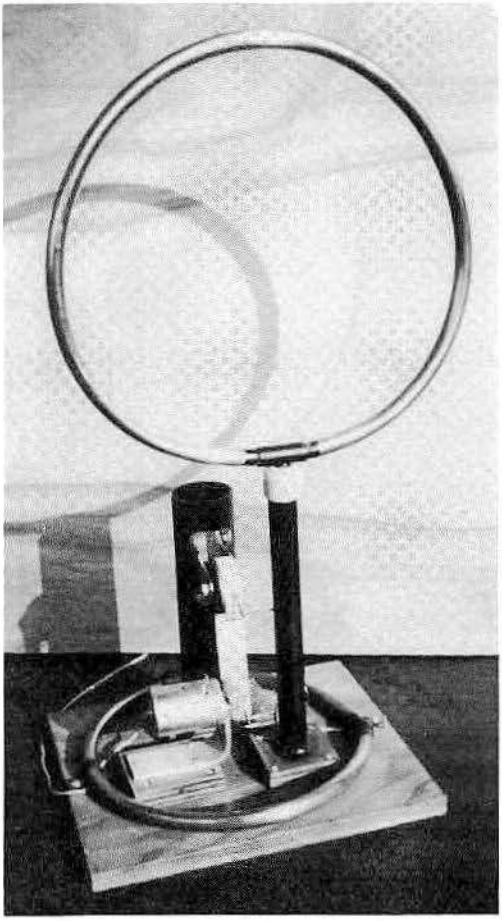

Figure 14 – External Van de Graaff Generator

A photograph of my external Van De Graaff and its schematic are shown in figures 14 and 15. I found this to be an extremely simple and inexpensive design. In operation it will produce a reliable voltage in the range of 10 to 30 thousand volts using an ordinary rubber band. It will produce a solid and exhilarating "pop" when it discharges to your hand. It can throw sparks to ground more than an inch in length. Neon bulbs such as a Ne2 will glow brightly and continuously when near the generator. A blue corona glow can be readily observed from the collecting brushes in a darkened room once your eyes have adjusted to the dark. Negative and positive corona can also be identified by their appearance using this device as discussed later.

The external Van De Graaff shown has some minor performance limitations due to the relatively high friction of the top wheel (no ball bearings) and excessive flexing of the rubber band. A rubber band should be used as a belt only for demonstration purposes. For more

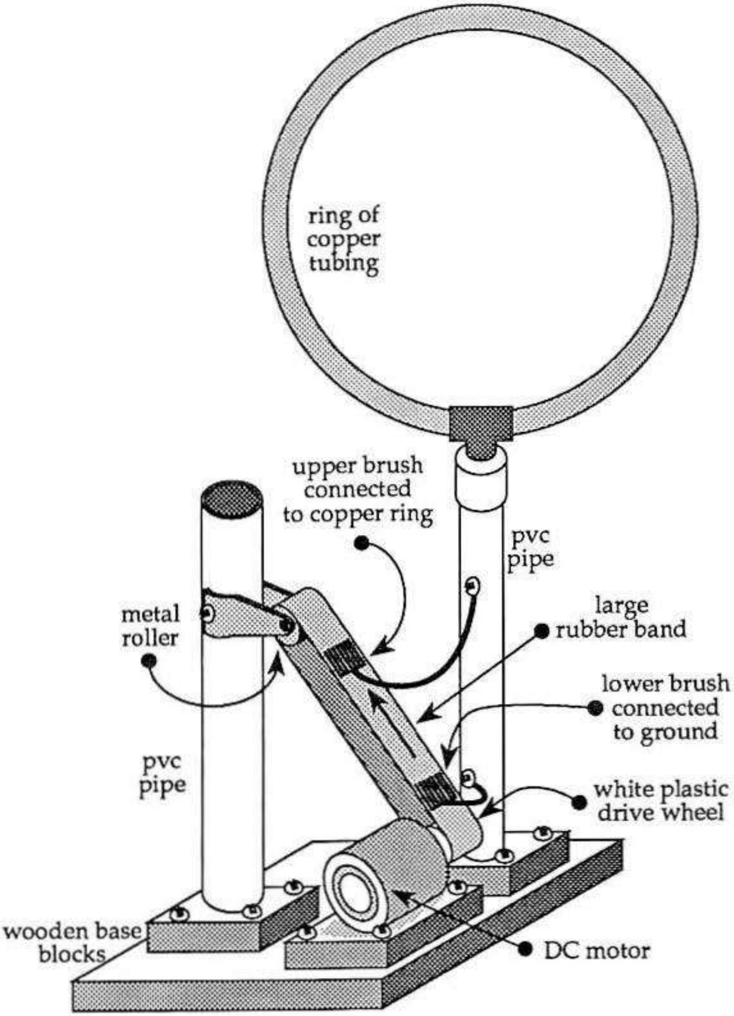
ring of copper tubing
upper brush connected to copper ring
pvc pipe
metal roller
large rubber band
lower brush connected to ground
pvc pipe
white plastic drive wheel
wooden base blocks
DC motor

External
Van de Graaff Generator

Figure 15

serious experimenting, replace the belt with a less flexible belt such as made from a bicycle inner tube or, ideally, neoprene. Also, without ball bearings, the top wheel required occasional lubrication that frequently brought about the risk of oil contamination. Except for these minor problems, this generator has proven to be an excellent and reliable performer. It still remains the most easily modified and observed of the Van de Graaff designs I'll describe.

The construction of the external Van De Graaff is so simple and straightforward that little description is required outside of the drawing. A few tricks can simplify and improve construction.

An easy way to get your motor, roller and belt in alignment is to run your generator while holding the motor with your hand, experimenting until you find the right position for it. If your belt slips off, shift the motor slightly to change its attack or grip on the belt.

When connecting a belt between two rollers in any Van De Graaff type generator, it is natural to assume that pulleys should have flanges on their edges to provide a groove in which the belt will run preventing it from slipping off the driving wheel. Surprisingly enough, the belt will have a tendency to stay centered on the pulleys even when they are somewhat out of alignment! I assume this has something to do with the way the pressures change in the belt as it approaches the edge of a pulley. You may find that a small "hump" in the drive wheel can even be helpful at high rpm. Once aligned, you should have no problems at all with belts slipping off.

Use the largest diameter of copper tubing that you conveniently can for the top terminal. The problem is that when larger diameter copper tubing is bent into small diameter circles, undesireable ripples form. Larger gauge tubing must be bent into larger diameter circles forcing you to compromise between size and efficiency.

When soldering the two ends of the copper tube into the "T" connector, first clean them with extra fine steel wool, then coat (tin) them with a little solder before inserting them. This will give an exceptionally strong and smooth bond. You can use any ordinary propane torch for heating and soldering the tube.

If you use the same type of brushes as shown in the drawing, it might be tempting to shorten the copper wires leading to them from the PVC tubing or otherwise alter this mounting. Don't do it! It is the flexibility in these long copper wires that provides the "spring" needed to keep the brushes flat against the belt.

The plastic bottom pulley is simply one of the white plastic wheels used on cheap furniture casters. It is attached to the motor shaft by carefully drilling out the wheel's center hole to a size slightly smaller than the diameter of the motor axle. You then apply a few drops of super glue to the axle and press or tap the wheel into place.

If the hole in the wheel is larger than the diameter of the axle, glue progressively larger diameter pieces of rubber tubing to the motor axle until the wheel fits snugly in place.

One last note: Since your top terminal is copper which quickly tarnishes, you might want to apply a little Silicone cleaner/protectant such as that sold for cleaning chrome automotive parts. This will keep your top terminal bright and shiny at all times. It also seems to limit leakage and produce an interesting double "pop" when touched.

The Evils of Oil

The only major problem that I encountered in building and operating Van De Graaff generators regardless of design is oil contamination. Since a Van De Graaff generator relies on friction between the plastic wheel and rubber belt to produce a charge, you can imagine the devastating effect oil or any lubricant can have on belt performance. Even the oil from your fingers can reduce the charge carried by the belt. Handle your belts sparingly, and hold them only from the outside edges as you might a record album.

The worst lubricant is a petroleum-based oil such as used to lubricate motor axles and bearings. Angular momentum from the motor axle can actually pull a drop of oil all the way along the axle, along the side of a wheel, and out onto a belt. *Therefore, **Do not oil your motor or bearings excessively.***

If you want a graphic demonstration of this effect, place a drop of fast drying super glue on your motor's axle. Before it has a chance to dry, run your motor for several seconds. If your motor is like mine, you will have a little trail of super glue from the axle to the outside of the wheel where it will form a little blob.

Once your belt and wheels are contaminated, they can be very difficult to clean. The best way to clean the belt is not to clean it all, but to throw it in the trash and get another belt.

The conductive metal top roller can be cleaned with a residue-free cleaner like ammonia, followed by a light sanding with extra fine sand paper and then by another ammonia cleaning.

Cleaning the plastic wheel is the most difficult. The plastic wheel usually builds up a static charge of its own during the time the generator is run. This charge seems to act like a sponge for oil! The only way I have found to remove it is it by attaching a wire to ground and running the motor while dragging the wire along the outside of the wheel. After discharging the pulley in this way, ground a piece of metallic sand paper and again run the motor while sanding away some of the outside of the wheel. This is accompanied with or followed by ammonia cleanings. Other solvents may work, but watch out for those that may dissolve the plastic. I realize this procedure may sound extreme, but I have even tried soaking the plastic wheels in industrial degreasers not available to the general public and have had no luck at removing this oil in any other fashion. If you are careful in your use of oil you will never encounter this problem.

A Few Words On Brushes

In both of the Van De Graaff generator designs described in this book, I used a simple tuft of wire soldered to the end of a thicker solid copper wire spring which provides the force necessary to keep the brush riding firmly against the belt. By far the biggest problem you are likely to encounter will be in keeping the brush riding solidly on the belt. If your brushes are not well placed, or your belt has undue flexibility, the brushes may have to be readjusted frequently.

Many configurations, some very elaborate, have been tried in an attempt to perfect a fool-proof brush assembly. I haven't found one yet. You may want to try a hinge, a spring, a piece of metal screen, or even a conductive rubber or foam such as those use to package integrated circuits. In designing a brush assembly, keep in mind that, ideally, it should be comprised of many fine points like the ends of wires because the density of electrons in a fine point is higher and electron transfer is more likely. Nevertheless I have had very good luck with non-pointed objects such as conductive foams. Most texts, and my own experience has shown that a good old tuft of wire usually wins out, but, as mentioned, this is not a perfect solution to the problem. So, experiment. And have fun!

A Few Words On Belts

A good belt can be made from virtually any insulating material. Rayon, Dacron, rubber, paper, silk, plastic and even ordinary cloth have been used with success. Wear is probably the most important factor to consider in choosing a material. Rubber, for instance, rots in my home town due to the ozone contained in the smog here. Cloth will fray at high speed, although I have coated cloth with an extremely thin coating of Silicone rubber and have made good belts. Neoprene, joined with a diagonal splice is a long-time favorite.

In my opinion, the easiest all-around belt is made from an old bicycle inner tube, or even better, the rubber "liner" that covers the heads of spokes under the rim. One of these belts can be carefully spliced with inner tube cement from a tire patch kit. Even a little "Super Glue" will work if the splice is clamped.

In fabricating a belt also remember that the wider the belt, the faster a charge accumulates (within reason). And, you may also want to give a little thought to the noise generated by certain materials. At 5000 rpm a poorly made belt will make considerable noise.

One last belt option is to purchase a ready made belt from a scientific supply house such as Analytical Scientific, Post Box 675, Helotes, TX 78023. Such companies can often supply many different sizes and types of belts. I would suggest ordering the thickest, and strongest belt offered for a home-made generator. I find the price of commercial belts a bit too expensive for my bank account, but you might find their prices more acceptable depending on your project.

THE HIGH POWER "CLASSIC" VAN DE GRAAFF

Although these voltages are impressive, keep in mind that the classic Van De Graaff will easily produce 60 to 100 thousand volts or more and is only slightly more difficult to build.

The classic design for the Van De Graaff is shown in photograph 16. The one minor concession I made in my design was to use two stainless steel bowls for the high potential collector instead of the traditional sphere. I did this simply because I could not readily find a

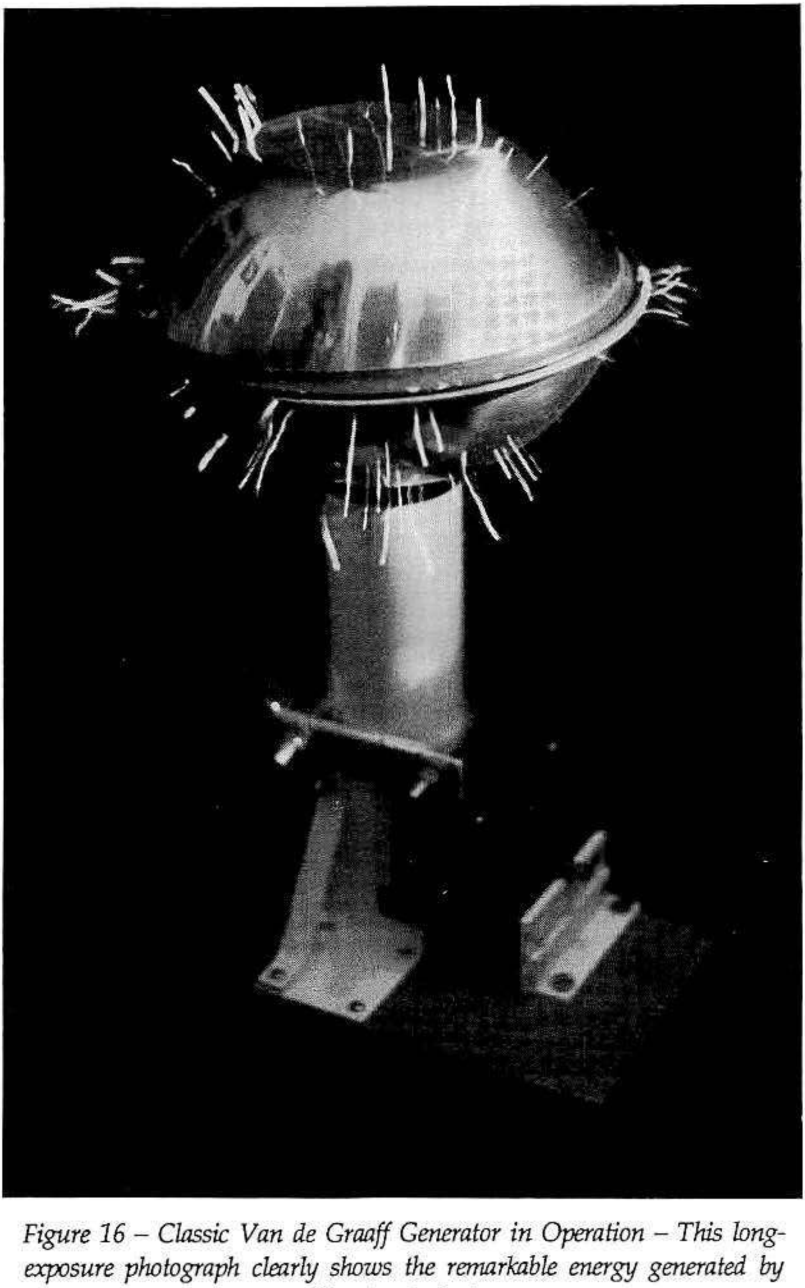

Figure 16 – Classic Van de Graaff Generator in Operation – This long-exposure photograph clearly shows the remarkable energy generated by this simple device.

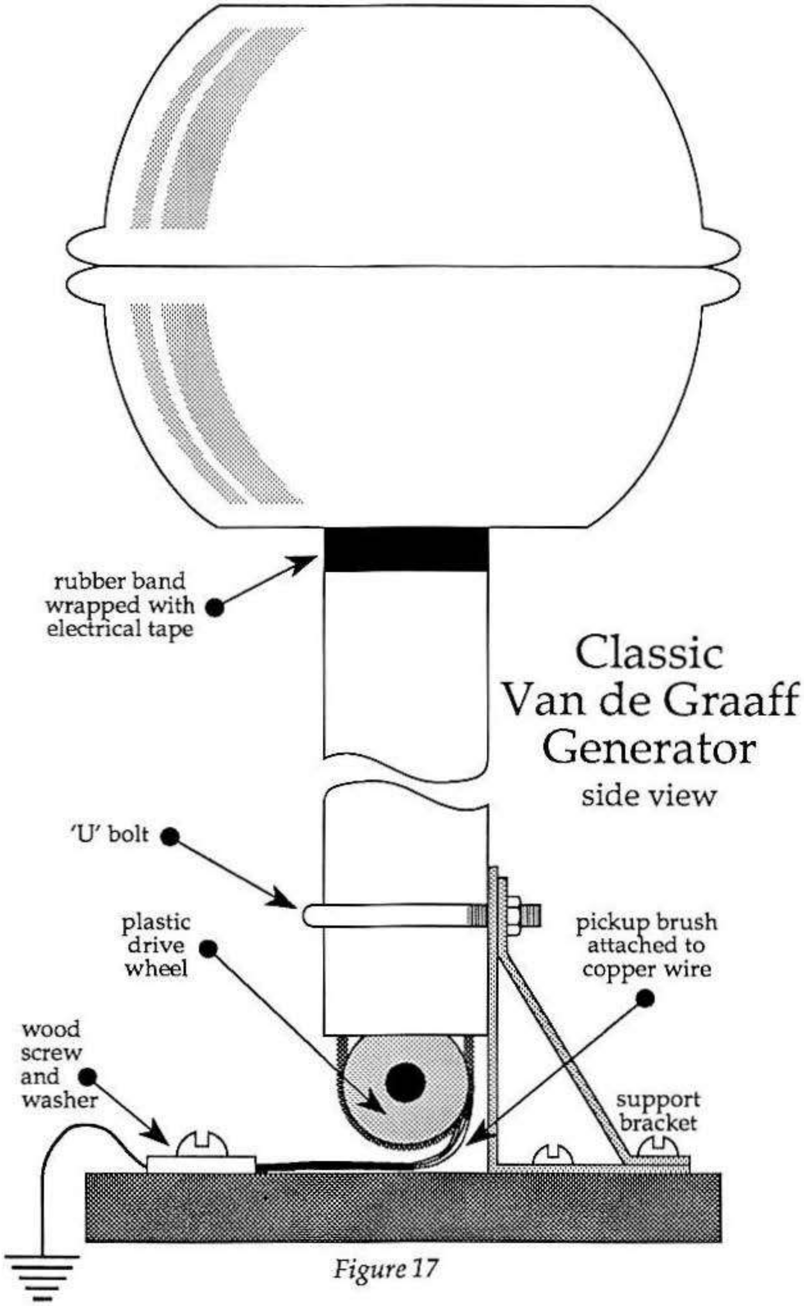
rubber band wrapped with electrical tape
Classic Van de Graaff Generator
side view
'U' bolt
plastic drive wheel
pickup brush attached to copper wire
wood screw and washer
support bracket

Figure 17

sphere of the proper dimensions. Besides, two bowls allow easy access to the top roller and brushes, and I wanted to test my hunch that almost as much power could be obtained from the bowls if the edges were heavily doped and insulated as could be obtained from a traditional spherical terminal.

All of the experiments to be described can be performed with this unit. It is an all-around solid machine and is fun to operate. If you plan to build only one generator from this book, and are wondering which generator to select, this machine is the best all-around performer. If carefully built, this design will exceed a voltage of 100,000 volts. It will arc thick, blue sparks well over four inches to an outstretched hand and deliver an invigorating, if not slightly uncomfortable, shock.

This generator is capable of charging up a person's body to a high voltage. It makes possible many experiments such as causing a person's hair to stand on end, or the rapid charging of large capacitors. The device will even produce an ominous "crackling" as your hand approaches the high potential terminal.

By standing on a large plastic insulator such as an overturned detergent bucket, a friend of mine was able to charge her body to such a high potential that her hands would "crackle" any time she pointed to a person nearby in the room. I think a physics or chemistry teacher would have little trouble keeping his/her students attention if the mere pointing of their hands produced this ominous cracking!

Though I am certainly not recommending the following approach, this generator was also inadvertently useful in training a dog in a single lesson not to jump on people. My dog, Merfi, had the bad habit of greeting people by jumping all over them. Nothing, it seemed, could discourage this behavior. One afternoon some friends and I were standing around a completed Van De Graaff and commenting on what an unusually large charge we had managed to bestow upon a person. Suddenly Merfi bounded into the shop. Despite our pleas, Merfi shot across the shop with his usual enthusiasm and leapt nose first into my highly charged friend, no doubt, intending to bestow upon him his traditional greeting.

We could do nothing but watch as a large blue spark jumped to Merfi's moist, waiting nose. Merfi stumbled back, temporarily lost control of his bladder, and then slunk out the door. A few moments later everyone was petting and reassuring Merfi, yet, to this day we have had virtually no problems with Merfi's traditional greetings. He still eyes the friend from whom he received the shock with suspicion. And with good reason!

I suppose it is needless to say that it is wise to keep any and all pets a safe distance from your experiments. As with the external Van De Graaff generator, very little description outside of the drawings is required for constructing the classic Van De Graaff design.

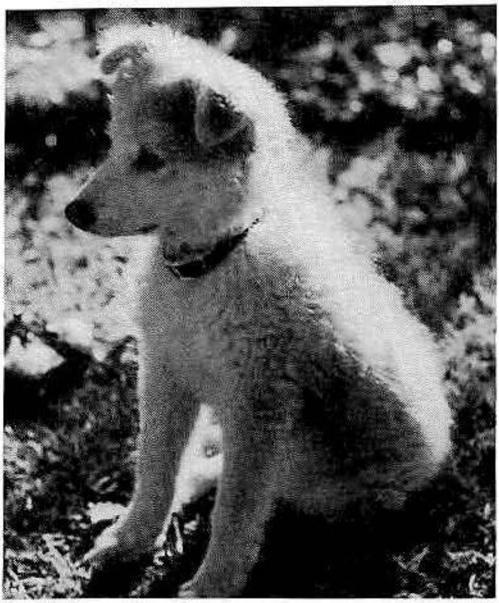
Figure 18 - "Merfi" – (An accidental electrostatic participant)

Things to keep in mind

The edges of the bowls are relatively sharp, even though the edges have been folded over. To prevent the charge from leaking through these sharp edges, they are coated with a special paint called "corona dope" which can be purchased in most T.V. and well-stocked electronic stores. It is available in both spray and brush-on forms. Corona dope has a high dielectric strength, and will prevent a charge from leaking off into the air. It can be used on any spots suspected of leakage on any of your high voltage projects. Keep in mind that the thicker the coating of corona dope, the better it will shield. I used around eight thick coats on my Van De Graaff and should have used more. Sparks fly right through it!

As with the last generator, the lower pulley is a white plastic wheel taken from an inexpensive furniture caster. It was mounted in the same way as was done in the external Van de Graaff.

The upper pulley is the front hub of a 10 speed bicycle wheel. I tried several different methods to reduce the friction of the hub bearings to make it turn more freely. I tried a variety of lighter oils and even

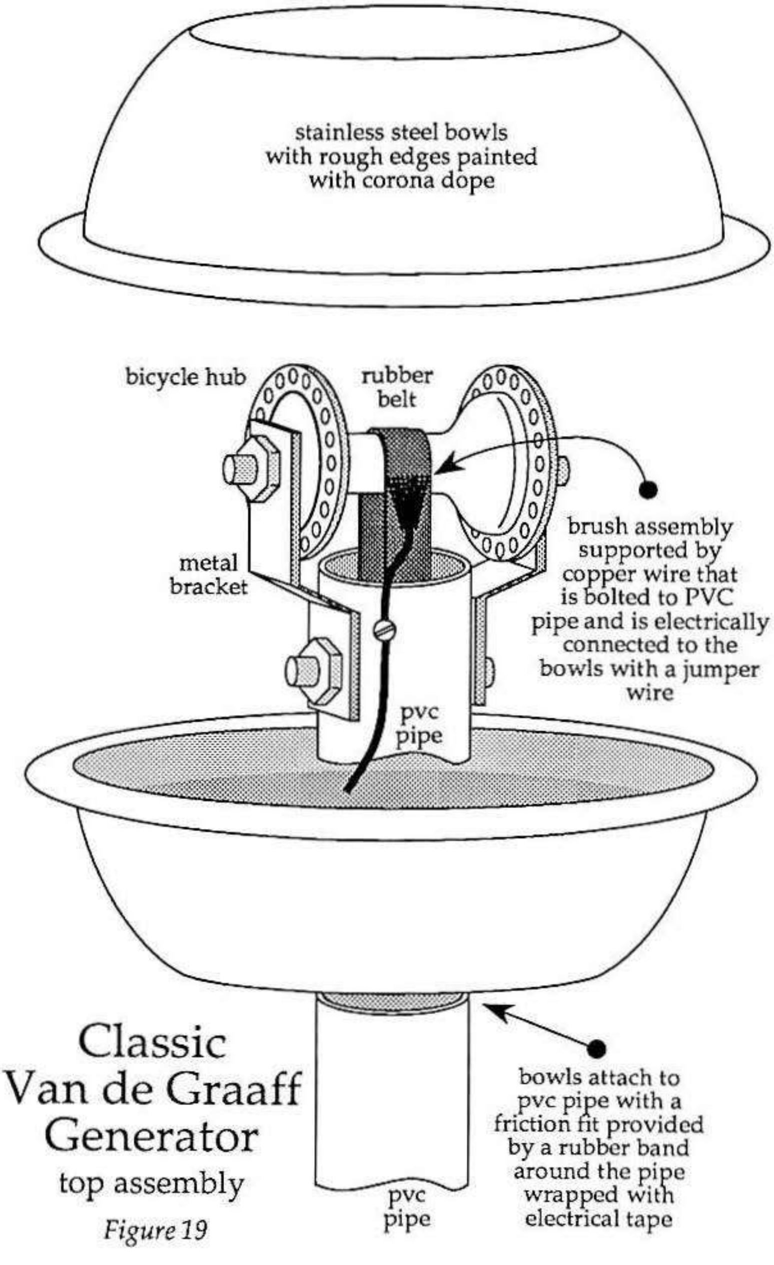
stainless steel bowls with rough edges painted with corona dope
bicycle hub
rubber belt
metal bracket
brush assembly supported by copper wire that is bolted to PVC pipe and is electrically connected to the bowls with a jumper wire
pvc pipe
bowls attach to pvc pipe with a friction fit provided by a rubber band around the pipe wrapped with electrical tape
pvc pipe

Classic Van de Graaff Generator top assembly

Figure 19

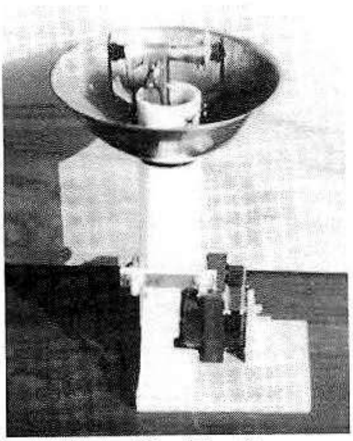

Figure 20 – Top Open

removed the rubber covers that protect the ball bearings from dirt contamination. I found that lighter oil works well, but because of the risk of oil contamination as discussed earlier, I ended up using grease. Removing the covers also helps, but when the generator is running and the roller is spinning, a sizable charge builds up on its surface, and every dust particle in the area is attracted to it! Although you may want to experiment, I have come to the conclusion that the hub should be left pretty much as it is when it comes off the bike.

The two top bowls can be hinged together from the inside with a wide strip of black tape, or can be bolted together by some nylon screws through the lip. Keep in mind that you want the surface of these bowls to remain as smooth as possible to prevent electrons from leaking off. Don't use metal hinges or any sharp objects in attaching them.

The main body tube is fabricated from three-inch diameter PVC drain pipe. (Remember that acrylic or phenolic pipe might be superior to

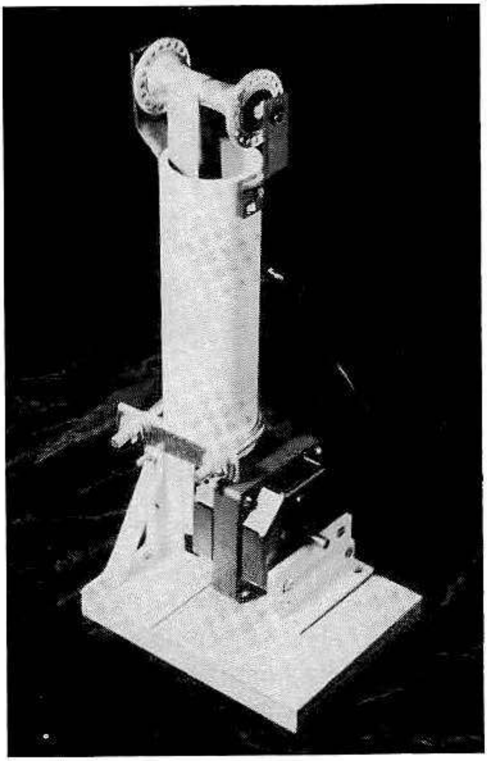

Figure 21 – Terminal Removed

Figure 22

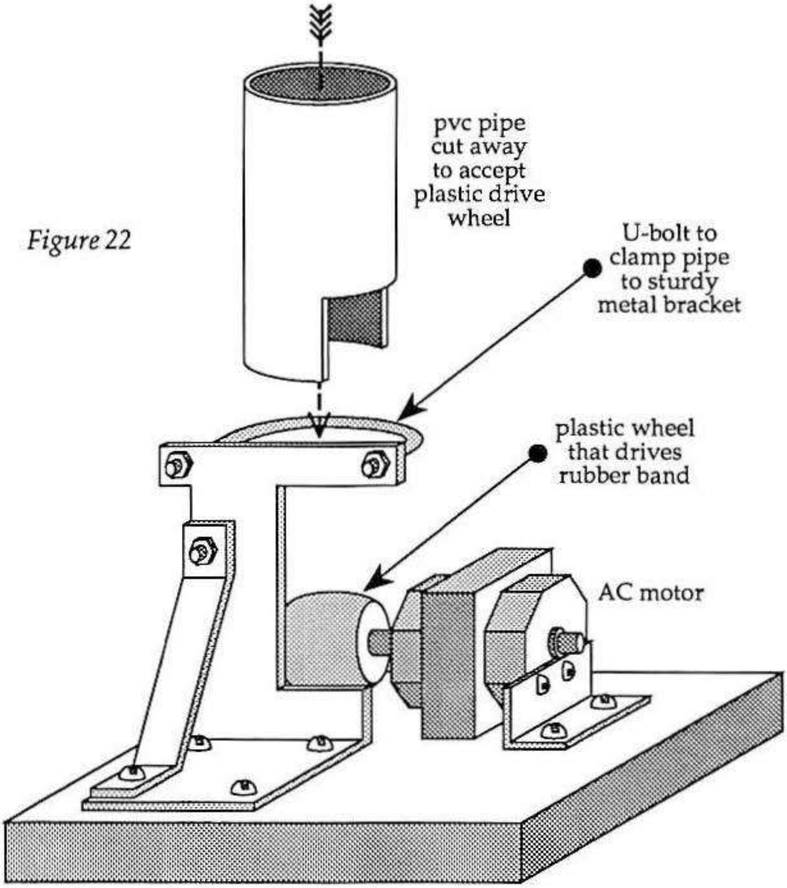
pvc pipe
cut away
to accept
plastic drive
wheel
U-bolt to
clamp pipe
to sturdy
metal bracket
plastic wheel
that drives
rubber band
AC motor

Classic Van de Graaff Generator
bottom assembly

PVC.) A slot is cut into the lower portion of the tube to allow the motor and plastic drive wheel to be inserted through wall. The PVC pipe is supported by a metal L-bracket and a U-bolt. Depending on the size of your motor, you may want to cut a notch into one side of the L-bracket to provide enough clearance for the motor to fully enter the bottom of the tube. Since this "L" bracket will support your whole generator, make sure it is sturdy and of a heavy gauge. You may even wish to add a metal brace as I did to increase rigidity.

When you have completed this generator, you will be able to perform any of the experiments that follow. This easy-to-build, high performance generator is a very exciting machine to own!

AN EXTREMELY HIGH POWER VAN DE GRAAFF

Before we move on to other types of static generators, I'd like to briefly describe one of the highest performance Van De Graaff generators I have ever seen. Shown in figure 24 is a design for an extremely high power Van De Graaff generator. I have not constructed this unit. I am providing details on this machine should you want to explore machines beyond the limits of my humble designs. This schematic will provide the essential design details.

At maximum efficiency this design could generate voltages exceeding 1,000,000 volts without using an auxiliary storage capacitor!

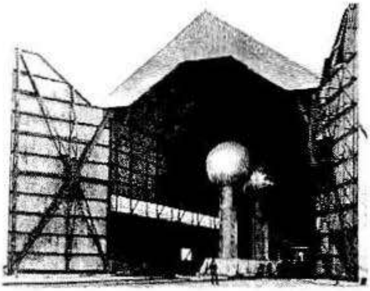

Figure 23 – Ten million volt Van de Graaff generators were built in the 1930's. One generator here is negatively charged, and the other is positive. Generators of this size are not recommended as a first project. (Perhaps not even as a second! ...but call me if you build one!)

Voltages of this order can be dangerous, possibly lethal. This is not a toy and should only be constructed by a serious and cautious experimenter. Because of its high output, I think this unit would make an outstanding laboratory power source for a homemade atom smasher.

The high power unit uses an external high voltage power supply to "spray" an initial charge on the rubber belt. Notice in the schematic that the ends of the spray wires are pointed. A rectified high voltage source between 5,000 to 10,000 volts is ideal. For low cost and ease of assembly you may want to try using an automotive coil as a transformer. For details see Lindsay Publications *"Build a 40,000 Volt Induction Coil."* (Written by an author who is generally a good sort.)

In this design, both top and bottom rollers are metal. The top roller itself is insulated from the high potential sphere. Charge is sprayed

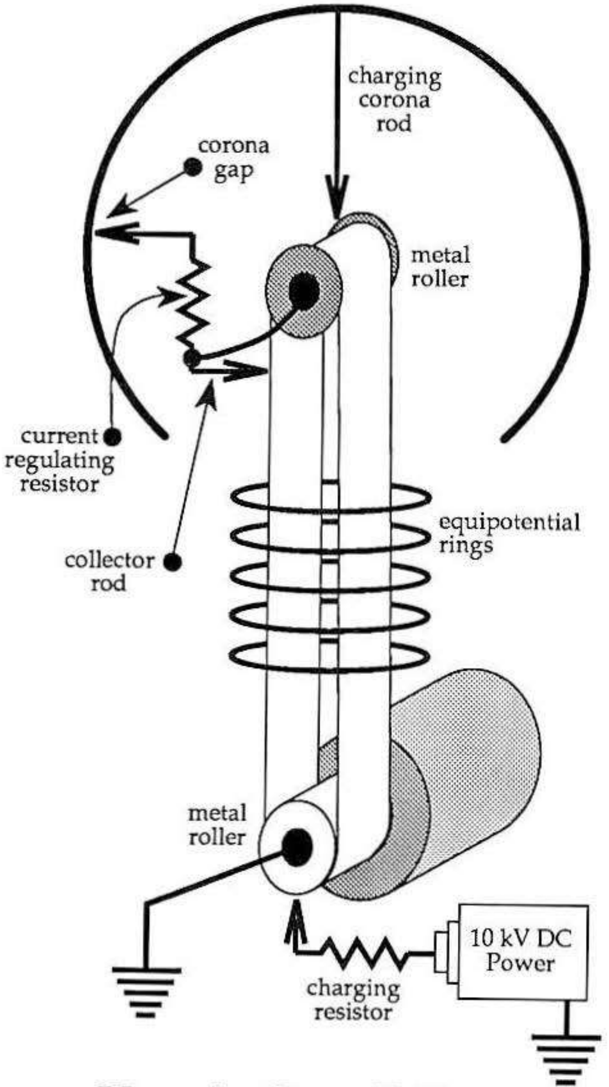
charging
corona
rod
corona
gap
metal
roller
current
regulating
resistor
equipotential
rings
collector
rod
metal
roller
10 kV DC
Power
charging
resistor

Van de Graaff Generator
high power configuration
using external excitation
(schematic diagram)

Figure 24

Figure 25

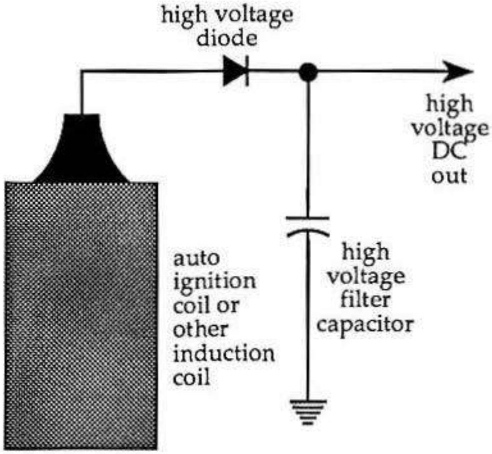
high voltage diode
high voltage DC out
auto ignition coil or other induction coil
high voltage filter capacitor

Induction Coil Exciter for Van de Graaf Generator

output must be recitified and filtered

onto the carrying belt as it passes through the corona discharge between the high voltage power supply and the lower roller which is grounded. A second set of points located just inside the upper high potential terminal picks up the charge and conducts it to the upper pulley. Once highly charged, current flows to the high potential terminal through the current regulating resistor. This configuration can also include a corona gap as shown near the inside surface of the high potential terminal.

An additional set of spray points, labeled "charging corona rod" on drawing, connect directly to the high potential terminal and are situated directly above the pulley.

According to *The Amateur Scientist*, *"The difference in potential between the upper pulley, made 'live' by the voltage drop across the current regulating resistor and corona gap, and the high-voltage terminal causes these points to 'spray a charge of opposite sign onto the downward run of the belt.' "*

The value of the current-regulating resistor can be computed roughly by Ohm's law. In small machines it is usually on the order of

40 megOhms.

Optimum belt speed for a high power unit is between 4,000 and 6,000 feet per minute. Most high power units use belts made from rubberized fabric.

If your design is to exceed 200,000 volts, charge distribution along the insulating column becomes important. Machines using belts more than four inches wide should be equipped with equipotential rings as shown spaced every two inches.

A large Van De Graaff generator intended to reach the million volt range is often sealed in a steel tank containing inert gases such as carbon dioxide and kept at many atmospheres of pressure. Keeping the gases at high pressure serves to increase their insulating properties and improves both the voltage and current output of the generator.

Owning a generator of this size would be incredible!

INDUCTIVE ELECTROSTATIC GENERATORS

Wouldn't it be wonderful to have a generator that could produce almost 90,000 volts while being virtually immune to changes in humidity? What if that same generator could be run extensively without concern for wear or maintenance? What if the same generator could be "daisy chained" to other generators to increase their output? What if it could be run as a motor, operating on electrostatic charges, just as a common D.C. generator can also by run as a motor? Well, what we would certainly have is a Dirod generator.

The Dirod generator is another invention of Mr. A. D. Moore. The Dirod gets its name from the main components of the generator: a high grade Plexiglas disk and a series of metal rods. "Dirod" is short for disk and rods. Its operation is most easily understood by observing a schematic representation of its workings.

Figure 26 shows a schematic of a basic Dirod generator. A number of metal rods are attached to an insulating disk (Lucite, glass, plastic, etc.) which is mounted on a bearing that allows it to rotate. In the center of the disk is an axle which is stationary, and attached to this axle is a wire with two brush ends as shown. As the disk rotates, each pair of metal rods is momentarily electrically connected to the stationary brushes as the rods pass by them.

Outside of the wheel we have two smooth metal plates each with

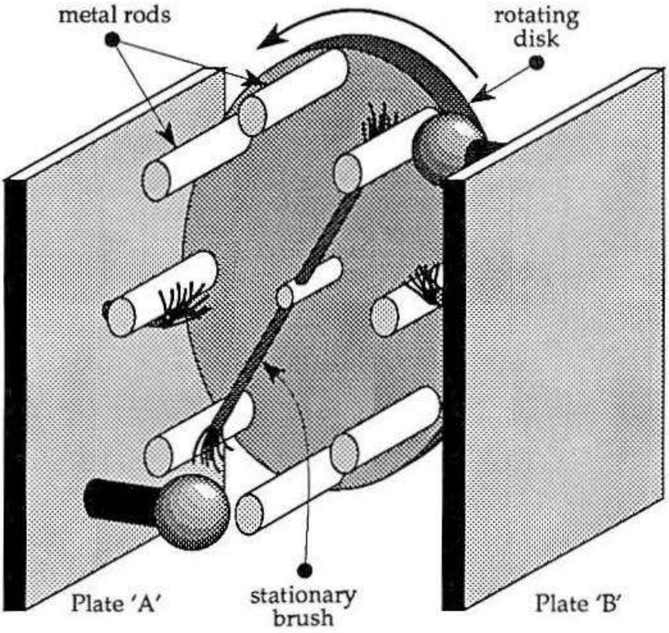

Induction Generator
schematic diagram

Figure 26

an electrode attached as shown and each having a pair of metal brushes as well. Positive and negative charges will form on these plates in the following manner. When the wheel begins its rotation, plate A and plate B will have similar amounts of negative and positive charges – similar, but of course, not exactly the same! Let's say that plate B has a slightly negative charge, and plate A has a slightly positive charge. Now, as a metal rod moves by the electrode from plate A, it is momentarily connected to the rod moving past the electrode from plate B, connected by the stationary brushes from the center axle as shown in figure 27a. At this point, since plate B has a slightly negative charge, the positive charges in the two bars rush into the bar closest to plate B (negative and positive charges attract), and the negative charges rush into the bar closest to plate A. This is comparable to one electrophorus charging the other without having them

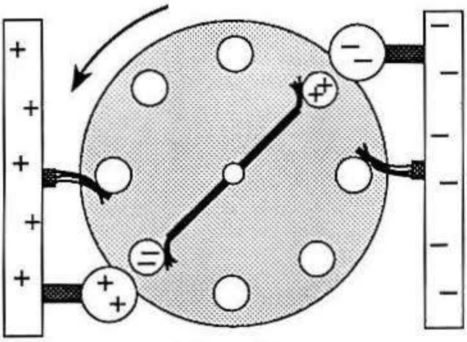

Figure 27a

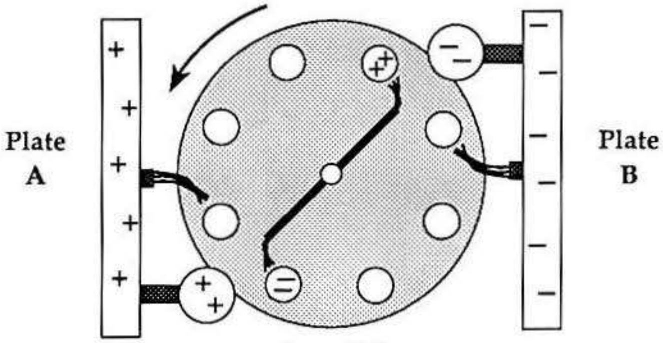
Plate A
Plate B

Figure 27b

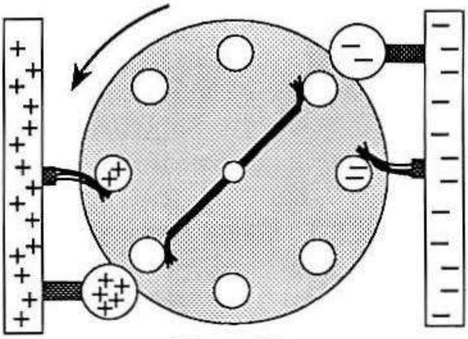

Figure 27c

Induction Generator Operation

actually touch.

As the disk continues to rotate, the two bars move past the brushes and break electrical contact. Now we have one bar with a mostly positive charge and one with a mostly negative charge. Figure 27b

As the bars continue to revolve around to the opposite plate, they briefly make a connection with the brushes from that opposite plate and bestow upon that plate their charge. Figure 27c. Now the plate has an even higher charge than it did originally. It charges the next rod passing by it even more forcefully. In this fashion the plates rapidly build up a larger and larger charge.

When you consider how fast these rods can be spun, and how fast these charges double, you realize how quickly this generator can produce high voltage!

To be a little more technical in describing the difference between our previous friction generators and an induction generator of this type, remember that friction generators charge by contact between two bodies and by the sharing of electrons from the resulting contact. Inductive generators make no demand on free electrons, but on the field they produce. This field causes electrons in the charging material to veer from their orbits. This displacement sets up or induces a charge in the formerly uncharged material. The series of figures 27a through 27c show not only how the machine moves but also how charges are built up during each step of the charging process described.

It is important to understand these simple principles of induction since they are critical to understanding and building the electrostatic motors that follow.

It is interesting to note that the Dirod generator is virtually the same as the world's first electric motor known as the "Franklin" motor for its inventor, Benjamin Franklin, and discussed later on.

Construction of the Dirod Generator

Construction of the Dirod generator is, again, straightforward and simple when viewing the drawings. Keep in mind that the metal rods can be mounted horizontally as shown or vertically, radiating out from the center. Vertical mounting is much trickier. However it is much less stressful on the Lucite disk allowing higher rotational speed to be achieved.

Also very important in achieving maximum voltages is a corona shield which is not shown in the drawing. A corona shield can be fabricated simply by placing a small sheet of Plexiglas between the metal spheres and the rotating rods. I would suggest that you first try your generator without a corona shield, and then with. If your design is optimum, your voltages will increase with the shield.

Because the brushes in this type of design can really take a beating I strongly suggest using a conductive foam such as the foam integrated circuits are packaged in or a conductive rubber brush. If forced to use a metal brush, use a metal that is very springy.

I built an induction generator by modifying the Franklin motor shown later on. By adding brushes and two stainless steel collectors to the Franklin motor, I had both an electrostatic motor and generator! Although I plan to build a much larger, more traditional Dirod machine, this small motor/generator still holds my fascination and continues to teach me the principles I will need to employ in devising the larger unit.

MEASURING VOLTAGES AND POLARITY

WARNING: As mentioned before, the electrostatic voltages produced by the generators in this text are generally safe for the experimenter even should he be accidentally shocked. The operation of the following devices is described for use with electrostatic devices. Should you choose to measure the voltage of other devices, more stringent safety precautions should be taken. Use extreme caution at all times.

Information in this chapter is useful not only in the field of electrostatics, but can also be valuable to those who work with Tesla coils, Oudin coils, induction coils, or other high voltage power generators. One of the most surprisingly difficult things to do with a high voltage is to accurately measure it. I was frustrated to find that text after text said simply that measurement of high voltages was impossible for the amateur. As a result of experimentation and having studied the writings of extraordinary experimenters such as Professor A.D. Moore, I can show you several methods of measuring high voltage that seem quite dependable and accurate. You can also use one of several means to determine the polarity of a high DC voltage – something which is rarely mentioned in any text.

Finding Polarity and Simple Voltage

There are several very simple means for determining the polarity of a high potential. The first, and by far the most sensitive is an FET based "electroscope." Figure 45 shows an electroscope of my own design. The device is simple to assemble and is shown in both schematic and component layout forms. The device works extremely well.

I highly recommend that you build one of these. It can show you the polarity of both very high voltages as well as low electrostatic voltages. In fact, this electroscope will respond to the static build up found in clothes taken from a dryer, or the charge that builds up on a hard rubber comb.

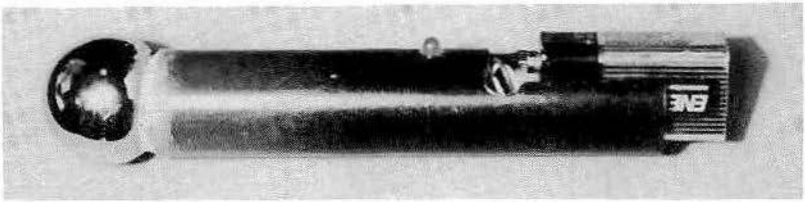

Figure 45 – The FET Electroscope

You may want to modify my design. You should consider replacing the 10 megohm resistor with one of a smaller or larger value. It controls the sensitivity of the circuit. The larger the resistor used, the more sensitive the device becomes, the smaller the resistor, the less

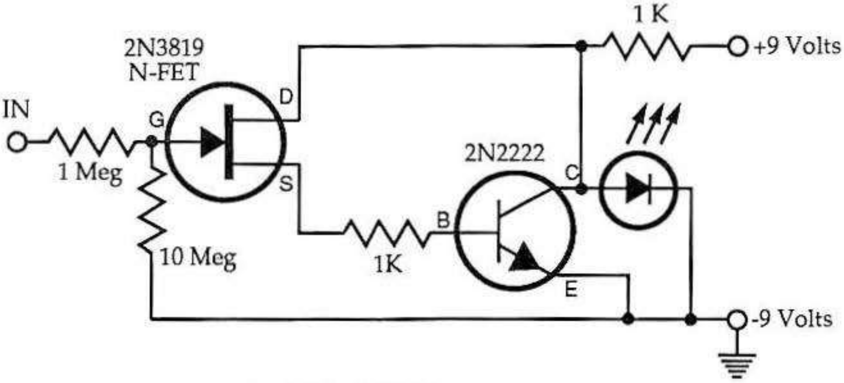
1 K
+9 Volts
2N3819
N-FET
IN
G
D
S
1 Meg
2N2222
C
B
E
10 Meg
1K
-9 Volts

FET Electroscope

Schematic Diagram

Figure 29

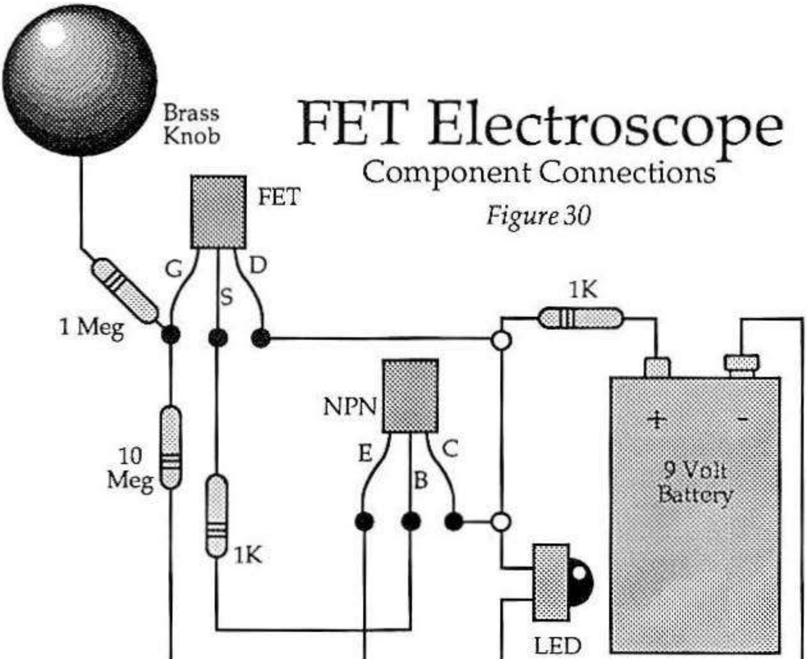
Brass
Knob
FET
G
S
D
1 Meg
1K
NPN
E
B
C
10
Meg
1K
+
-
9 Volt
Battery
LED

FET Electroscope

Component Connections

Figure 30

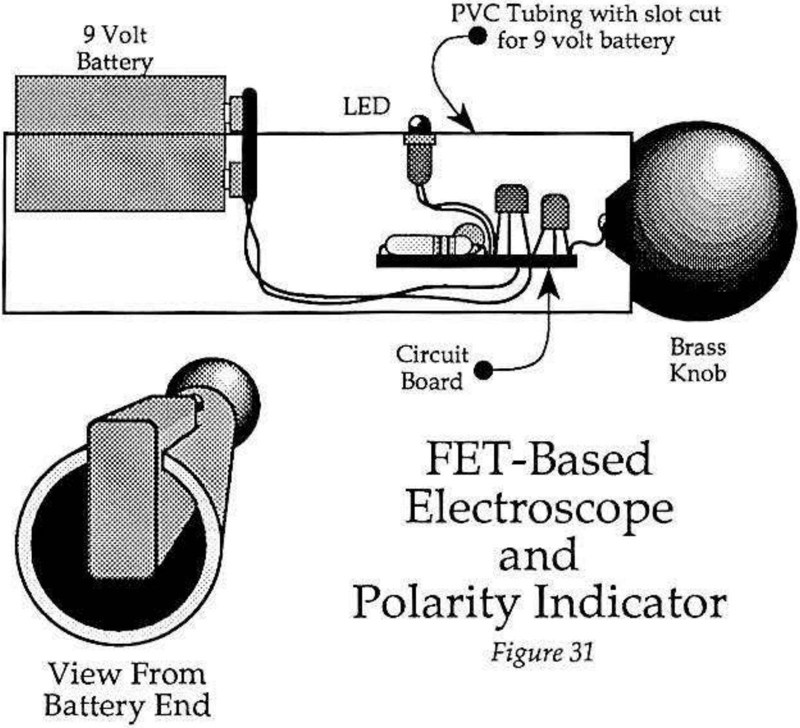
9 Volt
Battery
PVC Tubing with slot cut
for 9 volt battery
LED
Circuit
Board
Brass
Knob
View From
Battery End

FET-Based
Electroscope
and
Polarity Indicator

Figure 31

sensitive – within reason, of course. The resistor can also be removed all together for maximum sensitivity, but the device becomes unreliable because even stray radio emissions will be picked up.

When properly adjusted, the FET scope will light when placed in contact with a positive static charge. Keep in mind that the FET transistor, which is the heart of the circuit, could be damaged by an extremely high voltage, however, I have used my scope without difficulty to more than 30 thousand volts (30 Kv). This voltage was measured by placing the scope directly against the Van de Graaff with an appropriate spark gap.

A MECHANICAL ELECTROSCOPE

A mechanical version of the electroscope can be made quite easily from a glass jar, a little wire, and some aluminum foil. A mechanical electroscope is shown in figure 32.

When the top electrode on this scope is touched to a static voltage source, the aluminum leaves in the jar will separate. This happens because the foil in the jar takes on a uniform charge. Since like charges repel each other, the two leaves of foil are forced apart.

The demonstration of an electroscope can be quite impressive. Although it would be difficult, if not impossible, to calibrate, a

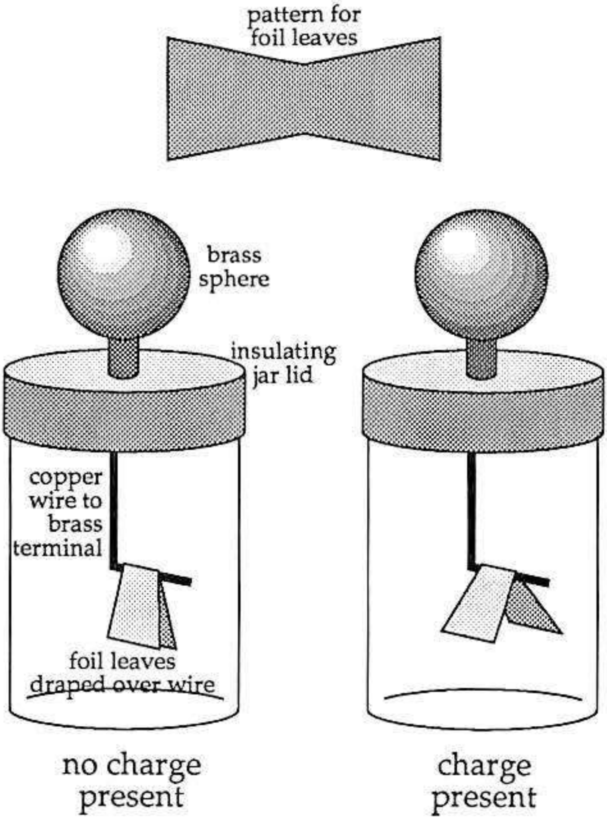
pattern for foil leaves
brass sphere
insulating jar lid
copper wire to brass terminal
foil leaves draped over wire

no charge present

charge present

Mechanical Electroscope

Figure 32

mechanical electroscope of this type can also be used to measure voltage. Heavier leaves require greater voltages to separate them. Therefore, the distance the leaves move away from one another indicates a relative voltage.

The Neon Lamp Bank

Another less sensitive but much simpler way to measure low electrostatic voltages is with the use of a neon lamp bank. Figure 33 shows the assembly of such a bank. Ne2 neon lamps are wired in series along a Plexiglas panel as shown in the drawing. Since each lamp fires at about 70 volts, five bulbs wired in series will read a voltage of up to 350 volts. You can quickly see how larger or smaller banks can be easily built. The cost of these bulbs is quite low, so even fairly large banks can be built without great cost.

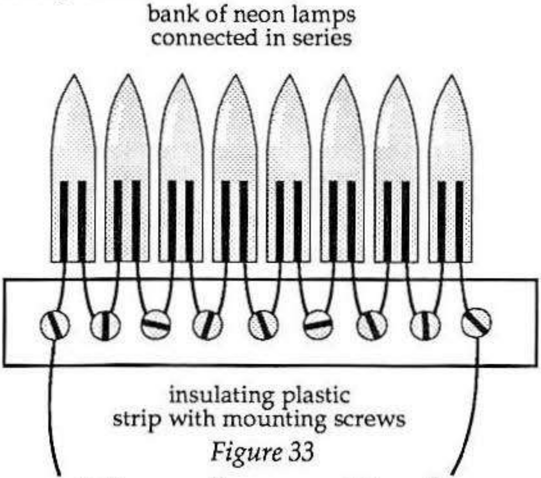

Figure 33

Neon Lamp Bank

When connected to a source of high voltage, it is quite easy to see which bulbs are lit. Sometimes, however, you may have to darken the room to get a more accurate reading. Neon lamps are not always easy to see.

You can increase visibility by leaving a small gap between the lamp bank and your high voltage source. Letting a spark periodically jump the gap will trigger the lamp bank. Attaching the bank directly to generator can draw enough current from the generator (load it down) such that it never reaches peak voltage. With a small Van de Graaff a steady stream of bright flashes will occur every two or three seconds.

Though this means of voltage measurement is crude, you could improve it somewhat by using bulbs with very close firing voltages.

The best way to find such bulbs is with the circuit such as shown in figure 34. Note that a volt-ohm meter can also be connected across the leads shown in order to gauge the exact voltage at which the bulbs will fire. By slowly bringing up the input voltage to a set of bulbs wired in parallel you can find which bulbs fire together and at what voltage.

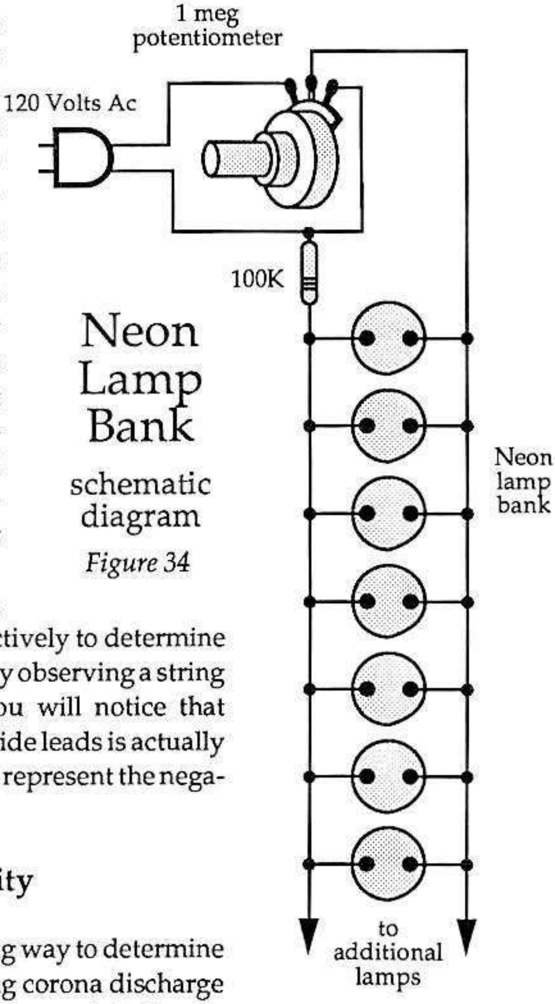

Neon Lamp Bank

schematic diagram

Figure 34

A neon bank can also be used very effectively to determine polarity. After carefully observing a string of flickering bulbs you will notice that only one of the two inside leads is actually lighting. This lead will represent the negative electrode.

Corona and Polarity

Another interesting way to determine polarity is by observing corona discharge from your high voltage potential. Corona discharge is that strange very faint blue glow that surrounds portions of your high potential terminal. To observe it, you will need to be in a darkened room with your eyes fully adjusted to the dark and your generator running.

One might assume that positive corona and negative corona look the same, but this is not true! Positive corona is even and uniform in appearance while negative corona is bushy, often with streaks or streamers that move around.

I first noticed this while using my external Van de Graaff, the same one described earlier. Not having had much experience with this generator, I immediately assumed that the light blue glow I was seeing in the darkness was due to arcing in the D.C. motor. It was too dark to see where the corona was coming from. Fearing the contacts in the motor would eventually be damaged, I tried every method I could think of to stop motor arcing, but the pesky blue glow went on unabated. Finally, I reached out in the darkness thinking I was going to touch the motor, but much to my surprise ended up touching the top of the generator! I had not expected to observe a corona discharge on this machine.

Another interesting fact is that negative corona can be heard! If you find that a hissing sound is coming from your generator, this will be generated by the negative corona discharges!

Using Spark Gaps To Determine Voltage

A spark gap is the best way I know of to accurately measure high voltage without having to resort to very expensive laboratory equipment.

Figure 35 shows the basic configuration of a "sphere gap" capable of accurately measuring up to 90,000 volts. A sphere gap consists simply of two hollow stainless steel spheres 1/2 inches in diameter which are attached to a length of threaded rod with epoxy glue that has

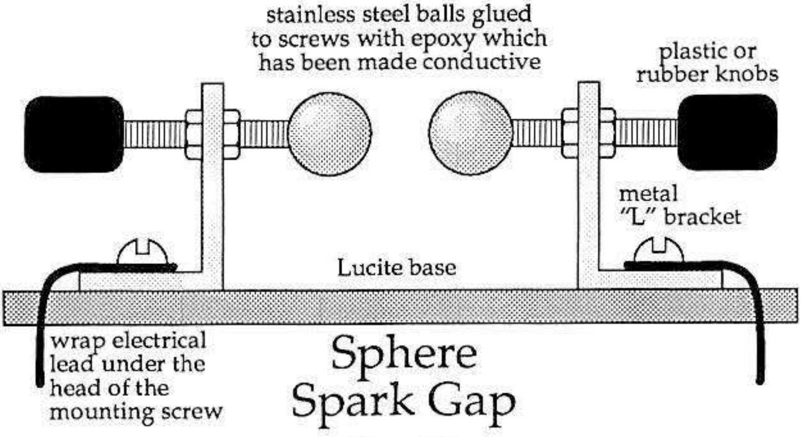
stainless steel balls glued to screws with epoxy which has been made conductive
plastic or rubber knobs
metal "L" bracket
Lucite base
wrap electrical lead under the head of the mounting screw
Sphere Spark Gap

Figure 35

been made conductive by some additive or has been coated with a little silver paint in order to make electrical contact between the sphere and the rod. The rod is threaded through a metal stand, and plastic handles are added to allow the spheres to be turned while the device is operating. The spheres must be at least 3 sphere diameters above the base to prevent distortion of the electrical field which can give inaccurate results.

Another type of gap you can build is a rod gap consisting of two aluminum rods 3/4 inch in diameter. The rods are lined up with each other just like the sphere gap. The ends of the rods are rounded by turning on a lathe. Precision is important. The ends of the finished rods must be as close to hemispherical in shape as possible and must be highly polished. A rod gap of this type will give fairly accurate measurements up to 60,000 volts.

Unfortunately, there are too many variables to give a simple equation for determining at what voltage a spark will bridge a given gap. The only practical solution is to either calibrate your gap by using devices that generate known voltages, or to use a set of measurements which have already been taken.

What follows is a chart of known values for the apparatus described. These charts are excerpts from *Electrostatics* by A.D. Moore.

kV	Rod gap	Sphere gap
5	0.056	
10	0.112	
15	0.170	
20	0.228	0.234
25	0.288	
30	0.355	0.370
35	0.430	
40	0.520	0.531
45	0.625	
50	0.750	0.710
55	0.903	
60	1.122	0.950
70		1.102
80		1.380
90		1.700

Note: Gaps listed are in inches.

One limitation of measuring high voltages with a sphere gap is that you will loose accuracy dramatically if the gap between the spheres becomes larger than the diameter of one sphere. Because of this, you might need a general way to calculate voltages for spheres of different sizes. Also, you may not be able to find a sphere of the exact size mentioned and will need to have a reference for adjusting your figures.

What follows is a listing of spark gaps for spheres of several different sizes. By using these numbers and plotting along a curve, you can remain fairly accurate in determining the characteristics of spheres whose sizes fall between those shown.

Gap Distances in Inches
for a variety of voltages and sphere sizes

	sphere sizes in cm				
kV	2.5	3	4	5	10
10	0.30	0.30	0.30	0.30	0.30
20	0.61	0.61	0.61	0.61	0.61
30	0.95	0.95	0.95	0.95	0.95
40	1.40	1.32	1.30	1.30	1.30
50	2.00	1.82	1.73	1.71	1.65
60	2.81	2.40	2.21	2.16	2.02
70	4.05	3.16	2.80	2.68	2.41
80		4.40	3.50	3.26	2.82
90			4.40	3.93	3.28
100				4.76	3.75

THE "NOON" ION HIGH VOLTAGE METER

OK, I have to admit that it is terribly arrogant for me to name one of my devices after myself, but what the heck. For the record, I won't complain if you affix your name to one of your inventions! I think that you will find this meter which has solved so many of my voltage measuring problems to be an extremely useful tool in experimenting with high voltage.

As we have discussed, the surfaces of a high potential terminal should be smooth and round, ideally spherical, in order to prevent electrons from flying off into space. If we intentionally direct these electrons onto a sharply pointed terminal, the density of electrons will increase to the point where they begin to jump out into the atmosphere that surrounds the terminal. As these electrons stream away, the pointed terminal moves in an opposite direction by virtue of Newton's second law of physics (ie. for every action there is an equal and opposite reaction.)

This reaction is extremely small. However, it is measurable, and can even drive a small motor (described later) or, as I have discovered, push the pointer on a standard panel meter! Since the strength of the reaction is directly related to the voltage accelerating the electrons, it is possible to construct a meter capable of measuring high voltage.

One difficulty with this meter is that, if constructed from off-the-shelf parts, no two meters will read exactly the same due to variations in the exact dimensions of the pointed terminal and the meter itself. It is necessary to use other methods to initially calibrate your meter such as comparing meter readings to voltages ascertained by the spark gap method.

When using your meter remember to keep it well clear of ground or any object that might attract the ions streaming off your high voltage generator. Anything that upsets the stream of electrons can affect the meter reading.

HIGH VOLTAGE AND SAFETY

Because storage devices can accumulate enough electrical charge to be lethal, we must discuss safety.

Most texts simply warn the experimenter to stay away from a high voltage discharge, and, of course, this is a good idea. No shock is completely safe, and people have been injured and even killed by remarkably small voltages.

Even low voltage can deliver high damaging current. A capacitor carrying just a few volts has been known to melt the end of a screwdriver that accidentally bridged its terminals!

If your experiences with electricity are anything like mine, no matter how careful you try to be, electricity will occasionally find some clever way to zap you when you aren't looking. Because of this, you should know how to determine when any source of high voltage you are experimenting with is safe.

The U.S. Atomic Energy Commission helps us by telling us that a discharge to the human body of 10 joules is known to be hazardous to life. A discharge of a mere 1/4 joules will deliver a heavy shock.

Use the following equation for figuring the charge in Joules:

$$\text{Joules} = \frac{1}{2}CV^2$$

where:

C = Capacitance in Farads
V = Voltage in Volts

What follows (again from A. D. Moore) is a short list of capacitors and the number of kilovolts needed to store up a dangerous 10 joule charge.

microfarads (uF)	kilovolts (kV)
0.002	100
0.20	10
20	1
80	.5
320	.25
2000	.1

From this you'll see that even small capacitors can store lethal amounts of electrical charge. Just because you may be dealing with less than 10 Joules does NOT mean you are safe! Again, remember that no shock is completely safe. If you are careful and use good judgement, you will have little to fear from electricity.

Most of the generators in this book will produce more than enough power to run any of the experiments outlined directly. High voltage static electricity is safe because there is so little current available. A Van de Graaff machine moving 50 square inches of belt per second generates only one microampere. Although the voltage (the pressure behind current) can be very high, there is very little current to be accelerated. Power is the product of voltage times current. One microampere would have to be accelerated by one million volts to be able to deliver one watt of power. With such low current levels, power levels are necessarily low and relatively safe.

If, however, small currents accelerated by high voltage are stored in a capacitor or bank of capacitors, it is possible to accumulate enough charge at high voltage that when it quickly discharges it can produce deadly levels of current. Some commercial high power static generators go as far as to present the warning "Do not attach capacitors." If you are experimenting with larger capacitors than those described, remember that it is no longer safe to verify static charges by touching terminals! Capacitors can store charges from *minutes to months* depending on the dielectric! Careless or thoughtless use can prove fatal!

Electrostatic generators are unusually safe. Most generators will produce no worse a shock than a good jolt resulting from shuffling across the carpet. However, in coupling your generators to energy storage capacitors you're moving into dangerously high voltages. Be very careful!

ENERGY STORAGE WITH THE LEYDEN JAR

A Leyden jar is simply a capacitor, that is, a device that stores electrical energy. Like all capacitors it consists of two conducting metal plates separated by an insulator. Figure 36

A brass knob such as those used on the handles of bathroom drawers is attached to an insulator, in this case, a flat piece of Lucite. You could also use cork, plastic, wax, or almost any insulator. The Lucite is cemented to the lid of a jar with a little Silicone glue. A large hole cut in the lid of the jar allows a wire to pass from the bottom of the knob to the bottom of the jar. The inside and outside of the jar are coated with a layer of aluminum foil. The foil should be worked with your fingers until it smoothly conforms to the sides of the jar. A wire is attached to the outside foil, and the entire outside is then wrapped with black tape.

To use the Leyden jar, the wire lead coming from the outside foil is attached to ground. A cold water pipe makes an excellent ground. The brass knob is connected to the output terminal of your generator.

Leyden jars can be assembled in any size, larger jars having larger capacitance. My personal favorite Leyden jar was made from a 5

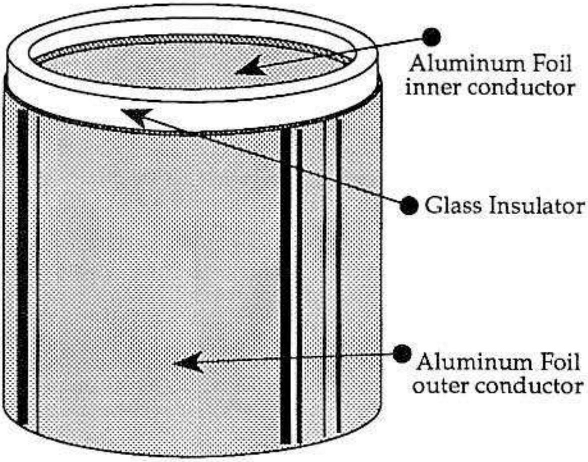

Leyden Jar & Other Capacitors
showing the
conductor-insulator-conductor layers

Figure 36

gallon white plastic pail that had contained detergent. This huge capacitor has been charged by the "Classic Van de Graaff" generator. Under good conditions it yields a very heavy, thick, multi-colored, spark of more than 4 inches accompanied by a "crack" that could be heard for some distance.

A word of warning: I made the error of touching this capacitor when I had thought it fully discharged, and though it was partially discharged at the time, I received a remarkably painful shock. This capacitor should be treated with respect. You can never be too careful.

Leyden Jar Tool

One tool you should build for your Leyden jar is a discharge rod. After experimenting with your Leyden jar, you will always want to be sure that it is fully discharged before handling or storing it. This is a very good habit to get into to avoid getting zapped when you are in a forgetful mood.

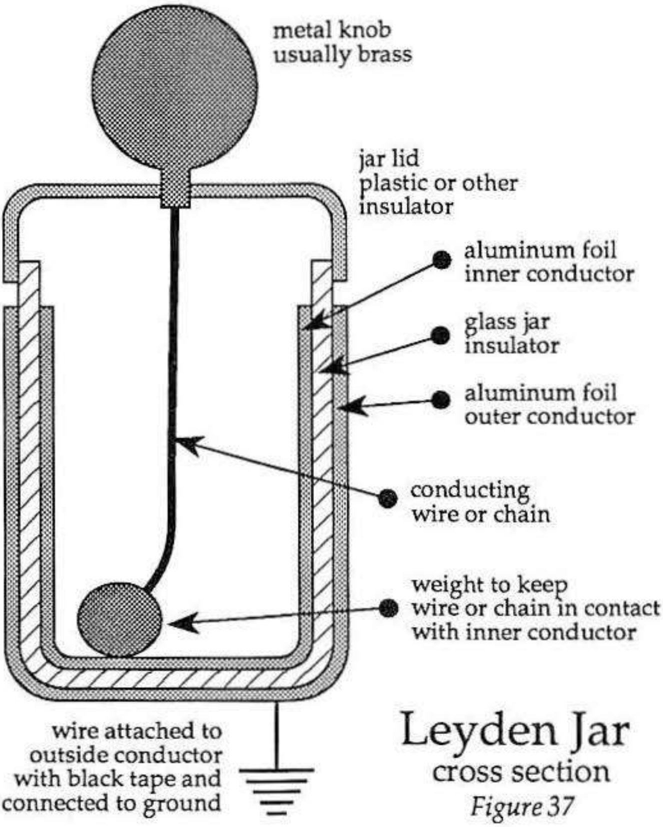
metal knob
usually brass
jar lid
plastic or other
insulator
aluminum foil
inner conductor
glass jar
insulator
aluminum foil
outer conductor
conducting
wire or chain
weight to keep
wire or chain in contact
with inner conductor
wire attached to
outside conductor
with black tape and
connected to ground

Leyden Jar
cross section
Figure 37

Electricity often seems to find some way to get into the jar without telling you. I was once shocked by a Leyden jar that was standing a few feet from a generator. It was not until I observed the generator and Leyden jar in a completely darkened room that I noticed a blue glow around a length of wire which had been acting as an aerial above the Leyden jar and had partially charged it although it was not even directly connected to the generator!

To make a discharge rod, simply take a length of heavy copper wire and bend it in a "U" shape. This "U" should cover the distance between the brass ball and the aluminum side of your Leyden jar. Attach an insulating handle to the center of the "U". Make sure this handle is long enough that a spark won't find its way along the handle to your hand. Holding it by the plastic handle, you can bridge the gap between the brass ball and side of the Leyden jar, and safely discharge it.

Discharging a large Leyden jar with this tool can be lots of fun since a large noisy spark can be safely produced. This provides a terrific demonstration for those who doubt the power of static electricity!

THE WATER CAPACITOR

Believe it or not, ordinary water and plastic bowls can be used to form an extraordinarily high dielectric, and make a good capacitor. Figure 38. A water capacitor is nothing more than two plastic bowls set inside one another each being partially filled with tap water. Any tap water will be somewhat conductive like the metal plates of the Leyden jar because of impurities. Dissolved salts in the water form positive and negative ions which move readily. The plastic wall of the inside bowl forms the dielectric while the water in each bowl forms the electrodes. A small capacitor of this type having 53 inches of useful surface area has a capacitance of approximately 100 pF.

This type of capacitor was supposed to have been a favorite of inventor, Nikola Tesla. Tesla, it is said, would wire together banks of

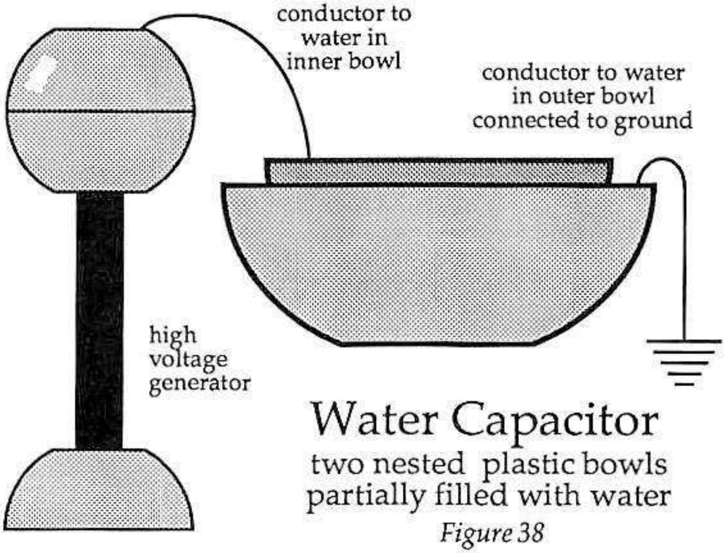
conductor to water in inner bowl
conductor to water in outer bowl connected to ground
high voltage generator

Water Capacitor
two nested plastic bowls partially filled with water
Figure 38

insulated metal trays which were filled partially with water. By connecting these water capacitors in parallel, Tesla could reach very large storage potentials with a minimum of effort.

Plate Capacitors

As you have undoubtedly gathered, all that is required for a simple capacitor to hold a charge are two conductive plates with an insulator (dielectric) in between. There are an infinite number of configurations to be experimented with. Figure 39 shows what is probably the simplest useful capacitor possible.

The largest plate capacitor I attempted during the writing of this text was one made from aluminum foil and my sliding glass door. As soon as I became interested in electrostatics, and making high voltage capacitors, I started seeing the makings of capacitors everywhere. Every day I looked at my sliding glass door, and every day it looked less like a door and more like a capacitor. Finally I broke down and

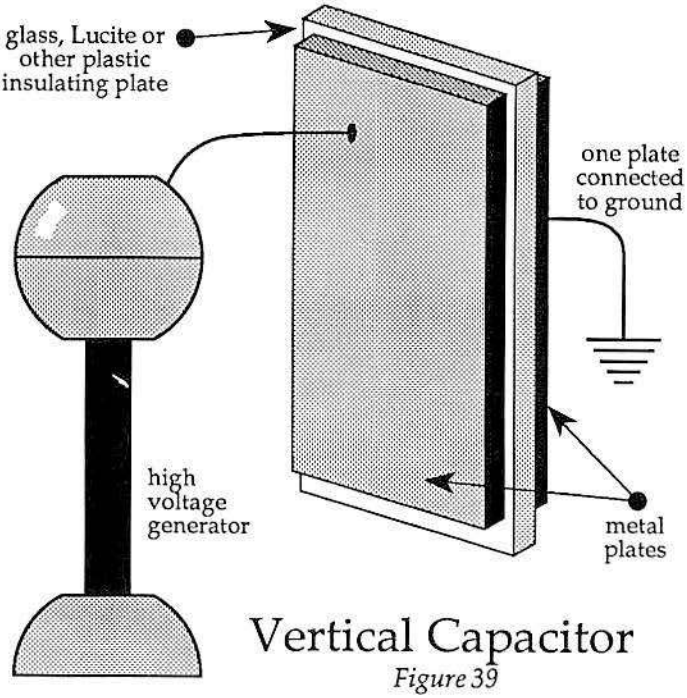
glass, Lucite or other plastic insulating plate
one plate connected to ground
high voltage generator
metal plates

Vertical Capacitor
Figure 39

bought a large roll of aluminum foil and some tape. After a thorough cleaning, I transformed my door into a giant capacitor, by taping sheets of foil to each side of the glass. But alas! The results were disappointing. Temperature differences between the inside and outside of the glass caused condensation and all the problems that moisture brings, and the large sharp edges of the foil which I attempted to insulate bled off charge rapidly into the air. The foil did odd things such as crinkling and making sounds when charged, but failed to yield an interesting spark or hold a charge for any length of time at all.

I'm sure you'll come up with as many ideas as I do. Most recently, a television special showed the huge metal domes on buildings in Moscow square giving me visions of tremendous high potentials. As I watched, the buildings kept looking more and more like high potential tops of Van de Graaff generators and less and less like buildings. I think this high voltage stuff is affecting my mind! I think it's a good thing that I don't live in Moscow.

Calculating Capacitance

Once you begin designing and building capacitors, you will undoubtedly want to know what charges they are capable of storing. Capacitance is fairly easy to calculate, although homemade capacitors will be much less than perfect and probably perform more poorly than calculations might indicate. The following equations will give you a good idea of what can be expected from your designs.

Air is a good dielectric, and is accounted for in calculations. If we have two metal plates standing parallel to one another with air between, the equation for finding capacitance in picofarad is:

$$C = .08854 \frac{A}{S}$$

Where

A is the area of one side of one plate in square centimeters

S is the distance between plates in centimeters.

The same equation can also be revised for inches:

$$C = .2249 \frac{A}{S}$$

Note that these equations remain fairly accurate as long as the smallest dimension of the plate is ten or more times "S". In other words, the area of the plate should be much, much larger than the distance between them.

Different materials have different insulating capabilities which are reflected in their dielectric constant. Since you will want to use materials other than air as a dielectric, you will need to look up the dielectric constant from standard tables. You'll find that Plexiglas provides about four times the dielectric strength of air. A capacitor built from two conducting plates would have four times the capacitance of an air capacitor.

Another example: We might calculate from the equation above that a capacitor with a desired plate size has a capacitance of 100 pF using air as the dielectric. Using the chart below we see that if dry paper were substituted for air, the capacitance will be raised to 200 pF.

Shown below are some typical dielectrics relative to air borrowed from Attwood's book *Electric and Magnetic Fields:*

Dielectric Constant	Material
1	Air
25	Ethyl alcohol
2.5	Transformer oil
4.5 - 7	Various types of glass
6.8-8.4	Plate Glass
8.0	Window Glass
4.1-6.1	Pyrex
5.4-5.8	Bakelite
2	Dry paper
2.3	Paraffin
4	Plexiglas
2.5	Pure rubber
4	Wood

Capacitors can be connected in parallel or in series. Parallel connection increases total capacity. The total capacity is the sum of the individual capacities.

$$C_{Total} = C_1 + C_2 + C_3 + \ldots$$

For example, three capacitors having capacity of 10, 20 and 30 pF each would act as one capacitor having 60 pF when connected in parallel. The breakdown voltage of the combined capacitors is only as great as the lowest breakdown voltage of any single capacitor.

Calculating the capacitance of capacitors connected in series is more complicated, being the reciprocal of the sum of the reciprocals of the individual capacitances.

In series you would figure:

$$C_{Total} = \frac{1}{\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \ldots}$$

For example, three capacitors have capacitance of of 2, 4 and 4 pF. When connected in series, the total capacitance would be 1 pF. Although capacitance has decreased, the maximum working voltage has increased. If each capacitor had a maximum working voltage of 10,000 volts, the series connection would have a maximum working limit of 30,000 volts.

Power Capacitors

You can build a very powerful capacitor by creating a stack of alternating sheets of dielectric material and metal plates. On a large sheet of dielectric place a metal plate, then another sheet of dielectric, another metal plate and so on. The plates must be absolutely clean and free of even the tiniest contaminant or the capacitor will not work. To complete the capacitor connect alternating metal plates to one lead and the remaining metal plates to another.

For your first attempt, build a power capacitor with 5 sheets of dielectric and 3 metal plates. Connect the top and bottom metal plate to one lead, and the middle metal plate to another lead.

Benjamin Franklin — The Turkey Killer

Benjamin Franklin, an incurable prankster, was an ardent experimenter with electrostatic machines. Amazing as it may seem, according to his biography, the major reason Ben Franklin tried his dangerous kite experiment was that he was tired of charging his Leyden jar by hand cranking his static electricity generator. He wanted to charge it with lightning which he was able to do without getting himself killed! (Note: this method of charging a Leyden jar is not recommended for those other than "founding fathers" and other experimenters who seem to have a supernatural gift for avoiding harm. In other words, don't you try it.)

Franklin's hand-cranked static machine was large enough when coupled to Leyden jars to produce a lethal voltage. In fact, Franklin often used his static machine to kill his Thanksgiving turkey. In his own words *"... to the amazement of many... A turkey is to be killed for our dinner by electrical shock, and roasted by the electric jack, before a fire kindled by the electric bottle..."*

Franklin's pranks were not without risk. Just a short time before his famous kite experiment Franklin accidentally took the full force of his turkey killing device. He wrote, *"Being about to kill a turkey by the shock from two large glass jars, containing as much electrified fire as forty common phials, I inadvertently took the whole through my own arms and body, by receiving the fire from the united top wires with one hand while the other held a chain connected with the outside of both jars. The company present (whose talking to me, and to one another, I suppose occasioned my inattention to what I was about) say that the flash was very great and the crack as loud as a pistol; yet, my senses being instantly gone, I neither saw the one nor heard the other; nor did I feel the stroke of my hand.... I then felt what I know not well how to describe: a universal blow throughout my body from head to foot, which seemed within as well as without; after which the first thing I took notice of was a violent quick shaking of my body, which gradually remitting, my sense gradually returned.... That part of my hand and fingers that held the chain was left white, as though the blood had been driven out, and remained so eight to ten minutes after, feeling like dead flesh; and I had a numbness in my arms and the back of my neck that continued till the next morning..."*

Franklin's account of his own mishap should certainly give thought to those who do not take the dangers of static electricity seriously!

ELECTROSTATIC EXPERIMENTS

The experiments here are divided into two categories. The first group describes electrostatic experiments that can be performed for an audience. In other words they're merely for entertainment.

The second section features experiments which possible useful applications of electrostatics which are not often discussed. Many of these "useful" experiments could be or are being developed into very real, marketable industries and should be particularly interesting to the serious experimenter.

Hair Raising Experiences

Using electrostatic voltages to make a person's hair stand on end is a classic experiment every audience should see. It can be done in a number of ways. Find a subject with hair that is light and clean. With some experience you will be able to recognize the best subjects immediately. Next, convince the subject that the ominous apparatus that they have just seen throwing four inch sparks will not harm them.

You may want to explain why the experiment is perfectly safe by describing the "skin effect" of high voltages. This effect refers to the strange fact that voltages of this order travel only on the outside or "skin" of the object they pass over and will not travel through the inside of a body where they could do any harm. For alternating currents this is a true skin effect, but for direct currents there is a similar effect that occurs probably because of the repulsion of like charges rather than due to a true skin effect.

The subject should stand within reach of the generator and should be well insulated from any ground. This is accomplished by having the subject stand on a large insulator such as an overturned plastic bucket, or, as a joke, a stack of college text books since they are usually very dry. Instruct the subject to place one hand on the high potential terminal, and then turn the generator on. Remember to remind a nervous subject not to remove his hand from the terminal until the experiment is completed since removing his hand could provide a

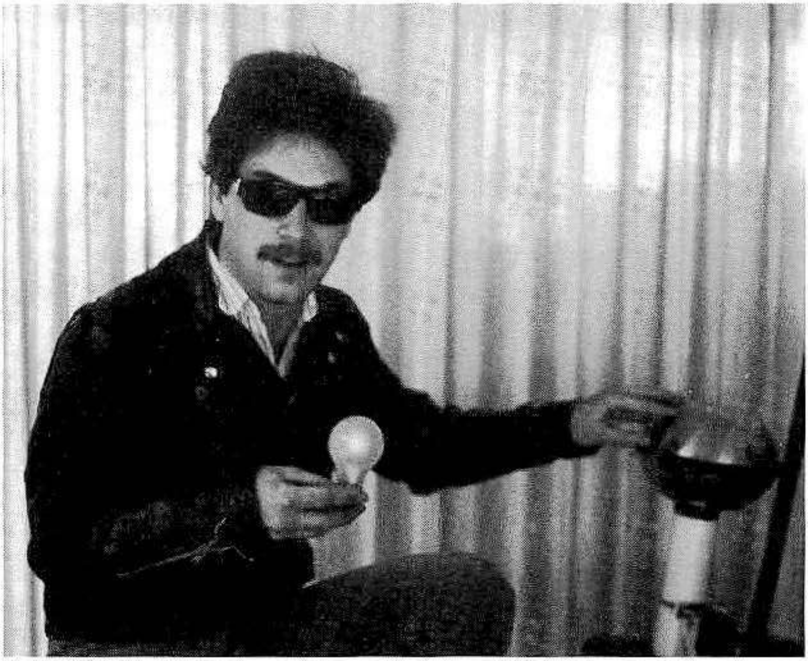

Figure 40 – The author's hair is usually flat and combed somewhat back. After a few minutes connected to a running Van de Graaff generator, and he's ready for a punk-rock concert!

shock. Depending on the output of your generator you may need to allow it to run for several minutes before your subjects hair will begin to rise. Having your subject shake his head occasionally with speed up the process.

During this whole procedure the subject will not feel the electricity passing over him other than perhaps a light tickling as his hairs stand on end. What is happening is that the subject's hair and whole body is taking on a charge from the generator. Since all of the hairs have the same charge whether it be positive or negative, and because like charges repel each other, every hair is repelled from every other hair and any nearby parts of the body. Warnings: **1) Don't use capacitors. 2) Hairspray is flammable. 3) Voltages over 500,000 are unsafe.**

Strangely enough, your generator will actually have a higher potential with someone is touching it than when it is operated alone since it can now use the person's body as a high potential collector. The success of this experiment depends largely on the subject's hair. Some women with short frizzy hair will have their hair stand out nearly straight from their head!

After the experiment is complete, have your subject remain standing on the insulator. This will allow the charge accumulated in their body to bleed off and will reduce the possibility of a large spark occurring when they eventually touch something that is grounded.

There are a few simpler although perhaps less dramatic ways to perform the same experiment. One is to make a wig out of fuzzy yarn and to place it over the generator's high potential terminal. You can place it on the spherical terminal of a Van de Graaff generator to simulate a person's head. Some Mr. Wizard impersonators even like to put a pair of plastic sun glasses under the wig to give the generator a little character. After the generator has run for a time, the hairs of the wig will stand on end.

A very quick way to demonstrate this effect is to simply take a few hairs from an old hair brush and place them on the generator's surface. They will immediately stand on end, and will often fly off the generator all together! Many different light weight materials can be used for these type of experiments. Bits of styrofoam work extraordinarily well.

The Magic Wand

As discussed, smooth, round objects tend to hold their charges while sharp pointed objects will allow electrons to fly off in a stream sometimes referred to as an "electric wind." The effects of this wind can easily be seen. Connect a short length of wire to your high potential terminal and hold the free end of the wire with a handle made of an insulating materials such as plastic. Point the end of the wire toward a lighted candle, and observe the flame. The flame will bend mysteriously away from the invisible "electric wind." In a well darkened room with ones eyes well adjusted the "wind" may actually be visible.

Another interesting experiment is to use the wand to chase a soap bubble. Because there are two forces acting on the bubble, the outcome of the experiment may be unpredictable. One force is the charge of the generator itself which draws the bubble toward the wand. The other force is the electric wind from the wand that pushes the bubble away. How the bubble will act is anyone's guess! It depends on which force is dominant.

A little later on we'll use this surprisingly powerful wind to create a small pinwheel motor.

Levitation

There are a number of ways that you can levitate lightweight materials using your generator. The easiest ways to demonstrate this is to simply spread very small shavings of styrofoam along the top of your generator. When the power is on, the pieces will take on the same charge as the surface of your generator and be repelled by it since like charges repel one another. The bits of styrofoam will levitate off into the air after the generator has been on a few seconds.

An entertaining experiment can be done by placing a number of tiny styrofoam balls in a short section of clear glass or plastic tubing about 2 1/2 inches in diameter and a few inches in length setting on your high potential terminal. Cover the top of the tubing with a metal plate, and connect it to ground. When your generator is turned on, the balls become charged and levitate to the top of the plastic tube. When they touch the metal plate at the top of the tube they lose their charge to ground and fall back to the generator again. The balls will oscillate between the ends of the tube until the generator is shut off. You should coat these balls with a conductive paint for a more complete charge transfer.

If you enjoy reproducing electrostatic experiments as they were originally done, you might want to try making your own old-fashioned pith balls instead of using styrofoam. Pith balls were used for these experiment before modern products like styrofoam were available. Pith balls are made by removing, shaping, and drying the insides of certain plants such as dandelions. Details can be found in old physics textbooks before 1900. I'll stick with good old styrofoam.

Place larger bits of styrofoam on your generator, and turn your generator on while touching the bits with your hand. The styrofoam will assume that you are ground and that your generator is a high voltage source. One side of the styrofoam will be charged one way, and the other side of the styrofoam the other. After removing your hand, a number of things can happen. The styrofoam may flip over and be stuck to your generator like a magnet. The styrofoam may stick to your hand like a magnet. Or you may find the styrofoam floating on end as the hair did in the last experiment.

An interesting trick can be performed by passing your hand over the styrofoam remaining on your generator terminal. As you do so, the electrical charges in the foam and in your hand will cause the styrofoam to jump about.

I have meant for some time to place a metal plate on a table underneath a table cloth and to energize that plate from a hidden electrostatic machine. I could then sit at the table and "mysteriously" move light weight objects such as bits of hair, sawdust, styrofoam or other similar materials around the table simply by passing my hands over them. I think this would make a very convincing illusion of mental telekinesis. Try it.

Lighting Without Wires

If you or your friends doubt the power of your generator, have on hand a supply of Ne2 neon bulbs which are available for less than a dollar at any electronics store. Remind those onlooking that it takes at least 70 volts to light a neon bulb. Hold one lead of the neon bulb in your hand and move the bulb toward your generator. Within inches, or even feet, the bulb will flicker or light. I have seen some experimenters hold the bulb in their mouth while reaching toward a generator with their hand. Seeing a 70 volt bulb light in a persons mouth with no wires attached is an impressive demonstration.

You may want to try large florescent bulbs as well. When my generator has been running I have seen flickers in the florescent lights in the ceiling of my home. These bulbs are many feet above the top of my generator yet respond well in a darkened room.

Perpetual Motion

Anything that people assume can't be done interests me. Many of the greatest inventions of all time would never have come into being if their inventors hadn't at one point been told that it couldn't be done. Currently, perpetual motion qualifies as such.

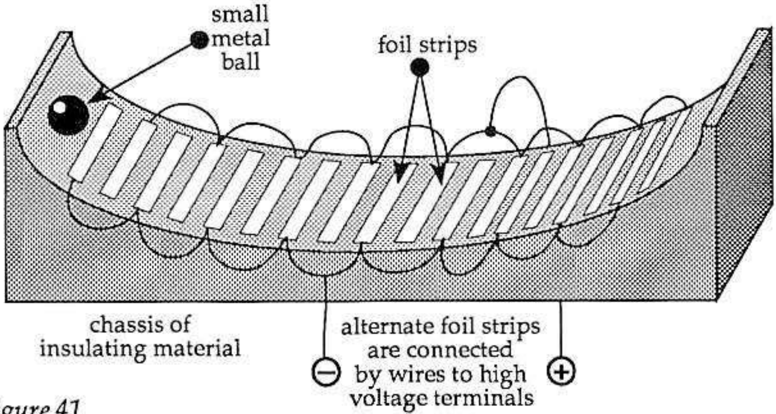
small metal ball
foil strips
chassis of insulating material
alternate foil strips are connected by wires to high voltage terminals
⊖
⊕

Figure 41

"Perpetual Motion"

The following device doesn't really qualify as perpetual motion, but it appears to be! Figure Pl shows the device. The idea is simple. A few light-weight conductive balls are released onto a U-shaped track on which the balls rock back and forth. It is normal to assume that the balls will slowly lose their energy and come to a stop in the center of the track. Instead, because of the electrodes in the track,the balls continue back and forth indefinitely and even pick up some speed along the way.

What has happened is that you have constructed simple electrostatic motor. As the balls pass over the foil strips they pick up the charge from that strip. Having the same charge as the strip, they are repelled by that strip and are simultaneously attracted to the next strip which has a dissimilar charge. Inertia will cause the balls to roll over the dissimilarly charged strip and on to the next. The electrostatic forces replace the energy lost in rolling friction. In this manner the balls will remain in motion for as long as power is supplied.

If you wish to really complete the illusion given by this device you will naturally want to hide the foil strips from view. To do this you could either use a conductive material other than foil that blends into the background, or, if your generator is powerful enough, you could probably bury the foil strips under the ramp material and still transfer a sufficient charge.

Trick Gun

We know that like charges repel one another. This has been demonstrated with the 'hair' experiments and others. What would happen if we introduced a charge into a liquid, and then shot that liquid out in a stream? We can find out with the use of a squirt gun. Purchase an inexpensive water gun that stores its water in the handle. Drill a small hole into the handle and twist a brass sheet metal screw into the hole. You may want to put a little Silicone glue around the top of the screw to make sure that this point will not leak when the gun is later filled with water. The brass screw provides electrical contact between your hand and the water.

Fill the gun with water, and pump out a few squirts. The water will come out in its customary straight stream probably covering the length of the room. Then stand on an insulator and charge yourself up as described earlier. When you squeeze the trigger, a stream of scattered droplets will be ejected only a few feet. What has happened is that all water droplets have a similar charge so each droplet is repelled by every other droplet. Other than being just an interesting way to get an unsuspecting friend wet, this mechanism can be applied anywhere that a liquid must be atomized.

I've wondered about making an electrostatic chamber on the carburetor of my car to see if better atomization of the fuel would result in better combustion. But first I need to check into the dangers of a spark occurring! I don't really want to blow myself up!

An Ion Motor

There are several different electrostatic motors that can be run by your generator, the simplest being an ion motor. You can make one by simply bending a length of light weight wire into a "Z". At the center of the "Z" shape, place a conductive bearing. Nothing fancy is required. In fact, just looping the center of the Z around an axle may work. Make sure however, that the Z will rotate very easily. A simple schematic of the ion motor can be seen in figure 42.

When power is applied to ion motor electrons stream off the pointed ends of the wire creating an electric wind. The force of the wind pushes the wire in the opposite direction. Again, for every action

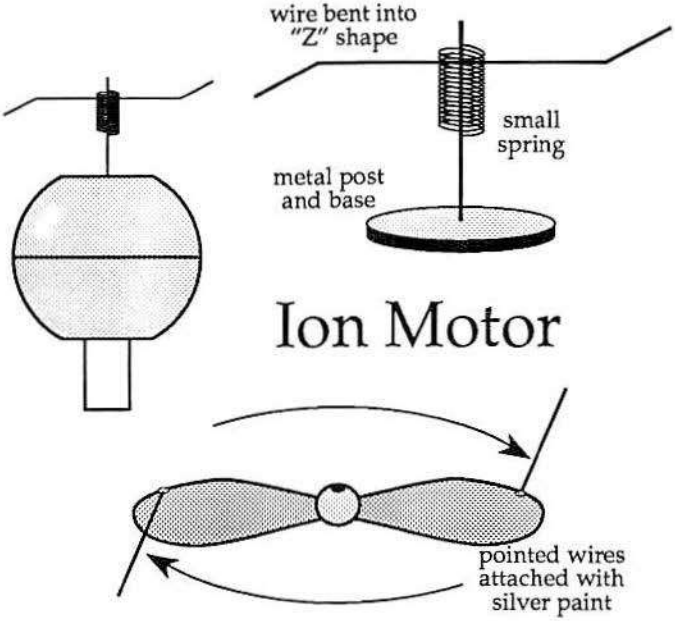
wire bent into
"Z" shape
small
spring
metal post
and base
Ion Motor
pointed wires
attached with
silver paint

Plastic Toy Airplane Propellor
Figure 42

there is an equal and opposite reaction. If your rotor is well balanced and the bearing friction is very low, your motor can and will achieve speeds of several thousands revolutions per minute! Of course, the torque generated by this motor will be very, very small. Nevertheless, it is fascinating to watch a motor propelled by a stream of electrons!

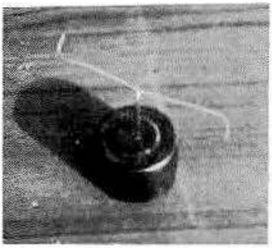
Figure 43 –Simple Ion Motor

Although the thrust of this motor is very small, I am certainly not the type of experimenter to let this hard-earned stream of electrons get away without doing some useful work along the way. I have attached one of my pinwheel motors to a tiny plastic airplane propeller as shown to blow a little air cool air my way. I think this is probably one of the few fans that I know of that must be connected to 100,000 volts or so to operate!

Another thought that I had for this fan was to drive it with a negative charge. Negative ions are thought to have many health benefits, and to cut down on pollutants in the air. Negative ions apparently attach themselves to specks of airborne dust and cause them to fall from the air. There are a number of negative ion generators being marketed for this purpose. I thought that a spinning propeller of this sort would be an absolutely outstanding way of both distributing and producing negative electrons in a room environment.

Ion motors also have serious applications in space travel. The Soviets have used ion motors to propel some of their long range probes. The problem with these motors is that the thrust is very small. However, ion motors offer at least two enormous advantages. First, they can be powered by a nuclear or solar electrical source which can provide power far longer than conventional chemical rockets. Second, an ion motor can provide higher velocity thrust than current chemical rockets because the nozzle velocity of a rocket is limited by the internal pressures that can be safely generated inside the rocket motor. An efficient ion motor can accelerate ions to speeds approaching the speed of light! Although the thrust is small, over time the space craft could itself approach the speed of light, at least, in theory.

A few tips for getting the best operation out of your ion motor—Make sure that your rotor is well balanced especially if you want your motor to be self-starting. Your rotor should spin easily for several seconds when tapped lightly by your finger. The more freely your rotor can spin, the faster it will revolve. The bearing used in a radiometer would make an excellent ion motor bearing.

Run your generator/motor in an area free of objects that might provide sources of ground. My first ion motor was well built and carefully balanced, yet it refused to spin. When I reached down to give it a tap to get it started, it refused to move. In fact it felt as though it were being held in place by a magnet. It was then that I realized that one of the arms of the rotor must have been attracted to an object with an opposite charge. That object turned out to be me! When I moved a short distance away, the ion motor sprang to life and immediately whirled up to several hundred RPM.

For this motor, voltage is more important than current. The motor will run and do useful work on currents under a microampere. Electrostatic motors have been run with currents of less than .001

microamperes, and have even from the earth's electrical charge itself.

Another interesting use that I have found for an ion motor is that of a primitive high voltage meter. The higher the voltage, the greater the thrust from the motor. This may be used to move a pointer such as in a panel meter or may be used simply to a relative indication of voltage.

I like to place a small ion motor on top of my Van de Graaff generator when I am charging capacitors. When the capacitors are charging, the motor will not move since the charge is being bled off from the generator to the capacitors. When the capacitors are fully charged, they stop drawing current from the generator. The terminal voltage rises, and the motor begins to spin. This is a very useful application of the ion motor since it is usually impossible to tell when a capacitor is fully charged any other way.

I am told that ion motors have been used in scientific toys such as those that can influence a balsa airplane in flight. Gustave LeBon's electrified rocket was a small aluminum foil rocket that could actually fly on electrostatic forces (Patent No. 2,018,585). I have not seen these devices, but they sound like something worth investigating.
Experimenting with "micro" sources of power can bring out the engineer in everyone.

An Ion Blower

The last experiment uses a stream of ions to propel a small rotor. Since in essence what we really have is a little jet engine, why not put our blower into a tube and see if we can pull a little air along with it. Figure 44 shows one possible configuration.

One end of a small, clear tube such of glass, Plexiglas, or other insulating material is fitted with a sharp metal point. A metal screen covers the other end. Electrons reach a high density at the metal point, break free into the air, and stream through the tube to the screen at the far end. As they move rapidly through the tube, they pull some of the air inside the tube with them. Obviously, a blower of this type has very low efficiency probably in the neighborhood of 1% under ideal conditions, Nevertheless, a puff of smoke at the intake side of the tube will easily be drawn through.

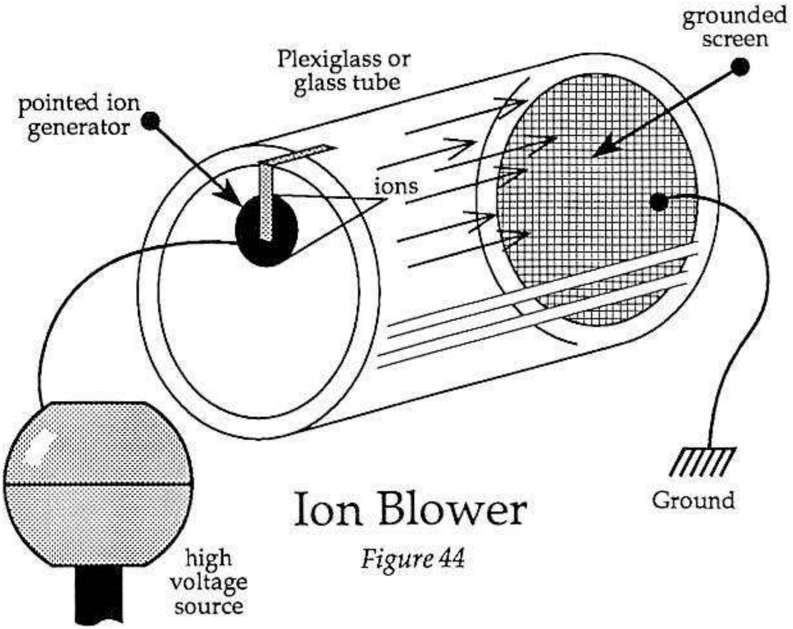

Ion Blower

Figure 44

Clearing The Air

Before moving on, here's another interesting smoke experiment you can try. Get a clear plastic or glass jar having no metal parts. Drive a pointed nail through the center of the lid. Blow a little smoke into the jar, and seal the jar with the lid. Next, connect the head of the nail to the output terminal of one of your high voltage generators.

As charge is applied to the smoke, it will begin to swirl. Soon the smoke will seem to disappear, although in reality it has been redistributed along the sides of the jar.

Variations of this effect are used in industry to place very fine films on surfaces. Similar electrostatic methods are being used to make sandpaper. When the tiny metallic abrasive particle are distributed along the surface of the sand paper they are given a charge that causes them to stand on end creating sandpaper with a better abrasive action.

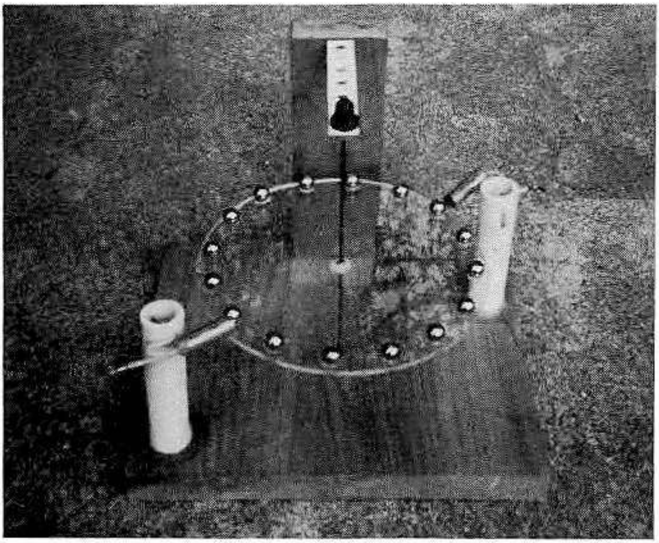

Figure 45 - Franklin Motor – Using large sling-shot balls and a simple Lucite disk, this little motor quickly speeds to over 600 rpm on less than 1 microampere of current!

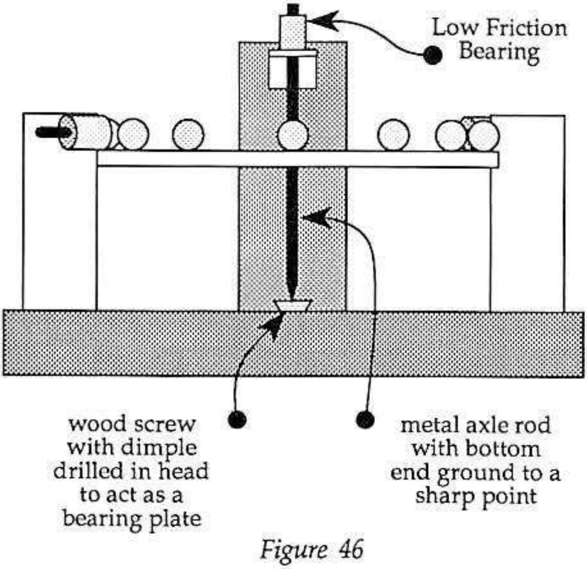
Low Friction Bearing
wood screw with dimple drilled in head to act as a bearing plate
metal axle rod with bottom end ground to a sharp point

Figure 46

Franklin Motor

Front View

Figure 47

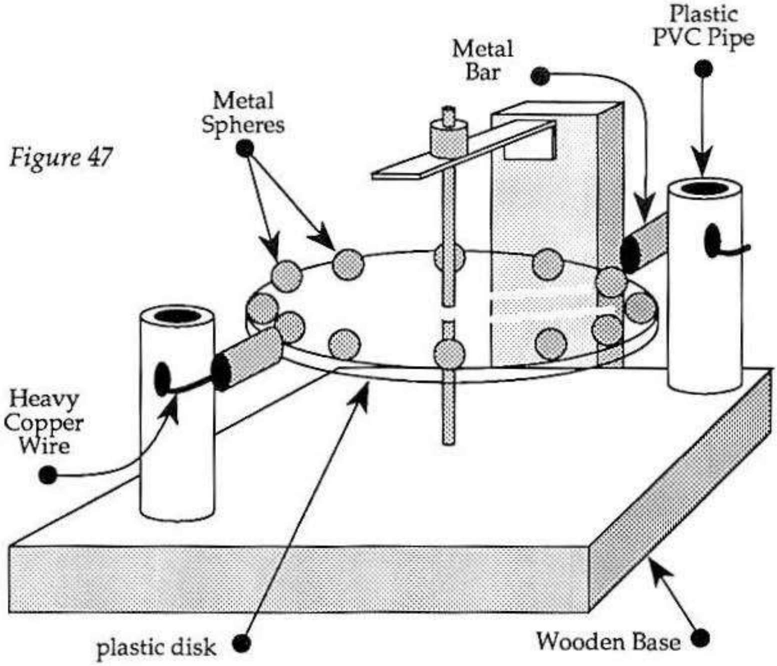

Franklin Motor

THE FRANKLIN ELECTROSTATIC MOTOR

The Franklin motor, named after its inventor Benjamin Franklin, is often said to be the first electric motor ever made. Its construction is simple, and, in fact, it is very similar in operation to the Dirod generator described earlier which can be run as a motor although inefficiently. Figure 47 shows the basic Franklin motor.

A series of lightweight spheres is glued in place along the outside of an insulating disk of Plexiglas, glass, etc. As these spheres rotate, they pass through the corona emanating from a highly charged high potential terminal at each side of the disk. One terminal is charged negatively and the other positively. As a sphere passes a terminal, it picks up a charge that will be attracted to other terminal causing the whole assembly to turn until the high voltage is removed.

One modification that you might like to experiment with is to make the Franklin motor self-starting. This can be accomplished by adding points on the high potential sources to create an electric wind that gives the spheres just enough of a nudge to start the rotor revolving.

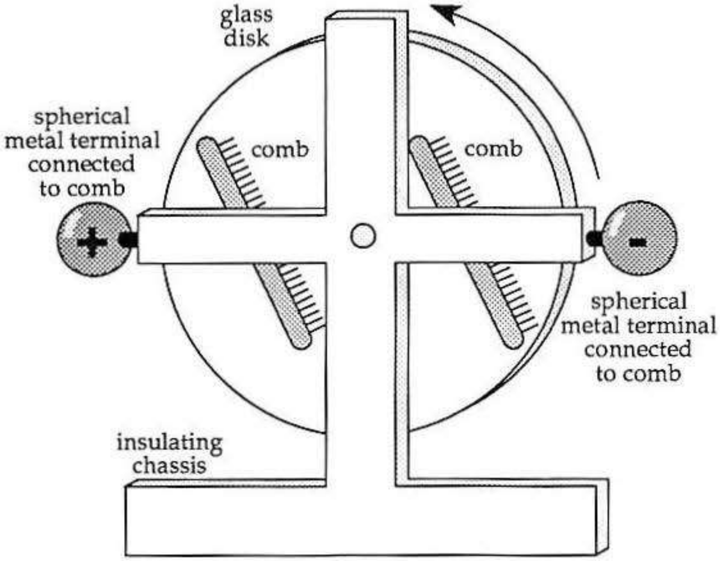
glass disk
spherical metal terminal connected to comb
comb
comb
+
-
spherical metal terminal connected to comb
insulating chassis

Corona Motor
schematic diagram
Figure 48

Poggendorff's Corona Motor

Franklin's motor works by charging a series of spheres as they pass by alternating high potentials. It is also possible to alternately charge an insulating surface as it passes near high potential terminals. Plexiglas, glass, and probably other materials will work. A drawing of Poggendorff's motor is shown in figure 48.

Here is how it works. Notice that there are two sets of brushes or combs with sharply pointed ends mounted at an angle. One brush sprays a negative charge onto the disk and the other a positive charge. As the disk continues to revolve, the charges on its surface find themselves attracted to the entire oppositely charged comb – not just the tiny spray points. It is important that the brushes be mounted at an angle, so that charges on the disk may be attracted to the comb of opposite polarity before they are replaced with an opposite charge as the disk rotates under the points of the comb. Understanding how a motor like this works is simple. Getting one balanced well enough to be self-starting and smoothly running takes quite a bit of patience. It is nonetheless an exciting and interesting project.

Atmospheric Electricity

The earth itself is a huge electrostatic generator. In the air at any time there are some 200 volts of electricity encircling you all of the time. The reason that we aren't being zapped all the time is simply that we are all at roughly the same potential. That is we all have the same level of charge, so, the electricity has no reason to move from one point to another.

Figure 49

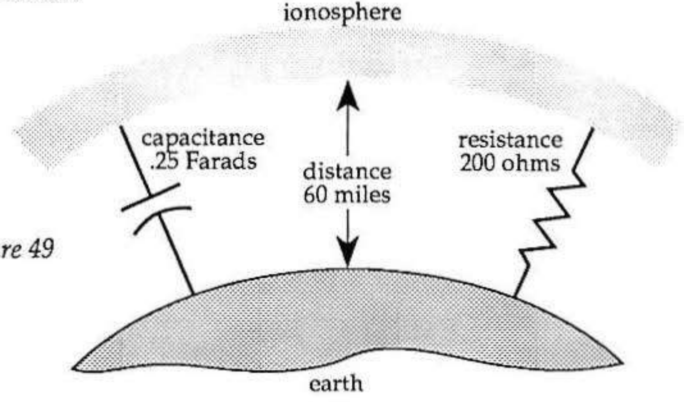

Earth's Atmosphere
electrical parameters

Between the earth and the electrically conductive ionosphere, both of which store electrical charge, exists an electrical potential. The same sort of buildup exists between clouds and the earth. When the potential builds to a level great enough, the air ionizes, loses its insulating capacity and becomes an excellent conductor. The result is a huge arc of static electricity we call lightning. As current flows between the cloud and the earth, the potential is reduced until the arc can no longer be sustained, and the lightning bolt is extinguished. It fascinates me to think that I live on one of two highly charged electrodes.

As an electrostatic generator the earth is fairly impressive. Between two and six thousand thunderstorms are in progress at any one time. Worldwide, a hundred bolts of lightning strike the earth every second. It is estimated that this gives the earth a high voltage potential of approximately 360,000 volts between ground and ionosphere with an average lightning bolt conducting 1800 amperes. The density of the electric field decreases as altitude increases. Near the earth's surface, however, the field is estimated to be approximately 100 volts per yard of height.

I say that the earth is fairly impressive because for all its size it is limited to these voltages by the fact that the atmosphere is relatively conductive having a resistance of approximately 200 Ohms. This is a relatively low resistance which quickly bleeds away charges.

Although it can take a significant amount of time depending on weather conditions, any conductive material insulated from ground can acquire charge from the surrounding air, and become charged to a high potential. Although the charge accumulated is usually negative in polarity, thunderstorms and other atmospheric conditions may create a positive potential with respect to ground. Radio antennas and flag poles insulated from ground have been known to throw hefty sparks and even start fires.

Plugging In To The Earth's Charge

If you're as curious as I am, you're asking yourself, "Can I tap into this charge?" The answer is yes! The first order of business is to measure and verify that a charge is present. If you are a skeptic like me, you've "gotta see that needle move" before you are going to believe there is something there!

The best antenna for capturing atmospheric charge is a metal structure, insulated from ground, pointing straight up into the air. The simplest, and probably easiest way to accomplish this is to simply attach a length of very light, bare, copper wire to a large helium balloon. This can be dangerous! Remember to avoid power lines and airports if you plan to launch such an energy-gathering balloon!

Having your bare copper wire come in contact with a high-tension power line is not a pleasant way to die! Thunderstorms are extremely dangerous. (Ben Franklin was lucky to have survived his experiment.)

The bottom of the wire should be attached to a high quality insulator well above ground. Attach one lead of a extremely sensitive meter to ground. Occasionally attach the other lead to the vertical metal conductor. Be careful! Do not take for granted that the potential will be harmless! If left unchecked it could easily reach a level that will bite you. You may find that it can take an hour or more to build up a significant charge, however.

At one time I tried using a kite flown with bare copper wire but had mixed results. I experienced significant difficulty with the wire kinking as I let it out and reeled it in, in response to changing winds.

Another time I hooked a diode into the line and got surprising voltages, but I had been fooled. The diode and coil in the meter had made a primitive crystal radio. I was extracting power from the atmosphere that had been placed there by radio broadcasting stations. Although this is not static electricity, it is nonetheless interesting because it is a demonstration of wireless transmission of power that Tesla was trying to perfect.

An antenna of this type has been used quite adequately to run a corona motor from the natural potential that exists between the earth and its atmosphere! A motor of this type must be extremely well built having great sensitivity and very low friction bearings. I have seen such a motor run, but I have not run one of my motors from this charge.

Because of weather, it can take a long time to build the potential up to a level capable of doing useful work such as moving a sensitive meter or a small corona motor. Several methods have been devised to speed up the charging process. Early experiments devised by Lord Kelvin used an insulated metal tank filled with water suspended above the ground constructed in such a way so as to allow a steady stream of water droplets to reach the ground. Appropriately, he called the setup a "water dropper." He reasoned that the surface of each water drop would be an extension of the surface of the water in the tank at the moment it dropped away. As it did, the drop would carry charges away with it. This continues until equilibrium was reached between the tank and the air.

Using this same idea, charging time can also be reduced by spraying a fine mist of water along a bare copper wire a hundred feet or more in length which is insulated from and suspended above ground. The falling water droplets can help to charge this wire to a measurable level in less than a minute's time.

Another method of speeding up the charging process is to increase the number of ions in the vicinity of the wire itself. A flame applied to the wire will increase the number of ions in the area of the flame and can bring a wire up to the same potential as the air around it in a matter of seconds. To accomplish this easily, wrap the wire with a flammable substance such as string or paper soaked in lead nitrate, and ignite it. It is also possible to get good results by simply waving a torch lightly across the wire.

In Conclusion

I hope that you will enjoy experimenting with electrostatics as much as I have! I find few subjects more fascinating than high voltage and its generation. As a technically minded person, you will undoubtedly see many improvements that can be made in my designs. I know that this is true because at times it seemed I might never complete this text because there was always one more improvement to make and one more experiment to try. No doubt, you will conceive of dozens of new ways to apply the basic principles of electrostatics. Like all experimentation, every change, every test, every improvement will suggest many more questions to be answered and new improvements to be made. You'll probably devise many new inventions!

Electrostatics is also a subject that everyone is absolutely fascinated by – even those who can't change flashlight batteries,. When I was experimenting with different friction generator materials, I purchased quite a number of little swatches of material from a local fabric store. The young ladies working in the store demanded to know what the fabric was for, and, were so fascinated by the idea of generating a spark that I emerged from the store piled high with swatches of material which they insisted be tried on the generator.

My own mother, who despises anything technical, shocked me to no end by demanding that I explain to her the workings of the atom immediately after seeing the classic Van De Graaff in operation. I do not think anyone can watch a powerful Van De Graaff generator at work and not be curious about the forces at work. (And you and I will always want control of those forces!)

So dig in, get to work! Mysteries await!

SELECTED BIBLIOGRAPHY

B.I. Bleaney & B. Bleaney, *Electricity and Magnetism.* 1957 p.762

Tricker, R.A.R., *The Contributions of Faraday & Maxwell to Electrical Science.* 1966

Moore, Arthur Dearth, *Electrostatics.* 1968 p.240

Simon and Schuster, *The Way Things Work.* 1967 p.590

R.A. Ford, *Tesla Coil Secrets.* 1985 p.74

Lodge, Oliver, Sir, *The Nature and Properties of Negative Electricity.* 1906

Adams, Charles K., *Nature's Electricity.* 1987 p.147

Sootin, Harry, *Experiments With Static Electricity.* 1969 p.85

Van Doren, Carl, *Benjamin Franklin.* 1938 p.845

Ammen, C. W., *The Electroplater's Handbook,* 1986, p. 213

ADDENDUM

Electroplating High Potential Terminals

As you have certainly realized by now, a tremendous amount of your generator's efficiency rests in the quality of your high potential terminal and in your ability to effectively store the charges produced. I think you will be very happy with the performance of any of the classic generators and storage devices described earlier. However, for those of you seeking to go beyond these boundaries of performance, I'd like to let you in on new area of experimentation: electroplating.

Electroplating is simply a process wherein metal salts such as copper sulfate or nickel ammonium sulfate dissolved in an acidic aqueous solution can be deposited as elemental metal on conductor by applying an electrical current. This means simply that any conductive object can be covered through the process of electrodeposition with a layer of metal!

Since electrostatic charges can be distributed quite nicely on even very thin conductive surfaces, this is a perfect system for making a flawless Leyden jar or a huge high potential terminal. I am currently working on creating very large, near-perfect high potential terminals by plating large plastic globes sold in many lighting stores for under $15.00!

Anything can be plated – plastic, glass, pvc, etc. The only trick is in making their surface conductive initially so that a current can be applied.

The basic procedure for producing a perfect Leyden jar is as follows.

Begin by making the glass jar extremely clean. No fingerprints, water, or oil of any kind can remain. Use masking tape to cover the top 20% of the jar to prevent plating in this area and to preserve a well-proportioned jar. Paint the exposed glass with a resin such as white shellac thinned with denatured alcohol, if necessary. I've even used a spray lacquer.

From here there are several routes that you can take. The easiest method is to take a finely powered metal such as iron or aluminum (something around 400 mesh) and sprinkle it on the sticky, wet shellac. Since these finely powdered metals will float in the air much like talcum powder, I usually put the metal powder and jar in a box and shake the two around around. Remember! These metals SHOULD NOT BE INHALED! Wear a safety mask, and if working with heavy metals, wearing protective clothing is a good idea. The metal powder should create a beautiful, even, light coating over the painted surface of the jar. Allow the paint to fully dry, and repeat the procedure on the inside of the jar. This is a little trickier, but not impossible.

At this point you can actually stop, and use the jar quite successfully. The tiny sandpaper edges of this powder metal coating will bleed off a stored charge when the jar is not in use. This makes a good natural safety feature.

The only drawback to the powdered-metal approach is that if you choose to copper plate it, the copper plating will be just as bumpy as the powdered metal surface itself. As a result, this method will always allow some leakage.

Another simple method of making the jar conductive is to cover the shellac with powdered graphite. This can be rubbed smoothly onto the surface when the shellac is partially dry and will provide a smoother surface which in turn will produce smoother plating. On the other hand, graphite surfaces often require considerably longer plating times and the process is less reliable.

Although both of the above methods work well and have merits, the very best way to make the jar or any non-conductor conductive is to allow the shellac to dry completely, then paint or dip it in a solution of

1/2	ounce	silver nitrate
4	ounces	distilled water
6	ounces	alcohol

While this is still wet, suspend the jar inside a larger jar or other acid-proof non-metallic vessel, and place a few lumps of potassium sulfite or iron pyrite in the bottom of the larger jar. Cover this with a solution made up of one part sulfuric acid and eight parts distilled water, and close the lid. You now have a jar covered with silver nitrate solution hanging above the potassium sulfite-acid solution in a sealed vessel. This will produce hydrogen sulfide fumes that will convert the nitrate to a sulfide which is a conductive coating for you jar. Within a

A glass jar for the Leyden jar capacitor is selected and thoroughly cleaned.

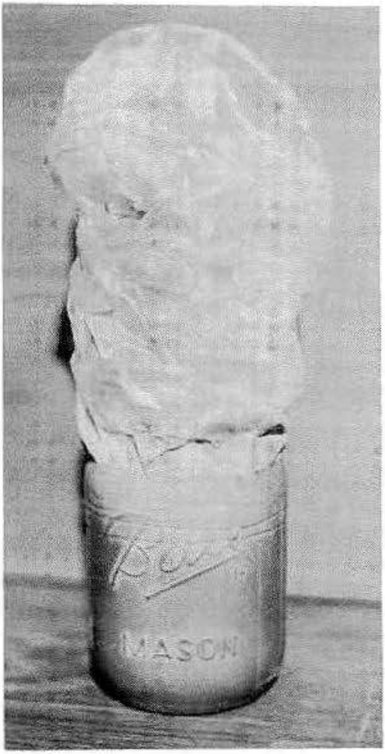

The jar is masked off and coated with conductive graphite.

few minutes the chemical reaction will stop, and your jar will be ready to be plated.

Now that the outside of the jar has been made conductive, it can easily be plated with copper. In an appropriate acid-proof vessel, make a solution of

25	ounces	copper sulfate
6	ounces	sulfuric acid
1	gallon	distilled water

Use warm water when mixing to dissolve the copper sulfate more quickly. Do not allow this solution to touch your skin or clothes.

After a few hours, the deposit of copper is easily seen.

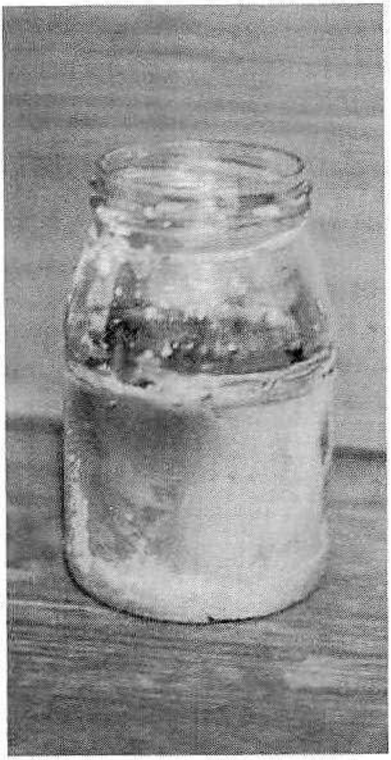

The plated jar is ready to be cleaned up and put into use.

Attach a wire from the negative lead of a 2 to 5 volt battery or transformer to the conductive part of your jar. A rubber band works well for this. Immerse the portion of your jar to be plated in the copper sulfate solution. Finally, attach a wire from the positive lead of your power source to a copper electrode having a large surface area such as a thin copper sheet. Set this electrode in your solution.

All that you have to do now is watch the jar as the plating takes place. Be sure you keep the solution warm, or the plating will proceed very slowly. I like a thick copper coat and usually allow my plating to take place overnight. However in a fairly short period of time you should see a reddish copper coating forming on the outside of your jar! Simply repeat this process for the inside of your jar.

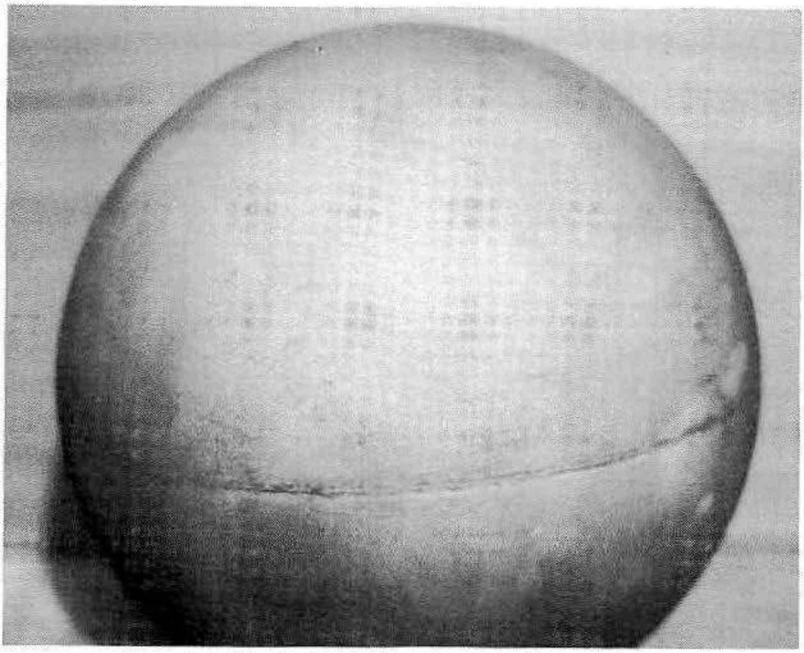

A twelve inch plastic sphere has been coated with graphite and is ready to be plated.

Very large spheres and other shapes can also be plated by scaling up this process. I have seen many inexpensive large glass and plastic globes since I began experimenting with this process that would lend themselves to this technique. For more information, refer to books on electroplating and electroforming that can give you advanced details on plating superior finishes.

As you can see, simple electroplating can make possible a vast number of new shapes and designs for high voltage terminals!

RSGB RADIO SOCIETY OF GREAT BRITAIN

D-i-Y RADIO
Yearling

RADIO & ELECTRONICS COOKBOOK

Newnes

Contents

Preface ix

1. A medium-wave receiver 1
2. An audio-frequency amplifier 4
3. A medium-wave receiver using a ferrite-rod aerial 9
4. A simple electronic organ 12
5. Experiments with the NE555 timer 17
6. A simple metronome 21
7. What is a resistor? 24
8. Waves – Part 1 27
9. A beat-frequency oscillator 31
10. What is a capacitor? 34
11. Waves – Part 2 38
12. An LED flasher 41
13. Waves – Part 3 44
14. Choosing a switch 46
15. An aerial tuning unit for a receiver 49
16. A simple 2 m receiver preamplifier 52
17. Receiving aerials for amateur radio 54
18. The Colt 80 m receiver – Part 1 58
19. A crystal radio receiver 62
20. The varactor (or varicap) diode 64
21. A portable radio for medium waves 65
22. The Colt 80 m receiver – Part 2 70

23. A simple transistor tester 73
24. An introduction to transmitters 77
25. The Colt 80 m receiver – Part 3 81
26. A two-way Morse practice system 88
27. The Colt 80 m receiver – Part 4 91
28. A simple crystal set 95
29. A crystal calibrator 100
30. A simple short-wave receiver – Part 1 104
31. A fruit-powered medium-wave radio 106
32. A capacitance bridge 109
33. A simple short-wave receiver – Part 2 113
34. A basic continuity tester 117
35. A charger for NiCad batteries 119
36. An 80 metre crystal-controlled CW transmitter 123
37. A solar-powered MW radio 129
38. A receiver for the 7 MHz amateur band 133
39. Diodes for protection 137
40. An RF signal probe 140
41. An RF changeover circuit 142
42. A low-light indicator 146
43. A J-pole aerial for 50 MHz 149
44. Measuring light intensity – the photometer 153
45. A 70 cm Quad loop aerial 156
46. A UHF field strength meter 160
47. Christmas tree LEDs 162
48. An audio signal injector 166
49. Standing waves 168
50. A standing-wave indicator for HF 170
51. A moisture meter 174
52. Simple aerials 177
53. A breadboard 80 cm CW transmitter 182

54. A 7-element low-pass filter for transmitters 186
55. Radio-frequency mixing explained 189
56. A voltage monitor for a 12 V power supply 192
57. A 1750 Hz toneburst for repeater access 196
58. A circuit for flashing LEDs 201
59. Digital logic circuits 205
60. A resistive SWR indicator 210
61. An audio filter for CW 213
62. An electronic die 215
63. The absorption wavemeter 222
64. An HF absorption wavemeter 224
65. A vertical aerial for 70 cm 228
66. A UHF corner reflector aerial 230
67. A switched dummy load 234
68. A simple Morse oscillator 238
69. A bipolar transistor tester 240
70. The ‘Yearling’ 20 m receiver 245
71. Adding the 80 metre band to the Yearling receiver 251
72. How the Yearling works 255
73. A field strength meter 258
74. Preselector for a short-wave receiver 261
75. An audible continuity tester 265
76. An experimental 70 cm rhombic aerial 268
77. Water level alarm 272
78. A delta loop for 20 metres 275
79. A simple desk microphone 279
80. Morse oscillator 284
81. A simple 6 m beam 287
82. An integrated circuit amplifier 291
83. A novice ATU 293
84. CW QRP transmitter for 80 metres 297

85. An audio booster for your hand-held 303
86. A grid dip oscillator 306
87. A CW transmitter for 160 to 20 metres 312
88. Matching the end-fed random-wire aerial 315

1 A medium-wave receiver

Introduction

Let us start off with something that is really quite simple and yet is capable of producing a sense of real satisfaction when complete – a real medium-wave (MW) radio receiver! It proves that receivers *can* be simple and, at the same time, be useful and enjoyable to make. To minimise the confusion to absolute beginners, *no* circuit diagram is given, only the constructional details. The circuits will come later, when you have become accustomed to the building process. In the true amateur spirit of ingenuity and inventiveness, the circuit is built on a terminal strip, the coil is wound on a toilet roll tube (as amateur MW coils have been for 100 years!), and the receiver is mounted on a piece of wood.

Putting it together

Start by mounting the components on the terminal strip as shown in **Figure 1**, carefully checking the position and value of each one. The three capacitors are all the same, and so present no problem. They (and the resistors) may be connected either way round, unlike the two semiconductors (see later). The resistors are coded by means of coloured bands. You can refer to Chapter 7 if you have difficulty remembering the colours and their values.

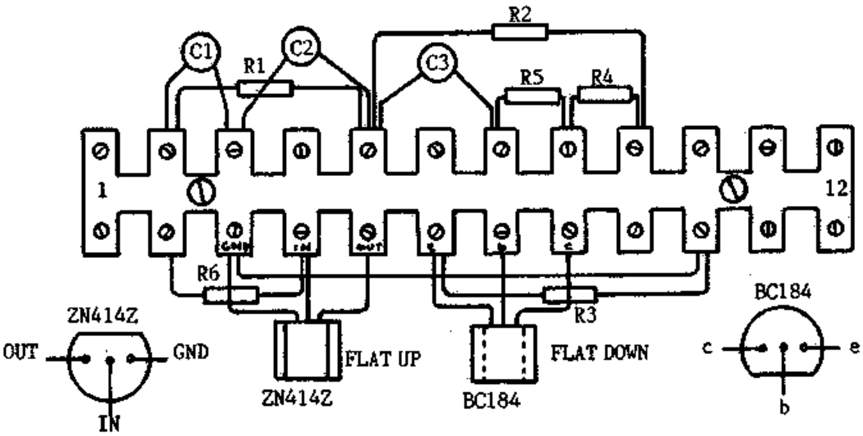

Figure 1 Terminal strip – position of components

1. Brown, Black, Yellow	100 000 ohms	(R1, R5, R6)
2. Green, Blue, Brown	560 ohms	(R2)
3. Red, Violet, Brown	270 ohms	(R3)
4. Brown, Black, Orange	10 000 ohms	(R4)

The integrated circuit (the ZN414Z) and the transistor (the BC184) *must* be connected correctly. Check Figure 1 carefully before fitting each device.

Now wind the coil. Most tubes are about 42 mm diameter and 110 mm long. Don't worry if your tube is slightly different; it shouldn't matter. Make two holes, about 3 mm apart, about 40 mm from one end, as shown in **Figure 2**. Loop your enamelled wire into one hole and out of the other, and draw about 100 mm through; loop this 100 mm through again, thus anchoring the wire firmly. Now wind on 80 turns, keeping the wire tight and the turns close together but not overlapping. After your 80th turn, make another two holes and anchor the wire in the same way as before. Again, leave about 100 mm free after anchoring. Using another piece of enamelled wire (with 100 mm ends as before), loop one end through the same two holes which contain the end anchor of the last winding, wind two turns and anchor the end of this short winding using another pair of holes. Figure 2 shows the layout.

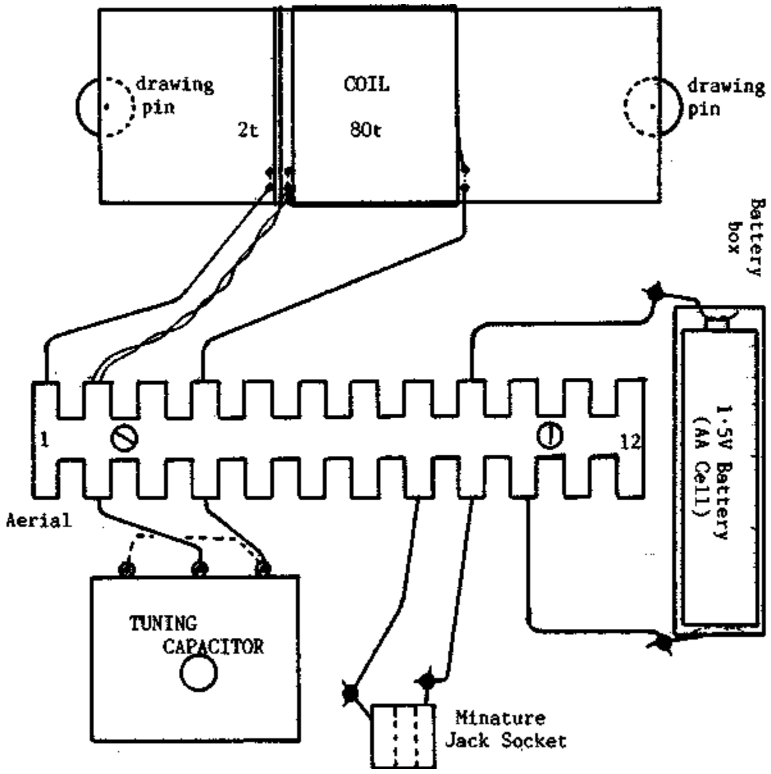

Figure 2 The layout of the parts on the wooden base

With some glass paper, remove the enamel from the ends of both pieces of wire which go through the same holes (i.e. the bottom of the large coil and the top of the small coil), then twist these bare ends together. Remove the enamel from the remaining ends of the coil. The coil is now finished!

The baseboard can be any piece of wood about 150 mm square. Fix the coil near the back edge using drawing pins and connect the wires from the coils to the terminal strip as shown in Figure 2. Using short pieces of PVC-insulated wire (and with assistance if you have never soldered before), solder one piece across the two outer tags of the variable capacitor, shown by the dotted line in Figure 2, and then two longer pieces to the centre tag and one outside tag. Connect these to the terminal strip. Then solder two more insulated wires on to the jack socket (into which you will plug your crystal earpiece), the other ends going to the terminal strip. The last two wires (one must be red) need to be soldered on to the battery box, their other ends going to the terminal strip also. Make sure the red wire goes to the *positive* terminal on the battery, and is connected to terminal 9. The other connection to the battery goes to terminal 10.

Attach the terminal strip to the baseboard with small screws or double-sided sticky tape. The other parts can be mounted the same way.

Listening is done ideally with the recommended crystal earpiece. Don't be tempted to use your Walkman earpieces; they are not the same and will not perform anything like as well. The receiver should work without an extra aerial, but one can be attached to terminal 1 if necessary. A long piece of wire mounted as high as possible is ideal. The *Audio-frequency Amplifier* project will enable you to use a loudspeaker with your receiver, using the signal from the jack socket. No circuit modifications will be needed!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R5, R6	10 kilohms (kΩ)
R2	560 ohms
R3	270 ohms
R4	10 kilohms (kΩ)

Capacitors

C1, C2, C3	100 nanofarads (nF)
	500 picofarads (pF)

Semiconductors

ZN414Z, BC184

Additional items
- 12-way 2 A terminal strip
- 22 metres of 28 SWG enamelled copper wire
- A few short pieces of coloured PVC-insulated wire
- Crystal earpiece
- 3.5 mm jack socket
- 1.5 V AA-size battery and box
- Toilet roll tube
- Double-sided sticky tape or selection of screws

Tools required

Small screwdriver, soldering iron.

2 An audio-frequency amplifier

Introduction

This simple amplifier can be built by anyone who is able to solder reasonably well. It doesn't require any setting up and, provided our instructions are followed exactly, will work very well. The circuit diagram is included for the benefit of our more advanced readers, but it is not needed in the construction process. Please practise your soldering before you start, and don't use a printed circuit board (PCB) until you are confident that your soldering is up to scratch.

The amplifier can be used with other projects; it will provide plenty of sound from the MW Radio or from the Morse Sounder projects. It will usually be built into other pieces of equipment, so a box is not supplied with the kit. There is no reason why it shouldn't be put into a box and used as a general-purpose amplifier to help test other projects.

The components

Before you start, you should check that you have *all* the components to hand. A list and some helpful hints are given below.

1. PCB. The plain side is the *component side* and the soldered side is the *track side*. **Figure 1** shows the track side full size. Make the PCB from the pattern given in Figure 1. Otherwise, build the circuit on a matrix board.
2. Three resistors. Locate the gold or silver band around the resistor, and turn the resistor until this band is to the right. There are three coloured bands at the left-hand end of the resistor. Find the resistor whose colours are YELLOW, VIOLET, RED, and look at the resistor colour code chart which you will find in Chapter 7. From this, you will see that YELLOW indicates the value 4, VIOLET the value 7, and RED the value 2. The first two colours represent real numbers, and the last value is the number of zeros (noughts) which go *after* the two numbers. So, the value is 47 with two zeros, i.e. 4700 ohms. In this way, the resistor coloured BROWN, GREY, BROWN has a value of 180 ohms, and the last one, BROWN, RED, GREEN, has a value of 1 200 000 ohms. The ohm (often written as the Greek letter omega (Ω)) is the unit of resistance. If you do not yet feel confident in identifying resistors by their colours, use the Resistor Colour Codes given in Chapter 7.
3. Four capacitors. The two small 'beads' are tantalum capacitors and will be marked 4.7 μF or 4μ7, with a '+' above one lead. A tubular capacitor with wires coming from each end should be marked 220 μF, with one end marked '+' or '–'. This is called an *axial* capacitor because the wires lie on the axis of the cylinder. This is in contrast to the final capacitor, where both wires emerge from the same end. This is a *radial* capacitor, and will be marked 47 μF. Again, one lead will be marked '+' or '–'. Capacitors marked like this are said to be *polarised*, and it is vital that these are placed on the PCB the right way round, so take notice of those signs!
4. Two diodes. These are tiny glass cylinders with a band around one end, and may be marked 1N4148; this is their type number. Like polarised capacitors, they *must* be put on the PCB the correct way round!

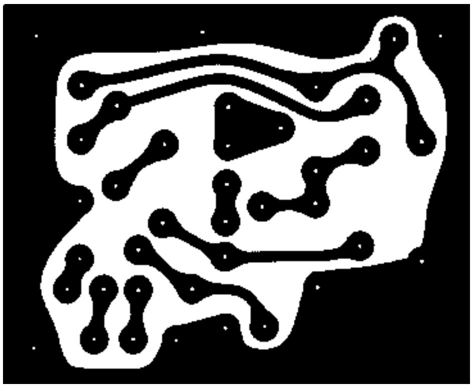

Figure 1 The toil pattern of the PCB – looking from the track side

5. Three transistors. One should be a BC548 (or a BC182), the other two should be BC558 (or BC212).
6. One volume control with internal switch.
7. One loudspeaker. This is quite fragile – don't let anything press against the cone.
8. One PP3 battery clip with red and black leads.

Putting it together

Lay the PCB on a flat, clean surface with the track side downwards. It is always useful to compare the layout with the circuit diagram, given here in **Figure 3**. Although you can't see it, the D-i-Y Radio sign should be at the top. Compare the hole positions with those shown in **Figure 2**. Bend the resistor wires at right angles to their bodies so that they fit cleanly into the holes in the PCB. Push each resistor towards the board so that it lies flat on the board. Then supporting each one, turn the board over and splay out the wires *just* enough to prevent the resistor falling out. Then, solder each wire to its pad on the PCB, and cut off the excess wire. When you have more confidence, you can cut of the excess wire *before* soldering; it often makes a tidier joint.

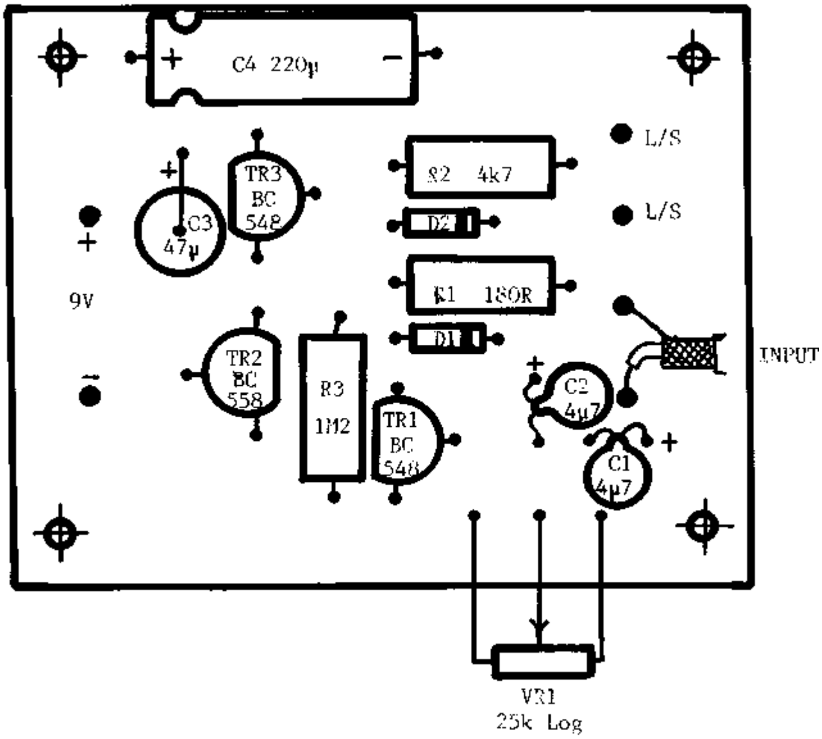

Figure 2 Positions of the components on the printed circuit board (PCB)

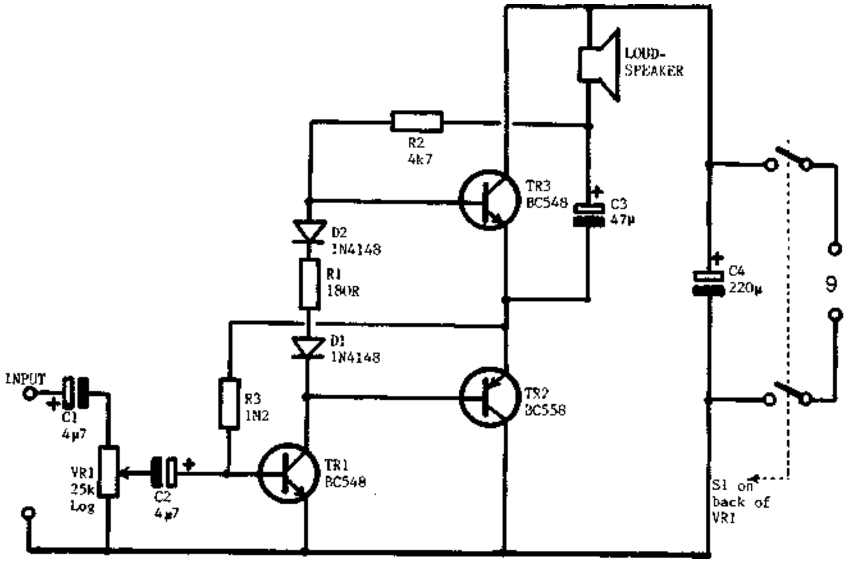
LOUD-
SPEAKER
R2
4k7
TR3
BC548
C3
47µ
D2
1N4148
R1
180R
D1
1N4148
C4
220µ
+
9 V
−
INPUT
+
C1
4µ7
R3
1N2
TR2
BC558
VR1
25k
Log
C2
4µ7
+
TR1
BC548
S1 on
back of
VR1

Figure 3 The amplifier's circuit diagram

Now fit the four capacitors. Each must be connected the right way round, so look at each component, match it up with the diagram of Figure 2, bend its wires carefully and repeat the soldering process you performed with the resistors, making sure that the components are close to the board and not up on stilts! Fit the two diodes the correct way round, and solder then as quickly as you can – they don't like to be fried!

Mount the transistors about 5 mm above the PCB. Make sure the correct transistors are in the correct places, and that the flats on the bodies match up with those shown in Figure 2.

Mount the volume control so that the spindle comes out from the front of the board. Use a piece of red insulated wire to the pad marked + on the PCB, and a black piece to the pad marked –, and solder these to the tags on the back of the control, as shown in **Figure 4**. Connect the two leads from the battery clip to the other tags on the switch; Figure 4 will help you. Finally, use two pieces of insulated wire about 100 mm long, twisted together, to connect the loudspeaker to the PCB.

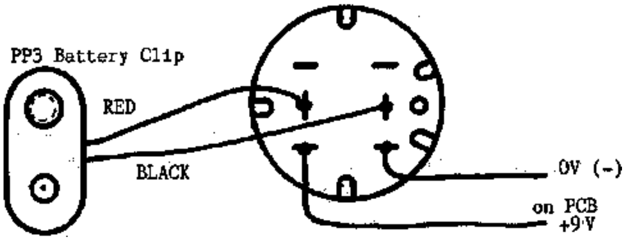
PP3 Battery Clip
RED
BLACK
0V (-)
on PCB
+9V

Figure 4 Connections to switch on back of VR1

Box clever!

If you wish to put the amplifier into a box, there is no problem; almost any box that is big enough will do. All that is needed is one hole big enough to accept the bush of the volume control; the PCB will be supported by the volume control. The prototype was not fitted into a box, but mounted on an odd piece of aluminium, bent into an L-shape and screwed on to a wooden base. The loudspeaker was mounted on the aluminium panel by two small pieces of aluminium with 3 mm holes drilled in them, which acted as clips around the edge of the speaker. Drill a few holes in the panel in the position of the speaker to let the sound get out!

Your input signal can be connected to the amplifier with two short pieces of wire, but if the connection needs to be long, use screened cable, with the braid connected as shown in Figure 2.

If you decide to use a different loudspeaker, make sure that its impedance (the resistance value marked on the back of the magnet) is at least 35 ohms. Anything lower may damage TR2 and TR3, and will certainly run down your battery very quickly. You will be surprised at the uses you can find for this little amplifier!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	180 ohms (Ω)
R2	4.7 kilohms (kΩ)
R3	1.2 megohms (MΩ)
VR1	25 kilohms (kΩ) log with DPST switch

Capacitors: all rated at 25 V minimum

C1, C2	4.7 microfarads (μF)
C3	47 microfarads (μF)
C4	220 microfarads (μF)

Semiconductors

TR1, TR3	BC548 npn
TR2	BC558 pnp
D1, D2	1N4148

Additional items
- PCB
- Speaker >35 ohms
- PP3 battery clip and battery

3 A medium-wave receiver using a ferrite-rod aerial

Introduction

This design came from the Norfolk Amateur Radio Club, and enables you to build a simple Amplitude Modulation (AM) receiver for frequencies between 600 kHz and 1600 kHz. It should take you around 2 hours to build, and can be used with Walkman-type earpieces. **Figure 1** shows the circuit diagram.

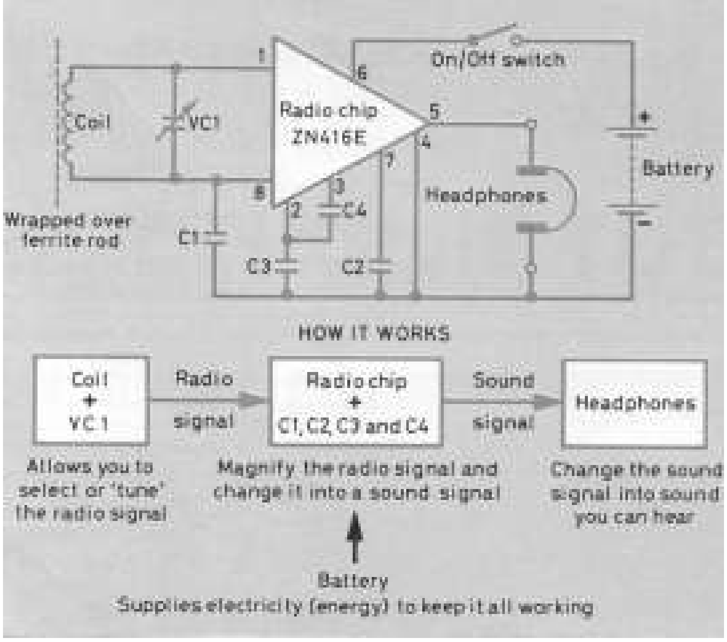
1
6
On/Off switch
Coil
VC1
Radio chip
ZN416E
5
4
7
+
Battery
−
8
3
2
C4
Headphones
Wrapped over
ferrite rod
C1
C3
C2
HOW IT WORKS
Coil
+
VC 1
Radio
signal
Radio chip
+
C1, C2, C3 and C4
Sound
signal
Headphones
Allows you to
select or 'tune'
the radio signal
Magnify the radio signal and
change it into a sound signal
Change the sound
signal into sound
you can hear
Battery
Supplies electricity (energy) to keep it all working

Figure 1 Circuit and block diagrams of the radio

Description

The whole circuit is built on a 50 mm by 50 mm printed circuit board (PCB) designed to fit on the inside of the lid of a plastic box, and is stuck there using sticky pads, the shaft of the variable capacitor going through a hole in

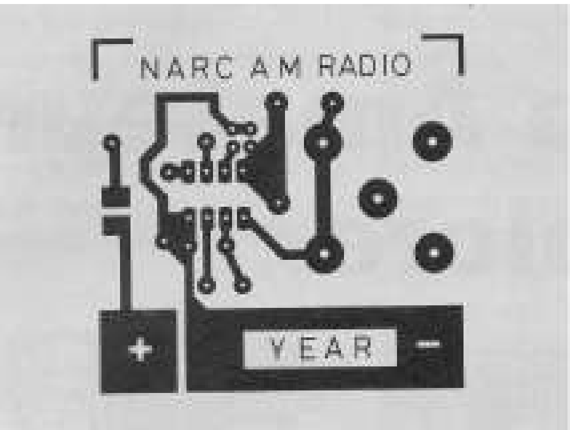

Figure 2a The PCB, solder side

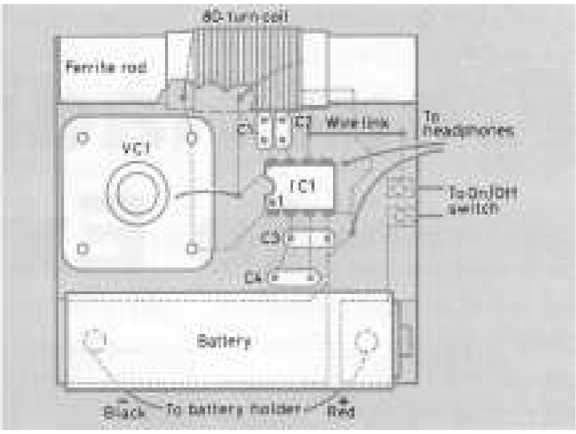

Figure 2b The PCB, component side

the lid. Only two pairs of leads are soldered to the board – one pair goes to the 1.5 V battery in its holder, and the other to the earphone socket. **Figures 2a** and **2b** show the printed circuit and the component layout double size for clarity. You are not obliged to build the circuit on a PCB.

Building it

1. **Check and identify components**. Tick the parts list.
2. **Carefully unwind the wire.** Use paper to make an insulating tube (called a 'former') around the centre of the ferrite rod and secure it with Sellotape. Now, close-wind *all* the wire (leave no gaps between adjacent turns) around the paper former. Secure the winding with more Sellotape, leaving 50 mm of wire free at each end for connection to the circuit. See **Figure 3a.**
3. **Solder in VC1.**
4. **Solder in the integrated circuit holder.** There is a notch in one end of the holder; this should face VC1. Solder also the wire link and the capacitors. Be careful to avoid solder 'bridges' between adjacent tracks on the PCB.
5. **Solder the battery leads.** These must be connected properly – the red battery lead to the + (positive) area and the black lead to the – (negative) area.
6. **Strip bare 1 cm of insulation from the ends of two wires.** Solder them between the PCB and the headphone socket (see **Figure 3b**). Use the end tabs on the socket. Using another pair of insulated wires connect the ON/OFF switch to the PCB tabs shown in Figure 2b.

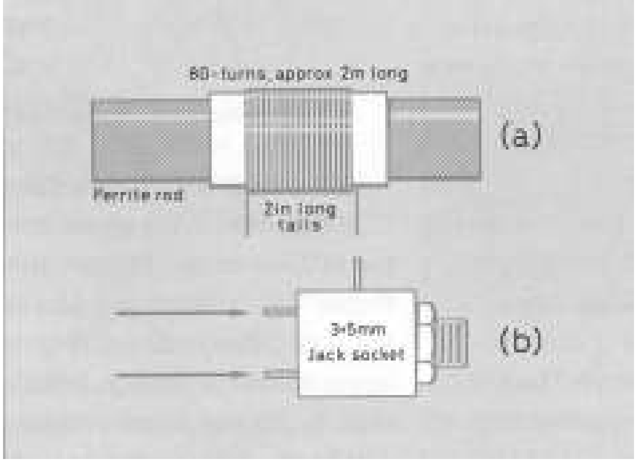

Figure 3 Details of coil and headphone socket

7. **Fix the elastic band.** This goes through the holes at the top of the PCB, with the ferrite rod being slipped through the two end loops. (Note: although the coating on the copper wire is designed to melt away during soldering, it is quite common for difficulty to be experienced in obtaining a good soldered joint; to be on the safe side, remove the coating *before* soldering (with a small piece of sandpaper).) Carefully place the wire ends of the coil through the PCB just above VC1, and solder on the track side.
8. **Fit IC1 into its holder.** This should be done carefully, making sure that *all* the pins are located above their respective clips *before* applying any pressure! Make sure also that the notch on the IC (as shown in Figure 2b) matches the notch in the holder, and faces VC1.
9. **Put battery in its holder.** Listen for some noise in the headphones as VC1 is rotated. Make sure the headphone plug is fully inserted into its socket.
10. **Fix the working board to the lid.** Use the sticky pads and apply *gentle* pressure. Fit the tuning knob, the ON/OFF switch and the earphone socket.
11. **Test again.** If all is still working, fit the lid screws and admire your completed radio!

Parts list

Capacitors	
C1, C2	0.01 microfarad (μF)
C3, C4	0.1 microfarad (μF)
VC1	500 picofarads (pF)
Semiconductor	
IC1	ZN416E

Additional items
- Plastic box (recommended size 76 × 64 × 50 mm internal)
- 8-pin DIL socket for IC1
- Printed circuit board
- Tuning knob for VC1
- Wire link for PCB
- 2 m of 30 SWG copper wire, self-fluxing
- Piece of paper 25 × 50 mm, to make the coil former
- Ferrite rod 70 mm long by 10 mm diameter, approximately
- Battery, AA size 1.5 V, with holder and attached wires
- Miniature earphone socket (3.5 mm stereo jack)
- ON/OFF switch (push-button SPST latched or slide switch)
- 4 off 100 mm insulated connecting wires, for jack socket and ON/OFF switch
- Pair Walkman-type earphones
- Elastic band, to attach ferrite rod to PCB
- 4 off sticky pads for securing PCB to box lid

Kits

Ready-made PCBs may be available from Alan J. Wright, G0KRU, Hewett School, Cecil Road, Norwich NR1 2PL.

4 A simple electronic organ

Introduction

This project has nothing to do with radio but, let's admit it, *any* electronics project is good experience! Why not build this little organ – it will keep the children amused at least! It uses the popular NE555 integrated circuit, which contains a circuit which will periodically switch the voltage on the output pin between the supply voltage and zero. Just how frequently this switching occurs depends upon the components external to the integrated circuit. If this switching occurs several hundred or thousand times a second, the change in voltage produced will generate a musical note when connected to a small loudspeaker. The circuit is shown in **Figure 1**.

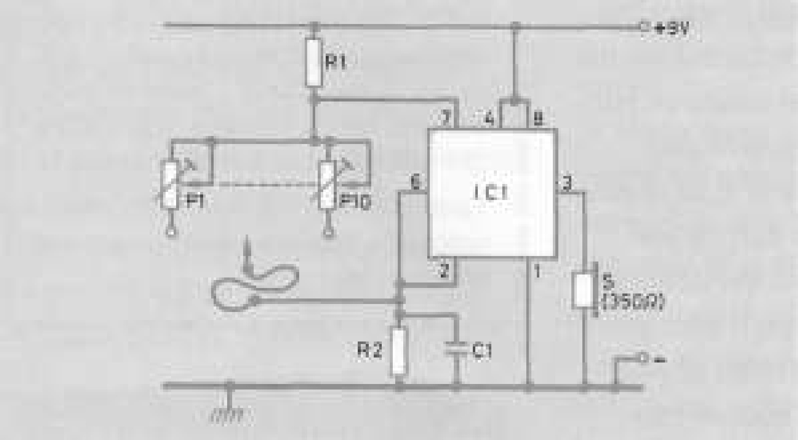
+9V
R1
7
4
8
6
IC1
3
P1
P10
2
1
S
(350Ω)
R2
C1
−

Figure 1 Circuit diagram

Putting it together

(a) Using a PCB. The job is very simple. The placement of components on the unsoldered side of the board is shown in **Figure 2** and the design on the copper track is illustrated in **Figure 3**. Put each component, in turn, on the board, making sure that it lies flat on the board with its tags or wires going cleanly through the holes provided for it; then, solder the wires to the board, cropping them before or after the soldering,

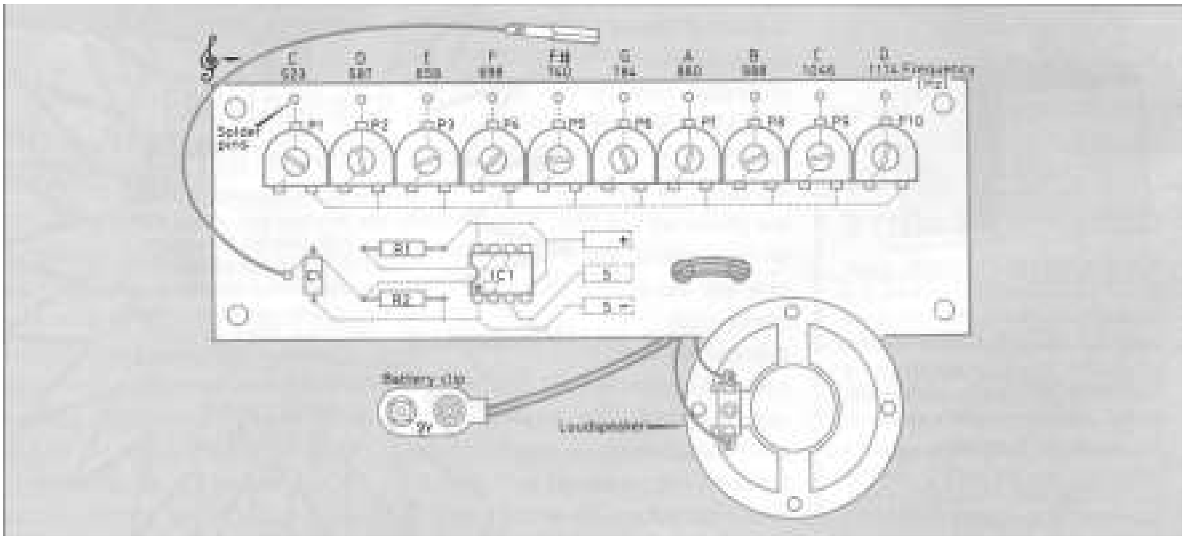
C
523
D
587
E
659
F
698
F#
740
G
784
A
880
B
988
C
1046
D
1174 Frequency (Hz)
Solder pins
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
R1
IC1
R2
C1
+
S
S −
Battery clip
9V
Loudspeaker

Figure 2 Position of components on the printed circuit board (PCB)

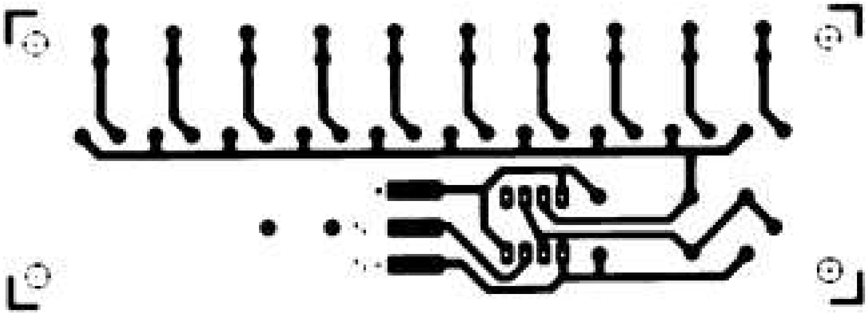

Figure 3 The connections

depending on your preference. If you choose to use a holder for your integrated circuit (highly recommended if your soldering is less than perfect), make sure that the end with a notch in it faces R1 and R2, as shown in Figure 2. Solder the two leads to the speaker to the tabs marked S (either way round), having looped them through the two holes to the right of the tabs in Figure 2. Looping them through the holes acts as a strain relief, ensuring that the soldered joints are not subjected to pulling and bending as you move the wires about. Do the same with the battery leads, the red lead going to the + tab and the negative lead to the – tab (which also has one speaker lead already attached to it). **Figure 4** shows this in detail. Treat the loudspeaker with care – the cone is quite fragile and must not be touched.

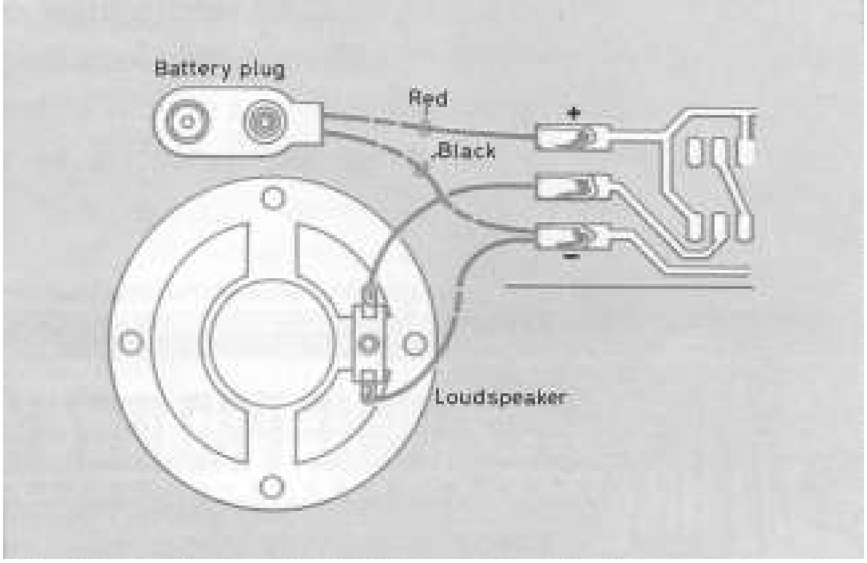

Figure 4 Battery plug and loudspeaker connections

(b) Without a PCB. This is more difficult, and you may need to enlist some help. Using some matrix board (such as Veroboard) is probably the best way of replacing the PCB. You could arrange your circuit in exactly the same way as in the PCB in Figure 3, using wires to replace the copper track.

A simple ‘keyboard’

The keyboard is a row of solder pins along the rear edge of the PCB, one for each note covering the range shown in Figure 2. A flying lead with a small spade on it is provided to touch any of the pins in turn, producing any one of ten different notes.

Testing

Check first that each component is in the correct place. When inserting the NE555 chip, first make sure that the end carrying the notch lies over the end of the holder with the notch; then, make sure each pin of the chip lies directly above the hole into which it fits, before pressing *gently* to insert the chip into the socket. Make sure the battery connections are correct, and insert the battery into the clip. Nothing should happen, except for a click from the loudspeaker; touching the spade on any of the pins should produce a coarse note from the speaker. If nothing happens, check everything again; don’t *assume* that wires go where you think they go!

After you get the first note, all the others should work, too, but they will sound off-tune at first. The organ needs tuning up by adjusting the 10 preset variable resistors P1 to P10. The approximate frequency to which each note should be tuned is given in Figure 2; if you can beg, borrow or steal a frequency counter, setting up is easy. If you have a piano, the organ can be tuned by comparison of the notes with those on the piano. The frequencies are given in Hertz (abbreviation Hz), and represent the number of times the IC switches on and off every second. If the sound coming from the loudspeaker is too loud or very distorted, then try putting an 330 Ω resistor (colour code orange, orange, brown) in series with the loudspeaker. This is done by taking the resistor and cutting its leads to about 5 mm; then, disconnect one speaker lead from the tab on the PCB (it doesn’t matter which). Solder one end of the resistor to the vacated speaker tab, and the free speaker lead to the other end of the resistor. This will limit the volume of sound from the speaker, and lengthen the life of your battery. If it is still too loud, try a resistor of a larger value, or use a smaller resistor to make it louder.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	2.7 kilohms (kΩ)
R2	1 Megohm (MΩ)
P1–P5	Preset resistor 100 kilohms (kΩ)
P6, P7	Preset resistor 50 kilohms (kΩ)
P8	Preset resistor 25 kilohms (kΩ)
P9, P10	Preset resistor 10 kilohms (kΩ)

Capacitor

C1	100 nanofarads (nF) or 0.1 microfarad (μF)

Integrated circuit

IC1	NE555 timer chip

Additional items

S	Loudspeaker >60 ohms (Maplin)
	1 off battery clip (for PP3 battery)
	1 off spade terminal
	12 off solder pins 'Veropins'
	3 off 10 cm lengths of 'hook-up' wire

(This article is based on projects originally designed by Radio Scouting, Netherlands.)

5 Experiments with the NE555 timer

Introduction

Several of the projects in this book use the NE555 timer, an integrated circuit which is at the heart of many circuits whose processes are determined by time intervals. **Figure 1** shows the circuit diagram of an audio oscillator using the 555. The timing voltages (governing the frequency of oscillation) are produced by R1, R2 and C2; a voltage appears at pin 3 which 'switches' at this frequency between zero and a voltage close to the supply voltage, which in this case can be anywhere between 6 V and 14 V. The output current, when applied through R3 to a small loudspeaker, produces an audible tone, provided that there is a DC path between the two test leads.

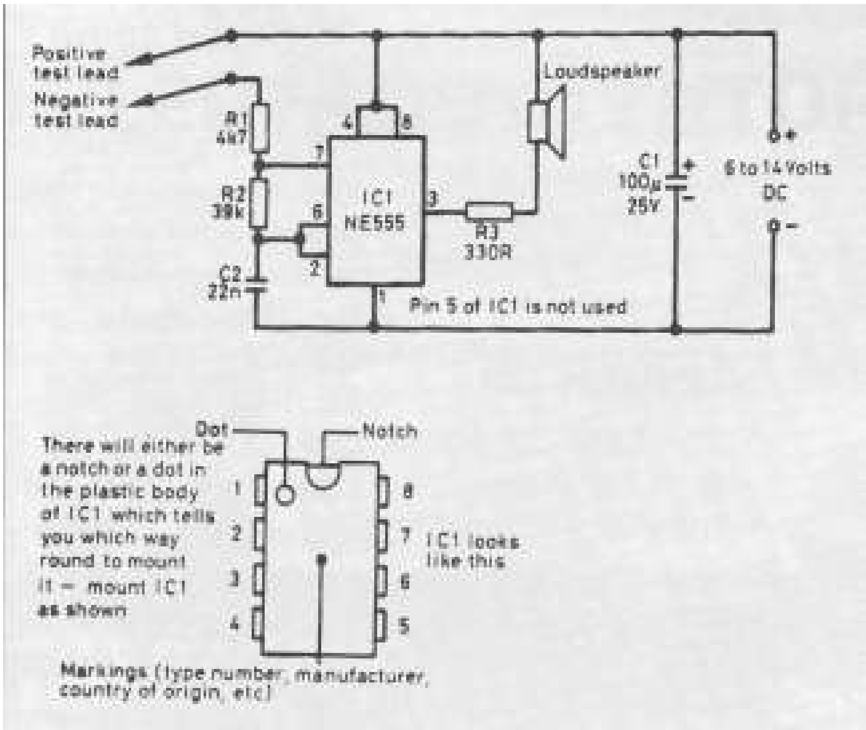
Positive test lead
Negative test lead
R1 4k7
R2 39k
C2 22n
4
8
7
6
2
1
3
IC1 NE555
R3 330R
Loudspeaker
C1 100μ 25V
+
−
+
6 to 14 Volts DC
−
Pin 5 of IC1 is not used
Dot
Notch
There will either be a notch or a dot in the plastic body of IC1 which tells you which way round to mount it − mount IC1 as shown
1
2
3
4
8
7
6
5
IC1 looks like this
Markings (type number, manufacturer, country of origin, etc)

Figure 1 The circuit diagram

Construction

The simplest way to mount the components is on a piece of matrix board (Veroboard), available from any of the good suppliers. The prototype of this circuit used the type of board with copper strips along the underside; these strips are used like the copper tracks on a PCB, to join components together. Firstly, cut the four strips between the positions of the pins of the IC socket, as shown in **Figure 2**. You can buy a tool for this purpose, but a small twist drill (about 3 mm diameter) is just as good. Turn it between your fingers – if you use a drill you will end up with holes right through the board! Then solder in the IC socket (with the notch in the position shown), followed by the four links made with single-conductor insulated wire. Put in each component as shown, ensuring that C1 (an *electrolytic* or *polarised* capacitor) is connected correctly. When all the components have been soldered in, take the 555 chip and lay it on its socket, with its own notch lying above that of the holder. Then, making sure that each pin lies *directly* above its corresponding socket, press down *gently* on the chip, with the board supported on a flat surface.

Testing

Connect the circuit to a battery or small power supply, ensuring that the positive and negative leads are the right way round. Always use red and black leads here, then you are less likely to get it wrong! Switch on. Nothing

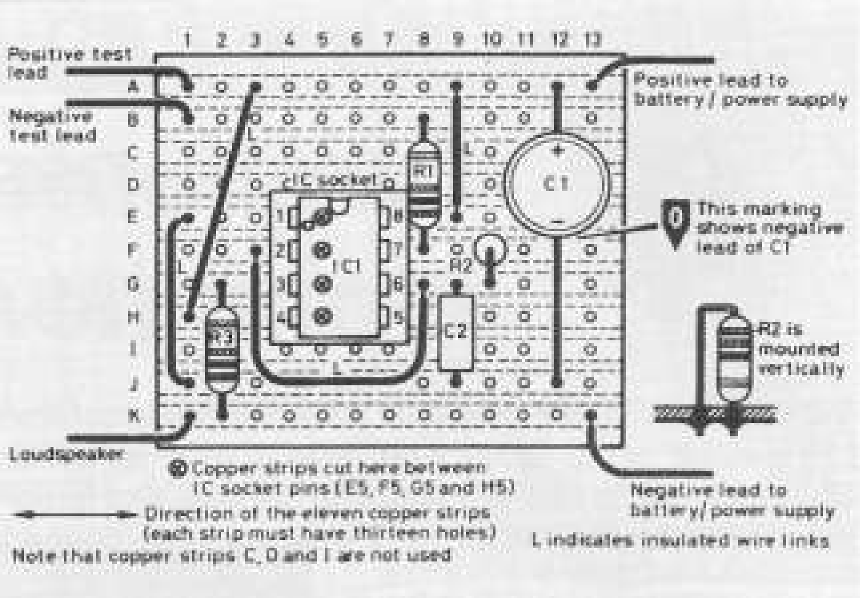
1 2 3 4 5 6 7 8 9 10 11 12 13
A B C D E F G H I J K
Positive test lead
Negative test lead
Loudspeaker
L
IC socket
1 2 3 4 5 6 7 8
IC1
R1
R2
R3
C1
C2
+
–
Positive lead to battery / power supply
This marking shows negative lead of C1
R2 is mounted vertically
Negative lead to battery/power supply
L indicates insulated wire links

⊗ Copper strips cut here between IC socket pins (E5, F5, G5 and H5)

←——→ Direction of the eleven copper strips (each strip must have thirteen holes)

Note that copper strips C, D and I are not used

Figure 2 Veroboard layout. If you can read a circuit diagram, the project can be built using other methods

should happen until you short together the two test leads, when there should be a note from the loudspeaker. If this doesn't happen, switch off, disconnect the circuit and check your wiring and soldering. Is it exactly like Figure 2? Are the soldered joints round and shiny? If any are dull, then 'sweat' them briefly with a hot soldering iron until the solder runs, remove the iron, and check that they are as shiny as the rest. Check that there are no solder 'bridges' between adjacent tracks by holding your board up to a strong light. Then, reconnect, switch on and touch the test leads together. All should now work!

Uses of your circuit

1. As a Morse practice oscillator. Simply connect the two test leads to your key and, each time the key is pressed, you should hear a note from the speaker. The frequency of the note may be altered by putting a resistor in series with the key. To do this, remove one test lead from the key and select a resistor; connect one end of the resistor to the free test lead and the other to the empty terminal on your key. Selecting the value of resistor that you need will be a useful experiment in itself.
2. As a continuity tester. You can check fuses and lamp bulbs by connecting them across your test leads. If the speaker remains silent, the fuse or bulb has blown.
3. To indicate changes of resistance. Hold the ends of the test leads in each hand; you should hear a low note, because of the high resistance of your body. Squeeze the ends harder, and the frequency of the note should rise, because you are now making better contact. Repeat this with damp hands and the frequencies will be higher still.
4. As a thermometer. Connect the test leads to a thermistor (a device whose resistance changes with temperature) and warm it with a hair-dryer, or even in your hands, and you will hear the pitch changing with the temperature of the thermistor. A suitable 'bead' thermistor is available from Maplin (order code FX21).
5. As a diode tester. Use any diode, and connect the negative test lead to the end of the diode marked with the ring. This is the cathode of the diode. The other end, the anode, should be connected to the positive test lead, and a note should be heard from the speaker. This does not necessarily mean that the diode is working – yet. Reverse the connections and *nothing* should be heard. If this is the case, the diode *is* working.
6. As a light meter. Use a photoconductive cell (a device whose resistance changes with light intensity) connected between the test leads. A note should be heard. Shading the device with your hand will increase its resistance and the note should decrease in frequency. A suitable device is the ORP12 cell from Maplin (order code HB10).

There are many more applications. **Do not connect the test leads to other circuits that are switched on.** Your circuit, or the circuit you are connecting it to could be damaged. Think of a *passive* device or circuit (i.e. one *not* requiring a power supply or battery) where changes of resistance occur, and you have found another application!

Parts list

Resistors: all 0.25 watt carbon film types

R1	4.7 kilohms (kΩ) – yellow, violet, red
R2	39 kilohms (kΩ) – orange, white, orange
R3	330 ohms (Ω) – orange, orange, brown

Capacitors

C1	100 microfarads (μF) 25 V radial electrolytic
C2	22 nanofarads (nF) or 0.022 microfarad (μF) polyester type with 10 mm lead spacing

Integrated circuit

IC1	NE555

Additional items

Miniature loudspeaker (preferably 35, 40 or 80 ohm)
8-pin DIL socket for IC1
0.1 inch Veroboard (‘stripboard’), size 11 strips by 13 holes
PVC-covered stranded wire for test leads, loudspeaker and battery connections
PVC-covered solid wire for links on the board
A power source of between 6 V and 14 V, such as a 9 V battery (PP3)

Component sources

Cirkit
Tandy – many high street shops

6 A simple metronome

Introduction

A metronome is a device used by musicians to indicate the tempo of a piece of music. Until electronics came on the scene, this 'beating of time' was achieved in much the same way as a clock keeps time, i.e. with a pendulum device, the clicking of the escapement indicating the beats of the music.

Those of you who have already built the Morse Key and Buzzer from the designs in this book, will recognise the circuit of this metronome – it is exactly the same as was used to produce the note of the buzzer. This circuit is shown in **Figure 1**.

The circuit

Three components determine the speed at which the circuit oscillates – the speaker (LS), the resistors (VR1 + R1) and the capacitor (C1). VR1 is a variable resistor, so that the speed at which the oscillator operates can be varied. Compared with the component values of the Morse Buzzer (which operated at around 800 Hz), these components now give an oscillation frequency of around 1.25 Hz, which is far too low to be heard as a note. What we *do* hear, however, is a series of clicks, as the voltage across the speaker changes quickly from 0 to 9 V and back again.

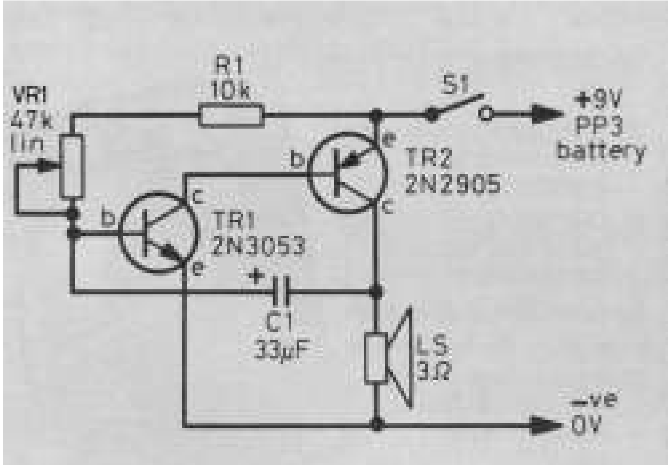

Figure 1 The metronome circuit is rather like the Morse oscillator

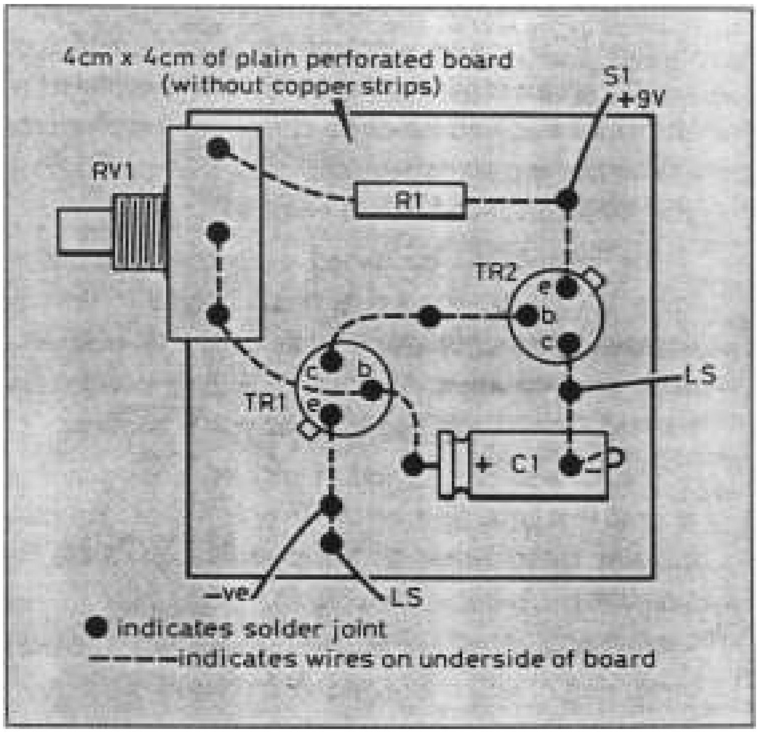
4cm x 4cm of plain perforated board
(without copper strips)
S1
+9V
RV1
R1
TR2
e
b
c
LS
c
b
e
TR1
+ C1
−ve
LS
● indicates solder joint
– – – –indicates wires on underside of board

Figure 2 The component wires are pushed through holes in the circuit board and joined together underneath

Variation of speed could be achieved by varying resistance or capacitance. However, as you may already know, variable capacitors have values in the picofarad range, not the tens of microfarads used here, so it is very simple to employ a variable resistor (potentiometer) to control the oscillator. You could use a multi-way switch to switch in one of several capacitors, as well as having the variable resistor, but this was found to be an unnecessary complication. This design operates between about 100 clicks per minute and 200 clicks per minute.

Making the prototype

A single piece of plain matrix board (no copper strips) measuring about 40 × 40 mm is sufficient to hold all the components except the potentiometer and switch (see later). The case can be plastic or aluminium, and one measuring 65 × 100 × 50 mm is about right. Make sure there are holes in the case beside the speaker cone to let the sound out, and larger holes for the potentiometer and switch. If a potentiometer is used with a combined ON/OFF switch, then the extra hole for the switch is not necessary! It is advisable to construct the circuit *before* putting it in the box, so that it can be tested to ensure that everything is working. If it is, then you can exercise your ingenuity in mounting the speaker, battery and board inside the box. A final test can be made before starting the calibration process.

Calibration

There is no 'easy' way to do this. The frequencies involved are too low to be measured with the average frequency counter, so you will need to resort to using a stopwatch and counting the number of clicks per minute.

Parts list

Resistors: 0.25 watt, 5% tolerance

R1	10 kilohms (kΩ)	
VR1	47 kilohms (kΩ) linear potentiometer	

Capacitor

C1	33 microfarads (μF) electrolytic	

Transistors

TR1	2N3053	npn
TR2	2N2905	pnp

Additional items

S1	SPST	ON/OFF switch
LS	3 ohms (Ω)	loudspeaker
	Knob with pointer for VR1	
	PP3 battery and connector	
	Aluminium case, 65 × 100 × 50 mm	
	Matrix board (plain), 40 × 40 mm	

7 What is a resistor?

Introduction

Materials that carry electricity easily are called **conductors**. They include all metals and salt water, for example. We use wire as a conductor, and the ease with which it passes an electric current depends upon the material, its thickness and its length. Silver (symbol Ag), gold (Au), copper (Cu) and aluminium (Al) are the best metallic conductors. Most wires are made of copper, although the best conductor, weight for weight, is aluminium.

Materials that *don't* carry current (or, at least, do so very badly) are called **insulators**, and they include dry wood, rubber, plastic and glass among their number. Wires are often coated with a layer of insulator to prevent adjacent wires touching and causing an accident.

Resistors

If there wasn't such a thing as resistance, the subject of electronics wouldn't exist; only infinite currents would flow and voltages wouldn't exist either! We need to reduce the flow of current if we are to make current do something useful for us. Components that resist the flow of current are called **resistors**, and they are said to have a **resistance** which is measured in ohms (Ω), named after Georg Ohm, who formulated the law (also named after him) by which the voltage and current through a conductor are related. His law gave rise to the formula *everyone* remembers:

$$I = \frac{V}{R},$$

where I is the current flowing, measured in amps,
V is the voltage across the conductor, and
R is the resistance of the conductor, measured in ohms.

From this equation, you can see that, for a constant value of voltage, V, if the resistance goes up, the current will go down, and vice versa. The circuit symbols for resistors are shown in **Figure 1**. You will find the upper symbol in older magazines; it is still preferred by many engineers. The lower symbol is the prevalent standard symbol.

Resistors are made in several ways, the cheapest using carbon; another type is usually made from a ceramic cylinder (used only as a support) on which

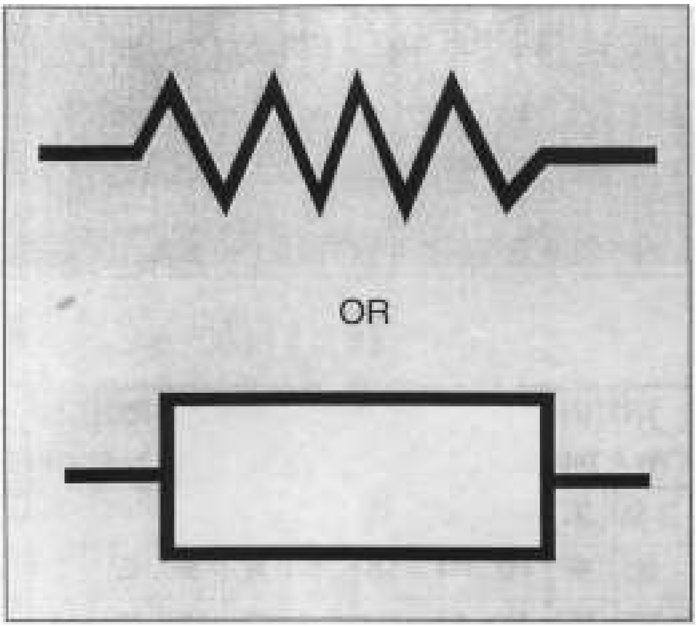

Figure 1 Circuit symbols for resistors

is placed a very thin film of metal – the thinner the film, the greater the resistance. All resistors are coated with a thin film of insulation, for the same reason we discussed earlier.

The colour code

Each resistor has coloured bands on it which enable us to see what value of resistance it has. There are normally three (but sometimes four) at one end, and a single one at the other (see **Figure 2**). The colours indicate figures, according to the list below.

Colour	*Value*	*Colour*	*Value*
Black	0	Green	5
Brown	1	Blue	6
Red	2	Violet	7
Orange	3	Grey	8
Yellow	4	White	9

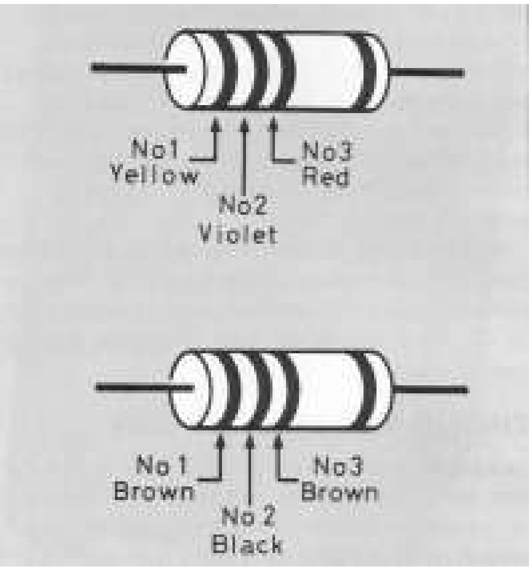

Figure 2 Some examples of resistor colour codes; top 4700 Ω (4.7 kΩ) and bottom 100 Ω

Using the colour codes is easy, once you see the logic behind it. Hold the resistor so that the single band is towards the right. The three colours on the left are read in the normal order from left to right. The first two bands *always* indicate numbers; the third band gives the number of zeros to add to the right of these two numbers. So, looking at the top resistor in Figure 2, yellow, violet, red means 4, 7, and two zeros, giving 4700 ohms! Looking at the lower resistor, brown, black, brown means 1, 0, and one zero, giving 100 ohms.

Remembering the order of the colours may be difficult at first. The colours from red to violet are the colours of the rainbow, in order, so if you know those, you're almost there! Around those colours are black and brown below the red, and grey and white above the violet, which you can imagine as getting brighter from black to white (well, almost!). It won't be long before you don't need to remember them at all.

The isolated band on the right-hand side is not part of the resistor's value; it indicates its *tolerance*, i.e. how close it might be to the indicated value. A brown band indicates ±1%, a red band ±2%, a gold band ±5% and a silver band ±10%. For example, a resistor marked as being 100 ohms with a ±5% tolerance will have an *actual* value somewhere between 95 ohms and 105 ohms.

8 Waves – Part 1

Introduction

Waves are responsible for most of the processes in life where energy is transferred from one place to another. Heat and light energy from the sun, for example, come to us as *electromagnetic waves*. Sound travels through the air as a wave; it is not the same sort of wave as light or heat, but it obeys many of the same properties. Damage is caused to coastal margins by the waves of the sea – again, another type of wave, but still obeying many of the same properties. Radio waves are of the same type as heat and light waves, as are gamma rays, X-rays, ultra-violet waves and infra-red waves. So, once we begin to understand what radio waves do, we are also learning about a huge chunk of physics at the same time! All these waves are part of the *electromagnetic spectrum*. The word 'spectrum' simply means a 'range', so what we have is a range of electromagnetic waves – that's all!

Sensing things

Light waves are invisible, but our eyes can detect the *effect* they have on different materials because the waves produce an effect on the retina of the human eye which the brain can interpret. We cannot see heat waves either, but we can feel the *effect* they have on our skin. Gamma rays and X-rays are also invisible, but their detrimental *effects* on human tissue are well known. It is not surprising, then, that we cannot see radio waves. We cannot sense them, either, until we produce a device upon which they have an *effect*. That device is a **radio receiver**; it is able to process certain characteristics of radio waves, and make these characteristics audible by generating sound waves from the loudspeaker or headphones. Other characteristics of the same waves may be turned into light as a TV picture on a cathode-ray tube, or as a fax image on a sheet of paper.

Visible waves

Let's start our description with some waves that we can actually see! When a small stone is thrown into a pond, we see circular water waves *radiating* from the point where the stone fell into the water, as **Figure 1** shows. (Notice that we use the word *radiating*, even with water waves; it is not a radio term, but one which describes any motion where the *radius* of a circle is increasing. In this case, it is the radius of the circular waves which is

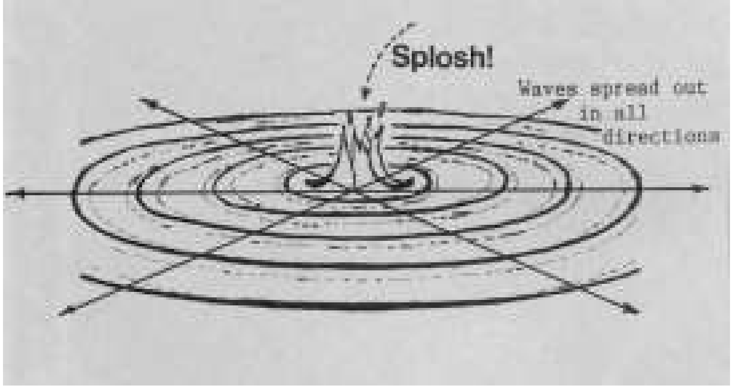
Splosh!
Waves spread out in all directions

Figure 1 The waves move out from the point where the stone landed

increasing.) If you were to watch the water waves down at the water level, perpendicular to the direction in which they are travelling, you would see something like the illustration in **Figure 2**.

The horizontal line represents the water level before the wave started, and the vertical line represents the direction in which the water is displaced at any instant.

All waves are described in the same way

'Freezing' the motion of the water in this way allows us to define two very important characteristics of a wave, characteristics which we talk about every day – *wavelength* and *amplitude*. The wavelength of a wave is simply the distance (measured in metres) from any point on one wave to the same point on the adjacent wave. Look at the diagram and you will see what is meant by 'the same point on the adjacent wave'. The amplitude of a wave

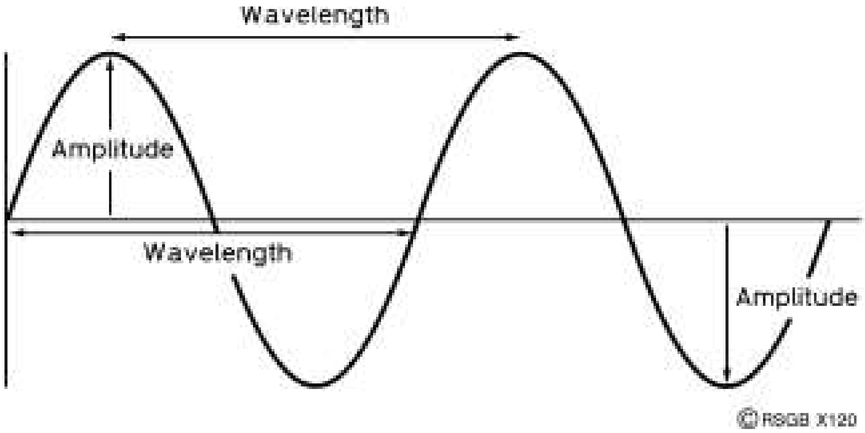

Figure 2 A water wave, viewed in cross-section

is always measured from the centre (undisturbed) position of a wave to the peak (or the trough) of the wave. Both these positions are shown, the arrow indicating that the measurement is taken *from* the centre *to* the peak or trough. The amplitude of a wave is defined as the maximum displacement of the wave from the centre position – the direction (up or down) of that displacement does not matter. Waves of greater amplitude carry more energy with them.

If we now 'unfreeze' the wave, we will see it travel from left to right (or right to left, depending on where we are looking). The speed at which it moves is called its *velocity*, and is measured in metres per second.

Another useful word is *propagate*; it means *travel*. We talk about radio waves propagating from a transmitter to a receiver. This velocity of propagation (for electromagnetic waves) is very fast indeed – they will cover 300 million metres in one second. This is virtually incomprehensible, so think of a radio wave travelling around the earth – it can travel 7½ times round the earth in one second! We use the symbol c for the velocity of radio waves (which is the same as the velocity of light, of course – all electromagnetic waves travel at this speed through air and space).

The last thing we need to know about the wave is its *frequency*. Imagine a cork floating on the water in the path of the wave; it will bob up and down. If we were able to count the number of times it went through its highest position in one second, then that number would be its frequency. Any periodic motion like this is said to go through one *cycle* each time one complete wave passes a point (in this case, our cork). We are thus counting the number of cycles per second of the cork's motion. The unit of frequency is thus 'cycles per second'; this unit is now named after Hertz, a radio pioneer, and is abbreviated to Hz.

Our description of the wave is now quite simple – we need only four quantities:

(a) **Frequency** symbol f – unit, hertz (Hz)
(b) **Wavelength** symbol λ (Greek letter lambda, pronounced 'lamb-da') – unit, metre (m)
(c) **Amplitude** symbol a – unit depends on application
(d) **Velocity** symbol c – unit, metres per second (m/s or ms^{-1})

The basic formula

Whatever may happen to a wave while it travels through different media (vacuum, air, brick, wood, etc.), one thing and **only** one thing remains constant – its frequency. Its wavelength, amplitude *and* velocity may change, but its frequency never does. Three of the four characteristics already identified are connected by the simple relationship

$$c = f \times \lambda.$$

Remember that *c* is constant if the wave travels in air or in a vacuum. This means that waves having higher frequencies (*f* large) must have smaller wavelengths (λ small) and vice versa. You can imagine frequency and wavelengths being on opposite ends of a see-saw!

Divisions of units

Because the frequencies of radio waves are so high (despite them having the lowest frequencies in the electromagnetic spectrum!) we have a problem with writing them down. Do you write in your log book that you have just heard a station on 14 100 000 Hz? Of course not, you write it as 14.1 MHz, knowing that the prefix mega (M) means ‘one million’. The prefixes which you need to know (when applied to frequency) are:

kHz	kilohertz	meaning	1000 Hz
MHz	megahertz	meaning	1 000 000 Hz
GHz	gigahertz	meaning	1 000 000 000 Hz.

Notice that the ‘k’ in kilohertz is a lower case letter. It is **incorrect** to write it as an upper case letter. ‘K’ is a computer-related prefix meaning **not** 1 000 but 1 024!

When we come on to discuss heat and light waves, we will use wavelengths rather than frequencies, because of the see-saw effect – as the frequencies get larger and larger, the wavelengths get smaller, and hence are numbers which are more manageable, both to talk about and to write down!

Bands

Gamma rays, X-rays, ultra-violet waves, light waves, infra-red waves are all part of the electromagnetic spectrum, but we divide them up because they have different properties. This is why we divide up our radio frequencies into different bands. The radio waves of top-band signals (around 2 MHz) have completely different properties compared with those in the 20 metre band, so we are dividing up the *radio spectrum* in the same way – by property.

9 A beat-frequency oscillator

Introduction

Many readers will know that, although they have a short-wave radio which covers at least one of the amateur bands (e.g. 7 MHz or 14 MHz), they are unable to listen to SSB or Morse signals. This is because the receiver lacks a Beat-Frequency Oscillator (BFO). We need the ‘carrier’ frequency of a BFO to replace the carrier that has been removed from the signal at the transmitter. When listening to Morse signals, the BFO signal ‘beats’ with the incoming signal to produce a note in the loudspeaker. If you are a musician, you will be familiar with the method of using ‘beats’ to tune one musical instrument from another; in the BFO, the beat frequency produced is the tone signal you hear.

In the more complex amateur radio receiver, a BFO is incorporated as part of the whole system. In our model, it is an external circuit that sits alongside your radio. The circuit diagram is shown in **Figure 1**.

Construction

Built on a small piece of matrix board about 80 × 50 mm, the circuit can be fitted inside a small plastic box. For once, we *don’t* want to screen the circuit to prevent it interfering with other equipment; we *want* it to interact

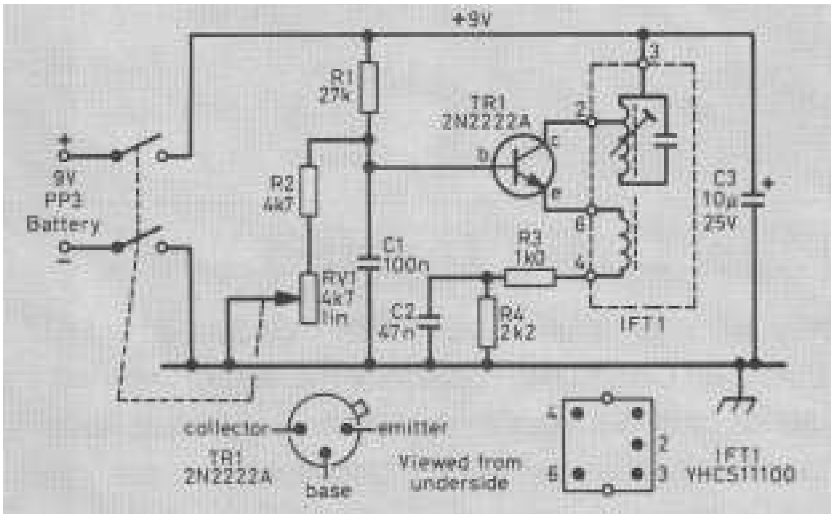
+9V
R1
27k
TR1
2N2222A
b
c
e
2
3
6
4
+
9V
PP3
Battery
–
R2
4k7
C3
10µ
25V
C1
100n
R3
1kΩ
RV1
4k7
lin
C2
47n
R4
2k2
IFT1
collector
emitter
base
TR1
2N2222A
Viewed from
underside
4
2
5
3
IFT1
YHCS11100

Figure 1 Circuit diagram of the BFO

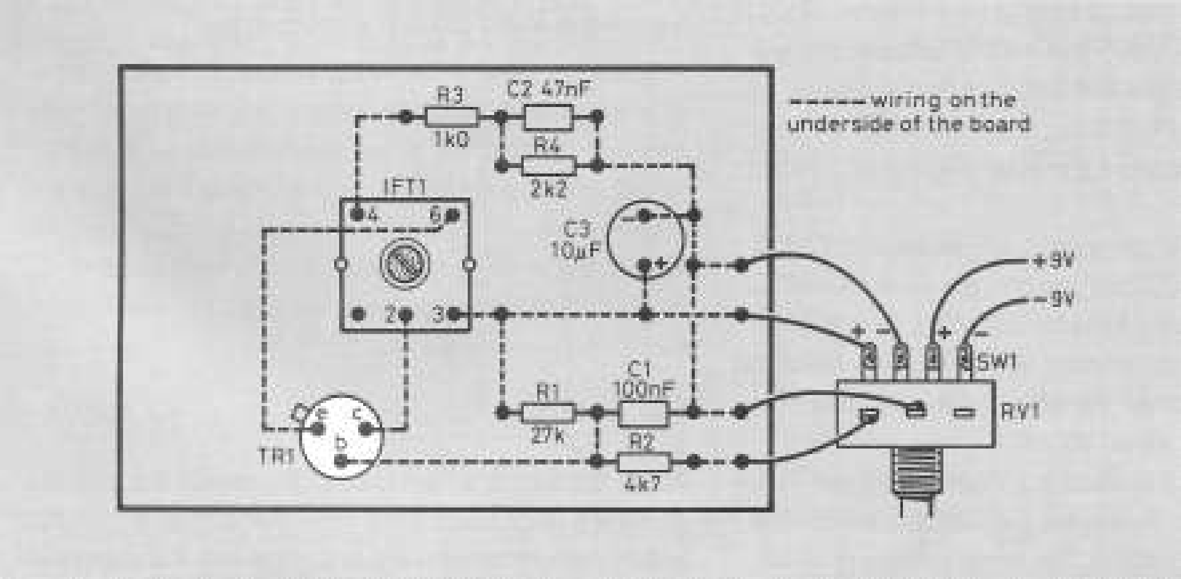

Figure 2 Matrix board layout shown from the component side. Adjust IFT1 carefully for the best results

with our receiver! This is why we use a plastic box. Maplin Electronic Components supply a suitable box, complete with the matrix board to fit inside (order code YU46).

Look at **Figure 2** carefully before you start to build the circuit, so that you can position the components correctly. Firstly, inspect the intermediate frequency (IF) transformer, IFT1, and remove its metal screening can very carefully. Again, this is to allow some signal to escape from our circuit and enter our radio. Having done this, solder the components, using the matrix board as a support. Underneath the board, the components are linked by single-conductor, insulated wire. Take particular care with the polarity of the electrolytic capacitor, C3, and the connections to the transistor, TR1 (see Figure 1).

The variable resistor, VR1, has a switch mounted behind the control itself, and the insulated leads to it from the battery should be about 10 cm long. Connect these before fitting VR1 into the case, so that the BFO can be calibrated (adjusted) correctly.

Calibration

After a final check that all the components have been fitted and soldered correctly, connect the battery, switch on, and hold the transistor between your fingers, to check that it is not getting hot. Place the circuit close to your receiver, and set RV1 to mid-position. Tune your receiver to find an amateur SSB transmission; the frequencies listed below will help you in knowing where to look. It may sound very strange, but don't worry. *Slowly* turn the core of IFT1 with a small, non-metallic screwdriver or with the correct 'trimming tool'. The core into which the blade fits is very fragile, so attempt this process with care. When the speech sounds as natural as you can get it, leave the core at this position, and use RV1 to make the speech sound natural.

Using the BFO

For best results, you may have to move the BFO nearer or further away from your radio. At the lower end of most bands (for instance just above 7.000 MHz or 14.000 MHz) you should be able to resolve Morse code (CW) signals. If you find that the BFO signal is a little weak, solder a 15 cm length of insulated wire to pin 2 of IFT1, and place it alongside your radio. This should improve signal intelligibility. When you are happy with the performance, switch off, drill a 10.5 mm hole in the box and fit RV1, followed by the matrix board assembly. Screw the base to the box, fit the knob, and the BFO is complete!

Where to listen

Band	**Frequencies (MHz)**
15 m	21.000–21.450
17 m	18.068–18.168
20 m	14.000–14.350
30 m	10.100–10.150
40 m	7.000–7.100
80 m	3.500–3.800

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	27 kilohms (kΩ)
R2	4.7 kilohms (kΩ)
R3	1 kilohm (kΩ)
R4	2.2 kilohms (kΩ)
VR1	4.7 kilohms (kΩ), linear, with DPST switch

Capacitors

C1	100 nanofarads (nF) or 0.1 microfarad (μF), ceramic
C2	47 nanofarads (nF) or 0.047 microfarad (μF), ceramic
C3	10 microfarads (μF), 25 V radial, electrolytic

Additional items

TR1	2N2222A npn
IFT1	Toko type YHCS11100
Box	plastic, approximate size 100 × 70 × 45 mm
Board	matrix, to fit inside the box
Connector	for PP3 battery
Knob	for RV1

10 What is a capacitor?

Introduction

A capacitor is a device that stores energy in the form of electricity. Much less energy than a battery, and for a shorter time, however. The simplest form of capacitor takes the form of two flat metal sheets separated by air; connections are made to each plate, as **Figure 1** shows. If you imagine a pair of these plates, 30 cm square and separated from each other in air by 1 mm, the capacitance of this device would be almost exactly 80 picofarads (pF), i.e. 80 million-millionths of the unit of capacitance, the farad. Now this is quite a small value, you will agree, and it comes about because the farad is such a large unit. Nevertheless, as you will probably know, we may have capacitors of value 10 000 microfarads (μF) in our radio equipment, and they can be smaller than your little finger, so they are obviously not made the same way!

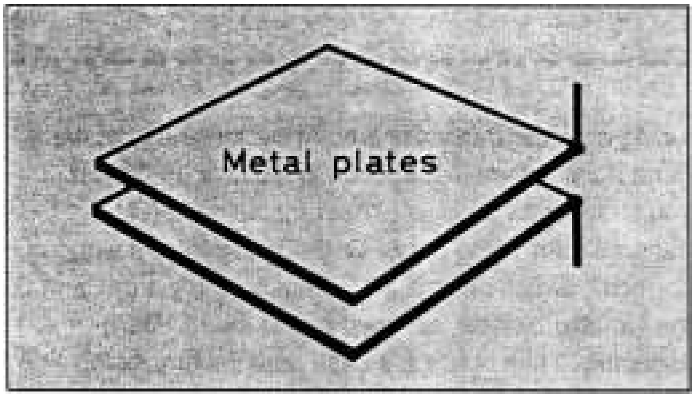

Figure 1

Large and small

We cannot go into the manufacture of capacitors here – after all, we are users of the devices, not the designers! First of all, beware of incorrect statements; the words 'capacitor' and 'capacitance' are *not* the same. For example, a large capacitor would be a description of one the size of a toilet roll. It need not have as large a capacitance as one the size of your little finger. A 'large capacitor' is one which is physically *big*, a 'large capacitance' refers to a capacitor which can store a larger amount of energy when a certain voltage is applied between its plates. The capacitors in a mains power supply are usually big **and** have large capacitances. High-power RF amplifiers may have large capacitors with small capacitances!

Electrolytics . . .

Electrolytic capacitors usually have capacitances of 1 μF or above. They differ from other capacitors in that they **must** be connected the right way round (i.e. they have positive and negative connections, just like a battery). They may explode if the connections are reversed!

. . . and the others

Other capacitors may be connected either way round, despite their names. We have *polystyrene*, *ceramic*, *silver-mica* and *tantalum*. Each has its own advantages and disadvantages, and the parts list for a project will always tell you which type is best.

Storing energy

If you were to connect a large capacitance across a 12 V power supply, nothing would appear to happen. Removing the capacitor from the supply and connecting it to a voltmeter would show that the capacitor has 12 V between its ends. This shows that, while the capacitor was connected to the supply, energy flowed from the supply into the capacitor. We say that the capacitor was *charged up* by the supply. If you are using an *analogue* voltmeter (i.e. one with a meter and pointer), you will notice that the indicated voltage slowly drops until, eventually, there is no voltage across the capacitor. This is because the capacitor has *discharged* its energy into the voltmeter. If you had used a smaller capacitance, the same would happen, except that the voltage would drop to zero more quickly – the capacitor stores a smaller amount of energy because its capacitance is smaller. Capacitors behave like other things in life – a small car can move more quickly than a large bus – a small piccolo emits a higher note than a flute – the voltages in a circuit containing a small capacitance will change more quickly than those in a circuit with large capacitance.

Varying the capacitance

Some capacitors are capable of having their capacitance changed manually; these are called *variable capacitors*. They work like the basic capacitor of Figure 1. Imagine moving the top plate of the pair a little to one side; the capacitance is determined, not just by the size of the two plates, but by their *area of overlap*. As this decreases, so does the capacitance. Such devices are limited in their capacitance, about 500 pF being the maximum value.

AC and DC

Because the plates of a capacitor do not touch each other, a direct current (DC) cannot pass between them. However, an alternating voltage on one plate can induce an identical alternating voltage on the opposite plate, and thus a capacitor appears to pass an alternating signal, even though currents as such, do not pass between the plates. This property of passing AC and not DC is very important, and a capacitor used in this way is called a *DC blocking capacitor* or, simply, a *blocking capacitor.* A blocking capacitor can be used at the same time, to couple a signal from one circuit to the next; here it would be known as a *coupling capacitor. Decoupling capacitors* are to be found where the capacitor is employed to remove an AC signal while retaining a DC component.

Finally . . .

Unlike resistors, the manufacture of capacitors renders them susceptible to excess voltage, so if you find a capacitor labelled 10 μF 16 V, it means that operating it above 16 V may fatally damage the device (and the circuit around it). This voltage is called the *working voltage* of the capacitor; on some electrolytics, you may find it expressed as *volts working* (i.e. 8 μF 450 V WKG).

Many smaller capacitors have their properties marked on them in a colour code, like resistors. **Figure 2** shows these codes, and their meaning, and the table below summarises the values of the colours.

Table 1

Colour	*Value*	*Voltage (tantalum capacitor)*	*Voltage (polyester capacitor)*
Black	0	10	–
Brown	1	–	100
Red	2	–	250
Orange	3	–	–
Yellow	4	6.3	400
Green	5	16	–
Blue	6	20	–
Violet	7	–	–
Grey	8	25	–
White	9	3	–

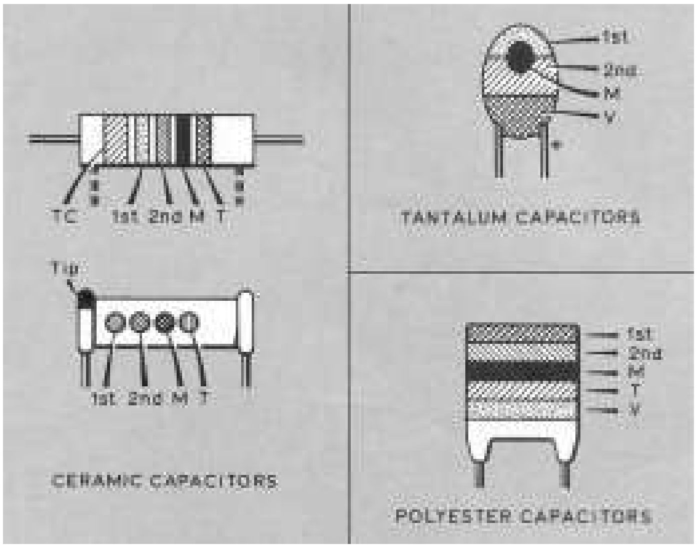
TC
1st 2nd M T
1st
2nd
M
V
TANTALUM CAPACITORS
Tip
1st 2nd M T
CERAMIC CAPACITORS
1st
2nd
M
T
V
POLYESTER CAPACITORS

Figure 2 Some capacitors have coloured bands or stripes, rather like resistors. The colour code, which is the same as the resistor code, is shown in Table 1. The band shown on the chart as ‘1st’ is the first number of the capacitor’s value in pico-Farads, ‘2nd’ is the second number and ‘M’ is the Multiplier or number of noughts. For example, a capacitor reads from the top: Brown, Black, Yellow, Black, Red. Its value is One, then Nought, then Four more noughts = 100 000 pF (also referred to as 0.1 μF or 100 nF). Its tolerance (Black) is 20% and the working voltage (Red) is 250 V. The ‘V’ means the maximum working voltage. The band marked ‘T’ shows the tolerance, just like resistors, and the one marked ‘TC’ is only used on special capacitors designed to change their value with temperature

11 Waves – Part 2

Introduction

We left Part 1 with the concept that radio waves are divided up into bands which have different properties. Not *all* the properties are different, though. We need to discuss several wave properties, so we will start with what happens to waves as they propagate over long distances.

Getting weaker

Imagine a torch battery connected to a bulb by wires about 1 metre long. The bulb lights normally. If we now take the bulb 100 metres away from the battery and wire it up, we would expect the bulb to be somewhat dimmer, which is exactly what would happen. It happens because of the resistance of the wires – the wires do not form a perfect conductor. A similar situation occurs with radio waves.

All waves suffer from *attenuation* – they get weaker the further they travel. In cases of extreme attenuation, we need to apply some *amplification* before the attenuated wave can be used in a receiver.

Carrying information

When we speak over the telephone, the range of frequencies in our voices extends from very low frequencies up to about 15 or 20 kHz. In audio terms, this is quite a large *bandwidth* (meaning a wide band of frequencies). For communications purposes, however, most of this bandwidth is not needed, and in the telephone system (and in our transceivers), this is cut down so that it extends from about 200 Hz to 3 kHz, a reduction in bandwidth from 20 kHz to about 3 kHz. A bandwidth of 3 kHz has been found to be sufficient to convey speech intelligibly which, after all, is just what we need!

The radio waves coming from an amateur transmitter convey our speech signals over long distances. By themselves, the speech signals do not travel very far, so they have to be combined with a radio signal that *will* travel long distances. This extra signal is called the *carrier wave* (or just the *carrier*), because its job is to *carry* the speech signals along with it! The process of combining the speech (or Morse code) with the carrier is called *modulation*.

Wider and wider

Sending Morse code is achieved simply by switching (or *keying*) the carrier on and off. The bandwidth of the transmitted signal is only about 100 Hz. Speech, with its reduced bandwidth of 3 kHz, will produce a *single-sideband* (SSB) transmitted signal with a bandwidth of 3 kHz. If the same speech signal were used to produce an *amplitude-modulated* (AM) signal from the transmitter, it would have a bandwidth of about 6 kHz. Perhaps you can now understand why the bandwidths needed to produce hi-fi broadcasts need to be so large. TV signals need bandwidths running into tens of megahertz!

Waves need aerials

Radio waves are produced whenever changing currents flow through a wire, and when that wire is made in such a way as to maximise the *radiation* from the wire, it is called an *aerial* or *antenna*. The same piece of wire will receive radiation from other aerials; an aerial will transmit and receive. This is an important property of the aerial: when a current flows through it, electromagnetic waves are launched into the air; when electromagnetic waves in the air encounter the aerial, currents are produced in it.

From the simplest transmitting aerial, waves travel in all directions, like the waves on the pond that we considered in Part 1. They will travel a long way through air and space before they become too week to be received. They won't travel very far into the earth, however! The earth will *reflect* some of the wave and will *absorb* the rest. That portion of the wave which is reflected will again travel through air and space until it is totally attenuated.

Look at **Figure 1**; A represents a radio transmitter, with B1, B2 and B3 being receiving stations. The two arrows pointing 'downwards' from A represent two of the waves from A which just graze the earth's surface. Waves above

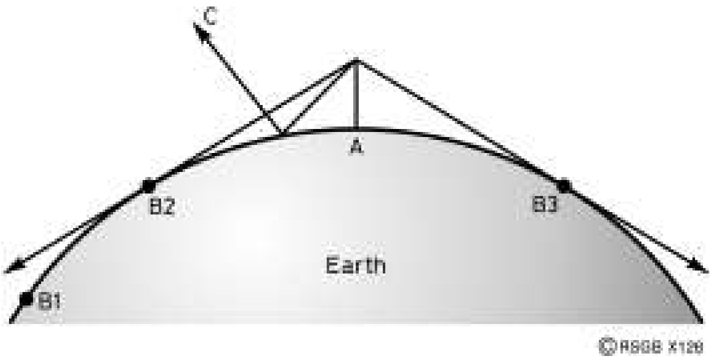

Figure 1 Stations between B2 and B3 receive the *ground wave*

these will travel on into space; waves below them will either be absorbed and reflected by the earth or received by aerials. B1 will not receive any signals from A, because it is below A's horizon. B2 and B3 can receive A's signals because they are *just* on A's horizon. Any stations between B2 and B3 will also receive A's signals, which are known as *ground-wave signals*. The wave at C represents one which is reflected by the ground and travels into space.

This description begs the question of how signals are received from stations well beyond the ground-wave range.

Mirrors in space

Suppose that there was something, out in space, that would reflect radio waves. Waves from A that travel out into space could be reflected off it and return to earth, enabling stations such as B1 to receive A's signals. The situation just described is illustrated in **Figure 2**.

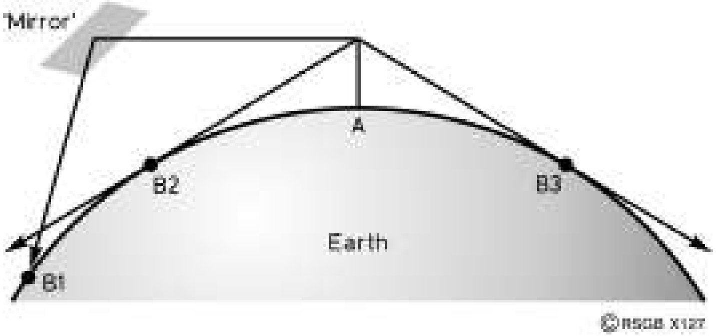

Figure 2 Reception of A's signals at B1 using the *sky wave*

Just such a mirror in the sky really *does* exist. It is not man-made, of course! The earth is surrounded by the atmosphere, a mixture of many gases, such as nitrogen and oxygen. The energy contained in the radiation from the sun is more than sufficient to *ionise* these gases, thus making them into electrical conductors. When a gas is ionised, some of its electrons are physically stripped out of the atoms and are free to move about, just as the electrons of a metal do in a wire. Consequently, we can regard this part of the atmosphere (the part illuminated by the sun) as acting like a sheet of metal, which *reflects* radio signals! It is not a perfect reflector, but is sufficient to produce long-range (DX) radio propagation via the *sky wave* under the right conditions. (A more down-to-earth example of ionised gases conducting electricity can be found in the fluorescent tube and the neon sign – many gases glow when they are continuously ionised.)

This conducting region at the extremity of the atmosphere is called the *ionosphere*, and it exists in layers between 60 km and 700 km above the earth's surface. When the ionosphere is sufficiently ionised, it glows; this is the natural phenomenon known as the *aurora borealis*, or the *northern lights*.

The property of the ionosphere that enables radio waves to be reflected does not act in a uniform way; it is very selective about which waves it reflects, and which waves go straight through it and into outer space. In general, it reflects only those waves with frequencies below about 30 MHz – the HF bands!

In Part 3 we will look at families of waves.

12 An LED flasher

Introduction

The LM3909 is an integrated circuit (IC) which will flash a light-emitting diode (LED). Using only two extra components and a battery, the circuit is cheap and has a very low current drain from a 1.5 V cell. The circuit can be used as a novelty flasher, an indicator for a dummy alarm bell box, or it could be attached to a torch so that it could be found easily in the dark! The simple circuit is shown in **Figure 1**.

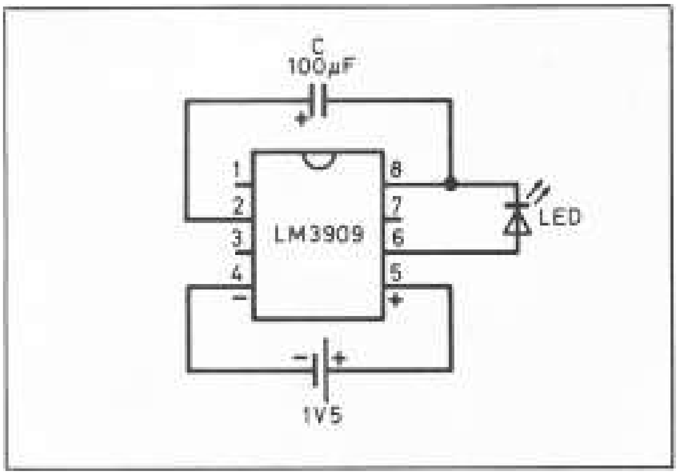

Figure 1 Circuit diagram of the LED Flasher. Pins 1, 3 and 7 of the IC are not used

Assembly

The circuit can be built on a small piece of Veroboard (the piece shown in **Figure 2** measures 15 holes by 10 strips). Using such a board, follow these instructions.

1. Depending on how far away you want the LED from the circuit board, solder a length of insulated wire to each lead of the LED. Use different colours of insulation – say, red and black, connecting the *red* lead to the *anode* (a) lead (the longer one) of the LED, and the black one to the *cathode* (k). Figure 2 shows these leads.
2. Cut the copper tracks as shown in Figure 2, using a 3 mm (⅛ inch) diameter drill, rotated between thumb and forefinger, or use the proper tool. Make *absolutely* sure that the tracks are completely broken!
3. Fit the IC holder in the correct position, using the cut tracks as guides, and make sure the small notch is facing towards the top of the board. Solder the pins to the copper tracks.
4. Mount the capacitor, positive end to the left, so that the positive lead is soldered to track F, which connects it to pin 2 of the IC; the negative lead is soldered to the right-hand side of track E, this being connected to pin 8 of the IC.
5. Solder on the battery leads, positive to the right, and the extended LED leads, positive downwards.

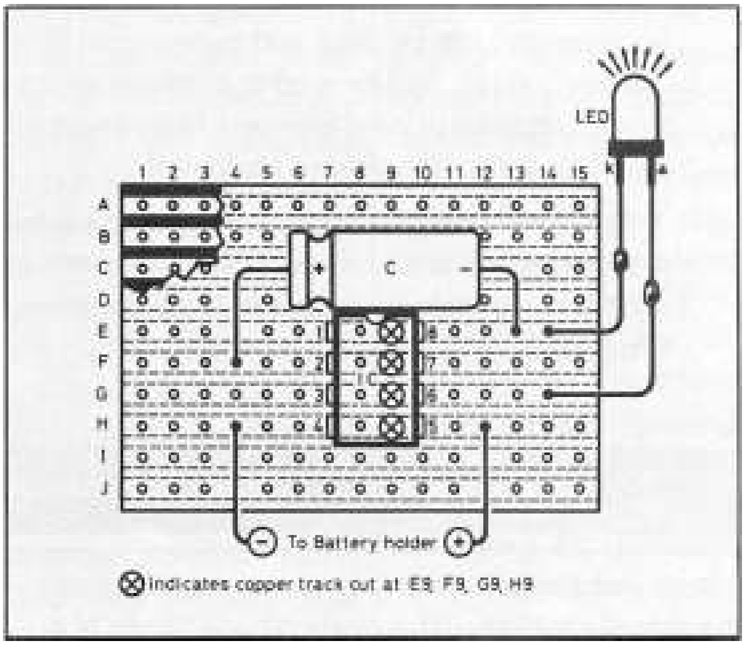

Figure 2 Board layout viewed from the component side. The tracks are cut under the board where shown

6. Check the circuit, and hold up the board to a bright light and look carefully for solder bridges between the tracks and pieces of copper swarf which may have escaped your inspection in 2 above! Remove whatever you find.
7. When all seems well, put the IC into the socket, ensuring that the notch or dot on the upper surface of the IC lines up with the notch on the holder. Line up each pin on the IC with the hole below it before pressing *gently* on the IC with the board supported on a firm surface.
8. Connect the battery; the LED should start to flash. The circuit is complete and working!

If you prefer, the whole circuit (battery included) can be mounted in a small plastic box, with the LED mounted on a clip and protruding through the panel. There are many other possibilities, and it is up to you to find an application for your own use.

Parts list

		Maplin code
LM3909	Integrated circuit	WQ39N
IC socket	8-pin DIL	BL17T
LED	5 mm diameter	WL27E
100 microfarad (μF)	Electrolytic capacitor (10 V)	FB48C
Battery holder	For AA-size cell	YR59P
Battery	1.5 V AA cell	

Small piece of Veroboard (15 holes by 10 strips)
Small plastic box (if required)
LED clip (if required)
Two lengths of coloured, insulated wire for LED (as required)

Availability

All parts can be obtained from Maplin Electronics Ltd.

13 Waves – Part 3

Introduction

When we talk about the spectrum being divided up into bands, this is just for our convenience; there are no natural divisions although, as we have seen, some of the properties of different bands really *are* different! Let's look at **Figure 1**, and see how the frequencies are divided up.

The divisions

- **Very low frequencies** (VLF) cover the range from a few kilohertz up to 30 kHz. Very long-range communication is possible, but at *very* small bandwidths. It is used for special purposes.
- **Long waves** (LW) are used for medium-distance commercial broadcasting and have frequencies from 30 kHz to 300 kHz.
- **Medium waves** (MW) are used for commercial broadcasting, and use frequencies from 300 kHz to about 1.5 MHz (1500 kHz). Typical range is about 200 km.
- **Short waves** (SW) encompass both the low-frequency (LF) and high-frequency (HF) amateur radio bands. There are nine narrow amateur bands in the SW spectrum between 1.8 MHz and 30 MHz. Some of these bands give round-the-world communication.
- **Very high frequencies** (VHF) span the range between 30 MHz and 300 MHz. Relatively short-range communication is possible. They were once used for broadcast TV before it moved to UHF. There are now three

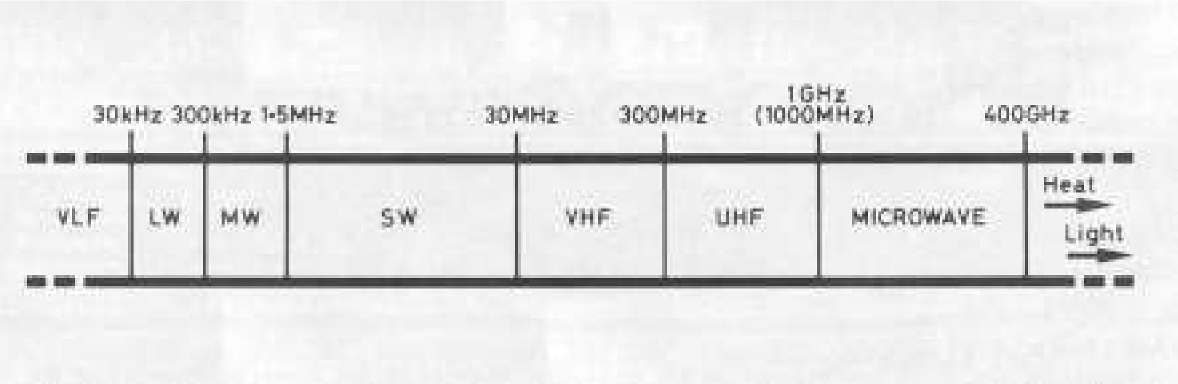

Figure 1 Diagram of radio frequency spectrum

amateur bands here – 6 m, 4 m, and 2 m. Repeaters are used to extend the usable range of mobile stations. VHF waves are not usually reflected by the ionosphere, but when they are, ranges of several thousand kilometres are possible. Weather affects these waves on a regular basis, however. In addition to amateur users, the VHF part of the spectrum is also used by the police, the fire and ambulance services, weather satellites, and many others,

- **Ultra-high frequencies**, sometimes called *centimetre waves*, cover the range from 300 MHz to 1000 MHz (or 1 GHz). The only amateur band in this range is the 70 cm band, and we share it with radar, TV and cellular telephone users as well.
- **Microwaves** begin at 1 GHz and extend to about 400 GHz. They are *never* reflected by the ionosphere, are partially attenuated by buildings, and are reflected from aircraft and cars. Microwave absorption in the atmosphere is quite significant, and rain and fog can attenuate microwaves quite heavily.
- **Heat, light** . . . Above 400 GHz we run into the infra-red bands and on into the visible light and ultra-violet bands. We generally take 400 GHz as being the limit of what we class as radio waves.

Bandwidth again

Complex signals need more bandwidth than simple signals. Even if it were possible, we would not be able to transmit a single television channel in the whole of the MW broadcast band! When TV used part of the VHF band, only five channels were possible in the range from 45 MHz to 68 MHz. By moving TV to the UHF band, we now have 47 channels between 470 MHz and 855 MHz!

It’s your choice!

The number of permutations you have amongst all the modes and all the bands is enormous! Only you can decide what you are interested in and what you want to learn about. That is the attraction of amateur radio!

14 Choosing a switch

Simple, but be careful!

A switch is the simplest electronic component, but care is needed in choosing the correct one for the job. Here is a list of the main characteristics of a switch, which will help you to select what you want.

1. **Rating**. This gives the maximum voltage and current that a switch can handle. For example, a 250 V 1.5 A switch will switch mains voltage at a current not exceeding 1.5 A. If the current is greater than 1.5 A, the switch may get hot and fail. If the voltage is too great, the switch may arc each time it is switched off, thus wearing away the contacts.
2. **Number of poles**. A single switch can control many circuits; the number of poles tells you *how many* different circuits it can handle. See **Figure 1**.
3. **Number of throws**. This tells you the number of positions each pole can have. This is best illustrated in Figure 1. The simplest ON/OFF switch is a Single-Pole, Single-Throw (SPST) switch. A Single-Pole,

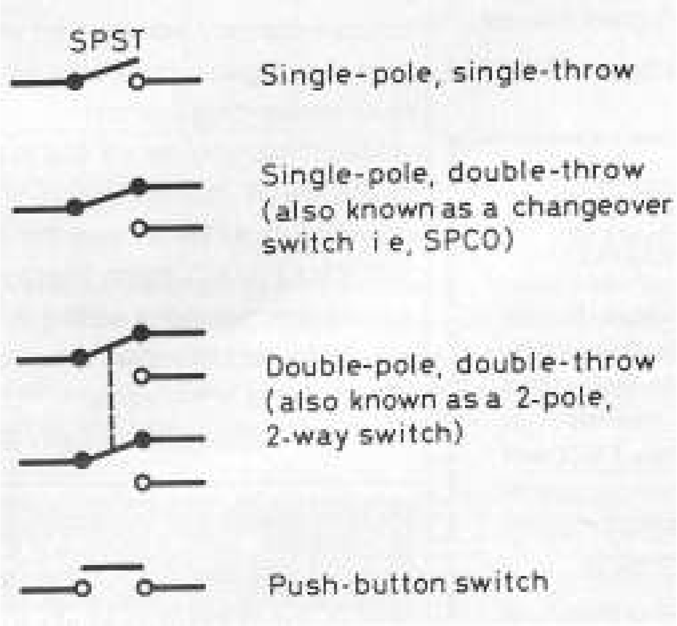
SPST
Single-pole, single-throw
Single-pole, double-throw (also known as a changeover switch i e, SPCO)
Double-pole, double-throw (also known as a 2-pole, 2-way switch)
Push-button switch

Figure 1 Some circuit symbols for different types of switch

Double-Throw (SPDT) switch may also be used as an ON/OFF switch, but is used mainly to change between two parts of a circuit, and is commonly known as a Single-Pole Change-Over (SPCO) switch. Two or more SPCO switches can be operated at once; Figure 1 shows an example. The two switches are said to be *ganged*. See also 'Types of switch' below.

4. **Number of ways**. When a switch has more than one throw, we tend to use the word 'ways' instead. This means that if a switch has one pole and six throws, we would normally call it a '1-pole 6-way' switch. Such switches tend to be rotary switches, as are described below.

Types of switch

(a) **Push-button**. These are found on calculators, telephones, electronic games and most equipment with a digital display.

(b) **Rotary**. These are switches controlled by a knob, and are turned instead of moved up and down. **Figure 2** shows the rear of such a switch and its circuit symbol. The commoner types of rotary switch are: 1-pole, 12-way; 2-pole, 6-way; 3-pole, 4-way; 4-pole, 3-way; 6-pole, 2-way. All these switches have 12 click positions, as you may have guessed, but each one comes with an adjustable end-stop so that you can set the correct number of ways according to the contacts on the switch.

(c) **Slide**. This switch is common on the cheaper types of radio, mainly as an ON/OFF or band-changing switch. They are not very rugged, but are small and cheap to produce. Very small types are manufactured for use on PCBs.

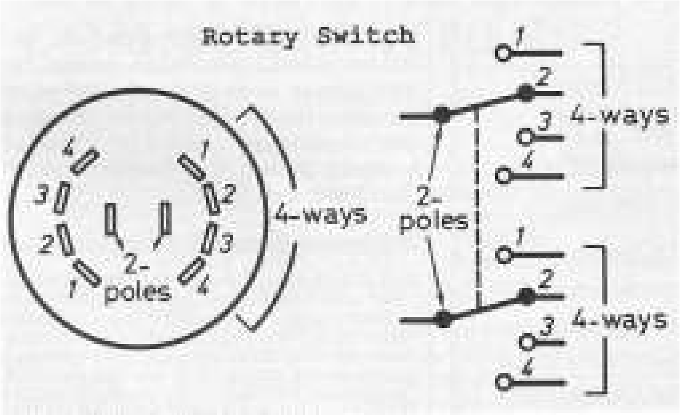

Figure 2 The rotary switch can select several different circuits at once

(d) **Toggle**. Available as two-way types or three-way, with a centre-off position. Can be manufactured to handle very high currents.

(e) **Micro-switch**. These are devices which are usually operated indirectly, such as when the cover is removed from a high-voltage power supply, or when the door of a fridge is opened. The 'micro' part of the name doesn't refer to the size of the switch, but to the *small* movement that is required to activate it.

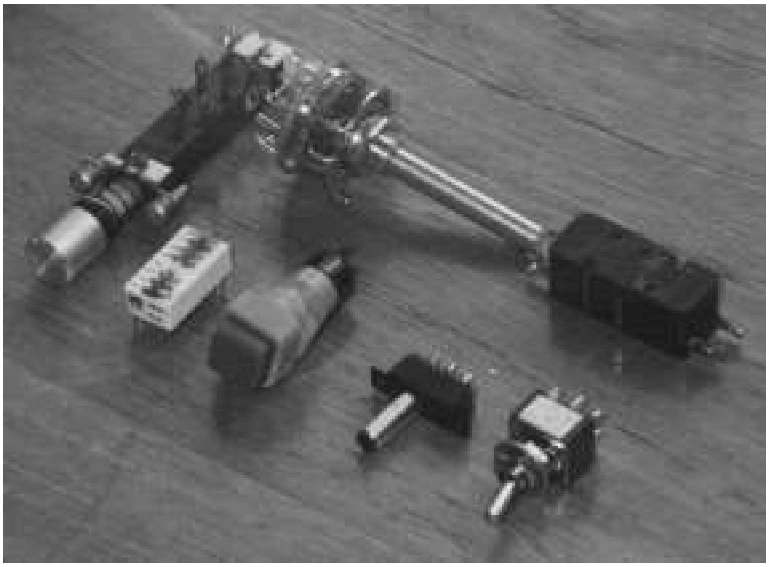

15 An aerial tuning unit for a receiver

Introduction

Any length of wire will act as an antenna (or aerial) but, to get the best results from a transceiver or receiver, an aerial tuning unit (ATU) is required. This matches the *impedance* of your aerial to the impedance at the aerial socket of your radio. Impedance is like resistance, and is measured in ohms, but it is used for alternating currents, and hence is common in audio and RF engineering. Most receivers have an impedance at the aerial socket of about 50 ohms (Ω); aerial impedance, on the other hand, can be anywhere between 20 Ω to over 1000 Ω, depending on its length and its height above ground.

Fortunately, you don't *need* to be able to calculate your aerial's impedance; all you need is a device that will perform the matching operation for aerials with a large range of impedances, and this is exactly what is described here! The subject of *matching* is a complex one, but all you need to know is that most signals will become clearer, and that there will be less noise and interference. Stations will become louder, so you will probably be able to reduce the setting of your RF gain control (always a good thing to do).

This design of ATU covers all amateur and broadcast bands from 10 m (28 MHz) to 80 m (3.5 MHz), and is very easy to build. The circuit is shown in **Figure 1**.

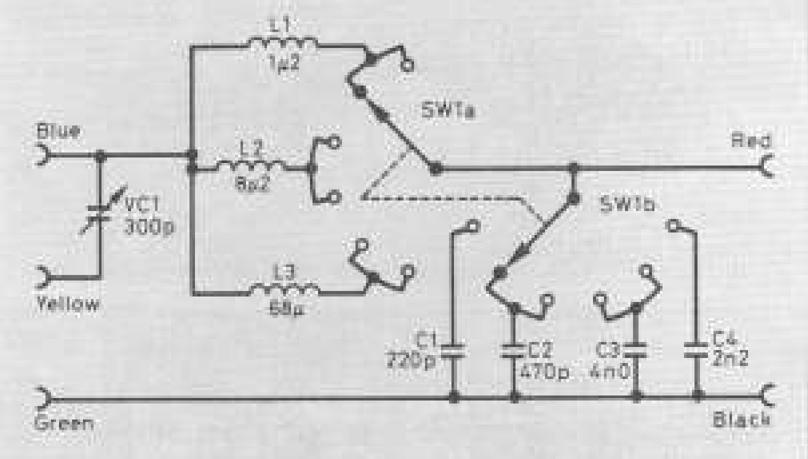

Figure 1 Circuit diagram of the Antenna Tuning Unit, showing the use of a 2-pole, 6-way rotary switch to select inductors (L) and capacitors (C)

Construction

1. Firstly, you will need a simple plastic case in which to house the ATU. The size should be approximately 85 × 145 × 50 mm.
2. Start by drilling two 10.5 mm holes in the front of the case; these are for the 6-way switch and the tuning capacitor.
3. Drill three 8 mm diameter holes in the left-hand side of the box, for the three sockets, coloured blue, yellow and green.
4. On the right-hand side of the box, drill two 8 mm diameter holes for the red and black sockets.
5. Now, fit the 6-way switch (SW1), the tuning capacitor (VC1), and all the sockets to the case. Check that the vanes of the capacitor rotate smoothly when the shaft is turned.
6. Wire up the inductors (coils). Figure 1 and the wiring diagram of **Figure 2**, will help with this. As you can see, each one side of each coil is connected to two switch connections, the other end going to VC1.
7. Solder in the fixed capacitors. One end of each goes directly to the ground socket (black), and the other end goes to the switch.
8. Solder a wire between the green and black sockets. The output from the ATU comes from the red socket, and this is connected to the two tags in the centre of the switch, as Figure. 2 shows clearly.
9. Finally, connect the blue and yellow sockets to the tuning capacitor, and the ATU construction is complete.

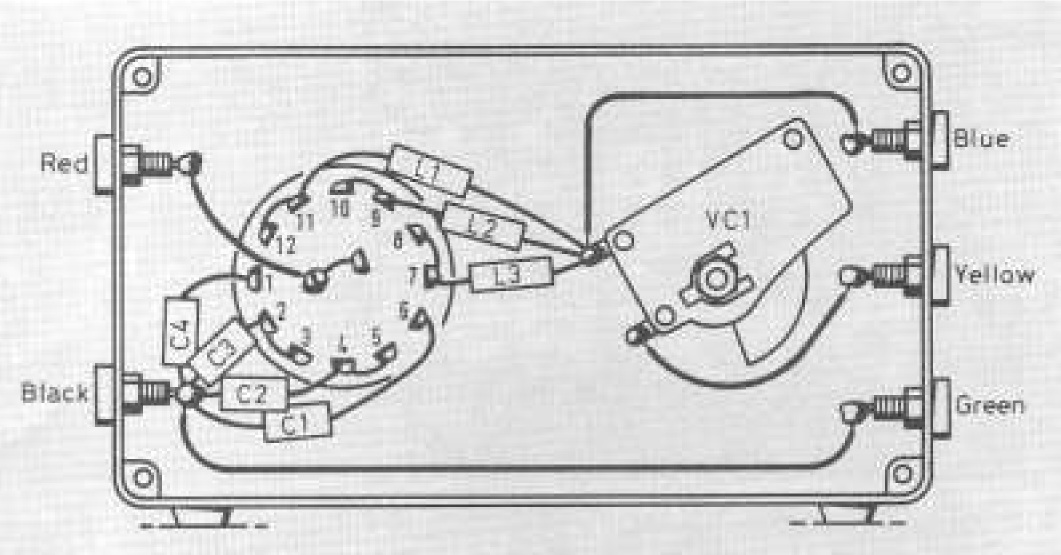

Figure 2 The internal view of the case shows the main tuning capacitor, VC1. This is a solid dielectric type, which has adjustable brass plates

In practice . . .

Figures 3 and **4** show two different ways of connecting your aerial to your ATU. In each case you will need to select each switch position in turn, and rotate the tuning capacitor through its full range while listening to a station. You should find that one switch position enables VC1 to produce a peak in the signal strength in the loudspeaker. At this point, your aerial and receiver are said to be *matched*. Stations in the same band will probably peak with VC1 at the same setting of SW1, but different bands will almost certainly require different positions of SW1.

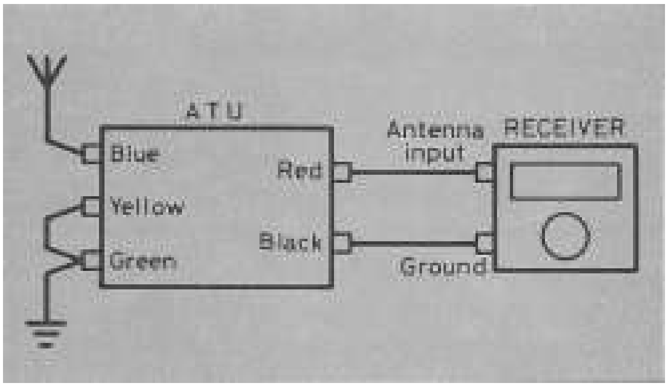
ATU
Blue
Yellow
Green
Red
Black
Antenna input
RECEIVER
Ground

Figure 3 You may find that parallel tuning gives best results with your antenna

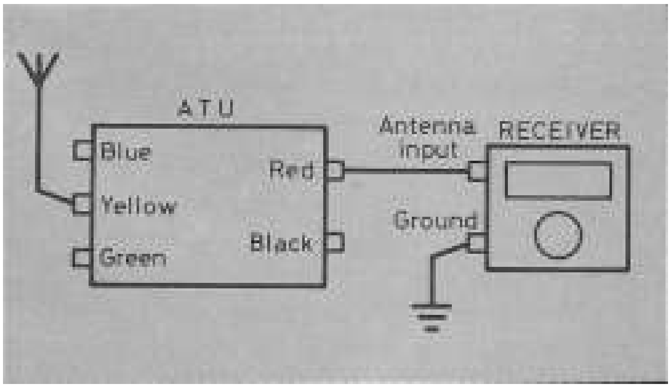
ATU
Blue
Yellow
Green
Red
Black
Antenna input
RECEIVER
Ground

Figure 4 In other cases, series tuning could be the most effective arrangement

Parts list

Capacitors (all rated at 16 V or more)

C1	220 picofarads (pF) polystyrene
C2	470 picofarads (pF) polystyrene
C3	1000 picofarads (pF) or 1 nanofarad (nF) polystyrene
C4	2200 picofarads (pF) or 2.2 nanofarads (nF) polystyrene

Inductors

L1	1.2 microhenries (μH)
L2	8.2 microhenries (μH)
L3	68 microhenries (μH)

Switch

SW1	2-pole 6-way rotary

Sockets

4 mm type, one each of red, black, yellow, blue, green

Additional items

Plastic or metal case, e.g. Maplin type YU54
Two large knobs for SW1 and VC1

16 A simple 2 m receiver preamplifier

Introduction

Designed specifically to complement the modified air-band portable (also described in this series), this can be used with some success on many receivers suffering from 'deafness' on VHF.

The circuit

An RF preamplifier is a device which improves the input signal to an existing receiver, enabling it to work more effectively. Because of the noise which is added to the signal by the preamplifier, *very* weak signals may not be usefully enhanced, but stronger signals will be improved considerably.

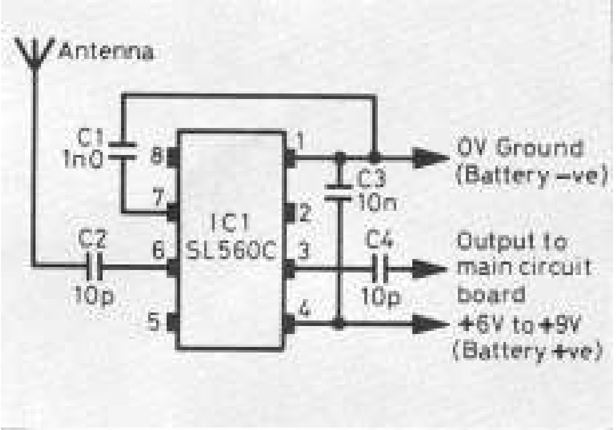

Figure 1 The integrated circuit is mounted upside down. Make sure you identify pin 1 which has a small dot next to it

This little circuit (shown in **Figure 1**) uses a GEC/Plessey integrated circuit type SL560C. With the addition of four capacitors, it is used between your aerial and the aerial input of the radio's PCB.

Putting it together

1. Use a small piece of prototype (matrix) board about 25 mm square. Use an 8-pin DIL socket for the integrated circuit (don't risk soldering the chip – it is seldom a risk worth taking). Figure 1 shows the connections to the socket, *looking from underneath*.
2. Make special note of the pin numbers, so that you know how to put the chip into the socket when you have finished. The positive and negative connections to the circuit are taken from the main PCB *after* the ON/OFF switch – so that the switch operates the preamplifier, too.
3. Unsolder the lead to the radio's telescopic aerial and connect it to the free end of C4, as shown. Then solder a short lead between the telescopic aerial and C2.
4. Sometimes it is possible to cajole your little preamplifier into the radio's plastic case, provided there is room and that you make sure that none of the soldered joints on your little board touch any of the metal inside the case when you replace the back and screw it on again.
5. If there is not enough room inside, then put the preamplifier into its own box, with battery and switch, and its own aerial. Keep the lead from the preamplifier to the aerial connection of the radio as short as possible – perhaps using screened cable.

Because your preamplifier is untuned, you will find not only that it helps with reception on 2 metres, but also that reception on the FM broadcast band is improved!

Parts list

Integrated circuit	
IC1	GEC/Plessey SL560C
Capacitors	
C1	1 nanofarad (nF) ceramic
C2, C4	10 picofarads (pF) ceramic
C3	10 nanofarads (nF) ceramic

Additional item

Prototype broad approx. 25 × 25 mm

17 Receiving aerials for amateur radio

Introduction

For any radio receiver to work well, it must have some form of antenna, or aerial. In almost all domestic transistor radios, the aerial is built into the set, either as a ferrite rod (which looks like a rod of dark grey metal) on which are wound coils of wire, or as a chromium-plated telescopic metal rod. Some radios have both forms of aerial, using the ferrite rod aerial for long waves (LW) and medium waves (MW) and the metal rod for very high-frequency (VHF) stations using frequency modulation.

There's broadcast reception . . .

No aerial is perfect, and these two types are *far* from perfect! As in most mass-produced equipment, *they serve their purpose*, which is not critical, and they allow the radio to be carried around easily, because they are not big and bulky. The broadcast stations are very powerful, usually quite close, and the circuits in the radio are quite sensitive, so the need for large aerials disappears.

. . . and there's amateur DX

Here we have an application which, in comparison with the broadcast situation, couldn't be much more different. The stations do not use much power, they may be half the world away, and the requirement for good aerials and very sensitive receivers is paramount. The aerials must be large and they must be as high as possible, away from buildings and trees, which cause reduction in signal strength, and away from man-made sources of interference.

You may not have thought about this but, in general, the larger an aerial becomes, the longer the cable (or 'feeder') must be in order to reach your shack and the receiver inside it. Cables reduce the received signal, so what your aerial gains by being large, the feeder (if you're not careful) will lose!

A simple aerial . . .

We now know that a 'good' aerial is essential. But what is a 'good' aerial? It depends on your purse, your property and your enthusiasm! One of the simplest (and, incidentally, one which is *not* subject to the cable loss problem discussed above) is the **Long Wire**, shown in **Figure 1**. This is, quite literally, a long piece of wire going from a chimney stack to a tree or pole

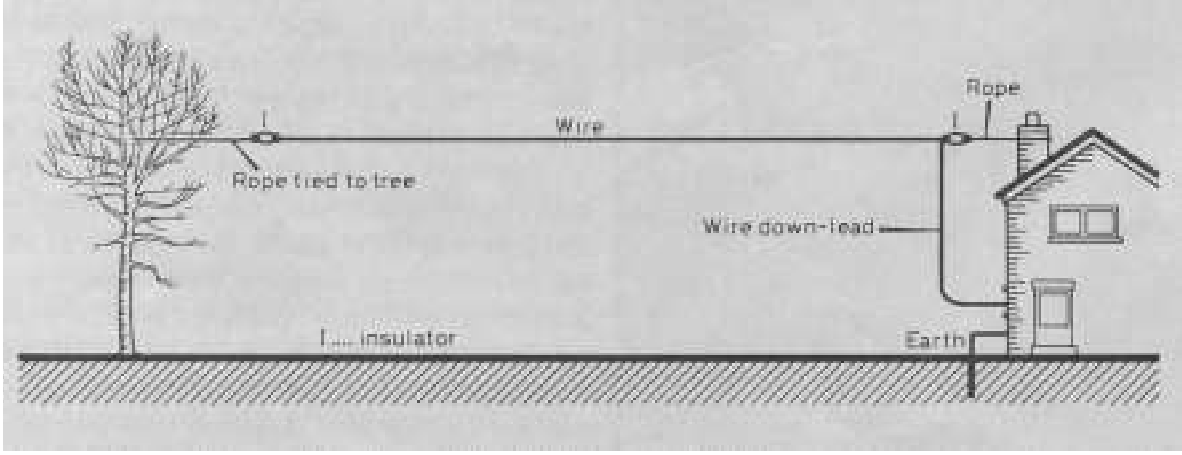

Figure 1 A tree at the bottom of the garden can provide a useful antenna support. Insulators can usually be obtained from Tandy stores or TV aerial suppliers

at the end of the garden. The longer it is, the better. Notice that the wire itself is *not* used to loop around the tree or chimney. Rope is used for both, and is secured to the end of the wire using an insulator of the 'dog bone' or 'egg' variety, to be found in profusion at rallies.

The wire is brought into the house through a window into the shack where the receiver is situated. The long wire aerial is best used with an aerial tuning unit (ATU), which is also described in this book. One of the advantages of the long wire is that it can be used on several frequency bands.

. . . and more complicated ones

Next up the ladder of complexity is the **dipole** (meaning two poles, or two elements). One form of dipole for lower frequencies would take the same basic form as the long wire, except that the feed to the receiver is taken, not from the end, but from the centre. The wire is essentially cut into two halves, and the two ends at the centre are connected to one end of a coaxial cable, which is then taken to the receiver. A smaller form of this is shown in **Figure 2**, which is conveniently mounted in a house loft. Aerials should always be mounted outside for best results, but will work when mounted inside, and the loft space is the logical situation.

The total length of the dipole should ideally be one-half of the wavelength of operation – hence the term 'half-wave dipole'. For example, a dipole for use on 20 m should be about 10 m long. Dipoles are thus 'single-band' aerials. They can be modified for use on several bands, and then become known as 'trapped dipoles', having coils and capacitors at certain points along their lengths.

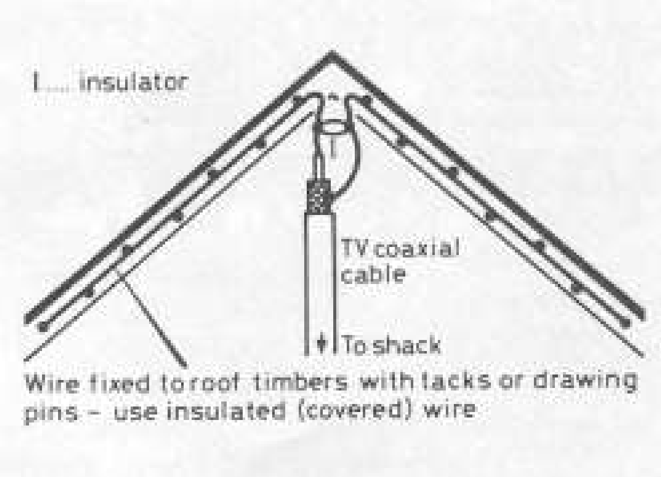
I..... insulator
I
TV coaxial cable
To shack
Wire fixed to roof timbers with tacks or drawing pins – use insulated (covered) wire

Figure 2 An indoor antenna should be mounted as high as possible

Indoor aerials

For flat- or apartment-dwellers, an indoor aerial is often the only solution, in which case the dipole of Figure 2 may be of great use. Other popular types include the **loop aerial**, of which several types are shown in **Figure 3**. The shape of the loop is of secondary importance, but it should be as large as possible.

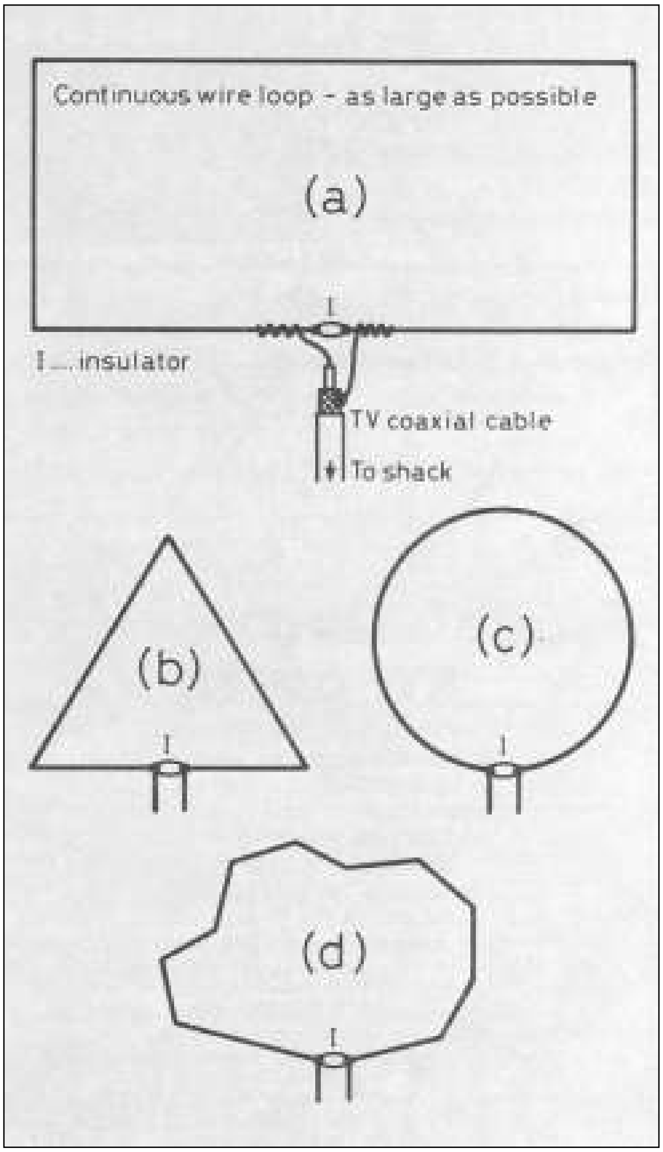
Continuous wire loop – as large as possible
(a)
I
I ... insulator
TV coaxial cable
To shack
(b)
I
(c)
I
(d)
I

Figure 3 Try different shapes of loop antenna to see which works best

Ingenuity is called for when considering indoor aerials. A dipole can be mounted along a picture rail or pelmet. A vertical loop of aluminium foil can be stuck to an inside wall and covered with wallpaper! The designs are limited only by your imagination, and will give you hours of listening pleasure

Outdoor aerials

If you want to venture beyond the simple long-wire discussed earlier, then the erection of outdoor aerials can be a bit tricky; they can be more than a one-man job, and can involve seeking planning permission. Get some help advice from someone who has done it before – what about your local radio club? Plenty of designs for aerials appear in *RadCom*, so you should never be short of ideas!

18 The Colt 80 m receiver – Part 1

Introduction

This is the first of four parts detailing the design and construction of a simple radio capable of good reception of amateur radio signals on the 80 metre band. If you can solder and have some basic hand tools, at the end of Part 4 you will have a working receiver, in which you can take pride, and start some serious listening on 80 m! For testing, all you will need is another receiver and a multimeter.

Description

The radio will be built in three modules, or stages, as illustrated in **Figure 1**. Each of these is built on a printed-circuit board (PCB) or matrix board and can be tested in its own right. The case will need some holes drilled, and care will be needed to mount the PCBs in the case.

Get to it!

Surprisingly enough, when you are building a radio section by section (which is *always* the best way), it is easiest to start at the output and work

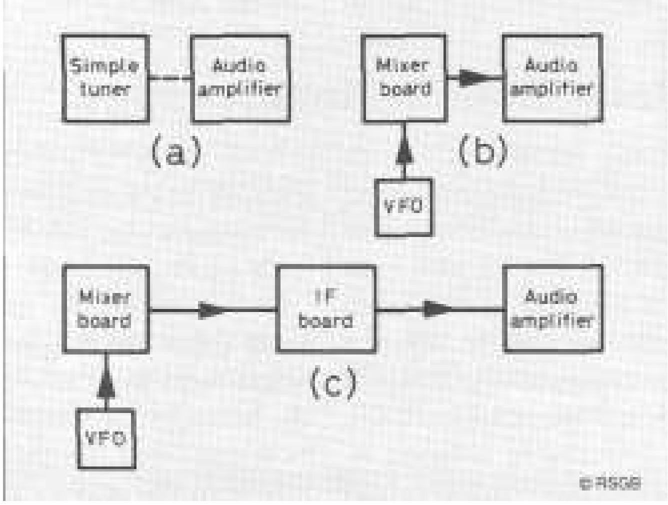
Simple tuner
Audio amplifier
(a)
Mixer board
Audio amplifier
VFO
(b)
Mixer board
IF board
Audio amplifier
VFO
(c)
© RSGB

Figure 1 Build your Colt and watch it grow. A simple crystal set (a) becomes a direct conversion receiver (b) and finally an 80 metre amateur band superhet (c)

towards the input, and you can test what you have done stage by stage. You will see what this means as you progress with your construction.

Every receiver needs some audio frequency (AF) amplification to make the sound signals big enough for you to hear. The circuit for the AF amplifier is shown in **Figure 2**. It uses the TDA7052 integrated circuit (IC) plus a handful of extra components. R1 in conjunction with C2 and C3 *decouple* the battery supply, preventing any audio signals getting through to it and affecting other parts of the radio, when they are connected. C1 acts to prevent high frequencies (above the 3 kHz bandwidth) going into the amplifier input. A volume control, VR1, is connected across the amplifier input, so that the amplifier can accept signal inputs over a wide range. **Figure 3c** shows the connections, which are made with screened cable. The

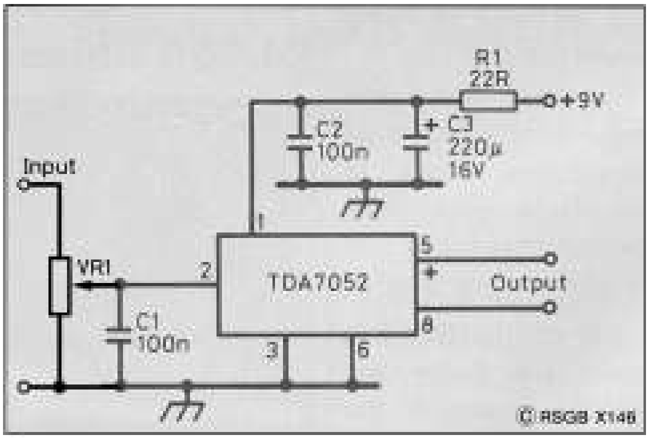
R1
22R
+9V
C2
100n
+ C3
220μ
16V
Input
1
5
+
VR1
2
TDA7052
Output
C1
100n
8
3
6
© RSGB X146

Figure 2 The Philips TDA7052 integrated circuit (IC) used in the audio amplifier needs very few extra components. It has a signal voltage gain of 100 times and the output is suitable for a loudspeaker or headphones

centre conductor of the left lead goes to the point marked 'input' in Figure 2, the braid being connected to the amplifier earth (0 V) tag. The right lead to VR1 goes to whatever signal source you have for testing – see later. VR1 is *not* mounted on the PCB.

The circuit is constructed on a small PCB or matrix board. Make sure that the electrolytic capacitor, C3, is soldered into the board the correct way – its positive and negative connections are shown in **Figure 3b** for reference. If you are **at all** concerned about soldering the IC into the board, enlist some help, *or*, obtain an 8-pin DIL socket, which you can solder in and then carefully insert the chip into the socket, making sure that is the correct way round. The markings on the chip are shown in **Figure 3a**.

The output leads from the amplifier go to a plastic 6.3 mm (¼ inch) mono jack socket, so that *neither* output lead is connected to the metal case. The amplifier will drive a pair of headphones or a small 8 Ω loudspeaker.

The battery leads must be the right way round also; the battery itself can be a PP3 or PP9, or you can use a small DC power supply.

Testing

First, check that all the components are in the right places, that your soldering is good, and that you have headphones or a loudspeaker connected. Set VR1 about halfway along its travel. Connect the battery.

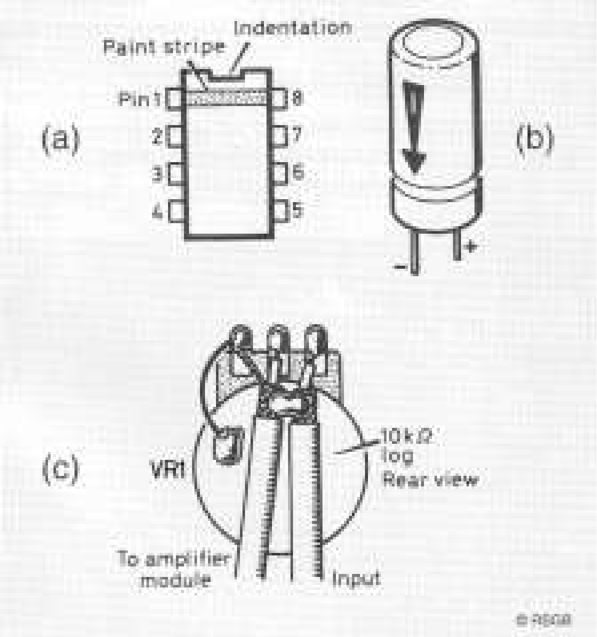

Figure 3 It is important to check the component connections carefully. The diagram shows (a) top view of the IC, (b) electrolytic capacitor and (c) volume control (VR1) connected across the input of the amplifier

A slight hissing noise should be heard; touching the input lead to the amplifier (the centre of the three connections on VR1) should produce a loud buzz. Touching the shaft at the same time will make the buzz quieter. This is the quickest way of confirming that your amplifier *seems* to be working. The only real test is to give it something meaningful to amplify! See the design of our *Crystal Radio Receiver* for full details.

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	22 ohms (Ω)
VR1	10 kilohms (kΩ) log
Capacitors	
C1, C2	0.1 microfarad (μF)
C3	220 microfarads (μF) electrolytic 16 V
Integrated circuit	
IC1	TDA7052 audio amplifier
Additional item	
PCB	(see below)
Component suppliers:	
Maplin	

The next part . . .

The metal case will be marked out ready to receive the completed modules.

19 A crystal radio receiver

Introduction

Here's a quick project to fill in a winter's evening! It was originally designed as a piece of test gear for the *Colt* receiver, but can be successfully used by anyone as a first radio project. Another use for it would be as test gear for any audio amplifier project requiring an audio input for test purposes; this is ideal, as it does not require any power supply!

Details

The initial rough-and-ready test for the audio amplifier of the *Colt* receiver will have whetted your appetite; you will want to prove more conclusively that your amplifier works, and in a way that others in the household will appreciate. Buzzing noises are not convincing in this respect! You are going to put together a simple crystal set – the simplest type of radio that there is – and use it as a signal injector for your amplifier. In this way, you build a real medium-wave (MW) receiver which drives a loudspeaker, as an intermediate product of the construction of an 80 metre amateur band receiver!

The circuit diagram is shown in **Figure 1**. It has a 60-turn coil mounted on a small piece of paper or card wrapped round a ferrite rod. The coil has a connection made to its centre-tap (the middle turn of the coil). The tuning capacitor, VC1, is the most expensive part of the circuit but don't worry, it will be used in the final design of the receiver also! Connect the diode, D1, from the centre-tap to the input to the potentiometer of the amplifier circuit of the *Colt*. Be careful to connect the aerial to the *vanes* of VC1, and not to its *frame*, or you will experience some strange effects when you are tuning. If your amplifier is working correctly, you should be able to receive local stations on medium-waves quite well.

If you think the crystal set will be of use to you in the future as a signal injector, all you will need will be another variable capacitor! If you do not intend to use the amplifier, a small crystal earpiece will allow you to listen. Walkman-type headphones will *not* work!

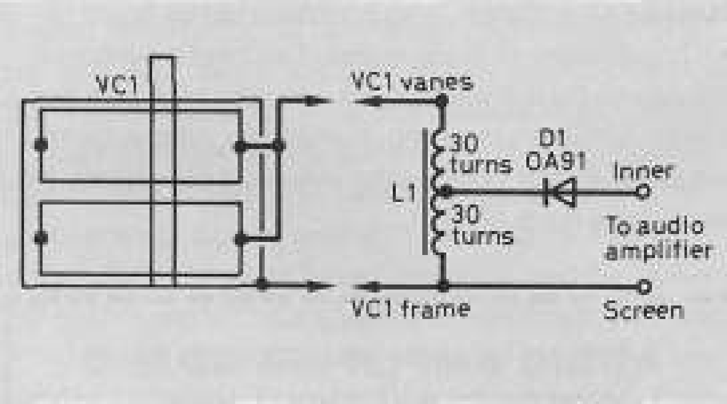
VC1
VC1 vanes
30 turns
D1
OA91
Inner
L1
30 turns
To audio amplifier
VC1 frame
Screen

Figure 1 Both sets of moving vanes are joined as shown. L1 is wound on a ferrite rod with 32 SWG wire and centre-tapped. A single winding (no tap) can be used, joining D1 to the top of the winding

Parts list

Variable capacitor

VC1	125 + 125 picofarads (pF) twin-gang

Ferrite rod

About 100 mm to 140 mm long

Diode

D1	OA91 germanium diode

Additional items

Wire	2 m of enamelled copper wire, between 23 SWG and 32 SWG
Earpiece	High-impedance crystal type (only needed if you are not using the amplifier)

20 The varactor (or varicap) diode

Introduction

Many of the circuits for receivers and transmitters presented in this series rely upon the variable capacitor as a means of tuning. Another method of varying capacitance (without any moving parts) is provided by the *varactor diode*, sometimes called a varicap diode. This is a component which changes its capacitance as the voltage across it is varied.

The details

Figure 1 shows how a varactor diode might be connected to demonstrate its operation. Its symbol is that of an ordinary diode, with a capacitor symbol next to it. A variable voltage is applied across it in such a way that the diode is *reverse-biased*. This means that virtually no current passes through it – the positive voltage is applied to the cathode. Varactors are cheaper than variable capacitors, and they are tiny in comparison, very suitable for today's miniature circuits. If A and B were connected across the tuning coil in a simple receiver (with a series capacitor to block the DC from the battery reaching the coil), the tuning operation would be accomplished by turning the knob on the 10 kilohm potentiometer.

Varactors are available with different values, from less than 20 picofarad (pF) for VHF applications to 500 pF for medium-wave radios. They are

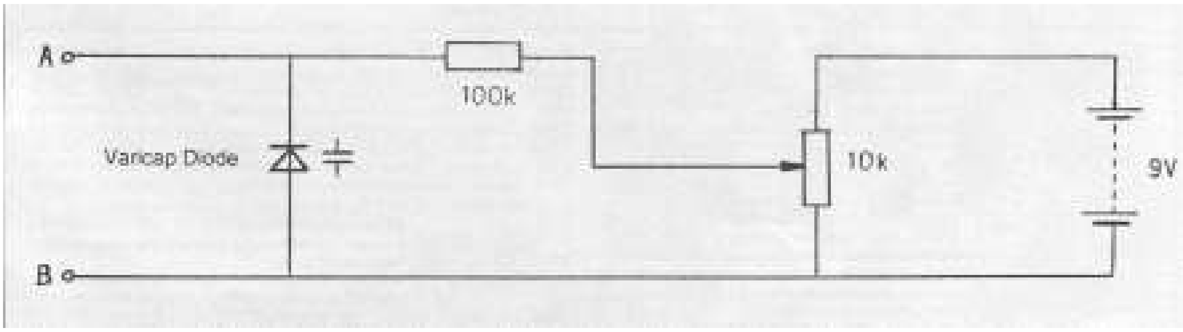

Figure 1 The capacitance of the varicap diode (between A and B) increases as the voltage is reduced, using the variable resistor

tuned usually by voltages between 2 V and 9 V. For a real application of varactors, you should consult the circuit diagram of the *Yearling* 20 metre receiver, elsewhere in this book.

In some circuit designs, several circuits are all tuned to the same frequency in order to improve the overall *selectivity* (the ability of the circuit to reject signals very close in frequency to the wanted signal). Special dual- and triple-varactors are available for circuits like this. Having been made at the same time from the same materials makes their individual characteristics virtually identical. Like all other diodes, they must be correctly wired into the circuit – their *polarity* is important.

Changes in temperature will cause the capacitance to change which, if it were part of an oscillator circuit, would cause the oscillator to drift – you would have to keep retuning the radio! This can be corrected by using a special integrated circuit called a *phase-locked loop* (PLL). Modern TV sets and satellite receivers use varactors and PLLs in this way.

Some useful varactor types

Type No.	*Tuning range*		*Description*
	pF/V	*pF/V*	
BB204B	42/2.0	15/12	Dual VHF
BB212	560/0.5	22/8	AM tuning
KV1235	450/2.0	30/8.5	Triple AM
KV1236	450/2.0	30/8.5	Dual AM
MV1404	120/2.0	9/10.0	HF tuning

21 A portable radio for medium waves

Introduction

The ZN415E integrated circuit (IC) can be used to make a very efficient AM portable MW broadcast radio with a built-in loudspeaker. Here’s how!

The circuit

Figure 1 shows the circuit diagram of the portable radio. It’s not as complicated as it may appear, especially after you have got started. L1 is a coil of wire mounted on a ferrite rod, acting as an aerial; VC1 is a variable capacitor which works, with L1, to tune in different stations. IC1 contains circuits of its own which boost the selected signal and it includes a *detector* which extracts the audio signal from the incoming RF signal. Earphones could be connected to the output of IC1 (between pins 4 and 5), but the output would not be powerful enough to drive a loudspeaker.

More sound

This is where IC2, an LM386 comes in. This is a small audio *power* amplifier which produces audio signals with enough power to drive a small

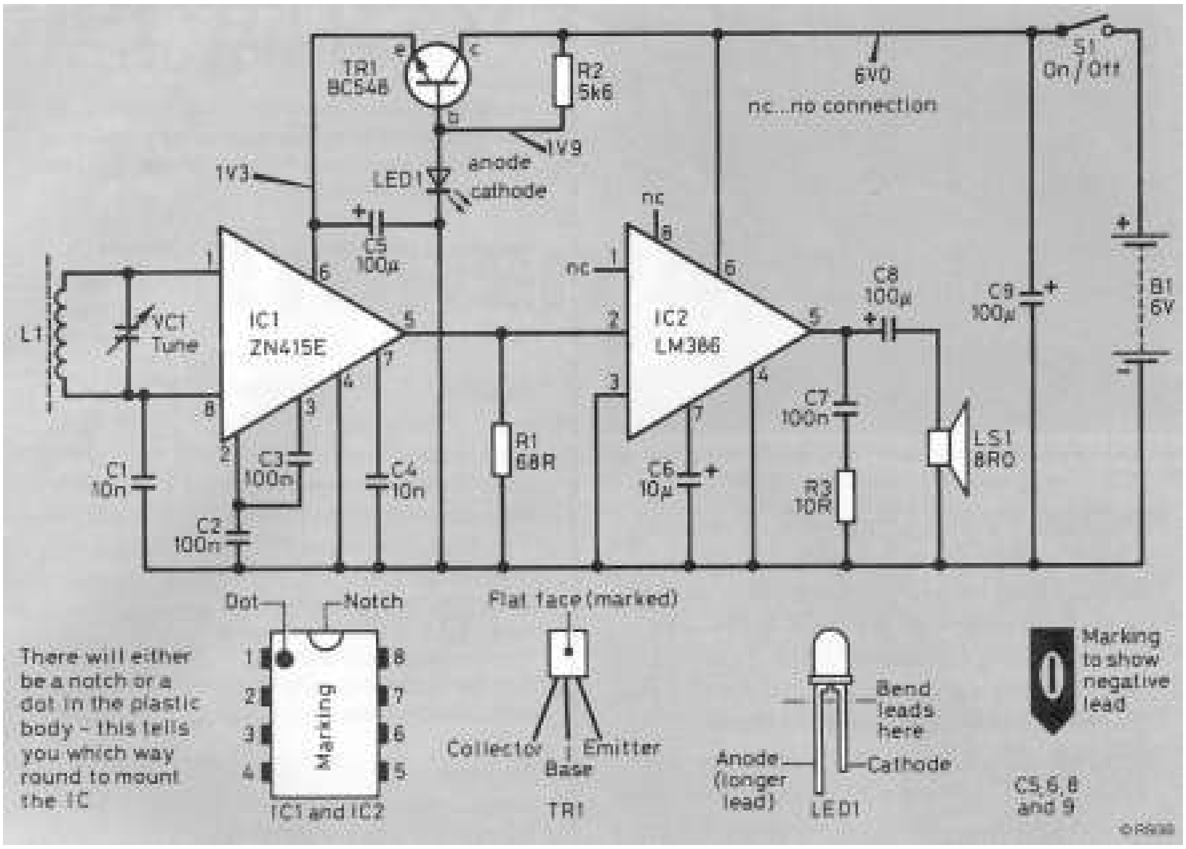
TR1
BC548
e
c
b
R2
5k6
6V0
nc...no connection
S1
On / Off
1V9
1V3
LED1
anode
cathode
C5
100µ
nc
8
nc
1
6
C8
100µ
C9
100µ
B1
6V
L1
VC1
Tune
IC1
ZN415E
1
6
5
7
4
3
8
2
IC2
LM386
2
5
3
4
7
C1
10n
C3
100n
C4
10n
R1
68R
C6
10µ
C7
100n
R3
10R
LS1
8R0
C2
100n
Dot
Notch
Flat face (marked)
There will either be a notch or a dot in the plastic body – this tells you which way round to mount the IC
1
2
3
4
8
7
6
5
Marking
IC1 and IC2
Collector
Base
Emitter
TR1
Bend leads here
Anode (longer lead)
Cathode
LED1
Marking to show negative lead
C5,6,8 and 9
© RSGB

Figure 1 The circuit diagram of our easy-to-build portable radio. Take care to mount the ICs and LED the correct way round

loudspeaker, LS1. The radio uses a 6 volt battery, which is made by connecting four 1.5 volt AA cells in series (4 × 1.5 V = 6 V) using a battery holder designed for this purpose. Although a 6 V supply is ideal for IC2, it is far too great for IC1, which needs only about 1.3 V. This lower voltage is provided from the 6 V supply by TR1 (an npn transistor), R2 and LED1 (a light-emitting diode). When current passes through an LED (see the description of the LED in this series) a reasonably constant voltage of 1.9 V appears between the anode and the cathode. Because of the voltage (0.6 V) that *always* exists between the base and emitter of a working transistor, the voltage on the emitter is about 1.9 V – 0.6 V = 1.3 V, and this is used as the power supply for IC1.

To keep the radio as simple as possible, no volume control has been fitted. Instead, you can use the directional properties of the ferrite rod aerial (see the information on ferrites in this book) to reduce the volume by rotating the set about a vertical axis using the handle provided.

Putting it all together

1. Start by covering the ferrite rod with Sellotape, or alternatively wrap a piece of paper tightly around it, and secure it with Sellotape. Then, with at least 2 metres of 24 SWG enamelled copper wire, wind 75 turns tightly around the rod. To be safe, leave about 50 mm of wire at the ends of the coil, then wrap the whole coil with Sellotape to hold the turns in place, leaving only the ends free. Then, using a small piece of sandpaper, remove the enamel from the last centimetre of each end of the coil.
2. Most of the components are mounted on a piece of Veroboard (the type with parallel copper strips on one side). The piece used on the prototype measured 32 holes by 10 strips, as **Figure 2** shows. Before you start fitting components, cut the copper strips as shown. It is easier to do it now than when the board is littered with components! The strips may be cut with a 3 mm (⅛ inch) twist drill rotated between thumb and forefinger. Resist the temptation to use a hand drill – the idea is just to cut the copper, **not** to drill right through the board!
3. Solder the IC sockets and the other components on the board as shown in Figure 2. Make sure that the IC sockets are fitted with their notches towards the *top* of the board, as viewed in Figure 2. Do not insert the chips yet. Always keep the wires left over from cropping resistors and capacitors, they will come in handy at times like this: make the wire links that are clearly shown in Figure 2. Connect the electrolytic capacitors (C5, C6, C8 and C9), the transistor and the LED the correct way round; then check it again when you have done it!
4. Finally, solder lengths of stranded insulated wire to act as ‘flying leads’ for future connection to L1, VC1, LS1, S1 and the battery connector.
5. Apart from the battery holder, everything is mounted on the case *lid*. This makes assembly and testing much easier, and eases fault-finding if

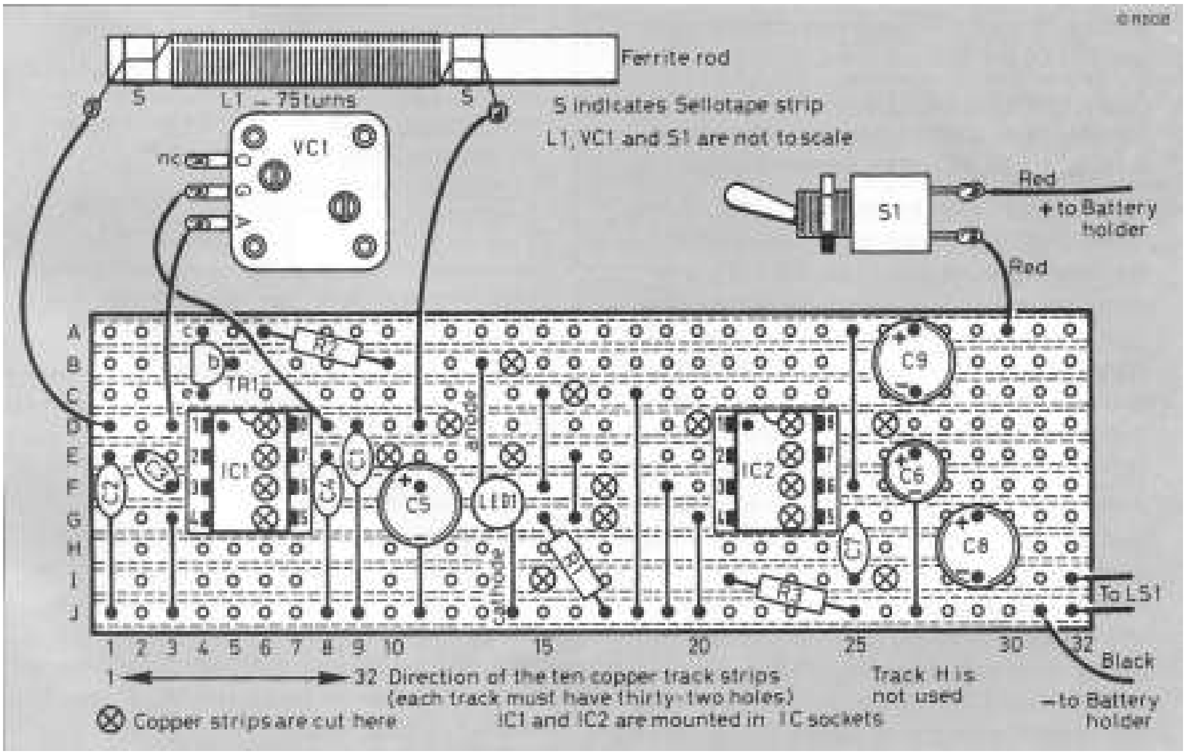
©RSGB
Ferrite rod
S
L1 = 75 turns
S
S indicates Sellotape strip
L1, VC1 and S1 are not to scale
VC1
nc
O
G
A
S1
Red
+ to Battery holder
Red
A
B
C
D
E
F
G
H
I
J
c
b
e
TR1
R2
IC1
C1
C2
C3
C4
C5
anode
cathode
LED1
R1
IC2
R3
C7
C9
C6
C8
To LS1
1 2 3 4 5 6 7 8 9 10
15
20
25
30
32
Black
− to Battery holder
1 ← → 32 Direction of the ten copper track strips (each track must have thirty-two holes)
Track H is not used
⊗ Copper strips are cut here
IC1 and IC2 are mounted in IC sockets

Figure 2 Veroboard layout for the Portable Radio. Make sure that all the wire links are included

the need arises! At the speaker position, make one large hole or a series of small holes to let the sound out. The ferrite rod may be stuck to the lid, as may the loudspeaker. Drill holes of the correct size to fit the particular types of variable capacitor (VC1) and switch (S1) that you are using. The Veroboard may be held in position by Blu-Tack or double-sided sticky tape.

6. Before inserting IC1 and IC2, connect the battery and switch S1 on. The LED should glow dimly (you may have to shield it with your hand in order to see it). If you have a test meter, check that there is about 1.3 V between pins 6 and 4 of IC1. If the reading is around 6 V or there is no glow, you may have connected the LED the wrong way round! When everything seems normal, switch off and disconnect the battery. Insert IC1 and IC2, making sure that the pins are straight and lie immediately above their corresponding holes in the sockets, and that the notches line up with the notches in the holders. Then push *gently* downwards on each IC in turn until the chip is firmly seated in its socket.
7. Switch on! By rotating the tuning capacitor, VC1, you should now be able to tune in many stations, rotating the radio to give you some volume control.

Final touches . . .

The handle was made with part of an old leather belt, secured to the case with 'number plate' nuts and bolts from Halfords. The loudspeaker grille is the lid from a *pot-pourri* container, and some extra holes were drilled in the back to improve the sound. See what you can find to finish off your radio!

Parts list

Maplin order codes are given for most of the parts, but you should get used to using the 'beg, borrow or steal' technique, or to use your ever-expanding junk box.

		Maplin code
Resistors: all 0.25 watt, 5% tolerance		
R1	68 ohms (Ω)	M68R
R2	5.6 kilohms (kΩ)	M5K6
R3	10 ohms (Ω)	M10R
Capacitors		
C1, C4	10 nanofarads (nF) or 0.01 microfarad (μF) ceramic	BX00A
C2, C3, C7	100 nanofarads (nF) or 0.1 microfarad (μF) ceramic	YR75S
C5, C8, C9	100 microfarads (μF) electrolytic, at least 10 V	FF10L
C6	10 microfarads (μF) electrolytic, at least 25 V	FF04E
Semiconductors		
IC1	ZN415E radio chip	
IC2	LM386 audio power amplifier	UJ37S
LED1	3 mm green LED	WL33L
TR1	BC548 npn transistor	QB73Q
Additional items		
LS1	Miniature 8 ohm loudspeaker	WB08J
S1	Miniature SPST toggle switch	FH97F
Ferrite rod	Length approx. 100 mm	YG20W
	24 SWG enamelled copper wire	BL28F

Additional items (*continued*)

VC1	Tuning capacitor 140 to 300 picofarads (pF)	FT78K
	Tuning knob	FK41U
	8-pin DIL IC sockets (two required)	BL17T
	4 × AA-size battery holder (long)	HF94C
	PP3-type clip for battery holder	HF28F
	Plastic box approx. 158 × 95 × 54 mm	LH51F
	0.1 inch Veroboard, min. size 32 holes × 10 strips	JP46A

Plus
Stranded insulated conductor for flying leads
Multicore solder
Materials for handle and speaker grille
Double-sided sticky tape or Blu-Tack
Sellotape
Glue
Four AA-size 1.5 V batteries

22 The Colt 80 m receiver – Part 2

Introduction

In Part 1 we constructed the audio amplifier module for the system and tested it in a very simple way. If you did what was suggested and built the simple Crystal Set to use as a signal source, you will know just how well the amplifier works.

The case

Metal cases for the project are available from Maplin, telephone 01702 554 161 (code XB67). From the photograph on p. 72 you can see the way the components are mounted. The audio amplifier is seen at the top right of the base, to the right of the tuning capacitor VC1. The next in this series will deal with the variable-frequency oscillator (VFO) and VC1. The current part deals with preparing the case to receive the components.

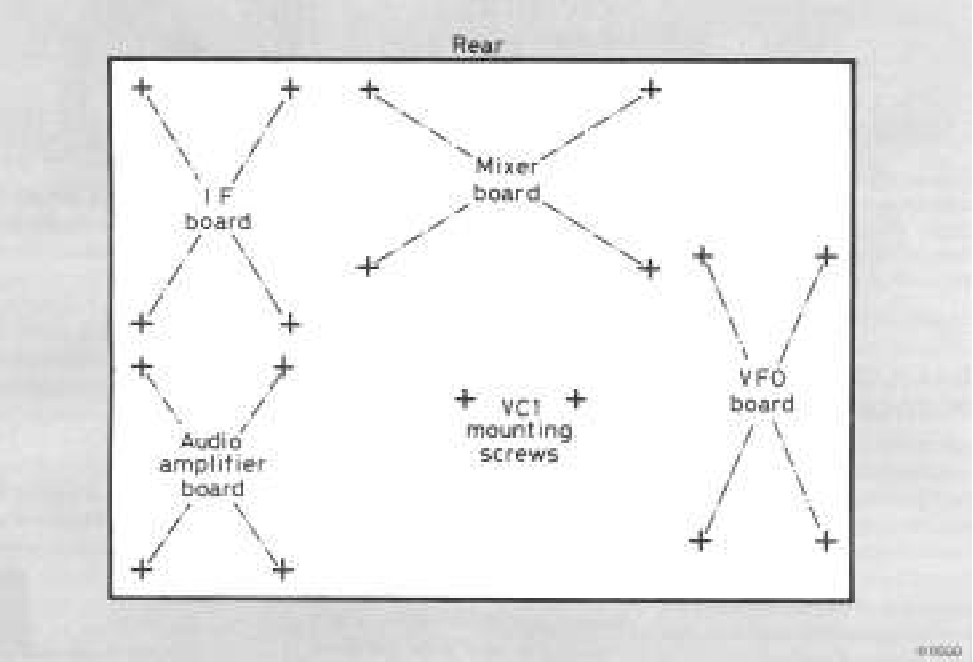
Rear
I F
board
Mixer
board
Audio
amplifier
board
VC1
mounting
screws
VFO
board

Figure 1 Fixing holes for each module are best measured from each printed circuit board or matrix board

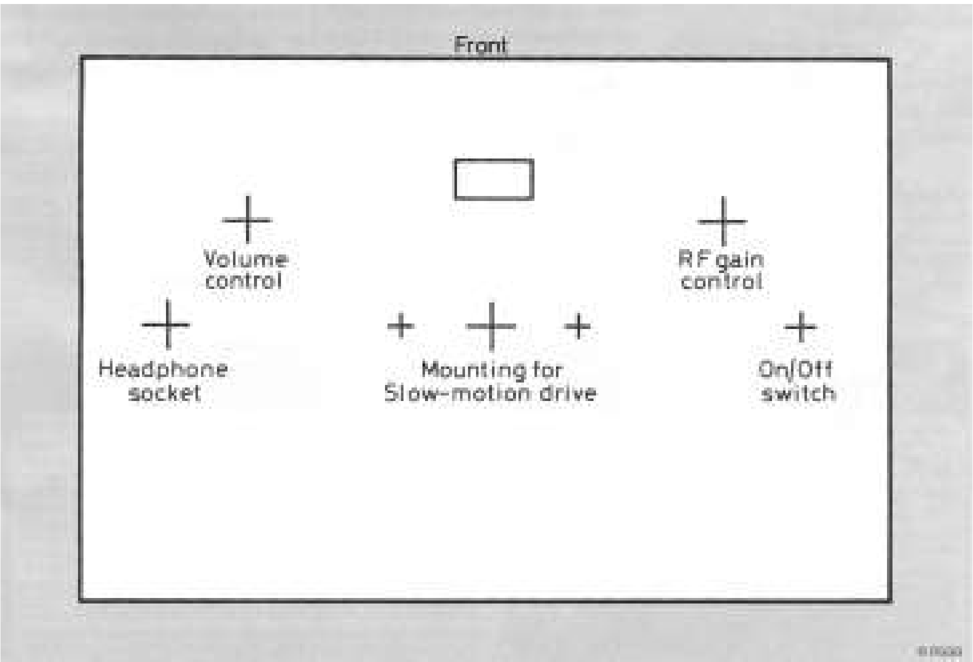
Front
Volume
control
RF gain
control
Headphone
socket
Mounting for
Slow-motion drive
On/Off
switch

Figure 2 Position the slow motion drive to allow viewing of the tuning dial

Figures 1 and **2** show the markings for preparing the front panel and base. All the circuit boards and the tuning capacitor are mounted on the base using 10 mm stand-off pillars with 6BA bolts. The board locations are shown in Figure 1. The front panel control positions are shown in Figure 2, together with the small rectangular hole for viewing the tuning dial.

The best way to mark out the holes for the boards is to lie the boards flat on the base (before you've started soldering the components in) and marking the base through the holes in the boards. This minimises the scope for errors!

A *reduction drive* is used between the tuning knob and the capacitor shaft. This is simply a gear mechanism that slows down the capacitor shaft by a factor of six compared with the tuning knob, and makes tuning very much easier. The recommended variable capacitor also has a pulley wheel mounted on the shaft. Glued to this wheel will be a scale marked with frequency and is visible through the rectangular hole in the front panel.

The next part . . .

The variable-frequency oscillator and mixer will be added to the project.

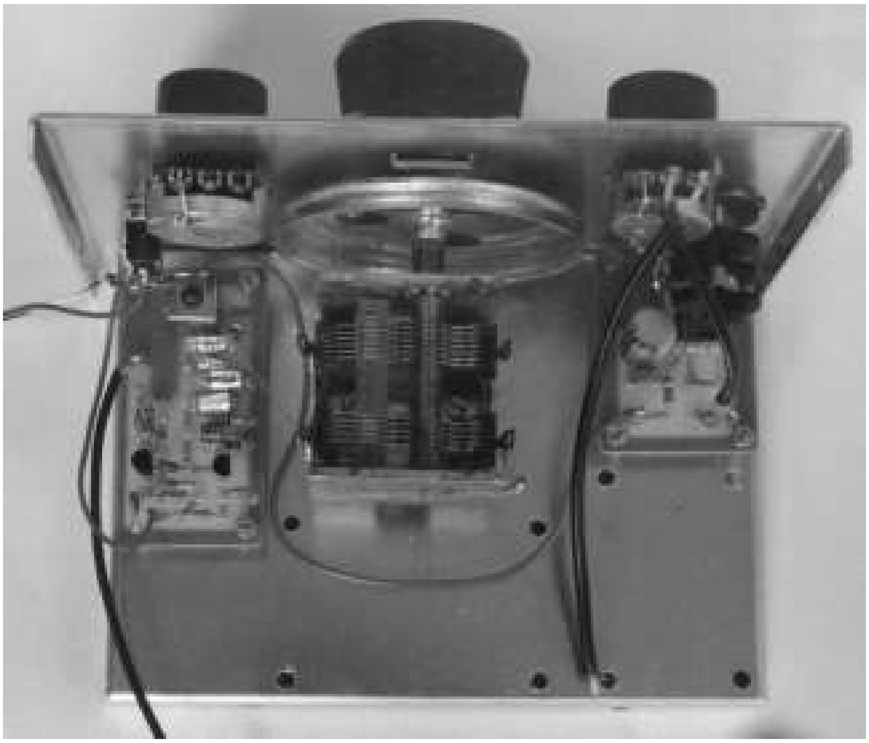

23 A simple transistor tester

Introduction

Although transistors aren't used as much as they were before integrated circuits came along, a transistor tester is still a useful piece of test equipment to have around the shack. This design is about the simplest possible and will produce an indication of whether a transistor is giving any current gain; this does not necessarily mean that the transistor is perfect but that it is working to a certain extent. This tester will **not** test field-effect transistors (FETs). If you buy a bag of transistors at a rally, this tester is useful for giving a yes/no indication of which ones go straight in the bin and which are kept for further use.

How it works

Figure 1 shows the simple wiring circuit. In order to explain the working of the circuit, a circuit diagram is shown in **Figure 2**, with the npn transistor, TR1, under test shown as part of the circuit. The pnp/npn selector switch, SW1, is omitted for clarity.

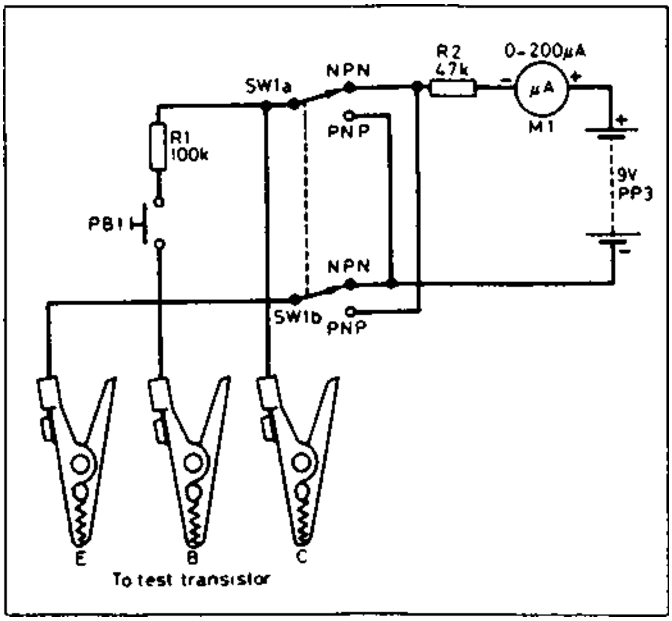

Figure 1 Circuit diagram of the transister tester

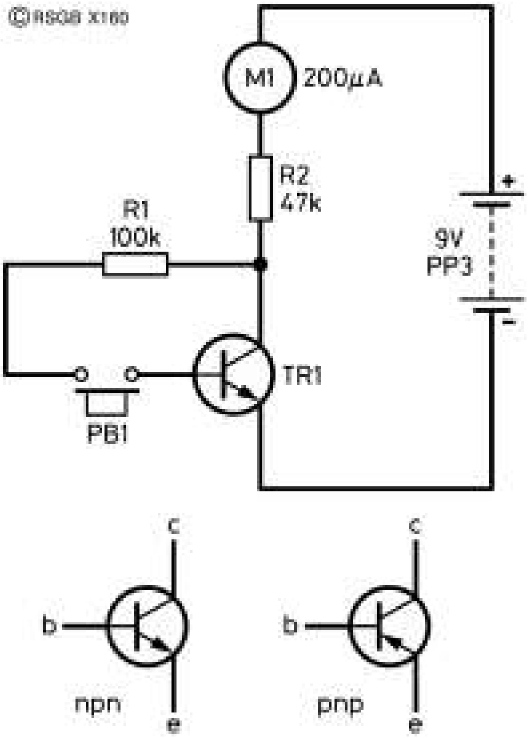

Figure 2 Testing an npn transistor

Any current that flows through the transistor, TR1, must flow from the battery, through the meter, M1, and through the protective resistor, R2. R2 prevents excessive current flowing through the meter and damaging it. Even if there is a short-circuit between emitter and collector, the maximum current that will flow is given by the simple equation

$$I = \frac{V}{R},$$

where I is the current flowing in amps,
V is the battery voltage, and
R is the total circuit resistance in ohms.

Putting in the correct values, gives

$$I = \frac{9}{47\ 000} = 0.000191 \text{ A, or 191 microamps } (\mu\text{A}).$$

The resistance of the meter, M1, will cut this down a little more, but it is within the indicating range (200 μA) of the meter.

As shown, with the push-button switch, PB1, *open*, a good transistor will not draw any current from the battery, and M1 will thus remain at zero. When PB1 is *pressed*, a *very small* current is injected into the base of the transistor. If the transistor is working, it will produce a much larger current between the collector and emitter, and this current will also flow through M1 and R2, giving a significant reading on M1, showing that all appears to be well. If an appreciable current flows when PB1 is *open*, then your transistor is suspect.

Don't be put off by the apparently complicated switch, SW1. It is there to allow the other type of transistor, the pnp type, to be tested. All it does is reverse the battery connections, so that the emitter goes to the *negative* battery terminal for testing an npn transistor, and to the positive terminal for a pnp type!

Most transistors in common use are of the npn type, which is why Figure 2 shows the testing of an npn type. The connections to the different transistor *encapsulations* (shapes) are given in any good component catalogue. Avoid the trial-and-error method to discover the connections to a transistor. This is unscientific, and can be very frustrating, particularly if the transistor is faulty in the first place!

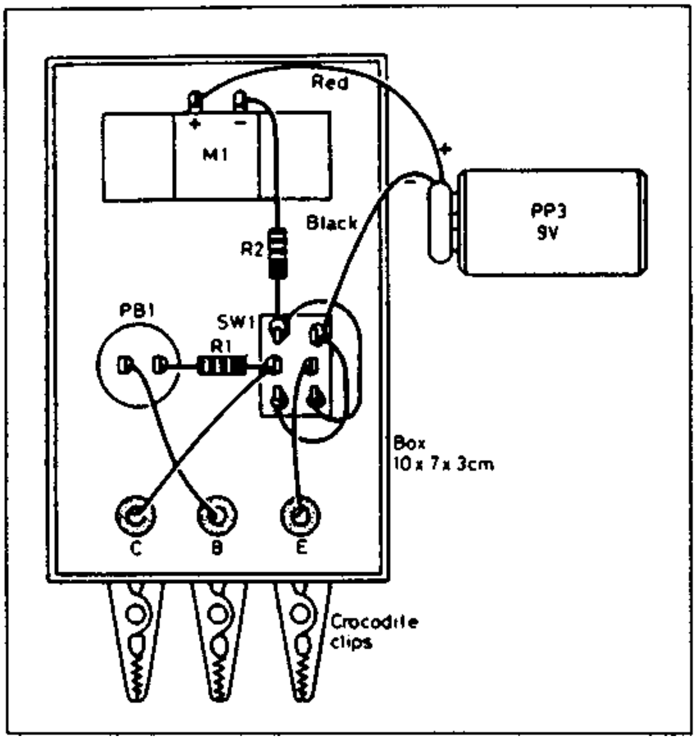

Figure 3 This shows how to wire up the tester

Construction

The unit can be built into a small plastic box, the components being soldered directly to the fixed terminals of the meter and the two switches; no circuit board is necessary! Use different colours of wire for the three test leads, and make sure you know which is which! Check the connections against the wiring diagram of **Figure 3**. When you are confident that all is correct, connect the battery and make sure there is no reading on the meter when nothing is connected to the crocodile clips!

Find any transistor for which you know the connections and the type (npn or pnp). Set the npn/pnp switch accordingly. Connect the three clips, making sure that they do not touch each other. A small reading on the meter at this stage means the transistor is suspect; a large reading means it is not working and should be thrown away! Press PB1 and watch the meter; a reading greater than half of full-scale indicates a good transistor. If it is less than half, you may have a transistor with 'low gain'; it may be usable for non-critical applications, but if you are in any doubt – throw it away!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	100 kilohms (kΩ)
R2	47 kilohms (kΩ)

Additional items

PB1	Push-button switch
SW1	DPDT (double-pole double-throw) switch
M1	Micro-ammeter – *not* less than 200 μA full-scale deflection
	Crocodile clips (3 needed)
	PP3 battery and connector
	Plastic box about 10 × 7 × 3 cm

Common types of transistor

BC108, BC109, 2N2369A

These are small-signal npn types, used in audio amplifiers. They are in metal cases (called TO18) and have a tab next to the emitter lead. Common types have a B or C suffix (e.g. BC109C). The C suffix indicates a higher current gain than those with a B suffix. The 2N2369A is specially designed for radio use at high frequencies.

2N3703, BC212L, BCY71
These are pnp transistors, and so must be used with the collector and base *negative* with respect to the emitter. These three types are used in small amplifiers and audio oscillators. The first two have plastic encapsulations (TO92), while the BCY71 has a metal case (TO18).

BFY50, BFY51, BFY52
For slightly higher powers, these are ideal. They have been used in novice transmitters up to 600 milliwatts (mW). The TO5 case is a scaled-up version of the TO18 case. They are all npn types.

2N3055, 2N3773, TIP35C
These are high-power transistors in bigger encapsulations. The thick metal TO3 case of the 2N3055 and 2N3773 is designed to bolt to a heat sink, a large piece of metal which conducts the heat into the air more rapidly than the transistor itself can. The TIP35C is made of plastic but has a thick metal tab by which it, too, can be bolted to a heat sink.

24 An introduction to transmitters

Introduction

We usually think of a transmitter as being a 'black box'. However, that is the form a transmitter takes for our use on the amateur bands. Many electrical circuits are transmitters, even though transmitting may not be their primary function!

Anything that emits electromagnetic energy at *any* frequency is a transmitter, from radio at the low-frequency end of the spectrum, to gamma rays at the high-frequency end. We all know that a magnet will attract certain metals and that a comb rubbed on your coat sleeve will pick up small pieces of paper. The former is an example of the effect of a magnetic field, the latter of an electric field. Electromagnetic fields are combinations of both types of field, and are produced whenever an electromagnetic wave is transmitted.

What frequency?

Many everyday objects have a natural frequency of oscillation. This is called their *resonant frequency*. A wine glass will ring when struck gently; an empty wine bottle will sound if you blow across the top; a guitar string will vibrate when plucked. These are all examples of resonance, and the resonant frequencies will **not** change unless the objects themselves are changed physically in some way. These are resonances in *sound*; we are primarily interested in electrical resonances.

The basic electrical resonant circuit is the combination of an inductor (coil) and a capacitor, as shown in **Figure 1**.

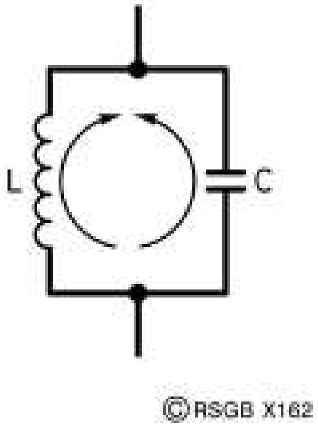

Figure 1 The basic electrical resonator. The energy in the circuit alternates between the inductor and the capacitor

A pulse of energy applied to this *tuned circuit* will make it *ring* (oscillate) at its resonant frequency. The energy in the circuit transfers between the inductor and the capacitor every cycle of the oscillation. Just like the wine glass, its oscillation dies away because it is losing some energy to its surroundings – it is transmitting! The frequency of the resonance depends on the values of *L* and *C*.

Keeping it going

If we want to keep the circuit oscillating, rather than having it die away, we must supply the circuit with just enough energy to replace the energy lost both by radiation and by losses in the circuit itself. Because of this, you will find in all oscillator circuits, a transistor, valve or FET working with the tuned circuit to provide this extra energy.

As it stands, of course, even with its transistor, our oscillator will not radiate very far. Connecting an aerial to it, and a Morse key to interrupt the power supply, it would become a very low-power CW transmitter. Add a couple more transistors to form a radio-frequency (RF) amplifier, and you have the basis of a simple low-power (QRP) transmitter.

Resonant circuits can also be made using quartz crystals; these work at the crystal frequency only, and this is marked on the crystal case.

A tiny spark transmitter

This is a simple piece of test gear that will increase your knowledge and understanding of resonance. You can use it to estimate the resonant frequency of most of the inductor/capacitor (LC) tuned circuits that you build. The circuit is shown in **Figure 2**. It operates around a relay. Any relay that operates from a 6 V to 9 V source and has contacts which are normally closed (i.e. closed when the battery is not connected). Fit the relay, a toggle switch and the battery in a metal box, and connected up as shown in the diagram. Some foam rubber inside the box may help to reduce the escaping noise of the relay. A small hole in the side of the box enables the 2-turn loop to emerge. This should be about 40 mm diameter, made with insulated wire. Switch on; there should be a loud buzzing noise from the relay. If not, you have probably chosen the wrong contacts on the relay!

When it is working, bring the loop close to the aerial of a radio – it should produce a loud noise from the speaker!

How it works

When you switch on, current flows through the relay contacts and through the relay coil. The relay operates and opens the contacts, causing the relay to 'drop out'. When it does, the circuit is completed again and the contacts are opened, and the cycle repeats. Each time the relay contacts open, there is a small spark between them, causing very rapid current surges through the wire loop. This makes the loop transmit RF energy, very briefly. In the early days of radio, this type of circuit was known as a *spark transmitter.*

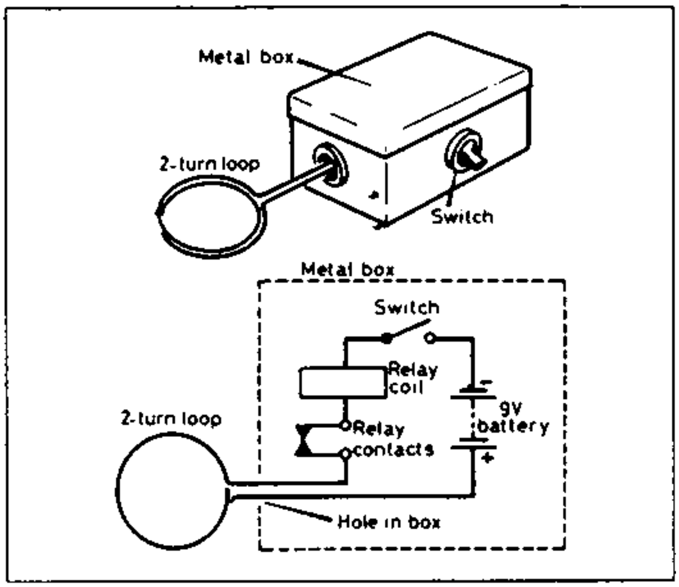

Figure 2 The current path is interrupted when relay is energised as shown above

Make a tuned circuit

Use a discarded toilet-roll centre, and wind about 10 turns of enamelled copper wire round it, keeping each turn close to the next. Scrape off the enamel for about 1 cm at each end, and solder a 100 picofarad (pF) capacitor (or a variable capacitor of about the same value) between the ends. The resonant frequency should be about 10 MHz. If you have used fewer turns or a smaller capacitor, the frequency will be higher.

Measuring the resonant frequency

Set up the buzzer as shown in **Figure 3**, with the loop around one end of your coil. Then make a similar loop, solder it to the end of a piece of coaxial cable going to the aerial socket of a calibrated receiver. Set the buzzer going, tune the receiver around 10 MHz, and search for the maximum noise level from the speaker. When you have found it, move the two loops as far away as possible from the main coil. This is called reducing the *coupling* between the coils, and it may result in a slightly different, but more accurate, resonant frequency.

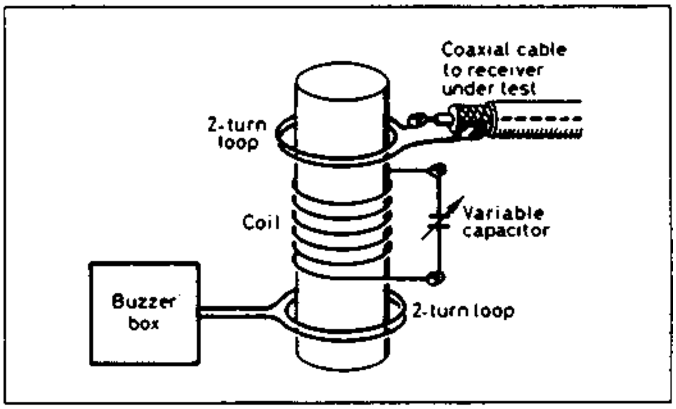

Figure 3 Experiment with the spacing between the loop and the tuned circuit

Parts list

Any small relay which operates between 6 V and 9 V
Metal box – do not use a plastic box!
9 volt battery and connector
On/off (SPST) toggle switch
Plastic foam, as required

25 The Colt 80 m receiver – Part 3

Introduction

In Part 2 we marked out the case ready for installation of the modules when they are completed. In this part, the building of the variable-frequency oscillator (VFO) and mixer will be described. This will produce a type of receiver known as *direct-conversion*, because it converts the radio-frequency (RF) signal directly into an audio-frequency (AF) signal which we can hear in a loudspeaker after amplification. A block diagram of the system is shown in **Figure 1**. The modifications needed to make a full superheterodyne receiver will be left until later.

The direct-conversion receiver covers the 80 metre amateur band and will receive both Morse (CW) and speech (SSB) signals. The audio amplifier was covered in Part 1, so your *Colt* is rapidly taking shape! By the time your construction has reached the end of this part, you will have a receiver ready to use, even if the project is not yet complete!

The direct conversion process

Like most things in radio, the principles of direct conversion are not difficult. From the aerial, the signal we want to hear is selected by the *tuned filter*, which rejects the signals we don't want. The signal then enters the

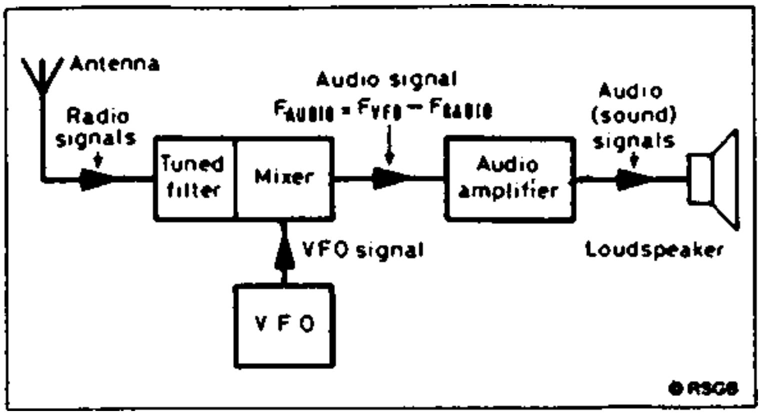
Antenna
Radio signals
Tuned filter
Mixer
Audio signal
$F_{AUDIO} = F_{VFO} - F_{RADIO}$
Audio amplifier
Audio (sound) signals
Loudspeaker
VFO signal
V F O
© RSGB

Figure 1 Stages of direct-conversion receiver

mixer, along with the signal from the *VFO*. The VFO produces a sine wave whose frequency can be varied across the whole of the 80 metre amateur band (3.5–3.8 MHz), by turning the knob on the tuning capacitor, VC1. The mixer produces, at its output, *two* signals; one signal is at the frequency of the **sum** of the signal and VFO frequencies, the other is at the **difference** of the two frequencies. It is the latter that we want. Let's look at the numbers involved. If the signal is at 3650 kHz and the VFO is at 3651 kHz, then the sum frequency is 7301 kHz, and the difference frequency is 1 kHz. If we feed the output of the mixer into our audio amplifier, the 7301 kHz signal is automatically removed (it is far too high to be considered an audio signal!) and the resulting 1 kHz signal is amplified and fed to the speaker, producing a note which we can hear!

Building the VFO

Figure 2 shows the circuit of the VFO. It is a tried and tested circuit, and should work first time. It uses a *field-effect transistor* (FET) for TR1, the oscillator itself. RFC is a *radio-frequency choke*, a coil of wire which will pass a direct current (DC) but which will prevent radio-frequency (RF) signals getting through.

All VFOs have a tuned circuit which, in this case, is formed by the coil L1 and the capacitors C1 and VC1. The frequency will also be affected to some extent by C2, C3, C4 and C5. Transistor TR2 is an *emitter follower*, a stage which gives no voltage gain but provides a good *buffer* stage, isolating the VFO from the effects of the stages that follow it. When building a VFO, the

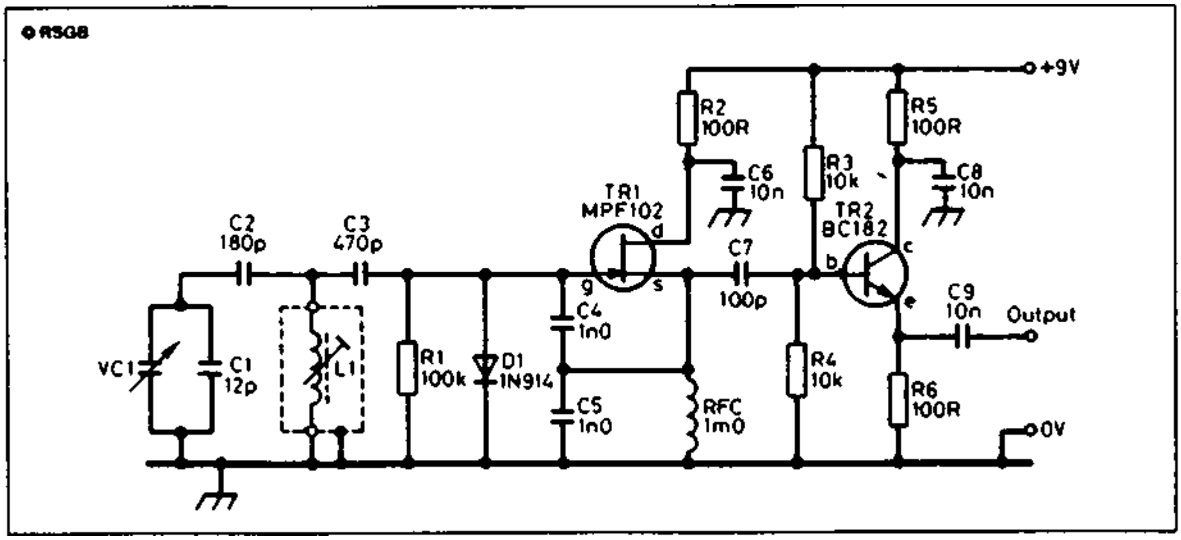

Figure 2 The variable-frequency oscillator uses a field-effect transistor (FET)

parts must be securely mounted. If components move, so does the frequency! At worst, the oscillation will become unstable and the VFO will be useless. Keep the component leads as short as possible – this improves their mechanical stability as well as their electrical stability!

Mount the parts on the printed-circuit board (PCB) or matrix board and, when completed, the VFO should look like the one shown on the left in the photograph on p. 72. Make sure that TR1, TR2 and D1 are the right way round.

On completion, check the component positions then mount it in the case as shown in the photograph, to the left of the tuning capacitor when viewed from the rear. The VFO coil, L1, will need some adjustment, but that will have to wait until the mixer is built. Connections to the other boards are made with screened cable.

The mixer board

So far, we have an audio-frequency (AF) amplifier and a VFO; the addition of a mixer board gives us a complete direct-conversion receiver for 80 m. The mixer circuit diagram is shown in **Figure 3**. Let's follow the signal path.

- From the aerial, the RF signal goes to the gain control potentiometer, RV1. This reduces very strong signals, to prevent them overloading the mixer.
- To select the required band of signals, a *bandpass* filter is made up of RF transformers T1 and T2; these are tuned by C1 and C3, and are coupled together by C2.

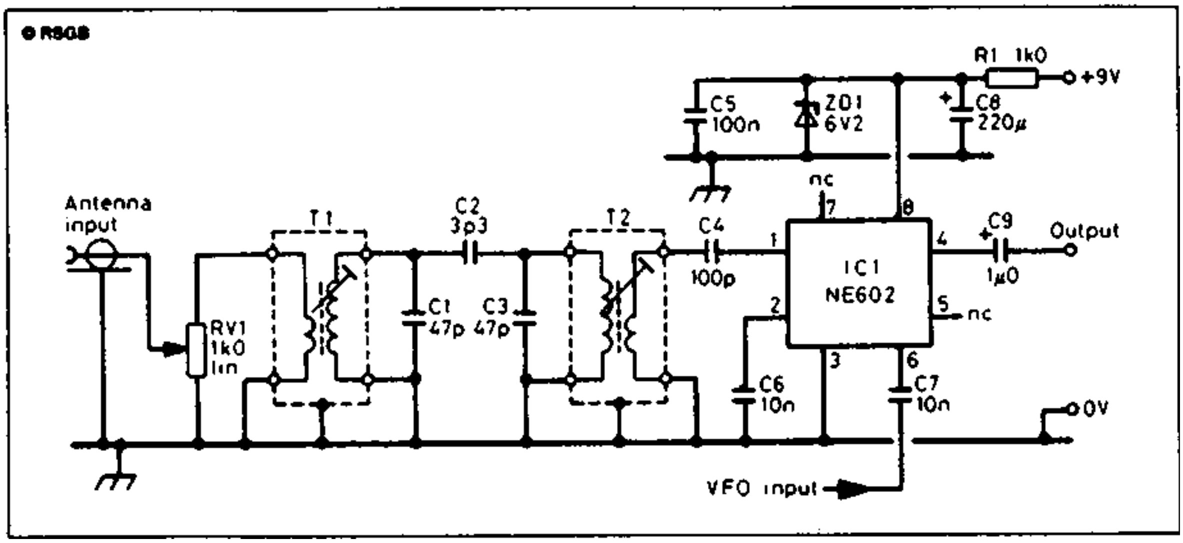

Figure 3 The mixer board has a bandpass filter and stabilised supply

- After the filter, the signal is coupled into the integrated circuit (IC) mixer type NE602, by the capacitor C4. Capacitors C5 and C8 *decouple* the supply line, to prevent unwanted signals on the supply from disturbing the VFO operation. The use of the term *decouple* is exactly the opposite of *couple*; when two circuits are *coupled* together, the signal passes from one to the other; when two circuits are *decoupled*, signals **cannot** pass from one to the other.
- The NE602 works with a 6 V supply; it is produced here from the 9 V supply by the *Zener diode* ZD1 and the resistor R1. ZD1 operates at 6.2 V, and gives a steady output for the mixer. The audio output from the mixer appears at pin 4 or IC1. This is taken to the audio amplifier board via C9 and the volume control (see Part 1).

Care must still be taken to insert some components the right way round. These are the electrolytic capacitors, C8 and C9, the Zener diode, ZD1, and the integrated circuit, IC1. Check all component positions and make sure all your soldered joints are bright and shiny.

Putting it together

The interboard wiring, shown in **Figure 4**, uses screened cable; ideally, this should be thin coaxial cable, but screened microphone cable is suitable. The diagram shows how the two controls, the RF Gain and Volume, are connected to the boards. The leads marked '+9 V' are all connected to the battery supply via a miniature on/off toggle switch. Double check all connections before connecting the battery.

Setting the VFO

Very little adjustment is needed to get the receiver going. Firstly, the VFO must be adjusted to cover the required band, in this case 3.5–3.8 MHz.

If you have a frequency counter, connect it to the output of the VFO. If you haven't, read this part anyway so you understand the process, then another means of setting the VFO will be given especially for you! Rotate VC1 until the vanes are fully meshed. Very carefully, adjust the core of the VFO coil (L1) with a plastic trimming tool so that the frequency approaches and settles at 3.500 MHz. When you rotate VC1, the frequency should increase to **at least** 3.800 MHz at the far end of its travel.

In the absence of a frequency counter, borrow a communications receiver, set it for SSB reception (USB or LSB) on exactly 3.500 MHz. Set VC1 with the vanes fully meshed and turn the core of L1 in both directions until you hear a whistle in the communications receiver. Rotate the core so that the whistle reduces in frequency. It will eventually fade out at around 200 Hz;

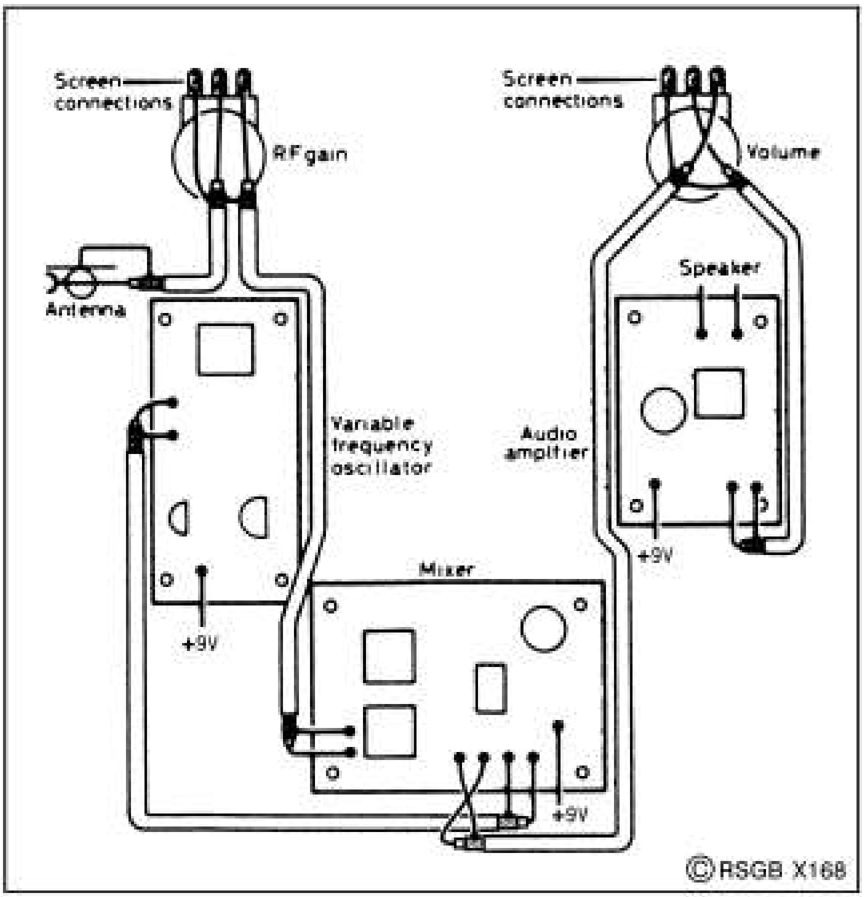
Screen connections
RF gain
Antenna
Screen connections
Volume
Speaker
Variable frequency oscillator
Audio amplifier
+9V
Mixer
+9V
+9V
©RSGB X168

Figure 4 Make sure that interconnections between the boards are correct, including cable screens

turn the core a little further and then leave it at that position. Rotate VC1 right to the other end of its travel, and search for the whistle with the tuning knob of the communications receiver. Check that the VFO frequency is at least 3.800 MHz.

Then, whichever method you are using for frequency measurement, mark the dial with frequency steps of 50 kHz. Setting and calibration are finished!

Setting the mixer

Again, there are various ways of doing this. If you have a signal generator, inject a signal at a frequency within the 80 m band, and adjust the cores of T1 and T2 sequentially for maximum output.

If you haven't a signal generator, connect an aerial to the mixer input, set the RF Gain to maximum (fully clockwise), and find a consistent signal. Adjust the volume control to a comfortable level. Rotating the core of T2 with your trimming tool, maximise the output. Then do the same with T1, although this will have much less effect. Find another station, and check that the positions of the cores aren't too different for a maximum signal.

You may find that your receiver benefits from the insertion of an aerial tuning unit (ATU) between the aerial and the input, to compensate for the impedance of your aerial not being 50 Ω. A design for such an ATU is presented in another part of this series. If the signals are still weak, connect the ATU to the junction of C1 and C2 via a 100 pF capacitor.

Try listening!

Remember that 80 m is a variable band. During daylight hours, your will hear Morse signals at the lower end of the band, and some British and closer continental stations between 3.7 and 3.8 MHz. In the evenings, stations up to 1000 miles away should be heard. Look for Novices around 3.7 MHz!

Parts list – VFO board

Resistors: all 0.25 watt, 5% tolerance		
R1	100 kilohms (kΩ)	
R2, R5, R6	100 ohms (Ω)	
R3, R4	10 kilohms (kΩ)	
RV1	1 kilohm (kΩ) linear	
Capacitors		
C1	12 picofarads (pF) polystyrene	
C2	100 picofarads (pF) polystyrene	
C3	470 picofarads (pF) polystyrene	
C4, C5	1 nanofarad (nF) polystyrene	
C6, C8, C9	10 nanofarads (nF) polystyrene	
C7	100 picofarads (pF) min. ceramic	
VC1	140 + 140 picofarads (pF) variable	
Semiconductors		
TR1	MPF102	FET
TR2	BC182	npn
D1	1N914	silicon
Inductors		
L1	Toko KANK3334	
RFC	1 mH	RF choke

Parts list – mixer board

Resistors: 0.25 watt, 5% tolerance

R1	1 kilohm (kΩ)
VR1	1 kilohm (kΩ) potentiometer (linear)

Capacitors

C1, C3	47 picofarads (pF) min. ceramic
C2	3.3 picofarads (pF) min. ceramic
C4	100 picofarads (pF) min. ceramic
C5	100 nanofarads (nF) min. ceramic
C6, C7	10 nanofarads (nF) min. ceramic
C8	220 microfarads (μF) electrolytic 16 V
C9	1 microfarad (μF) electrolytic 16 V

Integrated circuit

IC1	Philips NE602 or NE602A

Additional items

T1/T2	Toko KANK3333
	On/off switch
	Miniature toggle switch

The next part . . .

The IF amplifier and the Beat-Frequency Oscillator will be added to convert the *Colt* into a superheterodyne receiver.

26 A two-way Morse practice system

Introduction

It is said that the pleasure of sending and receiving Morse code more than compensates for the learning effort needed. Transmitters for Morse code can be much simpler than those needed for any of the speech modes.

Many people choose to learn sending and receiving on their own. It can be much more fun if you have someone to learn with you, and it is for this reason that the following project arose. It comprises just one small circuit, and you build an identical circuit for each of the people who want to learn with you. All the individual circuits are connected with their output leads in *parallel*, as the two-way circuit of **Figure 1** shows.

A simple circuit

This diagram shows two identical circuits, as would be used if two people wanted to learn together. The circuit centres around our old friend the NE555 integrated circuit (IC). As it is connected here, it works as an *astable multivibrator*, a daunting name for what is essentially an oscillator. Each circuit is self-contained, having its own battery, Morse key, sounder and

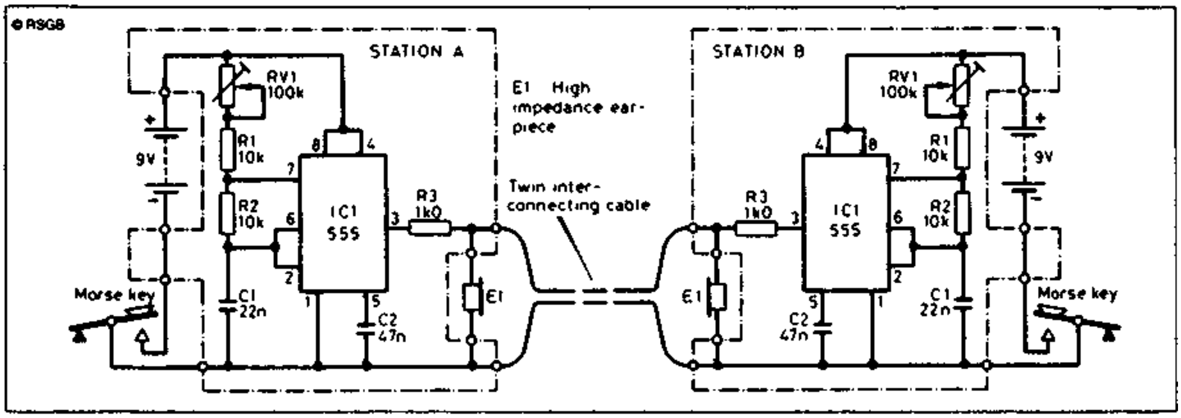
© RSGB
STATION A
STATION B
RV1
100k
R1
10k
R2
10k
R3
1kΩ
C1
22n
C2
47n
IC1
555
9V
+
−
8
4
7
6
2
1
5
3
Morse key
E1
E1 High impedance ear-piece
Twin inter-connecting cable

Figure 1 The circuit diagram shows two stations, but you could connect several more if needed

plastic box. The Morse key makes and breaks the power supply to the circuit, thus turning on and off the note emitted by the earpiece. The frequency of the note is controlled by the variable resistor RV1, resistors R1 and R2, and the capacitor C1. The range given by RV1 is approximately from 500 Hz to 2200 Hz.

R3 protects the output of the IC against accidental shorting either to the 0 V or to the 9 V supply rails. For best results, a crystal earpiece or high-impedance headphones should be used. You could try Walkman-type headphones, but the volume may be too low for you.

The units may be interconnected with thin twin cable (the sort used for wiring doorbells) and can be comfortably separated by 25 m, more than enough to communicate between rooms, or even with a neighbour!

In use

When an operator presses his key, the note is heard by himself and by everyone else who is connected. If more than one key is pressed at the same time, two or more notes are heard by everyone! It helps if each person adjusts his own potentiometer, RV1, to give a note which is different from the others. When one operator stops sending, the other can start immediately, without pausing to change from transmit to receive. In practice, this technique is known as *full break-in*.

The current drawn by the circuit is only 10 milli-amps (mA) when the key is down, meaning that the life of a typical PP3 battery will be virtually its shelf-life! No switch is needed, because the key acts as the switch. The current drain will be even less if the NE555 is replaced by an ICM7555 IC. It has exactly the same pin connections as the NE555, so no circuit modifications are needed.

The circuit board

The circuit is constructed easily on a small piece of Veroboard of the copper-strip type. **Figure 2** shows the layout of the prototype, the board measuring 18 holes by 10 strips. **Be aware that there is no row ‘I’ when you are transferring mental images of where the parts are to real positions on the board!** Break the copper tracks where shown with a 3 mm ($\frac{1}{8}$ inch) twist drill, rotated carefully between thumb and forefinger. Hold the board up to a bright light to make sure that the tracks are completely broken and that there are no fragments of copper swarf shorting adjacent tracks. Then, solder in the wire links shown in Figure 2, followed by the resistors, RV1, and the capacitors, C1 and C2. Solder the battery connector leads to tracks A and G, G being positive. Use insulated wire to connect from the board to the jack sockets used as connectors for the key and the earphones. If you use

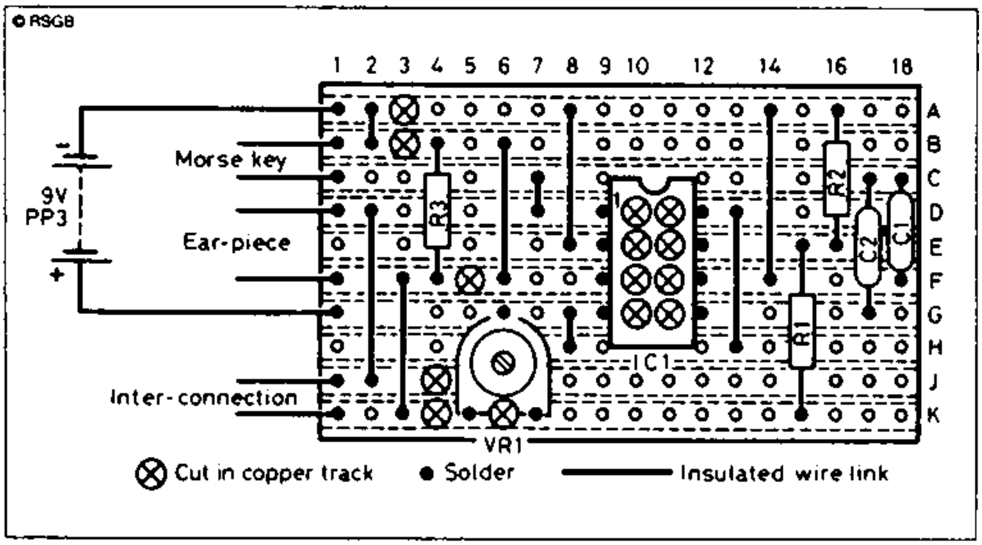
© RSGB
1 2 3 4 5 6 7 8 9 10 12 14 16 18
A B C D E F G H J K
Morse key
9V
PP3
-
+
Ear-piece
Inter-connection
R1
R2
R3
C1
C2
IC1
1
VR1
Cut in copper track
Solder
Insulated wire link

Figure 2 The Morse Duet is an easy project to build on stripboard (Veroboard)

different sized jack sockets (i.e. 2.5 mm and 3.5 mm) you won't plug the earphones into the key socket! A two-screw terminal strip is useful to connect the cable runs between operators.

The key

The circuit is designed to be used with a *straight key* (one that moves up and down). It is a requirement of the UK Morse test that a straight key is used, so it is very sensible to learn sending with a straight key before you try anything more complex! You will find many to choose from at rallies.

Parts list

For each board, you will need:		**Maplin order codes**
Resistors: all 0.25 watt, 5% tolerance		
R1, R2	10 kilohms (kΩ)	M10K
R3	1 kilohm (kΩ)	M1K0
RV1	100 kilohm (kΩ) min. preset (horizontal mounting)	UH06G
Capacitors		
C1	22 nanofarads (nF) ceramic, 25 V	WX78K
C2	47 nanofarads (nF) ceramic, 25 V	RA47B

Integrated circuit	
IC1 NE555	QH66W
Additional items	
8-pin DIL socket	BL17T
Jack plug and socket 2.5 mm	HF76H, HF78K
Jack plug and socket 3.5 mm	HF80B, HF82D
Battery type PP3	
Battery connector	HF28F
Plastic box with lid 114 × 76 × 38 mm	LH14Q
Veroboard 0.1 inch with copper strips 18 holes × 10 strips	JP46A
Terminal strip (two section)	FE78K
Crystal earpiece	LB25C
Morse key	

27 The Colt 80 m receiver – Part 4

Introduction

In Parts 1 to 3 of this series, the design of an 80 metre direct-conversion receiver has been described. In this final part, we are going to change the circuit to operate as a superheterodyne receiver, or *superhet*. Most radio receivers are superhets, and they change the incoming signal to another frequency, known as the *intermediate frequency*, or IF, before producing an audio signal. The use of a superhet in a good receiver is mandated by the requirements for good *sensitivity* and *selectivity.*

Sensitivity and selectivity

Figure 1 shows the block diagram of the receiver. If you look closely, you will see that it is two similar circuits, one after the other. The incoming

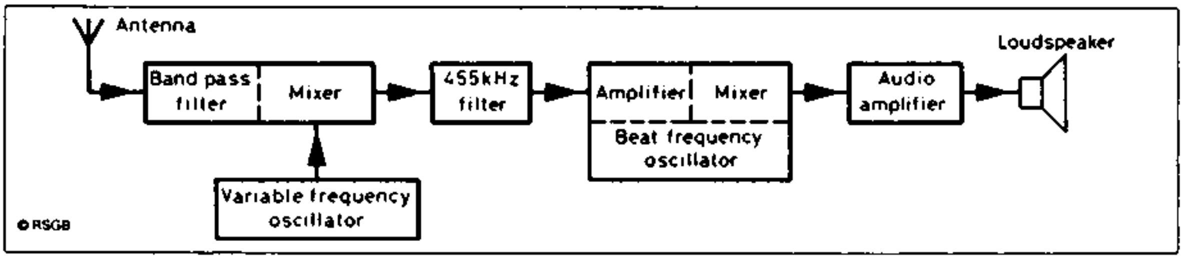

Figure 1 This is how the different stages of the Colt go together to make a superhet receiver

signal is filtered and fed to a mixer where it is combined (mixed) with the signal from the variable-frequency oscillator (VFO). This oscillator operates at a frequency which is 455 kHz higher than the incoming signal from the aerial, and the mixer output is therefore at a frequency of 455 kHz. If you are not sure about this, please refer to the section ‘The direct conversion process’ in Part 3 of this project. This new frequency is called the *intermediate frequency*, or IF. This frequency doesn’t change; the tuning is accomplished by the VFO, and the mixer output is *always* at 455 kHz. The extraction of the audio signal from the IF signal is identical with the direct conversion process which is used in your existing receiver.

This may seem a long-winded way of doing things, but it has its advantages. A receiver must have a good *sensitivity*, or gain, so that it can receive very weak signals. In very general terms, it is easier to handle low-frequency signals than it is to handle high-frequency signals. We are changing our signal frequency from around 3.6 MHz down to 0.455 MHz (455 kHz), which is much lower and can be filtered and amplified relatively easily. A second advantage is that it is easier to provide gain at a fixed frequency than at a variable frequency. Remember that the IF is fixed, and providing gain is, again, relatively simple.

Our receiver also needs good *selectivity*, the ability to separate (or select) one station from another very close to it in frequency. This requirement is significantly simplified by the fact that the IF is fixed, and a good filter in the IF circuits can do wonders for the rejection of adjacent-frequency stations! Several stages of IF amplification and filtering are possible in more adventurous designs.

The filtered signal at 455 kHz passes to another mixer which has an associated oscillator, usually called a *beat frequency oscillator* (BFO). When receiving CW (Morse) signals, the BFO is usually tuned about 1 kHz above or below the IF (i.e. at 454 or 456 kHz) to produce a 1 kHz beat note as the audio signal. This signal is then amplified and fed to a loudspeaker or headphones.

The circuit

Figure 2 shows the circuit of the IF section shown in the photograph on p. 94; this section is added to the existing circuit to make it a superhet. The existing mixer board (described in Part 3) is used as the first mixer. Between it and the audio amplifier is connected the new IF board.

The signal output from the first mixer (at 455 kHz IF) is fed into a *crystal filter*, the most expensive part in the whole receiver. It provides the selectivity which makes the Colt such a good receiver. Another NE602 mixer/oscillator chip follows the filter. The oscillator section is controlled by the tuned circuit in T1, the frequency of which can be altered by rotating the core inside the coil. Once set, it remains fixed.

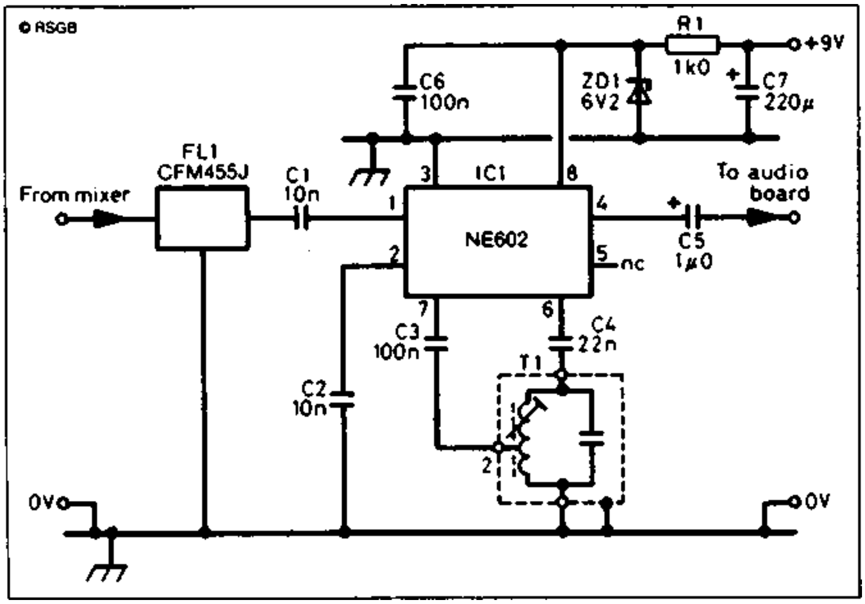

Figure 2 The IF board is connected between the mixer and audio amplifier

At this stage you should have built the first mixer and the audio amplifier, and proved that they *both* work by using the circuit as a direct-conversion receiver. When you have finished constructing and checking the IF board, you will need to add it to your existing circuits.

Adjusting and testing

There are three pairs of connections to the IF board – the 9 V supply leads, the input leads and the output leads. The IF board should be mounted on the metal baseplate, along with the other boards. **In making the following connections, make absolutely sure that the braid of each piece of coaxial**

cable is soldered to the correct connection on each board. The same applies to the polarity of the 9 V leads. Disconnect the screened lead which presently goes from the first mixer board to the AF amplifier at the amplifier end. Connect it instead to the *input* of the IF board. Connect a new piece of screened cable from the IF output to the vacated AF input. Now connect the 9 V supply leads to the same points as the supply leads from the other boards. Check these newly made connections.

You should have confidence at this point that things should be right. After all, you *have* tested the direct-conversion process and you *know* it works. All that you are now testing is the IF board. This is the attraction of building a receiver in modules, working from the speaker backwards, and testing as you go!

The VFO and BFO need to be correctly adjusted. Mesh the vanes of the tuning capacitor and, using the same frequency measurement method as you did originally, set the VFO frequency to 3.955 MHz (which is 3.500 MHz + 0.455 MHz, if you hadn't guessed!). The BFO can be set using a frequency counter, but it is just as good to set it by listening to SSB or CW signals. A high-pitched hissing sound should be heard in the speaker. As you rotate the core, the pitch should reduce, go through a minimum, then increase again. Set the core at the minimum pitch position. You may want to readjust the two cores a little as your listening skills improve but, once you are happy, they will never need to be altered again!

In conclusion . . .

You should now have built a superhet receiver capable of excellent results. It uses the same type of circuit as that found in far more advanced receivers.

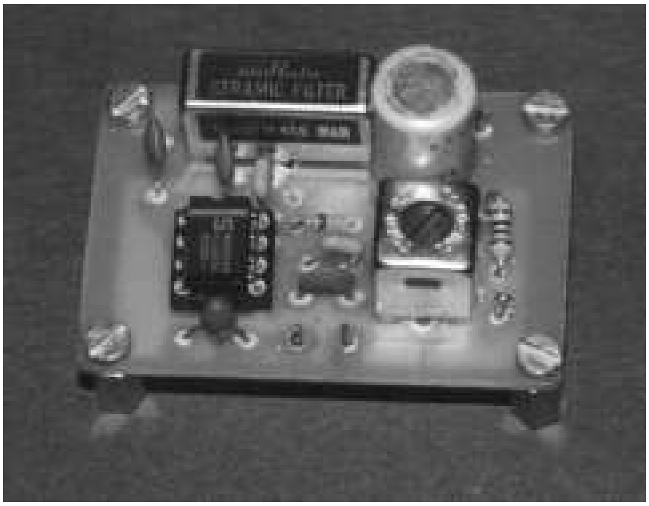

The superhet is far more sensitive than the direct-conversion type, and can weed out those elusive DX stations. You may have found that a station will appear at two places on the dial of the direct-conversion receiver; you will have no such problem with the superhet.

Parts list

Resistor	
R1	1 kilohm (kΩ), 0.25 watt, 5% tolerance
Capacitors	
C1, C2	10 nanofarads (nF) min. ceramic
C3, C6	100 nanofarads (nF) min. ceramic
C4	22 nanofarads (nF) min. ceramic
C5	1 microfarad (μF) 16 V electrolytic
C7	220 microfarad (μF) 16 V electrolytic
Integrated circuit	
IC1	NE602
Additional items	
ZD1	Zener, 6.2 V 0.5 watt
FL1	Crystal filter, Murata CFM455J
T1	Tuned inductor, Toko YHCS11100AC

28 A simple crystal set

Introduction

There are many designs of crystal set – they are all ‘simple’, and even the most seasoned radio amateur will build one of these every so often because it is something that never ceases to amaze! Although using modern components, the design is a period piece!

Design and construction

The circuit diagram of **Figure 1** shows the simplicity of the circuit. The prototype was constructed on a wooden baseboard with an aluminium front panel, although this could be made of wood if you prefer.

Winding the coil always causes the most groans but when finished, gives the most satisfaction. The construction of the coil is detailed in **Figure 2**. A discarded toilet roll centre is an ideal *former* on which to wind the coil. The coil will take up about 7 cm of the length of the tube, so assuming the tube is about 11 cm long, the ends of the coil will be about 2 cm in from each end of the tube.

The aerial coil, the *primary winding* of this radio-frequency (RF) transformer, is made from about 30 turns of 26 SWG enamelled copper wire, wound on the matchstick ribs. These ribs should be about 2 cm long, and are glued on **top** of the end of the secondary coil *after* the secondary coil has been wound. The *secondary winding* has 140 turns with *taps* every 10 turns for 70 turns. This coil covers most of the length of the cardboard tube. To make your coil-winding easier, here are some tips for completing the coil and still having some hair left at the end!

- Support the reel of wire on a dowel rod or pencil, clamped in a vice or held rigid in a vertical position by some other means. Don't leave it to trail around on the floor.
- Have a small piece of sandpaper handy to remove the enamel from the copper wire, starting with the end coming from the reel. Remove about 1 cm only, until the shiny copper is visible all round the end of the wire.

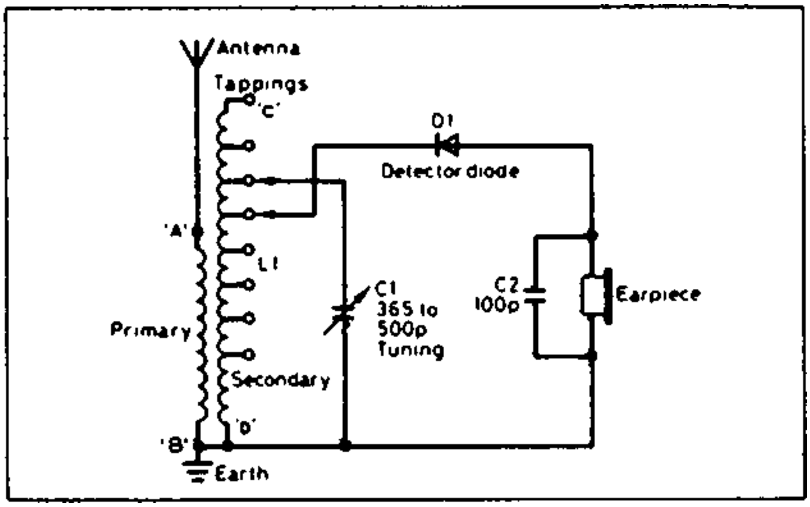

Figure 1 Circuit diagram of the crystal set

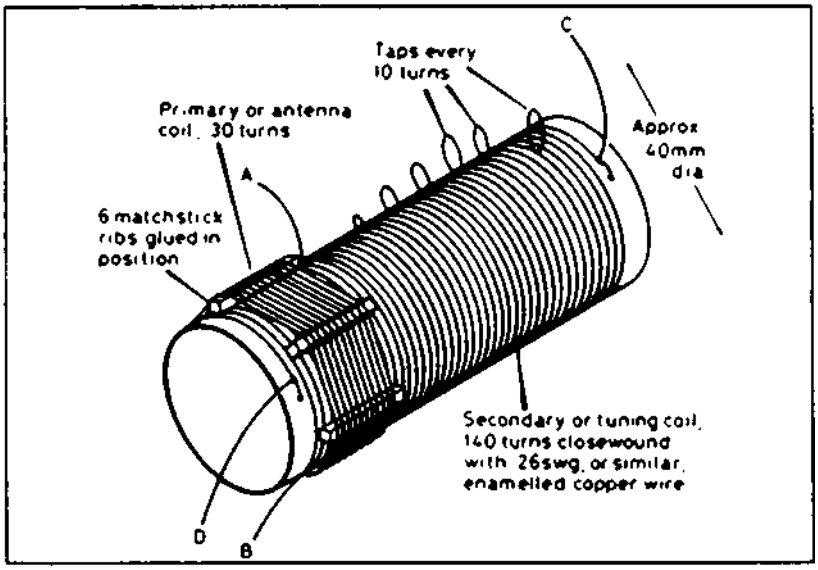
C
Taps every
10 turns
Primary or antenna
coil, 30 turns
Approx
40mm
dia
A
6 matchstick
ribs glued in
position
Secondary or tuning coil,
140 turns closewound
with 26swg, or similar,
enamelled copper wire
D
B

Figure 2 The coil is mounted on a cardboard tube as shown. Exact size is not too important. Try different taps for best results

- This part is the most difficult of all. Seasoned coil winders may have their own favourite ways of doing this, but for first-timers, play it safe and do it this way! Have a second bobbin (or a piece of stout card) ready, on which you will need to wind the wire loosely while you prepare the taps. Then, starting from the free end of the wire, measure 125 cm and, at this point, remove about 2 cm of the enamel with the sandpaper. Bend the wire firmly back on itself in the middle of the bared area, and *tin* (cover with solder) the exposed copper. This is your first tap. Then repeat this six more times at 125 cm intervals, thus giving you seven taps; wrap the wire as you go, on to the second bobbin or piece of card. After the seven taps, you are half-way down your coil so, to begin winding it in earnest, wind the wire carefully from the second bobbin or card **back** on to the original reel of wire.
- Looking at Figure 2, make two small holes about 3 mm apart, about 2 cm down from the top of the tube, where wire C will be entering. Poke the end of the wire into one hole, then bring it out again through the other; leave about 10 cm of wire on the free end. Loop this wire in and out once more, thus anchoring the wire firmly.
- Then, begin winding; keep the wire tight and make adjacent turns touch neatly. As you pass each tap, make sure it sticks outwards while avoiding flexing the wire too much; the wire is inherently weak at each tap.
- When you reach the last 10 cm of wire, stop. Put two more holes, like the first, beside your stopping point, and anchor the end, D, of the coil in the same way as you did with end C. The secondary winding is complete. Take a break *after* you have completed the next step!

- Prepare the six matchsticks for supporting the primary coil and glue these every 60 degrees around the end of the coil at D. Remove the enamel from the free end of the new copper wire on your reel and, using your favourite quick-drying cement, put a drop over the wire on one of the matchsticks, leaving 10 cm free as before. If you are using 26 SWG enamelled copper wire, its diameter is 0.46 mm, so 30 turns should occupy 14–15 mm of the 20 mm matchstick length. This will allow you to position the glued point such that the coil will be roughly in the centre of the matchsticks. When the glue has set, wind the 30 turns and anchor this end of the coil in the same way. Then remove the enamel at the end of the 10 cm lead.

Assembly

The coil former can be screwed to the baseboard at each end, *before* the front panel is screwed to the baseboard! As the photograph on the next page shows, the tuning capacitor is mounted in the centre of the front panel. Five solder tags (or drawing pins as a last resort) should be screwed into the baseboard. The two to the left of the photograph are the connections to the earphones, the two along the rear edge have the diode, D1, soldered to them. One end of the diode (and, for once, you can connect the diode either way round!) is connected to one earphone tag. The other end has a flying lead of about 10 cm attached to it and terminated in a crocodile clip which is used to connect to one of the taps on the coil. The fifth tag secures the end, A, of the primary coil, to which you will attach the aerial. The earth tag on the tuning capacitor, C1, serves as an anchor point for the ground connections to B and D and to the other earphone tag.

Use

The longest piece of wire you have available to use as an aerial should be connected to the aerial tag just mentioned. If you can connect the earth tag of the tuning capacitor to a good electrical earth, this will help also. You should be able to hear something as you turn the tuning knob. Try adjusting the tapping connections on the coil – change only one at a time, or you will never find the optimum positions. You now have a fully operational crystal set!

Simple it may be, but this circuit illustrates some important principles which are used even in the most expensive receivers. Firstly, the coil, L1, and the capacitor, C1, form a *tuned circuit* which *resonates* at the frequency of the station you have tuned in. This *selects* the signal you want to hear. The *modulation* on this signal is removed by the *detector*, D1, and fed to the earphones, which act as *transducers*, turning the electrical energy into sound energy which you can hear. It doesn't need a battery or other power supply either!

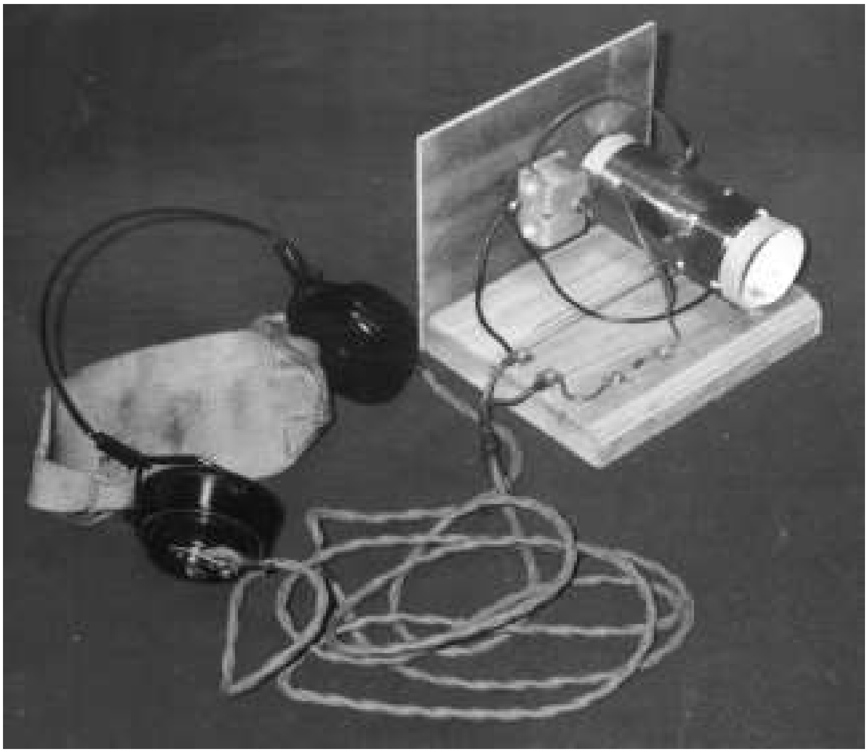

Parts list

Capacitors

C1	Variable capacitor of between 200 and 500 picofarads (pF) maximum
C2	100 picofarads (pF) min. ceramic

Additional items

D1	Germanium, type OA47 or OA90
	High impedance crystal earphone (not Walkman type)
	Tuning knob
	Wooden base, approx. 110 mm square
	Aluminium (or wood) front panel, approx. 80 × 100 mm
	Cardboard tube (toilet roll centre), about 110 mm long, 40 mm diameter
	Reel of 26 SWG enamelled copper wire for both windings of L1

29 A crystal calibrator

Introduction

A crystal calibrator is a device that produces oscillations of a precise frequency which are rich in *harmonics*. A harmonic is an integer multiple of the fundamental frequency; for example, if our fundamental frequency was 100 kHz (as it is in this circuit), there would be harmonics of this frequency at 200 kHz, 300 kHz, 400 kHz, and so on. These harmonics can be used, when picked up by any receiver, to calibrate that receiver, as the harmonics are quite accurate in frequency (see later for an assessment of accuracy). However, a gap of 100 kHz between harmonics is rather wide for most purposes, so we reduce this frequency to 25 kHz, so that the harmonics are then 25 kHz apart, thus producing a much more useful set of *marker points*. A calibrator such as this is often called a *crystal marker*, producing these marker points from 25 kHz to beyond 30 MHz.

The circuit

The circuit diagram is shown in **Figure 1**. The circuit around TR1 is the fundamental oscillator, and its frequency is controlled by the quartz crystal, X1. Even crystal oscillators are not 100% accurate, and the small trimmer capacitor, TC1, is able to 'pull' the frequency to one which is nearer the

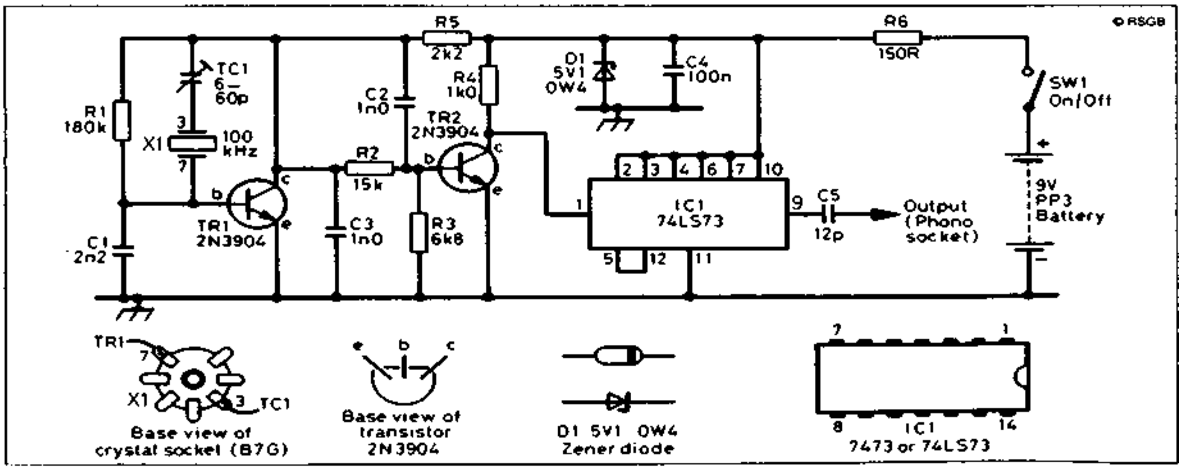
© RSGB
R1
180k
TC1
6–
60p
3
X1
100
kHz
7
b
c
e
TR1
2N3904
C1
2n2
C2
1n0
R2
15k
C3
1n0
R3
6k8
R5
2k2
R4
1k0
TR2
2N3904
D1
5V1
0W4
C4
100n
R6
150R
SW1
On/Off
+
9V
PP3
Battery
−
2
3
4
6
7
10
1
IC1
74LS73
9
5
12
11
C5
12p
Output
(Phono
socket)
TR1
7
X1
3
TC1
Base view of
crystal socket (B7G)
e
b
c
Base view of
transistor
2N3904
D1 5V1 0W4
Zener diode
7
1
8
IC1
14
7473 or 74LS73

Figure 1 Circuit diagram of the calibrator which uses easily obtainable components

correct one. TR2 is a *buffer* stage, which isolates the oscillator from the rest of the circuit. It acts as a switch, and applies a good signal (switching between 0 and 5 V) to the input of IC1.

IC1 is an integrated circuit which can be connected to do several things. It is connected here to divide the incoming frequency by a factor of 4, producing on pin 9 an output frequency of 25 kHz.

The combination of R6, C4 and D1 produces a supply of 5 V for IC1; it would be damaged if the battery voltage were applied to it. You will no doubt be ready to assemble the circuit, so here is some information for you.

Construction

If possible, always build a circuit in individual stages, which you can test as you go along. It is not always easy in small projects like this, but even the crystal calibrator can be split into two for construction and testing.

It is an ideal project for assembly on Veroboard of the copper-strip type; the prototype layout is shown in **Figure 2**, on a board measuring 11 holes by 24 strips. First of all, remove the copper strip at the locations shown, using a 3 mm drill rotated between the fingers. Hold the board up to the light to ensure that there are no pieces of copper swarf bridging adjacent strips, and that the copper is completely removed where it should be! Assemble the circuit from left to right, but do not wire up anything around IC1. Be aware that the diagram shows the board *from the component side*. Although you

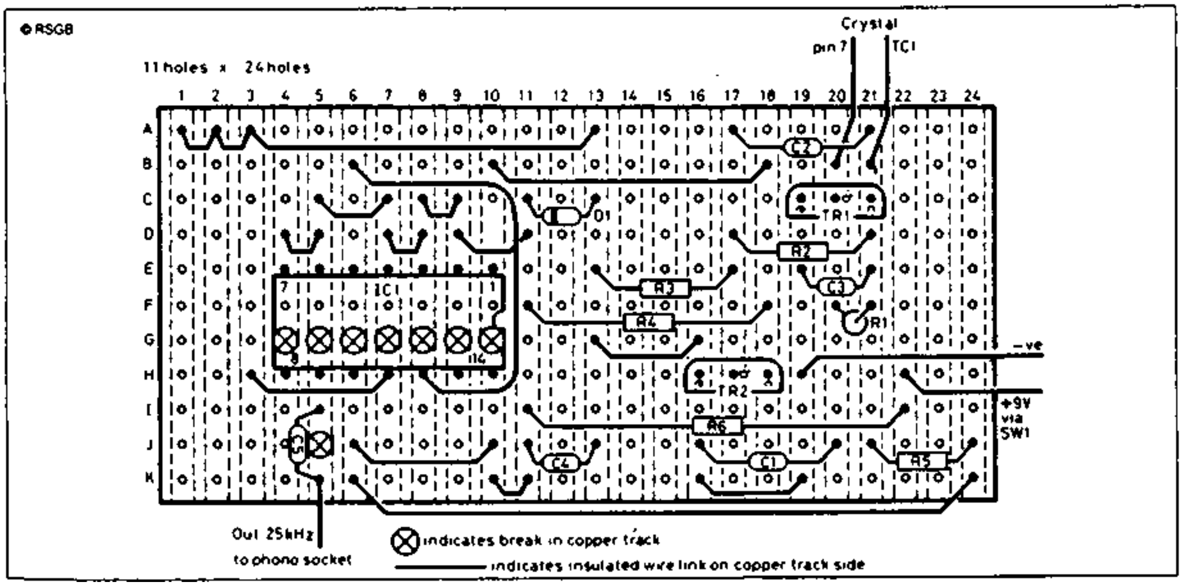
©RSGB
11 holes x 24 holes
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
A B C D E F G H I J K
Crystal
pin 7
TC1
C2
TR1
D1
R2
R3
C3
R4
R1
IC1
7
9
14
−ve
TR2
+9V via SW1
R6
C4
C1
R5
Out 25kHz to phono socket
⊗ indicates break in copper track
—— indicates insulated wire link on copper track side

Figure 2 Stripboard, such as Veroboard, should be cut as shown to approx. 62 × 28 mm

can use virtually any kind of crystal, you will need to mount it firmly. If you use a small one, it can be soldered directly in position on the board. With one in a valve envelope, you will need a B7G valveholder. Be prepared to mount the circuit in a **metal** container; the prototype used an (empty) tin of tuna! The valveholder mounts well on any metal case.

Disconnect the aerial of your receiver, and replace it with about 30 cm of wire laid near your oscillator circuit. Connect the battery and switch on. Tune the receiver around until you can hear a whistle on SSB. Rotate the tuning knob to reduce the frequency of the whistle and, as the whistle becomes too low to hear, you have reached one of the calibration markers, and your frequency will be an integer multiple of 100 kHz. If you are already using a calibrated receiver, you will be able to verify this. Going up or down in frequency should locate another marker 100 kHz away, and so on right through the receiver's tuning range. If the signals coming from the circuit are very weak, switch off the calibrator and connect another 30 cm piece of insulated wire to the collector of TR2, and lay it close to (but *not* touching) the wire from the aerial connector. Switch on an try again; you should not now have a problem! If you have access to a multimeter, check that the voltage at the collector of TR1 is very close to 5 V. If it is close to 9 V, you have connected the diode, D1, the wrong way round!

Having verified that the oscillator is working, you can now wire up the integrated circuit socket, being careful to put the notched end of the socket pointing towards R4, the collector resistor for TR2.

Check your connections around IC1, and when you are satisfied that they are correct, line up IC1 with its socket, making sure that the notched ends are together, and press down *gently* to insert the IC into its socket. Insert the 30 cm piece of insulated wire into the output socket, connect the battery and switch on. You should still hear whistles in your receiver, but now they should be 25 kHz apart, rather than 100 kHz apart.

All that now remains to be done is to mount the circuit rigidly inside whatever casing your have chosen. Make sure that none of the connections under the board touch the metal case, and secure the valveholder, on/off switch and output connector to the case. You now have a completed crystal calibrator.

Calibration

The simplest way to calibrate your circuit is with a frequency counter. Most clubs will have one of these and, if not, will know someone who has! Connect it to the collector of TR2, where the frequency should be 100 kHz. Do *not* connect it to any part of the circuit around TR1, or you may alter the frequency you are trying to measure! If the frequency is not exactly 100.000 kHz, rotate TC1 until it is (or is as close as you can get it). Now your calibrator is as accurate as the counter with which you have calibrated it.

Accuracy

Despite your best efforts at calibration, by whatever means, your crystal will *never* have a constant frequency. Such a thing is a scientific impossibility. It is usual to express the accuracy of a crystal in parts-per-million (ppm), and it is governed by many things, principally its temperature. You will be *very* fortunate if your circuit maintains an accuracy better than about ±10 ppm. Expressed in figures, it means that the true frequency can be anywhere between 99.999 kHz and 100.001 kHz, i.e. within 1 Hz of the correct frequency.

Although this may seem more than adequate, it is as well to remember that the accuracy degrades as the frequency increases. At 1 MHz the error will be ±10 Hz and at 10 MHz it will be ±100 Hz. At 30 MHz it will be 300 Hz. Even so, this should be acceptable for most non-critical applications.

Parts list

Resistors: all 0.25 watt, 5% tolerance		**Maplin code**
R1	180 kilohms (kΩ)	M180K
R2	15 kilohms (kΩ)	M15K
R3	6.8 kilohms (kΩ)	M6K8
R4	1 kilohm (kΩ)	M1K0
R5	2 kilohms (kΩ)	M2K0
R6	150 ohms (Ω)	M150R
Capacitors		
C1, C2, C3	1 nanofarad (nF) or 1000 picofarads (pF)	WX68Y
C4	100 nanofarads or 0.1 microfarad (μF)	YR75S
C5	12 picofarads (pF)	WX45Y
TC1	60 picofarads (pF) trimmer	WL72P
Semiconductors		
TR1, TR2	2N3904 npn	QR28F
IC1	7473 or 74LS73N dual JK flipflop	YF30H
D1	5.1V 500 mW Zener	QH07H
Additional items		
100 kHz crystal		
B7G valveholder (found at most rallies)		
14-pin DIL IC socket		BL18U
Phono socket		YW06G
PP3 battery connector		HF28F
PP3 battery		
On/off switch		FH00A
Veroboard		JP47B
Metal case as required		

30 A simple short-wave receiver – Part 1

Introduction

No apologies for yet another design of short-wave radio. This is the beauty of our hobby – there is always something new to try. Not all circuits operate in the same way; not all circuits work equally well; sometimes a simple design suits the operator better than a complicated design. So here is another one for your consideration; it's a good project for a novice, *and* a good one-evening project for someone more experienced.

A basic description

Figure 1 gives the block diagram of the receiver, which employs a *regenerative detector*, one of the earliest techniques by which excellent *selectivity* (the ability to separate two stations very close in frequency) could be combined with good *sensitivity* (the ability to pick up very weak stations) using a very simple circuit. The receiver falls into the category known as *tuned radio frequency* (TRF), meaning that the whole circuit (prior to the extraction of the modulation from the carrier) operates at the incoming radio frequency. In other words, it is *not* a superhet.

As you can see, the regenerative detector has what is called a *feedback loop*, which feeds a small amount of the output signal back to the input. You have heard the effects of feedback with a public address system, when the microphone gain is too great – everything becomes very loud and then bursts into a deafening squeal! This is exactly what the feedback does here, except that it is carefully controlled, thus providing both gain and selectivity. The resulting audio is then amplified for use with headphones.

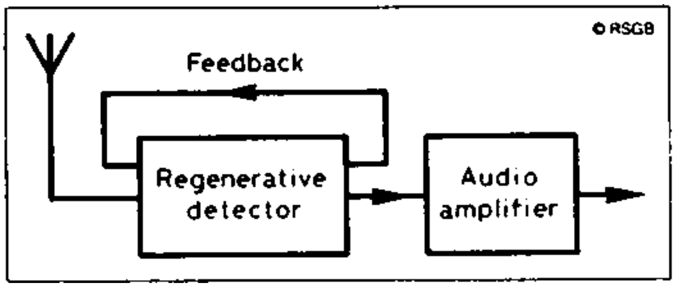

Figure 1 Just two simple stages make up the circuit

The circuit

The circuit of **Figure 2** shows the complete system. TR1 is an untuned *field-effect transistor* (FET) stage, and is used to match the aerial to the next stage. Occasionally, a regenerative detector produces unwanted signals, and TR1 also prevents them from reaching the aerial and being transmitted! The smaller of the two windings on T1 (the *primary* winding) will match a low-impedance aerial, the capacitor input matching a high-impedance aerial.

Ignore TR2 for the moment – the next stage in the signal path is TR3, the detector, using another FET. The tuned circuit is formed by T2, VC1 and VC2 (remember TR1 is untuned). The primary winding on T2 taking the output from TR1. The reason for having two variable tuning capacitors, one large, the other small, is that the large capacitor is the main tuning capacitor, while the small one is used as a *bandspread* control (for very fine tuning). The tuned signals are then detected (converted to audio frequencies) by TR3.

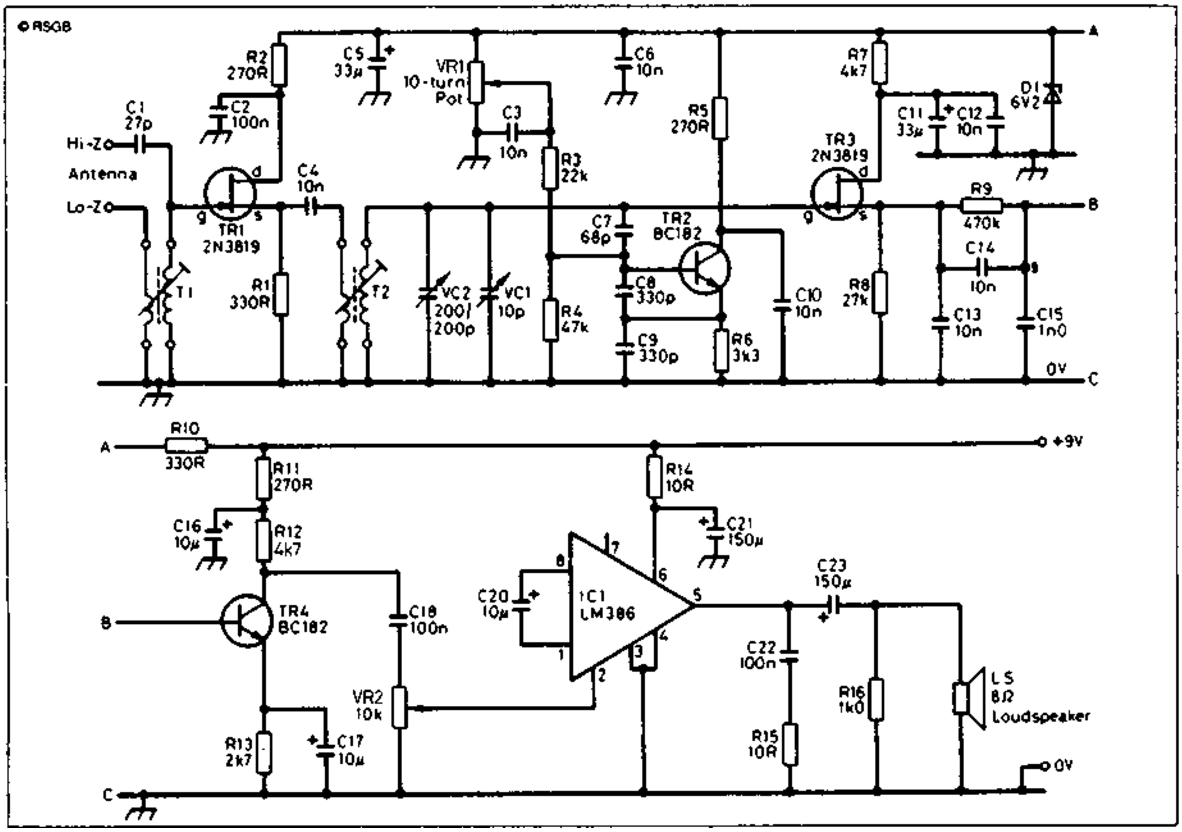
©RSGB
C1 27p
Hi-Z
Antenna
Lo-Z
T1
R2 270R
C2 100n
TR1 2N3819
g
d
s
R1 330R
C4 10n
C5 33µ
VR1 10-turn Pot
C3 10n
R3 22k
T2
VC2 200/200p
VC1 10p
R4 47k
C6 10n
C7 68p
TR2 BC182
C8 330p
C9 330p
R5 270R
R6 3k3
C10 10n
R7 4k7
C11 33µ
C12 10n
D1 6V2
TR3 2N3819
R8 27k
R9 470k
C14 10n
C13 10n
C15 1n0
A
B
0V
C
R10 330R
+9V
R11 270R
C16 10µ
R12 4k7
TR4 BC182
C18 100n
VR2 10k
R13 2k7
C17 10µ
R14 10R
C21 150µ
C20 10µ
IC1 LM386
1
2
3
4
5
6
7
8
C23 150µ
C22 100n
R15 10R
R16 1k0
LS 8Ω Loudspeaker
0V

Figure 2 The circuit diagram showing the 10-turn potentiometer VR1. This is the regeneration control for the circuit

TR2 is a *Q-multiplier* stage. It is a Colpitts-type oscillator which uses C8 and C9 to give feedback and produce the oscillation. C7 couples the tuned circuit (and the input to TR3) to TR2. The tuned circuit thus controls the frequency of oscillation of TR2 and the signal passed on to TR3.

The secret of the ease of operation of this receiver is VR2, a 10-turn potentiometer (or *helipot*). The resistance wire is wound in the form of a helix, giving a much greater wire length than in a normal potentiometer, and the shaft must be turned ten times to cover the whole length. Helipots are very useful when very fine adjustments have to be made. Here, VR2 sets the regeneration (or *reaction*) level, depending on the type of signal you are receiving, as will be discussed in Part 2.

TR4 provides the first stage of audio amplification and, after the volume control, VR2, the audio amplifier integrated circuit, IC1, will drive a small loudspeaker or headphones.

In Part 2, the construction will be discussed, together with the choice of aerial, the parts list, and advice on using the receiver.

31 A fruit-powered medium-wave radio

Introduction

This is a one-evening project that will result in a working medium-wave (MW) radio, and will also teach you a little about the way electricity can be generated from the right metals and a little (safe) acid. All you need are three lemons or other citrus fruit, three pieces of copper and three pieces of zinc (or galvanised metal) for your power supply.

Construction

Figure 1 shows the circuit and **Figure 2** its layout on a simple ‘plug-in’ prototype board. The six pieces of metal are connected as shown, to wire the three lemons in series; use ordinary wire between each lemon and the next. If you have a meter to measure the total voltage, it should be about 1.8 V. Use a standard ferrite rod, and wind on it about 40 turns of single-conductor PVC-insulated wire.

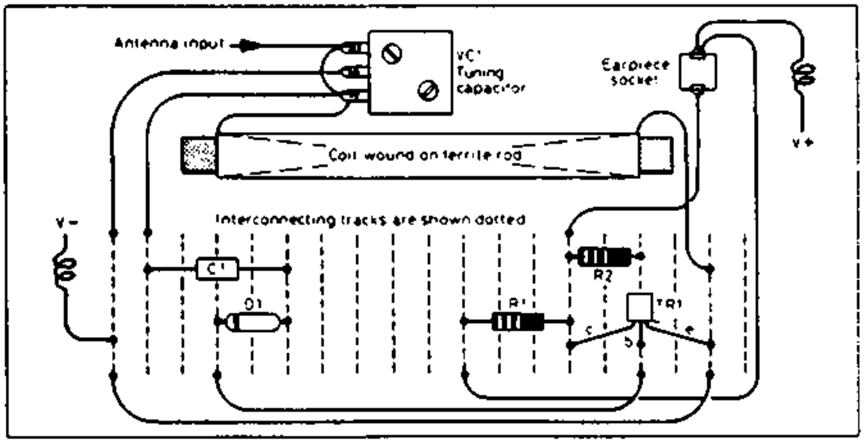
Antenna input
VC1
Tuning
capacitor
Earpiece
socket
V+
Coil wound on ferrite rod
V−
Interconnecting tracks are shown dotted
C1
D1
R1
R2
TR1
c
b
e

Figure 1 Most parts are plugged into the board as shown – soldering is not required

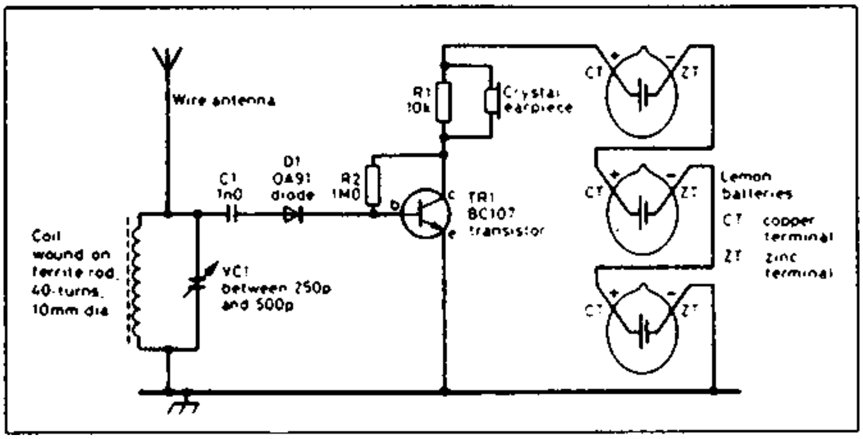
Wire antenna
C1
1n0
D1
OA91
diode
R2
1M0
R1
10k
Crystal
earpiece
b
c
e
TR1
BC107
transistor
Coil
wound on
ferrite rod,
40-turns,
10mm dia
VC1
between 250p
and 500p
CT
ZT
+
−
Lemon
batteries
CT copper
terminal
ZT zinc
terminal

Figure 2 Three lemons power the radio, which gives good results for such a simple circuit

Wire up the circuit on the board as illustrated in Figure 2. Soldering is not required with this type of board – just plug in the components and the wires. Only one transistor is needed. The tuning capacitor, VC1, selects the station you want to hear, and D1 helps to remove the carrier from the RF signal. The resulting audio signal is fed to TR1, a small transistor audio amplifier, which makes the signal big enough to drive a crystal earpiece comfortably. Walkman-type earphones will not work, so invest in a crystal earpiece which you can use in several other projects, too! If you use a smaller capacitor than that specified for VC1, you will need more turns on the aerial coil.

Tests on the prototype indicated that the radio will run for about a week on three lemons!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	1 megohm (MΩ)
R2	10 kilohm (kΩ)

Capacitors

C1	1 nanofarad (nF) min. ceramic
VC1	250–500 picofarad (pF) variable

Semiconductors

TR1	BC107 npn (or BC108, BC109C)
D1	OA90, OA91 germanium (not silicon)

Coil

L1	2 metres of single-conductor insulated wire on a standard ferrite rod

Additional items
- Plug-in prototype board, e.g. Maplin YR84F
- Wire aerial at least 3 m long
- Crystal earpiece
- Three juicy fruits

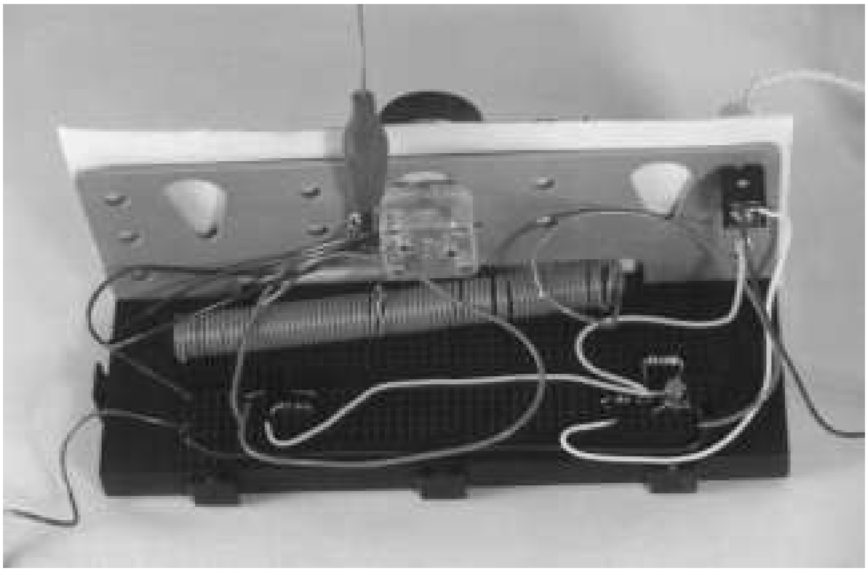

32 A capacitance bridge

Introduction

It is always valuable to have a means of measuring capacitors amongst your test gear. Universal LCR bridges (i.e. systems that can measure inductance, capacitance and resistance) can be found at most rallies, and you can work your way up to owning one of these. In the meantime, a piece of home-made equipment gives you experience as well as resulting in a useful measuring instrument.

The circuit described here is called a *capacitance bridge*, because it balances the effects of one resistor/capacitor pair against another; if one capacitor has an unknown value, then the other can be calculated. The basic bridge circuit is shown in **Figure 1**. To avoid having calculations to perform, this instrument will be calibrated by using capacitors of known values. The bridge is a useful way of performing measurements, because a knob is turned until there is a null in the signal from an earpiece or loudspeaker. The ear is very precise in being able to perceive nulls, which makes the bridge easy to use and reasonably accurate. At the null, the bridge is said to be *balanced*.

How does it work?

Figure 2 is the circuit diagram for this capacitance bridge. Transistors TR1 and TR2 form an oscillator. This is the audio oscillator shown in Figure 1, and produces an alternating voltage which is fed to the bridge. RV1 (in the collector lead of TR2) replaces both R1 and R2 in Figure 1 – that part of

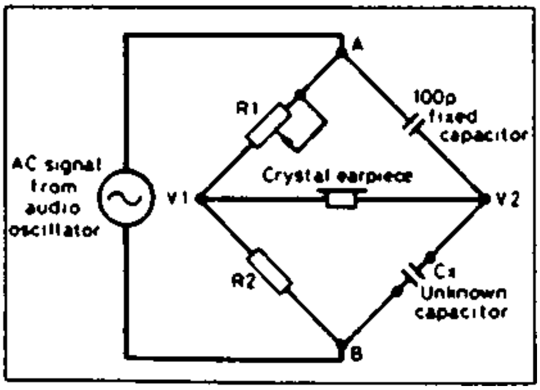

Figure 1 Simplified circuit of a capacitance bridge. R1 is adjusted for minimum sound

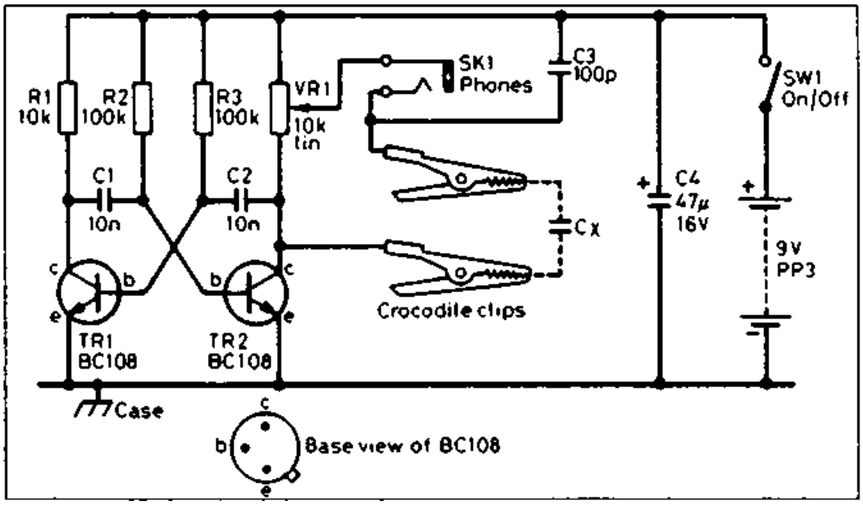
R1
10k
R2
100k
R3
100k
VR1
10k
lin
SK1
Phones
C3
100p
SW1
On/Off
C1
10n
C2
10n
C4
47µ
16V
Cx
9V
PP3
c
b
e
TR1
BC108
TR2
BC108
Crocodile clips
Case
Base view of BC108

Figure 2 Transistors TR1 and TR2 give an audio signal which is adjusted by variable resistor RV1

RV1 *above* the wiper represents R1, while that part *below* the wiper represents R2. The voltage on one side of the earpiece is determined by the ratio of these values, and is adjusted by rotating RV1. The voltage on the other side of the earpiece is determined by the ratio of C3 to Cx, where Cx is the unknown-valued capacitor. When these two voltages are the same, the bridge is balanced and there is no current through the earpiece.

Figure 3 shows the layout of the components on the matrix board of the prototype. It measures 10 holes by 14 holes, and is of the plain type, i.e. no copper strips. All earth connections are taken to a single solder tag. When

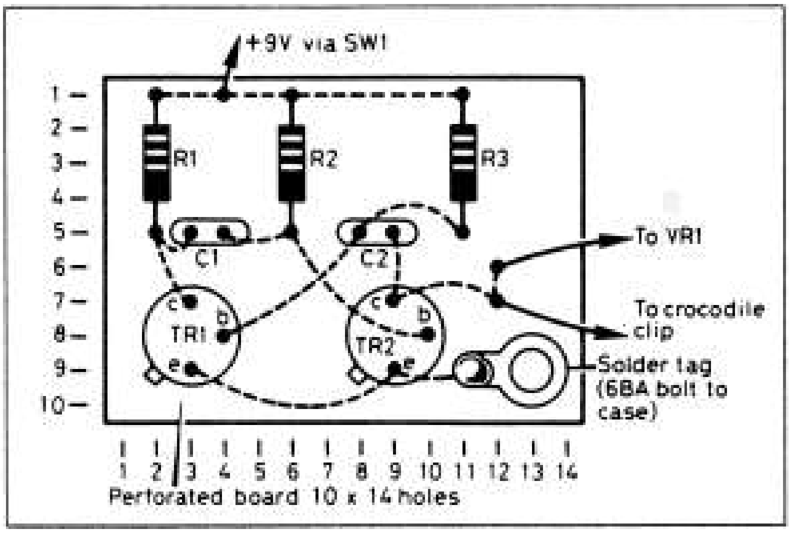
+9V via SW1
1
2
3
4
5
6
7
8
9
10
R1
R2
R3
C1
C2
To VR1
To crocodile clip
c
b
e
TR1
c
b
e
TR2
Solder tag
(6BA bolt to case)
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Perforated board 10 x 14 holes

Figure 3 Components are soldered together on a small prototype board

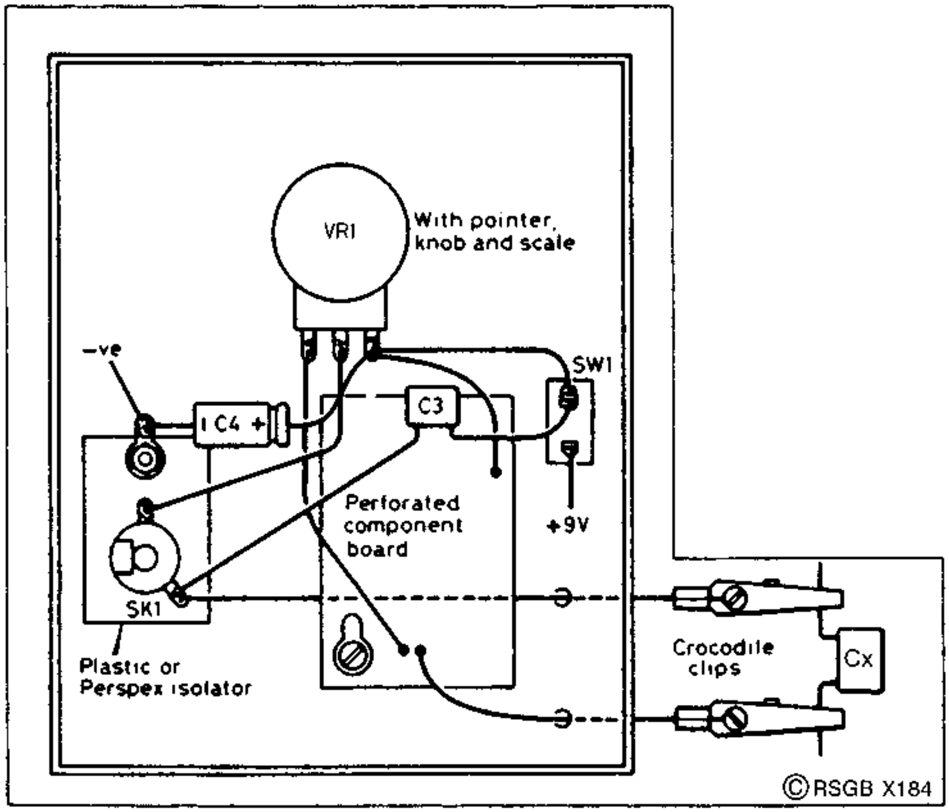
VR1
With pointer, knob and scale
–ve
C4
C3
SW1
Perforated component board
+9V
SK1
Plastic or Perspex isolator
Crocodile clips
Cx
©RSGB X184

Figure 4 Layout of the parts inside an aluminium box. The earpiece socket is insulated from the case

mounting the board inside a metal box, a long screw is used with extra nuts, to earth the board to the case while providing a stand-off, thus preventing any unwanted short-circuits between the board connections and the case. Be aware that a *crystal* earpiece must be used; low-impedance headphones of the Walkman type are not satisfactory. Solder two flying leads as shown in **Figure 4**, about 15 cm long, terminated in small crocodile clips for attaching to the unknown capacitor.

Calibration

After checking the circuit carefully, attach the battery and switch on; a buzzing noise should be heard in the earpiece. This is the first sign that everything should be OK. If there is no buzz, switch off and recheck the connections.

You will now need a range of close-tolerance (1%) silver-mica capacitors covering the range 10 to 1000 picofarads (pF). Arrange the capacitors in ascending order and connect them to your bridge in sequence. After having prepared a neat piece of card or paper mounted behind the knob on the front panel, mark the dial at the positions of the nulls for all the capacitors. You have now calibrated your capacitance bridge. If you are more likely to want to measure larger capacitors, replace C3 by a 1 nanofarad (nF) capacitor, and the bridge will measure up to 10 nF approximately.

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	10 kilohms (kΩ)
R2, R3	100 kilohms (kΩ)
RV1	10 kilohms (kΩ) linear
Capacitors	
C1, C2	10 nanofarads (nF) ceramic
C3	100 picofarads (pF) silver mica or polystyrene
C4	47 microfarads (μF) 16 V electrolytic
Semiconductors	
TR1, TR2	BC108 npn
Additional items	
SW1	SPST on/off switch
	Battery connector, PP3 type
	Earpiece, crystal type
SK1	3.5 mm jack socket for earpiece
	Battery, 9 V PP3
	Matrix board, approx. 10 holes by 14 holes
	Aluminium case, approx. 10 × 8 × 5 cm

33 A simple short-wave receiver – Part 2

Introduction

In Part 1, the design of this receiver was discussed in some detail. Now we are going to put it all together and see how it works. The receiver is laid out on a printed-circuit board (PCB) or on a matrix board.

Construction

The layout of the components is shown in **Figure 1**. Identify each part separately, insert it into the holes in the board and solder carefully. Long leads may be cropped before or after soldering, depending on your skill and preference. All electrolytic capacitors, T1, T2 and IC1, must be connected correctly. The front-panel controls are connected to the PCB terminals shown in Figure 1.

The layout of the controls and the placing of the board inside the case are matters of personal preference. The size of the prototype front panel is shown in **Figure 2**. The prototype had a slow-motion drive fitted to VC2, the main tuning capacitor. This required the capacitor to be fitted on its own small panel. The bandspread control does not need any form of

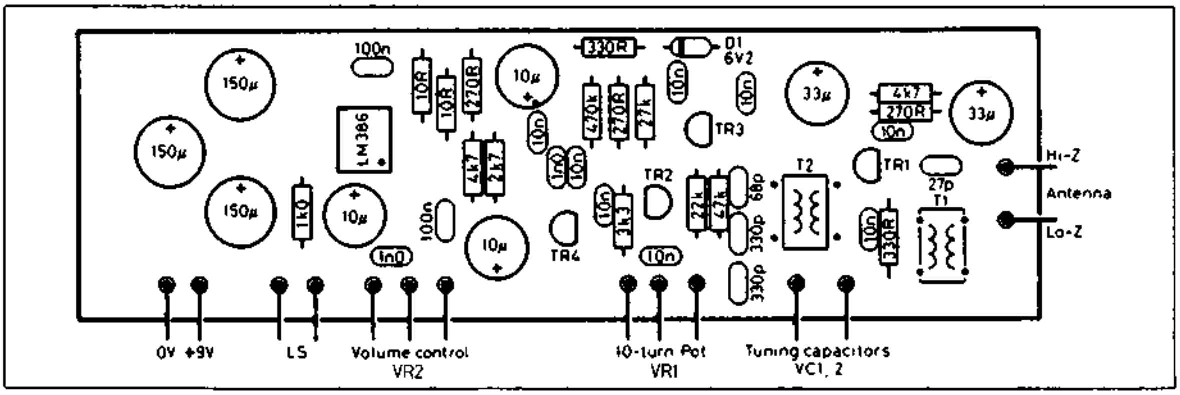
150µ
150µ
150µ
100n
LM386
1kΩ
10µ
1n0
100n
10R
10R
270R
4k7
2k7
10µ
10µ
10n
1n0
10n
TR4
330R
470k
270R
27k
10n
3k3
TR2
10n
D1
6V2
10n
10n
TR3
22k
47k
68p
330p
330p
T2
33µ
4k7
270R
10n
33µ
TR1
27p
T1
10n
330R
Hi-Z
Antenna
Lo-Z
0V
+9V
LS
Volume control
VR2
10-turn Pot
VR1
Tuning capacitors
VC1, 2

Figure 1 Layout of the receiver circuit board

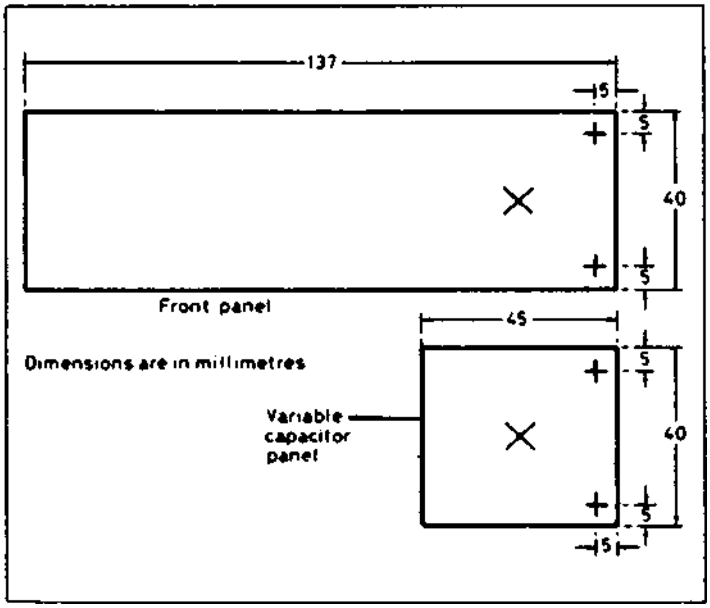

Figure 2 The case can be metal or made from PCB material

slow-motion drive. A tuning dial may be fitted behind the tuning knob on VC2 if required. The volume control, VR2, has its own on/off switch for the battery, which can be mounted behind the back panel, on which are mounted the aerial input sockets and the headphone/speaker socket. Check your soldering carefully, together with all the wiring between the board, the potentiometers and variable capacitors. When you are confident that everything is perfect, connect the battery and switch on.

The first tests

If you have a good aerial and are using an aerial tuning unit (ATU), use the low-impedance input. Excellent results are possible, though, with about 3 metres of wire connected to the high-impedance input. Set the volume control to give a gentle hiss in the headphones or speaker. Advance the reaction control to give a definite hiss. As you tune in an AM station, the hiss will change to a whistle; back off the reaction until the oscillation *just* stops. The receiver is now correctly set for AM reception.

On the amateur bands, the stations will be SSB or CW, and the reaction needs to be set *just above* the oscillation point, and will need slight adjustment from time to time for different qualities of signal. Juggling with the volume, reaction and tuning is part of the pleasure of using regenerative receivers!

In action

Practice is needed for best results. The regenerative receiver is renowned for its versatility in being able to be set up *exactly* for all types and strengths of signal. The basic receiver tunes from 6.5 MHz to 11 MHz approximately; this includes two amateur bands at 7.0 MHz and 10.1 MHz and two broadcast bands.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R10	330 ohms (Ω)
R2, R5, R11	270 ohms (Ω)
R3	22 kilohms (kΩ)
R4	47 kilohms (kΩ)
R6	3.3 kilohms (kΩ)
R7, R12	4.7 kilohms (kΩ)
R8	27 kilohms (kΩ)
R9	470 kilohms (kΩ)
R13	2.7 kilohms (kΩ)
R14, R15	10 ohms (Ω)
R16	1 kilohm (kΩ)
VR1	10 kilohm (kΩ) linear 10-turn potentiometer
VR2	10 kilohm (kΩ) log potentiometer

Capacitors

C1	27 picofarads (pF)
C2, C18, C22	100 nanofarads (nF)
C3, C4, C6, C10, C13, C14	10 nanofarads (nF)
C5, C11	33 microfarads (μF) 16 V electrolytic
C7	68 picofarads (pF)
C8, C9	330 picofarads (pF)
C15	1 nanofarad (nF)
C16, C17, C20	10 microfarads (μF) 16 V electrolytic
C21, C23	150 microfarads (μF) 16 V electrolytic
VC1	10 picofarads (pF) variable
VC2	200 picofarads (pF) variable

Semiconductors

TR1, TR3	2N3819
TR2, TR4	BC182
IC1	LM386
D1	6.2 V 0.5 W Zener

Inductors

T1	3 turns primary, 15 turns secondary, wound on 2-hole ferrite bead, with 28 SWG wire
T2	2 turns primary, 17 turns secondary, wound on a T68-2 toroidal former, with 28 SWG wire

Additional items
- Printed-circuit board
- Battery connector
- PP3 battery
- 8-pin DIL socket
- 8 ohm speaker or headphones (Walkman type)
- Case to suit

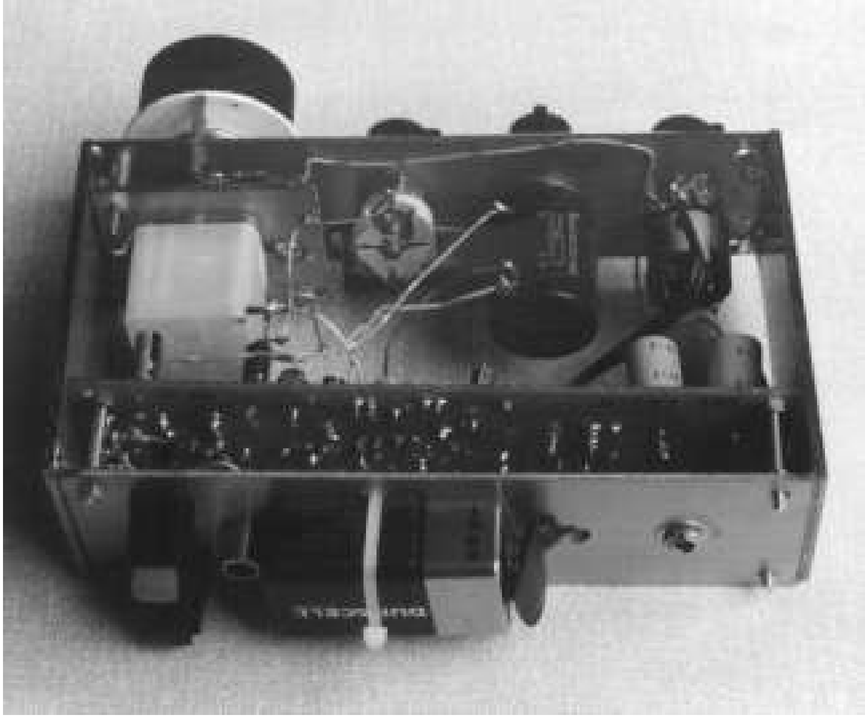

34 A basic continuity tester

Introduction

This little device can be built on a plug-in breadboard, and is ideal for testing fuses and cables, as well as doubling as a signal source for testing amplifiers, etc. It can even be used for testing npn transistors by replacing either of the transistors in the circuit, and seeing if the circuit still works!

Simple and quick to build

Using a plug-in breadboard, this circuit is so simple it could almost be built when you need it, and then dismantled again! If you want the circuit to make a different sound, the components to change are R2, R3, C1 and C2. Always make sure that R2 = R3 and C1 = C2, or the sound may be excessively 'edgy' and lacking in volume.

Basically, the circuit is an oscillator which drives a little loudspeaker directly. The fuse or cable you are testing is connected between the crocodile clips. If there is a current path between the two clips, the current also flows through the circuit, thus operating the oscillator and producing a sound from the loudspeaker.

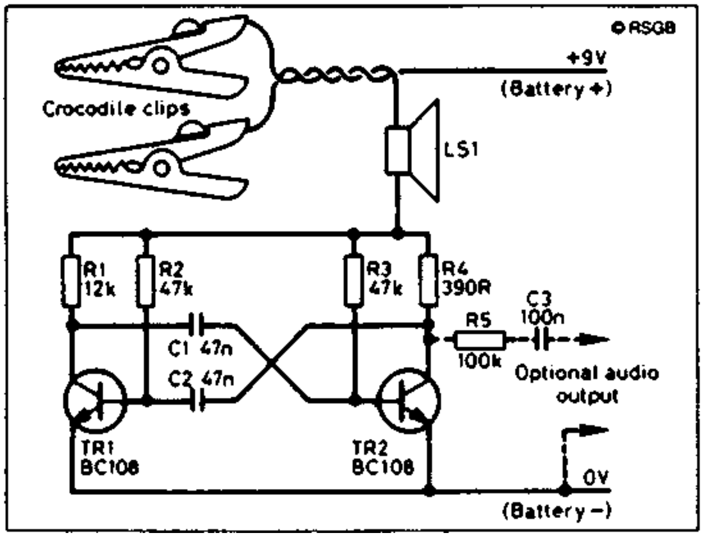
© RSGB
+9V
(Battery +)
Crocodile clips
LS1
R1
12k
R2
47k
R3
47k
R4
390R
C3
100n
R5
100k
C1 47n
C2 47n
Optional audio
output
TR1
BC108
TR2
BC108
0V
(Battery −)

Figure 1 The circuit diagram of the continuity tester

If you want to use the circuit to produce a signal to test an amplifier, for instance, connect the croc clip which comes directly from the battery to the junction of R1, R2, R3 and R4. This supplies current to the circuit while, at the same time, bypassing the loudspeaker. Use the two components R5 and C3 between the oscillator and your amplifier. The loudspeaker is bypassed because you will want to listen to the output from your amplifier, *not* from the oscillator!

Remember, NEVER test any equipment which is still connected to the mains electricity supply. Avoid testing anything which is switched on and has its own power supply. You may damage both your tester ***and*** the circuit you are 'testing'!

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	12 kilohms (kΩ)
R2, R3	47 kilohms (kΩ)
R4	390 ohms (Ω)
R5	100 kilohms (kΩ) optional
Capacitors	
C1, C2	47 nanofarads (nF) ceramic
C3	100 nanofarads (nF), 0.1 microfarads (μF) ceramic optional
Semiconductors	
TR1, TR2	BC108 npn
Additional items	
LS1	Loudspeaker (8 Ω)
	2 crocodile clips
	Prototype circuit board
	PP3 battery connector
	PP3 battery

35 A charger for NiCad batteries

Introduction

The cost of replacing dry batteries can be alleviated to a great extent by the use of rechargeable batteries such as those containing nickel and cadmium (called *NiCads*). They are more expensive than the batteries they replace, but they can be charged hundreds of times and thus prove cheaper in the long run.

NiCads produce only 1.2 V per cell, compared with the 1.5 V of ordinary cells, so before you rush out and buy lots of these, please make sure that the equipment on which you plan to use them will work on the reduced voltage! For example, if you are using four cells to give you a 6 V supply, NiCads will give you only 4.8 V, which is quite a reduction. Using six cells to replace a 9 V battery will give you only 7.2 V. Not all equipment is happy with these reductions!

However, we all use them when we can, and they save substantial amounts of money. Here is the circuit of a charger to keep them in prime condition.

Charging NiCads – the ampère-hour

NiCads require charging at constant current, which means that connecting one across a normal power supply (constant *voltage*) is useless and can destroy it. They need pampering to the extent of needing a long charge (around 16 hours) at a rate dependent upon the *capacity* of the battery. By the capacity of a battery, we mean how much energy it can store. You will probably know that energy is measured in joules. For the purposes of storing energy in batteries, the joule is *not* the ideal unit, so we use one that is! This unit is the ampère-hour (Ah), and must be interpreted with some realism. For example, if the battery is rated at 2 Ah, it will deliver a current of 0.5 amp for 4 hours, or 0.25 amp for 8 hours. Provided the current is not too high, the product of the current (in amps) and the time for which it will flow (in hours) before the battery is flat will always be around 2 Ah. A workable 'rule of thumb' for calculating the charging current is that its value should be around one-tenth of the numerical value of the capacity; so, for our 2 Ah battery, a charging current of around 200 mA (2 ÷ 10 = 0.2 A or 200 mA) would be used.

Constant voltage to constant current

Many integrated circuit (IC) chips are available for use as voltage regulators, i.e. they supply a constant voltage. Most of these can be persuaded to become constant current supplies with one external resistor!

The voltage regulator IC usually has only three connections – 'input', 'output' and 'common'. It is designed (in the case of the LM7805) to produce a constant 5 V output between the 'output' and the 'common' connections, at currents of up to 1 A. If a resistor is connected between these, the IC will maintain 5 V across it. If you look at **Figure 1**, you will see the circuit performing this conversion.

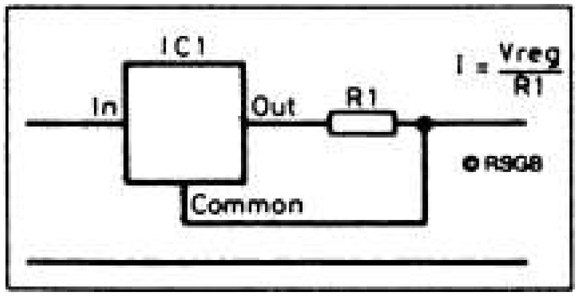

Figure 1 Voltage regulator arranged to produce a constant current

For the previous example, we derived a charging current of 200 mA, so we now need to calculate the value of resistor that will produce this current. Using the equation which is derived from Ohm's law:

$$I = \frac{V}{R}, \text{ from which } R = \frac{V}{I},$$

where R = R1, the resistance in ohms that we are calculating,
V is the voltage across R1 (5 V), and
I is the current flowing (200 mA).

So,

$$\text{R1} = \frac{5}{0.2} = 25\,\Omega.$$

25 ohms is not a 'common' or 'preferred' resistor value, so we must choose the next *largest* value, which is 27 Ω. This reduces the current, but only slightly – it is now 185 mA. When calculating resistor values in power supply circuits, we must *always* check the power that they dissipate and make sure we specify and fit a suitable resistor.

Power (in watts) is the product of the voltage across and the current through a device, so in this case it is given by:

$$\text{Power} = V \times I = 5 \times 0.185 = 0.925\,\text{W}.$$

Rather than use a 1 watt resistor operating very near its limit, it is safer to use a 2 watt resistor operating well within its limits.

Looking again at Figure 1, we now have a constant-current source producing 185 mA, when R1 is a 27 Ω, 2 W resistor. For use with NiCads requiring charging currents other than 200 mA, you will need to repeat the two equations above, using a new value for *I*.

The full circuit and its assembly

This is shown in **Figure 2**, and can be broken down into two parts. The first is the AC to DC conversion produced by the mains transformer, T1, the bridge rectifier, BR1, and the smoothing capacitor, C1. The second is the constant-current section already discussed.

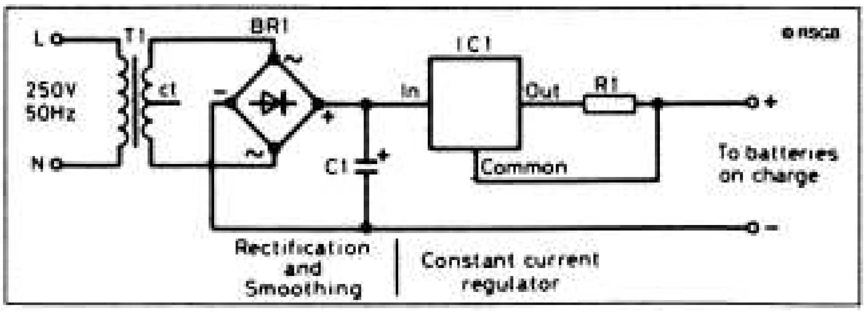

Figure 2 The complete circuit including mains components

The prototype was assembled on matrix board measuring 18 holes by 12 strips, although, as **Figure 3** shows, this is much larger than is strictly necessary. No strip cutting is needed, but make sure that IC1 and C1 are inserted correctly.

Warning! Before you attempt to wire up the transformer and the bridge rectifier, be aware that you will eventually be connecting the circuit to the mains supply, so there are three possibilities for you: (1) get a qualified friend to supervise your completion and testing of the circuit; (2) get your qualified friend to complete and test the circuit for you; (3) replace the transformer and bridge rectifier with a mains adapter.

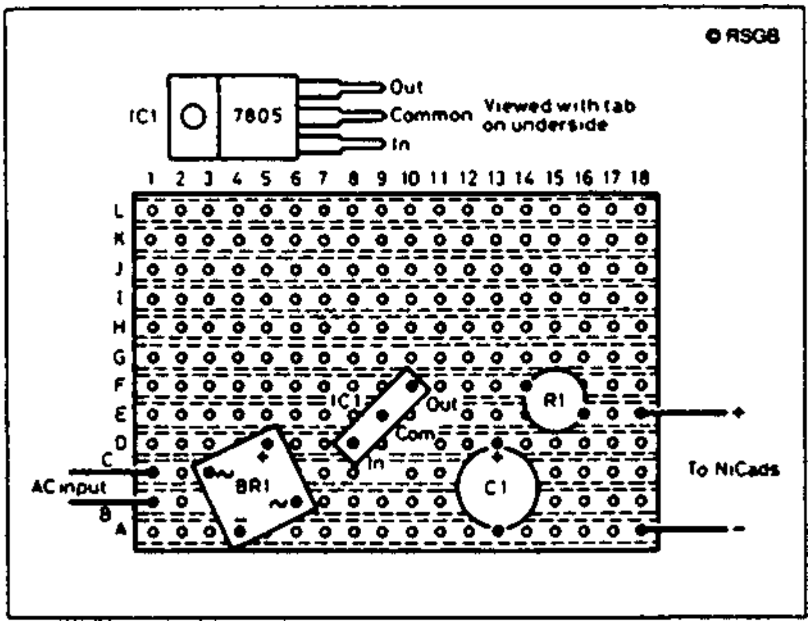

Figure 3 The charger can be built on Veroboard

If you decide to use the mains adapter, its output is connected directly across C1, because T1 and BR1 are now no longer needed. Make sure the polarity (positive and negative) is correct and that the adapter output is set to 12 V.

A quick test

If you have built the mains version, make sure all connections are correct, and that there are no soldered joints which will touch other parts of the circuit. The box must be securely closed before tests begin. The RSGB cannot be held responsible for damage to equipment or batteries! The version using the mains adapter need not be closed during tests.

Switch on. With nothing connected to the output, the unit should run cold. If this is not the case, switch off and recheck your circuit. If all is well, switch on again and connect a DC multimeter (on the current range) across the output. It should indicate only a slight difference from the calculated value of 185 mA. You can now charge your NiCads!

Other charging currents can be set by having different values for R1, perhaps selectable by a rotary switch. Remember to make sure that the values of both resistance *and* power dissipation are correct, and don't exceed the 250 mA rating of the transformer (or the 1 amp rating of the IC if you are using a bigger transformer).

Parts list

		Maplin code
Resistor		
R1	As required – see text	
Capacitor		
C1	1000 microfarads (μF) 35 V electrolytic	FF18U
Semiconductors		
IC1	LM7805 5 V, 1 A regulator	CR14Q
BR1	W005 50 V, 1 A full-wave rectifier	QL37S
Transformer		
T1	9–0–9 V, 250 mA sub-min. transformer	YN15R
Additional items		
Case		FG41U
Veroboard, 18 holes by 12 strips		
Plug to suit NiCads		
Double-sided sticky tape as required		
Insulated wire for battery connections		

36 An 80 metre crystal-controlled CW transmitter

Introduction

A simple transmitter is ideally suited to anyone venturing into our marvellous hobby for the first time. If you are put off by the complexities and prices of the ‘black boxes’ then this is the transmitter you’ve been looking for!

The circuit

The complete circuit is shown in **Figure 1**. It is a single-transistor crystal oscillator which is *keyed* (switched on and off) by the Morse key in the +12 V supply rail. The circuit first appeared in *QST* (the American equivalent of *RadCom*) in 1982, and has since appeared in a modified form in the *ARRL Handbook*. The circuit can produce about 2 watts on 80 m, and can be constructed on a piece of plain matrix board; the prototype board measured 7 cm by 3 cm, and its layout is shown in **Figure 2**.

Note that a bare copper wire runs along the bottom edge of the board to act as an earth wire for the relevant components. The only 'difficult' part of the construction is the winding and wiring of T1. The main winding is 38 turns of 26 SWG enamelled copper wire, and there are two link windings of four turns each. Make sure that all windings are wound the correct way round the toroid – Figure 1 shows this and should be studied carefully. If the windings do not have the correct *sense* (i.e. a clockwise coil has been wound anticlockwise, or vice versa), or have been connected incorrectly to the rest of the circuit, the oscillator will not work!

When putting components on the board, wire in the crystal socket *without the crystal in it*. Crystals do not like to be subjected to the horrors of a soldering iron, so keep your crystal to one side during the construction process!

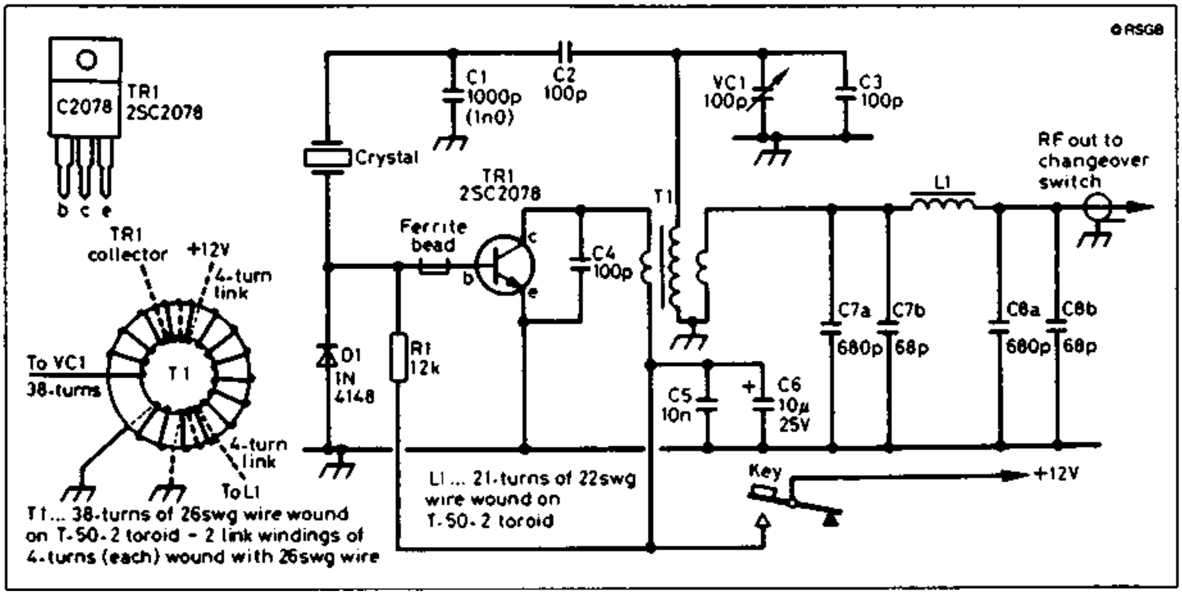
©RSGB
C2078
TR1
2SC2078
b c e
TR1 collector
+12V
4-turn link
To VC1
38-turns
T1
4-turn link
To L1
T1... 38-turns of 26swg wire wound on T-50-2 toroid - 2 link windings of 4-turns (each) wound with 26swg wire
C1 1000p (1n0)
C2 100p
VC1 100p
C3 100p
Crystal
TR1 2SC2078
Ferrite bead
b
c
e
C4 100p
T1
L1
RF out to changeover switch
C7a 680p
C7b 68p
C8a 680p
C8b 68p
D1 1N 4148
R1 12k
C5 10n
C6 10µ 25V
Key
+12V
L1 ... 21-turns of 22swg wire wound on T-50-2 toroid

Figure 1 The simplest form of transmitter is a keyed crystal oscillator. Note that L1, C7 and C8 make up a *low-pass filter* which reduces unwanted *harmonics* (outputs at the transmitter frequency multiplied by 2, 3, 4, etc.)

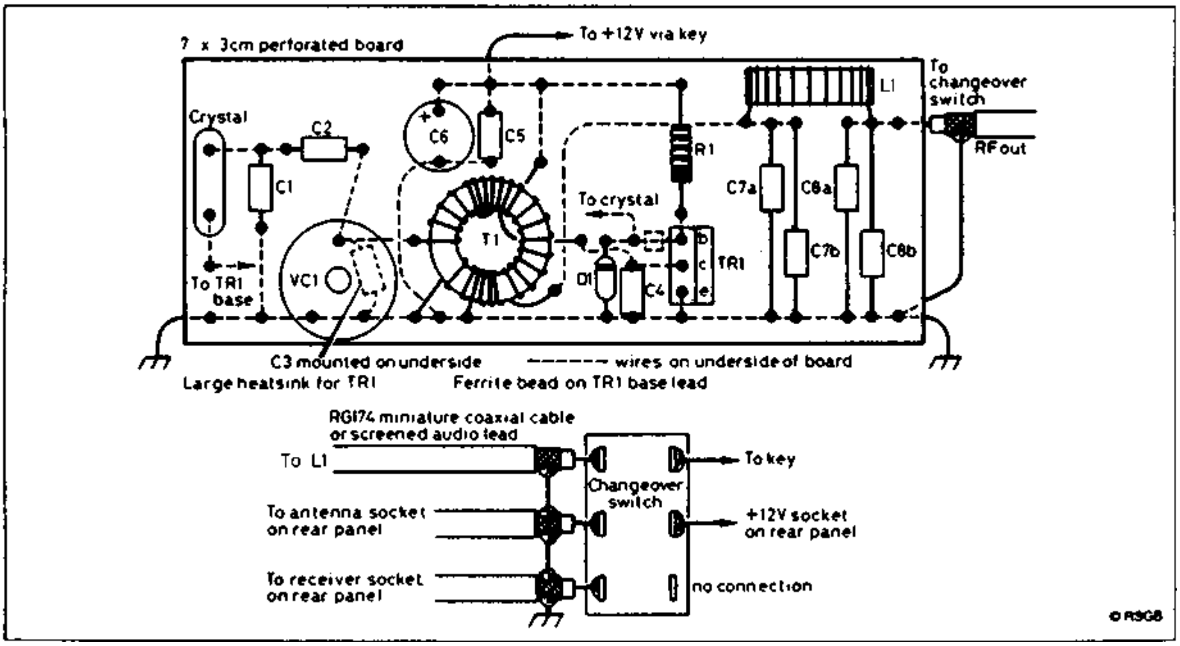
7 x 3cm perforated board
To +12V via key
Crystal
C2
C6
C5
L1
To changeover switch
RF out
C1
R1
C7a
C8a
To crystal
T1
b
c
e
TR1
C7b
C8b
To TR1 base
VC1
D1
C4
C3 mounted on underside
------ wires on underside of board
Large heatsink for TR1
Ferrite bead on TR1 base lead
RG174 miniature coaxial cable or screened audio lead
To L1
To antenna socket on rear panel
To receiver socket on rear panel
Changeover switch
To key
+12V socket on rear panel
no connection
© RSGB

Figure 2 Component layout is straightforward on a 'matrix board' (it has holes but no copper strips). The dotted lines are the connecting wires. The lower part shows the connections for the transmit/receive switch

The output filter, which comprises C7, C8 and L1, is a *low-pass* filter, which helps reduce any *harmonics* present in your signal. Harmonics are integral multiples of your transmitter frequency, so if you are transmitting on a frequency, *f*, harmonics will be present at frequencies 2*f*, 3*f*, 4*f*, . . . and so on. L2 is another inductance using 22 SWG enamelled copper on a ferrite toroid. The changeover switch is external to the transmitter board, and is used to switch your aerial between the transmitter and the receiver; its wiring is shown in Figure 2.

Use a dummy load

A dummy load enables you to test your circuit without actually transmitting a signal. If you haven't such a thing already, it is easy to construct one to use with this transmitter. Don't use it for transmitters of more than 2 watts output, though. Use two 100 ohm, 1 watt resistors, connected in parallel across the end of a short piece of coaxial cable, terminated in a BNC, PL259 or N-type free plug, as shown in **Figure 3**. Plug this into the aerial socket on your transmitter, plug in your crystal and connect the transmitter to a 12 V supply. Have another receiver switched on and tuned to the crystal frequency. Although the radiation from your dummy load is minimal, it will

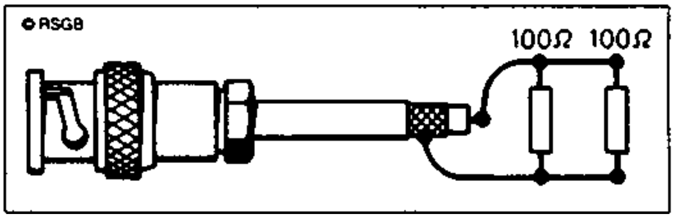

Figure 3 A simple 2 W dummy load can be made from two 100 Ω (ohm) resistors in parallel. The plug should match the socket on your transmitter

be enough to be picked up by a receiver in the same room. Send dashes with the Morse key, and adjust VC1 until the received note is *clean*. It should not sound rough, or have a *chirp* (change its frequency during a dash or dot). Avoid tuning for maximum power; this is seldom the correct setting!

You will need to put your completed transmitter in a metal box, using sockets for the power supply, aerial, receiver and Morse key. The sockets can be chosen to match your existing equipment.

Figures 4–7 are taken from the RSGB book *Practical Antennas for Novices*, and may give you some ideas on the type of aerial to be used with your transmitter.

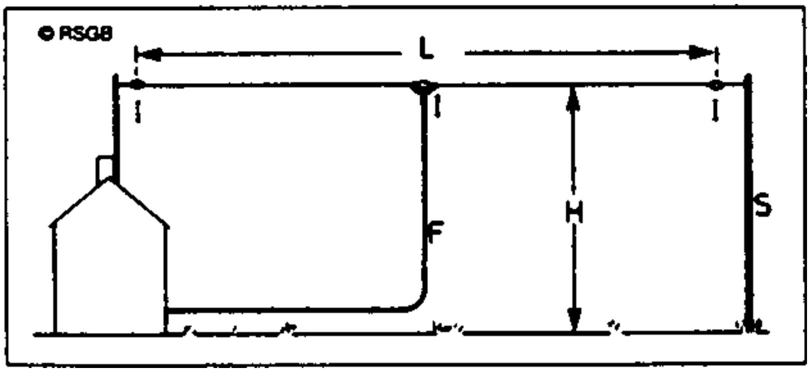

Figure 4 A simple dipole can be very effective. For the 3.5 MHz band, length L is 40 metres and height H should be as large as possible. The far support S can be a tree, pole or building. Insulators I may be home made from strong plastic and the feeder F should be 50 Ω (ohm) coax cable

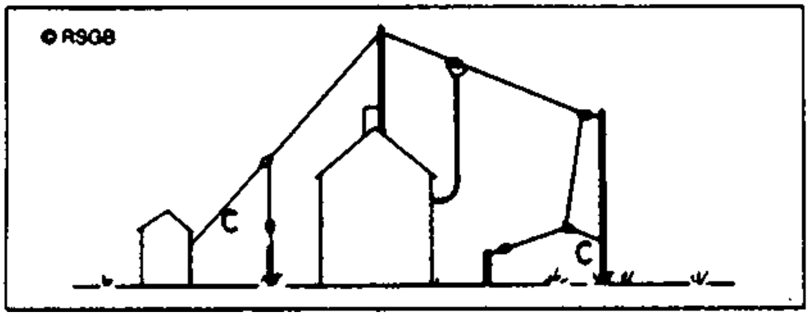

Figure 5 Your signal is radiated mostly from the centre of the dipole so the ends can droop or even be bent but the length may need shortening by a few centimetres because the ground and the bends will detune the dipole. Cords C are best made from strong plastic rope from a sailing or camping shop

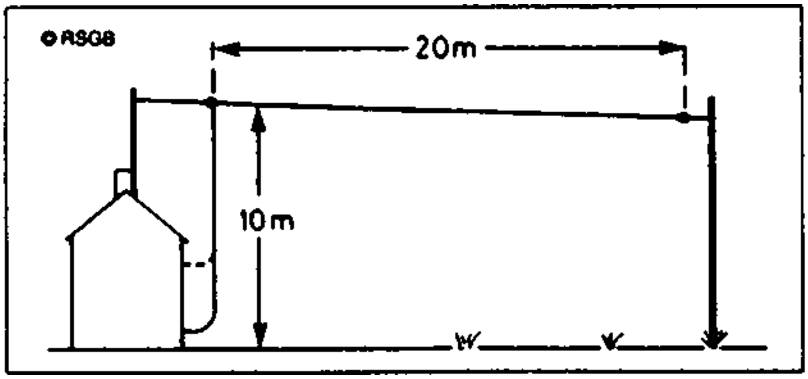
© RSGB
20m
10m

Figure 6 An 'inverted-L' takes up less space than a dipole and doesn't need coax cable. Like the dipole, the end can droop or be bent to save space as in this case most of the radiation comes from the area around the top of the vertical part

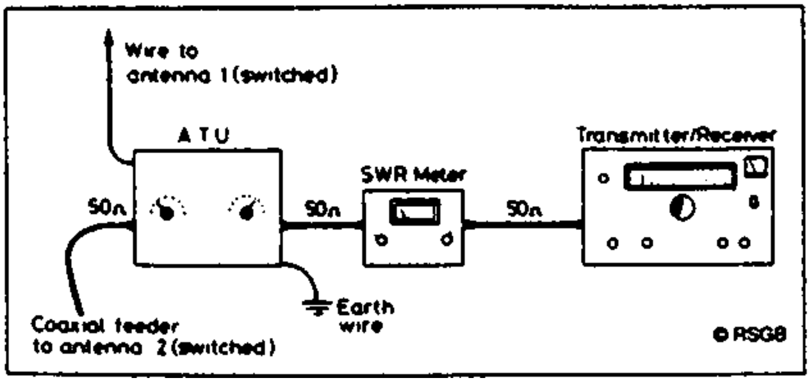
Wire to antenna 1 (switched)
ATU
Transmitter/Receiver
SWR Meter
50Ω
50Ω
50Ω
Earth wire
Coaxial feeder to antenna 2 (switched)
© RSGB

Figure 7 Almost any length of wire more than 10 m or so long will work (though it will work better the longer and higher it is) but an Aerial System Tuning Unit (ASTU or ATU) will be needed

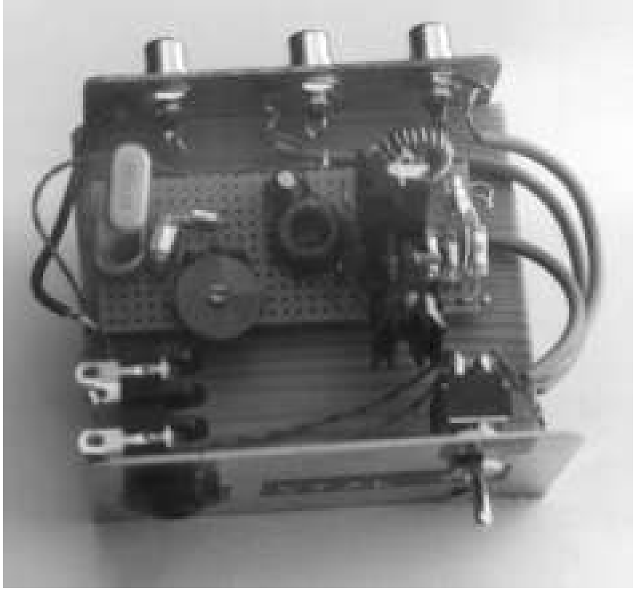

Parts list

Resistor

R1	12 kilohms (kΩ), 0.25 watt, 5% tolerance

Capacitors

C1	1000 picofarads (pF) polystyrene
C2	100 picofarads (pF) polystyrene
C7a, C8a	680 picofarads (pF) polystyrene
C7b, C8b	68 picofarads (pF) polystyrene
C3, C4	100 picofarads (pF) ceramic
C5	10 nanofarads (nF) polyester
C6	10 microfarads (μF) electrolytic 25 V
VC1	100 picofarads (pF) trimmer

Semiconductors

D1	1N4148
TR1	2SC2078 (see sources list)

Inductors

T1	38 turns 26 SWG enamelled copper on T–50–2 toroid, with two link windings of four turns
L1	21 turns 22 SWG enamelled copper on T–50–2 toroid

Additional items
- Ferrite bead
- Crystal (e.g. 3.579 MHz) and holder
- Metal box
- Socket for Morse key.
  This must be totally isolated from the metal of the box, as *both* connections can be at +12 V.
- Sockets for 12 V supply, aerial and receiver
- Switch – DPDT
- Heat sink for TR1
- RG174 miniature coaxial cable for signal leads (see Figure 2)

Component sources

Special components

2SC2078	Cricklewood Electronics Ltd, 40 Cricklewood Broadway, London, NW2 3ET.

37 A solar-powered MW radio

Introduction

What could be more ecologically friendly than a radio powered by the sun's energy? This design is quite standard, and if you have built any of the other medium-wave radios in this series, then this one should present few problems.

The solar panel

The solar panel is to the solar cell as the battery is to the cell; in other words a solar panel is several solar cells connected in series. The solar panel quoted for this radio will generate about 9 V at a current of around 30 mA on a sunny day. The circuit will operate on a supply of around 2 V, so bright sunshine is *not* necessary for satisfactory operation. The volume will be less, of course.

The circuit

The radio uses the ZN415E integrated circuit (IC), connected as shown in **Figure 1**, the complete circuit diagram. The signal is tuned in by the combination of L1 and VC1. L1 is made by winding about 35 turns of 24 SWG enamelled copper wire on a ferrite rod. A standard ferrite rod of 10 cm length and 1 cm diameter is used.

Signals selected by the tuned circuit are passed to IC1, which amplifies the signals and removes the audio component, which is then amplified further by IC2 for driving a small loudspeaker. The removal of the audio component is the process we call *detection*. In addition to this, IC1 provides *automatic gain control* (AGC), which helps to keep the audio signal constant, even when the incoming RF signal may vary due to fading.

The prototype board

Veroboard (also known as matrix board or stripboard) is ideal for the construction of the radio. The layout is shown in **Figure 2**. The board size is 11 strips by 30 holes (**please note that there is no row 'I', so take care with your counting!**). Using a 3 mm (⅛ inch) twist drill rotated between

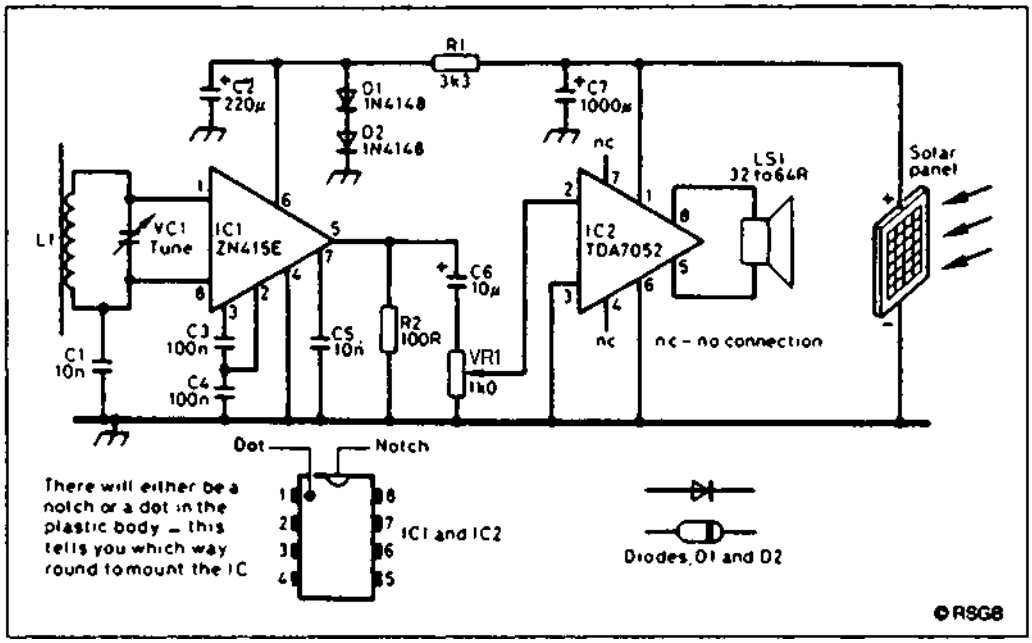
R1
3k3
C2
220µ
D1
1N4148
D2
1N4148
C7
1000µ
nc
LS1
32 to 64R
Solar panel
+
−
L1
VC1
Tune
IC1
ZN415E
C1
10n
C3
100n
C4
100n
C5
10n
R2
100R
C6
10µ
VR1
1k0
IC2
TDA7052
nc
nc – no connection
Dot
Notch
1
2
3
4
8
7
6
5
IC1 and IC2
Diodes, D1 and D2
©RSGB

There will either be a notch or a dot in the plastic body – this tells you which way round to mount the IC

Figure 1 The radio uses just two integrated circuits (chips) and operates at any voltage from 2 to 9 V

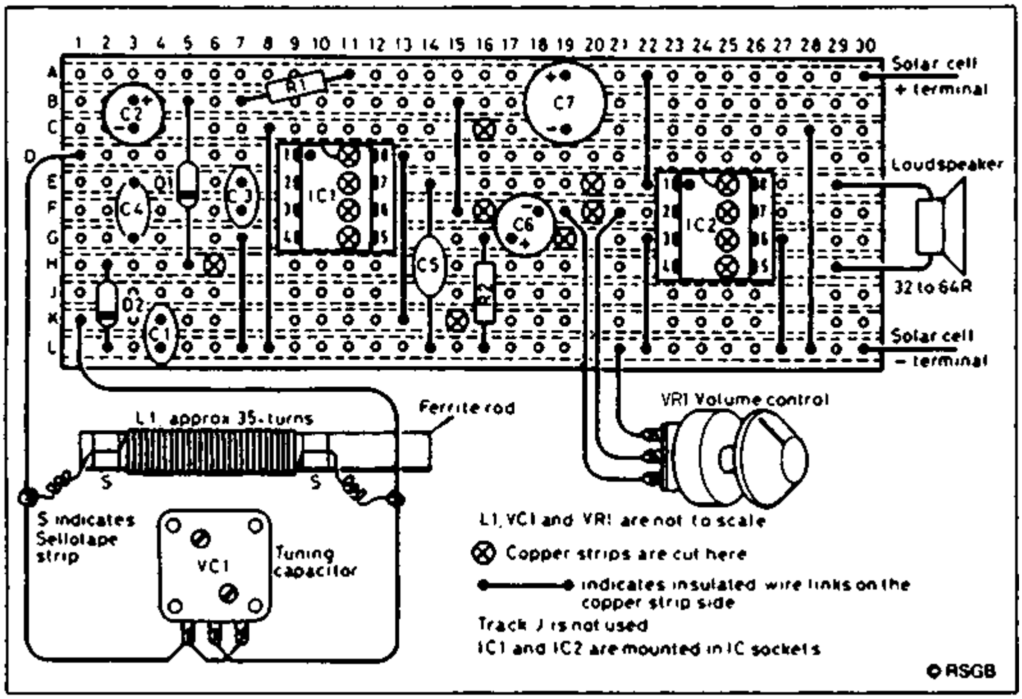
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
A
B
C
D
E
F
G
H
J
K
L
R1
C2
C7
D1
C4
C3
IC1
C5
C6
R2
IC2
D2
C1
Solar cell + terminal
Loudspeaker
32 to 64R
Solar cell – terminal
VR1 Volume control
Ferrite rod
L1 approx 35-turns
S
S
VC1
Tuning capacitor
©RSGB

S indicates Sellotape strip

L1, VC1 and VR1 are not to scale

⊗ Copper strips are cut here

●——● indicates insulated wire links on the copper strip side

Track J is not used

IC1 and IC2 are mounted in IC sockets

Figure 2 Follow the layout carefully, making sure that all connections are neatly soldered

your thumb and forefinger, break the strips at the points marked with a cross inside a circle. Hold the board up to the light to ensure that the tracks *have* been broken and that there is no copper swarf shorting tracks together.

Firstly, solder in the 8-pin DIL sockets for the ICs, making sure that the notches in the sockets are facing upwards, as shown on the diagram. Then, solder in the wire links, resistors, capacitors and diodes, making sure that the electrolytic capacitors (C2, C6 and C7) and the diodes (D1 and D2) are connected the correct way round. Use different colours of wire for the connections to the volume control, VR1, to avoid incorrect connections. Note the wiring of the tuning capacitor (VC1) shown in Figure 2; a two-section type is used, and both sections are wired in parallel to give twice the capacitance of a single section.

There is no on/off switch – just turn the volume down when you are finished using the set! The solar panel can be mounted parallel to the top of the case, or angled to receive the maximum energy from the sun, as shown in the photograph. You could have a battery available as a standby source to use the radio after dark; any battery of between 6 V and 9 V will do. Wire it with a simple changeover switch, so you can switch between solar and battery power. Ask a friend for help with this if you are not sure how to do it.

You may need to adjust the number of turns on L1 to get the best results, but it should be possible to receive at least five stations at good volume with your sun-powered radio!

SOLAR PANEL

Parts list

		Maplin code
Resistors: 0.25 watts, 5% tolerance		
R1	3.3 kilohms (kΩ)	M3K3
R2	100 ohms (Ω)	M100R
VR1	1 kilohm (kΩ) log	FW21X
Capacitors		
C1, C5	10 nanofarads (nF) ceramic	BX00A
C2	220 microfarads (μF) electrolytic (10 V)	FB60Q
C3, C4	100 nanofarads ceramic	YR75S
C6	10 microfarads (μF) electrolytic (10 V)	FB22Y
C7	1000 microfarads (μF) electrolytic (10 V)	FB81C
VC1	140–300 picofarads (pF)	FT78K
Semiconductors		
IC1	ZN415E (or ZN416E)	UR70M
IC2	TDA7052	UK79L
D1, D2	1N4148	QL80B
Solar panel	9 V at 50 mA	RK23A
Additional items		
LS1	32–64 ohm miniature loudspeaker	YT28F
	Ferrite rod	YG20W
	24 SWG enamelled copper wire	BL28F
	Plastic box, approx. 220 × 140 × 70 mm	YN39N
	Veroboard, cut to size	JP47B
	8-pin DIL sockets, 2 required	BL17T
	Knobs, 2 required	FD67X
	Material for speaker grille	
	Connecting wire	

38 A receiver for the 7 MHz amateur band

Introduction

Listening on the 40 metre band (from 7.0 to 7.1 MHz) can be very rewarding – it is a popular haunt for HF Special Event stations, and at night there are signals to be heard from all over Europe. This receiver is designed purely for the 40 m band, and is ideal for those who have built the simpler receivers and are looking for something a little more challenging. The more experienced constructor may prefer to build this on prototype board.

The circuit and its construction

Figure 1 shows the circuit diagram. The receiver will work well with headphones or loudspeaker. Walkman-type headphones and speakers are ideal for use here.

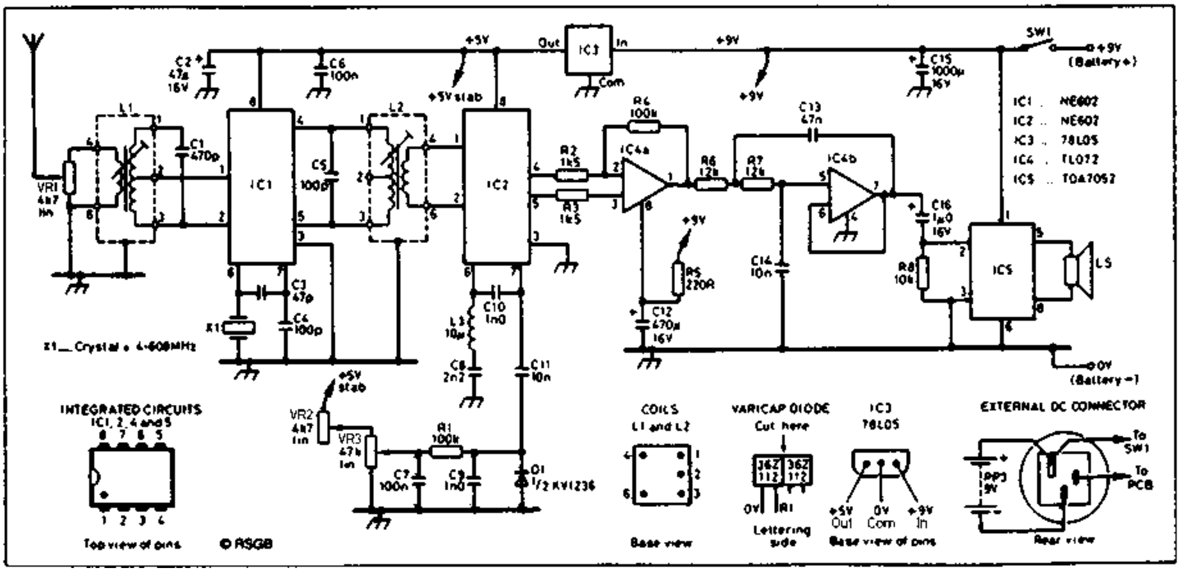
C2 47µ 16V
C6 100n
+5V
Out
IC3
In
Com
+5V stab
+9V
SW1
+9V (Battery+)
C15 1000µ 16V
L1
C1 470p
VR1 4k7 lin
IC1
C5 100p
L2
IC2
R2 1k5
R3 1k5
R4 100k
IC4a
R6 12k
R7 12k
C13 47n
IC4b
C16 1µ0 16V
IC5
LS
R8 10k
R5 220R
C12 470µ 16V
C14 10n
0V (Battery −)
IC1 .. NE602
IC2 .. NE602
IC3 .. 78L05
IC4 .. TL072
IC5 .. TDA7052
C3 47p
C4 100p
X1
X1 Crystal 4·608MHz
L3 10µ
C10 1n0
C8 2n2
C11 10n
+5V stab
VR2 4k7 lin
VR3 47k lin
R1 100k
C7 100n
C9 1n0
D1 1/2 KV1236
INTEGRATED CIRCUITS IC1, 2, 4 and 5
Top view of pins
© RSGB
COILS L1 and L2
Base view
VARICAP DIODE
Cut here
0V
R1
Lettering side
IC3 78L05
+5V Out
0V Com
+9V In
Base view of pins
EXTERNAL DC CONNECTOR
To SW1
To PCB
PP3 9V
Rear view

Figure 1 The receiver gives good performance on the 7 MHz amateur band as well as being simple to construct and align

Signals arriving at the aerial are coupled into IC1 via gain control VR1, which also functions as the on/off switch. Tuning is provided by varying the voltage on the *varactor diode* (or varicap), D1. VR3 is the main tuning control, and VR2 is the *bandspread* (fine tuning) control. The varactor diode is supplied as a dual device, which must be cut down the middle *carefully* with a sharp knife; with the lettering upwards, the ground lead (0 V) is on the left-hand side, as Figure 1 illustrates.

Solder in the IC sockets first, followed by the coils. After this come the links, resistors, capacitors and varactor diode. Ensure that IC3, the voltage regulator, is wired correctly, and check the polarity of the electrolytics. The crystal, X1, is very fragile, so take extra care with it. The wiring of the three controls is shown in **Figure 2**.

Before putting the ICs in their sockets, connect up the battery and check the following voltages with the negative voltmeter lead connected to the negative terminal of the battery:

Pin 8	IC1	5 V
Pin 8	IC2	5 V
Pin 8	IC4	9 V
Pin 1	IC5	9 V

When all these have been found to be correct, switch off and put the ICs carefully into their sockets. Use wire of different-coloured insulation to wire up the front-panel controls.

The case can be a small plastic box of size 22 cm by 15 cm by 8 cm, with three 10.5 mm holes drilled in the front and two 8 mm holes in the side for the aerial and earth connections. On one side are a 6 mm hole for the speaker socket and an 11 mm hole for the optional external power supply.

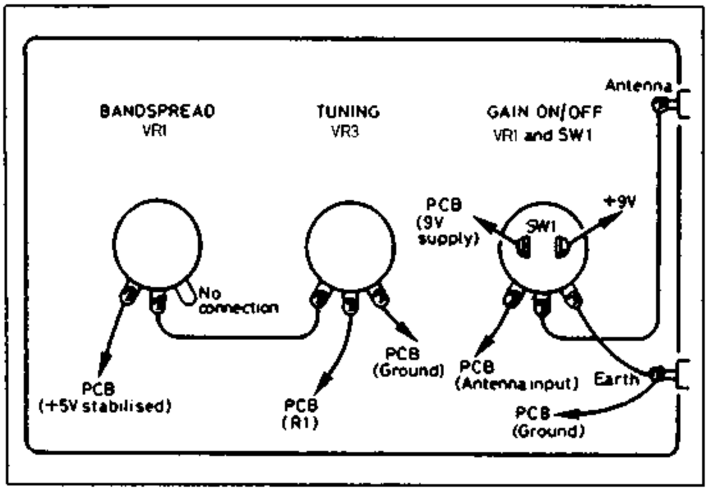

Figure 2 Rear view of the variable resistors. Check the connections carefully to make sure the wires are fitted correctly

Testing and tuning

The aerial for the receiver should be between 30 and 70 feet of wire, mounted as high as you can make it, away from trees and buildings if possible. Connect the battery and switch on. Adjust L1 and L2 for the best results. Tune slowly with VR3; you should find CW stations at the lower end of the band (anticlockwise) and SSB stations at the upper end (clockwise). You may find that it is easiest to make these adjustments *before* mounting the board in the case with double-sided sticky tape or pads. If you are planning to use an external DC supply, make sure it is a safety approved stabilised 9 V type, and **disconnect the battery before you use such a supply!**

If you suspect that the tuning doesn't quite cover the lower CW end of the band, try increasing C9 to 1200 pF. If it is the upper SSB end which is missing, decrease C9 to 820 pF.

It is always advisable to use an aerial tuning unit (ATU) between your aerial and the receiver. A suitable design of ATU is included as a project in this book.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R4	100 kilohms (kΩ)
R2, R3	1.5 kilohms (kΩ)
R5	220 ohms (Ω)
R6, R7	12 kilohms (kΩ)
R8	10 kilohms (kΩ)
VR1	4.7 kilohms (kΩ) linear, with SPST switch
VR2	4.7 kilohms (kΩ) linear
VR3	47 kilohms (kΩ) linear

Capacitors: all rated 16 V or more

C1	470 picofarads (pF) polystyrene 5%
C2	47 microfarads (μF) electrolytic
C3	47 picofarads (pF) polystyrene 5%
C4, C5	100 picofarads (pF) polystyrene 5%
C6, C7	100 nanofarads (nF) ceramic
C8	2.2 nanofarads (nF) polystyrene 5%
C9, C10	1 nanofarad (nF) polystyrene 5%
C11, C14	10 nanofarads (nF) ceramic
C12	470 microfarads (μF) electrolytic
C13	47 nanofarads (nF) ceramic
C15	1000 microfarads (μF) electrolytic
C16	1 microfarad (μF) electrolytic

Inductors

L1	Toko KANK3335R
L2	Toko KANK3333R
L3	10 μH 5%, e.g. Toko 283AS–100

Semiconductors

IC1, IC2	NE602 or NE602A
IC3	78L05 5 V, 100 mA
IC4	TL072
IC5	TDA7052

Additional items

D1	Toko KV1236 cut into two sections (one half used)
X1	4.608 MHz (available from Cirkit)
	3 × silver knobs, one with pointer
	Plastic case approx. 22 × 15 × 8 cm
	Speaker 8–32 Ω, or headphones
	4 × 8-pin DIL sockets for IC1, IC2, IC4, IC5
	2 × 4 mm sockets (red and black) for aerial and earth
	3.5 mm chassis-mounting jack socket for speaker
	DC power socket for external supply (if required)
	Prototype board

Kits

A complete kit is available from JAB Electronic Components.

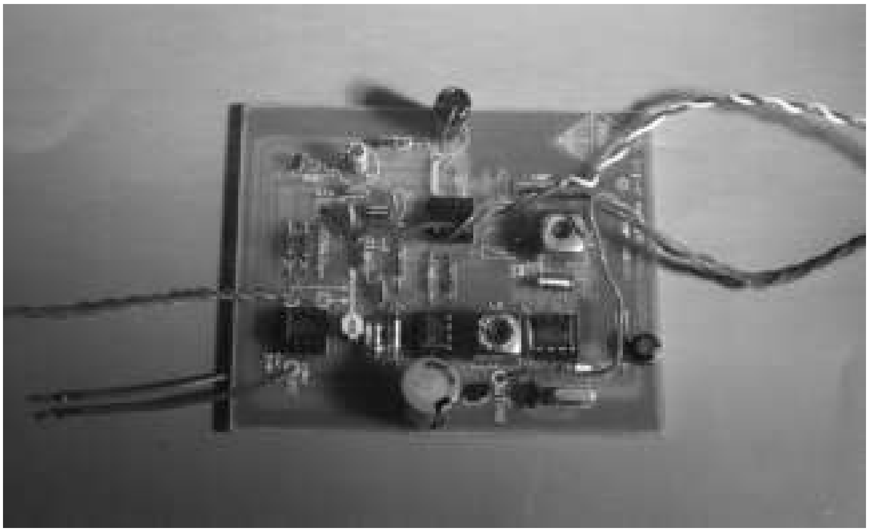

39 Diodes for protection

Introduction

Many semiconductor devices can be destroyed in an instant if their supply is reversed. With the use of batteries as power sources, it is quite common for the battery to be connected ‘the wrong way round’, and the scope for damaging equipment is significant and very real. Diodes can protect equipment in several ways, and you may do worse than to consider one of these approaches to protect your next expensive project.

Choose wisely!

All the diode circuits given here are so simple as to invite calamity. The circuits are not foolproof but, with a little care, will work first time. Remember that a semiconductor diode has a forward voltage drop of between 0.5 V and 0.7 V, depending on its type and the current flowing. This will be mentioned later.

The first thing you need to do is to insert a good multimeter *in series* with the circuit you want to protect; set the range to Amps DC, and switch on. Check that the circuit works properly. Then, decrease the current range on the meter until a good reading is indicated. Make a note of this current, as it is the normal running current of your circuit.

To choose a diode, you must consult the catalogues or data sheets and find one where the quoted *maximum forward current* exceeds the current you have measured; preferably it should be at least twice your measured current. Secondly, the diode will have a *peak inverse voltage* (PIV); this is the maximum voltage it can withstand when the cathode is made positive with respect to the anode, i.e when it is *reverse-biased* and not conducting. This voltage must be greater than your battery voltage, again by a factor of about 2. Except in the case of the bridge rectifier (see later), these criteria will enable the selection of a suitable diode to be made easily.

The series diode

The simplest and most obvious way to protect equipment is to insert a diode in the positive supply lead, as shown in **Figure 1**, with the diode passing current *only* when the supply is of the correct polarity. Because of the 0.6 V that exists across the diode, your equipment will normally operate on a

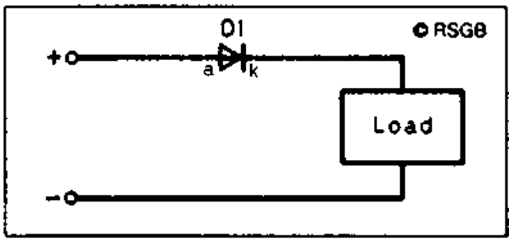

Figure 1 Series diode protection

slightly lower voltage. If you imagine reversing the supply to this circuit, you will see that the negative terminal of the battery is connected to the anode of the diode; the diode becomes reverse-biased and will not conduct any current, thus protecting the load. Your equipment will not operate when the battery connections are reversed.

The parallel diode

The circuit of **Figure 2** overcomes this voltage drop. It places the diode in parallel with the load (your equipment) but in a normally reverse-biased condition so that it draws no current when the battery is correctly connected. Reverse the battery connections, however, and a very large current will flow through the diode, thus blowing the fuse! For this technique to work successfully, the current drawn by the diode when the battery connections are reversed *must* be much greater than the maximum current drawn by the equipment, in order to blow the fuse. This is usually not a problem, however.

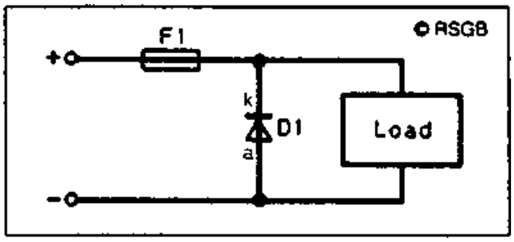

Figure 2 Parallel diode and fuse

The diode bridge

For sheer elegance, the circuit of **Figure 3** takes the biscuit! It uses four diodes connected in the form known as a *bridge rectifier*. Such rectifiers exist, and do not have to be made up from four discrete diodes.

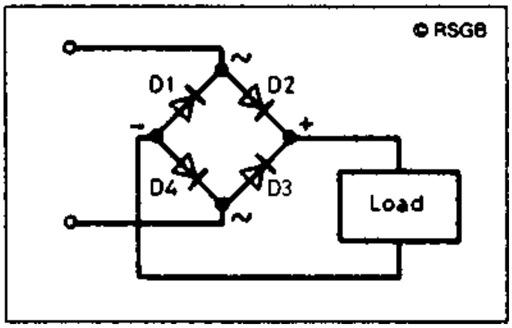

Figure 3 The diode bridge

Follow the current round the circuit from the supply, assuming initially that the top wire is positive and the bottom wire negative. It flows from the positive supply

(a) through D2
(b) through the load (top to bottom)
(c) through D4

and back to the negative of the supply.

Now assume that the bottom supply lead is positive and the top lead negative. The current flows from the positive supply

(a) through D3
(b) through the load (top to bottom)
(c) through D1

and back to the negative of the supply.

So, whichever way round the battery is connected, the current will *always* flow the same way through your equipment!

The circuit does have a drawback, however. Whichever way round the battery is connected, there are always **two** diodes conducting the current at any time. In the first case it is D2 and D4; in the second case, D1 and D3. This means that there is a total voltage drop of about 1.2 V. If your equipment can tolerate that reduction in voltage, then you will not have a problem.

Decoupling

Whenever the supply rail to a piece of equipment, or even to an individual stage of a circuit, is broken for the insertion of a device that will drop voltage, strange things can happen. This is because the supply for any circuit is assumed to have a low resistance to DC and a low *impedance* to AC. (Impedance is the AC equivalent of DC resistance.) These two are not the same, and the insertion of a diode or diodes is certain to make a big difference to them both. To overcome any irregularities in the operation of the circuit you are protecting in any of the ways described previously, the protected supply needs to be *decoupled* with a capacitor.

If the circuit handles audio frequencies only, placing an electrolytic capacitor directly across the load in all three circuits should solve the problem. The parallel circuit of Figure 2 is less at risk than the other circuits. The size of the capacitor will be determined by the current taken by the circuit, and may need to be chosen within the range 100 μF to 10 000 μF, with a working voltage greater than the supply voltage.

If the circuit is mainly handling RF currents, placing a capacitor of 0.01μF across the load should prevent any problems. A second capacitor, also across the load, of between 10 μF and 100 μF may be needed. Again, the parallel circuit of Figure 2 is less at risk than the other two.

Don’t be afraid to experiment, but confine your experimenting (at first) to small equipment and low currents, until you get a ‘feel’ for the technique.

40 An RF signal probe

Introduction

A radio-frequency (RF) diode probe is a simple device which, when used with a conventional multimeter, enables the measurement of RF voltages in a circuit. When constructed, this will be one of the most useful pieces of test equipment for the experimenter who revels in the construction of transmitters and receivers.

The circuit

Figure 1 shows the simple circuit diagram. It is *almost* the same as a common diode rectifier circuit, but with a simple change to make it more sensitive. The circuit is known as a *voltage-doubler*, and is often found in high-voltage supplies, with beefier capacitors and diodes, of course! Because we are dealing with high frequencies and smaller voltages, the diodes and capacitors can be physically very small. The diode circuit of D1 and D2 *rectifies* or *detects* the RF from the probe, and any remaining AC is removed (short-circuited to ground) by C2. This produces, at the output, a constant voltage proportional to the peak-to-peak RF voltage present at the input; the output voltage is fed to an ordinary multimeter (on a voltage range).

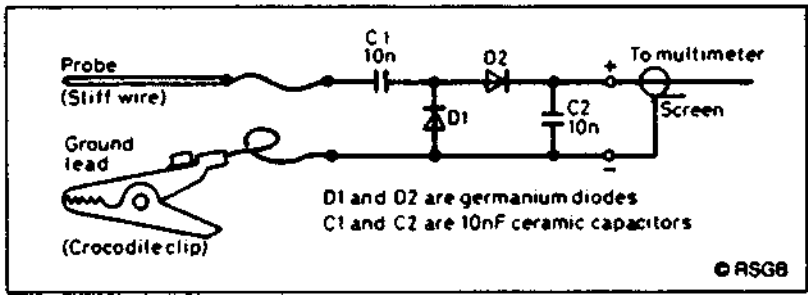
Probe
(Stiff wire)
C1
10n
D2
+
To multimeter
D1
C2
10n
Screen
Ground
lead
(Crocodile clip)
–
D1 and D2 are germanium diodes
C1 and C2 are 10nF ceramic capacitors
© RSGB

Figure 1 The diode probe can be made from four electronic components

Construction

Although the circuit layout is not critical, a description of the prototype is given here for your information. **Figure 2** has the details. The components are soldered to square pieces of copper-clad printed-circuit board (PCB) glued to a larger piece of PCB. The larger piece serves as a ground connection. Use a stiff copper wire as the probe, and an insulated flexible wire with a crocodile clip to connect the probe to the ground of the circuit under test.

Cut a piece of plain PCB, 30 mm by 45 mm and another of 15 mm by 45 mm. Cut the smaller piece into three measuring about 15 mm square. Stick the three small pieces to the larger piece, ensuring that there are small gaps between each, as Figure 2 shows. Solder the components in place. The connection to the multimeter should be thin coaxial cable or screened microphone cable. If you use unscreened cable, there may be RF pickup here which can lead to false readings.

Simple to use

Using the probe is simple. Connect it to the multimeter and set the meter to around 10 V DC – you may need to reduce this, depending upon the magnitude of the RF voltage you are trying to measure. Hold the probe by

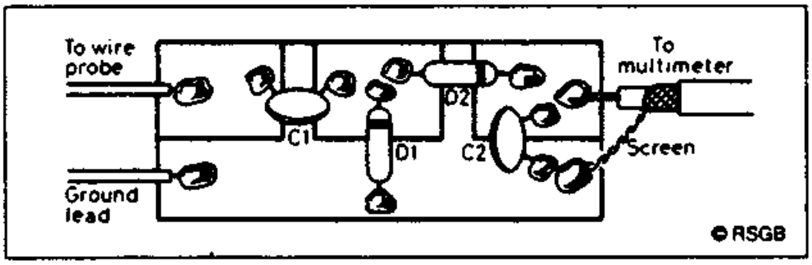
To wire
probe
Ground
lead
C1
D1
D2
C2
To
multimeter
Screen
© RSGB

Figure 2 The components are mounted on a copper-clad PCB

its base, being careful *not* to touch any of the components. Connect the croc clip to a ground point on the circuit and touch the probe on the circuit point to be tested. If an RF signal is present, there should be a reading on the multimeter. That's all there is to it!

Parts list

Capacitors

C1, C2	10 nanofarads (nF), or 0.01 microfarad (μF) ceramic

Semiconductors

D1, D2	Any *germanium* signal diode, e.g. OA91 or AA119

Additional items
- Stiff copper wire
- Insulated flexible wire
- Copper-clad PCB
- Crocodile clip
- Coaxial or screened cable

Source

All components are available from Maplin.

41 An RF changeover circuit

Introduction

The simplest entry route to operating an HF station is to buy or build a receiver, and to add a simple CW (Morse) transmitter. In theory, this sounds so simple, yet in practice there is one significant hurdle to be overcome. How can they both share the same aerial? A manually operated changeover switch is the obvious solution, but it takes time to perform the switching operation, by which time someone else has squeezed in before you and contacted the DX station! This small circuit accomplishes the changeover automatically, as soon as the transmitter is keyed.

How it’s done

The circuit is shown in **Figure 1**, and uses the signal from the transmitter to operate a relay. A relay is a switch which is operated electrically in the following way. A coil of wire with an iron core is used as an electromagnet. When current flows through the coil, it produces a magnetic field which, in turn, is used to pull a set of switch contacts. These contacts will be used to switch the aerial from transmitter to receiver and vice versa. The relay used here has a double-pole changeover (or DPDT, or DPCO) switch. If you are confused by the different types of switch, look at the basic descriptions elsewhere in this book. The *pole* of a switch is the part that doesn’t move; in Figure 1, one pole is connected to L1 and the other to the aerial output. In a circuit diagram, the switch contacts are always shown in their *normal* state, i.e. when *no* current flows in the relay coil. The changeover switch is normally in the receive position, so we say the receive switch is *normally closed*, and the transmit switch is *normally open*. We choose to have the circuit in the receive position normally, because everyone spends a lot more time receiving than transmitting. The relay is energised only when you are transmitting.

There are three RF sockets – one for the transmitter, one for the aerial, and one for the receiver. When the transmitter is not in use, there is a direct connection (via the Rx normally closed contacts of the relay) between the aerial and the receiver. The transmitter (which is not keyed at this time) is connected to a 50 ohm dummy load, R1.

When the transmitter is activated by pressing the Morse key, an RF signal appears at the transmitter socket. The wire carrying the signal to the relay

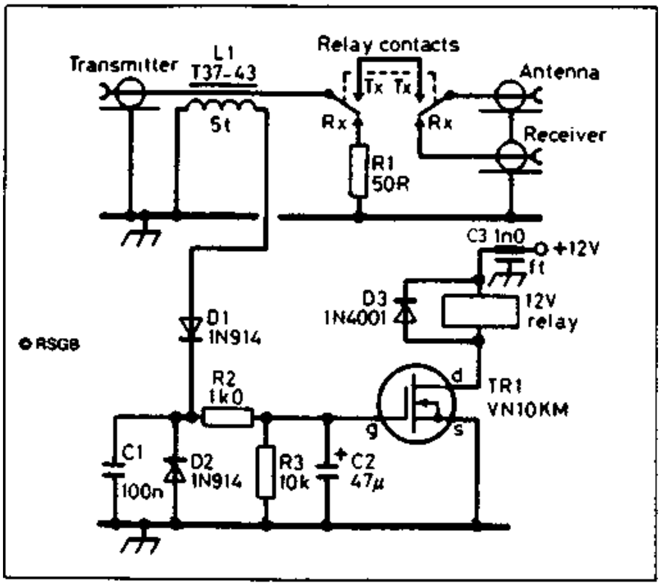

Figure 1 RF activated relay

Tx contacts passes through a toroid (a ferrite coil former, looking like a black Polo mint), which has a five-turn coil wrapped around it. This five-turn coil picks up the signal, which is then detected by D1 and D2, and converted to a steady voltage. C1 removes any remaining RF and C2 provides *hang*. Without it, the circuit would detect the gaps between every dot and dash being sent, and switch the aerial over very rapidly and very frequently! This is *not* what we want. We need the relay to remain in the transmit condition at normal Morse keying speeds, and then return to receive when the sending is complete. The combination of C2 and R3 achieves this.

The voltage appearing across C2 is fed to a VMOS field-effect transistor (FET), TR1, which acts as an electronic switch. When a voltage appears across C2, the FET switches on and passes a current through the relay coil, changing the contacts from receive to transmit. The diode, D3, across the relay winding protects the FET from being damaged by the large reverse-voltage spike which occurs across the coil when the FET switches off.

Construction

This is very simple – a circuit board is not required. The prototype was built into an old 50 g tobacco tin. All ground leads are soldered directly to the tin, thus supporting the components automatically. Phono sockets were used for the aerial, receiver and transmitter; you could use whatever connectors matched the rest of your station. The 12 V supply is fed through the tin using a 1000 pF feed-through capacitor, C3.

The ‘dummy load’, R1, must be able to withstand the RF output power of the transmitter. For novice use, 4 watts is adequate. Two suggestions for making up R2 from standard resistors are shown in **Figure 2**. Six 330 ohm, 1 watt resistors in parallel have a combined resistance of 330/6 = 55 Ω at 6 W. Two 100 ohm, 2 watt resistors in parallel have a combined resistance of 100/2 = 50 Ω at 4 W.

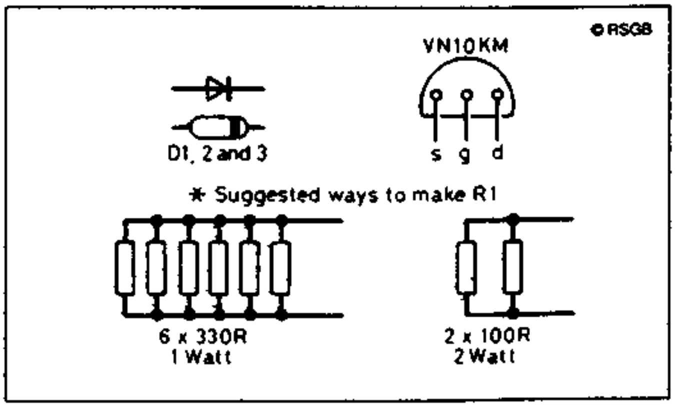

Figure 2 Component details

Make sure all the contacts to the relay are correct; also ensure that TR1, C2, D1, D2 and D3 are correctly wired.

Winding the toroid, L1, is very simple. It is made from PVC-insulated hook-up wire. Each time the wire passes through the centre of the toroid counts as one turn. The wire from the transmit socket to the relay simply passes through the centre! Use thin 50 Ω coax for the leads to the three sockets.

What happens to the receiver?

On transmit, the receiver is not connected to the aerial, but is only the separation of the switch contacts (about 0.5 mm!) away from the transmit lead. The receiver *will* pick up the transmitted signal, and there is a possibility that this signal will be enough to damage the receiver's sensitive input circuits. This can be prevented with the simple addition shown in **Figure 3**. Two diodes are connected back to back across the receiver socket. These act as a *limiter*, reducing the amount of signal that can enter the receiver. Solder these directly between the receiver socket and the tin.

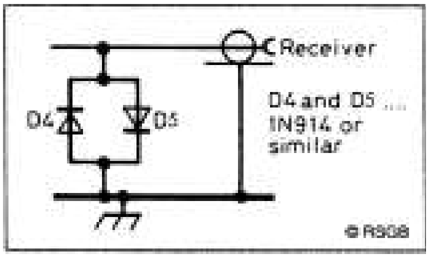

Figure 3 Receiver input protection

Warning

The wiring **must** be thoroughly checked. If the relay switch connections are wrong, there is a **great probability** that the full RF power will be applied to your receiver! You can check these connections as follows. With the 12 V supply connected, check that the relay energises (clicking sound heard) when the FET drain (d) is shorted to earth. Check the aerial and receiver connections with an ohmmeter or continuity tester. Disconnect the 12 V supply, and check that there is continuity between the centre pins of the aerial and receiver sockets. Reconnect the 12 V supply, short the FET drain to earth again (this will not damage the FET) and check that there is continuity between the centre pins of the aerial and transmitter sockets. If these tests show correct operation, disconnect the drain shorting wire (*most* important!) and your aerial auto-changeover is ready for use! Avoid using the device without the lid fitted.

Parts list

Resistors: 0.25 watt, 5% tolerance, unless otherwise stated

R1	50 ohms (Ω) – see text and Figure 2
R2	1 kilohm (kΩ)
R3	10 kilohms (kΩ)
Capacitors	
C1	100 nanofarads (nF), 0.1 microfarad (μF)
C2	47 microfarads (μF) 15 V electrolytic
C3	1000 picofarads (pF) feedthrough
Semiconductors	
D1, D2, D4, D5	1N914
D3	1N4001
TR1	VN10KM
Additional items	
L1	T37–43 toroid
	Relay 12 V DPDT (DPCO) relay
	Three sockets to suit (phono, SO239, etc.)
	Thin coax cable and hook-up wire

42 A low-light indicator

Introduction

This is a simple one-evening project that can be built for the pure fun of it, or to use as the basis of a more complex project to switch your shack lights on when it gets dark! In its prototype form, it simply flashes an LED when the ambient light level drops to a preset point.

Operation

The heart of the circuit shown in **Figure 1** is a *photo-conductive cell*, also called a *light-dependent resistor* (LDR), a device whose resistance changes according to the amount of light falling on it. In bright light, the resistance is low (about 1 kΩ), whereas in the dark, its resistance is very high (up to 10 MΩ). The cell is made from a semiconducting material known as cadmium sulphide (CdS), and is enclosed in a small plastic container. The semiconductor is laid on a flat insulating surface in the form of a small flat ribbon. The ribbon construction gives a good area of surface for a given

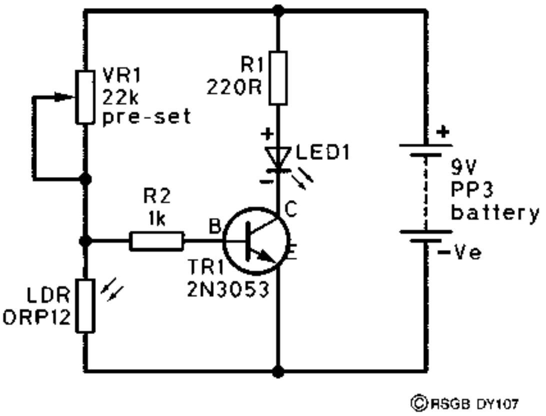
VR1
22k
pre-set
R1
220R
LED1
9V
PP3
battery
R2
1k
B
C
E
TR1
2N3053
LDR
ORP12
-Ve
©RSGB DY107

Figure 1 Circuit diagram

length of ribbon, and the length of the ribbon is maximised by laying it out in a zig-zag pattern, as can be seen diagrammatically in **Figure 2**. In the dark, CdS is an insulator; when light falls on it, electrons are released inside the CdS, making it conduct. The more light there is, the more electrons there are, and the resistance falls.

In this circuit, the LDR is connected across the 9 V supply in series with a variable resistor, VR1. In this arrangement, the voltage that exists across the LDR will be determined by the light level. As the light intensity increases, the resistance of the LDR falls, dropping a smaller voltage across it. The

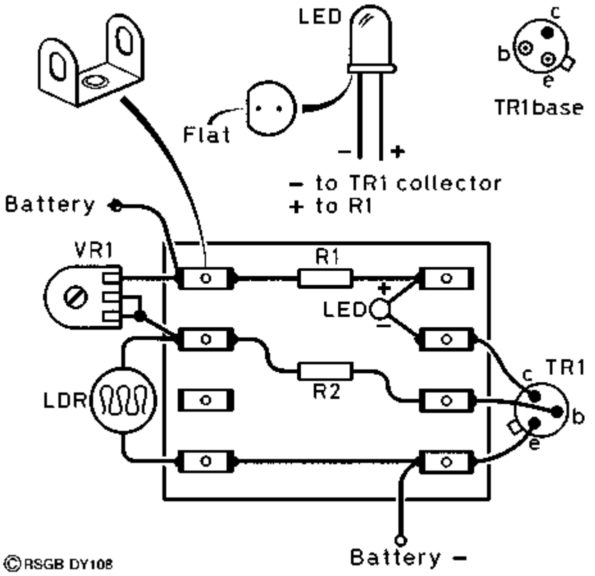
LED
Flat
– +
– to TR1 collector
+ to R1
c
b
e
TR1base
Battery +
VR1
R1
LED
+
–
LDR
R2
TR1
c
b
e
Battery –
©RSGB DY108

Figure 2 Component layout on tagstrip

reverse happens when the light intensity falls – the voltage across the LDR increases. This voltage is used to drive an npn transistor, TR1, connected as an electronic switch. As the voltage on the base (B) rises, it will reach a point where TR1 will suddenly start to pass current, just as an ordinary switch does when pressed. The current flowing through TR1 also flows through a flashing light-emitting diode (LED), D1 and its series resistor, R1. D1 can be a steadily glowing type, if preferred. If you want to make the circuit switch the LED on at a different ambient light level, adjust VR1.

Construction

The prototype was made on an eight-tag tagboard (Figure 2). The resistors R1 and R2 can be laid on the tagboard for soldering, the rest of the components lying above or to the side of the board. Check the circuit after you have soldered everything on, then connect the battery. The LED should flash if you put your hand over the LDR, and VR1 can be adjusted to vary the point at which the LED lights. The project can be housed in a small plastic box with holes provided for the LDR and LED and an on/off switch if you want one.

Parts list

Resistors: all 0.25 ohm, 5% tolerance

R1	220 ohms (Ω)
R2	1 kilohm (kΩ)
VR1	22 kilohms (kΩ) linear pre-set

Semiconductors

LDR	ORP12
TR1	2N3053 npn
LED1	Flashing LED

Additional items
- Plastic box
- Tagboard with 8 tags
- PP3 battery connector

43 A J-pole aerial for 50 MHz

Introduction

For FM communication (i.e. voice and data) on the VHF bands, a vertical aerial is used to give all-round (non-directional) coverage. This is a half-wave aerial which can be fed at the end, thus removing the principal problem with the conventional vertical centre-fed half-wave dipole, which is that the feeder should leave the dipole at right angles. This is no problem when the dipole is horizontal, but can be difficult for the vertical dipole.

Basic facts

A feeder must be connected to an aerial at a point where the *impedance* (AC 'resistance') of the aerial closely matches that of the feeder. The difference between the two impedances gives rise to the voltage standing-wave ratio (VSWR), which is unity only when the two impedances are the same. With 50 Ω feeders, the feed point of a half-wave aerial is at the centre, where the aerial impedance is around the same value. At the end of a half-wave aerial, the impedance is high, so it is *not* a suitable point to connect a 50 Ω feeder.

Connection at this point can be effected using an RF transformer. RF transformers act in the same way as ordinary transformers, except that they are much smaller, and usually comprise wires of particular lengths adjacent to each other. **Figure 1** is a good starting point. It shows the aerial in its diagrammatic form. Notice that the aerial is in the form of an elongated letter 'J'; this shape gives rise to its nickname – the *J-pole*. The quarter-wave RF transformer is the lower 'U' section below the half-wave element. At the bottom of the U section, the impedance is zero (this may become clearer later) and at the top of the U section it is high, thus matching the aerial impedance. The coaxial feeder cable is connected part-way up the U section, where the impedance is around 50 Ω.

The practicalities

Figure 2 shows the aerial as constructed. As it is about 4.5 metres high, it may be too high for the average house loft, but is ideal for mounting outside, supported by a non-metallic pole or hung from a tree branch. The upper half-wave section is made from 1.5 mm insulated copper wire, as used in domestic mains wiring. The quarter-wave transformer below the half-wave section is made from 300 Ω balanced line ('ribbon cable'). The

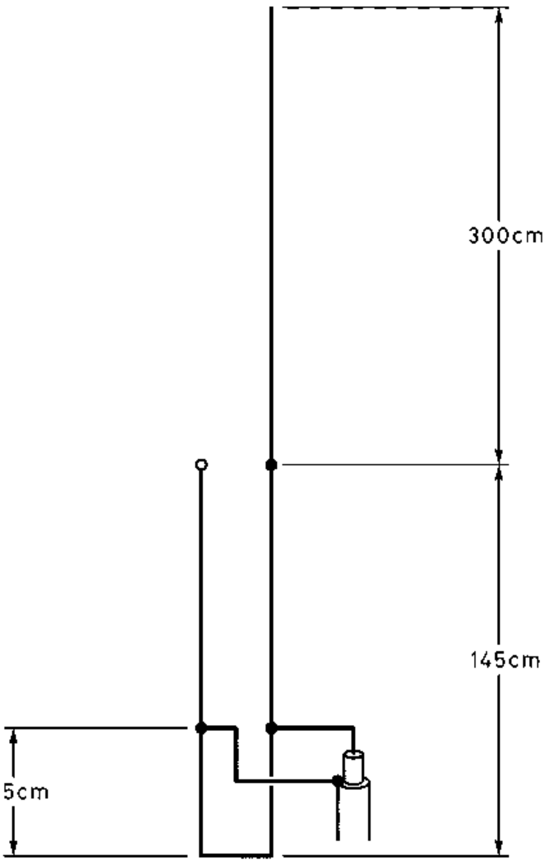

Figure 1 Overall dimensions of the 6 metre J-antenna

wires at the bottom of the transformer section are stripped of their insulation, twisted and soldered. At the upper end, only one wire of the balanced line is soldered to the bottom of the half-wave section. The other wire of the pair is not connected and is left insulated.

At the feed point of the transformer, the insulation needs to be carefully stripped from the balanced line. You will need a standing-wave meter (VSWR meter) in the coaxial line between your transmitter and the aerial, and you will need to adjust the position of the feed point. 45 cm from the bottom was the best point on the prototype, but this position is dependent upon the immediate surroundings of the aerial, and *must* be done when the aerial is in its final operating position. **Warning: *Never* make adjustments to the feed point when the transmitter is on. Make a VSWR measurement, switch off, move the feed point, switch on again, make another measurement, and so on**. You will need to aim for the lowest VSWR you can – certainly better than 2:1. Having found the best position, wrap all the exposed wires with self-amalgamating tape, to seal them against the ingress of moisture.

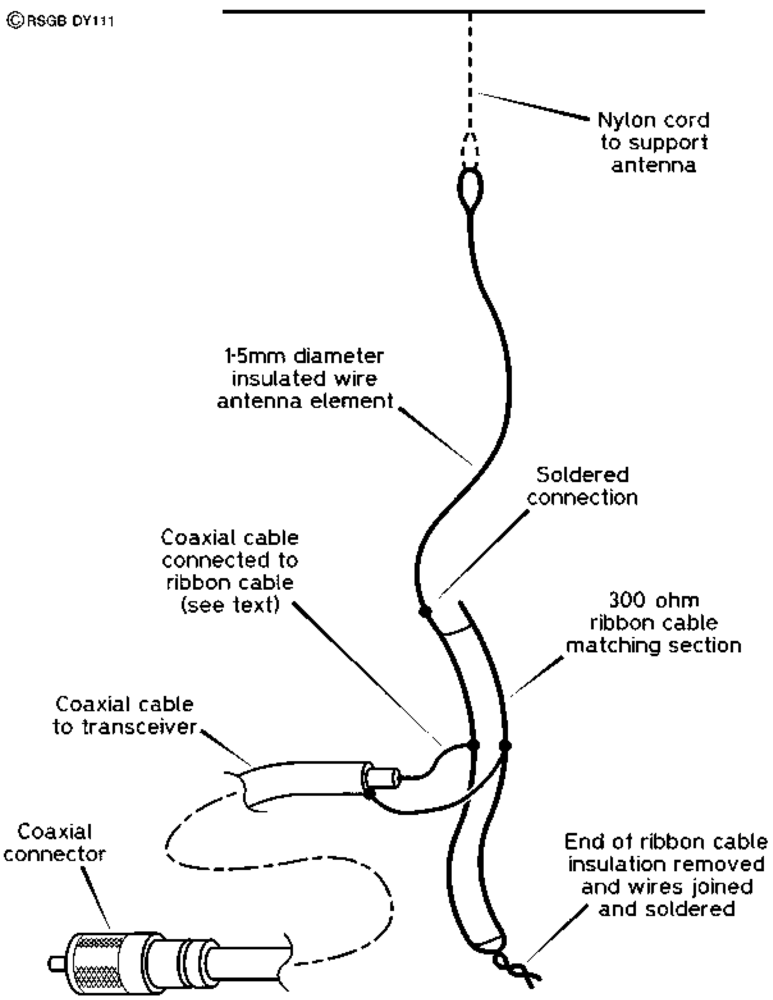
©RSGB DY111
Nylon cord to support antenna
1·5mm diameter insulated wire antenna element
Soldered connection
Coaxial cable connected to ribbon cable (see text)
300 ohm ribbon cable matching section
Coaxial cable to transceiver
Coaxial connector
End of ribbon cable insulation removed and wires joined and soldered

Figure 2 Construction details

How it performs

Figure 3 shows a computer prediction of how the J-pole radiates. It is called a *polar diagram*, and shows the distribution of your transmitted power when viewed 'from the end of your garden'. Most of your signal is sent at a fairly small angle to the horizontal; very little signal goes upwards, which is a good thing, of course. This also shows why the J-pole (or any other vertical aerial) should not be called 'omnidirectional', which means it radiates in all directions. It is omnidirectional *only* in the horizontal plane.

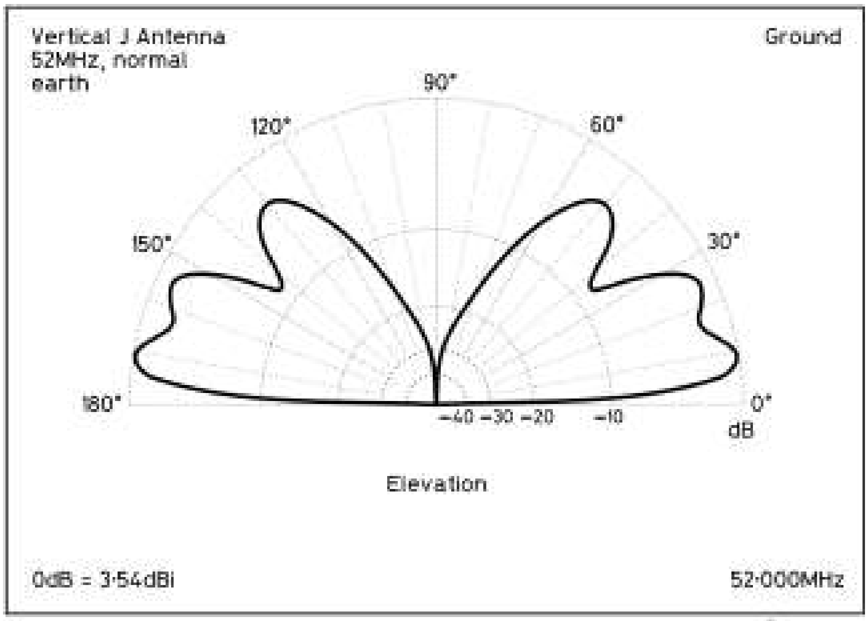

©RSGB X210

Figure 3 This polar diagram computed for the J-pole

Safety

Where you mount your aerial is a matter of personal preference and the restrictions of height and space, but the following safety rules *must* be applied.

1. Never fix an aerial where it may come into contact with power lines or telephone lines.
2. When climbing a ladder to put up an aerial outside, make sure the ladder is safe and that it is secured.
3. Don't do this alone. Preferably have someone with you. If this is not possible, make sure someone knows where you are.

Parts list

3.00 metres	1.5 mm insulated copper wire
1.50 metres	300 ohm balanced line
As required	50 ohm coaxial cable
As required	Self-amalgamating tape

44 Measuring light intensity – the photometer

Introduction

Before the days of automatic ‘point-and-shoot’ cameras, a photographer would use a light meter or *photometer* to measure the light level, then manually convert this reading into shutter speed and lens aperture settings to ensure a correctly exposed negative. Modern cameras have quite sophisticated photometers, which control the shutter speed and iris settings automatically.

A short explanation

A simple photometer circuit is shown in **Figure 1**, and is based on a device called a *light-dependent resistor*, or LDR. As its name tells us, its resistance depends upon the amount of light falling on it. In bright light, the resistance is relatively low (about 1 kΩ), whereas in the dark, its resistance is very high (up to 10 MΩ). The cell is made from a semiconducting substance known as cadmium sulphide (CdS), and is enclosed in a small plastic container. The semiconductor is laid on a flat insulating surface in the form of a small flat ribbon. The ribbon construction gives a good area of surface for a given length of ribbon, and the length of the ribbon is maximised by laying it out in a zig-zag pattern. In the dark, CdS is an insulator; when light falls on it, electrons are released inside the CdS, making it conduct. The more light there is, the more electrons there are, and the resistance falls.

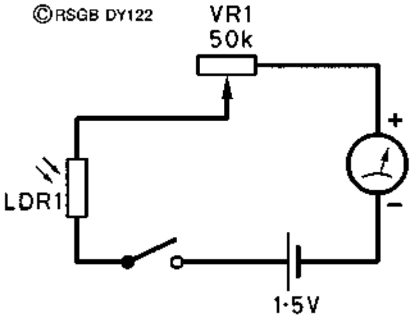

Figure 1 Circuit diagram of photometer

The photometer circuit

The circuit is simply a series connection of four things – the battery, the LDR, a variable resistor and a meter. A switch is also provided. The combination of the LDR resistance and that of VR1 determines the current flowing, which is indicated on the meter. Altering the resistance of VR1 sets the *sensitivity* of the photometer – you may want full-scale deflection of the meter needle for a bright light, or for a dim light.

It is simple to make, and a plug-in type matrix board is ideal to test the circuit, so that you can decide if you want to make a permanent version. Connect all the components in series; the only change you may want to make is the connection to VR1. If you find that the sensitivity control seems to work 'backwards', simply unsolder the wire from the end tag of VR1 and solder it to the opposite end tag. Problem solved!

In use

As soon as you connect up the battery, you will probably have a meter reading because of the daylight falling on the LDR. Shading it with your hand should reduce the reading. If the meter needle is hard over against the end-stop, turn VR1 until it indicates about half-scale. The LDR is very sensitive, and will read zero only in a dark room. If you put on a torch to see what the meter reading is, the LDR will detect the torch light, and will give a reading!

Here is a simple project where you can use the photometer in an experiment which has an analogy in radio. Draw a circle on a large (A3 or bigger) sheet of paper and divide it up into 30-degree sectors, as shown in **Figure 2**. Draw a series of smaller circles which divide the maximum radius into five. Look at the figure if you're not sure about this. Bring the LDR away from the circuit by using two long, flexible wires. The experiment must be performed in a darkened room (preferably in total darkness). Prepare a table with two columns, the left-hand one headed 'Angle (degrees)' and the right-hand one 'Meter reading'. Fill in the left-hand column 0, 30, 60 . . . and so on up to 360°.

Place the torch in the position shown, with its lens at the centre of the circle and pointing along the 0° line; switch it on. Place the LDR facing the torch and adjust VR1 until you have full-scale deflection of the meter needle. Suppose the meter indicates 10 units at this point. Enter this into your table in the 0° row. Keeping the torch the same distance from the circle centre, and pointing at it, move the LDR round all 30° positions and record the meter readings. Switch off the torch and take the sheet of paper and your tabulated results into daylight!

Lay the large sheet of paper on a table with your results beside it; then, at each 30° interval, plot the point along the radius corresponding to the meter

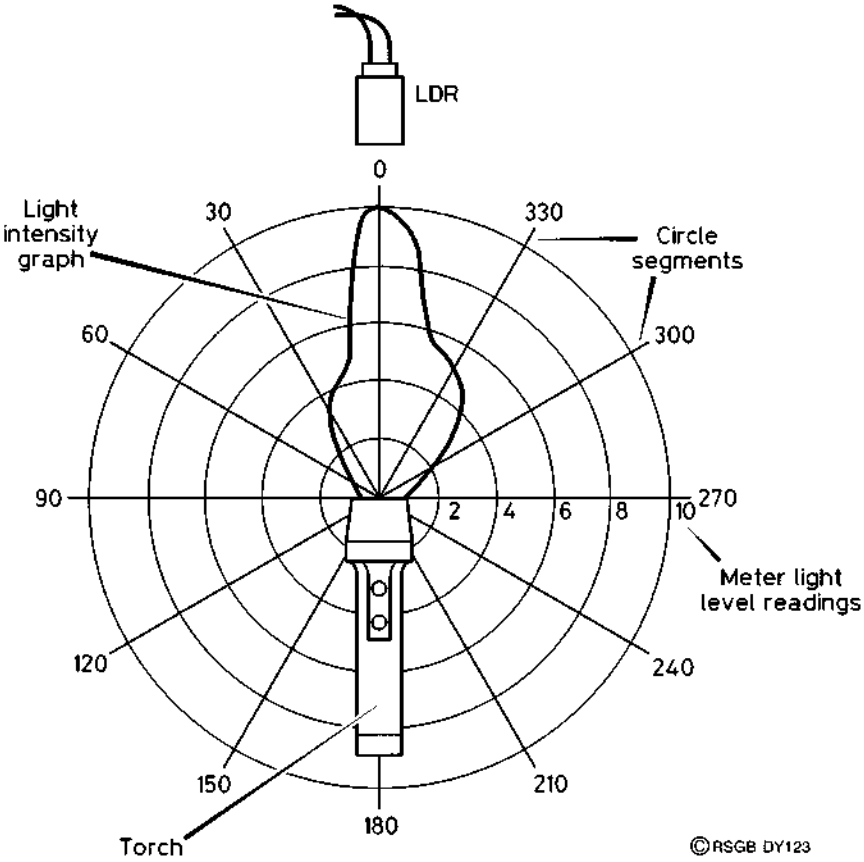

Figure 2 Torch light intensity pattern plotted at RSGB HQ

reading. Then, join up all the points and you have what is called a *polar diagram* of light intensity. The use of the word 'polar' implies that the readings have been taken in a circle and plotted that way.

Light waves and radio waves are both examples of *electromagnetic radiation*. The torch is designed to 'beam' its radiation in a particular direction, just like an aerial does. Hence the use of the word 'beam' for a directional aerial. If a similar polar diagram is drawn for a Yagi-type aerial, it will show the same general characteristics as does Figure 2, namely a main direction (or 'lobe') where most of the energy is concentrated, with evidence of *sidelobes*, indicated by 'lumps' on the otherwise smooth main lobe.

Also in this book you will find a project for the construction of a UHF field-strength meter, which you could use to carry out a measurement of the polar diagram of a UHF aerial.

Parts list

Resistor

VR1	50 kilohms (kΩ) linear

Semiconductor

LDR1	ORP12

Additional items
- Meter, 50 or 100 μA
- Battery clip for single AA cell
- AA cell

45 A 70 cm Quad loop aerial

Introduction

This is a description of how to make an aerial for the 70 cm band which has gain compared with the 'rubber duck' or the dipole aerial. It can easily be dismantled and reassembled, making it ideal for contest use.

The principles

Most aerials which comprise several similar *elements* arranged along a *boom* are variations of the design originated by Yagi and Uda, and which takes the name (for historical reasons) of the former, and is know as the Yagi array. Let us suppose we have a Yagi aerial beaming left to right in front of us. The elements get progressively shorter from the left (look at almost any TV aerial to see this). All the elements on a Yagi aerial are classified as follows:

- The *reflector* – the leftmost element as we look at the array. It is the longest. Next to it is:
- The *driven element* – this is the element connected to the feeder, which in turn runs down to the transceiver. It is slightly shorter than the reflector.
- All the elements beyond the driven element are called *parasitic elements*, or *directors*. They are shorter than the driven element and usually get progressively shorter as we progress to the right along the boom. The directors are mainly responsible for the *directivity* (or *beamwidth*) of the array.

The progression from a simple dipole (a driven element in isolation) to a Yagi array is simple, but is nevertheless important. To make an aerial of two elements, a reflector (*not* a director) is added to the driven element. For three or more elements, directors are added to the two-element design. Adding more and more directors soon becomes impractical, the reduction in beamwidth (such as it is) does not warrant the extra expense, weight and wind resistance that is incurred.

Instead of using linear (straight) elements, as in the generic Yagi, this design uses loops. Designs using squares of wire instead of loops are known as *Quad* aerials, and HF designs require large X-shape frames to support the large squares of wire. At 70 cm, however, the use of wire loops is easier, and they are self-supporting.

Construction

This is quite simple. Any type of material (metal, plastic, wood) can be used for the *boom* (the support for the elements) and for the mast. The elements are made from 14 SWG enamelled copper wire. 16 SWG hard-drawn aerial wire, which is not enamelled, has also been used with success. Thinner wire might result in a rather 'floppy' aerial! The separate parts of the aerial are held together with *jubilee clips* (hose clips).

The driven element is secured to the boom with a jubilee clip and a three-connector plastic connector block as shown in detail in **Figure 1**, and in the photograph. Cut the wire for the driven element 70 mm longer than the 700 mm indicated in Figure 1. Then, using sandpaper, remove the enamel from one end to a distance of 20 mm, and from the other to a distance of 50 mm. After forming the loop of the driven element, bend both stripped ends through 90°, and insert them into the first to holes of the plastic connector block (Figure 1). Do not tighten the screws yet. Push both ends into the block as far as they will go, then bend the 50 mm end back on itself and pull the ends back through the connector block so that the end you have just bent goes into the third hole in the block. Now tighten the screws in the block and in the jubilee clip.

Each director and the reflector should be made 40 mm longer than the circumferences shown in Figure 1. Strip the enamel, as before, from the last 20 mm at each end. Form the wire into the loop, slip the ends under the jubilee clip (Figure 1) and tighten it. You may find that it helps to solder the stripped ends together before securing the jubilee clip.

The boom is fixed to the mast using jubilee clips and wire, as shown in Figure 1. Solder the feeder cable to the driven element, with the braid soldered to the end of the driven element which is connected to the boom (this applies to metal booms; with plastic or wooden designs, the feeder connections are not critical).

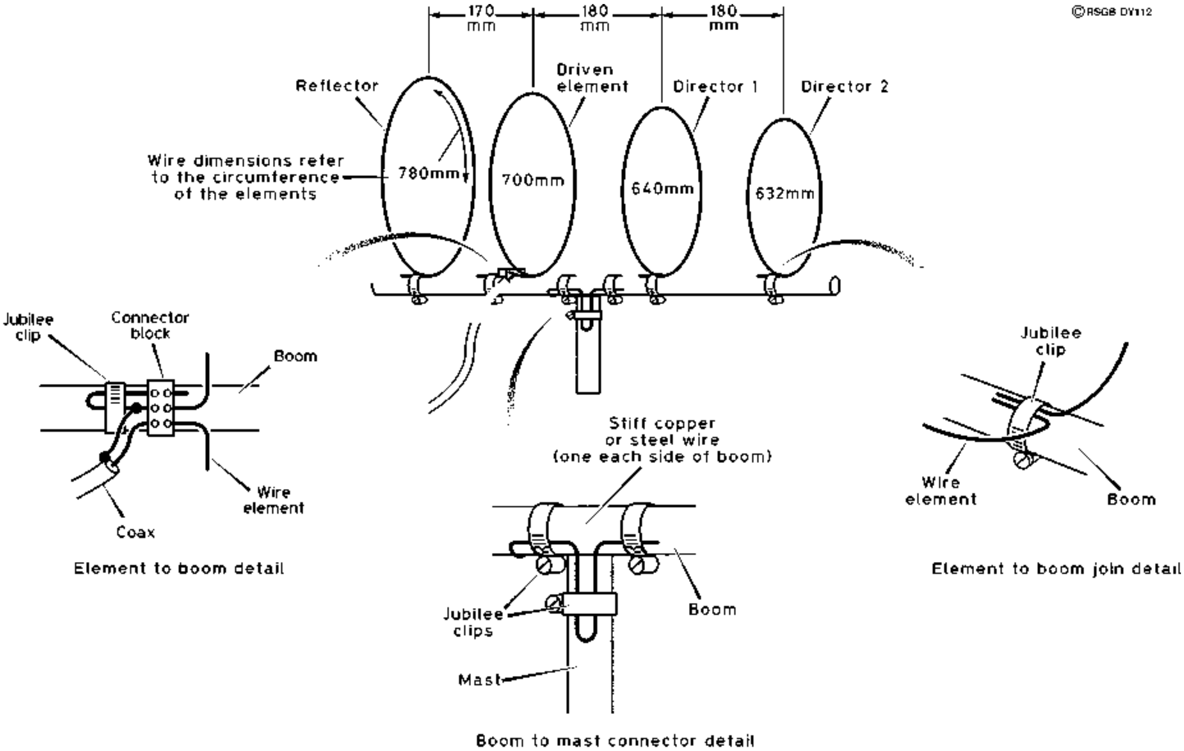
170 mm
180 mm
180 mm
©RSGB DY112
Reflector
Driven element
Director 1
Director 2
Wire dimensions refer to the circumference of the elements
780mm
700mm
640mm
632mm
Jubilee clip
Connector block
Boom
Wire element
Coax
Element to boom detail
Jubilee clip
Wire element
Boom
Element to boom join detail
Stiff copper or steel wire (one each side of boom)
Jubilee clips
Boom
Mast
Boom to mast connector detail

Figure 1 70 cm 4 element quad construction, with detail of how hose clamps are used in the construction

You will need a plug on the shack end of your feeder to suit the transceiver, aerial tuning unit (ATU) or the standing-wave meter (SWM) you are using. Always connect these in the order: transceiver – SWM – ATU – aerial.

Testing

Always test aerials outside and away from buildings (if possible!). This avoids getting misleading results.

Use a rubber duck, or whatever aerial you usually use, and tune around to find a repeater or beacon signal which is consistent. Note the reading on the S-meter. Then, connect your new Quad loop. Rotate it to give the strongest S-meter reading (don't forget it is directional). Verify that the meter reading varies as you rotate it. How does the S-meter reading compare with the original reading? Much depends on the siting of your original aerial; if you are comparing your Quad loop at ground level with a vertical on the chimney, you wouldn't expect your new aerial, even with its gain, to outperform a vertical which is well elevated!

Now a test on transmit is called for. The use of an SWR meter is essential here. Find a clear frequency and check that it *really is clear* before announcing your presence and carrying out the test. A reading of 1:1 is excellent, but any value less than about 1.8:1 would be acceptable. You can measure the directivity of your aerial using the field strength meter, also described in this series.

Materials

4 metres of 14 SWG enamelled copper wire
Material for boom and mast
15 amp connector block
7 Jubilee clips

The enamelled copper wire is available from AA&A Ltd, Sycamore House, Northwood, Wem, Shropshire SY4 5NN. Everything else is available from most hardware stores.

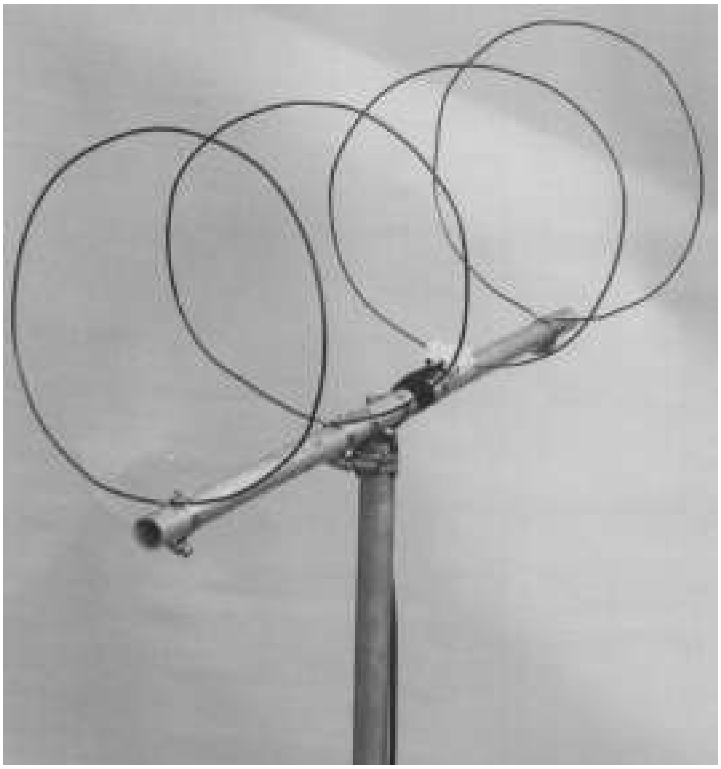

46 A UHF field strength meter

Introduction

It is always interesting, and often useful, to know just *where* the radiation is going from your aerial. How much of your radiated energy is going in the general direction of the station you are in contact with, and how much is being effectively wasted? Some of these questions can be answered with the use of a *field strength meter*. A field strength meter is simply a receiver, stripped down to its bare essentials, such that it responds only to the magnitude of the carrier. The use of a field strength meter assumes that the aerial under test is radiating a continuous carrier. Don't forget to find a clear frequency and identify your transmissions at least every quarter-hour, in order to comply with the terms of your licence.

Description

Two types of field strength meter are shown in **Figure 1**. You will recognise both circuits (Figure 1b particularly) as being types of 'crystal set' with a meter replacing the headphones. Figure 1a is a broad-band HF design (there is no tuning provision) and Figure 1b is tuned in the same way as the crystal set; with a loop of wire as an aerial, it will perform well in the VHF/UHF range.

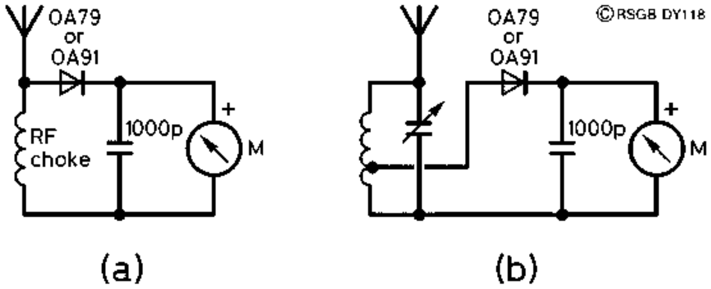

Figure 1 Construction of the UHF field strength meter

A tuned field strength meter also doubles as an *absorption wavemeter* if it is well calibrated. Such devices are useful for detecting transmitter harmonics also. This design is intended for use in the UHF band, so it will have to be sensitive around 432 MHz.

Construction

The field strength meter comprises a loop of wire, 600 mm long, which acts both as aerial and as the tuning inductance, a diode, a capacitor, a connector block, a meter and a length of twin wire. All the components, with the exception of the meter, are fixed to a pole with a jubilee clip, as shown in **Figure 2**. The meter should have a sensitivity of between 50 μA and 100 μA, or a multimeter can be used. The multimeter is more flexible, as you can select different current ranges, giving you a range of sensitivities.

Using the field strength meter

Connect your meter to the ends of the twin wire from the pole. Place a hand-held transceiver about 2 metres away, and press Transmit. If there is no meter reading switch off the transmitter and check the wiring. If the needle attempts to go negative, simply reverse the wires to the meter. If the reading is too high, either move the transmitter further away, or increase the current range on the multimeter. Try changing the orientation of the transmitter aerial, and note how the signal varies.

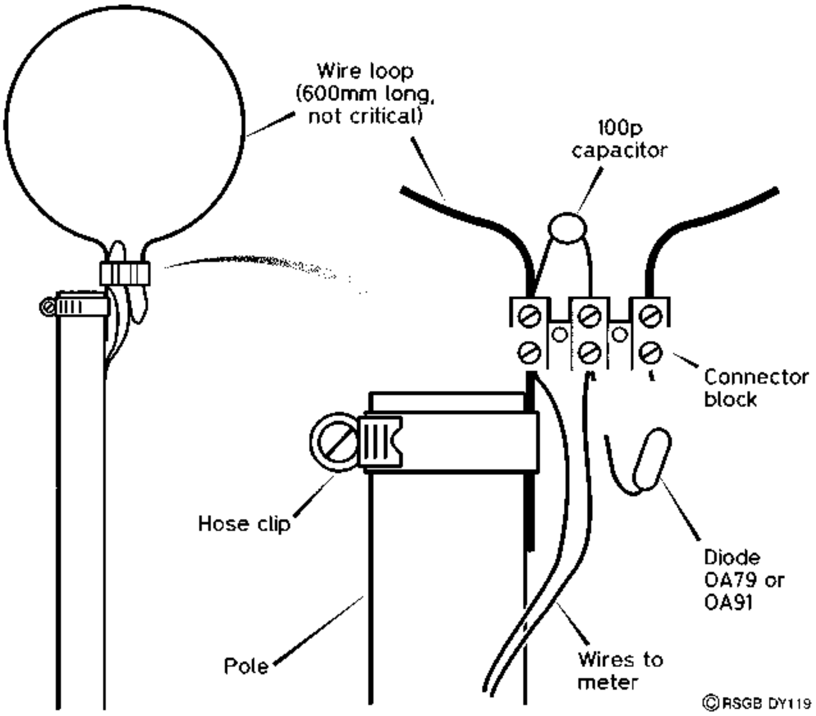
Wire loop
(600mm long,
not critical)
100p
capacitor
Connector
block
Hose clip
Diode
OA79 or
OA91
Pole
Wires to
meter
©RSGB DY119

Figure 2 Assembly of the field strength meter

To measure the *polar diagram* (a graph of the signal strength against the angle between the aerial boom and the field strength meter, plotted on polar axes) of a beam aerial, mount the loop as far away from the aerial as you can, preferably at the same height, and rotate the aerial, say 15 degrees at a time, and record the signal strength, until the aerial has been turned through 360 degrees. If you need help plotting the graph, enlist the help of a friend who has done it before, or one who knows about polar graphs! If you have already built the Photometer project, you will have measured and plotted the polar diagram of the light intensity from a torch. Now you will see the great similarity!

Parts list

Capacitor	100 picofarads (pF)
Diode	Germanium, OA79 or OA91
Connector block	10 A, 3-way
Wire	600 mm of 16 SWG enamelled copper
	Length of twin cable for meter connection
Clip	Jubilee (hose) clip

47 Christmas tree LEDs

Introduction

A novelty ideal for the festive season, this circuit causes one LED at a time to light up around a small cardboard Christmas tree.

Warning

This circuit uses members of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which each IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains-earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**
3. **Repeat the process with the second IC.**

The circuit is safe from destruction while it is connected to the battery. However, when the battery is removed, the same care should be exercised with its handling, because there are no supply decoupling capacitors across each IC.

Description

Figure 1 shows the layout, and how the wires are connected from the circuit board to the LEDs around the tree. The circuit, shown in **Figure 2**, is quite complicated, so you need to be confident in your logical approach to circuit-building before you attempt this one! It uses two common integrated circuits (ICs). IC1 is simply an oscillator which provides timing pulses for IC2, which 'counts' up to a maximum of 10. The outputs of the counter are indicated by light-emitting diodes (LEDs); red, orange, yellow and green are common colours which you can use.

The circuit is built on a single piece of Veroboard measuring 29 holes by 12 strips. **Be aware that there is no strip labelled 'I', so don't make mistakes in your counting!**

First of all, cut the tracks in the positions shown in Figure 1 using a 3 mm (⅛ inch) twist drill rotated between thumb and forefinger. Then, solder in the IC sockets, the notches facing row M. Wire up and solder in the links and then the Veropins for the connections to the LEDs and battery. Having done this, the resistors and the capacitor should be fitted. When wiring the LEDs, each cathode (the lead adjacent to the 'flat' on the LED encapsulation) is connected to R4, and the anodes go to separate pins on IC2.

Testing

Hold up your circuit board to the light and check carefully for solder bridges between adjacent tracks. Then check again that the wiring is correct. Place the ICs in their holders, with the notched ends lining up with those on the holders. Connect the battery, and the LEDs should illuminate in sequence.

If you have no success, you are now wishing you had checked the circuit more carefully! Learn something from your mistake and it will not have

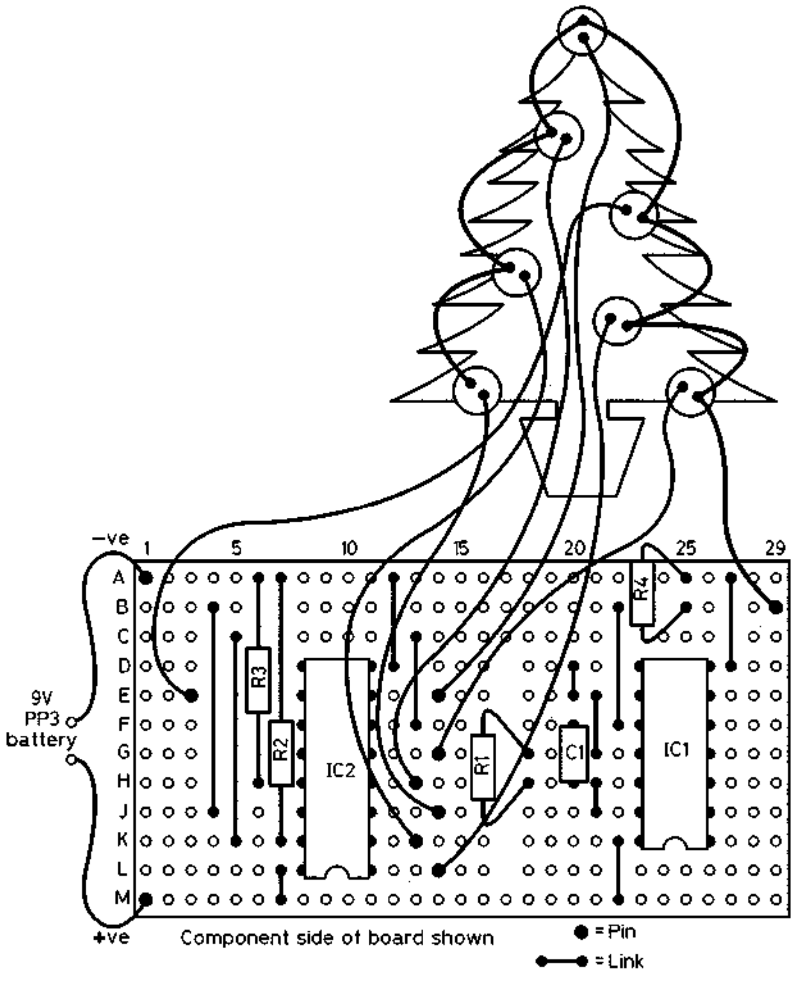
−ve
1
5
10
15
20
25
29
A
B
C
D
E
F
G
H
J
K
L
M
9V
PP3
battery
R3
R2
IC2
R1
C1
IC1
R4
+ve
Component side of board shown
= Pin
= Link
= Cut in copper track

©RSGB DY137

Figure 1 Christmas tree, component layout

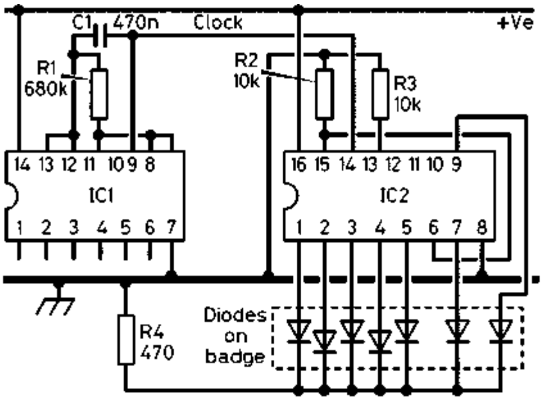
C1
470n
Clock
+Ve
R1
680k
R2
10k
R3
10k
14 13 12 11 10 9 8
IC1
1 2 3 4 5 6 7
16 15 14 13 12 11 10 9
IC2
1 2 3 4 5 6 7 8
R4
470
Diodes
on
badge
D6 D2 D1 D3 D7 D4 D5

©RSGB DY138

Figure 2 Christmas tree, circuit diagram

been in vain! The first thing to check is that the LEDs are the correct way round. If that is OK, then check the positions where the tracks are broken (intentionally!). After that, are the wire links in the right places – had you forgotten there is no row 'I'? Check *all* the wiring, then check *again* for solder bridges, and switch on again. One of these tests should have revealed a fault. If it *still* refuses to work, perhaps one (or both) of your ICs were damaged by static electricity, despite your precautions – or did you choose to ignore them?

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	680 kilohms (kΩ)
R2, R3	10 kilohms (kΩ)
R4	470 ohms (Ω)
Capacitor	
C1	0.47 microfarad (μF) min. metallised polyester film
Semiconductors	
IC1	4011
IC2	4017
Additional Items	
D1–D7	5 mm LEDs in choice of colours
	14-pin DIL socket for IC1
	16-pin DIL socket for IC2
	PP3 battery clip and battery
	Veroboard, Veropins
	Insulated wire for links

48 An audio signal injector

Introduction

An audio signal injector is a device used to test audio frequency circuits. It is simply an oscillator running at a frequency in the audio range, so that when its output is fed to the input of an amplifier, it will produce a sound in the loudspeaker if the amplifier is working. The oscillation is so rich in harmonics that the signal can also be heard (sounding rather different) when injected into an RF circuit.

The design

The circuit is shown in **Figure 1**, and is a basic *astable multivibrator*, a free-running oscillator producing a roughly rectangular-wave output. The two transistors, TR1 and TR2, operating as switches, switch on and off alternately at a frequency around 500 Hz. The prototype was constructed on plain matrix board (no copper strips), as illustrated in **Figure 2**.

Both transistors are type BC108, which are only a few pence each new, and can be found at almost any rally. You can add an on/off switch, or simply disconnect the battery when you are finished using it. To make the unit in one piece, the battery can be taped to the board, as the diagram shows.

The probe itself is made from a short piece of stiff insulated wire, soldered to a tag on the board; an earth lead is also soldered to the board, and terminated in a crocodile clip to attach to the ground lead of the equipment under test.

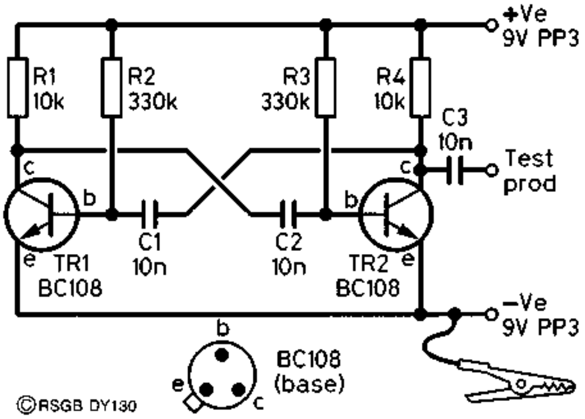

Figure 1 Circuit diagram of the signal injector

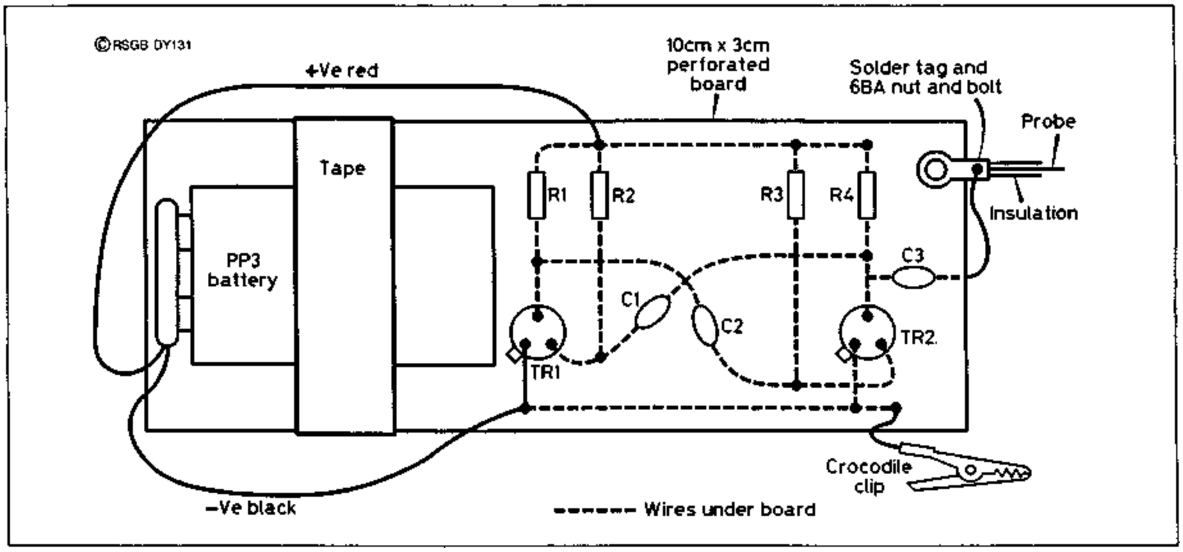
©RSGB DY131
+Ve red
10cm x 3cm perforated board
Solder tag and 6BA nut and bolt
Probe
Tape
R1
R2
R3
R4
Insulation
PP3 battery
C3
C1
C2
TR1
TR2
Crocodile clip
–Ve black
Wires under board

Figure 2 Component layout of the signal injector

Using it

After checking the wiring of the circuit, connect the battery. Find a transistor radio that is known to work. Carefully remove the case, switch on and apply the probe to the centre tag of the volume control. A *very* loud buzz should be heard from the speaker, indicating that the audio circuits of the radio are working. Using the injector to fault-find equipment you have made yourself is rather more instructive and rewarding, because you *know* where to inject the signal, and you should *know* what to expect when you do.

Warning

Do *not* work on any equipment connected to the AC mains. Work only on battery-powered circuits, for your own safety.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R4	10 kilohms (kΩ)
R2, R3	330 kilohms (kΩ)

Capacitors

C1, C2, C3	10 nanofarads (nF) or 0.01 microfarad (μF) ceramic

Semiconductors

TR1, TR2	BC108

Additional items
- PP3 battery and connector
- Matrix board 10 cm by 3 cm
- 6BA solder tags
- Thick insulated wire for probe
- Crocodile clip

49 Standing waves

Introduction

Everyone has, or *should* have, a standing-wave ratio (SWR) meter as part of his/her array of test gear. Most people know how to use it, but what does it really do?

Before attempting to answer that question, we need to look at some aerial fundamentals. An aerial is a *transducer*, the word meaning 'to lead across'. We use it whenever one form of energy is converted into another form. A bulb is a transducer; it converts electrical energy in the filament to radiated heat and light energy. **Figure 1** shows the situation.

An aerial, or antenna, is also a transducer; it converts radio-frequency (RF) energy in the feeder (or *transmission line*, to give it its proper name) into radiated electromagnetic energy, in the manner shown in **Figure 2**. The aerial has resistance, just like the bulb filament, and if the filament had zero resistance, there would be no radiated energy. In the same way, if the aerial had no resistance, there would be no radiation from it.

Transmission lines should convey RF energy from transmitter to aerial with the minimum power loss. A common form of transmission line is the *coaxial*

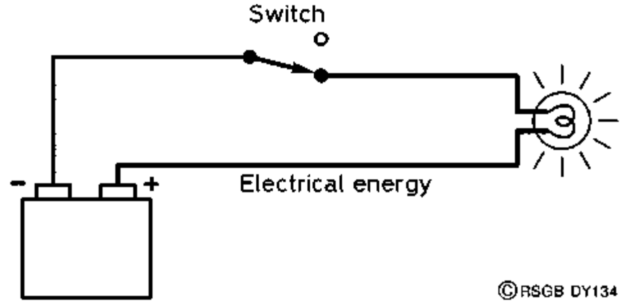

Figure 1 The light bulb converts electrical energy to electromagnetic light energy

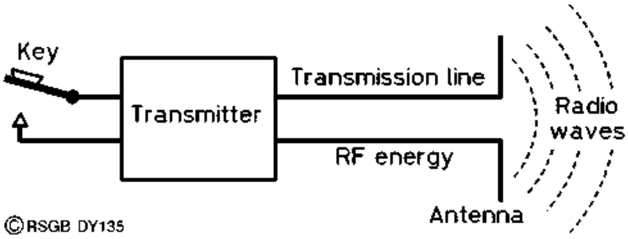

Figure 2 The antenna converts RF current to electromagnetic radio waves

cable. As you might expect, any feeder cable has a DC resistance caused by the resistance of the copper wire from which it is made. It also has an AC resistance, caused by the capacitance between the centre conductor and the braid, and by the inductance of the cable itself. This means that the feeder has an *impedance*, which is constant for the particular type of cable. This is what we call the *characteristic impedance* and, for the cables used in most amateur radio applications, it is 50 Ω.

If this impedance can be made the same as that of the aerial (it is already the same as the impedance of the transmitter output), then the transfer of energy will occur with minimum loss. If the aerial and cable impedances are *not* the same, then there is a *mismatch*, which causes some of the RF energy to be reflected back towards the transmitter. We now have a situation where RF energy is flowing along the cable from the transmitter to the aerial (the *forward* wave) and, at the same time, flowing from the aerial to the transmitter (the *reflected* wave). The two waves interact along the cable and form a stationary pattern of voltage and current. The pattern is known as a *standing wave*, and can be visualised from the waveforms in **Figure 3**. The ratio of the maximum voltage to the minimum voltage on a given wave defines the *voltage standing-wave ratio* (VSWR), or just *standing-wave ratio* (SWR) for short.

The SWR meter is easy to use. It is positioned in the feeder between the transmitter and the aerial tuning unit (ATU) if there is one, or between the transmitter at the aerial otherwise. Most meters have a single meter and a four-position switch. The transmitter is first keyed and, with the switch in

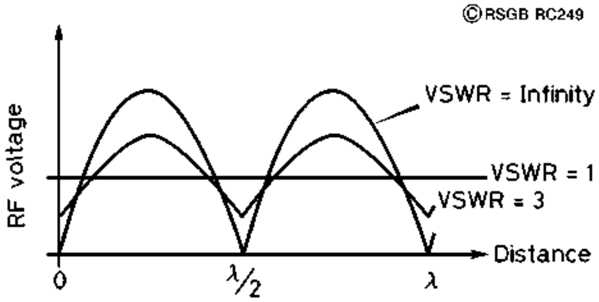

Figure 3 Standing waves on a transmission line

the 'calibrate' position, the sensitivity control is adjusted to give full-scale deflection on the meter. The switch is then changed to 'forward' to read the forward power, to 'reflected' to read the reflected power, and to 'SWR' to read the value of the standing-wave ratio. When there is no reflection (see Figure 3), the meter should read 1:1 or, simply, 1.

Most SWR meters remain in the feeder line while the transmitter is operating, so the condition of the aerial and feeder can be constantly monitored. Problems with the aerial (such as water entering the feeder at its junction with the driven element) are immediately shown up. Without the use of the SWR meter, the situation would slowly deteriorate over several months and you would be left wondering why so few stations were answering your calls!

50 A standing-wave indicator for HF

Introduction

The standing-wave ratio (SWR) meter shows how well the aerial system, including the feeder, is *matched* to the output of the transmitter. This design does not *measure* SWR, but it gives an indication of when the SWR is minimum for a given system of aerial and feeder. The design is usable on the HF bands from 1.8 to 28 MHz, and can be used at 50 MHz with reduced sensitivity.

How it works

There are two types of wave in any feeder: the *forward* wave, which travels from the transmitter to the aerial; the *reflected* wave, which travels back to the transmitter from the aerial. The presence of a reflected wave is evidence that some of your transmitted power is not being radiated, but is being returned to the transmitter to be lost as excess heat. When aerial and feeder are perfectly matched, there is *no* reflected wave, and all the power from the transmitter is radiated.

Referring to the circuit of **Figure 1**, a tiny fraction of the signal is removed by the transformer, T1, and by the capacitors, VC1 and C1. It is then detected by the germanium diodes, D1 and D2, and any residual RF removed by the capacitors, C2 and C3. The currents through the diode and meter (depending on the position of switch, S1) represent the forward and reflected signals. VR1 acts as a sensitivity control for the meter.

It pays to shop around for a suitable meter. Surplus types from tape recorders and hi-fi equipment are usually ideal for this purpose. A new one would cost several pounds. The more sensitive the meter, the more sensitive your indicator will be. Meter sensitivity is given by the current required to give full-scale deflection (FSD) of the pointer. One with an FSD of between 50 and 200 micro-amps (μA) is suitable for this circuit. The higher the FSD, the less sensitive the circuit.

Construction

The meter circuit and the sampling transformer (see **Figure 2**) are built and mounted on Veroboard of the copper-strip variety. It simplifies construction

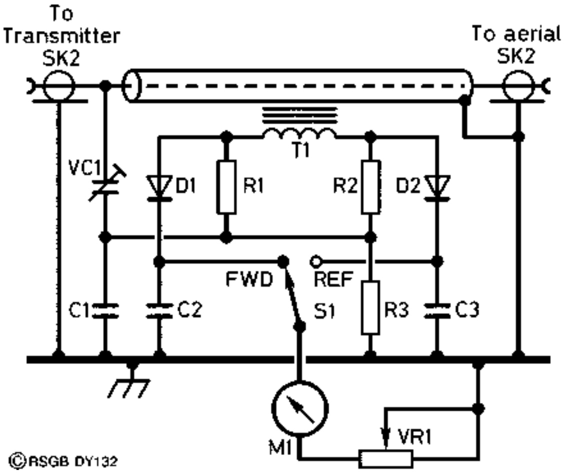

Figure 1 Circuit diagram of the SWR meter

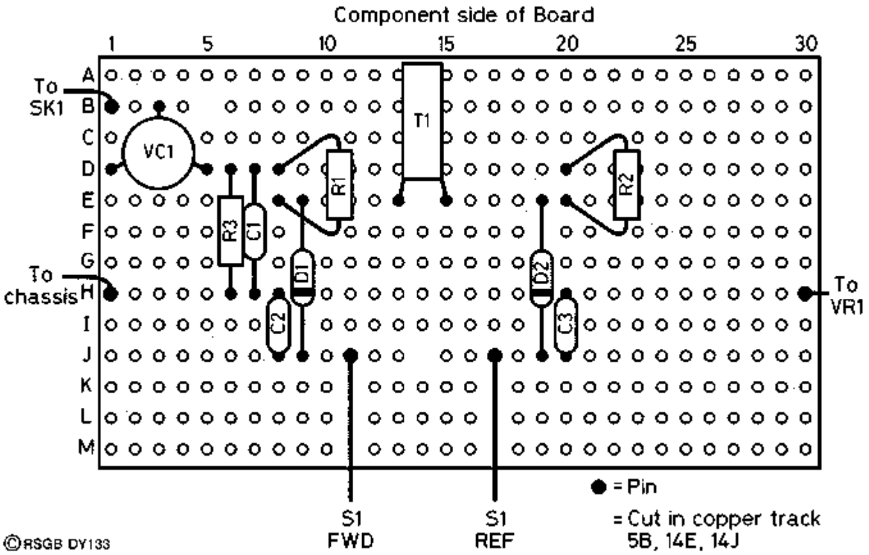

Figure 2 Component layout of the SWR meter

but reduces the operational range of the meter to below 30 MHz because of the capacitive coupling between strips. The board has 13 strips by 30 holes, although you can reduce this if you have a smaller case.

Firstly, cut the tracks at the three points shown. Then insert and solder Veropins for connections to the external components, the switch, variable resistor and the meter. Solder in the components starting with the resistors and followed by the capacitors and the diodes, ensuring that the diodes are inserted correctly.

Now you have to wind the transformer, T1, on a small toroidal ferrite core. Wind the secondary with 15 turns of 36 SWG enamelled copper wire, spaced evenly over about two-thirds of the former. The turns should not overlap, and considerable care must be taken; the wire is very thin, will kink easily and will break if you apply too much tension. The ‘primary’ is an 8 cm length of 50 Ω coaxial cable which passes through the toroid on its way between the input and output connectors. The braid of the cable is connected to the case at only *one* of the connectors (see Figure 1); this prevents the screen and the metal case between the two sockets forming a single, shorted turn.

The ends of the secondary winding must be carefully stripped of their enamel with sandpaper, before attaching the toroid to the board with cotton or nylon fishing line. On no account must wire be used for this!

Solder the secondary connections of T1 to the board and thread through the coaxial cable ready for soldering to the connectors.

The case used was an aluminium box (Maplin LF02C), but any suitable metal box could be used. Aluminium is preferable, as it is easily drilled with simple tools. Use standoff insulators to mount the board in the case. Once this has been done, the leads from the board to the chassis-mounted components can be soldered. So can the coaxial cable passing through the toroid. Make the lead from the input socket to VC1 as short as possible.

Setting up

You will need a 50 Ω dummy load and a transmitter to set up your indicator. Connect the transmitter to SK1 and the dummy load to SK2. Set the toggle switch, S1, to *forward* and the sensitivity control, VR1, to mid-travel. Switch on the transmitter, and set VR1 for maximum meter deflection. Switch to *reflected* and adjust VC1 until the reading is minimum (ideally zero). This completes the setting up!

Using the indicator

For setting up an aerial, connect your circuit between the transmitter and the cable leading to the aerial. With S1 in the *forward* position, key the transmitter and adjust VR1 for maximum reading on the meter. Switch to *reflected*, and then adjust your ATU to give minimum reflected power. If your adjustments are to be made to the aerial itself, to give minimum reflected power, you must make a note of the reflected reading, switch off the transmitter, make a change to the aerial, key the transmitter, and note whether the reflected power is greater or less than before. Then, make more changes to the aerial. **Never adjust your aerial with the transmitter on**. Make your adjustments on an unused frequency, and do it as quickly as possible, thus avoiding (or minimising) interference to other stations.

Parts list

Resistors: all 0.25 watt, carbon 5% tolerance
(or Maplin 0.6 watt metal film)

R1, R2	27 ohms (Ω)
R3	2.2 kilohms (kΩ)
VR1	10 kilohms (kΩ) linear

Capacitors

C1	220 picofarad (pF) disc ceramic 50 VDC
C2, C3	0.1 microfarad (μF) disc ceramic 50 VDC
VC1	20 picofarad (pF) trimmer

Semiconductors	
D1, D2	OA91 germanium
Additional items	
S1	Single-pole changeover (SPDT or SPCO)
SK1, SK2	Coaxial sockets to suit station standards
	Veroboard – 13 strips by 30 holes
	Veropins (7 off)
	Amidon FT 50–43 ferrite toroid
	Meter, less than 200 μA FSD
	36 SWG enamelled copper wire
	Short length of UR43 or RG58 coaxial cable
	Insulated stranded wire
	Aluminium box
	Standoff insulators for mounting the board
	Knob for the sensitivity control

51 A moisture meter

Introduction

Dry rot (*Merulius Lacrymans*) can strike havoc in buildings, causing the timbers to decay and crumble to dust – hence the term *dry rot*. Wood is attacked only if its moisture content rises above 20%.

Construction

The circuit of the moisture meter is shown in **Figure 1**. The two probes touch the wood, and the current that flows between them depends on the moisture content of the wood. If the moisture is sufficiently high, the current, after amplification, will be enough to light the LED.

The meter can be made on a piece of plain matrix board (no copper strips), as **Figure 2** shows. The board is big enough (10 cm by 2.5 cm) to accommodate the PP3 battery, taped on. No case is needed, unless you want

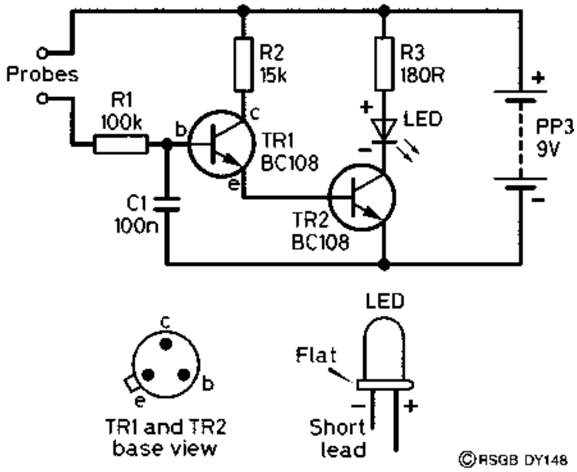
Probes
R1
100k
b
R2
15k
c
TR1
BC108
e
C1
100n
R3
180R
+
LED
−
TR2
BC108
+
PP3
9V
−
LED
Flat
−
+
Short lead
c
b
e
TR1 and TR2 base view
©RSGB DY148

Figure 1 Circuit diagram of a moisture meter

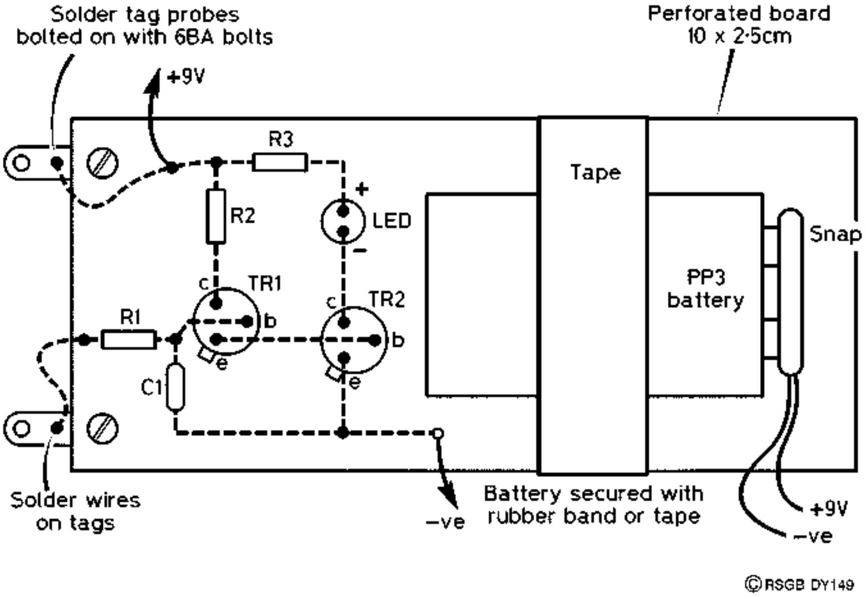
Solder tag probes bolted on with 6BA bolts
+9V
Perforated board 10 x 2·5cm
R3
+
R2
LED
−
Tape
Snap
PP3 battery
c
TR1
b
e
c
TR2
b
e
R1
C1
Solder wires on tags
−ve
Battery secured with rubber band or tape
+9V
−ve
©RSGB DY149

Figure 2 Moisture meter, component layout

to leave the meter in a damp location for a protracted period! Make sure that the transistors and LED are mounted correctly. In Figure 2, the connections as shown to TR1 and TR2 are illustrated as if the transistors were transparent. An on/off switch is not really necessary, as only a very small current flows when the probes do not touch anything. Use solder tags, screwed to the board, to act as probes.

After the assembly is completed, check your circuit one final time, and then connect the battery. Nothing should happen at first. If you lick your forefinger and hold it across the probes, the LED should light.

Using it

The prototype was compared with a commercial moisture meter, and the LED lit when the moisture was around 20%. This was quite fortuitous, as the point at which the LED lights depends both on the separation of the probes and on the gain of the two transistors.

In addition to searching for dry rot, the instrument may be used to monitor the moisture in the soil of household plants. In this case, probes made of 16 SWG copper (*not* enamelled) should be soldered on to the two tags, and should penetrate the soil to a depth of several centimetres, and R1 may require adjusting so that the LED extinguishes if the soil is too dry, and lights if the soil is sufficiently moist. If you wanted to leave the meter with the probes in the soil, an on/off switch *would* be necessary.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	100 kilohms (kΩ)
R2	15 kilohms (kΩ)
R3	180 ohms (Ω)

Capacitor

C1	0.1 microfarad (μF) polyester

Semiconductors

TR1, TR2	BC108
LED	Any shape or colour will do

Additional items
PP3 battery and connector
Solder tags (2 off) for probes
Matrix board 10 cm by 2.5 cm

52 Simple aerials

Introduction

The performance of *any* receiver or transceiver, no matter how expensive it is, is limited by the aerial that feeds it. Two of the most frequently asked questions are:

- Which is the best sort of aerial to use?
- Where is the best place to locate an amateur radio aerial?

To answer these questions, you must ask yourself what sort of operation you want to do. Are you interested in local, chatty contacts on the lower bands or VHF, or are you more disposed towards long-distance (DX) contacts, and on what band?

A house with a moderately sized garden is assumed in the diagrams here, to illustrate the configurations of some simple aerials. You would not need all these aerials festooned around your house, because one or two would be sufficient for your needs. The problems incurred by properties with more restricted space will be covered later.

VHF aerials

For VHF operation, the aerial should be mounted as high as possible, either on a mast or on a chimney. For all-round coverage on FM and the local repeaters, a vertical colinear is a good choice. For SSB and CW DX operation, a horizontal rotatable beam is needed. If satellite working is envisaged, you will need to contemplate mounting an elevator on top of your rotator, so that your beam can point in *any* direction, including vertically upwards! An advantage of satellite working is that the aerials do not *necessarily* have to be up in the air, provided you have a relatively uncluttered site. Your rotator and elevator can be at ground level, which is good!

If the VHF aerial is mounted on the chimney, use a double mounting bracket, particularly if you have a beam and rotator. Keep the TV, broadcast FM and amateur aerials as far apart as possible, and keeping the feeders separated is also a good plan.

The dipole aerial

One of the simplest types of aerial for single-band operation is the half-wave dipole. (The name ‘dipole’ simply means ‘two poles’ or ‘two elements’, and

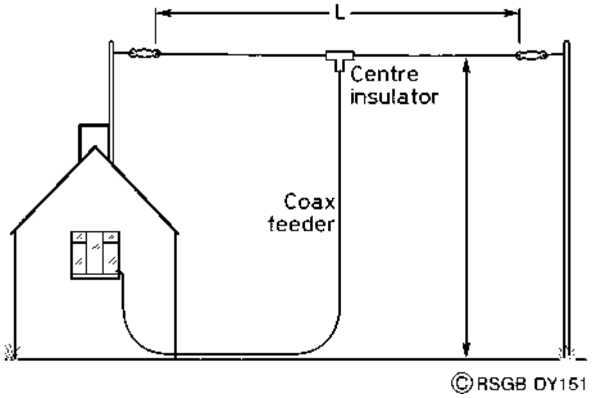

Figure 1 Layout for a dipole aerial

in this case the total length of the dipole is approximately half a wavelength at the operating frequency.) It is usually fed in the centre by coaxial cable as shown in **Figure 1**. The length of the dipole for the lowest frequency in each band is shown in **Table 1**. Normally, the length of the aerial will be 'trimmed' to be tuned to the centre frequency of the part of the band in which you will operate. This is done using the data in the right-hand column of Table 1. As an example, suppose you wanted your aerial to be resonant at 3.7 MHz. The table gives an overall dipole length of 42.86 m for 3.5 MHz. To resonate the aerial 200 kHz *higher*, then this length must be shortened by 2 × 0.595 m = 1.190 m. Your dipole would thus be 41.67 m long. Remember to allow extra wire for fixing the dipole ends to the insulators.

Table 1 Dipole lengths for lowest frequency of each band and the length to be trimmed from each to raise the resonant frequency by 100 kHz

Band (MHz)	*Dipole length (m)*	*Trim each end (mm/10 kHz)*
1.8	83.33	2190
3.5	42.86	595
7	21.43	150
10	14.85	70
14	10.71	35
18	8.33	20
21	7.14	15
24	6.03	12
28	5.36	10
50	3.00	6

On the lower-frequency bands, the lengths become rather large. In this case, you can 'bend' your dipole, as illustrated in **Figure 2**. The length of wire required to give an acceptable value of SWR (less than 2:1 on transmit) may need to be different from the calculated value, so be prepared to experiment!

Dipoles are single-band aerials, although they will often work acceptably on the third harmonic of their design frequency: a 7 MHz dipole often operates reasonably well on 21 MHz. It is possible to operate several dipoles in parallel, as **Figure 3** shows. Interaction between the elements can occur if the spacing between them is less than about 10 cm. A multi-band dipole, as shown in Figure 3, has the elements separated with plastic spacers, and drooping ends to produce maximum spacing between the elements' ends.

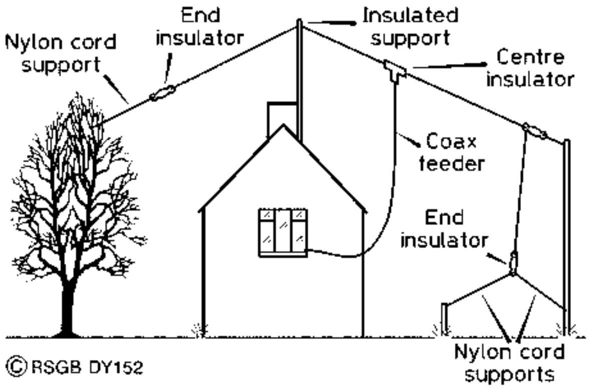
Nylon cord support
End insulator
Insulated support
Centre insulator
Coax feeder
End insulator
Nylon cord supports
©RSGB DY152

Figure 2 Possible layout for a dipole aerial in a confined space

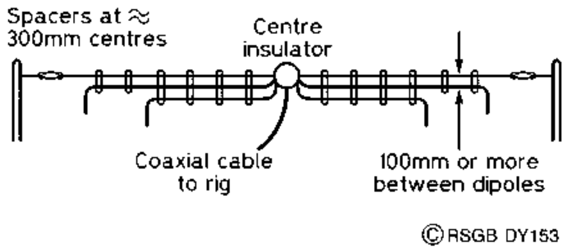
Spacers at ≈ 300mm centres
Centre insulator
Coaxial cable to rig
100mm or more between dipoles
©RSGB DY153

Figure 3 Multi-band dipole aerial

The long-wire aerial

This aerial is simple, cheap, easy to erect, and suits most houses and gardens, as **Figure 4** shows. Using an aerial tuning unit (ATU), an end-fed long wire can function on several bands when used with a set of radials or a counterpoise. **Figure 5** illustrates the setup. The length of the aerial will determine the bands which will be covered.

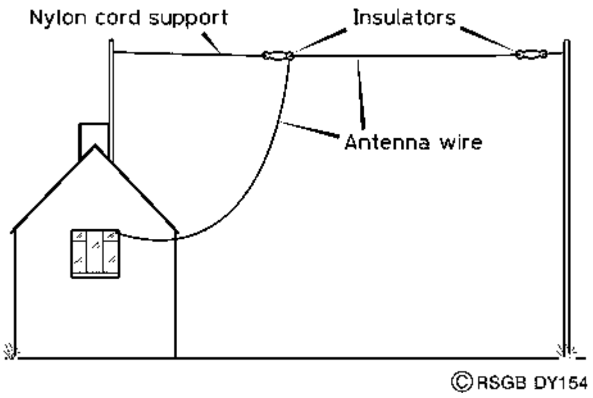
Nylon cord support
Insulators
Antenna wire
©RSGB DY154

Figure 4 Long-wire or inverted-L aerial

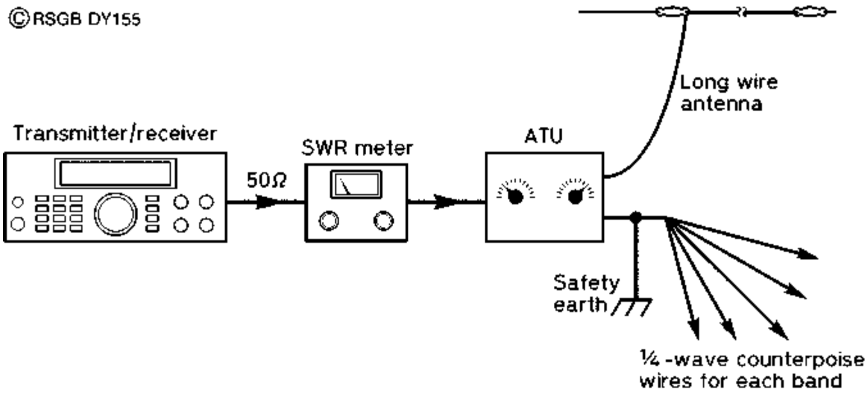
©RSGB DY155
Long wire antenna
Transmitter/receiver
SWR meter
ATU
50Ω
Safety earth
¼ -wave counterpoise wires for each band

Figure 5 How to connect a radio to a long-wire aerial

> A wire length of 10.5 m will work on the 40, 30, 17, 15 and 12 m bands.
>
> A wire length of 15.5 m will work (with an ATU) on the 80, 40, 20 and 12 m bands and possibly (depending on your ATU) on the 17 and 15 m bands.
>
> A wire length of 26.5 m will operate on all bands, but may be difficult to load on 10 m.

The wire lengths given here may need some adjustment because of the geometry of your particular house and garden. For receive-only purposes, the lengths are far less critical.

In general, you cannot get a good radio-frequency (RF) earth from a first-floor (or higher) shack. Unless a good RF earth exists within a small fraction of a wavelength of the transceiver, an *artificial ground* comprising a single $\frac{\lambda}{4}$ radial or counterpoise will be needed. You will need one counterpoise for each band you intend to use, and the wire can be concealed around the skirting-board of the shack, or under the carpet. Make sure that the free end

Table 2 Lengths of elements for vertical antennas, radials for verticals and counter-poises for end-fed long wire antennas

Band (MHz)	*Element length (m)*
1.8	39.66
3.5	20.40
7	10.20
10	7.14
14	5.1
18	3.96
21	3.4
24	2.95
28	2.55

of each counterpoise is well insulated; this point can carry a *very high voltage* when you transmit; anyone coming into contact with this can suffer very severe RF burns. Counterpoise lengths can be read from **Table 2**.

The vertical aerial

The single-band vertical aerial is sometimes used by DX operators because it has a low angle of radiation, which favours long-distance propagation. However, it must be sited clear of obstructions and must have a *good* counterpoise or radial system. Illustrations of the vertical aerial are shown, and the lengths of the vertical and radial sections are given in Table 2. The centre of the coaxial feeder is connected to the vertical section, and the braid to the counterpoise or radial system, which is made up of four or more wires buried just below the surface and joined together near the base of the aerial.

Cable entry to the house

Bringing coaxial cable into the house by an open window must be regarded as a temporary measure. Wooden window frames can be drilled, one hole for each feeder. Make the holes slope downwards from inside to out to prevent rain entering, and treat these with wood preservative. Leads from long-wire and inverted-L aerials should be kept separate from other cables.

Alternatively, a plastic pipe large enough to take all your feeders could be fitted into the brickwork (again, sloping downwards towards the outside). You may want to let a friendly builder do this for you.

53 A breadboard 80 m CW transmitter

Introduction

In the early days of radio, many circuits were built on a wooden baseboard, the parts being screwed down on the board. This was called breadboard construction, because it was a breadboard that was frequently commandeered for the process! Wives have always been generous in this respect, it appears!

This circuit was originally designed by GM3OXX, and became known as the *Oner*, because it was built on a circuit board one inch square! The circuit appeared in the G QRP Club journal *Sprat* and, since that time, many hundreds of *Oner* circuits have been built and used on the air. It is a well-proven circuit.

The transmitter has no tuned circuits in the power amplifier (PA) and thus has a rather high *harmonic* content. It **must** be used with the ***low-pass filter*** described elsewhere in this book. Without the low-pass filter, interference will be caused to other stations.

Simple aerial changeover switching is provided, which allows this circuit to be used with any of the 80 m receivers, such as the *Colt*, described in this book. It can also be used with any kit or commercial receiver for the 80 metre band.

The circuit

The transmitter circuit is shown in **Figure 1**. TR1 is a crystal oscillator, the frequency of which is controlled by crystal X1. A small trimmer capacitor, TC1, is added to allow the frequency of X1 to be varied by a small amount. If adjustment of this trimmer is made possible from the front panel, it is useful to adjust the transmit frequency to avoid other stations already on the crystal frequency. The collector load resistor, R2, of the oscillator transistor, TR1, determines the power output; a value of 3.3 kΩ seems to work well in producing an output of 3 watts.

TR1 is directly coupled to TR3, a VMOS transistor (a type of field-effect transistor (FET)). This acts as the power amplifier (PA) stage. TR3 should

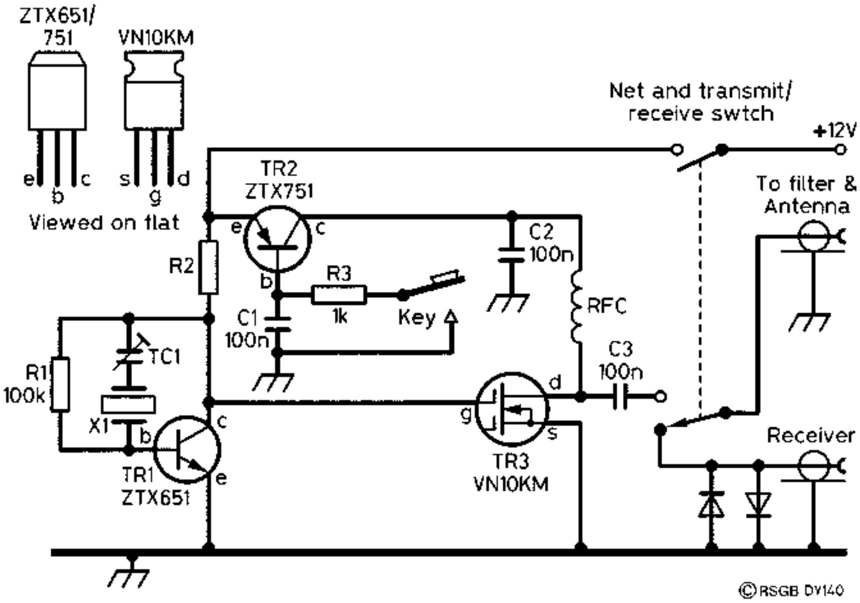
ZTX651/
751
VN10KM
e
b
c
s
g
d
Viewed on flat
Net and transmit/
receive swtch
+12V
TR2
ZTX751
e
c
b
To filter &
Antenna
R2
C2
100n
R3
1k
Key
RFC
C1
100n
C3
100n
R1
100k
TC1
X1
b
c
e
d
g
s
TR1
ZTX651
TR3
VN10KM
Receiver
©RSGB DY140

Figure 1 Circuit diagram of the breadboard transmitter

give about 3 W output, which is then coupled to the output by C3. The radio-frequency choke (RFC) providing the drain load of TR3 is simply a few turns of wire on a ferrite bead.

TR2 is an interesting addition to the circuit. It is used as a switch to 'key' the PA, TR3. The transmitter *could* be built without TR2, just placing the Morse key between the top of the RFC and the 12 volt supply. Adding TR2 is helpful, because it means that one side of the Morse key can be grounded (always a good thing), and some degree of *shaping* of the output RF waveform is provided by R3 and C1. This makes the transmission sound a little better and reduces the possibility of spurious frequencies being generated and transmitted. TR2 is a pnp transistor; note that it is the *emitter* of this transistor which is connected to the positive side of the supply.

Some form of changeover switching is needed for the aerial. A double-pole changeover toggle switch can be used. See the chapter on switches, later on in this book. One pole is used to switch the aerial between transmitter and receiver; the other pole is connected in the 12 volt supply line, and is labelled RECEIVE/TRANSMIT – NETTING. Its use will be described later. In this simple circuit, the PA cannot work when the key is open, because the key breaks its supply (via the RFC). When the key closes, TR2 switches on and applies the 12 volt supply to the top of the RFC. C2 is a *decoupling* capacitor, which prevents any residual RF signals at the top of the RFC reaching TR2.

Building

The prototype was put together as follows. Take a piece of plain printed-circuit board (PCB) measuring 5 cm by 4 cm. Then, with a *new*, *sharp blade* in a junior hacksaw, draw the blade horizontally across the surface of the copper in order to make a pattern of 6 squares along the 5 cm side and 5 squares along the 4 cm side. No more pressure should be applied than is necessary to cut through the copper! All the parts will be soldered on these pads in a form of surface-mount construction. To do this, each active pad (i.e. one that is going to have a component soldered to it) needs to be *tinned*. This means coating the pad's surface with solder, and is carried out as follows. Place the hot tip of the soldering iron on to the pad, and hold it there for a second or so. Then, with the tip still in place, touch the end of your reel of solder *on the pad*, not the tip of the iron. The solder should flow evenly all over the pad, and you can remove the iron. The solder should solidify in a rounded, shiny blob! This provides a good surface for making soldered joints.

To join component leads to the pads, cut each lead about 1 cm long, and then bend the last 2 mm at right angles to the rest of the lead. As you did before, tin the 2 mm length of each lead. Place the tinned portion on to the pad, and place the tip of your iron *on the pad*, close to the lead. The solder on the pad and on the lead will melt and run together; remove the iron and *hold the component still* until the solder solidifies. When the joint has cooled, give the lead a gentle tug to make sure you have a good joint (a good *mechanical* joint is usually a good *electrical* joint, too!). Each transistor straddles three pads, so the centre lead will need to be shorter than the other two. Take care here to get the lead lengths right – if you do, you will be surprised how much more firmly the transistor is held than if you just botched the lead lengths by bending them to fit! Make sure the connections to the transistors are correct.

Winding the RFC is quite simple. Seven (or more) turns of thin (32 SWG) enamelled copper wire are threaded through a small ferrite bead. This requires care, because the bead is small and the wire is thin. Trim the ends to within about 1 cm of the bead, remove the enamel carefully with sandpaper and tin the bare ends, prior to soldering the choke to the board.

After completion of the wiring, check the circuit against Figure 1. Bread-boarding a circuit like this has its advantages, but it can have disadvantages, too. One of these disadvantages is that it can make circuit checking difficult. For a simple circuit like this, it is not too bad! Check that no solder has run between the pads. Plug in your crystal for the 80 m band, and connect up the 12 V power supply. Do not connect the Morse key yet and do not switch on the power.

Testing and operating

Clear your workbench of all metallic objects, slivers of copper, bits of wire, etc., switch on your power supply. With an external receiver, listen on and around your crystal's frequency for a signal. Remember that the oscillator runs all the time and, because you haven't yet connected the Morse key, your receiver is close enough to pick up the signal from the oscillator. This confirms that your oscillator is running. Switch off.

Connect the station aerial to the transmitter's aerial socket, and the receiver to the transmitter's receiver socket. Connect the Morse key, put the Receive/Transmit–Netting switch in the receive position and switch on. You should be able to hear stations in the normal way. Now put the switch in the Transmit–Netting position. Signals in the receiver should almost disappear, as the circuit has disconnected the receiver's aerial.

Tune the receiver until you can hear your own crystal oscillator signal. This is known as *netting*, tuning your receiver and transmitter to the same frequency. Pressing the key will now transmit your signals when the switch is in the Transmit–Netting position; switch back to the receive position to listen for stations answering your call. As soon as you are happy that your circuit is functioning properly, **you must build the low-pass filter circuit** before using the transmitter regularly.

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	100 kilohms (kΩ)
R2	See text
R3	1 kilohm (kΩ)
Capacitors	
C1, C2, C3	100 nanofarads (nF), or 0.1 microfarad (μF)
TC1	3–60 picofarads (pF) trimmer
Semiconductors	
TR1	ZTX651
TR2	ZTX751
TR3	VN10KM
Additional items	
RFC	7 turns of 32 SWG enamelled copper on a ferrite bead
Switch	Double-pole changeover (DPDT or DPCO) toggle
Crystal	For 80 m band
Crystal holder	HC25 type
Sockets	According to station fittings

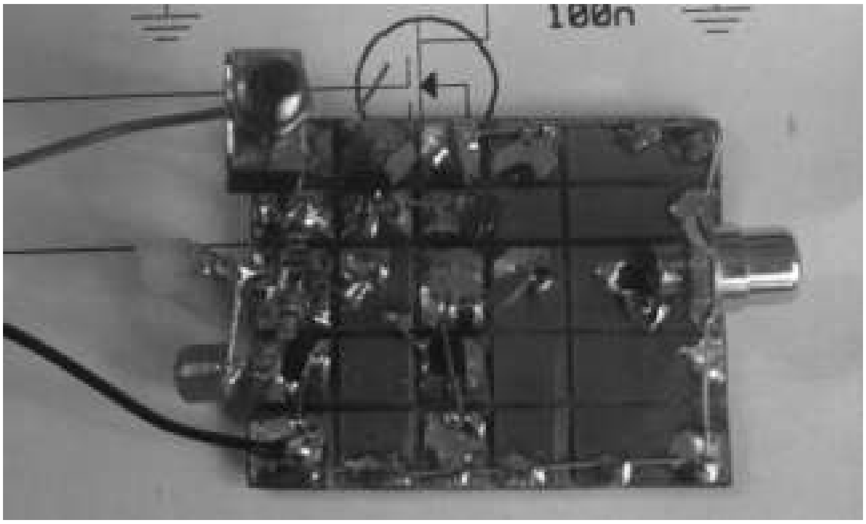

54 A 7-element low-pass filter for transmitters

Introduction

The simpler the transmitter, the more likely it is to radiate *harmonics* of its fundamental frequency. Harmonics are integral multiples of the frequency on which the transmitter is *designed* to operate. If you think you are transmitting on a frequency, *f*, for instance, you will also be radiating the harmonics of *2f*, *3f*, *4f*, . . . and so on. This results in your signals being heard on *several frequencies*, spread over a very wide frequency range. You are also contravening the terms of your licence. To avoid this, it is always advisable to use a *low-pass* filter between your transmitter and aerial. This is a filter which will *pass* your signal frequency, *f*, and all frequencies *below* it, but will *not pass* (attenuate) frequencies above *f* to any significant extent. The value of *f* is known as the *cutoff frequency* of the filter.

A design of 7-element low-pass filter

A 7-element low-pass filter (LPF) is so called because it has seven components, as the circuit diagram of **Figure 1** shows. Filters containing any odd number of elements are possible: a 3-element filter would

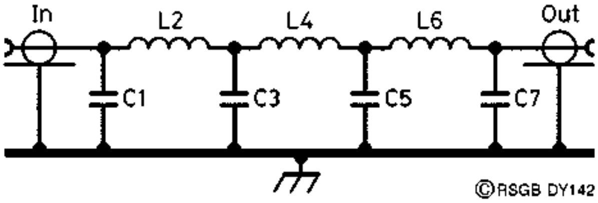

Figure 1 Circuit diagram of filter

comprise C1, L2 and C3 only, and is sometimes called a pi-network because the element disposition resembles the Greek letter pi (π); a 5-element filter would comprise C1, L2, C3, L4 and C5 only, and so on. In general, the more elements the filter has, the more effectively it attenuates signals above *f*.

The circuit of Figure 1 is designed to have an input and an output impedance of 50 Ω, which means that it can be placed in the aerial feed of any common transmitter. Filter design is a very complex business, and is best left to the experts. One such expert is W3NQN, who produced a number of computer designs of LPF using commonly available (*preferred value*) capacitors, and aimed specifically for use on amateur frequencies. The results of this work are condensed into **Table 1**. The inductors are wound on standard toroidal cores, and their details are included in the table.

Table 1 Filter component values for each brand

Band metres	*C1,7 pF*	*C3,5 pF*	*L2,6 turns*	*L4 turns*	*Core type*	*Wire SWG*
80	470	1200	25	27	T37-2	28
40	270	680	19	21	T37-2	26
30	270	560	19	20	T37-6	26
20	180	390	16	17	T37-6	24
15	82	220	12	14	T37-6	24
10	56	150	10	11	T37-6	22

Making the filter

The filter was made originally as an adjunct to the *Breadboard 80 m CW transmitter*, which you will also find in this book. It uses the same constructional technique, based on a single piece of plain, copper-clad

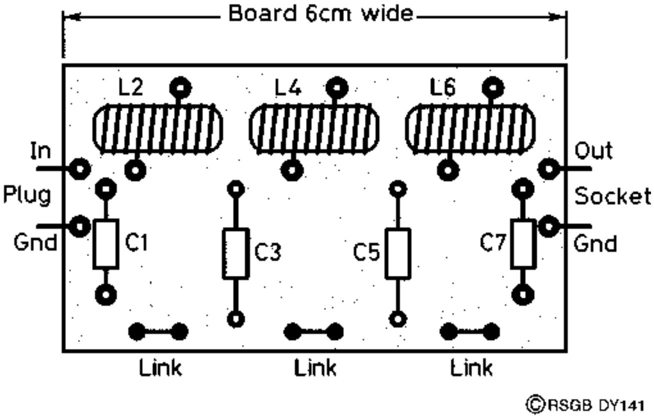

Figure 2 Layout and construction of filter

PCB, with 'pads' created by using a sharp blade in a junior hacksaw. The cuts in the copper are shown in the layout diagram of **Figure 2**. There are two ways of mounting the components: the first way is to drill small holes in each pad, as shown in Figure 2, and mount the components through the holes in the normal PCB manner; the second way is to solder the components directly to the pads, in the way that was described for the *Breadboard transmitter.*

Winding the inductors is quite simple. All you need to remember is that each time the wire passes through the core counts as one turn. Cut off the spare wire at each end of each coil to about 1 cm, scrape off the enamel with sandpaper, and tin the exposed copper. See the transmitter description if you are unsure of how to do this. Note the wire links between each of the lower pads, forming a solid 'ground' for the elements. The prototype had a plug and socket on the ends, to match the transmitter and aerial terminations.

The type of capacitor used in the design is not critical; the polystyrene type works well.

55 Radio-frequency mixing explained

Introduction

Mixers find widespread use in electronic circuitry. Many of the projects in this book, together with every TV set and radio in the home, contain mixer circuits – a good indication of their usefulness.

Confused?

Audio mixers (as used in recording studios and radio broadcast stations) are used to **add** or 'balance' the signals from various sources such as microphones, CD players, etc. These have nothing whatsoever to do with radio-frequency (RF) mixers, and should never be confused with them.

RF mixers and beat frequencies

Instead of **adding** signals (as in the audio mixer), the RF mixer **multiples them together.** As you might expect, this has an entirely different effect. The two signals entering the mixer *beat* or *heterodyne* with each other to produce *signals on other frequencies*. One example of this occurs in sound, when two musical notes of almost the same frequency are heard together. Instead of hearing two separate notes, the listener hears one note whose *intensity* (*loudness*) appears to increase and decrease. This intensity variation is called a *beat*, and its frequency is equal to the difference in the frequencies of the two original notes. The technique is used by musicians to tune their instruments. If one note is known to be a correct frequency, the other can be tuned to it by making the beat frequency as close to zero as is possible.

Multiplying together

The process of mixing presupposes that we have a device which will automatically multiply two signals together. Fortunately, this is easy; so easy, in fact, that it often occurs when we do not want it! Multiplying is achieved by any device which is *non-linear*; this means a device whose output is not a constant factor larger than its input, something that can be achieved by many electronic devices and circuits.

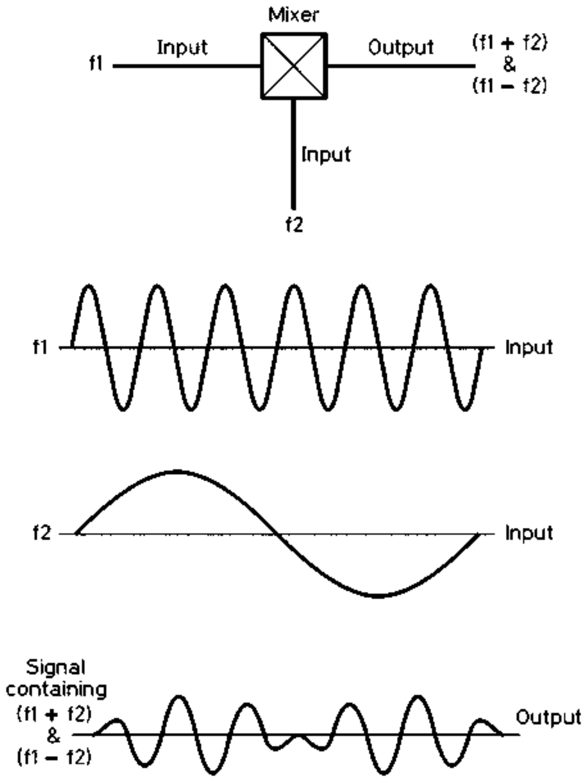

C RSGB DY145

Figure 1 The effect of multiplying (or mixing) two signals together

Let us look now at what a mixer does in concrete terms. Suppose two signals, of frequencies *f*1 and *f*2 go into our mixer. These signals are shown in **Figure 1**. Putting numbers in, to make the situation clearer, suppose *f*1 is 1.000 MHz and *f*2 is 160 kHz. The beat frequency is the *difference* of these: 1.000 MHz – 0.160 MHz = 0.840 MHz, or 840 kHz. A mixer *also* produces an output at the *sum* of these frequencies; in this case the new frequency would be 1.000 MHz + 0.160 MHz = 1.160 MHz.

Suppose you fed the output of your mixer, operating with these input frequencies, into a receiver and tuned around to find what frequencies were present. You would find two signals, one at 840 kHz and one at 1.160 MHz, showing that the two ‘new’ frequencies were very real!

In addition to drawing out the waveform of the resultant signal, as in Figure 1, we can draw the inputs and outputs on a frequency axis, to form a *spectrum* of the signal components. This is done in **Figure 2**. The top two diagrams show the input signals at *f*1 and *f*2. The bottom diagram shows the output signals in relation to the input signals. Depending on the type of mixer used, one or both of the input signals would be removed.

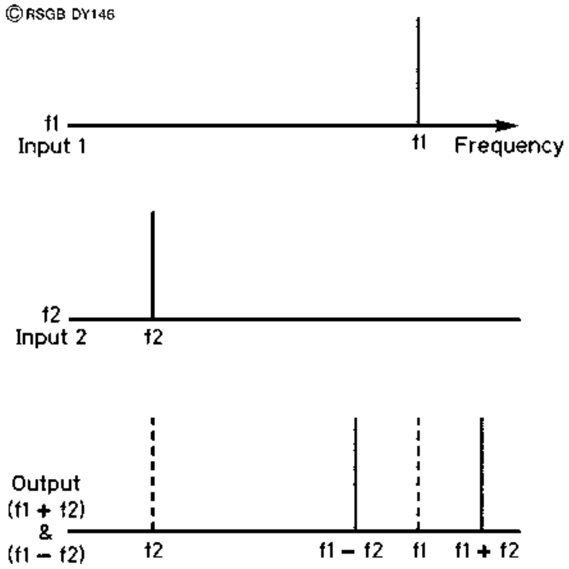

Figure 2 The result of multiplying or mixing two signals together as seen on a spectrum analyser

A mixer in every radio

Basically, a mixer is used to change a signal from one frequency to another, something it does without altering the characteristics of the incoming signal. If the incoming signal is amplitude modulated (AM), then the frequency-changed signal would be AM also. The same applies to FM, SSB, CW, and all other modulation forms you can think of. This explains why mixers are often called *frequency changers*.

Frequency changing is the key process in the type of radio known as a *superheterodyne* (or *superhet*). By mixing the incoming signal with a variable-frequency *local oscillator* as **Figure 3**, shows, the signal can be converted to the fixed frequency of a filter and amplifier. This is useful

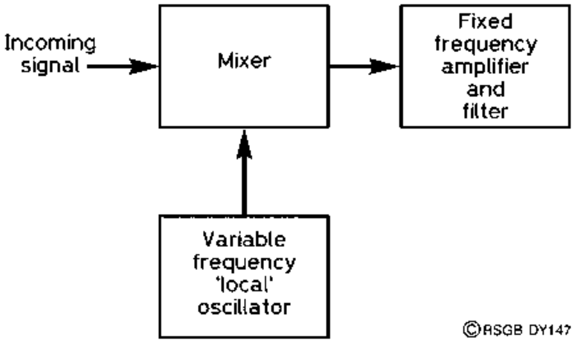
Incoming signal
Mixer
Fixed frequency amplifier and filter
Variable frequency 'local' oscillator
©RSGB DY147

Figure 3 The basic idea of a superhet receiver

because it is easier to make a very high-quality filter and amplifier at a single fixed frequency, than at a variable frequency.

All TV receivers and virtually all radio receivers (and transmitters) use mixers. Both the *Yearling* and *Colt* receivers (see the relevant projects) are superhets, and use mixers.

56 A voltage monitor for a 12 V power supply

Introduction

If for any reason, the stabilisation of your main 12 volt power supply unit (PSU) breaks down, it is possible that a voltage much higher then the nominal 13.8 V will be applied to your precious equipment. If you would like to know the instant that this occurred, and hence be able to switch things off before it was too late, then this circuit is what you need. It will give audible *and* visual indications if the voltage rises above 14.4 V, and a visual indication only if the voltage is reduced.

The circuit

The circuit uses three ICs and is shown in **Figure 1**. The circuit is powered by the PSU whose output is being monitored, and the circuit's immunity to supply line variations is secured by the 6 volt regulator, IC1. The heart of the circuit is IC2, an LM3914; it is a *bargraph driver*, which operates ten LEDs in a display resembling a thermometer – the string of lit LEDs increases in length as the voltage on pin 5 increases.

The input voltage range on pin 5 is 1.2 V maximum, making each LED correspond to one-tenth of this, which is 0.12 V, the *step size*. R1 and R2 act as a voltage divider, so that voltages of up to the maximum of 14.4 V may be applied to R1 without exceeding 1.2 V at pin 5. R3 sets the brightness of the LEDs and R4 determines the step size.

IC3 is an *opto-isolator*, a device containing an LED and a phototransistor in one package. This enables the piezoelectric sounder to operate without affecting the operation of the bargraph driver. The input to IC3 is provided by the voltage on D8, so that if any of the LEDs at or above D8 are lit, the

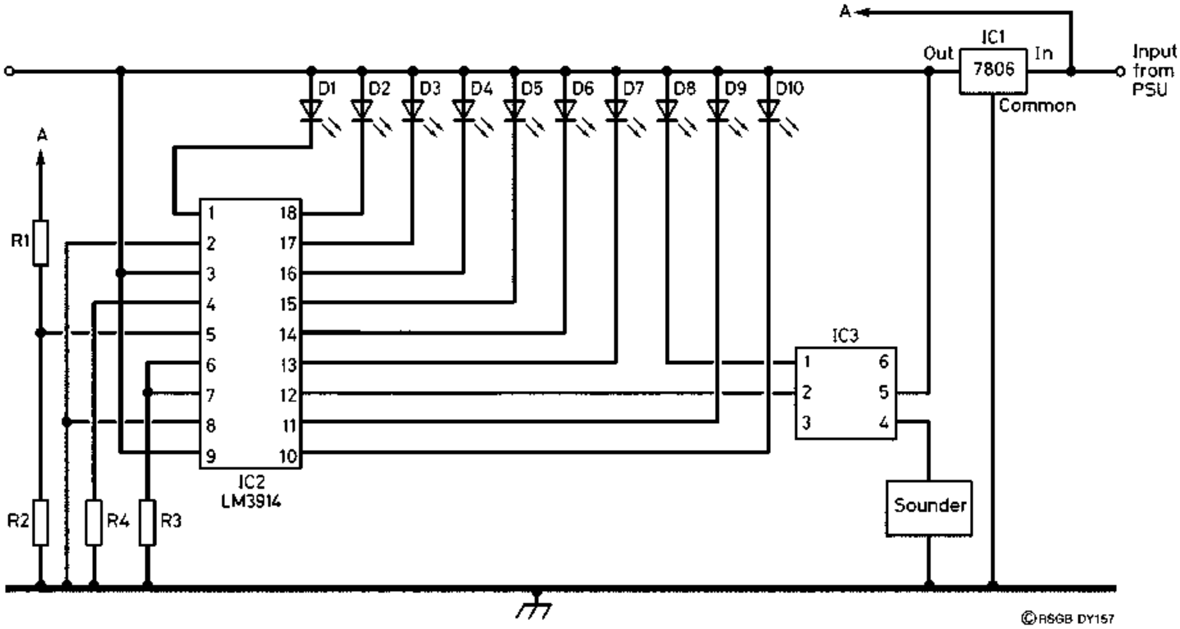

Figure 1 PSU monitor, circuit diagram

sounder will operate, indicating overvoltage. Despite the fact that some of your other equipment might be damaged by this overvoltage, the monitor circuit itself is unaffected.

Construction

The prototype was built on two pieces of Veroboard of the copper-strip variety. The main (circuit) board measured 15 strips by 25 holes, and the display (LED) board measured 4 strips by 30 holes. Track cuts are necessary in each board. The correct places are shown for the main board in **Figure 2**, but there is sufficient flexibility in the layout of the display board for a prescriptive layout not to be needed. All the anodes of the LEDs are connected to the same strip, which makes things comparatively simple!

For the main board, insert the Veropins and the wire links first, and solder them to the copper strips. Then fit the IC holders and the resistors. Fit the IC holders with their notches towards the top of the board. When fitting the voltage regulator, IC1, note than the centre lead (the ‘common’ lead in Figure 1) does *not* go to position A3; the track should be cut at A3, and the common lead soldered at B3. This is indicated in Figure 2.

Wire up the LED board with its ten LEDs and 11 connecting wires, each about 10 cm long. This length depends on how far away from the main

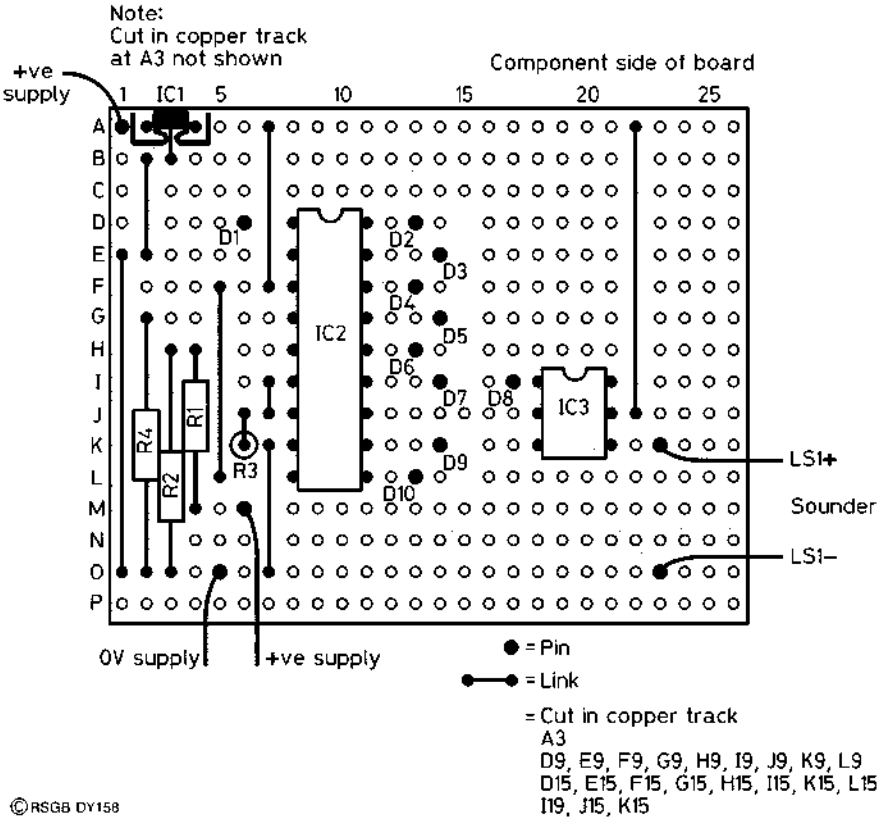
Note:
Cut in copper track
at A3 not shown
+ve supply
Component side of board
1
IC1
5
10
15
20
25
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
IC2
IC3
R1
R2
R3
R4
LS1+
Sounder
LS1−
0V supply
+ve supply
= Pin
= Link
= Cut in copper track
A3
D9, E9, F9, G9, H9, I9, J9, K9, L9
D15, E15, F15, G15, H15, I15, K15, L15
I19, J15, K15
©RSGB DY158

Figure 2 PSU monitor, PCB layout

board you are planning to mount the display. The LEDs should have different colours: three orange, four green and three red, to indicate ‘low’, ‘medium’ and ‘high’ voltage.

Now solder the other ends of these leads to the Veropins on the main board, making sure that the order is correct. Finally, connect the piezoelectric sounder to the main board; the polarity is important, so make sure the red lead goes to pin 4 of IC3 and the black lead to the ground rail. If you were careful to fit the IC holders with their notches in the correct positions, match these up with the notches on the ICs before pushing home the ICs *gently*. Check your circuit for solder bridges and unwanted pieces of copper swarf before screwing the small heat sink to the voltage regulator, IC1.

Testing

(a) **With a variable-voltage PSU**. Set the PSU for 10 V, connect the circuit and switch on. Increase the voltage slowly, and check that each LED lights up after the one before it. As the voltage exceeds about 14.4 V, the first red LED should light and the sounder should operate. If an LED does not illuminate, you should immediately suspect either a dry joint or an incorrect LED polarity. The voltage at which the first red LED lights can be adjusted by varying R1; increase R1 if the LED comes on too early; decrease R1 if the LED comes on too late.

(b) **Without a variable-voltage PSU**. For your ‘variable supply’, you can use several AA-type 1.5 V cells (or 1.2 V NiCad cells) in series. The voltages produced by a range of cells is shown in **Table 1** – because of the lower voltage of NiCad cells, more of them are needed to produce a given voltage.

Once the operation of the circuit has been checked, it can be fitted into a plastic or metal box. Only two connections are needed for the PSU. You may want to drill some holes in the case to increase the apparent loudness of the sounder.

Table 1 Test voltages available from batteries in series

No. of batteries	*Voltage, nicads*	*Voltage, dry cells*
7	8.4	10.5
8	9.6	12
9	10.8	13.5
10	12	15
11	13.2	–
12	14.4	–
13	15.6	–

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	11 kilohms (kΩ) – see text
R2	1 kilohm (kΩ)
R3	1.2 kilohms (kΩ)
R4	18 kilohms (kΩ)

Semiconductors

IC1	L7806
IC2	LM3914
IC3	Opto-isolator – Maplin code WL35Q
D1–D3	3 mm LED, orange
D4–D7	3 mm LED, green
D8–D10	3 mm LED, red

Additional items

LS1	Piezoelectric sounder, wire leads
	6-pin DIL socket for IC3
	18-pin DIL socket for IC2
	Veroboard – two pieces for main and display boards, see text for sizes
	Veropins
	Heat sink for IC1
	Single-core insulated wire for links
	Insulated stranded wire for interconnecting the boards
	Case as required

57 A 1750 Hz toneburst for repeater access

Introduction

Repeaters across the UK and much of Europe need an access tone to switch the transmitter from standby ready for use. Commonly, this is a 1750 Hz tone of duration no greater than half a second. Although many UK repeaters may now be accessed using the *continuous tone-coded squelch system* (CTCSS – see the *RSGB Yearbook*), you may wish to access a repeater whose CTCSS frequency you don't know; in this case, using the universal 1750 Hz tone will gain you access. Commercial transceivers are usually fitted with an automatic toneburst, but if you are using a home-made design, then you may want to incorporate this little circuit.

Warning

This circuit uses a member of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current

and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from destruction while it is connected to the battery.

Circuit description

The circuit is shown in **Figure 1**. The tone is generated by an integrated circuit oscillator (IC1), whose frequency is controlled by a *ceramic resonator*, XL1. Its frequency is very high, and is divided down to the 1750 Hz needed by the same chip. The ceramic resonator is designed to operate at 455 kHz, the intermediate frequency of many receivers. Because all divider circuits use powers of 2, we need the oscillator to run at 448 kHz so that when it is divided by 256 ($256 = 2^8$), we end up with 1750 Hz. Try it on your calculator:

$$\frac{448\ 000}{256} = 1750.$$

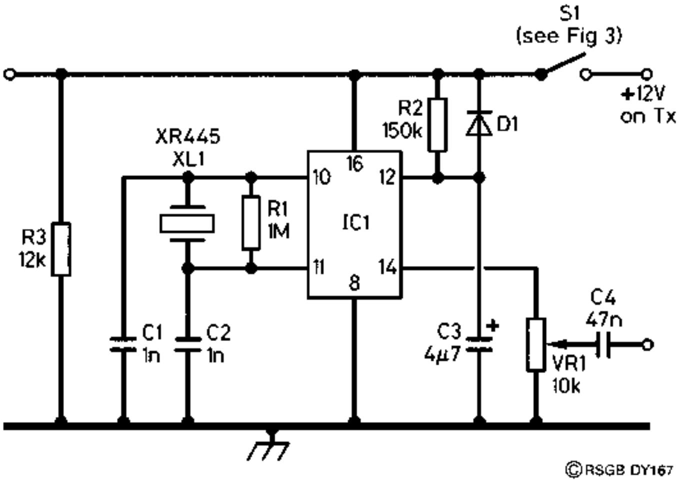

Figure 1 Toneburst module, circuit diagram

We use C1, C2 and R1 to *pull* the frequency of the oscillator away from 455 kHz to 448 kHz. The divider chain has eight *counters* in it, and each counter divides the frequency of the signal it sees by two, giving the final division of 256.

This counting process can be stopped at any time by taking the voltage on the *reset* pin (pin 12) up to the supply voltage. When the circuit is switched on by closing S1, pin 12 is at 0 V because C3 is discharged. The oscillator runs, producing the output frequency of 1750 Hz. As time progresses, C3 charges up through R2 and the voltage on pin 12 rises. When this has risen sufficiently, and in a time determined by the values of C3 and R2, the counter resets and stays in the reset state; no division takes place and there is no output. The duration of the toneburst is thus governed by C3 and R2.

When S1 is opened, the circuit is switched off, and C3 is discharged through D1 and R3, ready for the next toneburst. If you have used a repeater, you will know that a toneburst is needed **only** to activate a repeater in the standby condition; it is not needed once a contact has been established.

VR1 adjusts the amplitude of the tone fed to the microphone, and C4 prevents any voltage that may be present on your microphone connector from damaging the integrated circuit.

Construction

The prototype circuit was built on Veroboard of the copper-strip variety, measuring 20 holes by 14 strips. The layout is shown in **Figure 2**. Make the track cuts first, and check that there are no slivers of copper wedged between adjacent tracks. Then, solder in the IC socket (with the notched end

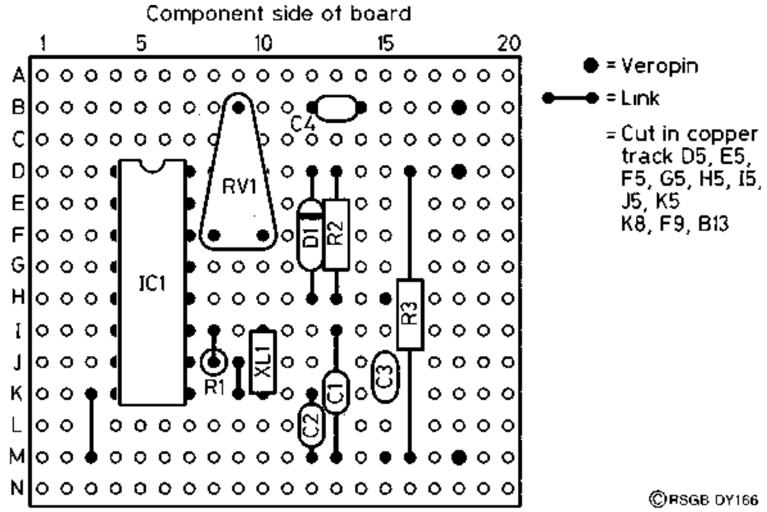

Figure 2 Veroboard component layout

facing towards track A), the wire links and the three Veropins. Having done this, solder in the resistors, capacitors and diode, making sure that D1 and C3 are the right way round! Using your best soldering technique, solder in the ceramic resonator quickly, to prevent heat damage. Recheck your circuit, check for solder splashes and bridges, and then gently insert IC1 into its socket, matching up its notch with that of the socket.

Testing

Set VR1 to half-way and connect a crystal earpiece to the output; apply power to the circuit. You should hear the tone, lasting for about half a second. If there is no tone, disconnect your circuit from the power supply, and check for dry joints in the vicinity of pin 12. Is the diode, D1, the correct way round? Is C3 the correct way round? Did you choose to ignore the CMOS safety precautions given earlier?

Once the circuit is working, you need to decide how you are going to connect it to your transmitter. Two options are shown in **Figure 3**. If you have access to a point in your transmitter circuit that has between 9 V and 12 V positive on it during transmit, you can use this to power your circuit. As the toneburst is needed only for repeaters, the switch, S1, disconnects it when not needed, as shown in Figure 3a. If you want the circuit to be self-powered, then a 9 volt PP3 battery may be used; Figure 3b shows this configuration.

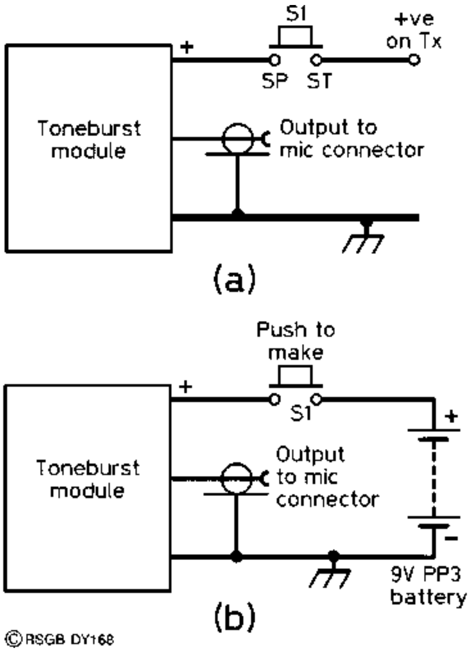

Figure 3 Toneburst module: two alternative switching arrangements

The output from the circuit board is fed directly into the microphone socket, in parallel with the microphone itself; use thin coaxial or screened cable for this lead, or you may induce hum into the microphone circuit and suffer from RF breakthrough into the audio circuits. To adjust the setting of VR1, start with it at the zero output position and connect a dummy load to your transmitter. Slowly, increase the output while monitoring your transmitted signal on another nearby receiver. Make sure you do **not** increase the output so far that the signal sounds distorted. If you would prefer that the tone was on continuously while you made this adjustment, simply connect a wire across C3 remembering, of course, to remove it as soon as you have completed the test!

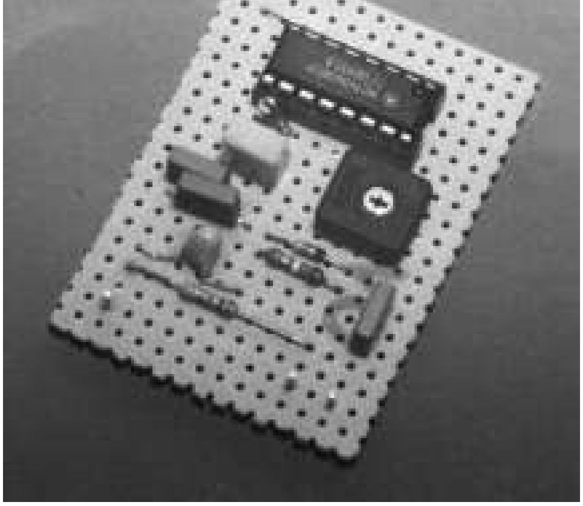

Parts list

Resistors: all 0.25 watt, 10% tolerance (or better)

R1	1 megohm (MΩ)
R2	150 kilohms (kΩ)
R3	12 kilohms (kΩ)
VR1	10 kilohms (kΩ) horizontal preset

Capacitors: all 16 V WKG or higher

C1, C2	1 nanofarad (nF) disc ceramic
C3	4.7 microfarad (μF) tantalum bead
C4	47 nanofarads (nF) disc ceramic

Semiconductors	
IC1	4060
D1	1N4148
Additional items	
XL1	XR455
	Veroboard (see text for size)
	Veropins (3)
S1	Switch (momentary action push-to-make SPST)
	16-pin DIL socket for IC1
	Single-core insulated wire for links
	Coaxial or screened cable for microphone connection

58 A circuit for flashing LEDs

Introduction

There are many occasions when one's attention needs drawing to the fact that something important has happened. A single red light coming on is seldom sufficient to attract attention, particularly if it is surrounded by other lights and indicators. The eye is known to be very sensitive to *changes* in its peripheral vision; such changes can be brought about by movement or by differences in light level – a flashing light, for example. So, a circuit that flashes a single LED or a pair of LEDs finds plenty of uses in the amateur station.

Warning

This circuit uses a member of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton

destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from destruction while it is connected to the battery. However, when the battery is removed, the same care should be exercised with its handling, because there is no supply decoupling capacitor across the IC.

Basic description

LEDs can be made to flash (switch on and off) by driving them from sources that switch on and off. Such a source is an *astable multivibrator*. If you have built or read about *A basic continuity tester*, elsewhere in this book, you will have come across such a beast before. That circuit used an astable multivibrator made from two transistors. This new circuit achieves the same behaviour from a single integrated circuit, the CMOS 4011. To give it its full description, the 4011 is a *quad 2-input NAND gate*. Quite a mouthful, but all it means is that inside the chip are four NAND gates, each with two inputs.

A NAND gate needs a positive voltage (known as a *logic 1*) on *both* inputs in order to produce zero volts (known as *logic 0*) at the output. Two NAND gates can be connected, as are A and B in **Figure 1**, to make our astable multivibrator. The combination of A and B has been described as the most perverse circuit in electronics; as soon as the output goes to logic 1, the circuit decides that it would prefer to have a logic 0 there, and switches over. With logic 0 at the output, the circuit now prefers to have logic 1 there, and so it goes on! We are going to use this continuous switching backwards and forwards to flash two LEDs. The rate at which A and B 'change their minds' is the frequency at which our LEDs will flash, and is controlled by the charging and discharging times of C1 through R2 and C2 through R1. As the values of R1 and R2 are the same, and those of C1 and C2 are the same, the ON and OFF states of the circuit are the same.

Gates C and D do not contribute to the flashing action; they act as *buffers* to isolate the LEDs from the multivibrator circuit itself. You will find in electronics that an oscillator is seldom used to drive another device *directly*; there is usually a buffer between it and the stage it drives.

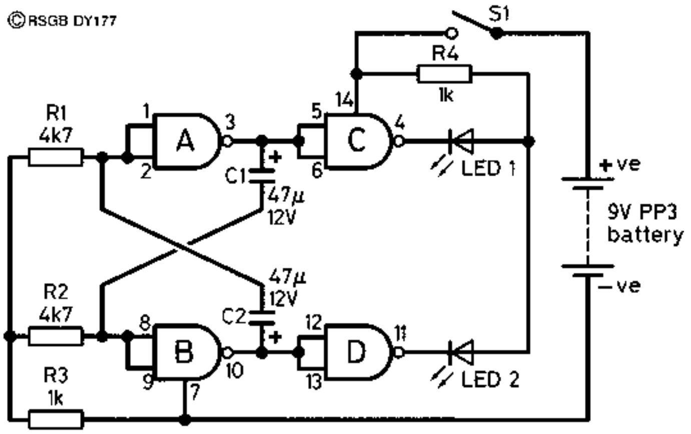

Figure 1 Flashing LEDs, circuit diagram

A characteristic of all multivibrators, astable or not, is that they have two outputs. In this case, those outputs are at pins 3 and 10, which are then buffered and appear at pins 4 and 11, respectively. When one output is at logic 1, the other is at logic 0, and vice versa. This means that LED1 is off when LED2 is on, and LED1 is on when LED2 is off, the two states switching backwards and forwards at the frequency of the oscillator.

Construction

Read the warning at the beginning of this article again. It is not intended to scare you off from building this, but is a genuine piece of advice which can save you time and irritation when all your labours result in a circuit that doesn't work! That extra bit of care can make all the difference!

Veroboard (the copper strip type) is used for the layout, shown in **Figure 2**. It measures 20 holes by 12 strips. **Be aware that there is no row 'I' in the layout, so don't miscount when you are placing components on the board!**

Firstly, cut the tracks using a 3 mm (1⁄8 inch) twist drill held between thumb and forefinger; check that there are no slivers of copper bridging any of the tracks, and that the tracks have been completely cut by the drill. Solder in the components carefully. Leave the IC in its carrier for the time being, and solder in the IC socket, with the notched end facing row A. On completion, check the circuit carefully. If you are happy that it is correct, follow the instructions given earlier and fit IC1 into its socket, matching up the two notches. Connect the battery and switch on. The two LEDs should flash on and off alternately. If only one LED flashes, you have probably connected the other one the wrong way round. Switch off, check and correct if necessary. If neither LED flashes, you must have a significant error in your circuit, which will need checking *again*! Or did you choose to ignore the handling precautions for the CMOS chip?

Figure 2 Flashing LEDs, component layout

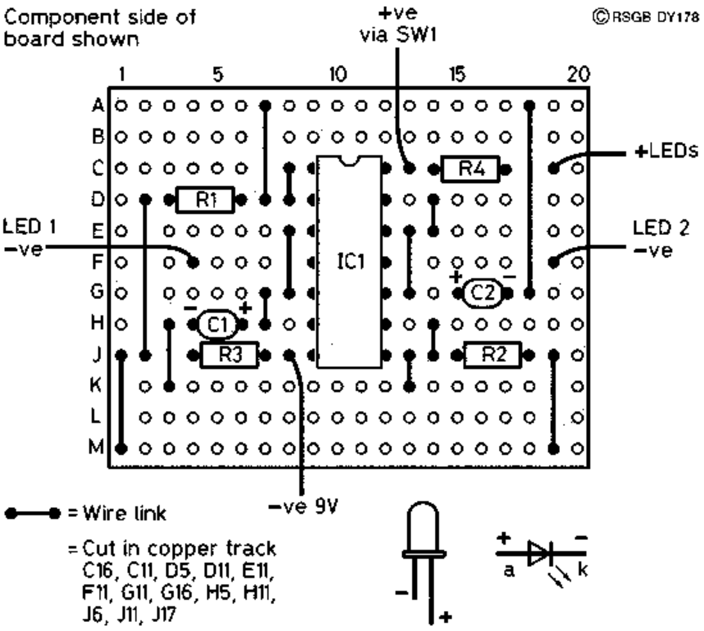
Component side of board shown
+ve via SW1
©RSGB DY178
1
5
10
15
20
A
B
C
D
E
F
G
H
J
K
L
M
R1
R4
IC1
C1
C2
R3
R2
+LEDs
LED 1 −ve
LED 2 −ve
−ve 9V
= Wire link
= Cut in copper track C16, C11, D5, D11, E11, F11, G11, G16, H5, H11, J6, J11, J17
−
+
a
k

The board can be mounted near to the point where you want your flashing LEDs to be seen, although long leads to the LEDs are acceptable. The LEDs can be different colours – it's all up to you now!

Parts list

Resistors: all 0.25 watt, 10% tolerance or better	
R1, R2	4700 ohms (Ω)
R3, R4	1 kilohm (kΩ)
Capacitors	
C1, C2	47 microfarads (μF) 12 V WKG
Semiconductors	
IC1	4011
Additional items	
LED1, LED2	Any size of LED, any colour
S1	SPST on/off
	Plastic box if needed, 8.5 by 5 by 2.5 cm

Source

Components are available from Maplin.

59 Digital logic circuits

Introduction

Logic circuits form the backbone of even the most advanced computer, yet their basic operation can be demonstrated by a couple of switches, a battery and a bulb.

Logic using switches

Everyone reading this article will look at **Figure 1** and know immediately how it works and be able to write down something like 'When switch A and switch B are closed, the light will come on'. Without knowing it, you have written down a logic statement involving the so-called **AND** operation; the light comes on only when switches A **AND** B are ON. Below the circuit in Figure 1, is a table showing the only possible positions of the two switches and the state of the bulb for each position. This is called a *truth table*, and is frequently used in logic analysis.

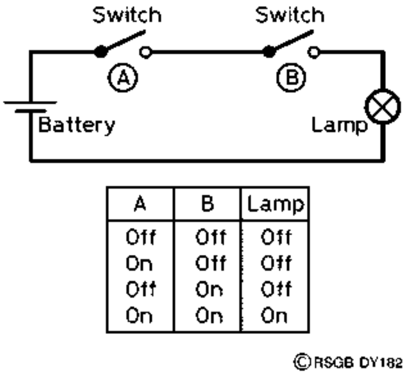

A	B	Lamp
Off	Off	Off
On	Off	Off
Off	On	Off
On	On	On

Figure 1 Switches and lamp AND gate

Figure 2 shows a different circuit. Here, the two switches are in parallel rather than in series, as was the case in Figure 1. Again, if you analyse the circuit in words, you would say that the light will be on when switch A **OR** switch B is ON. This is an example of the **OR** operation, and its truth table is shown in Figure 2. The statement above is not complete, however; can you see why? The truth table will show you. The light comes on if A is ON, **OR** if B is ON, ***OR*** if A **AND** B are both ON. That third condition is easy to miss, but don't worry about it!

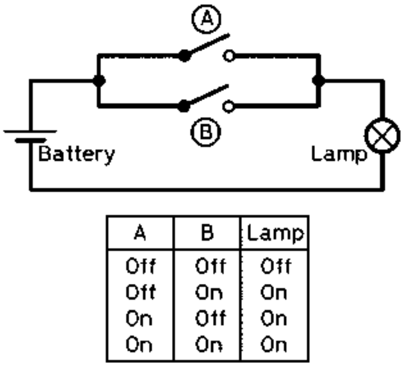

A	B	Lamp
Off	Off	Off
Off	On	On
On	Off	On
On	On	On

©RSGB DY184

Figure 2 Switches and lamp OR gate

Believe it or not, some very complicated logic is possible (in theory) using switches and lights, but it is highly impractical and would be very slow. This is where electronic logic circuits come in.

Switches with no moving parts

You may have come across projects in this book where a statement is made such as '. . . the transistor is being used as a switch . . .'. Transistors *can* be used as switches, as were thermionic valves in the world's first programmable computer *Colossus*, at Bletchley Park. However, technology has moved on from valves, through transistors to *logic gates*, combinations of electronic switches designed specifically to perform logic functions.

These act on voltage levels as their inputs and produce changes in voltage levels as their outputs. A positive voltage is called *logic 1*, and corresponds to a switch being ON in our previous descriptions; a zero voltage is called *logic 0*, and corresponds to a switch being OFF. The output from a logic gate (normally labelled Q) is also logic 1 or logic 0, corresponding to our light being ON or OFF, respectively, in our switch analogy.

Many logic devices operate from a stabilised 5 V supply, and this determines the *ideal* voltages corresponding to the two logic states:

logic 0 ≡ 0 V,

logic 1 ≡ 5 V.

The world isn't an ideal place, so the *real* voltage ranges used by the logic gates are:

logic 0 ≡ 0.0 to 0.4 V,

logic 1 ≡ 3.0 to 5.0 V.

The **AND** circuit of Figure 1 is now called an **AND gate**, and requires logic 1 inputs on A **AND** B to produce a logic 1 at the output. **Figure 3** shows this. The truth table is identical with that of Figure 1 – logic 1 replaces ON and logic 0 replaces OFF. Now compare Figure 2 with **Figure 4** – circuits and truth tables for the **OR** function. Again, we have exact similarity.

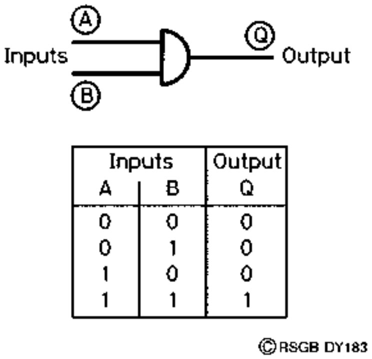

Inputs		Output
A	B	Q
0	0	0
0	1	0
1	0	0
1	1	1

©RSGB DY183

Figure 3 Electronic AND gate

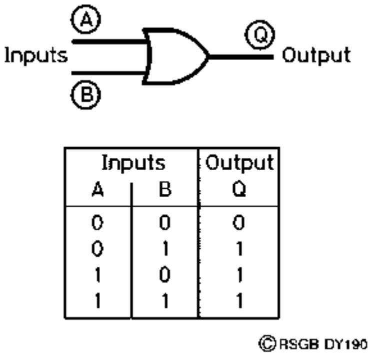

Inputs		Output
A	B	Q
0	0	0
0	1	1
1	0	1
1	1	1

©RSGB DY190

Figure 4 Electronic OR gate

There is another very common logic gate, which performs the **NOT** function. It is easy to understand. Just ask yourself the question 'What is **NOT** logic 0?', and the answer is obviously 'logic 1'. Similarly, logic 0 is **NOT** logic 1. A **NOT** gate simply changes the logic state of the input; it is also known (because of this behaviour) as an *inverter*. Its symbol and truth table can be found in **Fig 5**. Note the little circle on the output of the gate in Figure 5. In logic circuits, this symbol always implies inversion, or the presence of a **NOT** gate. Keep an eye open for it!

So far, the logic functions we have discussed have all been words which we use in everyday language, which has made the electronic interpretation of them relatively easy. Now we must introduce a function for which there is *no* analogy in normal speech – the **NAND** function. This means a

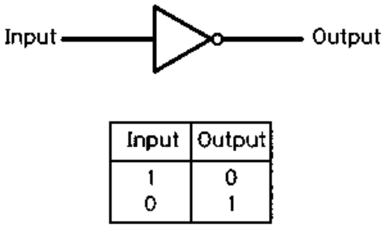

Input	Output
1	0
0	1

©RSGB DY185

Figure 5 Electronic NOT gate

combination of an **AND** gate and a **NOT** gate, and the sharp-eyed reader will have spotted the little circle added on to the normal **AND** gate symbol in **Figure 6**!

To make things easier to understand, the truth table in Figure 6 has four columns, not three as in previous tables. The third column is the standard **AND** output – compare it with the third column in Figure 3. That is the output from the **AND** gate *before* it encounters the little circle that inverts it, so the final output from the **NAND** gate is the output of the **AND** gate, inverted! The third and fourth columns are the inverse of each other.

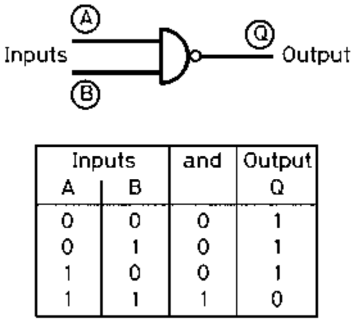

Inputs		and	Output
A	B		Q
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

©RSGB DY186

Figure 6 Electronic NAND gate

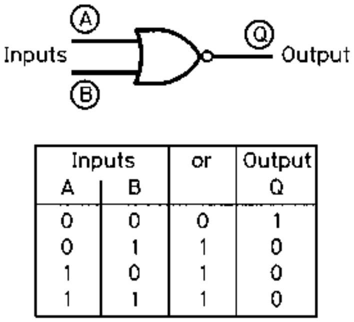

Inputs		or	Output
A	B		Q
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

©RSGB DY187

Figure 7 Electronic NOR gate

We can add the inverting operation to the output of an **OR** gate also, producing a **NOR** gate! The symbol and its truth table are shown in **Figure 7**. Again, notice that the final output is that of the ordinary **OR** gate, inverted!

These new functions of **NAND** and **NOR** are used more than the **AND** and **OR** functions because it makes other circuits easier to design using combinations of these gates.

A taste of Boolean algebra

The design of circuits using combinations of logic gates usually begins with a little mathematics, where the functions to be implemented are analysed. The mathematics used is surprisingly simple, and is a slightly changed version of ordinary algebra called *Boolean algebra*, which allows manipulation of logic functions to be made. Normal algebra has operations in it such as addition, subtraction and multiplication and division. The mathematician Boole found that the logical **AND** operation could be handled by the algebraic operation of *multiplication* (symbols × or •), and the **OR** operation by the algebraic operation of *addition* (symbol +). The **NOT** operation involved a new symbol, that of a bar over the input being inverted, such as $\overline{A}$.

So, our five basic logic operators can now be written in a mathematical form:

AND $Q = A \times B$

OR $Q = A + B$

NOT $Q = \overline{A}$

NAND $Q = \overline{A \times B}$

NOR $Q = \overline{A + B}$.

Using logic operations in a mathematical form enables the most complex logic to be designed, simplified and converted into circuit diagram form in a very efficient and rapid way. There are more logic operators than the five we have considered here but, in general, they can all be broken down into combinations of these gates alone!

60 A resistive SWR indicator

Introduction

When a transmitter produces some output power, we want to make sure that as much as possible of this power is radiated by the aerial. This often requires the use of an *aerial tuning unit* (ATU), which matches the aerial impedance to that of the transmitter.

How do you know when this matching has been achieved? The most usual way is to use a *standing-wave ratio* (SWR) indicator. If the impedance of the aerial does not match that of the transmitter output, some of your transmitter power (also known as the *forward* power) is reflected back along the aerial feeder and back into the transmitter, where it causes excess heating. The forward and reflected waves interact along the feeder to produce a wave whose position remains constant, and which is therefore called a *standing wave* or a *stationary wave*. An SWR meter simply indicates forward power and reflected power, and adjustments are made to your ATU until the reflected power is as small as possible (ideally zero, of course). If there is no reflected power then, by a process of elimination, all your forward power is reaching the aerial!

Sampling the RF

Whatever type of SWR indicator you use, it must use some sort of sampling circuit to pick up the forward and reflected waves. The project *A standing-wave indicator for HF*, elsewhere in this book, uses a toroidal transformer to separate the readings for the forward and reverse waves. This design differs in that it measures the voltages across resistors through which the RF current is passing. Its advantages are:

(a) it uses cheap parts – four resistors, two capacitors and a diode, together with a rotary switch, a surplus meter, a preset potentiometer and two sockets;
(b) within this SWR indicator, there is always a resistive path for the RF current from the transmitter, formed by R1, R2 and R3; this can prevent damage to simpler home-made transmitters, which may be damaged during adjustment of the ATU when using more conventional SWR indicators.

The only disadvantage of this form of indicator is that it must be switched out of circuit once the ATU has been adjusted for a particular band.

Construction

The SWR indicator is very simple to build, as most of the components can be mounted on the back of the 3-way rotary switch. This is shown in **Figure 1**. The switch is a 4-pole, 3-way rotary type, of which only two poles are used.

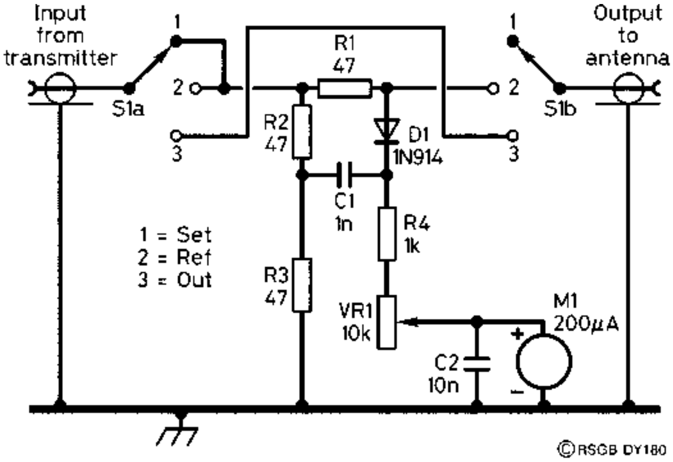

Figure 1 Resistive SWR meter, circuit diagram

Because the other switch contacts are not being used, they can be employed as support tags for other components. The ground wires are all soldered on to the metal frame of the switch. If your switch frame is of all-plastic construction, then a 12 SWG copper wire run around the switch will make a good earth connection to the metal case for the leads shown in Figure 1. The preset potentiometer used to control the sensitivity of the circuit can be mounted directly on the meter tag.

Resistors R1, R2 and R3 handle the RF power during the tuning-up process. If you have them, use 1 watt resistors; otherwise, you can use two 100 ohm half-watt resistors in parallel for each of R1, R2 and R3. The meter, M1, can be any DC type of sensitivity around 200 μA.

In use

First, find a clear frequency, and without the indicator in circuit, check that the frequency really *is* clear by asking and identifying yourself. If it is, connect the indicator between the transmitter and the ATU which, in turn, is connected to your aerial. Turn SW1 to the SET position and key the transmitter. Adjust VR1 until the meter reads full scale. Switch off the transmitter. Turn SW1 to the REF position and key the transmitter again. Adjust the ATU until the lowest reading is obtained on the meter. Switch off the transmitter. For the chosen frequency, you have adjusted your ATU for minimum reflected power and hence the lowest SWR. You will need to repeat the process when you change bands, and possibly when you change frequency within the same band. Switching SW1 to the OUT position, you are ready to transmit. You may have noticed that it is good practice to switch the transmitter off when operating SW1. Get into that habit!

Parts list

Resistors	
R1, R2, R3	47 ohm (Ω) 1 watt (or 2 × 100 Ω, ½ watt, see text)
R4	1 kilohm (kΩ) 0.25-watt
VR1	10 kilohms (kΩ) preset
Capacitors	
C1	1 nanofarad (nF)
C2	10 nanofarads (nF)
Semiconductors	
D1	1N914 or similar
Additional items	
SW1	4-pole 3-way rotary switch, of which only 2 poles are used
M1	200 μA DC
	Case

61 An audio filter for CW

Introduction

This is a simple passive circuit (it has no power supply) that adds some audio selectivity for Morse code reception and also includes a very simple noise limiter that gives a visual indication of when noise spikes are being removed!

The circuit

Figure 1 shows the complete circuit. The tuned circuit of C1 and L1 resonates very close to 800 Hz, so initially you will have to tune a signal in carefully until it sounds loudest – you will soon be able to do this without thinking. The two LEDs connected back to back across the signal path act as a noise limiter, reducing the amplitudes of static crashes and noise from car ignition systems, etc. The LEDs blink when they conduct – this is not necessary to the operation of the circuit, but adds a little colour to your listening! The noise limiter *does* make listening more comfortable, though.

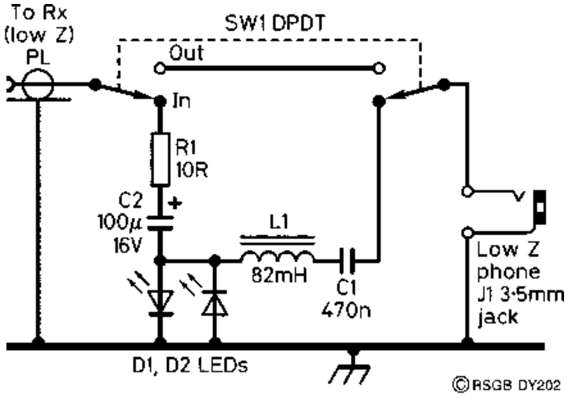

Figure 1 CW filter, circuit diagram

Construction

The circuit layout is shown in **Figure 2**. Point-to-point wiring is used, with a small tag-strip being the only item used for the extra support of C2 and R1. The LEDs and SW1 support the other components. An aluminium box (**Figure 3**) is used to make the circuit tidy and usable.

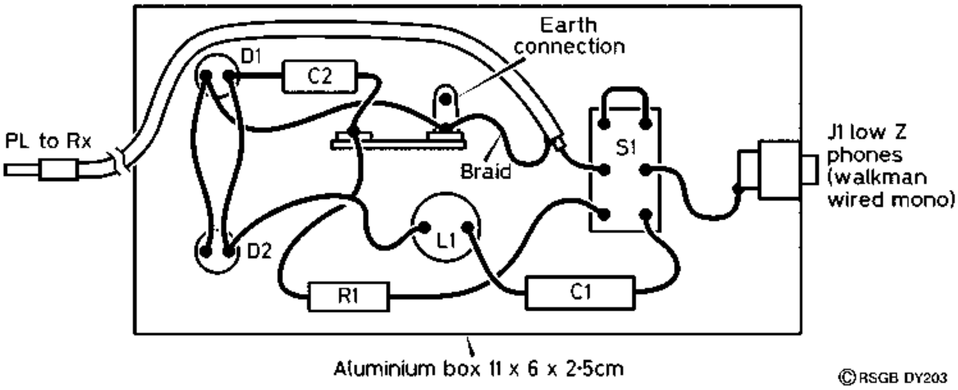
Earth connection
D1
C2
PL to Rx
S1
Braid
J1 low Z phones (walkman wired mono)
L1
D2
R1
C1
Aluminium box 11 x 6 x 2·5cm
©RSGB DY203

Figure 2 CW filter, component layout

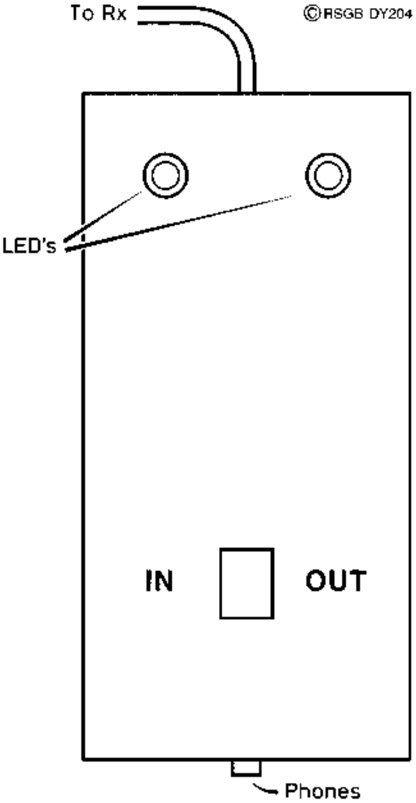
To Rx
©RSGB DY204
LED's
IN
OUT
Phones

Figure 3 CW filter, front panel

You must use low-impedance headphones for the circuit to perform properly. Plug the completed unit into the headphone socket of your receiver and adjust the volume so that the LEDs are just *not* blinking on normal audio. The unit should be switched out of circuit for speech listening; this is the purpose of the toggle switch, SW1.

Parts list

Resistor	
R1	10 ohms (Ω), 0.25 watt, 10% tolerance
Capacitors	
C1	470 nanofarads (nF)
C2	100 microfarads (μF) electrolytic, 16 V WKG
Inductor	
L1	82 millihenries (mH)
Semiconductors	
D1, D2	LEDs
Additional items	
SW1	DPDT
J1	3.5 mm jack socket
PL	Jack plug to suit receiver
	Aluminium box, approx. 11 by 6 by 2.5 cm

Source

Components are available from Maplin.

62 An electronic die

Introduction

Throwing a die is a venerable way of generating a ‘random’ number between 1 and 6. A die is easy to lose, so here is an electronic die-throwing circuit which brings the technique up to date and serves as another application of logic circuits. If you need a reminder of the basics of logic, refer to *Digital logic circuits*.

Warning

This circuit uses members of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the

magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains-earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from damage while it is connected to the battery.

Description

The circuit is shown in **Figure 1**. IC1a and IC1b form an *astable multivibrator*, similar to that used in 'An LED Flasher' also in this book. It runs constantly, and its output is fed to IC2 via a single NAND gate, IC1c. The other input of IC1c is held at 0 V by R3 and C3. Whenever one input of a NAND gate is logic 0 (or 0 V), there can be *no output* from the gate, irrespective of what is happening at other inputs. So, despite the fact that the oscillator (or *clock*) is running all the time, its square-wave output never

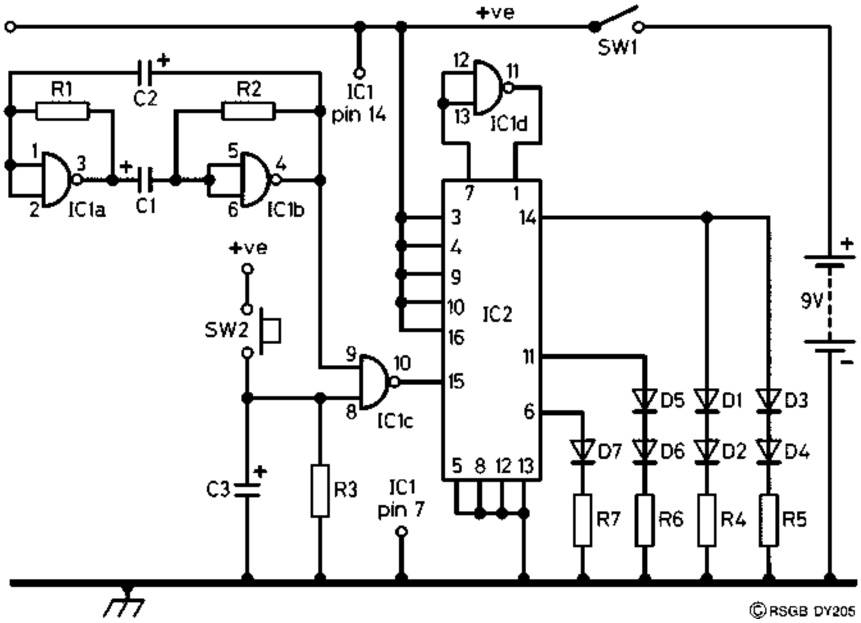
+ve
SW1
R1
C2
R2
IC1 pin 14
12
11
13
IC1d
1
3
2
IC1a
C1
5
4
6
IC1b
7
1
3
14
4
9
10
16
IC2
11
6
5 8 12 13
+ve
SW2
9
10
15
8
IC1c
C3
R3
IC1 pin 7
D5
D1
D3
D7
D6
D2
D4
R7
R6
R4
R5
9V
©RSGB DY205

Figure 1 Electronic dice, circuit diagram

reaches IC2 until the ‘other’ input of IC1c is brought away from logic 0. This happens when SW2 is pressed. Pressing SW2 momentarily, as is normally done, charges up C3 to 9 V, putting a logic 1 on the second input of IC1c and allowing the clock signal through to IC2. As the switch is released, C3 begins to discharge through R3, gradually lowering the voltage on pin 8 of IC1c. As this voltage crosses 4.5 V (half the supply voltage), IC1c then treats this input as logic 0, which cuts off the clock signal from IC2 again. Thus, when SW2 is momentarily pressed, the clock signal is fed to IC2 for a short interval of time, before being blocked again.

IC2 is a binary counter, and the outputs that concern us are from pins 14, 11 and 6. Pin 14 is the *most significant bit* and pin 6 the *least significant bit*. See the panel for an explanation of what happens here.

To understand these outputs, you should be able to count in binary, using three bits. At the start of the counting process, all the bits have zero values, i.e. 000. The left-most bit is called the *most significant bit*, and the right-most bit is called the *least significant bit*. (When we write numbers normally, a number 1 in the left-hand position represents one hundred, 1 in the middle column represents ten, while the number 1 in the right-hand position means one. One hundred is a more significant number than one, hence the nomenclature.)

The three bits of our binary number have values, from left to right, of 4 ($=2^2$), 2 ($=2^1$) and 1 ($=2^0$). This means that a number 1 in the left-hand column signifies the normal number four, in the middle column it would represent two, and in the right-hand column, it would represent one. This should help you understand the patterns of bits which emerge as the clock waveform is counted, as the next paragraph explains.

As each cycle of the clock enters IC2, it increments its internal counter and the values of that counter are shown by the states of pins 14, 11 and 6. After the first pulse, these three pins would have states corresponding to 001; after the second, 010; after the third, 011; after the fourth, 100; after the fifth, 101; after the sixth, 110. This sequence of 3-bit numbers represents a binary count from 1 to 6 in ‘normal’ parlance. These six states are used to illuminate the conventional pattern of dots on a die, using LEDs.

As the clock cycles are counted, the LEDs flicker as the die is ‘rolled’. Resistors R4 to R7 are used to limit the current through the LEDs. When the counter stops, it can be in any of the positions shown in **Table 1**. Because of

Table 1 Dice number relative to counter output

Step	*Pins*			*Dice number*
	14	*11*	*6*	
1	0	0	1	4
2	0	1	0	2
3	0	1	1	6
4	1	0	0	1
5	1	0	1	5
6	1	1	0	3

the way in which the dots are grouped on the faces of a die, the wiring of the LEDs is simpler than it might otherwise be. You can see from Figure 1 that there are really only three sets of connections to the LEDs – one from each bit of the counter output. When pin 14 is at logic 1, four LEDs are lit, corresponding to the number four. For the number five, pins 14 and 6 will be at logic 1. For the number one, only pin 6 is at logic 1. For three, pins 11 and 6 are at logic 1, for six, pins 14 and 11 are at logic 1, and for two, only pin 11 is at logic 1. These conditions are summarised in Table 1.

Construction

Two pieces of Veroboard (of the strip type) are needed – one for the main circuit (**Figure 2**) and the other for the display LEDs (**Figure 3**). The display board is easier to build, so we will do that first. It measures 20 holes by 18 strips.

Cut the tracks as shown in Figure 3, using a 3 mm (⅛inch) twist drill rotated between thumb and forefinger. Insert and solder the Veropins, resistors and the wire links. Then solder the LEDs in place, making sure their polarities are correct. The LEDs are mounted proud of the board (**Figure 4**, which will help when you come to fix the board into a case.

The main board measures 32 holes by 18 strips. As before, remove the tracks in the places shown in Figure 2. Note that track 'L' is *not* broken under the position for IC2. Solder in the IC holders and the links. Solder in R1 and R2 vertically, and R3 horizontally. Then solder in C1, C2 and C3; these all being electrolytics, check their polarities. Solder in the Veropins and prepare the connections to those components not on the main board, using stranded, insulated wire. Check the circuit carefully, and read the handling precautions given earlier before carefully inserting the two ICs.

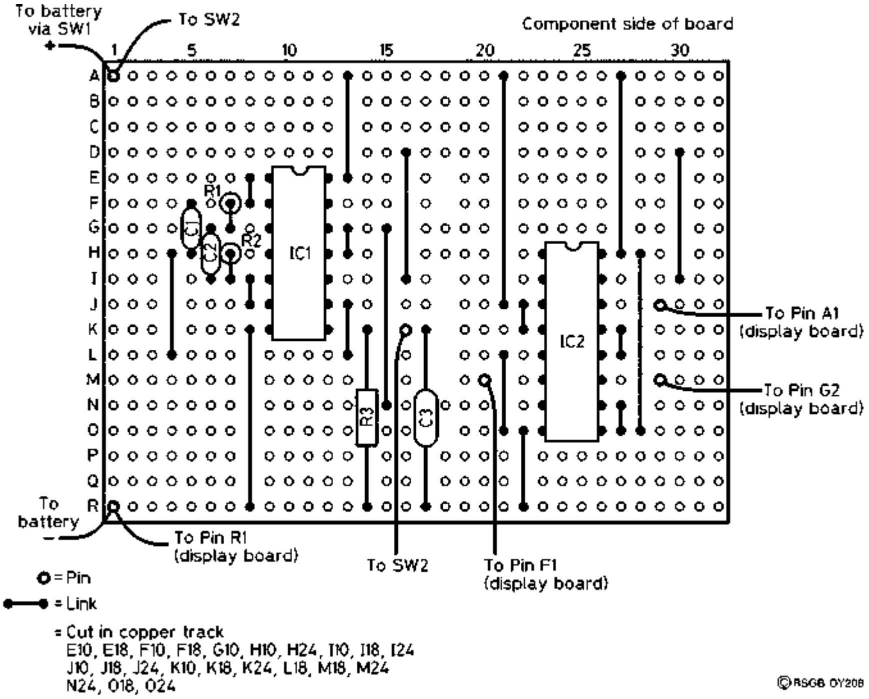
To battery via SW1
+
To SW2
Component side of board
1
5
10
15
20
25
30
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
R1
R2
C1
C2
IC1
IC2
R3
C3
To Pin A1 (display board)
To Pin G2 (display board)
To battery
–
To Pin R1 (display board)
To SW2
To Pin F1 (display board)
O = Pin
= Link
= Cut in copper track
E10, E18, F10, F18, G10, H10, H24, I10, I18, I24
J10, J18, J24, K10, K18, K24, L18, M18, M24
N24, O18, O24
©RSGB DY208

Figure 2 Main board, component layout

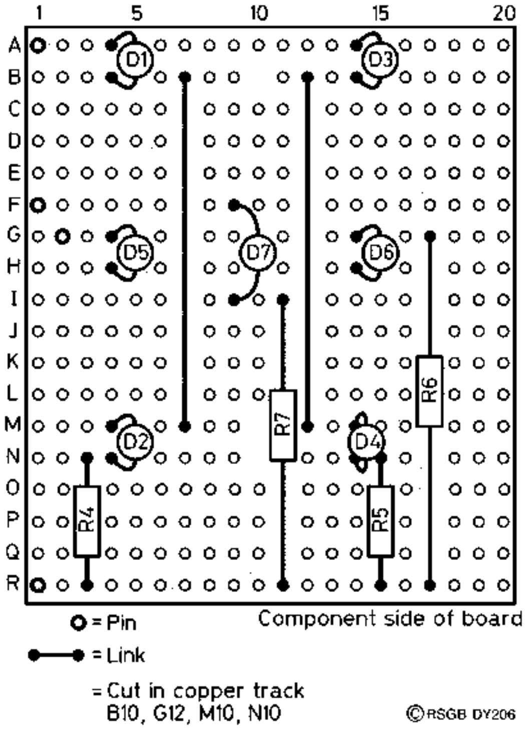
1
5
10
15
20
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
D1
D3
D5
D7
D6
D2
D4
R4
R5
R6
R7
O = Pin
= Link
Component side of board
= Cut in copper track
B10, G12, M10, N10
©RSGB DY206

Figure 3 Board for the six LEDs

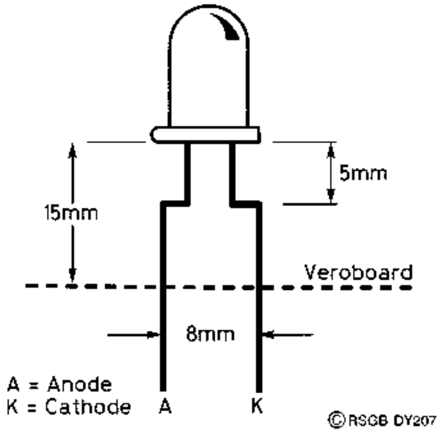

Figure 4 D7 lead bending

Testing and use

When the circuit is switched on, it is likely that the centre LED only will be lit. Press SW2, the 'throw' switch. All the LEDs should flicker, even after the switch is released, but only for a second or so. Then a standard pattern should show, signifying numbers between one and six. If this does not happen, switch off and check your circuit again. Are there any missing links? Are there any obvious dry joints (usually dull instead of shiny)? Can you see any solder bridges between tracks? Did you choose to ignore the handling precautions? If none of these results in a working circuit, try some fault-finding.

- Disconnect the lead to D7 on the main board (marked 'to pin F1' on Figure 2). Touch the lead on to the positive supply rail. D7 should light. Now transfer this lead on to pin 4 of IC1. If it flickers, then the oscillator is running (which it should).
- Transfer this lead to pin 10 of IC2. D7 should flicker when SW2 is pressed, but should be on permanently when it is released. This means that the correct signals are reaching IC2 from IC1.
- If one of the chains of LEDs (i.e. D1 and D2, D3 and D4, D5 and D6) does not light, you may have one *or both* LEDs the wrong way round.

When the die is working, you may care to experiment with the values of R1 and R2, but you should keep their values the same, i.e. R1 = R2. The larger the values of these, the more slowly the die will appear to 'roll'.

Finishing touches

On completion of the project, you will want to mount it in a smart case; any plastic box is suitable for this, with the display board mounted so that the

LEDs protrude through the top and are fixed to it using LED clips. The way in which you mount both boards to the case is entirely up to you.

Parts list

Resistors: all 0.25 watt, 10% tolerance, or better

R1, R2	5600 ohms (Ω)
R3	15 kilohms (kΩ)
R4–R7	470 ohms (Ω)

Capacitors

C1, C2	1 microfarad (μF) electrolytic, 16 V WKG
C3	68 microfarads (μF) electrolytic, 16 V WKG

Semiconductors

IC1	4011
IC2	4029
D1–D7	LEDs, any size and colour

Additional items

SW1	On-off switch SPST toggle
SW1	Push-to-make, non-latching
	Veroboard, 2 pieces, see text for sizes
	Veropins, 10
	Stranded insulated wire for general wiring
	Single-core insulated wire for links
	PP3 battery clip
	PP3 battery
	Plastic box to suit
	LED mounting clips (7)
	Means of mounting boards to box

Source

Components are available from Maplin.

63 The absorption wavemeter

Introduction

The purpose of the absorption wavemeter is to check that a transmitter is radiating within the correct waveband, and to detect any spurious *harmonic* emissions. A harmonic frequency is an integral multiple of the carrier frequency – e.g. if the carrier is at a frequency *f*, the harmonics are at frequencies 2*f*, 3*f*, 4*f* . . . and so on.

It can also be employed as a relative field-strength indicator, being used in experimentation with aerials. What it *cannot* do is to measure the transmitted frequency accurately; all it can do is to place the signal within a particular band, say the 7 MHz or the 18 MHz band (or their harmonics).

How it works

You may recognise the circuit of **Figure 1** as a crystal set with a meter replacing the headphones. It is simply the parallel combination of an inductance with a variable capacitance, which constitutes a resonant circuit with a variable frequency. A dial on the variable capacitor can be roughly calibrated with frequency. The fixed capacitor across the meter helps to improve the meter reading by making it read the peak value of the carrier wave rather than the average value.

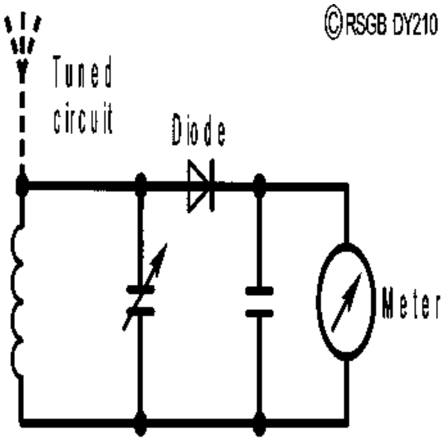

Figure 1 Absorption wavemeter circuit

When the wavemeter is situated in the vicinity of a relatively strong radio-frequency (RF) field, the inductor absorbs a small amount of energy from that field. If the circuit is resonant at the same frequency as that of the transmitter, an RF voltage is produced across the coil and is proportional to the strength of the RF field. The RF voltage is *rectified*, or *detected* (turned from AC into DC) by the diode and the second capacitor, and is displayed by the meter.

The tuned circuit

The formula for calculating the resonant frequency of a coil and capacitor is relatively simple. It is

$$f = \frac{1}{2\pi\sqrt{LC}},$$

where f is the resonant frequency (Hz),
L is the inductance (H), and
C is the capacitance (F).

If this formula puts you off, there is this slightly easier version:

$$f = \frac{159}{\sqrt{LC}},$$

where f is the frequency in **MHz**,
L is the inductance in **microhenries** (**µH**), and
C is the capacitance in **picofarads** (**pF**).

When a variable capacitor is specified, it is usual to quote its maximum capacitance, i.e. the value when the vanes are fully meshed. When the value of the capacitor is small, it is usual to quote its minimum capacitance also, the value when the vanes are fully open. If you use a variable capacitor of 10–100 pF with a 100 µH inductor, you will use the value of 10 pF to calculate the maximum resonant frequency, and the value of 100 pF to calculate the minimum resonant frequency, giving a tuning range of 1.6 MHz to 5 MHz. Check it yourself!

64 An HF absorption wavemeter

Introduction

An absorption wavemeter is basically a simple tunable detector circuit, such as would be used in a crystal set, but with the headphones replaced by a meter, in order to indicate the strength of the received signal.

The circuit

Figure 1 shows the circuit of the wavemeter. It is designed to cover all the HF bands from 1.8 MHz to 28 MHz in four switchable ranges. There is no built-in method of amplification, so a sensitive meter is needed in order to indicate sufficiently using the available absorbed energy. The meter does not need calibrating, it serves to produce only an indication of the absorbed power, *not* its absolute value.

An absorption wavemeter does not have outstanding selectivity, mainly because of the loading effect of the meter. This problem is ameliorated here (but only slightly) by the inclusion of R1 in series with the meter. It gives

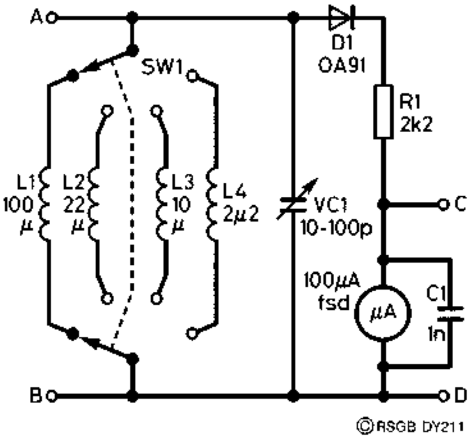

Figure 1 HF absorption wavemeter, circuit diagram

improved selectivity at the expense of a slight loss of sensitivity. The appearance of capacitor C1 across the meter is discussed in *The Absorption Wavemeter*, in the previous section.

Coils

These can be purchased (see the parts list) or hand-wound on short lengths of PVC water pipe or conduit, and the coil-winding details are given in **Table 1**. In theory, the 100 pF tuning capacitor would cover all the HF bands with only three coils, omitting the 22 μH coil. Sample calculations are explained in the previous article. However, its inclusion avoids the common problem of having some bands cramped at the extreme edges of the tuning scales.

Table 1 Close-wound coil values

L (μH)	No of turns	Span (mm)	SWG
PVC FORMER OF OUTSIDE DIAMETER 25 mm			
2.2	9	9	20
10	24	24	20
22	44	44	20
100	110	60	26
PVC FORMER OF OUTSIDE DIAMETER 20 mm			
2.2	10	10	20
2.2	9	5	26
10	32	32	20
10	24	12	26
22	63	62	20
22	41	21	26
100	134	70	26
PVC FORMER OF OUTSIDE DIAMETER 19 mm			
2.2	11	11	20
2.2	9	5	26
10	35	34	20
10	25	13	26
22	68	67	20
22	44	22	26
100	157	79	26

Construction

Hand-wound coils, using 19–25 mm PVC formers, have single-layer windings, with their ends secured by threading the wires through two small holes drilled in the formers at each end of each winding.

You may have wondered about the use of the two-pole switch SW1, as all the lower ends of the coils are grounded anyway. The answer lies in the ease of construction. Each coil can be mounted on the tags of the switch for support. If you are using the commercial coils, they are small enough to be accommodated on a miniature rotary switch; a larger rotary switch is required for hand-wound coils. The specification of a 6-way switch allows for the addition of extra frequency ranges at little extra cost. Rotary switches such as the 2-pole, 6-way variety have an adjustable end-stop which allows the number of ways to be set anywhere between 2-ways and 6-ways. Remove the nuts from the switch, and a ‘washer’ will fall out. It is not a washer, really, and has a little piece of metal bent over at right angles to the washer. This fits into one of 12 holes in the body of the switch, and prevents the shaft from turning through 360° and selecting the number of ways.

Calibration

This process needs a commercial amateur-bands transceiver, if it is to be done quickly and accurately.

- Connect the transceiver to a dummy load, and connect a piece of flexible wire from point A in Figure 1 and wrap two or three turns around the cable between the transceiver and the dummy load.
- Turn the transceiver power level control to minimum, and set the frequency at 1.81 MHz and the mode to FM or AM. Switch to transmit.
- With the wavemeter set to its lowest frequency range (i.e. 1.8–3.5 MHz), rotate the tuning capacitor until maximum deflection is obtained from the meter needle. Write the frequency on the inner ring of the wavemeter scale at that point.
- Switch off and retune the transmitter to 3.5 MHz. Repeat the above process.
- Repeat, using switch range 2 for 3.5, 7 and 10 MHz, range 3 for 7, 10 and 14 MHz, and range 4 for 14, 18, 21, 24 and 28 MHz.

An alternative source of calibration signal could be an HF signal generator, with a single-turn loop of wire at the remote end of its cable, and the wire from the wavemeter brought close to the loop.

Remember that the marks on the scale represent frequency *bands*, not precise frequencies.

Extending the range

If you would like to experiment with increasing the range to the lower VHF band, then 50 MHz should be achievable with an additional self-supporting coil (no former). Try three turns of 20 SWG enamelled copper wire of 25 mm ID (internal diameter). Wind it on some 25 mm PVC pipe, as before, then slide out the pipe! The same result could be achieved with four turns of 19 or 20 mm ID, or 12 turns of 7.1 mm ID.

Parts list

Resistor	
R1	2200 ohms (Ω) 0.25 watt, 10% tolerance
Capacitors	
C1	1 nanofarad (nF) min. ceramic
VC1	100 picofarad (pF) variable
Inductors	
L1	100 microhenries (μH)
L2	22 microhenries (μH)
L3	10 microhenries (μH)
L4	2.2 microhenries (μH)
Semiconductors	
D1	Germanium OA91 or 1N4148
Additional items	
SW1	Rotary 2-pole, 6-way
	Moving-coil meter, 50 or 100 μA FSD (any meter of this sensitivity will do)
	Plastic box, approx. 150 by 80 by 50 mm
	Connecting wire, coloured

65 A vertical aerial for 70 cm

Introduction

If the range of your hand-held transceiver is very limited, what you need is a vertical aerial mounted outside. This design is a half-wave dipole fed at the end instead of the middle. Because the impedance of the dipole is high at its ends (and low in the centre), a matching circuit is needed so that this high impedance can be matched to that of the low-impedance coaxial cable. All that is needed is a coil, which increases the *electrical length* of the aerial to ⅝-wavelength. (Note that the *electrical* length (i.e. the length as it appears to an RF signal) is not necessarily the same as the *actual* length.)

Construction

The aerial element and the coil are made from a single piece of 1.5 mm welding (brazing) rod, and the dimensions are given in **Figure 1.** Wind the coil around a 4 mm rod or the shank of a twist drill. The lower end is filed to a point and then soldered into the centre conductor of a 4-hole panel-mounting BNC socket. (Try to obtain a good-quality BNC socket with PTFE insulation – the insulation of cheaper sockets is easily damaged.) Trim the element to 427 mm (top of element to top of coil) *after* the wire has been soldered to the socket.

The base coil causes the aerial to be rather ‘whippy’, so a piece of 5 mm plastic knitting needle can be cut to the length of the coil and then forced into it.

The radials are made from four lengths of 3 mm welding rod. These are bent and soldered into the four mounting holes of the socket, and then cut to the lengths shown in Figure 1.

Testing

Using a standing-wave-ratio (SWR) meter connected between your aerial and the transceiver, measure the SWR (or obtain some indication of the reflected power) on transmit. If it is greater than 1.5, switch off the transmitter, trim about 3 mm off the end of the element, and try again. Repeat the process until the measured SWR (or the reflected power) is as low as you can get it.

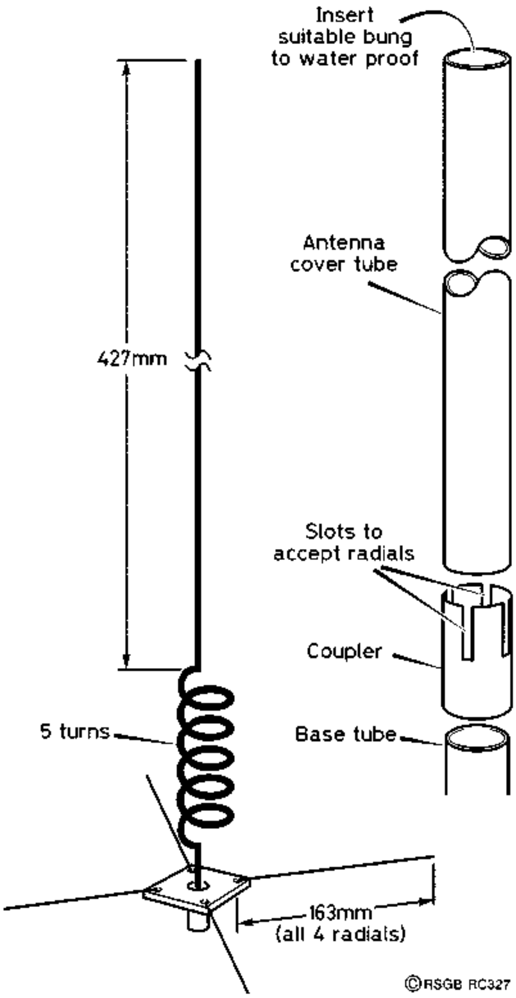
Insert suitable bung to water proof
Antenna cover tube
Slots to accept radials
Coupler
Base tube
427mm
5 turns
163mm (all 4 radials)
©RSGB RC327

Figure 1 Construction of the 70 cm antenna

Installation

The aerial is made waterproof by enclosing it in a 22 mm diameter PVC waste water pipe. It is 'weldable', and available at plumbers' merchants, usually by the metre. A coupler is slotted to take the radials (Figure 1). File the BNC socket as required, so that it slides inside the coupler until the radials poke out of the slots. Cut a length of plastic tubing which is 30 mm longer than the aerial, and push it into the coupler. You will then need a plastic bung or screw-top to waterproof the top end.

Plastic welding solution is now applied to the joints, sealing the aerial inside the tube. The coupler has been weakened as a result of making the slots, so it is worthwhile applying PVC tape around this joint until the welding solution sets.

A support for the aerial is made from an off-cut of tube pushed into the lower end of the coupler and held with a self-tapping screw.

Parts list

1	4-hole panel-mounting BNC socket
1 m	3 mm brazing rod (may be available from a small garage)
1 m	1.5 mm brazing rod
1 m	22 mm PVC waste water pipe
1	22 mm straight coupler
	Plastic welding solution

66 A UHF corner reflector aerial

Introduction

The corner reflector is a well-known design and is capable of good performance on the VHF and UHF bands. At UHF, the practical implementation of the corner reflector is an ideal constructional project.

Some details

Expressed quite simply, the aerial consists of a $\frac{\lambda}{2}$ dipole (where λ is the standard symbol for wavelength, making a '$\frac{\lambda}{2}$ dipole' a half-wave dipole).

Nothing new in that, you might say. However, the interesting feature is the reflector, which is not the usual single element, but a 90° metal ‘corner’, acting rather like a parabolic dish as used for satellite signal reception. The wind resistance of this type of reflector makes it impractical so, to reduce the ‘windage’ quite significantly, we make the ‘corner’ from closely spaced rods, as illustrated in **Figure 1**.

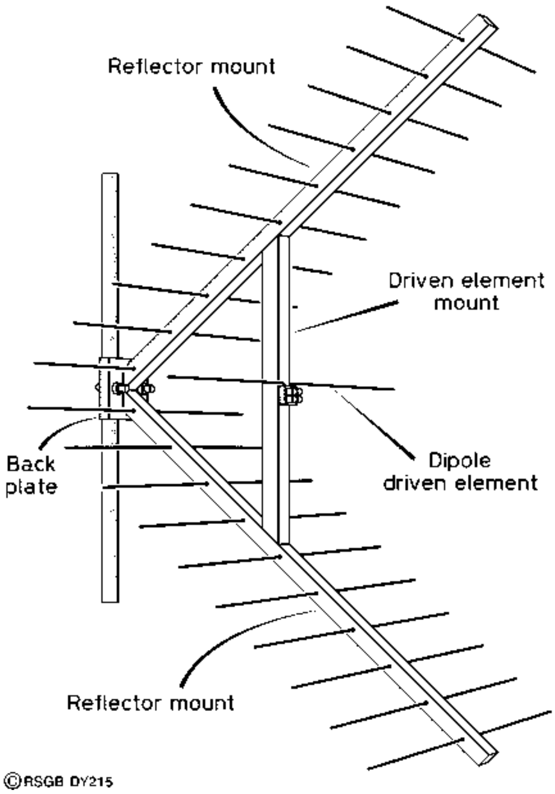

Figure 1 70 cm corner reflector antenna

The reflector consists of a number of 0.6λ rods, spaced from each other by 0.1λ. The aerial frame can be made of metal or wood, but wood is easier to work with, and mounting the elements to the frame is simpler. The prototype was made with wood of 20 mm by 15 mm cross-section, as **Figure 2** shows. The wood was varnished for protection. The elements were made from 1.5 mm diameter copper wire, because a large reel of the wire happened to be available. The wire diameter is not critical; tubing could be used just as successfully. 14 SWG hard-drawn copper aerial wire would be even better than that used in the prototype.

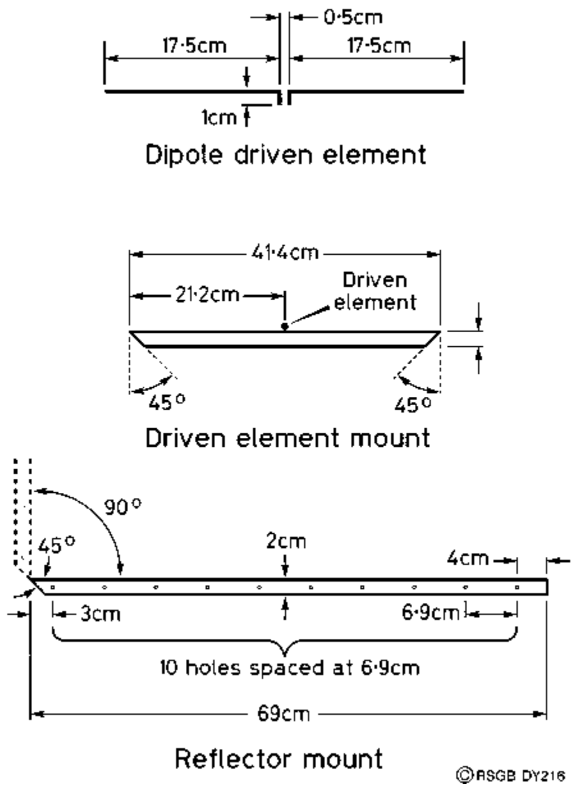
0·5cm
17·5cm
17·5cm
1cm
Dipole driven element
41·4cm
21·2cm
Driven element
45°
45°
Driven element mount
90°
45°
2cm
4cm
3cm
6·9cm
10 holes spaced at 6·9cm
69cm
Reflector mount
©RSGB DY216

Figure 2 Driven element dimensions, together with boom dimensions for driven element and reflectors

Construction

This project is just as much a woodworking project as a radio project! Follow the instructions carefully, and you should have little trouble.

- Cut the booms for the reflectors, as shown in Figure 2. A mitre block is invaluable here in producing the 45° corners.
- Using the dimensions given on the diagram, mark the hole positions for the reflector elements, and then drill holes of a size which holds the elements firmly.
- Cut the driven element boom according to the diagram, and mark the point midway along the longer side, which will assist you later in positioning the driven element.
- Cut the back plate to size (about 120 mm by 80 mm). You may need to alter this size depending on the size of the U-bolt you will be using to clamp the aerial to the mast.

- If you want to be extra cautious in your construction, use the belt-and-braces approach, commonly known as ‘screw-and-glue’ to fix the booms to each other and to the back plate.
- Fix the reflector booms to the back plate first, then slide in the driven-element boom until it will go no further, then apply the wood glue and screw the two ends tightly to the reflector booms. Leave for the period prescribed by the glue manufacturers for the glue to harden.
- Varnish the whole structure.
- Cut the driven element to the correct size plus a couple of centimetres (the reason for this will be evident in the *Testing* section), and fix it to the centre of its boom (at the position you marked earlier) with a ‘chocolate block’ connector to which the coaxial feeder cable will eventually be connected.
- Cut and fix the reflector elements in place. If you find that these are a loose fit in the holes then, *for each element*, drill a pilot hole through the boom to intersect the hole for the element. File off the point of a woodscrew, and screw it gently into the pilot hole until it meets the element and grips it in place. You will now see why the point was filed off! Alternatively, you can glue the elements in place.

Testing

Place the aerial on a mast, clear of obstructions. Connect it to a transceiver with a length of coaxial cable, with a standing-wave-ratio (SWR) meter in circuit. Find a clear frequency, identify your transmission and ask if the frequency *really is* clear. If so, key the transmitter again and note the SWR. **Do not stand in front of any aerial when it is radiating!** The length of the driven element must be adjusted to obtain an SWR of less than 2. If you have to shorten the dipole, bend the ends over rather than cut them off. That way, if you go too far, your can lengthen them again! The dipole was initially cut too long intentionally, to allow for adjustment here! Bending the ends over also reduces the risk of physical damage to clothing, skin and eyes. You may like to consider applying the same technique to the reflector elements for that reason alone.

Moving on . . .

Once you have warmed to the idea of the corner reflector as an aerial, you might like to ring the changes regarding the reflector. How about replacing the 20 reflector elements with a wire mesh, such as garden centres sell as ‘chicken wire’? Choose the finest mesh if there is a choice. Some extra support may be needed around the edges of the mesh, but you could go on to make a comparison of aerial gain between the two types, using the *UHF Field Strength Meter* described elsewhere in this book.

Materials

Stiff wire or thin-walled tubing for dipole and reflector
Frame – wood, 15 mm by 20 mm cross-section, lengths given in text
Back plate – stout plywood, dimensions given in text
U-bolt to suit mast
Wood screws
Wood glue
50 W coaxial cable for feeder
2-terminal 'chocolate block' for dipole connection to feeder
Varnish

67 A switched dummy load

Introduction

A dummy load is a pure resistor of value 50 ohms which can replace your transmitting aerial and enable you to operate the transmitter for test purposes without radiating a signal. It sounds simple enough, but there are two main problems. Firstly, it is impossible to delve into your junk box and emerge with a resistor that will dissipate 100 W PEP and still retain its marked resistance value. Secondly, a 'pure resistance' is very difficult to achieve. A pure resistance is a device which has resistance but no reactance. All common resistors have significant reactance at radio frequencies, particularly the wire-wound varieties, which have a helical (i.e. wound like a coil) construction. This is particularly annoying, because wire-wound construction is normally used for large-wattage resistors.

Although all resistors have *some* reactance, not all are quite as bad as the wire-wound type. Carbon film resistors are made by depositing a thin film of carbon on the surface of a small, hollow ceramic cylinder, the thickness of the film of carbon determining the value of the resistor. Provided the lead lengths are kept short, these resistors have a tolerably small reactance, and will be used in this project.

Bearing the load

A 2 W carbon film resistor is hardly going to withstand our 100 W PEP of SSB, so it is obvious that the design of our dummy load must be a little more complex than a single resistor and a switch, despite what **Figure 1** might suggest! In fact, it uses 20 resistors, each of value 1000 Ω (1 kΩ). How does this solve our problem?

Perhaps a little theory is in order here, but no more than is required by the Radio Amateurs' Examination.

When two *equal* resistors of value r are combined in parallel (i.e. side by side), the total resistance, R_T, is given by:

$$\frac{1}{R_T} = \frac{1}{r} + \frac{1}{r}.$$

Adding $1/r$ to $1/r$ gives $2/r$, therefore:

$$\frac{1}{R_T} = \frac{2}{r}, \quad \text{i.e. } R_T = r/2.$$

So, by connecting *two* equal resistors in parallel, we get a combined resistance which is *half* the individual resistances. If we combine *three* in parallel, we get a *third* of the resistance, and so on.

Here, we are connecting *20* resistors of 1 kΩ in parallel, so we will produce an overall resistance of *one-twentieth* of the individual resistance, i.e. 1000/20 = 50 Ω, which is what we set out to achieve!

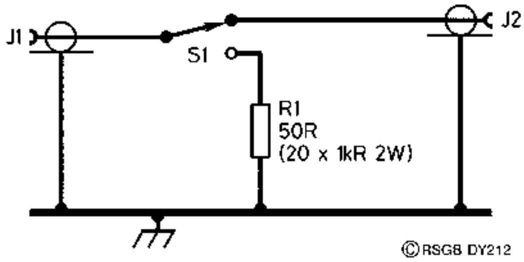

Figure 1 Basic circuit diagram

This is not the only advantage, though. Each resistor is capable of dissipating 2 W, so 20 of them will safely dissipate 40 W, for short periods at least. This power dissipation is approximately the same as 100 W PEP of normal speech, so the design should be capable of use in an 'average' amateur station. The power-handling ability of any dummy load can be improved by providing a 'heat sink' which helps to conduct the heat energy away from the resistors, thus lowering their temperature. One popular heat sink is a can of transformer oil, into which the resistors are immersed. This design uses a rather more mundane heat sink, but which is adequate for the job in hand.

Figure 2 PCB resistor supports

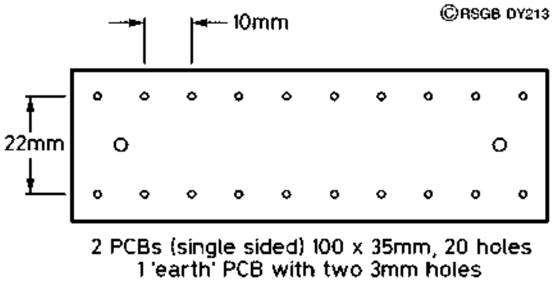
10mm
©RSGB DY213
22mm
2 PCBs (single sided) 100 x 35mm, 20 holes
1 'earth' PCB with two 3mm holes

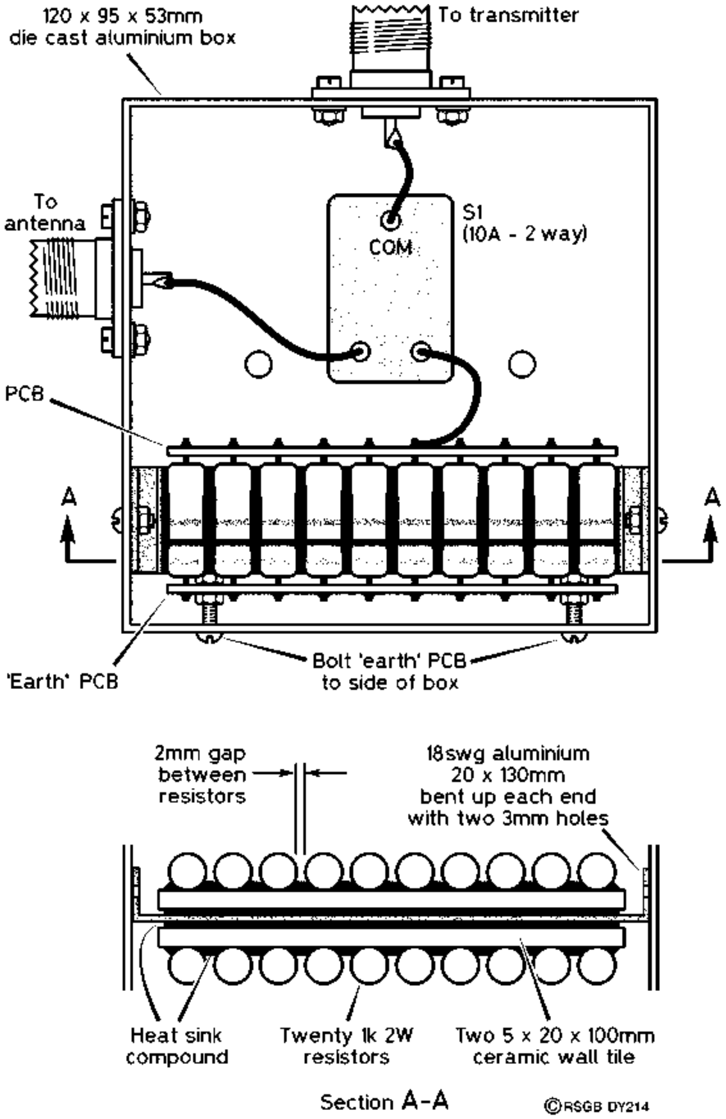
120 x 95 x 53mm
die cast aluminium box
To transmitter
To antenna
S1
(10A - 2 way)
COM
PCB
A
A
'Earth' PCB
Bolt 'earth' PCB to side of box
2mm gap between resistors
18swg aluminium 20 x 130mm bent up each end with two 3mm holes
Heat sink compound
Twenty 1k 2W resistors
Two 5 x 20 x 100mm ceramic wall tile
Section A-A
©RSGB DY214

Figure 3 Switched dummy load, component layout

Pouring the heat away

Two basic PCBs are needed, the details being given in **Figure 2**. Both measure 11 × 35 mm. Each has a set of 20 holes of diameter about 1 mm to take the ends of the resistors. Only one of them has two 3 mm holes which are used for mounting the completed load. **Figure 3** shows how the heat sink is assembled. Solder one end of each resistor into the 'earth' PCB, and then mount this to the box using the two bolts as shown in the diagram. Between each rank of resistors is a 'sandwich' consisting of an aluminium strip and two pieces of ceramic wall tile, to act as a heat sink for the resistors. Heat sink compound is used to provide good thermal contact between the resistors and the tiles, and between the tiles and the aluminium strip. This is shown in the lower part of Figure 3. Thread the loose wires of the resistors through the holes in the second PCB, solder into place, and crop the protruding wires.

Switching

A changeover switch must be used so that the transceiver can be switched between the dummy load and the aerial without unscrewing connectors. For most purposes, an ordinary 10 A 230 V changeover switch will suffice, as found in many electrical shops and DIY stores.

Wire this into the circuit as shown in Figure 1 and Figure 3. Check your wiring. Put S1 in the 'dummy load' position. If you have a multimeter which includes an ohmmeter, measure the resistance across the socket, J1, before connecting it to the transceiver. It should be very near 50 Ω. With S1 in the 'aerial' position, there should be an infinite resistance across J1. Move your ohmmeter to read the resistance across J2. It should be infinite for both positions of S1. If all seems correct, close the box and your dummy load is ready for use!

Parts list

Resistors
  1000 Ω (1 kΩ) 2 W carbon film, 20 required
Additional items
  10 A 230 V changeover (SPDT) switch
  SO–239 sockets, 2 required
  Aluminium box 120 × 95 × 53 mm or similar
  Ceramic wall tile cut as required
  Aluminium strip, 18 SWG
  Nuts and bolts as required

68 A simple Morse oscillator

Introduction

This is an excellent project which uses the 'junk box' as its source of components. If you have trouble in finding the bits for this one, a good source of the components for this and many other similar projects is to be found with the parts list at the end of the project.

The circuit

This is shown in **Figure 1**, and uses an 'unknown' Plessey chip, which makes the overall circuit extremely easy to build. A 0.1 μF capacitor is connected between pins 7 and 8, a speaker (in the popular 8 to 25 Ω impedance range) is connected between pins 8 and 9. If a 9 V battery is connected with its positive terminal to pin 8 and its negative terminal to pin 1, 3 or 5, a tone will be produced in the speaker.

To make this circuit into a good Morse practice oscillator, it is necessary only to insert a Morse key into the supply rail from the battery.

However, there is another refinement which you may care to build into the circuit. The tone from the loudspeaker is different, depending upon which of pins 1, 3 or 5 you use. In the prototype, a single-pole changeover switch was used to select the tones from pin 1 or pin 3, and the Morse key would be connected to the circuit via a small jack socket. There is no need for an on/off switch, as the Morse key performs that function. The switch and the jack socket can be seen in the photograph.

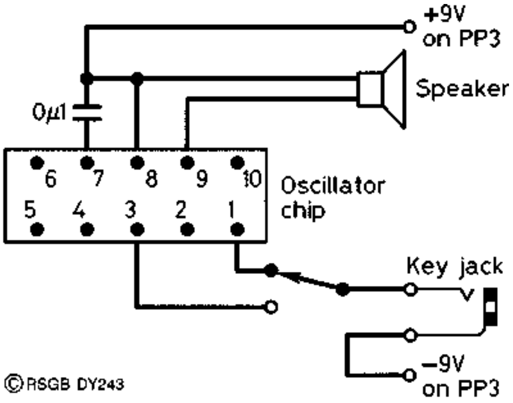

Figure 1 Morse oscillator, circuit diagram

The case

Almost any small loudspeaker will do; the higher the impedance the better. The 'impedance' is the figure usually printed on the rear of the speaker magnet. It is a value given in ohms (Ω). For simple circuits like this, you will usually find that the higher the impedance (within reason), the louder the sound it will produce. Speakers from old transistor radios will work, although their impedances can be rather low sometimes.

Any case big enough to house the components and the battery will be suitable. The prototype used a 'Walkman'-type speaker and case, and is shown in the photograph.

To use the circuit, simply attach a 3.5 mm jack plug to your Morse key, and insert it into the socket. Nothing should happen until you press the key, when a tone should be heard from the loudspeaker.

Another application

Try soldering the two wires of a 'twisted pair' to the 3.5 mm jack plug. Touching the wires together produces a tone from the speaker. This simple circuit can then be used as a 'continuity tester'. Touch the two wires to the ends of a fuse. If there is no sound, then the fuse is blown. There are many other such tests you could perform with this device – to check whether there really *is* a connection between one end of a wire and the other, for example.

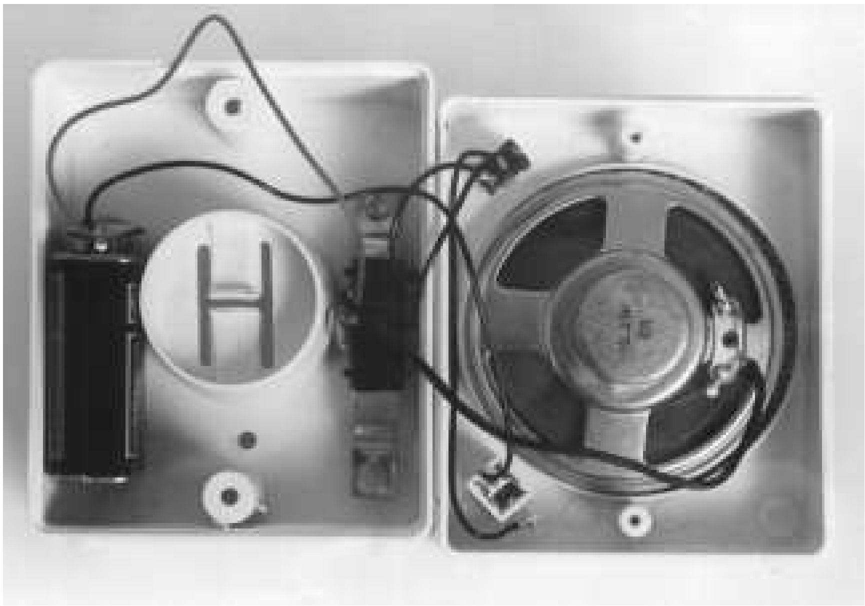

Warning

You must **never** make such tests on any equipment which is connected to the mains supply, even if it is switched off. If you want to make any such tests, make sure you are supervised by someone who understands what you are doing and is competent to advise and supervise you.

Parts list

Plessey oscillator chip
0.1 µF capacitor (disc ceramic)
Loudspeaker 8 to 25 Ω
3.5 mm jack socket (and 3.5 mm jack plug if needed)
Single-pole changeover switch

The Plessey oscillator chip is available from J. Birkett, 25 The Strait, Lincoln LN2 1JF, tel: 01522 520 767.

John Birkett may also provide a kit of parts (the chip, capacitor and loudspeaker).

69 A bipolar transistor tester

Introduction

This is a circuit which will test normal transistors, i.e. npn or pnp. It has the advantage of being able to test devices while they are still connected in their original circuits. However, when such tests are made, **the circuit containing the transistor under test must not be switched on**.

The circuit and how it works

The circuit runs from a 9 V battery such as a PP3 or six AA-type 1.5 V cells. Alkaline cells are to be preferred, as their electrolyte leakage properties are better. The circuit shown in **Figure 1** uses a single CMOS integrated circuit type 4001 or 4011. CMOS circuits require special handling precautions which are described in the project *Christmas Tree LEDs*, elsewhere in this book.

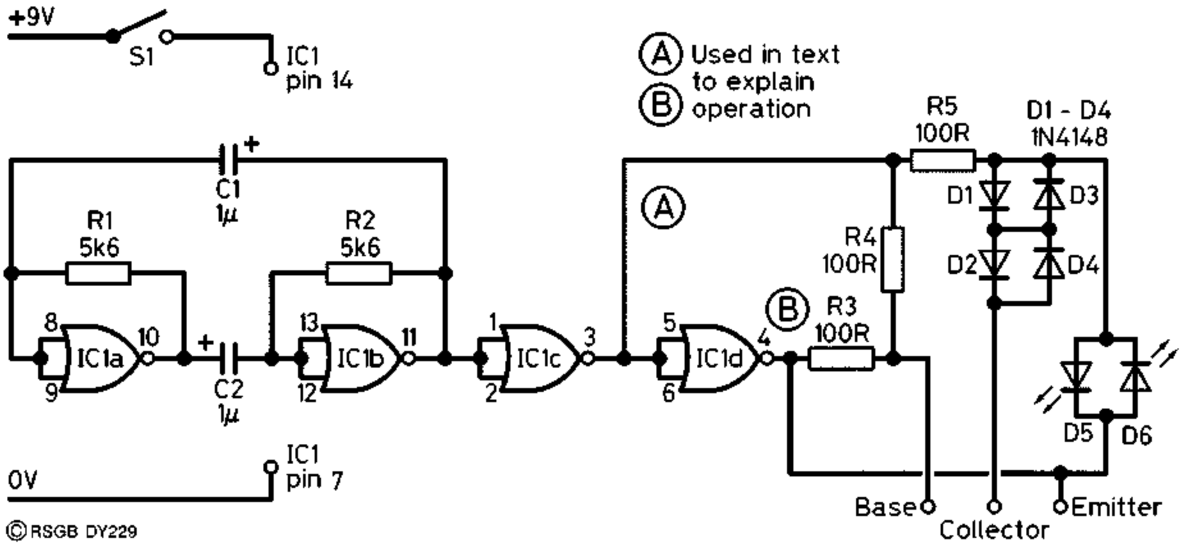
+9V
S1
IC1
pin 14
Ⓐ Used in text
to explain
Ⓑ operation
R5
100R
D1 - D4
1N4148
+
C1
1µ
R1
5k6
R2
5k6
Ⓐ
D1
D3
R4
100R
D2
D4
Ⓑ
R3
100R
8
10
IC1a
9
+
C2
1µ
13
IC1b
11
12
1
IC1c
3
2
5
IC1d
4
6
D5
D6
IC1
pin 7
0V
Base
Collector
Emitter
©RSGB DY229

Figure 1 Transistor tester, circuit diagram

Inside the IC are four logic gates (see *Digital Logic Circuits*) which are all connected as inverters, which means that the output signal is always the logical 'opposite' of the input. The first two gates are connected as an oscillator; the circuit being the same as that used in *An Electronic Die*.

The output of the oscillator, at pin 11, is connected to the input of a buffer stage, IC1c, which helps to isolate the oscillator from the circuit that follows it. The buffer output appears on pin 3, which we shall label as *test point A* for future use. Another inverter, IC1d, follows this, its output at pin 4 being labelled *test point B*.

There are two LEDs connected back to back at the circuit output. These are D5 and D6, D5 being red and D6 being green. Whatever the output of the oscillator at any instant, one of the LEDs must be lit and the other unlit. With point A positive and point B zero, the red LED is lit, when A is zero and B is positive, the green LED is lit. Because the oscillator output is repeatedly switching from one polarity to the other, the lit LED is alternately red and green. They switch between the two colours much faster than we can see, so what we think we see are both LEDs lit together.

The two 100 ohm resistors, R3 and R4, provide the bias to the base of the transistor under test. When A is positive and B is zero, the base-emitter junction of the transistor will be forward-biased, and the transistor will switch on (if it is a working npn type). When the transistor is on, it effectively short-circuits D5 (the red LED) and it extinguishes. When point A is zero and B is positive, an npn transistor will be switched off and the green LED (D6) will light. Thus, for a working npn transistor, only the green LED is lit.

If the transistor under test is a pnp type, it will switch on when A is zero and B positive, thus short-circuiting the red LED (D6). When A is positive and B zero, the transistor is off and the red LED (D5) is lit. Thus for a working pnp transistor, only the red LED is lit.

To summarise, the states of the LEDs indicate the following conditions:

- Both LEDs apparently lit: no transistor connected, or transistor permanently open circuit.
- Neither LED lit: a collector-emitter short-circuit is almost certain.
- Red LED alone: pnp transistor in working order.
- Green LED alone: npn transistor in working order.

Construction

The prototype tester was built on a piece of Veroboard measuring 15 strips by 18 holes. Cut the tracks using a track-cutting tool or a 3 mm (⅛ inch) twist drill, as shown in **Figure 2**. Notice that, in this diagram, there is no

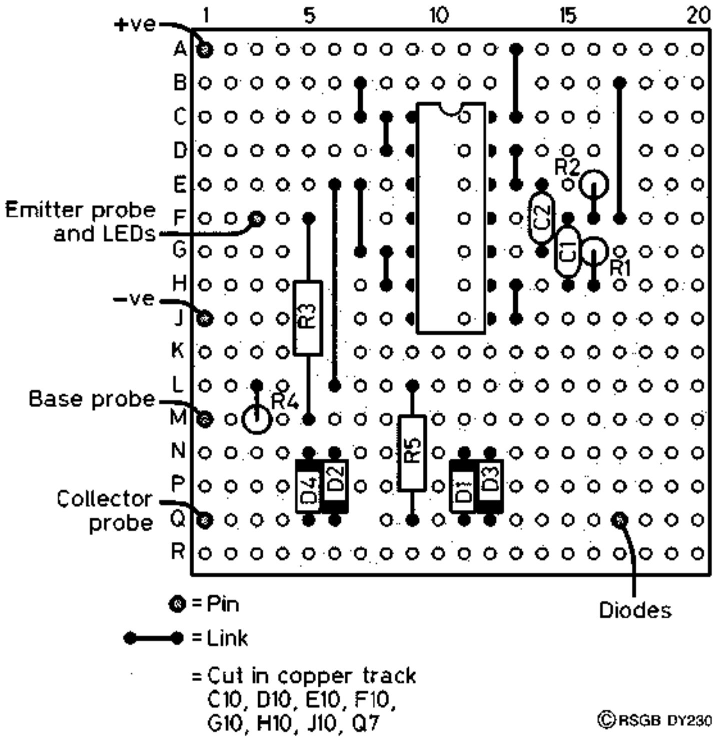
1
5
10
15
20
+ve
A
B
C
D
E
F
G
H
J
K
L
M
N
P
Q
R
Emitter probe and LEDs
−ve
Base probe
Collector probe
R1
R2
R3
R4
R5
C1
C2
D1
D2
D3
D4
Diodes
= Pin
= Link
= Cut in copper track
C10, D10, E10, F10,
G10, H10, J10, Q7
©RSGB DY230

Figure 2 Transistor tester, component layout

track ‘I’, so try to avoid miscounting when you translate diagram positions to real positions on the board. Hold the board up to a strong light to ensure that there is no copper swarf shorting adjacent tracks together, and that you have made the cuts in the correct places.

Having done this, insert and solder Veropins for all the connections to be made to components not on the board itself. Then solder in the IC holder and the wire links. Then insert the other components in the order resistors, capacitors and diodes D1–D4. Some resistors are mounted vertically so that their connections are on adjacent tracks. Double check the diode polarities – it is easy to make a mistake when wiring diodes in anti-parallel! Next, connect up the off–board components, again making sure that the LEDs have the correct polarity. The probe leads for the emitter, base and collector should be made from different colours of wire and terminated in probe clips (small insulated crocodile clips).

Check carefully for dry joints and errant blobs of solder. Plug the IC into its holder, ensuring that it is inserted the right way round, as shown in Figure 2.

Testing

Without a transistor in circuit, and the battery connected, both LEDs should be lit. Connect a known good npn transistor and verify that the green LED lights. Now simulate two transistor faults: disconnect the base lead and both LEDs should light; remove the transistor and connect the emitter and collector leads together. Neither LED should light.

Repeat the tests with a known good pnp transistor. The results should be the same, except that the correct indication should now be a lit red LED. On your computer, make a small label of the bulleted list above, to fit on your tester showing the states of the LEDs and what they mean. It will act as a useful *aide mémoire* when you use the tester in future.

Using

The circuit will test transistors in isolation or in an existing circuit, i.e. prior to use. You can check the lead identifications in component catalogues such as the Maplin catalogue. The tester is ideal for going through the large bags of unmarked transistors that you can buy for a song at rallies. You can sort them into three piles – npn, pnp and dud!

Parts list

Resistors: all 0.25 watt, 10% tolerance or better

R1, R2	5600 ohms (Ω)
R3, R4, R5	100 ohms (Ω)

Capacitors

C1, C2	1 microfarad (μF) electrolytic, 16 V WKG

Semiconductors

D1–D4	1N4148 general-purpose silicon diodes
D5	Red LED
D6	Green LED

Integrated circuit

IC1	CMOS 4001 or 4011

Additional items

Veroboard, 15 strips by 18 holes
Veropins
PP3 battery and connector (or 6 × AA cells in PP3 clip holder)
Switch, SPST
Connecting wire
One each of known working npn and pnp transistors for test purposes.

Source

Components are available from Maplin.

70 The ‘Yearling’ 20 m receiver

Introduction

Published to celebrate the first anniversary of *D-i-Y Radio*, this excellent receiver design forms a suitable ‘second receiver project’ for those who have successfully completed the MW receivers earlier in this series. The receiver is powered from a PP3 battery or from a mains adaptor, and can be built with the help of an experienced constructor, on a prototype board. The circuit diagram and some of the components used are shown in the separate diagram. Headphones or a loudspeaker can be used and, once the radio is completed, a few simple adjustments will make the *Yearling* spring to life!

Building the receiver

Before starting the constructional process, start by identifying all the parts. One by one, tick them off against the parts list. Are their values correct? The varactor diode is a twin type (see circuit diagram overleaf), and must be cut *carefully* down the middle, producing two devices, D1 and D2, with two wires each.

First, solder the IC sockets, followed by the coils (inductors); L1 is pink inside the top, and L2 is yellow inside. Then, solder in the varactors; the lettering on D1 should be next to coil L1, and the lettering on D2 should face resistor R7. After those, the capacitors, wire links and resistors should be soldered to the board. Take care to wire the voltage regulator, IC3, correctly. Solder in the crystal X1 as quickly and deftly as you can – crystals do *not* take kindly to having their leads bent and being fried with a soldering iron! Make sure that the electrolytic capacitors C2, C12, C15 and C16 are fitted the right way round. Most electrolytic capacitors have only the negative lead marked.

Figure 1 shows the rear of the front panel, illustrating the connections from the board and antenna socket to the controls. All normal connecting wires are 22 SWG or thereabouts, with insulation. Their lengths should be about 15 cm, except for the battery lead to the switch, which is about 8 cm. It is recommended that you use different-coloured wire for each connection to a control. **Figure 2** shows this. The variable resistor section of VR5, the AF gain control, uses single screened cable connected to 0 V (ground) at the

Amateur Radio

D-i-Y

Year

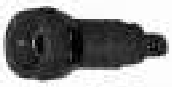

Antenna Socket

Variable Inductor

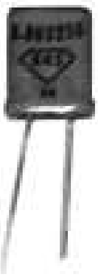

Crystal

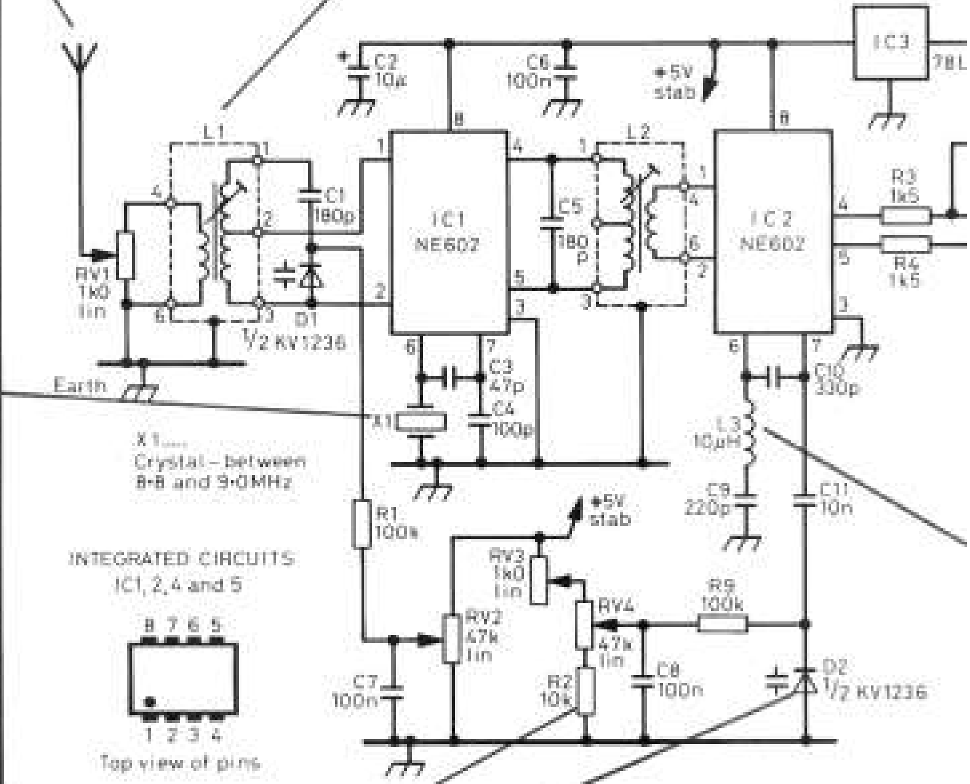

L1
1
2
3
4
6
RV1
1k0
lin
C1
180p
D1
1/2 KV1236
Earth
C2
10µ
C6
100n
+5V
stab
8
1
2
4
5
3
6
7
IC1
NE602
C5
180
p
L2
IC3
78L
IC2
NE602
R3
1k5
R4
1k5
C3
47p
C4
100p
X1
X1 .....
Crystal – between
8·8 and 9·0MHz
C10
330p
L3
10µH
C9
220p
C11
10n
R1
100k
+5V
stab
RV3
1k0
lin
RV4
47k
lin
R9
100k
RV2
47k
lin
C7
100n
R2
10k
C8
100n
D2
1/2 KV1236
INTEGRATED CIRCUITS
IC1, 2, 4 and 5
8 7 6 5
1 2 3 4
Top view of pins

Resistor

Varicap Diode

Most radio amateurs enjoy building par
is Do-it-Yourself (D-i-Y) Radio, and it is
successfully built yourself. Licensed
allowed (by the Government) to use ho
can see how radio components are dra
circuit diagram for the

D-i-Y Radio, RSGB, Lambda House, Cra

– Do it Yourself

RADIO

rling

Polyester Capacitor

Electrolytic Capacitor

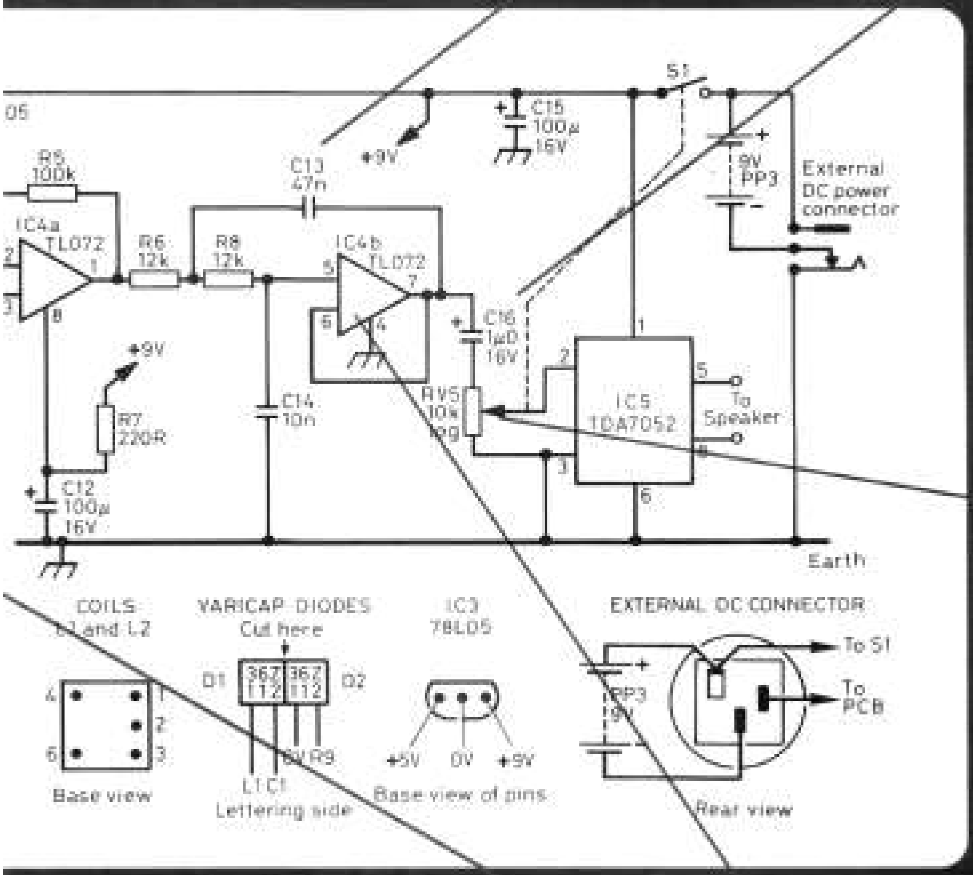

05
R5
100k
IC4a
TL072
2
1
3
8
R6
12k
R8
12k
C13
47n
+9V
C15
100µ
16V
S1
+
9V
PP3
–
External
DC power
connector
A
IC4b
TL072
5
7
6
4
+9V
R7
220R
C12
100µ
16V
C14
10n
C16
1µ0
16V
RV5
10k
log
1
2
IC5
TDA7052
5
To
Speaker
8
3
6
Earth
COILS
L1 and L2
4
1
2
6
3
Base view
VARICAP DIODES
Cut here
D1
367
112
367
112
D2
L1 C1
0V R9
Lettering side
IC3
78L05
+5V
0V
+9V
Base view of pins
EXTERNAL DC CONNECTOR
+
PP3
9V
To S1
To
PCB
Rear view

AF Gain Control

of their radio stations themselves, this
very satisfying to use a radio you have
radio amateurs are the only people
me-built transmitters. On this page you
wn on a circuit diagram. This is the full
D-i-Y Radio Yearling.

Fixed Inductor

Integrated Circuit

nborne Road, Potters Bar, Herts. EN6 3JE

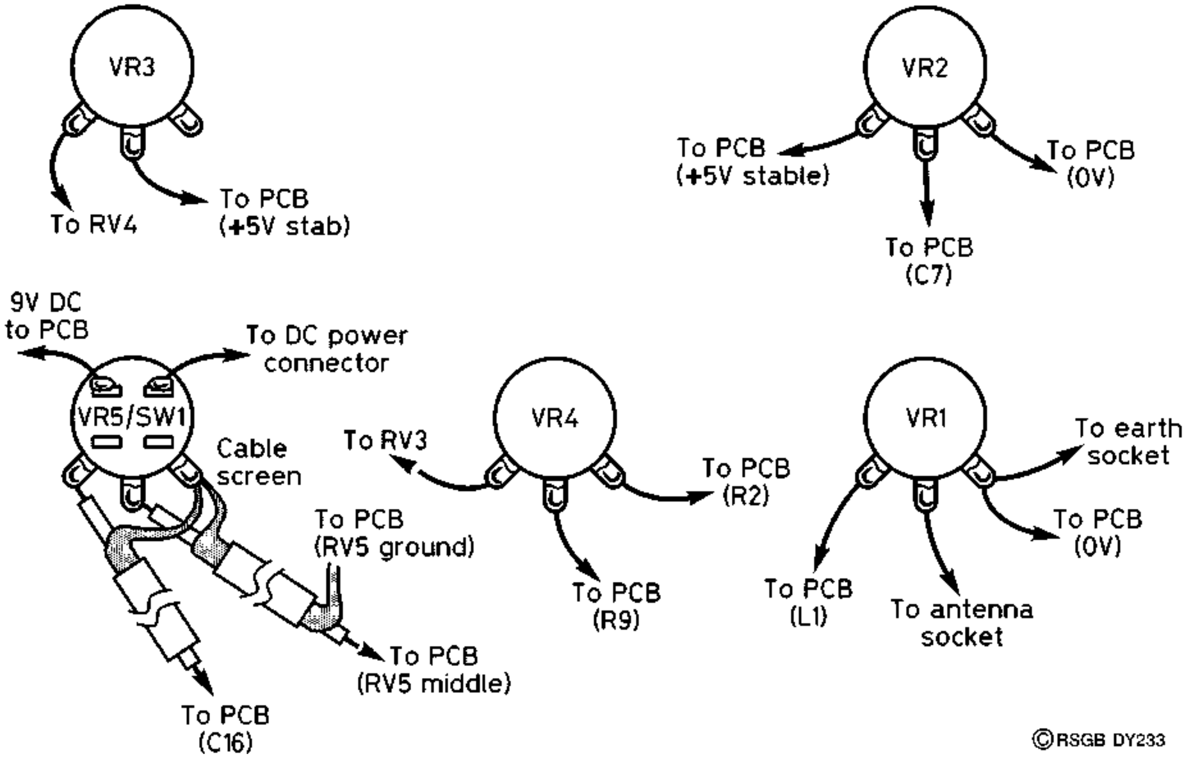
VR3
To RV4
To PCB
(+5V stab)
VR2
To PCB
(+5V stable)
To PCB
(0V)
To PCB
(C7)
9V DC
to PCB
To DC power
connector
VR5/SW1
Cable
screen
To PCB
(RV5 ground)
To PCB
(RV5 middle)
To PCB
(C16)
To RV3
VR4
To PCB
(R2)
To PCB
(R9)
VR1
To earth
socket
To PCB
(0V)
To PCB
(L1)
To antenna
socket
©RSGB DY233

Figure 1 Rear view of the variable resistors. Check the connections carefully to make sure the wires fit the correct holes

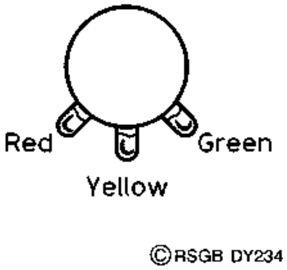
Red
Green
Yellow
©RSGB DY234

Figure 2 It is helpful to use different colour insulated connecting wires. Wires between each variable resistor and the board should be twisted together to give a neat wiring outfit.

board end. Now fit the ICs into their sockets. Make sure that they are the right way round (see circuit diagram) and that each pin lies directly above its corresponding socket before applying *gentle* pressure with the back of the board firmly supported.

Check that all the connections are correct (don't *assume* this – check the ends of each wire against the circuit diagram) and that all your soldered joints are shiny. Lastly, drill the five 10.5 mm diameter holes for the main controls. On the side of the case, a 6.3 mm diameter hole is needed for the speaker socket, and 8 mm holes for the antenna and earth sockets. The external power socket requires an 11 mm diameter hole.

Adjusting it

Before you can make adjustments, and in order to hear *anything* on your *Yearling*, you will need to connect an antenna (aerial) to the antenna socket. About 8 metres of wire, preferably outdoors and as high as possible, is all you need to connect to the socket. Connect the 3.5 mm jack socket for the speaker, and a 3.5 mm jack to your speaker leads. Connect a battery and switch on.

1. Using a very small screwdriver or, better still, a non-metallic 'trimming tool', *gently* screw in the core of L2 as far as it will go, but *don't force it*. Then unscrew it by three turns anticlockwise.
2. Set VR1, VR2 and VR4 to mid-position and rotate the core of L1 anticlockwise until the hissing noise you hear reaches a maximum intensity. Then adjust L2 for maximum noise.
3. If you now tune carefully with the main tuning control, VR4, you should hear some amateur Single-Sideband (SSB) speech signals. You may have to adjust the bandspread (fine tuning) control, VR3, to make the speech sound normal.
4. Having verified that everything is working, switch off and mount the controls on the front panel and the sockets on the side. To do this, it is much safer to disconnect all the controls and sockets, mount them in their final positions, and then wire them up again.
5. Fit the front panel knobs, connect your aerial and switch on again, checking that everything is working. Then, locate the cluster of CW (Morse) signals to be found at the bottom of the 20 metre band, set the bandspread control to mid-position and slacken off the main tuning knob. Turn the knob (but *not* the control!) until the pointer lies a little clockwise of 14.0 MHz. The SSB signals should now lie roughly between the dial centre and 14.35 MHz. Tighten up the knob.
6. Finally, fix the board to the rear panel, and secure the battery (if you are using one). Attach the rear panel to the back of the box, and you are finished!

Listening!

Make a habit of keeping the bandspread control in its centre position when searching for stations; then you can adjust it either way to make the signals readable. Unless you have a very big aerial, it is best to have the 'RF Gain' control, VR1, at maximum. Use the 'Antenna Tune' control, VR2, to give the best signal, and control the volume with the 'OFF/AF Gain' control. You will find some excellent DX stations with your *Yearling* receiver, and it will serve you well.

Parts list

Resistors (all 0.25 W, 5%)

R1, R5, R9	100 kilohms (kΩ)
R2	10 kilohms (kΩ)
R3, R4	1.5 kilohms (kΩ)
R6, R8	12 kilohms (kΩ)
R7	200 ohms (Ω)

Capacitors (all rated at 16 V or more, tolerance *at least* what is quoted)

C1, C5	180 picofarads (pF) polystyrene 5%
C2	10 microfarads (μF) electrolytic
C3	47 picofarads (pF) polystyrene 5%
C4	100 picofarads (pF) polystyrene 5%
C6, C7, C8	100 nanofarads (nF) or 0.1 microfarad (μF) ceramic
C9	220 picofarads (pF) polystyrene 2%
C10	330 picofarads (pF) polystyrene 2%
C11, C14	10 nanofarads (nF) or 0.01 microfarad (μF) ceramic
C12, C15	100 microfarads (μF) electrolytic
C13	47 nanofarads (nF) or 0.047 microfarad (μF) polyester, 5%
C16	1 microfarad (μF) electrolytic

Variable resistors

VR1, VR3	1 kilohm (kΩ) linear
VR2, VR4	47 kilohms (kΩ) linear
VR5	10 kilohms (kΩ) log with switch

Inductors

L1	Toko KANK3335R
L2	Toko KANK3334R
L3	10 microhenries (μH), 5%

Semiconductors

IC1, IC2	Philips/Signetics NE602 or NE602A
IC3	78L05 5 V 100 mA regulator
IC4	TL072 Dual Op-Amp
IC5	Philips TDA7052 audio amplifier

Additional items

D1, D2	Varactor diode Toko KV1236 (cut into two sections – see text)
X1	Crystal 8.86 MHz type (from Maplin, etc.)
4 off	8-pin DIL sockets for IC1, IC2, IC4 and IC5
2 off	4 mm sockets aerial (red) and earth (black)
1 off	3.5 mm chassis-mounting speaker jack socket
1 off	DC power socket for external supply (if required)
4 off	Red knobs with pointers
1 off	Tuning knob with pointer (e.g. 37 mm PK3 type)
1 off	Printed-circuit board or prototype board
1 off	Plastic case approx 170 × 110 × 6 mm (e.g. Tandy number 270-224)
1 off	Speaker 8–32 Ω impedance (or headphones)

71 Adding the 80 metre band to the Yearling receiver

Background

You will have noticed that your *Yearling* receiver has a dial which shows coverage of the 80-metre amateur band (3.5–3.8 MHz). This band is used for local contacts during the day, and contacts up to about 1600 miles in darkness. Longer distances are possible, particularly in the middle of winter.

The modifications

Only a few extra parts are required, as you may have noticed from the parts list. A low-pass filter, FL1 (one which passes low frequencies and rejects higher frequencies), is switched into the circuit on 80 m. The circuit of the

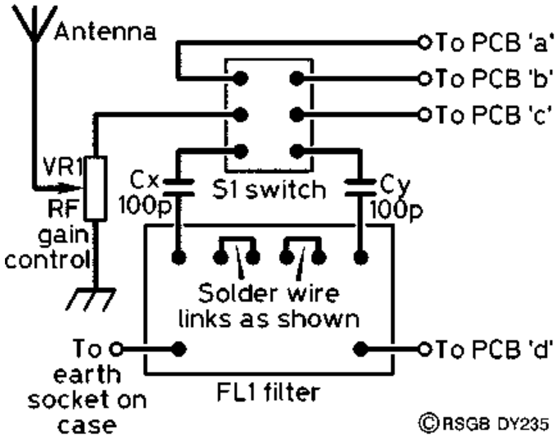
Antenna
To PCB 'a'
To PCB 'b'
To PCB 'c'
VR1
RF gain control
Cx 100p
S1 switch
Cy 100p
Solder wire links as shown
To earth socket on case
To PCB 'd'
FL1 filter
©RSGB DY235

Figure 1 The circuit diagram shows the extra components for 80 m operation. Note the connections to the PCB 'a', 'b', 'c' and 'd'

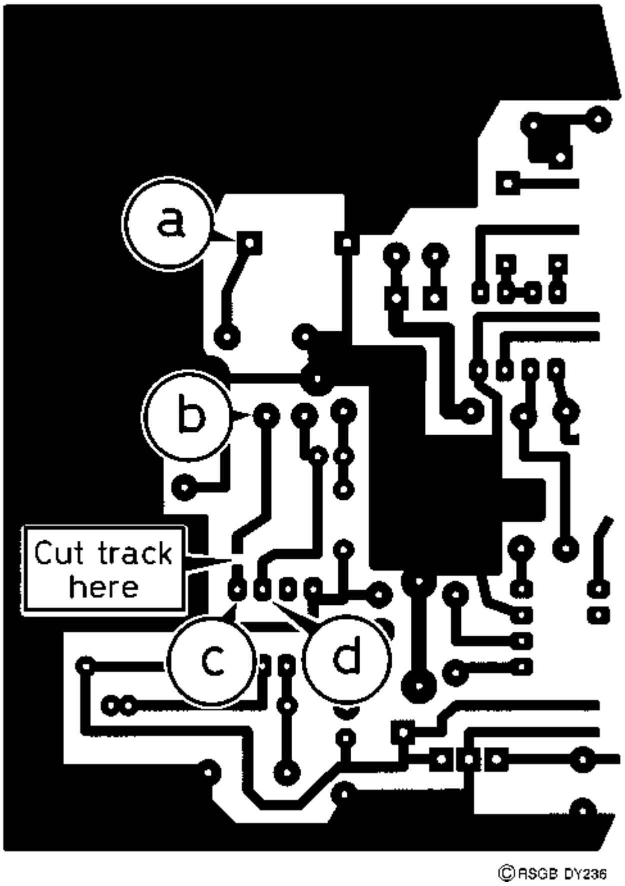
a
b
Cut track here
c
d
©RSGB DY236

Figure 2 The underside of the PCB. Wires are connected from the switch and filter as shown

switch and its connections is shown in **Figure 1**. Before making the modifications, dismantle the receiver so that you have easy (and safe) access to the case and the track side of the PCB.

1. The first thing to do is to drill a 6.5 mm diameter hole in the side of the case into which the switch fits.
2. Then, using a sharp Stanley knife or scalpel, carefully cut the track on the PCB as shown in **Figure 2**, making a gap about 1 mm wide.
3. Using 10 cm lengths of different-coloured insulated wire, make the four connections, a, b, c and d, to the PCB, as shown in Figure 2.
4. Solder the two links on the filter, and then make the connections to the switch, capacitors and PCB. You will have to disconnect the existing wire between the RF gain control, VR1, and pin 4 of L1 on the PCB.
5. Lastly, check your new connections carefully, then mount the switch in the new hole and fix the filter to the bottom of the case with a little glue, as shown in **Figure 3**. Reassemble the circuit, and replace the back of the case.

More testing!

Firstly, switch your new switch, S1, into the 20 m position, to check that the original circuit still works! If you find that the ‘Antenna Tune’ control peaks at a slightly different position, don’t worry.

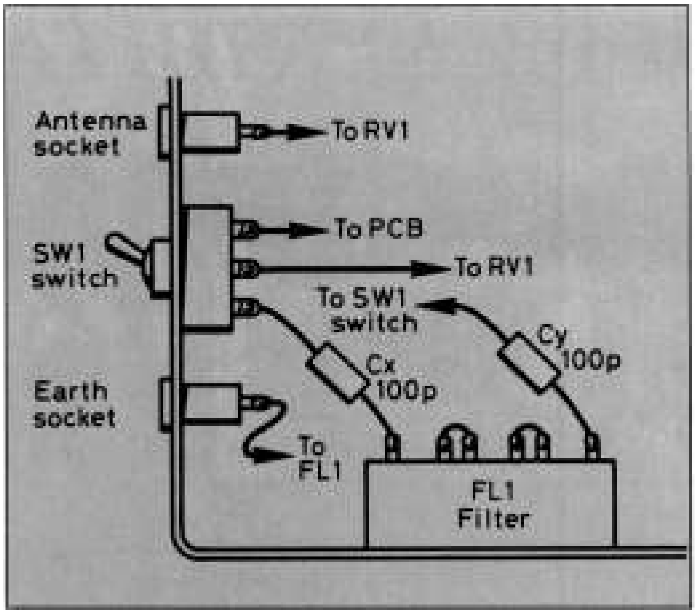

Figure 3 Internal view of the Yearling case. The filter FL1 is attached to the base with glue

Now switch to 80 m, and tune around the anticlockwise end of the dial; you should hear some SSB stations, particularly in the evenings and at weekends, when may people are on the air. At the other end of the travel of the tuning control, you should hear CW (Morse) stations.

The Radio Society of Great Britain broadcasts amateur radio news every Sunday morning on or about 3.65 MHz; the table below has the details. Finally, a good antenna is more important than ever for 80 metre reception – aim for more height and length, and then consider the project concerned with making an Antenna Tuning Unit (ATU)!

Parts list

Capacitors (all rated at 16 V or more, tolerance 10% or better)

Cx, Cy	100 picofarads (pF) polystyrene

Filter

FL1	Toko 237LVS1110 low-pass filter

Additional items

SW1	2-pole 2-way (changeover) toggle switch
7 off	Short lengths of insulated wire of different colours

The RSGB news broadcasts, GB2RS – Sunday mornings

Frequency (MHz)	*Local time*	*Reception area*
3.650	0900	SE England
3.650	0930	Midlands
3.650	1000	SW England
3.650	1100	Yorkshire
3.640	1130	Aberdeen
3.660	1130	Glasgow

The Midlands transmission is repeated at 1800 (6pm) local time on 3.650 MHz. All frequencies are approximate in order to avoid interference, and use lower sideband (LSB). If you also have a 40 metre receiver, there are GB2RS news broadcasts on 7.048 MHz at 0900 local time from Northern Ireland and from 1100 local time from the north midlands.

72 How the Yearling works

Introduction

The *Yearling* was designed to provide an introduction to *Amateur Radio* on the 20 m amateur band. Let's look at how the different sections (or 'stages') of a radio work, and how they fit together to form a complete receiver. **Figure 1** shows a block diagram which you can follow and compare with the circuit diagram of your *Yearling* receiver.

The antenna (or aerial)

Connected to your receiver, it will pick up not only amateur signals, but all other signals as well! This means that the receiver has to select the one signal that interests you, while rejecting all the others. The following stages do just that.

The RF filter

This stage (centred around L1) selects the band of radio frequencies (RF) containing the signal you want, in this case, those having wavelengths around 20 m. Signals from the 40 m band, for example, would not get through.

The crystal oscillator

This is an oscillator circuit designed around a quartz crystal (X1), and has a very stable frequency. It produces a single, very pure frequency to feed into the mixer. A crystal having a frequency between 8.800 MHz and 9.000 MHz is suitable for this circuit. The oscillator and mixer functions are both carried out inside IC1.

The first mixer

Yes, this stage 'mixes' two signals together. In this case, the two signals are (i) from the aerial via the RF Filter, and (ii) from the crystal oscillator. Two

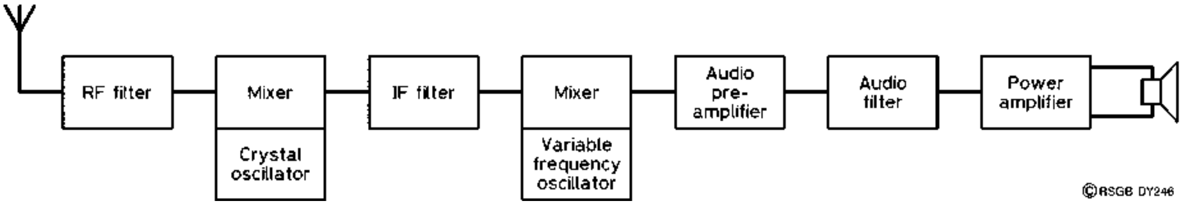

Figure 1 Block diagram of the Yearling, showing how the various stages fit together to make a complete radio receiver

bands of signals emerge from the mixer. The first is centred upon a frequency equal to the incoming signal frequency **added** to the crystal frequency, and the second is centred upon a frequency equal to the incoming signal frequency **subtracted** from the crystal frequency. Look at an example – if the signal is at 14 MHz and the oscillator at 9 MHz, then the mixer outputs will be 14 + 9 = 23 MHz and 14 – 9 = 5 MHz.

Intermediate frequency (IF) filter

It is the purpose of the IF filter (centred around L2) to select only one of these two bands of frequencies emerging from the mixer. In this case, it is the lower band of frequencies (around 5 MHz) which we select. This is because, in general, lower frequencies are easier to handle than higher ones.

Variable-frequency oscillator (VFO)

The VFO (part of IC2) enables us to tune into a particular station, and operates over a band of frequencies between 5 MHz and 5.35 MHz in this receiver. You will notice that IC1 and IC2 are the same type of chip, so that you will be expecting another mixer stage to be associated with the VFO. You are quite right!

The second mixer

This mixer obeys exactly the same rules as those of mixer 1. Sum and difference frequencies are produced, like this. Mixing is between the incoming IF signals (around 5 MHz) and the VFO signals (around 5 MHz), producing output frequency bands centred upon 10 MHz and 0 MHz. The use of the words ‘band of frequencies’ throughout this explanation is intentional. If all the signals were pure, there would be no bands; the bands are produced because of one thing – the modulation imposed on the pure frequencies at the transmitter. So, the ‘bands’ contain the one thing that we want to extract from the signal, and that is the speech or Morse code that the signal contains. The band of frequencies at 0 MHz is just that – the audio frequencies we want in the loudspeaker. Because of this, the audio frequency output of the second mixer is selected and passed on for amplification.

The audio preamplifier

Preliminary amplification of the minute audio signal which emerges from the second mixer is provided by IC4a, which will respond only to audio signals, automatically rejecting the 10 MHz signal.

The audio filter

The bandwidth of normal speech when transmitted by an amateur station is around 3 kHz, so there is no advantage to be gained in amplifying frequencies greater than this. IC4b is known as a low-pass filter, because it passes (lets through) lower frequencies and rejects higher ones.

The power amplifier

IC5 produces the final audio amplification and provides enough power (about 350 milliwatts (mW) to drive a small speaker.

How does it work on 80 m?

If you have fitted the 80 m modification to your receiver, you are probably wondering how the circuit works at this different frequency. Firstly, the filter which you fitted selects the 80 m band instead of the 20 m band. The only other slight difference lies in the way the first mixer stage works. Its job is to produce sum and difference frequencies from the incoming signal and crystal frequencies. On 20 m, it did this by subtracting the crystal frequency (9 MHz) from the incoming frequency (14 MHz) to produce an IF output of 5 MHz. On 80 m, the incoming frequency (3.5 MHz) is subtracted from the crystal frequency (9 MHz) to produce an IF output of 5.5 MHz, which is still within the tuning range of the VFO in the next stage.

73 A field strength meter

Introduction

Having a reliable field strength meter is always useful. There are two basic types – the tuned variety (also known as an absorption wavemeter) and the broad-band variety. A project centred on the tuned variety can be found in this book (*An HF absorption wavemeter*), and the way it works is also presented (*The absorption wavemeter*). This project describes an untuned or broad-band type which is every bit as useful as the tuned type, when it is the level of RF in which you are principally interested.

With the possible exception of the meter, you may have all the parts needed in your junk box. The prototype was distinguished by its meter, found at a rally, and marked 'Safe', 'Dangerous' and 'Explosive' on its scale.

How it works

The basic field strength meter uses the circuit of a crystal set, but with a meter replacing the headphones. A better design, which is used here, is that of a voltage doubler, giving more sensitivity. **Figure 1** illustrates the voltage doubler circuit.

The AC input shown will be our RF input, which will be explained soon. Diode D1 will pass the positive half of the signal and use it to charge up C1 to the peak value of the signal. D2 uses the negative half of the signal to charge up C2 to the same value. Because C1 and C2 are in series, the peak voltage appearing across *both* of them (which is the DC output voltage) is equal to *twice* the peak input voltage, hence the name voltage doubler. If you're wondering why DC is present at the output when AC comes in at the

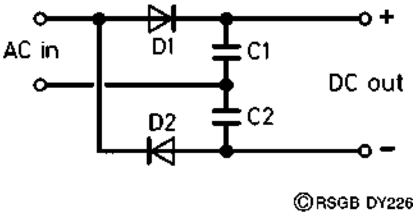

Figure 1 Voltage doubler circuit

input, remember that capacitors pass AC (RF) but block DC; thus the RF is shorted out through C1 and C2, but the steady voltage (DC) remains across the two capacitors. The voltage is thus proportional to the size of the RF signal applied at the input.

The circuit

The voltage doubler is converted into a field strength meter using the circuit of **Figure 2**. A piece of wire serves as the aerial to provide an RF signal across the radio-frequency choke (RFC). A choke is an inductor which is large enough to prevent the RF passing through it – it 'chokes' the RF. This produces the maximum RF signal at the input to the voltage doubler, and the DC output from it is measured on the meter.

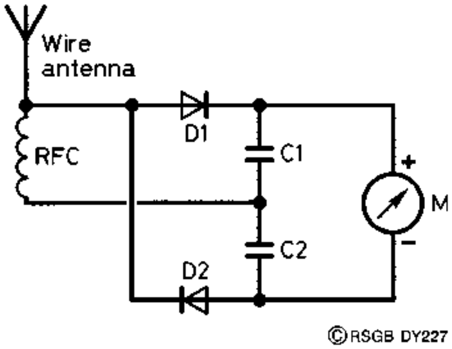

Figure 2 Field strength meter

The parts used are not very critical. The RF choke can have any value between about 1 mH and 2.5 mH. Almost any common diode such as the 1N914 or 1N4148 can be used for D1 and D2. The two capacitors could be any value between 1 nF and 100 nF (0.001 μF and 0.1 μF). The meter should be reasonably sensitive, with a full-scale deflection (FSD) in the range 50 μA to 100 μA. Look for VU meters at rallies – these are ideal.

Construction

The prototype circuit was made on matrix board. If you don't want to use pins with the board, simply push the component leads through the holes and make the connections on the underside of the board, either with the excess component leads themselves, or with ordinary connecting wire.

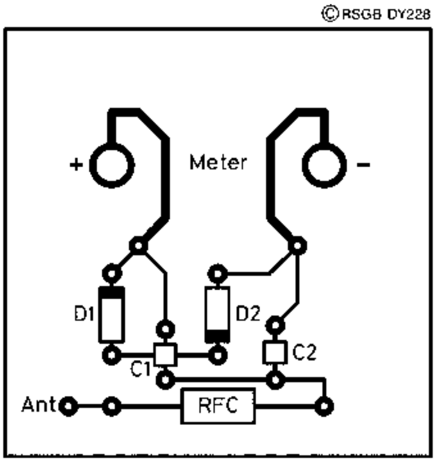

Component side of board shown.
All connections are under board

Figure 3 Board layout and interconnections

Figure 3 shows how the parts are placed and connected. The matrix board can be mounted directly on the meter using its terminal bolts!

In use

Needing no power supply other than an RF signal, just connect it all up and leave it to work! A short length of insulated wire is enough to pick up some RF and display it on the meter. Using a 200 μA meter, about 3 metres of wire gave a good deflection on the meter. To increase deflection, the wire can be wrapped around the aerial lead, provided that the wire is PVC covered and doesn't come into contact with the aerial wire.

This requires a little experimentation. Try a long piece of wire first and adjust its position until the meter needle kicks whenever there is a transmission. It is very reassuring to see the meter moving during a transmission. Although SWR meters also indicate power, they are usually set to read reverse power, and show little or no movement during transmission.

Parts list

D1, D2	1N914, 1N4148 or similar
C1, C2	10 nF disc ceramic
RFC	Miniature axial choke (1 mH)
Meter	Surplus VU meter or similar
Matrix board or similar	

Components are available from Maplin.

74 Preselector for a short-wave receiver

Introduction

A preselector is a simple RF tuned amplifier which is inserted between the aerial and the receiver. It provides some extra gain and may improve the overall performance of the receiver. This project uses a Field-Effect Transistor (FET) amplifier in grounded-gate mode.

The design has a tuned circuit at both the input and output which, with excessive gain and poor construction, would produce only one thing – oscillation! So, to avoid this happening, we will have only a low gain, and use a circuit which provides good isolation between input and output. The grounded-gate FET amplifier fulfils both these criteria. It will also cover a frequency range from about 7 to 30 MHz, which includes most of the HF amateur bands.

The circuit

This is shown in **Figure 1**. The signal from the aerial arrives at an RF transformer, the secondary of which is tuned with capacitor VC1a. The output from the tuned circuit is taken from a tap on the secondary to the source of the FET. The gate is grounded (earthed) and the amplified signal appears at the FET drain, which is then fed to the primary of another RF transformer, which is tuned by VC1b. The output to the receiver comes from the secondary of the RF transformer.

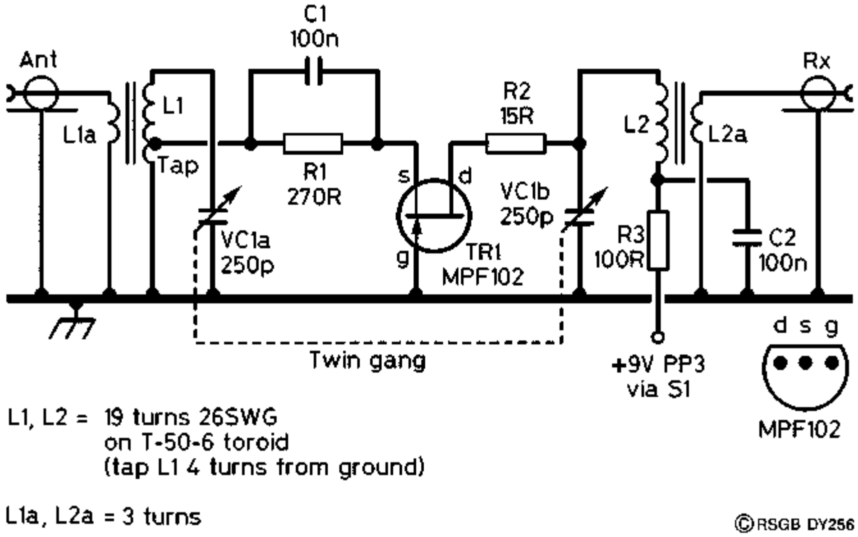
C1
100n
Ant
Rx
L1
L1a
Tap
R2
15R
L2
L2a
R1
270R
s
d
VC1b
250p
VC1a
250p
g
TR1
MPF102
R3
100R
C2
100n
d s g
Twin gang
+9V PP3
via S1
MPF102
L1, L2 = 19 turns 26SWG
on T-50-6 toroid
(tap L1 4 turns from ground)
L1a, L2a = 3 turns
©RSGB DY256

Figure 1 HF preselector, circuit diagram

Notice that the two RF transformers are identical, but they are used ‘back to back’, with the secondary of the first and the primary of the second being tuned. They are tuned with identical capacitors, fitted on the same shaft of a variable capacitor. We say that the two capacitors are ‘ganged’. Because L1 and L2 are the same, and VC1a and VC1b are the same, both RF transformers should be resonant at the same frequency, no matter what that frequency is.

Construction

The final layout should look something like that shown in **Figure 2**. The external connectors and controls being two SO-239 sockets for connection to your receiver and aerial, a tuning control and its associated scale, and an on/off switch.

The circuit can be put together on a plain matrix board, using pins to anchor the components, or simply by pushing the component leads through the board and making connections on the underside. The layout of the prototype is shown in **Figure 3**. Mounting the board to the aluminium box is accomplished with bolts, solder tags and stand-off insulators.

Check your construction against the circuit diagram and against the layout diagram. Wire in the PP3 battery clip, put the switch in the ‘off’ position, and fit the battery. Testing can be carried out without fitting the top of the box.

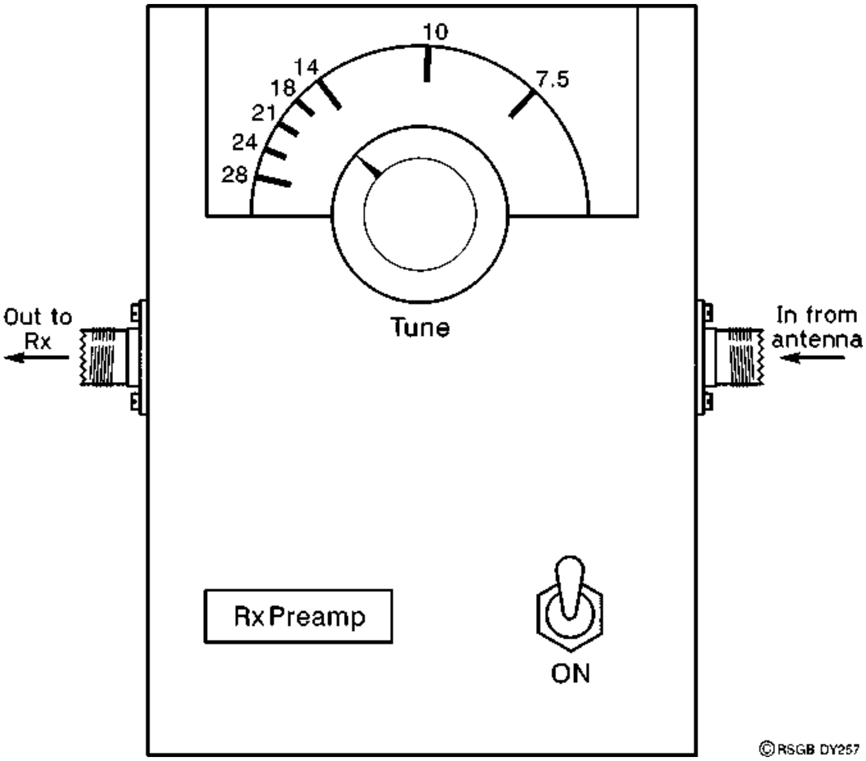

Figure 2 HF preselector, front panel layout

Testing

Don't fit your preselector yet. Tune your radio to a broadcast station, preferably a fairly weak one. Disconnect the aerial and fit your preselector between the aerial and receiver. Switch it on and rotate the tuning knob slowly. You should find a position where your original station is received more clearly than before. If it doesn't work at all, recheck your wiring. Is there a positive voltage on the drain of the FET? If not, work back towards the positive battery terminal. Is there a voltage at the junction of L2 and R2? Is there a voltage at the junction of L2 and R3? Is there a voltage at the junction of R3 and the battery lead? If there isn't a voltage at that point, then you have probably mounted your switch upside down, and it is off, not on! It's a common mistake.

Calibration

This is not obligatory, surprisingly enough. However, if having a frequency scale appeals to you, then using an RF signal generator (or using the services of a friend who has one) is the simplest solution. Feed in a weak modulated

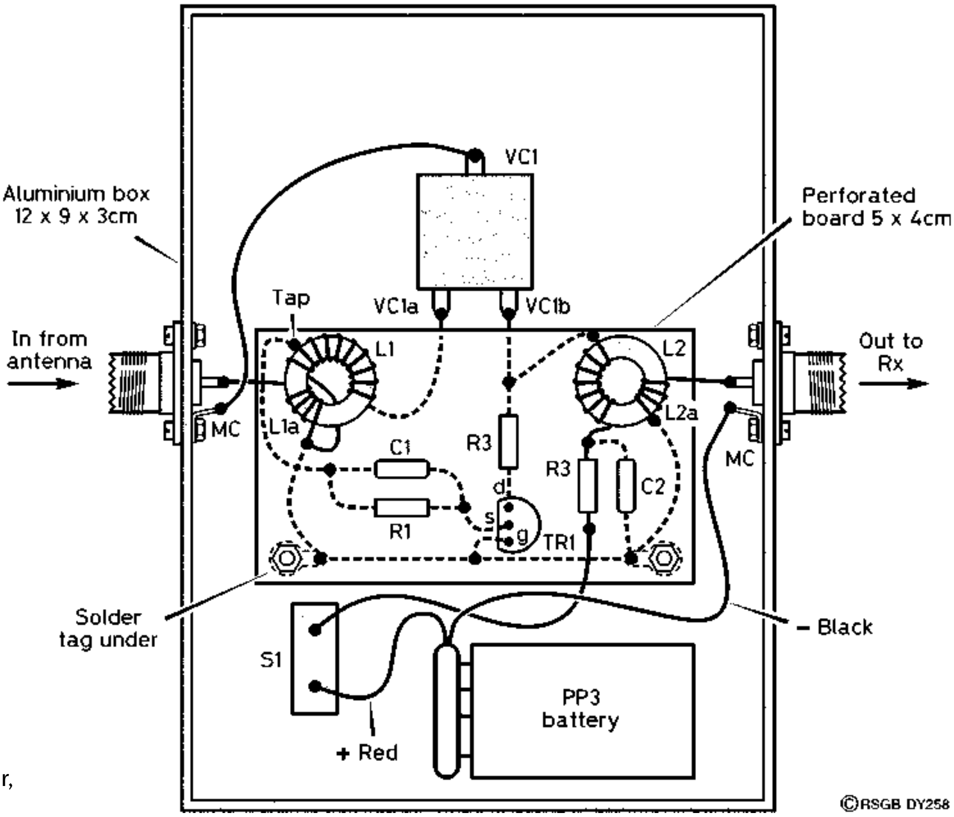
VC1
Aluminium box
12 x 9 x 3cm
Perforated
board 5 x 4cm
Tap
VC1a
VC1b
In from
antenna
L1
L2
Out to
Rx
MC
L1a
L2a
C1
R3
R3
C2
MC
d
s
g
R1
TR1
Solder
tag under
– Black
S1
PP3
battery
+ Red
©RSGB DY258

Figure 3 HF preselector, component layout

signal from the generator to the preselector, and rotate VC1 until the signal is maximum. Mark this frequency on your dial. Repeat the process for the frequencies shown on the dial in Figure 2.

If you have a commercial transceiver, feed its output into a dummy load (such as the type described in *A Switched Dummy Load*, elsewhere in this book) using a distinctive modulating signal such as an idling RTTY signal. Set the receiver to the same frequency with the preselector out of circuit. Insert it into the aerial lead, and search for the signal with VC1 until it gives the maximum deflection on your S-meter, then mark the frequency on your dial.

Parts list

Resistors: all 0.25 W carbon film or better

R1	270 Ω
R2	15 Ω
R3	100 Ω

Capacitors

C1, C2	100 nF (0.1 μF) ceramic
VC1	250/250 pF polyvaricon

Inductors

L1, L2	19 turns 26 SWG enamelled copper on T.50.6 toroid L1 has a tap 4 turns from ground end L1a and L2a – 3 turns wound over previous winding

Semiconductors

TR1	MPF102 FET

Additional items

Matrix board	To fit aluminium box (see Figure 3)
Aluminium box	12 × 9 × 3 cm
Battery and connector	PP3 9 V
S1	SPST on/off switch
Knob	As required
SO239	Sockets – 2 required

75 An audible continuity tester

Introduction

This is not the only continuity tester in this book. This alone attests to their use, so you may well want to experiment with several designs, then come up with one of your own! The very simplest form of continuity tester is probably a battery and a bulb in series, with the circuit being closed by connecting it to a fuse or other object being tested for continuity. The bulb could be replaced by a buzzer to give an audible indication. The current taken by the buzzer could damage some components, however. An ohmmeter can also be used, and is very popular for the purpose, as it indicates whether the circuit is low, medium or high resistance. This project has the advantages of indicating whether there is no continuity, some resistance or good continuity, and making an audible sound, so that you don't have to move your eyes from the circuit while making the test.

The circuit

Using only one integrated circuit, four components and a battery, this is a particularly simple circuit, as **Figure 1** shows. IC1 is usually used to flash an LED from a 1.5 V source, and to have a low current consumption. By changing the component values, IC1 is made to oscillate at audio frequencies, and we can hear these through the loudspeaker, LS1. Low-impedance speakers can be used, but result in an increased current drain, so the higher the impedance, the better.

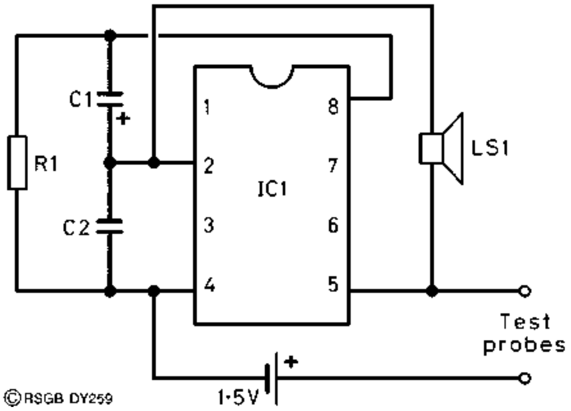

Figure 1 Continuity tester, circuit diagram

Construction

The layout of the circuit on Veroboard measuring 7 strips by 15 holes is shown in **Figure 2**. Start by making the four track cuts which will lie underneath the IC. Use a track cutter or a 3 mm (⅛ inch) twist drill rotated between thumb and forefinger for this. Solder the Veropins in place, followed by the wire link. Then solder in the IC holder, the resistor and the capacitors, making sure that the electrolytic capacitor, C1, is connected the correct way round.

Check, with the board against a bright light, that there are no shorted tracks, either by large blobs of solder or by copper swarf from the track-cutting process. Then insert IC1 into its holder the right way round. The probes can be ordinary connecting wire, the free ends being tinned with solder to prevent wire whiskers from touching components other than the one you are testing.

If you are happy that the circuit and the wiring appear to be correct, fit the battery into its clip. Nothing should happen until the probes are touched together, when you should hear a note from the loudspeaker. If nothing happens, all you can do is to recheck your circuit, as there is nothing else to go wrong!

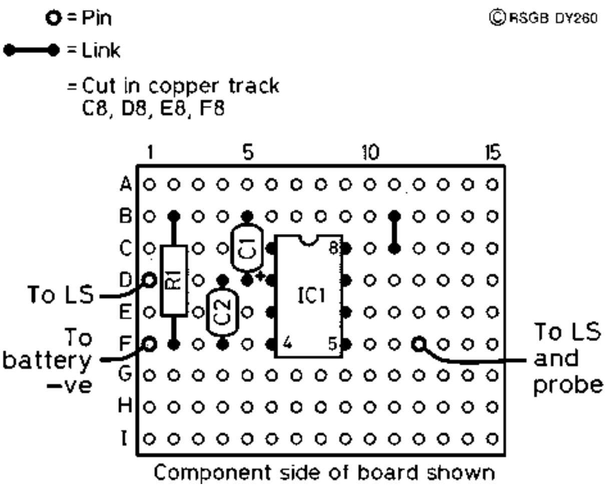

O = Pin
= Link
= Cut in copper track C8, D8, E8, F8
©RSGB DY260
1
5
10
15
A
B
C
D
E
F
G
H
I
R1
C1
C2
IC1
8
4
5
To LS
To battery −ve
To LS and probe
Component side of board shown

Figure 2 Continuity tester, component layout

Any box can be used; there is no justification for a metal case, unless you want to make use of something, which is to hand, such as a tobacco tin. Alternatively, any suitable plastic box will do.

Use

The tester will give different pitch notes for resistances of different values. The higher the resistance, the higher the pitch from the speaker. Try it with small inductors, and you will learn to recognise the different tones produced by the IC.

Safety notice

Using the continuity tester on components *in situ* is not advisable, as the results could be misleading. It can be dangerous to make measurements *in situ* on equipment, which is operating. If you must make such tests, always disconnect the equipment form its power source first.

Parts list

Resistor

R1	1000 Ω (1 kΩ), 0.25 W carbon, 10% or better

Capacitors

C1	10 μF electrolytic, 16 V
C2	0.1 μF subminiature polyester or ceramic

Integrated circuits

IC1	LM3909N

Additional items

	Veroboard 7 strips by 15 holes
LS1	64 Ω miniature loudspeaker
	IC holder
	Veropins
	Connecting wire
	Test leads
	1.5 V AA alkaline battery

76 An experimental 70 cm rhombic aerial

Introduction

Most commonly used aerials can be classed as *resonant* or *standing-wave* aerials. There is another class known as *non-resonant* or *travelling-wave*. Resonant aerials, such as the dipole, are narrow-band; this occurs because resonance occurs only over a narrow band of frequencies. Travelling-wave aerials, on the other hand, can operate over a wide band of frequencies.

The *rhombic* is an example of a non-resonant or travelling-wave aerial. It is often employed for fixed commercial and military short-wave radio links. Made with wire, it has a diamond shape when looking down on it from above. The four corners are supported on four masts. It is a very effective aerial, and has good gain, a quality which can be judged from the *polar diagram* shown in **Figure 1**. This has been obtained from a computer program, and so is the perfect shape for a rhombic aerial.

Theory (but only a little)

A polar diagram shows graphically the ability of an aerial to radiate (or receive) more effectively in one direction at the expensive of the radiation in other directions. Figure 1 shows the polar diagrams of two aerials, a simple 70 cm dipole and the rhombic described in this project. The dipole has the

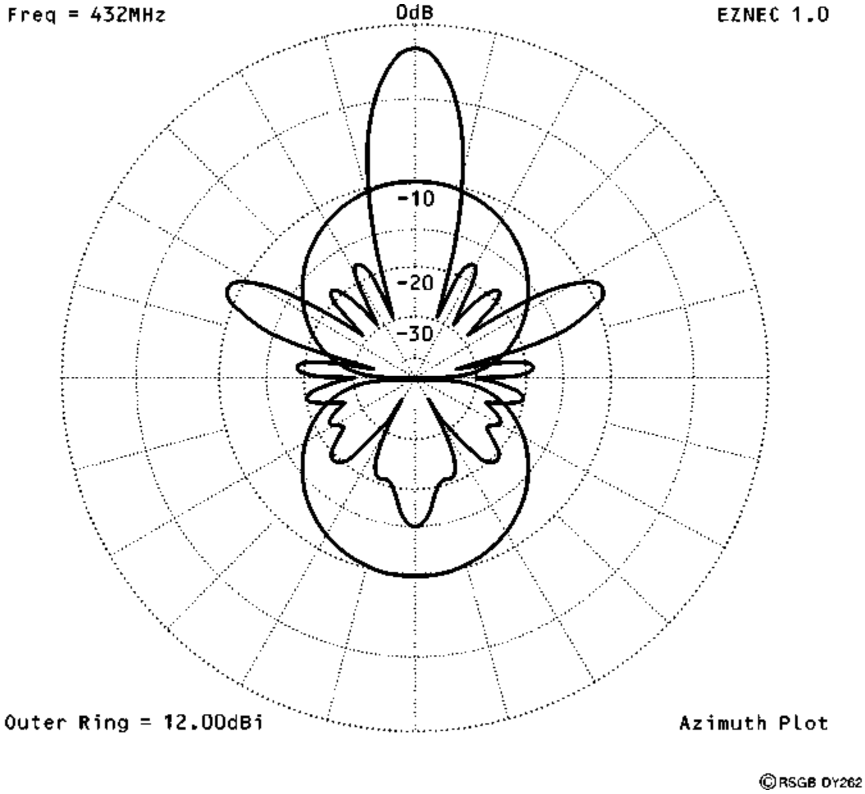

Figure 1 The horizontal polar diagrams of the small rhombic antenna and a dipole compared

well-known figure-of-eight shape, showing a symmetrical angular behaviour about the direction of the aerial and about a direction perpendicular to the aerial. That of the rhombic, on the other hand, is quite irregular by comparison, but is still symmetrical about one axis only, not two. It's this asymmetry that gives the rhombic its gain, by virtue of its *front-to-back ratio*. This is the ratio of the power radiated forwards to that being radiated backwards. Notice the large *lobe* (lump) at the top of the polar diagram; this is the direction in which the aerial transmits best. The lobe in the opposite direction has been reduced significantly, allowing more power to be directed forwards, not backwards

Problems

The rhombic is not found in every amateur's back garden, despite its attractions. To work best each edge of the diamond shape should be about two wavelengths long. Hence, a rhombic for the 20 metre band could be about 80 m from tip to tip! Another disadvantage is that, because of its size, it cannot be rotated.

It will operate well over a range of frequencies. An HF rhombic could work on the 7, 10, 14, 18, 21 and 28 MHz bands. One for the lower VHF frequencies could operate on the 50, 70 and 144 MHz bands.

Its large size is less of a problem at UHF. A portable rhombic can be made for 70 cm which can be used with a hand-held transceiver. The aerial design to be described here will give a gain of up to 9 dB relative to a dipole. This is written as 9 dBd, the second 'd' meaning 'relative to a dipole'. This is equivalent to improving your signal by 1.5 S-points at the receiver or (and you may be surprised by this) by fitting a linear amplifier to your transceiver that would take 5 W input and produce 40 W output to a normal dipole! Consider the relative costs of the two approaches to producing the same received signal. You will also receive everyone else's signals 1.5 S-points better than before!

Unlike the popular Yagi aerial, this design has no critical dimensions and can be folded up for transport by car or bicycle. It has a fairly high input (or feed) impedance, being fed usually by balanced twin feeder. To feed it with standard 50 Ω coaxial cable, a matching transformer in the form of a *balun* is required. A balun will convert the *bal*anced (symmetrical) aerial impedance to the *un*balanced and lower impedance of the coaxial cable. The details of how to build the balun, which in this case is a half-wave transformer, are included in the constructional details.

Construction

The aerial frame is made up of 1 cm × 2 cm strips of wood fixed to a plywood centre using 30 mm long M4 bolts, as shown in **Figure 2a**. The outer bolts fixing the front and side supports can be removed for folding prior to transportation, as shown in **Figure 2b**.

The wires are fixed to the front and rear supports using screw connectors, sometimes called 'chocolate block' connectors. Detail X and Y of Figure 2 show how this is done. The side supports have holes in the ends through which the wire is threaded.

The aerial must be mounted in the horizontal plane using a small shelf bracket attached to the centre plate, using the same bolts that hold the rear support. The other half of the bracket may be mounted to a vertical mast using screws or jubilee clips (hose clips).

The balun

Cut a 23 cm length of coaxial cable; this is to be our half-wave transformer. Cut and remove 2 cm from each end of the sheath. Make the braid into a pigtail at each end. Cut and remove 1 cm of the inner insulator from each

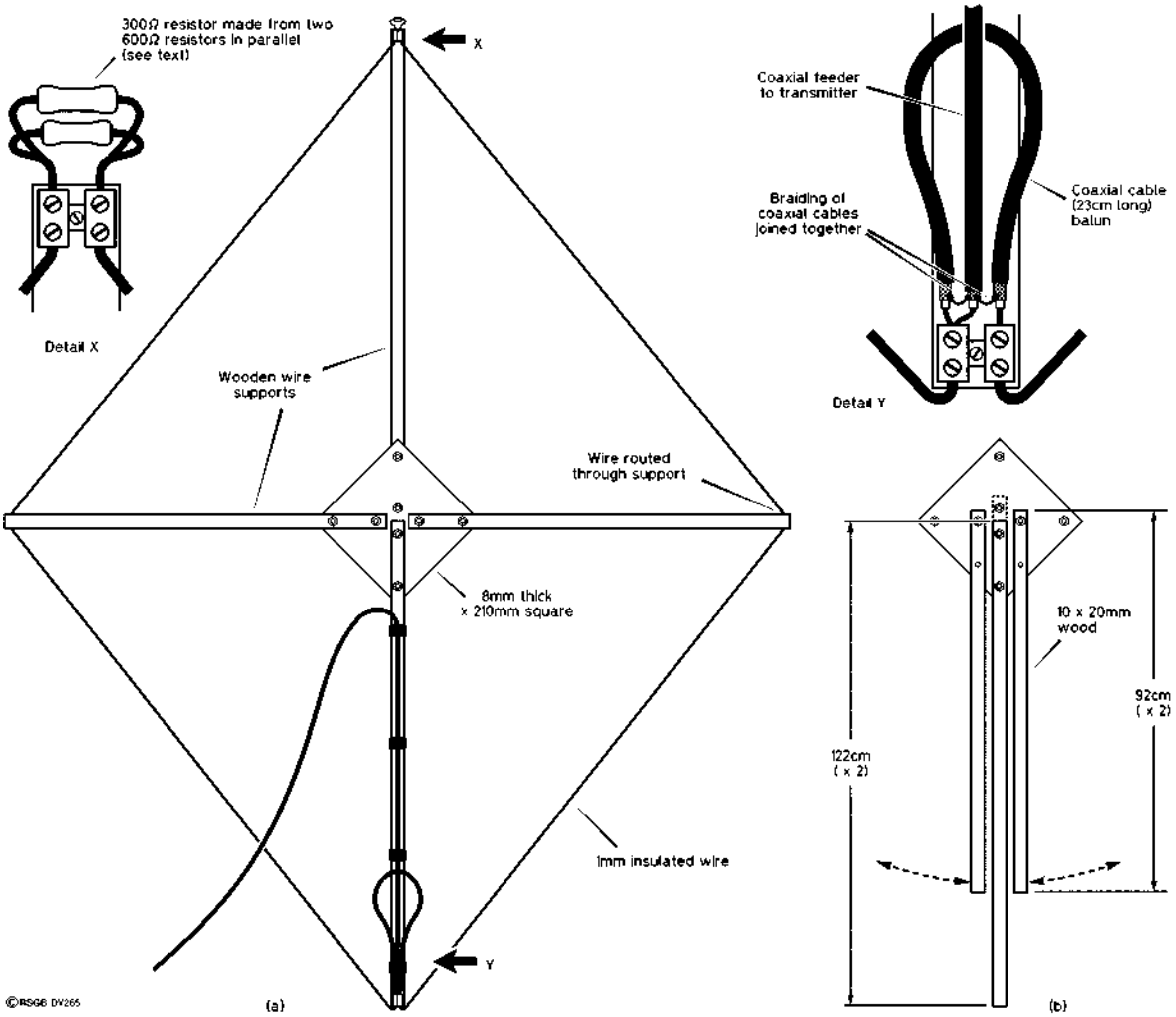
300Ω resistor made from two 600Ω resistors in parallel (see text)
X
Detail X
Wooden wire supports
Wire routed through support
8mm thick x 210mm square
1mm insulated wire
Y
©RSGB DV265
(a)
Coaxial feeder to transmitter
Braiding of coaxial cables joined together
Coaxial cable (23cm long) balun
Detail Y
10 x 20mm wood
92cm (x 2)
122cm (x 2)
(b)

Figure 2 Construction of the UHF rhombic antenna

end, leaving 1 cm of the centre conductor exposed. Fasten this piece of cable to the chocolate block connector together with the coaxial feeder cable as shown in Detail Y of Figure 2. The braids of all three prepared ends are soldered together, but are not connected to anything else. The inner conductor of the feeder from the transceiver is soldered to one end of the centre conductor of the 23 cm piece, and is connected to one end of the rhombic by one side of the chocolate block. The other end of the half-wave transformer also goes to the chocolate block, where it is connected to the other side of the rhombic loop.

The opposite corner of the rhombic, as shown in Detail X of Figure 2, shows that the loop is broken at the chocolate block, and is 'terminated' by two 600 Ω resistors in parallel. This makes the aerial a broad-band travelling-wave device, and gives it its directivity and gain. There is a rule of thumb, which includes a safety factor, which says that the terminating resistor must be able to absorb one-half of the maximum transmitter output power. So, if you use two 2 W 600 Ω carbon resistors (*not* wire-wound resistors) in parallel, you can use a transmitter with an output of 8 W, which is more than adequate for 70 cm hand-helds.

Using the rhombic

Fit the aerial to a pole or mast in the horizontal plane, if you intend to use CW or SSB, but in the vertical plane if you want to concentrate on FM work. In the latter case, a wood or fibreglass pole is mandatory. Tune to a local repeater, whose signal strength you know. Rotate the aerial to face the repeater, and you should see that the signal strength is greatly improved! Verify the directional properties of the aerial by rotating it and observing the changes in signal strength on your S-meter.

77 Water level alarm

Introduction

This is a simple device which, when you have built it, you begin to wonder how you ever managed without it! It will sound a buzzer when the level of water reaches a particular point, which you can set and vary at will.

Detecting water

Water is not too difficult to detect electrically, because it is a conductor of electricity. Not a good one, but sufficiently so to carry enough current to control a warning circuit.

The circuit is shown in **Figure 1**. If you imagine a resistor placed between the ends marked 'water detector', the circuit is then recognisable as a transistor used as an electronic switch. Whether the switch is off or on depends on the value of this resistor. If the resistance is low, the transistor switches on and the buzzer sounds. If the resistance is high, the transistor switches off and the buzzer stops.

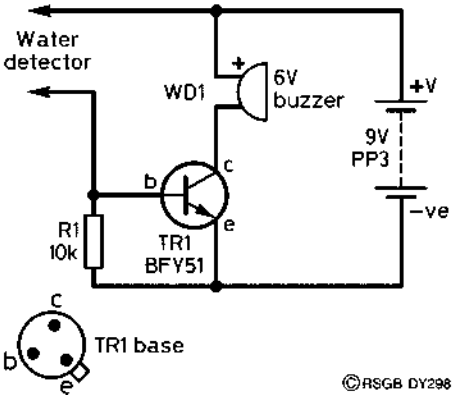
Water detector
WD1
+
6V buzzer
+V
9V PP3
−ve
b
c
e
R1 10k
TR1 BFY51
c
b
e
TR1 base
©RSGB DY298

Figure 1 Circuit diagram of the water level detector

Construction

The unit is made in two separate sections – the detector board (**Figure 2**) and the main board (**Figure 3**). The main board is plain perforated board (without copper strips). The components are pushed through holes in the main board and connected together on the underside. The battery is held on with sticky tape. No on/off switch is used, as the circuit consumes virtually no current when it is not sounding the buzzer.

Unlike some buzzers, this type is polarised, and the red lead must go to the supply as shown in Figs 1 and 3.

The detector board is a piece of standard matrix board – the type with copper strips along one side. It measures 9 strips by 15 holes, and is wired in such a way that alternate strips are connected together, as Figure 2 shows. This gives a greater surface area for the water to touch, thus increasing the sensitivity of detection.

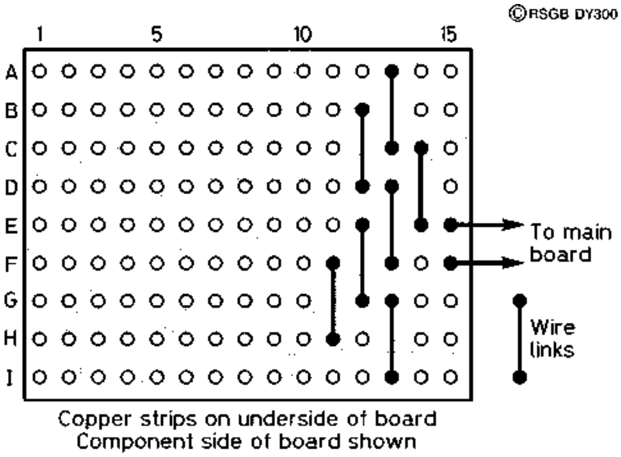
©RSGB DY300
1
5
10
15
A
B
C
D
E
F
G
H
I
To main board
Wire links
Copper strips on underside of board
Component side of board shown

Figure 2 Layout of the main board

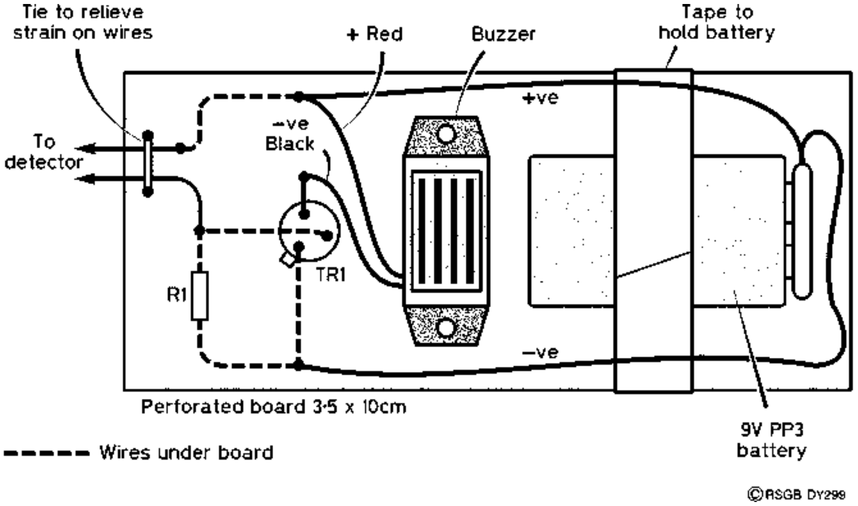
Tie to relieve strain on wires
+ Red
Buzzer
Tape to hold battery
+ve
To detector
−ve Black
TR1
R1
−ve
Perforated board 3·5 x 10cm
9V PP3 battery
Wires under board
©RSGB DY299

Figure 3 Water detector constructed from stripboard

The two boards are connected by ordinary twisted pair, enabling the detector to hang over the edge of the water container, with the buzzer circuit safely out of the way. When positioning the detector board, the left-hand edge (as shown in Figure 2) should face downwards, with the connecting-wire edge facing upwards. When the water level rises, the water causes current to flow between adjacent strips, and the higher the water, the lower the resistance between them. Eventually, the resistance will be low enough to turn on the transistor and sound the alarm.

Parts list

Resistor

R1	10 000 Ω (10 kΩ), 0.25 W

Semiconductor

TR1	BFY51

Additional items

WD1	6 V miniature buzzer (polarised)
	Battery PP3
	Battery connector
	Strip board 9 strips by 15 holes
	Perforated board 3.5 × 10 cm

78 A delta loop for 20 metres

Introduction

HF aerials take up a lot of room when they are straight. Space can be saved by bending them and, at the same time, giving them properties which are quite different from their original linear forms. The delta loops shown in the accompanying diagrams have the shape of the Greek upper-case letter delta (Δ) upside down. It can be upright, except that it is more practical to have the feed point at the bottom than at the top.

Putting the loop together

This is the design for a loop to be used on the 20 metre band. It is a very popular band and carries most of the amateur radio DX traffic. The aerial is light in weight and operates best if the top section is at least 30 ft above the ground, with the feeder point being about 6 ft above ground. All you need is some good wire, some polypropylene rope, some insulators and a method of matching the output of this 75 Ω balanced aerial to the 50 Ω unbalanced coaxial input of your transceiver or receiver.

Each of the three sides must be 7.2 m long. Using tent pegs or six-inch nails, mark out the three corners on the lawn, making sure that each pair of pegs or nails is exactly 7.2 m apart. Put the wire around the nails, together with two insulators on the wire for the two top corners. These insulators can be of the 'dog bone' variety, or of the home-made type, using a piece of flat plastic with holes in each end. The use of these is shown in **Figure 1**. Located at the positions of two of the tent pegs or nails, they must be secured at those points by any convenient means.

Now attach a support rope to the free end of each of the insulators, ready for suspension, as shown in the diagrams, from any convenient supports (house, tower, tree, etc.). The top section of the loop could be taped to a continuous piece of rope between the two supports. You might like to try this if you think it is easier. Then, when you find some dog-bone insulators, you can change the design and see if there is any noticeable change of performance.

Connecting to the radio

If you are the proud possessor of an aerial tuning unit (ATU) with a balanced output, all you will need is a length of twin cable soldered to the

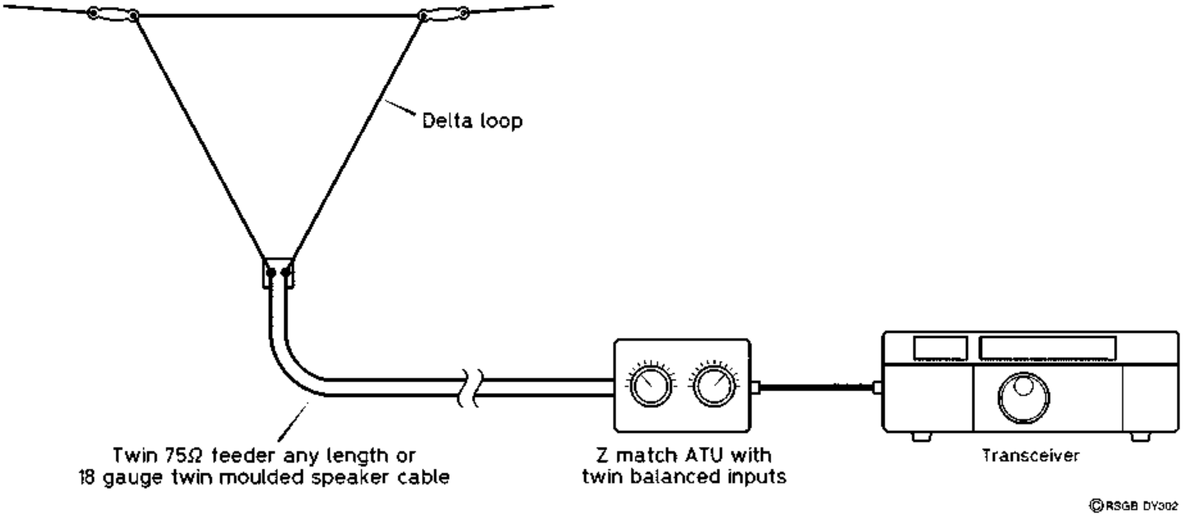
Delta loop
Twin 75Ω feeder any length or
18 gauge twin moulded speaker cable
Z match ATU with
twin balanced inputs
Transceiver
©RSGB DY302

Figure 1 Multiband version of the delta loop connected to the ASTU using 75 Ω twin feeder

ends of your delta loop and connected to the balanced output terminals on the ATU, as Figure 1 shows.

Another way is to buy a ferrite-cored 1:1 balun, and use it as indicated in **Figure 2**. This produces an aerial which will operate over the whole range of amateur and commercial short-wave frequencies, when used with an ATU.

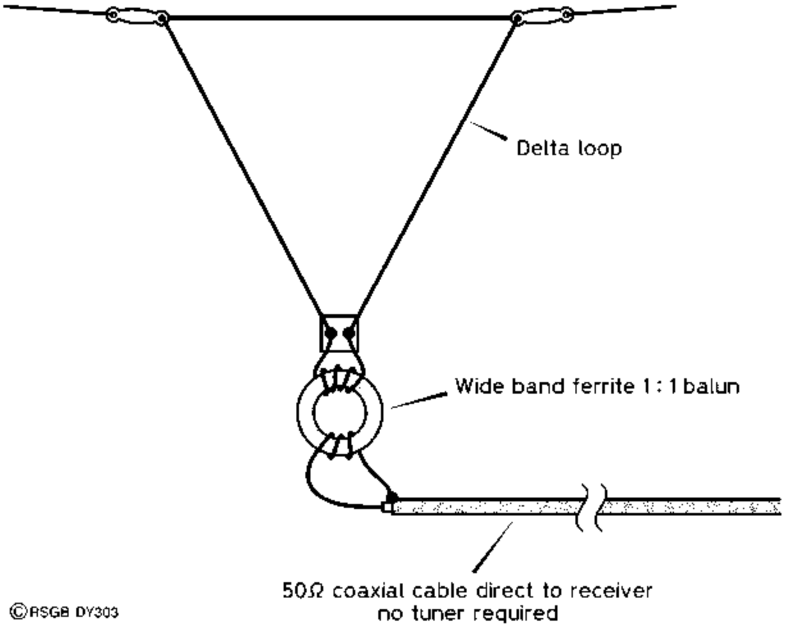
Delta loop
Wide band ferrite 1 : 1 balun
50Ω coaxial cable direct to receiver
no tuner required
©RSGB DY303

Figure 2 Multiband version of the delta loop using 75 Ω coaxial cable. This arrangement requires an ASTU if used with a multiband antenna

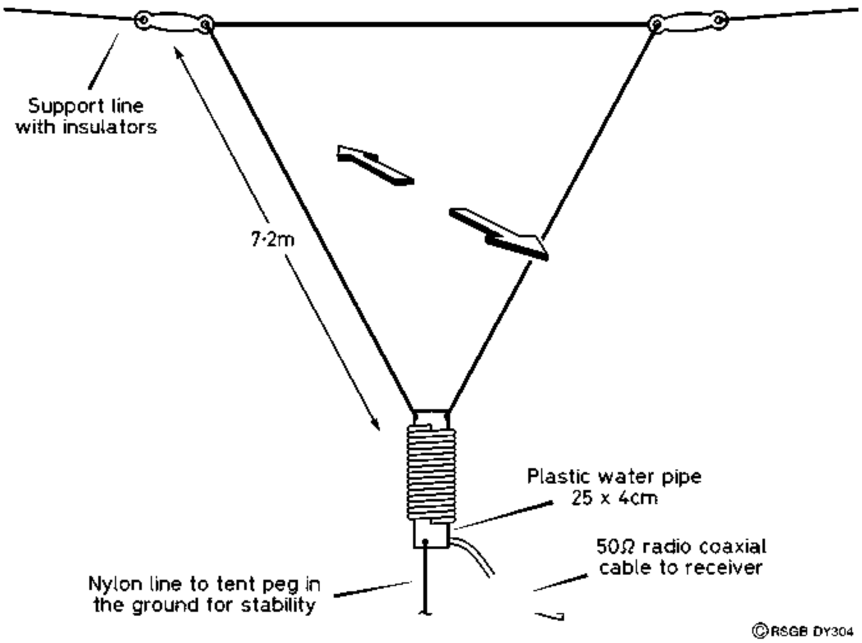

Figure 3 Single band version of the loop antenna using a coaxial balun. The direction of maximum signal strength is indicated by the arrows

However, since this aerial is designed to be a single-band type, there is a simpler answer to the problem of matching the aerial to your 50 Ω coaxial cable – **Figure 3** shows this. Make a tuned balun with a length of plastic water pipe, 25 cm long and 4 cm diameter and a piece of good-quality TV aerial cable. TV aerial cable has a characteristic impedance of 75 Ω, compared with the 50 Ω of the coaxial feeder that we usually use with amateur radio equipment. Use the type with a brown sheath and a closely knit earth braid, not the type having an earth foil inside the sheath.

The length we need for the balun is 3.8 m, but always allow between 3 and 4 cm extra for preparing the ends. Drill two small holes diametrically opposed in the top of the tube; these will be used to anchor the two ends of the delta loop, as shown in Figure 3. Drill another single hole in the bottom end of the tube, which will be used to anchor a nylon line going to the ground to add stability to the loop. Then, after drilling a pilot hole, drill a 5 or 6 mm hole near the top end of the tube, as shown in **Figure 4**. Prepare both ends of the coaxial cable, then feed one end into the tube, far enough for its ends to be soldered to the ends of the loop when the assembly reaches that stage.

Now, close-wrap the cable around the tube until only about 3 cm remain. Holding the cable tightly, drill another 6 mm hole beside the free end of the cable and feed it into this hole. At this point, feed the two ends of the loop into the top two holes and twist it back on itself. Figure 4 shows how this is done. Then, solder the ends of the coaxial balun to the ends of the loop.

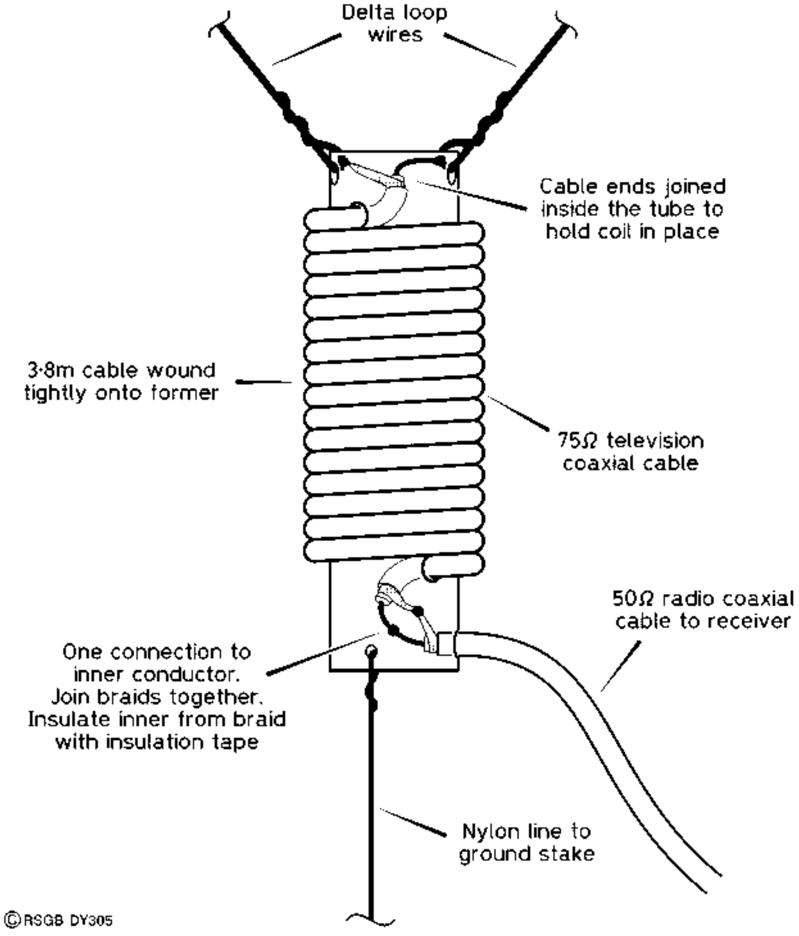
Delta loop wires
Cable ends joined inside the tube to hold coil in place
3·8m cable wound tightly onto former
75Ω television coaxial cable
50Ω radio coaxial cable to receiver
One connection to inner conductor. Join braids together. Insulate inner from braid with insulation tape
Nylon line to ground stake
©RSGB DY305

Figure 4 Construction of the balun

The bottom ends of the balun are then soldered to the 50 Ω coax which goes to your shack and to the transceiver.

Hoist the aerial into position carefully, being careful not to pull too hard on the support lines. Then, take the nylon line from the bottom of the balun to a peg in the ground. This adds stability to the aerial.

Using the delta loop

It is a directional aerial, as Figure 3 shows. It produces maximum power (and has maximum receive sensitivity) along a direction perpendicular to its own plane. Don't be too concerned with which direction it is pointing at first. Give it a try 'on the air' and see how it performs. Then you can contemplate how to point it in your favourite direction, to the USA, or Australia, for example.

Experimenting

You may want to enclose the balun in some sort of weatherproof container. Plastic ice cream containers are favourites for this sort of job. Seal all the holes where wires enter it with silicone sealant or self-amalgamating tape.

If you have used ordinary single-strand or multi-strand wire for your aerial, it will stretch over time under its own weight and that of the balun. Its operating frequency will fall slightly as a result. If you notice a significant difference, then dismantle it, remeasure and fix the sides and erect it again, perhaps facing a different direction. You can buy pre-stretched or hard-drawn wire for such purposes, if you feel that periodic tweaking of your aerial is a chore.

You can make a delta loop for different frequencies simply by scaling the lengths of wire for the loop and for the balun according to the design frequency. If your maths isn't quite up to this, enlist the help of someone well versed either in maths or aerial design!

79 A simple desk microphone

Introduction

A hand-held transceiver is usually the first type used by most novices. It is ideal when used out of doors, in the shack or mobile. However, using one in the shack with an outside aerial is not the best of solutions.

This article describes the design of a desk microphone which can be used either for transceivers using a common microphone and push-to-talk (PTT) line, or for those using a separate PTT line.

It is a simple design, using the barrels of two plastic pens, the lid of an old coffee jar, and the plastic sheath of a DIN plug.

Construction

The end caps and the ink tubes should be removed from two ball-point pens, leaving only the two plastic shells. One of them is used for the vertical microphone support, while the other is used to hold the microphone element.

The vertical support should have the pointed end shaped so that it holds the microphone support at a convenient angle, as **Figure 1** shows.

The base should be prepared next – see **Figure 2**. Four holes are required. Two, A and B, are small holes for the cable to be fed through. A third, C, is slightly larger and is used to take one of the plastic end caps. The fourth hole, D, is determined by the type of switch that you want to use for the PTT function.

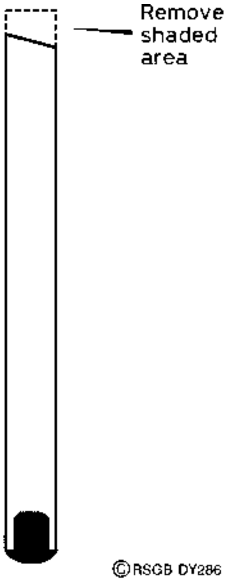
Remove
shaded
area
©RSGB DY286

Figure 1 Vertical support made from ball-point pen shell

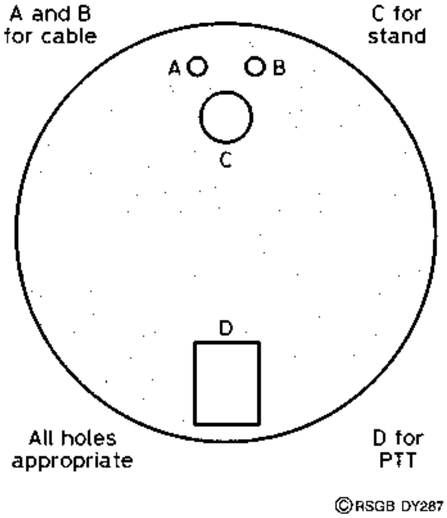
A and B
for cable
C for
stand
A
B
C
D
All holes
appropriate
D for
PTT
©RSGB DY287

Figure 2 The microphone stand base

Some epoxy resin such as Araldite Rapid is used to mount the separate parts. Mix only a small amount at a time, and glue each part of the base assembly and allow to harden before moving on. Be sure to read and follow the instructions for the use of the glue to prevent accidents.

Firstly, glue the end cap to the base, feeding it underneath and pushing it up through hole C. Then, the vertical pen body is placed over the protruding end cap and glued into place, making sure that the top (shaped) end is pointing in the right direction to hold the other plastic tube carrying the microphone. While the glue hardens, place the DIN plug sheath on the second pen body. Place and glue this pen body on the shaped top of the vertical support, making sure that the whole assembly is supported while the glue sets. What you should have now ought to resemble **Figure 3**, but without the microphone and wiring. Leave everything to set for 24 hours.

Any switch of the press-to-make variety can be used for your PTT switch, but choose one which does not apply too much force to operate! If you find the whole assembly is too light and slides over your table, some plasticine can be pushed into the base when all the wiring has been completed.

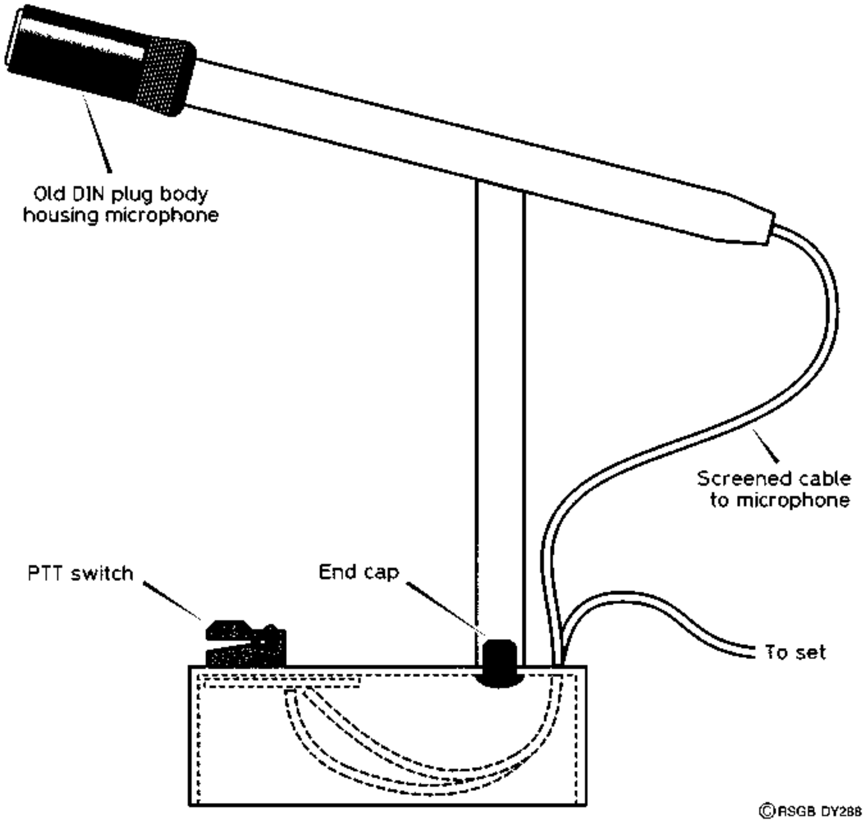
Old DIN plug body housing microphone
Screened cable to microphone
PTT switch
End cap
To set
©RSGB DY288

Figure 3 Complete microphone stand assembly

The circuit

A small electret condenser microphone is used, powered by a DC supply. Most hand-held radios use a single screened lead for both the PTT line and the microphone audio lead. In such cases, the circuit of **Figure 4** will operate the PTT function. Note that the PTT switch is in series with the audio lead from the microphone. If you find that the PTT switch does not switch your radio into transmit, reduce R1 from its normal value of 33 kΩ to 27 kΩ. You should find that the original value works with most hand-helds.

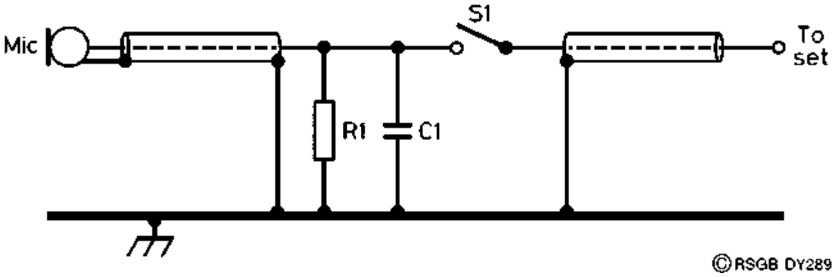

Figure 4 Wiring diagram for hand-held radios

For radios with a separate PTT lead, the circuit of **Figure 5** is used. However, a power supply is needed for the electret microphone. This can be a PP3 battery, or can (in most cases) be derived from the microphone socket on the transceiver. You will need to consult the makers' handbook for this information. Figure 5 shows the PTT switch wired in a 'ground-to-transmit' configuration, which is correct for most base-station transceivers. If you're in any doubt about what your radio needs, consult an experienced friend.

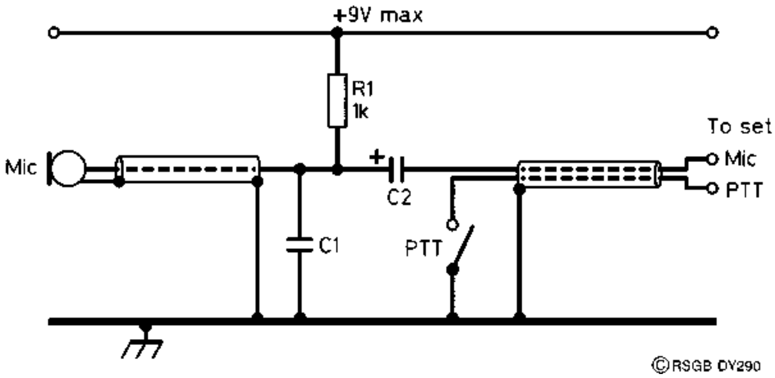

Figure 5 Wiring diagram for radios which use separate PTT lines

The last connections

The microphone element should be connected with screened cable. Tape (masking or insulating) should be wound around the electret insert until it is a snug fit inside the DIN plug sheath. Feed the screened cable through the barrel of the pen until it emerges from the far end. Poke it through hole B and make the connections under the coffee jar lid. The components can all be mounted on the PTT switch or on a small piece of Veroboard mounted under the top surface. The output lead emerges through hole A and is of sufficient length to reach your transceiver. A suitable plug needs to be fitted to it.

Parts list

Resistors: all 0.25 W carbon, 5% tolerance

R1 (Figure 4)	33 000 Ω (33 kΩ)
R1 (Figure 5)	1000 Ω (1 kΩ)

Capacitors

C1 (Figure 4 and 5)	0.1 microfarads (0.1 μF) disc
C2	10 microfarads (10 μF) electrolytic

Additional items

S1	Push-to-make
Mic	Electret microphone (Maplin type FS34W)
Screened cable	As required
Plug	As required by your radio
Veroboard	If required
PP3 battery and clip	If required

Two old plastic ball-point pens
Lid of coffee jar
Plastic sheath from DIN plug
Araldite or similar epoxy resin glue

80 Morse oscillator

Introduction

This is not the simplest Morse oscillator to build, but it differs from the simple circuits in that it produces a pure note, not a coarse, rasping sound. People who have practised Morse using a non-sinusoidal oscillator sometimes find that they have trouble copying Morse code with a pure tone. As the pure tone is the correct way to receive Morse code, it is important that you should learn to listen to the code from a pure oscillator – so here's one!

The twin-T

There are many oscillator circuits, and there are many variations of the twin-T oscillator that we are going to use. **Figure 1** shows one version of a very useful circuit. All oscillators must have positive feedback in order to work. The feedback determines the frequency of the note produced by the oscillator.

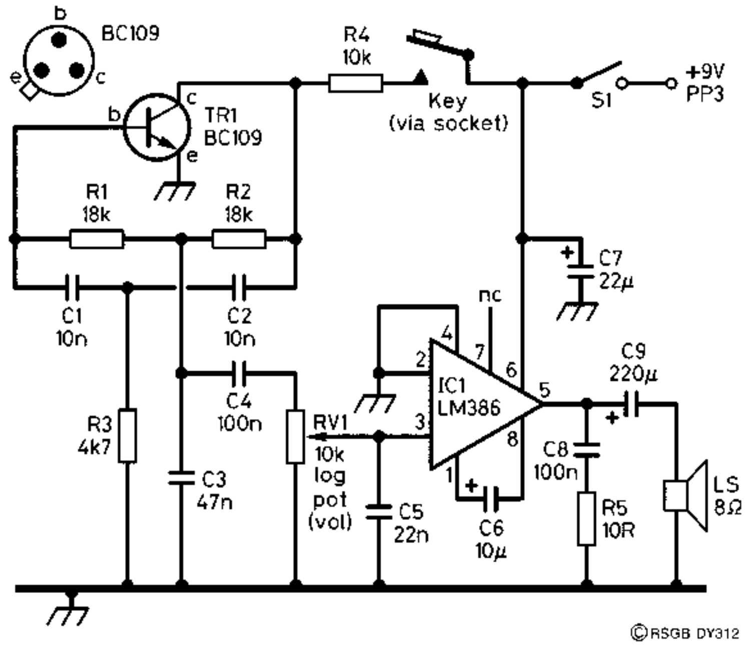
b
BC109
e
c
R4
10k
Key
(via socket)
S1
+9V
PP3
c
b
TR1
BC109
e
R1
18k
R2
18k
+
C7
22µ
nc
C1
10n
C2
10n
C9
220µ
4
7
2
6
IC1
LM386
5
C4
100n
3
RV1
10k
log
pot
(vol)
+
8
C8
100n
R3
4k7
C3
47n
1
+
C5
22n
C6
10µ
R5
10R
LS
8Ω
©RSGB DY312

Figure 1 Circuit diagram of the sinewave oscillator and amplifier

Here, the feedback circuit looks like two letter Ts. If you look at Figure 1, one T is formed by R1, R2 and C3, the other by C1, C2 and R3 – hence the name 'Twin-T'. Notice that the two Ts are connected in parallel between the collector and base of TR1, so any signal appearing at the collector is fed into both Ts. What emerges is then fed back into the base, producing in turn a signal at the collector. And so it goes on, producing a sine wave output.

The oscillator output is fed into an integrated circuit amplifier for output via a small loudspeaker.

Putting it together

The prototype was constructed on plain matrix board (the type *without* the copper strips), as shown in **Figure 2**. The components have their leads pushed through the holes in the board, and connections are made underneath.

Build the amplifier circuit first, using a socket for IC1. Connect the 9 V supply and touch pin 3 with an ordinary piece of wire. If a buzz is heard from the speaker, all should be well. If not, check your circuit and make changes until it does.

Build the oscillator circuit, and connect its output to the volume control VR1 via C4. Set VR1 half-way along its travel and switch on. A note should be heard from the speaker when the Morse key is depressed. The component values making up the twin-T determine the frequency of the note. Try varying them if you think your note is too high or too low. Whatever changes you make, either to the resistors or the capacitors, always ensure that R1 = R2 and that C1 = C2.

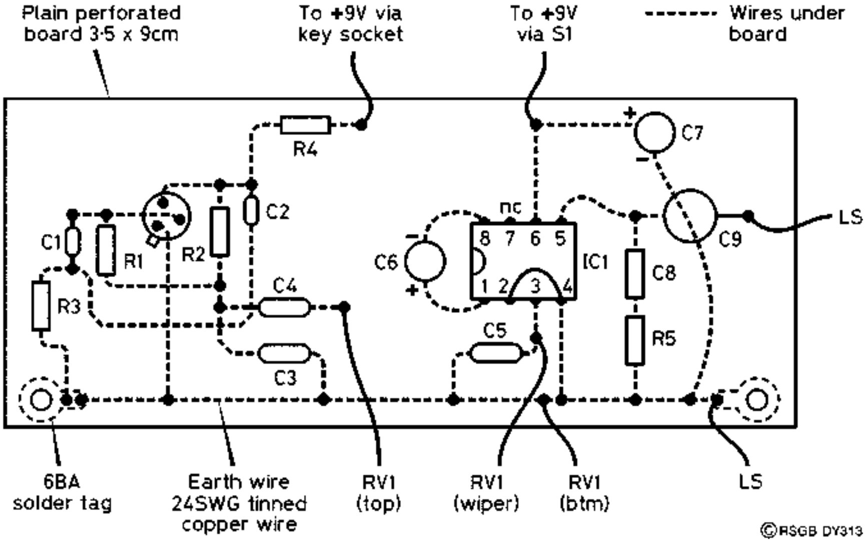
Plain perforated board 3·5 x 9cm
To +9V via key socket
To +9V via S1
------- Wires under board
R4
C7
+
-
C2
nc
LS
C1
R1
R2
8 7 6 5
C6
IC1
C8
C9
R3
C4
1 2 3 4
+
C5
R5
C3
6BA solder tag
Earth wire 24SWG tinned copper wire
RV1 (top)
RV1 (wiper)
RV1 (btm)
LS

©RSGB DY313

Figure 2 Component layout and interconnection diagram

Parts list

Resistors: all 0.25 W carbon film

R1, R2	18 000 ohms (18 kΩ)
R3	4700 ohms (4.7 kΩ)
R4	10 000 ohms (10 kΩ)
R5	10 ohms (10 Ω)

Capacitors

C1, C2	10 nF (0.01 μF)
C3, C8	47 nF (0.047 μF)
C4	100 nF (0.1 μF)
C5	22 nF (0.022 μF)
C6	10 μF electrolytic, 16 V WKG
C7	22 μF electrolytic, 16 V WKG
C9	220 μF electrolytic, 16 V WKG

Semiconductor

TR1	BC109

Integrated circuit

IC1	LM386

Additional items

VR1	10 kΩ log potentiometer with switch
LS	8 Ω loudspeaker

Matrix board 3.5 × 9 cm
Small jack socket for key input
Box
PP3 battery and clip

81 A simple 6 m beam

Introduction

The attraction of building your own aerials is an abiding feature of our hobby. You can *buy* almost any shape or size of aerial, but one you have made yourself can often work every bit as well as a commercial device costing ten times as much.

The design

This aerial, designed for use on the 6 m band, is essentially a two-element Yagi, with the elements bent in order to reduce the physical size. It is known

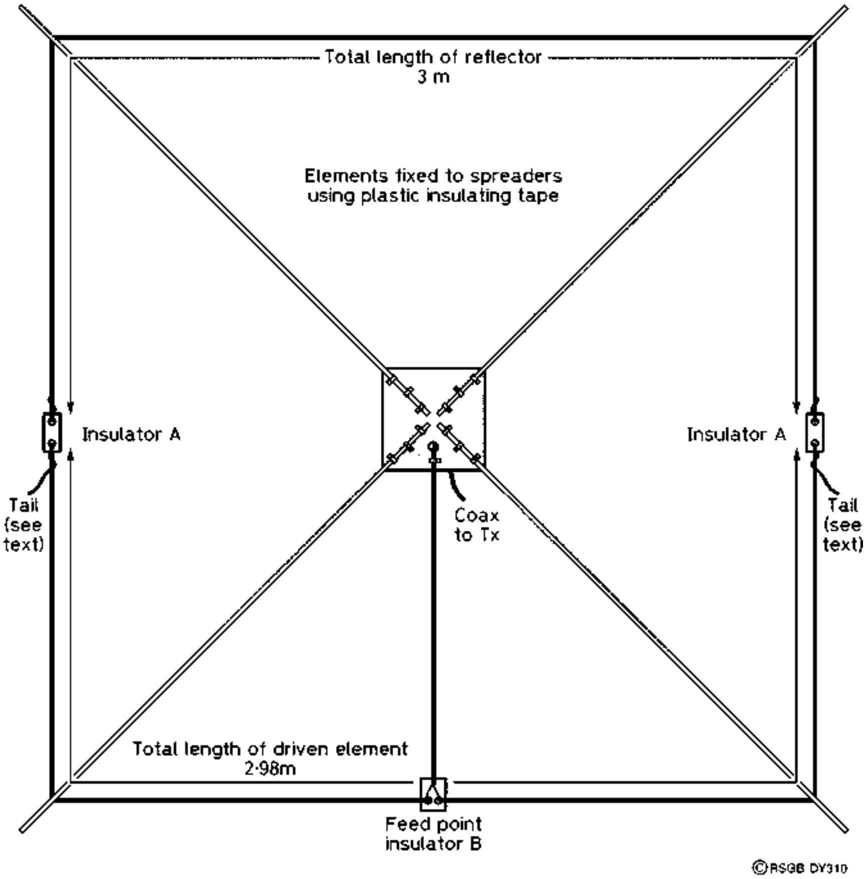
Total length of reflector
3 m
Elements fixed to spreaders
using plastic insulating tape
Insulator A
Insulator A
Tail
(see
text)
Tail
(see
text)
Coax
to Tx
Total length of driven element
2·98m
Feed point
insulator B
©RSGB DY310

Figure 1 Plan view of the complete 50 MHz VK2ABQ antenna

as the VK2ABQ beam, and was designed originally for the 20, 15 and 10 m bands, principally because of its space-saving qualities. It is made using a wooden frame and wire elements, and is ideal for portable operation.

Tools ready?

The beam is shown in **Figure 1**. The driven element is the one whose centre is fed by the coaxial cable, and lies between the two insulators marked A and the feedpoint at B. The reflector is also anchored at the points A, and lies over the upper half of the frame.

The wooden centrepiece is used to support the cross-pieces and to mount the aerial on the mast, using a common shelf bracket. The cross-pieces, known as *spreaders*, can be wooden canes or dowelling, and are mounted to the centrepiece using cable clips and adhesive. **Figure 2** shows how this is done.

If the aerial is to be a permanent installation, the spreaders should be weatherproofed using a good-quality exterior varnish. The wire elements are PVC covered and fed through holes in the spreaders.

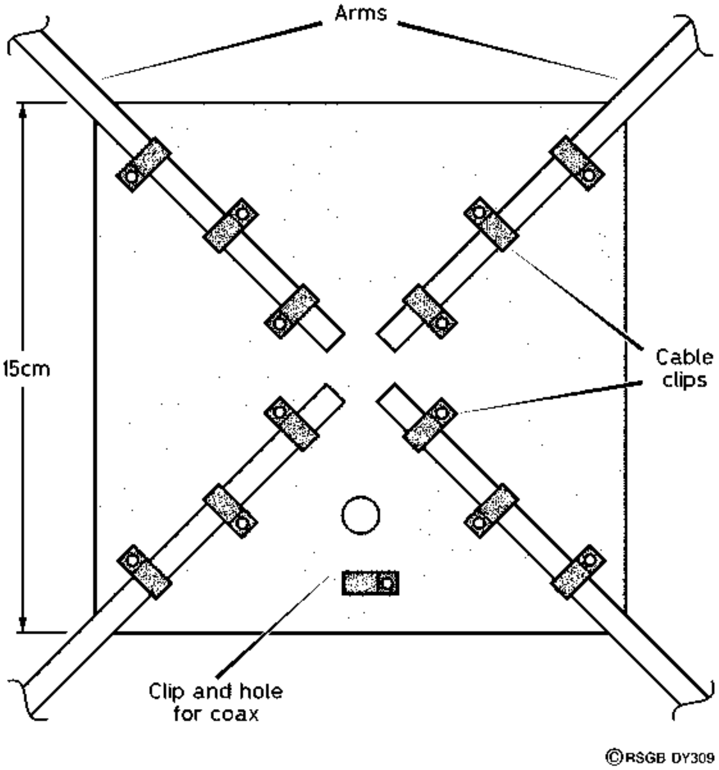

Figure 2 Centre support piece

The end insulators are made of drilled perspex, and the wire passed through the two holes and twisted, as shown in **Figure 3**.

Note: if you have not drilled holes in perspex before, take care! The drill bit must be well-lubricated because it generates a lot of heat (enough to melt the perspex and jam the drill). Never turn the drill for more than a few seconds at a time, and moisten the bit in between. Start with a pilot hole and use bits of increasing size until the hole is the size you want.

Another perspex insulator is used to secure the feeder to the aerial, as shown in Figure 3. The feeder then passes directly to the centrepiece, where it is fastened with a cable clip and then passes down the mast.

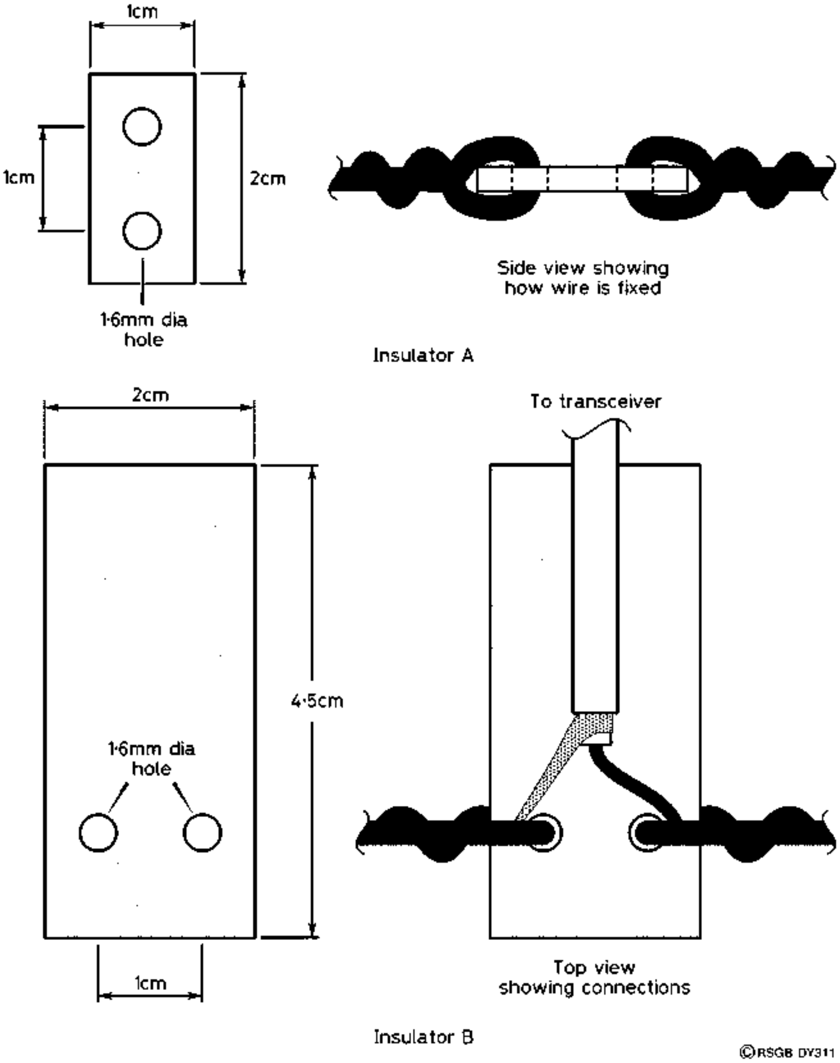
1cm
1cm
2cm
1·6mm dia hole
Side view showing how wire is fixed
Insulator A
2cm
To transceiver
4·5cm
1·6mm dia hole
1cm
Top view showing connections
Insulator B
©RSGB DY311

Figure 3 Details of insulators

Adjustment

The driven element (A–B–A) in Figure 1 is fixed to the end insulators in such a way as to have 'pigtails' which are about 10 cm long. Using an SWR meter between the aerial and the transmitter, trim the pigtails equally at each end for minimum SWR in that portion of the 6 m band in which you plan to work. Make sure that the transmitter is off when you trim the ends, as high voltages can be present there. Always listen on the frequency before you transmit and, when you do, ask if the frequency is in use and identify yourself. Use as little power as possible.

The prototype had an SWR of 1.2 at 50.2 MHz and performed well. If you look at Figure 1, which is a view of the aerial from above, you will immediately see that when the aerial is horizontal, it radiates with horizontal polarisation in a direction from the top of the page to the bottom. A metal pole or mast can be used for horizontal polarisation, but if you intend to use the aerial for vertical polarisation, it is much better to use a wooden or fibreglass pole.

Portable use

If you plan to operate portable with this aerial, the only real modifications you need are to the centrepiece and how it supports the spreaders. Instead of using glue and cable clips, nuts and bolts through the spreaders and centrepiece would allow the spreaders to be 'hinged' closed for transport.

Materials

Centrepiece	Hardwood, 15 × 15 × 25 mm
4 spreaders	110 cm long (cane or 6 mm dowel)
Wire	1.5 mm PVC-covered copper
50 Ω coaxial cable	RG58 or similar
Cable clips	6 mm (12 off)
Varnish	Polyurethane for waterproofing
Tape	Self-amalgamating, to waterproof all soldered joints
Insulators	See text

82 An integrated-circuit amplifier

Introduction

A simple audio-frequency amplifier is a very useful building-block in many more advanced electronic circuits. It is also a very useful piece of test equipment. To have one spare in the shack can be a life-saver at times.

Planning

Get into the habit of planning your project. How do you want it to look when it is finished. Do you want it in a box? Do you want it 'open-plan', with all the components on view?

The 'minimalist' approach to any project is simply a front panel and a baseboard, on to which all the components are fitted. In this case, the front panel would accommodate the loudspeaker, the volume control and the input and output jack sockets, and the baseboard would support the circuit board. Once you have decided these things, and know the size of the board, speaker and volume control, you can decided how big the panel and baseboard should be.

The amplifier

Instead of building the amplifier from discrete components (i.e. transistors), as was done in the project *An audio-frequency amplifier* (which you will find elsewhere in this book), we are going to use an integrated-circuit (IC) amplifier. External resistors and capacitors are still needed but, compared with the number of components *inside* the chip, these are very few indeed!

The IC we are going to use is the TBA820M. The circuit may look complicated, but with the use of a matrix circuit board it becomes quite simple. **Figure 1** shows the circuit diagram and **Figure 2** the layout on the board. The external connections to the PCB are shown in **Figure 3**.

To avoid having a separate switch to switch off the power supply when the amplifier is not being used, a volume control which incorporates a switch is used.

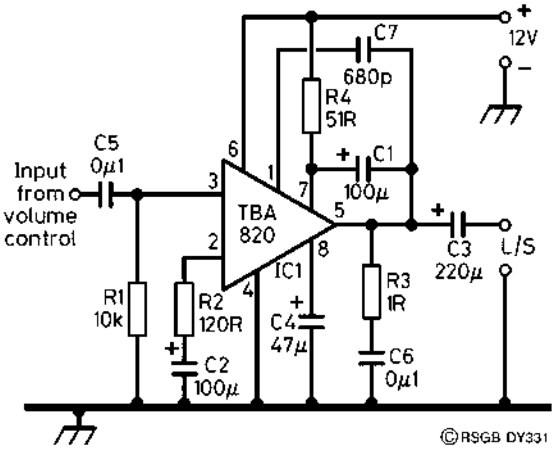
+
12V
−
C7
680p
R4
51R
C5
0µ1
Input
from
volume
control
6
3
1
7
+
C1
100µ
TBA
820
5
2
8
4
IC1
+
C3
220µ
L/S
R1
10k
R2
120R
+
C2
100µ
+
C4
47µ
R3
1R
C6
0µ1
©RSGB DY331

Figure 1 Audio amplifier, circuit diagram

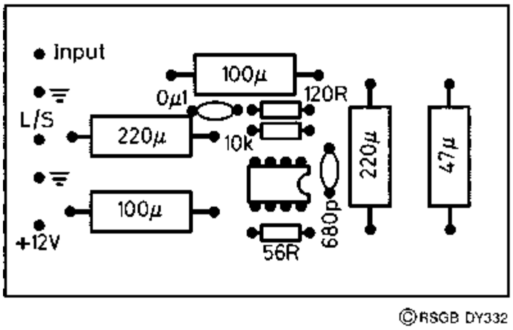
Input
L/S
+12V
100µ
0µ1
120R
220µ
10k
100µ
56R
680p
220µ
47µ
©RSGB DY332

Figure 2 Audio amplifier, component layout

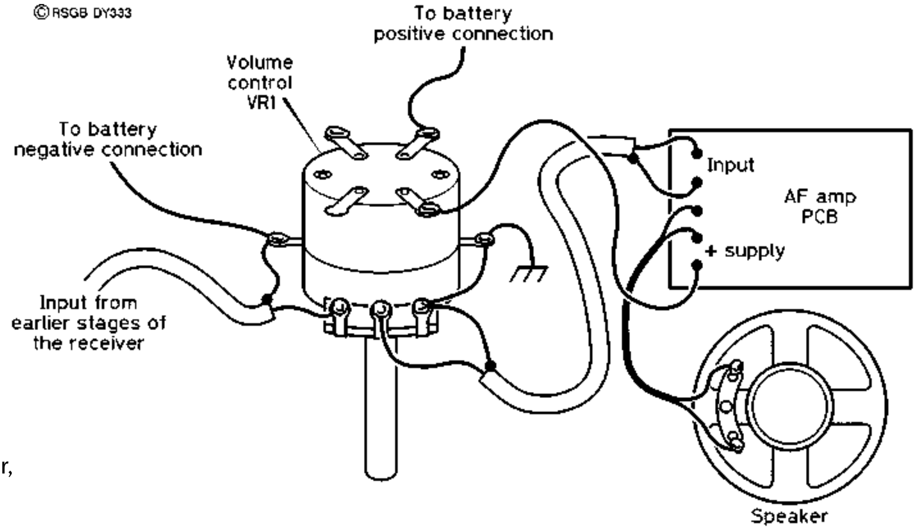
©RSGB DY333
To battery
positive connection
Volume
control
VR1
To battery
negative connection
Input
AF amp
PCB
+ supply
Input from
earlier stages of
the receiver
Speaker

Figure 3 Connecting the circuit board to the speaker, volume control/switch and battery

What power supply?

If the amplifier is to be used only for short periods, as a test instrument, for example, then it can be run from a PP3 battery, which can be mounted on the baseboard. If you intend to use it often, then a connection to an external power supply is preferable. For this option, you may want to consider fixing two terminals to the baseboard for this connection.

83 A novice ATU

Introduction

Having an aerial tuning unit (ATU) is always useful. It is used for adjusting the aerial impedance, as ‘seen’ by the transceiver, to be the same as that of the transceiver itself, usually 50 Ω. This process is called *matching*, and ensures that the receiver and the power amplifier (PA) stage of the transmitter work efficiently.

This design is due to the late Doug DeMaw, W1FB, and uses readily available components. It will handle up to about 5 W, and operates over the frequency range 1.8 to 30 MHz.

Circuit evolution

The basic circuit of one type of ATU is shown in **Figure 1**. On transmit, the input signal at L1 is coupled into L2, which forms a resonant circuit with C1. The signal across the resonant circuit is fed to the aerial by C2. The combination of L2 and C1 helps to remove signal harmonics, because they are not at the resonant frequency and are shunted to earth. On receive, only those signals which are within the pass-band of L2 and C1 will pass into the receiver. This improves receiver performance by rejecting many out-of-band signals which the simpler receiver doesn’t like.

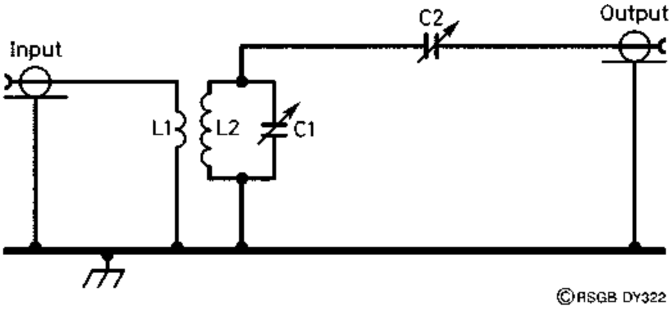

Figure 1 Basic circuit of the ATU

For this design, the all-important resonant frequency is chosen to be in the 80 metre band. If we want to make the circuit operate on other bands, we need to change either L2 or C1. It is easy to change L2 by using a set of switched inductors (L3–L5) as shown attached to S1 in **Figure 2**. Forget about S2 and L6 for the moment.

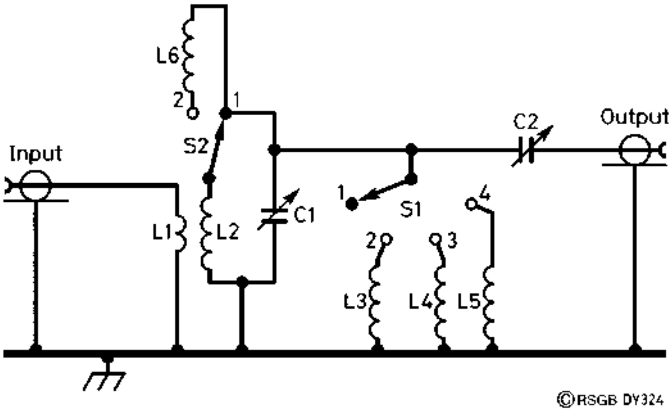

Figure 2 L6 added for 160 m

When S1 is in position 1, the circuit reverts to that of Figure 1. S2 should be brought into circuit only when S1 is in position 1. Looking at the way the circuit is drawn, you can see that when S2 is in position 2, we have two coils in series, which is equivalent to adding more turns to L2. More turns means more inductance, which lowers the resonant frequency still more, and gives coverage of the 160 m band.

Rather than having to remember to flick S2 to the correct position *and* move S1 to position 1 when we want to operate on 160 m, both functions can be combined into one rotary switch with two wafers. This circuit is shown in **Figure 3**.

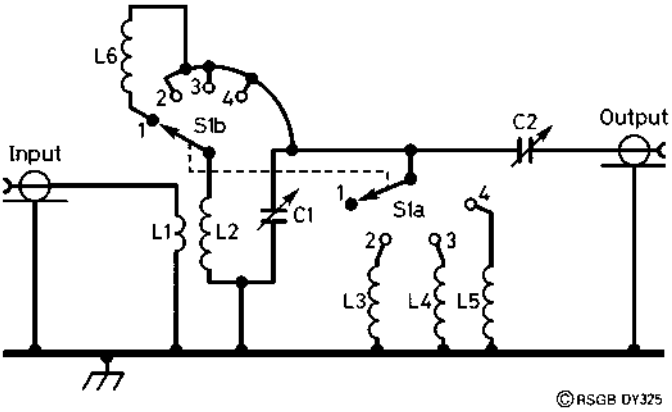
L6
2
3
4
1
S1b
Input
L1
L2
C1
C2
Output
1
S1a
2
3
4
L3
L4
L5
©RSGB DY325

Figure 3 Two-pole switch for all bands

In all positions of S1a *except* position 1, the extra coil, L6, is shorted out. The two halves of the switch, S1a and S1b, move together, as the shaft is turned, so both halves are in position 1 at the same time, position 2 at the same time, and so on. The two switches, S1 and S2, are said to be *ganged*.

Construction

L1 is formed by winding four turns of 22 SWG enamelled copper wire over L2, as in **Figure 4**. L2 already exists on the purchased former. After scraping the enamel off the ends of the wire (with a sharp knife or sandpaper), one end of L1 must be soldered to one end of L2, as shown in Figure 4. The free end of L1 goes to the transceiver aerial socket.

All components except the capacitors are assembled on the switch. Note that the rotor of C1 is earthed, but neither side of C2 is earthed. This means that the metal shaft of C2 is not earthed, and touching it will detune the ATU. Using a plastic knob for C2 will minimise this effect. If you decide to use a metal box, take precautions to ensure that no part of C2 is in electrical contact with the box.

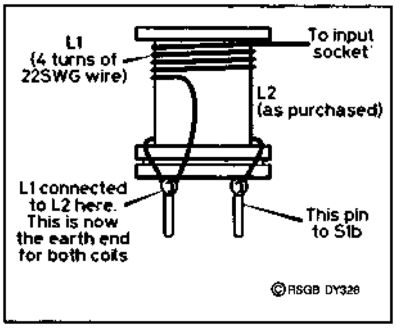
L1
(4 turns of
22SWG wire)
To input
socket
L2
(as purchased)
L1 connected
to L2 here.
This is now
the earth end
for both coils
This pin
to S1b
©RSGB DY328

Figure 4 L1 is wound over L2 as shown

The capacitors are mounted using M2.5 screws. Make sure that the screws do not foul the vanes of the capacitor. If your screws are too long, a few washers between the box and the capacitor will solve the problem!

In use

The best indication of a good match is obtained with a standing-wave meter between the ATU and the transceiver. The controls are adjusted alternately to ‘feel’ your way to a better and better match.

For receive-only use, the same alternate adjustments are used, watching for the maximum signal strength on the S-meter or, for a very weak signal, making the signal from the loudspeaker as large as possible.

Parts list

Capacitors	
C1, C2	350 pF variable
Inductors	
L1	See text
L2	27 μH
L3	10 μH
L4	2.2 μH
L5	1 μH
L6	65 μH
Additional items	
SO239 sockets (2 off)	
S1	2-pole 6-way rotary
Box	as required
Plastic knobs (2 off)	
Stick-on feet (4 off)	
M2.5 screws for capacitors	
Screws, nuts and washers for mounting the input and output sockets.	

84 CW QRP transmitter for 80 metres

Introduction

This is a relatively simple transmitter design having an output of 1 W. The design is not new, having been described before in other amateur radio publications. The components are all available new and the total cost should not exceed £15.

The circuit

Like other simple transmitters (see *An 80 Metre Crystal-Controlled CW Transmitter* and *A Breadboard 80 m CW Transmitter* elsewhere in this book) this one is crystal controlled. This assures frequency stability, but limits the usefulness of the transmitter. The key to increased frequency coverage without a conventional Variable Frequency Oscillator (VFO) is the use of a low-cost 3.58 MHz *ceramic resonator*. The 'pulling' range of a 3.58 MHz ceramic resonator covers the UK novice 80 m sub-band and some of the CW segment below 3.525 MHz.

A ceramic resonator is like a crystal, but not quite as stable in frequency. Its main advantage is its large pulling range.

The block diagram is shown in **Figure 1**. It is very similar to a crystal-controlled transmitter, and includes an oscillator, buffer and final amplifier.

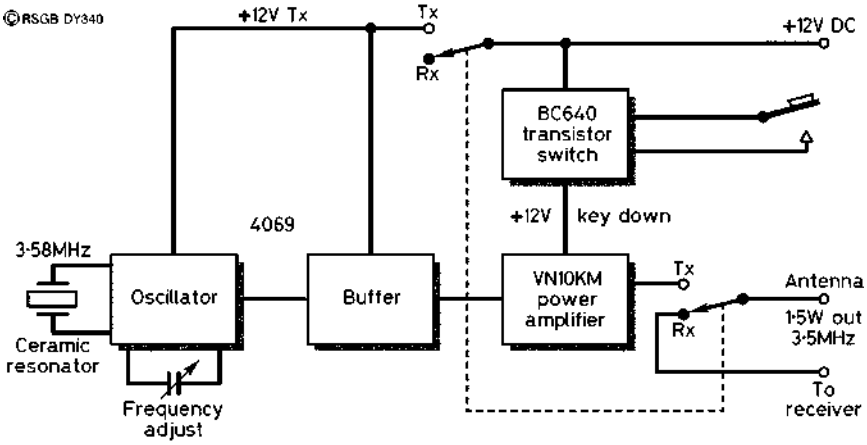
©RSGB DY340
+12V Tx
Tx
Rx
+12V DC
BC640 transistor switch
+12V
key down
4069
3·58MHz
Oscillator
Buffer
VN10KM power amplifier
Tx
Antenna
Rx
1·5W out 3·5MHz
Ceramic resonator
Frequency adjust
To receiver

Figure 1 Block diagram of the ceramic resonator controlled 80 m CW transmitter

This amplifier is keyed, the oscillator remaining running all the time. This improves frequency stability because the oscillator is not being continuously stopped and started by the key. It is switched off while receiving, though, to avoid interference with the received signal. Transmit/receive switching is accomplished by a panel-mounted switch controlling both the aerial, oscillator and buffer switching.

Figure 2 shows the transmitter circuit diagram. An unusual aspect of this transmitter is the use of a digital CMOS integrated circuit (IC) type 4069 for the buffer and oscillator stages. The IC houses six inverters, four of which are used in the circuit. One is used as the oscillator, two are used for the buffer stage, and the fourth provides an output for a direct-conversion receiver, should one be added at a later date.

The frequency of the oscillator is changed by varying the capacitance in the ceramic resonator circuit. This is provided by VC1.

The power amplifier (PA) is a small MOSFET (metal oxide semiconductor field-effect transistor), TR1. This is capable of providing an output power of 2 W but, in this circuit, it is run conservatively to give 1.5 W. The output can be varied by changing the resistance (R5 + R6) in the gate circuit. Attempts

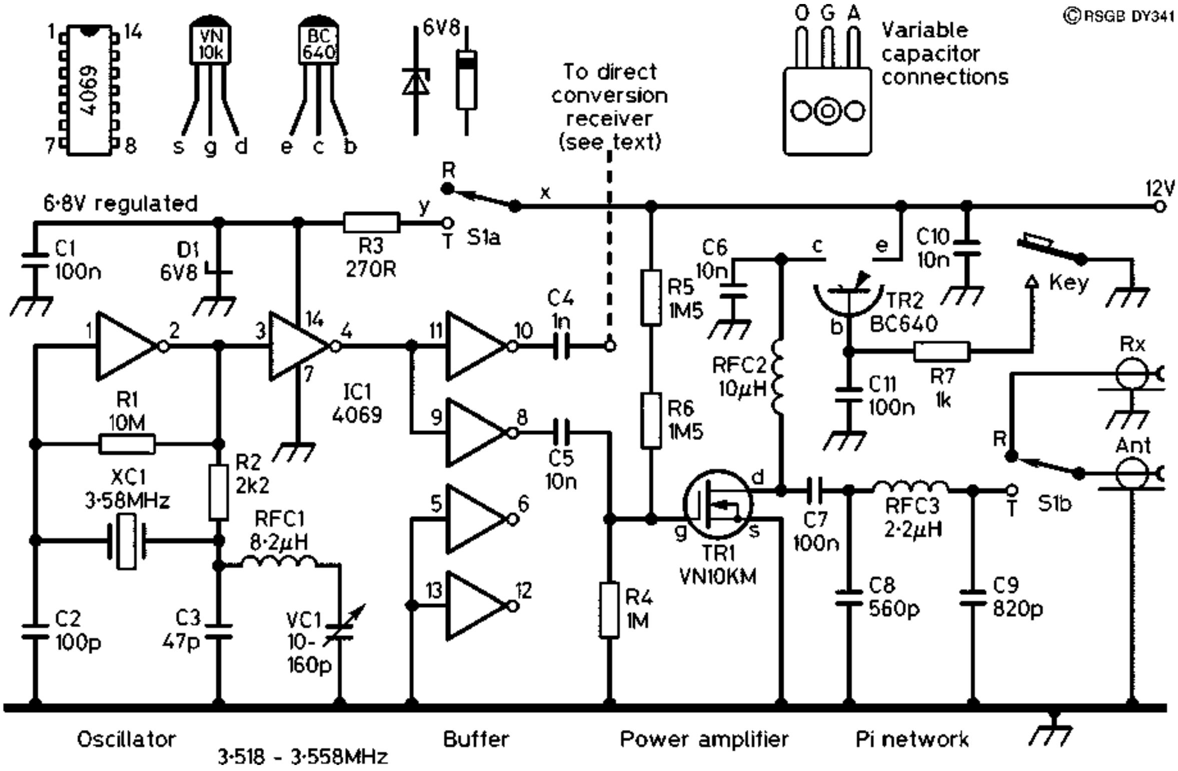
©RSGB DY341
1
14
4069
7
8
VN 10k
s g d
BC 640
e c b
6V8
To direct conversion receiver (see text)
O G A
Variable capacitor connections
6·8V regulated
C1 100n
D1 6V8
R3 270R
R
y
T
x
S1a
12V
C6 10n
c
e
C10 10n
Key
R5 1M5
TR2 BC640
b
C4 1n
1
2
3
14
4
11
10
7
IC1 4069
R1 10M
R6 1M5
RFC2 10µH
C11 100n
R7 1k
Rx
9
8
C5 10n
R
Ant
R2 2k2
XC1 3·58MHz
d
g
s
TR1 VN10KM
C7 100n
RFC3 2·2µH
T
S1b
RFC1 8·2µH
5
6
13
12
R4 1M
C8 560p
C9 820p
C2 100p
C3 47p
VC1 10-160p
Oscillator
3·518 - 3·558MHz
Buffer
Power amplifier
Pi network

Figure 2 Circuit diagram of 80 metre transmitter

to raise the output power by decreasing the values of these resistors may result in immediate MOSFET failure.

A pi-network (C8, RFC, C9) provides impedance matching to 50 Ω, together with harmonic suppression. Like all inductors in this transmitter, the pi-network inductor is a pre-wound RF choke. A pi-network is so called because the components are arranged in the shape of the Greek letter pi (π).

Keying is carried out by a pnp transistor switch, TR2. Closing the key earths the base and supplies 12 V to the collector of TR2 and to the drain of the MOSFET, TR1, allowing the PA to operate.

Construction

You must house your transmitter in a metal box, to avoid hand-capacity effects and the radiation of spurious frequencies. Size is not important, provided it is large enough to accommodate the transmitter without cramping the components. You may want to allow space for future additions such as a direct-conversion receiver, break-in keying, sidetone or a small power amplifier. A good size is 5 × 15 × 15 cm. You can make your own box, buy it, or even use a biscuit tin!

Front and rear panel connectors can be fitted first. The choice of these is a personal matter, but a good working choice would be:

(a) Power socket – 2.1 mm panel socket – centre pin positive.
(b) Key socket – ¼ inch jack socket.
(c) Aerial and receiver connectors – panel-mounting SO239 type.

Particular attention must be paid to the mounting of the variable capacitor, VC1. Make sure the hole for the shaft is amply big enough, and if you use screws to mount the capacitor on the front panel, then make sure they are not too long, otherwise they will touch the vanes of the capacitor! Mounting can be by means of glue, sparingly applied and kept well away from the shaft.

A board size of about 6 × 10 cm is adequate. Component layout on the board is suggested in **Figure 3**. The prototype used ordinary matrix board, which is preferable to stripboard for a design like this; stripboard has undue capacitance between adjacent strips. Component leads are fed through holes in the board and are connected underneath. Make sure that leads and connections are rigid because, if they can move, there is always the danger of short-circuiting, and capacitance changes.

To facilitate construction, servicing and testing, it is advisable to use Veropins for connections to the variable capacitor, transmit/receive switch,

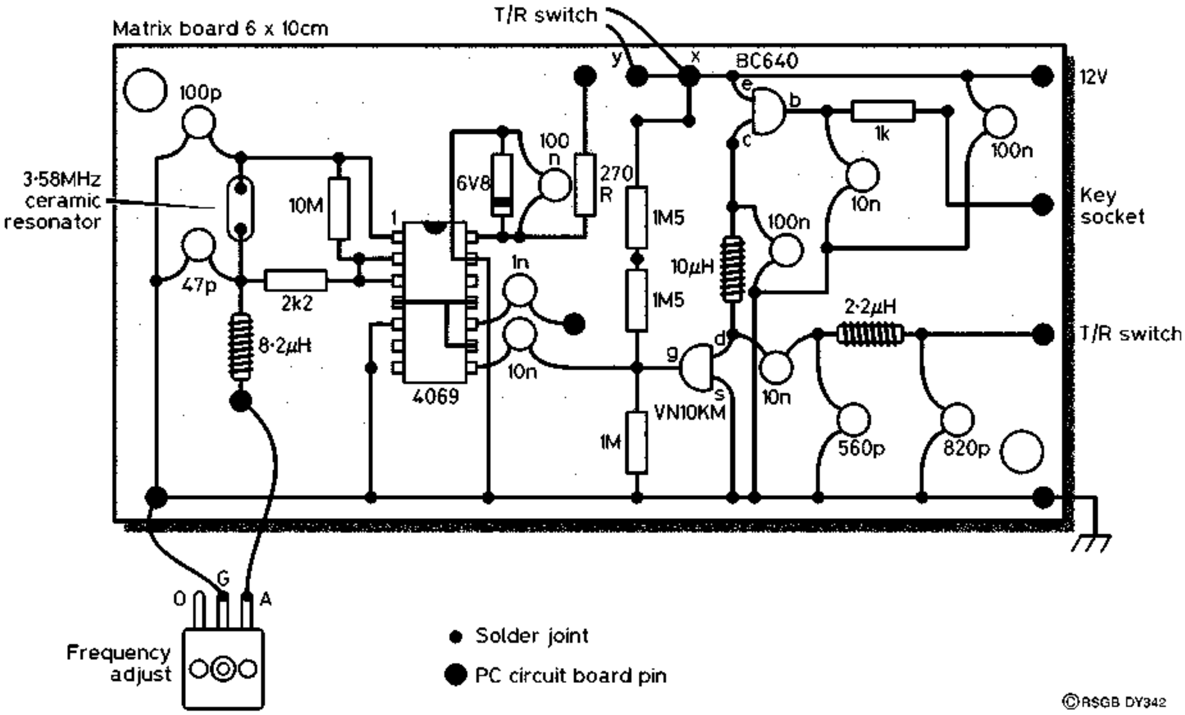
Matrix board 6 x 10cm
T/R switch
y
x
BC640
e
b
c
12V
100p
3·58MHz
ceramic
resonator
10M
1
6V8
100
n
270
R
1M5
1k
100n
10n
Key
socket
100n
10µH
47p
2k2
1n
1M5
2·2µH
T/R switch
8·2µH
4069
10n
g
d
s
VN10KM
10n
1M
560p
820p
O
G
A
Frequency
adjust
● Solder joint
● PC circuit board pin
©RSGB DY342

Figure 3 Component layout of the 80 metre transmitter. The transmit/receive switch is not mounted on the board and is not shown in this diagram

aerial and power sockets. Use screws and spacers to mount the circuit board to the box. Mounting the board horizontally assists troubleshooting.

Use a socket for the IC, and observe the CMOS handling precautions given in *An Electronic Die*, elsewhere in this book. When soldering the leads to the ceramic resonator, do it quickly – excessive heat damages the device. The earth lead running acoss the bottom of the board must be connected to the metal case by a short length of stout wire.

Testing

After carefully checking your wiring, both against the circuit diagram and the layout diagram, it is time to test your circuit. You will need a multimeter, an 80 m SSB receiver and a 50 Ω dummy load. A good design of dummy load can be found in the project *A Switched Dummy Load*, also in this book. An RF power meter and frequency counter will also be useful, although if your receiver has a digital frequency readout and S-meter, the latter two items are not really necessary. You will also need a 12 V 1 A power supply unit (PSU) to power the transmitter.

Switch the transmitter to receive and switch on the transmitter. No current should be consumed. Switch to transmit and check that pin 14 of IC1 is 6.8 V positive. With the dummy load connected to the aerial socket, press the key. The voltage on TR2 collector should now be 12 V, dropping to zero when the key is released.

Now check the operation of the oscillator. In transmit mode, you ought to be able to find a strong carrier signal with the receiver, even though the dummy load is connected. Adjusting the variable capacitor should change the frequency. At the lower end of the frequency range, you may find that the oscillator is unreliable in starting, because the circuit is attempting to pull the resonator too low in frequency. If this is the case, set the trimmer at the back of VC1 to minimum capacitance. If your version of VC1 has two trimmers, and you don't know which one to set, set them *both* to minimum capacitance. If there is still a problem, reduce the value of RFC1 to 6.8 or 4.7 μH.

In all probability, the unmodified circuit of Figure 2 will not require any of the changes outlined here. A coverage of 3.518 to 3.558 MHz should be possible, while preserving good frequency stability and reliable oscillation.

A signal probe (see *An RF Signal Probe*, elsewhere in this book) is useful for checking the operation of the oscillator and PA. Alternatively, an RF power meter or the receiver's S-meter can be used. With the PA running, the unit should draw between 200 and 300 mA. If TR1 becomes too hot to touch after a few seconds of transmitting, increase R5 or R6 to limit the transistor's heat dissipation. A small 6.3 V bulb connected across the aerial output is a simple way to check that the PA is working. An orange/white glow when the key is pressed is indicative of correct operation.

The final test is to monitor keying 'quality'. With your dummy load connected, press the key and listen to the note on the receiver's loudspeaker. Then operate the key, sending a string of dits, for example. What you hear should be free of chirps and clicks, as well as being stable in frequency. This test is sometimes better performed with no aerial connected to the receiver, thus preventing receiver overload and its associated plops. No problems should be encountered here.

Frequency tuning

A peculiarity of ceramic resonators is that, every now and again, their frequency changes abruptly by 100 Hz or so, then remains stable for some time. This is certainly noticeable in the received signal, but does not detract from the QSO and no characters are lost as a result. Try to keep the area around the ceramic resonator cool, to avoid temperature variations.

Parts list

Resistors: all 0.25 W, 5% tolerance

R1	10 megohms (MΩ)
R2	2200 ohms (2.2 kΩ)
R3	270 ohms (270 Ω)
R4	1 megohm (1 MΩ)
R5, R6	1.5 megohms (1.5 MΩ)
R7	1000 ohms (1 kΩ)

Capacitors

C1, C7	100 nF
C2	100 pF
C3	47 pF
C4	1 nF
C5, C6, C10	10 nF
C8	560 pF
C9	820 pF
VC1	10–160 pF variable

Inductors

RFC1	8.2 μH
RFC2	10 μH
RFC3	2.2 μH

Semiconductors

IC1	4069
TR1	VN10KM
TR2	BC640
D1	6V8 Zener

Additional items

- Matrix board (see text)
- 14-pin DIL socket
- Pointer knob
- Sockets (see text)

85 An audio booster for your hand-held

Introduction

The audio output from many hand-held transceivers and receivers usually leaves much to be desired, so this little amplifier was designed to increase the output at minimal expense.

All that is needed is a separate amplifier and bigger loudspeaker. This is accomplished using a single integrated circuit (IC), a few components, and a loudspeaker from the junk box. This circuit will enable the output from your hand-held to be heard easily in a car.

The circuit

This is shown in **Figure 1**. It uses only those components necessary to operate the IC amplifier. VR1 is the preset volume control, and varies the signal coming from the 'External speaker' jack on the hand-held before feeding it into the IC for amplification. C1 blocks any constant voltage present on the input.

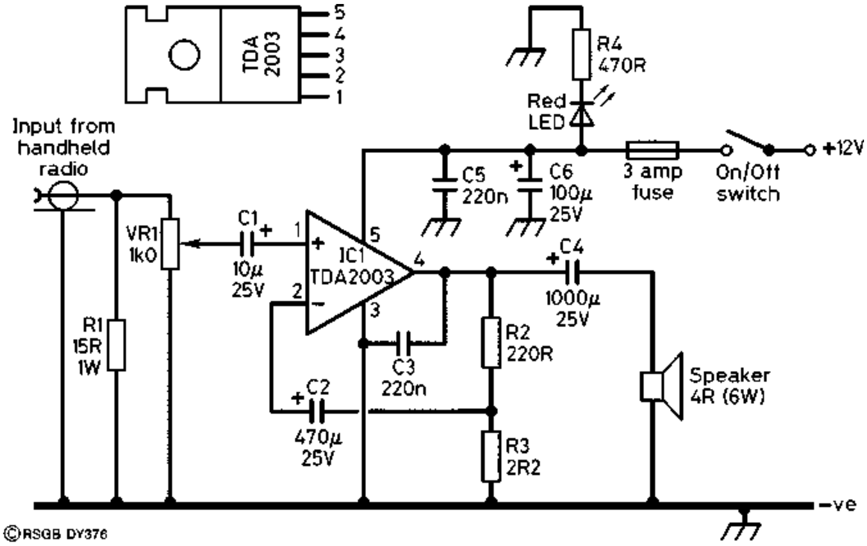
TDA 2003
5
4
3
2
1
R4 470R
Red LED
Input from handheld radio
+12V
C5 220n
C6 100µ 25V
3 amp fuse
On/Off switch
VR1 1k0
C1 10µ 25V
IC1 TDA2003
C4 1000µ 25V
R1 15R 1W
R2 220R
C3 220n
C2 470µ 25V
R3 2R2
Speaker 4R (6W)
–ve
©RSGB DY376

Figure 1 Circuit diagram of the audio amplifier. The power is derived from the cigarette lighter socket and the fuse is in-line with the lead

The IC output comes from pin 4 and is fed via the electrolytic capacitor, C4, to the loudspeaker. The circuit is provided with an on/off switch, fuse and LED to indicate when the circuit is switched on.

Construction

The box is made of aluminium. This is necessary to help to dissipate some of the heat generated by the IC. Do not build the circuit inside a plastic box unless you take special precautions! The IC has a metal mounting tab with a hole, specifically designed to be mounted to a metal box or other metal heat sink. Apply plenty of heat sink compound between the tab and the box, tighten the nut and bolt, and then wipe off any excess compound. The box will get slightly warm in operation.

The size of the box is not specified. You may want to decide on this when you find a loudspeaker. Choose one which will be able to handle 6 W

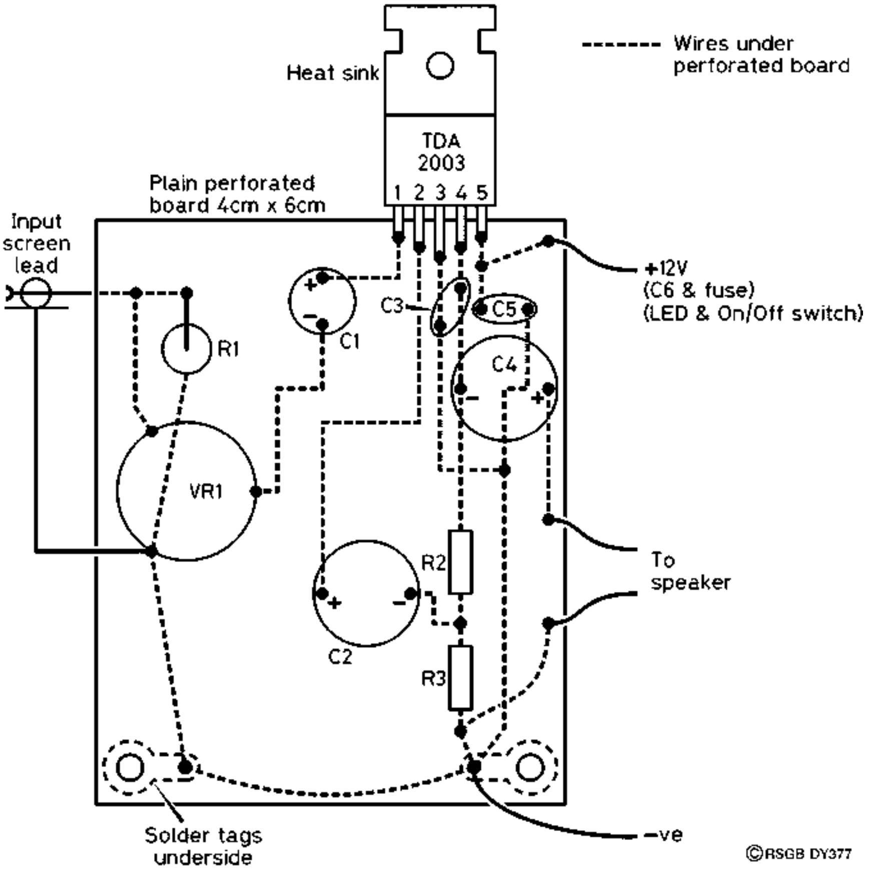
Heat sink
Wires under perforated board
TDA 2003
1 2 3 4 5
Plain perforated board 4cm x 6cm
Input screen lead
+12V (C6 & fuse) (LED & On/Off switch)
C1
C3
C5
C4
R1
VR1
R2
C2
R3
To speaker
Solder tags underside
–ve
©RSGB DY377

Figure 2 Layout of the components within the box

output. Drill all the holes in the box first. Holes for the speaker, input phono socket and the LED. The amplifier can be constructed on ordinary matrix board, which can be mounted inside the box with screws and spacers.

The layout is shown, for your guidance, in **Figure 2**. The components are mounted by pushing their leads through the holes in the board and making connections on the underside. The preset volume control, VR1, is set such that the hand-held's volume control is sufficient to control the final output over a good volume range. Use a screened lead from the 'External speaker' jack socket to the phono plug.

An external power supply is needed for this circuit. The normal dry battery which we usually use for small projects in this book will *not* work here, so you will need a proper mains power supply producing a stabilised 12 V. If you are going to use the unit principally in a car, then the cigar lighter socket can supply this voltage easily. Do make sure that the polarity is correct before you switch on!

When you plug the jack plug from your booster into the 'External speaker' socket on your hand-held, its internal speaker will be muted, so don't think that something dire has gone wrong! Adjust VR1 for a good volume range on your booster, when the volume control is turned on the hand-held.

Parts list

Resistors: all 0.25 W, 5% tolerance, unless otherwise stated

R1	15 ohms (15 Ω) 1 W
R2	220 ohms (220 Ω)
R3	2.2 ohms (2.2 Ω)
R4	470 ohms (470 Ω)
VR1	1000 ohms (1k Ω)

Capacitors

C1	10 μF 25 V
C2	470 μF 25 V
C3, C5	220 nF (0.22 μF) Mylar
C4	1000 μF 25 V
C6	100 μF 25 V

Semiconductor

LED	5 mm Red

Integrated circuit

IC1	TDA2003

Additional items
  Heat sink compound
  Nuts and bolts
  Loudspeaker 4 Ω 6 W
  3 A fuse
  On/off switch (SPST)
  Matrix board 4 × 6 cm
  Solder tags
  Plugs and screened cable for connecting lead
  Aluminium box

86 A grid dip oscillator

Introduction

When an inductor and a capacitor are connected, whether in series or parallel, they form a circuit with a natural (or *resonant*) frequency. The circuit stores energy, and this energy is being constantly shifted from the inductor to the capacitor and back again.

The dip oscillator is a simple instrument used to measure the resonant frequency of a tuned circuit without having to make any direct connection to the circuit. The circuit is more commonly known as the *grid dip oscillator* (GDO), from the days when the active device in the circuit was a valve. The FET or Field-Effect Transistor operates in a way which is very similar to that of the valve, so it is not quite a misnomer to call this instrument a grid dip oscillator, too.

The circuit

The GDO uses a calibrated, tunable FET oscillator in the circuit of **Figure 1**. It has a frequency range of 1.6 to 35.2 MHz in four ranges using a set of plug-in coils, shown in **Figure 2**. When the oscillator coil, L1, is placed near an external resonant circuit, some of its RF energy is coupled into the external circuit. A gain in energy of the external circuit must mean a loss of energy in the GDO circuit, resulting in a change of current through TR1, which is measured by the meter, M1.

The current through TR1 is of the order of 5 to 8 mA, but the change of current may be only a few microamps. To measure a very small

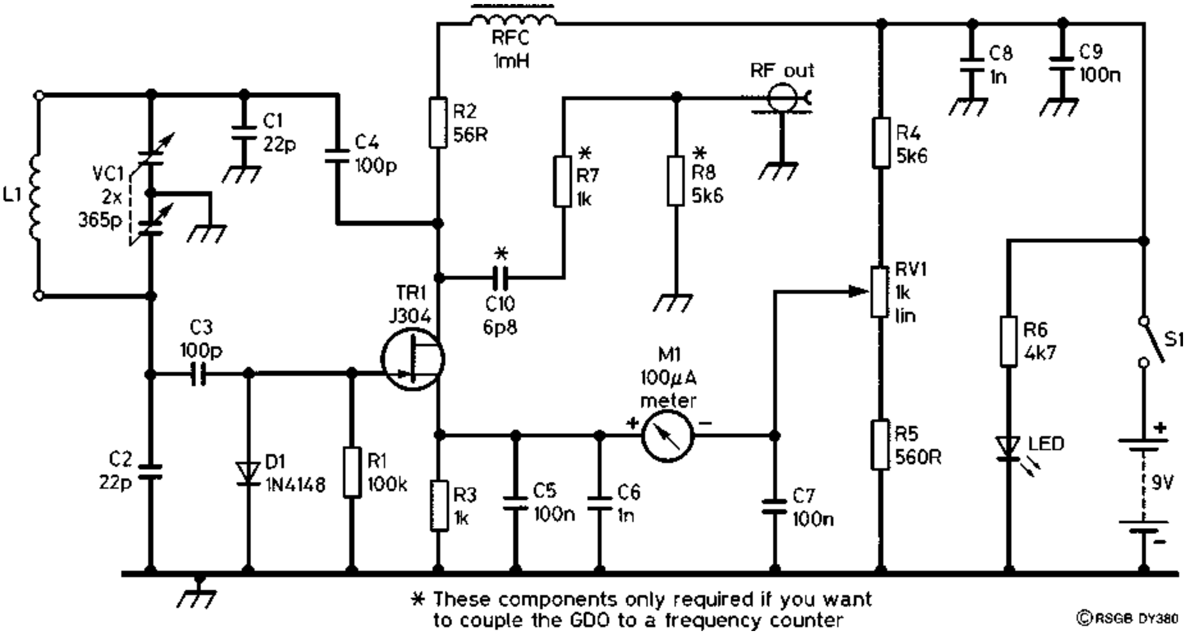
RFC
1mH
RF out
C8
1n
C9
100n
C1
22p
C4
100p
R2
56R
*
R7
1k
*
R8
5k6
R4
5k6
L1
VC1
2x
365p
RV1
1k
lin
*
C10
6p8
TR1
J304
C3
100p
M1
100µA
meter
+
–
R6
4k7
S1
LED
+
9V
–
R5
560R
C2
22p
D1
1N4148
R1
100k
R3
1k
C5
100n
C6
1n
C7
100n
* These components only required if you want to couple the GDO to a frequency counter
©RSGB DY380

Figure 1 Circuit of an FET GDO. The coils are wound on DIN speater plugs, which provide both a plug-in base and a coil former

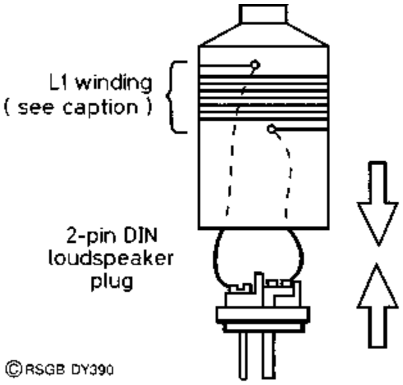
L1 winding
(see caption)
2-pin DIN
loudspeaker
plug
©RSGB DY390

Figure 2 Details of coil construction:
Range 1: 1.6–4.0 MHz 55 turns of 30 SWG
Range 2: 3.3–7.9 MHz 27 turns of 30 SWG
Range 3: 6.3–4.0 MHz 14 turns of 26 SWG
Range 4: 11.9–4.0 MHz 7 turns of 24 SWG.

change superimposed on a much larger standing current, the method of *offset* can be used.

One connection to the meter goes to the source of the FET, while the other goes to a variable offset voltage set by VR1. M1 has a full-scale deflection (FSD) of 100 µA. If the current through TR1 changes, the voltage across R3 changes. When there is no resonance, the voltage at the wiper of VR1 is set to be very slightly greater than that across R3, and there is a 75% FSD meter deflection. When the voltage across R3 decreases very slightly, due to external circuit resonance, a significant ‘dip’ in the meter deflection is produced, hence the name of the instrument.

The circuit is not difficult to make on standard matrix board. Provided you can follow a circuit and translate it into a good component layout, then this project is probably only an evening’s work.

The most important part of the GDO is the tuning capacitor and its associated frequency-calibrated dial. New, air-spaced tuning capacitors can cost you up to £20, so it is worth delving around in junk boxes, or scouring the tables at a local boot sale or rally. The tuning capacitor from an old transistor radio should be perfect. It may even have a slow-motion drive and a dial which can be remarked for the project.

Choose a coil plug and socket arrangement that is practical. Think about crystal holders or phono plugs and sockets. The prototype shown in Figure 1 and Figure 2 used 2-pin DIN plugs, with the coil wound on the outside of the plastic plug cover. Figure 2 shows the coil construction and the winding details.

If you use a variable capacitor, VC1, with a value different from that shown in the parts list, then the frequency ranges will be different. This does not matter, as it will be taken into account during calibration.

Position VC1 so that the dial will be easy to see and to operate, while locating the coil socket as close to it as possible. **Figure 3** shows the traditional layout of the GDO.

Calibration

Because the GDO also radiates a very small amount of energy, a general coverage receiver can be used to calibrate the dial. Don’t try to aim for great accuracy and clutter the dial with marks and figures! If you include C10, R7 and R8 in your circuit, you can connect a frequency counter directly to the GDO and leave it in circuit all the time.

The GDO in use

Always try to place the external tuned circuit with its coil coaxial with the plug-in coil, as shown in Figs 1 and 3. If the coils are at right angles, the GDO may not produce any resonance. Set the offset control to give about 75% FSD and slowly tune L1 through its whole range. If no dip occurs, you may have the wrong coil plugged in. When you eventually find a dip, move the external coil further away until only a minute dip is seen. You may have to retune the dip meter as you do this, but it gives a much more accurate reading of frequency.

Remember that you cannot ‘dip’ a coil by itself – there must always be a capacitor present.

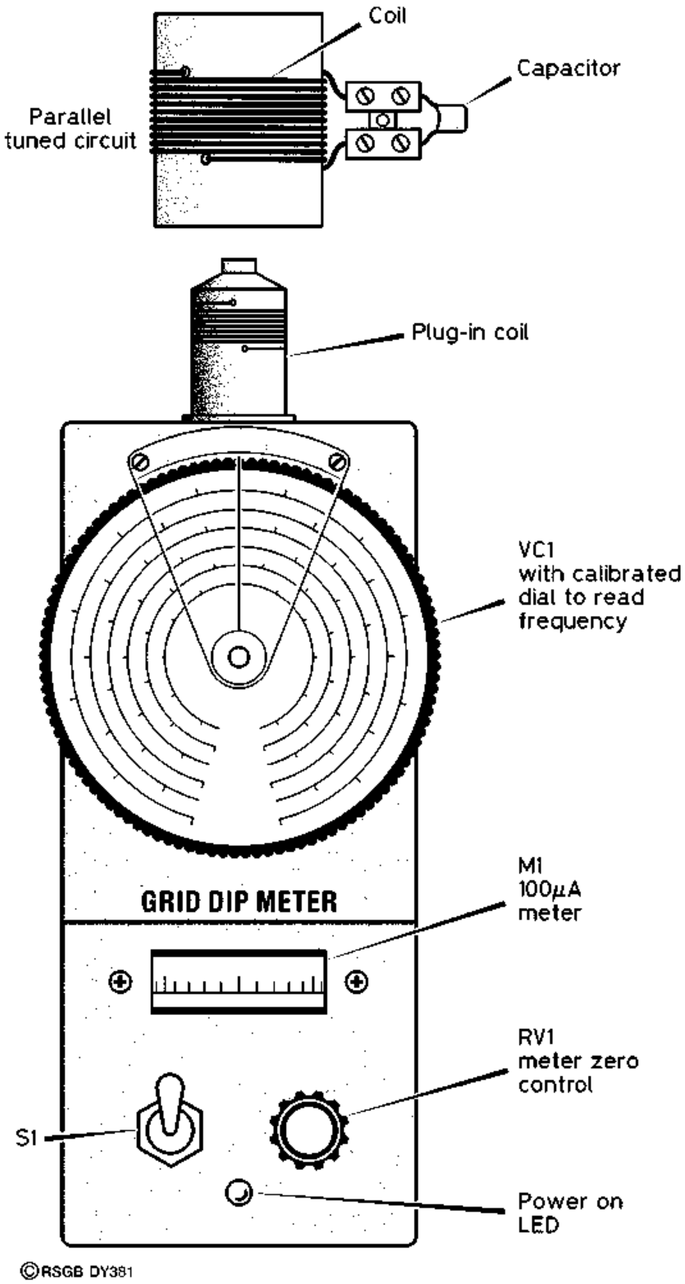
Coil
Capacitor
Parallel tuned circuit
Plug-in coil
VC1 with calibrated dial to read frequency
GRID DIP METER
M1 100μA meter
RV1 meter zero control
S1
Power on LED
©RSGB DY381

Figure 3 Layout of a typical GDO. The dial, meter and the location of the coil to the circuit under test can all be viewed at the same time. It is shown measuring resonance of a tuned circuit

Aerial resonance

Figure 4 shows how to check the resonance of a dipole aerial. Disconnect the coax feed at the aerial and place a short piece of wire, terminated with crocodile clips, across the centre insulator to short together the two ends of the aerial. By placing the GDO close to the shorting link, a dip should be seen on the meter while VC1 is turned. Alternatively, a loop in the element can be made around the coil, as in Figure 4, or the shorting link can be made long enough to loop over the coil. The latter method does not require tampering with the mounting and tensioning of the dipole wires.

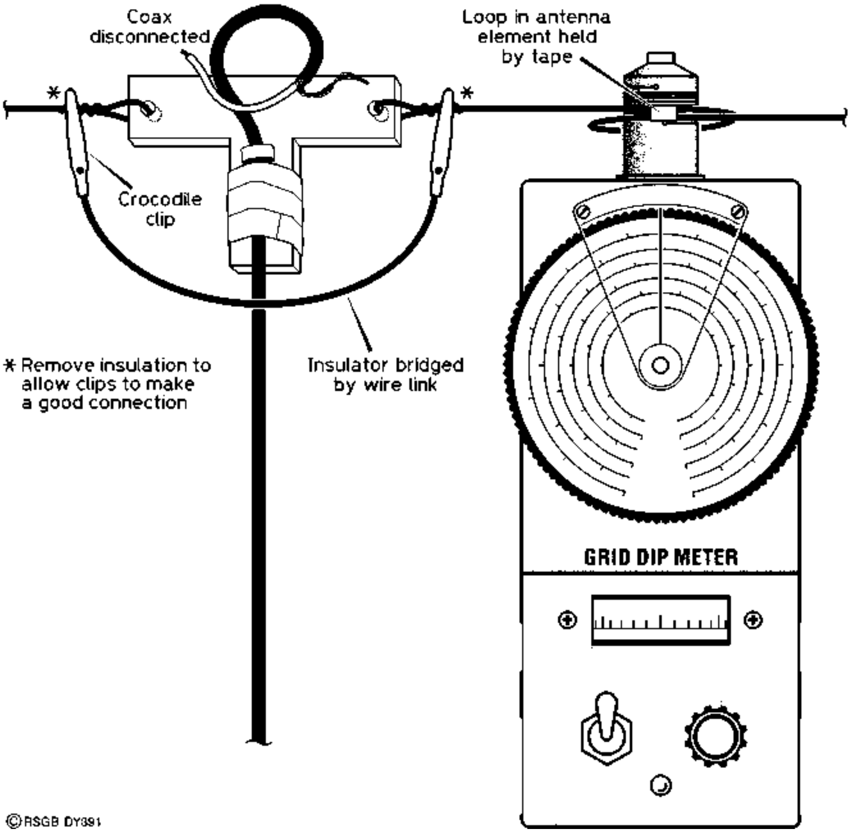
Coax disconnected
Loop in antenna element held by tape
*
*
Crocodile clip
* Remove insulation to allow clips to make a good connection
Insulator bridged by wire link
GRID DIP METER
©RSGB DY391

Figure 4 A wire antenna element can be looped into a single turn coil for increased coupling to the GDO

Parts list

Resistors: all 0.25 W, 5% tolerance

R1	100 000 ohms (100 kΩ)
R2	56 ohms (56 Ω)
R3	1000 ohms (1 kΩ)
R4	5600 ohms (5.6 kΩ)
R5	560 ohms (560 Ω)
R6	4700 ohms (4.7 kΩ)
R7	1000 ohms (1 kΩ)
R8	5600 ohms (5.6 kΩ)
VR1	1000 (1 kΩ) linear

Capacitors

C1, C2	22 pF
C3, C4	100 pF
C5, C7, C9	100 nF (0.1 μF)
C6, C8	1 nF (1000 pF)
C10	6.8 pF
VC1	2 × 365 pF

Semiconductors

TR1	J304 or similar
D1	1N4148
LED	

Additional items

L1	See Figure 2
RFC	1 mH
S1	SPST
M1	100 μA

Source

Components are available from Maplin.

87 A CW transmitter for 160 to 20 metres

Introduction

This very small transmitter is designed to work on any band from top band (160 m) to 20 m, with an RF output of 1 W. It will work on higher frequencies but with a reduced output.

The circuit

The three-transistor circuit is shown in **Figure 1**. It comprises a crystal oscillator using a BC182 transistor which drives a 2N3866 power amplifier (PA) keyed by a ZTX750 PNP transistor. The oscillator and PA are coupled by a capacitor and resistor; this provides a very small amount of positive bias to the PA.

The oscillator can be used as a basic crystal oscillator but, by including a variable series capacitor as shown in Figure 1, the crystal frequency can be 'pulled' slightly, making the oscillator a *variable crystal oscillator* (VXO).

Construction

The PCB layout is shown in **Figure 2** and in the photograph. Although the prototype was built around a PCB, this circuit is equally amenable to

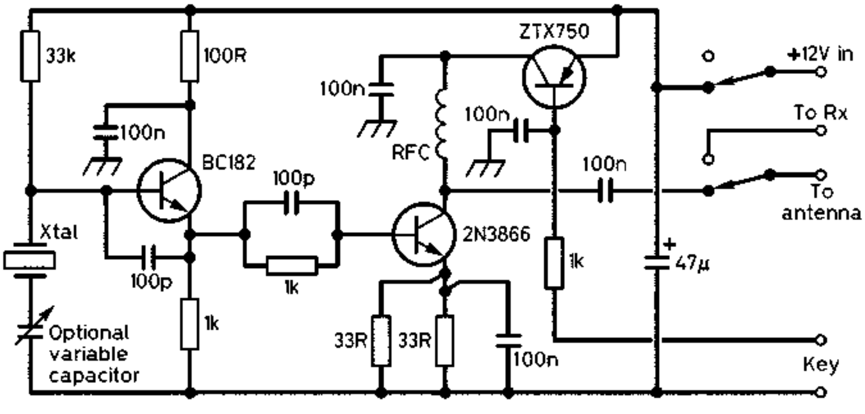
33k
100R
100n
BC182
100p
Xtal
100p
1k
1k
Optional
variable
capacitor
ZTX750
100n
RFC
100n
2N3866
33R
33R
100n
1k
100n
47μ
+12V in
To Rx
To
antenna
Key

©RSGB DY396

Figure 1 Transmitter circuit diagram

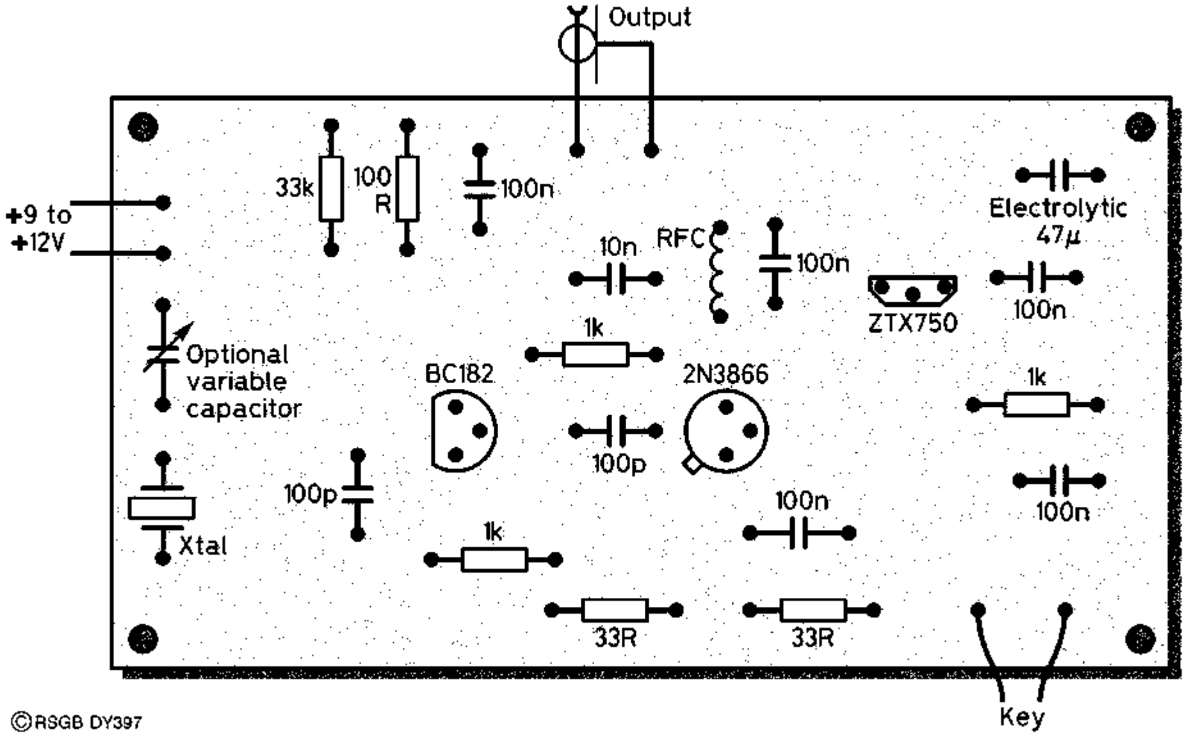
Output
+9 to
+12V
33k
100
R
100n
10n
RFC
100n
Electrolytic
47µ
100n
ZTX750
1k
Optional
variable
capacitor
BC182
2N3866
1k
100p
100p
100n
100n
Xtal
1k
33R
33R
Key
©RSGB DY397

Figure 2 Component placing on the PCB. The external connections are also shown

construction on a matrix board. Populating the PCB is very simple, and you can expect to be able to do this in about one hour. The radio-frequency choke (RFC) is made by winding 10 turns of 33 SWG wire on a ferrite bead. The enamel coating of the wire is intended to vaporise when soldered into the board, thus obviating the need to remove the enamel manually with a knife or sandpaper. However, if you do have problems with the PA either not working or keying intermittently, it is suggested that you investigate the RFC connections immediately!

If you decide to make the VXO version, you will have to cut the track between the crystal and earth, and connect the variable capacitor (250 pF) across the break.

In use

After performing the usual checks on the accuracy of your circuit building and the wiring of the external components, it is time to connect a 12 V battery between the points shown in Figure 2. Do not switch on yet. An aerial needs to be connected to the output via an ATU and a crystal, of frequency matching that of the aerial, fitted. The variable capacitor should give you a tuning range from about 14.058 to 14.064 MHz. 'Netting' (the process of tuning your transmitter to the same frequency as that of a received station) is achieved simply, because the oscillator is always running, and the leakage of the signal (despite the fact that the PA is not powered) is sufficient to bring the two frequencies to zero beat.

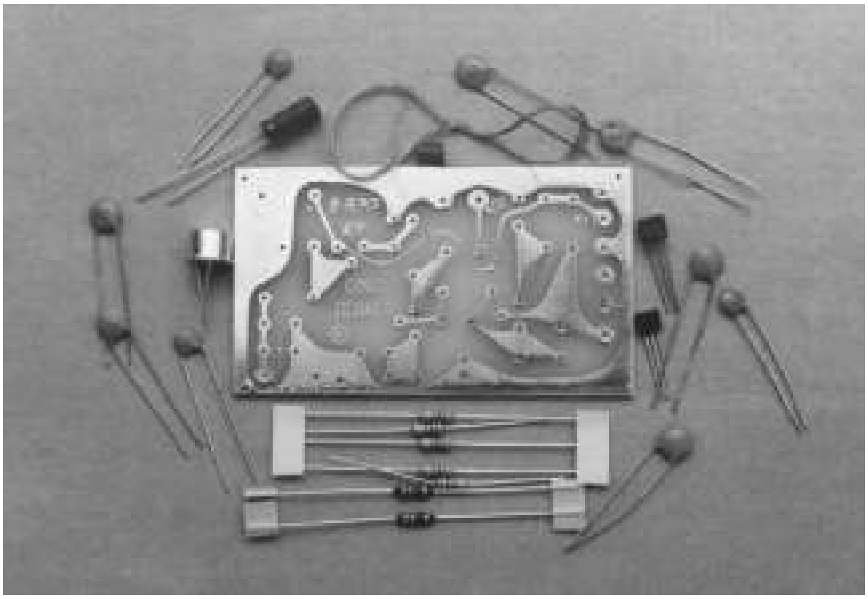

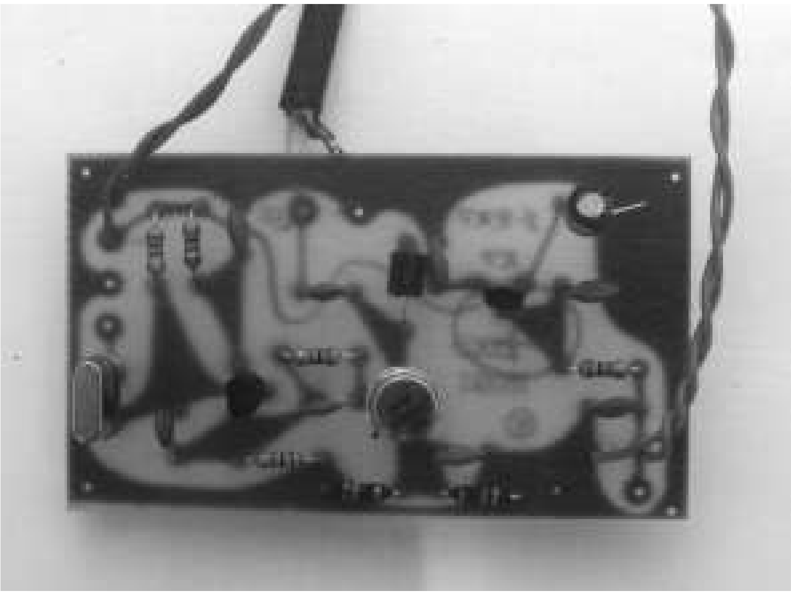

Warning

Note that the transmitter has no filtering; harmonics are not suppressed. It is strongly recommended that you use this transmitter in conjunction with the excellent low-pass filter described in the project *A 7-element low-pass filter for transmitters*, described elsewhere in this book.

88 Matching the end-fed random-wire aerial

Introduction

Many amateurs who do not have the space (or money) for a multi-band beam aerial, make use of the simplest possible alternative – the longest piece of wire that they can erect, with its end connected to the transceiver or receiver by an aerial tuning unit (ATU).

The length of the wire is not of major importance. Any length between 10 m and 80 m, with bends if necessary, will suffice. A good earth connection to the radio is just as important. Bends in the aerial wire can have some interesting effects on the directional properties of the aerial; V- or L-bends, or even a square shape are permitted. The only thing *not* permitted is to fold the wire back on itself in a tight hairpin bend!

Longish wire aerials

The term ‘long wire’ is usually used (incorrectly) to describe an end-fed aerial. How long is a piece of string? It depends what you mean by ‘long’. In aerial parlance, it means ‘long with respect to one wavelength’. Again, this depends on the band you are using. A long wire at 20 m is somewhat different from a long wire at 160 m. However, if you have sufficient real estate for a long wire on the 160 m band it must, by definition, be a long wire on all the other bands, too!

This should put you on your guard when analysing published data about feed-point impedance and the directional properties of a long wire aerial. Such theoretical data relate to a *real*, *ideal* long wire which is straight, horizontal, very high above perfect (conducting) ground, and not obstructed in any way. So your aerial doesn't quite match these criteria? Join the club!

Don't let this dampen your ardour when it comes to evaluating what the longish wire can do for you. The following should explain why.

Feed-point impedance

The impedance at the end of a longish-wire aerial can vary from a few tens of ohms to several thousand ohms, depending on the frequency in use and the wire length. It is also affected by factors such as bends, height above ground, proximity to buildings and wire diameter.

The actual value doesn't matter, provided we can make the aerial *appear* to have a 50 Ω impedance at the aerial socket of the transmitter. This process is what we call *impedance matching*, or simply *matching*. It maximises the power transfer from the transmitter to the aerial, and from the aerial to the receiver.

That is why an aerial tuning unit, or ATU, is almost (but not necessarily) obligatory.

The ATU

Many commercially produced HF receivers and transceivers have single 50 Ω coaxial sockets as their one and only means for connecting an external aerial. This means that an external aerial should have a 50 Ω feed impedance if it is to work efficiently, and it rules out most of the aerials being used by amateurs on the HF bands. Some means is necessary to change the aerial feed impedance to 'match' that of the transceiver. Such impedance-matching, or Z-matching (because Z is the symbol for impedance, just as L is the symbol for inductance) is the rôle of the ATU.

These can be bought and will accommodate either an end-fed or a coax-fed aerial. They can be bought ready for use or in kit form. Whether you want one for receiving only, or for use with a low-power (QRP) or high-power (QRO) transmitter, will determine what you need and how much you pay.

A simple single-band ATU

The simplest form of ATU is shown in **Figure 1**. It is simply a parallel LC (coil and capacitor) circuit, resonant at the chosen frequency, with taps on the coil for the aerial and the coaxial feed to the transceiver. If we assume the circuit is resonant, a high impedance exists at the top of the coil, and a low impedance at the bottom.

We said earlier that the end-fed longish wire presented an impedance which was high (or at least higher than 50 Ω). This explains why the aerial is tapped to the coil near the top, where the impedance is high, and the 50 Ω coax is tapped near the bottom, where the impedance is low.

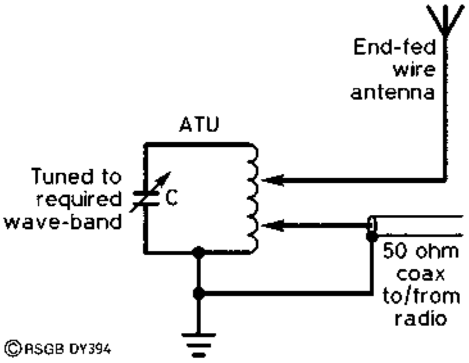

Figure 1 Parallel tuned circuit as single-band ATU for end-fed longish-wire antenna

Because the feed-point impedance of the aerial changes with frequency, so must the point at which the aerial is tapped to the coil to achieve impedance matching. The value of C must be changed also, to ensure that the circuit is resonant, and the 50 Ω tap will require tweaking also.

Setting up an ATU is quite simple. Make up an LC parallel-tuned circuit consisting of 50 turns of enamelled copper wire on an empty 35 mm film plastic container (or similar), tuned with a 500 pF variable capacitor. Make sure the enamel is removed from the ends of the wire before soldering.

Solder the inner wire of the coaxial cable from the radio to the first or second turn of the coil from its grounded end. Then solder the braid to the grounded end. Connect the aerial about one-third of the way down the coil from the top, removing the enamel at the connection point.

Adjust the variable capacitor for maximum noise or signal strength in the receiver. Then, try different tapping points from the aerial, to maximise the signal again. This matches the aerial impedance to that of the tuned circuit. Repeat the process with the coax tap, thus matching the impedance of the radio and the cable to that of the tuned circuit.

You will no doubt find that the tapping process on the coil was not easily accomplished, especially when the enamel must be removed at each tapping point without shorting adjacent turns together. It is therefore logical to produce a design where the taps have been prepared during the winding of the coil, and are selected with a rotary switch. To this end, the following multi-band design is described.

A multi-band ATU

The same type of coil design around a 35 mm film container is used (see **Figure 2**). The tapping points can be prepared in advance by a little judicious planning. Dismantle your original coil and measure the length of wire on it. It will be a little more than a calculation of $50 \times \pi D$ would suggest (where D is the diameter of the film container), due to the lack of tension in the coil and the wire diameter itself. You will need aerial tapping points at turns 1, 2, 3, 5, 10, 15, 20, 25, 30, 35, 40 and 45, corresponding to all the bands from 28 MHz to 1.8 MHz. The coaxial cable tap is fixed at turn 2 (an acceptable compromise). All the turns are counted from the earthy end.

Cut another piece of enamelled copper the same length as you used originally. Then, with the aid of an ordinary calculator, work out the positions of the points where the enamel must be removed for the taps. For example, turn 15 will have to be made $\frac{15}{50}$ of the way along the wire, turn 20 tap made at $\frac{20}{50}$ of the way along. So, if the length of wire is, say 4.7 metres, the two taps in question will be made at $\frac{15}{50} \times 4.7 = 1.41$ m and $\frac{20}{50} \times 4.7 = 1.88$ m from one end. This must be repeated for each of the tap positions, and the enamel removed ready for the wire to be soldered to it. With a

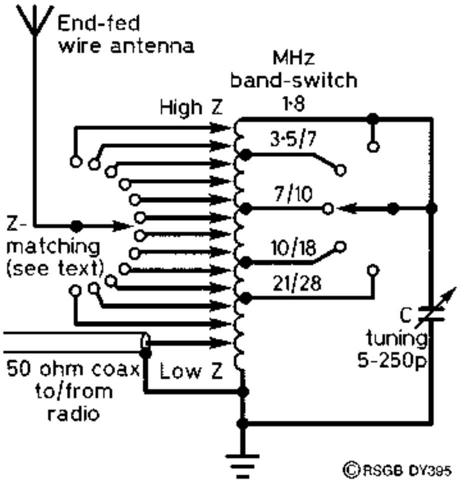

Figure 2 Simple multi-band ATU for end-fed longish-wire antennas

soldering iron, tin each tap point while the copper is shiny, thus ensuring a good, low-resistance connection. Do not solder on the taps yet.

The coil can now be wound as before. The taps can be soldered on, taking the lead from each one to the wafer of a single-pole 12-way rotary switch, the pole being connected to the aerial.

The tuning of the ATU is carried out by the same 250 pF capacitor, with a single-pole 5-way rotary switch used to select the band. Its tapping points will need to be chosen manually, using the method described earlier. Don't attempt to make new tapping points on the coil for this – use the taps available on the wafer of the other rotary switch, and find which is optimum for each band.

Notes

- For the aerial, use PVC-covered stranded tinned copper, of size 16/0.2 mm or 24/0.2 mm.
- Make the wire as long as possible, but anything over 10 m should be OK.
- Keep the wire as high as you can, in the clear and away from obstructions.
- Don't worry about bends, but don't use hairpin bends.
- Use a good insulator to attach your aerial – anything plastic will do.
- Anchor the wire near the point of entry to the building, but use a U-bend to prevent ingress of water.
- The wire can be brought in through the corner of a window, the PVC acting as an insulator. If you must drill a hole in the brickwork, make sure it slopes upwards from outside, so that water is deterred from entering.
- Use a good RF earth (as opposed to an electrical earth) such as half a dozen bare copper wires buried under the lawn in a fan shape. They should be joined together at the point of the fan and strapped to the earth connector of your ATU.

L2443
N87
999

UNIVERSITY OF N.C. AT CHAPEL HILL
00018843218

THE ANKH

African Origin of Electromagnetism

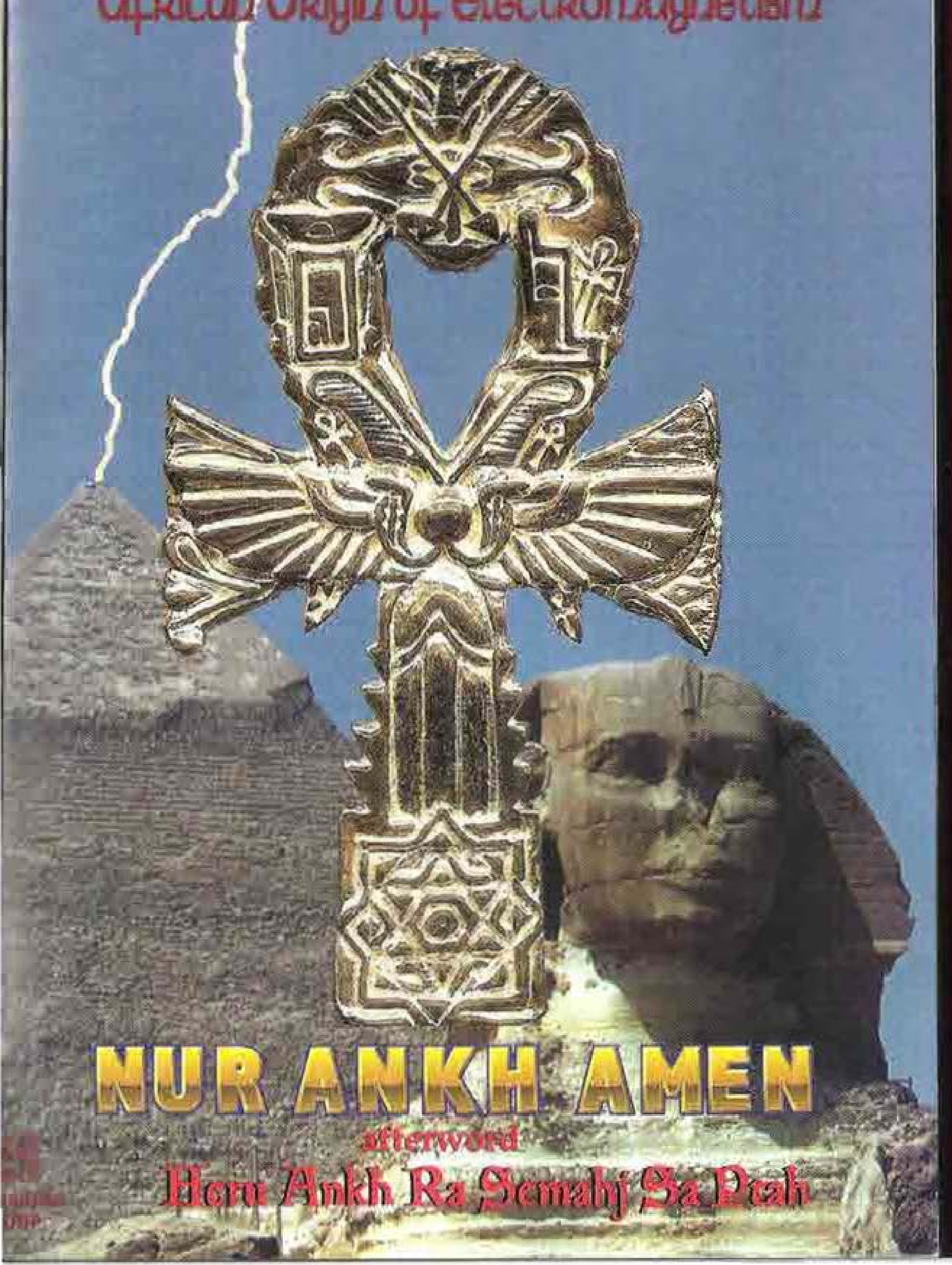

NUR ANKH AMEN

afterword

Heru Ankh Ra Semahj Sa Ptah

Adrift upon an electron sea
Dying of thirst how could this be?
With water everywhere,
I search for money without a care
Never once did I stop to think
maybe it's time to take a drink
Nur

BL2443
.N87
1999

THE ANKH

3-14-01AD

African Origin of Electromagnetism

NUR ANKH AMEN

AFTERWORD
HERU ANKH RA SEMAHJ SA PTAH

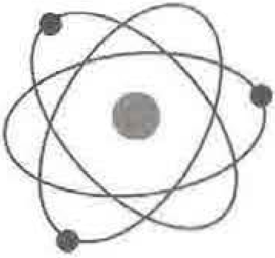

A&B Publishers Group
1000 Atlantic Avenue
Brooklyn, New York
11238

THE LIBRARY
THE UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL

Acknowledgments

The Author wishes to thank the following for their contribution, although the views expressed in this publication does not represent their own.

Lee Boltin Picture Library
National Council of Churches of Christ
The Egyptian Organization of Antiquities

Contents

Introduction 9

Author's Notes 13

The Ankh and Popular Myth 16

Ankh Science and the
History of European Electronics 25

The Oscillator 45

Melanin 55

The Spiritual or Electromagnetic Worlds 60

Osiris the First Resurrected Christ.................64

The Heavens 85

African Physics: 101

The Science of Kemit 107

The Ankh of Science 117

Conclusion 121

Afterword 123

Glossary 129

References 131

Index 133

Introduction

DUE TO THE INCREASING INTEREST IN Egyptian science, the achievements of the Nile Valley civilization of Africa have been brought into focus. Depictions on tombs, pyramids and temples, provide a glimpse of the rich spiritual, cultural and scientific wisdom of Kemit.

These astonishing achievements point to the fact that Africa invented the Modern World. The origin of every scientific discipline, from algebra to zoology can be traced to the Nile Valley civilization in Africa. However in the face of overwhelming evidence to support an African origin of science and technology, the racist scientific community continues to deny its recognition. African Egyptologists fighting to return Egypt to Africa, have only been partially successful against the repressive colonial Academic Power that endeavors to hold our minds in slavery. The truth is the Real Estate over which we must be prepared to die, for a lie is far more dangerous, not only to ourselves but to our collective consciousness.

Now the war has taken a new turn on a different front: that of electronics. The most powerful tool of modern technology is electronics and its origin has, like most things, been attributed to the European minds of the early eighteenth century. With damaging consequences for our youth, the discovery of electronics has been claimed as the sole domain of Europeans without a single mention of any prior knowledge in Africa.

The European concept of discovery should be examined in the light of its political consequences on others, for a pattern exists which invariably results in domination and exploitation. The pattern eradicates the truth, and replaces it with lies designed to promote White supremacy. This is the prime directive: to maintain control over all other civilizations, by academic interference and distortion of traditional religious beliefs.

The discovery of the New World by Christopher Columbus is an example of this concept in action. Its awful brutality is fully documented and the results are evident on the native population even today, five centuries later. Discovery has nothing to do with original use, it is a means by which a Claim can be pursued and adjudicated by Colonial Powers.

The strategy is still used in all areas of endeavor, to allowing the expropriation of wealth, whether it be land, inventions, art or culture.

Evidence is mounting which proves that the so-called discoverers of electronics like Volta, Hertz and Benjamin Franklin, were merely Columbus's of the electronic frontier. The evidence also points to an extensive cover-up, which became imperative because of the legal aspects governing discovery. They are today known as Patent Laws. Subsequent knowledge of an invention could preclude any financial gains, rendering them useless to the new profiteer.

The cover-up and denial of prior electronic use in Africa allowed many inventions to be patented creating massive profit for the inventors and this new industry. One lie begets another and academia's role is to perpetuate these lies for future generations. However, to quote Professor Leonard Jeffries, "the truth is our most powerful weapon."

The parallel between the sudden rise in electronic inventions and European Egyptological excavations (grave-robbery) has been suspiciously ignored. It is no coincidence that the most important discovery in electronics, the pile or battery, and Napoleon's scientific expedition to Egypt occurred almost simultaneously.

Foremost in Egyptian electronics is the ☥ Ankh, and although the European Egyptologists claim that what it represents is still unknown, this is convenient ignorance.

The implication of Ankh Science is enormous and can only be fully understood through the spiritual awareness of the African, which allows him to appreciate and develop the technical sophistication with regard to its importance. I say Africans, because the ☥ Ankh is a symbol of our unique spiritual heritage. Although physical evidence of African electronic devices have long decayed or were expropriated for later examination by the plunderers of Egypt in the 18th and 19th centuries, there still remains sufficient proof of wide-spread knowledge and use in Kemit, prior to the Greek and Assyrian presence in the Nile Valley.

This book will be of special interest to those of us who have been educated in traditional electronics and science courses, and felt disenfranchised as White inventors were lauded for their contributions to a science our ancestors originated.

Author's Notes

IT IS THE AUTHOR'S HOPE IN writing this book, that serious investigation of this subject will follow, with the aim of restoring our traditional greatness and spirituality in these original African sciences.

My introduction to the ☥ ankh came in 1986 when I, as a jeweler, was asked, to make a silver ankh for a customer. I became extremely fascinated with the symbol that seem to communicate something about my ancestral past that resonated with my very soul. Trips to the local library and questions to so-called experts all gave a limited view on the subject. With the aide of my spiritual guide, Sankhamen, I investigated the Ankh for six years; studying every mention, drawing or article of jewelry. I came to the realization that the ☥ Ankh was the ultimate spiritual icon.

My traditional educational history was a disaster. I had difficulty understanding some concepts, because my professors were more concerned to label me as incompetent in science, rather than trying to educate me. The curriculum was designed to promote White supremacy and discourage deeper understanding. There was something they did not want me to know, something that seemed to arouse great fear and anxiety in them. Somehow they knew I was different, that I was not buying "the great White scientist routine;" I wanted the truth. A truth that I later found them incapable of supplying. Was I asking to be educated or spoon fed?

I decided to wean myself from the American Educational System and pursue a deeper understanding of science through spiritual discipline and divine inspiration.

Through these methods I was taken to the heights of celestial wisdom and knowledge so vast in scope, that they could not be explained in the scientific precepts or language used today. My impressions are that spiritual precepts are a higher order of the same scientific concepts but separated by a yet unrecognized form of sensory perception. It was like trying to see with one's nose or attempting to describe a scene based solely on one's sense of hearing. No wonder the spiritualists were frustrated, they could not express to anyone what they knew to be true from these intensely personal spiritual experiences.

I have come to the conclusion that we can no longer allow ourselves to be educated in the same manner, by a system based on cranial capacity, while ignoring our superior spiritual faculties, simply because they cannot be expressed in language to the satisfaction of our academic slave-masters. Divine inspiration is a supernatural form of communication and the ultimate goal of a truly liberated African Education System, a system that I contend is still intact today. Our professors and teachers are our ancestors, alive in a realm where belief is the door and the ☥ Ankh is the key.

What about credibility? Why do we need the Established Academic Apparatus to legitimize our theories, and why do we need degrees to expound on those we present? Why? Let the truth speak for itself.

They do not try to disprove the substance of what we are saying, instead they mount personnel attacks on us to discredit us, or worse , they will use one of our own. Maybe a respected Black Egyptologist, whose authority will be threatened by this revolutionary view of Kemit, which he, with all his scholarly acclaim could not bring to light.

The Ankh and Popular Myth

Anyone familiar with the ☥ Ankh will tell you it is the key of life and for some that explanation is sufficient, but for others the next question is, "what is life ?"

1. *Life:*
   - a: the quality that distinguishes a vital and functional being from a dead body.
   - b: a principle or force that is considered to underlie the distinctive quality of animate being compare
2. *Vitalism:*
   - a: an organismic state characterized by capacity for metabolism, growth, reaction to stimuli, and reproduction.
   - b: the sequence of physical and mental experiences that make up the existence of an individual.
   - c: one or more aspects of the process of living
3. *Biography*
   - a: spiritual existence transcending physical death.

Life is harmonious coordination of energy and matter, of such sophisticated construction and efficiency as to appear divinely ordained and self sustaining.

How could an object like the ☥ Ankh come to be representative of life? The only conclusion is that the Ankh is a mechanism that possesses a living characteristic.

An examination of the word and its usage in Egypt will reveal that the Ankh meant more than organic life.

By permission. From Webster's Ninth Collegiate Dictionary 1991 by Merriam Webster Inc., publisher of the Merriam Webster dictionaries.

ankh	𓋹𓈖𓐍	life, to live, living
ankh-	𓋹𓊽𓌀	"life, stability, blessing [or, good fortune]
ankh-	𓋹𓆓	ever-living," living forever
ankhu	𓋹𓀀	human life, male.
ankh	𓋹𓀭	the living incarnation of God.
ankh-	𓏙𓋹	"to whom life is granted." life giving.

ankh 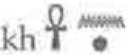 a living entity.

ankh 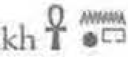 the house of life

Ankhet 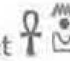 the land of life, the etheric realm.

ankh 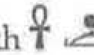 life in a tomb.

The Encyclopedia of Hieroglyphics states:
The ankh is commonly known to mean life in the language of Ancient Egypt, where numerous examples have been found that were made from metal, clay and wood. It is usually worn as an amulet to extend the life of the living and placed on the mummy to energize the resurrected spirit. The gods and kings are often shown carrying the ankh to distinguish them from mere mortals. The ☥ ankh symbolized eternal life and bestowed immortality on anyone who possessed it.

There is no consensus among Egyptologists as to what object the ankh represents.

It is believed that life energy emanating from the ☥ ankh can be absorbed by anyone within a certain proximity. An ankh serves as an antenna or conduit for the divine power of life that permeates the universe.

The Amulet is a powerful talisman that provides the wearer with protection from the evil forces of decay and degeneration.

The Encyclopedia of Ancient Egypt defines the ankh as:

The symbolic representation of both physical and eternal life. The ☥ ankh is the original cross with a loop that was held by the gods. It is associated with Isis and Osiris in the Early Dynastic Period, and although the knowledge of what object it signified was lost, it remained a hieroglyphic symbol of life to the Christian era.

The ankh was used in various religious and cultural rituals involving royalty. In the Treasures of Tutankhamen, the ankh is explained away as:

A stylized version of a sandal strap.

Zolar's Encyclopedia of Omens, signs and superstition says:

The circle symbolizes eternal life and the cross below it represents the material plane. The ☥ ankh is called the "Crux Ansata," it is of Egyptian origin and can be traced to the Early Dynastic Period, appearing frequently in artwork of various material and in relief, depicting the gods.

It is usually held to the nose of the deceased king by the gods to represent the breath of life given in the after-world. The ☥ ankh resembles a key and is considered the key to eternal life after death.

Its influence was felt in every dynastic period and survives as an icon possessing mystical power throughout the Coptic Christian Era.

Amen: The secret waters of the Great Pyramid.

The ankh was the key used to control the flow of water in the underground flow of the Nile, the way a tap is used in today's faucets.

The wire ankh has been used in Dowsing, which is the pseudoscience of detecting water underground.

Stonehedge: Large volts of energy have been emitted from a megalith to a wire shaped in the form of an ankh.

The ankh possessed by each god had power associated with that god. The ankh of Anubis related to the protection of the dead, that of Sekmet war, Hapi related to the living waters of the Nile and Amen, the spirit god, the breath of life. It is being suggested that the ankh box found in the tomb of Tutankhamen was actually a mirror case. Although no such mirror was found and mirror cases were not completely enclosed, as only the mirror face needed protection, it is still referred to as a mirror case.

The ankh box of Tutankhamen

The box contained Tutankhamen's ankh device, which today is no doubt sitting in the private collection of some wealthy European collector. Some say it's in the Vatican.

One very curious theory suggest the ankh ☥ produced high pitch ultrasound inaudible to our human ear but effective in frightening wild animals like jackals and hyenas. This would have been especially useful in the wilderness at night, protecting travelers from unseen dangers.

Not only were the living protected but the dead as well. Ultrasound has the effect of frightening insects, rodents and maggots that would devour the bodies of the deceased. By producing tiny air bubbles in the body of pests, the ultrasonic environment is very uncomfortable.

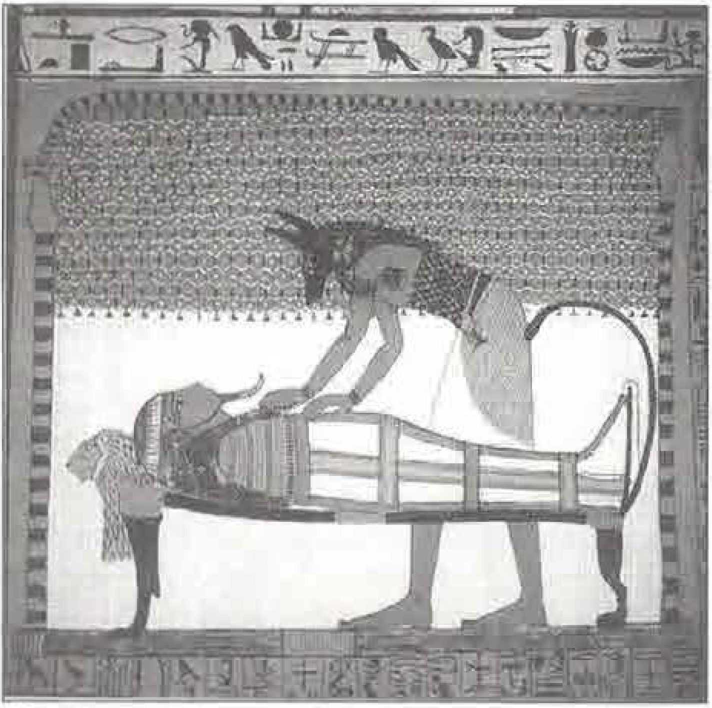

The deity associated with the dead, Anubis, is a jackal-headed god with his ears pointed up, able to hear what is inaudible to man.

Although the basic shape has remained constant, closer examination reveals subtle differences in design and construction of ankhs at different periods. The different shapes suggest refinement or varied usage.

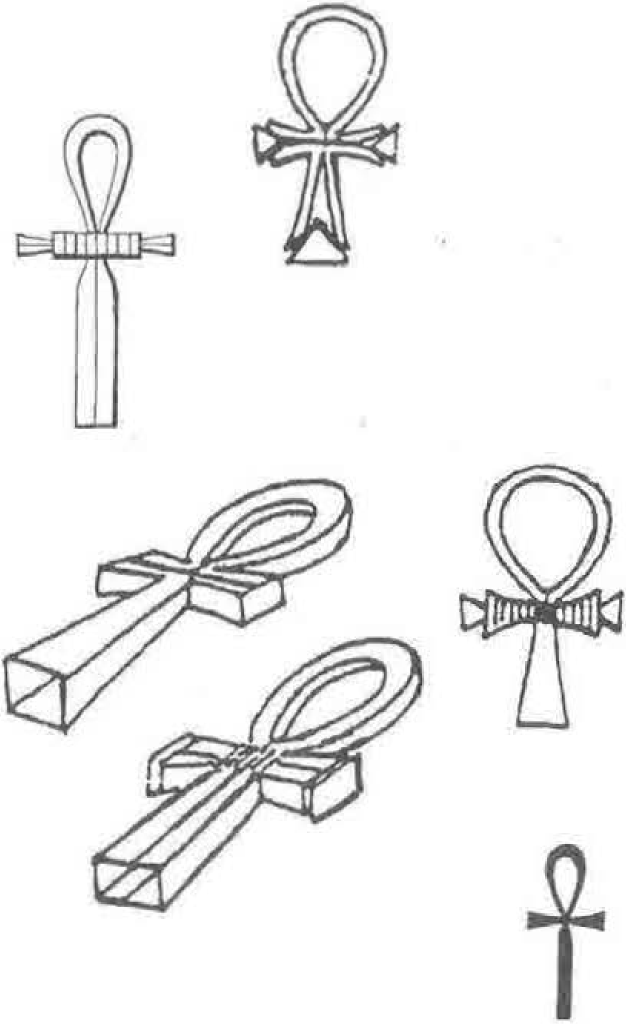

The ☥ ankh is often depicted in combination with other symbols, but it is usually attached to its power source, the djed.

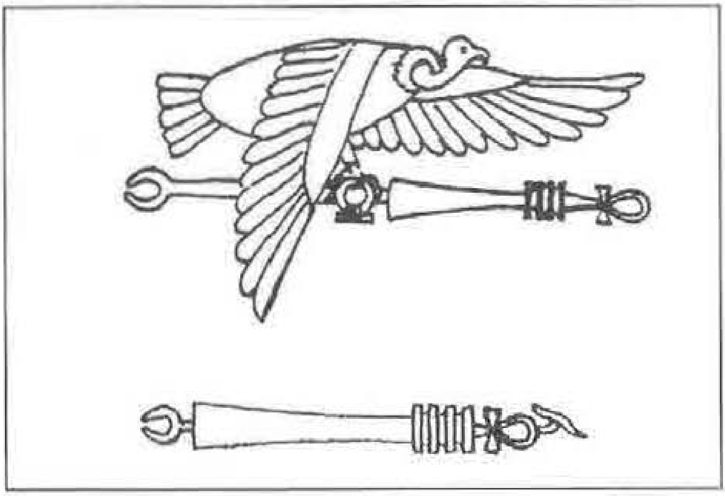

The gods Ptah and Khonsu hold this combination scepter.

Ankh Science and the History of European Electronics

THE EARLIEST RECORD OF studies in electrical and magnetic phenomena by Europeans, were done by priests of the Jesuit Schools in the 17th Century. The mysterious properties of amber, which attract small pieces of paper when charged, and the lodestone or magnet were investigated.

The core knowledge of these schools relied on the works of Greek philosophers like Pythagoras and Aristotle, who were students at the feet of the Africans in Kemit. The electron is named after the Greek word for amber.

Experimentations of the 17th Century culminated in the discovery of means by which large static electrical sparks could be produced by various contraptions which reproduced the effect of rubbing one's shoes on the carpet and touching a door knob.

An accidental discovery enabled electrical charges to be stored in what is called a leyden jar or capacitor in 1742. This device stimulated interesting experimentation all over Europe and as far away as Japan. Electrical current could be passed from person to person in a chain and a terrific lightning bolt could be produced by the discharge.

A leyden jar or capacitor is constructed by placing an insulating material, usually wood or glass, between two sheets of metal. This invention was known 5000 years before in Ancient Egypt. Shrines were built of wood and covered with sheets of gold inside and out, then charged to protect sacred objects placed inside.

One of the most famous of these shrines was the Ark of the Covenant. Built by Moses the Egyptian, to protect the tablets of the Ten Commandments and to serve as a reminder of the presence of God. Once charged this portable shrine was potentially lethal if touched, as was evident in the Old Testament account. A priesthood specially trained to handle the shrine wore special clothing for protection and used a rod to ground or discharge the Ark. When the shrine was discharged by the priest it could be handled safely.

Exodus 25

10 "They shall make an ark of acacia wood; two cubits and a half shall be its length, a cubit and a half its height

11. And you shall overlay it with pure gold, within and without shall you overlay it, and you shall make upon it a molding of gold round about.

12 And you shall cast four ring of gold for it and put them on its four feet, two rings on the one side of it, and two rings on the other side of it.

13 You shall make poles of acacia wood, and overlay them with gold.

14 And you shall put the poles into the rings on the sides of the ark, to carry the ark by them.

15 The poles shall remain in the rings of the ark; they shall not be taken from it.

16 And you shall put into the ark the testimony which I shall give you.

17 Then you shall make a mercy seat of pure gold; two cubits and a half shall be length, and a cubit and a half its breath.

18 And you shall make two cherubim of gold; of hammered work shall you make them , on two ends of the mercy seat.

19 Make one cherub on the one end, and one cherub on the other end; of one piece with the mercy seat shall you make the cherubim on its two ends.

20 The cherubim shall spread out their wings above, overshadowing the mercy seat with their wings, their face one to another; toward the mercy seat shall the face of the cherubim be.

21 And you shall put the mercy seat on the top of the ark; and in the ark; and in the ark you shall put the testimony that I shall give you.

22 There I will meet with you, and from above the mercy seat, from between the two cherubim that are upon the ark of the testimony, I will speak with you of all that I will give you in commandment for the people of Israel.

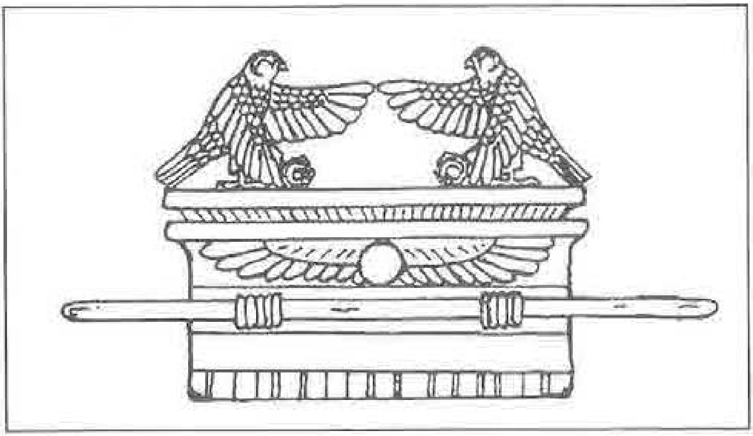

The Ark of the covenant

> And when they come to the threshing floor of Chidon, Uzzah put out his hand to steady the ark, when the oxen stumbled. And the wrath of the Lord was kindled against Uzzah; and he smote him because he put forth his hand to the ark; and he died there before God.
>
> *1 Chronicles 13: 9*

When the positive and negative terminals of a charged capacitor or shrine where brought close to each other, a spark could occur and if the capacitor is attached to a stable source of current, charging and discharging occur at a specific rate.

The high voltage necessary to charge the shrine or ark was provided by the atmospheric potential difference between the sky and the ground. The electrical gradient above the earth is 60 to 100 volts per meter in fair weather and increases to 6000 to 12,000 volts per meter under a thunderstorm.

Shrines were usually placed on high mountains, natural or artificial, to take advantage of this effect. Moses was no doubt an eye witness to the ark in the King's Chamber of the Great Pyramid at Giza. The sarcophagus contained the electrified mummy of the God-King Khufu as part of a high voltage circuit.

With this tremendous power, the God-king could still communicate his will in this world from the ankhet (the other-world). So powerful was this achievement that certainty of an afterlife was firmly engraved in the psyche of Egypt, lasting thousands of years. Communication was not necessarily in words, like radio, it was a spiritual communion or revelation which was impressed upon the mind resulting in a prophetic experience.

They were communications of a highly moral nature, coming from beings who had been freed from the sinful bodily nature and viewed us as being almost like children.

The pillaging and grave robbing in Egypt continued for hundreds of years by foreigners of every description. Total disregard for the African was the only rule, so much so that the mummies or corpses were ground and boiled to make potions for sale. The European trade in mummies in the 16th and 17th century took on a feverish pace, because they were thought to have medicinal value. The melanin rich corpses of your ancestors were eaten by these savages.

By the 18th Century, the plundered skeletal remains of Egypt were under the control of intellectual scavengers in search of occult wisdom, by which they could feed the already mesmerized and addicted European aristocracy. Collectors were willing to spend fortunes to acquire these objects of antiquity, especially if they were rumored to possess some mysterious supernatural powers. Many of these curious objects are still kept secretly in private collections throughout the world.

In 1798 Napoleon recruited 167 scientists and technicians to accompany the French fleet in the conquest of Egypt. And in 1801 Volta, the so-called discoverer of the pile or battery, was given a medal and a pension by Napoleon for his contribution. The British invasion of Egypt in 1801 seemed primarily aimed at stopping the French savants from exporting anymore artifacts. It was viewed as a matter of National Security in those days.

Volta's discovery of the principles by which the djed produced a stable electrical current, is considered to be the greatest invention of the 19th Century.

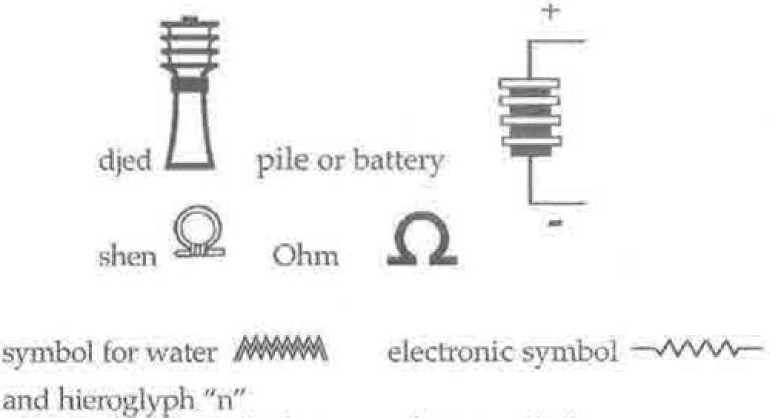

Note the similarity between these symbols.

The Encyclopedia of Ancient Egypt defines the djed as:

The Egyptian symbol of stability. The djed is a pillar with bands across the top and base. This hieroglyphic symbol is associated with the Osirian mortuary cult practice of mummification. Serving as a powerful symbol in Egyptian magic and reincarnation, where it was used to symbolize the metamorphosis which the body underwent at death. The transformation of the physical being into its eternal celestial form.

The djed is thought to represent the backbone of Osiris, who's sparks were as bright as the Sirius Star.

It is also associated with Jacob's Ladder.

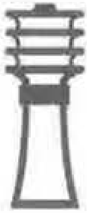

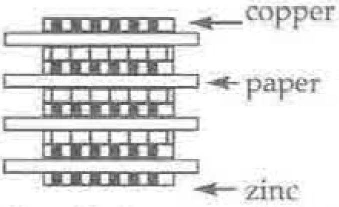

A djed or battery is made by placing two dissimilar metals on either side of a piece of papyrus soaked in salt water or acid. Greater power can be realized by increasing the number of units. For instance, a single unit of zinc and copper can produce 0.759 volts and the current has greater magnitude than that produced by electrostatic machines. Napoleon gave the Ecole Polytechnic University a battery with 600 units.

Passing voltage through all sorts of material, experimenters using these batteries were able to produce many chemical reactions and split molecules like water into hydrogen and oxygen.

Egyptologists have long suspected the use of direct current in Egypt by the many examples of electroplated jewelry and their knowledge of other chemical processes. The Egyptian word for electroplating of silver or gold is sem (sem means: a kind of, similar i.e. likeness of gold) which contains the ideographs conveying the entire process. The uas: an indicator of electric current and 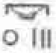 nub, a gold in solution. However rather than accept the scientific evidence, these artifacts are often attributed to accidental occurrence or some still unknown process.

A good example is their promotion of iron as a Hittite technology, despite evidence of prior usage in Kemit. Chiekh Anta Diop claims iron usage originated in the Ancient Empire (2700 B.C.) and that the steel blade of Tutankhamen's dagger, could only have been produced by very deliberate means. He theorized that such a purity of iron would require the type of sophisticated processes used in modern electrolytic refinement. The ideograph for pure gold 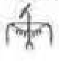 is proof that the Egyptians knew that the electrolytic process led to the purest metals.

Plating was in common use during ancient times as evidenced by the many words for it. Yet no such words in Coptic (a Greek-Egyptian language) have survived, except in a reference to Alchemy from a variation of the word NHB neb; to smelt or work in metal (gold).

Alchemy, the transmutation of base metals, i.e. lead, to gold, was the most sought after secret of the Medieval Period. The origin of the word chemistry, it is a reference to Egypt and the philosophy held by the Blacks concerning the transmutation of the body to an immortal spirit.

Had Egyptology been in the hands of Africans or people without a racist agenda, more knowledge would have been gathered concerning these batteries. For instance, the origin of the lead acid batteries in popular use today, may well be Ancient Egypt. Both lead and lead oxide, the elements of its construction were widely used by 3000 B.C. But what is most interesting is the word tet (battery or djed), it sounds like 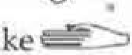 teht, Copt ⲦⲀϨⲦ the word for lead.

With the principle of the djed fully understood, the principles of the Ankh remained the greatest challenge to European scientist, however it would not yield its secrets for another 87 years.

Early attempts to unravel its secrets were made by Henry, now credited with the discovery of self- induction.

Henry reports that...

When a short wire is connected across a battery, no spark is perceived when the connection is formed or broken. But if a wire 30 or 40 feet long is used instead of the short wire, although no spark will be seen when the connection is made, yet when broken a vivid spark is produced. The effect appears somewhat increased by coiling the wire into a helix; and it also seems to depend in some measure on the length and thickness of wire...

Henry had rediscovered self-induction, the method of producing electricity from a changing magnetic field.

Hans Christian Oersted (1777-1851) observed in 1820 the magnetic action of an electric current, opening the study of electromagnetism. He observed the deflection of a magnetic needle during a thunderstorm.

In 1842, Henry observed that the effects of sparks in one location could be detected some distance away by their magnetic effects in another location. The analogy between the rippling waves, caused by a stone dropped in the center of a pool, and the spark transmitting rippling waves in an ether-like medium, were attributed to Henry. He also noted that the effects of a single spark could be detected thousands of feet away.

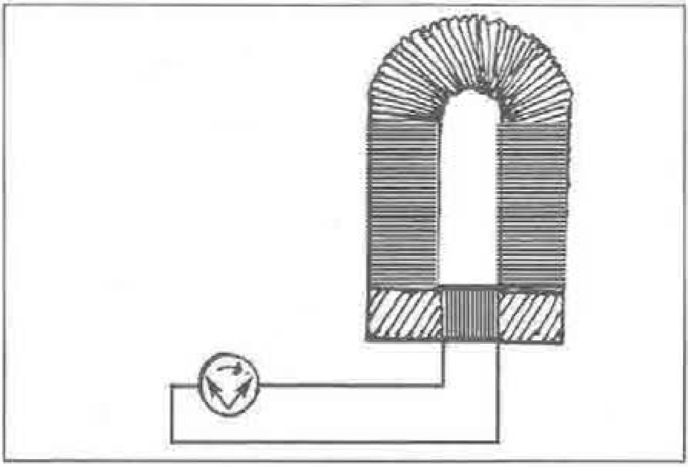

Hertz, experimenting with various sizes of loop detectors, was able to detect electromagnetic waves and determine quantitatively the effect known as resonance. Using the spark from an induction coil, electromagnetic waves could be detected in a loop some distance away. In other words, a spark in one location produces a spark between the gap of a loop detector in another location.

A hertz oscillator consists of a coil and a capacitor, made from two flat sheets of metal B and a spark gap G. One can easily see the similarity to an ankh oscillator, which produces a higher frequency because the coil is one turn or loop.

As mentioned before, the English defeated the French in Egypt, taking over the scientific explorations and

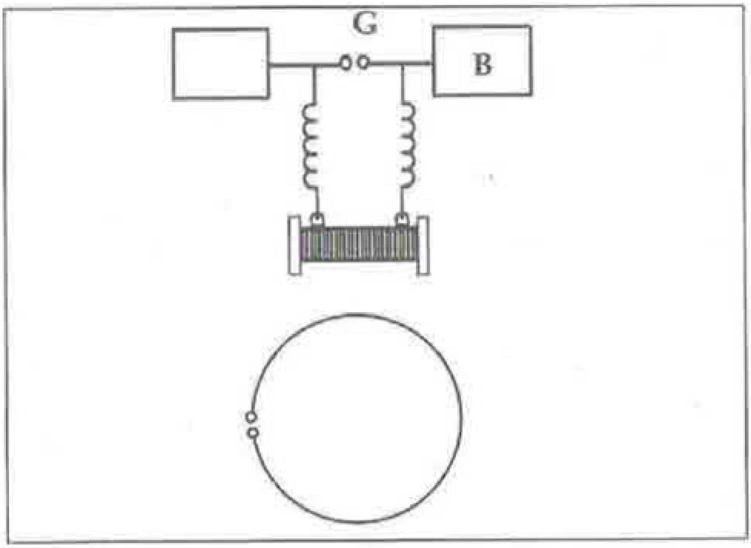

Apparatus used by Hertz to demonstrate the wave like nature of electromagnetism.

other antiquities excavated by the French. Among these were strange devices, numerous scientific and mathematical papyri and most notably, the "Rosetta Stone."

As a result of this scientific coup d'état, Newton, the famous mathematician provided the mathematical language to codify the theory of wave transmission and the action-at-a-distance properties of electromagnetic waves. Budge was able to translate the so-called Book of the Dead, and Sir Oliver Lodge discovered the properties of the ☥ Ankh.

Lodge was the first European to prove that lightning and sparks between the gap, were high-frequency discharges. Current would rather flow through the high resistance path across the gap, than the low dc-resistance wire loop. He was able to show that this occurs because of the high impedance presented by the wire loop at higher frequencies. Lodge was fascinated with the subject of psychic phenomena and communication with the dead. He spent much of his time trying to give scientific legitimacy to Spiritualism.

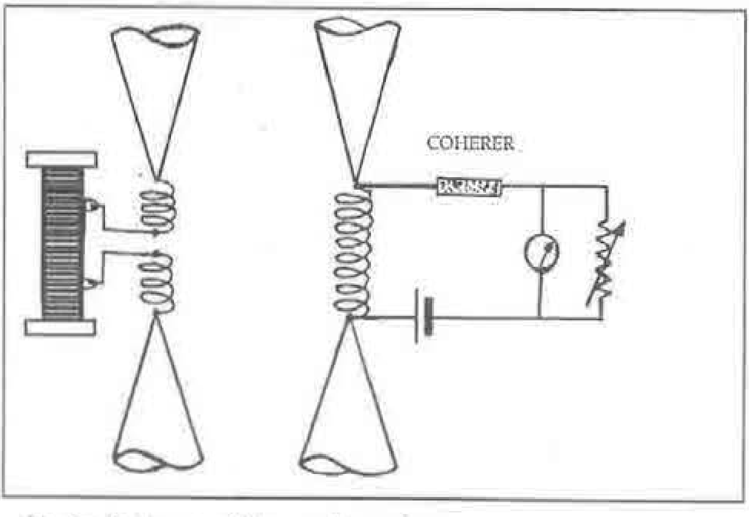

Lodge's transmitter and receiver.

You may have noticed a set of grooves or lines in the center of some ankhs, especially the one belonging to

Tutankhamen. Well, after the discoveries of Hertz, a Professor Righi invented a very sensitive spark gap made from a sheet of glass covered with tin foil. This foil was divided into several strips by very fine razor cuts. When used in a Hertz loop detector, sparks are easily observed jumping across each gap in the foil. I would think Professor Righi was 4000 years late.

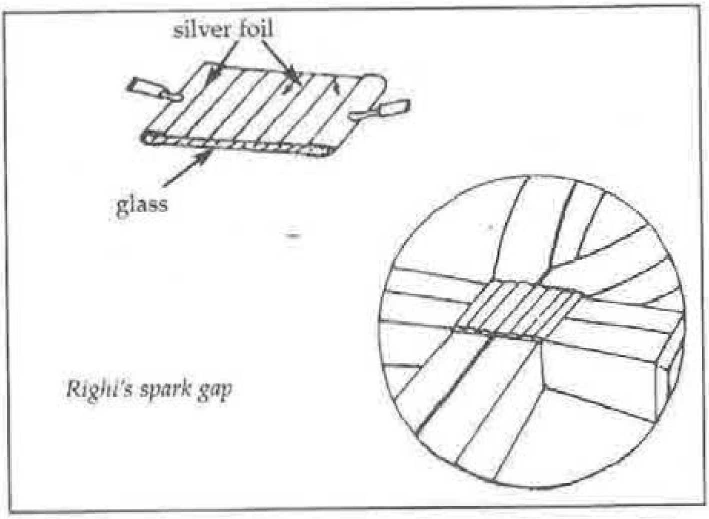

Righi's spark gap

The early telegraph and radio circuits used detectors of various construction and material. Detectors are required because the high frequency of the oscillator is inaudible. A very sensitive detector was galena (lead sulfide) crystals, some of which were found in Tutankhamen's tomb.

The hieroglyphic symbol for stone ▥ in the center of the ☥ ankh, suggests the usage of crystal or even semiconductor technology. For instance, continuous wave oscillation can be produced by attaching oscillating components (capacitor and inductor) and a battery across a galena-galena contact. This black stone, kam 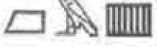 a representative of Egyptian Spirituality, is the most sensitive crystal, detecting electromagnetic waves well into the infrared frequencies known as heat.

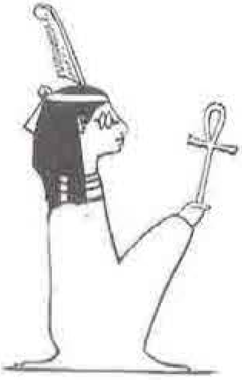

Maat with her eyes closed

If you subscribe to the belief that the ankh of a god had power associated with that god, then the ankh of Maat would be instrumental in the realm of law and righteousness. Maat is the goddess of truth, indispensable in the cause of justice. Her symbol is the feather of truth, against which the heart is weighed in the judgment. Therefore, she is representative of the solar archetype Libra.

The ankh she is holding, is the oracle of truth, operating on a principle known as galvanic skin response (GSR) named after the Italian experimenter Luigi Galvani, who in 1750 discovered the reflex action of a frogs leg to electrical stimulus, it is widely used as a part of a modern polygraph test.

The conductivity of the skin varies with one's emotional state and what makes (GSR) so effective is how totally involuntary these skin responses are. An electrode is placed in the hand or taped to the finger of the subject and the conductivity or resistivity of the skin is measured in response to questioning. The skin is the key to the soul, portraying subtle psychological changes, it does not lie.

Vesica Piscis

Changes in conductivity of the skin alters the capacitance of the ankh, creating frequency changes as a result. By holding the ankh in this way, the emotional and spiritual responses can be communicated as a frequency modulated transmission.

A high frequency oscillator, like the ankh, has the advantage of producing high voltages at low current, so they can be handled safely. Tesla was known to have attached a high frequency coil to himself and produced corona-like discharges from his body. Luminous effects produced by high voltage high frequency coils are used today in Kirlian photography. Objects appear to have a glowing aura around them, when subjected to high voltage.

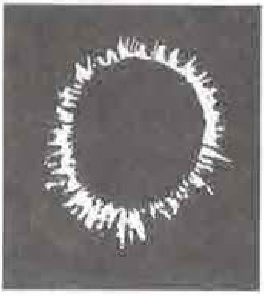

Kirlian photograph of a fingertip

The Europeans have skillfully rewritten history in such a way as to present themselves as the sole provenance of all science. Weaving a concoction so convincing, that one can only conclude that they alone have brought progress to the world. This part of their racial policy is important if they are to convince the world to tolerate their murderous onslaughts and exploitation as a necessary side effect in the cause of civilization.

One chord seems consistent in the propaganda of their discovery: a diversionary claim of ignorance and the naming of other Europeans on whose work they hold up as the basis or source material for their discovery.

Columbus used the diversion of seeking a route to

India, even though he was in possession of maps and accurate information gleaned from West African sailors who had been traveling across the Atlantic for 200 years. He claimed the Portuguese as his information source. It is the same with Volta and Galvani in the concealment of the piles Egyptian origin.

This stranglehold caused by European editorial censorship of Egyptology, is preventing the spiritual reconnection with our glorious past. Without the oppressive scrutiny of the Academic Establishment, you are free to develop your own conclusions about the ☥ ankh using your knowledge of electromagnetism as a guide.

It is imperative for us as Africans to develop a way of teaching electronics, that is both interesting and culturally relevant, if we are ever to compete successfully in the new millennium

FREE THE PEOPLE!!

The Ankh is essential to the liberation of African peoples throughout the world. Without the Ankh and it's associated technology we will remain dependent on the inferior technology permitted by the maintainers of the status quo.

As a Free Energy Generator the Ankh represents the symbol of our liberation. It produces electricity indefinitely without any fuel or other energy sources such as wind, solar or hydro power.

Ankhs of different sizes or capacities may be used to power boats, cars or airplanes giving them infinite range. Imagine, never running out of fuel or paying another utility bill; enjoying the freedom to live in remote regions of earth and sky as UFOs that are powered by ☥ Ankhs.

The Oscillator

AN OSCILLATOR IS A circuit which produces or responds to electromagnetic waves, usually within a range of frequencies. The circuit may consist of only a single wire, crystal or transistor semiconductor device. They respond most strongly to vibration at their resonant frequency.

Everything in nature is an oscillator and God is the spark projecting all frequencies to which everything responds. Atoms and molecules are perfect oscillators, which respond to light by absorbing and emitting at specific frequencies. Even complex organic molecules like melanin exhibit quantum states once thought to be limited to atoms and molecules. An oscillator's ability to absorb and emit energy, explains a host of phenomena which were once considered mysterious.

Oscillator: 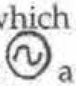 a sine wave within a circle, is the modern symbol that corresponds to while the uraeus 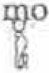 which is the ancient symbol, representing the serpents undulating or oscillating movements across the sand.

A circuit containing a capacitor, inductor and resistor are the components of an oscillator, which operates within a defined range of frequencies depending on their values.

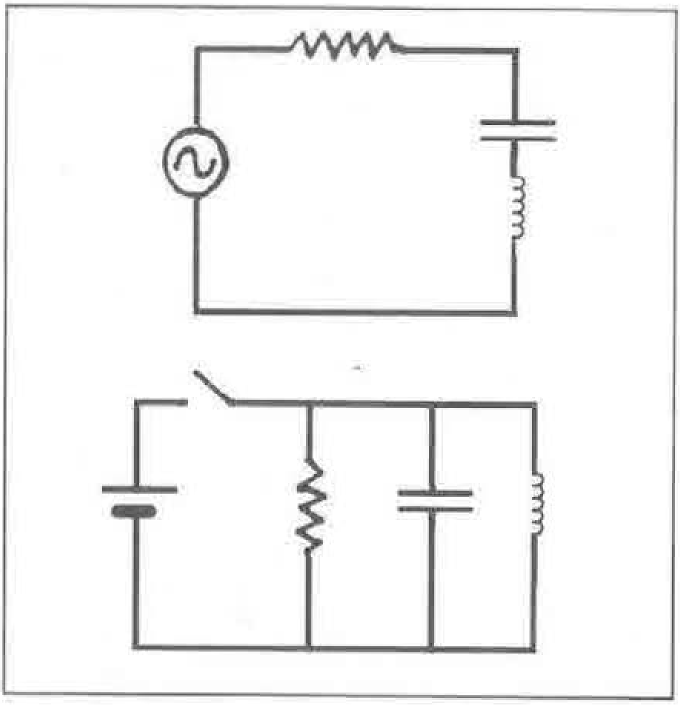

Because of the frequency-restricted nature of these oscillators they are called filters. They are also called tank circuits because of their ability to store energy.

When the switch is opened, oscillations are produced by the opposing nature of the reactive components. However, without a method of amplification in the LC circuit, the resistance eventually quells these oscillations. The result is a dampened waveform.

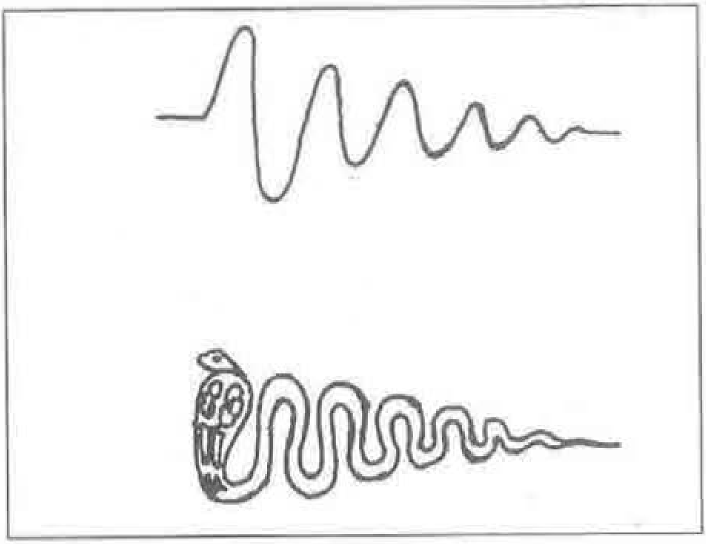

The components of a typical oscillator can be traced to the Amen Priesthood, which wielded power based on knowledge kept secret from the masses.

The symbols in many hieroglyphic reliefs are in fact elements of electronic circuit design, meant to convey knowledge of the hidden force of electromagnetism, to the future generations of this Order. These symbols would hold supreme religious significance only to the descendants of a race of Photonic Beings.

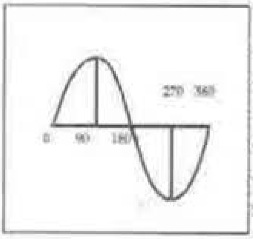

The mathematical description of a sine wave is attributed to Fourier, one of the 167 French savants who accompanied Napoleon in the conquest of Egypt. His formula describing a wave as a distinctive sum of sine waves, defined by their frequency, amplitude and phase, predates the European discovery of electromagnetic waves by almost a century.

Like Volta's discovery of the djed, Fourier's mathematical description of the serpent, has made a profound impact on all areas of analysis from sound to quantum physics.

The scepter (rod, conductor or connector) is an important electronic component, used in the conveyance of current through a circuit and in discharging a capacitor (shrine).

The gods are often shown holding the scepter to symbolize their power over the lethal force of electrical discharge. This rod, having a loop at its base which allowed a free swinging connection, is a switch. Once planted by the loop, it was thrown against the shrine to affect an electrical discharge flow to ground.

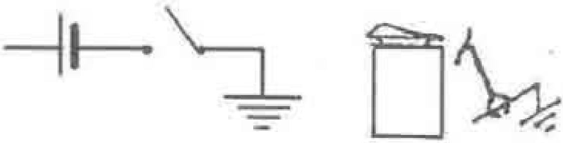

Hence the phrase "throw the switch."

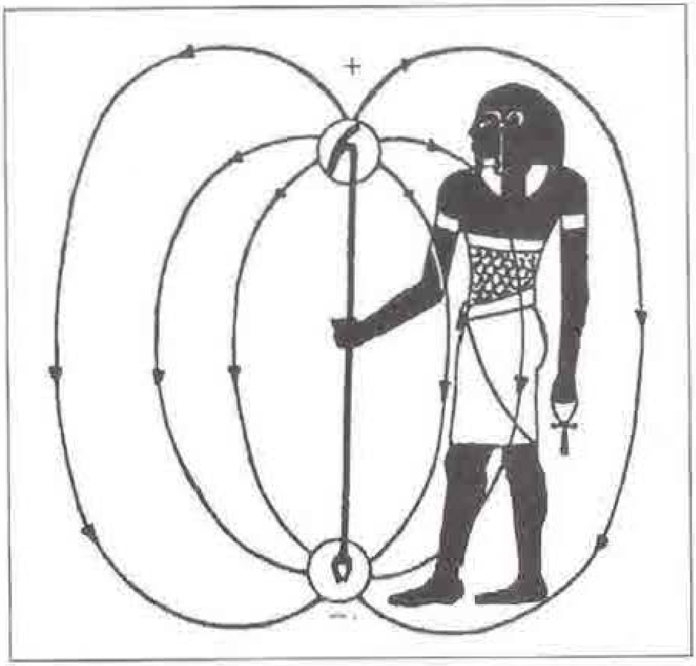

The electric field around a dipole antenna.

The mark of the beast Set (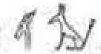 , a canine) and the mark of Cain are one and the same, the symbol of Christendom (+). The ion only the Setian could consider positive. While the ideograph of the earth 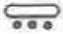 is the ground ⏚ and a source of the electron (-).

The Egyptians were fully aware of the fact that a spark produced high-frequency discharge and understood the inductive properties of rods and wires.

> So Moses and Aaron went to Pharaoh and did as the Lord commanded; Aaron cast down his rod before Pharaoh and his servants, and it became a serpent. Then Pharaoh summoned the wise men and sorcerers; and they also, the magicians of Egypt, did the same by their secret arts. For every man cast down his rod, and they became serpents.
>
> *Exodus 7: 10*

A conductor wrapped in coiled fashion, as a serpent is known to coil itself around a rod, is an electromagnet.

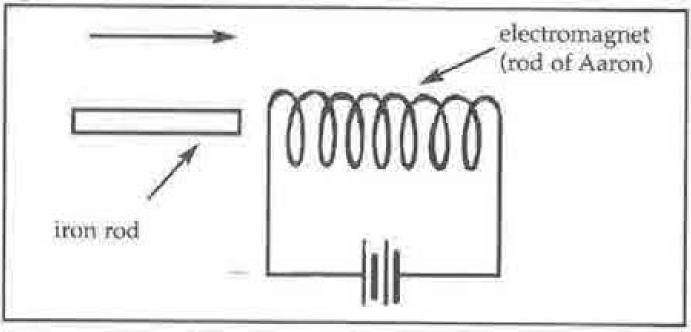

... But Aaron's rod swallowed their rods.

The Egyptian name for this coil circuit is 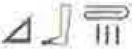 qeb, which is also the name of the multi-coiled mythological serpent 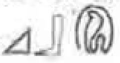 Qeb. The magnetic attraction of a coil was well known to the Egyptians, 4000 years before the European named Gauss, Henry or Weber as units of magnetic force.

Another device attributed to Moses is the Helicial antenna, made in the form of a copper serpent and suspended on a pole. It was supposedly built in the wilderness (Egypt), and as the golden calf, represented the goddess Hathor, the serpent was symbolic of the goddess Isis The serpent built by Moses was destroyed by Hezekiah during a period when Israel was under intense pressure from the Assyrians to abandon their spiritual, cultural and military alliances with Egypt.

> He removed the high places, and broke the pillars, and cut down the Ashe'rah. And he broke in pieces the bronze serpent that Moses had made, for until those days the people of Israel had burned incense to it; it was called Nehushtan.
>
> *2 Kings 18:4*

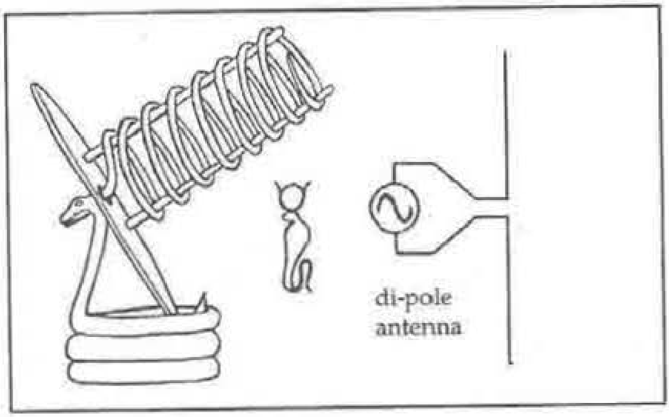

Antennas are resonant circuits which detect a specified range of frequencies. Two forms of these are the half-wave dipole and loop antenna. Used extensively by Hertz to prove the existence of electromagnetic waves, the loop antenna was used to detect the waves emitted from a dipole antenna.

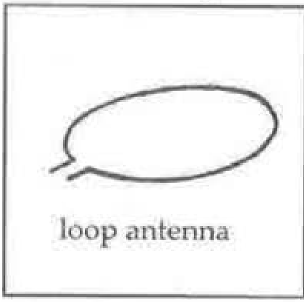

At the turn of the century the basic concepts of the Ankh were understood, as evidenced by the Duddell's musical Arc. The arc or spark-gap has negative resistance, where the resistance to electron flow is greatly reduced. If an oscillating circuit (inductor and capacitor) is placed across a spark-gap, oscillations are induced in that circuit.

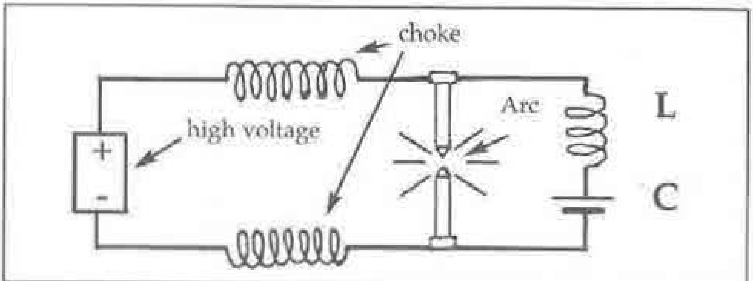

A metal loop in close proximity to the spark-gap could be excited into oscillation, due to the ultraviolet rays ionizing the air between the gap of the loop.

Perhaps the closest circuit that describes the ☥ ankh's operation is the Liebowitz Mercury Generator, invented in 1914. Using a mercury vapor tube instead of a spark-gap, continuous oscillation was produced.

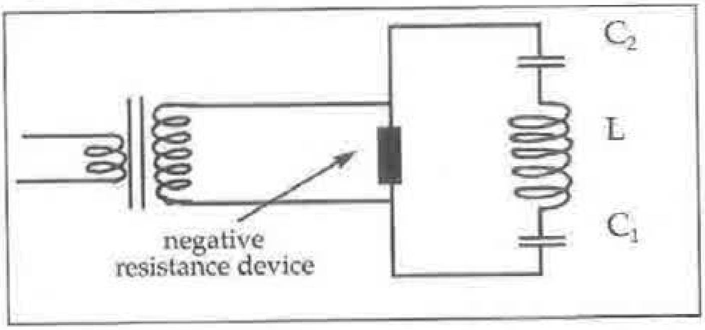

With the advent of other negative resistance devices, like the vacuum tube and the transistor, spark-gap oscillators were no longer used. They were considered noisy for the purpose of communication.

The question still remains, whether a method of amplification existed in Ancient Kemit. An essential component in the production of continuous wave oscillation, amplification is necessary in overcoming the dampening effects of circuit resistance. It is my belief that both evacuated glass tubes (vacuum tubes) and semiconductor type active components were known

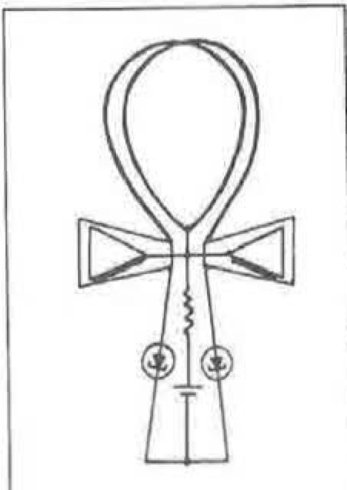
The modern equivalent of the ankh, using Tunnel Diodes.

and used by the Egyptians.

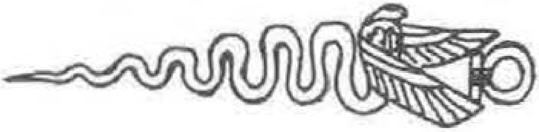

symbolic representation of amplification.

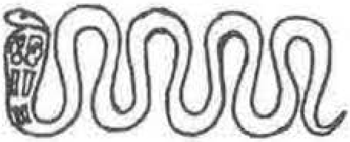

a representation of continuous waves

Continuous waves have been produced in circuits using an ark light bulb as an active device. And it should be remembered, that the first vacuum tubes were invented through experimentation with ordinary light bulbs.

You maybe asking yourself why the Europeans did not use the ☥ Ankh for its intended purpose, or how could an object so popular in the Nile Valleys of Africa be virtually ignored in Europe.

This is because the Europeans lacked melanin, which is an organic semiconductor, that acts as a detector (the sixth sense). Without an electric skin, an ankh is practically useless. However the principles on which the ankh operated would give birth to the many electronic gadgets in use today.

Melanin

MELANIN IS A PIGMENT IN THE skin of people of color, which is produced by melanocyte cells, and deposited in the epidermal tissue. Melanocytes are neuron-like cells which produce melanin and numerous proteins in response to electromagnetic radiation.

The production of melanin starts with the conversion of tyrosine by the enzyme tyrosinase to 5, 6-iodole quinone.

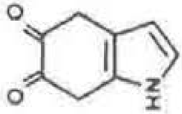

chemical stucture of
5, 6-iodole quinone

Tyrosinase is a copper-containing enzyme which catalyzes the conversion of tyrosine (an amino acid) and stabilizes the conformation of the melanin structure.

The metal ion acts as a backbone for the polymer structure of melanin, resulting in a metal-organic complex. The amino acid forms peptide-linked formations with the metal ions. The ligands are attached at the nitrogen atoms.

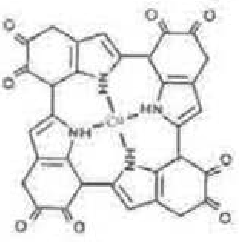

The proposed structure resembles a swastika, with the interactions occurring between the central Copper (Cu) ions. This complex metal compound is the only substance in the body that qualify as a organic semiconductor.

Black and brown melanin granules are oval in shape, forming a small dipole antenna. The field due to one dipole can induce a dipole in another melanin granule nearby.

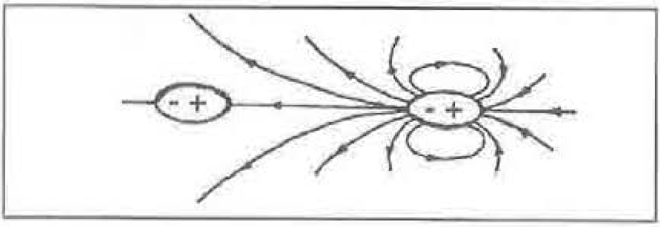

Melanin granules act as tiny primitive eyes, forming a large neural network structure, whose function is to absorb and decode electromagnetic waves. Neural-network computers are learning machines which are made with a number of receptors that can adjust their weights (quantitative properties) to produce a specific output.

The bodies of Africans contain massive amounts of melanocytes that encode all life experiences in their

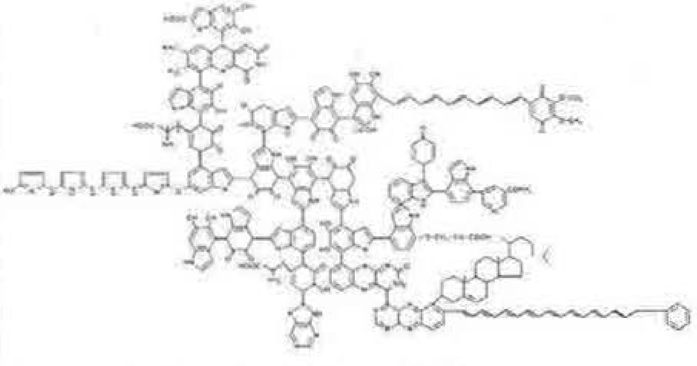

chemical stucture of melanin molecule

melanin production, with the aim of creating an actual- reality state after death. During life, visions appear frequently and ESP is common.

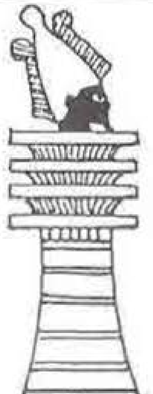

A representation of a higher form of life, in the superconducting body of Osiris (the mummy)

The reason for mummification was to preserve the skin which contained a living neural-network of melanin. The conductivity of melanin increases with age, so Tutankhamen's mummy is more alive in the spiritual plane than our own. Consider the difference in flammability between living green trees and coal.

As a semiconductor, melanin has an energy gap. The absorption of energy is required before electrons can jump into the conduction band and make melanin conductive. An increase in conductivity increases the sensitivity of melanin to the electromagnetic world of etheric beings, astral projections and spiritual entities.

At low frequencies, the conductivity of melanin is small, but at ultra high frequencies (UHF), melanin is a superconductor. Maximum current flows only on the skin, due to the skin-effect, at melanin's UHF resonant frequency.

King Tutankamen on his throne

Melanin is the most important substance in the human body. It is an oxidized form of RNA, which enables the body to coordinate the production of Proteins needed in cellular repair. Wherever there is cell damage melanin is seen surrounding the site, functioning as a neurotransmitter in coordination with melanocyte protein production for the repair of damaged DNA.

Knowledge of the medicinal value of melanin is suppressed by the Medical Establishment in order to deny its supremacy.

The most damaging attack on African health is the promotion of albino domesticated animals for consumption. In fact only albino animals are considered domesticated. The yellow pigment in chicken and the lack of tyrosinase in albino animals, are responsible for premature aging.

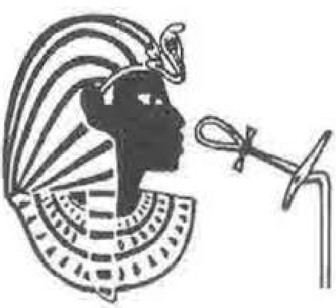

Illustrates the curative properties of negative ions, the electromagnetic energy of mental concentration can be directed through the breath of life, by a competent Psychic healer.

The laying of hands on which is known to emit energy from the aura.

The Spiritual or Electromagnetic Worlds

THE MYSTERIOUS NATURE of God is central to Christian doctrine. Most believers and their preachers maintain the mystery through ignorance and confusion. The various interpretations of the Bible leaves one wanting for a simpler understanding and a more practical version of God. Despite the many scientific discoveries into the nature of the world, a separation between science and religion has prevented the understanding of anything spiritual, from an objective point of view. This separation is the result of a pathological condition among non- pigmented people, which renders them incapable of becoming in tune or sensing these higher forces directly. Imprisoned by this condition, our white ruler's viewpoint on spiritual affairs cannot be trusted despite its domination over more traditional nature-based religions of the world.

It is obvious that the early Christians' religious concepts came to Europe from Ethiopia, as evidenced by the language used to describe the spiritual realm.

ethereal	1. a: of or related to the regions beyond the earth b: celestial, heavenly c: unworldly, spiritual. 2 a: lacking material substance: immaterial, intangible.
ether	1 a: the rarefied element formally believed to fill the upper regions of space b: the upper regions of space: heavenly 2 a: a medium that in the undulatory theory of light permeates all space and transmits transverse waves b: the medium that transmits radio waves
ethic	the discipline dealing with what is good and bad and with moral duty and obligation. 2 a: a set of moral principles or values b; a theory or system of moral values (the present-day materialistic)
Ethiopic	1 a: a Semitic language formally spoken in Ethiopia and still used as the liturgical language of the Christian church in Ethiopia 2 a: the Ethiopic group of Semitic languages.
utopia	1: an imaginary and indefinitely remote place 2 a: a place of ideal Perfection esp. in · laws, government and social conditions.

By permission 1991 by From Webster's Ninth New Collegiate Dictionary by y Merriam Webster Inc., Publisher of the Merriam Webster dictionaries

The rise of materialism brought new language, electromagnetic replaced the word "ether," in describing the heavenly medium, thus erasing the link between spiritual and natural phenomena. The pathological condition of race caused them to enforce this separation academically and religiously, by claiming that the classical ether was unnecessary and the vacuum was somehow endowed with odd and unusual characteristics. With their strong emphasis on materialism, all electromagnetic theory was co-opted for material implementation, therefore losing all spiritual application.

Anyone displaying an interest in electromagnetic spiritualism was considered to be a pseudo-scientist and ridiculed. There were many devices which claimed to photograph, or record the voices of the dead. These ghost detectors were routinely rejected by the materialist' scientific community.

The seventies brought renewed interest in Psychic Phenomena and ESP around the world. Knowing that belief is the prerequisite to the use of ones psychic ability, the scientific community conspired to take over this research with the aim of disproving and suppressing any attempt at legitimizing this field. So most people continue to experience vivid dreams, visions, premonitions and a host of strange occurrences, which scientists refuse to recognize because of their threat to the materialistic power structure.

African culture and history are rich with examples of the unity of spirit and matter; heaven has always been our birthright. Today those faculties that put us in tune with the spiritual world are weak from lack of exercise. We have allowed the TV, radio and the telephone to replace our telepathic abilities, few now believe or would accept the notion of astral-travel.

Service to the materialist's society requires the denunciation of the spiritual. When bills are on our minds, we dream less as we focus on the depressing condition of this earthly prison, with our tormentor on our backs. The ☥ Ankh is the key to the door of eternal life, strengthening our faith in a world unseen, where our souls can be free.

Osiris The First Resurrected Christ

Statue of Osiris

THE CORRELATION BETWEEN the myth of Osiris and the story of the Christ have been illustrated by many prominent historians, who have shown that the evolution of Christianity is rooted in the religion of Isis, Horus and Osiris. This religion survived in its last outpost in Nubia until the Temple at Philae was destroyed by Justinian in 6th century A.D.

Although there are many versions, the essential theme of the Osirian legend, states that Osiris was once king of Egypt. He came to an untimely death at the hand of his brother Set, who had the parts of the dismembered body of Osiris hidden in various locations all over Egypt. Isis, wife and sister of Osiris found all the parts and was

Horus, Osiris and Isis

able to reconstruct the body with the exception of the penis, which was thrown in the Nile and eaten by fish. Her lamentations were heard by Ra, who sent his fourth son Anubis to wrap the body with bandages and perform all rites due one of his stature. Isis caused the breath of life to enter into the body by the rapid beating of her wings, whereupon Osiris was resurrected and became the king of the other world. While hovering over the body she became pregnant and later conceived a son Horus, who would avenge his fathers death.

Hence-forth all the dead of Egypt were considered Osiris, wrapped in the bandages of the familiar mummy, with the hope of resurrection in the spirit world of the ankhet 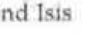 "the land of life."

If the Egyptians seemed to be preoccupied with the dead, it is from a belief that life on Earth was merely a preparation for an eternal life. The proof of this could be demonstrated by the ankh, which become the symbol of life, because of its ability to detect the energies of that realm.

Further proof that life on Earth was preparation for another life is the manner in which we dream. Dreams are our fledgling entrance into the spirit world while we are still alive. Our earliest attempt at spiritual flight are short. While dreaming the penis is usually erect (the phallus of Osiris is unnecessary) and the body is kept rigid by powerful chemical secretions from the brain; which prevent it from acting out the experience. During those brief seconds we are as if dead, yet alive in the other world.

> The days of our years are three score years an,reason of strength they be four score years, yet I labor and sorrow; for it is soon cut off, and we fly away.
>
> *Psalm 90: 10*

According to Gerald Massey's book *Ancient Egypt Light of the World* , Osiris is the Corpus Christi. The word Christ comes from the Greek word "Kristos" which means "anointed." In Egypt "krst" means; to embalm, to knot, to make the mummy.

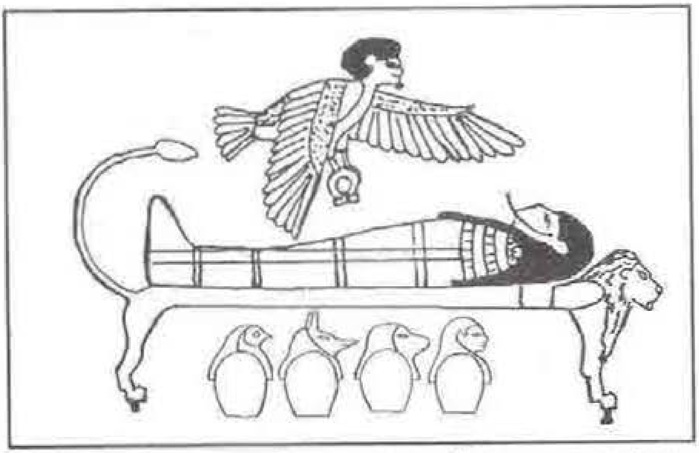

Central in Christian theology is the resurrection of Christ from the dead. This concept has an undying appeal for the masses of humanity from the dawn of the human experience.

> The thief cometh not, but for to steal, and to kill, and to destroy: I am come that they might have life, and that they might have it more abundantly.
>
> *John 10: 10*

Acceptance of the Egyptian definition of "Krst" could put the confusion within Christianity to rest. For Christianity did not start with the Jews, but was of Egyptian origin, revolving around mummification and the spiritual resurrection of the dead. The followers of Jesus, after his supposed death and resurrection, were only later called Christians because their beliefs were similar in nature to what was practiced in Egypt for thousands of years.

Unfortunately the misinterpretation of this mystical gospel would ultimately inspire the outrageous behavior of ingesting the flesh of mummies (Krst, Christ) in hope of attaining immortality. This felonious religion let loose among a canine race, already predisposed to cannibalism, found easy acceptance in a god who offered his flesh and blood as a means of salvation.

> And he said to me, "It is done! I am the Alpha and the Omega, the beginning and the end. To them that thirst I will give water without price from the fountain of the water of life."
>
> *Revelation 21:6*

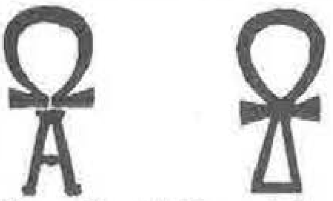

These Greek letters, alpha and Omega, have come to represent Christ.

In the original myth, Horus lost an eye in his victorious battle against Set. After his wound was healed by Thoth, he gave the eye to Osiris to eat, which vivified and strengthen him. The eyes of Horus are the sun and moon. Our sun is living water, made mostly of hydrogen that has not suffered the death of oxidation.

There is much evidence to conclude that life energy may be harnessed to prolong the life of the living. Our ancestors took full advantage of this wholistic science, allowing some to live hundreds of years. This technology predates the Early Dynastic Period in a time when the Earth, according to the Egyptian, was inhabited by demigods and kings. Obviously the Ancients had different priorities. The accumulation of wealth was superseded by a desire for wisdom and life.

The mythological story of the creation of man, recorded in Genesis, was taken from the original Egyptian story of the creator god Khnum who formed man and his Ka on a potter's wheel.

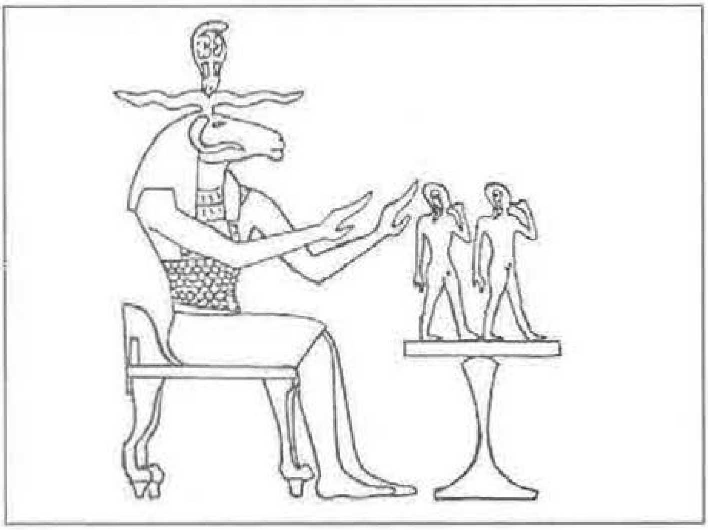

The hieroglyphic symbol which represents Khnum consists of a Tet between the raised arms of the Ka symbol and the word 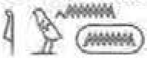 aun, which means light-tower, 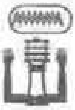 another name for the God Amen.

> "Allah" is the light of the heavens and the earth. A likeness of his light is as a pillar on which is a lamp. The lamp is a glass, the glass is as it were a bright shining star-lit from a blessed olive tree, neither eastern nor western, the oil whereof giveth light, though fire touch it not - light upon light. Allah guides to his light whom he pleases."
>
> *Surah 24:35*

The Tet is the orthodox illustration of a tree and according to Gerald Massey, it represents the eternal life after death.

> The Lord God formed man from the dust of the ground and breathed into his nostrils the breath of life, and man became a living being. The Lord God planted a garden in Eden away to the east and in it he put the man he made. The Lord God made trees to grow up from the ground, every kind of tree pleasing to the eye and good for food; and in the middle of the garden he set the tree of life and the tree of the Knowledge of good and evil.
>
> *Genesis 2:7-9*

Notice that the tree of life was not planted but set in the middle of the garden. After Adam and Eve ate of the fruit of knowledge of good and evil they became knowledgeable.

> The man has become like one of us, knowing good and evil; what if he now stretch forth his hand and takes the fruit of the tree of life also and eats it and live forever? So the Lord God banished him from the garden to till the ground from which he has been taken.
>
> *Genesis 3:22*

The tree of life was no ordinary tree.

> She replied. "We may eat the fruit of any tree in the garden. But the tree in the middle of the garden, we may not eat or even touch it lest we die."
>
> *Genesis 3:2*

The Tet is a high voltage power source, touching it may have been lethal to those without knowledge. Adam, who now possessed knowledge knew that the ankh portion of the tree was a high-frequency source that could be handled safely.

Tet 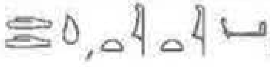 fruit tree., Danger do not touch.

Now we have come full circle and in our midst is the tree of good and evil, the modern computer and the Tree of Life, which is the Ankh Science of our Spiritual Nature. The moral question faces our generation, whether to trust Satan's path or the path which leads to Eternal Life as gods.

For it is not simply a question of a machine, but the programming of falsified information aimed at seducing you from your Divine Nature.

Who said you were ignorant, for you were created clad in Divine Wisdom and light.

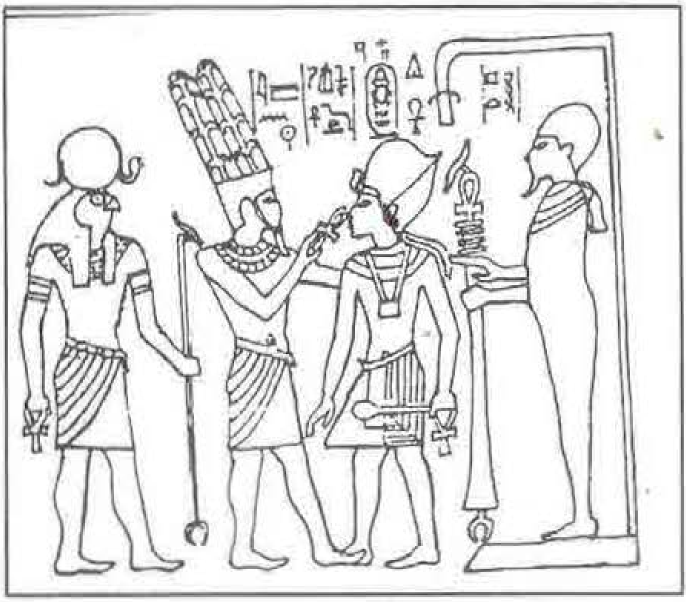

The depiction above shows Ra, Sun-god and his etheric double, Amen-Ra, the hidden force behind the sun. Our sun has a surface temperature of 6000 degrees, yet its corona (spirit or aura) exceeds 2 million degrees. The difference in temperature is caused by laser-like effects produced by the strong electric and magnetic fields within ionized gases over the highly reflective surface of the sun.

King Tut is also depicted with his etheric double and like Ra his name Tutankhamen expresses the hidden source of life. All living systems have electromagnetic phantoms. The phantoms of atoms and molecules are called photons, they are the essence of nuclear life, governing the energy states of the material world.

The ankh is held to the nose area because of its sensitivity to electromagnetic waves. The area between the fifth and sixth chakra is centered around the nasal cavities, where a massive quantity of neural receptors, responsible for the sense of smell, are exposed to the environment. The sensations are often confused with sound because of their proximity to the ears, sounding like the noise you hear from a TV set tuned to an unused channel or like the sound of the ocean.

Sensitivity can be increased and with practice one can detect the presence of high frequency waves close by. Some highly spiritual people, are so sensitive to these waves, that their neurons can demodulate radio broadcasts, causing them to hear voices. Many end up in psychiatric wards because they hear voices other can't.

The gods, who are a spiritual race of photonic beings, made man in their own image and likeness, to serve them on Earth. However man became aware of his own divinity and sought to be like the gods while in the flesh.

The story in Genesis also relates the longevity of Adam and his sons, whose average life span was 900 years.

> Then the Lord said, "My spirit shall not abide in man for ever, for he is flesh, but his days shall be a hundred and twenty years"
>
> *Genesis 6:3*

So it is the spirit (photonic or electromagnetic energy) which gives man the ability to live an extended life span. A tree deprived of a important nutrient like sunlight or water would wither and die in a shorter time. Without this vital ingredient we grow old and die prematurely.

The first signs of aging are apparent in the skin and is related to some melanin dysfunction. There is an increase in yellow pigment and among Caucasians there is premature wrinkling and melanoma. If the resonant frequency of melanin was known, the appropriate energy could be absorbed to vitalize the skin and delay aging.

Electromagnetic radiation may cause cancer among Whites because non-pigmented skin, has no defense against a high frequency wave which penetrates deeply in human tissue.

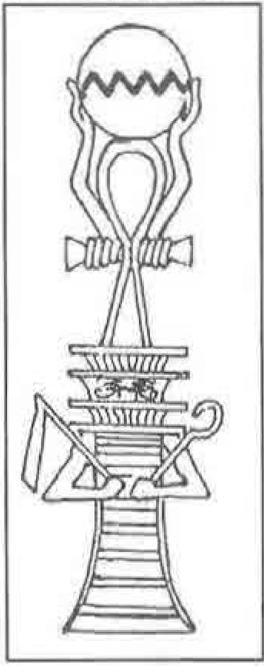

A drawing of the World's First electronic clock, built by Nile Valley Africans to represent their ancestral spirit Tet Ankh Ka Ra, The Ankhi (Khnum, chronos). The glowing globe was slowly rotated by the precise electromechanical movement of the arms.

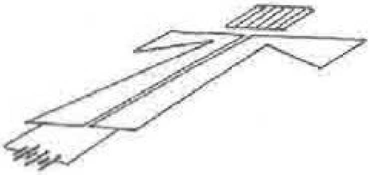

The Tree of Life provided sufficient power for the creation of electromagnetic radiation (angelic or spiritual beings) and the rotation of the world. The Egyptians not only knew the world was round, they knew the force which turned it was a perpendicular electric and magnetic field. They used elements of electronic circuitry to represent the spiritual nature of life. The circuit on the right of the tree is a relaxation oscillator, charging and discharging through the spark gap, to provide excitation for the loop circuit in the figure below. A minimum of 300 volts is required to produce a spark across the gap.

The relaxation oscillator is an inefficient method of excitation, because the discharged cannot be synchronized. However, with the use of kam (galena), a more effective continuous wave excitation oscillator can be made that requires significantly lower voltage.

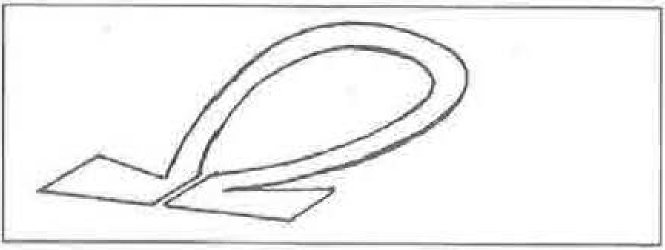

The loop and plates above are also a tank circuit which is shunted across the gap and capacitively coupled to the input resonant circuit. This transmitter radiates UHF power to ankhs of comparable size within the vicinity.

The spark was considered to be an indication of Gods presence because within a spark all frequencies are generated, therefore, all circuits are compelled to respond. The large capacitance of the shrine or ark could provide powerful sparks of lightning sufficient to excite an ankh into oscillation, especially when this discharge was through the gap of a transmitting ankh circuit.

A circuit consisting of a loop and plates could also provide lighting at night or in the darkness of the tomb. It has always been a mystery as to what method of lighting was used to allow those intricate paintings to be produced on the walls of some of those tombs. No soot was found on the ceiling suggesting that some other method besides the torch was used. Again we are asked to believe it was done with mirrors reflecting sunlight, because the truth of electric lighting in tombs would suggest widespread use similar to our modern use, but superior, because power would be wireless transmitted..

The pectoral of King Senusret II embodies the spiritual and scientific principles governing the universe. The ankhs illustrated have coils on either side of the loop, and Heru (falcon) is standing on the senk Ω , a divine light.

When this type of ankh was constructed and attached to a battery, it operated like a motor and produced high voltage spikes at audible frequencies.

The coils on both sides of the loop, form an electromagnetic switch, which switches on and off under the influence of magnetic attraction and the flexibility of the loop. The high voltage produced is a result of the sudden interruption of current flow through the coil.

This by itself did not seem strange, for on closer examination of the output, it was discovered that the output power exceeded the input power. With the use of rectifiers and rechargeable batteries, the excess power kept the ankh in a state of perpetual motion. The latent energy within the magnetic field was harnessed to produce electricity.

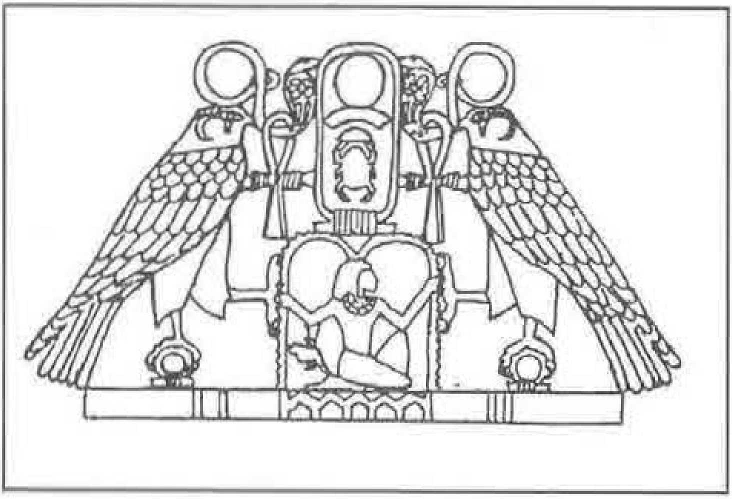

The kneeling figure in the center of the pectoral is Heh the god of Eternity, holding the notched chain of infinity.

This type of ankh was undoubtedly the power transmitter for the loop circuit of the senk.

snk , to see the light

snk ⊙ light rays

snk , night or darkness (note the senkh is upside down over the symbol for sky)

sen Ω or Ω kh 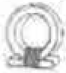 senk or senkh.

sen means, similar, dual, two or alike.

Similar to today's fluorescent lamp, the senk was used extensively before the numerous invasions put this light of Egypt out forever.

Since there are no known surviving examples of the senk, its basic operation can only be extrapolated from fluorescent light theory and examples depicted in Egyptian Art. The primary elements of the senk's construction are:

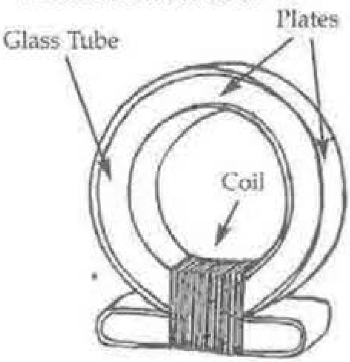

1. partially evacuated glass tube
2. coil or loop (inductance)
3. plates (capacitor)
4. DC. voltage

A partially evacuated glass tube can be made to glow when placed within a rapidly changing electrostatic field. The air in the tube is ionized by the presence of high-frequency oscillation in the coil and plates, at 10 to 50 MHz. The ionized gas in the tube becomes conducting and establishes direct current flow from the battery through the tube, further increasing the intensity and duration of the glow.

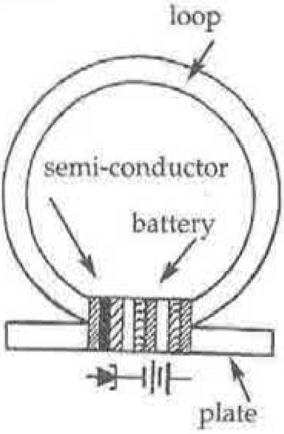

The ionized gas tube of the senk can be adapted for use as a diode (rectifier) or as an active device by exploiting it's negative resistance for amplification. By adding an extra set of plates to isolate the DC voltage, the senk could easily be modified to produce continuous wave oscillation. However since the Egyptians had knowledge of galena, it is likely that the vacuum tube was abandoned in favor of more efficient semiconductor technology. If all the components of modern radio transmission and reception were known and used in Kemit, 4000 years before their re-emergence in this century, then the super technology of the UFO's need not have come from other planets, but could have developed right here on Earth.

These are the questions which will arise to disturb the status quo if knowledge of ☥ Ankh science is made public. It would necessitate a complete re-examination of Kemitic religion to explain the connection between the African and these electromagnetic circuits, which were of such high spiritual significance. This would provide a direction for research into what we truly are as a people.

The atmosphere is a rarefied ocean inhabited by numerous spiritual entities living in realms which vary in activity levels between night and day or during cloudy weather conditions which causes the spirits to take on a higher density.

The electronic circuitry of the Egyptians was used to facilitate greater contact with the spirit world, known as the Kingdom of Heaven. The lords of this Kingdom are called the "grays" and many consider them negative aliens because of the fear they arouse and their enigmatic behavior.

"The fear of the Lord is the beginning of wisdom."

Ankh Circuits

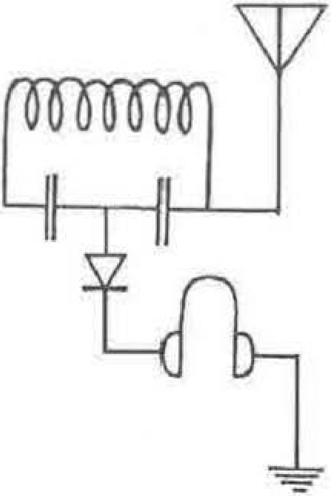

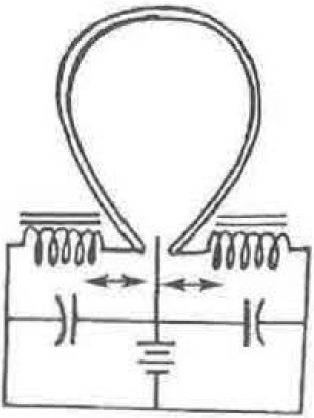

Ankh Circuits continued

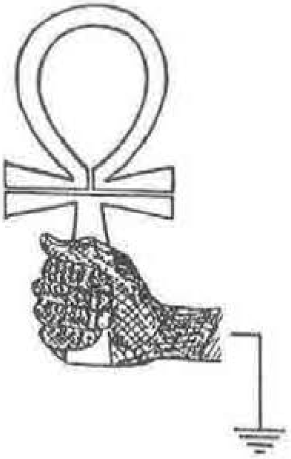

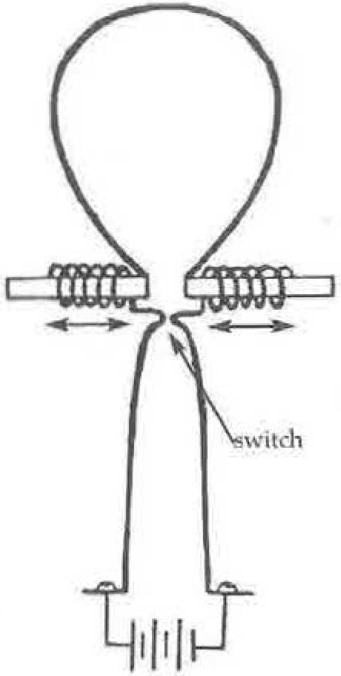

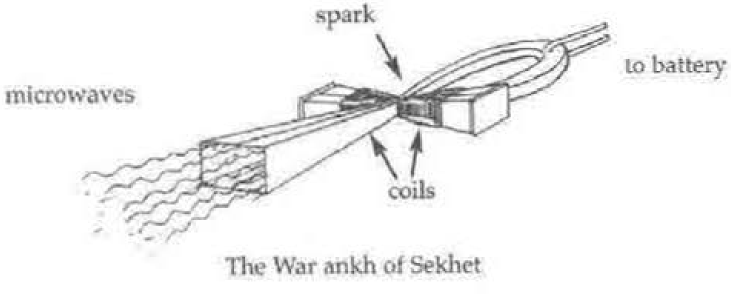
spark
to battery
microwaves
coils

The War ankh of Sekhet

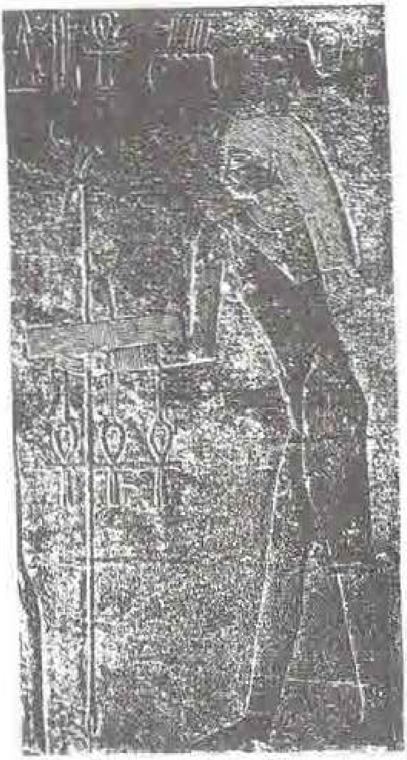

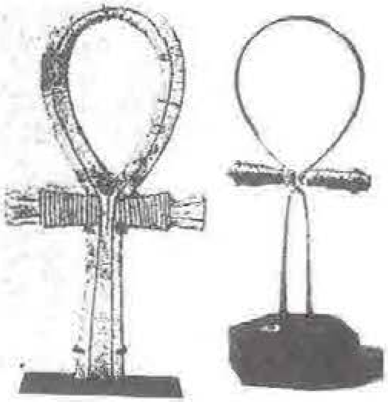

The Heavens

Although the Egyptians considered heaven to be the abode of the eternally blessed, it contained all the elements of a blissful earthly life. The dead were usually buried with their earthly possessions and provisions to sustain a life similar to their earthly one.

Everyone seems to agree that this realm is located in the sky above the clouds somewhere, but no one wants to specify its exact location. Here also, the separation of matter and spirit create

the usual confusion characteristic of modern religious theology and the materialist scientific aristocracy. If both spiritual and scientific evidence concerning heaven are examined objectively by someone with a unified perception, they would conclude that heaven was in the ionosphere. Our ancestors, who were not afflicted with this modem schism, knew this also.

> In the beginning God created the Heavens and the Earth.
>
> *Genesis 1: 1*

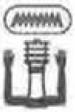

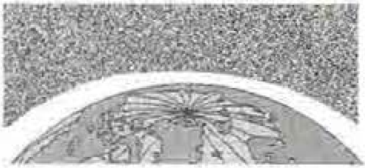

The aureole of the Earth created by the breath of Ra.

The Egyptian symbol of this Creator-God Khnum is the most profoundly esoteric, containing all the elements which could explain the nature of divine existence.

This is the god that created night and day by the rotation of the earth and made man on this revolving potter's wheel.

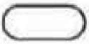

-island vacuum glass tube ... au

-breakdown discharge, "N" water, aurora.

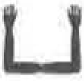

-Ka, spirit, power, image-double.

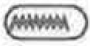

-Horizon, the God Amen, aun; which means lamp or light post

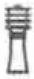

-Tet, djed, battery or voltage Osiris, Jacob's ladder.

As we have discussed before, the evacuated tube provides an easy path in the flow of electrons, creating the ionization and negative resistance used for continuous waves, and glowing when these electrons and ions recombine to form neutral atoms.

All these activities occur in the near-vacuum of the ionosphere, which is a region of high electron density, about 300 km above the Earth's surface. This layer is maintained by ultra violet radiation and high energy particle emissions from the sun. The high voltage (potential difference or electric field) between the earth and the ionosphere, represented by the Tet, is responsible for the rotation of the planet. This is due to the Earth's magnetic field being perpendicular to that electric field and in exactly the same way as the force which rotates the shaft of an electric motor created. The four bands at the top of the Tet represents the fact that the ionosphere is divided into four layers, the two uppermost levels F1 and F2 merge into one layer at the first hour of the Tuat (night).

Because of its high electron content, this layer reflects electromagnetic waves, allowing extended range to radio transmission.

> And he who talked to me had a measuring rod of gold to measure the city and its gates and walls. The city lies foursquare, its length the same as its breadth; and height are equal. He also measured its wall, a

> hundred and forty-four cubits by a man's measure, that is, an angel's
>
> *Revelation 21:15*

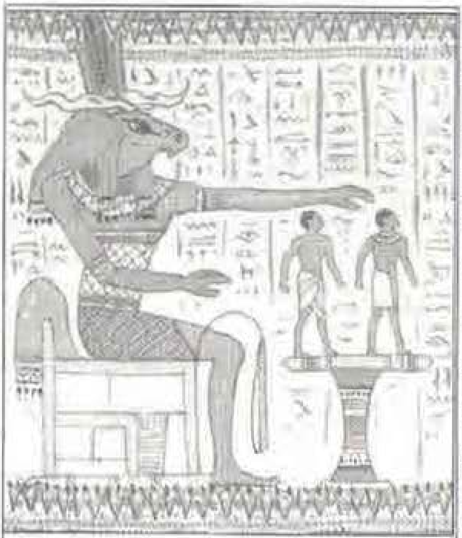

Khnum at his potter's wheel fashions a person out of clay. One figure represents the physical body while the other represents the spirit. Both came together to create one new being.

The description of the measurement are those of a sphere whose height is defined by the cut-off frequency of 1,440 Mhz. Because the measure of a man, that is of the angel is the Ka 𓂓, ten million.

> And he dreamed that there was a ladder set up on the earth, and the top of it reached to heaven; and behold, the angels of God were ascending and descending on it! And behold, the Lord stood above it and said, "I am the Lord, God of Abraham your father and the God of Isaac; the land on which you lie I will give to you and to your descendants; and your descendants shall be like the dust of earth ...
>
> *Genesis 28:12-14*

Heaven	Earth
Genesis Chap. 1-1 And the spirit of God was moving over the face of the waters. The word "ankh" also means mirror, "that which sees the face," in the language of ancient Egypt. **Genesis Chap. 1-27** So God created man in his own image, in the image of God he created him; male and female he created them.	**Chap. 2-7** then the Lord God formed man of dust from the ground, and breathed into his nostrils the breath of life; and man became a living being.

Phonetic Value	Hieroglyphic character	Meaning of the ideograph
ankh	𓋹	life
n	𓈖	water
kh	𓐍	dust (symbolized by a sieve)

An examination of the word 𓋹𓈖𓐍 ankh, reveals the nature of our earthly and heavenly incarnation in the eternal cycle of life. The word contains three ideographs, each representing an aspect of the Trinity, Father, Son and Holy Spirit. If you can accept the concepts ... Our

Father who art in Heaven,... I and the Father are one and... Son of man, you will be well on your way to an understanding of eternal life.

From the Egyptian point of view, ;(pet) heaven was made from a material called 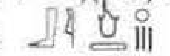 baa. This baa (baa; metallic substance i.e. black meteoric iron) was known to have metallic properties such as conductivity and reflectivity.

The ionosphere is a mirror image of the electromagnetic emissions from the Earth, no matter how faint. The spiritual characteristics of every object on Earth is duplicated in Heaven, because their emissions cause the electrons there to vibrate in such a way as to emit an exact electrical image of all earth transmission.

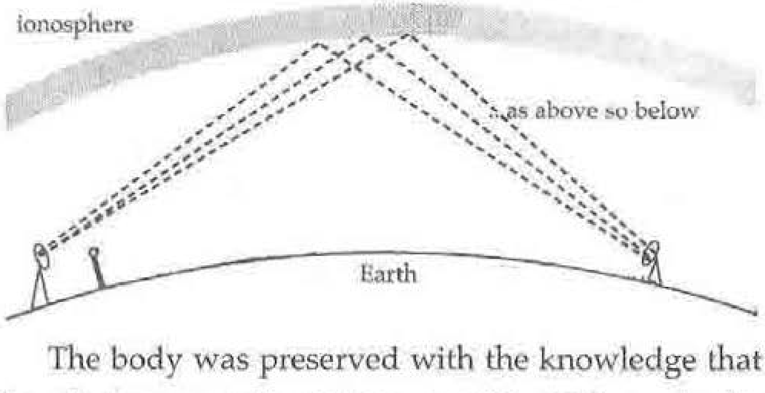

The body was preserved with the knowledge that its electromagnetic nature would continue to be reflected in the Heavens. The concept of a spiritual abode in heaven is as perplexing to the materialist of today as it was in the time of the Messiah. In John chapter 3, the Messiah attempts to explain to a White man why he cannot go to or understand heaven.

> Jesus answered, "Truly, truly, I say to you, unless one is born of water and spirit, he cannot enter the kingdom of God. That which is born of the flesh is flesh, and that which is born of spirit is spirit. Do not marvel that I said to you, 'You must be born anew.' The wind blows where it wills, and you hear the sound of it, but you do not know whence it comes or whither it goes; so it is with everyone who is born of Spirit." Nicode' mus said to him, "How can this be?" Jesus answered him, "Are you a teacher of Israel, and yet you do not understand this? Truly, truly, I say to you, we speak of what we know, and bear witness to what we have seen; but you do not receive our testimony. If I have told you earthly things and You do not believe, how can you believe if I tell you heavenly things? No one has ascended into heaven but he who descended from heaven, the Son of man. And as Moses lifted up the serpent in the wilderness, so must the Son of man be lifted up, that whoever believes in him may have eternal life.
>
> *John 3:5-14*

It would be fruitless at this point to attempt to untangle the enormously confusing theology of the Trinity proposed by Whites who, having discovered themselves excluded, sought to deny the Divinely Blessed the knowledge of their birthright. The secret ingredient, for which our race has been made to suffer is melanin, the dust of the Kha. The word Kha ranges in meaning, from dead body or corpse to powdered medicine or simply the dust of the dead.

Allow me to digress for a moment to examine the use of dust in the early days of radio. A phenomenon, which permitted the detection of radio waves, was the response of dust particles to electromagnetic waves caused by a spark. Particles of dust tend to stick together (coherence) aligning themselves in such a way as to increase the current flow in a certain direction. The early rectifiers, known as coherers, were made by filling a tube with an electrically responsive dust or filings, (See Lodge's transmitter and receiver). When there was a passage of electromagnetic waves the coherer would conduct, however once aligned they were no longer responsive and needed to be tapped periodically to loosen the particles to respond to the next incoming wave. This is the reason why the dreamer's body must be kept rigid, like the dead, for movement disrupts the alignment reducing conductivity.

Christ said to Mary, "Touch me not, for I am not yet ascended unto my father."

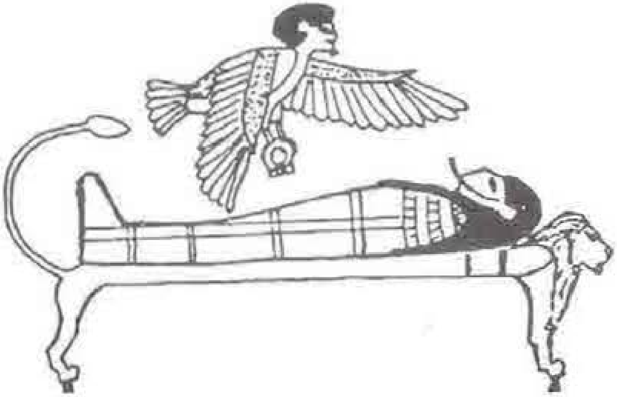

The senkh held by the Ba represents the electromagnetic emissions of the Kha.

The majority carriers (type of charge transport) of melanin are electrons as opposed to (+) holes. Therefore, in the Judgment of Osiris, described by the Millikan's oil drop experiment, the etheric body rises.

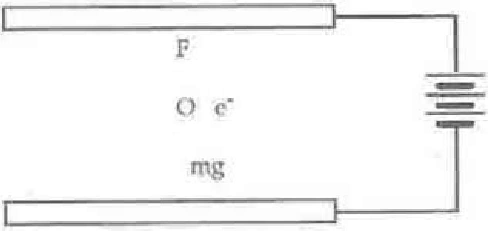

In Millikan's oil drop experiment, an electrically charged oil drop is suspended between two plates by the upward force of an electric field against it's weight mg.

> Who Knoweth the spirit of man that goeth upward, and the spirit of beast that goeth downward to the earth?
>
> *Ecclesiastes 3:21*

The confusion as to the direction of charge flow displays Greek ignorance and Egyptian wisdom. Electrons flow from the so-called negative potential, known as the cathode, which canes from the Greek word "Kathodos" meaning 'the way down'. The Egyptian root of the word "Ka," is represented by the raised arms 𓂓 representing the true direction of electron flow, from the ground upwards.

Each melanin granule is a tiny feather and, when energized by the ankh of levitation, the gods were given the power of flight.

The ankh of levitation was related to the goddess of the Upper Kingdom, the Divine Vulture. These vultures are often seen soaring around large clouds during a thunderstorm. The reason for this behavior was thought to be related to upward convection currents of air. And it was believed that these birds used these currents to attain great heights without effort. This assumption neglected the fact that thunderclouds possessed strong electrostatic charges which exerted a force of attraction on the birds in the same way that a charged plastic comb attracts small pieces of paper.

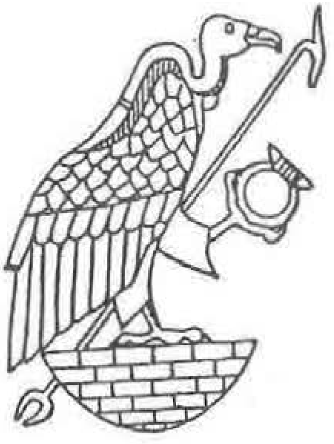

The goddess Nekhbet represented this principle, by which the Messiah was taken up into the clouds. By holding a high frequency emitter like the tet-ankh scepter, one could soar into the sky as the Medieval witches were said to do. At sufficiently high frequencies, the dipole of the melanin granule cannot respond rapidly enough to remain in step with the changing electric field. Therefore, a displacement current arises and the granule behaves as a dielectric, subject to the force of electrostatic attraction and repulsion

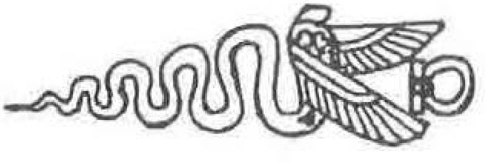

The winged serpent is the most compelling argument in the link between spiritual and electromagnetic phenomenon. While representing all the elements of airborne electromagnetic wave transmission, it evokes the wisdom and spirituality associated with Isis.

SETI (Search for ExtraTerrestrial Intelligence)

In October 1992, NASA will engage in the study of the heavens under the guise of a search for extraterrestrial intelligence in the galaxy. This project, with the very Egyptian sounding acronym, will use the World's largest radio telescopes and advanced multichannel scanners to canvas the sky in an attempt to discover the source of angelic communications.

It is no coincidence that the frequencies of interest to these researchers is around the 21 cm wavelength of the constituents of water (hydrogen: H and hydroxyl: OH), called the "water hole." This is the most prudent choice since the Messiah referred to these beings as composed of water and spirit.

These beings have the some color as the clouds, a glowing gray, and it is said that their appearance is like unto the "the sons of men," that being like children.

Despite the sophistication of the computerized equipment, the project is doomed to failure. Even if angelic communications were detected among the noise of terrestrial transmission, the language of the angels is unknown. I am not suggesting that communication with Angelic Beings is impossible. What is required is nothing less than Ankh Science in the hand of the Oracle.

The synthesis of amino acids (the building blocks of life) under the condition of a primordial atmosphere, with the use of electrical discharge, confirmed the existence of extraterrestrial life. If the chemical composition of amino acids (hydrogen, oxygen, carbon, nitrogen and other trace elements), can be found

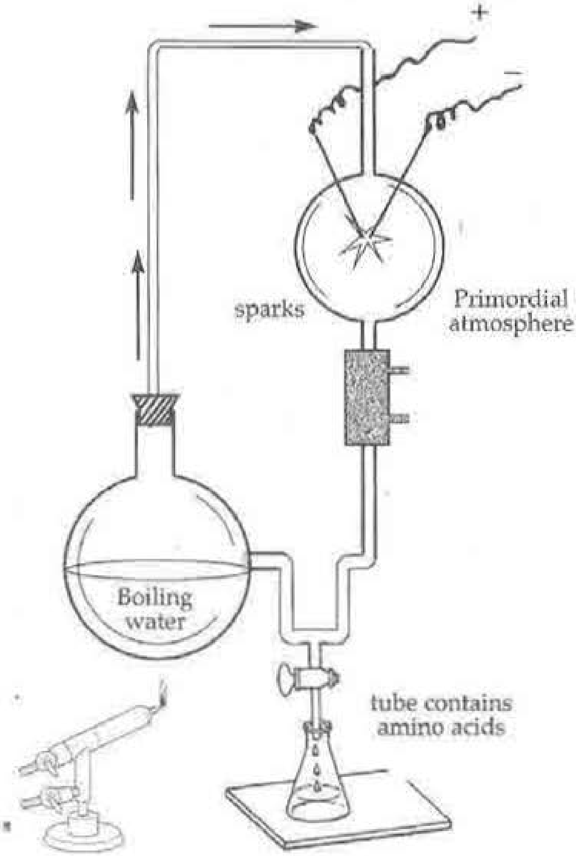
+
–
sparks
Primordial atmosphere
Boiling water
tube contains amino acids

in the upper- atmosphere, then why should life be limited to the earth. The synthesis of more complex molecules (protein and nucleic acids) by linking together amino acids, involves the bonding of an amino group (H_2N) and an acid group ($COOH$). The C—N bonding results in the removal of (H and OH) a water molecule, exactly what SETI is searching for. In this hospitable environment, of the heavenly ocean, the holographic projection of the recently dead is reconstructed, giving birth to the Son of man. This is what is meant by the return of Christ in the clouds, now revealed in the present age of Aquarius.

It is said that these beings traverse the heavens in boats or barges. These UFO's are the Fishers of men. The Government is fully aware of the existence of these beings, and has launched an extensive campaign of disinformation to prevent its citizens from knowing the truth. How could they explain to their people that the lowly Africans are merely beings in a caterpillar stage, destined to be transfigured into an image of splendor and glory. Creatures whose total existence is so eloquently expressed in the life cycle of the Scarab beetle.

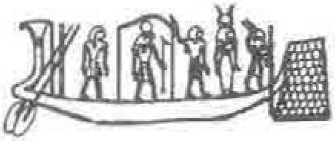

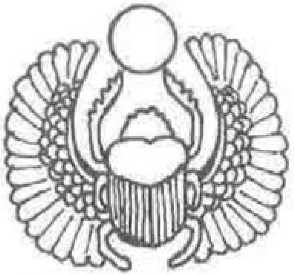

The Scarab Beetle

Symbol of Christianity
the fish or UFO's

The Scarab, symbolic of the god Kheper 𓆣𓂋𓇋𓀭 is a dung beetle commonly found in Egypt. It. is usually seen rolling a ball of dung which contains it's larvae, around the countryside, creating in the mind the Egyptians an analogy with solar orbit. In Egyptian cosmology, the universe is represented as a cow called Nut.

The ball of cow dung is obviously the earth, although the European Egyptologist would like us to believe that our ancestors were as ignorant as theirs were of the Earth's orbit of the Sun.

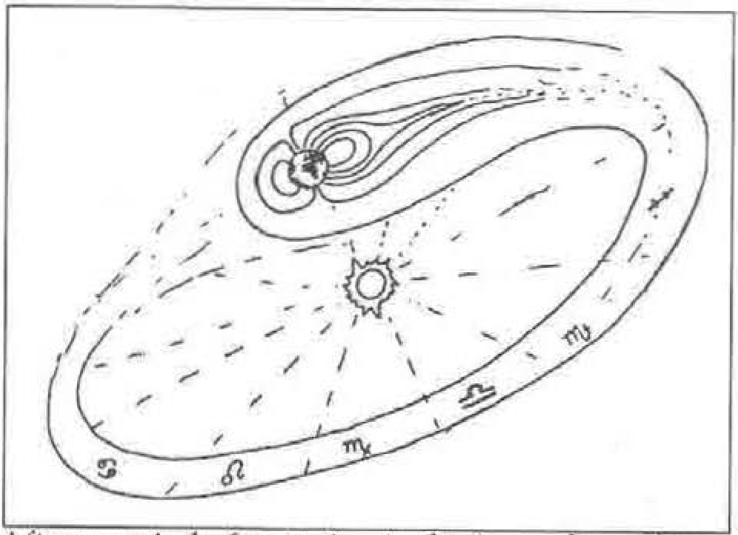

After a period of gestation in the ionosphere, the spiritual body returns to it's astral homeland, one of twelve regions, from which it came. There it will find a glorious existence beyond description, an existence where there is no night only day and it's every desire will be fulfilled in Amentet.

African Astrophysics

THE THEOSOPHICAL CONFLICT between the Big-Bang and the Steady-state model of the universe, is unresolved after 4000 years. It is deeply rooted in the Osirian and Amenite perspectives concerning the origin of the universe.

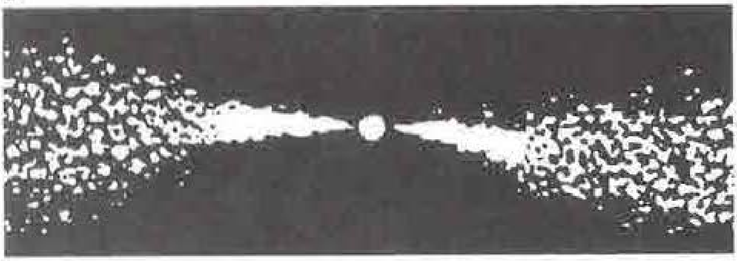

Fig. 1 Radiograph of a typical radio galaxy with jets projecting from its whirling black hole nucleus.

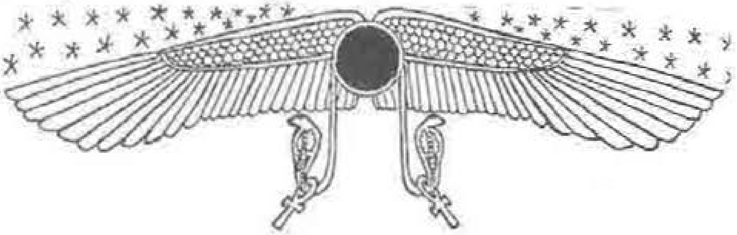

Fig 2 The twin serpents illustrate the unique feature of a radio galaxy which appear as two separate radio sources.

Osiris, God of continuity through the resurrection of the dead, is the Steady-state universe. A philosophy which arose from the African's view of himself as a reflection of the God Osiris, the Ultimate Black. The Ancient African Astrophysicist, studying the universe, also looked inside himself and found that the connection between melanin and the black hole, was the electron (the breath of life 4).

The Grand Unification Theory, which eluded Albert Einstein, has already been solved. The Africans discovered that subatomic particles, like the electron, were created when ordinary matter (energy) passes through a black hole. However, since the Europeans cannot accept Osiris, their scientists have sought diverse theories in order to circumvent the truth of an all powerful Black God.

African
Plasma Physics

In the original myth, Horus lost an eye in his victorious battle against Set (Typhon, typhoon). After his wound was healed by Thoth, he gave the eye to Osiris to eat, which vivified and strengthened Him.

This myth, handed down from Ancient Times, relate the events occurring in the eye of a radio galaxy. Where stars and planets are swallowed by the central black hole and resurrected as fourth-dimensional matter (plasma).

A plasma is characterized by its high electron content. About 95% of the matter in the universe is in this state. A low pressure gas plasma need not be strongly ionized to respond to or produce electromagnetic effects. Black skin, which is composed of a layer of organic semiconductors can be considered a plasma or fourth dimensional matter. And since neuro-melanin and melanocytes are the basis of higher mental activity, it stands to reason that our moral nature is an expression of a more direct contact with God through the spirit. In other words, melanin is your link with the Universal Mind of God and morality is the Universal Law of reconciliation characteristic of our God Osiris. Osiris is the God of Judgment and the cross-road, whether to go right or left on the Path that leads to Him and Peace in the eye of the storm. The Egyptian symbol of a plasm 𓈖 is the same as "n" or water, but it must be considered as living water. It is also illustrated in the symbol of the aurora (𓈖) which most people can recognize as an electrical discharge in what is now called a cathode-ray tube.

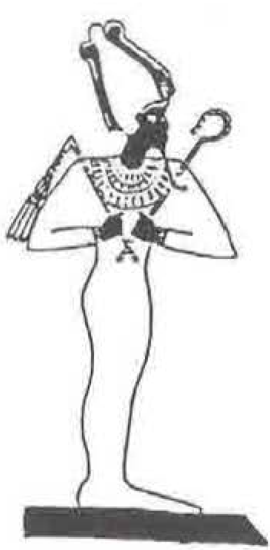

Osiris

African Quantum Physics

SINCE THE DISCOVERY of the electron, modern physicists have found themselves in lockstep with a more ancient reality already defined by our African ancestors. The Tuat is the fourth- dimensional world of the electron and the core of the particle- wave paradox in quantum physics. Only the names have been changed to protect the guilty.

Our present reality is defined by the speed of light, beyond which nothing material can be transmitted, however in the actual reality world of the Tuat, the eye of the Ka sees another light. The light described in the Quran as 'light upon light' and in the Bible as an incomprehensible light that shines in darkness. The Bible is replete with descriptions of this immaterial world, providing a source for physicists, who behind closed doors, are probing the tenents of spirituality in search of answers.

> But, beloved, be not ignorant of this one thing, that one day is with the Lord as a thousand years, and a thousand years as one day.
>
> *2 Peter 3:8*

Here the existence of a non-local reality is defined by a difference in time reference. For if a thousand years is as a day, then something is traveling 365,000 times faster.

The intermediary between this world and the Tuat is the electron, the substance of the soul. It is indivisible, eternal and connected to the mind of God. This alternate reality can be understood in the analogy of the whale, whose dorsal fin emerges when the whale surfaces for air. What we observe is the point-like character of the electron or fin, but beneath the surface lies the greater glory of the whale. The Ka, which surfaces periodically from the sea of eternity, is represented not only as a particle but also a wave (the mythological serpent Qeb. The human body is merely the skin of this mystical serpent which is shed periodically and in which the god walks upon the Earth.

Nur Ankh Amen
The Religious Institute of Kemitronics.

The Science of Kemit

THE RELIGIOUS CONCEPTS of Ancient Egypt were structured on scientifically verified electromagnetic theory, not unsubstantiated traditional beliefs, as ours are today. Separation of science and religion is a modern phenomenon where clashes between cultures in a shrinking world, force compromises with the perception of truth.

Most of us are unprepared to cope with the idea of a superior African science. Years of slavery and colonial domination have damaged our belief in the ability of our race to have accomplished these feats of mental genius. In spite of the pyramids and other colossal monuments, which are a testimony to our greatness, most of us are still in disbelief.

Electromagnetic religion in Kemit can be seen to have played a role in the evolution of monotheism during the reign of Amenhotep IV in the 18th Dynastic Period. The reason for the conflict between the polytheistic doctrine of the past and the singular universal god Aten may have been the result of the discovery of a new source of electromagnetic radiation called Cosmic Background Radiation,

In 1964, experimenters at Princeton University studying the possibility of leftover radiation from the "Big-Bang," discovered the universal microwave radiation hypothesized by Akenaten 3500 years before. The Aten was represented by a disc with many rays ending in hands holding the ankh (symbol of life). Because this source of radiation was universal, it was difficult to depict, therefore, the instrument used to detect its microwave emissions was used.

The Disk of aten was an ancient solar device, used to focus sunlight to a point or beam of extreme intensity. These large domed disks were made of highly polished metal (the word aten means mirror) to concentrate the heat of the sun on a focal point for a variety of uses, which include the splitting of large stone blocks. A stone block was exposed to a narrow beam and, when heated, cold water was poured causing the stone to split in precise dimensions lines.

With the rise of the Amenites, the technology concentrate other invisible electromagnetic wave obsessed with the building of larger and more requiring massive quantities of gold and silver shrines of the ancient gods, which ruined the national economy and created enemies throughout Egypt. The disks were dismantled during the reign of Tutankhamen, to fund the restoration of the ruined temples of the Ancient gods.

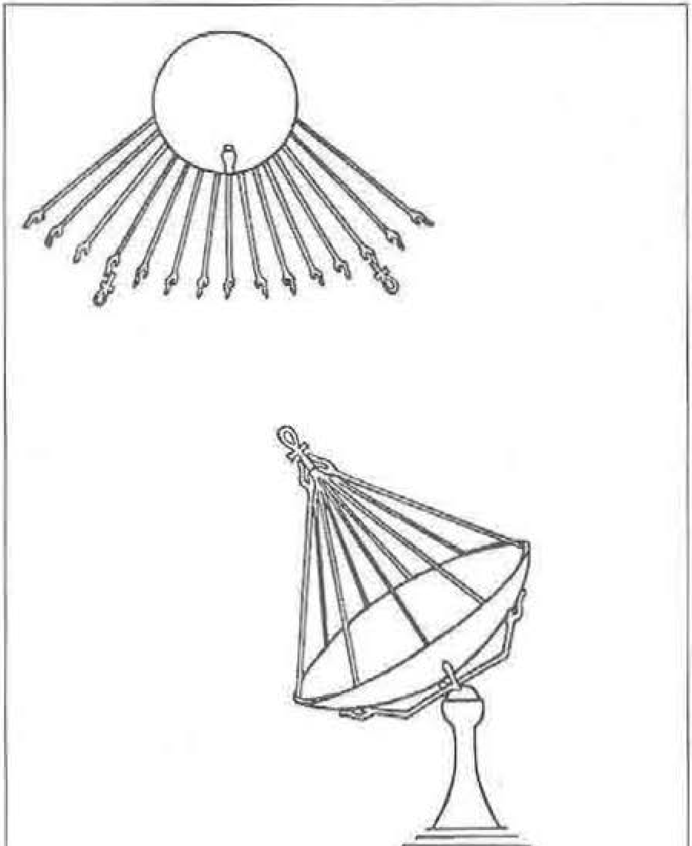

This may be difficult to accept if you are unfamiliar with the two dimensional convention of Egyptian Art as opposed to the modern three dimensional representation.

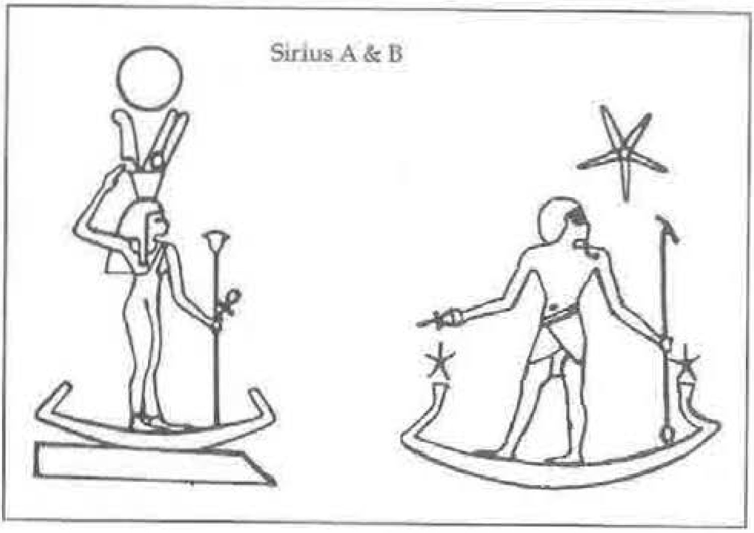

Invisible Sirius B was "discovered" in 1970. She is composed of superdense matter created by an encounter with a black hole.

The Aten was rejected as the universal source of all life because the faintness of the emissions suggested weakness in comparison to the more powerful gods of the past.

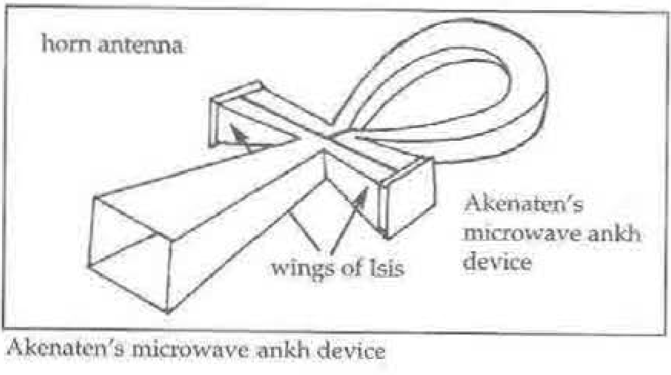
horn antenna
Akenaten's microwave ankh device
wings of Isis

Akenaten's microwave ankh device

The disc of Aten, concentrated the microwave radiation into the horn antenna, exciting the wings of Isis (resonant cavity) into oscillation. The natural frequency of her wings and the incoming microwaves are mixed (heterodyned) and the lower difference frequency, which resulted from the mixing, is detected by the loop. An explanation of the loop's function in this FM discriminator requires an excessive use of technical jargon, which would only detract from the spiritual essence of this divinely inspired device.

This radio telescope could detect radiation from stars and other stellar activities in the cosmos. Akenaten was dismissed as a heretic and is not a favorite among European Egyptologists because of his distinctly African appearance. but he was undoubtedly one of the greatest astronomer- scientists in human history. For he not only discovered the remnants of the Original Spark (the Aten, the Big bang) from which all life originated, but he founded the first religion based on the Oneness of God.

Higher on the electromagnetic frequency scale above microwaves are the invisible infrared rays also known as heat. And by exploring infrared we come to the real reason the pyramids were built.

The word "pyramid" means, fire within (pyr; fire, heat & mid; middle), not the familiar flames, but the Ankhet living fire." A special fire whose flames were life giving. These rays were concentrated or focused by the pyramid into the sacred chamber under its base.

The pyramid was simply a lens which focused a specific frequency of infrared waves on the sacred chamber.

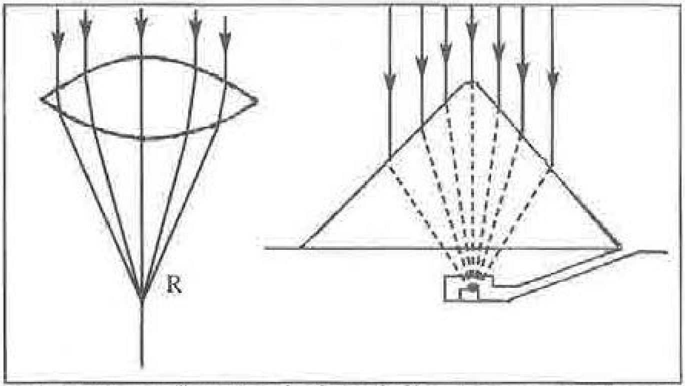

In order to understand this fully, let us explore the question of "living fire."

A by product of both life and combustion is carbon dioxide. If living fire is to exist, carbon dioxide must be eliminated or exhaled and fresh oxygen supplied for further combustion of fuel (organic, fossil or sugar). A flame in an enclosed area would die from the lack of oxygen, which would have been consumed and converted into carbon dioxide.

If carbon dioxide could somehow be converted to it's constituents, carbon and oxygen, the oxygen could be re-used to further continue combustion and produce a living fire.

Carbon dioxide molecules can be split into carbon and oxygen by absorbing energy at specific resonant frequencies. The molecule vibrates in response to the energy absorbed and if there is sufficient concentration of these photons, the molecule splits and oxygen is liberated.

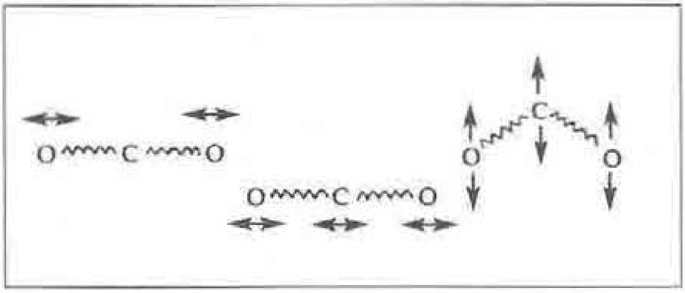

modes of vibration of CO_2

Every object with a temperature above zero degrees Kelvin (K), emits infrared by what is known as Black Body Radiation. The frequency and intensity of the radiation is related to the temperature of the object. At the pyramid's natural temperature of 300 K, there is sufficient infrared at the proper frequency, which the shape of the pyramid can concentrate on the sacred chamber to effect dissociation of carbon dioxide. If the incorrect frequency is used the molecule could split into poisonous carbon monoxide.

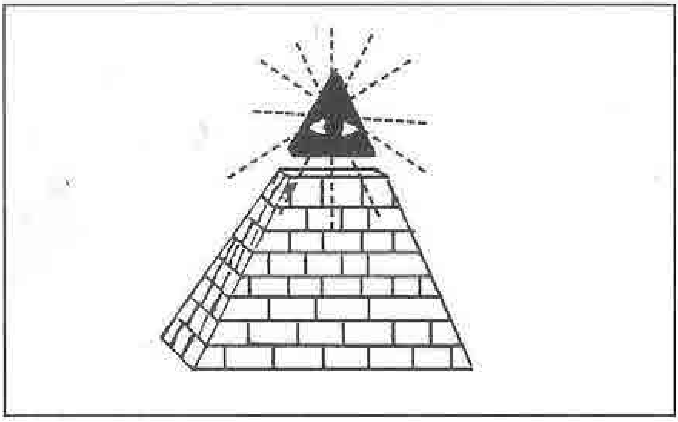

The Sacred Chamber was a special life promoting

Imhotep, builder of the first pyramid

environment where the dead were revived and incredible surgical procedures could be carried out safely. This technology is specific to Africans, simply because white skin is too reflective and would not allow enough energy to be absorbed.

The first pyramid was built by Imhotep, the African multi-genius, whom the Greeks called Asclepius, their god of medicine.

ASCLEPIUS

Legend has it that Asclepius, god of healing and son of Apollo, was delivered by cesarean section performed on his dead mother Coronis by Thoth (Hermes) and taught the art of medicine. Another version states that he was abandoned on a mountain and suckled by goats. The shepherd who discovered him was frightened when he observed a mysterious aura of glistening corona discharges emanating from the child's skin.

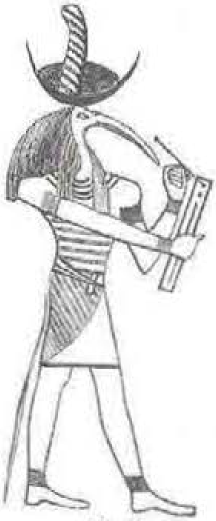

Thoth

His wisdom and skill in the healing arts were unsurpassed, for Asclepius discovered the secret of the Gorgon's blood, which was both a medicine and a deadly poison. With this knowledge he developed a method of bringing the dead back to life. After reviving many people he himself was killed with a lightning bolt from Zeus, for fear that he might reveal the secret which would give immortality to human beings. Apollo, angered by the murder, responded by killing the thunder-making Cyclopes of Zeus.

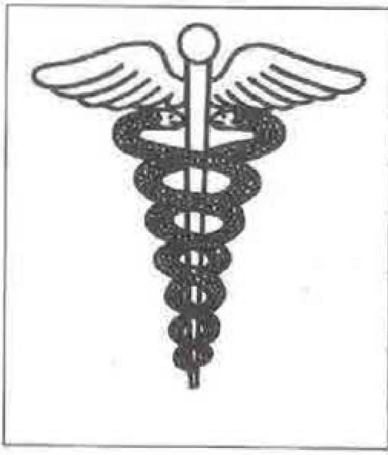

Asclepius was associated with serpents and his emblem of twin winged snakes is still used as the insignia of the medical profession.

Focused infrared rays were also used to dehydrate the bodies of the dead. This sophisticated form of mummification, usually reserved for nobility, produced the most life-like mummies because of the precision of the dehydration process.

However the ultimate use of infrared rays, involved the semiconductor nature of melanin, which requires an infusion of energy to bridge the gap between it's levels of conduction. With this constant source of power the Kha of the Pharaoh would remain an electrically active antenna, supporting his heavenly reflection in divine splendor.

The Ankh of Science

WHEN SANKH WAS HIGH PRIEST, Ankh Science had long ago reached it's zenith in Kemit. The Priesthood of the ▭☥ Ankh, an established University of Ankh Science, maintained the temples and shrines of numerous ankh possessing gods. The most notable of these were the shrine of khonsu, miraculous healing, and Ptah, the god of creative insight. However the Priests of Amen possessed the most powerful shrine, that of spiritual flight or teleportation. There power arouse from a knowledge of themselves, the knowledge that the photonic essence of man which emanated from the soul was indeed his truest self.

> In him was life, and the life was the light of men. And the light shineth in darkness; and the darkness comprehended it not.
>
> *St. John 1:4-5*

> The light of the body is the eye: if therefore thine eye be single' thy whole body shall be full of light.
>
> *Matthew 6:22*

After receiving this revelation, I beseeched my lord and master, the High Priest Sankhamen, continually to grant me the knowledge of the secret workings of the ankh of light, which was referred to as the Neb ankh. Understanding of the Neb ankh was achieved at the culmination of a lifetime of study and religious discipline. Although I am only a messenger, with a message already too complex for those who would receive it, I felt compelled to include some mention of it's enormous importance.

The Neb ankh was an awesome experience which would change forever, the initiate's perception of reality. For it was a foretaste of life after death, a discovery of an existence outside the mass of the body, a feeling of total weightlessness and an inter-dimensional access to the Tuat (the otherworld).

Superior African Technology has already achieved levels which are still thought to be science fiction.

The rewards of African Nationalism will be the ability to mount large scale research efforts, akin to the Space Program or the Manhattan Project, to revive ancestral wisdom. Imagine the cohesive power and national unity in Kemit during the Era of Pyramid Building. Modern technology is still in its infancy and can only scratch the surface of this enormously complicated issue. We should abandon the Pale god and return to the worship (study) of Osiris (the Black God, the electronics of Black skin).

Ptah, portrayed as a mummy, with his hands protruding from the wrappings, holding a staff that combines the djed pillar, ankh sign and sceptre.

Although the active component is the melanin granule, its random orientation renders it ineffective. The electromagnetic emissions from the ankh and the uas (scepter) were used to align these tiny dielectric resonator antennas, in order to produce a coherent radiation pattern. The position of a granule can be adjusted. Because, as shown on page 56, the field of a melanin granule could exert a dipole moment on its neighbor, creating the proper alignment of this microscopic antenna array.

Coherence is the essential property of a laser (Light Amplification by Stimulated Emission of Radiation), and Black skin can be made to lase.

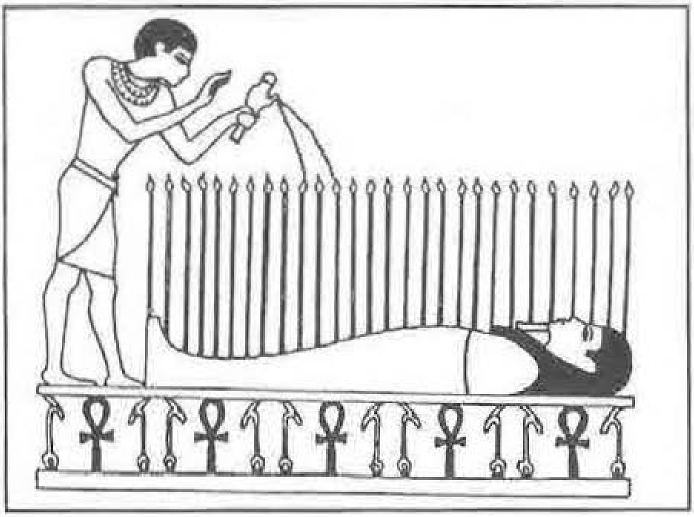

Light rays, in the form of beams tipped by tiny flames, emitted from the body of Osiris, the Neb Ankh.

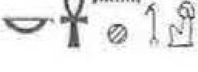

Conclusion

THE BASIC CHARACTER OF African Science is the promotion of life, as opposed to the Euro-centric preoccupation with destructive or weapons based technology. Because of this mind-set, the view that all science has some weapons potential, Europeans are intent on suppressing Ankh Science as a matter of National Security. Their need to mis-educate African children results from a very real fear, that the genetic memory of these mental giants will be awakened. This policy has been in effect since the Greeks first encountered this advanced extraterrestrial civilization of the Nile Valley.

Because of the barbaric nature of the Greeks, the Amen Priests were killed, libraries burned, religion distorted and hieroglyphic writing forbidden. This arms race mentality is a permanent feature of Euro-science, having the effect of placing all science under the scrutiny and directorship of the governing body, in order to prevent a departure from science that is only beneficial to Whites. The encounter with an alien civilization, possessing science developed over thousands of years around unique genetic differences, forced Whites to institute policies of mis-education.

The photo-active nature of melanin is an evolutionary quantum-leap, generating a science, culture and religion totally alien to the invaders of the Nile Valley. If allowed unimpeded development, this advantage could be exploited with miraculous consequences for the African, the greatest benefits being in the medical arena, where laser technology and the protein-producing melanocytes will be wedded to create cures unheard of since the time of Imhotep And speaking of Imhotep, could reviving the dead by using infrared laser techniques to dissociate carbon dioxide in the body be far away? This might not be our greatest achievement, for within the genetic memory of an African child is the knowledge to deoxidize melanin and put flesh on dry bones.

Afterword

WHEN I FIRST READ *The Ankh:African origin of electromagnitism,* I was elated, because all positive information about the Ankh is necessary to awaken the Nubian Kamitik Giants sleeping in the belly of Amenta (the West). The Ankh is the old rugged sign that has been despised by the world. The four major world religions: Hebrew-Israelite, Buddhism, Christianity and Islam were all nurtured by the Ankh people of the Nubian Hapi (Nile) Valley. In this scintillating book, the electromagnetic power of the Ankh has been scientifically investigated by Nur Ankh Amen.

The Egyptological would-be suppressers of AfrakaNubian Spiritual High Kulture continue to deny the primacy of Afrakan Kosmik thought and its application in the mundane realms. These humble Black farmers of the Nile Valley could not possibly have knowledge of electricity and atomic energy. Long before the atom was discovered and misused by western man, Atum appeared on the temple walls of Khamit (Egypt) as the primordial principle of unmanifested potential. It was Atum, cradled on the head of Ptah, the unmoved mover who copulated within her-himself to bring forth Shu-Nifu, the fire fed wind and Tefnut-the materialization of H_2O. These complementary elements sustain the Ankh (Life) principle of the wombniverse.

Nur Ankh Amen has not been criticized because of fear that the reclamation of Ankh Spiritual Science by Afrakan people, will make the foreign imposed religions obsolete. The oppressor knows this much, that the Ankh sign itself is a primordial mental key to unlock the gates of ancestral memory. In his book, *Black Seminarians Without a Black Theology*, Dr. Yosef Ben Jochannan issued a call for restoration of the Ankh to its place of primacy in Afrakan Spiritual High Kulture.

The Ankh is also carried by Nubian Kamitik initiates in the major cities of the West. Bearers of the Ankh are very courageous people who are undaunted by the slander against the Ankh. They have seen beyond the lie that the original AfrakaNubian bearers of the Ankh were 'pagan' and 'heathen' people who 'worshiped idols.' Yet this slander has caused many people of Afrakan ascent to turn away from Ankh science. Bearers of the Ankh are keepers of a sacred legacy which calls them to service as healers in the earth. On the temple walls of Aton-Ra Akhet (called El-Armana today) the Aton Ra (Sun) is chiseled in stone projecting rays terminating in Ankh (Life). The rays of Aton-

Ra are the food of plants. To consume green plants is the way we feed our internal sun within the pineal gland.

The first Ankh bearer was called Heru (in Greek Horus). The aim of all initiates in Ankh Sacred Science is to become transformed into Heru. Heru is the Ankh (Life) advocate ruling the upper realms of your Konsciousness. Heru Asr, the Nubian Falcon of the Holy Spirit, is pictured on temple walls flying with Ankhs in its talons. Heru - the Ankh (Life) messenger came with words of healing and life.

The Socio-Spiritual Meaning of the Ankh

In woman the loop of the Ankh represents the womb, the crossbar the fallopian tubes, the staff the birth canal. In man the loop is his prostate, the crossbar his testes, the staff his penis. Ankh is a life wand. Khamitically speaking our ecological Ankh is that loop of mist drawn up through the power of Aton-Ra's (the Sun's) rays to make clouds which is purified by the Kosmik kiss of Tefnut (moisture). It falls back to earth (the crossbar) as Nun (Rain) or Hapi (die inundated river) fertilizing and bringing the seed of new life (the staff) out of the ground as vegetal food of life in the first times. Ptah's face was green with the promise of renewal and regeneration of life through eating and drinking the green. Queen Afua who as Mut Nebt-Het is Chief Priestess of Purification says, "Eat green, wear green, walk on green" and Priestess Tehuti adds, "Put green in your pocket." Remember electromagnetic Ankhs utilized copper which is oxidized as green. This is the great drama of the rebirth of Asar the Resurrected One. Ankh is the perennial symbol of the Divine Afrakan family in perpetuity. However you look at it, Ankh is life.

Special Note: For Bearers. Wearers and Keepers of Ankh

Ankhians are not selfish, we share the gifts of healing and life with all people who respect the primordial Aftakan source of all earth life. The mitochondrial Eve is a Nubian Khamit woman - Ta-Urt Sekemet Mut Ast.

The Ankh is an Afrakan re-birth concept and symbol which Afraka gave to the world when she preserved the Ankh on temple walls and as amulets. Ankh may be worn by all who adore NTR the one life source as NTR Ankh (Divine Supreme Womb/Seed Source of Life) the Mother/Father Creator. A wearer of the Ankh is not racist, is not violent, does not eat dead or live animals, nor flesh, avoids any foods that cause putrefaction or fermentation stinking up the body and does not eat creatures with eyes. An initiate wearer, carrier or user of the Ankh honors the feminine (womb) principle as the first and eldest manifestation of divine in Kosmos and on earth. She/he is cognizant of the need of Maat in a patriarchal world. Ankh supporters practice Maatiarchy - balanced

Execution of power in female/male relationships. Ankh demands focus and attention to spiritual and material needs simultaneously.

However you see the Ankh either as a Spiritual, Kultural, healing or scientific tool or wand, it is essential to remember that the Creators of the Ankh instrument caused the rise of the longest lasting most righteous civilization in the Hapi (Nile) Khamitic Valley and sent messengers of Ankh (Life) to all people of earth. So tenacious is humanity's hold on Ankh (Life) that Ankh subliminally creeps up in all English words of continuity. All gerund (ing) words/sounds conceal the Hesi (mantra) Ankh - "ing" is continuity throughout all time and all space. The anchor which secures the lives of passengers on a stormy sea conceals the hesi (word/sound) Ankh. When we thank the Creator, we use the word Ankh. Connect conceals Ankh in the letters "onnec." Our interconnectedness with NTR (Divine) is the greatest lesson taught by Ankh (Life Eternal) science. Ask your k-nee (reverse the sound and the movement). Ask your ankle, ask your tongue which stimulates the waters which form the sound of your voice. Ankh the very word in MTU NTR begins with the wave, (Nu) water. According to Gerrald Massey in Ancient Egypt Light of the World, the very word King conceals the Khamite hope of life eternal. Nkh, Ankh - -g may he live. -g Live the -g.

The Ankh unifies us all in spite of religious and philosophical differences because it is the sign of the Wombniversal AfraKosmik Spiritual source of all religions. Without Ankh (life) of what use is religion? Meditate on ankh as you hesi (chant) Ankh over and over. Come to Ankh. Come to Life.

May this book serve to fill in open gaps in your knowledge. Add this scientific and spiritual portion to your store of Ankh knowledge (reverse the first three letters in know what do you get? Onk, life.) Ink makes the word permanent. The written words of wisdom gives life. I offer this humble addendum as I have been asked. May you take your Ankh back. RISE AFRAKANUBIANKHAMITIKUSHITE PEOPLE, RISE!

Forever yours in Ankhtuity,
For Ankhternity,
Ankhtually,
Heru Ankh Semahj sa Ptah
Studio of Ptah
155 Canal Street, suite 9
New York, NY, 10013
1-212-226-8487
1-212-343-9706

Glossary

ankh: A tau cross with a loop on top.

diety: A person or thing revered to as supremely powerful or beneficent.

djed: Cruciform symbol with at least three cross bars.

lodestone: naturally occurring mineral that is strongly magnetic.

aurora: Intermittent electrical discharge that occurs in the rarefied upper atmosphere.

ionosphere: An ionized layer in a planetary atmosphere where free electrons and ions with thermal energies exist under the control of earth's gravitational and magnetic fields.

ions: electrically charged atoms or group of atoms.

electron: negatively charged elementary particle of the atom.

DNA: The nucleic acid forming the genetic material of all cells.

Heh: God of infinity.

chakras: The points located along the body that are known as energy centers. They are seven of them.

diode An electrical device that allows electricity to flow in one direction only.

SETI: Search for Extraterrestrial Intelligence

UFO Unidentified Flying Object.

Khnum: A Ram-headed creator god often depicted at a pot-

SPRINGER BRIEFS IN
ELECTRICAL AND COMPUTER ENGINEERING

Saleh Faruque

Radio Frequency Modulation Made Easy

Springer

SpringerBriefs in Electrical and Computer Engineering

More information about this series at http://www.springer.com/series/10059

Saleh Faruque

Radio Frequency Modulation Made Easy

Springer

Saleh Faruque
Department of Electrical Engineering
University of North Dakota
Grand Forks, ND
USA

ISSN 2191-8112 ISSN 2191-8120 (electronic)
SpringerBriefs in Electrical and Computer Engineering
ISBN 978-3-319-41200-9 ISBN 978-3-319-41202-3 (eBook)
DOI 10.1007/978-3-319-41202-3

Library of Congress Control Number: 2016945147

© The Author(s) 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

By inventing the wireless transmitter or radio in 1897, the Italian physicist Tomaso Guglielmo Marconi added a new dimension to the world of communications. This enabled the transmission of the human voice through space without wires. For this epoch-making invention, this illustrious scientist was honored with the Nobel Prize for Physics in 1909. Even today, students of wireless or radio technology remember this distinguished physicist with reverence. A new era began in Radio Communications.

The classical Marconi radio used a modulation technique known today as “Amplitude Modulation” or just AM. This led to the development of Frequency Modulation (FM), amplitude shift keying (ASK), phase shift keying (PSK), etc. Today, these technologies are extensively used in various wireless communication systems. These modulation techniques form an integral part of academic curricula today.

This book presents a comprehensive overview of the various modulation techniques mentioned above. Numerous illustrations are used to bring students up-to-date in key concepts and underlying principles of various analog and digital modulation techniques. In particular, the following topics will be presented in this book:

- Amplitude Modulation (AM)
- Frequency Modulation (FM)
- Bandwidth occupancy in AM and FM
- Amplitude shift keying (ASK)
- Frequency shift keying (FSK)
- Phase shift keying (PSK)
- N-ary coding and M-ary modulation
- Bandwidth occupancy in ASK, FSK, and PSK

This text has been primarily designed for electrical engineering students in the area of telecommunications. However, engineers and designers working in the area of wireless communications would also find this text useful. It is assumed that the student is familiar with the general theory of telecommunications.

In closing, I would like to say a few words about how this book was conceived. It came out of my long industrial and academic career. During my teaching tenure at the University of North Dakota, I developed a number of graduate-level elective courses in the area of telecommunications that combine theory and practice. This book is a collection of my courseware and research activities in wireless communications.

I am grateful to UND and the School for the Blind, North Dakota, for affording me this opportunity. This book would never have seen the light of day had UND and the State of North Dakota not provided me with the technology to do so. My heartfelt salute goes out to the dedicated developers of these technologies, who have enabled me and others visually impaired to work comfortably.

I would like to thank my beloved wife, Yasmin, an English Literature buff and a writer herself, for being by my side throughout the writing of this book and for patiently proofreading it. My darling son, Shams, an electrical engineer himself, provided technical support in formulation and experimentation when I needed it. For this, he deserves my heartfelt thanks.

Finally, thanks are also to my doctoral student Md. Maruf Ahamed who found time in his busy schedule to assist me with the simulations, illustrations, and the verification of equations.

In spite of all this support, there may still be some errors in this book. I hope that my readers forgive me for them. I shall be amply rewarded if they still find this book useful.

Grand Forks, USA Saleh Faruque
May 2016

Contents

1 **Introduction to Modulation** 1
  1.1 Background 1
  1.2 Modulation by Analog Signals 3
    1.2.1 AM, FM, and PM 3
    1.2.2 AM and FM Bandwidth at a Glance 4
  1.3 Modulation by Digital Signal 5
    1.3.1 Amplitude Shift Keying (ASK) Modulation 5
    1.3.2 Frequency Shift Keying (FSK) Modulation. 6
    1.3.3 Phase Shift Keying (PSK) Modulation. 7
  1.4 Bandwidth Occupancy in Digital Modulation 7
    1.4.1 Spectral Response of the Encoded Data 8
    1.4.2 Spectral Response of the Carrier Frequency Before Modulation. 9
    1.4.3 ASK Bandwidth at a Glance. 10
    1.4.4 FSK Bandwidth at a Glance 11
    1.4.5 BPSK Bandwidth at a Glance. 12
  1.5 Conclusions 14
  References 15

2 **Amplitude Modulation (AM)** 17
  2.1 Introduction 17
  2.2 Amplitude Modulation 19
  2.3 AM Spectrum and Bandwidth. 20
    2.3.1 Spectral Response of the Input Modulating Signal. 20
    2.3.2 Spectral Response of the Carrier Frequency 21
    2.3.3 AM Spectrum and Bandwidth. 21
    2.3.4 AM Response Due to Low and High Modulating Signals 23
    2.3.5 AM Demodulation 24
    2.3.6 Drawbacks in AM. 24

2.4 Double Sideband-Suppressed Carrier (DSBSC) 25
    2.4.1 DSBSC Modulation. 25
    2.4.2 Generation of DSBSC Signal 26
    2.4.3 DSBSC Spectrum and Bandwidth 27
    2.4.4 DSBSC Drawbacks 28
2.5 Single Sideband (SSB) Modulation 29
    2.5.1 Why SSB Modulation? 29
    2.5.2 Generation of SSB-Modulated Signal. 29
    2.5.3 SSB Spectrum and Bandwidth 30
2.6 Conclusions 32
References 32

3 Frequency Modulation (FM) 33
3.1 Introduction 33
3.2 Frequency Modulation (FM). 34
    3.2.1 Background 34
    3.2.2 The Basic FM 35
3.3 FM Spectrum and Bandwidth 37
    3.3.1 Spectral Response of the Input Modulating Signal. 37
    3.3.2 Spectral Response of the Carrier Frequency 38
    3.3.3 FM Spectrum 39
    3.3.4 Carson's Rule and FM Bandwidth. 40
    3.3.5 Bessel Function and FM Bandwidth 41
    3.3.6 FM Bandwidth Dilemma 42
3.4 Conclusions 44
References 44

4 Amplitude Shift Keying (ASK). 45
4.1 Introduction 45
4.2 ASK Modulation. 46
4.3 ASK Demodulation 48
4.4 ASK Bandwidth 49
    4.4.1 Spectral Response of the Encoded Data 49
    4.4.2 Spectral Response of the Carrier Frequency Before Modulation. 51
    4.4.3 ASK Bandwidth at a Glance. 51
4.5 BER Performance 53
4.6 Conclusions 54
References 55

5 Frequency Shift Keying (FSK) 57
5.1 Introduction 57
5.2 Frequency Shift Keying (FSK) Modulation. 58
5.3 Frequency Shift Keying (FSK) Demodulation 60

5.4 FSK Bandwidth. 61
5.4.1 Spectral Response of the Encoded Data 61
5.4.2 Spectral Response of the Carrier Frequency Before Modulation. 63
5.4.3 FSK Bandwidth at a Glance 63
5.5 BER Performance 65
5.6 Conclusions 67
References 67

6 Phase Shift Keying (PSK) 69
6.1 Introduction 69
6.2 Binary Phase Shift Keying (BPSK) 70
6.2.1 BPSK Modulation. 70
6.2.2 BPSK Demodulation 73
6.3 QPSK Modulation 74
6.4 8PSK Modulation 75
6.5 16PSK Modulation 75
6.6 PSK Spectrum and Bandwidth 77
6.6.1 Spectral Response of the Encoded Data 77
6.6.2 Spectral Response of the Carrier Before Modulation 79
6.6.3 BPSK Spectrum 79
6.7 Conclusions 82
References 82

7 N-Ary Coded Modulation. 85
7.1 Introduction 85
7.2 N-Ary Convolutional Coding and M-Ary Modulation 86
7.2.1 Background 86
7.2.2 Generation of Complementary Convolutional Codes 86
7.2.3 2-Ary Convolutional Coding with QPSK Modulation 88
7.2.4 4-Ary Convolutional Coding with 16PSK Modulation 89
7.3 N-Ary Convolutional Decoder. 91
7.3.1 Correlation Receiver 91
7.3.2 Error Correction Capabilities of N-Ary Convolutional Codes 93
7.4 N-Ary Orthogonal Coding and M-Ary Modulation 94
7.4.1 Background 94
7.4.2 Orthogonal Codes 95
7.4.3 2-Ary Orthogonal Coding with QPSK Modulation 95
7.4.4 4-Ary Orthogonal Coding with 16PSK Modulation 97
7.4.5 2-Ary Orthogonal Decoding 97
7.4.6 4-Ary Orthogonal Decoding 99
7.4.7 Error Correction Capabilities of N-Ary Orthogonal Codes 99
7.5 Conclusions 103
References 104

Chapter 1
Introduction to Modulation

Topics

- Background
- Modulation by Analog Signal
- AM and FM Bandwidth at a Glance
- Modulation by Digital Signal
- ASK, FSK and PSK Bandwidth at a Glance

1.1 Background

Modulation is a technique that changes the characteristics of the carrier frequency in accordance with the input signal. Figure 1.1 shows the conceptual block diagram of a modern wireless communication system, where the modulation block is shown in the inset of the dotted block. As shown in the figure, modulation is performed at the transmit side and demodulation is performed at the receive side. This is the final stage of any radio communication system. The preceding two stages have been discussed elaborately in my previous book in this series [1, 2].

The output signal of the modulator, referred to as the modulated signal, is fed into the antenna for propagation. Antenna is a reciprocal device that transmits and receives the modulated carrier frequency. The size of the antenna depends on the wavelength (λ) of the sinusoidal wave where

$\lambda = c/f$ m
c = Velocity of light = 3×10^8 m/s
f = Frequency of the sinusoidal wave, also known as "carrier frequency."

Therefore, a carrier frequency much higher than the input signal is required to keep the size of the antenna at an acceptable limit. For these reasons, a high-frequency carrier signal is used in the modulation process. In this process, the

© The Author(s) 2017

S. Faruque, *Radio Frequency Modulation Made Easy*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-41202-3_1

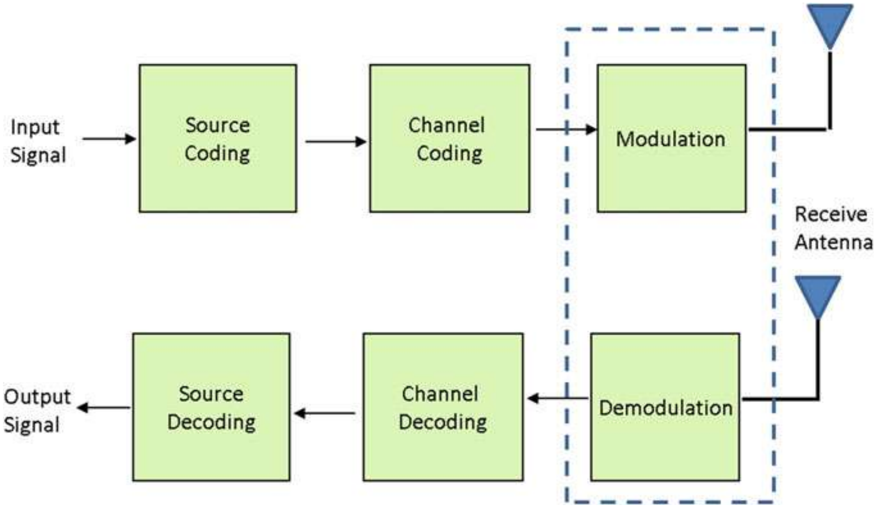

Fig. 1.1 Block diagram of a modern full-duplex communication system. The modulation stage is shown as a *dotted* block

low-frequency input signal changes the characteristics of the high-frequency carrier in a certain manner, depending on the modulation technique. Furthermore, as the size and speed of digital data networks continue to expand, bandwidth efficiency becomes increasingly important. This is especially true for broadband communication, where the digital signal processing is done keeping in mind the available bandwidth resources.

Hence, modulation is a very important step in the transmission of information. The information can be either analog or digital, where the carrier is a high-frequency sinusoidal waveform. As stated earlier, the input signal (analog or digital) changes the characteristics of the carrier waveform. Therefore, there are two basic modulation schemes as listed below:

- Modulation by analog signals and
- Modulation by digital signals.

This book presents a comprehensive overview of these modulation techniques in use today. Numerous illustrations are used to bring students up-to-date in key concepts and underlying principles of various analog and digital modulation techniques. For a head start, brief descriptions of each of these modulation techniques are presented below.

1.2 Modulation by Analog Signals

1.2.1 AM, FM, and PM

For analog signals, there are three well-known modulation techniques as listed below:

- Amplitude Modulation (AM),
- Frequency Modulation (FM),
- Phase Modulation (PM).

By inventing the wireless transmitter or radio in 1897, the Italian physicist Tomaso Guglielmo Marconi added a new dimension to the world of communications [3, 4]. This enabled the transmission of the human voice through space without wires. For this epoch-making invention, this illustrious scientist was honored with the Nobel Prize for Physics in 1909. Even today, students of wireless or radio technology remember this distinguished physicist with reverence. A new era began in Radio Communications. The classical Marconi radio used a modulation technique known today as "Amplitude Modulation" or just AM. In AM, the amplitude of the carrier changes in accordance with the input analog signal, while the frequency of the carrier remains the same. This is shown in Fig. 1.2 where

- $m(t)$ is the input modulating audio signal,
- $C(t)$ is the carrier frequency, and
- $S(t)$ is the AM-modulated carrier frequency.

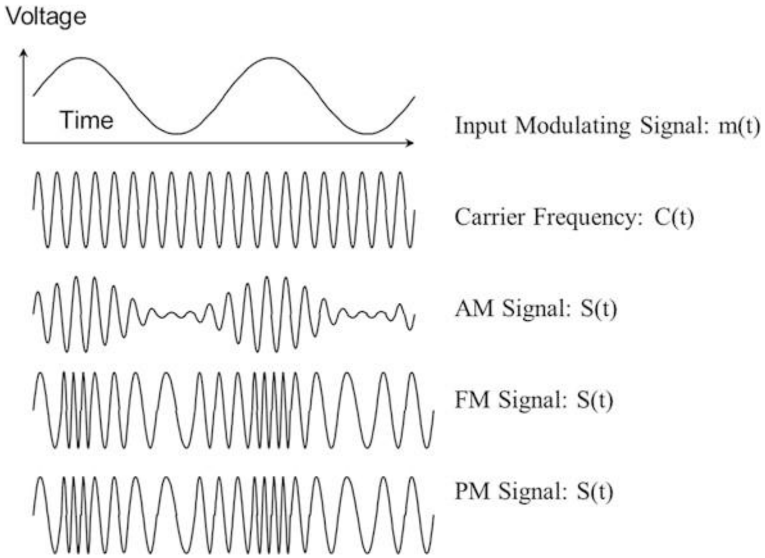

Fig. 1.2 Modulation by analog signal

As shown in the figure, the audio waveform changes the amplitude of the carrier to determine the envelope of the modulated carrier. This enables the receiver to extract the audio signal by demodulation. Notice that the amplitude of the carrier changes in accordance with the input signal, while the frequency of the carrier does not change after modulation. It can be shown that the modulated carrier $S(t)$ contains several spectral components, requiring frequency-domain analysis, which will be addressed in Chap. 2. It may be noted that AM is vulnerable to signal amplitude fading.

In Frequency Modulation (FM), the frequency of the carrier changes in accordance with the input modulation signal as shown in Fig. 1.2 [5]. Notice that in FM, only the frequency changes while the amplitude remains the same. Unlike AM, FM is more robust against signal amplitude fading. For this reason, FM is more attractive in commercial FM radio. In Chap. 3, it will be shown that in FM, the modulated carrier contains an infinite number of sideband due to modulation. For this reason, FM is also bandwidth inefficient.

Similarly, in Phase Modulation (PM), the phase of the carrier changes in accordance with the phase of the carrier, while the amplitude of the carrier does not change. PM is closely related to FM. In fact, FM is derived from the rate of change of phase of the carrier frequency. Both FM and PM belong to the same mathematical family. We will discuss this more elaborately in Chap. 3.

1.2.2 AM and FM Bandwidth at a Glance

The bandwidth occupied by the modulated signal depends on bandwidth of the input signal and the modulation method as shown in Fig. 1.3. Note that the unmodulated carrier itself has zero bandwidth.

In AM:

- The modulated carrier has two sidebands (upper and lower) and
- Total bandwidth = 2 × base band.

In FM:

- The carrier frequency shifts back and forth from the nominal frequency by Δf, where Δf is the frequency deviation.
- During this process, the modulated carrier creates an infinite number of spectral components, where higher-order spectral components are negligible.
- The approximate FM bandwidth is given by the Carson's rule:
  - FM BW = $2f\,(1 + \beta)$
  - f = Base band frequency
  - β = Modulation index
  - $\beta = \Delta f/f$
  - Δf = Frequency deviation.

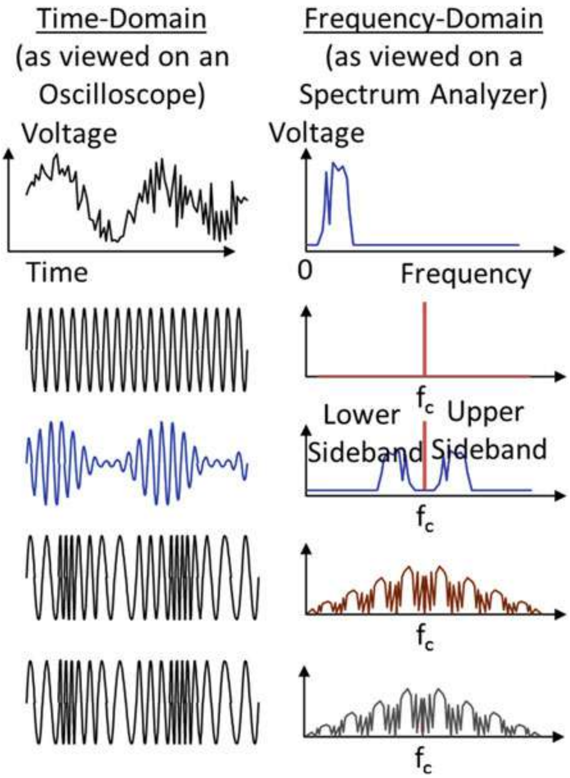
Time-Domain
(as viewed on an Oscilloscope)
Voltage
Time
Frequency-Domain
(as viewed on a Spectrum Analyzer)
Voltage
0
Frequency
f_c
Lower Sideband
Upper Sideband
f_c
f_c
f_c

Fig. 1.3 Bandwidth occupancy in AM, FM, and PM signals

1.3 Modulation by Digital Signal

For digital signals, there are several modulation techniques available. The three main digital modulation techniques are as follows:

- Amplitude shift keying (ASK),
- Frequency shift keying (FSK), and
- Phase shift keying (PSK).

Figure 1.4 illustrates the modulated waveforms for an input modulating digital signal. Brief descriptions of each of these digital modulation techniques along with the respective spectral responses and bandwidth are presented below.

1.3.1 *Amplitude Shift Keying (ASK) Modulation*

Amplitude shift keying (ASK), also known as on–off keying (OOK), is a method of digital modulation that utilizes amplitude shifting of the relative amplitude of the

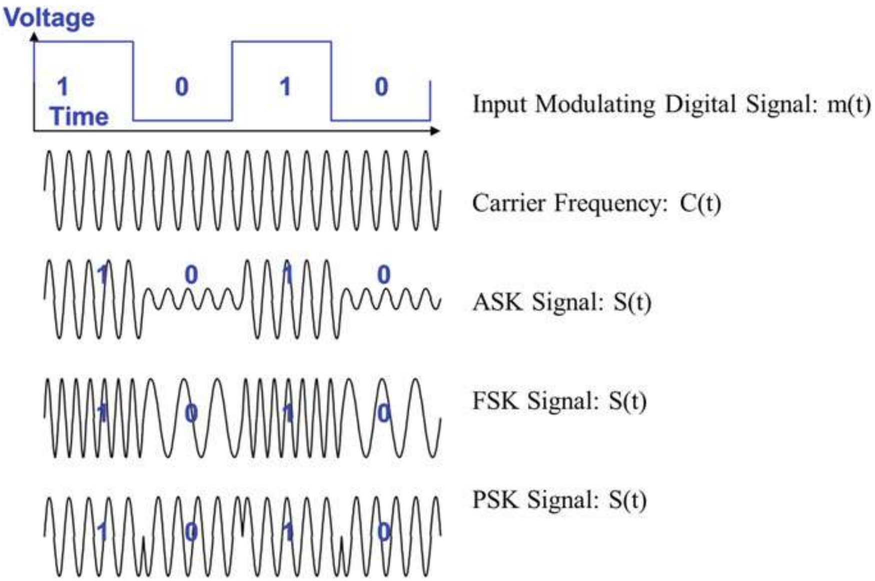
Voltage
1
0
1
0
Time
Input Modulating Digital Signal: m(t)
Carrier Frequency: C(t)
1
0
1
0
ASK Signal: S(t)
1
0
1
0
FSK Signal: S(t)
1
0
1
0
PSK Signal: S(t)

Fig. 1.4 Modulation by digital signal

career frequency [6–8]. The signal to be modulated and transmitted is binary; this is referred to as ASK, where the amplitude of the carrier changes in discrete levels, in accordance with the input signal, as shown.

- Binary 0 (bit 0): Amplitude = Low and
- Binary 1 (bit 1): Amplitude = High.

Figure 1.4 shows the ASK-modulated waveform where

- Input digital signal is the information we want to transmit.
- Carrier is the radio frequency without modulation.
- Output is the ASK-modulated carrier, which has two amplitudes corresponding to the binary input signal. For binary signal 1, the carrier is ON. For the binary signal 0, the carrier is OFF. However, a small residual signal may remain due to noise, interference, etc.

1.3.2 *Frequency Shift Keying (FSK) Modulation*

Frequency shift keying (FSK) is a method of digital modulation that utilizes frequency shifting of the relative frequency content of the signal [6–8]. The signal to be modulated and transmitted is binary; this is referred to as binary FSK (BFSK), where the carrier frequency changes in discrete levels, in accordance with the input signal as shown below:

- Binary 0 (bit 0): Frequency = $f + \Delta f$.
- Binary 1 (bit 1): Frequency = $f - \Delta f$.

Figure 1.4 shows the FSK-modulated waveform where

- Input digital signal is the information we want to transmit.
- Carrier is the radio frequency without modulation.
- Output is the FSK-modulated carrier, which has two frequencies ω_1 and ω_2, corresponding to the binary input signal.
- These frequencies correspond to the messages binary 0 and 1, respectively.

1.3.3 Phase Shift Keying (PSK) Modulation

Phase shift keying (PSK) is a method of digital modulation that utilizes phase of the carrier to represent digital signal [6–8]. The signal to be modulated and transmitted is binary; this is referred to as binary PSK (BPSK), where the phase of the carrier changes in discrete levels, in accordance with the input signal as shown below:

- Binary 0 (bit 0): $\text{Phase}_1 = 0°$.
- Binary 1 (bit 1): $\text{Phase}_2 = 180°$.

Figure 1.4 shows the modulated waveform where

- Input digital signal is the information we want to transmit.
- Carrier is the radio frequency without modulation.
- Output is the BPSK-modulated carrier, which has two phases φ_1 and φ_2 corresponding to the two information bits.

1.4 Bandwidth Occupancy in Digital Modulation

In wireless communications, the scarcity of RF spectrum is well known. For this reason, we have to be vigilant about using transmission bandwidth. The transmission bandwidth depends on the following:

- Spectral response of the encoded data,
- Spectral response of the carrier frequency, and
- Modulation type (ASK, FSK, PSK), etc.

Let us take a closer look!

1.4.1 Spectral Response of the Encoded Data

In digital communications, data is generally referred to as a non-periodic digital signal. It has two values:

- Binary-1 = High, Period = T.
- Binary-0 = Low, Period = T.

Also, data can be represented in two ways:

- Time-domain representation and
- Frequency-domain representation.

The time-domain representation (Fig. 1.5a), known as non-return-to-zero (NRZ), is given by

$$\begin{aligned} V(t) &= V \qquad < 0 < t < T \\ &= 0 \qquad \text{elsewhere} \end{aligned} \tag{1.1}$$

The frequency-domain representation is given by "Fourier transform" [9]:

$$V(\omega) = \int_0^T V \cdot e^{-j\omega t} dt \tag{1.2}$$

$$|V(\omega)| = VT\left[\frac{\sin(\omega T/2)}{\omega T/2}\right]$$

$$P(\omega) = \left(\frac{1}{T}\right)|V(\omega)|^2 = V^2 T\left[\frac{\sin(\omega T/2)}{\omega T/2}\right]^2 \tag{1.3}$$

Here, $P(\omega)$ is the power spectral density. This is plotted in (Fig. 1.5b). The main lobe corresponds to the fundamental frequency and side lobes correspond to harmonic components. The bandwidth of the power spectrum is proportional to the frequency. In practice, the side lobes are filtered out since they are relatively insignificant with respect to the main lobe. Therefore, the one-sided bandwidth is given by the ratio $f/fb = 1$. In other words, the one-sided bandwidth $= f = f_b$, where $f_b = R_b = 1/T$, T being the bit duration.

The general equation for two-sided response is given by

$$V(\omega) = \int_{-\alpha}^{\alpha} V(t) \cdot e^{-j\omega t} \mathrm{d}t$$

In this case, $V(\omega)$ is called two-sided spectrum of $V(\mathrm{t})$. This is due to both positive and negative frequencies used in the integral. The function can be a voltage

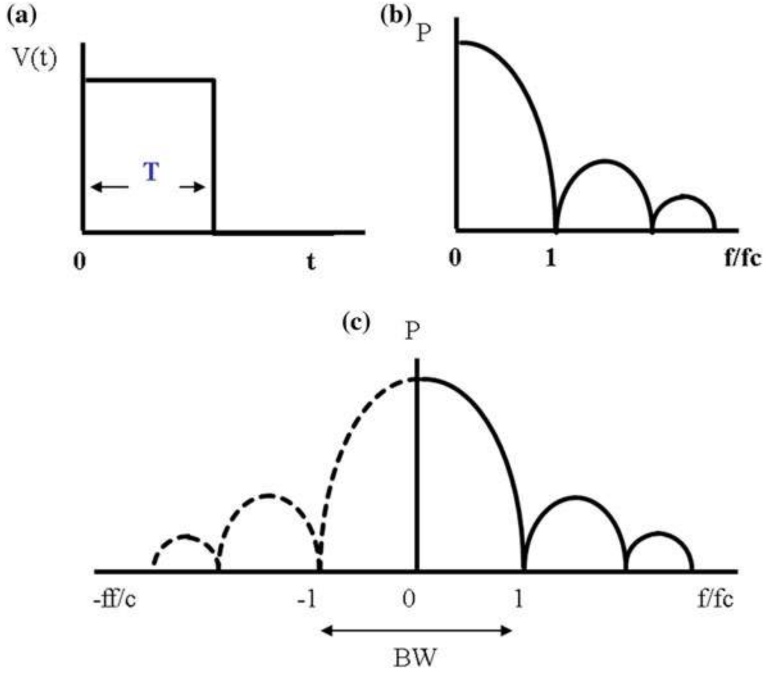

Fig. 1.5 **a** Discrete time digital signal, **b** it is one-sided power spectral density, and **c** two-sided power spectral density. The bandwidth associated with the non-return-to-zero (NRz) data is $2R_b$, where R_b is the bit rate

or a current (Fig. 1.5c) shows the two-sided response, where the bandwidth is determined by the main lobe as shown below:

$$\text{Two sided bandwidth (BW)} = 2R_b \; (R_b = \text{Bit rate before coding}) \tag{1.4}$$

1.4.2 *Spectral Response of the Carrier Frequency Before Modulation*

A carrier frequency is essentially a sinusoidal waveform, which is periodic and continuous with respect to time. It has one frequency component. For example, the sine wave is described by the following time-domain equation:

$$V(t) = V_p \sin(\omega t_c) \tag{1.5}$$

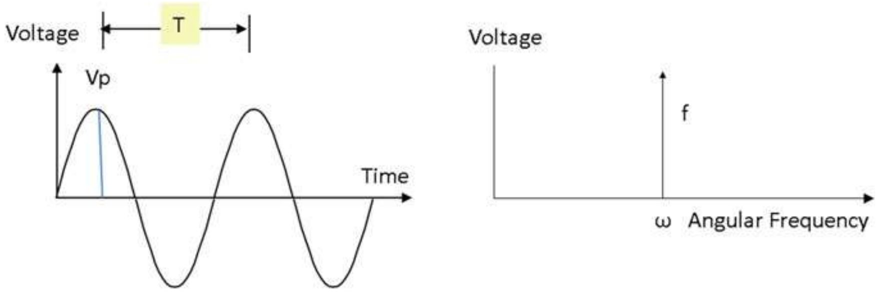

Fig. 1.6 A sine wave and its frequency response

where

$$V_p = \text{Peak voltage}$$

- $\omega_c = 2\pi f_c$
- f_c = Carrier frequency in Hz

Figure 1.6 shows the characteristics of a sine wave and its spectral response. Since the frequency is constant, its spectral response is located in the horizontal axis and the peak voltage is shown in the vertical axis. The corresponding bandwidth is zero.

1.4.3 ASK Bandwidth at a Glance

In ASK, the amplitude of the carriers changes in discrete levels, in accordance with the input signal where

- Input data: $m(t) = 0$ or 1.
- Carrier frequency: $C(t) = A_c \cos(\omega_c t)$.
- Modulated carrier: $S(t) = m(t)C(t) = m(t)A_c \cos(\omega_c t)$.

Since $m(t)$ is the input digital signal and it contains an infinite number of harmonically related sinusoidal waveforms and that we keep the fundamental and filter out the higher-order components, we write:

$$m(t) = A_m \sin(\omega_m t)$$

The ASK-modulated signal then becomes:

$$\begin{aligned} S(t) &= m(t)S(t) = A_m A_c \sin(\omega_m t)\cos(\omega_c) \\ &= A_m A_c \cos(\omega_c \pm \omega_m) \end{aligned}$$

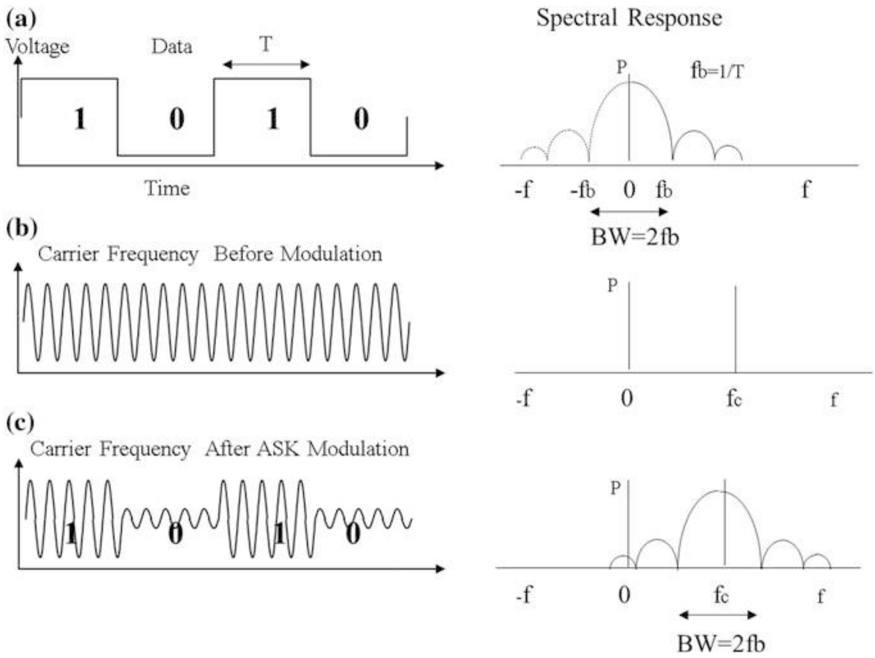
(a)
Voltage
Data
T
1
0
1
0
Time
Spectral Response
P
fb=1/T
-f
-fb
0
fb
f
BW=2fb
(b)
Carrier Frequency Before Modulation
P
-f
0
fc
f
(c)
Carrier Frequency After ASK Modulation
1
0
1
0
P
-f
0
fc
f
BW=2fb

Fig. 1.7 ASK bandwidth at a glance. **a** Spectral response of NRZ data before modulation. **b** Spectral response of the carrier before modulation. **c** Spectral response of the carrier after modulation. The transmission bandwidth is $2f_b$, where f_b is the bit rate and $T = 1/f_b$ is the bit duration for NRZ data

The spectral response is depicted in Fig. 1.7. Notice that the spectral response after ASK modulation is the shifted version of the NRZ data. Bandwidth is given by,

BW = $2R_b$ (coded), where R_b is the coded bit rate.

1.4.4 FSK Bandwidth at a Glance

In FSK, the frequency of the carrier changes in two discrete levels, in accordance with the input signals. We have:

- Input data: $m(t) = 0$ or 1
- Carrier frequency: $C\ (t) = A\ \cos\ (\omega t)$
- Modulated carrier: $S(t) = A\ \cos(\omega - \Delta\ \omega)t$, For $m(t) = 1$
  $S(t) = A\ \cos(\omega + \Delta\omega)t$, For $m(t) = 0$

where

- $S(t)$ = The modulated carrier,
- A = Amplitude of the carrier,

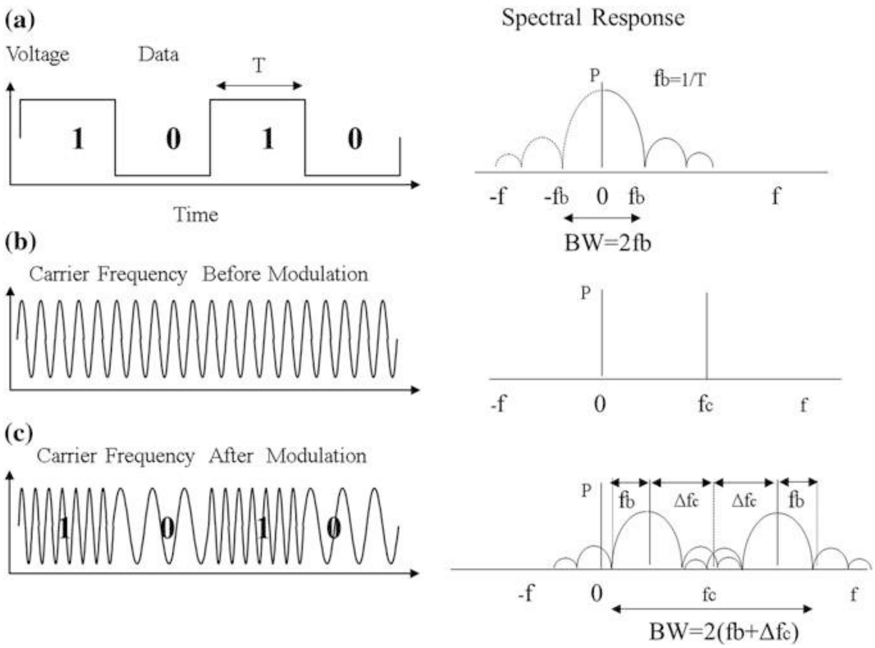
(a)
Voltage
Data
T
1
0
1
0
Time
Spectral Response
P
fb=1/T
-f
-fb
0
fb
f
BW=2fb
(b)
Carrier Frequency Before Modulation
P
-f
0
fc
f
(c)
Carrier Frequency After Modulation
1
0
1
0
P
fb
Δfc
Δfc
fb
-f
0
fc
f
BW=2(fb+Δfc)

Fig. 1.8 FSK bandwidth at a glance. **a** Spectral response of NRZ data before modulation. **b** Spectral response of the carrier before modulation. **c** Spectral response of the carrier after modulation. The transmission bandwidth is $2(f_b + \Delta f_c)$. f_b is the bit rate and Δf_c is the frequency deviation = $1/f_b$ is the bit duration for NRZ data

- ω = Nominal frequency of the carrier frequency, and
- $\Delta\omega$ = Frequency deviation.

The spectral response is depicted in Fig. 1.8. Notice that the carrier frequency after FSK modulation varies back and forth from the nominal frequency f_c by $\pm\,\Delta f_c$, where Δf_c is the frequency deviation. The FSK bandwidth is given by

$$\text{BW} = 2(f_b + \Delta f_c) = 2f_b(1 + \Delta f_c/f_b) = 2f_b(1 + \beta),$$

where $\beta = \Delta f/f_b$ is known as the modulation index and f_b is the coded bit frequency (bit rate R_b).

1.4.5 BPSK Bandwidth at a Glance

In BPSK, the phase of the carrier changes in two discrete levels, in accordance with the input signal. Here, we have:

- Input data: $m(t) = 0$ or 1
- Carrier frequency: $C(t) = A \cos(\omega t)$
- Modulated carrier: $S(t) = A \cos(\omega + \varphi)t$

where

- A = Amplitude of the carrier frequency,
- ω = Angular frequency of the carrier, and
- φ = Phase of the carrier frequency.

Table below shows the number of phases and the corresponding bits per phase for MPSK modulation schemes for $M = 2, 4, 8, 16, 32, 64$, etc. It will be shown that higher-order MPSK modulation schemes ($M > 2$) are spectrally efficient.

Modulation	Number of phases φ	Number of bits per phase
BPSK	2	1
QPSK	4	2
8PSK	8	3
16	16	4
32	32	5
64	64	6
:	:	:

Figure 1.9 shows the spectral response of the BPSK modulator. Since there are two phases, the carrier frequency changes in two discrete levels, one bit per phase, as follows:

$\varphi = 0°$ for bit 0 and
$\varphi = 180°$ for bit 1.

Notice that the spectral response after BPSK modulation is the shifted version of the NRZ data, centered on the carrier frequency f_c. The transmission bandwidth is given by

$$\mathrm{BW(BPSK)} = 2R_b/\mathrm{Bit\,per\,Phase} = 2R_b/1 = 2R_b$$

where

- R_b is the coded bit rate (bit frequency).
- For BPSK, $\varphi = 2$, one bit per phase.

Also, notice that the BPSK bandwidth is the same as the one in ASK modulation. This is due to the fact that the phase of the carrier changes in two discrete levels, while the frequency remains the same.

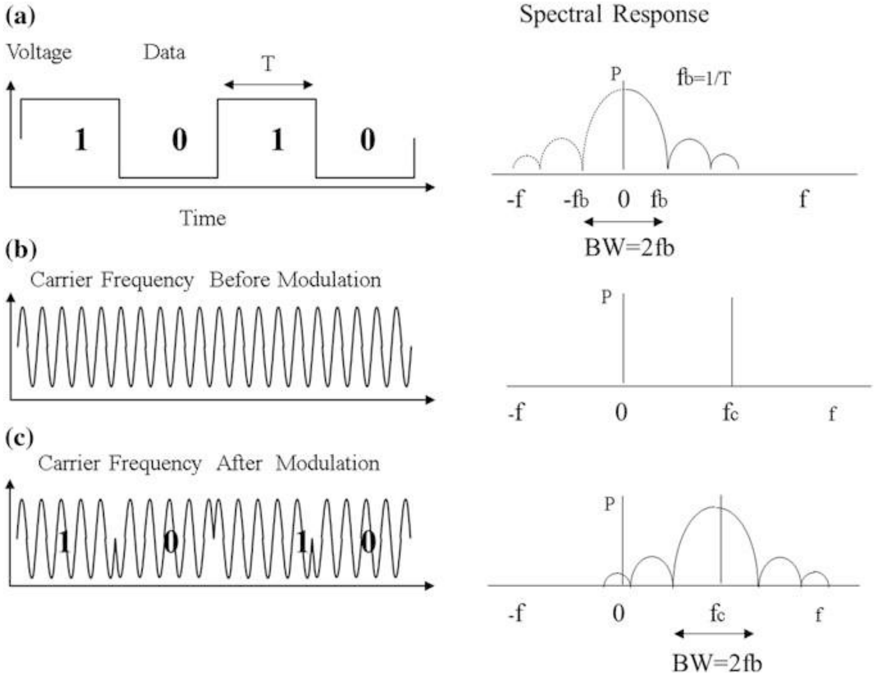
(a)
Voltage
Data
T
1
0
1
0
Time
(b)
Carrier Frequency Before Modulation
(c)
Carrier Frequency After Modulation
1
0
1
0
Spectral Response
P
fb=1/T
-f
-fb
0
fb
f
BW=2fb
P
-f
0
fc
f
P
-f
0
fc
f
BW=2fb

Fig. 1.9 PSK bandwidth at a glance. **a** Spectral response of NRZ data before modulation. **b** Spectral response of the carrier before modulation. **c** Spectral response of the carrier after modulation

1.5 Conclusions

This chapter presents a brief overview of modulation techniques covered in this book. Numerous illustrations are used to bring students up-to-date in key concepts and underlying principles of various analog and digital modulation techniques. In particular, following topics will be presented in this book:

- Amplitude Modulation (AM),
- Frequency Modulation (FM),
- Bandwidth occupancy in AM and FM,
- Amplitude shift keying (ASK),
- Frequency shift keying (FSK),
- Phase shift keying (PSK), and
- Bandwidth occupancy in ASK, FSK, and PSK.

References

1. Faruque S (2014) Radio frequency source coding made easy. Springer, New York
2. Faruque S (2014) Radio frequency channel coding made easy. Springer, New York
3. Marconi G (1987) Improvements in transmitting electrical impulses and signals, and in apparatus therefor. British patent No. 12,039 . Date of Application 2 June 1896; Complete Specification Left, 2 Mar 1897; Accepted, 2 July 1897 (later claimed by Oliver Lodge to contain his own ideas which he failed to patent)
4. Marconi G (1900) Improvements in apparatus for wireless telegraphy. British patent No. 7,777. Date of Application 26 Apr 1900; Complete Specification Left, 25 Feb 1901; Accepted, 13 Apr 1901
5. Armstrong EH (1936) A Method of reducing disturbances in radio signaling by a system of frequency modulation. Proc IRE (IRE) 24(5):689–740. doi:10.1109/JRPROC.1936.227383
6. Smith DR (1985) Digital transmission system. Van Nostrand Reinhold Co. ISBN: 0442009178
7. Leon W, Couch II (2001) Digital and analog communication systems, 7th edn. Prentice-Hall Inc, Englewood Cliffs. ISBN 0-13-142492-0
8. Sklar B (1988) Digital communications fundamentals and applications. Prentice Hall, Upper Saddle River
9. Joseph Fourier JB (1878) The analytical theory of heat (trans: Freeman A). The University Press, London

Chapter 2
Amplitude Modulation (AM)

Topics

- Introduction
- Amplitude Modulation (AM)
- AM Spectrum and Bandwidth
- Double Side Band Suppressed Carrier (DSBSC)
- DSBSC Spectrum and Bandwidth
- Single Side Band (SSB) Carrier
- SSB Spectrum and Bandwidth

2.1 Introduction

By inventing the wireless transmitter or radio in 1897, the Italian physicist Tomaso Guglielmo Marconi added a new dimension to the world of communications [1, 2]. This enabled the transmission of the human voice through space without wires. For this epoch-making invention, this illustrious scientist was honored with the Nobel Prize for Physics in 1909. Even today, students of wireless or radio technologies remember this distinguished physicist with reverence. A new era began in Radio Communications.

The classical Marconi radio used a modulation technique known today as "Amplitude Modulation" or just AM, which is the main topic of this chapter. In AM, the amplitude of the carrier changes in accordance with the input analog signal, while the frequency of the carrier remains the same. This is shown in Fig. 2.1, where

- $m(t)$ is the input modulating audio signal,
- $C(t)$ is the carrier frequency, and
- $S(t)$ is the AM-modulated carrier frequency.

© The Author(s) 2017

S. Faruque, *Radio Frequency Modulation Made Easy*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-41202-3_2

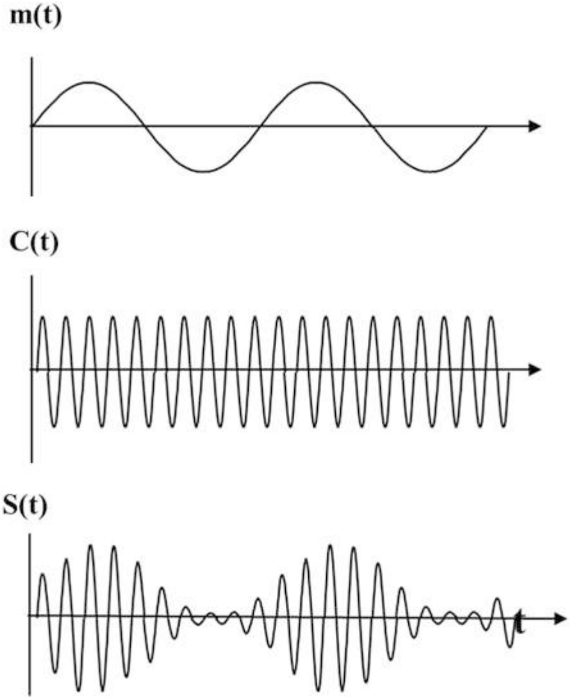

Fig. 2.1 AM waveforms. The amplitude of the carrier changes in accordance with the input analog signal. The frequency of the carrier remains the same

As shown in the figure, the audio waveform changes the amplitude of the carrier to determine the envelope of the modulated carrier. This enables the receiver to extract the audio signal by demodulation. Notice that the amplitude of the carrier changes in accordance with the input signal, while the frequency of the carrier does not change after modulation. However, it can be shown that the modulated carrier $S(t)$ contains several spectral components, requiring frequency domain analysis. In an effort to examine this, this chapter will present the following topics:

- Amplitude Modulation (AM) and AM spectrum
- Double Sideband-suppressed carrier (DSBSC) and DSBSC spectrum
- Single sideband (SSB) carrier and SSB spectrum

In the following sections, the above disciplines in AM modulation will be presented along with the respective spectrum and bandwidth. These materials have been augmented by diagrams and associated waveforms to make them easier for readers to grasp.

2.2 Amplitude Modulation

Amplitude Modulation (AM) is a method of analog modulation that utilizes amplitude variations of the relative amplitude of the career frequency [3–5]. The signal to be modulated and transmitted is analog. This is referred to as AM, where the amplitude of the carrier changes in accordance with the input signal.

Figure 2.2 shows a functional diagram of a typical AM modulator for a single tone. Here, $m(t)$ is the input analog signal we want to transmit, $C(t)$ is the carrier frequency without modulation, and $S(t)$ is the output AM-modulated carrier frequency. These parameters are described below:

$$\begin{aligned} &\bullet\quad m(t) = A_m \cos(2\pi f_m t) \\ &\bullet\quad C(t) = A_c \cos(2\pi f_c t) f_c \gg f_m \\ &\bullet\quad \begin{aligned} S(t) &= [1 + m(t)]C(t) \\ &= C(t) + m(t)C(t) \end{aligned} \end{aligned} \tag{2.1}$$

Therefore,
When $m(t) = 0$:

$$\bullet\quad S(t) = A_c \cos(2\pi f_c t) \tag{2.2}$$

When $m(t) = A_m \cos(2\pi f_m t)$:

$$\bullet\quad S(t) = A_c \cos(2\pi f_c t) + A_c A_m \cos(2\pi f_m t) \cos(2\pi f_c t) \tag{2.3}$$

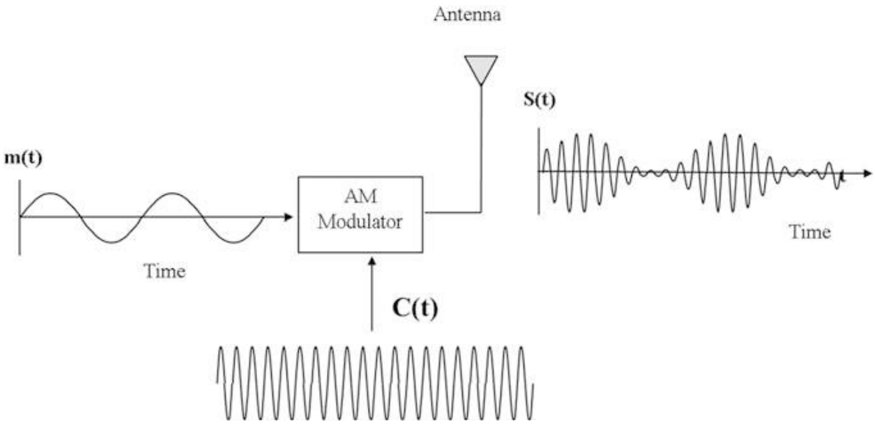

Fig. 2.2 Illustration of amplitude modulation. The amplitude of the carrier $C(t)$ changes in accordance with the input modulating signal $m(t)$. $S(t)$ is the modulated waveform which is transmitted by the antenna

In the above equation, we see that:

- The first term is the carrier only, which does not have information
- The second **term** contain**s** **the** information, which has several spectral components, requiring further analysis to quantify them.

2.3 AM Spectrum and Bandwidth

In wireless communications, the scarcity of RF spectrum is well known. For this reason, we have to be vigilant about using transmission bandwidth and modulation. The transmission bandwidth depends on the following:

- Spectral response of the input modulating signal
- Spectral response of the carrier frequency and
- Modulation type (AM, FM ASK, FSK, PSK, etc.)

Let us take a closer look!

2.3.1 Spectral Response of the Input Modulating Signal

In AM, the input modulating signal is a continuous time low-frequency analog signal. For simplicity, we use a sinusoidal waveform, which is periodic and continuous with respect to time. It has one frequency component. For example, the sine wave is described by the following time domain equation:

$$V(t) = V_p \sin(\omega_m t) \tag{2.4}$$

where

- Vp = Peak voltage
- $\omega_m = 2\pi f_m$
- f_m = Input modulating frequency in Hz

Figure 2.3 shows the characteristics of a sine wave and its spectral response. Since the frequency is constant, its spectral response is located in the horizontal axis at f_m, and the peak voltage is shown in the vertical axis. The corresponding bandwidth is zero.

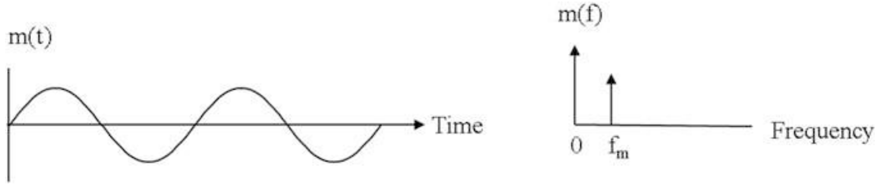

Fig. 2.3 A low-frequency sine wave and its frequency response

2.3.2 Spectral Response of the Carrier Frequency

A carrier frequency (f_c) is essentially a sinusoidal waveform, which is periodic and continuous with respect to time. It has one frequency component, which is much higher than the input modulating frequency ($f_c \gg f_m$). For example, the sine wave is described by the following time domain equation:

$$V(t) = V_p \sin(\omega_c t) \tag{2.5}$$

where

$$V_p = \text{Peak voltage}$$

- $\omega_c = 2\pi f_c$
- f_c = Carrier frequency in Hz

Figure 2.4 shows the characteristics of a sine wave and its spectral response. Since the frequency is constant, its spectral response is located in the horizontal axis at f_c and the peak voltage is shown in the vertical axis. The corresponding bandwidth is zero.

2.3.3 AM Spectrum and Bandwidth

Let us consider the AM signal again, which was derived earlier:

$$S(t) = A_c \cos(2\pi f_c t) + A_c A_m \cos(2\pi f_m t) \cos(2\pi f_c t) \tag{2.6}$$

Using the following trigonometric identity:

$$\cos A \cos B = 1/2\cos(A+B) + 1/2\cos(A-B) \tag{2.7}$$

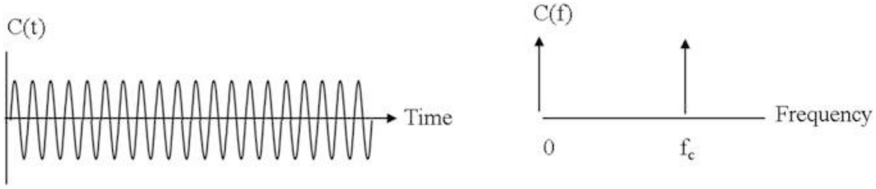

Fig. 2.4 A high-frequency sine wave and its frequency response

where

- $A = 2\pi\ (f_c + f_m)t$
- $B = 2\pi\ (f_c - f_m)t$

we get

$$S(t) = A_c \cos(2\pi f_c t) + (1/2)A_c A_m \cos[2\pi(f_c + f_m)t] + (1/2)A_c A_m \cos[2\pi(f_c - f_m)t] \tag{2.8}$$

This is the spectral response of the AM-modulated signal. It has three spectral components:

- The carrier: f_c
- Upper sideband: $f_c + f_m$
- Lower sideband: $f_c - f_m$

where f_c is the carrier frequency and f_m is the input modulating frequency. This is shown in Fig. 2.5. The AM bandwidth (BW) is given by

$$\mathrm{BW} = 2f_m \tag{2.9}$$

Notice that the power is distributed among the sidebands and the carrier, where the carrier does not contain any information. Only the sidebands contain the information. Therefore, AM is inefficient in power usage.

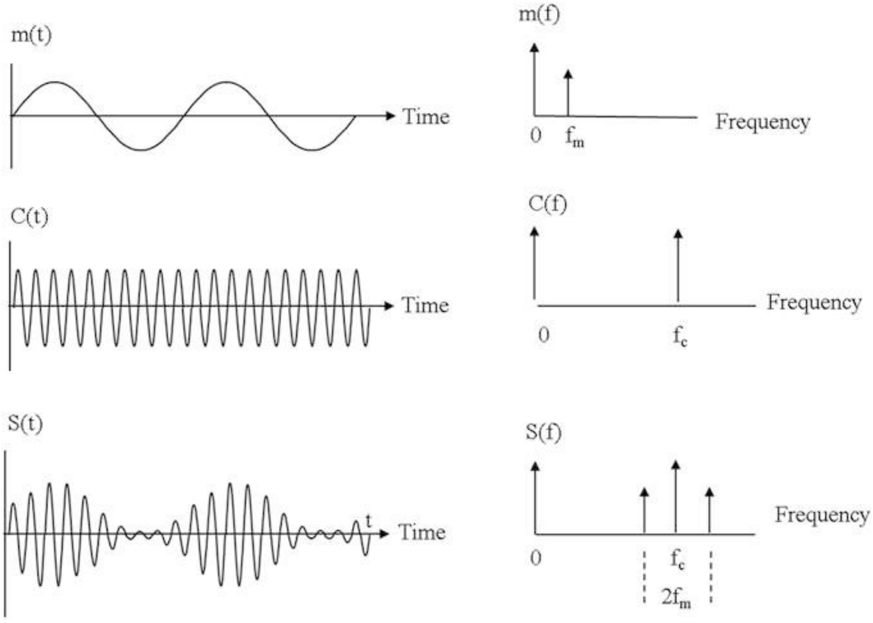
m(t)
Time
m(f)
0
f_m
Frequency
C(t)
Time
C(f)
0
f_c
Frequency
S(t)
t
Time
S(f)
0
f_c
2f_m
Frequency

Fig. 2.5 AM spectrum. The bandwidth is given by 2 f_m

2.3.4 AM Response Due to Low and High Modulating Signals

If $m(t)$ has a peak positive value of less than +1 and a peak negative value of higher than −1, then the modulation is less than 100 %. This is shown in Fig. 2.7.

On the other hand, if $m(t)$ has a peak positive value of +1 and a peak negative value of −1, then the modulation is 100 % [3–5]. Therefore,

- $$\text{For}\, m(t) = -1: \quad S(t) = A_c[1-1]\cos(2\pi_c t) = 0 \tag{2.10}$$
- $$\begin{aligned}\text{For}\, m(t) = +1: S(t) &= A_c[1+1]\cos(2\pi f_c t)\\ &= 2A_c\cos(2\pi f_c t)\end{aligned} \tag{2.11}$$

This is called 100 % modulation, as shown in Fig. 2.6. The percent modulation is described by the following equation:

$$\begin{aligned}&\text{The overall modulation percentage is:}\\ &\%\,\text{Overall Modulation} = \frac{A_{\max} - A_{\min}}{A_c} \times 100\\ &= \frac{\max[m(t)] - \min[m(t)]}{2A_c} \times 100.\end{aligned} \tag{2.12}$$

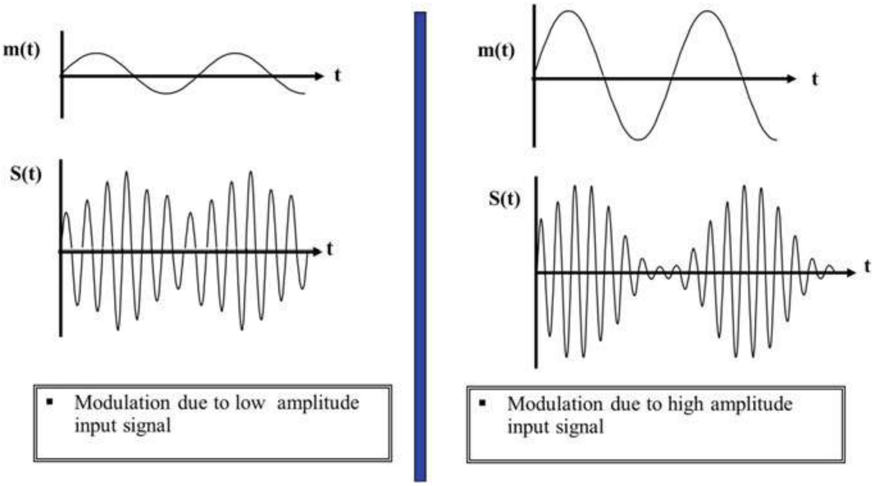

- Modulation due to low amplitude input signal

- Modulation due to high amplitude input signal

Fig. 2.6 Amplitude modulation due to low and high modulating signals

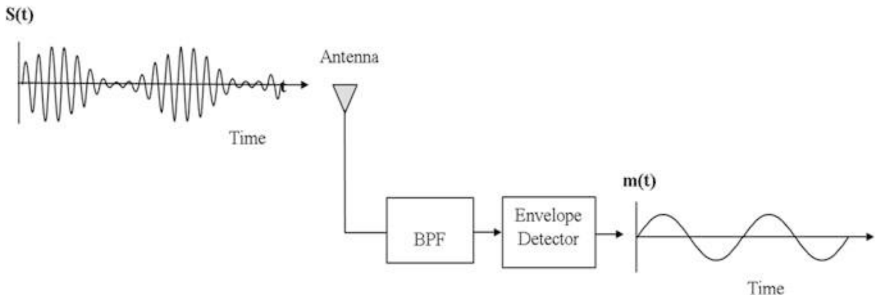

Fig. 2.7 AM demodulation technique. As the signal enters the receiver, it passes through the band-pass filter, which is tuned to the carrier frequency f_0. Next, the recovered signal is passed through an envelope detector to recover the original signal that was transmitted

2.3.5 AM Demodulation

Once the modulated analog signal has been transmitted, it needs to be received and demodulated. This is accomplished by the use of a band-pass filter that is tuned to the appropriate carrier frequency. Figure 2.7 shows the conceptual model of the AM receiver. As the signal enters the receiver, it passes through the band-pass filter, which is tuned to the carrier frequency f_0. Next, the recovered signal is passed through an envelope detector to recover the original signal that was transmitted.

2.3.6 Drawbacks in AM

- The modulated signal contains the carrier; carrier takes power and it does not have the information
- Therefore, AM is inefficient in power usage
- Moreover, there are two sidebands, containing the same information
- It is bandwidth inefficient
- AM is also susceptible to interference, since it affects the amplitude of the carrier.

Therefore, a solution is needed to improve bandwidth and power efficiency.

Problem 2.1 Given:

- Input modulating frequency f_m = 10 kHz
- Carrier frequency f_c = 400 kHz

Find

- Spectral components
- Bandwidth

Solution

Spectral Components:

- $f_c = 400$ kHz
- $f_c + f_m = 400$ kHz + 10 kHz = 410 kHz
- $f_c - f_m = 400$ kHz − 10 kHz = 390 kHz

Bandwidth

- BW = $2\,f_m = 2 \times 10$ kHz = 20 kHz.

2.4 Double Sideband-Suppressed Carrier (DSBSC)

2.4.1 DSBSC Modulation

Double sideband-suppressed carrier (DSBSC), also known as product modulator, is an AM signal that has a suppressed carrier [3–5]. Let us take the original AM signal once again, as given below:

$$\begin{aligned} S(t) &= A_c\cos(2\pi f_c t) + (1/2)A_cA_m\cos[2\pi(f_c + f_m)t] \\ &\quad + (1/2)A_cA_m\cos[2\pi(f_c - f_m)t] \end{aligned} \tag{2.13}$$

Notice that there are three spectral components:

- The first term is the carrier only, which does not have any information
- The second and third terms contain information.

In DSBSC, we suppress the carrier, which is the first term that does not have any information. Therefore, by suppressing the first term we obtain the following:

$$S(t) = (1/2)A_cA_m\cos[2\pi(f_c + f_m)t] + (1/2)A_cA_m\cos[2\pi(f_c - f_m)t] \tag{2.14}$$

Next, we use the following trigonometric identities:

- $\cos(A + B) = \cos A\,\cos B - \sin A\,\sin B$
- $\cos(A - B) = \cos A\,\cos B + \sin A\,\sin B$

With $A = 2\pi f_m t = \omega_m t$ and $B = 2\pi f_c t = \omega_c t$, we obtain:

$$S(t) = A_cA_m\cos(\omega_m t)\,\cos(\omega_c t) \tag{2.15}$$

Now, define

- $m(t) = A_m\cos(\omega_m t)$
- $C(t) = A_c\cos(\omega_c t)$

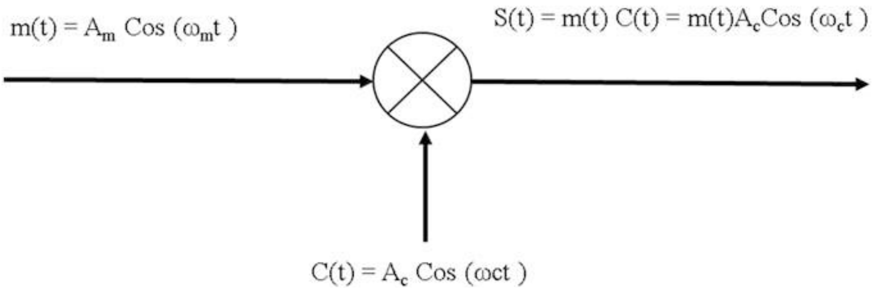
m(t) = A_m Cos (ω_m t)
S(t) = m(t) C(t) = m(t)A_c Cos (ω_c t)
C(t) = A_c Cos (ωct)

Fig. 2.8 Symbolic representation of DSBSC, also known as product modulator

Then, we can write the above equations as:

$$S(t) = m(t)\,C(t) \tag{2.16}$$

This is the DSBSC waveform. Since the output is the product of two signals, it is also known as product modulator. The symbolic representation is given in Fig. 2.8, where *m*(*t*) is the input modulating signal and *C*(*t*) is the carrier frequency.

2.4.2 Generation of DSBSC Signal

A DSBSC signal can be generated using two AM modulators arranged in a balanced configuration as shown in Fig. 2.9 [3–5]. The outcome is a cancellation of the discrete carrier. Also, the output is the product of two inputs: $S(t) = m(t)\,C(t)$. This is why it is called “product modulator.”

Proof of DSBSC

Consider the DSBSC modulator as shown in Fig. 2.9. Here, the AM modulators generate $S_1(t)$ and $S_2(t)$, which are given by:

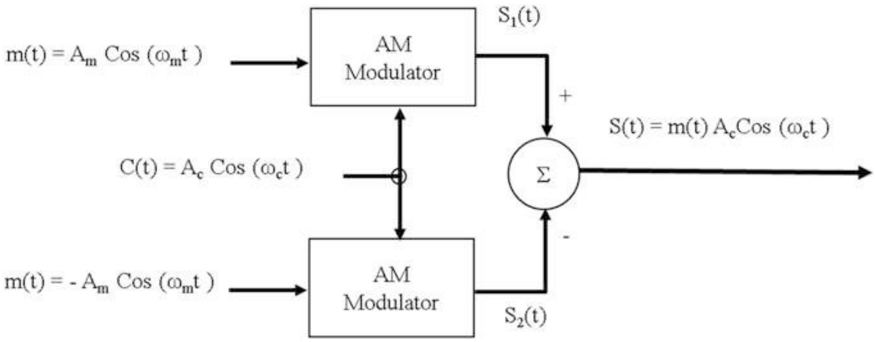
m(t) = A_m Cos (ω_m t)
AM Modulator
S_1(t)
+
S(t) = m(t) A_c Cos (ω_c t)
C(t) = A_c Cos (ω_c t)
Σ
-
m(t) = - A_m Cos (ω_m t)
AM Modulator
S_2(t)

Fig. 2.9 Construction of DSBSC modulator. The output is the product of two signals

$$S_1(t) = A_c[1 + m(t)]\cos(\omega_c t) \tag{2.17}$$

$$S_2(t) = A_c[1 - m(t)]\cos(\omega_c t)$$

Subtracting $S_2(t)$ from $S_1(t)$, we essentially cancel the carrier to obtain:

$$\begin{aligned} S(t) &= S_1(t) - S_2(t) \\ &= 2\,m(t)\,A_c\,\cos(\omega_c t) \end{aligned} \tag{2.18}$$

Therefore, except for the scaling factor 2, the above equation is exactly the same as the desired DSBSC waveform shown earlier, which does not have the carrier. In other words, the carrier has been suppressed, hence the name double sideband-suppressed carrier (DSBSC).

2.4.3 DSBSC Spectrum and Bandwidth

We begin with the DSBSC-modulated signal:

$$S(t) = 2\,m(t)\,A_c\cos(\omega_c t) \tag{2.19}$$

where

$$m(t) = A_m\cos(\omega_m t)$$

Therefore,

$$S(t) = 2\,A_c A_m\cos(\omega_m t)\ \cos(\omega_c t) \tag{2.20}$$

This is the desired DSBSC waveform for spectral analysis.
Now, use the trigonometric identity:

- $\cos A\ \cos B = 1/2\ \cos(A + B) + 1/2\ \cos(A - B)$ (2.21)

where

$$A = \omega_m t = 2\pi f_m t \text{ and } B = \omega_c t = 2\pi f_c t$$

Therefore,

$$\begin{aligned} S(t) &= A_c[\cos(\omega_c + \omega_m)t + \cos(\omega_c - \omega_m)t] \\ &= A_c[\cos\ 2\pi(f_c + f_m)t + \cos\ 2\pi(f_c - f_m)t] \end{aligned} \tag{2.22}$$

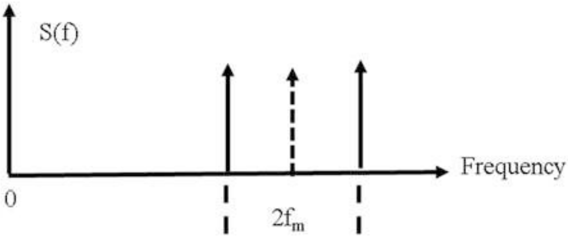

Fig. 2.10 DSBSC spectrum where the carrier frequency is suppressed. The bandwidth is given by 2 f_m

Notice that the carrier power is distributed among the sidebands (Fig. 2.10). Therefore, it is more efficient. The bandwidth is given by:

$$\mathrm{BW} = 2f_m \tag{2.23}$$

2.4.4 DSBSC Drawbacks

- There are two identical sidebands.
- Each sideband contains the same information
- Bandwidth is 2 f_m
- Unnecessary power usage

Therefore, a solution is needed to improve bandwidth efficiency.

Problem 2.2 Given:Two product modulators using identical carriers are connected in a series, as shown below:

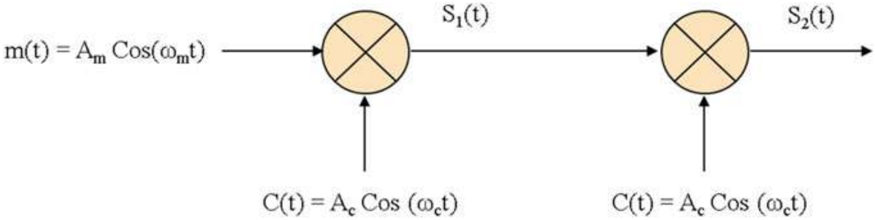
$m(t) = A_m \cos(\omega_m t)$
$S_1(t)$
$S_2(t)$
$C(t) = A_c \cos(\omega_c t)$
$C(t) = A_c \cos(\omega_c t)$

Find

(a) The output waveform $S_2(t)$
(b) What is the function of this circuit?

Solution

(a)
$$S_1(t) = m(t)\,C(t)$$
$$S_2(t) = S_1(t)\,C(t) = m(t)C^2(t) = A_m \cos(\omega_m t)\,Ac^2 \cos\,2(\omega_c t)$$

(b) The function of the circuit is to demodulate DSBSC signals, where the carrier frequency is filtered out.

2.5 Single Sideband (SSB) Modulation

2.5.1 *Why SSB Modulation?*

- The basic AM has a carrier which does not carry information—*Inefficient power usage*
- The basic AM has two sidebands contain the same information—*Additional loss of power*
- DSBSC has two sidebands, containing the same information—*Loss of power*
- Therefore, the basic AM and DSBSC are bandwidth and power inefficient
- SSB is bandwidth and power efficient.

2.5.2 *Generation of SSB-Modulated Signal*

Single sideband (SSB) modulation uses two product modulators as shown in Fig. 2.11 [3–5], where

$$\bullet \quad m(t) = A_m \cos(\omega_m t) \tag{2.24}$$

$$\bullet \quad m(t)^* = A_m \sin(\omega_m t) - (\text{Hilbert Transform}) \tag{2.25}$$

$$\bullet \quad C(t) = A_m \cos\,(\omega_c t) \tag{2.26}$$

$$\bullet \quad C(t)^* = A_c \sin(\omega_c t) - (\text{Hilbert Transform}) \tag{2.27}$$

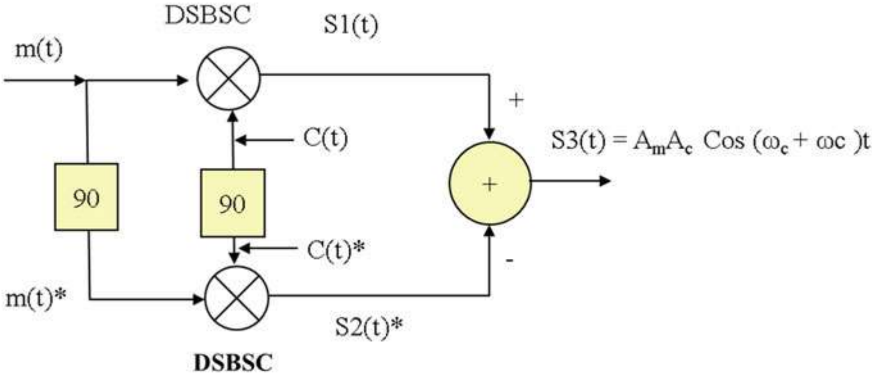

Fig. 2.11 Generation of SSB signal

Solving for S_1, S_2, and S_3, we obtain:

- $S_1(t) = A_c A_m \cos(\omega_m t)\cos(\omega_c t)$ (2.28)
- $S_2(t) = A_c A_m \sin(\omega_m t)\sin(\omega_c t)$ (2.29)

$$\begin{aligned} S_3(t) &= S_1(t) - S_2(t) \\ &= A_c A_m \cos(\omega_m t)\cos(\omega_c t) - A_c A_m \sin(\omega_m t)\sin(\omega_c t) \end{aligned} \tag{2.30}$$

Using the following formula:

- $\cos A \cos B = 1/2\cos(A + B) + 1/2\cos(A - B)$
- $\sin A \sin B = 1/2\cos(A - B) - 1/2\cos(A + B)$

Solving for S_3, we get:

$$S_3 = A_c A_m \cos(\omega_c + \omega_m)t \tag{2.31}$$

In the above equation, $S_3(t)$ is the desired SSB signal, which is the upper sideband only.

2.5.3 SSB Spectrum and Bandwidth

Let us consider the SSB signal again:

$$\begin{aligned} S_3 &= A_c A_m \cos(\omega_c + \omega_m)t \\ &= A_c A_m \cos\ 2\pi(f_c + f_m)t \end{aligned} \tag{2.32}$$

Here, we see that the SSB spectrum contains only one sideband. Therefore, it is more efficient. The SSB bandwidth is given by:

$$\text{SSB BW} = f_m \tag{2.33}$$

Figure 2.12 displays the SSB spectrum.

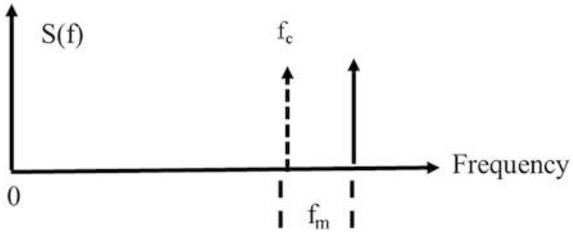

Fig. 2.12 SSB spectrum showing the upper sideband. The SSB bandwidth is f_m

Problem 2.3

Given:

- $m(t) = A_m \cos(\omega_m t)$
- $m(t)^* = A_m \sin(\omega_m t)$ − (Hilbert Transform)
- $C(t) = A_c \cos(\omega_c t)$
- $C(t)^* = A_c \sin(\omega_c t)$ − (Hilbert Transform)

Design an SSB modulator to realize the lower sideband. Sketch the spectral response.

Solution

Solving for S_1 and S_2, we obtain:

- $S_1(t) = A_c A_m \cos(\omega_m t) \cos(\omega_c t)$
- $S_2(t) = A_c A_m \sin(\omega_m t) \sin(\omega_c t)$

Obtain S_3 as:

$$\begin{aligned} S_3(t) &= S_1(t) + S_2(t) \\ &= A_c A_m \cos(\omega_m t) \cos(\omega_c t) + A_c A_m \sin(\omega_m t) \sin(\omega_c t) \end{aligned}$$

Using the following formula:

- $\cos A \cos B = 1/2\cos(A + B) + 1/2\cos(A - B)$
- $\sin A \sin B = 1/2\cos(A - B) - 1/2\cos(A + B)$

Solving for S_3, we get:

$$\begin{aligned} S_3 &= A_c A_m \cos(\omega_m - \omega_c)t \\ &= A_c A_m \cos 2\pi(f_c - f_m)t \end{aligned}$$

In the above equation, S_3(t) is the desired SSB signal, which is the lower sideband only. The spectral response, showing the lower sideband, is presented below.

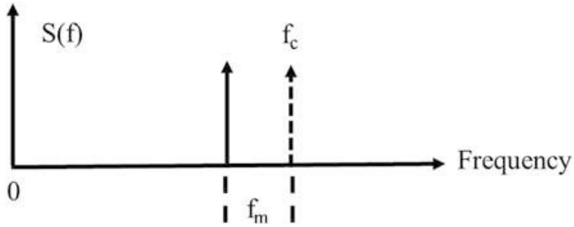

2.6 Conclusions

This chapter presents the key concepts and underlying principles of Amplitude Modulation. It was shown how the audio waveform changes the amplitude of the carrier to determine the envelope of the modulated carrier. It was also shown that the modulated carrier contains several spectral components that lead to DSBSC and SSB modulation techniques. In particular, the following topics were presented in this chapter:

- Amplitude Modulation (AM)
- AM spectrum and bandwidth
- Double sideband-suppressed carrier (DSBSC)
- DSBSC spectrum and bandwidth
- Single sideband (SSB)
- SSB spectrum and bandwidth

These materials have been augmented by diagrams and associated waveforms to make them easier for readers to grasp.

References

1. Marconi G (1897) Improvements in transmitting electrical impulses and signals, and in apparatus therefor. British patent No. 12,039. Date of Application 2 June 1896; Complete Specification Left, 2 Mar 1897; Accepted, 2 July 1897 (later claimed by Oliver Lodge to contain his own ideas which he failed to patent)
2. Marconi G (1900) Improvements in apparatus for wireless telegraphy. British patent No. 7,777. Date of Application 26 Apr 1900; Complete Specification Left, 25 Feb 1901; Accepted, 13 Apr 1901
3. Leon W, Couch II (2001) Digital and analog communication systems, 7th edn. Prentice-Hall Inc, Englewood Cliffs. ISBN 0-13-142492-0
4. Godse AP, Bakshi UA (2009) Communication engineering. Technical Publications, p 36. ISBN 978-81-8431-089-4
5. Silver W (ed) (2011) Chapter 14 transceivers. The ARRL handbook for radio communications, 88th edn. American Radio Relay League. ISBN 978-0-87259-096-0

Chapter 3
Frequency Modulation (FM)

Topics

- Introduction
- Frequency Modulation (FM)
- FM Spectrum
- Carson's Rule & FM Bandwidth
- Bessel Function & FM Bandwidth
- FM bandwidth Dilemma

3.1 Introduction

In Frequency Modulation (FM), the frequency of the carrier changes in accordance with the input analog signal, while the amplitude of the carrier remains the same [1–5]. This is shown in Fig. 3.1, where

- $m(t)$ is the input modulating audio signal,
- $C(t)$ is the carrier frequency, and
- $S(t)$ is the FM-modulated carrier frequency.

As shown in the figure, the audio waveform changes the frequency of the carrier. This enables the receiver to extract the audio signal by demodulation. Notice that the frequency of the carrier changes in accordance with the input signal, while the amplitude of the carrier does not change after modulation. However, it can be shown that the modulated carrier $S(t)$ contains an infinite number of spectral components, requiring frequency domain analysis [3]. In an effort to examine this, this chapter will present the following topics:

- The basic Frequency Modulation (FM),
- FM spectrum, and
- FM bandwidth.

© The Author(s) 2017

S. Faruque, *Radio Frequency Modulation Made Easy*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-41202-3_3

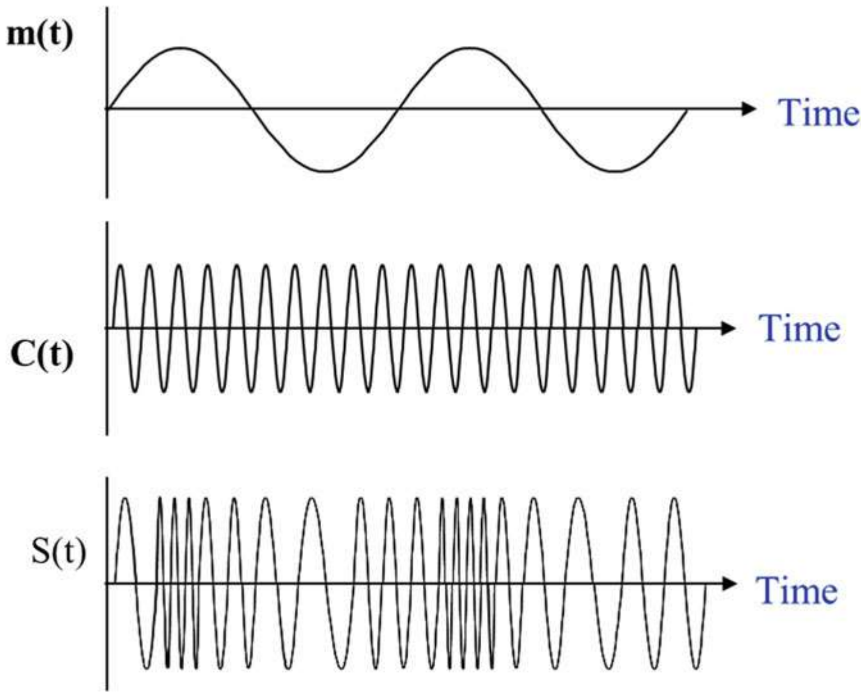

Fig. 3.1 Illustration of FM

3.2 Frequency Modulation (FM)

3.2.1 Background

FM is a form of angle modulation, where the frequency of the carrier varies in accordance with the input signal. Here, the angle refers to the angular frequency (ω). The angular frequency ω is also recognized as angular speed or circular frequency. It is a measure of rotation rate or the rate of change of the phase of a sinusoidal waveform as illustrated in Fig. 3.2.

Magnitude of the angular frequency ω is defined by one revolution or 2π radians:

$$\omega = 2\pi f = \mathrm{d}\theta/\mathrm{d}t \quad \text{Radians per second} \tag{3.1}$$

where,

- ω = Angular frequency in radians per seconds,
- f = Frequency in Hertz (Hz) or cycles per second, and
- θ = Phase angle.

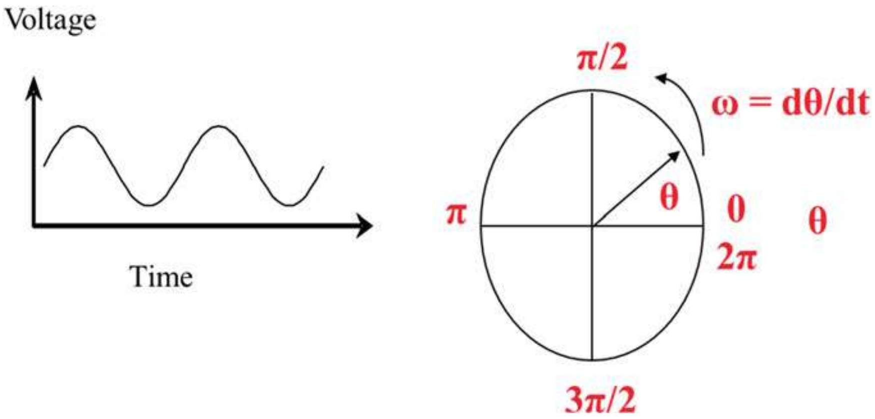

Fig. 3.2 A sinusoidal waveform in the time domain and its representation in the phase domain

Notice that in Eq. 3.1, the angular frequency ω is greater than the frequency f by a factor of 2π. Now solving for the phase angle, we obtain,

$$\theta_i(t) = 2\pi \int_0^t f_i dt \tag{3.2}$$

- θ_i = Instantaneous phase angle and
- f_i = Instantaneous frequency.

This forms the basis of our derivation of FM as presented in the following section.

3.2.2 The Basic FM

Frequency Modulation (FM) is a method of analog modulation that utilizes frequency variation of the relative frequency of the career [1]. The signal to be modulated and transmitted is analog. This is referred to as FM, where the frequency of the carrier changes in accordance with the input signal. The modulated carrier frequency f_c varies back and forth and depends on amplitude A_m and frequency f_m of the input signal.

Figure 3.3 shows the functional diagram of a typical FM, using a single-tone modulating signal. Here, $m(t)$ is the input analog signal we want to transmit, $C(t)$ is the carrier frequency without modulation, and $S(t)$ is the output FM-modulated carrier frequency. These parameters are described below.

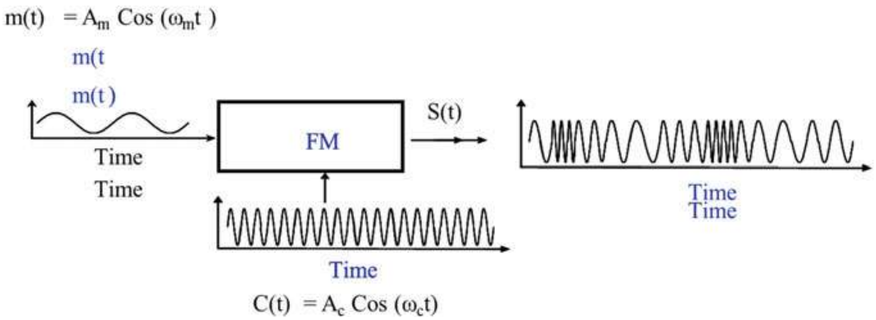

Fig. 3.3 A functional diagram of a typical FM modulator using a single-tone input modulating signal. Here, $m(t)$ is the input analog signal we want to transmit, $C(t)$ is the carrier frequency without modulation, and $S(t)$ is the output FM-modulated carrier frequency

We examine this by means of a single-tone input modulating signal and its angular frequency as shown in Fig. 3.2. Since the frequency of the carrier varies in accordance with the input signal, the instantaneous frequency of the carrier is given by

$$f_i(t) = f_c(t) + D_f m(t) \tag{3.3}$$

where,

- f_i = Instantaneous frequency,
- f_c = Carrier frequency,
- D_f = Constant, and
- $m(t) = A_m \cos(wmt)$.

The FM-modulated signal is given by

$$S(t) = A_c \cos(\theta_i) \tag{3.4}$$

where Ac is the amplitude of the carrier frequency and θ_i is the instantaneous angle. Substituting Eq. (3.2) into Eq. (3.4) for θi, we get

$$S(t) = A_c \cos[2\pi \int_0^t f_i dt] \tag{3.5}$$

where f_i is the instantaneous frequency. Substituting Eq. (3.3) into Eq. (3.5) for fi, we obtain

$$S(t) = A_c \cos[2\pi \int_0^t \{fc(t) + Df\, m(t)\} dt] \tag{3.6}$$

where $m(t) = A_m \cos(\omega_m t)$ is the input modulating signal. Integrating the above equation, we obtain the desired FM signal as follows:

$$\begin{aligned} S(t) &= A_c \cos\left[2\pi f_c t + \left(\frac{\Delta f}{f_m}\right)\sin(2\pi f_m t)\right] \\ &= A_c \cos[2\pi f_c t + \beta \sin(2\pi f_m t)] \qquad (3.7) \\ \beta &= \left(\frac{\Delta f}{f_m}\right) = \text{Modulation Index} \quad \Delta f = DfAm = \text{Freq. Deviation} \end{aligned}$$

where,

- $S(t)$ = FM-modulated carrier signal,
- f_c = Frequency of the carrier,
- A_c = Amplitude of the carrier frequency,
- $\Delta f = D_f A_m$ = Frequency deviation,
- f_m = Input modulating frequency,
- Am = Amplitude of the input modulating signal,
- D_f = A constant parameter, and
- $\beta = \Delta f / f_m$ = Modulation index.

Note that the modulation index β is an important design parameter in FM. It is directly related to FM bandwidth. It may also be noted that FM bandwidth depends on both frequency and amplitude of the input modulating signal. Let us take a closer look.

3.3 FM Spectrum and Bandwidth

In wireless communications, the scarcity of RF spectrum is well known. For this reason, we have to be vigilant about using transmission bandwidth and modulation. The transmission bandwidth depends on the following:

- Spectral response of the input modulating signal,
- Spectral response of the carrier frequency, and
- Modulation type.

Let's take a closer look!

3.3.1 Spectral Response of the Input Modulating Signal

In FM, the input modulating signal is a continuous time low-frequency analog signal. For simplicity, we use a sinusoidal waveform, which is periodic and

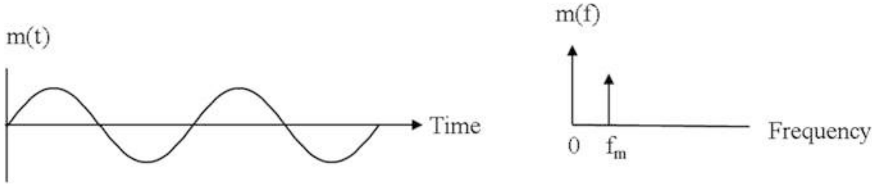

Fig. 3.4 A low-frequency sine wave and its frequency response

continuous with respect to time. It has one frequency component. For example, the sine wave is described by the following time domain equation:

$$V(t) = V_p \sin(\omega_m t) \tag{3.8}$$

where,

- Vp = Peak voltage,
- $\omega_m = 2\pi f_m$, and
- f_m = Input modulating frequency in Hz.

Figure 3.4 shows the characteristics of a sine wave and its spectral response. Since the frequency is constant, its spectral response is located in the horizontal axis at f_m and the peak voltage is shown in the vertical axis. The corresponding bandwidth is zero.

3.3.2 Spectral Response of the Carrier Frequency

A carrier frequency (f_c) is essentially a sinusoidal waveform, which is periodic and continuous with respect to time. It has one frequency component, which is much higher than the input modulating frequency ($f_c \gg f_m$). For example, the sine wave is described by the following time domain equation:

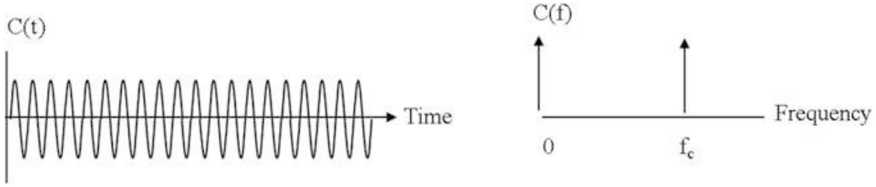

Fig. 3.5 A high-frequency sine wave and its frequency response before modulation

$$V(t) = V_p \sin(\omega_c t) \tag{3.9}$$

where

$$V_p = \text{Peak voltage}$$

- $\omega_c = 2\pi f_c$ *and*
- f_c = Carrier frequency in Hz.

Figure 3.5 shows the characteristics of a high-frequency sine wave and its spectral response. Since the frequency is constant, its spectral response is located in the horizontal axis at f_c and the peak voltage is shown in the vertical axis. The corresponding bandwidth is zero.

3.3.3 FM Spectrum

In FM, the frequency of the carrier changes in accordance with the input signal. Here, we have:

$$\begin{array}{lll} \bullet & \text{Input Signal :} & m(t) = A_m \cos(\omega_m t) \\ \bullet & \text{Carrier Frequency :} & C(t) = A_c \cos(\omega_c t) \\ \bullet & \text{Modulated Carrier :} & S(t) = A_c \cos[(\omega_c t) + \beta \sin(\omega_m t)] \end{array} \tag{3.10}$$

where

- $S(t)$ = The modulated carrier,
- Ac = Frequency of the carrier,
- ω_c = Nominal frequency of the carrier frequency,
- $\beta = \Delta f/fm$ = Modulation index,
- $\Delta f = D_f A_m$ = Frequency deviation,
- D_f = Constant, and
- A_m = Amplitude of the input modulating signal.

By inspecting the modulated carrier frequency, we observe that $S(t)$ depends on both frequency and amplitude of the input signal. The spectral response is given in Fig. 3.6. Notice that the carrier frequency after modulation varies back and forth from the nominal frequency f_c as depicted in the figure. In Fig. 3.6, we see that:

- As time passes, the carrier moves back and forth in frequency in exact step with the input signal.
- Frequency deviation is proportional to the input signal voltage.
- A group of many sidebands is created, spaced from carrier by amounts $N \times f_i$.
- Relative strength of each sideband depends on Bessel function.
- Strength of individual sidebands far away from the carrier is proportional to (freq. deviation × input frequency).
- Higher order spectral components are negligible.

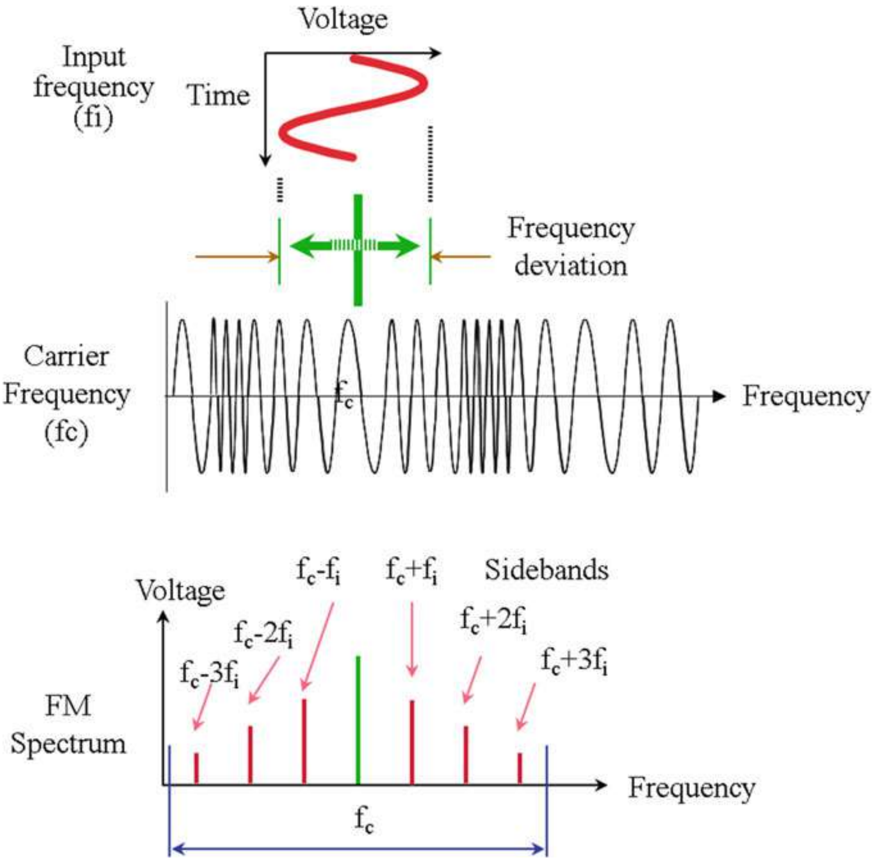
Voltage
Input frequency (fi)
Time
Frequency deviation
Carrier Frequency (fc)
f_c
Frequency
Voltage
f_c-f_i
f_c+f_i
Sidebands
f_c-2f_i
f_c+2f_i
f_c-3f_i
f_c+3f_i
FM Spectrum
Frequency
f_c

Fig. 3.6 FM spectrum. As time passes, the carrier moves back and forth in frequency in exact step with the input signal and generates an infinite number of sidebands

- Carson's rule can be used to determine the approximate bandwidth: bandwidth required = 2 × (highest input frequency + frequency deviation).

3.3.4 Carson's Rule and FM Bandwidth

The Carson's rule, a rule of thumb, states that more than 98 % of the power of FM signal lies within a bandwidth given by the following approximation:

$$\text{FM Bandwidth (BW)} = 2f_m(1 + \beta) \tag{3.11}$$

where

Table 3.1 Bessel function table

β	J0	J1	J2	J3	J4	J5
0	1					
0.25	0.98	0.12				
0.5	0.94	0.24	0.03			
1	0.77	0.44	0.11	0.02		
1.5	0.51	0.56	0.23	0.06	0.01	
2	0.22	0.58	0.35	0.13	0.03	

- $\beta = \Delta f/fm$ = Modulation index,
- Δf = Peak deviation of the instantaneous frequency from the center of the carrier frequency, and
- fm = Highest frequency of the modulating signal.

3.3.5 Bessel Function and FM Bandwidth

FM bandwidth can be estimated by means of Bessel function of the first kind. For a single-tone modulation, it can be obtained as a function of the sideband number and the modulation index.

For a given β, the solution for $S(t)$ is given as follows:

$$\begin{aligned} S(t) &= A_c\cos[(\omega_c t) + \beta\sin(\omega_m t)] \\ &= J_0(\beta)\cos(\omega_c t) \\ &\quad + J_1(\beta)\cos(\omega_c t + \omega_m t) + J_2(\beta)\cos(\omega_c t + 2\omega_m t) + J_3(\beta)\cos(\omega_c t + 3\omega_m t) + \ldots \\ &\quad - J_1(\beta)\cos(\omega_c t - \omega_m t) - J_2(\beta)\cos(\omega_c t - 2\omega_m t) - J_3(\beta)\cos(\omega_c t - 3\omega_m t) + \ldots \end{aligned} \tag{3.12}$$

Here, J's are the Bessel functions, representing the amplitude of the sidebands. $J_0(\beta)$ is the amplitude of the fundamental spectral component, and the remaining spectral components are the sidebands. Each sideband is separated by the input modulating frequency. These values are also available as a standard Bessel function table.

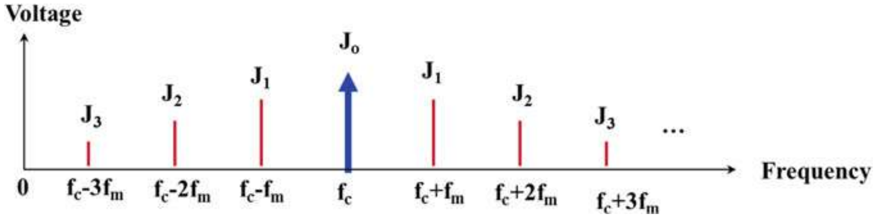

Fig. 3.7 FM sidebands. Power is taken from the carrier $J0$ and distributed among the sidebands $J1$, $J2$, $J3$,…, etc. Each sideband is separated by the modulating frequency f_m

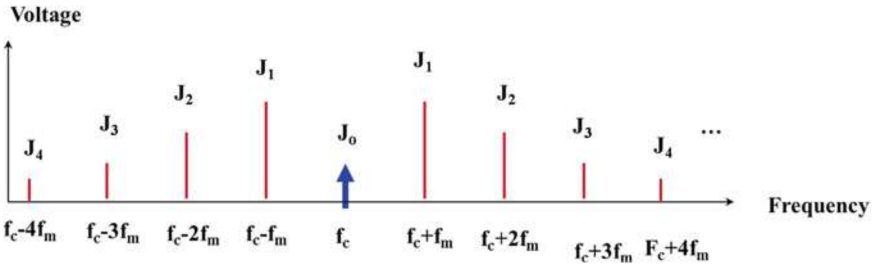

Fig. 3.8 FM sidebands for large β. More power is taken from the carrier $J0$ and distributed among the sidebands $J1$, $J2$, $J3$, $J4$,... Each sideband is separated by the modulating frequency f_m

As an example, Table 3.1 provides a few Bessel parameters to illustrate the concept. In this table, the carrier and sideband amplitude powers, $J0$, $J1$, $J2$,..., etc., are presented for different values of β. Here, $J0$ is carrier power before modulation. After modulation, with a given value of β, power is taken from the carrier and distributed among the sidebands. Also, there is a unique value of β, for which the carrier amplitude becomes zero and all the signal power is in the sidebands.

Note that, in FM, the sidebands are on both sides of the carrier. Therefore, the total bandwidth includes spectral components from both sides of the carrier. For a given low β, this is shown in Fig. 3.7. Here, J0 is the spectral component of the carrier. $J1$, $J2$, $J3$, etc., are the sidebands. Each sideband is separated by the input modulating frequency f_m. After modulation, power is taken from the carrier J0 and distributed among the sidebands, depending on the modulation index β.

Figure 3.8 shows another scenario where β is large. In this case, more power is taken from the carrier and distributed among the sideband, while creating more significant sidebands, requiring more bandwidth.

3.3.6 FM Bandwidth Dilemma

In FM, we notice that if the modulation index is low, the occupied bandwidth is low and the sidebands take less power from the carrier, making the modulation less efficient. On the other hand, if the modulation index is high, the occupied bandwidth is also high, while the sidebands retain most of the power, making the modulation more efficient at the expense of bandwidth. This is a dilemma in FM. However, FM is more popular because it is less sensitive to noise.

Problem 3.1
Given:

$$f_m = 1\,\text{kHz}$$
$$f_c = 1\,\text{MHz}$$
$$\beta = 1$$

Find:

(a) The FM bandwidth using Carson’s rule and
(b) The FM bandwidth using Bessel function

Solution:

(a) FM BW (Carson’s rule) = $2 f_m (1 + \beta) = 2 \times 1$ kHz $(1 + 1) = 4$ kHz.
(b) For $\beta = 1$, See Table 3.1:

- $J0 = 0.77$, $J1 = 0.44$, $J2 = 0.11$, $J3 = 0.02$

where $J0$ is the carrier and $J1$, $J2$, and $J3$ are the sidebands. Here, we can neglect $J3$ since it has a negligible power. Therefore, there are two significant upper and lower sidebands, and each sideband is 1 kHz apart. Therefore, FM bandwidth is 2 × Number of significant sidebands = 2 × 2 = 4 kHz. See figure below.

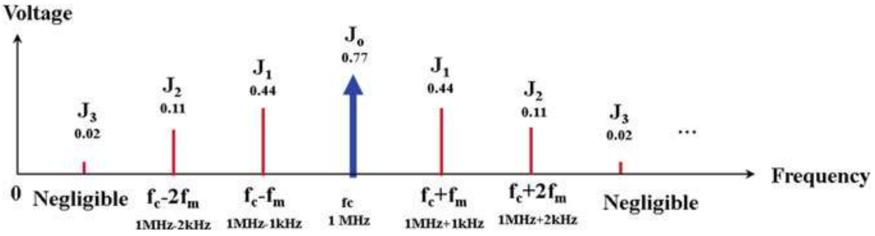

Problem 3.2
Given:

$$fm = 1\,\text{kHz}$$
$$fc = 1\,\text{MHz}$$
$$\beta = 2$$

Find:

(a) The FM bandwidth using Carson’s rule and
(b) The FM bandwidth using Bessel function

Solution:

(a) FM BW (Carson’s rule) = $2 f_m (1 + \beta) = 2 \times 1$ kHz $(1 + 2) = 6$ kHz.
(b) For $\beta = 2$, See Table 3.1:

- $J0 = 0.22$, $J1 = 0.58$, $J2 = 0.35$, $J3 = 0.13$, $J4 = 0.03$

Here, we have 4 upper and 4 lower sidebands, and each sideband is separated by 1 kHz. Since the J4 is negligible, we take 3 upper sidebands and 3 lower sidebands. Therefore, the FM bandwidth is 2 × 3 = 6 kHz. See figure below.

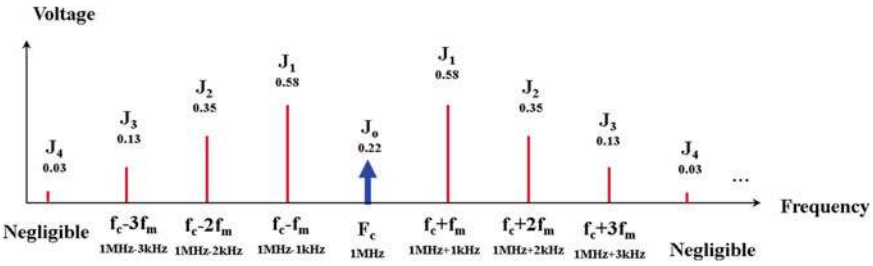

3.4 Conclusions

- This chapter presents a brief overview of Frequency Modulation (FM) and its attributes.
- The key concept and the underlying principle of FM are presented with numerous illustrations to make them easier for readers to grasp.
- FM spectrum is addressed with illustrations.
- FM bandwidth is estimated by means of Carson’s rule and Bessel function.
- FM bandwidth dilemma is explained, and problems are given to illustrate the concept.

References

1. Armstrong EH (1936) A method of reducing disturbances in radio signaling by a system of frequency modulation. Proc IRE (IRE) 24(5):689–740. doi:10.1109/JRPROC.1936.227383
2. Lathi BP (1968) Communication systems. Wiley, New York, pp 214–217. ISBN 0-471-51832-8
3. Leon W, Couch II (2001) Digital and analog communication systems, 7th edn. Prentice-Hall Inc, Englewood Cliffs. ISBN 0-13-142492-0
4. Haykin S (2001) Communication systems, 4th edn. Wiley, New York
5. Felix MO (1965) FM systems of exceptional bandwidth. Proc IEEE 112(9):1664

Chapter 4
Amplitude Shift Keying (ASK)

Topics

- Introduction
- Amplitude Shift Keying (ASK)
- ASK spectrum
- ASK Bandwidth
- Performance Analysis

4.1 Introduction

In amplitude shift keying (ASK), the amplitude of the carrier changes in discrete levels in accordance with the input digital signal, while the frequency of the carrier remains the same. This is shown in Fig. 4.1, where

- $m(t)$ is the input modulating digital signal,
- $C(t)$ is the carrier frequency, and
- $S(t)$ is the ASK-modulated carrier frequency.

As shown in the figure, the digital binary signal changes the amplitude of the carrier in two discrete levels. This enables the receiver to extract the digital signal by demodulation. Notice that the amplitude of the carrier changes in accordance with the input signal, while the frequency of the carrier does not change after modulation. However, it can be shown that the modulated carrier $S(t)$ contains several spectral components, requiring frequency domain analysis.

In the following sections, the above disciplines in ASK modulation will be presented along with the respective spectrum and bandwidth. These materials have been augmented by diagrams and associated waveforms to make them easier for readers to grasp.

© The Author(s) 2017

S. Faruque, *Radio Frequency Modulation Made Easy*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-41202-3_4

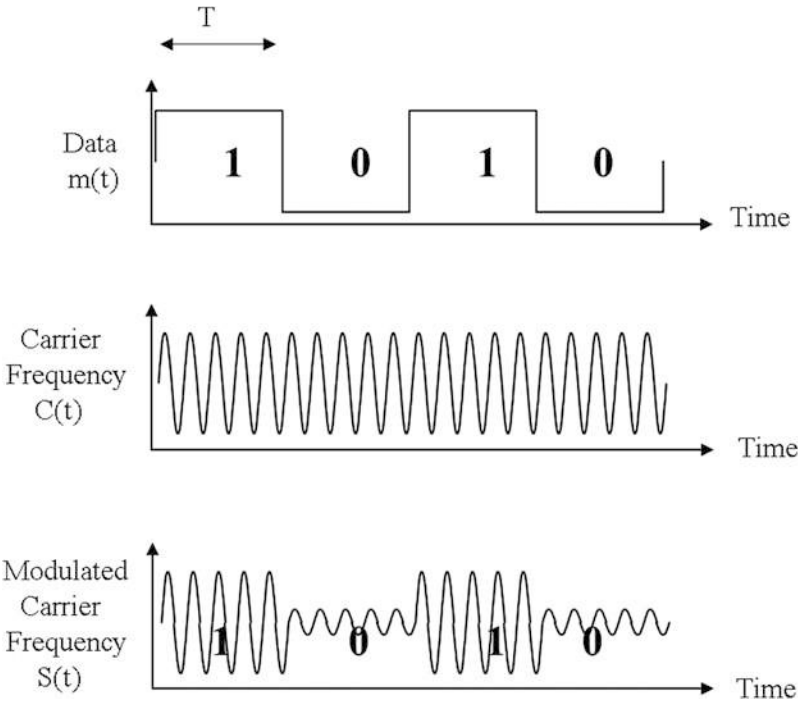

Fig. 4.1 ASK waveforms. The amplitude of the carrier changes in accordance with the input digital signal. The frequency of the carrier remains the same

4.2 ASK Modulation

Amplitude shift keying (ASK), also known as on–off keying (OOK), is a method of digital modulation that utilizes amplitude shifting of the relative amplitude of the career frequency [1–3]. The signal to be modulated and transmitted is binary, which is encoded before modulation. This is an indispensable task in digital communications, where redundant bits are added with the raw data that enables the receiver to detect and correct bit errors, if they occur during transmission [4–16].

While there are many error-coding schemes available, we will use a simple coding technique, known as "Block Coding" to illustrate the concept.

Figure 4.2 shows an encoded ASK modulation scheme using (15, 8) block code where an 8-bit data block is formed as M-rows and N-columns ($M = 4, N = 2$). The product $MN = k = 8$ is the dimension of the information bits before coding. Next, a horizontal parity P_H is appended to each row and a vertical parity P_V is appended to each column. The resulting augmented dimension is given by the product $(M + 1)(N + 1) = n = 15$, which is then ASK modulated and transmitted row by row. The rate of this coding scheme is given by

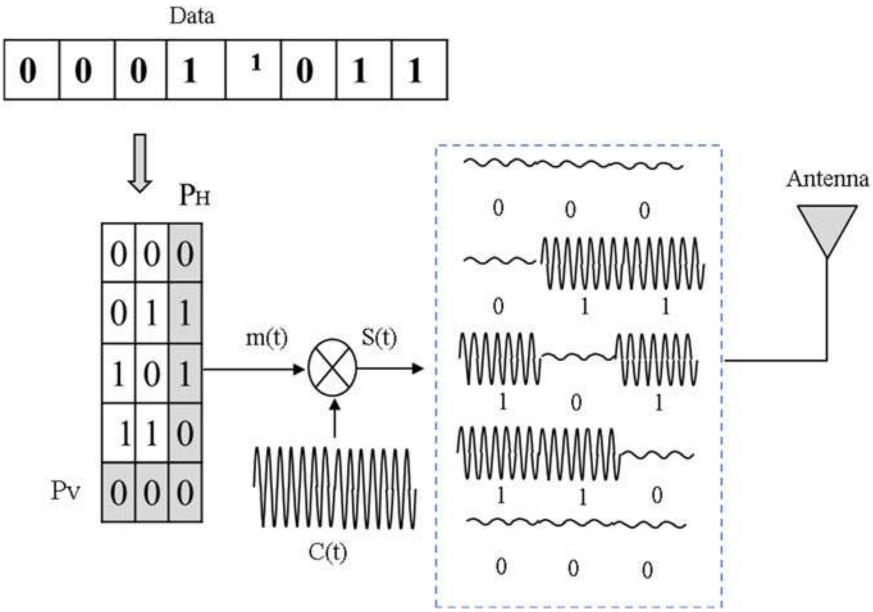

Fig. 4.2 Amplitude shift keying (ASK) is also known as on–off keying (OOK). The input encoded data block is transmitted row by row. The amplitude of the carrier frequency changes in accordance with the input digital signal

$$\text{Code Rate: } r = (MN)/[(M+1)(N+1)] = (4 \times 2)/(5 \times 3) = 8/15 \tag{4.1}$$

The coded bit rate R_{b2} is given by

$$R_{b2} = \text{Uncoded Bit Rate}/\text{Code Rate} = R_{b1}/r = R_{b1}\,(15/8) \tag{4.2}$$

Next, the coded bits are modulated by means of the ASK modulator as shown in the figure. Here,

- The Input digital signal is the encoded bit sequence we want to transmit
- Carrier is the radio frequency without modulation
- Output is the ASK-modulated carrier, which has two amplitudes corresponding to the binary input signal. For binary signal 1, the carrier is ON. For the binary signal 0, the carrier is OFF; however, a small residual signal may remain due to noise, interference, etc., as indicated in the figure.

As shown in the figure, the amplitude of the carrier changes in discrete levels, in accordance with the input signal, where

$$\begin{array}{lll} \bullet & \text{Input Data :} & m(t) = 0 \text{ or } 1 \text{ (coded data)} \\ \bullet & \text{Carrier Frequency :} & C(t) = A\cos(\omega t) \\ \bullet & \text{Modulated Carrier :} & S(t) = m(t)C(t) = m(t)A\cos(\omega t) \end{array} \tag{4.3}$$

Therefore,

$$\begin{array}{l} \text{For } m\,(t) = 1 :\ S(t) = A\cos(\omega t) \text{ i.e. the carrier is ON} \\ \text{For } m\,(t) = 1 :\ S(t) = 0, \text{i.e. the carrier is OFF} \end{array} \tag{4.4}$$

where A is the amplitude and ω is the frequency of the carrier.

4.3 ASK Demodulation

Once the modulated binary data has been transmitted, it needs to be received and demodulated. This is often accomplished with the use of a band-pass filter. In the case of ASK, the receiver needs to utilize one band-pass filter that is tuned to the appropriate carrier frequency. As the signal enters the receiver, it passes through the filter and a decision as to the value of each bit is made to recover the encoded data block, along with horizontal and vertical parities. Next, the receiver appends horizontal and vertical parities P_H* and P_V* to check parity failures and recovers the data block. This is shown in Fig. 4.3 having no errors. If there is an error, there will be a parity failure in P_H* and P_V* to pinpoint the error.

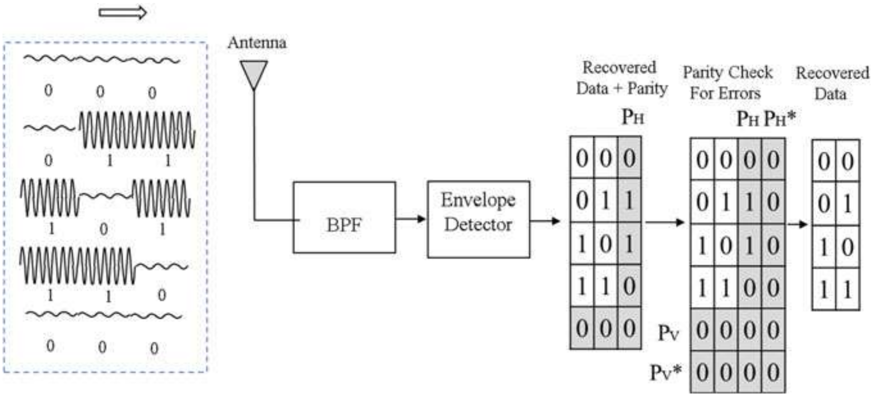
0 0 0
0 1 1
1 0 1
1 1 0
0 0 0
Antenna
BPF
Envelope Detector
Recovered Data + Parity
PH
0 0 0
0 1 1
1 0 1
1 1 0
0 0 0
Parity Check For Errors
PH PH*
0 0 0 0
0 1 1 0
1 0 1 0
1 1 0 0
PV 0 0 0 0
PV* 0 0 0 0
Recovered Data
0 0
0 1
1 0
1 1

Fig. 4.3 Data recovery process in ASK, showing no errors. If there is an error, there will be a parity failure in P_H* and P_V* to pinpoint the error

4.4 ASK Bandwidth

In wireless communications, the scarcity of RF spectrum is well known. For this reason, we have to be vigilant about using transmission bandwidth in error control coding and modulation. The transmission bandwidth depends on:

- Spectral response of the encoded data
- Spectral response of the carrier frequency, and
- Modulation type.

4.4.1 Spectral Response of the Encoded Data

In digital communications, data is generally referred to as a non-periodic digital signal. It has two values:

- Binary-1 = High, Period = T
- Binary-0 = Low, Period = T

Also, data can be represented in two ways:

- Time domain representation and
- Frequency domain representation

The time domain representation (Fig. 4.4a), known as non-return-to-zero (NRZ), is given by:

$$\begin{aligned} V(t) &= V \qquad < 0 < t < T \\ &= 0 \qquad \text{elsewhere} \end{aligned} \tag{4.5}$$

The frequency domain representation is given by "Fourier transform":

$$V(\omega) = \int_0^T V \cdot e^{-j\omega t} dt \tag{4.6}$$

$$\begin{aligned} |V(\omega)| &= VT\left[\frac{\sin(\omega T/2)}{\omega T/2}\right] \\ P(\omega) &= \left(\frac{1}{T}\right)|V(\omega)|^2 = V^2 T\left[\frac{\sin(\omega T/2)}{\omega T/2}\right]^2 \end{aligned} \tag{4.7}$$

Here, $P(\omega)$ is the power spectral density. This is plotted in Fig. 4.4b. The main lobe corresponds to the fundamental frequency and side lobes correspond to harmonic components. The bandwidth of the power spectrum is proportional to the frequency. In practice, the side lobes are filtered out since they are relatively insignificant with respect to the main lobe. Therefore, the one-sided bandwidth is given by the ratio $f/f_b = 1$. In other words, the one-sided bandwidth $= f = f_b$, where $f_b = R_b = 1/T$, T being the bit duration.

The general equation for two-sided response is given by:

$$V(\omega) = \int_{-\propto}^{\propto} V(t) \cdot e^{-j\omega t} dt$$

In this case, $V(\omega)$ is called two-sided spectrum of $V(t)$. This is due to both positive and negative frequencies used in the integral. The function can be a voltage or a current. Figure 4.4c shows the two-sided response, where the bandwidth is determined by the main lobe as shown below:

$$\text{Two-sided bandwidth (BW)} = 2R_b \; (R_b = \text{Bit rate before coding}) \tag{4.8}$$

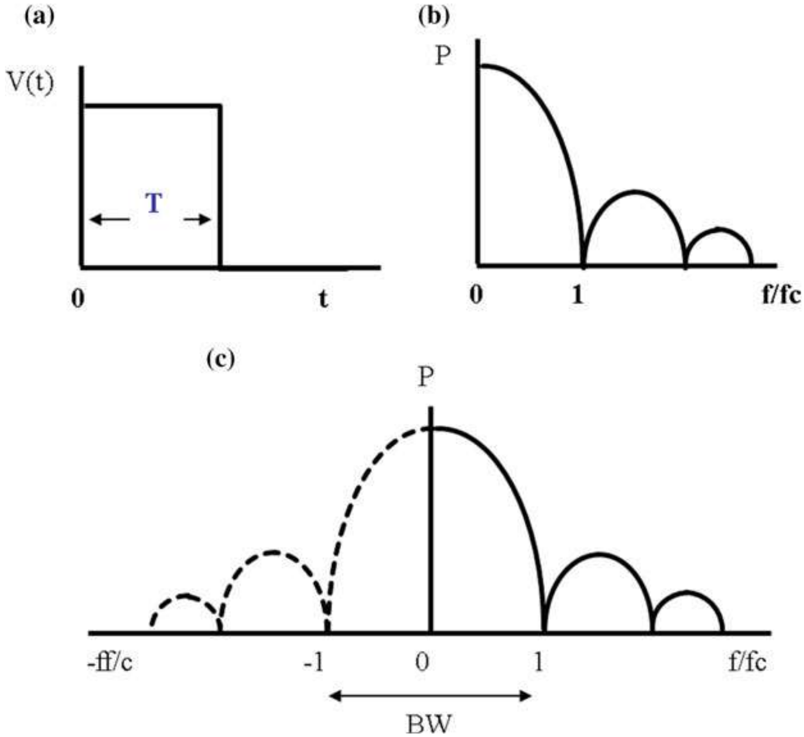

Fig. 4.4 **a** Discrete time digital signal **b** it is one-sided power spectral density and **c** two-sided power spectral density. The bandwidth associated with the non-return-to-zero (NRZ) data is $2R_b$, where R_b is the bit rate

Important Notes

1. If R_b is the bit rate before coding, and if the data is NRZ, then the bandwidth associated with the raw data will be $2R_b$. For example, if the bit rate before coding is 10 kb/s, then the bandwidth associated with the raw data will be 2×10 kb/s = 20 kHz.
2. If R_b is the bit rate before coding, code rate is r, and if the data is NRZ, then the bit rate after coding will be R_b (coded) = R_b (uncoded)r. The corresponding bandwidth associated with the coded data will be $2R_b$ (coded) = $2R_b$ (uncoded)/r. For example, if the bit rate before coding is 10 kb/s and the code rate $r = 1/2$, then the coded bit rate will be R_b (coded) = R_b (uncoded)/r = 10/0.5 = 20 kb/s. The corresponding bandwidth associated with the coded data will be $2 \times 20 = 40$ kHz.

4.4.2 Spectral Response of the Carrier Frequency Before Modulation

A carrier frequency is essentially a sinusoidal waveform, which is periodic and continuous with respect to time. It has one frequency component. For example, the sine wave is described by the following time domain equation:

$$V(t) = V_p \sin(\omega t_c) \tag{4.9}$$

where

$$V_P = \text{Peak voltage}$$

- $\omega_c = 2\pi f_c$
- f_c = Carrier frequency in Hz

Figure 4.5 shows the characteristics of a sine wave and its spectral response. Since the frequency is constant, its spectral response is located in the horizontal axis and the peak voltage is shown in the vertical axis. The corresponding bandwidth is zero.

4.4.3 ASK Bandwidth at a Glance

In ASK, the amplitude of the carriers changes in discrete levels, in accordance with the input signal, where

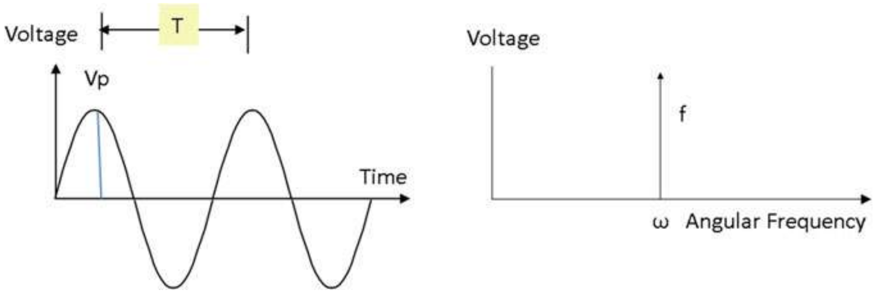

Fig. 4.5 High-frequency carrier frequency response

- Input Data : $m(t) = 0 \text{ or } 1$
- Carrier Frequency : $C(t) = Ac\cos(\omega_c t)$
- Modulated Carrier : $S(t) = m(t)C(t) = m(t)A_c\cos(\omega_c t)$

Since $m(t)$ is the input digital signal and it contains an infinite number of harmonically related sinusoidal waveforms and that we keep the fundamental and filter out the higher-order components, we write:

$$m(t) = A_m \sin(\omega_m t)$$

The ASK-modulated signal then becomes:

$$\begin{aligned} S(t) = m(t)S(t) &= A_m A_c \sin(\omega_m t)\cos(\omega_m t) \\ &= 1/2\, A_m A_c\, [\sin(\omega_c - \omega_m)t + \sin(\omega_c + \omega_m)t] \end{aligned} \tag{4.10}$$

The spectral response is depicted in Fig. 3.14. Notice that the spectral response after ASK modulation is the shifted version of the NRZ data. Bandwidth is given by: BW = $2R_b$ (coded), where R_b is the coded bit rate (Fig. 4.6).

Problem 1
Given:

- Bit rate before coding: R_{B1} = 10 kb/s
- Code rate: r = 8/15
- Modulation: ASK

Find:

(a) the bit rate after coding: R_{B2}
(b) Transmission bandwidth: BW

Solution

(a) Bit rate after coding (R_{B2})
$R_{B2} = R_{B1}/r$ = 10 kb/s (15/8) = 18.75 kb/s
(b) Transmission BW = 2 × 18.75 kb/s = 37.5 kHz.

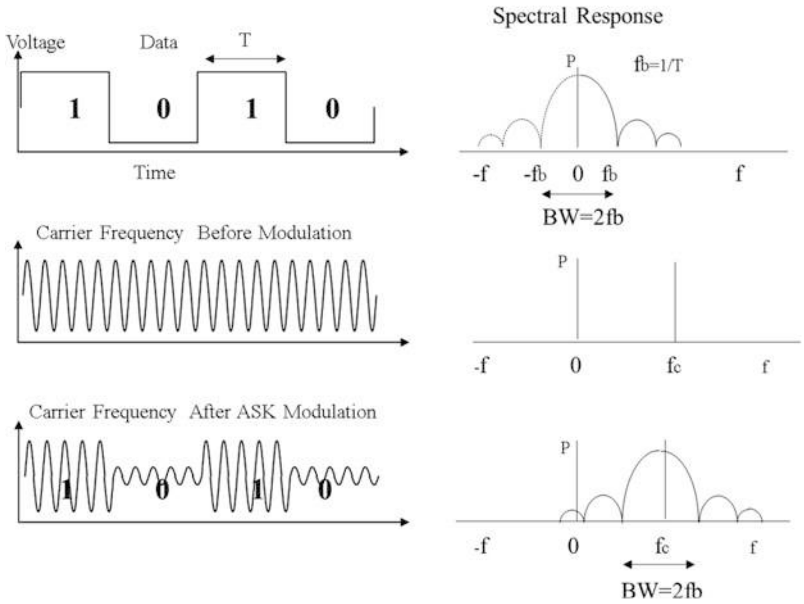
Voltage
Data
T
1
0
1
0
Time
Spectral Response
P
fb=1/T
-f
-fb
0
fb
f
BW=2fb
Carrier Frequency Before Modulation
P
-f
0
fc
f
Carrier Frequency After ASK Modulation
1
0
1
0
P
-f
0
fc
f
BW=2fb

Fig. 4.6 ASK bandwidth at a glance. **a** Spectral response of NRZ data before modulation. **b** Spectral response of the carrier before modulation. **c** Spectral response of the carrier after modulation. The transmission bandwidth is $2f_b$, where f_b is the bit rate and $T = 1/f_b$ is the bit duration for NRZ data

4.5 BER Performance

It is well known that an (n, k) bock code, where k = number of information bits, n = number of coded bits, can correct t errors [4, 5]. A measure of coding gain is then obtained by comparing the uncoded word error WER_U with the coded word WER_C. We examine this by means of the following analytical means.

Let the uncoded word error be defined as (WER_U). Then, with ASK modulation, the uncoded BER will be given by:

$$\mathrm{BER}_U = 0.5\,\mathrm{EXP}(-E_B/2N_0) \tag{4.11}$$

The probability that the uncoded word (WER_U) will be received in error is 1 minus the product of the probabilities that each bit will be received correctly. Thus, we write:

$$\mathrm{WER}_U = 1 - (1 - \mathrm{BER}_U)^k \tag{4.12}$$

Let the coded word error be defined as (WER_C). Since $n > k$, the coded bit energy to noise ratio will be modified to E_c/N_O, where $E_c/N_O = E_b/N_O + 10\log(k/n)$. Therefore, the coded $\mathrm{BER}c$ will be:

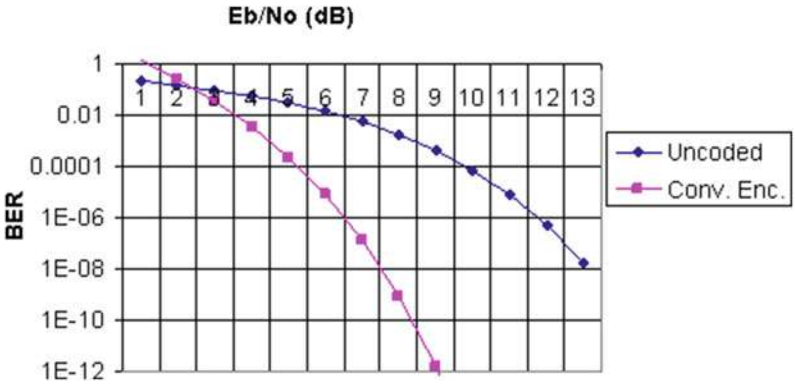

Fig. 4.7 Typical error performance in AWGN

$$\mathrm{BER}_C = 0.5\,\mathrm{EXP}(-E_C/2N_0) \tag{4.13}$$

The corresponding coded word error rate is given as:

$$\mathrm{WER}_C = \sum_{k=t+1}^{n} \binom{n}{k} \mathrm{BER}_C^K (1 - \mathrm{BER}_C)^{n-k} \tag{4.14}$$

When $\mathrm{BER}_c < 0.5$, the first term in the summation is the dominant one; therefore, Eq. (4.14) can be simplified as

$$\mathrm{WER}_C \approx \binom{n}{k} \mathrm{BER}_c^k (1 - \mathrm{BER}_c)^{n-k} \tag{4.15}$$

Using (15, 8) block code ($n = 15$, $k = 8$, $t = 1$), we obtain the coded and the uncoded WER as shown in Fig. 4.7. Coding gain is the difference in E_b/N_0 between the two curves. Notice that at least 3–4 db coding gain is available in this example where $r = k/n = 8/15$.

4.6 Conclusions

- This chapter presents ASK modulation and its attributes.
- Numerous illustrations are provided to show how amplitude of the carrier changes in discrete levels in accordance with the input digital signal, while the frequency of the carrier remains the same.
- The Fourier transform is used to derive the spectral components and ASK bandwidth is calculated.

- Bit error rate (BER) performance is presented.
- These materials have been augmented by diagrams and associated waveforms to make them easier for readers to grasp.

References

1. Smith DR (1985) Digital transmission systems, Van Nostrand Reinhold Co. ISBN: 0442009178
2. Leon W, Couch II (2001) Digital and analog communication systems, 7th edn. Prentice-Hall, Inc., Englewood Cliffs. ISBN: 0-13-142492-0
3. Sklar B (1988) Digital communications fundamentals and applications. Prentice Hall, Englewood Cliffs
4. Faruque S (2016) Radio frequency channel coding made easy. Springer. ISBN: 978-3-319-21169-5
5. Clark GC et al (1981) Error correction coding for digital communications. Plenum press, New York
6. Ungerboeck G (1982) Channel coding with multilevel/multiphase signals. IEEE Trans Inf Theory IT28:55–67
7. Lin S, Costello DJ Jr (1983) Error control coding: fundamentals and applications. Prentice-Hall, Inc., Englewood Cliffs
8. Blahut RE (1983) Theory and practice of error control codes. Addison-Wesley, Reading
9. van Lint JH (1992) Introduction to coding theory. GTM 86, 2nd edn. Springer, p 31. ISBN 3-540-54894-7
10. Mac Williams FJ, Sloane NJA (1977) The theory of error-correcting codes. North-Holland, p 35. ISBN 0-444-85193-3
11. Huffman W, Pless V (2003) Fundamentals of error-correcting codes. Cambridge University Press. ISBN 978-0-521-78280-7
12. Ebert PM, Tong SY (1968) Convolutional Reed-Soloman codes. Bell Syst Tech 48:729–742
13. Gallager RG (1968) Information theory and reliable communications. Wiley, New York
14. Kohlenbero A, Forney GD Jr (1968) Convolutional coding for channels with memory. IEEE Trans Inf Theory IT-14:618–626
15. Reddy SM (1968) Further results on convolutional codes derived from block codes. Inf Control 13:357–362
16. Reddy SM (1968) A class of linear convolutional codes for compound channels. Technical report, Bell Telephone Laboratories, Holmdel, New Jersey

Chapter 5
Frequency Shift Keying (FSK)

Topics

- Introduction
- Frequency Shift Keying (FSK)
- FSK spectrum
- FSK Bandwidth
- Performance Analysis

5.1 Introduction

In frequency shift keying (FSK), the frequency of the carrier changes in discrete levels in accordance with the input digital signal, while the amplitude of the carrier remains the same. This is shown in Fig. 5.1 where

- $m(t)$ is the input modulating digital signal,
- $C(t)$ is the carrier frequency, and
- $S(t)$ is the FSK-modulated carrier frequency.

As shown in the figure, the digital binary signal changes the frequency of the carrier on two discrete levels. This enables the receiver to extract the digital signal by demodulation. Notice that the frequency of the carrier changes in accordance with the input signal, while the amplitude of the carrier does not change after modulation. However, it can be shown that the modulated carrier $S(t)$ contains several spectral components, requiring frequency-domain analysis.

In the following sections, the above disciplines in FSK modulation will be presented, along with the respective spectrum and bandwidth. These materials have been augmented by diagrams and associated waveforms to make them easier for readers to grasp.

© The Author(s) 2017

S. Faruque, *Radio Frequency Modulation Made Easy*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-41202-3_5

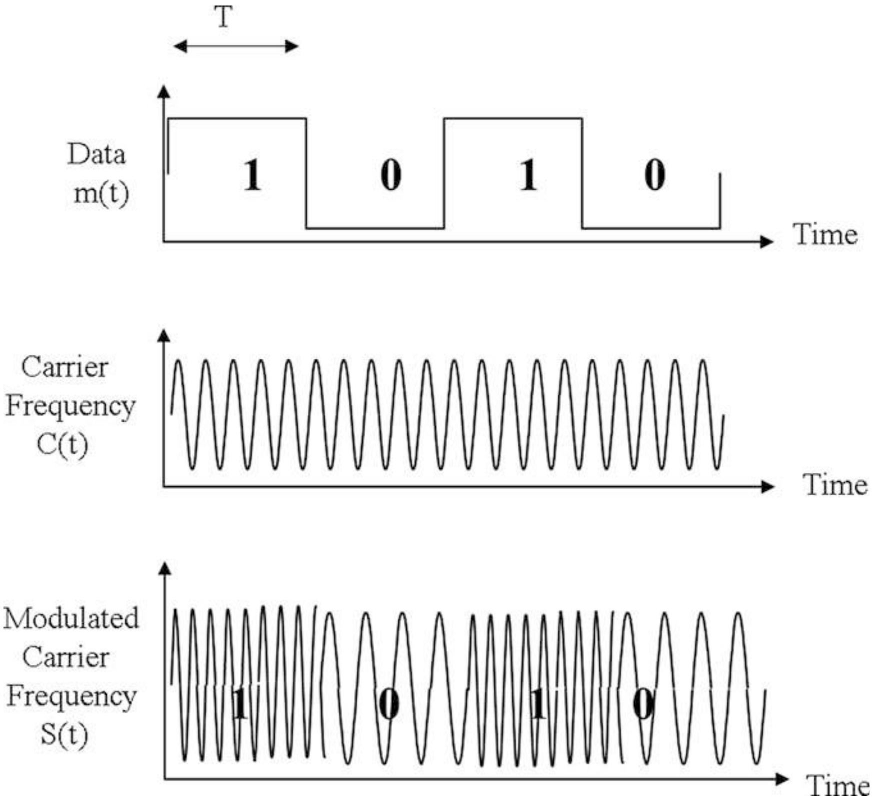

Fig. 5.1 FSK waveforms. The frequency of the carrier changes in accordance with the input digital signal. The amplitude of the carrier remains the same

5.2 Frequency Shift Keying (FSK) Modulation

Frequency shift keying (FSK) is a method of digital modulation that utilizes frequency shifting of the relative frequency content of the signal [1–3]. The signal to be modulated and transmitted is binary, which is encoded before modulation. This is an indispensable task in digital communications, where redundant bits are added with the raw data that enables the receiver to detect and correct bit errors, if they occur during transmission [4–16]. While there are many error-coding schemes available, we will use a simple coding technique, known as "block coding" to illustrate the concept.

Figure 5.2 shows an encoded FSK modulation scheme using (15, 8) block code where an 8-bit data block is formed as M rows and N columns ($M = 4$, $N = 2$). The product $MN = k = 8$ is the dimension of the information bits before coding. Next, a horizontal parity P_H is appended to each row and a vertical parity P_V is appended to each column. The resulting augmented dimension is given by the product $(M + 1)$

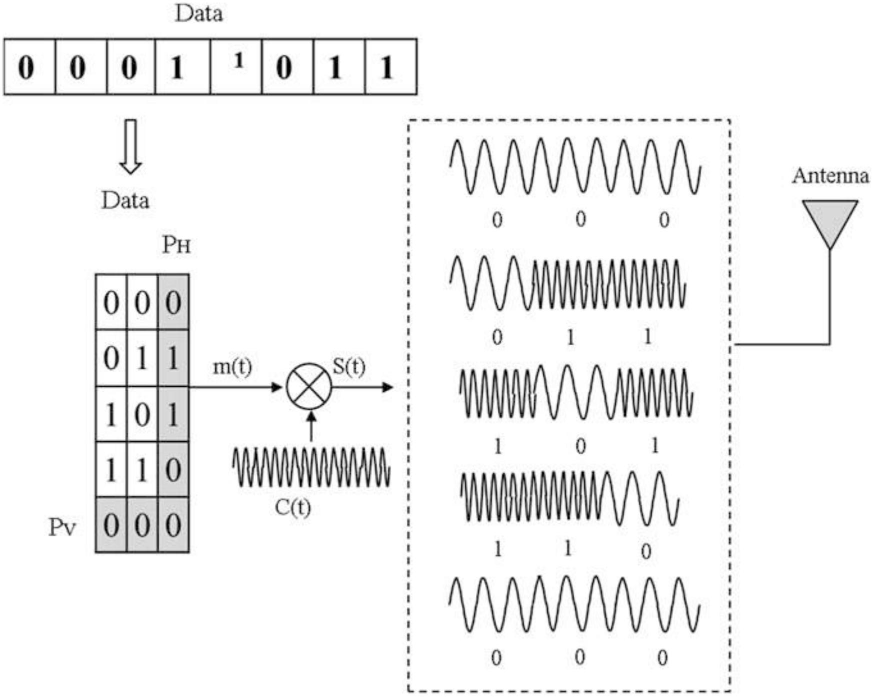

Fig. 5.2 Binary frequency shift keying (BFSK) modulation. The input encoded data block is transmitted row by row. The frequency of the carrier changes in accordance with the input digital signal

$(N + 1) = n = 15$, which is then FSK-modulated and transmitted row by row. The rate of this coding scheme is given by:

$$\text{Code Rate} : r = (MN)/[(M+1)(N+1)] = (4 \times 2)/5 \times 3) = 8/15 \tag{5.1}$$

The coded bit rate R_{b2} is given by:

$$R_{b2} = \text{Uncoded Bit Rate}/\text{Code Rate} = R_{b1}/r = R_{b1}(15/8) \tag{5.2}$$

Next, the coded bits are modulated by means of the FSK modulator as shown in the figure. Here,

- The input digital signal is the encoded bit sequence we want to transmit.
- Carrier is the radio frequency without modulation.
- Output is the FSK-modulated carrier, which has two frequencies corresponding to the binary input signal.
- For binary signal 1, the carrier changes to $f_c - \Delta f$.
- For binary signal 0, the carrier changes to $f_c + \Delta f$.
- The total frequency deviation is $2\Delta f$.

As shown in Fig. 3.8, the frequency of the carrier changes in discrete levels, in accordance with the input signals. We have:

$$\begin{array}{lll} \bullet & \text{Input Data:} & m(t) = 0 \text{ or } 1 \\ \bullet & \text{Carrier Frequency:} & C(t) = A\cos(\omega t) \\ \bullet & \text{Modulated Carrier:} & S(t) = A\cos(\omega - \Delta\omega)t, \text{For } m(t) = 1 \\ & & S(t) = A\cos(\omega + \Delta\omega)t, \text{For } m(t) = 0 \end{array} \tag{5.3}$$

where

$$\begin{array}{l} \bullet \; A = \text{Frequency of the carrier} \\ \bullet \; \omega = \text{Nominal frequency of the carrier frequency} \\ \bullet \; \Delta\omega = \text{Frequency deviation.} \end{array} \tag{5.4}$$

5.3 Frequency Shift Keying (FSK) Demodulation

Once the modulated binary data has been transmitted, it needs to be received and demodulated. This is often accomplished by the use of band-pass filters. In the case of binary FSK, the receiver needs to utilize two band-pass filters that are tuned to the appropriate frequencies. Since the nominal carrier frequency and the frequency deviation are known, this is relatively straightforward. One band-pass filter will be centered at the frequency ω_1 and the other at ω_2. As the signal enters into the receiver, it passes through the respective filter and the corresponding bit value is made. This is shown in Fig. 5.3. In order to assure that the bits are decoded correctly, the frequency deviation needs to be chosen with the limitations of the filters in mind to eliminate crossover.

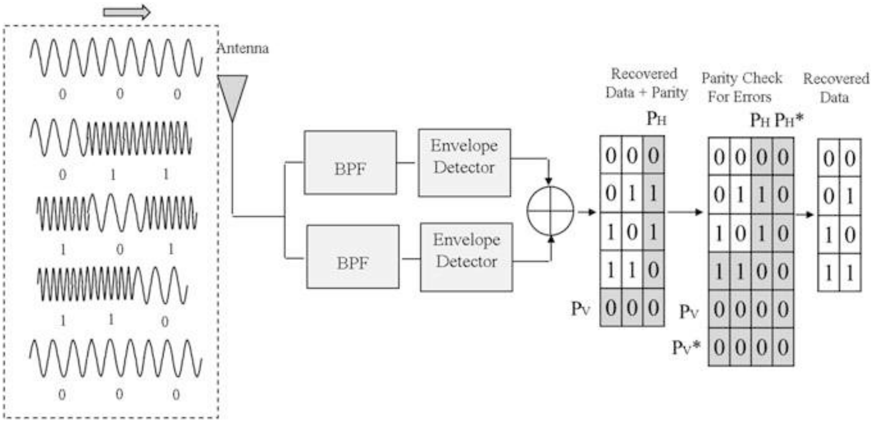
Antenna
0 0 0
0 1 1
1 0 1
1 1 0
0 0 0
BPF
Envelope Detector
BPF
Envelope Detector
Recovered Data + Parity
PH
000
011
101
110
PV 000
Parity Check For Errors
PH PH*
0000
0110
1010
1100
PV 0000
PV* 0000
Recovered Data
00
01
10
11

Fig. 5.3 Binary FSK detector utilizing two matched band-pass filters

5.4 FSK Bandwidth

In wireless communications, the scarcity of RF spectrum is well known. For this reason, we have to be vigilant about using transmission bandwidth in error control coding and modulation. The transmission bandwidth depends on the following:

- Spectral response of the encoded data
- Spectral response of the carrier frequency and
- Modulation type.

5.4.1 Spectral Response of the Encoded Data

In digital communications, data is generally referred to as a non-periodic digital signal. It has two values:

- Binary 1 = high, period = *T*
- Binary 0 = low, period = *T*

Also, data can be represented in two ways:

- Time-domain representation and
- Frequency-domain representation

The time-domain representation (Fig. 5.4a), known as non-return-to-zero (NRZ), is given by:

$$\begin{aligned} V(t) &= V \quad < 0 < t\ < T \\ &= 0 \quad ;\text{elsewhere} \end{aligned} \tag{5.5}$$

The frequency-domain representation is given by “Fourier transform”:

$$V(\omega) = \int_0^T V \cdot e^{-j\omega t} dt \tag{5.6}$$

$$|V(\omega)| = VT\left[\frac{\sin(\omega T/2)}{\omega T/2}\right] \tag{5.7}$$

$$P(\omega) = \left(\frac{1}{T}\right)|V(\omega)|^2 = V^2 T\left[\frac{\sin(\omega T/2)}{\omega T/2}\right]^2$$

Here, $P(\omega)$ is the power spectral density. This is plotted in Fig. 5.4b. The main lobe corresponds to the fundamental frequency, while side lobes correspond to harmonic components. The bandwidth of the power spectrum is proportional to the

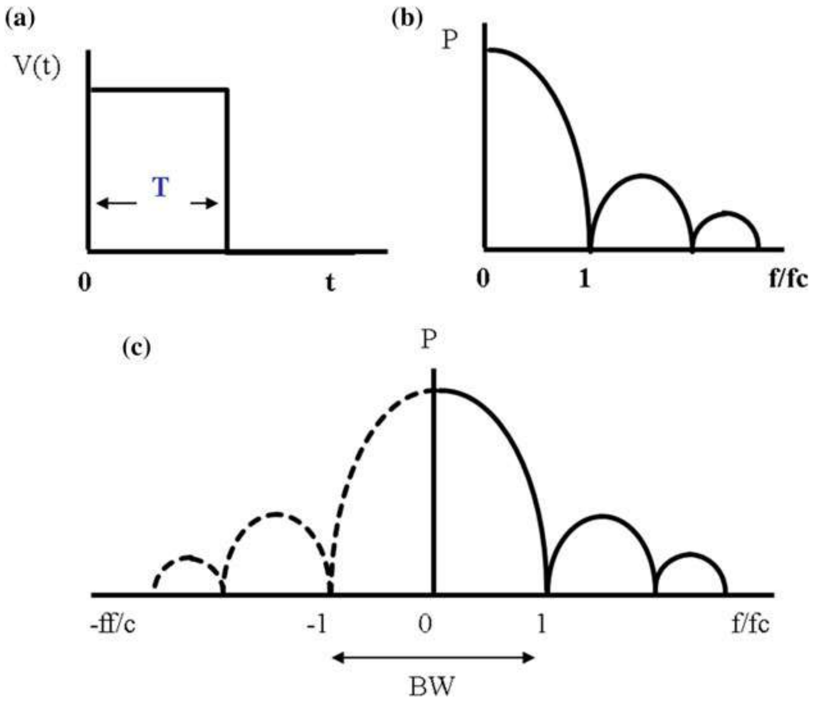

Fig. 5.4 **a** Discrete-time digital signal, **b** its one-sided power spectral density, and **c** two-sided power spectral density. The bandwidth associated with the non-return-to-zero (NRz) data is $2R_b$, where R_b is the bit rate

frequency. In practice, the side lobes are filtered out, since they are relatively insignificant with respect to the main lobe. Therefore, the one-sided bandwidth is given by the ratio $f/f_b = 1$. In other words, the one-sided bandwidth $= f = f_b$, where $f_b = R_b = 1/T$, T being the bit duration.

The general equation for two-sided response is given by:

$$V(\omega) = \int_{-\propto}^{\propto} V(t) \cdot e^{-j\omega t} dt$$

In this case, $V(\omega)$ is called the two-sided spectrum of $V(t)$. This is due to both positive and negative frequencies used in the integral. The function can be either a voltage or a current. Figure 5.4c shows the two-sided response, where the bandwidth is determined by the main lobe as shown below:

$$\text{Two sided bandwidth (BW)} = 2R_b (R_b = \text{Bit rate before coding}) \tag{5.8}$$

Important Notes:

1. If R_b is the bit rate before coding, and if the data is NRZ, then the bandwidth associated with the raw data will be $2R_b$. For example, if the bit rate before coding is 10 kb/s, then the bandwidth associated with the raw data will be 2×10 kb/s = 20 kHz.
2. If R_b is the bit rate before coding, code rate is r, and if the data is NRZ, then the bit rate after coding will be R_b (coded) = R_b(uncoded)r. The corresponding bandwidth associated with the coded data will be $2R_b$ (coded) = $2R_b$ (uncoded)/r. For example, if the bit rate before coding is 10 kb/s and the code rate $r = 1/2$, the coded bit rate will be R_b (coded) = R_b (uncoded)/r = 10/0.5 = 20 kb/s. The corresponding bandwidth associated with the coded data will be $2 \times 20 = 40$ kHz.

5.4.2 Spectral Response of the Carrier Frequency Before Modulation

A carrier frequency is essentially a sinusoidal waveform, which is periodic and continuous with respect to time. It has one frequency component. For example, the sine wave is described by the following time-domain equation:

$$V(t) = V_p \sin(\omega t_c) \tag{5.9}$$

where

$$V_P = \text{Peak voltage}$$

- $\omega_c = 2\pi f_c$
- f_c = Carrier frequency in Hz

Figure 5.5 shows the characteristics of a sine wave and its spectral response. Since the frequency is constant, its spectral response is located in the horizontal axis and the peak voltage is shown in the vertical axis. The corresponding bandwidth is zero.

5.4.3 FSK Bandwidth at a Glance

In FSK, the frequency of the carrier changes in two discrete levels, in accordance with the input signals. We have:

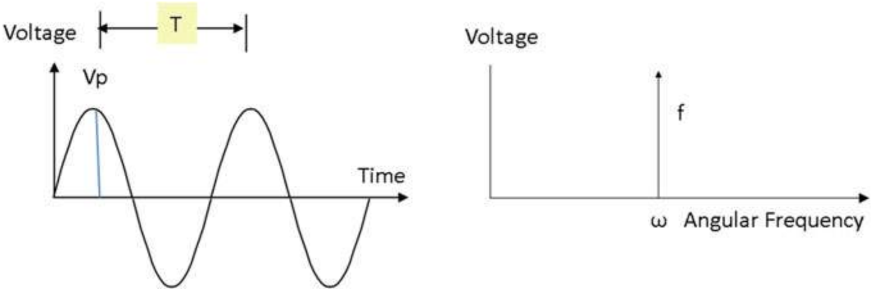

Fig. 5.5 High-frequency carrier response

- Input Data: $m(t) = 0 \text{ or } 1$
- Carrier Frequency: $C(t) = A\cos(\omega t)$
- Modulated Carrier: $S(t) = A\cos(\omega - \Delta\omega)t, \text{For } m(t) = 1$
  $S(t) = A\cos(\omega + \Delta\omega)t, \text{For } m(t) = 0$

where

- $S(t)$ = The modulated carrier
- A = Amplitude of the carrier
- ω = Nominal frequency of the carrier
- $\Delta\omega$ = Frequency deviation

The spectral response is depicted in Fig. 5.6. Notice that the carrier frequency after FSK modulation varies back and forth from the nominal frequency f_c by $\pm\Delta f_c$, where Δf_c is the frequency deviation. The FSK bandwidth is given by:

$$\begin{aligned} \mathrm{BW} &= 2(f_b + \Delta f_c) \\ &= 2f_b(1 + \Delta f_c/f_b) \\ &= 2f_b(1 + \beta) \end{aligned} \tag{5.10}$$

where $\beta = \Delta f/f_b$ is known as the modulation index and f_b is the coded bit frequency (bit rate R_b). The above equation is also known as "Carson's rule."

Problem
Given:

- Bit rate before coding: R_{B1} = 10 kb/s
- Code rate: r = 8/15
- Modulation: FSK
- Modulation index β = 1

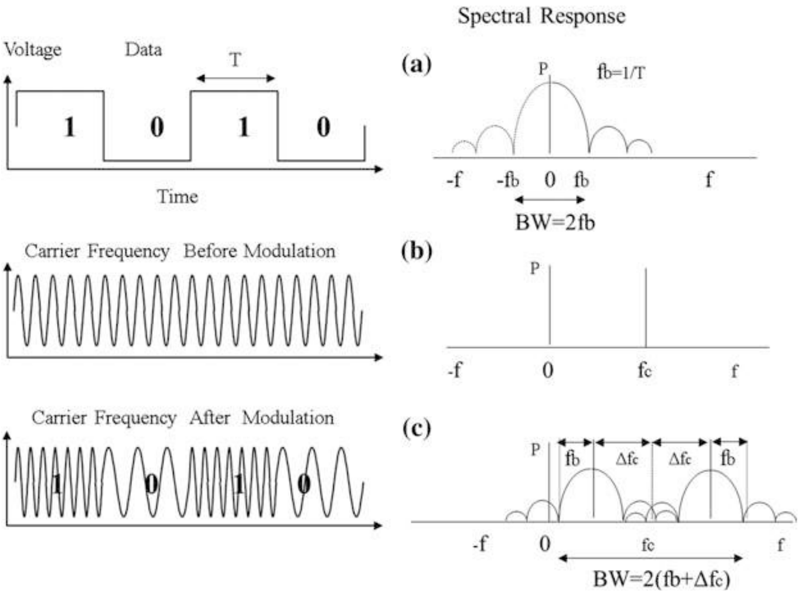
Spectral Response
Voltage
Data
T
1
0
1
0
Time
(a)
P
fb=1/T
-f
-fb
0
fb
f
BW=2fb
Carrier Frequency Before Modulation
(b)
P
-f
0
fc
f
Carrier Frequency After Modulation
1
0
1
0
(c)
P
fb
Δfc
Δfc
fb
-f
0
fc
f
BW=2(fb+Δfc)

Fig. 5.6 FSK bandwidth at a glance. **a** Spectral response of NRZ data before modulation. **b** Spectral response of the carrier before modulation. **c** Spectral response of the carrier after modulation. The transmission bandwidth is $2(f_b + \Delta f_c)$. f_b where f_b is the bit rate and Δf_c is the frequency deviation $=1/f_b$ is the bit duration for NRZ data

Find:

(a) the bit rate after coding: R_{B2}
(b) Transmission bandwidth: BW

Solution

(a) Bit rate after coding (R_{B2}):
$R_{B2} = R_{B1}/r = 10$ kb/s $(15/8) = 18.75$ kb/s
(b) Transmission BW = 2 $R_{B2}(1+\beta)$
$= 2 \times 18.75$ kb/s $(1 + 1) = 75$ kHz.

5.5 BER Performance

It is well known that an (n, k) bock code, where k = number of information bits and n = number of coded bits, can correct t errors [4, 5]. A measure of coding gain is then obtained by comparing the uncoded word error WER_U, to the coded word WER_C. We examine this by means of the following analytical means.

Let the uncoded word error be defined as (WER_U). Then, with FSK modulation, the uncoded BER will be given by:

$$\mathrm{BER}_U = 0.5\,\mathrm{EXP}(-E_B/2N_0) \tag{5.11}$$

The probability that the uncoded word (WER_U) will be received in error is 1 minus the product of the probabilities that each bit will be received correctly. Thus, we write:

$$\mathrm{WER}_U = 1-(1-\mathrm{BER}_U)^k \tag{5.12}$$

Let the coded word error be defined as (WER_C). Since $n > k$, the ratio of coded bit energy to noise will be modified to E_c/N_0, where $E_c/N_0 = E_b/N_0 + 10\log(k/n)$. Therefore, the coded BER_c will be:

$$\mathrm{BER}_C = 0.5\,\mathrm{EXP}(-E_c/2N_0) \tag{5.13}$$

The corresponding coded word error rate is:

$$\mathrm{WER}_C = \sum_{k=t+1}^{n} \binom{n}{k} \mathrm{BER}_C^K (1-\mathrm{BER}_C)^{n-k} \tag{5.14}$$

When $\mathrm{BER}_c < 0.5$, the first term in the summation is the dominant one; therefore, equation can be simplified as

$$\mathrm{WER}_C \approx \binom{n}{k} \mathrm{BER}_c^k (1-\mathrm{BER}_c)^{n-k} \tag{5.15}$$

Using (15, 8) block code ($n = 15$, $k = 8$, $t = 1$), we obtain the coded and the uncoded WER as shown in Fig. 5.7. Coding gain is the difference in E_b/N_0 between the two curves. Notice that at least 3–4-db coding gain is available in this example where $r = k/n = 8/15$.

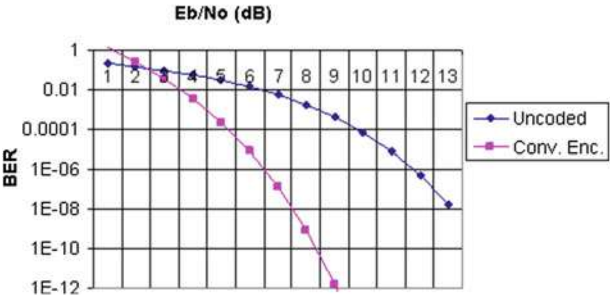

Fig. 5.7 Typical error performance in AWGN

5.6 Conclusions

- This chapter presents FSK modulation and its attributes.
- Numerous illustrations are provided to show how frequency of the carrier changes in discrete levels in accordance with the input digital signal, while the amplitude of the carrier remains the same.
- The Fourier transform is used to derive the spectral components, and FSK bandwidth is calculated.
- Bit error rate (BER) performance is presented.
- These materials have been augmented by diagrams and associated waveforms to make them easier for readers to grasp.

References

1. Smith DR (1985) Digital transmission systems, Van Nostrand Reinhold Co. ISBN: 0442009178
2. Leon W, Couch II (2001) Digital and analog communication systems, 7th edn. Prentice-Hall Inc., Englewood Cliffs. ISBN 0-13-142492-0
3. Sklar B (1988) Digital communications fundamentals and applications. Prentice Hall, Englewood Cliffs
4. Faruque S (2016) Radio frequency channel coding made easy. Springer. ISBN: 978-3-319-21169-5
5. Clark GC et al (1981) Error correction coding for digital communications. Plenum press, New York
6. Ungerboeck G (1982) Channel coding with multilevel/multiphase signals. IEEE Trans Inf Theory IT28:55–67
7. Lin S, Costello DJ Jr (1983) Error control coding: fundamentals and applications. Prentice-Hall, Inc., Englewood Cliffs
8. Blahut RE (1983) Theory and practice of error control codes. Addison-Wesley, Reading
9. van Lint JH (1992) Introduction to coding theory. GTM 86, 2nd edn. Springer, p 31. ISBN 3-540-54894-7
10. Mac Williams FJ, Sloane NJA (1977) The theory of error-correcting codes. North-Holland, p 35. ISBN 0-444-85193-3
11. Huffman W, Pless V (2003) Fundamentals of error-correcting codes. Cambridge University Press. ISBN 978-0-521-78280-7
12. Ebert PM, Tong SY (1968) Convolutional Reed-Soloman codes. Bell Syst Tech 48:729–742
13. Gallager RG (1968) Information theory and reliable communications. Wiley, New York
14. Kohlenbero A, Forney GD Jr (1968) Convolutional coding for channels with memory. IEEE Trans Inf Theory IT-14:618–626
15. Reddy SM (1968) Further results on convolutional codes derived from block codes. Inf Control 13:357–362
16. Reddy SM (1968) A class of linear convolutional codes for compound channels. Technical report, Bell Telephone Laboratories, Holmdel, New Jersey

Chapter 6
Phase Shift Keying (PSK)

Topics

- Introduction
- Binary Phase Shift Keying (BPSK) Modulation
- QPSK Modulation
- 8PSK Modulation
- 16PSK Modulation
- PSK spectrum and Bandwidth
- SSB Spectrum and Bandwidth

6.1 Introduction

In phase shift keying (PSK), the phase of the carrier changes in discrete levels in accordance with the input digital signal, while the amplitude of the carrier remains the same. This is shown in Fig. 6.1, where

- $m(t)$ is the input modulating digital signal,
- $C(t)$ is the carrier frequency, and
- $S(t)$ is the PSK-modulated carrier frequency.

As shown in the figure, the digital binary signal changes the phase of the carrier on two discrete levels. This enables the receiver to extract the digital signal by demodulation. Notice that the phase of the carrier changes in accordance with the input signal, while the amplitude of the carrier does not change after modulation. However, it can be shown that the modulated carrier $S(t)$ contains several spectral components, requiring frequency domain analysis.

In the following sections, the above disciplines in PSK modulation will be presented, along with the respective spectrum and bandwidth. These materials have been augmented by diagrams and associated waveforms to make them easier for readers to grasp.

© The Author(s) 2017

S. Faruque, *Radio Frequency Modulation Made Easy*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-41202-3_6

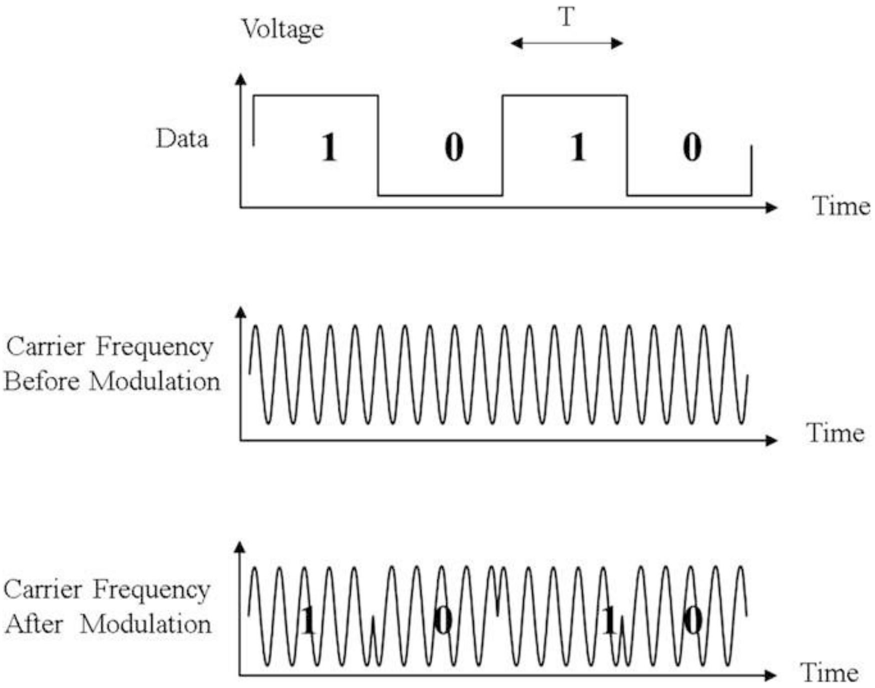

Fig. 6.1 Binary PSK (BPSK) waveforms. The phase of the carrier changes in accordance with the input digital signal. The amplitude of the carrier remains the same

6.2 Binary Phase Shift Keying (BPSK)

6.2.1 BPSK Modulation

Phase shift keying (PSK) is a method of digital modulation that utilizes phase shifting of the relative phase content of the signal [1–3]. The signal to be modulated and transmitted is binary, which is encoded before modulation. This is an indispensable task in digital communications, where redundant bits are added with the raw data that enable the receiver to detect and correct bit errors, if they occur during transmission [4–16]. While there are many error-coding schemes available, we will use a simple coding technique, known as “Block Coding” to illustrate the concept.

Figure 6.2 shows an encoded BPSK modulation scheme using (15, 8) block code where an 8-bit data block is formed as M-rows and N-columns ($M = 4$, $N = 2$). The product $MN = k = 8$ is the dimension of the information bits before coding. Next, a horizontal parity P_H is appended to each row and a vertical parity P_V is appended to each column. The resulting augmented dimension is given by the product $(M + 1)$ $(N + 1) = n = 15$, which is then PSK modulated and transmitted row by row. The rate of this coding scheme is given by

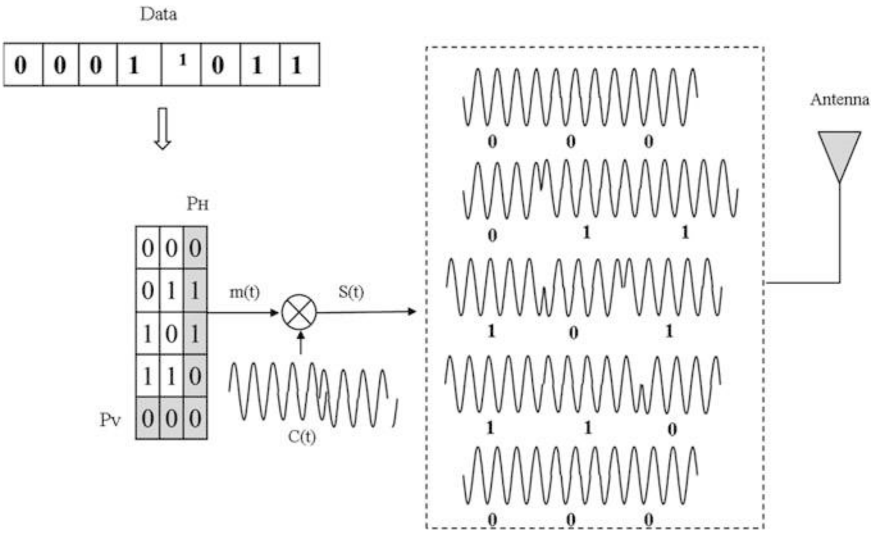

Fig. 6.2 Binary phase shift keying (BPSK) modulation. The input encoded data block is transmitted row by row. The phase of the carrier changes in accordance with the input digital signal

$$\text{Code Rate}: r = (MN)/[(M+1)(N+1)] = (4\times 2)/5\times 3) = 8/15 \tag{6.1}$$

The coded bit rate R_{b2} is given by:

$$R_{b2} = \text{Uncoded Bit Rate/Code Rate} = R_{b1}/r = R_{b1}(15/8) \tag{6.2}$$

Next, the coded bits are modulated by means of the PSK modulator as shown in the figure. Here,

- The Input digital signal is the encoded bit sequence we want to transmit;
- Carrier is the radio frequency without modulation;
- Output is the PSK-modulated carrier, which has two phases corresponding to the binary input signals;
- For binary signal 0, $\varphi = 0°$; and
- For binary signal 1, $\varphi = 180^{\circ}$.

As shown in Fig. 6.2, the phase of the carrier changes in two discrete levels, in accordance with the input signals. Here, we have the following:

- Input Data : $m(t) = 0 \text{ or } 1$
- Carrier Frequency : $C(t) = A\cos(\omega_{ct})$
- Modulated Carrier : $S(t) = A_c\cos[\omega_c t + 2\pi/M)m(t)] \quad m(t) = 0,\ 1,\ 2,\ 3,\ldots M-1$ (6.3)

where,

- A_c = Amplitude of the carrier frequency
- ω_c = Angular frequency of the carrier
- $M = 2, 4, 8, 16, \ldots$
- In BPSK, there are two phases 1 bit/phase($M = 2$) (6.4)
- In QPSK, there are four phases, 2-bits/phase, $M = 4$
- In 8PSK, there are 8 phases, 3-bits/phase, $M = 8$
- In 16PSK, there are 16 phases, 4-bits/phase, $M = 16$

We can also represent the BPSK modulator as a signal constellation diagram with $M = 2$, 1 bit per phase. This is shown in Fig. 6.3, where the input raw data, having a bit rate R_{b1}, is encoded by means of a rate r encoder. The encoded data, having a bit rate R_{b2}, = $R_{b1}r$ ($r < 1$), is modulated by the BPSK modulator as shown in Fig. 6.3.

The BPSK modulator takes one bit at a time to construct the phase constellation having two phases, also known as "Symbols," where each symbol represents one bit. The symbol rate is therefore the same as the encoded bit rate R_{b2}.

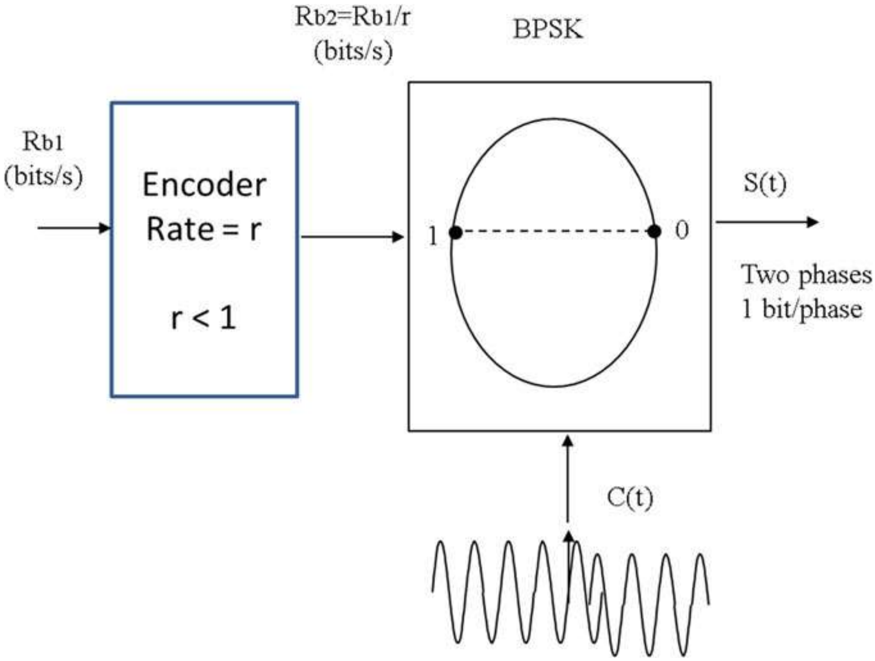

Fig. 6.3 BPSK signal constellation having 2 symbols, 1-bit per symbol

Therefore, the BPSK modulator has the following specifications:

- 2 phases or 2 symbols;
- 1-bit/symbol,

The above specifications govern the transmission bandwidth, as we shall see later.

6.2.2 BPSK Demodulation

Once the modulated binary data has been transmitted, it needs to be received and demodulated. This is often accomplished with the use of a phase detector, typically known as phase-locked loop (PLL). As the signal enters the receiver, it passes through the PLL. The PLL locks the incoming carrier frequency and tracks the variations in frequency and phase. This is known as coherent detection technique, where the knowledge of the carrier frequency and phase must be known to the receiver.

Figure 6.4 shows a simplified diagram of a BPSK demodulator along with the data recovery process. In order to assure that the bits are decoded correctly, the phase deviation needs to be chosen with the limitations of the PLL in mind to eliminate crossover.

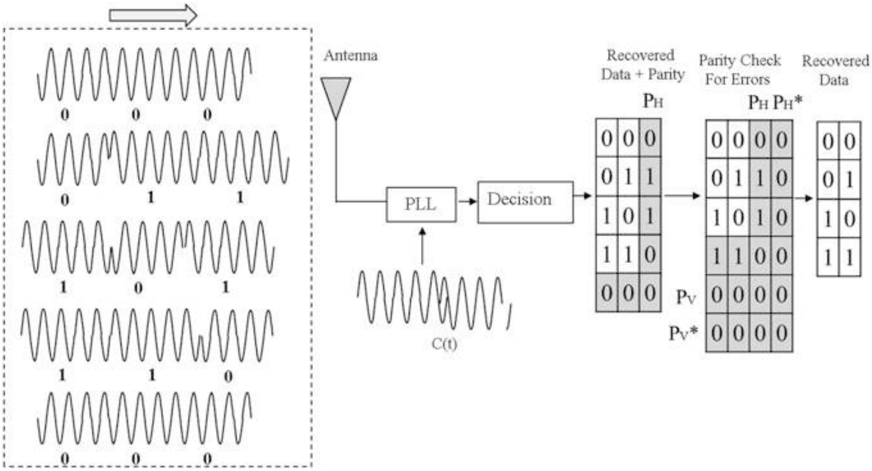
0 0 0
0 1 1
1 0 1
1 1 0
0 0 0
Antenna
PLL
Decision
C(t)
Recovered Data + Parity
PH
0 0 0
0 1 1
1 0 1
1 1 0
0 0 0
Parity Check For Errors
PH PH*
0 0 0 0
0 1 1 0
1 0 1 0
1 1 0 0
PV 0 0 0 0
PV* 0 0 0 0
Recovered Data
0 0
0 1
1 0
1 1

Fig. 6.4 Binary PSK detector showing data recovery process

6.3 QPSK Modulation

In QPSK, the input raw data, having a bit rate R_{b1}, is encoded by a rate r ($r < 1$) encoder. The encoded data, having a bit rate $R_{b2} = R_{b1}/r$, is serial to parallel converted into two parallel streams. The encoded bit rate, now reduced in speed by a factor of two, is modulated by the QPSK modulator as shown in Fig. 6.5.

The QPSK modulator takes one bit from each stream to construct the phase constellation having four phases, also known as "Symbols," where each symbol represents two bits. The symbol rate is therefore reduced by a factor of two. The QPSK modulator has four phases or 4 symbols, 2-bits/symbol as shown in the figure.

Therefore, the QPSK modulator has the following specifications:

- 4 phases or 4 symbols
- 2-bits/symbol

The above specifications govern the transmission bandwidth, as we shall see later.

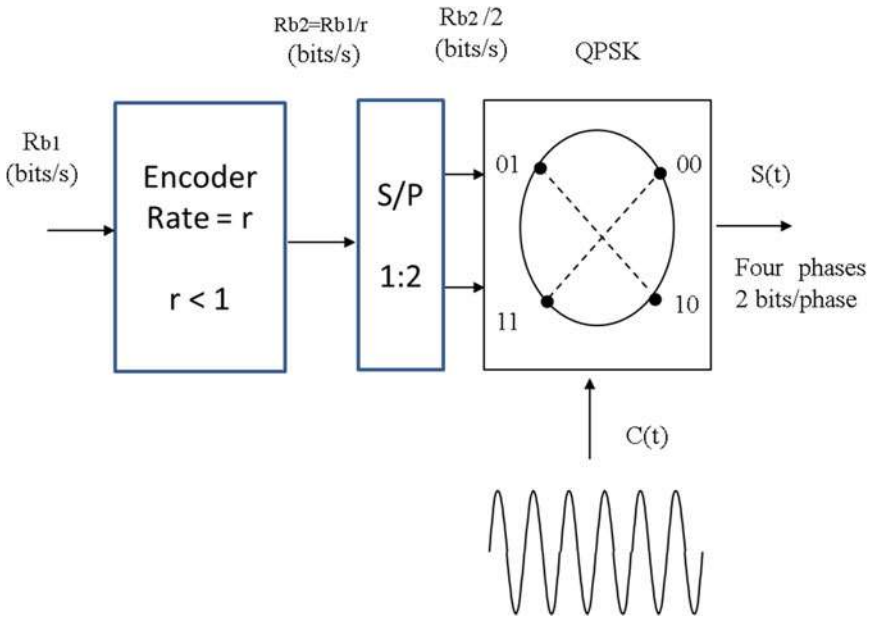
Rb2=Rb1/r
(bits/s)
Rb2 /2
(bits/s)
QPSK
Rb1
(bits/s)
Encoder
Rate = r
r < 1
S/P
1:2
01
00
11
10
S(t)
Four phases
2 bits/phase
C(t)

Fig. 6.5 QPSK signal constellation having 4 symbols, 2-bits per symbol

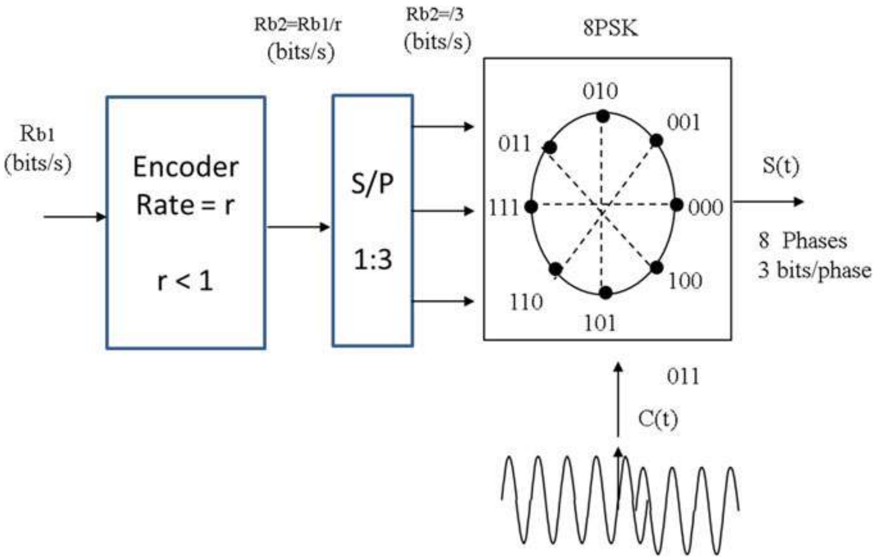

Fig. 6.6 8PSK signal constellation having 8 symbols, 3-bits per symbol

6.4 8PSK Modulation

In 8PSK, the input raw data, having a bit rate R_{b1}, is encoded by a rate r ($r < 1$) encoder. The encoded data, having a bit rate $R_{b2} = R_{b1}/r$, is serial to parallel converted into three parallel streams. The encoded bit rate, now reduced in speed by a factor of three, is modulated by the 8PSK modulator as shown in Fig. 6.6.

The 8PSK modulator takes one bit from each stream to construct the phase constellation having 8 phases, also known as "Symbols," where each symbol represents 3-bits. The symbol rate is therefore reduced by a factor of 3. The 8PSK modulator has 8 phases or 8 symbols, 3-bits/symbol as shown in the figure.

Therefore, the 8PSK modulator has the following specifications:

- 8 phases or 8 symbols;
- 3-bits/symbol.

The above specifications govern the transmission bandwidth, as we shall see later.

6.5 16PSK Modulation

In 16PSK, the input raw data, having a bit rate R_{b1}, is encoded by means of a rate r ($r < 1$) encoder. The encoded data, having a bit rate $R_{b2} = R_{b1}/r$, is serial to parallel converted into four parallel streams. The encoded bit rate, now reduced in speed by a factor of four, is modulated by the 16PSK modulator as shown in Fig. 6.7.

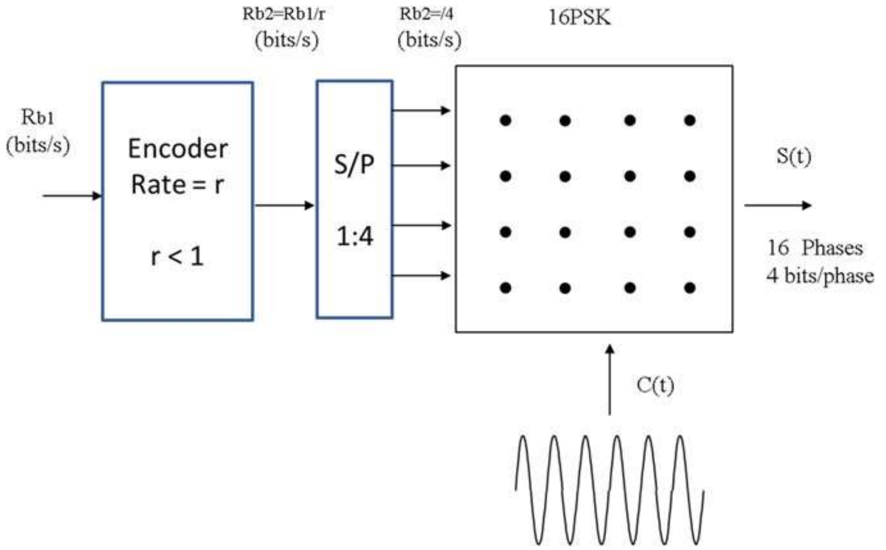

Fig. 6.7 16PSK signal constellation having 16 symbols, 4-bits per symbol. Here, each symbol is represented by a *dot*, where each dot represents 4-bits

The 16PSK modulator takes one bit from each stream to construct the phase constellation having 16 phases, also known as “Symbols,” where each symbol represents four bits. The symbol rate is therefore reduced by a factor of four. Therefore, the 16PSK modulator has 16 phases or 16 symbols, 4-bits/symbol as shown in the figure.

Therefore, the 16PSK modulator has the following specifications:

- 16 phases or 16 symbols;
- 4-bits/symbol.

The above specifications govern the transmission bandwidth, as we shall see later.

Table 6.1 shows the number of phases and the corresponding bits per phase for MPSK modulation schemes for $M = 2, 4, 8, 16, 32, 64$, etc.

Table 6.1 MPSK modulation parameters. $M = 2, 4, 8, 16$, and 32

Modulation	Number of phases φ	Number of bits per phase
BPSK	2	1
QPSK	4	2
8PSK	8	3
16	16	4
32	32	5
64	64	6
:	:	:

6.6 PSK Spectrum and Bandwidth

In wireless communications, the scarcity of RF spectrum is well known. For this reason, we have to be vigilant about using transmission bandwidth in error control coding and modulation. The transmission bandwidth depends on:

- Spectral response of the encoded data;
- Spectral response of the carrier phase; and
- Modulation type.

Let us take a closer look:

6.6.1 Spectral Response of the Encoded Data

In digital communications, data is generally referred to as a non-periodic digital signal. It has two values:

- Binary-1 = High, Period = T
- Binary-0 = Low, Period = T

Also, data can be represented in two ways:

- Time domain representation and
- Frequency domain representation

The time domain representation (Fig. 6.8a), known as non-return-to-zero (NRZ), is given by:

$$\begin{aligned} V(t) &= V \quad < 0 < t < T \\ &= 0 \quad \text{elsewhere} \end{aligned} \tag{6.5}$$

The frequency domain representation is given by "Fourier Transform":

$$V(\omega) = \int_{0}^{T} V \cdot e^{-j\omega t} dt \tag{6.6}$$

$$\begin{aligned} |V(\omega)| &= 2\left(\frac{V}{\omega}\right)\sin\left(\frac{\omega T}{2}\right) = VT\left[\frac{\sin(\omega T/2)}{\omega T/2}\right] \\ P(\omega) &= \left(\frac{1}{T}\right)|V(\omega)|^2 = V^2 T\left[\frac{\sin(\omega T/2)}{\omega T/2}\right]^2 \end{aligned} \tag{6.7}$$

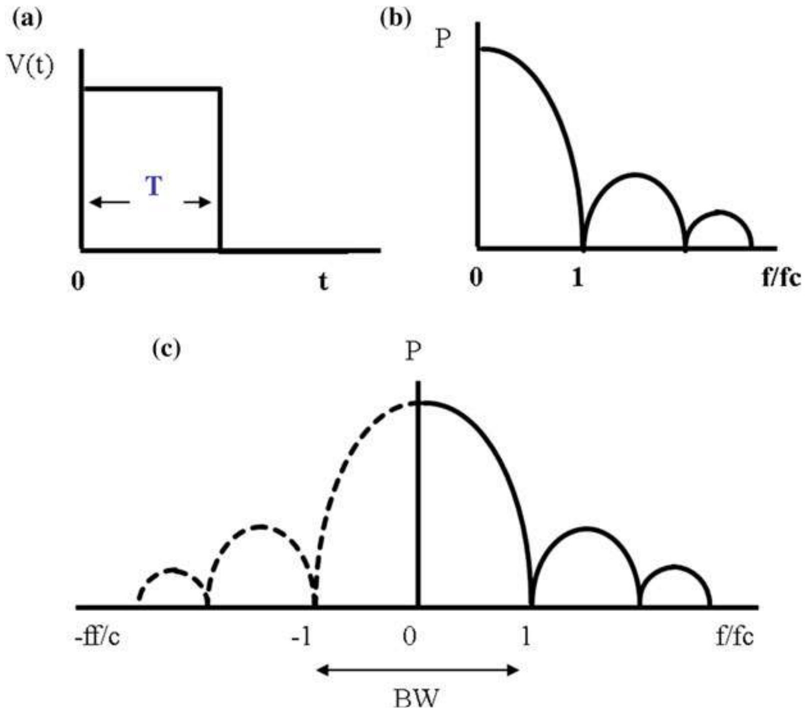

Fig. 6.8 **a** Discrete time digital signal **b** it is one-sided power spectral density and **c** two-sided power spectral density. The bandwidth associated with the non-return-to-zero (NRz) data is $2R_b$, where R_b is the bit rate

Here, $P(\omega)$ is the power spectral density. This is plotted in Fig. 6.8b. The main lobe corresponds to the fundamental frequency, while the side lobes correspond to harmonic components. The bandwidth of the power spectrum is proportional to the frequency. In practice, the side lobes are filtered out, since they are relatively insignificant with respect to the main lobe. Therefore, the one-sided bandwidth is given by the ratio $f/fb = 1$. In other words, the one-sided bandwidth $= f = f_b$, where $f_b = R_b = 1/T$, T being the bit duration.

The general equation for two-sided response is given by:

$$V(\omega) = \int_{-\propto}^{\propto} V(t) \cdot e^{-j\omega t} dt$$

In this case, $V(\omega)$ is called the two-sided spectrum of $V(t)$. This is due to both positive and negative frequencies used in the integral. The function can be either a voltage or a current. Figure 6.8c shows the two-sided response, where the bandwidth is determined by the main lobe as shown below:

$$\text{Two-sided bandwidth (BW)} = 2R_b \, (R_b = \text{Bit rate before coding}) \tag{6.8}$$

6.6.2 Spectral Response of the Carrier Before Modulation

A carrier frequency is a sinusoidal waveform, which is periodic and continuous with respect to time. It has one phase component. For example, the sine wave is described by the following time domain equation:

$$V(t) = V_p \sin(\omega_c t) \tag{6.9}$$

Where

$$Vp = \text{Peak voltage}$$

- $\omega_c = 2\pi f_c$
- f_c = Carrier phase in Hz

Figure 6.9 shows the characteristics of a sine wave and its spectral response. Since the phase is constant, its spectral response is located in the horizontal axis and the peak voltage is shown in the vertical axis. The corresponding bandwidth is zero.

6.6.3 BPSK Spectrum

In BPSK, the input is a digital signal and it contains an infinite number of harmonically related sinusoidal waveforms. This is given by (see Sect. 6.6.1):

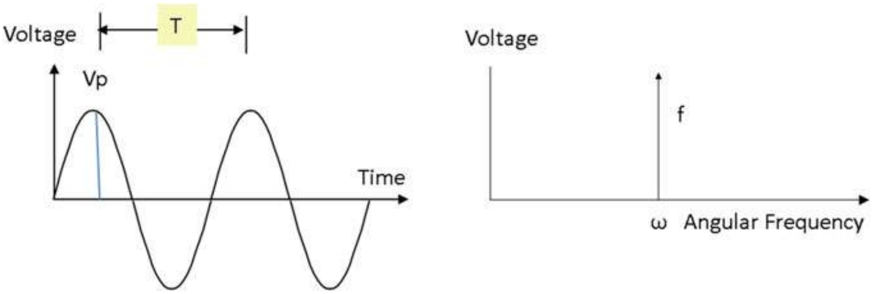

Fig. 6.9 High-frequency carrier response

$$|V(\omega)| = 2\left(\frac{V}{\omega}\right)\sin\left(\frac{\omega T}{2}\right) = VT\left[\frac{\sin(\omega T/2)}{\omega T/2}\right] \tag{6.10}$$

Here, $V\ (\omega)$ is the frequency domain representation of the input digital signal, which has a $\sin(x)/x$ response that governs the phase of the carrier frequency.

With $V(t) = m(t)$, we write the following as:

$$S(t) = A_c \cos[\omega_c t + \beta\, m(t)] \tag{6.11}$$

where β is the phase deviation index of the carrier and $m(t)$ has a $\sin(x)/x$ response, which is given by

$$m(t) = VT\left[\frac{\sin(\omega m T/2)}{\omega m T/2}\right] \tag{6.12}$$

Therefore, the spectral response after BPSK modulation also has a $\sin(x)/x$ response, which is the shifted version of the NRZ data, centered on the carrier frequency f_c, as shown in Fig. 6.10. The transmission bandwidth associated with the main lobe is given by

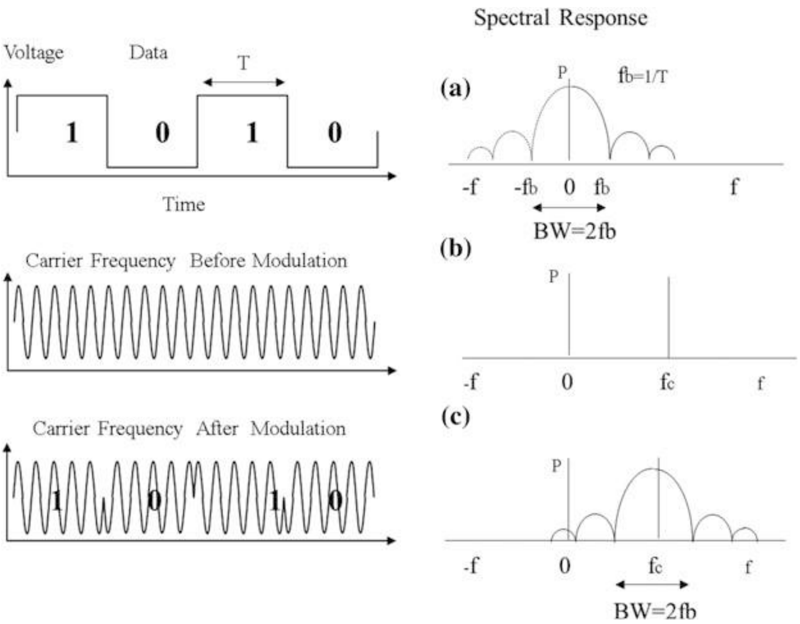
Spectral Response
Voltage
Data
T
1
0
1
0
Time
Carrier Frequency Before Modulation
Carrier Frequency After Modulation
1
0
1
0
(a)
P
fb=1/T
-f
-fb
0
fb
f
BW=2fb
(b)
P
-f
0
fc
f
(c)
P
-f
0
fc
f
BW=2fb

Fig. 6.10 BPSK bandwidth. **a** Spectral response of NRZ data before modulation. **b** Spectral response of the carrier before modulation. **c** Spectral response of the carrier after modulation

$$\begin{aligned}\mathrm{BW}\,(\mathrm{BPSK}) &\approx 2R_{b2}/\text{Bit per Phase}\\ &\approx 2R_{b2}/1 \approx 2R_{b2}\end{aligned} \tag{6.13}$$

where R_{b2} is the coded bit rate (bit frequency). Notice that the BPSK bandwidth is the same the ASK bandwidth. It may be noted that the higher-order spectral components are filtered out.

Problem 6.1
Given:

- Uncoded input bit rate: R_{b1} = 10 kb/s
- Code rate: r = 8/15
- Carrier frequency fc = 1 MHz
- Modulation: BPSK, QPSK, 8PSK, and 16PSK

Find:

(a) Coded bit rate R_{b2}
(b) BPSK bandwidth (BW)
(c) QPSK bandwidth (BW)
(d) 8PSK bandwidth (BW)
(e) 16PSK bandwidth (BW)

Solution:

(a) Coded bit rate: $R_{B2} = R_{B1}/r$ = 10 kb/(15/8) = 18.75 kb/s
(b) BPSK bandwidth: $BW = 2R_{B2}/1$ = 2 × 18.75 = 37.5 kHz
(c) QPSK bandwidth: $BW = 2R_{B2}/2$ = 2 × 18.75/2 = 18.75 kHz
(d) 8PSK bandwidth: $BW = 2R_{B2}/3$ = 2 × 18.75/3 = 12.5 kHz
(e) 16PSK bandwidth: $BW = 2R_{B2}/4$ = 2 × 18.755/4 = 9.375 kHz

NOTE: Higher-order PSK modulation is bandwidth efficient.

Problem 6.2
Given:

- Input bit rate R_{b1} = 10 kb/s
- Code rate r = 1/2
- Modulation: BPSK, QPSK, 8PSK, 16PSK

Find:

(a) Bit rate after coding R_{b2}
(b) Transmission bandwidth BW

Solution:

(a) $R_{b2} = R_{b1}/r$ = 10 kb/s × 2 = 20 kb/s
(b) Transmission bandwidth:

- BPSK $BW = 2R_{b2}$/Bits per Symbol = 2 × 20 kb/s/1 = 40 kHz
- QPSK $BW = 2R_{b2}$/Bits per symbol = 2 × 20 kb/s/2 = 20 kHz
- 8PSK $BW = 2R_{b2}$/Bits per symbol = 2 × 20 kb/s/3 = 13.33 kHz
- 16PSK $BW = 2R_{b2}$/Bits per symbol = 2 × 20 kb/s/4 = 10 kHz

NOTE: Higher-order PSK modulation is bandwidth efficient.

6.7 Conclusions

- This chapter presents PSK modulation and its attributes.
- Numerous illustrations are provided to show how phase of the carrier changes in discrete levels in accordance with the input digital signal, while the amplitude of the carrier remains the same.
- The Fourier transform is used to derive the spectral components and PSK bandwidth is calculated.
- These materials have been augmented by diagrams and associated waveforms to make them easier for readers to grasp.

References

1. Smith DR (1986) Digital transmission systems. Van Nostrand Reinhold Co. ISBN: 0442009178
2. Couch II LW (2001) Digital and analog communication systems, 7th edn. Prentice-Hall, Inc.: Englewood Cliffs. ISBN: 0-13-142492-0
3. Sklar B (1988) Digital communications fundamentals and applications. Prentice Hall
4. Faruque S (2016) Radio frequency channel coding made easy. Springer, ISBN: 978-3-319-21169-5
5. Clark GC et al (1981) Error correction coding for digital communications. Plenum press
6. Ungerboeck G (1982) Channel coding with multilevel/multiphase signals. IEEE Trans Inf Theory IT28:55–67
7. Lin S, Costello DJ Jr (1983) Error control coding: fundamentals and applications. Prentice-Hall, Inc., Englewood Cliffs
8. Blahut RE (1983) Theory and practice of error control codes. reading. Addison-Wesley, Massachusetts
9. van Lint JH (1992) Introduction to coding theory. GTM 86. 2nd ed. Springer. p 31. ISBN 3-540-54894-7
10. Mac Williams FJ, Sloane NJA (1977) The theory of error correcting codes. North-Holland. p 36. ISBN 0-444-85193-3
11. Huffman W, Pless V (2003) Fundamentals of error-correcting codes. Cambridge University Press. ISBN 978-0-521-78280-7
12. Ebert PM, Tong SY (1968) Convolutional Reed-Soloman codes, Bell Syst Tech, pp 729–742
13. Gallager RG (1968) Information theory and reliable communications. John Wiley, New York

14. Kohlenbero A, Forney GD Jr (1968) Convolutional coding for channels with memory, IEEE Trans. Information Theory IT-14, 618–626. 1968
15. M.A. Reddy, S. M, “Further results on convolutional codes derived from block codes. Inf Control 13:357–362
16. Reddy SM (1968) A class of linear convolutional codes for compound channels. Technical Report, Bell Telephone Laboratories, Holmdel, New Jersey

Chapter 7
N-Ary Coded Modulation

Topics

- Introduction
- N-ary Complementary Convolutional Coding and M-ary Modulation
- N-ary Complementary Orthogonal Coding and M-ary Modulation
- Coding Gain and Bandwidth
- Conclusions

7.1 Introduction

N-ary coded modulation is a multi-level channel coding and multi-level modulation technique, where instead of coding one bit at a time two or more bits are encoded simultaneously and then modulated by means of M-ary modulation. In its most basic construction, the input serial data is converted into several parallel streams. These parallel bit streams, now reduced in speed, are mapped into a bank of *N* unique convolutional or orthogonal codes. The coded information bits are then modulated by an M-ary PSK modulator and transmitted through a channel. At the receiver, the decoder recovers the data by means of code correlation. A lookup table at the receiver contains the input/output bit sequences. Upon receiving an encoded data pattern, the receiver validates the received data pattern by means of code correlation.

In this chapter, we will present the key concept, underlying principles and practical application of N-ary coded modulation schemes, offering spectrum efficiency with improved error correction capabilities. Construction of N-ary coded modulation schemes based on convolutional and orthogonal codes will be presented to illustrate the concept.

© The Author(s) 2017

S. Faruque, *Radio Frequency Modulation Made Easy*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-41202-3_7

7.2 N-Ary Convolutional Coding and M-Ary Modulation

7.2.1 Background

In convolutional coding, a sequence of data signals is transformed into a longer sequence that contains enough redundancy to protect the data [1–6]. This type of error control is also classified as forward error control coding (FECC), because these methods are often used to correct errors that are caused by channel noise. In a typical convolutional encoder, k information bits enter into the encoder sequentially. The convolutional encoder generates n parity bits as encoded bits ($n > k$). The code rate is defined as $r = k/n$. The coded information bits are modulated and transmitted through a channel.

Another class of convolutional codes, known as parallel concatenated codes or turbo codes [7–9], is popular in wireless communications because of its superior error control capabilities. Yet, turbo codes are limited to two convolutional codes connected in parallel, where delays are introduced due to concatenation and interleaving, thereby limiting high-speed data communications in wireless communications. In addition, it is also too complex to realize.

In this section, a modified technique, defined as N-ary complementary convolutional coding, along with M-ary modulation, is presented to overcome these problems. N-ary complementary convolutional coding is a multi-level convolutional code, where more than two convolutional codes can be used simultaneously to further enhance coding gain without concatenation and interleaving.

7.2.2 Generation of Complementary Convolutional Codes

In a typical convolutional encoder, k information bits enter into the encoder sequentially. The encoder generates n parity bits as encoded bits ($n > k$). The code rate is defined as $r = k/n$. The proposed N-ary convolutional codes are a block of n-convolutional codes and their complements.

In its most basic construction, as shown in Fig. 7.1, a typical convolutional encoder is taken as the basis to generate a block of n-convolutional codes. Next, each convolutional code is inverted to construct a block of n-antipodal codes. Therefore, a block of N-ary convolutional code has n-convolutional codes and n-antipodal codes, for a total of $2n = N$ complementary convolutional codes. In short, we shall define this as N-ary complementary convolutional codes.

Let us consider Fig. 7.1 once again and assume that the input bit sequence to the encoder is:

$$m(t) = 1\ 0\ 1 \tag{7.1}$$

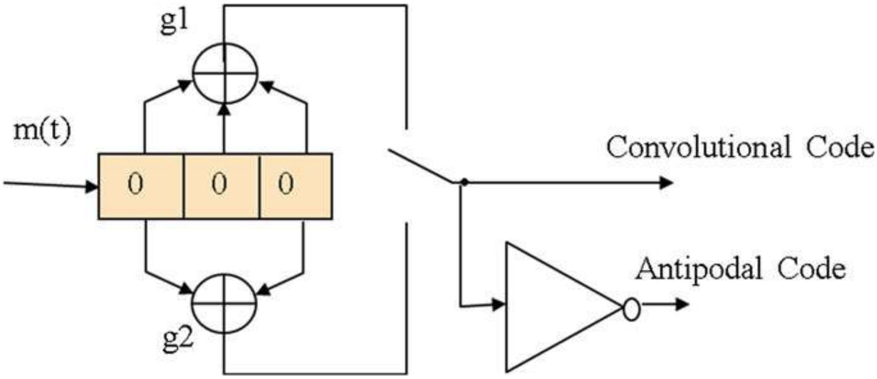

Fig. 7.1 Generation of N-ary complimentary convolutional codes

Which is described by the following polynomial?

$$\begin{aligned} m\,(X) &= 1 + 0X + 1X^2 \\ &= 1 + X^2 \end{aligned} \tag{7.2}$$

The encoder is described by the following generator polynomials:

$$\begin{aligned} g1(X) &= 1 + X + X^2 \\ g2(X) &= 1 + X^2 \end{aligned} \tag{7.3}$$

Then, the product of polynomials for the encoder can be described as follows:

$$\begin{aligned} m\,(X)g1(X) &= (1 + X^2)\,(1 + X + X^2) = 1 + X + X^3 + X^4 \\ m\,(X)g2(X) &= (1 + X^2)\,(1 + X^2) = 1 + X^4 \end{aligned} \tag{7.4}$$

with $X^2 + X^2 = 0$, the output bit sequence can be found as $U(X) = m(X)g1(\mathrm{X})$ multiplexed with $m(X)g2(X)$, where $m(X)$ is the input bit sequence [7]. We write the above two equations as follows:

$$\begin{array}{lcccccccccc} \mathrm{m(X)g1(X)} & = & 1 & + & 1\mathrm{X} & + & 0\mathrm{X}^2 & + & 1\mathrm{X}^3 & + & 1\mathrm{X}^4 \\ \mathrm{m(X)g2(X)} & = & 1 & + & 0\mathrm{X} & + & 0\mathrm{X}^2 & + & 0\mathrm{X}^3 & + & 1\mathrm{X}^4 \\ \hline \mathrm{U1U2} & = & (1,1) & + & (1,0) & + & (0,0) & + & (1,0) & + & (1,1) \end{array} \tag{7.5}$$

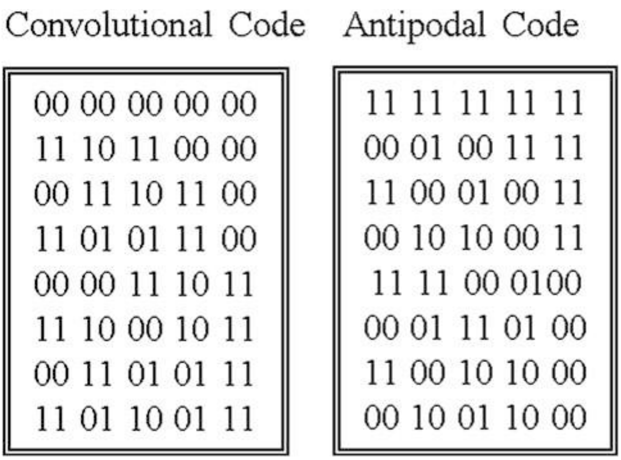

Convolutional Code	Antipodal Code
00 00 00 00 00	11 11 11 11 11
11 10 11 00 00	00 01 00 11 11
00 11 10 11 00	11 00 01 00 11
11 01 01 11 00	00 10 10 00 11
00 00 11 10 11	11 11 00 0100
11 10 00 10 11	00 01 11 01 00
00 11 01 01 11	11 00 10 10 00
11 01 10 01 11	00 10 01 10 00

Fig. 7.2 N-ary complimentary convolutional code blocks

Taking only the coefficients, we obtain the desired convolutional code sequence corresponding to the input bit sequence $m(T) = 101$ as follows:

$$\text{Convolutional Code} = 1\,1 \quad 1\,0 \quad 0\,0 \quad 1\,0 \quad 1\,1 \tag{7.6}$$

The corresponding antipodal code is obtained simply by inverting the convolutional code as shown in the figure. This is given by:

$$\text{Antipodal Code} = 0\,0 \quad 0\,1 \quad 1\,1 \quad 0\,1 \quad 0\,0 \tag{7.7}$$

As can be seen, each 3-bit data corresponds to a unique 10-bit convolutional code. Antipodal code is just the inverse of the convolutional codes. Since a 3-bit sequence has 8 combinations, the above convolutional encoder generates 8 unique convolutional codes and 8 unique antipodal codes, for a total of 16 complementary convolutional codes. These code blocks are displayed in Fig. 7.2.

7.2.3 2-Ary Convolutional Coding with QPSK Modulation

A 2-ary ($N = 2$) coded QPSK modulator can be constructed by inverse multiplexing the incoming traffic into 6-parallel streams as shown in Fig. 7.3. These bit streams, now reduced in speed by a factor of 6, are partitioned into two 8 × 3 data blocks. The first 8 × 3 data block maps the 8 × 10 convolutional code block, and the next 8 × 3 data block maps the next 8 × 10 antipodal code block. These code blocks are stored in two 8 × 10 ROMs. The output of each ROM is a unique 10-bit

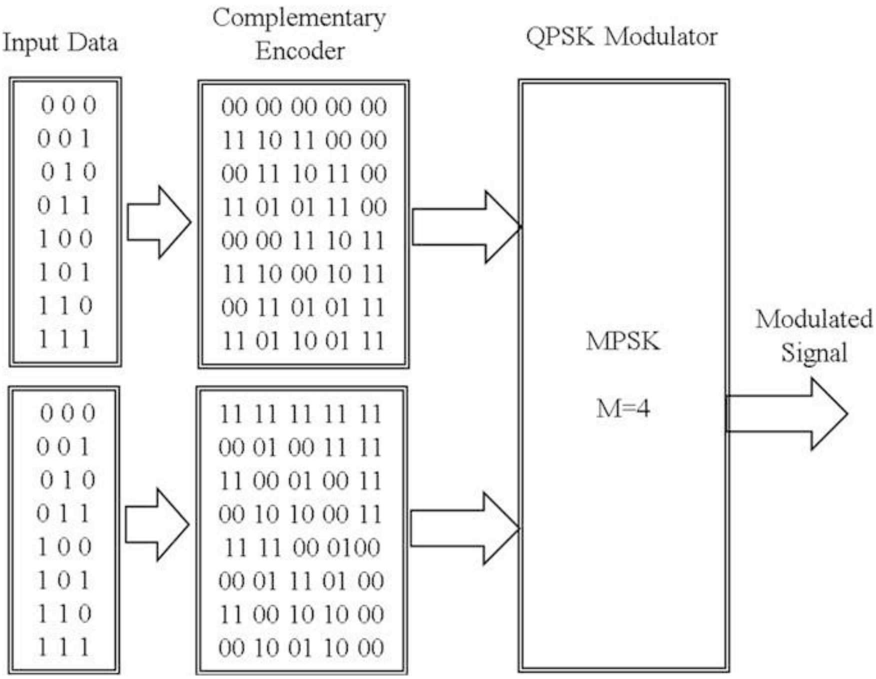
Input Data
Complementary Encoder
QPSK Modulator
000
001
010
011
100
101
110
111
00 00 00 00 00
11 10 11 00 00
00 11 10 11 00
11 01 01 11 00
00 00 11 10 11
11 10 00 10 11
00 11 01 01 11
11 01 10 01 11
000
001
010
011
100
101
110
111
11 11 11 11 11
00 01 00 11 11
11 00 01 00 11
00 10 10 00 11
11 11 00 0100
00 01 11 01 00
11 00 10 10 00
00 10 01 10 00
MPSK
M=4
Modulated Signal

Fig. 7.3 2-ary convolutional coding with QPSK modulation

convolutional/antipodal code, which is modulated by a QPSK modulator using the same carrier frequency. The code rate is given by:

$$\text{Code Rate}: r = 6/10 = 3/5 \tag{7.8}$$

Since there are two convolutional waveforms (one convolutional and one antipodal), the number of errors that can be corrected is doubled. Moreover, the transmission bandwidth is also reduced, which is given by:

Transmission bandwidth: $BW = 2R_{b2}$/bits per symbol Hz

$$B = 2R_{b2}/2 = R_{b2} \tag{7.9}$$

where $R_{b2} = R_{b1}/r$. R_{b2} is the coded bit rate, and R_{b1} is the uncoded bit rate.

7.2.4 4-Ary Convolutional Coding with 16PSK Modulation

A 4-ary ($N = 4$) coded 16PSK modulator can be constructed by inverse multiplexing the incoming traffic into 8-parallel streams as shown in Fig. 7.4. These bit streams, now reduced in speed by a factor of 8, are partitioned into four 4 × 2 data

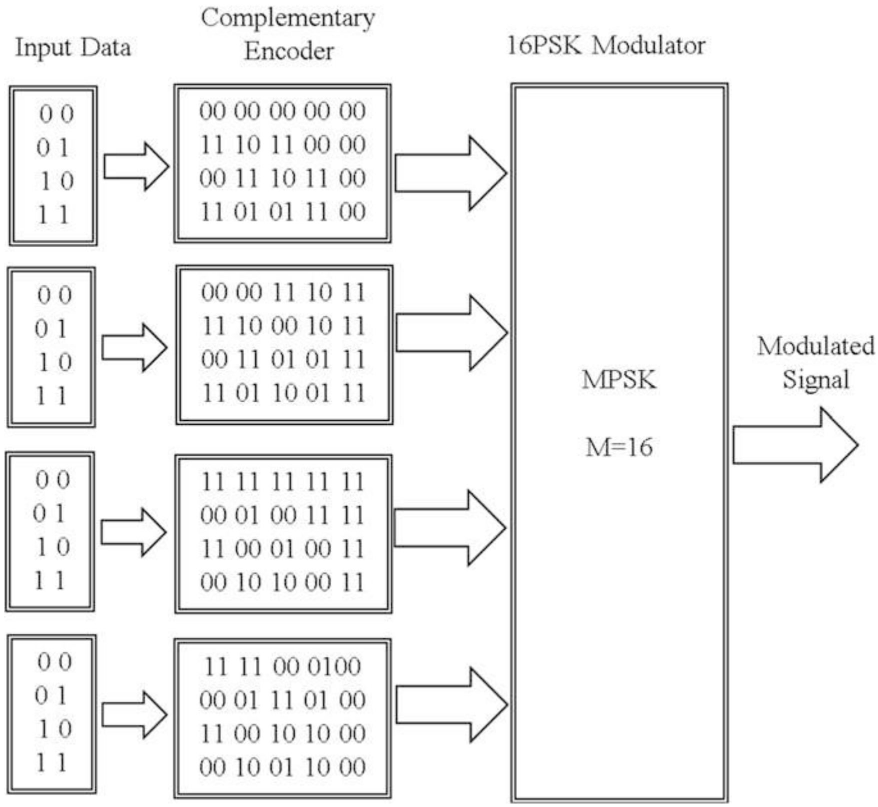
Input Data
Complementary Encoder
16PSK Modulator
0 0
0 1
1 0
1 1
00 00 00 00 00
11 10 11 00 00
00 11 10 11 00
11 01 01 11 00
0 0
0 1
1 0
1 1
00 00 11 10 11
11 10 00 10 11
00 11 01 01 11
11 01 10 01 11
0 0
0 1
1 0
1 1
11 11 11 11 11
00 01 00 11 11
11 00 01 00 11
00 10 10 00 11
0 0
0 1
1 0
1 1
11 11 00 0100
00 01 11 01 00
11 00 10 10 00
00 10 01 10 00
MPSK
M=16
Modulated Signal

Fig. 7.4 4-ary convolutional coding with 16PSK modulation

blocks. Each 4 × 2 data block maps corresponding 4 × 10 convolutional code blocks. These code blocks are stored in four 4 × 10 ROMs. The output of each ROM is a unique 10-bit convolutional/antipodal code, which is modulated by a 16PSK modulator using the same carrier frequency. The code rate is given by:

$$\text{Code Rate:} r = 8/10 = 4/5 \tag{7.10}$$

Since there are four convolutional waveforms (two convolutional and two antipodal), the number of errors that can be corrected is quadrupled. Moreover, the transmission bandwidth is further reduced, which is given by:

Transmission bandwidth: $BW = 2R_{b2}/\text{bits per symbol Hz}$

$$= 2R_{b2}/4 = R_{b2}/2 \tag{7.11}$$

where $R_{b2} = R_{b1}/r$. R_{b2} is the coded bit rate, and R_{b1} is the uncoded bit rate.

7.3 N-Ary Convolutional Decoder

7.3.1 Correlation Receiver

Decoding is a process of code correlation. In this process, the receiver compares the received data with the expected data set to recover the actual data. The expected data is stored into a lookup table [see Table 7.1]:

- The lookup table at the receiver contains the input/output bit sequences.
- For *m* = 3, there are 8 possible output combinations of 3-bit data.
- For each combination of 3-bit data, there is a unique encoded 10-bit data (see table).
- The receiver receives one of 8 output sequences.
- Upon receiving an encoded data pattern, the receiver validates the received data pattern by means of code correlation.

Let us examine the correlation process using the following example:

- The input bit pattern *m* = 0 1 1
- Encoded transmit data: *U* = 00 11 01 01 11
- Received data with errors *U** = 00 11 01 01 **00**

Notice that the last two bits are in error, identified in bold. Now, let us determine how the receiver recovers the correct data, where the actual input data is *m* = 0 1 1. This is a correlation process, requiring several tests to validate the actual data. The correlation process is described below:

Test 0:

This test compares the received data with the 1st row of data stored in the lookup table and counts the number of positions it does not match. This is accomplished by MOD2 operation (EXOR operation). The result is presented below:

- Received data: 00 11 01 01 **00**
- 1st row of data in the lookup table: 00 00 00 00 00
- Mod-2 Add: 00 11 01 01 00

Table 7.1 Lookup table a

Input (*m*)	Output (*U*)
0. 0 0 0	00 00 00 00 00
1. 0 0 1	00 00 11 10 11
2. 0 1 0	00 11 10 11 00
3. 0 1 1	00 11 01 01 11
4. 1 0 0	11 10 11 00 00
5. 1 0 1	11 10 00 10 11
6. 1 1 0	11 01 01 11 00
7. 1 1 1	11 01 10 01 11

- Correlation value = 4 (count the number of 1's in MOD2 Add)
- Verdict: No match, continue search.

Test-0: Look Up Table

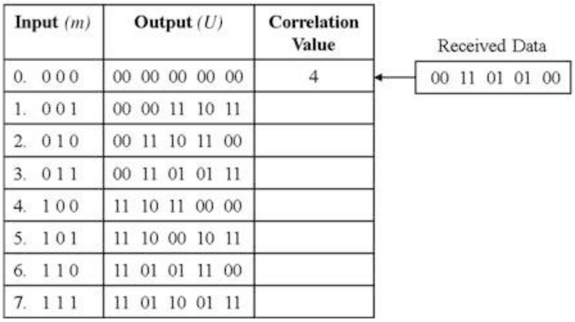

Input (*m*)	Output (*U*)	Correlation Value
0. 0 0 0	00 00 00 00 00	4
1. 0 0 1	00 00 11 10 11	
2. 0 1 0	00 11 10 11 00	
3. 0 1 1	00 11 01 01 11	
4. 1 0 0	11 10 11 00 00	
5. 1 0 1	11 10 00 10 11	
6. 1 1 0	11 01 01 11 00	
7. 1 1 1	11 01 10 01 11	

Received Data

00 11 01 01 00

Test 1:

This test compares the received data with the 2nd row of data stored in the lookup table and counts the number of positions it does not match. This is accomplished by MOD2 operation (EXOR operation). The result is presented below:

- Received data: 00 11 01 01 00
- 2nd row of data in the lookup table: 00 00 11 10 11
- Mod-2 Add: 00 11 10 111 11
- Correlation value = 7
- Verdict: No match, continue search

Test 1: Look Up Table

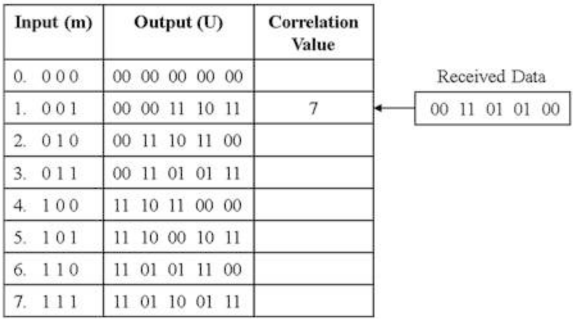

Input (m)	Output (U)	Correlation Value
0. 0 0 0	00 00 00 00 00	
1. 0 0 1	00 00 11 10 11	7
2. 0 1 0	00 11 10 11 00	
3. 0 1 1	00 11 01 01 11	
4. 1 0 0	11 10 11 00 00	
5. 1 0 1	11 10 00 10 11	
6. 1 1 0	11 01 01 11 00	
7. 1 1 1	11 01 10 01 11	

Received Data

00 11 01 01 00

The remaining tests may be conducted by using the same method to reach the final verdict as given below:

Final Verdict:

Collect the correlation values and validate the data that indicates the lowest correlation value. This is presented in the following lookup table.

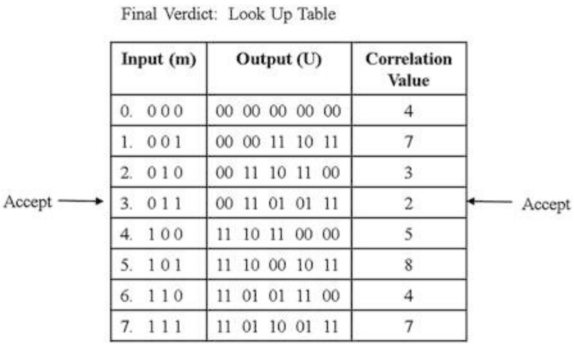

Final Verdict: Look Up Table

Input (m)	Output (U)	Correlation Value
0. 0 0 0	00 00 00 00 00	4
1. 0 0 1	00 00 11 10 11	7
2. 0 1 0	00 11 10 11 00	3
3. 0 1 1	00 11 01 01 11	2
4. 1 0 0	11 10 11 00 00	5
5. 1 0 1	11 10 00 10 11	8
6. 1 1 0	11 01 01 11 00	4
7. 1 1 1	11 01 10 01 11	7

Accept
Accept

In examining the above table, we find that:

- The lowest correlation value is 2.
- The corresponding data is $m = 0\ 1\ 1$.
- This is the data which has been transmitted to the receiver.

7.3.2 *Error Correction Capabilities of N-Ary Convolutional Codes*

An n-bit convolutional code is generated by a rate ½ encoder of constraint length K where

$$n = 4K - 2 \tag{7.12}$$

In the above equation, K is the length of the shift register and the factor −2 is due to the initial content of the shift register, which is 000. For example, a rate ½ convolutional encoder with $K = 3$ generates 00 when the initial content of the shift register is 000. Therefore, the value of n is reduced by 2. Therefore, each 3-bit data sequence corresponds to a unique 10-bit convolutional code. The antipodal code is just the inverse of the convolutional codes. Since a 3-bit sequence has 8 combinations, the above convolutional encoder generates 8 unique convolutional codes and 8 unique antipodal codes, for a total of 16 complementary convolutional codes.

These code blocks are displayed in Fig. 7.2. As can be seen, each code is unique and there is a minimum distance $d_{\min}$ between any two convolutional codes where

$$d_{\min} \leq n/2 \tag{7.13}$$

This distance property can be used to detect an impaired received code by setting a threshold midway between two convolutional codes. It can be shown that an n-bit convolutional code can correct the errors, where

Table 7.2 Error correction capabilities of N-ary convolutional codes

Constraint length (K)	Number of errors corrected if the modulation scheme is (BPSK)	Number of errors corrected if the modulation scheme is (QPSK)	Number of errors corrected if the modulation scheme is (16PSK)
3	2	4	8
5	4	8	16
7	6	12	24
9	8	16	32

$$t \leq d_{\min}/2 \leq n/4 \leq (4K - 2)/4 \tag{7.14}$$

where

- t = Number of errors that can be corrected by means of a single convolutional code
- n = Code length
- K = Constraint length

The proposed N-ary coded modulation is a multi-level channel coding technique, where instead of coding one bit at a time two or more bits are encoded simultaneously. Therefore, we modify the above equation to obtain

$$N\,t \leq N\,d_{\min}/2 \leq Nn/4 \leq N(4K - 2)/4 \tag{7.15}$$

Table 7.2 displays error correction capabilities of N-ary convolutional codes.

7.4 N-Ary Orthogonal Coding and M-Ary Modulation

7.4.1 Background

Orthogonal codes are used in CDMA cellular communications for spectrum spreading and user ID [10–13]. The use of orthogonal codes for forward error control coding has also been investigated by a limited number of authors [14–16]. This chapter presents a method of channel coding based on N-ary orthogonal codes [17–20]. N-ary orthogonal coding is a multi-level channel coding technique, where instead of coding one bit at a time two or more bits are encoded simultaneously. This type of channel coding is bandwidth efficient and further enhances the coding gain with bandwidth efficiency. In the proposed scheme, the input serial data is converted into several parallel streams. These parallel bit streams, now reduced in speed, are mapped into N-ary orthogonal codes. The coded information bits are then modulated by means of MPSK modulator and transmitted through a channel. At the receiver, the decoder recovers the data by means of code correlation. The proposed coded modulation scheme is bandwidth efficient and offers a coding gain near the Shannon's limit.

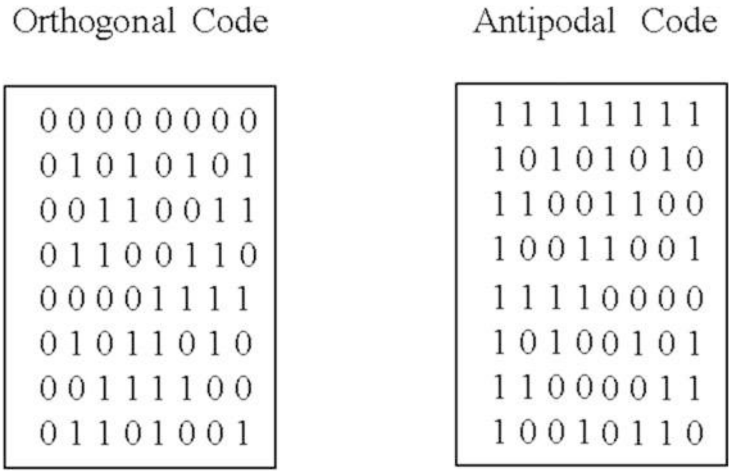

Orthogonal Code	Antipodal Code
0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
0 1 0 1 0 1 0 1	1 0 1 0 1 0 1 0
0 0 1 1 0 0 1 1	1 1 0 0 1 1 0 0
0 1 1 0 0 1 1 0	1 0 0 1 1 0 0 1
0 0 0 0 1 1 1 1	1 1 1 1 0 0 0 0
0 1 0 1 1 0 1 0	1 0 1 0 0 1 0 1
0 0 1 1 1 1 0 0	1 1 0 0 0 0 1 1
0 1 1 0 1 0 0 1	1 0 0 1 0 1 1 0

Fig. 7.5 Bi-orthogonal code set for $n = 8$. An 8-bit orthogonal code has 8 orthogonal codes and 8 antipodal codes for a total of 16 bi-orthogonal codes

7.4.2 *Orthogonal Codes*

Orthogonal codes are binary-valued and have equal numbers of 1's and 0's. Antipodal codes, on the other hand, are just the inverse of orthogonal codes. Antipodal codes are also orthogonal among them. Therefore, an n-bit orthogonal code has n-orthogonal codes and n-antipodal codes, for a total of 2n bi-orthogonal codes. For example, an 8-bit orthogonal code has 8 orthogonal codes and 8 antipodal codes, for a total of 16 bi-orthogonal codes, as shown in Fig. 7.5 [12]. Similarly, a 16-bit orthogonal code has 16 orthogonal codes and 16 antipodal codes for a total of 32 bi-orthogonal codes, as shown in Fig. 7.6.

We note that orthogonal codes are essentially (n, k) block codes, where a k-bit data set is represented by a unique n-bit orthogonal code ($k < n$). We now show that code rates such as rate ½, rate ¾, and rate 1 are indeed available out of orthogonal codes. The principle is presented below.

7.4.3 *2-Ary Orthogonal Coding with QPSK Modulation*

A 2-ary orthogonal coded QPSK modulator with $n = 8$, having 16 complementary orthogonal codes, can be constructed by inverse multiplexing the incoming traffic into 6-parallel streams, as shown in Fig. 7.7. These bit streams, now reduced in speed by a factor of 6, are partitioned into two data blocks. Each data block is mapped into an 8 × 8 code block, as depicted in the figure.

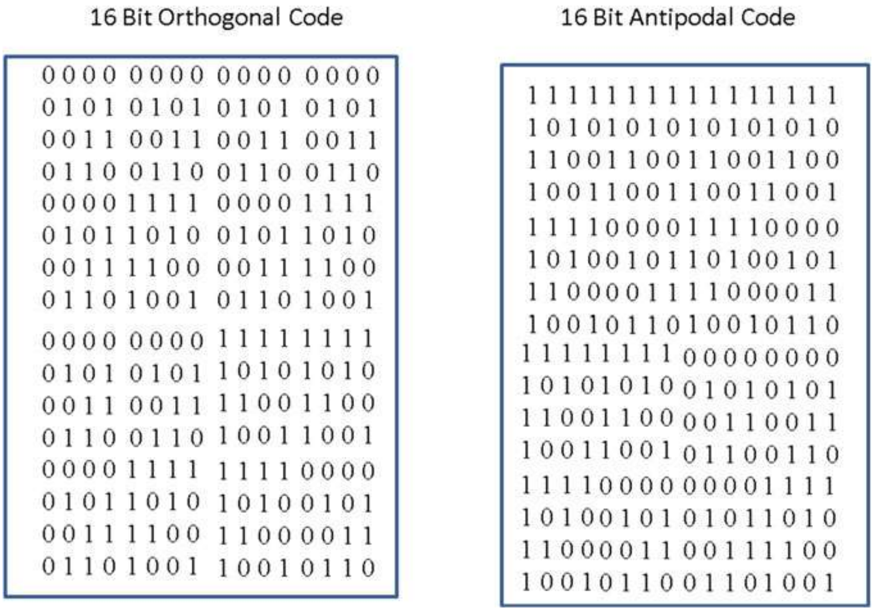

16 Bit Orthogonal Code

```
0000 0000 0000 0000
0101 0101 0101 0101
0011 0011 0011 0011
0110 0110 0110 0110
00001111 00001111
01011010 01011010
00111100 00111100
01101001 01101001
0000 0000 11111111
0101 0101 10101010
0011 0011 11001100
0110 0110 10011001
00001111 11110000
01011010 10100101
00111100 11000011
01101001 10010110
```

16 Bit Antipodal Code

```
1111111111111111
1010101010101010
1100110011001100
1001100110011001
1111000011110000
1010010110100101
1100001111000011
1001011010010110
11111111 00000000
10101010 01010101
11001100 00110011
10011001 01100110
1111000000001111
1010010101011010
1100001100111100
1001011001101001
```

Fig. 7.6 Bi-orthogonal code set for *n* = 16. A 16-bit orthogonal code has 16 orthogonal codes and 16 antipodal codes for a total of 32 bi-orthogonal codes

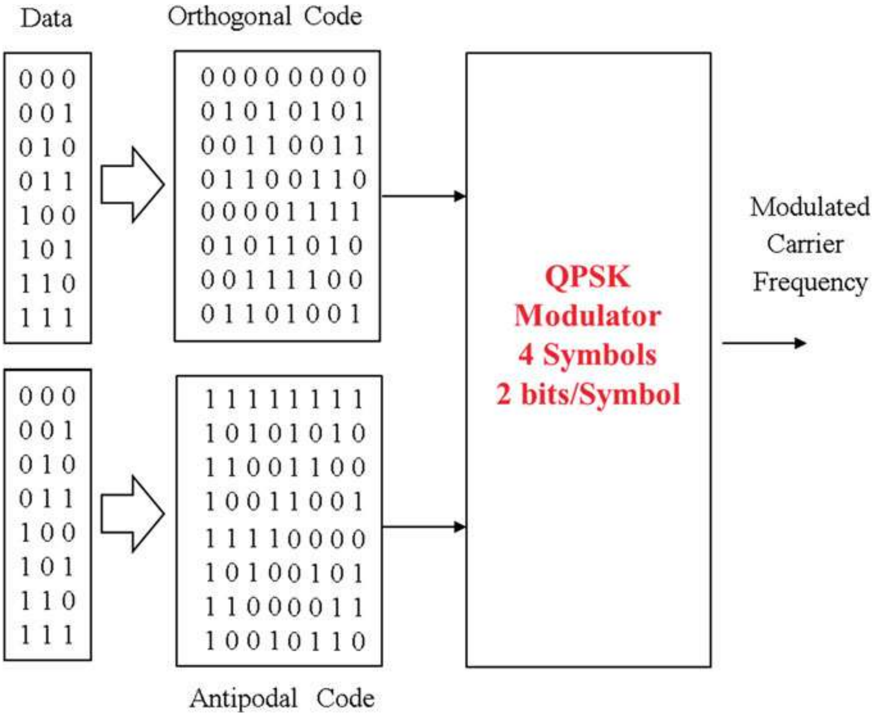

Data
Orthogonal Code
000
001
010
011
100
101
110
111
00000000
01010101
00110011
01100110
00001111
01011010
00111100
01101001
000
001
010
011
100
101
110
111
11111111
10101010
11001100
10011001
11110000
10100101
11000011
10010110
Antipodal Code
QPSK
Modulator
4 Symbols
2 bits/Symbol
Modulated
Carrier
Frequency

Fig. 7.7 2-ary orthogonal coding with $N = 2$

According to the communication scheme, when a 3-bit data pattern needs to be transmitted, the corresponding orthogonal/antipodal code is transmitted instead, requiring a QPSK modulator. The modulated waveforms are in orthogonal space and have fewer errors. The code rate is given by:

$$\text{Code Rate: } r = 6/8 = 3/4$$

Since there are two orthogonal waveforms (one orthogonal and one antipodal), the number of errors that can be corrected is doubled. Moreover, the transmission bandwidth is also reduced, which is given by:

Transmission Bandwidth: $BW = 2R_{b2}$/bits per symbol Hz

$$= 2R_{b2}/2 = R_{b2}\,\text{Hz} \tag{7.16}$$

where $R_{b2} = R_{b1}/r$. R_{b2} is the coded bit rate and R_{b1} is the uncoded bit rate.

7.4.4 4-Ary Orthogonal Coding with 16PSK Modulation

A 4-ary orthogonal coded 16PSK modulator, with $n = 8$, having 16 complementary orthogonal codes, can be constructed by inverse multiplexing the incoming traffic into 8-parallel streams, as shown in Fig. 7.8. The bit streams, now reduced in speed by a factor of 8, are partitioned into four data blocks. Each data block is mapped into a 4×8 code block as depicted in the figure. According to the communication scheme, when a 2-bit data pattern needs to be transmitted, the corresponding orthogonal/antipodal code is transmitted instead, requiring a 16PSK modulator. The modulated waveforms are in orthogonal space and have fewer errors. The code rate is given by:

$$\text{Code Rate: } r = 8/8 = 1 \tag{7.17}$$

This is achieved without bandwidth expansion.

Since there are four orthogonal waveforms (two orthogonal and two antipodal), the number of errors that can be corrected is quadrupled. Moreover, the transmission bandwidth is further reduced, which is given by:

Transmission bandwidth: $BW = 2R_{b2}$/bits per symbol Hz

$$= 2R_{b2}/4 = R_{b2}/2\ \text{Hz} \tag{7.18}$$

where $R_{b2} = R_{b1}/r$. R_{b2} is the coded bit rate, and R_{b1} is the uncoded bit rate.

7.4.5 2-Ary Orthogonal Decoding

Decoding is a correlation process. Notice that the entire biorthogonal code block is partitioned into two code blocks. Each code block represents a data block as shown in Fig. 7.9. Upon receiving an impaired code, the receiver compares it with each

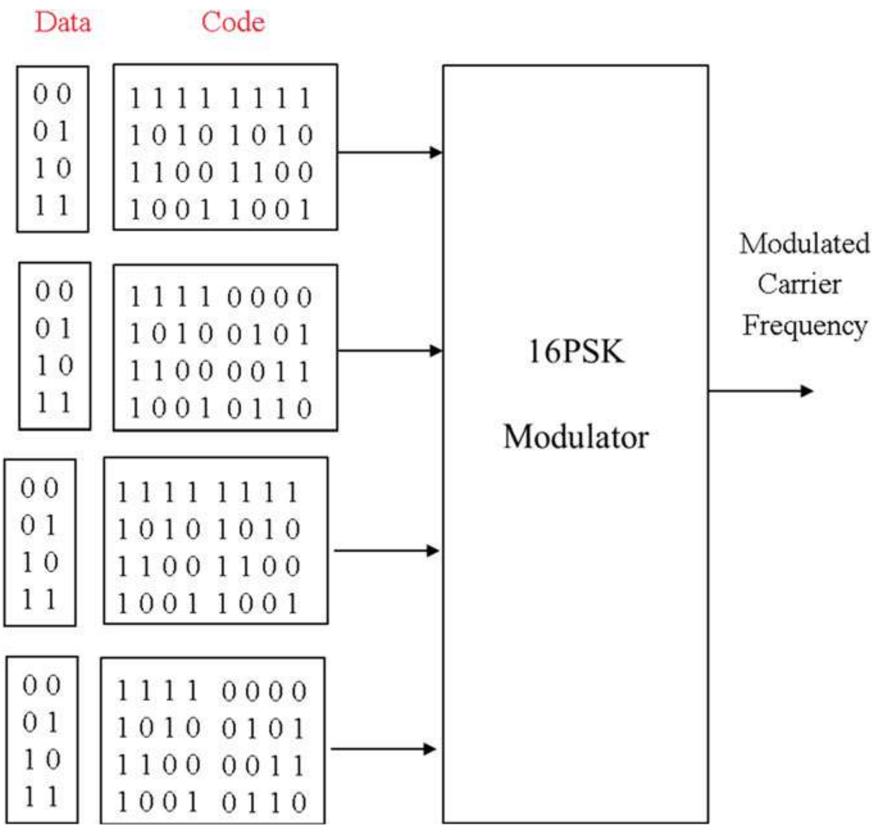
Data
Code
00
01
10
11
1111 1111
1010 1010
1100 1100
1001 1001
00
01
10
11
1111 0000
1010 0101
1100 0011
1001 0110
00
01
10
11
1111 1111
1010 1010
1100 1100
1001 1001
00
01
10
11
1111 0000
1010 0101
1100 0011
1001 0110
16PSK
Modulator
Modulated
Carrier
Frequency

Fig. 7.8 N-ary orthogonal coding with $N = 4$

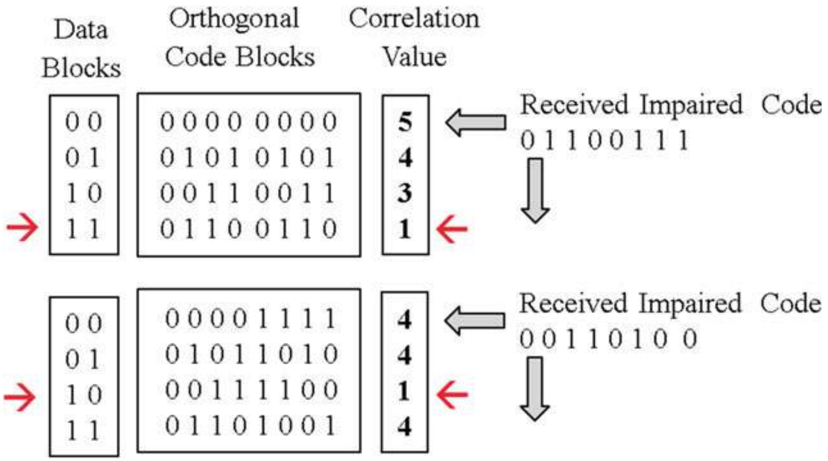
Data
Blocks
Orthogonal
Code Blocks
Correlation
Value
00
01
10
11
0000 0000
0101 0101
0011 0011
0110 0110
5
4
3
1
Received Impaired Code
0 1 1 0 0 1 1 1
00
01
10
11
00001111
01011010
00111100
01101001
4
4
1
4
Received Impaired Code
0 0 1 1 0 1 0 0

Fig. 7.9 2-ary orthogonal decoding

entry in the code block and appends a correlation value for each comparison. A valid code is declared when the closest approximation is achieved. As can be seen, the minimum correlation value in each block is 1 as depicted in the figure. The corresponding data is 1 1, 1 0, 1 1, 1 0, respectively.

Since there are two orthogonal waveforms (one orthogonal and one antipodal), the number of errors that can be corrected is given by 2 × 1 = 2. Moreover, the bandwidth is also reduced, as presented below:

$$BW = 2Rb2/\text{bits per symbol} = 2_{Rb2}/2 = \mathrm{R}_{b2} \tag{7.19}$$

7.4.6 4-Ary Orthogonal Decoding

Decoding is a correlation process. Notice that the entire biorthogonal code block is partitioned into four code blocks. Each code block represents a data block, as shown in Fig. 7.10. Upon receiving an impaired code, the receiver compares it with each entry in the code block and appends a correlation value for each comparison. A valid code is declared when the closest approximation is achieved. As can be seen, the minimum correlation value in each block is 1 as depicted in the figure. The corresponding data is 1 1, 1 0, 1 1, 1 0, respectively.

Since there are four orthogonal waveforms (two orthogonal and two antipodal), the number of errors that can be corrected is given by 4 × 1 = 4. Moreover, the bandwidth is further reduced as presented below:

$$BW = 2Rb2/\text{bits per symbol} = 2_{Rb2}/4 = R_{b2}/2 \tag{7.20}$$

7.4.7 Error Correction Capabilities of N-Ary Orthogonal Codes

Error correction capabilities of orthogonal codes have been discussed earlier [16]. We present it again for convenience. An n-bit orthogonal code has $n/2$ 1's and $n/2$ 0's; i.e., there are $n/2$ positions where 1's and 0's differ. Therefore, the distance between two orthogonal codes is $d = n/2$. This distance property can be used to detect an impaired received code by setting a threshold midway between two orthogonal codes as shown in Fig. 7.11, where the received code is shown as a dotted line. This is given by:

$$d_{th} = \frac{n}{4} \tag{7.21}$$

where n is the code length and d_{th} is the threshold, which is midway between two valid orthogonal codes. Therefore, for the given 8-bit orthogonal code, we have

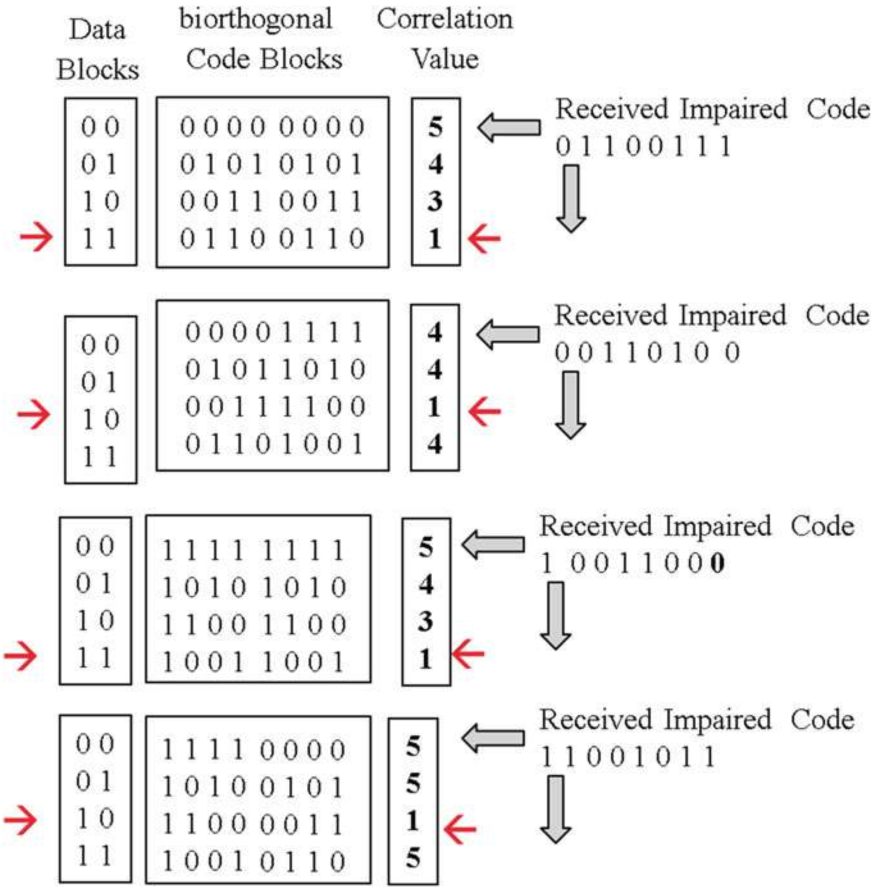
Data Blocks
biorthogonal Code Blocks
Correlation Value
00
01
10
11
0000 0000
0101 0101
0011 0011
0110 0110
5
4
3
1
Received Impaired Code
01100111
00
01
10
11
00001111
01011010
00111100
01101001
4
4
1
4
Received Impaired Code
00110100
00
01
10
11
1111 1111
1010 1010
1100 1100
1001 1001
5
4
3
1
Received Impaired Code
1 00110000
00
01
10
11
1111 0000
1010 0101
1100 0011
1001 0110
5
5
1
5
Received Impaired Code
11001011

Fig. 7.10 4-ary orthogonal decoding

Fig. 7.11 Distance properties of orthogonal codes

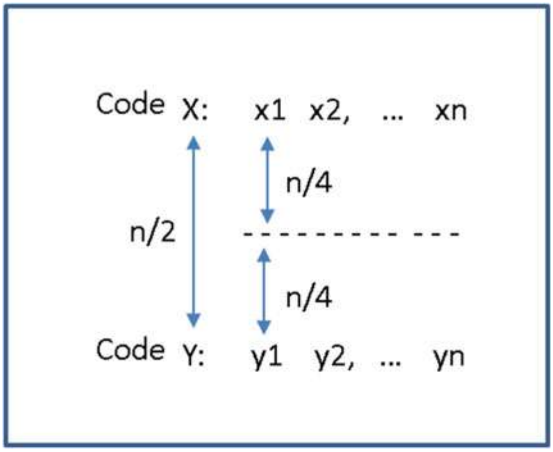
Code X: x1 x2, ... xn
n/4
n/2
n/4
Code Y: y1 y2, ... yn

$d_{th} = 8/4 = 2$. This mechanism offers a decision process, where the incoming impaired orthogonal code is examined for correlation with the neighboring codes for a possible match.

The received code is examined for correlation with the neighboring codes for a possible match. The acceptance criterion for a valid code is that an n-bit comparison must yield a good autocorrelation value; otherwise, a false detection will occur. The following correlation process governs this, where an impaired orthogonal code *i*s compared with a pair of n-bit orthogonal codes to yield:

$$R(x, y) = \sum_{i=1}^{n} x_i y_i \geq (n - d_{th}) + 1 \tag{7.22}$$

where x and y are two n-bit orthogonal codes, $R(x, y)$ is the autocorrelation function, n is the code length, and d_{th} is the threshold as defined earlier. Since the threshold is in the midway between two valid codes, an additional 1-bit offset is added to Eq. 7.21 for reliable detection. The average number of errors that can be corrected by means of this process can be estimated by combining Eq. 7.21 and Eq. 7.22, yielding:

$$t = n - R(x, y) = \frac{n}{4} - 1 \tag{7.23}$$

In the above equation, t is the number of errors that can be corrected by means of an n-bit orthogonal code. For example, a single-error-correcting orthogonal code can be constructed by means of an 8-bit orthogonal code ($n = 8$). Similarly, a three-error-correcting orthogonal code can be constructed by means of a 16-bit orthogonal code ($n = 16$) and so on. Table 7.3 shows a few orthogonal codes and the corresponding error-correcting capabilities.

Problem 7.1 This problem relates to N-ary convolutional codes and M-ary PSK modulation.

Given:

- Input bit rate before coding: $R_{b1} = 10$ kb/s
- $K = 3$, Rate ½ complementary convolutional codes
- 4-ary convolutional coding ($N = 4$)

Table 7.3 Error correction capabilities of N-ary orthogonal codes

Code length n	Number of errors corrected per code t	Number of errors corrected by N-ary code Nt $N = 1, 2, 4, 8, \ldots$
8	1	N
16	3	3 N
32	7	7 N
64	15	15 N

Find:

(a) Modulation level M
(b) Code rate r
(c) Bit rate after coding R_{b2}
(d) Transmission bandwidth BW
(e) Number of errors corrected

Solution:

(a) A 4-ary ($N = 4$) coded modulator can be constructed by inverse multiplexing the incoming traffic into 8-parallel stream. These bit streams, now reduced in speed by a factor of 8, are partitioned into four 4 × 2 data blocks. Each 4 × 2 data block maps a corresponding 4 × 10 convolutional code blocks. These code blocks are stored in four 4 × 10 ROMs. The output of each ROM is a unique 10-bit convolutional/antipodal code, which is modulated by a $2^4 = 16$ PSK modulator using the same carrier frequency. Therefore, $M = 16$.

(b) The code rate is given by:

$$\text{Code Rate:}\, r = 8/10 = 4/5$$

(c) Bit rate after coding $R_{b2} = R_{b1}/r = (5/4)$ 10 kb/s = 12.5 kb/s

(d) Transmission bandwidth: $BW = 2R_{b2}$/bits per symbol Hz

$$= 2R_{b2}/4 = R_{b2}/2 = (12.50\,\text{kb/s})/2 = 6.25\,\text{kHz}$$

(e) Since there are four convolutional waveforms (two convolutional and two antipodal), the number of errors that can be corrected is quadruple,
$4t = 4 \times 2 = 8$ [Each code corrects 2 errors.]

Problem 7.2 This example relates to N-ary orthogonal codes and M-ary PSK modulation.

Given:

- Input bit rate before coding: $R_{b1} = 10$ kb/s
- $n = 8$ orthogonal codes having 16 complementary orthogonal codes
- 4-ary orthogonal Coding

Find:

(a) Modulation level M
(b) Code rate r
(c) Bit rate after coding R_{b2}
(d) Transmission bandwidth BW
(e) Number of errors corrected

Solution:

(a) A 4-ary orthogonal coded modulator, with $n = 8$, having 16 complementary orthogonal codes, can be constructed by inverse multiplexing the incoming traffic into 8-parallel streams. The bit streams, now reduced in speed by a factor of 8, are partitioned into four data blocks. Each data block is mapped into a 4×8 code block. According to the communication scheme, when a 2-bit data pattern needs to be transmitted, the corresponding orthogonal/antipodal code is transmitted instead, requiring a $2^4 = 16$ PSK modulator. Therefore, $M = 416$.

(b) The code rate is given by:

$$\text{Code Rate: } r = 8/8 = 1$$

(c) The bit rate after coding is given by:

$$R_{b2} = R_{b1} r = 10\,\text{kb/s}\ (r = 1)$$

(d) The transmission bandwidth is given by:

$$\begin{aligned} BW &= 2Rb2/\text{bits per symbol Hz} \\ &= 2R_{b2}/4 = 2 \times 10\,\text{kb/s}/4 = 5\text{ k Hz} \end{aligned}$$

(e) Since there are four orthogonal waveforms (two orthogonal and two antipodal), the number of errors that can be corrected is quadrupled, i.e.,

$$4t = 4 \times 1 = 4$$

7.5 Conclusions

A method of coded modulation, based on N-ary complementary codes and M-ary PSK (MPSK) modulation, is presented. N-ary coded modulation is a multi-level channel coding and multi-level modulation technique, where instead of coding one bit at a time two or more bits are encoded simultaneously and then modulated by means of M-ary modulation. At the receiver, the decoder recovers the data by means of code correlation. A lookup table at the receiver contains the input/output bit sequences. Upon receiving an encoded data pattern, the receiver validates the received data pattern by means of code correlation. Construction of N-ary coded modulation schemes based on convolutional and orthogonal codes is presented to illustrate the concept. The proposed coded modulation schemes are bandwidth efficient and offer a coding gain near the Shannon's limit.

References

1. Ebert PM, Tong SY (1968) Convolutional Reed-Soloman codes, Bell Syst Tech, pp 729–742
2. Gallager RG (1968) Information theory and reliable communications. Wiley, New York
3. Kohlenbero A,.Forney GD Jr (1968) Convolutional coding for channels with memory, IEEE Trans Inf Theory IT-14:618–626
4. Reddy SM (1968) Further results on convolutional codes derived from block codes. Inf Control 13:357–362
5. Reddy SM (1968) A class of linear convolutional codes for compound channels. Technical Report, Bell Telephone Laboratories, Holmdel, New Jersey
6. Reddy SM, Robinson JP (1968) A construction for convolutional codes using block codes. Inf Control 12:55–70
7. Berrou C, Glavieux A, Thitimajshima P (2010) Near Shannon limit error—correcting (PDF). Retrieved 11 February 2010
8. Berrou C (2010) The ten-year-old turbo codes are entering into service, Bretagne, France. Retrieved11 February 2010
9. McEliece RJ, MacKay DJC, Cheng J-F (1998) Turbo decoding as an instance of Pearl's "belief propagation" algorithm. IEEE J Sel Areas Commun 16(2):140–152, doi:10.1109/49.661103, ISSN 0733-8716
10. IS-95 Mobile station—base station compatibility standard for dual mode wide band spread
11. Spectrum Cellular Systems, TR 55, PN-3115, March 15 (1993)
12. Samuel CY (1998) CDMA RF system engineering. Artech House Inc
13. Vijay KG (1999) IS-95 CDMA and cdma 2000. Prentice Hall
14. Bernard S (1988) Digital communications fundamentals and applications. Prentice Hall
15. Faruque S, Faruque S, Semke W (2013) Orthogonal on-off keying for free-space laser communications with ambient light cancellation. SPIE Optical Eng J 52(9), September 26
16. Faruque S (2016) Radio frequency channel coding made easy, Springer, ISBN: 978-3-319-21169-5
17. Faruque S (1999) Battlefield wideband transceivers based on combined N-ary orthogonal signaling and M-ary PSK modulation, SPIE proceedings, vol. 3709 digitization of the battle space 1 V, pp 123–128
18. Faruque S et al (2001) Broadband wireless access based on code division parallel access, US Patent No. 6208615, March 27
19. Faruque S et al (2001) Bi-orthogonal code division multiple access system, US Patent No. 6198719, March 6
20. Faruque S (2003) Code division multiple access cable modem, US Patent No. 6647059, November

RADIO TESLA

the secret of Tesla's radio and wireless power

by George Trinkaus

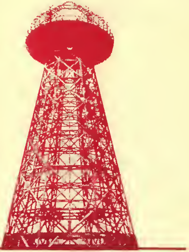

In clear English, 66 illustrations, and a minimum of math, this booklet details the peculiar radio technology of Nikola Tesla. Early elemental radio devices like Tesla's are fascinating and worthy of study because they remind us that powerful radio technologies can be so simple and accessible. Tesla's transmitters produced high-voltage, sudden-pulse disturbances conducted through the earth at low frequencies. His radio is as taboo in official science as his wireless power, which works on the very same principles. Tesla said the radio physics of Hertz, still in vogue, was an "abberation," a "fiction." Modern radio took a lot from Tesla, but what did it ignore? Also in these pages: spark reconsidered, grounding rediscovered, low frequency revived, lament on solid state, underground radio, license-free radio, the suppression of amateur radio, carrier-current radio, radio-free energy...

$7.75

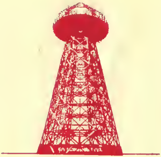

"Oscillations of low frequency are ever so much more effective in transmission." — Tesla, page 10

Radio Tesla
ISBN 0-9709618-3-9
9 780970 961839
90000

High Voltage Press, P.O. Box 1525
Portland, OR 97207, USA
toll-free 877-263-1215
teslapress@aol.com
www.teslapress.com

Thanks to Lee Dodson for all kinds of sustenance and
to Jim Davis for his patient teaching.

Copyright © 1993 by George B. Trinkaus
All rights reserved.

ISBN 0-9709618-3-9

photos by Jan L. Kahn

Contents

Introduction 1
1. High Voltage, Sudden Pulse 3
2. Low Frequency 10
3. Conduction through the Ground 15
4. Resonance 19
5. Sensitive Device 25
6. Aerial Capacity 32
Tank Tables 36

HIGH VOLTAGE PRESS

Introduction

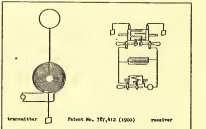

Tesla radio system

When radio was born around the turn of the 20th century, various inventors, who are not celebrated today, created their own peculiar radio technologies, which are largely ignored today. Among these inventors is Nikola Tesla, though there are others, like Nathan Stubblefield and Mahlon Loomis, who are even more obscure. The radio technology that is peculiar to Tesla, though it got a few years of public exposure in its time, gets even less acceptance in today's technology than Tesla's disk turbine, his tesla coil, or his high-frequency lighting, and Tesla's radio is as taboo in official science as his wireless power, which works on the very same principles.

In 1943, only a few months after Tesla's death, the U.S. Supreme Court, yielding finally to the pressure of a suit fought over many years, declared that Tesla's radio patents were among those that had been infringed upon by Marconi and thus, in effect, wrote into the official record Tesla's status as a founder of radio. This was a purely symbolic victory, for Tesla's radio was suppressed, and the technology that developed is distinctly different in many essential respects.

During the period of radio's most rapid growth (1915-1940), Tesla watched quietly from the sidelines, for by this time he had fallen out of favor with the media, or rather with the establishment that controls it. Still, some of his comments have made it into the published record. In 1927 he said that broadcasting "is now carried out with unfit apparatus and on a commercially defective plan." Of radio technology generally Tesla said in 1932 that "the transmitting and receiving apparatus is ill-conceived and not well adapted for selection. The transmitter generates several systems of waves, all of which, except one, are useless. As a consequence only an infinitesimal amount of energy reaches the receiver, and dependence is placed on extreme amplification."

"Radio experimenters of this age," Tesla said of the hams of 1934, "are following ancient theories." By this he meant backward theories. Tesla's favorite backward theorist was Heinrich Hertz, who saw the phenomenon of radio as some kind of straight-line radiation akin to light. Tesla said Hertz' theory, which is still in vogue today, was "one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history." This was not a reckless statement, for Tesla reports that he had carefully reviewed Hertz' experiments, conducted comparative tests with his own brand of radio and had come to a different set of conclusions. Tesla said, "I considered this so important that in 1892 I went to Bonn, Germany to confer with Dr. Hertz in regard to my observations. He seemed disappointed to such a degree that I regretted my trip and parted from him sorrowfully." Tesla made subsequent tests in 1900 with the same results and kept abreast of articles on Hertzian radio-telegraphy, which, he said, always impressed him "like works of fiction."

suppressed information

In this book I attempt to break down Tesla's radio into a set of specific principles and to survey the whole of radio technology from Tesla's perspective. While I have studied some out-of-print literature, much of my information comes from reprints still available. These titles I cite at a chapter's end together with addresses for ordering. (Addresses given on first mention only.)

For the core material on Tesla, what you have in print is one fat hardcover book, *Colorado Springs Notes*, eight U.S. patents, and a few magazine articles by Tesla himself, one of which is titled "The True Wireless." The biggest single source, *Colorado Springs Notes*, embraces only one year (1899) of Tesla's many years in radio, albeit a very intense year. Though this work is rich in information, it represents only a tiny fragment of Tesla's total legacy, which has been said to amount to some 100,000 documents, including 34,552 pages of scientific material and 5,297 pages of technical drawings and plans. Though much of Tesla's research into radio took place in New York City in years prior to his Colorado adventure, there is no published volume called "New York Notes."

Colorado Springs Notes itself has a bizarre publishing history. Tesla's papers, confiscated by the U.S. government upon his death, surfaced years later in a Tesla Museum in Belgrade, Yugoslavia, or at least a large chunk of the expropriated material did. When the Museum, thirty or so years after Tesla's death, published the Colorado fragment, it did so in Serbo-Croatian, a curious choice since Tesla, though of Serbian origin, wrote these notes in English. What the Museum published as *Colorado Springs Notes* in 1978 (under the imprint NoLit) had been translated back into English from the Serbo-Croatian. The translator, in his appended notes, tends to discredit Tesla's version of radio. The discrepancies between the original and the translation have prompted a study by Tesla scholar John Ratzlaff called *Serbo-Croatian Diary Comparisons*.

These frustrations noted, you still have to be grateful for the relatively abundant documentation of Tesla's radio relative to what you would find in print on any other deviant radio pioneer, such as Loomis or Stubblefield.

To put Tesla in perspective, I have also read extensively in the literature of conventional radio, and, by way of grounding this information, have invested perhaps an inordinate amount of time in designing, building, and experimenting with elemental transmitters and receivers, a couple of which are illustrated in this book.

Tesla's radio principles

Tesla believed that radio was conduction, not radiation. A radio transmitter should put out high voltage delivered in pulses sharp and sudden. Tesla's radio was grounded rather than aerial. He saw radio as most potent in the lower frequencies. For both transmitting and receiving, the magnifying of effects, he believed, should be achieved with resonance and capacitive discharge rather than amplifier stages. He believed radio was a disturbance of a pervasive medium called the ether. Rather than a radiating antenna, he employed an aerial capacity. Tesla's radio inventions, like his others, show a striving for simplicity rather than complexity. He eschewed the miniaturization fashionable today, conducting his experiments on a fearless scale. Fundamental to resonant tuning was Tesla's quarter-wave principle. Tesla planned, promoted, and began construction of a World System of radio founded on these principles.

On these same transmitting principles, and with about the same equipment, Tesla experimented with his wireless power. This was a radio-like system that was to be a successor to the wired grid, which was built on Tesla's own 60-cycle alternating-current patents. Tesla's wireless power system (also patented) was designed to deliver by means of earth-electrical vibration power sufficient for industrial demands.

Tesla boasts of the simplicity of his wireless-power receiving system: "Any person skilled in the mechanical and electrical arts can utilize to advantage the practical applications of the system," by which he means that any person so skilled can construct the apparatus, receive the energy, and put it to use. Thus Tesla's system short-circuits the techno-priesthood, inviting suppression.

Tesla's wireless power has never been duplicated. Tesla is on record as saying that he lit up 10,000 watts worth of Edison bulbs wirelessly 26 miles from his magnifying transmitter at Colorado Springs, but if Tesla wrote an entry on this remarkable experiment, it is not to be found in the English version of the published *Notes*. Whatever the viability of Tesla's wireless power, his conviction that such a thing was possible testifies to his belief in the special power of his peculiar radio technology.

for more information

Tesla's **Colorado Springs Notes** (No-Lit) is distributed by Twenty First Century Books (P.O. Box 2001, Breckenridge, CO 80424, www.tfcbooks.com) as are John Ratzlaff's **Complete Patents** of Tesla, **Serbo-Croation Diary Comparisons**, and Ratzlaffs' **Tesla Bibliography.**

1. High Voltage, Sudden Pulse

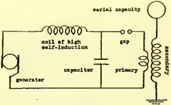

basic Tesla transmitter circuit

Tesla-style radio transmitting applies high electromotive force to the aerial-ground system – "hundred thousands or millions of volts," said Tesla. These extraordinary-sounding pressures can be achieved using the tesla coil, and fact is that Tesla's transmitters were tesla coils.

The tesla coil, with the addition of a resonant third coil and a few other alterations, becomes Tesla's magnifying transmitter, which Tesla boasted "would enable the obtainment of practically any emf, the limits being so far remote that I would not hesitate to produce sparks thousands of feet long in this manner." In Tesla radio you are happening only when your aerial glows blue.

In modern transmitters potentials in the hundreds of thousands, much less millions, of volts are unheard-of. Rarely does the voltage anywhere in vacuum-tube transmitter circuits exceed ten thousand and is usually under 1500. Antenna voltage rarely exceeds 1000. In calculating the performance of a modern transmitter, focus is on antenna current, not voltage.

sudden pulse

In addition to the high voltage, Tesla-style transmission requires sudden pulsings, in Tesla's words, "an immense rate of momentary energy delivery." This also comes easily to the tesla coil transmitter, driven as it is by the sudden capacitive discharges of Tesla's spark-gap oscillator.

Tesla's transmitters

"With electromotive impulses not greatly exceeding 15 to 20 million volts, the energy of many thousands of horsepower may be transmitted over vast distances, measured by many hundreds and even thousands of miles." So writes Tesla in a basic radio and wireless-power patent. Among the uses for this electromotive energy Tesla cites the transmission of "intelligible messages to great distances," power for "industrial purposes," the production of nitric acid and fertilizer by electrifying nitrogen in the atmosphere, and the illumination of "upper strata of air," this last item referring to Tesla's outrageous scheme to light up areas of the planet at night by beaming oscillating electric energy into the sky.

"It is a form of oscillator of great simplicity," Tesla wrote. "A coil of high self-induction is connected in series with a condenser, and across the condenser is placed a break generally in series with the primary of the coil. Very sudden discharges are produced when using a fine stream of electrolyte or mercury to effect short circuit," says Tesla, referring to one of his patented circuit controllers.

engineering

The engineering of the Tesla transmitter is the engineering of the tesla coil, and I offer, as supplement to my booklet on the subject, the following guidelines from Tesla's patents and Colorado notes:

The primary coil's inductance, and its resistance, should be as low as possible. Just one turn of heavy double cable sufficed for the 51-foot diameter primary of Tesla's Colorado transmitter, and the coil with which Tesla demonstrated wireless before the patent examiners, a coil with a flat spiral secondary, had a primary of a single turn.

The primary and secondary should have equal weights of metal. This principle helps determine the thickness of the primary conductor.

Of the secondary, Tesla wrote that its resistance, too, should be as low as possible, but the self-induction should be as high as possible. The flat spiral secondary of the coil mentioned above had 50 turns of number 10 wire. The helical secondary of the Colorado transmitter had 24 turns of number 8 on its 51-foot diameter.

A loose coupling between primary and secondary is desirable. "The mutual inductance should be small so as to permit free oscillation," said Tesla; a coil oscillates most vigorously when it is "inductively free of other circuits." The incredible excitation of a magnifying transmitter's tertiary coil Tesla attributes to its freedom from the damping effects that would result from a more closely coupled secondary. The tertiary, or "extra coil," as Tesla called it, can be located some distance from the coil stimulating it. Within the Colorado transmitter's 51-foot diameter primary-secondary coil system, Tesla would set down a tertiary, not necessarily centered. One Colorado photo shows several of these extra coils, ranging in diameters from seven inches to eight feet, placed here and there within the 51-foot enclosure, each spouting sparks.

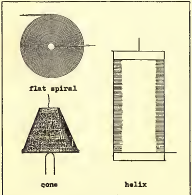
flat spiral
cone
helix

secondary coils

Loose coupling for a more free and swinging oscillation is an engineering objective behind the flat-spiral secondary design. As the coil spirals in, turns become more and more remote from the primary and its inertial influence. Also, since the secondary takes on higher voltage as it spirals inward, the outer turns nearer the primary are at relatively low potential, and this helps to thwart arcing between the coils. A cone-shaped secondary offers similar advantages and is also considered superior to the customary helical.

The magnifying effect of the interacting coils, Tesla said, is "directly proportionate to the inductance and frequency and inversely to the resistance of the secondary system." This need for low resistance invites the builder to consider conductors superior to conventional copper wire. A video from Borderland Sciences demonstrates a Tesla-style transmitter with a flat-spiral secondary made of silver-teflon coaxial cable; the shielding is the conductor. The same coil has a primary of bronze strap.

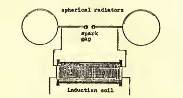
spherical radiators
spark gap
induction coil

Hertz resonator

Tesla tuned his transmitters for "maximum rise of

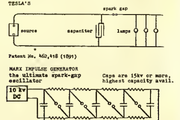
TESLA'S
spark gap
source
capacitor
lamps
Patent No. 462,418 (1891)
MARX IMPULSE GENERATOR
the ultimate spark-gap oscillator
Caps are 15kv or more, highest capacity avail.
10 kv DC
fires first
DANGER: shock hazard. Discharge caps after shut-off.
set gaps to fire consecutively

spark-gap oscillators

emf on the excited system," as indicated by a small incandescent bulb connected in a single loop around the final coil. Tuning was aided by a regulating coil and by an adjustable spark gap in the ground circuit or by a gap shunting part of the secondary. He observed that "the tuning is remarkably exact, 1/8 turn of the self-induction box reducing the effect very much." He said, "It becomes easy to locate the maximum rise within one-quarter of one percent." He described his transmitter's tuning as "very sharp."

spark

Spark radio is high-voltage, sudden-pulse. Spark is as old as radio itself. Radio was supposedly discovered by somebody wondering about spark noises in a telephone. Hertz' experiments used spark. The induction coil with spark gap is the first radio transmitter. The next advance put a capacitor in the circuit, as in Tesla's spark-gap oscillator. This Tesla invented not for radio but for an early high-frequency lighting system. For radio, Tesla used his all-purpose oscillator, the tesla coil.

As radio moved into the future, leaving Tesla behind, the spark transmitter that came into general use was not a tesla coil but a truncated version in which the magical resonant secondary was lopped off. The power supply was an induction coil which one could build or buy. One manufacturer advertised eleven models providing sparks from 1/4 inch to 8 inches. Antenna and ground were connected to a primary-like helical coil called a transmitting helix. The rarity of true tesla-coil transmitters in the literature suggests that this potent oscillator may have been a well-kept secret even in the spark era.

"spark forever"

"Using spark allows obtaining of great suddenness," said Tesla. No modern tube or semiconductor can match disruptive discharge for rapid switching of high energy without blowing apart. The history of radio reports that operators were satisfied with spark. "The pervading spirit was one of complete complacency with regard to the technical status of the art," according to one historian. A 250-watt spark set went 300-400 miles, and with spark that power was easy to come by. The circuitry is simple, the parts easy to find even today for those in the know about tesla-coil building. If you needed to throw together a telegraphic radio signalling device in a pinch, spark might be the way to go, and even voice modulation is possible, as we shall see.

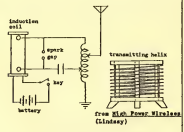

from High Power Wireless (Lindsay)

spark-gap transmitter with helix

The sudden-pulse spark-gap oscillator could be used to drive a microwave cavity resonator (page 7).

Could a self-powered spark transmitter be created from a static machine? (Also page 7)

Radio operators, both commercial and amateur, resisted mightily the movement to suppress spark and replace it with tube oscillators, which they saw as costly, complex, and relatively ineffectual. The anti-spark movement was imposed from above by industry and government. Imprinted on some of the QSL cards

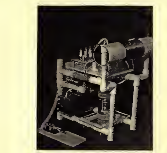

author's experimental spark transmitter hams sent through the post to confirm contact was the slogan "spark forever." Resistance continued into the 1930's, and, who knows, there may still be a few covert spark operators out there in the night.

cleaning up spark

The problem imputed to spark was broadness of signal and harmonic noise in radio bands which were becoming progressively more crowded with competing communicators. Desired was a sharp, clean signal that occupied only the thinnest slice of the busy radio spectrum. There is nothing in the FCC Code for amateurs that prohibits spark itself. The Code does set strict emission standards that the rudimentary transmitter of the spark era had trouble conforming to.

The common historical impression is that spark transmitters were uniformly and inevitably crude and that the new vacuum-tube transmitters, using feedback principles of oscillation, came along like white knights and cleaned up the spectrum. Fact is, though, that toward the end of the spark era transmitters using evolved spark technologies were developed, manufactured in numbers, and saw long service on land and sea, sharing the bands with the incoming tube (and arc) transmitters. Although any spark transmitter, no matter how sophisticated, dissipated some of its power in harmonics, these advanced spark rigs produced relatively clean and narrow fundamental signals and the smooth, damped, "continuous waves" boasted by the promoters of the new tube transmitters. Even after ships were routinely equipped with tube rigs, radio operators insisted on keeping on board as auxiliaries their old reliable units with spark.

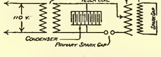

from High Power Wireless
(Lindsay)

tesla-coil transmitter (1910)

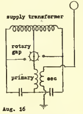

Cols. notes, Aug. 16

Tesla's push-pull oscillator

The engineering of superior spark transmitting focused on improving the quenching (shutting off) of the gap: the technology of airtight series gaps, magnetic blow-outs, and rotary gaps. Perfecting of the spark gap got a lot of attention from Tesla, and perhaps his patents were studied by radio engineers, but not necessarily.

Tesla anticipated the push-pull circuit common in

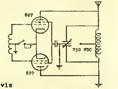

simplified schematic

with Jim Davis

author's push-pull tube transmitter

amplifiers and transmitters. Tesla ultimately replaced the spark gap altogether and achieved disruptive discharge with ingenious high-energy mechanical "breaks," including rotary mercury switches and mercury jets. Tesla said that his transmitters produced continuous waves. Evolved spark transmitters used synchronous rotary gaps which coordinated the charging rhythm of the capacitor with the frequency.

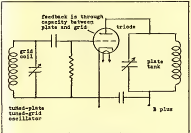
feedback is through
capacity between
plate and grid
triode
grid
coil
plate
tank
tuNed-plate
tuned-grid
oscillator
B plus

basic tube transmitter

A feature of Tesla's magnifying transmitter design that discourages undamped waves is the single-turn primary. This reduces the induction of the primary circuit so that unwanted oscillation cannot occur there. The same provision shows up in the only magnifying transmitter circuit I have seen in radio outside of Tesla's.

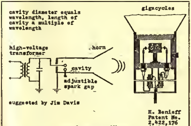
cavity diameter equals
wavelength, length of
cavity a multiple of
wavelength
gigacycles
high-voltage
transformer
horn
cavity
adjustible
spark gap
suggested by Jim Davis
H. Benioff
Patent No.
2,422,176
(1947)

spark-gap oscillator
drives cavity resonator

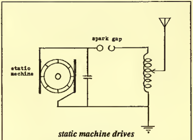
spark gap
static
machine

static machine drives
spark transmitter (putative)

Called the impact transmitter, this circuit from the spark era featured quenched gaps in tandem. The name belies an understanding of Tesla-style sudden-pulse transmitting. Knowledge of the impact transmitter, patented by Oliver Lodge, was never widely circulated to the public and remained a secret.

Can you modulate spark?

The general impression is that spark is strictly telegraphic, but, fact is, spark transmitters can be modulated.

For telegraphy, tone modulation is inherent in spark, the tone being determined by the frequency of the gap. Rotary and series gaps produce characteristic tones

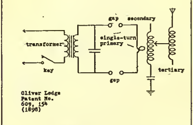
gap
secondary
single-turn
primary
transformer
key
tertiary
gap
Oliver Lodge
Patent No.
609, 154
(1898)

impact transmitter

that help telegraphic signals cut through static noise. (Tube and transistor transmitters require separate audio-oscillator circuits to impose a tone, yet another complexity.) Voice modulation requires that this spark sound be eliminated, that the frequency of spark discharges be raised to at least 5000 sparks per second. As early as 1900, Reginald Fessenden achieved bona-fide radio telephony with a spark transmitter. He probably used a very simple and now-forgotten method known as absorption.

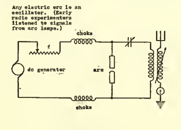
Any electric arc is an oscillator. (Early radio experimenters listened to signals from arc lamps.)
choke
f
dc generator
arc
choke

arc transmitter

A precedent for absorption modulation is to be found in the technology of alternator transmitters. Tesla, the inventor of the ac generator, or alternator, had experimented with units having many poles and capable of frequencies of up to 20 or 30 kc. These he intended to use not for radio but for his high-frequency lighting. Tesla abandoned the alternator early on in his experimenting because of its frequency limitations and developed his spark-gap oscillator and tesla coil for his lighting. Ironically, it was an alternator quite similar to Tesla's that became the first overseas radio transmitter. Fessenden developed a 50-kilowatt dynamo. Alexanderson scaled this up to 200 kilowatts. These huge alternators (they looked like power-plant dynamos) were put into service parallel with spark. Alternator radio necessarily remained low-frequency (often so low that voice was not possible) with long antennas and deep groundings. Unlike spark, the high-RPM alternator could not be turned on and off for telegraphy. So how did they key an alternator? The solution was a magnetic shut-off, which works on a principle similar to that of absorption modulation. Notice that the shut-off is in the ground circuit, suggesting that alternator radio was construed as a grounded radio. Absorption modulation for radio-telephony was installed in the circuits of either the antenna or the ground. Fessenden, in 1906, broadcast a program of music and speech with an 80 kc alternator.

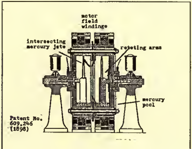
motor field windings
intersecting mercury jets
rotating arms
mercury pool
Patent No. 609,246 (1898)

Tesla's mercury break

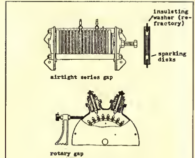
insulating washer (refractory)
sparking disks
airtight series gap
rotary gap

series (quenched) gap and rotary gap

In vacuum-tube transmitting, expensive high-power audio amplifiers are necessary for plate modulation, the respectable mode; grid- and screen-modulation also require audio amplifiers, though of lesser power. The inexpensiveness and simplicity of absorption modulation tempts experimenters who seek a simpler radio technology. I have seen the method described as "crude" in a 1924 radio book where it pops up, but no specifics were given. Is it necessarily low-fi? Does it cause unwanted FM effects, widening the band-width unacceptably? I have seen reference to the carbon microphone overheating in multi-kilowatt rigs. The cure was a water-cooled microphone. KQW in San Jose, California broadcast voice and music circa 1912 on an arc transmitter heard for about a thousand miles, and WLW in Chicago boomed on the AM band in the early '30's with a half-megawatt spark transmitter. Both used the antenna-absorption method.

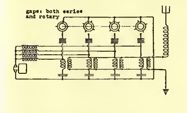

continuous-wave
spark transmitter

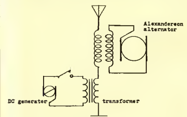

magnetic shut-off

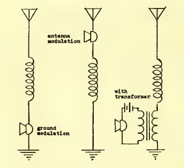

absorbtion modulation

Tesla's available notes and patents tell us very little about his particular modulation schemes (his Colorado notes show an "arc controller"), but as the promoter of a world broadcasting system he seemed supremely confident that his magnifying transmitters could be modulated.

The passing of spark marked the end of Tesla-style high-voltage, sudden-pulse transmitting.

for more information

Tesla Coil by George Trinkaus. A how-to for the nonexpert. How Tesla did it. How you can from off-the-shelf parts. (High Voltage Press). **The Tesla Coil Designer** by Walt Noon is a computer program. (3283 Belvedere, Riverside, CA 92507). **Theory of Wireless Power** (book) and **Tesla's Longitudinal Electricity** (video) are among many Tesla titles from Borderland Sciences (P.O. Box 6250, Eureka, CA 95502). **Lindsay Publications** has many pertinent titles. (P.O. Box 12, Bradley, IL 60915).

2. Low Frequency

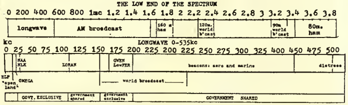
THE LOW END OF THE SPECTRUM
0 200 400 600 800 1mc 1.2 1.4 1.6 1.8 2 2.2 2.4 2.6 2.8 3 3.2 3.4 3.6 3.8
longwave
AM broadcast
160 m ham
120m. world b'cast
90m world b'cast
80m. ham
kc
LONGWAVE 0-535kc
0 25 50 75 100 125 150 175 200 225 200 225 250 275 300 325 400 450 475 500
NAA NLK
LORAN
GWEN LowFER
beacons: aero and marine
distress
ELF "open land"
OMEGA
world broadcast
GOVT. EXCLUSIVE
government shared
government exclusive
GOVERNMENT SHARED

"Oscillations of low frequency," said Tesla in 1919, "are ever so much more effective in transmission, which is inconsistent with the prevailing idea." Radio at low frequencies (below the AM broadcast band) delivers more distance per watt. It is less vulnerable to atmospherics than shortwave. It does not suffer from the fade-in-fade-out syndrome that plagues shortwave. Low frequency radio is less affected by the sun. While much of shortwave is blotted out in daytime, low-frequency propagation is strongest at midday.

In early radio, the superiority of low-frequency was taken for granted. Shortwave was considered useless. But as radio attracted more practitioners, there inevitably developed a struggle for elbow room within the limited confines of the recognized radio spectrum.

who owns the ether?

Not surprisingly, the very first entity to claim ownership of the increasingly precious spectrum real estate was Government, in particular that of the U.S. through its navy. The U.S. Navy between 1900 and 1910 made several attempts to put radio under its sole control. The Navy, which during this period was already developing global marine communications and radio-navigation systems, was so assured of its right to rule the ether that it campaigned Congress for a monopoly of all of radio, both point-to-point and broadcast on sea and land worldwide.

The Navy also fought for the phase-out of spark and the development of a more refined radio. It set the specs for industry to develop the new continuous-wave technology.

Of course, the spectrum had not been waiting silently for government to step in and regulate it. The ether was already abuzz with activity, much of it conducted by the independent citizen-experimenters who ultimately became labelled (did they think it condescending?) as "amateurs."

The Navy introduced into Congress a series of bills that had the standard formula of providing for different classes of stations that would be registered and licensed, and then making it illegal for any outsider to interfere with these stations. No mention was made of the amateurs, whose transmitting stations (largely spark) were to number about 6000 by 1916. Thus, in these early legislative attempts, the hams were rendered de-facto outlaws. The Navy's campaign was ultimately joined by the Department of Commerce and by such commercial interests as Marconi's United Wireless. The hams, to defend themselves, organized into the still-active American Radio Relay League.

The legislation that finally passed (1912) did recognize the hams but banished them from any activity in the low frequencies. This was in the spirit of yes-you-can-go-swimming-but-don't-go-in-the-water, for low-frequency radio was all the radio there was. The short waves were then considered not only inferior but unworkable, the desert real estate of the spectrum.

The Government's final solution to the problem of a peoples' experimental radio was achieved, albeit briefly, in World War I when all amateur radio was flatly outlawed for the duration. The Emergency Order, which came from the Dept. of Commerce and was signed by the Navy, banned not only transmitting but short- and long-wave listening as well. The order put radio under the full control of the Navy, which then tried to persuade Congress to make this situation permanent. Ultimately, federal control of radio was to pass out of the control of the Navy and into the Dept. of Commerce, until the Radio Act of 1934 created the FCC.

WWI also provided the opportunity for the government's destruction by dynamite of Tesla's magnifying transmitter tower, which still stood sturdily at Wardencliff, Long Island, a curiosity to passers-by. The excuse given was that the tower could be used by "spies," but the intention must have been to erase this monument to Tesla's alien radio and wireless power.

During WWII the government ordered another blackout of amateur radio.

The hams are still on the defensive. As I write the ARRL is dealing with a threat to yet another amateur territory way up in the spectrum at 222 megacycles, a band United Parcel Service and others are pressuring the FCC to reallocate to land-mobile uses.

The low frequencies, under the control of government, have been allocated to military and other bureaucratic functions, to navigation beacons, like Omega (10-14 kc) and LORAN (100 kc), and for weather stations, and time registers. Tesla had suggested low-frequency navigation systems. Some of the military transmitters are humungous, like the Navy's 3000 acre NAA command facility (24 kc) that runs 2 million watts and ELF down at 76 cycles, which uses antennas over 50 miles long.

citizens' radio

The government-military takeover of radio (nearly 100 percent of the low bands and much of the upper spectrum as well) has impacted on the amateurs, who have had to justify themselves as a quasi-government "service" in order to retain the privilege of radio. The amateurs' awkward posture is: "We are an emergency service, but, until an emergency comes up, we'll just be jawboning here to keep these bands open."

As early as 1925 the hams were building close functional relationships with the military. During the amateur blackout of WWII, many hams, as a way of staying active, enlisted in a military-run civil-defense program called the War Emergency Radio Service. Before government had a police radio, there was an amateur service that assisted the police in matters such as recovering stolen cars.

The fact that there is any nongovernment or noncommercial radio is obscured from the public consciousness. While hams sometimes get a few seconds on the tv news when they are performing their public service during an earthquake, the general public is given little awareness via the mass media that such a thing as a citizen's radio exists. On the tv the only character holding a radio mike is a cop.

A colorful wall-poster chart of the entire radio spectrum published by the Office of Spectrum Management (a bureaucracy within the U.S. Department of Commerce's National Telecommunications and Information Administration) provides elaborate color coding of 28 radio services, including broadcast, radionavigation, amateur, etc., but has no code for a citizens radio, and the 40-channel citizens band at 27 mc is ominously absent from the chart.

Regulation of the ether can never be absolute. Some governments, like Italy's, have had to abdicate some of their control over broadcasting to independents ("pirates"). Pirate radio and tv flourish where a state radio tries to dominate, and, by the same dynamic, where broadcasting is dominated by a handful of media corporations acting as one, piracy will predictably rise up. (Some people believe in keeping a gun in the house by way of providing for some impending political catastrophe. I'm not a gun person, but I do advocate keeping a transmitter handy. Does the First Amendment guarantee the right to bear transmitters?) The FCC once thought that it controlled CB, but when CB exploded with the advent of inexpensive transceivers made possible by the phase-lock loop frequency synthesizer, the government found it could not enlist the cooperation of all operators in a licensing procedure and threw up its hands. CB may be anarchy, but it still works, and the FCC can't do much about the kilowatt linears anyway.

black-box consumers

There is little encouragement for hams or anybody else to explore radio experimentally. The literature testifies to a time when way opened more easily for the curious and the experimenter. Published between 1900 and 1930 were at least twenty different monthlies for the electrical experimenter. These publications were full of projects, including transmitters and receivers, that one could build in the basement workshop from parts easily available at independent dealers. Tesla, who published in the popular press, while avoiding the academic, was a star contributor to some of these magazines. The reader was assumed to be an experimenter, a builder, and capable of sustained thought, judging from the prose. Today he is assumed to be, above all, a consumer.

The typical ham through the 1930's built his own station, transmitter and receiver. By the mid-1940's he was still likely to build his own transmitter (though manufactured units were becoming available), but by then he had been persuaded that the manufactured receiver was superior to anything he himself could build. Since the '60's the ham has bought both transmitter and receiver (even most of the kits popular in the '50's are gone), and today he is encouraged to buy, for $800 and up, both units boxed into one solid-state transceiver. Once an experimenter and builder, the ham of today is encouraged to become just another dependent consumer of black-box electronic components, which he himself often cannot even repair. In the case of broadcasting, the FCC prohibits a licensed operator of an AM, FM, or tv station building his own transmitter, and the FCC dictates design, hence cost. Today, among the few who build their own, is the pirate broadcaster, though he is also likely to convert an old ham rig. Many of the radio listeners of yesteryear built their own receiving sets for broadcast AM listening as well as shortwave, but by 1930 this became unfashionable due to the abundance of manufactured units.

DO NOT REMOVE COVER.
NO USER-SERVICABLE PARTS INSIDE.

With today's integrated circuits, even the do-it-yourself discreet-component builder and experimenter becomes a consumer of tiny black boxes whose concealed workings he may only dimly understand.

Sealing up electronics in black boxes discourages the curious, while it promotes a priesthood, and so does the conventional electronics education and its textbooks. The innocent student, from the first day, is snowed under with mathematical formulae – information that could be useful someday if he were allowed to engineer a circuit, but it cannot tell him how electricity works. (Your author in his checkered past was an editor of a series of those "basic" electronics texts; his Tesla work may be an act of penance.) Tesla, a consummate mathematician, was skeptical of mathematical explanations. He called Relativity "a mass of deceptions wrapped in a beautiful mathematical cloak." Tesla had no good words for the electronics theory that grew out of quantum and Einsteinian physics. He is on record as saying that "there is no such thing as an electron."

low-frequency revival

While conventional histories congratulate the amateurs for turning shortwave into viable radio (the hams invented the shortwave technology the Navy and others would adopt), the allocation of hams to frequencies above 1500 kc meant their exile from superior turf. The banishment of hams from long-wave along with the romancing of the shortwaves has resulted in the almost complete lapse from public memory of low-frequency radio and what it means. This is particularly true in the U.S., for there is some long-wave broadcasting abroad, which, incidentally, can be received here to some extent on both coasts.

But in recent years there has been a revival of longwave. Low Frequency Experimental Radio people (LowFERs) transmit license-free in the 1750-meter band (160-190 kc) with government sanction under Part 15 of the Federal Communications Act, the same little-known part of the law that permits flea-power community broadcasting. The law limits LowFERs to just one watt and to a 50-foot antenna-ground system. Under

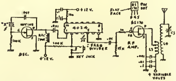

from The Low Frequency Scrapbook by Ken Cornell

low-frequency transmitter (160-190 kc) ideal conditions, a legal LowFER signal can be heard for up to 1000 miles, testimony to the power of longwave. But there are so many competing disturbances in these frequencies from broadcast harmonics, powerline harmonics, powerline-carried noises from motors and light dimmers and many other inadvertent transmitters, that reception is almost impossible in urban environments. Most LowFERs set up telegraphic beacons, listen for each other, and report reception in their networking newsletters, publications that are in the avant-garde of experimental radio. LowFERs are among the few who build their own. LowFERs remain true to the experimental spirit that marked early radio. They are willing, as one LowFER put it, to "reinvent the wheel." LowFER reception requires the experimenter to go mobile into the boondocks where it's electronically quiet, but even then signals may be overwhelmed by GWEN, a nuclear-doomsday Air Force system having about one hundred sites across the country. GWEN operates in the same band assigned to the one-watt LowFERs and its stations rehearse daily, emitting bursts of heavily encrypted radio teletype. (The same federal spectrum chart that forgets CB assigns a "government exclusive" code to all of LowFERland.)

Most LowFERs are resigned to working experimentally within the challenging constraints of the government-mandated one watt, but some are agitating for more practicable levels of power.

Closely allied to the LowFERs are the MedFERs, medium-wave experimenters who operate just above and below the AM band with a 100 milliwatt power limit, the same flea-power limit applied to community broadcasters. The upper MedFER territory has been opened to commercial broadcasters, but these experimenters will continue in the band until displaced. A few broadcasters are already operating in the band; their viability depends upon the manufacture of receivers with extended AM-band coverage.

carrier-current radio

When amateur radio was outlawed in WWII, the only activity open to hams, outside of enlisting in the Wartime Radio Service, was a low-frequency radio propagated over the power lines. Called carrier-current or wired wireless, this little known mode of radio is particularly effective at low frequencies. Line transformers tend to choke out high frequencies, but low frequencies conduct through. The WWII hams worked voice and code down at 160 to 200 kc.

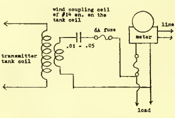

a carrier-current coupling to power lines

Power companies use such a low-frequency system today to transmit information between stations. Powerline carriers are another nuisance to low frequency listeners.

For the independent broadcaster, carrier current is an option. Within flea-power watt restraints, it's even legal under Part 15. Independent and community broadcasters use carrier current for transmitting both AM and FM. Receivers need not be coupled to the lines; even car radios can get the signal from the overhead wires. Though the AM band is five times higher than the frequency a power company or a WWII ham would use, AM probably does some penetrating of transformers. I recall that my college AM carrier-current station did a good job of getting around the entire campus as well as much of the adjoining small town.

open land

Did you know that in the basement of the very crowded, highly controlled radio spectrum, down at 0 - 9000 cycles, there is no government regulation? Thus, this unallocated very low frequency band is open to unlicensed experimenters at any power level. Conceivably, one could set up point-to-point communications using just everyday audio amplifiers. An underwater diver communication system uses 10-watt audio amplifiers. Output is connected to widely spaced submerged electrodes as antennas. The idea might be tempting to audiophiles who would be radiophiles. A problem would be the noises, both civilized and natural, that abound in this band, particularly the 60-cycle power-line hum and its harmonics. Another problem is that any one voice-modulated signal would consume the entire band within its range for the duration of the transmission.

low-frequency listening

When shortwave came into fashion, the lower frequencies became de-facto off-limits to the listening public. But lately, along with the LowFERs, the low-frequency listener is coming back. There is a growing curiosity about what goes on in the forbidden territory below the AM broadcast band.

To tune in reliably to distant LowFER beacons takes a sophisticated receiver with crystal filters and one kilocycle dial resolution, but I can testify to the enjoyment provided by seeking stronger beacons on my simple, one-transistor, uncalibrated home-built regenerative unit shown later herein. If you own a shortwave receiver, you may want to try a solid-state converter like the one shown above. Beacons are radio-telegraphic lighthouses that repeat slow alphabetic identifiers (good code practice). Called NDB's in LowFER jargon, marine non-directional beacons run only about 25 watts, while the FAA's aeronautic beacons can go to a kilowatt. There are directories that can tell you the location of beacons heard, if you want to take it that far. Also you may hear maritime ship-to-shore telegraph as well as big military transmitters, along with a lot of other strange signals you just have to wonder about. Receiving the megawatt foreign low-frequency broadcast stations is a possibility if you live on either U.S. coast. Any low-frequency listening is subject to suppression by civilized electrical noise and benefits from a nonurban environment. It is in urban listening that the expensive well-filtered receiver pays off.

natural radio

With a very low-frequency receiver (below 30 kc), you can hear the sounds of natural radio: spherics, hisses, pops, chirps, tweaks, whistlers, as well as something called the dawn chorus. Tesla listened to the sounds of VLF radio and conjectured that among the disturbances were signals from Mars. ("A clear suggestion of number and order," he noted; "impossible to think accidental...a purpose was behind these signals.") Experimenters today who tune in to nature's radio use receivers of various degrees of sophistication, but, according to Michael Mideke in his *Sounds of Natural Radio*, "given a large enough antenna, even amplification can be dispensed with; signals can be heard using nothing more than headphones connected between antenna and ground."

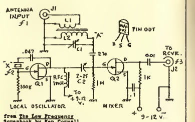

from The Low Frequency Scrapbook by Ken Cornell

low frequency converter

for more information

200 Meters and Down, The Story of Amateur radio by Clinton DeSoto, 1936, (ARRL, 225 Main St. Newington, CT 06111). **U.S. Frequency Allocations, The Spectrum** is a full-color wall poster chart (U.S. Govt. Printing Office, Washington D.C. 20422). **High Power Wireless** (1910) and **The 1934 Official Shortwave Manual**, Hugo Gernsbach, ed. (all three from Lindsay). **The Lowdown** is the LowFER and MedFER newsletter, Bill Oliver, ed. (Longwave Club of America, 45 Wildflower Rd., Levitown, PA 19057). **Carrier Current Techniques** by Ernest Wilson (Pan Com, P.O. Box 130, Paradise, CA 95969). **Low and Medium Frequency Scrapbook** by Ken Cornell, the bible of LF tech (225 Baltimore Ave., Point Pleasant Beach, NJ 08742). **Tesla: Contact with Mars** (Rex Research, P.O. Box 19250, Jean NV 89019).

3. Conduction Through the Ground

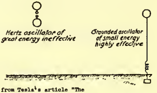
Hertz oscillator of great energy ineffective
Grounded oscillator of small energy highly effective

from Tesla's article "The True Wireless" (1919)

Hertzian vs Tesla radio

The propagation of radio signals and wireless power Tesla saw as a matter of conduction. Tesla's radio-conductive path was not a circuit in the conventional sense but was modelled on the single-wire-without-return principle that he demonstrated in his high-frequency, high-voltage lighting system. Electric researcher Eric Dollard calls this one-way conductive effect "longitudinal electricity."

Tesla's radio-conduction can happen either through the earth or through the sky. Tesla said, "The earth behaves simply as an ordinary conductor." He said the earth was superior conductor; but, pertinent to sky transmission, Tesla is on record in a patent as saying that "transmission through elevated strata encounters possibly less resistance than copper wire."

Conventional theory never hints at radio propagation being anything like conduction but represents it as some kind of radiation. This is more than a textbook convention; a whole science of antenna design is based on this assumption, which Tesla dismissed.

an etheric medium

What is radio, anyway? The texts, like my *Radio Amateur's Handbook,* speak of radio as "radiation." But, at the same time, it is a "wave." Then again, in the same paragraph of the *Handbook*, which is just echoing respectable physics, we are told that it is "traveling electrostatic and electromagnetic fields." Whatever it is, some sort of almost material emanation is presumed to be issuing from the antenna into a void.

Tesla, however, saw radio as a "disturbance of the medium," or a "commotion in the medium." He writes of "transmitting an electrical movement to the environing medium." What is this medium?

In Tesla's time it was an article of faith that there existed a unified field that permeated all being, all matter, including solids, liquids, and even what we call space or vacuum. This pervasive "ether," as it was called, is the medium of radio. The transmitter sets up a disturbance that produces, says Tesla, "alternate compressions and rarefactions in the medium." This suggests some kind of elastic continuum that conducts standing waves.

Coming from this view, all electric phenomena — charge, polarity, oscillation — involve some kind of strain or vibratory disturbance of the equilibrium of the ether.

The etheric perception of radio came naturally to early experimenters. Loomis writes of his aerial telegraphy as "causing electrical vibrations or waves to pass around the world, as upon the surface of some quiet lake one wave circle follows another from the point of disturbance to the farthest shores."

In Tesla's radio, effective transmitting is achieved by setting this elastic medium vibrating with a sudden high-voltage whack, like that created by the discharge of a capacitor.

grounded radio

Hertzian radio has us all conditioned to thinking in terms of aerial radio: "the air waves," "on the air." Extreme importance is placed on the antenna and its configurations, which can become byzantine. But Tesla's radio is not aerial. It is ground-conduction. The lower end of the energized coil is rooted in the earth. Pure Herzian radio has no such natural load. Modern radio regards earth ground as an electronic toxic waste dump to which noise is conducted. Tesla said radio "should be designed with due regard to the physical properties of the planet and the electrical conditions obtaining in same."

The electrical vibrations of the transmitter are "communicated to the ground," says Tesla. Here they set up standing waves — "outgoing crests and hollows in parallel circles." He said, "The terrestrial conductor is thrown into resonance with the oscillations impressed on it like a wire." Tesla was convinced that the waves were not "electromagnetic" since such waves were not likely to travel through the earth. Tesla's ground wave, unlike Hertz' aerial, does not travel uniformly at the speed of light and can even reach a velocity that is infinite.

Oscillating the entire electrostatic Earth was among Tesla's many ambitions, and he observed that "the planet behaves like a perfectly smooth and polished conductor of inappreciable resistance." He went on to characterize the globe as having "capacity and self-induction uniformly distributed along the axis of symmetry of wave propagation." In the same text (a patent applied for in 1902), Tesla added that these standing waves propagate "without attenuation." That is, he is saying they do not diminish in intensity over distance.

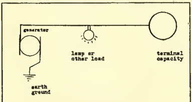

single wire without return

no ionosphere

Tesla believed Hertzian aerial radio was a waste of energy. Tesla suggests that there is some radiation component in a radio transmission but that it is a weak and incidental output. He thought the theory that high-frequency signals travelled long distances by bouncing off some high-altitude, radio-reflective layer called the ionosphere was "an utter impossibility." Tesla was not alone in disputing the existence of the Heavyside Layer, as the ionosphere was called then; it was quite controversial for a time before it was frozen into official science. Tesla said that Hertz wave theory "by its fascinating hold upon the imagination, has stifled creative effort in the wireless art and retarded it for 25 years."

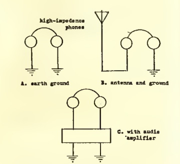

audio-frequency grounded radio

audio radio

In World War I, when field telephones were used that stretched long wires over the surface of the ground, it was discovered that one could eavesdrop on opposing forces' conversations just by connecting headphones to ground rods. This discovered, effective monitoring systems were developed by both sides that consisted of two widely separated ground rods connected to a sensitive audio amplifier. Here was a Tesla-like radio: low-frequency conduction through the earth. There is no reason you cannot use this principle for an audio-frequency radio in the unregulated open land below 10 kc. When you consider that radio can happen in the audio band, you have to wonder about the term "radio frequency" (rf), as in "rf energy," etc., which seems to imply that radio is a function of frequency.

Disturbances caused by the various electrical happenings in nature, such as lightning-created static crashes, an occasional whistler, and other strange sounds can also be heard using the above system, and it so happens that the WWI eavesdropping apparatus was sometimes totally jammed by such natural activity.

underground radio

A follower of Tesla, James Harris Rogers, patented (1920) a radio system in which both sending and receiving antennas were sunk completely underground or underwater. Rogers had his own doubts about Hertzian radio. He wondered, "If 50 units of power are passed into the aerial, then what becomes of the equal amount of energy which passes into the ground?"

Roger's was working at low frequencies, but nothing in his patents limits his grounded radio to the low band. The Rogers underground achieved superior signal strength. According to accounts, reception was free from atmospheric static, and there was no diminishing of the signal in daytime.

The Navy used the Rogers system secretly in WWI and announced it to the public in 1919. An editorial by the famous editor Hugo Gernsbach in the March 1919 issue of the *Electrical Experimenter* predicted that "the greatest pride of the radio amateur, the aerial on top of his house...is doomed... As for commercial stations their towers are doomed shortly for the scrap heap." Tesla himself had said, "The great amount of energy which can be conveyed to a [receiving] circuit by conduction through the ground, makes it appear possible that the necessity of elevating terminals...may be dispensed with." Gernsbach, who published much on Tesla in his time, said, "All our pet theories on wireless are thrown into chaos," and he predicted "a war to the knife between our wave-propagation theorists and the new school of ground-impulse savants."

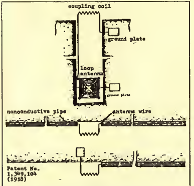

Rogers underground radio

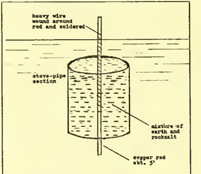
heavy wire wound around rod and soldered
stove-pipe section
mixture of earth and rocksalt
copper rod abt. 5'

stove-pipe ground

This noted, the Rogers underground vanished from the media, and Rogers' career took a downturn reminiscent of Tesla's.

Stubblefield

In 1902 Nathan Stubblefield demonstrated a grounded wireless-telephone system of his own invention. Sticking iron rods into the ground at the Virginia shore of the Potomac River, Stubblefield communicated with a steamship half a mile away. Stubblefield's grounded radio was powered by the high-frequency output of his Stubblefield battery (patent No. 600,457), a type of space-energy receiver, which shows how the occult science of free energy nudges up against that of radio, another instance being the free-energy work of T. Henry Moray. Loomis, too, used a power supply that obtained electricity from the atmosphere.

grounding technology

"The ground should be made with great care with the object of reducing its resistance," said Tesla. To ground his Colorado magnifying transmitter, Tesla buried a 20-x-20-inch copper screen 12 feet down in the arid soil. Over the top of the screen he spread a layer of coke. He flowed water over the spot continuously. Beneath Tesla's Wardencliff tower, a shaft descended 120 feet into the earth. Out from the foot of the 12-x-12-foot shaft, side tunnels extended radially. They were carbon-blackened and hence conductive. Situated near the Long Island shore, the giant transmitter was thus grounded to the oceans.

Such solid grounding has become another lost radio art. Until the 1950's even commercial broadcast receivers had a ground terminal on the chassis that one was encouraged to connect to a cold-water pipe or other ground; some houses had a radio ground wired into a wall socket.

A 1930's shortwave magazine recommended a stovepipe ground: a buried stove-pipe section filled with a mixture of soil and rock salt, which attracts moisture. Charcoal can be mixed in, as Tesla used coke. Down the center runs five feet of solid copper rod or galvanized pipe around which is wrapped the heavy lead-in wire.

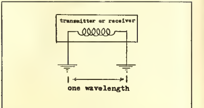

nodal grounding

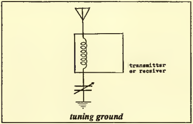

tuning ground

Any conductor used to connect radio and ground should be as short as possible and with the heaviest conductor, such that an excessively long run of, say, twenty feet is best done with a wire gauge of zero or bigger or maybe with copper tubing or galvanized pipe.

Clamping securely onto cold-water plumbing near to its ground-entry point is good grounding but this can be supplemented by other grounds, like earth rods or existing fence posts. Any body of water makes a good ground. Connect to a large submerged metal object. Any radio, transmitter or receiver, becomes hyperactive near or connected to a body of water.

Older schematics show a variable capacitor between output coil and ground, comparable to an antenna tuner. I've adopted this practice of tuning ground with good results, including noise control in receiving.

Tesla's radio assumes standing earth waves, and Tesla recommends grounding the two ends of the receiving coil one-half wavelength apart at a wave's nodal points, the location of which one determines experimentally. This arrangement is particularly important for reception of wireless power.

Having established a solid ground, some interesting experiments can be conducted in receiving. Disconnect the antenna. Does the signal vanish? What happens when you connect ground to a receiver's sacred antenna terminal? Will the Hertz police arrest you if you try any of this?

for more information

Underwater Communication, includes some of Rogers patents (Rex Research). **Subsurface Antennas and the Amateur** by Richard Silberstein, an article in Vol. 1 of *The Antenna Compendium* (ARRL). **The Complex Secret of Dr. T. Henry Moray** by Jorge Resines, radio and TV circuits from 1928 applied to Moray's free-energy discoveries (Borderland).

4. Resonance

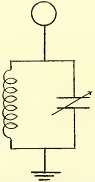

The idea of properly harmonized coils and capacitors vibrating with sympathetic reinforcement and thereby magnifying effects by tremendous ratios is at the heart of Tesla's technology.

Tesla's basic radio tuning "tank" circuit (coil plus capacitor between aerial and ground) is, all by itself, a powerful resonant signal amplifier and a beautifully simple one. Tesla suggests the power of a coil to vorticize energy. But as radio developed over the years, the tank coils in both transmitters and receivers shrank in size, and the result was a loss in gain that was compensated for by the addition of stage after stage of complex amplification circuitry using vacuum tubes.

Tesla watched this development with bewilderment. For transmitting, the resonantly tuned circuits of the tesla coil provided the high electromotive force Tesla thought necessary for long-distance radio-conduction. For receiving, the big loading and tuning coils "magnified effects," as Tesla liked to put it. So why depend on complex, multi-stage circuitry to amplify? It didn't make sense to Tesla. He said, "My plans involved the use of a highly effective and efficient transmitter conveying, at whatever distance, a relatively large amount of energy. The receiver itself is a device of elementary simplicity... In such a system resonant amplification is the only one necessary."

Taking the tank circuit another step out, Tesla conceived of a receiving coil made of glass tubing that was filled with a rarified gas, thus offering almost no impedance to the signal. This superconductivity would, he reasoned, result in tremendous gain.

I've heard tell of a radio experimenter in Portland, Oregon who built the basic crystal set on an obscene scale with a broadcast AM-band tank coil four feet in diameter and and six feet long, wound of copper tubing. Tuned with a big old commercial variable capacitor, the single-stage set, void of amplifier stages, drove an eight-inch loudspeaker with room-filling volume.

earth resonance

Tesla calculated the frequencies and pulsings of his transmitters with an eye to resonating the earth. Earth resonance is fundamental to Tesla's grounded radio and wireless power. Tesla suggested that the electrostatic earth resonated at a particular frequency, and seems to be suggesting that the closer the vibration is to that frequency the greater the magnification of effects.

What is this magical earth-resonant frequency? In his Colorado notes, Tesla says the wavelength is 5200 feet or 1737.7 meters, which works out to about 173 kc. Tesla's Colorado transmitter ranged from 60 kc to 190 kc; 170 kc was a typical operating frequency, according to the *Notes*.

In a 1920 patent, however, Tesla specifies that the frequency should be "smaller than 20 kc, though shorter waves may be practicable. "The lowest he would allow would be 6 cycles per second, to which he adds that "paradoxical as it may seem, the effect will be greatest in a region diametrically opposite the transmitter." In another fundamental radio patent, Tesla says he has learned to set up standing waves in the earth. "I found," he says, "their length to vary approximately from 25 to 70 kilometers." This works out to 428 cycles to 12

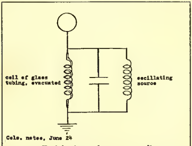

Tesla's glass-tube vacuum coil

kilocycles. If earth resonance happened for Tesla below 60 kc, then he did not achieve it on his Colorado transmitter unless it was by harmonics.

In the 1920 patent Tesla said, "The most essential requirement is...that, irrespective of frequency, the wave or wave-train should continue for a certain interval of time, which I have estimated to be not less than 1/2 or probably .08484 of a second — the time taken in passing to and returning from the opposite pole at a mean velocity of about 471,240 kilometers per second."

Recent thinking on the earth-resonance mystery hypothecates an earth-ionosphere waveguide, a Hertzian notion that would set Tesla spinning in his grave. Tesla spoke only of resonating the earth itself. The Schuman Cavity, as this waveguide is called after its theorist, is said to resonate at exactly 7.83 cycles.

the resonant tank

Of his Colorado transmitter, Tesla wrote, "The vibrating system is formed by a continuously variable and exactly determinable inductance and a capacity standard, or by an inductance standard and a continuously adjustable condenser, or by a system in which both these elements are continuously adjustable."

Modern radio technology has settled on the second of these options for tuning — fixed coil and variable capacitor. This is the arrangement since the 1930's for both transmitters and receivers. But early radio shows a mix of options to accomplish tuning. Sometimes there is no capacitor, and tuning is accomplished with a variable inductance alone. Variable coils include the tapped coil, the slider, and the variometer in which one coil rotates within another. The tradition of the tapped and slider coils is preserved among builders of the crystal set. Almost forgotten is the capacity slider, a section of tubing which encircles the coil except for a narrow slit. There is a revival among LowFERs of the elegant variometer.

Variable capacitors include the familiar rotary plate (also called "air variable" because the dielectric is air), the simply constructed "book" type, (p. 22) and the tubular high-voltage type for transmitting. Old circuits show variable capacitors consisting simply of a cluster of fixed capacitors and a rotary switch. Tesla's variable capacitor consisted of two opposing conductive disks whose distance apart was adjustable (p. 22). Transmitters sometimes use little neutralizing variables of this design, but I've seen it nowhere else. The book type was manufactured by Crosley in the 1920's. The tubular shows up in the transmitter literature circa 1910.

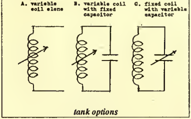
A. variable coil alone
B. variable coil with fixed capacitor
C. fixed coil with variable capacitor

tank options

Even the popular rotary-plate type of variable ca-

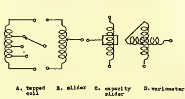

types of variable coil

where L is the coil's inductance, N is the number of turns of wire, and A is the length of the coil winding, B the diameter, both in inches.

Divide the oscillation constant for the frequency you want by the capacitance in microfarads and you'll get the inductance in microhenries that you'll design the coil for. By transposing the coil formula, you can yield the number of turns. By plugging in lengths and diameters, the physical characteristics of a coil with the desired inductance will begin to emerge.

$$N = \sqrt{\frac{L(9A+10B)}{A}}$$

quarter-wave principle

Tesla advised a simple method of calculating the length of a tuning coil. Just divide the wavelength by four to get the length of the coil wound up. This length includes the length of the entire aerial-ground circuit. Odd multiples of the quarter wave, including, I assume, odd fractions, will also resonate at the frequency.

Tesla stated the rule: "In order to attain the best results it is essential that the length of each wire or circuit, from ground connection to the top, should be equal to one-quarter of the wavelength of the electrical vibration in the wire, or else equal to that length multiplied by an odd number."

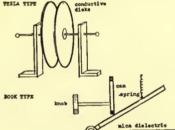

book type variable and Tesla type

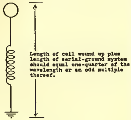
Length of coil wound up plus length of aerial-ground system should equal one-quarter of the wavelength or an odd multiple thereof.

Tesla's quarter wave

Tesla called his quarter-wave principle "the secret of tuning." He said that "without the observation of this rule it is impossible to prevent the interference and insure the privacy of messages."

Tesla's quarter-wave formula is still honored in calculating the length of antenna elements and is still useful in calculating an antenna-loading coil and for tuning circuits that have a coil alone without a capacitor.

building capacitors

If you are building a variable (or any other) capacitor from scratch this formula is useful. It also helps you calculate the value of unmarked variables that you find in surplus electronics stores:

$$C = .224KA/d\,(n-1)$$

C is the capacity in picofarads (pf); A is the area of one plate in square inches; d is the distance between the plates in inches; n is the number of plates. K is the dielectric constant, an index of the insulative (dielectric) power of the material, air being the standard with a K of 1. You can more than double the K of a variable if you are clever enough to find a way to submerge the device in mineral oil, which has a K of 2.2 and you can more than quadruple the K in castor oil (4.7). Oil also ups the voltage-handing ability by a factor of ten or more.

Tesla built a salt-water capacitor for his Colorado magnifying transmitter. It consisted of salt-water in large mineral-water bottles immersed in a tank of salt-water. The "plates" are the salt-water, inside the bottles and outside. The dielectric is the bottle-glass. Tesla called the bottles "jars" (probably in the tradition of the Leyden jar) and measured capacity by how many jars he connected into the circuit. You can build an adjustable salt-water capacitor out of beer bottles (See *Tesla Coil.*)

World System

When the magnifying transmitter tower at Wardencliff was still under construction (1902), Tesla published a brochure advertising the project as a prototype for a global radio communications corporation. Called the World System in the brochure, it would serve as a multi-frequency wireless traffic center for all existing telephone, telegraph, and stock-ticker services around the planet. Built on Tesla's earth-resonant radio principles, the System would also carry a universal time register, navigation beacons, and even facsimile transmissions. It would also be a global system of broadcasting. Tesla was among the first to suggest broadcasting of news and entertainment to the public; only point-to-point signalling had been experimented with up to then.

Tesla's tower was never completed beyond the rugged framing, and the World System fantasy collapsed when Tesla's backing, from J. P. Morgan, was pulled suddenly from under him. This crushing event signalled the end of Tesla's official career.

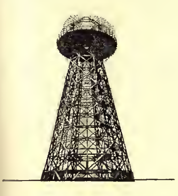

Wardencliff tower

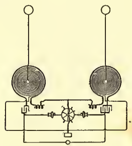

Patent No. 723,188 (1901)

Tesla's multiplex

The system Tesla describes represents a huge jump forward from any radio technology of Tesla's documented in his Colorado notes, his patents, or anywhere else. The brochure suggest the achievement of precisely tunable, high-Q, limitlessly powered transmitting of multiple channels from a single point, transmitting that could be voice-modulated as well.

Little technical detail is given, but the promotional literature suggests that Tesla may have been planning to rely heavily on multiplex techniques to cut through noise. "My individualized system with transmitters emitting a wave complex and receivers comprising separate tuned elements cooperatively associated." He called the technique a "combination lock" and boasted that any degree of safety against statics or other kinds of disturbance can be obtained." The receiver is so designed that it responds "only through the joint action of the tuned elements."

Was Tesla a fascist?

You have to wonder since his World System would have taken radio right off the bat into global centralization, a fantasy of control beyond the dreams even of the U.S. Navy. As it turned out, a quarter century after Tesla proposed his System, broadcast radio, particularly within the U.S. was still quite diversified with hundreds of AM stations, most only running 100 to 500 watts. National networks were still undeveloped. Stations were owned by entrepreneurs, local newspapers, colleges, churches, retailers. Global radio became the BBC and VOA. Even today, as multinational monopoly media is putting into place a decadent Hertzian equivalent of Tesla's World System using satellites, including direct-satellite tv, the broadcasting of signals across national boundaries is hotly resisted at diplomatic and other levels, though this battle gets no coverage in the mainstream media.

Tesla saw his World System as a civilizing force: "It will be very efficient in enlightening the masses, particularly in still uncivilized countries and less accessible regions." What did Tesla mean by civilization? He said, "No community can exist and prosper without rigid discipline." He said, "Law and order absolutely require the maintenance of organized force." Tesla said government "should prevent the breeding of the unfit by sterilization and the deliberate guidance of the mating instinct."

Tesla is admired for his technological purism, his insistence that a machine possibility be carried to its logical conclusion. Society, too, was a machine, and it needed perfecting. The World System fitted into Tesla's ideals of social order. His logical conclusion for organizing radio was a system that was centralized, omnipotent, and global.

for more information

Henley's 222 Radio Circuit Designs, 1922 vacuum-tube technology, fascinating (Lindsay) **Tesla's Work with AC and Wireless**, edited by Tesla Researcher Leland Anderson (Twenty First Century Books)

From High Voltage Press

TESLA, The True Wireless edited by George Trinkaus **$7.75**

In Nikola Tesla's own words and in 21 illustrations, this is the inventor's final published statement on how radio, at the radical, really works. What is the true wireless? Tesla says the orthodox Hertzian radio we've been taught is a "fiction." He insists that the amount of energy that can be transmitted is "billions of times greater" than conventional radio would allow. Can we transmit electric power to our homes and workplaces without wires? Can the unsightly and fragile grid come down? Tesla claims his radio can deliver wireless power. This article of 1919 has been completely reset and redesigned for clarity. Includes an introduction by Trinkaus, explanatory notes, and information sources.

High Voltage Press, P.O. Box 1525
Portland, OR 97207, USA
toll-free 877-263-1215
teslapress@aol.com

5. Sensitive Device

Tesla gave a lot of attention to the development of the ideal apparatus for detecting disturbances in the medium. Tesla let his imagination run free in his quest for the optimal "sensitive device," as he called it.

the coherer

Tesla's detector research was paralleled by many others at the time. The popular detector among radio experimenters was the coherer. This is simply a short glass tube partly filled with small metal chips or filings. Strained to near-conduction by battery voltage, this early semiconductor mysteriously switches on when an oscillating disturbance is present. A tap is needed to reset the coherer back to nonconduction. Breaking the battery circuit also works. Tesla improved the coherer by setting it into constant rotation at about 16 RPM so it would automatically reset. Changing the rate of rotation controls sensitivity. In a patent he also mentions a vibrating coherer. Constant motion suspends the coherer's chips in space, making them more susceptible to disturbances.

The coherer's internal chips or filings are ideally of uniform size. They are cleaned thoroughly in alcohol. Chips are ideally a mixture of nickel and silver, but other conductive materials can be used. Ideally, the chamber is evacuated, but not necessarily.

Until 1902 the coherer was the only detector in wide use. It dropped out of use about 1912. Tesla must have regarded the coherer as a passably reliable sensitive device because he used it in his robot (radio-controlled) boat (patented 1898) and in his Colorado lightning-tracking experiments (1900).

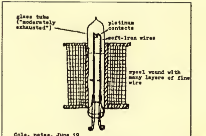

Colo. notes, June 19

Tesla's magnetic detector

Tesla's detectors

Tesla explored the possibilities of many other sensitive devices. His patents show a rotating rectifier, a precursor to such static rectifying diodes as the crystal detector. Tesla replaced his rotating rectifier with a vacuum-tube diode. This is mentioned in the Colorado notes but there is no patent. I've seen mention of Tesla exhibiting, at the 1893 Chicago World's Fair, a vacuum-tube receiver for voice and music. This must have used for its detector a vacuum diode.

The idea of straining a device to near-conduction, as in the coherer, he applied to a diode vacuum tube and to semiconductors having thin-film dielectrics.

Another Tesla detector, limited to telegraphy, was visual. It used the deposit of a thin film on a glass surface. An iodine solution, upon being stimulated by a radio disturbance, releases a conductive haloid film onto the glass screen. Battery current, conducted through the film, destroys it, thus erasing the screen. A sort of telegraphic tv, this is a precursor of the liquid-crystal display.

Tesla's Colorado notes show a magnetic detector. It consists of a coil of many hundreds of turns of fine wire around glass-enclosed wires of soft iron which are magnetically stimulated by a disturbance. Marconi used a "magnetic detector" that worked on another principle. A pair of horseshoe magnets slowly revolved over an electro-magnet the windings of which were connected

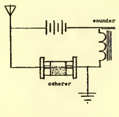

coherer receiver

to earphones. This detector was the immediate successor to the coherer.

Tesla noticed that the megavolt streamers from his Colorado coils were extremely responsive to the slightest changes in etheric conditions and wondered how this phenomenon might be applied to the sensitive device. Later, working with high-voltage, high-frequency currents in vacuum, he discovered the "rotating brush." This is an eerie emanation, a brush discharge, from a spherical conductor exactly centered in a glass bulb. The device resembles the familiar plasma globes, but these contain gases. The result is different in a vacuum. The brush resolves into a rotating ray so sensitive that if you approach it from a few paces it will turn away from you. Tesla found that a small one-inch magnet "will affect it visibly at a distance of two meters, slowing down or accelerating the rotation according to how it's held relative to the brush." There is no information on how Tesla planned to harness the rotating brush as a sensitive device, but he wrote that it was "undoubtedly the most delicate wireless detector known."

Tesla's receivers

Over 50 different receiver circuits are to be found in Tesla's Colorado notes. These are various configurations of sensitive device, tuning coil, capacitor, rotary interrupter, and battery source. If early radio engineers had studied these notes (But how could they have since they were not published until 1978?), they would have found circuits that lay the foundation for two receivers which would dominate the history of radio: the regenerative, and the superhetrodyne.

Most of Tesla's receivers have one or more continuously moving parts, an offensive idea in this age of solid state. Tesla's coherer rotated, and so did his rectifying diode. Central to Tesla's receiver designs is another continously moving part, a rotary "break" or switch used to discharge the capacitor at the proper intervals. The break discharges the capacitor through a sounder. It is a device that magnifies effects.

All the moving parts notwithstanding, Tesla allows for semiconductors to do the job of the mechanical break. "The devices," he says in an 1898 patent, "may consist of merely two stationary electrodes separated by a feeble dielectric layer of minute thickness." Tesla's experience with spark gaps must have attuned him to the possibilities. He writes of thin insulative films serving as dielectrics. The patents offer no drawings or verbal detail. How far Tesla took these thin-film semiconductor ideas experimentally I do not know.

radio-free energy

Central to Tesla's receivers was the use of capacitors to store and release energy and to magnify effects: "However feeble or attenuated the impulses received, enough energy may be accumulated from them by storing up the energy of succeeding impulses for a sufficient interval of time to render the sudden liberation of it highly effective in operating a receiver." By "receiver" here

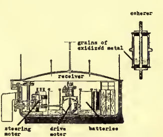

Patent No. 613,809 (1898)

Tesla's robot boat

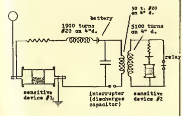
50 t. #20
on 4" d.
battery
1900 turns
#20 on 4"d.
5100 turns
on 4" d.
relay
sensitive
device #1
interrupter
(discharges
capacitor)
sensitive
device #2
Colo. notes, July 28

Tesla's coherer receiver

Tesla means sounder. It probably took a fair amount of energy to drive the sounder, especially if what Tesla had in mind was the old magnetic click sounder used in wired telegraphy.

Tesla's free-energy patent (685,957) was filed at about the same time (1901) as a string of his radio receiver patents and is, in fact a kind of radio receiver. As in his radio receiving circuits Tesla is using precisely timed capacitive discharges to magnify effects. Timing is tuning and is critical whether receiving signals or collecting energy at levels sufficient to do work.

The capacitors suitable for energy applications, Tesla says, should be "of considerable electrostatic capacity," and the dielectrics made of "the best quality mica."

the crystal set

The simplest detector diode is the crystal, a sensitive device with a long and illustrious history in radio. It is not clear how close Tesla came to this kind of solid-state rectifying diode; he may have gone from the rotating rectifier to the vacuum diode unaware of the crystal principle, but I doubt it.

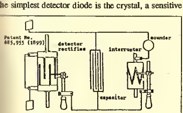
Patent No.
685,955 (1899)
detector
rectifies
sounder
interrupter
capacitor

Tesla's rotating rectifier

The crystal set is the simplest receiver in radio, and the tradition of crystal-set building persists even into this "high tech" era wedded to impressive complexity. You can build a crystal set using an inexpensive diode from Radio Shack. The traditional rock crystal with cat's whisker is still available from Antique Radio Supply, as is the high-impedence headset that enables you to listen to crystal and other elemental receivers without amplification. The cat's-whisker crystal was mass-produced in the Crystal Age enclosed in a glass envelope and with the cat's whisker welded to the active spot. Although germanium became the dominant rectifying material for diodes, similar detectors show up in the literature where the same rectification is accomplished with junction to less exotic materials: an acid solution, an electrolytic solution, a piece of strap iron. Soon after WWII there appeared a "foxhole" radio which used for its detector a safety pin and a razorblade of the no-longer-available blued-steel type. I built one, and it worked.

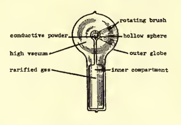
rotating brush
conductive powder
hollow sphere
high vacuum
outer globe
rarified gas
inner compartment

Tesla's rotating brush

The crystal set is the simplest receiver you can build. While a single tuning (tank) circuit might work in the country, in the city one strong station may overwhelm all others unless multiple tuning circuits are used. A crystal set with two or more tuned circuits is the simplest practical radio receiver.

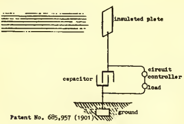
insulated plate
circuit controller
capacitor
load
ground
Patent No. 685,957 (1901)

Tesla's free-energy device

the regenerative

Tesla laid the foundation for the regenerative receiver. The vacuum-tube version, credited to Edwin Armstrong, was the receiver that succeeded the crystal set in the development of radio. "Regenerative" refers to the recirculating of the signal vibrating in the coil back into the sensitive device so as to further excite it. This feedback is another of Tesla's strategies for magnifying effects, and he knew about it early on from the self-excitation circuits he used in some of his dynamos.

The regenerative feedback link in Tesla's regenerative is a coil adjacent to the main tuning coil. In the vacuum-tube regens this became known as the "tickler coil." In some regen circuits the feedback link was through a capacitor, which could be a variable. Tesla did not patent the regen but shows many circuit variations of the idea in his Colorado notes, where he observes that it had "many valuable uses since by its means effects too feeble to be recorded in other ways may be rendered sufficiently strong to cause the operation of any suitable device."

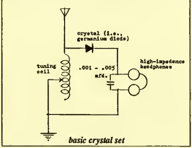
crystal (i.e., germanium diode)
tuning coil
.001 - .005 mfd.
high-impedence headphones

basic crystal set

The regen is one of the forgotten magical circuits of radio, but only recently forgotten. Familiar to every student of radio at least up through the '60's, it is not to be found in recent editions of *The Radio Amateurs Handbook*.

Though a regen often has a stage or two of amplification, an unamplified, single-stage regen having only a dozen parts is an effective listening tool for a wide range of frequencies from low through short wave. (As in other elemental receivers, high-impedence head phones are used. These differ from other phones in that there are many more turns of finer wire on the electromagnets, a technique to magnify effects, as in Tesla's magnetic detector.) The regen is sensitive and discriminating. As is, without any additional circuitry, a regen can receive code and both AM and single-sideband phone.

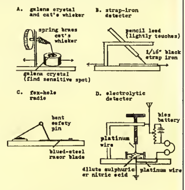
A. galena crystal and cat's whisker
spring brass cat's whisker
galena crystal (find sensitive spot)
B. strap-iron detector
pencil lead (lightly touches)
1/16" black strap iron
C. fox-hole radio
bent safety pin
blued-steel razor blade
D. electrolytic detector
bias battery
platinum wire
dilute sulphuric or nitric acid
platinum wire

some diode detectors

In my youth I built a one-tube, 4-band, shortwave delicate junctions and tenuous conductors within integrated circuits and other solid-state devices are particularly prone to vaporization. (Vacuum-tube devices are said to be ten-thousand times more resistant.) It's been speculated that a nuke exploded 200 miles above Nebraska would dud all unprotected solid-state circuitry in the continental United States. Particularly vulnerable are components connected to the power grid, to telephone lines, and to antennas. Magnetic memory could also be erased. While large institutions have been cued to this contingency and are moving ahead with the hardening of computer and communications facilities, the general public is largely oblivious to the fact that the entire "high tech" electronic culture is EMP-cancellable at a stroke.

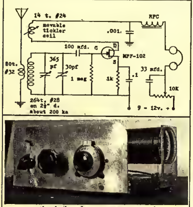
14 t. #24
movable tickler coil
RFC
.001
100 mfd.
G
D
S
MPF-102
80t. #32
365 pf
30pf
1 meg
1k
.1
33 mfd.
10K
264t. #28 on 2½" d. about 200 kc
9 - 12v. +

authar's low-frequency regenerative

This same vulnerability of solid-state devices to shock makes them a headache for the builder experimenter. The ability to plug these little items into breadboards is a big step forward in convenience, but, when a circuit fails, the experimenter is left wondering whether his hook-up is flawed or has the IC or transistor blown due to excess heat in soldering, some miniscule excess of current or voltage, from reversed polarity, some mini-EMP-like static discharge, or kickback from some related high-voltage component.

CAUTION: STATIC-SENSITIVE DEVICE

label on Radio Shack IC package

The same cheap mass production of transistors and IC's that has made possible the world of digital has also encouraged the corruption of the quality of electronic components across the board. Switches, pots, audio transformers, variable capacitors, once built with integrity have become cheaply made mini-junk, and this is often the only stuff readily available to experimenters. Parts suitable for transmitters, tesla coils, and other high-voltage, high-current work, like power transformers, heavy-duty wire-wound pots, chokes, high-watt

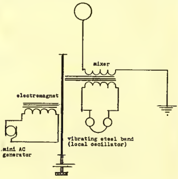
mixer
electromagnet
vibrating steel band
(local oscillator)
mini AC
generator

Tesla's beat receptar

resistors, transmitter variables, vacuum tubes, and insulators, must be obtained from surplus sources. For experimenters mini-junk electronics also means tiny, brittle, vexatious, finger-puncturing, eye-straining connection terminals where there used to be hefty lead wires or sturdy posts.

Miniaturization did not become an obsession until the 1950's and the advent of the miniature and "acorn" tubes. The fashion ultimately reduced the size of all components to the minimum, including coils. But formerly, the size of a coil was correlated with its power. "Note the difference in size," said an ad in an *Electrical Experimenter* magazine of 1917. The ad illustrated a 15,000-meter ham antenna-loading coil against a competitor's smaller counterpart. The advertised coil's diameter was 10 inches, its length 32. A typical Tesla receiving coil was 25 inches in diameter. Tesla built with rugged components and on a shameless scale.

for more information

Some coherer technology is to found in **Induction Coils** (Borderland) For Tesla's free-energy patent and other suppressed Tesla technology: **Tesla: The Lost Inventions** by George Trinkaus (High Voltage Press). For crystal set projects **Radios That Work for Free** by K. E. Edwards (Lindsay). **EMP** is from Rex Research. **Antique Radio Supply** is a source of crystal detectors, tubes, variable caps, and more. (P.O. Box 27468, Tempe, AZ 85285). **All Electronics** for all sorts of experimenter stuff. (P.O. Box 567, Van Nuys, CA 91408). A good radio surplus source: **R5D3** (Portland, OR, 503-513-0410).

High Voltage Press Catalog

What is a 3rd-generation Tesla coil? What is a magnetic amplifier? What is the true wireless?
by George Trinkaus:

Tesla Coil. How Tesla built it. How you can from off-the-shelf parts. Electric energy magnifiers, free energy, beer-bottle capacitor. Lots of background. 26 illustrations. ISBN 0-9709618-0-4 $7.75

Son of Tesla Coil. Sequel to the classic. Utilitarian Tesla, build a Tesla lighting plant, solid-state drivers, oil magic, home power, 3rd generation Tesla coils. 37 illustrations. ISBN 0-9709618-1-2 $7.75

Tesla: The Lost Inventions. Disk-turbine rotary engine, magnifying transmitter, wireless power, high-frequency lighting, free-energy receiver. 42 illus., many Tesla patents. ISBN 0-9709618-2-0 $7.75

Radio Tesla. Who owns the ether? Powerful elemental radio devices, spark reconsidered, grounding rediscovered, low frequency revived, underground radio, license-free radio, carrier current. 66 illustrations. ISBN 0-9709618-3-9 $7.75

Tesla: The True Wireless. Authored by Tesla (1919), ed. by Trinkaus. Tesla's final public statement on the issue. How radio, at the radical, really works. What ionosphere? Everything you know is wrong. Reset and redesigned. 21 illustrations. ISBN 0-9709618-4-7 $7.75

Magnetic Amplifiers. Another lost technology. By the U.S. Navy (1951), ed. by Trinkaus. This facile, rugged device was going to replace the vacuum tube in all functions up to a MHz. Completely reset and redesigned. 43 illus. ISBN 0-9709618-5-5 $7.75

Magnetic Amplifiers Bibliography. A supplement to *Magnetic Amplifiers*. 500 citations of patents, journal articles, institute proceedings, military documents, etc. ISBN 0-09709618-6-3 $7.75

any title $7.75, complete set $39.95

Add for postage and handling: $1.75 for first title, 75 cents each additional. (Outside USA, double that.)

High Voltage Press, P.O. Box 1525, Portland, OR 97207, USA; 877-263-1215; teslapress@aol.com

teslapress.com

Get on our mailing list for news of future titles on Tesla technology.

6. Aerial Capacity

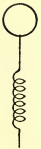

That ball sticking up in the air that is so symbolic of Tesla's radio: What is it? Some sort of antenna? Actually, Tesla never referred to it as an antenna but as an "aerial" or "air capacity" or as an "elevated capacity." Tesla did not see the elevated ball as a radiator, which is how the transmitting antenna of conventional Hertzian radio is construed. The aerial capacity corresponds to the terminal capacitor of a tesla coil. Tesla said that in radio the aerial capacity "heightened the effect" of what is essentially a grounded system. Electrically, in transmitting, it appears to provide a capacitive leverage against which to pump ground. The ball shape also holds high voltage, minimizing coronal discharge and loss.

The ball aerial appears on Tesla's receivers as well as on his transmitters. Tesla understood that a long wire (as in a long-wire antenna) had capacity but he believed the sphere was the efficient geometry. "By using a body of considerable surface...better results are obtained than a wire leading to a height alone... The system is more economical in providing an electrical vibration in the ground." This would be especially true at low frequencies that would require a wire of inordinate length. Tesla's Colorado transmitter when operating at 60 kc would have required a half-wave wire antenna 2500 meters long if modern antenna conventions had been observed.

Tesla also believed that the idea of polarizing, or putting into parallel, transmitting and receiving antennas was nonsense. On his Colorado magnifying trans-

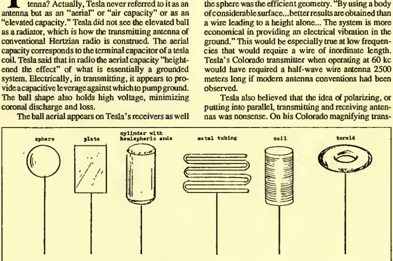

aerial capacity geometries

mitter, Tesla used a hollow copper sphere only 30 inches in diameter. It was thickly coated with rubber insulation. Insulated or not, the ball terminal reduces to a minimum the problem of streamers breaking out at high electrical pressures, since those jump more readily from angular surfaces. Tesla experimented extensively with the effect of the ball's height. An increase in height caused an increase in the effective capacity of the ball, Tesla discovered. (However, of Hertzian radio, Tesla said, "The actions at a distance cannot be proportionate to the height of the antenna or the current in same.")

The ball worked for Tesla but so did other geometries. In Colorado Tesla experimented with a structure of iron pipes as an aerial capacity. He said the aerial capacity could be a cylinder with hemispheric ends, or it could be a toroid. Tesla suggested that a coil of insulated wire put aloft would suffice. He said any hollow vessel, like a ball, could be filled with a gas like hydrogen at low pressure for better effect. Some of Tesla's receiver schematics show the aerial capacity as a simple metal plate.

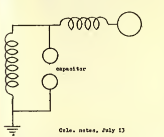

Tesla's opposing-ball capacitor

The ball capacitor shows up in other Tesla circuits besides aerial, consisting of a capacitor of two opposing hollow balls.

capacitive antennas

The "capacitive hat" appears as an occasional element in modern antenna design. Although the texts explain it as something that "improves radiating efficiency," it's still in the nomenclature as a "capacitive hat." Capacitive hats appear frequently in low-frequency radio: in the antenna systems of navigational beacons, of LORAN, of GWEN, and among LowFERs. The capacitive hat appears on top of the helical antenna, which, as a coil of wire aloft, might itself qualify as an aerial capacity.

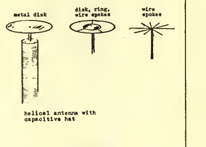

capacitive hats

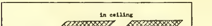

capacity antenna

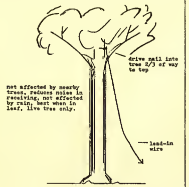
drive nail into tree 2/3 of way to top
not affected by nearby trees, reduces noise in receiving, not effected by rain, best when in leaf, live tree only.
lead-in wire

tree as antenna

Is the loading coil on a center- or top-loading whip "seen" as an aerial capacity?

In the radio literature of the 1920's there are references to a "capacitive antenna." This consisted of a sheet of metal or a wire screen aloft or, alternatively, where a ground connection was distant, a sheet of metal or screen aloft and another below, the two separated by 10 to 15 feet. Recommended as an indoor alternative where an outdoor antenna cannot be put up, the upper element could be placed in an attic, the lower, if any, under the carpet.

Hertzian antennas, like the ham array and the CB whip, are flags announcing the presence of a transmitter. Are they necessary?

Loomis sent kites aloft as aerials; they were covered with copper gauze. Loomis also observed that a tree could be used as a receiving antenna, and others have discovered the tree as a transmitting antenna as well. I've seen no mention of frequencies in the tree-as-antenna literature, which comes from the low-frequency era. (I've tried it for receiving, and it works on long-, medium-, and shortwave bands.) Does the transmitter or receiver "see" the tree as a Tesla-style aerial capacity?

loops

Loop antennas abound in radio and suggest Tesla's elevated-coil idea. You can easily build your own using PVC tubing for the structure. For strength, use one-inch schedule 40 tubing instead of the 1/2 inch I used (photo).

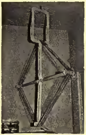
PVC:
1 cross
4 tees
4 els
odd fraction of the quarter wave

author's loop antenna

The shielded loop appears in the LowFER receiver literature as a noise-reduction strategy. The shielding excludes the magnetic component of disturbances while the loop inside responds only to the electrostatic, ac-

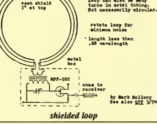
coax; cable tv hardline, RG-59, etc. Loop can also be many turns in metal tubing. Not necessarily circular.
open shield 1" at top
rotate loop for minimum noise
length less than .08 wavelength
metal box
MPF-102
coax to receiver
by Mark Mallery
See also QST 3/74

shielded loop

cording to the literature. A preamp is a good idea with any loop.

studded mushroom

Tesla's magnifying transmitter patent shows a toroid-shaped aerial capacity. It is studded with half-spherical metal plates, presumably to enlarge the surface without increasing by much the bulk. The ultimate Tesla aerial capacity is the studded mushroom that topped his Wardencliff tower. Its diameter was 68 feet. The studding may have been parabolic rather than strictly hemispherical. There is not much information about the tower, and Tesla's reasoning behind the strange geometry is not fully understood.

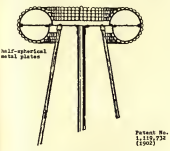

Tesla's magnifying transmitter aerial capacity

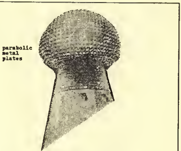

Wardencliff aerial capacity (artist's conception)

The notion of aerial capacity has little currency outside of Tesla-land. What would antennas be like if they were reinvented, not as senders and receivers of Hertzian radiation, but as aerial capacitors?

for more information

Electric Spacecraft Journal, ed. by Charles Yost, has some interesting speculations on Wardencliff's terminal. (P.O. Box 18387, Ashville, NC 28814). **Trees as Antennas** (Rex Research).

WAVELENGTH, FREQUENCY, AND OSCILLATION CONSTANT

Wave-length Meters	Frequency	LC	Wave-length Meters	Frequency	LC
1	300,000,000	.0000003	100	1,579,000	.01016
2	150,000,000	0600011	195	1,538,000	01071
3	100,000,000	.0000025	200	1,500,000	.01126
4	75,000,000	.0000045	205	1,463,000	.01183
5	80,000,000	.0000070	210	1,429,000	.01241
6	50,000,000	.0000101	215	1,395,000	.01301
7	42,860,000	.0000138	220	1,364,000	.01362
6	37,500,000	.0000180	225	1,333,000	01425
0	33,333,000	.0000228	230	1,304,000	.01489
10	30,000,000	.0000282	235	1,277,000	.01555
15	20,000,000	.0000634	240	1,250,000	.01622
29	15,000,000	.0001126	245	1,225,000	.01690
25	12,000,000	.0001760	250	1,200,000	.01760
30	10,000,000	.0002533	255	1,177,000	01831
35	8,571,000	.0003448	260	1,154,000	.01903
40	7,500,000	.0004503	265	1,132,000	.01977
45	6,667,000	.0005700	270	1,111,000	.02052
50	6,000,000	.0007039	275	1,091,000	02129
55	5,454,000	.0008519	280	1,071,000	.02207
60	5,000,000	.001014	200	1,034,500	02366
05	4,615,000	.001188	295	1,017,000	.02450
70	4,286,000	.001378	300	1,000,000	.02533
75	4,000,000	.001583	310	967,700	.02705
60	3,750,000	.001801	320	937,500	.02883
85	3,529,000	.002034	330	909,100	.03066
00	3,333,000	.002280	340	882,400	.03255
95	3,158,000	.002541	350	557,100	.03448
100	3,000,000	.002816	360	833,300	.03646
105	2,857,000	.003105	370	810,800	.03854
110	2,727,000	.003404	380	769,500	.04005
115	2,609,000	.003721	300	769,200	.04277
120	2,500,000	.004052	400	750,000	.04503
125	2,400,000	.004397	410	731,700	.04733
130	2,308,000	.004757	420	714,300	.04966
135	2,222,000	.005130	430	697,700	.05204
140	2,144,000	.005516	440	881,800	.05446
145	2,069,000	.005919	450	666,700	.05700
150	2,000,000	.006335	460	652,200	.05960
155	1,935,000	.006780	470	638,300	.06219
160	1,675,000	.007204	480	625,000	.06485
165	1,818,000	.007662	490	612,200	.06759
170	1,765,000	.008134	800	800,000	.07039
175	1,714,000	.008620	510	588,200	.07327
160	1,667,000	.009120	520	576,000	.07606
185	1,622,000	009634	530	566,000	.07003

WAVELENGTH, FREQUENCY, AND OSCILLATION CONSTANT—*(Continued)*

Wave-length Meters	Frequency	LC	Wave-length Meters	Frequency	LC
540	555,800	.08208	900	303,100	.2759
550	545,400	.08519	1,000	300,000	.2316
560	535,700	08836	1,010	297,000	.2870
570	526,300	.09139	1,020	294,100	.2927
580	517,200	09467	1,030	291,300	.2966
500	508,500	.09801	1,040	288,400	.3045
600	800,000	.1014	1,050	285,700	.3105
610	491,800	.1047	1,060	283,600	.3161
620	483,000	.1082	1,070	280,400	.3222
630	476,200	.1117	1,080	277,800	.3283
640	468,700	.1154	1,090	275,200	.3345
650	461,500	.1188	1,100	272,700	.3404
660	454,500	.1225	1,110	270,300	.3467
670	447,800	.1263	1,120	267,000	.3531
680	441,200	.1302	1,130	265,500	.3595
690	434,800	.1341	1,140	263,100	.3660
700	428,600	.1378	1,150	260,900	.3721
710	422,500	.1419	1,160	258,600	.3767
720	410,700	.1459	1,170	256,400	.3853
730	411,000	.1501	1,180	254,200	.3921
740	405,400	.1540	1,100	252,100	.3988
750	400,000	.1583	1,200	250,000	.4052
780	394,800	.1626	1,210	247,900	.4121
770	389,800	.1668	1,220	245,000	.4190
780	384,800	.1712	1,230	243,900	.4260
700	379,800	.1756	1,240	241,900	.4326
800	375,000	.1801	1,250	240,000	.4397
610	370,400	.1847	1,260	238,100	4469
820	365,900	.1893	1,270	236,200	.4541
630	361,400	.1941	1,280	234,400	.4610
840	357,100	.1985	1,290	232,600	.4683
650	352,000	.2034	1,300	230,800	.4757
860	348,800	.2082	1,310	229,000	.4831
670	344,800	.2132	1,320	227,300	.4906
860	340,000	.2179	1,330	225,600	.4978
800	337,100	.2229	1,340	223,900	5053
000	333,800	.2280	1,350	222,200	.5130
910	329,700	.2332	1,360	220,800	.5206
920	326,100	.2381	1,370	218,900	.5281
930	322,600	.2434	1,380	217,400	.5359
940	319,100	.2487	1,390	215,800	.5438
950	315,000	.2541	1,400	214,300	.5516
960	612,500	.2595	1,410	212,800	.5596
970	309,300	.2647	1,420	211,300	.5674
980	306,100	.2704	1,430	209,800	.5755

tank tables

To calculate the appropriate values for any tank coil and capacitor to resonate at a particular frequency, I find these tables from a 1928 technical-school text clear and useful.. They are also useful for engineering a tesla coil. The column headed "LC" contains numbers representing the "oscillation constant." This is the number you get by multiplying the capacitance in microfarads by the inductance in microhenries. If you know one value, you can get the other by division. The frequency in column two is in cycles; lop off three zeros for kilocycles. See pages 21 –22.

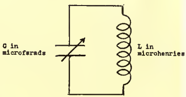

Usually the variable capacitor will be the known and the coil the unknown. Variables are rated in picofarads. Move the decimal point six places to the left to get microfarads.

WAVELENGTH. FREQUENCY, AND OSCILLATION CONSTANT—*(Continued)*

Wave-length Meters	Frequency	*LC*	Wave-length Meters	Frequency	*LC*
1,440	208,300	.5837	1,890	158,700	1.006
1,450	206,900	.5919	1,900	157,900	1.016
1,400	205,500	.5998	1,910	157,100	1.027
1,470	204,100	.6081	1,920	156,300	1 038
1,480	202,700	6165	1,930	155,400	1.049
1,400	201,300	.6250	1,940	154,600	1 060
1,500	200,000	.6335	1,950	153,800	1.071
1,510	198,700	.6418	1,960	153,100	1 081
1,520	197,400	.6502	1,970	152,300	1 092
1,530	196,100	6590	1,980	151,500	1.104
1,540	194,800	6677	1,990	150,800	1.115
1,550	193,600	.8700	2,000	150,000	1.128
1,560	192,300	.6849	2,050	146,300	1.183
1,570	191,100	.6938	2,100	142,900	1.241
1,580	189,900	.7028	2,150	139,500	1.301
1,590	188,700	.7118	2,200	136,400	1.362
1,600	187,500	.7204	2,250	133,300	1.425
1,610	186,300	.7296	2,300	130,400	1.489
1,620	185,200	.7387	2,350	127,700	1.555
1,630	184,100	.7480	2,400	125,000	1.822
1,640	182,900	.7573	2,450	122,500	1.690
1,650	161,800	.7662	2 500	119,000	1.700
1,660	180,700	.7756	2,550	117,700	1.661
1,670	179,600	.7852	2,600	115,400	1.903
1,680	178,600	.7946	2,650	113,200	1 977
1,600	177,500	.8037	2,700	111,100	2 052
1,700	176,500	.8134	2,750	109,100	2 129
1,710	175,400	.8231	2,800	107,100	2.207
1,720	174,400	.8320	2,850	105,300	2.287
1,730	173,400	.8422	2,900	103,500	2.366
1,740	172,400	.8520	2,950	101,700	2 450
1,750	171,400	.8620	3,000	100,000	2.533
1,760	170,500	.8720	3,100	96,770	2 705
1,770	169,400	.8821	3,200	93,750	2 883
1,780	168,500	.8916	3,300	90,910	3.066
1,790	167,600	.0018	3,400	88,240	3 255
1,800	166,700	.9120	3,500	85,910	3.446
1,810	165,700	.9223	3,600	63,330	3.648
1,820	164,800	.9327	3,700	81,080	3.854
1,830	163,900	.9425	3,800	76,950	4.065
1,840	163,000	.9530	3,900	78,920	4.277
1,850	162,200	.9634	4,000	75,000	4.503
1,860	181,300	.9741	4,100	73,170	4.733
1,870	100,400	.9841	4,200	71,430	4.966
1,880	159,600	.9943	4,300	69,770	5.204

WAVELENOTH, FREQUENCY, AND OSCILLATION CONSTANT—*(Continued)*

Wave-length Meters	Frequency	*LC*	Wave-length Meters	Frequency	*LC*
4,400	68,180	5 446	8,800	34,000	21.79
4,500	66,670	5.700	8,900	33,710	22.29
4,600	65,220	5.960	9,000	33,330	22 80
4,700	63,830	6 219	9,100	32,970	23.32
4,800	62,500	6 485	9,200	32,810	23 81
4,900	61,220	6.759	9 300	32,200	24.34
5,000	60,000	7.039	9,400	31,010	24.87
5,100	58,820	7.327	9,500	31,500	25.41
5,200	57,600	7.606	9,600	31,250	25 95
5,300	66,600	7 905	9,700	30,930	26 47
5,400	55,560	8.208	9,800	30,610	27.04
5,500	54,550	8 519	9,900	30,310	27 59
5,600	53,570	8.836	10,000	30,000	28 16
5,700	52,630	9.139	10,500	28,570	31 05
5,800	51,720	9 467	11,000	27,270	34.04
5,900	50,850	9 801	11,500	26,090	37 21
6,000	50,000	10 14	12 000	25,000	40 52
8,100	49,180	10.47	12,500	24,000	43 97
8,200	48,550	10.82	13,000	23,080	47 57
6,300	47,620	11.17	13,500	22,220	51.30
6,400	46,870	11.54	14,000	21,440	55.18
6,500	46,150	11.88	14,500	20,690	59 19
6,600	45,450	12.25	15,000	20,000	63 35
6,700	44,780	12.63	15,500	19,350	87 60
6,800	44,120	13.02	16,000	18,750	72 04
6,900	43,480	13.41	18,500	18,180	78 62
7,000	42,860	13.78	17,000	17,650	81 34
7,100	42,250	14.19	17,500	17,140	86 20
7,200	41,870	14.59	18,000	16,670	91 20
7,300	41,100	15 01	18,500	16,220	96.34
7,400	40,540	15 40	19,000	15,790	101 64
7,500	40,000	15.83	19,500	15,380	107 06
7,600	39,470	16 28	20,000	15,000	112 56
7,700	38,960	16.68	21,000	14,290	124 12
7,800	38,460	17.14	22 000	13,640	136 24
7,900	37,080	17 56	23,000	13,040	148 93
8,000	37,500	15 01	24,000	12,500	162 18
8,100	37,040	18 47	25 000	12,000	175 97
8,200	36,500	18 93	26,000	11,540	190 26
8,300	36,140	19 41	27,000	11,110	205 20
8,400	35,710	19 85	28,000	10,710	220 70
8,500	35,290	20 34	29,000	10,350	236 63
8,600	34,880	20 82	30,000	10.000	253.32
8,700	34,480	21 32			

High Voltage Press Catalog

What is a 3rd-generation Tesla coil? What is a magnetic amplifier? What is the true wireless?
by George Trinkaus:

Tesla Coil. How Tesla built it. How you can from off-the-shelf parts Electric energy magnifiers, free energy, beer-bottle capacitor. Lots of background. 26 illustrations. ISBN 0-9709618-0-4 $7.75

Son of Tesla Coil. Sequel to the classic. Utilitarian Tesla, build a Tesla lighting plant, solid-state drivers, oil magic, home power, 3rd generation Tesla coils. 37 illustrations. ISBN 0-9709618-1-2 $7.75

Tesla: The Lost Inventions. Disk-turbine rotary engine, magnifying transmitter, wireless power, high-frequency lighting, free-energy receiver. 42 illus., many Tesla patents ISBN 0-9709618-2-0 $7.75

Radio Tesla. Who owns the ether? Powerful elemental radio devices, spark reconsidered, grounding rediscovered, low frequency revived, underground radio, license-free radio, carrier current. 66 illustrations. ISBN 0-9709618-3-9 $7.75

Tesla: The True Wireless. Authored by Tesla (1919), ed by Trinkaus. Tesla's final public statement on the issue. How radio, at the radical, really works. What ionosphere? Everything you know is wrong. Reset and redesigned. 21 illustrations. ISBN 0-9709618-4-7 $7.75

Magnetic Amplifiers. Another lost technology. By the U.S. Navy (1951), ed. by Trinkaus. This facile, rugged device was going to replace the vacuum tube in all functions up to a MHz. Completely reset and redesigned 43 illus. ISBN 0-9709618-5-5 $7.75

Magnetic Amplifiers Bibliography. A supplement to *Magnetic Amplifiers.* 500 citations of patents, journal articles, institute proceedings, military documents, etc. ISBN 0-09709618-6-3 $7.75

any title **$7.75** complete set **$39.95**

Add for postage and handling. $1.75 for first title, 75 cents each additional. (Outside USA, double that.)

High Voltage Press, P.O. Box 1525, Portland, OR 97207, USA; 877-263-1215; teslapress@aol.com

www.teslapress.com

Keely's Laws of Being

- on the Nature and Dynamics of Vibration, Oscillation and Toroids

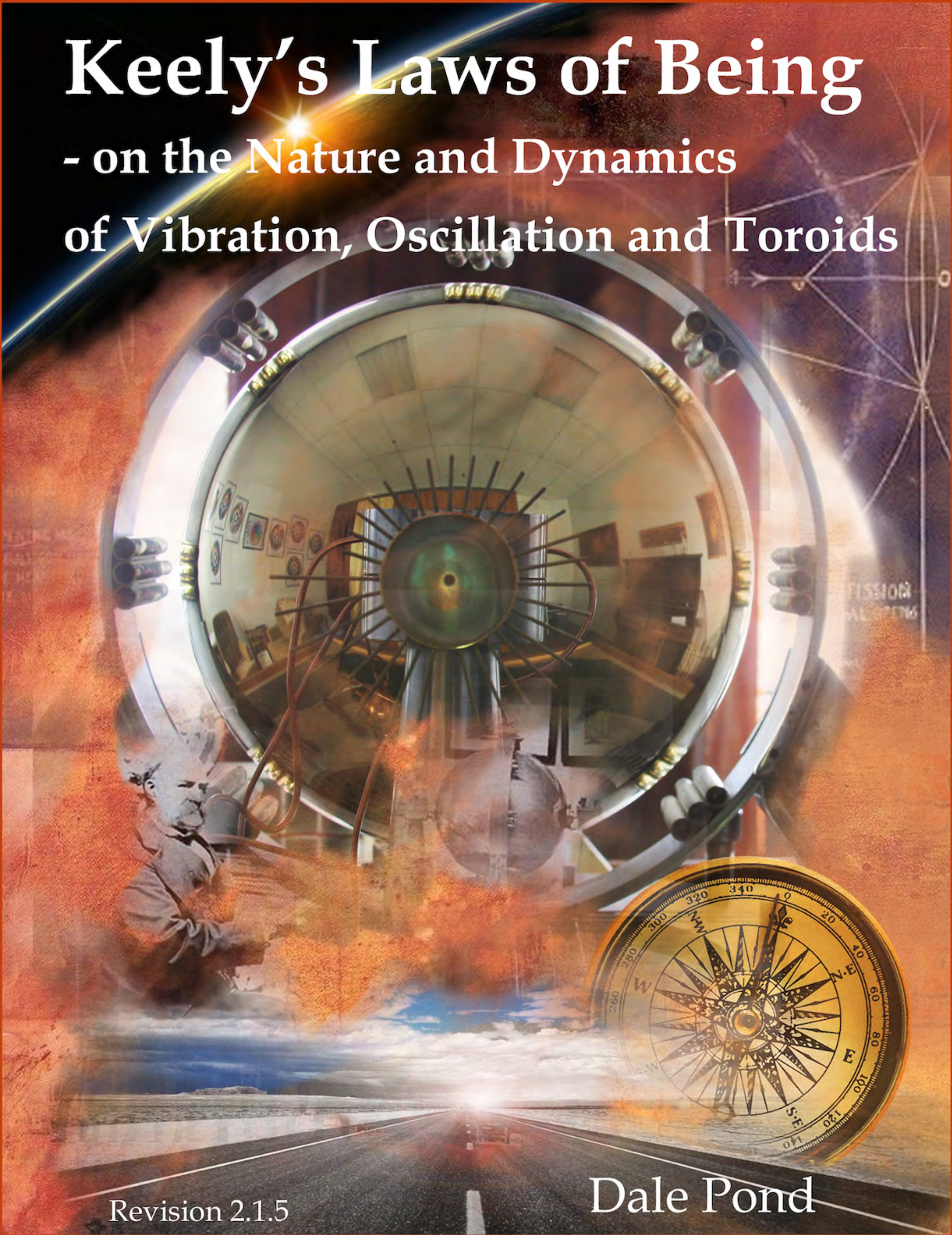
340
320
300
280
260
N.W
W
N.E
E
S.E
20
40
60
80
100
120
140

Dale Pond

Revision 2.1.5

Keely's Laws of Being - on the Nature and Dynamics of Vibration, Oscillation and Toroids, version 2.1.5

by Dale Pond

©Dale Pond, 2013, 2014

''*If you only knew the magnificence of the 3, 6 and 9, then you would have a key to the universe.*" Nikola Tesla

What did Tesla consider as "magnificent" about these three numbers? These three numbers represent unseen dynamics and are the very essence of the nature, structure and dynamics of complex periodic motions. This work describes these three unseen (scalar) aliquot components of vibration, oscillation, vortices and toroids (or any periodic motion) as also their relationships and dynamical functions.

Published by
Delta Spectrum Research
921 Santa Fe Avenue
La Junta, Colorado 81050
720-249-2731
svpwiki.com
svpvril.com
svpbookstore.com

Dedication

This work is dedicated to those rare and courageous individuals who have boundless curiosity in their hearts and minds intensely pursuing a Greater Truth about themselves and the universe in which they live and have their being. They are critical thinking rebels recognizing the status quo does not satisfy their constant and persistent questioning of every thing or idea whether trivial or held as sacred dogma.

"Here's to the crazy ones. The misfits. The rebels. The troublemakers. The round pegs in the square holes. The ones who see things differently. They're not fond of rules. And they have no respect for the status quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do is ignore them. Because they change things. They push the human race forward. And while some may see them as the crazy ones, we see genius. Because the people who are crazy enough to think they can change the world, are the ones who do." Apple Computers

Forward

“There are times in life when all the stars seem to align for us. This work, for Dale Pond, is one of those moments! Dale has worked to recapture John Keely’s fragmented work for over three decades and collate this collection of fragmented footnotes back into John Keely’s coherent science called Sympathetic Vibratory Physics [SVP]. I have known Dale personally for twelve years and studied along side him the works of John Keely. It took me many years to understand in a coherent manner what John Keely was saying. A lot of this confusion was due to his works being so fragmented. Dale has completed and is still working on his SVPwiki.com, which integrates so many others’ works that support the threads of John Keely. Still one must be deeply committed to the study to pick up the basic corollary concepts to develop a clear image of this new science discipline. “Well, no longer! This new publication of Dale Pond’s is surely his greatest work and most fundamental. Here he has laid out basic building blocks of Sympathetic Vibratory Physics in a sequential coherent pattern that helps the reader circumvent years of study to reveal an understanding of this science.

“John Keely was a thousand years ahead of today’s quantum physicist understanding of the basic building blocks of the Universe. John Keely saw it as a three dimensional construct of radially vibrating energy systems from centers in proportional progressive values filling all the plenum of the Universe. He understood through repeatable experimentation the character of this primal energy construct and how to manipulate it’s present stable resonate fundamental states - known matter into new ones - yet undiscovered by modern man. Also, along the way he discovered many wonderful, what would seem magical, characteristics of these new states of matter and force.

“The possible progression from here as I see it, for Dale Pond, is forward into applied application of this now defined science for humanity’s new age.”

Jerry Williams

Preface

0.1 What this book is

This book is a culmination of a two year epiphany. In early 2011 I began sensing there is more to the concept and state of *polarity* than previously thought by anyone save Keely. In July of that year I attempted to speak of this new insight at the Extraordinary Science Conference in Albuquerque. That attempt failed.

Then in October of 2013 the idea gelled - I got it! This book then is an expression of that two year epiphany and over a year of writing.

This work attempts to explain the unrecognized nature, construction and dynamics of vibration, oscillation and toroids with their polar dynamics as also to clarify Professor Daniel Brinton's[1] description of Keely's ideas within his article "Laws of Being"[2] revealing Keely's technique for understanding and manipulating the periodic nature of matter and energy[3]. This Brinton article (Chapter 7 - Laws of Being) is central to this document.

Two prior versions of this book have been issued. This version (2.2 or 4th edit) is a rewrite of the previous versions and has been expanded, reorganized and retitled. The chapters coming before Chapter 7 are foundational materials to understanding Chapter 7. The chapters coming after Chapter 7 are consequential to it.

Keely developed what seems to be entire fields of research mostly akin to our non-classical quantum physics and quantum mechanics decades before these fields existed[4]. Keely was splitting the atom in the 1880s[5] working on his own and with few resources fifty years before mainstream science had a clue. Conventional science at that time period considered the atom to be indivisible. Yet here was Keely splitting the atom and releasing its awesome powers. Because the scientists and engineers "knew" this was impossible most would not could not believe what Keely was demonstrating to them. Political correctness (dogma and doctrine) didn't work well then either.

---

[1] [Daniel Brinton]

[2] Appendix I and II

[3] see Appendix III - Ultimate Constitution of Matter

[4] [Sympathetic Vibratory Physics]

[5] [Keelys Accomplishments]

It should further be noted Keely worked with the forces and laws that govern matter and energy (non-classical physics). Conventional science and engineering works with things and their motions (classical physics). The distinction between these two very different fields is all important. Most people who have witnessed or read Keely and Russell look at their work from the classical viewpoint and miss everything that was intended. Their first reaction is to condemn Keely and Russell for no other reason than they did not understand - not because Keely and Russell were wrong which they weren't. They were simply too far ahead of their time and everyone else's comprehension.

Unfortunately, or fortunately, Keely's research lab notes, books[6] and publications have not been available wherewith all his work could be studied. Therefore this book is limited to exploring Brinton's paper as also the supporting definitions and explanations required to grasp the subject matter.

0.2 Hindsight is mostly 20/20

Within Keely's philosophy and science are several fundamental concepts substantially different from conventional understandings. It must be remembered Keely developed these ideas years and sometimes decades before similar concepts were imagined or considered by orthodox science. After years of study it has been found these points to be not only essential to understanding Keely's work but they are also more accurate conceptualizations. His work has a consistency, continuity, completeness and homogeneity not seen in other earlier models.

Essentially Keely discovered quantum mechanics and quantum physics decades before these fields were suspected to exist. He further devised means to power and operate mechanical devices[7] with these quantum energy level forces - a feat that has yet to be fully achieved by modern science and engineering. Keely's focus was primarily on 1) power development and production,[8] whether by pressure from dissociated matter or rotation 2) discovering and harnessing the ultimate Source[9] of all energy 3) demonstrating Mind is that ultimate source and 4) developing and proving out his theories through his many devices.[10]

---

[6] [Keelys Lost Books]

[7] [Keelys Mechanical Inventions and Instruments]

0.3 Original Concepts

A few of Keely's original concepts that are somewhat different than conventional concepts would be but are not limited to:

- What vibration and oscillation (periodic motion) are.
- Causative (Scalar) constituent components of vibratory motion.
- Dynamical and structural nature of matter and its association with force and energy.
- Mind Force as primordial source of matter and energy.
- Fundamental definitions of basic manifestations of forces and energies such as

1. Electricity
2. Magnetism
3. Gravity
4. Mind / Consciousness
5. Vibration and Oscillation

While many correspondences have been identified between Keely's work and conventional science some like the ones above have no clear correspondence or translation. These are new (for us) concepts and constitutes a new paradigm. It has been found the theories and philosophy of Walter Russell[11] coincides with, compliments and supports in countless ways Keely's ideas. Hence some Russell ideas and artwork are used throughout this book. The few Russell ideas incorporated herein are not all that compliment Keely's work as it would require several volumes to illustrate Russell's work alone. Both their works are vast if not outright encyclopedic in scope. The focus of this work is fairly concentrated and does not attempt to cover the entire gamut of their philosophies and sciences.

---

[8] [Apergy - Power Without Cost]

[9] [Source]

[10] [Keelys Mechanical Inventions and Instruments]

[11] University of Science and Philosophy http://www.philosophy.org

0.4 Analog Signal Generators

Keely developed his science and machines in a time before there were electronic or digital instruments such as signal generators, microphones, amplifiers, signal conditioners, synthesizers, transducers, accelerometers, speakers, etc. Therefore he made use of what was at hand - music instruments and simple disk, rod, cavity and string resonators were his signal generators. He invented devices to produce and test for frequencies in the high ranges of etheric and mind vibrations. A violin, zither, harmonica, stretched strings, tube resonators and chladni plates (gongs) among other simple and complex devices supplied his acoustic signals (simple frequencies) and chords (complex wave forms). He also developed precious metal wires and beads[12] that somehow multiplied multiple frequencies and conducted them without resistance. Peppered throughout the literature are many references to his copious use of diverse music instruments employed to excite various responses from his many devices. Below is a picture of one such device using a zither as part of a much more complicated device that also has within it rod, sphere, tube and plate resonators and a simple horn adorning its top.

---

[12] [Trexar], [Bixar], [Trextrinar]

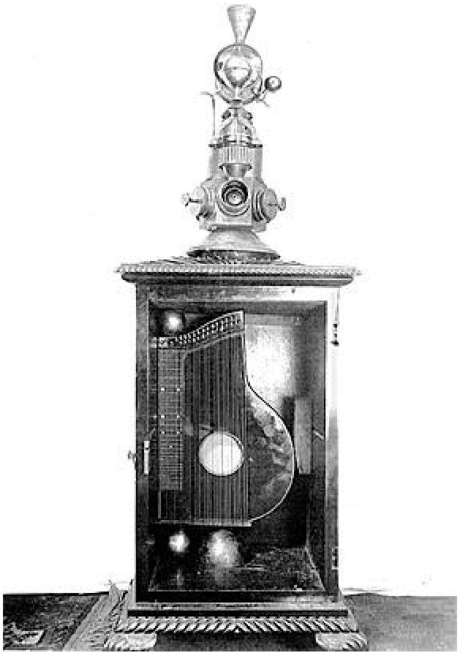

Figure 1: Example of an analog zither as acoustic source
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/zither.jpg

0.5 Music as an Engineering Tool

Music is orderly organized sound. Noise is disorderly unorganized sound. Music allows the intelligent use of simple (notes) and complex tones (chords), time, amplitude, concord and discord (the basis of all vibration), wave form control and how these tools can be used. The study of music science incorporates numbers, arithmetic, geometry, history, philosophy and acoustics. Music is not unlike a shorthand approach to acoustics. Music is of the Whole Mind connecting physical science with higher states of consciousness.

'*'It (the study of music) gives (a knowledge of) a rhythm that is as necessary as law, in making any success in a material experience. Some knowledge of music and some knowledge of law are necessary in the experience of every male individual.*" Cayce (903-3,4)

"That was a happy inspiration which led the Quintet Club, of Philadelphia, to pay a visit to the workshop of Keely a few weeks ago. Its members had been told that the illustrious inventor had employed the power of music to develop the wonderful forces of nature, and evolve by a law of sympathetic vibrations a mighty energy through the disintegration of a few drops of water. Naturally they were anxious to go. They were familiar with the claim by Paganini that he could throw down a building if he knew the chord of the mass[13] *of masonry, and wanted to know if it were possible that the dreams of the great violinist is realized at last.*

"So nearly as can be made out from the mysterious language of the man of many promises, there is a harmony of the universe[14] *that is controllable by the strains of music. Each of the molecules composing a mass of matter is in a state of incessant oscillation, and these movements can be so much changed by means of musical vibration that the matter will be disintegrated*[15]*, its constituent molecules fly apart, and a propulsive force be generated similar to that which is evolved by the touching of a match to a single grain of powder stored in a magazine. He holds that matter is nothing but forces held in equilibrium, and that if the equilibrium be once destroyed*[16] *the most tremendous consequences will ensue.*

"According to the report, he proved to the satisfaction of more than one member of the club that he has already discovered the means of calling out this force, and is able to partially control it. In their presence he caused a heavy sphere to rotate rapidly or slowly, according to the notes given by the instrument on which he played. The sphere was so isolated as to prove that it could not be acted on by electricity or in any other way than by the sound waves. He disintegrated water[17] *into what he calls "etheric vapor"*[18] *by means of a tuning fork and a zither. The disintegration of only four drops of water produced a pressure of 27,000 pounds*[19] *to the square inch, and three drops of*

---

[13] [chord of mass]
[14] see [Sympathy]
[15] [DISINTEGRATION OF MATTER - THREE SYSTEMS]
[16] [Disturbance of Equilibrium]
[17] [THEORY AND FORMULA OF AQUEOUS DISINTEGRATION]
[18] [Etheric Vapor]
[19] [PRESSURES PRODUCED BY VIBRATION]

the harmless liquid fired off a cannon[20] *with a tremendous roar."*[21]

Music incorporates ratio and proportion - the very base constructive elements of vibrations and oscillations.

0.6 A word of warning

It is highly unlikely the reader will grasp the entirety of this material on the first or second read through. One must learn the background ideas, basic or underlying definitions and concepts before the Light dawns in consciousness. This is not unlike learning a new language which does take some time, effort and consequent shifts in consciousness. Do not give up if you don't get SVP within the first five minutes! What you are learning here is PhD level or higher. It takes years of study to earn a PhD - SVP is no different in its level of commitment and materials except that SVP is far more interesting and satisfying.

Do not now or in the future presume what is presented here is the Truth carved in stone. Some is undoubtedly accurate while some not so much. Use your own discernment at all times. This book is my take on these ideas. It is my 'best guess' after years of careful study and deep contemplation. No doubt many will interpret Keely's work differently which is completely OK with me. It will, in my opinion, take many more years if not decades before Keely's work is sorted out, understood and applicable for humanity's benefit. The same goes for Russell's many books and drawings. To do a thorough effort to cover and decipher all that material would take a properly focused and funded team decades of dedicated effort - in my opinion.

0.7 How to use this book

Where terms or phrases are enclosed in brackets ([ ]) the reader is encouraged to look these up within the svpwiki.com. These bracketed terms are mostly in the footnotes at the bottom of each page. If a concept, word or phrase is not understood look it up. A good dictionary resource is onelook.com[22]. Learning Keely is learning a new jargon describing a new paradigm - it takes time to assimilate these new-to-us ideas. On top of that one is obliged to learn the jargons of several fields such as quantum mechanics and physics, chemistry, math, music, acoustics, metaphysics and several others. It then is a hard climb to the peak of understanding but the view will be worth it.

---

[20] [Cannon]

[21] [Keely and His Discoveries]

Due to the nature and expense of the printing process, especially color, imbedded images are black and white smaller low resolution facsimiles of the originals to which have been included below each image a link to the original color and/or higher resolution image.

0.8 Generalities and Idealism

The concepts covered herein are mostly generalities. There are some specifics. Usually a general concept is idealized for simplicity's sake. All phenomena are subject to countless influences, some major and some quite subtle all causing some degree of modulation in the simple models presented. A good example covered in 'Chapter 12 - Vortices and Toroids'. A torus is usually presented looking somewhat like a common donut. This is a gross over-simplification as the structural and functional elements in toroids are many, complex and dynamic. The treatment on these complex toroid structures and dynamics takes the basic idea then ventures into a bit more detail. One's lifetime is not enough to cover all the details of all the subjects touched on within Sympathetic Vibratory Physics [SVP][23]. I'll do what I can to provide details as I am able. Other unresolved details are left for later writings and to those so interested and driven to ferret them out.

Dale Pond[24]
La Junta, Colorado

---

[22] It has been found the older dictionaries are better to understand Keely and the science of that time period than newer dictionaries.

[23] [Sympathetic Vibratory Physics]

[24] [Dale Pond]

Contents

0.1 What this book is 4
0.2 Hindsight is mostly 20/20 5
0.3 Original Concepts 6
0.4 Analog Signal Generators 7
0.5 Music as an Engineering Tool 8
0.6 A word of warning 10
0.7 How to use this book 10
0.8 Generalities and Idealism 11

1 Ether as a Working Fluid **27**
1.1 Something old something new 27
1.2 Energy Levels 28
1.3 Classification of Energy Levels, States or Phases 29
1.4 Michelson-Morley 31
1.5 Etheric or Quantum Carriers 32

2 Love in Science and Physics **35**
2.1 First and foremost Love is sympathy 35
2.2 Love in Science and Physics 35
  2.2.1 Sympathy essential 35
  2.2.2 The state or condition of Sympathetic Association . . . 36
  2.2.3 Looking at the substance of a rock 36
  2.2.4 Sympathy Connects Everything 37
  2.2.5 Sympathy is simply another word for Love 37
  2.2.6 Sympathy is Attractive 37
  2.2.7 Individuated Complex Wave Functions 38
  2.2.8 The state of pure Sympathy 38

3 Sympathy **41**
3.1 Sympathy is Love 41
3.1.1 Infinitely Small Centers 42
3.1.2 Atomoles 43
3.1.3 Records of Motion are Seeds 43
3.2 Neutral Center of the Earth 44
3.2.1 Force of Sympathy 44
3.2.2 Sympathy is the Essence of Life 44
3.2.3 Sympathy is as Syntropy and Electricity 45
3.2.4 The rate of acceleration of Sympathy is enormous . . . 47
3.2.5 Sympathetic Associative Forces 48
3.2.6 Sympathy is Entanglement 49
3.3 Sympathy defined 50
3.4 Recommended reading 52

4 Neutral Center **53**
4.1 Neutral, Latent Neutral and Neutral Center 53
4.2 Neutral means 53
4.3 Coordination or coincidence of forces 55
4.4 Neutral Center 58
4.4.1 Disturbance of Equilibrium 59
4.4.2 Supreme Neutral Center 61
4.4.3 Keely further describes the Neutral Centers 63
4.4.4 Law of Transmissive Vibraic Energy 64
4.5 Divine Will - The Wheelwork Source of all Force 70
4.5.1 Center of Sympathetic Attraction 74
4.5.2 Center of Sympathetic Coincidence 74
4.5.3 Center of Association and Concentration 74
4.5.4 Center of Focalization and Transformation 75
4.5.5 Nucleal Center 75
4.5.6 Resistance 77
4.5.7 Prime Neutral Center 80
4.5.8 Supreme Neutral Center 80

5 Mind and Mind Force **83**
5.1 Mind is Force 83
5.2 Whole vs Split Mind 84
5.3 Mind is of the Sixth and Seventh Subdivisions 85

5.4 Keely - Substance and Nature of Mind Force 86
5.5 Russell on Mind and Mind Substance 89
5.6 Mind is a Substance 90
5.7 Mind is also as an aggregate of consciousness 91

6 What is Vibration? 93
6.1 Quotes - vibration defined 93
6.2 A Vibration Is 94
6.2.1 Three Forces action on all Matter and Energy 95
6.2.2 Law of Attraction (syntropy) 96
6.2.3 Law of Repulsion (entropy) 96
6.2.4 Law of the Dominant (Celestial Mind Force) 96
6.3 Rad-Energy - entropic dispersion 96
6.3.1 Introductory Impulse 97
6.4 Vibration v Oscillation 97
6.5 What a Vibration is NOT 98
6.6 Wave Propagation 99
6.7 Frequency 100
6.8 Law of Cycles 100

7 Laws of Being 103
7.1 Introduction 103
7.2 Rhythmical Relations 104
7.2.1 Principle of Rhythmic Balanced Interchange 104
7.2.2 Every individualized object assimilates itself 104
7.2.3 These modes are today much like "phase" 107
7.2.4 'Thirds' is usually a mathematical relationship of proportion 109
7.3 Threefold Relations 109
7.3.1 Each of the above dynamics (IV) is a $\frac{1}{3}$ part or current of a $\frac{3}{3}$ stream 110
7.3.2 Ratios of Harmonic and Enharmonic Potentials 111
7.4 Three Actions and Laws of Being 111
7.4.1 Law of Sympathetic Association 112
7.4.2 Law of Assimilation 113
7.4.3 Law of Attraction 113
7.4.4 Law of Sympathetic Oscillation 113
7.4.5 Infinite neutral Mind centers and controls all motion . 114

  - 7.4.6 One Substance is Mind . . . 115
- 7.5 Three Modes of Vibration . . . 117
  - 7.5.1 Polarity, Duality, Bias and Potential . . . 117
- 7.6 Three Laws of Being . . . 118
  - 7.6.1 Dominant Mode . . . 118
- 7.7 Table of Modes Correspondences . . . 120
  - 7.7.1 The above dynamics are to the following effects . . . 120
  - 7.7.2 Relations of Thirds . . . 121
- 7.8 Three Modes, Currents and Streams . . . 121
  - 7.8.1 Ratios and Proportions of Thirds . . . 121
- 7.9 Evolution of Matter . . . 123
- 7.10 Neutral Center . . . 123
  - 7.10.1 Center of Sympathetic Coincidence or Neutral Center . 124
- 7.11 Syntropy or Mutual Affinity . . . 125
- 7.12 Subdivisions of Individualized Being . . . 126
  - 7.12.1 Table of Subdivisions of Matter and Energy . . . 126
- 7.13 Arithmetical and Geometric Progressions . . . 127
- 7.14 Progressions of Orders . . . 127
- 7.15 Infinite Range of Vibration . . . 127
- 7.16 Progression Types . . . 128
  - 7.16.1 Arithmetic and Geometrical Progressions . . . 128
- 7.17 Three Major Orders . . . 129
  - 7.17.1 The above are not “things” . . . 129
- 7.18 Controlling Mode . . . 130
  - 7.18.1 Modes combine in unique proportions . . . 130
- 7.19 Modes as actual components . . . 130
- 7.20 Foundational Concepts . . . 131
  - 7.20.1 Balance or Equation of Father-Mother Forces . . . 131
- 7.21 Latent Force . . . 132
  - 7.21.1 Matter is the result of three forces or modes of vibration 132
  - 7.21.2 Latent Force, Keely . . . 135
- 7.22 Sympathy and Resonance . . . 136
  - 7.22.1 Attraction and a Repulsion between all things . . . 136
- 7.23 Matter composed of Molecules . . . 137
  - 7.23.1 Attraction is an effect of Syntropy . . . 137
- 7.24 Molecules are sensitive to modes . . . 138
- 7.25 Oscillations of Molecules . . . 138
- 7.26 Disturbance of Equilibrium . . . 139

  - 7.26.1 Disturbance of Equilibrium, Gravity 140
- 7.27 Sixths induces Neutral Radiation 141
  - 7.27.1 Seeming Paradox 142
  - 7.27.2 Exciting the Molecular Mass 142
- 7.28 Third is the Dominant 142
  - 7.28.1 Harmonic Resonant Impulse 142
- 7.29 The Dominant or Etheric Mode 143
  - 7.29.1 Etheric Current 143
- 7.30 Centers of Harmonized Vibration 144
- 7.31 Sympathetic Outreach (sympathy/resonance field effect) . . . 144
- 7.32 Etheric Outreach 145
  - 7.32.1 Law of Transmissive Vibraic Energy 146
  - 7.32.2 Etheric vibration is the cause of sympathetic association 146
- 7.33 Chord of Mass 147
  - 7.33.1 Qualifiers 147
  - 7.33.2 Law of Harmonic Pitch 147
- 7.34 Controlling Chord of Mass 147
  - 7.34.1 Equilibrium or Balance cannot exist 148
  - 7.34.2 Law of Cycles 148
  - 7.34.3 Table of Transformation over Time 148
  - 7.34.4 Law of Transformation of Forces 149
- 7.35 Harmonizing Chord of Mass 149
- 7.36 Molecular Oscillation 149
- 7.37 Transmitting Concordant Impulses 150
  - 7.37.1 Law of Vibrating Atomolic Substances 150
  - 7.37.2 Frequencies of the Modes 151
- 7.38 Vibratory Rotation 151
  - 7.38.1 Microwaved Radiometer 151
- 7.39 Rotation 152
  - 7.39.1 Medium for inducing a Neutral Center 152
  - 7.39.2 Keely on Rotation 153
  - 7.39.3 Power of Rotation 155
  - 7.39.4 Receptive Transmittive Concordance induces Rotation 156
  - 7.39.5 Polar and Depolar Differentiation result in Motion . . . 156
  - 7.39.6 Neutral Center or fulcrum represents Focalization . . . 157
  - 7.39.7 Rotation of the ether produces condensation 157
  - 7.39.8 Controlling “negative attractive” 158
  - 7.39.9 Testing Positive and Negative Rotation 158

7.39.10 Russell on rotation 158
7.40 Controlling negative attraction 159
7.41 Notes 159

8 Modes of Vibration 161
8.1 Two models of three modes 161
8.2 Three Modes of Vibration 162
8.2.1 Compression or Longitudinal Wave or Mode 162
8.3 Three Modes of Vibration 162
8.3.1 Transverse or Shear Wave or Mode 163
8.3.2 Rayleigh, Lamb or Love Wave or Mode 163
8.4 Introduction to Keely's Wave Function 165
8.5 Feynman Morphology 166
8.6 Footnotes 169

9 Keely Wave Function 171
9.1 Introduction to Keely's Wave Function 171
9.1.1 Law of Transformation of Forces 173
9.2 Thirds as Ratios within a Whole 174

10 Proportions (Ratios) of Relative Values 179
10.1 Proportion 179
10.2 Reciprocating Proportionality 180
10.3 Reciprocating Proportionality 181
10.4 Reciprocal Proportion 182
10.5 Reciprocals and Proportions of Motions and Substance 184
10.6 Thirds as Ratios within a Whole 185

11 Universal Heart Beat 187
11.1 The Love Cycle 188

12 Vortices and Toroids 201
12.1 Vortices 201
12.2 Vortex Components 203
12.3 Vortex Dynamics 204
12.4 Vortex Action is rotary 204
12.5 Vortex Center 205
12.6 Vortex Polar Forces 207

12.7 Vortex Periphery . . . 207
12.8 Vortex Wall . . . 208
12.9 Toroid . . . 208
  12.9.1 Sympathetic Associative Attraction . . . 209
  12.9.2 Law of Cycle . . . 210
  12.9.3 Neutral Center Repulsion . . . 211
  12.9.4 Periphery Neutralization or Depolarization . . . 212
12.10 Plane of Inertia . . . 212
12.11 Free Space . . . 213
12.12 Etheric Envelopes . . . 215

13 Electricity and Magnetism **219**
13.1 Electricity is a Polar State Dipole . . . 219
  13.1.1 Voiding the Dipole or Polarity . . . 219
  13.1.2 Conduction . . . 220
  13.1.3 Polar and Depolar States . . . 220
13.2 Keely on Electricity . . . 220
13.3 Russell on Electricity . . . 226
13.4 Tesla on Electricity . . . 230
  13.4.1 Summary . . . 232

14 Gravity **233**
14.1 Keely on Gravity . . . 233
14.2 Keely and anti-gravity . . . 235
14.3 Levitation experiments . . . 237
14.4 Russell on Gravity . . . 240

15 Bibliography and References **243**

A I - Laws of Being **247**

B II - Laws of Being **255**

C III - Ultimate Constitution of Matter **263**

List of Figures

1 Example of an analog zither as acoustic source 8

3.1 Opposite polar forces void/combine as One 47

4.1 Universal Seesaw balancing seeming opposites 54
4.2 From the One comes the many 63
4.3 Coincidental forces meet at center 66
4.4 Neutral Center - creator of nebula and galaxies 67
4.5 Keely's Compound Disintegrator 68
4.6 Liberating and measuring Etheric Vapor pressures 69
4.7 An earlier Liberator design 70
4.8 Sympathy Links all Centers to all other Centers 71
4.9 Neutral Center is Mind or Consciousness of God 72
4.10 Neutral Center is as mind or consciousness 73
4.11 Syntropy winding up force to 4++ center 76
4.12 Entropy unwinding energy to 0 Plane of Inertia 77
4.13 Russell's 4 plus plus center of polar opposition 78
4.14 Cavitation is an example of the Law of Cycles 79

5.1 Whole and Split Mind 85
5.2 Red and Blue halves of the visible light spectrum 89
5.3 Red and Blue halves of Polar states 90

6.1 Four Phases of a Wave or Oscillation 100

7.1 Sympathetic Vibration - Entanglement 107
7.2 Chart showing location of the three mode frequencies. 108
7.3 Triune Forces/Modes/Proportions combined as One Force . . 110
7.4 Typical Triplet Chords in Keely's charts. 111
7.5 From the One comes the many. 115

7.6 Mind centers all motion. 116
7.7 Every point in the universe is a gravitative center. 124
7.8 Table of seven subdivisions or orders. 126
7.9 Keely Motor activated by slow released etheric vapor 134
7.10 Keely's cannon fired by quick release of etheric vapor 135
7.11 Table of Transformation of Forces over Time 148
7.12 Globe motor actuated by rotary vibratory dynamics 153
7.13 Keely's globe motor and provisional engine 154
7.14 Keely and his globe motor or dynasphere 154
7.15 Pond's replication of the globe motor or dynasphere 155

8.1 Coiled Spring showing Longitudinal Wave 162
8.2 Compression and Rarefaction in a Wave 163
8.3 Transverse Wave 163
8.4 Relationship of Three Modes 164
8.5 Suppositional Math and Symbolic Structure 165
8.6 Feynman's triplet structures of the proton and neutron 166
8.7 Table of ranges of forces, forms and laws 167

10.1 One Balanced Whole and Two Reciprocating Dynamics 181
10.2 Plane of Inertia 183
10.3 Scale of 0 to 4 to 0 183
10.4 Nested cubes or increasing and decreasing proportions 184

11.1 Universal Heartbeat 187
11.2 Neutral Center 4++ extention to periphery 0 191
11.3 The basic formula of Creation 198
11.4 Formula of Rhythmic Balanced Interchange 199

12.1 Differential densities, pressures and masses 202
12.2 Tornado Vortex 203
12.3 Vortex showing a few of its many components 204
12.4 The depolar One divides into duality/polarity 206
12.5 Triune Forces of a vortex 207
12.6 Typical idealized torus, toroid or "donut" concept 208
12.7 Ideally toroids are two vortices meeting apex to apex 209
12.8 Wave elements of Time, Compression and Dispersion 210
12.9 Compression to center then equatorial dispersion 211
12.104++ at center of transformation 211

12.11 Depolar state at plane of inertia 212
12.12 Triune polar forces as One Force about a common center . . . 213
12.13 Three double vortices 214
12.14 Keely's three envelopes or shells 215
12.15 Three overlapping envelopes or shells sealing poles 216

13.1 Neutral Center governing seeming opposite potential states . . 222
13.2 Triune Forces of Magnetic Field in Magnet 225
13.3 White Light of Mind divided into red and blue halves 227
13.4 Attractive (syntropic) power of electricity 228
13.5 Polar dynamics of electricity 229

14.1 Keely's experimental air ship 237
14.2 Levitating weights 238
14.3 Gravity wave bar and poles 241

List of Tables

5.1 Keely's Seven Subdivisions of Matter and Energy 86

7.1 Three Polarity States and Effects 118
7.2 Table of Mode Correspondences 120
7.3 Some Mode Effects 120
7.4 Thirds, Sixths and Ninths 121
7.5 Mode Frequency Ranges 128
7.6 Hypothetical controlling Modes and Proportions 130
7.7 Thirds, Sixths and Ninths Intervals 140

8.1 Modes of Vibration 166
8.2 Table of attributes of the three modes 168
8.3 Keely's three modes; subdivision, octave and law 168

9.1 Ranges of forces vibration forms types and laws 173
9.2 Neutral (thirds); energy radiates from center (sixths); force contracts to center (ninths) 176

Chapter 1

Ether as a Working Fluid

1.1 Something old something new

Has the Victorian Ether[1] been rediscovered? I believe it has. Science has been brow-beat to never ever use the term "ether" or *else*. Such a threat has perpetuated a clear state of ignorance in direct proportion to holding back true scientific progress. The too often used blanket statement *"Ether does not exist"* is nothing short of dogma and therefore decidedly unscientific. But no one (openly) questions that statement out of reactionary inquisitional fear. Had any questioned it they would have seen the term "ether" is 1) a general concept and not specific and 2) there is obviously 'something' unseen and unacknowledged operating at finer levels of 3) classifications or states of matter and energy. It would require a book-length expose to show not only does the ether exist and has always existed but it is now being seen by conventional science under different eyes because the general catch-all term '*elemental particle*' is being applied to a multitude of new classifications of sub-atomic or quantum entities both elementary and composite. We could take all these 'new' quantum entities or higher classifications of matter and energy states and interactions and lump them under the term "ether" and we would then know what ether is. Ether to SVP is simultaneously a 'classification' or 'state' and depending on context a 'thing'. This puts us into a nice position of understanding what Keely called "Ether", "Etheric Vapor" [2], *etc.* and still be on a solid scientific footing which none could successfully assail.

---

[1] [Ether]

[2] [Etheric Vapor]

Quark-Gluon Plasma (QGP) "The strength of the color force means that unlike the gas-like plasma, quark-gluon plasma (QGP) behaves as a near-ideal Fermi liquid, although research on flow characteristics is ongoing. In the quark matter phase diagram, QGP is placed in the high-temperature, high-density regime; whereas, ordinary matter is a cold and rarefied mixture of nuclei and vacuum, and the hypothetical quark stars would consist of relatively cold, but dense quark matter."[3]

Compare the above description to

Liquid Ether *"The atomolic[4] substance is what is termed the ether which fills all space and is the transmitting medium for all celestial and terrestrial forces. This is the liquid ether of occult science."* Keely

Both named substance are 1) subatomic entities, 2) liquids and are 3) conducting or transmitting[5] mediums of/for diverse forces. I do not at this point consider QCP as elementary as it is a composite conventionally classified within the "Very high energy state" of matter and energy. It therefore ought to be classified to correlate with Keely's Interetheric subdivision[6]. To ferret out all this "new" information one could begin with the svpwiki.com pages "State" and "Subdivision" then proceed to Wikipedia, States of Matter, *etc*. Again many volumes could be researched and written about this one aspect of SVP correlated to conventional science.

1.2 Energy Levels

For the purposes of this document "Ether" (quantum energy level substances) is considered as a *Working Fluid* not unlike refrigerants in a cooling system where liquid is converted into gas state then back to liquid state by varying pressure and temperature. Molecular substances that can undergo this type of repeating phase change without changing their chemistry are called

---

[3] Wikipedia/Quark-Gluon Plasma

[4] [etheric]

[5] [Connecting Link]

[6] [Subdivision]

azeotropic fluids of which there are a great variety on the market. The only thing that changes within such a system is the Wave Function[7] (mathematically defined state) of the Working Fluid. In this context Ether, being and viewed as an azeotropic fluid, may be (generally speaking)

Molecular, as in liquid water or water vapor (composed of molecules)

Atomic, as in Hydrogen and Oxygen and (composed of atoms)

Etheric, as in Electrons, Protons, Neutrons, Photons, Quarks and Gluons, etc. of which there are a great variety. (composed of quantum entities)

Hence there are many types and variations of these three main classifications each possessing their own identifying state or Wave Function. These states or Wave Functions are interchangeable according to Keely's Law of Transformation of Forces[8]. The first two above states operate within what is called *classical* physics (sometimes called Newtonian Physics, 3D, materiality, maya, *etc.*) and obey its laws. The third state or classification operates within *non-classical* physics (sometimes referred to as Quantum Physics, Quantum Mechanics, Quantum Electrodynamics, Subtle Energy Physics, Mind Physics, Scalar Physics or Spiritual Physics). As is well known the third state or classification of non-classical physics does not operate according to principles of classical physics even though many foolishly attempt to apply classical physics laws to these subtler (non-classical) states of matter and energy.

1.3 Classification of Energy Levels, States or Phases

SVP establishes a seven-level matter and energy classification having the three main divisions of molecule, atom and ether. The modern classification is not as simple or neat. Modern science has identified multiple classes or types of *associations of quantum entities*. Several of these associative states are theoretical meaning they are hypothesized but not yet "proven". Several others are well established and accepted.[9] The modern states that are listed below may be associated with Keely's "Etheric" state or subdivision are (but not limited to)

---

[7] [wavefunction]

[8] [Law of Transformation of Forces]

I - Low Temperature states

1. Superfluid
2. Bose-Einstein condensate
3. Fermionic condensate
4. Rydberg molecule
5. Quantum Hall state
6. Strange matter
7. Photonic matter

II - High-energy states

1. Degenerate matter
2. Quark-gluon matter [10]
3. Color-glass condensate

III - Very high energy states

The High-energy states may be considered in this document as 'states of ether' of which there are countless variations. It is not our intent to delve into the intricate details of these states. If is however our intent to explore and clarify how some aspects of these etheric states associate, behave and have their being. These high energy states as defined as ether or *'atomolic substances'*:

"The atomolic substance is what is termed the ether which fills all space and is the transmitting medium for all celestial and terrestrial forces. This

---

[9] [Wikipedia; States of matter]

[10] [svpwiki.com; Quark-gluon plasma]

is the liquid ether of occult science." Keely

"*For convenience' sake we will use the term atomolic in place of etheric in our subsequent definitions.*" Keely[11]

1.4 Michelson-Morley

Keely's quantum discovery work preceded even the discovery of the electron in the early 1890s[12]. Keely's pioneering work with *splitting the atom* into its constituent quantum elements predates quantum mechanics and quantum physics by decades. This accomplishment[13] predates any officially recognition fields such as these existing within quantum realms of matter and energy - because they didn't exist to victorian science unless it be referred to as *Ether*. In fact the atom was considered as *indivisible* by orthodox science of that day. No scientist or engineer of that time period had a clue what Keely was doing because his ideas were never suspected by science or engineering. In consequence he developed his own terms which worked for him as he applied his findings to design and operate his innovative hybrid quantum/mechanical devices[14][15] of diverse designs and purposes - mostly to prove out his theories as he constructed his paradigm of Sympathetic Vibratory Physics which he sometimes called Vibratory Etheric Physics or simply Vibratory Physics. Of course as there were no other scientists or engineers aware of quantum mechanics and physics in Keely's day there was no one who could understand or verify what he was doing. The best they could do (the honest ones of integrity) was to witness Keely's demonstrations and report what they thought they saw without offering any explanations.[16] The dishonest ones or those who felt threatened by knowledge they knew nothing about (despite their lofty degrees) condemned everything Keely did as fraud or fakery[17].

From the above and with 20/20 hindsight we can see the Michelson-Morley experiment did not prove the ether did not exist. It simply proved the ether or the now known to exist high-energy states could not be detected or measured with his crude equipment built and operated with their incomplete classical physics paradigm that did not include non-classical physics even though is what they were hoping to verify.

---

[11] [Dashed Against the Rock, p. 47]

[12] Depends who you ask.

[13] [Keelys Accomplishments]

[14] etheric or quantum forces powering mechanical machines

[15] [Keelys Mechanical Inventions and Instruments]

[16] [Eye Witness Accounts]

[17] [Was Keely a Fraud]

Keely said etheric substances are so tenuous as to pass through seeming solid matter as wind through *"a coarse sieve"*. If his view is correct there was no possibility Michelson and Morley were going to impede or measure any resistance caused by ether.

"The negative sympathetic polar stream is the magnetic flow proper, and it is in sympathetic coincidence[18] *with the second atomic flow, the electric current is the first and second order of atomic vibration, a dual force, the flow of which is too tenuous to displace the molecules. It can no more do so than the flow from a magnet can displace the molecules of a glass plate when it is passed under it. The flow from a magnet is too fine to disturb the plate molecules, but passes as freely between them as a current of air would through a coarse sieve."* (underlines added)[19]

1.5 Etheric or Quantum Carriers

These high-energy states or tenuous substances are *carriers* or conductors of subtle imperceptible forces or disturbances that once excited will manifest as other perceptible detectible phenomena. These streaming forces are Keely's Sympathetic Streams[20]. These fine subtle forces are therefore the unseen and often unsuspected scalar or causative forces manifesting as other lower frequency coarser effects of attraction, repulsion and rotation. This book is about identifying and manipulating these subtle unseen constituent forces that we may be enabled to create desirable effects such as magnetism, electricity, health, gravity effects, rotation and all of their attributes and variations. As carriers of excitation or disturbances[21] they are therefore as connecting links[22] between the seven subdivisions[23] of matter and energy states thus affecting transformations or transmutations of these energy states[24] between themselves whether moving into or out of higher or lower energy ranges manifesting as *centralizations* or as more commonly called - *particles.*[25] These unseen and unacknowledged subtler forces are the constructive constituent elements making up the wave function of all known states of matter and energy.

---

[18] [sympathetic coincidence]
[19] [Keely], [ATTRACTION PROPULSION ETC], [16.21 - Magnetic Flow]
[20] [Sympathetic Streams]
[21] [13.12.1 - Disturbance of Equilibrium]

One of the more significant attributes of this Dirac Sea-like[26] substratum of the universe is it's vibratory uniformity and harmony establishing a sympathy between centers or centralizations which modern science now calls *quantum entanglement.* Keely called this ubiquitous harmonic state of being between entities *sympathetic association* and is governed by the Law of Sympathetic Association[27]. The active link between centralizations or particles is their inherent vibrations and oscillations. When these periodic motions form unisons or harmony between secondary and tertiary harmonics are called sympathetic vibration and sympathetic oscillation. It is this sympathy between periodic motions that is the cause and link seen in "action at a distance"[28] phenomena such as quantum entangle[29]ment, remote viewing, mind to mind[30] communication, telepathy, telekinetic actions, *etc.*[31] As we will see another synonym for this universal sympathetic attribute of and between all centralized energy states (centralizations, particles) is Love (as one).

---

[22] [connecting link]

[23] [Subdivision],[State]

[24] [Law of Transformation of Forces]

[25] [connecting link], [Vibratory Physics - The Connecting Link between Mind and Matter], [The Connecting Link Between Mind and Matter - Keelys Progress - Part 2]

[26] [Dirac Sea]

[27] [Law of Sympathetic Association]

[28] [action at a distance]

[29] [telepathy]

[30] [mind to mind]

[31] [action at a distance]

Chapter 2

Love in Science and Physics

2.1 First and foremost Love is sympathy

First and foremost Love is sympathy or that which connects (entangles) all things to all else. In SVP, sympathy does NOT mean “feeling sorry” but means “feeling the same as”, “as one” or “one with”. Some might call this “empathy”, “harmonization”, “resonance”, “bonding”, “coupling” or “entanglement”.[1]

> “*Love is the Creative Force of All that Is.*”[2]

2.2 Love in Science and Physics

2.2.1 Sympathy essential

This harmonious high-energy state of matter and energy is everywhere the same. It, by its nature, is that which connects everything to everything else. Without this unity of being the universe would be fragmented, separate and discontinuous. It is a universal bonding agent without discontinuities. Should anything be disturbed everything else responds in like fashion as their essences (constituent elements) are entangled.

---

[1] [Sympathetic Association], [Law of Sympathetic Association], [Law of Assimilation], [Harmony], [Affinity], [Attraction], [Awareness], [Knowing], [Universal Love], [Christ], [Science], [Truth], [Coupling]

[2] [Dialogue on Awakening], page 160

"Where there is no sympathy there can be no love, for love is sympathy."[3]

2.2.2 The state or condition of Sympathetic Association

The state or condition known as Love is important to science and physics because this state or condition is what underlies and connects all seeming individuated things. Love is a condition or state of vibration that links or entangles all things to all other things everywhere, essentially eliminating Time and Space. Which is to say all things are quantum entangled. This statement appears false as we do not perceive it to be the case when we see every thing as seemingly being separate and distinct. Our perception (opinion) is of the physical or molecular state, form and color of things. For instance, a rock is a molecular mass appearing to have an individual existence all by itself. So we see the mass which we can pick up and move it around and not physically see a connection to any other thing or rock not realizing the rock is a portion of the earth and is resonant to the earth else it would not remain either on the earth or in a molecular state or both. *Should we consider just the electrons and photons making up every thing that is, regardless of location, as being essentially the same, vibratorily speaking, how can they not be entangled?*

2.2.3 Looking at the substance of a rock

Looking at the substance of a rock (for instance)
We know
1 - it is composed of molecules which are composites of atoms.
2 - these atoms are composed of electrons, protons and neutrons.
3 - these elementary entities are composed of photons.
4 - photons are composed of quarks[4].
5 - all these above entities vibrate and oscillate perpetually at given rates and amplitudes and as *modulated by their near and far vibratory environment.*
6 - that individual objects of the same frequencies or chords of frequencies are sympathetically linked or entangled to and with each other. As every type of

---

[3] [Vera Vita the Philosophy of Sympathy]

[4] https://en.wikipedia.org/wiki/Photon_structure_function

quark, photon and electron are essentially the same[5] no matter where they are we can say all these particles are sympathetically linked or entangled to each other to varying degrees or gradients. They have the same or very similar Wave Function[6].[7]

2.2.4 Sympathy Connects Everything

Thus this underlying sympathy that dynamically connects these seeming individual entities is the unseen Connecting Link[8] that connects them all. This Connecting Link functions on the 6th and 7th subdivisions[9] of Matter and Energy and permeates all forms of matter whether in solid or fluid form.

2.2.5 Sympathy is simply another word for Love

Sympathy is simply another word for that state of being known as Love, Harmony, Coupled or Quantum Entanglement.

2.2.6 Sympathy is Attractive

It is this sympathy or sympathetic affinity of vibration or oscillation rates that binds, unites or coheres like entities to one another whether they are quarks, atoms or people individually or in groups. This property of vibratory attraction was demonstrated by Carl Bjerknes[10] in 1881 and is a fundamental dynamic of SVP.[11] Likewise vibratory attraction is the cause of syntropy[12] (negative entropy[13]) another basic and important dynamic of SVP.

---

[5] notwithstanding near and far, seen and unseen vibratory influences and/or modulations

[6] [Wave Function]

[7] [Entanglement], [Subdivision]

[8] [Connecting Link]

[9] [Subdivision]

[10] [Carl Anton Bjerknes]

[11] [Bjerknes Effect], [Law of Attraction], [Syntropy]

[12] [syntropy]

[13] [negentropy]

2.2.7 Individuated Complex Wave Functions

Having said all that it must be also stated that no two electrons are exactly the same compound chord of frequencies (complex wave function or spectra) at any given moment in Time and location in Space. They do all possess certain 'standard' chord configurations in common else they would be something other than electrons. Each has a slight and ever changing variation dependent upon modulating forces such as local gravity, electric and magnetic influences, pressures, temperatures, *etc.* If these subtle differences were not present then destroying one electron would destroy them all and as we see a continuity happening with the universe we can see this completely entangled sympathy does not occur on the electron subdivision. Some aspects of electrons are fully entangled other aspects not so much. On the other hand should we progress to the next subdivision, the photonic or light, we see a higher degree of sympathy or harmony between the photons than the sympathy that exists between electrons. This is because there are less frequency parameters (simpler chord or signature) to a photon than has a larger group of photons manifesting as an electron. Likewise we can advance to the next higher subdivision of quarks and notice there is again another higher degree or "Q" of sympathy than with the lower grosser photon or electron subdivisions. There is yet another higher level of subdivision which Keely called the Compound Interetheric or Undifferentiated Mind Force and substance. He indicated this level is not differentiated and is wholly sympathetic unto and within itself. A pure (latent) sympathy if you will being of course pure Love or Oneness of vibratory Being. Religions have called this state or condition of total Love or sympathy many names from God, Nirvana, Allah, Jehovah, Heaven, *etc.*

2.2.8 The state of pure Sympathy

This state of pure sympathy acting in a highly sympathetic medium (etheric, quantum subdivisions) means no matter what happens anywhere in this sympathetic medium such is communicated to all of the medium simultaneously and the velocity of communication is near instantaneous.[14] Now, is not this instant awareness or knowing of what is going on regardless of distance (space) an omniscience? Is this not a "Self Awareness" of the underlying universal matrix we call the universe? Does this interconnectedness not account for shared Mind phenomena such as remote viewing, clairvoyance, spooky action at a distance, and telepathy?[15]

---

[14] [Quasiparticles carry entanglement to near infinite speeds]

This background or ground state sympathy was established when quiescent Divine self-awareness occurred and before active thought or thinking began. This state is not unlike the initiatory Void of Genesis. An undifferentiated state of "perfect continuity without extension". The originating state of neutrality of a Neutral Center cannot change but its effects once disturbed are always changing.

---

[15] [Mind Force the hidden Scalar Force]

Chapter 3

Sympathy

3.1 Sympathy is Love

Sympathy is Love or that which connects (entangles) all things to all else. When the frequency of vibrations are one there is unison and sympathetic resonance[1]. In SVP, sympathy does NOT mean "feeling sorry" but is "feeling the same as", "one with", "as one", concordant, harmonic or entangled with identical or near identical states with near identical wave functions.[2] With this insight we can see Sympathy is a vibratory state of matter and energy with its own waveform, spectra and wave function. It is a real state or condition existing throughout nature and is a condition we can know, understand and engineer[3]. For instance between two tuned tuning forks exists this state of sympathy, sympathetic vibration or sympathetic association. The sympathetic vibrations and oscillations originating from each fork and active between them creating this shared state, can be modified to increase or decrease the state of sympathy existing between them. The fact that they are sympathetic and co-vibrate together is called *resonance* which only occurs within this state or condition of sympathetic association via sympathetic vibrations and oscillations.

---

[1] [Sympathetic Resonance]

[2] [Concord], [Harmony], [Affinity], [Attraction], [Awareness], [Knowing], [Universal Love], [Christ], [Science], [Truth], [Coupling]

[3] "Engineering with Love", http://svpbookstore.com/

3.1.1 Infinitely Small Centers

The infinitely small strings of the String Theory are a classical physics concept of oscillating physical strings extrapolated to the non-classical quantum realm and is a *non sequiter*. Such is an attempt to apply known or assumed classical oscillation dynamics associated with an oscillating 3D or physical string, which are also poorly understood and in error, as an answer or postulate to address presumed quantum dynamics.[4] Should we replace the hypothetical infinitely small oscillating strings as held in the String Theory with infinitely small constructive (virtual or actively created by virtue of dynamics and purpose) centers that vibrate, pulsate and oscillate like tuning forks it can be seen how every center is sympathetic (in sympathy) with every other center and are therefore quantum entangled; *i.e.*, (sympathetic to one another). All of these centers form One Continuous Continuum of sympathetic (entangled) centers and is everywhere throughout Space permeating every seeming separate or seemingly individualized object. What connects all these sympathetic centers is a state of etheric level high-grade sympathy creating a condition of sympathetic association - one with all others[5]. This would of course imply a single fundamental tone or frequency at the very heart/core of the universe. Keely called these centralizations of orchestrated motion “atomoles”[6]. Therefore what is experienced by one center is to a degree experienced by all other centers regardless of Time and Space. Being mutually responsive to excitation where all are aware of all others we may term this continuum the Whole Undifferentiated Mind of the universe. Mind or Consciousness being synonyms for Keely’s Compound Interetheric state. Some over the years have referred to this continuum and its sensitive responsive properties as God and have erroneously anthropomorphized it even though it is the root of Awareness and cognition. Being sympathetic these centers can be stimulated or excited by an appropriate exciter[7]. That exciter being clear focused Mind Force, Thought, Volition or idea.[8] Other different state excitations (stimulations) on coarser molecules and atomic substances could be acoustical, electrical, microwave or physical or similar.

---

[4] see ‘Chapter 6 - What is Vibration?’

[5] [Law of Assimilation]

[6] [atomoles]

[7] [Exciter]

[8] [What is Sympathy]

3.1.2 Atomoles

Atomoles (atom = unit + mole = power) = Power Unit, are primordial particles (preferably 'centralizations') filling all space.

"Atomoles are elementary units of matter uniform in size and weight, and exist in solid, liquid, gaseous, and isolated forms." (plural 'atomolini')[9]

"The atomoles are made up of atomolini (singular atomolinus); the subdivision of matter from this point is beyond man's power, as at this point it escapes all control of apparatus, passing through glass and hardened steel as a luminous flame without heat, which is hardly seen before it vanishes, - a perpetual flame coldly luminous."[10]

''*The atomolic substance is what is termed the ether which fills all space and is the transmitting medium for all celestial and terrestrial forces. This is the liquid ether of occult science.*"[11][12]

3.1.3 Records of Motion are Seeds

Walter Russell refers to these excitable (plastic) centers, when considered as One Continuous Continuum, as points of 4++ (voiding and dispersion) and as "seeds" or "ideas". These seed centers correspond to Keely's Neutral Centers conditions of birth to all seeming things in a progressive evolutionary manner.

"Records of Motion are seeds for repetition of motion. In the seed is desire for manifesting imaged forms of idea. All forms are wave forms. All wave forms unfold from seed records of those wave forms."[13]

---

[9] [Keely]

[10] Keely, 1893

[11] Keely, 1893

[12] [Connecting Link]

[13] Russell, [The Secret of Light], page 137

3.2 Neutral Center of the Earth

"In setting the conditions of molecular sympathetic transmission by wire, "writes Keely, "the same law calls for the harmonious adjustment of the thirds, to produce a non-intermittent flow of sympathy. Intermission means failure here. That differential molecular volume is required, in two different mediums of molecular density, to destroy differentiation[14] *of sympathetic flow, seems at first sight to controvert the very law established by the great Creator, which constitutes harmony - a paradoxical position which has heretofore misled physicists who have propounded and set forth most erroneous doctrines, because they have accepted the introductory conditions, discarding their sympathetic surroundings. The volume of the neutral centre of the earth is of no more magnitude than the one of a molecule: the sympathetic conditions of one can be reached in the same time as the other by its coincident chord."*[15]

3.2.1 Force of Sympathy

"There is no force in all creation more real or more powerful that the force of sympathy."[16]

Sympathy may be considered a force because it provides a harmonizing modulation. Where there is discord or enharmonicity Sympathy exerts an influence, of its own accord and nature, converting or transforming that imbalance back to balance or harmony. Discord, entropy or differentiation may be considered as a disturbance of equilibrium which imbalance is restored by Sympathy - harmony or balance being the natural eigenfrequency (wave form and wave function) of the universal matrix - with preponderance of syntropy during growth and a preponderance of entropy during death.

3.2.2 Sympathy is the Essence of Life

"Sympathy, like a human electricity, is the essence of life."[17]

---

[14] [differentiation]

[15] [Keely], [Law of Sympathetic Association]

[16] [David Sinclair], [A New Creed the book]

[17] [David Sinclair], [A New Creed the book]

"There is a celestial mind force, a great sympathetic force which is life itself, of which everything is composed."[18] Keely[19]

"Life is creative, and is the manifestation of that energy, that oneness, which may never be WHOLLY discerned or discovered in materiality, and yet is the basis of all motivative forces and influences in the experiences of an individual."[20]

For life or Life Force[21] to be creative it must be syntropic - that building process of assimilating elements its environment - to itself. As this Life Force cannot be discerned strictly in materiality or the 3D classical world it must be of the unseen non-classical scalar forces such as Mind. Hence Keely's use of the term *celestial mind force*[22] also known as *celestial radiation*[23].

3.2.3 Sympathy is as Syntropy and Electricity

"The powers of electricity have not hitherto been observed, nor have the origin and importance of sympathy. The electric marvels of the present day were in electricity thousands of years ago, but there was no practical observer to utilize them. Sympathy is as everlasting and all-powerful as electricity is. Men, in their honest research for human happiness, will put it to the test, and the new insight it will give to life will be more marvelous than the marvels of electricity."[24]

What David Sinclair[25] is saying here is sympathy as a force or state exists and can be engineered. That science has not recognized this state of being existing and operating as an ubiquitous construct of the universe is lamentable. Keely recognized it to such an extent he has able to utilize its properties in his science and incorporated it as a usable function in their dynamics. It is a simple fact that unless a property is acknowledged to exist it cannot be utilized. We can look at this state of being as a restorative force. Natural systems are in a state of equilibrium. Sympathy, as a pre-existing condition of dynamic systems, will restore that balance or equilibrium should the system be disturbed. This self-acting and self-correcting functional dynamic, unseen by classical and non-classical science and lay alike, is a core functional principle in Sympathetic Vibratory Physics. Taking the metaphor of an undisturbed seesaw quietly sitting in a horizontal position we can see it will tend to re-level itself once disturbed. It will return to its horizontal position. Or a pendulum will come to rest (balance) is moved off center (disturbed out of rest or balance). Likewise we know a given volume of atmosphere is essentially in balance between dry hot expanding air and cold contracting moist air. When a cold wet front approaches a volume hot dry air the localized atmospheric system is then out of balance between the two volumes. The re-balancing occurs when the two volumes come together with the unseen desire[26] to re-balance themselves by one attempting to become the other. This seeming antagonistic action is what Walter Russell called Voiding[27]. This voiding action can be gentle or it can be violent as a tornado depending on Time and polarity or bias difference between the two states or conditions. This desire to void or re-balance (to become *as one*) is born of sympathy and the nature of sympathetic associations where all unlike (seemingly opposite) participant forces or states desire to be alike (in balance or harmony). Hence we see throughout nature seeming opposite poles of electricity voiding against or through each other in a spark or load, opposite poles of a magnetic adhere as one and opposite sexes drawing together and assimilating each other.

---

[18] [Life]
[19] [Celestial Radiation]
[20] Cayce (2012-1)
[21] [Life Force]
[22] [celestial mind force]
[23] [Celestial Radiation]
[24] [David Sinclair], [A New Creed the book]
[25] [David Sinclair]

---

[26] [Desire]

[27] [Voiding]

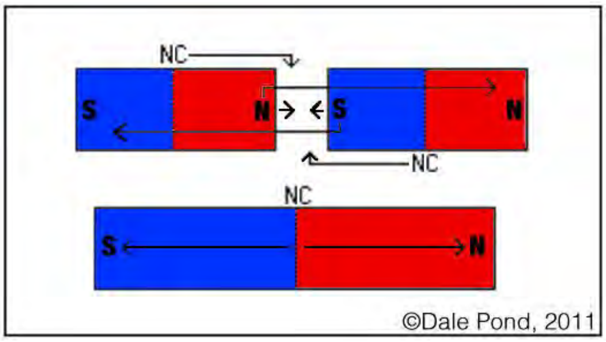

Figure 3.1: Opposite polar forces void/combine as One
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/bar_magnet1.jpg

After voiding (Keely's *neutralization*[28]) or becoming depolar or neutral each pole or participant in the voiding action returns to its original state of unbalance or near so.[29] This return from depolar to polar is of course Time dependent.

3.2.4 The rate of acceleration of Sympathy is enormous

Sympathy exists between all centers. It is as an etheric connecting link[30] stronger and denser than hardened steel. Because transmission or propagation increases with density its rate of extension or propagation is enormous.

"It is a law of falling bodies that they steadily increase their speed, so is it with this great power of sympathy when once it is started. Its rate of acceleration is enormous. A force that is once created can never be destroyed, and there is no force in all creation more real or more powerful than the force of sympathy. It may be stored up and hidden as energy is in fossil vegetable matter. It is a great created power, latent in the hand of the Creator, and at the command of the creature."[31]

---

[28] [neutralization]

[29] In Figure 3.1 'NC' = [Neutral Center]

[30] [connecting link]

[31] [David Sinclair], [A New Creed the book]

3.2.5 Sympathetic Associative Forces

"*What modern cosmologists would have us believe was evolved by blind chance out of the inherent potentialities of primordial chaos, this glorious system of spiritual physics demonstrates as dependent upon the relationship between mind and matter; showing whence the initial impulse comes which sets in motion machinery of unimagined complexity, on given lines, toward a prescribed end. Chance, as has been said, is in no sense a force, but merely the sway of a balance of forces already in action. Sir Isaac Newton taught that there is such consistency in nature that what lies wholly beneath the region of visibility may be safely inferred to be similar to that which is gross enough to be palpable to sense; and Keely, reasoning on this line, has copied nature in his mechanical work. "I call this indefinable, latent element," he writes, "the soul of the sympathetic elements in which it manifests, itself; and which until now has been locked up in their interstitial embrace. It is the leader of all triple streams, associated with the polar negative envelope of our planet and the one most sympathetically concordant to celestial radiation. In our individual organisms, the latent soul-forces, existing in the cerebral domain, are sympathetically subservient to the celestial radiating force whereby they are stimulated into action in controlling the movements of our bodies. Take away this latent element from the brain and the physical organism becomes an inert, dead mass; on the same order as a mechanical device without an energy to operate it.*

"*The polar negative machine is a mechanical brain, with all the adjuncts associated with it to sympathetically receive and distribute the polar negative force. Its sympathetic transmitter (corresponding to our sun in our planetary system, transmitting all energy from the central sun of the universe) is the medium whereby sympathetic concordance is established between it and polar sympathy. The requisites for polarizing and depolarizing keep up the action of the machine as long as it is associated with the transmitter. The force which operates the mechanical is the same as that which operates the physical brain; purely mental, emanating from celestial outreach. There is nothing in the range of philosophy which so satisfies the intellect as the comprehension of this wondrous system of sympathetic association, planned by the Creator of the celestial and terrestrial universe, for the government of all forms of matter.*

"*Nature cannot rebel against herself. The flowers of spring cannot resist the sympathetic force which calls them into bloom, any more than the latent force in intermolecular spaces can rebel and remain in neutral depths when sympathetic vibration calls it forth.*

"*What is the soul but life in latent suspension? The motion exhibited in matter shows that its soul is ever present; and yet there are men of great learning, as taught in the schools, who, after spending their lives in researching all forms of matter, deny that all living things depend on one everlasting Creator and Ruler, in whom they live and move and have their being through all time, as much as when He first breathed into them the breath of celestial radiation; and to whom they are as closely allied, still, by the workings of the great cosmical law of sympathetic association, as when the evolutionary work of creation commenced.*

"The ancients were far better schooled in spiritual philosophy than are we of the present age. Their mythological records, in their symbolical meaning, prove this fact. They recognized this latent element as the very breath of the Almighty; the sympathetic outflow of the trinity of force, the triple spiritual essence of God Himself. Their conceptions of Deity were greater and truer than our own. From them we learn that when God said 'Let there be light,' He liberated the latent celestial element that illuminates the world: that when He breathed into man the breath of life, He impregnated him with that latent soul-element that made him a living and moving being." [32]

3.2.6 Sympathy is Entanglement

When object A is in sympathy with object B, A is aware (responsive) of what happens to B and B is aware of what happens to A. As there is an innate or natural vibratory sympathy (sympathetic association of vibration and oscillation) between all things or objects all things and objects are aware (responsive) of all that is happening, everywhere and all the time regardless of Time and Space. This property or condition os sympathetic association has been called by many names throughout humanity's long evolution: Entanglement, Omniscience, Coupling, God (and all its variations), non-locality, Mind, Love, Remote Viewing, Clairvoyance, Celestial Sympathetic Radiation, Brahma, Cosmic Consciousness, *etc.*[33]

---

[32] Keely from [The Veil Withdrawn]

3.3 Sympathy defined

It is important to understand what sympathy is as it plays a tremendous and an all important role in how the universe and all it contains works. Sympathy is a state or condition that occurs between all centers and is an inherent property of these centers.

1 - ‘‘*Where there is no sympathy there can be no love, for love is sympathy.*”[34]

2 - noun: a relation of affinity or harmony between people or things; whatever affects one correspondingly affects the others.[35]

3 - noun: In natural history, a propension of inanimate things to unite, or to act on each other. Thus we say, there is a sympathy between the lodestone and iron.[36]

4 - Sympathy is a Force, as material as Sound, Light and Heat - *“For the sake of being popularly understood, sympathy as a force is here, in a general sense, sometimes spoken of in a kind of figurative way as a message on the telegraph wire, etc. In a more particular way it will now be scientifically explained as a force, as material as sound, light and heat. We speak of the sun’s rays as if each ray were some simple element, and yet easily carry in our minds the fact that each ray is composed of light and heat; so this philosophy speaks of this ether as a mere lineal connecting-rod between the creature and Creator. If the analogy between the sun’s rays and the ether be closely followed, it will be seen that wisdom, truth, and sympathy are the same thing, and travel on the ether lines*[37] *exactly as light and heat do on the sun’s rays.*”[38]

5 - Pertaining to or produced by sympathy.

---

[33] [Dale Pond], 11/21/10

[34] [Vera Vita the Philosophy of Sympathy]

[35] [Sympathetic Vibration]

[36] Webster’s 1828 Dictionary

[37] [connecting link]

[38] [David Sinclair], [Vera Vita the Philosophy of Sympathy]

6 - *"The mutual relation between parts more or less distant, whereby a change in the one has an effect upon the other."*[39]

7 - A real or supposed affinity between certain things, by virtue of which they are similarly or correspondingly affected by the same influence, affect or influence one another (especially in some hidden, unseen way), or attract or tend towards each other.

8 - A relation between two bodily organs or parts (or between two persons or things) such that disorder, or any condition, of the one induces a corresponding condition in the other.

9 - Agreement, accord, harmony, consonance, concord, agreement in qualities, likeness, conformity, correspondence.

10 - Conformity of feelings, inclinations, or temperament, which makes persons agreeable to each other; community of feelings; harmony of disposition.

11 - The quality or state of being affected by the condition of another with a feeling similar or corresponding to that of the other; the fact or capacity of entering into or sharing the feelings of another or others; fellow-feeling. Also, a feeling or frame of mind evoked by and responsive to some external influence.[40]

12 - Sympathy, n.; pl. Sympathies. (F. sympathie, L. sympathia, Gr. ; with + suffering, passion, fr. , to suffer. See Syn-, and Pathos.)

13 - Feeling corresponding to that which another feels; the quality of being affected by the affection of another, with feelings correspondent in kind, if not in degree; fellow-feeling.

14 - An agreement of affections or inclinations, or a conformity of natural temperament, which causes persons to be pleased, or in accord, with one another; as, there is perfect sympathy between them.

---

[39] Blakiston's New Gould Medical Dictionary. The Blakiston Company, 1949. 1st edition

[40] (source unknown, underline added)

15 - Kindness of feeling toward one who suffers; pity; commiseration; compassion.

16 - (Physiol.) (a) The reciprocal influence exercised by the various organs or parts of the body on one another, as manifested in the transmission of a disease by unknown means from one organ to another quite remote, or in the influence exerted by a diseased condition of one part on another part or organ, as in the vomiting produced by a tumor of the brain. (b) That relation which exists between different persons by which one of them produces in the others a state or condition like that of himself. This is shown in the tendency to yawn which a person often feels on seeing another yawn, or the strong inclination to become hysteric experienced by many women on seeing another person suffering with hysteria.

17 - A tendency of inanimate things to unite, or to act on each other; as, the sympathy between the loadstone and iron.

18 - Similarity of function, use, office, or the like.

3.4 Recommended reading

[David Sinclair], [A New Creed the book]
[David Sinclair], [Vera Vita the Philosophy of Sympathy]

Chapter 4

Neutral Center

4.1 Neutral, Latent Neutral and Neutral Center

Scattered throughout the following pages are the terms "neutral", "latent neutral" and "neutral center". These are central concepts in SVP (Keely's physics) and any study of polarity and its dynamics.

4.2 Neutral means

Neutral means: Where two phases from two vibrations are opposite in polarity and cancel each other out or "sum to zero" is called *phase conjugation* and is one form of neutral. Another form is the balance point between polarities much like a fulcrum under a seesaw. This balance point within a vibration or oscillation is neutral, depolar or potential. The power (potential) in both above cases of the phases or polarities has been equated and becomes latent thus establishing two poles and one neutral - a trinity of states of forces. There are other forms or associative dynamics constituting a neutral state or neutral center and will be discussed later.

"When a force becomes neutral it is inactive, and is no longer a force; consequently the name given is misleading, Keely calls this third element the "latent neutral," which is a better name for it."[1]

---

[1] Bloomfield-Moore [CJBM], [What Electricity Is - Bloomfield Moore]

Latent means depolar. The other two forces or states are polarized forces dynamic relative to each other. These are the Harmonic (syntropy) and Enharmonic (entropy). There exists a neutral state from which these two polar forces originate and to which they return. This neutral (depolar) state is "latent" or quiescent until stimulated (polarized).

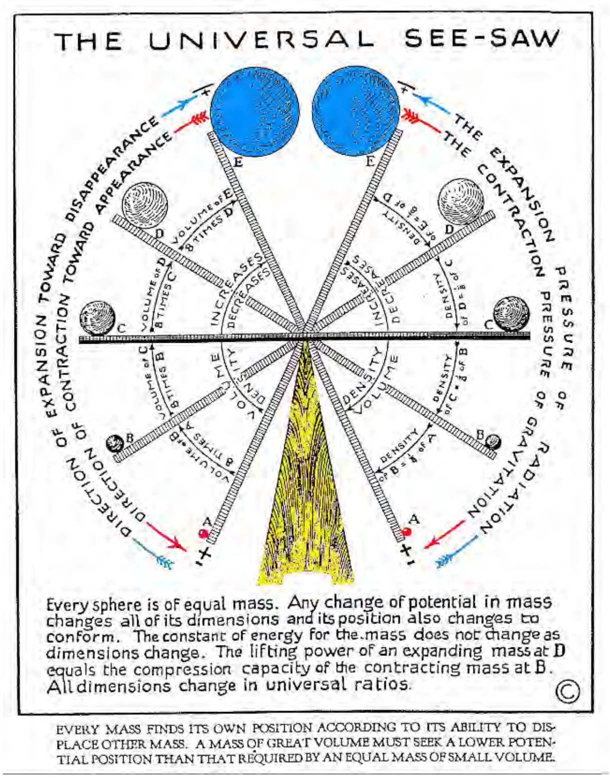
THE UNIVERSAL SEE-SAW
DIRECTION OF EXPANSION TOWARD DISAPPEARANCE
DIRECTION OF CONTRACTION TOWARD APPEARANCE
THE EXPANSION PRESSURE OF RADIATION
THE CONTRACTION PRESSURE OF GRAVITATION
+
-
A
B
C
D
E
VOLUME of E 8 TIMES D
VOLUME of D 8 TIMES C
VOLUME of C 8 TIMES B
VOLUME of B 8 TIMES A
DENSITY of E = ⅛ of D
DENSITY of D = ⅛ of C
DENSITY of C = ⅛ of B
DENSITY of B = ⅛ of A
VOLUME INCREASES
DENSITY DECREASES
DENSITY INCREASES
VOLUME DECREASES
©

Every sphere is of equal mass. Any change of potential in mass changes all of its dimensions and its position also changes to conform. The constant of energy for the mass does not change as dimensions change. The lifting power of an expanding mass at D equals the compression capacity of the contracting mass at B. All dimensions change in universal ratios.

EVERY MASS FINDS ITS OWN POSITION ACCORDING TO ITS ABILITY TO DISPLACE OTHER MASS. A MASS OF GREAT VOLUME MUST SEEK A LOWER POTENTIAL POSITION THAN THAT REQUIRED BY AN EQUAL MASS OF SMALL VOLUME.

Figure 4.1: Universal Seesaw balancing seeming opposites
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/universal-seesaw.jpg
(courtesy University of Science and Philosophy)

A latent center is a center that has not been stimulated or excited. It is quiescent (depolar). Latent Forces are forces not stimulated (polarized) into motion or activity but held in stable (harmonic) balance. A scalar[2] force is a non-moving non-dynamic latent force or potential - not kinetic.

---

[2] [scalar]

In the above Russell image he uses the phrase "potential position" which is the same as Keely's "level of tenuity". Thus indicating it is the potential or tenuity of relative position that constitutes mass being a relative quantity. The fulcrum is the neutral balance point between two potentials or polarity, bias or bipolar polarities.

Neutral State or condition and Neutral Center are two different things or dynamics. A Neutral State is a condition independent of location. A Neutral Center is a location or foci, relatively speaking. A Neutral State has among other attributes Mutual Affinity between participating streams or conditions, sensitivity to appropriate stimuli, equated polarizations, point or state of equation. There are diverse orders, kinds or types of Neutral Centers. The two sides of the seesaw represent polarization. The fulcrum represents the neutral sustaining point, principle or condition where the polarizations meet by mutual affinity in mutual antagonism. *"All forces are Mind Forces"* meaning all forces originate from Undifferentiated Mind[3] or the Compound Interetheric[4] subdivision. As sections or segments of Undifferentiated Mind are polarized (differentiated) they become forces or energies. The Ultimate Differentiator is thought or idea. These polar forces become neutral or latent by removal of polarization (discords). Keely called this process "harmonization".

4.3 Coordination or coincidence of forces

Consequently when and where various forces meet and cancel (void) or agree there is established a Neutral Center. Sometimes this point may be called a center of neutrality, center of attraction, center of sympathetic attraction, center of sympathetic coincidence, center of association and concentration, center of focalization, nucleal center, Prime Neutral Center and Supreme Neutral Center, depending on its general dynamical state. Some of the types of Neutral Centers are:

- mass - every mass of whatever substance or size has a neutral center.
- molecular - neutral centers in molecules.

---

[3] [Undifferentiated Mind]

[4] [Compound Interetheric]

- atomic - neutral centers in atoms.
- etheric neutral centers - neutral centers in quantum/etheric centralizations.
- Prime Neutral Centers - neutral centers that govern and enliven secondary centers not unlike the Sun (primary) and its planets (secondary).
- Supreme Neutral Center - The one Dominant Center that controls all other neutral centers.
- center of centers - an object may have more than one neutral center, this is the coordinated center between them all.
- focalized centers - centers can have more than one center. This idea is what we see in a circle (one foci) and an ellipse (two foci). A focalized center is where all these centers are coincident or focalized together as one center.
- sympathetic concordant focalized neutral center - a focalized center that is concordant (in harmony of parts) that is sympathetic to something else, usually an outside stimulant or exciter or other center(s).
- focalized negative attractive centers - A focalized neutral center that is in the syntropic phase of assimilation, attraction or concentration. It's opposite would be a dispersing positive propulsive center (radiating or radiant). See [Law of Assimilation].
- centers of sympathetic coincidence - A dynamic center where forces that are sympathetic to each other coincide.
- centers of neutrality - a neutral center that is in a quiescent state of neutral latency.
- centers of focalization - centers where forces are mutually attracted or focused to interact.
- centers of distribution - centers where forces are mutually repelled and dispersed.

Centers are intermittently attractive and repulsive or dispersive. When a Neutral Center is disturbed it becomes intermittently attractive and repulsive. When attractive it becomes a center of focalization. When repulsive it becomes a center of propulsion or dispersion. In between one state and its opposite is a moment of pause not unlike the furthest point in the swing of a pendulum or seesaw. This null point or fulcrum can endure both polarities without itself being affected.[5]

The Neutral Centers give birth to quantum centralizations Keely called atomoles. The Neutral Centers are then the cause that give rise to their effects - these atomoles - and are synchronistic meaning "can't have one without the other' (the chicken and egg riddle). On the other hand Keely and Russell both said there is an infinite number of Neutral Centers implying there is an infinitely small space between these centers which makes the universe quite dense (986,000 denser than steel[6]) and births the logical idea matter is more of a tenuous 'bubble' in this dense etheric substratum than the way ego sees matter as more dense than seeming vacuity.

Matter evolves from vacuity of space

"*Pure sympathetic concordants are as antagonistic to negative discordants as the negative is to the positive; but the vast volume the sympathetic holds over the non-sympathetic, in ethereal space, makes is at once the ruling medium and re-adjuster of all opposing conditions if properly brought to bear upon them.*" Again Russell and Keely concur on the origin of matter. Sympathetic streams aggregate at a neutral center about which forms centralized motions; *i.e.*, form or appearance of matter. Non-sympathetic conditions are discordant and dispersive resulting in the appearance of seeming empty space. Solidity devolves towards vacuity through the action of non-sympathetic (discordant) conditions; *i.e.*, entropy. Outer space is the domain of the non-sympathetic or dispersed matter.

In one of many ways Russell describes the dual nature of space (cube) and matter (sphere) he says this about that dynamic state: "*The sphere is the*

---

[5] See [Philosophers Stone]

[6] [Law of Matter and Force]

result of the concentrative pulsation of Gods thinking which pulls inward from within (syntropy) to focus the light of Gods thinking to a point of stillness, around which a thought body (form or matter) can be formed. This causes that universal effect, which we know as gravitation, and the still point of concentration is a center of gravity. GRAVITATION MULTIPLIES POWER. [See ([Neutral Center])

"*The cube is the result of the decentrative pulsation of Gods thinking which thrusts outward (entropy) from within to the stillness of the cube boundaries to produce the opposite effect which we know as radiation. RADIATION DIVIDES POWER.*

"*Gravitation and radiation are the opposite ends of the piston which motivates the heartbeat of this eternally living dual body of God, which is our dual body. Its continual giving and regiving*[7] *manifests the love principle*[8] *in both the action and reaction not just in one of them.*"[9]

4.4 Neutral Center

Neutral Center = Dominant = Fulcrum = Pure (Undifferentiated) Consciousness = Love = 0 (Russell) = First Cause. There are neutral centers on all seven subdivisions of matter and energy. These Neutral Centers possess differing states; *i.e.*, properties, functions and geometries.[10]

We do not yet know all the properties of the neutral center but Keely does list a few of them. Keely considered his discovery of the Neutral Center to be one of his greatest discoveries as also its control through machinery. Our rediscovery of it will be one of humanity's greatest achievements. Keely's description of the neutral center follows:

"Every molecule, every mass, every moving body in space, every solar system, every stellar system, EVERY ROTATORY SYSTEM, is built about a

---

[7] [Giving-Regiving]

[8] [Law of Sympathetic Association]

[9] Russell, Home Study Course, GOD CREATED ONLY ONE FORM; see [Cube Sphere]

[10] [Barycenter], [Center], [Center of Gravity], [Center of Mass], [Sympathetic Coincidence], [MINERAL DISINTEGRATION - Snell]

NEUTRAL CENTER.”[11]

“The neutral center represents only focalization and distribution of the streams of energy.” [12]

Thus there are discrete mass, molecular, atomic and etheric neutral centers as well as Prime Neutral Centers, “center of centers”, “focalized centers”, “sympathetic concordant focalized neutral center”, “focalized negative attractive centers”, “centers of sympathetic coincidence”, “centers of neutrality” and “centers of focalization”, thus defining its state depending on dynamical function, location, association, etc.

“The neutral center is that protean, uncreated, indestructible, forever-existing FIRST CAUSE[13]*. Without hands, without tools, without thought, without emotion, without love, without form, without substance, it, of itself, created all these. All that we see or can see in the objective Universe exists because of and by means of the properties and powers of the NEUTRAL CENTER.”*[14][15]

“Where the vibrations under this mode meet, and are maintained in a state of mutual affinity or equilibrium, there is established what is called a “neutral center”, or, as otherwise expressed, ‘a center of sympathetic coincidence’.”[16]

4.4.1 Disturbance of Equilibrium

This is why a meditative state (neutral or undifferentiated) of mind is important (sympathetic) to “knowing” these interlaced concepts. Because they are All One. To be it is to know it. The thinking ego or intellect is differentiated conscious and is highly discordant which is to say it destroys harmony as is present in the meditative or quiet mind. This does not make the intellect thinking an enemy - it is simply discordant, a disturber of the harmony. This discordance is the Infinite Exciter[17] ‘*‘that moulds and makes*”[18] something from seeming nothing thus making humanity as a Creator. We use our crystal clear thinking and focus to create the lives we live. As neutral states are pure unmanifested potential (latent) in perfect harmony it is from where and whence all manifestation of form originates - *“from harmony is generated discord*”[19]. These manifestations become reality when the neutral state is disturbed. This disturbance of equilibrium[20] [21]is a key function in SVP.

---

[11] [Keely]

[12] [ROTATION - Snell]

[13] [First Cause]

[14] [Keely]

[15] [First Cause]

[16] [KEELYS PHYSICAL PHILOSOPHY - Snell]

Logically, if all three forces are modulated the same degree at the same time there would be imbalance in the equation. Disturbance of equilibrium means the three forces, in an established triune or three-way equilibrium, are thrown out of balance or equation by changing one or two of the three so the equation of harmony becomes one of discord; *i.e.*, thrown into motion which then seeks to re-establish the originating harmony or balance. For instance, this principle is active in spiritual seekers as they are all seeking to return to the original Peace and Harmony with the One (wake up or self realization. Such a disturbance in a molecule being intermittent will cause the molecule to deform in its motions into activity then recede back to its original quiescent state.

The Neutral Center is a point or center of sympathetic coincidence. It has many properties. The creation and perpetuation of the Neutral Center is predicated on the inflow of what some call God/Divine Force (Divine Will or sympathetic celestial streams) or as Keely called it *Celestial Sympathetic Radiation*[22] originating from the Supreme Neutral Center, establishing its progressive evolution, as it were, and through the Principle of Regeneration[23] is radiated outward as refracted or differentiated materiality via the Law of Harmonic Pitch[24], Law of Harmonic Vibrations[25], and Law of Cycles[26]. This process is in other words a ubiquitous (everywhere in all Time and Space) or non-local “Micro Bang” (syntropic implosion/cavitation) process that then radiates and eventually manifests as what we perceive as matter, *etc.* In a word, the Neutral Center is a Compound Interetheric sympathetic force or state of Undifferentiated Mind (scalar). Therefore it is this inflow of the Divine Will (sympathetic celestial streams, Mind Force) as CAUSE of materiality; I feel this needs to be clearly stated. See Figure 12.8 for an animated graphic showing this spontaneous generation or production of discords that breaks down into dispersion the formative harmony.

---

[17] [Infinite Exciter]
[18] [As a Man Thinketh]
[19] [8.23 - Law of Cycles]
[20] [disturbance of equilibrium]
[21] see [7.26 Disturbance of Equilibrium]
[22] [Celestial Sympathetic Radiation]
[23] [Principle of Regeneration]
[24] [Law of Harmonic Pitch]
[25] [Law of Harmonic Vibrations]
[26] [Law of Cycles]

4.4.2 Supreme Neutral Center

The Compound Interetheric (Undifferentiated Mind) is outside of Time and Space and is unaffected by either. So one could rightly say the Supreme Neutral Center has always existed and always will exist even though it only exists in the NOW moment like everything else. The physical universe is as a Thought in this Undifferentiated Mind. Thoughts are polar. Expression is polar. Manifestation is polar. All materiality is polar. Non-materiality is depolar (neutral) in relation to materiality. Hence all polar conditions emanate from depolar or neutral latency until stimulated into being by desire, idea or thought.

The egoic intellect creates the ideas of Time and Space. As time is an illusion, a creation of our polar egos, it has not relevance to actuality. There never was a “time” prior to the creation of the “universe”. For lack of a better expression the universe has always existed and had no beginning because if time is an illusion then it might be said there never was a past and there is no future. There is only Now. Our fractured egos cannot Know the Undifferentiated Mind Knowing (God’s knowing). The countless attempts to anthropomorphize ‘creation’ and ‘God’, *etc.* are doomed as feeble attempts at rearranging dream elements (fabrications of the ego). Only our whole undifferentiated minds (Cosmic Consciousness) can know these infinite things, states or conditions which cannot be adequately expressed through finite intellectualism. The Mind of Deity[27] is cause while the universe is the effect of that Cause. A cause and its effects cannot be separated. The connecting link is the etheric domain. As Cause is Undifferentiated Mind then all its effects are polarized (split) mind thinking.

---

[27] [Mind of God]

Keely indicated there is a single Supreme Prime Neutral Center governing all other neutral centers whether mass, molecular, atomic or etheric regardless of location. Prime Neutral Centers are as the Sun (prime) to its planets (secondaries). As neutral centers are effects as well as cause this Supreme Neutral Center is Dominant and an integral of all neutral centers - they are all one and the same in substance and dynamic with the exception One is sum and source of all others. Russell referred to it as the Universal One[28] from which all else originated. The intellect wants to separate every thing into parts. The universe is one whole acting through what appears to be individuated parts. All the parts are governed by one Supreme Neutral Center, each manifesting as though individual. There is one supreme state of neutrality that permeates every thing. This omnipresent state or condition of Neutrality is perfect concordance or Sympathy (Love) and everywhere present in all things and actions - the Universal One.

Outflow and Inflow

Each Neutral Center has an intermittent outflow and an inflow of sympathetic streams originating from the Supreme Neutral Center redirected from the Prime Neutral Center. These occur according to Russell's Rhythmic Balanced Interchange[29] and Macvicar's Law of Assimilation[30].

There is an outflow: ‘*‘..communicates direct by means of its outflow of sympathy, with every planetary mass in the Universe.*”
There is an inflow: ‘*‘Through its inflow of sympathy, through the solar intermediate, the sun, it receives the life flow from the Supreme Neutral Center..*”

---

[28] [The Universal One]
[29] [Rhythmic Balanced Interchange]
[30] [Law of Assimilation]

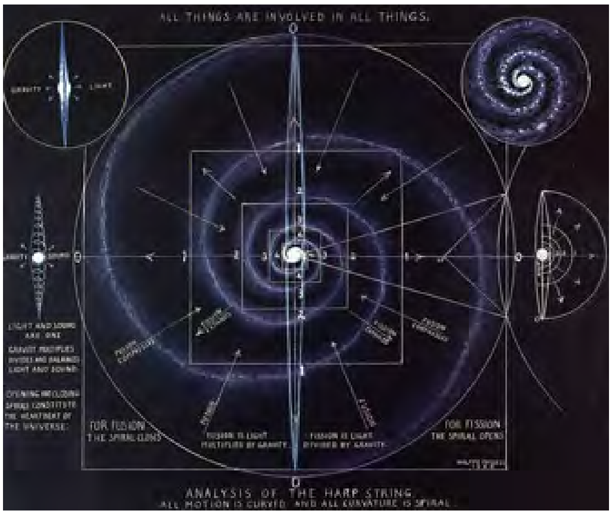
ALL THINGS ARE INVOLVED IN ALL THINGS.
GRAVITY
LIGHT
SOUND
LIGHT AND SOUND ARE ONE
OPENING AND CLOSING SPIRALS CONSTITUTE THE HEARTBEAT OF THE UNIVERSE.
FOR FUSION THE SPIRAL CLOSES
FUSION IS LIGHT MULTIPLIED BY GRAVITY.
FISSION IS LIGHT DIVIDED BY GRAVITY.
FOR FISSION THE SPIRAL OPENS
ANALYSIS OF THE HARP STRING
ALL MOTION IS CURVED AND ALL CURVATURE IS SPIRAL

Figure 4.2: From the One comes the many
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/russell-neutral-center2-300.jpg
(courtesy University of Science and Philosophy)

4.4.3 Keely further describes the Neutral Centers

"All the Dominant conditions of nature represent the focal centers towards which like surrounding ones become sympathetically subservient."[31]

The dominant condition of nature is harmony or Sympathy (Love). This harmonic or attractive state has a field of influence (sympathetic outreach) which casts its influence or its controlling effects as a modulation of everything in its reach (its Sympathetic Outreach[32]). Other surrounding Neutral Centers which can be few or countless depending on the chord of the Dominant Neutral Center fall under this influence. Sympathy extents as a state or condition of forces. This extension has harmonizing effects and will modulate any other chord to, in some degree, resemble this originating chord of harmony. This is what and how people "feel" energy around the Dynaspheres. The calming or harmonizing power is extended by sympathetic vibrations and/or oscillations (quantum entanglement) on the sixth and seventh subdivisions. This power and dynamic has been anthropomorphized throughout history as the"Love of God"[33]. This dynamic is described in Keely's

---

[31] [Keely and His Discoveries], pg 179

[32] [Sympathetic Outreach]

4.4.4 Law of Transmissive Vibraic Energy

"All oscillating and vibrating coherent aggregates create, in the media in which they are immersed, outwardly propagated concentric waves of alternate condensation and rarefaction, having a period-frequency identical with the pitch of the aggregate."
Scholium: *"All forms of transmissive energy can be focussed, reflected, refracted, diffracted, transformed, and diminished in intensity inversely as the square of the distance from the originating source."*

Which law and its dynamics sets the stage for:

"Neutral centers are the focalized seat of sympathetic concordance for controlling any differentiation that may exist outside, or in the mass that surrounds them."[34]

Every individual "thing" has its own chord of mass. Apparently no two chords of mass are identical. A chord of mass will have numerous frequencies within itself resulting from its own chordal evolutions (overtone series) and as modulated from countless outside FM and AM modulations (summation and difference tones). It is these modulations that cause differentiations[35]. They are the differentiations. One center will control all the neighboring neutral centers being as it is or becomes the Dominant Center. Centers are not entirely individuated in their existence because they are ultimately controlled or governed by sympathetic streams, concurrently arriving to each, from the Supreme Neutral Center which created them and in which they have their being or existence perpetuated.

"Neutral centers are the center of Sympathetic Coincidence."[36]

Which dynamic and conditions establishes Sympathy as being the controlling state or condition over everything in its reach or field of influence.

---

[33] Law of Sympathetic Association]

[34] [Keely and His Discoveries], pg 255

[35] [Differentiation]

[36] [Keely and His Discoveries], pg 220

‘‘*Pure sympathetic concordants are as antagonistic to negative discordants as the negative is to the positive; but the vast volume the sympathetic holds over the non-sympathetic, in ethereal space, makes it at once the ruling medium and readjuster of all opposing conditions if properly brought to bear upon them.*”[37]

‘‘*The effects of the law of sympathetic association, which Mr. Keely demonstrated as the governing medium of the universe, find illustrations in inanimate nature.*”[38]

“A center of introductory action is necessary in all operations of Nature. All structures require a foundation. This neutral center is the foundation.

“Every molecule, every mass, every moving body in space, every solar system, every stellar system, EVERY ROTATORY SYSTEM, is built about a NEUTRAL CENTER. It is the indestructible unit around which all that we recognize as matter is built. Immovable itself, it moves all things. Indestructible itself throughout infinity of time, it creates all things. It produced and preserves the incalculable energy of motion of the entire Universe. It bears the unthinkable burden of the mass of the Universe. It is the most wonderful thing Man has discovered in the Universe since he discovered fire.

“If we should take a planet of say 20,000 miles diameter and should displace a portion of the interior so as to have a crust of say 5,000 miles thickness, and at the center of the planet, place a billiard ball, that small mass, immeasurably smaller than the bulk of the earth, would bear the entire burden of the mass of the crust 5,000 miles thick and would keep it equidistant from itself. No power, however great, could possibly displace this central mass so as to bring it into contact with the crust. Furthermore, to move this central mass in any direction, would require a force sufficient to move the entire mass of the planet, and in propagating or continuing any such motion the neutral center, this billiard ball, will at all time periods remain still in the exact center, bearing the same equidistant relation to its hollow shell. The mind staggers in contemplating the burden borne by this neutral center, where weight ceases.

---

[37] [Keely, Chapter 7 of Keely and His Discoveries]

[38] [Bloomfield-Moore, Keely and His Discoveries]

"No less wonderful are other properties of the neutral center.

"It is the cause of the physical Universe. Its attraction condensed that which we recognize as substance. Matter was evolved from the affinity of this neutral center for Sympathetic Streams and since it is immovable, it caused, through negative attraction the formation of nodes in these streams, where the vibrations thereafter continued to meet in a center of Sympathetic Coincidence causing the permanence of form and matter.

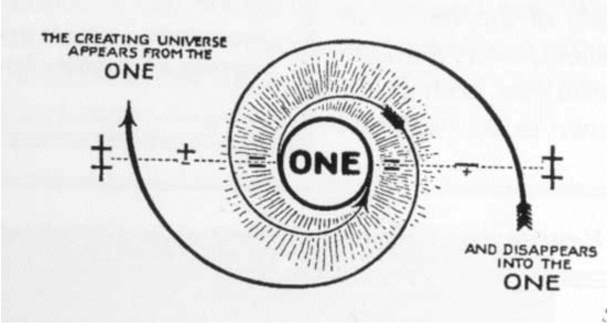
THE CREATING UNIVERSE APPEARS FROM THE ONE
ONE
AND DISAPPEARS INTO THE ONE

Figure 4.3: Coincidental forces meet at center
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/syntropy-entropy-meet.gif
(courtesy University of Science and Philosophy)

"Every nebula, an embryonic world, is acted upon, created and preserved by this neutral center, and at the termination of its cycle, it is ultimately also destroyed by it, causing its absorption into the Unknown from whence it came."

Every molecule and atom disintegrates into its constituent components and disperses into the vacuity of space - seeking their density coincident or 'level of tenuity' (of density). If they all be depolarized they lose their form and dynamics. They become as 'nothing'; *i.e.*, no 'thing'. Polarity gives form and motion. Depolarity (neutrality) has no form or motion. All form evolves from the vacuity of space (dispersed matter). Neutral Centers do not of themselves exist as a 'thing'. They are dimensionless points of the confluence of forces. How can neutrality have dimension if it has no form? But it does have function not unlike Mind or Consciousness itself.

This latent sympathetic plenum (sympathy) throughout the universe gives rise to all form and motion when disturbed, stimulated or excited. A thought or desire excites this neutrality into a polar condition or state. Polarity causes motion and motion appears as form. Wherever forces synchronize or equate there is established an activated neutral center about which 'matter' may aggregate creating form.

Figure 4.4: Neutral Center - creator of nebula and galaxies
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/ringnebula.jpg

"The actual neutral center of the earth is, in fact, even infinitely smaller than the billiard ball referred to above. It consists of a compound interetheric point in space, so small that were we to magnify a pin head to the size of the sun, and from that substance take a particle of matter the same size, again magnifying it to the size of the sun, the neutral center would still be invisible, even though the structure of this last substance was examined through the highest powered microscope ever created, or to be created. For the neutral center is INDIVISIBLE[39]*. Its attributes do not belong to matter, and pertain in no way to matter, which is but its exterior manifestation.*[40]

---

[39] How can neutrality (Undifferentiated Mind) have dimension? One cannot divide zero.

[40] [FORM OF THE ATOM]

"Every aggregate mass consists of molecules, each of which has its neutral center where the three modes of vibration, dominant, harmonic, and enharmonic, meet in a center of Sympathetic Coincidence and are equated without cancellation of their energy.[41]

"Neutral centers are the focalized seat of sympathetic concordance.

"The proof of this assumption is that all matter responded to Keely's disintegration process[42] *and must therefore consist of these fundamental modes of vibration.*

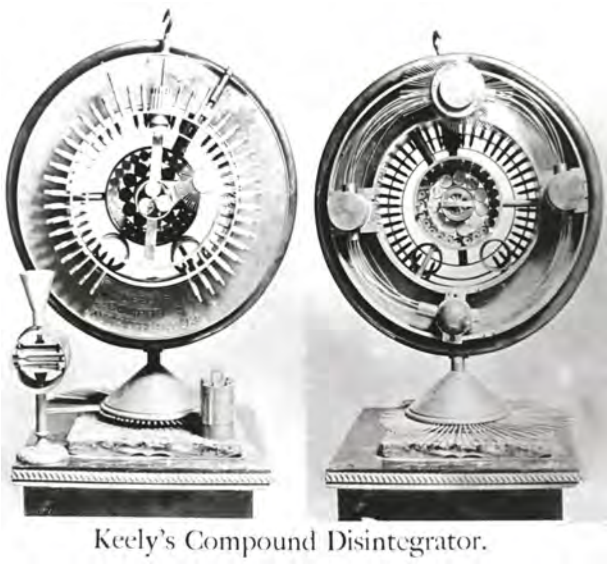

Figure 4.5: Keely's Compound Disintegrator
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/compound-disintegrator.jpg

The above Compound Disintegrator,[43][44] later called Liberator[45], is presumed to be one of the devices Keely used to progressively dissociate or disintegrate water into etheric vapor (plasma). In modern terms this device would be not unlike a multi-frequency "analog signal generator" in function. See section 7.21 Latent Force. The term 'Liberator' describes its function of liberating the Latent Force held within the confines of the rotating etheric envelopes of the atom. See section 12.12 Etheric Envelopes. In the photo below we see the Liberator connected to the pressure lever for measuring pressure of the liberated etheric vapor[46].

---

[41] [Scalar]

[42] disintegrator picture gallery: http://www.svpvril.com/DisPix/Disin1.html

[43] [Compound Disintegrator], [Disintegrator]

[44] more photos http://www.svpvril.com/DisPix/Disin1.html

[45] [Liberator]

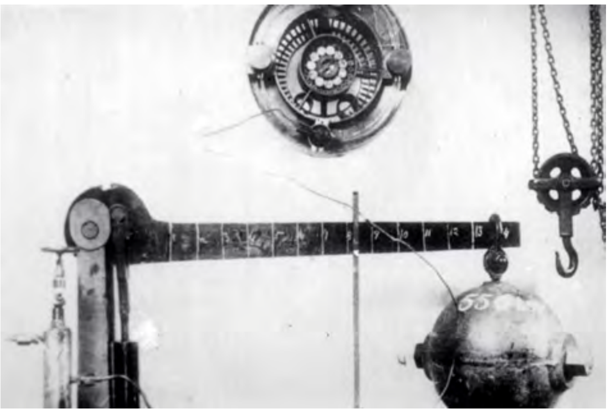

Figure 4.6: Liberating and measuring Etheric Vapor pressures
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/liberating-pressure.jpg

---

[46] [etheric vapor], [Ether Generator Producing High Pressures]

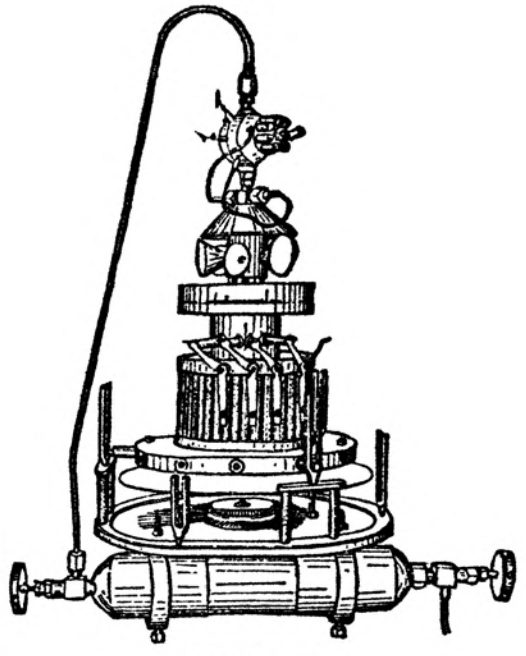

THE LIBERATOR.

Figure 4.7: An earlier Liberator design
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/liberator1.jpg

4.5 Divine Will - The Wheelwork Source of all Force

"The fixed neutral center of the earth is the concentration or totalized power of all the several molecular neutral centers in the earth's mass. This neutral center, which is absolutely WITHOUT WEIGHT, an interetheric point in space, communicates direct by means of its outflow of sympathy, with every planetary mass in the Universe. Through its inflow of sympathy, through the solar intermediate, the sun, it receives the life flow from the Supreme Neutral Center that enables it to perpetuate its existence. Thus through the outflow from this Supreme Neutral Center that pivoting point of the Universe controls the existence and motion of not only every stellar, solar and planetary mass in space, but also the rotatory vibration, in every individual molecule,

intermolecule, etc. through all the subdivisions of matter, thereby sustaining their existence and motion with the life flow.

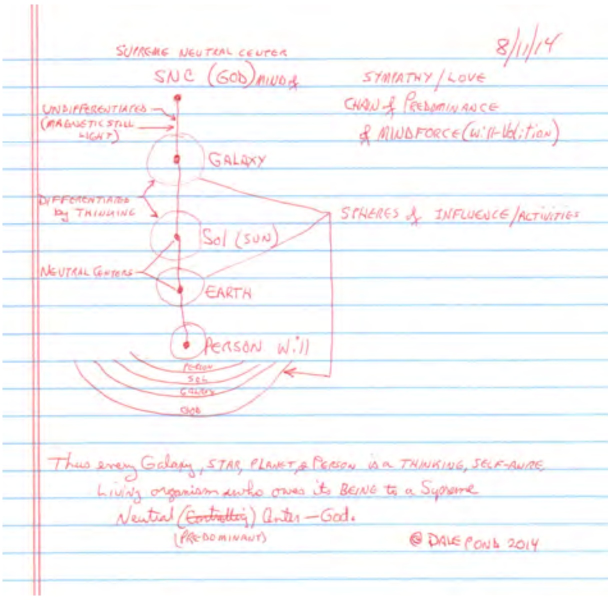
8/11/14
SUPREME NEUTRAL CENTER
SNC (GOD) MIND &
SYMPATHY / LOVE
CHAIN of PREDOMINANCE
& MINDFORCE (Will-Volition)
UNDIFFERENTIATED
(MAGNETIC STILL LIGHT)
GALAXY
DIFFERENTIATED by THINKING
SPHERES of INFLUENCE/ACTIVITIES
Sol (SUN)
NEUTRAL CENTERS
EARTH
PERSON Will
PERSON
SOL
GALAXY
GOD

Thus every Galaxy, STAR, PLANET & PERSON is a THINKING, SELF-AWARE Living organism who owes its BEING to a Supreme Neutral (~~Controlling~~) Center—God.
(PREDOMINANT)

© DALE POND 2014

Figure 4.8: Sympathy Links all Centers to all other Centers
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/universal-oneness.jpg

"All foundations must be sufficient to bear their burden. Conceive then the Universe centered upon and resting the burden of is incalculable mass and kinetic energy on a vacuous interetheric point in space, so minute that it is actually INDIVISIBLE. This conception can only be fully comprehended by an infinite mind. Independent of time, because indestructible in its unity, independent in space, because through its properties space itself exists and without it would not exist, independent of matter because its properties in an external direction created all that we know as matter and gave it seeming permanence, the neutral center is that protean, uncreated, indestructible, forever-existing FIRST CAUSE. Without hands, without tools, without thought, without emotion, without love, without form, without substance, it, of itself, created all these. All that we see or can see in the objective Universe exists because of and by means of the properties and powers of the NEUTRAL CENTER.[47]

---

[47] [God]

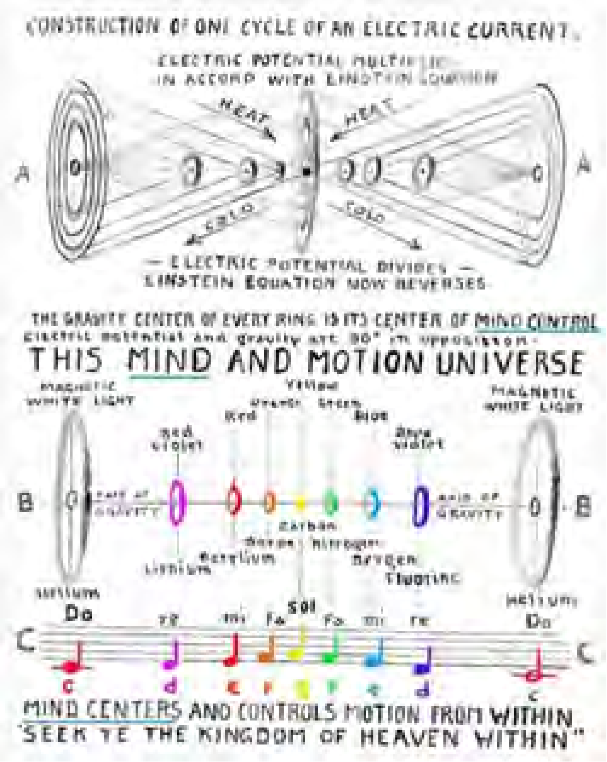
CONSTRUCTION OF ONE CYCLE OF AN ELECTRIC CURRENT.
ELECTRIC POTENTIAL MULTIPLIES IN ACCORD WITH EINSTEIN EQUATION
HEAT
HEAT
A
A
COLD
COLD
— ELECTRIC POTENTIAL DIVIDES — EINSTEIN EQUATION NOW REVERSES
THE GRAVITY CENTER OF EVERY RING IS ITS CENTER OF MIND CONTROL
electric potential and gravity are 90° in opposition.
THIS MIND AND MOTION UNIVERSE
MAGNETIC WHITE LIGHT
Yellow
Orange
Green
Red
Blue
Red violet
Blue violet
MAGNETIC WHITE LIGHT
B
AXIS OF GRAVITY
AXIS OF GRAVITY
B
Carbon
Boron
Nitrogen
Beryllium
Oxygen
Lithium
Fluorine
Helium
Sol
Helium
Do
re
mi
Fa
Fa
mi
re
Do
C
C
c
d
e
f
g
f
e
d
c
MIND CENTERS AND CONTROLS MOTION FROM WITHIN
"SEEK YE THE KINGDOM OF HEAVEN WITHIN"

Figure 4.9: Neutral Center is Mind or Consciousness of God
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/mind-motion400.jpg
(courtesy University of Science and Philosophy)

"No machine heretofore constructed has been made with a neutral center. This conception of mechanics has never before dawned on man's thought field. Had this been done, perpetual motion[48] *would have become a demonstrated fact. Were a machine so constructed as to use its properties, an introductory impulse*[49] *would suffice to run it for centuries. However, this would not be a useful mechanical contrivance for no more energy could be obtained from it than was originally given, and its only value would be as a timekeeper.*[50]

"Keely did not seek to invent, nor did he claim to have invented, perpet-

---

[48] [Perpetual Motion], [Continuous Motion]
[49] [Introductory Impulse]
[50] [Continuous Motion]

ual motion.

"The physicist tells you that "you cannot make something out of nothing;" that "in the economy of nature profit and loss must balance;" that "no matter what the nature of the force may be, its production must necessarily be accompanied by a corresponding expenditure of force in some form or other," etc., etc. But, in the prodigality of nature, this energy flows, without measure and without price, from the great storehouse of the Infinite Will. From the sympathetic portion of the etheric field, all visible aggregations of matter emanate, and on the same order that molecular masses of all living organisms are vitalized by the sympathetic flow from the brain."[51]

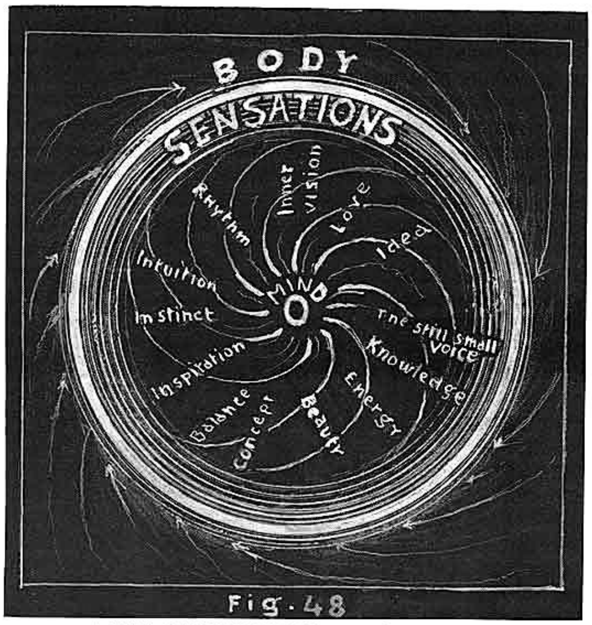
BODY
SENSATIONS
Rhythm
Inner vision
Love
Idea
Intuition
Instinct
MIND
The still small voice
Knowledge
Energy
Beauty
Concept
Balance
Inspiration

Fig. 48

A MIND-CENTERED THOUGHT-RING OF MOTION

Figure 4.10: Neutral Center is as mind or consciousness
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/ring-of-motion.jpg
(courtesy University of Science and Philosophy)

"When this great scientific and religious truth has been made known, and established by demonstration, all controversy as to the source of energy will

---

[51] Bloomfield-Moore ([CJBM]) in [Vibratory Sympathetic and Polar Flows]

be for ever silenced. If I am the chosen instrument to develop this knowledge, and to make known the conditions which surround this pure truth, it is only that I may hand the key to those who will use it to enter the doorway that opens into the inaudible, and thus gain an insight into the now invisible region of the operation of Nature's most powerful governing forces, in the control over terrestrial matter by celestial mind."[52]

4.5.1 Center of Sympathetic Attraction

"I find that there is no medium in the range of vibratory philosophic research, that is as unerringly exact, towards the center of sympathetic attraction, as the negative attractive influence of a certain triple association of the metallic masses of gold, silver and platina[53]*. In fact, they are as accurate indicators of the earth's terrestrial sympathetic envelope, and its triple focalized action towards the earth's neutral center, as the magnet is an indicator of the diversion of the attractive flow of the dominant current of the electric stream. Although much has been written on the subject, the conditions attending the continuous flow of the magnet remains a problem that has never been solved by any other theory. Yet the solution is very simple when harmonic vibratory influence is brought to bear upon it.*[54]

4.5.2 Center of Sympathetic Coincidence

"Where the vibrations under this mode meet, and are maintained in a state of mutual affinity or equilibrium, there is established what is called a "neutral center", or, as otherwise expressed, "a center of sympathetic coincidence"."[55]

4.5.3 Center of Association and Concentration

A shared point, point of convergence or center of sympathetic coincidence (Neutral Center) or where syntropic forces meet, coincide, centralize or come together or harmonize.

---

[52] [Keely] in [Keely and His Discoveries] pg. 365-367

[53] [Trexar], [Trextrinar]

[54] [Keely], [ANSWERS MADE IN LETTERS FROM MR. KEELY - TO QUESTIONS ASKED OF HIM]

[55] [KEELYS PHYSICAL PHILOSOPHY - Snell]

4.5.4 Center of Focalization and Transformation

In the following quote we see the evolution or creation of perceptible light when thought (imperceptible) is directed to a center, undergoes transformation via Law of Cycles, then is manifest as perceptible light.

"Light is induced by electromagnetic percussion emanating from the ether, and in its action represents the plane of magnetism. In fact, it is the plane of magnetism when under polarization. (Platina wires the thickness of a fine hair associated with each of the nine nodal beads, and concentrated towards a general center of focalization, attaching the other end of the wires to the focal center, will determine, by the magnetic conduction, the number of corpuscular oscillations per second induced by a thought[56]*, either positive or negative, in the central centers. These are the only conditions those of magnetic conduction whereby the evolution of a thought can be computed in regard to its force under propagation, as against the amount of latent energy set free to act as induced by such thought on the physical organism.) Some scientific theories of the past have taught us that electricity and magnetism are one and the same thing. Sympathetic vibratory philosophy teaches that they are two distinct forces of one of the triune sympathetic family."*[57]

4.5.5 Nucleal Center

Term used by Russell to denote the point in a wave of maximum polar opposition or the 4++[58] state of polar opposition and when depolar as the Plane of Inertia[59]. Also understood to be the same as Keely's Neutral Center where Keely mentions 'antagonism'[60], *etc.*

In between one state and its opposite is a moment of pause not unlike the furthest point in the swing of a pendulum or seesaw. This null point or fulcrum can endure both polarities without itself being affected. See [Philosopher's Stone] There are in actuality two fulcrum points or points of transformation. The first is the 4++ center position where opposing polarities seek to Void at maximum opposition. The other fulcrum or transformation point is what Russell called the Plane of Inertia where polarities are at their least opposition. The sympathetic streams wind or spiral towards the center of 4++ and unwind spiraling away from that center to the six Planes of Inertia.

---

[56] [Thought Force]

[57] [Keely], [Snell Manuscript - the book], [ELECTROMAGNETIC RADIATION - Snell]

[58] [4plusplus]

[59] [Plane of Inertia]

[60] [antagonism]

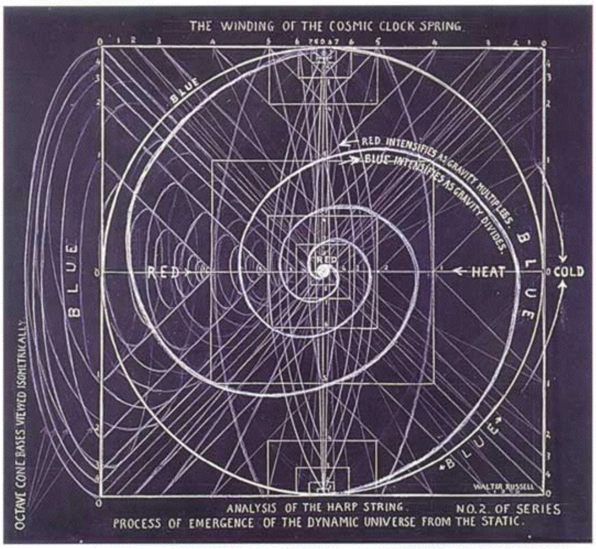
THE WINDING OF THE COSMIC CLOCK SPRING.
BLUE
RED INTENSIFIES AS GRAVITY MULTIPLIES.
BLUE INTENSIFIES AS GRAVITY DIVIDES.
BLUE
RED
RED
HEAT
COLD
BLUE
BLUE
OCTAVE CONE BASES VIEWED ISOMETRICALLY.
WALTER RUSSELL
ANALYSIS OF THE HARP STRING.
NO.2. OF SERIES
PROCESS OF EMERGENCE OF THE DYNAMIC UNIVERSE FROM THE STATIC.

Figure 4.11: Syntropy winding up force to 4++ center
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/67-Winding.GIF
(courtesy University of Science and Philosophy)

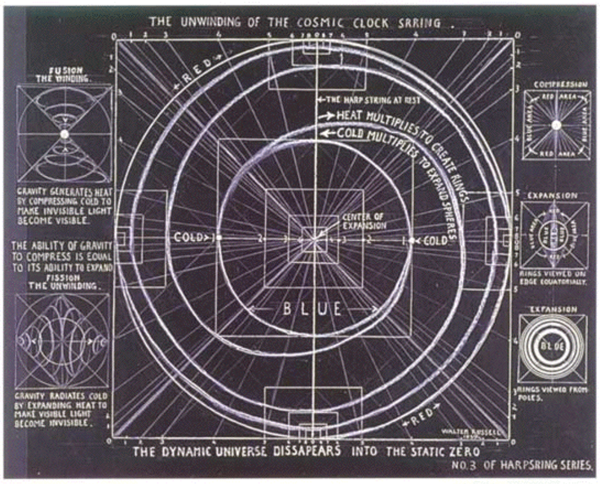
THE UNWINDING OF THE COSMIC CLOCK SRRING .
FUSION
THE WINDING.
GRAVITY GENERATES HEAT BY COMPRESSING COLD TO MAKE INVISIBLE LIGHT BECOME VISIBLE.
THE ABILITY OF GRAVITY TO COMPRESS IS EQUAL TO ITS ABILITY TO EXPAND
FISSION
THE UNWINDING.
GRAVITY RADIATES COLD BY EXPANDING HEAT TO MAKE VISIBLE LIGHT BECOME INVISIBLE.
RED
THE HARP STRING AT REST
HEAT MULTIPLIES TO CREATE RINGS
COLD MULTIPLIES TO EXPAND SPHERES
CENTER OF EXPANSION
COLD
COLD
BLUE
RED
COMPRESSION
RED AREA
BLUE AREA
BLUE AREA
RED AREA
EXPANSION
RINGS VIEWED ON EDGE EQUATORIALLY.
EXPANSION
BLUE
RINGS VIEWED FROM POLES.
WALTER RUSSELL
THE DYNAMIC UNIVERSE DISSAPEARS INTO THE STATIC ZERO
NO.3 OF HARPSRING SERIES.

Figure 4.12: Entropy unwinding energy to 0 Plane of Inertia
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/68-unwinding.gif
(courtesy University of Science and Philosophy)

4.5.6 Resistance

Resistance in DC circuits is equivalent to impedance in AC circuits. It is quite clear the concepts of resistance in the writings of Keely and Russell are more akin to AC impedance than DC resistance (as commonly understood). Resistance or antagonism is the presence of discords. As two poles harmonically accumulate force through mutual syntropic assimilation and at the 4++ point of convergence resistance or antagonism is created ultimately resulting in activating the Law of Cycles by way of run-away development of discords.[61] These discords multiply in number and accumulate power until the unifying harmonic forces revert from harmonic assimilation to enharmonic dispersion. No doubt this dynamic is the cause of implosion and cavitation processes and activities.

---

[61] [overtone series]

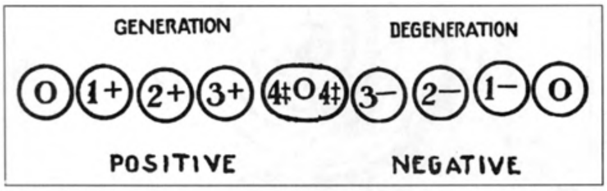
GENERATION
DEGENERATION
O
1+
2+
3+
4‡O4‡
3−
2−
1−
O
POSITIVE
NEGATIVE

Fig. 68
Two-way journey from zero—through zero—to zero

Figure 4.13: Russell's 4 plus plus center of polar opposition
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/4plusplus2.jpg
(courtesy University of Science and Philosophy)

About resistance Keely says: *"..In crystalline structures, heat, which expands the atoms, by twisting them produces striae*[62]*, increases the resistance, etc."*[63]

---

[62] [striae]

[63] Keely, [Law of Vibrating Atomolic Substances]

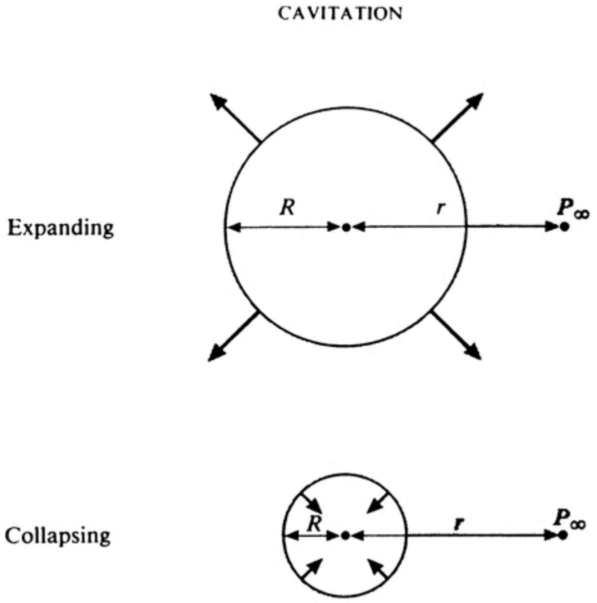

Figure 4.14: Cavitation is an example of the Law of Cycles
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/Cavitation_Bubbles.jpg
(courtesy University of Science and Philosophy)

Keely continues: *"The electric current destroys cohesion and chemical tension directly as square of current in amperes, inversely as the resistance in ohms, inversely as the chemical equivalent, and conversely as the coefficient of the difference between the freezing and volatilizing temperature of mass acted upon."*[64]

Russell describing the opposing spirals: *"Positive electricity pulls inward spirally from within against the opposing resistance of negative electricity which thrusts outward spirally from within."*[65]

Russel continues: *"In this electric universe, sensation is the strain of resistance to the separation which exists between all separated masses. All matter is one. Separated particles desire to find that oneness."*[66]

---

[64] Keely, [Law of Variation of Atomic Oscillation by Electricity]

[65] Russell, [The Secret of Light], page 148

[66] Russell, [The Secret of Light], page 85-86

4.5.7 Prime Neutral Center

Prime Neutral Centers are primary centers about which other subservient secondary neutral centers are attracted and aggregate. In the quote below these Prime Neutral Centers are the neutral centers as stars or suns about which aggregate the secondary neutral centers of their planets. Hence they, the neutral centers are the source of gravity.

"The luminous, etheric, protoplastic element, which is the highest condition[67] *of the ether, fills the regions of infinite space, and in its radiating outreach*[68] *gives birth to the prime neutral centers that carry the planetary worlds through their ranges of motion."*[69]

The actions of Neutral Centers (almost) dictates the existence of rotating shells surrounding centralizing (gravitic) neutral centers. This not unlike so-called electron shells. Hence if every thing or mass has a neutral center and all neutral centers behave according to the same laws then planets would by those same laws (almost) have to be hollow. If it walks like a duck and quacks like a duck then it is a duck. I'm not saying they are but logic points to such a conclusion.

It is not physicality that causes negative attraction but the activated neutral centers that cause negative attraction. It is an illusion of the physical to ascribe causative properties to it when it is in fact nothing more than an effect of the cause - neutral center activity. The neutral centers formed the instant Divine self awareness occurred. It was all created in an instant of no duration and continues to recreate itself as it evolves. *The human intellect has a hard time with the concept of no time.*[70]

4.5.8 Supreme Neutral Center

Every molecule, atom and sub-quantum centralization everywhere existent within the universe has a Neutral Center. All of these neutral centers are sympathetic each to (entangled with or 'as one' with) all others. What keeps all these Neutral Centers "in order" (having one shared state) is subservience to One Supreme Neutral Center of the universe. Some would call this One Supreme Neutral Center, God. Taking out the egocentric anthropomorphic[71] assignments we then have an Intelligence (Consciousness) intimately associated with (in sympathy or entangled with) everything that is. The connecting link[72] between all these centers is Compound Interetheric sympathy which Keely said above '*'fills the regions of infinite space*". These ubiquitous Neutral Centers constitute a state of perfect continuity and living extensions of the One that through progressive evolution[73] or devolution as the case may be manifest as seeming solid matter in all its motions and activities which motion is Life.[74]

---

[67] [Interetheric]

[68] [outreach]

[69] [Keely], [Keely and His Discoveries] pg 270

[70] [Time]

It is the nature of neutrality as Undifferentiated Mind Force that fills every thing everywhere. This Undifferentiated Mind Force is everywhere present as "Sympathy" within which, as motion (form), is latent. Being of infinite density and extenuation (although not having dimension) it is ubiquitous and reactive to and action upon every neutral center everywhere without itself being changed. See [Philosophers Stone]. It is the Supreme Prime Neutral state or condition (Cause) governing all else, manifested or not. The state of Neutrality is "perfect continuity without extension". Density and extenuation are dimensions according to Walter Russell so they do not actually exist except in the intellect of Man; *i.e.*, within his definition of matter and energy.

A neutral center is a conjunction of coincidence and focalization of forces. If those centralized forces (molecules, planets, galaxies) cease to exist so too does their controlling neutral center which created them and by which they have their being. They, the cause and its effects, exist concurrently and codependently. In other words "can't have one without the other".

"All motion is thought, and all force is mind force."[75]

This then clarifies the adage all, the many, comes from One. The state of the one Supreme Neutral Center does not change yet it births and governs all else. No state of a material thing is fixed or stable within the universe as everything is in constant compound rhythmical (vibratory) motions. These states of motion are governed by invisible yet knowable principles and laws. It could be said the principles and laws governing these motions are fixed or stable.

---

[71] made in Man's egoic image, anthropomorphizied

[72] [connecting link]

[73] [Progressive Evolution]

[74] [Dale Pond], [Prime Neutral Center]

[75] Keely, [Keely and His Discoveries], pg 252

"Macvicar[76] foreshadowed the teachings of this new philosophy when he wrote, *"All motion in the universe is rhythmical[77]. This is seen in the forward and backward movement of the pendulum, the ebb and the flow of the tides, the succession of day and night, the systolic and diasystolic action of the heart[78], and in the inspiration and expiration of the lungs. Our breathing is a double motion of the universal ether, an active and a reactive movement. This androgyne principle, with its dual motion, is the breath of God[79] in man."* The writings of the ancients teem with these ideas, which have been handed down to us from generation to generation, and are now flashing their light, like torches in the darkness, upon mysteries too long regarded as "*lying outside the domain of physical science[80].*"[81]

How can mind be devoid of thought/awareness? Desire is an aspect of mind, for instance. A quiet mind of knowing contains thought (depolar motionless awareness) but it does not contain egoic or intellectual thinking (polar motion). The physical universe is a recording of thinking dynamics; its existence is manifestation of the originating knowing thought activated by desire. Just as a seesaw contains a fulcrum that does not move it also contains extensions that do move. A seesaw is a trinity of states. Same thing with Divine Mind (the fulcrum or Cause) and everything else (the extensions or effects) in the universe.

---

[76] Dr. John Gibson [Macvicar] went to the university of St. Andrews in 1814, where he excelled in mathematics and natural philosophy, and thence to Edinburgh, where he studied chemistry, anatomy, and natural history under John Knox's tutelage, together with rhetoric, Hebrew and church history (see Dictionary of National Biography).

[77] same as Russell's [Rhythmic Balanced Interchange]

[78] [Universal Heart Beat], [Dynaspheric Force]

[79] [breath of life]

[80] classical physics

[81] Bloomfield-Moore ([CJBM]), see [Father-Mother Principle], [Universal Heart Beat]

Chapter 5

Mind and Mind Force

5.1 Mind is Force

Universal Undifferentiated Mind is the One Force from which all other forces are derived. This premise is born out in the writings of many great philosophers, thinkers and sundry historical inspired texts.

"All forces in nature are mind forces: magnetic, electric, galvanic, acoustic, solar, are all governed by the triune streams of celestial infinity; as also the molecular, intermolecular, atomic, and interatomic[1]. The remote depths of all their acoustic centers become subservient to the third, sixth, and ninth position of the diatonic, harmonic and enharmonic chords; which, when resonantly induced, concentrate concordant harmony[2], by reducing their range of corpuscular motion, drawing them as if towards each other's neutral center of attractive infinity."[3]

Keely writes *"Luminiferous ether[4], or celestial mind force, a compound interetheric[5] element, is the substance of which everything visible is composed. It is the great sympathetic protoplastic element, life[6] itself. Consequently, our physical organisms are composed of this element ... its seat in the cerebral*

---

[1] see [Subdivision]

[2] [Syntropy]

[3] Keely, see [Mind Force the hidden Scalar Force]

[4] [Luminiferous ether]

[5] [compound interetheric]

[6] [life]

convolutions ... This sympathetic outreach[7] *(sympathetic radiation) is* *mind flow* *proper, or* *will force* *- sympathetic polarization to produce action, sympathetic depolarization to neutralize it. Polar and depolar differentiation*[8] *result in motion. The true protoplastic element ... permeates all forms and conditions of matter, having for its attendants, gravity, electricity and magnetism, the triple conditions born in itself. In fact, it is the soul of matter*[9]*, the element from which all forms of motion*[10] *receive their introductory impulse*[11]*.*"[12] (underline added)

"The luminiferous ether the compound interetheric element in other words, *celestial mind force*[13] *is the substance of which all visible and invisible things are composed.*"[14] (underline added)

That Mind Force is a real force has been studied over a 28 year period by the engineering department of Princeton University in their PEAR Proposition project.[15] Their findings were revolutionary with far ranging implications.[16]

"Nearly three decades of intense experimentation leave little doubt that the anomalous physical phenomena appearing in the PEAR studies are valid, and are significantly correlated with such subjective variables as intention, meaning, resonance, and uncertainty."[17]

5.2 Whole vs Split Mind

Mind or Whole Mind means total Awareness unless otherwise specified. The term Mind or Whole Mind does not mean "intellect"[18]. Intellect is of the ego and is a subset of the general Consciousness or Mind and is that limited awareness we have created as our *interpretation* of reality. The term "Mind" as used in SVP means the general Consciousness, Awareness, Whole or Undifferentiated Mind. *Thinking* is of the differentiated intellect, ego or Split Mind while *knowing* is of the Mind, heart or Whole Mind. Whole Mind knowing is knowing with the feeling of the heart while Split Mind thinking or intellect is of the limited duality of the ego.

---

[7] [sympathetic outreach]
[8] [Differentiation]
[9] [soul of matter]
[10] [Motion]
[11] [introductory impulse]
[12] [THE NEUTRAL CENTER]
[13] [celestial mind force]
[14] [Ether - Snell]
[15] http://www.princeton.edu/ pear/
[16] http://www.princeton.edu/ pear/implications.html
[17] video http://vimeo.com/4359545
[18] [intellect], [ego]

"*Your ego is a limitation within your Mind. Your ego exists as an expression of limitation.*"[19]

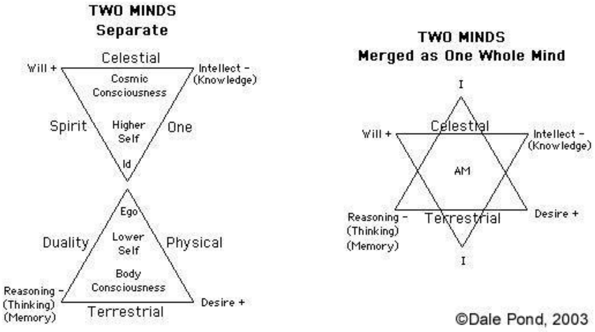
TWO MINDS
Separate
Celestial
Will +
Intellect -
(Knowledge)
Cosmic
Consciousness
Spirit
Higher
Self
One
Id
Ego
Duality
Lower
Self
Physical
Body
Consciousness
Reasoning -
(Thinking)
(Memory)
Terrestrial
Desire +
TWO MINDS
Merged as One Whole Mind
I
Will +
Celestial
Intellect -
(Knowledge)
AM
Reasoning -
(Thinking)
(Memory)
Terrestrial
Desire +
I
©Dale Pond, 2003

Figure 5.1: Whole and Split Mind
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/two-minds.jpeg

5.3 Mind is of the Sixth and Seventh Subdivisions

Mind is of the Compound Interetheric[20] or seventh subdivision[21] of matter and energy. See Table 5.1. According to Keely's molecular and atomic morphology a molecule may be dissociated into its constituent parts or atoms. Atoms may then be further dissociated into their constituent parts and so forth for all finer subdivisions until the seventh subdivision where the rarefied ubiquitous sympathetic substance of Mind exists.

---

[19] Tom Carpenter, [Dialogue on Awakening], pg. 29

[20] [Compound Interetheric]

[21] [subdivision]

Dimension	**Subdivision**	**Relation**
1	Molecular	separation of air brings us to the first subdivision only.
2	Intermolecular	to the second; see [supercritical fluid]
3	Atomic	to the third;
4	Interatomic	to the fourth;
5	Etheric	to the fifth; and
6	Interetheric	to the sixth subdivision, or positive association with luminiferous ether;
7	Compound Interetheric	as the seventh subdivision.

Table 5.1: Keely's Seven Subdivisions of Matter and Energy

It will be noted these seven states of matter and energy are not unlike the ancient states of matter and energy: Earth, Water, Air and Fire.

5.4 Keely - Substance and Nature of Mind Force

"There is a celestial mind force[22], a great sympathetic force which is life itself, of which everything is composed."[23]

"All motion is thought, and all force is mind force."[24]

"There is no dividing of matter and force into two distinct terms, as they

---

[22] [Mind Force]

[23] Keely

[24] Keely, [Keely and His Discoveries], pg 252

both are ONE. FORCE is liberated matter. MATTER is force in bondage."[25]

The above last two sentences sound like something Albert Einstein would have said roughly twenty years later.

"The mighty forces latent in corpuscular matter, by which we are surrounded, are all held in oscillating vortex action by the Infinite Designer[26] *of workings hidden from us, until the time is ripe for their disclosure. This latent, registered power*[27] *interchanges sympathetically with the celestial radiating streams, whereby light, heat, electricity, magnetism and galvanic action are propagated in their different orders, vitalizing all nature with their life-giving principles.* (underline added)

"When this great scientific and religious truth has been made known, and established by demonstration, all controversy as to the source of energy will be for ever silenced. If I am the chosen instrument to develop this knowledge, and to make known the conditions which surround this pure truth, it is only that I may hand the key to those who will use it to enter the doorway that opens into the inaudible, and thus gain an insight into the now invisible region of the operation of Nature's most powerful governing forces, in the control over terrestrial matter by celestial mind."[28] (underline added)

"The action of the mind itself is a vibratory Etheric evolution, controlling the physical, its negative power being depreciatory in its effects, and its positive influence elevating."[29]

"The substance of the brain[30] *is molecular. The mind which permeates the brain is interetheric in substance; it is the element by which the brain is impregnated. This element, when excited into action, controls all physical motion as long as the necessary sympathetic conditions are maintained. Then conditions are no more immaterial in their character than are light and heat. Electricity, magnetism, gravity, and heat are latent in all aggregations of mat-*

---

[25] [Keely]
[26] [God]
[27] [Latent Force]
[28] Keely in [Keely and His Discoveries] pg. 365-367
[29] [Keely]
[30] [brain]

ter. They are not obtained from terrestrial influences. Celestial radiation[31] *is the true impregnating medium in all these forces. The brain is the high resonating receptacle where the sympathetic celestial acts, and where molecular and atomic motion are induced, as according to the intensification brought to bear upon it by radiation.*"[32]

"The sympathetic conditions that we call mind are no more immaterial in their character than light or electricity. The substance of the brain is molecular, while the substance of the mind that permeates the brain is interetheric and is the element by which the brain is impregnated, exciting it into action and controlling physical motion. In order to trace the successive triple impulses[33]*, taking the introductory one of sympathetic negative outreach*[34]*, towards the cerebral neutrals, which awaken the latent element to action, we find that mind may be considered a specific order of interatomic motion sympathetically influenced by the celestial flow and that it becomes when thus excited by this medium a part and parcel of the celestial itself. The brain is not a laboratory. It is as static as the head of the positive negative attractor*[35]*, until influenced by certain orders of vibration, when it reveals the true character of the outreach so induced. The brain is the high resonating receptacle where the sympathetic celestial acts, and where molecular and atomic motion exhibits itself, as according to the intensification brought to bear upon it by the celestial mind flow. The question arises, Why is this condition of the (luminous) ether always under a state of luminosity of an especial order? Its characteristics are such, from its infinite tenuity and the sympathetic activity with which it is impregnated, that it possesses an order of vibratory, oscillatory velocity, which causes it to evolve its own luminosity. This celestial, latent power, that induces luminosity in this medium, is the same that registers in all aggregated forms of matter, visible and invisible. It is held in the corpuscular embrace until liberated by a compound vibratory negative medium. What does this activity represent, by which luminosity is induced in the high etheric realm, if not to indicate that even this order of ether is bounded by a greater region still beyond? The activity of the corpuscles represents its outflow from the luminiferous track towards the molecular centers of*

---

[31] [Celestial Radiation]

[32] Keely, [The Operation of the Vibratory Circuit]

[33] [Musical Triplet]

[34] [sympathetic outreach]

[35] [Keelys Mechanical Inventions and Instruments]

neutrality, revealing the connecting link[36] *between mind and matter. These conditions of luminosity have no thermal forces associated with them, although, paradoxically, all thermal conditions emanate from that source. The tenuity of this element accounts for it. It is when the streams come in conflict with crude matter that heat is evolved from its latent state and a different order of light from the etheric luminous is originated, the sun being the intermediate transmitter. All planets and systems are sympathetic intermediates, the whole of one system, connected in sympathy for each other, the Brain of Deity."*[37] See Figure 4.7.

5.5 Russell on Mind and Mind Substance

Keely uses the term 'luminosity' or 'luminiferous ether'[38] which Russell calls 'white still magnetic light of gravity' or simply 'magnetic light'.

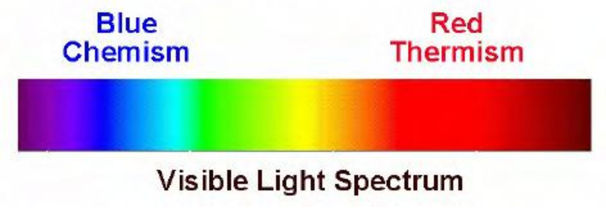

Figure 5.2: Red and Blue halves of the visible light spectrum
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/visible-light-spectrum.jpg

"Electricity is the divider of the invisible white still magnetic light[39] *of gravity - which Mind is - into the red and blue halves of the visible light spectrum for the purpose of creating male and female bodies which alone constitute all matter."*[40]

---

[36] [connecting link]
[37] Keely, [Dashed Against the Rock]
[38] [7B.09 - Luminiferous Ether or Light]
[39] [Magnetic Still Light]
[40] Walter Russell, notation on chart, 1955

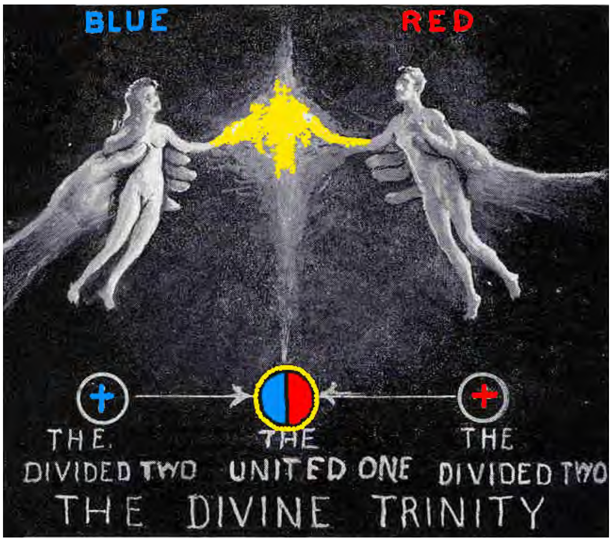

Figure 5.3: Red and Blue halves of Polar states
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/female-male-color.jpg

"This universe of matter in motion is but the electric record of thought."[41]

"Light is the living substance of Mind in action. It is the creating principle of the One substance.

"The One substance is the etheric "spiritual" substance of the One universal Mind.

"The entire "created" universe of all that is, ever has been, or ever will be, is but the One substance in motion, light."[42]

In other words matter is ultimately constituted of the polar concentrated and solidified substance of thought or Mind Force.

5.6 Mind is a Substance

Mind is substance or matter. Mind is a tenuous (ultra-gaseous) matter like ether or plasma (quantum substances). Mind is an etheric (quantum) substance (ultra-plasma). Mind therefore possesses vibratory, chemical and other knowable and engineerable attributes. As Mind is substance it can be regulated, harnessed and manipulated as any other gaseous substance.

---

[41] Walter Russell, [The Secret of Light]
[42] [Ether - Russell]

5.7 Mind is also as an aggregate of consciousness

Mind is also as an aggregate of consciousness (self awareness) and is comprised of three distinct subsets:

Compound Interetheric

Superconscious.[43] Also known as Christ Consciousness[44], Cosmic Consciousness[45], Knowing, Wisdom, awareness and science.

Interetheric

Consciousness. Also known as intellect, ego, reason, thinking.

Subconscious. Also known as Soul, body-mind and instinct.

Unconscious. Belief, Habit, Natural Man

Etheric

Tenuous matter that composes the plasma gases of the body actuated (vivified) via our conscious and unconscious thinking and (beliefs, cognized or not).

Interatomic

The liquids and solids composing the body actuated by the etheric or quantum mind forces.

Atomic

Atoms

Molecular

Simple, Compound and Complex Molecules

It is because of this hitherto unsuspected property of mind and consciousness as a form of tenuous matter the link is established between so-called hard science (3D, materiality, classical physics) and so-called sacred science or spiritual fields of study[46] now loosely called non-classical physics. They are in fact one field of study when this connection is recognized and admitted.

---

[43] [Superconscious]

[44] [Christ Consciousness]

[45] [Cosmic Consciousness]

[46] [telekinesis], [telepathy]

Chapter 6

What is Vibration?

6.1 Quotes - vibration defined

"Vibration is the rhythmical motion of a body within itself."[1]

"Vibration is a periodic interchange of state."[2]

"Rhythmic Balanced Interchange."[3]

"All force is vibration..." Cayce (900-422) *"So is matter."* Cayce (1861-16)

"All comes from one central vibration - taking different form." Cayce (900-422)

"Everything is vibratory." Cayce (195-54)

"As we see manifest in the electrical forces as used by man. This becoming only an atom in motion, and as the atomic force gathers this, producing such vibration as to create heat, light, and of the various natures, by the kind, class or nature of resistance[4] *met in its passage in the cycle given, reducing or raising the velocity, or better by the class of atomic force it vibrates, either*

---

[1] Keely, see [Oscillation]

[2] [Dale Pond]

[3] [Russell]

[4] [Law of Vibrating Atomolic Substances]

with or against. These are examples of portions of universal forces."[5]

"Vibration is movement. Movement is activity of a positive and negative force."[6]

"Electricity[7] *or vibration is that same energy, same power, ye call God."*[8]

"Life in its manifestations is vibration. Electricity is vibration. But vibration that is creative is one thing. Vibration that is destructive is another. Yet they may be from the same source. As in the electrical forces in the form or nature prepared even for use in the body. Remember life is vibration. So is mind. So is matter."[9]

6.2 A Vibration Is

A vibration is a rhythmic periodic expansion (entropy/dispersion) and contraction (syntropy/aggregation) change of state; *i.e.*, a periodic inter-exchange of seemingly opposite yet co-joined and inseparable polar states. The two polar states are entropy (expansion) and syntropy (contraction). In each wave or vibration there are two distinct yet related unseen sets of dynamic aliquot parts or constructive currents (when considering a wave train or continuous vibration as a vibratory stream or vibratory flow). One set of aliquot attributes or parameters brings about the periodic concentration or aggregation of the vibrating media while the other set of aliquot attributes or parameters causes periodic expansion or dispersion of the media. If these unseen influences (aliquot or scalar components) were not there and a dynamical constituent of each wave or vibration there could be no delta or change of state as both states or phases would be identical and unchanging. The perceived wave or vibration is then the effect of these unseen causative (scalar) influences, attributes, parameters or currents. The details of these unseen causative (scalar) influences, attributes, parameters or currents are the subject of this presentation and are also presented in Laws of Being[10], Laws of Being - Annotated[11], Wavefunction[12] and Part 12 - Russells Locked Potentials[13]. These two states are governed in part by the Law of Harmony[14], Law of Harmonic Pitch[15], Law of Harmonic Vibrations[16] and the Law of Repulsion[17]. See also [Bjerknes Effect] for a mechanical experiment demonstrating these two tendencies of vibratory attraction and repulsion.

---

[5] Cayce (900-17)

[6] Cayce (281-29)

[7] [Law of Vibrating Atomolic Substances]

[8] Cayce (2828-4)

[9] Cayce (1861-16)

Vibrations manifest as three-dimensional (along x, y and z planes) expressions of these unseen creative currents. Vibrations are not well represented by the typical sine wave. The use of the sine wave graphic has become so ubiquitous it is often mistaken for the actual motions of a vibration. These three-dimensional expressions are a periodic expansion and contraction of the elements of the vibrating medium; *i.e.,* the air particles are mutually attractive then mutually repellant according to the Law of Attraction[18] and Law of Repulsion[19]. The contraction is seen as the compressive wave front while the repulsion is seen as the vacuous backside of the compression front. This same dynamic can also be seen as the compressed air at the front of an airfoil and the rarefaction on top of that same airfoil or airplane wing or propeller.

6.2.1 Three Forces action on all Matter and Energy

"All conditions of dispersion and focalization are accompanied by the "celestial mind force" acting upon "terrestrial matter" - (corresponding to the mind force acting on the brain, which is only its molecular instrument.) This force is the first seal of the Book of Vibrational Philosophy - the first stepping stone toward solution of the Source of Life.

"All Nature's forces are mind forces: magnetic, electric, etheric, acoustic,

---

[10] [Laws of Being]
[11] [Laws of Being - Annotated]
[12] [Wavefunction]
[13] [Part 12 - Russells Locked Potentials]
[14] [Law of Harmony]
[15] [Law of Harmonic Pitch]
[16] [Law of Harmonic Vibrations]
[17] [Law of Repulsion]
[18] [Law of Attraction]
[19] [Law of Repulsion]

solar. Any metallic mass can be so impregnated with certain vibrations that it will assume mental attributes - attraction and repulsion. We must first understand the triune conditions or laws of that sympathetic medium which interconnects matter with matter, the triune conditions or laws of Sympathetic Streams and unites resonance of each of the seven subdivisions before we can understand the induction by means of acoustic generators, of magnetic antagonisms in matter and the different forms of energy thereby liberated."[20]

6.2.2 Law of Attraction (syntropy)

"Juxtaposed coherent aggregates vibrating in unison, or harmonic ratio, are mutually attracted."[21]

6.2.3 Law of Repulsion (entropy)

"Juxtaposed coherent aggregates vibrating in discord are mutually repelled."[22]

6.2.4 Law of the Dominant (Celestial Mind Force)

"All wave propagations, electromagnetic or otherwise by being thus refracted can be measured in regard to the time of their propagation all of which are introductorily subservient to the luminiferous ether."[23]

6.3 Rad-Energy - entropic dispersion

Rad-energy[24] is radiant energy - expanding from a center. Sometimes this energy is referred to as dispersive, entropic or discordant as it is both. According to the Bjerknes Effect[25] radiation or radiant energy disperses (because affected media is mutually repulsive according to the Law of Repulsion[26]) from a center losing its amplitude and degree of discord following the Square

---

[20] [Snell Manuscript]
[21] Keely, [Bjerknes Effect]
[22] Keely
[23] Keely, [Luminiferous Ether]
[24] rad-energy
[25] Bjerknes Effect
[26] Law of Repulsion

Law[27] until it reaches zero amplitude and zero discordant content at Russell's Plane of Inertia[28]. The range of dispersion is dependent upon the Introductory Impulse[29] and conductivity of the media. This dispersive energy effects the media in which it travels through modulation and heterodyning[30]. The overall effects of this process is governed by the Law of Transformation of Forces[31].

6.3.1 Introductory Impulse

The introductory impulse sets the tone so to speak of the acoustic sound mentioned above. The introductory impulse may establish an entropic dispersion or syntropic condition according to its overtone content. When the partials are discords the resultant sound will be entropic dispersion. When the partials are concords the resultant sound will be syntropic concentration. The development of discordant overtones is a simple and low powered occurence. All sounds developing from an introductory impulse will create discords. It appears the more power initiating an introductory impulse the more discords are developed according to the Law of Cycles[32].

6.4 Vibration v Oscillation

These two rhythmic motions of vibration and oscillation are not the same. Without vibration and oscillation made distinct seeming unfathomable paradoxes arise. For these paradoxes to be understood the difference between vibration and oscillation has to be clarified.

"Vibration is the rhythmical motion of a body within itself."[33] Not unlike the rhythmic motions of or within the heart without change in spatial location of the heart.

---

[27] Square Law
[28] Plane of Inertia
[29] Introductory Impulse
[30] Heterodyne
[31] Law of Transformation of Forces
[32] Law of Cycles
[33] [Keely]

"Oscillation is a rhythmically recurring translatory movement."[34] Back and forth motion over space not unlike a child's swing or penduluum.

In the ground state, at the moment of inflow of the sympathetic celestial streams or Divine Permeation (spark of Life), vibration is, symbolically, one cps and oscillation is one cps. From that moment on in the process of progressive rhythmic devolution or evolution, due to the Law of Harmonic Pitch[35], Law of Harmonic Vibrations[36], Law of Transformation of Forces[37] and Law of Cycles[38], acting together as one complex or orchestrated law, the One is refracted or differentiated into the multiplicity of materiality (the One becomes the Many) in the natural formation of overtones sometimes called harmonics[39]. These secondary and tertiary overtones are not separate from their source (the fundamental) but are inextricably linked to it as Cause is linked to its Effects. Demonstrating everything-that-is has a common origin or One Source and state of Being (sympathy; *i.e.*, Love or Harmony) regardless of outer appearance (opinion) of separateness and individuality.[40]

6.5 What a Vibration is NOT

A vibration is not a two-dimensional expression as a sine wave. Sine wave patterns are developed from measuring a wave front passing by a measuring device such as a microphone, accelerometer or transducer. As the amplitude changes - at the point of monitoring - a sine wave is traced of the varying amplitude over Time. I spoke about this over simplification in my 1994 SVP presentation in Milwaukee.[41] The sound wave front actually expands simultaneously along all three dimensions (x, y and z). A simple sine wave then is but a trace of what is happening at a given xyz point in space but says next to nothing about what is happening within the three dimensional sound wave as per is originating forces, expanding and contracting dynamics and what is occurring with the periodic or aperiodic motions within the substance of the transmitting medium. The compression portion of the wave front is in actuality the mutual periodic attraction and subsequent concentration between the medium particles due to an increase in syntropic force. Conversely, the less dense (rarefied) wave section is an effect of the entropic forces causing the media particles to be mutually repulsive dispersing them into a relatively rarefied condition. Therefore the amplitude of a wave is predicated on the quantity of syntropy or entropy brought to bear on the vibrating media. The response within the media to an external excitation largely depends on the character (time, duration, condition and location) of the introductory impulse[42] of the excitation. From the above sketch it can be readily seen how Macvicar's Law of Assimilation[43] (Russell's Rhythmic Balanced Interchange[44]) is active in every periodic motion.

---

[34] [Keely]

[35] [Law of Harmonic Pitch]

[36] [Law of Harmonic Vibrations]

[37] [Law of Transformation of Forces]

[38] [Law of Cycles]

[39] In SVP 'harmonics' are divisible by 2. All other developed secondary and tertiary tones are divisible by other divisors such as 3, 5, 7, etc. and are called 'overtones'. Collectively they are called 'partials'.

[40] [Dale Pond]

[41] see videos on page [Basic Principles]

6.6 Wave Propagation

In the air the molecular wave front does not actually propagate as a wave. The wave is an effect of subtler causes which when viewed accumulatively have the *appearance* of a wave. Thus we are fooled when relying on our senses. The elements of the molecular medium (air particles) primarily rotate and oscillate about their centers depending on the nature of the disturbance. It is this disturbance that propagates causing periodic expansion and contraction of that which it encounters along the way and before dissipation. The quantum scalar component propagates as a quantum disturbance through the molecular medium. This disturbance is rhythmical; *i.e.*, (polar) intermittently syntropic then entropic in its effects upon the molecular receptive centers. There occurs an increasing compression to the 4++ potential[45] level where a 90° transformation, via the Law of Cycles, converts syntropic influence from syntropic compression into entropic dispersion (gradients of vacuity). Logically then the time it takes for the invisible internal disturbance to become a visible external phenomenon is the apparent propagation rate. This is a periodic disturbance (time to reach the critical 4++ state plus relaxation time plus reformation time) and takes on the appearance of a wave or vibration. Of course, it should be noted not all rhythmic motions result in a complete 4++ transformation as not all rhythmic motions have sufficient power to achieve transformation. Hence there are low energy waves and vibrations that simply repeat themselves without self-destruction and rebirth in a new energy state.

---

[42] [Introductory Impulse]

[43] [Law of Assimilation]

[44] [Rhythmic Balanced Interchange]

[45] [4plusplus]

6.7 Frequency

It can be argued Keely counted frequency as a dual motion of IN then OUT or OUT then IN depending where one begins counting. If this is true then the apparent oscillation of a pendulum is two oscillations from side to side. It is not certain this method of counting carries throughout the Keely materials but it is good to keep this in mind.

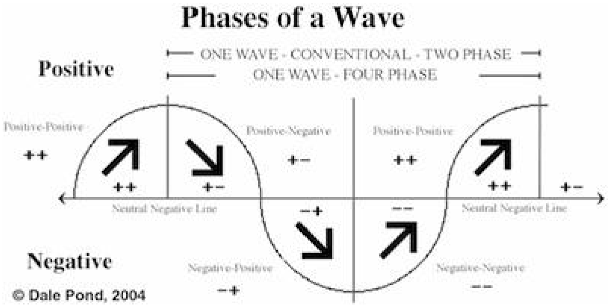
Phases of a Wave
ONE WAVE - CONVENTIONAL - TWO PHASE
ONE WAVE - FOUR PHASE
Positive
Positive-Positive
++
++
+−
Positive-Negative
+−
Positive-Positive
++
++
+−
Neutral Negative Line
−+
−−
Neutral Negative Line
Negative
Negative-Positive
−+
Negative-Negative
−−
© Dale Pond, 2004

Figure 6.1: Four Phases of a Wave or Oscillation
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/fourphases400.gif

6.8 Law of Cycles

"Coherent aggregates harmonically united constitute centers of vibration bearing relation to the fundamental pitch not multiples of the harmonic pitch, and the production of secondary unions between themselves generate pitches that are discords, either in their unisons, or overtones with the original pitch; from harmony is generated discord, the inevitable cause of perpetual trans-

formation."[46]

---

[46] [Keely]

Chapter 7

Laws of Being

7.1 Introduction

Professor Daniel Brinton[1], author of the following reviewed article, spent considerable time in Keely's nearby laboratory and workshop. It is believed this article is the clearest attempt to explain the basic concepts Keely discovered, developed and worked with and how he used them. After all, Keely was present to directly answer Brinton's questions. Therefore it behooves us to closely examine this article to discern understanding and hopefully practical use.

The terms and phrases in this article are linked to supporting materials within the svpwiki.com where the bulk of the known Keely materials are located. Comments are given immediately below a Brinton quote. These comments are quite important because they represent a synthesis of years of learning Keely's and Russell's jargons as well as comparative contemporary concepts and research. A careful reader will see an intermingling of materials from various sources that support each other. Most notably is the mutual affinities between the works and ideas of John W. Keely and Walter Russell.

Brinton's words or quotes are preceded by **Roman Numerals in bold type**. Numbers and brackets ([]) contain comments and links about Brinton's writings and Keely's concepts. These terms and phrases in brackets

---

[1] Daniel Garrison Brinton (May 13, 1837 July 31, 1899) was an American archaeologist and ethnologist at the University of Pennsylvania.

([]) can be searched in www.svpwiki.com. The original unaltered Brinton text can be seen in "Keelys Physical Philosophy - Appendix-I" and "Laws of Being, Appendix-II" at the end of this book or within the svpwiki.com.

7.2 Rhythmical Relations

I. *"The fundamental conception of the Universe is force manifesting itself in rhythmical relations."* (continues to II);[2]

7.2.1 Principle of Rhythmic Balanced Interchange

"This principle of rhythmic balanced interchange[3] *between Father*[4] *(syntropy) - Mother*[5] *(entropy) lights of gravitation (syntropy) and radiation (entropy) is fundamental in all creating things."*[6]

In other words the principle of Rhythmic Balanced Interchange[7] is a universal principle[8][9], as Russell repeatedly emphasized, active in all things from molecular to Mind energy subdivisions (states). Rhythmic Balanced Interchange is after all what every vibration and oscillation does that the ubiquitous and over simplified sine wave does not reveal.

7.2.2 Every individualized object assimilates itself

Macvicar[10] writes: *"Every individualized object assimilates itself to itself in successive moments of its existence and all objects tend to assimilate one another."*[11]

---

[2] [Rhythmic Balanced Interchange]

[3] [rhythmic balanced interchange]

[4] [Father]

[5] [Mother]

[6] Russell, [The Secret of Light], page 150; see 'Chapter 6 - What is Vibration?'

[7] [Rhythmic Balanced Interchange]

[8] [Principles]

[9] see Chapter 11

[10] Dr. John Gibson [MacVicar] professor at university of St. Andrews 1830.

[11] [Snell Manuscript], [Three Laws of Being], [Law of Assimilation], [Modes of Vibration - Annotated]

Please note this law as expressed by Macvicar is the same universal law in its essence and dynamic as expressed by Jesus wherein he says:

"Jesus said unto him, Thou shalt love the Lord thy God with all thy heart, and with all thy soul, and with all thy mind. This is the first and great commandment." (Matthew 22:3738).

"And the second (commandment) is like unto it; Thou shalt love thy neighbor as thyself. On these two commandments hang all the law and the prophets." (Matthew 22:3940).

SVP (nor this author) does not subscribe to or promote organized religions but we are interested in universal laws and principles (applicable universally) regardless who revealed them or wherever they may be found.

Love and Sympathy are synonyms for affinity, assimilation or being One With. These same dynamics apply everywhere and with everything thus establishing sympathetic association[12] between all things, everywhere and at all times.

To repeat when the constituent frequencies of A and B are identical A transmits to B and B transmits to A sympathetic association or mutual resonance (Love or Sympathy) occurs. These dynamics are as expressed in the following of Keely's forty laws[13]:

Law of Harmonic Vibrations[14]

"All coherent aggregates are perpetually vibrating at a period-frequency corresponding to some harmonic ratio of the fundamental pitch of the vibrating body; this pitch is a multiple of the pitch of the atomoles[15]." See Overtone Series[16].

---

[12] [sympathetic association], entangled

[13] [Keelys Forty Laws]

[14] [Law of Harmonic Vibrations]

[15] [atomoles]

[16] [Overtone Series]

Law of Transmissive Vibraic Energy[17]

"All oscillating and vibrating coherent aggregates create, in the media in which they are immersed, outwardly propagated concentric waves of alternate condensation and rarefaction, having a period-frequency identical with the pitch of the aggregate."

Scholium: *"All forms of transmissive energy can be focussed, reflected, refracted, diffracted, transformed, and diminished in intensity inversely as the square of the distance from the originating source."* See also Square Law[18].

Law of Sympathetic Oscillation[19]

"Coherent aggregates immersed in a medium pulsating at their natural pitch simultaneously oscillate with the same frequency, whether the pitch of the medium be a unison, or any harmonic of the fundamental pitch of the creative aggregate."

As the atomole is the constructive constituent of all atoms all atoms are therefore in sympathetic association at, by and through these atomoles. All is one at this level or state of being even though at the physical level of our perception everything appears to be separate and individual.

---

[17] [Law of Transmissive Vibraic Energy]

[18] [Square Law]

[19] [Law of Sympathetic Oscillation]

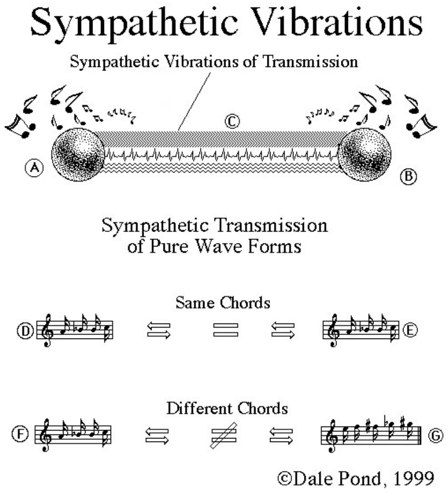
Sympathetic Vibrations
Sympathetic Vibrations of Transmission
Ⓐ
Ⓑ
Ⓒ
Sympathetic Transmission of Pure Wave Forms
Same Chords
Ⓓ
Ⓔ
Different Chords
Ⓕ
Ⓖ
©Dale Pond, 1999

Figure 7.1: Sympathetic Vibration - Entanglement
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/sympathetic-vibrations.jpg

7.2.3 These modes are today much like "phase"

These modes, or periodic change of state, are today much (but not exactly) like what we call "phase" of a wave. Usually conventional 'phase' is undistinguished in construction as to its causative (aliquot/scalar) forces and dynamics being primarily intermittent syntropic and entropic. Phase in modern usage is almost exclusively linked to the positive or negative halves of a sine wave which does not convey all that is occurring in the wave under investigation.[20]

II. ***"This definition [from I] is exhaustive, including both thought[21] and extension[22], mind and matter. The law for the one is the law for the other. The distinction between them is simply relative, i.e., quantitative, not qualitative.***

As above so below - all forms of matter and energy regardless of frequency, subdivision or state are involved in a periodic Rhythmic Balanced Interchange between their polar states where one polar state becomes the other intermittently.

---

[20] see chapters 'Chapter 6 - What is Vibration?' and 'Chapter 8 - Modes of Vibration'

[21] [thought]

[22] [Sympathetic Outreach]

III. ***"The rhythmic relations in which force acts are everywhere, under all conditions, and at all times, the same. They are found experimentally to be universally expressible by the mathematical relations of thirds."***[23]

"Rythmic relations" indicates ALL vibratory and oscillatory motions are effects of this periodic interchange of intertwined poles between syntropic contraction or assimilation and entropic expansion or dispersion. This interchange occurs as these two states are in the relative proportion (ratio) of third parts to one another.

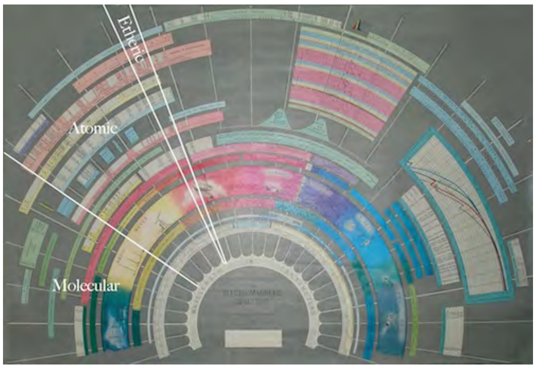

Figure 7.2: Chart showing location of the three mode frequencies.
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/electromagneticspectrummodes500.jpg

That three or threeness is fundamental is now seen to be valid in the Rule of Threes[24] re-discovered by Vitaly Efimov and is called the 'Efimov trimer' or 'Efimov state'[25]. Likewise both Keely in the 1880s and Russell in the 1920s clearly comment on the value of threes[26].

"*Rest and action are three. Space and matter are three. Equilibrium and*

---

[23] [14.35.1 - Keely 3 6 and 9] and [Rhythmic Balanced Interchange]
[24] [Rule of Threes proven]
[25] Wikipedia/Efimov state
[26] [Three]

motion are three. Dimensions and pressures are three. The heartbeat of the universe, and yours, are three. Likewise, its breathings and yours, its temperatures and yours, and all things else of the universe, and you, are three."[27]

7.2.4 'Thirds' is usually a mathematical relationship of proportion

Within this and related SVP documents the concept and use of the term 'thirds' is usually a mathematical relationship of proportion of wave function parameters[28] (of volume, force, density, *etc.*) and not always a musical relation. A given whole being One Whole, has three parts always in varying yet relative proportion to one another. The overall or resultant expression of the One Whole is dependent on the ratios being applied. More syntropy makes the whole more syntropic. The more entropy the more entropic will be the One Whole. Therefore and thereby *polarity is not fundamental but may be modified or manipulated by altering the ratios of these proportions held as thirds or third parts.*[29]

7.3 Threefold Relations

IV. "These threefold relations may be expressed with regard to their results (appearances) as,

I. Assimilative (attractive, aggregating)
II. Individualizing (dispersing, dissociating)
III. Dominant or Resultant (controlling, determining)

---

[27] Russell, [Atomic Suicide]?, Chapter 5 - Prelude - The Transformation of Man

[28] [Eighteen Attributes or Dimensions]

[29] Chapter 10 - Proportions (Ratios) of Relative Values

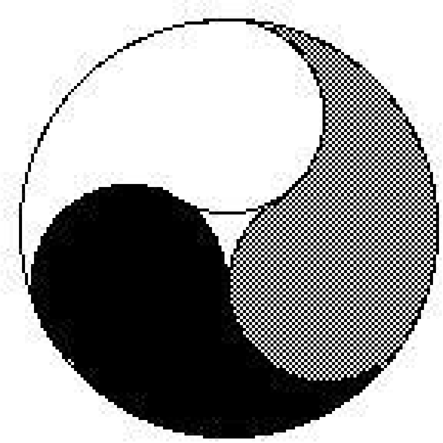

Figure 7.3: Triune Forces/Modes/Proportions combined as One Force
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/Triune-Forces.jpg

7.3.1 Each of the above dynamics (IV) is a $\frac{1}{3}$ part or current of a $\frac{3}{3}$ stream

Each of the above dynamics is a $\frac{1}{3}$ portion or current of a $\frac{3}{3}$ stream (rhythmic flow of force or energy[30]). These three-fold relations are resultant dynamics of three-fold relations of these three currents acting *"as one"* - constituting a single compound stream or chord tone; *i.e.*, combined proportions of each with the others. This combination of three entities to make one compound entity works the same way three musical notes may be sounded at one time as a triplet chord, hence Keely's constant use of triplet chord notations throughout his charts. It is further presumed the ratios of the musical notes within a triplet chord on his charts[31] are the same ratios when specifying any other applied proportional value such as frequency, amplitude, *etc. regardless of frequency range or subdivision.*

---

[30] *"In this system (Sympathetic Vibratory Physics), force and energy are classified as opposites, working in antagonism to each other: Force as a positive power which initiates aggregative motion, and resists separative motion, in three postules, of ponderable matter in the etheric medium; Energy as a negative power, which initiates separative motion, or disintegration, and resists aggregative motion, in three postules of ponderable matter, also of the etheric medium."* Bloomfield-Moore, [What Electricity Is - Bloomfield Moore]

[31] [Part 25 - Keelys Wonderful Charts of Vibratory Etheric Science]

7.3.2 Ratios of Harmonic and Enharmonic Potentials

It is further assumed (until bench testing shows otherwise) these triplet chords when considered as a singular tone each represents a different degree of concord and discord. Tables would have to be developed from bench testing. These ratios of concord and discord would also parallel in some regards Russell's concepts of Locked Potentials because that is what they are.

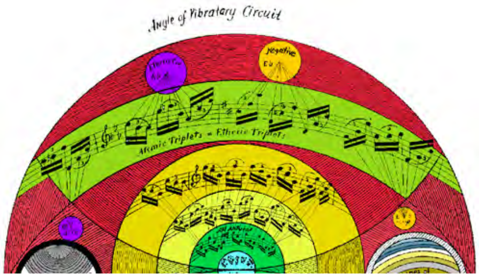
Angle of Vibratory Circuit
Negative
Atomic Triplets = Etheric Triplets

Figure 7.4: Typical Triplet Chords in Keely's charts.
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/Angles-Segment-500.jpg

7.4 Three Actions and Laws of Being

V. ***"From these three actions are derived the three fundamental LAWS OF BEING:***

I. **Law of Assimilation**: every individualized object assimilates itself to all other objects.[32]

II. **Law of Individualization**: every such object tends to assimilate all other objects to itself.[33]

III. **Law of the Dominant**: every such object is such by virtue of the higher or dominant force which controls these two tendencies.[34]

---

[32] [Law of Assimilation]

[33] [Law of Individualization]

[34] these three laws act according to and with the [Law of Sympathetic Association] and [Law of the Dominant]

7.4.1 Law of Sympathetic Association

"The effects of the Law of Sympathetic Association[35]*, which Mr. Keely demonstrated as the governing medium of the universe, find illustrations in inanimate nature. What else is the influence which one string of a lute has upon a string of another lute when a stroke upon it causes a proportionable motion and sound in the sympathizing consort, which is distant from it*[36]*, and not perceptibly touched? It has been found that, in a watchmaker's shop, the timepieces, or clocks, connected with the same wall or shelf, have such a sympathetic effect in keeping time, that they stop those which beat in irregular time; and, if any are at rest, set those going which beat accurately. Norman Lockyer*[37] *deals with the Law of Sympathetic Association as follows: - "While in the giving out of light we are dealing with molecular vibration taking place so energetically as to give rise to luminous radiation, absorption phenomena afford no evidence of this motion of the molecules when their vibrations are far less violent." . . . "The molecules are so apt to vibrate each in its own period that they will take up vibrations from light*[38] *which is passing among them, provided always that the light thus passing among them contains the proper vibrations.*[39]*" . . . "Let us try to get a mental image of what goes on. There is an experiment in the world of sound which will help us." . . . "Take two large tuning-forks, mounted on sounding-boxes, and tuned to exact unison. One of the forks is set in active vibration by means of a fiddle-bow, and then brought near to the other one; the open mouths presented to each other. After a few moments, if the fork originally sounded is damped to stop its sound, it will be found that the other fork has taken up the vibration, and is sounding distinctly. If the two forks are not in unison, no amount of bowing of the one will have the slightest effect in producing sound from the other."*[40]

---

[35] [Law of Sympathetic Association]

[36] spooky [action at a distance]

[37] Sir Joseph Norman Lockyer, FRS (17 May 1836 - 16 August 1920), known simply as Norman Lockyer, was an English scientist and astronomer.

[38] etheric [disturbance of equilibrium]

[39] [Law of Force]

[40] Clara Bloomfield-Moore([CJBM]), [Keely and His Discoveries], [Tuning Forks], [Sympathy]

7.4.2 Law of Assimilation

The Law of Assimilation[41] acts according to the laws of sympathetic attraction and oscillation as also the Law of Sympathetic Oscillation.[42] - *"this law is to the effect that every individualized object tends to assimilate itself to itself, in successive moments of its existence, and all objects to assimilate one another."*[43]

7.4.3 Law of Attraction

"Juxtaposed[44] *coherent aggregates vibrating in unison, or harmonic ratio, are mutually attracted."*[45]

Here then is the great law of attraction properly explained. When two or more objects are 'vibrating in unison' they are sympathetic to each other and what occurs to one occurs to both. The objects are 'linked' by sympathetic vibration and/or sympathetic oscillation. This phenomenon is recognized today as '*entanglement*' or 'spooky action at a distance'.[46] Where objects are 'equally and uniformly spaced' we see the basis for geometric shapes such as tetrahedrons, *etc.* thought so much of by so many these days. Uniformity of design elements or dimensions (geometry) is therefore an attribute of harmony and therefore attraction or assimilation lacking entropic (discordant) elements of dispersion (numbers and ratios other than those forming unisons or harmonic ratios).

7.4.4 Law of Sympathetic Oscillation

"Coherent aggregates immersed in a medium pulsating at their natural pitch simultaneously oscillate with the same frequency, whether the pitch of the medium be a unison, or any harmonic of the fundamental pitch of the creative aggregate."[47]

---

[41] [Law of Assimilation]

[42] [Law of Sympathetic Oscillation]

[43] [Law of Assimilation]

[44] equally and uniformly spaced

[45] Keely, 1894; see [Bjerknes Effect]

[46] [Sympathetic Vibration], [Sympathetic Oscillation], [Entanglement]

[47] Keely, 1894

This coincidental and synchronous action is seen in the coordinated motion of seeming separated clock pendulums all immersed in the common medium of the atmosphere or sharing contact through a wall or shelf. In a similar fashion this is the basis of dissolving foreign materials (aggregates) by solvents in solutions. Curiously it is the same dynamic at work in the Stockholm Syndrome[48]. See II above why this is so.

7.4.5 Infinite neutral Mind centers and controls all motion

Should we replace the non-existent infinitely small strings as held in the String Theory with infinitely small centers[49] to which are applied the above three laws it can be seen how every center is sympathetic (in sympathy) with every other center and are therefore quantum entangled; *i.e.*, sympathetic one to all others and clearly reflects the above three laws and their dynamics. This would of course imply a single fundamental chord at the very heart or core of the universe. This highly sympathetic universal chord could be the *'Cosmic microwave background'*[50] that vibrates and oscillates everything in the universe at a common pitch thus sympathetically connecting (entangling) every seeming individuated thing to all other seeming individuated things regardless of space. At certain subdivision levels it may therefore be presumed all things regardless of location are sympathetically linked and behave as One Thing thus all things may be considered as One yet having individual exist and experience.[51]

Cayce mentions this seeming duality of One yet individual state of affairs as described by the above laws, states and circumstances thusly:

"That it, the entity, may know itself to be itself and part of the Whole; not the Whole but one with the Whole; and thus retaining its individuality, knowing itself to be itself, yet one with the purposes of the First Cause that called it, the entity, into being, into the awareness, into the consciousness of itself. That is the purpose, that is the cause of being."[52](underlines added)

---

[48] Wikipedia/Stockholm Syndrome

[49] [Center Theory]

[50] [Cosmic microwave background]

[51] [What is Sympathy]

[52] Cayce (826-11)

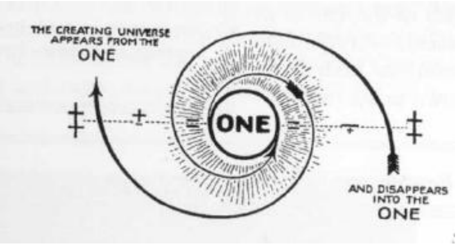
THE CREATING UNIVERSE APPEARS FROM THE ONE
ONE
AND DISAPPEARS INTO THE ONE

Figure 7.5: From the One comes the many.
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/One-Many.jpg
(courtesy University of Science and Philosophy)

And of course this illusion of separateness carries on into the affairs of each individual and hence into society.

As can be seen from the above the totality of the universe is One totality with seeming separate individualizations. Science sees it when it insists on creating experiments isolated from all outside influences. This of course can work in purely mechanical experiments but once the threshold of the atom is breached into the quantum world nothing can be isolated from these higher and finer outside forces and influences. Indeed even the atom is affected by these higher finer forces - we simply do not know how to pay attention to the effects which includes the existence of the atom itself being created and maintained by these unseen forces[53].

How the many forms of matter and substance develop through evolution[54] from the seeming void of Mind or Consciousness is not the topic of this book. Perhaps we'll get around to writing such a book in the future.

7.4.6 One Substance is Mind

There is but One Substance in the universe and that substance is Mind which some have called "Light" as in the Light of Mind. Being of the Light or Etheric order of subdivided being the ancients called it "Luminiferous

---

[53] [7.9 - Ultimate Constitution of Matter and Action of Force Regulating its Phenomena]

[54] svpbookstore.com - Hughes, F.; Harmonies of Tones and Colours - Developed by Evolution

Ether"[55]. Walter Russell expressed it thusly:

"Light units have the appearance of individuality, even as you and I.
"There is no individuality in this universe of Mind. There is but an appearance of individuality.
"This is a universe of the One Thing.
"That One Thing is Mind.
"The substance of thinking Mind in action is light.
"There is but One substance.
"The One substance cannot be divided into many substances, or many parts.
"Individuality is but an appearance, an effect of potential in the periodicity of thinking."[56]

Thus it can be seen Mind centers[57] all things because all things are motions centralized about a common center originating of Mind as Source[58] or the Interetheric realm or subdivision (6th subdivision) of being.

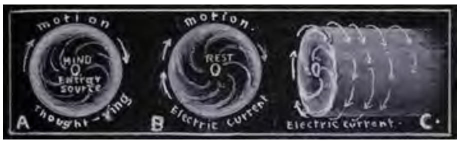

Figure 7.6: Mind centers all motion.
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/thought-rings.jpg
(courtesy University of Science and Philosophy)

VI. *"Applying these fundamental laws to an explanation of the universe, as it is brought to human cognition, all manifestations of force may be treated as modes of vibrations."*

Simplistically everything vibrates and oscillates perpetually.[59] But not all vibrations and oscillations are equal bringing us to the subject of this book. The great secret to understanding is to see and know the unseen working constituents of vibrations and oscillations such that we can engineer them to create the precise waveforms and effects we want. These periodic motions are not solely about sine, square, sawtooth and pulse shaped waves. Carefully sculpting a waveform synthesizing specific partials and harmonics allows a fine precision in applying these waveforms. Likewise understanding how waves interact with each other, as shown in Keely's forty laws,[60] gives us mastery over such vibrating and oscillating dynamics and media.

---

[55] [Luminiferous Ether]

[56] Russell, [The Universal One], Book 02 - Chapter 01 - Dynamics of Mind - Concerning Light Units of Matter

[57] [Figure 12.06 - Mind Centers and Controls Motion]

[58] very important article by Keely [7.9 - Ultimate Constitution of Matter and Action of Force Regulating its Phenomena]

[59] [Law of Harmonic Vibrations]

7.5 Three Modes of Vibration

VII. ***"The essential differences give rise to three modes of vibration:***

I. **The Radiating**: called also the "Dispersing", the "Propulsive", the "Positive", and the "Enharmonic".[61]

II. **The Focalizing**: called also the "Negative", the "Negative Attractive", the "Polarizing", and the "Harmonic".[62]

III. **The Dominant**: called also the Etheric, or the Celestial.[63]

7.5.1 Polarity, Duality, Bias and Potential

Polarity is usually understood as two seemingly opposing forces or states as in for instance the two poles of magnetism. Neither Keely nor Russell saw polarity in this simplistic manner. To them there are two composite poles and a centering fulcrum or neutral point, neutral center or neutral line. All three states are *effects* of subtler unseen forces united in discrete proportions of three united and manifesting as one. A key here is *preponderance* as in one composite force is preponderant over the other within a triplet configuration or pole. Each pole then is a result determined by predominance of one composite force over the other two *within each pole.*

---

[60] [Keelys Forty Laws]

[61] [Entropic]. [Russell] also calls this '[Mother]' or '[Magnetic]'.

[62] [Syntropic] or [Negentropic]. Russell also calls this '[Father]' or '[Electric]'.

[63] [Mind], [Etheric Elements], [Scalar], [Spirit]

7.6 Three Laws of Being

VIII. ***"These, it will be noted, correspond to the three Laws of Being"*** (continues to **IX.**)

Syntropy	**Neutral**	**Entropy**
The Focalizing	**The Dominant**	**The Radiating**
Absorbing	Centering	Emanating
Accumulating	Balancing	Dissipating
Plus	Fulcrum	Minus
Assimilating	Neutralizing	Dispersion
Heating	Zero	Cooling

Table 7.1: Three Polarity States and Effects

6465

7.6.1 Dominant Mode

"By use of the dominant mode, which is allied to the "order" of etheric vibrations, we can induct, "sympathetic negative attraction" or "sympathetic positive propulsion" in any mass, according to its mass chord.[66]

The Dominant mode is of the etheric and interetheric subdivisions that governs the two lower order (molecular and atomic) subdivisions party to the polar triplet.

---

[66] This quote is from [SYMPATHETIC INDUCTION - Snell]. It says there is a polarity to the etheric (quantum) range of vibrations (row four in the above Table). This also implies the two polar streams are different and therefore must be created and applied as separate streams.

"All the dominant conditions of nature represent the focal centers towards which like surrounding ones become sympathetically subservient."[67]

"By use of the dominant mode, which is allied to the "order" of etheric vibrations, we can induct, "sympathetic negative attraction" or "sympathetic positive propulsion" in any mass, according to its mass chord."[68]

"The magnetic cannot lead the electric, nor the electric the gravital, nor the gravital the magnetic. All are subservient to the dominant, as a train of cars is subservient to the locomotive which pulls it along; the only difference between the two is that one is sympathetic, the other mechanical. Though this is a crude illustration, it conveys a great truth in sympathetic philosophy. All sympathetic flows have this triune condition associated with them, the same as the molecular, atomic and etheric aggregations of all forms of visible matter; the compound etheric, or dominant, being the leader and yet one of the constituents of the molecule itself. The dominant we may call the etheric portion of the molecule; the harmonic, the atomic; and the enharmonic, the molecule itself. The dominant parts of the triune combination of the sympathetic streams are the leaders, toward which all co-ordinate to make up the sympathetic terrestrial envelope of the earth; the cerebral being the high dominant, or compound etheric, the luminiferous proper."[69] (underline added)

---

[67] [Keely and His Discoveries], page 179

[68] [SYMPATHETIC INDUCTION - Snell]

[69] [Keely and His Discoveries]

7.7 Table of Modes Correspondences

Source	**From Center**[70]	**To Center**[71]	**Balancing** or **Controlling**
Keely / SVP[6]	Enharmonic	Harmonic	Dominant
Modern wave form	Longitudinal[2]	Transverse[3]	Rayleigh[10]
Angle of Incidence	0°- 45°	45°- 90°	<90°
Path/Motion[3]	Straight	Zig-zag	Circular
Keely / SVP	Dispersing	Syntropic	Dominant/Depolar
Keely / SVP	Entropic	Syntropic	Dominant /Depolar
Russell[4]	Plus	Minus	Fulcrum
Russell[4]	Gravitation	Radiation	Balance, Void
Charles Henry[7]	Gravitation	Electro-magnetic	Biopsychic
Possible Correspondence[5]	Thirds	Sixths	Ninths

Table 7.2: Table of Mode Correspondences

7.7.1 The above dynamics are to the following effects

Many	to	**One**	to	**Many**
Polar	to	Depolar	to	Polar
Unbalanced	to	Balanced	to	Unbalanced
Split	to	Whole	to	Split
Dispersed	to	Concentrated	to	Dispersed
Differentiated	to	Voided/Undifferentiated	to	Differentiated
Discord	to	Harmony	to	Discord

Table 7.3: Some Mode Effects

7.7.2 Relations of Thirds

Thirds refers to molecular subdivision.	Physical movements, classical physics and mechanics.
Sixths refers to atomic subdivision.	Sympathetic impregnation of electromagnetism in matter. Operations of thermism[72] and low light.
Ninths refers to etheric subdivision.	Sympathetic transfer from the celestial luminous or celestial radiation[73], also chemism.[74]

Table 7.4: Thirds, Sixths and Ninths

[75]

7.8 Three Modes, Currents and Streams

IX. "It is not to be understood that any one of these three modes of vibration can exist independently. Each by itself is called a 'current', and all three must be present in every 'stream' or 'flow' of force. The relations of the currents in every flow are expressible in thirds, and it is experimentally demonstrable that the relation of the three are in the order named: as $33\,\frac{1}{3} : 66\frac{2}{3} : 100$.

7.8.1 Ratios and Proportions of Thirds

The ratios and relations below are logical determinations (best guess) and are not to be taken as established or proven fact. Lots of lab work is needed here.

---

[75] see 'Chapter 9 - Modes of Vibration' for more details on structure of thirds. See also [Newton of the Mind], [Law of Cycles].

$E = 33\ \frac{1}{3}$= Dispersing, Positive, Enharmonic, Depolarizing, Entropic, Mother, Magnetic.

$S = 66\ \frac{2}{3}$= Concentrating, Negative, Harmonic, Polarizing, Syntropic, Father, Electric.

D = 100 = Etheric, Celestial, Mind, SpiritWhen defined as proportional parts:

$E = 33\ \frac{1}{3} = \frac{1}{3}$ parts of A
$S = 66\ \frac{2}{3} = \frac{2}{3}$ parts of A
D = 100 = 100 parts of A.

so if

D = 100 parts then
$S = (\frac{2}{3}$ x A) or $66\ \frac{2}{3}$ parts
and
$E = (\frac{2}{3}$ x A) or $33\ \frac{1}{3}$ parts.
(when applied to vibratory or oscillatory rates)

if

D = 100 cps

then

$S = 66\ \frac{2}{3}$ cps
$E = 33\ \frac{1}{3}$ cps

Certainly this applies to AMPLITUDE and POWER as well.[76]
These proportions most assuredly apply to “states” of compression (syntropy) and dispers

Where

E = Entropy (dispersion)
S = Syntropy (contraction)
D = Dominant (controlling or regulating)

---

[76] [Potential]

7.9 Evolution of Matter

X. “The evolution of what is called matter from the different modes of vibration is through the action of the second law, that of focalization, or negative attraction, or ‘negative affinity’.[77]

Conventionally and generally speaking this assimilating and accumulating into seeming solid matter process would be akin to the forces involved in the formation of bosons and in Bose-Einstein condensation nuclear fusion (BECNF) or boson cluster state nuclear fusion (BCSNF).[78]

7.10 Neutral Center

XI. “Where the vibrations under this mode meet, and are maintained in a state of mutual affinity or equilibrium, there is established what is called a neutral center, or, as otherwise expressed, ‘a center of Sympathetic Coincidence’.[79]

---

[77] [Syntropy], [Negentropy], (Harmonic concentration or aggregation) via [Sympathetic Affinity], [Inverse Square Law] and [Sympathetic Association]

[78] Wikipedia/Bose-Einstein condensation

[79] [KEELYS PHYSICAL PHILOSOPHY - Snell], [Neutral Center], [Center of Sympathetic Attraction]

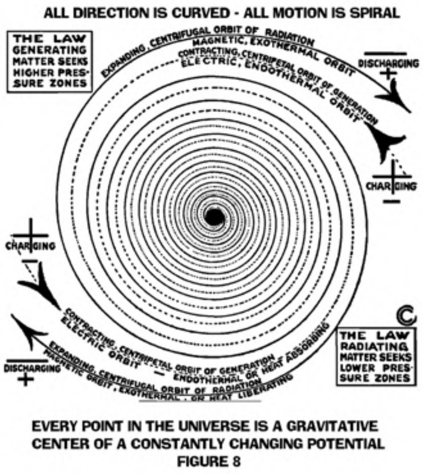
ALL DIRECTION IS CURVED - ALL MOTION IS SPIRAL
THE LAW GENERATING MATTER SEEKS HIGHER PRESSURE ZONES
EXPANDING, CENTRIFUGAL ORBIT OF RADIATION MAGNETIC, EXOTHERMAL ORBIT
CONTRACTING, CENTRIPETAL ORBIT OF GENERATION ELECTRIC, ENDOTHERMAL ORBIT
DISCHARGING
+
CHARGING
+
−
CHARGING
+
−
CONTRACTING, CENTRIPETAL ORBIT OF GENERATION ELECTRIC ORBIT − ENDOTHERMAL OR HEAT ABSORBING
EXPANDING, CENTRIFUGAL ORBIT OF RADIATION MAGNETIC ORBIT, EXOTHERMAL, OR HEAT LIBERATING
DISCHARGING
+
©
THE LAW RADIATING MATTER SEEKS LOWER PRESSURE ZONES
EVERY POINT IN THE UNIVERSE IS A GRAVITATIVE CENTER OF A CONSTANTLY CHANGING POTENTIAL
FIGURE 8

Figure 7.7: Every point in the universe is a gravitative center.
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/gravity4.jpg
(courtesy University of Science and Philosophy)

7.10.1 Center of Sympathetic Coincidence or Neutral Center

"The neutral center is that protean, uncreated, indestructible, forever-existing FIRST CAUSE[80]*. Without hands, without tools, without thought, without emotion, without love, without form, without substance, it, of itself, created all these. All that we see or can see in the objective Universe exists because of and by means of the properties and powers of the NEUTRAL CENTER."*[81]

"Neutral centers are the focalized seat of sympathetic concordance for controlling any differentiation[82] *that may exist outside, or in the mass that surrounds them."*[83]

I like to think of the power of the Neutral Center to be much like the power of a conductor wielding control over his orchestra. He sets the tone with his pitch pipe and moderates the time, coordinates activity (harmony) and amplitude of each section and individual within his domain - the orchestra.

---

[80] [First Cause]

[81] Keely, [First Cause], [Neutral Center]

[82] [differentiation]

[83] [Keely and His Discoveries], pg 255

"Neutral centers are the center of Sympathetic Coincidence[84]."[85]

7.11 Syntropy or Mutual Affinity

XII. "The terms 'neutral attraction', 'neutral affinity', 'negative attraction', or 'polar negative attraction', are employed to express the property of a mode of vibration to direct its components towards such center.[86]

These center directing forces are by definition syntropic derived from mutual affinity itself born of harmony. They are born of harmonic (sympathetic) vibrations as discussed in the Bjerknes Effect[87], Law of Harmonic Vibration, Law of Attraction and Law of Cycles. The etheric or Dominant level frequencies appear to possess a predominant proportion of syntropic (center seeking) forces which increase with velocity. Hence the centralizing power of tornadoes and hurricanes - all vortex motion.

"Taking into consideration even the introductory conditions of the etheric stage, etheric vibration has proved to me that the higher the velocity of its rotating stream the greater is its tendency towards the neutral center or center of sympathetic coincidence. Were it otherwise, how could there ever be any planetary formations or the building up of visible structures? If a billiard ball were rotated to a certain velocity, it would separate in pieces, and the pieces would fly off in a tangent, but if it were a ball of ether, the higher the velocity of rotation, the stronger would be the tendency of its corpuscles to seek its center of neutrality, and to hold together."[88] (underline added)

---

[84] [Sympathetic Coincidence]

[85] [Keely and His Discoveries], pg 220

[86] [Syntropy], [Negentropy], [Father Forces]

[87] [Bjerknes Effect]

[88] [Keely], [ATTRACTION PROPULSION ETC]

7.12 Subdivisions of Individualized Being

XIII. "As no current or flow of force can be composed of one mode of vibration only, but must always be composed of three modes uniting in varying thirds, we have 1 x 2 x 3 = 6 as the total possible forms of Sympathetic Coincidence, or, to speak in ordinary terms, there can be six; and six only, possible forms of individualized being. These are what Keely calls the six orders of atomic subdivision, or orders of vibratory motion, and he names them as follows:

I. Molecular
II. Intermolecular
III. Atomic
IV. Interatomic
V. Etheric
VI. Interetheric
VII. Compound Interetheric (added by editor)[89]

7.12.1 Table of Subdivisions of Matter and Energy

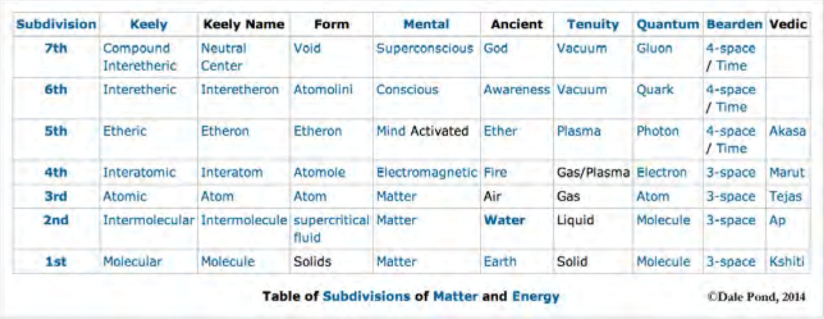

Subdivision	Keely	Keely Name	Form	Mental	Ancient	Tenuity	Quantum	Bearden	Vedic
7th	Compound Interetheric	Neutral Center	Void	Superconscious	God	Vacuum	Gluon	4-space / Time	
6th	Interetheric	Interetheron	Atomolini	Conscious	Awareness	Vacuum	Quark	4-space / Time	
5th	Etheric	Etheron	Etheron	Mind Activated	Ether	Plasma	Photon	4-space / Time	Akasa
4th	Interatomic	Interatom	Atomole	Electromagnetic	Fire	Gas/Plasma	Electron	3-space	Marut
3rd	Atomic	Atom	Atom	Matter	Air	Gas	Atom	3-space	Tejas
2nd	Intermolecular	Intermolecule	supercritical fluid	Matter	**Water**	Liquid	Molecule	3-space	Ap
1st	Molecular	Molecule	Solids	Matter	Earth	Solid	Molecule	3-space	Kshiti

Table of Subdivisions of Matter and Energy ©Dale Pond, 2014

Figure 7.8: Table of seven subdivisions or orders.
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/table-of-subdivisions.jpg

---

[89] [Compound Interetheric], [Subdivision], [Etheric Elements], [Magnetic Still Light]

7.13 Arithmetical and Geometric Progressions

XIV. "In this list forms of matter are arranged in the mathematical sequence of the rapidity of the oscillations of their constituent members; the proportion being proved by experiment to be as follows for the molecular orders:[90]

1 : 3 : 9 : 27 : 81 : 243 ...
$1 : 3^1 : 3^2 : 3^3 : 3^4 : 3^5$...

7.14 Progressions of Orders

XV. "The arithmetical progression changes in the atomic orders to a geometrical progression as follows:

3 : 9 : 81 : 6561 : 43046721...
$3^1 : 3^2 : 3^4 : 3^8 : 3^{16}$...[91]

7.15 Infinite Range of Vibration

XVI. "This same method of progression is believed to hold in all the orders of vibrations above the molecular, and soon passes into mathematical infinity.

---

[90] [Figure 8.14 - Some Basic Waveforms and their constituent Aliquot Parts]

[91] [Part 12 - Russells Locked Potentials] covering Russell's [Locked Potentials], [Square Law].

Mode	**Subdivision**	**Octave - Hertz**	**Controlling Law**
Enharmonic *Centrifugal*	Molecular Intermolecular	1st - 21st octaves 1×10^0– 2.09715×10^6	Law of Oscillating Atomic Substances[92]
Harmonic *Centripetal*	Atomic Interatomic	21st - 42nd octaves 2.09715×10^6– 4.39805×10^9	Law of Sonothermity[93]
Dominant *Controlling*	Etheric Interetheric	42nd - 63rd octaves 4.39805×10^9– 9.22337×10^9	Law of Oscillating Atoms[94]

Table 7.5: Mode Frequency Ranges

[95]

7.16 Progression Types

7.16.1 Arithmetic and Geometrical Progressions

Arithmetic Progression - In mathematics, an arithmetic progression or arithmetic sequence is a sequence of numbers such that the difference between the consecutive terms is constant.[96]
Geometric Progression - In mathematics, a geometric progression, also known as a geometric sequence, is a sequence of numbers where each term after the first is found by multiplying the previous one by a fixed, non-zero number called the common ratio.[97]
Linear Progression - In music, a linear progression (Auskomponierungszug or Zug, German: train, abbreviated: Zg.) is a passing note elaboration involving stepwise melodic motion in one direction between two harmonic tones.[98]
Melodic Progression may be explored in *"Harmonies of Tones and Colours*

---

[96] Wikipedia; Arithmetical Progression
[97] Wikipedia; Geometric Progression
[98] Wikipedia; Linear Progression

- Developed by Evolution"[99] and *"The Scientific Basis and Build of Music*" by Ramsay[100]

Music (vibrations and oscillations of a string) is linear arithmetic progression[101] while energy expansion and contraction is volumetric with geometric progression[102] as is described in Russell's Scale of Locked Potentials[103], the Square Law and Inverse Square Law.[104]

7.17 Three Major Orders

XVII. "Actually, however, all matter of which we are capable of cognition through the medium of our senses[105] is in one of three forms of aggregation:

I. Molecular
II. Atomic
III. Etheric

7.17.1 The above are not "things"

The above states constitute our sensorial world. They are not solely "things" but are also considered as states, phases, forms or dynamic structures resulting from specific combinations of the dynamic three modes of motion. Each state has a respective combination or proportion of the three modes. The higher three forms of aggregation are not perceptible to our senses and are therefore considered extrasensory, unseen, spiritual, subtle, scalar, nonclassical, *etc.*

---

[99] svpbookstore.com
[100] svpbookstore.com
[101] [arithmetical progression]
[102] [geometric progression]
[103] [Scale of Locked Potentials]
[104] [Progressive Evolution]
[105] [Sense]

7.18 Controlling Mode

XVIII. "In each of which the controlling mode of vibration is respectively:

I. The Enharmonic (entropy, dispersive) - controls Molecular
II. The Harmonic (syntropy, attractive) - controls Atomic
III. The Dominant (neutral, controlling) - controls Etheric

7.18.1 Modes combine in unique proportions

The above three modes combine in unique proportions to create orthogonal[106] stabilized centralizations of forces that have the appearance of seeming solid matter.

The following table is a supposition about the mixing of mode proportions in the 3rd and 4th columns for each aggregation. They seem reasonable but only bench testing will proven them out.

Aggregation	Governing Mode	Dominant Mode	Subservient Mode
Molecular	100 x Enharmonic	66 $\frac{2}{3}$ Harmonic	33 $\frac{1}{3}$ Dominant
Atomic	100 x Harmonic	66 $\frac{2}{3}$ Enharmonic	33 $\frac{1}{3}$ Dominant
Etheric	100 x Dominant	66 $\frac{2}{3}$ Harmonic	33 $\frac{1}{3}$ Enharmonic
Thirds	3	2	1
Proportion	9	6	3
Portion	Ninths	Sixths	Thirds[107]

Table 7.6: Hypothetical controlling Modes and Proportions

7.19 Modes as actual components

XIX. "But it must be understood that each of these modes is a Positive and real constituent of every atom and molecule.

---

[106] ideally speaking as is demonstrated in carbon [crystallization].

Considering an atom or molecule as a balanced dynamic of syntropic and entropic forces they must then by definition have these three component modes as essential to their existence. This concept is not unlike Buckminster's Tensegrity[108] idea.

7.20 Foundational Concepts

XX. "It will be seen that as every form of material aggregation is to be considered as a "neutral center of attraction", where the vibratory force of all three orders are held in Sympathetic Coincidence, that is, in balanced activity or harmonized motion, and not by any means cancelled or mutually destroyed, there is no diminution of force, but only temporary suspension of its radiating or propulsive activity or expression.[109] (underline added)

7.20.1 Balance or Equation of Father-Mother Forces

This balance or equation of Father-Mother Forces[110] is not a perfect balance where there is no motion. There is in all seeming solid and still matter a near imperceptibly tiny periodic motions Russell referred to as "Rhythmic Balanced Interchange". Macvicar called this balanced periodic motion the cosmical Law of Assimilation[111]. The range of motion of this periodic motion when considered as stable does not exceed 50% of the molecular dimension.[112] Conventional science refers to this natural residual vibration as vacuum fluctuations, quantum jiggle or Zero-Point Energy. [see **XXV** below]

"If access to the zero-point-energy (ZPE) reservoir is successful, one needn't worry about either depletion of this resource or creating an imbalance in the local environment. It is the electromagnetic equivalent of scooping cupfuls of water out of the ocean, with replacement occurring at the velocity of propagation of electromagnetic waves, the velocity of light. As to the ultimate

---

[108] Wikipedia/Tensegrity

[109] [Neutralization]

[110] [Father-Mother Principle]

[111] [14.32 - Law of Assimilation]

[112] [AQUEOUS DISINTEGRATION - Snell]

origin of the ZPE, two views are discussed in the physics literature: one, that it is simply part and parcel of the energetic legacy that emerged with the Big Bang, and another that it is an energetic substratum the preceded even the Big Bang, with our universe emerging as the result of a giant vacuum fluctuation.[113] *In any case an argument can be made that it is sustained by a cosmological feedback cycle in which charged particles radiate due to their "quantum jiggle," and the particles "jiggle" due to being caught up in the collective radiation of all the other particles, an electromagnetic equivalent of placing a microphone near a speaker and generating a squeal."*[114]

7.21 Latent Force

XXI "This is the foundation of Keely's doctrine of 'latent force'[115], and of the indefinite power which can be obtained by breaking up the harmonious balance or equation of forces of every mode, which exists in every neutral center, that is to say in every mass of matter."

7.21.1 Matter is the result of three forces or modes of vibration

In SVP every form of matter is the result of three forces or modes of vibration that are balanced or equated to a determinable degree. Disturbing this balance or equation of forces liberates them - not the breaking of a solid appearing atom. Increasing syntropic force, for instance, will instantly evoke the Law of Cycles thereby causing transformation into intense dispersion (emission, radiation). Liberated entropy will instantly disperse. An example of this process of liberation is the well known atomic and nuclear detonations as seen in cavitation or implosion processes.

The use of these principles in atomic weapons is an extreme liberation of the contained latent force within a very short time span. The designers' intent of such weapons is to get the biggest bang possible. In the real world we have no need of such gigantic liberations of energy. It is possible and desirable to perform smaller releases over longer periods of time such as is done with dissociation of water via electrolysis and similar means. In this case it would be prudent to use materials containing less latent energy on the molecular level and perhaps the atomic order. These releases of energy would be more than sufficient for most any application serving humanity's needs. Such liberations would be saner, simpler, safer, non-ionizing and sustainable.

---

[113] A third view is this persistent periodic motion is that which maintains the physical universe.

[114] H. E. Puthoff, "On the Source of Vacuum Electromagnetic Zero-Point Energy," Phys. Rev. A, vol. 40, p. 4857,1989; Phys. Rev. A vol. 44, p. 3382,1991.

[115] [Latent Force]

Keely was able to release these forces in a gentle or violent manner depending on the stage of the Law of Cycles progression.

"Sudden unition induces the same effect; but demonstration shows that the concentration of this triple force is as free of percussion as is the breath of an infant against the atmosphere; for the three currents flow together as in one stream, in the mildest sympathetic way, while their discharge after concentration is, in comparison to their accumulation, as the tornado's force to the waft of the butterfly's wing."[116]

Water molecules (Keely's favorite medium for creating his all powerful and useful 'etheric vapor') can be easily dissociated in the molecular order with an electric current (AC or DC), X-rays, cavitation[117], ultrasound[118], microwaves[119], radio waves[120] and even vacuum. This molecular dissociation liberates a gas composed of the atoms of Hydrogen and Oxygen. These as gases constitute the next higher level or order of being - the atomic level of being. This atomic level can be dissociated with an electric spark[121] 'igniting' or rendering (dissociating) the atomic order into the next 'etheric' level or order being plasma; *i.e.*, flame made of ionized gases possessing great pressure as discordant radiating entropy mode preponderates. The process is more complicated than the above but you get the essential points. As humanity masters these low levels of dissociation the higher levels liberating greater quantities of useful energy may be accessed.

---

[116] [Keelys Contributions to Science]

[117] [15.12 - Dissociating Water with Acoustic Cavitation]

[118] [15.09 - Dissociating Water with Ultrasonic Vibration - Puharich]

[119] [Dissociating Water with Microwave]

[120] [15.08 - Dissociating Water with X-Rays - Radiolysis]

[121] But then what is a spark if not the electric stream preponderant with the individualizing (syntropic) mode? see [Spark], [Introductory Impulse]

It was this etheric vapor under immense pressure derived from dissociated water that Keely had intended to use to power his original Keely Motor. After designing and building over 120 models and 30 years of effort he gave up on that approach when he discovered the etheric vapor is not the *source* of power but instead is a *carrier*[122] of the force he wished to liberate.

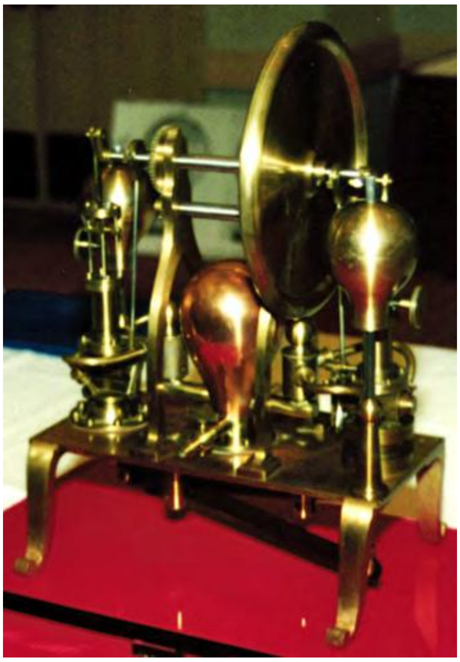

Figure 7.9: Keely Motor activated by slow released etheric vapor
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/keely-motor.jpg

There was one instance where Keely did use this etheric vapor under immense pressure and released or liberated quickly was in his etheric cannon[123]. He demonstrated this cannon on many occasions and once before the US Navy at Sandy Hook Naval Yard, New Jersey. The Navy turned it down because they thought the mechanism “too complex”.

---

[122] [connecting link]
[123] [cannon]

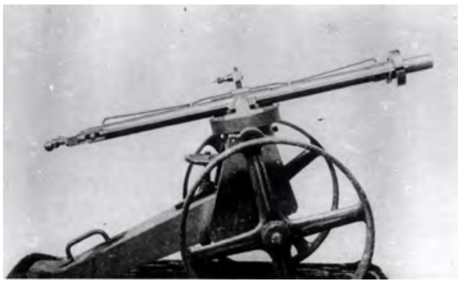

Figure 7.10: Keely's cannon fired by quick release of etheric vapor
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/etheric-cannon.jpg

7.21.2 Latent Force, Keely

"With our present [1880s] knowledge," writes Mr. Keely, *"no definition can be given of this latent force; which, possessing all the conditions of attraction and repulsion associated with it, is free of magnetism. If it is a condition of electricity, robbed of all electrical phenomena, or a magnetic force, rebellant [sic] to the phenomena associated with magnetic development, the only philosophical conclusion I can arrive at is that this indefinable element is the soul of matter*[124]*. Were not every form of matter, even to the cerebral convolutions of the brain, impregnated with this latent element of force, which is sympathetically subservient to celestial radiation, nature would be like a still-born child, or a marble statue - dead to the sympathetic association that induces motion. Matter could not exist without this element, this spiritual essence, this impregnation from the Deity, which is its Soul, any more than a man with an ossified brain could have motion or life."*[125]

This new higher state liberated energy, not having any detectible magnetism or electricity, must by definition be a higher subtler subdivision than the atomic or etheric orders of being both having these two attributes. These new subdivisions then would be the sub-quantum levels of light (photonic) and finer quantum divisions such as quarks, gluons, *etc.*

---

[124] [Soul of Matter]

[125] [What Electricity Is - Bloomfield Moore]

7.22 Sympathy and Resonance

XXII. "Insomuch as every mass of matter consists thus, in fact, of vibrations in harmonic equilibrium, related by simple proportions of thirds, it follows that every mass of every description stands in harmonic relation to every other mass. This is, in part, what is meant by the sympathy of all forms of matter and of motion (*i.e.*, entanglement, bonding, coupling or love); and it is through the study of the methods of increasing or diminishing[126] this sympathy that we reach practical results in this field of research. At present this is best accomplished by resonance[127]; that is, through the harmonic vibrations created by musical instruments (or signal generators), bringing out the acoustic world as the microscope reveals the hidden visual world.

7.22.1 Attraction and a Repulsion between all things

There exists, by virtue and attributes of harmonic vibrations or sympathy (acoustic entanglement), an attraction and a repulsion between all things. The nature of this attraction and repulsion is discussed in 'Chapter 2 - Love in Science and Physics'. This sympathetic association can be modified at will according to the Law of Sympathetic Association.[128] Keely referred to this process as 'Disturbance of Equilibrium'[129] where he disrupted the harmonic equilibrium thus exciting the dynamics of balanced systems. Centralized forces in a state of coincidental harmony contain Latent Force[130]. The quantity of that Latent Force is determined by the subdivision or order of the material. This force is liberated into an useable force by an outside acoustic disturbance upsetting that balance. This would be not unlike a frequency and amplitude modulation (FM and AM) of existing chordal (complex frequency) patterns.

---

[126] [Disturbance of Equilibrium]

[127] established with sympathetic vibration or oscillation, see [Law of Force]

[128] [Bjerknes Effect], [Love in Science and Physics], [Three Laws of Being], [Law of Assimilation]

[129] [Disturbance of Equilibrium]

[130] [Latent Force]

7.23 Matter composed of Molecules

XXIII. "Every visible or tangible mass of matter must be regarded as an aggregation of molecules; the molecules being the true centers of the equated forces of "neutralized attraction".

'Neutralized attraction' is where the two forces of syntropy and entropy are equated to a balanced state or condition. A stable molecule has such a center of neutralization of forces and every molecule, according to Keely, whether singularly or in a larger mass have this distinction. Disturbance of one molecule of such a mass is disturbing all molecules of the mass. Disturbance is accomplished with an 'exciter'[131] being an acoustic pulse[132] possessing certain characteristics such that it is received and re-modulates the molecule into a discordant condition.

7.23.1 Attraction is an effect of Syntropy

We now know that attraction is an effect of syntropy which is an effect of sympathetic association or harmonized frequencies or unisons. Syntropy unopposed is pure active attraction or affinity. "Neutralized Attraction" then would be syntropy balanced by entropy when the rates of vibration or oscillation form an unison with the fundamental or harmonics.[133]

Regarding attraction or bonding as an effect of harmonized vibrations and oscillations is decidedly different than force carrier particles. The first condition is a dynamic condition of the forces involved in the interaction. The second is an isolationist viewpoint requiring an unnecessary additional player on stage. Although a unit measure of attractive force - a quantum (least measure or quantity such as a gluon) of attractive force - could be assigned to these entities for mathematical purposes which I feel is how all this quantum business got started in the first place.

---

[131] [Exciter]

[132] or any other frequency within the three main modes

[133] [Law of Attraction], [Bjerknes Effect], [Law of Sympathetic Oscillation], [Law of Force], [Law of Chemical Affinity] and [Voiding]

7.24 Molecules are sensitive to modes

XXIV. "These molecules have been experimentally proved by Keely to be formed of all three modes of vibration; the proof being that they respond to all three modes when subjected to the tests of compound concordant impulses (chords).[134]

An example of a 'compound concordant impulse' would be a musical triplet of say C, E and G sounded together with a sharp attack and little decay within its signature envelope. Logic indicates at least one of those frequencies is concordant with natural tones (eigenfrequencies) of the target molecule. In which case power from the exciter chord is added (sympathetic and syntropic addition) to the molecule thus offsetting its original concordant equation of forces or balance.

7.25 Oscillations of Molecules

XXV. "When in that state of neutral aggregation which we know as matter[11]**, each molecule is in perpetual oscillation, the range of the oscillation being one-third of the molecule, and its rapidity 20,000 oscillations in a second.**[135]

This has always been a curious statement - that all molecules oscillate about their center at a fixed rate of 20,000 Hz. Why hasn't this been found by conventional science? Perhaps it has but I've not found any reference to it (yet). In today's world there are countless types of molecules and many of them are artificial such as are found in plastics and new metal alloys. Does this statement apply to all of them? One would presume - to err on the safe side - that Keely is referring to naturally occurring molecules. That would be a big and perhaps dangerous assumption...[136] Nevertheless it is something to explore. The effort may be well worth it.

---

[134] see XXVI, [Law of Force], [Law of Chemical Affinity], [15.18 - Keely's Process for Liberating Ether from Water]

[135] [Molecular Dissociation]

[136] see section '6.5 Frequency' where frequency of oscillation is twice the ordinary count.

7.26 Disturbance of Equilibrium

XXVI. "It is through the disturbance of this oscillatory equilibrium, by means of resonant impulses, that Keely alters the relations of the vibratory impulses which constitute matter[11]. This he does by striking the same chord in three octaves, representing the third, sixth, and ninth of the scale.[137]

The above quote is the beginning of the description of how Keely modulated the eigenfrequency or chord of the mass[138] being acted upon. The following sections give insight into how he approached and did this using the thirds, sixths and ninths of a scale constructed upon the tonic or fundamental of the mass being acted upon using the quantization method[139]. Here lies perhaps one of Keely's greatest secrets - the application of the much verbalized "3, 6 and 9" so often attributed to Tesla[140]. Somewhere I read Tesla stated he had figured our Keely's secret. This may be the secret he referred to. I do not have that quote nor a link to the original material.

In music theory intervals are expressed as ratios. In standard music theory there are Major and Minor Thirds and Keely does not specify which. Normally a Major Third is considered as the quantity 5:4. Translated this quantity is the fundamental frequency divided by 4 which quotient is then multiplied by 5 gives the frequency of a Major Third. A Minor Third is considered as the quantity 6:5. Translated this quantity is the fundamental frequency divided by 5 which quotient is then multiplied by 6. The same treatment applies to all music intervals.

These intervals form a three tone chord (musical triplet) when sounded simultaneously. Multiply each of these frequencies by two and four creates Keely's Triple Triplet[141] of *'same chord in three octaves'*.

---

[137] [Triple Triplet], [Triple Triplet Flows], [Inversion]

[138] [chord of the mass]

[139] [Quantization - SVP]

[140] [14.35 - Teslas 3 6 and 9]

[141] [Triple Triplet Flows], [Triple Triplet]

Interval	Ratio
Minor Third	6:5
Major Third	5:4
Minor Sixth	8:5
Major Sixth	5:3
Augmented Sixth	125:72
Minor Ninth	32:15
Major Ninth	9:4

Table 7.7: Thirds, Sixths and Ninths Intervals

7.26.1 Disturbance of Equilibrium, Gravity

"Disturbance of equilibrium, like gravity, is inherent, an eternal existing force."[142]

It is possible these two phases, gravity and disturbance of equilibrium, are simply the two opposite poles of matter, the physical Universe, the positive and negative poles of matter, controlled by the third or controlling principle, the Universal Will.

"Unstable equilibrium, like gravity, is a condition born in each neutral center at the time of birth, and thus designed by the Architect as the connective link[143] *between the "dispersing positive"*[144] *and the "attractive negative."*[145] *The action inducing this "link" I call "sympathetic planetary oscillation".*

"The enharmonic portion of the electric current carries the power of propulsion that induces disturbance of negative equilibrium.

"In his (Keely's) experiments of aqueous disintegration[146] *disturbance of equi-*

---

[142] [13.12.1 - Disturbance of Equilibrium]

[143] [Connecting Link]

[144] [Radiation]

[145] [Contraction]

[146] [Aqueous Disintegration]

librium was brought about by using mediums of differential gravity in the disintegrator[147]*, air as one, water as the other.*"[148]

Forces are held in a dynamic state of equilibrium, a neutral state, until disturbed by an excitation or stimulation. Adding either harmonic (syntropy) or enharmonic (entropy) excitations will evolve into opposite conditions either attraction or repulsion. This dynamic holds true on all subdivisions. On the lower subdivisions these disturbances must be sympathetic to that being stimulated which is to say acoustic waves or pulses. On the higher etheric subdivisions a sympathetic excitation would be the "infinite exciter" as thought or a form of thought such as desire, intent, etc.

Matter is not directly created by these incoming sympathetic streams. The incoming etheric (celestial) streams impact (bombard) a Neutral Center which differentiates them and matter results from the effects of this impact or antagonism developing summation and difference modulations (tones) during its progressive evolution into being. The incoming excitations need not be all frequencies as countless frequencies may develop during the impact or antagonism. Any tone instantly generates an overtone series. These secondary and tertiary partials or tones spring into existence the instant any tone is sounded. These many tones are effects of the one causative tone and cannot be separated from their cause. They are entangled everywhere experiencing summation and difference modulations (tones) originating from everywhere. Each experience is of course unique because of local influences or modulations.

7.27 Sixths induces Neutral Radiation

XXVII. "Of these, the sixth reduces the range of molecular oscillation, and thus tends to give greater tenuity to the mass. It induces "trajectile velocity" from neutral centers, or "neutral radiation".

---

[147] [disintegrator]

[148] [Snell Manuscript]

7.27.1 Seeming Paradox

The first part of the first sentence of XVII appears to contradict the second part of the same sentence thus creating a seeming paradox. However, the two sentences taken together illustrate quite clearly the function and dynamics of the Law of Cycles, said dynamics elsewhere discussed.

"Experiment shows that molecular dissociation does not take place until the molecule attains an oscillation approaching, if not fully reaching two-thirds of its diameter. This can be effected by means of the action of the enharmonic or "radiating" current applied to the mass, after its molecules have been disturbed by an "introductory impulse"; that is, by the musical note[149] *above mentioned.*

7.27.2 Exciting the Molecular Mass

First excite the molecular mass with a strong pulse, then continue this disturbance at resonance with its natural eigenfrequency (chord of mass) until a certain amplitude of oscillation occurs where dissociation takes place. These quantities of applied excitation (amplitude, duration, attack, sustain and decay) could be determined through experiment.

7.28 Third is the Dominant

XXIX. "The third represents the dominant, and when brought under control of a harmonic resonant impulse induces a complete rearrangement of the modes of vibration and oscillation; in other words, will transform the mass either into its component initial forces, or into some other form of matter.

7.28.1 Harmonic Resonant Impulse

"harmonic resonant impulse" means an impulse that is harmonic (in tune with; *i.e.*, sympathetic) not unlike a singer singing the correct pitch that is in synch or establishes a chord of resonance[150] with that of a wine glass to

---

[149] a musical note is a specific pitch in mathematical relationship with other pitches.

[150] [Chord of Resonance]

be shattered. Also called *"resonant impulse"*.[151]

Of course Brinton/Keely is talking about none other than transmuting the mass being experimented upon whether that be water or some other molecular substance - *'transform the mass either into its component initial forces, or into some other form of matter'*. 'Component initial forces' means liberating these awesome atomic forces whether slowly or instantly. Of course the slower the liberation the better.

> *"Matter is bound up energy and energy is liberated matter."* Keely, 1893

It is astonishing these ideas were first published in 1894 well before orthodox science acknowledged splitting the molecule or atom was even possible. Not only does Keely say this is possible but he says how to do it and that he had been doing it for some years already prior to 1894. Realizing the above it is obvious why no one in Keely's day had any idea what he was doing. In Keely's 'Ultimate Constitution of Matter and Action of Force Regulating its Phenomena'[152] he lays out the basic concepts of matter, force and their interactions such as converting one into the other.

7.29 The Dominant or Etheric Mode

XXX. "It is the study of the dominant[153] to which Keely has devoted his recent researches. He aims to control the power he evolves by altering the dominant or etheric mode of vibration in the triplicate flows of force.

7.29.1 Etheric Current

The etheric (Dominant) current is the controlling medium in the triune flows.[154] It is the switch or regulator that actuates the power of the other two currents or forces thus disturbing the triune balance. The key phrase in the above is '*control the power he evolves*' indicating he doesn't want an all at once release as in a destructive explosion.

---

[151] [Law of Transformation of Forces], [Transmutation]

[152] [[7.9 - Ultimate Constitution of Matter and Action of Force Regulating its Phenomena.]

[153] [Dominant]

[154] Keely's [Three Modes of Vibration] for range of frequencies of the etheric current, One Universal Controlling Medium

7.30 Centers of Harmonized Vibration

XXXI. "As all molecules and masses are mere centers of harmonized vibrations,[9] temporarily held in suspension by simple Laws[155] identical with those of resonance, it follows that these centers can be broken up or divided by certain orders of vibration impinging upon and disturbing them.[156]

See the above sections and subsections describing this equilibrium and its disturbance. The idea that the control of the forces of a molecule can be addressed via the molecule's center of balanced forces represents a major diversion from conventional thought. It is far easier to modulate this center of balanced forces than any other condition of motion of a molecule be it electrons or atoms forming part of the molecule. The point or center of balance of forces is not unlike that point in a child's swing when the highest or furthest point of motion is reached and all motion stops for an instant then reverses direction. This point coincides with Russell's Zero Plane of Inertia[157]. This instant of depolar non-motion is where the kinetic out swing and in swing are neutralized (conjugated, sum to zero) and energy of the system becomes latent (potential). At this point it is easiest to add additional outside power to maintain or increase motion. It is the optimum point of modulation. Attempting to add power at other points of the swing are either impossible or difficult.

7.31 Sympathetic Outreach (sympathy/resonance field effect)

XXXII. "It is a familiar fact that a cord in vibration tends to produce a similar vibration in a cord placed near it.[158] This property belongs to all vibrations, whether resonant or not, and they exert it in proportion to the order to which they belong. The distance in Space to which this power extends, or can be extended, is what

---

[155] [Keelys forty laws]

[156] [Disturbance of Equilibrium], [DISTURBANCE OF EQUILIBRIUM - Snell], [13.12.1 - Disturbance of Equilibrium], [7.1 - Matter is centralized motion]

[157] [Plane of Inertia]

[158] [Sympathetic Oscillation], [Tuning Fork]

is called 'the sympathetic outreach'[159] of the current or flow."

Sometimes the effects of this property are referred to as "entrainment"[160] (sometimes as "coupling"[161]) as in the case where multiple oscillating mechanisms will synchronize their motion in and over time. The linking between mechanisms in sympathetic oscillation[162] is referred to as the "connecting link"[163] or the unseen etheric communication mechanism spanning space and time. This connecting link is the purveyor of so-called "spooky action at a distance"[164]. These sympathetic properties can be quantified and therefore engineered. They have for instance an "outreach" or 'field effect' that can likewise be quantified and engineered. This outreach or field effect is an extension of sympathy from the mass and affects other masses within reach of its field effect depending on the degree of sympathetic vibration or oscillation of other nearby masses. Degrees of sympathetic association can be determined by the degree of harmony or discord in their respective chord of mass (vibration or oscillation signature). This property is governed by the Law of Sympathetic Oscillation[165].

7.32 Etheric Outreach

XXXIII. "In this manner we have 'sympathetic negative attraction', and 'sympathetic positive propulsion', with reference to the outreach of the third or dominant current of the stream, which is allied to the order of etheric vibrations."[166]

The effect of this field or outreach can be either attractive or repulsive on nearby objects according to the degree the creative and receptive chords are harmonic or enharmonic with each other.[167] The outreach or field effect of the etheric current is as mentioned above the connecting link between masses.

---

[159] [Sympathetic Outreach]
[160] [Entrainment]
[161] [Coupling]
[162] [sympathetic oscillation]
[163] [connecting link]
[164] [action at a distance]
[165] [Law of Sympathetic Oscillation]
[166] [Sympathetic Transmission], [Subdivision]
[167] [Law of Force]

This condition of synchronous motions is described in the

7.32.1 Law of Transmissive Vibraic Energy

"All oscillating and vibrating coherent aggregates create, in the media in which they are immersed, outwardly propagated concentric waves of alternate condensation and rarefaction, having a period-frequency identical with the pitch of the aggregate."
Scholium: *"All forms of transmissive energy*[168] *can be focussed, reflected, refracted, diffracted, transformed, and diminished in intensity inversely as the square of the distance from the originating source."*

These "outwardly propagating waves" can be in harmony or discord with other masses creating sympathetic resonance. If the waves from the masses are in tune or harmonic they will be attracted to one another. If they are discords the masses will be repelled. These tendencies and motions of attraction are governed by the laws of attraction[169] and repulsion[170]. If these waves are in harmony the masses are then 'coupled' sympathetically and what occurs to one occurs to the other. They become *as one.*

7.32.2 Etheric vibration is the cause of sympathetic association

Etheric (quantum) vibration is the cause of Sympathetic Association, Sympathetic Vibration and Sympathetic Oscillation also known as Entanglement, Coupling, Quantum Entanglement, *etc.*. This etheric vibratory force[171] can be either syntropy (sympathetic negative attraction) or entropic (sympathetic positive propulsive). This phenomena may also be related to Russell's lens[172] idea.

---

[168] [Law of Force]
[169] [Law of Attraction]
[170] [Law of Repulsion]
[171] [Etheric Vibration. - The Key Force]
[172] [Figure 3.4 - Focalizing Lenses at nested Cube faces]

7.33 Chord of Mass

XXXIV. "Each molecule of a given mass of matter represents the same harmonic chord or note in its oscillatory motion. The chord of the mass is, therefore, the chord of every molecule of [the] mass."

7.33.1 Qualifiers

The above sentence (XXXIV) contains several qualifiers:

1 - *"Each molecule of a given mass of matter"* - each and every molecule of a specific mass. Therefore a given mass has a single chord being the same as each molecule participating in that mass.

2 - *"same harmonic chord or note in its oscillatory motion"* - to the degree of syntropy the molecule/mass is (*i.e.*, represented by a musical note or pitch) in the mass oscillatory (translation) motion, NOT it's internal vibration.

3 - The chord of mass[173] is of course the vibration or oscillation signature, a collection of the fundamental and all partials resulting from that fundamental as per the Law of Harmonic Pitch.

7.33.2 Law of Harmonic Pitch

"Any aggregate in a state of vibration develops in addition to its fundamental pitch a series of vibration in symmetrical sub-multiple portions of itself, bearing ratios of one, two, three, or more times its fundamental pitch."[174]

7.34 Controlling Chord of Mass

XXXV. "But as the condition of absolutely stable equilibrium is theoretical only, and does not exist in nature, the chord of the mass is constantly changing. Yet we must learn to control this chord of the mass by resonant induction[175], if we would gain command of the molecular forces."

---

[173] [chord of mass]

[174] Keely, 1894

[175] [resonant induction]

7.34.1 Equilibrium or Balance cannot exist

The states or conditions of equilibrium or balance cannot exist in the material world of seen effects as every thing or force is in perpetual transformation because of the ever present Law of Cycles and the Law of Transformation of Forces. Everything is always in motion as motion is Life activity a resultant of dynamic polarity of force and energy in their perpetual and persistent efforts to void.

7.34.2 Law of Cycles

"Coherent aggregates harmonically united constitute centers of vibration bearing relation to the fundamental pitch not multiples of the harmonic pitch, and the production of secondary unions between themselves generate pitches that are discords, either in their unisons, or overtones with the original pitch; from harmony is generated discord, the inevitable cause of perpetual transformation."[176]

7.34.3 Table of Transformation over Time

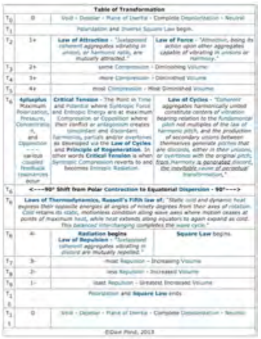

Figure 7.11: Table of Transformation of Forces over Time
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/Table-of-Transformation300.jpg

---

[176] Keely, 1894

7.34.4 Law of Transformation of Forces

"All forces are different forms of Universal Energy unlike in their period-frequency, merging into each other by imperceptible increments; each form representing the compass of 21 octaves. Each form or pitch may be transformed into an equivalent quantity of another pitch above or below it in the scale of 105 octaves. The transformation can occur only through its static effect, developing vibrations of harmonic pitches above or below their fundamental vibration, or developing with juxtaposed aggregates, resultant and difference, or third order, as the case may be.
Scholium: *"A table of the intervals and harmonics of the normal harmonic scale will indicate the ratios in which the transformation of forces will occur."*[177]

7.35 Harmonizing Chord of Mass

XXXVI. "Keely believes he has solved this problem, by the invention of a mechanical device which brings the chords of all masses within the conditions of a few simple acoustic tests."[178]

It is not known what this mechanical device is, specifically. It is presumed by this writer this device is the Trexar[179] and similar wires Keely made with specific proportions of specific metals. These wires appear to have the property of superconductivity at room temperature. See section 7.37 below.

7.36 Molecular Oscillation

XXXVII. "The range of molecular oscillation is affected differently in different substances when submitted to the same vibratory impulse, and these ranges can be measured."

Again, more careful bench work is called for to determine these effects or modulations.

---

[177] Keely, 1893

[178] [Chord of Mass], [Keelys Mechanical Inventions and Instruments], [MECHANICAL INVENTIONS AND INSTRUMENTS]

[179] [Trexar], [Trexnonar]

7.37 Transmitting Concordant Impulses

XXXVIII. "In the three metals, silver, gold, and platina, we obtain the proportions - 3 : 6 : 9 - As this is the primary relation of the modes of vibration, a wire[8] made of these three metals is peculiarly adapted to transmit concordant impulses; and nodes made of these substances placed upon a wire, transmitting resonant vibrations, indicate, by the different orders of vibration induced in them, the rate of oscillations of the atomic constituents."[180]

The first part of the above concerning transmission of concordant impulses is determined by the

7.37.1 Law of Vibrating Atomolic Substances

"Atoms are capable of vibrating within themselves at a pitch inversely as the Dyne (the local coefficient of Gravity), and as the Atomic volume, directly as the Atomic weight, producing the creative force (Electricity), whose transmissive force is propagated through atomolic solids, liquids, and gases, producing induction and the static effect of magnetism upon other atoms of attraction or repulsion, according to the Law of Harmonic Attraction and Repulsion."
Scholium: *"The phenomenon of Dynamic Electricity through a metallic conductor and of induction are identical. In a metallic conductor, the transmission is from atom to atom, through homologous interstices, filled with ether, presenting small areas in close proximity. In crystalline structures, heat, which expands the atoms, by twisting them produces striae, increases the resistance, etc. Between parallel wires and through air the induction takes place from large areas through a rarefied medium composed of a mixture of substances, whose atoms are separated by waves of repulsion of various pitches, discordant to electric vibrations; the said atoms sympathetically absorb the vibrations and dissipate from themselves, as centers, concentric waves of electric energy which produces heat and gravism[181]."*[182]

The second part of the above concerns measuring frequencies. The exact method of this detecting process is not known. Below are some frequencies apparently detected using this or a similar method.

---

[180] [Trexar], [Law of Force], [Law of Chemical Affinity], [Disturbance of Equilibrium]

[181] [gravism]

[182] Keely, 1894]

7.37.2 Frequencies of the Modes

"He [Keely] mentions the following frequencies, measured in sound colors[183]:

Molecular frequency

inducted in homexar[184] by harmonic thirds, 100,000,000, [100 MHz]

that inducted by enharmonic sixths, 300,000,000, [300 MHz]

that induced by diatonic ninths, 900,000,000, [900 MHz]

that inducted by dominant etheric sixths, 8,100,000,000, [8.1 GHz (microwave)]

that inducted by interetheric ninths, 24,300,000,000. [24.3 GHz (microwave)]

In the trexar[185] the ''*compounding of the triple triple*[186] *will give a frequency from the ninth node that, set down, would make a string of figures a mile long.*"[187]

7.38 Vibratory Rotation

XXXIX. "The phenomenon of rotation arises from the harmonic interaction of the dominant and enharmonic elements of the flow; in other words, the first and third, the third and ninth, *etc.*; those whose vibrations bear the proportions to each other 33 $\frac{1}{3}$: 100."[188][189]

7.38.1 Microwaved Radiometer

According to the listing of frequencies above (7.37.2) microwaves were part of Keely's vibratory tool box. Today we have ready access to these frequencies.

---

[183] [music note or sound colors]

[184] [Homexar]

[185] [Trexar]

[186] [Triple Triplet]

[187] from [VIBRATIONAL FREQUENCIES]

[188] see '7.7.2 Relations of Thirds'

[189] see 'Chapter 10 - Proportions (Ratios) of Relative Values'

For instance the effects of these waves can be readily seen by placing an ordinary radiometer or even a typical incandescent light bulb into a typical high power kitchen microwave oven. The radiometer vanes in such an environment and under microwave excitation spin too rapidly to see. The incandescent bulb flashes brilliantly white light. The effects are quite dramatic when these 'dominant' waves interact with the 'enharmonic' or molecular/atomic elements within these bulbs.

Experiment: A microwaved radiometer (with sufficient power) rotates at high velocity, flashes brilliantly and gets very hot very quickly. 'Enharmonic' or molecular/atomic elements are Argon gas, metal of vanes, black and white coatings. Each of these has their own vibration frequency and figure into the process.

7.39 Rotation

7.39.1 Medium for inducing a Neutral Center

"In the conception of any machine heretofore constructed, the medium for inducing a neutral center has never been found. If it had, the difficulties of perpetual motion seekers would have ended, and this problem would have become an established and operating fact. It would only require an introductory impulse[190] *of a few pounds, on such a device, to cause it to run for centuries. In the conception of my vibratory engine*[191]*, I did not seek to attain perpetual motion*[192]*; but a circuit is formed that actually has a neutral center, which is in a condition to be vivified by my vibratory ether*[193]*, and while under operation, by said substance, is really a machine that is virtually independent of the mass (or globe), and it is the wonderful velocity of the vibratory circuit*[194] *which makes it so. Still, with all its perfection, it requires to be fed with the vibratory ether*[195] *to make it an independent motor. . . ."*[196]

---

[190] [introductory impulse]
[191] [Compound Vibratory Engine]
[192] [perpetual motion]
[193] [etheric vapor]
[194] [The Operation of the Vibratory Circuit]
[195] [Vibratory Ether]
[196] Keely, [Keely and His Discoveries], [More Science]

7.39.2 Keely on Rotation

"Rotation arises from the harmonic interaction of the dominant and enharmonic modes of vibration, which are in the ratio of 3:9." [see **XXXVIII**]

"Harmonic interaction"; where the frequencies of the dominant and enharmonic are harmonized or made coincident.[197]

Keely is here referring to the dynamics operating in and through his globe motor[198] or Dynasphere[199] which was his experimental base or model to demonstrate and prove out his theories on vibratory rotation and power development. It was in this design he first began to use his etheric vapor as a carrier[200] of these higher rotating vibratory forces herein explored.

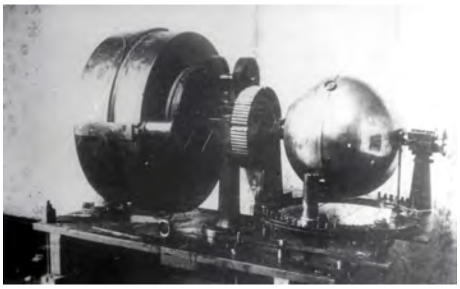

Figure 7.12: Globe motor actuated by rotary vibratory dynamics
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/keely-globe-motor.jpg

---

[197] [Figure 9.4 - Radiation and Absorption interactions with Neutral Center]

[198] [globe motor]

[199] [Dynasphere]

[200] [connecting link]

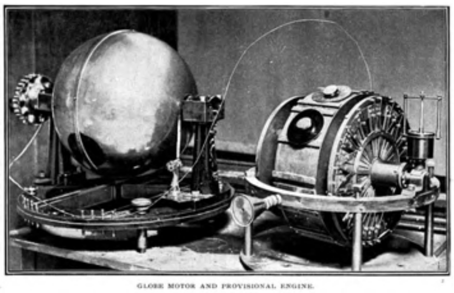
GLOBE MOTOR AND PROVISIONAL ENGINE.

Figure 7.13: Keely's globe motor and provisional engine
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/globe-motor.jpg

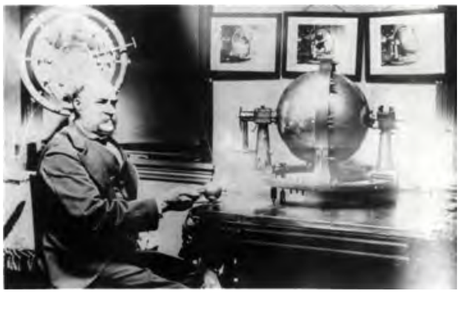

Figure 7.14: Keely and his globe motor or dynasphere
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/keelyn.jpg

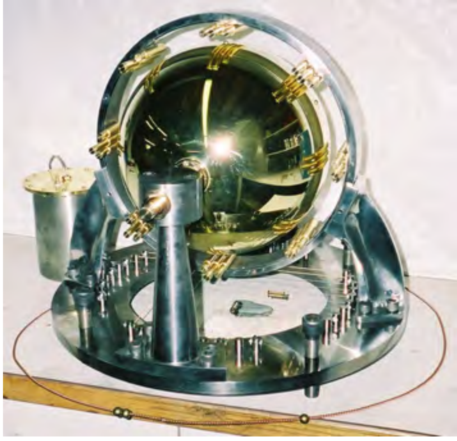

Figure 7.15: Pond's replication of the globe motor or dynasphere
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/altea.jpg

7.39.3 Power of Rotation

"Power of rotation comes on the positive power of negation, arising when the receptive flows become independent of the circular chord of resonation[201] *(set up mechanically or otherwise) breaks up the rotational power. Rotation is caused by the receptive concussion of the positive and negative forces as they come together at the neutral center and as each seeks its "coincident" by harmonic waves, flows or streams."*[202]

I - Two concepts here:

A: Power of Rotation (torque)

"Power of rotation comes on the positive power of negation, arising when the receptive flows become independent of the circular chord of resonation (set up mechanically or otherwise) breaks up the rotational power."

B: Rotation itself. (rotary motion)

---

[201] [Figure 20.13 - Two Rings - Outer and Inner]

[202] [Keely; [Snell Manuscript], PHILOSOPHY OF TRANSMISSION AND ROTATION OF MUSICAL SPHERE]

"Rotation is caused by the receptive concussion of the positive and negative forces as they come together[203] *at the neutral center and as each seeks its 'coincident'*[204] *by harmonic waves, flows or streams."*

7.39.4 Receptive Transmittive Concordance induces Rotation

"Every body capable of rotation is susceptible to the operation of force, which, applied, impels motion. Receptive transmittive concordance[205] *induces rotation. This simply means - the reception of streams of energy by the neutral center, and the transforming of them into radiant outflowing streams of energy causes rotation of the molecule or mass. All the magnets in the world, no matter how differentiated cannot induce rotation, but polar negative attraction*[206] *induces rotation. The earth's rotation is caused and kept in continuance rotation by the action of the positive and negative sympathetic celestial streams*[207]*, that is, by the inflowing celestial*[208] *from the sun and the outflowing radiant celestial dispersing this same energy to all planetary masses in space."*[209][210]

7.39.5 Polar and Depolar Differentiation result in Motion

"Polar and depolar differentiation[211] *result in motion. The compound interetheric or seventh subdivision is the Soul of Matter*[212]*, from which all*

---

[203] syntropically and [Law of Cycles]

[204] [VIBRATIONAL COINCIDENTS]

[205] [receptive transmittive concordance]

[206] [polar negative attraction]

[207] [sympathetic celestial streams]

[208] [celestial]

[209] [Law of Cycles]

[210] [Keely; [Snell Manuscript], PHILOSOPHY OF TRANSMISSION AND ROTATION OF MUSICAL SPHERE]

[211] [differentiation]

[212] [Soul of Matter]

forms of matter receive their introductory impulse.[213]"[214]

Polar and depolar is Keely's way of saying having two poles (as in a magnet or electric circuit) or not.

7.39.6 Neutral Center or fulcrum represents Focalization

"The neutral center (fulcrum) represents focalization (of syntropic forces) and distribution of the streams of (entropic) energy. It is not associated with magnetism. When the radiant elements generated by the focalizing chord are submitted to compound vibration of their mass thirds, those radiant elements become magnetic and rotation ceases. Rotation is induced by submitting the mass to three different orders of vibration simultaneously, giving the majority to the harmonic third."[215]

7.39.7 Rotation of the ether produces condensation

"When we rotate a mass with sufficient rapidity, the particles of that mass ultimately overcome cohesion by dispersing as fragments at a tangent but should we cause rotation of the ether, this would produce condensation, which is opposite in effect. This condensation effect increases with the velocity of rotation of the ether. This is the direct cause of formation of molecules as well as planetary masses.[216]

The above describes the difference between centrifugal (dispersive) and centripetal (attractive) forces. The first belongs to the molecular and atomic domains or orders while the latter is a function of the etheric order or domain. In this we can see when the etheric is the exciter rotation of the molecular results - when proponderance is given the etheric order or third.

---

[213] [introductory impulse]

[214] [Keely; [Snell Manuscript], PHILOSOPHY OF TRANSMISSION AND ROTATION OF MUSICAL SPHERE]

[215] [Keely; [Snell Manuscript], PHILOSOPHY OF TRANSMISSION AND ROTATION OF MUSICAL SPHERE]

[216] [Keely;[Snell Manuscript], PHILOSOPHY OF TRANSMISSION AND ROTATION OF MUSICAL SPHERE]

7.39.8 Controlling "negative attractive"

"To control rotational force or produce commercialized energy, we must control through its properties the 'negative attractive' or 'enharmonic' current. This will solve the problem up to any limit of power."[217]

7.39.9 Testing Positive and Negative Rotation

"At the same time Keely was completing his third system[218]*, he was also completing an experimental sphere in which he intended to test the combination of the positive and negative rotation. This experiment was at least entirely successful which showed the explanation of rotation given above, to be correct. The sphere even rotated with physical vibrations from the positive and negative interchange of positive and negative waves - not streams in this case. He had a desperate struggle in seeking to learn these laws of polarization and depolarization*[219]*. It was necessary for him to understand these laws before he could unfailingly secure rotation and control the reversions*[220] *which so often had made wrecks of his machines."*[221]

7.39.10 Russell on rotation

"Mass is an accumulation of the universal constant of energy[222] *into higher potential. The greater the mass the greater the potential. Mass accumulates inductively along the centripetal lines of closing spirals, and is rotated by the impact of genero-active force. Mass dissipates conductively along centrifugal lines of opening spirals, and is rotated by the expulsion of radio-active*[223] *force. These opposite forces are the cause of rotation of mass."*[224]

---

[217] [Keely; [Snell Manuscript], PHILOSOPHY OF TRANSMISSION AND ROTATION OF MUSICAL SPHERE]

[218] [Keelys Three Systems]

[219] refers to the Laws of Being herein explained

[220] [reversion]

[221] [Snell Manuscript]

[222] [universal constant of energy]

[223] [radio-active]

[224] Russell, [The Universal One]

7.40 Controlling negative attraction

XL. "A practical example of rotation is a wheel in revolution on its axis. This is force in its commercial or economic aspect. To accomplish this result by molecular vibratory action, we must gain control of the negative attractive[225] or enharmonic current[226] of the triple flow, and the problem is then solved up to any limit of power."

End of Brinton article.

7.41 Notes

[1] Also called Circular wave, Curl, Love Wave, Lamb Wave, Surface wave, Etheric. *"Power of rotation comes on the positive and power of negation, arising when the receptive flows become independent of the circular chord of resonation (set up mechanically or otherwise) breaks up the rotational power."*[227]

[Rayleigh Wave][228]

[Vortex][229]

[2] Also called Compression wave. *"The vibratory velocity governing the magnetic flow ranges from 300,000 to 780,000 per second and comes under the first interatomic. This is the first order above odor and permeates the molecules of glass in the compass cover as air passes through a sieve. Being governed by the full harmonic chord this flow moves in straight lines free from molecular interference.*[230] ... *"when free of this differentiation are in straight lines."* see [Discordant]

[Figure 8.3 - Coiled Spring showing Longitudinal Wave]

---

[225] [negative attractive], [Sympathetic Negative Attraction]

[226] [Enharmonic]

[227] [ROTATION - Snell]

[228] [Rayleigh Wave]

[229] [Vortex]

[230] [MAGNETIC ENGINE - Snell]

[Longitudinal]
[Longitudinal Wave]
[Longitudinal Waves in Vacuum]

[3] *"Discordance in any mass is the result of differentiated groups, induced by antagonistic chords, and the flight or motions of such, when intensified by sound, are very tortuous and zig-zag;"* see [Discordant]
[Figure 8.4 - Transverse Wave]
[transverse]
[Transverse Wave]

[4] Wave Field]
[12.05 - Three Main Parts of a Wave]
[12.09 - The MINUS side]
[12.08 - The PLUS side]
[Figure 2.12.1 - Polarity or Duality]
[Figure 15.03 - Opposing Forces of Gravity and Radiation]
[7B.16 - Polarity], [Principle of Polarity]

[5] Artwork by [Jeremy Pfieffer] et al.

[6] [The Seven Subdivisions of Matter and Energy], [VACUUMS FROM VIBRATORY INDUCTION]

[7] Jose Arguelles, Earth Ascending, page 21; 1996 (3rd edition) [See Binary Triplet]

[8] [Trexar], [Trextrinar], [TREXAR - Snell], [Trexnonar]

[9] [7.1 - Matter is centralized motion]

[10] [Etheric Vibration. - The Key Force]

[11] [7.9 - Ultimate Constitution of Matter and Action of Force Regulating its Phenomena]

Chapter 8

Modes of Vibration

8.1 Two models of three modes

There are three distinct modes of vibration in the standard acoustic model. These are the compression or longitudinal wave, shear or transverse wave and Rayleigh, Lamb or Love wave.

There are three distinct modes of vibration in the Keely/SVP model. These are the

Harmonic,
Enharmonic and
Dominant modes.

Each of these modes have distinct attributes such as

mode of motion,
angle of incidence,
relative frequency or pitch and
creative aliquot component signatures.

The two sets of three modes do not appear to be directly translatable. However they do possess common characteristics.

8.2 Three Modes of Vibration

8.2.1 Compression or Longitudinal Wave or Mode

Longitudinal Wave[1] same as Compression Wave or Sound Wave.

1) A wave incident or normal to a surface of a medium.

2) A vibration moving directly away from a sound source in a straight line.

3) Generally, considered as a motion to and from a center or source.

When harmonic is generally considered syntropic and when enharmonic considered entropic.

"Vibration in which the principal motion is in the direction of the longest dimension."[2]

"A vibrating medium must lengthen and shorten two times per each lateral (transverse) oscillation; so one longitudinal motion equals two lateral or (transverse) motions. Longitudinal frequencies are independent of string tension."[3]

8.3 Three Modes of Vibration

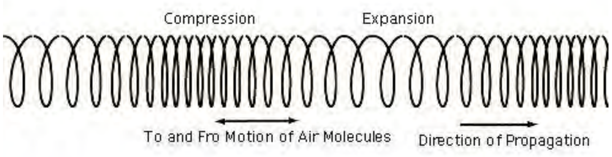

Figure 8.1: Coiled Spring showing Longitudinal Wave
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/longitudinal.jpeg

---

[1] [Longitudinal Wave]

[2] Rossing, Thomas D. The Science of Sound. Addison-Wesley Publishing Company, 1982

[3] Tyndall, John; Sound; Longmans, Green, and Co., London, 1893

The above and below are but an illustration commonly held of how a longitudinal wave behaves. This is a good working model but it does not explain the complex dynamics in vibrations which dynamics are touched upon in Chapter 6 and 7 where particles in a vibrating medium mutually attract (syntropy) then mutually disperse (entropy) periodically.

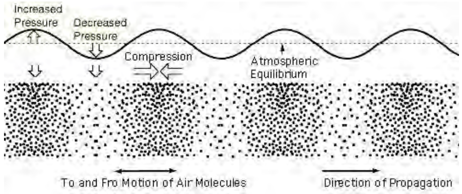
Increased Pressure
Decreased Pressure
Compression
Atmospheric Equilibrium
To and Fro Motion of Air Molecules
Direction of Propagation

Figure 8.2: Compression and Rarefaction in a Wave
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/compress-wave.jpeg

8.3.1 Transverse or Shear Wave or Mode

Waves 45° and higher to normal incidence.
Waves moving at 90° to the direction of (longitudinal) propagation.

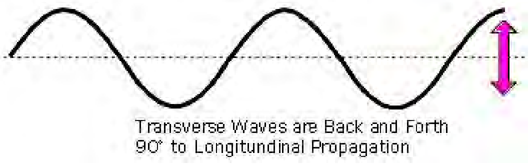
Transverse Waves are Back and Forth
90° to Longitundinal Propagation

Figure 8.3: Transverse Wave
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/transverse.jpeg

8.3.2 Rayleigh, Lamb or Love Wave or Mode

Rayleigh waves are circular vibrations. They travel along the surface of a vibrating object when the angle of incidence exceeds 90° or nearly parallel to the surface. This behavior is not unlike the idea electrons or electricity travels along the outside surface of a conducting wire and not through the conducting wire. This is the same form and motion of the 'curl' suggested by Maxwell[4] in 1871 and so often spoken about by Bearden[5]. Thus this circular form and motion would be universal principle of all periodic motion. *With this idea we can say with some confidence power travels longitudinally 'through' the wire, amplitude and polarity forms* 90° *to the longitudinal and the Rayleigh wave is the circulating path (skin effect*[6]*) around a conducting wire - all three wave forms are* 90° *to each other.* Think about that for awhile.....

Triple axis of the three modes of vibration drawn in reference to a circle and sphere. The three modes working together develop rotation by and through each other's motion - the rotation is the Rayleigh Wave or Surface Wave mode component. The first two modes cause (or are) straight line and zig-zag motions only. The third or Rayleigh Wave travels in a circular motion.

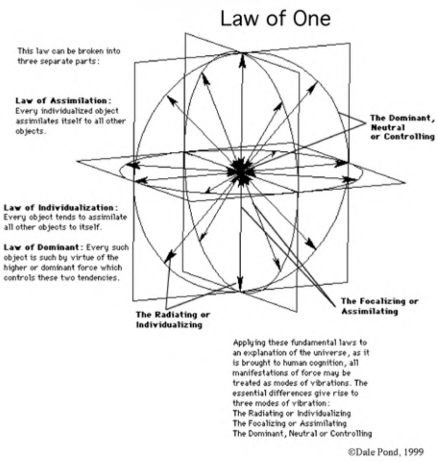
Law of One
This law can be broken into three separate parts:
Law of Assimilation: Every individualized object assimilates itself to all other objects.
Law of Individualization: Every object tends to assimilate all other objects to itself.
Law of Dominant: Every such object is such by virtue of the higher or dominant force which controls these two tendencies.
The Dominant, Neutral or Controlling
The Radiating or Individualizing
The Focalizing or Assimilating
Applying these fundamental laws to an explanation of the universe, as it is brought to human cognition, all manifestations of force may be treated as modes of vibrations. The essential differences give rise to three modes of vibration:
The Radiating or Individualizing
The Focalizing or Assimilating
The Dominant, Neutral or Controlling
©Dale Pond, 1999

Figure 8.4: Relationship of Three Modes
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/Triune-Modes-Motion.jpg

---

[4] James Clerk Maxwell (13 June 1831 5 November 1879) was a Scottish theoretical physicist and mathematician. See [James Clerk Maxwell]

[5] Thomas E. Bearden, Lieutenant Colonel, U.S. Army (Retired). Born Dec. 17, 1930, Cheniere, Louisiana. See [Bearden]

[6] Wikipedia/skin effect

8.4 Introduction to Keely's Wave Function

Perhaps one of the more interesting aspects of Keely's work is his Wave Function. The Wave Function is comprised of those foundational forces and attributes creating all oscillation and vibration which are seen effects of those unseen modal causes forces and attributes. Fundamentally these forces are in the 3, 6 and 9 proportions discussed in Chapter 7.

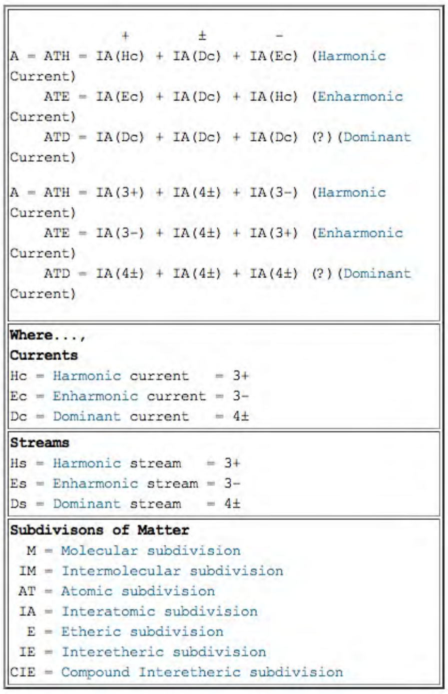

```
              +        ±        -
A = ATH = IA(Hc) + IA(Dc) + IA(Ec) (Harmonic Current)
    ATE = IA(Ec) + IA(Dc) + IA(Hc) (Enharmonic Current)
    ATD = IA(Dc) + IA(Dc) + IA(Dc) (?)(Dominant Current)

A = ATH = IA(3+) + IA(4±) + IA(3-) (Harmonic Current)
    ATE = IA(3-) + IA(4±) + IA(3+) (Enharmonic Current)
    ATD = IA(4±) + IA(4±) + IA(4±) (?)(Dominant Current)
```

Where...,

Currents

Hc = Harmonic current = 3+
Ec = Enharmonic current = 3-
Dc = Dominant current = 4±

Streams

Hs = Harmonic stream = 3+
Es = Enharmonic stream = 3-
Ds = Dominant stream = 4±

Subdivisons of Matter

M = Molecular subdivision
IM = Intermolecular subdivision
AT = Atomic subdivision
IA = Interatomic subdivision
E = Etheric subdivision
IE = Interetheric subdivision
CIE = Compound Interetheric subdivision

Figure 8.5: Suppositional Math and Symbolic Structure
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/table-13-01.jpg

The above table is an attempt to correlate some of what has been discussed.Elements include the three modes, their combinations, seven subdivisions and Russell's Indig numbers[7]. The following table is another presentation of a portion of the above data.

---

[7] [Indig numbers]

Aggregation	**Governing Mode**	**Dominant Mode**	**Subservient Mode**
Molecular	100 x Enharmonic	$66\frac{2}{3}$ Harmonic	$33\frac{1}{3}$ Dominant
Atomic	100 x Harmonic	$66\frac{2}{3}$ Enharmonic	$33\frac{1}{3}$ Dominant
Etheric	100 x Dominant	$66\frac{2}{3}$ Harmonic	$33\frac{1}{3}$ Enharmonic
Thirds	3	2	1
Proportion	9	6	3
Portion	Ninths	Sixths	Thirds

Table 8.1: Modes of Vibration[8]

8.5 Feynman Morphology

Richard Feynman's morphology[9] from his book, QED: The Strange Theory of Light and Matter[10], depicts Keely's $\frac{1}{3}$ and $\frac{2}{3}$ nested triune structure concept.

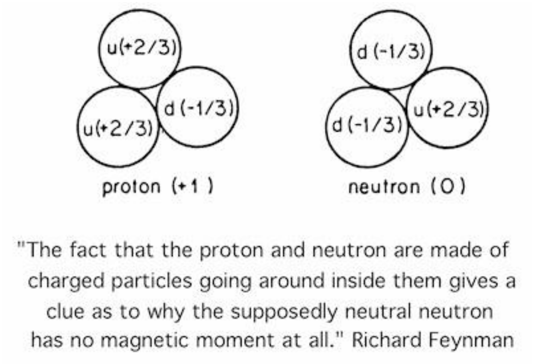

"The fact that the proton and neutron are made of charged particles going around inside them gives a clue as to why the supposedly neutral neutron has no magnetic moment at all." Richard Feynman

Figure 8.6: Feynman's triplet structures of the proton and neutron
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/feynmandiagram400.jpg

---

[9] [Figure 4.14 - Feynmans Triplet Structures of the Proton and Neutron]

[10] QED: The Strange Theory of Light and Matter; Princeton Science Library, ISBN-10:0691125759

Thus we see again Keely anticipated the quantum world decades before conventional science even had a glimmer of its existence as we also saw in the Rule of Threes (section 7.2.3 the Rule of Threes) and elsewhere.

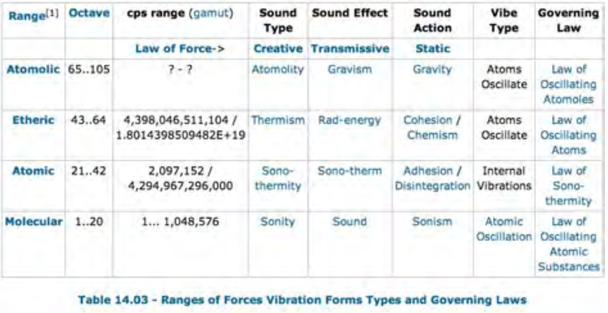

Range[1]	Octave	cps range (gamut)	Sound Type	Sound Effect	Sound Action	Vibe Type	Governing Law
		Law of Force->	Creative	Transmissive	Static		
Atomolic	65..105	? - ?	Atomolity	Gravism	Gravity	Atoms Oscillate	Law of Oscillating Atomoles
Etheric	43..64	4,398,046,511,104 / 1.8014398509482E+19	Thermism	Rad-energy	Cohesion / Chemism	Atoms Oscillate	Law of Oscillating Atoms
Atomic	21..42	2,097,152 / 4,294,967,296,000	Sono-thermity	Sono-therm	Adhesion / Disintegration	Internal Vibrations	Law of Sono-thermity
Molecular	1..20	1... 1,048,576	Sonity	Sound	Sonism	Atomic Oscillation	Law of Oscillating Atomic Substances

Table 14.03 - Ranges of Forces Vibration Forms Types and Governing Laws

Figure 8.7: Table of ranges of forces, forms and laws
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/table-14-03.jpg

(See Law of Transformation of Forces[11] and subdivision[12] for explanation of the above table)

---

[11] [Law of Transformation of Forces]

[12] [Subdivision]

Source	**From Center**	**To Center**	**Controlling**
Keely / SVP[6]	Enharmonic	Harmonic	Dominant
Modern wave form	Longitudinal[2]	Transverse[3]	Rayleigh[1]
Angle of Incidence	0° – 45°	45° – 90°	> 90°
Path/Motion[3]	Straight	Zig-zag	Circular
Keely / SVP	Focalizing	Radiating	Dominant
Keely / SVP	Syntropic	Entropic	Dominant
Russell[4]	Plus	Minus	Fulcrum
Russell[4]	Gravitation	Radiation	Balance, Void
Charles Henry[7]	Gravitation	Electromagnetic	Biopsychic
Possible Correspondence[5]	Thirds	Sixths	Ninths

Table 8.2: Table of attributes of the three modes

Mode	**Subdivision**	**Octave** [Hz = 2^{octave}]	**Law**
Enharmonic	Molecular / Intermolecular	1-21	Law of Oscillating Atomic Substances[13]
Harmonic	Atomic / Interatomic	21-42	Law of Sonothermity[14]
Dominant	Etheric / Interetheric	42-63	Law of Oscillating Atoms[15]

Table 8.3: Keely's three modes; subdivision, octave and law

‘*‘By use of the dominant mode, which is allied to the “order” of etheric vibrations, we can induct, “sympathetic negative attraction” or “sympathetic positive propulsion” in any mass, according to its mass chord.”*[16]

---

[16] [SYMPATHETIC INDUCTION - Snell]

8.6 Footnotes

[1] Also called Circular wave, Curl, Love Wave, Lamb Wave, Surface wave, Etheric. *"Power of rotation comes on the positive and power of negation, arising when the receptive flows become independent of the circular chord of resonation (set up mechanically or otherwise) breaks up the rotational power."*[17]

[Rayleigh Wave]
[Vortex]

[2] Also called Compression wave. "*The vibratory velocity governing the magnetic flow ranges from 300,000 to 780,000 per second and comes under the first interatomic. This is the first order above odor and permeates the molecules of glass in the compass cover as air passes through a sieve. Being governed by the full harmonic chord*[18] *this flow moves in straight lines free from molecular interference.*[19] *"when free of this differentiation are in straight lines."*[20]

[Figure 8.3 - Coiled Spring showing Longitudinal Wave]
[Longitudinal]
[Longitudinal Wave]
[Longitudinal Waves in Vacuum]

[3] *"Discordance in any mass is the result of differentiated groups, induced by antagonistic chords, and the flight or motions of such, when intensified by sound, are very tortuous and zig-zag;"*[21]

[Figure 8.4 - Transverse Wave]
[transverse]
[Transverse Wave]

[4] See [Wave Field]

[Three Main Parts of a Wave]
[12.09 - The MINUS side]
[12.08 - The PLUS side]

---

[17] [ROTATION - Snell]
[18] [full harmonic chord]
[19] [MAGNETIC ENGINE - Snell]
[20] [Snell Manuscript], [Discordant]
[21] [Keely and His Discoveries, pg 275], [Discordant]

[Figure 2.12.1 - Polarity or Duality]
[Figure 15.03 - Opposing Forces of Gravity and Radiation]
[7B.16 - Polarity, Principle of Polarity]

[5] Artwork by [Jeremy Pfieffer] et al.

[6] [The Seven Subdivisions of Matter and Energy], [VACUUMS FROM VIBRATORY INDUCTION]

[7] Jose Arguelles, Earth Ascending, page 21; 1996 (3rd edition) See [Binary Triplet]
See [MUSICAL CHORDS], [Triplet]

Chapter 9

Keely Wave Function

9.1 Introduction to Keely's Wave Function

One of the keys to Keely's mastery of matter and energy lies in what I call the Keely Wave Function[1] and his ability to manipulate it. A wave function is a mathematical description of the factors involved in a sine-like wave, vibration or oscillation expressed over time. Generally, science considers a wave function of one dimension or within a closed and defined range of frequency (to put it simply) which includes the constructive attributes of amplitude, Time and Space (wave length). In Keely's case the wave function is composite three dimensional wave function that operates within three levels, orders or ranges of frequencies, simultaneously and having additional constructive attributes or aliquot parts within those orders or levels. Thus it is more complicated but also more inclusive of the factors operating naturally within a wave, vibration or oscillation. The Keely Wave Function would include some if not all of the dimensions of the Russell Wave Function[2].

Perhaps one of the more interesting aspects of Keely's work is his Wave Function. The Wave Function is comprised of those creative foundational forces creating all oscillation and vibration which are seen effects of those unseen causes. Fundamentally these forces are in the 3, 6 and 9 proportions discussed in the [Laws of Being - Annotated] and [Modes of Vibration - Annotated] and hinted at by Nikola Tesla and discussed in this book primarily in Chapter 7 and 8.

---

[1] [Keely WaveFunction]

[2] [Eighteen Attributes or Dimensions]

The wave function preforms not too unlike complex impedance[3] calculations over time but with more variables or dimensions. The Keely Wave Function operates/is on three levels or orders simultaneously:

Molecular
Atomic
Etheric

Each level or order has three motions/dynamics:

Harmonic = Syntropic
Enharmonic = Entropic
Dominant = Neutral

This table below shows the three main levels or orders plus a fourth level or order and the forces active in each as also the law governing each specific frequency range.

---

[3] [impedance]

Range[1]	Octave	cps range	Sound Type	Sound Effect	Action	Vibe Type	Governing Law
		Law of Force>	**Creative**	**Transmissive**	**Static**		
Atomolic	65-105	$? - ?$	Atomolity	Gravism	Gravity	Atoms Oscillate	Law of Oscillating Atomoles
Etheric	43-64	4,398,046,511,104 - 1.8014398509E+19	Thermism	Rad-energy	Cohesion / Chemism	Atoms Oscillate	Law of Oscillating Atoms
Atomic	21-42	2,097,152 - 4,294,967,296,000	Sono-thermity	Sono-therm	Adhesion / Disintegration	Internal Vibrations	Law of Sono-thermity
Molecular	1-20	1 - 1,048,576	Sonity	Sound	Sonism	Atomic Oscillation	Law of Oscillating Atomic Substances

Table 9.1: Ranges of forces vibration forms types and laws

Table 9.1 shows the break-out into orderly detail the components, activity and functions of Keely's Law of Transformation of Forces[4].

9.1.1 Law of Transformation of Forces

"All forces are different forms of Universal Energy[5] *unlike in their period-frequency, merging into each other by imperceptible increments; each form representing the compass of 21 octaves. Each form or pitch may be transformed into an equivalent quantity of another pitch above or below it in the*

---

[4] [Law of Transformation of Forces]

[5] [12.02 - Universal Energy Unit Constant or Impulse]

scale of 105 octaves[6]*. The transformation can occur only through its static effect, developing vibrations of harmonic pitches above or below their fundamental vibration, or developing with juxtaposed aggregates, resultant and difference, or third order, as the case may be.*

Scholium: *"A table of the intervals and harmonics of the normal harmonic scale will indicate the ratios in which the transformation of forces will occur."*[7]

It is said energy cannot be created or destroyed however energy can be transformed or transmuted into different forms. These vibratory forces, states and conditions have reciprocity and are expressed in ratios of relative values or 'modes' thus they can be changed one into the other under special circumstances as described in Keely's Law of Transformation of Forces.

9.2 Thirds as Ratios within a Whole

In Brinton's "Laws of Being" (Chapter 7) we can see the description of Keely's Thirds, Sixths and Ninths as learned, understood and presented by Professor Brinton during his visits to Keely's laboratory. We can also see these terms are not specifically musical intervals of thirds, sixths or ninths, as considered by musicians, but they are musical in terms of relativity, proportion and value. They are relative proportions within a whole. Flows and currents can be understood by the metaphor of a river or stream. The stream flows as a whole and has within this stream many currents. Herein the syntropic flow to a center is composed of three currents. Each current is a third $\frac{1}{3}$ of the whole flow or stream. These currents are expressed as ratios within a whole of 100; *i.e.*, $33\ \frac{1}{3} : 100$ or $66\ \frac{2}{3} : 100$. The whole wave, flow or stream, regardless of polarity, is 100. Exploring the literature a bit further we find the following relationships:

"Silver represents the third, gold the sixth and platinum the ninth, in their respective relative molecular oscillating ranges. This triune condition will equate thirds in vibratory frequencies, that is, chords in intervals of thirds will set up disturbances in the Trexar and these disturbances will be equated

---

[6] [Scale of the Forces in Octaves]

[7] [Keely]; [Law of Cycles], [12.06 - Mid-Tones and Neutral Centers], [Scale of the Forces in Octaves]

so as to be transmitted as thirds from the positive or farthest end of the wire.”[8]

---

[8] [TREXAR - Snell]

Thirds	**Sixths**	**Ninths**
Silver	Gold	Platinum
33 $\frac{1}{3}$	66 $\frac{2}{3}$	100
neutral	depolar	polar
33 $\frac{1}{3}$	>>Attractive - Equation<<	100
<<Repulsive - Equation>>	66 $\frac{2}{3}$	100
neutral	entropic	syntropic
celestial	terrestrial - matter dissolving	terrestrial - matter forming
dominant	individualizing	assimilative
Mind as Intent, Will, Volition, Desire	I want to be One with Everything	I want Everything for myself
Ideals	Service to Others	Service to Self
Fulcrum, Neutral, Balance	Radiating from Center	Contracting to Center
Controller	Dispersing, Dissociating	Concentrating, Associating
Neutral	Minus	Plus
Dominant	Enharmonic	Harmonic
Neutral	Magnetic	Electric
Fulcrum, Void	Mother	Father
Now	Life Reducing (death)	Life Enhancing
Neutral Chords	Dispersing Chords	Assimilating Chords

Table 9.2: Neutral (thirds); energy radiates from center (sixths); force contracts to center (ninths)

The above three levels and three forces acting in each are delineated in the pages listed below, especially within the [Laws of Being - Annotated] and

[Modes of Vibration - Annotated] all subject of this book especially Chapter 7. The action of each mode is determined by their composite proportions or modal preponderance. Hence a mode may have different action in different contexts, use or circumstance.

Chapter 10

Proportions (Ratios) of Relative Values

10.1 Proportion

Proportion is the ratio (relativity) of two or more numbers or quantities to each other as have been presented in the preceding chapters. The same of course applies to notes in a music chord. While the individual frequencies are important they take on a new meaning according to their proportion within the scale, chord and to each other. This gives us an insight into why Keely used music notation to such an extent. On Keely's charts can be seen the following terms of numeric relationship.

Proportion is in three kinds: (1) multiplex. (2) Superparticularis. (3) Superpartiens. Proportio multiplex is when the larger number contains the smaller so many times without a remainder, as

2:1 (dupla)
3:1 (tripla)
4:1 (quadrupla).

Proportio superparticularis is when the larger number exceeds the smaller by one only as

3:2 (sesquialtera)

4:3 (sesquitertia)
5:4 (sesquiquarta).

Proportio superpartiens is when the larger number exceeds the smaller by more than one, as

5:3 (superbipartienstertias)
7:4 (supertripartiensquartas)
9:5 (superquadripartiensquintas).

Thus, it will be understood, that instead of giving simply the ratio between two numbers, early writers on arithmetic and geometry, as well as music, coined a single word to express that ratio; for example, 17:5 was said to be Triplasuperbipartiensquintas, *i.e.*, that the larger number contained the smaller number three times (tripla) with two remainder (bipariens). Again, Triplasupertripartiensquartas proportio, signified that the larger contained the smaller three times and three over, as 15:4, 27:8, *etc.*, the last part of the compound word always pointing out the smaller of the numbers compared, or an exact multiple of it. Lastly, the addition of sub showed that the smaller number was compared to the larger, *e.g.*, 4:15 would be called Subtriplasupertripartiensquartas proportio. This system of proportion was used not only with reference to intervals but also to the comparative length of notes (time).[1]

An excellent booklet on proportion is 'Propositions of Proportions'[2] in the svpbookstore.com online catalog.

10.2 Reciprocating Proportionality

Reciprocating Proportionality means what is added to one side must be simultaneously subtracted from the other. A children's seesaw is a perfect analogy. Whatever quantity or rate of change added to one side must be subtracted from the other. Another good analogy is a mathematical equation where what is added to one side must be subtracted from the other side or where one side is multiplied the other is divided. A dynamic balance where the whole is always one while the two balanced forces in opposition (*e.g.*; syntropy and entropy) are constantly changing in direct proportion to one another. The Square Law and Inverse Square Law is another good example. This equal and balanced reciprocating dynamic is at the root of Russell's Rhythmic Balanced Interchange[3] and Macvicar's cosmical Law of Assimilation[4].

---

[1] Stainer, John; Barrett, W. A.; A Dictionary of Musical Terms; Novello, Ewer and Co., London, pre-1900

[2] svpbookstore.com

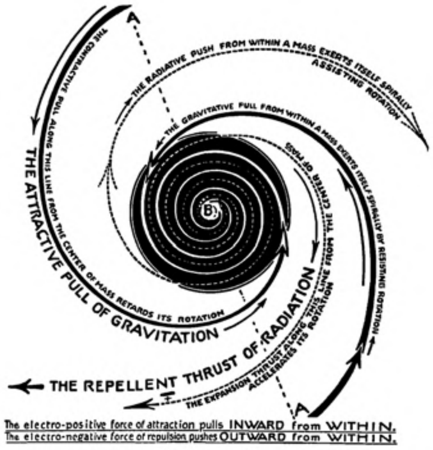
A
THE CONTRACTIVE PULL ALONG THIS LINE FROM THE CENTER OF MASS RETARDS ITS ROTATION
THE ATTRACTIVE PULL OF GRAVITATION
THE RADIATIVE PUSH FROM WITHIN A MASS EXERTS ITSELF SPIRALLY ASSISTING ROTATION
THE GRAVITATIVE PULL FROM WITHIN A MASS EXERTS ITSELF SPIRALLY BY RESISTING ROTATION
CENTER OF MASS
B
THE EXPANSION THRUST ALONG THIS LINE FROM THE CENTER ACCELERATES ITS ROTATION
THE REPELLENT THRUST OF RADIATION
A
The electro-positive force of attraction pulls INWARD from WITHIN.
The electro-negative force of repulsion pushes OUTWARD from WITHIN.

FIGURE 7

Figure 10.1: One Balanced Whole and Two Reciprocating Dynamics
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/reciprocating.jpg
(courtesy University of Science and Philosophy)

10.3 Reciprocating Proportionality

"The underlying law of Creation is RHYTHMIC BALANCED INTERCHANGE in all transactions in Nature. It is the one principle upon which the continuity of the universe depends. Likewise, it is the one principle upon which the continuity of man's transactions, his health and happiness depend. It is the manifestation of God's two opposing desires in all processes of creation.

"BALANCE is the principle of unity, of oneness. In it is the stability which lies in CAUSE. BALANCE is the foundation of the universe.

---

[3] [Rhythmic Balanced Interchange]

[4] [Law of Assimilation]

"BALANCED INTERCHANGE simulates oneness by interchange between pairs of opposites. It is the principle of equal giving between all moving pairs of unbalanced opposites which constitute this dual electric universe. In it is the instability of EFFECT. Instability is forever seeking to find stability. It can never find it, but it can simulate it by balancing its instability through equality of interchange.

"RHYTHMIC BALANCED INTERCHANGE is the principle of continuity of EFFECT. Balanced interchange between opposites repeats simulation of stability, and rhythmic balanced interchange continues that repetition."[5]

The degree to which a vibrating or oscillating system is either one polarity or the other can be pictured or expressed in relative and accumulative numbers. Russell used the + sign to indicate accumulating, concentrating (syntropic) quantities and a − sign to indicate dispersive, radiating (entropic) quantities. Russell used his Scale of Locked Potentials[6] to measure degrees of polarity and this author believed Keely similarly used his mysterious thirds, sixths and ninths as part of his referencing method where proportions or parts of a whole were given proportioned or ratioed values.[7]

10.4 Reciprocal Proportion

Increases and decreases (establishing polar preponderance) are in inverse or direct proportion to rate of change or *delta*.

Energy at Plane of Inertia is neutral or depolar. $-0\ 0+$.

---

[5] Russell, [The Secret of Light], page 106-107

[6] [Scale of Locked Potentials]

[7] [3.13 - Reciprocals and Proportions of Motions and Substance]

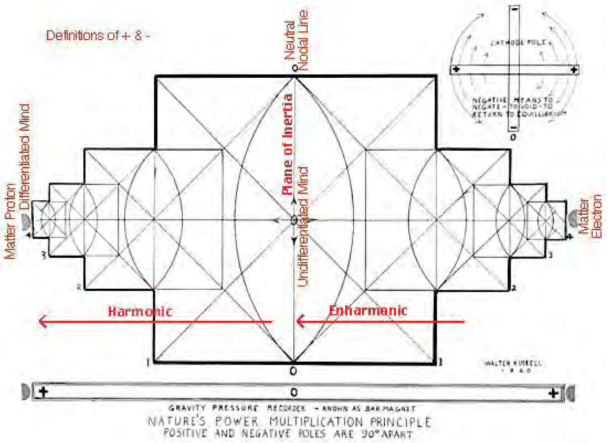
Definitions of + & -
Neutral Nodal Line
Plane of Inertia
Undifferentiated Mind
Matter Proton
Differentiated Mind
Matter Electron
CATHODE POLE
NEGATIVE MEANS TO NEGATE - TO VOID - TO RETURN TO EQUILIBRIUM
Harmonic
Enharmonic
WALTER RUSSELL
GRAVITY PRESSURE RECORDER - KNOWN AS BAR MAGNET
NATURE'S POWER MULTIPLICATION PRINCIPLE
POSITIVE AND NEGATIVE POLES ARE 90° APART

Figure 10.2: Plane of Inertia
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/plane-of-inertia.jpeg
(courtesy University of Science and Philosophy)

Degree of Concord or Discord is determinative anywhere from −0 0+ to 4 + +[8].

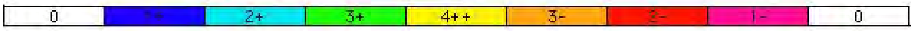

Figure 10.3: Scale of 0 to 4 to 0
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/lockedpotentials.jpeg

When preponderantly harmonic a larger volume −0 0+ condenses to smaller volume of −1 1+ progressively to 4 + +.

When preponderantly enharmonic a smaller volume 4 + + dissociates and expands to larger volume of −3 3+ progressively to −0 0+[9].

---

[8] [4plusplus]

[9] [4plusplus]

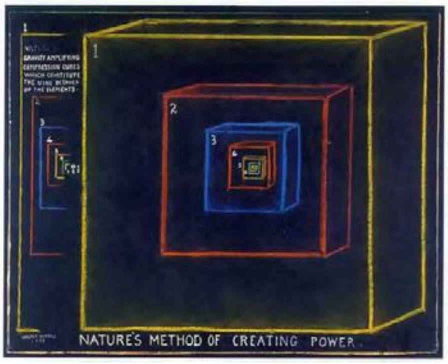
1
2
3
4
NATURE'S METHOD OF CREATING POWER.

Figure 10.4: Nested cubes or increasing and decreasing proportions
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/nested-cubes.jpg
(courtesy University of Science and Philosophy)

Volume decreases with increased potential.

Volume increases with decreased potential. For numeric details and relationships on the above see [cube-sphere] and [Cubing the Sphere].

Transformations occur at the −0 0+ and 4 + + points, states or conditions. The first according to Russell's Principle of Regeneration[10] (Keely's "static effect") and the latter by Keely's Law of Cycles[11]. See Figure 12.8 for an animated graphic show progressive generation of discords until dissociation.

10.5 Reciprocals and Proportions of Motions and Substance

"The following fundamental laws governing mathematical relations will briefly convey the idea of the Principles governing universal ratios which is in every effect of motion the direct and inverse square of the distance, area or time dimensions and the cube of volume or potential dimensions. The direct and the inverse are the two way indicators of direction, the latter belonging to

---

[10] [Principle of Regeneration Russell], [13.13 - Principle of Regeneration] and [Principle of Regeneration Bearden]

[11] [Law of Cycles]

gravitational and the former to radiational effects."[12]

Notice the reference to the reciprocals Square Law and Inverse Square Law in the above Russell quote. These two laws are reciprocals of each other and reflect the natural ways force and energy[13] interchange in nature when forming pressures, vacuity, rotation, gyroscopic motions, or vortices.

"All dimensions[14] simultaneously expand and contract in opposite directions of the same ratio."[15]

10.6 Thirds as Ratios within a Whole

In the above quote we can see the description of Keely's Thirds, Sixths and Ninths as presented by Professor Brinton after his visits to Keely's lab on Chestnut street, Philadelphia. We can also see these terms are not specifically musical intervals of thirds, sixths or ninths, as considered by musicians, but they are musical in terms of relativity, proportion and value. They are proportions within a whole. Flows and currents can be understood by the metaphor of a river. The river flows as a whole stream and has within this flow many currents. Herein the syntropic flow to a center is composed of three currents. Each current is a third ($\frac{1}{3}$) of the whole flow. These currents are expressed as ratios within a whole of 100; i.e., $33\ \frac{1}{3} : 100$ or $66\ \frac{2}{3} : 100$. The whole wave or stream, regardless of polarity, is 100. Exploring the literature a bit further we find the following relationships:

"*Silver represents the third, gold the sixth and platinum the ninth, in their respective relative molecular oscillating ranges. This triune condition will equate thirds in vibratory frequencies, that is, chords in intervals of thirds will set up disturbances in the Trexar[16] and these disturbances will be equated so as to be transmitted as thirds from the positive or farthest end of the wire.*"[17]

---

[12] Russell, [Genero-Radiative Concept], part 6b

[13] In SVP [Force] is considered Syntropic while [Energy] is considered Entropic.

[14] [Dimension]

[15] Russell

[16] [Trexar]

[17] [TREXAR - Snell]

Chapter 11

Universal Heart Beat

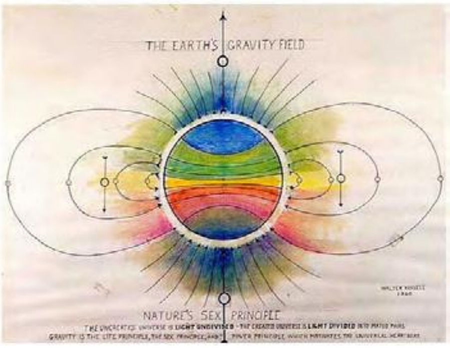

Figure 11.1: Universal Heartbeat
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/heartbeat.jpg
(courtesy University of Science and Philosophy)

Universal Heart Beat =
Dynaspheric Force[1] =
Rhythmic Balanced Interchange[2] =
Triune Polar Flows[3]

"The eternally conscious entity - call it by whatever name we please - moves

---

[1] [Dynaspheric Force]
[2] [Rhythmic Balanced Interchange]
[3] [Triune Polar Flows], [Figure 13.12 - Triune Polar Stream Characteristics]

in cycles as eternal and infinite as itself; it oscillates and vibrates perpetually."[4]

The universal heartbeat is the periodic flow of a polar system from a preponderance of one pole then to preponderance of its seeming opposite pole. This is the action of all vibration, oscillation and waves no matter what order or level (frequency or media) it is. Russell mentions "*in the wave lies the secret of Creation.*"[5] The secret then of Creation[6] is the unseen dynamics of the wave, vibration or universal heartbeat - the essence of this book. The following are primarily Russell quotes describing this universal dynamic or dance of polarities.

11.1 The Love Cycle

"Energy accumulates during genero-active inhalation by rising potential, and is dissipated during the radio-active[7] *exhalation by lowering potential. The periodicities of inhalation and exhalation in all mass are absolute."*[8]

"All the force of electricity is exerted in the attempt to create the illusions of form and dimension.

"All the force of magnetism is exerted in the attempt to destroy all illusion, all form and all dimension.

"Neither force completely fulfills its desire, for each partially thwarts the other.

"The energy of magnetism is the elastic energy of expansion, a straining energy ever pushing toward the inertial line of equalized pressures which lies between any two masses, while the energy of electricity is ever pulling toward the pulsing heart, the gravitational nucleus of every mass."[9]

---

[4] Keely's description of that state of being commonly known as God, [Dashed Against the Rock], see [14.30 - Effect of Preponderance] and [14.31 - Preponderance Russell]

[5] [In the Wave lies the Secret of Creation]

[6] [Creation]

[7] [radio-active]

[8] Russell, [The Universal One]

[9] Russell, [The Universal One]

"*To these periodic oscillations, alternating in sequence, is due the revolution and rotation of all mass.*[10]

"*All evolving and devolving mass beats time in accord with its periodic sequences of alternating electro-magnetic oscillations.*

"*All mass beats time in accord with its varying potential.*

"*The electro-magnetic oscillations of the universe are the heart-beats, the life pulsations of the universal One.*"[11]

"*The misconception of gravity as a force of attraction which pulls inward from within. The only force which motivates the heart beat of this universe is the cyclic wave force of gravity which expands and compresses, heats and cools, integrates and disintegrates, centripetally and centrifugally.*"[12]

"*Waves of motion spring from the stillness of the universal equilibrium. They are the universal heartbeat which manifests eternal life and power in God's stillness by eternal repetitions of simulated life and power as expressed in waves of motion.*"[13]

"*Each interchange between the invisible omnipresent universe where motion begins and ends, and the visible transient universe, which multiplies and divides the speed and power of motion, is a cycle. Pulsing cycles constitute the heart beat of this universe, which simulates Mind-Idea through pulsing cycles of two-way motion. Every pulsation of the life principle of multiplying motion creates divided electric male and female bodies, which seek rest in each other from the strains and tensions of their division into pairs of opposites.*"

"*This shows how the heartbeat piston operates between the vacuity of the*

---

[10] [See 7.39.4 Receptive Transmittive Concordance induces Ro- tation]

[11] Russell, [The Universal One], Book 02 - Chapter 03 - New Concepts of Electricity and Magnetism

[12] Russell, [Russells Optic Dynamo-Generator]

[13] Russell, [The Secret of Light], page 115

Creator's Magnetic Light[14] *universe of Mind-energy*[15] *and the potential of the electrically pressured universe to create cycles of two-way motion.*"[16]

"*The sexless Father-Mother Creator is One. His extended sex-conditioned, male and female bodies are the completion of His Trinity.*

"*Rest and action are three. Space and matter are three. Equilibrium and motion are three. Dimensions and pressures are three. The heartbeat of the universe, and yours, are three. Likewise, its breathings and yours, its temperatures and yours, and all things else of the universe, and you, are three.*"[17]

"*To account for this unnatural phenomena The Coulomb Law was adopted, which says that opposites attract, meaning oppositely sexed mates and likes repel, meaning similarly sexed pairs. This law is invalid, for oppositely sexed mates do not attract. They are forced into collision by the electric action of dividing Oneness into unit pairs. The instant that they do collide they use their utmost endeavors to re-attain the sex-divided condition by re-charging their discharged condition. Nature helps in this process by assisting to recharge with every heartbeat, every breathing cycle, and the food one eats.*"[18]

"*By turning on the current it will be found that the spot between the two electrodes is of zero potential, and that many lines of radiation extend from that center and cause a moving electric current to spin around the zero fulcrum center where the ring is located. Not only that but an electrode placed in contact with the ring, will convey current away from it. This effect of electric current extending from a still center to convey its power to the ring, and other outside conductive matter, is identical with the mental example quoted in Fig. 48.*

---

[14] [Magnetic Still Light]

[15] [Mind Force]

[16] Russell, [Atomic Suicide]?, Chapter 4 - The True Nature of This Mind and Motion Universe

[17] Russell, [Atomic Suicide]?, Chapter 5 - Prelude - The Transformation of Man

[18] Russell, [Atomic Suicide]?, Chapter 8 - The Oneness of Gravity and Magnetism; see also [PoL - Chapter 9]

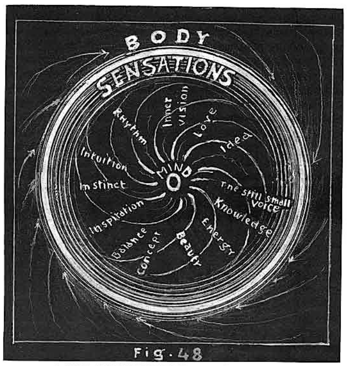

Fig·48

A MIND-CENTERED THOUGHT-RING OF MOTION

Figure 11.2: Neutral Center 4++ extention to periphery 0
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/ring-of-motion.jpg
(courtesy University of Science and Philosophy)

"*Anyone who may question this fact may not only prove it by this simple experiment but can also move the two electrodes away from the center of the rim so that their centering zero occupies an eccentric position, as the human heart does in relation to the chest, and as the zero center of the heart itself is eccentrically placed in relation to its mass. A different wave pattern will result but the electric qualities will remain the same. There is always a neutral center which centers the birth point of any action. Motion spins around that neutral center, but there is no motion at that center.*"[19]

"*Say also thou to man who thinketh of his heart beat as his own alone - unbound to all things else - that all creating things unfold with the pulse beat of My cyclic thinking. As My thinking is universal, so likewise, is My pulse beat universal.*

"*And furthermore I say that every blade of grass and leaf of oak in mighty forest pulseth with the heart beat of man and all things else, yea even down to microcosmic cells of forming things of earth, and sea, and sky; for even the most minute of these pulse with the electric waves of My cyclic thinking as the mightiest of stars in My heavens likewise pulse.*

"*For I, God of Light, thy Father-Mother God of Love, am the One Whole.*

---

[19] Russell, [Atomic Suicide]?, Chapter 9 - The Mind Nucleus of the Atom

From Me all unfolding-refolding things extend, and to Me they return for re-extension.

"*All things go from very heart of Me into My imaged universe; and when they disappear from there I also take them back to very heart of Me.*

"*For I am the Soul of art. To him whose Soul would touch My Soul, and feel the heartbeat of its mighty rhythms, I say, in so far as thou knoweth thy Self as Light, thou shalt know Me as Light.*"[20]

"*God has two desires: the desire to unfold His creations to give them form and set them into action; and the other desire is to refold His forms and take them back to Him for rest and resurrection so that they may repeat their action. That is the basis of the entirety of creation - the two desires of God, to act, to create, and to take back for re-creation. That is why God divided everything into pairs of opposites: the desire to express action, the desire to separate from Oneness, God's Oneness desiring to be manifested as many ones - the idea of separation and action; and the desire for return to rest from which that action sprang for the purpose of repeating the action in cycles, countless millions of cycles of action and rest, in-breathing, out-breathing; the pulse beat of the universe; the heartbeat of all things pulsing in all things to manifest life in action and reaction from the stillness of God's One Light in which there is no action and reaction.*"[21]

"*God's universe is entirely composed of microcosmic and macro-cosmic masses of pulsing electric waves, which we call matter. These pulsing waves constitute the universal heartbeat, or universal breath of God's body. God's body continues its manifestation of the life principle by breathing outward and inward sequentially, just as you breathe out and in in balanced sequences to continue your appearance of existence.*

"*Every mass in the universe breathes out and in, rhythmically. There are no exceptions to this law which every nebula, sun, planet or electron of an atom obeys, but in varying frequencies appropriate to their potential. Your breath frequency may be ten cycles per minute while the sun's frequency is*

---

[20] Russell, [The Message of the Divine Iliad]

[21] Russell, Fifth Kingdom, The Beginning of Man

one cycle in eleven years.

"*The most obvious fact of Nature is its repetition in reverse of every effect of motion in two-way pulsing cycles. It is unaccountably strange that science has never observed this most obvious of all of Nature's characteristics. Every cycle in Nature is a two-way, equal interchange between pairs of opposite conditions. That interchange between the equally balanced anodes and cathodes of this electric wave universe constitute its pulsing heartbeat, which likewise is cyclic, otherwise it would not continue.*"[22]

"*The test for ascertaining the average number of those who are ready for this higher step upward in the unfolding of the man-idea is not a religious one, it is a cultural one. In New York City, for example, about 7,000 people out of 8,000,000 people love the higher inspirational rhythmic creations of the world's great geniuses who interpret the heartbeat of Nature for the Souls of men. These same few shun noisy crowds, shudder at Jazz music, which distresses them acutely, in order to seek aloneness in the forests or oceanś shores where they can hear Nature's silent whisperings and rhythmic sounds within their very Souls.*"[23]

"*There is an indescribable glory all along the road which leads from your active thinking to the stillness of deep meditation. If you cannot actually go to the forest or sea, the better to commune with God, go there in your imagination. If some problem demands your thoughts let the imagined tones of the sea drown them out. If you imagine the sea and hear the rhythms of its waves pounding upon the shores as echoes of the heartbeat of the universe resounding within you, or if you hear the breezes whispering in the pines with your inner ears, the glory of that ecstasy will soon drown out dross thoughts of earth.*

"*Earth music is melodic and chains you to earth. The music of the spheres which you hear with inner ears, is deeply tonal, not melodic. The rhythms and tones of Nature's music which sing its octaves in the language of Light have always within them the pulsing rhythms of the universal heartbeat, always the one-two, one-two pulsing of the Father-Mother heartbeat of borning-reborning forms of things, which manifest the Father-Mother pulsing rhythms*

---

[22] Russell, [Home Study Course], Unit One - Prelude

[23] Russell, [Home Study Course], Unit One - Lesson 1.1

of the Soul's high heavens.

"*Always in your deep meditation the music of the silences from which sounds of earth spring is like unto the aftermath of the great symphonies by cosmic masters, which have refolded from sounds of earth which your ears have heard, into your very Soul, which still hears them in the pulse-beat of your imagining.*"[24]

"*And as love is rhythmically repeated in its givings and regivings to synchronize with the givings of your outward breath to the heavens and their regivings to you, and as such balanced rhythmic interchange between your heartbeat pulsings and all other electric extensions from the stillness of your Soul is LAW - the basic fundamental LAW of all creating things in all this electric universe - your own masterpiece is the measure of how divinely you have conceived it and how worthily you have rendered it in balance with Universal Law.*

"*To the extent that you can extend the beauty of your imagined rhythms to material bodies built in the images of your spiritual conceptions those rhythms which you create are masterpieces of great art. The very measure of the quality of your material interpretation lies in your ability to translate imagined forms and rhythms of the universal heartbeat into bodily forms and rhythms which can reinspire others with the ecstasy of your inspiration.*"[25]

"*Your transcendent genius is the result of your communion with God. By your masterly interpretation of the rhythmic heartbeat of His thinking you have caused others to commune with you and God. Your meditation and theirs are one. You have made them to forget their sensed bodies and dwell in the Mind kingdom of your high heavens with you and all illumined Souls who dwell in God's kingdom of Light.*"[26]

"*Imagine yourself communing with God indirectly, such as listening to the heavenly rhythms of God's heartbeat in a masterly symphony. While thus enraptured by the divine rhythms reaching your very Soul can you possibly*

---

[24] Russell, [Home Study Course], Unit One - Lesson 1.2
[25] Russell, [Home Study Course], Unit One - Lesson 2.1
[26] Russell, [Home Study Course], Unit One - Lesson 2.2

imagine yourself thinking evil, or thinking sinfully by planning to cheat, or lie, or steal? No - of course you cannot. The reason you cannot is because you are reflecting the love nature of the universe in your spiritual Self in its entirety, and there is no sin in that love nature to reflect."[27]

"*The entirety of Creation is the manifestation of love. Every effect of motion in nature is a manifestation of the nature of God, for Nature gives of its all in every action and regives equally in every reaction. If that were not so the movements of the planets would not be predictable.*

"*The electric heartbeat of Nature expresses love in both halves of each cycle of the universal heartbeat. Nature never takes for God never takes. Nature never gives unequally for God never gives unequally.*

"*God's equally balanced giving for regiving is never violated in Nature. It could not be violated for Nature is an extension of God's nature. For this reason the balance of Nature can never be upset.*

"*You can KNOW music and THINK it without having acquired the technique of expressing it, or without producing sounds to give it a body. Your inner ears can hear the music in the silence of God's eternal rhythms, and you can interpret your moods into the rhythms of the universal heartbeat, as Beethoven interpreted the mood of the moonlight into the rhythms of it without having any teacher but his God-Self.*

"*Consider the life of any great musical genius. What he finally produces in a written composition first begins in his Soul. He must first find aloneness with his Self by getting away from all other people, and even from his own body awareness. He must suppress all outer influences to seek that stillness from within from which the inner ears of his Soul can hear the rhythms of the universal heartbeat of God's creative thinking.*

"*To find that stillness in the Light of knowing he must stop thinking, in order that he can conceive idea from the Light of all-knowing. From the ecstasy of that stillness the inspiration is born in his very Soul. The motive for his symphony gradually takes form through the heartbeat of his thinking*

---

[27] Russell, [Home Study Course], Unit One - Lesson 3.1

which he extends from the stillness of his Soul, then ceases to extend it to again find the stillness of the Light which is gradually giving him his motive by its inspiring illumining within him."[28]

"*SECOND. Your desire is written upon your heart. It is registered in your Soul. It is no longer your body desire, it is your Soul's desire. You are in communion with God about it - not in words but in Light. You have an increasing sense of joyousness. Your desire is written into the universal heartbeat rhythmically. You are responding to it rhythmically. You are no longer just body - you are Soul - slightly aware of body - slightly aware of earth, and the heavens above, and of creating things. You are not thinking of extraneous things. You cannot think of objective things of earth while in the Spirit.*"[29]

"*The key to that answer is that God's universe is founded upon balance - one balanced condition which He divides into two equal and opposite conditions of His electric thinking. This electric universe of matter is composed entirely of these two equal and opposite conditions.*

"*Every effort of every creating thing is expressed by the motion of interchange between those pairs of opposite conditions, such as matter and space which constantly interchange equally by breathing into and out of each other equally. Without that equality of rhythmically repeated interchange the universe could not continue. That rhythmic interchange constitutes the heartbeat of God's body. Its rhythm and balance are absolute because the division of God's electric thinking is equally balanced.*"[30]

"*Meditation transports one from the transient world of matter to the real world of dreamings, visions, and imaginings where idea is and concepts are born. That is the world where sounds are heard in the silence of your Soul where no sound is where rhythms of symphony and poem are the rhythms of cosmic pulsings of God's thinking manifested in the heartbeat of His universe.*"[31]

---

[28] Russell, [Home Study Course], Unit One - Lesson 3.2

[29] Russell, [Home Study Course], Unit One - Lesson 4.1

[30] Russell, [Home Study Course], Unit One - Lesson 4.2

[31] Russell, [Home Study Course], Unit One - Lesson 4.4

"When you pray for God to show His love for you by giving your field good crops you must yourself give your love to your fields to regive His giving. Your love given to them is God's love given to you for regiving. Your love and God's love are ONE. Its manifestation is TWO in giving and regiving but that TWO is the ONE HEARTBEAT of God's body, even as the swinging of the pendulum is the TWO of the ONE source of their manifestation.

"He who knows the Light of love in him does not beseech the Father-Mother of the universe for his next breath, nor for his next heartbeat, for he fully knows that his heartbeat is one with the universal pulse, and that for which he would needlessly ask is already his.

"And as there is not a time when God centers you, and an interval elapses in which He again centers and controls your balanced heartbeat to synchronize with His, so is there not a time when you should pray to Him, and another time when you should not."[32]

"The love born symphony is prayer for love fulfilled in the composer and again fulfilled in the hearts of those whose heartbeat feels the ecstasy of God's heartbeat as expressed in those heavenly rhythms."[33]

"I instantly, and timelessly, knew the still magnetic Light which is the fulcrum of life and power, also I knew the heartbeat of the electric universe which manifests love, life and power in matter.

"I also instantly and timelessly knew that the magnetic Light of the God of love was all that is - and that the sensed electric wave universe of motion which simulates love, life and power, had no reality whatsoever. I knew it for what it was - God's thinking - God's imagining - pure illusion - simulation - self-voiding forms thrown on the screen of space to manifest changeless idea by setting it in motion to produce the effect of change.

"Thus knowing the universal heartbeat I knew the electric wave, or universal pulse, to which all moving, creating things were geared. Thus knowing the wave and its octave pulse beat in cyclic rhythms which spelled their rhythms

---

[32] Russell, [Home Study Course], Unit One - Lesson 6.1

[33] Russell, [Home Study Course], Unit One - Lesson 6.2

out for me always as

+zero +one +two +three +four +zero- -four -three -two -one -zero

I then knew all cause and comprehended all effect.

"*There was nothing to know but God's one whole idea of Creation, which is represented in the following diagram by the changeless ZERO of the spiritual Mind universe of REST - and nothing to comprehend but the heartbeat of this pulsing electric wave universe of motion which is represented in the diagram by the NUMBERS.*

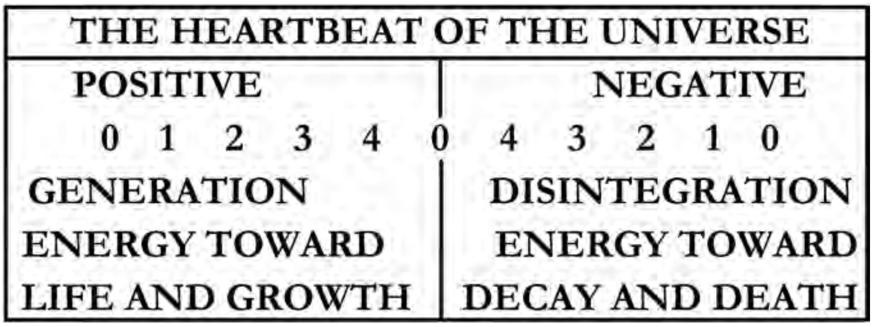

THE HEARTBEAT OF THE UNIVERSE	
POSITIVE	NEGATIVE
0 1 2 3 4 0	4 3 2 1 0
GENERATION	DISINTEGRATION
ENERGY TOWARD	ENERGY TOWARD
LIFE AND GROWTH	DECAY AND DEATH

Figure 11.3: The basic formula of Creation
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/positive-negative-scale.jpg
(courtesy University of Science and Philosophy)

"*This diagram is given to you now as the first simple step toward your acquisition of all knowledge of CAUSE and all comprehension of EFFECT. During the whole construction of this electric universe in this entire course of study the basic principle above diagramed will never be departed from, or exceeded, for there is nothing in Nature, nor in your own life, nor in your thinking, nor in any action of yours which can vary one whit from this basic formula of Creation.*

"*Nor is there anything the chemist, physicist, metallurgist or engineer can exceed beyond this formula or vary it in their work. Likewise the poet, painter, inventor, musician or architect is limited to this simplicity beyond which he CANNOT GO, and to which HE MUST CONFORM.*

"Likewise the healthy growth of your body, or the conduct of your friendly or business relations must conform to this principle of BALANCE or else suffer the consequence of whatever unbalanced residue remains from lack of conformity with the heartbeat of the universe."

"The reason why I have been able to do so many things during life, and do them in a masterly manner, was because of my lifetime growing knowledge of the universal heartbeat which fully unfolded in one timeless flash in 1921.

"It may not suddenly unfold to you, as it did to me, but you can very greatly accelerate its unfolding if you will open your heart to these lessons and get what I am telling you into your Consciousness instead of just photographing them into your brain."[34]

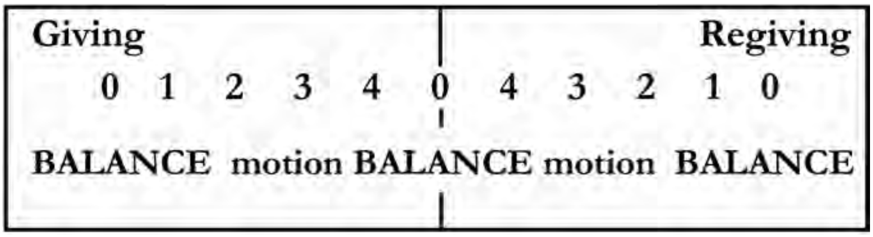

Figure 11.4: Formula of Rhythmic Balanced Interchange
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/giving-regiving.jpg
(courtesy University of Science and Philosophy)

"As you see I have again used the same formula of rhythmic balanced interchange in all transactions of Nature. I have but used different words to express different ideas."[35]

---

[34] Russell, [Home Study Course], Unit One - Lesson 7.3

[35] Russell, [Home Study Course], Unit One - Lesson 7.4

Chapter 12

Vortices and Toroids

12.1 Vortices

Vortices combine their many complex motions to create compound vortices or toroids composed of several parts and dynamics. In other words a toroid of rotating motion is the effect of vortex dynamics. These dynamics are the combined effects of differential densities[1], temperatures, pressures and tensions. These interactions between differences in density, potential, pressure, temperature and mass are the cause of motion which is the system attempting to regain balance in all its parts. These dynamic systems are in turn governed by simple laws as are discussed in this book. The forces creating vortices are the same forces creating vibration and oscillation but with different parameters.

---

[1] [differential densities], [2.7 - Differential Densities], [16.13 - Differential Densities], [3.6 - Differential Densities Begin to Form]

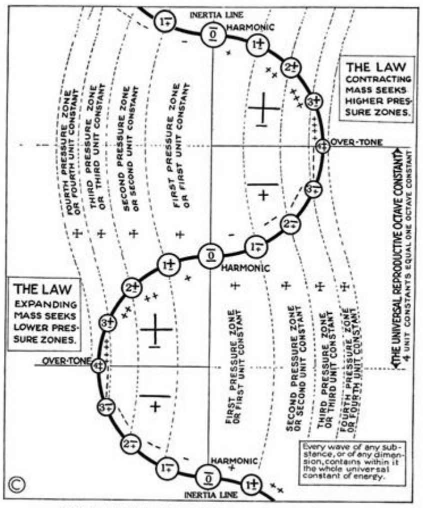
INERTIA LINE
HARMONIC
THE LAW
CONTRACTING MASS SEEKS HIGHER PRESSURE ZONES.
OVER-TONE
FOURTH PRESSURE ZONE OR FOURTH UNIT CONSTANT
THIRD PRESSURE ZONE OR THIRD UNIT CONSTANT
SECOND PRESSURE ZONE OR SECOND UNIT CONSTANT
FIRST PRESSURE ZONE OR FIRST UNIT CONSTANT
THE UNIVERSAL REPRODUCTIVE OCTAVE CONSTANT
4 UNIT CONSTANTS EQUAL ONE OCTAVE CONSTANT
HARMONIC
THE LAW
EXPANDING MASS SEEKS LOWER PRESSURE ZONES.
OVER-TONE
FIRST PRESSURE ZONE OR FIRST UNIT CONSTANT
SECOND PRESSURE ZONE OR SECOND UNIT CONSTANT
THIRD PRESSURE ZONE OR THIRD UNIT CONSTANT
FOURTH PRESSURE ZONE OR FOURTH UNIT CONSTANT
Every wave of any substance, or of any dimension, contains within it the whole universal constant of energy.
HARMONIC
INERTIA LINE
©

PRESSURE ZONE DIAGRAM. THE UNIVERSAL OCTAVE CONSTANT DIVIDES ITSELF INTO FOUR UNITS OF SEVEN TONES. EACH UNIT OF TWO TONES IS DIVIDED BY PRESSURE ZONES, WHICH CAUSE ALL MASS TO ROTATE AND TO REVOLVE IN SPIRAL ORBITS.

Figure 12.1: Differential densities, pressures and masses
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/differentials.jpg
(courtesy University of Science and Philosophy)

In examining the dynamics of a vortex it is obvious there are many subtle polar dynamics converging to create a typical vortex such as are witnessed in a tornado or hurricane. These polar striations[2] (flows of differing energy states) would include differences in pressure[3], mass, potential, velocities and tensions caused by the differences of these states.

"...vortex action (is) induced by differential conflict between the low and high tensions..."[4]

"All such experiments invariably resulted in vortex motion, whether induced sympathetically or otherwise. All corpuscular action in Nature is vortex motion. The undulatory theory[5] *of light is only hypothetical. The conditions of*

---

[2] [striations], [stria]

[3] [Figure 3.6.1 - Tornadoes also have a Down Draft caused by its vacuous center], [Figure 3.6 - Tornado Vacuums up Everything for Redistribution]

[4] Keely, [LIGHT - Snell]

[5] wave theory

electro-magnetic radiation, on the same plane of matter as light, disprove the undulatory theory in many particulars. The vortex action induced by differential conflict between the low and high tensions shows conditions analogous to those in the molecular dissociation of water[6] *into hydrogen and oxygen in other words, vortex action of the highest order but peripheral only. Were it otherwise, the ether could not be held suspended or enclosed in the molecular or atomic envelopes*[7]*.*"[8]

Figure 12.2: Tornado Vortex
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/tornado-hf.jpg

12.2 Vortex Components

1. Cone
2. Cone Base (large opening)
3. Cone Apex (Gravity Center, Neutral Center)
4. Syntropic Still Magnetic Light, Gravity or Mind Shaft (hollow center)
5. Etheric Vacuous Center
6. Entropic Rotating Periphery (striated with high and low pressures and tensions)
7. Pressure Striations
8. Entropic Equator 90° to formative forces

---

[6] [Aqueous Disintegration], [Part 15 - Dissociating Water], [AQUEOUS DISINTEGRATION - Snell], [THEORY AND FORMULA OF AQUEOUS DISINTEGRATION]

[7] [rotating envelope]

[8] [LIGHT - Snell]

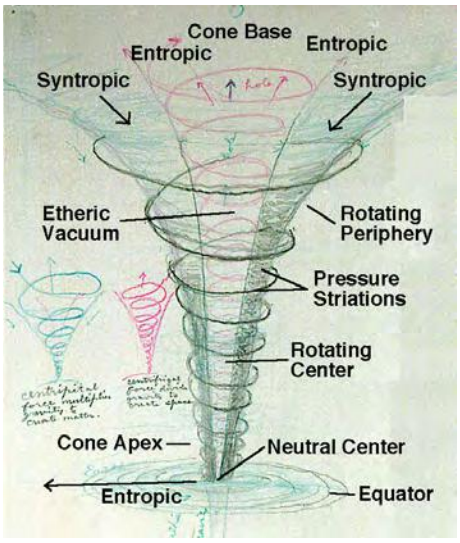
Cone Base
Entropic
Entropic
Syntropic
Syntropic
Etheric Vacuum
Rotating Periphery
Pressure Striations
Rotating Center
centripetal force multiplies gravity to create matter.
centrifugal force divides gravity to create space
Cone Apex
Neutral Center
Entropic
Equator

Figure 12.3: Vortex showing a few of its many components
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/vortex-nd.jpg
(courtesy University of Science and Philosophy)

12.3 Vortex Dynamics

Syntropic (Sympathetic) Formative Forces[9] attracting to center.
Entropic (Repellant) Dispersive Forces dispersing from center.
Neutral Center Dynamics (Dominates/controls above two dynamics).

12.4 Vortex Action is rotary

"All such experiments invariably resulted in vortex motion, whether induced sympathetically or otherwise. All corpuscular action in Nature is vortex motion." [Keely]

A vortex rotates on a divided horizontal plane along its extension, length or altitude from the ground to the clouds. The high and low tensions[10] (pressures, temperatures and densities) are in constant conflict seeking balance or voidance through and against each other. This assertive seeking for balance or voidance is the cause of the vortex's violent motions - a sex act if you will allow such term. The base universal state is 'desire for' balance, harmony or equilibrium. Desired because no state of perfect balance can permanently exist. *A system disturbed from this balance of forces will reassert itself in an effort to restore a stable state of balance or harmony (Love or sympathetic association) but such action is like a dog chasing its tail - and never catching it.*

---

[9] [Figure 4.12 - Keelys Formative Structural Dynamic Morphology]
[10] [antagonism]

12.5 Vortex Center

A tornado's center is an etheric vacuum having centripetal forces powerful enough to suction asphalt from paved roads[11] or lift cows, people and houses. This central vacuous column has been described as being luminous by those fortunate enough to have seen it and survived to tell about it. Walter Russell referred to this center as the Magnetic Still Light of Mind[12]. It may also be the fabled Luminiferous Ether[13] of the ancients.

"Know thou that God does not extend His Self into his moving universe, for the God-Light is still. Its stillness centers all things and it, likewise, centers the shafts of all motion which turns around it, shafts which are levers of fulcrums and end at poles which measure extensions."[14]

"Electricity (a polar condition) is the divider of the invisible white[15] *still magnetic light of gravity - which Mind is - into the red and blue halves of the visible light spectrum for the purpose of creating male and female bodies (in motion) which alone constitute all matter."*[16]

---

[11] [Figure 3.6 - Tornado Vacuums up Everything for Redistribution]

[12] [Magnetic Still Light]

[13] [Luminiferous Ether]

[14] [Atomic Suicide]?; Chapter 5 - Prelude - The Transformation of Man, part 2 of 2

[15] In Figure 12.4 the white light is shown as yellow. Hughes in her book "Harmonies of Tones and Colours" considered yellow as the Source from which discord (form) originates and progresses.

[16] Walter Russell, notation on chart, 1955. see [First Cause]

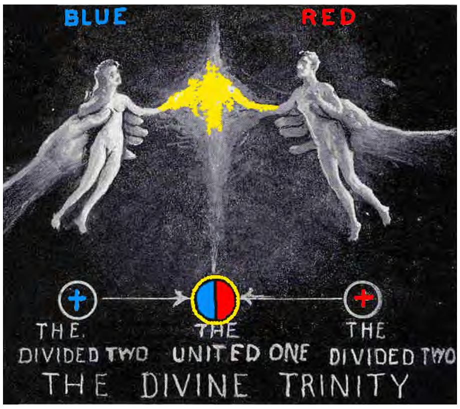

Figure 12.4: The depolar One divides into duality/polarity
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/female-male-color.jpg
(courtesy University of Science and Philosophy)

"The One begat the Two which begat the Three and from the Three all things are made." Lao Tse[17]

The two initial polar states being 1) to separate or be individual and 2) be Unified or One are nothing more than tendencies or desires. Entropy tends in a direction away from the center, under certain circumstances, *in a spiral path.* Syntropy tends to a direction towards the center, under certain circumstances, *in a spiral path.* Interestingly the two opposing states, forces or poles become each other intermittently[18]. Since there are only two states they are initially in seeming perfect equilibrium one against the other. There is no motion or substance with the desire and tendency to direction and motion held latent. In this state of perfect latency on the etheric level there is pure potential waiting to be transformed into motion and matter by introducing idea (desire) or disturbing excitation.[19] But such a static state cannot persist. On the molecular and atomic levels disturbance of this static equilibrium can occur with simple acoustic impulses.

---

[17] 'The Tao te Ching' [http://en.wikiquote.org/wiki/File:8orbitals.jpg]

[18] [Law of Assimilation]

[19] [2.11 - Beginning as Undifferentiated One]

12.6 Vortex Polar Forces

Russell's reference to - creating male and female bodies - refers to the still magnetic light as dividing itself creating the two seeming polarities - the two seeming opposing forces of entropy and syntropy whose mutual conflict[20] to void or balance results in the motions of the vortex. The two poles' prime objective is to void against/through each other that they may regain the Oneness (balance, harmony or love) from whence they came. These are the dynamics of the Universal Heartbeat, Rhythmic Balanced Interchange or Dynaspheric Force. In the simplest terms *the vortex is comprised of three currents acting as a whole stream or flow* - the same as has been described in Chapter 7 and as shown in Figure 12.5.

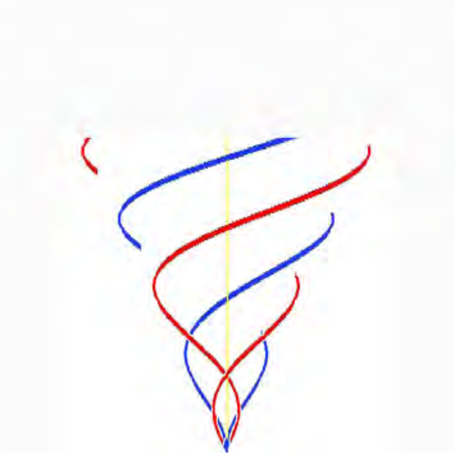

Figure 12.5: Triune Forces of a vortex
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/spiral-stria.jpg

12.7 Vortex Periphery

The periphery is molecular rotation governed by centrifugal forces counterbalanced by centripetal forces. In this section of the vortex molecular substances (gases, evacuated debris, *etc.*) *orbit* around the vortex rotating center at high velocities and constitutes the visible form of the vortex.

---

[20] [antagonsim]

12.8 Vortex Wall

The wall or border condition between the still magnetic light and whirling molecular gases is where centripetal forces (syntropy) are transformed to centrifugal energies (entropy). Obviously this conversion zone is created by a dynamic balance between the inward pulling centripetal or syntropic forces of the vacuous etheric center and the outward flinging centrifugal entropic forces maintained by sufficient velocities.[21]

If we picture three tornadoes rotating at 90° to each other would the combined forces tend enclose themselves to create a sphere much like Keely's description of corpuscles formed via rotating envelopes[22]? Of course they would.

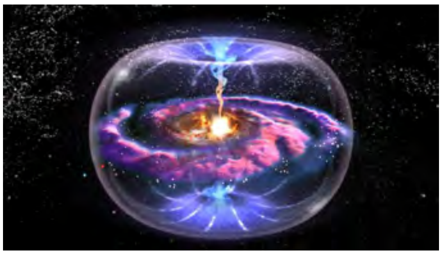

Figure 12.6: Typical idealized torus, toroid or "donut" concept
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/donut-idealized.jpg
(courtesy anonymous web artist)

12.9 Toroid

Idealized toroids are essentially composed of two opposing vortices meeting at their two apices forming a mutual center of neutrality, voidance or transformation. Being resultants of vortex structural components and dynamics they exist by virtue of these defining attributes. Of course, this is a preliminary study and does not include all the myriads of structural and dynamical properties of all the different vortex forms. Please note these dynamics are inherent causative properties and not a mathematical description of phenomena which would be elegant in and of itself[23] yet just a facsimile of what is really taking place.

---

[21] [Law of Cycles], [Law of Transformation of Forces]
[22] [rotating envelopes]

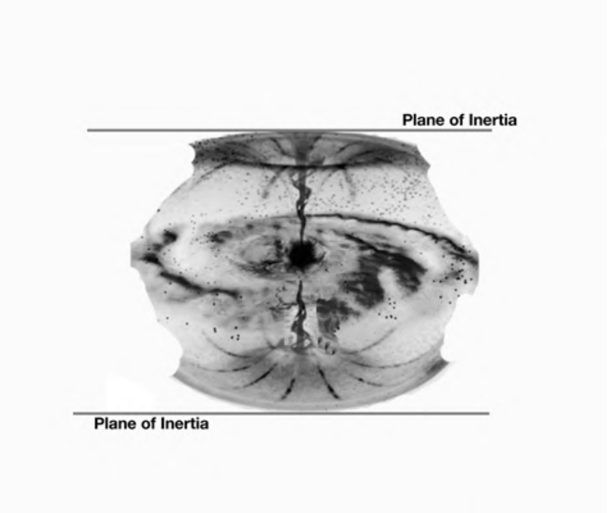

Figure 12.7: Ideally toroids are two vortices meeting apex to apex
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/donut-inverted-mod.jpg

12.9.1 Sympathetic Associative Attraction

The two apices come together by virtue of sympathetic associative attraction or syntropy according to the Law of Harmony. They meet in this fashion at Russell's 4++ wave crest where the Law of Cycles kicks in and the Law of Repulsion assumes preponderance.

---

[23] for a good example: http://vortexmath.webs.com/

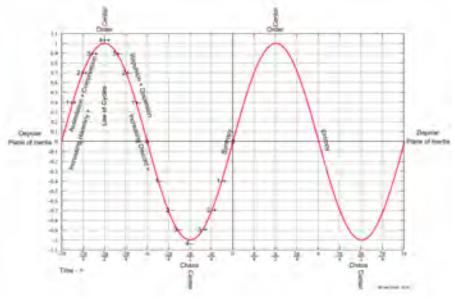

Figure 12.8: Wave elements of Time, Compression and Dispersion
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/sine-indig-wave1.jpg

Figure 12.8 shows an idealized structure of a wave being a sine wave. Some of the functional elements are included for identification purposes. In reality the left upward moving slope is syntropic (Law of Harmony) and will usually be a longer time interval than from upper Center to lower Center (Law of Repulsion) which time may be greatly reduced. This is the time interval when compression (harmony) reverts to dispersion (repulsion) according to the Law of Cycles and may be of negligible duration. Hence the overall sine wave form will become a sawtooth wave as harmonics are generated and synthesized into the wave form according to the same law. An excellent animation of this process can be found on Wikipedia/Sawtooth_wave here:

http://en.wikipedia.org/wiki/Sawtooth_wavemediaviewer/File:Synthesis_sawtooth.gif

12.9.2 Law of Cycle

This sympathetic associative attraction increases until the conditions for the Law of Cycles to assume dominance at which point the sympathetic associative attraction reverts to anti-sympathetic mutually repellant discords[24] of increasing amplitudes. These discords will increase until their power of repulsion exceeds the power of attraction at which point the associating forces become repellant and the center discharges at 90° to the formative sympathetic associative forces. Thus forming the expanding energies outward through the equator. Russell proposed a four step compression to the 4++

---

[24] By formation of secondary and tertiary pitches not harmonic with the fundamental.

point of transformation discussed in previous chapters[25].

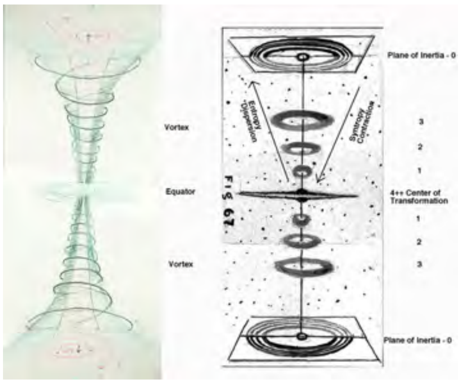
Plane of Inertia - 0
Vortex
Entropy Dispersion
Syntropy Contraction
3
2
1
Equator
FIG 67
4++ Center of Transformation
1
2
3
Vortex
Plane of Inertia - 0

Figure 12.9: Compression to center then equatorial dispersion
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/vortex-side-by-side-sm.jpg
(courtesy University of Science and Philosophy)

12.9.3 Neutral Center Repulsion

The now highly anti-sympathetic (enharmonic) centralized bipolar center goes 'critical' so to speak when the participant substances mutually repel each and all other participant substances as they override the harmonic syntropic forces of attraction and compression. The mutual repulsion occurs 90° to the sympathetic associative forces outwards towards and through the equatorial plane.

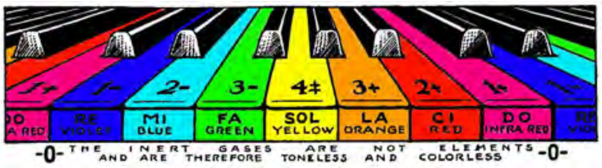
1+
1-
2-
3-
4‡
3+
2+
1+
DO
RE VIOLET
MI BLUE
FA GREEN
SOL YELLOW
LA ORANGE
CI RED
DO INFRA RED
-0- THE INERT GASES ARE NOT ELEMENTS -0-
AND ARE THEREFORE TONELESS AND COLORLESS

Figure 12.10: 4++ at center of transformation
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/4plusplus.jpeg
(courtesy University of Science and Philosophy)

---

[25] Figure 10.3: Scale of 0 to 4 to 0

12.9.4 Periphery Neutralization or Depolarization

As these substances move away from the center and nearing the end of influence of sympathetic outreach they begin losing their polar condition or state until such point as they become again depolarized or neutralized. The center represents Russell's 4++ state. The polar substances becoming depolar at the 0th state of Russell's Scale of Locked Potentials.[26] Beginning at the vortex center with 4++ maximum polarization expanding outward (dispersing according to the Square Law) gradually losing polarization until all polarization is lost (depolarization) at the 0th state of Russells Scale of Locked Potentials which he defines as the Cube Wall of Neutrality, Plane of Inertia or the limit of Keely's sympathetic outreach[27]. At this Plane of Inertia a reversion takes place where the depolar state becomes polarized and begins the journey back to the 4++ state or condition.

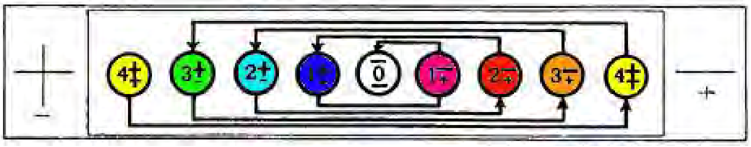
No state of motion ever began or ever ended.

Figure 12.11: Depolar state at plane of inertia
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/0-inertia.jpeg
(courtesy University of Science and Philosophy)

12.10 Plane of Inertia

We can see from photographs of hurricanes the tops are generally flat. These flat tops appear to coincide with Russell's plane of inertia which coincides with the wall of his cube structures. Physically this is a zone where there is no polarity or is depolar. This extending downward as polarity increases to where the maximum (4++) zone is reached at or near the surface of the earth where most polarity occurs and hence the most violent action or motion.

---

[26] [12.01 - Scale of Locked Potentials]

[27] [sympathetic outreach]

12.11 Free Space

With hurricanes and tornados the plane of inertia is in higher altitudes if not space itself and the vertex at the surface of the earth are mostly molecular structures and are bound by gravity and other terrestrial dynamics. In the case of free forming quantum entities these bounds do not exist as these formations form in etheric space or vacuum. According to Keely and Russell there are two vortices per axis coming together at their mutually attractive centers or vortices making six in number. Do all six have the exact same wave function? No, else they would all be exactly the same which is an impossibility.

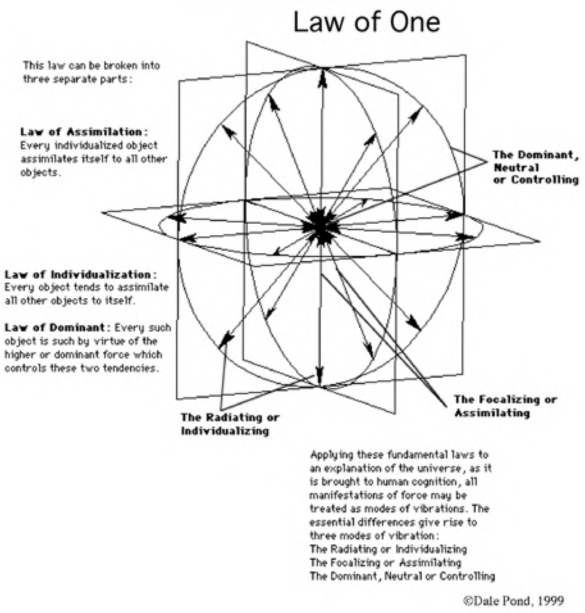

This law can be broken into three separate parts:

Law of Assimilation: Every individualized object assimilates itself to all other objects.

Law of Individualization: Every object tends to assimilate all other objects to itself.

Law of Dominant: Every such object is such by virtue of the higher or dominant force which controls these two tendencies.

Applying these fundamental laws to an explanation of the universe, as it is brought to human cognition, all manifestations of force may be treated as modes of vibrations. The essential differences give rise to three modes of vibration:
The Radiating or Individualizing
The Focalizing or Assimilating
The Dominant, Neutral or Controlling

©Dale Pond, 1999

Figure 12.12: Triune polar forces as One Force about a common center
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/triune-modes-motion.jpg

Following the logic of Russell's cosmogony there are virtual 'cubes' in Free Space having six sides. These six sides are the planes of inertia or depolarization. These correspond to Keely's juxtapositioning of atomolic centers of mutual attraction which would be an alignment with the X, Y and Z vectors each forming their own respective virtual cubes as these would align with the planes of said virtual cubes. Therefore there would be three of these double vortices forming along these vectors, each 90° to the others. Each would possess its own center line of mutual attraction. At the compound center there would therefore be three atomolic (etheric) centers of mutual attraction and six vortices.

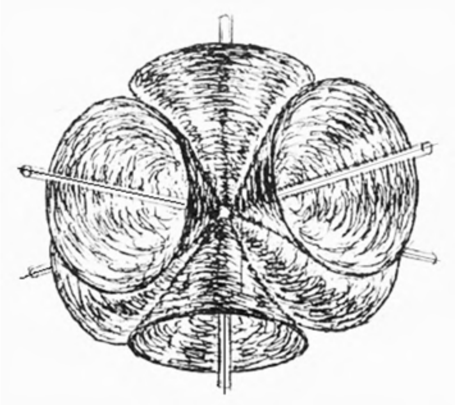

Figure 12.13: Three double vortices
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/3d-vortices.jpeg
(courtesy anonymous web artist)

According to Keely these atomolic centers bear the burden of all structure and motion it controls. Consequently each of the three centers is autonomous and independent of the other two though the three mutually interact to some degree and manner yet to be determined as they mutually occupy the same coordinates at the center of the structure.

Therefore a quantum or etheric torus would have three orthogonal axes each with two mutually attracted vortices. This compound torus would in structure be a compound spheroid with a surface of complex motions or encapsulated rotating envelops[28] so long as their three centers are coincident.[29]

---

[28] [rotating envelope]

[29] [Etheric Capsule], [AFFINITY OF ETHERIC CAPSULES], [DISINTEGRATION OF ETHERIC CAPSULE], [ROTATION OF THE ETHERIC CAPSULE], [MEASURING THE VELOCITY OF THE ETHERIC CAPSULE]

12.12 Etheric Envelopes

Here is one of the concepts where Keely and Russell diverge in their theories. Russell has the three planes and six vortices. Keely has the three planes as axis where the etheric substance rotates at 90° to the axis resulting in an envelope or shell spinning at tremendous velocities establishing the circumference of the centralization[30]. This results in three envelopes or shells spinning at 90° to each other thus forming the entirety of the corpuscle. Each shell acts together geometrically to fully enclose the corpuscle as they seal each others' non-rotating poles.

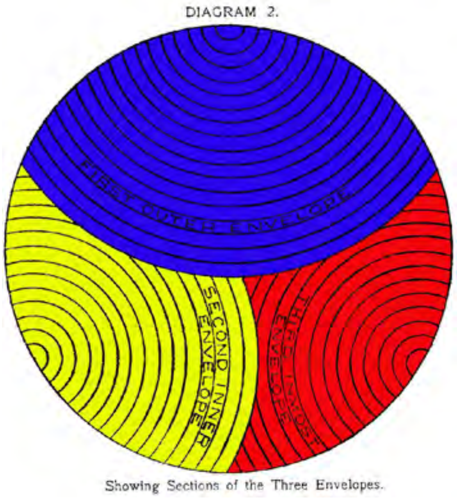

Showing Sections of the Three Envelopes.

Figure 12.14: Keely's three envelopes or shells
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/envelopes.jpg

---

[30] [Figure 7.3 - Step 3 - Sphere Forms Orthogonally Triple Compressing Shell Layers]

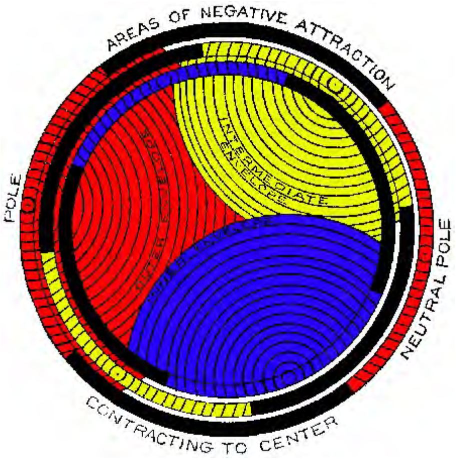

Figure 12.15: Three overlapping envelopes or shells sealing poles
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/sealedpoles.jpg

This configuration[31] is fundamentally important to Keely's work with splitting the atom or molecule releasing its latent forces bound within. He said the ether (quantum substances) when given sufficient velocity will rotate and centripetally seek its controlling neutral center. The greater the velocity the greater the syntropic force the greater the compression inward and greater the contained latent energy to be liberated. He did this by manipulating these envelopes or shells rupturing them whether instantly or gradually.[32][33][34]

It is not the purpose of this book to go into the details of how Keely manipulated these molecular and atomic envelopes (splitting the atom) to produce his etheric vapor and other forces and quantum substances. To do so would require a doubling or more of its volume (and time to write) and it is already near too much to digest. Perhaps a subsequent volume would be appropriate covering his dissociation and disintegration processes[35] with water and other elements.

---

[31] [Figure 4.12 - Keelys Formative Structural Dynamic Morphology]

[32] [15.18 - Keelys Process for Liberating Ether from Water]

[33] [Vibratory Physics - True Science]

[34] [Etheric Vapor]

[35] [Aqueous Disintegration], [THEORY AND FORMULA OF AQUEOUS DISINTEGRATION], [DISINTEGRATION OF ETHERIC CAPSULE], [Disintegration], [Dissoci-

---

ation], [DISINTEGRATION OF MATTER - THREE SYSTEMS]

Chapter 13

Electricity and Magnetism

13.1 Electricity is a Polar State Dipole

Contrary to popular belief the common generator of electricity flow or current does not generate electrons then pushes them along a conductor. The generator creates a polarity, potential, dipole or bias between the positive electricity and negative electricity streams where prior to this condition of separation neither pole seemed to exist. This creation of this bias or potential is in fact a differentiation or splitting into three flows of an undifferentiated compound triune flow.

"I believe electricity to be a substance, not a force."[1]

"Electricity is the result of three differentiated sympathetic flows, brought together by combining the celestial flow with the terrestrial flow through a certain degree of "negative attractive assimilation". Electricity is one of Nature's efforts toward restoring "attractive differentiation" for it has the highest degree of assimilative affinity."[2]

13.1.1 Voiding the Dipole or Polarity

Once the two (dipole) electric streams are created, thus creating a potential, they have a natural tendency to re-emerge or assimilate each other through voiding as Russell calls it and neutralization or depolarization as Keely calls it. This voiding, neutralization or depolarization is accomplished by providing a conductor and a place or load where the separated streams can and will recombine, re-emerge, unite, assimilate or void the created polarity, dipole or bias - returning back to their natural undifferentiated state. Heat, light and rotation are produced at the point of voiding, neutralization or depolarization by the interaction (antagonism) of these streams against each other.

---

[1] Keely quoted in [Pittsburg Dispatch April 07-1890]

[2] [Keely in SYMPATHETIC STREAMS - Snell]

13.1.2 Conduction

The conduction of the electric current along a conductor is governed by Keely's Law of Electric Conductivity.

13.1.3 Polar and Depolar States

An electric dipole is mutually attractive and will do everything possible to neutralize itself.

A magnetic dipole is mutually repulsive and will retain its polar state.

The difference between these two states of attraction and repulsion is explained in Keely's Law of Attraction and Law of Repulsion - governed by or as explained in the Laws of Being. See also [Bjerknes Effect]

13.2 Keely on Electricity

Should we bring our understanding from previous chapters to bear upon electricity and magnetism a paradigm forms somewhat different than the standard model but which appears to make sense and evidently works as Keely produced electricity from etheric space (vacuum) in 1893.[3]

"Nikola Tesla, for reasons that need not be explained, would not have been able to serve on the committee. It is well known that he is striving to draw power from space by purely scientific methods, which Keely succeeded in doing

---

[3] [Dogmatism of Science], [Keely - Electricity from Space], [Newton of the Mind], [Free Energy], [The Operation of the Vibratory Circuit]

in 1893 by "unscientific methods."[4]

"Electricity is the result of three differentiated sympathetic flows, combining the celestial and terrestrial flows by an order of assimilation negatively attractive in its character. It is one of Nature's efforts to restore attractive differentiation. In analyzing this triple union in its vibratory philosophy, I find the highest order of perfection in this assimilative action of Nature. The whole condition is atomic, and is the introductory one which has an affinity for terrestrial centres, uniting magnetically with the Polar stream; in other words, uniting with the Polar stream by neutral affinity. The magnetic or electric forces of the earth are thus kept in stable equilibrium by this triune force, and the chords of this force may be expressed as 1st, the dominant, 2nd, the harmonic, and 3rd, the enharmonic. The value of each is, one to the other, in the rates of figures, true thirds. E flat - transmissive chord or dominant; A flat - harmonic; A double flat - enharmonic. The unition of the two prime thirds is so rapid, when the negative and the positive conditions reach a certain range of vibratory motion, as to be compared to an explosion. During this action the positive electric stream is liberated and immediately seeks its neutral terrestrial centre, or centre of highest attraction."

---

[4] Bloomfield-Moore in [Dogmatism of Science], [Keely - Electricity from Space]

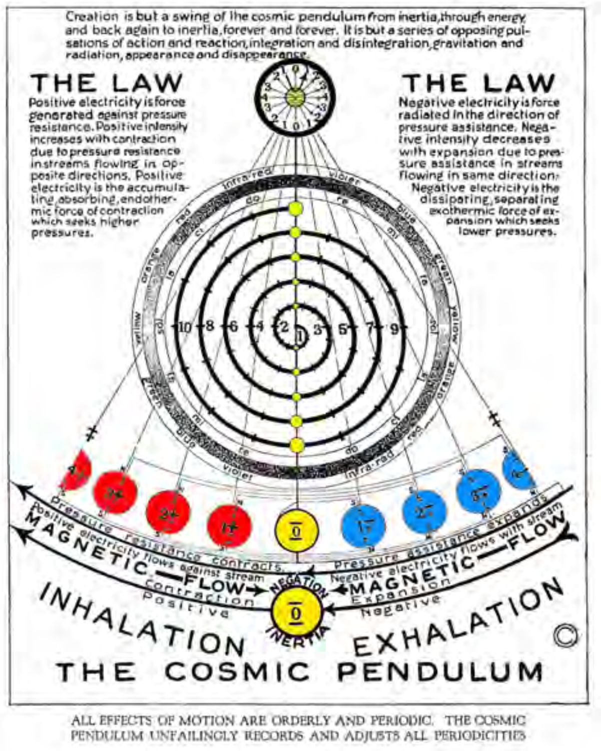

Creation is but a swing of the cosmic pendulum from inertia, through energy and back again to inertia, forever and forever. It is but a series of opposing pulsations of action and reaction, integration and disintegration, gravitation and radiation, appearance and disappearance.

THE LAW

Positive electricity is force generated against pressure resistance. Positive intensity increases with contraction due to pressure resistance in streams flowing in opposite directions. Positive electricity is the accumulating, absorbing, endothermic force of contraction which seeks higher pressures.

THE LAW

Negative electricity is force radiated in the direction of pressure assistance. Negative intensity decreases with expansion due to pressure assistance in streams flowing in same direction. Negative electricity is the dissipating, separating exothermic force of expansion which seeks lower pressures.

infra-red
violet
do
re
mi
fa
sol
la
ci
red
orange
yellow
green
blue
10
8
6
4
2
1
3
5
7
9
4+
3+
2+
1+
0
1−
2−
3−
4−
Pressure resistance contracts.
Positive electricity flows against stream
Pressure assistance expands
Negative electricity flows with stream
MAGNETIC — FLOW
Contraction
Positive
MAGNETIC — FLOW
Expansion
Negative
NEGATION
INERTIA
INHALATION
EXHALATION
©
THE COSMIC PENDULUM

ALL EFFECTS OF MOTION ARE ORDERLY AND PERIODIC. THE COSMIC PENDULUM UNFAILINGLY RECORDS AND ADJUSTS ALL PERIODICITIES

Figure 13.1: Neutral Center governing seeming opposite potential states
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/cosmic-penduluum.jpg
(courtesy University of Science and Philosophy)

"The power of attractive vibration of the solar forces is the great coincident towards which the terrestrial magnetic sympathetic flow is diverted. This force is the celestial current that makes up the prime third of the triple association. It also induces aqueous disintegration[5] *and thermal concentration, the two prime conductors towards this coincident chord of sympathy with itself. Without this aqueous disintegration there would be no connective link*[6] *between the celestial and terrestrial. There would exist nothing but a condition of luminous radiation on the order of the aurora - a reaching out for the concordant without any sympathetic diversion to create unstable equilibrium of terrestrial magnetism. In fact under such a condition, the absence*

---

[5] [aqueous disintegration]

[6] [connective link]

of the sun on one side, or the absence of water on the other, the magnetic or electric force would remain in a stable state of equilibrium, or the highest order of the chaotic. Disturbance of equilibrium[7] *and sympathetic equation*[8] *constitute the dual power that governs all the varied forms of life and motion which exist terrestrially, of which the electric or magnetic is the prime mover and regulator. All electrical action, no matter of what character, has its sympathetic birth by the intervention of that current of the triune flow, which I call the dominant, with the polar harmonic current; all sympathetic flows being composed of three currents. They become associative one with the other only near the junction of terrestrial interference. The great vacuous field which exists between the planetary ranges holds this portion of the etheric flow free of all antagonism*[9]*, molecularly or otherwise, till the associative point is reached; so wonderfully planned by the Great Creator, for instant electric evolution and assimilation with terrestrial centres of attraction. I call this intervention, atomic intermolecular and molecular density. The combination of the action of the triune sympathetic celestial stream with the same intervening medium induces heat and light as the resultant of these corpuscular confliction with sympathetic celestial and terrestrial focalized centres of neutral radiation. I do not recognize electricity, nor light, nor heat as coming from the sun. These conditions, according to my theories, emanate from atomic and interatomic interference on induced molecular vibration, by sympathetic etheric vibration, the celestial attractive being the prime mover. In my estimation this is not at all phenomenal; it is only phenomenal as far as the knowledge of its action in mechanical physics is concerned. Physicists have been working in the wrong direction to lead them to associate themselves with Nature's sympathetic evolution. The expression "Electricity attracts at a distance" is as bad as, if not worse than, the "microbe of the magnet." Clerk Maxwell seems, when theorizing on sound transmission by an atmospheric medium, not to have taken into consideration the philosophy attending the phenomena of the origination of electric streams in celestial space. Light is one of the prominent evolved mediums in electric action, and is evolved by corpuscular bombardment*[10] *induced by sympathetic streams acting between the neutral centres of planetary masses, all of which are under a condition of unstable equilibrium. These unstable condition were born in them, and were*

---

[7] [Disturbance of equilibrium]

[8] [sympathetic equation]

[9] [antagonism]

[10] [INTERATOMIC BOMBARDMENT], [bombardment]

thus designed by the Architect of Creation in order to perpetuate the connective link between the dispersing positive and the attractive negative. The action that induces this link I call sympathetic planetary oscillation.[11]

"Light is induced by electromagnetic percussion emanating from the ether, and in its action represents the plane of magnetism. In fact, it is the plane of magnetism when under polarization. (Platina wires the thickness of a fine hair associated with each of the nine nodal beads, and concentrated towards a general center of focalization, attaching the other end of the wires to the focal center, will determine, by the magnetic conduction, the number of corpuscular oscillations per second induced by a thought, either positive or negative, in the central centers. These are the only conditions - those of magnetic conduction - whereby the evolution of a thought can be computed in regard to its force under propagation, as against the amount of latent energy set free to act as induced by such thought on the physical organism.) Some scientific theories of the past have taught us that electricity and magnetism are one and the same thing. Sympathetic vibratory philosophy teaches that they are two distinct forces of one of the triune sympathetic family."[12] (underline added)

"Maxwell's theory[13] *is correct that the plane of polarized light is the plane of magnetic force. The sympathetic vibrations associated with polarized light constitute the pure coincident of the plane of magnetism. Therefore, they both tend to the same path, for both are interatomic, assimilating sympathetically in a given time, to continue the race together, although one precedes the other at the time of experimental evolution. The time is approaching when electromagnetic waves with an outreach of two feet will be produced, having an energy equal to that now shown up on the magnet when it is about to kiss its keeper, and showing a radiating force too stupendous for actual measurement."*[14]

"Professor Fitzgerald's[15] *lecture on electromagnetic radiation shows that sci-*

---

[11] [Vibratory Physics - The Connecting Link between Mind and Matter]

[12] Snell, [Snell Manuscript]

[13] [Maxwell Equations]

[14] [Snell], [Snell Manuscript]

[15] [Professor George Francis Fitzgerald, Irish professor 1851-1901, specializing in electromagnetic theory and physics.]

entific men are beginning to realize and that fairly, the truths appertaining to the new philosophy. The professor admits that electricity and magnetism are of differential character and he is right. The progressive subdivision, induced by molecules by different orders of sympathetic vibration, and the resultant conditions evolved on the intermolecule[16] *and interatom*[17]*, by introductory etheric dispersion, prove that the magnetic flow of itself is a triple one, as is also the electric."*[18]

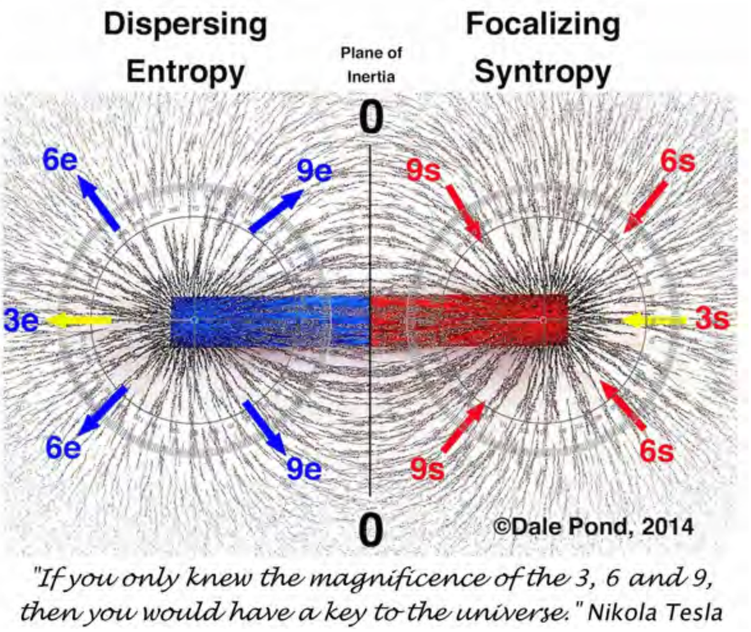
Dispersing
Entropy
Plane of
Inertia
Focalizing
Syntropy
0
6e
9e
9s
6s
3e
3s
6e
9e
9s
6s
0
©Dale Pond, 2014
"If you only knew the magnificence of the 3, 6 and 9, then you would have a key to the universe." Nikola Tesla

Figure 13.2: Triune Forces of Magnetic Field in Magnet
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/Forces.jpg

Keely says the iron of the magnet has no magnetism of itself. The iron atoms of the magnet diverts a portion of the terrestrial streams. In the above image this diversion is clearly shown where the streams come into the South pole and exit via the North pole. Also shown are the two modes of Syntropy (s), Entropy (e) as also the 3, 6 and 9 modes. His explanation is:

---

[16] [intermolecule]

[17] [interatom]

[18] [Snell], [Snell Manuscript]

"The explanation of the magnetic flow is very simple. The harmonic attractive chord, thirds, induces nodal interference on the harmonic current in the earth's polar stream, and it moves toward the negative pole to flow out through the positive end. The diversion of the harmonic portion of the stream from the dominant portion causes the magnet to assume its position."[19]

"Again, the professor says that electricity and magnetism would be essentially interchangeable if such a thing existed as magnetic conduction, adding: "It is in this difference that we must look for the difference between electricity and magnetism." Thus you see how plain it is that progressive scientists are approaching true science. The rotation of the magnetic needle[20]*, as produced in my researching experiments, proves conclusively that the interchange spoken of, in Professor Fitzgerald's lecture, is a differentiated vibratory one, in which the dominant and enharmonic forces exchange compliments with each other, in a differential way, thus inducing rotation, in other words polarization and depolarization."*[21]

"In electric lighting, the velocity of the dynamos accumulates only the harmonic current - by atomic and interatomic conflict - transferring one-two hundred thousandth (1/200,000) of the light that the dominant current would give, if it were possible to construct a device whereby it could be concentrated and dispersed."[22]

13.3 Russell on Electricity

"Electricity is the servant of the God-Mind. Electricity expresses the desire in the God-Mind for creative expression by seemingly dividing the One still light into transient waves of spectrum divided positive-negative colors of light."[23]

"Electricity is the motivative force which projects the One Light of Mind two ways to create cycles of light waves for the purpose of expressing thought

---

[19] [POLARIZATION AND DEPOLARIZATION]

[20] [4.2.5 - Rotation of Compass Needle], [DISTURBANCE OF MAGNETIC NEEDLE]

[21] [Snell], [Snell Manuscript]

[22] [Snell], [Snell Manuscript]

[23] Russell, [The Secret of Light], page 50

cycles."[24] (underline added)

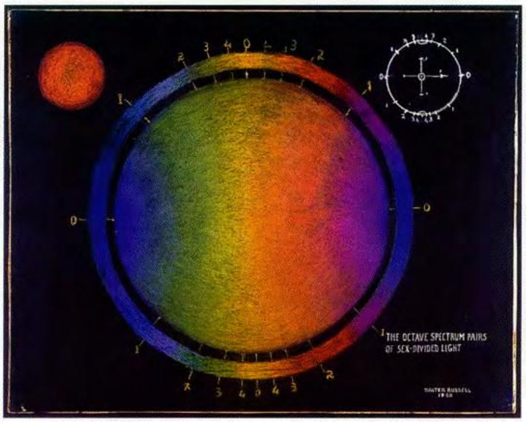

Figure 13.3: White Light of Mind divided into red and blue halves
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/red-blue.jpg
(courtesy University of Science and Philosophy)

"Electricity is the divider of the invisible white still magnetic light of gravity - which Mind is - into the red and blue halves[25] *of the visible light spectrum for the purpose of creating male and female bodies which alone constitute all matter."*[26]

"As servant of Mind, electricity gives moving form to the idea by performing the work necessary to produce it."[27]

"Electricity is forever winding light up into hot spheres, surrounded by cold cube wave-fields of space, and likewise sequentially unwinding them for rewinding. While winding them into compressed light masses, they are simultaneously unwinding to a lesser extent. Conversely, while unwinding they are simultaneously rewinding to a lesser extent."[28]

---

[24] Russell, [The Secret of Light], page 57
[25] [chemism], [rad-energy]
[26] Walter Russell, notation on chart, 1955.
[27] Russell, [The Secret of Light], page 114
[28] Russell, [The Secret of Light], page 124

"The farther removed from rest, the greater the strain or tension. That which we call high electric potential is merely great strain to maintain a condition which is far from the condition of rest."[29]

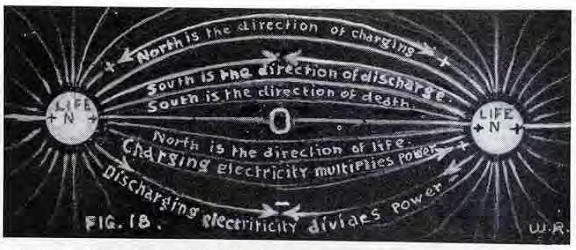
North is the direction of charging
South is the direction of discharge.
South is the direction of death.
LIFE N
0
LIFE N
North is the direction of life.
Charging electricity multiplies power
Discharging electricity divides power
+
−
FIG. 18.
W.R.

A study of Fig. 18 together with the bar magnet diagrams and other polarity diagrams shown in Figs. 11 to 16 will prepare you for a better understanding of the nature of electricity and the electric current.

Figure 13.4: Attractive (syntropic) power of electricity
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/attractive-electricity.jpg
(courtesy University of Science and Philosophy)

"Electricity is the strain or tension set up by the two opposing desires of universal Mind thinking - the desire for balanced action and the desire for rest."[30]

"Electricity creates Life by dividing stillness into two-way [polar] motion. It creates death by rest (depolar) from motion."[31]

---

[29] Russell, [The Secret of Light], page 139
[30] Russell, [The Secret of Light], page 139
[31] Russell, [Atomic Suicide]?; Figure 9 and 10, page 88]

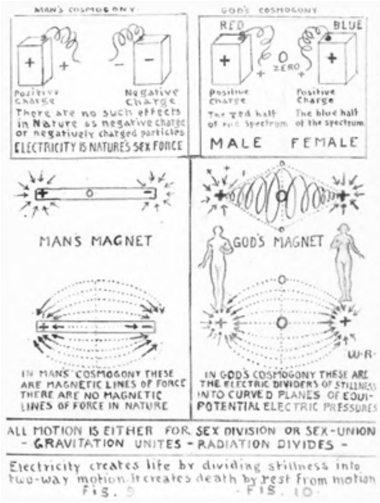
MAN'S COSMOGONY
GOD'S COSMOGONY
+
+
−
−
Positive Charge
Negative Charge
There are no such effects in Nature as negative charge or negatively charged particles
ELECTRICITY IS NATURE'S SEX FORCE
RED
BLUE
+
+
0
ZERO
+
+
Positive Charge
Positive Charge
The red half of the Spectrum
The blue half of the spectrum
MALE
FEMALE
MANS MAGNET
GOD'S MAGNET
IN MANS COSMOGONY THESE ARE MAGNETIC LINES OF FORCE THERE ARE NO MAGNETIC LINES OF FORCE IN NATURE
IN GOD'S COSMOGONY THESE ARE THE ELECTRIC DIVIDERS OF STILLNESS INTO CURVED PLANES OF EQUI-POTENTIAL ELECTRIC PRESSURES
W·R·
ALL MOTION IS EITHER FOR SEX DIVISION OR SEX-UNION
- GRAVITATION UNITES - RADIATION DIVIDES -
Electricity creates life by dividing stillness into two-way motion. It creates death by rest from motion
FIG. 9
FIG. 10

Figure 13.5: Polar dynamics of electricity
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/fig-9-10-88-bw.jpg
(courtesy University of Science and Philosophy)

"Electricity is the strain or tension set up by the two opposing desires of universal Mind thinking - the desire for balanced action and the desire for rest."[32]

"As rest cannot be unbalanced save by illusion, electricity which causes that illusion has no existence."[33]

"Electricity is the servant of Mind. It does all of the work of creating this light-wave universe in unfolding-refolding sequences which Mind desires. The universal Mind has two desires - the desire for creative expression through the action of concentrative (polar) thinking and the desire for rest from action through decentrative (depolar) thinking."[34]

---

[32] Russell, Walter; [The Secret of Light], page 139

[33] Russell, [The Secret of Light], page 141

[34] Russell, [The Secret of Light], page 144

13.4 Tesla on Electricity

"The day when we shall know exactly what electricity is will chronicle an event probably greater, more important, than any other recorded in the history of the human race."[35]

Below is a quote from William Lyne's book[36] discussing Tesla's ideas about electricity and magnetism. We will see Tesla's ideas were decidedly different than Keely's or Russell's. Incidentally Tesla knew of Keely and he knew Russell personally s mentioned in Russell's 'Atomic Suicide?' book.

In his 1891 A.I.E.E. lecture at Columbia College, Nikola Tesla said in pertinent part: *''What is electricity, and what is magnetism "...We are now confident that electric and magnetic phenomena are attributable to the ether, and we are perhaps justified in saying that the effects of static electricity are effects of ether in motion". "...we may speak of electricity or of an electric condition, state or effect". "...we must distinguish two such effects, opposite in character neutralizing each other". "...for in a medium of the properties of the ether, we cannot possibly exert a strain, or produce a displacement or motion of any kind, without causing in the surrounding medium an equivalent and opposite effect." "...its condition determines the positive and negative character." "We know that it acts like an incompressible fluid;" "...the electro-magnetic theory of light and all facts observed teach us that electric and ether phenomena are identical." "The puzzling behavior of the ether as a solid to waves of light and heat, and as a fluid to the motion of bodies through it, is certainly explained in the most natural and satisfactory manner by assuming it to be in motion, as Sir William Thomson*[37] *has suggested." "Nor can anyone prove that there are transverse ether waves emitted from an alternate current machine; to such slow disturbances, the ether, if at rest, may behave as a true fluid."*

In his statements, Nikola Tesla was balancing the various arguments in preparation for his decision: *"...Electricity, therefore, cannot be called ether in the*

---

[35] [Nikola Tesla]

[36] Summation of Tesla's Dynamic Theory of Gravity; An excerpt from: Occult Ether Physics by William R. Lyne

[37] [William Thomson], 1st Baron Kelvin OM, GCVO, PC, PRS, PRSE, (26 June 1824 - 17 December 1907) was a Belfast-born British mathematical physicist and engineer.

broad sense of the term; but nothing would seem to stand in the way of calling electricity ether associated with matter, or bound ether; or, in other words, that the so-called static charge of the molecule is ether associated in some way with the molecule."

'*'...It cannot differ in density, ether being incompressible: it must, therefore, be under some strain or in motion, and the latter is the most probable.*" Nikola Tesla obviously believed in an ether (quantum substances) which was in motion relative to earth, because the earth is in motion.

The thing which Nikola Tesla had realized, was that ether possesses electric charges which are deposited on or contained within atoms. In supporting the "dynamic" ether concept, he was supporting the "stationary ether" concept, since the "motion" he referred to was "apparent" motion of the ether perceived by an observer on earth, relative to a stationary ether.

The importance of cosmic motion to the electromagnetic effects of static charges was brought up by Nikola Tesla in his lecture: *"About fifteen years ago, Professor Rowland*[38] *demonstrated a most interesting and important fact, namely, that a static charge carried around produces the effects of an electric current." "...and conceiving the electrostatically charged molecules in motion, this experimental fact gives us a fair idea of magnetism. We can conceive lines or tubes of force which physically exist, being formed of rows of directed moving molecules; we can see that these lines must be closed, that they must tend to shorten and expand, etc. It likewise explains in a reasonable way, the most puzzling phenomenon of all, permanent magnetism, and, in general, has all the beauties of the Ampere theory without possessing the vital defect of the same, namely, the assumption of molecular currents. Without enlarging further upon the subject, I would say, that I look upon all electrostatic, current and magnetic phenomena as being due to electrostatic molecular forces.*"

In these statements, Nikola Tesla showed he was aware that any "stationary" locale on earth is actually in fantastic motion ("70,000 mph"). The electrostatic charges "carried around" are currents between atoms and the ether, which produce magnetism. The phenomena of 'permanent magnetism' or 'cosmically induced' magnetism are apparently due to electrostatic charges 'carried around' by cosmic motion, in the universal ether field.

---

[38] [Henry Augustus Rowland] (November 27, 1848 April 16, 1901) was a U.S. physicist. Between 1899 and 1901 he served as the first president of the American Physical Society. He is remembered today particularly for the high quality of the diffraction gratings he made and for the work he did with them on the solar spectrum.

Since no one can hold an atom or molecule perfectly still - because it is in fantastic motion - all atoms and molecules carry currents producing magnetic fields. Since a magnetic field is the product of a current, no one can produce a magnetic field without electricity, moving through or along a conductor, or as electrostatic charges in local or cosmic motion."[39]

"Now, I must tell you of a strange experience which bore fruit in my later life. We had a cold snap observed even drier than before. People walking in the snow left a luminous trail. As I stroked Macak's back, it became a sheet of light and my hand produced a shower of sparks. My father remarked, this is nothing but electricity, the same thing you see on the trees in a storm. My mother seemed alarmed. Stop playing with the cat, she said, he might start a fire. I was thinking abstractly. Is nature a cat? If so, who strokes its back? It can only be God, I concluded. I can not exaggerate the effect of this marvelous sight on my childish imagination. Day after day I asked myself what is electricity and found no answer. Eighty years have gone by since and I still ask the same question, unable to answer it." Tesla [40]

13.4.1 Summary

Keely's definition of electricity in this writer's opinion seems to address the issue best of the three because he details causative forces - when allied with some of Russell's ideas. More research into Tesla's ideas is required. We know ether is quantum and supports electric and magnetic phenomena. To fully understand Keely's definition requires a re-education to more fully grasp his concepts discarding (or at least setting aside) descriptions of phenomena (effects) to the contrary.

---

[39] Summation of Tesla's Dynamic Theory of Gravity; An excerpt from: Occult Ether Physics by William R. Lyne

[40] Nikola Tesla - Lord of Lightning; by Nikola Tesla , edited by William Kern

Chapter 14

Gravity

14.1 Keely on Gravity

"Gravity is nothing more than an attractive, sympathetic stream[1]*, flowing towards the neutral center of the earth, emanating from molecular centers of neutrality; concordant with the earth's centers of neutrality, and seeking its medium of affinity with a power corresponding to the character of the molecular mass. Gravity, he defines as transmittive interetheric force under immense etheric vibration. He continues: - The action of the mind itself is a vibratory etheric evolution, controlling the physical, its negative power being depreciatory in its effects, and its positive influence elevating."*[2]

"Gravity is transmittive interetheric force under immense etheric vibration."[3]

"Gravity is the mutual attraction of atomoles."[4]

"Gravity is an eternal existing condition in etheric space, from which all visible forms are condensed. It is inherent in all forms of matter, visible and invisible. It is not subject to time or space. It is an established connective link[5] *between all forms of matter from their aggregation. Time is annihilated by it, as it has already traversed space when the neutral centers*

---

[1] [sympathetic stream]

[2] [Keely and His Discoveries], Chapter 5

[3] Keely

[4] Keely

[5] [connecting link]

of the molecules were established. It is nothing more than an attractive, sympathetic stream, flowing towards the neutral center of the earth, emanating from molecular centers of neutrality, concordant with the earth's centers of neutrality and seeking its medium of affinity with a power corresponding to the character of the molecular mass."[6]

"Nature has established her sympathetic concordants from the birth of the neutral centres of the planets. This is gravity; therefore gravity is fixed, inherent. There is no flight of gravity."[7]

"Gravity is polar propulsion the sympathetic concordant of the ninths and but one of the triune combinations. Magnetism is polar attraction, while gravity is polar propulsion. By proper vibrations, the action of both magnetism and gravity can be intensified or accelerated."[8]

"The sympathetic outreach[9] *of negative attraction*[10] *is the power that holds the planetary masses in their orbital ranges of oscillatory action.*"[11]

"Gravity is not subject to time or space. It pervades the universe without reference to time or space, instantaneously and without intermission. It is, however, a sympathetic flow, proceeding from the molecular or mass neutral centers to the earth's neutral center with a power corresponding to the character of each individual molecular mass." He [Keely] believed gravity dependent on the medium of the polar stream, for he says *"If the sympathetic negative polar stream were cut off from the earth the molecular neutral centers would float away into space like a swarm of bees.*"

"Under my system the gravital flow comes under the order of the "sympathetic concordant of the ninths" and is that third of the triune combination called the "polar propulsive." Gravity is polar propulsion while magnetism is polar attraction. Both magnetism and gravity can be accelerated by proper

---

[6] John Worrell Keely circa 1886, [Snell Manuscript]

[7] [Snell Manuscript]

[8] MAGNETIC ENGINE - Snell]

[9] [sympathetic outreach]

[10] [7B.18 - Sympathetic Negative Attraction], [negative attraction], [Neutral Negative Attraction], [polar negative attraction], [Sympathetic Negative Attraction]

[11] Keely, [Snell Manuscript]; [Sympathetic Negative Attraction]

vibrations.

"He believed gravity to be the result of a law which provides for a reversion of attraction in repulsion, and which acts by transmission of force under immense etheric vibration through the "interetheric subdivision." He also showed that gravitation acts "as a lever." MacVicar states "Every individualized object assimilates itself to itself in successive moments of its existence and all objects tend to assimilate one another." Keely says "Gravity is an ever existing eternal force, coexistent with the compound etheric or high luminous (seventh subdivision) since it entered into, and is an inherent property of, all forms of aggregated matter from their birth. It is the source from which all matter originated and each substance-unit or neutral center is a concordant link, attractive and dispersive, to all other neutral centers. Each neutral center is the nucleus of what we recognize as substance and is potentially the nucleus of a planet."[12]

14.2 Keely and anti-gravity

Keely not only had a working theory of what gravity is and how it functions he also developed experimental models which he demonstrated to scientists of his time and an actual heavier-than-air flying craft.[13]

From Dashed Against the Rock - *"A small instrument, having three gyroscopes as a principal part of its construction, is used to demonstrate the facts of aerial navigation*[14]*. These gyroscopes are attached to a heavy, inert mass of metal, weighing about one ton. The other part of the apparatus consists of tubes, enclosed in as small a space as possible, being clustered in a circle. These tubes, represent certain chords, which were coincident to the streams of force acting upon the planet, focalizing and defocalizing upon its neutral center. The action upon the molecular structure of the mass lifted was based upon the fact that each molecule in the mass possessed a north and south pole, - more strictly speaking, a positive and negative pole, - situated through the center, formed by the three atoms which compose it. No mat-*

---

[12] [GRAVITY - Snell]

[13] [air-ship], [Air-ship model], [Figure 19.17 - Keelys Air-ship], [17.15 - Theory of Vibratory Lift for Air-ships]

[14] [aerial navigation]

ter which way the mass of metal is turned, the poles of the molecule point undeviatingly to the polar center of the earth, acting almost exactly as the dip-needle when uninfluenced by extraneous conditions, electrical and otherwise. The rotation of the discs of the gyroscopes produces an action upon the molecules of the mass to be lifted, reversing their poles, causing repulsion from the earth in the same way as like poles of a magnet repel each other. This repulsion can be diminished and increased according as the mechanical conditions are operated. By operating the three discs, starting them at full speed, then touching two of them, so as to bring them, according to the tone they represented by their rotation, to a certain vibratory ratio, the weight then slowly sways from side to side leaves the floor, rising several feet in the air, remaining in that position, and as the discs gradually decrease their speed of rotation the weight sinks to the floor, settling down as lightly as thistle-down. Where one molecule can be lifted, there need be no limit as to the number in a structure that may be operated upon as easily as one. The vessel in contemplation, the aerial navigator, will be over two hundred feet long, over sixty feet in diameter, tapering at both ends to a point, made of polished steel, and will be capable of being driven under the power of depolar repulsion, at the rate of three hundred miles an hour. It can be far more easily controlled than any instrument now in use for any phase of transit. Another very remarkable feature connected with this system of aerial navigation, is that the vessel is not buoyed up or floated in the air through the medium of the air, so that if there were no atmosphere it would float just as readily; hence, under mechanical conditions most certainly capable of production, involving massive strength of resistance to interstellar vacuity this can be made capable of navigating even the remote depth of space, positions between planets where polarity changes being controlled by other adjuncts of concentration for that purpose.

"Safely enclosed within this structure, a man possessing the chemical knowledge these new laws[15] *give, with sufficient supply of material from which to make oxygen, by the enormously increased rate of speed attained by such navigator where atmospheric friction is avoided, the time occupied in traveling from one planet to another would be amazingly brief, and one can travel to other planets in this system of worlds as easily as the same ship could navigate the depths of the ocean.*

---

[15] [Keelys Forty Laws]

"The great obstacle hitherto preventing the solution of this problem has been the strength of structure needed under conditions above presented. With this knowledge of matter, the size of structure is unimportant; the heaviest can be as easily controlled as the lightest."[16]

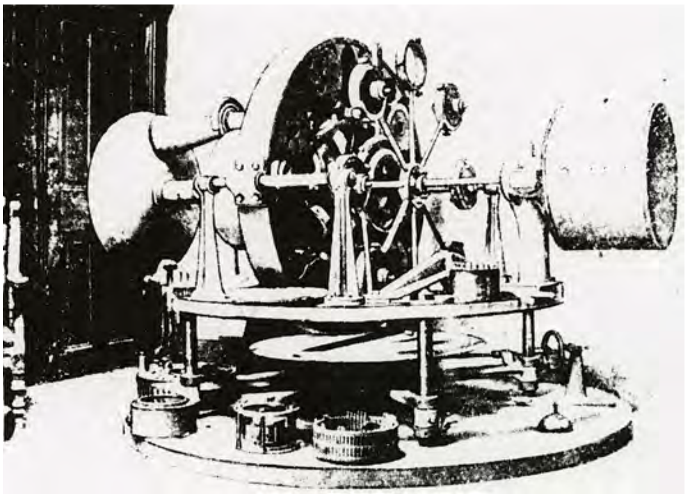

Figure 14.1: Keely's experimental air ship
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/air-ship-photo.jpeg

14.3 Levitation experiments

No less intriguing are Keely's experiments showing principles of levitation.

---

[16] [Levitating Gyroscopes], [Dashed Against the Rock]

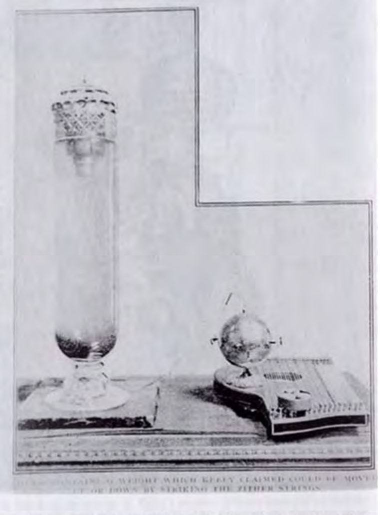

Glass containing weight which Keely claimed could be moved up or down by striking the zither strings.

Figure 14.2: Levitating weights
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/anti-gravity-keely.jpg

From Pittsburg Dispatch April 07, 1890 - Mr. Willcox[17] and Dr. Leidy[18] then examined the platinum wire, the metal disk on which the magnetic needles rested, the cupboard, the base of the cupboard and the table on which the disk and needle rested. When asked if there was, in his opinion, any possibility that the force which made the needle revolve[19] was electricity, compressed air or steam, Dr. Leidy looked thoughtful and said he could not see the possibility of any of those forces producing the result attained.

"What you have seen was shown you in order to illustrate the ease with which this force can be made to do work," said Keely.

"Of course the work just done was trifling. But I hope now to show you what will look very differently."

---

[17] Dr. James M. Wilcox, PhD, [Wilcox]

[18] [Joseph Leidy] (September 9, 1823 April 30, 1891) was an American paleontologist. Leidy was professor of anatomy at the University of Pennsylvania, and later was a professor of natural history at Swarthmore College.

[19] [4.2.5 - Rotation of Compass Needle], [DISTURBANCE OF MAGNETIC NEEDLE]

He pointed out then two glass jars, such as chemists use, on a table near by. The jars were of the same size, about 40 inches high, and 10 inches in diameter. They were filled with what is said to be, and certainly smelled and tasted like, Schuylkill[20] water. In the bottom of one jar lay a copper globe, cut in half to show, Keely said, just what it was; and filled, in each hollow half, with iron nails.

Another Peculiar Experiment.

In the other jar were three brass balls of different sizes. The copper globe and nails were weighed by Dr. Leidy and found to kick the beam at five pounds and six ounces. The brass balls or eggs they were egg-shaped, weighed less. Everybody sat down alter the weighing and Keely fastened another "platinum silver wire" to the cylinder on the cupboard and, detaching the one already in use from the magnetic needle disk was about to fasten the loose ends of each to the metal disks that covered the tops of the jars when some one asked if the wires were "hollow." The suggestion was followed by a smile from Keely, who at once cut off the end of one of them and handed it around.

"Prof. Rowland, of Baltimore, [had] declared that this was a fraud, because the wires were hollow,' said the woman, *"but Keely asked' him how he could explain what Keely did even on the hypothesis that the wires were hollow and he didn't answer. Then Keely got mad and would not let him cut the wire, as he wanted to do."*

Dr. Leidy followed everything closely. Keely fastened the loose ends of both wires into the metal caps of the cylinders.

The String an Important Feature.

Again he spun the spindle on his cupboard with the twine string he had used before. Again with his gnarled fingers, the joints of the first two fingers of his right hand being as big as walnuts, be pounded the "harp" in the cupboard and the resonant bars on top of it.

"What are you doing now?" asked Dr. Leidy.

---

[20] river that passes through Philadelphia

"I am trying" said Keely, *"to get the mass chord of that copper sphere full of nails. Every aggregation of molecules or of matter, I claim, or, in other words, every mass of matter, has a sympathetic chord, through the medium of which I can operate my vibratory force."*

The chord was not found for some minutes. Again the spindle was spun by the help of the twine, and its whizz [sic]was distinct in the silence of the room. The search for the mass chord continued on the "harp" and the resonant rods. A deep, clear note resounded from both at the same time, and at the instant it broke on the ear the heavy copper globe quivered as it lay at the bottom of the water, rolled over, reluctantly as it were abandoning the ties by which gravity held it to the bottom of the jar, floated at first slowly, and then more swiftly and steadily to the top of the jar, against which it impinged with an audible concussion.

Gravity Easily Overcome.

"Why, there is the force of gravity as plainly overcome and indeed annihilated as it is possible for a human being to imagine," exclaimed the woman.

Dr. Leidy was asked this question: *"Doctor, is it true that this unknown force, or what is here manifested as such, has actually before your eyes overcome the force of gravity with which we are all familiar?"* and the answer, slowly, deliberately, was: *"I see no escape from that conclusion."*[21]

14.4 Russell on Gravity

"Gravitation is the generative force of increasing potential and the regenerative force of decreasing potential.

"Gravitation is the power within the electric force of action to attract the electric force of action.

"Gravitation is the contractive power within electricity to divert the univer-

---

[21] [Pittsburg Dispatch April 07-1890], [Leidy and Wilcox Visit Keelys Lab], [Figure 17.18 - Keelys Levitation Demonstration], [Eye Witness Accounts]

sal constant of energy into centripetal vortices of closing spirals of increasing speed, thereby attracting similar states of motion into an accumulation of mass, the pressure of which increases toward its center.

"Gravitation is an expression of the power of electricity to accumulate by induction and, by so doing, to force magnetism to increase its resistance to that accumulation.

"Gravitation is the (electric) inductive force."[22]

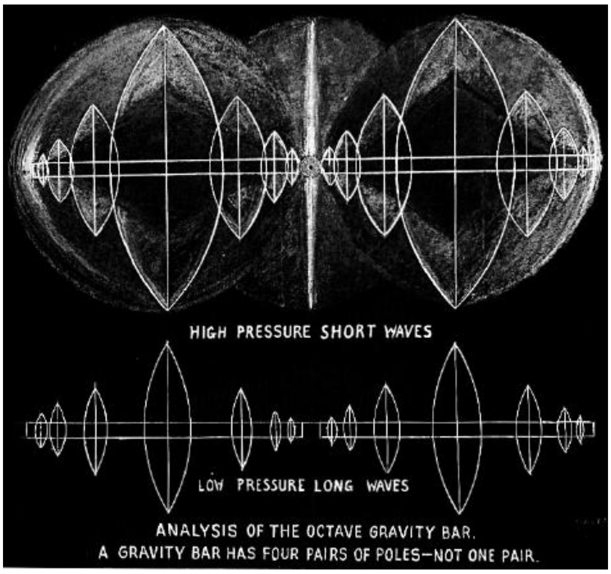
HIGH PRESSURE SHORT WAVES
LOW PRESSURE LONG WAVES
ANALYSIS OF THE OCTAVE GRAVITY BAR.
A GRAVITY BAR HAS FOUR PAIRS OF POLES—NOT ONE PAIR.

Figure 14.3: Gravity wave bar and poles
Image: http://pondscienceinstitute.on-rev.com/imagesNatDyn/gravitybar.jpg
(courtesy University of Science and Philosophy)

---

[22] Russell, [The Universal One]

Chapter 15

Bibliography and References

This bibliography pertains to this book alone. It does not include the hundreds of bibliographical references and links to authoritative materials accumulated throughout the svpwiki.com web site supporting the SVP paradigm.

Arguelles, Jose; Earth Ascending, page 21; 1996 (3rd edition)
Bearden, Thomas E.; Lieutenant Colonel, U.S. Army (Retired). Born Dec. 17, 1930, Cheniere, Louisiana. See [Bearden]
” ” Analysis of Scalar/Electromagnetic Technology
” ” Energy from the Vacuum - Concepts and Principles; 2002
” ” Gravitobiology - A New Biophysics
” ” The New Tesla Electromagnetics and The Secrets of Electrical Free Energy
Bible, Mathew
Bjerknes, Carl Anton
Blakiston's New Gould Medical Dictionary. The Blakiston Company, 1949. 1st edition
Bloomfield-Moore, Clara Jessup; Keely and His Discoveries, Delta Spectrum Research
” ” The Operation of the Vibratory Circuit
” ” Vibratory Sympathetic and Polar Flows
” ” What Electricity Is, Delta Spectrum Research
” ” and many other of her writings. See [Bibliography].
Brinton, Professor Daniel Garrison; (May 13, 1837 July 31, 1899) was an American archaeologist and ethnologist at the University of Pennsylvania
Carpenter, Tom; Dialogue on Awakening, The Carpenter's Press

Cayce, Edgar; the Edgar Cayce readings, The Edgar Cayce Foundation, Virginia Beach, Virginia

Colville, William; Dashed Against the Rock, 1894, Delta Spectrum Research

Efimov, Vitaly; is a Russian theoretical physicist. He proposed the existence of a novel and exotic state of matter now dubbed the ‘Efimov State’ as a researcher in A.F. Ioffe Physico-Technical Institute, USSR Academy of Sciences, Leningrad, USSR in his 1970 paper “Energy levels arising from resonant two-body forces in a three-body system”. It was announced in 2006 that the existence of this state of matter had been confirmed.

Feynman, Richard; QED: The Strange Theory of Light and Matter; Princeton Science Library

Fuller, Buckminster; SYNERGETICS - Explorations in the Geometry of Thinking, Volumes I II, New York, Macmillan Publishing Co, 1975, 1979.

Harris, T. F., B.Sc., F.C.S.; Handbook of Acoustics, 1903, Delta Spectrum Research

Hughes, Fitzgerald; Harmonies of Tones and Colours - Developed by Evolution, Delta Spectrum Research

Keely, John Ernst Worrell, 1827-1898, discovered and developed the science of Sympathetic Vibratory Physics. Various articles. See [Bibliography].

LaRouche, Lyndon Hermyle, Jr. (born September 8, 1922) is an American self-styled economist, political activist, and the founder of several political organizations known collectively as the LaRouche movement. Various articles.

Leidy, Joseph (September 9, 1823 April 30, 1891) was an American paleontologist. Leidy was professor of anatomy at the University of Pennsylvania, and later was a professor of natural history at Swarthmore College. **Lyne**, William R.; Summation of Tesla’s Dynamic Theory of Gravity; An excerpt from: Occult Ether Physics

Lockyer, Sir Joseph Norman; FRS (17 May 1836 16 August 1920), known simply as Norman Lockyer, was an English scientist and astronomer

Maxwell, James Clerk ; (13 June 1831 5 November 1879) was a Scottish theoretical physicist and mathematician. See [James Clerk Maxwell]

Pond, Dale; Speaking Freely, Delta Spectrum Research
” ” svpwiki.com
” ” Universal Laws Revealed, Delta Spectrum Research
” ” video Basic Principles, Delta Spectrum Research

Puthoff, H. E.; “On the Source of Vacuum Electromagnetic Zero-Point Energy," Phys. Rev. A, vol. 40, p. 4857,1989; Phys. Rev. A vol. 44, p. 3382,1991.

Ramsay, Dougald; The Scientific Basis and Build of Music, Delta Spectrum Research

Rossing, Thomas D. The Science of Sound. Addison-Wesley Publishing Company, 1982

Russell, Walter; Atomic Suicide?, University of Science and Philosophy

" " Fifth Kingdom, The Beginning of Man, University of Science and Philosophy

" " Genero-Radiative Concept, University of Science and Philosophy

" " Home Study Course, University of Science and Philosophy

" " notation on chart, 1955, University of Science and Philosophy

" " Russells Optic Dynamo-Generator, University of Science and Philosophy

" " The Message of the Divine Iliad, University of Science and Philosophy

" " The Secret of Light, University of Science and Philosophy

" " The Universal One, University of Science and Philosophy

Sinclair, David; A New Creed, Delta Spectrum Research

" " Vera Vita the Philosophy of Sympathy, Delta Spectrum Research

Snell, C. W.; The Snell Manuscript, Delta Spectrum Research

Stainer, John; Barrett, W. A.; A Dictionary of Musical Terms; Novello, Ewer and Co., London, 1903

Tesla, Nikola; Nikola Tesla Lord Of Lightning: The Experiments, Inventions, Lectures, Writings And Patents of Nikola Tesla

Thomson, William; 1st Baron Kelvin OM, GCVO, PC, PRS, PRSE, (26 June 1824 - 17 December 1907) was a Belfast-born British mathematical physicist and engineer. **Tyndall**, John; Sound; Longmans, Green, and Co., London, 1893

Webster's 1828 Dictionary

White, Carol; Energy Potential: Toward a New Electromagnetic Field Theory; (with essays by Bernhard Riemann trans. from German by J. J. Cleary, Jr.), Campaigner Publications, New York, 1977.

Wikipedia/Arithmetical Progression

" " /Bose-Einstein condensation

" " /Dirac Sea

" " /Efimov state

" " /Geometric Progression

" " /Linear Progression

" " /Quark-gluon plasma
" " /Richard Phillips Feynman
" " /States of matter
" " /Stockholm Syndrome

Appendix A

I - Laws of Being

from 'Keely and His Discoveries', Conclusion

Keely's Physical Philosophy

"Mr. Keely begins with sounds whose vibrations can be known and registered. I presume that the laws of ratio, position, duality, and continuity, all the laws which go to mould the plastic air by plastic bodies into the sweetness of music, will also be found ruling and determining all in the high silence of interior vibrations, which hold together or shake asunder the combinations that we call atoms and ultimate elements."[1]

"What Keely has discovered in physics, I am in some measure credited with discovering in metaphysics: this is nothing strange, according to this philosophy, which shows that many people may divine the same original truth at the same time by means of the etheric element which connects the Deity, the source of all truth, with all His creatures."[2]

Abstract of Keely's Physical Philosophy in its main features up to the point of practical application;

by Professor Daniel G. Brinton, of the Pennsylvania University; subject to modifications and additions when Keely has made public his system.

---

[1] [The Scientific Basis and Build of Music]; [Dougald Carmichael Ramsay]. Edited by the Rev. [John Andrew]. Marcus Ward and Co.

[2] Preface to [Vera Vita; or, The Philosophy of Sympathy]. [David Sinclair]. Author of [A New Creed]. Digby, Long and Co., London.

The fundamental conception of the Universe is force manifesting itself in rhythmical relations. This definition is exhaustive, including both thought and extension, matter and mind. The law for the one is the law for the other. The distinction between them is simply relative, i.e. quantitative, not qualitative.
The rhythmical relations in which force acts are everywhere, under all conditions, and at all times, the same. They are found experimentally to be universally expressible by the mathematical relations of thirds.

These threefold relations may be expressed with regard to their results as,-

I. Assimilative
II. Individualizing.
III. Dominant or Resultant.

From these three actions are derived the three fundamental LAWS OF BEING.

I. Law of Assimilation: every individualized object assimilates itself to all other objects.
II. Law of Individualization: every such object tends to assimilate all other objects to itself.
III. Law of the Dominant: every such object is such by virtue of the higher or dominant force which controls these two tendencies.

Applying these fundamental laws to an explanation of the universe, as it is brought to human cognition, all manifestations of force may be treated as modes of vibrations.

The essential differences give rise to three modes of vibration:-

I. The Radiating: called also the “Dispersing,“ the “Propulsive,“ the “Positive,“ and the “Enharmonic.” see entropy
II. The Focalizing: called also the “Negative,“ the “Negative Attractive,“ the “Polarizing,“ and the “Harmonic.” see syntropy
III. The Dominant: called also the “Etheric,“ or the “Celestial.”

These, it will be noted, correspond to the three laws of being. It is not to be understood that any one of these three modes of vibration can exist independently. Each by itself is called a "current," and all three must be present in every "stream" or "flow" of force. The relations of the currents in every flow are expressible in thirds, and it is experimentally demonstrable that the relations of the three are in the order named: as $33\ \frac{1}{3} : 66\ \frac{2}{3} : 100$.

The evolution of what is called "matter" from the different modes of vibration is through the action of the second law, that of focalization, or "negative attraction," or "negative affinity."[3]

Where the vibrations under this mode meet, and are maintained in a state of mutual affinity or equilibrium, there is established what is called a "neutral center," or, as otherwise expressed, "a center of sympathetic coincidence."

The terms "neutral attraction," "neutral affinity," "negative attraction," or "polar negative attraction," are emplyed to express the property of a mode of vibration to direct its components towards such centre. see syntropy

As no current of flow of force can be composed of one mode of vibration only, but must always be composed of three modes uniting in varying thirds, we have 1 x 2 x 3 = 6 as the total possible forms of sympathetic coincidence, or, to speak in ordinary terms, there can be six; and six only, possible forms of individualized being. These are what Keely calls the six orders of atomic subdivision, or orders of vibratory motion, and he names them as follows:

Molecular
Intermolecular
Atomic
Interatomic
Etheric
Interetheric

In this list the forms of matter are arranged in the mathematical sequence of the rapidity of the oscillations of their constituent members; the proportion being proved by experiment to be as follows: for the molecular orders:

---

[3] see [syntropy]

1 : 3 : 9 : 27 : 81 : 243.

This arithmetical progression changes in the atomic orders to a geometrical progression as follows:

3 : 9 : 81 : 6561 : 43,046,721; etc.

This same method of progression is believed to hold in all the orders of vibrations above the molecular, and soon passes into mathematical infinity.

Actually, however, all matter of which we are capable of cognition through the medium of our senses is in one of three forms of aggregation:

Molecular
Atomic
Etheric

in each of which the controlling mode of vibration is respectively,

The Enharmonic.
The Harmonic.
The Dominant.

But it must be understood that each of these modes is a positive and real constituent of every atom and molecule.

It will be seen that as every form of material aggregation is to be considered as a "neutral center of attraction," where the vibratory force of all three orders are held in "sympathetic coincidence," that is, in balanced activity or harmonized motion, and not by any means cancelled or mutually destroyed, there is no diminution of force, but only temporary suspension of its radiating or propulsive activity or expression. [4]

This is the foundation of Keely's doctrine of "latent force," and of the indefinite power which can be obtained by breaking up the harmonious balance or equation of forces of every mode, which exists in every "neutral center," that is to say in every mass of matter.

---

[4] see [Rhythmic Balanced Interchange]

Insomuch as every mass of matter consists thus, in fact, of vibrations in harmonic equilibrium, related by simple proportions of thirds, it follows that every mass of every description stands in harmonic relations to every other mass. This is, in part, what is meant by the sympathy of all forms of matter and of motion; and it is through the study of the methods of increasing or diminishing this sympathy that we reach practical results in this field of research. At present this is best accomplished by resonance; that is, through the harmonic vibrations created by musical instruments, bringing out the acoustic world as the microscope reveals the hidden visual world.

Every visible or tangible mass of matter must be regarded as an aggregation of molecules; the molecules being the true centres of the equated forces of "neutralized attraction."

These molecules have been experimentally proved by Keely to be formed of all three modes of vibration; the proof being that they respond to all three modes when subjected to the tests of compound concordant impulses.

When in that state of neutral aggregation which we know as matter, each molecule is in perpetual oscillation, the range of the oscillation being one-third of the molecule, and its rapidity 20,000 oscillations in a second.

It is through the disturbance of this oscillatory equilibrium, by means of resonant impulses, that Keely alters the relations of the vibratory impulses which constitute matter. This he does by striking the same chord in three octaves, representing the third, sixth, and ninth of the scale.

Of these, the sixth reduces the range of molecular vibrations or oscillations; and, by thus bringing nearer to each other the neutral centers, increases solidification.

The ninth extends the range of molecular oscillation, and thus tends to give greater tenuity to the mass. It induces "trajectile velocity" from neutral centers, or "neutral radiation." Experiment shows that molecular dissociation does not take place until the molecule attains an oscillation approaching, it not fully reaching two-thirds of its diameter. This can be effected by means of the action of the "enharmonic" or "radiating" current applied to the mass, after its molecules have once been disturbed by an "introductory impulse;" that is, by the musical note above mentioned.

The third represents the "dominant," and when brought under control of a harmonic resonant impulse induces a complete rearrangement of the modes of vibration and oscillation; in other words, will transform the mass either into its component initial forces, or into some other form of matter.[5]

It is the study of the dominant to which Keely has devoted his recent researches. He aims to control the power he evolves by altering the dominant or etheric mode of vibration in the triplicate flows of force.

As all molecules and masses are mere centres of harmonized vibrations, temporarily held in suspension by simple laws identical with those of resonance, it follows that these centres can be broken up or divided by certain orders of vibration impinging upon and disturbing them.

It is a familiar fact that a cord in vibration tends to produce a similar vibration in a cord placed near it. This property belongs to all vibrations, whether resonant or not, and they exert it in proportion to the "order" to which they belong. The distance in space to which this power extends, or can be extended, is what is called "the sympathetic outreach" of the current of flow.

In this manner we have "sympathetic negative attraction," and "sympathetic positive propulsion," with reference to the "outreach" of the third or dominant current of the stream, which is allied to the order of etheric vibrations.

Each molecule of a given mass of matter represents the same harmonic chord or note in its oscillatory motion. The "chord of the mass" is, therefore, the chord of every molecule of the mass.

But as the condition of absolutely stable equilibrium is theoretical only, and does not exist in nature, the chord of the mass is constantly changing. Yet we must learn to control this "chord of the mass" by resonant induction, if we would gain command of the molecular forces.

---

[5] see [Law of Transformation of Forces]

Keely believes he has solved this problem, by the invention of a mechanical device which brings the chords of all masses within the conditions of a few simple acoustic tests.

The range of molecular oscillation is affected differently in different substances when submitted to the same vibratory impulse, and these ranges can be measured.

In the three metals, silver, gold, and platina, we obtain the proportions - 3 : 6 : 9

As this is the primary relation of the modes of vibration, a wire made of these three metals is peculiarly adapted to transmit concordant impulses: and nodes made of these substances placed upon a wire, transmitting resonant vibrations, indicate, by the different orders of vibration induced in them, the rate of oscillations of the atomic constituents.

The phenomenon of rotation arises from the harmonic inter-action of the dominant and enharmonic elements of the flow: in other words, the first and third, the third and ninth, etc.; those whose vibrations bear the proportions to each other $33\frac{1}{3} : 100$.

A practical example of rotation is a wheel in revolution on its axis. This is force in its commercial or economic aspect. To accomplish this result by molecular vibratory action, we must gain control of the "negative attractive" or "enharmonic" current of the triple flow, and the problem is then solved up to any limit of power.

Appendix B

II - Laws of Being

"The brain of an Aristotle was needed to discern and grasp Keely's meaning to interpret and define it. Dr. Brinton never touches a subject without throwing light upon it, and his penetrating mind perceived the ideas to be defined in all their relations. His keen logical acumen separated and classified them in their order, in a true, sound, and scientific manner. In the words of Sir James Crichton Browne, who heard Professor Brinton read this abstract in London, "Professor Brinton's synopsis is an able, lucid and logical paper." [Progressive Science]

Laws of Being

by Professor Daniel Brinton, University of Pennsylvania

Abstract of Keely's Physical Philosophy in its main features up to the point of practical application; by Professor Daniel G. Brinton, of the Pennsylvania University; subject to modifications and additions when Keely has made public his system. [1]

"The fundamental conception of the Universe is force manifesting itself in rhythmical relations.

"This definition is exhaustive, including both thought and extension, mat-

---

[1] from Keely and His Discoveries, Conclusion, Keelys Physical Philosophy

ter and mind. The law for the one is the law for the other. The distinction between them is simply relative, i.e., quantitative, not qualitative.

"The rhythmic relations in which force acts are everywhere, under all conditions, and at all times, the same. They are found experimentally to be universally expressible by the mathematical relations of thirds.

"These threefold relations may be expressed with regard to their results as,

I. Assimilative
II. Individualizing
III. Dominant or Resultant

"From these three actions are derived the three fundamental LAWS OF BEING:

I. Law of Assimilation: every individualized object assimilates itself to all other objects.

II. Law of Individualization: every such object tends to assimilate all other objects to itself.

III. Law of the Dominant: every such object is such by virtue of the higher or dominant force which controls these (above) two tendencies.

"Applying these fundamental laws to an explanation of the universe, as it is brought to human cognition, all manifestations of force may be treated as modes of vibrations.

"The essential differences give rise to three modes of vibration:

I. The Radiating: called also the "Dispersing", the "Propulsive", the Positive, and the Enharmonic.

II. The Focalizing: called also the Negative, the Negative Attractive, the "Polarizing", and the Harmonic.

III. The Dominant: called also the Etheric, or the Celestial.

"*These, it will be noted, correspond to the three laws of being. It is not to be understood that any one of these three modes of vibration can exist independently. Each by itself is called a "current", and all three must be present in every "stream" or "flow" of force. The relations of the currents in every flow are expressible in thirds, and it is experimentally demonstrable that the relation of the three are in the order named: as* $33\,\frac{1}{3} : 66\,\frac{2}{3} : 100$*.*

"*The evolution of what is called matter from the different modes of vibration is through the action of the second law, that of focalization, or negative attraction, or "negative affinity".*

"*Where the vibrations under this Mode meet, and are maintained in a state of mutual affinity or equilibrium, there is established what is called a neutral center, or, as otherwise expressed, "a center of Sympathetic Coincidence".*

"*The terms "neutral attraction", "neutral affinity", "negative attraction", or "polar negative attraction", are employed to express the property of a Mode of vibration to direct its components towards such center.*

"*As no current or flow of force can be composed of one Mode of vibration only, but must always be composed of three modes uniting in varying thirds, we have* $1 \times 2 \times 3 = 6$ *as the total possible forms of Sympathetic Coincidence, or, to speak in ordinary terms, there can be six; and six only, possible forms of individualized being. These are what Keely calls the six orders of Atomic subdivision, or orders of vibratory motion, and he names them as follows*:

I.	Molecular
II.	Intermolecular
III.	Atomic
IV.	Interatomic
V.	Etheric
VI.	Interetheric

"*In this list forms of matter are arranged in the mathematical sequence of* *the rapidity of the oscillations of their constituent members; the proportion being proved by experiment to be as follows for the molecular orders*:

1 : 3 : 9 : 27 : 81 : 243 ...

"*The arithmetical progression changes in the atomic orders to a geometrical progression as follows*:

3 : 9 : 81 : 6561 : 43046721, etc.

"This same method of progression is believed to hold in all the orders of vibrations above the molecular, and soon passes into mathematical infinity.

"*Actually, however, all matter of which we are capable of cognition through the medium of our senses is in one of three forms of aggregation*:

I. Molecular
II. Atomic
III. Etheric

"*In each of which the controlling Mode of vibration is respectively*:

I. The Enharmonic
II. The Harmonic
III. The Dominant

"*But it must be understood that each of these modes is a Positive and real constituent of every atom and Molecule.*

"*It will be seen that as every form of material aggregation is to be considered as a "neutral center of attraction", where the vibratory force of all three orders are held in Sympathetic Coincidence, that is, in balanced activity or harmonized motion, and not by any means cancelled or mutually destroyed, there is no diminution of force, but only temporary suspension of its radiating or propulsive activity or expression.*

"*This is the foundation of Keely's doctrine of latent force, and of the indefinite power which can be obtained by breaking up the harmonious balance* *or equation of forces of every Mode, which exists in every neutral center, that is to say in every mass of matter.*

"*Insomuch as every mass of matter consists thus, in fact, of vibrations in harmonic equilibrium, related by simple proportions of thirds, it follows that every mass of every description stands in harmonic relation to every other mass. This is, in part, what is meant by the sympathy of all forms of matter and of motion; and it is through the study of the methods of increasing or diminishing this sympathy that we reach practical results in this field of research. At present this is best accomplished by resonance; that is, through the harmonic vibrations created by musical instruments, bringing out the acoustic world as the microscope reveals the hidden visual world.*

"*Every visible or tangible mass of matter must be regarded as an aggregation of molecules; the molecules being the true centers of the equated forces of "neutralized attraction".*

"*These molecules have been experimentally proved by Keely to be formed of all three modes of vibration; the proof being that they respond to all three modes when subjected to the tests of compound concordant impulses.*

"*When in that state of neutral aggregation which we know as matter, each Molecule is in perpetual oscillation, the range of the oscillation being one-third of the Molecule, and its rapidity 20,000 oscillations in a second.*

"*It is through the disturbance of this oscillatory equilibrium, by means of resonant impulses, that Keely alters the relations of the vibratory impulses which constitute matter. This he does by striking the same chord in three octaves, representing the (musical) third, sixth, and ninth of the scale.*

"*Of these, the sixth reduces the range of molecular oscillation, and thus tends to give greater tenuity to the mass. It induces "trajectile velocity" from neutral centers, or "neutral radiation". Experiment shows that molecular dissociation does not take place until the Molecule attains an oscillation approaching, if not fully reaching two-thirds of its diameter. This can be effected by means of the action of the enharmonic or "radiating" current applied to the mass, after its molecules have been disturbed by an "introductory impulse"; that is, by the musical note above mentioned.*

"The third represents the dominant, and when brought under control of a harmonic resonant impulse induces a complete rearrangement of the modes of vibration and oscillation; in other words, will transform the mass either into its component initial forces, or into some other form of matter.

"It is the study of the dominant to which Keely has devoted his recent researches. He aims to control the power he evolves by altering the dominant or etheric mode of vibration in the triplicate flows of force.

"As all molecules and masses are mere centers of harmonized vibrations, temporarily held in suspension by simple Laws identical with those of resonance, it follows that these centers can be broken up or divided by certain orders of vibration impinging upon and disturbing them.

"It is a familiar fact that a cord in vibration tends to produce a similar vibration in a cord placed near it. This property belongs to all vibrations, whether resonant or not, and they exert it in proportion to the order to which they belong. The distance in Space to which this power extends, or can be extended, is what is called "the sympathetic outreach" of the current or flow.

"In this manner we have "sympathetic negative attraction", and "sympathetic positive propulsion", with reference to the outreach of the third or dominant current of the stream, which is allied to the order of Etheric vibrations.

"Each Molecule of a given mass of matter represents the same harmonic chord or note in its oscillatory motion. The chord of the mass is, therefore, the chord of every Molecule of mass.

"But as the condition of absolutely stable equilibrium is theoretical only, and does not exist in nature, the chord of the mass is constantly changing. Yet we must learn to control this chord of the mass by resonant induction, if we would gain command of the molecular forces.

"Keely believes he has solved this problem, by the invention of a mechanical device which brings the chords of all masses within the conditions of a few simple acoustic tests.

“The range of molecular oscillation is affected differently in different substances when submitted to the same vibratory impulse, and these ranges can be measured.

“In the three metals, silver, gold, and platina, we obtain the proportions - 3 : 6 : 9 : - As this is the primary relation of the modes of vibration, a wire made of these three metals is peculiarly adapted to transmit concordant impulses; and nodes made of these substances placed upon a wire, transmitting resonant vibrations, indicate, by the different orders of vibration induced in them, the rate of oscillations of the atomic constituents.

“The phenomenon of rotation arises from the harmonic interaction of the dominant and enharmonic elements of the flow; in other words, the first and third, the third and ninth, etc.; those whose vibrations bear the proportions to each other $33\,\frac{1}{3}$ *: 100.*

“A practical example of rotation is a wheel in revolution on its axis. This is force in its commercial or economic aspect. To accomplish this result by molecular vibratory action, we must gain control of the negative attractive or enharmonic current of the triple flow, and the problem is then solved up to any limit of power.”

Appendix C

III - Ultimate Constitution of Matter

The Ultimate Constitution of Matter and the Action of the Force Regulating its Phenomena

by John W. Keely

First: Matter is capable of infinite subdivision.

Second: In the aggregation of matter, force or energy is stored up or conserved.

Third: In the dissociation of matter, force is liberated.

Fourth: All matter is in a state of perpetual activity, whether the substance under consideration be inanimate or animated, visible or invisible.

Fifth: There is no dividing of matter and force into two distinct terms, as they both are ONE. FORCE is liberated matter. MATTER is force in bondage.

Sixth: All motion is synchronous; no sound or movement can be made but all that moves or sounds does so in harmony with something.

Seventh: All structures, whether crystalline or homogeneous, have for their unit structures minute bodies called molecules. It is the motion of these molecules with which we have particularly to deal; as in experimental research and demonstration, when we produce an action upon one molecule we do so upon all the molecules constituting the mass operated upon.

Eight: These molecules have an envelope, rotating with inconceivable rapidity, formed of a high tenuous ether, whose place in the order of subdivision ranks third, the three divisions being, - first, molecular; second, atomic; third, atomolic. (Atomolic is the same as Etheric)

Ninth: This atomolic substance has a density approximately 986,000 times that of steel, enabling it to permeate steel as light penetrates glass; this rotating envelope of atomolic substance is in a liquid condition. There are four conditions of matter; viz. solid, liquid, gaseous, and ultra-gaseous. These conditions result from greater or lesser range of oscillation of the composing units individually; this is equally true, whether the units are molecules, atoms, atomoles, planets, or suns. But one LAW governs all matter.

Tenth: This molecular envelope, rotating with such great velocity, holds in its embrace the next subdivision of matter, the atomic. There cannot ever be more or less than three atoms in any molecule. These are placed so as to form a triangle in the interior; they rest in a condition of substance, or matter, we will term intermolecular. In this intermolecular substance we find an enormous energy or force in bondage, held thus by the rotating envelope enclosing it. Were we to rotate a spun brass shell, say nine inches in diameter, at a very much less rate of speed than that at which the molecular envelope rotates, - say nine hundred revolutions per second, - its equator would first bulge out, then form into an oval disk. A solid block of wood subject to such revolution would swiftly fly to pieces. The rotating envelope of a molecule, unlike these, the greater the velocity of rotation, the greater is its compression toward the center of the molecule. The rotation of this envelope is of such a nature as to produce an internal pressure upon every portion from every point of the molecule as a sphere. Were we to consider a rotating envelope as ordinarily understood, it would be one in which the envelope rotated around an equator having poles of no rotation; i.e., the poles would not possess the compressing force of the equator; the result would therefore be a compressed equator, and the intermolecular substance would pass out without resistance at the poles.

Eleventh: If it be possible let us conceive of an envelope with an equator, but destitute of poles, a number of these rotating over the sphere, this atomolic envelope possessing an almost infinite attractive force toward the center of the molecule, pressing in the intermolecular substance, where it is held until this revolving envelope becomes negatized by a certain order of vibration, when the enclosing matter rushes out to its natural condition of concordant tenuity, as in the case of gunpowder, dynamite, and nitroglycerine. This force, we must see, has been held in the embrace of the rotating envelopes of the unit structures, or where does it come from? This force at the time of an explosion was liberated by shock or fire, both being orders of imparted motion or vibration. How much greater the result would be were we to associate a scientific instrument now completed, and shortly to be given to the world, with such an agent as nitroglycerine; one pound of nitroglycerine would have its destructive force augmented beyond all possible control. These instruments are carefully concealed by wise masters from all persons save the few who are already prepared to study their potency with the exclusive end in view of aiding the real scientific progress of humanity; and, furthermore, it may be truly stated that a ferocious sensualist, however powerful his intellect, would be utterly unable to either comprehend or operate one of these marvelous constructions.

Twelfth: Next in order of consideration is the second subdivision of matter - the atomic. The atom has the same rotating envelope as the molecule, governed by the same laws of rotation and compression. The rotating envelope holds in its embrace the interatomic substance and three atomoles resting in it, the atomoles within the atom being constructed after the same pattern as the atom and the molecule, obedient to the same laws; the atomolic being simply the third subdivision of matter. The threefold order is absolutely universal.

Thirteenth: The atomolic substance is what is termed the ether which fills all space and is the transmitting medium for all celestial and terrestrial forces. This is the liquid ether of occult science.

Fourteenth: The atomoles are made up of atomolini (singulara tomolinus); the subdivision of matter from this point is beyond man's power, as at this point it escapes all control of apparatus, passing through glass and hardened steel as a luminous flame without heat, which is hardly seen before it vanishes, - a perpetual flame coldly luminous.

Fifteenth: (Not in book, may have been a mis-numbering. editor)

Sixteenth: This again, from previous analysis, is made up according to the triple order, and may again be subject to subdivision, even into infinity." [Dashed Against the Rock]

Dale Pond

Author of many books.
Creator of svpwiki.com and
Maker of Dynaspheres.

Keely's Laws of Being

This book describes the unseen (scalar) components of vibration, oscillation and toroids as also their relationships and dynamical functions. This book is a follow-up (volume II) of The Physics of Love published in 1994. The forces described in this book are the 3, 6 and 9 unseen (scalar) forces Keely used to achieve some of the remarkable feats of science and engineering accredited to him yet not understood by the rest of us.

Published by Delta Spectrum Research

B01130

Musical Numerology

By Don Robertson, 1970 with revisions in 2003

Copyright 1970, 2003 by Don Robertson

(First appeared in the book *Kosmon)*

I wrote the article Musical Numerology in 1970 and published it in Denver, Colorado in a book called Kosmon, a book created by myself and two other people. It was in this book that I first published my findings about the duochord and proclaimed that rock and roll music was taking a downward turn.One day, after we had started distributing the book to local bookstores, I received a letter from a young man who lived nearby and wanted to meet the author of this article. I went to see him and looking around the one-bedroom apartment, I saw that everywhere there were large pieces of white paper on which he had been making mathematical calculations: on tables, chairs, and taped to all of the walls. He was ecstatic, "I had to meet the author because I have applied what was stated in the article to every known science and found that the same formula proves itself in every case." I went back a week later to discuss the matter with him more thoroughly , but the apartment was empty and he was gone. Now after 23 years, I have revised the article, and am publishing it again on DoveSong.Com. I hope this information can be as useful to others as it was for this young man and has been for me. I believe that musical scales and harmony are based on alchemical realities that can be proved by math. Math demonstrates the formation of the laws of the universe, and music, color, and all manifestation follow the same laws.

Don Robertson, May, 2003

* *

To our knowledge, there is no religion on the earth that does not incorporate some kind of symbolic use of **the quantities three and seven**.

Our objective is to illustrate some of the mysterious relationships between the quantity three, the quantity **seven**, and what these quantities have in common with the basic physical structure of the universe.

The direct relationship between **three** and **seven** is as follows:

If we have three objects, which we will call 1, 2, and 3, we will observe that the absolute number of different combinations of these 3 objects is 7;

that is, we may arrange them in no more than seven different combinations, regardless of the order in which they are arranged.

Combination 1	**1, 2, 3**
Combination 2	**1, 2**
Combination 3	**1, 3**
Combination 4	**2, 3**
Combination 5	**1**
Combination 6	**2**
Combination 7	**3**

There are **three primary colors** in the color spectrum, plus four more natural colors that are created from the combinations of the three colors similarly to the three objects above.

All colors that exist are either :

a) One of the primary colors, or
b) A resultant mixture of two or three of the primary colors.

The primary colors cannot be created by mixing other colors, so they are rather special.

The three primary colors are **red**, **yellow**, and **blue**.[1]

The ***natural***, or ***secondary*** colors are those colors that are created by mixing equally the three primary colors according to the chart above.

The resultant seven natural colors the colors of the rainbow, a naturally occurring phenomena that can be created using a prism to break down light.

They are arranged in the spectrum as follows:

The Color Spectrum

Red	**1**	**Red**
Orange	**1, 2**	**Red, Yellow**
Yellow	**2**	**Yellow**
Green	**2, 3**	**Yellow, Blue**
Blue	**3**	**Blue**
Indio	**1, 2, 3**	**Red, Yellow, Blue**
Violet	**1, 3**	**Red, Blue**

Thus we see that the color spectrum is based on the relationship between 3 and 7.

There are three primary tones of the musical scale and seven natural tones:

The Musical Scale

C	**1**	**C**
D	**1, 2**	**C, E**
E	**2**	**E**
F	**2, 3**	**E, G**
G	**3**	**G**
A	**1, 2, 3**	**C, E, G**
B	**1, 3**	**C, G**

The notes that correspond to the primary colors are **C, E, G**: the **major triad**, which is **the most important chord in music**, and the groundwork of the scale.

It is important to realize the correspondence between the colors of the spectrum and the notes of the scale, as the musical scale is based on the same principals of nature (the natural overtone series) and it has been adapted in some form or another most all cultures.

There are three root-position triads contained in the **major scale (C,D,E,F,G,A,B).** These are:

The major triad that is based on the correspondence with the primary colors (C, E, G) and **two minor triads, based on three of the natural, or secondary, tones (D, F, A and E,G,B).**

Major and minor triads are the fundamental underpinnings of harmony used in the music of Western cultures.

Three Forces at the heart of Nature

The simplest physical structure in the material universe is the hydrogen atom. It has the least number of forces at work: **three**.

These **three forces** are:

1) At the center of the atom resides the nucleus, composed of a **positive charge** of electricity called a **proton**.

2) Orbiting around this proton in an elliptical orbit is a negative charge of electricity called an **electron**. Since this electron has a negative charge, and since a **negative charge** is attracted to a positive charge, the electron is attracted to the proton.

3) Since the electron is attracted to the proton, there needs to be a force that causes equilibrium between the two, as the electron would collide with the proton if no such force were in place. This force then is the force that governs bodies in motion, causing them to travel in an orbit. It is ***centrifugal force***. In the atom, this is the **neutron**.

Some Aspects of Three

If there is a given object that is fixed in space, this object has, as its location, three coordinates. An example of this would be the relative position of an airplane above the earth. It is calculated by longitude, latitude, and altitude, its three coordinates.

If someone looks into the corner of a room, where the room meets the walls, they will see **three intersecting planes that are considered to extend into infinity**. These are the three dimensions of space.

There are three time zones that man senses, past, present, and future.

If there is space with no motion, then time does not exist, only space. But if there is no motion, then there can be no space, for there has to be space to contain the motion. The motion through space causes sequentially. The earth's orbit around the sun and revolution on its axis causes a time sequence on earth. **Thus we have three aspects: time, space, and motion**.

The Law of Three

The fact that events, laws, and all creation result from three principals is referred to as the "**Law of Three**" or the "**Law of the Triangle**." This law is illustrated by the example that every effect has two causes: **one passive, and one active**.

What does the Russian philosopher Georges Gurdieff have to say about the Law of Three? In the book *In Search of the Miraculous*, PD Ouspensky quotes him as follows:

"This is the **'Law of Three'** or the law of the three principles or the three forces. It consists of the fact that every phenomenon, on whatever scale and in whatever world it may take place, from molecular to cosmic phenomena, is the result of the combination or the meeting of three different and opposing forces.

Contemporary thought realizes the existence of two forces and the necessity of these two forces for the production of a phenomenon: **force and resistance**, **positive and negative magnetism**, **positive and negative electricity, amel and femail cells, and so on**. But it does not observe even these two forces always and everywhere. No question has ever been raised as to the third, or if it has been raised it has scarcely been heard.

"According to real, exact knowledge, once force, or two forces, can never produce a phenomenon. The presence of a third force is necessary, for it is only with the help of a third force that the first two can produce what may be called a phenomenon, no matter in what sphere.

"The teaching of **the three forces is at the root of all ancient systems**. The first force may be called active or positive, the second, passive or negative; the third, neutralizing. But these are merely names, for in reality all three forces are equally active and appear as active, passive and neutralizing, only at their meeting points, that is to say, only in relation to one another at a given moment.

The first two forces are more or less comprehensible to man and the third may sometimes be discovered either at the point of application of the forces, or in the medium, or in the result. But speaking in general, the third forces is not easily accessible to direct observation and understanding.

The reason for this is to be found in the functional limitations of man's ordinary psychological actively and in the fundamental categories of our perception of the phenomenal world, that is, in our sensations of space and time resulting from these limitations. People cannot perceive and observe the third force directly any more than they can spatially perceive the forth dimension.

"The first fundamental law of the universe is the law of three forces,or three principles, or as it is often called, the law of three. According to this law, every action, every phenomenon in all worlds without exception, is the result of a simultaneous action of three forces: the positive, the negative, and the neutralizing. Of this we have already spoken…. The next fundamental law of the universe is the law of seven or the law of octaves.

Some Aspects of Seven

During the late 1700's, some chemists began identifying certain chemical elements. About 1800, they began to determine the atomic weights of some of them. In 1808, John Dalton suggested that atoms were physical objects with specific weights. In 1829, Johann Doebereiner showed that particular elements could be arranged in groups of three. In each group, the elements had similar properties and the weight of the middle atom of the group was close to the average of the other two. These were said to be chemical equivalents to the musical triad.

The Law of Octaves

In 1864, John A. Newlands grouped all the known elements in the order of their atomic weights. He then divided them into groups of seven elements each. He showed that when the atoms were put into order with their weights increasing, there was a repetition in the similarities according to the musical octave, the same found on the keyboard. When he talked to chemists of this time about his newly found law of octaves, they laughed.

But octaves are the substance of the musical scale, the same musical scale that was demonstrated by Pythagoras and used today, with slight modifications according to culture. It is a naturally occurring phenomena. The basis of the musical scale is the *octave*, the repetition of the vibration of the set of seven notes that occurs as the fundamental vibration of the first note, the basic note of that scale, is doubled.

In 1869, Mendeleev compiled a "Periodic Table of the Elements." This table was arranged according to atomic weights. Mendeleev found, as had Newlands, that the chemical properties of the elements reoccur at definite intervals. He concluded that these were the periodic functions of their atomic weights. According to Daniel Morris, in his article *Music of the New Spheres* in the December, 1969 Chemistry Magazine, Mendeleev's arrangement of the elements came to him while listening to a performance of Schumann's exquisite piano quintet. He was seated upon a sofa, apparently mulling over Newland's and Doebereiner's arrangements of the elements, when all of a sudden he jumped up, sat at his desk, and arranged all of the elements accordng to a new plan. Morris' feeling was that there was a definite relationship between the quintet and the periodic table in the repetition and development of the seven-note melody of the quintet.

Vibrations in Different Octaves

It was in 1925 that French scientist Louis Victor Pierre Raymond duc de Broglie discovered that the electron was actually a wave of vibration. These waves properly follow the same property and characteristics of vibration as the vibrating string of a musical instrument, the waves being complete with overtones. (We are familiar with the presence of overtones on the vibrating string from the experiments of Pythagoras and his *monochord* – a single-stringed instrument he built to display how musical scales were produced by the natural overtone series).

An atom such as uranium has more that 90 electrons, each capable of its own harmonies, and all this can be recorded by a spectroscope. From this, one can conclude that all matter is made of waves. Therefore, we, and the world we live in, are vibrations…just as the children of the 1960s discovered while taking large doses of LSD! But this should not be too surprising. Our understanding of the world around us is through perception of the five senses. What we experience is actually taking place in our brain, interpretations of waves of light, taste, smell, feeling, and sound. Are we really living in a world, as we suspect, or is this simply a world created in our brain, similar to that when we dream? And if spectral and aural vibrations obey the same law of octaves that applies to the physical makeup of the universe, then perhaps it is all just a part of one large keyboard of vibrations, with particular sensations and manifestations taking place in particular octave ranges. Light, heat, magnetic and chemical vibration are all subject to the law of octaves.

The ancient Greek philosopher Democritus wrote:

Sweet and bitter
Cold and warm,
As well as all the colors.
All these things exist
But in the opinion
And not in reality.

What really exists
Are unchangeable articles: atoms,
And their motions in empty space.

The great German mathematician Gottfried Wilhelm Leibniz wrote: “I am able to prove that not only light, color, heat, and the like, but motion, shape and extension too are mere apparent qualities.”

Many scientists and philosophers have come to the conclusion that everything exists in the mind: stars, planets, atoms. Einstein showed us that even space and time were forms of intuition.

Georges Gurdjieff, the important Russian spiritual teacher said that the law of octaves was based upon the discontinuity of all vibrations in nature, as they do not develop uniformly, but with periodic accelerations and retardations.

The original impulse becomes alternately stronger and weaker. This could be a phenomenon similar to the AC cycle. This cycle of alternating electrical vibrations can be viewed by means of an oscilloscope or with a computer program such as Sound Forge. Called the duty cycle, it appears as such.

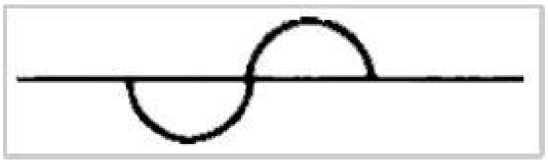

All sounds are made up of these cycles.

The cycles take place in the ***time domain***. By using a brilliant mathematical formula known as the Fourier transform, these cycles can be translated into the frequency domain, where we find that the timber of the sound depends upon the strength of overtones, the higher harmonics of the fundamental tone.

The duty cycle can be equated to a circle broken at its diameter with the bottom half twisted 180 degrees or to the continuous movement of a spiral.

Alternating current (AC) obeys the laws of the duty cycle: energy alternates with no energy.

Home electricity in the United States alternates sixty times per second, thus it is called 60 cycles, or 60 Hertz.

An electric light bulb does not continually burn, but flashes sixty times in one second, faster than our eyes can follow, so we perceive the appearance only of continuous light. This would not be the case when light is powered by a battery, as batteries produce direct current (DC).

Musical tones can be represented using the mathematical symbols of trigonometry, signs and cosigns, and the imaginary numbers of calculus.

Computer systems that generate musical tones, such as synthesizers, do so using these mathematical formulas. All of this is based on the alternation of the duty cycle.

“Shocks”

While teaching in Russia during the first world war, Gurdjieff described something about octaves that had not been publically known before. He stated that there were two places in the octave where retardation of the vibrations takes place.

These places are between the notes **E** and **F** and between the notes **B** and C. These are the two places on the keyboard where there are no sharp or flat note (black) keys between the white keys.

Octaves, according to Gurdjieff, develop according to whether or not these places where retardation of vibrations take place are “filled.”

In the right development of these octaves, these places are filled by the energy of octaves that run in parallel to the given octave.

This process of filling in Gurdjieff calls a shock. Processes, institutions, laws and other things all progress through time according to the law of octaves, and therefore these two shocks in the octave are a part of the development process.

Gurdjieff stated that if development continues without the shocks being filled in, then the line of development constantly changes.

The way he described it, if octaves – as a principal of nature – develop unimpeded, the development looks as if it is proceeding along a straight line, but every time one of the two points in the octave is reached where a shock occurs, even though development of the process that the octave represents appears to be progressing forward along a straight line, the development of the process actually jogs a little, causing what would be compared to a jog in the line.

The development continues and after the progress through the series of octaves has gone on for a particular amount of time, what was considered to be development along a straight line has actually turned into a development in the ***opposite direction***! Further continuation will bring the development back to the same direction that it started in. This is actually a full duty cycle, or one full revolution on a spiral.

The law of octaves, with this intrinsic circular motion generating a spiral, governs all activity and shows how organizations, religions, and governments -- if allowed to develop without conscious knowledge concerning the application of the ‘filling in’ of these shocks -- always develop into something that is opposite to what was originally intended.

The Octave with Two Added Shocks

Gurdjieff explained how an octave of seven ‘notes’ was transformed into one of nine, the seven notes and the two shocks, by applying the following fascinating bit of arithmetic.

He divided the number **‘1’** into seven equal fractions: **1/7, 2/7, 3/7, 4/7, 5/7, 6/7, 7/7**

The decimal equivalents of these seven fractions are arrived at using the time-honored tradition of dividing the numerators by the denominator:

1/7 = .142857
2/7 = .285714
3/7 = .428571
4/7 = .571428
5/7 = .714285
6/7 = .857142

Of course 7/7 equals 1

Studying the numbers that result from our division is a fascinating process.

We find that we have a reoccurring sequence of numbers that is present in every case!

This is the sequence 7142857142857….that continues on and on ad infinitum.

.142857142857142857 etc….
7/1.0000000000000000

(Numerical patterns such as these are very telling.

They describe a situation that is naturally occurring and should be taken seriously by the researcher).

The nature of this pattern is very interesting. It starts with the number **'7'**, which is the number of the octave, the number that we are dealing with here. Next in line is double that number, **'14'**.

Following that we have another doubling**: '28'**, and after that another doubling: **'56'**…whoops, not **56**, but **57**, the **'7'** being the start of another sequence!

It takes some thought about this to fully appreciate the beauty of this sequence and this change from **56** to **57** to start another series. This just really ***breathes*** "cycle," does it not?

There are other interesting things about this series of numbers. If we add each line in the set of six decimals above, we always get the number **'28'** as a result.

This is because each of the six decimal sequences is ordered in such a way that when you add downwards, only the six numbers of the sequence appear, there are no duplications.

We have six pieces of the whole **(7/7),** each with a sequence of six numbers, the same sequence in each piece, ordered in the very unique fashion. This is something to contemplate.

Now we will turn to something called horizontal addition. This is a process used in numerology.

Although we don't endorse the common understanding of what numerology represents, the principles behind manipulation of numbers are simple aspects of natural mathematics, which demonstrate the laws of nature, and this understanding of numbers and what they are all about is the key to understanding nature and what it is all about.

That is why trigonometry and calculus are so important to engineering.

Both these sciences give us an understanding of nature, and hence, an understanding of sound, and the natural laws that govern the use of sound in music…principals that are not arbitrary, but work according to natural law, the law of three, the law of octaves, and now the law of nine (the octave plus the two shocks).

By using horizontal addition on the sequence, we arrive at another number:

1 + 4 + 2 + 8 + 5 + 7 = 27

2 + 7 = 9

Adding the sequence numbers together, we end up with '9'. This demonstrates one way that seven ties into nine.

The Periodicity of Nine

The notes of the octave combined with the two shocks show us what is called the ***periodicity of nine****.*

Creating a circle with nine positions and connecting the numbers from our sequence above, we have the figure that Gurdjieff calls the ***enneagram*****.**

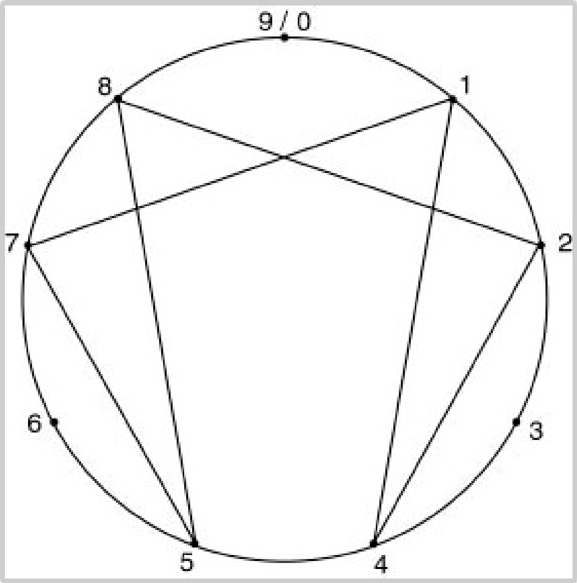
9 / 0
1
2
3
4
5
6
7
8

The numbers that are not in the sequence are 3, 6, and 9. When a they are connected, the result is a triangle. The numbers 3 and 6 are the shocks.

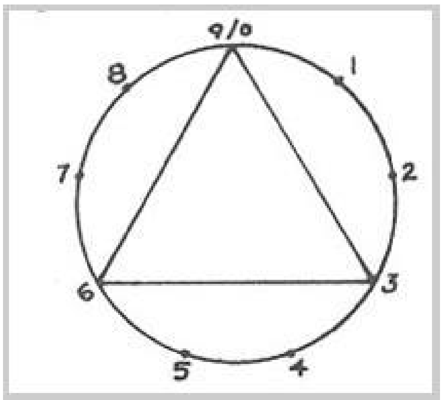
9/0
1
2
3
4
5
6
7
8

By applying the color spectrum and musical notes to the enneagram, we have the following:

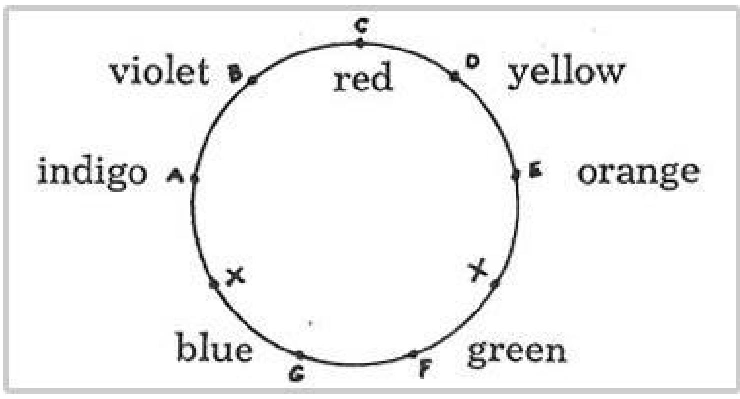
C
red
D
yellow
E
orange
X
F
green
G
blue
X
A
indigo
B
violet

When asked why one shock correctly appeared between the E and the **F**, but the other, instead of appearing between **B** and **C**, appeared between G and **A,** Gurdjieff stated that this demonstrated the nature of the shock that occurred when passing from one octave to the next.

He also said that every completed whole, every organism, is an enneagram, but that each of these enneagrams does not have an inner triangle (**3,6,9**).

142857 is the working of the corporeal scheme, the triangle is the presence of higher elements in the organism. This inner triangle is possessed by plants, for example, from which come our “drugs”: hemp, poppy, hops, tea, coffee, tobacco.

When any number is added to nine, the sum, added horizontally to one significant figure, always equals that same number (also added horizontally to one significant figure).

Examples:

8 + 9 = 17 and 1 + 7 = 8

6 + 9 = 15 and 1 + 5 = 6

36 + 9 = 45 and 4 + 5 = 9 and 3 + 6 = 9

When any number is multiplied by nine, the product added horizontally to one significant figure, always equals nine.

Examples:

9 x 7 = 63 and 6 + 3 = 9

9 x 12 = 108 and 1 + 0 + 8 = 9

Casting Out Nines

Further use of horizontal addition is applied in a mathematical process that is called *casting* ***out nines.*** The sum of any number added horizontally is equal to the remainder of that same number when it is divided by nine. Thus, to cast out all nines from a number, add it horizontally:

Example: **16,378 = 1 + 6 + 3 + 7 + 8 = 25 = 7**

If you were to divide **16,378 by 9**, you will find that the **remainder is 7.**

The Enneagram with Combinations of Nine

This interesting drawing shows how multiplication of nine by the other single-digit numbers produces a sum that fits nicely into our enneagram. We have equated nine with zero because nine becomes zero when it is cast. Notice how the sums are equal to the two numbers at that level of the drawing (for example, the sums on the first line are **18** and 81, on the second **72** and **27** and so on).

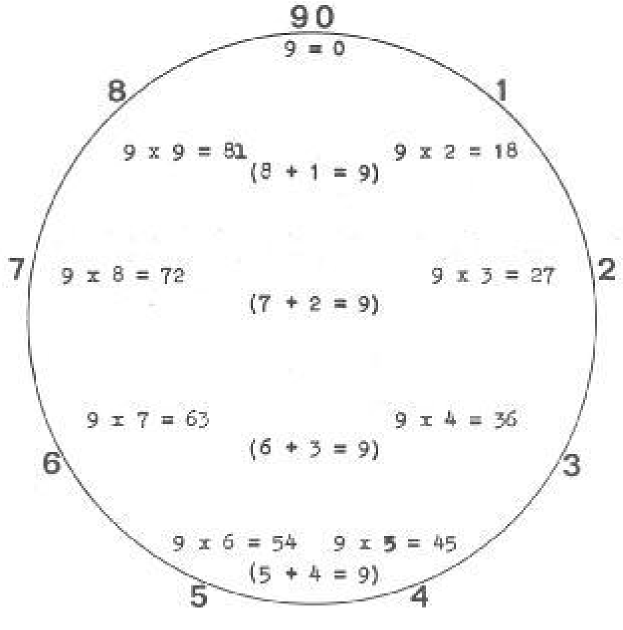

We see that by horizontal addition of the sums, in each case, we always end up with nine. If we now take each number from one to eight and multiply it by numbers **1** through **9** – we will see some interesting numerical patterns develop.

We will start with the first line from the circle above:

8 x 1 = 8
8 x 2 = 7 (16, or 1 + 6 = 7)
8 x 3 = 6 (24, or 2 + 4 = 6, and so on…)
8 x 4 = 5
8 x 5 = 4
8 x 6 = 3
8 x 7 = 2
8 x 8 = 1
8 x 9 = 9

We end up now with a sequence of 876543219.

A similar sequence is produced by doing the same thing with 1.

This sequence is the reverse of the other:

1 x 1 = 1
1 x 2 = 2
1 x 3 = 3
1 x 4 = 4
1 x 5 = 5
1 x 6 = 6
1 x 7 = 7
1 x 8 = 8
1 x 9 = 9

This is how this sequence looks on the circle:

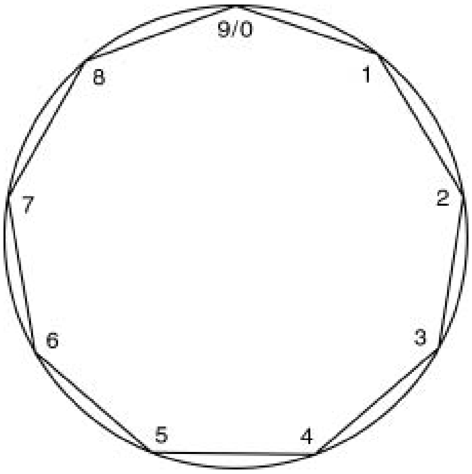

Notice that this sequence was generated by the numbers **8** and **1**, the two numbers at the top of the enneagram.

The number series that resulted from these two was the same, however one series is the reverse of the other. **For 8** we have: **876543219**, and **for 1** we have **12345789**.

To be completely the opposite of each other, the second series would have had to have been **987654321**, but interestingly enough, this isn't the case.

Now we will generate another pair of sequences using seven and two, the next set of numbers in the enneagram.

7 x 1 = 7	**2 x 1 = 2**
7 x 2 = 5	**2 x 2 = 4**
7 x 3 = 3	**2 x 3 = 6**
7 x 4 = 1	**2 x 4 = 8**
7 x 5 = 8	**2 x 5 = 1**
7 x 6 = 6	**2 x 6 = 3**
7 x 7 = 4	**2 x 7 = 5**
7 x 8 = 2	**2 x 8 = 7**
7 x 9 = 9	**2 x 9 = 9**

This is how this sequence looks on the circle:

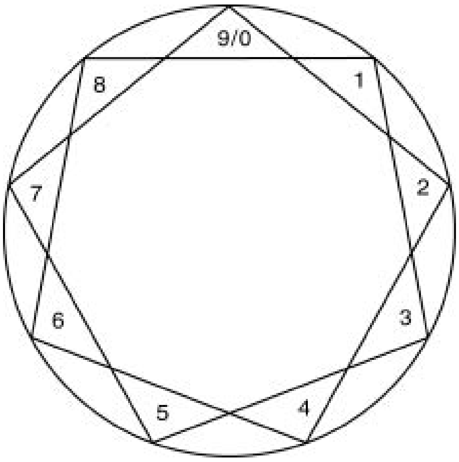

Again we have a resultant series of numbers where one is the opposite of the other.

For 7 we have 753186429, and for 2 we have 246813579.

As is the case **with 8 and 1 above, both series always ends with a 9!**

Also, the reverse series is one number off from the original.

This series of numbers is one that skips every other number.

Now we will generate another pair of sequences using five and four:

5 x 1 = 5	**4 x 1 = 4**
5 x 2 = 1	**4 x 2 = 8**
5 x 3 = 6	**4 x 3 = 3**
5 x 4 = 2	**4 x 4 = 7**
5 x 5 = 7	**4 x 5 = 2**
5 x 6 = 3	**4 x 6 = 6**
5 x 7 = 8	**4 x 7 = 1**
5 x 8 = 4	**4 x 8 = 5**
5 x 9 = 9	**4 x 9 = 9**

This is how this sequence looks on the circle:

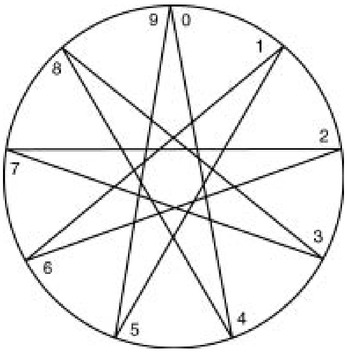

When we look at these, we see again that we have the same kind of pattern where the series from the second set is the opposite of the first, however offset by a single number, and that both end in **9**.

The series **for 5 is 516273849** and the series **for 4 is 483726159**. This series of numbers is one where three numbers is skipped each time.

Finally we will take a look at the sequence generated by the two shock values **three** and **six**:

6 x 1 = 6 **6 x 2 = 3** **6 x 3 = 9** **6 x 4 = 6** **6 x 5 = 3** **6 x 6 = 9** **6 x 7 = 6** **6 x 8 = 3** **6 x 9 = 9**	**3 x 1 = 3** **3 x 2 = 6** **3 x 3 = 9** **3 x 4 = 3** **3 x 5 = 6** **3 x 6 = 9** **3 x 7 = 3** **3 x 8 = 6** **3 x 9 = 9**

This is how this sequence looks:

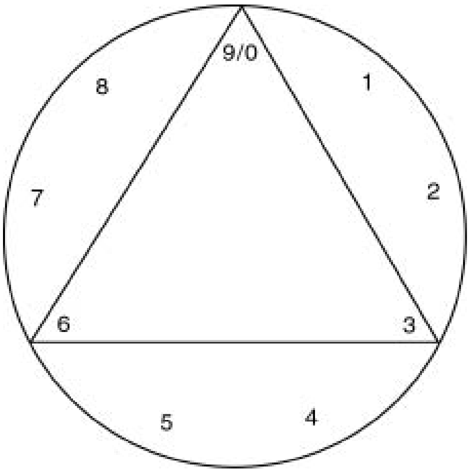

Interestingly enough, **the number 6 yields a series of 639, 639, 639**, and the series from 3 yields the opposite: **369, 369, 369. Who says the universe isn't the result of mathematics!**

We have now horizontally multiplied each number from 1 to 9 by the numbers 1 through 9 and have arrived at a series that is in each case has been illustrated using the enneagram circle, the circle with 9 points.

The numbers **1,2,4,5,7,** and **9** are the numbers found in the period **(142857).** When we look at the tables that we have just created using these numbers (that is the tables for "**8 and 1**", "**7 and 2**", and "**5 and 4**") we can see that a series of numbers has been created in each table.

For example, in the 8 table, the results from horizontally **multiplying 1 through 9 by 8** resulted in a pattern (reading down the sum column) of **876543219** and the results from horizontally **multiplying 7 by 1 through 9 is the sequence 753186429.**

Now, taking the **third**, **sixth**, and **ninth** numbers in each of these series, we discover that we have a repeating sequence of either **639** or **369**. These are the "**inner triangles**" of each one of these sets

8 x 1 = 8	**1 x 1 = 1**	**7 x 1 = 7**	**2 x 1 = 2**	**5 x 1 = 5**	**4 x 1 = 4**
8 x 2 = 7	**1 x 2 = 2**	**7 x 2 = 5**	**2 x 2 = 4**	**5 x 2 = 1**	**4 x 2 = 8**
8 x 3 = 6	**1 x 3 = 3**	**7 x 3 = 3**	**2 x 3 = 6**	**5 x 3 = 6**	**4 x 3 = 3**
8 x 4 = 5	**1 x 4 = 4**	**7 x 4 = 1**	**2 x 4 = 8**	**5 x 4 = 2**	**4 x 4 = 7**
8 x 5 = 4	**1 x 5 = 5**	**7 x 5 = 8**	**2 x 5 = 1**	**5 x 5 = 7**	**4 x 5 = 2**
8 x 6 = 3	**1 x 6 = 6**	**7 x 6 = 6**	**2 x 6 = 3**	**5 x 6 = 3**	**4 x 6 = 6**
8 x 7 = 2	**1 x 7 = 7**	**7 x 7 = 4**	**2 x 7 = 5**	**5 x 7 = 8**	**4 x 7 = 1**
8 x 8 = 1	**1 x 8 = 8**	**7 x 8 = 2**	**2 x 8 = 7**	**5 x 8 = 4**	**4 x 8 = 5**
8 x 9 = 9	**1 x 9 = 9**	**7 x 9 = 9**	**2 x 9 = 9**	**5 x 9 = 9**	**4 x 9 = 9**
order: 639	**order: 369**	**order: 369**	**order: 639**	**order: 639**	**order: 369**

We will now show each number and the order of its series.

This order, as you have just seen, was determined by the inner triangle of each series (**the points of 3, 6 and 9 on the circle**) and number present at those point turn out to be either a sequence of **639** or **369**.

639 represents energy moving counter clockwise and **369** is the **energy moving clockwise.**

For purposes of demonstration, we will fix polarity designations to the **639** and **369**.

We will call **369 positive** and **639 negative**.

The numbers **3**, **6** and **9**, not being part of the periods, we will call neutral **(=)**:

Note	Number	Order	Polarity
C	0		=
D	1	369	+
E	2	639	-
(shock)	3		=
F	4	369	+
G	5	639	-
(shock)	6		=
A	7	369	+
B	8	639	-
C	9		=

We will turning again to the sequences that we created above.

For simplicity sake, we will just use the forward sequence from each set.

These sequences were:

123456789 (from 8-1),
135792468 (from 7-2)
639639639 (from 6-3) and
159483726 (from 5-4).

Applying the polarities we assigned in the chart above to these sequences we get:

8-1	**1 2 3 4 5 6 7 8 9**
	+ - = + - = + - =
7-2	**1 3 5 7 9 2 4 6 8**
	+ = - + = - + = -
6-3	**6 3 9 6 3 9 6 3 9**
	(all neutral)
5-4	**1 5 9 4 8 3 7 2 6**
	+ - = + - = + - =

[1] It is noted that modern theory of colors has adapted new colors to represent the subtractive domain of primary colors, and some writers go so far as to dispute the standard primary colors, but we stand by what we have written: red, yellow and blue are the primary colors.

For more information about modern color theories, see: http://www.goatrance.de/mahak/colors.html
http://www.physicsclassroom.com/Class/light/U12L2d.html

UNIVERSAL
LAWS
NEVER BEFORE
REVEALED:

KEELY'S
SECRETS

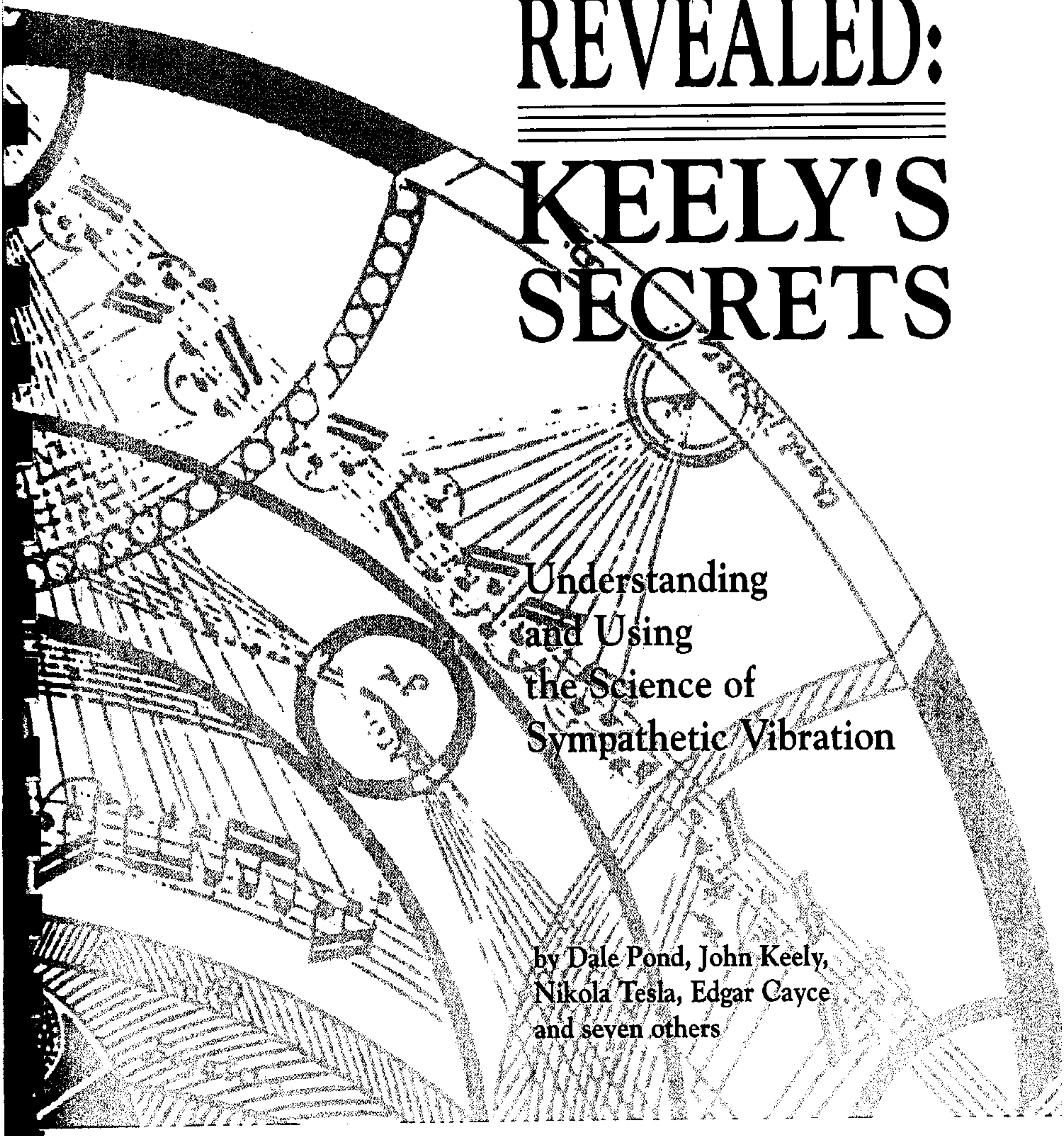

Understanding
and Using
the Science of
Sympathetic Vibration

by Dale Pond, John Keely,
Nikola Tesla, Edgar Cayce
and seven others

NEW REVISED EDITION

Copyright © 1990, 2000 by Dale Pond
All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical or other means now known or hereafter invented, including information storage and retrieval systems, photocopying and recording without the written permission of the publisher except for brief quotations embodied in critical articles, reviews or works of scholarship. The authors, editor and publisher disclaim all liability incurred from the use or misuse of information in this book.

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 94-77879
ISBN: 1-57282-003-9 (Paperback)
Cover Design by Janice St. Marie
Book Design by James Berry

Publisher's Cataloging in Publication

Universal laws never before revealed : Keely's secrets : understanding
and using the science of sympathetic vibration / by Dale Pond...
[et al.] ; edited by Dale Pond. -- 2nd ed., new rev. ed.
p. cm.
Includes bibliographical references and index.
Preassigned LCCN: 94-77879.
ISBN: 1-57282-003-9 (pbk.)

1.Keely, John Ernest Worrell, 1827-1898. 2. Vibration. 3.
Keely motor. I. Pond, Dale.

QC231.U65 1995 620.2
QB195-20616

Published and distributed by:

The Message Company
4 Camnino Azul
Santa Fe, NM 87505
505-474-0998
505-474-7604

Printed in the United States of America
Printed on acid-free recycled paper using soy ink.

ACKNOWLEDGEMENTS

I wish to thank all those who have helped directly and indirectly in the preparation of this work. Especially those lonely geniuses of old who bore the brunt of derision and ridicule, never faltering in the thankless task of bringing the Light of Understanding to the Minds of Men. With special gratitude to the memory of Mrs. Clara Sophia Jessup Bloomfield-Moore in her courage and strength to stand by such a Man as John Ernst Worrell Keely, even in the face of public scorn, and humiliation. No less thanks go to The Theosophical Society* for their Light-Bearing efforts, The Rosicrucian Order AMORC*, for their centuries of thankless toil and to the Association for Research and Enlightenment and its employees for their support and encouragement. No less do we thank those unnamed ones who have withstood the fury of ignorance of past ages in order that Man may Evolve to his rightful place. It is with truth and humility that "The few carry the many."

This book contains a vast collection of material on Keely and his work. In it has been included all the pictures and charts known to exist as of this date. I have also included all that we have been able to decipher of this wonderful new science and philosophy. There will be more discoveries and these will be published at a later date. The discoveries mentioned here demonstrate that Keely's science of *Sympathetic Vibratory Physics* is a wholistic approach to science and nature and reveals a wonderful order and simplicity.

It is hoped readers will find a number of ideas and methods in these pages which they can apply to their current scientific work and research.

Dale Pond

***Author's Note:** There has been a great deal of discussion concerning the motives and belief systems of these two organizations. It is taken herein that the *original* intent and purpose of these groups was the preservation and disseminations of *real TRUTH* from the ancient days to the present. We, as human beings, can now more properly partake and use the forces and knowledge contained in their *original* knowledge bases. It is further recognized that these organizations may have more or less "lost their way" during the past two or three centuries. Perhaps the present day motivations and directions being assigned to them are in error – perhaps not. It is not to these I give thanks but to their original intent and accomplishments. For, it is, I believe, through the efforts of these *original* organizations that certain specific and valid information was indeed preserved and brought through the centuries of ignorant persecution that we, through Keely's efforts (and others'), may understand and use same. *Dale Pond*

ABOUT THE AUTHOR/EDITOR

Dale Pond is truly a renaissance man with a diversified background in chemistry, physics, mechanical engineering, computers, mathematics, acoustics, hydrodynamics, geometry, music, and common law.

His vast experience includes technical writing, inventing, lecturing, corporate training, business, manufacturing, publishing and authoring scientific and philosophical journals and books.

At the leading edge of new paradigm thinking, he has been able to show that vibration is the principle underlying all things. Vibration is the link between spirit and matter - the magic key for creating new worlds and universes.

As an internationally renowned lecturer, Dale has presented workshops and seminars at the Swiss Association for Free Energy Conference, the International Keely Conferences, Sound and Vibration Conference, International Symposium on New Science, International Forum on New Energy, United States Psychotronics Association Conference, the Alchemy of Sound Conference and many others.

Other books by Mr. Pond:

Sympathetic Vibratory Physics - A Compendium of Terms and Phrases
Sympathetic Vibratory Physics Certification and Training Course - Level I
Nikola Tesla's Earthquake Machine
HyperVibes (a Hypercard research reference library for SVP)
A History of the American Constitutional or Common Law
Our American Common Law
The Physics of Love
John Keely's Musical Dynasphere

Table Of Contents

Illustrations List 6
Disintegration of Stone 9
by R. Harte
Etheric Force Identified as Dynaspheric Force 17
by Mrs. Clara Jessup Bloomfield-Moore
One Phase of Keely's Discovery
in its Relations to the Cure of Disease 29
by Mrs. Clara Jessup Bloomfield-Moore
On Dynaspheric Force 39
by Laurence Oliphant
What is Matter and What is Force? 43
The Astral Light 51
by Louise A. Off
Mr. Plum's visit to Keely's Laboratory 54
from Dashed Against the Rock
Pyramids, Telescopes and Light 62
by John Keely, from Dashed Against the Rock
Levitation 64
by John Keely, from Dashed Against the Rock
Amplitude of Force 66
by John Keely, from Dashed Against the Rock
The Chord Settings of Life 70
by John Keely, from Dashed Against the Rock
The Dynamics of Mind 72
by Henry Wood
The Neutral Center 78
by John Ernst Worrell Keely
Theory and Formula of Aqueous Disintegration 80
by John Keely
Keely's Motor Secret Revealed at Long Last 84
By Dale Pond
Sympathetic Vibratory Physics 86
by Dale Pond
Keely's Definitions of His Own Terms 91
by John Keely
The Basis of a New Science
by John Keely 94
Law of Matter and Force 96
Law of Corporeal Vibrations 100
Law of Corporeal Oscillations 102
Law of Harmonic Vibrations 106
Law of Transmissive Vibraic Energy 108
Law of Sympathetic Oscillation 112
Law of Attraction 114
Law of Repulsion 114
Law of Cycles 116

Latent Force in Interstitial Spaces 116
by John Keely
The Role of Atomic Forces in Healing 118
by John Keely
Law of Harmonic Pitch 122
Law of Force 124
Law of Oscillating Atomic Substances 126
Law of Sono-thermity 128
Law of Oscillating Atoms 130
Law of Vibrating Atomolic Substances 132
Gravity Quotes 134
by Dale Pond
Polarity of Science 136
by Dale Pond
Law of Oscillating Atomoles 140
What is Gravity? 140
Law of Transformation of Forces 142
Law of Atomic Pitch 146
Law of Variation of Atomic Pitch by Rad-energy 150
Law of Variation of Atomic Pitch by Electricity and Magnetism 154
Law of Variation of Atomic Pitch by Temperature 156
Law of Pitch of Atomic Oscillation 158
Mathematical Demonstration of the Size of an Atom, Weight, Volume 158
By John Keely
Law of Variation of Pitch of Atomic Oscillation by Pressure 160
Law of Variation of Atomic Oscillation by Temperature 162
Law of Variation of Atomic Oscillation by Electricity 164
Keely's Discoveries Vindicated? 165
by Dale Pond
Law of Variation of Atomic Oscillation by Sono-thermism 166
Law of Chemical Affinity 168
Law of Chemical Dissociation 170
Law of Chemical Transposition 172
Law of Chemical Substitution 174
Law of Catalysis 176
Law of Molecular Synthesis and Combination 178
Relativity of Circles and Triangles 178
by Dale Pond
Neutral Center Defined 179
by Dale Pond
Law of Chemical Morphology 180
Keely's Chords and Synthesized Wave Forms 181
by Dale Pond
Mr. Keely's New Philosophy 183
by Clara Jessup Bloomfield-Moore
Man's Greatest Achievement 187
by Nikola Tesla
Law of Atomic Dissociation 188
Law of Atomolic Synthesis of Chemical Elements 190
Mastery of Self is Mastery of Science 191
by Dale Pond
Dangers of Ultrasonic Energy 192
by Dale Pond
Law of Heat 194
Law of Electro-Chemical Equivalents 196
Law of Cohesion 198
Law of Refractive Indices 200

Law of Electric Conductivity 202
Music: A Vibration Analysis Shortcut 204
by Dale Pond
Scale of Forces in Octaves 208
by John Keely
Induction, Sympathy & Resonance 212
by Dale Pond
Gravity 214
by Edgar Cayce & John Keely
Music of the Molecule 218
by Dale Pond
Much Ado About Nothing 220
by Dale Pond
Music - The Realm of Activity 224
by Dale Pond
Vibration vs Oscillation 225
by Dale Pond
The Neutral Force 225
by Dale Pond
Polarity 226
by Dale Pond
Clara Jessup Bloomfield-Moore 228
by Dale Pond
Keely's Trexar - A Superconductive Wire 232
by Dale Pond
Musical Intervals 233
by Dale Pond
John Keely - A Personal Interview 236
by Dr. Hall, Scientific Arena
A Second Visit to Mr. Keely 240
by Dr. Hall, Scientific Arena
Mr. Keely's Researches - Sound Shown to be a Substantial Force 244
by Henry B. Hudson
The Keely Motor Illustrated 248
by Henry B. Hudson
Deciphering Keely's Harmonic & Inharmonic Chords 254
by Dale Pond
Keely's Triple Chords 258
by Dale Pond
Keely's Acoustic Theoretical Charts 260
by John Keely

ILLUSTRATIONS

Chart defining angles of radiation and the full line of quadruple settings for full Etheric circuits, as also the Chords of Neutral Centre. .... Front Cover

Chart Showing Key to Vibratory Rotation 16
Triple Molecule Configuration 21
Chart defining angles of radiation and the full line of quadruple settings for full Etheric circuits, as also the Chords of Neutral Centre. 28
Keely's Compound Disintegrator 38
Keely in His Lab 42
Etheric Force Generator and Etheric Motor 50
Circle of Musical Notes with Colors 54
Gravity Overcome with Vibration 60
Keely's Vibratory Switch 61
External View of Atom 63
Three Dynaspheric Envelopes 63
Triple Atoms in Molecule 63
Lines of Interference of Triple Envelopes 63
Chart Showing the triple condition of the vibratory flows governing the molecule during disruption. 1st Order of Vibration 65
Original Keely Patented Motor Drawing (top view) 68
Original Keely Patented Motor Drawing (front view) 69

In the Beginning was the Word 83
Etheric Vapor Lifting Weight Experiment 93
Keely's Musical Sphere Motor 95
Keely and His Musical Sphere Motor 99
Vibration 103
Oscillation 103
Vibration Illustrations 104
Compound Disintegrator 105
Vibratory Globe and Resonator 105
Medium for testing vibration under different orders of evolution 105
Law of One 107
Sympathetic Negative Attractor 111
Rotating Resonator, weight 780 lbs. and Vibratory Globe 120
Vibratory Globe and Accelerator 121
Vibratory Planetary Globe 139
Keely, His Final Motor and Dynamo 145
Etheric Chord 8th Octave 148
Quadruple Chord 3rd Octave 148
Musical Sphere Motor (internal view) 149
Chord of Equation 5th Octave 152
Sympathetic Chord of Transmission 152
Chord of Radiation 1st Octave 152

Chord of A flat 1st Octave 152
Neutral Chord any Octave 152
Hydro Pneumo Vacuo Motor (figure 1) 153
Hydro Pneumo Vacuo Motor (figure 2 & 3) 155
Hydro Pneumo Vacuo Motor (figure 4) 157
Hydro Pneumo Vacuo Motor (figure 5, 6, 7, 8 & 9) 159
Computation of Etheric Force exerted on lever 161
Levitation Apparatus? 163
Keely's Atomic Triplets 165
Feyman's Atomic Triplets 165
Hydro Pneumo Vacuo Motor (still existing) 167
Dynaspheric Shells on Molecule 168
Hydro Pneumo Vacuo Motor (still existing) 169
Triangle and Circle Relationships 171
Compound Disintegrator 173
Keely's Provisional Engine 175
Liberator 177
Nestled Triangles and Circles 179
Music Staffs 179
Neutral Center 179
Three Different Waveforms 182
Harmonic Series 182
Amplitude Modulation 182
Frequency Modulation 182
Keely's Molecular Morphology 189

Vibratory Indicator 195
Attractive Disks 195
Musical Globe 195
Medium for testing Centripetal Force 195
Pneumatic Gun 199
Modern Molecular Morphology 201
Keely with Vibrodyne Motor 207
Keely's Lab 210
Unidentified Machinery 211
Globe Motor and Provisional Engine 217
Musical Ratios of Sphere 218
Musical Ratios of Ovoid 219
Globe Motor and Provisional Engine 223
Keely's Provisional Engine 231
Keely's Lab 235
Generator of Etheric Vapor 247
Keely and His Etheric Vapor Cannon 252
Sketch of Motor 253
Waveform 254
Music Staff 254
An Organ Pipe Curve and its Components 255
Three different waveforms 255
Harmonic Series 255
Musical Scale 255
Compound Waveform 256
Zither and Anti-gravity Experiment 257

KEELY'S ACOUSTIC THEORETICAL CHARTS

Chart showing the triple conditions of the vibratory flows governing the molecule during disruption. 1st Order Vibration. 65
Title Page of booklet 260
John Keely in His Laboratory 261
Symbols and Conditions of Vibratory Streams 262
Angle of Vibratory Circuit 263
Chart of the Angles of Interference and Coincidence Trigeminal 264
Molecular Radiation, first line of interference. 265
Chart showing the conditions governing the transmittive link of sympathy between neutral centre and periphery of vibratory circuit. 266
Chart showing the conditions governing the discordants on all their combinations 267
Chart showing the conditions governing harmonious chords on the different octaves quadruple and double quadruple also positive and negative radiation from neutral centre and equation of concordants under antagonism. 268
The Brain as applied to Vibratory Etheric Science 269
Chart defining the different chord associations on the three Octaves, as also interferences on single and double chord bar and theory of setting chords of masses 270
Chart defining the angles of radiation on the full line of quadruple settings for full Etheric circuits, as also the Chords of Neutral Centre 271
Key to Vibratory Rotation 272
Chart of Differentiation for setting Chords on Vibratory Bar 273
Chart defining the arrangement of the different atoms and corpuscles of matter as contained in their respective envelopes free of intensified vibration. 274
Chart with symbols defining the relative simple and compound sympathetic association between the different orders of etheric chords as associated with the Molecular, Atomic and Interetheric to induce dispersion under the progressive orders of vibration both positive andnegative as also the different combined orders of radiation to induce luminosity 275
Vibrophonic system of setting vibrators on the diminishing line to neutral centre of full vibratory circuit embracing three lines of sub-division, viz: molecular, atomic and etheric. 276
Chart of the Sinuses and nerves of the Skull vibro-etherically considered as associated with the Liberator. 277
Chart showing Base of the Brain 278

Disintegration of Stone

1888

A short time ago the mining world in America was seized with an inexplicable excitement. The value of gold mines in particular suddenly rose. Mines long since abandoned on account of the expense of working, awoke, and rubbing their eyes made their way again into the stock list. Presently it leaked out that a syndicate of the longest-headed and wealthiest mining capitalists were quietly buying up all the cheap and apparently worthless gold mines they could hear of, and people at once concluded that something was up. Then everyone of a speculative turn, very knowingly began to buy worthless gold mining shares at ever-increasing prices, and when the little speculators had gorged themselves to the full extent of their financial capacity, they asked: What next? No one knew exactly what he was after; and everyone looked to the Syndicate for the next move; but the Syndicate smilingly put its hands in its pockets and whistled! After the fever came prostration. The small fry had not, like the Syndicate, bought to hold, so they got first uneasy, then alarmed, and finally panic-stricken. The tide of credulity turned and began to run out even more quickly than it had set in, and thousands of the unlucky, but greedy little grudgeon of the Stock Exchange were left stranded in a short time by the receding tide of speculation, kicking and gasping in the mortal agonies of financial asphyxia. The panic is easily accounted for by the general laws that govern the movements of the Stock Exchange; but not so the action of the syndicate. The problem remains: Why did the long-headed millionaires buy up worthless mines? That is the point of interest, and the explanation thereof is as follows:---

A few weeks before the panic occurred, twelve solid men - millionaires - met by appointment in a certain laboratory in Philadelphia to witness an exhibition of the disintegration of quartz by a new method. They were mining magnates, who had a tremendous interest in getting the gold out of quartz rock quickly and cheaply. The inventor obliged them by simply touching some blocks of quartz with a little machine he held in his hand; and as he touched each block it instantly crumbled into atomic dust, in which the specks of gold it had contained stood out like boulders in a bed of sand. Then the twelve solid men solidly said: Mr. Keely, if you will in the same manner disintegrate some quartz for us in its natural place, we will each of us give you a cheque for --- dollars. So off they all went to the Katskill mountains, and there the twelve solid men pointed out a reef of quartz on the side of a mountain, as solid as themselves; and Mr. Keely took out his little machine and said: Gentlemen, please take the time. In eighteen minutes there was a tunnel in that quartz mountain eighteen feet long and four and a half feet in diameter. Then Mr. Keely quietly returned to Philadelphia with his cheques in his pocket, and the twelve solid men went from New York to San Francisco to gather in the seemingly worthless stock of mines long disused because of the working expense, thus producing the disastrous effect upon the mining world, which we have just seen. (All these men bound themselves to secrecy; and this is the first time that this incident has been made public.)

How was the quartz disintegrated?--That is one of Keely's secrets.

The disintegration of the rock is, however, a very small and accidental effect of that tremendous force that lies behind the secret. Indeed, that particular application of the force was a chance discovery. One day the inventor was studying the action of currents of ether playing over a floor upon which he had scattered fine sand,---the ether was rolling the sand into ropes,---when a block of granite, which was used for fastening back a door, disintegrated under his eyes. He took the hint, and in a few days he had made a vibratory disintegrater.

Who is this man, and what is this force? to whom, or to which, boring a tunnel into the mountain side is mere child's play? Surely, were such things true, science would long ago have filled the world with the renown of such a man---the man who has discovered a force in nature compared to which all known motor or mechanical forces are like the scratch of a nail, or the breath of a child. Surely the press, the platform, and even the pulpit would have resounded with the glad tidings of so great a victory over the stubborn powers of nature, a victory which goes so far towards making man the master of things in this material plane!

Those who argue like that know little of modern science and its votaries. An Anglican bishop never ignored a dissenting preacher with more dignified grace than the professor of orthodox science ignores the heterodox genius who has the audacity to wander beyond the limitations which received opinion has placed upon the possibilities of nature. The fact is that men of science have persistently ignored, and know absolutely nothing about, the great department of nature into which Keely penetrated years ago, and in which he has now made himself at home. Not long ago a Fellow of the Royal Society of Edinburgh, Major Ricarde-Seaver, went to Philadelphia to convince himself as to the nature of Keely's discovery. He returned, saying that Keely was working with, and had the apparent command over forces, the nature, or even the very existence, of which was absolutely unknown to him, and so far as he is aware, to modern science.

Beyond disintegration lies dispersion, and Keely can just as easily dissolve the atoms of matter as disintegrate its molecules. Dissolve them into what? Well,---into ether, apparently; into the hypothetical substratum which modern scientists have postulated, and about whose nature they know absolutely nothing but what they invent themselves, but which to Keely is not a hypothesis, but a fact as real as his own shoes; and which ether, indeed, seems to be the protoplasm of all things. As to the law of gravity, it appears very like a delusion, in the light of Keely's experiments, or, at least, but one manifestation of a law of very much wider application---a law which provides for the reversion of the process of attraction in the shape of a process of repulsion. One of Keely's little scientific experiments is to put a small wire round an iron cylinder that weighs several hundred weight, and when the force runs through the wire, to lift the cylinder up on one finger and carry it as easily as if it were a piece of cork. Not long ago he moved, single-handed, a 500 horse-power vibratory engine from one part of his shop to another. There was not a scratch on the floor, and astounded engineers declared that they could not have moved it without a derrick, to bring which in operation would have required the removal of the roof of the shop. Of course it is but a step in advance of this to construct a machine which, when polarized with a negative attraction, will rise from the earth and move under the influence of an etheric current at the rate of 500 miles an hour, in any given direction. This is, in fact, Keely's air ship.

Lately, he has applied his force to optics, and by means of three wires placed across the lens of a microscope he makes its magnifying power equal to that of the great telescope in the Lick observatory - the largest in the world. Why don't all astronomers and opticians run to look through Keely's microscope, and to examine into the process? Perhaps if Galileo were alive he might express an opinion!

But, the reader may naturally exclaim, how long has this been going on, and we to know nothing about it? Mr. Keely is now over 60 years of age, and he has worked since he was a boy, at times, upon various inventions before his discovery of ether. For the last 18 years he has been constantly employed with experiments upon the ether; for eighteen long years he has worked day and night, with hand and brain, in the face of discouragements that would long ago have killed the owner of a less heroic soul; and he has worked almost single handed. Slander, ridicule, open accusations of fraud, charlatanry, insanity---everything evil that it could enter the head of the knave of the heart of the fool to conceive, every mean insinuation, every malicious lie that prejudice, bigotry, ignorance, self-conceit, vested interests, greed, injustice, dishonesty, and hypocrisy could concoct---these have been the encouragement which, so far, the world has bestowed upon the discoverer of the profoundest truths and laws of nature that have ever been imparted to the profane, or even hinted at, outside of the circle of Initiates. And now it has been proved in a hundred ways, and before thousands of persons competent to judge of the merits of his machines, that he has really discovered previously unknown forces in nature, studied them, mastered some of their laws, invented, and almost perfected, apparatus and machinery that will make his discoveries of practical application in a hundred ways---now that he has actually done all this, how does the world treat him? Does Congress come forward with a grant to enable him to complete his marvelous work? Do men of science hail him as a great discoverer, or hold out the hand of fellowship? Do the people do honour to the man whose sole entreaty to them is to receive at his hands a gift a thousand times more precious to them than steam engine or telegraph? It is a literal fact that the world to-day would tear Keely to pieces if it had the power to do so, and if he fell exhausted in the terrible struggle he has so long maintained, his failure to establish his claims would be received with a shout of malignant delight from nearly every lecture hall, pulpit, counting-house, and newspaper office in the so-called civilized world! The world has hardly ever recognized its benefactors, until it has become time to raise a statue to their memory; 'in order to beautify the town.' Jealousy, stupidity, the malignity which is born of conscious inferiority, are at this moment putting in Keely's road every impediment which law and injustice can manufacture. Two hundred years ago he would have been burned, a century since he would have been probably mobbed to death, but thank God we are too civilized, too humane to burn or mob to death those who make great discoveries, who wish to benefit their fellow men, or whose ideas are in advance of their age - we only break their hearts with slander, ridicule, and neglect, and when that fails to drive them to suicide, we bring to bear upon them the ponderous pressure of the law, and heap upon them the peine forte et dure of injunctions, and orders, and suits, to crush them out of a world they have had the impertinence to try to improve and the folly to imagine they could save from suffering without paying in their own persons the inevitable penalty of crucifixion. Had it not been for the obligations incurred by Mr. Keely, writes Mrs. Bloomfield-Moore in the Philadelphia Inquirer, of Jan. 20th of this year, in accepting the aid of the Keely Motor Company - in other words, had scientists, instead of speculators, furnished him with the means necessary to carry on his 'work of Evolution,' the secrets which he has so carefully guarded would now have been public property, so little does he care personally for financial results. As it is, those who have witnessed his beautiful experiments in acoustics and sympathetic vibration were often too ignorant to comprehend their meaning, and, consequently, even after expressing gratification to him, went away from his workshop to denounce him as a Cagliostro, while others, competent to judge, have refused to witness the production of the ether, as Sir William Thomson and Lord Raleigh refused when they were in America a few years since. The company here mentioned has been a thorn in the inventor's side ever since it was organized. It has been bulled and beared by greedy speculators, in whose varying interests the American newspapers for years have been worked, the results of which the inventor has had to bear. For many years the Company has contributed nothing towards Mr. Keely's expenses or support, and in the opinion of many lawyers it is virtually dead. How far it is entitled to his gratitude may be gathered from the fact, as stated in Mrs. Bloomfield-Moore's article above quoted, that when Mr. Keely abandoned his old generator of etheric force, baffled in his attempts to wrest from nature one of her most carefully guarded secrets, harassed by his connection with the Keely Motor Company, some of the officers and stockholders of which had instituted law proceedings against him, which threatened him with the indignity of imprisonment, he destroyed many of his marvelous models, and determined that, if taken to prison, it should be his dead body and not himself.

When the history of his discoveries and his inventions come to be written there will be no more pathetic story in the annuals of genius than that of John Worrell Keely. The world hereafter will find it hard to believe that in the last quarter of the 19th century a man with an insight into the secret workings of nature, and a knowledge of her subtler forces, which, whenever it is utilized, will relieve mankind from much of the grinding toil that now makes bitter the existence of the vast majority of mortals, that such a man should have been left to starve, because in all the ranks of Science there was not found one man capable of understanding his colossal work - because in all the ranks of religion there was not found one man able to realize the enlarged conception of Deity immanent in Keely's great thoughts - because in all the ranks of commerce, of speculation, of literature, of art, there was not found one man large enough, generous enough, unselfish enough, to furnish money for a purpose that did not promise an immediate dividend.

It is to a woman, not a man, that the eternal honour is due of having come to Keely's rescue, and saved humanity from once more disgracing itself by doing genius to death with broken-hearted want and neglect. That woman's name will go down the centuries inseparably connected with Keely's discoveries. Probably no more romantic incident ever happened in the history of invention than the connection between this wealthy and large-minded woman and this slandered and persecuted genius, and no stranger one than the way in which she was led, by a series of most unfortuitous events, to offer her aid. From that day this lady has been not only his benefactor, but also his co-worker, trusted friend, and courageous defender. With the exception of his friend, those who have occupied themselves with Keely's discoveries have confined their attention to its commercial value. This was to be expected, for Science now is the hand-maid of trade, and Religion has become the fawning follower of Science. There is, however, a higher aspect to Keely's discoveries, and that their value as contributions to man's knowledge of Nature and natural laws. So far as that is concerned, Keely's success is an accomplished fact. His work, explaining his whole system, is now in the Press, and were he to die tomorrow he will be just as great a figure in the world's history as he would be were a thousand speculators to clear ten million dollars apiece by his inventions. Fancy honouring Copernicus or Galileo because the yelping jackals of speculation, who were their contemporaries, grew fat by feeding on their brains!

Whether Keely's inventions will be commercial success at present is another matter. The force, or, rather forces, which Keely handles, are the same as those known under other names in Occultism, and it is the belief of Occultists that these forces cannot be introduced into the practical life of men, or fully understood by the uninitiated, until the world is fit to receive them with benefit to itself - until the balance of the good and the evil they work is decidedly on the side of the good. Keely himself is persuaded that the world will derive almost unmixed benefit from his discoveries; but an Occultist would prefer to say that inventions and discoveries are disclosed to man, rather than to credit genius with the elaboration of ideas - disclosed, that is to say, through the brain of the ostensible inventor by one of the higher powers that guide the destinies of humanity. The discoveries of Keely have an occult side, which perhaps he himself may not fully perceive, but it is upon that side that it depends whether those discoveries themselves are fitted, by reason of sympathetic vibration of a still more inner ether than Keely has publicly spoken of, to harmonize with the mass chord of our present civilization, and manifest in the material life of man. Occultists believe that there are intelligent powers behind the visible things and events of life, which powers alone can say So far shalt thou go, and no further; but they do not believe that these powers act as a deus ex machina, for in themselves they are part of the natural order of things, and act in and through material and immaterial nature. We at present in our normal state of consciousness know these powers only as forces and laws, and when we become conscious of them as intelligent entities, we perceive at the same time that they themselves are governed by higher wills and intelligences, which act through them, as they act through us, and are to them their forces and their Laws.

*Occultists see in everything the (to us) eternal action of two opposing powers or principles, which are ever seeking equilibrium, and never find it, for behind them there is a definite tendency towards that which we call progress, which tendency gives the preponderance to one of these powers, and thus prevents the establishment of equilibrium, in other words of stagnation and death. Now all great discoveries are manifestations of one of these powers or forces only, and, however good in themselves, tend to disturb the equilibrium of terrestrial life more than is required for the normal rate of universal progress; and therefore they produce a disproportion of parts, and the opposite power or force gathers strength to resist and check the exaggeration. Already, in the estimation of an ever-growing number of thinking men, the inventions and discoveries of the present century have proved themselves a curse rather than a blessing. The have raised the world's standard of comfort, and at the same time they have lowered the power of purchasing these very comforts, a desire for which they have generated. The advantages that accrue from steam and from machinery have not been distributed, but have become the property of a small minority. Year by year competition is becoming fiercer, and labor more arduous and continual, and men are growing more and more like living machines, and the helpless slaves of machinery and of institutions. An operative, in these days of steam power, has less liberty than a slave ever had, ex-

---

* Franck, in Die Kabballa, says: We learn, by the last three Sephiroth, that the Universal Providence, that the Supreme Artist, is also Absolute Force, the all-powerful Cause; and that, at the same time, this cause is the generative element of all that is. It is these last Sephiroth that constitute the natural world, or nature in its essence and in its Active principle, Natura naturans. This passage is quoted in Isis Unveiled (Vol.I,p.40), the authoress adding: "This Kabalistic conception is thus proved identical with that of the Hindu philosophy. Whoever reads Plato and his Dialogue Timaeus, will find these ideas as faithfully re-echoed by the great philosopher." They are all, in fact echoes of the archaic Secret Doctrine of the Occultists; and it is somewhat singular (pour le dire en passant) that the newspaper press so persistently refers to Theosophy as a new, or new-fangled, religion or philosophy, whereas the ideas now called theosophical are the oldest in the world, and may be found, more or less disguised or mutilated, and under many different forma and names, in all the great philosophical and religious sytems of antiguity. One is forced to conclude that the complaint, so often heard, that competition now compels young people to earn their bread while they ought still to be at school, applies with a good deal of force to a certain class of writers for the press.

cept in one particular - he has full liberty to starve, or to work himself to death, neither of which privileges an owner would allow him. Keely, however, thinks his discoveries will restore this disturbed equilibrium.

The direct effect of modern discoveries and inventions has been the rise of the commercial and economic system; and the inevitable consequence of that system has been to deepen the gulf between the poor and the rich. The natural effect of this is an antagonism between the two poles of society, which has its roots deep down in human nature and human passions, and this antagonism is becoming better recognized, and growing in intensity, year by year, in so much that it is almost universally felt that the only possible outcome from it is a social overturning, the date of the actual occurrence of which will depend chiefly upon the activity of the school-boards, and the thoroughness of their work. Hardly a thinking man of the present day but foresees, sooner or later, a great social cataclysm, in which all mere political and financial considerations will be as straws in a whirlwind. Now, it would seem that Keely's discoveries tend to develop power over material nature in the same direction in which that power has been growing during the last hundred years. If it be a power into the exercise of which there enter no moral considerations whatever, then it is applicable alike for good purposes and for evil; and it will be as ready to the hand of the bad man as to that of the good. Were such inventions given to the world in their completeness, the whole of the enormous power they gave over human life and destines would, it would seem, fall into the possession of the same small minority who at present control the power conferred by our present inventions and discoveries - the capitalists. If so, that section of the community would then, under our present institutions, obtain almost absolute power over the great majority - those who depend upon their labor for their support. The capitalists who owned the tremendous powers implied in a monopoly of Keely's inventions would be practically the absolute masters of the people; and obedience to their will would be far more really, than even now, the condition upon which those who were not capitalists also would be allowed the means of continued existence.

Occultists believe that the world is not yet ready for the appearance of such tremendous forces on the stage of human life. Mankind is too selfish, too cruel, too stupid, too pitiless, too animal, to be entrusted with what, in sober reality, are minor divine powers. Such powers could not at present be employed for the benefit of mankind and for the advancement of the race; on the contrary, they would tend to the further brutalization and virtual enslavement of the poor, and also to the further materialization and moral degradation of One the rich. In a word, the human qualities of justice, mercy, love, generosity, unselfishness, have not yet grown strong enough in the race, and the animal qualities of revenge, anger, jealousy, tyranny, hatred, selfishness, are still too powerful in man to make the acquisition of almost absolute power over nature, and over one another, anything but a curse to mankind. It would be less disastrous to give dynamite cartridges to monkeys for playthings.

For this reason Occultists, in general, do not regard Keely's discoveries as likely to succeed in the commercial sense. And at present things have certainly a look that is in accordance with that opinion. The powers that might be expected to intervene in order to prevent Keely's inventions from becoming factors in human life, are, as has been said, through human means, and the stolid stupidity of the scientists in regard to Keely's discoveries, the bovine indifference of theologians, the silly ridicule of the press, the hostility of vested interests, the suicidal greed of some of the largest shareholders, and the paralyzing influence of the law, which apparently lends itself in this case to those whose object is simple robbery. All these things seem very like the operation of the higher controlling powers, acting with a consciousness other than our consciousness for the attainment of ends that transcend our narrow calculations.

Be this as it may, Keely's discoveries, and Keely's personality also, have a peculiar interest for Theosophists, for the force with which he is working is without doubt the ether of the ancient philosophers, which is one aspect of the Akasa, the underlaying great force in nature, according to the Secret Doctrine, a force whose existence has been recognized from time immemorial under various fanciful names, and whose property is sound, whether audible or inaudible to us; or, in more modern language, whose characteristics are vibration and rhythm. It corresponds to the seven-fold Vach of Hindu Philosophy, and is the raison d'etre of spells and Mantrams. It is the basis of harmony and melody throughout Nature. This force is alluded to many times in Madame Blavatsky's Isis Unveiled. On page 139, vol I., we read: The Akasa is a Sanscrit word which means sky, but it also designates the imponderable and intangible life principles, the astral and celestial lights combined together, and which two form the anima mundi, and constitute the soul and spirit of man; the celestial light forming his nous, pneuma, or divine spirit, and the other his psyche, soul, or astral spirit. The grosser particles of the latter enter into the fabrication of his outward form, the body. The Akasa is connected on the one hand with physical matter and on the other with WILL, that intelligent, intangible, and powerful something which reigns supreme over all inert matter. Of the Akasa in this respect we read on page 144, vol.I., of Isis Unveiled: The mysterious effects of attraction and repulsion are the unconscious agents of that will; fascination, such as we see exercised by some animals, by serpents over birds, for instance, is a conscious action of it, and the result of thought. Sealing-wax, glass, and amber, when rubbed, *i.e.*, when the latent heat which exists in every substance is awakened, attract light bodies; they exercise unconsciously Will; for inorganic as well as organic matter, however infinitesimally small it may be, possesses a particle of the divine essence in itself... What is, then, this inexplicable power of attraction but an atomical portion of that essence that Scientists and Kabalists equally recognize as the 'principle of life' - the Akasa? Granted that the attraction exercised by such bodies may be blind; but as we ascend higher the scale of organic beings in nature, we find this principle of life developing attributes and faculties which become more determined and marked with every rung of the endless ladder. Man, the most perfect of organized beings on earth, in whom matter and spirit - i.e., Will - are the most developed and powerful, is alone allowed to give a conscious impulse to that principle which emanates from him, and only he can impart to the magnetic fluid opposite and various impulses without limit as to the direction. Isis Unveiled was published nearly eleven years ago; and in her forthcoming work, The Secret Doctrine, the authoress enters more fully into this and other matters only sketched or hinted at in her former volumes.

It is the fact that Keely is working with some of the mysterious forces included under the name Akasa that makes his discoveries interesting to Theosophists. It is the fact that he has shown magnificent courage and fixity of purpose under every kind of opposition, and the fact also that he has been supported all through by the generous belief that his discoveries will be of inestimable benefit to mankind that make his personality of interest. If he can succeed in making his marvelous discoveries pay dividends, science may begin to give attention to them; for men of science, like other men, require a sign before they can accept as truth the things that are beyond their comprehension, and the value of a scientific discovery is now determined by its market value.

R. Harte (Sec. T.P.S.)
July, 1888

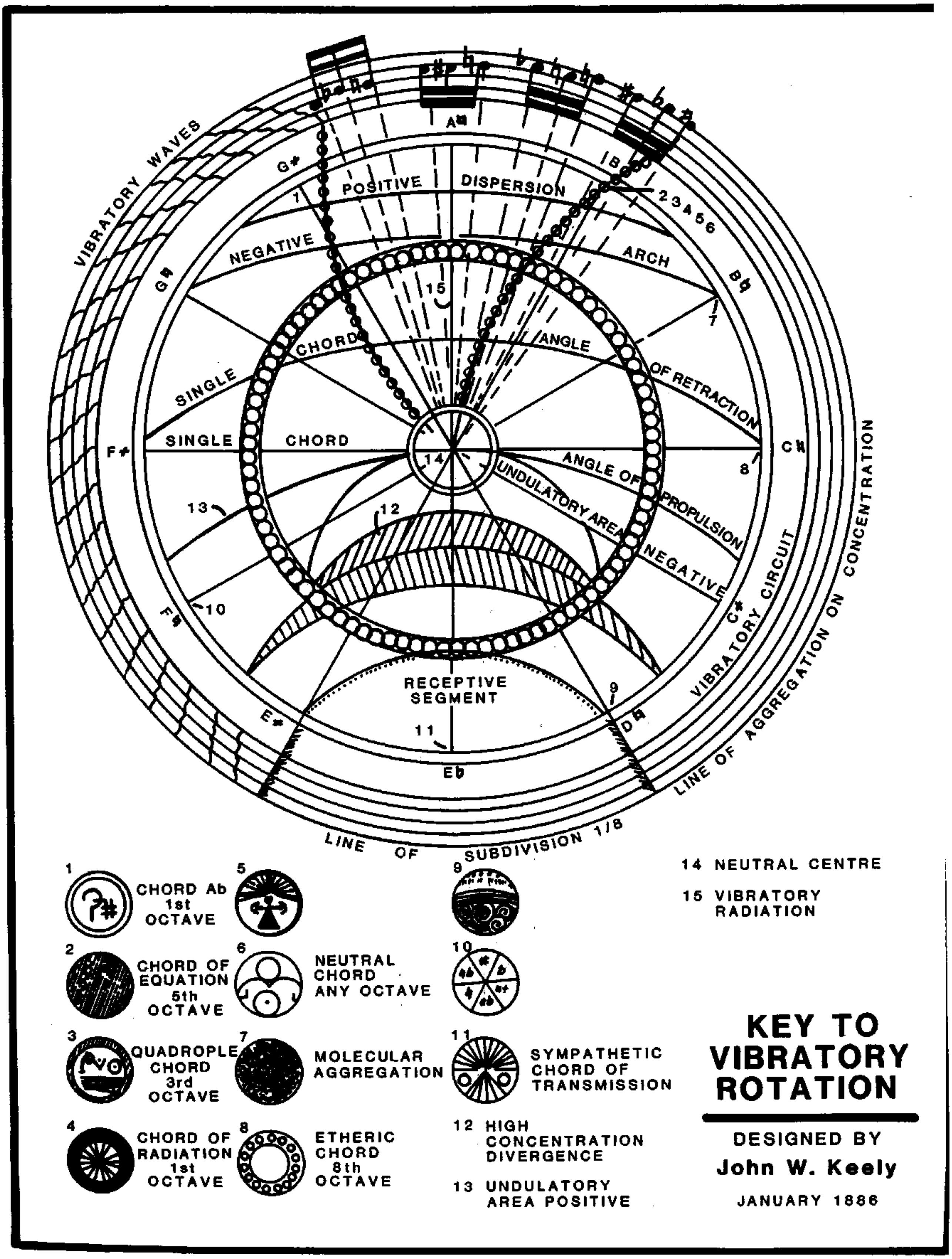
VIBRATORY WAVES
A♮
G♯
B
POSITIVE
DISPERSION
1
2 3 4 5 6
NEGATIVE
ARCH
G♮
B♭
15
CHORD
ANGLE
7
SINGLE
OF RETRACTION
SINGLE
CHORD
F♯
14
C♮
ANGLE OF
8
UNDULATORY AREA
PROPULSION
13
12
NEGATIVE
F♮
10
C♯
VIBRATORY CIRCUIT
LINE OF AGGREGATION ON CONCENTRATION
RECEPTIVE SEGMENT
9
E♯
D♮
11
E♭
LINE OF SUBDIVISION 1/8
1 CHORD A♭ 1st OCTAVE
2 CHORD OF EQUATION 5th OCTAVE
3 QUADROPLE CHORD 3rd OCTAVE
4 CHORD OF RADIATION 1st OCTAVE
5
6 NEUTRAL CHORD ANY OCTAVE
7 MOLECULAR AGGREGATION
8 ETHERIC CHORD 8th OCTAVE
9
10
11 SYMPATHETIC CHORD OF TRANSMISSION
12 HIGH CONCENTRATION DIVERGENCE
13 UNDULATORY AREA POSITIVE
14 NEUTRAL CENTRE
15 VIBRATORY RADIATION

KEY TO VIBRATORY ROTATION

DESIGNED BY
John W. Keely
JANUARY 1886

Etheric Force Identified as Dynaspheric Force

Clara Jessup Bloomfield-Moore

"Science is to know things."
Herodotus.

"Knowledge is developed by experience from innate ideas."
Plato

"Truth is not attained through reflection, but through immediate intuition. We neither originate thought nor its form."
Aryan Teachings

"It may be said that if all things come from only one cause or internal source, acting within itself, then motion and matter must be fundamentally and essentially one and the same, and we look upon matter as being latent force and upon force as being free matter."
Franz Hartmann, M.D.

JOHN WORRELL KEELY---the discoverer of compound inter-etheric force, as operating in the animal organism, man - is a great thinker, and a great student of the capabilities of nature in offering to man's intelligence the means whereby he may discover for himself the secrets she often veils without entirely concealing.

The result of more than twenty years of persistent effort to apply this force to the operation of machinery has, at last, been enabled to produce continuity of motion in his engine; but, up to this time, he has not so mastered this subtle force as to control reversions. The development of his various discoveries has been one uninterrupted work of evolution, reaching, within the last year, he thinks, the sphere of perfect vibratory sympathy, both theoretically and practically. The proof of this is found in the fact that he now transmits vibrations along a wire, connected at one end with the vibratory machine which is the source of power, and at the other end with the engine or cannon, as the case may be, which is operated by such vibratory power. Until recently, comparatively speaking, Keely stored force, as he generated it, in a receiver; and experiments were made by him in the presence of thousands, at various times, for the purpose of testing the operations of this force, liberated in the presence of his audience and stored up in this small receiver. The editor of the Scientific Arena thus describes what took place at one of these exhibitions, when he was present:---"The confined vapor was passed through one of the small flexible tubes to a steel cylinder on another table, in which a vertical piston was fitted so that its upper end bore against the underside of a powerful weighted lever. The superficial area of this piston was equal to one-half of a square inch, and it acted as a movable fulcrum placed close to the hinged end of the short arm of this lever, whose weight alone required a pressure of 1,500 pounds to the square inch against the piston to lift it.

After testing the pressure by several small weights, added to that of the lever itself, in order to determine how much power had already been accumulated in the receiver, the maximum test was made by placing an iron weight of 580 pounds, by means of a differential pulley, on the extreme end of the long arm of the lever. To lift this weight, without that of the lever supporting it, would require a pressure against the piston of 18,900 pounds to the square inch, counting the difference in the length of the two arms and the area of the piston, which we, as well as several others present, accurately calculated. When all was ready, and the crowded gathering had formed as well as possible to see the test, Keely turned the valve-wheel leading from the receiver to the flexible tube, and through it into the steel cylinder beneath the piston, and simultaneously with the motion of his hand the weighted lever shot up against its stop, a distance of several inches, as if the great mass of iron had been only a cork. Then, in order to assure ourselves of the full 25,000 pounds to the square inch claimed, we added most of our weight to the arm of the lever without forcing the piston back again.

After repeating this experiment till all expressed themselves satisfied, Keely diverted his etheric gas to the exciting work of firing a cannon, into which he placed a leaden bullet about an inch in diameter. He conveyed the force from the receiver by the same kind of flexible copper tube, attaching at one end of it to the breech of the gun. When all was again in readiness he gave a quick turn to the inlet valve, and a report like that of a small cannon followed, the ball passing through an inch board and flattening itself out to about three inches in diameter, showing the marvelous power and instantaneous action of this strange vapor.

The difficulty encountered by Keely in his old generator of etheric force grew out of the fact, in part, that the vaporic power produced was so humid that he could not, when he attempted to utilize it, obtain its theoretical value in work. This difficulty has been entirely overcome by dispensing with the water which he used in liberating etheric force, by his old generator; and, by this departure, he has attained a success beyond that which we anticipated by himself, when he abandoned his original line of experiment. Ignorant, indeed, of the nature of Keely's work must those men be who accuse him of "abandoning his base" or "principle," each time that he discovers his mistakes:---using them as stepping-stones to approach nearer and still nearer to his goal. Reproaching him, even, for keeping his own counsel, until certainty of success rendered it prudent for him to make known that he had changed his field of experiment from positive attraction to negative attraction, he had succeeded in his efforts to produce continuity of motion. Equally ignorant are those, who would wrench by force his secrets from him before the time is ripe for their disclosure. Let us suppose that Faraday, when he discovered radiant matter in 1816, had formed a "Faraday Phospho-Genetic Radiant Company," to enable him to experiment: fully cognizant of all that Crookes has since discovered, and had taken for his base in experimenting the principle involved in Crookes' discovery. Not succeeding at first, we will suppose that the Company became clamorous for returns, and demanded that his secret principle should be made public. Had he been driven into making known, who would have credited what Crookes is now able to prove? The effect would have been upon the Faraday Company the same as if a balloon were punctured just as it was soaring heavenward. The same with the Keely Motor Company, had Keely obeyed the order of the Court in 1882, and made his marvelous secret public. It would have collapsed. Therefore, he has maintained his secret in the interests of the stockholders of the Keely Motor Company with a firmness worthy of a Christian martyr. The one person to whom alone Keely then disclosed it thought him under a delusion, until he had demonstrated its soundness. When he said, in all solemnity, "Now, I feel as if you and I were the god and goddess of this world," the effect upon the hearer was no more than it would have been had a patient in a hospital for the insane spoken the words. Charles B. Collier, Keely's patent lawyer, writes as follows, concerning the difficulties attendant upon "the supposed duty" of his client's imparting his "secrets," as ordered by the Court to do, some time since:---

If to-day, for the first time in your lives, you saw a harp attuned and being played upon, you would hardly expect, without considerable time and study, to be able to reproduce the harp, attune its strings in proper relation to each other, and to play upon it so as to produce the harmonies which you had listened to. Mr. Keely's work is analogous to the illustration which I have presented, inasmuch as he is dealing with the subject of sound, or acoustics, but in a much more involved form than as applied simply for the production of harmonies for the delight of the ear. Mr. Keely's engine is analogous to the mechanism of the human ear, in the respect that it is a structure operated upon, and its notion induced by vibration; and to the end of securing and attaining, in and by it, uniformity or regularity of motion, there must be perfect unison or synchronism, as between it and his structure which is the prime source of vibration. To attain this perfect unison or synchronism, has involved unparalleled research and experiment upon his part---experiments that have varied from day to day. No one, in my opinion, who had not stood by his side, as his shadow, watching every experiment, could have kept fully abreast of him. To pursue my simile, I may say that his harp (engine) is not yet perfectly attuned ("graduated"); when it is so, it will produce nothing but harmony (regularity of motion), and his work will be finished.

At such time, I doubt not that he will be able to give to Mr. Boekel, myself or another, the scale with which to reconstruct and attune another apparatus so as to produce like results with it; but to go over the ground that he has gone over, to explore the wilderness in which he has been the pioneer, in other words, the study, to a full understanding of them, of his experiments and researches, as recorded in his writings and illustrated in the beautiful charts which he has produced, will be a work rather for scientists than for mechanicians or engineers

Mr. Keely's "Theoretical Expose'" is in preparation for the press; and, when these volumes are issued, we may look for a change of attitude towards him in all men who hold themselves ready to abandon preconceived notions, however cherished, if they be found to contradict truths; which, Herbert Spencer says, is the first condition of success in scientific research. The Rev. J. J. Smith, M.A., D.D., tells us that the only way the great problem of the universe can ever be scientifically solved is by studying, and arriving at just conclusions with regard to the true nature and character of force. This has been Mr. Keely's life study; and he is able to demonstrate all that he asserts.

The author of No. 5 of the pamphlets issued by The Theosophical Publication Society, "What is Matter and What is Force, " says therein, "The men of science have just found out 'a fourth state of matter,' whereas the occultists have penetrated years ago beyond the sixth, and therefore do not infer, but know of, the existence of the seventh, the last. "This knowledge comprises one of the secrets of Keely's so-called "compound secret." It is already known to many that his secret includes "the augmentation of energy, "the insulation of the ether," and the "adaptation of dynaspheric force to machinery."

Laurence Oliphant writes: Recent scientific research has proved conclusively that all force is atomic --- that electricity consists of files of particles, and that the interstellar spaces contain substances, whether it be called ether or astral fluid (or by any other name), which is composed of atoms, because it is not possible to dissever force from its transmitting medium. The whole universe, therefore, and all that it contains, consists of matter in motion, and is animated by a vital principle which we call God.

"Science has further discovered that these atoms are severally encompassed by an ethereal substance which prevents their touching each other, and to this circumambient, inter-atomic element they have given the name of dynasphere; but, inasmuch as has further been found, that in these dynaspheres there resides a tremendous potency, it is evident that they also must contain atoms, and that these atoms must in their turn be surrounded by dynaspheres, which again contain atoms, and so on *ad infinitum.* Matter thus becomes infinite and indestructible, and the force which pervades it persistent and everlasting.

"This dynaspheric force, which is

also called etheric, is conditioned as to its nature on the quality of the atoms which form its transmitting media; and which are infinite both in variety and in their combinations. They may, however, be broadly divided into two categories; *viz.* the sentient and the non-sentient atoms. Dynaspheric force, composed of non-sentient atoms, is the force that has been already mechanically applied by Mr. Keely to his motor; and which will probably ere long supersede the agencies now used for locomotive, projectile, and other purposes. When the laws which govern it come to be understood, it will produce materially a great commercial and industrial revolution. There is no hard and fast line between the animal and vegetable creation, so there is a graduated scale of atoms, between which although animated by the divine life are not sentient, and those which are as highly developed, relatively to them, as man is to a cabbage."

"The most remarkable illustration of the stupendous energy of atomic vibratory force is to be found in that singular apparatus in Philadelphia --- which for the last fifteen years has excited in turn the amazement, the skepticism, the admiration, and the ridicule of those who have examined it --- called 'Keely's Motor.'" ... "in the practical land of its origin, it has popularly been esteemed a fraud. I have not examined it personally, but I believe it to be based upon a sound principle of dynamics, and to be probably the first of a series of discoveries destined to revolutionize all existing mechanical theories, and many of the principles upon which they are founded. Those who are sufficiently unprejudiced to connect the bearings of this discovery, of what must be dynaspheric force, with phenomena which have hitherto been regarded as supernatural by the ignorant, will perceive how rapidly we are bridging over the chasm which has divided the seen from the unseen, obliterating the distinction between 'matter' and what has most erroneously been called 'spirit.'"

In 1882 a woman, coversing with Mr. Keely, said, "You have opened the door into the spirit world." He answered, "do you think so? I have sometimes though I might be able to discover the origin of life." At this time Mr. Keely had given no attention whatever to the occult bearing of his discovery; and it was only after he had pursued his researches, under the advantages which his small Liberator afforded him for such experiments, that he realized the truth of this women's assertion. It was then, in 1887, that a "bridge of mist" formed itself before him, connecting the laws which govern physical science with the laws which govern spiritual science, and year by year this bridge of mist has solidified, until now he is in a position to stand upon it, and proclaim that its abutments have a solid foundation --- one resting in the material and visible world, and the other in the spiritual and unseen world; or, rather, that no bridge is needed to connect the two worlds, one law governing both in its needed modifications.

"The physical thing," writes a modern scientist, "which energizes and does work in and upon ordinary matter, is a separate form of matter, infinitely refined, and infinitely rapid in its vibrations, and is thus able to penetrate through all ordinary matter, and to make everywhere a fountain of motion, no less real because unseen. It is among the atoms of the crystal and the molecules of living matter; and, whether producing locked effects or free, it is the same cosmic thing, matter in motion, which we conceive as material energy, and with difficulty think of as only a peculiar form of matter in motion."

Oliphant, commenting upon this view of energy, says: "This is nothing more or less than what we have been in the habit of calling 'spirit.'" ... "Mind is also composed of this extraordinary matter; so is will; so is every emotion." Jacob Boehme calls it "heavenly substantiality," and Swendenborg calls it "natural and spiritual atmospheres, composed of discrete substances of a very minute form." Professor Crookes has invented the word "protyle." Professor Cones calls it "soul-stuff," or biogen; while Occultists call it "astral fluid."

Etheric Force Indentified as Dynaspheric Force

To all who are conversant with Mr. Keely's theories a similarity of views will be evident.

The President of the British Association, Sir Henry Roscoe, in his address before that body, said: "In nature there is no such thing as great or small; the structure of the smallest particle, invisible even to our most searching vision, may be as complicated as that of any of the heavenly bodies which circle round our sun." As to the indivisibility of the atom, he asks this question: "Notwithstanding the properties of these elements have been studied, and are now known with a degree of precision formerly undreamt of, have the atoms of our present elements been made to yield?" He continues: "A negative answer must undoubtedly be given, for even the highest of terrestrial temperatures, that of the electric spark, has failed to shake any one of these atoms in two."

This is an error, for it is well known by those who are fully acquainted with the principle involved in Keely's inventions that the intense vibratory action which is induced in his "Liberator" has accomplished what the retort of the chemist has failed to do, what the electric spark has left intact, and what the inconceivably fierce temperature of the sun and the volcanic fires has turned over to us unscathed. The mighty Genii imprisoned within the molecule, thus released from the chains and fetters which Nature forged, has been for years the the tyrant of the one who rashly intruded, without first paving the way with the gold which he has since been accused of using in experiments with reckless and wanton waste! For more than a score of years has Keely been fighting a hand-to-hand fight with this Genii; often beaten back by it, paralyzed at times, even, by its monstrous blows; and only now so approaching its subjugations as to make it safe to harness it for the work that is calling for a power mightier than steam, safer and more uniform in operation than electricity; a power which, by its might and beneficence, will ameliorate the condition of the masses, and reconcile and solve all that now menaces our race; as it was never menaced before, as has been said.

The structure of the air molecule, as believed in by Keely, is as follows: --- Broken up, by vibratory action, he finds it to contain what he calls an atomic triplet. The position of a molecule, on the point of a fine cambric needle sustains the same relation to the point of the needle that a grain of sand sustains to a field of ten acres.

Although, as Sir H. Roscoe has said, "In nature there is no such thing as great or small," the human mind cannot conceive such infinitesimal minuteness. We will, then, imagine a molecule magnified to the size of a billiard ball, and the atomic triplet magnified to the size of three marbles, in the triangular position, within that molecule, at its center; unless acted upon by electricity, when the molecule, the billiard ball, becomes oblate, and the three atoms are ranged in a line within, unless broken up by the mighty force of vibratory action. Nature never gives us a vacuum; consequently, the space within the molecule not occupied by the atomic triplet must be filled with something. This is where the Genii- "the all-prevading ether"---has made its secret abode through untold eons, during which our world has been in course of preparation for its release, to fulfill its appointed task in advancing the

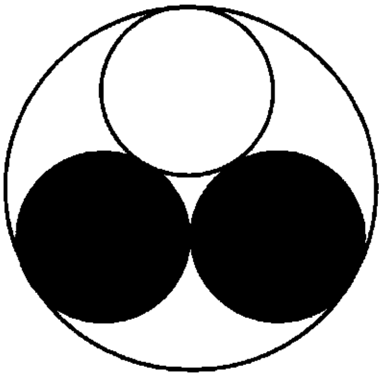

progress of the human race.

Step by step, with patient perseverance which some day the world will honour, this man of genius has made his researches, overcoming the colossal difficulties which have raised up in his path what seemed to be insurmountable barriers to further progress: but never before has the world's index finger so pointed to an hour when all is making ready for the advent of the new form of force that mankind is waiting for. Nature, always reluctant to yield her secrets, is listening to the demands made upon her by her master, necessity. The coal mines of the world cannot long afford the increasing drain made upon them. Steam has reached its utmost limits of power, and does not fulfill the requirements of the age. Electricity holds back, with bated breath, dependent upon the approach of her sister colleague. Air ships are riding at anchor, as it were, waiting for the force which is to make aerial navigation something more than a dream. As easily as men communicate with their offices from their homes by means of the telephone, so will the inhabitants of separate continents talk across the ocean. Imagination is palsied when seeking to foresee the grand results of this marvelous discovery when once it is applied to art and mechanics. In taking the throne which it will force steam to abdicate, dynaspheric force will rule the world with a power so mighty in the interests of civilization, that no finite mind can conjecture the results. Laurence Oliphant, in his preface to "Scientific Religion," says: "A new moral future is dawning upon the human race - one, certainly, of which it stands much in need." In no way could this new moral future be so widely, so universally, commenced as by the utilizing of dynaspheric force to beneficial purposes in life, thus revealing to all men another phase of God's "underlying purpose."

In 1746, when Franklin's attention was drawn to the phenomena of electricity, little more was known on the subject than Thales had announced two thousand years before. Von Kleist in Leyden, Collinson in London, and others in as widely-separated cities in Europe, were experimenting in the same field of research. What our last century has done toward subduing this tyrant which Franklin succeeded in bringing down to earth, from the clouds, the next century will see surpassed beyond man's widest conjectures, should Keely's utilization of this unknown force of nature bestow upon humanity the costless motive power, which he anticipates it will. Reynolds predicted that those who "studied the mysteries of molecular vibration would win the victorious wreaths of successful discovery." After such discoveries as Mr. Keely has made in this field of research, it matters not to him whether he succeeds commercially or not. His work of discovery commenced when, as a boy of twelve, he held the seashells to his ear as he walked the shore and noted that no two gave forth the same tone. From the construction of his first crude instrument, his work of evolution progressed slowly for years; but within the last five years he has made giant strides towards the "Dark Tower" which is his last fortress to take. When he is ready, "Dauntless the slug-horn to his lips" he will set; and the world will hear the blast, and awaken from its slumber into new life.

Molecular vibration is thus seen to be Keely's legitimate field of research; but more than once has he had to tear down portions of the vibratory scaffolding which aided him in the building up of his edifice of philosophy; therefore, he is ever ready to admit that some of the present scaffolding may have to be removed. The charge of "abandoning his base," recently brought against him by one of the editors of The New York Times, could only have been made by one who is utterly ignorant of the subject upon which he writes. Under the heading "A Cool Confession," this editor asserts that Keely has "given up the Keely Motor as a bad job," and that he admits that he is a "bogus inventor" and a "fraud." This is not true.

What Keely does admit is that, baffled in applying vibratory force to mechanics, upon his first and second lines of experimental research, he was obliged either to confess a commercial failure, or to try a third departure from his base or principle; seeking success through another channel of experiment. While experimenting upon this third line, until his efforts were crowned with success, he kept his secret from all men; with the approbation of the one who furnished the money for these experiments. There is a time when silence is golden; and the charge made by the same editor that Keely had been "receiving money from the Keely Motor Company on false pretences from the time that he abandoned his original plans," could only have been made by one who knows nothing of the facts of the case: for years have passed away since the Keely Motor Company broke its contract with him, and since it has furnished him with any money for his experiments.

But let Keely speak for himself in reference to his work: --- *"In considering the operation of my engine, the visitor, in order to have even an approximate conception of its modus operandi, must discard all thought of engines that are operated upon the principle of pressure and exhaustion, by the expansion of steam or other analogous gas which impinges upon an abutment, such as the piston of a steam-engine. My engine has neither piston nor eccentrics, nor is there one grain of pressure exerted in the engine, whatever may be the size or capacity of it.*

"My system, in every part and detail, both in the developing of this power and in every branch of its utilization, is based and founded on sympathetic vibration. In no other way would it be possible to awaken or develop this force, and equally impossible would it be to operate my engine upon any other principle.

"All that remains to be done is to secure a uniform speed under different velocities and control reversions. That I shall accomplish this is absolutely certain. Some few years ago, I contemplated using a wire as a connective link between two sympathetic mediums, to evolve this power as also to operate my machinery --- instead of tubular connections as heretofore employed --- I have recently succeeded in accomplishing successfully such change. This, however, is the true system; and henceforth all my operations will be conducted in this manner---that is to say, the power will be generated, my engines run, my cannon operated, through a wire.

"It has been only after years of incessant labour, and the making of almost enumerable experiments, involving not only the construction of a great many most peculiar mechanical structures, and the closest investigation and study of the phenomenal properties of the substance "ether," per se, produced, that I have been able to dispense with complicated mechanism, and to obtain, as I claim, mastery over the subtle and strange force with which I am dealing.

"When my present process of adjustment is completed, the force, the mechanism, and all that pertains to it, will be explained in a theoretical exposition of the subject, with appropriate diagrams, which I shall publish to the world; through the medium, and my patents, when taken out, a knowledge of all that is required for its commercial employment will be more easily acquired than is the necessary skill required to enable one to safely operate a steam-engine.

"My power will be adapted to engines of all sizes and capacities, as well to an engine capable of propelling the largest ship as to one that will operate a sewing machine. Equally well and certain is it that it will be adapted as a projectile force for guns and cannons of all sizes, from the ordinary shoulder-piece to the heaviest artillery."....

When Keely in 1887 obtained continuity of motion (for a time) in his engine he thought that his last difficulty had been overcome: but, up to the present time, he has not succeeded in governing its speed nor in controlling reversions. He has, however, again reduced in size the instrument with which he produces force. From 1882 to 1884 the "Generator" was a structure six feet long and correspondingly wide and high; but, failing in his attempt to make an automatic arrangement upon which its usefulness in mechanics depended, Keely found a new standard for research in an experiment often made by himself, but never before successful, which resulted in the production of a machine in 1885 which he named a "Liberator" --- not so large as a lady's small round work-table. Continuing his labor of evolution Keely within one year made such astonishing progress, from experiments with this beautiful piece of vibratory mechanism, as to combine the production of the power, and the operation of the cannon, his engine and his disintegrator in a machine no larger than a dinner plate, and only three or four inches in thickness. This instrument was completed in 1886, up to which time his experiments had been conducted upon a principle of sympathetic vibration, for the purpose of liberating a vapour or etheric product. His later experiments have been confined to another modification of the vibratory sympathy; and the size of the instrument used now, '88, for the same purposes is no larger than an old-fashioned silver watch, such as we see in Museum collections. The raising of a lever with an apparent uplifting expansive force of between 20,000 and 30,000 pounds to the square inch, the running of the engine, the firing of the cannon, are conducted without one ounce of pressure in any part of the apparatus, and without the production or presence of what has been known as Keely's ether. The force is now transmitted along a wire (of platinum and silver), and when the lever is lowered there is no exhaustion, into the atmosphere of the room, of any up-lifting vapor, as was always the case when ether was used in this experiment; nor is there any vapor impinging upon the piston under the lever to raise it.

Mr. Keely has named this new [illegible] force in nature [illegible] been experimenting with, and the phenomena attending them, are the very antithesis of each other. Keely does not feel the shadow of a doubt as to his eventual success in producing engines of varying capacities; small enough, on the one hand, to operate sewing machines with, and large enough, on the other hand, to propel the largest ships that plough the seas. Every fact and feature surrounding the case warrants the belief, notwithstanding the incredulity of all who have not witnessed the progress of Mr. Keely, step by step, that his success will be complete, and his work stand as the most colossal example of the survival of the fittest, in the process of inventive evolution. Cox says: "not one of the great facts which science now accepts as incontrovertible truths but was vehemently denied by the scientists of its time:---declared to be *a priori* impossible, its discoverers and supporters denounced as fools and charlatans, and even investigation of it refused as being a waste of time and thought." "History repeats itself," and Amiel's definition of science gives the key to the incredulity of scientists in reference to Mr. Keely's discovery; for if, as Amiel has said, "science is a lucid madness occupied with tabulating its own hallucinations," it is not strange that men of science should refuse to investigate what they consider the hallucinations of others.

It is an undisputed fact that "too much has been conceded to science, too little to those sublime laws which make science possible." But the one law which regulates creation, and to which all other laws are made subservient, keeping in harmony the systems upon systems of worlds throughout space, developing sound and colour, animal and vegetable growth, the crystallization of minerals, is the hidden law, which develops every natural science throughout the universe; and which both Kepler and Newton anticipated would be revealed in our age. "You can even trace the poles in sound," writes Mrs. F. J. Hughes, a niece of Darwin, in her work upon the "Evolution of Tones and Colours." The experiments made by Mrs. Watts Hughes, at the annual Reception of the Royal Society, and the Pendulograph* writings by Andrew of Belfast, have a bearing upon Keely's discovery; illustrating the workings of this hidden law of nature.

---

* A system of pendulums tuned to swing the various ratios of the musical scale, form a "Silent Harp" of extraordinary interest. This "Silent Harp," D.C.Ramsay, of Glasgow, has shown to his students of harmony for many a year. A pen, placed by means of a universal-jointed arrangement between any two pendulums of this "Silent Harp," so as to be moved by a blend of their various mo-[illegible] writes, with all the precision of gravitation, a [illegible] which two corresponding [illegible]

Of all women Mrs. Hughes approaches nearest to the theories of Mr. Keely. Concerning them she writes to a friend well versed in music, as music is taught: - "From ignorance of the present Science of Musicians, which you know so thoroughly, it is easier for me to grasp his meaning, than for you. I have lately been proving by scriptural types how Nature's laws in the lower creation develop by fifths below and by fourths above the key-note, the two meeting in one harmony: art mingling both, creating discords, and undulating them into harmony. Dr. C. Martin says: - 'Musicians must have discords; the ear is educated to them: but every one must allow that the nearer Art follows nature, the more perfect it is.' Mrs. Hughes adds: "I think Keely must have caught the center where both laws unite, or act upon Nature's law only."

Of the law of periodicity, Hartmann writes: *"Its actions have long ago been known to exist in the vibrations producing light and sound, and it has been recognized in chemistry by experiments tending to prove that all so-called simple elements are only various states of vibration of one primordial element, manifesting itself in seven principal modes of action, each of which may be subdivided into seven again. The difference which exists between so-called single substances appears, therefore, to be no difference of substance or matter, but only a difference of the function of matter in the ratio of its atomic vibration."* It is by changing the vibrations of cosmic ether that Mr. Keely releases this energy, and Dr. Kellner in Austria produces electricity in the same way; while it is said that a chemist in Prague produces magnetism; also Dr. Dupuy, of New York, who has been for years experimenting in this field without meeting with Keely's progressive successes.

Horace Wemyss Smith, in commenting upon the fact that, at the time of Franklin's discovery, men in France, in Belgium, in Holland, and in Germany were pursuing the same line of experiment, says that there is something worthy of observation in the progress of science and human genius, inasmuch as in countries far distant from each other men have fallen into the same tracks, and have made similar and corresponding discoveries, at the same period of time, without the least communications with each other.

Laurence Oliphant's recent works give us the clue to an explanation of this fact; and Lowe, in his Fragments of Physiology, condenses the answer in these words: "Man is not the governor and commander of the created world; and were it not for superhuman influence constantly flowing into created forms, the world would perish in a moment. It is this superhuman influence, felt most by those who have educated the hidden sense (with which all men are born), which inspires all discovery, all invention, all poetry, all of truth, let it take whatever form it may. This sixth sense is as much undeveloped in the mass of mankind as sight would be had we been born with our eye-lid sealed; able to distinguish nothing beyond the period of day-light from the reign of night, and remaining sealed all the years of our life upon earth. We know that some spiritually minded persons seem to possess powers unknown to those who are spoken of in Scripture as the carnal minded ; and it may be that with dim vision they are able to discern as in a glass darkly, without education of this hidden sense, truths which are hidden from others. Of such are men and women of genius. Again, there are others who possess uncertain, unreliable powers, which often lead astray those who commit themselves to the direction of these powers. What infallible medium of superhuman influence? Spiritualism represents a great truth, behind the spiritism which stands in the same relation toward it that counterfeit coin holds to sterling gold. The operations of our sixth sense are as liable to be deceptive as are the operations of our other senses; and are limited or governed by law in the same way. We cannot see in total darkness, and this hidden power, susceptible of education, can only be brought into use by an illuminated mind; a mind that has studies the laws of evolution and involution, the descent of spirit into matter, and the re-ascension of matter to spirit - laws of the life-impulse beginning in the elemental kingdom and ending in an evolution of man, far beyond the comprehension of man of the present day. Man and woman as units, continues Oliphant, are still so ignorant of the great powers which they themselves inherit that they wholly fail to see them, though they sweep like mighty seas throughout all human nature.

When mankind has become sufficiently spiritualized by the process of evolution laid down in the plan of the great Master, then shall we know ourselves and our powers as we are known to Him. True Science must first open wider the path of religion - a religion of progress, a religion suited to the wants of humanity, as well as a humane religion - the religion, taught by our Holy Master, of love for our fellow-men, of harmony with all that is good - at war only with evil; not with those who, warped by transmitted tendencies, commit evil. An eye for an eye, a tooth for a tooth, is the old Jewish law. Christ's law is the law of love, which is God's law. Do unto others as ye would have others do unto you ; and this is the law which we need to fulfil, in order to purify and regenerate mankind.

Hitherto we have been, in one respect, like the laborer in Tolstoi's Confession, doing the work assigned him, in the space assigned, without understanding where he was, or what the result would be, and unable to judge whether the arrangements for this work were reasonably planned by his master or not. The laborer worked the handle of a pump, he saw the water flow into numerous channels for irrigation of the soil; little by little, shrubs grew and blossomed and bore fruit,, and the laborer passed on to more important work, understanding better and better the arrangement of the establishment, and never doubting that its Lord had planned all for the best. Our race has been pumping water for generations, planting the seed, watching the growth of the plants, the shrubs, and the trees; not always satisfied, however, that the Creator of the universe had planned as well as all might have been planned, or that there was any plan, or any Creator.

There are men in various parts of the world, unknown even by name to each other, who tell us by the sign of the times that the season of harvesting is approaching; the season for gathering the fruit, which has been deferred, century after century, because mankind is yet not ready, in the opinion of many, to share the fruit with one another.

Hyndman says that capitalism has been as necessary as serfdom in the progress of the human race; and its stores will continue to be garnered and used to spread the great net-work of railways, steam navigation, telegraph and telephone lines, which have given bread to millions and millions; encouraging paid labor and bringing nearer and nearer the age of love and harmony, which, it has been predicted that the twentieth century will usher in.

Renan, the French theologian, writing of the advances to be made in science by the discovery of nature's secrets, said that, although he had ceased to look forward to anything very unexpected, he envied those who should live to see the wonders of the twentieth century. For, he added they who live then will see things of which we have no conception. Already the light of a new dawn is breaking upon the world of science. These foretold advances are heralded by the modifications of force, before alluded to. The instrument invented by Dr. Kellner collects and produces electricity directly from the ether of the atmosphere without any friction of solid corporeal substances and without any chemical agency; the invention of one of the Prague professors, which, it is said, collects and gives out magnetism, seeming to be derived from changing the vibrations of cosmic ether, as in Keely's and Kellner's experiments, and other more recent discoveries.

It has been said that when Keely's vibratory force shall have taken the place of steam-engines, the millions of working men who gain with difficulty their daily bread by the work of their hands, will find themselves without occupation. The same prediction was made in regard to steam, but instead we find the city of Boston giving work to thirty thousand men in one manufactory of boots and shoes by steam, in place of the three thousand shoemakers who were all that were occupied in this branch of labor in that city when the work was done by hand.

Dr. Kellner's colleague, Franz Hartmann, M.D., writing in reference to Keely's discovery, says: I have taken great interest in him ever since I first heard of him in 1882. As gaslight has driven away, in part, the smoky petroleum lamp, and is about to be displaced by electricity, which in the course of time may be supplanted by magnetism, and as the power of steam has caused muscular labor to disappear to a certain extent, and will itself give way before the new vibratory force of Keely, likewise the orthodox medical quackery that now prevails will be dethroned by the employment of the finer forces of nature, such as light, electricity, magnetism, *etc.*

When the time is ripe, these are of the true scientists who will come to the front to lead as Progress leads, men who know how to wait upon God, *viz.*, to work while waiting; and to such the end is, sooner or later, victory! God never hurries. He counts the centuries as we count the seconds, and the nearer we approach to the least comprehension of His underlaying purpose the more we become like Tolstoi's laborer, who knew that the fruit was ripening for him and his fellow-men, trusting implicitly in the superior wisdom of his master.

"Evermore brave feet, in all ages,
  Climb the heights that hide the coming day,
Evermore they cry, these seers and sages,
  From their cloud, "Our doctrines make no way."
All too high they stand above the nations,
  Shouting forth their trumpet-calls sublime,
Shouting downwards their interpretations
  Of the wondrous secrets born of time. "

No man, whose spiritual eyes have been opened to "discern the signs of the times," can doubt that we are on the eve of revelations which are to usher in the dawn of a brighter day than our race has yet known; and no prophecy of this brighter day, foretold by prophets, apostles, and inspired poets, was ever made in truer strains than in these glorious lines of Elizabeth Barrett Browning:---

"Verily many thinkers of this age,
  Aye, many Christian teachers,
    half in heaven,
Are wrong in just my sense who understood
  Our natural world too insularly, as if
No spiritual counterpart completed it,
  Consummating its meaning, rounding a
To justice and perfection, line by line,
  Form by form, nothing single nor alone;
The Great Below clenched by the Great
  Above."

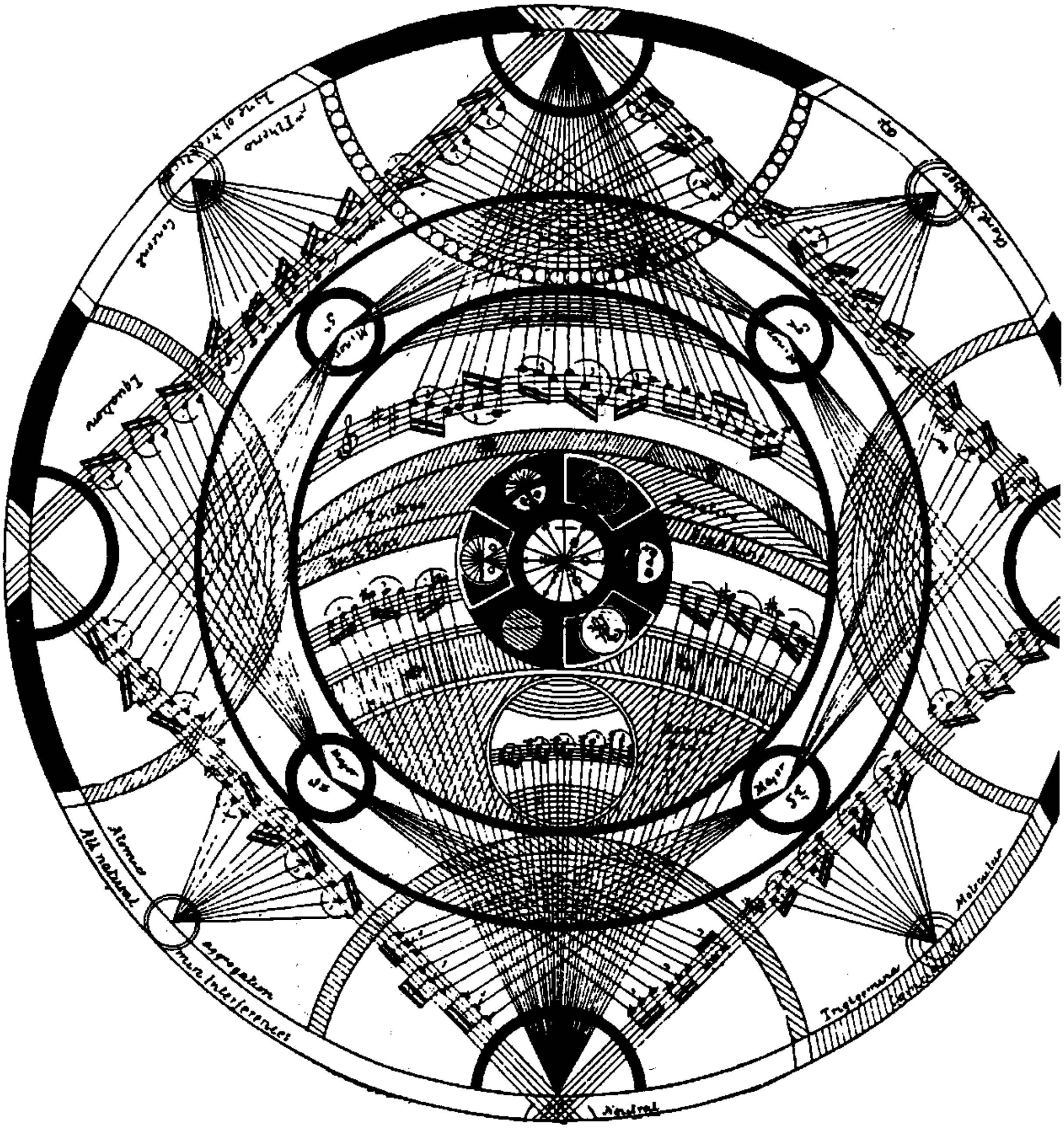

Atomic
All natural
aggregation
Interferences
Molecular
Major
Major
Neutral

MUSICAL CHART BY WHICH KEELY CLAIMED TO HAVE DISCOVERED HIS "ETHERIC" FORCE.

One Phase of Keely's Discovery in Its Relations to the Cure of Disease

Clara Jessup Bloomfield-Moore

"I know medicine is called a science. It is nothing like a science. It is a great humbug! Doctors are mere empirics when they are not charlatans. We are as ignorant as men can be. Who knows anything in the world about medicine? Gentlemen, you have done me the honour to come here to attend my lectures, and I must tell you now, frankly, in the beginning, that I know nothing about medicine, nor do I know anyone who does know anything about it. Nature does a great deal, imagination does a great deal, doctors do devilish little when they do not do harm. Sick people always feel they are neglected, unless they are well drugged, les imbeciles!"

Professor Magendie;
(before the students of his class in The Allopathic College of Paris)

In the year 1871, the writer was sent from Paris to Schwalbach, by Dr. Beylard, and recommended to the care of Dr. Adolph Genth. She said to the physician, I wish for your opinion and your advice, if you can give it to me without giving me any medicine. He replied, With all my heart, madam; and I wish to God there were more women like you, but we should soon lose most of our patients if we did not dose them.

This is a terrible excuse for the use of those agencies which Dr. John Good says have sent more human beings to their graves than war, pestilence and famine combined. Keely holds the opinion that Nature works under the law of Compensation and Equilibrium---the law of Harmony; and that when disease indicates the disturbance of this law Nature at once seeks to banish the disease by restoring equilibrium, He seeks to render assistance on the same plan; replacing grossly material agencies by the finer forces of nature; as has been so successfully done by Dr. Pancoast and Dr. Babbitt in America. It was the intention of Dr. Hartmann to establish a Theosophical Sanatorium at Goritz, in Austria, this summer, where medicine would be dispensed with: but his plans have been interfered with by his visit to America, where he went last March for the purpose of ascertaining Mr. Keely's views in reference to the best manner of conducting experimental research in reference to the restoration of equilibrium in the human system; the disturbance of which occasions local disorders and all disease, according to his ideas. Paracelsus held that *Man is made out of the four elements, and is nourished and sustained by magnetic power, which is the Universal Motor of Nature.* He treated disease in two ways - Sympathetically and Antipathetically; but only a fragmentary trace of his system can now be found. Nature, says Dr. Pancoast, author of The True Science of Light, works by antagonism in all her operations: when one of her forces overdoes its work, disease, or at least a local disorder, is the immediate consequence; now, if we attack this force, and overcome it, the opposite force has a clear field and may re-assert its rights---thus equilibrium is restored, and Equilibrium is health.

The Sympathetic System, instead of attacking the stronger force, sends recruits to the weaker one, and enables it to recover its powers; or, if the disorder be the result of excessive tension of Nerves or Ganglia, a negative remedy may be employed to reduce the tension. Thus, too, equilibrium is restored.

Dr. Hartmann disclosed to me in one of his letters that he knew the most important secret involved in Mr. Keely's compound secret. But he had not in any way connected this so-called secret with Mr. Keely. In one of Dr. Hartmann's letters to me, he writes -

Mr. Keely is perfectly right in saying

that all disease is a disturbance of the equilibrium between positive and negative forces. In my opinion, no doctor ever cured any disease. All he can possibly do is to establish conditions under which the patient (or nature) may cure himself. The universal power which Mr. Keely calls the ether, and, which Dr. Kellner calls transitory element, was known to medieval philosophers as prima materia, will and thought; or, according to Schopenhauer, will and imagination are substance. I recognize only one universal and fundamental power, which I call consciousness, acting within matter by means of thought; and I have no doubt that you already know that we agree all around, although we may not all use the same terms to signify the same objects. In your most important papers, I have found my own sentiments and views reflected; and I have in my books on Magic, Paracelsus, and the Rosicrucians, attempted to explain these identical views. Why will our scientists insist on refusing to see the self-evident fact that all visible material substances, animal organs, etc., are nothing else than the ultimate products of pre-existing psychic (interior and invisible) forces? These facts were all known to the ancient philosophers; while the moderns insist on mistaking the effects for the cause. They reject the idea of God (the primordial cause of all in its highest aspect of spiritual consciousness) because they formed a misconception of that which is intellectually inconceivable; they found that God could not be that which they had imagined, and they logically(?) concluded that there could be no Divine power at all. But this subject is too grand, too sublime, and extensive to be more than alluded to in this letter, and I merely write these remarks to show you that your views, those of Mr. Keely, and my own are all identical, as they, indeed, must be with those who are capable to perceive self-evident truth; for the truth is only one, and all who know it possess that same identical knowledge. Mr. Keely's power seems to be derived by changing the vibrations of cosmic ether. The machine which my friend Dr. Kellner has invented seems to be based upon the same principle, only, while Mr. Keely transforms these vibrations into some force connected with sound, Dr. Kellner's machine transforms them into elctricity. Again, Dr. Hartmann writes: Even to the superficial observer, the fact that the world is becoming more and more spiritualized, from top to bottom, begins to be evident. The crude scientific opinions which were prevailing in the beginning of the century are disappearing before a higher knowledge in regard to the laws of nature; the materialism which flourished twenty years ago, the offspring of animalism and ignorance, has almost disappeared from view, and has to descend to the lowest strata of society to find admirers. The iron rod, with which a self-conceited and arrogant sacerdotalism ruled the people, has been broken, and its remnants exist only in those countries where priestcraft is upheld and abetted by kings and governments.

If you enter the field of therapeutics and medicine, we find a decided fermentation of new ideas; not among the fossil specimens of antediluvian quackery, but among those who are called irregulars, because they have the courage to depart from the tracks trodden out by their predecessors. The more intelligent class of physicians have long ago realized the fact that drugs and medicines are perfectly useless, excepting in cases where diseases can be traced to some mechanical obstruction, in some organ that may be reached by mechanical action. In all other cases our best physicians have become agnostics, leaving nature to have her own way, and observing the expectative method, which, in fact, is no method of cure at all, but merely consists in doing no harm to the patient. Recently, however, light, electricity, and magnetism have been employed; so that even in the medical guild the finer forces of nature are taking the place of grossly material, and therefore injurious substances. The time is probably near when these finer forces will be employed universally. Everywhere the leaven is working, and many are asking, 'What causes it to work?' The answer is 'It is spirit working in matter. But the term 'spirit' is to the majority of mankind a term without any meaning, a nonentity. Nevertheless, the action of that power which is called cohesion, and which is equally invisible, but which really holds the atoms of all bodies together and prevents them from dissolving into tangible ether, is continually manifested before their eyes. Why should not the opposite form of activity, that which enters between the atoms and separates them, likewise be a reality? The scientist will answer, 'We know this activity, and we call it heat. What has heat to do with spirit?' It has been demonstrated long ago that heat is a mode of motion, and likewise every other form of energy (including spiritual activity) is nothing else but a mode of motion. Motion is that universal agent which is fundamentally and essentially only one, but whose mode of manifestation differs according to the conditions under which it manifests itself. Acting without relative consciousness, it is known as gravitation, attraction, heat, light, sound, electricity, magnetism, etc. In a higher state it is known as life, and becomes endowed with relative consciousness, acting in the highest plane of existence it becomes self-conscious and self-existent, and is called spiritual power. But there is no motion thinkable without a substance to move; we cannot imagine a force without matter, nor matter without energy. There must, therefore, be one original substance, or primordial matter, although of a kind very different from the form in which it appears to us on the externally visible plane. The existence of this primordial substance was known to the spiritual perception of the ancient Rosicrucians, and some of them have, by logical conclusions, arrived at a belief in its existence, and named it Cosmic Ether; while by the Eastern sages it was called Akasa. We therefore see that there is one primordial and universal power , which is Motion; and, likewise, one primordial and universal substance which we may call Ether, or Matter; and that all existing forms can be nothing else but various shapes of that Ether in various states of density, and existing under various conditions, while all forms of energy, from the most grossly material up to the highest spiritual seem to be merely modifications of motion in Ether, manifesting themselves in various conditions and under various circumstances, unconsciously, consciously, and with self-consciousness. Furthermore, it may be said that if there is only one God, that is to say, if all things come from only one cause of internal force acting within itself, then motion and matter must be fundamentally and essentially one and the same, and we may look upon matter as being latent force and upon force as being free matter. Finally, if the great first cause is not to remain eternally in a state of inactivity, or, in other words, if it is to manifest itself as matter and motion, and if motion is to act within matter, then there must be a cause why such an activity takes place, and this cause can be nothing else but the eternally active Great First Cause itself, because there can be only one universal cause and no other. This is a self-evident truth to all who are able to see it. There can be no special name for that cause, because it is in itself the all and cannot be specified, for it is, in itself, everything that exists. It, however, appears to us in manifold aspects, and according to the aspects under which it appears, we may give to it different names. Looked at in its aspect as an universal power, which causes action and reaction, we may call it the will, existing within all forms in an active and latent condition. Whenever it becomes active, it may act unconsciously, consciously, or with self-consciousness, according to the conditions under which it is active.

The great and universal trinity of cause, motion, matter---or, as others call it, will, thought, and manifestation---was known to the ancient Rosicrucians and adepts as prima materia. Paracelsus expressly states that each of the three is also the other two, for nothing can possibly exist without cause, matter, and energy, i.e., spirit, matter, soul, the ultimate cause of existence being that it exists. We may, therefore, look upon all forms of activity as being an action of the universal or Divine will upon the ether...It would be useless for us to speculate about the spiritual power of the will if acting through the organism of an adept; but we may study the effects of that same will-power when it acts within a more material plane, where it is known to us as causing heat, light, sound, electricity, and magnetism. All these forms of energy may theoretically be transformed one into another, because they all manifest themselves as various rates of vibrations or undulations of the ether which is contained in everything; and if we can change the rate of these vibrations, we may transform one form of energy into another.

For a long time it has been known to modern science that one form of energy can be transformed into another, although with a certain amount of loss; and it was believed impossible that one amount of energy, if transformed into another, would cause more than the same amount to become manifest. The cause of this false conclusion rests in the still prevailing misconception that a form or substance creates or produces an energy, while, in fact, the form is only an instrument through which the universal and pre-existing motion acts....

Worlds and planets are the products of the pre-existing cosmic ether or space, and not the ether the products of the planets! The same fundamental law evidently exists in all departments of nature, manifesting itself differently according to the difference of conditions under which it acts. Universal forces are bound into forms, and the forms dissolved into forces. Every form, on giving up its ghost, renders to the universal storehouse that which has been entombed in the form, but no more; in the same sense as steam, cooled off into water and frozen into an icicle, will, if heated, produce the same amount of heat again. The universal forces exist not merely in the form, but also in the universal storehouse in nature. By means of a glass lens we may collect the heat which exists in the light of the sun's rays and set a piece of wood on fire. No heat exists in the wood; it is merely a certain motion of the ether, which has been latent, and which is rendered free by the process of burning. As in heat, so in sound. No sound exists in a fiddle; it is the ether in the atmosphere which is transformed into vibration of sound by the instrumentality of the fiddle. No light exists in the fire; it is merely the ether which, by the process of combustion, is transformed into certain vibrations which ultimately produce the phenomenon called light. No magnetism exists in iron; but ether, in a certain state of vibration which we call magnetic, acts through the instrumentality of the iron. No life is produced through a vegetable or animal organism; but they are instruments through which the universal element may manifest itself as life. No thought is created with the brain; but the brain is an instrument through which the universal mind operates. No love, will, faith, or any other spiritual power is created by the soul; but the sole is an organism through which these eternal and self-existent powers may become manifest. There is before me a little electrical instrument, invented by a well-known Austrian inventor, which collects and produces electricity directly from the ether of the atmosphere without any friction of solid corporeal substances and without any chemical agency. Moreover, the amount of electricity produced by it is far greater than that produced by a great engine with friction; a continuous stream of electric fire proceeds from it five to seven inches in length. It clearly proves that the electricity does not reside in the substance by means of which it is produced, but in the ether contained in the atmosphere, from which it is collected by means of the instrument and rendered perceptible to our senses. It also shows that electricity(i.e., the ether in that state of atomic vibration which we call electricity; as this is Keely's definition of electricity, it should not be attributed to Hartmann;) is something substantial, for it produces an electric gush of wind similar to the vapour produced by an atomizer; or still more resembling the cold gushes known to spiritualists, and which often occur at the beginning of some so-called spiritual manifestation....If we had any means to induce certain vibrations of ether in the air, or in the ether of space, by producing them in some substance able to communicate them to the ether of space, we might set the whole atmosphere, or even all the ether of space, into certain vibrations, and exercise a power whose limits cannot be estimated by our present comprehension. On the material plane we can only deal with those powers which we can insulate or store up in a form. We can store up heat, light, electricity, magnetism, and motion; but we cannot store up ether in its original form, because it prevades all known substances. There is nothing which offers any resistance to it. We can, therefore, deal with ether only when it becomes manifest to us through the instrumentality of a substance or for; that is to say, we can deal with it when transformed into heat, electricity, etc. Then it has entered into a state which renders it capable to be insulated by certain substances which offer resistance to it. We must, therefore, conduct our physical experiment with ether stored up in material forms. Everybody knows that a note struck upon an instrument will produce sound in a correspondingly attuned instrument in its vicinity. If connected with a tuning fork, it will produce a corresponding sound in the latter; and if connected with a thousand tuning forks, it will make all the thousand sound, and produce a noise far greater than the original sound, without the latter becoming any weaker for it. Here, then, is an augmentation of multiplication of power, as it has been called by the ancient Rosicrucians, while modern scientists have called it the law of induction. If we had any means to transform sound again into mechanical motion, we would have a thousand-fold multiplication of mechanical motion. It would be presumptuous to say that it will not be as easy for the scientists of the future to transform sound into mechanical motion, as it is for the scientist of the present to transform heat into electricity. Perhaps Mr. Keely has already solved the problem. There is a fair prospect that in the very near future, we shall have, in his ethereal force, a power far surpassing that of steam or electricity. Nor does the idea seem to be Utopian if we remember that modern science heretofore only knew the law of the conservation of energy; while to the scientist of the future the law of the augmentation of energy will be unveiled. As the age which has passed away has been the age of steam, the coming era will be the age of induction. There will be a universal rising up of lower vibrations into higher ones, in the realm of motion, emotion, and thought. Mr. Keely will, perhaps, transform sound into mechanical motion by applying the law of augmentation and multiplication of force known to the ancient Rosicrucians; as we will apply the same law in the realm of thought, and induce people to think. Thus matter will become more subject to the action of the finer forces of nature, and the minds of men will become less gross and easier to be penetrated by the light of Divine wisdom. All this has been predicted eighty-eight years ago, at the beginning of the century.

Mr. Keely, finding in his first interview with Dr. Hartmann about ethereal force, or dynaspheric force, was so well understood by that learned gentlemen, expressed great pleasure in meeting, for the first time, one who comprehended so much more of its nature than any man whom he had ever met; and Dr. Hartmann expressed himself as equally charmed and satisfied with Mr. Keely in this interview; although he gained from him nothing in the way of information that was new to him.

Before the second meeting took place, one of Mr. Keely's papers upon disturbed equilibrium (in the brain) was given to Dr. Hartmann to read with the request that he would limit his next conversation with Mr. Keely to the proper method of re-adjusting opposing conditions in the brain --- or, in other words, ascertaining how the ruling medium could be brought to bear upon these opposing conditions, in the brain, in order to restore equilibrium. Mr. Keely's paper amply treats the cause of disturbance of equilibrium in the brain, producing insanity; and reads as follows:---

Brain Disturbance

"In considering the mental forces as associated with the physical, I find, by my past researches, that the convolutions which exist in the cerebral field are entirely governed by the sympathetic conditions that surround them.

The question arises, what are these aggregations and what do they represent, as being linked with physical impulses? They are simply vibrometric resonators, thoroughly subservient to sympathetic acoustic impulses given to them by their atomic sympathetic surrounding media, all the sympathetic impulses that so entirely govern the physical in their many and perfect impulses (we are now discussing purity of conditions) are not emanations properly inherent in their own composition. They are only media-the acoustic media-for transferring from their vibratory surroundings the conditions necessary to the pure connective link for vitalizing and bringing into action the varied impulses of the physical.

All abnormal discordant aggregations in these resonating convolutions produce differentiation to concordant transmission; and, according as these differentiations *exist in volume, so the transmissions are discordantly transferred, producing antagonism to pure physical action.*

Thus, in Motor Ataxis, a differentiation of the minor thirds of the posterior parietal lobule produces the same condition between the retractors and extensors of the leg and foot; and thus the control of the proper movements is lost through this differentiation. The same truth can be universally applied to any of the cerebral convolutions that are in a state of differential harmony to the mass of immediate cerebral surroundings. Taking the cerebral condition of the whole mass as one, it is subservient to one general head center, although as many neutrals are represented as there are convolutions.

The introductory minors are controlled by the molecular; the next progressive third by the atomic; and the high third by the Etheric. All these progressive links have their positive, negative, and neutral position. When we take into consideration the structural condition of the human brain, we ought not to be bewildered by the infinite variety of its sympathetic impulses; inasmuch as it unerringly proves the true philosophy that the mass chords of such structures are governed by vibratory etheric flows---the very material which composes them. There is no structure whatever, animal, vegetable, mineral, that is not built up from the universal cosmic ether. Certain orders of attractive vibration produce certain orders of structure; thus, the infinite variety of effects---more especially in the cerebral organs. The bar of iron or the mass of steel, have, in each, all the qualifications necessary, under certain vibratory impulses, to evolve all the conditions that govern that animal organism---the brain: and it is as possible to differentiate the molecular conditions of a mass of metal of any shape so as to produce what you may express as a crazy piece of iron or a crazy piece of steel; or, vice versa, an intelligent condition in the same.

I find in my researches, as to the condition of molecules under vibration, that discordance cannot exist in the molecule proper; and that it is the highest and most perfect structural condition that exists; providing that all the progressive orders are the same. Discordance in any mass is the result of differentiated groups, induced by antagonistic chords, and the flight or motions of such, when intensified by sound, are very tortuous and zig-zag; but when free of this differentiation are in straight lines. Tortuous lines denote discord, or pain; straight lines denote harmony, or pleasure. Any differentiated mass can be brought to a condition of harmony, or equation, by proper media, and an equated sympathy produced.

There is good reason for believing that insanity is simply a condition of differentiation in the mass chords of the cerebral convolutions, which creates an antagonistic molecular bombardment towards the neutral or attractive centers of such convolutions; which, in turn, produce a morbid irritation in the cortical sensory centers in the substance of idea-

tion; accompanied, as a general thing, by sensory hallucinations, ushered in by subjective sensations; such as flashes of light and colour, or confused sounds and disagreeable odours, etc., etc.

There is no condition of the human brain that ought not to be sympathetically coincident to that order of atomic flow to which its position, in the cerebral field, is fitted. Any differentiation in that special organ, or, more plainly, any discordant grouping tends to produce a discordant bombardment---an antagonistic conflict; which means the same disturbance transferred to the physical, producing inharmonic disaster to that portion of the physical field which is controlled by that especial convolution. This unstable aggregation may be compared to a knot on a violin string. As long as this knot remains it is impossible to elicit, from its sympathetic surroundings, the condition which transfers pure concordance to its resonating body. Discordant conditions, i.e., differentiation of mass, produce negatization to coincident action.

The question now arises. What condition is it necessary to bring about in order to bring back normality, or to produce stable equilibrium in the sympathetic centers?

The normal brain is like a harp of many strings strung to perfect harmony. The transmitting conditions being perfect, are ready, at any impulse, to induce pure sympathetic assimilation. The different strings represent the different ventricles and convolutions. The differentiations of any one from its true setting is fatal, to a certain degree, to the harmony of the whole combination.

If the sympathetic condition of any physical organism carries a positive flow of 80 per cent on its whole combination, and a negative one of 20 per cent, it is the medium of perfect assimilation to one of the same ratio, if it is distributed under the same conditions to the mass of the other. If two masses of metal, of any shape whatever, are brought under perfect assimilation, to one another, their unition, when brought into contact, will be instant. If we live in a sympathetic field we become sympathetic, and a tendency from the abnormal to the normal presents itself by an evolution of a purely sympathetic flow towards its attractive centers. It is only under these conditions that differentiation can be broken up, and a pure equation established. The only condition under which equation can never be established is when a differential disaster has taken place, of 66 2/3 against the 100 pure, taking the full volume as one. If this 66 2/3 or even 100 exists in one organ alone, and the surrounding ones are normal, then a condition can be easily brought about to establish the concordant harmony or equation to that organ. It is as rare to find a negative condition of 66 2/3 against the volume of the whole mass, as it is to find a coincident between differentiation; or, more plainly, between two individuals under a state of negative influence. Under this new system, it is as possible to induce negations alike as it is to induce positives alike.

Pure sympathetic concordants are as antagonistic to negative discordants as the negative is to the positive; but the vast volume the sympathetic holds over the non-sympathetic, in ethereal space, makes it at once the ruling medium and re-adjuster of all opposing conditions if properly brought to bear upon them."

(signed) Keely

Until Keely's Theoretical Expose is given to science, there are few who will fathom the full meaning of these views. So little did Dr. Hartmann comprehend the principle involved that he ignored them altogether, and in the more than one hour's conversation with Mr. Keely which followed, instead of keeping to this subject matter as requested, he made no illusion to it whatever, and confined his investigations to the mechanical work of Mr. Keely in its application to machinery. In leaving, Mr. Keely again expressed his great delight in meeting one who knew so much of the hidden working of some of nature's laws; whereas, after his departure, Dr. Hartmann announced it as his opinion that, although Mr. Keely had made the greatest discovery of this or of any other age, he would never be able to utilize the force in mechanics, and that his mission would be to spiritualize the world instead of advancing its material progress.

Some days later, when Mr. Keely was asked why it was that Dr. Hartmann no longer believed in the mechanical success of Mr. Keely's inventions, the reply was made that Dr. Hartmann, in disclosing his own views and theories and philosophy, had prevented Mr. Keely from any attempt to point out the errors in these views and theories: feeling, as Mr. Keely did feel, that he would be wanting in humility to dispute with one so learned as Dr. Hartmann, and preferring to wait until the court had removed the injunction placed upon him (Keely), when he would be at liberty to demonstrate to Dr. Hartmann the nature of his errors by the operation of his inventions. However, this delay was not necessary, inasmuch as upon the occasion of Dr. Hartmann's first visit to the workshop, where he saw the old generator, the old Liberator and other machinery, his knowledge that, by means of the vibrations of Ether called Sound, the molecular structure of bodies may be changed, even though these vibrations are not audible to the human ear, caused Dr. Hartmann to confess his error, and to assert that his confidence in Mr. Keely's mechanical success was reestablished and stronger than it had ever been before. Those scientists who, because they could not hear the vibrations of sound, in Mr. Keely's Liberator, denied its operation, saying that one could not make something out of nothing, seem to forget that there are inaudible vibrations of sound as there are invisible rays of light.

Dr. Hartmann knows that everything in nature has its own appropriate 'sound', 'color', and 'number', and can be acted upon as soon as we are in possession of its 'key-note'. This knowledge enabled him to grasp the principle of Mr. Keely's inventions, as soon as the action of the mechanism was explained to him. Although Dr. Hartmann then and there expressed his intention of sacrificing some of his property in order to invest in the new company, in process of organization, it was from no sordid motive that he was so intensely interested in the practical part of Mr. Keely's work; but, having seen such marvelous effects produced in occult experiments, while residing in India, he was inclined to attribute to Mr. Keely natural occult powers which could never be made available in mechanics. Mr. Keely's financial success depends upon the prolongation of his life until his work of evolution is completed. Therefore, the writer of this paper has never advised anyone to invest on such an uncertainty and she requested Dr. Hartmann not to do so.

Mr. Keely's discovery embraces the manner or way of obtaining the key-note, or chord of mass, of mineral, vegetable, and animal substances; therefore, the construction of instruments, or machines, by which this law can be utilized in mechanics, in arts, and in restoration of equilibrium in disease, is only the question of the full understanding of the operation of this law. Herein lies Mr. Keely's work of evolution.

The principal point of difference existing between Mr. Keely and Dr. Hartmann, in their views respecting force, lies in the former attributing the so-called forces of nature to various modes of vibration, as to the length and direction of the vibrations; while the latter attributes all forces to various modes of vibration, as to the number of vibrations in a second. Electricity Mr. Keely defines as a certain form of atomic vibration.

Keely estimates that, after the introductory impulse is given on the harmonic thirds, molecular vibration is increased from 20,000 per second to 100,000,000. On the enharmonic sixths, that the vibration of the inter-molecule is increased to 300,000,000.

On the diatonic ninths, that atomic vibration reaches 900,000,000; on the dominant etheric sixths, 8,100,000,000; and on the inter-etheric ninths, 24,300,000,000; all of which can be demonstrated by sound colours.

In such fields of research, Mr. Keely finds little leisure. Those who accuse him of dilly-dallying, of idleness, of always gong to do and never doing, of visionary plans, etc., etc., know nothing of the infinite patience, the persistent energy, which for a quarter of a century

One Phase of Keely's Discovery in Its Relations to the Cure of Disease

has upheld him in his struggle to attain this end. Still less, if possible, is he understood by those who think he is seeking self-aggrandizement, fame, fortune, or glory.

The time is approaching when all who have sought to defame this discoverer and inventor, all who have stabbed him with unmerited accusations, all who have denounced him as a bogus inventor, a fraud, an impostor, a charlatan, a modern Cagliostro, will be forced to acknowledge that he has done a giant's work for true science, even though he should not live to attain commercial success. But history will not forget that, in the nineteenth century, the story of Prometheus has been repeated, and that the greatest mind of the age, seeking to scale the heavens to bring down the light of truth for mankind, met with Prometheus's reward.

Mrs. Clara F. Bloomfield-Moore
July 1, 1888

NOTE.---Dr. Hartmann, in a report, or condensed statement, in reference to Keely's discovery, writes as follows: He will never invent a machine by which the equilibrium of the living forces in a disordered brain can be restored.

As such a statement would lead the reader of the report to fancy that Keely expected to invent such an instrument, it is better to correct the error that Dr. Hartmann has fallen into. Keely has never dreamed of inventing such an instrument. He hopes, however, to perfect one that he is now at work upon, which will enable the operator to localize the seat of disturbance in the brain in mental disorders. If he succeeds, this will greatly simplify the work of re-adjusting opposing conditions; and will also enable the physician to decide whether the differential disaster has taken place which prevents the possibility of establishing the equation that is necessary to a cure.

According to Keely's theories it is that form of energy known as magnetism - not electricity - which is to be the curative agent of the future, thus reviving a mode of treatment handed down from the time of the earliest records, and made known to the Royal Society of London more than fifty years since by Professor Keil, of Jena, who demonstrated the susceptibility of the nervous system to the influence of the natural magnet, and its efficacy in the cure of certain infirmities, as thousands can testify in our day who are indebted to Parke's Compound Magnets for relief; trying them as the last resort after having suffered much at the hands of many physicians, as St. Paul said. A grandson of Goethe, after calling upon Robert Browning many years since, returned to inquire if he had dropped the magnet there which he was wearing, as he has missed it after leaving the house. The effect of the magnet is one of the effects of the law of sympathetic association, which Keely demonstrates as the governing medium of the universe throughout animate and inanimate nature.

As Cheston Morris, M.D., has well said in his paper on Vital Molecular Vibrations, We are entering upon a new field in biology, pathology, and of course, therapeutics, whose limiits are at present far beyond our ken.

The adaptability of drugs, says Dr. Henry Wood, to heal disease is becoming a matter of doubt, even among many who have not yet studied deeper causation. Materia Medica lacks the exact elements of a science. The just preponderance, for good or ill, of any drug upon the human system is an unsolved prolem, and will so remain ... After centuries of professional research, in order to perfect the art of healing, diseases have steadily grown more subtle and more numerous ... Only when internal, divine forces come to be relied upon, rather than outside reinforcement, will deterioration cease. Said Plato, 'You ought not to attempt to cure the body without the soul.'

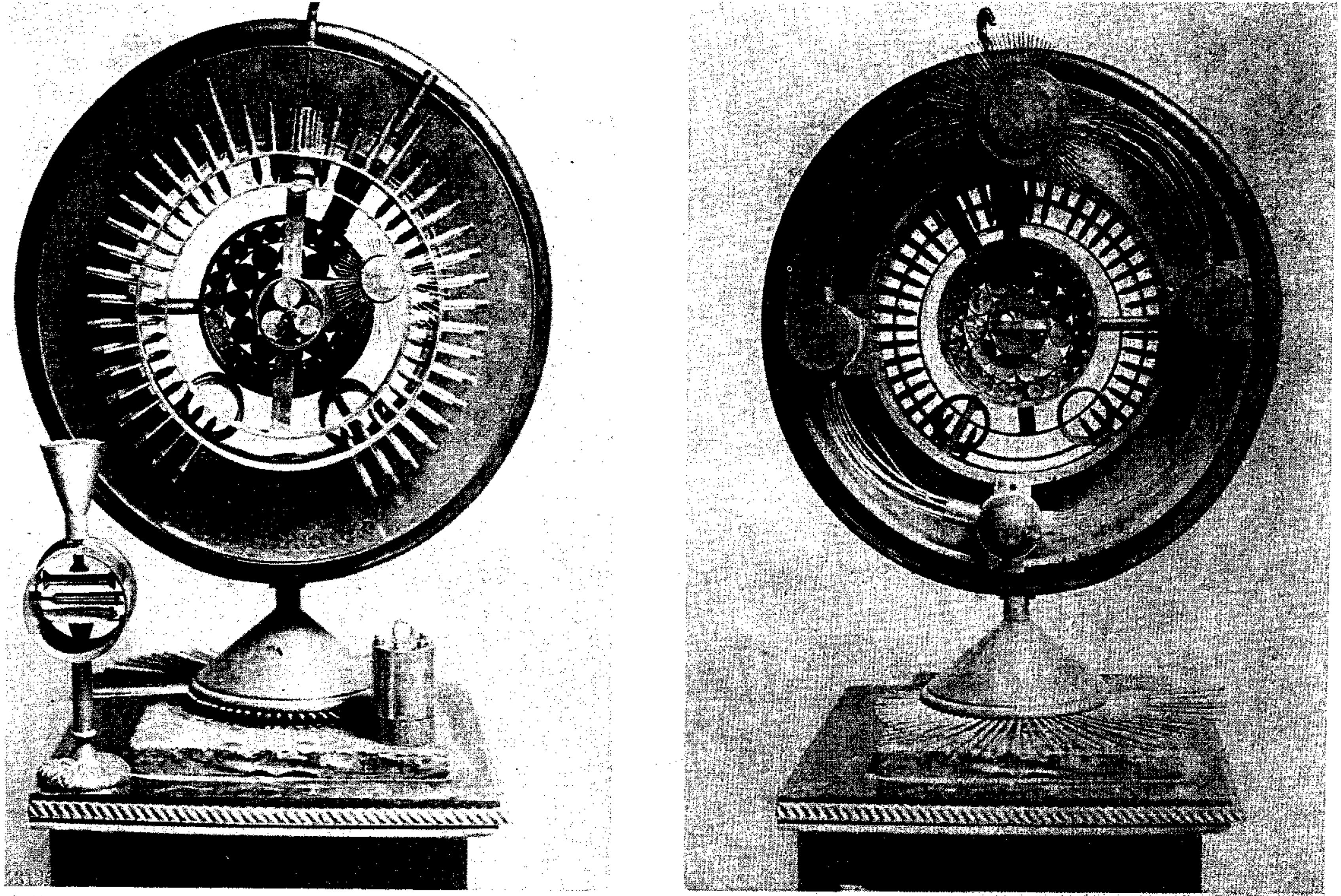

C. COLLIER,
GRAPHER,
[illegible]

Keely's Compound Disintegrator.

On Dynaspheric Force

by Laurence Oliphant, 1888

Recent scientific research has proved conclusively that all force is atomic. That electricity consists of files of particles, and that the interstellar spaces contain substance, whether it be called ether, or astral fluid, or by any other name, which is composed of atoms, because it is not possible to dissever force from its transmitting medium. The universe, therefore, and all that it contains, consists of matter in motion, and is animated by a vital principle which we call God. Science has further discovered that these atoms are severally encompassed by an ethereal substance which prevents their touching each other, and to this circumambient interatomic element they have given the name dynasphere--- but inasmuch as it has further been found that in these dynaspheres there resides a tremendous potency, it is evident that they also must contain atoms, and that these atoms must in their turn be surrounded by dynaspheres, which again contain atoms, and so on ad infinitum. Matter thus becomes infinite and indestructible, and the force which pervades it, persistent and everlasting.

This dynaspheric force, which is also called etheric, is conditioned as to its nature on the quality of the atoms which form its transmitting media, and which are infinite both in variety and in their combinations and permutations. They may, however, be broadly divided into two categories, the sentient and the non-sentient atoms.

Dynaspheric force, composed of non-sentient atoms, is the force that has been already mechanically applied by Mr. Keely to his motor, and which will probably ere long supersede the agencies now used for locomotives, projectiles and other purposes; when the laws which govern it come to be understood, it will produce materially a great commercial and industrial revolution. There is no hard-and-fast line between the sentient and non-sentient atoms; just as zoophytes are a connecting link between the animal and vegetable creation, so there is a graduated scale of atoms between those which, although animated by the divine life, are not sentient, and those which are as highly developed relatively to them as man is to a cabbage. For the highest class of sentient atoms through which divine force is transmitted are in the perfect human form. They are infinitesimal bi-sexual innocences male and female, two in one. The tradition of fairies is the lingering consciousness, come down from a remote past, of this fact.

Owing to the unhappily debased condition of our planet, this force is not now operant upon it, except to a very limited and imperfect degree - it is struggling, however, to penetrate into the human organism through the channel provided for it, and this channel must, of necessity, partake of the nature of the forces operant within it---in other words, it must be a bi-sexual channel. It was this bi-sexual channel which Christ came to restore by his mission to earth; and thus inaugurate a process by which man should regain his lost bi-une condition. That process has now partially achieved its consummation in the advent of the complementary half of man whom we call the sympneuma. It is only through the sympneuma that the dynaspheric force, consisting of bi-sexual atoms, can be projected into nature. It comes for the healing of the nations, and is all the more necessary now because the conditions of nature have of late years undergone such a change as to render possible the invasion of the human organism by forces similar in character, with this one difference, that the atoms of which they are composed are not bi-sexual. These forces exhibit themselves in the phenomena of hypnotism, thought-reading, telepathy, mesmeric healing, spiritualistic manifestations, and in divers other ways, and depend for their quality on the source of their projection in the invisible and the human medium through whom they are transmitted---where both are bad, the atoms are in the form of infusoria, or predatory animalculae, who prey upon each other, and work moral and physical malady. Where both are relatively good, they are in the form of separate uni-sexual beings, depending for their quality upon the medium, and partaking of what moral taint his nature may possess. It must be said that the same remark applies to the bi-sexual atoms of the sympneumatic force; but, although imperfect, there is this guarantee for their superior quality, that it is not possible for a human being to enter upon sympneumatic conditions, excepting after a long and arduous discipline and self-sacrifice for his neighbor, and of great sufferings.

The sympneuma visits none who have not been thus prepared, and who do not live exclusively for the service of humanity, to the extinction of private affections, personal ambitions, or worldly considerations of any kind.

A false sympneuma may, however, visit those who are wholly engrossed by self; such are the succabi and incubi---well-known by the Church---and the force acting through them is the most fatal which can operate upon earth.

There are methods, however, not necessary to enter upon here, by which the true can be distinguished from the false with absolute certainty. All human emotion is atomic, and it has never been possible that it should be otherwise. The peculiarity of the atomic force of the present day is, that it has received an immense accession of energy, through changes which have operated in the invisible.

It is these changes which render will-force, and magnetic influence so much more powerful now than they were formerly; and hence it becomes of such transcendent importance that persons who find themselves in possession of this reinforced energy, and able to operate upon others hypnotically or for curative purposes, should realize the character of the agency they are dealing with---for it is quite impossible for them to project this will-force, or magnetic influence, into the organism of another, without projecting the atoms with it. Now, these atoms vary in quality from the predatory animalculae to the human form through an infinite variety of types; none of them pure and good, though some are far purer and better, relatively, than the others---still no magnetiser is so perfect that his magnetism does not convey to his patients the atoms of the vices and defects peculiar to his own nature, of which they may have been comparatively free.

It may thus happen that a magnetiser, while healing the body of a patient, may work irreparable moral injury to his soul, and this while animated with the best intentions, and quite unconsciously to himself. It often happens, moreover, that the progress of the soul can only be achieved by an attenuation of the external structural atoms, thus producing physical disease; to heal a person thus under going moral treatment, directed from the unseen world, by a sudden and premature exercise of will-force in this one, applied to his surface organism, is to render him a fatal service. Again, it may be that the welfare of a person's soul is dependent upon its removal from the body at a certain juncture; here, again, human interference, by the operation of the human will being free, and yet under specific law, that free operation cannot be arbitrarily hindered in defiance of the law under which it acts.

The reason why material remedies of all kinds may be employed with safety and propriety, is because the curative forces they contain are not composed of sentient atoms, and can be controlled from the unseen in quite a different manner from those which are---which may, to a certain extent, be influenced by them, but cannot be controlled. When a person has reached the point, which may be attained after a long sympneumatic training, and a life passed under the influence of that training, of having no will but that of God operating freely in him, as his own, he may, under a pressure, which he will recognize as a divine impulse, put forth a healing power, but he will have no personal desire connected with it; the healing force will be put through him irrespective of any conscious will he used; the energy he projects will convey bisexual atoms, which may prove a seed sown as a preparation for a sympneumatic descent.

At such moments the operator will hold himself exclusively open to Christ, for it cannot be too earnestly insisted upon, that Christ is the one source and channel of sympneumatic life, and the healing which comes through it, when a person's moral condition renders such physical healing desirable.

In the presence of the rapid development which dynaspheric force is acquiring, and of the great interest which it is attracting, especially among good and earnest truth seekers, who are only investigating it with the object of turning it to account for the benefit of humanity, it has seemed to me necessary to make these remarks. I have done so in the hope that they might serve as a warning and an encouragement---as a warning of the dangers that beset the unwary explorers into these little known and almost untrodden regions; and as an encouragement as indicating the immense potentialities now descending upon the world for its succor in the hour of its approaching need.

If I seem to have written with the certainty of conviction, it is with no desire to impose my authority arbitrarily upon my readers, but in all humility to give them the facts as they have been revealed to me, after an arduous struggle and investigation into the methods of operation of these forces, which has lasted nearly twenty-five years.

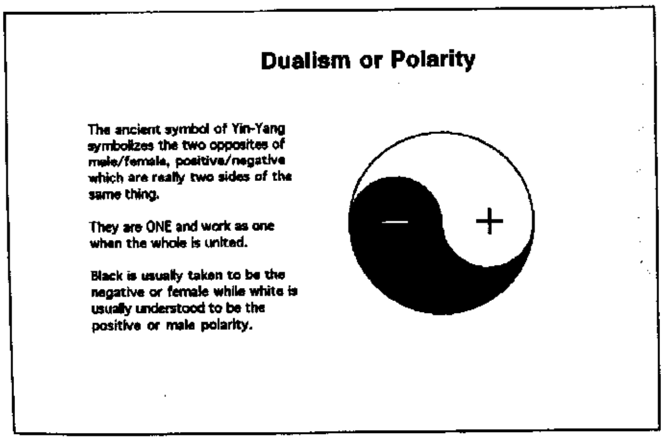

Dualism or Polarity

The ancient symbol of Yin-Yang symbolizes the two opposites of male/female, positive/negative which are really two sides of the same thing.

They are ONE and work as one when the whole is united.

Black is usually taken to be the negative or female while white is usually understood to be the positive or male polarity.

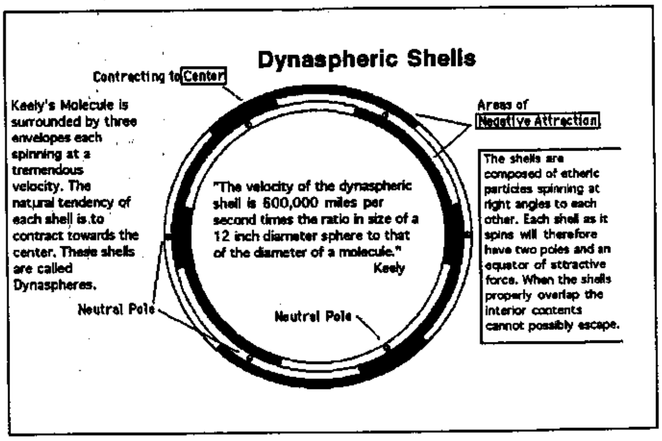

Dynaspheric Shells

Keely's Molecule is surrounded by three envelopes each spinning at a tremendous velocity. The natural tendency of each shell is to contract towards the center. These shells are called Dynaspheres.

"The velocity of the dynaspheric shell is 600,000 miles per second times the ratio in size of a 12 inch diameter sphere to that of the diameter of a molecule."
Keely

The shells are composed of etheric particles spinning at right angles to each other. Each shell as it spins will therefore have two poles and an equator of attractive force. When the shells properly overlap the interior contents cannot possibly escape.

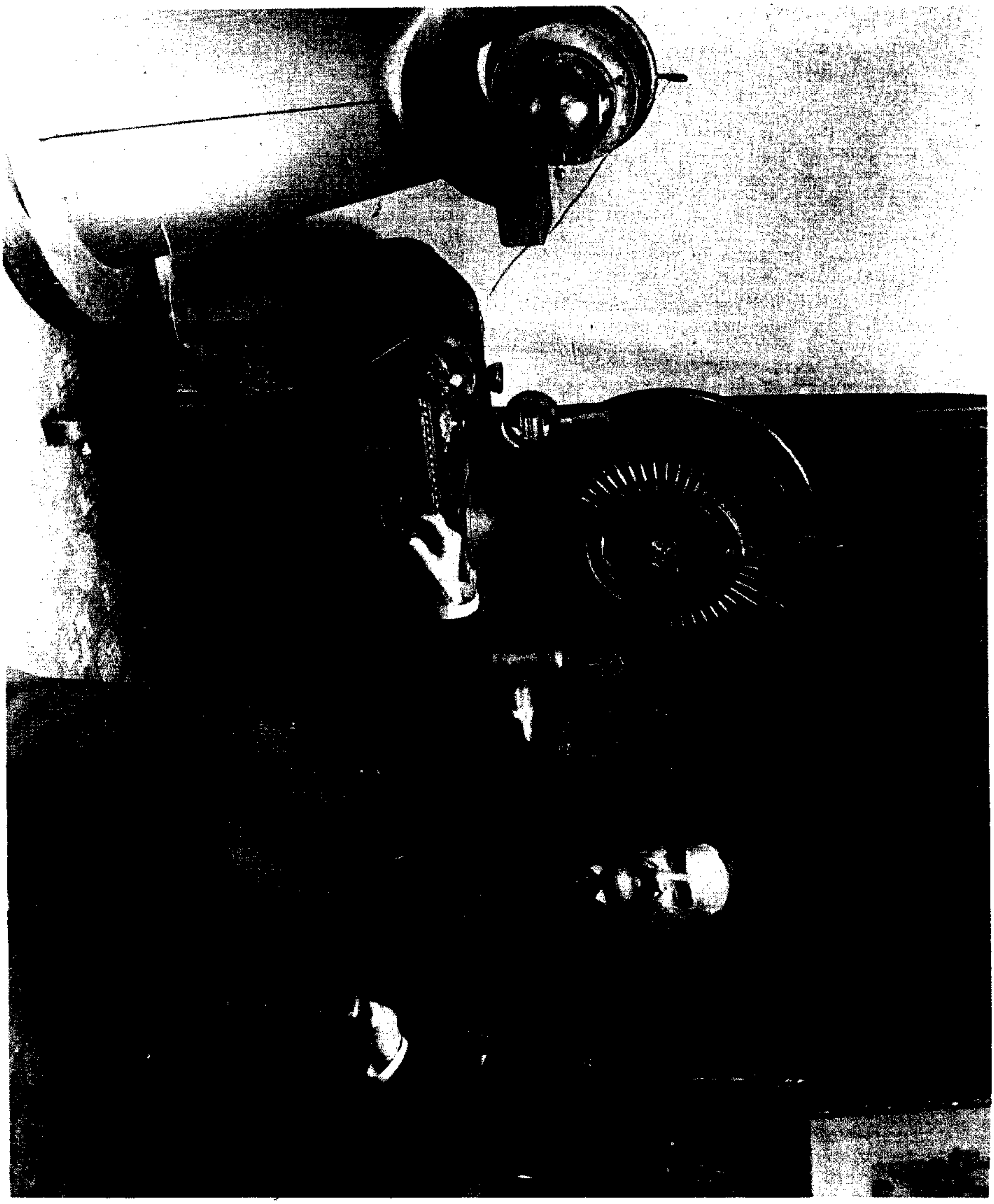

What Is Matter And What Is Force

(from The Theosophist, September, 1888)

[The continual discoveries with which modern science astonishes and bewilders the world create a presumption in the many minds that an article dealing with scientific subjects becomes unreliable when it is a few years old, because most probably superseded by new facts or theories, discovered or formulated in the meantime by the heirophants of the laboratory or the lecture hall, but not yet generally known. Some of these discoveries would in legal phraseology be termed retroactive - they upset theories hitherto upheld as axioms of science or laws of nature; and science that has become out of date is quite as much to be avoided as last year's fashions. This, of course, is a danger which necessarily threatens anyone who, not being in the inner circle of scientists, quotes from even the latest editions of scientific works, and the uncertainty thereby created does much to keep up the unquestioning faith which so many persons feel in the dicta of modern science, paradoxical as that may sound. The reason is plain enough. Men of science stand to the vulgar in the position of revelators. It does not so much matter if they are mistaken today, since they will most likely correct their mistakes tomorrow; and this uncertainty does for the men of science what the constant promise of a new revelation did for Joe Smith and Brigham Young - it creates expectant attention and happy anticipation. This uncertainty of science tends also to discourage undue curiosity and criticism on the part of the public, and to favor the autocratic assumptions of scientific authorities; for naturally there does not seem to be much use for outsiders to spend their time in learning abstruse things that may be discovered, in a few months or years, to have after all been altogether falsely conceived and wrongly explained. In the case of the following article, however, most of the scientific theories and data quoted are still standing, although the article is over five years old; and the occult views of nature it puts forth are true for all time. For this reason it has been thought well to republish it without delay, before the scientific views it deals with have any more time to turn sour or explode. The article, it should be mentioned, was written in answer to some objections raised by a correspondent in a previous number of The Theosophist, as to the designation of electricity as a form of matter; and the writer does not imagine that the question which forms its title is fully answered in the text. There remain several aspects of force almost wholly unnoticed, and it would, of course, be ridiculous to suppose that the subject of Force has been exhausted by dealing with some of its manifestations. Had it been the intention of the article to settle the questions at issue, the writer would not have taken for a title the query, What is Matter and What is Force? but would probably have adopted the usual style of teachers of science, and headed the article, What Matter and Force are. Enough is said in the article, however, to answer its original purpose, namely, of showing that the all-important questions, What is matter? and What is Force? have received no satisfactory answer from modern science, and of proving that no one has any right, in the present state of scientific knowledge, to assert, in the face of the occultists, that electricity is not a form of matter, more especially so since Helmholtz has declared it to be as atomic as ponderable matter itself.]

It is a question of science, which, as such, has to be strictly kept within the boundaries of modern materialistic science. All discussion on the subject, however desirable, would prove, on the whole, unprofitable. Firstly, because science confines herself only to the physical aspect of the conservation of energy or correlation of forces; and, secondly, because, notwithstanding her own frank admission of helpless ignorance as to the ultimate causes of things, judging by the tone of our critic's article, I doubt whether he would be willing to admit the utter inaptness of some of the scientific terms as approved by the Dwija, the twice-born of the Royal Society, and obediently accepted by their admirers. The fact that modern science has been pleased to divide and subdivide the atmosphere into a whole host of elements, and to call them so for her own convenience, is no authoritative reason why Occultists should accept that terminology. Science has never yet succeeded in decomposing a single one of the many simple bodies miscalled elementary substances, for which failure, probably, the latter have been named by their elementary. And whether she may yet, or never may, succeed in that direction in time, and thus recognize her error, in the meanwhile, we Occultists permit ourselves maintain that the alleged primordial atoms would be better classified under any other name but that one. With all respect due to the men of science, the terms element and elementary applied to the ultimate atoms and molecules of matter of which they know nothing do not seem in the least justified. It is as though the Royal Society agreed to call every star a Kosmos, because each star was supposed to be a world like our own planet; and then would begin taunting the ancients universe. So far, however, science admits herself that the words element and elementary, unless applied to primordial principles or self-existing essences, out of which the universe was evoluted, are unfortunate terms, and remarks there upon that experimental science deals only with legitimate deductions from the facts of observation, and has nothing to do with any kind of essences except those which it can see, smell, or taste. Professor J.P. Crooke tells us that science leaves all others to the metaphysicians (New Chemistry, 1887). This stern pronunciamento, which shows the men of science refusing to take anything on faith, is immediately followed by a very curious admission made by the same author, Our theory, I grant, may be all wrong, he adds, and there may be no such thing as molecules...The new chemistry assumes as its fundamental postulate that the magnitudes we call molecules are realities, but this is only a postulate. We are thus made suspect that the exact science of chemistry needs to take as well as transcendental metaphysics something on blind faith. Grant her the postulate - and her deductions make of her an exact science; deny it - and exact science falls to pieces! Thus, in this respect, physical science does not stand higher than psychological science, and the Occultists need fear but very little of the thunderbolts of their exact rivals. Both are, to say the least, on a par. The chemist, though carrying his subdivision of molecules further than the physicist, can no more than he experiment on individual molecules. One may even remind both that neither of them has ever seen an individual molecule. Nevertheless, and while priding themselves upon taking nothing on faith, they admit that they cannot often follow the sub-division of molecules with the eye, but can discern it with the intellect. What more, then, do they do than the Occultists, the alchemists, the adepts? While they discern with the intellect, the adept, as he maintains, can as easily discern the sub-division ad infinitum of that which his rival of the exact methods pleases to call an elementary body, and he follows it - with the spiritual in addition to the physical intellect...We must pass to the more important question now, and see how far science is justified in regarding electricity as a force and . . . Eastern Occultists in maintaining that it is still matter. Before we open the discussion I must be allowed to remark that, since a Theosophist wants to be scientifically accurate, he ought to remember that science does not call electricity a force, but only one of the manifestations of the same; a mode of action or motion. Her list of the various kinds of energy which occur in nature is long, and many are the names which she uses to distinguish them. With all that, one of her most eminent adepts, Professor Balfour Stewart - one of the authorities he quotes against our President - warns his readers (see The Forces and Energies of Nature) that their enumeration has nothing absolute or complete about it, representing, as it does, not so much the present state of our knowledge as our want of knowledge, or rather profound ignorance, of the ultimate constitution of matter. So great is that ignorance, indeed, that, treating upon heat, a mode of motion which is supposed to be better understood than electricity, that scientist confess that if heat be not a species of motion, it must necessarily be a species of matter, and adds that the men of science have preferred to consider heat as a species of motion to the alternative of supposing the creation of a peculiar kind of matter. And, if so, what is there to warrant us that science will not yet find out her mistake some day, and recognize and call electricity, in argument with the Occultists, a species of a peculiar kind of matter. Thus, before the dogmatic admirers of modern science take the occultists to task for viewing electricity under one of its aspects - and for maintaining that its basic principle is - matter, they ought first to demonstrate that science errs when she herself, through the mouthpiece of her recognized high priests, confesses her ignorance as to what is properly force and what is matter. For instance, the same Professor of Natural Philosophy, Mr. Balfour Stewart, LL.D., F.R.S., in his lectures on the Conservation of Energy, tells us as follows: - We know nothing or next to nothing of the ultimate structure or properties of matter, whether organic or inorganic, and . . . it is in truth only a convenient classification and nothing more.

Furthermore, one and all, the men of science admit that though they possess a definite knowledge of the general laws, yet they have no knowledge of the individuals in the domains of physical science. For example, they suspect a large number of our diseases to be caused by organic germs; but they have to avow that their ignorance about these germs is complete. And in the chapter What is Energy? the same great naturalist staggers the too-confiding profane by the following admission: - If our knowledge of the nature and habits of organized molecules be so small, our knowledge of the ultimate molecules of inorganic matter is, if possible, still smaller. It thus appears that we know little or nothing about the shape or size of molecules, or about the forces which actuate them. The very largest masses of the universe sharing with the very smallest this property of being beyond the scrutiny of the human senses. Of physical human senses he must mean, since he knows little, if anything, of any other senses. But let us take note of some further admissions, this time by Professor Leconte, in his lecture on the distinction between force and energy is very imperfectly, if not at all, defined in the higher forms of force, and especially in the domain of life. . . Our language cannot be more precise until our ideas in this department are far clearer than now.

Even as regards the familiar liquid - water - science is at a loss to decide whether the oxygen or hydrogen exist as such in water, or whether they are produced by some unknown and unconceived transformation of its substance. It is a question, says Mr. J.P. Crooke, Professor of Chemistry, about which we may speculate, but in regard to which we have no knowledge. Between the qualities of water and the quality of these gases there is not the most distant resemblance. All they know is that water can be decomposed by an electrical current; but why is it so decomposed and then again recombined, or what is the nature of that they call electricity, *etc.*, they do not know. Hydrogen, moreover, was still very lately one of the very few substances which was known only in its aeriform condition. It is the lightest form of matter known.

There is not an atom in nature but contains latent or potential electricity which manifests under known conditions. Science knows that matter generates* what it calls force, the latter manifesting itself under various forms of energy - such as heat, electricity, magnetism, gravitation, *etc.*, - yet that same science has hitherto been unable as we find from her own admissions, as given above, to determine where it is that matter ends and force (or spirit, as some erroneously call it) begins. Science, while rejecting metaphysics and relegating it through her mouthpiece, Professor Tyndall, to the domain of poetry and fiction, unbridles as often as many metaphysicians her wild fancy, and allows mere hypotheses to run riot on the field of unproved speculation. All this she does, as in the case of the molecular theory, with no better authority for it than the paradoxical necessity for the philosophy of every science to arbitrarily select and assume imaginary fundamental principles; the only proof offered in the way of demonstrating the actual existence of the latter being a certain harmony of these principles with the observed facts. Thus, when men of science imagine themselves subdividing a grain of sand to the ultimate moleculе they call oxide of silicon, they have no real but only an imaginary and purely hypothetical right to suppose that, if they went on dividing it further (which, of course, they cannot), the molecule, separating itself into the chemical constituents of silicon and oxygen, would finally yield that which has to be regarded as two elementary bodies - since the authorities so regard them. Neither an atom of silicon nor an atom of oxygen is capable of any further subdivision into something else, they say. But the only good reason we can find for such a strange belief is because they have tried the experiment and - failed. But how can they tell that a new discovery, some new invention of still finer and more perfect apparatuses and instruments, may not show their error some day? How do they know that those very bodies or molecules, which, when analyzed with still greater minuteness, may show containing in themselves the real primordial elementary globules, the gross encasement of the still finer atom-spark, the spark of life, the source of electricity - matter still! Truly has Henry Kunrath, the greatest of the alchemists and Rosicrucians of the middle ages, shown spirit in man, as in every atom - as a bright flame enclosed within a more or less transparent globule - which he calls soul. And since the men of science confessedly know nothing of (a) the origin of either matter or energy; (b) nor of electricity or life; and (c) that their knowledge of the ultimate molecules of inorganic matter amounts to a cipher. Why, I ask, should any student of Occultism, whose great masters may know, perchance, of essences which the professors of the modern materialistic school can neither see, smell, or taste - why should he be expected to take their definitions as to what is Matter and what is Force as the last word of unerring, infallible science?. . .The term imponderable agents is now regarded as a scientific absurdity. The latest conclusions at which modern chemistry has arrived, it seems, have brought it to reject the word imponderable, and to make away with those text books of pre-modern science which refer the phenomena of heat and electricity to attenuated forms of matter. Nothing, they hold, can be added to or subtracted from bodies without altering their weight. This was said and written in 1876, by one of the greatest chemists in America. With all that, have they become any wiser for it? Have they been able to replace by a more scientific theory the old and tabooed phlogiston theory of the science of Stahl, Priestley, Scheele, and others? Or, because they have proved, to their own satisfaction, that it is highly unscientific to refer the phenomena of heat and electricity to attenuated forms of matter, have they succeeded at the same time in proving what are really Force, Matter, Energy, Fire, Electricity, Life? The phlogiston of Stahl - a theory of combustion taught by Aristotle and the Greek philosophers - as elaborated by Scheele, the poor Swedish apothecary, a secret student of Occultism, who, as Professor Crooke says of him, added more knowledge to the stock of chemical science in a single year than did Lavoisier in his lifetime, was not a mere fanciful speculation, though Lavoisier was permitted to taboo and upset it. But, indeed, were the high priests of modern science to attach more weight to the essence of things than to mere generalizations, then, perhaps, they would be in a better position to tell the world more of the ultimate structure of matter than they now are. Lavoisier, it is well known, did not add any new fact of prime importance by upsetting the phlogiston theory, but only added grand generalization. The Occultists are perfectly aware, it need hardly be said, of modern theories of combustion, and fully recognize the part which oxygen plays therein. They prefer, however, to hold to the fundamental theories of ancient sciences, knowing well that a very large balance of real knowledge lies on that side, when the ancient and the modern are compared. No more than the authors of the old theory do they attach to phlogiston - which has its specific name as one of the attributes of Akasa - the idea of weight which the uninitiated generally associate with all matter. And though to us it is a principle, a well defined essence, yet no more than we did they view it as matter in the sense it has for the present men of science. As one of their modern professors put it - Translate the phlogiston by energy, and in Stahl's work on Chemistry and Physics, of 1731, put energy where he wrote phlogiston, and you have . . . our great modern doctrine of conservation of energy. Verily so it is the great modern doctrine, only plus something else, let me add. Hardly a year after these words had been pronounced, the discovery by Professor Crookes of radiant matter - of which farther on - has nigh upset again all their previous theories.

---

*Using the term generated in the lesser sense of calling forth into manifestation. Force or energy is known to be eternal, and cannot of course be 'generated' in the same sense of being created.

Force, energy, physical agent, are simply different words to express the same idea, observes our critic. I believe he errs. To this day the men of science are unable to agree in giving to electricity a name which would convey a clear and comprehensive definition of this very mysterious agent, as Professor Balfour Stewart calls it. While the latter states that electricity or electrical attraction may probably be regarded as peculiarly allied to that force which we call chemical affinity; and Professor Tyndall calls it only a mode of motion; Professor A. Bain regards electricity as one of the five chief powers or forces of nature:--- One mechanical or molar, the momentum of moving matter, the others molecular, or embodied in the molecules, also supposed(?) in motion - these are heat, light, chemical force, electricity. (The Correlations of Nervous and Mental Forces.) Now, these three definitions would not gain, I am afraid, by being strictly analyzed . . . Light was never regarded as a force. It is, says science, a manifestation of energy, a mode of motion produced by a rapid vibration of the molecules and transmitted by the undulations of ether. The same for heat and sound, the transmission of the latter depending, in addition to the vibrations of ether, on the undulations of an intervening atmosphere. Professor Crookes thought at one time that he had discovered light to be a force, but found out his mistake very soon. The explanation of Thomas Young of the undulatory theory of light, holds now as good as ever in the theories of modern science at least, and according to this explanation that which we call light is simply an impression produced on the retina of the eye by the wave-like motion of the particles of matter. Light, then, like heat - of which it is the crown - is simply the ghost, the shadow of matter in motion! . . The men of science have just found out a fourth state of matter, whereas the Occultists have penetrated years ago beyond the sixth, and, therefore, do not infer, but know of the existence of the seventh, the last. Professor Balfour Stewart, in seeking to show light to be an energy or force, quotes Aristotle, and remarks that the Greek philosopher seems to have entertained the idea that light is not a body, or the emanation of any body (for that, Aristotle says, would be a kind of body), and, that, therefore, light is an energy or act. To this I respectfully demur, and answer that if we cannot conceive of motion without force, we can conceive still less of an energy or act existing in boundless space from the eternity, or even manifesting without some kind of body. Moreover, the conceptions about body and matter of Aristotle and Plato, the founders of the two great rival schools of antiquity, opposed as they were in many things to each other, are nevertheless still more at variance with the conceptions about body and matter of our modern men of science. The Theosophists, old and modern, the Alchemists, and Rosicrucians have ever maintained that there were no such things per se as light, heat, sound, electricity, least of all could there be a vacuum in nature. And now the results of old and modern investigation fully corroborate what they had always affirmed, namely, that in reality there is no such thing as a chemical ray, a light ray, or a heat ray. As far as can be ascertained by those whose observations and experiments are confined to the material plane, there is nothing but radiant energy; or, as a man of science expresses it in the Scientific American, radiant energy, motion of some kind, causing vibrations across space of something between us and the sun; something which, without understanding fully (verily so!), we call 'ether,' and which exists everywhere, even in the vacuum of a radiometer. The sentence, for being confused, is, none the less for it, the last word of science. Again: We have always one and the same cause, radiant energy, and we give this one thing different names - 'actinism,' 'light,' or 'heat.' And we are told also that the miscalled chemical or actinic rays, as well as those which the eye sees as blue, or green, or red, and those which the thermometer feels, are all one thing, the effects of motion in ether. (The Sun's Radiant Energy, by Professor Langley.)

Now, the sun and ether being beyond dispute material bodies, necessarily every one of their effects - light, heat, electricity, *etc.* - must be, agreeably to the definitions of Aristotle (as accepted, though slightly misconceived, by Professor Balfour Stewart), also a kind of body, ergo - matter.

Now what is in reality matter? We have seen that it is hardly possible to call electricity a force, and yet we are forbidden to call it matter under the penalty of being called unscientific. Electricity has no weight - ergo, it cannot be matter. Well, there is much to be said on both sides. Mallet's experiment, which corroborated that of Pirani (1878), showed that electricity is under the influence of gravitation, and must have, therefore, some weight. A straight copper wire, with its ends bent downwards, suspended at the middle to one of the arms of a delicate balance, while the bent ends dip in mercury. When the current of a strong battery is passed through the wire by the intervention of the mercury, the arm to which the wire is attached, although accurately balanced by a counterpoise, sensibly tends downwards, notwithstanding the resistance produced by the buoyancy of the mercury. Mallet's opponents, who tried at the time to show that gravitation had nothing to do with the fact of the arm of the balance tending downward, but that it was due to the law of attraction of electric currents, and who brought forward to that effect Barlow's theory of electric currents, and Ampere's discovery that electric currents, running in opposite directions, repel one another, and are sometimes driven upwards, only proved that men of science will rarely agree, and that the question is far an open one. This, however, raises a side issue as to what is the law of gravitation. The scientists of the present day assume that gravitation and attraction are quite distinct from one another. But the day may not be far distant when the theory of the Occultists that the law of gravitation is nothing more or less than the law of attraction and repulsion will be proved scientifically correct.

Science may, of course, if it so pleases her, call electricity a force. Only by grouping it together with light and heat, to which the name of force is decidedly refused, she has either to plead guilty of inconsistency, or to tacitly admit that it is a species of matter. But whether electricity has weight or not, no true scientist is prepared to show that there is no matter so light as to be beyond weighing with our present instruments. And this brings us directly to the latest discovery, one of the grandest in science; I mean Mr. Crookes' radiant matter, or, as it is now called, the fourth state of matter. That the three states of matter, the solid, the liquid, and the gaseous, are but so many stages in an unbroken chain of physical continuity, and that the three correlate or are transformed one into the other by insensible gradations, needs no further demonstration, we believe. But what is of far greater importance to us Occultists is the admission made by several great men of science in various articles upon the discovery of that fourth state of matter. Says one of them in Scientific American:--- 'There is nothing any more improbable in the supposition that these three states of matter do not exhaust the possibilities of material condition than in supposing the possibilities of sound to extend to aerial undulations to which our organs of hearing are insensible, or the possibilities of vision to ethereal undulations too rapid or too slow to affect our eyes as light.

And as Professor Crookes has now succeeded in refining gases to a condition so ethereal as to reach a state of matter fairly describable as ultra-gaseous, and exhibiting an entirely novel set of properties, why should the Occultists be taken to task for affirming that there lie beyond that ultra-gaseous state still other states of matter; state so ultra-refined, even in their grosser manifestations - such as electricity under all its forms - as to have fairly deluded the scientific senses, and let the happy possessors thereof call electricity - a force! They tell us that it is obvious that if the tenuity of some gas is very greatly increased, as in the most perfect vacua attainable, the number of molecules may be so diminished that their collisions under favourable conditions may become so few in comparison with the number of masses that they will cease to have a determining effect upon the physical character of the matter under observation. In other words, they say: The free-flying molecules, if left to obey the laws of kinetic force without mutual interference, will cease to exhibit the properties characteristic of the gaseous state, and take on an entirely new set of properties. This is radiant matter. And still beyond lies the source of electricity - still matter. . . Speaking of his discovery, Professor Crookes justly remarks that the phenomena he has investigated in his exhausted tubes reveal to physical science a new field for explanation, a new world, a world wherein matter exists in a fourth state, where the corpuscular theory of light holds good, and where light does not always move in a straight line, but where we can never enter, and in which we must be content to observe and experiment from without. To this the Occultist might answer: If we can never enter it with the help of our physical senses, we have long since entered and even gone beyond it, carried thither by our spiritual faculties and in our spirit bodies.

And now I will close this too lengthy article with the following reflection. The ancients never invented their myths. One acquainted with the science of occult symbology can always detect a scientific fact under the mask of grotesque fancy. Thus one who would go to the trouble of studying the fable of Electra - one of the seven Atlantides - in the light of occult science, would soon discover the real nature of Electricity, and learn that it signifies little whether we call it force or matter, since it is both, and so far, in the sense given it by modern science, both terms may be regarded as misnomers. Electra, we know, is the wife and daughter of Atlas the Titan, and the son of Asia and of Pleione, the daughter of the Ocean. . . As Professor Leconte well remarks, there are many of the best scientists who ridicule the use of the term vital force or vitality as a remnant of superstition, and yet the same men use the words gravity, magnetic force, physical force, electrical force, *etc.*, and are unable withal to explain what is life, or even electricity; nor are they able to assign any good reason for the well-known fact that when an animal body is killed by lightning, after death the blood does no coagulate. Chemistry, which shows to us every atom, whether organic or inorganic, in nature susceptible to polarization, whether in its atomic mass or as a unit, and inert matter allied with gravity, light with heat, etc., hence as containing latent electricity, that chemistry still persists in making a difference between between organic and inorganic matter though both are due to the same mysterious energy, ever at work by her own occult processes in Nature's laboratory, in the mineral no less than in the vegetable kingdom. Therefore do the Occultists maintain that the philosophical conception of spirit, like the conception of matter, must rest on one and the same basis of phenomena, adding that force and matter, spirit and matter, or viewed as opposite poles in their respective manifestations, yet they are in essence and in truth but one, and that life is present as much in a dead as in a living body, in inorganic as in organic matter. This is why, while science is searching still, and may go on searching for ever, to solve the problem What is life? the Occultist can afford to refuse taking the trouble, since he claims, with as much good reason as any given to the contrary, that life, whether in its latent or dynamical form, is everywhere, that it is as infinite and indestructible as matter itself, since neither can exist without the other, and that electricity is the very essence and origin of life itself. Purush is non-existent without Prakriti; nor can Prakriti, or plastic matter, have being or exist without Purush, or spirit, vital energy, life. Purush and Prakriti are, in short, the two poles of the one eternal element, and are synonymous and convertible terms. Our bodies as organized tissues are indeed an unstable arrangement of chemical forces, plus a molecular force - as Professor Bain calls electricity - raging in it dynamically during life, tearing asunder its particles at death, to transform itself into a chemical force after the process, and thence again to resurrect as an electrical force or life in every individual atom. Therefore, whether it is called Force or Matter, it will ever remain the omnipresent Proteus of the universe, the one element, Life, Spirit or Force at its negative, Matter at its positive pole; the former the Materio-Spiritual, the latter the Materio-Physical Universe, Nature, Swabhavat or Indestructible Matter.

Unsigned

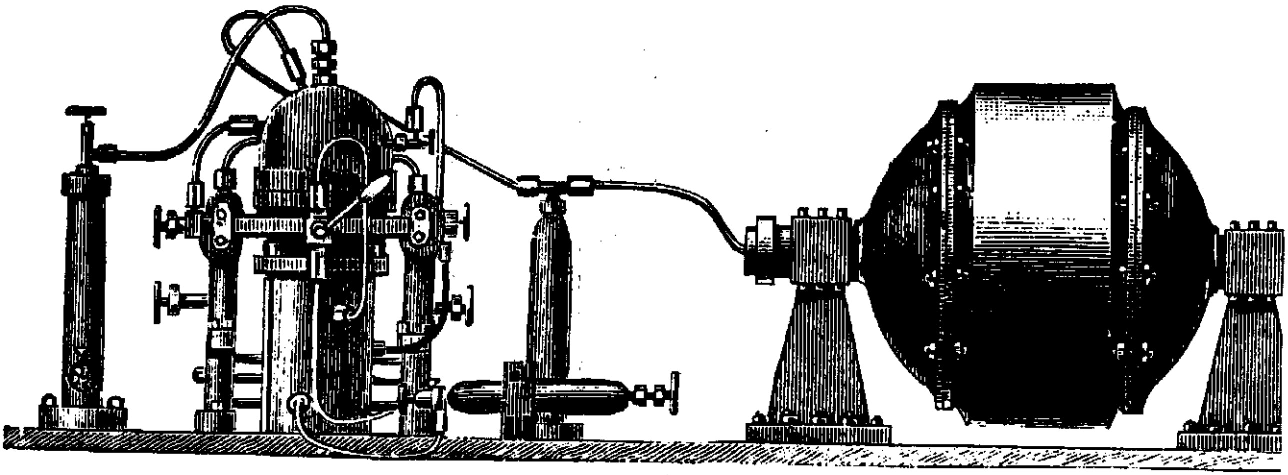

The Astral Light

by Louise A. Off

The Astral Light, as the source of all world phenomena, is a theme of no little importance to the student of occultism. The root of the word Astral is to be found in the Assyrian Istar, signifying star, and was applied to this element by the Kabbalists and later mystics, because they considered the heavenly bodies as the concrete crystallizations of the Astral Light. Some Theosophic writers have confounded the nature of this element with that of Akasa, while in fact the latter comprehends infinitely more both in quality and quantity. Literally the Sanscrit term Akasa means the sky, but occultly the impalpable Ether or the Soul within the Ether. Our most logical authority, The Secret Doctrine, defines it as the immortal spirit, the progenitor of Cosmic life and universal intelligence whose characteristic property is Buddhi. Akasa is the sphere of the pure undifferentiated Monad, the essence of wisdom, while the Astra Light at its opposite pole is the abstract atom of matter, the plane of generation, and the great womb out of which issues all planetary life. Ether, which is the highest vibration of the Astral Light, is but as a vehicle for Akasa, a gross body in comparison. The functions of the Astral Light are as manifold as the expressed universe. Its nature is dual - the highest Ether forming its positive, and the concrete, or differentiated elements, its negative pole. Its cause reaches back to the root of all causes, and its effects involve all our physical and psychical experiences. We deal with its familiar phenomena in every breath and every motion, while the rare and abnormal phases are as strictly subject to its laws. It is not substantially identical with any one of the material elements of Cosmic matter, but is one degree superior to Prakriti (Nature as apprehended by the senses), and it impenetrates and vitalizes each atom. It is itself the one underlying element in which all other known elements have their source and supply. In its physical aspects it includes the Ether of modern scientists, but in the metaphysical sense they scarcely touch its borderland. For while it is the reservoir of Heat, Light, Magnetism, and Electricity - the field of all degrees of vibration - it is also the sphere of all intellectual life, and the ruling agent in the alchemical process which frees the cerebral atom and converts it into thought. Its vibratory rate determines individual mental tendencies, and also establishes our intimate relations in body with the stars. Paracelsus maintains that, as fire passes through an iron stove, so do the stars pass through man with all their properties, and go into him as the rain into the earth, which gives fruit out of that same rain. While the modern spectroscope reveals the identity in substance of infinitesimal man, and the greatest luminiferous body that glides in vast revolutions through space, no instrument has, as yet, been discovered so sensitive as to register the subtle and evanescent fluid which, by its uniform nature, makes astronomical research and thought-transference possible. Keely's Motor has, however, already foreshadowed such a discovery.

The Astral Light is the great record-book upon whose pages every thought and act of differentiated consciousness is engraven, there to be read by the individual who has learnt the secret of exalting his vibrations until they become synchronous with the waves of this finer element. The definition of Memory which has ever been the enigma of science, a function with an inapprehensible cause, is relegated to the domain of the occultist, who may briefly define it as the correlative vibration of the cerebral center with the Astral Light. Within this correlation reside all the possibilities of consciousness from the horizon of Maya (illusion) to the zenith of pure Ether of transcendental life. Madame Blavatsky sweepingly states: The psychic forces, the 'ideomotor' and 'electro-biological powers,' latent thought' and even 'unconscious cerebration' theories can be condensed into two words, the Kabbalistic 'Astral Light.' Quesnes treated of it as a universally diffused fluid permeating all things, and differing in action only according to the mobility of the organism by which it is confined.

The differentiated will to live accompanying each primary monad is the sculptor of the astral images which constitute individual experience. An intense power of concentration makes these images subjective, in which case they are realities only to the operator, but under still more acute and intelligent development, these images may assume a concrete objective form with power of duration proportioned to the original impulse or determined purpose of the projector. Determined will, says a fire-philosopher, is a beginning of all magical operations, and the great magician, Abbe Constant, states: To acquire magical power, two things are necessary: to disengage the will from all servitude and to exercise it in control. He alone can become a Master whose physical and psychical organization is attuned to the Astral Key-note - whose self-consciousness has outgrown the limits of personal slavery, and whose will is so cultivated as to act without fear and without desire - intelligent, determined, self-possessed and confident. While the majority of mankind is occupied in mere negative registration of sense-impressions the occultist classifies these, and admits only those most useful to his purpose. Colonel Olcott refers to the manipulation of the Astral Light in his statement, that the efficacy of all words used as charms and spells lies in what the Aryans call the vach, a certain latent power resident in Akasa. Physically we may describe it as the power to set up certain measured vibrations, not in the grosser atmospheric particles whose undulations beget light, sound, heat and electricity, but in the latent spiritual principle or force - about the nature of which modern science knows almost nothing. As an illustration of this we have the word Aum, which, as all students know, has an equilibrating effect which resists the inroad of passion.

The symbol used to express the astral realm by the mystics of all ages is the serpent, or the fiery dragon. It is stated that long before our globe or even our universe became egg-shaped, a long trail of Cosmic dust (or fine mist) moved and writhed like a serpent in space. This was the beginning of our Eternity, exoterically expressed by a serpent with its tail in its mouth, or in the act of incubating the Mundane Egg with its fiery breathe. The Chaldean Oracles refer to the Astral Light as winding in form, which qualification refers to the vibratory motion that characterizes it. The intense rate of its pulsation may be faintly realized in the rapidity with which successive images are recorded in the dreaming or hypnotized state. An illustration is recounted of a student who was making scientific experiments in this psychological field with a friend. While a drop of water was descending down upon his forehead, he closed his eyes and dreamed that he started out from a harbor upon the wide main, soon passing several beautiful islands full of villages, cities, verdant fields and mountains. The sun was beaming generously, but little by little the sky grew darker and heavier, and drifts of black clouds swept upon the scene. A great gal arose. Consternation and horror pervaded the crew. The waves began to rise higher and higher, until finally the heavens and the sea were blended into one dense chaos.

The crises was at hand. The dreamer suddenly felt as though the whole sea had burst asunder and drawn him in - he awoke just as the drop of water touched his skin, fallen from the hand of his friend and held but a few inches above his head. He had registered the full episode of dramatic changes during the instant in which the sensor nerve flashed its irritation to the brain. Far more remarkable than this are illustrations on record which cover an extent of years and are recorded in an instant of time, experimentally proven to be less than the period required for transmitting a nervous current. The wondrous rapidity with which feeling and thinking may be condensed has also been frequently analyzed in the experience of persons nearly drowned, and as Helmholtz has demonstrated that the period in which a nerve-current may be conducted is a definite one, we have no alternative but to assume that a far subtler element or vehicle is employed in the registering psychic experiences. By the astral current throughout the ayras, around the nerve-centers and nerve-tubes, mind contacts mind, thoughts are flashed in upon us and emotions conveyed. It radiates from the individual man as an impalpable but intelligent aura - the medium of his psychic and intuitional life, by means of which he evinces sympathy and understanding; while in its higher aspect it is the sensitive plane that records Cosmic ideation, and conveys impressions of truth and of universal law to the human mind.

Experiment with trained sensitives has proven that when an intellectual concept is formed, the astral aura photographs it instantaneously, while, when an emotion is evinced, the aura changes in intensity of color, and, on volition becoming manifest, there is a positive increase of vibration. We, therefore, carry with us all we have ever thought and felt - and self is the ever-refining essence of this thought and feeling. From this point of view responsibility assumes enormous proportions, and we perceive why the great Teachers of the world have ever agreed that Man is his own heaven and his own hell. The great mesmerizer, Du Potet, declares that the faculty of directing this fluid is inherent in certain organizations, that it passes through all bodies, and that everything can be used as a conductor - no chemical or physical forces are able to destroy it. Treating thus with the merest initial of this latent power, who can say where its further unfoldment will lead the collective consciousness of Humanity?

In the present age, our science informs us, we perceive in the main only the lowest vibrations and inferior qualities of the Astral Light, because we are in the lower arc of our present cycle, and the energy of the life-wave is at its ebb. Mme. Blavatsky says of Plato's method of expression that he divides the intellectual progress of the universe during every cycle into fertile and barren periods . . . When those circulations, which Eliphas Levi calls 'currents of the Astral Light' in the universal Ether, . . take place in harmony with the divine spirit, our earth and everything pertaining to it enjoys a fertile period. The occult powers of plants, animals, and minerals magically sympathize with its superior natures, and the divine soul of man is in perfect intelligence with these inferior ones. But during the barren periods the latter lose all their magic sympathy, and the spiritual sight of the majority of mankind is so blinded as to lose every notion of the superior power of its own divine spirit. We are in a barren period. Ignoring our latent inherent forces, we drift in negative submission to the lower laws of Nature, suffer deprivation, want of thought, emotion, and volition, while the precious fluid, in mute anticipation, haunts our dreamless sleep, and awaits the dawning of a higher Consciousness.

Keely's Secrets

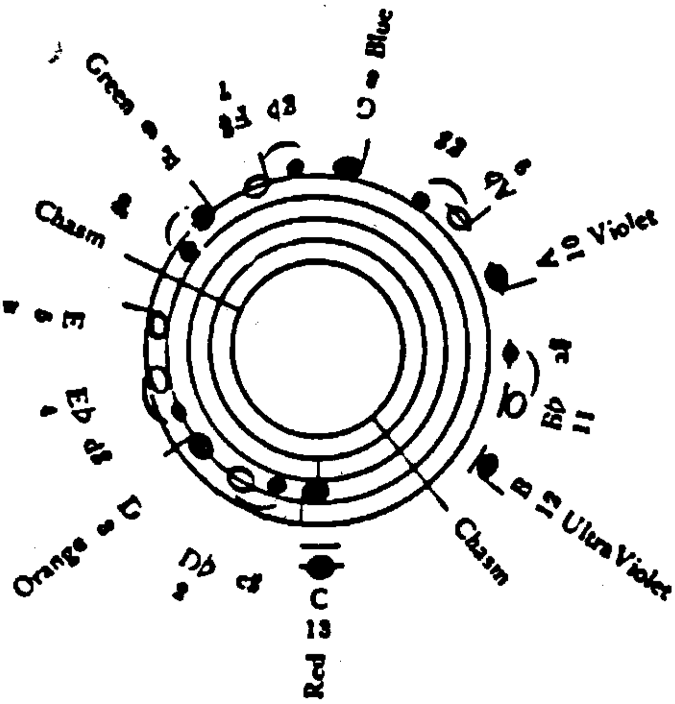
Green 6 F
F# 7 Gb
G 8 Blue
G# 9 Ab
A 10 Violet
A# Bb 11
B 12 Ultra Violet
Chasm
C 13 Red
C# Db 2
D 3 Orange
D# Eb 4
E 5 Yellow
Chasm

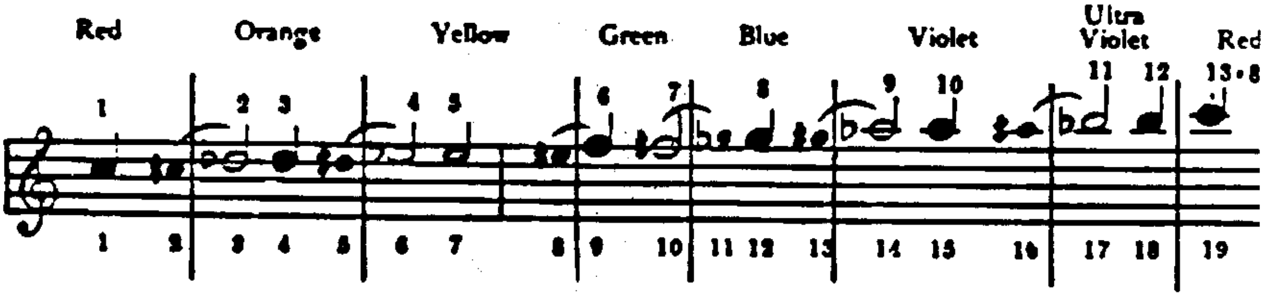
Red
Orange
Yellow
Green
Blue
Violet
Ultra Violet
Red
1
2 3
4 5
6 7
8
9 10
11 12
13•8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Mr. Plum's Visit to Keely's Laboratory

(sometime in late 1893)
from Dashed Against the Rock

"I have seen a spectacle I would have pronounced impossible according to all accepted theories of physics with which I am familiar. Without apparent exhibition of heat, electricity, or any other form of energy hitherto operated by man, I have seen a strong metallic wheel, weighing seventy-two pounds, in swift and steady revolution by the hour, and absolutely without cost. It is but a subsidiary engine, made and used simply to help equip with similar mysterious capacity of movement the large commercial engine by its side. And that is a most strange and complex mechanism, which perhaps no one but the inventor can even understand at present, and which, but for too frequent previous unauthorized fixing of dates, might be said to give promise of being itself in motion very soon. What is 'very soon' in such an undertaking? Another thirty years of patient, lonely plodding on this labyrinth path would be nothing, if then this explorer could reach his goal. How long after Franklin's kite did the world wait, and how many hundred great experimenters, before a dynamo engine kindled our lamps and whirred our wheels? Yet this solitary pioneer, grown half blind by groping in these dim intricacies so long, again and again hurled aside, broken and almost dieing by the terrific force with which he is seeking to cope, is met with the sneer, 'Why don't you do something?' He has done much, of ignorant, senseless, and cruel abuse. His immortality, however, is sure. For the world at length honors an honorable purpose, persistently pursued in a high undertaking. And he has already so enlarged the domain of human knowledge, he has lifted man into such a new world of facts, the truths his experiments unveil are so novel, suggestive, and inspiring, that whether all this is ever turned to practical account or not, his name will never die. But if he should turn out to be a prophet, if he is a seer, and does really discern a promised land of lightened toil into which mankind will eventually enter, even though he may not live to lead them in, then the world will gratefully build his tomb.

But the world asks, who is the witness that testifies so boldly to these surprising things? Is he competent and worthy of trust? The witness is not a capitalist, and he has no relations with investors, and is free to say that if Keely were to die tomorrow, it might be a hundred years before another mind would arise able to complete his work; if indeed, it is capable of being completed at all, which no one at present knows. Impelled by a life-long interest in the wonders of natural science, and honored by the personal friendship of Keely and a few of his advisers, I have followed the course of this investigator for years with the intensest interest and sincerest admiration. I spent more of my vacation this season in the Philadelphia laboratory, and saw greater wonders there, than in the Chicago Fair.

In whose judgement greater? Is a layman in physical science competent to judge in such matters? Confessedly not, on some questions. To most men the learned physicists speak an unknown tongue. Too profound for the common apprehension are the mathematical formulas, even, with which their works abound, though their theories and arguments are full of interest. And many would confess also that they can no more understand the ground of Keely's assertions concerning the number of millions of oscillations taking place in a given substance each second, nor his fluent discourse upon clustered thirds and introductory ninths, upon nodal transmitters and neutral centers, and upon streams and waves of polar and depolar influence. On these declarations this witness has no testimony to offer. In electrical science the world gladly accepts the terminology and the philosophy by which the specialists creditably seek to gain some practical apprehension of the elusive mystery with which they deal; elusive, for through all their technical terms and fine-spun theories, the futility of their endeavor to gain any exhaustive comprehension of it plainly appears. Experts have their field, but as Mr. Gladstone says of the Hebraist and the scientist in reference to the higher criticism and the scriptural cosmogony, 'their title to speak with authority is confined to their special province, nor are they inerrable there; and if we allow them to go beyond it, and still to claim their authority, when they are what is called at school "out of bounds," we are much to blame, and may suffer for our carelessness.' 'My contention is,' he says, 'that there is a ground which the specialist is not entitled to occupy in his character as a specialist, and on which he has no warrant for entering, except in so far as he is a just observer and reasoner in a much wider field.'

It is into this wider field of fact, where any can go whose general training fits him to be in any wise 'a just observer and reasoner,' that this witness deems it not improper to enter, especially as he follows in the wake of not a few who rank high as experts in mechanical engineering, in chemistry, in electricity, and other departments of superior culture. For, not only has Keely's legal counsellor, Charles B. Collier, an experienced patent lawyer, acute, cultured, and discerning, given him from the first his sincere and hearty support, but numbers of other men of honorable character and position, many of them eminent for scientific attainments, have given their unqualified testimony that Keely is an original and able investigator in an interesting and promising, though wholly novel, field - a wonder-worker, whose work seems to overturn certain accepted theories, and has puzzled and baffled their learned advocates. Yet, partly, perhaps, because Keely is not in the fraternity of college bred men, but has educated himself (though his writings show familiarity with scholarly works), partly because his claims are so astonishing and his methods so incomprehensible, and partly because of premature predictions of a practical issue of his labors, and because also of unfortunate differences reported in respect to the business side of his enterprise, there are comparatively few men of public prominence who seem to be willing to be known as believers in the importance of his investigations, or even in the integrity of the man. At any rate, ridicule and contempt continue to be thrown at him and at the faithful friends who have long and nobly stood by him. Only lately a prominent journal intimated that 'an interruption of Keely's personal freedom' ought to result from what it calls his 'gigantic jugglery.' It is these unworthy flings, together with a sense of the public importance of the whole matter, which have prompted my voluntary and unsolicited testimony in the interest of truth.

For though scores of assemblies, comprising learned scientists, skillful engineers, and men of large success in the practical conduct of affairs, have witnessed various experiments by Keely during the past dozen years, and although their clear and positive statements of the interest and value of his researches have been repeatedly published in leading newspapers, with the names and professional titles of the witnesses given, yet the general public appears either to overlook or forget all these testimonies, and to be rudely impatient of every undertaking that does not immediately issue in commercial success. Seldom does any public journal refer to Keely in terms of appreciation and respect. As his labors have now reached some new results which only a few persons have witnessed, this further testimony is offered as information upon a matter of scientific interest, certainly, and with a possible bearing upon industrial advance.

What, then, is the testimony that the present witness has to give? After some ten years of acquaintance with Keely, and after personally seeing many of his experiments, 'witness deposeth and of one of the most beneficent revolutions in the history of human progress; a man of wonderful insight and truly amazing fertility of inventive genius in overcoming obstacles and in contriving appliances for attaining his mechanical ends; that he is dealing with and trying to employ in practical mechanics a force absolutely new among all the forces hitherto handled by man, although its presence in nature is affirmed by the theories of scientists, and demonstrated by various observed phenomena; a force of mysterious and awful energy, boundless in extent, and literally costless as the air. Electricity is subtle and powerful and illimitable in supply, but it requires constant and costly expenditure of energy to call it into exercise and keep it at work. This new force, beyond the curious and complicated mechanism which this wonderful wizard has contrived for it to employ, the harness he has fashioned for it to wear, seems to require but a few slight musical sounds, the sonorous vibrations of certain metallic appliances, to set it in motion, and then it will keep in motion - for all that at present appears, in steady, noiseless, and almost resistless motion - till the solid metals of which it is composed wear out.

What! one and all exclaim, is the absurdity of perpetual motion to be revived again? But the physicists tell us there is perpetual motion all around us in nature, intense and all-prevading, and always has been, since the hour 'when the morning stars sang together, and all the sons of God shouted for joy.' Here we touch the robe of the Infinite One, who 'upholdeth all things by the word of his power.' Of him the Unerring One declared, 'My Father worketh hitherto and I work.' Aye, works unceasingly now, in the incessant and intense molecular vibration all the time going on in all matter; in the solid table by which we sit, in the firm granite of the building which encloses us. Action, motion in everything, by everything, everywhere, all the time, and swift, more nimble-footed sometimes than thought almost, but with such a soft and easy pace that no footfall is heard, no movement discerned save as we take observation by the distant heavenly orbs among which we all here on the earth are traveling, hurled along our pathway over a thousand miles a minute. Movement of everything from here to there, and movement in everything while here or there. And so harmonious is the movement, on such delicate anti-friction cushions do the bearings rest, that it is all inaudible, save to that One alone whose ear discerns the music of the spheres - the spheres immensely great and infinitesimally small -

'Forever singing as they shine,
The hand that made us is divine.'

And only now, after thousands of years of unheard song, this great magician arises and strikes the chord of sympathy to which this vibrating force responds, and lo! it comes forth from its secret chambers like the mighty Genius unloosed by the Arabian fishermen from the copper flask, and waits on man to do his bidding, bending its tough sinews and plying its facile fingers to perform his humblest tasks.

And what proof can there be that this dream of poetry and fancy of story is in any degree an accomplished fact? Look and see. Here is a wooden table, sometimes covered by a heavy slab of glass. Standing on the glass or on the wood, and capable of being moved freely upon it, is a metal standard say a foot high, bearing a copper globe about a foot in diameter. Around the base of the standard project horizontally numbers of small metal rods a few inches long, of different sizes and lengths, vibrating like tuning forks when twanged by the fingers. In the hollow globe is a Chladni plate and various metal tubes, the relation which can be altered by turning a projection like a door-knob, on the outside of the globe, at the outer end of a small shaft, round and round to the right or left. This construction is called a 'sympathetic transmitter.' Some two or three feet distant on the table stands a movable metallic cylindrical case, some six inches by eight in size, composed of certain metal resonating tubes, and certain other metal fixtures. You take it all apart and see there is no magnet there. You place on top of this cylinder a small pocket compass, a brass cup of two inches in diameter with its glass face. The six inches by eight in size, composed of certain metal resonating tubes, and certain other metal fixtures. You take it all apart and see there is no magnet there. You place on top of this cylinder a small pocket compass, a brass cup of two inches in diameter with its glass face. The needle points north. From the periphery of the globe of the 'sympathetic transmitter' extends a wire of the size of a common knitting-needle, made of gold and silver and of platinum. The free end of this wire is now attached to the cylinder. The needle is still true to the pole. Then the vibrating rods are twanged, the knob is turned, and on a rude harmonicon trumpet for a moment or two certain sounds are made, when lo! the needle is invisible, it is whirling on its pivots fast. The operator talks of the variant length of waves and of a continuous stream, and in some instances it is half a minute, sometimes three minutes, before the needle comes to rest, and it has kept in swift revolution for many hours; but when it pauses it points no longer to the north, but to a particular part of the mechanism. You leave it there, and are busy with other wonders for an hour or so. Returning, you find the needle still points to its new master. You lift the compass off, and at once it resumes its normal position. You slowly lower it towards the silent cylinder, and when within an inch or two it obeys the new impulse again, and points as before. So also it veers from the north when you carry it near the knob of the copper globe. As Gladstone says, 'Our hands can lay hold of truths that our arms cannot embrace,' and though it takes a physicist to comprehend this miracle, any careful observer can apprehend it, and, after seeing it repeated many times, if he is measurably well read, is competent to testify that here is a new, subtle, silent, continuous influence, and that it is called into exercise in connection with certain brief musical sounds.

Look again. On this rude harmonicon trumpet this magician blows through a small window into the next room towards a common zither some ten feet distant, held upright on a table by a small standard composed of a group of metal tubes. The two musical instruments have been carefully attuned to each other. Attached to the back of the zither is a common silk thread loosely hanging and extending some eight feet away, where it is tied to a movable framework of half-inch iron rods, supporting and bracing in position, on an isolated table of glass, a metal globe, fifteen inches in diameter, capable of turning freely in either direction, on its axis, which bears inside the globe certain resonant tubes and plates, the table standing at an angle of 45 degrees from the face of the zither. Louder sounds the horn, till in a minute or two the metal globe begins to revolve. The horn stops, the globe stops. Again the horn resounds, again the globe turns, and the stronger and more continuous the blast, the more swiftly whirls the globe. You snip the thread apart with your scissors, and the ear of the globe has grown dull; no sound can awake it to motion again. Does a man need to be an expert in physics after he has seen that marvel repeated a few times, and has moved all that apparatus freely hither and thither, to testify that the rapid revolution of that metal globe was not caused by compressed air, coming in concealed tubes from a hidden reservoir, or that a silk thread is not the highway usually cast up for electricity to travel?

But these are philosophical toys. What about an engine with power to help human toil?

I have in my study a paper weight - a disc, said to be composed of an alloy of three metals. It looks like steel, measures two and a half inches by three-quarters of an inch, weighs about a pound, is enclosed in a brass ring, and exhibits no magnetic power. I am told that shut up in a glass chamber and connected with the wire which seemed to affect the compass, it absorbed some seven pints of hydrogen gas. The story runs that it was also rapidly whirled by a steam engine a certain number of hours, still in connection with the apparatus from which seemed to flow that subtle influence which the needle of the compass obeyed. Whatever may be thought of this, it is a fact that the disc thus 'vitalized in its atomic or molecular constitution' adheres to the under side of a certain metallic resonant structure as if held there by magnetic attraction, and also supports a weight hung to itself of over two hundred and thirty pounds. Dissociated from the peculiar vibrating apparatus, it falls like any other heavy body, and though that apparatus attracts the disc, even with the attached weights, it is incapable of attracting anything else; it will not support the smallest iron filing. Here then is a strong pulling power in exercise in certain circumstance when two bodies are in contact. Can it pull bodies together which are not in contact?

I see before me on a table a glass jar, ten inches in diameter and forty-eight inches high, filled with water. At the bottom lie three metal balls like the one I hold in my hand, which weighs about two pounds. The jar has a metal cap to which is attached the gold and platinum wire reaching from the copper globe. I am told each ball, like every mass of matter, has its peculiar musical chord. I am reminded of well-known facts of sympathetic vibration; *e.g.*, a large mill trembling in response to the note of a neighboring waterfall, and only quieted and rendered safe by building on an addition, changing its musical chord. And now again the rods are twanged, the knob is turned, the trumpet sounds and keeps sounding till, in a moment or two, I see one ball begin to sway from right to left, then slowly leave the bottom of the jar and rise through the water till with a bump it strikes the metal cap, rebounds a few inches and comes to rest in contact with it on the surface of the water. Still the horn blows, and by this time the second ball responds in like manner, and then the third. Then the music ceases, and we turn to other experiments, but as long as I stayed in the shop that day something made that metal swim. My companion said he had often seen the weights brought slowly down, or held midway, as shown by photographs, by sounding other chords. On the top of the jar lay certain pieces of metal. Keely said, 'Do not remove those. I once did that, and crash went the balls through the bottom of my jar.' Now here was a pulling power acting at a distance of four feet, not capable of lifting the weights through the air, but before all eyes lifting them through water. Can this pulling power turn a wheel?

Here is a wheel of stout metal weighting, as stated, seventy-two pounds, free to move either way on its stationary axis. Its hub is a cylinder containing certain resonant tubes parallel to the axis. It has eight spokes, each carrying one of the 'vitalized discs' at its outer end, the face of the disc at right angles with the spoke. There is no rim to the wheel, but there is a stationary metal rim some six inches wide and thirty-two inches in diameter, within which the wheel turns without touching it. This rim carries on its inner surface nine similar discs, and on the outside, attached to each disc, a resonating cylinder. The requisite amount of the metallic volume of this cylinder is obtained by enclosing in its tubes a few cambric needles, more or less as required, and curiously enough, some of these needles at length become magnetic. Attached to this engine is a gold and platinum wire, some ten feet in length, running through the small window to the copper globe in the other room, where sits the man who has fashioned all this. He twangs the rods of the sympathetic transmitter on the table at his side, he turns its knob, the musical instruments sound for a moment, and peering through the window along the line of the wire his face lights up with a smile of triumph. He settles back in his chair, and all is still. That wheel at the end of the wire is in rapid revolution before your eyes. You turn and look with amazement upon Orpheus returned to earth again and outdoing his fabled exploits of old. For by the enchantment of the subtle harmonies he evokes, too fine for human ear to catch, you see the untamed forces of nature obey his behest; that most constant of all things, the magnetic needle, you see charmed into fickleness by his magic spell; you see balls of iron swim; you see insensate matter - as you thought it, but sensitive now to his call - leap forward into instant rotation, continuous and swift. Long we stand around that flying wheel. The friend who photographed it at rest again levels his camera upon it. In vain; its spokes cannot tarry long enough to be caught by his snare. It is still as death, and almost as mysterious. We listen to long dissertations upon the reason for the relative position of the eight discs on the wheel and the nine on the stationary rim, and how the adjustment can be so altered that, instead of a revolution, there will be a violent oscillation back and forth. We are shown the corresponding wheel and the rim of the large engine close by, which is to bear the discs not singly, but in groups, the steel resonating drums with their circles of tubes inside, and thirty-five inch Chladni plate underneath the 'sympathetic transmitter' on top; the extra wheel bearing on its spokes cylindrical cases, each filled solid with a hundred thin-carved plates of steel, to get the utmost superficial area, we are told, and it is all so utterly beyond comprehension, that we can see no reason why it should have been made as it is, or how any one can be sure it will ever run. But we turn around and look again on that noiseless wheel, still running rapidly all alone, and confess we should have said the same thing about that. And we are inclined then to trust the word of the inventor when he says the running of the smaller insures the running of the larger; that the wheel you see spinning so fast cannot be stopped by any force except one that would tear it into fragments, unless with thumb and finger you loosen that golden wire along which 'the stream of sympathetic vibration' is said to flow, and that there is no reason why the wheel should not keep in motion till the bearings wear out.

I say nothing now of the wonders of which other witnesses can speak, and which are said to have appeared in the slow progress this incomprehensible man has been making all these years; of a pressure obtained from the disintegration of water by vibration of twenty thousand pounds to the square inch; of a slowly revolving drum which went no slower when winding tightly upon itself a stout inch and a half rope fastened to a beam, and no faster when the rope parted under the strain; of the disintegration of rock into impalpable powder; of raising heavy weights by aid of a 'vibratory lift,' recalling the 'negative gravity' of our modern story-teller.

The engine you have been looking upon requires as part of itself for some mysterious purpose certain heavy tubular copper rings. Skillful artisans failed in various endeavors, by electrical deposit and otherwise, to make them right. The inventor contrived machinery for bending into semicircles sections of copper tube, one and a half inch bore, three-eights of an inch thick, forcing a steel ball through them to keep the tube in shape. To make a ring, he placed two of those half-circles together and joined the ends in some way (without heat), by what he calls sympathetic attraction, so the resonant properties of the ring are satisfactory, and though you see the line of union, the two parts cannot be severed. You see one of these rings, some fifteen inches in diameter, hanging by block and tackle from the ceiling, and lashed to the lower half swings a big iron ball weighing five hundred and fifty pounds, and there it has swung for weeks. Has the man who has done simply that, and done it merely to furnish a subsidiary adjunct to his main contrivance, won no place among the great artificers? Is it worthy business to revile him as a swindling charlatan? The end is not yet. We shall see what we shall see, or some one will. One thing, however, we see clearly now, and that is that John Worrell Keely deserves the esteem and admiration of his fellow-men. Who does not hope that he has solid grounds for the persistent belief which has been his star of hope these many years; that a merciful Providence is about to confer a new boon upon the suffering industries of mankind; that the time at length has come when man is wise enough to fashion and strong enough to handle the beneficent gift of a costless motor to ease the burdens of human toil?

Wise enough and strong enough, perhaps, some may say, but is man trusty? For the question has arisen whether a force of such fearful energy as some of these experiments disclose can safely be entrusted to such a being as man, who can destroy as well as build. But why should man have been set to discover and harness it? 'I take great comfort in God,' said James Russell

Lowell, in one of his recently published letters, 'I think. ... He would not let us get at the match-box as carelessly as he does unless he knew that the frame of his universe was fireproof.'

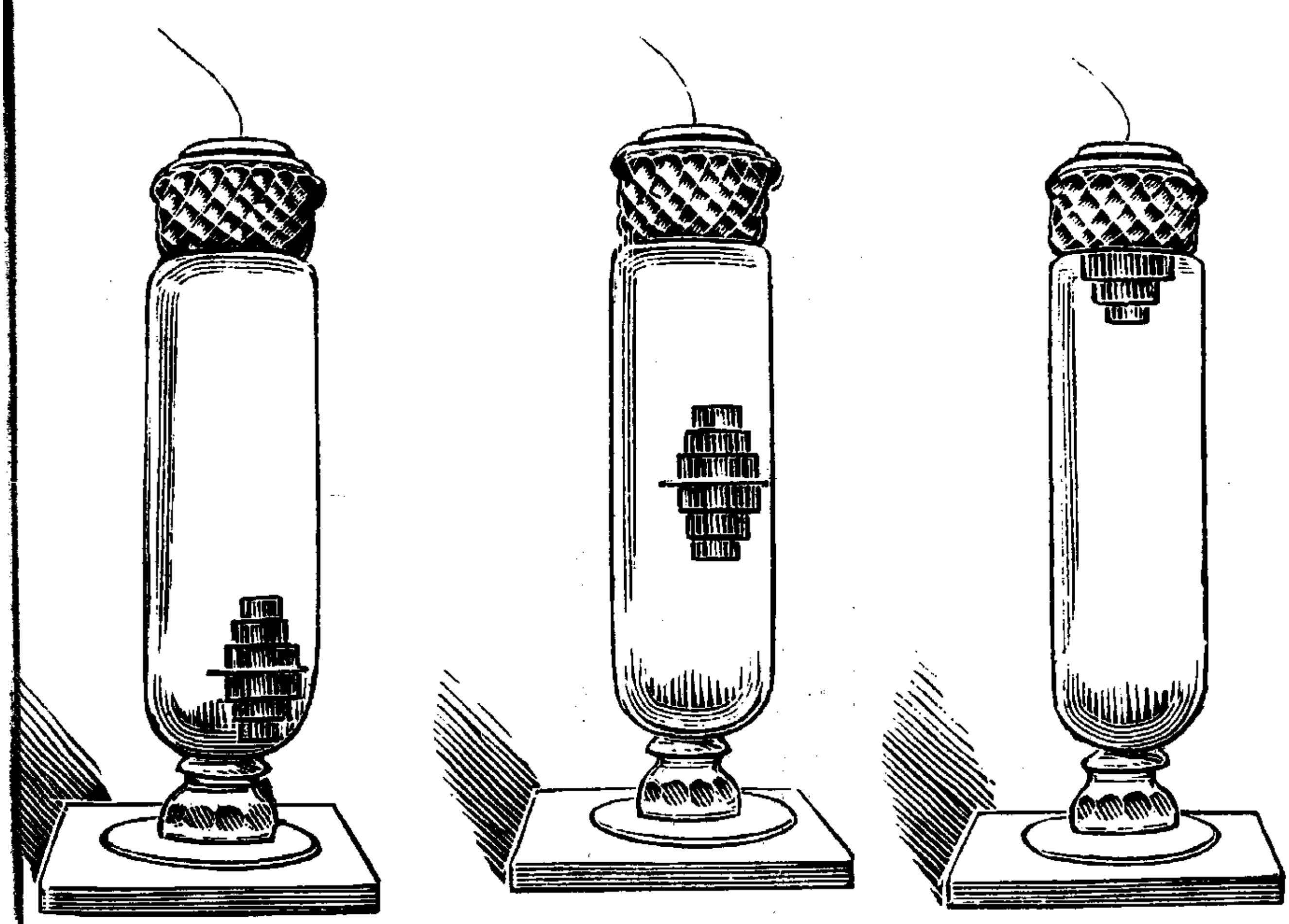

GRAVITY OVERCOME

XPERIMENTAL MODELS FOR RESEARCHING.

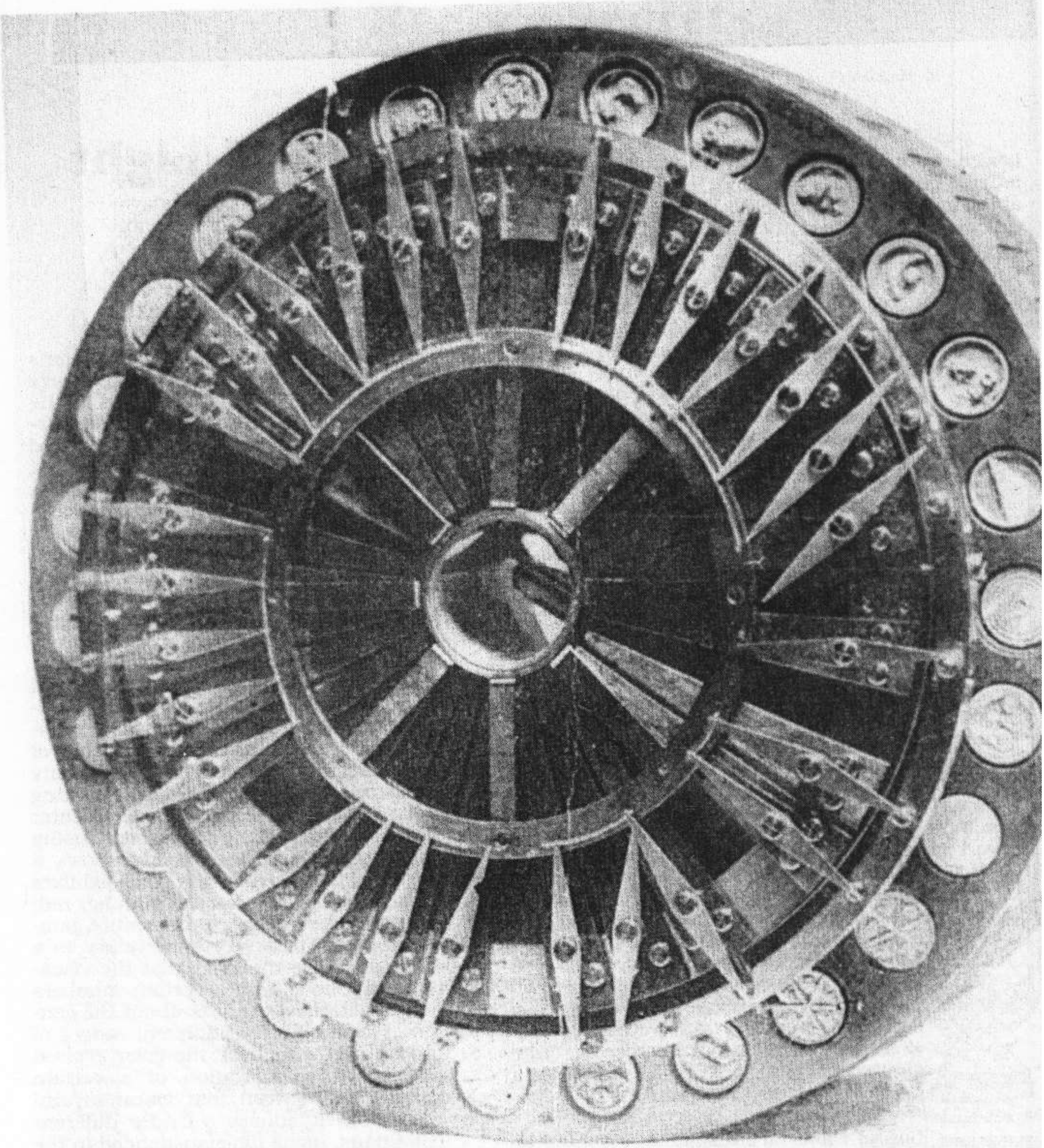

Photo by Collier.

KEELY'S VIBRATORY SWITCH.

No one understands it but himself.

Pyramids, Telescopes and Light

John Ernst Worrell Keely
from Dashed Against the Rock

A building, pyramidal in shape, two hundred feet high, one hundred feet at the base, and having at the apex a disc with a minute aperture in its center, and a triple combination of reflectors, which must concentrate upon one center which must be focalized upon the minute aperture in the disc, - the image being received at the base of the pyramid upon a white surface prepared to receive it, - would yield results beyond the dreams of the most sanguine astronomer of the present day. The distinctness of the image taken would be the most wonderful part of the phenomenon, and the size of the magnification would be limited only to the diameter of the base of the pyramid. This probably explains the great Egyptian pyramid, with its circular opening through its center and cavity at the bottom. The apparatus from its summit has probably been destroyed during some of the many ravages to which the land of strange vicissitudes has so frequently been subjected. The Great Pyramid proves that, among the ancient Egyptians, the knowledge of astronomy was amazingly great, although they need not necessarily have known of the modern telescope to have obtained it. The almost impossible labor of making lenses is done away with in this system, which embraces also a microscope on a similar principle without lenses, far superior to anything now in use.

Vibrations necessary to the production of colors are very high, reaching to millions of vibrations per second. The true relation between the vibration and the color can only be a harmonic one, as colors commence in the millions and end not lower than in trillions of vibrations per second. The conditions essential of this class of phenomena are, first, a condition relating to the structure in which they take place, free from all extraneous vibrations; *the experimenter's presence even influences to a great extent the motions of the molecules* used in the experiments. A bath is employed, arranged so that light rays can be projected upon it at certain angles, and from that upon a screen. The bath must have what are termed centers of association and concentration, and suspended across its surface in such a manner that the centers approach very near the surface of the liquid, which is preferably water, whose surface must be free from all films. Upon this is dropped a single drop of naptha dissolved in ether, which constantly spreads over the surface and gradually evolves most beautiful variations of color. The vibrations are transmitted from an instrument capable of producing vibrations of varying intensity and pitch, across the line of resonating centers, nine in number. The first center shows a *light* straw color, condensing about the center; the second shows a *very dark* straw color; the third exhibits an orange red; the fourth, a bluish red; the fifth is largely green. The entire gamut ranges from a very light yellow to a dark blue. The theory is that the vibrations produced gather certain numbers of the molecules together about the centers, which represent different ranges of motion, and which by the color evolved affords a demonstration of a certain connection between that vibration and the colors it educes. Under different conditions, using films suspended in the air, and light passing through these films under polarization, (polarized light) the vibrations influence the molecules in the film, producing again exhibitions of varying colors when projected upon a screen. Experiments of this kind are most difficult, on account of the unstable conditions found everywhere.

Keely's Secrets

DIAGRAM 1.

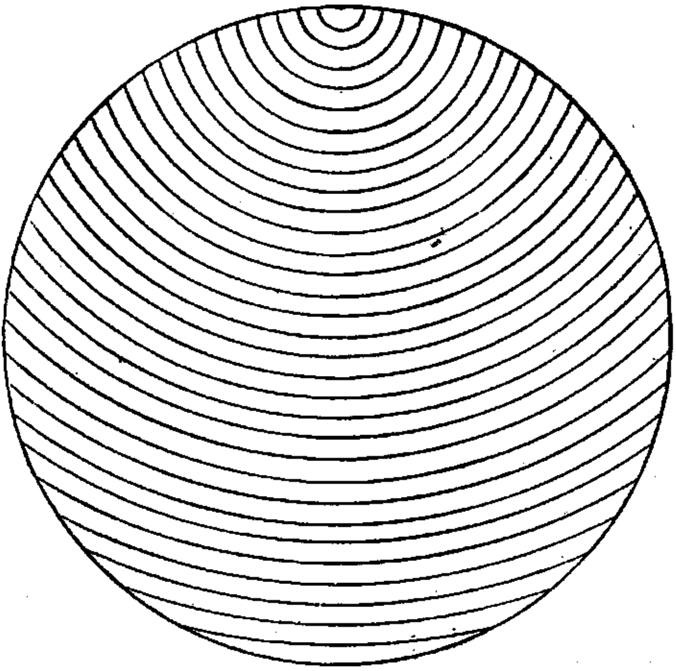

External View of Atom.

DIAGRAM 2.

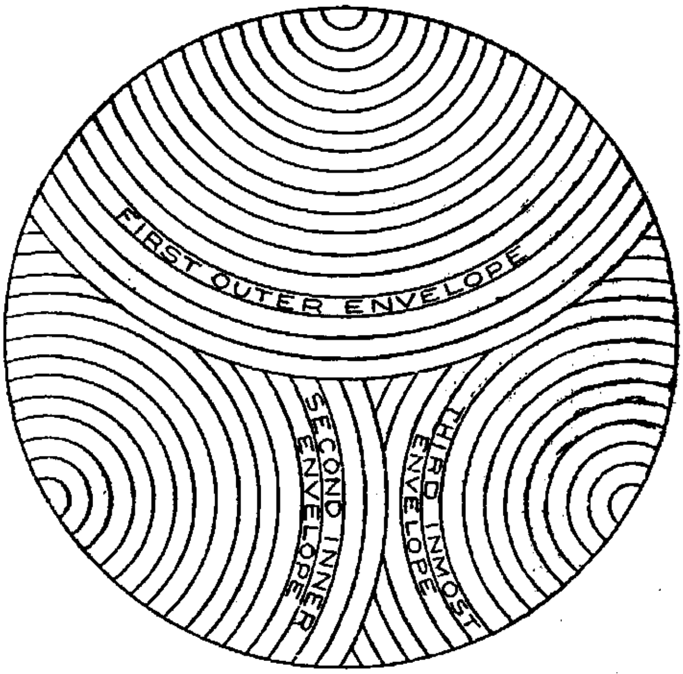
FIRST OUTER ENVELOPE
SECOND INNER ENVELOPE
THIRD INMOST ENVELOPE

Showing Sections of the Three Envelopes.

DIAGRAM 3.

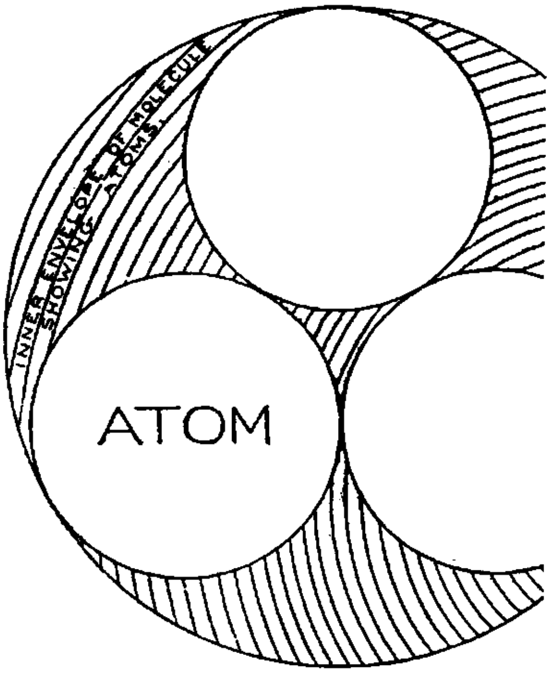
INNER ENVELOPE OF MOLECULE SHOWING ATOMS.
ATOM

Showing Positions of Atoms in Inner Envel

DIAGRAM 4.

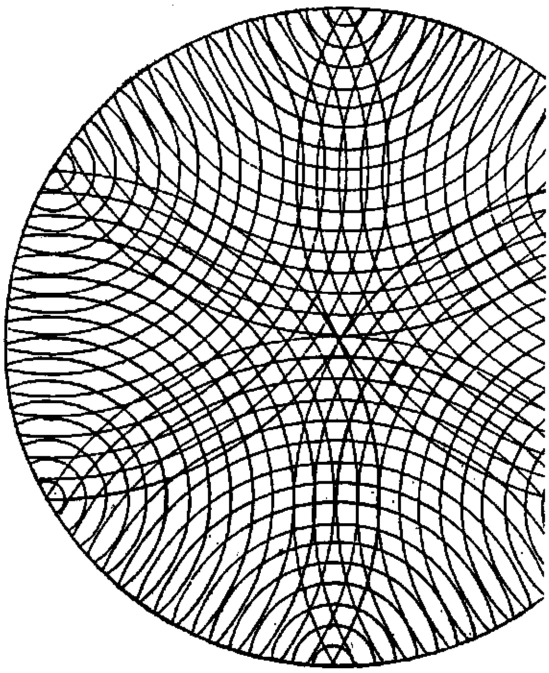

Showing Lines of Interference of Triple Envel

Levitation

John Ernst Worrell Keely
from Dashed Against the Rock

A small instrument, having three gyroscopes as a principal part of its construction, is used to demonstrate the facts of aerial navigation. These gyroscopes are attached to a heavy, inert mass of metal, weighing about one ton. The other part of the apparatus consists of tubes, enclosed in as small a space as possible, being clustered in a circle. These tubes, represent certain chords, which were coincident to the streams of force acting upon the planet, focalizing and defocalizing upon its neutral center. The action upon the molecular structure of the mass lifted was based upon the fact that each molecule in the mass possessed a north and south pole, - more strictly speaking, a positive and negative pole, - situated through the center, formed by the three atoms which compose it. No matter which way the mass of metal is turned, the poles of the molecule point undeviatingly to the polar center of the earth, acting almost exactly as the dip-needle when uninfluenced by extraneous conditions, electrical and otherwise. The rotation of the discs of the gyroscopes produces an action upon the molecules of the mass to be lifted, reversing their poles, causing repulsion from the earth in the same way as like poles of a magnet repel each other. This repulsion can be diminished and increased according as the mechanical conditions are operated. By operating the three discs, starting them at full speed, then touching two of them, so as to bring them, according to the tone they represented by their rotation, to a certain vibratory ratio, the weight then slowly sways from side to side leaves the floor, rising several feet in the air, remaining in that position, and as the discs gradually decrease their speed of rotation the weight sinks to the floor, settling down as lightly as thistle-down. Where one molecule can be lifted, there need be no limit as to the number in a structure that may be operated upon as easily as one. The vessel in contemplation, the aerial navigator, will be over two hundred feet long, over sixty feet in diameter, tapering at both ends to a point, made of polished steel, and will be capable of being driven under the power of depolar repulsion, at the rate of three hundred miles an hour. It can be far more easily controlled than any instrument now in use for any phase of transit. Another very remarkable feature connected with this system of aerial navigation, is that the vessel is not buoyed up or floated in the air through the medium of the air, so that if there were no atmosphere it would float just as readily; hence, under mechanical conditions most certainly capable of production, involving massive strength of resistance to interstellar vacuity this can be made capable of navigating even the remote depth of space, positions between planets where polarity changes being controlled by other adjuncts of concentration for that purpose.

Safely enclosed within this structure, a man possessing the chemical knowledge these new laws give, with sufficient supply of material from which to make oxygen, by the enormously increased rate of speed attained by such navigator where atmospheric friction is avoided, the time occupied in traveling from one planet to another would be amazingly brief, and one can travel to other planets in this system of worlds as easily as the same ship could navigate the depths of the ocean.

The great obstacle hitherto preventing the solution of this problem has been the strength of structure needed under conditions above presented. With this knowledge of matter, the size of structure is unimportant; the heaviest can be as easily controlled as the lightest.

Chart showing the triple condition of the vibratory flow
governing the molecule during disruption. 1st order of Vibratio

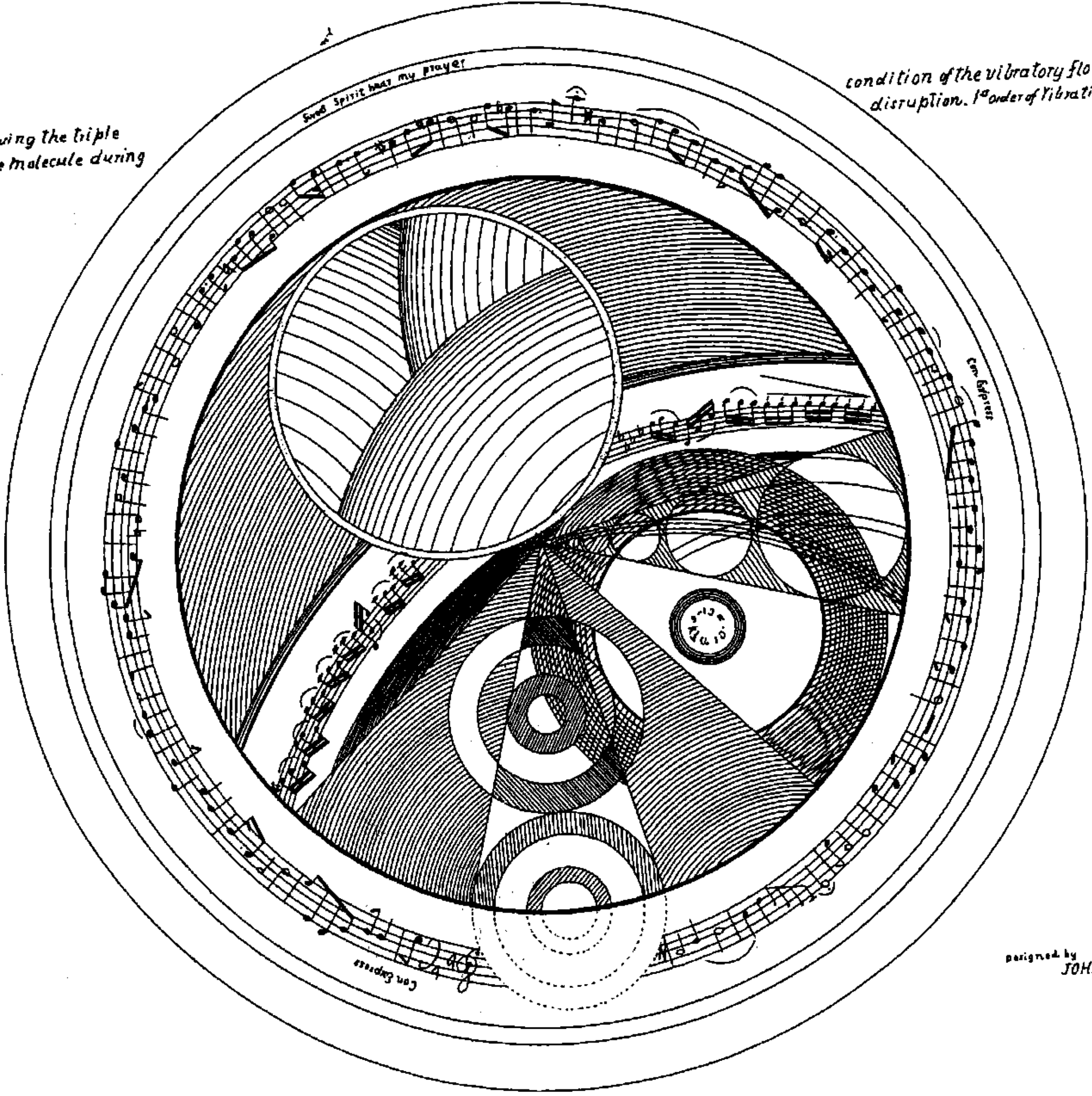
Sweet Spirit hear my prayer
Con Espress
Con Espress

Designed by
JOHN
Ja

Amplitude of Force

John Ernst Worrell Keely
from Dashed Against the Rock

The amplitude of vibrations is directly increased or diminished by increasing or diminishing the size or number of creative aggregates.

The human EGO, subject to the forces of love or hate, kindness or cruelty, forgiveness or revenge, is according to circumstances ruled or dominated by these forces in proportion to their intensity.

The intensity of a force is precisely proportionate to the number of units vibrating at that particular pitch. For instance, let fear assail one man, and according to its intensity will be the effect; but let a crowd of men experience fear, see the result in the augmentation of fear, though its source be relatively insignificant. A curtain in a theater, for example, takes fire; one or two persons, cowardly at heart, become afraid through the dominance of the purely animal instinct of bodily preservation; there is actually no real danger, but these two or three persons are sufficient to arouse the unreasoning dread which lies latent in every breast, with perhaps a very few remarkable exceptions. The fire burns nobody; but blind fear, which is extremely contagious among people mutually sympathetic, by reason of the rapidity with which etheric waves transmit all feeling, occasions a terrible panic, during which many severe accidents and many instances of fierce cruelty occur, all because of this sympathetic transfer of feeling starting from one or two augmented or intensified fear-centers, each person being a center emanating the feeling of fear. Were there no counteracting centers of influence in an audience, radiating contrary feelings, the result of a panic would be the total bodily extinction of a very large percentage of the assembled multitude.

Thus the human race is immersed in forces whose intensity is vast in proportion to the number of EGOS adding each its quota to the already intense vibration, tending either to love or hate, kindness or cruelty, timidity or bravery. Those who intensify the force of cruelty in the place where they reside, may be strengthening a murder's hand to strike the deadly blow in a distant land. This result is brought about through the agency of etheric waves, which transmit forces with undiminished intensity even to uncalculated distances. This phenomenon may be termed *transympathetic*.

They who feel that force called *love*, which on higher planes is known as *sympathy*, thrill with waves of force which are already strong, augmenting them or increasing their intensity. They who indulge such sentiments and encourage such forces may stop the falling hand on evil sped.

In order to protect ourselves effectually from becoming the dispensers or propagators of deadly force, we must consciously and deliberately relate ourselves by resolute determination, to awaken within us such centers only as are concordantly sympathetic with all force radiating in the interest of universal goodwill, thereby aiding the establishment of universal brotherhood.

All ye who feel a longing for a better life or nobler existence draw to yourselves streams of force which they alone feel who have attuned their bodies to the higher harmonies. For a moment you feel as they who dwell perpetually in communion with higher harmonies, living immersed in that higher force; they are the true heirophants, and you, O neophytes, struggling to attain the goal which they have reached, do not despair though at present you find yourselves unable to maintain this high altitude for long together. Though you fall many times, be not discouraged, for as yet your organisms and all their centers of resonation are not yet concordant to the focalized vibrations of the higher harmonies; being still related to the mass, you are drawn again and again into the whirlpool of the vibrations which affect the mass, for these you cannot resist. But know that you can change all this rapidly or slowly as your purpose is steadily intense or vacillating.

With the cessation of your lower desires comes the cessation of the action upon you of the lower forces; the resonating centers which formerly distributed this force, no longer active, become latent and are absorbed back to an embryonic condition.

Let him realize fully the transient character of his own personality and contrast therewith his eager longing to know the immortal.

Every man contains, developed or embryonic, all conditions of the Infinite; therefore no height is too great to reach. Impossibility is a meaningless word to the man who apprehends the fathomless contents of his own nature. Thou comest here, O man, with the instrument thou hast graduated in thy many past existences; how few of thy chord-settings, if thou art numbered with the many, respond to the higher harmonies! Universal unity or fraternity has been absorbed to almost embryonic conditions by the prevalence of material self-regard. Charity has been rendered almost latent, that beautiful chord-setting found even in the lowest forms of creation - LOVE, the dominant chord of the cycles. Love has an amplitude of action in the brute which may well make the selfish man ashamed, but until the crust of selfishness is broken through, the beauty of love is obscured, and though it exists all about him, the poor blind egotist has no eye to discern it. The centers of love, brotherhood, charity, voice their music loud and clear, yet the masses will not listen. I do not mean the immortal EGO when I say man will not listen; I refer to the personality which is the resultant of all the ages of action in this, now rapidly closing, cycle.

You who exist to-day, to-morrow would exist no longer in your present personalities did you but dare to yield to these higher harmonies. I say YIELD because it is a *yielding* process for this personality. In a moment your outer life would end, and you, the warrior, would enter peace.

The immortal EGO is an entity of which man can become thoroughly conscious while here on earth, but to arrive at this consciousness necessitates the entire abandonment of all the petty considerations involved in the transient and subordinate EGO, which is the only self of which the unenlightened man is conscious. Let him who desires to reach this inner consciousness enter his inner sanctuary, wherever that sanctuary may be; it matters not whether it be his own chamber, the open field, the mountain top, the seashore, the stately cathedral, or the humble village chapel. Let him realize fully the transient character of his own personality and contrast therewith his eager longing to know the immortal. Let him concentrate his whole consciousness upon his personality, fully arousing all his personal conditions as a distinct individual; then with all the aspiration of which this personality is capable, let him beseech of the immortal EGO - which is eternal and does not incarnate, but overshadows all incarnations, waiting until one is formed capable of illumination, to whom it may reveal itself - to consider him worthy of illumination, and according to his preparedness to receive illumination will it then be granted. He who asks this, knows not what he asks; for were the prayer answered, life henceforth for such an one would be a weary round, as Hamlet says: "to-morrow and to-morrow and to-morrow brings in this weary round of life"; for, having seen the glory of this immortal EGO, all else seems so base, so commonplace and mean, so inglorious, that oftentimes the personality has utterly collapsed when thrown back from the radiant vision of this glorious immortal entity possessed by all alike, though scarcely dreamed of by any save the very few who, discontented with the ignorance and emptiness of terrene existence, aspire to know the great reality of the supernal. As the incarnations of every entity, passing through certain orders of experience through numerous lives, inevitably culminate in this moment of conscious realization of the immortal entity; the Buddha says: "All shall reach the sunlit snows."

You who through your daily life move on unthinking, not caring, inactive, you shall hear when your supplications reach this high entity, *"Lo! thou didst not even try, knowing that even thy failures were acceptable to me."*

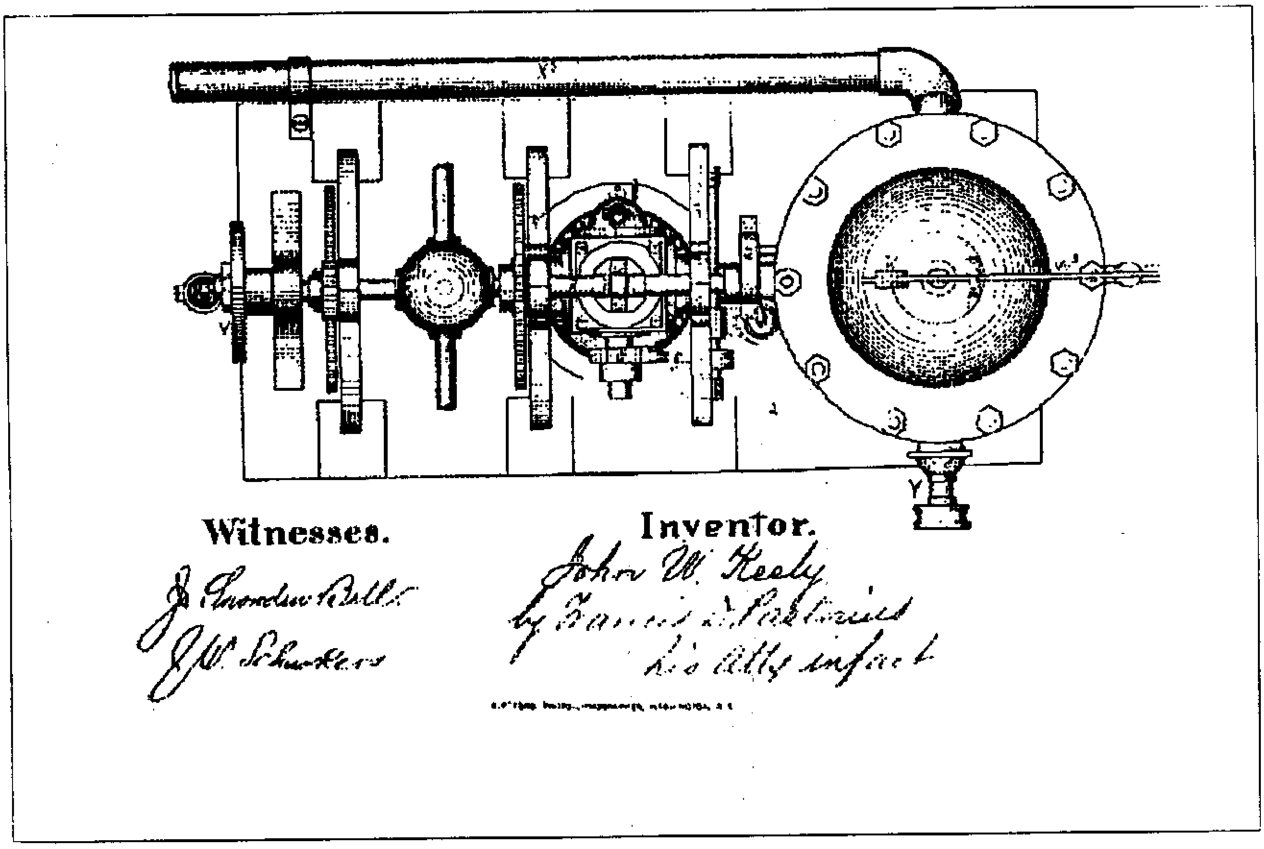

Top view looking down on the original Keely motor.
Scanned image from original patent drawings.

Front view looking straight on the original Keely motor.
Scanned image from original patent drawings.

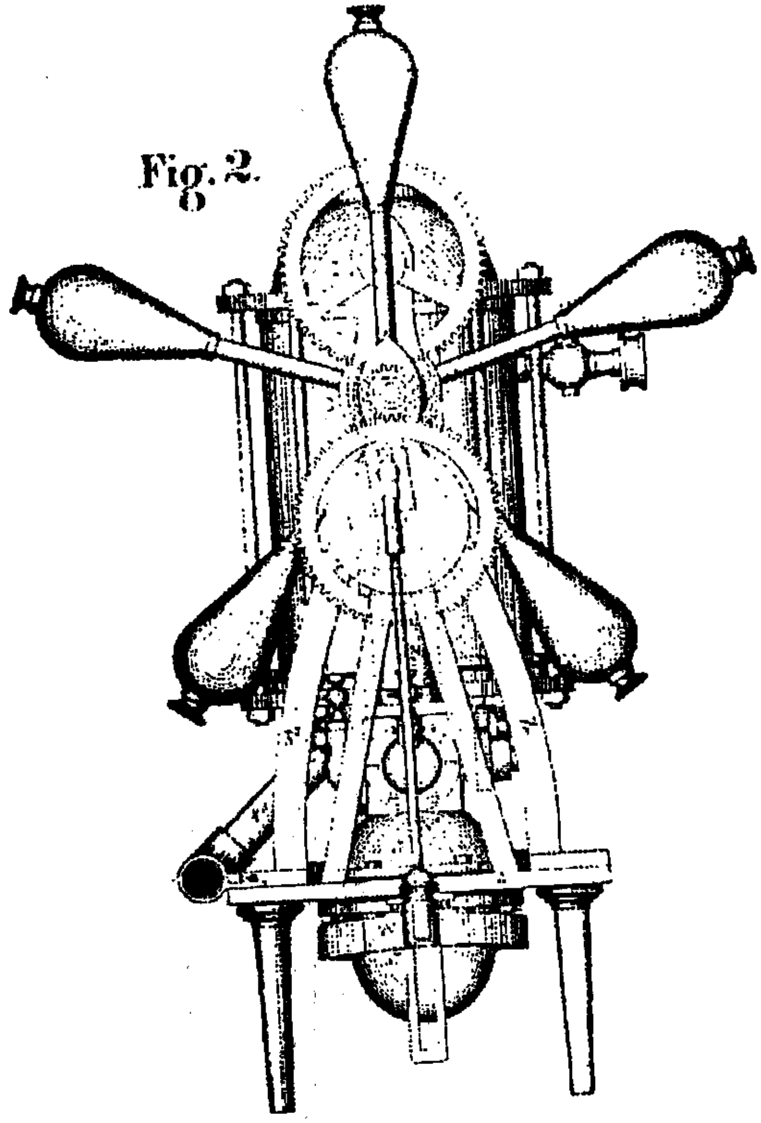

Fig. 2.

The Chord Settings of Life

John Ernst Worrell Keely
from Dashed Against the Rock

Question: "By what term shall we define that force which, when differentiated, expresses itself on the lower planes of manifestation as charity, self-forgetfulness, compassion, and the tendency of all illuminated ones to association in universal brotherhood?"

Keely's reply: "I hold that ONE SUPREME FORCE, which we may term the incomprehensible, holds within itself all these sublime qualities, as an octave embraces its many tones. This force expressed in the human organism, has what may be termed CONCORDANT CHORDSETTINGS dominated by one or other of the above-named differentiations of this supreme force.

Now let us ask what makes human beings differ the one from the other. The reason we give for this striking natural phenomenon is that in one individual these chord-settings are allowed full amplitude in action, while in others they are suppressed, and by suppression rendered latent. To illustrate: we will picture a beggar asking alms of a richly dressed gentleman, who passes by entirely oblivious of the suppliant's needs; but here the wonderful law of sympathetic action intercedes, making the wealthy individual comprehend the necessitous condition of the pauper.

At this point the *ego* enters the chamber of the ordeal; here, in commonplace life, in everyday surroundings, man is tried; this is *initiation*.

The well-to-do man goes on his way, not caring to stop, hurried perchance by the urgency of worldly affairs; the chord-setting representing that differentiation of infinite force called by us *compassion*, acting upon its concordant chord-setting, loudly proclaims to the *ego* what is right action, and the opportunity passing when it should be seized, powerfully exerts its force against the will of the personality that would suppress its action. Here is where the battle is fought; simple though the illustration may appear, it forcibly sets forth the actual conflict continually waging between divine wisdom and mortal error, carelessness, and ignorance.

The man of wealth and position goes on, perhaps, for some distance, the battle all the while continuing; finally, he stops and turns back, he yields to the dominating influence of that chord-setting; he gives the beggar alms and goes on his way with the skies bright above him; he has won a battle he will not have to fight again. Understand that victory is won, *not by the giving of alms*, but by *yielding to that divine force-differentiation*. If this chord-setting had not retained its dominance, it would only require a few instances of the above type to render it latent, and when latent the person is no longer amenable to influences calculated to arouse compassion. In like manner, other centers may be rendered latent by repeated suppression, until we find a person so dead to all appeals from the various chord-settings that his whole course in life is represented by the sum-total of the antagonisms internally produced; results proving this are seen everyday.

You ask why do people commit such blunders and perform such acts as they do, all the while seemingly unable to help themselves. The reason is that they have rendered latent these centers which otherwise would have given them the power to rightly control their deeds instead of being, as they now are, dominated exclusively by the forces of aggregated matter which we usually call the self-will of the outer personality, as distinquished from the distinctly humane individuality which always responds to a divine appeal.

It can be readily seen from this example that a man can mould himself practically as he chooses; though he may have to encounter many obstacles erected by himself in past periods of earthly existence, as he comes to earth anew with these chord-settings latent, or developed to the extent they were so, at the conclusion of his earth-embodiment; a man has therefore only to carefully examine the condition of these settings to learn whether they are latent or developed; if latent, he knows well that if he yields to the dominance of the chord-settings of the supreme force - and he needs no monitor other than these to instruct him, their voice being loud-toned, full, strong, deep, and high - to carry him on to the consummation of his highest ideal, their suppression leading him to the lowest condition in which we behold that section of humanity which is dominated by the action of the blind forces inherent in aggregated matter, - forces which are at all times powerfully and intimately associated with everyone who possess a material body.

The work of arousing to activity the latent chord-settings is sometimes equivalent to giving birth to an entirely new condition in the person, the intellect and the imagination having to actively cooperate in the endeavor to produce even the minutest degree of activity.

In such persons their hardest experiences may be the greatest benefit to them, if rightly taken; for it is through certain orders of experience - not unattended with suffering - that these centers are powerfully acted upon. The foregoing illustrations of the merchant and the beggar introduces one in whom the action of the centers is to a certain degree active.

In experiences where the emotions are *intensely* aroused their action is far greater, and we may see the result of the conflict, in the event of a person of hitherto unnoticeable traits being developed either into a person of crabbed, irritable disposition or into one of considerable moral beauty and attractiveness, and all because a center hitherto dormant has been powerfully aroused or more completely suppressed.

Who knows but that all the varieties of disposition we see expressed every hour in the persons of those about us is due to HABIT; in the case of the sweet and lovable, to the habit of constant yielding to the dominance of what we may term super-celestial force, while on the other hand, the sour and morose are but the suffering victims of their own habitual suppression of these same divine centers of radiation, which are continually dispersing the divine energy focalized upon them throughout the eternal ages of unwasting life.

Concerning *circumstance* let it be most emphatically stated that they never need be permitted to suppress the upward tendencies of our nature; that they have, like *all* experiences, and opportunity contained within them, to act either for the more perfect dominance of the celestial or the terrestrial, none need deny; but it rests with every individual to *embrace* or *reject* opportunities as he will, to accept the honors of initiation or undergo the regrets consequent upon failure.

Only the keen, sensitive soul can understand when these opportunities come and go, for only such have won this right by successive victories gained through yielding to these celestial streams of force, and the conqueror over himself is the victor *always*, though he may seem sometimes to yield obedience to a force greater than himself. Let the supreme desire of each one of us ever be that these resonating centers, permanent throughout measureless cycles of time, graduated by the all-wise builders of the universe to perfect concordance with the Divine Force which is in essence incomprehensible, - shall vibrate to fullest amplitude of action; so shall we each escape from the pains, sorrows, and disappointments associated with their suppression and *inevitably* resulting from it."

The Dynamics of Mind

Henry Wood
reprinted from ARENA, #LXIII February, 1895

In the light of recent psychical demonstrations, it has been said that thoughts are things, perhaps a more exact statement would be that they are forces.

In physical science, the present trend of teaching is distinctly from the former accepted atomic basis, which included the solidity and potency of matter, towards a hypothesis in which energy is regarded as the underlying principle of all phenomena. Thus the atom, which has never been discovered, and is not likely to be, is no longer recognized as the real unit in the physical economy, energy being now accepted as the primal starting point.

Manifestations to our senses, which we call light, heat and sound, are only differentiated modes of vibratory forces. Primal energy, unitary in its essence, and always conserved in the aggregate, takes on, to us, one of several qualitative appearances, according to the form of its waves, or rather, perhaps, the rapidity of its vibrations. Under certain circumstances and through the action of laws yet imperfectly understood, these various modes of manifestation are interchangeably transformed in constant repetition.

Modern science has accepted the conclusion that vibration is a universal law, and the recognition of this fact is the key which is unlocking mysteries and solving phenomena hitherto unexplainable. It has furnished an all-comprehensive working hypothesis. Beginning with an inter-molecular rhythm of inconceivable rapidity in all bodies, even those that appear to be solid and at rest, its domain of wave movements extends through all space, and its impulses are coursing in every conceivable direction. They are ceaseless and endless. The cosmos may truly be said to be "all of a quiver."

The basic medium of these enumerable wavy motions is undoubtedly the universal ether, the nature of which can only be dimly conjectured through its multiform manifestations. Who can say that this is not the boundless common meeting ground between the spiritual and the material? Unaccountably enough, the myriads of vibrations of different kinds and velocities that are sweeping through space do not appear to disturb or neutralize each other in the least. There is a clear path for all.

These late developments in physical science, which have only been hinted at in the most general of terms, carry with them necessary inferences and correlations, the scope of which can yet hardly be imagined. The dematerialization, or perhaps what may even be called the *spiritualization of physics*, as a science, is one of the marked logical tendencies. There is also a growing demonstration and conviction of the deceptive and utterly unreliable nature of sensuous appearances. Science, before finally accepting any proposition as proven, has always insisted upon material and mathematical demonstration. This is well in its place but it is not all, in fact it is only the lower and cruder side.

Matter, as formerly regarded, seems to be consciously melting into mind or spirit. It is no longer inert or dead, but instinct with life. Its transformations are ceaseless and mysterious. Can any one explain just how and why a visible solid can take the form of an invisible gas and *vice versa*?

The theoretical boundary line between the immaterial and the material is getting very faint if not actually disappearing. Let us drop our crude, childish materialism and rise easily and reasonably to the grand conception that differentiated forces are being traced back, even through the methods of the physicist to the *One Primal Energy* - INFINITE MIND. The veils which in our infantile development we have hung around external nature are growing so attenuated that we can almost discern with unaided vision the active operation of Supreme Intelligence, Goodness and Beneficence.

All profound discernment and analogy lead back to the grand fundamental premise, that behind all manifestations, energy is One, that it is an Intelligent Energy, and is therefore Omni-present Mind. Monisn, or the inherent unity of all things, is the growing inspiration of science. It is thereby confirming the impressions already received through the delicate vision of the unfolded interior faculties. Paul's immortal aphorism, that "in Him we live and move and have our being" has waited long for scientific endorsement, but it is apparently soon to be realized. We behold the universe as soulful and not mechanical. This is no ancient superstitious pantheism resurrected. Rather the Deity is infinitely honored as compared with any and all past human concepts.

If all energy, in its last analysis, be Intelligent Mind, and vibration the universal method, we may reasonably infer that human mind or volition, being in, and a part of the whole, should form no exception in the working plan of its orderly activities. If essential, potential and ideal man be the "offspring," "image" and manifestor of God, nothing unlike it could be normal. As the former ideals of a Deity, localized, personified, changeable and in every way unconsciously limited, are slowly replaced by the transcendent ideal of the unconditioned "All in All," the interrelation of all things, to and in God, is being grasped. When man refines, enlarges and elevates his consciousness of Divinity, he does the same for his own deeper and generic spiritual nature, which though temporarily obscured, is in reality, himself. If God be spirit, man, His reflection and likeness, must also be spirit and not dust. By a traditional and distorted self consciousness he has thought himself to be a poor, sinful, *material* being, and the formative power of his mental specification has externally actualized his model. He *is* mind or spirit, but his physical expression, which should be of ideal quality, outpictures his perverted estimate of himself. Not recognizing his true being, he has drawn a mistaken outline and then naturally filled it out. He has thus unwittingly hidden his own potential and divine forces, though they are still within. The mirror of false consciousness has reflected a doleful image which he has seriously taken for himself.

The purer and higher trend of science is characterized by a gradual refinement and immateriality. The laboratory should become a sanctuary, for in it are gained glimpses of the Eternal. Man himself is being more truly interpreted as the highest expression of divinity. He is a concrete manifestation of the One Mind, finite, but with unlimited possibilities. He is inconceivably great, though ignorantly unaware of it. But a significant indication of his growing consciousness of the possession of supernal power is found in the recent discovery of the dynamic and formative potency of his thought. The Infinite Uncreate is the primal and universal energy, but man is its embodier and manifestor. His mental forces cannot create *de novo*, but they can mould, utilize and express. The unfolded soul having developed a self-consciousness of its transcendent power, intelligently sends out its own vibrations from its own center. Conforming to the divine plan and chord, it becomes a reflection, or secondary radiator of rhythms which are concordant with the Original.

We are logically led to the conclusion that the recent recognition of the potency and utility of the projective vibration of thought, is an unprecedented and immense step in scientific achievement, human unfoldment and spiritual evolution. Man is finding his rightful dominant place in nature, in the arcana of soul force and expression, and in his relation to the Infinite.

Before considering specifically the dynamic relations between mind and mind, it may be well to note briefly these relations as they exist between a human mind and its physical counterpart. Man is mind, and this statement implies that the physical organism is not man, but only his visible index or expression. To attempt to prove this is like demonstrating an axiom, but yet mankind at large indicate by their action that they do not practically believe it. Nine-tenths of the care, labor, and attention of the world is bestowed upon the body and its gratification, or upon those subordinate mental powers, the product of which will command the greatest commercial value. Most of the prevailing systems of education, so called, have the same end more or less directly in view. The trained intellect, including not only technical and professional attainment, but also the powers of literary, poetic and dramatic ability, eloquence and wit, are largely rated and valued on an economic and material basis. To train, control and uplift the mind, and develop its higher faculties for its own sake, and that of others, is not common. The world is still endeavoring to "live by bread alone."

Prevailing systems of philosophy, science, theology, therapeutics, sociology and charity, including Darwinian evolution, all proceed upon the general hypothesis that man is intrinsically a *material* being. He has an attenuated quality called a soul, dependent upon fleshly brain cells.

The "fall," not historic but continuous, is from the ideal, potential and in mostly actual, into the external of appearances, and this comprises the Adamic consciousness. Men cling to the sensuous Eden until they are started and driven from it by the loud calling of the divine voice within. The beneficent expulsion from that Eden, and the succeeding necessary restlessness, furnish the true and only impetus for voluntary moral and spiritual evolution. The world is still largely peopled with Adams who practically believe that they are made of red earth or dust.

Is man to grasp, mould and rule that little portion of dust that he has temporarily taken on, and which before has often been used to express and embody other qualities of life, or must he believe himself in bondage to it? Shall the shadow, even though real as a shadow, dominate the substance? Not forever, even in what is called this life. So soon as man recognizes the fact that he is a mental and spiritual dynamo he will no longer remain a vassal in his own legitimate kingdom. But the more specific treatment of the relations of mind to body must be reserved for a subsequent paper.

A dominant vibration in the thought-atmosphere is able to arouse a nation, or a continent. Great minds, as well as those of less development, are submerged and swept along by it. Crusades, reformations, revolutions, and reforms, furnish numberless illustrations of psychic upheaval and contagion. *Through sympathetic vibration a vast number of responsive mental strings are stirred into action.* As the rhythmical step of a regiment will powerfully shake a strong bridge, so the concerted energy of mind will generate tidal waves of tremendous import. The result is not merely from a contemporaneous logical process, carried on respectively by many individuals, but from a great immaterial gulf stream, deep and mighty, though silent and unconscious.

Mind, as a *force*, is no more unintelligible or unthinkable than other vibrations of unseen energy. And there lies the tremendous significance of the new psychology or recognition of soul force. Till recently conventional science, as taught in all accepted textbooks, recognized no extension of the dynamics of thought beyond the confines of the physical organism. The mind, with feeble domination, through nerve channels, could transmit its orders to different parts of its visible counterpart, but it was not believed that it could go one inch beyond that limit. Any suggestion that telepathy, or thought transference, could take place at a distance of a thousand miles, or even one mile, would have been pronounced impossible.

We shall waste no time in the mere attempt to prove the fact that thought is, and can be, projected through space, both consciously and unconsciously. No well-informed individual who has given any adequate attention to the subject now questions it. Scores of pages might be filled with examples, now on record, which are entirely beyond collusion or coincidence. Every one of thousands of hypnotic experiences proves it, and every case of healing through mental treatment attests it. There is no fact in physical science better assured.

And how has the world received this transcendent truth which is transforming in its potency, all-inclusive in its sequences, and divine in its possibilities? Very much as it would a new *curio* or an ingenious toy. The institutional psychologist fondles it, turns it over, weighs and measures its properties in his laboratory, speculates *about* it, and makes a profession of it. But the last thing to be thought of, is to make it useful to mankind. That would be unprofessional. To harness and utilize this force of all forces for the good of humanity would lower it from the select and charmed circle of professional theory and speculation to the broad plane of practical and beneficent agencies.

The average psychical researcher shows much of the same indifference as to any utilization of his favorite principles of pursuits. He is engaged in a never-ending pursuit of phenomena. He will strain his investigative powers, and burn midnight oil in testing, comparing and recording curious manifestations, and in interpreting their methods and laws, but as to their practical application in ethical culture, therapeutic potency or spiritual unfoldment he is as innocent as a child. It has not occurred to him. These reflections are made in no impatient spirit as applied to individuals, but rather to show the negative character of systems of thought into which we have allowed ourselves to become crystallized. How much freedom, originality and progress would at once be manifest if the fear of being called unprofessional and unconventional, which now holds men in bondage, could be eliminated!

Besides the classes already noted there are many excellent people, lovely in character and pure in motive, whose temperamental fondness for the mystical leads them to seek visions, dream dreams, and to cultivate an order of phenomena more dramatic than profitable. Abstract truth and vivid demonstration are well, but the world is hungering for their application to its woes.

If we gained some knowledge of the laws which govern a force inconceivably grander and higher than electricity, may we not dismiss undue sensitiveness as to deviations from traditional scholasticism, and for the sake of humanity, step out of the ruts which have been grooved by the schoolment of the darker and narrower past? All great advances in their earlier aspects have been irrational innovations.

Regarding the fundamental basis of psycho-dynamics, not only as admitted but overwhelming proven, let us now concisely sum up a few of the results which logically should be realized. They are of stupendous significance, but surrounded as we are by the blank walls of our self-imposed and traditional limitations we can hardly picture them even to the imagination.

Thoughts being forces, every mind is a creative center from which rhythms of qualitative energy are going out in all directions. By their impact upon corresponding chords in other minds, these are also swept into active vibration. Throw a pebble into a lake and the placid surface at once becomes vibrant with a series of ever-widening circles which go out to its utmost boundary. They are never quite lost, or neutralized, though we may be unable to trace them to their final destination. So every soul is the seat of a great centrifugal current, which is generated and set free in the simple process of thinking. This is true - though less in degree - of desultory or aimless thought, as well as of that which is concentrated and projected with definite intent. *Every thinker is a battery of positive forces even though he utter never a word.*

The soul - which is the man - is a resonant instrument with innumerable tremulous strings of the most delicate quality. The water in the lake responds to the pebble, but the medium through which thought-waves pass is infinitely more subtle and elastic.

What volumes of potential energy are wasted and far worse, in negative and discordant mental activities! We are not thinking for ourselves but for the world. With the shuttle of thought in the loom of the mind, we are weaving the multicolored fabric of conditions, and these not merely immaterial but to be outwardly actualized and manifested. If one in his own soul strikes the discordant notes of anger, envy, avarice, selfishness or even those seemingly more harmless ones of simple fear, weakness, grief, pessimism or depression, he is creating and vibrating those conditions far and near, thereby stirring the corresponding chords in other souls into sympathetic activity. The sphere of outward action is limited, while that of thought is boundless. Mere doing makes ephemeral reputation, while quality of thinking determines, or rather *is*, vital character.

Every one's thought-images are being constantly impressed both upon himself and others. His mind is a busy factory where conditions are positively manufactured. He weaves their quality, consciously or unconsciously, into every nerve, muscle and tissue of his own body. His materialistic thought tethers him in a little circle of limitation, while boundless green fields lie beyond waiting for occupation. His mental pictures of evil, disorder and disease, photograph themselves not only upon his own mind and body but upon those of his fellows.

One cannot afford to think much about evil, even for the well-intentioned purpose of its suppression. The true remedy is its displacement. Thought-space given to it confers realism, familiarity and finally dominion. To silence discordant strings in ourselves or others we must vibrate their opposites. To truly sympathize with a friend who is quivering with trouble or sorrow, is not to drop into his rhythm and intensifying it - as is usual - but to lift his consciousness by striking a higher chord in unison. The road to mental and physical invigoration lies through the dynamics of formative thought. Our way to elevate other lives is also through their creative mental energies.

When the art of projecting thought vibrations on a high plane is systematically cultivated, and the concentrative habit developed, potency for good is increased a hundredfold. Force is no longer squandered in worse than useless discordant negations, but intelligently conserved in positive vigor and exuberance. *Purposeful thought ministration, spiritual and pure in quality, accurately and scientifically projected, like an arrow towards a target, will be the great harmonizing and uplifting agency that will transform the world.* Vibrations of love, peace, spirituality, health, sanity and harmony, will be radiated in ever widening circles, striking responsive unisons that are only waiting for a well-directed concordant impulse.

The dynamics of mind, when generally utilized, will be the sovereign balm that with scientific accuracy will heal all the infelicities of society. It will usher in not only reform but regeneration. In its copious fulness it will overflow from the altitude of spiritual development, until the subordinate plains of intellectuality, ethics, therapeutics, sociology, economics and physics are swept purified and uplifted. The highest includes everything below. With the kingdom of heaven - which is subjective harmony - first sought, "all these things" will be added.

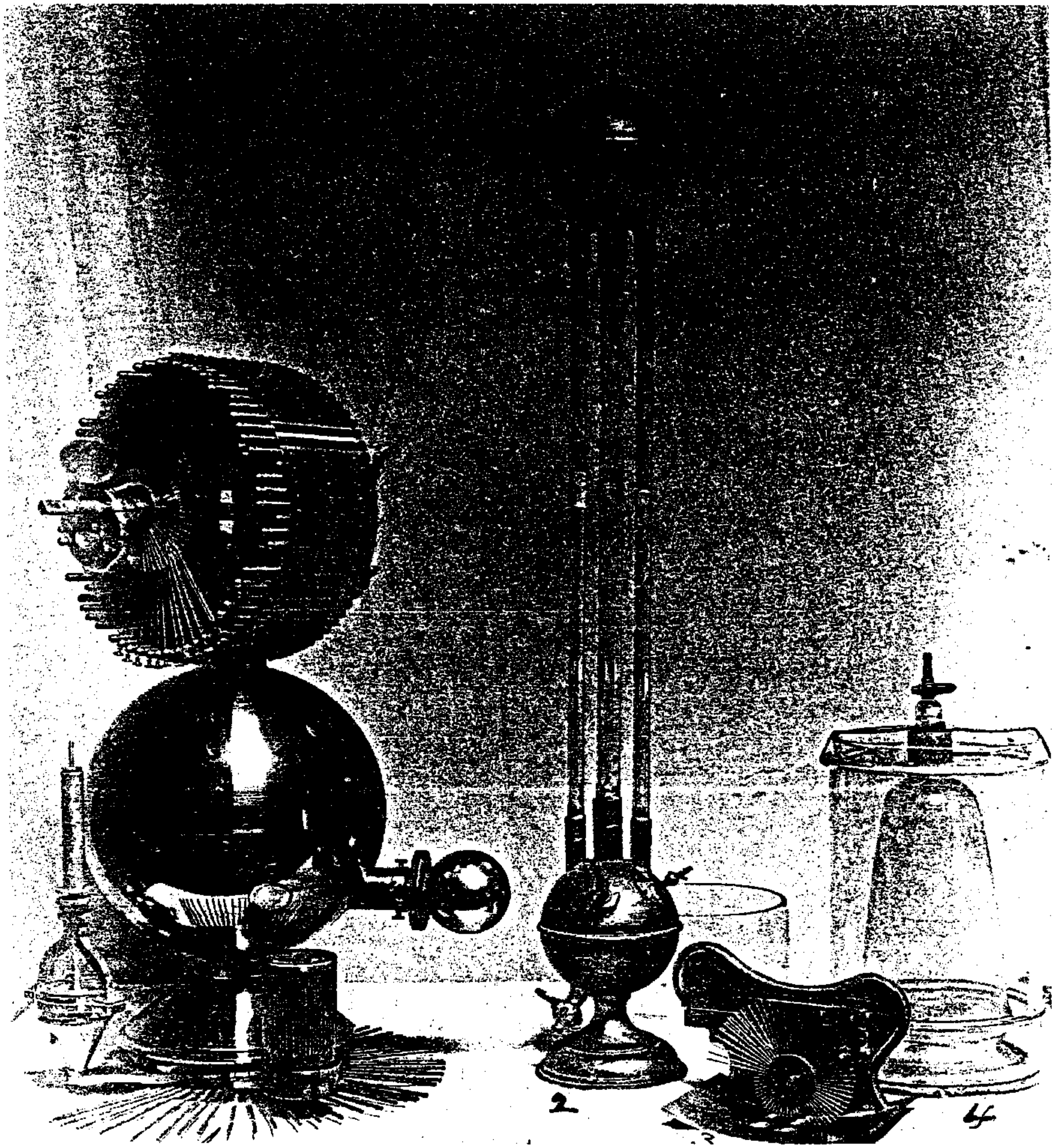

(*from left to right*) - Sympathetic Negative Transmitter; Triple Column for Measuring Specific Gravity of Gases; Dominat Scale and Resonator; Hydrogen Generator

The Neutral Center

John Ernst Worrell Keely

"A center of introductory action is necessary in all operations of Nature."

All structures require a foundation. This neutral center is the foundation.

"Every molecule, every mass, every moving body in space, every solar system, every stellar system, EVERY ROTATORY SYSTEM, is built about a NEUTRAL CENTER. It is the indestructible unit around which all that we recognize as matter is built. Immovable itself, it moves all things. Indestructible itself throughout infinity of time, it creates all things. It produced and preserves the incalculable energy of motion of the entire Universe. It bears the unthinkable burden of the mass of the Universe. It is the most wonderful thing Man has discovered in the Universe since he discovered fire.

"If we should take a planet of say 20,000 miles diameter and should displace a portion of the interior so as to have a crust of say 5,000 miles thickness, and at the center of the planet, place a billiard ball, that small mass, immeasurably smaller than the bulk of the earth, would bear the entire burden of the mass of the crust 5,000 miles thick and would keep it equidistant from itself. No power, however great, could possibly displace this central mass so as to bring it into contact with the crust. Furthermore, to move this central mass in any direction, would require a force sufficient to move the entire mass of the planet, and in propagating or continuing any such motion the neutral center, this billiard ball, will at all time periods remain still in the exact center, bearing the same equidistant relation to its hollow shell. The mind staggers in contemplating the burden borne by this neutral center, where weight ceases." No less wonderful are other properties of the neutral center.

"It is the cause of the physical Universe. Its attraction condensed that which we recognize as substance. Matter was evolved from the affinity of this neutral center for sympathetic streams and since it is immovable, it caused, through negative attraction the formation of nodes in these streams, where the vibrations thereafter continued to meet in a center of sympathetic coincidence causing the permanence of form and matter. Every nebula, an embryonic world, is acted upon, created and preserved by this neutral center, and at the termination of its cycle, it is ultimately also destroyed by it, causing its absorption into the Unknown from whence it came.

"The actual neutral center of the earth is, in fact, even infinitely smaller than the billiard ball referred to above. It consists of a compound interetheric point in space, so small that were we to magnify a pin head to the size of the sun, and from that substance take a particle of matter the same size, again magnifying it to the size of the sun, the neutral center would still be invisible, even though the structure of this last substance was examined through the highest powered microscope ever created, or to be created. For the neutral center is INDIVISIBLE. Its attributes do not belong to matter, and pertain in no way to matter, which is but its exterior manifestation.

"Every aggregate mass consists of molecules, each of which has its neutral center where the three modes of vibration, dominant, harmonic, and enharmonic, meet in a center of sympathetic coincidence and are equated without cancellation of their energy."

"The fixed neutral center of the earth

is the concentration or totalized power of all the several molecular neutral centers in the earth's mass. This neutral center, which is absolutely WITHOUT WEIGHT, an interetheric point in space, communicates direct by means of its outflow of sympathy, with every planetary mass in the Universe. Through its inflow of sympathy, through the solar intermediate, the sun, it receives the life flow from the SUPREME NEUTRAL CENTER that enables it to perpetuate its existence. Thus through the outflow from this Supreme Neutral Center that pivoting point of the Universe controls the existence and motion of not only every stellar, solar and planetary mass in space, but also the rotatory vibration, in every individual molecule, intermolecule, etc. through all the subdivisions of matter, thereby sustaining their existence and motion with the life flow.

"All foundations must be sufficient to bear their burden. Conceive then the Universe centered upon and resting the burden of is incalculable mass and kinetic energy on a vacuous interetheric point in space, so minute that it is actually INDIVISIBLE. This conception can only be fully comprehended by an infinite mind. Independent of time, because indestructible in its unity, independent in space, because through its properties space itself exists and without it would not exist, independent of matter because its properties in an external direction created all that we know as matter and gave it seeming permanence, the neutral center is that protean, uncreated, indestructible, forever-existing FIRST CAUSE. Without hands, without tools, without thought, without emotion, without love, without form, without substance, it, of itself, created all these. All that we see or can see in the objective Universe exists because of and by means of the properties and powers of the NEUTRAL CENTER.

"No machine heretofore constructed has been made with a neutral center. This conception of mechanics has never before dawned on man's thought field. Had this been done, perpetual motion would have become a demonstrated fact. Were a machine so constructed as to use its properties, an introductory impulse would suffice to run it for centuries. However, this would not be a useful mechanical contrivance for no more energy could be obtained from it than was originally given, and its only value would be as a timekeeper." Keely did not seek to invent, nor did he claim to have invented, perpetual motion.

"The neutral center is only established when rotation exceeds 100 revolutions per minute, which is sufficient, with the vibratory circuit running at 100,000 per second, to neutralize the influence of gravity on the volume of the neutral third of the sphere." The neutral center controls its sphere of operation, whether that mass controlled by it be homogeneous or complex. All differential mass antagonism is equated on the induction of certain orders of vibration. The structural conditions can be entirely adverse, even of unlike states --- gases with liquids, liquids with solids, solids with gases, its control will be as complete in one case as in the other.

"The condition of unstable equilibrium was born in each neutral center, that by means of this arrangement, the neutral center might become the connective link or controlling tendency, holding these two properties in balance and assuming either phase at will. Between the dispersing positive and the attractive negative it stands, the deciding factor, the Universal Will. "The action that induces this property I call the connective link is sympathetic planetary oscillation."

Theory And Formula Of Aqueous Disintegration

by John Keely

The peculiar conditions as associated with the gaseous elements of which water is composed, as regards the differential volume and gravity of its gases, make it a ready and fit subject of vibratory research. In submitting water to the influence of vibratory transmission, even on simple thirds, the high action induced on the hydrogen as contrasted with the one on the oxygen (under the same vibratory stream) causes the antagonism between these elements that induces dissociation. The differential antagonistic range of motion, so favoring the antagonistic thirds as to become thoroughly repellant. The gaseous element thus induced and registered, shows thousands of times much greater force as regards tenuity and volume than that induced by the chemical disintegration of heat, on the same medium. In all molecular dissociation or disintegration of both simple and compound elements, whether gaseous or solid, a stream of vibratory antagonistic thirds, sixths, or ninths, on their chord mass will compel progressive subdivisions. In the disintegration of water the instrument is set on thirds, sixths, and ninths, to get the best effects. These triple conditions are focalized on the neutral center of said instrument so as to induce perfect harmony or concordance to the chord note of the mass chord of the instruments full combination, after which the diatonic and the enharmonic scale located at the top of the instrument, or ring, is thoroughly harmonized with the scale of ninths which is placed at the base of the vibratory transmitter with the telephone head. The next step is to disturb the harmony on the concentrative thirds, between the transmitter and the disintegrator. This is done by rotating the siren so as to induce a sympathetic communication along the nodal transmitter, or wire, that associates the two instruments. When the note of the siren becomes concordant to the neutral center of the disintegrator, the highest order of sympathetic communication is established. It is now necessary to operate the transferable vibratory negatizer or negative accelerator, which is seated in the center of the diatonic and enharmonic ring, at the top of the disintegrator, and complete disintegration will follow (from the antagonisms induced on the concordants by said adjunct) in triple progression, thus: First thirds: Molecular dissociation resolving the water into a gaseous compound of hydrogen and oxygen. Second: sixths, resolving the hydrogen and oxygen into a new element by second order of dissociation, producing what I call low atomic ether. Third: ninths, the low atomic ether resolved into a new element, which I denominate high or second atomic harmonic. All these transmissions being simultaneous on the disturbance of sympathetic equilibrium by said negative accelerator.

EXAMPLE:

Taking the chord mass of the disintegrator B flat, or any chord mass that may be represented by the combined association of all the mechanical parts of its structure (no two structures being alike in their chord masses) taking B flat, the resonators of said structure are set at B flat first octave, B flat third octave and B flat ninth octave by drawing out the caps of resonators until the harmony of thirds, sixths, and ninths are reached; which is simple movement of the fingers on the diatonic scale, at the head, will determine by the tremulous action which is highly sensible to the touch, on said caps. The caps are then rigidly fixed in their different positions by set screws. The focalization to the neutral center is then established by dampening the steel rods, on the back of the scale, representing the thirds, sixths and ninths, drawing a piece of small gum tube over them, which establishes harmony to the chord mass of the instrument. Concordance is thus effected between the disintegrator and the ninths of the scale at base of transmitters with telephonic head.

This scale has a permanent sympathetic one, set on the ninth of any mass chord that may be represented, on any and all the multiple variations of mechanical combinations. In fact, permanently set for universal accommodations.

The next step is to establish pure harmony between the transmitter and the disintegrator, which is done by spinning the siren disk, then waiting until the sympathetic note is reached, as the siren chord, decreasing in velocity, descends the scale. At this juncture, the negative accelerator must be immediately and rapidly rotated, inducing high disturbance of equilibrium between the transmitter and the disintegrator by triple negative evolution, with the result that a force from five to ten, fifteen, twenty, and thirty thousand pounds to the square inch is evolved by the focalization of this triple negative stream on the disintegrating cell, or chamber, whether there be one, two, three, five, or ten drops of water enclosed within it."

"When moist air is subdivided by atomic vibration, the hydrogen is dissociated or separated from the oxygen but neither of them pass from the intermolecular state. Not until the intermolecular structure of hydrogen is subdivided by interatomic vibrations can it assimilate with the introductory etheric element."

"The latent force from liquids and gases differs from that liberated in metals in that it results from the breaking up of their rotating etheric envelopes, increasing simultaneously the range of their corpuscular action giving under confinement forces of almost infinite variety of pressures. When liberated from the tube it is confined in it seeks its medium of corresponding tenuity with a velocity exceeding that of light."

"The sympathetic neutral flow (from the molecular neutral centers to the mass neutral center) is the latent power liberated in the disintegration of water. In water this power is dispersive, liberating latent elastic energy, while in metals the latent force manifests itself only in negative attraction and when these vibrations are applied to minerals, there is evolved the primal ether which is dissipated, leaving behind only an impalpable intermolecular dust, in which is contained in virgin form an metals originally in the mineral mass."

A non-intermittent flow of sympathy must flow along the Trexar, consisting of harmoniously adjusted thirds. Differential molecular weight is required, that is, the gas must contain heavy atoms and light atoms. This condition is satisfied by water, which consists of two light atoms of hydrogen and one heavy atom of oxygen. The oxygen atom weighs sixteen times as much as the hydrogen.

It is absolutely necessary to release the molecular ether by reducing the water to the interatomic state, before the atoms within the intermolecules can be released. Also the etheric capsules of the atomic and interatomic must be ruptured before the interatomic force can be produced. The ether from one subdivision is essential to subdividing the next higher.

Keely systemized the proper vibratory chords progressively, from the introductory molecular to the interetheric, embracing seven distinct orders of triple subdivision. He proved to his own satisfaction that progressive subdivision evolves new and distinct elements "too multiple to enumerate."

Water is a ready and fit subject for vibratory research on account of its "differential volume" and gravity of its gases. Even on simple thirds the differential action between hydrogen and oxygen causes antagonism and dissociation. The differential of mass is such that the hydrogen and oxygen become thoroughly repellant and thereby exhibit thousands of times more force than could be induced by heat on the same amount of water. Vibration of antagonistic thirds, sixths and ninths on the mass chord will compel progressive subdivisions and to get the best effect on water, the instrument is set on all of these. First, the focalizing chord of sixths induces perfect harmony to the mass chord. Then the diatonic scale and enharmonic scale at the top of the instrument (or ring) is harmonized with the scale of the ninths at the base of his Trexar with the telephone transmitter. His next procedure was to disturb the harmony between the transmitter and disintegrator on the concentrative thirds or sixths, by rotating a siren. This he used to induce sympathetic communication along the Trexnonar between the vibrator and disintegrator. When the siren concorded to the neutral center it established the highest order of sympathy. He then operated the "transferable vibratory negatizer" or "vibratory accelerator" in the center of the "diatonic and enharmonic ring" at the top of the disintegrator and complete disintegration followed. The thirds first resolved the water molecules into hydrogen and oxygen, then the sixths resolved the hydrogen and oxygen into new elements by dissociation, producing "low atomic ether." Then the ninths resolved the low atomic ether into a new element or "high or second atomic harmonic." All these transmissions were simultaneous on the sympathetic equilibrium being disturbed by the negative accelerator.

Keely's first efforts toward disintegration of the elements of water were successful because of the differential weight of the respective atoms composing its molecules. Subsequently, he discovered that this same method, disintegration by differential mass, may be carried out with any gas, in other words, it must contain heavy atoms and light atoms. The heavy atom or oxygen in water weighs just sixteen times as much as either of the hydrogen atoms to which it is joined by chemical affinity.

"In the dissociation of water, molecular and intermolecular dissociation produces the first order of ozone, which is refreshing and vitalizing to breathe. Atomic and interatomic dissociation produces the second order of ozone, which is too pure to breathe, for it produces insensibility." The third order of ozone, produced by etheric and interetheric dissociation, Keely used in his "carbon register" to produce a high vibratory circuit to break up cohesion, which he recognized as molecular magnetism. It is possible that this "dissociation" in the "carbon register" depolarized the iron molecules by allowing the corpuscular outreach to return within the molecular embrace. The acceleration of vibration producing these different orders is governed by the introductory impulse and the subsequent chords, as arranged in his Liberator, by which he dissociated water. In molecular dissociation he used one fork of 620, setting chords on the first octave. In atomic separation, he used two forks, one of 620 and one of 630, setting chords on the second octave. In etheric separation he used three forks, one of 620, one of 630 and one of 12,000, setting the chords on the third octave.

"Not until the intermolecular structure of hydrogen is subdivided by interatomic vibrations, can it assimilate with the introductory etheric element."

Certain differential, dual, triple or quadruple chords act as an introductory impulse exciting action on molecular masses (liquid or gaseous) decreasing molecular oscillation. They are then in a receptive state for vibratory disintegration. The diatonic-enharmonic is sounded, increasing molecular oscillation. Molecular subdivision takes place when oscillation exceeds 50% of their diameters. (Molecular or intermolecular?) The gas is now molecularly subdivided and assumes a high velocity of rotation in any confining container, be it sphere or tube and becomes the medium or prepared subject for further disintegration. At this particular time Keely sought to further excite this preparatory medium by the use of an "illuminated revolving prism" a condenser (concave mirror or convex lens) and colored lenses, thereby giving the vibratory frequencies of their respective colors. In his disintegrator a glass tube of sufficient strength to withstand a pressure of 1000 lbs. per square inch was arranged leading to the neutral center and the Trexar was attached to the external end of this glass tube. His use of this tube was probably also to convince the skeptics that the production of his power was genuine and possibly also was used to transmit color vibrations of such frequencies as to release the energy in the neutral center of the sphere.

When the triple introductory impulse or chords in three octaves, are transmitted to the disintegrator it SUBSERVES OR RENDERS NON-OPERATIVE THE MOLECULAR CONCORDANT THIRDS AND ANTAGONIZES WITH DISCORDANT THIRDS, extending their range of oscillation and thereby inducing the highest degree of repellant antagonism or repulsion toward the neutral center of the sphere volume.

The etheric dominant or celestial current induces aqueous disintegration and thermal concentration which two prime conductors display a coincident chord of sympathy with the celestial current. These two prime conductors link the terrestrial to the celestial, without which electricity and magnetism would tend to
become static or stable, all life and motion are governed by a dual power disturbance of equilibrium and sympathetic equation, both of which are in turn moved and regulated by electricity and magnetism.

Progressive molecular and intermolecular dissociation reproduces on a small scale Nature's system of light production and also invariably results in vortex motion. All corpuscular action in Nature is vortex motion. The vortex action between the terrestrial and celestial streams, terrestrial condensation against solar tensions, shows conditions analogous to those displayed in the dissociation of water into hydrogen and oxygen, that is, vortex motion of the highest order, but peripheral only.

EXAMPLE OF DISINTEGRATION

If the mass chord of the disintegrator is B flat, the resonators of that structure are set at B flat first octave, B flat third octave and B flat ninth octave, by drawing out the resonator caps and clamping the set screws. A simple movement of the fingers on the diatonic scale at the head will determine by the tremulous action of the caps when exactly resonant. Neutral focalization is then established by dampening the steel rods on the back of the scale representing thirds, sixths, and ninths, by drawing a piece of small gum tube over them, which establishes harmony to the mass chord of the instrument. This effects concordance between the disintegrator and the ninths of the scale at base of transmitter with telephone head.

This instrument had a universal accommodator for all ninths in all multiple variations. This was possibly his sphere resonator.

Concordant harmony must now be established between the transmitter and the disintegrator by spinning the siren disk and waiting for the sympathetic note as its velocity decreases. As soon as this note sounds, the negative accelerator must be rapidly rotated inducing high disturbance of equilibrium between the disintegrator and the transmitter by triple negative evolution resulting in enormous pressure. By triple negative evolution he doubtless means disintegration to the etheric stage.

"In the beginning was the word"

The neutral negative center

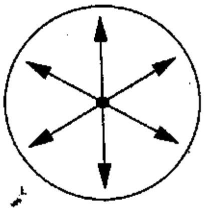

Sensitized - reaching out

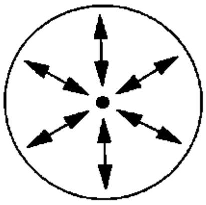

Establishing range of outreach

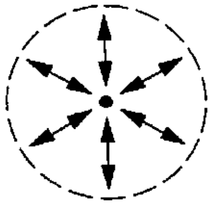

Positive Propulsive
Establishing Resonance

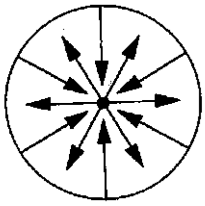

Equilibrium or
Harmony

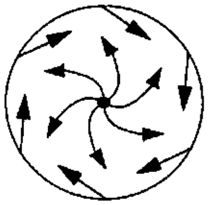

Introductory Impulse
Giving Vortex Motion

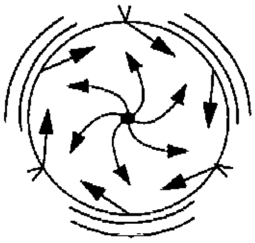

Vibration into
Thirds at Nodes

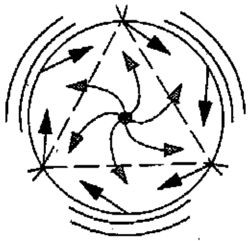

Nodes forming
Tensorial Lines of
Negative Force

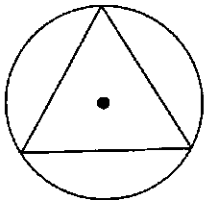

Completed Unit
of Duality

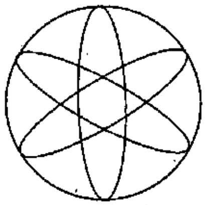

solid matter
in 3 dimensions

Keely's Motor Secret Revealed at Long Last

by Dale Pond

For over 120 years the secret concerning the function and operation of the Keely Hydro-Pneumo-Pulsating-Vacuo motor has baffled all the scientists and engineers who saw it. This mystery of mysteries has been solved.

I had developed a theory of operation over the past several years regarding its secret of operation. What was needed was to rebuild and run the motor to prove the theory.

During the second week of August, Vic Hansen and I resolved to recondition the original Keely motor in his possession. Extraordinary care was taken not to damage any part and all gaskets were replaced and all leaks fixed. We succeeded in recuperating only one of the two pistons as a valve in the second piston was frozen and we loathed to dismember the motor that this second piston would function. We were able to reestablish the functionality in only one piston. This proved to be sufficient for the test. Running on only one cylinder was not sufficient for the motor to run on its own power so we ran water through it in reverse. This was sufficient to prove the principle as we could clearly hear the "pop" of the one live cylinder coinciding with the timing and function of principles. As the piston "popped" the crank and connections responded to the momentary decrease in pressure. Yes, I said decrease in pressure.

This function of this motor represents the female or nightside forces at work. The force is one of implosion or centralization of vibratory forces. It is exactly opposite in function to every known type and manner used to today to derive motive power. This explains, in part, why it has remained a mystery all these years. Because they were looking for expansive force or the male, propulsive type forces, they remained baffled because no such force exists in its function. Two other notable things can be said about its function: one, it uses a lot of water to run, second, there is no heat whatsoever as it operates. The need for so much water can be resolved with better engineering - this should be simple feat as we know Keely eventually generated thousands of psi using but a few drops of water as he improved the design of the motor. The coldness of operation is a boon just the way it is. This means new models can be built of PVC or other plastics.

The importance of this discovery cannot be over stated: 1) It proves Keely wasn't a fraud as some allege. 2) It gives mankind a whole new energy source that is clean, powerful and simple. 3) It uses a plentiful and costless fuel - water. 4) It gets us into the doorway to reach for his greater and more wonderful discoveries based on his etheric vapor. This motor, as simple as it is, led Keely to the discovery and development of his etheric vapor and etheric force. By understanding its prime principle of function, we too, can make this giant step into the far reaches of applied physics.

So, how does it work? Where does the energy come from to propel itself? From cavitation of water. You are no doubt familiar with the phenomenon of water hammer. This is cavitation and is the very same source of energy in the Keely Motor. Water hammer is the development of a negative pressure cell in a flowing stream of water by cutting off its supply quite suddenly. The resulting partial vacuum collapses (implodes) with a force that can reach several thousand pounds per square inch instantly. In his motor this pressure was controlled to around 50 psi which was sufficient to run this little model.

Those of you working with implosion forces will be glad to see this practical application of the phenomenon of cavitation and the resulting implosive force.

Thus, the engineers investigating Keely's Motor were all looking for a male or explosive force and found none. During the collapse of the cavitation bubble formed in the piston chamber the piston was sucked down and not pushed up as in our "modern" engineering concepts.

The design of Keely's motor is so simple and basic that anyone with a modicum of talent can replicate and even improve on his design. He built his machine over 120 years ago and was hindered by not having the convenience we have of modern valves, ready made materials and 120 years of hydraulic engineering as support. Since the motor runs cold and with ordinary water the parts to build a unit can be had at your nearest hardware store. I suggest looking into PVC pipe and automatic valves such as are used in building sprinkler systems. Two valves, one for intake and one for exhaust coupled to a PVC cylinder should be all that is needed.

To go about building your own working model of his motor you should first acquire a copy of the book *Cavitation* by F. Ronald Young, MacMillan and Co. This book is now out of print. This book is the first and hence only book detailing cavitation. The different kinds of cavitation and their technical parameters are explored very thoroughly. The resulting effects of cavitation such as sonoluminescence and sonochemistry are all explored. The book further gives hundreds of references for source materials or previous research and lab work on this most fascinating of subjects. There is also a very good technical description of water hammer on page 249.

How the motor works

Water hammer is caused by a stream of water being shut off quickly. The trailing edge of the water stream (next to the shutoff valve) becomes rarefied (partial vacuum) and forms into a negative pressure area filled with a *sensitized* vapor composed of water, air and any gases that may have been present in the water. This vapor is the first stage or the basic form of Keely's etheric vapor. The negative pressure cell collapses when the water stream bounces back against the shutoff valve. The physics of these actions and reactions is fascinating and promises so much for development that we will all be very busy developing them for years to come.

In the Keely Motor the water is first drawn into the central pulsating vacuum chamber by a light vacuum formed when the pistons pull water into the system. This vacuum chamber maintains a steady negative pressure on the water from this point on. Between the vacuum chamber and the pistons are two valves. The first is a one-way check valve which serves to maintain the vacuum in the central pulsating vacuum chamber yet allows the pistons to pull water into itself as it is needed. The second valve is quite ingenious and simple. Its function is: 1) to allow water into the piston, 2) close for implosion and then 3) to allow drainage of the spent water.

The water then passes directly into the piston during its up (suction) stroke. The valve rotates shutting off the water stream just as the piston reaches its apex and cavitates which brings the piston down again. Instantly the valve rotates a little more opening two exhaust ports and the water is pushed and drained out through the bottom. This is all there is to the Keely motor. It is simple, elegant and functional. Rules of thumb are: 1) the more vacuum the better the cavitation, 2) the greater the velocity of the water the better the cavitation. The book "Cavitation" has more ideas on improving cavitation than we have space here to list.

Editor's Note (June, 1994): Since this article was written much more has been learned about this motor. Please refer to booklet *The Keely Motor* available from Delta Spectrum Research P. O. Box 316 Valentine, Nebraska 69201 (402) 376-1523.

Sympathetic Vibratory Physics

The Science and Technology of John Ernst Worrell Keely

Lecture given by Dale Pond before the
Swiss Association of Free Energy,
October 28, 1989, Einseideln, Switzerland

Mankind has placed himself in a grave situation. He has divorced himself from Nature with his current attitudes towards Science and Religion. The study of Nature is really a study of Science and Religion. Science is man's effort to know about Nature. Religion is man's effort to reconnect with that Nature. Hence through the arts of science and religion Mankind is attempting to know what Nature is in its truest sense and by way of this ever-expanding understanding to establish his kinship with Nature. These two subjects are bound together as a two sided unit. Science on the one side is ever expanding understanding and Religion is the assimilation of himself into this ever growing awareness of all that is around him. This concept is of course the ideal situation. Mankind has rarely attained this ideal state.

In actuality Mankind has separated himself from Nature with his ever increasing materialistic views embodied in both Science and Religion. His Science admits no connection between himself and his experiments. Atheistic attitudes separates him from Nature and natural processes. This is in error and has lead ultimately to the abysmal situation we find on earth today. Pollution is rampant and uncontrolled. Hunger, pain and suffering, and a lack of the basic material necessities plagues people in all countries. In fact, this century has been the bloodiest time in all of Mankind's history with nearly 100 million souls destroyed in all parts of the globe.

It would seem, at first glance, that with all the modernizations and increases in human science and technology, that we are living in the greatest age of history. But this is not so. To those who have perished and those who now suffer this is a terrible time in history. The question remains "why is this so?"

It is an easy thing to point one's finger at the "bad guys" and lay blame at someone else's feet. It is easy to say one political group is evil and one's own political group is innocent. The truth of the matter is these problems were not created by any one or several groups of politically oriented peoples. These problems were created and perpetuated by ignorance. This ignorance is a lack of awareness that all of Mankind is a part of Nature. This materialistic ignorance does not recognize that we are all part of a single natural scheme closely bound together in our actions and circumstances. A concept of Oneness, an idea that Mankind is a valuable part of Nature, has long been held and preserved through the centuries most conspicuously by the religious and metaphysical organizations. Materialistic science has cut off these tried and true ideas at the knees and left man bereft of his common foundation. Materialistic science has left mankind with nowhere to go except along the paths he has established in this century. These paths have lead to more bloodshed, pain and suffering than all previous paths he has ever choosen.

Free energy will not liberate Mankind from these evils. Neither will faster and more powerful machines free him of these self imposed burdens. Certainly more money, more people and more bureaucracies will not help either. We already have cheaper energy, more people, more bureaucracies and have spend vast amounts of human effort towards the solution of these problems and the net result has been an increase of the same problems. Mankind is on the wrong path.

What is needed is a comprehensive awareness that binds Mankind together within a wholistic knowledge that he is a part, an important part, in the natural processes of Nature in which he finds himself. What one man does has an effect on his neighbor and the natural processes surrounding them both. This dual view of nature is fundamental to science, religion and common sense. It is called the Law of Cause and Effect by science and karma by certain religions. This is an almost impossible law for most science or most religions to accept because it implies responsibility to one's self. It also implies that the perpetrator be charitable. He must concede that his neighbor have as much importance to him as himself. Another false premise of materialistic science is that of insufficiency. The establishment of monopolistic and totalitarian organizations is based on this false idea of there not being enough energy or materialistic substances to go around. With a broader paradigm that concedes the Universal Forces of Nature are limitless mankind can then be charitable. But to even concede this much many are hesitant because it implies their knowledge and power structures are based erroneously on false premises. These ideas of endless abundance implies sharing and most refuse to share because they believe in limited materialities as opposed to limitless potentialities. -- Hence, the exercise of charity has not been one of Mankind's greatest achievements.

My message to you today is that there is a paradigm of both Science and Religion that encompasses the highest of virtues expressible by Man. There is a field of science so powerful that it can build a new and better future for Mankind. There is a philosophy and science so all-encompassing that it can bring Mankind to a realization of his Oneness with his neighbors and with all of Nature. This paradigm has latent within itself, because of its breadth and scope, more than enough material and spiritual benefits for everyone because it recognizes and uses the very foundation principles of Nature itself. Because of its recognition and acceptance of its role in life it does not trample human rights or the rights of Nature and assumes full responsibility for itself within its own scope of possibilities. More than all this, it recognizes the valuable parts of all sciences and religions and merges them into a comprehensive and usable paradigm that has within itself seeds for unlimited growth and expansion. This paradigm of science is called Sympathetic Vibratory Physics.

What is Sympathetic Vibratory Physics? It is the science of Harmony, of Oneness. It holds the concept that all comes from One Source, One Force. It maintains that the fantastic array of things and activities throughout the universe are related in a simple manner. This simple basis of relativity is called vibration. It has been long recognized that everything in the universe vibrates. Thus, a study of vibrations is the study of the very foundation of Nature. For once we grasp the inner functioning of the vibratory Universal Creative Forces as they operate throughout the universe, then, and only then, can we correlate them into a comprehensive paradigm and bring them into practical modes of usefulness for the benefit of mankind. Unlike dogmatic, cold science, Sympathetic Vibratory Physics does not isolate one phenomena from another but accepts and shows the inter-connectivity between all things and forces. This interconnection between all things is vibration and related vibratory phenomena. To understand vibration we must develop a comprehensive organized science of vibration governed by laws and principles. Fortunately a great deal of this development work has already been done. We have a well developed science of music. Music is an organized art and science of vibrations found within the audible range of the infinite Electromagnetic Scale. A major segment of Sympathetic Vibratory Physics is therefore a study of music, its arithmetical and philosophical basis of development and how these may be

United States Patent [19] [11] **4,292,804**

Rogers, Sr. [45] **Oct. 6, 1981**

[54] METHOD AND APPARATUS FOR OPERATING AN ENGINE ON COMPRESSED GAS

[76] Inventor: **Leroy K. Rogers, Sr.,** #5 Capistrano Ct., Ft. Myers, Fla. 33908

[21] Appl. No.: **158,303**

[22] Filed: **Jun. 10, 1980**

[51] **Int. Cl.**^{3} **F15B 11/06**

[52] **U.S. Cl.** **60/407**; 91/187; 91/275

[58] **Field of Search** 60/407, 412; 91/187, 91/275, 364

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,881,399	5/1975	Sagi et al.	91/187 X
3,885,387	5/1975	Simington	60/407 X
4,018,050	4/1977	Murphy	60/412 X

Primary Examiner—Allen M. Ostrager

Attorney, Agent, or Firm—Burns, Doane, Swecker & Mathis

[57] **ABSTRACT**

The present invention relates to a method and apparatus for operating an engine having a cylinder and a piston reciprocable therein on compressed gas. The apparatus comprises a source of compressed gas connected to a distributor which distributes the compressed gas to the cylinder. A valve is provided to selectively admit compressed gas to the cylinder when the piston is in an approximately top dead center position. In one embodiment of the present invention the timing of the opening of the valve is advanced such that the compressed gas is admitted to the cylinder progressively further before the top dead center position of the piston as the speed of the engine increases. In a further embodiment of the present invention a valve actuator is provided which increases the length of time over which the valve remains open to admit compressed gas to the cylinder as the speed of the engine increases. A still further embodiment of the present invention relates to an apparatus for adapting a conventional internal combustion engine for operation on compressed gas.

22 Claims, 8 Drawing Figures

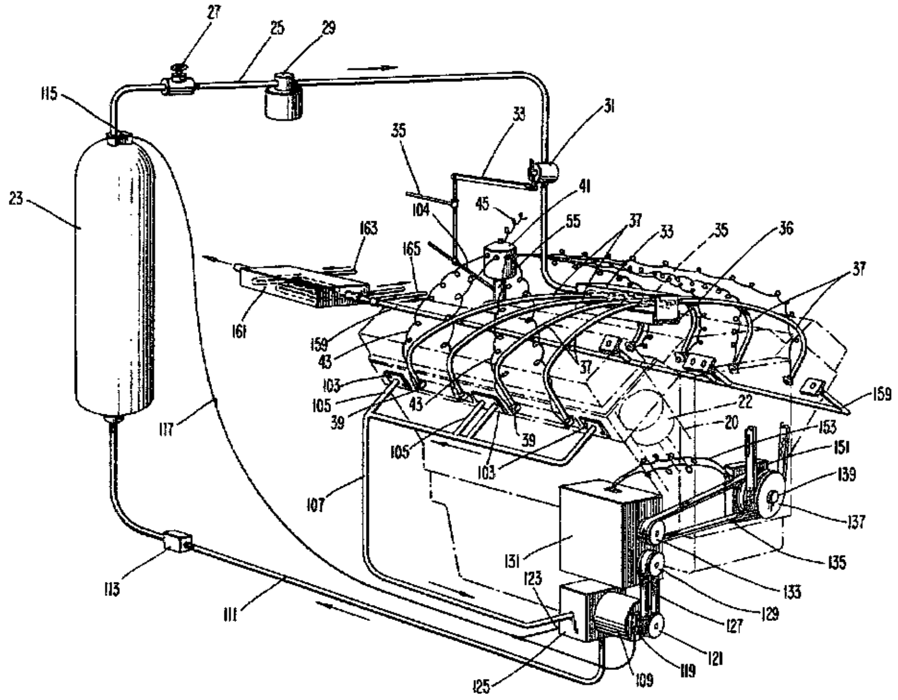

FIG-1

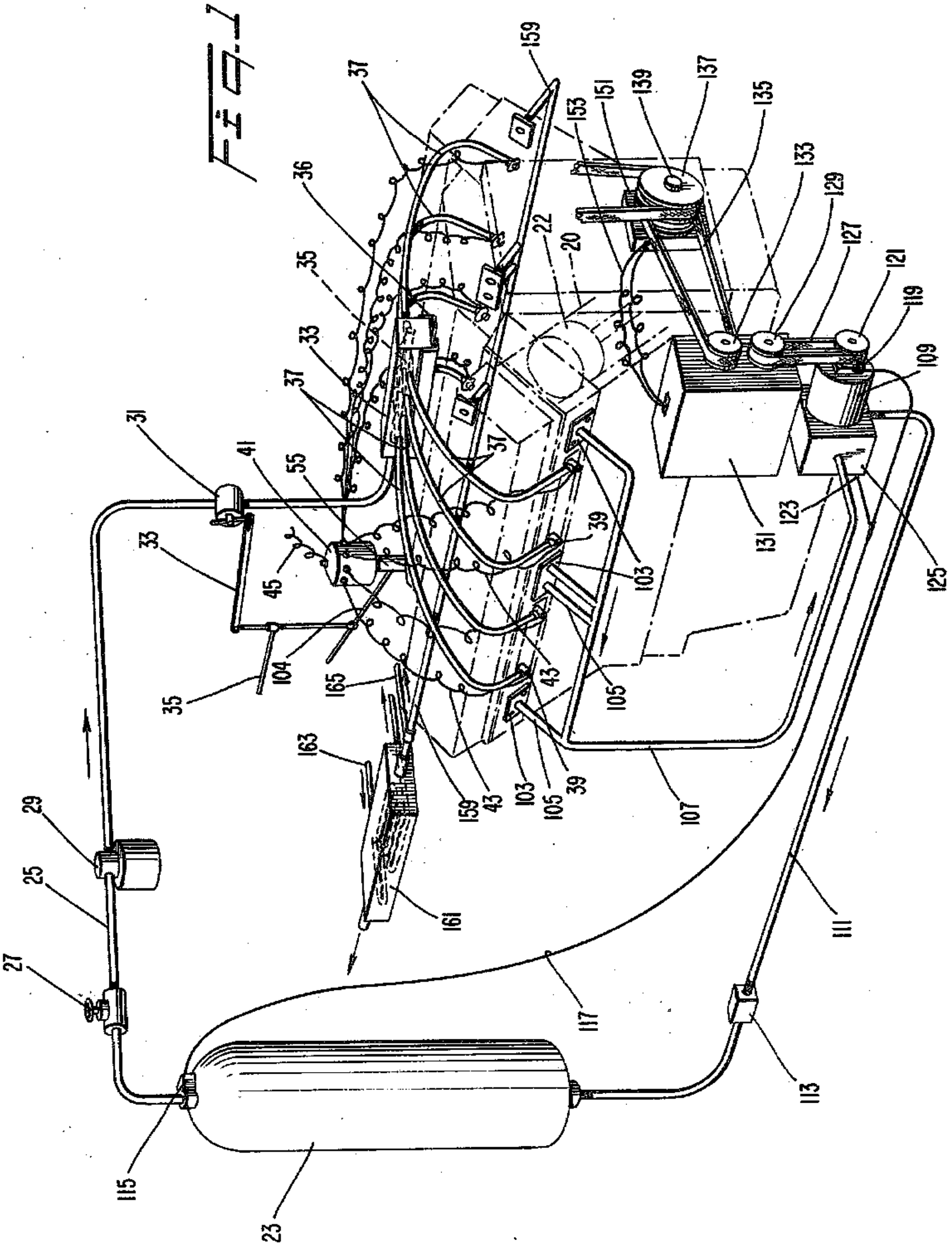

FIG. 6

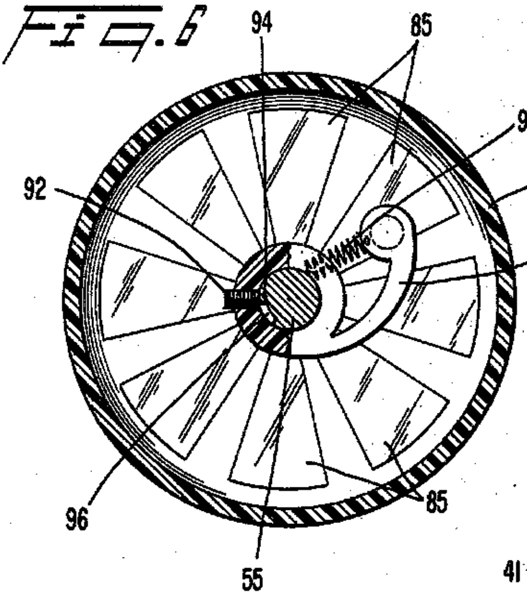
94
85
93
92
53
91
96
85
55

FIG. 7

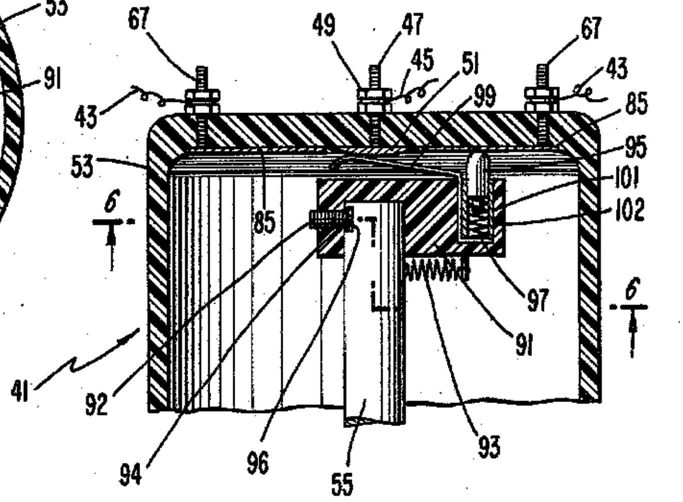
67
49
47
45
51
67
43
99
43
85
95
53
101
6
85
102
6
97
91
41
92
94
96
55
93

FIG. 3

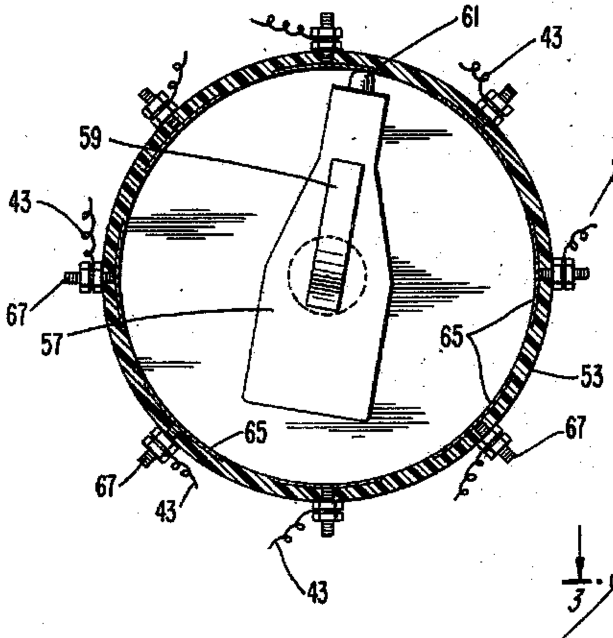
61
43
59
43
67
57
65
53
65
67
67
43
43

FIG. 2

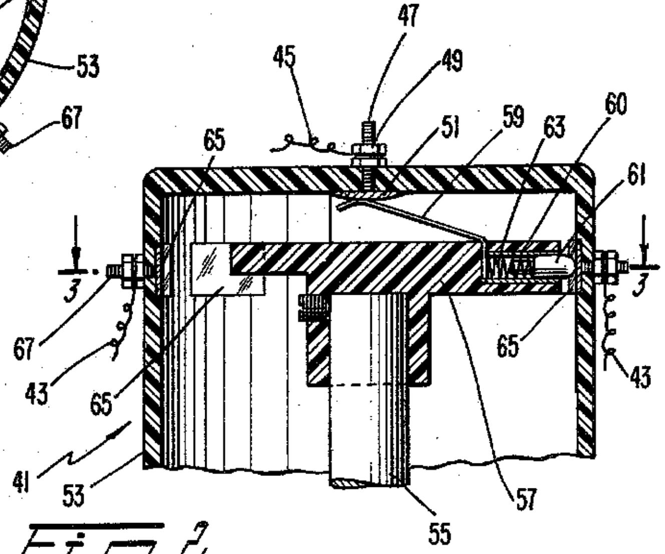
47
45
49
51
59
63
60
65
61
3
3
67
43
65
65
43
41
53
55
57

Fig. 5

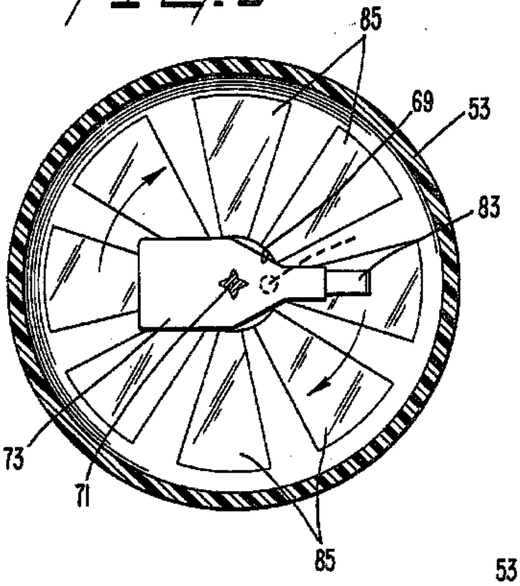
85
69
53
83
73
71
85

Fig. 4

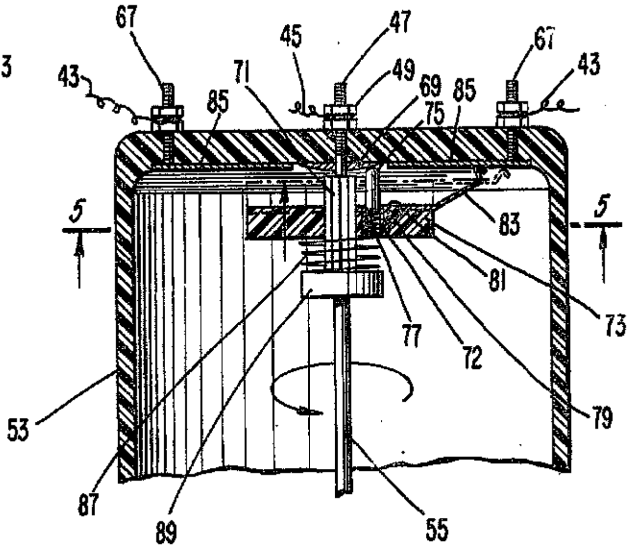
67
45
47
49
67
43
43
71
85
69
75
85
5
5
83
81
73
77
72
79
53
87
89
55

Fig. 6

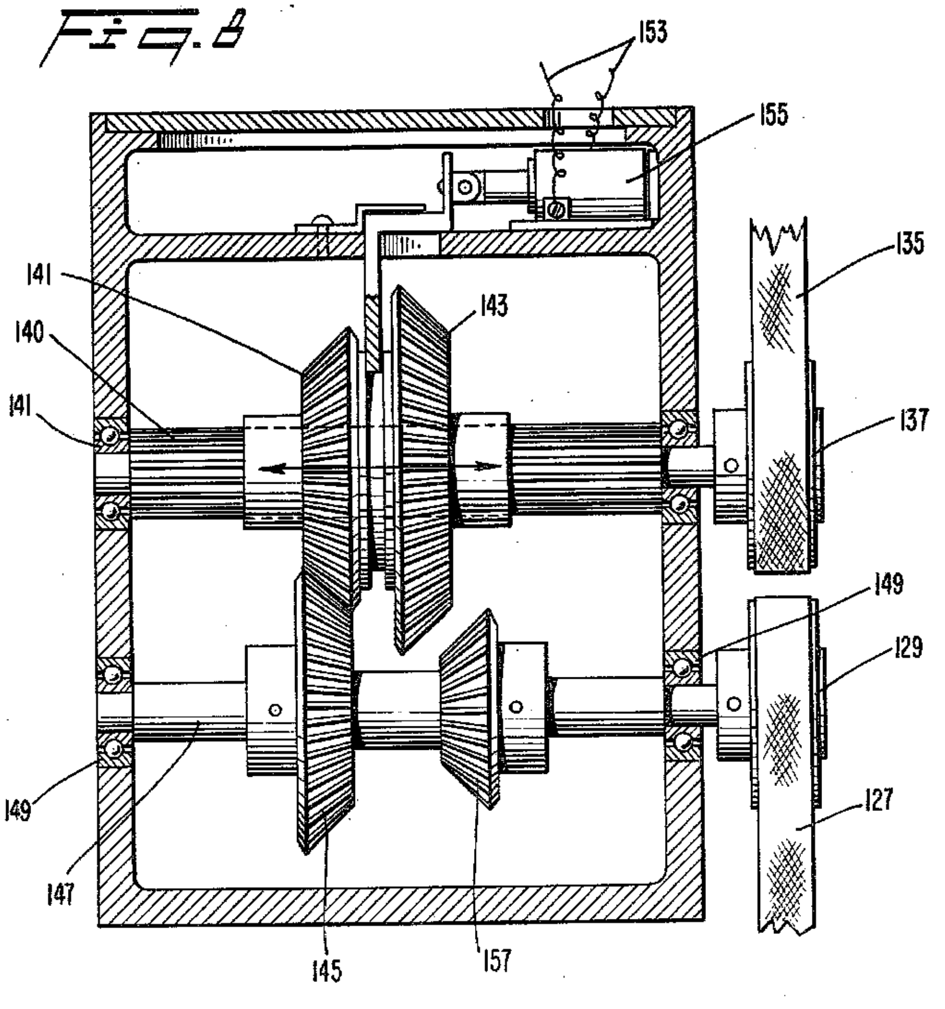
153
155
135
141
140
141
143
137
149
129
127
149
147
145
157

METHOD AND APPARATUS FOR OPERATING AN ENGINE ON COMPRESSED GAS

BACKGROUND AND SUMMARY OF THE PRESENT INVENTION

The present invention relates to a method and apparatus for operating an engine using a compressed gas as the motive fluid. More particularly, the present invention relates to a apparatus for adapting a pre-existing internal combustion engine for operation on a compressed gas.

Air pollution is one of the most serious problems facing the world today. One of the major contributors to air pollution is ordinary internal combustion engine which are used in most motor vehicles today. Various devices, including many items mandated by legislation, have been proposed in an attempt to limit the pollutants which an internal combustion engine exhausts to the air. However, most of these devices have met with limited success and are often both prohibitively expensive and complex. A clean alternative to the internal combustion engine is needed to power vehicles and other machinery.

A compressed gas, preferably air, would provide an ideal motive fluid for a engine since it would eliminate the usual pollutants exhausted from an internal combustion engine. An apparatus for converting an internal combustion engine for operation on compressed air is disclosed in U.S. Pat. No. 3,885,387 issued May 27, 1975 to Simington. The Simington patent discloses an apparatus including a source of compressed air and a rotating valve actuator which opens and closes a plurality of mechanical poppet valves. The valves deliver compressed air in timed sequence to the cylinders of an engine through adapters located in the spark plug holes. However, the output speed of an engine of this type is limited by the speed of the mechanical valves and the fact that the length of time over which each of the valves remains open cannot be varied as the speed of the engine increases.

Another apparatus for converting an internal combustion engine for operation on steam or compressed air is disclosed in U.S. Pat. No. 4,102,130 issued July 25, 1978 to Stricklin. The Stricklin patent discloses a device which changes the valve timing of a conventional four stroke engine such that the intake and exhaust valves open once for every revolution of the engine instead of once every other revolution of the engine. A reversing valve is provided which delivers live steam or compressed air to the intake valves and is subsequently reversed to allow the exhaust valves to deliver the expanded steam or air to the atmosphere. A reversing valve of this type however does not provide a reliable apparatus for varying the amount of motive fluid injected into the cylinders when it is desired to increase the speed of the engine. Further, a device of the type disclosed in the Stricklin patent requires the use of multiple reversing valves if the cylinders in a multi-cylinder engine were to be fired sequentially.

Therefore, it is an object of the present invention to provide a reliable method and apparatus for operating an engine or converting an engine for operation with a compressed gas.

A further object of the present invention is to provide a method and apparatus which is effective to deliver a

constantly increasing amount of compressed gas to an engine as the speed of the engine increases.

A still further object of the present invention is to provide a method and apparatus which will operate an engine using compressed gas at a speed sufficient to drive a conventional automobile at highway speeds.

It is still a further object of the present invention to provide a method and apparatus which is readily adaptable to a standard internal combustion engine to convert the internal combustion engine for operation with a compressed gas.

Another object of the invention is to provide a method and apparatus which utilizes cool expanded gas, exhausted from a compressed gas engine, to operate an air conditioning unit and/or an oil cooler.

These and other objects are realized by a method and apparatus according to the present invention for operating an engine having at least one cylinder and a reciprocating piston therein using compressed gas as a motive fluid. The apparatus includes a source of compressed gas and a distributor connected with the source of the compressed gas for distributing the compressed gas to the at least one cylinder. A valve is provided for admitting the compressed gas to the cylinder when the piston is in approximately a top dead center position within the cylinder. An exhaust is provided for exhausting the expanded gas from the cylinder as the piston returns to approximately the top dead center position.

In a preferred embodiment of the present invention a device is provided for varying the duration of each engine cycle over which the valve remains open to admit compressed gas to the cylinder dependent upon the speed of the engine. In a further preferred embodiment of the present invention, an apparatus for advancing the timing of the opening of the valve is arranged to admit the compressed gas to the cylinder progressively further before the top dead center position of the piston as the speed of the engine increases.

Further features of the present invention include a valve for controlling the amount of compressed gas admitted to the distributor. Also, a portion of the gas which has been expanded in the cylinder and exhausted through the exhaust valve is delivered to a compressor to be recompressed and returned to the source of compressed gas. A gear train is selectively engagable to drive the compressor at different operating speed depending upon the pressure maintained at the source of compressed air and/or the speed of the engine. Still further, a second portion of the exhaust gas is used to cool a lubricating fluid for the engine or to operate an air conditioning unit.

In a preferred embodiment of the present invention, the valve for admitting compressed gas to the cylinder is electrically actuated. The device for varying the duration of each engine cycle over which the intake valve remains open as the speed of the engine increase comprises a rotating element whose effective length increases as the speed of the engine increases such that a first contact on the rotating element is electrically connected to a second contact for a longer period of each engine cycle. The second contact actuates the valve whereby the valve remains in an open position for a longer period of each engine cycle as the speed of the engine increases.

Still further features of the present invention include an adaptor plate for supporting the distributor above an intake manifold of a conventional internal combustion engine after a carburetor has been removed to allow air to enter the cylinders of the engine through the intake manifold and conventional intake valves. Another adaptor plate is arranged over an exhaust passageway of the internal combustion engine to reduce the cross-sectional area of the exhaust passageway.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of a method and apparatus for operating an engine according to the present invention will be described with reference to the accompanying drawings wherein like members bear like reference numerals and wherein:

FIG. 1 is a schematic representation of an apparatus according to the present invention arranged on an engine;

FIG. 2 is a side view of one embodiment of a valve actuator according to the present invention;

FIG. 3 is a cross-sectional view taken along the line 3—3 in FIG. 2;

FIG. 4 is a cross-sectional view of a second embodiment of a valve actuator according to the present invention;

FIG. 5 is a view taken along the line 5—5 in FIG. 4;

FIG. 6 is a cross-sectional view of a third embodiment of a valve actuator according to the present invention;

FIG. 7 is a view taken along the line 7—7 in FIG. 6;

FIG. 8 is a cross-sectional view of a gearing unit to drive a compressor according to the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

With reference to FIG. 1, an engine block **21** (shown in phantom) having two banks of cylinders with each bank including cylinders **20** having pistons **22** reciprocable therein (only one of which is shown in phantom) in a conventional manner. While the illustrated engine is a V-8 engine, it will be apparent that the present invention is applicable to an engine having any number of pistons and cylinders with the V-8 engine being utilized for illustration purposes only. A compressed gas tank **23** is provided to store a compressed gas at high pressure. It may also be desirable to include a small electric or gas compressor to provide compressed gas to supplement the compressed gas held in the tank **23**. In a preferred embodiment, the compressed gas is air which can be obtained from any suitable source.

A line **25** transports the gas withdrawn from the tank **23** when a conventional shut off valve **27** is open. In addition, a solenoid valve **29** preferably operated by a suitable key operated switch (not shown) for the engine is also arranged in the line **25**. In normal operation, the valve **27** is maintained open at all times with the solenoid valve **29** operating as a selective shut off valve to start and stop the engine **21** of the present invention.

A suitable regulating valve **31** is arranged downstream from the solenoid valve **29** and is connected by a linkage **33** to a throttle linkage **35** which is operator actuated by any suitable apparatus such as a foot pedal (not shown). The line **25** enters an end of a distributor **33** and is connected to an end of a pipe **35** which is closed at the other end. A plurality of holes, which are equal to the number of cylinders in the engine **21**, are provided on either side of the pipe **35** along the length of the pipe **35**.

When the present invention is used to adapt a conventional internal combustion engine for operation on compressed gas, an adaptor plate **36** is provided to support the distributor **33** in spaced relation from the usual intake opening in the intake manifold of the engine after a conventional carburetor has been removed. In this way, air is permitted to enter the internal combustion engine through the usual passageways and to be admitted to the cylinders through suitable intake valves (not shown). The adaptor plate **36** is secured to the engine block **21** and the distributor **33** by any suitable apparatus, e.g., bolts.

Each of the holes in the pipe **35** is connected in fluid-tight manner to a single line **37**. Each line **37** carries the compressed gas to a single cylinder **20**. In a preferred embodiment, each of the lines **37** is $\frac{1}{2}$ inch high pressure plastic tubing attached through suitable connectors to the distributor **33** and the pipe **35**. Each of the lines **37** is connected to a valve **39** which is secured in an opening provided near the top of each of the cylinders **20**. In the case of a conversion of a standard internal combustion engine, the valves **39** can be conveniently screwed into a tapped hole in the cylinder **20** typically provided for a spark plug of the internal combustion engine. In a preferred embodiment, the valves **39** are solenoid actuated valves in order to provide a fast and reliable opening and closing of the valves **39**.

Each of the valves **39** is energized by a valve actuator **41** through one of a plurality of wires **43**. The valve actuator **41** is driven by a shaft of the engine similar to the drive for a conventional distributor of an internal combustion engine. That is, a shaft **55** of the valve actuator **41** is driven in synchronism with the engine **21** at one half the speed of the engine **21**.

A first embodiment of the valve actuator **41** (FIGS. 2 and 3) receives electrical power through a wire **45** which is energized in a suitable manner by a battery, and a coil if necessary (not shown) as is conventional in an internal combustion engine. The wire **45** is attached to a central post **47** by a nut **49**. The post **47** is connected to a conducting plate **51** arranged within a housing **53** for the valve actuator **41**. Within the housing **53**, the shaft **55** has an insulating element **57** secured to an end of the shaft **55** for co-rotation therewith when the shaft **55** is driven by the engine **21**. A first end of a flexible contact **59** is continuously biased against the conducting plate **51** to receive electricity from the battery or another suitable source. A second end of the contact **59** is connected to a conducting sleeve **60** which is in constant contact with a spring biased contact **61** which is arranged within the sleeve **60**. The contact **61** is biased by a spring **63** which urges the contact **61** towards a side wall of the housing **53**.

With reference to FIG. 3, a plurality of contacts **65** are spaced from one another and are arranged around the periphery of the housing **53** at the same level as the spring biased contact **61**. Each contact **65** is electrically connected to a post **67** which extends outside of the housing **53**. The number of contacts **65** is equal to the number of cylinders in the engine **21**. One of the wires **43**, which actuate the valves **39**, is secured to each of the posts **67**.

In operation, as the shaft **55** rotates in synchronism with the engine **21**, the insulating element **57** rotates and electricity is ultimately delivered to successive ones of the contacts **65** and wires **43** through the spring biased contact **61** and the flexible contact **59**. In this way, each of the electrical valves **39** is actuated and opened in the proper timed sequence to admit compressed gas to each of the cylinders **20** to drive the pistons **22** therein on a downward stroke.

The embodiment illustrated in FIGS. 2 and 3 is effective to actuate each of the valves 39 to remain open for a long enough period of time to admit sufficient compressed gas to each of the cylinders 20 of the engine 21 to drive the engine 21. The length of each of the contacts 65 around the periphery of the housing 53 is sufficient to permit the speed of the engine to be increased when desired by the operator by moving the throttle linkage 35 which actuates the linkage 33 to further open the regulating valve 31 to admit more compressed gas from the tank 23 to the distributor 33. However, it has been found that the amount of air admitted by the valves 39 when using the first embodiment of the valve actuator 41 (FIGS. 2 and 3) is substantially more than required to operate the engine 21 at an idling speed. Therefore, it may be desirable to provide a valve actuator 41 which is capable of varying the duration of each engine cycle over which the solenoid valves 39 are actuated, i.e., remain open to admit compressed gas, as the speed of the engine 21 is varied.

A second embodiment of a valve actuator 41 which is capable of varying the duration of each engine cycle over which each of the valves 39 remains open to admit compressed gas to the cylinders 20 dependent upon the speed of the engine 21 will be described with reference to FIGS. 4 and 5 wherein members corresponding to those of FIGS. 2 and 3 bear like reference numerals. The wire 45 from the electrical source is secured to the post 47 by the nut 49. The post 47 has a annular contact ring 69 electrically connected to an end of the post 47 and arranged within the housing 53. The shaft 55 rotates at one half the speed of the engine as in the embodiment of FIGS. 2 and 3.

At an upper end of the shaft 55, a splined section 71 slidably receives an insulating member 73. The splined section 71 of the shaft 55 positively holds the insulating member 73 for co-rotation therewith but permits the insulating member 73 to slide axially along the length of the splined section 71. Near the shaft 55, a conductive sleeve 72 is arranged in a bore 81 in an upper surface of the insulating element 73 generally parallel to the splined section 71. A contact 75, biased towards the annular contact ring 69 by a spring 77, is arranged within the conductive sleeve 72 in contact therewith. The conductive sleeve 72 also contacts a conductor 79 at a base of the bore 81.

The conductor 79 extends to the upper surface of the insulating element 73 near an outer periphery of the insulating element 73 where the conductor 79 is electrically connected to a flexible contact 83. The flexible contact 83 selectively engages a plurality of radial contacts 85 arranged on an upper inside surface of the housing 53. A weak spring 87 arranged around the splined section 71 engages a stop member 89 secured on the shaft 55 and the insulating element 73 to slightly bias the insulating element 73 towards the upper inside surface of the housing 53 to ensure contact between the flexible contact 83 and the upper inside surface of the housing 53. As best seen in FIG. 5, the radial contacts 85 on the upper inside surface of the housing 53 are arranged generally in the form of radial spokes extending from the center of the housing 53 with the number of contacts being equal to the number of cylinders 20 in the engine 21. The number of degrees covered by each of the radial contacts 85 gradually increases as the distance from the center of the upper inside surface of the housing 53 increases.

In operation of the device of FIGS. 4 and 5, as the shaft 55 rotates, electricity flows along a path through the wire 45 down through post 47 to the annular contact member 69 which is in constant contact with the spring biased contact 75. The electrical current passes through the conductive sleeve 72 to the conductor 79 and then to the flexible contact 83. As the flexible contact 83 rotates along with the insulating member 73 and the shaft 55, the tip of the flexible contact 83 successively engages each of the radial contacts 85 on the upper inside of the housing 53. As the speed of the shaft 55 increases, the insulating member 73 and the flexible contact 83 attached thereto move upwardly along the splined section 71 of the shaft 55 due to the radial component of the splines in the direction of rotation under the influence of centrifugal force. As the insulating member 73 moves upwardly, the flexible contact 83 is bent such that the tip of the contact 83 extends further radially outwardly from the center of the housing 53 (as seen in phantom lines in FIG. 4). In other words, the effective length of the flexible contact 83 increases as the speed of the engine 21 increases.

As the flexible contact 83 is bent and the tip of the contact 83 moves outwardly, the tip remains in contact with each of the radial contacts 85 for a longer period of each engine cycle due to the increased angular width of the radial contacts with increasing distance from the center of the housing 53. In this way, the length of time over which each of the valves 39 remains open is increased as the speed of the engine is increased. Thus, a larger quantity of compressed gas or air is injected into the cylinders as the speed increases. Conversely, as the speed decreases and the insulating member 73 moves downwardly along the splined section 71, a minimum quantity of air is injected into the cylinder due to the shorter length of the individual radial contact 85 which is in contact with the flexible contact 83. In this way, the amount of compressed gas that is used during idling of the engine 21 is at a minimum whereas the amount of compressed gas which is required to increase the speed of the engine 21 to a level suitable to drive a vehicle on a highway is readily available.

With reference to FIGS. 6 and 7, a third embodiment of a valve actuator 41 according to the present invention includes an arcuate insulating element 91 having a first end pivotally secured by any suitable device such as screw 92 to the shaft 55 for co-rotation with the shaft 55. The screw 92 is screwed into a tapped hole in the insulating element 91 such that a tab 94 at an end of the screw 92 engages a groove 96 provided in the shaft 55. In this way, the insulating element 91 positively rotates with the shaft 55. However, as the shaft 55 rotates faster, a second end 98 of the insulating element 91 is permitted to pivot outwardly under the influence of centrifugal force because of the groove 96 provided in the shaft 55. A spring 93 connected between the second end 98 of the element 91 and the shaft 55 urges the second end of the element 91 towards the center of the housing 53.

A contact 99 similar to the contact 59 (FIG. 2) is arranged such that one end of the contact 99 is in constant contact with the conducting plate 51 located centrally within the housing 53. The other end of the contact 99 engages a conductive sleeve 101 arranged in bore 102. A contact element 95 is arranged in the conductive sleeve 101 in constant contact with the sleeve 101. The bore 102 is arranged generally parallel to the shaft 55 near the second end of the arcuate insulating element **91**. The contact **95** is biased by a spring **97** towards the upper inside surface of the housing **53** for selective contact with each of the plurality of radial contacts **85** which increase in arc length towards the outer peripheral surface of the housing **53** (FIG. **6**).

In operation of the device of FIGS. **6** and **7**, as the shaft **55** rotates the arcuate insulating element **91** rotates with the shaft **55** and the second end **98** of the insulating element **91** tends to pivot about the shaft **55** due to centrifugal force. Thus, as the effective length of the contact **95** increases, i.e., as the arcuate insulating element **91** pivots further outwardly, the number of degrees of rotation over which the contact **95** is in contact with each of the radial contacts **85** on the upper inside surface of the housing **53** increases thereby permitting each of the valves **39** to remain open for a longer period of each engine cycle to admit more compressed gas to the respective cylinder **20** to further increase the speed of the engine **21**.

With reference to FIG. **1**, a mechanical advance linkage **104** which is connected to the throttle linkage **35**, advances the initiation of the opening of each valve **39** such that compressed gas is injected into the respective cylinder further before the piston **22** in the respective cylinder **20** reaches a top dead center position as the speed of the engine is increased by moving the throttle linkage **35**. The advance linkage **104** is similar to a conventional standard mechanical advance employed on an internal combustion engine. In other words, the linkage **104** varies the relationship between the angular positions of a point on the shaft **55** and a point on the housing **53** containing the contacts. Alternatively, a conventional vacuum advance could also be employed. By advancing the timing of the opening of the valves **39**, the speed of the engine can more easily be increased.

The operation of the engine cycle according to the present invention will now be described. The compressed gas injected into each cylinder of the engine **21** drives the respective piston **22** downward to drive a conventional crankshaft (not shown). The movement of the piston downwardly causes the compressed gas to expand rapidly and cool. As the piston **22** begins to move upwardly in the cylinder **20** a suitable exhaust valve (not shown) arranged to close an exhaust passageway is opened by any suitable apparatus. The expanded gas is then expelled through the exhaust passageway. As the piston **22** again begins to move downwardly a suitable intake valve opens to admit ambient air to the cylinder. The intake valve closes and the ambient air is compressed on the subsequent upward movement of the piston until the piston reaches approximately the top dead center position at which time the compressed gas is again injected into the cylinder **20** to drive the piston **22** downward and the cycle begins anew.

In the case of adapting a conventional internal combustion engine for operation on compressed gas, a plurality of plates **103** are preferably arranged over an end of the exhaust passageways in order to reduce the outlet size of the exhaust passageways of the conventional internal combustion engine. In the illustrated embodiment, a single plate having an opening in the center is bolted to the outside exhaust passageway on each bank of the V-8 engine while another single plate having two openings therein is arranged with one opening over each of the interior exhaust passageways on each bank of the V-8 engine. A line **105** is suitably attached to each of the adaptor places to carry the exhaust to an appropriate location. In a preferred embodiment, the exhaust lines **105** are 1½" plastic tubing.

In a preferred embodiment, the exhaust lines **105** of one bank of the V-8 engine are collected in a line **107** and fed to an inlet of a compressor **109**. The pressure of the exhaust gas emmanating from the engine **21** according to the present invention is approximately 25 p.s.i. In this way, the compressor **109** does not have to pull the exhaust into the compressor since the gas exhausted from the engine **21** is at a positive pressure. The positive pressure of the incoming fluid increases the efficiency and reduces wear on the compressor **109**. The exhaust gas is compressed in the compressor **109** and returned through a line **111** and a check valve **113** to the compressed gas storage tank **23**. The check valve **113** prevents the flow of compressed gas stored in the tank **23** back towards the compressor **109**.

A suitable pressure sensor **115** is arranged at an upper end of the tank **23** and sends a signal along a line **117** when the pressure exceeds a predetermined level and when the pressure drops below a predetermined level. The line **117** controls an electrically actuated clutch **119** disposed at a front end of the compressor **109**. The clutch **119** is operative to engage and disengage the compressor **109** from a drive pulley **121**. Also, the signal carried by the line **117** actuates a suitable valve **123** arranged on a compressor housing **125** to exhaust the air entering the compressor housing **125** from the line **107** when the clutch **119** has disengaged the compressor **109** from the drive pully **121**.

In a preferred embodiment, when the pressure is the tank **23** reaches approximately 600 p.s.i., the clutch **119** is disengaged and the compressor **109** is deactivated and the valve **123** is opened to exhaust the expanded gas delivered to the compressor **109** from the line **107** to the atmosphere. When the pressure within the tank **23** drops below approximately 500 p.s.i., the sensor **115** sends a signal to engage the clutch **119** and close the valve **123**, thereby operating the compressor **109** for supplying the tank **23** with compressed gas.

The pulley **121** which drives the compressor **109** through the clutch **119** is driven by a belt **127** which is driven by a pulley **129** which operates through a gear box **131**. With reference to FIGS. **1** and **8**, a second pulley **133** on the gear box is driven by a belt **135** from a pulley **137** arranged on a drive shaft **139** of the engine **21**. The pulley **137** drives a splined shaft **140** which has a first gear **141** and a second larger gear **143** arranged thereon for rotation with the splined shaft **140**. The splined shaft **140** permits axial movement of the gears **141** and **143** along the shaft **140**.

In normal operation (as seen in FIG. **8**), the first gear **141** engages a third gear **145** arranged on a shaft **147** which drives the pulley **129**. The shafts **140** and **147** are arranged in suitable bearings **149** arranged at each end thereof. When the speed of the engine **21** drops below a predetermined level, a suitable sensor **151** responsive to the speed of the drive shaft **139** of the engine **21** generates a signal which is transmitted through a line **153** to a solenoid actuator **155** arranged within the gear box **131**. The solenoid actuator **155** moves the first and second gears **141**, **143** axially along the splined shaft **140** to the right as seen in FIG. **8** such that the second, larger gear **143** engages a fourth smaller gear **157** which is arranged on the shaft **147**. The ratio of the second gear **143** to the fourth gear **157** is preferably approximately 3 to 1.

In this way, when the speed of the engine **21** drops below the predetermined level as sensed by the sensor **151** (which predetermined level is insufficient to drive the compressor **109** at a speed sufficient to generate the 500–600 pounds of pressure which is preferably in the tank **23**), the solenoid actuator **155** is energized to slide the gears **143**, **141** axially along the splined shaft **140** so that the second, larger gear **143** engages the fourth, smaller gear **157** to drive the pulley **129** and hence the compressor **109** at a higher rate of speed to generate the desired pressure. When the speed of the engine increases above the predetermined level, in a preferred embodiment approximately 1500 rpm, the solenoid actuator **155** is deactivated by the sensor **151** thereby moving the gears **143** and **141** to the left as seen in FIG. **8** such that the first gear **141** re-engages with the third gear **145** to effectuate a 1 to 1 ratio between the output shaft **139** of the engine **21** and the pulley **129**.

The other bank of the V-8 engine has its exhaust ports arranged with adapter plates **103** similar to those on the first bank. However, the exhaust from this bank of the engine **21** is not collected and circulated through the compressor **109**. In a preferred embodiment, a portion of the exhaust is collected in a line **159** and fed to an enlarged chamber **161**. A second fluid is fed through a line **163** into the chamber **161** to be cooled by the cool exhaust emmanating from the engine **21** in the line **159**. The second fluid in the line **163** may be either transmission fluid contained in a transmission associated with the engine **21** or a portion of the oil used to lubricate the engine **21**. A second portion of the exhaust from the second bank of the V-8 engine is removed from the line **159** in a line **165** and used as a working fluid in an air conditioning system or for any other suitable use.

It should be noted that the particular arrangement utilized for collecting and distributing the gas exhausted from the engine **21** would be determined by the use for which the engine is employed. In other words, it may be advantageous to rearrange the exhaust tubing such that a larger or smaller percentage of the exhaust is routed through the compressor **109**. It should also be noted that since the exhaust lines **105** are plastic tubing, a rearrangement of the lines for a different purpose is both simple and inexpensive.

In operation of the engine of the present invention, the engine **21** is started by energizing the solenoid valve **29** and any suitable starting device (not shown), e.g., a conventional electric starter as used on an internal combustion engine. Compressed gas from the full tank **23** flows through the line **25** and a variable amount of the compressed gas is admitted to the distributor **33** by controlling the regulator valve **31** through the linkage **33** and the operator actuated throttle linkage **35**. The compressed gas is distributed to each of the lines **37** which lead to the individual cylinders **20**. The compressed gas is admitted to each of the cylinders **20** in timed relationship to the position of the pistons within the cylinders by opening the valves **39** with the valve actuator **41**.

When it is desired to increase the speed of the engine, the operator moves the throttle linkage **35** which simultaneously admits a larger quantity of compressed gas to the distributor **33** from the tank **23** by further opening the regulator valve **31**. The timing of the valve actuator **41** is also advanced through the linkage **104**. Still further, as the speed of the engine **21** increases, the effective length of the rotating contact **83** (FIG. **4**) or **95** (FIG. **6**) increases thereby electrically contacting a wider portion of one of the stationary radial contacts **85** to cause each of the valves **39** to remain open for a longer period of each engine cycle to admit a larger quantity of compressed gas to each of the cylinders **20**.

As can be seen, the combination of the regulating valve **31**, the mechanical advance **104**, and the valve actuator **41**, combine to produce a compressed gas engine which is quickly and efficiently adaptable to various operating speeds. However, all three of the controls need not be employed simultaneously. For example, the mechanical advance **104** could be utilized without the benefit of one of the varying valve actuators **41** but the high speed operation of the engine may not be as efficient. By increasing the duration of each engine cycle over which each of the valves **39** remains open to admit compressed gas to each of the cylinders **20** as the speed increases, conservation of compressed gas during low speed operation and efficient high speed operation are both possible.

After the compressed gas admitted to the cylinder **20** has forced the piston **22** downwardly within the cylinder to drive the shaft **139** of the engine, the piston **22** moves upwardly within the cylinder **20** and forces the expanded gas out through a suitable exhaust valve (not shown) through the adapter plate **103** (if employed) and into the exhaust line **105**. The cool exhaust can then be collected in any suitable arrangement to be compressed and returned to the tank **23** or used for any desired purpose including use as a working fluid in an air conditioning system or as a coolant for oil.

When using the apparatus and method of the present invention to adapt a ordinary internal combustion engine for operation with compressed gas it can be seen that considerable savings in weight are achieved. For example, the ordinary cooling system including a radiator, fan, hoses, etc. can be eliminated since the compressed gas is cooled as it expands in the cylinder. In addition, there are no explosions within the cylinder to generate heat. Further reductions in weight are obtained by employing plastic tubing for the lines which carry the compressed gas between the distributor and the cylinders and for the exhaust lines. Once again, heavy tubing is not required since there is little or no heat generated by the engine of the present invention. In addition, the noise generated by an engine according to the present invention is considerably less than that generated by an ordinary internal combustion engine since there are no explosions taking place within the cylinders.

The principles of preferred embodiments of the present invention have been described in the foregoing specification. However, the invention which is intended to be protected is not to be construed as limited to the particular embodiments disclosed. The embodiments are to be regarded as illustrative rather than restrictive. Variations and changes may be made by others without departing from the spirit of the invention. Accordingly, it is expressly intended that all such variations and changes which fall within the spirit and the scope of the present invention as defined in the appended claims be embraced thereby.

What is claimed is:

1. An apparatus for operating an engine having at least one cylinder and a reciprocating piston therein comprising:

a source of compressed gas;

- distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;
- valve means for admitting the compressed gas to the at least one cylinder when the piston is in approximately a top dead center position within the cylinder;
- altering means for increasing the duration of each engine cycle over which the valve means admits compressed gas to the at least one cylinder as the speed of the engine increases; and
- exhaust means for exhausting gas as the piston subsequently approaches approximately the top dead center position.

2. The apparatus of claim **1** further comprising control means for controlling the amount of compressed gas admitted to the distributor means.

3. The apparatus of claim **1** wherein the valve means is a solenoid valve secured in an opening in the cylinder above the level of the piston at the top dead center position.

4. The apparatus of claims **1** or **2** further comprising means for advancing the timing of the valve means as the speed of the engine increases such that compressed gas is admitted progressively further before the top dead center position as the speed of the engine increases.

5. The apparatus of claim **4** wherein the means for advancing the timing comprises a mechanical linkage connected to an operator actuated accelerator linkage.

6. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is compressed in a compressor driven by an output shaft of the engine and is returned to the source of compressed gas.

7. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is used to cool transmission fluid for a transmission associated with the engine.

8. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is used as a working fluid in an air conditioning system.

9. The apparatus of claim **6** further comprising first gearing means interposed between the output shaft of the engine and the compressor for increasing the speed at which the compressor is driven.

10. The apparatus of claim **6** further comprising clutch means attached to the compressor both for disengaging the compressor from the output shaft of the engine when a first predetermined pressure at the source of compressed gas is exceeded and for engaging the compressor with the output shaft of the engine when the pressure at the source of compressed gas drops below a second predetermined pressure.

11. The apparatus of claim **9** further comprising means for both disengaging the first gearing means when a predetermined speed of the engine is exceeded and engaging a second gearing means for driving the compressor at a speed slower than the first gearing means when the predetermined speed of the engine is exceeded.

12. The apparatus of claim **1** wherein the valve means is electrically actuated and wherein the altering means comprises:

- a rotating member timed with the at least one cylinder and arranged within a housing;
- first and second contacts arranged on a first end of the rotating member and on an inside surface of the housing, respectively;
- means for increasing the distance of the first contact from the rotational axis of the rotating member as the speed of the engine increases such that the first contact moves radially outwardly within the housing; and
- said second contact presenting a longer arc length to the first contact as the distance of the first contact from the rotational axis of the rotating member increases.

13. The apparatus of claim **12** wherein the rotating member comprises an arcuate arm and wherein the means for increasing the distance of the first contact comprises pivotally mounting a second end of the arcuate arm about the axis of rotation of the rotating member and spring means for biasing the first end of the arcuate arm towards a radially inward position whereby the first end of the arcuate arm pivots radially outwardly as the speed of the engine increases.

14. The apparatus of claim **12** wherein the rotating member is axially slidably received on a rotating shaft for co-rotation therewith, said shaft having splines with a radial component in the direction of rotation, and wherein the first contact comprises a flexible contact located on an upper surface of the rotating member, said flexible contact being biased against the inside surface of the housing which carries the second contacts whereby as the speed of the engine increases the rotating member is urged axially along the splined shaft towards the inside surface of the housing such that the flexible contact is forced radially outwardly along the inside surface.

15. The apparatus of claim **12** wherein the second contact comprises of radially extending conductor arranged on an upper inside surface of the housing, said conductor increasing in arc length as the conductor extends radially outwardly from a central portion of the housing.

16. An apparatus for adapting an internal combustion engine for operation with compressed gas, the internal combustion engine having at least one cylinder, a piston reciprocable within the at least one cylinder, intake and exhaust means disposed in the at least one cylinder, and a tapped hole in the at least one cylinder adapted to receive a spark plug, the apparatus comprising:

- a source of compressed gas;
- distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;
- valve means arranged in the tapped hole for admitting the compressed gas to the at least one cylinder when the piston is in approximately a top dead center position within the cylinder; and
- altering means for increasing the duration of each engine cycle over which the valve means remains open to admit the compressed gas as the speed of the engine increases.

17. An apparatus as in claim **16** further comprising first adapter plate means for supporting the distributor means above an intake manifold of the engine, which adaptor plate means allows ambient air to enter through the intake manifold.

18. The apparatus of claim **16** further comprising second adapter plate means for reducing the exit area of the exhaust means.

19. A method of operating an engine on compressed gas, said engine having at least one cylinder and a piston reciprocable therein comprising the steps of:

delivering compressed gas from a source to a distributor;

distributing the compressed gas to the at least one cylinder;

admitting compressed gas to the at least one cylinder through an intake valve when the piston is at approximately a top dad center position;

increasing the duration of each engine cycle over which compressed gas is admitted to the at least one cylinder as the engine speed increases; and

exhausting the remaining gas when the piston subsequently reaches approximately the top dead center position.

20. The method of claim **19** further comprising the step of controlling the amount of compressed gas which is delivered to the distributor.

21. The method of claim **19** further comprising the step of advancing the timing of the opening of the intake valve as the speed of the engine increases.

22. An apparatus for operating an engine having at least one cylinder and a piston reciprocable therein on compressed gas comprising:

a source of compressed gas;

distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;

electrically actuated valve means secured in an opening in the at least one cylinder for selectively admitting compressed gas to the at least one cylinder when the piston is in approximately a top dead center position; and

means for advancing the timing of the valve means as the speed of the engine increases whereby compressed gas is admitted progressively further before the top dead center position as the speed of the engine increases.

* * * * *

PRELIMINARY PLANS TO RUN YOUR CAR ON TAP WATER!

IT ALSO WORKS ON YOUR
TRUCK / RV / MOTORCYCLE / AIRPLANE (ETC)

Will This Work?

These plans were sent to the Spirit of Ma'at anonymously, from someone who does not want his or her name printed (for obvious reasons).
We have had them checked by an expert who believes that they are real.
We also have talked with another individual who has patented a similar device, and we know by personal experience that the technology is sound.
So although we cannot guarantee it, we believe these plans will enable you to build a car that runs on water. If you test it out, though, do as the writer suggests and use an old car that doesn't represent a loss of value if you can't make it work. And leave everything intact so that you can always reconnect back to gas if you have to.

But if you do get it working, please send us your experience for our readers. You could be a national hero and help save our country and our world.

We know for certain that an automobile will run on water. So this could be an interesting project for you mechanical types, with a great reward of never having to purchase gasoline for the rest of your life - and helping humanity at the same time.

Re The Need to Rustproof Your Exhaust System

It is possible to make a hybrid of both gas and water (a system that is being tested now in Mexico), which would eliminate the need to open the head and remove the exhaust system. Just a thought. It takes only a small amount of gas to keep the system dry.

The text sent from the anonymous individual was edited slightly for better reading. The following is his/her words and drawings, which has been given into the public domain.
- Spirit of Ma'at LLC

Introduction

It is suggested you try this out to begin with on a second vehicle you own, one that you don't need to live with everyday, until you perfect this technology.

Do-it-yourself plans allow the individual (that's you and me, folks) to make a difference. This is the easiest and lowest-cost way to convert your car to run on (relatively) free energy.

Now, with existing technology, anyone can stand up and make a difference by reducing the local automotive pollution, eliminate gasoline expenses, help restore our atmosphere, and breathe a little easier.

In putting these plans into operation, you will be making use of your entire existing system except for the fuel tank and the catalytic converter.

The Plan

Build and install a low-cost alternative method for running your vehicle (internal combustion engine) on tap water, using off-the-shelf components.

This is simply an efficient way to convert ordinary tap water into gaseous hydrogen and oxygen, and then burn these vapors in the engine, instead of gasoline.

This "minisystem" runs easily from your existing battery and electrical system, and it plugs into your carburetor with simple off-the-shelf fittings.

You will be installing a plastic water tank, a control circuit, a reaction chamber, a hi-pressure carb/FI fitting, and 3 gauges, and then hooking into your existing carb/FI.

The simplicity comes from its being an "on-demand" system requiring no fancy storage or plumbing. You crank the gas pedal or throttle, and you electrically create more vapor for immediate consumption, on demand; low-high flow rate as needed, from idle to maximum power. The only real change is that you are using tap water as fuel, instead of the traditional petroleum-based fuel.

Given a choice, which way would you choose?

Frequently Asked Questions

Q: Does it really work ?
A: Yes; this is well-established technology dating back to stainless steel. But be sure to follow these instructions using the proper mechanical and electrical assembly techniques, as this plan incorporates the best qualities of several techniques.

Q: How does it qualify as "free energy"?
A: If you're paying someone for the water you use, then it is not strictly free. But the alternative is to keep buying into expen$ive ga$oline and its resultant hydrocarbon pollution.

Q: Is it safe?
A: Technically, it is safer than running on fossil fuel because you are no longer choking on your own emissions (health-wise). In general, it is practically as safe as your current gasoline arrangement. You will be installing a few simple safety devices, using current automotive standards.

Q: What kind of performance can I expect?
A: Properly adjusted, your modified vapor-only fuel system will run cooler, and at a modestly higher power level. The mileage performance expected from this design ranges from 50-300 mpg (of water), depending on your adjusting skills.

Q: Can I do the modification myself?
A: Why not? If you don't have any mechanical skills, and you know someone with basic mechanical and/or electrical skills, you can even delegate some of the construction. If you are using a fuel-injected engine, you may have to get a mechanic's opinion. [There will have to be an adapter inserted into the fuel-injection system, just as you would have to do if you were going to run on propane, hydrogen, or natural gas. Ed.)

Q: What is the environmental impact that my vehicle will have?
A: It will be producing H20 steam (water vapor) and unburnt O2 (Oxygen). Hence, it will be cleaning the environment, rather than dumping nauseous toxins into it. Plus you will be helping to save our dwindling supply of atmospheric oxygen. Any excess vapor in the reaction becomes either steam or oxygen. You can also expect to be receiving more than casual interest from those around you.

Q: Isn't this really a steam engine?
A: No. Really. Exceedingly high temperature and pressure are not used. This is strictly an internal-combustion engine (burning orthohydrogen) with residual steam in the exhaust as a by-product.

There are a few things you should know about gasoline:

Gasoline as a fuel is not necessary; it is optional.

Gasoline versus Water

There is a lot of thermochemical energy in gasoline, but there is even more energy in water. The DOE (Department of Energy) has quoted about 40%, so it is probably much more than that.

Most people are unaware that "internal combustion" is defined as "a thermo-vapor process" - as in "no liquid in the reaction." Most of the gasoline in a standard internal combustion engine is actually consumed, (cooked, and finally, broken down) in the catalytic converter after the fuel has been not-so-burnt in the engine. Sadly, this means that most of the fuel we use in this way is used only to cool down the combustion process, a pollution-ridden and inefficient means of doing that.

How It Works

Exceedingly simple. Water is pumped as needed to replenish and maintain the liquid level in the chamber. The electrodes are vibrated with a 0.5-5A electrical pulse which breaks $2(H_2O) \Rightarrow 2H_2 + O_2$. When the pressure reaches say 30-60 psi, you turn the key and go. You step on the pedal, you send more energy to the electrodes, and thus more vapor to the cylinders; i.e. fuel vapor on demand.

You set the idle max-flow rate to get the most efficient use of power, and you're off to the races.

In the big picture, your free energy is coming from the tap water in an open system, as the latent energy in the water is enough to power the engine and hence drive the alternator and whatever belt-driven accessories. And the alternator is efficient enough to run the various electrical loads (10 - 20 amps), including the additional low current to run this vapor reaction. No extra batteries are required.

STEP BY STEP CONSTRUCTION (Please refer to diagrams at the end of this document)

OVERVIEW - Here is the suggested sequence of steps:

1. Install the CHT (or EGT) gauge and measure your current operating temp range (gasoline), for comparison.

2. Build and test the controller to verify the correct pulse output.

3. Build the reaction chamber and test it with the controller (i.e pressure out).

4. Install the tank, controller, chamber, and pressure fittings.

5. Run engine and adjust the control circuit as necessary for best performance.

6. Install the stainless steel valves and get the pistons/cylinders coated with ceramic.

7. Coat the exhaust system with ceramic without the catalytic converter (or let it rust out and then replace the whole dang thang with stainless steel pipe sections).

YOU WILL NEED

plastic water tank with pump and level sensor.
control circuit, wiring, connectors, and epoxy.
reaction chamber with electrodes and fittings.
3/8" stainless steel flex-tubing, fittings and clamps.
carb/FI vapor-pressure fitting kit. - pressure, CHT (or EGT), & level gauges.
stainless steel valves.
copper mesh junction.
ceramic surface treatment for cylinders & pistons.
stainless steel or ceramic treated exhaust assembly.

BASIC TOOLS

drill, screwdriver and pliers
hole cutter
wire-wrap, solder-iron and clippers
DVM and oscilloscope.

REACTION CHAMBER

Construct as shown in the diagrams. Use a section of 4" PVC waste pipe with a threaded screw-cap fitting on one end and a standard end-cap at the other. Make sure to drill-and-epoxy or tap threads thru the PVC components for all fittings. Set and control the water level in the chamber so that it well submerses the pipe electrodes; yet leave some headroom to build up the hydrogen/oxygen vapor pressure. Use stainless steel wires inside the chamber or otherwise use a protective coating; use insulated wires outside. Ensure that the epoxy perfects the seal, or otherwise lay down a bead of water-proof silicone that can hold pressure.

The screw fitting may require soft silicone sealant, or a gasket; its purpose is to hold pressure and allow periodic inspection of the electrodes. No leaks, no problems. Make sure you get a symmetric 1-5mm gap between the 2 stainless steel pipes. The referenced literature suggests that the closer to 1mm you get, the better. You will want to get your chamber level sensor verified before you epoxy the cap on.

Make your solder connections at the wire/electrode junctions nice, smooth, and solid; then apply a water-proof coating, e.g. the epoxy you use for joining the pipes to the screw cap. This epoxy must be waterproof and be capable of holding metal to plastic under pressure. You will want to get your chamber level sensor verified before you epoxy the cap on.

CONTROL CIRCUIT

The diagrams show a simple circuit to control and drive this mini-system. You are going to make a 'square-pulse' signal that 'plays' the electrodes like a tuning fork; which you can watch on an oscilloscope. The premise given by the literature is: the faster you want do go down the road, the 'fatter' you make the pulses going into the reaction chamber. Duty cycle will vary with the throttle in the vicinity of 90%MARK 10%SPACE (OFF/ON).

There is nothing sacred about how the pulse waveform is generated; there are many ways to generate pulses, and the attached diagrams show a few. The diagram shows the NE555-circuit approach from the referenced patent. The output switching transistor must be rated for 1-5 amps @ 12VDC (in saturation).

Go with a plan that works for you or your friendly neighborhood technoid or mechanic, and go get all the circuit elements from your local electronics store, such as Radio-Shack or Circuits-R-Us, including the circuit board, IC sockets, and enclosure/box.

DigiKey has better selection, service, and knowledge; plus they have no minimum order. Be sure to use a circuit board with a built-in ground plane, and to accommodate room for mounting 2 or 3 of the gauges. Mounting the reaction chamber in the engine compartment will require running a stub to your pressure gauge where you can watch it.

You can easily make 30-gauge wire-wrap connections between the socket pins and thru-hole discrete components having wire leads. Also make sure to get spec sheets on any IC you use. More details of the best circuits to use will be announced pending prototype testing. You will want to get your chamber level sensor verified before you epoxy the cap on.

Throttle Control

If you have a throttle position sensor, you should be able to access the signal from the sensor itself OR from the computer connector. This signal is input to the circuit as the primary control (i.e. throttle level = pulse width = vapor rate).

If you don't have such a signal available, you will have to rig a rotary POT (variable resistor) to the gas linkage (i.e. coupled to something at the gas pedal or throttle cable running to the carb or FI. If you make the attachment at the carb/FI, be sure to use a POT that can handle the engine temp cycles. Don't use a cheezy-cheapy POT; get one rated for long life and mechanical wear; mount it securely to something sturdy and stationary that will not fall apart when you step on the gas.

Control Range. The full throttle RANGE (idle-max) MUST control the vapor rate, i.e. pulse-width (duty). The resistor values at the throttle signal must allow the throttle signal voltage, say 1-4 Volt swing, to drive the VAPOR RATE. You will be using this voltage swing to generate a 10% ON 'square' pulse. The patent implies using a 'resonant' pulse in the 10-250 KHz frequency range; but it is not explicitly stated so.

In this circuit, you will simply tune to whatever frequency makes the most efficient vapor conversion. You will have to get into the specs for each IC you use, to insure you connect the right pins to the right wires, to control the frequency and pulse width. You can use spare sockets to try out different discrete component values. Just keep the ones that are spec-compatible in the circuit, and get the job done.

You crank up the throttle signal and put more electrical energy (fatter pulses) into the electrodes; verify you can get 10% duty on the scope (2 - 100 usec on the horizontal time-base). Your averaging DVM will display the 90%-10% DC voltage across the output transistor (Vce or Vds or Output to Ground). Set and connect DVM in the supply current and measure .5 - 5 amps, without blowing the DVM fuse. Now verify that you got everything you wanted.

Verify your wiring connections using your DVM as a continuity detector. Check your wiring 1 at a time and yellow line your final schematic as you go. You can best use board-mount miniature POTs for anything you want to set-and-forget. The LEDs are there to give you a quick visual check of normal vs abnormal operation of your new creation. You will want to get your chamber level sensor verified before you epoxy the cap on.

CARB/FI CONNECTION

The diagram also shows that fittings are required to the carb/FI l. There are ready-made kits (such as by Impco) available for making your pressure fittings to the carburetor or fuel-injector as the case may be. You will necessarily be sealing the built-in vents and making a 1-way air-intake.

The copper mesh comprises the inadvertent backfire' protection for the reaction chamber. Make sure that all vapor/duct junctions are air-tight and holding full pressure without leakage. Your new 'system' is considered successful and properly adjusted when you get the full power range at lower temp and minimum vapor flow without blowing the pressure safety valve.

CHT (or EGT)

Monitor your engine temp with the CHT (cylinder head temp) or EGT (exhaust gas temp) instead of your original engine temp indicator (if any). Your existing gauge is too slow for this application and will not warn you against overheating until after you have burnt something. Make sure that your engine runs no hotter than in the gasoline arrangement. VDO makes a CHT gauge with a platinum sensor that fits under your spark plug against the cylinder head (make sure it is really clean before you re-install your spark plug (as this is also an electrical ground).

ENGINE/EXHAUST TREATMENT

Get the valves replaced with stainless steel ones and get the pistons/cylinders ceramic-treated ASAP when you have successfully converted and run your new creation. Do not delay as these items will rust, either by sheer use or by neglect (i.e. letting it sit). You could make max use of your current exhaust system by using it with your new deal until it rusts through, then have your mechanic or welder friend to fit a stainless steel exhaust pipe (no catalytic converter is required). But it could be easier and cheaper to send your existing exhaust system out for the ceramic treatment, and then simply re-attach it to the exhaust ports.

GENERAL

1. Do not discard or remove any of the old gasoline setup components, e.g. tank, carb/FI, catalytic converter, unless necessary. Better to always leave an easy way to revert back to something that at least runs, just in case. Some people are leaving their gasoline setup completely intact, and switching back and forth at will, just to have a backup plan.

2. Set your throttle circuit so that you get minimum vapor flow at idle, and maximum vapor flow at full power without blowing the pressure relief valve. In this way, you control how 'lean' your mixture is by the strength of the pulse (i.e. “fatness” at the optimum pulse frequency).

3. If you just don't get enough power (at any throttle setting), it means that you need to (1) change the pulse frequency, (2) change the gap between the electrodes, (3) change the size (bigger) electrodes, or (4) make a higher output pulse voltage (last resort). Always use an output transistor, such as a MOSFET, that is rated for the voltage and current you need to get the job done. OK so you might have to play around with it some. Isn't that where all the Fun is anyhow?

4. If you get any engine knock our loud combustions (not compensated by adjusting the timing), it means that you need to install an additional coil in the chamber, and drive the coil with an additional pulse signal (about 19 Hz on the .1sec time base (see diagram). Here,

you will be slowing down the burn rate just enough so that the vapors burn thru out the power stroke of the piston. Be sure to include a board-mount POT to set the correct strength of this 2nd pulse signal into the coil. This is a stainless steel coil of about 1500 turns (thin wire) that you can arrange like a donut around the center pipe (but NOT touching either electrode), directly over the circular 1-5mm gap. You want no knocking at any power/throttle setting; smooth power only, but also no excess hydrogen leftover from the combustion.

5. Build the canister(s) as tall as you can without compromising your ability to mount them conveniently near the dash panel, or in the engine compartment, as the case may be. This way, you can always make the electrodes bigger, if necessary without undue hardship. Remember that anything in the engine compartment should be mounted in a bullet-proof, vibration and temperature tolerant fashion.

6. If you have to drill a thru-hole for wiring or plumbing thru metal, make sure to also install a grommet for protection against chafing. Always watch your chamber pressure range from IDLE (15-25 psi) - FULL POWER (30-60 psi). Set your safety-pressure relief-valve to 75 psi and make sure it's rated for much higher.

7. Shut OFF the power switch and pull over if there is any malfunction of the system. Your engine will last longest when it still develops FULL POWER+ at some minimum temperature that we are sure you can find, by leaning back the Royal Vapor Flow and/or by making use of the water-vapor cooling technique (see diagram). Keep good mpg performance records, and periodic maintenance/inspection. Keep it clean; save some money; clean the air; heal the planet; happy motoring; tell a friend; enjoy your freedom and self-empowerment.

8. There lacks documented material for perfecting this vapor system thru a fuel injector; there may be some details you will discover on your own as working prototypes progress. For example, you may be restricted to inject the hydrogen/oxygen vapor without any water vapor, as it may rust the injectors. If engine temp and CHT is a problem, then you will want to re-think your plan, e.g. ceramic-coating the injectors. There is always “replacing the FI system with a Carb.”

9. If you install the water-vapor system (for lower operating temp/stress), you will want to lean the mixture (vapor/air) for minimum vapor flow rate to achieve any given throttle position (idle - max). Make sure that you get a minimum flow for IDLE and a modestly sufficient flow for MAX, that does the cooling job without killing the combustion.

10. If you cannot find stainless steel pipe combinations that yield the 1-5mm gap, you can always regress back to alternating plates of +/- electrodes.

11. If you are concerned about the water freezing in your system, you can (a) add some

98% isopropyl alcohol and re-adjust the pulse frequency accordingly; or (b) install some electric heating coils.

12. Do not let ANYONE ever compromise your dream, your freedom, your independence or your truth.

REFERENCES

Stephen Chambers 'Apparatus for Producing Orthohydrogen and/or Parahydrogen' US Patent 6126794, uspto.gov
Stanley Meyer 'Method for the Production of a Fuel Gas' US Patent 4936961, uspto.gov
Creative Science & Research, 'Fuel From Water', fuelless.com
Carl Cella “A Water-Fuelled Car” Nexus Magazine Oct-Nov 1996
Peter Lindemann “Where in the World is All the Free Energy”, free-energy.cc <http://www.free-energy.cc/>
George Wiseman “The Gas-Saver and HyCO Series” eagle-research.com <http://www.eagle-research.com/>
C. Michael Holler “The Dromedary Newsletter” and “SuperCarb Techniques”
Stephen Chambers “Prototype Vapor Fuel System” xogen.com <http://www.xogen.com/>

COMMON LAW COPYRIGHT #285714: All rights to the use and duplication of these plans are hereby reserved for the People, in their efforts to heal and restore the environment. Dare to express your uniqueness and environmental ideals. This technology is an exercise in responsible self-determination.

DISCLAIMER: The Spirit of Ma'at LLC and the Spirit of Ma'at ezine and the author of this document assumes no liability for the use or misuse of this information; which is made available as public-domain information and free of charge, for the purposes of education, ecology, health, well-being, freedom, liberty, and pursuit of happiness.

CONVERT YOUR ENGINE TO BURN HYDROGEN AND OXYGEN VAPOR FROM WATER....... ON DEMAND + POLLUTION-FREE

FIGURE 1

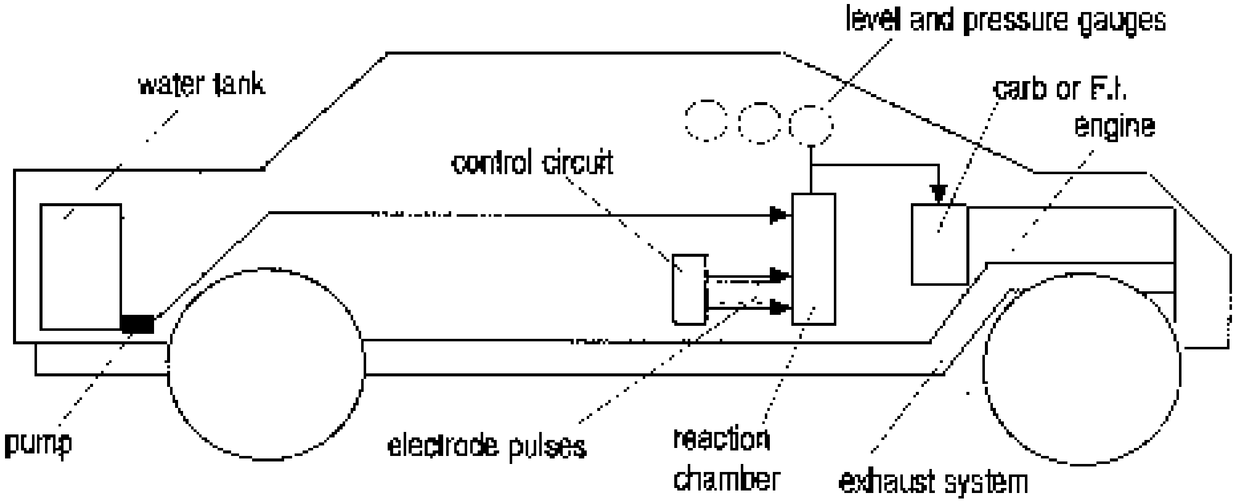
level and pressure gauges
water tank
carb or F.I.
engine
control circuit
pump
electrode pulses
reaction chamber
exhaust system

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

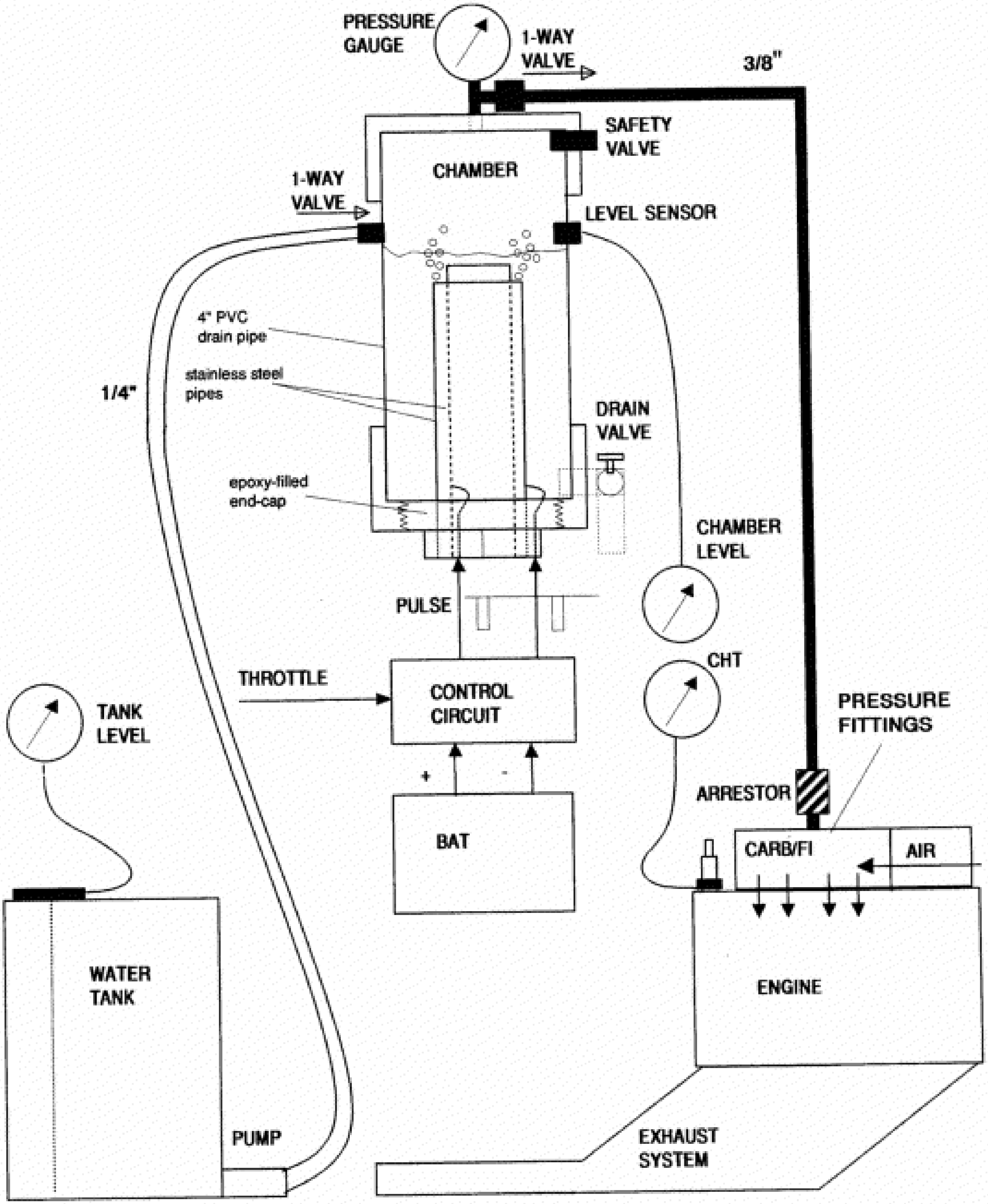
PRESSURE GAUGE
1-WAY VALVE
3/8"
SAFETY VALVE
CHAMBER
1-WAY VALVE
LEVEL SENSOR
4" PVC drain pipe
stainless steel pipes
1/4"
DRAIN VALVE
epoxy-filled end-cap
CHAMBER LEVEL
PULSE
CHT
THROTTLE
CONTROL CIRCUIT
PRESSURE FITTINGS
TANK LEVEL
+
-
ARRESTOR
BAT
CARB/FI
AIR
WATER TANK
ENGINE
PUMP
EXHAUST SYSTEM

FIGURE 2 BLOC DIAGRAM

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

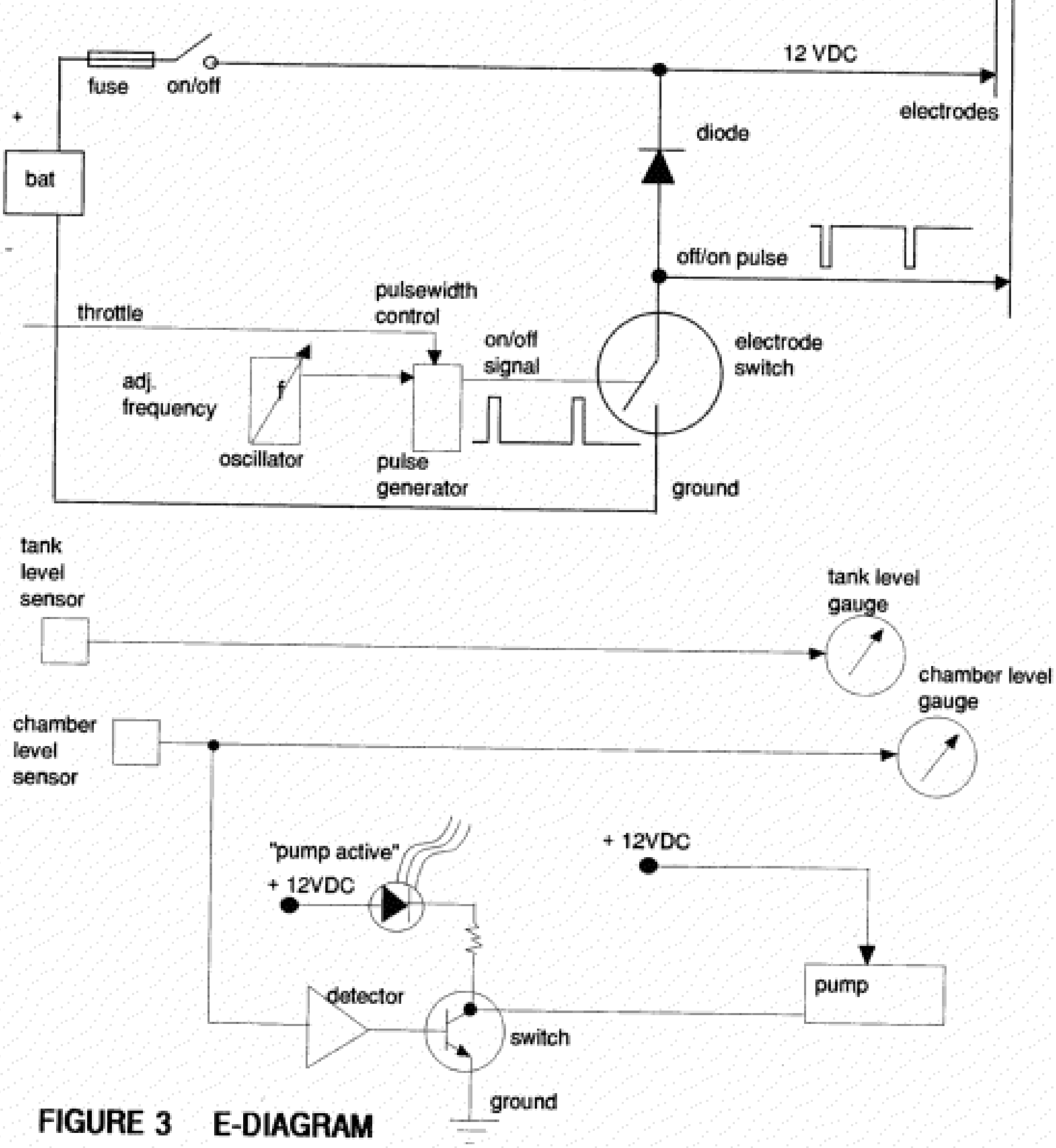
12 VDC
fuse
on/off
+
bat
-
electrodes
diode
off/on pulse
throttle
pulsewidth control
on/off signal
electrode switch
adj. frequency
f
oscillator
pulse generator
ground
tank level sensor
tank level gauge
chamber level gauge
chamber level sensor
"pump active"
+ 12VDC
+ 12VDC
pump
detector
switch
ground

FIGURE 3 E-DIAGRAM

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

REV. A

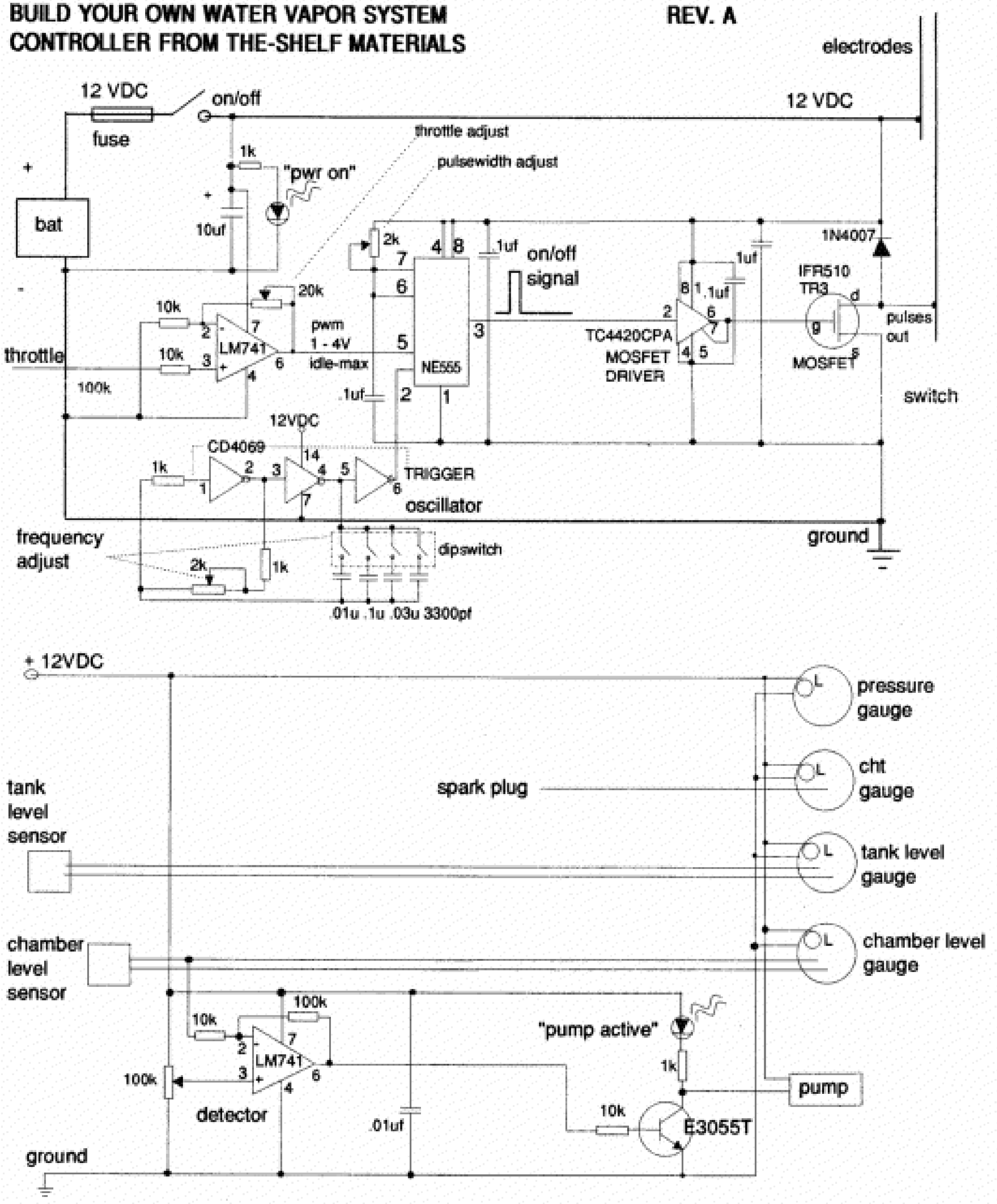
electrodes
12 VDC
on/off
fuse
12 VDC
+
bat
-
1k
"pwr on"
+
10uf
throttle adjust
pulsewidth adjust
2k
7
4
8
.1uf
on/off
signal
1N4007
.1uf
IFR510
TR3
d
20k
6
8
1
.1uf
10k
7
pwm
2
6
g
pulses
out
2
-
LM741
1 - 4V
5
3
TC4420CPA
7
throttle
10k
3
+
6
idle-max
NE555
MOSFET
4
5
MOSFET
s
100k
4
DRIVER
.1uf
2
1
switch
12VDC
CD4069
14
1k
2
3
4
5
TRIGGER
1
7
6
oscillator
ground
frequency
adjust
dipswitch
2k
1k
.01u
.1u
.03u
3300pf
+ 12VDC
pressure
gauge
L
L
cht
gauge
spark plug
tank
level
sensor
L
tank level
gauge
chamber
level
sensor
L
chamber level
gauge
100k
10k
7
"pump active"
2
-
LM741
1k
100k
3
+
6
4
pump
detector
10k
.01uf
E3055T
ground

FIGURE 4 E-SCHEMATIC

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

REV. A

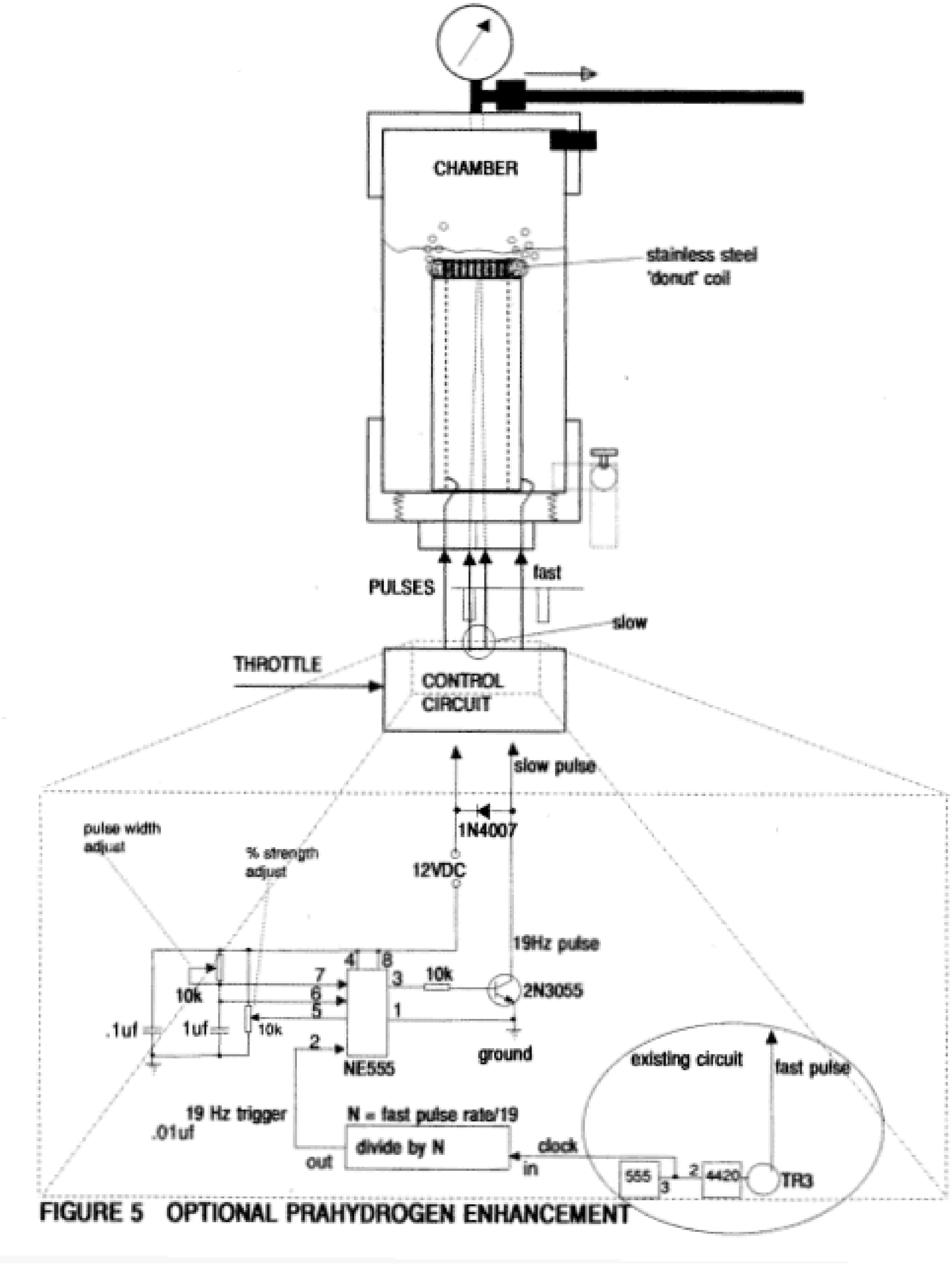
CHAMBER
stainless steel 'donut' coil
PULSES
fast
slow
THROTTLE
CONTROL CIRCUIT
slow pulse
1N4007
12VDC
pulse width adjust
% strength adjust
19Hz pulse
4
8
7
6
5
3
1
2
10k
2N3055
10k
.1uf
1uf
10k
NE555
ground
19 Hz trigger
.01uf
N = fast pulse rate/19
divide by N
out
clock
in
existing circuit
fast pulse
555
3
2
4420
TR3

FIGURE 5 OPTIONAL PRAHYDROGEN ENHANCEMENT

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

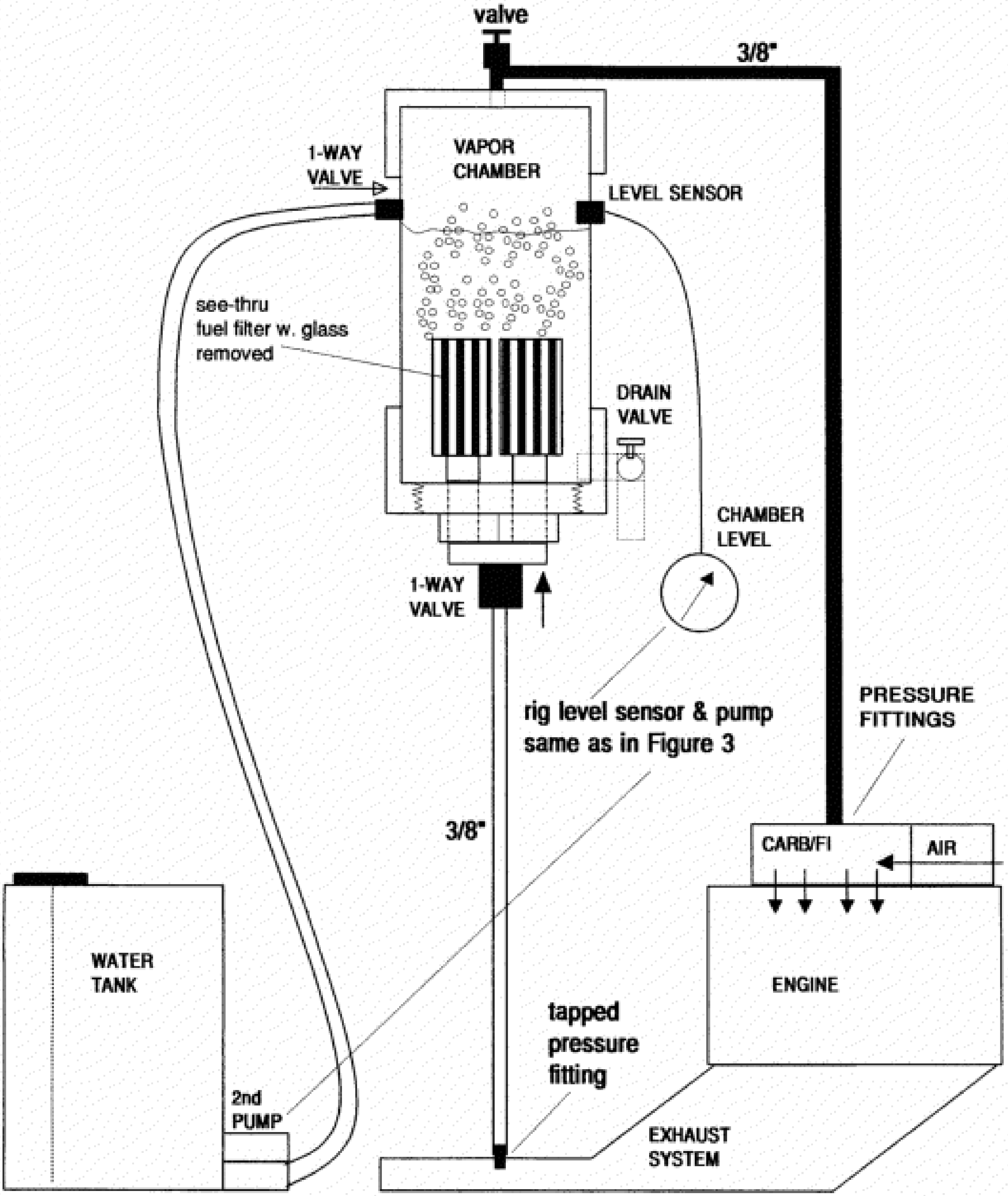
valve
3/8"
VAPOR CHAMBER
1-WAY VALVE
LEVEL SENSOR
see-thru fuel filter w. glass removed
DRAIN VALVE
CHAMBER LEVEL
1-WAY VALVE
rig level sensor & pump same as in Figure 3
PRESSURE FITTINGS
3/8"
CARB/FI
AIR
WATER TANK
ENGINE
tapped pressure fitting
2nd PUMP
EXHAUST SYSTEM

FIGURE 6 OPTIONAL WATER COOLING

An idle telephone line tempts designers to use its 48V potential as a power source. However, Part 68 of the US Federal Communications Commission's telecommunications regulations states that any device that connects to the phone line and is not actively communicating must present a resistance of at least 5 MΩ (**Reference 1**). To meet this requirement, a device's continuous-current drain must not exceed 10 μA. Fortunately, many devices that connect to the phone line do not require continuous power and can remain off for long intervals, awakening only for a short time before relapsing into power-off mode. Providing power for these applications from the phone line presents obvious advantages by eliminating the need for a battery or another power source and the cost of battery maintenance.

The circuit in **Figure 1** charges a 1.5F supercapacitor, C_1, from the phone line through a diode bridge and a 5.6-MΩ resistor. A **Maxim MAX917** nanopower comparator, IC_1, consumes only 0.75 μA from its power supply. Resistors R_2 and R_3 halve the voltage across C_1 and apply it to IC_1's positive input voltage at Pin 3 for comparison with its built-in 1.245V reference. For voltages across C_1 that do not exceed 2.49V, IC_1's output at Pin 6 remains low. When C_1's voltage reaches 2.5V, Pin 3's voltage exceeds the reference voltage, and IC_1's output goes high, turning on Q_1 and Q_2.

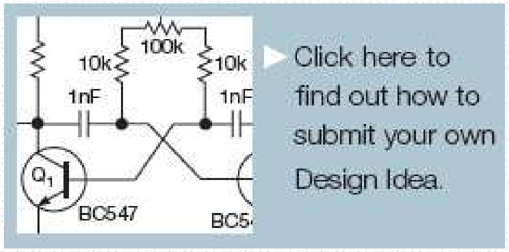
Click here to find out how to submit your own Design Idea.

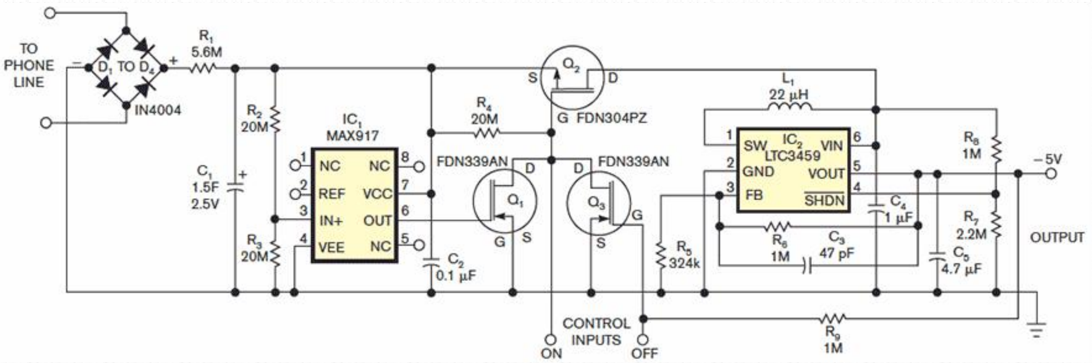
TO PHONE LINE
D1 TO D4
IN4004
R1 5.6M
R2 20M
C1 1.5F 2.5V
R3 20M
IC1 MAX917
NC
REF
IN+
VEE
NC
VCC
OUT
NC
C2 0.1 μF
R4 20M
Q2
S
D
G FDN304PZ
FDN339AN
Q1
Q3
CONTROL INPUTS
ON
OFF
R5 324k
L1 22 μH
SW
IC2 LTC3459
GND
FB
VIN
VOUT
SHDN
R6 1M
C3 47 pF
C4 1 μF
C5 4.7 μF
R8 1M
R7 2.2M
R9 1M
–5V
OUTPUT

Figure 1 This power-conversion circuit delivers intermittent bursts of regulated voltage from a supercapacitor charged by a trickle of current from a telephone line.

Several days must elapse before C_1 becomes fully charged, given its huge capacitance and a charging current of less than 10 µA. The voltage on C_1 can never exceed 2.5V because, once it reaches 2.49V, Q_1 and Q_2 turn on, connecting C_1 to a switched-mode-power-supply circuit. Because the power-supply current exceeds the charging current, the voltage across C_1 starts to decrease when Q_2 turns on. Transistor Q_3 holds Q_2 on when C_1's decreasing voltage causes Q_1 to turn off.

The switched-mode-power-supply circuit comprises a **Linear Technology LTC3459** micropower boost converter, IC_2, and its associated components, which deliver 5V at 10 mA. A fully charged C_1 can supply power to a 10-mA load for approximately 40 sec. With no load, the circuit can sustain its 5V output for more than 10 hours. For greater output current and shorter operating time, select another boost converter that can operate at a low input voltage.

READ MORE designideas

Mechanical switches, open-drain MOSFETs, open-collector transistors, or a microcontroller's open-drain output pins can drive two external control inputs to force the circuit on and off. Pulling the On input low forces Q_2 to turn on and deliver power from C_1 to the power converter, and pulling the Off input low turns off Q_2 and removes power from the converter. Note that the power converter's output-return line connects to the telephone line and thus should not connect to an earth ground or to grounded equipment.

Reference

1. "**Part 68**," Federal Communications Commission.

in Share G+ Tweet Like 0

Download Favorite

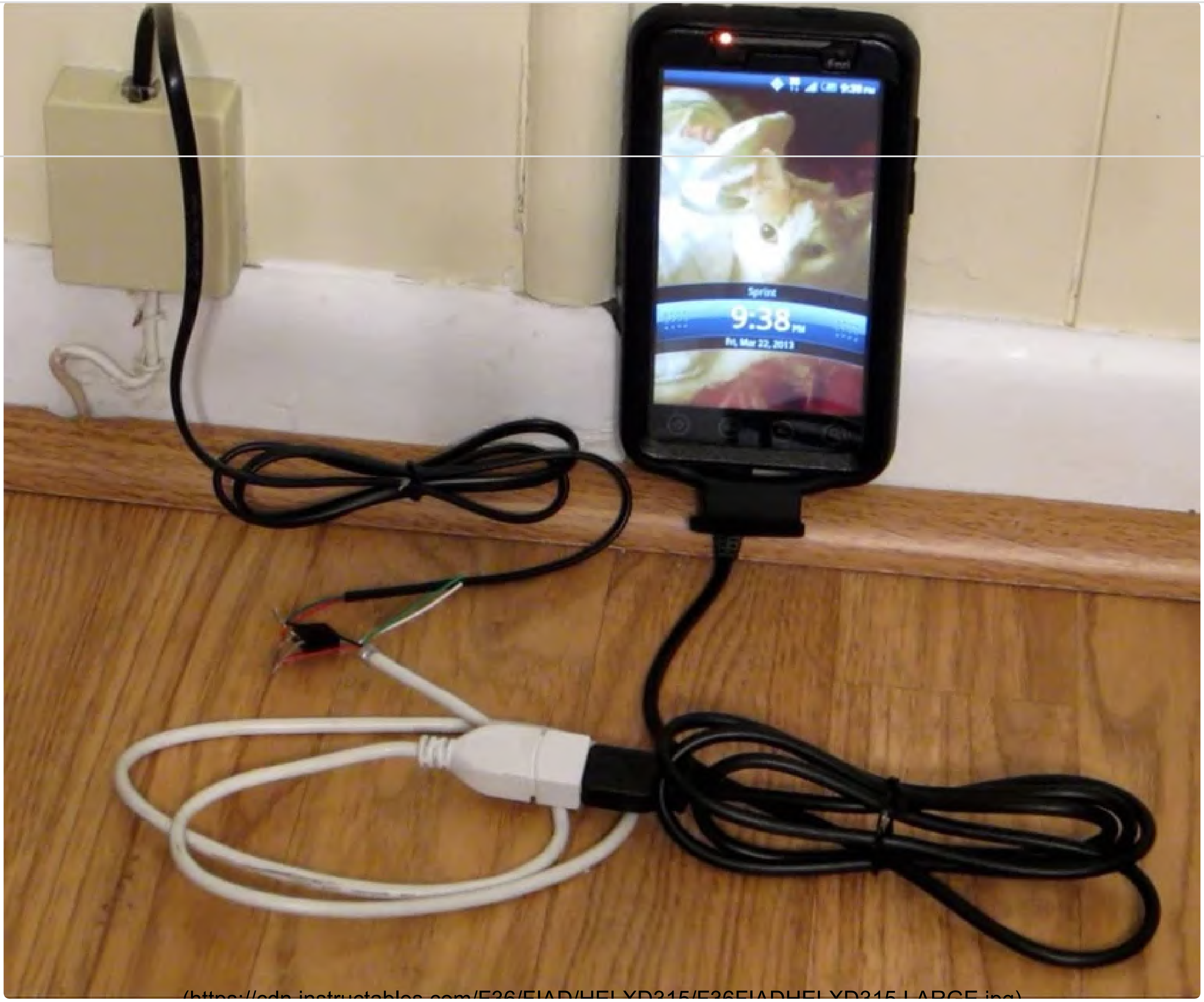

(https://cdn.instructables.com/F36/FIAD/HELXD315/F36FIADHELXD315.LARGE.jpg)

vimeo

Tons and tons of video storage.

Get up to 7TB

advertisement

(/member/DIY%20Hacks%20and%20How%20Tos/)

By **DIY Hacks and How Tos (/member/DIY%20Hacks%20and%20How%20Tos/)**

DIY Hacks & How Tos on YouTube (http://www.youtube.com/user/DIYHacksandHowTos)

Follow

More by the author:

HELLO
I AM BAYMAX
YOUR PERSONAL
HEALTHCARE
COMPANION

6 THINGS TO COO
A COFFEE MAKE

Festivu
Pole

not in use, this is a constant DC signal (about 50-60 volts). When the phone rings, the signal is a 20 hertz AC signal (about 90 volts). When in use it is a modulated DC signal (between 6 and 12 volts).

The phones lines even have power during a blackout in most cases. This is because the phone company maintains their own backup power system. Your phone lines may be powered even if you don't have a land line service set up.

Add Tip Ask Question Comment Download

Step 2: Check the Phone Line With a Multimeter

Download Favorite

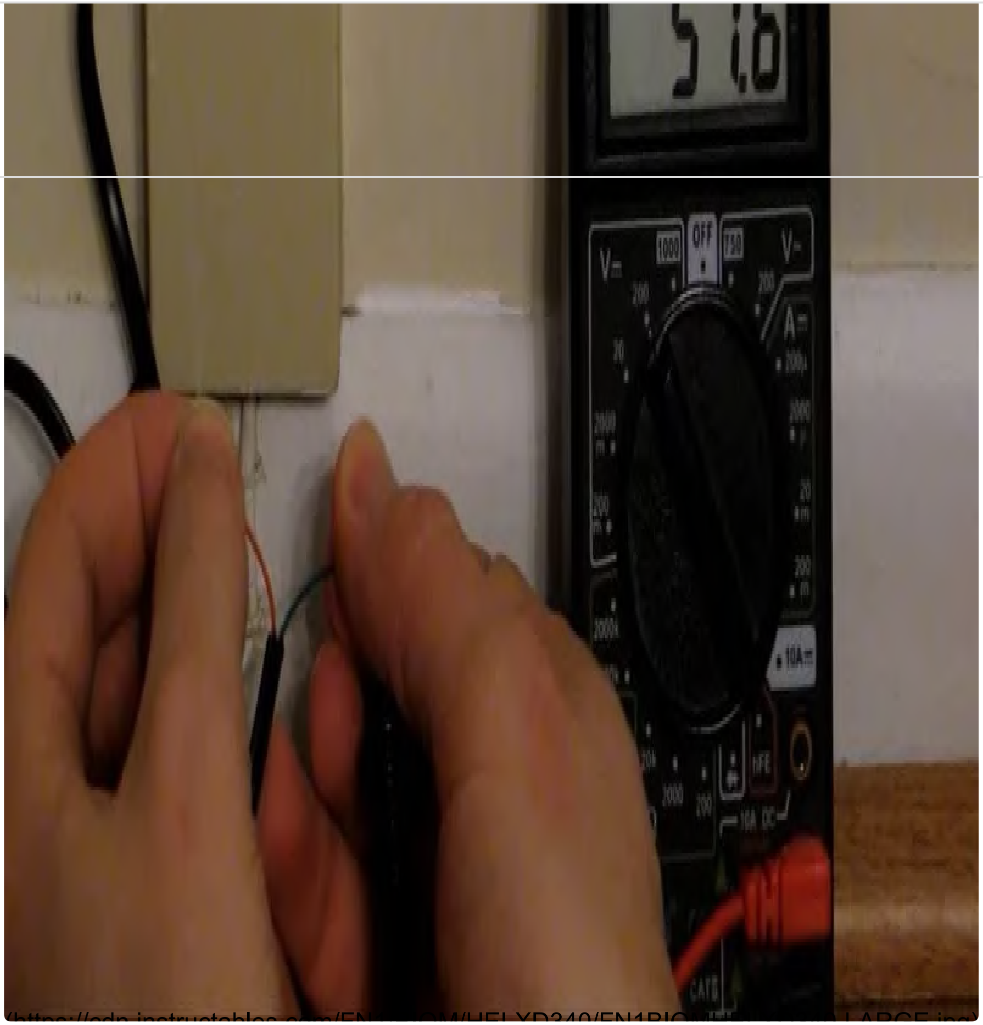

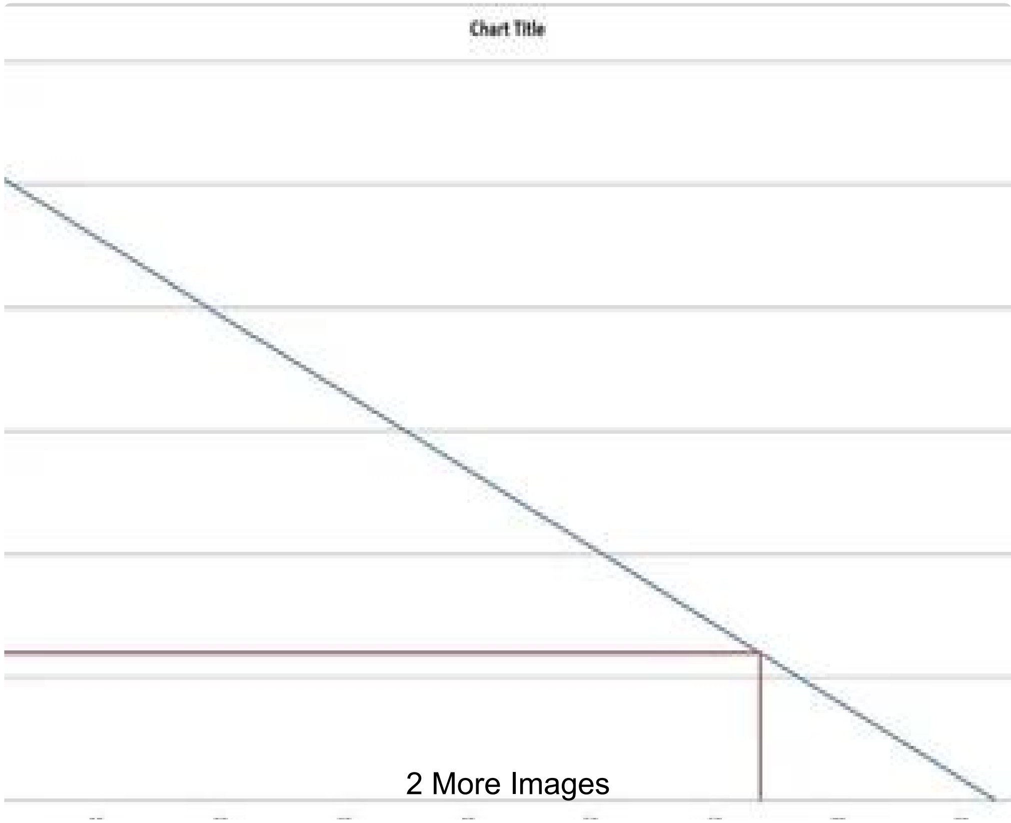

Before you try to tap into the electricity in the phone line, you should check it with a multimeter to see what you are working with.

Start by cutting open a phone cord and separating the internal wires. In most cases you will have one red wire and one green wire. Strip the insulation off the ends. Then plug the cord into a phone jack and use a multimeter to measure the output voltage. At my house, I measured an open-circuit (no load) voltage (http://en.wikipedia.org/wiki/Open-circuit_voltage) of 52 volts DC.

Then I hooked up various resistors to see what the output would be with different loads. I determined that the supply voltage isn't regulated. This means that the voltage changes depending on the resistance of the circuit that it is powering. After some calculating, I worked out that the base signal coming out of my phone jack pretty closely resembles a 52 Volt DC source (http://en.wikipedia.org/wiki/Voltage_source) with a 628 ohm internal resistance (http://en.wikipedia.org/wiki/Internal_resistance).

Basically this means that I can run a 12V circuit at 64mA, a 9V circuit at 68mA, or a 5V circuit at 75mA. This isn't a lot. But it is enough to charge a cell phone.

Add Tip Ask Question Comment Download

Step 3: Construct a Simple Voltage Regulator Circuit

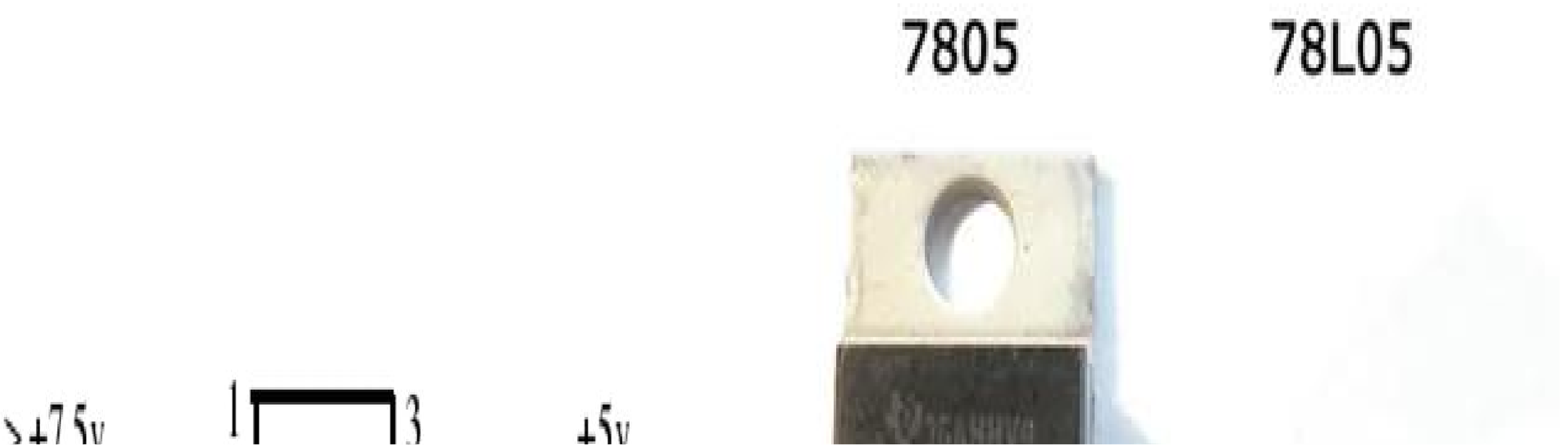

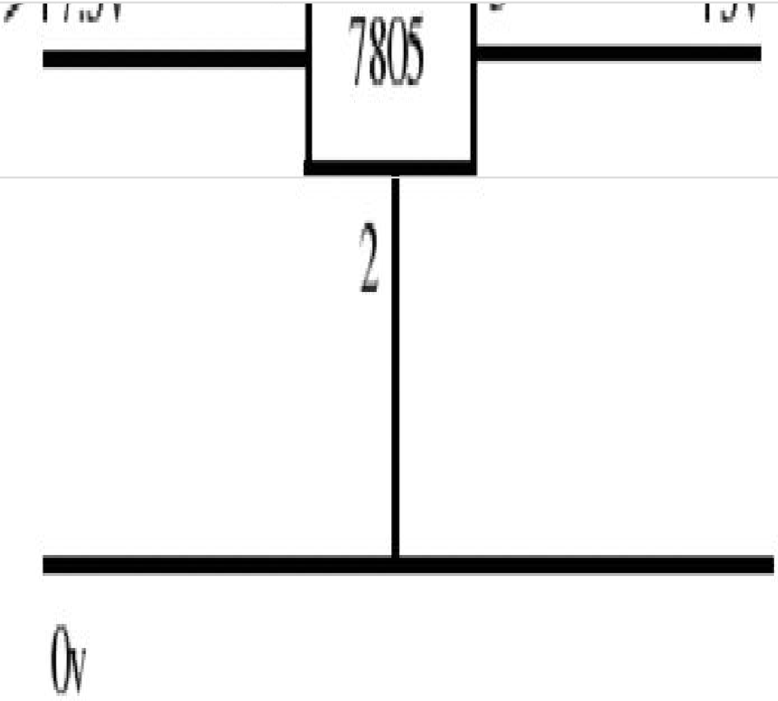

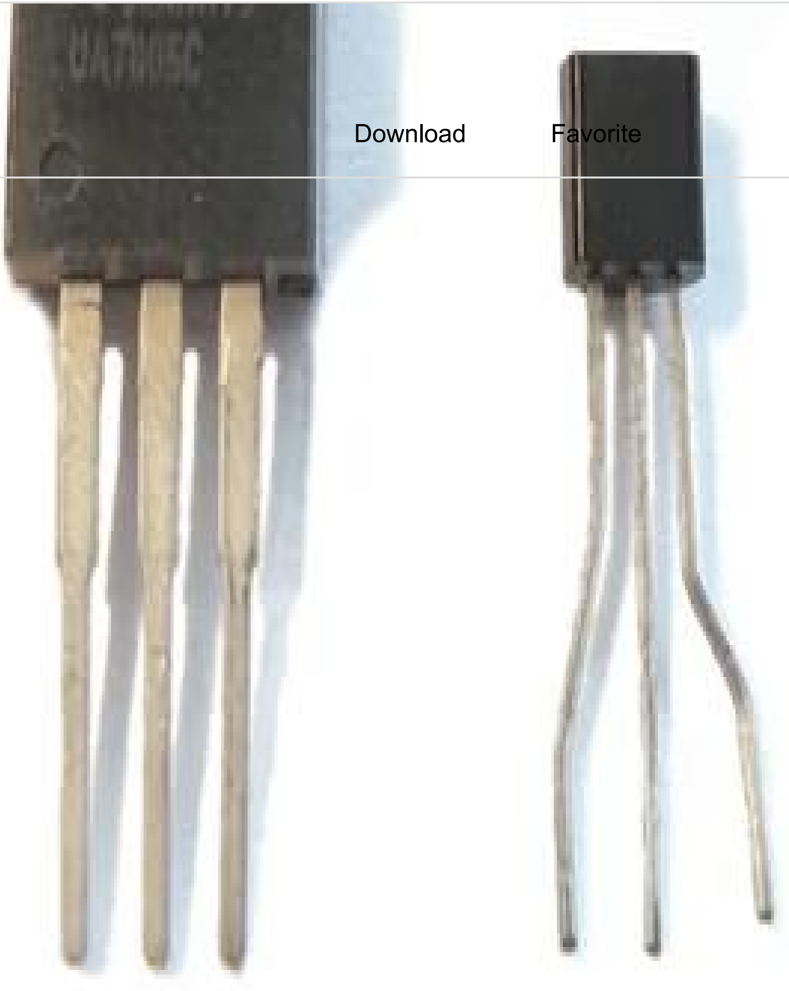

(https://cdn.instructables.com/FXB/5YW1/HFI4EJY4/FXB5YW1HFI4EJY4.LARGE.jpg)

(https://cdn.instructables.com/FEV/H9M3/HFLXD34J/FEVH9M3HFLXD34J.LARGE.gif)

We know that the phone needs 5 volts in order to charge. But much current it draws or it's equivalent [load resistance (http://en.wikipedia.org/wiki/Electrical_load)](http://en.wikipedia.org/wiki/Electrical_load). So we can't charg from the phone line. We need to use a [voltage regulator (http://en.wikipedia.org/wiki/Voltage_regulator)](http://en.wikipedia.org/wiki/Voltage_regulator) to bring the out

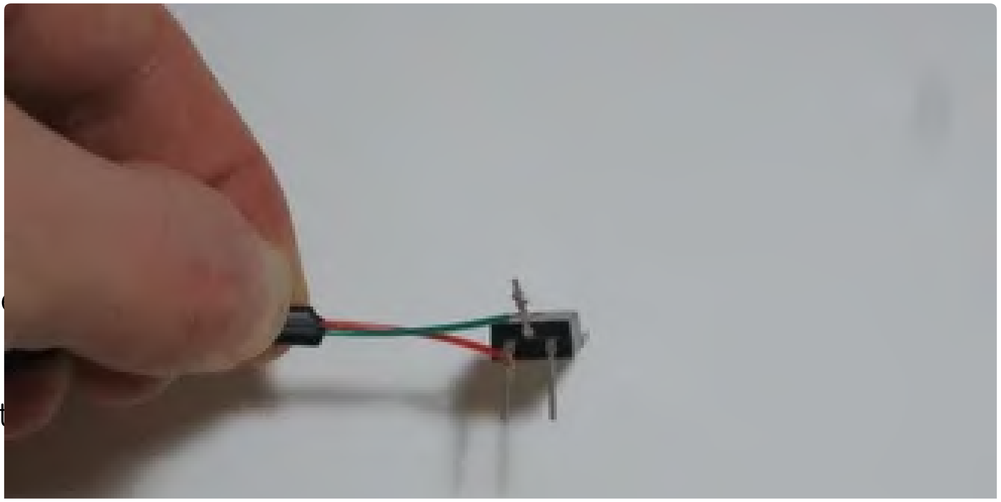

down to 5 volts and keep it there. A LM7805 5 volt regulator s

3 More Images
Download
Favorite
(https://cdn.instructables.com/FAP/XDW2/HFLXD34Y/FAPXDW2HFLXD34Y.LARGE.jpg)

To make this simple phone line adapter you will need the phon been working with, the 5V voltage regulator and a USB connector cable with a female end. Just connect the red wire from the phone line to the first lead on the regulator and connect the green wire from the phone line to the second lead. Then connect the black wire from the USB cable to the second lead on the regulator and connect the red wire from the USB cable to the third lead on the regulator. If you can't solder the wires together (because the power is out), you can just wrap the wires around each lead. If you do this, you should bend the leads of the regulator away from each other. This will help you avoid accidentally crossing the wires.

This simple regulator circuit is able to safely convert the base phone signal into something that can be used to charge your phone. However, many voltage regulators are not able to handle the AC signal that they would receive if the phone rang. So if you are worried that you might receive a call while your regulator is hooked up to the phone line , then you may wish to add a diode between the red wire from the phone line and the first pin on the voltage regulator. This will protect your circuit from problems that may be caused by reverse polarity.

Add Tip Ask Question Comment Download

Step 4: Use Other Regulators for Other Output Voltages

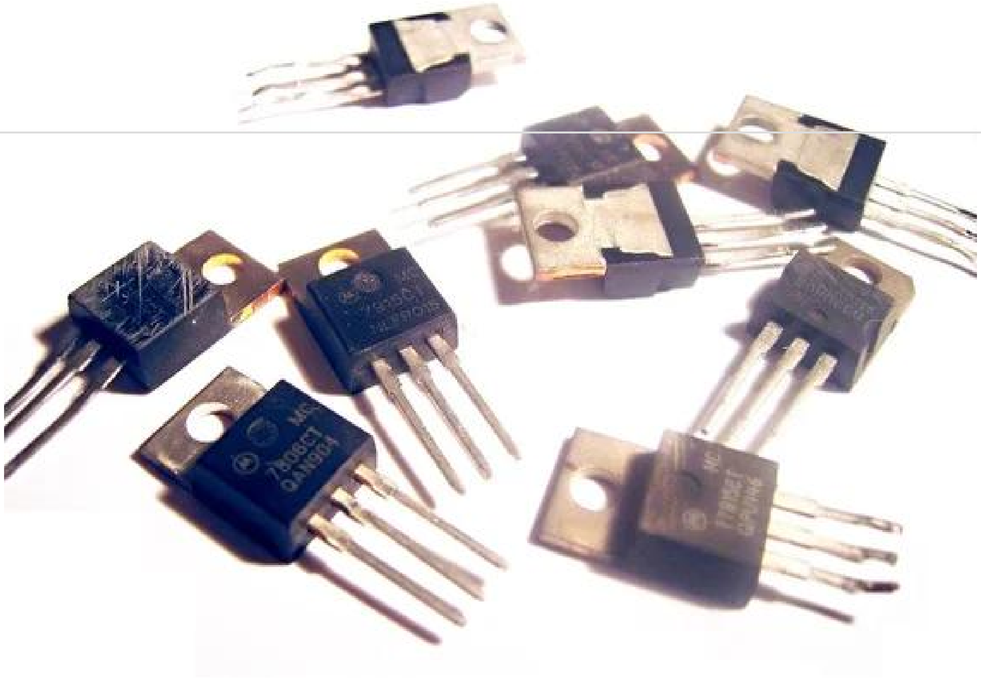

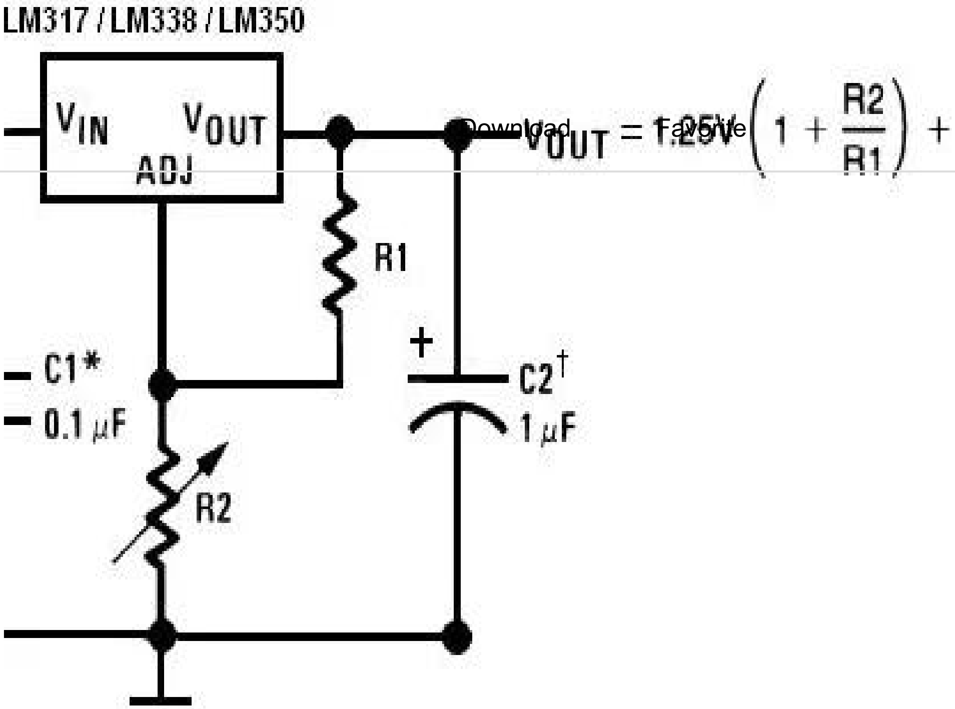
LM317 / LM338 / LM350
V_{IN}
V_{OUT}
ADJ
$V_{OUT} = 1.25V\left(1 + \frac{R2}{R1}\right) +$
R1
+
C1*
0.1 µF
C2†
1 µF
R2

A 7805 regulator will work if you need an output of 5 volts but other kinds of voltage of voltage regulators are also available. Other voltages in the 78xx series include 6V, 8V, 9V, 10V, 12V, 15V, 18V, and 24V. In addition to these fixed value regulators, there are also variable regulators that let you set the voltage level with the use of a few external components. One such variable voltage regulator is the LM317 (http://www.ti.com/lit/ds/symlink/lm117.pdf). These are what you would use if you needed a different output voltage.

Add Tip Ask Question Comment Download

Step 5: Finished Phone Line Adapter Tool

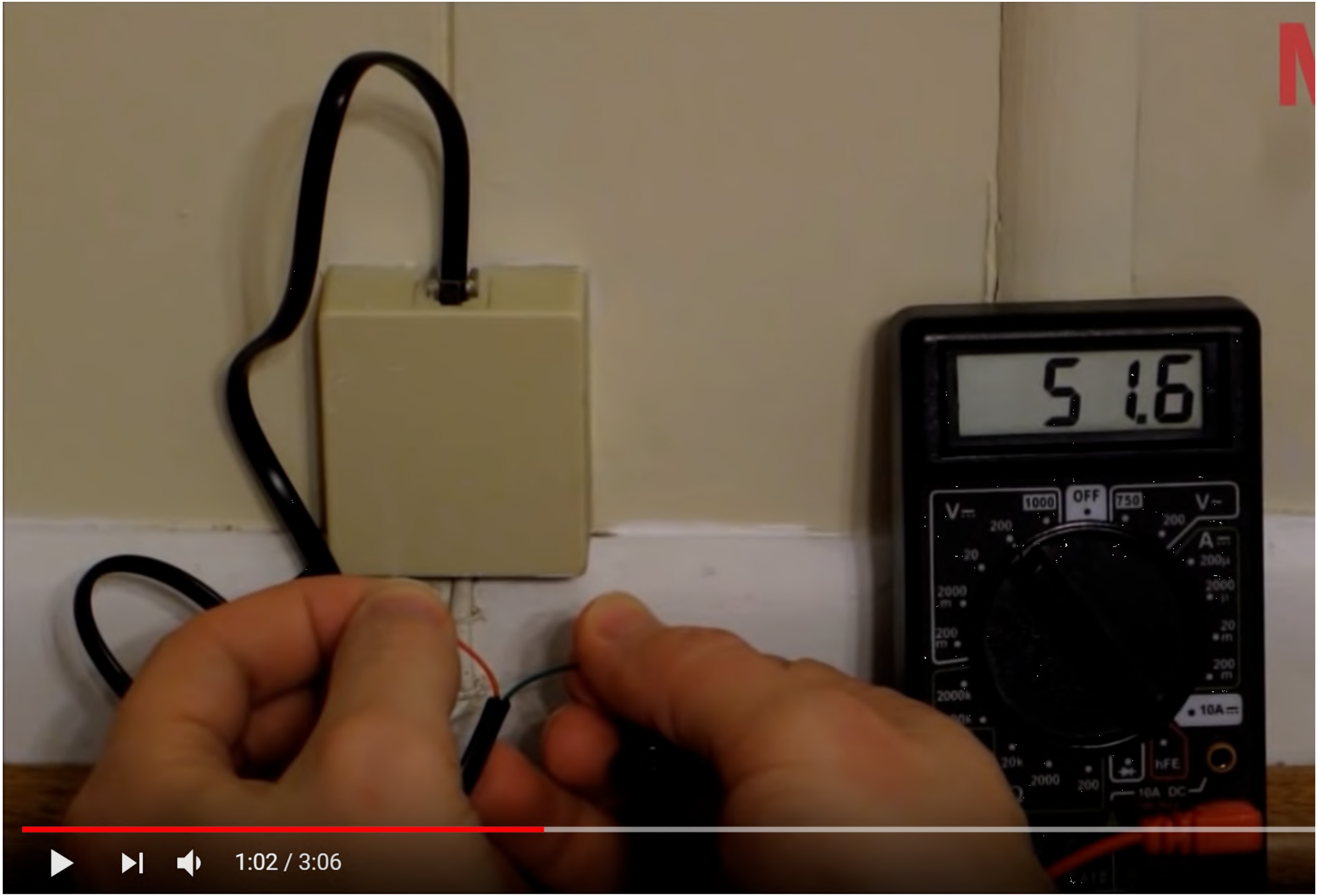
51.6
OFF
1000
750
V⎓
V~
200
20
2000 m
200 m
A⎓
200µ
2000 µ
20 m
200 m
2000k
10A⎓
20k
2000
200
hFE
10A DC
1:02 / 3:06

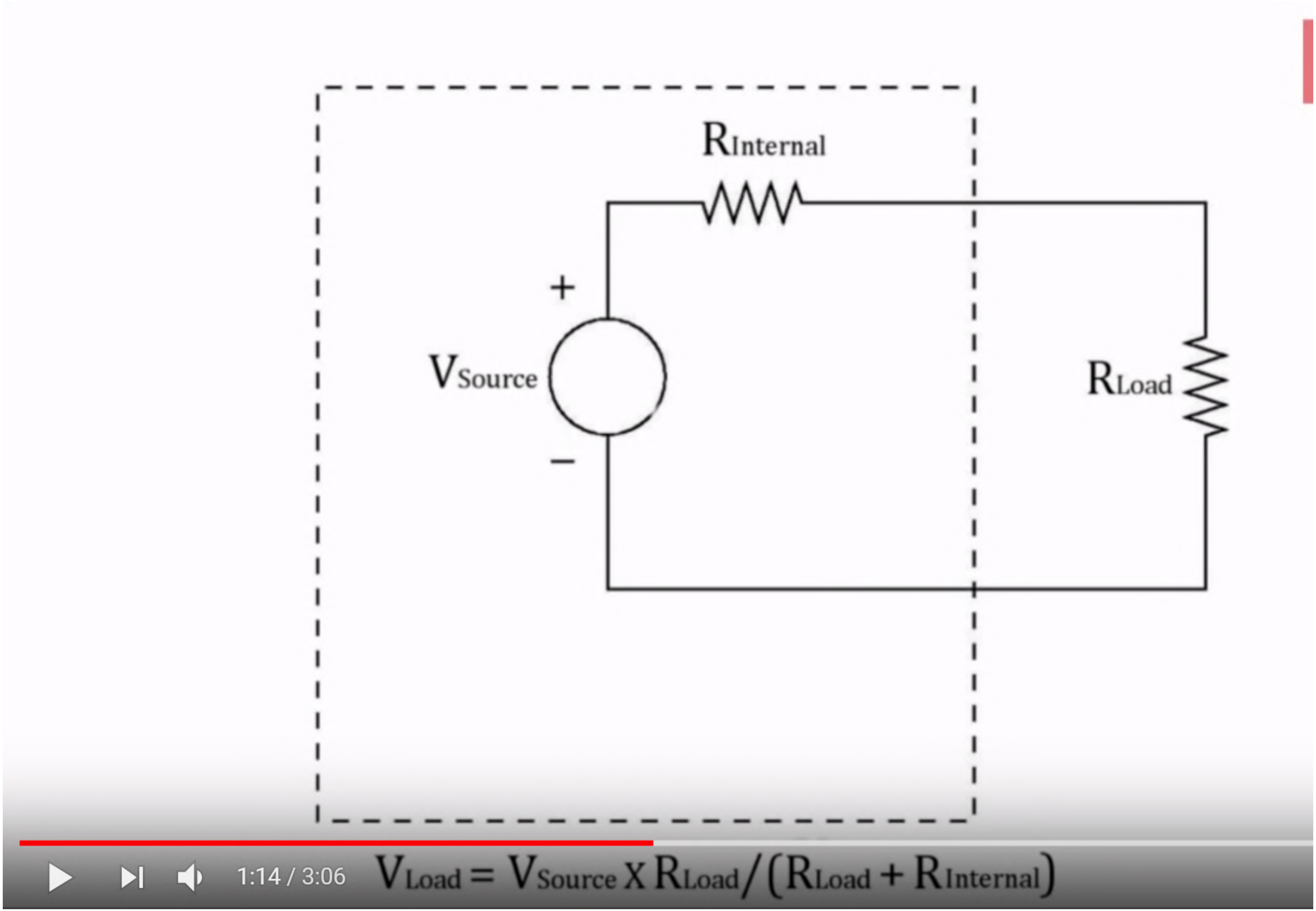
$R_{Internal}$
$+$
V_{Source}
$-$
R_{Load}
1:14 / 3:06

$$V_{Load} = V_{Source}\,X\,R_{Load}/(R_{Load} + R_{Internal})$$

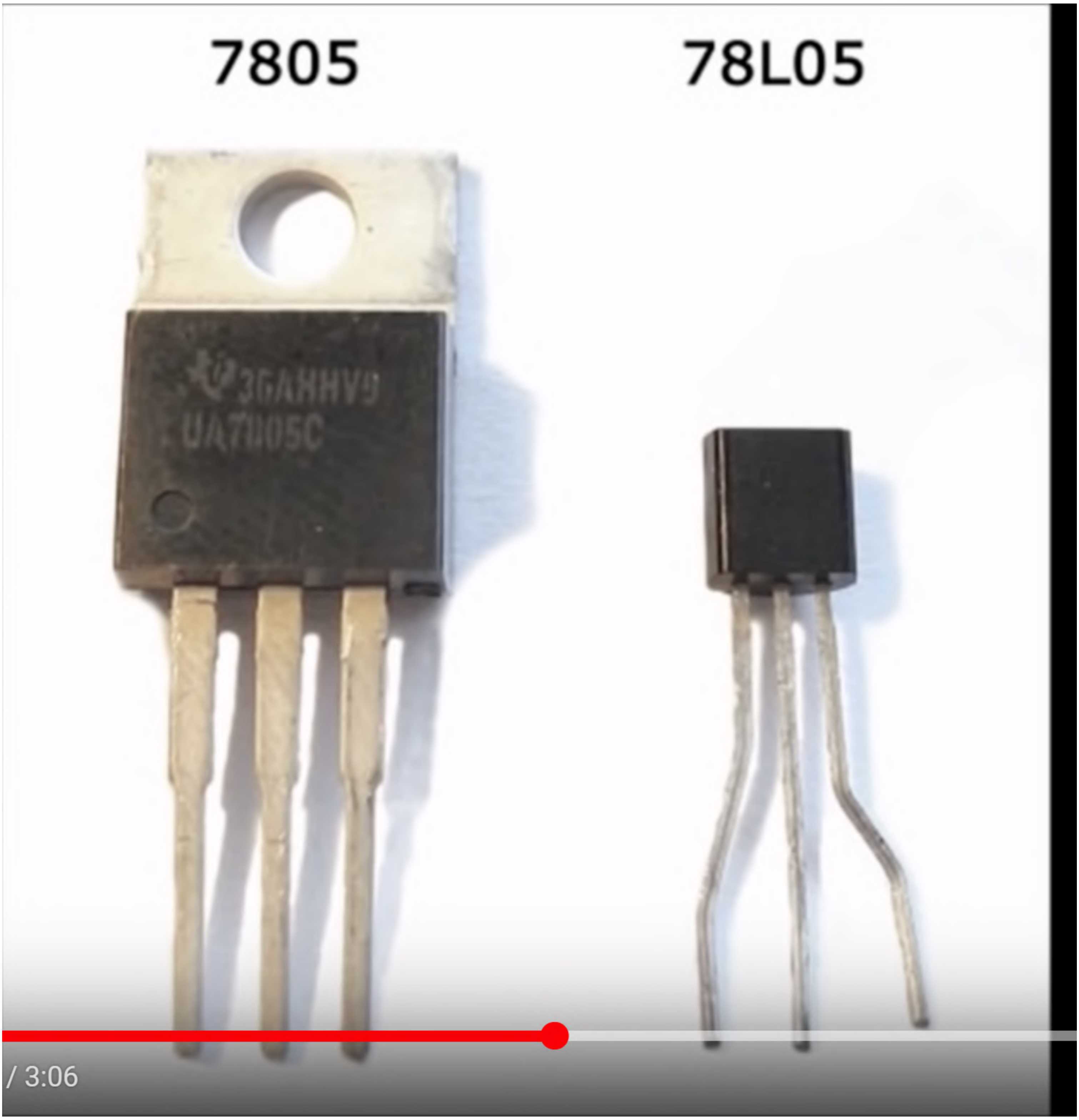
7805
78L05
3GAHHV9
UA7805C
/ 3:06

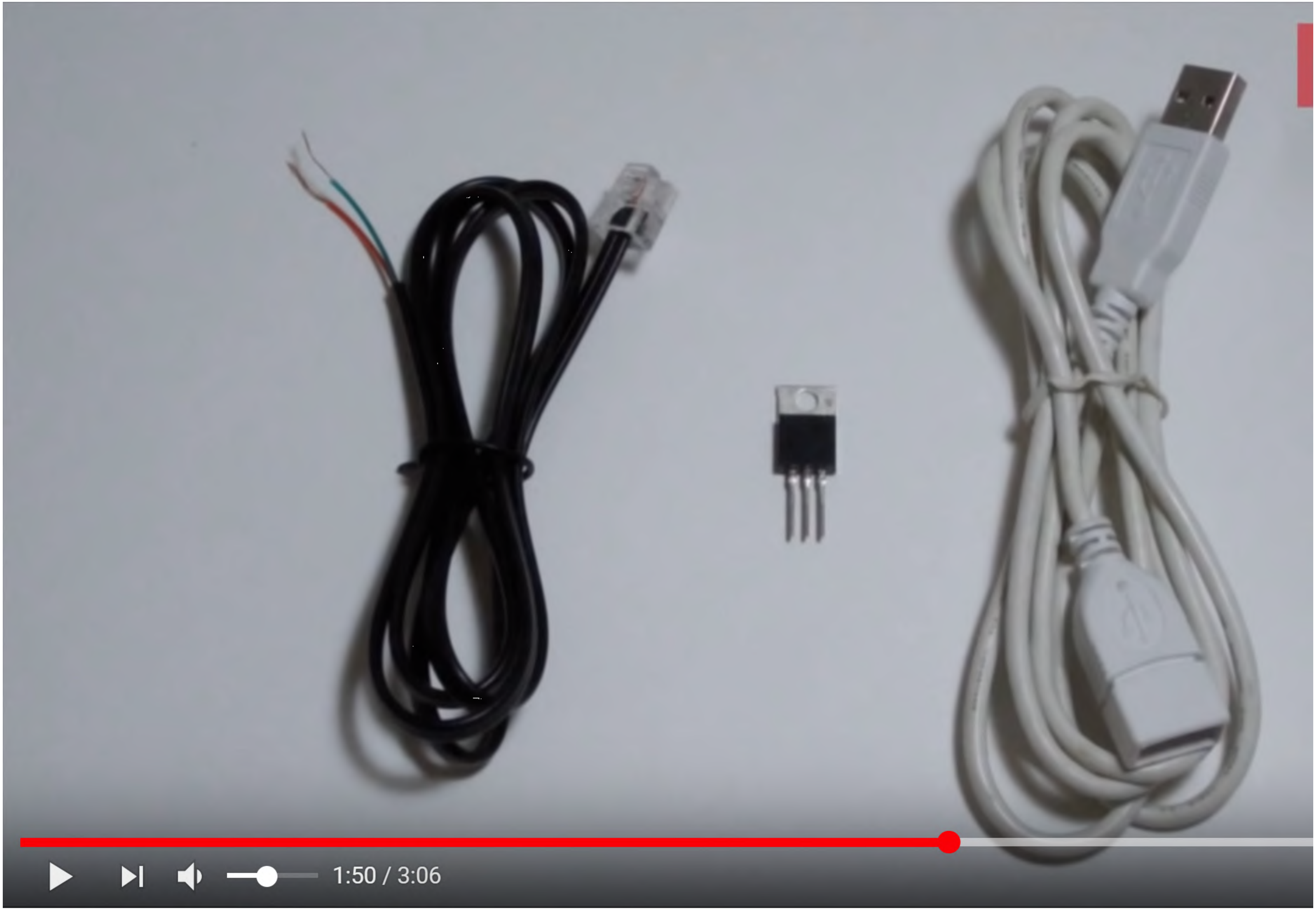
1:50 / 3:06

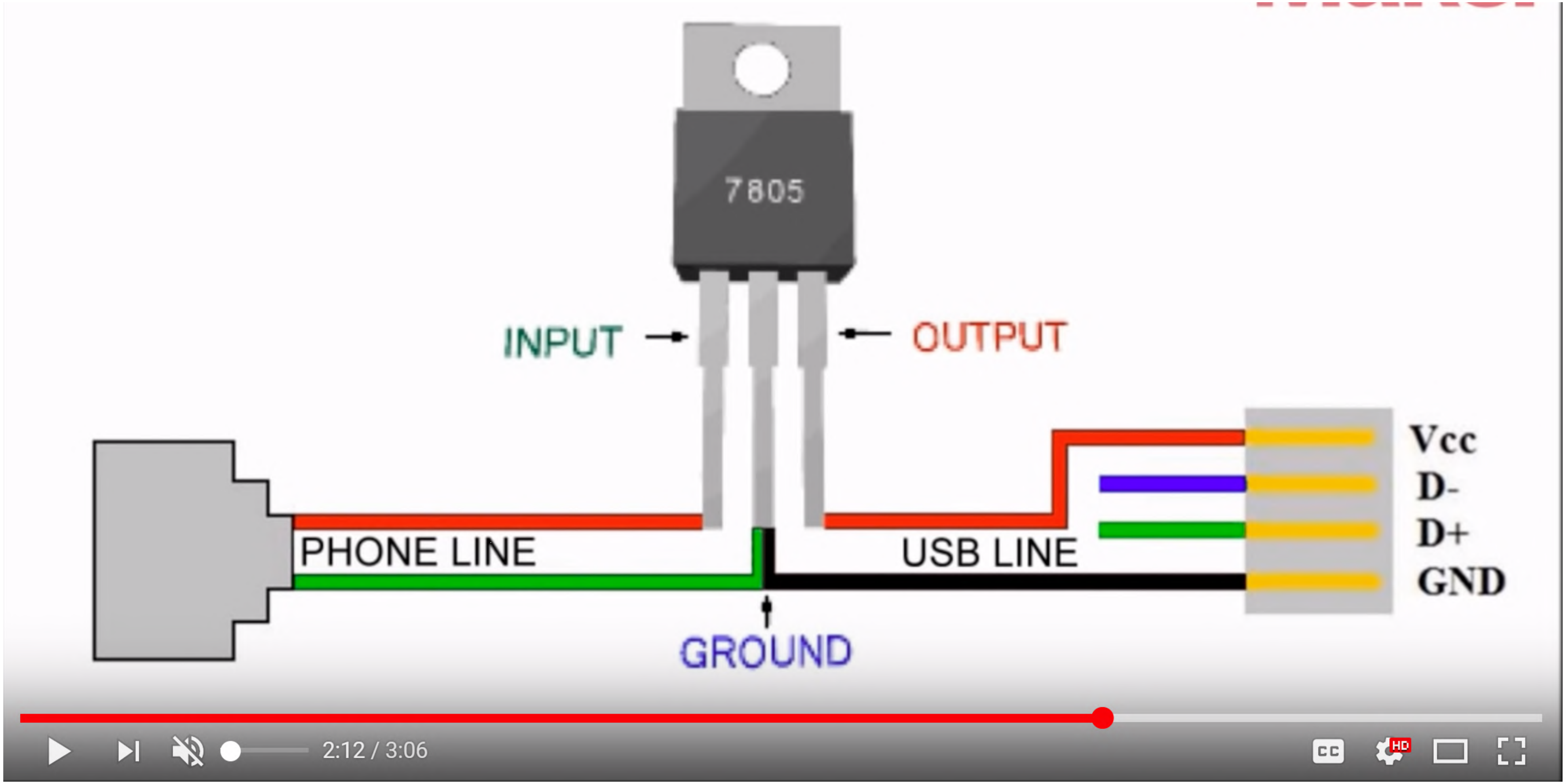
7805
INPUT
OUTPUT
Vcc
D-
D+
GND
PHONE LINE
USB LINE
GROUND
2:12 / 3:06
HD

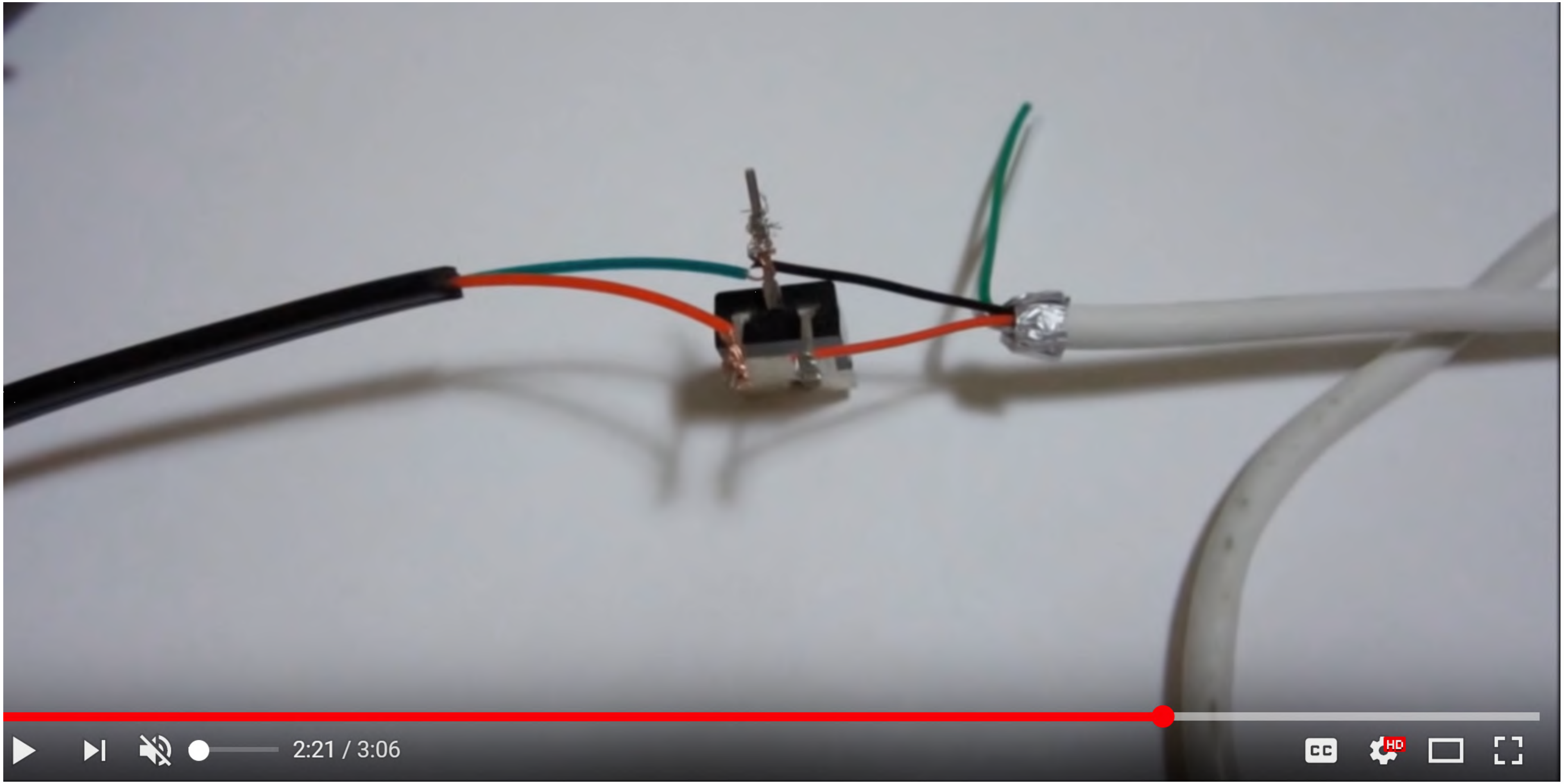
2:21 / 3:06
CC
HD

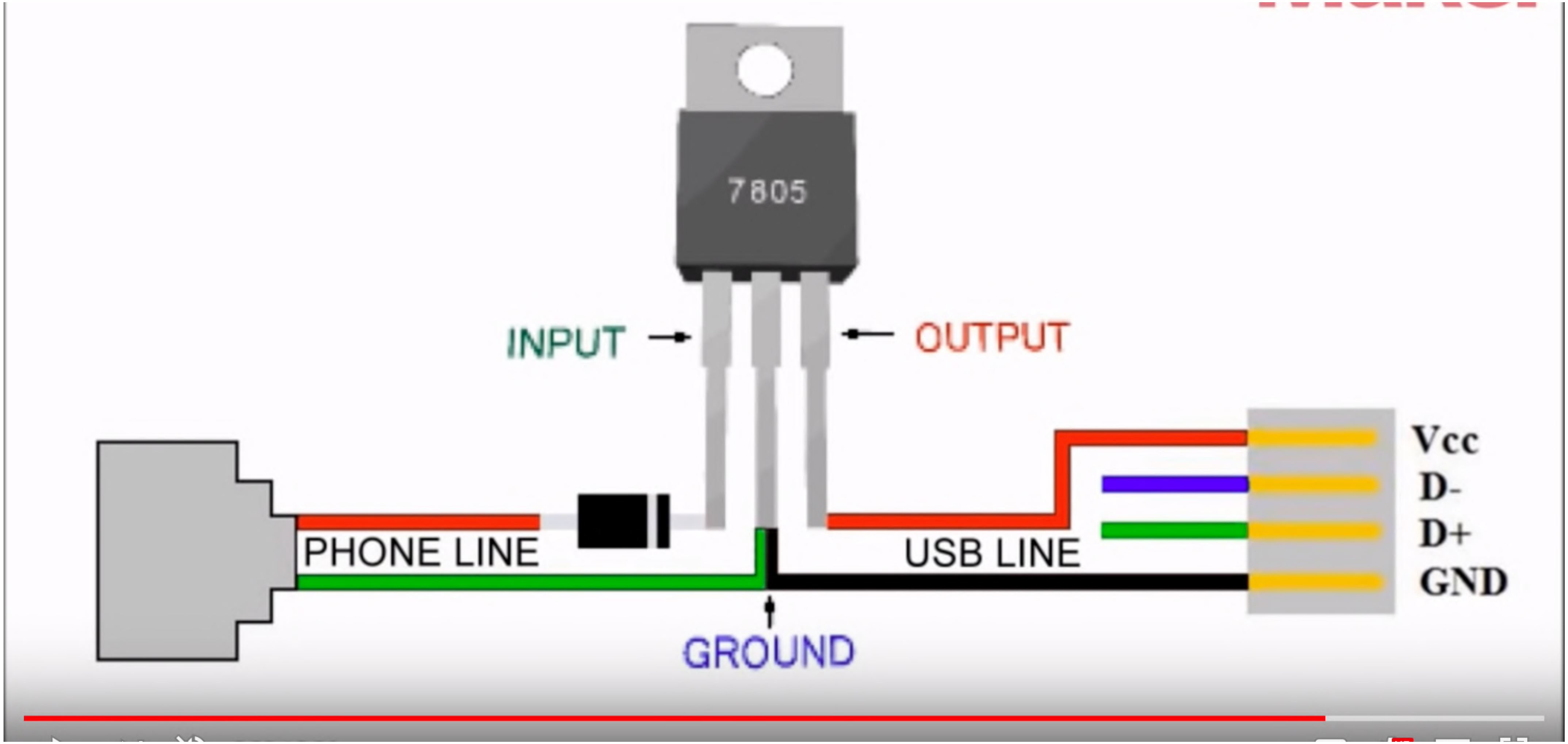
7805
INPUT
OUTPUT
PHONE LINE
USB LINE
GROUND
Vcc
D-
D+
GND

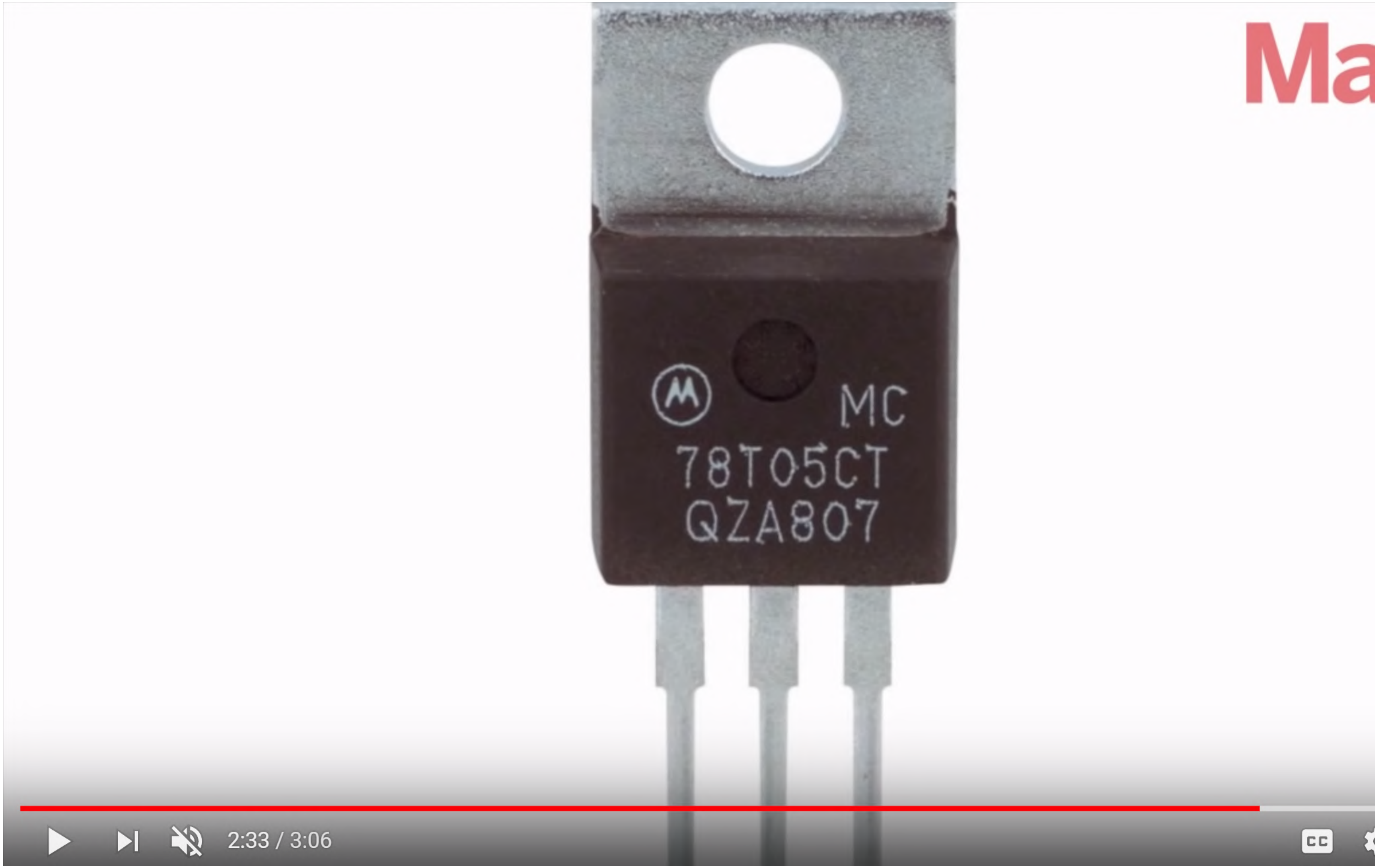
Ma
MC
78T05CT
QZA807
2:33 / 3:06

BUILDING A DIY VIDEO TRANSMITTER

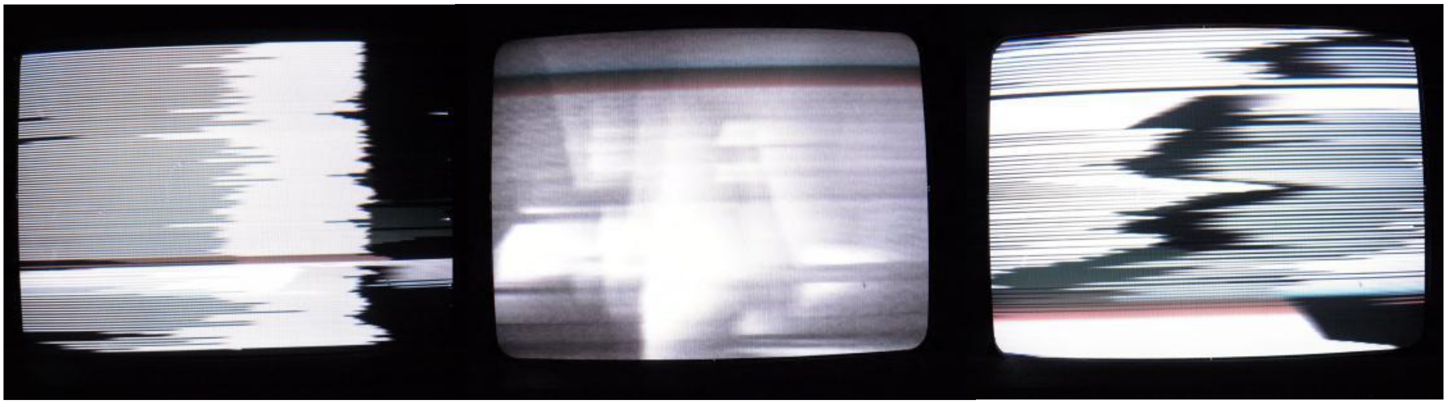

WHAT YOU NEED

- Blank PCB Copper Plate (jameco.com part #169279)
- Capacitors: 2x 0.1uf
    1x 100u
    2x 6-70pf Variable Caps (jameco.com part #32855)
- Resistors: 27kOhm
    10kOh
- Transistor-MPSA18 (jameco.com part #210681)
- 1k potentiometer
- Small amount (~4”) of magnet wire 24 AWG or close (jameco.com part #2098419)
- Sand paper
- Stranded hookup wire
- Hacksaw or dremel tool
-Headphones with 1/8” plug (from iPod or MP3 player)
- Hot glue or super glue
- Soldering Iron
- CRT “tube” TV with an antenna (the transmitter won’t transmit to newer flat screen TV’s)

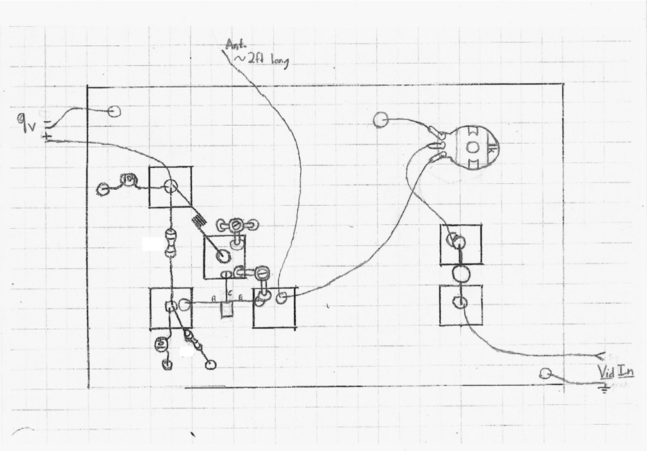

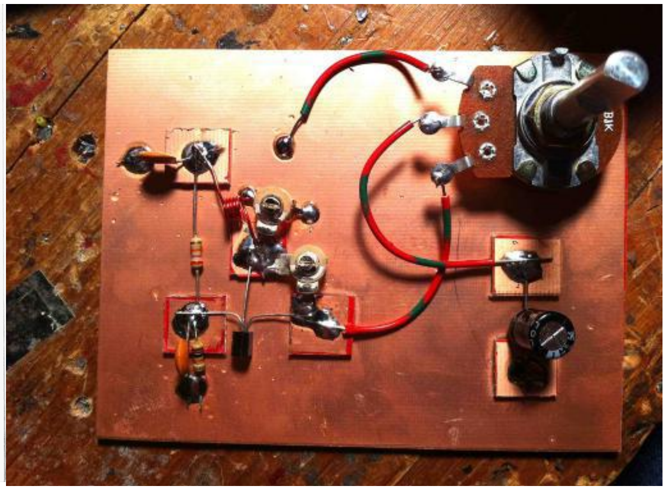

INTRODUCTION

This micro video transmitter was designed by the DIY radio and micro TV artist Tetsuo Kogawa. His website (http://anarchy.translocal.jp/) is full of great info on DIY radio, and all credit for this circuit goes to him.The transmitter can be used to transmit a relatively clear video signal or abstract visualizations, depending on the desired output. It will usually function at a distance of about 20 feet. If you have any previous experience with electronics or circuit building, you will notice that this method of building is different. The circuit is constructed on top of a copper “ground plate” which is common in radio applications. This tutorial uses step-by-step sketches and photos created based on Tetsuo Kogawa’s design.

BUILDING THE TRANSMITTER

1. PREPARING THE COPPER PLATE

- Take the bare copper plate and cut off 1inch of material using a dremel tool or hacksaw.
- Use this material to cut out 6 square ½ in. x ½ in. pieces.
- Arrange the 6 pieces on the larger copper plate as shown in the image below…

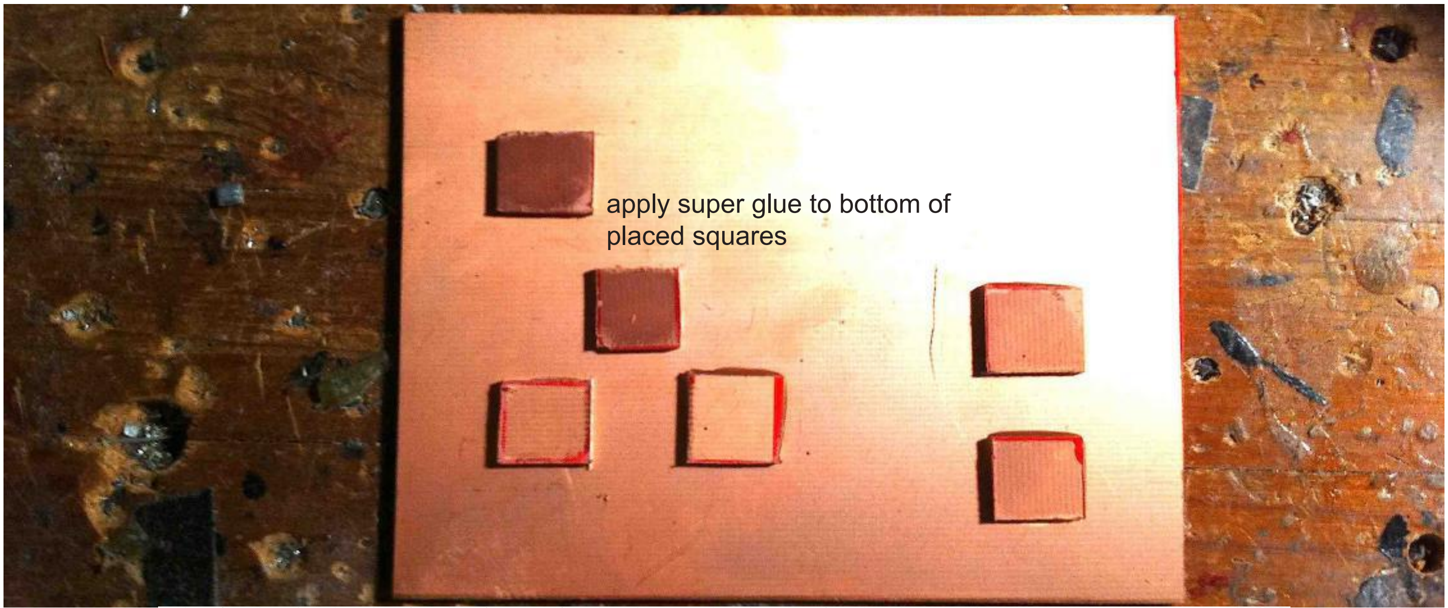

apply super glue to bottom of placed squares

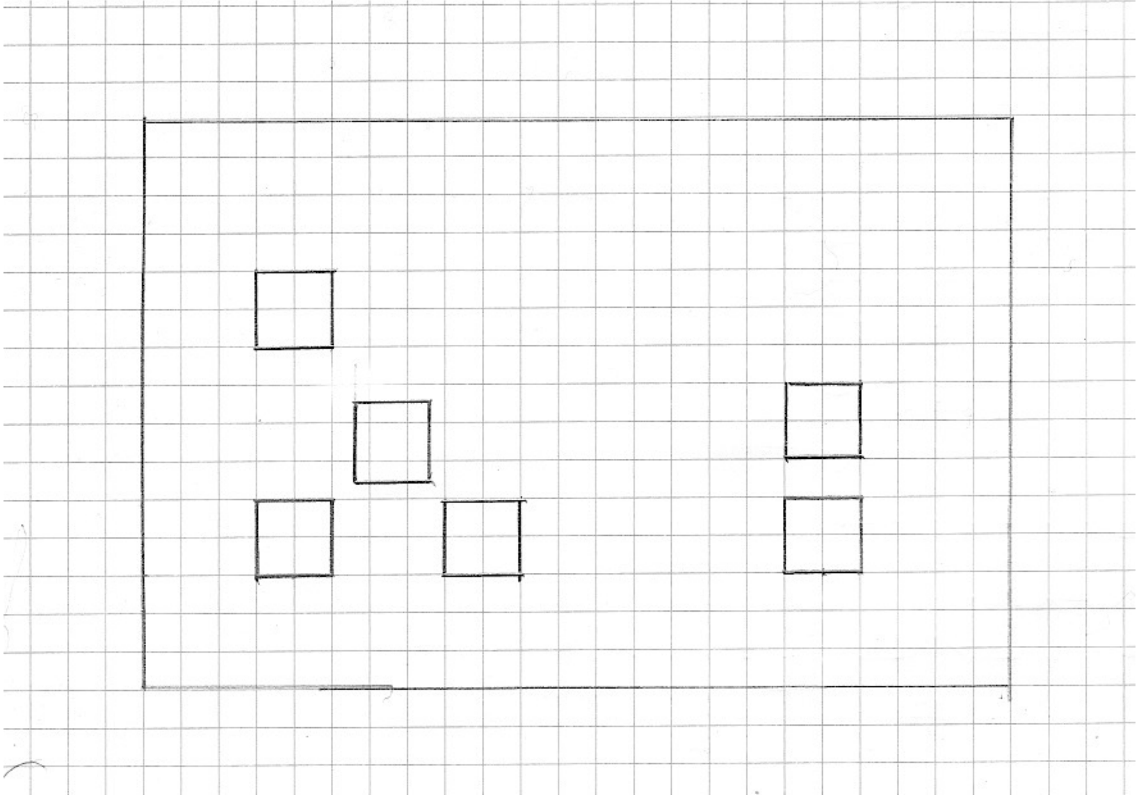

NOTE Be sure to arrange them so that all the component leads will be able to reach the squares. Look ahead to see what components will go where so you know exactly how to space them.

- Once you have them arranged, put a dab of hot or super glue on the bottom of each square and stick them in place on the copper plate.

2. ADDING THE 100uF CAPACITOR (the cylinder with “100uF” written on the side)

- Melt a blob of solder onto the two squares furthest to the right. To do this, apply your iron to each square to transfer heat for an extended period of time until the solder flows onto the surface.

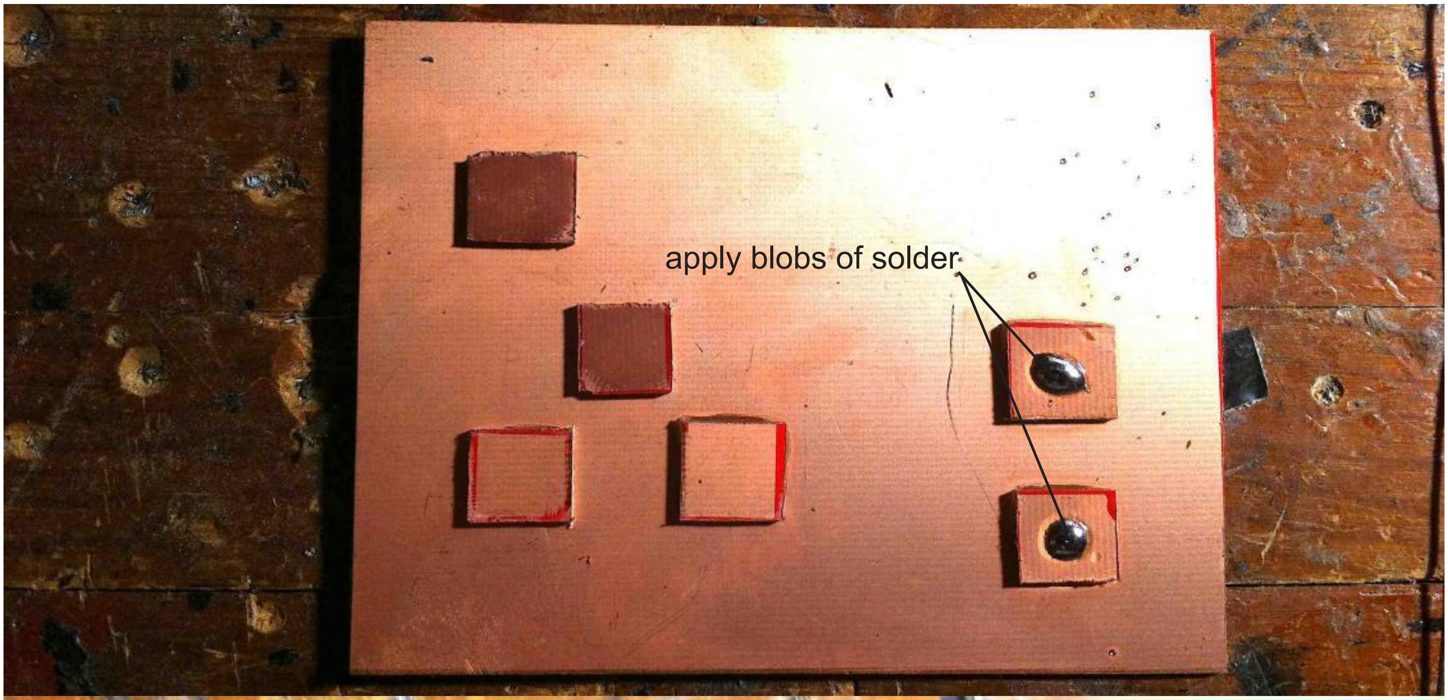
apply blobs of solder

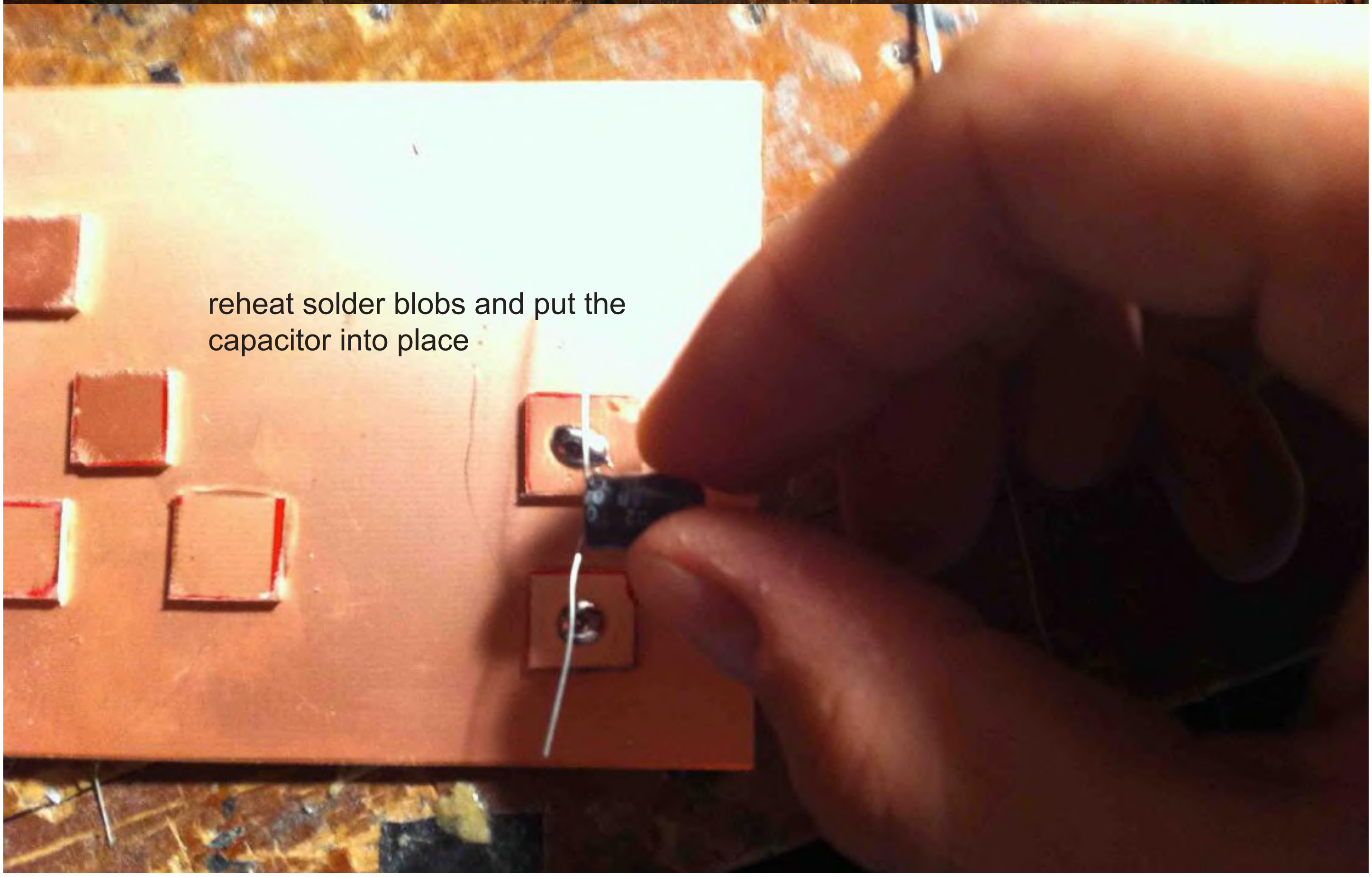
reheat solder blobs and put the capacitor into place

- Bend the legs out of the 100uf capacitor and straddle them between the two squares making sure that the leads are only touching the two small squares (not touching the ground plate).
- Cut the excess off of the leads.
- One at a time, re-heat the solder blobs and place the capacitor leads into them, allowing them to cool down until they stay in place.

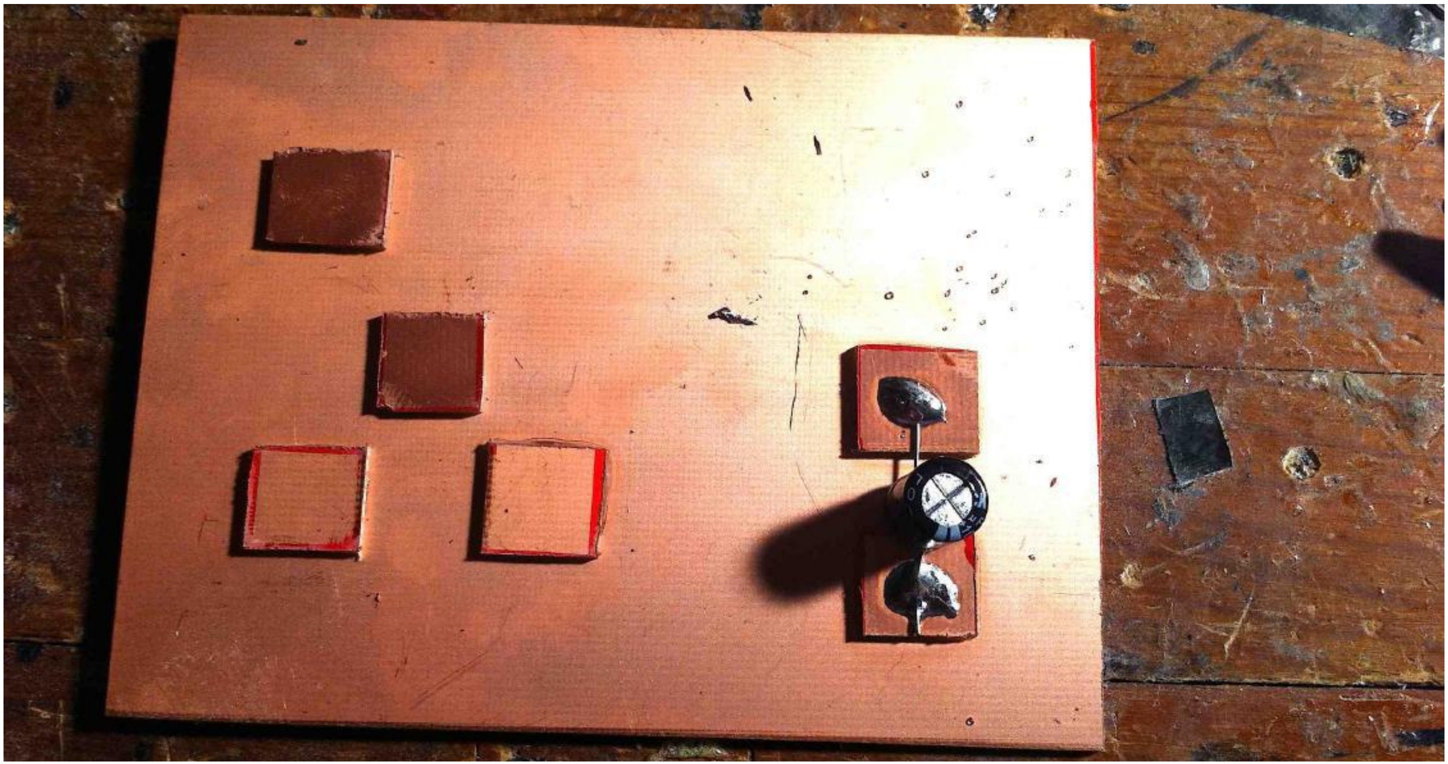

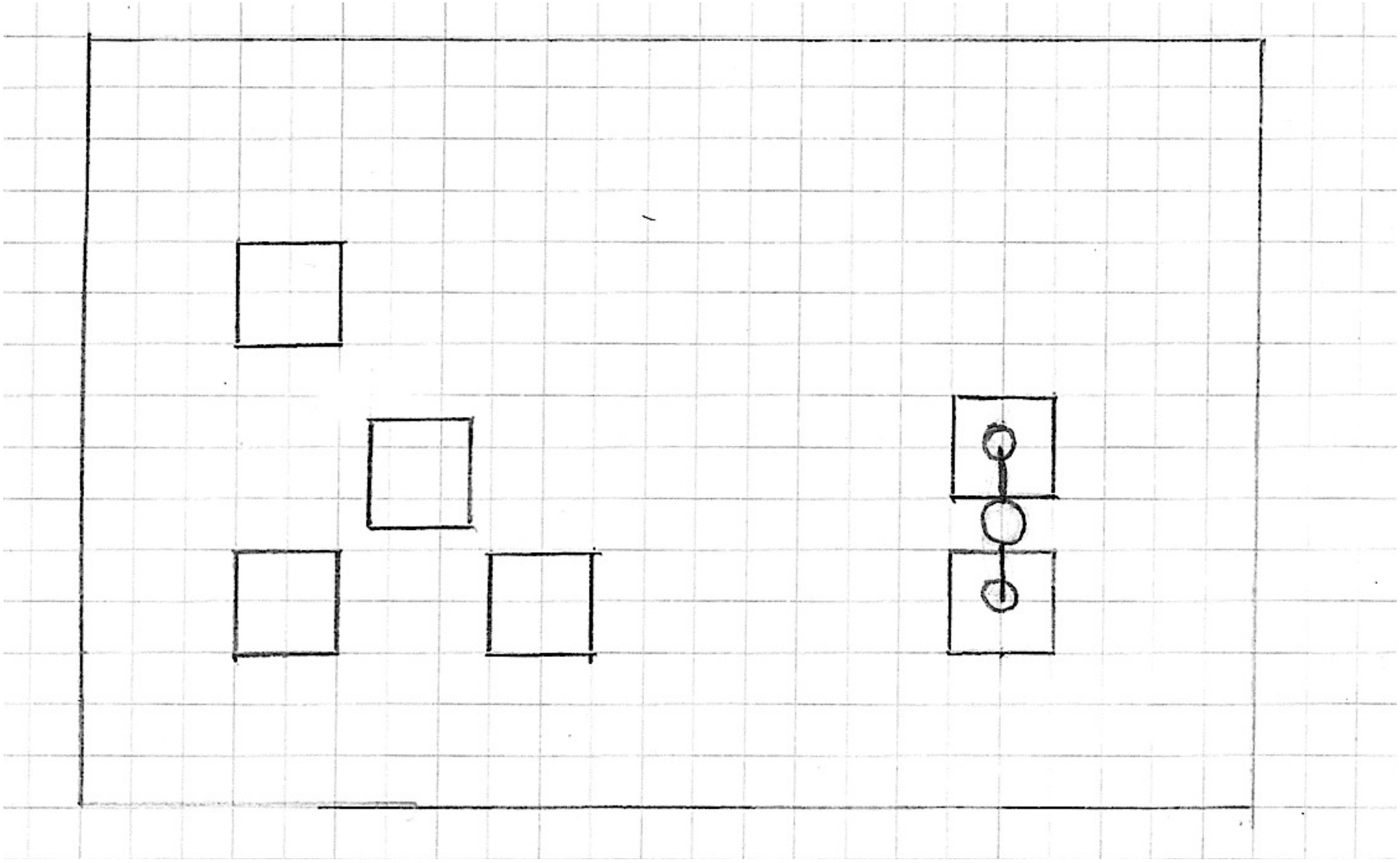

*****NOTE***** - This capacitor is polarized meaning that it has to be facing in the appropriate direction. Make sure you solder the NEGATIVE lead to the LOWER of the two squares. The negative side has the shorter lead and a vertical stripe along the side of the capacitor.

3. ADDING THE .1uF CAPACITORS (the two discs with “104” written on them)

- These capacitors will be positioned between the upper and lower left most squares and the ground plate. Melt a blob of solder on these two squares and two blobs on the ground plate beside each. (Reference the pictures for exact positioning).
- Re-heat the blobs and position the capacitors accordingly as in step 2.

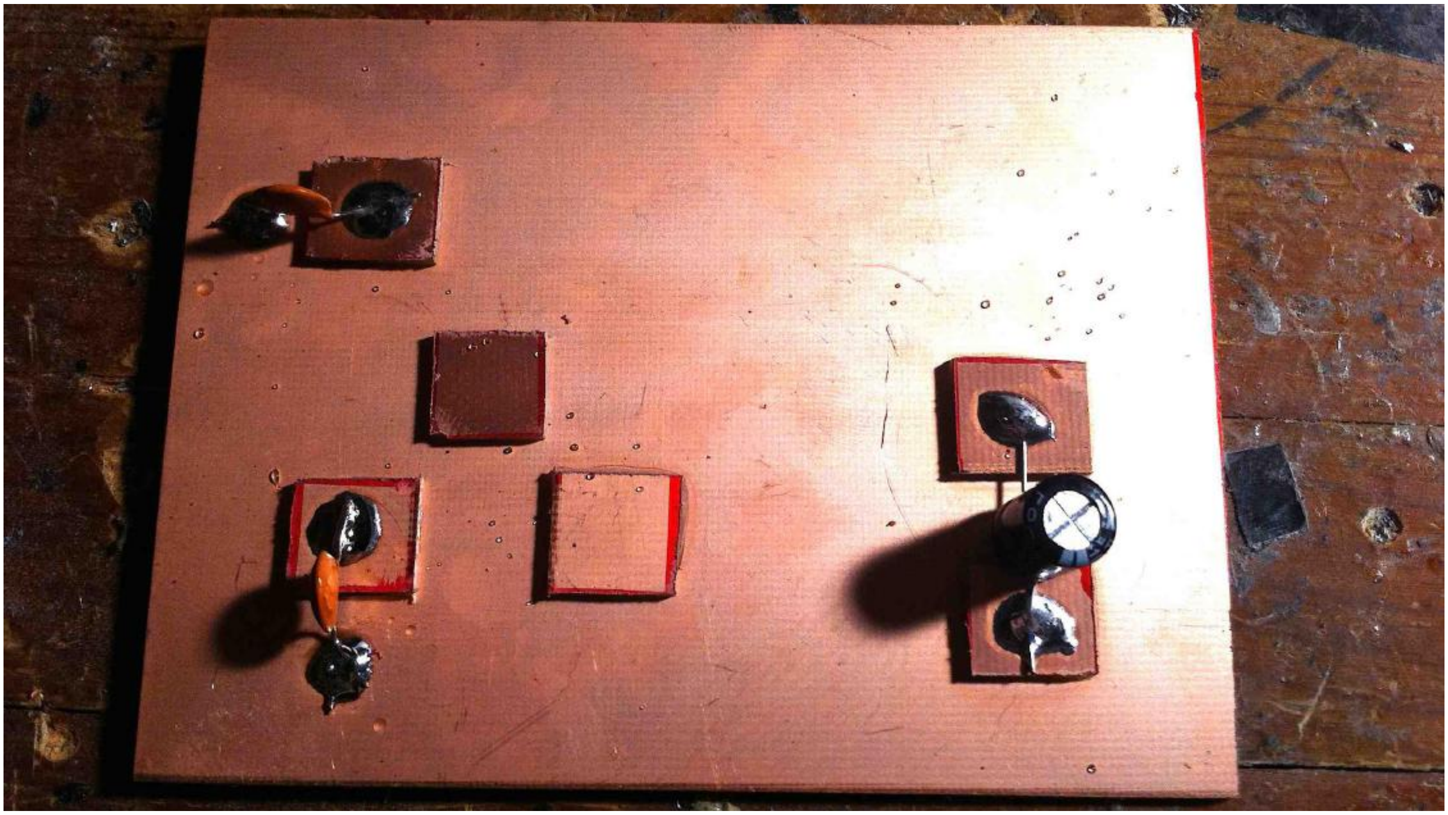

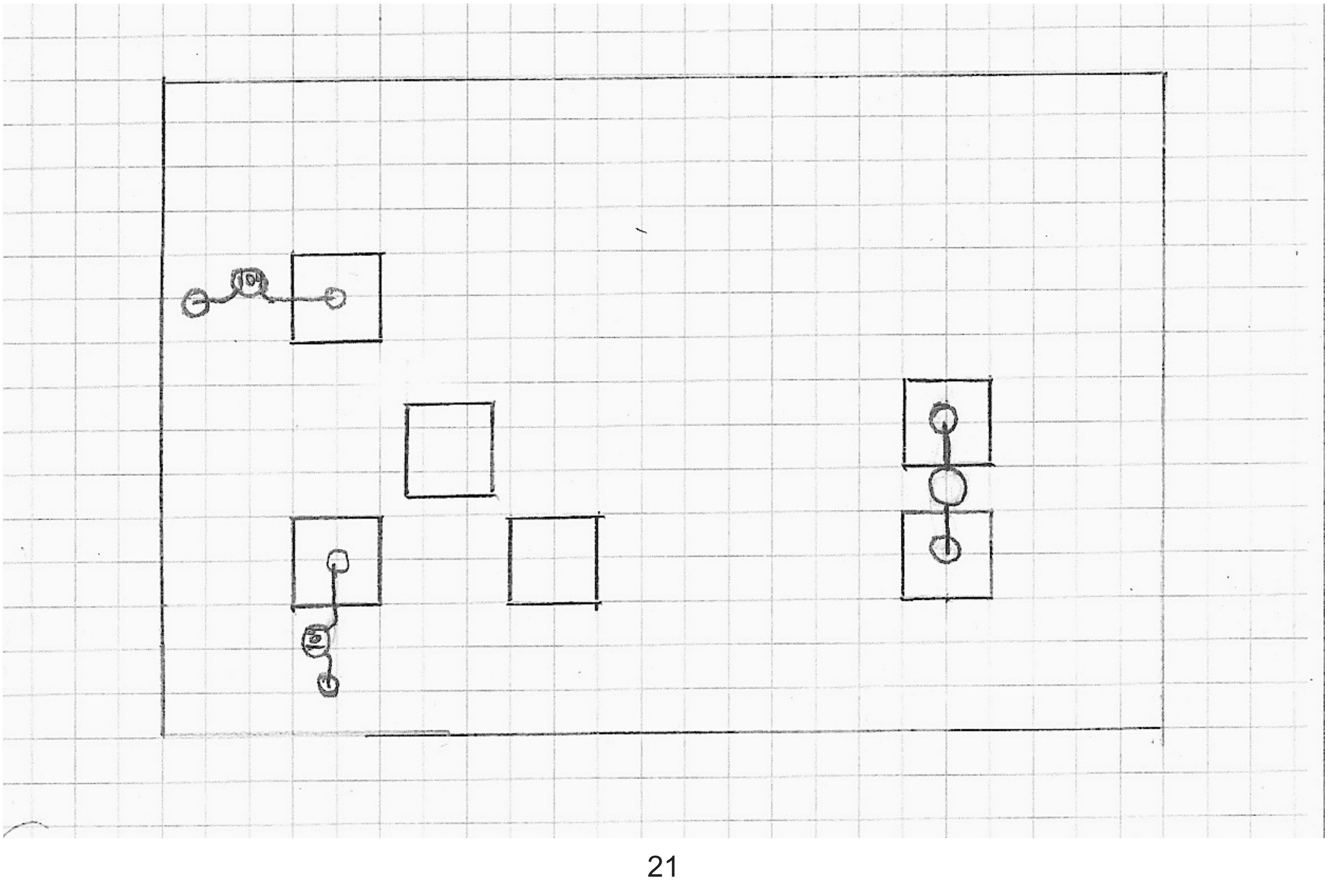

4. ADDING THE RESISTORS (10k and 27k)

- Solder the 10k resistor (brown, black and orange stripes) between the bottom left square and ground.
- Solder the 27k resistor (red, violate and orange stripes) between the two left squares.

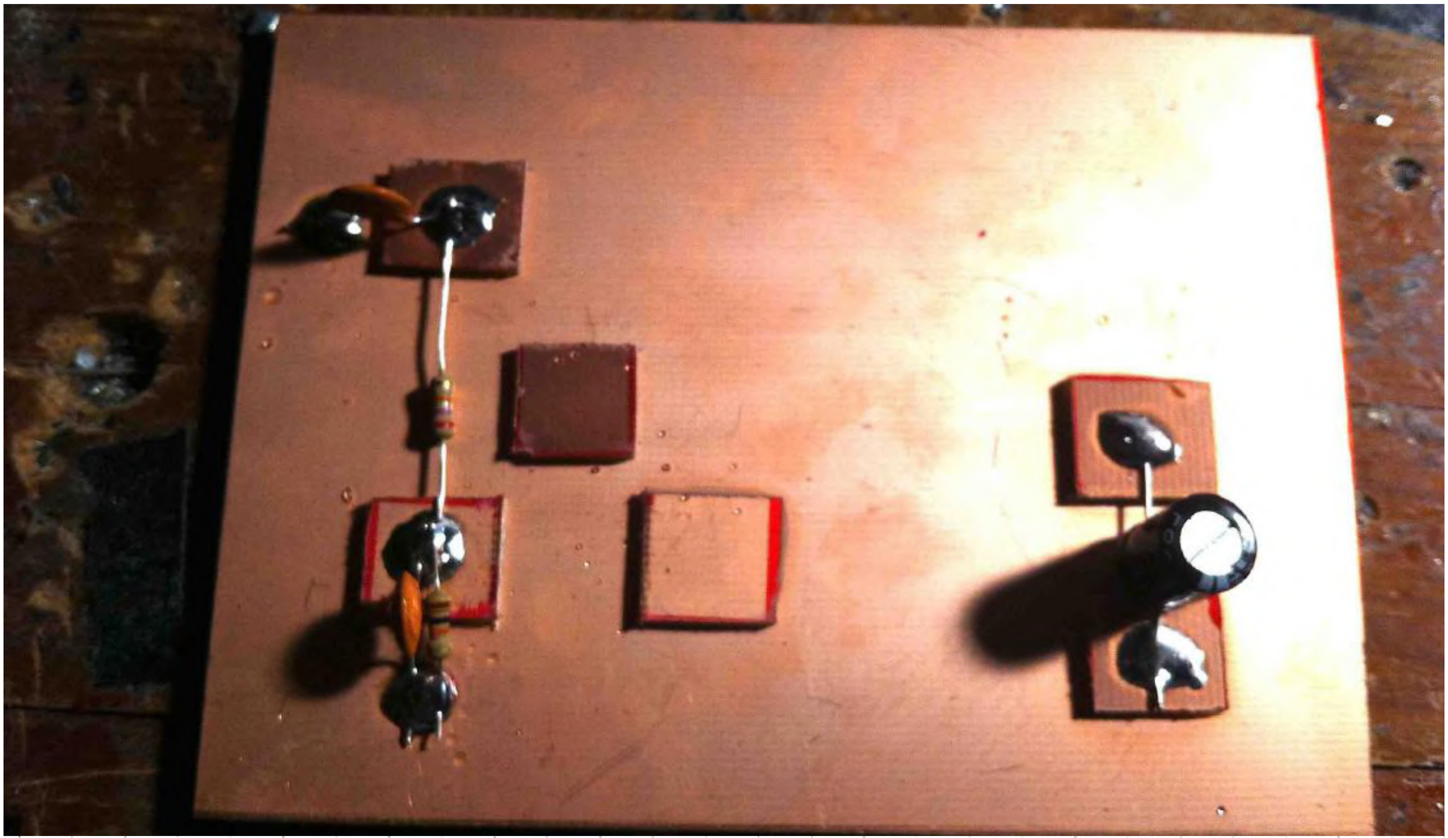

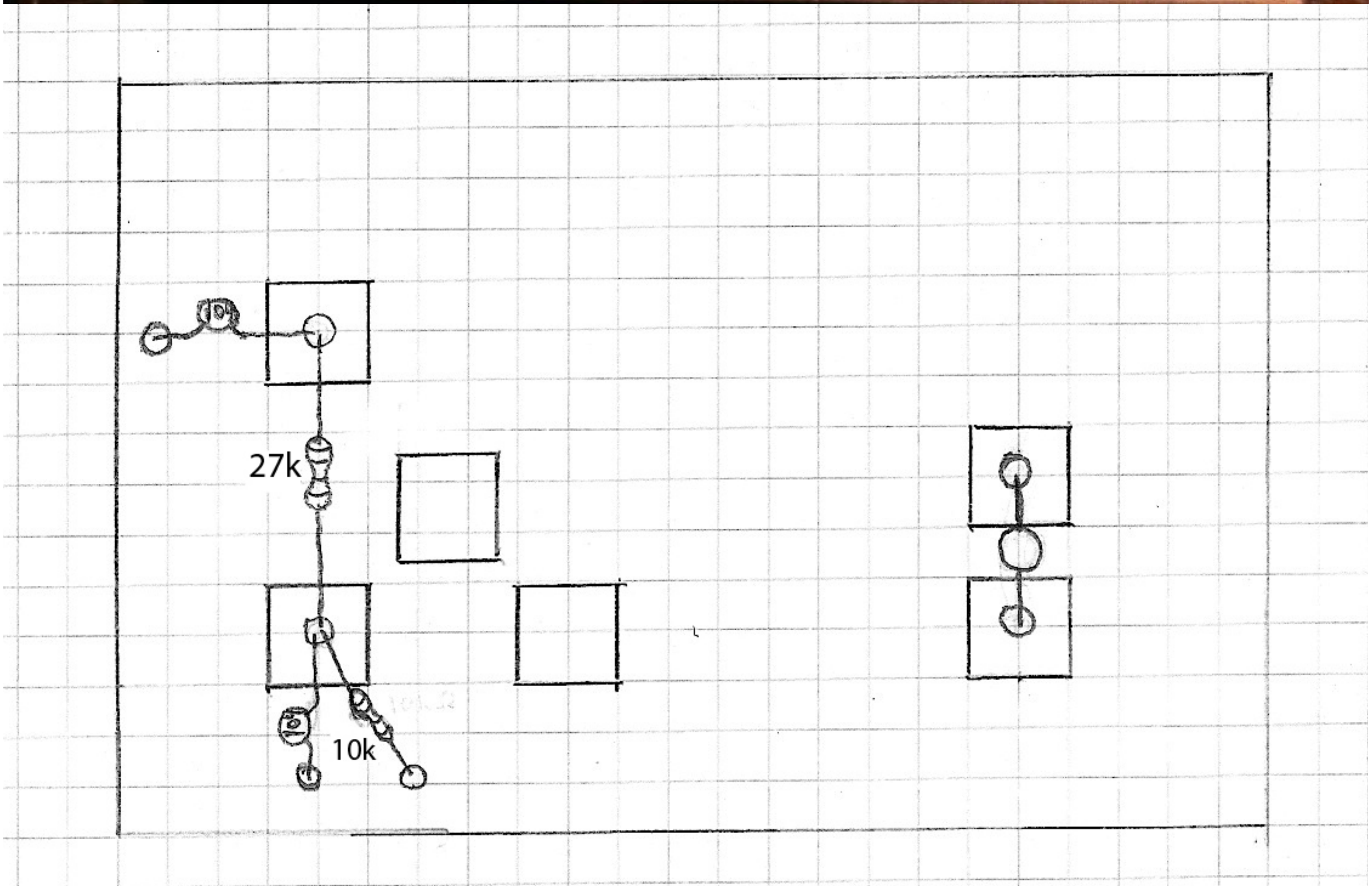

5. ADDING THE COIL

- Cut about 4” of the magnet wire (24 AWG or close)
- Use the plug from your headphones to wrap the wire around. Make 5 turns.

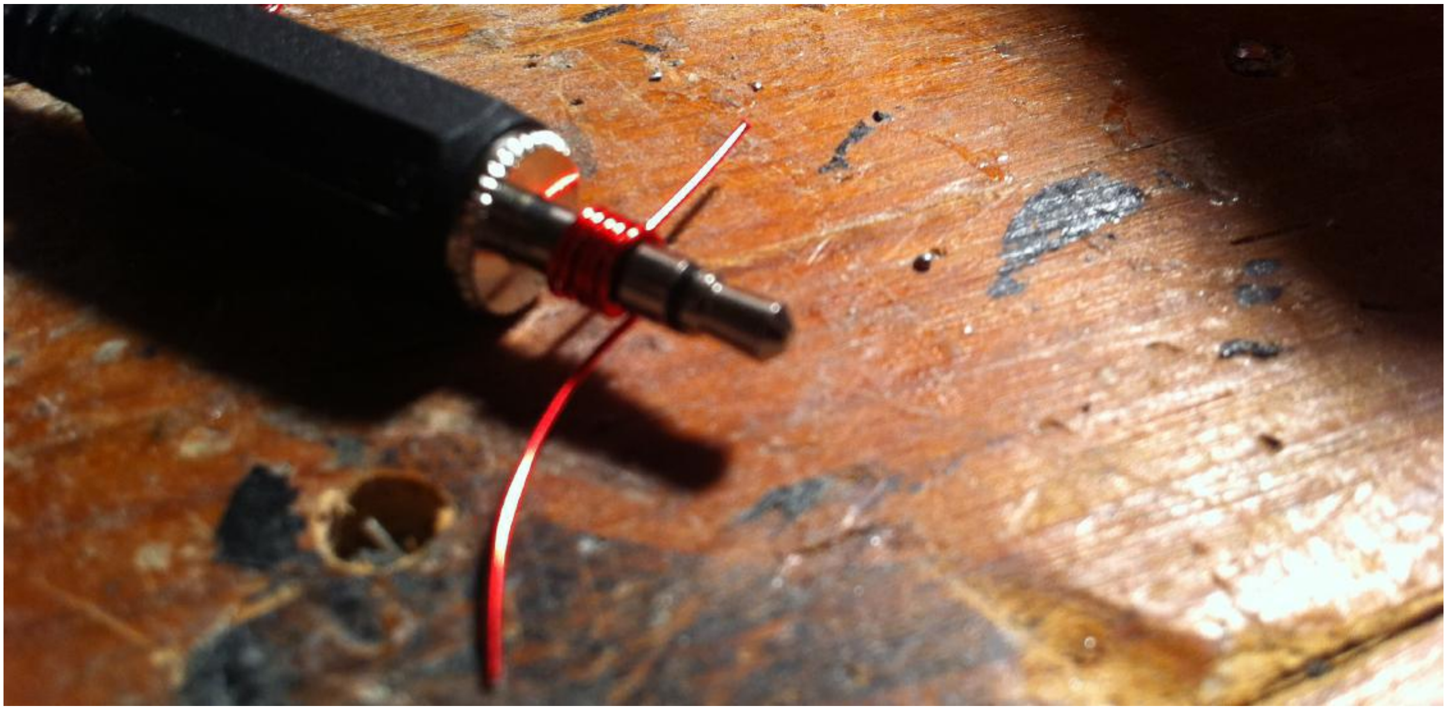

- With the wire still wrapped around the headphone plug, take a small piece of sand paper and rub off the colored insulation of each end until the bare copper is exposed (If the copper is not exposed, you won’t be able to make a connection).

- Solder the coil into place bridging the two squares as seen below

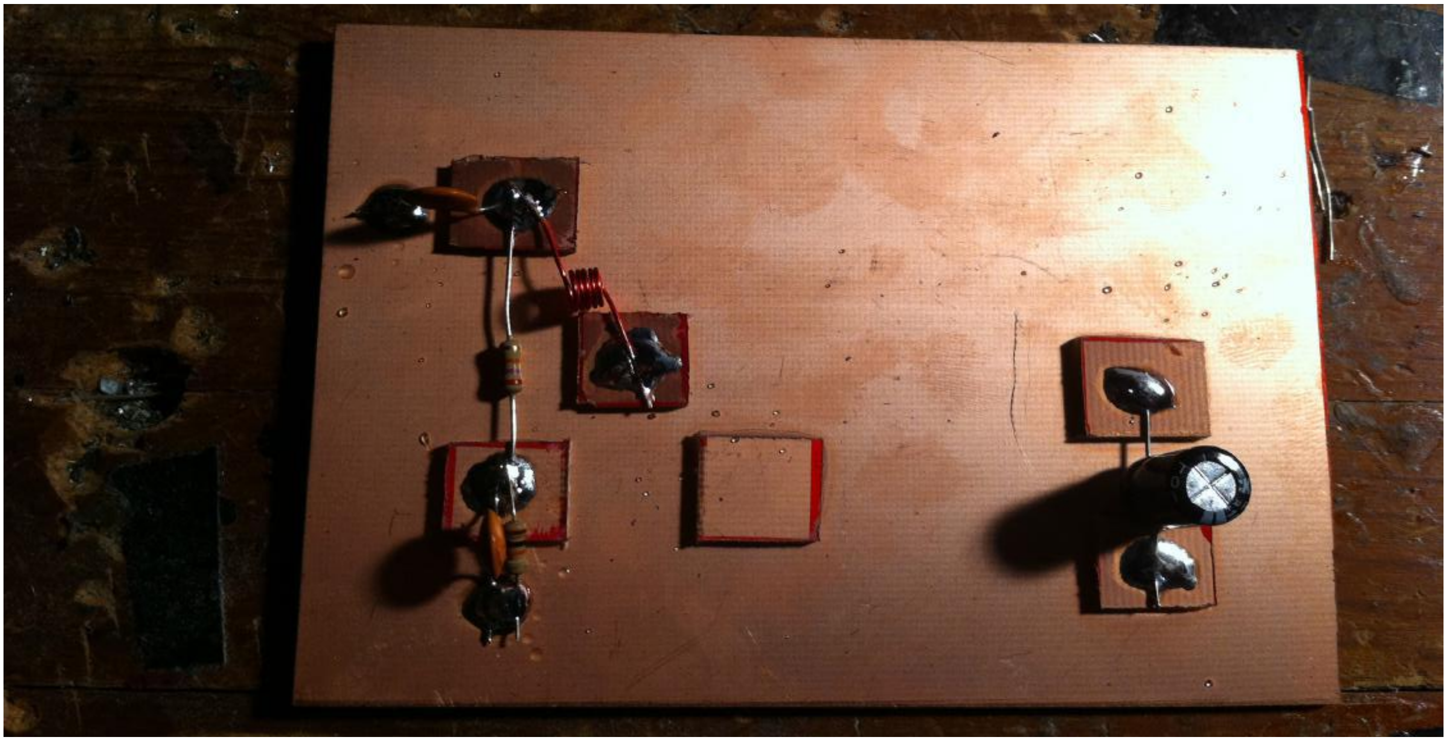

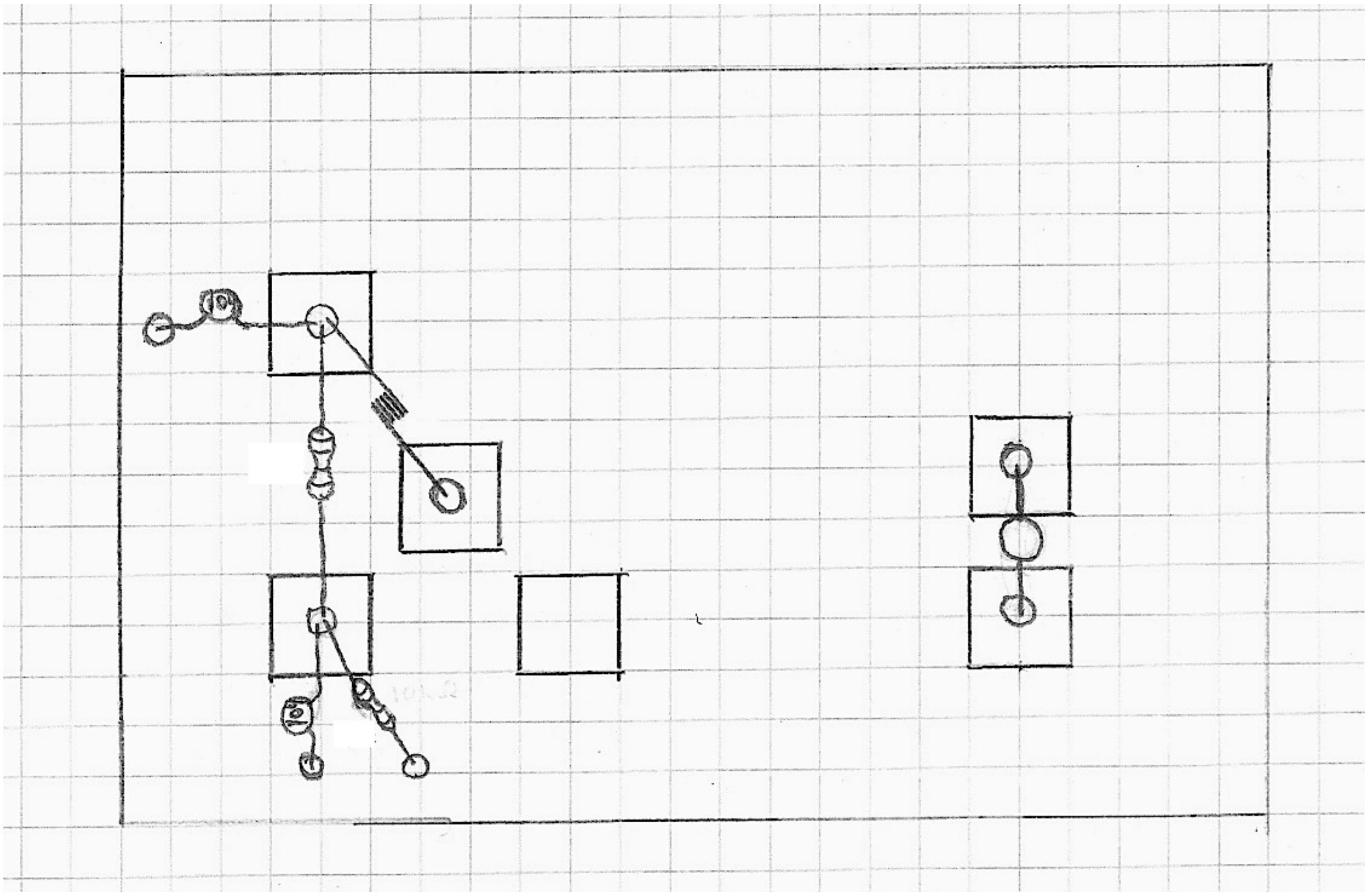

6. ADDING THE VARIABLE CAPACITORS

NOTE Once the component leads start to pile up on each other on each square, it helps to add additional solder.

- The first variable capacitor will be soldered between the top square of the three that make up the bottom triangle and the ground plate.
- Solder the “nose” of the capacitor (the lead that is different from the other two) to the square, and the other two side leads to the ground plate.

NOTE You will have to bend the leads in a certain way in order to get everything positioned properly. Just play around with it till you get it.

- The second variable capacitor will bridge between top and right squares of the triangle. Only two leads are needed here so use your clippers to remove one of the side leads, leaving the nose and only one side. (See the pic below)

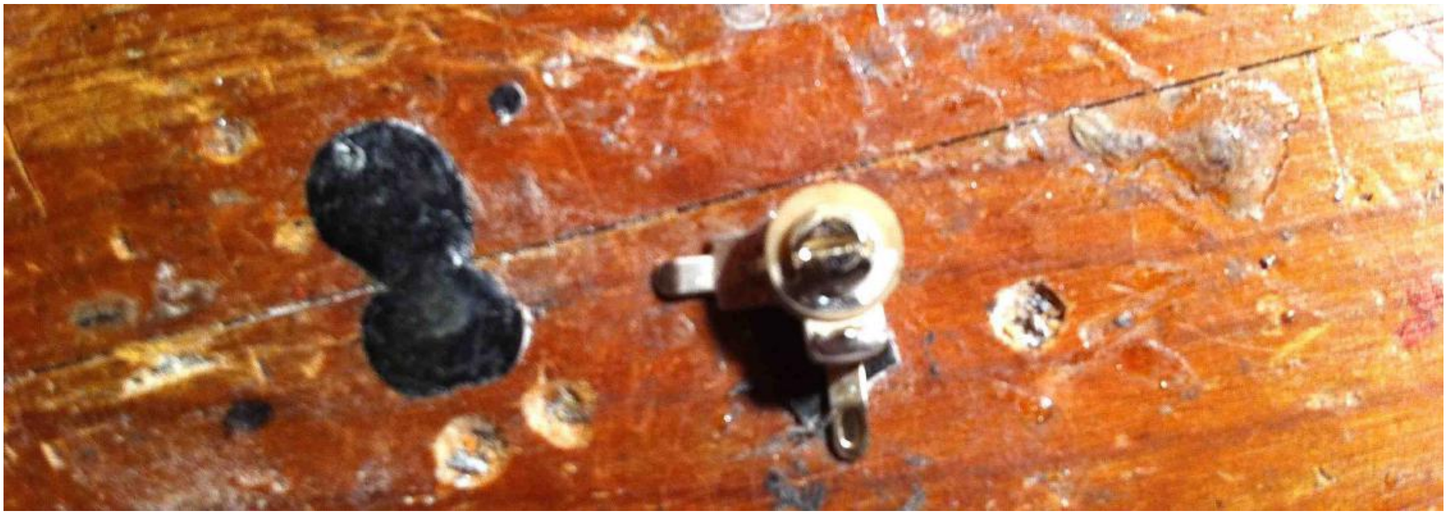

The variable capacitors soldered into place…

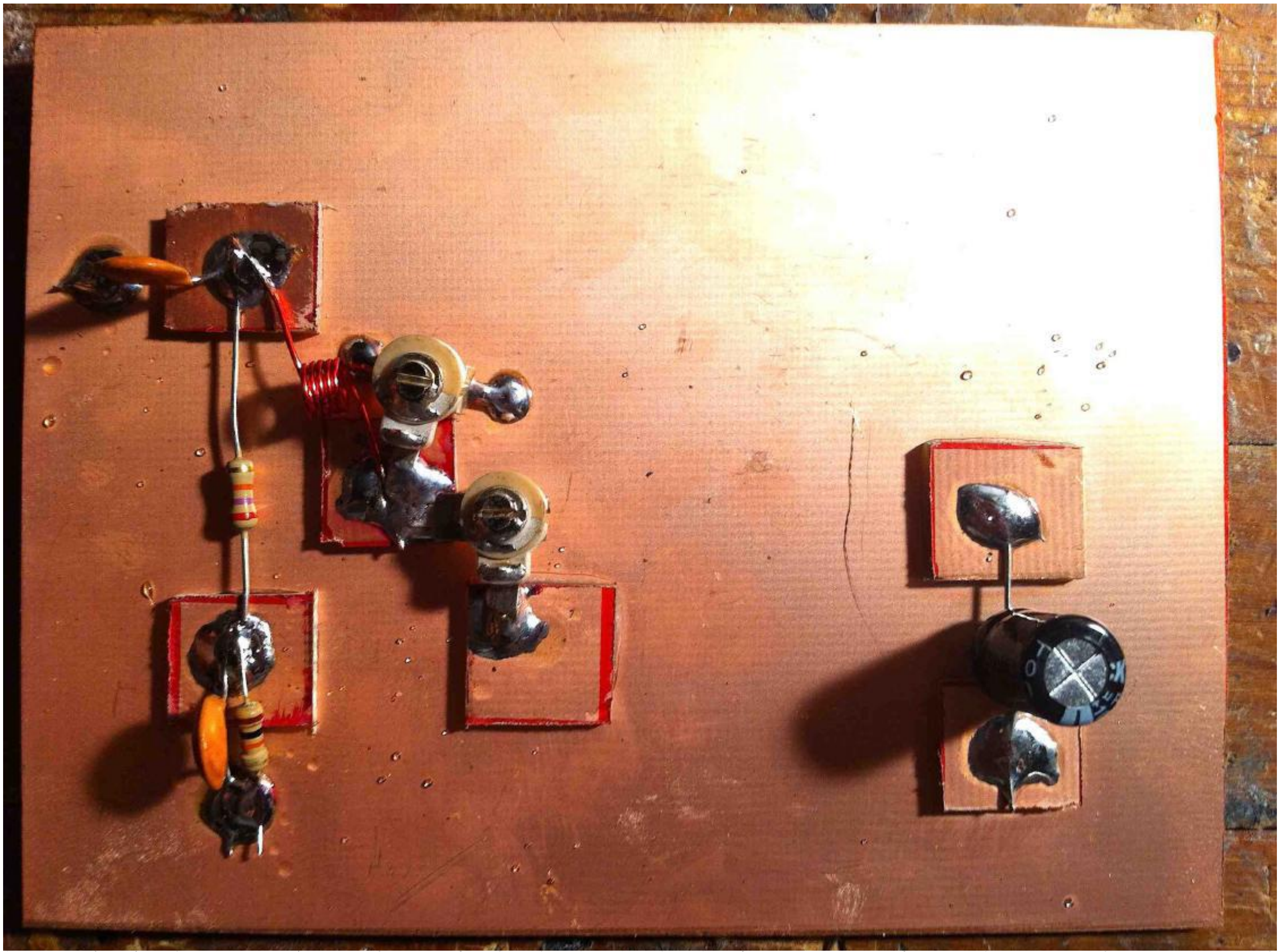

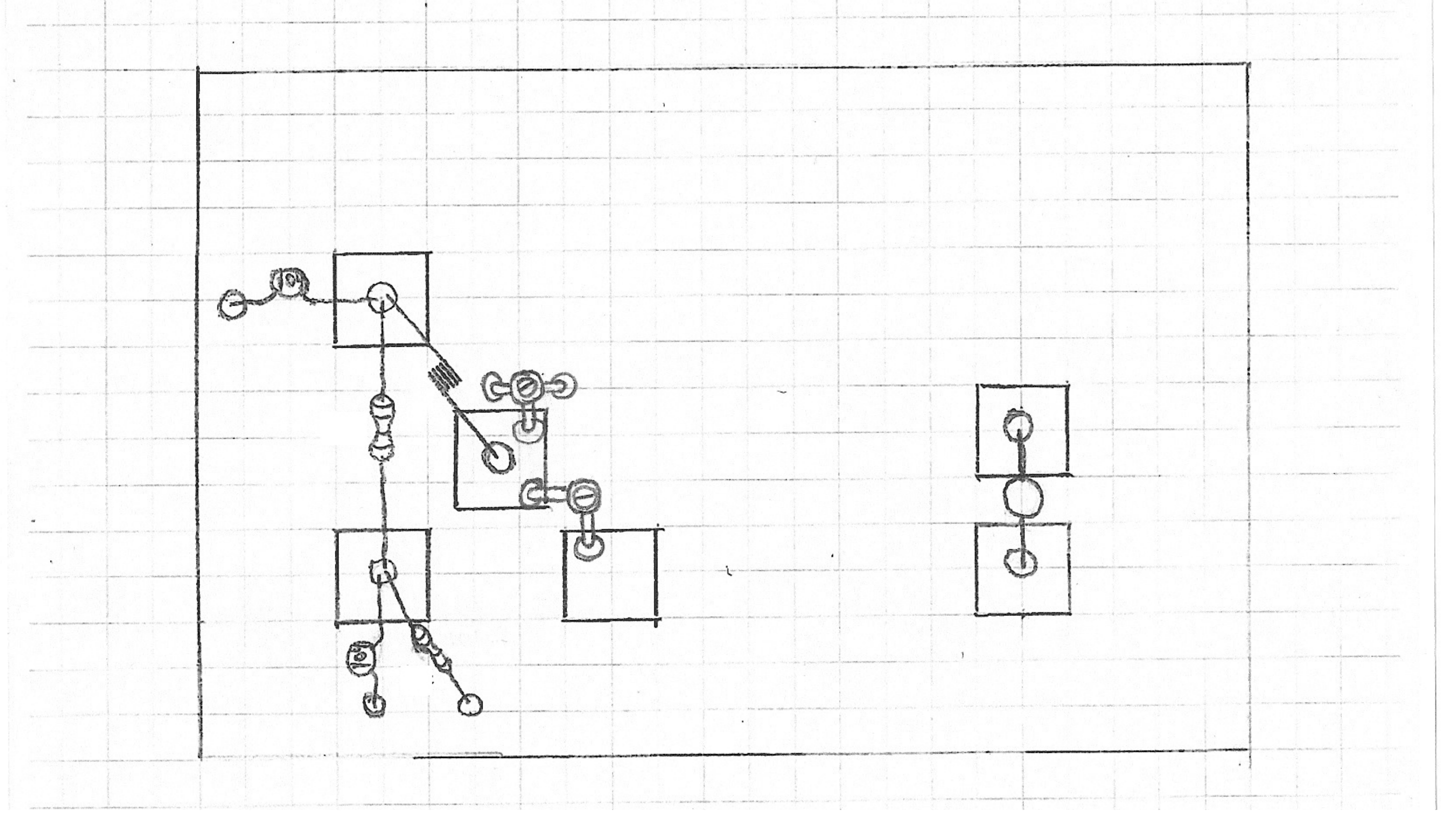

7. ADDING THE TRANSISTOR (MPSA18)

NOTE This transistor is sensitive to heat and can be destroyed by the heat from your iron. Be sure to follow the necessary precautions below. But don't worry if you do kill one, they are only ~$.10 a piece.

- The three leads of the transistor all do specific things and must go to the appropriate places. Here is a pic of the orientation of this transistor…

C=Collector B=Base E=Emitter

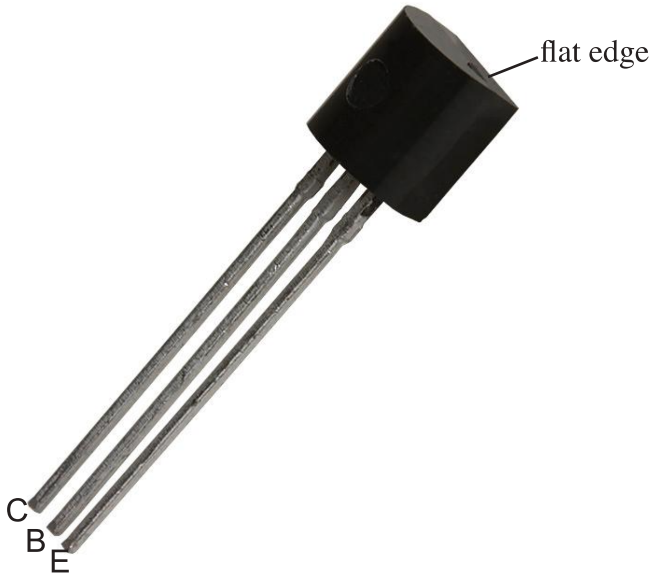
flat edge
C
B
E

- Each lead from the transistor will be soldered to the three different squares that make up the triangle, so you will have to bend the leads accordingly.
- The B lead (base) goes to the left square of the triangle, the C lead (collector) goes to the top square, and the E lead (emitter) goes to the right square, as in the image below.

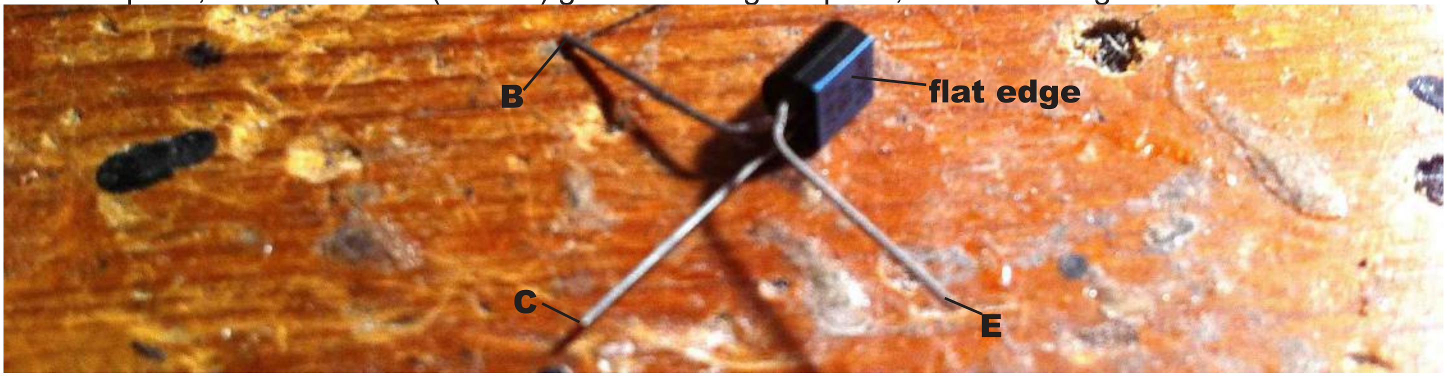
B
flat edge
C
E

- Before you solder the transistor in place, you need to have a "heat sink" attached to protect it. This will draw some of the heat from the transistor to something else. To do this, simply take an alligator clip and attach it to the lead you are soldering…

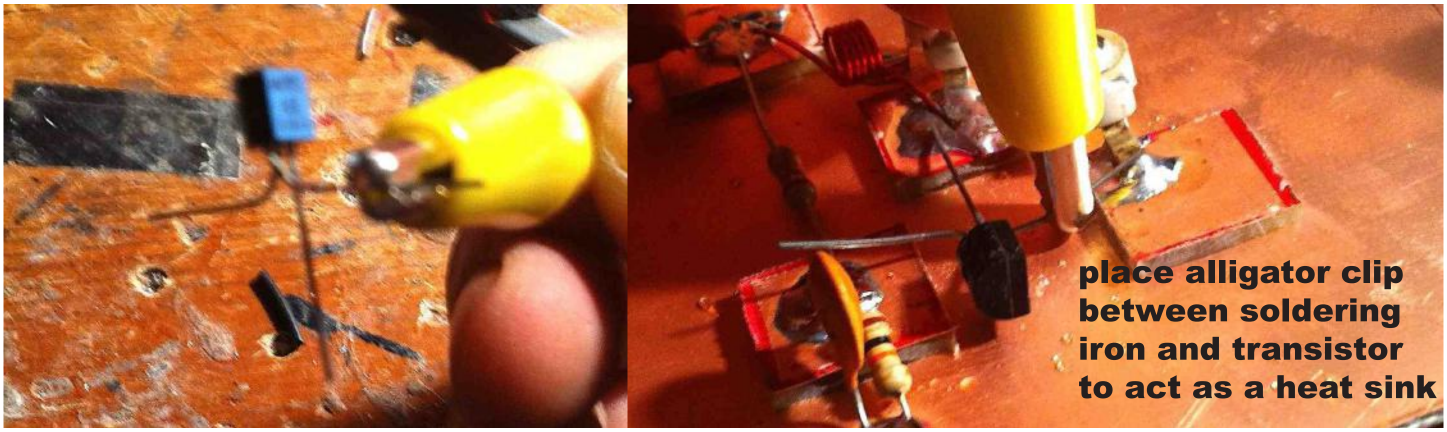
place alligator clip between soldering iron and transistor to act as a heat sink

- Solder each lead in place attaching the clip lead to each one before doing so.

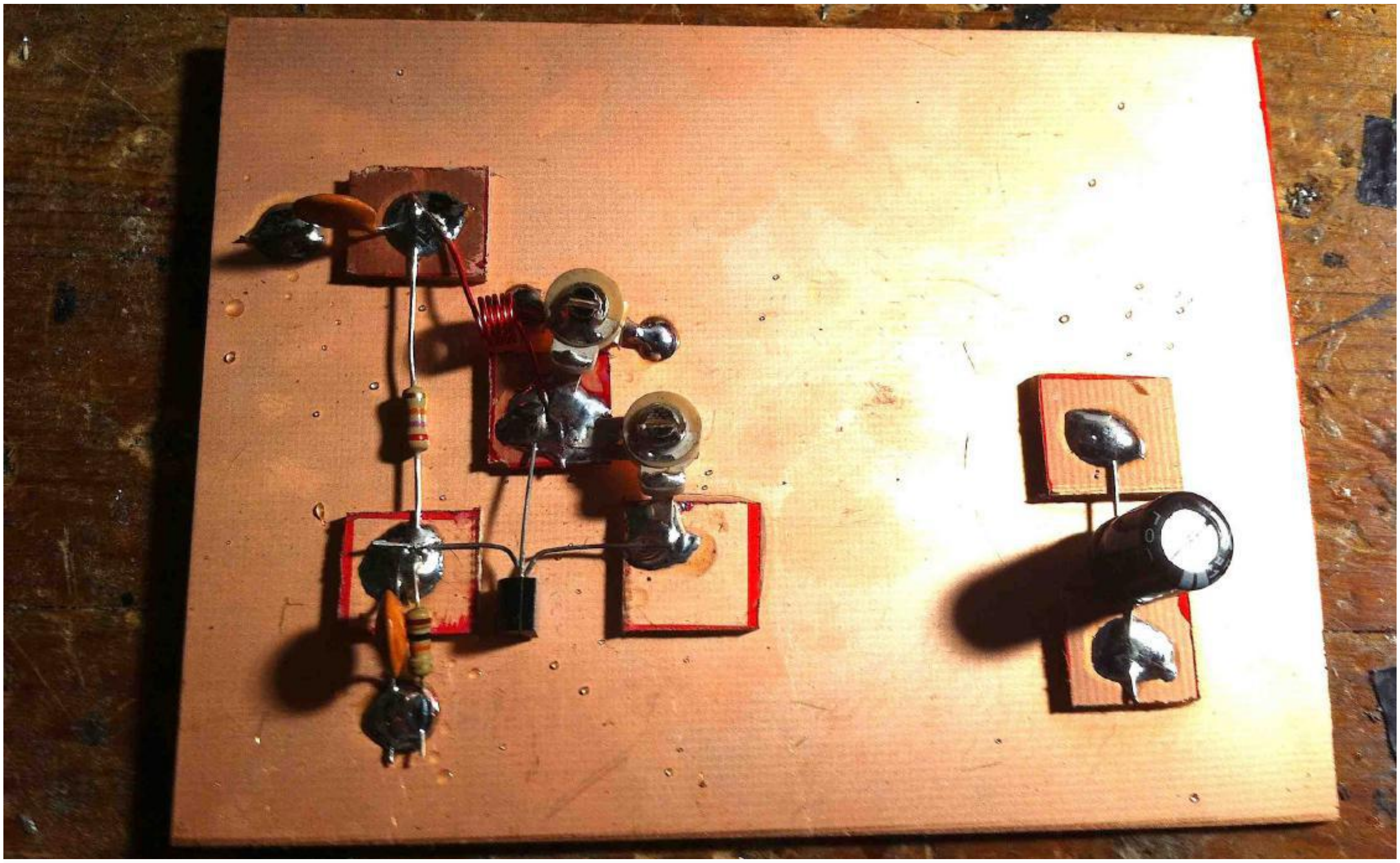

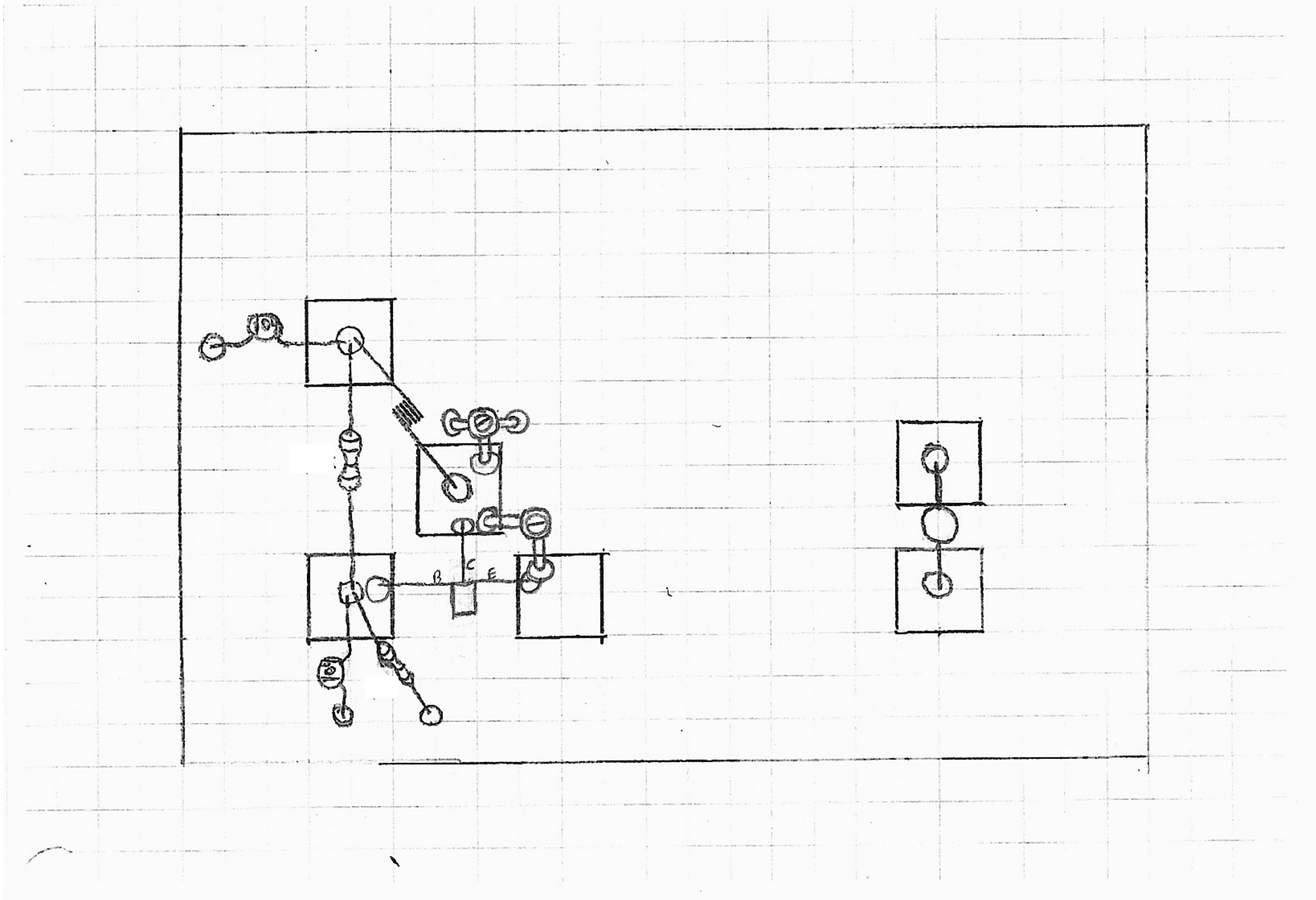

8. ADDING THE 1k POTENTIOMETER

- Put a bit of super glue or hot glue on the bottom of the potentiometer and stick it to the upper right corner with the leads facing towards the left of the plate.
- Take a small amount of stranded hookup wire and strip both ends. Solder one side to the upper most lead and the other side to the ground plate.
- Repeat the process by connecting the middle lead of the potentiometer to the upper most of the two right squares. Then connect the lower lead of the potentiometer to the right most square of the triangle.

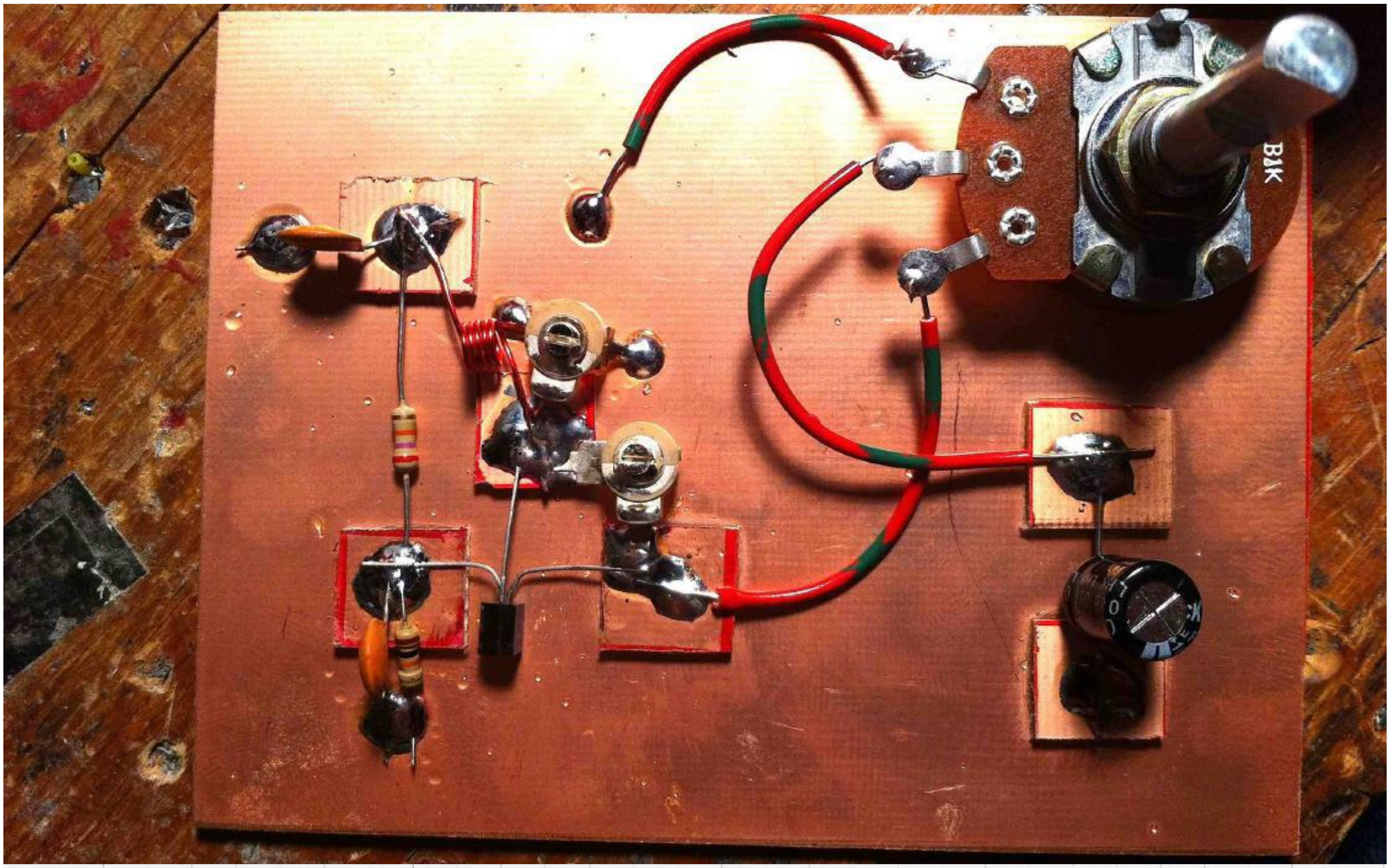

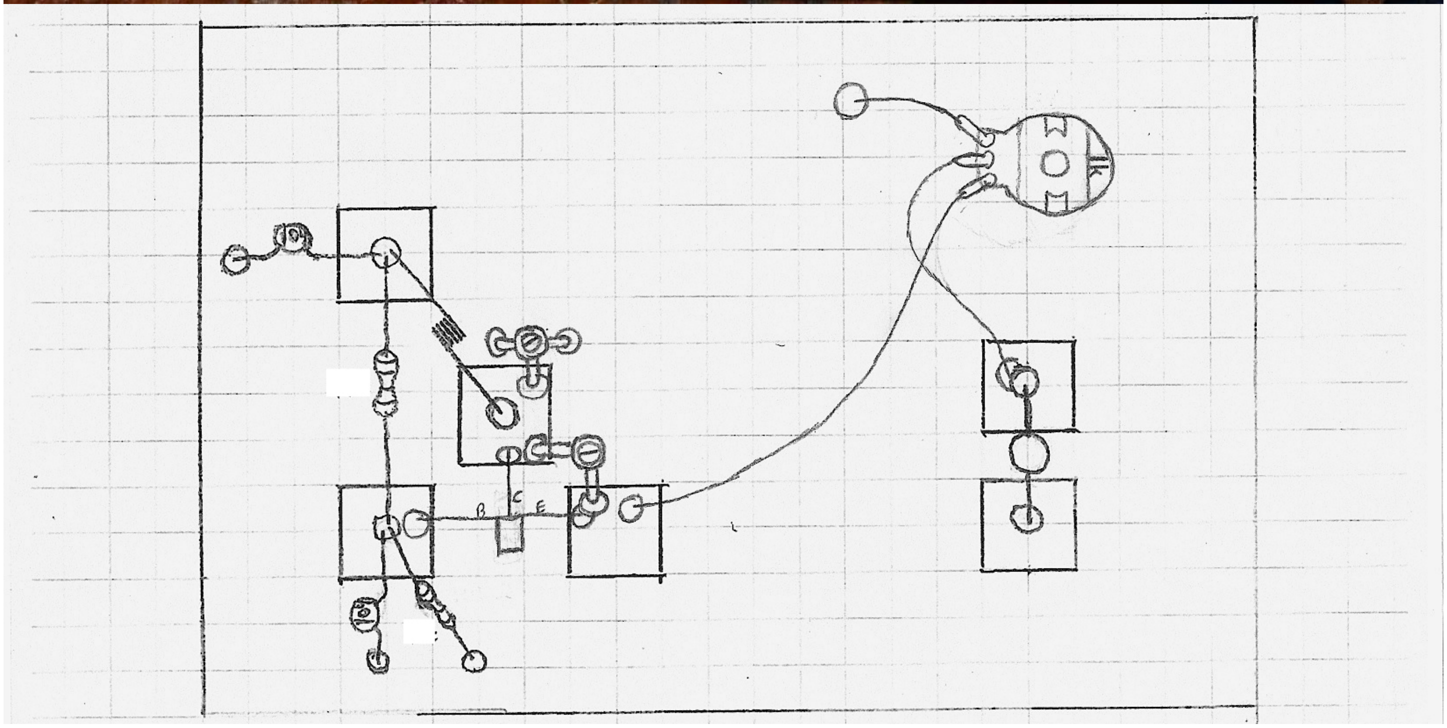

9. ADDING THE INPUT WIRES

- Cut two small lengths (~4-5”) of your stranded wire and strip both sides
- Solder one of the wires to the ground plate (this will be the ground input from your video source)
- Solder the other wire to the bottom square of the two right squares. (This will be the signal input from your video source)

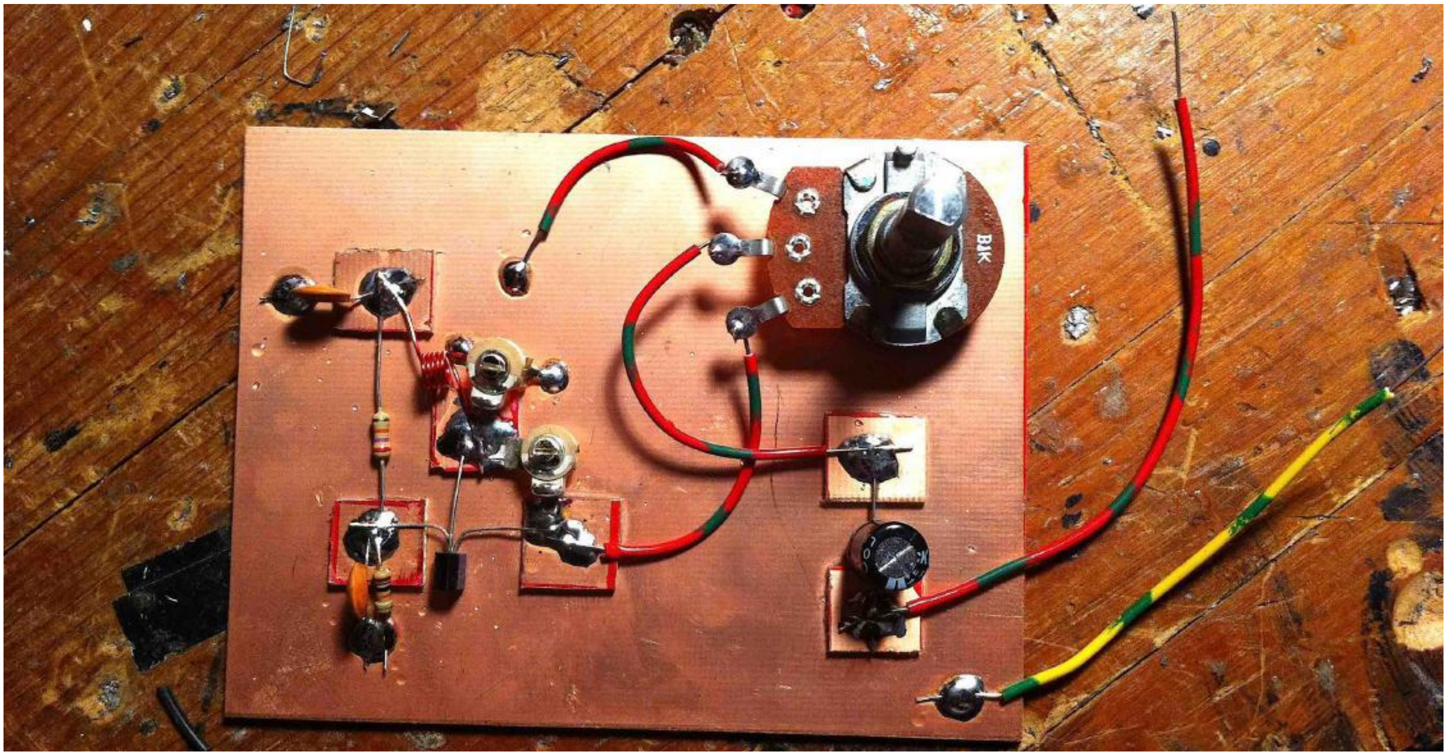

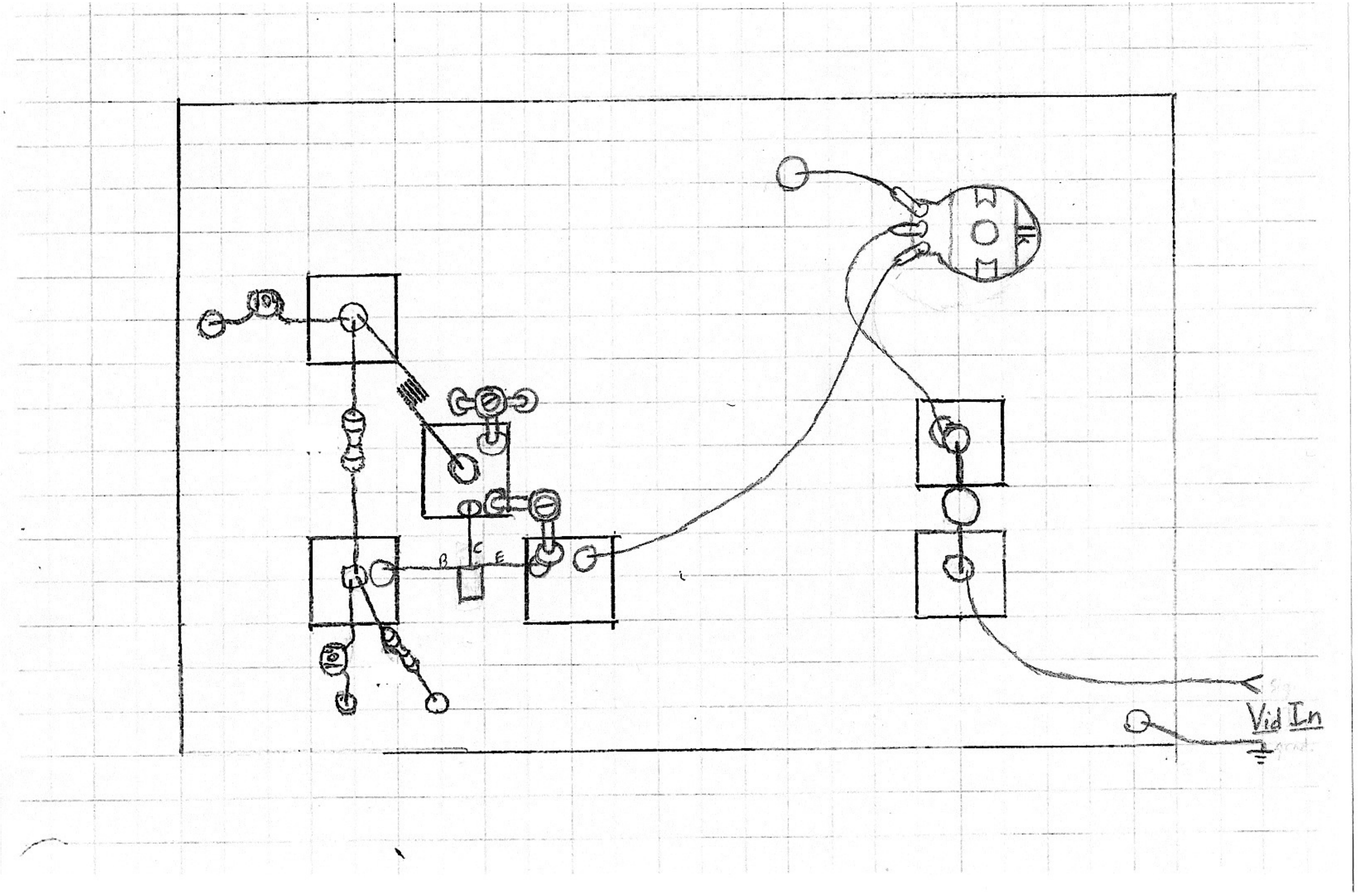

10. ADDING THE BATTERY INPUT WIRES

- Just like the last step, cut two shot pieces of wire and strip both ends.
- Solder one of the wires to the ground plate (this will be the negative input from your 9volt battery).
- Solder the other wire to the upper left square (this will be the positive input from your 9volt battery).

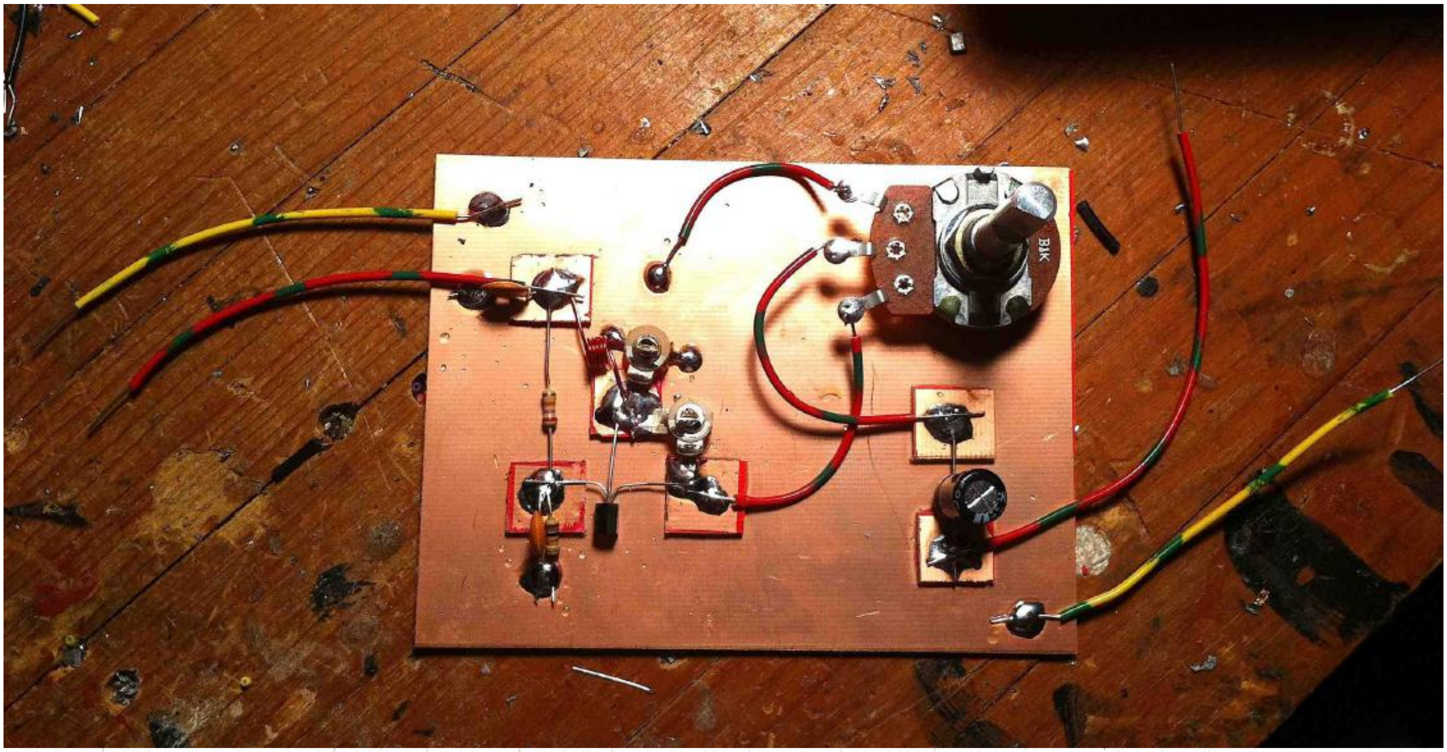

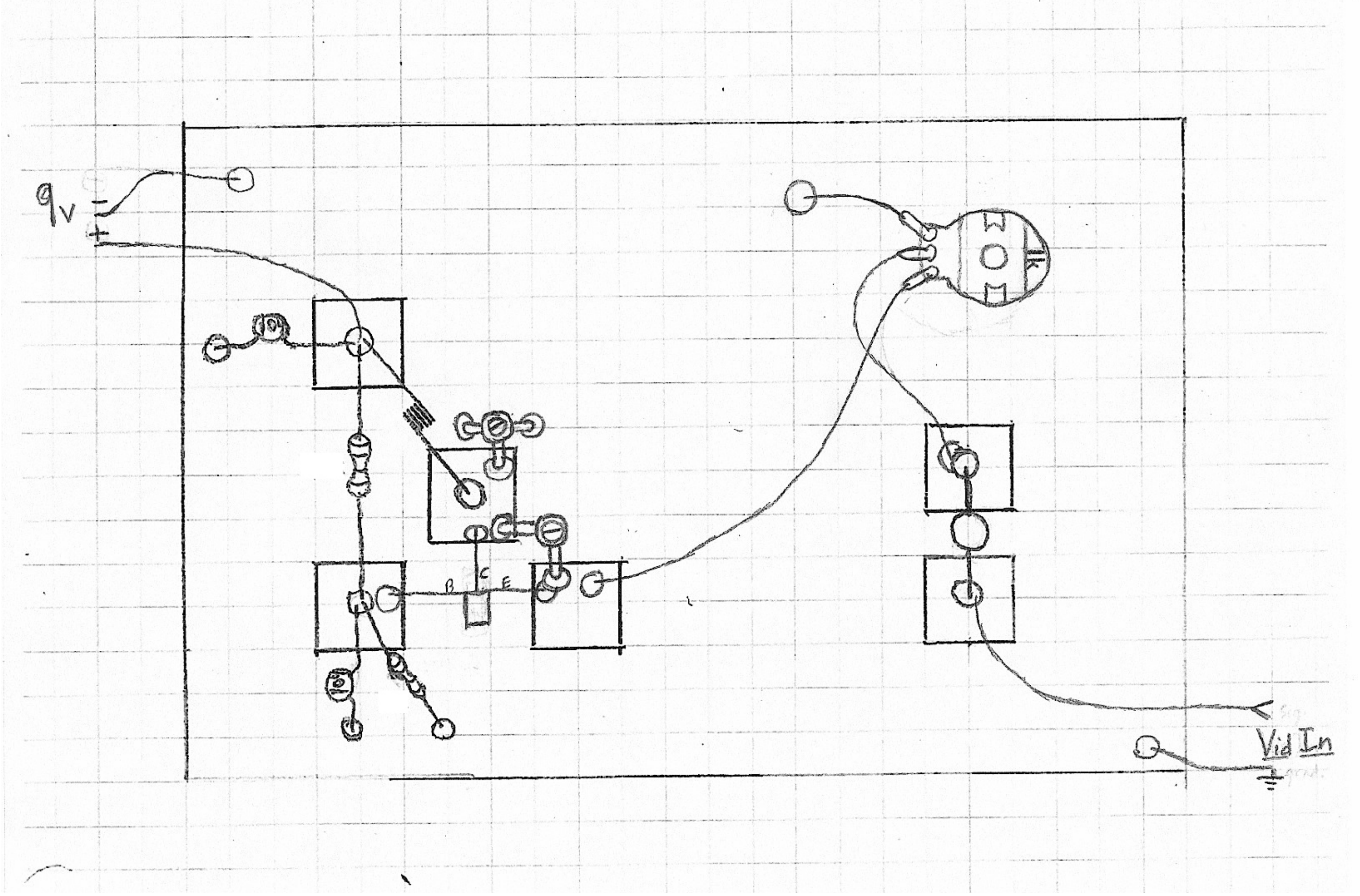

11. ADDING THE ANTENNA

- Cut about 2 feet of wire and strip one end.
- Solder the wire to the right most square of the triangle.

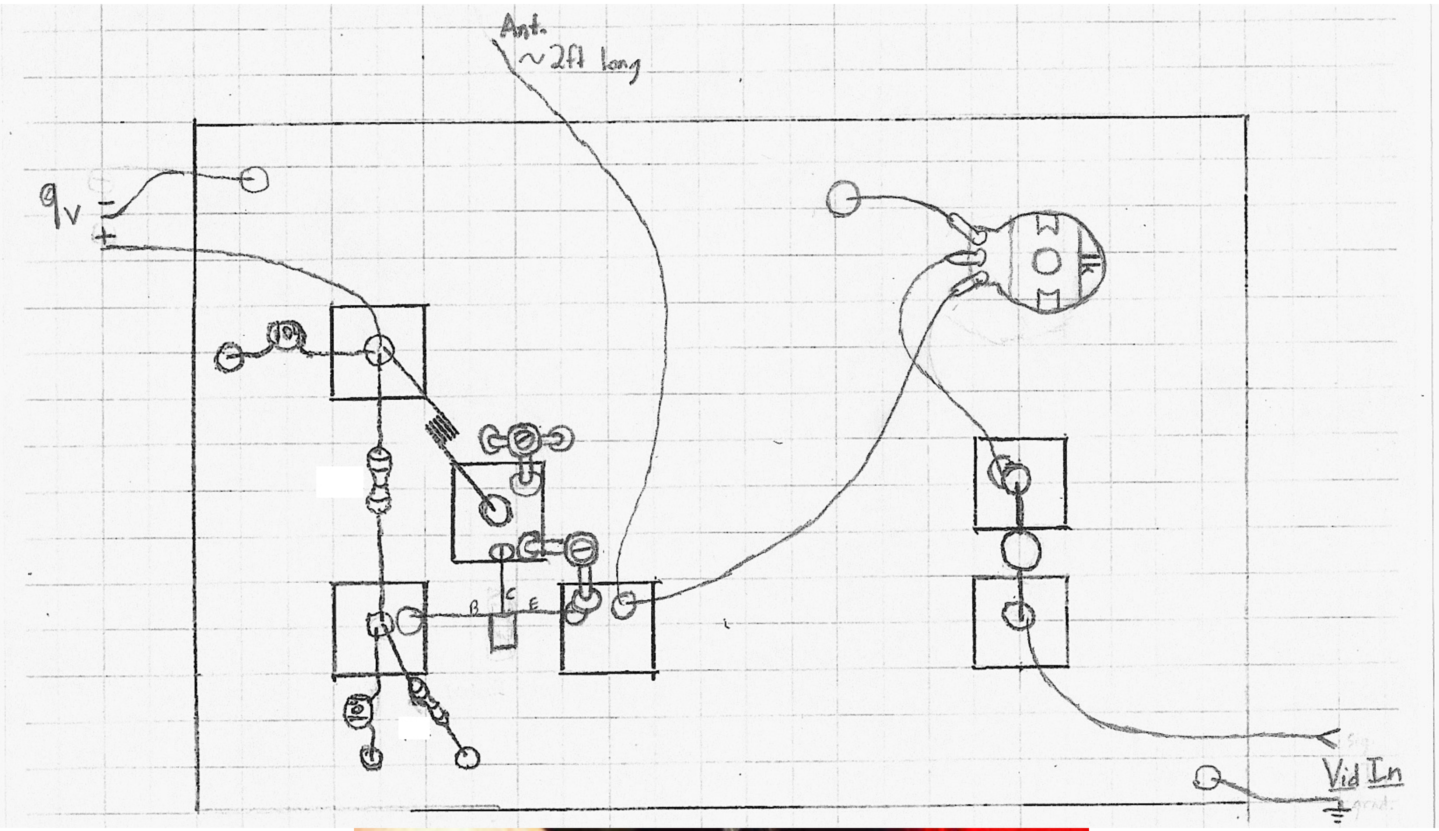
Ant.
~2ft long
9v
-
+
B
C
E
sig.
Vid In
gnd.

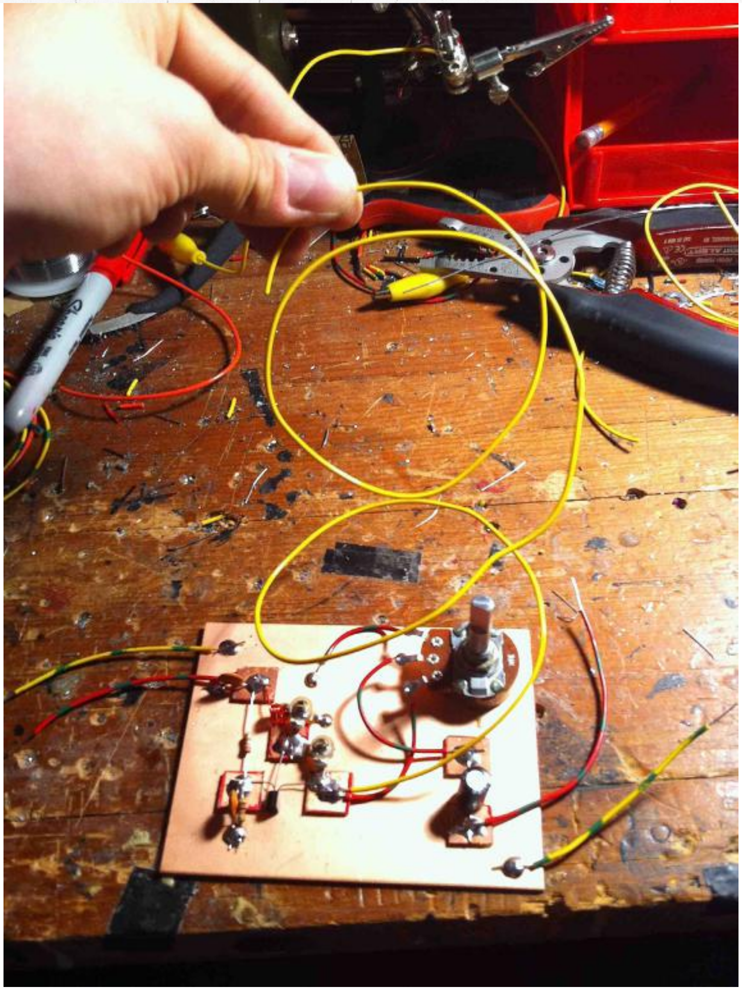

OPERATING

- Find a video source (DVD player, VHS player, iPod etc.) that has a composite video output.
- Use an alligator clip lead to connect the ground from your video source to the ground input of the transmitter. Using another alligator clip lead, attach the signal output from your source to the signal input of the transmitter.
- Use two more alligator clips to connect the positive and negative of a 9v battery to the positive and negative battery inputs of the transmitter.
- Turn on a CRT “tube style” TV and tune it to channel 2 or 3.
- Turn your potentiometer to somewhere around 50%
- Using a small screwdriver, tune the upper most variable cap slowly until you see the TV make a flicker or change.
- Tune the second variable capacitor until you start to see some strong signal coming through
- Tune the potentiometer and the two variable capacitors until you get the imagery you wan

Send any comments, questions or suggestions to yaktronix.online@gmail.com

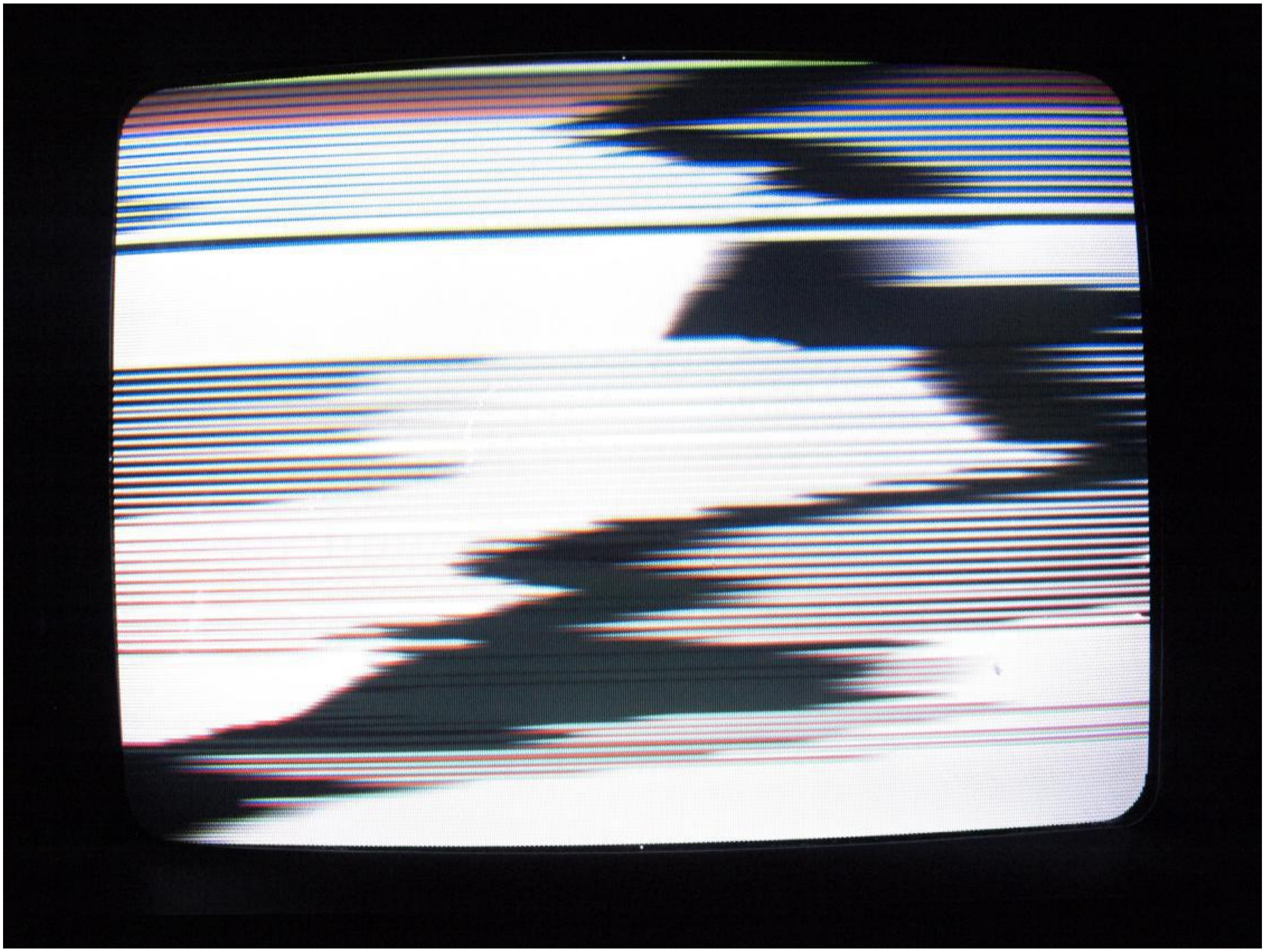

rexresearch.com

Paul BAUMANN

Testatika Generator

INNOVATIVE SEMICONDUCTOR SOLUTIONS
life.augmented
Digi-Key ELECTRONICS
Learn More

Methernitha Group: CH-3517 Linden, Switzerland

Photo Gallery & Videos

Links:

Methernitha ~ http://www.methernitha.com
http://utenti.tripod.it/testatikmachine/
Paul E. Potter ~ http://website.lineone.net/~aarekhu/index.html

---

Hans Holzerr: Testatika Demonstration Notes

Paul Potter: Back-Engineered Testatika

P. Potter: Principle Experiment

Anonymous: Believe It or Not --- Here It is!

Cyril Smith: Testatika Generator and Over-Unity

Methernitha: The Research Work of Methernitha in the Field of the so-called Free Energy"

P. Potter: Transcript of Methernitha Testatika Video

Testatika Demonstration (4 August 1999)

Translation by Stefan Hartmann < harti@harti.com > & Hans Holzherr

From: Hans Holzherr
To: Stefan Hartmann

Recently, over 30 technicians and engineers (most of them retired) were allowed to visit the Methernitha group in Linden, Switzerland, where they witnessed a demo of the different Testatika machines.

Here is a report from Hans Holzherr from Switzerland who was present:

Hello Mr. Hartmann,

To your questions: > Have you seen live a machine with a load? If so, what load?

I am referring in the following to the model with the 50-cm diameter disks. This machine was already running when the visitors stepped into the room, and was not halted during the whole time --- we were there for about 1.5 hrs. As a first load a 1000-Watt lamp was connected for approximately 10 seconds whose brightness did NOT diminish --- the corresponding sequence on the Testatika film is just an effect of the camera aperture's automatic adjusting to the sudden brightness! The second load was a U-shaped heating element, that Mr.Baumann handed to me. It became so hot within one second that I had to put it down immediately! What was particularly impressive was that while he pulled back one of the contact wires (that was with the lamp, I believe), a 1-cm long arc appeared between the output electrode and the connecting wire for approximately one second. The apparatus was under a plexiglass hood. Near the base it had two holes which Baumann used to insert the contact wires to touch the output electrodes.

How do you think the high wattage is produced ?

Good question! I'd love to know the answer, too!

Did the disks slow down when a load was placed across the output electrodes ?

I did not notice that (nobody else did), but of course you tend to turn your look to 'where the action is' (the lamp etc.) The disks turned with 15 rpm, which is quite slow. The spin rate was regulated magnetically.

What general impression did you have ?

It was really impressive! One can hardly believe it, with this slow rotation. In any case, this cannot be explained in terms of bare electrostatics in the sense of the Wimshurst machine. The perforated sheets seem to have a key function... Beside the pick-up and the drive electrodes there are a number of small plexiglass blocks with glued-on perforated sheets, whose function is unknown.

As Adolf Schneider already mentioned, my colleague Bernhard XXX and I want to try to copy the principle experiment shown by Baumann --- without much hope to find anything extraordinary, though.

The device consists of a horizontal swiveling plexiglass arm with a small rectangular plexiglass plate at both ends glued to the lower side of the arm. The lower side of the arm is covered with perforated aluminum sheets (square holes), while the bottom of the plates is covered with brass wire mesh. Beneath each plate five additional plates are glued onto the base plate. There is also wire mesh between each pair of plates in the two blocks. From the mesh layer between the lowest plate and the base a wire goes to the two capacitors, which are connected in parallel . Baumann seized the arm with both hands and turned it about ten times back and forth (a full rotation was not possible, because the capacitors were in the way), then measured the DC voltage with a digital measuring instrument: 60 Volts. Then, as he short-circuited the condensers a loud crack could be heard. I don´t know if that already is an abnormal result...

On my question Baumann replied that with metal foil (instead of wire mesh) the device would not produce that effect.

For more information, Contact: leo@zelator.in-berlin.de

"Back-Engineered Testatika"

by

Paul E. Potter

http://website.lineone.net/~aarekhu/index.html

(Copyright 2000)

That the Swiss Methernitha group's Testatika machine is thought to be based on a Wimshurst electrostatic generator, is only a sparse approximation of the truth --- of the great multitude of electrostatic influence machines developed around the 1900's it more closely follows the charge-separation-and-collection system used by the 1898 Pidgeon machine [note 1] for its electrical circuit.

Its 50-per-disc steel grilles or 'gitter-grilles' are plainly unique to the Methernitha (see fig.1) but in principle follow on from previous research and patents for corrugated sectors which were found to be more efficient charge carriers [note 2] than flat ones, and from a similar example in more recent times of aluminium rods extending out like wheel spokes from an insulating hub of perspex [note 3].

Another unique function of these perforated grilles attached to the discs is how they induce charge from the rotating discs onto the special collecting pads, or 'tasten' antennae keys (which are also perforated --- so as to more readily pick up charge); for in a Wimshurst you had conductive brushes or rails of sharp points which actually touched the discs or were placed very close to them, but in the Methernitha the charge has to be made to traverse a parallel air-gap to the pads and for this purpose the metal gitter-grilles are so designed to create miniature eddy-currents of charged air which circulate in and out of the perforated metal's surface charges, and are more easily bounced out to the collecting pads. This process is categorised as VARIABLE CAPACITANCE electrostatic generation.

Careful note needs to be made of how the Methernitha uses its basically Pidgeon setup with regard to its neutralising rods (that equalise and stabilise the opposite charges --- see fig.2), and how charges are picked up from one area and accumulated at others, so that the polarities of charge are distributed correctly to specific areas on both discs [note 4].

And although there have been some fanciful claims, or misinformation, that it uses all sorts of radioactive materials to achieve its pulsed output I most strongly believe that the auxiliary electromagnetic circuit, that wraps itself around the rotating discs, portrays a simple electronic approach; afterall, who would use radium radioactive emission alongside leyden jar capacitors ! Indeed, the more you look into certain elements of its construction the more they point to three main eras of electronics development, the 1900's, the 1920's and the 1950/60's. The authentic Methernitha was designed and developed by purists who believed they had discovered a previously unknown electronic phenomenon, but they wanted to keep an integrity to the early pioneering days of the Pidgeon, Wimshurst and Holtz electrostatic machinery; they would not use such modern devices as transistors or IC chips (more's the pity) --- but they do use some pretty uncommon electronic engineering in their circuit [note 5].

Obviously, the electronics are in two parts; one --- the electrostatic generator and its particular technologies of how to direct what charge where, and two --- the very unique auxiliary electromagnetic circuit of inductances, capacitances and rectification that mobilises that 'static' electricity. To understand how they convert static energy into an electromotive force you would do well to go back to the earliest years of radio. From the pages of spark radio you soon appreciate just how important oscillation circuits and their valve rectifiers were, and moreover, how difficult it proved to engineer them. For although radio transmitters and receivers from the 1900's used resonating circuits their oscillations were controlled by sparks between two contacts and, of course, they were relatively inefficient. Not until the 1920's did the first electric current oscillations become an observable controlled phenomenon when someone coupled a rectifier valve, a capacitor, and a resistor together [note 6]. The early 1920's also saw the best era of experimentation and invention for novel devices that turned static energy into useable electromagnetic energy; it was in a 1921 patent that we see a German physicist Hermann Plauson describe in great detail his methods to convert static power, not only from rotary influence machines but also from balloons collecting atmospheric electricity up in the sky; and by using thermionic rectifiers, leyden jar capacitors and inductor coils he proposed a free-energy network that was to power the whole of Germany [note 7] ! The thermionic rectifier valve heralded a new era for radio and high voltage physics, and as it was then subjected to such a broad array of experiments and modifications to improve its efficiency so it paved the way for all sorts of new avenues in electronics. Indeed, with such a technical catalogue of similarities with what we see in the available photographs of Testatika it can be assumed without doubt that the horizontal glass tube which sits on top of the Methernitha machines is exactly what a home-made vacuum thermionic rectifying valve would look like; with its internal anode mesh-plate, surrounded by a coiled copper grid, fed by a glowing (heated) cathode wire running horizontally across its centre and capped by two black end-pieces, which are too big and bulbous to be mere end-caps and must surely be black rubber vacuum seals to seal the glass tube and the input/output wires [note 8].

With such a rectifier, some induction coils, and some leyden jar capacitors you have a circuit that oscillates, and that's what has to happen with a Methernitha, the electromagnetic circuit has to oscillate for it to work, and then the oscillations have to be rectified (or even modulated) so that the resulting single-pole pulses can be channeled through the big cans, which are basically high-efficiency transformers, and outputted as reduced voltage higher current DC pulses (see fig.3).

The precise components used to oscillate the primary oscillating circuit are, I believe, not to be seen in any of the available photographs, but there are various hints for their approximate whereabouts on the machine. Firstly, according to electronic design there should be a capacitor and coil configuration in close proximity to the rectifier. Well, from the picture "3KWREAR" can be seen the two long upright tubes which, according to those who have seen them first-hand, comprise a spirally turned aluminium strip (which indicates they are chokes [note 9]) inside a glass tube, inside the same sort of outer shielding that the big cans have (which indicates they are electrostatic shields), inside yet another glass tube, and are terminated at the top with a brass connecting rod which does a right-angled turn and passes into the side of the tower --- but only two-thirds up the height of the tower. These two assemblies must form a connection to the rectifier, because the rectifier is at the top of the tower, so why don't these electrostatically sensitive tubes extend all the way up to it ? Again, from the photographs of the rear and front of the Methernithas there is a wire that comes out of the tower's side wall at about 4 inches above the upright's brass terminals and this wire then passes through a short black tube and on to the rectifier valve. This, of course, would happen on both sides of the tower, enabling a connection to both ends of the rectifier. But why have this 4 inch gap of connections at the top of the tower ? Something is placed inside the top of the tower in this intermediate space which is very necessary to the circuit, and I think it must be the location of the capacitor/inductor configuration to oscillate the circuit. This (fig.4) is how I would see the inside of the top of the towers [note 10].

I've seen some of the patented inventions that rotate discs --- by using magnets (i.e., H.Rosenberg's permanent magnet excited rotational machine, US Patent # 3,411,027), and by utilising inscribed metalised discs (US Patent # 3,239,705 for instance), but there simply isn't enough room for these to be located in the Methernitha disc setup --- also, you don't want to interfere with the ES fields that zip around the revolving discs: From the reports of those who have seen the small machines working it appears their discs were rotated by small DC electric motors after they were hand-started, some re-wound with thinner wire (to presumably increase their torque) and powered directly from the discs' generated electricity --- but I have also seen how two discs can continue to rotate simply by careful placement of curved electrodes [note 11] which would act on the charges on the discs --- like the 3kw Testatica Distatica generators.

After reading through the many early accounts of electrostatic rotary machines, and some of the more recent ones, you can't help but be puzzled by the Methernitha's incredibly low rotational speed of just 60 rpm (and in the 1999 engineers report as low 15 rpm !). Most other early experimenters boasted up to 3000 rpm. J.G.Trump in his work on high voltage generation in space [note 12] spun his rotary machine at 10,000 rpm (to produce 433 Watts at 24 KV no less). One reason for this low speed might be to do with the close proximity of the 50 lamellas (gitter-grilles) on the discs at their inner ends, they are very close together, I think too close. Air, normally an insulator, breaks down and conducts at around 25-35 KV (this figure has been fairly constant from day-one of electrostatic machine experiments right through to the present day --- because air has a breakdown field strength of 3x106 volts/metre) and short-circuits the circuit. I feel that because this design of grilles is prone to short-circuiting at high voltages the Methernitha people have limited their rotational speed so as to ensure a low operating voltage --- of what I'd guesstimate to be only 12 to 24KV.

But, is this a waste of extra potential ? Not necessarily, for I don't think that the main power output comes solely from what the two counter-rotating discs supply.

There is, I believe, a far more important power generator --- the electron cascade generator, and the Methernitha has two of them, held inside the two horseshoe magnets, and providing the circuits to the magnets are made to oscillate at the right frequency at a high enough voltage then these metalised-perspex laminated blocks can enmass A MUCH LARGER AMOUNT OF ELECTRICITY THAN WHAT IS PUT INTO THEM.

This, perhaps, is the previously unknown electronic phenomenon that the Methernitha group have so zealously been trying to protect against unscrupulous entrepreneurs. But I would say that this copious supply of free energy is already known to the world --- it is not readily available - and its principles are not fully understood, as yet, but it is known about.

As the descriptions say (on the Testatika website), between the horseshoe magnet legs are four blocks of transparent 'plexiglass' type material alternated with copper and aluminium plates (that may or may not be perforated), in the sequence c-p-a-c-p-a-c-p-a-c-p-a (also see fig.6). And according to the Linden Experiment, where Paul Baumann induces a resonance of about 80-140 MHz in a coiled horseshoe and then has an aluminium-insulator-copper block moved between the horseshoe legs, a voltage could be taken off the plates of the block which measured 700 volts (DC presumably) [note 13]. This incredible phenomenon has never been replicated by any 'outside researcher', and is said to be the basis by which the Methernitha machine could be understood how to work [the clue, possibly, to this Principle Experiment may be variable-capacitance and dielectric-absorpsion].

But what, I hear you say, is an electron cascade... Well, it was only by chance, very recently, that I happened to listen to an audio tape by a Dr. Flanagan about crystal water; when I switched the tape over after the end of side one Dr. Flanagan then began talking about an electronic configuration that applied a high frequency, high voltage alternating field across an insulator --- that created what he called an electron cascade effect --- Yes, I thought, here is the answer to the Methernitha Machine.

The electron cascade or avalanche effect is where air molecules are accelerated to the device at such a high velocity that they collide with other molecules and atoms in the air to liberate new electrons which in turn also collide and liberate even more 'free electrons' from other air molecules (see fig.5), all of which become accelerated by the electric field, and an avalanche of electron-multiplications progresses throughout the whole immediate environment [note 14]. It's a chain reaction, and an entirely safe one, it happens in a more ferocious way in lightning strikes, and is a natural phenomenon. And, as in this case, the environment actually becomes part of the circuit [note 15] because the process is actually negatively-ionising the air surrounding the Methernitha machines, and that is why those who have been near these generators when working say the air around about them is cool and fresh [note 16].

In view of the fact that it's designers have chosen to wind insulated wire (which may or may not be bifilar) around the horseshoe metal it is likely that the horseshoes are used for some form of induction [note 17], it would also be very possible to draw directly from this part of the circuit the extra electric current produced from the electron cascade blocks, with suitable connections that might lead downward into the wooden base (where it is believed that an alternate layering of perforated metal plates and insulating plates - making up a large high-voltage storage capacitor - is located). This power could then be discharged as a pulsed output of high wattage, especially if the final output part of the electronic circuit is configured as a Pulse Forming Network of multiple sections of inductor / capacitor combinations [note 18].

The two big cans at the side, the big capacitors, are probably not highly technical (see fig.7), once the fundamental formula has been decided upon all models of a Testatika generator would follow a similar construction process. The written descriptions are a little contradictory but they seem to suggest a central input rod, or tube, connecting at the bottom of the cans to a stack of inter-linked pancake coils, that are wound secondary-outside primary-inside, fitted around a core of 6 hollow donut-ring magnets stacked in such a way with plastic spacers as to allow air gaps between them, and then finally the output of each can is a connection from the top coil of the secondaries of the pancake coils to a brass ring around the centre of the black plastic top lid --- and from the photographs can be seen a large diameter wire or tube [note 19] connecting that polarity's output terminal to the top lid's brass ring via a brass screw terminal. I would suggest that the ring magnets (of anistropic ferrite perhaps) are gapped in this way to prevent the magnetic flux fields of the pancake primaries co-joining as one sprawling field, because it would be more advantageous, and safer, to have each separate pancake's magnetic flux cut it's own adjoining secondary coil, and divide the secondary output voltage into smaller amounts of potential, thus depending less on complicated insulating procedures that accompany high voltage single primary / single secondary transformers.

The use of aluminium mesh and solid copper sheeting is commonly used in electronic construction; the outer aluminium mesh cylinder would be used to shield stray electrostatic charges, and the solid copper cylinder is to shield the large amount of stray electromagnetic fields produced by the transforming process from high voltage/low current to lower voltage/higher current [note 20]. Obviously they don't want field contamination taking place between the sensitive electrostatic generator and the transformers.

Within these two outside shielding-cylinders are 'grid condensers' which, according to the 1999 report by the 30 engineers, can be as many as 20 layers of perforated sheet (presumably as concentric cylinders) - which I have indicated (in fig.7 for instance) as being electrically connected

BETWEEN each separate secondary winding - in the fashion of an old discovery from the early days of wireless telegraphy and based on the 'disruptive discharge coil' devised by Nikola Tesla, that such a condenser connected in the center of a secondary coil collects the maximum amount of voltage created by that secondary. This configuration of one condenser inside another inside another etc etc, has a striking similarity to the layout of a pulse forming network [see note 18].

In the red wired can the transformer is wired to output negative, and the blue wired can's transformer is wired to output positive polarity. Special note should be made of a similar arrangement for divided primary / secondary windings devised by Van de Graaff in his 'High Voltage Electromagnetic Charged-Particle Accelerator Apparatus Having an Insulating Magnetic Core' [note 21] with respect to magnetic reluctance gaps.

Whilst it has been said that the clear perspex disc was designated the 'cloud' disc, and the (rear) dark disc the 'ground' disc I would think this relates to different types of acrylics or plastics that might become charged to different polarities, as in the triboelectric series, where frictional charging of different plastics --- and then bringing them close together, might cause donation or acceptance from one to the other; I would think from the above that cloud represents a donator (positive charge) and that ground must mean an acceptor (negative charge). Has anyone tried the combination of a teflon disc (extremely negative charge) with a glass disc (highly positive charge) ?

Or discs doped with paramagnetic particles perhaps [note 22] ?

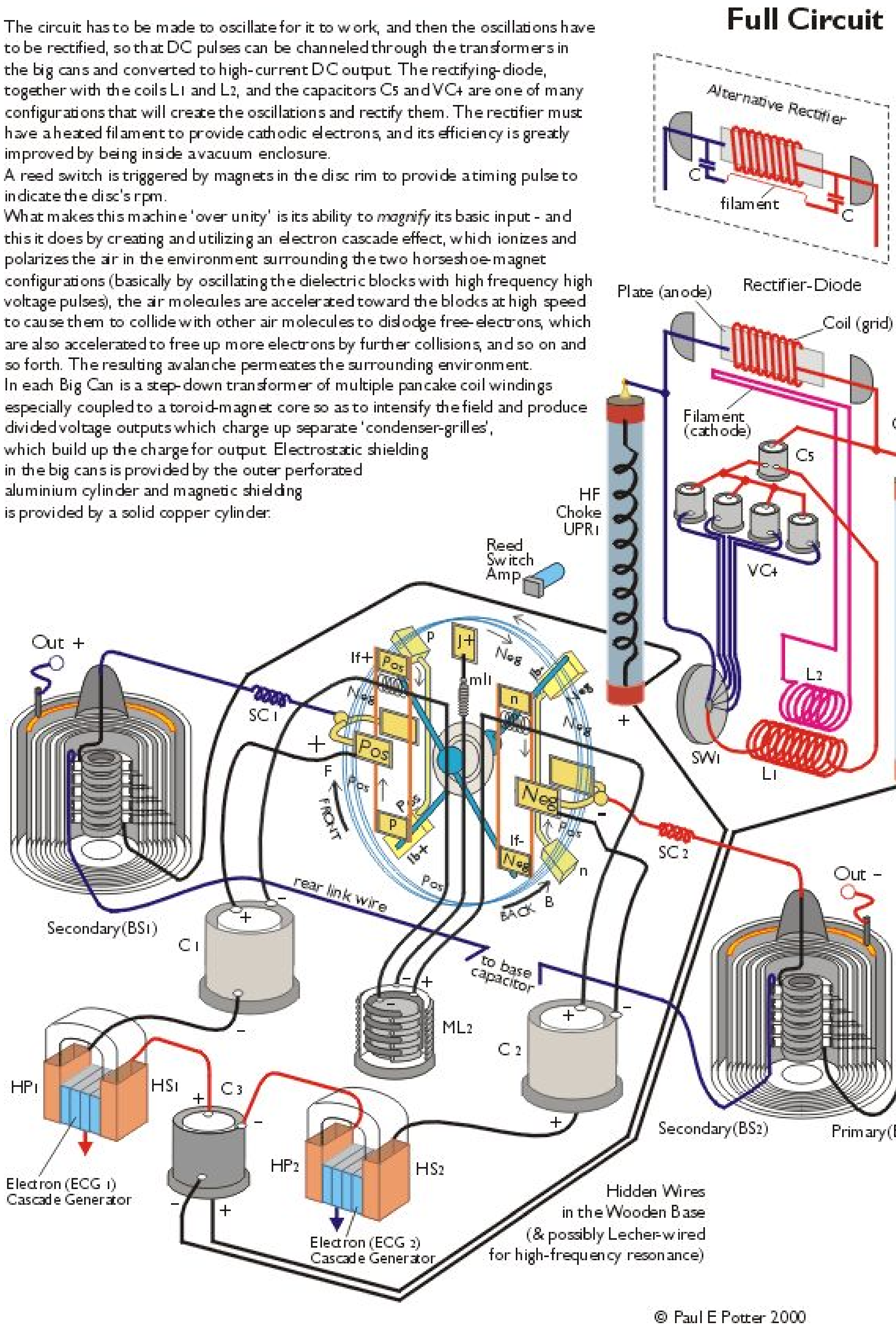

The circuit has to be made to oscillate for it to work, and then the oscillations have to be rectified, so that DC pulses can be channeled through the transformers in the big cans and converted to high-current DC output. The rectifying-diode, together with the coils L1 and L2, and the capacitors C5 and VC4 are one of many configurations that will create the oscillations and rectify them. The rectifier must have a heated filament to provide cathodic electrons, and its efficiency is greatly improved by being inside a vacuum enclosure.

A reed switch is triggered by magnets in the disc rim to provide a timing pulse to indicate the disc's rpm.

What makes this machine 'over unity' is its ability to *magnify* its basic input - and this it does by creating and utilizing an electron cascade effect, which ionizes and polarizes the air in the environment surrounding the two horseshoe-magnet configurations (basically by oscillating the dielectric blocks with high frequency high voltage pulses), the air molecules are accelerated toward the blocks at high speed to cause them to collide with other air molecules to dislodge free-electrons, which are also accelerated to free up more electrons by further collisions, and so on and so forth. The resulting avalanche permeates the surrounding environment.

In each Big Can is a step-down transformer of multiple pancake coil windings especially coupled to a toroid-magnet core so as to intensify the field and produce divided voltage outputs which charge up separate 'condenser-grilles', which build up the charge for output. Electrostatic shielding in the big cans is provided by the outer perforated aluminium cylinder and magnetic shielding is provided by a solid copper cylinder.

Full Circuit
Alternative Rectifier
C
filament
C
Plate (anode)
Rectifier-Diode
Coil (grid)
Filament (cathode)
HF Choke UPR2
C5
VC4
HF Choke UPR1
Reed Switch Amp
J+
P
Neg
If+
Pos
ml1
n
Neg
Neg
Pos
+
F
Pos
FRONT
P
Pos
Ib+
Neg
If-
Neg
Pos
n
Pos
BACK
B
+
-
Out +
SC1
SW1
L2
L1
-
SC2
Out -
Secondary (BS1)
rear link wire
+
-
C1
to base capacitor
-
+
ML2
+
-
C2
-
HP1
HS1
+
C3
-
HP2
HS2
Electron (ECG 1) Cascade Generator
-
+
Electron (ECG 2) Cascade Generator
+
Secondary (BS2)
Primary (BP2)
Hidden Wires in the Wooden Base (& possibly Lecher-wired for high-frequency resonance)

© Paul E Potter 2000

Back-Engineered Methernitha --- Notes

Note 1 --- For more information on the Pidgeon machine see "Electrical Influence Machines" by John Gray, 1903 p. 206 & "Philosophical Magazine" Dec 1898, p. 564, and of course the Pidgeon patents.

Note 2 --- See "Modern High Speed Influence Machines" by V.E. Johnson, 1921, p. 76. Johnson was not only a researcher of electrostatic machines but was also an innovative constructor of them, and as such was keen to try any technique that made his generators more powerful than even the specialised Wommelsdorf multi-disc condenser machines. This book is an absolute must for those who wish to work in this field. Another 'must' is the website of Antonio Carlos M. de Queiroz (http://www.coe.ufrj.br/~acmq/electrostatic.html) which is absolutely full of information about (and with links to) present-day developments in electrostatic machines.

Note 3 --- See "Self-Excited, Alternating, High-Voltage Generation Using A Modified Electrostatic Influence Machine" by M.Zahn, et al., American Journal of Physics, Vol 42 (1974) p. 289.

Note 4 The Methernitha designers have taken a basic Pidgeon electric field system and added a few modifications of their own, partly to lock a certain polarity of charge to a certain area so as to stabilise it, and also to boost certain areas with charge. As in their use, for example, of an extra field plate located at the top-centre in front of the front disc (just under the rectifying valve), note also that this plate, or antennae key, is indirectly coupled to the rest of the circuit, via a coil setup. Much the same occurs with the two plates slightly below it, these plates are connected to a brass terminal which connects to a copper wire that goes down and winds in a coil shape around a hollow plastic tube, and inside the tube will be another wire or small coil that draws off the electric charge. So these three plates are using not direct connection but induction to get their charge.

Note 5 ------- By looking at how each of the photographed machines have been constructed you can see that these are high quality crafted structures. I would think each would start off as sub-assemblies fitted together by pairs or small groups of members, those sub-assemblies of wooden base, big cans, perspex framework, discs with bearings and axles, when completed would be passed on to the electrical engineers of the community who would then fit the wiring connections, vacuum tube rectifier and make sure that not only did they work but that they looked like a work of art.

Note 6 --- The Fleming valve had been around since 1905 and while it progressed to the thermionic valve and audion, by 1922 the ‘Pearson and Anson Effect’ was discovered whereby oscillating currents could be produced with a resistor, capacitor and thermionic valve coupled together.

Note 7 --- See US Patent 1,540,998 (9 June 1925) Conversion of Atmospheric Electric Energy by Hermann Plauson. He also wrote a book of the subject titled "Gewinnung und Verwertung der Atmospharischen Elektrizitat" in 1920 in German (which is currently held in the British Library).

Note 8 --- Whilst some have seen the smaller 300 Watt machine’s discharger/rectifier quite open and not encased in a vacuum tube the vacuum tube models would be much more efficient and would waste less current. Also, the rectifier tube must have a heated filament (which on the 3KW machines can be seen as a glowing line running the whole length of the grid and coil assembly between the two black end caps, and in the films you can see faint flashes coming from behind the rectifier so possibly the filament is wrapped around the other side of the grid/coil assembly as well). Coolridge, back in the 1900’s, discovered that no discharge from the cathode to the anode would occur, even at 100,000 volts, unless the filament was heated (Physics Review, Vol. 2, Dec 1913, p. 418). Aluminium mesh will give off electrons quite readily and can be used as a cold cathode --- but a heated cathode offers the advantage of being able to control the oscillations.

Note 9 --- The two long upright tubes are without doubt choke coil assemblies in precisely the right place to slow down the current where it gets oscillated and rectified. In a choke the higher the flow of current the greater will be its resistance to that current flow. An even better form of choke will have some form of iron core inside it.

Note 10 --- I have come up with 6 different circuits for this oscillation section, some of which include small quartz crystals. (See notes 13 and 16 on frequency of oscillation). The black dial at the rear of the 3kw machine is most likely to select a variety of capacitances so as to control the oscillations of the circuit, which in turn control disc rotational speed.

Note 11 --- The phenomenon of electrostatic motors has been well researched over the years (see "Electrostatic Motors" O. Jefimenko in "Physics Teacher" Vol. 9, March 1971, p. 121-9, and in "Electrostatics And Its Applications" by A.D. Moore (1973) p. 131-147; "Electrostatic Motors" by B.Bollee in "Philips Tech. Review" Vol 30 (1969), p. 178-194). The Methernitha Testatika generators (see a recent report by 30 engineers) auto-rotate, after they have been started by hand, by the same principles of these ES motors.

Note 12 --- J.G.Trump worked for the US Air Force and pioneered some highly efficient electrostatic machines around the 1960's (see"Electrostatic Sources of Electric Power" in "Elec. Eng." 66:525, June 1947; and "High Voltage Generation in Space:The Parametric Electrostatic Machine" in "Progr. Astronaut. Rocketry" (vol 3 --- Energy Conversion for Space Power) 1961 p745).

Note 13 --- Although the 'Linden Experiment' was thought to register a frequency of 80-140 MHz this does not necessarily mean that the Methernitha generators would oscillate at that rate also. Such a frequency seems unnecessarily high.

Note 14 --- See "Plasma --- The Fourth State of Matter" by D.A. Frank-Kamenetskii (1972) pp10, and Dr.Patrick Flanagan's US patents 4,743,275 (May 10, 1988) and 4,391,773 (Jul 5, 1983).

Note 15 --- The effect is very similar to the converging forces in a non-uniform field, the oscillating perspex blocks become one 'electrode' and the surrounding air in the room becomes the opposite 'electrode', and by the processes of electrophoresis and dielectrophoresis the electrically charged particles in the air (the electrons and negative ions) are drawn toward the central electrode, which in this case is the perspex block assembly (see "Nonuniform Electric Fields" by Herbert A. Pohl in "Scientific American" (Dec 1960) p. 107-8). I am much more inclined to believe that the ingenuity of the design of these types of machines comes from physicists and not electronic engineers.

Note 16 --- Dr. Flanagan actually uses the electron field generator in his own special ionizer (see Method of Purifying Air and Negative Field Generator US Patent 4,391,773).

How does an electron cascade generator work, I would think that while you have an alternating electron movement (and Dr. Flanagan reckons this effect occurs with a high voltage field alternating at above 20 KHz) at the metal electrodes, the perspex blocks sandwiched between them would transfer the electricity not through their mass but around it, as surface charge --- actually in the layer of air right next to the insulator's surface. Its the same principle as dielectric absorption - the perspex blocks don't discharge themselves fast enough to keep up with the alternating voltage and so they accumulate more and more charge --- until it forms as a layer of charge on the insulator's surface. This means that at a high enough frequency the surface-air molecules polarize, with the more mobile electrons separating from the slower bulks of those molecules and while the electrons get thrusted back and forth a secondary layer of (slower) positive air ions develops, and so on, and the process of high voltage high frequency polarization triggers the electron avalanche effect.

In the event that the perspex blocks are indeed ELECTRETS (as free-energy researcher Geoff Egel and others suggest) I would think that they would work in a similar fashion to the above process, of dielectric absorption that charges up the blocks before they produce the cascade-effect. Because in the electret the electrons charged into the perspex/plastic, and the positive ions, would still be manipulated by the reversing electromagnetic field in such a way as to orientate (as with dipoles) back and forth, to eventually reach the point (if the whole circuit is tuned properly) where they would attain resonance with the immediate air surrounding them. And if this effect is similar to an inductance then possibly a back-emf will result also, to increase the output voltage. Either way I believe the effect will still be an electron cascade through the environment and the product of this oscillating output (at the blocks) could similarly be drawn off and accumulated in the multi-layered base capacitor network.

I would suggest that a test program to find the best type of blocks would be; One --- try different types of plastic/acrylic/ceramic materials for the blocks.
Two --- try different methods of electrifying the plastics (as with electrets). Three --- try plastics doped with semiconducting particles. Four - try plastics doped with paramagnetic particles. Five - try hollow plastic blocks containing an electrolytic fluid. More information on plastics will be found on the Electret vs Dielectric Absorption Page.

Note 17 --- There are several definitions of Bifilar, one where the wires cancel out their magnetic fields, and one where the wires are wound to ensure a tight low-loss magnetic flux coupling, in this case you need all the magnetic flux you can get so it must be the latter --- See "Transformers For Electronic Circuits" by Nathan R. Grossner (1967), p. 224 etc.

Most commonly used magnetic metal is Mumetal, which is an easily saturable magnetic material, routing magnetic flux through it rather than in the surrounding air, so as to enhance the mutual induction between the two coilings of red wire around the horseshoe legs.

Note 18 --- So that the machine's output voltage doesn't drain away when it is connected up to a large load, what is needed here is a Pulse Forming Network (or artificial delay line). "Such a network is an improvement on simple capacitor storage because of the cascading action from one capacitor to the next along the chain. At the beginning, all capacitors are charged to the same voltage but as soon as the first one starts to loose voltage, the one behind it is then free to discharge into it. This topping-up action, which trickles down the network from capacitor to capacitor, is the mechanism by which the voltage across the output terminals tends to hold onto its original level." (see "High Energy Discharge Systems" A.P.Stephenson, Electronics Today International, March 1992, pp. 24-26).

Note 19 --- When voltage of a high potential and high frequency flows along a wire it does so on the outer surface (called the ‘skin effect’) and so the Methernitha would use thick wiring or even 1/8" tubing to connect its circuit.

Note 20 --- Two references for shielding are: "A Shielded Loop" by S. Goldman in "Electronics" Vol 11 (1938), pp. 20-22; and "Measurements in Radio Engineering" by F.E. Terman (1935), p. 218 & p. 341.

Note 21 --- For information on maximum voltage in the center of a secondary coil see "A Handbook of Wireless Telegraphy" by J. Erskine-Murray (1913) p. 42; and an article called "Dielectric Hysteresis at Radio Frequencies" by E.F.W. Alexanderson in "Proc. I.R.E., Vol. 2 (June 1914) p137-157. For Van de Graaff's transformer see US patents 3,323,069 (May 30, 1967) and 3,187,208 (June 1, 1965). These patents were not just for a Van de Graaff high voltage generator, they were for a special system devised by Van de Graaff long after his generator had been in use to convert static electricity into current electricity. This system may be a little too complicated for the Methernitha but, nevertheless, the principles he used for multiple primary / secondary windings may be of some interest.

Note 22 --- Dr. Flanagan modified his insulator blocks, made of resin, by doping them with paramagnetic granules (such as silicon carbide) to enhance even more the electron cascade effect; which is an idea that the physicist Thomas Townsend Brown first experimented with (by using lead oxide granules) in his US Patent 3,187,206 (June 1, 1965) to good effect. The surrounding air could also be ‘enhanced’ in similar fashion to polarise it's electric charge and improve its side of the performance (for those interested in the ‘physics’ of this see an article by W.A.Douglas Rudge "On Some Sources of Disturbance of the Normal Atmospheric Potential Gradient" in Proc. Royal Soc. A, Vol. 90 (1914) pp. 571, etc).

Note 23 --- Some other generators with similarities to the Testatika machine are the "Electrostatic Energy Field Power Generating System" invented by William W. Hyde (US Patent 4,897,592 of Jan 30, 1990) is a rotor/stator variable capacitance machine capable of producing 300 KV. Other such generators are; "Parametric Electric Machine" invented by Ferdinand Cap (US Patent 4,622,510 of Nov 11, 1986) which has a series resonant (LCR) circuit structured into it so that it oscillates --- and indeed operates AT RESONANCE to ensure a high output; "Electrostatic Generator" invented by Dan B. Le May, et al. (US Patent 3,094,653 of Jun 18 1963) is a very ingenious system of variable capacitance; the "Electrostatic Machine" by Noel Felici (US Patent 2,522,106 of Sep 12, 1950) is a good standard which utilizes a valve rectifier; and the "Electrostatic Generator" by William S. Spencer (US Patent 1,415,779 of May 9, 1922) is an early rotor/stator generator which transferred its electric impulses through a transformer to produce a higher current output.

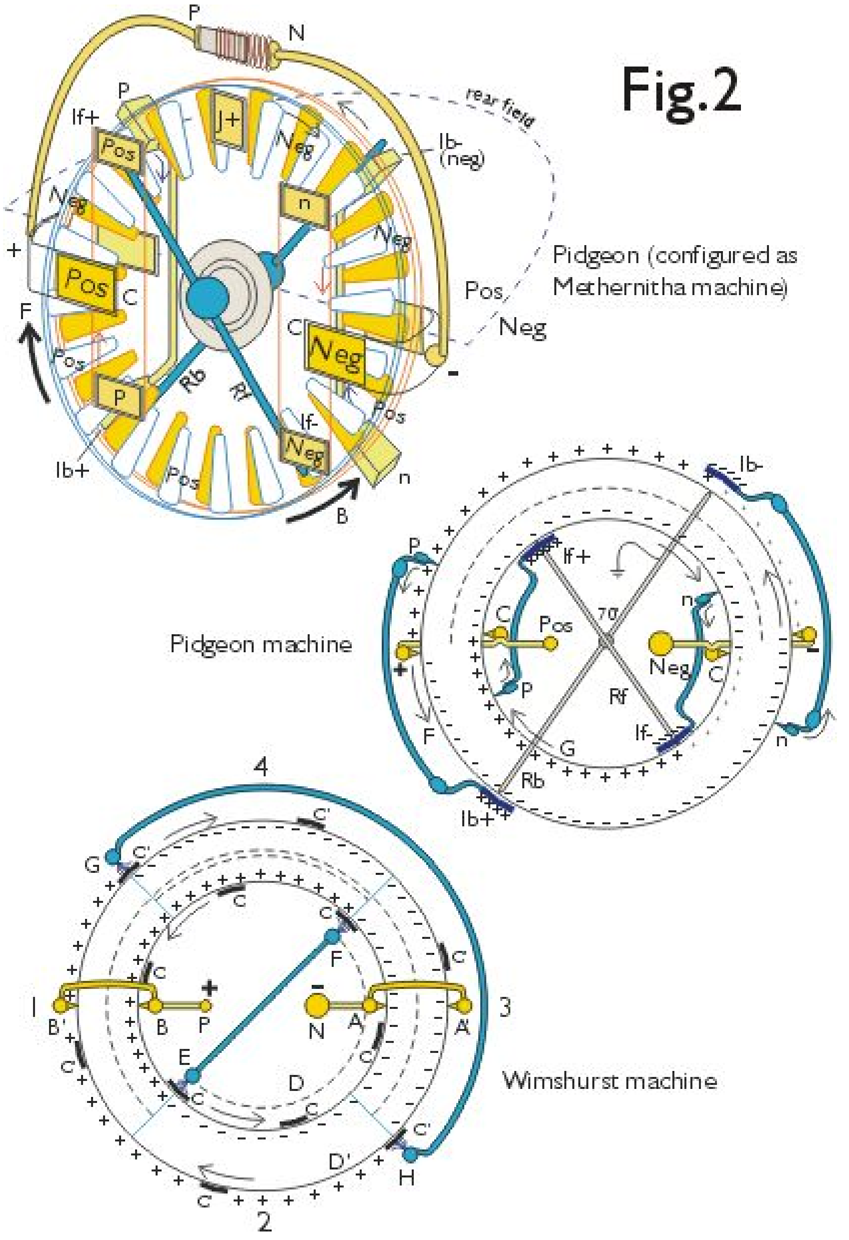

Fig.2

P
N
P
If+
J+
Neg
rear field
Ib-
(neg)
Pos
n
Neg
Neg
+
Pos
C
Pidgeon (configured as
Methernitha machine)
Pos
F
C
Neg
Neg
-
Pos
P
Rb
Rf
If-
Pos
Neg
Ib+
Pos
n
B
Ib-
P
If+
Pidgeon machine
C
Pos
70
n
Neg
C
P
Rf
If-
F
G
Rb
n
Ib+
4
C'
G
C'
C
C
C
I
B'
B
P
N
A
A
3
F
E
C'
C
D
C
C'
D'
H
C
2
Wimshurst machine

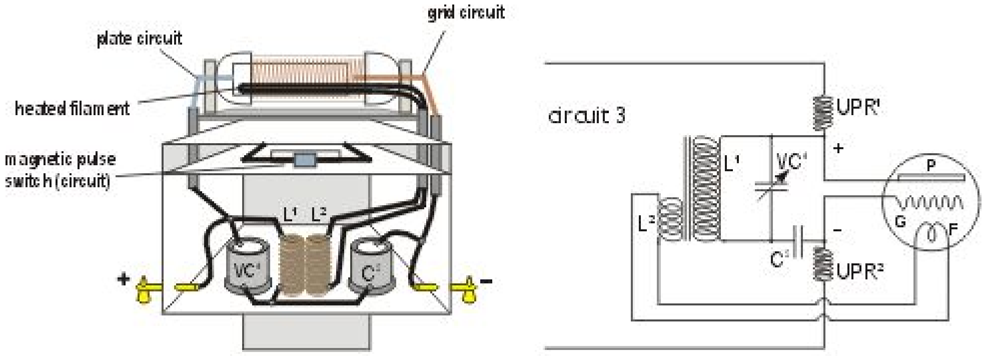
grid circuit
plate circuit
heated filament
magnetic pulse switch (circuit)
L^1 L^2
VC^1
C^1
+
-
circuit 3
UPR^1
+
L^1
VC^1
L^2
P
G
F
C^1
-
UPR^2

Fig. 4 - Tower Circuits

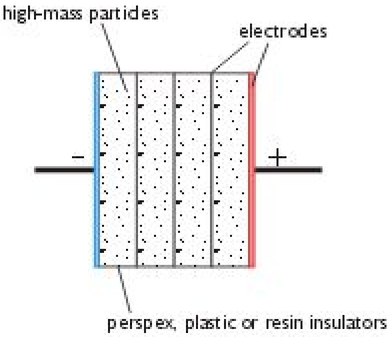
high-mass particles
electrodes
-
+
perspex, plastic or resin insulators

Fig. 6 - ECG Doped

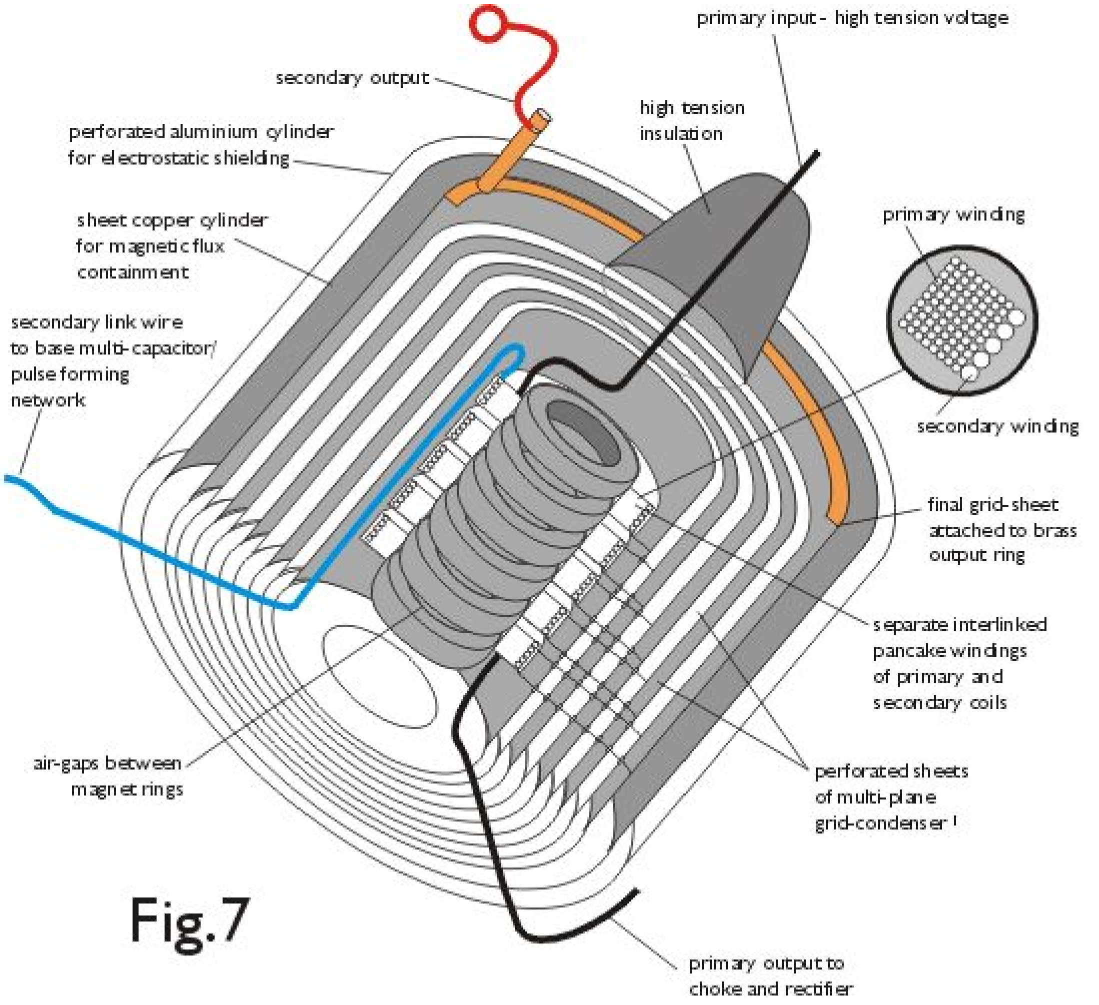
primary input - high tension voltage
secondary output
high tension insulation
perforated aluminium cylinder for electrostatic shielding
sheet copper cylinder for magnetic flux containment
primary winding
secondary link wire to base multi-capacitor/ pulse forming network
secondary winding
final grid-sheet attached to brass output ring
separate interlinked pancake windings of primary and secondary coils
air-gaps between magnet rings
perforated sheets of multi-plane grid-condenser ¹
primary output to choke and rectifier

Fig.7

[1] In the Testatika-Distatica Video it has... "By means of **grid-condensers** the energy is stored and it is then uniformly discharged, at the same time reducing the high-voltage and building up power with additional devices."

"Principle Experiment"

by

Paul E. Potter

With respect to the 30 engineer's report (of 1999) the 'principle experiment' was set up in such a way (see diag 1) that by swivelling the cross-piece over two other plexiglas blocks a current was transferred into two capacitors. Possibly when someone grabs hold of the swivelling cross-piece they transfer some static electricity from their body to the grilles of the cross-piece which, when vigorously turned back and forth will initiate a transfer of electric charge to the two stacks of blocks below them attached to the base. This may either be because anyone walking over a carpet or wearing man-made fibres will already hold thousands of volts of electrostatic charge - which will be enough to 'prime' this apparatus after touching it, to get it to work, much the same as when a rotating influence machine has to be primed with an electric charge before it will work; or, as many researchers have surmised already, that the plexiglas blocks glued to the swivelling cross-piece had been charged up with electricity as an ELECTRET.

The use of mesh around the cross-piece guarantees a variable capacitance to the moving arms --- and as these are moved quickly over the base blocks (again, alternated with mesh and Plexiglas) an amount of electric charge will be transferred to them EACH TIME the arms go backward-and-forward over them. Electronically, you have one variable capacitance acting upon another variable capacitance - which is why Baumann said that if metal foil were used instead of the mesh it would not produce the same effect.

What happens next is called "dielectric absorption", and the following extract, from "Dielectrics" by P.J.Harrop (1972) pp71, will explain;

"When a given dielectric [i.e., plexiglas] has a field put across it, and this field is then removed, the electrodes are briefly shorted out (not enough for all the ions and particles to relapse back) and the device left on open circuit one finds subsequently that it has partially charged up again.

This is due to the slow-moving charge carriers remaining in position. It can be a frightening phenomenon with practical dielectrics since a person who has briefly discharged a charged insulator may subsequently touch it and be severely shocked." Some dielectric materials being more prone to this effect than others, and so obviously some experimentation will be needed to choose the best perspex, plexiglas, plastic, or acrylic materials so as to take the most advantage from this dielectric absorption effect.

Once the blocks are being charged up the electric charge simply flows into the two interconnected capacitors and accumulates.

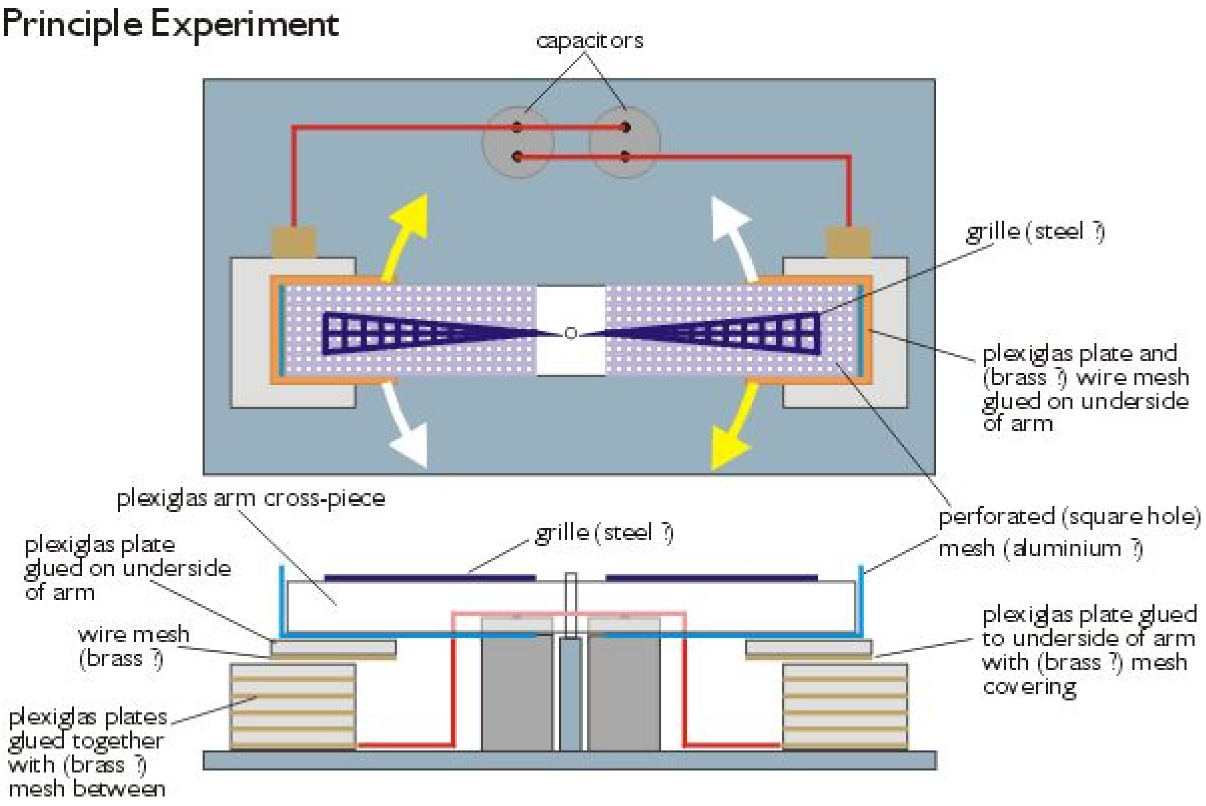
Principle Experiment
capacitors
grille (steel ?)
plexiglas plate and (brass ?) wire mesh glued on underside of arm
perforated (square hole) mesh (aluminium ?)
plexiglas arm cross-piece
plexiglas plate glued on underside of arm
grille (steel ?)
wire mesh (brass ?)
plexiglas plate glued to underside of arm with (brass ?) mesh covering
plexiglas plates glued together with (brass ?) mesh between

Linden Experiment ~

A look at the only available diagram of the Linden experiment (see diag 2) shows a horseshoe magnet coupled to what seems to be a closed-ended wire. It can also be seen in this diagram that the wire coils around the magnet are sufficiently spaced apart so as to provide capacitance (between each coil of the wire) in the circuit. So, with capacitance, coils (for induction), and magnetic flux, you have all the ingredients needed to make an oscillating circuit --- provided there is a supply of voltage and some sort of interruption mechanism - then you have a circuit that will resonate. And this leads us back to that oft repeated question --- is the block used in the Linden Experiment an electret. The first person to PUBLISH how they have duplicated this experiment will be the one to answer that question...

But there is something wrong with this diagram --- the resonant frequency of this simple circuit --- depending only on the values of inductance and capacitance included in the circuit would consequently enable it to oscillate at only a few kilohertz at best. No circuit of capacitance/inductance can be made to resonate at such a high frequency as claimed, not even if it included any form of quartz crystal. If, as the diagram says, it resonates at 140 MHz then there must be a Lecher wire system (the running of two conducting wires in a parallel line, about a foot apart or less, for several feet) to make it resonate at such a high frequency. It would seem to me also that a required modification of the Linden Experiment diagram would be to put a different metal into this circuit where the wire is said to be "closed-ended", say for instance a zinc plate, then there will be a contact voltage established between the copper wire and the zinc plate. Such a voltage would start the circuit resonating, for the reasons stated above.

And as with the Tini setup if a block comprising two metal plates separated by a suitable dielectric material (such as plexiglas) is placed in an oscillating electric field the electric charge on the metal plates will permeate into the dielectric, and for the same reasons as detailed above in the Principle Experiment, will temporarily accumulate voltage because of dielectric absorption. As I have already said elsewhere there is a great similarity between the phenomenon of the Testatika horseshoe blocks (and the Tini setup) and the principle behind Patrick Flanagan's Electron Field Generator which energizes an electron cascade mechanism throughout the immediate environment.

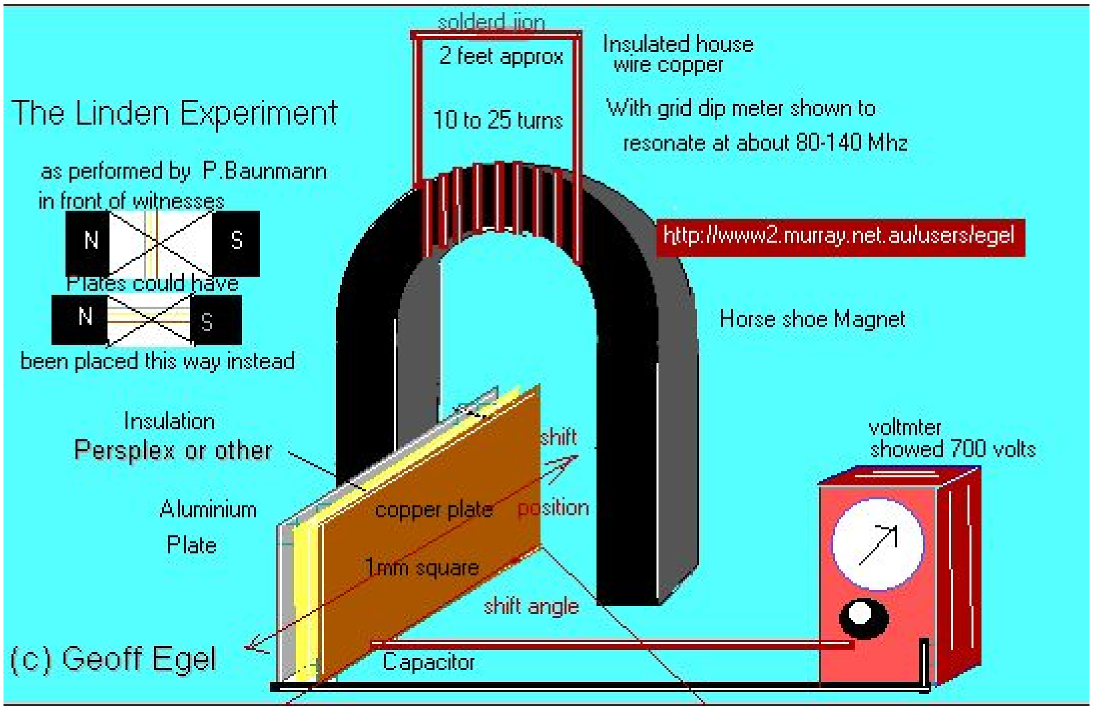
solderd jion
2 feet approx
10 to 25 turns
Insulated house
wire copper
With grid dip meter shown to
resonate at about 80-140 Mhz
The Linden Experiment
as performed by P.Baunmann
in front of witnesses
N
S
Plates could have
N
S
been placed this way instead
http://www2.murray.net.au/users/egel
Horse shoe Magnet
Insulation
Persplex or other
shift
position
voltmter
showed 700 volts
Aluminium
Plate
copper plate
1mm square
shift angle
(c) Geoff Egel
Capacitor

Geoff Egel (1997) --- http://www2.murray.net.au/users/egel ~

Rectifier Circuit ~

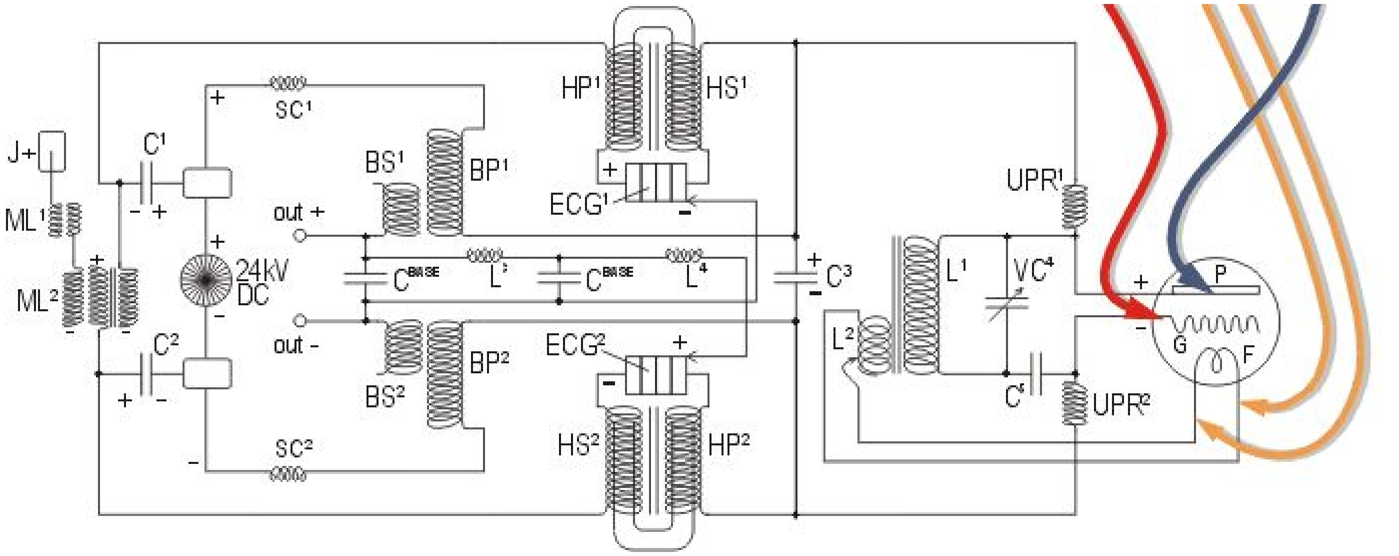
J+
ML¹
ML²
C¹
C²
SC¹
SC²
24kV DC
out +
out -
BS¹
BS²
BP¹
BP²
C BASE
L³
C BASE
L⁴
HP¹
HS¹
HS²
HP²
ECG¹
ECG²
C³
L¹
L²
VC⁴
C⁵
UPR¹
UPR²
P
G
F

The rectifier forms the basis for Testatika's operation For whilst the LC circuits L^1, VC^4, UPR^1, UPR^2 and C^5 set up oscillations from the generator's high voltage output the rectifier controls the oscillations to provide positive dc pulses (which are then directed through a step-down transformer BP^1/BS^1 and BP^2/BS^2), to supply power at a higher current rate and lower voltage.

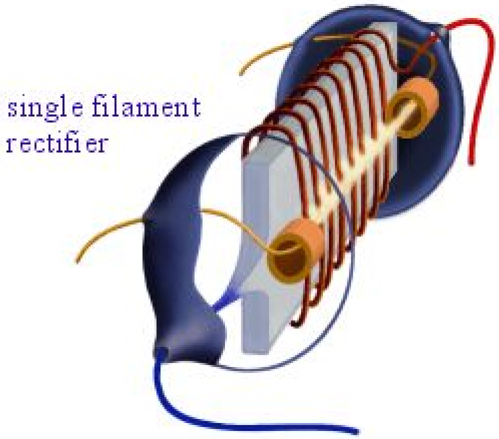
single filament
rectifier

Unidentified European Report ~

"Believe or Not, Here It Is!"

The first really running "free energy" machine coming from Switzerland, Europe. It was developed over a 20 years Research period by a religious group called: METHERNITHA. This group lives in CH-3517 Linden, Switzerland. The inventor of this superb machine, Mr.Paul Baumann claims, its running principle was found by studying the lightning effects from nature.

And here are the facts :

* Testatika is an influence-type Wimhurst machine which runs on it's own energy, once started by counterclock revolving it's 2 dics by hand !!!

* Testatika not only runs on it' own energy, but produces also a huge amount of excess power, at least 3 KWATTS of power! That is enough to supply a one-person apartment with one machine!

* The machine you see during this demo-animation is only about 70 cm wide, 40 cm deep and about 60 cm tall !

* It delivers DC-Voltage ranging from 270 til 320 Volts, only depending on the dryness of the air! At this voltage it can supply at least 10 Amperes of DC-current !

* Testatika is not a perpetuum mobile, but an energy machine that collects it's huge amount of "FREE" energy from the charged and ionized air particles. But there are still some technological tricks implemented to overcome the normal drag-resistance of a conventional Wimhurst machine, which is still the secret of the Methernitha group !!!

When will this technology be available for everyone?

The Testatika machines, (there already exists also some smaller units which only deliver about 200-300 Watts) are not yet mass-production type models! They are still laboratory prototype units, although they are build with a very good craftmansship !

For more info contact directly:

METHERNITHA, CH-3517 LINDEN, Switzerland
TEL.-Nr.: ++ 41 31 97 11 24

The pictures are from some photos and from a video-tape Methernitha sells. It shows the machines (also the smaler ones) and explains also Methernitha's spiritual aims. The tape has a running length of about 40 minutes and is really worth looking at it! It is also available in
VHS-NTSC and American language. It also shows Methernitha's earlier research developments like huge wind generators.

With the Testatika there is no fraud, hidden batteries or any other hidden energy source. It is really running! They already have the technology of ten years ahead! No more oil, no more pollution, no more atomic waste products, no more hunger in this world! This is the machine the mankind has dreamed about for centuries. Now it has come true!

More information about the Methernitha machine is also available in the German magazine *RAUM & ZEIT*, issue Nr.40 (8).

Jean L. Naudin Labs (4 February 1998) ~

"Testatika Generator and Over-Unity"

by

Cyril Smith

Having read Nelson Camus' article purporting to "explain" the Testatika over-unity generator (http://members.aol.com/overunity2/nelson/testatic.htm), I am of the opinion that the real explanation is hidden amongst the 19th century electrical mumbo jumbo which makes up much of the machine (Leyden Jars, Horseshoe Magnets, Wimshurst Generator and so on). From the description of spiral-wound foil capacitors containing radioactive material and placed within current carrying coils it strikes me that perhaps the real energy comes from Beta particles (fast electrons) absorbed into the electrical circuit.

To extract energy from Beta particles it is not just enough to capture them. Clearly this will result in a current flow according to the capture rate, but current by itself is not power. The kinetic energy of the particle must be captured and this will result in potential difference or voltage. If a Beta particle enters a thin foil conductor normal to its surface and is captured, then the potential difference associated with giving up its kinetic energy can be expected to occur across the opposite faces of the foil. If however the Beta particle were turned so as to enter at a shallow grazing angle to the surface, not only would the probability of capture be increased (since the particle will travel further within the material) but the potential difference will occur along the length of the foil. Beta particles travelling even at relavistic velocities can be turned within short distances by relatively weak magnetic fields, so the concept shown in the following Figure suggests itself.

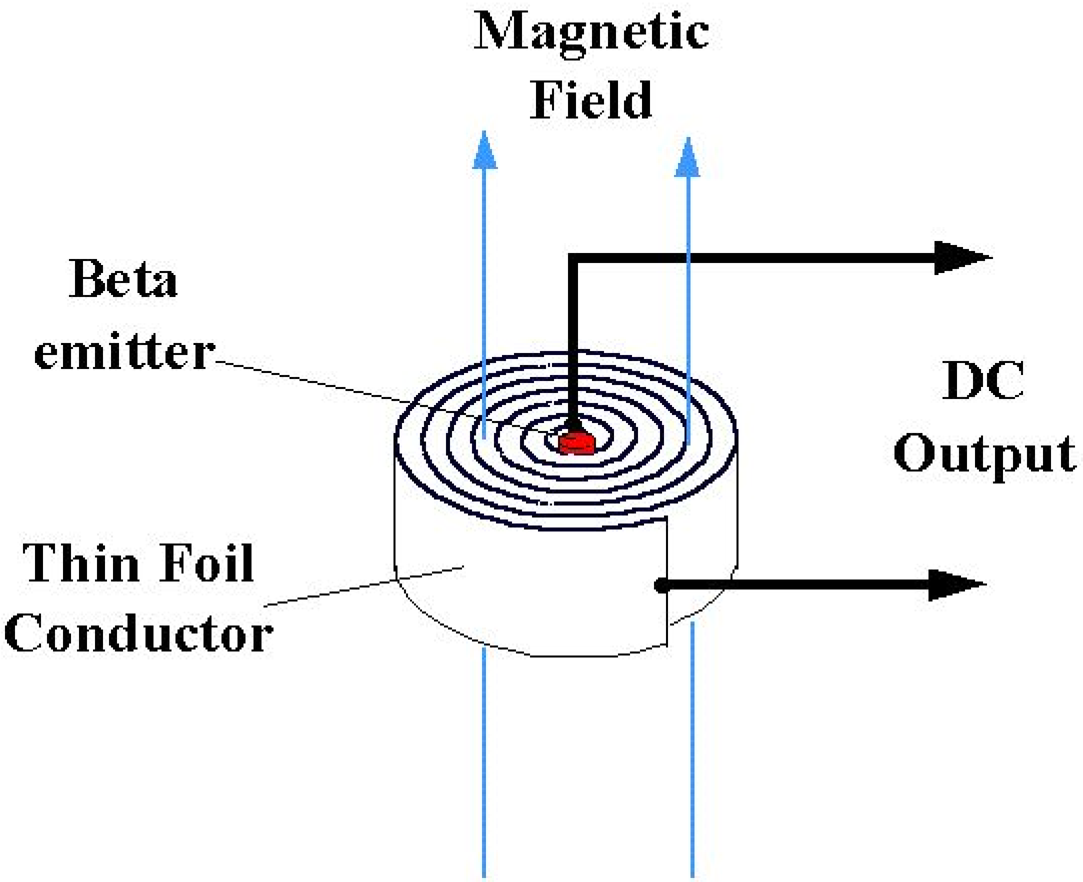

Methernitha Statement of Purpose ~

"The Research Work of Methernitha in the Field of the so-called Free Energy"

Ever since the foundation of Methernitha there existed a department for research, development and electronics, which was concerned with the problem of alternative energy sources, namely with technologies, which were suited to exploit the inherent forces of nature and thus to unlock sources of energy without disturbing natures ecological balance in any negative way. Any technology man invents should serve him in short as well as in long terms, and this condition is not fulfilled as soon as it opposes nature in any way.

This team of research within Methernitha works completely autonomous and is financed out of the co-operatives own resources, without any outside support.

The efficient utilisation of wind energy was one of the first objectives of Methernitha's research program. At the beginning generators with special excitation were developed, which allowed to load the cells of accumulators even at low RPM at times when there are only moderate air movements.

Utilising the kinetic energy of water currents was another field of interest of the development team, but it was pursued more as hobby. The key problem was here to transfer the slow revolutions of the waterwheel to an extent that the excitation threshold of the generator could be surpassed by a minimal loss of energy.

Also solar cells and solar heat collectors attracted the attention of our researchers since a long time. But since in these fields other institutions have attained outstanding results, Methernitha started, and this was already more than 30 years ago, to concentrate its efforts on lesser known and even generally unknown sources of energy. The result of this scientific work is the Thestatica machine.

The question arises: How it comes, that Methernitha, nothing more than a private organisation, could invest so much time, engagement, perseverance and financial resources in this kind of research.

Research and development are integrated parts of the general idealistic concept of Methernitha. To make you understand these ideals and goals, we will now introduce you into the practical sides of this spiritual community Methernitha, as it functions as a co-operative of people, living and working together.

Linden is a calm village of farmers just as dozens of others in the region of the valley called Emmental. However Linden is also the home of a special form of human life, which is capable of attracting more and more people from all over the globe, and which has repeatedly caused astonishment through its excellent functioning, even in extremely difficult situations.

There is no doubt that the formation of Methernitha is also a good part of the life story of Paul Baumann. Through his extraordinary technical capabilities, but also through his astonishing practicable wisdom he opened relations to all kind of people with ideals and in many places, and pretty soon, the idea to form a co-operative enterprise took shape.

"In the name of the God Almighty" people of equal thinking gathered and founded a co-operative, which could serve as the economic basis of this spiritual community.

"We want to be a united group of brothers, and never separate however severe the burden may be", this was the solemn vow.

Renunciation of alcohol and smoking and the will to realise a harmonious community life without dispute and discord like in the original Christian communities were the prerequisites to become a member. For the members Methernitha is an ideal opportunity to lead a fulfilled life by practising charity.

A workshop was erected and soon, one house after the other grew on these premises. Also all construction was done out of own resources and solely with the savings from the common work.

Today Methernitha is a social model proving the quality of its ideals through its well functioning just more than forty years. The people working here are almost without exception members of Methernitha.

Not even the prospering development that took place over the following decades could impair or even suppress the global idealistic aims. On the contrary, everybody works in his own interest with diligence and great joy towards the erection and the support of their new homeland and according to the principle:

One for all and all for one!

With this word one can conquer anything. Also the ancient truths appear again and attain respect.

The people living here feel themselves as members of a family, like a group sitting in the same boat and proud, but also grateful proprietors of their own beautiful homeland, which they may shape exactly as they wish to. Evidently, this form of human social life can function only on the base of idealistic principles.

The question arises, how the realisation of a sincere religious philosophy of life may be brought in harmony with a successful economic management. It is not obvious at all, that this is attainable.

Nowadays there are many amongst us that are caught in the world-wide and dense network of social and economic dependencies and obligations and many also feel the confrontation with its problems.

One member of the economic directory has recently phrased this in the following way:

"The fact that all essential functions of Methernitha are fulfilled without any external force, driven solely through inner conviction, which causes everybody to help and take care of the other, this is for me the most astonishing effect which is produced by this form of living together. It seems to be a miracle."

Another miracle within Methernitha is the Thestatica, which is the result of more than 20 years of research.

This wonder machine is lurked from nature, nothing else. Nature is the greatest source of power as well as knowledge which man has, and it still conceals many secrets, which are only revealed to those, who approach and tie in with them with highest respect and responsibility.

To understand nature and to perceive its voice, man is obliged to experience silence and solitude, and it was there, where the knowledge about this technology was obtained.

For these reasons it was always a great concern of Methernitha to acquire properties, untouched by man as far as possible, be it in valleys or forests, in the mountains or ashore of lakes, where one could study nature, ones own being and the creator of all this universe in silence and concentration and without being disturbed.

The public never understood this properly, rather interpreted it wrongly as an act of seclusion, assuming we had to hide something unclean. We actually had --- and still have --- to take great troubles to realise undisturbed all the things we intended to accomplish.

Such undertaking in research and development necessitate considerable financial expenditures. Therefore we are quite often obliged to construct things with most primitive means and materials. What was thrown away by our affluent society, we collect and possibly set up with it a cornerstone in the discovery of new forces and truth.

We are fortunate in gaining the experience that paradoxically the most beautiful and useful results can be achieved by just using the most simple means. Never did we use any borrowed capital because we want to stay free Swiss citizens and do not want to be hindered or even bound in any way in the pursuance of our aims.

The two conterrotating discs generate an electrostatic charge. One disc represents the earth, the other the cloud. Using grid electrodes the charges are bound. After that they are collected by non-contacting so-called antenna keys and then sorted.

After being initially turned on by hand, the discs rotate by themselves according to the electrostatic laws about attraction and repulsion. A rectifying diode keeps the cycles in steady state. Otherwise the impulses of attraction and repulsion would accumulate and cause the discs to run faster and faster. The correct speed is of great importance and for optimal power generation the discs have to run quite steady and slow.

By means of grid condensers the energy is stored and then uniformly discharged, at the same time reducing the high voltage and building up power with additional devices. Finally the machine supplies a uniform direct current, which varies according to the size of the model. The machine furnishes about 3-4kW permanent output, depending on humidity, whereby the electric potential ranges from 270 to 320Volt. High humidity of the atmosphere prevents the build-up of electric potential. The drier the air, the better.

No doubt, through the so far achieved results one main objective has been reached, namely to prove that it is possible to use Free Energy. Nevertheless the research work is not yet completed.

To the educated physicist many a thing of this machine may seem impossible, maybe even crazy. Maybe he is also offended by the conceptions used to explain the whole. Only partly we could use the concepts of conventional physical terminology to explain and define only approximately the functions and properties of the various parts of the machine.

After all it will be necessary to create some more new concepts like the one we have already used before, when we termed the non-contacting collectors of electric charges as antenna keys.

This machine puts experts, which are just trained in conventional physics to a very hard test, because its mode of action is not explainable with the state of the art of officially accepted physical knowledge, or at the most only partially explainable. However also a trained specialist should remain free and independent in his thinking, and should avoid to be limited by the temporal framework of publicly admitted knowledge in any science.

It has to be remembered that the established science was already many times forced to change or give up some of its very fundamental concepts. Think about Galilee, to name only one example. Our human society almost condemned this man as a sorcerer and magician, just because he investigated and discovered a truth that seemed unacceptable by the established science of the days.

The book knowledge of any times is not wrong, but it is incomplete, and therefore allows to draw wrong conclusions.

We are part of a new area which brings to light many new facts and new knowledge. The clothes of nowadays science have become too tight and should be stripped off, just as the larva of an insect does with its skin. Only this will allow a true metamorphosis to take its course, and finally, at the limits and at the destination of all worldly knowledge, an universal and unlimited spiritual science, radiating and beautiful, just as the completed imago of an insect, may give its blessings and benediction to a renovated humanity.

To ever reach there, a universal development of man has to take place. This however is only possible when man becomes aware of his true role within the whole creation, and again learns to recognise his true tasks, because the whole universe functions within a strict and precisely structured order according to the will and word of the creator. Therefore also man should recognise and realise these universal laws that are valid within the whole and within every part of this creation.

The hard facts rather show how far man has left the divine order through his self willed and authoritarian way of action and that he has become the actual cause of all discord and evil on this planet.

Unfortunately the ruling bodies that should be responsible for the well-being of the people work too often with the target to make life more and more difficult and to render impossible every free spiritual development.

Instead of utilising the achievements of science and technology for the benefit and preservation of all form of life, they are abused carelessly and irresponsibly to destroy and to kill, and thus turn them into a curse upon mankind.

To change all this, the evolution of a new technology is not enough, even if it were the most ecological and ingenious. To change this present status one has to go much deeper down, to the root-cause of all this evil, and this is mans way of thinking, his state of mind.

The ancient divine commandments are still valid today and also show today the way and direction humanity should go, as clearly spoken by the prophet Mica:

God has told you what is good. And what is it that the Lord asks of you? Only to act justly, to love loyalty, to walk wisely before your God.

"Testatika-Distatica Machine"

(Transcribed from the Methernitha Testatika video by Paul E Potter)

"... The efficient utilization of wind energy was one of the first objectives of Methernitha's research program.

At the beginning generators with special excitation were developed which allowed to load the cells of accumulators even at low rpm, at times when there are only moderate movements.

Utilizing the kinetic energy of water currents was another field of interest of the development team, but it was pursued more as hobby. The key problem was here to transfer the slow revolutions of the water-wheel to an extent that the excitation threshold of the generator could be surpassed by a minimal loss of energy. Also solar cells and solar heat collectors attracted the attention of our researchers since a long time. But since in these fields other institutions have obtained outstanding results Methernitha started, and this was already more than 20 years ago [since 1960], to concentrate its efforts on lesser-known and even generally unknown sources of energy, the result of this scientific work is the Testatika machine of which most of you may have heard already.

The two counter-rotating discs generate an electrostatic charge. One disc represents the earth, the other the cloud. Using grid-electrodes* the charges are bound. After that they are collected by non-contacted, so-called, antennae keys** and then sorted [shorted ?].

After being initially turned on by hand the discs rotate by themselves according to the electrostatic laws about attraction and repulsion.

A rectifying-diode keeps the cycles in steady state, otherwise the impulses of attraction and repulsion would accumulate and cause the disc to run faster and faster. The correct speed is of great importance, and for optimum power generation the discs have to run quite steady and slow.

By means of grid-condensers* the energy is stored and it is then uniformly discharged, at the same time reducing the high-voltage and building up power with additional devices.

Finally, the machine supplies a uniform direct current which varies according to size of the model.

The machine furnishes about 3 to 4 Kilo-Watt permanent output, depending on humidity, whereby the electrical potential ranges from 270 to 320 volt. High humidity of the atmosphere prevents the build up of electric potential. The dryer the air the better.

No doubt, through the so-far achieved results, one main objective has been reached, namely, to prove that it is possible to use free energy. Nevertheless, the research work is not yet completed: To finish a model which can be handed out more or less to anybody and without any haphazards, also to non-specialists, much work and also time will still be needed.

Only partly we could use the concepts of conventional physical terminology in order to explain and define only approximately the functions and properties of the various parts of the machine. After all, it will be necessary to create some more new concepts, like the one we have already used before, when we termed the non-contacting collectors of electric charge as antennae keys**. This machine puts experts who are just trained in conventional physics to a very hard test..."

* The German original has gitter-condensers, with gitter defined as wire lattice, grille, grate.

** the original word used is tasten or taster, which means antenna, probe, key.

Your Support Maintains this Service --

BUY

The Rex Research Civilization Kit

... It's Your Best Bet & Investment in Sustainable Humanity on Earth ...
Ensure & Enhance Your Survival & Genome Transmission ...
Everything @ rexresearch.com on a Data DVD !

ORDER PAGE

Started prototyping receiver to go with Michigan Mighty Mite Transmitter

wandrson Jul '16

In August, there will be two classes where we build a small one transistor morse code radio transmitter. The second class still has four tickets available, so if interested email me to reserve your seat!

Jul 2016
1 / 2
Jul 2016
May 17

August 20th Build a Michigan Mighty Mite Transmitter

In preparation for a couple of classes in September, I started prototyping a matching radio receiver for that transmitter. It will be a very basic direct conversion receiver with crystal control to match the frequency of the transmitter. With luck I will have the prototype completed by the first class on August 3rd

While waiting for some parts, I started assembling the front end to the receiver on some scrap pcb material.

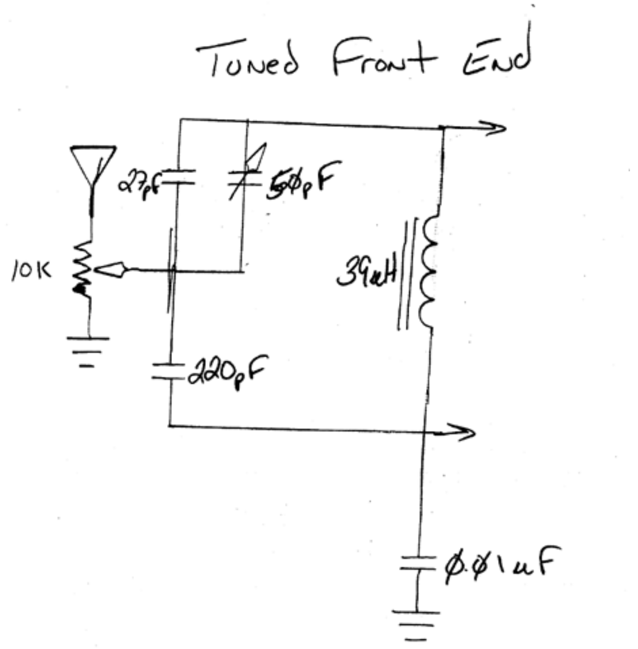

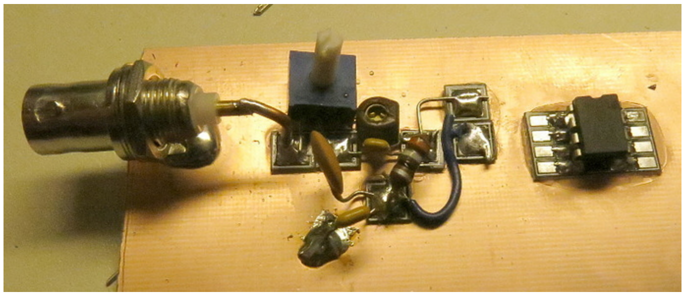

When working with projects like this, it is best to test (and if nescessary) fix the project in stages. So the first stage is a simple tuned front end (inductor-capacitor tank circuit). To ensure I did the math correctly, I plugged one of the crystals into a test oscillator I have and fed that into the antenna input and hooked up my scope to both the input and output of the tuned circuit as shown below.

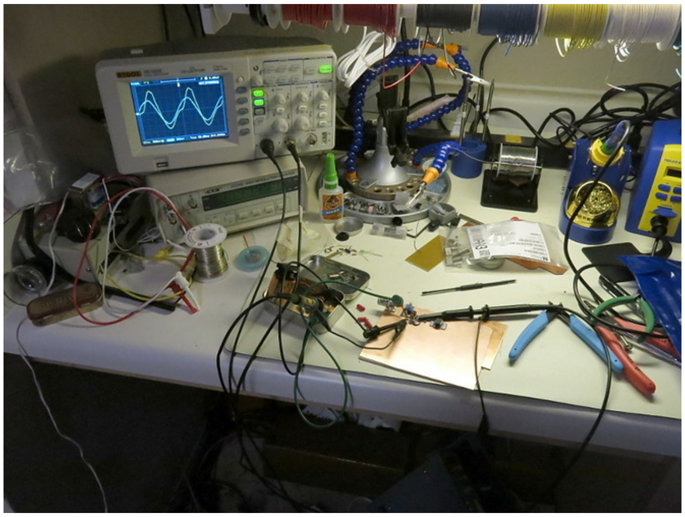

The circuit is aligned by using a non-conductive alignment tool (@artg_dms the Electronics committee could use a couple of sets of these) to rotate the little trimmer capacitor to maximize the amplitude of the output.

Here is a short video demonstrating the process:

If this kind of thing interests you, please sign up for the class (we can/will offer repeats of the transmitter class if there is interest) and come out to our monthly special interest group meetings for the Amateur Radio SIG!

Dallas Makerspace Show and Tell - August 2017

---

zmetzing May 17

> wandrson:
>
> If this kind of thing interests you, please sign up for the class (we can/will offer repeats of the transmitter class if there is interest) and come out to our monthly special interest group meetings for the Amateur Radio SIG!

I'm interested. 😉

Another interesting design for a 80m/40m receiver is here:

http://www.schripsema.org/pa3hdf/projects/dc-receiver/dc-rx.html

THE BOY'S BOOK of CRYSTAL SETS

BY

W. J. MAY

⁂

BERNARDS (PUBLISHERS) LTD.
THE GRAMPIANS
WESTERN GATE
LONDON W.6

INTRODUCTION

Before any attempt is made to construct a receiver, it is necessary to examine the problems which surround the crystal set, so that the best can be obtained from any of the designs attempted.

Firstly, it must be understood that the crystal set as it is to-day does **not** provide any amplification. It relies entirely on what is fed into it via the aerial and earth system and gives a very faithful replica of the original transmission.

From this it will be obvious that the aerial and earth system must be as efficient as possible if the final results are to be in any way outstanding. This because these are the only means by which the signals are fed to the receiver.

Secondly it is necessary to understand the nature of the transmitted signal, then it will be easy to understand the working of the set and to appreciate the function of each of the components.

When crystal sets first became popular, very little information was generally available, at least, not in a form that school-boys could understand.

We all built sets of all shapes and sizes, with coil designs that had to be seen to be believed, but very few of us had much idea of how they worked.

I well remember the case of a cousin of mine who, in those days acquired a magnificent variable capacitor or condenser as it was then called, this instrument had a most impressively engraved dial of polished ebonite, brass vanes and nickel plated end plates. Having observed similar dials on several highly priced commercial receivers in the town, he at once ripped out the somewhat tattered coil from the family receiver and replaced it with this device. The profound silence which ensued caused considerable amazement and dismay until a better informed adult explained the mysteries of L and C to him. Readers of this manual, however, will be better informed and there is no risk of failure with any of the designs described provided the instructions are carefully followed.

The Signal

To commence, we will assume that an orchestra is playing in a broadcasting studio. Since the principle of radio transmission is electrical it is necessary to change the sound produced by the orchestra into an electrical equivalent. This is carried out by the microphone, which picks up the sound and changes it into minute electric currents.

As they appear at the output of the microphone they are too small to be of use and accordingly are passed through a high power amplifier. These amplified currents could now be transmitted, but unfortunately, as we shall learn, owing to the inherent nature of the signal in this state, transmission over any useful distance would be impractical.

When the music from the orchestra is transformed into electrical currents they are in the form of alternating currents, usually called A.C., that is they rise to a maximum in one direction, fall to a minimum, rise to a maximum in the opposite direction and then fall to minimum again. This process is repeated over and over again. One complete rise and fall in each direction is called a cycle and is drawn in Fig. 1. Every time a note is struck on a piano, vibrations are sent out which reach the ear enabling you to hear it. These vibrations are also spoken of as cycles, they rise and fall in intensity the same way as an alternating current. The number of cycles radiated by any given note over a period of one second are referred to as its frequency. Middie C on the piano sends out 261 cycles every second and is known as having a frequency of 261. The microphone also " hears " the note and in the case of middie C produces minute A.C. at 261 cycles. This can be drawn as in Fig. 2, the only difference between Fig. 1 and 2 is, that the time factor is given so that the frequency can be identified. The higher the pitch of a note the higher the frequency and the lower the pitch the lower the frequency. On a piano the frequency of the top note is 3515 cycles and that of the lower 27 cycles. Those of you who have listened to an organ in a concert hall will have noticed that when a very deep note was played, it sounded like a growl to the ear, but the vibrating fre-

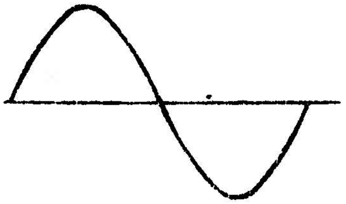

FIG 1

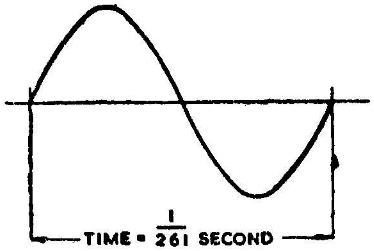

FIG 2

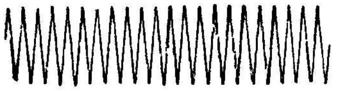

FIG 3

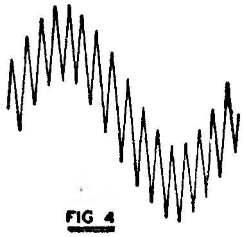

FIG 4

quency could be distinctly felt through the seat. Higher notes have too high a frequency to be observed in this manner.

The range of sounds which can be detected by the human ear are known as audio or low frequencies. From this you will understand that a low frequency amplifier is one which amplifies sound.

So that the transmitter will carry the programme over a useful distance it is necessary to radiate high frequencies. Now, as explained, the programme to be transmitted consists of low frequencies, and to overcome the difficulty, the transmitter generates A.C. of high frequency and combines it with the low frequencies. It will now be understood that the transmitted signal consists essentially of two different parts, a high and a low frequency content.

Fig. 3 gives a representation of the high frequency signal generated by the transmitter. In the case of the London Home Service, the frequency is 908000 cycles.

When referring to a high frequency signal on the medium or long wave-band it is usual to express the frequency in thousands of cycles, thus 908000 cycles becomes 908 kilo-cycles, which in turn may be abbreviated to 908 k/cs. It might at first be thought that by adding the low frequency or L.F. signal to the high frequency or H.F. carrier a form such as in Fig. 4 would result, such a combination is useless, and, so that the original L.F. content can be satisfactorily extracted by the receiver, the L.F. signal must vary the amplitude or output power of the H.F. signal as in Fig. 5.

It is in this form that the signal arrives at the receiving aerial. The aerial in itself is incapable of discriminating between one signal and another, and countless signals will be collected by the aerial at any one time. Many of these are too weak to be of use but the stronger ones must be sorted out since there is no point in receiving several programmes at once.

Fig. 6 shows the basic circuit of the input to a crystal. The coil L possesses a quality known as inductance, and the capacitor C, that of capacitance. If the coil had no capacitance whatever across it, all signals arriving at the aerial would be effectively short circuited to earth. As a matter of interest, it is impossible to obtain this state of affairs since even without any additional capacitor any coil **must** contain a certain amount of self capacitance.

By combining a coil and capacitor as in Fig. 6 a peculiar effect is observed, at one particular frequency, the signals are not short circuited to earth, but are developed across the coil. In other words the combined effect of L and C no longer provides a short circuit, **but only at one particular frequency.** If the value of C is altered the effect will be observed at a different frequency; likewise by altering L the frequency at which the effect will take place can be changed. There is a name for this phenomena, the frequency at which it occurs with any given L and C combination is known as the **resonant frequency.**

The values of the coils and capacitors shown in this manual have been carefully chosen so that resonance will be obtained at all frequencies where stations are broadcasting. Broadcasting stations work in bands of frequencies, those of major interest to crystal set constructors are the medium wave-band 1200 k/cs—600 k/cs and the long wave-band 300 kc/s—150 k/cs.

Usually a variable capacitor is used with a fixed inductance to cover one band and an additional coil switched in to increase the inductance to cover the other. In this way the L and C combination can be adjusted to provide resonance at the desired frequency of any given station. In other words you can **select** the station you want by varying C, that is turning the dial of the variable capacitor. This procedure is referred to as tuning. Having selected or tuned the required station it still remains necessary to change the form of the signal back to that of the original transmission. This process is called detection or de-modulation.

Examination of Fig. 5 will show that the signal has been duplicated, in other words, as it rises in one direction it also rises equally in the other. In this form the signal is useless since each half of the signal cancels the other, and if this signal is applied to a pair of headphones silence will result. Obviously some provision must be made to get rid of the unwanted half of the signal, and it is here that the crystal detector must be considered. This device will pass current in one direction only, ignoring any signal in the opposite direction, so that if the signal of Fig. 5 is passed through such a crystal, that of Fig. 7 will result. There is still the H.F. content to be reckoned

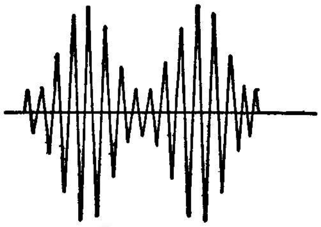

FIG 5

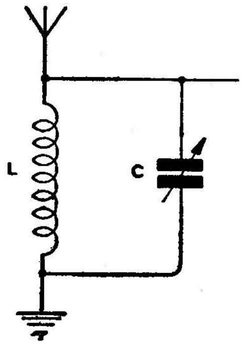

FIG 6

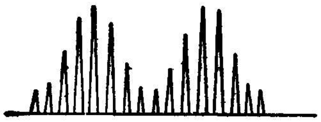

FIG 7

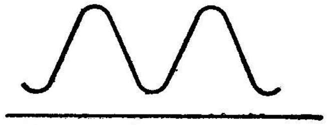

FIG 8

with, fortunately this is easily dealt with, a capacitor connected across the headphone terminals effectively disposes of this, leaving only the audio or L.F. content, as shown in Fig. 8. This audio content, which is a faithful replica of the original transmission is fed to the headphones. These in turn reverse the process of the microphone and transform the electrical currents into sound waves acceptable to the human ear.

Briefly then, your requirements are as follows:—

(1) A good aerial and earth installation, to make the most of the available signals.
(2) A receiver containing:—
  (a) Some form of coil and capacitor (L & C) combination to select or tune in the wanted station.
  (b) A crystal to get rid of the unwanted half of the signal (detection).
  (c) A fixed capacitor across the headphones to get rid of any remaining carrier.
(3) A pair of sensitive high resistance headphones.

Aerials

By this time the intending constructor will be able to appreciate the necessity of a good aerial. It is a point which cannot be over-emphasised. Assuming that you are in the fortunate position of being able to erect an outdoor aerial there are two main considerations, height and length.

One of the best that can be used is the inverted L shown diagramatically in Fig. 9.

It should be erected as high as is practical, every foot counts. The horizontal wire, that is the aerial proper, should have a minimum length of 60′ to which of course the length of the down lead is added. Where it is impossible to erect an aerial with an ideal horizontal length, a compromise must be effected.

Fig. 10 shows a three wire spreader aerial which gives quite a good effective length.

Suitable wire for a receiving aerial will not set any problems, stranded copper about 7/22 gauge is the best. 7/22 means that it consists of 7 strands of 22-gauge wire. This wire may be obtained covered, and for the present purpose is better than the plain or enamelled kind.

Note that insulators are used between the actual aerial wire and its anchoring supports, it is important that these are used, otherwise leakage will occur which will of course spoil its efficiency. Fig. 11 and 12 show how the wire may best be attached to two of the most common types of insulator available. Fig. 11 is of porcelain and is usually referred to as an egg insulator, whereas Fig. 12 shows a more modern (and more expensive) type in glass. If you are using the glass pattern, one is usually sufficient at each end of the aerial, but with the egg type two should be used. No doubt many readers will not be in a position to erect an out-door aerial, and must necessarily be content with an indoor installation.

The next best thing to a good out-door aerial is a replica constructed in a loft. If this form of construction is used, care must be taken when feeding the down lead to avoid close contact with the wall of the house. At the point where the lead feeds under the eaves, a length of rubber tubing can be used to cover the wire. The lead is fed through one of the small spaces left for ventilation purposes. A general idea is given by Fig. 13.

A less elaborate but quite effective aerial can be obtained by using a bed-spring. A length of 7/22 copper insulated wire is connected between the receiver and the spring. The spring should first be cleaned with emery cloth. Remove about 2″ of insulation from the wire and bind it tightly round the prepared spring. The joint may be covered with insulating tape which is obtainable from all electrical stores for a few pence.

Such aerials are quite popular since so many crystal sets are built for bedroom use.

When an aerial is required in the living room, the picture rail can be conveniently used. Insulated screw-eyes are fixed at intervals of about 3 feet along the rail, the wire is firmly anchored to one of them, and stretched right round the room until you arrive back at the starting point. By use of one of the insulated screw-eyes, the wire is secured and the down lead fed to the receiver. The idea is illustrated in Fig. 14.

Soldering

Before leaving the subject of aerials, a few words on soldering will not come amiss. Down leads on outdoor aerials should be soldered, and the same applies to loft types. Apart from aerial leads; earth leads and all the connections in the receiver will require soldering.

For the type of soldering necessary in radio construction an electric iron is the best solution. One of the small types marketed by Adcola or Henleys will be found admirable for the job. These are excellent for actual set construction but are hardly large enough for soldering the down lead to an outdoor type of aerial. Here the heat is dissipated much more quickly and a larger iron is required in order to get the solder to flow.

Ordinary irons which may be heated by a gas flame can be obtained very cheaply from most ironmongers' stores. Assuming the use of such an iron, first heat the iron until the copper bit is giving off a green coloured flame, the iron is now at the correct operating temperature. The bit is now discoloured or oxidised and its tip

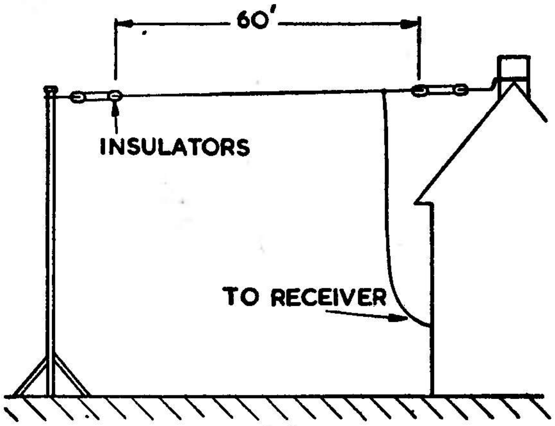
60'
INSULATORS
TO RECEIVER

INVERTED 'L' AERIAL

FIG. 9

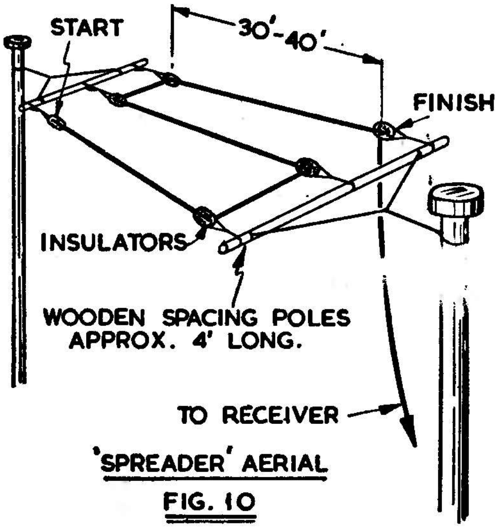
START
30'-40'
FINISH
INSULATORS
WOODEN SPACING POLES APPROX. 4' LONG.
TO RECEIVER

'SPREADER' AERIAL

FIG. 10

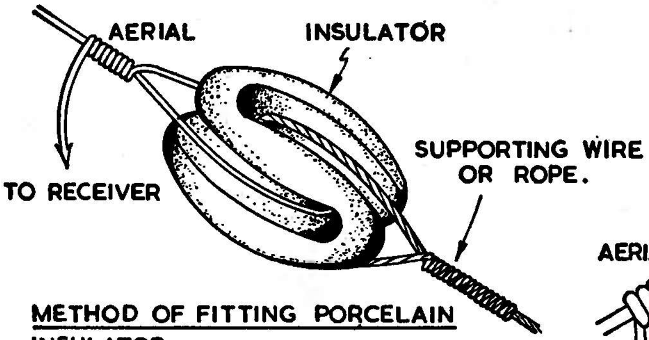
AERIAL
INSULATOR
TO RECEIVER
SUPPORTING WIRE OR ROPE.

METHOD OF FITTING PORCELAIN INSULATOR

FIG. 11

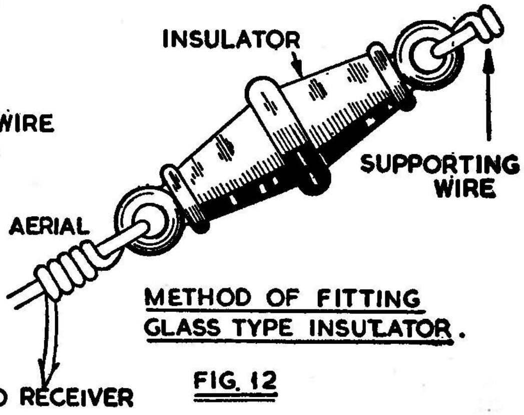
INSULATOR
SUPPORTING WIRE
AERIAL
TO RECEIVER

METHOD OF FITTING GLASS TYPE INSULATOR.

FIG. 12

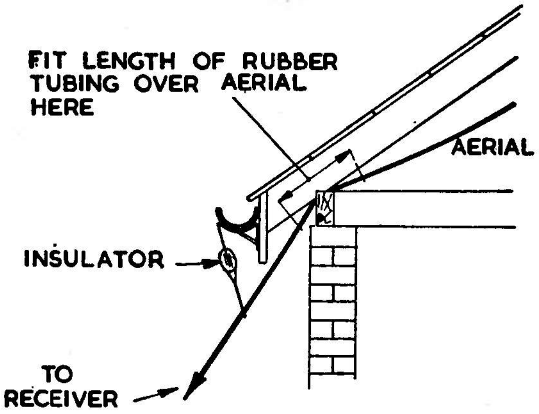
FIT LENGTH OF RUBBER TUBING OVER AERIAL HERE
AERIAL
INSULATOR
TO RECEIVER

LOFT AERIAL.

FIG. 13

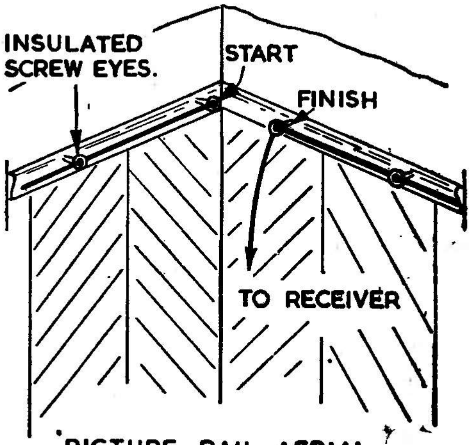
INSULATED SCREW EYES.
START
FINISH
TO RECEIVER

'PICTURE RAIL AERIAL.'

FIG. 14

should be quickly cleaned with an **old** file which should be kept specially for the purpose. A better idea is to obtain a small block of sal-ammoniac and rub the tip of the hot iron on it. This will clean the tip of the bit perfectly. Next take a length of cored solder such as Ersin Multicore, and melt a little on to the prepared tip faces, now smooth it evenly over the surface with a piece of old rag (be very careful not to burn your fingers) the iron is now "tinned" and ready for use. When heating the iron, be very careful not to let the bit overheat or get red hot, otherwise the tinned surface will be destroyed and the whole process will have to be gone through again.

An electric iron will not overheat, and the tinning will last much longer than with ordinary types, and since the heat does not deteriorate there is less likelihood of making faulty joints.

Having obtained a tinned iron, the process of soldering joints is remarkably simple and anyone with a little patience can acquire the art in a very short time.

To solder two copper wires together: clean the wire with emery cloth, apply a prepared iron and some cored solder to it, the solder will flow evenly over the wire thus tinning it. Repeat the process with the remaining wire and then twist the two together. Now apply the iron to the joint from the underside and the cored solder to the joint on top. Solder will flow evenly over the joint. Remove both iron and solder and allow to set. The solder will harden or set in a few seconds but during this period the joint must not be touched or moved, as otherwise the joint will be "dry" and quite useless mechanically or electrically. On radio components, tags for soldering are already tinned though if they are old, or discoloured it is best to re-tin them. Normally, however, it is only necessary to twist the connecting wire to it and apply the iron and solder as explained.

Remember: **never** apply the solder to the iron and then the iron to the joint, always apply the iron **and** the solder to the joint. It is however a good thing to apply a little solder to the iron tip even when it is perfectly tinned just before making a new joint.

One final "DON'T," you will have noticed that the solder referred to is "cored," that is, it contains resin and other substances through its centre, plain solder as used by electricians and plumbers will not do, as it is the resin or flux

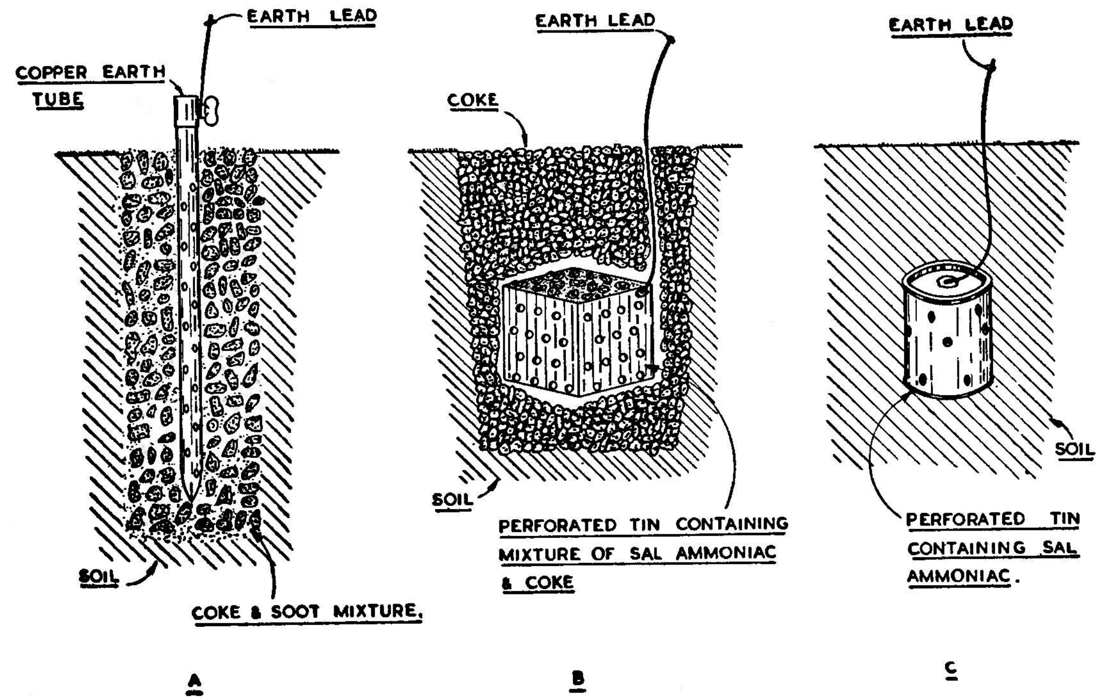
EARTH LEAD
COPPER EARTH TUBE
SOIL
COKE & SOOT MIXTURE.
A
EARTH LEAD
COKE
SOIL
PERFORATED TIN CONTAINING MIXTURE OF SAL AMMONIAC & COKE
B
EARTH LEAD
SOIL
PERFORATED TIN CONTAINING SAL AMMONIAC.
C

FIG. 15.

as it is called which makes the solder flow evenly and permits a good electrical joint.

Equally important do not attempt to use a separate flux or soldering fluid with plain solder or to "help" the cored solder; since these are almost certain to cause eventual corrosion and will destroy any components that have been contaminated.

Earths

The provision of a good earth is just as important as the rest of the installation.

Much disappointment would be avoided if this fact were not lost sight of, to avoid any slip up in this direction I am proposing to outline several well tried and efficient earth systems.

If it can be obtained, an earth rod, specially designed for the purpose provides the basis of a good earth connection. These were very popular some years ago when the majority of receivers were either crystal or battery operated, but in these days of modern mains driven sets they are not often used and consequently not always readily available.

Fig. 15 shows how it is used. First excavate a hole to a depth of three feet and fill with a mixture of soot and coke. Drive in the earth rod, which is a hollow copper tube, perforated. The lead to the receiver is connected at the top, and should be of covered 7/22 gauge, as used for the aerial system. Though not obvious it is important that this lead-in should be in covered wire, otherwise a number of indifferent earth contacts are likely to be made at various points along its length until it reaches the set. This is very undesirable and will spoil the efficiency of the system. It is essential to keep the soil surrounding the rod moist, which is one reason why the tube is hollow, and care must be taken to pour water into the tube at intervals.

A very efficient earth is the percolative type also popular at one time. Due to the chemicals used, it will extract moisture from its surroundings, thus maintaining a permanently moist earth.

It should be installed as in Fig. 15b.

The container is of copper or zinc, anything else will quickly rust away, again a good quantity of coke is used. Fill the container with a mixture of sal-ammoniac and coke, and then bury in coke as illustrated. This earth will not require further attention. It is possible to use powdered calcium chloride instead of sal-ammoniac but unless a chemist can be persuaded to make some up it is better to stick to sal-ammoniac.

If a zinc container is not available and you do not use or cannot easily get coke, Fig. 15c should be used. This makes a better earth than many so-called "earths" that I have come across. Obtain as large a tin as possible, make a number of holes as shown. Solder the lead in the bottom and fill it with sal-ammoniac. Replace the lid and bury in the ground. The tin will eventually rust away, but the replacement cost is negligible.

If it is quite impossible to make direct contact with the ground, a water pipe must be pressed into service. This should be a main pipe feeding straight to ground and not a hot water pipe or one fed from a tank. Scrape the pipe clean and twist the lead in tightly around it, a copper clip is even better. Do not attempt to solder on to the water pipe; since cold water is flowing through the pipe, it is extremely unlikely that your soldering iron will heat up the water supply sufficiently to allow a sound electrical joint, though you may spring a leak.

No attempt should be made to utilise gas pipes; the possibility of causing a fire is certainly very remote but, they make incredibly bad earth connections due to a number of joints made before true ground is reached. These joints are at best only semi-conductors, at least from an electrical standpoint.

Headphones

Since the late war there have been a large number of head-phone sets available on the surplus market. These may be roughly divided into two types, high-impedance and low-impedance. For the crystal sets detailed in this book high-impedance 'phones are required and the low-impedance pattern will not be suitable unless a matching transformer is used. As this is likely to cost more than the rest of the installation including the set, it will be as well to avoid them.

High-impedance types have an impedance of 2000Ω to 4000Ω whereas the low-impedance types are usually 600Ω.

The remaining consideration is weight; often cheap headphones are very heavy and uncomfortable to wear, every endeavour should be made to obtain 'phones as light in weight as possible.

Crystals

All the sets shown in this book have been designed to work with modern germanium crystals rather than the older galena crystal.

These germanium crystals require no adjustment, which, in itself, removes the main objection to this class of receiver. Suitable crystals are available from the following manufacturers and on the surplus market: Mullard, G.E.C., Brimar, Westinghouse and B.T.H. They are of robust construction, some in glass, others in ceramic or plastic, but must not be subjected to heavy knocks, otherwise the contact point may become dislodged.

From the information given, you should now be able to install an excellent aerial and earth system, and have some idea of how the crystal set works, so it is time to pass on to the actual receiver construction.

CONSTRUCTION 1

If you examine the following circuits you will find that each one is different. In most cases the difference lies in the coil design and/or the method by which the crystal and aerial is tapped into it. Each of these circuits has its own particular advantage to suit different conditions and the ideal circuit in some localities is not necessarily the best in others. It is not just a matter of a given circuit giving louder results than another, if it were there would be no point in showing more than one.

The main problem is to obtain adequate selectivity without reducing the volume level.

A receiver is said to be selective when it tunes sharply, a set with poor selectivity allows the stations to spread over the dial and when used near a transmitter will receive the local stations mixed together, which of course is useless.

Consider Fig. 16a, this is a very simple receiver, with no special attempt to provide any great amount of selectivity. In areas where signal

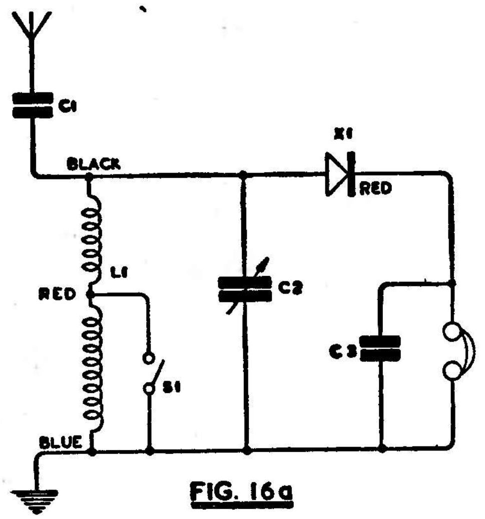

FIG. 16a

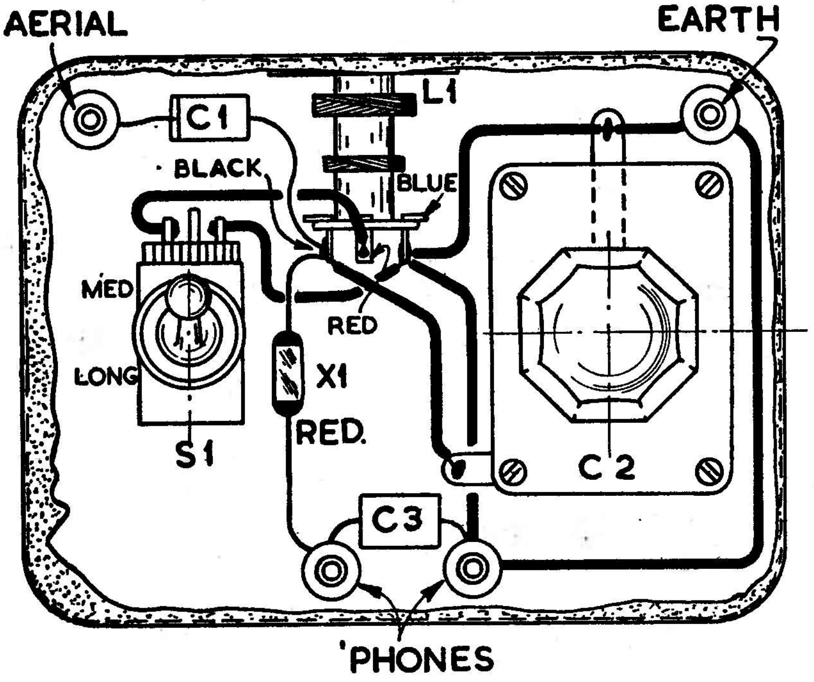

LID SHOWN BROKEN AWAY TO REVEAL POSITION OF COMPONENTS.

FIG 16b

strength is not high, or a short aerial is used, it will probably be ideal.

It would have been quite easy to increase the selectivity by providing a tap on the coil for the aerial as in Fig. 17a, but unfortunately, as the selectivity is increased overall volume is likely to decrease, so that unless you live close enough to the transmitter to have a large signal available and therefore need the selectivity, the circuit of Fig. 16a will be quite satisfactory.

Capacitor C1 is to prevent the aerial damping the circuit too heavily because this would flatten the tuning unnecessarily, however with small aerials it may be better to take the aerial direct to the black tag on L1. The signals are selected or tuned by L1 and C2, X1 is the crystal and C3 the capacitor across the 'phones to prevent unwanted carrier or R.F. reaching the 'phones.

A practical diagram (Fig. 16b) is provided showing the layout and all the wiring. You will require nuts and bolts to fix the coil, about ½″, 4BA size will do, the other parts have locking nuts provided. An old 2-oz. tobacco tin makes a very good container and keeps the size down. Note that the metal box is connected to earth. With the exception of AC/DC receivers the metal work on any receiver or amplifier is connected to earth.

When S1 is open as in the diagram, the set will tune in long-wave stations, but when closed the medium wave-band will be received.

Components List, Fig. 16a

C1	100pF mica capacitor.
C2	500pF tuning capacitor, solid dielectric.
C3	1000pF mica capacitor.
X1	Germanium crystal.
L1	Crystal Set Coil. R.E.P.
S1	Single Pole toggle switch.
4	Insulated wander-plug sockets and plugs.
1	2-oz. tobacco tin (or similar container).

Make sure that the wander-plug sockets are of the insulated type, otherwise the metal case will join all the sockets together electrically.

Try to follow the theoretical diagram when wiring, a little practice will soon enable you to wire up a set without a practical diagram, which is a great advantage because often only the theoretical diagram is given when circuits are detailed in the technical press.

Fig 28 on page 34 gives a list of symbols used on the theoretical diagrams, so that you can readily identify the components.

2

Fig. 17a circuit is similar in many respects to that of Fig. 16a. The difference is purely one of selectivity, tuning will certainly be sharper, and even with comparatively inefficient aerials the design will put up a very good performance. There is no reason why the construction should not follow the same lines as the previous receiver but, by enlarging the set a little and using an air spaced tuning capacitor, the efficiency is improved. Observe that the colour coding on the coil, an R.E.P. Dual Range (Blue Box) is different from that of the Crystal set coil used on the previous design, and make sure it is correctly wired in.

C1 is a mica compression capacitor, often referred to as a "trimmer" and because it is adjustable, permits the set to be matched to aerials of varying lengths. Construction is carried out on a square panel of bakelite, perspex or wood, as shown in Fig. 17b. Perspex being clear like glass presents a most attractive finish, provided the set is neatly wired. It must be drilled slowly, however, otherwise the generated heat of the drill will make the hole wander, and spoil the panel.

After the set has been wired and tested a small wooden container can be made to house the completed set.

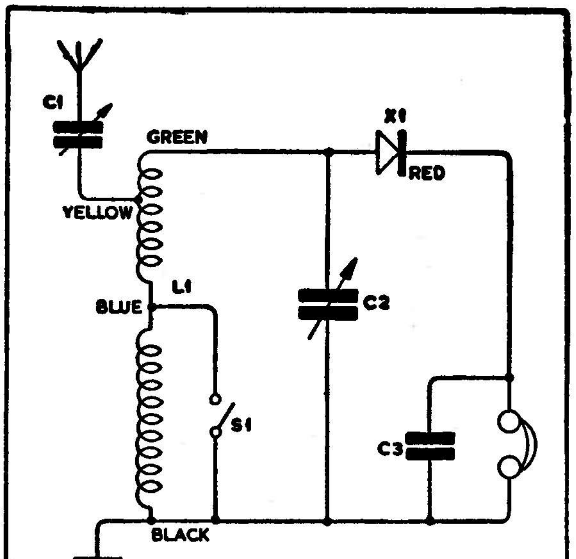

FIG. 17 a

When carrying out reception tests, a little experiment is well worth while, to get the best out of the set. Try removing the germanium crystal from the green tag on L1 and connecting it to yellow, at the same time removing the lead from C1 to yellow and connecting it to green. Once the best arrangement has been found the wiring can be left permanently in that position. Note that two tags on the coil are unused, this is intentional because on this circuit the extra winding connected to these tags is not required. With the switch S1 closed, the coil will cover the medium wave-band, and when open, long wave stations can be received. It is customary to abbreviate the expression medium wave-band to M.W. and long wave-band to L.W.

Components List, Fig. 17a

C1	100pF mica trimmer capacitor.
C2	500pF variable capacitor (air spaced).
C3	1000pF mica capacitor.
L1	Dual Range Coil R.E.P. (Blue Box).
S1	Single Pole toggle switch.
X1	Germanium Crystal.
4	Terminals or wander-plugs and sockets.
	Perspex, bakelite or wood for mounting panel.

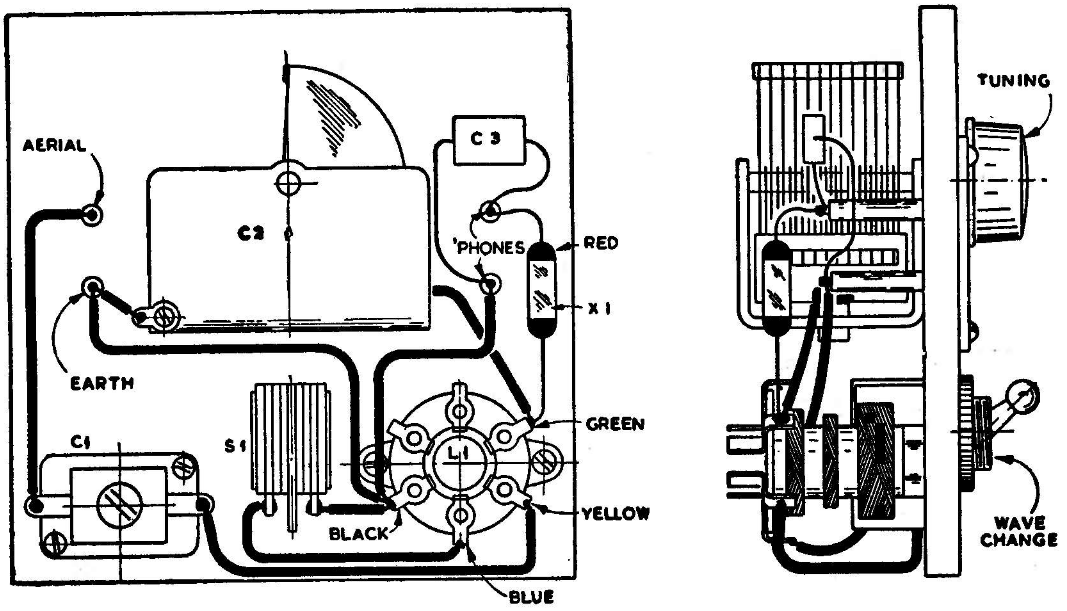

FIG. 17b

3

The design shown in Fig. 18a and 18b is more flexible than either of the preceding sets. That is, it can be varied to suit local conditions to a greater extent than the first two sets. This is made possible by the large number of taps provided on the coil.

You will notice that in this instance the coil is home-constructed on a cardboard former and is much larger than commercially produced coils. Coils wound in this manner are called **solenoids**, the turns are wound on side by side; in this case to a depth of some three inches. The diameter of the coil is made purposely large so as to obtain high efficiency. Modern commercial coils are invariably wave-wound and quite often are litzendraht which is the German for litz wire. This wire is made up of a number of strands of fine copper wire, each strand is enamelled to insulate it from the others, the whole is then silk covered. Litzendraht is more efficient than solid copper wire, and by using this and wave-winding, manufacturers can produce an efficient coil which is also small. Unfortunately wave-wound coils cannot be produced without a complex winding machine and home-constructed coils must take the solenoid form. It would be possible to use litzendraht but it is not easy to obtain by the reel, it is most expensive, and is difficult to handle. This last point is because at termination points, each strand must be cleaned of its enamel before a joint is made, and if one strand is broken, its advantage over plain copper wire is lost. How-

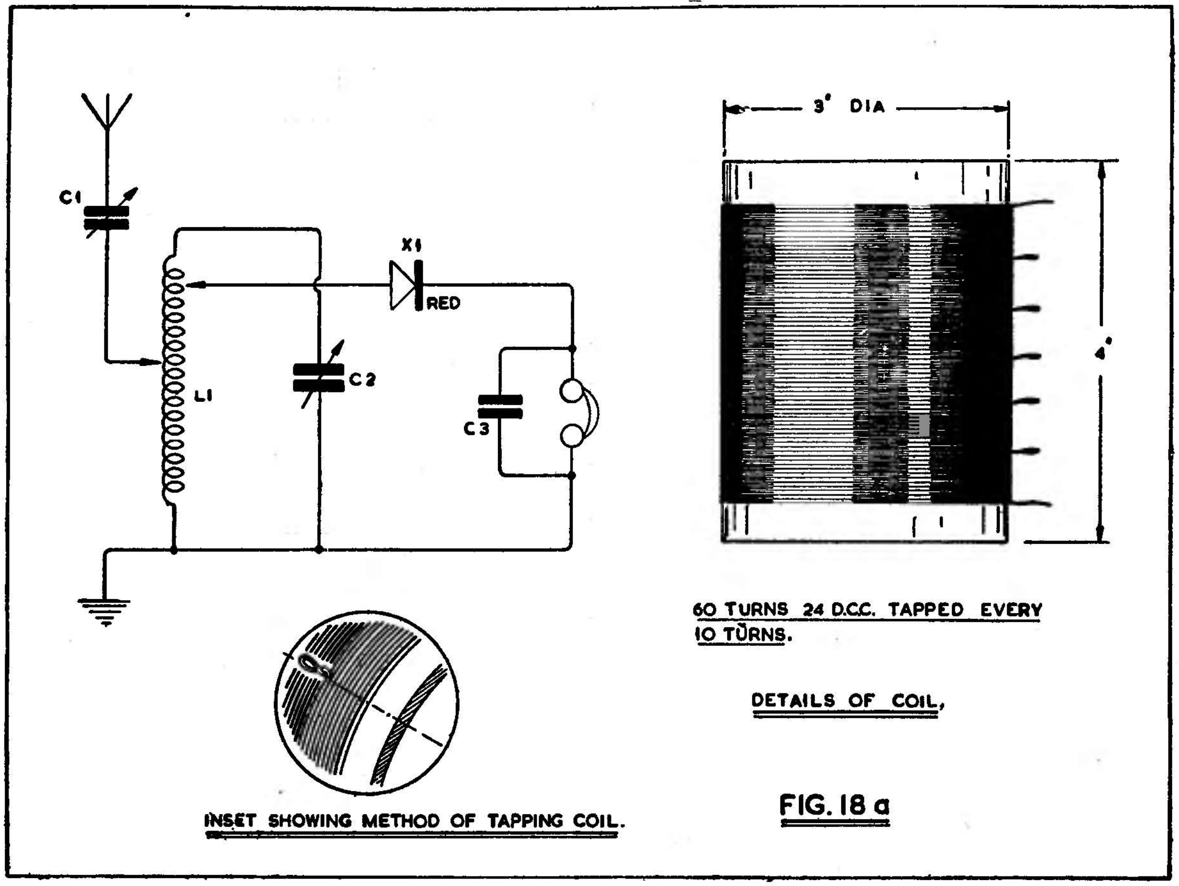
C1
X1
RED
C2
L1
C3
3" DIA
4"
60 TURNS 24 D.C.C. TAPPED EVERY 10 TURNS.
DETAILS OF COIL.
INSET SHOWING METHOD OF TAPPING COIL.

FIG. 18 a

ever, by using a former of reasonably large dimensions an efficient coil can be produced cheaply using ordinary copper wire.

The wire must be firmly secured at the start and finish of the windings. Pierce three small holes about ¼″ apart ½″ from the end of the former. Pass the wire through the first from the outside, return it through the second and pass it back again through the third. Leave some 6″ of wire at the end to make off the connection. It will now be possible to wind the turns on tightly without wire slipping.

Count on ten turns and make a loop 1″ long. The method of preparing loops is shown in Fig. 18a, loops or taps should always be made in this manner, never by baring the wire and soldering a further length of wire to it. Carry on with the winding, making off the taps every ten turns as directed until the coil is complete.

Construction is carried out on a wooden base-board and front panel. Secure the coil to the base-board as shown on Fig. 18b, mount the remaining components and proceed with the wiring.

When testing out, it will be found that the further the aerial is tapped down the coil towards the earth end, the greater the selectivity. A position should be found which permits separation of local stations without excessive loss of sensitivity.

The crystal tap is adjusted for best results and different settings of C1 tried out, when choosing the best position for the aerial tap.

Components List, Fig. 18a

C1	450pF padder.
C2	500pF variable capacitor.
C3	1000pF mica capacitor.
L1	See text and Fig. 18a.
X1	Germanium Crystal.
2	Crocodile Clips (to connect leads to coil taps).
4	Wander-plugs and sockets.

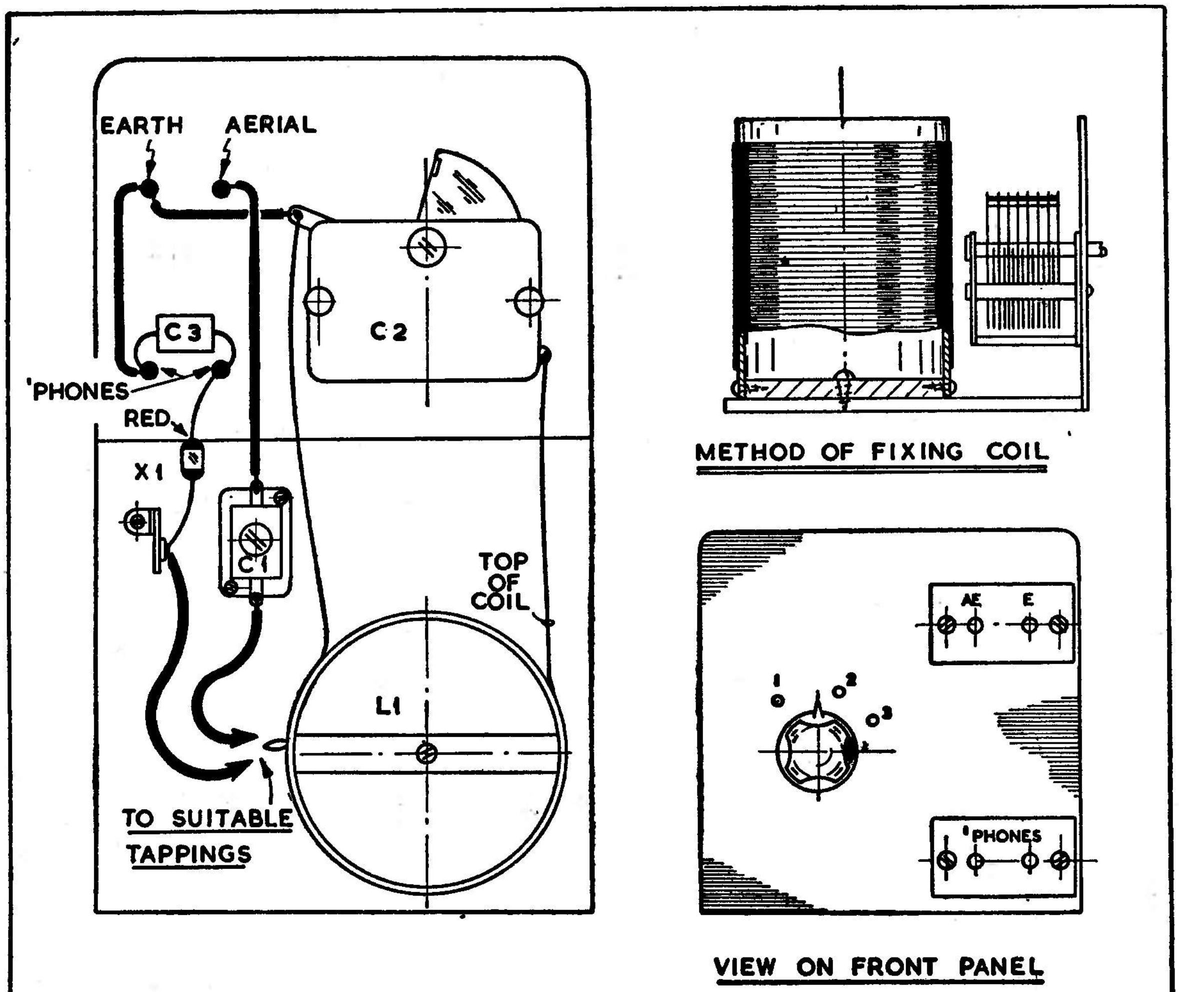

METHOD OF FIXING COIL

VIEW ON FRONT PANEL

FIG. 18 b

4

A most unusual design is reproduced in Fig. 19a.

It first appeared in America about three years ago, and it certainly does offer some advantage over more conventional sets. The tuning circuits are duplicated and two crystals are used, so the circuit may be described as a full-wave receiver. Signals received will certainly be louder than with simple sets, but care must be taken with the coil winding, as with the other receivers using home-made coils, otherwise results will be disappointing. Be very careful to ensure that all three windings are in the same direction, this is very important.

The distance between each winding should be ¼″. After the coil has been wound it is a good plan to warm it before a fire and paint the windings with "Durafix." Heating the coils makes the "Durafix" run freely. This substance sets quite hard and there will be no risk of the windings loosening. A small quantity of enamelled wire will cover the requirements of this coil, a 2 oz. reel will provide more than sufficient.

Before wiring in the germanium crystals, examine them carefully, note that one end is coloured red or in some cases marked with a positive sign thus +. You will notice, the sign is the same as the addition symbol used in arithmetic.

It is essential that both the red or positive ends are connected together, note that this is clearly marked on Fig. 19b. The receiver cannot work if

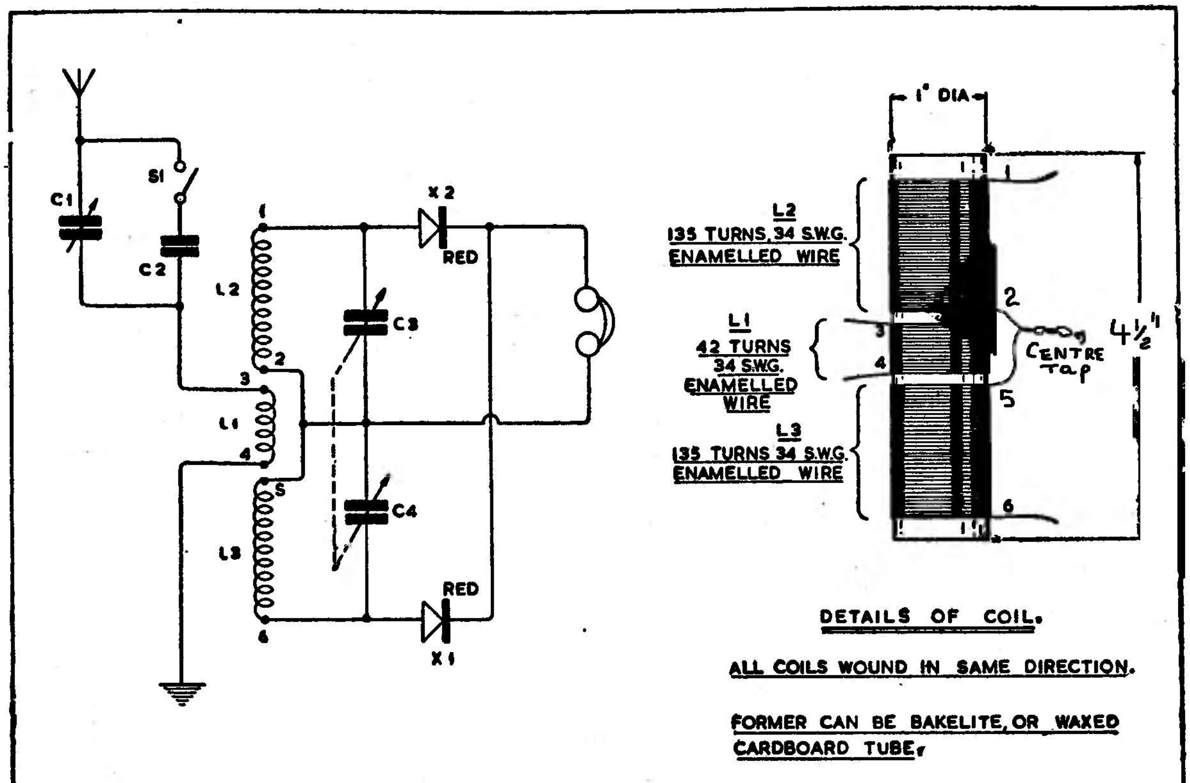

DETAILS OF COIL.

ALL COILS WOUND IN SAME DIRECTION.

FORMER CAN BE BAKELITE, OR WAXED CARDBOARD TUBE.

FIG. 19a

one of the crystals is connected the reverse way round.

A baseboard and panel form of assembly is used for this set which looks quite attractive if housed in a small polished cabinet. Dimensions are not at all critical but Fig. 19b gives a general idea of the layout that should be used. To avoid any unnecessary losses the coil should be mounted on perspex.

Obtain two strips of perspex 5½" long by ½" wide, drill fixing holes at both ends of each strip. It is best to clamp the two together when drilling these holes. Place one strip on the board in position, then put the coil and remaining strip over the first one and screw down. The coil will be clamped neatly and rigidly into position.

Twin socket bakelite strips are used to carry the aerial/earth and 'phone connections, if desired terminals mounted on pieces of bakelite could be used. These socket strips can be purchased from most shops selling components. If feet are not provided, drill two holes and screw into the edge of the base-board. Tag-strips are used to anchor some of the wires from the coil, and the germanium crystals. These also can be purchased for a few pence. At least two tags are necessary on one, and three on the other, though if the strips have more tags than required it is of no consequence. C3-C4, the ganged tuning capacitors have trimmers fitted, this should be stipulated when purchasing, and when first operating the set these small trimming capacitances are set about half-way. The idea of these trimmers is, that they compensate for any difference between the self-capacity of the tuning coils. With the switch S.1 open, that is in the OFF position, turn

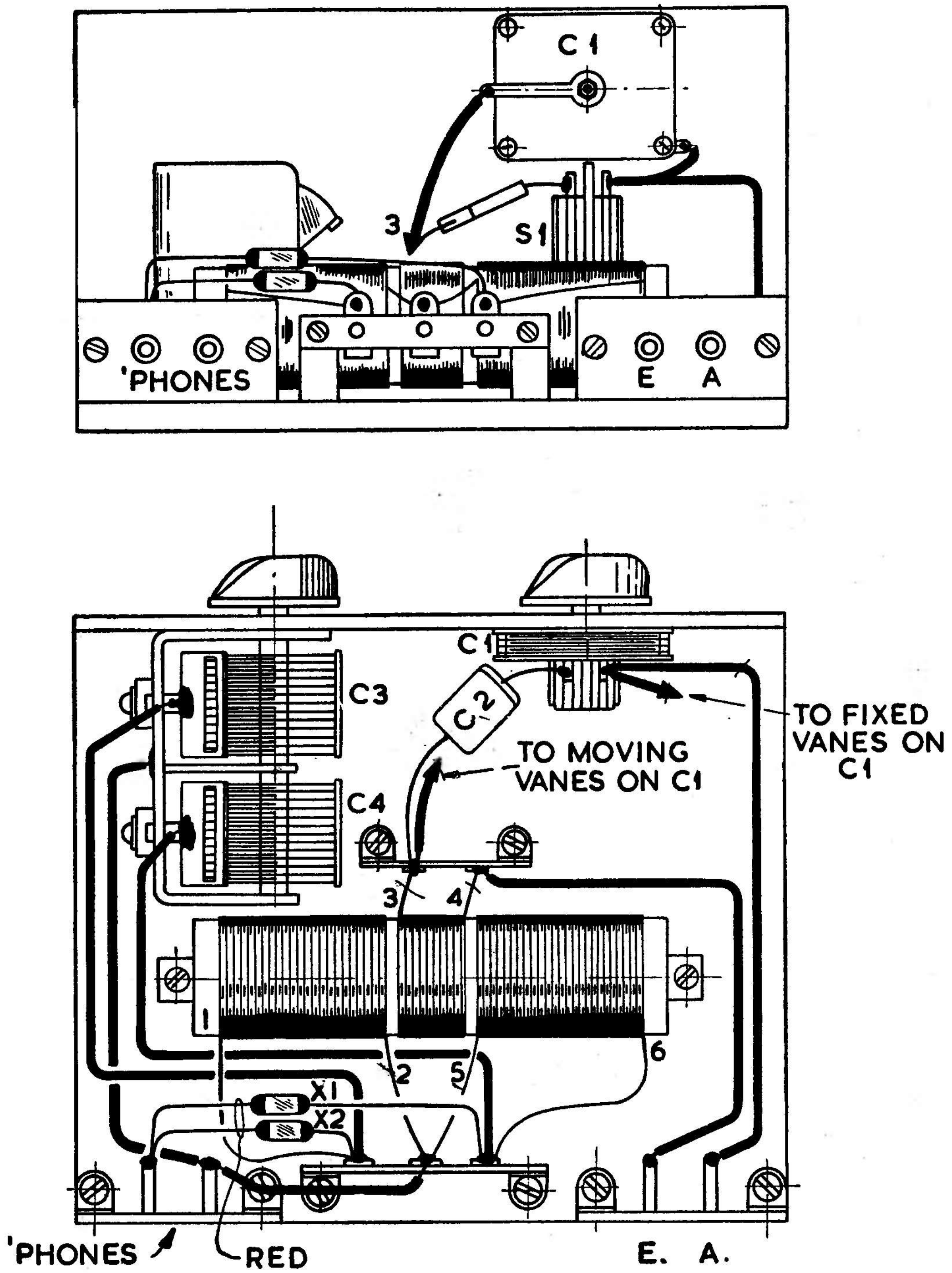
C1
3
S1
'PHONES
E
A
C1
C3
C2
TO FIXED VANES ON C1
TO MOVING VANES ON C1
C4
3
4
1
2
5
6
X1
X2
1
'PHONES
RED
E.
A.

FIG. 19 b

in a station at the high-frequency end of the band, that is with the vanes of C3-C4 towards the disengaged position. Adjust C1 for maximum volume, without interference from other stations. Now adjust the trimmers for loudest headphone strength. It may be necessary to run over the adjustments several times for best results. Once they are correctly set, the trimmers require no further adjustment. The tuning coil covers the M.W. band, 1500 kc/s. to 600 kc/s., as the capacitor C3-C4 is advanced and the vanes start to mesh, the frequency to which the receiver tunes **decreases** so that with the vanes disengaged the set is tuned to 1500 kc/s. approximately and at full mesh, 600 kc/s. Switch S1 is used as follows, for stations **lower** in frequency than 850 kc/s it should be in the ON position with the contact closed, but for stations **higher** in frequency, it is in the OFF position. As an example both the London Home Service and London Light Programme transmitters on the M.W. band operate at a higher frequency than 850 kc/s. If any doubt exists as to the operating frequency of a given station, reference should be made to the Radio Times which quotes both wave-lengths and frequency. If you know the wave-length in metres, it is simple to find the frequency. Divide 300,000 by the wave-length in metres, the dividend equals the frequency in kilocycles, i.e., 300,000 ÷ 300 metres = 1,000 kc/s. In the same way, dividing 300,000 by the frequency in kilocycles will produce the wave-length in metres.

To use the receiver, tune in the signal by C3-C4, adjust C1 for maximum volume without allowing stations to overlap.

Components List, Fig. 19a

C1	300pF solid dielectric variable capacitor.
C2	1000pF mica capacitor.
C3-4	500pF twin gang variable capacitor.
L1-2-3	See text and Fig. 19a.
X1-2	Germanium Crystal.
2	Tag-strips.
2	Twin socket strips.
S1	Single Pole toggle switch.

5

Whenever crystal sets are discussed, constructors are apt to think in terms of medium-wave reception. When you consider that at least 98% of published circuits are designed for this band (sometimes with the long-wave band thrown in as an afterthought) it is understandable. However, if you can provide a good outdoor aerial (and an equally good earth) there is a lot of fun to be had listening to the short-wave bands. This receiver is designed specially for short-wave reception.

A metal chassis is used for construction, these can be obtained ready made in aluminium from most good supply houses. It can be quite small and on the original model a 6″ x 4″ was used. Low loss components are used so as to obtain greatest efficiency on the short wave bands. The coil is a commercial product which plugs into a 4-pin base. Best results were obtained on the 3mc/s to 7mc/s band though coils covering other bands are available.

Note that the tuning capacitor C1 is smaller than normally used and has a maximum capacitance of 140pF.

Coil type 706/R covers the 3mc/s to 7mc/s band, but the set is equally satisfactory on the M.W. band, and to cover this coil type 706P should be used. This coil has an iron dustcore which can be adjusted to make the coil cover the required band. The effect of the core is as if turns were being added or removed from an ordinary coil. When dealing with the short waves it is customary to refer to the frequencies in terms of mega-cycles (mc/s) rather than kilo-cycles. One mega-cycle is equivalent to 1000 kilo-cycles, that is 1,000,000 cycles.

Layout and wiring can be clearly followed from Fig. 20b and a theoretical diagram is shown in Fig. 20a. Care should be taken when arranging a mounting for the aerial terminal. The best plan is to cut a 1″ hole in the chassis, mount the aerial terminal on a piece of perspex and screw into the chassis. This avoids any unnecessary losses between aerial and chassis. The earth terminal is screwed directly into the chassis.

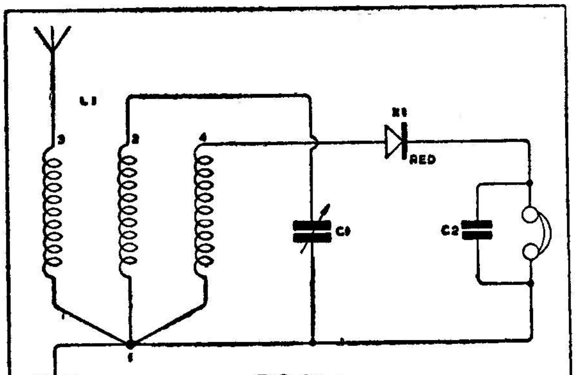

FIG. 20 a

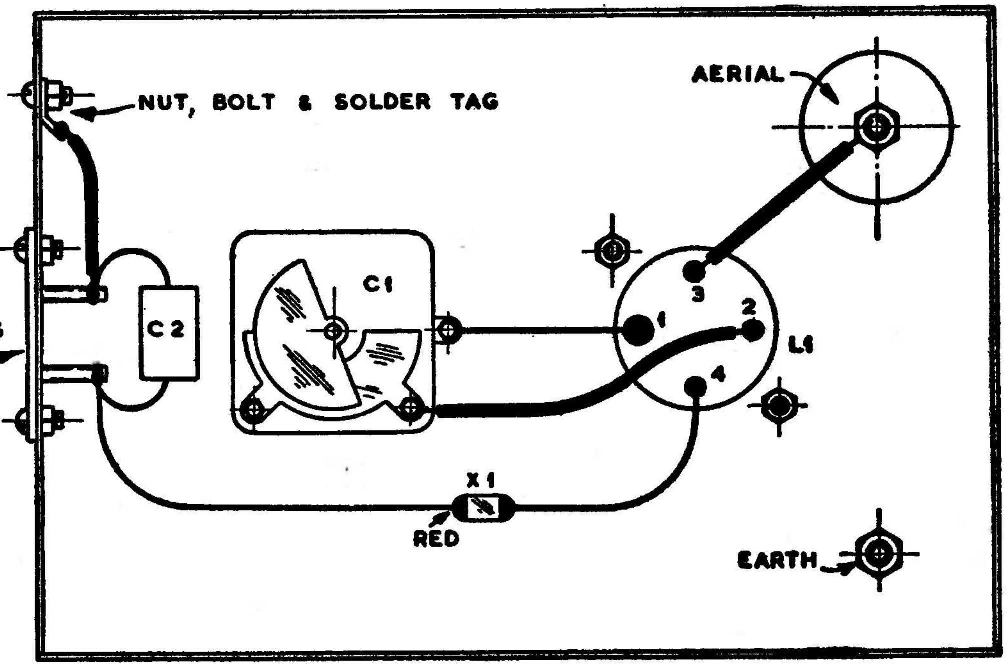

FIG. 20 b

Wiring as shown in Fig. 20a is likely to give the best results, but other arrangements are possible and to get the best out of the set they should be tried.

Refer to Fig. 20a. The aerial is connected to pin 3 on L1, C1 to pin 2 and X1 the crystal to pin 4. If you have a long aerial try this combination—Aerial to pin 4, C1 to pin 2, and the crystal to pin 3. Connections to pin 1 are not altered. In a few cases the following arrangement may prove best. Crystal **and** C1 to pin 2, aerial to pin 3, pin 4 left free and again pin 1 is unaltered.

Tuning on the short-wave bands is more critical than on the medium waves so tune very slowly over the band and remember signal strength is likely to vary from day to day.

Components List, Fig. 20a

C1	140pF variable capacitor Eddystone 586.
C2	1000pF mica capacitor.
1	Coil holder Eddystone 707.
L1	3m/c—7m/c coil Eddystone 706/R.
1	Engraved tuning dial.
1	6"x4"x2½" chassis.
2	Terminals (Aerial-Earth).
1	Twin socket strip ('phones).
X1	Germanium Crystal.

6

Fig. 21a-b is yet another design from America and appeared in Radio Craft some years ago, designed by Mr. W. J. Spain. The original used a silicon crystal which preceded the development of the germanium. Results are certainly very good though of course on the model built here, a germanium crystal was used. Selectivity can be adjusted to suit all conditions and even at short distances from local stations. Good volume can be obtained without the programmes overlapping. Home made coils are used and for best results a 75′ outdoor aerial is desirable.

Baseboard and panel construction is used, which can conveniently be housed in a small wooden cabinet.

Operation is a little more complicated than some of the other designs, but the results certainly merit the extra trouble taken. As with the other receivers much of the success obtainable is due to the coils and these must be carefully made otherwise results will be disappointing.

Layout and wiring can be followed quite easily from Fig. 21b and requires no special comment.

First make the coils. L1 is a tapped coil, the taps should be made in the same way as shown in Fig. 18b. It consists of 90 turns of 22 D.C.C. (double cotton covered) copper wire tapped at the following number of turns, 5, 10, 15, 25, 30, 40, 50, 60, 70 and 80. The former used has a diameter of 2". L2 is also wound on a 2" diameter former, this coil however is not tapped, it consists of 110 turns of 38 D.C.C. or enamelled copper wire. Great care must be taken when handling this wire, it is quite fine and in the hands of the inexperienced will easily break.

After wiring has been carried out the set can be put into operation.

Selectivity is controlled by S2, in position B tuning is broad but in position A it can be quite sharp.

When searching for a station switch S2 to B. Tune the station and switch S2 to A. The frequency range is controlled by the position of the tap on L1, a good plan is to start with the 5 turn tap on L1.

Components List, Fig. 21a

C1	500pF variable capacitor.
C2	500pF variable capacitor.
C3	1000pF Mica Capacitor.
X1	Germanium Crystal.
L1 & L2	See text.
S1	Single Pole toggle switch.
S2	Single Pole 2-way toggle switch.
2	Twin socket strips (Aerial/Earth and 'Phones).

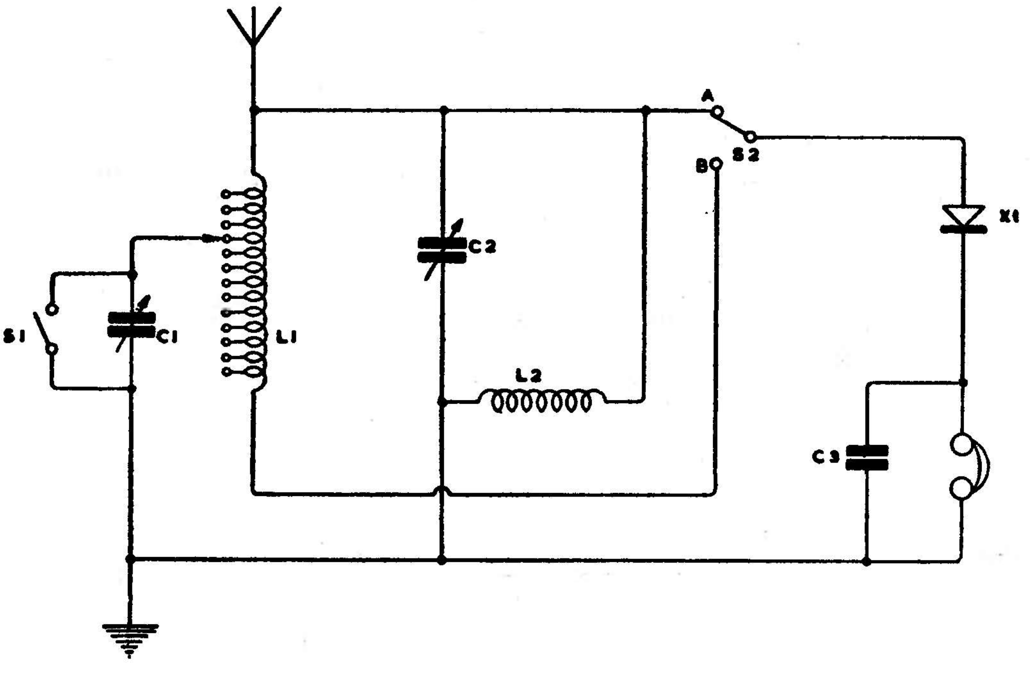

FIG. 21 a

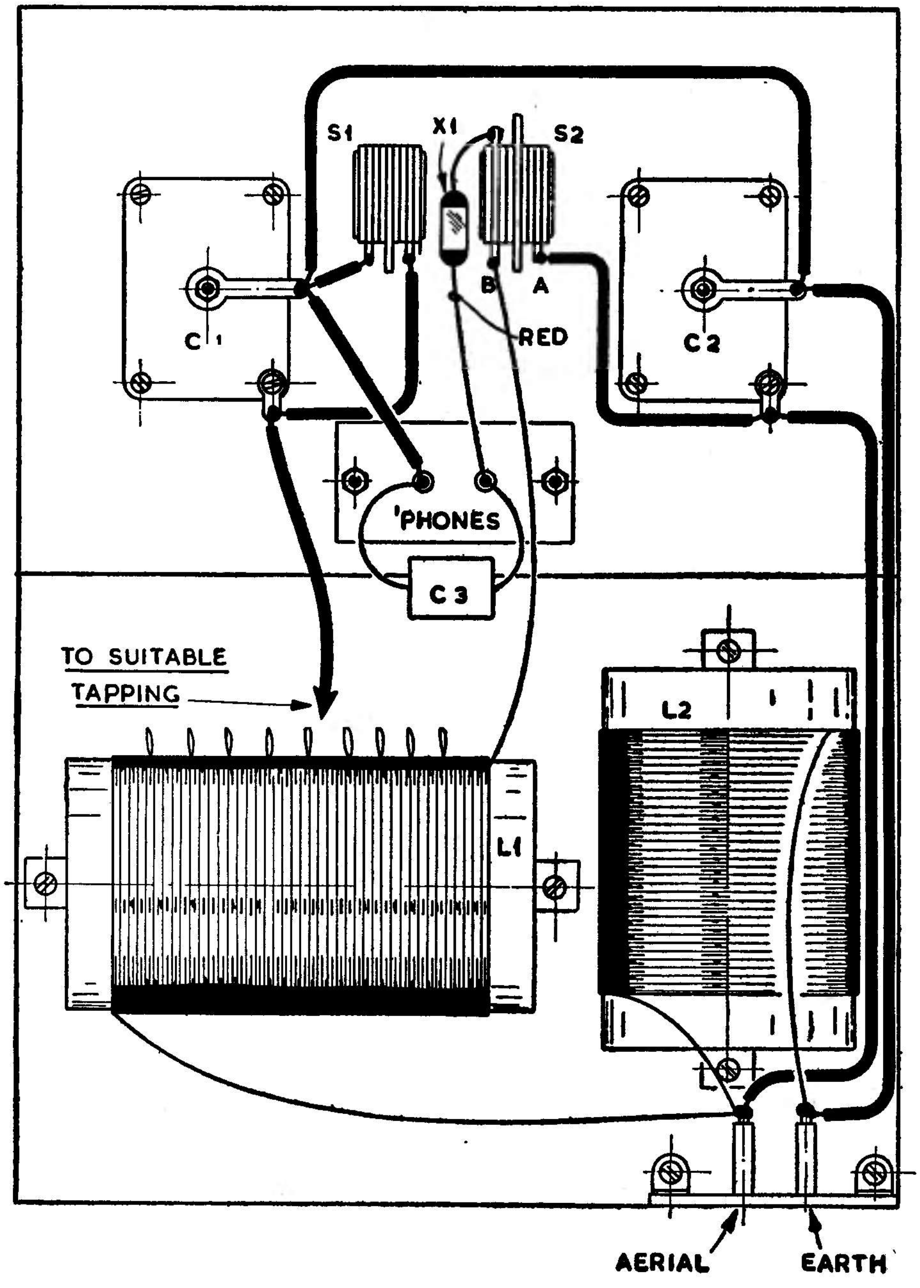
S1
X1
S2
B
A
C1
C2
RED
'PHONES
C3
TO SUITABLE TAPPING
L1
L2
AERIAL
EARTH

FIG. 21 b

Keep S1 open and C1 for tuning, adjusting the tap on L1 for best results. Very powerful signals are best brought in with S1 closed. When using C1 and adjusting the taps on L1, switch S2 is kept in the A position.

It will be found that C2 acts as a fine control on C1. One division of the tuning dial on C1 is roughly equal to a movement of 20 divisions on C2.

Note that the coils are mounted at right angles to one another and that in this instance wood has been used as far as possible, an aluminium or other metal chassis is not suitable.

7

A very useful receiver for both medium and long-wave reception is possible with the circuit of Fig. 22a-b. A home-made coil is used. On the original a 1¼″ diameter former was used but in fact this is not very critical and a 1½″ former could be used. An interesting point is the construction of the long-wave coil section. Solenoid coils as used for the medium wave-band are not efficient if the length is too great relative to the diameter. Now as the long-wave section has 300 turns, if wound as a solenoid the winding length would be very great and a lot of efficiency would be lost. To overcome the difficulty the coil is pile wound. If the whole coil were to be wound in one pile it would still be inefficient because of the capacity formed in the winding, but by dividing it into five sections a coil of reasonable efficiency results. The same remarks concerning winding apply to this coil as to all the others in

Components List, Fig. 22a

C1	300pF Mica Capacitor.
C2	500pF Variable Capacitor.
C3	1000 Mica Capacitor.
X1	Germanium Crystal.
L1-2	See Text.
S1	Single Pole Toggle Switch.

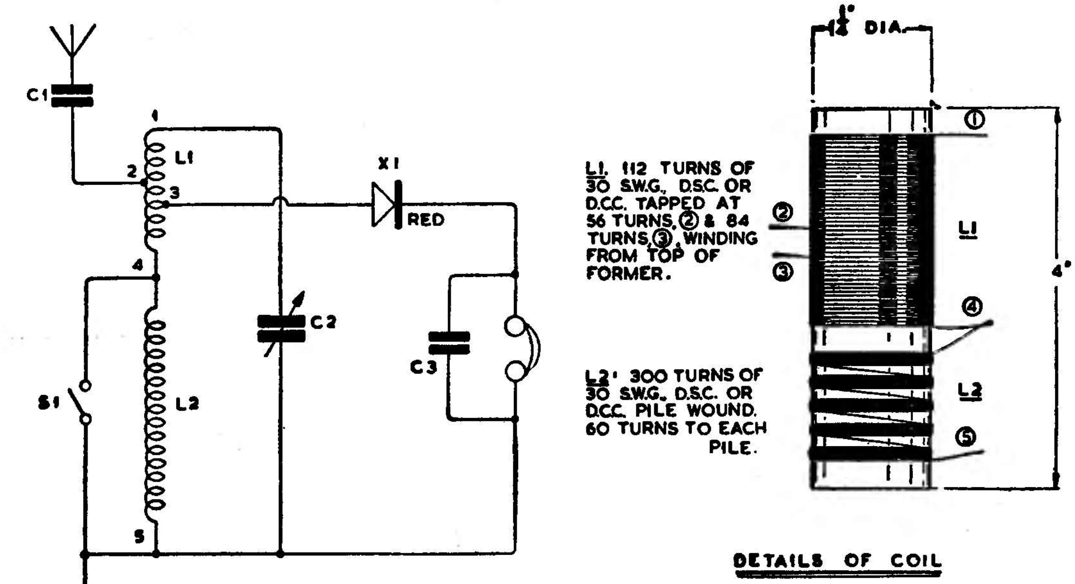
C1
1
2
3
4
5
L1
L2
S1
X1
RED
C2
C3
1¼″ DIA.
①
②
③
④
⑤
L1
L2
4″
L1. 112 TURNS OF 30 S.W.G., D.S.C. OR D.C.C. TAPPED AT 56 TURNS, ② & 84 TURNS, ③, WINDING FROM TOP OF FORMER.
L2. 300 TURNS OF 30 S.W.G., D.S.C. OR D.C.C. PILE WOUND. 60 TURNS TO EACH PILE.

DETAILS OF COIL

ALL COILS WOUND IN SAME DIRECTION.

FORMER CAN BE BAKELITE OR WAXED CARDBOARD TUBE.

FIG. 22a

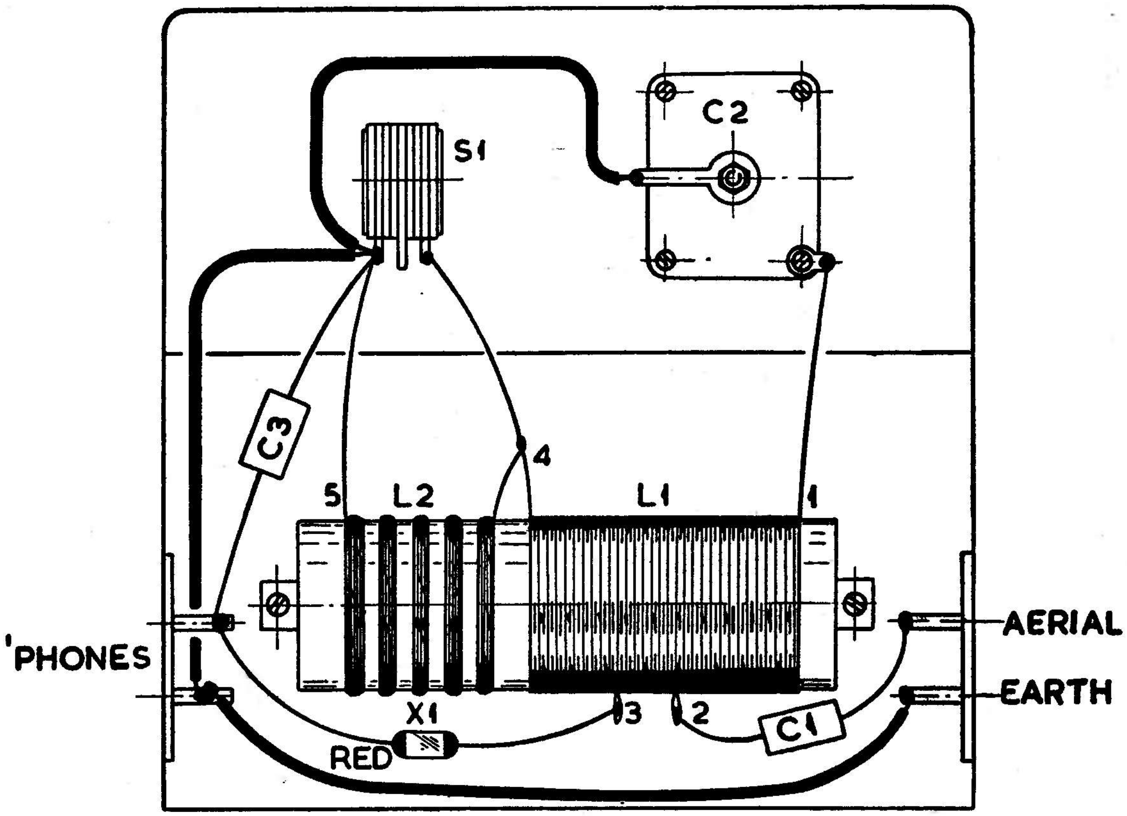

FIG. 22 b

this book, and readers should not, by this time, encounter any difficulties.

The design provides good selectivity with good sensitivity, and provided the set is not required to operate within a few miles of a powerful transmitter, no trouble will be encountered from interference between adjacent stations.

A wooden base-board and front panel is best for this design, and the wiring and layout can be clearly followed from Fig. 22b. Switch S1 is open for long-wave reception and closed for the medium-wave band.

Coil mounting can be on the same principle as for the full-wave design Fig. 19a-b.

8

If you look at all the other circuits shown in this book you will find that in each case the tuning capacitor and tuning coil are in parallel, this is by no means essential and the circuit of Fig. 23a shows a series arrangement that is the tuning capacitor is between the coil and earth instead of being wired up across it. This design also appeared in **Radio Craft** a few years ago.

An interesting point is that the coil is a conversion adapted from another type of component. To make this coil you must obtain an old I.F. transformer as used in superheterodyne receivers.

You will require one designed to work anywhere between 450 and 470 kc/s. Some of the very early ones worked at 110 kcs/ but these are unsuitable.

The I.F. transformer must also be of the type tuned by an iron dust core at each end.

Dismantle the transformer, remove the parallel capacitors and mounting wires so that you are left with two coil bobbins and their former. Be very careful when un-soldering the wires from the supports because the coils are usually wound in litzendraht and must remain intact. Remove 25% (approximately) of the turns from one coil which we will call the secondary, and completely remove the iron dust slug from this coil, now cut the former in half, and mount as in Fig. 23a. Warm the coils with heat from your soldering iron or from a fire and gently slide the coils into position at the ends of the former as in Fig. 23a. The wax will set again holding the coils in their new position.

Note that the wires from the coils are connected to tags on the mountings. These are pieces of bakelite, and usually the original end pieces used on the I.F. transformer can be used.

A word about soldering litzendraht. Do not attempt to use emery paper. Prepare the end of wire by removing the outer silk covering and dip the end in methylated spirits. Ignite with a match and after about five seconds wipe with a cloth held between finger and thumb. This will quench the flame and will remove all traces of enamel from the wire. It is a good plan to experiment with the odd length of wire removed from the secondary coil. If the coils prove to be constructed of ordinary copper wire, it can of course be prepared with fine emery cloth in the normal manner.

Components List, Fig. 23a

C1	500pF variable capacitor.
X1	Germanium Crystal.
L1	See Text.
4	Plugs and sockets or terminals.

Wiring is straight forward as can be seen from Fig. 23b and requires no explanation.

Operation is as follows, selectivity is controlled by varying the distance between coils with the wing nut, and to an extent by the position of the primary iron dust core. Remember that the aerial and earth installations are as important with this set as with any other crystal receiver. No parallel capacitance appears necessary across the 'phones, the set works quite happily relying on the self-capacitance of the head-phones.

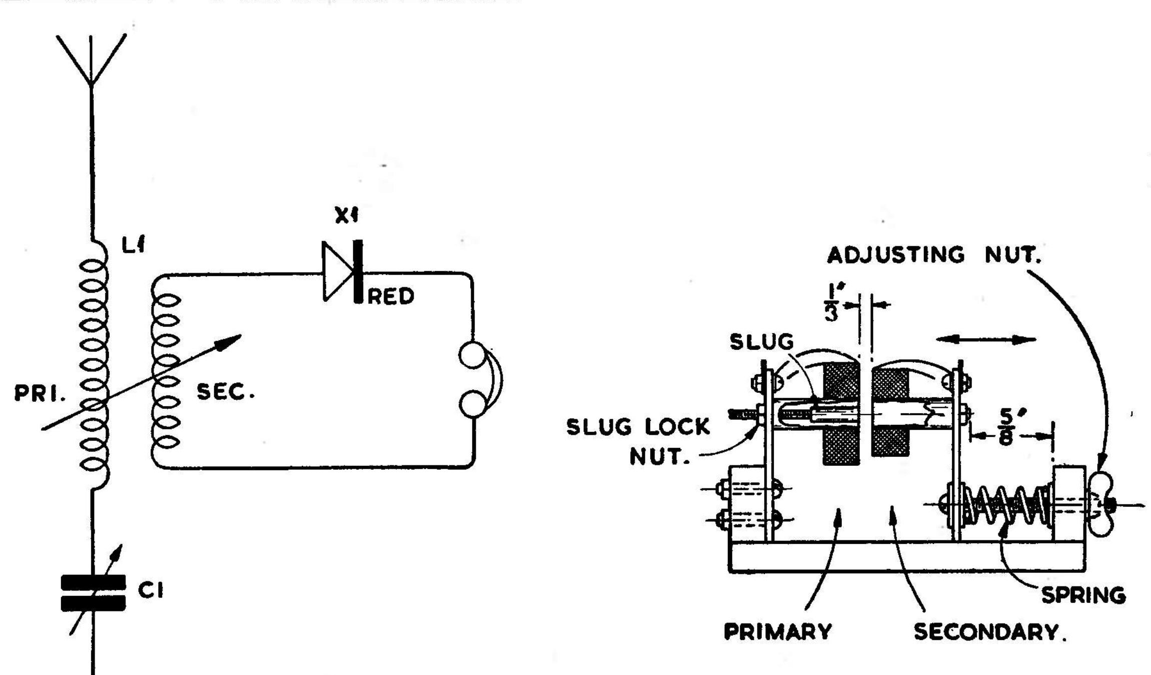

FIG. 23 a

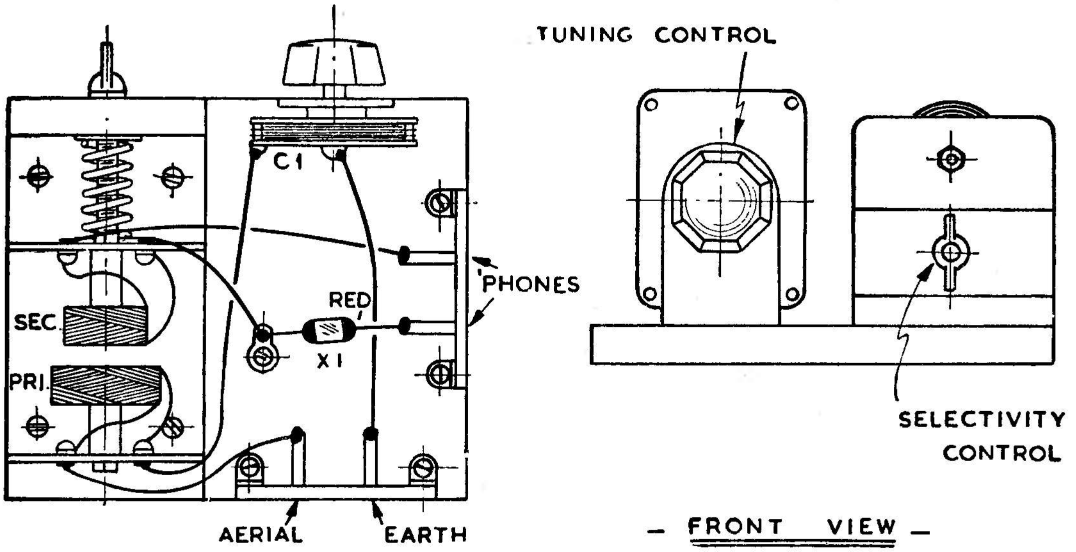
TUNING CONTROL
C1
SEC.
PRI.
RED
X1
PHONES
SELECTIVITY CONTROL
AERIAL
EARTH
– FRONT VIEW –

FIG. 23 b

9

So as to give a truly varied selection of circuits, some of the sets have been designed around commercial coils whilst others use coils which must be home-constructed.

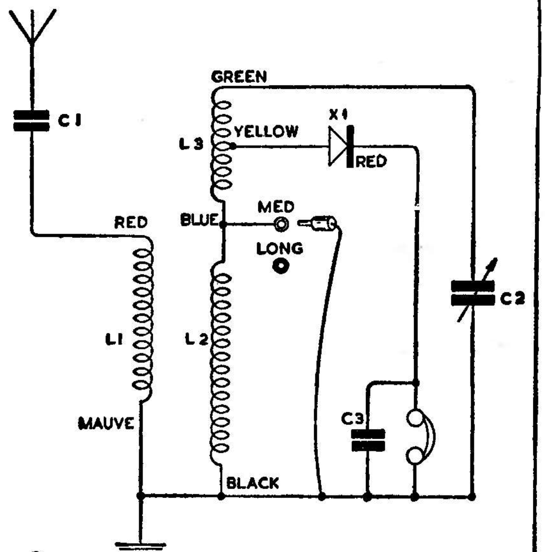
C1
GREEN
X1
L3
YELLOW
RED
RED
BLUE
MED
LONG
C2
L1
L2
MAUVE
C3
BLACK

FIG. 24 a

Most enthusiasts like to build their own coils but when it is a question of something really small combined with efficiency a commercial coil is the obvious choice.

Fig. 24a-b uses a tobacco tin or a similar container just as the design shown in Fig. 16a.

This receiver however is a little more selective and is more suitable than the earlier design, if you are near a transmitter or have a very long aerial. Take great care to connect the coil according to the colours shown on the diagrams. Sockets are used for wave-changing, this cheapens the cost a little and they are just as efficient.

Components List, Fig. 24a

C1	100pF Mica capacitor.
C2	500pF variable capacitor solid dielectric.
C3	1000pF Mica capacitor.
X1	Germanium crystal.
L1-2	R.E.P. dual range coil. Blue box.
6	Insulated wander-plug sockets.
5	Wander Plugs.
2oz.	Tobacco tin or similar container.

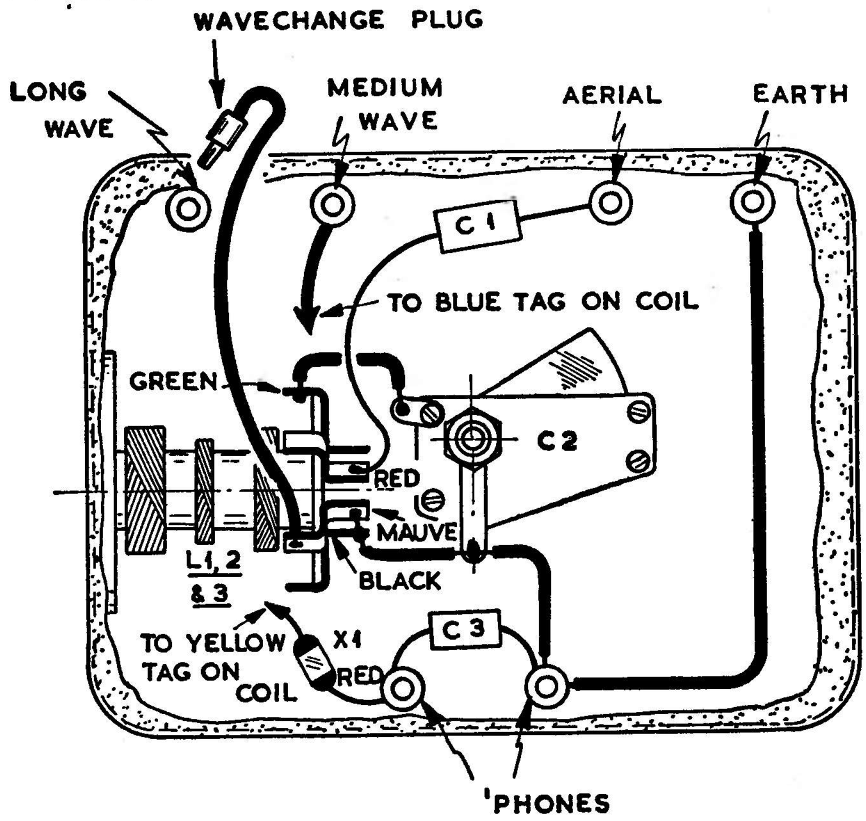

LID SHOWN BROKEN AWAY TO REVEAL POSITION OF COMPONENTS.

FIG. 24 b

10

In certain cases, signals from the local transmitter are too powerful to permit a crystal set with a normal coil to be of much use. I myself live in Hertfordshire a few miles from the London transmitters, and find a more elaborate circuit is necessary. Now in radio circles it is a well-known fact that the greater the number of tuned circuits, the greater the overall selectivity. Look at Fig. 25a, you will see that the usual L and C tuning arrangements have been duplicated. This type of circuit is known as a Band-Pass Filter. There are many kinds of such filters, this particular type is an inductively-coupled filter, note that energy from the first half (L2) is conveyed to L3 by way of two small inductors L1 and L4.

Tuning is quite sharp and the "feel" of the set is quite different from that of single coil designs. A small chassis is ideal for construction. 6"x4"x2½" was used on the original, though it could have been smaller.

Fig. 25b gives the layout and wiring.

Take particular notice that L1-2-3 is mounted horizontally and at right angles to L4-5-6, this is to prevent the coupling from being too "tight" which would ruin the idea of the filter. A good aerial and earth is of course necessary. It may be found on test that the selectivity is too great

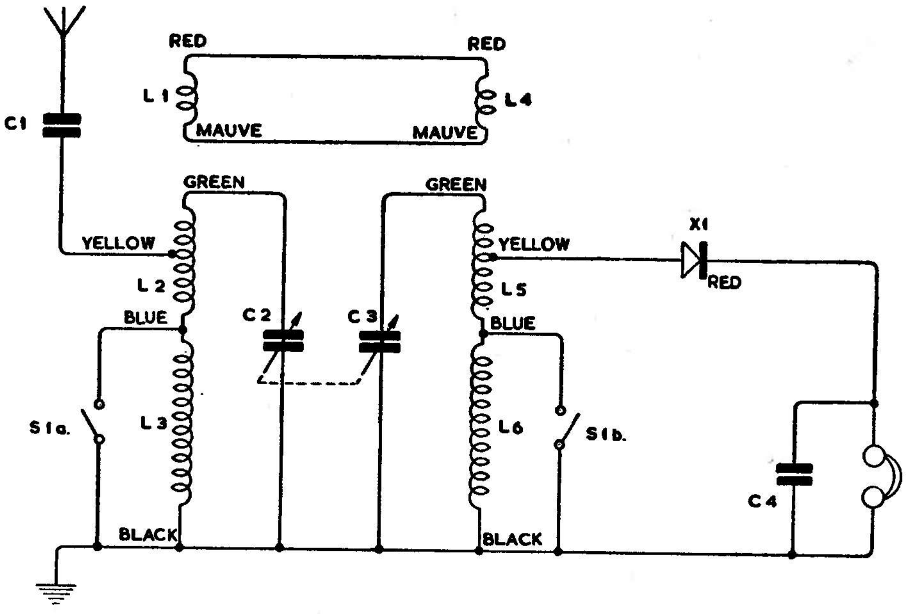

FIG. 25 a.

and that volume has suffered. This can be ingeniously overcome by connecting a very small capacitance between the two Green tags on the coils. The simplest way is to connect it across the fixed vane connections on C2-C3. Values for this extra capacitance are a matter for experiment usually a value between 10pF and 47pF is sufficient. A small 50pF trimmer could be used and adjusted to suit. The ideal setting is to enable powerful adjacent stations to be received just short of overlap so that as much volume as possible is obtained.

There is one point to be observed with the ganged capacitor. It should be of the type fitted with trimmers though of course, there is no reason why you should not fit them yourself. These are to allow for differences between coils to be balanced out, and consist of a small variable capacitance across each main section.

When setting up the receiver, first set each trimmer at half way and tune in a station near the high-frequency end of the band (vanes nearly out) adjust the trimmers for maximum volume. If extra coupling capacity is added as described, try adjusting them further.

Once properly set up, no further adjustment is necessary.

Components List, Fig. 25a

C1	100pF Mica capacitor.
C2-3	2 x 500pF variable capacitor (see text).
C4	1000pF Mica capacitor.
L1-2-3	Dual Range Coil R.E.P. (Blue Box).
L4-5-6	Dual Range Coil R.E.P. (Blue Box).
X1	Germanium Crystal.
S1a-b	2 pole single throw toggle switch.
4	Terminals or plugs and sockets.

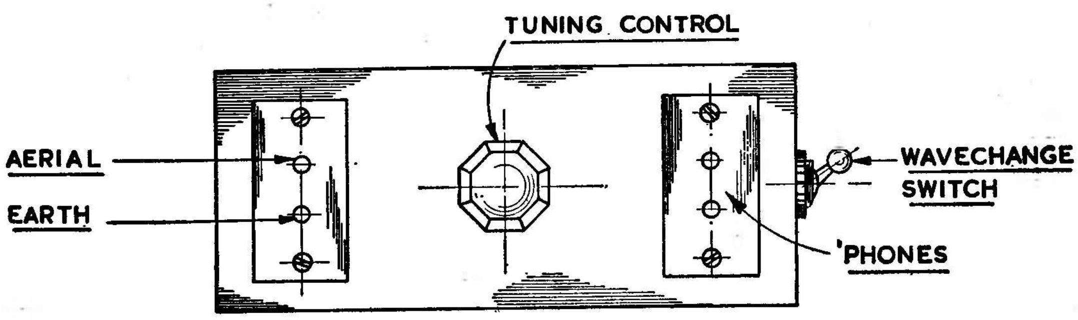
TUNING CONTROL
AERIAL
EARTH
WAVECHANGE SWITCH
'PHONES

FRONT OF COMPLETED RECEIVER.

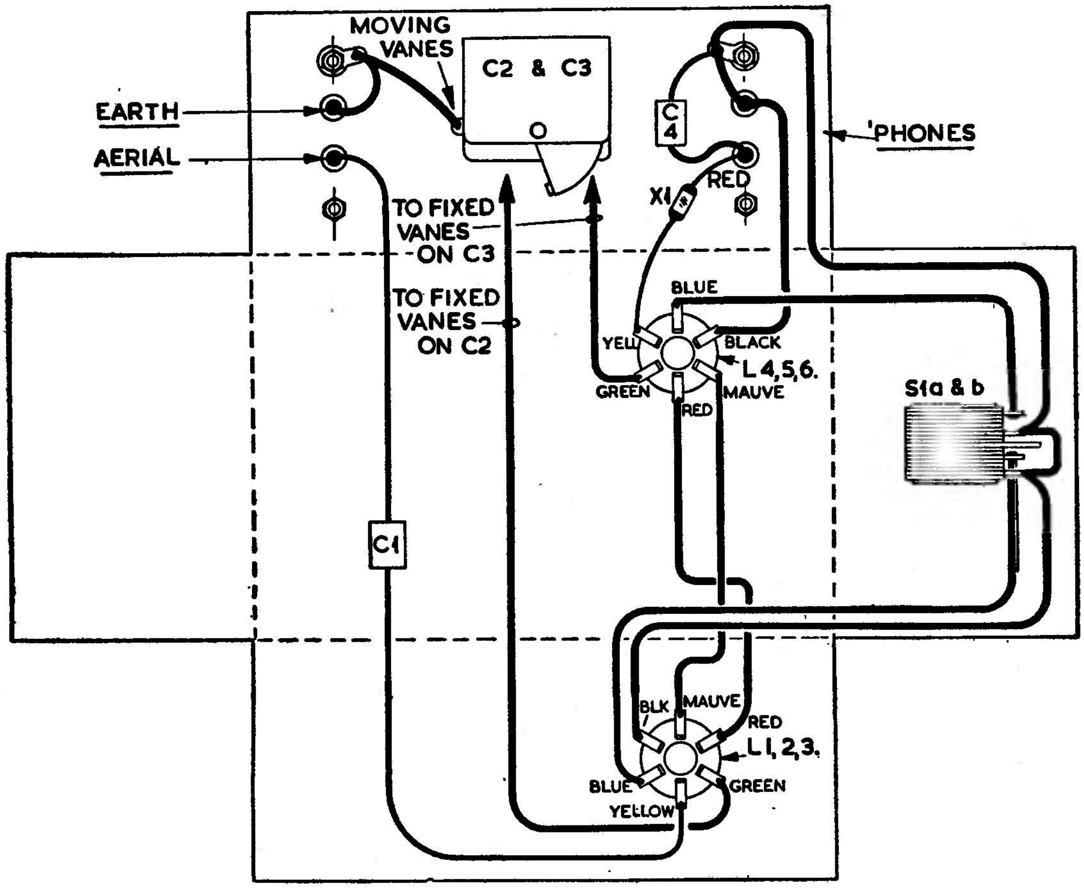
MOVING VANES
C2 & C3
EARTH
AERIAL
C 4
'PHONES
X1
RED
TO FIXED VANES ON C3
TO FIXED VANES ON C2
BLUE
YEL
BLACK
L 4,5,6.
GREEN
RED
MAUVE
S1a & b
C1
BLK
MAUVE
RED
L1,2,3.
BLUE
YELLOW
GREEN

EXPLODED UNDER-CHASSIS VIEW

FIG. 25 b

11

Another band-pass circuit is shown in Fig. 26a. This time home-made coils are used. This filter is not inductively coupled but entirely capacitive, relying on C3. As in the previous circuit it must be adjusted to give just sufficient " broadness " to the tuning to provide adequate volume without station overlap. To avoid unwanted coupling a screen is mounted between the coils, this is necessary because due to the size of the coils they would be bound to couple quite tightly without it. Such precautions were not necessary with the circuit of Fig. 25a, as the commercial coils are much smaller physically and the risk of unwanted coupling reduced.

The coils for this receiver are for medium wave reception only, this keeps the size down. Coil winding procedure has already been fully explained, and covers all the requirements of these coils. Base-board and panel construction is used, and for more accurate tuning since the coils are home-made, separate tuning capacitors are recommended. Fig. 26b gives the layout and practical wiring.

Components List, Fig. 26a

C1	100pF Mica capacitor.
C2	500pF variable capacitor.
C3	50pF Mica trimmer.
C4	500pF variable capacitor.
C5	1000pF mica capacitor.
L1-2	See text.
X1	Germanium Crystal.
4	Terminals or plugs and sockets.

For those who do not wish to bother with these a ganged capacitor can be used, but tuning may not be as good as with separate ones. The screen mounted between the coils is of aluminium or copper and should be about 1″ higher than the coils and the same length as the baseboard depth.

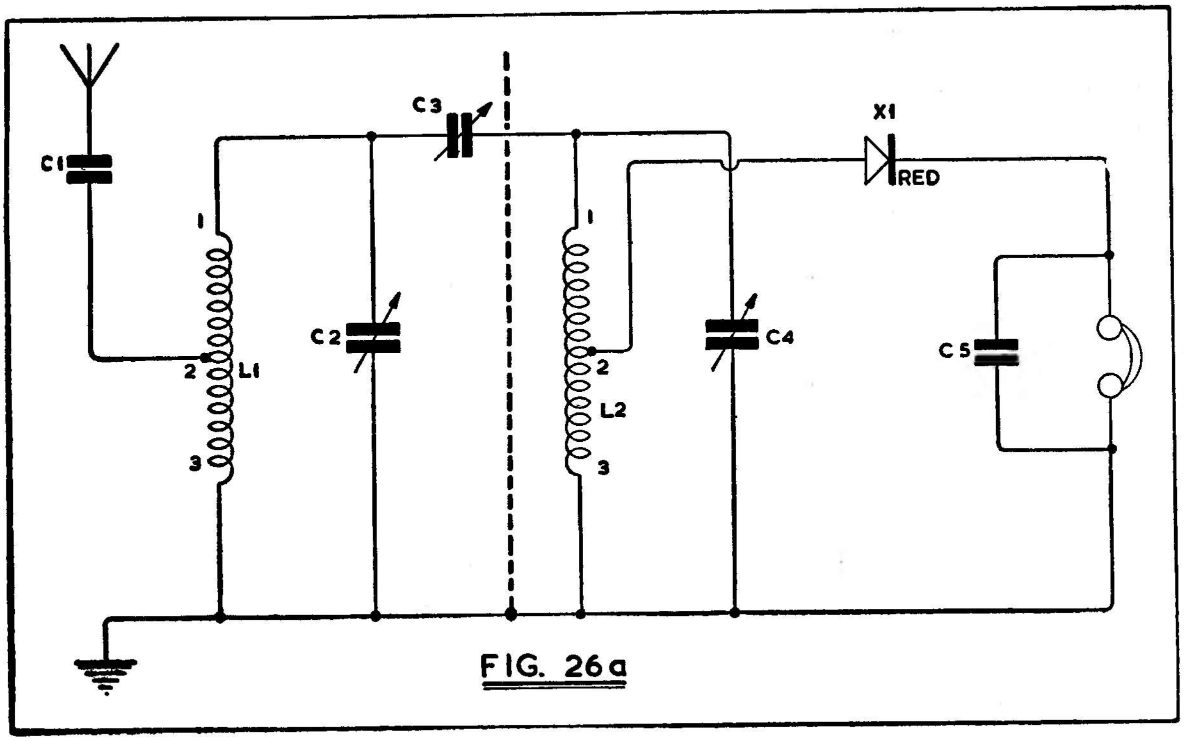

FIG. 26a

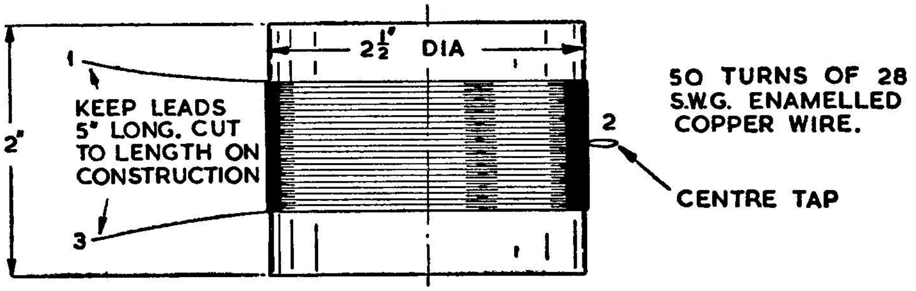
2½" DIA
2"
1
KEEP LEADS
5" LONG. CUT
TO LENGTH ON
CONSTRUCTION
3
2
50 TURNS OF 28
S.W.G. ENAMELLED
COPPER WIRE.
CENTRE TAP

COIL DETAILS.

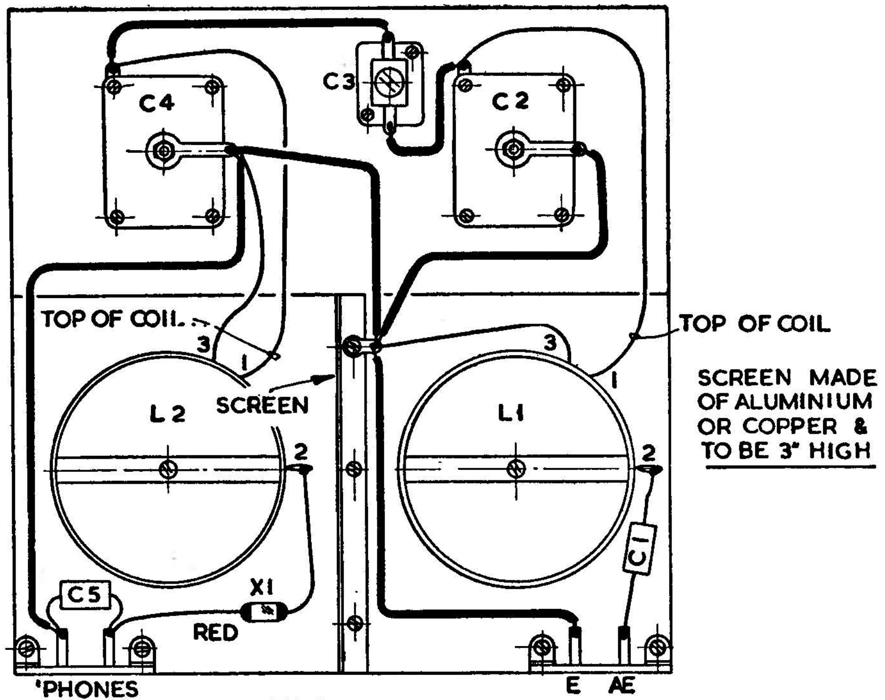
C4
C3
C2
TOP OF COIL
3
1
L2
SCREEN
2
L1
3
1
2
TOP OF COIL
SCREEN MADE
OF ALUMINIUM
OR COPPER &
TO BE 3" HIGH
C1
C5
X1
RED
'PHONES
E
AE

FIG. 26 b

12

Fig. 27a-b uses a band-pass filter for tuning, and the coils are home-constructed.

With this design a ganged capacitor is quite suitable as the coils tune quite accurately. It is essential that the coupling condenser C3 is **non-inductive,** there will be no difficulty in obtaining a new component of this pattern but it is as well to avoid old components which may not employ this form of construction.

Receivers using this type of filter are used to the best advantage when situated a short distance from a powerful transmitter and if selectivity is not important this particular design is not the most suitable.

Both medium-waves and long-waves are catered for: consider L1-2-3, this is the first half of the filter.

L1 is the aerial coupling coil, with S1 closed, the medium-waves are received. This coil is inductively coupled to L2 and L3. Again with S2 open L2 and L3 combined cover the long-waves and when S2 is closed the medium waves are covered.

Tuning for this half of the filter is by means of the capacitor C1, which is one half of the ganged capacitor. A screen is erected between the two sets of coils comprising the filter, and coupling is effected by C3, an earth return for the coils is provided by R1.

L4 and L5 operated in a similar manner to L2 and L3, tuning is by means of C2, the remaining section of the ganged capacitor. You will realise that the coils must be accurately wound otherwise tuning will not remain constant between the two sections over the band.

The ganged capacitor should be fitted with trimmer capacitors when purchased so that the two circuits can be balanced.

You will have noticed that this circuit uses three sets of switch contacts. A three-pole two-way wafer switch is the best choice otherwise either two or even three separate switches are necessary. All the necessary coil details are given in the diagrams, the tap on L4 is made in the same way as shown for other coils in the manual. One other component used only on this design, is the resistor R1, its size is not important because the current flowing is so small it can be ignored.

A ¼ watt rating is adequate, its precise value is not critical so that a 20% tolerance component is quite suitable.

Resistors are coded by colours and the 1000Ω type you require will bear the following sequence. At one end will be painted three coloured rings.

Components List, Fig. 27a

C1-2	500pF 2 gang, variable capacitor.
C3	0.05mfd. non-inductive paper capacitor.
C4	1000pF mica capacitor.
R1	1000 ohm resistor.
X1	Germanium Crystal.
S1-2-3	3-pole 2-way wafer switch.
L1-2-3 & L4-5	See text.
4	Terminals (Aerial, Earth and 'Phones).
	Sheet copper or aluminium for screen.

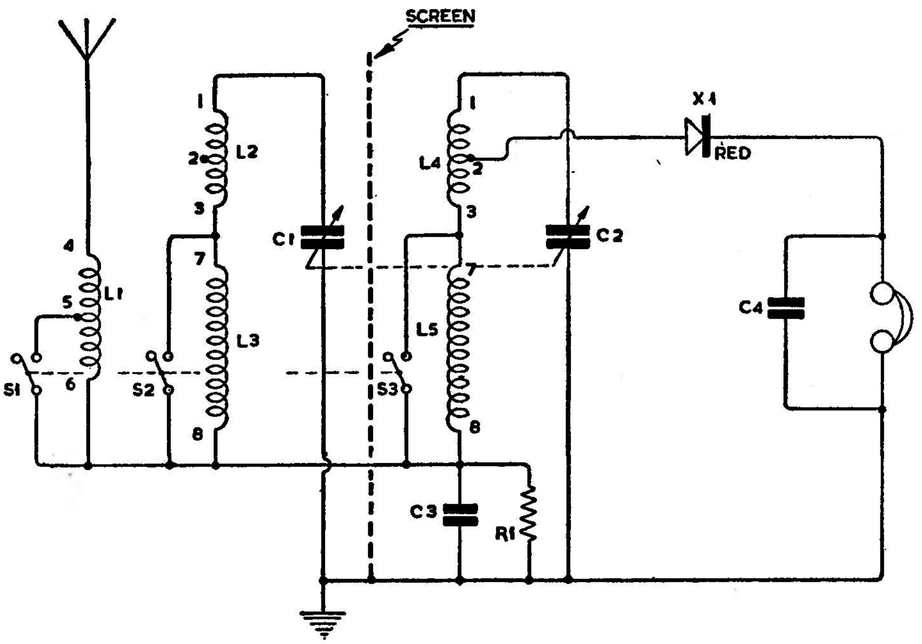

FIG. 27a

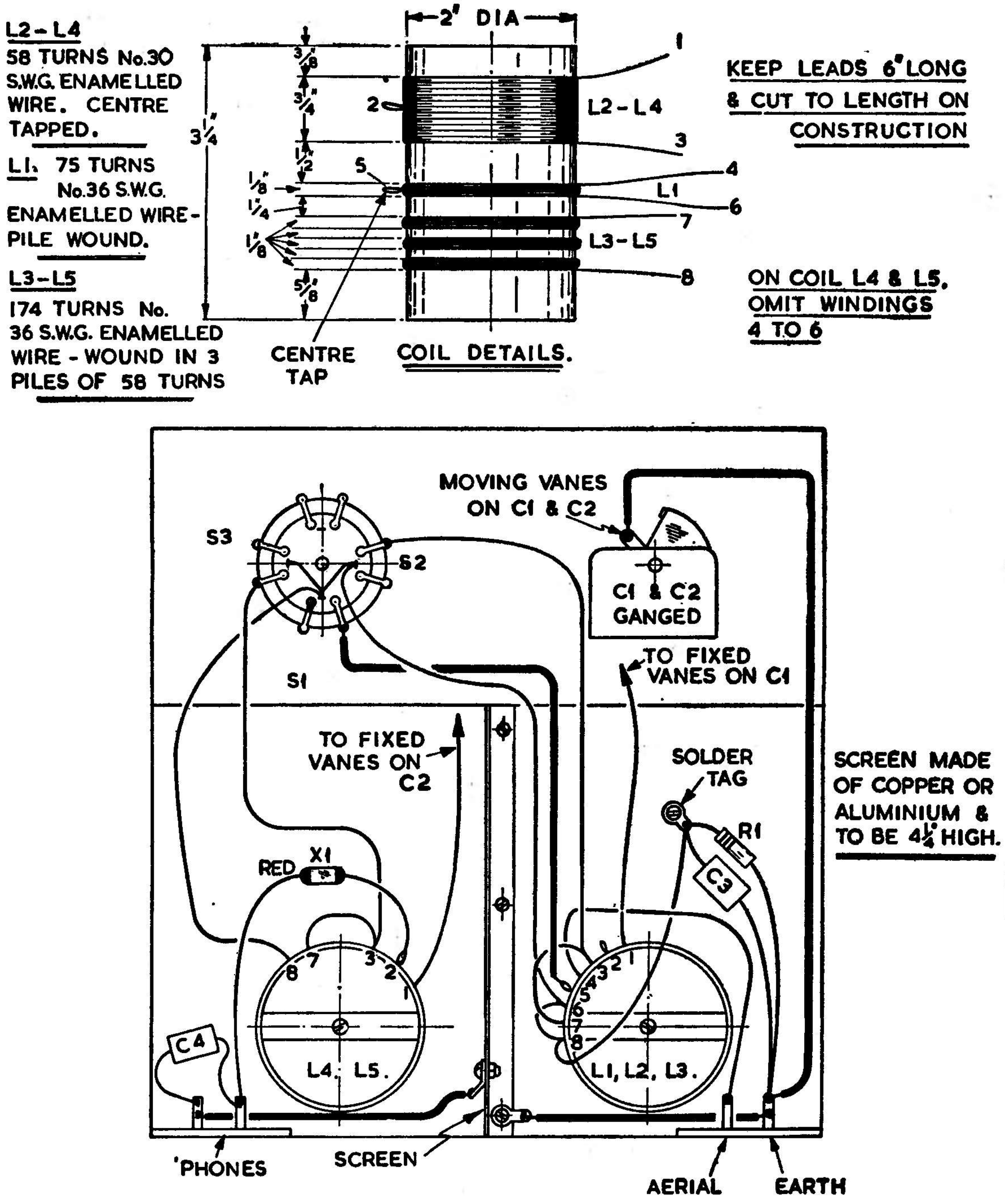
L2-L4
58 TURNS No.30 S.W.G. ENAMELLED WIRE. CENTRE TAPPED.
L1. 75 TURNS No.36 S.W.G. ENAMELLED WIRE-PILE WOUND.
L3-L5
174 TURNS No. 36 S.W.G. ENAMELLED WIRE - WOUND IN 3 PILES OF 58 TURNS
2" DIA
3/8"
3/4"
1/2"
1/8"
1/4"
1/8"
5/8"
3 1/4"
1
2
3
4
5
6
7
8
L2-L4
L1
L3-L5
CENTRE TAP
COIL DETAILS.
KEEP LEADS 6" LONG & CUT TO LENGTH ON CONSTRUCTION
ON COIL L4 & L5, OMIT WINDINGS 4 TO 6
MOVING VANES ON C1 & C2
S3
S2
S1
C1 & C2 GANGED
TO FIXED VANES ON C1
TO FIXED VANES ON C2
SOLDER TAG
R1
C3
SCREEN MADE OF COPPER OR ALUMINIUM & TO BE 4 1/4" HIGH.
RED
X1
C4
L4, L5.
L1, L2, L3.
'PHONES
SCREEN
AERIAL
EARTH

FIG. 27 b

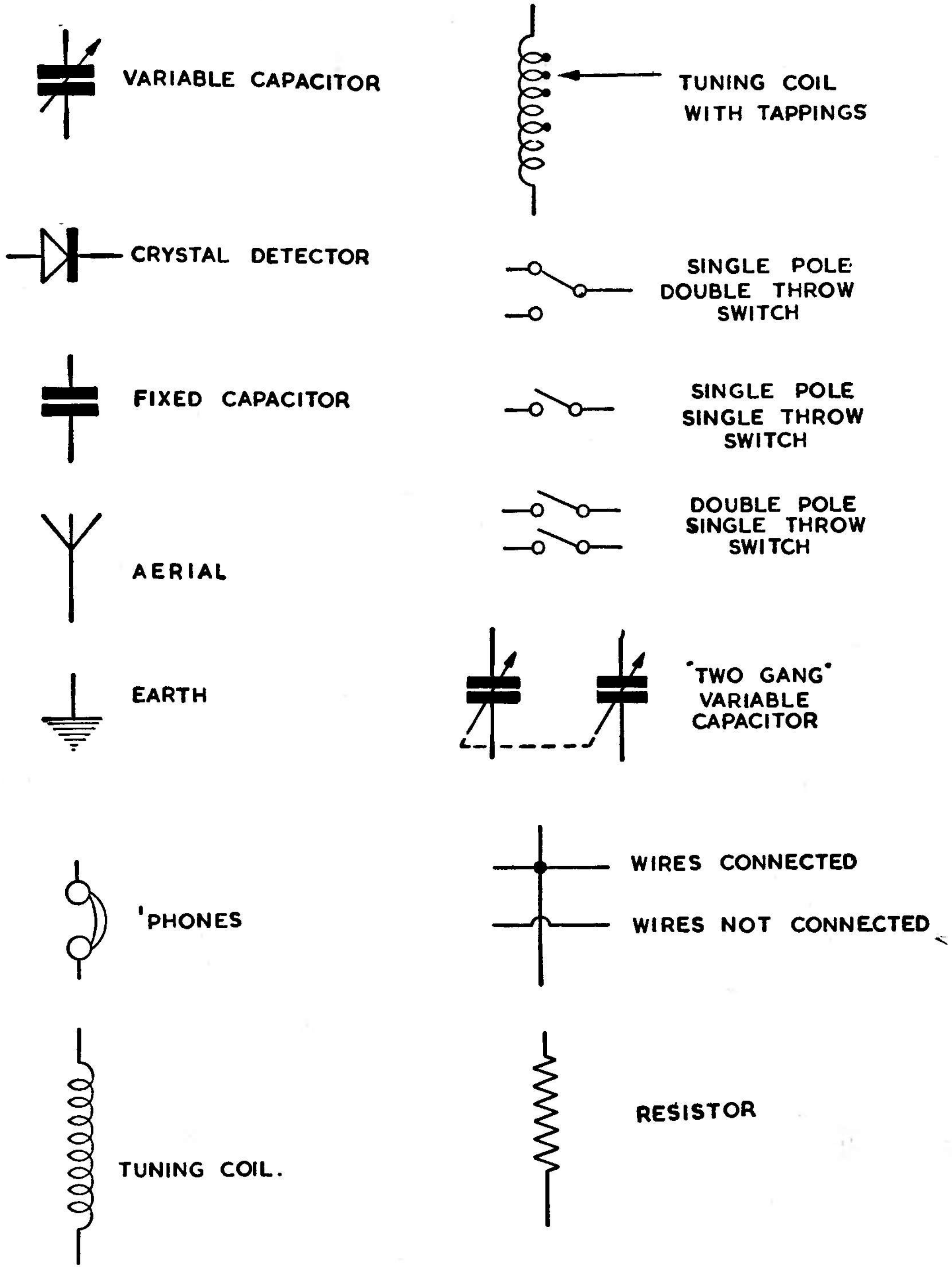
VARIABLE CAPACITOR
TUNING COIL WITH TAPPINGS
CRYSTAL DETECTOR
SINGLE POLE DOUBLE THROW SWITCH
FIXED CAPACITOR
SINGLE POLE SINGLE THROW SWITCH
DOUBLE POLE SINGLE THROW SWITCH
AERIAL
EARTH
'TWO GANG' VARIABLE CAPACITOR
WIRES CONNECTED
'PHONES
WIRES NOT CONNECTED
RESISTOR
TUNING COIL.

SYMBOLS USED IN THE CIRCUITS DESCRIBED.

FIG. 28

The first one is BROWN which represents 1, the first figure of the value, the second BLACK, this indicates the second numeral is nought, and the third colour is RED which tells us that two further ciphers or noughts are added to give the total value. From this you can see, we have 1 plus a nought equalling 10, plus two further noughts which total 1000. Some earlier types of resistor were painted differently, the whole body colour represented the first figure, one tip was coloured to represent the second figure, and a painted dot in the centre gave the number of ciphers to be added.

There are no other special points regarding construction except perhaps that if a long aerial is used, a small 100pF capacitor might be tried in the aerial lead to prevent damping the first tuned circuit.

When setting up the receiver, tune to a station at the high-frequency end of the M.W. band and set the trimmers for loudest volume, no other adjustments are necessary.

CONCLUSION

The twelve receivers described are representative of the best crystal set designs available to-day. Their construction will provide many hours of useful enjoyment, and the results will give lasting pleasure.

A final word about components, if no actual maker is specified, any good class component can be used. Switches, fixed and variable capacitors and crystals, are available from numerous manufacturers all of which are invariably of excellent quality. Where a particular manufacturer's product is called for, the specification should be adhered to. The commercial coils specified are freely available from most supply houses specialising in components for constructors, but in case of difficulty write to Bernards (Publishers) Ltd., The Grampians, Western Gate, London, W.6, who will, on receipt of a S.A.E. be pleased to supply the address of your nearest stockist. Coil formers are not always easy to obtain and in this case advice should be sought from Post Radio Supplies, 395, Queensbridge Road, London, E.8, who will be pleased to help in any way possible.

* *

The HF Current Probe: Theory and Application

KENNETH WYATT
Wyatt Technical Services
Woodland Park, Colorado, USA

This article describes one of the most valuable tools in the EMC engineers "bag of tricks" – the high-frequency current probe. Current probes are invaluable for measuring high-frequency common-mode (or "antenna") currents flowing on wires or cables. Experience has proven that poorly terminated (bonded or filtered) cables are the number-one cause for radiated emissions failures at a test facility. By measuring the common-mode (CM) currents (sometimes referred to as "antenna" currents) on these cables it's possible to troubleshoot and apply fixes to a product right there in your development lab. You can also predict, to a good degree of accuracy, whether a given cable current will pass or fail in the measurement chamber. This will save you tons of time trying to apply fixes at the test facility while the clock is ticking away your test time. I'll also show you several ways to create do-it-yourself (DIY) probes that are quick to make and very useful in a pinch.

COMMON-MODE CURRENTS

Let's consider CM currents and how they are generated, because it is not intuitive as to how current may travel the same direction through both the signal and signal-return wires in a cable or PC board. Referring to Figure 1, note that due to finite impedance in any grounding system (including circuit board signal/power return planes), there will be a voltage difference between any two points within that return plane. This is denoted by V_{GND1} and V_{GND2} in the figure. This difference in potential will drive CM currents through common cabling or circuit traces between circuits or sub-systems. In addition, unbalanced geometries - for example, different lengths or path routings for high-speed differential pairs - can create voltage sources that drive associated CM currents. Finally, routing a high-speed clock trace across a split in the return plane or referencing it to multiple planes, can also be a source of CM current. Because the current phasors in Figure 1 are additive, the resulting radiated phasor may be quite large compared to those generated by differential-mode (DM), or signal currents, which are opposite in direction, and so tend to cancel. Therefore, CM emissions tend to be more of an issue than DM emissions.

CURRENT PROBES: THEORY OF OPERATION

The RF current probe is an "inserted-primary" type of radio frequency current transformer. When the probe is clamped over the conductor or cable in which current is to be measured, the conductor forms the primary winding. The clamp-on feature of this probe enables easy placement around any conductor or cable. This is essentially a broadband high-frequency transformer. High-frequency currents can

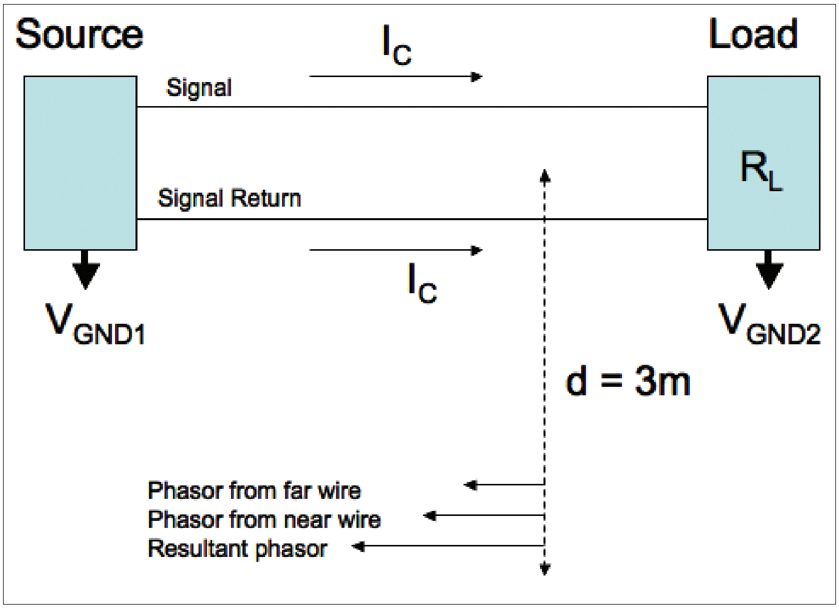

Figure 1. *Common-mode currents in a circuit loop. The source is a digital signal (with harmonics) and we'll assume a resistive load. Because the phasor current in the far wire is in the same direction as the phasor current in the near wire, the resultant phasor is relatively large compared to that produced by differential-mode current phasors. In this case, lowering the harmonic content (by slowing the digital rise/fall-times) or diverting/blocking the CM current is very important in limiting radiated emissions.*

be measured in cables without physically disturbing the circuit.

Since the current probe is intended for "clamp-on" operation, the primary shown in Figure 2 is actually the electrical conductor in which CM currents are to be measured. This primary is considered as one turn since it is assumed that the CM currents flow through the conductor and return to the source via a return conductor such as a frame, common ground plane, or earth. On some current probe models the secondary output terminals are resistively loaded internally to provide substantially constant transfer impedance over a wider frequency range.

COMMERCIAL CURRENT PROBES

While commercial current probes are pricey, the advantage is that they can open up and snap around a cable, rather than having to be threaded onto the cable to be measured. See Figure 3. They are also a lot more rugged and can take a lot of abuse as compared to the "do-it-yourself" (DIY) versions below. Finally, they are also accurately characterized, allowing very precise measurements of cable currents.

DIY CURRENT PROBES

In a pinch, you can make your own current probe. Examples of several DIY probes are shown in Figures 4 and 5. I typically try to find a ferrite toroid or clamp-on core that offers good high-frequency characteristics in the 10 to 1000 MHz range. Winding a few (not too critical) turns and terminating with a coax connector is all you need. Keeping the turns as far apart as possible (as in Figure 4) will reduce inter-winding capacitance and yield better results at the higher frequencies. This is one of the largest drawbacks in performance of the clamp-on ferrites (as in Figure 5).

TRANSFER IMPEDANCE

The CM current (Ic) in microamps in the conductor under test is determined from the reading of the current probe output (V) in microvolts divided by the current probe transfer impedance (Zт).

$$I_C = V/Z_T \qquad (1)$$

Or, in dB

$$I_C(\mathrm{dBuA}) = V(\mathrm{dBuV}) - Z_T(\mathrm{dB}\Omega) \qquad (2)$$

The typical transfer impedance of the current probe throughout the frequency range is determined by passing a known RF current (Ic) through the primary test conductor and noting the voltage (V) developed across a 50-Ohm load. Then,

$$Z_T = V/I_C \text{ (in standard units)} \qquad (3)$$

Or

$$Z_T(\mathrm{dB}\Omega) = V(\mathrm{dBuV}) - I_C(\mathrm{dBuA}) \qquad (4)$$

The Fischer F-33-1 probe is a commonly used troubleshooting tool and has a flat frequency response from 2 to 250 MHz (Figure 6). The transfer impedance is about 5Ω (approximately +14 dBΩ on the graph), therefore, a 1 uA current will produce a 5 uV output voltage from the current probe.

PROBE CALIBRATION

The accurate calibration of RF current probes is a complex process. Characterization is a more correct term to use than calibration. The probe must be properly characterized to reflect how the user uses the probe. Probe manufacturers usually sell a calibration fixture that attempts to maintain a 50Ω impedance. A 50Ω load is connected to the output port and a calibrated RF generator (or network analyzer) is connected to the input port. The probe to be characterized is clamped around the fixture and the frequency is swept while measuring

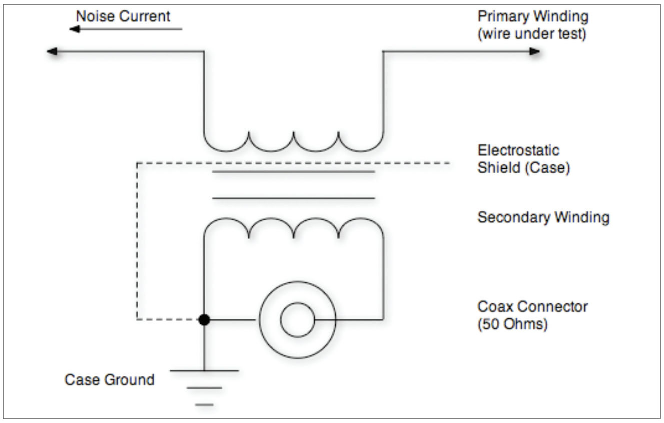

Figure 2. *The basic current probe (high-frequency current transformer).*

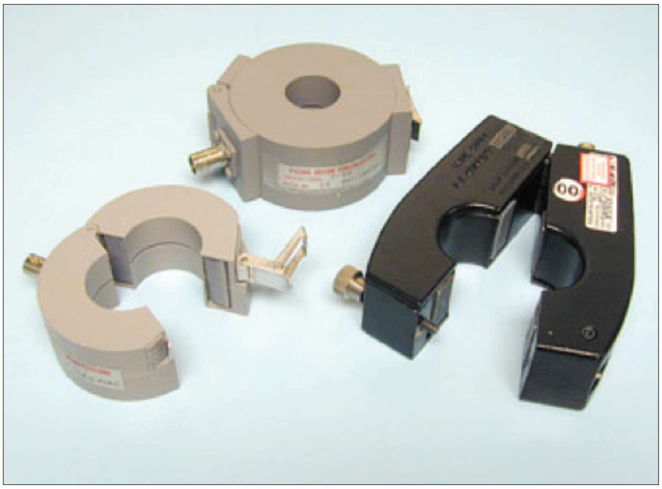

Figure 3. *Examples of commercial current probes.*

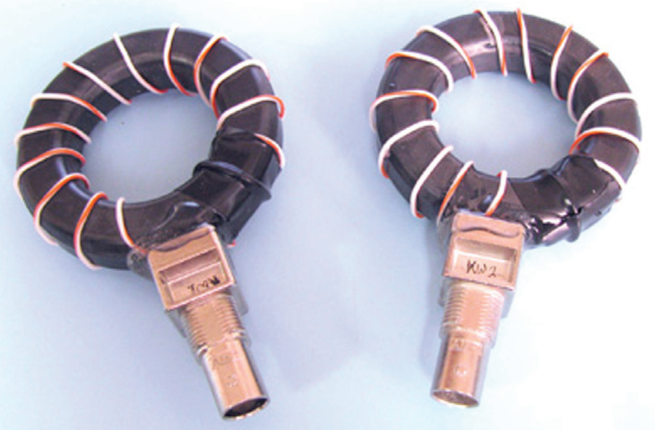

Figure 4. *Examples of DIY current probes based on a large toroid core. These photos were taken prior to installing the E-field shield which consists of a layer of copper tape over the windings, leaving a small gap around the inside of the toroid. 14 turns of Teflon-insulated wire wound around a Würth Electronik #74270097 ferrite core (4W620 material) was used, which is useful from 10 to 1000 MHz.*

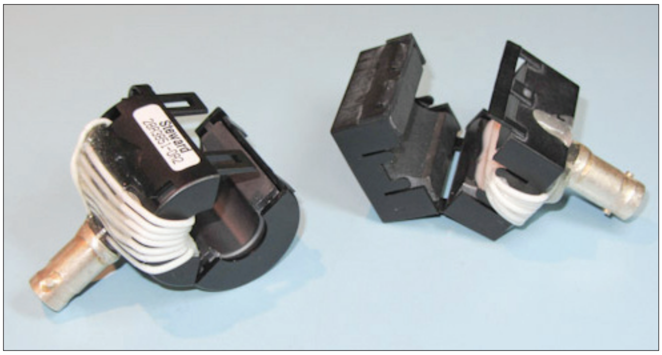

Figure 5. *Examples of DIY current probes based on clamp-on ferrite chokes. I used a couple sample Steward (now a unit of Laird Technologies) chokes – a round one (model 28A3851-0A2) and a square one (model 28A2024-0A2). They each had 7 turns of Teflon-insulated wire wound around one-half and glued down on the inside to hold the windings. I later epoxied a PC board-style BNC connector to the outside, making sure there was enough epoxy to hold the outer turns together. Type 28 material was used, which is useful from 10 to 1000 MHz.*

the probe output.

My test setup was a little more rudimentary (Figure 7), but for troubleshooting purposes, it's good enough. I used a short piece of stiff wire across the output port with a 50Ω resistive load in series. I then adjusted the generator for zero dBm – a convenient amount. This is equivalent to an output voltage of 224 mV (or 73 dBuA of current) into 50Ω. The actual generator output doesn't matter, so long as the resulting probe voltage is large enough to be seen readily in the receiver or spectrum analyzer. I monitored the probe output with a Thurlby Thander TTi PSA2701T handheld spectrum analyzer.

Knowing the current through the wire in dBuA and the probe output in dBuV, the transfer impedance may be plotted graphically by subtracting: $V(dBuV) - I_C(dBuA)$ (expressed in dB). In this case, $Z_T(dB\Omega) = V(dBuV) - 73$. While this may be useful for educational purposes, I wouldn't be too inclined to use the DIY probes to predict "pass/fail", as described further down. However, because they compare favorably to the commercial probes as far as output voltage, I believe (and have proven in practice) that they are completely suited for troubleshooting. You only need to know whether an EMC design fix made the cable current better or worse.

PREDICTING PASS/FAIL

It is possible to predict whether a particular cable will pass or fail radiated emissions by measuring the CM current at the offending frequency, reading off the transfer impedance of the probe, Zt (dBΩ) in Figure 6, and solving for I_C (using Equation 2 above). Plugging I_C(Amps) into Equation 5 will calculate the E-field level in V/m. The length of the cable is L(m) and the offending harmonic frequency is f(Hz). Use a test distance, d, of either 3 or 10m to predict the outcome at those test distances.

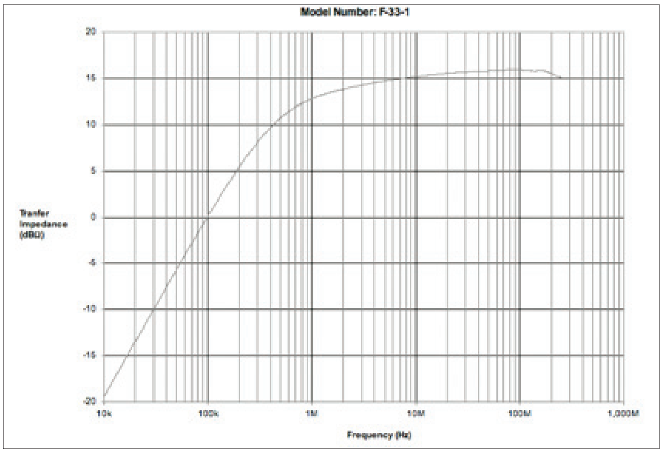

Figure 6. *Transfer impedance (ZT) graph of an F-33-1 current probe (courtesy of Fischer Custom Communications). The x-axis is frequency, while the y-axis is dBΩ. Use this to calculate the value of I_C (Equation 2), given the measured voltage at the probe terminals (V_{dBuV}) and Z_T.*

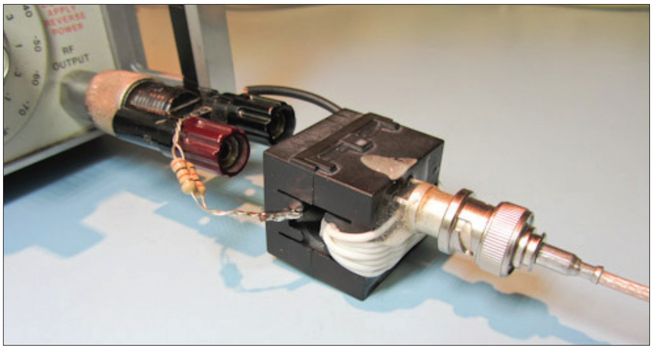

Figure 7. *I used a short wire and 50Ω load (two parallel 100Ω resistors) across the generator output for probe characterization. Obviously, there are shortcomings at higher frequencies, due to the inductance of the wire. In fact, the system impedance starts to go capacitive at 100 MHz and it's difficult to keep a fixed 224 mV across the load resistor with frequency.*

$$\left|\hat{E}_{C,\max}\right| = 1.257 \times 10^{-6} \frac{\left|\hat{I}_C\right| fL}{d} \tag{5}$$

Once you've determined a particular cable has CM currents that may cause a RE failure, you should to examine the connector where the cable is attached to the product enclosure. Very often, I find poor or non-existent bonding between the connector shield and enclosure shield. These points must be bonded well to permit the CM currents to flow back to their source within the product, avoiding associated cable radiation. Please refer to my previous articles on troubleshooting radiated emissions for more information (references below).

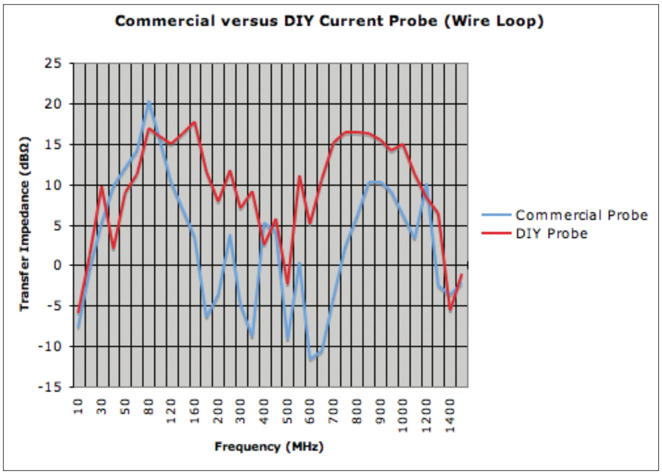

Figure 8 *Transfer impedance (ZT) graph of a commercial current probe versus the DIY toroidal probe. The x-axis is frequency, while the y-axis is dBΩ. Note that the commercial probe is only designed and characterized to 250 MHz, so the data above that, while interesting, is probably not valid. The DIY probe, as well, performs poorly above 200 MHz and frankly, the wire loop used to introduce a "calibrated" current (while as short as possible) affects the measurement, as well.*

REAL-WORLD TROUBLESHOOTING EXAMPLE

As previously mentioned, one of the most common sources of radiated emissions is due to poorly bonded connectors mounted on shielded product enclosures. This occurs especially if the connectors are circuit board mounted and penetrate loosely through the shielded enclosure. Poorly bonded connectors allow internally generated CM currents to leak out and flow on the outside of I/O, mouse or keyboard cables. This will also allow ESD discharges inside the product – more bad news. If these currents are allowed out of the enclosure, the attached cables will act as radiating antennas – often resonating around 300 MHz, due to their typical 1m length.

This was the case for a new digitizing oscilloscope prototype I worked on recently. The I/O connectors were all soldered onto the PC board and the board was fastened to the rear half of the enclosure. The connectors simply poked up through cutouts in the rear metal shield.

While using a current probe to measure the CM current flowing on the outside of the USB cable under test, I simply jammed the screwdriver blade of my Swiss Army knife between the connector bonding fingers and metal chassis enclosure and was able to drop the overall cable currents by 10 to 15 dB.

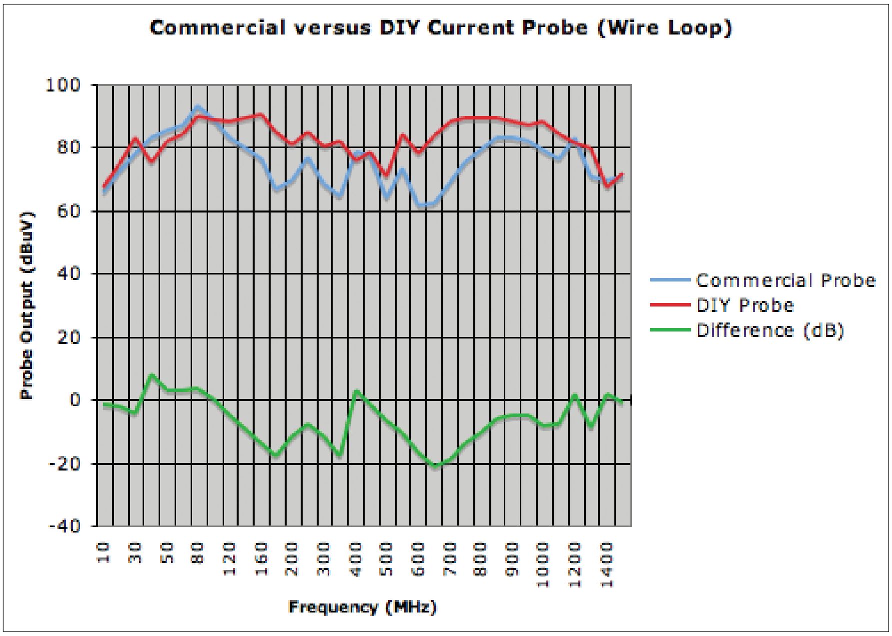

Figure 9. *Probe output voltage (V_{out}) graph of a commercial current probe versus the DIY toroidal probe. The x-axis is frequency, while the y-axis is dBuV. This shows that the probes are very comparable in output voltage versus frequency. For troubleshooting purposes, absolute accuracy is not required - just consistency in measurements. All one really needs to know is, "did the fix I implemented make the CM current go up or down?" The DIY probe works well for this.*

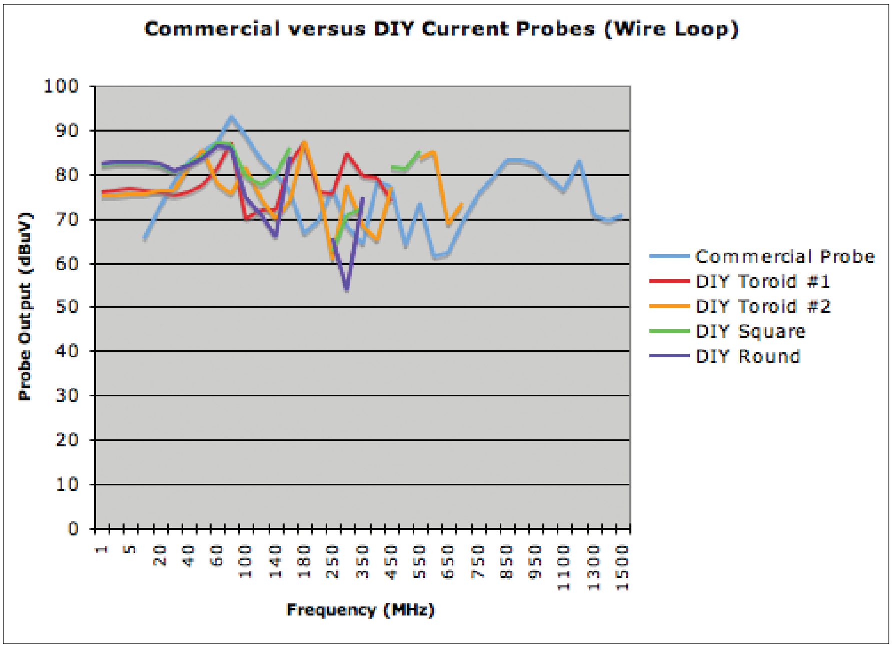

Figure 10. *Probe output voltage (V_{out}) graph of a commercial current probe versus two DIY toroidal probes and two different clamp-on probes. The x-axis is frequency, while the y-axis is dBuV. This shows that all these probes are very comparable in output voltage versus frequency and therefore, useful for troubleshooting purposes. Just don't try using the DIY probes to determine "pass or fail" predictions. Commercial probes are better-suited for that.*

The solution was to fabricate a custom shim with spring-fingers that would slip over all the connectors creating a firm bond between the connector ground shell and inside of the shielded enclosure. More and more low-cost products are relying on PC board mounted I/O connectors as a cost-cutting measure. Any time you see this, be prepared to carefully examine the bonding between the connector ground shell and the shielded enclosure.

TROUBLESHOOTING TIPS USING CURRENT PROBES

Here are a few troubleshooting tips using current probes.

1. When evaluating the harmonics on a cable by using a current probe, if sliding the probe back and forth changes the harmonic levels, part of the coupling may be near-field, rather than conducted.

2. When using a pair of current probes; one on each of two cables, if the harmonics are the same in each, the source is in the middle. If one cable has stronger harmonics, then you'll want to work on that side first. See Figure 12 below.

3. Measuring the currents on two suspect legs of a dipole should read the same. Placing the two suspect legs through the same current probe should cause a big decrease due to current cancellation. See Figure 12 below.

4. When measuring video cable currents and large cable movements cause big changes in amplitude, the coupling is likely inductive - otherwise, it's more likely conductive.

5. If you suspect inductive coupling, the phase at the victim will be 180-degrees from the source. This may be observed on an oscilloscope with H-field probes or current probes. Try syncing the scope trigger at the source using a scope probe.

My colleague, Doug Smith, has many more examples on how to use current probes for measuring cable and PC board resonances, injecting pulses for troubleshooting, interpreting the relative phase of common-mode currents and troubleshooting ESD issues. Refer to the references below.

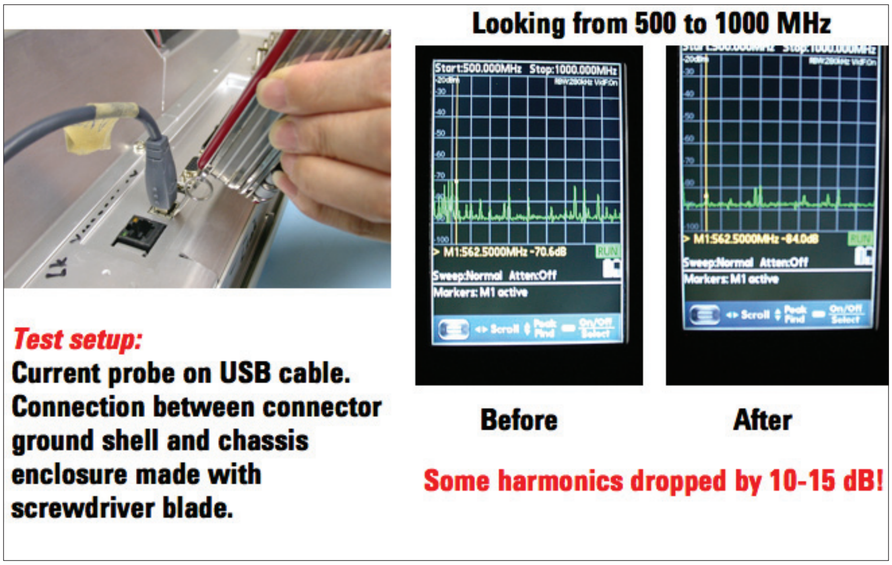
Looking from 500 to 1000 MHz
Start:500.000MHz Stop:1000.000MHz
M1:562.5000MHz -70.6dB
Sweep:Normal Atten:Off
Markers: M1 active
M1:562.5000MHz -84.0dB
Sweep:Normal Atten:Off
Markers: M1 active
Scroll
Peak Find
On/Off Select
RUN
Test setup:
Current probe on USB cable.
Connection between connector ground shell and chassis enclosure made with screwdriver blade.
Before
After
Some harmonics dropped by 10-15 dB!

Figure 11. *Cables should be tested individually. Here, I have a current probe clamped around the cable under test and am monitoring the harmonics with a simple hand-held spectrum analyzer. As I ground the connector shell to the chassis with the Swiss Army screwdriver blade, the harmonics were reduced 10-15 dB!*

SUMMARY

Use of a current probe is vital during the troubleshooting process. Poorly bonded cable connectors can be readily identified and fixed. The radiated E-field from a product I/O cable may be calculated by measuring the high-frequency common-mode currents flowing in the cable. All this may be performed right at the designer's workbench and without the expense of a third-party test facility or shielded chamber.

REFERENCES - PAPERS

- [1] Mat Aschenberg & Charles

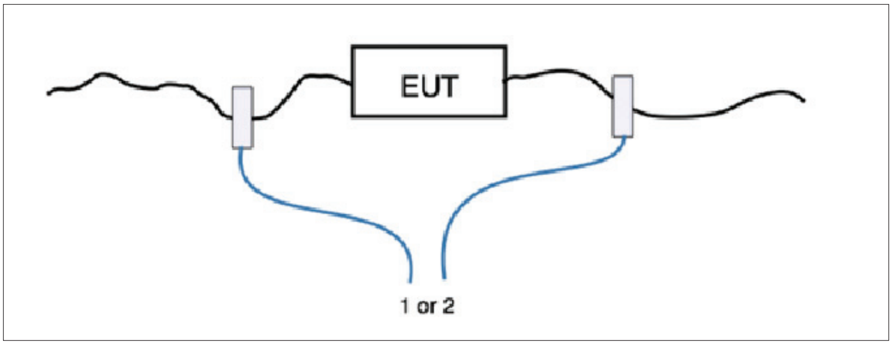

Figure 12. *When measuring two cables from a system and the harmonic currents are approximately the same (point 1 is the same as point 2), the source is at the center (the EUT) and the two cables are acting as a dipole antenna. You may notice a peak in harmonic strength at the half-have length of the two cables combined. If the harmonic currents are larger in one side or the other, then you'll want to troubleshoot just that cable.*

Grasso, Radiation from Common-Mode Currents – Beyond 1 GHz (Three Methods Compared)

- [2] Dave Eckhardt, Homebrew Clamp-On Current Probe, private correspondence (January 2009), Email: davearea51@wildblue.net.
- [3] Jasper Goodblood, Electromagnetic Compatibility, 1990, Prentice Hall, pages 31-34.
- [4] Michel Mardiguian, EMI Troubleshooting Techniques, McGraw-Hill, 2000, pages 39-49.
- [5] Montrose & Nakauchi, Testing for EMC Compliance, 2004, Wiley Interscience, pages 116-124, 143-145, and 159-161.
- [6] Henry Ott, Electromagnetic Compatibility Engineering, Wiley, 2009, pages 690-693.
- [7] Henry Ott, Measuring Common-Mode Currents on Cables, www.hottconsultants.com/techtips/tips-cm.html
- [8] Clayton Paul, Introduction to Electromagnetic Compatibility (2nd Edition), Wiley Interscience, 2006, pages 518-532.
- [9] Ridao, Carrasco, Galvin and Franquelo, Implementation of low cost current probes for conducted EMI interference measure in Power Systems, EPE 1999 (Lausanne).
- [10] H. Ward Silver, Hands-On Radio column, Detecting RF – Part 2, QST, August 2011, page 54-55.
- [11] Doug Smith, Current Probes, More Useful Than You Think, IEEE EMC Symposium 1998, http://emcesd.com/pdf/iprobe98.pdf.
- [12] Doug Smith, The Two Current Probe Puzzle, http://emcesd.com/tt061999.htm.
- [13] Doug Smith, Using Current Probes to Inject Pulses for Troubleshooting, http://emcesd.com/tt2007/tt120307.htm Part 1, http://emcesd.com/tt2009/tt030309.htm Part 2.
- [14] Doug Smith, Using Current Probes to Measure Cable Resonance, http://emcesd.com/tt2008/tt010108.htm.
- [15] Doug Smith, Measuring and Interpreting the Relative Phase of Common Mode Currents, http://emcesd.com/tt2008/tt030208.htm.
- [16] Doug Smith, Using a Comb Generator with a Pair of Current Probes to Measure Cable Resonance, http://emcesd.com/tt2009/tt110709.htm.
- [17] Doug Smith, Using Current Probes to Inject Pulses for Troubleshooting (Board Resonances), http://emcesd.com/tt2010/tt010110.htm.
- [18] Doug Smith, Predicting Cable Emissions from Common Mode Current, http://emcesd.com/tt2006/tt030106.htm.
- [19] Douglas Smith, High Frequency Measurements and Noise in Electronic Circuits, Van Nostrand Reinhold, 1993, pages 41-44, 159-182, 192-209.
- [20] Allen Wolff, Building a Ferrite Core Antenna Current Probe, Technical Correspondence, QST, August 2009, page 53.
- [21] Kenneth Wyatt, Troubleshooting Radiated Emissions Using Low-Cost Bench-Top Methods, Interference Technology (ITEM) – EMC Directory & Design Guide 2011, May 2011, page 10-21, http://www.interferencetechnology.com/uploads/media/Wyatt_DDG11.pdf.
- [22] Kenneth Wyatt, Troubleshooting Radiated Emissions – Three Case Studies, InCompliance Magazine, October 2011, http://www.incompliancemag.com/index.php?option=com_content&view=article&id=818:troubleshooting-radiated-emissions-three-case-studies&catid=27:testing&Itemid=136.

REFERENCES - SUPPLIERS

- [23] Fischer Custom Communications (FCC), Phone: (310) 303-3300, Email: sales@fischercc.com, Web: www.fischercc.com. They provide a very wide range of HF current probes – their specialty.
- [24] Laird Technologies, Web: http://www.lairdtech.com/Products/EMI-Solutions/. They offer a complete line of ferrite cores and chokes.
- [25] Pearson Electronics, Phone: (650) 494-6444, Email: sales@pearsonelectronics.com, Web: www.pearsonelectronics.com. They have a good selection of probes.
- [26] Rhode & Schwartz USA, Phone: (888) 837-8772, Email: info@rohde-schwarz.com, Web: www.rohde-schwartz.us. They have a very limited selection.
- [27] Solar Electronics, Phone: (800) 952-5302, Email: sales@solar-emc.com, Web: www.solar-emc.com. They have a limited selection.
- [28] Teseq USA, Phone: (732) 417-0501, Email: usasales@teseq.com, Web: www.teseq.us. They have a very limited selection.
- [29] Thurlby Thander Instruments, Phone: +44-1480-412451, Email: sales@tti-test.com, Web: http://www.tti-test.com/contact-tti.htm. They offer a low-cost handheld spectrum analyzer for under $2,000 USD.
- [30] Würth Electronics Midcom, (605) 886-4385, midcom@we-online.com, www.we-online.com. I used one of their large ferrite cores for my DIY current probe.

Kenneth Wyatt, Sr. *EMC Engineer, Wyatt Technical Services LLC, holds degrees in biology and electronic engineering and has worked as a senior EMC engineer for Hewlett-Packard and Agilent Technologies for 21 years. He also worked as a product development engineer for 10 years at various aerospace firms on projects ranging from DC-DC power converters to RF and microwave systems for shipboard and space systems. A prolific author and presenter, he has written or presented topics including RF amplifier design, RF network analysis software, EMC design of products and EMC troubleshooting techniques. He has been published in magazines such as, RF Design, EMC Design & Test, Electronic Design, InCompliance, Interference Technology, Microwave Journal, HP Journal and several others.*

Wyatt is a senior member of the IEEE and a long time member of the EMC Society where he serves as their official photographer. He is also a member of the dB Society and is a licensed amateur radio operator. Contact Wyatt at ken@emc-seminars.com. His website is www.emc-seminars.com.

No. 96

CRYSTAL SET CONSTRUCTION

SIMPLE INSTRUCTIONS FOR MAKING NINE CRYSTAL SETS

by B. B. BABANI

There is to-day a marked revival of interest in crystal receivers, and this book has been specially prepared to introduce the home constructor to the design and building of the various types. The text is simple and free from technicalities, so that even the young reader may confidently embark on the construction of any of the receivers shown.

CONTENTS

1. AERIALS.
2. EARTHS.
3. CRYSTAL DETECTORS.
4. COILS.
5. THE BEGINNERS' CRYSTAL RECEIVER.
6. FRAME AERIAL RECEIVER.
7. LONG DISTANCE RECEIVER.
8. HIGH GAIN RECEIVER.
9. HIGH SELECTIVITY RECEIVER.
10. BATTERY AIDED CRYSTAL RECEIVER.
11. MEDIUM AND LONG BAND RECEIVER.
12. ULTRA SENSITIVE ALL WAVE RECEIVER.
13. AN ADVANCED CRYSTAL DIODE RECEIVER.

THERE ARE 12 DIAGRAMS

LONDON: BERNARDS (PUBLISHERS) LIMITED

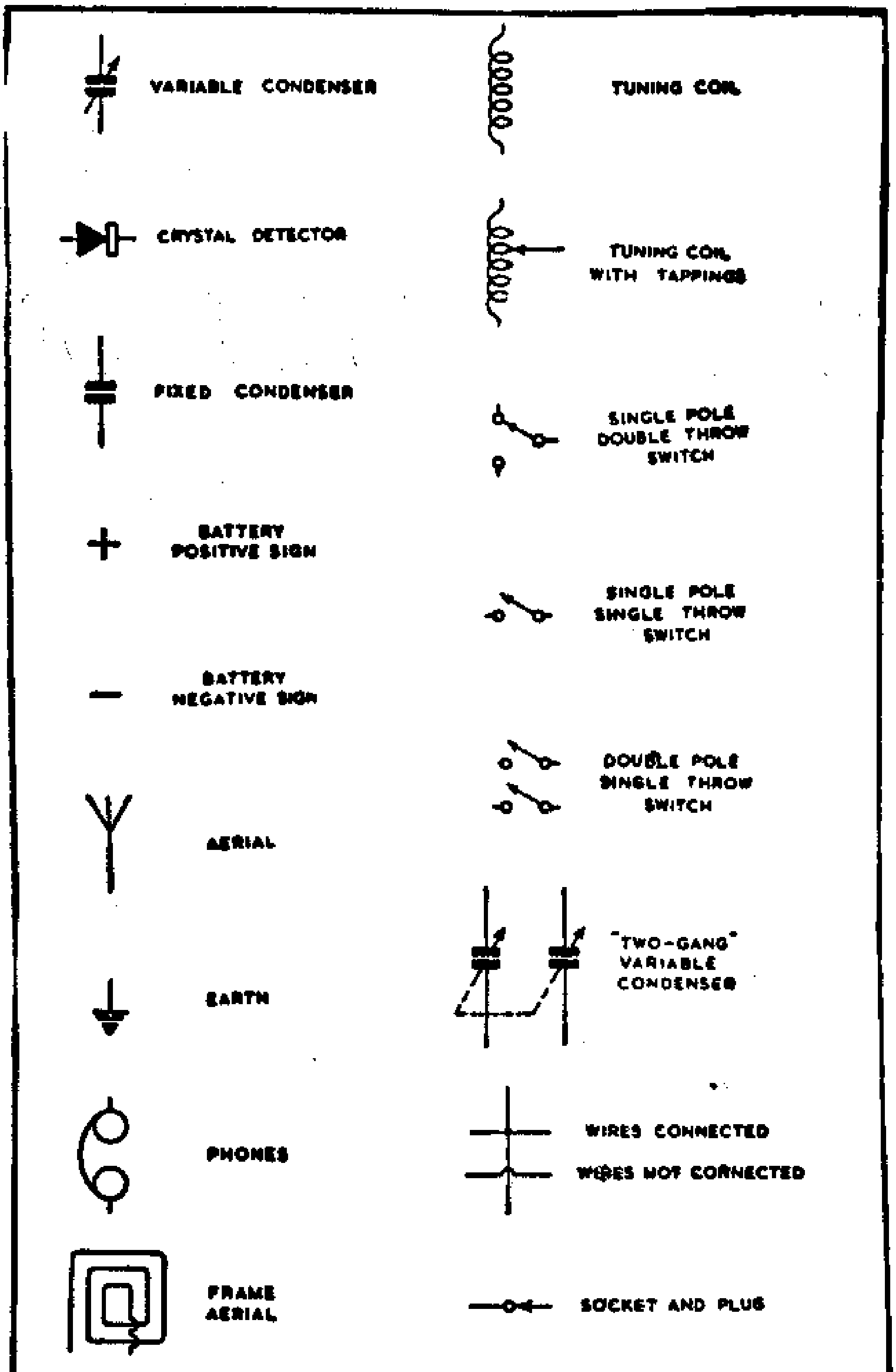
VARIABLE CONDENSER
TUNING COIL
CRYSTAL DETECTOR
TUNING COIL WITH TAPPINGS
FIXED CONDENSER
SINGLE POLE DOUBLE THROW SWITCH
BATTERY POSITIVE SIGN
SINGLE POLE SINGLE THROW SWITCH
BATTERY NEGATIVE SIGN
DOUBLE POLE SINGLE THROW SWITCH
AERIAL
"TWO-GANG" VARIABLE CONDENSER
EARTH
WIRES CONNECTED
WIRES NOT CONNECTED
PHONES
FRAME AERIAL
SOCKET AND PLUG

FIG. I SYMBOLS USED IN THE CIRCUITS DESCRIBED.

1. Aerials.

One of the essentials for satisfactory reception with crystal sets is an efficient aerial. Four different types are suggested and are shown in Fig. 2.

There are several important points to be noted when erecting an aerial for a crystal set.

(1) See that it is suspended as high as possible and is not screened by tall buildings, obstructions, etc.

(2) The aerial itself and the lead-in from the aerial must be adequately insulated. Be sure that the porcelain or plastic insulators do not permit the aerial or lead-in to touch the building, tree, or post from which they are suspended. The lead-in should be brought into the house through an insulated rod.

(3) It is exceedingly important to use insulated copper wire, either single or multi-stranded, for both the aerial and the lead-in.

2. Earths.

An adequate and efficient earth contributes to the good performance of any well-designed crystal set and Fig. 3 shows the best method of obtaining an efficient earth where there is a garden or patch of ground available at a short distance from the receiver.

Obtain a 3-ft. length of ½-in. galvanised iron pipe and at one end drill a hole right through. Insert a bolt with spring washer and locking nut in the hole. Connect the earth lead-in wire to the bolt; the other end of the lead-in to the earth connection on the receiver. Leave both ends of the pipe open. Insert the end opposite the bolt into the ground vertically until only 3-in. are showing above the surface of the surrounding earth.

It is important to see that the surrounding ground is kept damp, if necessary by occasionally pouring a pint or two of water down the open projecting end of the pipe.

Ready-made copper earthing rods with a pointed end for easy hammering into the earth are sometimes obtainable from radio dealers.

Should a garden not be available, an alternative earth is a cold water pipe coming from a rising main. Simply file very lightly the pipe at the back of the tap and wind the copper wire earth lead-in round the filed part and cover the joint with insulating tape.

In no circumstances should a gas pipe be used for an earth.

3. Crystal Detectors.

Most radio dealers can supply very efficient permanent or semi-permanent crystal detectors of proprietary makes. However, there are always available quite a number of ex-Government surplus units which are called crystal valves and, though these were originally designed for a very different use in service equipment, they are easily adaptable and will give extremely efficient results when used in circuits designed for them. To enable the constructor to purchase suitable types for this purpose—and they cost very little—the following list gives the Government reference numbers.

CV 101, CV 102, CV 103, CV 111, CV 112, CV 113, CV 226, CV 241, CV 246, CV 247, CV 253, CV 291, CV 351, CV 361, CV 364, CV 367, CV 727, CV 749, and CV 1785.

One of the satisfactory proprietary makes of this class of crystal detector is the BTH CS7A, which, incidentally, is the commercial equivalent of the CV 253.

For all these Government types, one connection should be made to each end by arranging a holder to fix them into or by carefully soldering the appropriate connecting wires to each end.

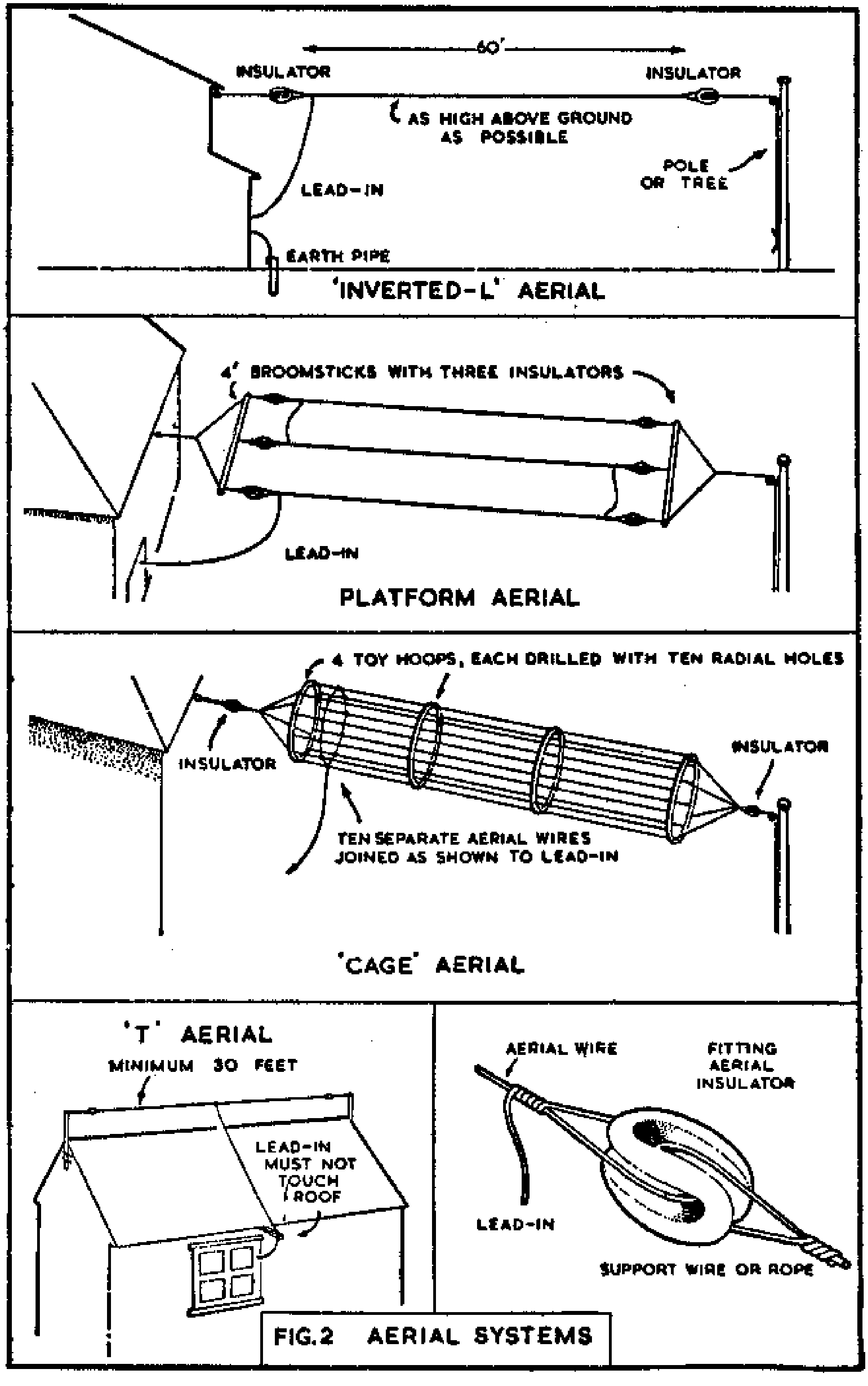
60'
INSULATOR
INSULATOR
AS HIGH ABOVE GROUND AS POSSIBLE
POLE OR TREE
LEAD-IN
EARTH PIPE
'INVERTED-L' AERIAL
4' BROOMSTICKS WITH THREE INSULATORS
LEAD-IN
PLATFORM AERIAL
4 TOY HOOPS, EACH DRILLED WITH TEN RADIAL HOLES
INSULATOR
INSULATOR
TEN SEPARATE AERIAL WIRES JOINED AS SHOWN TO LEAD-IN
'CAGE' AERIAL
'T' AERIAL
MINIMUM 30 FEET
LEAD-IN MUST NOT TOUCH ROOF
AERIAL WIRE
FITTING AERIAL INSULATOR
LEAD-IN
SUPPORT WIRE OR ROPE

FIG.2 AERIAL SYSTEMS

4. Coils.

In Fig. 3 two methods are shown of fixing the coil to the chassis. However, the ingenious constructor may evolve another arrangement.

The important thing to remember is that the coil should be fixed so that it is completely insulated from the metal chassis that may be used.

It is also important to remember that if a metal chassis is used, one must see that all components and wiring, with the exception of earth connections, are adequately insulated from this chassis so as to make sure that no short circuits take place. It is recommended that all crystal sets should be built on a panel, and that both panel and chassis should be of bakelite, paxolin or perspex, as these three materials are efficient insulators in themselves.

5. The Beginners' Crystal Receiver.

The popularity of the crystal receiver among beginners is undoubtedly due to the very modest outlay necessary, and the ease with which such receivers may be built. It is not always appreciated that, to obtain worth while results,

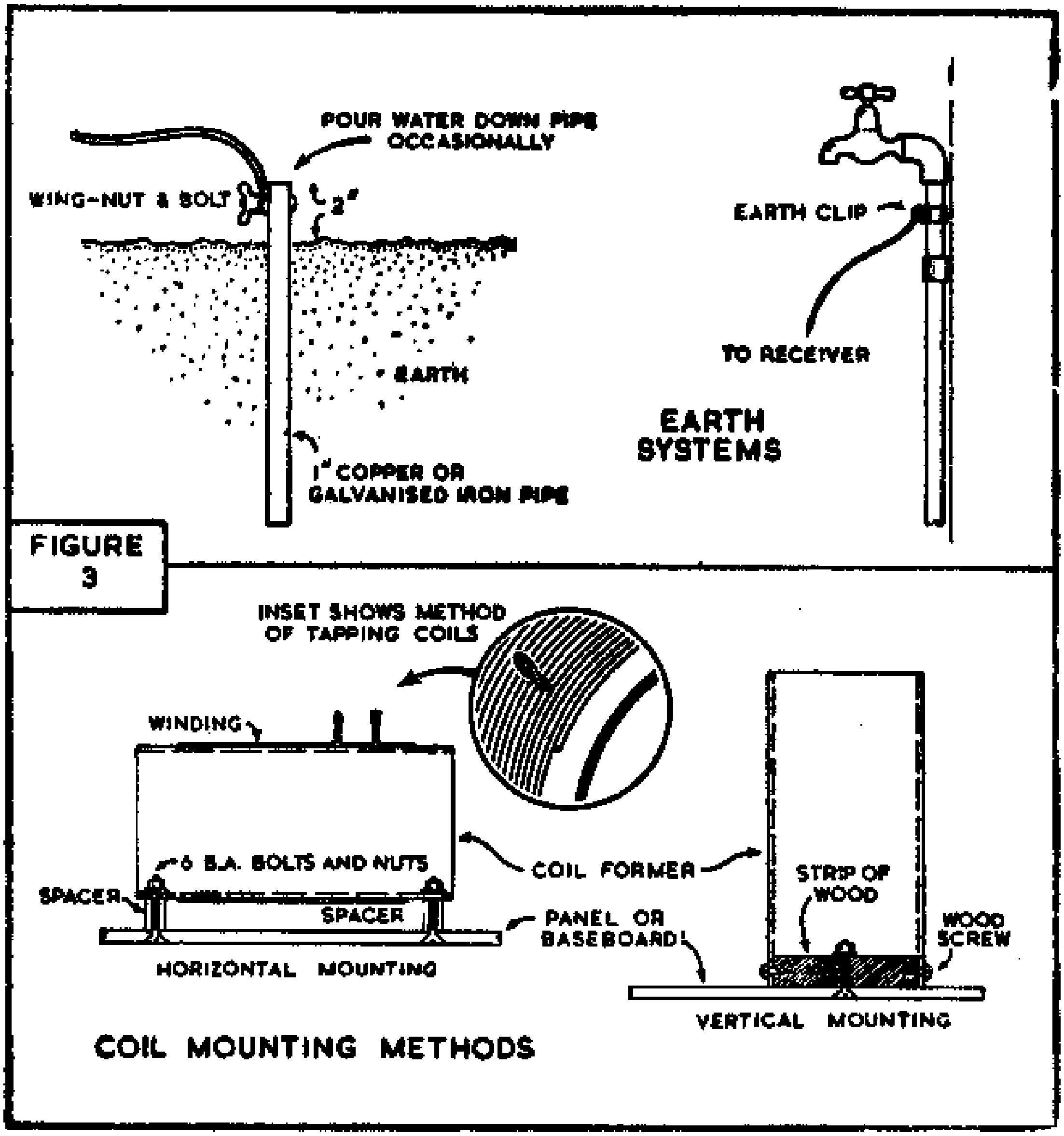
POUR WATER DOWN PIPE OCCASIONALLY
WING-NUT & BOLT
½"
EARTH
1" COPPER OR GALVANISED IRON PIPE
EARTH CLIP
TO RECEIVER
EARTH SYSTEMS
INSET SHOWS METHOD OF TAPPING COILS
WINDING
6 B.A. BOLTS AND NUTS
COIL FORMER
SPACER
SPACER
PANEL OR BASEBOARD
STRIP OF WOOD
WOOD SCREW
HORIZONTAL MOUNTING
VERTICAL MOUNTING
COIL MOUNTING METHODS

FIGURE 3

far more care in avoiding losses is necessary than with a three or four valve receiver. One reason for this is the fact that the crystal receiver can offer no help by way of amplification and it is therefore dependent on an efficient aerial/earth system, low loss design within the receiver itself, and the use of sensitive headphones. The experience gained is well worth the time expended since, when more ambitious receivers are attempted, the constructor will remember the benefits gained by efficient construction and installation with the result that each receiver attempted will be an instrument capable of giving a first class performance.

Losses within the receiver itself may be considerably reduced by efficient coil design. Unfortunately, selectivity is a factor which must not be overlooked; it is this factor which governs the sharpness of the tuning and permits the separation of adjacent powerful stations. The greater the selectivity, the greater the ease with which signals may be separated; but at maximum selectivity the received signals are at minimum strength. In Fig. 4, the coil is tapped at four points, which enables the best point between selectivity and sensitivity to be chosen. The nearer the aerial tap to the earth end of the coil, the greater the selectivity. The pre-set condenser C1 also aids selectivity; at minimum capacity the selectivity is high, and *vice versa.*

Used in conjunction with the coil taps, the constructor should have no difficulty in selecting a condenser setting and tapping point which permit the reception of strong signals and at the same time effect the required separation of powerful stations.

The coil is constructed on a former 3-in. diameter × 2-in., and consists of 50 turns 36 S.W.G. enamelled wire close wound and tapped at every tenth turn. These taps are brought out in the form of loops twisted firmly so that there is no actual break in the winding continuity. After completion the winding may be painted with Durafix to prevent any movement, and the taps and two free ends cleaned with emery cloth to remove the enamel. The layout and wiring is shown in Fig. 4. The headphones should have a resistance of some 4,000 ohms or more and should be of reputable manufacture. Any of the aerials previously described are suitable for use with this receiver.

C2 .0005 variable condenser.
C1 .0002 pre-set condenser.
C3 .001 mica condenser.
1 Crystal detector, semi-permanent pattern.
2 oz. 30 S.W.G. enamelled copper wire.
1 Coil former 3-in. dia. × 2-in.
1 Ebonite or Bakelite panel 6-in. × 5-in. × ⅛-in.
4 terminals, connecting wire, etc.

6. Frame Aerial Receiver for Local Station use on Medium wave bands.

This receiver was specially designed for constructors who are within 5 to 10 miles of a powerful station and who have no facilities for constructing a good outdoor aerial.

The frame itself consists of two ¾-in. × ¾-in. square section wooden rods each 36-in. in length. These are placed together at their centre line to form a cross; they are fastened by a nut and bolt and these in turn pass through an upright piece of wood ½-in. × 1-in. section 2-ft. in length. This upright support is then screwed into a baseboard 8-in. × 8-in. × 1-in. thick. At each of the arms of the 36-in. cross a 4-in. strip of ½-in. square section wood is placed. Reference to Fig. 5B will give the details of the construction of this aerial quite clearly.

It will be noted that approximately 85-ft. of 20 S.W.G. double cotton-covered wire are required for the inductance to cover the medium wave broadcast band. The wire should be wound around the four crossbars, thereby creating

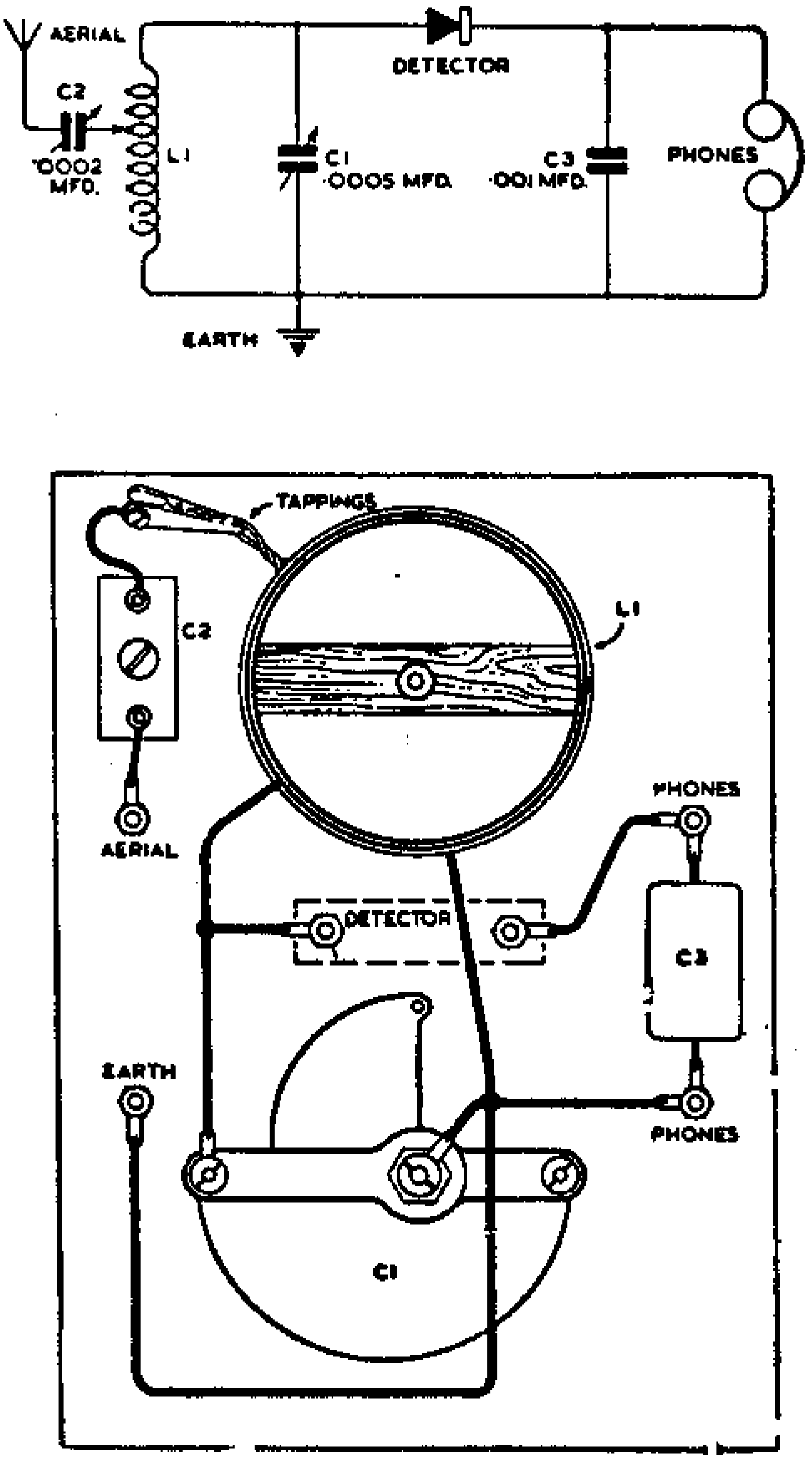
AERIAL
C2
·0002 MFD.
L1
C1 ·0005 MFD.
DETECTOR
C3 ·001 MFD.
PHONES
EARTH
TAPPINGS
C2
L1
AERIAL
PHONES
DETECTOR
C3
EARTH
PHONES
C1

FIG. 4 THE BEGINNER'S CRYSTAL RECEIVER.

a square wire frame, and each turn of the wire should be spaced approximately ⅛-in. The best way to effect this spacing accurately is to file on the 4-in. crossbars a number of small nicks at ⅛-in. intervals. The two ends of the wire should then be brought down on the two terminals which may be screwed into the square baseboard and two leads taken from there to the crystal set itself.

The actual set may be constructed on a small ebonite, paxolin or perspex panel 8-in. × 5-in. which may be screwed to one side of this base supporting piece of wood, thereby making the set virtually self-contained except for the earphones. Fig. 5A shows the circuit of the receiver.

To obtain the best results with this receiver, the aerial should be rotated by hand until the signals become loudest.

1 Semi-permanent crystal detector.
C1 1 .0005 mfd. variable condenser.
C2 1 .0005 mfd. fixed condenser.
1 condenser dial marked 0 to 100 or 0 to 180.
1 pair of high sensitivity earphones.
85-ft. to 90-ft. of 20 S.W.G. double cotton covered wire. (8oz.)
Sufficient connecting wire for wiring purposes.

7. Long Distance Receiver.

This receiver has been specially designed to achieve high sensitivity. With a sufficiently good outdoor aerial effective reception has been obtained up to 150 miles.

The circuit is quite straightforward, as can be seen from Fig. 6 and the only special point is in winding the coil L1.

This coil consists of a 2½-in. diameter former on which are wound a total of 51 turns using 24 S.W.G. double cotton wire. Starting at one end, tap every two turns until there are eight taps; wind a further 15 turns and tap; then a further 10 turns and tap; finally, add 10 turns to end the winding. The wire should be close wound, *i.e.*, with turns touching.

To obtain the best results, the two crocodile clips A and B should be tried on the various coil taps until the loudest reception is obtained. It may be found in use that the loudest reception of any particular station causes another station to be heard at the same time; care should, therefore, be taken to select the appropriate tapping point to separate any two powerful stations that may interfere with one another.

4 oz. 24 S.W.G. DCC copper wire.
Coil former 2½-in. dia. × 3-in.
C1-2 single variable condensers .00035 mfd.
2 Condenser dials marked from 0 to 100 or 0 to 180.
1 Permanent crystal detector.
C3 .001 fixed condenser.
1 pair of high sensitivity earphones (preferably between 4,000 ohms and 8,000 ohms impedance).
2 Crocodile clips.

8. High Gain Receiver.

This receiver has been designed to give good reception up to 50 miles distance from any normal power broadcasting station in the medium band. It is a very selective circuit and will therefore enable the user to separate closely situated stations.

It will be noticed in the circuit (Fig. 7) that the variable condenser C1, has an on/off switch S1 to throw it out of circuit if necessary. It is used to increase the selectivity of the receiver and, in operating, it should be tried with this condenser either in or out of circuit on the various tappings on the coil L1.

L1 consists of 90 turns of 22 S.W.G. double cotton covered wire wound on a 2-in. diameter former and tapped every ten turns. L2 separately wound on a 2-in. former, consists of 90 turns of 22 S.W.G. double cotton covered wire, also tapped as L1. The components required are given on page 9

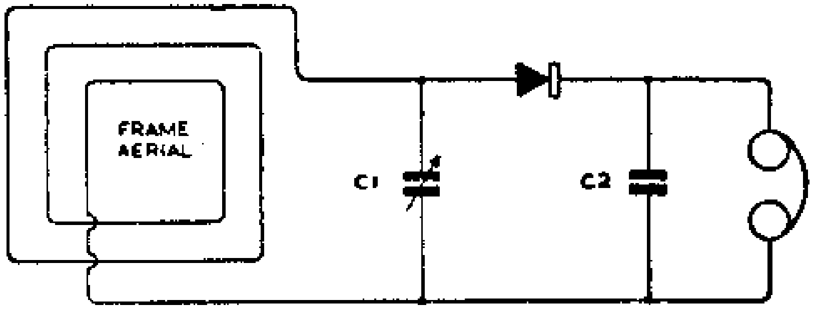
FRAME AERIAL
C1
C2

FIG. 5A FRAME AERIAL CRYSTAL RECEIVER

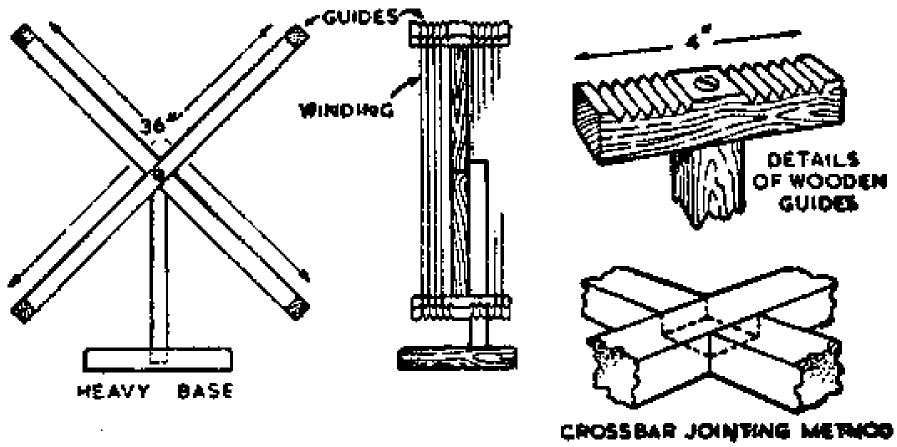
GUIDES
36"
WINDING
HEAVY BASE
4"
DETAILS OF WOODEN GUIDES
CROSSBAR JOINTING METHOD

FIG. 5B. DETAILS OF FRAME AERIAL

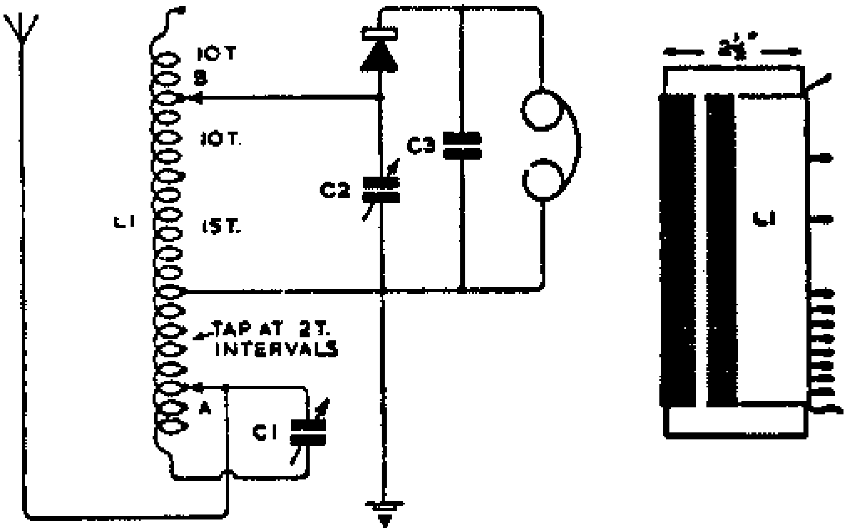
10T
B
10T.
15T.
L1
C2
C3
TAP AT 2T. INTERVALS
A
C1
2½"
L1

FIG. 6 LONG-DISTANCE CRYSTAL RECEIVER

C1-2 Single variable condensers .00035 mfd.
S1 Single pole single throw switch.
C5 .0001 preset condenser.
1 Permanent crystal detector.
C3-4 .001 mfd. fixed condenser.
1 Pair of high sensitivity earphones.
8 oz. 22 S.W.G. DCC.
2 Dials for variable condenser marked 0 to 100 or 0 to 180.

9. High Selectivity Receiver.

This circuit (see Fig. 8) is for use in high saturation strength areas where it is desired to separate powerful stations. The circuit is not complicated and is extremely effective.

It will be noted that C1 and C2, each .00035 mfd. capacity, are 2 gang unit.

The only special point in this set is the design of coils L1, 2 and 3, which are constructed as follows: L1, 60 turns of 22 S.W.G. double cotton covered wire on a 3 × 4 inches former tapped every five turns; L2, 15 turns of 22 S.W.G. double cotton covered wire on a 3 × 6 former; and L3, 85 turns of 22 S.W.G. double cotton covered wire wound on the same former as L2, the two windings being separated by approximately ¼-in. (see Fig. 8).

Note that the aerial lead is brought down to one of the tappings on wire L1. This should be tried on each of the tappings and the one giving the required selectivity should be used.

C1-2 2-gang variable condenser .00035 mfd.
2 3-in. formers for coils L1-2-3.
8 oz. 22 S.W.G. DCC.
C3 .0005 fixed condenser.
1 Semi-permanent crystal detector.

10. Battery-aided Crystal Receiver.

Fig. 9 shows a design which utilises a 9-volt grid bias battery to improve performance.

The charge built up in the .02 mfd. fixed condenser is changed to a positive current when it passes through the coil L1. This, in turn, is superimposed on the positive charge that is commencing to be rectified by the crystal detector and thereby increases the signal strength.

Coil L1 consists of a 4-in. diameter former wound with 85 turns of 24 S.W.G. double cotton covered wire tapped at 5, 25, 45 and 65 turns from the earth end. The earth end and these four taps are brought out to five small sockets and the connection from the earth is taken to a small socket. The plug should be tried in each of the five positions to find which one gives the greatest signal strength.

The 9-volt battery is connected across the .02 mfd. condenser and special care should be taken to see that the battery is connected correctly, *i.e.*, negative pole to earth.

1 Permanent crystal detector.
1 4-in. dia. × 3-in. former.
8 oz. 20 S.W.G. DCC copper wire.
C1 Variable condenser .0005 mfd.
C2 Fixed condenser .005 mfd.
C3 Fixed condenser .02 mfd.
1 pair of high sensitivity headphones.
1 Clix plug.
5 Clix sockets.
1 9-volt grid bias battery.

11. Medium and Long wave band Receiver.

This set has been designed for reception on two wave bands.

Care must be taken in winding the coils L1, 2, 3 and 4. These are wound on a single 3-in. diameter former. L1 consists of 30 turns of 28 S.W.G. DCC tapped at 8, 16 and 24 turns. L2, 60 turns of 32 S.W.G. DCC. L3, 40 turns of 28 S.W.G. DCC, and L4, 80 turns of 32 S.W.G. DCC. All are close wound and are separated from each other as follows: between L1 and L2, 3/16-in.;

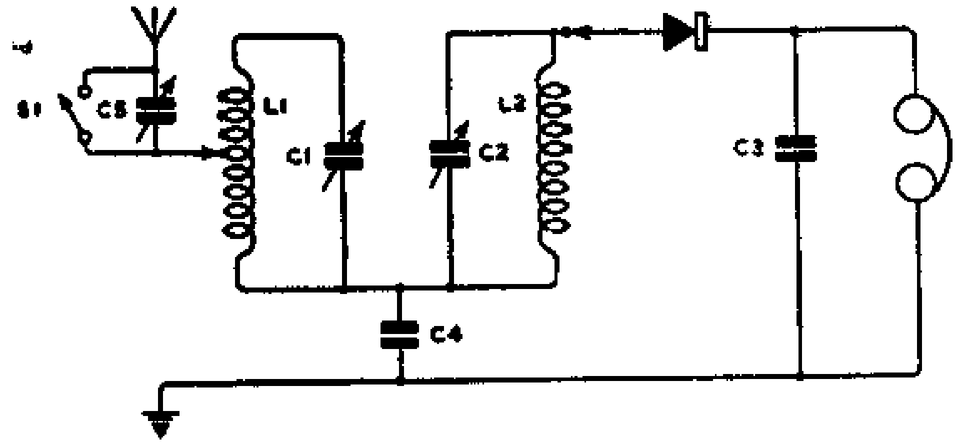
S1
C5
L1
C1
C2
L2
C3
C4

FIG. 7 HIGH-GAIN CRYSTAL RECEIVER

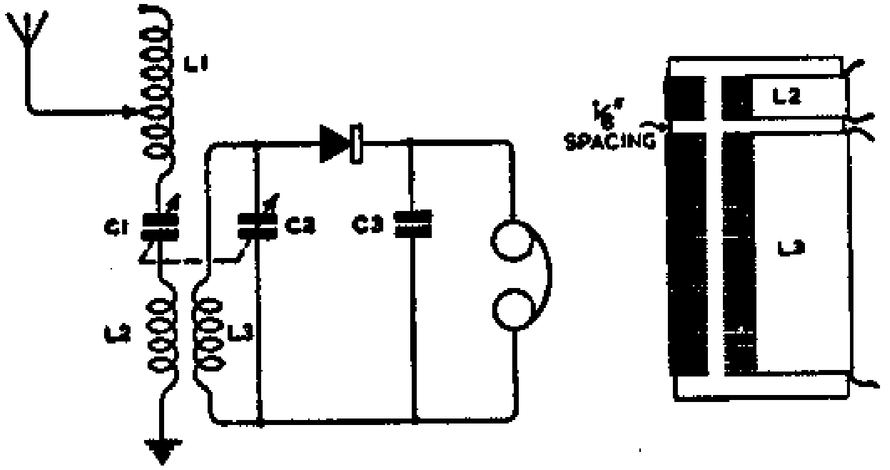
L1
C1
C2
C3
L2
L3
1/8"
SPACING
L2
L3

FIG. 8 HIGH SELECTIVITY CRYSTAL RECEIVER

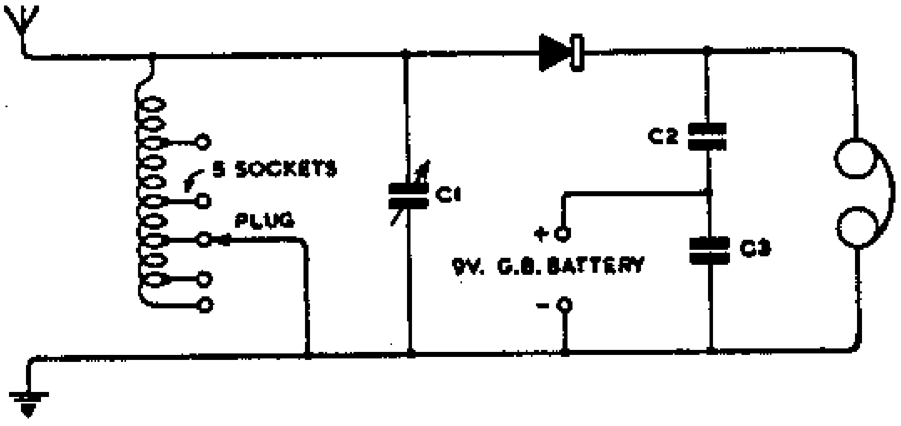
5 SOCKETS
PLUG
C1
C2
+
9V. G.B. BATTERY
-
C3

FIG. 9 BATTERY-AIDED CRYSTAL RECEIVER

between L2 and L3, 1-in.; and between L3 and L4, 3/16-in. See Fig. 10.

C1-2 single variable condensers
.0005 mfd.
2 Variable condenser dials marked
0 to 100 or 0 to 180.
1 Clix plug.
3 Clix sockets.
S1 Single pole double throw switch.
C3 .001 mfd. fixed condenser.
1 Pair of high sensitivity earphones.
1 3-in. dia. × 8-in. former.
2oz of both 28 and 32 S.W.G. double cotton covered wire to make these 4 coils.
1 Permanent crystal detector.

In use, the plug should be tried in each of the three sockets to find the point which gives the best results. The switch is used as a wave band change switch.

12. Ultra Sensitive All Wave Receiver.

The receiver shown in Fig. 11 is for long distance reception. It is a novel circuit and, though apparently complicated in construction, it is really very simple to assemble with the minimum of tools and labour.

It will be seen from the circuit diagram that there are two banks of sockets, five black and six red. When using this set the appropriate black and red plugs should be inserted in whichever of the similar coloured sockets give the best results. The coil is quite easy to construct and consists of a primary winding which is inserted at the earth end of the secondary winding. For the medium wave band, the secondary winding L2 consists of 54 turns of 20 S.W.G. enamelled or cotton covered wire. This coil has four taps in addition to the starting point and the end connection, and these taps are taken at six turns, 14 turns, 27 turns and 40 turns from the earth end. The taps are taken out and a connection is made to an appropriate socket, as shown in the diagram.

The coil is wound on a 3-in. diameter former which may be a paper tube, bakelite, paxolin or perspex, and spacing is approximately 17 turns to the inch.

For short wave work down to about 30 metres, L2, the secondary coil, will consist of the same diameter former wound with 15 turns of the same gauge wire, the four taps being taken at 3, 6, 9 and 12 turns from the earth end.

L1, the primary coil, is used for both wave bands and consists of 11 turns of similar gauge wire to that used for L2, wound on a 2-in. × 1-in. former, the winding to be spaced the same as L2. The coil is then inserted at the earth end of whichever secondary coil L2 is being used.

The crystal detector may be one of the semi-permanent types freely available to-day; the old-fashioned cat's whisker pattern; or one of the silicon permanent detectors which are sold by many surplus dealers.

The 'phones should be of a standard type 4,000 ohms impedance although 'phones of other values will probably operate quite satisfactorily with this powerful receiver. It should, however, be stressed that the better the headhones used, the more satisfactory the reception.

Coils for other wave bands may be experimentally designed by the constructor, remembering that four tappings are advisable for the secondary coil L2.

A long aerial is very desirable with this receiver, placed as high as possible. It is recommended that a minimum length of 70-ft. be allowed, to include the aerial and lead-in.

The values of the components are shown in the following components list. The constructor should experience no difficulty in obtaining excellent results from this cleverly designed receiver.

C1 .0005 variable condenser.
C2 .002 fixed condenser.
1 semi-permanent crystal detector.
1 coil former 3-in. dia. × 4½-in.
1 coil former 2-in. dia. × 1-in.
½ lb. 20 S.W.G. enamelled or DCC copper wire.
6 red wander plug sockets.
6 black wander plug sockets.
1 red and 1 black wander plug.

Headphones, terminals, connecting wire

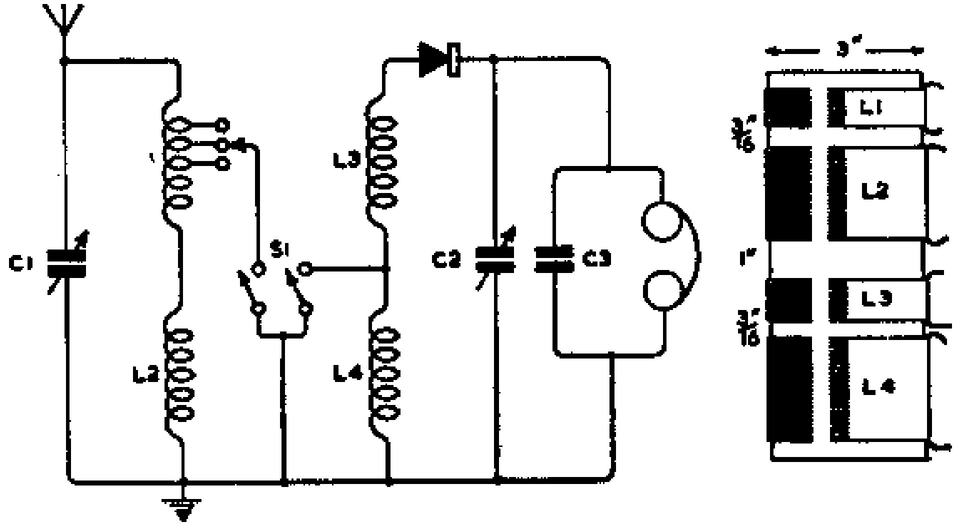
C1
L2
S1
L3
L4
C2
C3
3"
3/16"
L1
L2
1"
L3
3/16"
L4

FIG. 10 MEDIUM AND LONG-WAVE CRYSTAL RECEIVER

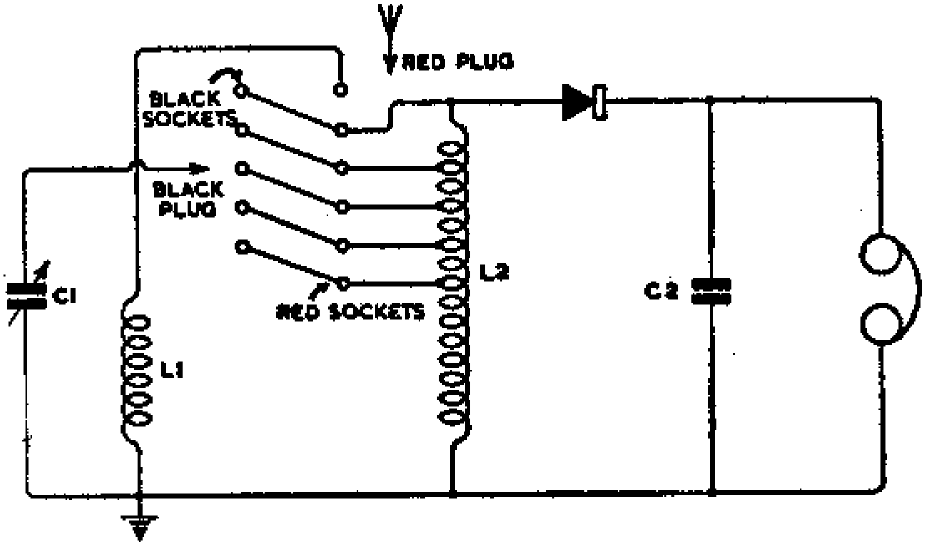
RED PLUG
BLACK SOCKETS
BLACK PLUG
C1
L1
RED SOCKETS
L2
C2

FIG. 11 ULTRA-SENSITIVE ALL-WAVE CRYSTAL RECEIVER

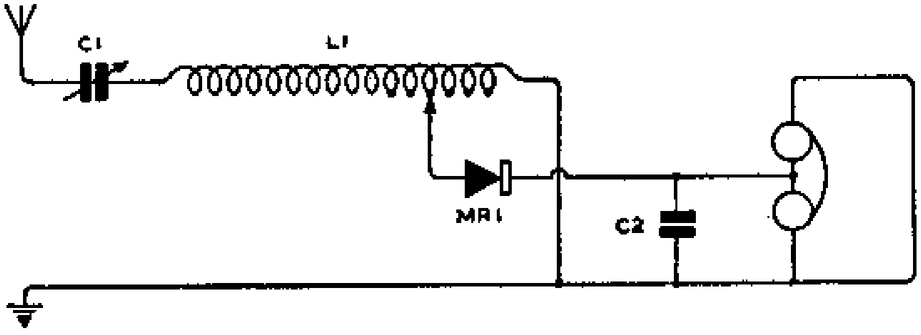

FIG. 12 ADVANCED CRYSTAL DIODE RECEIVER

13. An Advanced Crystal Diode Receiver.

The modern crystal diode of the radar type does not require any adjustment and thus has a distinct advantage over the ordinary crystal detector which needs delicate adjustment—even the semi-permanent type.

Such crystals are now obtainable on both the regular and the surplus markets, and a popular type is the B.T.H. CS7A silicon crystal (also coded as CV 253).

A rather different circuit from the usual crystal set arrangement is needed to suit the characteristics of a crystal diode. The diode must be tapped on to the tuning coil, and it is found that the tuning circuit itself gives best results if a series-tuned acceptor circuit is used. The circuit of a radar crystal receiver is shown in Fig. 12. The type of aerial used with this circuit has a very great bearing on the behaviour of the receiver for, in effect, the tuned circuit, with a series resistance equivalent to the reflected crystal load resistance, is in series with the capacitance of the aerial to earth and the aerial's effective series resistance.

At resonance—when the combination is tuned to any particular signal—the inductance resonates with the capacitances of the tuning condenser and the aerial in series, and the final effect is that the reflected load of the crystal, in series with the coil's R.F. resistance, is paralleled across the aerial's series resistance.

For maximum power transfer, the crystal load resistance must be made equal to the sum of the aerial series resistance and the coil R.F. resistance, and so the method in which the crystal is tapped into the tuning coil, and the exact capacitance required to tune the receiver to any required signal, must depend to a very great extent on the aerial itself.

At the same time the crystal resistance varies with the signal strength, the resistance being high for weak signals and dropping by as much as 50 per cent. and more for strong signals, so that this effect also has a bearing on the correct coil tap. The output impedance of the crystal also varies similarly, affecting the matching of the headphones into the crystal diode, and so for any set of conditions, the receiver requires to be matched up to both the aerial and the signal being received for best results. This would mean a series of coil taps and (theoretically, not practically) a matching transformer between the diode and the headphones; but in practice it will be found that a receiver using standard parts can be built up to give very good results under various conditions.

It has been shown already that the tuning of the receiver depends to a great extent on the characteristics of the aerial, and while a .0005 mfd. variable capacitor is shown in Fig. 12 as the tuner, the constructor must be prepared to experiment with different capacitance values until the required station is tuned in. The range of reception given by the receiver is quite good, if a really long and high aerial, and a good earth connection, are used; but no more than the local station signal can be expected, and the tuning therefore must be adjusted to suit the station frequency.

It must be mentioned that the headphones are shown parallel connected. High resistance headphones of the 4,000 ohm type must be used, and if two of these are connected in parallel rather than in the more usual series method, they will provide a roughly accurate match to the crystal. If more than one pair of headphones are to be connected in, then the pairs of headphones may be left series connected in the usual manner, the sets of headphones being connected in parallel.

The capacitor across the headphone terminals completes the crystal R.F. circuit, and any value between about 0.001 and 0.005 mfd. will serve. The higher capacitance will, of course, by-pass some of the higher audio frequencies, so that if good steady reception is obtained some experiment with this condenser is also worth while.

C1, .0005 mfd. variable condenser.
C2, .005 mfd. fixed condenser.
M.R.1 CV 253 Crystal diode.
Coil former 1½-in. dia. × 4½-in.
4oz. 26 S.W.G. enamelled copper wire.

Coil details are as follows: 150 turns of 26 S.W.G. tapped at 20, 25, 30, 35, 40 and 45 turns from the earth end. At each tapping-point the wire should be twisted up into a loop and the winding then continued without breaking the wire; when the coil is completed and the ends anchored the tapping loops can be bared.

BERNARDS RADIO BOOKS

No.		Price
56.	Radio Aerial Handbook	2/6
57.	Ultra-Shortwave Handbook	2/6
58.	Radio Hints Manual	2/6
64.	Sound Equipment Manual	2/6
68.	Frequency Modulation Receivers' Manual	2/6
73.	Radio Test Equipment Manual	2/6
83.	Radio Instruments and their Construction	2/6
96.	Crystal Set Construction	1/–
99.	One Valve Receivers	1/6
100.	A Comprehensive Radio Valve Guide, Book 1	5/–
103.	"Radiofolder" A. The Master Colour Code Index for Radio and T.V.	1/6
104.	Three Valve Receivers	1/6
107.	Four Valve Circuits	1/6
108.	Five Valve Circuits	2/6
121.	A Comprehensive Radio Valve Guide, Book 2	5/–
123.	"Radiofolder" F. The Beginners' Push-Pull Amplifier	1/6
126.	Boys' Book of Crystal Sets and Simple Circuits	2/6
129.	Universal Gram Motor Speed Indicator	1/–
134.	F.M. Tuner Construction	2/6
135.	All Dry Battery Portable Construction	2/6
138.	How to Make F.M. and T.V. Aerials, Bands 1, 2 and 3	2/6
141.	Radio Servicing for Amateurs	3/6
143.	A Comprehensive Radio Valve Guide, Book 3	5/–
145.	Handbook of AM/FM Circuits and Components	2/–
146.	High Fidelity Loudspeaker Enclosures	5/–
147.	Practical Tape Recording Handbook	5/–
148.	Practical Transistor Receivers, Book 1	5/–
149.	Practical Stereo Handbook	3/6
150.	Practical Radio Inside Out	3/6
151.	Transistor Superhet Receivers	7/6
155.	Portable Transistor Radio and Radiogram	2/6
156.	Transistor Circuits Manual, No. 1	2/6
157.	A Comprehensive Radio Valve Guide, Book 4	5/–
158.	Radio, Television, Industrial Tubes, Semiconductor and Diodes Equivalents Handbook (208 pages)	9/6
159.	Realistic High Fidelity	5/–
160.	Coil Design and Construction Manual	5/–
161.	Radio, Television and Electronics Data Manual	3/6
162.	High Fidelity Stereo Gramophone	5/–
163.	Transistor Circuits Manual, No. 2	2/6
164.	High Fidelity Tape Recorder for the Home Constructor	2/6
165.	Hartley on Hi-Fi, Book 1, Radio Tuners	5/–
166.	Public Address Systems	2/6
167.	Transistor Circuits Manual, No. 3	2/6
168.	Transistor Circuits Manual, No. 4	2/6
169.	Hi-Fi Transistor F.M. Tuner for the Home Constructor	3/6
170.	Transistor Circuits for Radio Controlled Models	7/6
171.	Super Sensitive Transistorised Pocket Set Construction	3/6
172.	International Radio Stations List	2/6
173.	Practical Transistor Audio Amplifiers	3/6
174.	Transistor Subminiature Receivers Handbook	5/–
175.	Transistorised Test Equipment and Servicing Manual	3/6
176.	Transistor Audio Amplifiers Manual	6/–
177.	Modern Transistor Circuits for Beginners	7/6
178.	Comprehensive Radio Valve Guide, Book 5	6/–
179.	Transistor Circuits Manual No. 5	5/–
180.	British Semiconductor Survey	15/–
181.	22 Tested Circuits Using Micro Alloy Transistors	5/–
182.	"At A Glance" Radio Valve & T.V. Tube Equivalents	3/6
183.	How to Receive Foreign T.V. Programmes on your Television Set by simple Modifications	5/–
	Resistor Colour Code Disc Calculator	1/6
	Engineers' Reference Tables	1/6
	International Radio Tube Encyclopaedia—3rd Edition	63/–

all about
CRYSTAL SETS

by
Charles Green

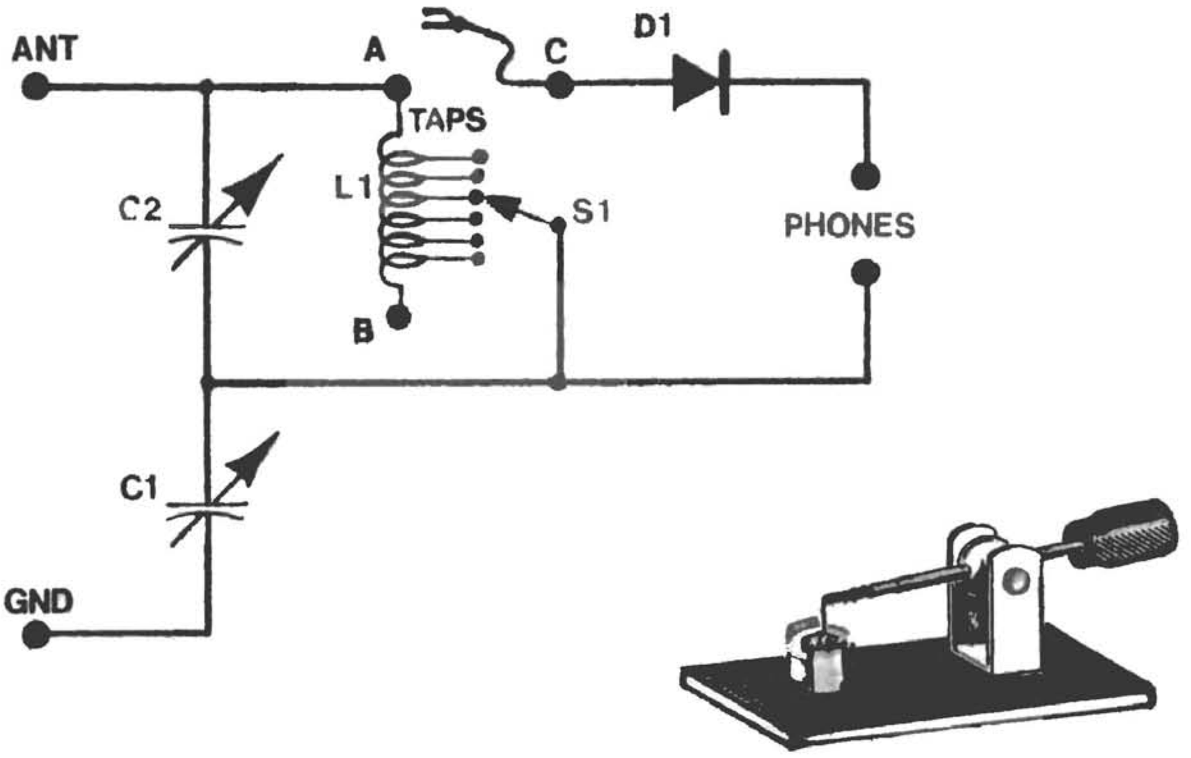
ANT
A
C
D1
TAPS
L1
C2
S1
PHONES
B
C1
GND

How-To-Build Simple
Crystal Set Radios

621.3841
G795A
c.1

allabout books no. 101

all about
CRYSTAL SETS

by

Charles Green

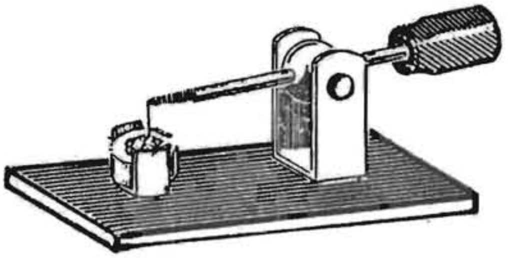

How - To - Build Simple Crystal Set Radios

FIRST EDITION

All rights reserved, including the
right of reproduction in whole or
in part in any form.

Copyright © 1984 by Charles Green

Published by Allabout Books
P.O. Box 22366
San Diego, CA 92192-2366

ISBN 0-930003-00-4

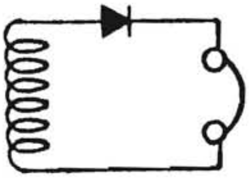

PREFACE

In the beginning of this century, when radio was called "Wireless", most of the early experimenters used a crystal set to receive radio signals.

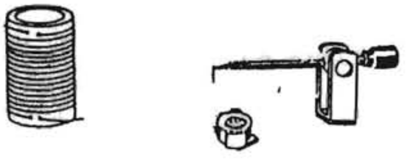

The crystal set is still the simplest type of radio receiver. A catwhisker (made from a thin wire) touching a sensitive spot on a piece of galena or silicon mineral crystal and a tuning coil, form the basic parts of this radio.

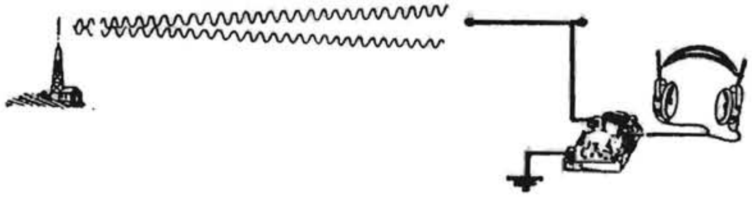

Radio signals from a nearby station are picked up via antenna and ground connections to the crystal set. The galena or silicon crystal detects the signals and the resultant audio is heard with headphones. No battery or electrical power is required.

This book tells you about crystal sets and shows you how to build them.

Charles Green

CONTENTS

1

THE CRYSTAL DETECTOR....................................4
How-it-Works, 4
Detection of Radio Waves, 6
Tuning the Crystal Set, 7
Capacitor, 8
RF Coil, 8
Tuned Circuit, 9
Antenna and Ground, 10
Headphones, 11

2

SIMPLE RECEIVER PROJECTS..............................12
Simple Crystal Set, 12
Slider Tuned Crystal Set, 17
Ferrite Coil Crystal Set, 23
Spiderweb Coil Crystal Set, 27

3

ADVANCED RECEIVER PROJECTS............................32
Dual-Tuned Crystal Set, 32
Selective Crystal Set, 37
Carborundum Crystal Set, 43

4

COMPONENT PROJECTS....................................48
Book Tuning Capacitor, 48
Rotary Switch, 53

5

WORKSHOP..54
Soldering, 54
Project Construction Tips, 55
Mounting Minerals, 56

INDEX...57

1

THE CRYSTAL DETECTOR

A crystal detector of radio waves can be either a traditional catwhisker and natural mineral crystal, or a modern semiconductor diode.

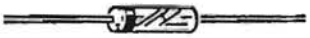

SEMICONDUCTOR DIODE

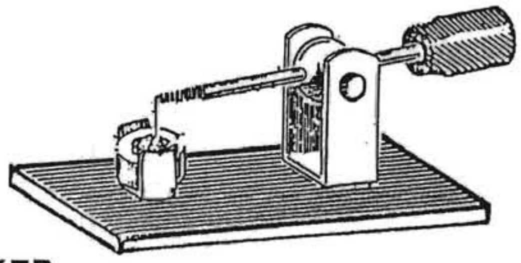

CATWHISKER DETECTOR

Certain natural minerals such as galena (lead sulphide) and silicon have electrically sensitive surface spots. In the beginning of this century, experimenters discovered that electrical current would flow into the sensitive spot on a mineral crystal from a catwhisker, but would not pass readily (high resistance) from the crystal to the catwhisker. This unique feature was used to detect (rectify) radio waves and extract the audio modulation.

HOW-IT-WORKS

There are two general types of crystal detctors;

. Point contact diode
. Junction diode

Both types use semiconductor material such as silicon, germanium, or galena. The basic semiconductor junction diode is made of two portions of the material "doped" in manufacturing by adding controlled amounts of chemicals. The doped portions have opposite electrical properties:

"P" TYPE, with an excess of positive electrical charges (holes).
"N" TYPE, with an excess of negative electrical charges (electrons).

Junction Diode

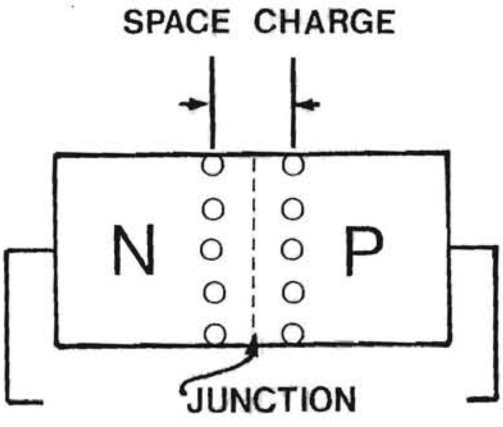

The boundary between the two doped portions is called a PN junction. The area on both sides of the PN junction is called the space charge region. When a positive voltage is applied to the "P" material (and negative voltage to the "N" material), the space charge region becomes narrow and electron current flows through the PN junction.

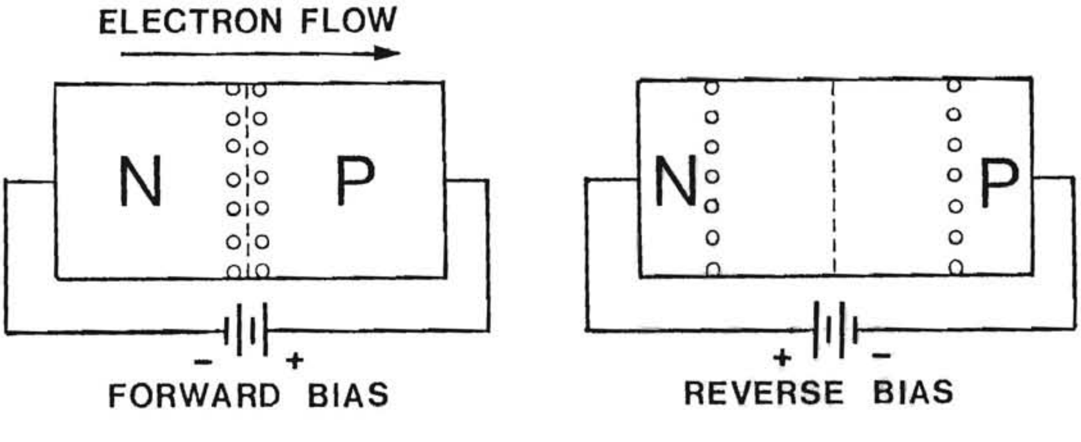

When the voltage polarities are reversed, negative to "P" and positive to "N", the space charge region widens (increasing the effective resistance of the PN junction) and the electron flow is very small (back current). If an ac voltage is applied to the PN junction, it will act as a rectifier (or radio wave detector).

Point Contact Diode

The point contact diode uses a "catwhisker" of thin wire contacting the surface of doped "N" semiconductor material. The pointed end of the wire produces a high intensity electric field in the area of the semiconductor surrounding the wire. When an electric current flows from the wire to the semiconductor material, a small region of "P" material forms around the area of the point contact. Thus a PN junction is formed which operates in the same way as the junction diode.

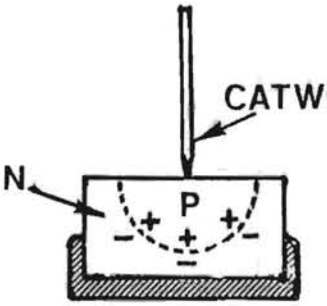

The sensitive spots of galena and silicon crystals used in traditional catwhisker detectors have naturally doped areas of semiconductor material, and operate in the same way as the point contact detector.

DETECTION OF RADIO WAVES

At the transmitting radio station, a microphone converts sound waves into electrical audio signals. The audio signals then modulate the transmitted radio waves.

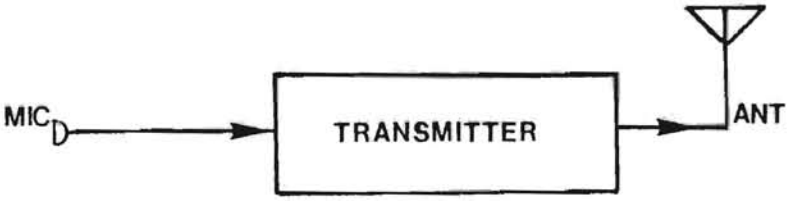

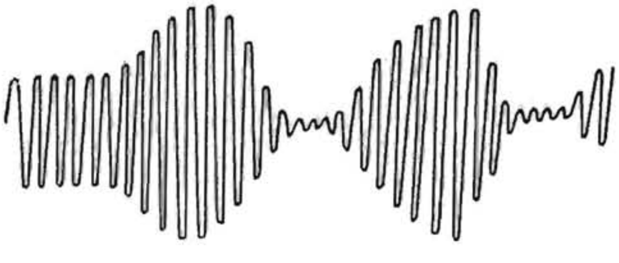

MODULATED RADIO WAVES ⟶

the crystal detector

The transmitted radio waves are intercepted by the receiving antenna and small rf currents are induced in the antenna wire. The rf currents (high frequency ac) flow through the detector to ground. As previously explained, the crystal detector will pass current in only one direction, therefore, pulsating dc voltages (detected audio signals) are developed by the crystal equivalent to the audio modulation of the transmitted radio waves.

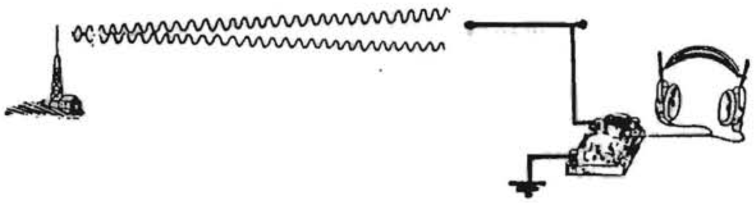

The detected audio signals flow through the headphone coils and affect the intensity of a magnetic field. This causes the steel discs of the headphones to vibrate and produce audible sound waves.

TUNING THE CRYSTAL SET

The antenna intercepts all radio waves in the area, and the signals will be heard jumbled up in the earphones. A nearby radio station may be received with the other stations heard in the background.

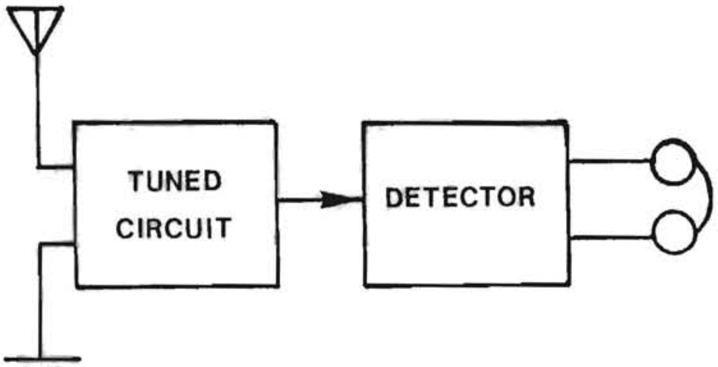

In order to receive a radio station properly, the crystal set must be tuned to the frequency of the transmitter. A tuned circuit composed of an rf coil (inductance) and a capacitor, is connected between the antenna/ground circuit and the detector.

Capacitor

The modern capacitor is evolved from the old Leyden jar. It consists of two or more metal sections separated by a dielectric (air, paper, or plastic). The capacitor stores energy in the form of an electrostatic charge in its dielectric. A variable capacitor (used in a tuned circuit) has movable metal sections that adjust the amount of capacity.

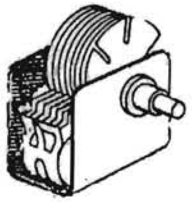

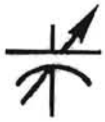

A capacitor will allow pulsating dc current or ac current to flow through it, but will block steady dc current. A capacitor has an oppositon or "reactance" to ac that varies with frequency. The higher the frequency, the lower the reactance of a given value of capacity. Capacity is measured in micro-farads (uF) or micro-micro-farads (picofarads) (uuF or pF).

RF Coil

An rf coil is made of insulated copper wire wound on a plastic or cardboard tube. When electric current flows through the coil, a magnetic field is created. This magnetic field resists any change in electric current flow. This opposition or "reactance" to ac varies with frequency. The higher the frequency, the higher the reactance of a given value of rf coil inductance. Inductance is measured in millihenrys (mH) and microhenrys (uH).

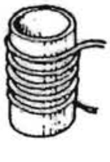

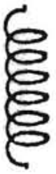

Tuned Circuit

A tuned circuit is composed of a capacitor and an rf coil. As previously described, the reactances of a capacitor and an rf coil change with frequency (but in opposite ways). At a particular frequency, the two reactances will cancel each other out. This condition is called "resonance". The impedance of a parallel tuned circuit is at its highest at resonance. Radio signals at the resonant frequency will be greatly magnified, while signals at other frequencies will be shunted to ground (rejected).

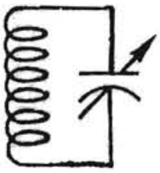

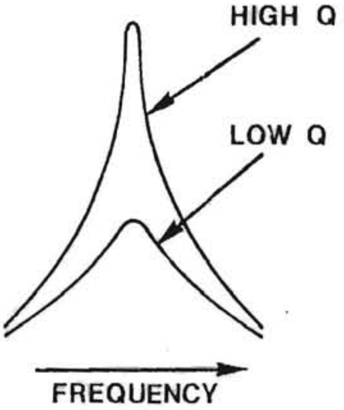

The resonant frequency of a tuned circuit can be changed by adjusting the capacitance (using a variable capacitor),or by varying the inductance (using a slider or taps on the rf coil). A ferrite core can also be used to change the inductance of a coil. The selectivity of a tuned circuit is dependent upon the quality (Q) of the components and the inductance /capacity ratio (L/C ratio). A high-Q tuned circuit is more selective than a low-Q circuit. A tuned circuit with a high L/C ratio (large inductance and small capacity) is more selective than a low L/C ratio circuit.

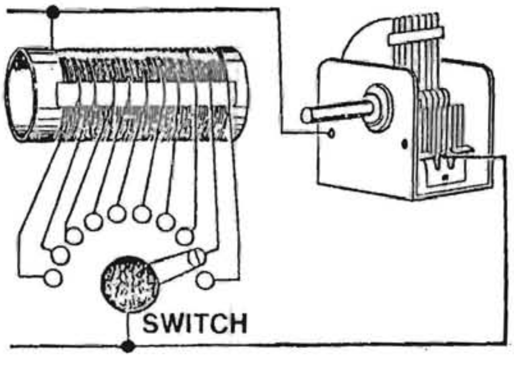

ANTENNA AND GROUND

A simple crystal set depends upon the strength of the radio waves received by the antenna to produce audio in the earphones. If you live in a strong signal area (a nearby radio transmitter), a crystal set can produce loud audio in earphones with only an inside antenna. But in lower signal areas, a good outside antenna with a ground connection is required.

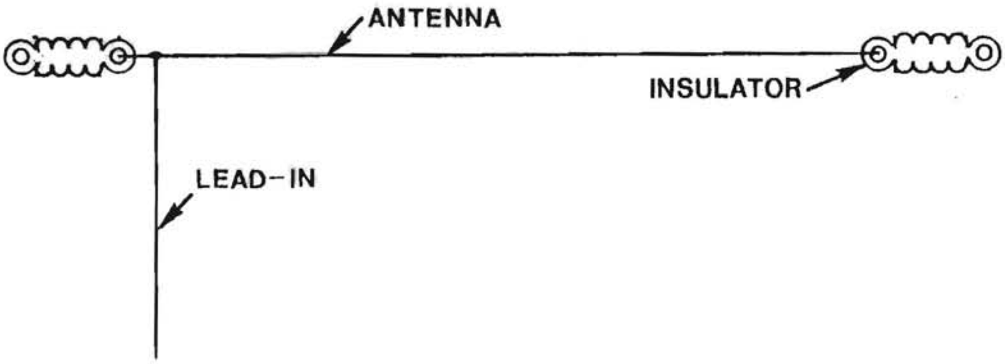

If you are fortunate enough to be able to put up an outside antenna, the basic rule of thumb is the higher, the better reception. Check your local electronics store for an antenna kit. You can construct an antenna using a glass or plastic insulator at each end of a length of copper wire (insulated or bare). Make the antenna as long as possible. It does not have to be in a straight line to work.

CAUTION

ANTENNA INSTALLATION
NEAR ELECTRICAL POWER
LINES IS DANGEROUS.

Make sure that your antenna
is far enough away from any
power line, so that it will
not touch a wire if it falls.

Connect a length of insulated wire between the end of the antenna and your crystal set antenna terminal. This antenna lead-in wire can be fed into your house through a window.

If you can not erect an outside antenna, run a length of insulated wire as long and high as possible through your house or apartment. It will work, but not as well as an outside antenna because of the shielding affect of the house walls.

Connect a ground lead to a cold water pipe. Make sure that it is a metal pipe. Or, you can buy a ground rod from your local electrical supply store and drive it into the ground near your house. Then connect your crystal set ground lead to the rod. For safety install a lightening arrestor between the antenna lead and ground (see the instructions with the device).

HEADPHONES

Headphones (also called earphones, or headset) are required to hear the received audio from a crystal set. The standard headphones are made with dual coils of fine wire connected in series in each phone. Horseshoe magnets are built into each phone case to provide a magnetic field. A disk of thin sheet iron (diaphragm) is mounted near the top of the coils. The phones are connected in series with a set of flexible leads terminated in metal "phone tips".

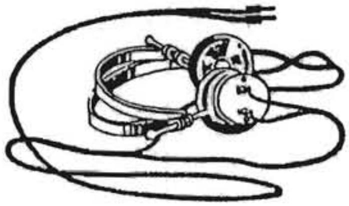

Detected audio signals flow through the coils and affect the intensity of the magnetic field from the horseshoe magnet. This changes the magnetic attraction of the diaphragm, and causes it to vibrate and produce audible sounds.

The total resistance of the series headphone coils is used as a measure of audio sensitivity. A resistance of 2000 ohms is sufficient. More sensitive headphones have a resistance of over 5000 ohms and are used to hear very weak signals.

The 2000 ohm headphones can be obtained at electronic parts mail order houses.

2

SIMPLE RECEIVER PROJECTS

SIMPLE CRYSTAL SET

Look at the schematic diagram. The crystal set uses a tunable ferrite coil L1, and a crystal diode D1. This type of tuning coil is called a "loopstick". The ferrite core is screwed into the coil form and increases the inductance of L1, thus tuning the receiver to a lower frequency.

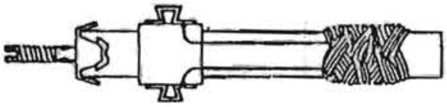

CONSTRUCTION

As shown in the illustration, the crystal set components are mounted on a 5 1/2-inch by 2 1/2-inch by 3/4-inch wood base. Layout and install the crystal set parts in the locations shown. The loopstick L1, is mounted with a metal strip supplied with it. Push the snap-in end of L1 into the large hole of the metal strip. Then bend 1/2-inch of the other end of the strip to form a bracket and mount it on the base with a wood screw. If the metal strip is not available, make one out of sheet metal approximately 2-inches long by 1/2-inch wide. Then drill a hole to fit L1 at one end, and a small hole to fit a wood screw at the other end.

Mount the four fahnestock clips and solder lugs with wood screws in the locations shown in the illustration.

WIRING

Use hookup wire to connect the crystal set components as shown in the schematic and illustration. The loopstick L1 may have three terminals. One of the terminals is a tap, which is not used in this circuit. Make sure that you connect only to the terminals at the start and end of the loopstick coil.

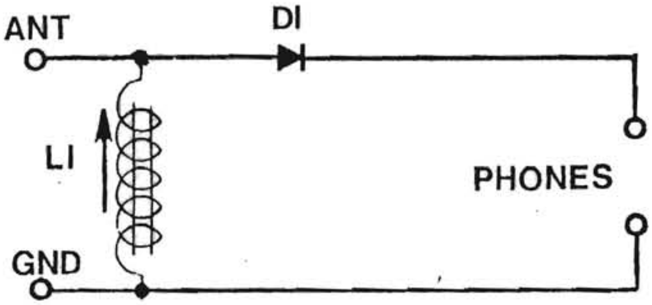
ANT
DI
LI
GND
PHONES

SIMPLE CRYSTAL SET

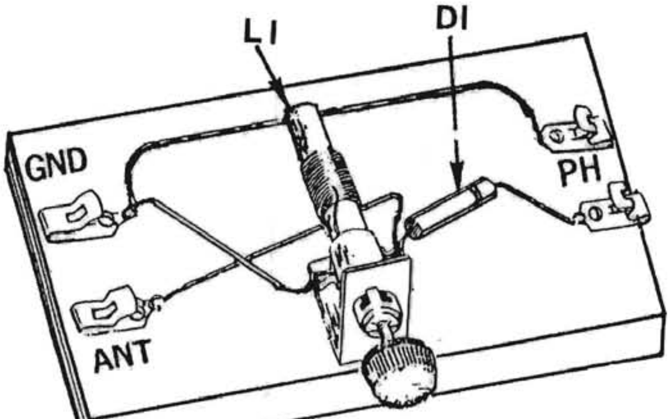
LI
DI
GND
PH
ANT

List of Materials

D1 - Germanium diode
(1N34A or equiv.)
L1 - Loopstick
(see text)
4 - Fahnestock clips
BASE - 5 1/2 X 2 1/2 X
3/4-inch wood section
PHONES - 2000-ohm headphones
MISC. - Solder lugs, wood screws,
knob and bracket to fit L1,
hook-up wire.

Install a small knob on the end of the L1 adjustment screw. Make sure that the screw turns freely with the knob. Position the 1N34A diode as shown before soldering the leads.

Experiment 1

Connect the antenna, ground, and 2000-ohm headphones to the crystal set terminals. Rotate the L1 knob counter-clockwise as far as it will go. Do not turn past this point, or you may break the ferrite rod inside the coil form. The ferrite is now outside of the coil winding. This is the highest frequency point of the tuning range of L1.

Slowly turn the L1 knob clockwise and listen for received radio stations. When the L1 knob is turned fully clockwise, the ferrite will be completly inside the coil (increasing the inductance) and it will be at the lowest frequency point of the tuning range.

Experiment 2

Connect capacitors, with values from 50 mmf to 300 mmf, in parallel with L1. Then tune L1 and note how the frequency range has changed.

AMPLIFIER CONNECTION

The simple Crystal Set (and the other receiver projects) can be connected to an amplifier in place of earphones. This amplifier can be a single channel input of your home stereo.

Parts values are not critical; C1 is approximately .005 mfd, and the volume control R1 can be any value from 1000 ohms to 10,000 ohms. Use shielded wire between R1 and the amplifier input jack.

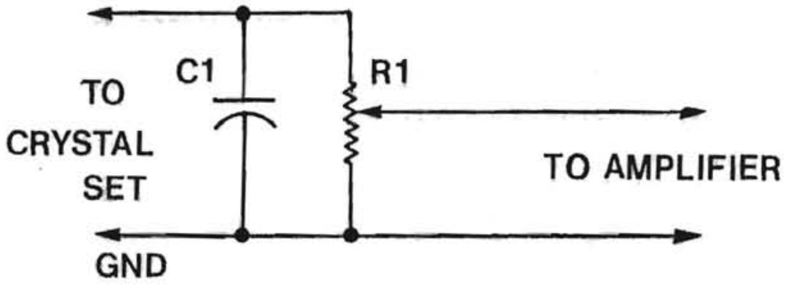
TO
CRYSTAL
SET
C1
R1
TO AMPLIFIER
GND

Amplifier Connection

Schematic Symbols

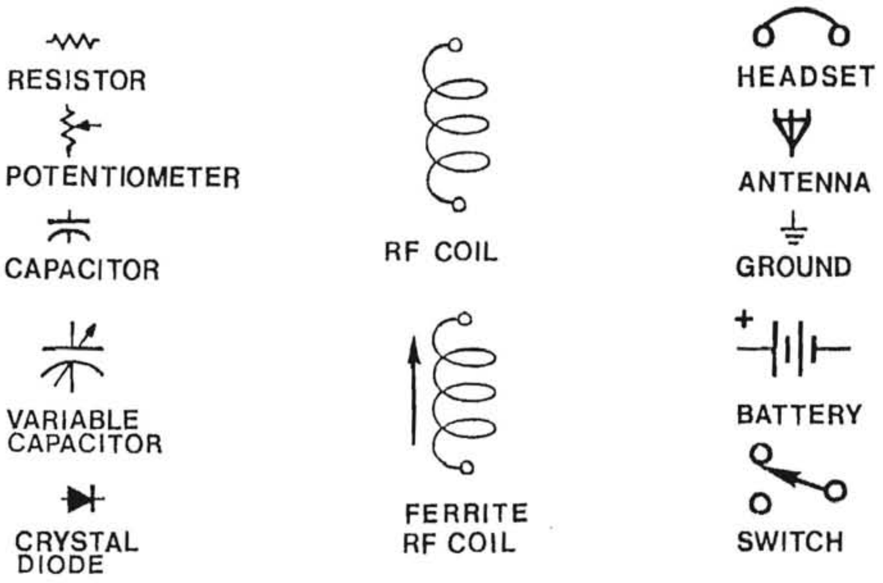
RESISTOR
POTENTIOMETER
CAPACITOR
VARIABLE CAPACITOR
CRYSTAL DIODE
RF COIL
FERRITE RF COIL
HEADSET
ANTENNA
GROUND
+
BATTERY
SWITCH

List of Materials

D1 - Germanium diode, type 1N34A, 1N60, 1N82A, or equiv.
L1 - Slider-tuned coil (see text)
4 - Fahnestock clips (terminals)
BASE - 6 x 3 x 1/2-inch section of wood (see text)
PHONES - 2000 ohms headphones
MISC. - Coil form (see text), No. 24 enameled copper wire, 3/32 inch dia. brass rod, brass ball slider (see text), solder lugs, two small brackets, hook-up wire, wood screws, machine screws and nuts.

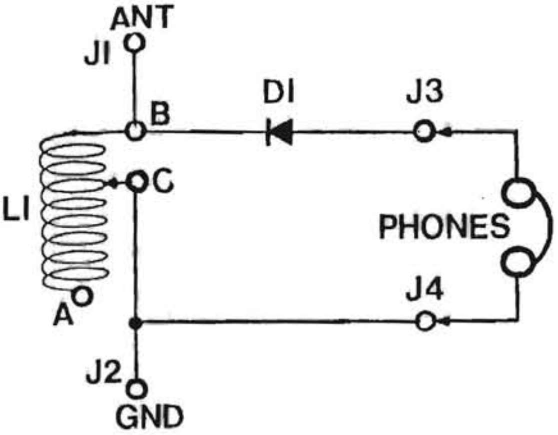
ANT
JI
B
DI
J3
C
LI
PHONES
A
J4
J2
GND

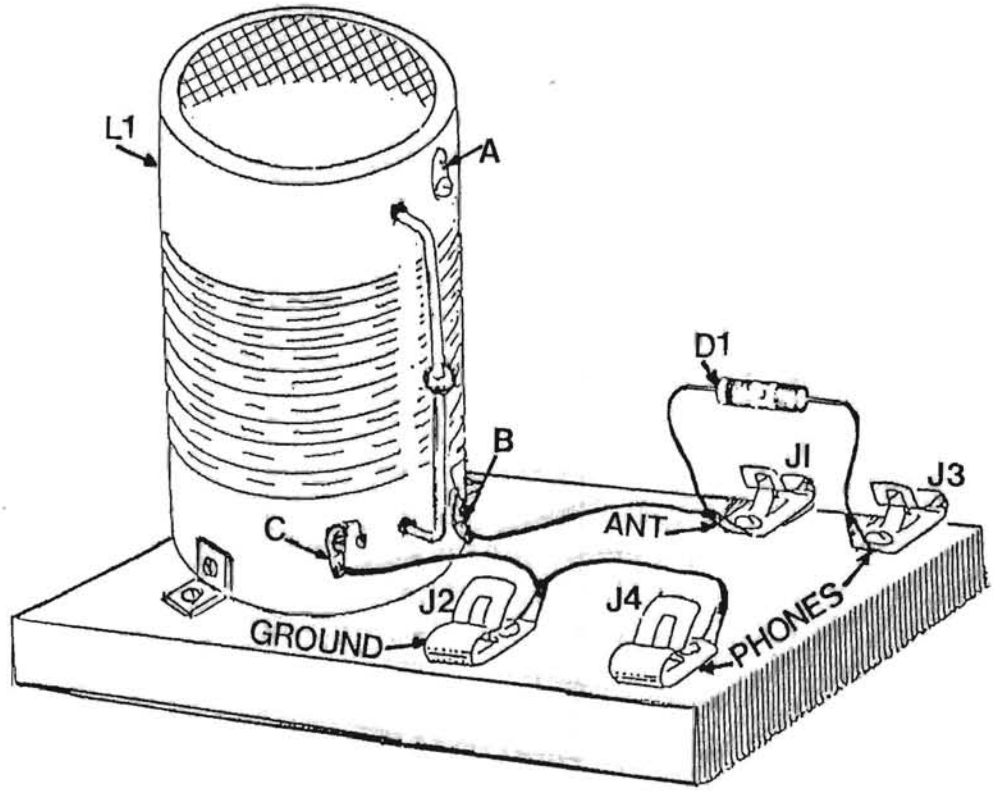
L1
A
D1
B
JI
J3
C
ANT
J2
J4
GROUND
PHONES

SLIDER TUNED CRYSTAL SET

SLIDER TUNED CRYSTAL SET

In the early days of radio, slider tuned coils were used in crystal sets. This method of tuning the inductance uses an adjustable metal contact (slider) that moves over the coil. The slider makes contact with individual coil wires to tune in the radio stations.

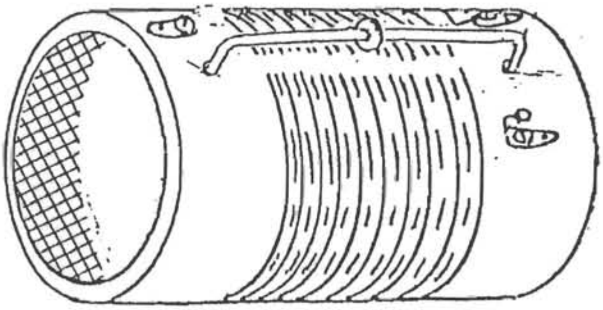

The enamel insulation of the coil wires in the slider path is removed for good electrical contact.

You can experiment with a slider tuned coil by building this crystal set. As shown in the illustration, the coil is mounted upright on a 6-inch by 3 1/2-inch by 1/2-inch wood base. A semiconductor diode is used for a detector.

TUNING COIL CONSTRUCTION

Look at the drawing of the tuning coil L1 construction details on the next page. The tuning coil is wound on a cardboard or plastic mailing tube section, 2 5/8-inches in diameter and 5-inches long. Make sure that you select a rigid, thick-walled tube for a coil form. The tube used in our model has a 3/16-inch wall. The coil form must be rigid to hold the slider ball and rod assembly.

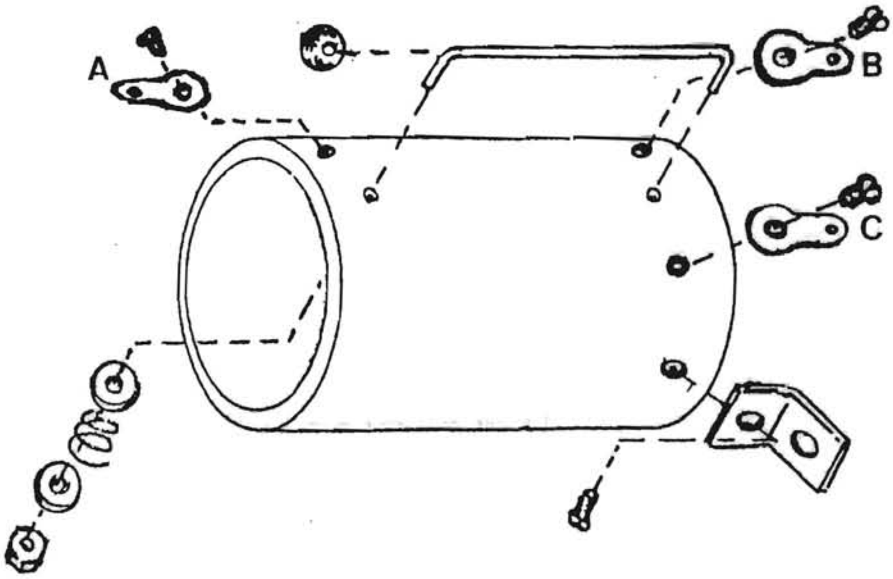

Before winding the coil, drill all the holes in the form as shown in the drawing, and then install the solder lugs. Position the solder lug ends toward the coil form edges.

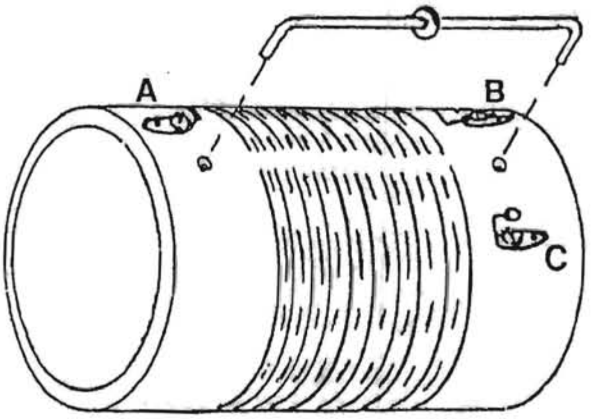

The coil is wound with No. 24 enameled copper wire. Remove the enamel from the wire end with sandpaper and then solder the wire to the lug. Start winding the coil, keeping the turns as tight and close together as possible. Wind the wire to the front of the bottom solder lug B. Remove the enamel and solder the wire end to the lug.

A brass ball approximately 3/8-inches in diameter is used as the moving contact slider. This brass ball can be found at your local electrical supply store. It is a type of ball that is used as a top nut for small table lamps. Drill a hole through the ball so that it will slide over a length of 3/32-inch diameter brass rod. Bend the brass rod at each end (after placing the ball in the center of the rod), to fit through the holes in the coil form as shown in the drawing. Cut the rod 1-inch down from the bend at each end.

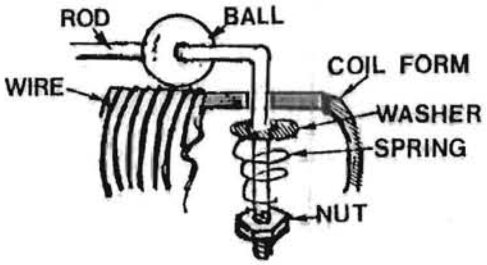

Insert the brass rod with the slider ball into the coil form holes. Then place a small washer and spring on one end of the rod (inside the coil form). Solder a nut on the end of the rod to hold the spring in place. Do the same to the other end of the rod. After soldering, slide the ball over the wires and check that it slides smoothly. Use sandpaper to remove the enamel insulation of the wires in the path of the ball slider. Connect a wire between the soldered nut on the bottom end of the brass rod and the solder lug near the rod.

BASEBOARD

All of the crystal set components are mounted on a wood baseboard as shown in the drawings. Cut the baseboard to size; a 6-inch by 3 1/2-inch section of 1/2-inch thick wood (maple is used in our model). Sand the edges after cutting. It is not necessary to paint or stain the baseboard.

Mount the four terminals and solder lugs as shown in the drawing. Label the four terminals with either rub-on lettering or cemented paper tags. Install two brackets on opposite sides of the bottom of the coil, and mount it on the baseboard with the lugs toward the terminals.

WIRING

Bare No.22 wire was used to wire our model, but any kind of hookup wire can be used. Connect the wires between the solder lugs as shown in the drawing, except do not connect the antenna terminal lug wire to the coil terminal at this time. Connect the crystal diode between the solder lugs as shown.

Experiment 1

This experiment will show how an untuned crystal detector functions.

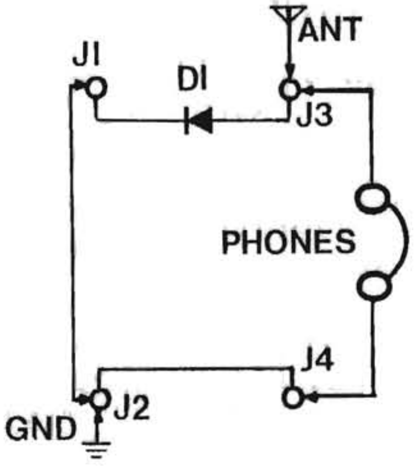

Connect the untuned detector circuit as shown in the drawing, and then hook-up your antenna and ground leads. Connect 2000-ohm headphones to the circuit and listen to the detected radio waves.

You will probably hear a strong local radio station, with other, weaker radio station transmissions all jumbled together.

Experiment 2

Reconnect the baseboard wires to the circuit of the Slider-Tuned Crystal Set as shown in the drawing. Listen to the detected radio waves as you tune the coil by sliding the brass ball over the sanded wire path. You should be able to receive the radio stations much better than the untuned detector circuit of Experiment 1.

LOADING COIL

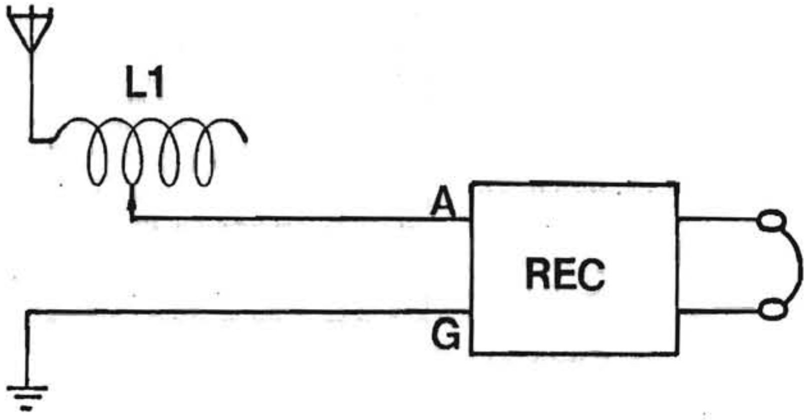

An additional slider tuned coil, L1, can be used as a Loading Coil to increase the electrical length of your antenna. This may allow better reception of the lower broadcast band frequencies. Connect L1 in series with the antenna lead to the receiver, and tune it for best reception.

WAVE TRAP

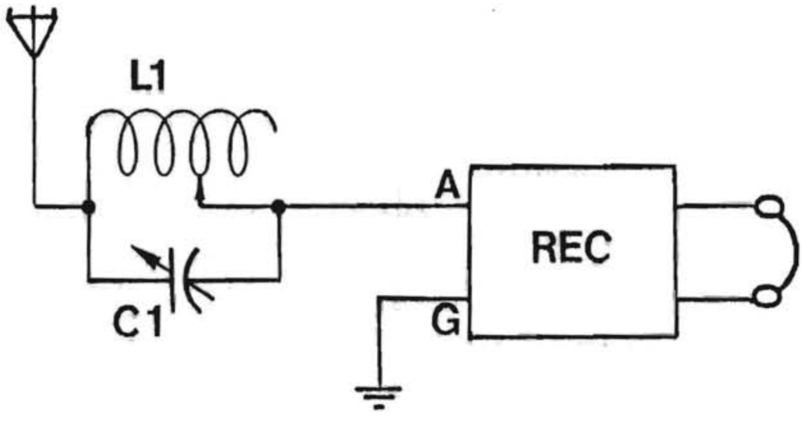

L1 can also be used in a Wave Trap circuit to tune out a strong local signal, and allow easier tuning of weaker stations. C1 is a 365 mmf variable capacitor connected in parallel with L1. Adjust L1 and C1 so that the interfering station is at minimum.

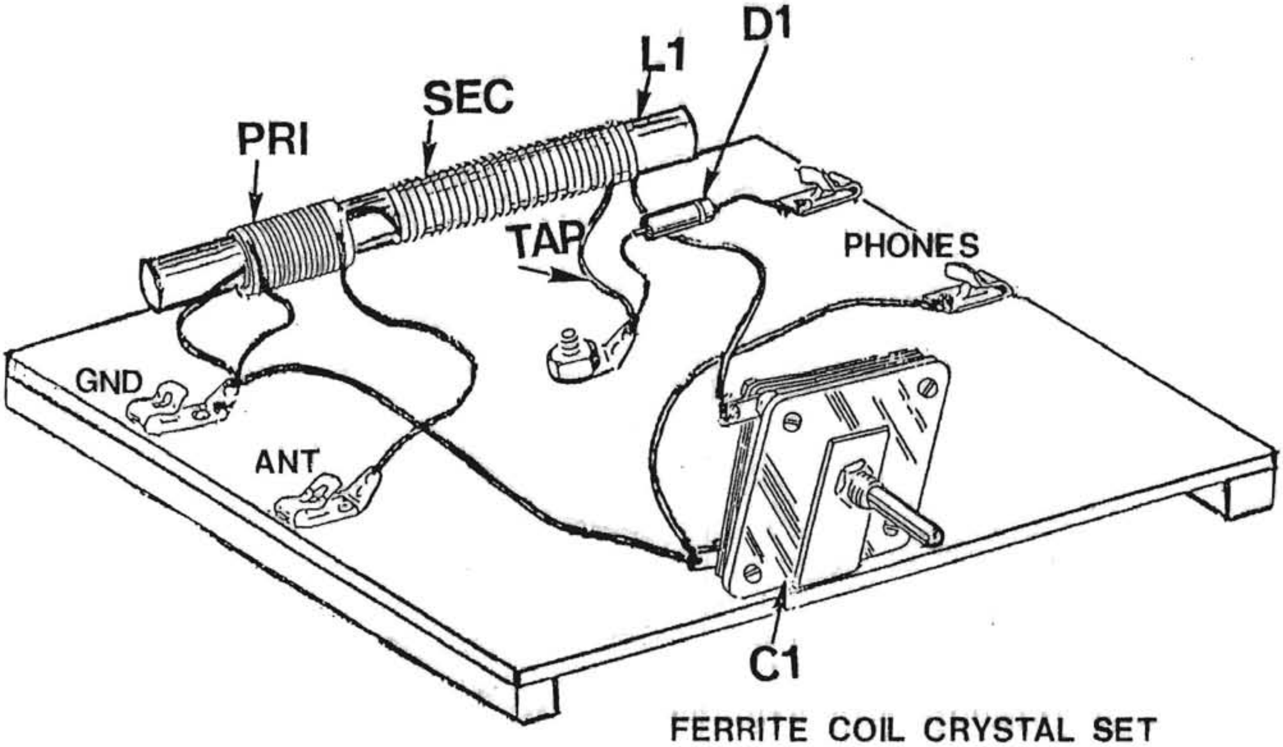
D1
L1
SEC
PRI
TAP
PHONES
GND
ANT
C1

FERRITE COIL CRYSTAL SET

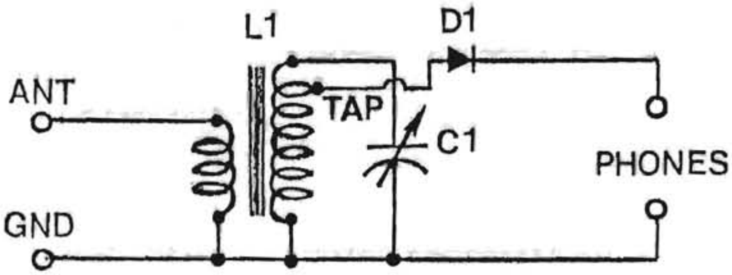
L1
D1
ANT
TAP
C1
PHONES
GND

List of Materials

C1 - 365 mmf variable capacitor.
D1 - Germanium diode (1N34A or equiv.).
L1 - Ferrite coil (see text).
4 - Fahnestock clips.
2 - wood sections; 6 X 1/2-inch square (see text).
BASE - 8 1/2 X 6-inch sheet plastic (see text).
PHONES - 2000-ohm headphones
MISC. - plastic tube section for L1 primary winding (see text), solder lugs, machine screws and nuts, wood screws, bracket for C1 (see text), knob for C1, hookup wire.

FERRITE COIL CRYSTAL SET

This crystal set has a large ferrite core coil L1 of a type that is normally used for a built-in antenna in portable radios. This type of coil has a very high "Q" and may be more selective than an air-core coil. As you can see in the schematic, a primary coil is wound around the ferrite coil. The original winding is used as the secondary.

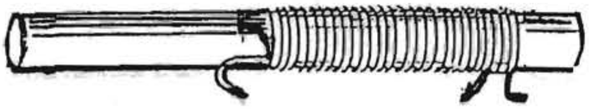

The primary coil (connected to the antenna and ground) can be moved close to, or away from, the secondary (original winding connected to D1). This variable coupling is used to control the selectivity of the crystal set.

PRIMARY COIL

The primary coil is wound on a 1-inch section of plastic tube that can fit loosely over the ferrite rod. A 1/2-inch diameter plastic tube is used on the receiver shown in the illustration. Wind 12-turns of hookup wire evenly over the tube and twist the ends together to prevent movement. Leave 4-inch leads from the primary coil winding. Slide the primary coil over the center of L1.

BASEBOARD

The crystal set components are mounted on a 8 1/2-inch by 6-inch section of sheet plastic. The thickness of the plastic is not critical, but it must be rigid enough to hold the parts without bending. A 6-inch long 1/2-inch square wood section is installed under each end of the baseboard. A section of wood can be used for the baseboard with wood screws to mount the parts, in place of machine screws and nuts.

The model shown uses a plastic variable tuning capacitor, but an air type capacitor can be used for C1 instead. Make a metal bracket approximately 1-inch wide by 1 1/4-inch high and a 1/2-inch right angle mounting foot from sheet aluminum. Drill a hole in the top of the bracket to fit C1 and two holes in the mounting foot to fit small machine screws. Locate the components on the baseboard as shown in the illustration and then drill mounting holes. If you have a cardboard or plastic mounting supplied with ferrite coil L1, you can mount it on the baseboard. If not, L1 can be taped loosely on top of the baseboard. Th ferrite coil L1 in our crystal set was obtained from an old radio. Install the components on the board with solder lugs, machine screws, and nuts.

WIRING

Connect the parts with hookup wire as shown in the illustration. Use the schematic diagram as a guide. Make sure that the ground lead from L1 secondary passes inside the primary coil. Also check that the primary coil slides freely over the L1 winding. Install a knob on the C1 shaft, and mark the terminal designations on the baseboard.

Experiment 1

Hookup the antenna, ground and 2000-ohm headphones to the crystal set terminals. Slide the primary coil over the center of the L1 winding. Tune C1 for radio stations. Then slide the primary coil towards the ground end of L1 while tuning C1. Note the difference in selectivity.

Experiment 2

Change the number of turns of the primary winding and repeat Experiment 1.

Operating Tips

After operating a crystal set for some time, you may notice a difference between day, evening, and night reception of radio stations. This is due to the two ways that radio signals travel between the transmitting antenna and a receiving antenna.

(a) The ground wave; which moves along the surface of the earth.
(b) The sky wave; which travels upward and is reflected by the ionosphere.

The reception of broadcast stations is dependent upon the time of day. During daylight hours, stations are received primarily by ground wave transmission. At night, the broadcast stations will be received via sky wave transmission. Also, some radio stations change to a lower power at night. Reception in the evening may be more difficult as the fading ground wave is interfered with by reflected waves. But later, it may be possible to hear far-away stations transmitted by sky wave.

Your geographical location to radio stations is very important for reception, as a crystal set receiver does not amplifiy signals. It is entirely dependent upon the strength of the radio waves. If you are in a deep valley surrounded by high mountains, reception will be difficult. If you live in a flat plain, with no obstacles to radio transmission, reception may be very good.

An outside antenna is necessary for for best reception. If possible, try erecting an extra antenna at a right angle to the original one, and switch it in to provide a directional effect for increased signal strength. Keeping a log of received stations for a week may be helpful in determining the radio reception conditions in your area.

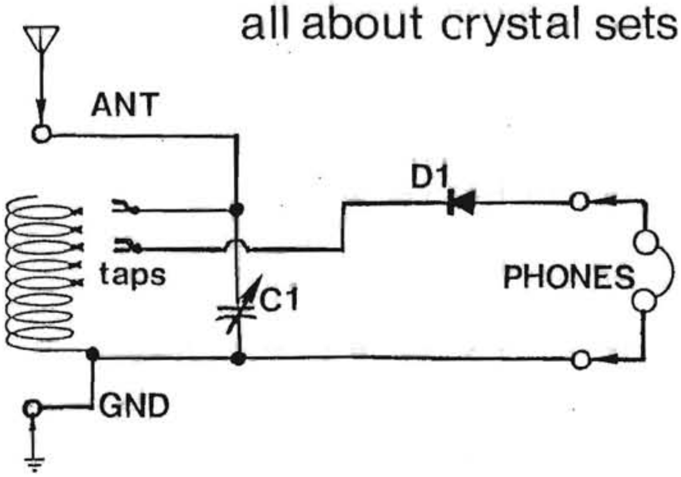
ANT
D1
L1
taps
C1
PHONES
GND

List of Materials

C1 - 365 mmf variable capacitor.
D1 - Catwhisker crystal detector
(or 1N34A or equiv. diode).
L1 - Spiderweb coil (see text).
4 - Fahnestock clips.
2 - small alligator clips
BASE - 8 X 6 X 1/2-inch section
of wood (see text).
MTG. POST - 5 x 3/4 x 3/8-inch
wood section (see text).
PHONES - 2000 ohms headphones.
MISC. - Coil form (see text),
No. 22 enameled copper wire,
solder lugs, hook-up wire,
wood screws, machine screws,
and nuts.

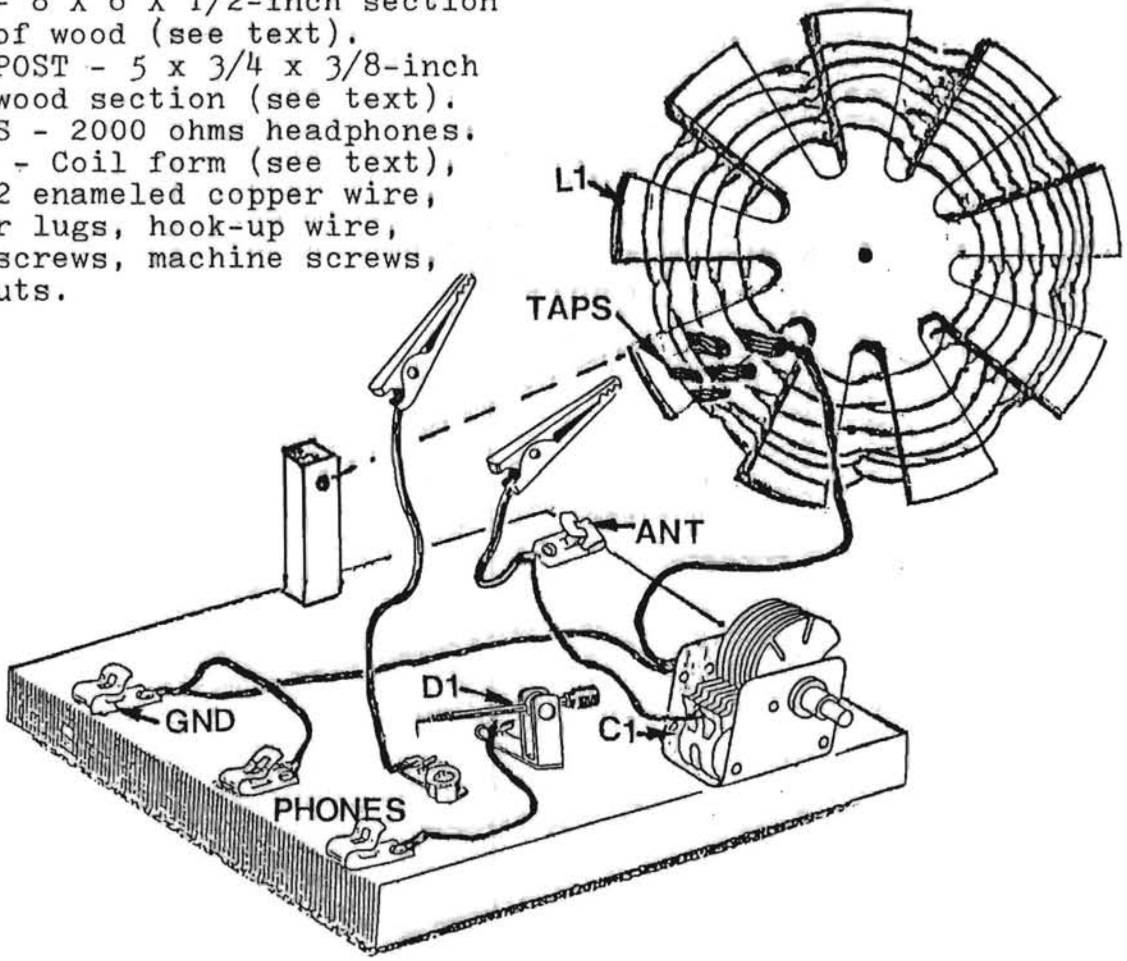
L1
TAPS
ANT
D1
C1
GND
PHONES

SPIDERWEB COIL CRYSTAL SET

SPIDERWEB COIL CRYSTAL SET

The spiderweb coil is a tuning coil that is wound on a flat wheel form that has an uneven amount of spokes. This type of coil was also used as a loop antenna in early radios.

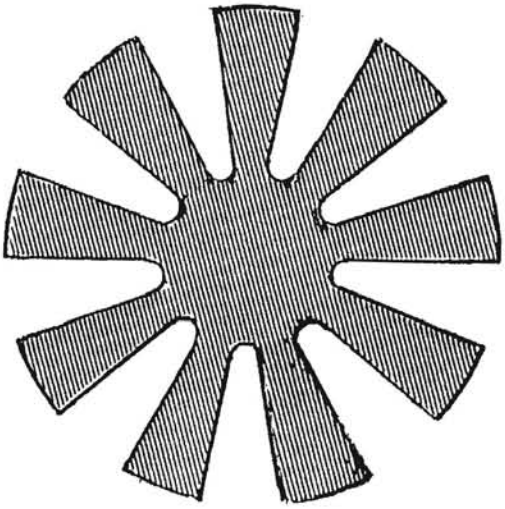

Our receiver uses a tapped spiderweb coil that is tuned by a variable capacitor. As shown in the illustration, the spiderweb coil is mounted at the rear of a 8-inch by 6-inch by 3/4-inch wood base. A catwhisker crystal detector and a variable capacitor are mounted near the front of the wood base.

SPIDERWEB

The spiderweb coil is wound on a 6-inch diameter flat form that has nine vanes. Make a tracing of the spiderweb coil form template on the next page. Then use carbon paper to transfer the tracing to a section of 1/8-inch fibreboard or sheet plastic. Cut the coil form out with a jeweler's saw, and drill a center hole to fit a small wood screw.

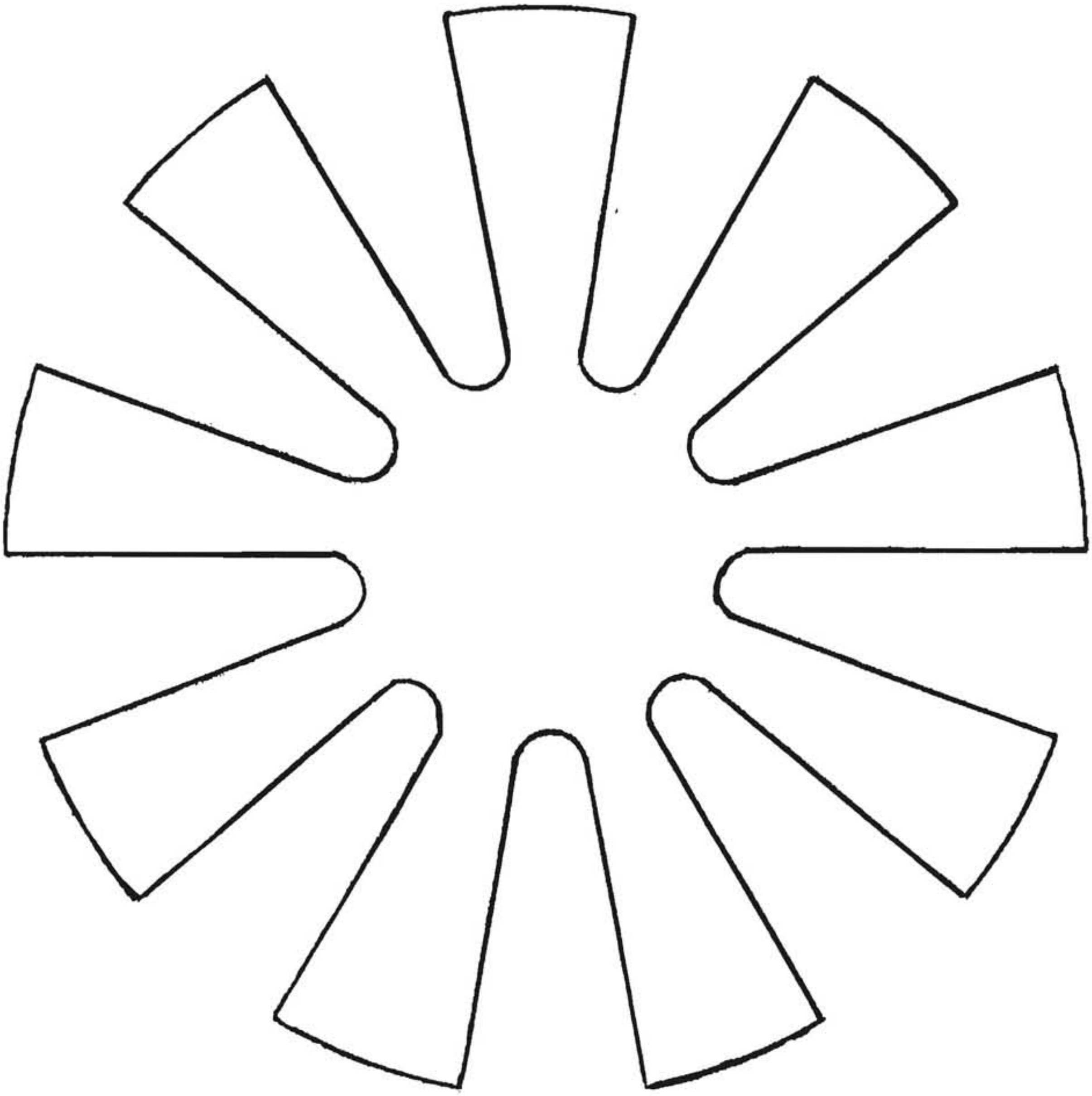

Spiderweb Coil Form Template

Wind the coil with No. 22 enameled copper wire, starting from the center and winding outward. Leave about 5-inches of the starting end of the wire free. It will be used as a connecting lead. Feed the wire behind a vane and over the front of the next vane as you wind. Wind the wire closely and evenly on the form. It is not necessary to count the turns. Wind the wire to about 1/4-inch from the end of a form vane. Then drill a small hole near the vane end, and bend a solder lug through it. Solder the wire end to the lug.

On one of the vanes, count down five wires and then carefully lift it outward. Remove about 1/4-inch of the enamel with a section of sandpaper and then bend the end of a solder lug aound the bared wire. Solder the lug to the wire, then cut off remaining portion of the lug (away from the bentover end). Repeat this every five turns until you have a total of six lug taps on the coil. Stagger the lugs away from each other to keep them from shorting.

BASEBOARD

As shown in the illustration, the Spiderweb Coil Crystal Set parts are mounted on a wood baseboard. Cut the 8-inch by 6-inch baseboard from a section of 1/2-inch thick wood. White pine was used in our model. Sand the surface of the wood after cutting.

Make the spiderweb coil mounting post by cutting out a 5-inch by 3/4-inch by 3/8-inch wood section. Then cut a notch in the center of the rear of the baseboard to fit the wood section. Install the wood section in the notch with small wood screws.

Mount the parts on the baseboard as shown in the illustration with small wood screws. If the mounting holes of the tuning capacitor C1 are on the bottom, drill matching holes in the baseboard and use flathead machine screws to hold the capacitor.

Install solder lugs with the four terminals, and mount the spiderweb coil to the mounting post with a wood screw through a center hole in the coil form. The coil should be positioned with the lugs facing forward (on the left side of the coil form).

WIRING

Connect the crystal set components with hookup wire as shown in the illustration. Use 6-inch lengths of insulated stranded wire to the two small alligator clips.

Experiment 1

Connect antenna, ground, and 2000-ohm headphones to the crystal set. Then connect the D1 alligater clip lead to the antenna terminal. Set the tuning capacitor C1 to minimum capacity (rotor blades all the way out). Connect the antenna clip lead to the tap on L1 near the center of the form.

Adjust the crystal detector D1 catwhisker to a sensitive spot and tune C1 for a radio station. It may be easier to tune for stations if you temporarily connect a crystal diode (1N34A or equiv.) in place of the catwhisker crystal detector.

After a station is heard, you can disconnect the diode and probe for a sensistive spot on the crystal with the catwhisker. Change the antenna clip lead to other L1 taps and continue tuning C1 for radio stations.

Experiment 2

Disconnect the D1 clip lead and reconnect it to a tap on L1 close to the center of the coil. This will load L1 more lightly than if the detector was connected to the top of the coil. Repeat the tuning procedure of Experiment 1 and note the difference in selectivity and volume of the received radio stations. Then tune in a radio station and move the D1 clip lead to other taps on L1. There will be a particular tap point for optimum selectivity and volume for each received station.

STATION LOG

DATE	TIME	DIAL	STATION	NOTES

3 ADVANCED RECEIVER PROJECTS

DUAL TUNED CRYSTAL SET

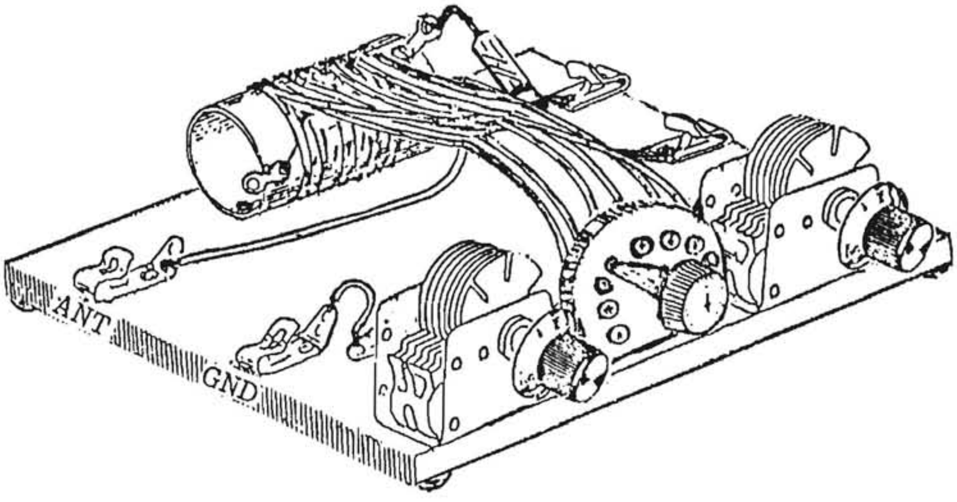

This crystal set circuit uses two variable capacitors C1 and C2. As you can see in the schematic drawing, C1 tunes the ground circuit, and C2 tunes a tapped coil L1. The taps of L1 are connected via the rotary switch S1.

TUNING COIL

Look at the drawing of the coil form construction details on page 34. The tuning coil L1 is wound on a section of 1/8-inch wall plastic tube, 2-inches in diameter, and approximately 4-inches long.

Drill holes in the form and install three solder lugs with machine screws and nuts. Position the solder lug ends towards the coil form ends as shown.

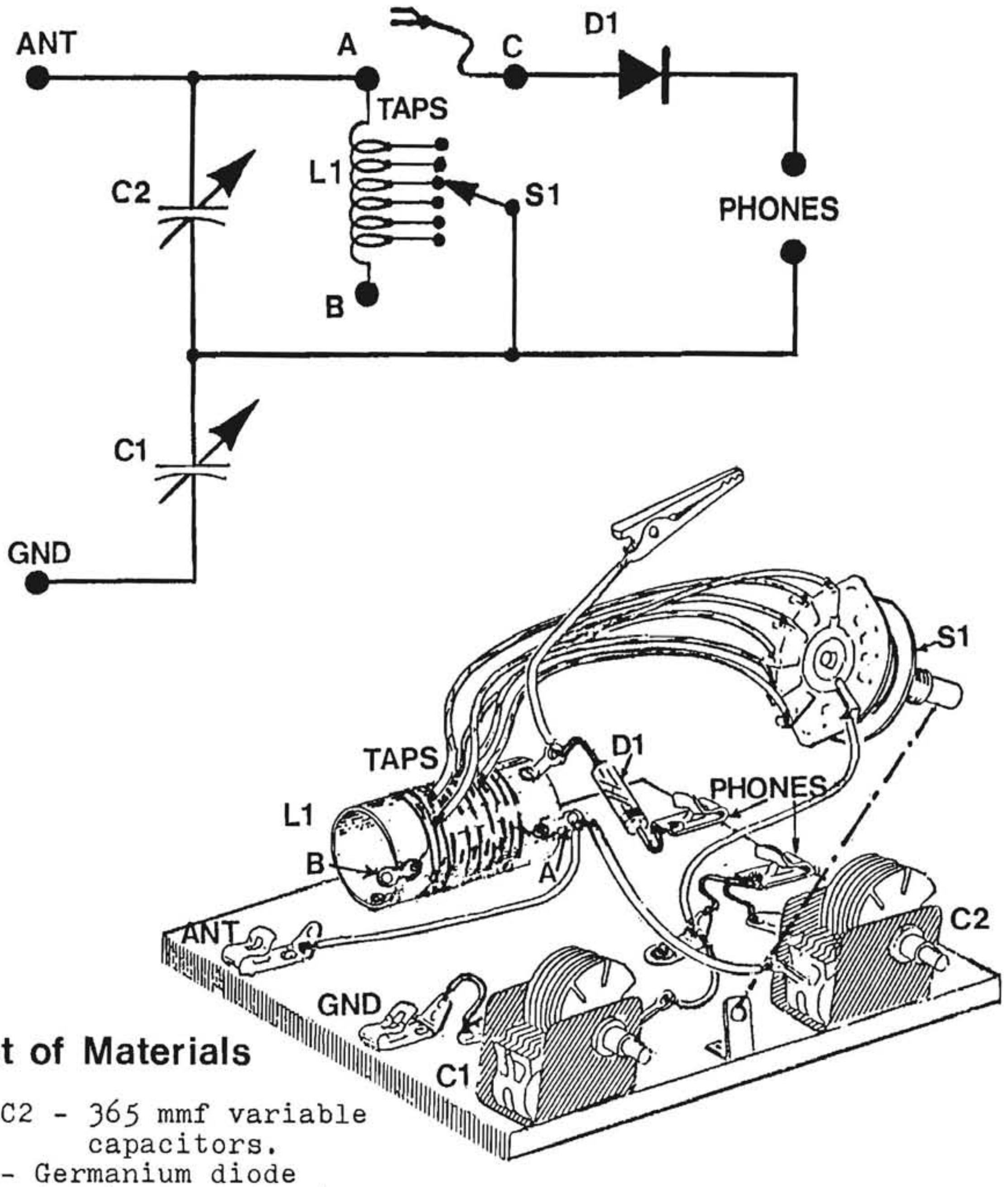
ANT
A
C
D1
TAPS
L1
S1
C2
PHONES
B
C1
GND
S1
TAPS
D1
PHONES
L1
B
A
C2
ANT
GND
C1

DUAL TUNED CRYSTAL SET

List of Materials

C1,C2 - 365 mmf variable capacitors.
D1 - Germanium diode (1N34A or equiv.).
L1 - Tuning coil (see text).
S1 - 1-pole, 10-position rotary switch (see text).
4 - Fahnestock clips.
1 - Small alligator clip.
BASEBOARD - 5 x 8 x 3/8-inch wood section (see text).
PHONES - 2000-ohm headphones.
MISC. - Coil form (see text), No. 24 enameled copper wire, solder lugs, hook-up wire, wood screws and nuts, knobs, bracket for S1.

The coil L1 is wound with No. 24 enameled copper wire, and is tapped every five turns for a total winding of 75 turns.

Drill a small hole in the coil form approximately 1 1/4-inch from the "A" lug. Remove the enamel from the end of the No. 24 wire and then feed the wire end through the hole. Solder the wire to lug "A".

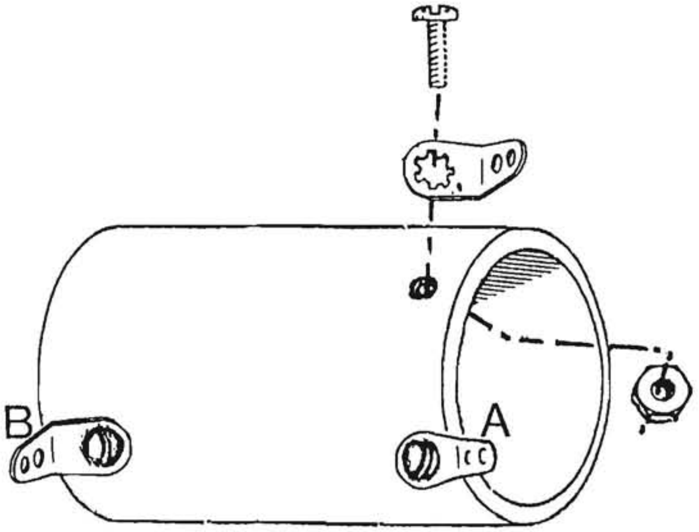

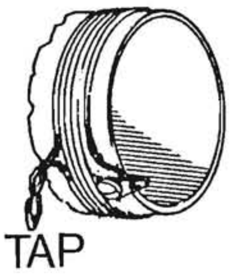

Wind five closely spaced tight turns around the coil form. Then bend back and twist together 1/2-inch of the wire to make a tap. Position the tap outward and then continue winding the coil, keeping the turns as tight and close together as possible. Make a tap every five turns and then drill a small hole in the coil form after the 75th turn. Feed the wire through the hole, remove enamel, and solder the end to lug "B".

BASEBOARD

The Dual-Tuned Crystal Set is built on a wood baseboard as shown in the illustration. Cut the 5-inch by 8-inch baseboard from a section of 1/4-inch thick plywood. Sand the edges and surface of the baseboard to remove any splinters. Drill holes in the baseboard and install the parts with small machine screws and nuts in the locations shown in the illustration. Drill mounting holes at each end of the tuning coil L1 and position it on the rear of the baseboard with the coil taps toward the front. The illustration shows a commercial rotary switch, but a wooden rotary switch can be used for S1 instead. (See page 53 for construction details.)

Install the four fahnestock clips (used as antenna, ground and headphone terminals) with solder lugs. Mount an additional solder lug on the baseboard midway between S1 and L1.

WIRING

Connect the crystal set parts with hookup wire as shown in the pictorial illustration and the schematic diagram. Use the center solder lug as a junction point for leads from C1, C2 and S1 as shown in the illustration. Connect the leads from S1 to the taps on L1 in numerical sequence; every 5 turns for 7 taps. Then connect the last three leads every 10 taps. Insert the lead ends into the taps and hold them with a drop of solder. This will make it easier to move leads to other L1 taps. Use a 5-inch length of stranded hookup wire for the alligator clip lead connected to D1 at solder lug C.

Experiment 1

Connect 2000-ohm headphones, antenna and ground leads, to the crystal set terminals. Then connect the D1 clip lead to the L1 "A" solder lug.

Set S1 to each tap position in turn and tune C1 and C2 for received radio stations. If all of the local stations can not be heard, experiment with different connections to the L1 taps. More turns may have to be added to L1 (with additional five-turn taps) to receive stations at the lower frequency end of the broadcast band.

Experiment 2

Remove the D1 clip lead from "A" and then connect it to the "B" solder lug on L1. Repeat the tuning method of Experiment 1 and note if the crystal set is more selective. This is possible, because L1 is now connected as an rf transformer. You can also connect the D1 clip lead to each of the S1 taps and note if there has been any improvement in reception.

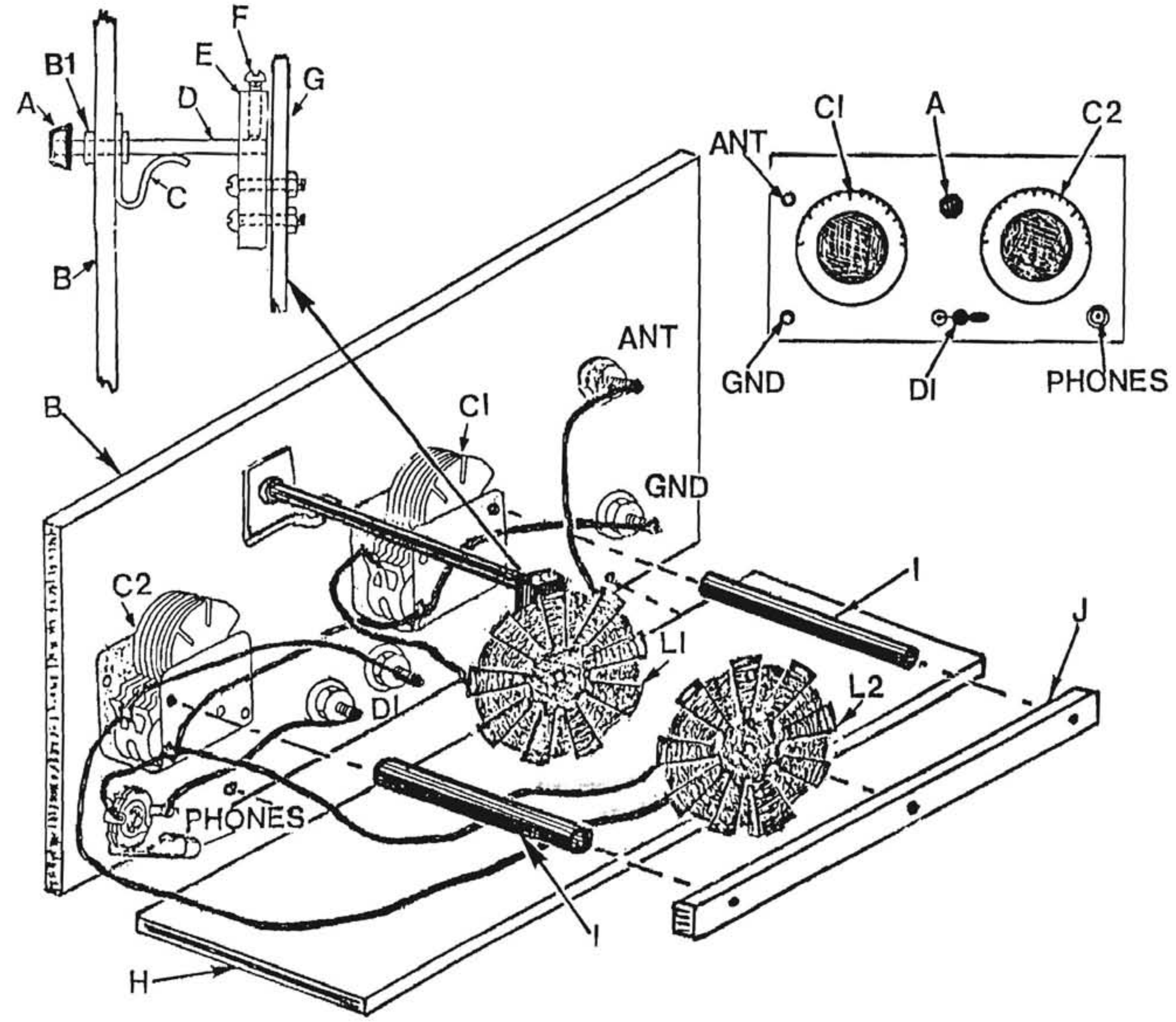
F
B1
E
G
A
D
C
B
CI
A
C2
ANT
GND
DI
PHONES
B
ANT
CI
GND
C2
I
J
LI
L2
DI
PHONES
I
H

SELECTIVE CRYSTAL SET

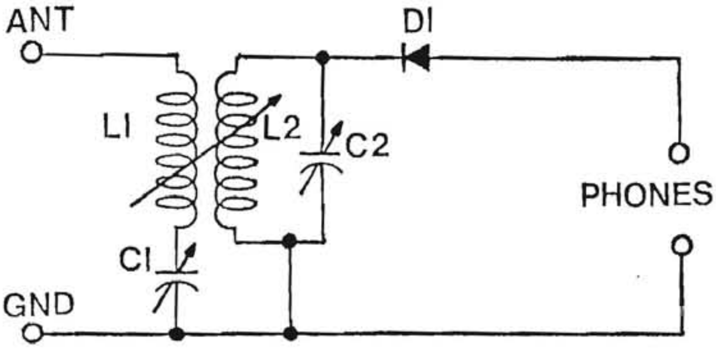
ANT
DI
LI
L2
C2
PHONES
CI
GND

SELECTIVE CRYSTAL SET

This receiver uses two spiderweb coils in a two-circuit tuner for good selectivity and sensitivity. As you can see in the schematic diagram, the antenna-ground rf input circuit is tuned by L1-C1 for maximum signal gain. L1 acts as the primary winding of an rf transformer, with L2-C2 as the tuned secondary. Coupling between L1 and L2 is variable for best selectivity. L2 is connected to a catwhisker crystal detector and the audio signals are fed through the phone jack to 2000-ohm headphones.

As shown in the illustration, the crystal set parts are mounted on the rear of a black plastic panel. This type of layout is similar to the type of construction of the old-time crystal sets.

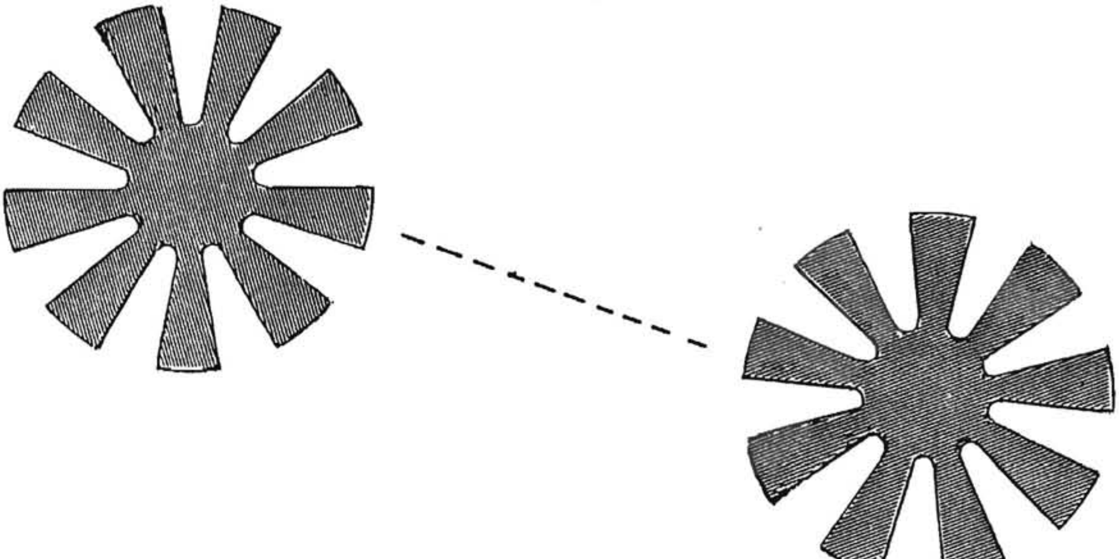

SPIDERWEBS

Both L1 and L2 spiderweb coils are made exactly the same way. L1 and L2 are wound on 17-vane coil forms 3 5/8-inches in diameter. The coil forms in our receiver are made from 1/16-inch sheet plastic of the same type used for printed circuit boards (but without the copper foil). Any kind of rigid sheet plastic should be suitable for the coil forms.

Make a tracing of the spiderweb coil template. Then use carbon paper to transfer the tracing on a section of 1/16-inch thick sheet plastic. Cut out the vanes with a jeweler's saw. Drill two holes as shown, and install the two solder lugs.

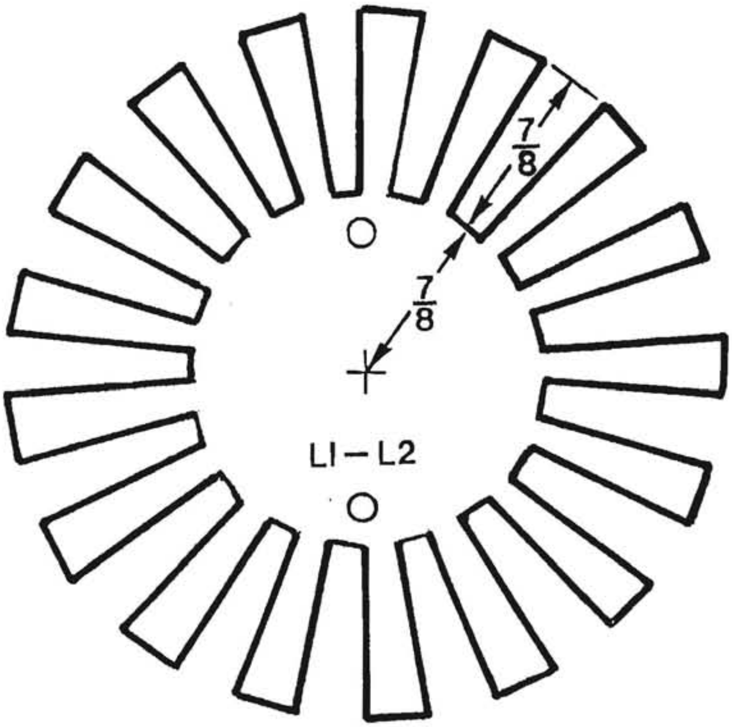

COIL FORM TEMPLATE FOR L1–L2

Wind as much No.22 enameled wire as possible on each coil form. Solder the wire to a solder lug before winding a coil. Then wind the wire over a vane and under the next vane. Wind the wire closely and evenly on the form. It is not necessary to count the turns. Solder the wire end to the remaining solder lug

CONSTRUCTION

The panel used in our receiver is a 7 3/4-inch by 4 1/2-inch section of 1/4-inch thick black acrylic plastic. The size is not critical, but it should be large enough to mount all of the parts as shown in the illustration.

Layout and drill the mounting holes of the components on the panel. Make sure that C1 and C2 are positioned so that the edges of the tuning knob dials are more than 1/2-inch apart. Install the catwhisker crystal detector assembly, antenna and ground terminals, and the phone jack in the locations shown.

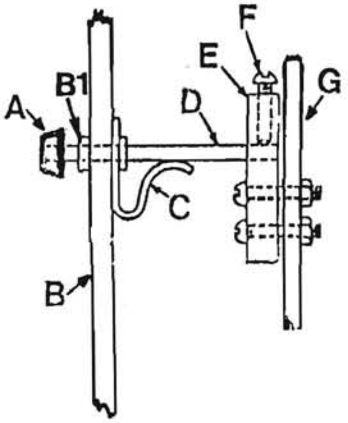

As you can see in the illustration, L1 is installed on the end of a variable coupling metal rod. This 1/4-inch diameter rod is 3-inches long with a flat side. Drill and tap a small hole close to the end of the rod. The rod used in our model came from a discarded volume control. Cut a 2-inch long by 1/2-inch wide by 3/8-inch thick wood section and drill holes to fit the rod, two mounting screws, and a hole for the rod set screw. Locate and mount the wood section in the center of a spiderweb coil L1 so that the rod will be 3/4-inch down from the edge of the top vane.

Hold L1 behind the panel and using the free end of the rod, locate the panel hole. Position L1 midway between C1 and C2, and make sure that the top edge of the coil is below the top of the panel. Drill a hole to fit the rod bushing in the panel hole location. The panel bushing in our model was removed from a discarded volume control.

Drill a hole to fit the bushing in one end of a 2-inch by 1/2-inch section of sheet aluminum. Bend the other end of the section to fit against the flat side of the rod. Install the metal section and the bushing in the panel hole.

Insert the rod into the bushing and make sure that the metal section end fits close to the flat side to prevent rotation of the rod. Install a knob on the rod end and make sure that the rod can be pushed in and out easily. If necessary, put a small amount of grease on the flat side of the rod.

Install the remaining spiderweb coil L2 on a wood section that is mounted on the rear of the panel with two spacers. The spacers should be long enough for L2 to be located 1/4-inch behind L1 with the rod pushed all the way in. The 5 1/2-inch by 1-inch by 1/4-inch wood section in our model is supported with screws installed through 1 1/2-inch metal spacers into tapped holes in the rear of C1 and C2. The screws and bushings can be made longer and mount into panel holes in your receiver.

Install a wood section on the bottom rear of the panel to keep the receiver upright. A 5-inch by 3-inch section of 1/4-inch thick wood is used in our receiver.

SET WIRING

Connect the receiver components with hookup wire as shown in the illustration and schematic diagram. Use flexible stranded wire for connections to L1. Make sure that the L1 leads do not touch any component as the variable coupling rod is moved in and out.

Install knobs on C1, C2, and the variable coupling rod. Then place the crystal in the detector assembly.

OPERATION

Connect antenna and ground leads to the receiver terminals. Plug in a set of 2000-ohm headphones into the panel phone jack. For easier initial operation of the receiver, temporarily connect a germanium diode (1N34A or equivalent) across the D1 connecting lugs. Make sure that the catwhisker is not touching the crystal at this time.

Push the variable coupling rod half way in and tune C1 and C2 for a station. Adjust the variable coupling rod and retune C1-C2 for best reception. Check operation over the entire broadcast band. It may be necessary to add or subtract turns of wire on L1-L2 for complete coverage. After checking operation with the germanium diode, remove it and try using the catwhisker crystal detector.

List of Materials

C1,C2 - 365 mmf variable capacitor.
D1 - Catwhisker crystal detector assembly.
L1,L2 - Spiderweb coils (see text).
2 - Terminals.
1 - Phone jack.
PHONES - 2000-ohm headphones.
MISC.- Plastic sheet for spiderweb coils, solder lugs, machine screws and nuts, wood screws, hookup wire.
CONSTRUCTION MATERIALS (See Text).
A - knob
B - Panel
B1- Bushing
C - Aluminum section
D - Rod
E - Wood section
F - Setscrew
G - L1 coil form
H - Wood section
I - Spacers (2)
J - Wood section

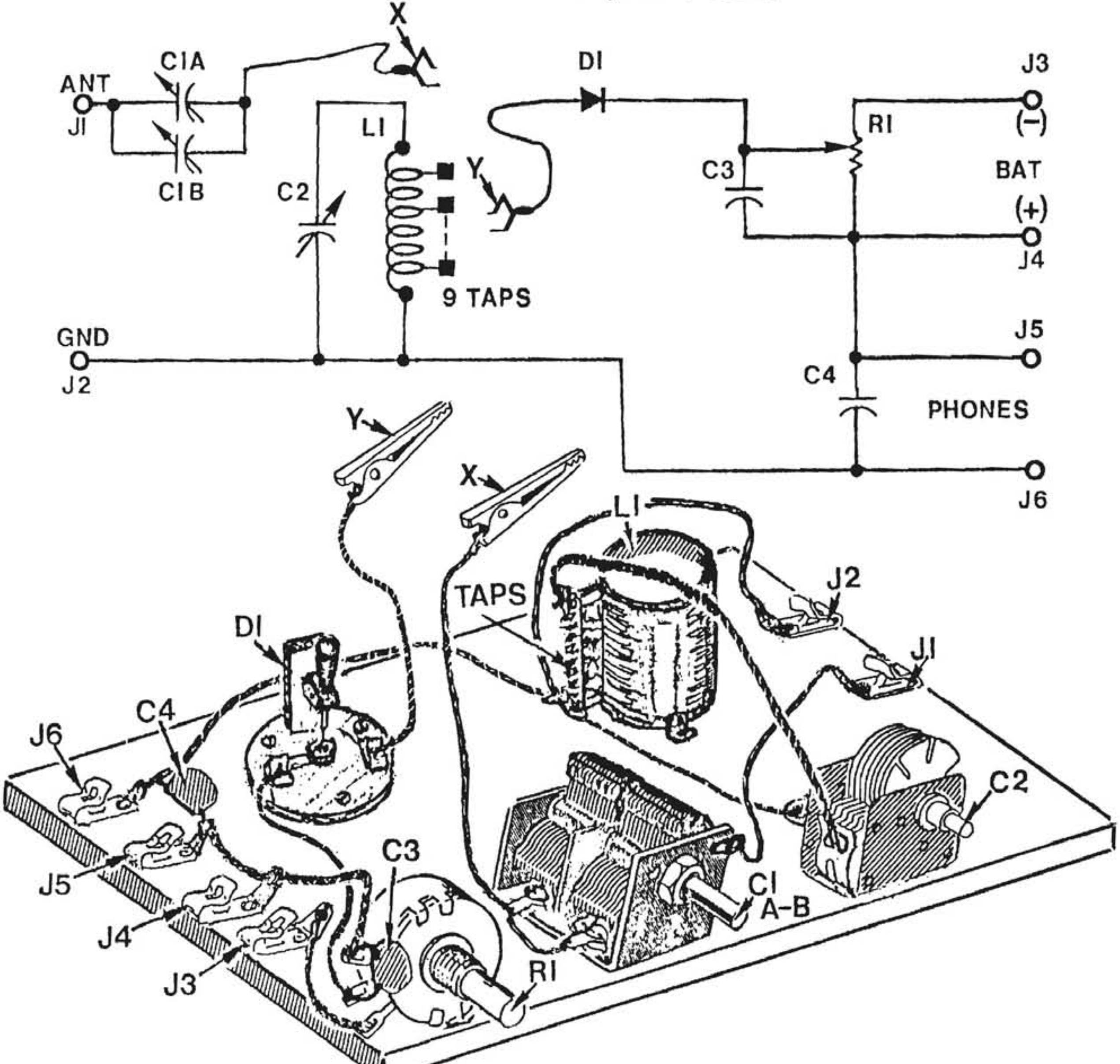
X
C1A
D1
J3
ANT
J1
C1B
C2
L1
Y
R1
(-)
C3
BAT
(+)
J4
9 TAPS
GND
J2
J5
C4
PHONES
J6
Y
X
L1
TAPS
J2
J1
D1
C4
J6
C2
C3
C1 A-B
J5
J4
J3
R1

List of Materials for Carborundum Crystal Set

C1A-C1B - Dual 365 mmf variable capacitor.
C2 - 365 mmf variable capacitor.
C3 - 0.1 mf capacitor.
C4 - 1000 mmf capacitor.
D1 - Carborundum crystal and catwhisker detector assembly (see text).
J1 to J6 - Fahnestock clips.
L1 - Tuning coil (see text).
R1 - 5,000-ohm potentiometer (linear taper).
BAT - 6-volt battery
PHONES - 2000-ohm headphones.
MISC. Coil form (see text), No.24 enameled copper wire, 1/2-inch long spacers, perfboard strip (see text), push-in clips, solder lugs, knobs, hook-up wire, 1N34A diode (or equiv.), wood for base and D1 detector assembly (see text), wood screws, machine screws and nuts, alligator clips X-Y, brackets for L1 and R1.

CARBORUNDUM CRYSTAL SET

Materials other than galena, silicon or germanium have the property of detection. In the beginning of radio, carborundum (the name given to a compound of silicon carbide) was also used as a detector. The required heavy catwhisker pressure made it suitable for the early radio stations on ships, as it did not fall out of adjustment. It would also handle the strong signal energy from nearby spark transmitters without burning out. But what makes this type of crystal detector different, is the necessity for a bias battery.

A crystal detector has a high current flow with voltage applied so that it conducts in the forward direction (catwhisker to crystal), and a very low current flow in the reverse direction. The amount of current flow in the forward direction depends upon the characteristics of the crystal material and the applied forward voltage.

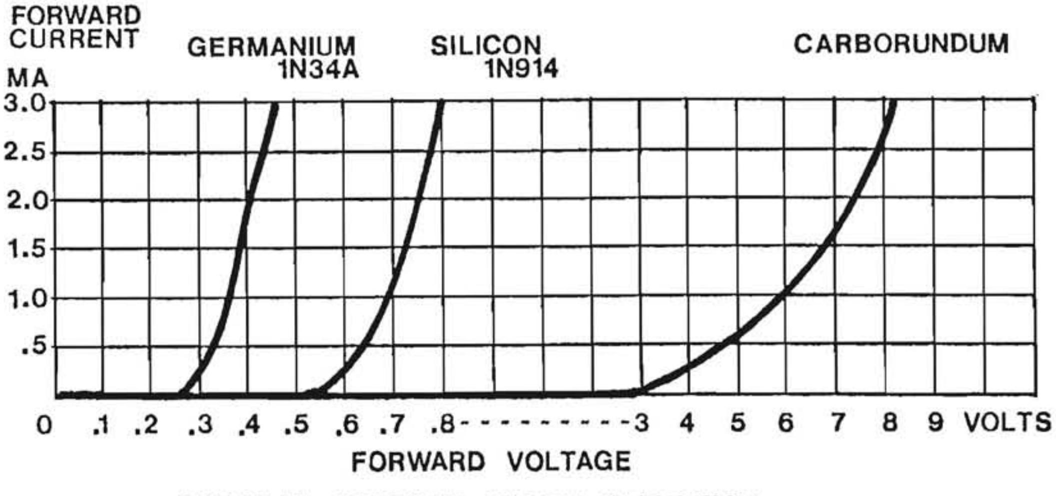

TYPICAL CRYSTAL DIODE CURRENT

The graph compares carborundum with germanium and silicon crystal detectors. Germanium minimum conduction voltage is about 0.3 V. Silicon is 0.6 V, and carborundum is 3 V. The high carborundum minimum conduction voltage is the reason a bias battery is used to move the threshold down so that weak radio signal voltages can be detected.

TUNING COIL

As shown in the illustration, the tuning coil L1 is wound on a cardboard (or plastic) mailing tube section 2-inches in diameter and 2 3/4-inches long.

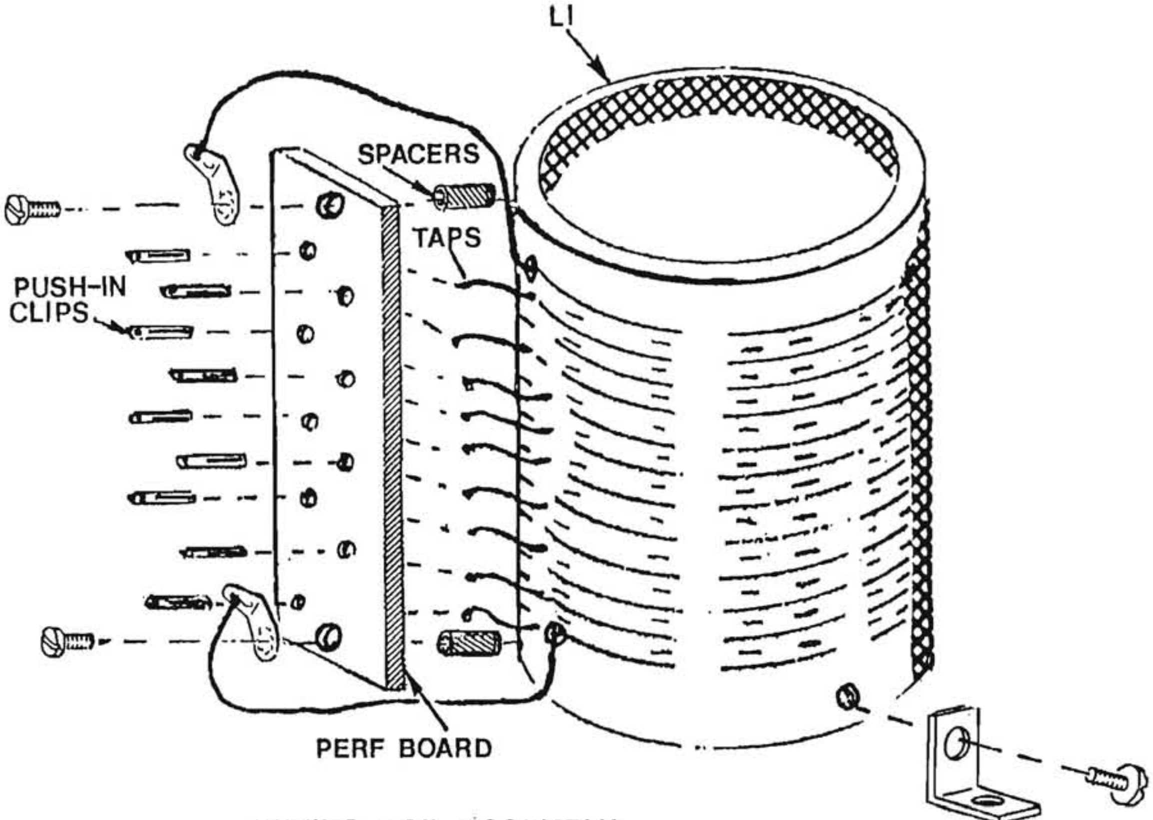

TUNING COIL ASSEMBLY

Locate and drill a hole 1/4-inch from each coil form end for machine screws to fit the two spacers. Also drill holes for two mounting brackets at one end. Then drill a small hole near each spacer hole to feed the wire on the form before you start winding.

Wind the coil with No. 24 enameled copper wire. Tap the coil every 10 turns for a total of 9 taps. An easy way to make the taps is to twist the wire together for a half inch and position the free end out. Make all the coil taps in the area between the two spacer holes. Thread the free end of the wire through the small hole at the end of the form near the spacer hole. Make sure that all of the turns are tightly wound and close together. Use a section of sandpaper to remove the enamel from the tap wire ends. Then tin the ends.

Mount 9 push-in clips in a 1/2-inch by 2 1/4-inch perf board section. Locate and drill holes on the board and install it on the coil form with two solder lugs, 1/2-inch spacers, machine screws, and nuts. Solder the coil taps to the ends of the push-in clips. Connect the coil leads to the solder lugs and mount the two 1/2-inch brackets on the coil form.

DETECTOR

The crystal detector assembly is vertical, (as shown in the illustration) instead of horizontal as in most catwhisker detectors. Begin construction by cutting the 2 1/2-inch diameter base and 2-inch by 4-inch vertical section from 1/4-inch thick wood.

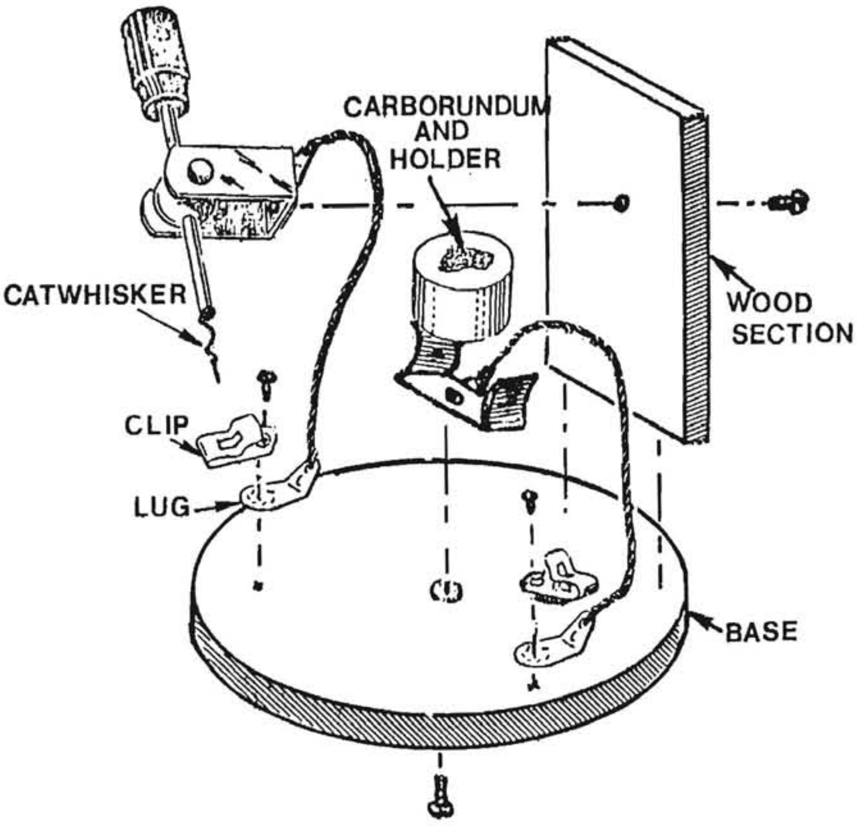

DETECTOR ASSEMBLY

Locate and install the two fahnestock clips on the base with solder lugs and small wood screws. Mount the vertical section on the edge of the base with wood screws (or cement). Locate and install the catwhisker holder with a solder lug, machine screw and nut near the top of the vertical section.

Temporarily position the crystal holder on the base so that the catwhisker will touch the center of the crystal. Locate and mount the crystal holder in this position with a solder lug and machine screw and nut. Connect the detector assembly with hookup wire as shown in the illustration. Solder a short section of coiled steel spring wire to the end of the positioning rod to serve as a catwhisker. Cut the end of the wire diagonally to give it a sharp point.

BASEBOARD

The receiver parts are installed on a 8-inch by 7-inch by 3/4-inch wood base. Wherever poossible, small wood screws are used to hold the components on the base.

Begin construction by laying out and installing the parts in the positions shown in the illustration. It will be necessary to make a metal mounting bracket to fit R1. C1A-B and C2 are mounted with machine screws in the tuning capacitor frames through countersunk holes in the base. Install solder lugs in the rear frame of both tuning capacitors. Also mount solder lugs with all of the terminals. Install knobs on the tuning capacitors and R1.

WIRING

Use hookup wire to connect the crystal set parts as shown in the schematic and illustration. Connect 5-inch leads of flexible stranded insulated wire to each alligator clip. Solder one lead to the stator lugs of C1A-B and connect the other lead to D1. Connect C3 and C4 as shown and keep their leads as short and direct as possible.

Experiment 1

Connect your antenna to J1, ground to J2, and 2000-ohm headphones to J5 and J6. Temporarily connect a germanium diode (1N34A or equiv.) in place of the carborundum crystal detector D1. The diode will make it easier to check out the receiver.

Do not attach the battery
to J3-J4 at this time, or
the diode may be damaged.

Connect **X** C1A-B clip lead to the top tap of L1, and **Y** D1 clip lead to the second tap from the top. Tune in a radio station with C2 and then adjust C1A-B for best volume. Change the position of the clip leads and retune C1A-B and C2 for best reception. Check operation of the receiver over the entire broadcast band.

Experiment 2

Tune the receiver to a radio station and then disconnect the germanium diode. Do not disturb the receiver tuning. Install a carborundum crystal into the D1 assembly. Set R1 to mid range, and connect a 6-volt battery to the receiver; negative lead to J3, positive lead to J4.

Adjust the D1 catwhisker for a sensitive spot on the carborundum crystal. Then set R1 for best volume of received signal. Retune C1A-B and C2 and change the clip leads on the L1 taps as necessary for reception of radio stations.

Note

Carborundum that will work as a detector may be hard to find. Grinding wheels are not suitable as a source, because the carborundum is usually in fine particles mixed with other material. Best bet is to contact collectors of antique radio parts. But you can use this receiver to experiment with minerals for semiconductor operation. See the Workshop Section for information on mounting mineral sections for detector experiments.

4

COMPONENT PROJECTS

BOOK TUNING CAPACITOR

During the roaring 20's,the Crosly Radio Corporation used a type of tuning capacitor that had two plates hinged together on one edge. The dielectric was a sheet of mica between the two plates. This type of tuning capacitor was called a Book Condenser.

The tuning capacitor shown in the illustration is similar to the Crosly Book Condenser. It uses aluminum foil plates cemented to fixed and movable wood sections, and a plastic sheet dielectric (cut from thin transparent plastic notebook protector material).

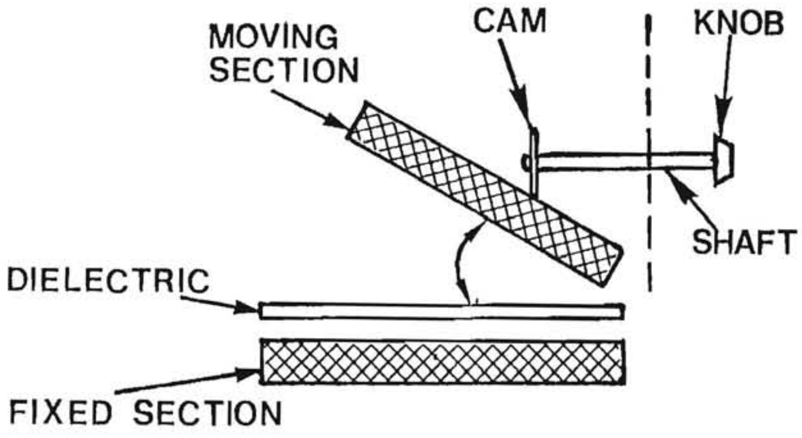

Adjusting the tuning knob causes a metal cam to rotate and control the movement of a spring loaded wood section. This varies the capacity between the two foil plates. As the plates are brought closer together, the capacity increases. As they are separated, the capacity decreases.

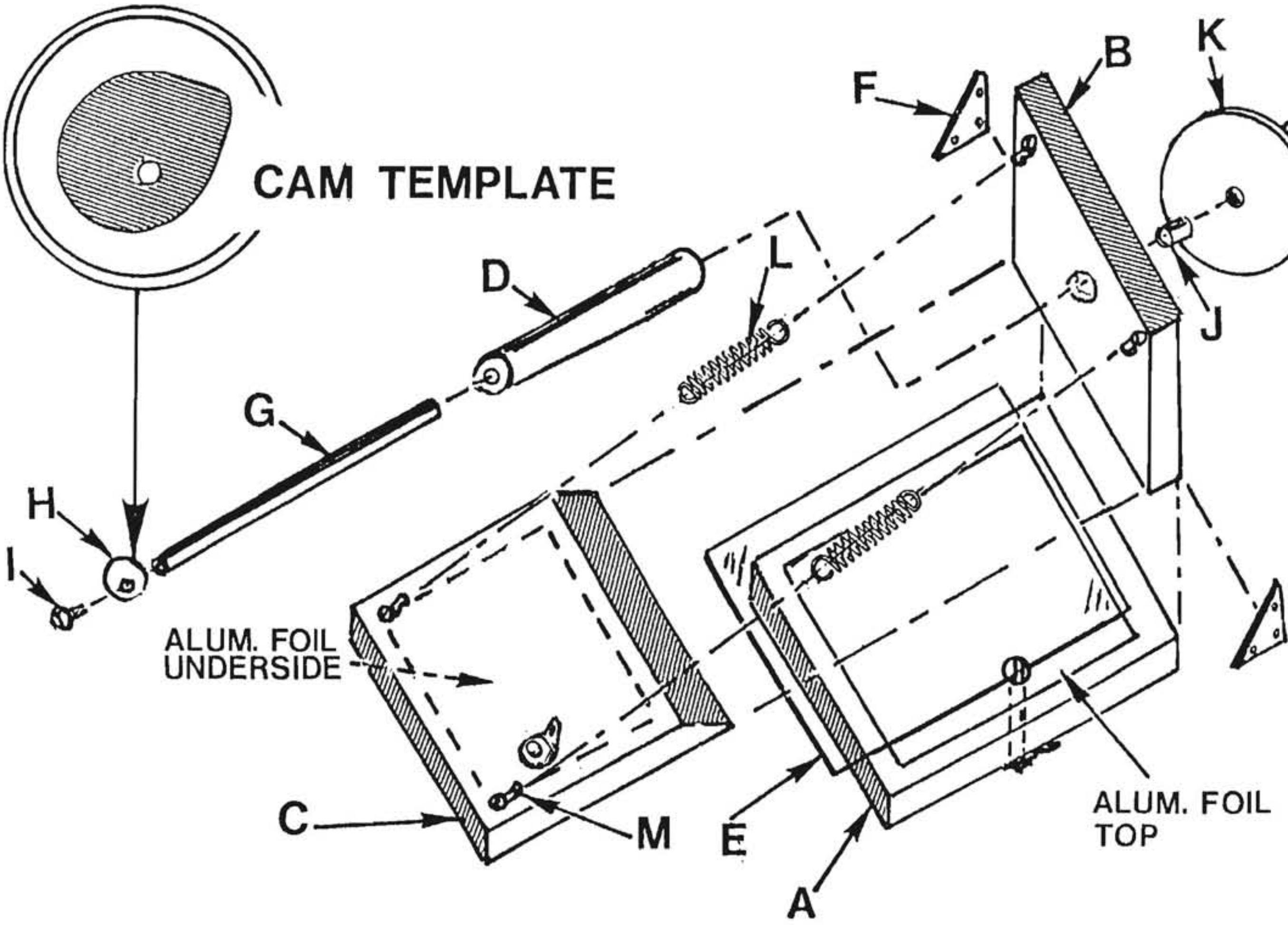

BOOK TUNING CAPACITOR

List of Materials

A - 3 x 3 1/2 X 3/8-inch wood section
B - 3 x 1 3/4 x 3/8-inch wood section
C - 3 x 3 1/8 x 3/8-inch wood section (one end beveled 45 degrees - see text).
D - 3/4 - inch diameter wood dowel, 2 1/4 inches long.
E - 3 1/2 x 3 -inch sheet plastic (See text).
F - 2 triangular 1-inch metal sections
G - 1/4-inch metal rod, 3 -inches long.
H - Metal cam (see text).
I - Machine screw (see text)
J - Bushing 3/8-inch long (to fit rod G).
K - Tuning knob (to fit rod G).
L - 2 springs 1/4-inch diameter by 1 1/4-inch long.
M - 4 solder lugs.
MISC. Small wood screws, nails, aluminum foil, machine screws, nuts, solder lugs and washers.

CONSTRUCTION

Start construction by cutting the wood sections from 3/8-thick wood to the sizes shown in the drawing. Section C has a 45 degree bevel at one end. This acts as a pivot point for the movable section. Drill and countersink a hole to fit a flathead machine screw in the places shown on sections A and C. Do not install the screws at this time.

Cement aluminum foil on the top of section A and on the bottom of section C. Then cut and remove a 1/2-inch strip of foil from the front edge of section A. Cut a 1/8-inch strip from the remaining edges. Also cut away a 1/8-inch strip from all the edges of the foil on section C.

Carefully slit the foil on top of the machine screw hole on section A and press the foil gently inside the hole. Install the flathead machine screw into the hole, making sure that it is in contact with the foil. Fasten the screw with a washer and nut. Make sure that the screw does not turn and rip the foil. Install a solder lug and nut on the free end of the screw. Repeat this operation with section C.

Cut the sheet plastic E to size and place it on top of section A. Then place section B on top of the plastic sheet and install it on the end of section A with wood screws. Install the two metal sections F with small nails.

Saw the metal rod G to size and tap one end to fit a machine screw. Cut the dowel D to size and center drill it to fit the rod G. Cut a flat on one side of the dowel.

Make a tracing of the cam H template and cement it on a section of heavy gauge sheet aluminumm. Cut the cam to size and drill a hole to fit the rod machine screw in the location shown. Install the cam with a lock washer on rod G. Insert the free end of the rod into dowel D. Place section C on top of the plastic sheet E on section A. Then temporarily position dowel D with rod G in the center of section B. Rotate the cam so that its long end touches the top of section C. Mark the placement of the dowel on section B. Drill a hole in section B to fit dowel D. Cement the dowel in place.

Locate and install solder lugs with wood screws on sections A and C as shown in the drawing. Bend over the ends of the solder lugs and install two springs L in the lugs.

Mount tuning knob K and bushing J on the free end of rod G. Adjust the knob and check that section C moves in and out with the cam rotation. changing the capacity between the foil plates.

You may have to experiment with the cam diameter for best operation of section C. File the cam edges in small increments only and recheck the operation often. Use very light springs to keep section C from pressing against the cam and moving it.

Reverse the cam on rod G to alter the direction of capacity (clockwise or counterclockwise). The capacitor can be installed vertically on a receiver project with a small metal bracket mounted on the bottom side of section A.

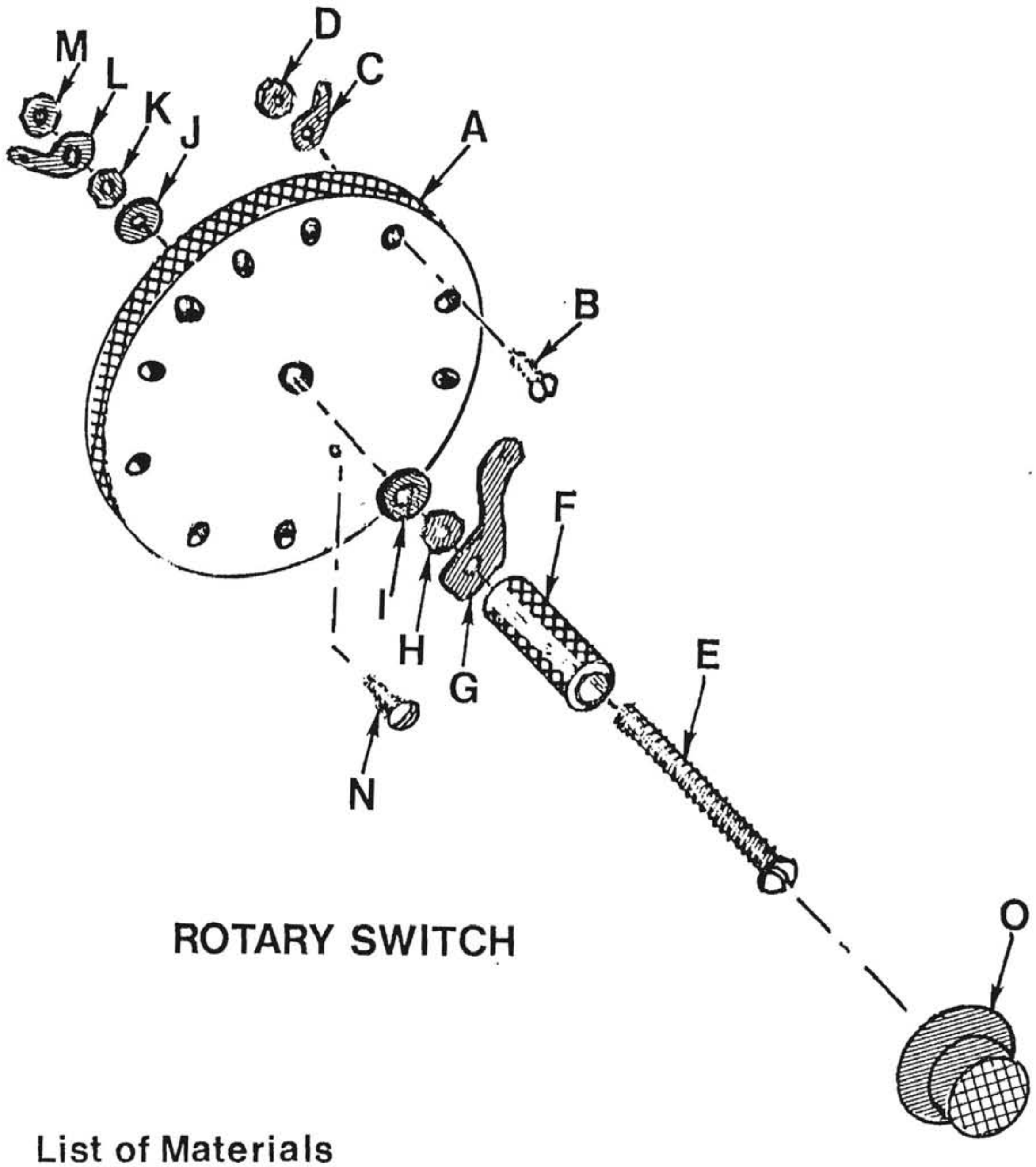

ROTARY SWITCH

List of Materials

A - 2 1/4 diameter x 3/8-inch wood section.
B - Flat head machine screws.
C,L - Solder lugs.
D - Nuts to fit B.
E - Machine screw 1 1/4-inches long.
F - 1/4 diameter x 1/2-inch long bushing (to fit E).
G - Switch arm 1 x 1/4-inch (sheet brass, see text).
H,K,M - Nuts to fit E.
I,J - Washers to fit B.
N - Small woodscrew.

ROTARY SWITCH

In the early days, radio experimenters could not buy parts that are readily available today. They had to build their radio components from wood sections and common hardware.The rotary switch ilustrated is similar to a type constructed for use with tapped coils of crystal sets.

The rotary switch shown uses flat head screws for contact points that are mounted on a wood section. The switch arm is cut from sheet brass and is installed on a machine screw.

CONSTRUCTION

Refer to the drawing of the rotary switch. Begin construction by cutting the 2 1/4-inch diameter section A from 3/8-inch thick wood. Locate, drill,and countersink the holes for the flathead machine screws B. The rotary switch shown has 10 machine screws, but you can space the hole locations for the number of switch contact points desired. Allow a space between the starting and ending contact points for the woodscrew N. This woodscrew is used as a stop for the rotary arm G. Locate and drill a center hole for the machine screw E. Install the screws B in section A with solder lugs C and nuts D. Install the woodscrew N.

Measure and cut the rotary arm G from a section of sheet brass. The arm should be long enough to reach the contacts. Make sure that it is not too wide, or it may short out to a nearby contact. Make a dent or "dimple" on the contact end, and drill a hole for screw E on the other end. Bend the rotary arm as shown in the drawing. Install screw E into bushing F, and arm G, and tighten the assembly with nut H. Place washer I on the screw and install screw E into the center hole of A. loosely fasten the screw with washer J and nut K. Then rotate the bushing F and arm G. Adjust the tightness of nut K as required to allow rotation of G with only a slight "drag". Install solder lug L on screw E with nut M. and bend the lug towards the screw end. Install knob O on bushing F, and recheck operation of the arm G. It should touch all of the contacts B and stop at screw N.

The rotary switch is used as shown in the Dual-Tuned Crystal Set Construction Project. Make sure that extra lead length is allowed for connection to the solder lug L. Bend the lead into a "pig-tail," so that it will not break as the solder lug L is rotated.

5

WORKSHOP

The construction projects in this book can be built with common hand tools found in an average home workshop.

Useful Hand Tools

Safety glasses or face shield
Steel rule
Square
Hack saw
Wood saw
Screwdrivers
Pliers
Wirecutters
Tin snips
Center punch
Files
Hammer
Soldering Iron

Wear safety glasses or face shield while working. Observe common sense safety precautions.

SOLDERING

Make sure that the part to be soldered is clean. Use only rosin core solder. Prepare the soldering iron tip as follows:

File each surface of the tip to remove oxidation. Plug in the iron and let it heat until you can melt solder on the tip (before the copper tip turns blue or bronze color). Melt enough solder until all sides of the tip are coated. Then wipe off excess solder with a damp sponge or cloth.

Make sure that the bare wire lead to be soldered is free of oxide. It is best to tin it first by melting a bit of rosin core solder on the lead end. Crimp the lead in the part hole to keep it from moving during soldering. Hold the iron tip under the part until it is hot enough to melt solder. Apply only enough rosin core solder to cover the surface of the part. Remove the iron. Do not move the part or wire while it is cooling.

PROJECT CONSTRUCTION TIPS

Tape a sheet of graph paper over a project wood section. Mark locations of the parts on the paper. Center punch mounting hole centers through the paper. Use a compass to draw circles the same size as the holes. Then drill holes through the paper.

Lettering on the projects can be done with a labelmaker that prints on strips of adhesive coated plastic. You can also use decals or dry transfer lettering sheets.

Small metal brackets can be made from heavy gauge sheet aluminum. Layout and drill the mounting holes and then cut it to size with a tin snips. Form the bracket by using a heavy mallet and a vise.

Tuning Capacitors, Knobs, and Dials

Old radios are a good source of components. The radios generally use two tuning capacitor sections ganged together. The receivers previously described will operate with one of the capacitor sections (365 mmf). You can cut out cardboard discs and cement them to radio knobs and use them as dials with the tuning capacitors. Mark the dials with dry transfer lettering, or black drawing ink.

Panels and Cabinets

The layout and construction of crystal sets is not critical. They can be fitted into many different types of wood cabinets. You can use almost any type of non-metallic box, but metal boxes are not recommended as sheet metal close to the tuning coil may lower its "Q". Black sheet plastic used for the front panel with white dry transfer lettering will give your receiver project an antique look.

MOUNTING MINERALS FOR CRYSTAL DETECTORS

Minerals (galena, etc) for experimentation as crystal detectors, will work best when mounted in a metallic base. For temporary mounting (while testing), wrap aluminum foil tightly around the sides and bottom of the mineral. Then place it into a sheet metal ring or cup that will fit into the detector assembly.

After testing, the mineral can be permanently mounted in a metal slug. Drill a shallow hole of a diameter equal to the detector assembly crystal in a block of wood. Fill the hole with melted solder. While the solder is still molten, carefully place your mineral section into the center. Hold the mineral in place with tweezers or long nose pliers until the solder cools. Then pry the metal slug out of the wood. Be sure that the top of the mineral section is free of solder.

INDEX

Advanced receiver projects,32
Amplifier connection,14,15
Antenna lead-in,10
Antenna,10

Book tuning capacitor,48
Boundary,5
Bias,5

Capacitor,8
Capacity,8
Carborundum crystal set,43
Carborundum,43
Catwhisker detector,4,6
Catwhisker,6
Component projects,48
Construction tips,55
Crystal detector assembly,45
Crystal detector,4

Detected audio signals,7
Detection,4,6
Diode electron flow,5
Dual spiderweb coils,37
Dual-tuned crystal set,32

Earphones,11
Enamel insulation,17
Electron current flow,5

Ferrite coil crystal set,23
Ferrite coil,12
Ferrite rod,23

Galena,4
Germanium,4
Ground lead,10
Ground wave,25
Ground,10

Hand tools,54
Headphones,7,11
Headset,11

Junction diode,4,5

Knobs and dials,55

Lettering,55
Loading coil,21
Loopstick,12

Modulated radio waves,6
Mounting minerals,56

N type material,4
Natural crystal,4

Operating tips,25

P type material,4
PN junction,5
Panels and cabinets,5
Parallel tuned circuit,9
Point contact diode,4,6
Primary coil,23

Reactance,8
Receiving radio waves,7
Resonant frequency,9
Rf coil,8
Rotary switch project,53

Schematic symbols,15
Secondary coil,23
Selective crystal set,37
Selectivity,9
Semiconductor diode,4
Silicon,4
Simple crystal set,12
Simple receivers,12
Sky wave,25
Slider tuned coil,17
Slider tuned set,17
Soldering,54
Space charge region,5
Spiderweb coil set,27
Station log,31

Tapped coil,32,44
Tapped spiderweb,27
Transmitted waves,6,7
Transmitting station,6
Tuned circuit,7,9
Tuning crystal set,7

Untuned detector,20

Variable capacitor,8
Variable coupling,39

Wave trap,21
Workshop,54

About the Author

Charles Green has been involved in the design and construction of electronic equipment for industry and the U.S. Government for 30 years. Much of his experience has been as a technical writer preparing operation and maintenance manuals. His continuing interest is in constructing radios and he has had over a hundred articles published in hobbyist magazines over the past two decades.

Radio Direction_Finder RDF Projects Joe WB2HOL

Sourse URL: http://home.att.net/~jleggio/projects/rdf/rdf.htm

Table of Contents

SIMPLE Time-Difference-Of-Arrival RDF....2
555 Time-Difference-Of-Arrival RDF 7
TAPE MEASURE BEAM OPTIMIZED FOR RADIO DIRECTION FINDING....10
RDF2 YAGI WITH TAPE MEASURE ELEMENTS....16
THE FOX - 40 milliwatt transmitter....21
THE FOX - 250mw transmitter with TIMER....25
THE FOX750 - 750 milliwatt transmitter....27
Style of PC board construction....29
SIMPLE ADJUSTABLE PASSIVE ATTENUATOR....32
ACTIVE ATTENUATOR....35
The HANDI-Finder®....37
The HANDI-Finder® Experimenter's Kit 39
Sense The Right Way To Go With The HANDI-Finder®....52
Frequently Asked Questions and Answers About T-hunting and RDF....56
Hardware Resources for RDF Home Builders....70
RF Sniffer with Tone Output....75
Audible S-Meter....76
The Montreal Audio FS-Meter 77

SIMPLE Time-Difference-Of-Arrival RDF

Joe Leggio WB2HOL

This T.D.O.A. RDF set is built into a die-cast Pomona box. I built it in October, 1989. As you can see from the photo, it has held up to lots of abuse as it gets knocked around in my RDF bag.

How it works

Time Difference of Arrival RDF sets work by switching your receiver between two antennas at a rapid rate. When both antennas are the same distance from the transmitter, the RF phase received by both antennas will be identical. If the two antennas are different distances from the transmitter the RF will have a different phase at each antenna. If we switch between the antennas 500 times a second, this phase change will be detected by an FM receiver as a 500 Hz tone.

By turning the antennas for a null in the tone, your two antennas will be perpendicular to the transmitter. Unfortunately, you can be facing the transmitter or facing away from the transmitter and get a null in the tone. By using a synchronous detector and processing the phase of the tone, it is possible to drive a center scale meter to indicate which direction to turn to face the transmitter.

If you have a null in the tone and turn slightly right, the meter should point to the left and direct you to turn left to face the transmitter. If instead, it points right as you turn right, you need to continue turning right to face the transmitter; you were facing away from it originally. (in other words, you need to turn 180 degrees around)

When you build this unit you must test it with *your specific receiver.* The audio circuits in a receiver sometimes invert the audio phase and if that is the case, the unit will read opposite of the true direction. If this happens to you simply turn the antenna unit around or reverse the connections to the meter.

If you look closely at the center of the antenna boom in the picture below, you will see my markings on the antenna for one of my HTs. (I used abbreviations, but it says "When using the Ten-Tec HT, this side of the boom should be on the bottom and away from the transmitter")

Circuit Description

This circuit is adapted from an idea which originally was described in QST in an article for a Double Ducky direction finder. It uses a CMOS 4066 as a synchronous detector. I used a CMOS 555 as a square wave oscillator to drive both the antenna switching diodes and the synchronous detector. I also added a stage of audio amplification to help drive a less sensitive meter. This unit was originally designed to be mounted directly on the antenna connector of my HT. This worked fine when I originally built it. Today, I have a Yeasu FT50R. It is smaller than the TDOA unit and has an SMA connector for the antenna. Somehow, I don't think it could support the weight.

Audio is input via a cable connected to the earphone jack of the HT. A small loudspeaker was mounted on the front of the T.D.O.A. RDF unit to allow me to hear the FOX. The loudspeaker is not illustrated in the schematic shown here. The T.D.O.A. RDF unit supplies voltage to the antenna unit through the center conductor of the coax. The voltage alternately causes the diodes connected to the antenna elements to conduct. As the diodes conduct, RF is able to flow from one of the antenna elements to the receiver The switching rate is set by the .01uf capacitor and the 47k resistor connected to pin 2 of the 555. If you would like a different tone, try adjusting the value of the 47k resistor.

This T.D.O.A. RDF unit uses a center scale meter to indicate which direction to turn in order to face the transmitter. I used a surplus meter movement in my TDOA RDF unit. It was not a center zero meter until I took it apart and moved the mechanical zero adjust lever to make it so. I would guess it was a 100 microamp meter movement originally. I think it was a signal strength meter from an old FM broadcast band receiver. A more sensitive meter would allow you to use less volume. I usually look for a 50 to 200 microamp zero center meter when building one of these units and can usually find a bargain at one of the ham radio flea markets I attend. I never spend more than five dollars for a surplus meter. The meter in this unit cost me about a dollar. I would guess that even a 500 microamp meter could be used. I would be less sensitive and require you to use a bit more volume.

The Printed Circuit Board

I used rub-on transfers (available from Radio Shack) to lay out the PC board in this unit. It does not have the audio amplifier stage which is present on the PC layout included on this web page. The amplifier is needed when using less sensitive meter movements.

The PC board layout is shown here looking at the bottom of the board as if you had X-ray eyes and could look through the board and see the parts on top. There is one wire jumper. It is at the lower right side of the PC board. The two transistors are 2N3904 or equivalent. The two diodes across the meter movement are 1N914 or equivalent. The three 25uf capacitors each have two different pads available for their leads. If you can obtain parts small enough to fit the set of pads spaced closer together, use them. There is nothing critical about the value of most of the parts. Anything plus or minus 20 percent of the value indicated will work. (don't add 20% to the 555 or the 4066. I doubt *that* would work!).

All of the resistors are 1/4 watt. The two diodes on the PC board are connected "back-to-back" The PC layout illustrates how the diodes should be mounted. The electrolytic caps should be rated for at least 10 volts. (assuming you are going to use a 9 volt battery). The 1.0 microhenry coil used to couple the output from the 555 to the antenna was fashioned by winding about 20 turns of number 22 AWG wire on a 3/16 inch drill bit. It was then slid off of the drill and mounted on the PC board.

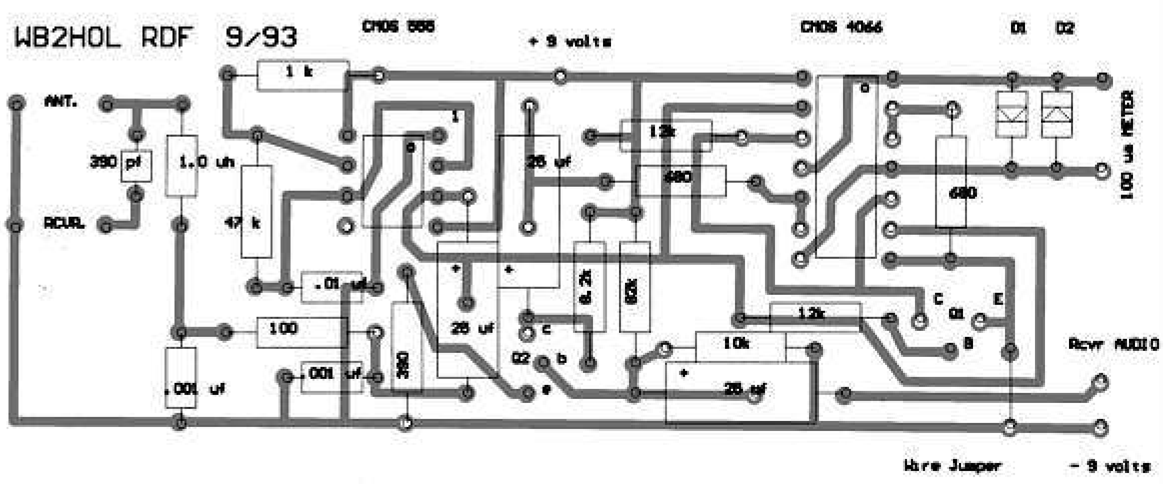

The Antenna Unit

Details of the antenna are shown below. The four antenna elements should be fashioned out of stiff wire or metal rods. One club member recycled a broken TV antenna and used 19 inch segments of its elements. The schematic shows two antennas; you might be wondering why there are four antenna elements. The extra two are connected to the coax shield at the ends of the boom. These form a counterpoise for the two active elements.

The boom of the antenna must be less than 1/2 wavelength long. The longer the boom, (up to 1/2 wavelenngth) the more RF phase difference will be detected. I found that spacing the antennas about 1/4 wavelength apart gave the best results. Further apart resulted in more modulation but at the expense of portability. It also seemed a bit less sensitive to multipath errors when I used the shorter boom. The antenna unit pictured here has a 15 inch long boom.

The antenna must be constructed as symmetrical as possible. The lengths of coax from the center BNC connector must be identical. A difference of 1/2 inch will cause an error of several degrees. If you use 1/2 wavelengths of coax you only need one set of diodes where the coax connects to the BNC connector at the center of the boom. If you find it easier to mount a flange style SO-239 connector on the boom instead of a BNC, use it instead. You will then need a cable with an PL-259 at one end and a BNC connector at the other to connect the antenna to the RDF unit.

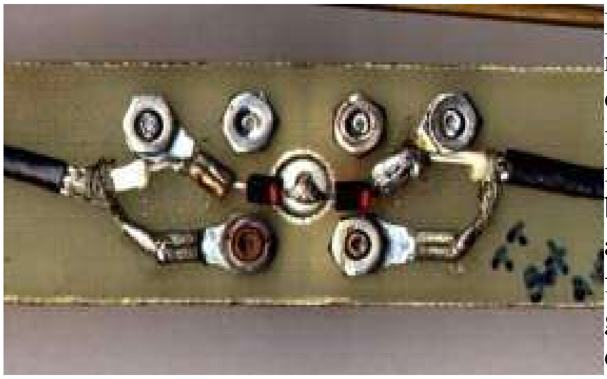

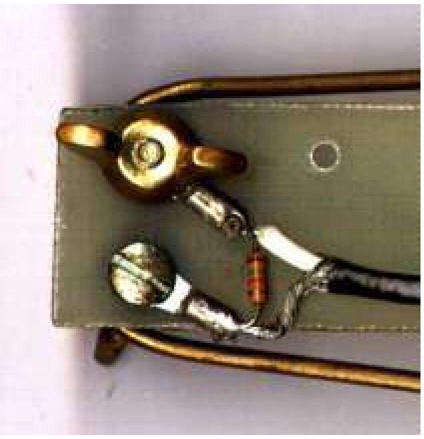

I have found that by mounting the antenna on a mast so it is up and in the clear will result in the most accurate bearings.(It is then away from reflections from your body and the ground) Whatever you do, make the unit collapsible so it can be stored when not in use.

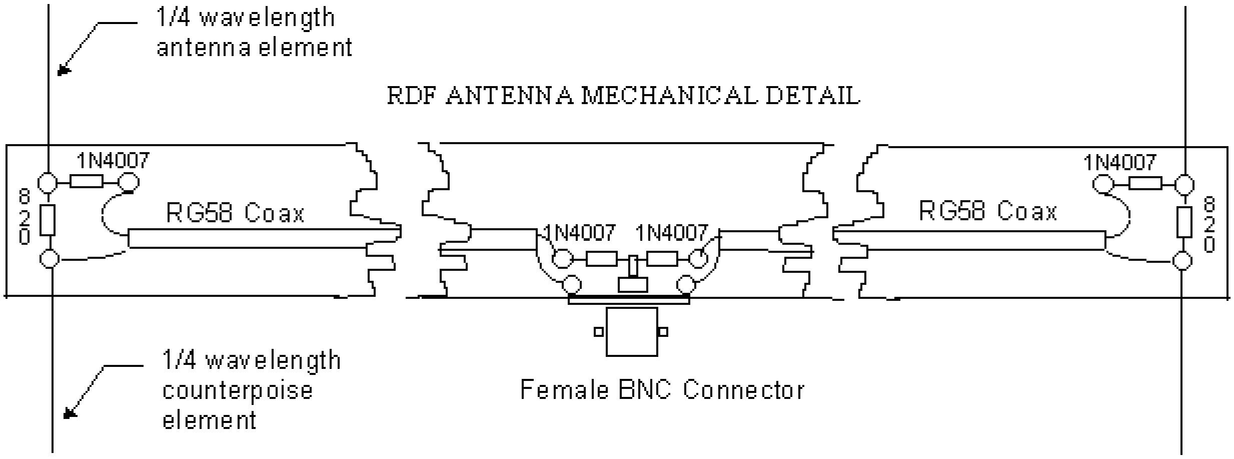
1/4 wavelength antenna element
RDF ANTENNA MECHANICAL DETAIL
1N4007
820
RG58 Coax
1N4007
1N4007
1N4007
RG58 Coax
820
1/4 wavelength counterpoise element
Female BNC Connector

It is very important that the diodes on the antenna be connected as shown on the schematic. Most diodes have a band which shows which end is the cathode. The best diodes for use in the antenna are RF PIN diodes. I used 1SS103 PIN diodes I purchased at a flea market. An equivalent diode is the MPN 3404. In a pinch, you can substitute a 1N4007 power diodes. They have a similar internal diode structure to the PIN diodes and will work almost as well. The MPN 3404 is advertised at Dan's small parts

RDF Antenna Mechanical Details

My antenna unit is designed to fold for storage. Wing nuts are used to secure the elements in position. I used brass brazing rod for the antenna elements on my latest hand-held unit. Earlier units used straightened out coat hanger. On the unit pictured, I used 1/2 wavelength lengths of coax between the center PIN diode and the antenna elements. This allowed me to use only one pair of PIN diodes in the center. I coiled the coax and secured it to the boom mid way between the center BNC connector and the ends. RG-174 coax was used on the unit illustrated here to reduce weight. The resistor in the picture is a 1.5k. The schematic below shows 820 ohms. Remember, in this circuit, few part values are critical.

The antenna boom was made from a piece of scrap glass epoxy circuit board material about 15 inches long and about 1 1/4 inch wide. I peeled the copper from the board, mounted a BNC connector at the center, and used small ring terminals secured by screws as tie points.The PIN diodes are the small rectangular black parts with the red stripe at one end. The picture shows them soldered to the back of a chassis mount BNC connector.

When you build your antenna make sure you mark one side as being the side towards the transmitter. Also mark the TOP. The first time I used this antenna, I started out walking away from the fox instead of towards it because the antenna was being held upside down. (My antenna was visually as well as physically symetrical) What amazed me was the number of other hunters who followed me thinking I knew where I was going. For the second fox hunt, I had marked the antenna. I won that event using this unit.

The Schematic

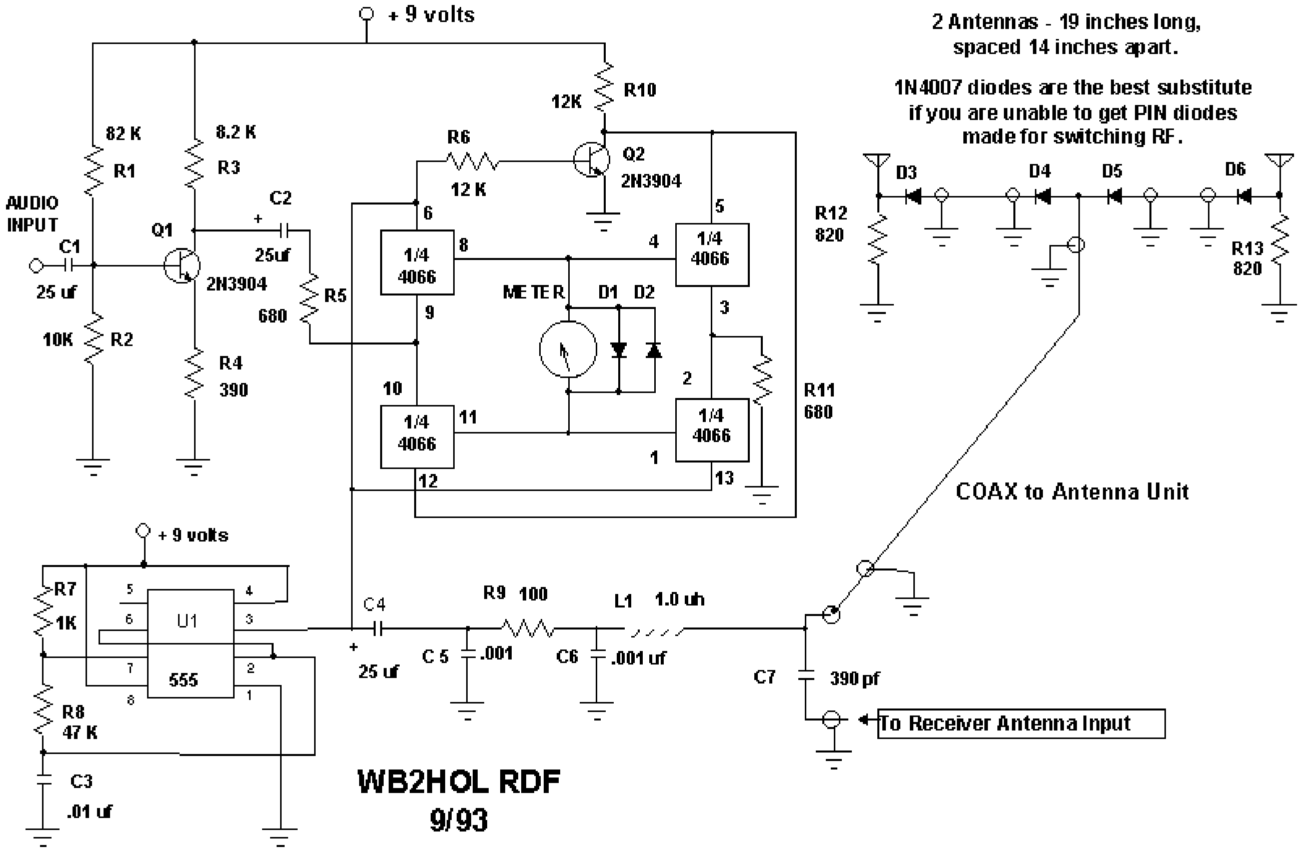
+ 9 volts
2 Antennas - 19 inches long, spaced 14 inches apart.
1N4007 diodes are the best substitute if you are unable to get PIN diodes made for switching RF.
12K R10
82 K R1
8.2 K R3
R6 12 K
Q2 2N3904
AUDIO INPUT
C1 25 uf
Q1 2N3904
C2 + 25uf
R5 680
10K R2
R4 390
1/4 4066
METER
D1 D2
R11 680
R12 820
R13 820
D3 D4 D5 D6
COAX to Antenna Unit
+ 9 volts
R7 1K
R8 47 K
C3 .01 uf
U1 555
C4 + 25 uf
R9 100
L1 1.0 uh
C 5 .001
C6 .001 uf
C7 390 pf
To Receiver Antenna Input
WB2HOL RDF
9/93

555 Time-Difference-Of-Arrival RDF

by Joe Leggio WB2HOL

This T.D.O.A. RDF set is built on a Radio-Shack experimenters board. It was featured on the cover of the May,June,July 1993 issue of the South Eastern Repeater Association, Inc Repeater Journal.

It uses a single 555 timer wired to produce a square wave output at about 500 Hz. 1N4007 power rectifier diodes were used in place of PIN diodes to switch between two dipole antennas at a 500 Hz rate. If you substitute another lower voltage power rectifier, it will work, just not as well. Of course, if you have RF PIN diodes, use them. (the 1N4007 is the only diode in the 1N400X series to contain a PIN diode structure) Power is supplied by a 9 volt battery.

I mounted the 555 oscillator circuit at the center of a 15 inch long piece of wood which acts as a boom to support two dipoles. The dimensions are not too critical. I used small brass screws as tie points at the ends of the boom and used wing-nuts to allow the dipole elements to fold against the boom when the unit was not in use.

I did not bother to add a power switch. I simply used a rubber band to secure the 9 volt battery to the boom and disconnect the battery from the circuit when I'm not using it.

How it works

Time Difference of Arrival RDF sets work by switching your receiver between two antennas at a rapid rate. When both antennas are the same distance from the transmitter, the RF phase received by both antennas will be identical. If the two antennas are different distances from the transmitter the RF will have a different phase at each antenna. If we switch between the antennas 500 times a second, this phase change will be detected by an FM receiver as a 500 Hz tone.

By turning the antennas for a null in the tone, your two antennas will be perpendicular to the transmitter. Unfortunately, you can be facing the transmitter or facing away from the transmitter and get a null in the tone. This circuit does not give you the ability to know if you are facing the transmitter or facing away from it. You must use triangulation to determine the correct direction. Although I have not tried it, you may also add a switched phasing line to your unit as shown here. This will result in a non-symmetrical receive pattern and allow an indication of which direction to travel.

The best method to tell the front from the back when using a TDOA RDF unit is by using a T.D.O.A. with a synchronous detector and processing the phase of the tone. It is then possible to drive a center scale meter to indicate which direction to turn to face the transmitter.

The Antenna Unit

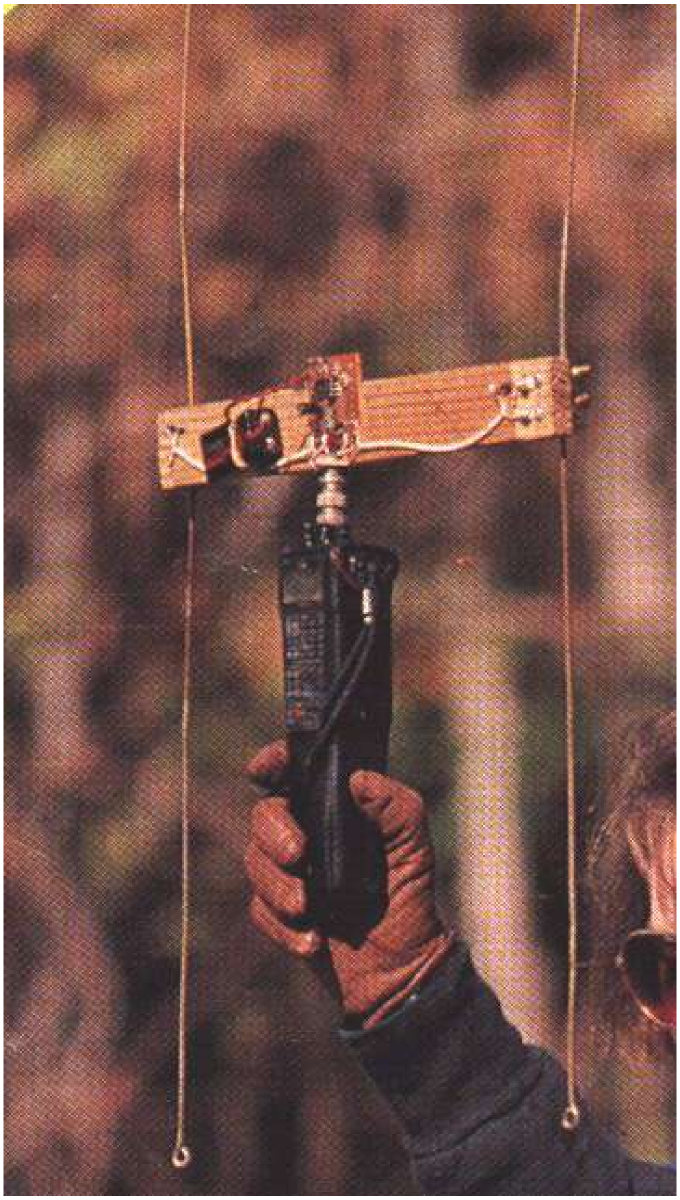

Details of the antenna are shown below. The four antenna elements should be fashioned out of stiff wire or metal rods. One club member recycled a broken TV antenna and used 19 inch segments of its elements. The schematic shows two antennas; you might be wondering why there are four antenna elements. The extra two are connected to the coax shield at the ends of the boom. These form a counterpoise for the two active elements. I found that 15 inch lengths of wire "coat-hanger" worked well as antenna elements for the unit pictured here. You can experiment to determine the optimal length and distance between the dipole elements. (until you run out of wire hangers, that is)

The boom of the antenna must be less than 1/2 wavelength long. The longer the boom, (up to 1/2 wavelength) the more RF phase difference will be detected. I found that spacing the antennas about 1/4 wavelength apart gave the best results. Further apart resulted in more modulation but at the expense of portability. It also seemed a bit less sensitive to multipath errors when I used the shorter boom. The unit pictured here has a 15 inch long boom.

The antenna must be constructed as symmetrical as possible. The lengths of coax from the center circuit board must be identical. A difference of 1/2 inch will cause an error of several degrees. Note that the layout of the antenna switching diodes on the experimenters PC board is symmetrical.

I have found that by mounting the antenna on a mast so it is up and in the clear will result in the most accurate bearings.(It is then away from reflections from your body and the ground) Whatever you do, make the unit collapsible so it can be stored when not in use. I used wing-nuts on the antenna elements to allow the unit to fold easily.

It is very important that the diodes on the antenna be connected as shown on the schematic. Most diodes have a band which shows which end is the cathode. The best diodes for use in the antenna are RF PIN diodes. I used 1SS103 PIN diodes I purchased at a flea market. An equivalent diode is the MPN-3404. In a pinch, you can substitute a 1N4007 power diodes. They have a similar internal diode structure to the PIN diodes and will work almost as well. The MPN-3404 is advertised at Dan's small parts

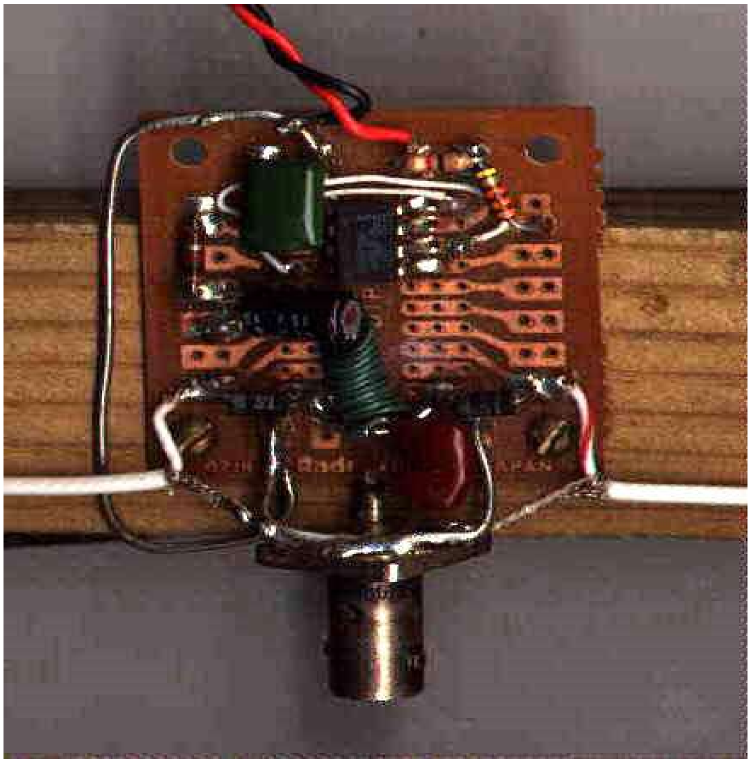

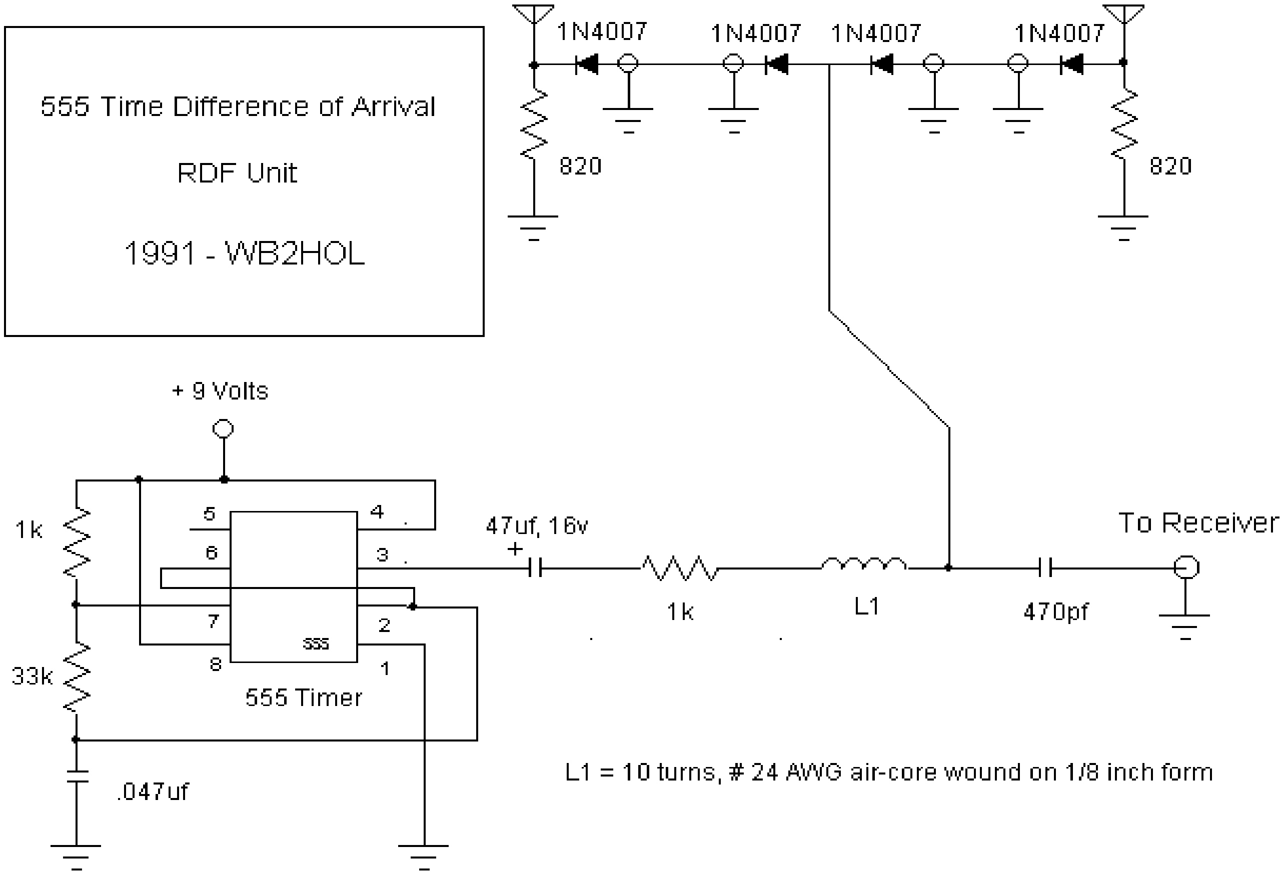

555 Time Difference of Arrival

RDF Unit

1991 - WB2HOL

1N4007
1N4007
1N4007
1N4007
820
820
+ 9 Volts
1k
33k
.047uf
5
6
7
8
4
3
2
1
555
555 Timer
47uf, 16v
+
1k
L1
470pf
To Receiver

L1 = 10 turns, # 24 AWG air-core wound on 1/8 inch form

TAPE MEASURE BEAM
OPTIMIZED FOR RADIO DIRECTION FINDING

Joe Leggio WB2HOL

Description

This antenna evolved during my search for a beam with a really great front-to-back ratio to use in hidden transmitter hunts. This design exhibits a very clean pattern and is perfect for RDF use. It trades a bit of forward gain in exchange for a very deep notch in the pattern toward the rear. (You could optimize the design for more forward gain, but at the expense of a really good notch in the pattern toward the rear.) It is a design that can be constructed using only simple hand tools (no machine shop needed) and still perform well. It has been duplicated several dozen times by other local hams and has been successfully used as a club construction project.

When I designed this antenna I had one basic idea in mind. It had to be easy to get in and out of the car when hunting for a hidden transmitter. This would be accomplished by the use of steel "tape measure" elements. These elements could fold easily when fitting the antenna into my car and yet still be self supporting. I decided to use three elements to keep the boom from getting too long.

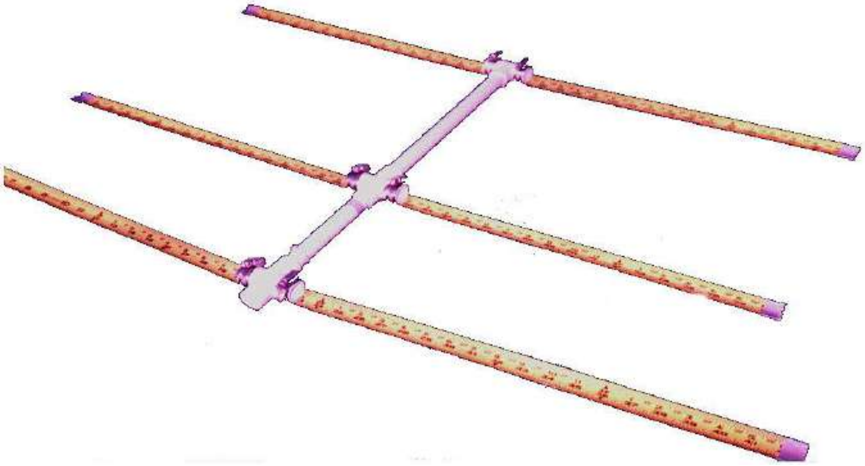

Another of my design goals was to use materials that were easy to obtain. I chose to use Schedule-40 PVC pipe and fittings available at my local hardware store for the boom and element supports. These kept the cost for the antenna very low. The element supports consist of PVC crosses and tees.

Since I had never seen any plans for an antenna using elements made from 1 inch wide steel "tape measure," I had to do the design myself. To assist in the design I used a shareware computer aided yagi design program written by Paul McMahon VK3DIP. It allowed me to optimize the antenna for the cleanest pattern combined with the best front-to-back ratio.

Performance Predicted by YAGI-CAD	
GAIN	7.3 dBd
Front-to-Back Ratio	>50 db
3 db Beamwidth	E = 67.5 degrees
3 db Beamwidth	H = 110 degrees

When I first built this beam I found it needed a matching network of some kind to have a low SWR. My first attempt was a Gamma match. This was unwieldy. The driven element could barely handle the weight and the Gamma match itself was not very flexible. The best matching network turned out to be a "hairpin match." This is simply a 5 inch length of wire that is connected across the feed points of the driven element. The antenna has some capacitive reactance without the matching network. The 5 inch length of wire has just enough inductance to cancel the capacitive reactance. This resulted in a better match than anything else I had tried.

The wire I used for the hairpin match was enamel insulated 18 gauge solid. Other hams who have duplicated this beam have used just about anything they had on hand. 14 gauge house wire works well, so does a length of 22 gauge hookup wire. It does not seem to matter if it is stranded or solid, use whatever you have available. This results in a very good match across the two meter band once you have adjusted the distance between the halves of the driven element for minimum SWR. (1 inch apart on my prototype).

I used a pair of shears to cut the tape measure elements to length. An old pair of scissors will probably do as well. No matter how you cut the elements be very careful. Those edges are very sharp and will inflict a nasty cut if you are careless. Use some sandpaper to remove the really sharp edges and burrs resulting from cutting the elements to size. I put some vinyl electrical tape on the ends of the elements to protect myself from getting cut. I encourage you to do the same. It will probably be best if you round the corners of the elements once you cut them. Wear safety glasses while cutting the elements. Those bits of tape measure can be hazardous.

The RG58 coax feedline is connected directly to the driven element. No matter what method you use to attach the feedline, make sure you scrape or sand the paint off the tape measure element where the feedline is attached. Most tape measures have a very durable paint finish designed to stand up to heavy use. You do not want the paint to insulate your feedline connection.

If you are careful, It is possible to solder the feedline to the element halves. Care must be taken since the steel tape measure does not solder easily and since the PVC supports are easily melted. You might want to tin the tape measure elements before mounting them to the PVC cross.

If you decide not to solder to the tape measure elements, there are two other methods that have been used to attach the feedline. One method employs ring terminals on the end of the feedline. The ring terminals are then secured under self tapping screws which hold the driven element halves. This method does not allow you to tune the antenna by moving the halves of the driven element. 6-32 bolts and nuts could be used if holes are drilled in the elements near the ends. If the bolt heads are placed nearest the PVC fitting, you could secure ring-terminals with nuts and lock washers. Another possibility is to simply slide the ends of the feedline under the driven element hose clamps and tighten the clamps to hold the ends of the coax. I know this is low-tech, but it works just fine.

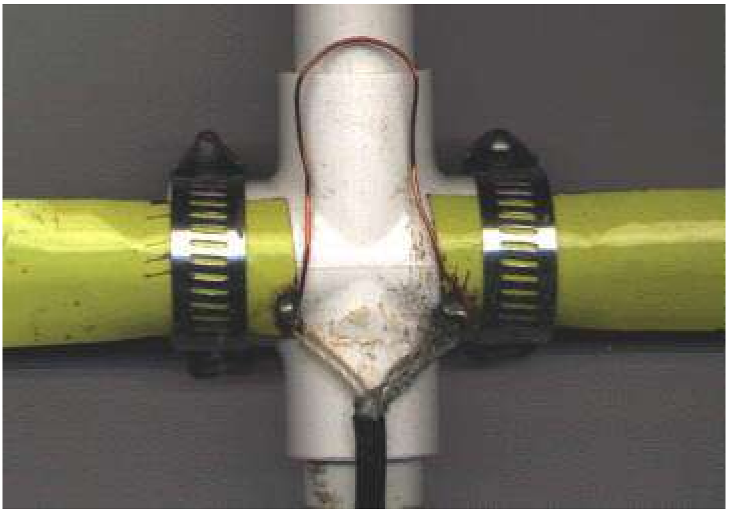

Stainless steel hose clamps are used to attach the driven element halves to the PVC cross which acts as its support. This has the added benefit of allowing you to fine tune your antenna for lowest SWR simply by loosening the hose clamps and sliding the halves of the driven element either closer or further apart. By using the dimensions specified, I found that the SWR was 1:1 at 146.565 Mhz (our Fox-Hunt frequency) when the two elements were spaced approximately 1 inch apart. Figure 1 shows the method used to attach the driven element to the PVC cross.

I used 1 1/2 inch hose clamps to attach all the elements on my prototype beam. Others who have duplicated my design have used self tapping screws to attach the elements to the PVC crosses and tees. Performance is the same using either method. The screws are much less expensive but they do not hold the elements as securely. If you do not use 1/2 inch PVC fittings but instead use 3/4 inch, make sure the hose clamps you buy are large enough to fit.

If you wish a slightly neater looking beam, use the self tapping screws. If you do not mind spending a few more dollars for the hose clamps, use them instead. If I were to build another beam I would use screws for the director and reflector, and hose clamps for the driven element. That would give me the best of both methods.

Rubber faucet washers have been used by some builders between the tape measure element and the PVC fittings on the director and reflector. These allow for the tape to fit the contour of the PVC fitting and will make the antenna look better. Now you know what to do with those washers left over from the assortment you once purchased; You know the ones I mean, the washers that do not fit the faucets you have in your house. If you are an apartment dweller, ask around, these things are stashed in almost every homeowners basement or garage.

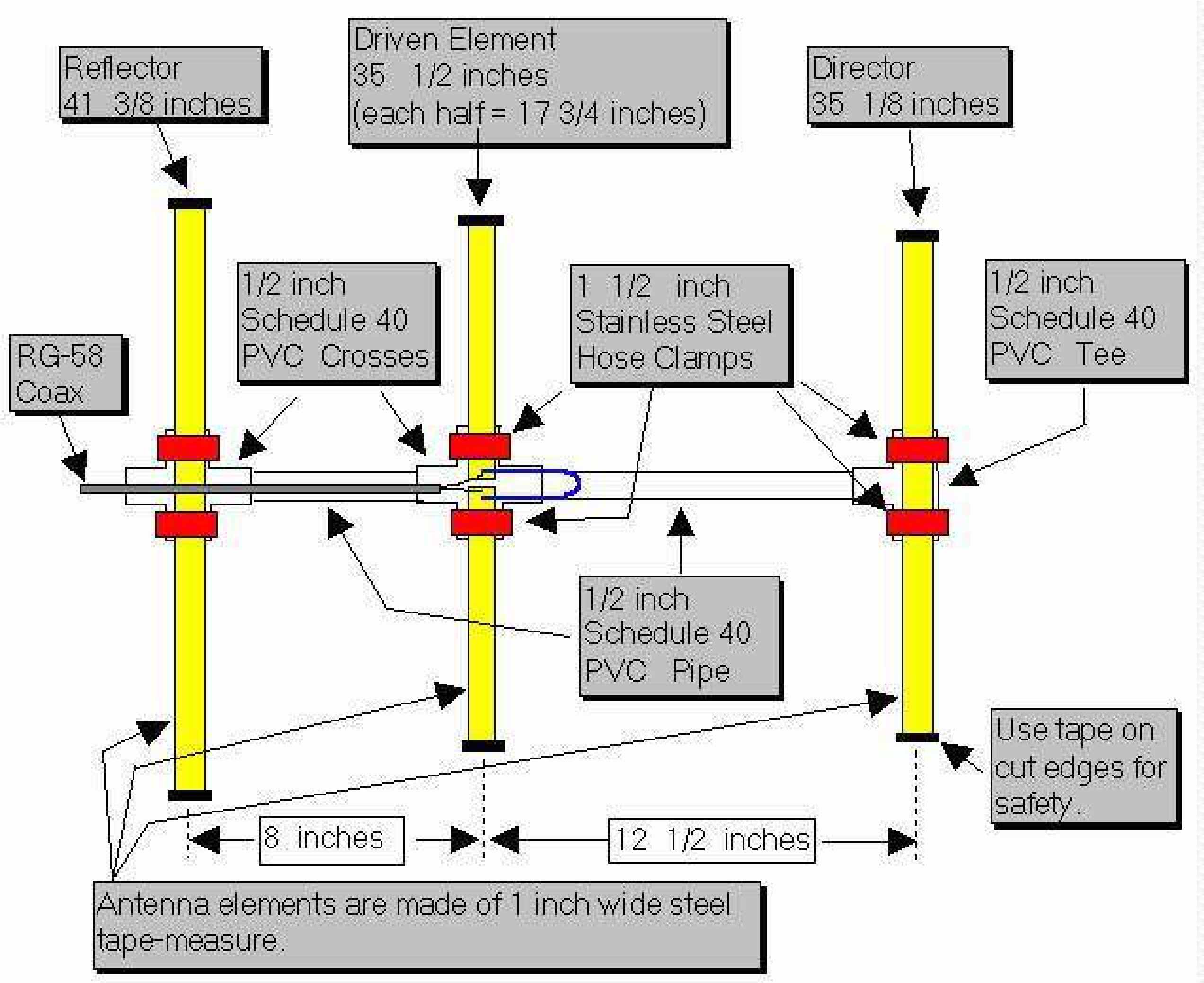
Reflector
41 3/8 inches
Driven Element
35 1/2 inches
(each half = 17 3/4 inches)
Director
35 1/8 inches
1/2 inch
Schedule 40
PVC Crosses
1 1/2 inch
Stainless Steel
Hose Clamps
1/2 inch
Schedule 40
PVC Tee
RG-58
Coax
1/2 inch
Schedule 40
PVC Pipe
Use tape on
cut edges for
safety.
8 inches
12 1/2 inches
Antenna elements are made of 1 inch wide steel
tape-measure.

Construction:

Cut a length of tape measure to 41 3/8 inches. It will be the Reflector element. Cut two lengths of tape measure to 17 3/4 inches. These will be used for the Driven element. Cut one length of tape measure to 35 1/8 inches. It will be used for the Director. Once you have cut the tape measure to length, put vinyl tape on the cut ends to protect yourself from the sharp edges. You will want to scrape or sand off the paint from one end of each of the driven element halves so you can make a good electrical connection to the feedline.

If you are planning to solder the feedline to the driven elements it is best to tin the elements first before attaching them to the PVC cross. If you don’t, the PVC will melt as you apply heat to the element. It would be a good idea to also take the time to form the wire used for the hairpin match into a “U” shape with the two legs of the “U” about 3/4 inch apart. Tin the ends of the hairpin if you plan on soldering it to the driven element. If you tin 1/4 inch of each end of the hairpin it will leave 4 1/2 inches to shape into the “U”.

You will need to cut two lengths of PVC pipe to use as the boom. One should be cut to 11 1/2 inches. It is used to form the boom between the Director and the driven element. The other piece of PVC should be cut to 7 inches. It will be used between the Reflector and the Driven element. Just about any saw will cut through the soft PVC pipe. I used a hacksaw. When we mass produced this antenna as a club project, we marked the pipe and used a portable jig saw to cut the lengths in assembly line fashion. It took longer to measure the pipe than to actually make the cuts. Since the pipe is available in ten foot lengths, you can make a few beams from a single 10 foot length. In any case, you might want to cut a few extras lengths for your friends. They will want to duplicate this once they see your completed antenna.

At this time you can pre-assemble the PVC boom, crosses and tee which will support the tape measure elements. I did not use any cement or glue when I assembled mine. The PVC pipe is secured in the fittings with a friction fit.

The hose clamps I used are stainless steel and have a worm-drive screw which is used to tighten them. They are about 1/2 inch wide and are adjustable from 11/16 inch to 1 1/2 inch diameter. Attach the tape measure elements to the PVC fittings as shown in the accompanying drawing. It is normal for the Reflector and Director elements to buckle a bit as it is tightened to the PVC Tee and Cross. You can eliminate this buckle if you use the washers and self tapping screws to attach these elements instead of the hose clamps. I do not think the beam will withstand as rough a treatment as when hose clamps are used.

How does it perform?

Once you have completed your beam you probably will be interested to see if it performs as well as the computer predicted. The SWR should be less than 2:1 across the entire two meter band. The front-to-back ratio is predicted to be very good with the antenna exhibiting a very deep notch in its pattern towards the rear. The YagiCad 4.1 program produced these antenna pattern graphs showing the pattern you should expect. If you would like to experiment a bit with this program, the yagi specification file for this tape measure beam is available for download here. Simply download the YAGI-CAD program and put the tape measure beam design file in the same directory. You will then be able to experiment with the design.

Note: under Windows95, only the first .yag file will show in the OPEN-FILE menu. You can either move all the other .yag files to a sub-directory or re-start the computer in MS-DOS mode. It works fine there. (I really do not know why this occurs but will blame Microsoft)

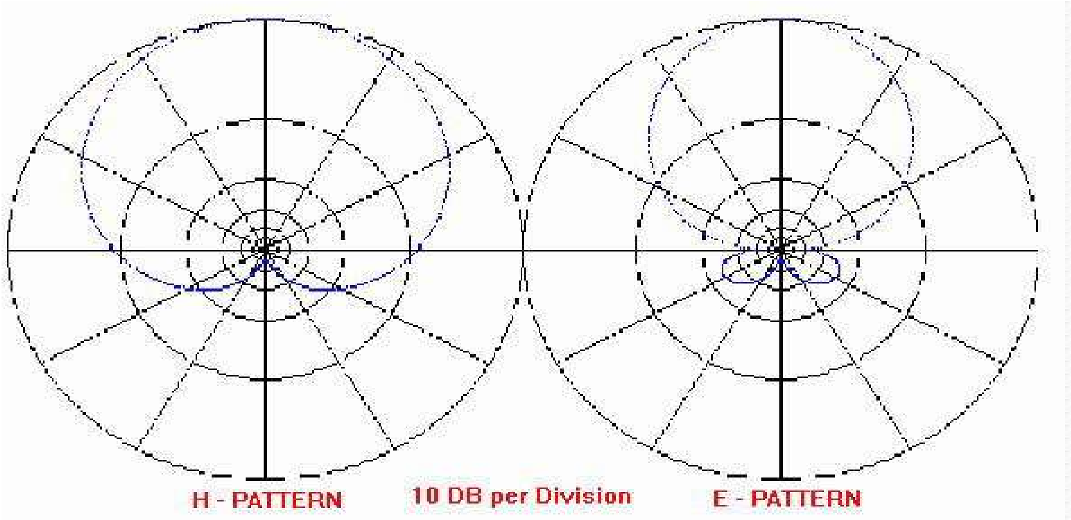

How does the tape measure beam "measure up?"

WB4SUV and WA6EZV used a storage scope connected to a copy of this antenna constructed by KC8FQY and provided the following picture of the actual antenna pattern. I am very happy to see that the computer prediction of a clean pattern with a really great front-to-back ratio was accurate. What do you think?

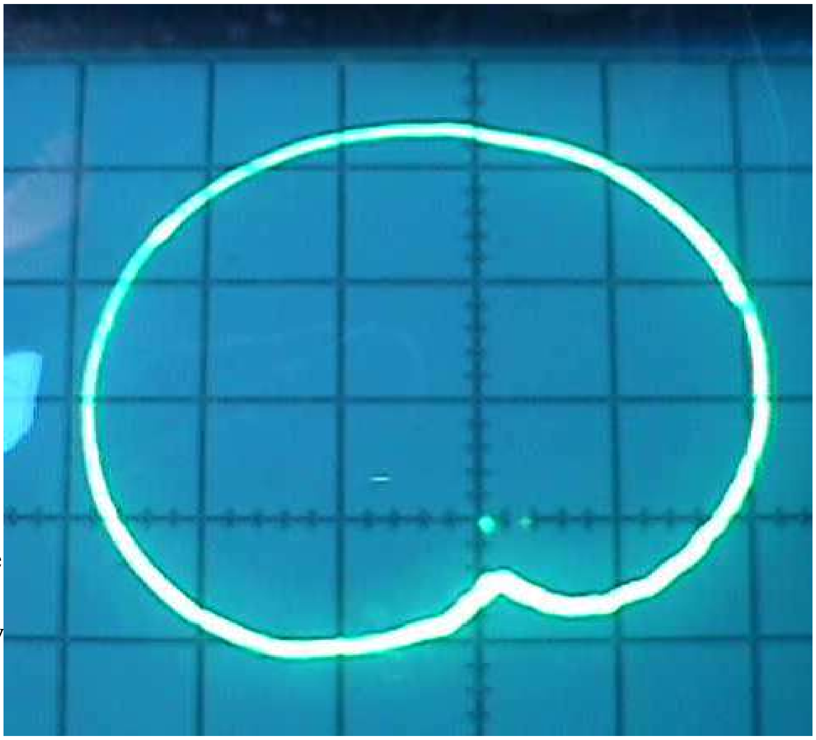

Summary

This beam has been used on Fox-Hunts, on mountain tops, at local public service events, outdoors, indoors in attics, just about everywhere. The SWR is typically very close to 1:1 once adjusted. Front to back performance is exactly as predicted. The null in the rear of the pattern is perfect for transmitter hunts. When tested using a sensitive field strength meter and a low powered fox transmitter, full scale readings were seen from a distance of ten feet. With the same field strength meter I was able to point the antenna away from the transmitter and move the reflector element to within a few inches of the transmitter antenna and still not see a reading. I don’t have the facilities to verify a 50 db notch as predicted by the Yagi-Cad software but It sure seems close. The flexible elements have taken a lot of abuse. My antenna has seen a lot of use and has held up quite well. Best of all, when on a fox-hunt, this beam is a breeze to get in or out of the car.

RDF2 YAGI WITH TAPE MEASURE ELEMENTS

Joe Leggio WB2HOL

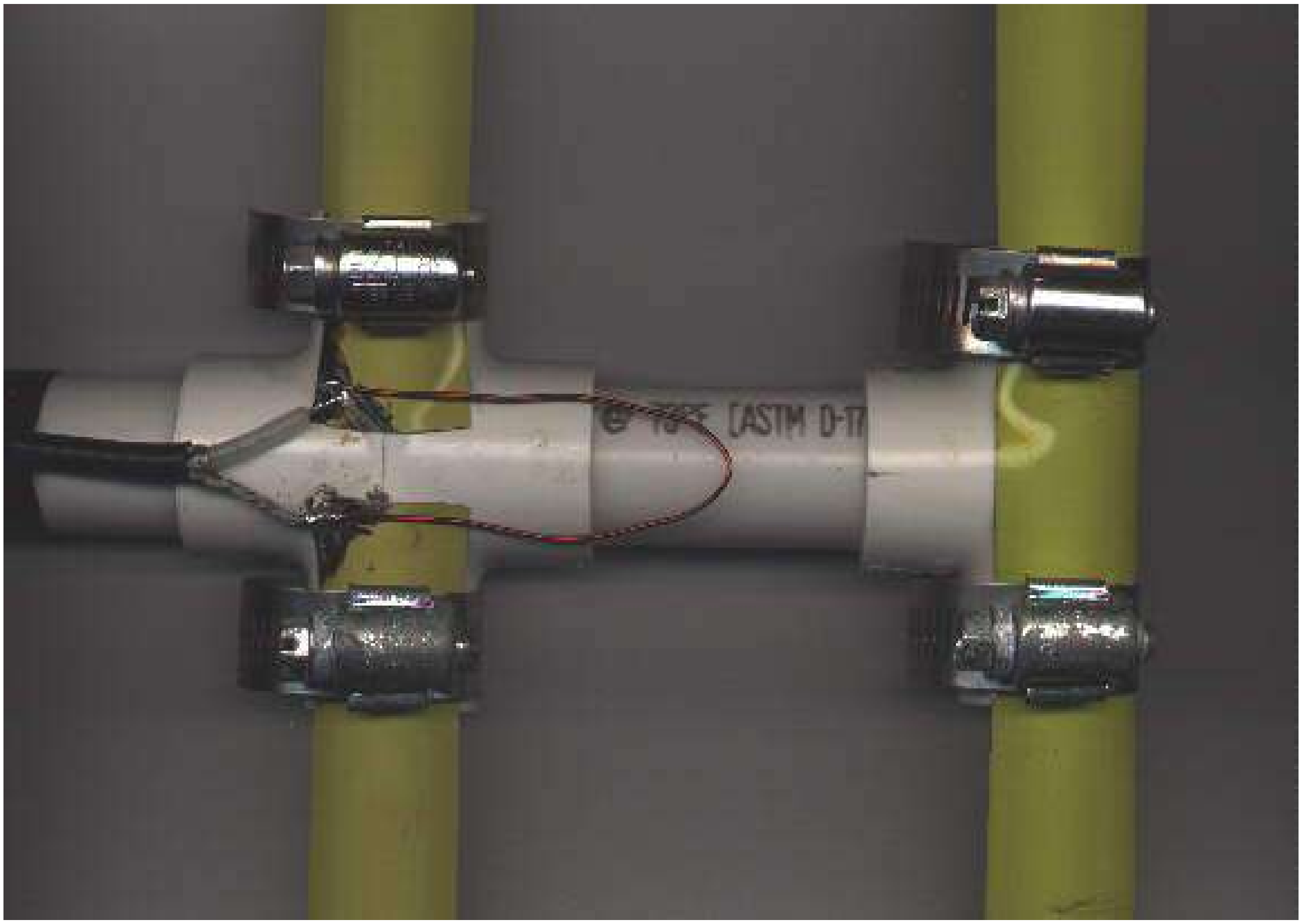

Description

This antenna is an adaptation of the RDF2 two element, close spaced Yagi antenna described in the May, 1996 issue of 73 Amateur Radio Today Magazine.

The original design, developed by N6WZI and featured in the Homing In column of 73, used 1/4 inch solid aluminum rods for antenna elements. If you don't subscribe to 73, find someone who does and ask to see the original article.

My version of the RDF2 antenna uses 1 inch wide steel "tape measure" elements. This is one antenna you can take into the woods without having to worry about Yagi Eating Trees. The flexible elements also make it very easy to get this antenna in and out of the car.

You do not need a machine shop to build this antenna. You only need common hand tools. It is much easier to construct than the original design and can be assembled in about 30 minutes once you gather the needed parts.

Performance Predicted by ELNEC	
GAIN	4.7 dBi
Front-to-Back Ratio	>8 db
3 db Beamwidth	E = 36 degrees
3 db Beamwidth	H = 70 degrees

N6WZI's original design used a "hairpin" matching network made from a length of TV antenna twinlead. I made a similar "hairpin match" but instead used a 5 inch length of wire connected across the feed points of the driven element. The antenna has some capacitive reactance without the matching network. The 5 inch length of wire has just enough inductance to cancel the capacitive reactance.

The wire I used for the hairpin match was enamel insulated 18 gauge solid. This results in a very good match across the two meter band once you have adjusted the distance between the halves of the driven element for minimum SWR at the desired frequency. (3/16 inch apart on my prototype). Experience has shown that the wire guage used for the hairpin match is not very critical. Use what you have on hand. 14 guage house wire works well, probably any insulated wire which will hold its shape will work just fine.

I used a pair of shears to cut the tape measure elements to length. An old pair of scissors will probably do as well. No matter how you cut the elements be very careful. Those edges are very sharp and will inflict a nasty cut if you are careless. Use some sandpaper to remove the really sharp edges and burrs resulting from cutting the elements to size. I put some vinyl electrical tape on the ends of the elements to protect myself from getting cut. I encourage you to do the same. It will probably be best if you round the corners of the elements once you cut them. Wear safety glasses while cutting the elements. Those bits of tape measure can be hazardous.

Replacement tape measure blades are available for reasonable prices at larger hardware stores. I found a 30 foot replacement blade at Sears for less than five dollars. If you use a replacement blade, be very careful when removing it from the package. The blade I purchased had a hard plastic cover which, when twisted, allowed access to the end of the replacement tape measure element. When you twist the cover, make sure you keep your fingers away from the "razor sharp" end of the blade. (I probably could have assembled this antenna in even less time if I didn't have to go upstairs to the medicine chest to get a band-aid. This is experience speaking. The spring steel popped out, inflicting a nasty cut on my thumb. I'll probably heal in a week or so...)

The RG58 coax feedline is connected directly to the driven element. No matter what method you use to attach the feedline, make sure you scrape or sand the paint off the tape measure element where the feedline is attached. Most tape measures have a very durable paint finish designed to stand up to heavy use. You do not want the paint to insulate your feedline connection.

If you are careful, It is possible to solder the feedline to the element halves. Care must be taken since the steel tape measure does not solder easily and since the PVC supports are easily melted. You might want to tin the tape measure elements before mounting them to the PVC cross. I used a bit of rosin flux to help with tinning the element connection points.

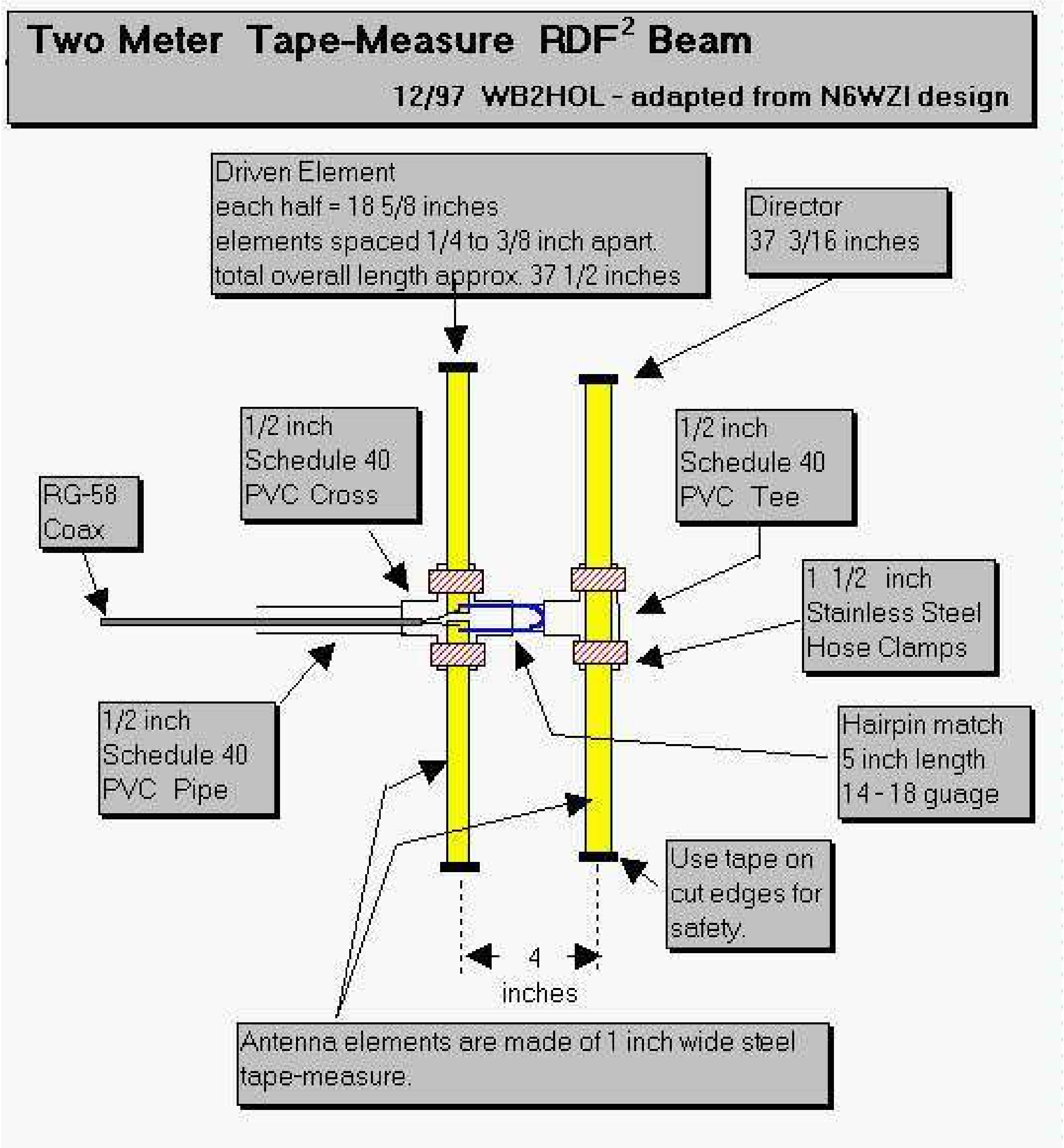
Two Meter Tape-Measure RDF² Beam
12/97 WB2HOL - adapted from N6WZI design
Driven Element
each half = 18 5/8 inches
elements spaced 1/4 to 3/8 inch apart.
total overall length approx. 37 1/2 inches
Director
37 3/16 inches
1/2 inch
Schedule 40
PVC Cross
1/2 inch
Schedule 40
PVC Tee
RG-58
Coax
1 1/2 inch
Stainless Steel
Hose Clamps
Hairpin match
5 inch length
14 - 18 guage
1/2 inch
Schedule 40
PVC Pipe
Use tape on
cut edges for
safety.
4
inches
Antenna elements are made of 1 inch wide steel tape-measure.

Stainless steel hose clamps are used to attach the driven element halves to the PVC cross which acts as its support. This has the added benefit of allowing you to fine tune your antenna for lowest SWR simply by loosening the hose clamps and sliding the halves of the driven element either closer or further apart. By using the dimensions specified, I found that the SWR was 1:1 at 146.565 Mhz (our Fox-Hunt frequency) when the two elements were spaced approximately 3/16 inch apart. Figure 1 shows the method used to attach the driven element to the PVC cross.

I used 1 1/2 inch hose clamps to attach all the elements on my prototype beam. if you do not use 1/2 inch PVC fittings but instead use 3/4 inch, make sure the hose clamps you buy are large enough to fit. Others who have duplicated my design have used self tapping screws to attach the elements to the PVC crosses and tees. Performance is the same using either method. The screws are much less expensive but they do not hold the elements as securely. In addition, the screws do not allow you to tune the antenna by sliding the halves of the driven element closer or further apart.

If you wish a slightly neater looking beam, use the self tapping screws. If you do not mind spending a few more dollars for the hose clamps, use them instead. If I were to build another beam I would use screws for the director, and hose clamps for the driven element. That would give me the best of both methods.

Construction:

Cut two lengths of tape measure to 18 5/8 inches. These will be used for the Driven element. Cut one length of tape measure to 37 3/16 inches. It will be used for the Director. Once you have cut the tape measure to length, put vinyl tape on the cut ends to protect yourself from the sharp edges. You will want to scrape or sand off the paint from one end of each of the driven element halves so you can make a good electrical connection to the feedline.

It is best to tin the elements first before attaching them to the PVC cross. If you don’t, the PVC will melt as you apply heat to the element. It would be a good idea to also take the time to form the wire used for the hairpin match into a “U” shape with the two legs of the “U” about 3/4 inch apart. Tin the ends of the hairpin to make it easier when soldering it to the driven element halves. If you tin 1/4 inch of each end of the hairpin it will leave 4 1/2 inches to shape into the “U”.

You will need to cut two lengths of PVC pipe. One should be cut to 3 inches. It is used to form the boom between the Director and the driven element. The other piece of PVC should be cut to 6 1/2 inches. It will be used as a handle so you can hold the antenna from the rear. Just about any saw will cut through the soft PVC pipe. I used a hacksaw. Since the pipe is available in ten foot lengths, you can make a few beams from a single 10 foot length. In any case, you might want to cut a few extras lengths for your friends. They will want to duplicate this once they see your completed antenna.

At this time you can pre-assemble the PVC boom, cross and tee which will support the tape measure elements. I did not use any cement or glue when I assembled mine. The PVC pipe is secured in the fittings with a friction fit. When assembled, the cross and tee will be 4 inches from center to center.

The hose clamps I used are stainless steel and have a worm-drive screw which is used to tighten them. They are about 1/2 inch wide and are adjustable from 11/16 inch to 1 1/2 inch diameter. Attach the tape measure elements to the PVC fittings as shown in the accompanying drawing. It is normal for the Director element to buckle a bit as it is tightened to the PVC Tee. You can eliminate this buckle if you use washers and self tapping screws to attach this element instead of the hose clamps but I do not think the beam will withstand as rough a treatment as when hose clamps are used.

How does it perform?

Once you have completed your beam you probably will be interested to see if it performs as well as the computer predicted. My prototype antenna does seem to have a pretty good front-to-back ratio. It is real easy to get a peak on the "S" meter. It does not hear cross polarized signals very well. This is an advantage as most reflections will not have the same polarization as the original transmitter. Gain seems to be between 3 to 5 dB less than the three element RDF beam I designed. This is probably an advantage when close to the transmitter. You will need a bit less attenuation. The short boom and light weight make this antenna attractive if you have to use it for any length of time. It is light enough for youngsters to carry without tiring. As long as you put vinyl tape on the ends of the elements, this antenna is pretty safe to carry. Descriptions of the N6WZI RDF2 Yagi performance when used on a foxhunt may be found at Jim Elmore's site.

Drawbacks

The curved, airfoil shaped, flexible elements will "flutter" and bend when in high wind situations. (Mobile flutter while stationary?) If you need additional support, you can add short lengths of PVC pipe to the TEE and CROSS and then tape the elements to the PVC pipe. I figure 6 to 12 inches of PVC would add lots of strength to the elements. The short boom does result in less gain than a more conventional Yagi. You may need more gain when starting on a foxhunt and still far from the fox transmitter.

THE FOX - 40 milliwatt transmitter

Joe Leggio WB2HOL

This two transistor 144 MHz transmitter uses a 48 MHz scanner crystal as an oscillator / tripler. A second transistor boosts the power output to 40 milliwatts. A 7-element filter is used to minimize output on the third harmonic. (I did not want it to be easy to RDF this on the 450 MHz third harmonic)

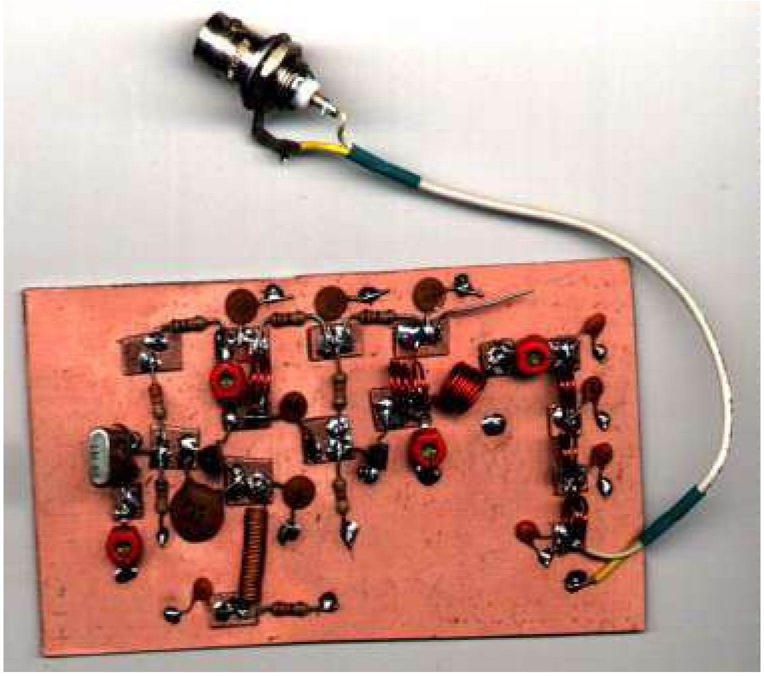

The scanner crystal I used was actually for a receive frequency 10.7 MHz higher than the two meter transmitter output. It was marked 156.21 MHz. (When you subtract the 10.7 MHz IF offset, it resulted in a transmit frequency of 145.51. Since this circuit is not the exact same circuit as the scanner, you may find that the crystal oscillates somewhere close to the marked frequency. The crystal marked as 156.21 MHz I used in this transmitter actually resulted in a signal on 145.53 MHz.)

I did get the salesperson at the local Radio Shack store a bit confused when I asked to look at his box of scanner crystals. The first thing he asked was which service I was trying to listen to... Police, Fire, Ambulance, etc.... When I said that I was looking for something for a homemade project and did not have a specific frequency in mind he understood. I did not try to tell him the crystal was to be used in a transmitter. He probably would have told me that I can't use a "receiver" crystal for a "transmitter."

As I looked at the available crystal selections, I simply took my calculator and subtracted 10.7 MHz from the marked receive frequency. I was lucky; several crystals ended up in places on the two meter band where a fox could transmit without interfering with other users. If you are not lucky, or live in an area where you must put your fox on a specific frequency, you can order a crystal through Radio Shack. Since there are still quite a few crystal controlled scanners out there, the odds are that someone in your town will have something you can use. Be aware, although most scanners use a 10.7 MHz IF, there are some with a 10.8 MHz IF. In that case, you may end up 100 kHz from where you expected.

You will need to wind several coils to build this transmitter. They are all self-supporting air-core coils made from enamel covered wire. If you follow the instructions illustrated here, you should not have any difficulty. The top of my workbench is littered with coils I have wound and tried in various circuits. I specifically avoided tapped coils or multi-winding coils to keep construction as easy as possible.

I use a simple diode detector across a 50-ohm dummy load to measure the power output while tuning up this transmitter. It is reasonably stable. It does not have a tendency to break into spurious oscillations. I did try about 5 or 6 different prototype circuits as I designed this transmitter; many were not stable. Their RF output did not smoothly change as the trimmer capacitors were adjusted but instead jumped as spurious oscillations occurred. I used an oscilloscope to monitor the output waveform and confirm the lack of spurious oscillations. This is not really necessary as you can usually hear the instability if you monitor the output signal with your 2-meter rig.

Style of PC board construction

I built all of the transmitters using copper clad PC board as a ground plane and 1/4 inch squares of PC board as tie points. The tie points were cemented to the ground plane using "crazy-glue." This technique made constructing the transmitter go very fast. It only took a drop of glue to secure the tie points. The glue set up in seconds. It made it easy to experiment with different circuits. In a pinch, a tie point could be removed or relocated by prying upward on its side with a screwdriver. I built three transmitters using this style of construction in the past month. All have worked exactly as expected. I used a pair of shears to cut a number of 1/4 inch squares of PC board material. I then used a wide file to remove any burrs created by the shear. This allowed the square of PC board material to lay flat against the PC ground plane. (Crazy-glue is not designed to fill voids. It works best when the surfaces to be glued fit together well) Be careful as you work with "crazy-glue." It is very effective at gluing fingers together.

I tried to use parts that could be obtained with little difficulty. You will need to wind several coils. I also kept these as simple as possible. The wire I used was recycled from old transformers.

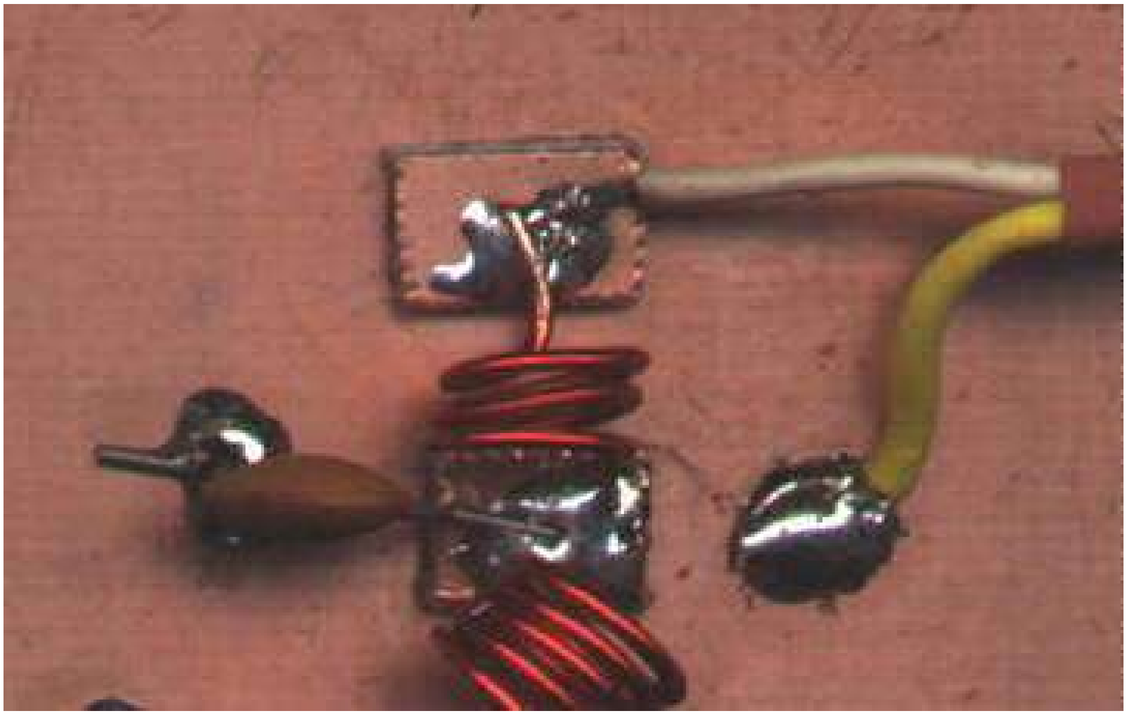

Tie-points are made from 1/4 inch squares of PC board materialglued with "crazy-glue" to a solid ground plane of PC board.

Winding the Coils

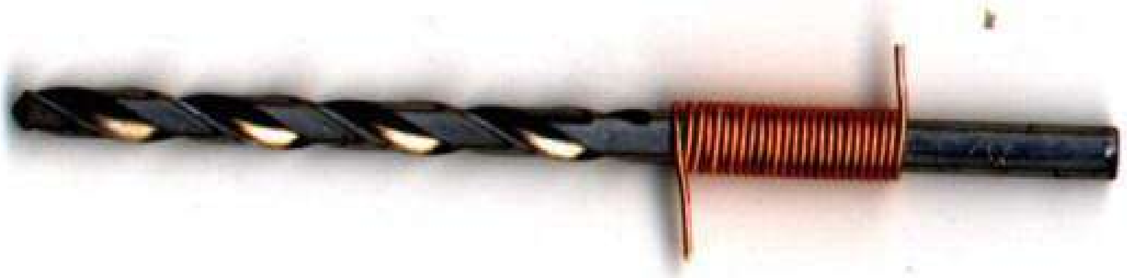

Enamel wire is close-wound on a drill bit to form the coils for this transmitter. It is easier to cut the leads to the coil and strip the enamel from the ends prior to removing the coil from the drillbit used as the form.

All the coils are wound in a single layer using a drill bit as a winding form. To help you visualize how to form the coils, I have included an illustration showing how the coils are wound..

All of the coils are "close wound" with "air cores." I used a drill bit as a form and then slipped the coil off of the bit after I had cut the free ends to about 1/4 inch. I used a hobby knife to scrape the enamel paint from the ends of the wire before I removed the coil from the drill bit used as a form. This made the coils easy to solder. With any luck, you too will be able to wind your own coils for this project.

Two different gauges of enamel covered wire are used for the coils in this project. If you have wire reclaimed from an old transformer, or from an old television deflection yoke, you can determine its gauge by winding a single layer on a drill bit. You will be able to fit about 46 turns per inch if the wire is 24 gauge. You will be able to fit about 37 turns per inch if the wire is 22 gauge. If you can't find an old transformer with the correct gauge wire to recycle, the best places I have found to buy enamel covered wire are shops which specialize in rewinding electric motors. In a pinch, if you live in the USA, Radio Shack sells "magnet wire" in both 24 and 22 gauge. You will only need a few feet of each gauge wire for this project.

I used the same transistor for both the oscillator and the amplifier stage. It is important that the transistors you use in your transmitter have sufficient gain at 2 meters. The 2N5770 is one of many I could have used. Please note: a 2N3904 would probably work fine as the oscillator (at 48 MHz), it would not work well as the amplifier as it does not work well at 144 MHz.

The tune-up of this transmitter is easy. Simply adjust each of the trimmer capacitors for maximum RF output to the antenna. (The trimmer in series with the crystal is the only exception, it is used to trim the transmitter frequency, use a frequency counter to adjust it)

I packaged one version of this transmitter in a very well shielded cast aluminum box and powered it with a 9 volt battery. The output power was reduced at the lower voltage to about 25 milliwatts. Even at that power level, it can be heard for miles. As an experiment, I tried voltages as low as 5 volts. The power output was down to below 10 milliwatts but the transmitter stayed on the air! This design works well even as the battery gets weak.

Schematic

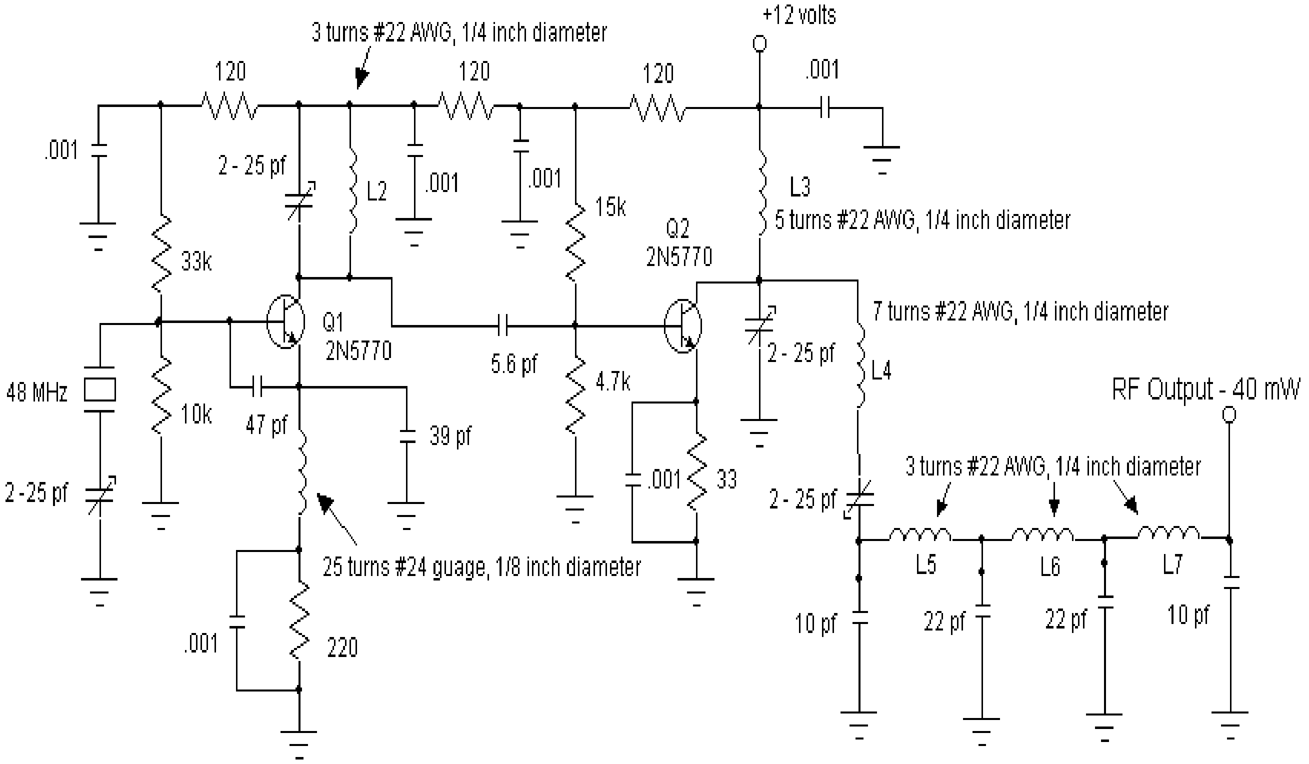

+12 volts
3 turns #22 AWG, 1/4 inch diameter
120
120
120
.001
.001
2 - 25 pf
L2
.001
.001
15k
L3
5 turns #22 AWG, 1/4 inch diameter
Q2
2N5770
33k
Q1
2N5770
5.6 pf
7 turns #22 AWG, 1/4 inch diameter
2 - 25 pf
L4
48 MHz
10k
47 pf
39 pf
4.7k
RF Output - 40 mW
.001
33
2 -25 pf
2 - 25 pf
3 turns #22 AWG, 1/4 inch diameter
25 turns #24 guage, 1/8 inch diameter
L5
L6
L7
.001
220
10 pf
22 pf
22 pf
10 pf

2 Meter - 40 mW FOX TRANSMITTER - 4/98 WB2HOL

THE FOX - 250mw transmitter with TIMER

Joe Leggio WB2HOL

This FOX transmitter is equipped with a timer. It can be set to start transmitting up to 24 hours later.

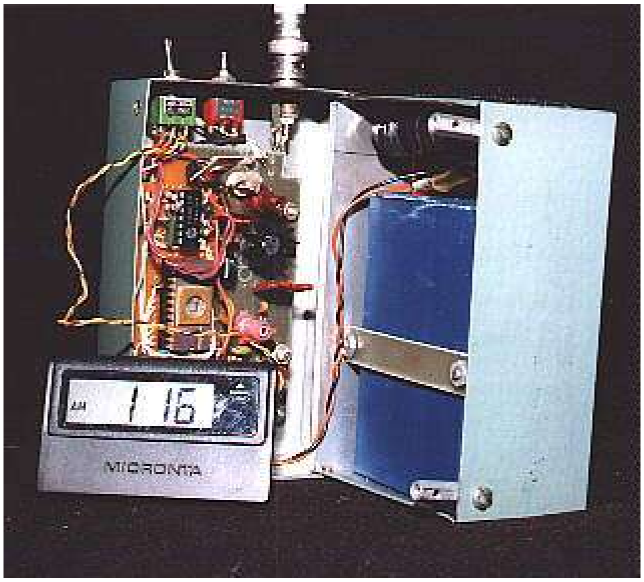

The timer consists of a Radio Shack travel alarm. I disconnected the pizeo-ceramic beeper and used the signal to set a flip-flop built from a CMOS 4011. The output of the flip flop is connected to a switching transistor which supplies voltage to the CW ID. I used an R/C combination on the flip/flop so it would be reset when first powered up. A push button is provided so the flip/flop can be set manually (turning on the transmitter)

The transmitter consists of three stages: a 45 MHz oscillator, a tripler, and a "power" amplifier. I know the heat sink on the final is a bit of overkill at 250 milliwatts output. I used a 6 volt gell cell for power. This thing can go for days.

The CW ID consists of a CMOS 4011 clock and a CMOS 4020 binary counter connected to a 2k EPROM. One bit on the EPROM is used for the CW ID, a second is used to control the switching transistor which supplies power to the transmitter oscillator. It is set up to transmit once every few minutes and stay on the air for about one minute.

Normally, the travel alarm timer is attached to the outside of the 3x5 file card box with velcro. This makes it easy to change the clock battery when needed.

A word of advice. Do not try to take a device like this through the security gate at an airport. It looks too much like something you would see in a "Die Hard" movie. Please make sure it is plainly labeled as an Amateur Radio Transmitter Location Device (or some other official title) and put your name and address on it in case it is discovered by someone not familiar with fox hunting. They *might* think the FOX looks suspicious and call officials to investigate. (You might want to notify officials in advance about your FOX-HUNT. We usually do just to eliminate problems when neighbors call in reporting *strange* people in the neighborhood)

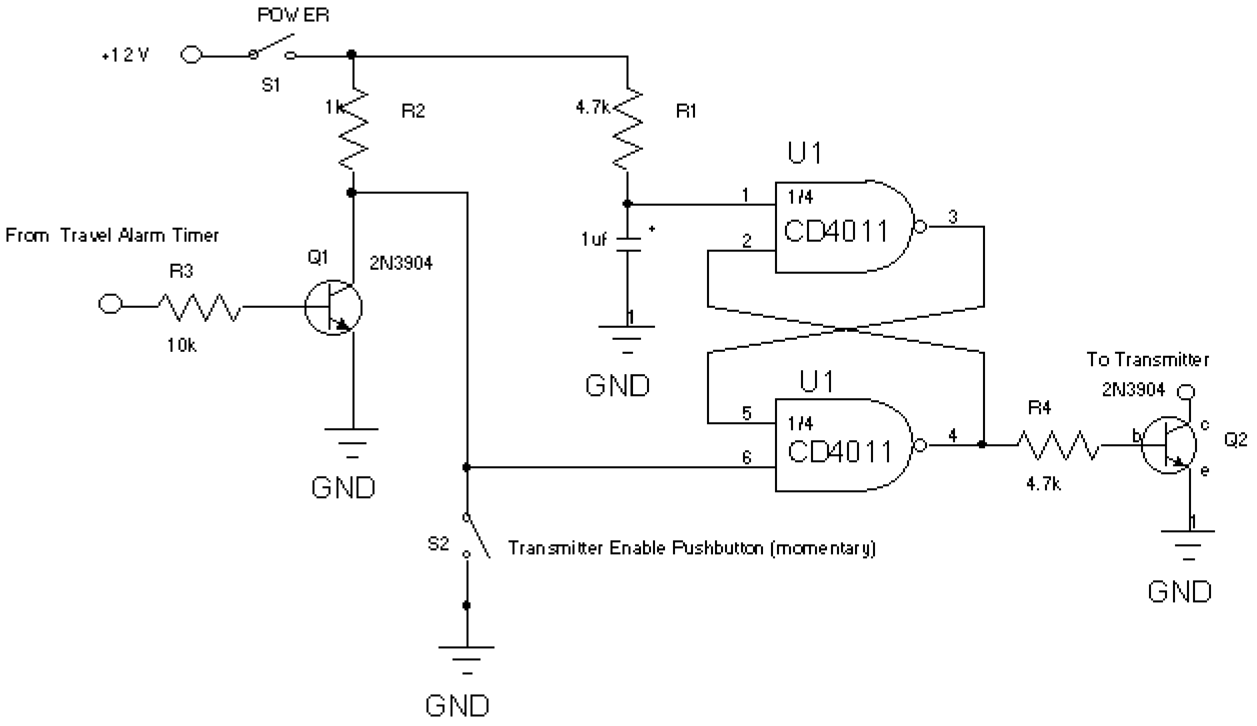
POWER
+12V
S1
1k
R2
4.7k
R1
U1
1
1/4
CD4011
3
2
1uf
+
1
GND
From Travel Alarm Timer
R3
Q1
2N3904
10k
GND
U1
5
1/4
CD4011
6
4
R4
4.7k
To Transmitter
2N3904
b
c
e
Q2
1
GND
S2
Transmitter Enable Pushbutton (momentary)
GND

THE FOX750 - 750 milliwatt transmitter

Joe Leggio WB2HOL

This three transistor 144 MHz transmitter uses a 48 MHz scanner crystal as an oscillator / tripler. A second transistor boosts the signal to 40 milliwatts, a third amplifies the power output to 750 milliwatts.

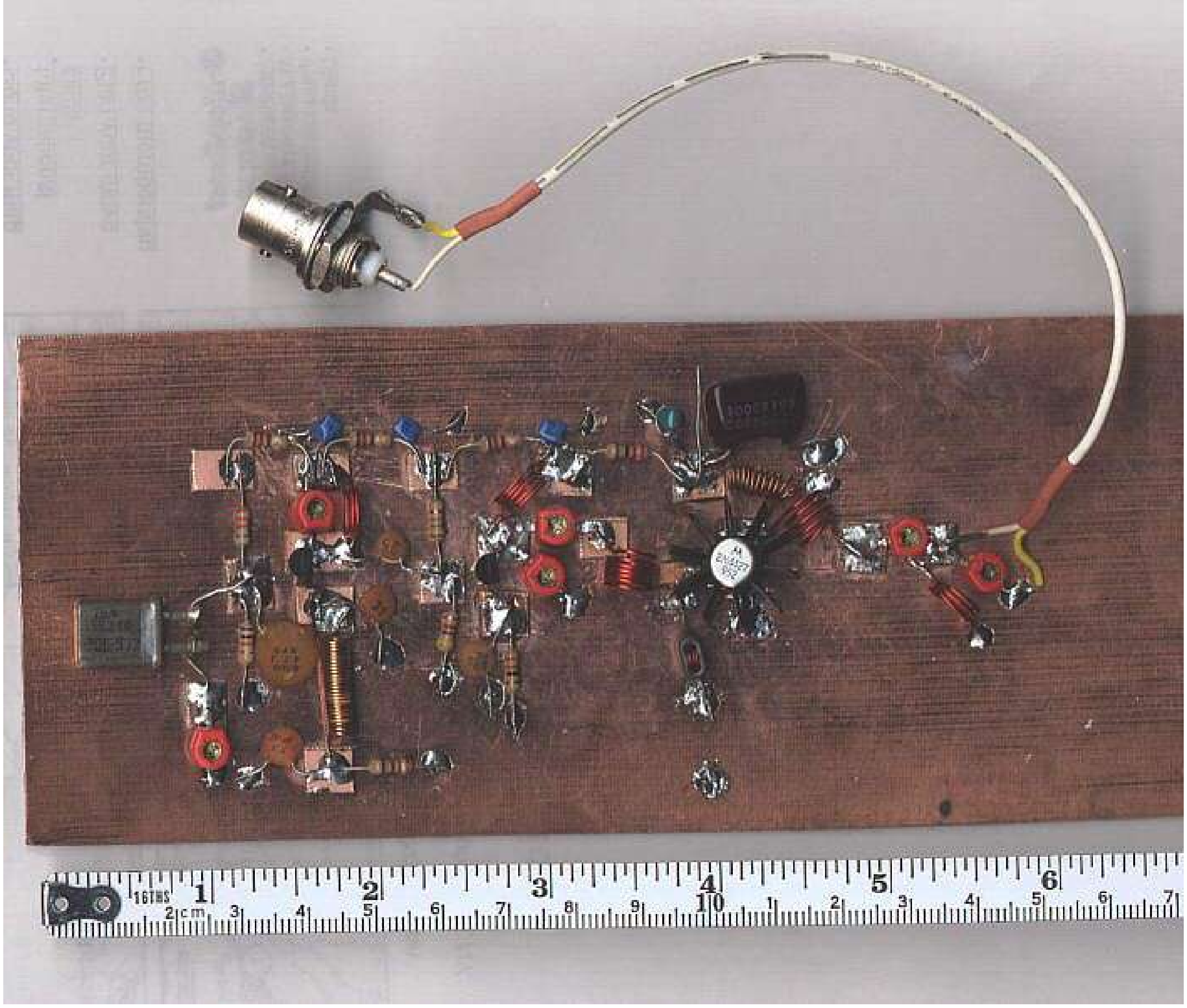

The scanner crystal I used was actually for a receive frequency 10.7 MHz higher than the two meter transmitter output. It was marked 156.21 MHz. (When you subtract the 10.7 MHz IF offset, it resulted in a transmit frequency of 145.51. Since this circuit is not the exact same circuit as the scanner, you may find that the crystal oscillates somewhere close to the marked frequency. The crystal marked as 156.21 MHz I used in this transmitter actually resulted in a signal on 145.53 MHz.)

I did get the salesperson at the local Radio Shack store a bit confused when I asked to look at his box of scanner crystals. The first thing he asked was which service I was trying to listen to... Police, Fire, Ambulance, etc.... When I said that I was looking for something for a homemade project and did not have a specific frequency in mind he understood. I did not try to tell him the crystal was to be used in a transmitter. He probably would have told me that I can't use a "receiver" crystal for a "transmitter."

As I looked at the available crystal selections, I simply took my calculator and subtracted 10.7 MHz from the marked receive frequency. I was lucky; several crystals ended up in places on the two meter band where a fox could transmit without interfering with other users. If you are not lucky, or live in an area where you must put your fox on a specific frequency, you can order a crystal through Radio Shack. Since there are still quite a few crystal controlled scanners out there, the odds are that someone in your town will have something you can use. Be aware, although most scanners use a 10.7 MHz IF, there are some with a 10.8 MHz IF. In that case, you may end up 100 kHz from where you expected.

You will need to wind several coils to build this transmitter. They are all self-supporting air-core coils made from enamel covered wire. If you follow the instructions illustrated here, you should not have any difficulty. The top of my workbench is littered with coils I have wound and tried in various circuits. I specifically avoided tapped coils or multi-winding coils to keep construction as easy as possible.

I use a simple diode detector across a 50-ohm dummy load to measure the power output while tuning up this transmitter. I did try about 5 or 6 different prototype circuits as I designed this transmitter; many were not stable. Their RF output did not smoothly change as the trimmer capacitors were adjusted but instead jumped as spurious oscillations occurred. This design is reasonably stable. It does not have a tendency to break into spurious oscillations. I used an oscilloscope to monitor the output waveform and confirm the lack of spurious oscillations. This is not really necessary as you can usually hear the instability if you monitor the output signal with your 2-meter rig.

Style of PC board construction

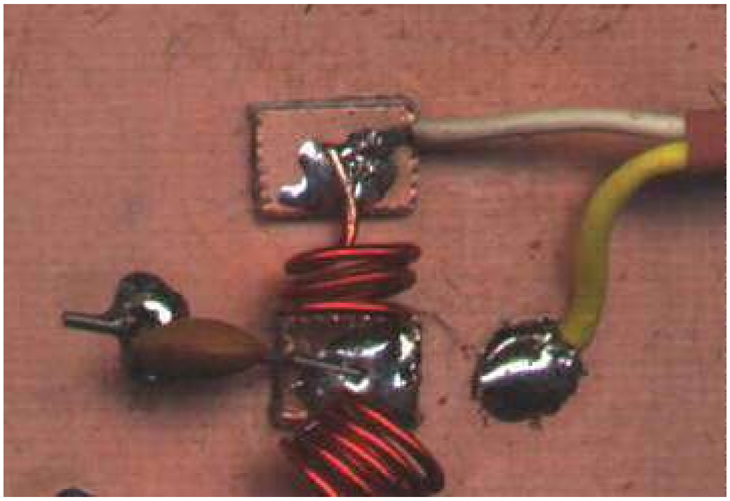

Tie-points are made from 1/4 inch squares of PC board materialglued with "crazy-glue" to a solid ground plane of PC board.

I built all of the transmitters using copper clad PC board as a ground plane and 1/4 inch squares of PC board as tie points. The tie points were cemented to the ground plane using "crazy-glue." This technique made constructing the transmitter go very fast. It only took a drop of glue to secure the tie points. The glue set up in seconds. It made it easy to experiment with different circuits. In a pinch, a tie point could be removed or relocated by prying upward on its side with a screwdriver. I built three transmitters using this style of construction in the past month. All have worked exactly as expected. I used a pair of shears to cut a number of 1/4 inch squares of PC board material. I then used a wide file to remove any burrs created by the shear. This allowed the square of PC board material to lay flat against the PC ground plane. (Crazy-glue is not designed to fill voids. It works best when the surfaces to be glued fit together well) Be careful as you work with "crazy-glue." It is very effective at gluing fingers together.

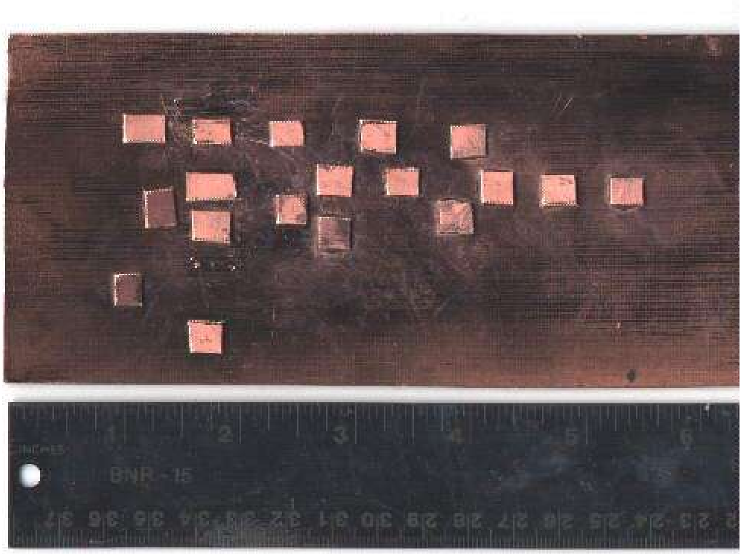

As I had already constructed several low power versions of this transmitter, I first glued squares of PC board material for the first two stages to the PC board ground plane. I then proceeded to add the components. I saved the output stage for last as I had not yet decided upon the interstage and output coupling curcuit design. The design I finally used was a combination of several from the ARRL handbook. I tried to use parts that could be obtained with little difficulty.

Construction is under way. The oscillator stage was constructed first and tested before proceeding.

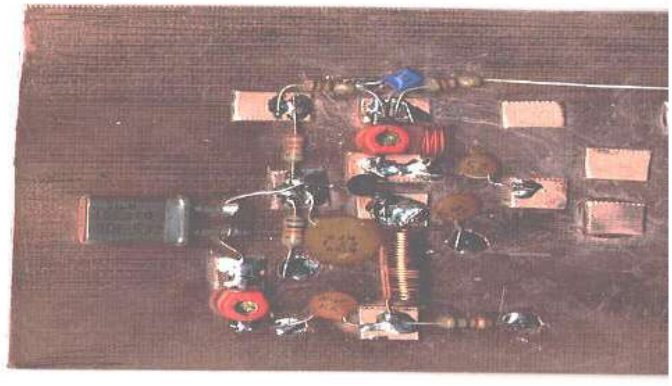

Winding Coils

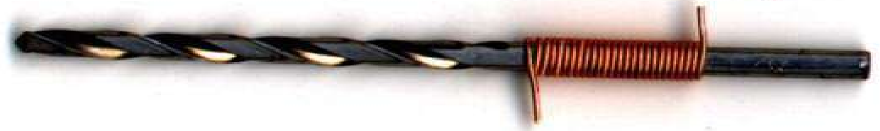

Enamel wire is close-wound on a drill bit to form the coils for this transmitter. It is easier to cut the leads to the coil and strip the enamel from the ends prior to removing the coil from the drillbit used as the form.Winding the Coils

You will need to wind several coils. I kept these as simple as possible. The wire I used was recycled from old transformers. All the coils are wound in a single layer using a drill bit as a winding form. To help you visualize how to form the coils, I have included an illustration showing how the coils are wound..

All of the coils are "close wound" with "air cores." I used a drill bit as a form and then slipped the coil off of the bit after I had cut the free ends to about 1/4 inch. I used a hobby knife to scrape the enamel paint from the ends of the wire before I removed the coil from the drill bit used as a form. This made the coils easy to solder. With any luck, you too will be able to wind your own coils for this project.

Two different gauges of enamel covered wire are used for the coils in this project. If you have wire reclaimed from an old transformer, or from an old television deflection yoke, you can determine its gauge by winding a single layer on a drill bit. You will be able to fit about 46 turns per inch if the wire is 24 gauge. You will be able to fit about 37 turns per inch if the wire is 22 gauge. If you can't find an old transformer with the correct gauge wire to recycle, the best places I have found to buy enamel covered wire are shops which specialize in rewinding electric motors. In a pinch, if you live in the USA, Radio Shack sells "magnet wire" in both 24 and 22 gauge. You will only need a few feet of each gauge wire for this project.

I used the same transistor for both the oscillator and the amplifier stage. It is important that the transistors you use in your transmitter have sufficient gain at 2 meters. The 2N5770 is one of many I could have used. Please note: a 2N3904 would probably work fine as the oscillator (at 48 MHz), it would not work well as the amplifier as it does not work well at 144 MHz.

You will need to use a heat-sink on the output transistor. I used a small press-on style. I gets warm to the touch after transmitting for a few minutes. If it gets very hot, you probably have something mis-tuned or the circuit is oscillating on it own If you follow the layout shown, you probably will not have any difficulty. If you try a different layout, be sure to keep the input and output coils of the various stages away from each other. It is too easy to create an accidental feedback path. If this occurs, the transmitter will be very unstable and almost impossible to tame. If you are experienced with VHF design, you may want to try winding the coils in toriod coil forms. Because toriods are self-shielding, you may be able to try other layouts without as much concern towards stability. Obviously, if you do this, you will have to determine the number of turns needed for the toriods. A grid dip meter would be helpful in that case.

When building the output stage, I realized that I needed to bend the base lead of the output transistor between the emitter and collector in order to install it properly. Although it did not affect performance, it did make soldering the transistor to the PC board pads a bit more difficult. If I build another version of this transmitter, I will lay out the board a bit differently so the transistor will fit better to the pads.

The tune-up of this transmitter is easy. Simply adjust each of the trimmer capacitors for maximum RF output to the antenna. (The trimmer in series with the crystal is the only exception, it is used to trim the transmitter frequency, use a frequency counter to adjust it)

Power output was between 750 milliwats and 900 milliwatts. I will probably add a three terminal voltage regulator so I can vary the output power for any specific hunt. As an experiment, I tried voltages as low as 5 volts. The power output was down to 100 milliwatts but the transmitter stayed on the air! This design works well even as the battery gets weak.

Schematic

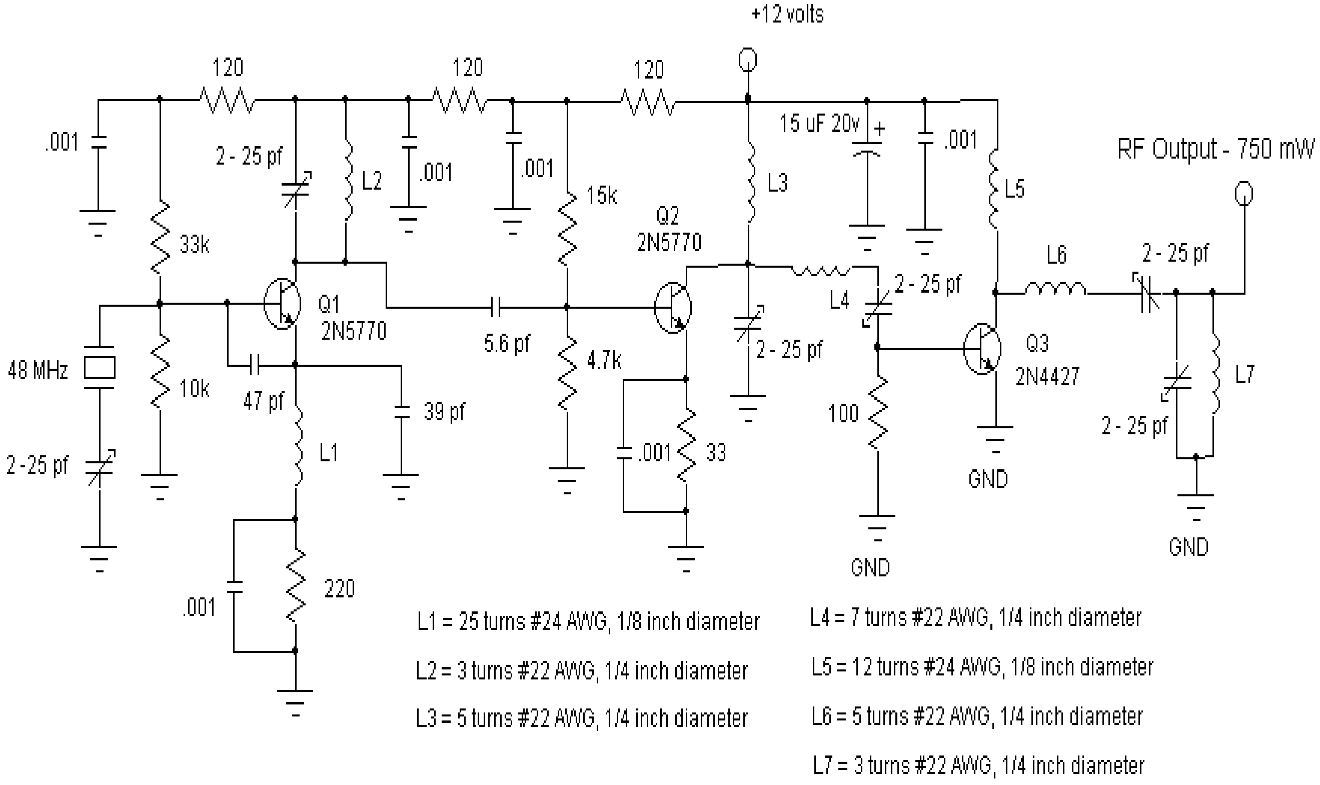
+12 volts
120
120
120
.001
2 - 25 pf
L2
.001
.001
15k
Q2
2N5770
L3
15 uF 20v
+
.001
L5
RF Output - 750 mW
33k
Q1
2N5770
L6
2 - 25 pf
L4
2 - 25 pf
48 MHz
5.6 pf
4.7k
2 - 25 pf
Q3
2N4427
L7
10k
47 pf
39 pf
100
2 - 25 pf
2-25 pf
L1
.001
33
GND
GND
GND
.001
220
L1 = 25 turns #24 AWG, 1/8 inch diameter
L2 = 3 turns #22 AWG, 1/4 inch diameter
L3 = 5 turns #22 AWG, 1/4 inch diameter
L4 = 7 turns #22 AWG, 1/4 inch diameter
L5 = 12 turns #24 AWG, 1/8 inch diameter
L6 = 5 turns #22 AWG, 1/4 inch diameter
L7 = 3 turns #22 AWG, 1/4 inch diameter

2 Meter - 750 mW FOX TRANSMITTER - 4/98 WB2HOL

SIMPLE ADJUSTABLE PASSIVE ATTENUATOR

Joe Leggio WB2HOL

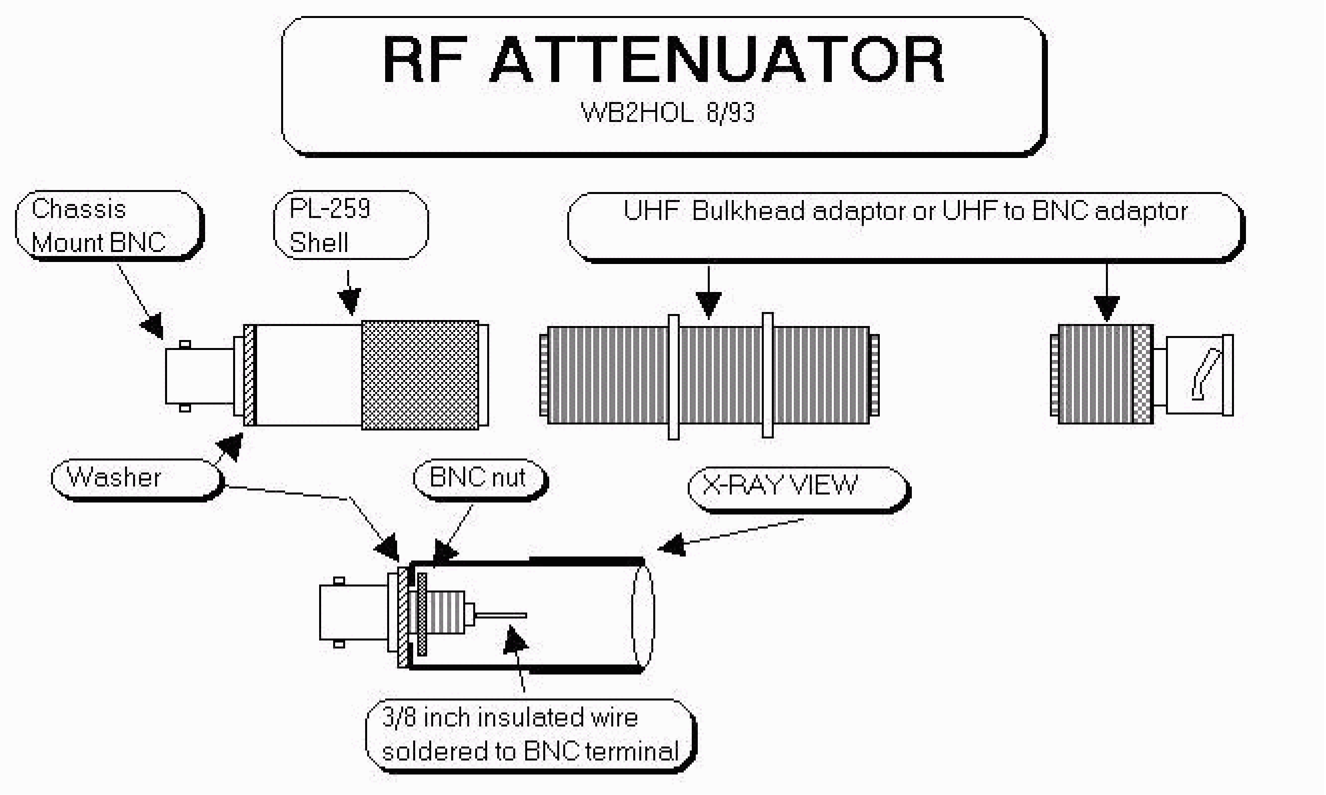
RF ATTENUATOR
WB2HOL 8/93
Chassis Mount BNC
PL-259 Shell
UHF Bulkhead adaptor or UHF to BNC adaptor
Washer
BNC nut
X-RAY VIEW
3/8 inch insulated wire soldered to BNC terminal

This adjustable passive attenuator works as a "waveguide beyond cutoff." It can be constructed in about ten minutes from parts available locally.

This attenuator was the result of a search for an inexpensive attenuator which could be used on an HT to help beginners equip themselves for a hidden transmitter hunt. The 3/8 inch piece of wire will extend into the center conductor of the bulkhead adaptor forming a small capacitor. The coupling between the two conductors will vary as the PL-259 shell is turned further onto the bulkhead adaptor.

If you are able to find a BNC to UHF adaptor with a long threaded portion, you may use it instead of the bulkhead adaptor.

The maximum attenuation available depends upon the length of the wire soldered to the chassis mount BNC connector. A service monitor measured a loss of up to 36 dB. With the cost so low and construction so simple, many hams have made several with different length wires.

Although 36 dB may sound like a lot of attenuation it will not be suffucient when on most transmitter hunts. This passive attenuator project is aimed at the beginner who is armed with only an HT and is attempting to use the "body-fade" method of Radio Direction Finding. The passive attenuator can be put in-line with the existing rubber-duck antenna and adjusted so the signal no longer is full scale on the S-meter.

Tuning off frequency by 5 or 10 kc, listening for the third harmonic, using a foil wrapped tube as an RF shield around the HT, or even wrapping aluminum foil directly around the HT are all methods which will help to some degree. One or more of these techniques will reduce the signal to where a "S" meter reading peak (or null) gives you a clue to the direction where the transmitter might be found. (If you do wrap aluminum foil around your HT, put some tape over any exposed battery charging contacts. Otherwise, the foil may get a bit warm as it rapidly discharges the battery)

When closer to the fox, more signal will enter the HT through the case than through the antenna. At this point, other methods must be used to locate the fox transmitter. The most effective attenuator in most cases is an "active attenuator." They can offer over 100 dB of attenuation. Plans for one are available here.

Construction Details

This BNC to UHF adapter has a raised knurled portion of the barrel which prevents the PL-259 shell from being tightened completely. (The BNC and insulated wire do not "bottom out" when tightened all the way) You can file that section of the adaptor down and enjoy more range in available attenuation.

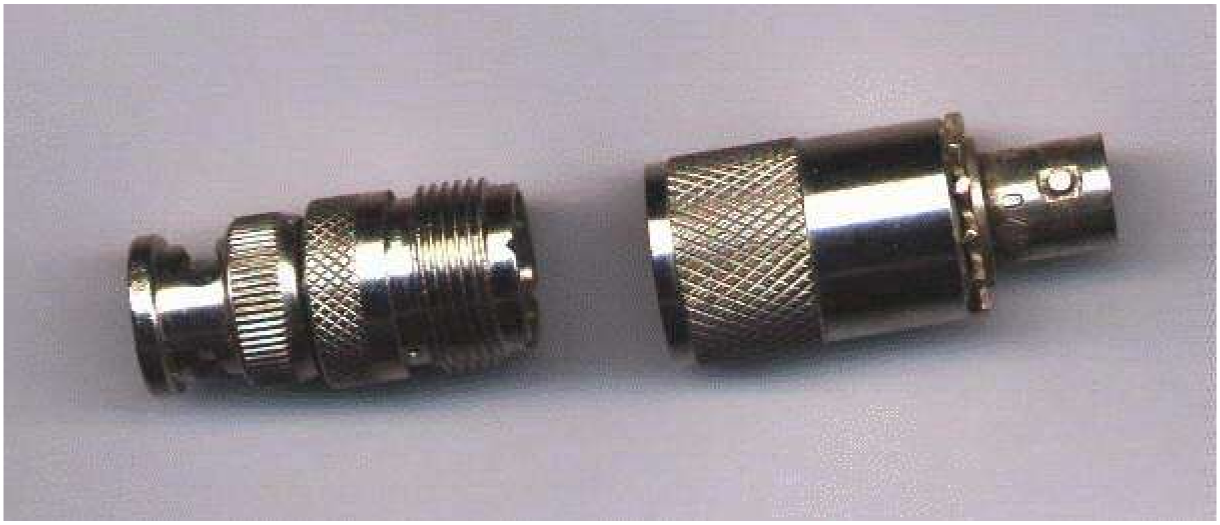

Looking in at insulated wire I soldered the wire to the BNC before mounting it in the PL-259 shell. I used a length of wire clipped from a 2 watt resistor. (Use any stiff wire) I then forced a small length of insulation stripped from another piece of wire onto it. By using this method, the insulation does not melt during assembly.

I used a "star" washer. It was just the right size. The washer prevents the BNC connector from falling through the PL-259 shell. The hardest part of the construction is tightening the nut on the BNC connector. I usually wedge a screwdriver in the shell to hold the nut and tighten the BNC from the outside.

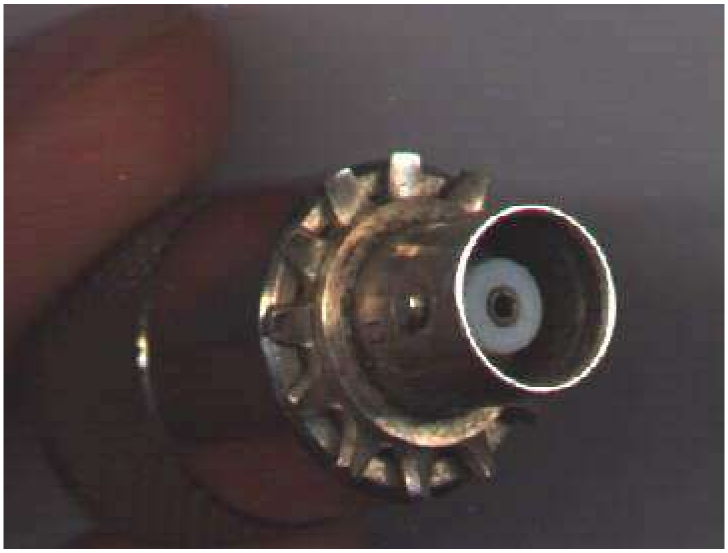

ACTIVE ATTENUATOR

Joe Leggio WB2HOL

ACTIVE ATTENUATOR
WB2HOL 6/96

How it works

This active attenuator works as a simple mixer. A signal from a 1 MHz oscillator is mixed with the received signal in a 1N914 diode. The 1000 ohm potentiometer adjusts the level of the local oscillator fed to the mixer diode.

It is possible to get over 100 db attenuation using this device. This circuit is a variation of one first described in the November, 1992 issue of QST. That circuit was designed by PA0ZR. It used a L/C tuned 500 KHz oscillator.

I used a 1 MHz crystal oscillator and it made construction even simpler than the original unit.

This attenuator is used by tuning the receiver 1 MHz above or below the frequency of the transmitter. By adjusting the oscillator level fed to the mixer diode, the signal output to the receiver will be attenuated.

Note: You may also tune your reciever 2 MHz above or below the frequency of the transmitter and get even more attenuation. (thanks to VE3EFY for this tip)

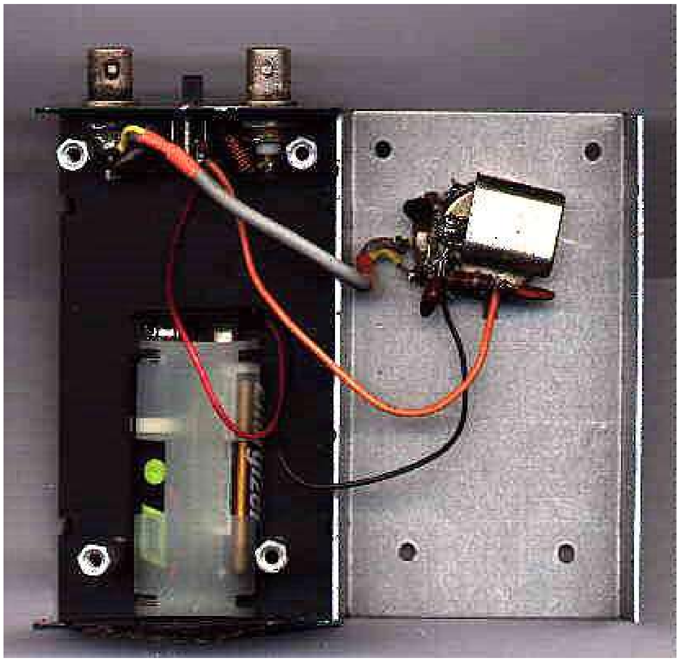

Technical Hints

If you find that this circuit does not oscillate with your crystal you might need to add a small capacitor (anything from 5pf to 25pf) from the gate of the FET to ground. It may also help to reduce the 150pf capicitor connected from the drain to ground to 100pf.

These two capicitors, along with the crystal, provide a 360 degree phase shift from the drain to the gate of the FET. The circuit pictured in the schematic below is a Pierce oscillator and relies upon the gate-to-source capacitance of the FET to form part of its phase shift network. If there is not sufficient stray capacitance to ground from the gate of the FET, you may need to add the additional capicitor. VE3EFY reported that his oscillator would not function below 4.5 volts. He used a 9 volt battery to power his unit. The battery should last a long time, he measured the current drain to be 1.8 mA.

Construction

I built this unit "dead-bug" style. Most of the components were mounted and supported by their own leads on the back of the 1000 ohm potentiometer. The RF Choke connected between the output connector and ground was made by winding about 20 turns of 22 guage wire on a 1/8 inch drill bit. The coil was then slid off of the form and wired in place. If you want to use a commercially available coil, a value of 3.3uh should work just fine. That was the value specified in the original article.

I mounted the 150pf capicitor, the 1N914 diode, and the RF Choke on the inside panel of the attenuator box between the two BNC connectors. I also mounted a slide switch between the BNC connectors so I could conserve the batteries when the unit was not being used on a hunt.

Note: The MPF-102 Field Effect Transistor has an unusual pin-out compared to general purpose transistors. The GATE is not the lead in the center. The pin-out for the MOTOROLA MPF-102 is illustrated below. (The Radio-Shack MPF-102 I used had the same pin-out)

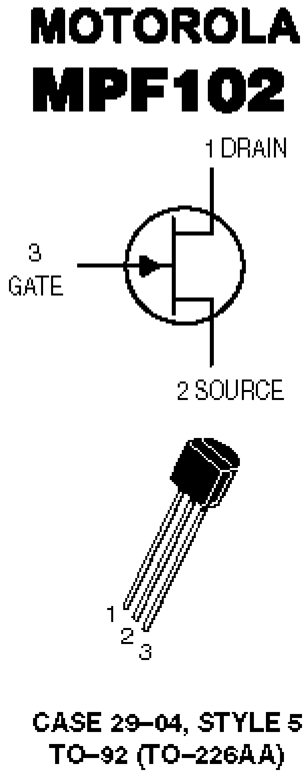
MOTOROLA
MPF102
1 DRAIN
3
GATE
2 SOURCE
1
2
3
CASE 29–04, STYLE 5
TO–92 (TO–226AA)

The Schematic

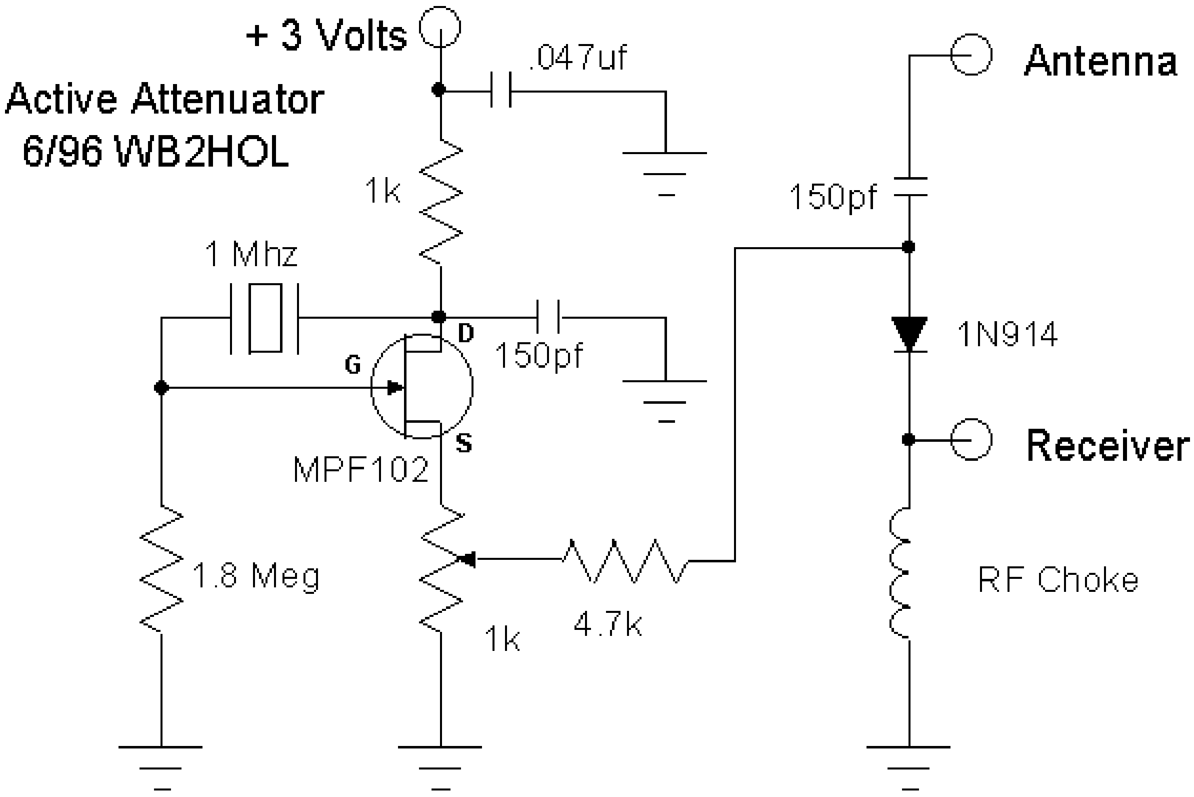
+ 3 Volts
Active Attenuator
6/96 WB2HOL
.047uf
1k
1 Mhz
D
G
S
150pf
MPF102
1.8 Meg
1k
4.7k
Antenna
150pf
1N914
Receiver
RF Choke

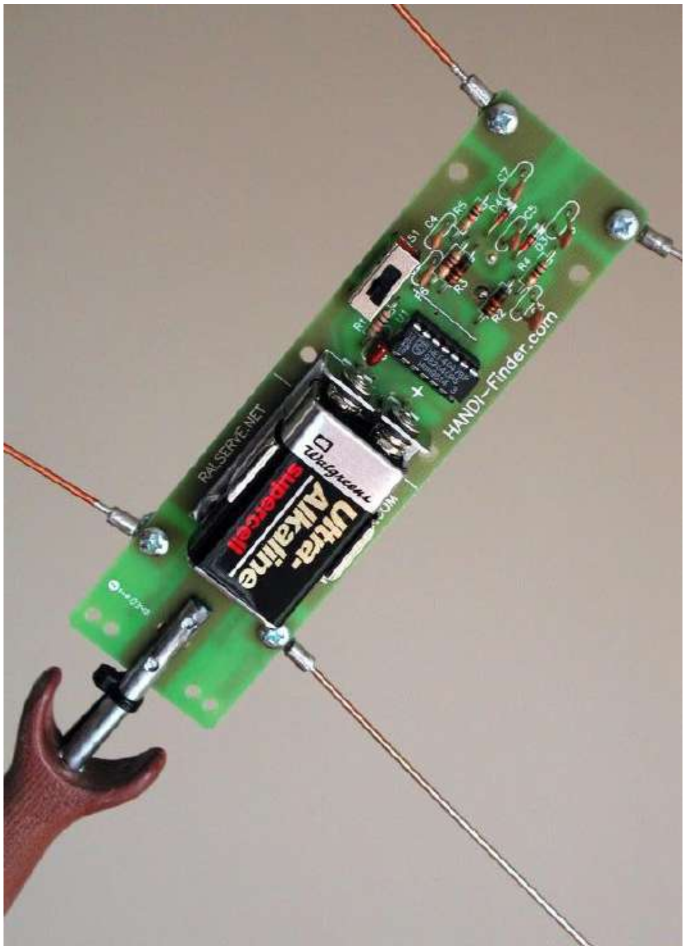

The HANDI-Finder®

The HANDI-Finder® is a HANdheld DIrection Finder which can be used to localize both AM and FM carrier-based sources using a single connection to the antenna input of an FM receiver tuned to the frequency of interest in the range of 45 to 450 MHz. It has been designed for bare-bone functionality as well as low power consumption, simplicity, and economy!

Because it is both an easy-to-build kit for the beginner and a convenient basis for further experimentation by those with more experience, it makes a great club project! Overall, it is a quick, inexpensive way to implement the concept of direction finding, and provides something relatively compact to keep readily available for locating sources of interference.

Except for adding a handle, fabrication is minimal. Two open-loop antennas are made from coat hanger wire bent into halves of a "bow-tie" shape and mounted directly to the circuit board, as illustrated in the "layout" drawing.

The coax downlead is connected to an FM receiver and detects the carrier regardless of whether the transmission mode is AM or FM. The unit is switched on (UP) and rotated for a null in the audio tone that it adds to the audio coming out the receiver. The signal direction is perpendicular to the plane of the antennas. There is 180-degree ambiguity, but this is not a problem in actual use because multiple "bearings" must be taken anyway in order to establish a "fix" on the location. Once close in, the operator works toward the general sense of the direction, and again the ambiguity is not a problem.

The 3-position switch is moved DOWN to stop the tone but still keeps the antenna activated for monitoring and standby purposes, --or it is moved to the CENTER position to shut off the unit. Refer to the extensive discussion in the manual which describes construction variations, and the concepts behind the design.

"Build the HANDI-Finder" appeared as an article in QST Magazine, May, 1993. All the information in that article and much more is included here! By clicking on the appropriate boxes in this web page, all the drawings and text that come with the kit can be printed out or downloaded from PDF files.

may be obtained by contacting:

Northern Ohio Amateur Radio Development

P. O. Box 271
Brunswick, OH 44212-0271 USA
Tel: 330-225-7373

The HANDI-Finder® Experimenter's Kit

by Bob Leskovec, K8DTS

6th Edition,

The HANDI-Finder® is a HANdheld DIrection Finder which can be used to localize both AM and FM sources using only a single connection to the antenna input of a VHF-FM receiver tuned to the frequency of interest.

The basic HANDI-Finder® when equipped with open-loop wire antennas and a short handle, stores flat and ready-for-use in a briefcase. It works well with just an HT (handy-talky).

The HANDI-Finder® has been designed for low power consumption, simplicity, and economy. The goal was to provide an easy-to-build kit for the beginner and a basis for further experimentation by those with more experience. Overall, it is a quick, inexpensive way to introduce users of FM communications equipment to the principle of direction finding and give them something with which they can help locate sources of malicious interference.

A HANDI-Finder® Experimenter's Kit has been put together to conveniently supply the essential parts necessary to build the control circuit powered by an on-board 9V battery. The user provides the coax, cable connector, and appropriate antennas. By using two open loop antennas made out of coat-hanger wire, the unit can be put into immediate use!

An equally valuable part of the kit is this instruction manual which describes variations in antennas, general construction and detailed discussions of the circuit and components, to encourage experimentation and ideas for improvements.

The HANDI-Finder® first came together in October of 1986, and continued to evolve. A write-up later appeared in May, 1993 QST Magazine, entitled “Build the HANDI-Finder”.

TECHNICAL OVERVIEW:

The electronic circuit is based on a design credited to Tom Feierabend SO/CM 03N18 circa 1979 which appeared in a manual published in May, 1980 by Van Field, DCP XVIII, entitled “VHF Radio Direction Finding Manual for Coast Guard Auxiliary Use”.

A similar circuit, referred to as the “Double-Ducky” direction finder (DDDF) designed by David Geiser, WA2ANU, is described in July, 1981 QST and reprinted in the 1983 ARRL Antenna Handbook.

The USCG AUX-03N18 version uses an LM555 driving two successive stages of 7404 TTL to provide complementary buffered outputs. Since TTL requires a 5-volt power supply, yet another IC, a 7805 or 78L05 is needed, to regulate the stated 6-30 volt input range.

One problem is that the LM555 does not easily put out a symmetrical square waveform, which is useful in this application. The antenna assembly consists of two vertical ½-wave dipoles (37” long) mounted 8” apart on a boom. This assembly is rather bulky and quite a bit of mechanical fabrication is required.

The QST circuit uses only one IC, the LM567 Phase-Locked-Loop Tone Decoder. This is a complex chip that contains an oscillator and other circuitry, including an output circuit which does put out a symmetrical square wave. However it does not have simultaneous complementary outputs. In the DDDF the single output is connected to the diode switches through a non-polar capacitor, and some adjustment is required to achieve the proper switching level. Data sheet specifications show that the LM567C typically draws 12mA. The antenna assembly consists of two "rubber duckies" mounted 10" apart on a 4.25" X 18.5" ground-plane. While the ground plane requires much less work to fabricate than the USCG-AUX circuit, two "rubber duckies" must be obtained.

The HANDI-Finder® is a good example of integrated simplicity, wherein one simplification contributes to another.

First, it uses a single CD4047B CMOS IC, which contains both an oscillator and a divider flip-flop to automatically provide complementary symmetrical square wave outputs without special adjustments. Only a single resistor and capacitor are needed to set the frequency. While many have not heard of this IC, it continues to be available from standard suppliers like Digi-Key, Mouser, and Jameco.

Second, very little current is used to bias the switching diodes so the total current draw is only 1.7mA at 9V. Good service can therefore be provided by a common alkaline "transistor radio battery" and there is no need for wiring to an external source such as a 12V vehicular supply. This, in turn, eliminates the need for noise filtering. The operation of the circuit is not dependent on battery voltage, so a regulator is not required. Supply voltage for the CD4047 can be anywhere from 3-18 volts. Finally, since all the parts, including the battery are mounted on a single circuit board, the board is designed to also serve as the mounting base for the two open-loop antenna elements easily made out of bent wire. Thus, there are no ground-plane or vertical dipole elements to make, nor "rubber duckies" to buy, and no case to drill, until later if you really want to!

HOW IT WORKS:

All three circuits described above are based on the same principle. An electronic switch alternately connects two antennas to the coax cable downlead going to the antenna input of an FM radio receiver tuned to the frequency of interest. First one antenna is connected, then the other, etc., back and forth with equal intervals. This is done at an audio rate, well within the audio bandpass of the receiver, and usually in the range of 400 to 1500 Hz. A good frequency is 1000Hz.

Of the two antennas, if one is slightly closer to the source, it receives the wave front slightly earlier in time (phase) than the other. There is a phase difference in the signal received by one antenna compared to the other. Since the receiver is being switched between the two antennas, the switching action imposes phase modulation on the incoming signal. This is detected in the FM receiver and is heard at the audio output as a tone equal to the switching frequency. The amplitude of the audio signal corresponds to the deviation, which depends on the physical separation of the two antenna elements, up to ½ wavelength. In other words, if the antennas are farther apart the circuit will impose a higher percentage of modulation or a larger deviation, producing a louder tone, but the modulating frequency will stay the same.

If the antenna is rotated so that the plane of the two elements is perpendicular or broadside to the direction of the signal, both elements receive the signal at the same time (phase) and there is no longer a difference in phase. Hence, the audio tone disappears. This is perceived as a rather sharp null in the audio as the antenna array is rotated into position perpendicular to the direction of the signal.

This type of direction finder has the disadvantage that it exhibits 180-degree ambiguity. However, it has several advantages:

1) It works on a nulling principle rather than a peaking principle. The null is sharp and much easier to detect than the peak from a directional or beam antenna.

2) When you null the superimposed audio, you are not nulling the carrier. This is unlike a conventional loop antenna or cardioid array, which nulls out the carrier. The problem with carrier-null, is that as you get closer to the null, the signal you are trying to hear in order to null out, is getting harder to hear! Also, when you null the superimposed audio, you can still hear the audio coming from the source.

3) Since audio is being nulled, the operator does not have to watch a field-strength meter. He only needs to listen, which is something he can do while driving, riding a bike or walking.

4) Since this method uses phase information, it works well with strong signals, so no attenuator is required. (By comparison, the signal from directional gain antennas must be progressively attenuated to keep the receiver RF within the range of the S-meter.)

REFERENCE INFORMATION:

The HANDI-Finder® is an evolving project which underwent several changes (hopefully for the better) during the early stages. However, the basic circuit remains essentially the same. This manual might be supplied to help you work on an earlier unit, as far back as 1986. If you have a model that does not correspond to the pictorial description in this document, you should be able to identify the components with only minimal tracing of the etched circuit pattern.

Your unit may also have component values that differ from the schematic. If you did not buy a kit, you may have obtained the circuit board and documentation from separate sources that copy and distribute such things and try to make them “public domain”.

The circuit board versions are identified by a prefix letter “A”, a 6-digit date code (YYMMDD), and an optional suffix. (The suffix, if present, indicates very minor changes.) As of this writing, the numbers used are: A861003, A860102, A870122, A890422, A01110, and A031123.

In general, if the board has a date code earlier than the documentation, the component values in the documentation take precedence, provided the integrated circuit type is pin-compatible. If the board has a date code later than the documentation, and/or a different IC pinout, you would do well to send a self-addressed-stamped-envelope (SASE) to the distributor for an update of the layout and parts list. Of course, be sure to furnish the number on your board, so you will be sent the correct information. Please indicate whether the RALTEC® or other trademark is present and we will help you figure it out. Try contacting the author at ral@ralserve.net. (If you have trouble, you can also find the author’s current snail mail and e-mail addresses through the internet Ham-Call database.) For information, try the website handi-finder.com.

CHECKING THE COMPONENTS

If you received this instruction manual as part of a packaged kit along with the circuit board and loose components, unpack all the parts and check the quantities and values against the Component Parts List before you do anything else because some parts have the same value but different mounting positions and lead lengths. Before proceeding with actual construction, it is suggested that you read this manual to see if there are any changes you would rather incorporate right away. Then proceed with "ASSEMBLING THE KIT" which is located at then end of this discussion.

NO PILOT LIGHT:

Considering that a small transistor radio draws about 10mA, it can be seen that the battery will last quite a while if the user remembers to turn it off. Unfortunately a pilot light would be self-defeating, since it would draw 5-10 times the current of the circuit itself!

How many times have you forgotten to turn of your HT? After a while you finally learn to double check. With the HANDI-Finder® it should even be a little easier, because you will probably be disconnecting it from your radio or otherwise storing it at the end of a "DF'ing" session. That action should help remind you to check that the switch is off. Better yet, remove the battery. That's why we have included a top grade battery holder in the bare board version.

ON-OFF SWITCH STATUS INDICATOR:

For models prior to A890422, here are a couple of things to help you tell OFF from ON in the absence of a pilot light. First, the "ON is UP" convention has been followed. This is fairly commonplace on most equipment. Second, you can make a "passive" indicator by using a dot of bright paint, for example, typewriter correction fluid such as "whiteout". Push the slider to the ON position, then paint a small dot on the lower part of the slider that is now exposed. Let the paint dry thoroughly so it doesn't rub off. When you slide the switch to OFF, the painted part should be hidden; when ON, it should be visible. On Model A890422 and later the switch has 3-positions: OFF is in the center, UP is for DF'ing, and DOWN is for straight receiving or Standby. (Notice the design date is embedded in this code as YYMMDD.)

ON-OFF SWITCH MECHANICAL VARIATION:

There are a couple of subminiature slide switches available from "experimenter" sources such as Mouser and Digi-Key, but they differ slightly in the spacing and style of the connector pins. Circuit boards A861003 and A860102 used the CW Industries switch available from Digi-Key as SW103-ND.

Starting with circuit board A870122, the switch pads were made larger and the spacing changed to use the Mouser 10SM007 or 10SP001. The larger pads allow larger holes to be drilled to allow for the wider flat pins. If the switch you have does not go right in, do not force it. Study the problem and carefully enlarge the holes only where necessary. A small modeling file is handy for this purpose. Don't make the holes any larger than necessary, or you will have trouble bridging the gap and getting a good flow joint when soldering. When you do solder it into position, use ample solder and heat it enough so the solder surrounding each terminal flows evenly into a nice even form similar to an "Indian tepee." A890422 and later use a DP3T switch with 6 pins, which will not fit the previous models, but that gave the ability to add the "standby" position which is very useful.

BATTERY HOLDER:

The 9-volt battery holder is fastened at one end by soldering the two terminal tabs that pass through the board. The other end of the frame can best be fastened by using 1/8" diameter "pop-rivets", but you may alternatively try hot-melt glue, epoxy, small screws, etc. However, make sure that whatever you use won't protrude and prevent the battery from seating in the clip.

QUICK HANDLE:

The bottom area of the circuit board contains an area where the handle is attached. The unit can be mounted on either a short handle or directly to a mast, whatever you desire. To get going quickly, take any convenient piece of wood or metal, lay the end against the board, mark the holes, drill them through the handle, and attach with screws. If you use a round handle, you will either have to make a flat cutout along one side, or cut a slot in the end so that the board can slide in. A round handle is best. The most utilitarian handle is described next. .

PAINT-ROLLER HANDLE:

Find an inexpensive paint roller, but one with a handle which is threaded for an extension pole. (Home supply stores sell them typically for $1.39). Study how the handle is attached and determine the best way to remove it or otherwise adapt it so the HANDI-Finder® can be mounted on it.

Most of the handles available these days are plastic, force-fit directly to the metal roller rod. Use a hacksaw and cut off the roller assembly rod so that about 3" of rod is left protruding from the handle. File the rod enough to have a flat area, center punch, then drill two #6 holes to match up with the circuit board, and attach. The flat side should of course go against the board.

Next, shop around for your choice of a wood or metal paint roller extension pole, preferably the type which is made of 3 sections which screw together. You will thus have a very flexible system. You can use the unit with just the handle alone, or screw on up to three lengths of additional "mast". You may even wish to obtain a second extension to have more lengths available. Be aware that some extension handles have different threads in the sections than they have at the roller end. In any event, the pieces are easily stored when disassembled.

UTILITY-DUSTER HANDLE:

There is a homemaker's utility duster on the market called a WEBSTER® which can be found for as little as $4 in discount stores. Unfortunately, only a small percentage of stores seem to stock it. However, this is an amazing value for our purposes because it contains not only a 6-inch detachable handle, but includes a removable, telescoping, extension pole!

The dusting head consists of an 8-inch diameter half-spherical array of bristles embedded in the end of the short handle. This end is easily sawed off to allow for the HANDI-Finder®. The extension pole is about 36" in the collapsed position and 60" when extended.

HANDLE, CAUTIONARY NOTE:

Mount the handle against the component side and put the screws through with the heads on the solder side, to keep a low profile for the coax cable which will come down the solder side

The HANDI-Finder® works best when the first ¼ wavelength of coaxial cable downlead to the receiver is kept vertical or parallel to the center line of the circuit board. (Measure the ¼-wave from the bottom antenna terminal.) If it waves around, it can throw off the bearing. Therefore, when using the unit with a short handle, make sure to hold it straight and high so that the cable hangs straight down. When using the unit with a long handle it is preferable to attach the coax to the mast for the first ¼ wavelength. Use tape, nylon cable ties, etc.

RESISTORS vs. INDUCTORS:

The HANDI-Finder® was designed mainly out of the need to inexpensively provide Radio Amateurs with something they could use to help locate interference on the 2-meter repeaters. But the unit operates over a much wider range of frequencies because the chokes "traditionally" used in such RF circuits have been replaced by resistors.

Inductors are usually used to feed direct current into some point in a circuit where it is desired to allow an alternating current signal to pass without attenuation. At radio frequencies these are called "RF chokes". Circuits which operate in the 150MHz range, for example, usually use small coils with a value of 1.0 to 1.2 microHenries. Using the expression Xc = 2(3.14)fL, where "f" is the frequency in MHz and L is the inductance in microHenries, the corresponding reactance works out to about 1000 Ohms. At 1/3 the frequency, or 50 Mhz, this same inductor would have a reactance of only 333 Ohms. At UHF this inductance would mathematically exhibit a proportionately higher reactance, but other difficulties arise. Certain assumptions about the construction of the coil are no longer valid and the math becomes more complicated. Factors like the "capacitance between turns" and the length of the connecting leads can no longer be ignored. At some frequencies the coil looks like a high impedance, but at others it may look like a short circuit!

Thus, the frequency band over which the circuit can operate is limited by the chokes. Different sets have to be installed to operate over different ranges of frequencies. Normally it is necessary to use chokes because they have a comparatively low resistance at d.c. so there won't be any significant voltage drop. The key word here is "significant".

In this circuit, the switching diodes are biased by current. As long as enough current is supplied to do the job, it doesn't matter if a little voltage drop occurs across the connecting element. The chokes can simply be replaced by 1000 Ohm resistors. The currents are so small that the voltage drop is negligible. Further, for all the complications involved with determining how chokes will act at different frequencies, it is no less risky, and a great deal easier, to assume that a resistor will exhibit the same resistance over a much wider frequency range.

In this circuit R4, R5, and R6 would have "traditionally" been specified as 1.0uH chokes for operation in the 150 MHz region. As you can see, by using 1K resistors, operation at 150 MHz should not be affected, and operation over a wider range of frequencies should now be possible. The value of the resistors is not that critical. Values in the range of 1K-1.3K will do, but make them all the same value. Using carbon-film resistors, we also get some useful inductance, but since the resistance is high and distributed with resistance, these are inherently free of self resonance, and broadband becaue they are low Q.

However, if you wish to experiment with inductors, you can either make them or buy them. According to the USCG AUX article, you can make these with one layer of #28 close wound on a 1 Meg ½-watt carbon composition resistor. Actually there is nothing magical about the 1 Meg value, it is the physical size and shape that is more important. The intention is that the resistance be at least a factor of 100 times greater than the inductive reactance. Therefore any value above 100k is acceptable.

Mouser Electronics has a line of good quality subminiature RF chokes that are quite reasonably priced, roughly $.23 each. Consult their catalog for the “43LQ” series. The 1uH value is part number 43LQ106. Keep in mind they probably will not fit on the HANDI-Finder circuit board supplied with the kit.

WIDEBAND OPERATION:

The HANDI-Finder® should work over a wider range when resistors are used in place of inductors, but there will still be problems with resonances (series LC) and anti-resonances (parallel LC) for any given set of antennas. However, ignoring those effects, let us discuss the effect of antenna spacing. Remember, you want the vertical members of the two antennas as far apart as practical to get maximum modulation.

This will make it easier to tune for a null.

For a given set of antennas, the upper frequency limit should be that which corresponds to a ½ wavelength equal to the widest spacing of the vertical members.

If the usual wavelength formula is multiplied by 12 inches/ft, a new “constant” factor is obtained: dividing 5616 by the frequency in MHz will give the ½ wavelength in inches. For the open-loop antennas you will be instructed to make elsewhere in this manual, the spacing between opposing vertical sections is typically 17.5 inches. Turning our formula around, and dividing 5616 by 17.5 gives 321 MHz, which is also the frequency at which maximum deviation would be available. Some have reported operation to include the 70cm Ham Band.

Moving to 1/10 that frequency or 32 MHz, only 1/10 the modulation would be available, but the units should still function. So it would appear that operation over a 10:1 range might be possible.

COUPLING CAPACITORS:

At the lowest limit, it is important that the coupling capacitors have a reactance of 50 Ohms or less so as not to attenuate the signal by more than 3dB. If we choose that limit to be 27MHz the value of capacitance having 50-Ohm reactance is 117pF. Values of .001uF, or 1000pf will be one tenth that or 5 Ohms, and work just fine.

SCREW TERMINALS:

Antenna connections to the board are made using screw terminals. On early units, these are formed by installing a clip-type “tinnerman nut” which is slipped onto the circuit board edge at each designated point, and then threading in a 6-32 x 3/8” or ½” binder head machine screw. Only four are needed for any given configuration. Later units use regular hex nuts, serrated lock washers, and flat washers against the board to keep the lock washers from tearing up the copper foil.

When fastening small diameter or stranded wires to these, it is suggested that you first install crimp-on spade lugs or ring lugs on the end of the wire. If you must wrap the wire around the screw, do it under a washer and wrap it in the direction that the screw tightens (clockwise).

BE CAREFUL NOT TO OVERTIGHTEN. If you experiment a lot, the tinnerman nuts will strip and wear out. If you need to have things very tight, you can use 6-32 screws with regular hex nuts. However, do not put serrated lock washers against the surface of the board since they really dig in and cause problems. It is better to use a slightly longer screw with flat washers to protect the surfaces. If you are tempted to just solder directly to the board, don't do it! The heavy heat may cause the foil to delaminate. Further, if the antennas are bumped, the foil may be pulled off the board. If you want to have some sort of "rounded off" nut on the thread end, inquire about "acorn nuts" in the screw specialty section of most hardware stores.

ANTENNAS:

The HANDI-Finder® is quite versatile in the way it can be used with different kinds of antennas. First of all, you can get it running immediately without extensive fabrication because two simple wire antennas can be attached directly at the circuit board. For 150 MHz, take two EQUAL lengths of stiff wire about 19 to 20 inches long and bend each one into a neat square "U" shape. The bottom of the "U" should be about 6". Form the ends into a hook and fasten them to the screw terminals on the circuit board. Looking at the component side of the board, you will see there are three terminals along the left side and three along the right side. Fasten one end of one antenna to the very top terminal on the right side. Fasten the other end of that same antenna to very bottom terminal on the same side. Then repeat this procedure for the left side.

Note that the bottom terminals are merely mounting points. They are electrically isolated. On some board designs there are circuit pads to allow installation of a grounding jumper. DO NOT INSTALL THIS JUMPER! It is desired that the antennas be open loops. If you ground the bottom of the loop, you will create a closed loop that will cause a carrier null in the direction of the signal. This is not desirable. On later circuit boards provisions for grounding was eliminated to avoid confusion.

The wire you use should be thick enough to provide desired rigidity, but thin enough to allow fastening under the screw terminals. If it is too thick, you might be better to first solder on some spade lugs or similar terminations. Of course, it helps if the wire is a good conductor, but steel coat-hanger wire or welding rod will work satisfactorily. A better choice would be brass rod or brazing rod, between 1/16" to 3/32" thick. If you wish, you can make the loops even larger.

Note the ACTIVE ELEMENT of each antenna in this application is the vertical part of the open loop supported in space by the horizontal part. A greater separation of the vertical elements will produce a larger deviation and more audio. However, the longer a vertical element is, the more signal it will receive, provided there isn't some gross impedance mismatch. This shows up as more carrier, better quieting, or a stronger S-meter reading. Of course, when that vertical section is maintained in space by an unshielded horizontal section, determining the resonance or tuning gets very complicated. Also consider that the horizontal part does receive some signal, and this degrades the intended signal. Therefor, larger loops may work worse. Feel free to experiment; that is the whole object of this project!

Alternatively, the circuit board is designed to accommodate connections by coax cable to other kinds of antenna arrays. Just below the top antenna mounting screw on each side, is a ground screw. This is not used with the open loop antennas, but is used for the coax shield. Thus, if you do wish to make a "Double-Ducky" direction finder as described in the ARRL Antenna Handbook, you can connect the two equal-length coax cables to the HANDI-Finder® circuit board. Similarly, you can fabricate the dual half-wave vertical dipole array described in the Coast Guard Auxiliary Manual and run it with the HANDI-Finder® board.

One suggestion for a more extravagant system is to position two multi-element Yagi antennas with vertical orientation at opposite ends of a horizontal boom. Support the boom at its center on a vertical mast so that it can be rotated. Use equal lengths of coax from each beam and connect them to the HANDI-Finder® board which should be mounted in a protective enclosure at the center of the boom. The beams will give greater forward gain and reduce the 180-degree ambiguity. It is just a little hard to use with the mobile!

Since the antennas have no path to ground, there is no need for DC blocking capacitors. On A890422, the positions of C6 and C7 are jumpered out with board foils. If an application is encountered which requires DC blocking, slit each foil at each end carefully with a sharp hobby knife, then heat the center section with a soldering iron and peel it off the board with the knife. Holes are provided to install the capacitors. Refer to parts list. Later boards don't have this connection, and come with capacitors. Check the board. If you don't have the capacitors, then be sure to insert some jumpers.

NO ENCLOSURE

As you can now appreciate from these discussions, there are many ways you may end up using the circuit board and enclose it accordingly.

In its simplest "quick & dirty" constructed form, the unit can be put to immediate use. The long narrow profile was purposely selected to minimize wind resistance for the benefit of a vehicle operator who may be trying to hold on to the unit mounted on a narrow mast protruding through the drivers window.

You will discover that an enclosure will only increase the drag and may not be worth the effort! With only minimal care, these units have rattled around in many trunks and back seats without anything getting shorted out. The battery in one of the demo lasted over three years, even though it had been left "ON" for several weekends during that time!

However, you might want to consider that numerous sharp edges from component leads could snag cloth upholstery, or scratch leather and vinyl.

Where you really need to, the circuit board can be mounted to surfaces on standoffs using 6-32 screws through existing holes.

MOUNTING IN AN ENCLOSURE:

Model A890422 and later are designed so that it can be more easily modified to fit into a case. This is somewhat irreversible since it requires that the bottom part of the board be cut off, unless you can find a really long, flat plastic case. Some have come and gone off the market, so it is there is no particular model case that can be recommended. It is strongly suggested that the circuit be completely built up on the existing board and thoroughly checked out. After you are familiar with the unit's operation, you may NOT wish to put it in a case.

TRYING IT OUT:

It is best to start off with a "test" situation where you know the location of the source, and experiment with "getting a feel" for the null. The null itself is fairly sharp, but it does not always manifest itself as a total null in the audio tone. Sometimes, you will observe instead, a jump in tone one octave up or down. (Refer to the ARRL Antenna Handbook article for more discussion.)

At other times, you may hear a "buzz" or a fast "twiddle". This is usually due to multipath, so moving just a few feet may help clear up the null. Also, if there is a strong transmitter in the area, such as a 1kw paging system, 3 or 4 MHz away, you may experience more "de-sense" than normal when the HANDI-Finder® is switched on in the DF-ing mode. The sharp edges of the diode switching waveform cause it to be rich in harmonics, all of which modulate that strong RF source and cause it to broaden its normal spectrum at the input of your receiver. (These are the same problems that bother the more complex "DOPSCANs" that switch 4 or 8 antennas.)

Using the unit with a synthesized scanner or one of the new extended-coverage HT's will not only allow you to work in the Ham bands, but other frequencies as well. The prototypes were tried over the range of 49 MHz to 450 MHz and worked well even though the 450 MHz test was well above the suspected ½-wave limit frequency of 321 MHz calculated previously. As you get considerably above or below the 150 MHz design frequency you may find that your unit has a "null" point which is no longer perpendicular to the plane of the circuit board. However this also sometimes occurs due to multi-path in high reflection areas, especially indoors. (Don't try to do anything meaningful indoors!) In any event, don't jump to conclusions. Take several measurements at different locations a few feet apart, and consider your findings carefully before deciding whether the problem is due to the location or a characteristic of the way you have set up your particular unit.

When driving through neighborhoods it is interesting to scan the bands and see what sort of frequencies are in use. For example, one of our experimenters found several homes with 49 MHz "baby monitor" intercoms. You may also hear cordless phones near that frequency.

Another application is to drive around and look for Cable-TV leakage on 145.250 MHz. You will quite easily pinpoint hotspots on poles and at the lawn boxes used for underground installations.

Perhaps you can promote some Ham Radio good will by helping local law enforcement officials find mobiles with "stuck mikes." Such a "free" service can help educate personnel about Amateur Radio and keep you from getting "pulled over" when you are out hunting.

If you are an avid radio-controlled model airplane enthusiast, and have had a plane get away from you and get lost in the woods, this could save you time. Put a low-power transmitter and "crash switch" in your plane. Then if it gets away or crashes in the woods, you can track it down!

Some areas have rather extensive park reserves or other situations where naturalists “tag” wild animals with radio transmitters to track their migration.

They are quite interested in reports on these animals, and will give out the frequencies to individuals with a genuine interest in helping. Invite your local naturalist to speak to your club and demonstrate how you can help. But keep in mind they DO NOT put radio collars on deer and turkey just to make it easy for game hunters to track and kill them!

CHECKING DIRECTION CALIBRATION:

A round handle is best, because it allows a calibration mark at any angle. You would do well to try your unit outdoors in an open field by walking in a circle around a central source. A person keying an HT held up at arm’s reach is easiest, but caution him to hold still. The null should always occur perpendicular to the plane of the antennas, but your radio or other factors may be introducing phase shift. If the error is totally consistent as you walk around the circle, then you will want to mark the true direction on the handle.

If such is the case, expect this calibration to change at different RF frequencies, or if you change the CD4047 oscillator frequency. If the error is not consistent, and changes as you walk around the source, you are experiencing (multi-path) reflections from the surroundings. Try reducing the power of the source. For example, if a 100 mW HT is too strong, remove the antenna and substitute a 51 Ohm, 2 watt carbon resistor. If you still get erratic readings you will have to go to another location.

Once you have established the calibration mark, fasten the cable along the side of the handle or mast so it runs over the mark. That way you can feel it in the dark. Now you are ready to do some serious DF-ing or Fox Hunting.

Notes:

1. The abbreviation “uF” is used throughout this text to denote “micro-Farads”.

2. Bob Leskovec, K8DTS, has been licensed since 1957 and promises not to change his call letters so you can always locate his mailing address via the Ham database!

3. A less detailed description of this project appeared as “Build the HANDI-Finder!” QST Magazine, May, 1993.

4. HANDI-Finder® and RALTEC® are Registered Trademarks.

The HANDI-Finder® Experimenter’s Kit

ASSEMBLING THE KIT: (Please read the whole manual before assembling)

1) Check the circuit board and make sure all the holes are drilled and it is otherwise finished and ready to accept the parts.

2) Having read the discussions about the ON/OFF switch, locate that item and make sure that the type you have does indeed fit the hole pattern on the circuit board. Do not install it yet.

3) Locate the integrated circuit socket, carefully straighten the pins, and insert it onto the component side of the board with the proper orientation. You may have to study the IC socket to determine how its design denotes the position of pin 1. Some sockets have a beveled corner at pin 1. Others have an indentation at the end between pins 1 & 14, some are not clear!

Note, for packing purposes, the IC may have been installed in the socket. It is NOT necessary to remove it, unless you can see it has been inserted in the socket incorrectly. The IC will incur much less handling and there will be less chance of damaging it if you just leave it in place when you solder the socket onto the board. Make sure the socket is properly oriented and seated flat against the board when you do the final soldering.

4) In the following order, mount the diodes, resistors, and capacitors, starting with the smallest parts first. If you have only one capacitor with long leads, SAVE that, and use the precut ones first. There is one location that needs the capacitor to be formed with longer leads.

5) Examine the ON/OFF switch and make sure that the terminals are clean and not tarnished, before soldering. If needed, carefully scrape them with a small hobby blade. Then mount the switch and solder it in position. The 6-pin switch also requires that the two diagonal frame tabs be soldered as well. NOTE: if you later use a solvent to remove flux from the board, be especially careful not to get any into the switch. Some literally melt. Some have four frame mounting pins. Carefully clip off only the two diagonal pins that prevent insertion.

6) Refer back to the discussion about the battery holder and mount that item.

7) Prepare the end of a length of RG-58/U or the miniature RG-174/U and attach it to the board. Note that the hole for the shield should be 1/8” so that the shield can fit through. The excess braid is cut off so that only about 3/16” lies flat against the circuit foil where it is soldered down. Install a nylon cable tie through the holes provided in the circuit board. Pull it tight so that it anchors the coax securely against the board to prevent flexing at the cable end. This is known as a “strain relief.” Clip off the excess. Depending on what type of handle you use, you will need to figure out how to route the coax so that it crosses to the center and comes down the handle equidistant from the two antennas for best symmetry. Later circuit board layouts have the coax routed down the centerline in the solder side of the board on the side opposite where the handles should be attached.

8) Slide on “tinnerman” nuts (if such nuts have been supplied) at the four points used to attach the open loop antennas. These are the two top-most and two bottom-most locations.

9) Fashion two open-loop antennas according to the instructions described previously and attach them to the board using the 6-32 x ½” screw hardware. That’s it! Refer to the section “TRYING IT OUT”.

The manual includes discussions of several other options and variations in the way the unit can be wired.

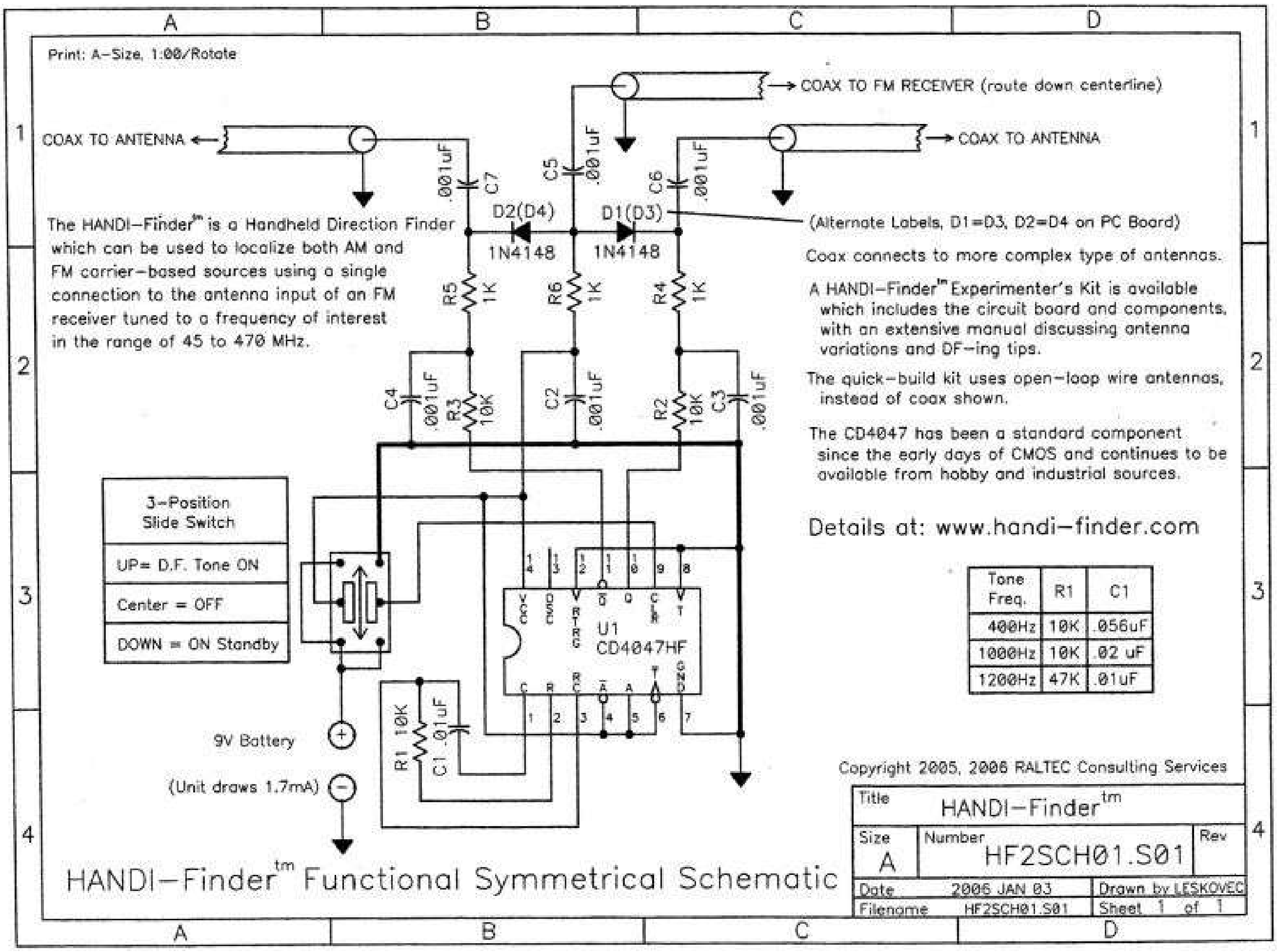
A
B
C
D
1
2
3
4
Print: A-Size, 1:00/Rotate
COAX TO ANTENNA
COAX TO FM RECEIVER (route down centerline)
COAX TO ANTENNA
C7 .001uF
C5 .001uF
C6 .001uF
D2(D4)
D1(D3)
1N4148
1N4148
R5 1K
R6 1K
R4 1K
C4 .001uF
R3 10K
C2 .001uF
R2 10K
C3 .001uF
14 13 12 11 10 9 8
VCC OSC RTRG Q̄ Q CLR +T
U1
CD4047HF
C R RC Ā A -T GND
1 2 3 4 5 6 7
R1 10K
C1 .01uF
9V Battery
(Unit draws 1.7mA)
+
-

The HANDI-Finder™ is a Handheld Direction Finder which can be used to localize both AM and FM carrier-based sources using a single connection to the antenna input of an FM receiver tuned to a frequency of interest in the range of 45 to 470 MHz.

3-Position Slide Switch
UP= D.F. Tone ON
Center = OFF
DOWN = ON Standby

(Alternate Labels, D1=D3, D2=D4 on PC Board)

Coax connects to more complex type of antennas.

A HANDI-Finder™ Experimenter's Kit is available which includes the circuit board and components, with an extensive manual discussing antenna variations and DF-ing tips.

The quick-build kit uses open-loop wire antennas, instead of coax shown.

The CD4047 has been a standard component since the early days of CMOS and continues to be available from hobby and industrial sources.

Details at: www.handi-finder.com

Tone Freq.	R1	C1
400Hz	10K	.056uF
1000Hz	10K	.02 uF
1200Hz	47K	.01uF

Copyright 2005, 2006 RALTEC Consulting Services

HANDI-Finder™ Functional Symmetrical Schematic

Title HANDI-Finder™
Size A
Number HF2SCH01.S01
Rev
Date 2006 JAN 03
Drawn by LESKOVEC
Filename HF2SCH01.S01
Sheet 1 of 1

Sense The Right Way To Go With The HANDI-Finder®

by Joe Moell KØOV
ARRL Technical Advisor

The following was submitted to the Technical Correspondence column in QST Magazine for October 1993 in response to the HANDI-Finder construction project in a previous QST issue. The original HANDI-Finder was created by Bob Leskovec K8DTS. It is documented at the HANDI-Finder Web site. A kit of parts is sold by the North Olmsted Amateur Radio Depot. "Clones" of the bow-tie bidirectional HANDI-Finder are being or have been offered for sale by some other companies and individuals. The KØOV unidirectional pattern modification and antenna improvements described below are applicable to all of these bow-tie sets.

Hats off to QST for its recent promotion of foxhunting and radio direction finding (RDF) activities. Simple RDF gear in the hands of a large number of hams will provide important preparedness for emergency searches and malicious interference location.

The HANDI-Finder® by K8DTS[1] holds promise for this purpose, but its bidirectional peaks and nulls limit it to short range "sniffing" in situations where time-to-locate is of little concern. Its 180-degree ambiguity makes it unsuitable for long-range or high-speed work.

If you were shopping for a compass to use on a hike, what would you think of one with an indicator that points either north or south, with no way of determining which? You probably wouldn't buy it, even if it was priced much lower than other compasses in the store. It's the same way with foxhunt equipment.

A bidirectional RDF set such as the basic HANDI-Finder presents a 50-50 chance of leading you away from the fox instead of toward it. So it's worthwhile only for occasions when boundaries are small, such as on-foot foxhunts in a small park. Fortunately, it's a simple matter to incorporate a second mode into your HANDI-Finder to resolve the ambiguity.

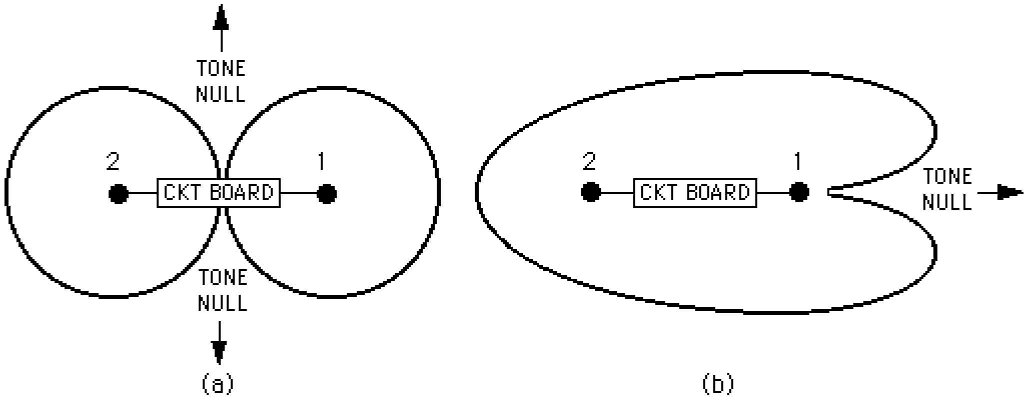

In this view looking down from the sky, the bidirectional HANDI-Finder unit (a) has two nulls, each perpendicular to the plane of the antennas. The added sense mode (b) has a single null on the Antenna #1 side.

Add a Sense Mode

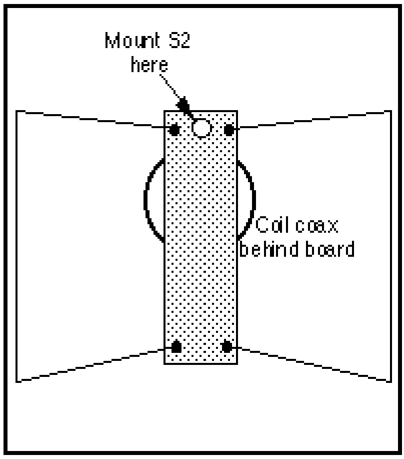

By adding time delay to the signal from one of the two HANDI-Finder antennas, the peak/null tone pattern of the unit is modified. If you add just the right amount of delay, the pattern changes from bidirectional (figure-8) with nulls perpendicular to the antenna plane to cardioid (heart-shaped) with a single null off one end.

This two-pattern technique forms the basis of the Handy Tracker, a RDF unit that I designed in 1989.[2] The cardioid pattern concept came from an unpublished 1980 project by Russ Andrews K6BMG. The Handy Tracker sense mode scheme can be applied to the K8DTS HANDI-Finder with the simple addition of a switch and a length of coaxial cable to provide delay.

As K8DTS points out, the bidirectional HANDI-Finder indication is independent of frequency. This is not true of the added sense mode. A delay line of a given length produces a precise cardioid pattern at only one frequency. A useful pattern is obtained over a frequency range of several per cent, however, so one delay line will provide ambiguity resolution over one full VHF amateur band.

Modification Details

The modification adds a DPDT micro-mini toggle switch without center-off (Radio Shack 275-626) and a precisely trimmed coax delay line. Mount the switch in the etch-free area of the circuit board between the two antennas.

On the rear side, cut the traces where C6 and C7 connect to the antenna terminals per the drawing. On the Antenna #2 side, wire the antenna connection through the switch per the schematic. On the Antenna #1 side, install an unshielded wire jumper 1-3/4 inches long, to compensate for the added switch wiring on the Antenna #2 side. This jumper is important, since one inch of difference in feed length to one antenna will cause 6.5 degrees bearing error in the figure-8 mode.

Cut the delay line according to the following formula:

$$l = (11808-df)v/f$$

where l = length of coax in inches, v = velocity factor of coax as a fraction, f = frequency in MHz, and d = spacing between the vertical antennas in inches. For example, assume we are adding the sense mode to a HANDI-Finder with "bow tie" open loop antennas for two meters. The vertical sections of the two loops are 15 inches apart. We are using RG-174 coax, which has velocity factor of 65.9 per cent. Computed for the center of the band (146 MHz), delay line length is

$$(11808-(15)(146))(0.659)/146 = 43.4 \text{ inches.}$$

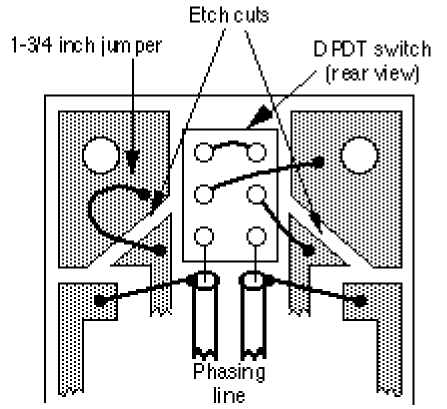

Rear view of modified board showing trace cuts and added switch wiring. Note loop of wire on left; this compensates for switch-lead lengths that would otherwise unbalance the system.

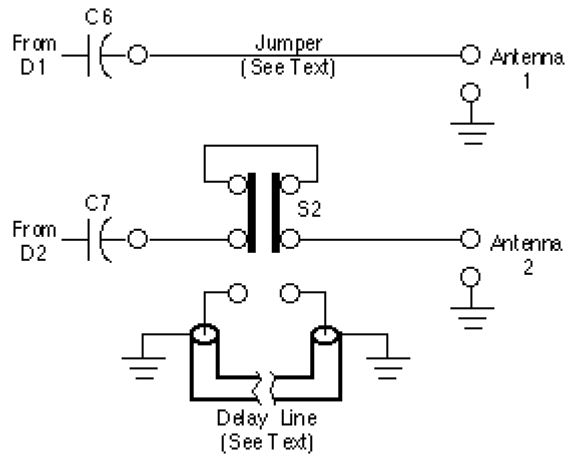

Partial HANDI-Finder schematic showing added jumper, toggle switch S2 and delay line to provide cardioid sense mode.

Coil up the coax and lace it to the board. Miniature coax such as RG-174 is best for the delay line because it makes a compact coil. Whatever coax you use, make sure to determine its characteristic velocity factor. It is different for foam and non-foam type cables.

Because of environmental multipath and stray coupling, a perfect cardioid tone pattern may not be achieved. The end null may not be complete or there may be a double null on one end. This doesn't matter, because the purpose of the sense mode is merely to determine which of the two figure-8 nulls to follow. There is plenty of peak/null amplitude difference for that.

Getting RDF bearings with the modified HANDI-Finder is a simple two-step process. In a clear location, with S2 set for normal (bidirectional) mode, listen for the tone nulls in the receiver to obtain a line of bearing as described by K8DTS. Next, switch S2 to the cardioid pattern mode and turn the unit 90 degrees so that one end points along your line of bearing. Note the tone level. Rotate the unit 180 degrees and note the tone level again. It should be distinctly different. The lower amplitude tone will occur when Antenna #1 is closest to the signal source. To help you remember, mark the peak/null directions on the board.

Other Observations

In the K8DTS QST article photo, the receiver feedline passes over the open end of one antenna loop. This unbalances the antenna pair and can cause major bearing errors on two meters and above. The coax should be routed away from the loop.

The horizontal sections of the bow tie antennas severely degrade RDF performance. Properly spaced coax-fed vertical dipoles are far better for serious use with this type of RDF set. Such an antenna set is easy to make with inexpensive whip antennas and PVC plumbing parts.

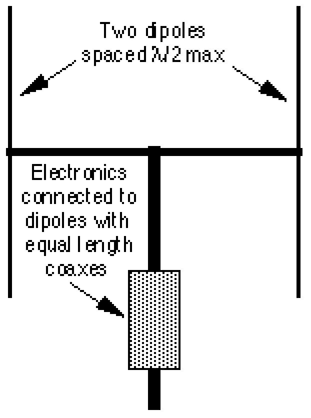
Two dipoles spaced λ/2 max
Electronics connected to dipoles with equal length coaxes

Dual vertical antennas fed with equal-length coax lines provide the best performance with switched-antenna RDF sets.

The two antennas should be spaced as widely as possible for maximum audible df tone, but spacing must not exceed one half free-space wavelength at the receiving frequency. Approximately 24 inch spacing is practical for both two meters and 125 cm. Coax cables from the HANDI-Finder board to each of the two vertical dipoles must be exactly equal in length.

Wide antenna spacing gives better performance, but it may increase the received tone deviation to the point that the receiver "squelches out." If this occurs, reduce the tone oscillator frequency to 400 Hz or lower by increasing the value of R1 and/or C1.

Footnotes:

[1] Leskovec, "Build the HANDI-Finder!" QST, May 1993, pp 35-38.

[2] Moell, "Homing In," 73 Amateur Radio, September 1989, pp 58-59 and November 1989, pp 52-53.

The spelling and form of HANDI-Finder constitutes a trademark registered to K8DTS.

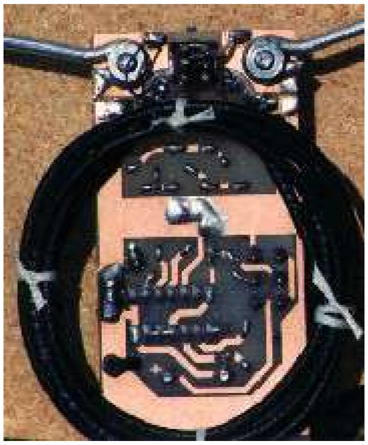

The added coax delay line is tied to the rear of the circuit board.

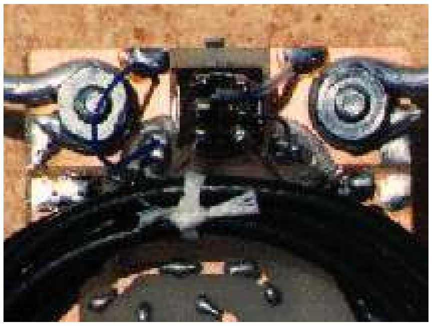

Close-up of top area of board, showing added DPST switch and wire jumper to provide equal signal paths in bidirectional mode.

Frequently Asked Questions and Answers About T-hunting and RDF

by Joe Moell KØOV

Over 98% of reader response to my articles is via e-mail. The 80-20 rule (or is it 90-10?) definitely applies, because the vast majority of questions are the same. Here they are, along with the answers. Now maybe I can spend more time building and T-hunting!

- Part 1 - Questions about Dopplers (scroll down)
- Part 2 - Questions about Other RDF Equipment
- Part 3 - Questions About Transmitters and and Covert Tracking
- Part 4 - Miscellaneous Questions

Part 1 - Questions About Dopplers

Q: What is a Doppler? What is a Roanoke Doppler?

A: Nineteenth Century physicist Christian Doppler is credited with creating equations that describe the apparent frequency shifts that result from the relative motion of the sender and receiver of a wave. The Doppler principle is now used for many applications, including dating the universe and catching highway speeders. In radio direction finding, a simulated rotating antenna can be used to determine the direction of an incoming signal using this principle.

When transmitter hunters speak of Dopplers, they usually mean a device with an electronically rotating array of 3 to 8 vertically oriented whips or dipoles and a circular display of 8 to 50 light-emitting diodes. These Doppler add-ons work with VHF and UHF narrowband FM receivers. There are other RDF devices with similar antenna sets and displays that are used with AM receivers (e.g Watson-Watt), but these are not Doppler sets.

The Roanoke Doppler was first published in my book "TRANSMITTER HUNTING---Radio Direction Finding Simplified" (THRDFS), along with an extensive writeup of the theory of operation. The Roanoke Doppler has become the most popular Doppler RDF home construction project. Originated by Chuck Tavaris N4FQ, then improved and documented by Joe Moell KØOV and Thomas Curlee WB6UZZ, the design was named by KØOV to honor the transmitter hunters of Roanoke, Virginia, where N4FQ tested his prototype. This Homing In site is the official support site of the Roanoke Doppler project:

- Three suppliers of circuit boards and parts for the Roanoke Doppler are listed on the Hardware Sources page.
- Updates to the THRDFS documentation of the Roanoke Doppler are in the THRDFS Update page.
- An improved multiband antenna switcher is fully described in the Wide-Range Antenna Array page.

There are other home-construction Doppler projects, but only the ones built from the design in THRDFS and this Web site are Roanoke Dopplers.

Q: I can't locate any 75492 ICs for the Roanoke Doppler display. What shall I do?

A: An improved 16-LED display circuit is in the THRDFS Update page at this site. The 75492 ICs have been eliminated. If you already have a circuit board for this project, you can still use it.

Q: The WA2EBY Doppler circuit in a recent ham magazine looks very much like the Roanoke Doppler. Is it an improvement?

A: That author has utilized many of the design attributes of the Roanoke Doppler, but his unit has no new functions or operating features.

Q: The WA2EBY article states that ECG-555 PIN diodes give poor UHF performance and that ordinary 1N4148 diodes are just as good for a Doppler antenna. Is that true?

A: First, The ECG-555 is an older PIN diode, with unspecified RF series resistance. Newer PIN diode types listed on the Hardware Sources page of this site should perform better. Second, the switcher circuit of the WA2EBY Doppler is not optimized for PIN diodes. For best isolation, RF diodes should be reverse-biased when OFF, not just at zero DC current. That's because the series capacitance of PIN diodes drops dramatically as OFF (reverse) bias goes from 0 to 4 volts. (Examples: HP 5082-3080 goes from 1.1 to 0.4 pF. MPN3404 goes from 2.5 to 1.4 pF.) The lower the capacitance, the less RF gets through (better isolation). The Wide-Range Roanoke Antenna at this Homing In site reverse-biases the OFF diodes, which is an important reason why it performs better than the original Roanoke antenna. The WA2EBY Doppler circuit does not reverse-bias the OFF diodes. I believe that the results of the WA2EBY's loss/isolation tests would have been quite different if it did.

I discuss PIN versus switching/rectifier diodes for Dopplers on page 128 of THRDFS. PIN diodes provide significantly less insertion loss when ON, if biased properly, meaning that a PIN switcher will generally have less loss. More importantly, stray RF rectification is much less likely with PINs. I encountered significant cross-modulation problems using a Doppler with non-PINs, especially when driving near strong RF sources. WA2EBY did not test for cross-modulation when he evaluated diodes. I would welcome input from anyone who has done such tests.

Q: Will the Roanoke Doppler find aircraft Emergency Locator Transmitters (ELTs)?

A: The Roanoke and other ring-antenna Doppler RDF sets can readily track signals in the 120 and 240 MHz aircraft bands, provided that a suitable receiver is used. The Doppler modulation produced by switched whips in a Doppler array is FM, not AM. Therefore, the receiver must have an FM detector to properly process it. I use a venerable Regency MX-7000 scanner for this purpose. Just tune it to the proper frequency (121.5 or 243.0 MHz for ELTs, 121.6 MHz for tests) and set the receiver to the Narrowband FM mode.

The new Wide-Range Antenna Array should be used with the Roanoke Doppler processor for greatest sensitivity and best rejection of multipath distortion. Whips for 121.5 MHz should be 22 inches long, spaced in a square pattern 21 inches on each side. Whip lengths and spacings should be halved for 243.0 MHz operation. The Doppler is intended for vehicle or fixed station use only, not for on-foot operation.

Some high-end scanning receivers (including Icom R1, R10, R100, R7100-2, AOR AR-3000A, AR-8000, Radio Shack Pro-60 and others) can be operated in the Narrowband FM mode on 121.5 MHz. Specification changes may occur, so check to be sure before purchasing. Another idea for experimenters is to install crystals for 121.5 and 121.6 into an old crystal-controlled FM scanner. Everybody wants synthesized scanners nowadays, so "rockbound" sets are cheap and plentiful at swap meets and yard sales. Retune the RF stages as necessary for best sensitivity, because these sets were usually designed for receiving 146 MHz and up.

While a Doppler may be useful in some ELT tracking situations (such as a false alarms from aircraft on an airport runway), it is not a universal tool. ELT signals from actual crashes are usually very weak and of mixed wave polarization. Experienced Civil Air Patrol (CAP) volunteers in my area prefer cubical quads and phased arrays (such as the L-Per) instead of a Doppler when in such weak-signal situations. The photo shows a dual frequency (121.5 and 243.0 MHz) stiff-wire quad built by Bob Miller N6ZHZ, Commander of Brackett Composite Squadron 64 in La Verne, California. He says, "With the quad, I am able to pick up a 243 MHz ELT two to five miles miles before the L-Per's antenna will hear it. When there are two or three ELTs chirping simultaneously in hangers at an airport, I use a high-gain quad to track one at a time, listening to the differences in tone sweep rate of each one."

An RF attenuator should be used with an RDF quad to allow S-meter bearing readout and to help determine how close you are to the ELT. Plans for suitable quads are in THRDFS. For more on CAP and tales of searches for ELTs, see Homing In for April 1994.

Q: Can I use the Roanoke (or other) Doppler on the 10-meter ham band and/or the 11-meter Citizens Band?

A: It could be done if:

- The receiver is in FM detection mode (not SSB or AM)
- Switcher inductors and capacitors are optimized for this frequency range
- Antennas are 8-foot whips in a square pattern 8 feet by 8 feet

Of course this is not practical on a car, but could be considered for a fixed installation. Remember, Dopplers need carrier-type signals, so you can track amplitude-modulated signals (in FM mode) but not SSB signals. I have experimented with the Roanoke Doppler on 6 meter (50 MHz) FM signals, using resonant whips spaced in a four-foot square. They fit on the roof of the van, but there wasn't enough ground plane area around the whips, and the whips were so tall that they waved around too much while driving. Compared to baseline two-meter performance, this didn't work very well, especially in motion.

Q: I'm thinking of mounting a Doppler RDF antenna on top of my home ham station tower. Any suggestions?

A: One problem with remote operation of the array will be rf losses in the coax. You will probably not have problems with long control lines, although some added protection from lightning-induced transients is in order. For tower mounting, I recommend using four vertical dipoles (which would look like the marine DF antenna photo in THRDFS) instead of whips on a ground plane. This will give better performance by minimizing mutual coupling between the whips.

Q: OK, how about putting the Doppler at our mountaintop repeater site to find jammers on the input?

A: There are many pitfalls in the way of remote mountaintop RDF systems. I get letters regularly from hams who tell me about the wonderful linked RDF systems at repeater sites that they'll be putting on line "real soon now." I encourage them to send me all the details of their results, with photos, so I can tell the world via Homing In. Then I never hear from them again.

You'll probably discover, as they do, that you won't be able to successfully operate a Doppler DF at your repeater site. Here's why:

1. The commutating antenna system, "moving" at an audio rate, will induce the annoying Doppler tone into your repeater receiver and probably into most other receivers at the site, even those on widely different frequencies. This will occur even if you use a separate receiver for the Doppler antenna. Note that this phenomenon is not caused by digital switching noise in the commutating antenna. It is a parasitic RF effect and can occur even if no digital switching noise exists.
2. The cross-modulation products caused by stong signals from your repeater transmitter and other VHF/UHF transmitters at the site into the switched antenna system will make the Doppler insensitive and prone to unreliable indications.

Fast-switched dual-antenna RDF sets such as TDOA's will probably have similar problems. You might consider a slowly rotating beam with some sort of direction indication through the repeater---this was done by the Happy Flyers organization many years ago.

I know of only one successful remote repeater-connected RDF system now in operation, and it is not a Doppler. (If you know of others, please enlighten me.) This system uses a phased array which is switched very slowly (every second or so) to avoid the problems mentioned above. That slow switching is OK in the user's application, because it is intended to track continuously-transmitting ELTs, not ham-band kerchunkers.

So if you want to use Dopplers for remote RDF, put them at sites that are not associated with repeaters or other VHF/UHF radio systems.

Q: I have read several articles on 2-meter Dopplers and they all specify different antenna element spacings. What is the best?

A: No matter how many whips or vertical dipole elements in a Doppler array, the adjacent elements must not be more than 1/2 free-space wavelength apart, to avoid ambiguous bearings due to phase steps of greater than 180 degrees. Furthermore, adjacent element spacings of greater than 1/4 wavelength will produce phase steps of more than 90 degrees, lowering the level of the recovered audio tone and worsening the signal-to-noise ratio. With that in mind, an optimum adjacent-whip spacing for a 4-whip mobile array for VHF or UHF is slightly less than 1/4 wavelength at the highest frequency to be used. Plans for the 4-whip Roanoke Doppler in THRDFS call for 18-inch whip-to-whip spacing, which is 0.22 wavelength.

The array doesn't have to be this big, but don't make it too small, either. Deviation of the recovered Doppler tone is a function of antenna size and speed of rotation, as given by the formula on page 121 of THRDFS. This works out to about 0.5 KHz deviation for the Roanoke Doppler 2-meter antenna, which is a good value. A super-small array such as the 0.07 wavelength-on-a-side mentioned in the Dick Smith literature won't work well, because the recovered Doppler tone out of the receiver would be only 0.1 KHz. It could be readily overpowered by modulation audio and noise.

If you use a mag-mount implementation, remember to place the whips in a perfect square pattern. Measure spacing carefully or make a template. A slightly rectangular or trapezoidal pattern will introduce significant bearing errors at some angles.

Q: What about the spacing if I use more than four whips?

A: Going to 6, 8, 12 or more whips, as provided for in some Doppler sets, provides little improvement if the array size (radius of rotation) is the same as for 4 whips. The Doppler tone deviation is unchanged, though there are fewer high-order harmonics in the recovered audio. (That's unimportant if a very narrow passband filter is used, as in the Roanoke Doppler). Furthermore, the closer element spacing of a same-size-more-whips array increases inter-whip RF coupling, which degrades performance in multipath.

To fully realize the advantages of more whips, the array size must be increased. Consider a two-meter 8-whip array in an octagon pattern of 18 inches per side (i.e. the same adjacent whip spacing as the Roanoke Doppler). Radius of rotation is more than double that of the 4-whip Roanoke, giving over 1 KHz Doppler tone deviation at the same rotation rate. The array aperture is more than doubled, which improves performance in multipath.

However, an array with a diameter of almost 4-1/2 feet won't fit on most vehicles, especially when you remember that you would need 7-1/2 feet total diameter to provide ground plane under the whips. A similar size (in wavelengths) array for the 70 cm band (450 MHz) would be only 2-1/2 feet diameter, which is practical.

Q: Why doesn't the Roanoke Doppler use RF preamps and/or "soft switching" in the antenna system to give more sensitivity and less noise?

A: Questions like this indicate three assumptions:

- Assumption #1: Soft switching eliminates noise
- Assumption #2: Monolithic amps provide increased system sensitivity
- Assumption #3: A Doppler with switched preamps provides the same directional performance as one using PIN diodes.

I disagree with all three assumptions. Here's why:

1. Any Doppler antenna commutation, whether "hard" or "soft," produces additional spectral components in the RF signal to the receiver. It is true that rectangular-wave switching can create some broadband noise, but if the noise is really bad in your system, it's probably because the diodes are biased or driven improperly. Wide-bandwidth receivers such as scanners and handi-talkies are much more susceptible to this problem than mobile radios with good input RF filtering (such as helical resonators). You may find that changing your receiver or adding a helical filter between antenna and receiver improves the performance of your Doppler markedly.

2. Monolithic preamps have gain, which can overcome the small amount of loss in the coax from antenna to receiver. But they usually have a poor noise figure, much worse than most receivers. RF gain ahead of the receiver will worsen any receiver overload and cross-modulation problems. Strong signals can also overload the preamps, distorting the incoming signal phase and degrading the bearing accuracy. (Long-time T-hunter Mike Musick NØQBF has observed this even on signals from one or two miles away!) So unless you are using a very insensitive receiver, antenna preamps will probably do more harm than good. In addition, some Doppler sets hard-switch the preamp supply voltages. This causes the same broadband noise problems in a wideband receiver that improperly biasing the PIN diodes does.
3. As I explained in detail in my Homing In of April 1995 and August 2003, the impedance match (or mismatch) presented to switched-off whips in a Doppler array determines the multipath performance of the system. Preamps usually "look like" 50 ohms, which is undesirable in this application.

In summary, I believe that a properly designed PIN-diode switcher (with correct ON and OFF biasing), when used with a sensitive narrow-range receiver, gives as good or better overall Doppler performance (in both sensitivity and multipath accuracy) than any preamp-based design does, whether hard or soft-switched.

Q: Our club wants to set up some RDF sites to automatically find jammers and display their locations on the Internet. How do we do it?

A: First, understand that you won't be able to use such a system to positively identify RF coming from a specific house or apartment at a distance. Let's say you set up a RDF system with three sites to cover a typical urban area of 20 X 20 miles. For good coverage of the area, the sites will have to be on the perimeter. Assume that you can find enough quiet sites and link them. Now assume that the jammer is right in the middle of the triangle of sites (best case). The intersection of your bearings and their uncertainties will create a polygon with an area of about 4 square miles! That certainly won't be good enough evidence for the FCC. There are about 10 hams per square mile in my city, probably the same as where you live. So you'll still have to do plenty of close-in DFing to verify the actual perpetrator's site and get eyewitness evidence.

Why the large polygon? Because the inherent uncertainty of a bearing from a typical VHF 4-whip Doppler in an urban environment is about +/- 5 degrees, even if you have 10-bit readout precision. Sometimes, the bearing error is even greater, especially at fixed sites. (See Chapter 18 of THRDFS.)

This is not intended to discourage you, just to add a healthy dose of reality. If you want to link up RDF stations, consider using APRS, which allows packet interconnection of a limitless number, both fixed and mobile, close-in or far away, each one displaying their own and everyone else's bearings. Bearings and station locations can be input to APRS either manually or automatically, so even base stations with beams can help out. For more info, see my Homing In columns in 73 Magazine for October 1994, January 1995, and February 1995.

Q: How can I interface my Doppler to APRS?

A: Any of the popular homebrew or kit-built Doppler RDF sets with 8, 16, or 32 LED indicators can be connected to computers using a circuit developed by Robert Swain N7LUE. This includes the popular Roanoke Doppler and the DoppleScAnt from QST in the 70's. Robert's circuit takes the binary information from the LED drivers and outputs a serial RS232 data stream of ASCII characters representing directions. For details, see Doppler-to-APRS interface at this Web site and Homing In for August 1995. A microcontroller-based serial interface that outputs in "Agrelo" format is described on the Web pages of Bob Simmons WB6EYV. He also sells the PicoDopp, a Doppler add-on that includes serial interface, in semi-kit form.

Some sets made by Doppler Systems Incorporated feature RS232 serial output consisting of 3-digit numerical readout of bearing in degrees. Doppler Systems sets in the 3000, 4000, and 5000 series can also be used with N7LUE's interface.

The MicroFinder Doppler set by AHHA! Solutions has serial output of bearing information in "Agrelo" format.

Q: Why are there only 16 LEDs on the Roanoke Doppler?

A: The Roanoke Doppler was designed for the best combination of simplicity and performance. A 16 LED display is usually quite adequate for tracking signals when mobile in urban areas. It gives a rapid indication of which way to drive when time is of essence. I still use my 16-LED unit on the local hunts with good results. When I need higher accuracy for triangulation, I switch to a beam.

Finer resolution is possible with some circuit changes. Tom Curlee WB6UZZ has designed a version with digital readout to the nearest degree. But the basic bearing accuracy of VHF Dopplers on vehicles is typically +/- 5 degrees, so such a display gives a false sense of accuracy. And all these improvements require more complexity. That means more parts to find and greater cost. We have found that most hams want maximum simplicity in projects they have to build themselves. (On the other hand, when they buy a ready-to-use piece of gear, they want maximum bells and whistles.)

Part 2 - Questions About Other RDF Equipment

Q: Can you help me get my (choose one) Agrelo, Ramsey, Dick Smith, or "bow-tie" RDF equipment (not in THRDFS) to work?

A: Mail about these units is so frequent that I have posted details at this site.

- Agrelo DFjr Doppler
- Ramsey SlyFox and Foxhound
- Dick Smith Doppler
- "Bow-tie" sets by K8DTS and others

Q: I can't get my Shrunken Quad to work. What's wrong?

A: Most likely you are using the wrong piston trimmers, but it could also be an improperly built balun. Go to More on the Shrunken Quad to find out.

Q: Are there any updates/corrections to ______ (fill in your favorite project in TRANSMITTER HUNTING---Radio Direction Finding Simplified)?

A: Click here to get the latest updates to THRDFS.

Q: Where can I find boards/kits/parts for your projects?

A: I try to locate inexpensive sources for uncommon parts and include them in all my project articles. For updates and additional parts/boards info, jump to the Hardware Resources page at this site.

Part 3 - Questions About Transmitters and Covert Tracking

Q: My club wants to start transmitter hunting but first we need a hidden transmitter. What is available?

A: No problem! The transmitting part of T-hunting is easy if you give it a little thought. Keep in mind that no transmitting setup is perfect for all hiding situations. Hiders use a wide variety of equipment, with power output and antenna type depending on the distance to the start point and the level of intended difficulty. You need to determine, for instance, whether you want to do short range on-foot hunting, longer range hunting on bicycles, or still longer range mobile hunts.

For your first mobile hunts, just have someone stay with the transmitter and and key it up at appropriate intervals. The hider can read into the mike from a book, or from the club newsletter, or just make comments and urge the hunters on, as NØQEC and KBØIEK of the Lincoln (Nebraska) Amateur Radio Club are doing in the photo at left.

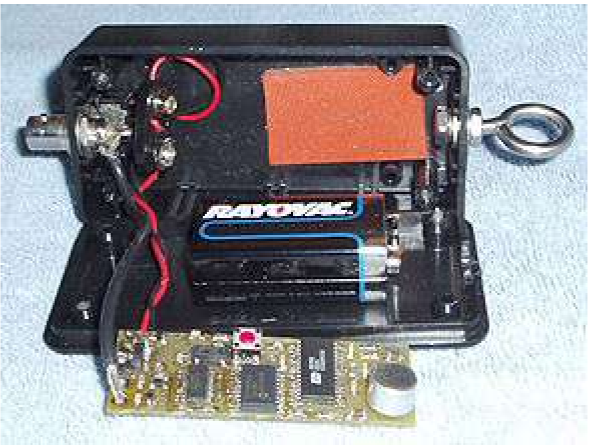

After a few hunts, you may want to put together an unattended transmitting setup, so the hider doesn't have to stay with the rig. For starters, connect your two-meter hand-held or mobile transceiver to a tape recorder playing an endless loop answering machine cassette with an appropriate message and ID. Some hiders like to use a dual-band hand-held and activate it on the subband from another transmitter. Later on, after you gain some experience and have a better idea what suits your needs, consider a dedicated home-built or commercial "foxbox" with tones and a cycling timer for hiding. See the Hardware Resources page at this site for transmitter controller ideas.

For short-range on-foot hunts in parks, consider complete mini-transmitters like the SquawkBox by Bob Simmons WB6EYV (photo at right). It records and sends short voice messages at adjustable periodic intervals.

More about transmitting gear for on-foot hunts can be found in the Equipment Ideas for Radio-Orienteering page of this site. And of course there is a complete chapter of clever transmitter hiding ideas in THRDFS.

Q: I need a miniature transmitter to put on my prized possession (sports car, TV set, motorcycle, coonhound) to find it after it wanders away or is stolen/abducted. What is available?

A: Yes, I have actually had inquiries about tracking devices for each of the above examples, and many more! Unfortunately, it is still true that you get what you pay for, and no matter what you pay, there's no assurance of success.

In the under-$40 category are "experimenter" micro-transmitters such as the TT7-K "Tracking Transmitter Homing Beacon" sold by Amazing Devices. Several miniature transmitter models are available from ElectroKits. These "bugs" typically use the 88-108 MHz FM broadcast band, where low-power unlicensed transmitters are permitted by FCC Part 15. Manufacturers sometimes claim range of three miles or more for these FM micro-transmitters. But that range assumes high-gain antennas on both transmitter and receiver. A practical antenna on a stealth radio tag will seldom be large and efficient enough for that. Furthermore, if one were to actually achieve three-mile range with such a transmitter, that fact would demonstrate that the field strength is in excess of that allowed by FCC Part 15 rules for unlicensed devices in the USA. In other words, it would be illegal operation of a radio transmitter in the eyes of the law.

Some of these experimenter FM-band transmitters use simple L-C tuning, so the transmitter tends shift frequency due to temperature changes, battery voltage droop or proximity to other objects. Futhermore, the FM broadcast band is not suitable for covert operations, as the tag can be tuned in on every home and car radio within range. Usable tracking range is severely limited by interference from powerful broadcast stations. The high level of audio FM deviation (+/-75 KHz) used by transmitters and receivers in the FM broadcast band is not compatible with narrowband Doppler RDF techniques. Have I convinced you that this is not a good option for most applications?

Also available for about $40 are micro VHF transmitters for the two-meter ham radio band, such as the MR6 Model Rocket Tracking Transmitter Kit by Ramsey Electronics. A similar postage-stamp sized transmitter project was detailed in Homing In for May and September 1993 (photo at left). At about 10 milliwatts, such transmitters are perfect for short-range ham radio transmitter hunts (see previous question). But two-meter frequencies may be used only by licensed ham radio operators, with proper transmitter identification, in accordance with FCC regulations. Plenty of ham radio T-hunters are ready and eager to track down intruders on these frequencies, and they won't be amused if they discover that a non-ham is using their spectrum space to keep track of his cat. There are restrictions and license requirements for the use of other VHF and UHF frequencies at these power levels as well.

A much better choice would be the tiny transmitters and tracking receivers made specifically for tracking mammals, reptiles, birds, fish and even insects. See my RDF links page for commercial suppliers of transmitters and receivers that are popular with two market groups: wildlife researchers and the owners of prize hounds. A good example is Communications Specialists of Orange, California, which sells waterproof tracking transmitters that are about the size of a stack of six quarters (photo at right). They may operate unlicensed under FCC Part 15 rules and are powered by a lithium-ion "coin-cell" battery that lasts about a month. Com-Spec is expanding the market for these radio tags by selling them as the LoCATor Pet Tracking System, the Radio Controlled Airplane ELT System, and the Law Enforcement Tracking System. Range of Part 15 compliant transmitters is a few blocks in the city and up to a mile in open terrain with transmitter and receiver elevated. For official law enforcement use, more powerful transmitters certified to FCC Part 95 are available for greater range.

To save battery life, these critter-tracking transmitters are pulsed for a few tens of milliseconds, every second or so. Their low-power signals are so weak at long range that dual-antenna sets and Dopplers are at a disadvantage. Sensitive narrowband receivers and beam or phased-array antennas work best. Professional grade tracking receivers are quite expensive, but the 30-channel Com-Spec Model PR-30 215 MHz range set with Moxon type directional antenna (photo at right) is a bargain at about $200. Another option is a consumer-type scanner. It must be multi-mode model such as the Icom IC-R10, Trident TR2400 or Sony ICF-PRO80, which include CW/SSB (BFO) modes. For a tracking antenna, you could build a small cubical quad from plans in THRDFS.

If money is no object, consider the complete RDF tracking systems manufactured by law-enforcement suppliers such as Tracking Products, Inc. Also at the high end of the cost range are "hostage tracking" systems sold by "spy market" distributors such as Surveillance Equipment Group and Airtight Investigations. Note that even with their high prices, there are practical limits on range and battery life. All advertising claims must be taken with more than a grain of salt. None of these providers offer any guarantees that you will locate whatever you're tracking or even that you will hear a trackable signal.

For vehicles, don't overlook LoJack and other commercial tracking systems. LoJack transmitters are installed in an undisclosed place in your vehicle or other important asset and activated automatically when it's reported stolen to the police. Then the signal is received and RDF-tracked by police and sheriff cruisers and aircraft (over 400 in the Los Angeles area alone). Ask the LoJack salesman if the guarantee of your car's return is still offered.

There are also non-RDF vehicle tracking systems such as the new GPS-based CyntrX by Enhanced Vehicle Applications of Wisconsin. It is sold by TechWholesale in Ohio. (Reports from users wanted, please let me know your experiences by e-mail.)

Q: What about a pager-sized tracking device to put on toddlers? Parents would carry a pocket-sized receiver so that they could locate the child in case he/she became lost.

A: As you can imagine, the tracking of children is a much more serious and difficult matter than the tracking of cats and sports cars. Here are just some of the many factors that must be considered if direction-finding technology is to be used:

- Size - For sufficient transmitter power and antenna size for tracking over a wide area, the child's device must be bigger and heavier than most people desire.
- Tracker - An effective tracker needs a "wide aperture" antenna for sensitivity and accuracy, so it cannot be pocket-sized. But parents won't want to haul around a big tracker on every trip to the amusement park.
- Battery Life - The longer the battery must last, the bigger and heavier the transmitter on the child must be.
- Antenna - How do you put an effective antenna next to a child's body without detuning it?
- RF Radiation - Is it safe to have a transmitting antenna next to a child's body?
- Liability - Will the maker be sued if parents cannot quickly find their child with the device?
- Security - How do you keep criminals from tracking lost children, too? What happens if a potential kidnapper finds the lost child before the parents do?

Companies such as Xmark Systems make "perimeter protection" systems that sound an alarm if a tag-wearing individual wanders or is taken out of a protected area. These systems are marketed to health care facilities as alarms to provide early notification of wandering Alzheimers or mental patients, or of infant abductions. They are based on RFID technologies (see next question), so they don't have any direction-finding capabilities. The protected perimeter is non-portable, and the tags can't be used to help find the person once he or she is outside the protected area. Care Trak, Incorporated sells a system of wrist/ankle transmitters that combines perimeter alarm and short-range (up to one mile) RDF tracking. In some localities, RDF-proficient volunteers assist authorities in tracking wandering patients under Project Lifesaver.

Some startup "personal tracking" companies, such as Protect Me Toys (now defunct?), eWorldTrack, Wherify Wireless, and Digital Angel (also defunct?) have received considerable press (and resulting investor attention) for their claims of long-range positioning and tracking. None of their systems use traditional direction finding (i.e. beacon transmitter and human-carried RDF gear). Instead, they are based on wearable Global Positioning Receivers with radio transmitters that beacon the GPS data to trackers directly or via cell/PCS phone systems. Currently available models are relatively large, have short battery life, and require a clear view of the sky for accurate positioning. A savvy child abductor would quickly remove such a device from the child or shield it to render it inoperative.

Other RDF-based systems now on the market for tracking children and adults include the medium- and high-end solutions described in my answer to the previous question, with the usual power, range, and battery limitation. If you have personal experience with any of these products, good or bad, or if you discover any other suitable (or unsuitable) systems, please let me know the details by e-mail.

Q: Our construction company is losing small tools, either by accident (falling into holes) or theft. Can a transmitter be inserted into items like wrenches, drills or grinders, with a RDF unit capable of detecting them inside a car trunk?

A: Radio Direction Finding (RDF) technology is typically done at considerable distance (from yards to miles and beyond). To do that, transmitters require long-term power sources (such as batteries) and antennas. They're usually too large to be placed on or in a hand tool. In addition, the transmitted signal can be detected by anyone with a receiver tuned to the proper frequency, so these systems are not covert.

RFID technology at the jobsite exit is probably more appropriate for this tool-detection application. RFID systems track property and objects at relatively close range with unique codes. For instance, chips (also called tags or transponders) can be implanted in pets to provide positive proof of ownership if the pet strays into the pound or is stolen. A reader device, passed over the chip, detects it and reads out the chip's unique ID code. Similar RFID systems sound an alarm when non-paid-for merchandise passes through the doors of a store. RFID done at greater distance in automatic toll collection systems on and elsewhere. At one time, the FasTrack transponders for southern California toll roads could even be used to quickly pay for a McDonald's burger at the offramp. (This feature was discontinued because so many transponders were being stolen out of cars and then used for this purpose!)

In an ideal situation, the chips are "passive," meaning that they don't require battery power. They usually cannot be detected with conventional receivers, but only by a reader or polling device designed to be used with them.

Part 4 - Miscellaneous Questions

Q: I'm ready to go on a mobile T-hunt. How do I find one?

A: If you live in southern California, see the Southern California T-Hunts for Beginners page at this site for monthly mobile T-hunts in the Los Angeles and Orange County area where first-timers are especially welcomed and encouraged. Web sites about other Los Angeles area Saturday night hunts and the southern California All-Day hunts are on the Links page of this site. That page also has links to Web sites and Homing In Correspondents who can tell you about hunts all over the USA, and in some other countries, too. If there are no hunts in your area now, get together with your local ham club members and start them!

Q: My friends and I chat on the local two-meter repeater. Somebody on there transmits beeps and mocks us. We can all hear him on the input of the repeater. Please listen with your direction finding equipment. We need to stop him.

(Believe it or not, requests for me to listen on VHF/UHF have come from hams that are hundreds of miles away from me!) You and your friends need to put some RDF gear on your vehicles and track down the offender yourselves. That's why I have this site and my book --- to help hams like you become proficient at RDF. It's not hard. Once you find him or her, a little friendly persuasion may be all that it takes to eliminate the problem. If not, the chapter on solving malicious interference problems in THRDFS may be helpful in developing a strategy.

You should also find out if the Amateur Auxiliary and Official Observer programs of the American Radio Relay League are active in your area. If so, join them now. Your ARRL Section Manager has more information.

Q: Is it true that transmitters can be "fingerprinted" for positive identification without the user being aware?

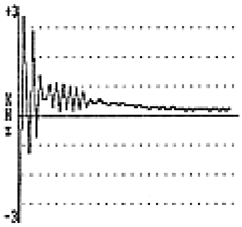

A: The concept of transmitter fingerprinting was invented and patented decades ago by Phil Farrell K7PF of Seattle. He licensed his patent/technique to The Boeing Company and to Motron Electronics, which sells equipment for it. The Motron implementation (called TxID) consists of a card that plugs into the expansion slot of a PC, plus software running under DOS. The asking price is higher than most individuals are willing to pay, but it is affordable for repeater clubs. I tested and reviewed the system in Homing In for November 1994. The review explains the principles of fingerprinting and shows good and bad points, including comparison of some same-model transmitters. Since then, Motron has added features such as receiver frequency control and readout for remote operation. TxID will control a tape recorder for evidentiary purposes.

A few hams have developed similar fingerprinting systems. The first was "XMIT_ID" by Richard Rager KB8RLN. It used PC software and an 8-bit Soundblaster card. I have lost contact with Richard and don't know the current status of this project.

"Sherlock" by Malcolm Mallette WA9BVS also runs on a PC and uses its sound card with commercial oscilloscope software. The newest version of Sherlock, which works with Windows XP and Vista, was detailed in a feature article by WA9BVS in the Winter 2006 issue of CQ VHF magazine. Click to download this article in PDF format (7 MB), courtesy of WA9BVS.

XMIT_ID and Sherlock are intended for home experimentation only. Commercial production of any fingerprinting system based on the turn-on frequency changes of a transmitter may infringe on US patent 5,005,210.

Q: What is a Crenshaw Factor and why do I need one?

A: The Crenshaw Factor has been used for many years in southern California T-hunts to normalize mileages and account for odometer differences between various vehicles. Your vehicle's Crenshaw Factor is the odometer mileage you get when you drive a standardized course that runs along Crenshaw Boulevard for about 9 miles. The course was chosen because it is relatively straight and goes up to the start point of the All-day hunts. This makes it easy for hunters to check their factors regularly on the way to a hunt.

A big problem right now is that the landmarks at the start of the Crenshaw have changed. The official start point is in front of a now-gone telephone booth in front of a now-gone Winchell's Doughnut shop on the southwest corner of Artesia Boulevard and Crenshaw. If you can find this point, then all you do is record your odo reading and drive in the left-hand lane along Crenshaw to the stop sign at Crest Road in Rancho Palos Verdes. Your elapsed mileage to that point is your Crenshaw Factor.

Nobody knows exactly what this distance is, but it doesn't matter. To normalize mileages on a T-hunt, each vehicle's elapsed odometer mileage is divided by its Crenshaw factor. The result is in "Crenshaw Units." The team with minimum Crenshaw Units wins.

Example: Say your odometer mileage on the hunt is 35.7 and your Crenshaw factor is 8.8. Your Crenshaw Units are 4.057. You think you won, but another hunter has odo mileage of 36.8 and Crenshaw factor of 9.3. That means his odo reads higher than yours for the same distance traveled. His Crenshaw Units are 3.957, so he beat you.

It is commonplace for two or more hunters to have low mileages within a few tenths of one another. Crenshaw Factors are needed to determine winners on many of our milege hunts, so it is important for every hunter to have one.

Q: Why has your book's price increased to $24.95?

A: I don't know. It was done by the new publisher in recent printing runs, without notice to the authors. There are no changes in these print runs, so look for previous-printing books at lower prices, if you can find them.

Q: I have a ham store and would like to carry your book. Do I have to get it from ARRL?

A: Retailers and dealers should get books directly from the publisher. Call McGraw-Hill Trade Order Desk at (800) 722-4726 to place quantity (5 or more) orders.

Q: Is there any difference between ISBN number 007-1560068 and ISBN number 0-8306-2701-4 for your book?

A: The book has two ISBN numbers because one was assigned to it by TAB Books originally, and the second was assigned by McGraw-Hill when McGraw-Hill acquired TAB Books. Booksellers might show stock in either ISBN number, depending on when they ordered the book and put it into their computer system. Sorry for the confusion, but I have had no control over this process.

Q: Why don't you sell copies of old Homing In columns or post them here on the Web?

A: For logistic and contractual reasons, I don't mail out or sell copies of magazine articles I have written. Furthermore, they cannot be posted here or on any other Web site, with only a few exceptions. Your fastest source for back issues of ham magazines is probably a visit to your local library. A nearby ham radio store may also have recent issues in stock. Many ham clubs maintain ham magazine libraries for their members.

Q: Can I make and sell kits and boards for projects in your books and articles?

A: Since Tom Curlee and I don't have the time to get into the board/kit business, we welcome other enterprising hams supplying these products. However, our publishers insist that we protect the copyrighted material appearing in the book and my magazine columns. Kit/board suppliers must not duplicate, sell, or give away any of the TAB/McGraw-Hill or 73 documentation of the projects. Kit/board buyers should be urged to purchase THRDFS for the tune-up, calibration, installation, and operating details. This works out well, because some of our kit suppliers are making extra money by retailing the book! To purchase books wholesale (5 or more copies), call the "trade order" desk at TAB/Mc-Graw-Hill. Please note that Tom and I support the circuits in our book, but we accept no responsibility for performance of third-party boards and products.

Hardware Resources for RDF Home Builders

Sources for the Roanoke Doppler project

Complete construction plans for the Roanoke Doppler are in Transmitter Hunting---Radio Direction Finding Simplified (THRDFS). The following are sources of printed circuit boards for the display unit. Some may also offer critical parts, complete kits, and/or wired/tested units.

- Directional Systems: Jim Sorenson KA4IIA, PO Box 81881, Conyers, GA 30094, ka4iia@radio.org
- FAR Circuits: Fred Reimers KF9GX, 18N640 Field Court, Dundee, IL 60118, farcir@ais.net or http://www.farcircuits.net (847) 836-9148. Users have reported errors on older versions of the display boards. They have been corrected in display boards now being shipped. FAR Circuits also has a circuit board for the Wide-Range Antenna Array. There is a minor correction to the marking on this board. Click here for more information.
- Stephen Douglas K6HZT of Sacramento, CA, k6hzt@arrl.net or http://www.marshwildman.com/doppler.htm. Please note the board errata information on his site.

Disclaimer: The information above, and elsewhere on this page, is provided for the convenience of readers only. These suppliers are independent entities. They have no business association with Joe Moell or Tab/McGraw-Hill. A listing on this page does not constitute endorsement of products. Contact these suppliers directly for offers and prices.

Boards are no longer available from Ed Greany KB6DOL or Tom Lewis AB5CK.

Sources for RF attenuator projects

Resistive RF attenuators are best suited for mobile HF and VHF T-hunting with directional antennas such as loops, yagis and quads. Pete Ostapchuk N9SFX sells parts for resistive attenuators, including a 6-section enclosure machined from 6061-T6 aluminum bar stock and engraved with attenuation values. Also available from him are switches, resistors and connectors. Plans for his project are in QST Magazine for May 1998. Send e-mail to n9sfx@aol.com for details and prices. The above disclaimer applies.

Another source of resistive RF attenuators is National RF, Incorporated (formerly Radio Engineers) of San Diego, CA.

For on-foot VHF/UHF "sniffing" and international-style foxhunting with a beam and handi-talkie or scanner, an "offset" or "active" attenuator is better because it eliminates the problem of signal leakage directly into the receiver's case. Plans for an active attenuator are at this Homing In site, click here.

Sources for fox-boxes

Hidden transmitter controllers, sometimes called "fox-boxes" are units that provide audio and timed key-up/key-down (push-to-talk) outputs. A fox-box turns an ordinary handi-talkie or mobile transceiver into a cycling hidden T. Here are four fox-box projects for home construction:

1. The Un-Music Box by KØOV is an easy-to-build project using common easy-to-obtain CMOS ICs. No microprocessor programming is required. Provides on-off cycling or continuous transmissions. Callsign is hand-programmed using two switches; shift register memory has battery backup. Complete plans in the book TRANSMITTER HUNTING---Radio Direction Finding Simplified. You can remote control the Un-Music Box with a surplus pager receiver using plans in Homing In for December 1990.
2. PicCon by Byon Garrabrant N6BG uses a Microchip PIC microcontroller and has a variety of programmed on/off modes and tone sequences for mobile T-hunts as well as IARU-standard MOE/MOI/MOS audio, timing and synchronization for international-style radio-orienteering. DTMF remote control and programming with your HT or external receiver. Delayed turn-on and turn-off timers. Callsign and programming info are held in EEPROM when power is removed. Reviewed in Homing In for March 1997. Complete kit is available from N6BG.
3. TBOX by Ron Seese N6MBR uses an 80C51 microprocessor to provide three programmed tone sequences and/or a CW message. User-programmable on and off times in intervals from seconds to hours. Programmed via RS-232 serial port using commands similar to packet TNCs. Non-volatile memory (EEPROM) holds programming and message when powered down. Optional SmartWatch clock turns fox on automatically at programmed time. Plans for basic unit are in Homing In for October 1991. Assembly, object and binary files for the microcontroller program (10K total) are available at this Homing In site. Contact N6MBR regarding availability of kits and critical parts.
4. Francois Tremblay VE2JX and Jacques Brodeur VE2EMM of Montreal have designed a new PIC-based foxbox for international-style hunts with MOE/MOI/MOS audio, synchronization of 2 to 7 foxes, delayed start of up to 3-1/2 hours in 1/2-hour increments, and automatic shutoff after four hours of operation. Callsign is held in non-volatile flash memory. Callsign changes and mode setup are easily done with dipswitches. This project is fully described in Homing In for April 1998.

Sources of PIN type RF diodes

For maximum receiving sensitivity and minimum problems with cross-modulation, good RF PIN diodes must be used in hard-switched dopplers such as the Roanoke Doppler. Breakdown voltage and power rating of the PIN diodes are not important in low-level receiving circuits. (You aren't going to transmit through the switcher, are you?) For low loss and good sensitivity, the diodes should have low RF resistance (Rs) at the expected forward DC bias current (7.5 mA for the wideband switcher). They should have low capacitance (Ct) at the reverse bias off-state voltage. Effective carrier lifetime and reverse recovery time (Trr) should be short. (PIN diodes with long carrier lifetime are ideal for current-controlled RF attenuators, but that's not what we want in this application.) Keep the diode leads very short to minimize inductance and make sure signal path lengths, including diodes, are equal for all elements in the array.

Suitable PIN diodes include:

- Narda KS3542 -- DO-34 package, click for data sheet. A ham with a supply of these diodes has offered to make them available in small quantities to home builders of RDF equipment. Send e-mail to me for more information.

  Philips NTE555 or ECG555 -- "Replacement" type semiconductor available at local electronics parts houses all over the country. Even some giant computer stores such as Fry's Electronics carry NTE parts. Priced $1.50 to $2.00 each. This is an older part with tab leads and unspecified series RF resistance.
- Agilent Technologies (formerly Hewlett-Packard) surface-mount PIN diodes -- Surface-mount parts are best if you're building a doppler for UHF (400 MHz and up). Three types used to be easy to get from RadioShack.com but are apparently no longer available there. Try Arrow Electronics, a major parts distributor in in Melville, New York. The Agilent HSMP-3892 or HSMP-3893 are good choices for a series doppler switcher such as the wideband array on this Web site. On the other hand, the HSMP-3820 is optimized for a shunt switcher such as the one in THRDFS.
- Motorola MPN3404 -- In TO92-style plastic package. 50 cents each or ten for $4.50 at Circuit Specialists, 220 S. Country Club Dr. #2, Mesa, AZ 85210 (800) 528-1417. $1.00 each from Debco Electronics, 4025 Edwards Road, Cincinnati, OH 45209 (513) 531-4499. Advertised price is 60 cents each at Dan's Small Parts and Kits, Box 3634, Missoula, MT 59806-3634 (406) 258-2782.
- TRW A5S303 -- Axial lead device formerly available from Tanner Electronics in Carrollton, TX (214) 242-8702. Does anyone have a new source?
- Agilent Technologies 5082-3077 -- Axial-lead low-Rs device available from Newark Electronics. Newark also carries Agilent 5082-3081 and 5082-3080, which are long-carrier-lifetime parts intended for variable RF attenuators and thus not suitable for doppler switches.
- Motorola MMBV3401LT1 -- Surface-mount replacement for the now-discontinued MPN3401. Available from Motorola industrial distributors. Not suitable for perf-board construction.

After I bought some Motorola PIN diodes from Debco Electronics, I mentioned their availability in Homing In for April 1995. A reader wrote to me after publication to say that Debco was astonished at the response. The order-taker said that they hadn't sold a PIN diode for over a year and then after the article came out they began selling like hotcakes. Who says hams don't build nowadays?

Other Sources of RF Parts

Hard-to-find RF parts such as air variable capacitors for tuning HF loop antennas can sometimes be found at dealers of "overstocked" or "industrial surplus" components. Here are three to try:

- Bogdan Electronic Research & Development (Jim & Carolyn), PO Box 62, Lakewood, CA 90712, (562) 634-0911. They sell RF parts by mail and are often at southern California swap meets, too. Note: Jim says that the margin on small quantities of RF parts is so low that it's not cost-effective for him to return calls. (I know the problem!) Therefore, please use e-mail or call back to catch him or Carolyn "live" on the phone if you get a recording when you call.
- All Electronics, PO Box 567, Van Nuys, CA 91408.
- Alltronics, 2300 Zanker Road, San Jose, CA 95131, (408) 943-9773.

For more information about sealed RF "piston" trimmer capacitors and sources, see the Shrunken Quad page at this site.

Ferrite beads and cores for the dual-polarization quad (see Homing In for February 1997) and other RF projects are available by mail order from Amidon Associates. Locals can save shipping charges by picking up orders at 240 Briggs Avenue, Costa Mesa, California, during normal business hours. Call in advance to 800-898-1883 or 714-850-4660.

ByteMark Corporation, 7714 Trent Street Orlando, FL 32807 (407) 679-3184 also carries Amidon ferrite components and has technical information about Amidon ferrite and iron powder products on its Web site. In addition, ByteMark sells a Windows-based program for calculating the turns for coils on various Amidon cores.

The Philmore line of packaged electronic parts includes unusual items such as Amidon T50-6 cores (used in the ATX80 80-meter fox transmitter), International Rectifier HEXFETs, MPF-102 JFETs, etc. Unfortunately the Philmore Web site does not have a complete product list, but it does include a list of local distributors of these parts.

Crystals for 80-meter foxboxes are available from Brian Carling AF4K in Sanford, FL. He also sells some crystal-can oscillators that may be suitable for building offset attenuators.

Jerry Boyd WB8WFK provides this feedback on ON7YD's Fox Transmitter for 80 meters (ATX80), which was featured in 73 Magazine for November 2000: "Motorola MC14001B Quad-NOR ICs from Radio Shack (part number 276-2401) didn't work. The oscillator would not start properly. I tried 6 different ones and they all acted the same. I put in an old Fairchild CD4001B Quad-NOR and now it works with no problem. I downloaded data sheets from the manufacturers' Web sites and the internal schematics are indeed different." Fairchild ICs are available from many suppliers, including Digi-Key. KØOV's six transmitters use Signetics (Philips) CD4001B's packaged by Jim-Pak and sold by local parts suppliers nationwide -- they all work fine.

Sources for building polar-plot bearing readouts

I have many requests for sources of unusual electromechanical parts used in two-dimensional bearing indicators such as the KK6CU motorized quad, the WB8WFK and KD6IFZ computer displays, and the NØMKJ Foxcopter. (See the Homing In Index page at this site for articles describing these RDF systems.)

For (relatively) inexpensive surplus sine-cosine potentiometers as used in the KK6CU motorized quad, try Servo Systems Company, 115 Main Road, PO Box 97, Montville, NJ 07045, (201) 335-1007, in USA (800) 922-1103. The last catalog I received from this company included sin-cos pots by Litton, Duncan, N.E.I., Samarius, and Gamewell. There were continuous turning and single-turn models with prices from US$18.50 to $69.50.

For a "Cadillac" system, check out these contactless sin-cos resolvers using internal magnets and Hall effect sensors, designed to be rotated at up to 10,000 RPM: Elweco Incorporated, PO Box 909, Painesville, OH 44077-0909, (440) 254-1716.

Small industrial rotary joints (slip rings) in the $30 price class are suitable for low-voltage uses. NØMKJ used one of these for the two-meter coax feed in his Foxcopter. They are available from McMaster-Carr Supply Company, PO Box 54960, Los Angeles, CA 90054-0960, (310) 692-5911.

Another promising source of rotary joints is Mercotac, Incorporated, 6195 Corte del Cedro #100, Carlsbad, CA 92009, (619) 431-7723. Unfortunately the Web site does not list prices. If you use one of these products in your RDF system, please let me know how it works and what it costs.

Inexpensive electronic compasses with digital output (Vector 2X series) are manufactured by Precision Navigation Incorporated, 1235 Pear Avenue, Suite 111, Mountain View, CA 94043 (415) 962-8777. They are sold by Jameco Electronic Components, 1355 Shoreway Road, Belmont, CA 94002, (415) 592-8097, in USA (800) 831-4242. They are also sold by JDR Microdevices, 1850 South 10th Street, San Jose, CA 95112, (408) 494-1400, in USA (800) 538-5000.

From Peter PB2RDF comes this suggestion for a source of obsolete and out-of-stock electronic components for replacement or for older designs: 1-Source Electronic Components, 591-B Bicycle Path, Port Jefferson Station, NY 11776 (800) 966-8826 or (631) 642-2479.

RF Sniffer with Tone Output

Here is a neat Sniffer designed by Don, KF6GQ. It produces an audible tone which rises in frequency as the signal gets stronger. It uses 5 transistors and a 555 timer IC.

For a 1 Watt transmitter, sniffing range is about 225 feet using a 3 element Quad beam, or about 100 feet with a 1/4 wave spike antenna.

The Zero Adjust pot is used to set the tone frequency when there is no RF. With knob, field adjustable.

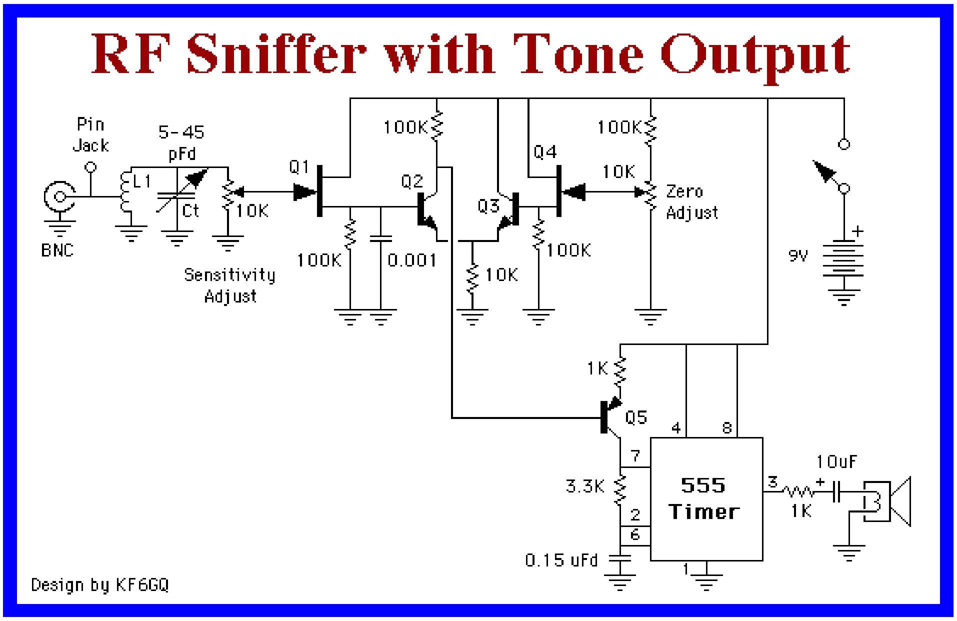
RF Sniffer with Tone Output
Pin Jack
5-45 pFd
L1
Ct
10K
BNC
Sensitivity Adjust
Q1
100K
0.001
Q2
100K
Q3
10K
Q4
100K
10K
100K
Zero Adjust
9V
+
1K
Q5
4
8
7
3.3K
2
6
555 Timer
3
10uF
+
1K
0.15 uFd
1
Design by KF6GQ

The Sensitivity Adjustment pot is used to control how much signal is required to produce the maximum tone frequency. With knob, field adjustable.

Parts Not Given on Schematic

L1 = 5 turns, 1/4 inch form tapped at 1.5 turns. (For 2 meters)
Ct = Small RF trimmer capacitor, adjustable over about 5 to 45 pFd. Not critical. Screwdriver adjustment is fine.
Q1 & Q4 = 2N5457 (silicon J-FET)
Q2 & Q3 = 2N3904 (silicon NPN)
Q5 = 2N3906 (silicon PNP)
NOTE: Just about any small-signal silicon transistor of the same general type as specified should work.

Contact Information

George (Russ) Andrews, K6BMG
BMG Engineering, Inc.
9935 Garibaldi Avenue
Temple City, CA 91780 USA
Phone: 1(626)285-6963
FAX: 1(626)285-1684 (dedicated line, automatic)
Web: http://members.aol.com/bmgenginc

Audible S-Meter

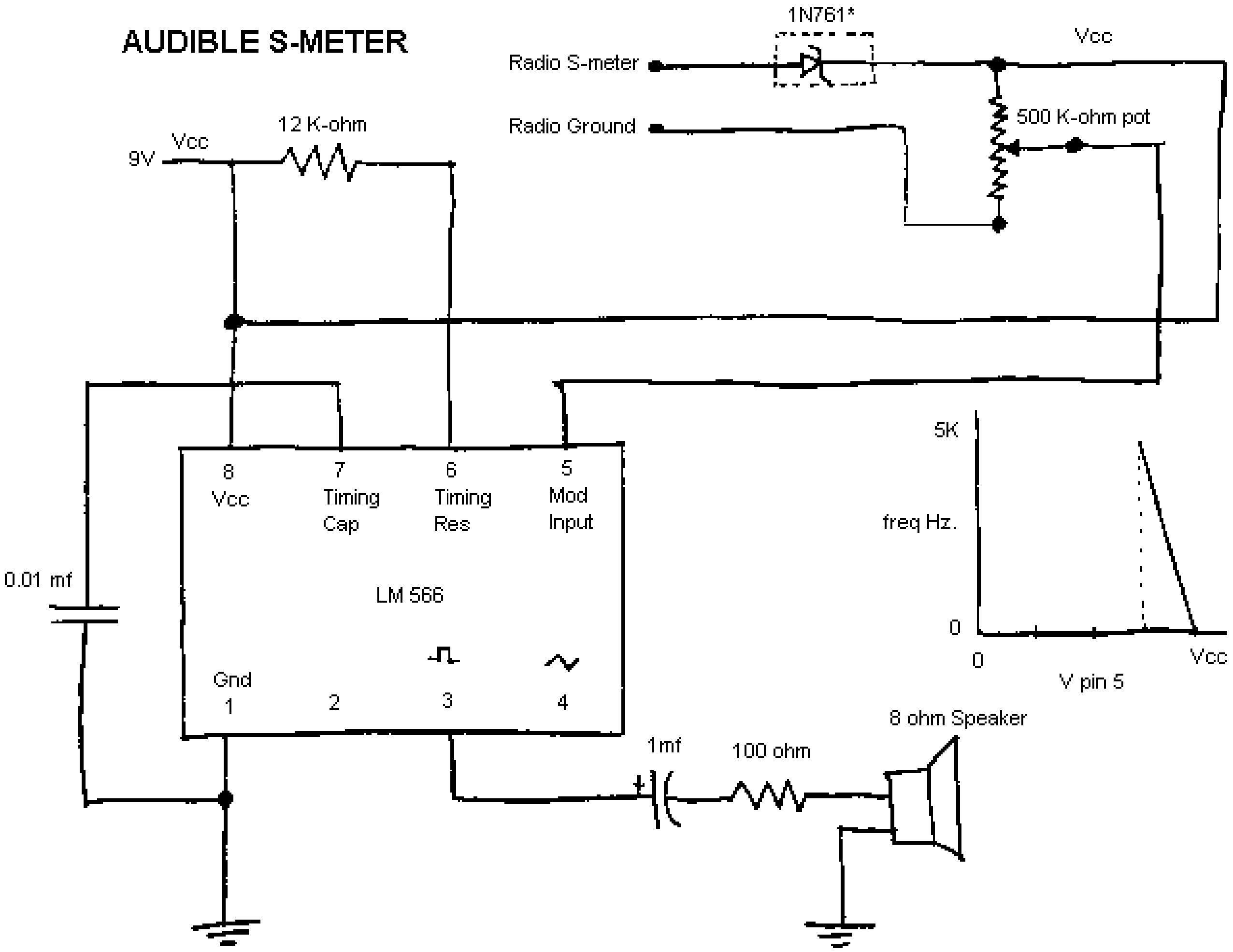
AUDIBLE S-METER
1N761*
Radio S-meter
Vcc
Radio Ground
500 K-ohm pot
12 K-ohm
9V
Vcc
8
Vcc
7
Timing
Cap
6
Timing
Res
5
Mod
Input
LM 566
Gnd
1
2
3
4
0.01 mf
5K
freq Hz.
0
0
Vcc
V pin 5
8 ohm Speaker
1mf
100 ohm

The Montreal Audio FS-Meter

PARTS LIST, 27 JULY 99

R1 = 2M7 1/4 W (2.7 mohms)
R2 = 47R 1/4 W (47 ohms)
R3 = 100K 1/4 W
R4 = 10K 1/4 W
R5 = 10K 1/4 W
R6 = 1K0 1/4 W or 1K0 trimm pot DK# D4AA12-ND for volume adjustment if needed (1 kohm)

RV1 = 1M0 to 5M0 POTENTIOMETER (GAIN) (1 to 5 mohms)
RV2 = 10K 1T TRIMM POT, DK# D4AA14-ND

C1 = 2-20 pF RED TRIMMING CAP DK# SG20015-ND
C2 = .001 uF CER
C3, C4 = .1uF CER
C5 = 1 uF ELECTROLITIC 16V
C6 = 100 uF ELECTROLITIC 16V, Mount laying down on the board
C7 = .01 uF CER

D1 = 1N34, 1N60 ETC GERMANIUM DIODE
IC1 = MAXIM MAX495CPA RAIL TO RAIL SINGLE SUPPLY OP-AMP, DK# MAX495CPA-ND
IC2 = MICROCHIP PIC12C671/JW, UPGRADABLE, DK# PIC12C671/JW-ND
OR FOR LOWER COST, PIC12C671-04/P, NOT UPGRADABLE, DK# PIC12C671-04/P

L1 = 2T, 1/4" ID ON COLD END OF L1 #20
L2 = 5T, 1/4" ID AIR WOUND #20

SW1 = SPST MINIATURE SWITCH

MISC:

3 = AAA BATTERY HOLDER
1 = 18 pins IC socket or 2 = 8 pins IC sockets
1/8" EARPHONE JACK MONO
SINGLE EARPHONE
BNC CHASSIS JACK
PUNCH BOARD OR PCB
METAL BOX for shielding

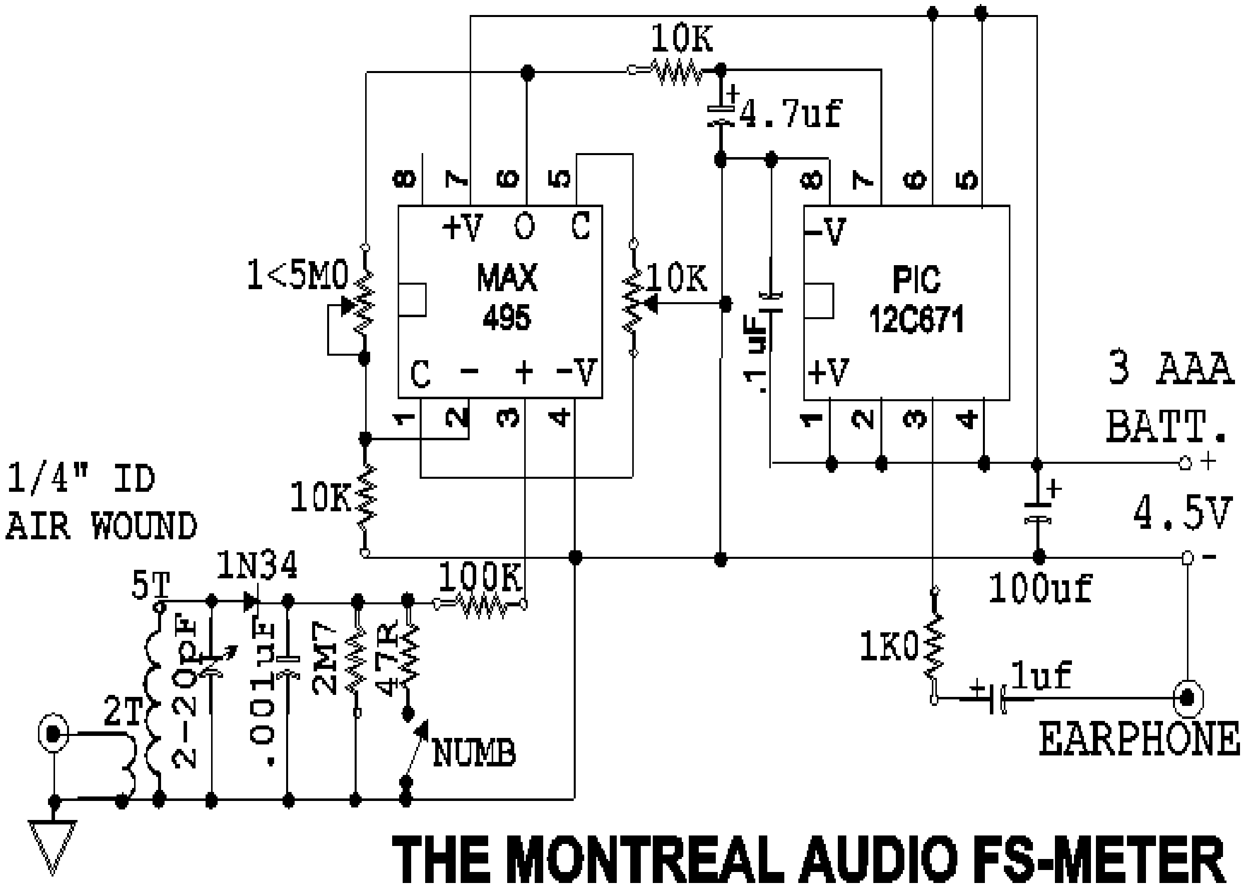
10K
4.7uf
8 7 6 5
+V O C
MAX
495
1<5M0
10K
C - + -V
1 2 3 4
8 7 6 5
-V
PIC
12C671
+V
.1 uF
1 2 3 4
3 AAA
BATT.
+
4.5V
-
100uf
1K0
1uf
EARPHONE
1/4" ID
AIR WOUND
10K
5T
1N34
100K
2T
2-20pF
.001uF
2M7
47R
NUMB

THE MONTREAL AUDIO FS-METER
JACQUES BRODEUR VE2EMM
V 1.0 JUNE 99

A Simple Crystal Radio Set

Whether you are a beginner or a seasoned expert, you would probably agree that there is nothing quite like using this simple design to make a crystal radio and enjoying it, listening to stations both near and far. Add to this the "something for nothing" advantage of not having to power this set with batteries or AC line and you can see why this simple design will be used and enjoyed for a long time to come. Almost like magic... Let' build one!

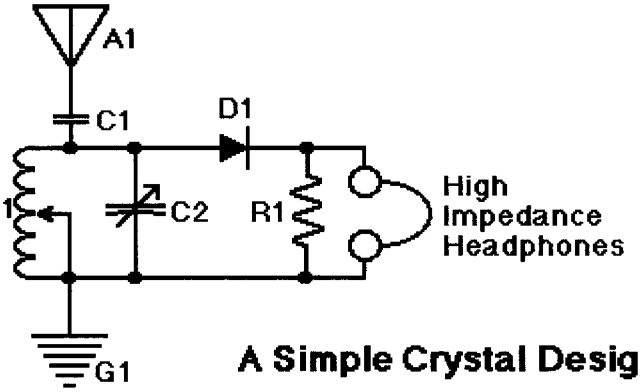
A1
C1
D1
L1
C2
R1
High Impedance Headphones
G1

A Simple Crystal Design

Components:

- **A1- 50-100 foot longwire antenna**
- **C1 - 50 pf ceramic capacitor OSE p/n CD-5**
- **C2 - 365 pf variable capacitor OSE p/n BC14400**
- **D1 - 1N34A diode OSE p/n 1N34A**
- **G1 - Ground (see instructions below)**
- **L1 - 100 turns of #20 AWG enamel coated copper wire OSE p/n MW20, adjustable tap (see instrctions below)**
- **R1 - a 100k ohm resistor OSE p/n RQ100K**
- **Crystal Earphones OSE p/n CE748**
- **Fahnestock clips OSE p/n 2090**

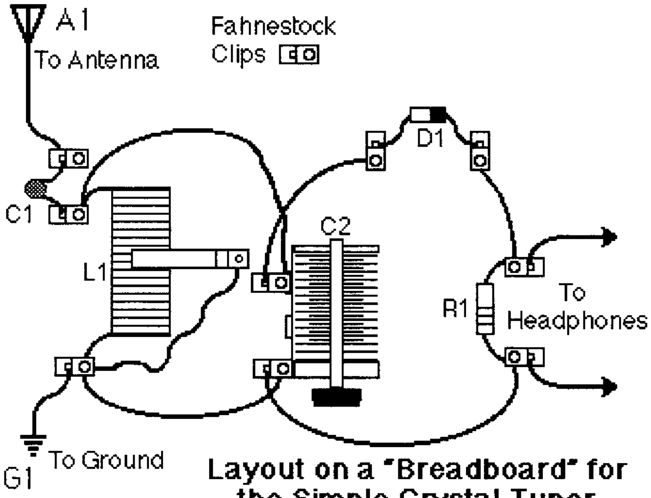
A1
To Antenna
Fahnestock Clips
D1
C1
C2
L1
R1
To Headphones
G1
To Ground

Layout on a "Breadboard" for the Simple Crystal Tuner

CHAPTER THREE

Crystal Sets

Crystal sets are simple to build and cheap! With six feet of aerial this set will pick up the BBC programmes at good strength; with fifty feet of aerial the volume is really loud. Build the set with care, remembering that reception depends entirely upon aerial and earth.

These crystal set designs are chosen as the first in our constructional series because of their simplicity and cheapness. The most popular receiver for years was the 'crystal set', the theoretical design of which was precisely the same, in effect, as that shown in Fig. 7, our first design. Those old crystal sets, however, used large coils of thick wire and detectors consisting of a piece of crystalline substance 'tickled' by a small cat's whisker, which was only a piece of thin wire wound up into a small spring. Today we use a miniature coil and a germanium diode detector.

Germanium is one of the modern metals which have been developed for industrial purposes in recent years. A small flake of it is sealed into a tiny glass tube, with its cat's-whisker contact cemented permanently into place (Fig. 6).

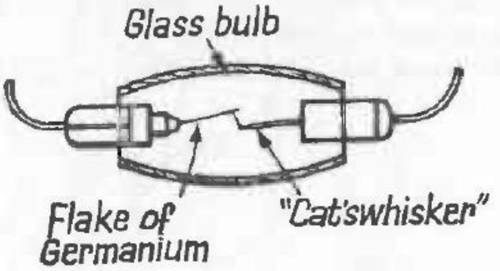

Figure 6 **The Germanium Diode Detector**

A crystal set has no power to amplify the signals which it receives via the aerial and the crystal detector merely converts the radio signals into currents capable of actuating the diaphragms of our headphones and producing sounds thereby.

It follows, therefore, that the strength of the sound we receive in the headphones depends on the size of the aerial we have – a larger aerial picking up 'bigger' signals – and the efficiency of the detector in converting them ('rectifying' them is the proper term) into 'sound' signals.

The germanium diode is much better as a detector in every way than the old crystal detector and is completely stable in use. The modern crystal set, therefore, must not be confused with that of fifty years ago; it is best to discard the term 'crystal set' really in favour of 'germanium diode receiver'. The receiver in the Fig. 7 circuit is small and compact and very simple to make. With a good aerial it will give an extremely loud headphone signal from two or, possibly, three stations. It is essential to have a good aerial with a crystal set, and a good earth as well if this is possible. In North London, I could receive both the BBC Radio 2 (Light) and Radio 4 (Home) services with about 6 feet of wire. With 50 feet of aerial in the loft I could obtain Radio 1, 2, 3 and 4, with Radio 2 and 4 at extremely loud volume. The trouble then is that it is difficult to separate Radio 2 from Radio 4, and each programme has a slight background of the other.

From this you see that on medium and long waves *small aerials mean good selectivity (i.e. power to separate stations) but small signals*, whereas large aerials give large signals but poor station separation. If you live in a country district a long way

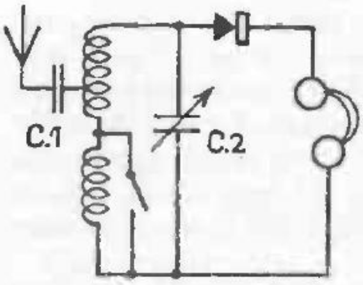

Figure 7 Crystal Set Theoretical Diagram

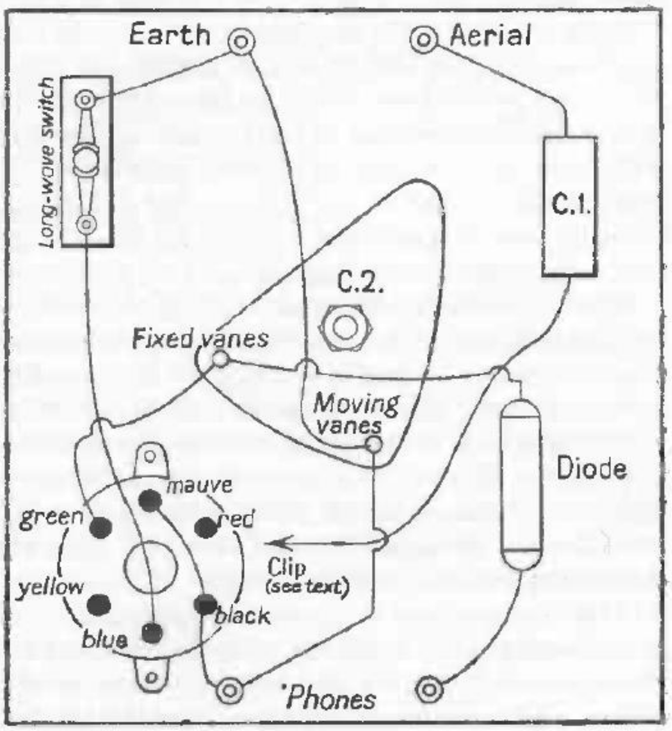

Figure 8 Crystal set Wiring Diagram

Components required:
·0005 mfd. solid dielectric variable capacitor (C.2)
On-off switch
DRR2 coil (Repanco)
·0002-mfd. fixed capacitor (C,1)
Germanium diode (Osram GEX 35 or Mullard OA 81)
4 terminals or 2 double mounts
Crocodile clip
Wood, wire, etc.
Tuning knob

from a BBC transmitter, and can erect a large aerial in the open air, you will probably get good results from the simple receiver; but if you live under twenty miles from a station for best results you need a more selective set.

Firstly, then, to build the set shown theoretically in Fig. 7 with a wiring diagram in Fig. 8, you need a piece of three-ply wood about 2½ inches square. If you have a particular box or cabinet to build the set into, and this requires a slightly larger size, it does not matter at all. You must drill holes for those components which are mounted on the panel; and these are the tuning capacitor in the centre, above and below it the 'A/E' and 'phones' terminals or terminal mounts. If you use the latter, the terminals affixed to them must protrude through holes to the other side, but terminals will automatically pass their shanks through the panel. The particular coil specified has a long-wave section and this requires an 'on-off' switch to short-circuit the long-wave section of the coil when medium waves only are required. As some of the holes may be a little larger than the drill you have available, an easy way to make them is to drill them out as large as possible and then 'reamer' them to size with the tang of a file or other tool.

After mounting those parts which require fixing to the panel, the soldering-up can be started; it will be found that remaining small items can be soldered into place and held into position by the wiring. Earlier remarks on soldering and wire to use may be referred to here, particularly the importance of clean joints made with a hot iron. Avoid 'dry joints' which look good but can be pulled apart quite easily and beware of applying a hot soldering iron for too long to the tags of components. In particular, do not let the iron get too near to the diode or remain too long on its connecting wires.

When soldering to the coil, be careful that the heat of the iron does not unsolder the wires already connected to the tags. Once the wiring has been completed and checked over, the headphones may be joined to the appropriate terminals, with aerial and earth also connected to their correct points.

Headphones must be of the 'high-resistance' type, i.e. 2000 or 4000 ohms impedance. If you have a pair of headphones which are low-resistance type, or yours are marked LR or 120 ohms, you can use them by interposing a special transformer (obtainable quite cheaply) between the set and the phones. Most dealers have these transformers.

After connecting the external apparatus, make sure the switch indicates the waveband required, that is, points *short-circuited*, cutting out the long waveband, if you want to receive on medium waves, and *open* if you want the whole coil in circuit for long-wave reception. Turn the knob slowly and you should pick up the stations available. A push-pull switch may not be available but any small on-off switch will do.

The coil specified is a Repanco DRR 2 and is really a coil with a reaction winding which is normally used in a one-valve or two-valve receiver. I have specified it for the crystal set because, in accordance with our promise of progressive set-building, it can be used for the receiver described in Chapter 4. In the design as given, using the DRR 2 coil, the 'reaction' winding is wired up so that it may be used as an aperiodic aerial winding which will give you a variation in selectivity. For this reason the lead from the aerial (through C.1) terminates in a crocodile clip.

This may be attached either to 'yellow' on the coil or 'red', in order to give the best position for the aerial in use. Try also fixing the diode to 'yellow' instead of to 'fixed vanes'.

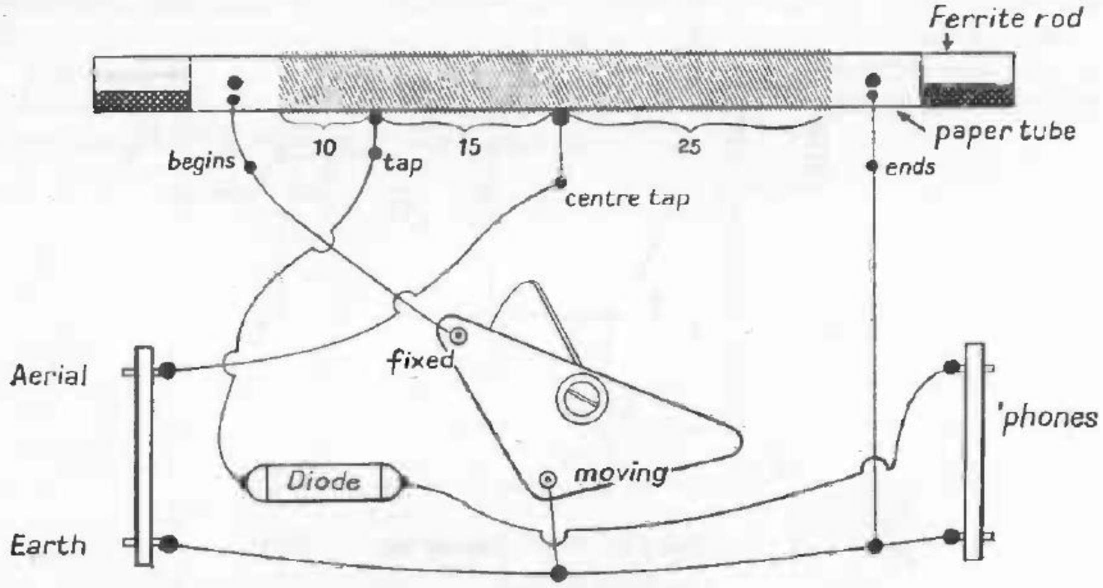

Figure 9 Crystal Set Wiring Diagram (home-made coil)

The second crystal receiver is very cheap and simple to make and uses the *BOP* coil which I first made up for use in two transistor receivers described in *Boy's Own Paper* in 1963. It is very easily wound on a ferrite rod which is 4 inches long and ¼ inch in diameter. At the end of this chapter I repeat the instructions which we gave in *BOP* for winding it. The receiver in Fig. 9 can be put together easily in any handy box and will tune medium waves only, so that if you must get your Light programme on long waves, you would do better to make the design in Fig. 8. Components for this ferrite-rod coil design should be the same as those listed below Fig. 8 except for the coil, of course, which will not be required, nor will the on-off wavechange switch or the crocodile clip. The ·0002 mfd. condenser (capacitor) will help selectivity if it is wired between the aerial terminal and centre-tap on coil. By the way, use two thicknesses of good writing paper or very thin card for making the coil. The wire should not be too close to the ferrite rod itself. You can experiment with this coil quite a lot, if you wish, by putting different windings on it. Readers have told me they have even been using it for short-wave reception in a transistor receiver by using fewer turns and adding a reaction winding. For this crystal receiver, it might be found better to use a tuning condenser of only 200 pf (i.e. ·0002 mfd.) instead of the ·0005 mfd. one specified.

If you build these crystal sets, remember that reception is *entirely* dependent upon the aerial and earth in use. If these are poor, good signals cannot be expected. *Poor signals cannot be amplified or adjusted to make them louder*. A crystal set connected to an amplifier such as the 'small high fidelity' described later makes a very satisfactory arrangement for local station reception. For 'reaching out' and for louder signals a valve detector is required. Do not take your crystal set to pieces unless you have no alternative. The next design shown in Fig. 10 uses the large home-made coil of Fig. 4.

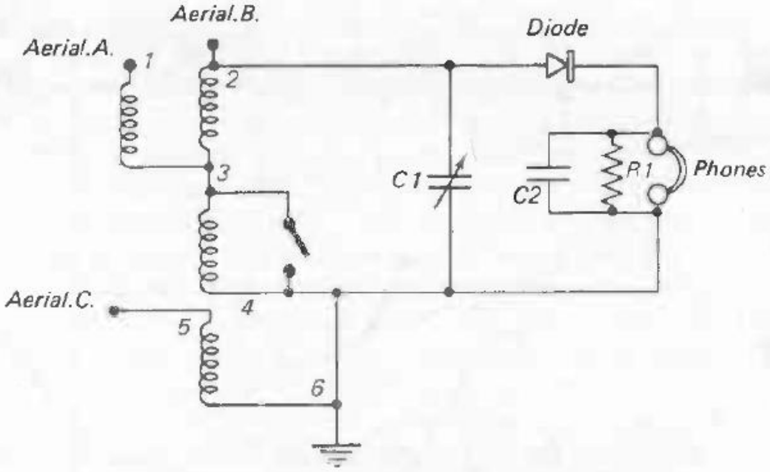

Figure 10 Diode 'Crystal Set'

Components required:
Coil: see text (home made type)
C.1. ·0005 mfd variable capacitor
C.2. ·001 mfd fixed capacitor
Diode: any suitable of good quality
R.1.: 1000 ohm resistor
Phones, wire, wood, etc.

The crystal set shown in Fig. 10 is very simple to make. Fix your coil on a small piece of wood in the manner shown. On one side of that piece of wood fix a small panel on which are mounted the tuning capacitor and two small terminals or a socket strip for phones. Also required will be a small on-off switch for wavechanging. Connect 2 on the coil to one side of diode and fixed plates on the tuning capacitor. Other side of diode to phones. Other side of phones to 4 on coil and 4 on coil also goes to one side of on-off switch plus 6 on coil and moving plates on capacitor. Other side of on-off switch to 3 on coil. Across the phones terminals connect a ·001 mfd. capacitor and if you are using a crystal earpiece you will also need a 1kΩ resistor there. For ordinary headphones (high-resistance) no resistor is needed. The aerial can be attached to a 'crocodile' clip and tried for best results on 1, 2 or 5 on coil. Earth can be clipped to 4 or 6, or you can add two more terminals to the panel for A & E. Your set should now work well and, having fixed a small knob to the shaft of the capacitor, you can tune in the available stations.

Figures 11 and 12 are rather special crystal set designs and I am indebted to Messrs Repanco Ltd, whose copyright designs they are, for allowing me to use them here. The coil featured is the DRR2 which they make and the first circuit is for a headphone crystal receiver which uses 2 diodes and 1½ volt cell. It is rather

CHAPTER 2

Crystal Sets

Crystal sets are simple to build and *cheap ! With six feet of aerial this set will pick up the BBC programmes at good strength; with fifty feet of aerial the volume is really loud. Build the set with care, remembering that reception depends* entirely *upon aerial and earth.*

THESE Crystal Set designs are chosen as the first in our constructional series because of their simplicity and cheapness. The most popular receiver for years was the 'crystal set', the theoretical design of which was precisely the same, in effect, as that shown in Fig 4, our first design. Those old crystal sets, however, used large coils of thick wire and detectors consisting of a piece of crystalline substance 'tickled' by a small cat's-whisker, which was only a piece of thin wire wound up into a small spring. To-day we use a miniature coil with a dust-iron core and a germanium diode detector.

Germanium is one of the modern metals which have been developed for industrial purposes in recent years, and is actually processed from soot-scrapings obtained from certain factory chimneys! A small flake of it is sealed into a tiny glass tube with its cat's-whisker contact cemented permanently into place (Fig. 3).

A crystal set has no power to amplify the signals which it receives via the aerial and the crystal detector merely converts the radio signals into currents capable of actuating the diaphragms of our headphones and producing sounds thereby.

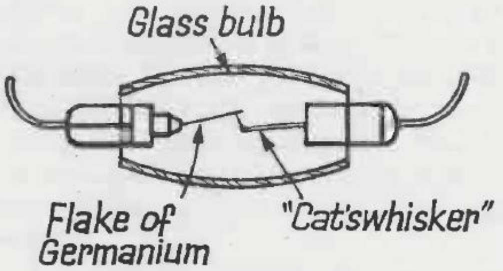

Fig. 3. The Germanium Diode Detector

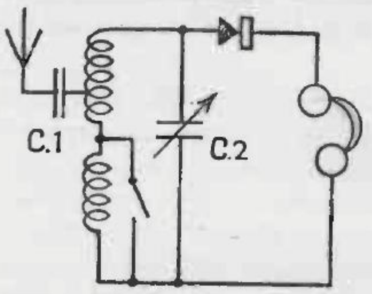

Fig. 4 Crystal Set Theoretical Diagram

It follows, therefore, that the strength of the sound we receive in the headphones depends on the size of the aerial we have—a larger aerial picking up 'bigger' signals—and the efficiency of the detector in converting them ('rectifying' them is the proper term) into 'sound' signals.

The germanium diode is much better as a detector in every way than the old crystal detector and is completely stable in use. The modern crystal set, therefore, must not be confused with that of thirty years ago; it is best to discard the term 'crystal set' really in favour of 'germanium diode receiver'. The receiver in Fig. 4 circuit is small and compact and very simple to make. With a good aerial it will give an extremely loud headphone signal from two or, possibly, three stations. It is essential to have a good aerial with a crystal set, and a good earth as well if this is possible. At my own home, which is ten miles from the BBC (Brookmans Park) transmitters in North London, I can receive both the BBC Home and Light service with about 6 feet of wire. With 50 feet of aerial in the loft I can obtain Home, Light and Third, the first two at extremely loud volume. The trouble then is that it is difficult to separate Home from Light, and each programme has a slight background of the other.

* * *

From this you see that on medium and long waves *small aerials mean good selectivity* (*i.e. power to separate stations*) *but small signals*, whereas large aerials give large signals but poor station separation. If you live in a country district a long way from a BBC transmitter, and can erect a large aerial in the open air, you will probably get good results from the simple receiver; but, if, like me, you live under twenty miles from a station, for best results you need the more selective set shown in our second design.

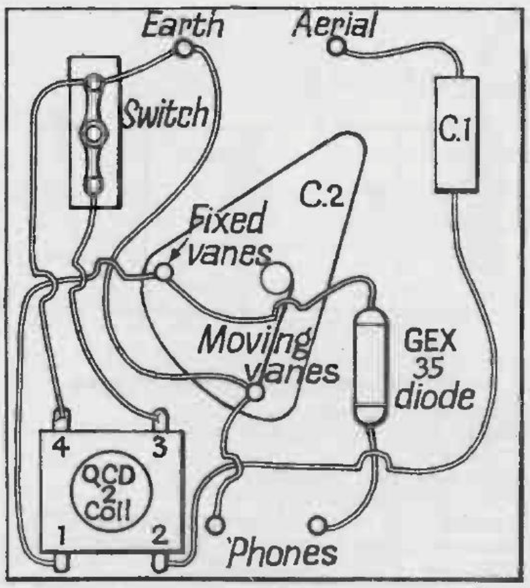

Fig. 5. Crystal Set Wiring Diagram

Components required:
·0005 mfd. solid dielectric variable capacitor (C.2)
On-off switch
QCD 2 Coil (Osmor)
·0002 mfd. fixed capacitor (C.1.)
Germanium diode (Osram GEX 35 or Mullard OA 60)
4 terminals or 2 double mounts
Wood, wire, etc.
Tuning knob
For better selectivity Teletron HAX coil can be used. Connections are given with coil.

Firstly, then, to build the set shown theoretically in Fig. 4 with a wiring diagram in Fig. 5, you need a piece of three-ply wood about 2½ inches square. If you have a particular box or cabinet to build the set into, and this requires a slightly larger size, it does not matter at all. You must drill holes for those components which are mounted on the panel; and these are the tuning capacitor in the centre, above and below it the 'A/E' and 'Phones' terminals or terminal mounts. If you use the latter the terminals affixed to them must protrude through holes to the other side, but terminals will automatically pass their shanks through the panel. The particular coil specified has a long-wave section and this requires an 'on-off' switch to short-circuit the long-wave section of the coil when medium waves only are required. As some of the holes may be a little larger than the drill you have available an easy way to make them is to drill them out as large as possible and then 'reamer' them to size with the tang of a file or other tool.

After mounting those parts which require fixing to the panel, the soldering-up can be started; it will be found that remaining small items can be soldered into place and held into position by the wiring. Earlier remarks on soldering and wire to use may be referred to here, particularly the importance of clean joints made with a hot iron. Avoid 'dry' joints which look good but can be pulled apart quite easily and beware of applying a hot soldering iron for too long to the tags of components. In particular do not let the iron get too

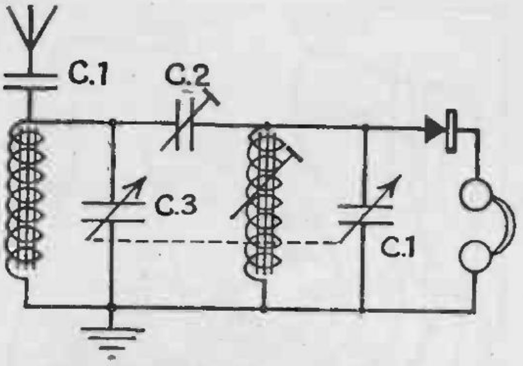

Fig. 6 **Crystal Set Theoretical Diagram**

Components required:

Midget ·0005 mfd. variable capacitor (C.3/4) dual gang
·00005 mfd. fixed capacitor (C.1)
2 Teletron Coils type BA 2
50 mmfd. (i.e., ·00005 mf.) variable trimmer (C.2)
Germanium diode (Osram GEX 35 or Mullard OA 60)
4 terminals or 2 terminal mounts
Knob or small dial

near to the diode or remain too long on its connecting wires.

When soldering to the coil be careful that the heat of the iron does not un-solder the wires already connected to the tags. Once the wiring has been completed and checked over, the headphones may be joined to the appropriate terminals, with aerial and earth also connected to their correct points.

Headphones must be of the 'high resistance' type, i.e. 2000 or 4000 ohms impedance. If you have bought a pair of Government surplus 'phones which are low-resistance type, or yours are marked LR or 120 ohms, you can use them by interposing a special transformer (obtainable quite cheaply) between the set and the 'phones. Most dealers have these transformers.

After connecting the external apparatus, make sure the switch indicates the waveband required, that is, points *short-circuited*, cutting out the long waveband, if you want to receive on medium waves, and *open* if you want the whole coil in circuit for long-wave reception. Turn the knob slowly and you should pick up the stations available.

The selective receiver in my next design is shown theoretically in Fig. 6, and fully wired in Fig. 7. It costs a little more than the set just described, but if a large aerial is available is well worth having because of the large signal the aerial brings in, which

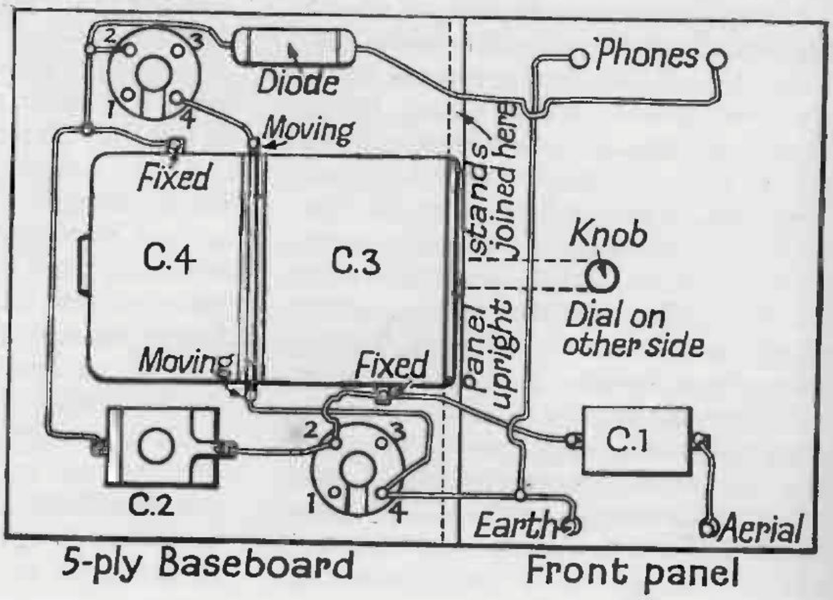

Fig. 7
Crystal Set Wiring Diagram

FREE-POWER AM RECEIVER (P,E)

This project uses a strange way to get its power. It uses power "extracted" from the tuned station to give to the amplifier transistor. With a reasonably strong applied modulation envelope, this type of detector produces a strong demodulated output.

Naturally, the closer you are to a strong station, the more current your radio will be able to supply.

For best results, you must do everything possible to deliver a strong signal to the transistor detector. We recommend a good antenna and ground, the latter preferably being made to a water pipe or solid external ground composed of a pipe driven at least 5 feet into moist earth. This is important for ensuring maximum signal pickup.

If you have plenty of space available, the longer the antenna (up to about 100 feet), the better the results.

Figure 1 shows the schematic diagram of the Free-Power AM receiver. Observe that the heart of the circuit is a germanium transistor that works as a detector and audio amplifier.

Components placement on a terminal strip used as a chassis is shown in ***Figure 2.*** The terminal strip can be fixed on an experiment board that can be constructed using some common tools and materials.

Q1 is any germanium transistor, such as GE-2 or 2N107. You can find germanium transistors in old nonworking AM transistor radio receivers. Nonworking AM radios can also supply the variable capacitor and the loopstick (ferrite rod).

Earphone must be a high-impedance crystal type. Low-impedance earphones don't function in this circuit.

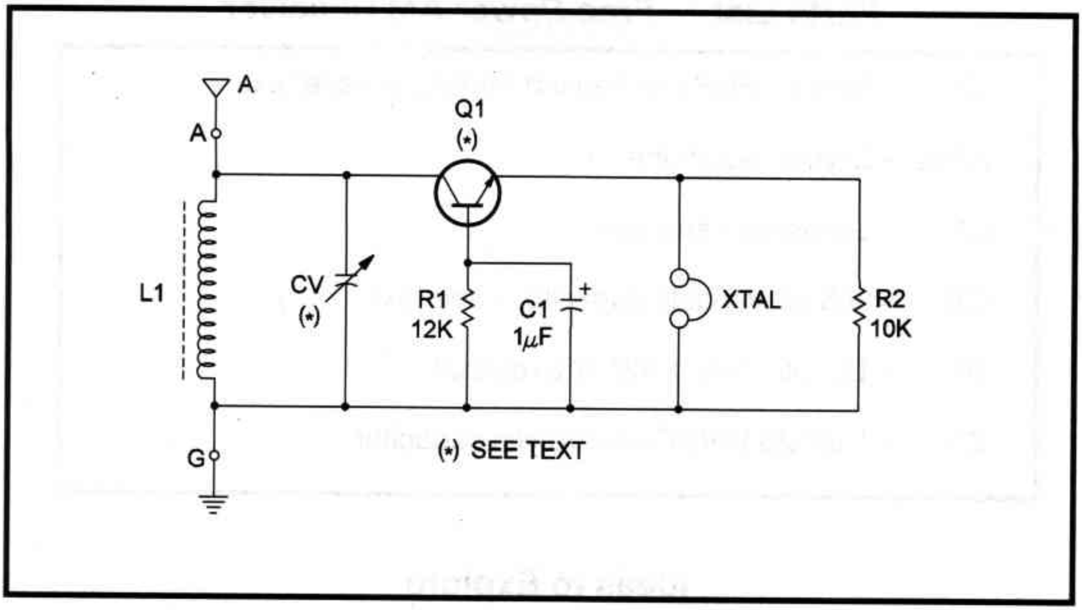
A
A
Q1
(*)
L1
CV
(*)
R1
12K
C1
1μF
+
XTAL
R2
10K
G
(*) SEE TEXT

Figure 1

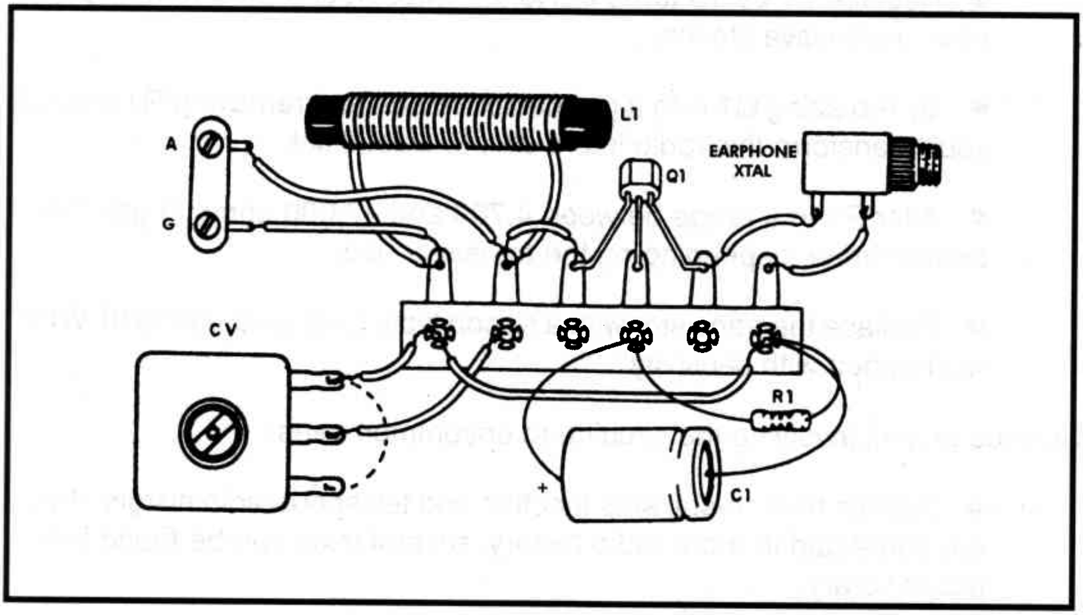
L1
A
G
EARPHONE
XTAL
Q1
CV
R1
+
C1

Figure 2

Parts List — Free-Power AM Receiver

Q1	- NPN or PNP germanium transistor - see text
XTAL	- Crystal earphone
L1	- Loopstick - see text
CV	- 365 pF variable capacitor - see text
R1	- 12,000 ohm, 1/4W, 5% resistor
C1	- 1 uF, 25 WVDC electrolytic capacitor

Ideas to Explore

To learn more about the circuit:

- Replace L1 by 15 turns of No. 28 wire, wound on a ferrite rod, to tune short-wave stations.
- By replacing Q1 with a germanium diode and removing R1 and C1 you'll transform this radio into a simple crystal set.
- Alter R1 in a range between 4,700 and 47,000 ohms to get better performance, depending on the transistor used.
- Replace the transistor with a silicon type, such as the BC548. What will happen with sensivity?

Science project involving the circuit and uncommon uses:

- Explain how crystal sets function and tell about radio history. If you are interested in more radio history, several texts can be found in the public library.
- If you want to learn more about radio receivers, try to find information about heterodyne and super-hereterodyne types.

BENJAMIN AND DAVID GOLDENBERG 'S VERY-FINE

OLD-TIME CRYSTAL RADIO

Parts List:

Item	Qty.
PVC pipe, 1" outside diameter (3/4” size) x 4" long	1
Aluminum or steel sheet metal, 0.5” wide x 5” long	1
Coil wire (#24 magnet wire)	36 ft
Hook-up wire (#22)	25 ft
Crystal earphone (GC electronics, cat. no. 30-5302)	1
Jack for the earphone plug	1
Crystal diode (1N34A)	1
Capacitor, 470 picofarad	1
#8 sheet metal screws	8
8-32 machine screw, 1" long	1
8-32 nuts	2
flat washers	6
Ring terminal	1
Board, 3/4" thick, approximately 7" x 9"	1

What else you need:

Vise-Grip pliers
Wire cutters
Soldering iron and solder
Drill, with 1/8", 3/16” and 3/8” bits
File or motorized grinder (e.g. Moto-Tool)
Screwdriver
Scotch tape

Instructions

I. CONSTRUCTION OF SPECIAL PARTS.

The following steps involve making a few special parts for the radio. Some of these steps are most easily performed using power tools and some involve soldering. This work should be done by someone old enough and knowledgeable enough to handle these tools safely. Once these parts are made, however, the rest of the project requires only hand tools and can be completed by nearly anyone.

1. Coil Form

The coil form is a piece of PVC pipe, 1” in diameter (outside) and 4” long. (For reasons fully understood only by plumbers, pipe of this diameter is referred to as 3/4”.)

To facilitate mounting the coil to the board, drill two sets of holes, on opposite sides of the form. On one side of the form, drill two holes, 3/16” in diameter and 3/8” from the ends. Directly opposite from these holes, drill two more holes 3/8” in diameter. Screws will be placed through the smaller holes, while the larger ones make it easy to reach the screws with a screwdriver.

2. Slider

Drill two 3/16” holes in the piece of sheet metal, each 1/4” from an end. After the holes are drilled, place a bend in the metal 3/4” from each end, as shown in the figure on page 3. Use a vise or pliers to bend the metal. Use a file to smooth all of the edges and corners of the metal.

CAUTION: Drilling sheet metal can be very dangerous! If the bit catches on the metal, you can have a very sharp edge flying through your hands. Hold the metal firmly with a pair of pliers (preferably Vice-Grip type), not with your hands! Also, place a piece of wood under the metal as you drill. This will help keep the metal from tearing and grabbing the bit as the passes through.

3. Pointed screw for slider

Using either a file or a grinder, grind a point on the 8-32 machine screw. The point doesn’t need to be very sharp; about like the tip of a medium ball-point pen is ideal.

4. Earphone jack with wires.

Cut two pieces of hook-up wire 2” long. Remove 1/4” of insulation form one end of each wire and solder the wires to the terminals of the earphone jack. Remove 1/2” of insulation from the other ends of the wires.

Note: Headphone jacks often have more than two terminals. Make sure that the terminals you solder to are the ones that will be connected to the tip and side of the plug when it is inserted.

5. Ring terminal with two wires.

Cut two pieces of hook-up wire, one 7” long and the other 2”. Remove 1/4” of insulation from one end of each wire. Solder the two wires to the ring terminal. Remove 1/2” of insulation from the other ends of the wires.

6. Board

Tape the layout template (the last page of this document) to a board, approximately 7” x 9”, and drill a 1/8" hole at each position marked with a cross.

II. ASSEMBLY

1. Wind the coil.

The coil is made by wrapping the #24 magnet wire onto the coil form, which is made from a piece of PVC pipe. The magnet wire is the wire that looks like it has no plastic insulation on it. But, it is actually coated with a thin layer of enamel. This coating must be removed from the ends of the wire before making any connections.

The best way to wind the coil is with two people, one to hold the wire taught (the holder) and the other to wind it onto the coil form (the wrapper). First, thread the end of the coil wire from the outside of the form through one of the large holes and out the end of the form. Leave about 4" of wire going out the end. Hold onto the other end of the wire and begin wrapping it around the form, starting right next to the hole. After you have wound a few turns on, push the wire together so it is nice and neat and put a piece of Scotch tape over it. The coil should look like this:

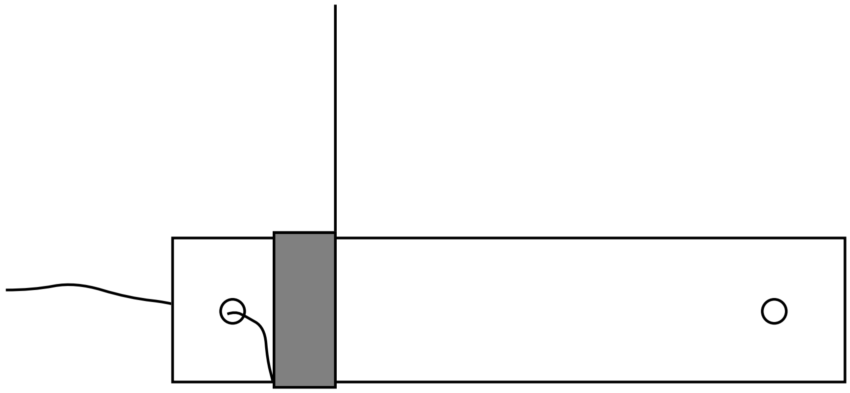

Now, the "holder" should hold the spool of wire loosely while the "wrapper" walks across the room unwinding the wire. Then, the holder should hold the spool very tightly so that the wrapper can wind the wire onto the pipe evenly. Each turn should be wrapped on right next to the previous turn, but shouldn't overlap. As the wire is wrapped on, the wrapper slowly walks across the room keeping the wire tight. When the wrapper reaches the holder, hold the wire that has been wrapped onto the form tightly, while the holder lets the wire on the spool unwrap. The wrapper then walks back across the room unwinding the wire from the spool, and the same procedure is repeated.

When the coil is within about 1/8" of the holes at the other end of the coil, stop winding. Put a piece of Scotch tape on the end of the wrapping, and cut the wire a few inches from the end of the coil. Thread the end of the wire through the large hole in the form.

2. Attach the coil to the board.

Use two of the #8 sheet metal screws to attach the coil to the board, using holes B and C that you drilled. Place the screws through the small holes in the coil form from the inside. You can then use the large holes on the other side of the form to reach the screws with a screwdriver. Make sure the coil is attached snugly to the board.

3. Assemble the slider.

Find the slider, the 8-32 screw and the two 8-32 nuts. Thread one of the nuts onto the screw until it is about 3/8" from the head of the screw. Put the screw through one of the holes in the slider and thread the other nut on, so that the slider looks like this from the side:

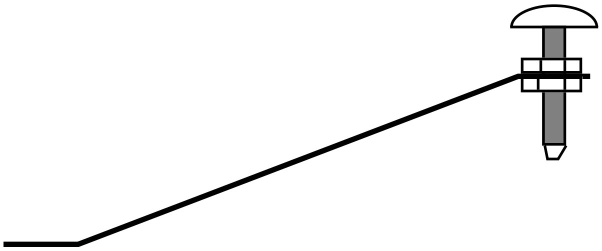

Tighten the second nut against the slider and the first nut with a wrench.

4. Attach the slider to the board.

Find a sheet metal screw, a washer and the ring terminal with two wires attached to it. Put the screw through the ring terminal, the washer and finally the open hole in the slider assembly (from the top of the slider as shown above). Attach the entire assembly to the board with the screw at hole E. Tighten the screw so that it is snug but still allows the slider to move back and forth. The pointed end of the machine screw at the other end of the slider should press firmly against the wire on the coil. If necessary bend the slider so that it makes firm contact with the coil but can still move.

Move the slider back and forth on the coil to scrape the enamel insulation from the coil wire where the pointed end of the machine screw touches it. The screw should be able to make an electrical contact with the coil wire at any position along the length of the coil.

5. Put the remaining screws in the board.

Screw the remaining sheet metal screws into holes A, D, F, G and H, with a washer on each screw. Do not tighten the screws yet.

6. Wire the components of the radio together.

Use the circuit drawing as a guide. For each connection, make sure that the insulation is removed from the end of the wire. Wrap the bare ends of the wires around the screw between the screw head and the washer underneath.

a. Wrap the two wires from the earphone jack around screws F and H. Do not tighten these screws yet.

b. Take the short wire from the slider and wrap the bare end around screw F (which should already be attached to one of the earphone jack wires). Do not tighten the screw yet.

c. Take the long wire from the slider and wrap the bare end around screw A. Do not tighten this screw yet.

d. Find the crystal diode, the small cylinder with two bare wires attached to it. Wrap one of the wires around screw H (attached to the other earphone jack wire), and tighten the screw.

e. Take the other wire from the diode and wrap it around screw G. Do not tighten this screw yet.

f. Find the capacitor, the small disc with two bare wires attached to it. Wrap one of the wires around screw F and tighten the screw.

g. Wrap the other wire from the capacitor around screw G, but do not tighten this screw yet.

h. Cut a piece of hook up wire about 6" long and remove about 1/2" of insulation from each end. Wrap one end around screw G and tighten the screw. Wrap the other end of the wire around screw D, but do not tighten this screw yet.

i. Take the wire from the right hand end of the coil and remove the insulation from about 1/2" from the end of the wire. To remove the insulation, either lightly scrape the wire with a knife or use a small piece of sandpaper. Wrap the bare end of the wire around screw D, but do not tighten the screw yet.

7. Connect the radio to ground.

In order to work, the crystal radio must be wired to a good ground connection. Often, the easiest way to make a ground connection is to use the grounding circuit for the house's electrical system. Cut a piece of hook-up wire long enough to reach from the radio to a nearby electrical outlet. Remove 1/2" of insulation from each end of the wire. Wrap one end around screw A of the radio and tighten the screw. On the electrical outlet, loosen the screw that holds the plate on the box. Wrap the other end of the wire around this screw and tighten it.

CAUTION: Be sure not to let the wire or anything else touch the 120 V socket!

Alternatively, you can use a cold water pipe to connect the radio to ground. Remove a few inches of insulation from the wire from screw A of the radio and wrap the bare end around a cold water pipe.

8. Connect an antenna to the radio.

Use the rest of the hook-up wire to make an antenna. Remove 1/2" of insulation from one end of the wire and wrap it around screw D. Tighten the screw. Unwind the wire from the spool and try to place it as high as possible in the room.

In areas with many powerful radio stations nearby, an inside antenna should be adequate to receive a few stations. If you have trouble hearing anything, it may help to set up a long outside antenna. Radio Shack sells a kit with the necessary wire and other parts (cat. no. 278-758).

9. Try the radio!

Plug the earphone into the jack. Listen to the earphone and slowly move the slider across the coil. Try to find the position where a station is loudest. You should be able to hear a few different stations with the slider at different positions.

10. If the radio doesn't work, check the following:

a. Make sure all of the connections are correct and that the insulation has been removed from the wires where they are wrapped around screws. Also check that there aren't any short circuits between the screws or bare wires.

b. Make sure that you scraped the insulation from the coil wire with the slider so that the slider can make good contact with the coil.

c. Make sure that you have a good ground connection. Try a different electrical outlet or a cold water pipe.

d. Place the antenna higher up or try a longer wire.

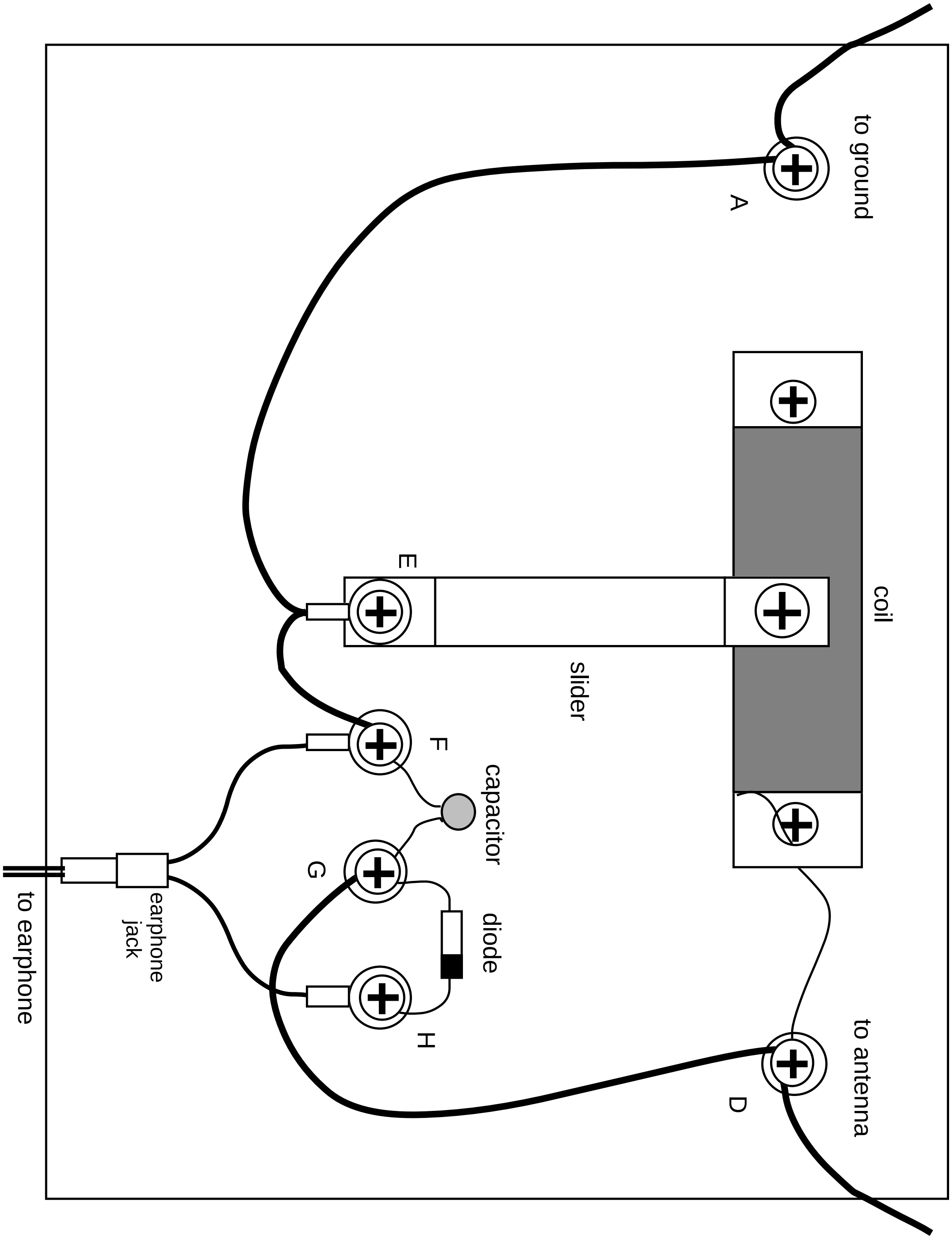
to ground
A
coil
E
slider
F
capacitor
G
diode
H
to antenna
D
earphone jack
to earphone

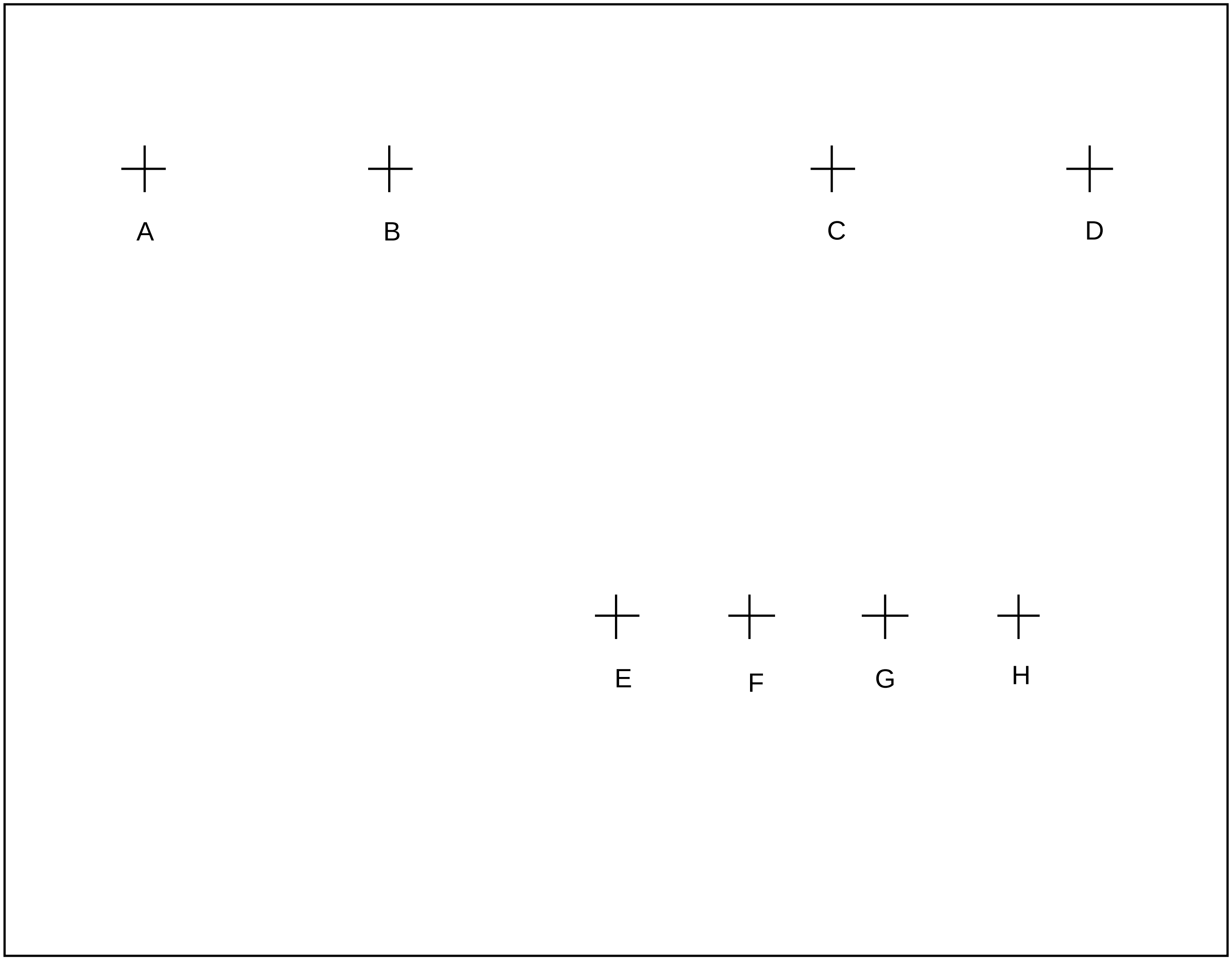
A
B
C
D
E
F
G
H

Electronic Projects for Beginners

F. G. RAYER, T.Eng. (CEI), Assoc. IERE

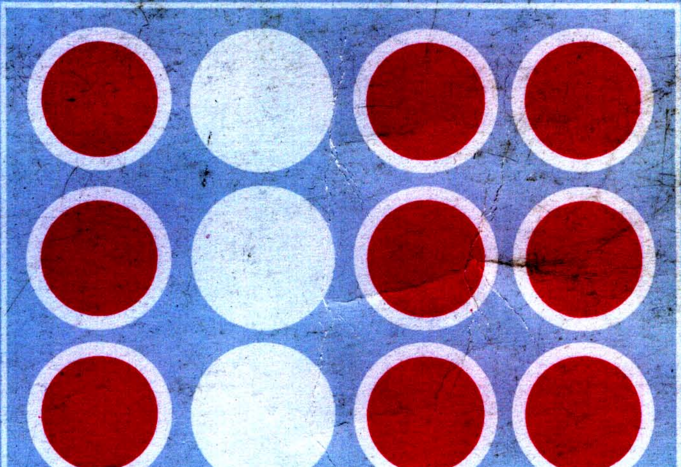

speaker SP2. SP1 then acts as a microphone, and sounds are reproduced in SP2.

With the switch in the opposite position, S1A selects SP2 as microphone, and amplified signals pass via S1B to SP1, so that speech can be in the other direction.

The switch, amplifier, battery and SP2 are assembled as the master control unit, and twin leads run to SP1. The person at SP1 should say "Over" so that the other person can operate the switch, and he in turn should say "Over" to signify that SP1 is again acting as microphone.

A simplified home-made change over switch is shown in Fig. 5. This is two strips of metal, pivoted so that they can be swung from side to side, and bearing on two of the three round-headed screws. A strip of insulating material is fitted so that both strips move together. In Fig. 5, this switch is in its "Off" position.

Ready made 2-pole 2-way switches can be obtained and used instead (see Fig. 49). Spring-loaded push switches are also available, and can be wired to provide "press to talk". A switch without a central "Off" position can also be used, a single pole on-off switch then being added in one battery connection.

The master unit can call SP1 by switching to speak, and allowing SP2 to be used as microphone. If calling in the other direction is required, a third connection, with bell push, battery and buzzer can be used. (An alternative, requiring more circuitry, but avoiding the need for a third wire, will be found in "Two Transistor Electronic Projects", (BP30, Babani Press.)

Fig. 5B shows assembly of the intercom in a wooden case. A 3½ inch speaker is suitable, so the whole can be quite a compact unit. The home-constructed switch, using the bolts below the speaker, can be replaced by a ready-made switch, as described, if preferred.

Best results are obtained with speaker units of quite high impedance – say 45 ohm to 75 ohm or so. The few other components are assembled on a small insulated board, as for earlier circuits.

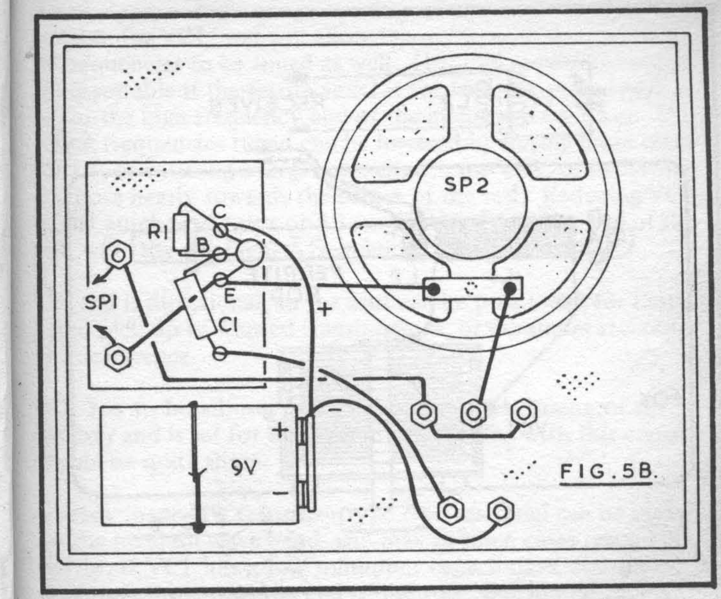

FIG. 5B.

The speaker at the remote point also requires a small case. Dimensions are not important, and it is possible to utilise plastic boxes of suitable type for each unit. Connections between units are by twin bell wire, or any similar twin conductor. With a transistor of other than the type shown, the value of R1 can be chosen to give maximum amplification.

Tuned Ferrite Aerial

This simple device will allow "portable" reception with the old type of receiver which normally requires an external aerial, and it can also give improved reception in terms of freedom from interference on the 160 metre amateur band and adjacent frequencies. As shown in Fig. 6, it consists of a tuned circuit with ferrite rod, and a coupling winding to connect to the receiver aerial and earth sockets. It will be found very useful with a communications type receiver, where a long extended outdoor or similar aerial results in a high level of noise and interference.

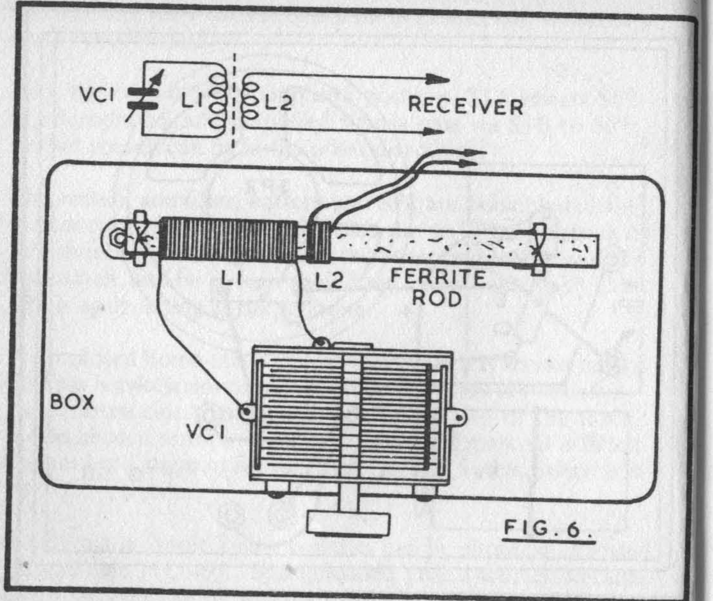

FIG. 6.

The device is constructed in a plastic lunch box about 6 x 4 x 4 inch or other *insulated* case. Some boxes of this kind, though strong, are of brittle material. These must be drilled without too much pressure, or they may crack. Drill points should also be sharp.

The rod is 6 inch long and 3/8 inch in diameter, and it is supported by two strips of insulating material, which are fixed with brackets. Each strip has a V-shaped notch, so that the rod can be held with thin string. L1 is the tuned winding, and it had sixty turns of 24 swg enamelled wire, side by side. Connections can be made to the frame (moving plates) and fixed plates of the capacitor VC1 by means of small bolts, or may be lightly twisted.

L2 is the coupling winding. Many communications type receivers have a low impedance input circuit, and twelve turns will be adequate for L2. The same wire can be used as for L1. With other receivers it may be worth while to experiment with the number of turns on L2, for best reception.

An air-spaced single gang capacitor of about 315pF is most suitable for VC1, and will allow the usual medium wave band of frequencies to be tuned as well. However, a smaller capacitor is suitable if the ferrite aerial is for 160m reception only, or for the high frequency end of the medium wave band. The actual frequencies tuned can be lowered by having more turns for L1, or by using a larger capacitance for VC1, or by having L1 more nearly towards the centre of the rod. Reducing VC1, or the number of turns on L1, or moving L1 to the end of the rod, will raise the highest frequency tuned.

The rod is directional, so the unit can be positioned for best signal pick up of wanted transmissions, or minimum reception of interference.

VC1 has to be adjusted in conjunction with tuning of the receiver and is set for best reception. Tuning with this control should be quite sharp.

A ready made MW transistor type receiver aerial can be useful for the medium wave band, and may in some cases reach 160 metres, if VC1 has a low minimum capacitance, and the winding is quite near the end of the rod. But should tuning not reach a high enough frequency, some turns will have to be removed.

Diode Radio

A lot of interest can be obtained from a diode radio receiver. It requires few components, no battery or other supply, and can give good headphone reception of a few of the more powerful transmitters. It can also be used as a tuner, as shown later, and will provide excellent audio quality for local station reception.

The circuit in Fig. 7 is intended for medium wave reception. That is, from about 200 metres to 500 metres, or 1500 kilohertz to 600 kilohertz.

The coil L1 is tuned by the variable capacitor VC1. At resonance with the wanted transmission, a signal voltage is developed across the coil. Detection or demodulation by the diode D1 allows the audio component of the signal to be recovered, and taken to the phones or to an amplifier. Such "crystal sets" were used in very many homes, in the early days of radio.

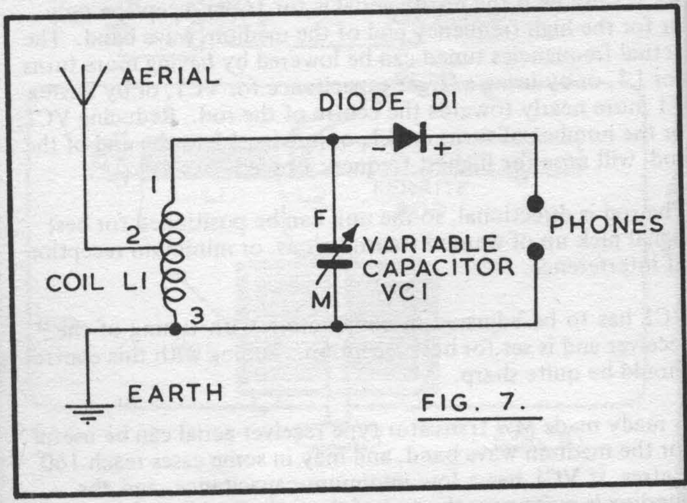

FIG. 7.

Fig. 7B shows the practical layout. It is convenient to use an insulated board larger than necessary, so that a transistor amplifier can be added later. A board about 5 x 8 inch is suggested, and it can be supported on 1 inch high runners (strips of wood) so that there is clearance for screws or terminals.

Coil L1 is wound with 26swg or similar wire, and has seventy turns, side by side. Secure the wire by passing it through two small holes, wind on about thirty five turns, and form a loop. Continue to wind in the same direction for a further thirty five turns, cut off the wire, and secure the end through two holes. The tube used can be 1½ inch to 2 inch in diameter, and can be paxolin, or can be made by winding thin card round a suitable object, smearing a little adhesive on to form a tube.

Point 1, the beginning of the coil, is taken to the fixed plates tag F of the variable capacitor, which is also connected to diode D1. The tapping 2 goes to the aerial terminal. End 3 is taken to the moving plates connection or metal frame of the capacitor. Ends 1 and 2, and loop 3, are left long enough to reach these points, and the insulation is carefully scraped off the wire where connections are to be made. The capacitor VC1 can best be about 500pF, but this exact value is not essential.

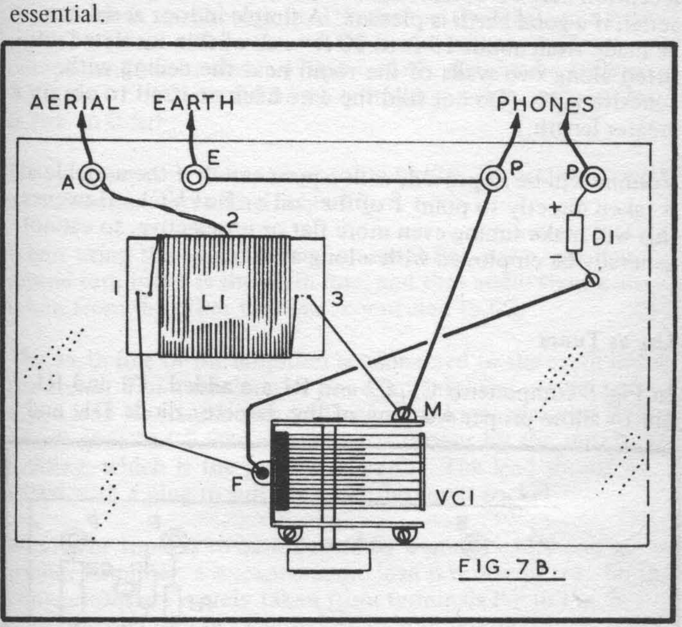

FIG.7B.

Four terminals, for Aerial, Earth, and Phones, will complete the receiver. Best results will be had by connecting a good headset, of about 500 ohm to 2,000 ohm. Very low impedance earpieces, such as are employed with some transistor receivers, are not suitable here.

With a simple receiver of this kind it is necessary to employ a reasonably effective aerial and earth. Only in some circumstances will volume be great enough with no earth, or with a short aerial. The earth connection can be taken to a metal spike in the ground, or to a metal object buried in soil, or to any other available earth other than a mains circuit earth. (No connection must be made to any mains plug or other mains circuit.) A cold water pipe may provide a good earth. Gas pipes should not be used.

For the aerial, a length of aerial wire some 30 ft to 75 ft or so in all may form the aerial itself and lead-in. If this can be suspended quite high, and clear of buildings or earthed objects, signal strength will be improved. However, satisfactory reception may be obtained from an indoor aerial, or shorter aerial, if a good earth is present. A simple indoor aerial can be made from about 15 ft to 20 ft or so of thin insulated wire, fitted along two walls of the room near the ceiling with adhesive tape. (Do not fold the wire back on itself to obtain a greater length.)

Volume will be improved, with a poor aerial, if the aerial lead is taken directly to point 1 of the coil or F at VC1. However, this will make tuning even more flat or unselective, so cannot generally be employed with a long aerial.

Use as Tuner

In Fig. 8 components C1, C2 and R1 are added. C1 and R1 are to allow proper working of the detector diode D1, and may be considered as replacing the headphones. C2 is a coupling capacitor passing audio signals to the following amplifier.

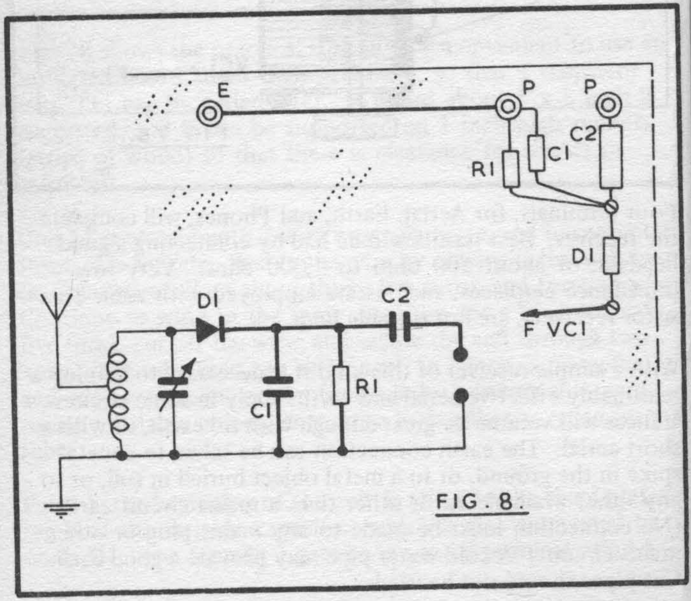

FIG. 8.

As with the previous circuit, any general purpose detection type diode is satisfactory for D1, such as the OA91 and many similar devices. The value of C1 is not critical, and this component can be about 470pF or 1nF. R1 can be 100k to 220k or so. C2 should be of quite large value, and can be 0.1μF to 0.5μF.

D1 can be moved from the position in Fig. 7B, and the few extra components added, as in Fig. 8.

When using the tuner to feed an amplifier, note that one phone terminal P is the earth line, and that audio signals are taken from the other terminal, connected to C2.

The earth line of the amplifier is connected to the earth line of the tuner. For a sensitive or powerful amplifier, a screened audio lead is recommended. This has an inner screened lead, which goes to C2, and a conductor former by the outer brading, which is the ground or earth. The lead should be fitted with a plug to suit the amplifier input socket.

Where the tuner is to be followed by a small headphone or similar amplifier, a screened audio lead is not required. So then connections are merely taken from terminals P-P in Fig. 8, remembering to join amplifier earth and tuner earth circuits as explained.

Medium and Long Wave

The addition of long wave coverage is worthwhile in some areas, and the coil can be modified as in Fig. 9 so that both MW and LW can be tuned.

A tube about 3 inch long and approximately 1½ inch in diameter will be necessary. The section L1 is wound as before. L2 consists of about two hundred turns of 34swg or 36swg wire, in two compact piles of one hundred turns each. All turns throughout are in the same direction. The junction of L1 and L2 goes to the waveband switch S1. The second switch terminal goes to 4 and earth line. MW tuning is obtained with S1 closed, and LW tuning with S1 open.

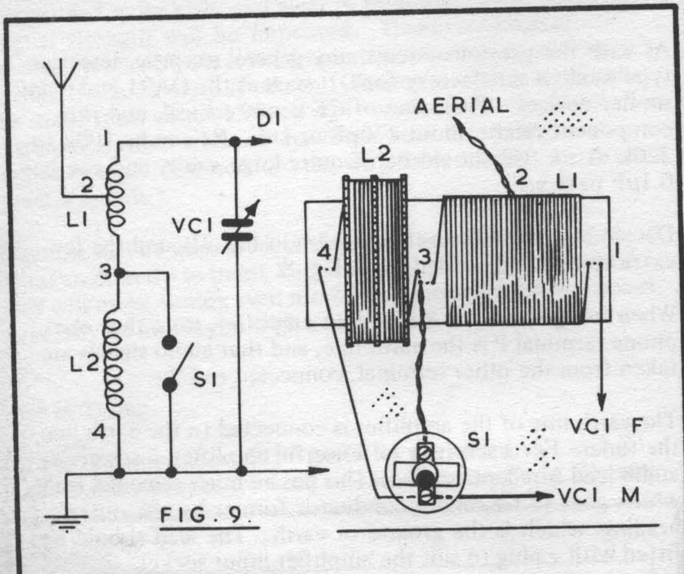

FIG. 9.

For accurate coverage of the usual MW and LW bands, and using a 1¼ inch diameter tube, L1 can be eighty-four turns of 32swg enamelled wire, with tapping 2 sixty turns from point 1. Leave a space of 5/8 inch and glue on a card disc cut to fit the tube. Also glue on two further discs, leaving spaces of ¼ inch between them. L2 is wound with 32swg double cotton covered wire, with one hundred and twenty turns in each slot (making 240 in all for L2). The coil can be fixed with two screws or small bolts, with spacers or extra nuts to hold it a little clear of the board. The windings should not be varnished, covered with adhesive, or wax.

With circuits of this kind there is no ganged tuning, or need for exact tuning of a particular frequency band. This means that coil winding details (wire gauge and covering, and diameter, or number of turns) can be varied to suit materials to hand. Ready-made coils can also be tried. Some commercially made coils with cores can give a slightly better performance than the air-cored home made type of coil. Coils can also be wound in a similar way to that in Fig. 6.

One Transistor Amplifier

Headphone volume obtained from the crystal diode receiver can be considerably boosted by adding a single transistor audio amplifier. Space for this can be found on the circuit board.

In Fig. 10, resistor R2 and the transistor have been added, with a switch S2 for on-off switching of the battery. R1, C1, D1 and C2 are already present, as shown in Fig. 8.

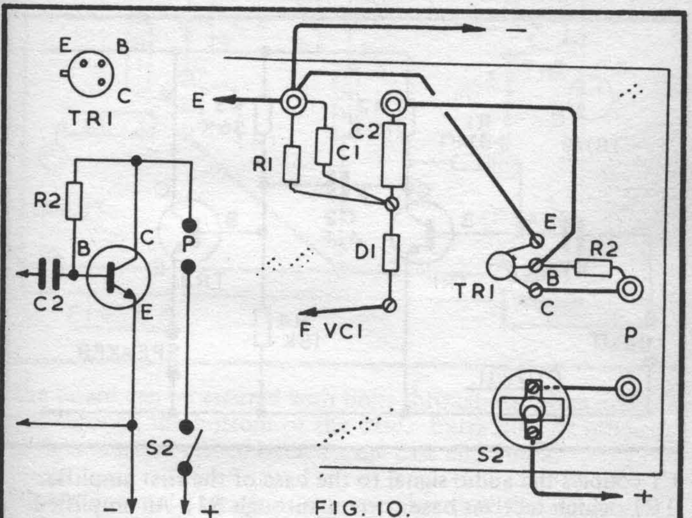

FIG. 10.

TR1, R2 and S2 are connected as shown. For TR1, a BC108, 2N706, or similar general purpose NPN transistor will be satisfactory. R2 can generally be 1 megohm with satisfactory results, but it can be worth while trying values from about 100k to 2.2 megohm here, as performance depends on the transistor, phones, and battery voltage.

Medium and high impedance phones, such as a 600 ohm DC headset, will be most suitable. The battery can be 4.5v to 9v.

If the phones are first connected to terminals P-P in Fig. 8, then taken to terminals P in Fig. 10, a very considerable increase in volume should be obtained.

If a PNP transistor is to be tried in this circuit, the polarity of the battery *must* be reversed.

Two Transistor Amplifier

Fig. 11 is the circuit for an easily made 2-transistor amplifier which can provide reasonable loudspeaker volume. The audio signal is taken to the input terminals, battery negative being the grounded side of this circuit.

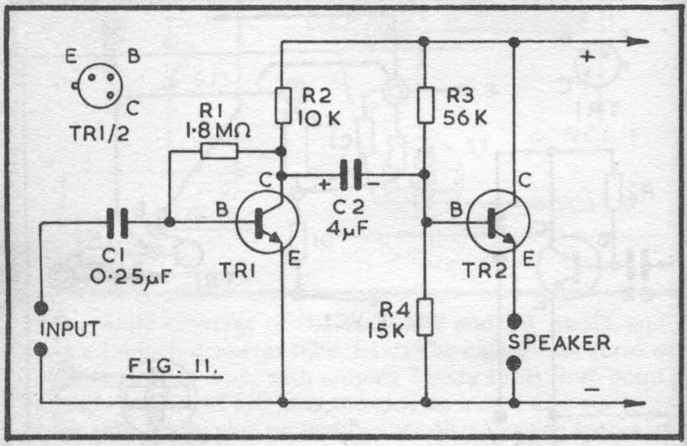

FIG. 11.

C1 couples the audio signal to the base of the first amplifier TR1, which receives base current through R1. An amplified signal is developed across the collector resistor R2, and is coupled to the base of the following amplifier TR2 by means of C2.

The two resistors R3 and R4 set the base operating conditions for TR2, and the output is taken from TR2 emitter.

Fig. 11B shows assembly of the components on a small insulated board. If preferred, connections between components can be on the underside of the board, in the way shown for later circuits. Fit red and black flexible leads for battery connections. The amplifier, with its battery, can be fixed in an insulated box or metal case. Sockets can then be used for input and output circuits, and an on-off switch can be placed in one battery lead.

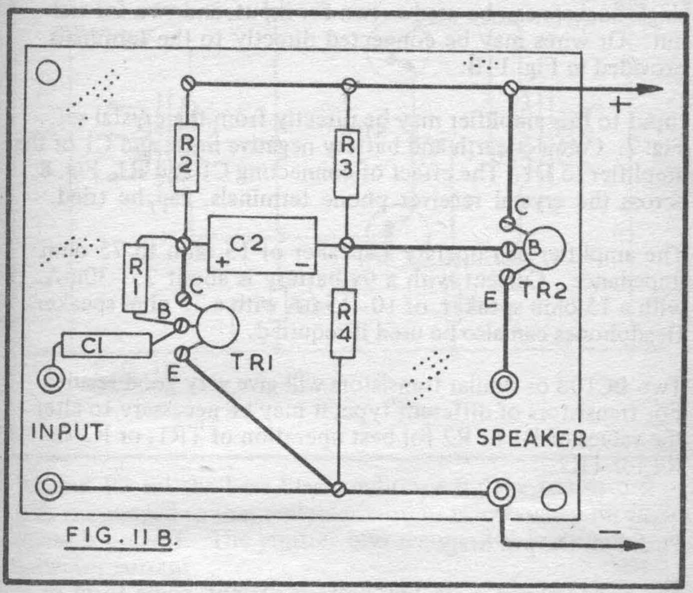

FIG. 11 B.

The board can be secured with bolts through its fixing holes, and through the bottom of the case. Extra nuts or other spacers must be placed between the case and board.

Whenever a metal box is used, remember that all connections and other circuits except the earth line must be insulated from the metal. The metal case itself is normally connected to the earth line, and a lead can be put under one of the fixing screws for this purpose. In Fig. 11, both input and speaker are returned to the earth line. So 3.5mm jack sockets of the type in which the outer or sleeve contact is common to the fixing bush could be used, as these points will be connected to battery negative and the earth line and metal box. But with some circuits it is necessary to use an insulated jack socket, or to insulate this from the metal. An example of this would arise if the speaker were placed in the collector circuit of TR2, as the outer or sleeve circuit would then need to go to battery positive.

Instead of the type of socket which takes a jack plug, ordinary single sockets can be used – two for input, and two for output. Or wires may be connected directly to the terminals provided in Fig. 11B.

Input to this amplifier may be directly from the crystal set, Fig. 7. Connect earth and battery negative lines; and C1 of the amplifier to D1. The effect of connecting C1 and R1, Fig. 8, across the crystal receiver phone terminals, can be tried.

The amplifier can operate a speaker of 15 ohm to 75 ohm impedance. Current with a 9v battery is about 25–30mA with a 15 ohm speaker, or 10–15mA with a 75 ohm speaker. Headphones can also be used if required.

Two BC108 or similar transistors will give very good results. For transistors of different type, it may be necessary to alter the values of R1 or R2 for best operation of TR1, or R3 or R4 for TR2.

Push-Pull Output Stage

For good volume, with low battery current, some form of push-pull output stage is generally used. This has two transistors, and each amplifies one half of the audio signal. Base bias conditions of the two transistors are so arranged that only a low current flows during silent intervals. Overall battery drain is thus much lower than if a single transistor were used to obtain a similar level of volume.

Fig. 12 is a typical push-pull output stage. T1 is the input or driver transformer, and an audio signal is taken to its primary P. The secondary S is centre tapped. So for one half of the audio cycle one transistor base is driven positive and the other transistor base is driven negative, this being reversed for the remaining half cycle. Thus each transistor conducts alternately, and the centre-tapped output transformer T2 combines the collector currents, to operate the speaker connected to the secondary S.

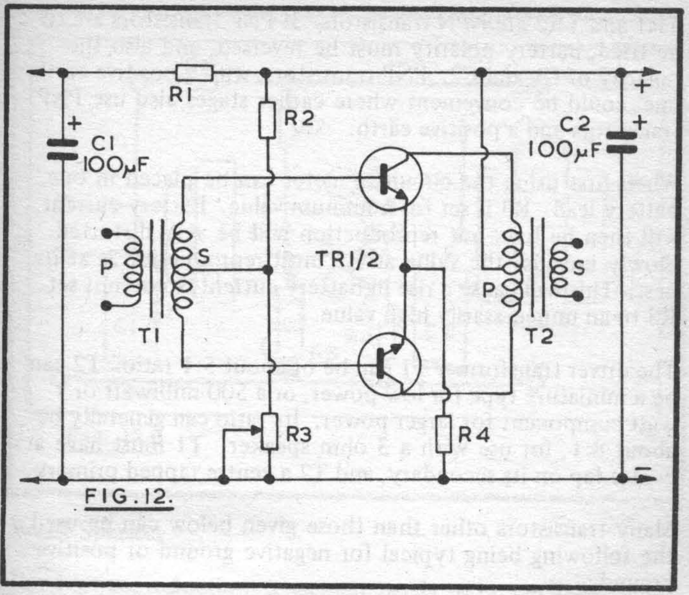

FIG.12.

R2 and R3 set the base bias conditions for the transistors. Bias is arranged so that collector current is low, when no audio signal is present. The emitter bias resistor R4 limits peak collector current.

If R3 is a pre-set resistor, as in Fig. 12, operating conditions can be arranged to suit the transistors fitted. For best results TR1 and TR2 should be a matched pair. This means that they will have similar gain and current characteristics. However, two transistors of the same type number will generally be satisfactory.

It is assumed that the same battery will supply earlier stages of the amplifier, and current for this purpose is taken from C1 and R1. R1 and C1 prevent audio signals present in the positive line passing back to earlier amplifier stages. The by-pass capacitor C2 also helps in this respect. With simple circuits not giving much overall amplification, or small power outputs, R1, C1 and C2 might all be omitted.

TR1 and TR2 are NPN transistors. If PNP transistors are to be tried, battery polarity must be reversed, and also the polarity of C1 and C2. PNP transistors, with a positive earth line, could be convenient where earlier stages also use PNP transistors and a positive earth.

When first using the circuit, a meter can be placed in one battery lead. R3 is set for minimum value. Battery current will then be low, but reproduction will be very distorted. Slowly increase the value at R3 until reproduction is at its best. This will cause a rise in battery current, so do not set R3 to an unnecessarily high value.

The driver transformer T1 can be of about 5:1 ratio. T2 can be a miniature type for low power, or a 500 milliwatt or 1 watt component for larger power. Its ratio can generally be about 8:1, for use with a 3 ohm speaker. T1 must have a centre tap on its secondary, and T2 a centre tapped primary.

Many transistors other than those given below can be used, the following being typical for negative ground or positive ground.

Negative earth (NPN transistors). 2 x BC108. R1 1.2k. R3 100 ohm fixed or 250 ohm pre-set. R4 4.7 ohm.

Positive earth (PNP transistors). 2 x AC128 or 2 x AC142. R2 4.7k. R3 82 ohm or 250 ohm pre-set. R4 4.7 ohm.

R1 will generally be 1k to 2.2k, depending on the earlier amplifier. The primary of T1 can be connected to terminals P in Fig. 10; or alternatively to any audio source where a reasonably strong signal is present.

Where signals are too strong, an audio gain control needs to be fitted. In Fig. 8 this can be done by omitting R1, and disconnecting C2 from D1. Connect the outer tags of a 100k volume control or log pot from D1 to earth line. Take C2 to the centre tag or slider of the potentiometer. Any required audio level can then be taken off by adjusting the control knob.

Fig. 12B shows the board layout.

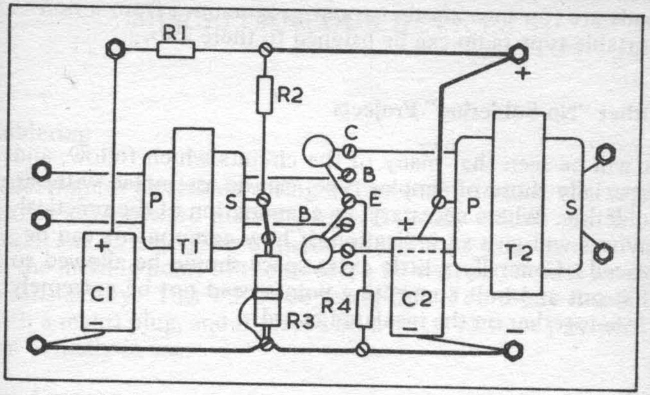

FIG. 12 B.

Boost Speaker

For best reproduction, a speaker needs to be enclosed in a cabinet or fitted to a baffle board. A cabinet can be made from wood, and can be about 6 x 10 inch and 4 inch deep, or as required for the speaker unit. Cut an aperture in the cabinet front to match the cone diameter, and glue thin fabric behind this opening. The speaker can be fixed with nuts and bolts, or short woodscrews driven in from behind.

Connect flexible leads, with separate plugs or a jack plug, to suit the equipment.

Small pocket receivers, which have a miniature speaker, can give much improved volume and reproduction when used with a larger speaker. If a headphone socket, operated from the receiver output stage, is present, it is only necessary to plug the speaker into this. A fairly large speaker, of sensitive type, will give best results. A unit 5 inch to 6 inch in diameter, or about 4 x 7 inch, will be convenient. The speaker has to be of approximately the correct impedance, which should be found in the receiver operating instructions. Where the speaker impedance is wrong, a matching transformer has to be employed between receiver and speaker.

Extension speakers are less frequently used today, but if two leads are run into another room programmes from a non-portable type radio can be listened to there also.

Other "No Soldering" Projects

It will be seen that many of the circuits which follow, and especially those of simpler type, can be assembled without soldering. Where necessary, an examination of the practical layouts will give an indication of how components can be placed. Generally a little extra space should be allowed so that nut and bolt connecting points need not be extremely close together on the insulated board.

SECTION 2
MISCELLANEOUS DEVICES

Soldering

Soldering for electronic wiring will be found to be a very easy and straightforward matter. An electrically heated iron is usual, and as large areas of metal do not have to be raised to the melting point of the solder, a 20 watt to 25 watt iron is satisfactory. This will be perfectly safe if correctly fitted with a mains plug, and if placed on a metal plate or tray when not actually in use.

It is best to use cored solder, as made for electronic wiring.

Leads, tags and other items to be joined should be clean and bright. They have to be raised in temperature to the melting point of the solder at the actual points where a joint is to be made, and the active core of the solder, or flux, must be present here when the joint is being made.

The iron should be tinned – that is, have melted solder on its tip – and have reached its correct temperature. Leads can be hooked or looped into position. The joint is then heated with the iron, and the solder is simultaneously applied to the joint, so that it melts and flows over the surfaces to be soldered. This usually takes place in a second or two. The solder and iron should be removed at once, as excessive heating may damage some components, and especially transistors or other semiconductors. The joint cools in a few seconds, and should be rigid, and solder having flown onto the items to be joined.

A probable cause of poor joints is insufficient heating of items to be soldered, or carrying solder to the joint on the iron, so that the activity of the flux is wasted.

Electronic items such as tags and the wire ends of components are generally designed for easy soldering. They will be copper, tinned copper, and similar metals and alloys which solder readily. For this reason, difficulties are unusual.

Some metals do not solder readily, or need special methods. Aluminium is one of these, and it is not used where popular cored solders intended for copper or general wiring are employed.

Make sure that heat from the iron does not damage furniture or other items, and keep the flexible cord away from the heated tip. Always switch the iron off at once when the work in progress is to be left.

One Transistor Audio Oscillator

This can be used instead of a bell or buzzer warning, or for code practice and similar purposes. It needs very few components, Fig. 13, and can be made to work satisfactorily with a wide range of transistors.

In Fig. 13, an NPN transistor is shown in the theoretical circuit, so battery polarity is as indicated, with negative to emitter. If a PNP transistor is fitted, battery polarity must be reversed.

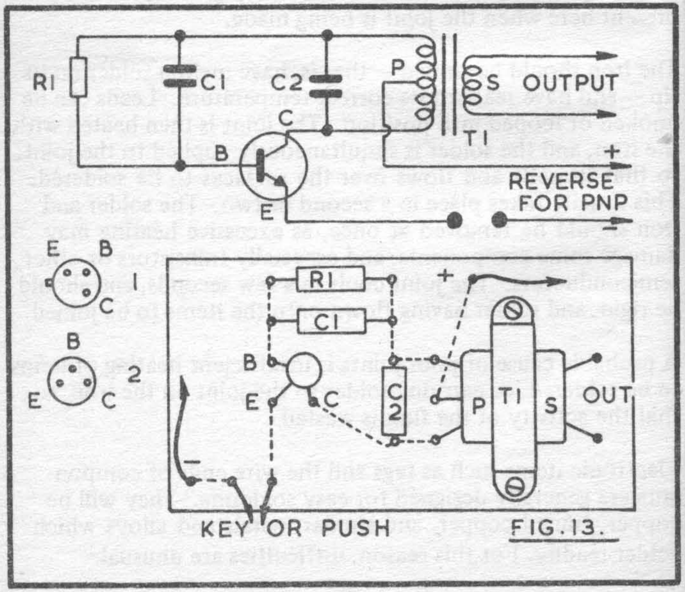

FIG.13.

(1) shows leads from a typical NPN transistor, such as the BC108, 2N706, and similar types. (2) is the base of the PNP AC128 transistor.

T1 can be a small speaker matching transformer, as used for push-pull stages. A centre tapped primary is necessary. The secondary S can be taken to a loudspeaker or headphones.

The values of R1, C1 and C2 can be varied considerably. C1 can be 0.1μF or so. R1 determines to some extent the power obtained, and can be about 22k, increased to 47k or 100k for lower power. The value of C2 has considerable influence on the note, or frequency of oscillation. Increasing the value of C2 lowers the pitch. C2 can be about 47nF to 4.7nF or thereabouts. A 9v battery can be used. Changes to the supply voltage will influence the note produced.

Fig. 13 also shows a layout for components, using either a fully perforated board, or a panel drilled to suit. Transformer T1 is fixed by screws. Small transformers of this type are usually wire ended, or have pins which pass down through holes in the board.

The wire ends of the resistors and capacitors are passed down through suitable holes, and bent over. The board is then turned so that the underside wiring can be carried out. In most places the wire ends of the components will be long enough to reach connecting points. Elsewhere, 24swg tinned copper wire, or similar connecting wire, will be used. Make the soldered joints as described, and snip off unwanted ends of wire. It is as well to leave the transistor to last, as this avoids heating it unnecessarily when soldering other items. Lengths of flex, anchored at holes in the board, are used for battery, key (or push) and speaker connections.

Multivibrator

Fig. 14 is the circuit of the multivibrator assembled in Fig. 2. Reference to Fig. 14B will show how this can also be prepared on a tag board.

The tag board has two rows of tags fitted to an insulated board, and it is five tags long in Fig. 14B. Boards of this type can be obtained in many sizes. It is also easy to cut a section from a long tag board, to suit the needs of the item to be assembled. Such boards are very handy indeed for electronic assemblies, as the tags support resistors, capacitors and other items, and a neat layout is readily achieved.

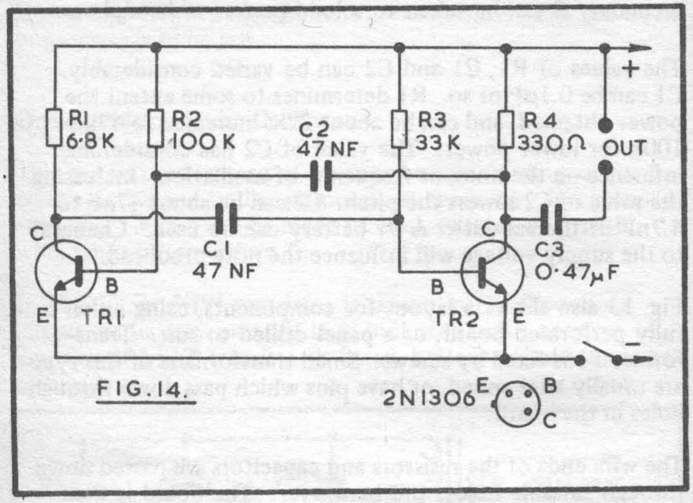

FIG. 14.

Tag strips are also useful for projects, and may have any number of tags required. Long tag strips can be cut down as necessary. Short tag strips have one tag for mounting the strip, longer strips having two tags formed as brackets for this purpose.

Place the components approximately as in Fig. 14B, and solder the wires to the tags. If any component values may be changed, do not twist the leads tightly round the tags before soldering, or removal will be difficult.

Wires which cross can be kept clear of each other, and then no insulated sleeving will be necessary. The transistors can be soldered on last, and can readily be changed without disturbing other parts.

The same layout can be used for the next project, though the component values are different.

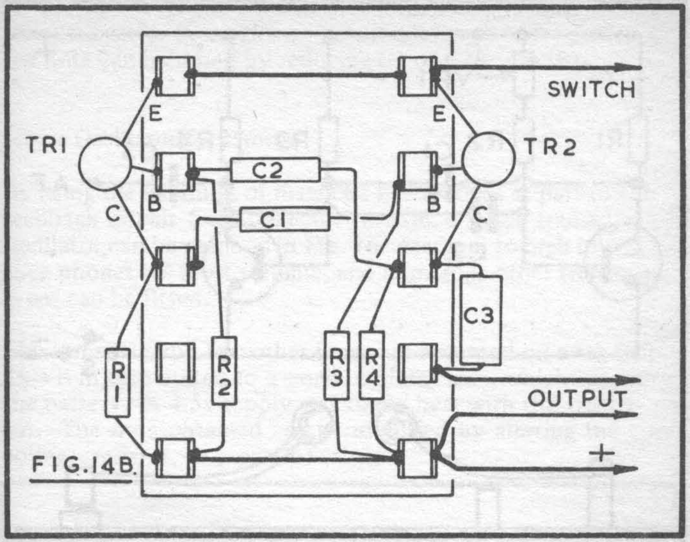

FIG. 14B.

Glissade

This device will produce swooping, rising and falling tones, for "effects" purposes, or can with additional skill be used to obtain musical notes through several octaves. The circuit is shown in Fig. 15, and reference can be made to Fig. 14B for a practical layout on a tag board.

In Fig. 15, suitable component values are:

R1	10k	C1	22nF
R2	12k	C2	22nF
R3	100k	C3	22nF
R4	10k	VR1	100k linear pot.

Transistors: 2 x BC108 or similar types.

The circuit is a wide frequency audio oscillator, in which the pitch can be controlled by VR1, and "on" switching is obtained by S. The oscillator is placed in a box, so that switch S and VR1 can be mounted on the top. The switch is a strip of springy metal, fitted with a spacer and bolt, so that it is just clear of a further bolt. A small downwards movement of the arm or knob completes the circuit.

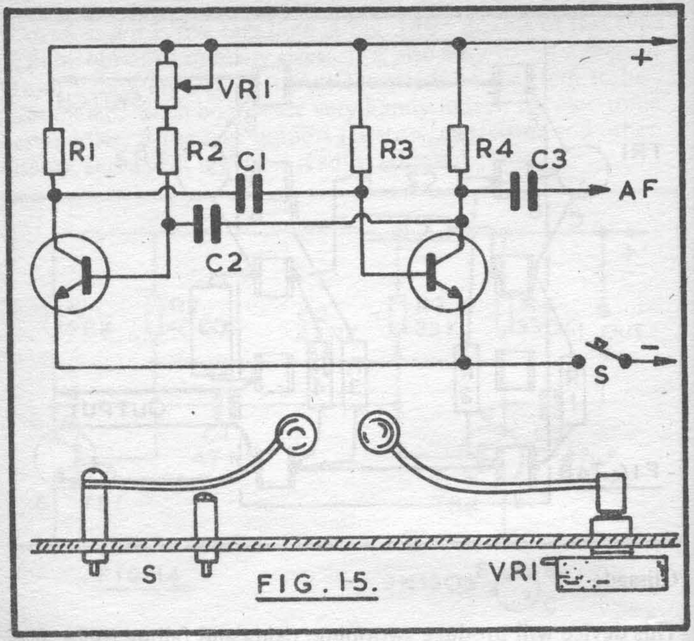

FIG. 15.

VR1 has a ¼ inch shaft coupler, into which is screwed a length of 4ba threaded rod. A knob is also fitted to the end of this rod. It can be swung backwards or forwards, to change the resistance of VR1. Connect VR1 so that movement of the knob towards the right raises the pitch.

Output from C3 can go to high resistance headphones, for amusement, but should be taken to an audio amplifier. Gain of the amplifier will have to be turned back somewhat, as the output from the glissade is at a quite high level.

Swooping tones are produced by holding down the switch S and swinging the control knob attached to VR1 from side to side. For notes which do not swoop to frequency, switch S must be released while VR1 is rotated. A slight quivering motion of the hand controlling VR1 will produce a vibrate effect. It is useful to have an ordinary switch in parallel with the knob control S, and to have the amplifier audio gain control near the left hand, so that fading effects can be produced as well.

About 4.5v to 9v can be used, this influencing volume. Should other transistor types give an unsuitable range of frequencies, the note can be raised by reducing C1 or C2 (or both).

Morse Oscillator for Phones

By using the windings of magnetic headphones as part of the feedback circuit from collector to base, a single transistor oscillator can be made as in Fig. 16. Medium to high impedance phones are most suitable, and numerous other transistor types can be fitted.

The transistor and few other items are mounted on a tag strip. This is in turn bolted to a small insulated box, which can carry the battery. A 4.5v supply was found best with this oscillator. The note obtained can be modified by altering the voltage, resistor, or capacitor values.

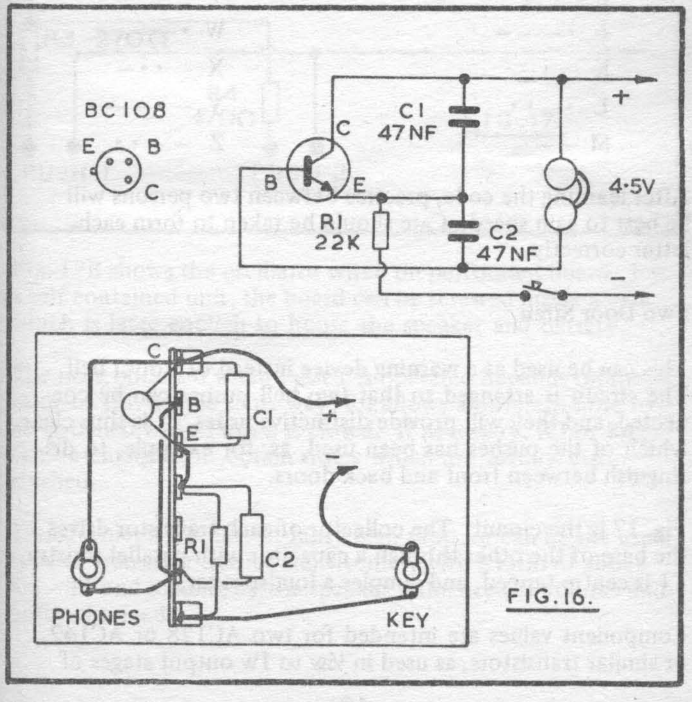

FIG.16.

For serious code practice a good Morse key is required. It is adjusted so that only a small movement of the wrist is needed to operate it. Each letter consists of short and long sounds, shown as (•) and (–). The (–) has three times the length of the (•)

A •–
B –•••
C –•–•
D –••
E •
F ••–•
G ––•
H ••••
I ••
J •–––
K –•–
L •–••
M ––

N –•
O –––
P •––•
Q ––•–
R •–•
S •••
T –
U ••–
V •••–
W •––
X –••–
Y –•––
Z ––••

After learning the code, practice between two persons will be best to gain speed. Care should be taken to form each letter correctly.

Two Door Siren

This can be used as a warning device instead of a door bell. The circuit is arranged so that two bell pushes can be connected, and they will provide distinctive notes. It is thus clear which of the pushes has been used, as, for example, to distinguish between front and back doors.

Fig. 17 is the circuit. The collector of each transistor drives the base of the other through a capacitor with parallel resistor. T1 is centre tapped, and couples a loudspeaker.

Component values are intended for two AC128 or AC142, or similar transistors, as used in ½w to 1w output stages of

receivers and small amplifiers. Other transistors will of course operate in this circuit. The pitch obtained can be modified by changing C3. The circuit shown is for PNP transistors, and battery polarity has to be reversed for NPN transistors.

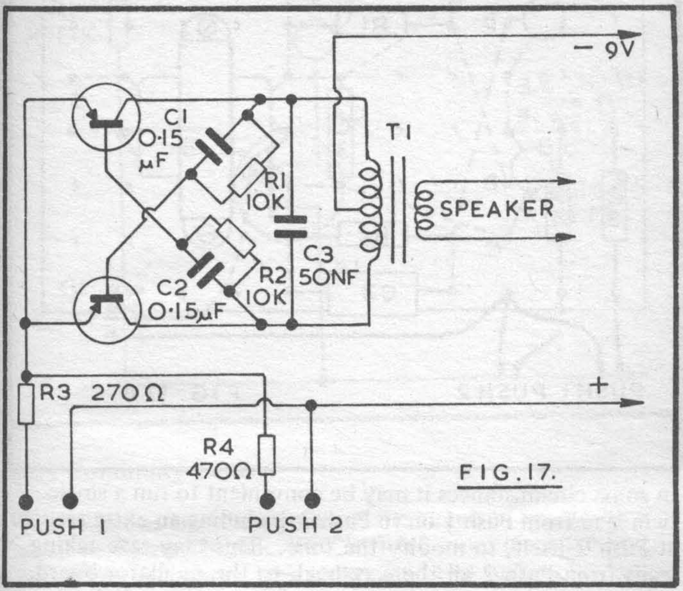

FIG.17.

Fig. 17B shows the oscillator wired on perforated board. For a self contained unit, the board can be screwed inside a case which is large enough to house the speaker and battery.

The note obtained when Push 1 is operated depends on the value of R3. In the same way, R4 allows a different note to be obtained when Push 2 is closed. If necessary, R3 or R4 can be changed, to obtain notes which are readily distinguished.

Transformer T1 is best of fairly high ratio, or can be the usual type of transistor output transformer, with a ratio of about 8:1. T1 and loading by the speaker influence the volume and notes obtained.

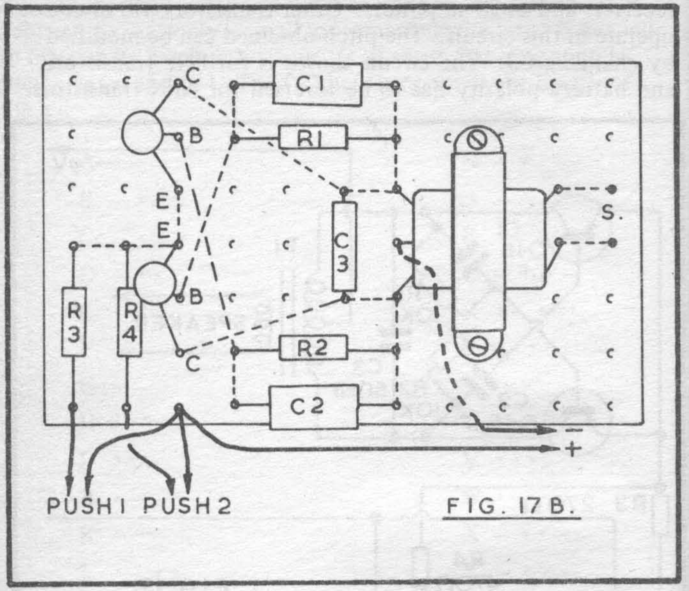

FIG. 17 B.

In some circumstances it may be convenient to run a single twin lead from Push 1 on to Push 2, including an extra resistor at Push 2 itself, to modify the tone. This may save taking leads from Push 2 all the way back to the oscillator board.

UJT Oscillator

The uni-junction transistor is a device having emitter, Base 1, and Base 2, Fig. 18. B2 receives positive bias, in this case through the headphones. When forward bias on the emitter E is reached the UJT conducts.

In Fig. 18, the emitter potential rises as C1 charges through the resistor R1. When the UJT conducts, C1 is discharged, and the process is repeated. This can take place over a very wide range of frequencies, depending on the values used for R1 and C1. Here, these are chosen to produce an audio tone.

A battery supply of 3v to 9v will be convenient. For use of the UJT oscillator during code practice, place the key in one battery lead.

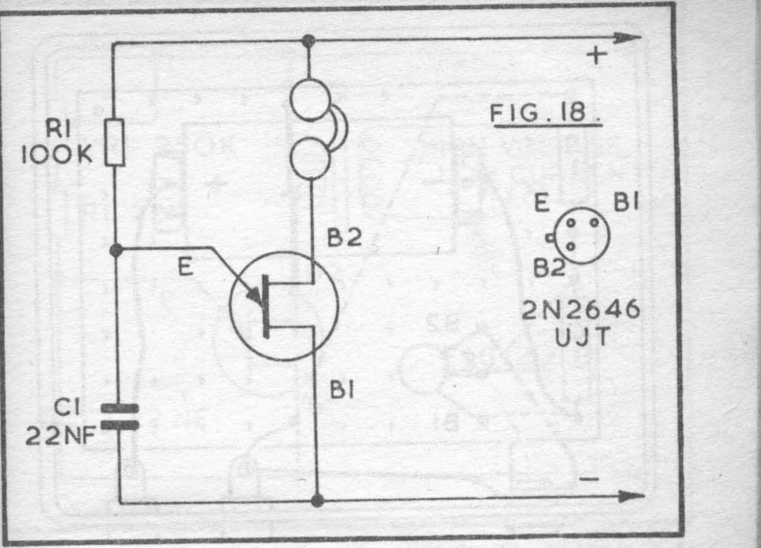

FIG. 18.

UJT Continuity Tester

Fig. 19 shows actual assembly of the circuit in Fig. 18 on a small perforated board. This is fitted in an insulated box, which carries a phone jack socket, and two sockets for Morse key or test leads. A 9v battery is used, secured with string or elastic.

For continuity tests, take prods or clips on the test leads to the suspected item. If the circuit is in order, the UJT oscillator will produce an audio tone. Oscillation will be obtained with a series resistance between the prods of up to 20k or more, so various components can also be tested, and some idea of the resistance in circuit can be gained from the sound obtained. The tester can also be used to check bulbs, lamps, fuses, extension and other leads, and similar items – these must of course be removed or disconnected from the equipment, before being tested.

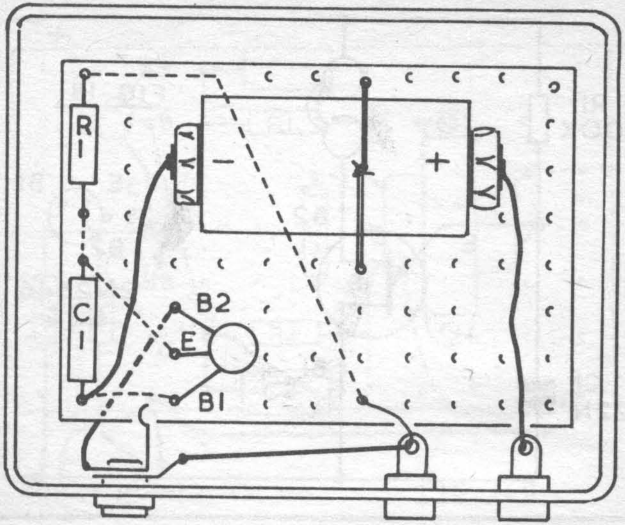

PHONES FIG. 19. KEY OR TEST LEADS

UJT High Voltage Producer

The UJT oscillator can be used to produce a high voltage, at very low current, by connecting a transformer as in Fig. 20. A loudspeaker transformer, as would be used in valve equipment, is suitable, and will have a ratio of about 45:1 or so. The usual secondary, or low-resistance winding, is connected to B2 and positive, so that a step-up of voltage is obtained. This should be found to light a small neon lamp.

VR1 is fitted to allow some adjustment, to find the most suitable frequency of oscillation. For maximum possible high voltage output, if required, operating voltage and current should be raised. No rating of the UJT must be exceeded. For the 2N2646, the supply should not exceed 30v, emitter current should not be over 50mA, and dissipation of power in the device should not be more than 300 milliwatts. (Dissipation can be found from Voltage x Current.)

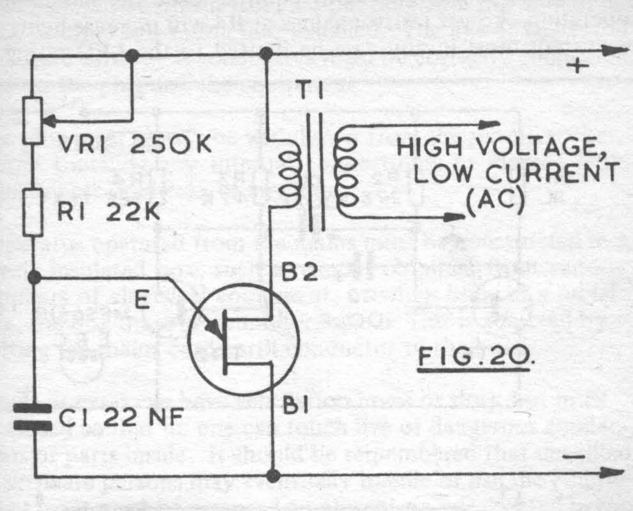

FIG.20.

A small mains transformer can be used for T1, with a low voltage winding, such as that for a 6.3v heater circuit, as primary, and the 200/250v winding as secondary. The high voltage output is alternating current.

Signal Blinker

Fig. 21 is a multivibrator in which the feedback capacitors C1 and C2 are of high value, so that conduction alternates between TR1 and TR2 at a low rate. A light emitting diode is placed in the collector circuit of TR2, and so flashes on and off.

This device can be used for various instruments and other equipment, and a flashing indicator draws attention to a greater extent than does one continuously illuminated. The LED can be located at any required point, with leads running to it.

Transistors other than the types shown can be utilised. R4 can usually be about 150 ohm for a 3v supply, 470 ohm to 1.2k for 6v, or 680 ohm to 1.5k for 9v. This depends to some extent on the amount of light required when the indicator is operating. Lower resistor values at R4 will increase light, but current must in any case be limited to the LED rating.

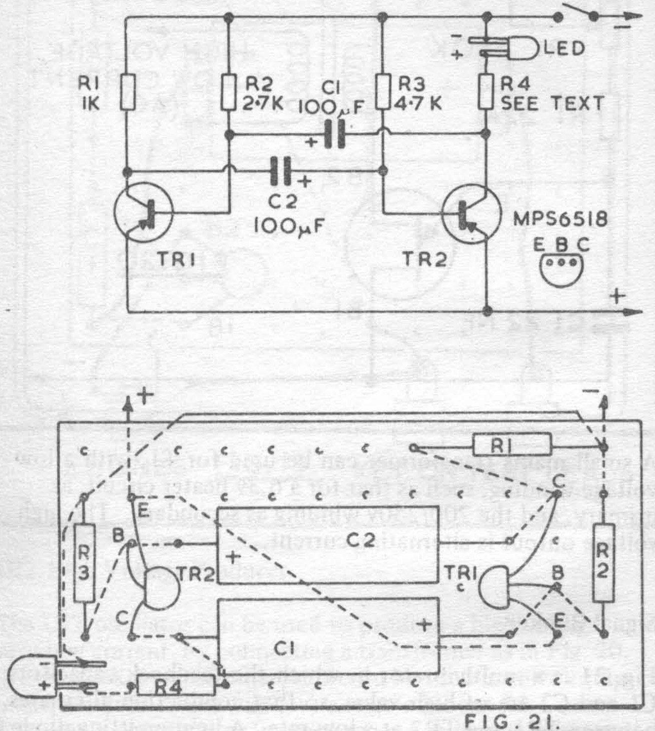

FIG. 21.

C1 and C2 need only be 6.4v working, and the device can be assembled on a very small perforated board. See p.61 for LED polarity.

Mains Wiring

All the devices described so far have operated from low voltage batteries, so than no shock hazard exists with them. However, it is necessary to employ mains power for some units, and proper care must then be taken to avoid any danger.

Power should be drawn from a 3-pin plug, and this should have a 3A or other low rating fuse installed. The live or L, neutral or N, and earth or E conductors must be correctly connected, both at the plug and the equipment.

The plug must always be withdrawn from the supply socket, before touching any internal connections, or making any adjustments or checks of any kind.

Apparatus operated from the mains must be constructed in a strong insulated box, such as can be obtained from various suppliers of electrical equipment; or when built in a metal box, the box must be soundly earthed. This is achieved by bolting the mains cord earth conductor to the box.

Boxes or cases can have ventilation holes or slots, but must be closed so that no one can touch live or dangerous connections or parts inside. It should be remembered that unskilled or unaware persons may eventually handle or use the equipment, so it must be arranged to present no unexpected hazards such as live terminals or parts, or openings for the fingers.

If these points are remembered, mains operated equipment can be safe to use and handle. No switch or other means of interrupting the circuit is put in the earth conductor, and any switch or equivalent device is placed in the live or L conductor. This leaves both earth and neutral permanently connected, for maximum safety.

Dimmer/Heat Control

It is often convenient to dim a filament lamp, or reduce the heat of a soldering iron to be left on for long periods, or the power of a low wattage heater. The silicon controlled rectifier circuit in Fig. 22 can be used for these purposes.

When the switch marked "Full" is closed, current for the load passes through this, so that full power is obtained.

When this switch is opened, current passes through the SCR. This device conducts only when the gate has been triggered, and the instantaneous point in the alternating current cycle where this arises is determined by the setting of the potentiometer VR1. Triggering is advanced as the wiper of VR1 is moved towards R1. The average power supplied to the load can thus be adjusted between wide limits. Since the device is operating as a half wave controller, full power is not obtained with VR1 at maximum, so the switch is closed as described.

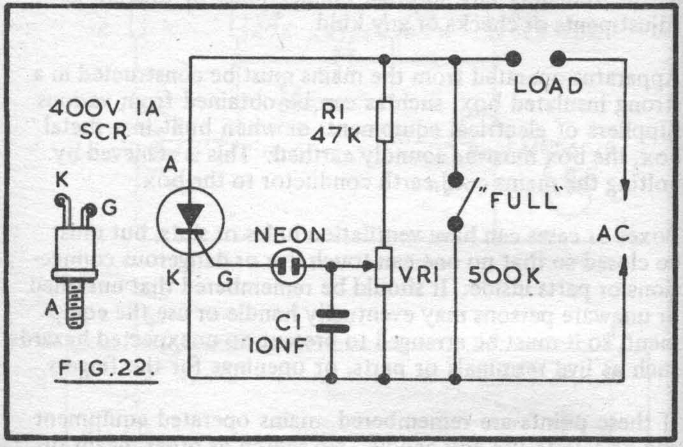

FIG.22.

Fig. 22B shows construction of the SCR controller in a metal box. The front of this carries a socket outlet, and the mains cord has a 3-pin plug, wired as explained earlier. It is thus easy to plug the controller into an outlet, and insert the plug of the lamp or other item to be used in the new outlet.

The mains cord is secured with a clamp, and passes through an insulated grommet. The tag strip supporting the small items is mounted by an unused tag. A strip of metal is cut to clamp round the body of the SCR, and is soldered to one tag, as shown. Run earth and neutral leads from the cord up through a hole, and connect them to the E and N terminals of the socket outlet. Also earth the case, at E. A well insulated wire runs from the L terminal of the outlet, to the "Full" switch. The outlet can be secured with bolts and nuts.

R1 should be 1 watt, and C1 is a 500v capacitor. The neon is a small type as used for indicators, etc.

After checking that the controller operates correctly, the back of the box should be screwed in place. A suitable size is about 4 x 5 x 2 inch deep.

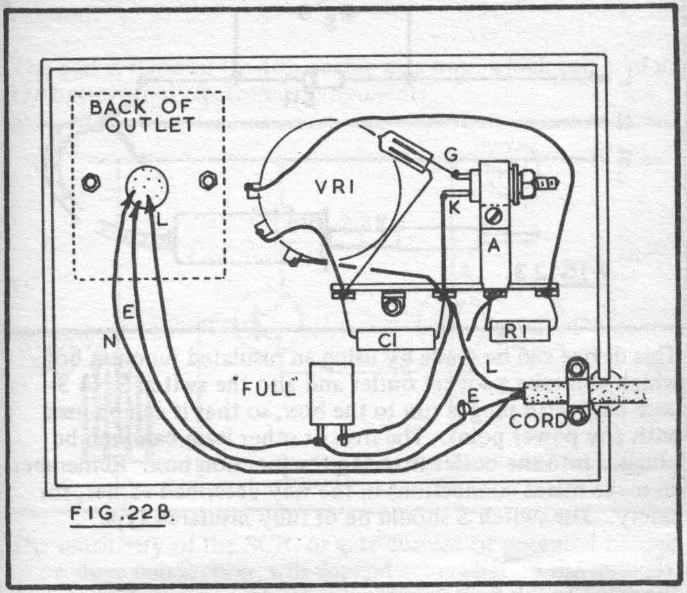

FIG.22B.

If a fully insulated box is used instead, then the bush of VR1 and of the switch should both be earthed by wires, except when these items are constructed wholly with insulated material. (This precaution is necessary to avoid any chance of metal parts outside the box becoming "live".)

Two Power

A circuit which will allow half or full power can be useful to reduce the heat of a soldering iron, or for dimming a filament lamp.

Fig. 23 is an easy circuit of this type. Full power is obtained when the switch S is closed. With S open, the silicon rectifier allows current to flow in one direction only, so that power is approximately halved. A 600v 1A or 1000v 1A rectifier is suitable.

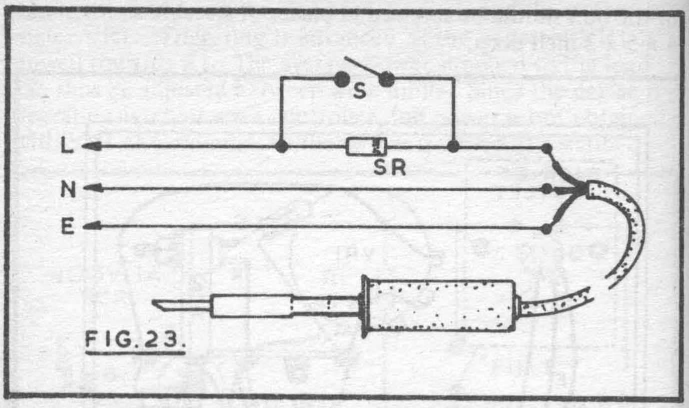

FIG. 23.

This device can be made by using an insulated junction box which will take a socket outlet and also the switch S. A 3-core cord with plug is run to the box, so that it can be used with any power point. The iron or other item can then be plugged into the outlet fitted to the junction box. Remember to make mains connections in the way described earlier, for safety. The switch S should be of fully insulated type.

Sensitive Touch Switch

In Fig. 24, finger contact with the touch pad switches on a low voltage bulb. This circuit is suitable for a battery operated portable bedside light, and for some other devices.

Normally the silicon controlled rectifier is not conducting, so the lamp is extinguished. Finger contact with the pad provides positive base bias for the transistor so that it conducts. This moves the transistor emitter and SCR gate positive, triggering the SCR. Current then flows from cathode K to anode A, and through the bulb. The SCR remains in the conducting condition until the battery circuit is interrupted by opening the switch. A push switch, normally closed, is most suitable here. Pressing it then switches off the bulb.

This touch pad is easily made from a piece of circuit board with foil strips. It can be about 2 x 1 inch, or of any convenient size. Join together alternate foils – e.g., 1, 3, 5 and 7 – by soldering a connecting wire through the holes. This will provide one connection to the pad. The other foils – 2, 4 and 6 – are similarly joined together for the other pad connection.

The pad is fixed to the top or side of a box, which can contain the battery and electronic components.

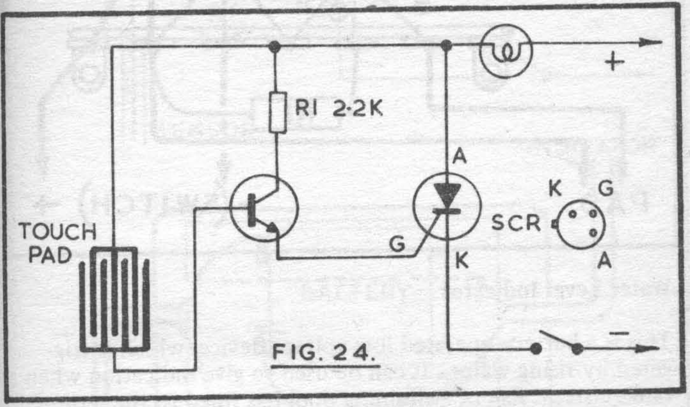

FIG. 24.

The sensitivity of the SCR, or gate current or potential needed to produce conduction, will depend somewhat on the individual device, while the gain of individual transistors will also be different. Numerous NPN types are suitable. If it is found that the switch is too sensitive, this can be corrected by placing a resistor from transistor base to SCR cathode or negative line. This resistor can be about 470k to 1 megohm or so. Excessive sensitivity will be apparent if the circuit is switched on by dust, damp air, or similar effects. The pad should be kept clean by wiping it with a dry cloth.

Fig. 24B shows assembly of the components on a tag strip. A small 50v 1A SCR is more than adequate. The bulb can be a screw in type, and can be 6.3v 0.15A, or 6.3v 0.3A. A 6v 3W (0.5A) bulb is also suitable, provided the battery is fairly large.

It is convenient to make a wooden box which will take the battery, and leave a little space for the tag strip assembly.

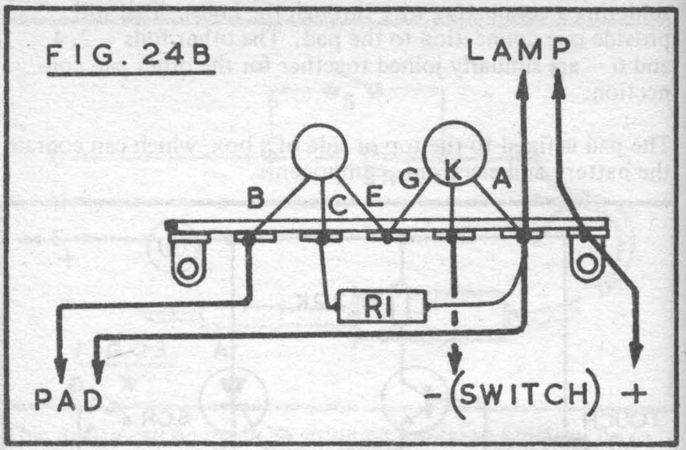

FIG. 24B

Water Level Indicator

This is a battery operated low voltage device, which is triggered by rising water. It can be used to give indication when a tank, cistern, fish or swimming pool has filled to the required level.

Fig. 25 shows the circuit, and also assembly on an insulated board. This also carries the battery, and it is fitted in a metal or plastic box.

In use, the base normally has no positive bias, as the sensor passes no current. When water reaches the sensor, it conducts, and the transistor passes collector current, to operate the indicator. R1 limits base current.

This circuit will operate satisfactorily with a wide range of PNP or NPN transistors, but battery polarity must be reversed for the PNP types.

The indicating or warning device can be one of the multivibrator or other audio oscillators shown earlier. These will give an audible warning, and will draw only a small current, so that a large transistor is not required.

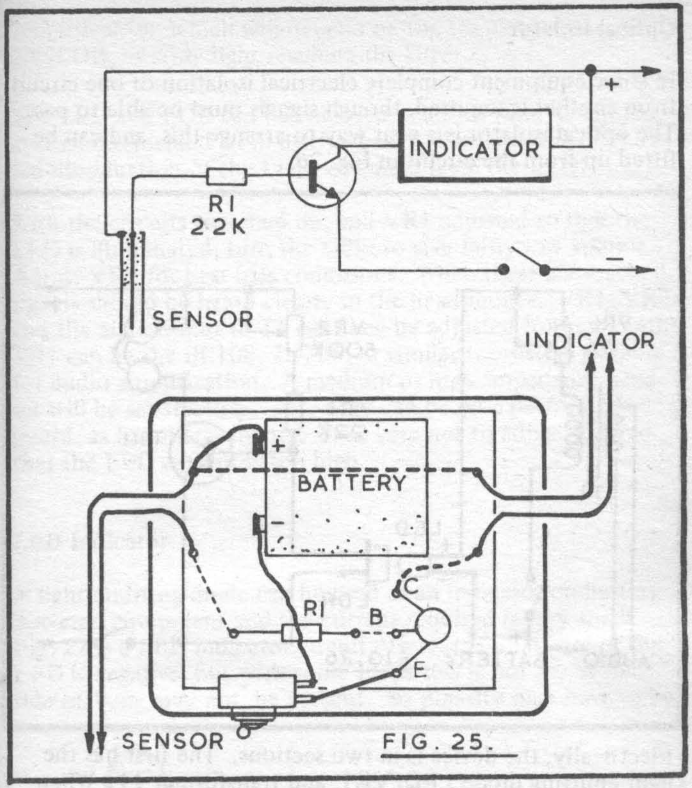

FIG. 25.

For indoor use or in a location protected from rain, the sensor can be two bare wires, about ¼ inch apart, bolted to a strip of insulated material. Where rain could operate the device, it needs to be covered. Alternatively, it can be made from a piece of metal tube, with a central wire, as in Fig. 25. The wire is bound with insulating tape for about half its length, to fit tightly in the tube. Flexible leads are soldered on, and the joints covered with tape. The sensor can be suspended inside the tank or pool, with its bottom at the required water level.

If the transistor is a high gain one, so that damp operates the indicator, sensitivity can be reduced by connecting a 100k resistor from base to emitter.

Optical Isolator

In some equipment complete electrical isolation of one circuit from another is required, though signals must be able to pass. The optical isolator is a neat way to arrange this, and can be fitted up from the circuit in Fig. 26.

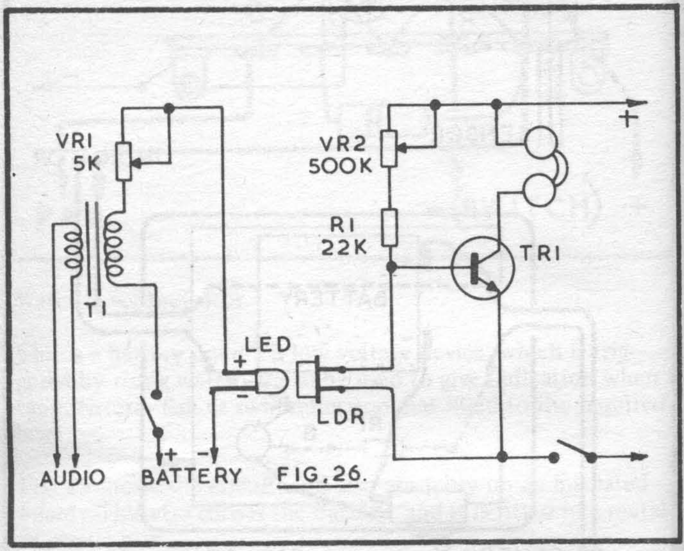

FIG. 26.

Electrically, the device is in two sections. The first has the light emitting diode LED, VR1, and transformer T1. When the battery is switched on, current flowing through the LED is adjusted by VR1, so that the LED glows dimly. A signal voltage in the secondary of T1 then modulates the LED light intensity.

To operate this circuit from a radio having a 2/3 ohm or similar speaker, a speaker matching transformer with a ratio of about 45:1 can be used. (The actual ratio is not very important.) Connect the usual transformer secondary to the radio speaker, so that T1 provides a step up effect.

The second part of the equipment has the light dependent resistor LDR, forming part of the network providing TR1 with base bias. VR2 allows adjustment of conditions for best amplification, which will depend on the transistor, and also the LDR, or stray light reaching the latter.

The LED and LDR can be enclosed in a card tube, or should at least be shielded from light, which will swamp the modulated illumination of the LDR, obtained from the LED.

With the circuits switched on, and VR1 adjusted so that the LED is illuminated, turn the radio to give fairly low volume. Adjust VR2 for best bias conditions. When these are reached, signals should be heard clearly in the headphones. VR1, VR2 and the audio input to T1 can then be adjusted for best results. TR1 can be the BC108, 2N706, or similar transistors suitable for audio amplification. A medium or high impedance headset will be satisfactory. Assembly can be on a perforated board, as for other circuits. Take care not to adjust VR1 so that the LED current is too high.

LED Indicator

A light emitting diode can be used as an indicator on battery powered equipment and the current required is very small. Fig. 27 is a LED indicator circuit. Here, the long wire of the LED is negative, but with some LEDs this is not so. A flat side may, or may not, be present. So polarity may have to be

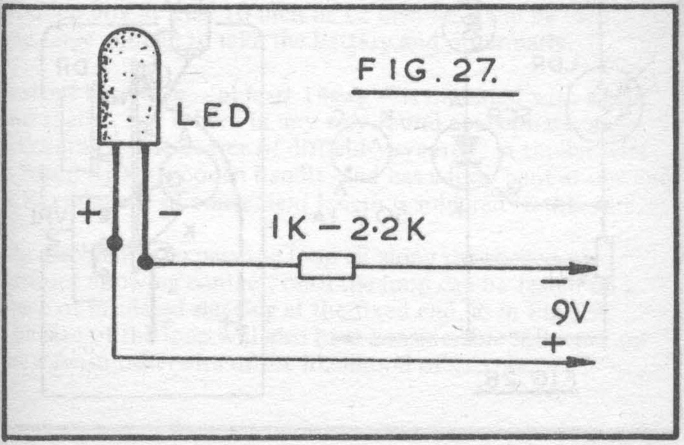

FIG. 27.

found by trial – the correct connections are those allowing the LED to light. The series resistor may be about 150 ohm to 1k or more, depending on the actual voltage, and brightness required. Current drawn varies with different LEDs, and with voltage and series resistance, but can be about 5mA or less, without reducing the brightness too far.

The LED indicator with resistor is connected to the circuit where voltage is present when the equipment is on. LEDs are available in red, yellow, green and other colours, with grommets or clips for panel mounting. Different circuits may be identified by a range of colours.

"Magic Candle"

The triggering of a silicon controlled rectifier can be used for slave photoflash and similar purposes. In Fig. 28 it is employed to switch on a light when a match or lighter is brought near. Sensitivity is best in somewhat subdued light, and it will then be found that striking a match at some distance can bring on the light.

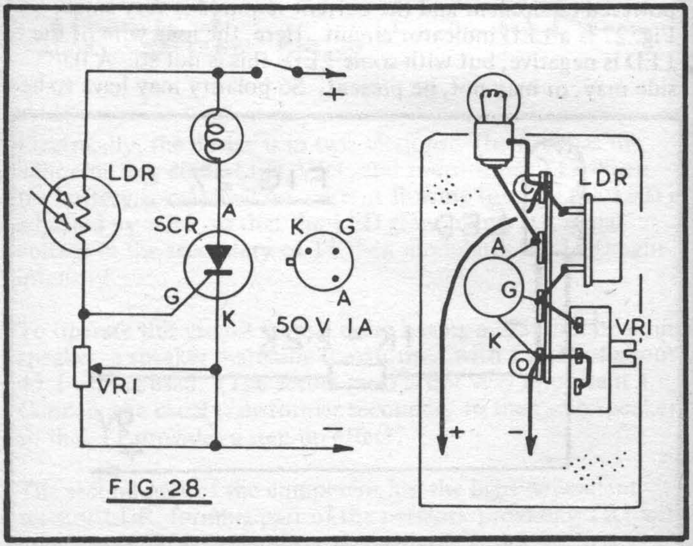

FIG.28.

The LDR has a high resistance when in darkness or subdued light, so that the gate potential of the SCR is not sufficiently high to trigger this device. When light reaches the LDR, its resistance falls, and the SCR gate voltage rises. The SCR then moves into its conducting condition, so that current flows through the bulb. The sensitivity of the circuit, or level of illumination needed to trigger the SCR, is set by VR1.

Components can be assembled on a small tag strip. This, with a battery, can occupy a tubular case, with the lamp on top. An opening is necessary opposite the LDR, and a hole is also made so that VR1 can be adjusted with a screwdriver. Here, a miniature pre-set 50k or 100k linear potentiometer is most suitable. It is necessary to have a small push switch to interrupt the battery current, as the SCR remains in the conducting condition until the circuit is momentarily broken. A 6V battery and bulb can be used. SCRs may fail to operate at all if attempts are made to use them with extremely low voltages, such as a single 1.5v cell.

"Nerve Tester"

This device tests the nerve, concentration, and steadiness of hand, and always provides amusement.

Connections and construction will be seen from Fig. 29. A wooden box at least 10 inch or 12 inch long will be used, and large enough to take the battery and other parts.

A stout bare wire – at least 14swg – is mounted with a bolt and spacer, and is bent in any way found appropriate, and according to the degree of difficulty wanted. A similar wire is fitted with a wooden handle, and has a loop bent at one end. A flexible lead of convenient length is soldered to this wire.

The problem is to pass the loop all along the shaped wire, without allowing contact, until the loop can be rested on a piece of insulated sleeving at the fixed end, as in Fig. 29. The size of the loop will also have considerable influence on the ease or otherwise of the likelihood of success.

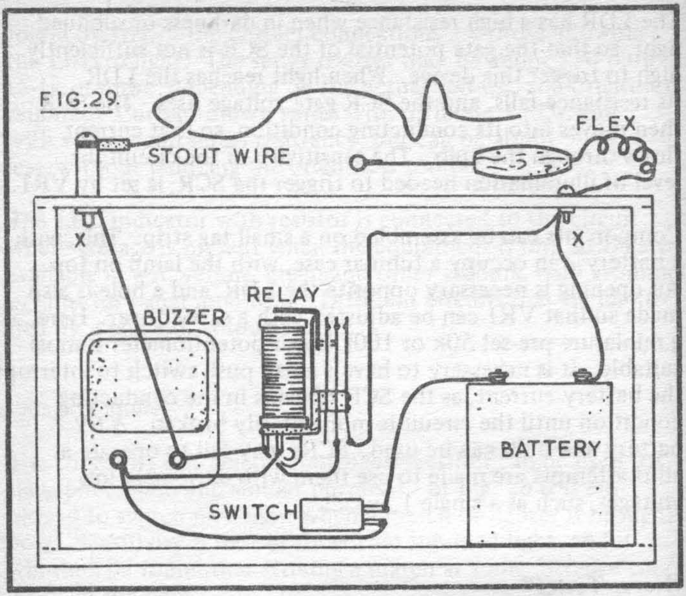

FIG.29.

If loop and wire touch, current flows from the battery to the buzzer, which sounds. Current also flows to the relay winding, so that the relay contacts close. Even a momentary contact between loop and wire will thus cause the buzzer to sound until the switch is used to break the circuit.

Any 4.5v, 6v or similar battery and buzzer will be satisfactory, with a relay which operates on a similar voltage. Connections are to the relay contacts which remain open until the relay is energised.

In use, the device can stand upright. A back is attached with small screws.

Burglar Alarm

The same circuit arrangement as in Fig. 29 will provide a burglar alarm operated by pressure mats. The latter are available in various sizes, and consist of a thin pad which can be placed under a carpet or ordinary mat.

Leads from the mat or mats will run to the terminals at X–X. The buzzer should be changed for a bell, which need not be fitted in the box, but can be situated where it is best heard. Momentary weight on any of the mats will then switch on the bell, which will be locked on by the relay as described, and will ring until the unit is switched off.

Transistor Tester

A very useful check of transistors can be made with the aid of a multi-range test meter, with a selection of base resistors to provide base current, as in Fig. 30. This method does not give an exact indication of transistor gain, but it does allow defective transistors to be eliminated, and provides a good idea of the performance to be obtained from surplus or other unknown transistors.

A 6-way switch allows the base to be open circuit, or to receive current through one of the resistors R1 to R5. These values are R1 330k, R2 110k, R3 33k, R4 11k and R5 3.3k. After allowing for the base-emitter potential, switch positions and corresponding base currents are near the following: 1, zero; 2, 10μA; 3, 30μA; 4, 100μA; 5, 0.3mA, and 6, 1mA.

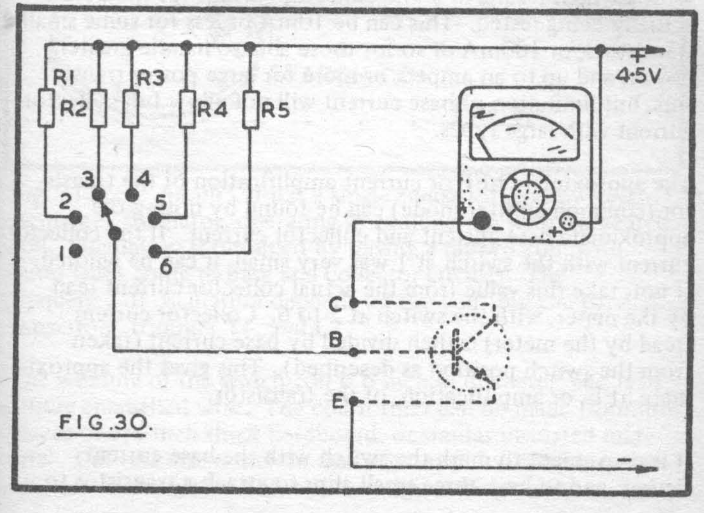

FIG.30.

Polarity is shown for NPN transistors. Reverse polarity of both meter and battery for PNP transistors.

For all tests, initially set the switch to 1, and the meter to its highest range (say 1 ampere). Afterwards, switch the meter to a lower range, after adjustment of the 6-way switch, as may prove to be necessary to obtain an easy reading.

With the switch at 1, a high reading on the meter shows that the transistor has a short circuit from emitter to collector, and is defective.

With germanium transistors, collector current (with the switch at 1) may be 0.1mA or so up to perhaps 0.5mA for small types, and up to some milliamperes for larger types. This small leakage current does not show that the transistor is defective.

In the case of silicon transistors, the leakage current (with switch at 1) will be zero or extremely small, if the device is in good condition.

The switch is then turned to one of the positions 2 to 6. If collector current does not rise, the transistor is defective.
If current rises at each step, do not go beyond the switch setting which results in a safe collector current for the device actually being tested. This can be 10mA or less for some small transistors, or 100mA or so for those able to handle greater power, and up to an ampere or more for large power transistors, but limitation of base current will not allow full collector current with large types.

The approximate hFE or current amplification of the transistor (common emitter mode) can be found by noting the approximate base current and collector current. If the collector current with the switch at 1 was very small, it can be ignored. If not, take this value from the actual collector current read by the meter, with the switch at 2 to 6. Collector current (read by the meter) is then divided by base current (taken from the switch position as described). This gives the approximate hFE, or amplification, of the transistor.

It is convenient to mark the switch with the base current figures, and to have three small clips to attach a transistor to

the circuit. Switch, clips, and battery can all be fitted to a small box. A development of the circuit would be to include a current meter, with switch and shunts, to replace the external general purpose test instrument.

Treasure Locator

This locator is among the simplest that can be constructed, as it uses an ordinary small portable radio as part of the detection circuitry. No modifications or connections to the radio are needed.

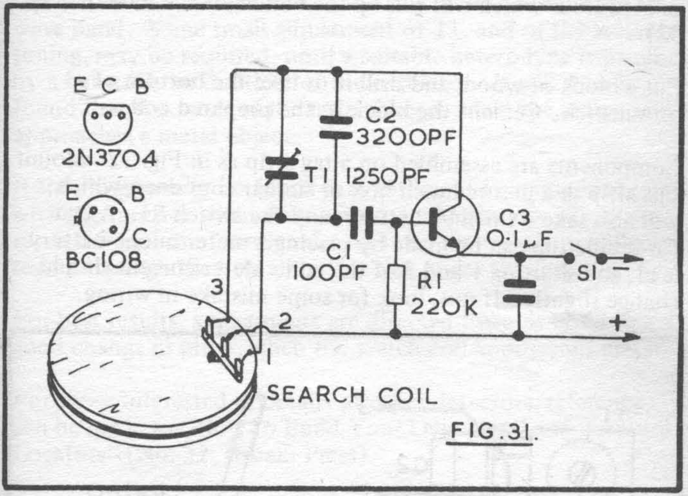

FIG.31.

Fig. 31 is the circuit. It has a single transistor as a low radio-frequency oscillator, operating in conjunction with the search coil. Oscillation is obtained by having a tapped coil, with outer leads going to base and collector circuits. The frequency of oscillation depends on the coil, and the parallel capacitors, trimmer T1, and C2.

The winding of the search coil is 6 inch in diameter, and is of 34swg enamelled wire. The coil former can be made from three pieces of 1/8 inch thick hardboard, or similar insulated material. Cut one disc 6 inch in diameter, and two others about 6½ inch in diameter. Smooth the edges with glasspaper, and cement them securely together, with the small disc between the larger discs. This provides a former with a channel to take the winding. The adhesive should be allowed to dry with the discs under pressure, so that there is no gap into which turns can fall. The former should also be varnished, and allowed to dry, before winding.

The wire is passed up through a small hole, this being point 1. Wind on fifteen turns. Make a loop and pass this up through a hole, for point 2. Wind a further twenty-five turns in the same direction, finishing off at point 3. These leads are scraped and soldered to a small tag strip. This is an anchor point for insulated leads which will run up the handle of the locator (Fig. 32).

Cut a block of wood, and drill it to take the bottom of a broomstick. Cement the block to the prepared coil.

Components are assembled on a tag strip as in Fig. 32. Mount this strip in a plastic lunch box or similar container, which will also take a small 9v battery, and the switch S1. A test for oscillation can be made by placing a meter in one battery lead, and shorting 1 and 3 of the coil. Meter current should change slightly. If not, look for some mistake in wiring.

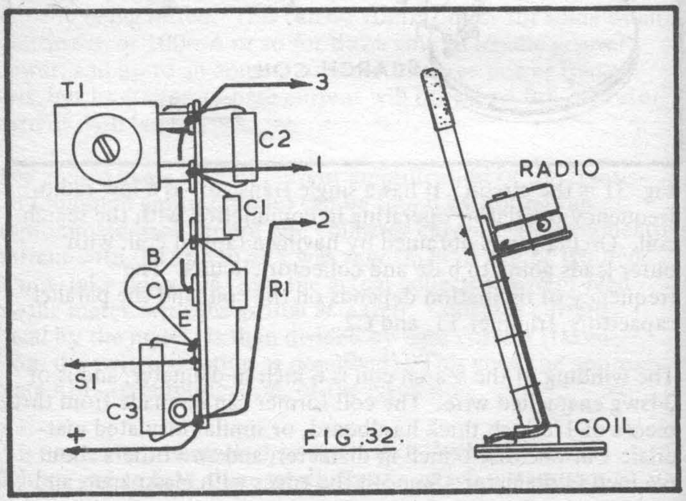

FIG.32.

A shelf is bolted to the handle, and the detector rests on this, with connections passing down to the search coil. A few loops of tape or string will secure these to the handle.

Operation in Great Britain can be on frequencies under 100kHz. Initially tune the radio receiver to 200kHz on long wave, and place it near the coil. Adjust T1 until a strong whistle is heard in the receiver speaker. This is the 2nd harmonic of the oscillator. T1 is then screwed down very slightly from this position.

The small portable radio is placed on top of the locator box, and can be held here with elastic. It is tuned to the medium wave band. Some small adjustment of T1, and of the receiver tuning, may be required, until a suitable heterodyne is produced by a harmonic or multiple of the oscillator. It will then be found that the pitch of the note changes when the search coil approaches a metal object.

It should generally be found that sufficient coupling is obtained with the receiver when they are situated as in Fig. 32. If not, run a few inches of insulated wire from the transistor collector tag up through the box and place this under the receiver.

For best results, adjustments are directed towards obtaining a good change in pitch, when the search coil approaches metal.

For those interested in details of other detectors, reference can be made to "How to Build Your Own Metal and Treasure Locators" (No. 32, Babani Press).

SECTION 3
RADIO AND AUDIO FREQUENCY

MW Booster

This unit can be employed in conjunction with any ordinary portable having a ferrite rod aerial, and will provide a considerable increase in range and sensitivity to weak signals. It has two functions – to allow the use of an external aerial and earth, and to amplify and couple signals to the portable receiver.

Tuning coverage is for medium waves. That is, about 200–550 metres, or 1500–550kHz. Many overseas and other stations can be picked up on this band, and distant reception can be particularly common on the higher frequencies, during the hours of darkness.

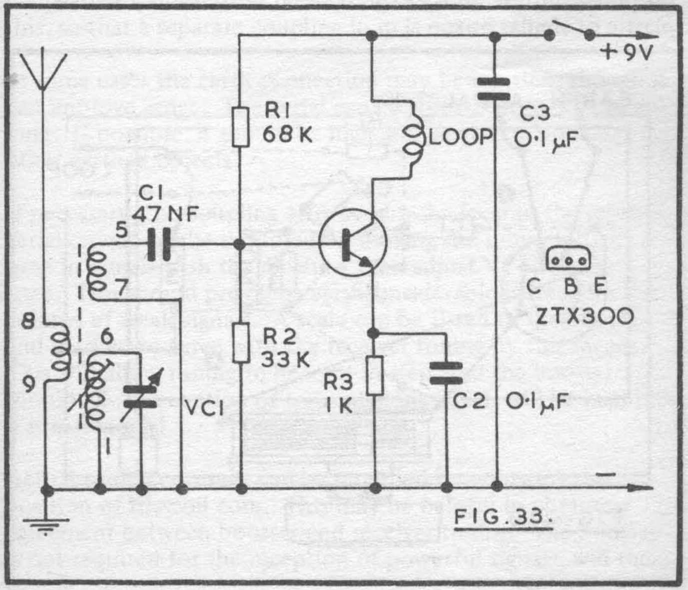

FIG. 33.

Fig. 33 is the circuit, and it uses a cored tuning coil with aerial coupling winding 8–9, tuned winding 6–1, and base coupling coil 5–7. These numbers are for the pins of the Denco (Clacton) "Blue" Range 2 component. It will be apparent that other coils, or surplus or home-wound inductors, can be substituted.

For the coil mentioned, VC1 is 315pF. However, this value is not too critical, provided the capacitor fitted allows tuning the wanted frequencies. A solid dielectric capacitor can be used, though the air spaced type is likely to be a little more efficient.

Coupling to the base is by C1, and resistors R1 and R2 set the base bias conditions, with R3 for emitter bias. These values will suit numerous RF transistors.

Fig. 33B shows assembly of the components on an insulated board. The coil is mounted by its threaded bush. When the board has been wired, it can be mounted by means of bolts with extra nuts. The case is preferably metal, for screening purposes, but the booster can be used satisfactorily with a plastic or similar box.

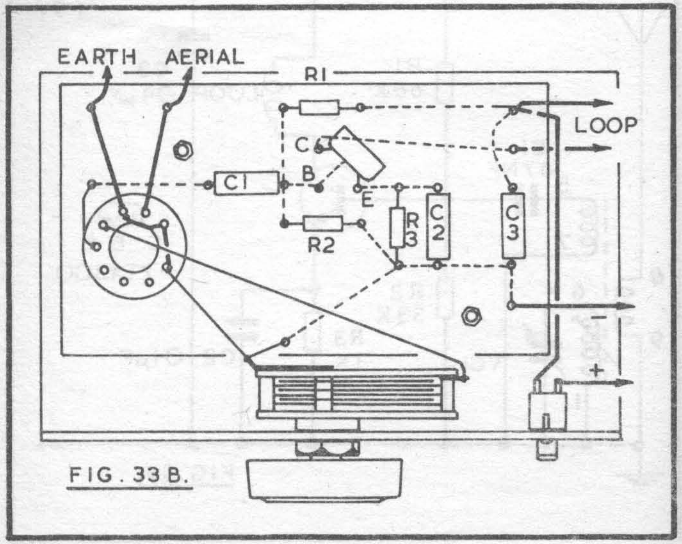

FIG. 33 B.

Fit a small slide switch to the box front, and aerial and earth sockets to the back, near the coil. A 9v battery is accommodated in the case.

Coupling to the transistor receiver is by means of a loop which is placed on the receiver ferrite rod, or near the rod. The loop can be about twenty turns of thin insulated wire, about ½ inch in diameter, and bound with cotton to keep the turns together The ends can be left long enough to reach the two connecting points in Fig. 33B.

If the case is metal, the booster and receiver can be near each other, so the loop leads will only need to be a few inches long. But when an insulated case is used, coupling from the receiver ferrite aerial or other circuits, to the booster coil, may cause instability. This will become apparent as whistles accompanying most signals. If this arises, the booster and receiver should be moved some inches apart. The aerial lead must not run near the loop connections.

Some receivers have an internal winding which is intended to couple an external aerial, and it may be possible to connect to this, so that a separate coupling loop is not required.

In some cases the earth connection may be omitted, though it can improve range. The aerial can be some 20ft to 40ft or so long. If possible, it should be high and clear of walls and other earthed objects.

If necessary, the coupling afforded by the loop to the receiver aerial circuit can be modified by moving the loop. Initially, tune in signals with the receiver, then adjust VC1 for resonance. This should provide a very considerable increase in the volume of weak signals. A scale can be fitted to the booster, and marked to agree with the receiver tuning by this means. This will allow tuning to be kept in step, and the booster should allow reception of transmissions which will be inaudible if it is removed.

Actual tuning coverage can be modified by adjusting the position of the coil core. This may be helpful in obtaining agreement between booster and receiver tuning. The booster is not required for the reception of powerful signals, and the receiver automatic gain control circuit will tend to mask any increase in signal strength obtained with such transmissions.

Midget MW Diode Radio/Tuner

A small, compact tuner can be of utility in allowing radio reception through an amplifier, and a diode receiver can give excellent quality of reception. Such a unit can also be employed alone, with headphones, for personal listening.

Fig. 34 is a receiver/tuner which can be built in a very small plastic or other insulated box. VC1 is a compression type trimmer, fitted with a small knob for adjustment by hand. To do this, it is necessary to remove the short screw (generally 6ba) and replace this with a longer one, which projects to take the knob. The latter can be a terminal head, or the cap from an empty tube of adhesive, secured with lock nuts.

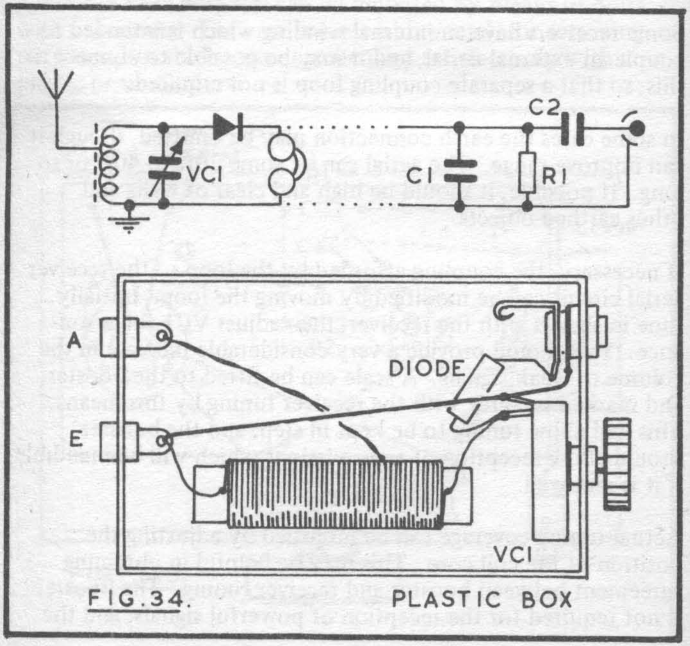

FIG.34.

For best efficiency and small size, the tuning coil is wound on a piece of ferrite rod. It consists of seventy turns of 24swg or similar enamelled wire, on a rod about 1½ inch long and 3/8 inch in diameter. Actual details will not be critical. A piece can be obtained from a longer rod by cutting a notch round the rod with a file and snapping at this point with the fingers. A few touches of adhesive will hold the ends of the winding secure. The loop for the aerial connection is about one-third the total number of turns from the diode end of the coil. (Moving this tapping towards the earthed end of the coil will increase selectivity or sharpness of tuning, but reduce volume.)

Aerial and earth sockets, and a 3.5mm jack outlet, complete the receiver. Headphones can be plugged directly in here.

The aerial and earth can be provided as explained for the diode radio in Fig. 7.

To use the radio as a tuner, C1, R1 and C2 may be added. Suitable values are C1 470pF, R1 220k, and C2 0.25μF. A screened lead with jack plug should be made up to connect the tuner to the amplifier input socket.

One Transistor VHF Receiver/Converter

The super-regenerative receiver is well known for its great sensitivity, which equals that of a multi-stage superhet. It has no alignment difficulties, and few components, so is an interesting project. Its main disadvantages are its background hiss, but this ceases when a signal is tuned in, and its lack of selectivity. The latter is however relatively unimportant on the VHF bands. The circuit here is particularly intended for 2m reception, but is easily modified. It is not intended for FM signals.

In Fig. 35 the detector is a VHF FET, and the field effect transistor will be found very satisfactory in circuits of this kind. T1 is the pre-set tuning capacitor, which allows adjustment over a fairly wide band of frequencies. Tuning is by VC1, which does not require a reduction drive.

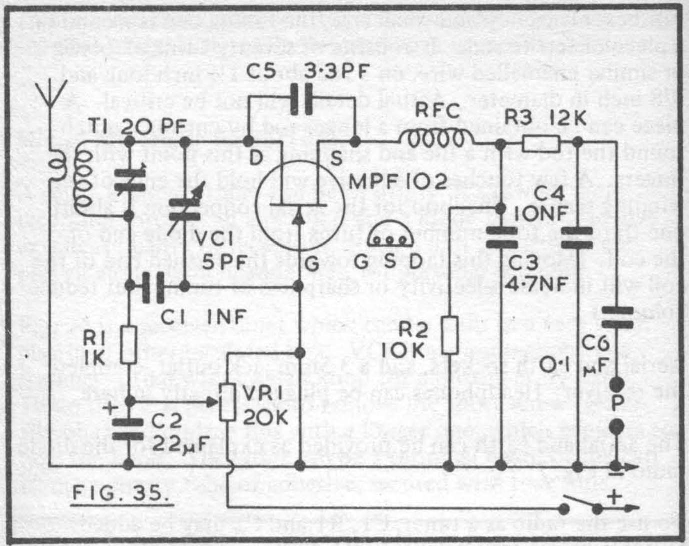

FIG. 35.

Control of super-regeneration is by means of the potentiometer VR1, which adjusts the FET drain potential. The value of VR1 is not very important, but a linear potentiometer is necessary.

C5 provides feedback from source to drain, for regeneration. R3 and the associated components form the quench network, with C6 feeding headphones, or an audio amplifier.

For best results, VR1 should enable the transistor to go smoothly into super-regeneration, which is shown by the hiss mentioned. Achieving this depends somewhat on the transistor, value of C5, adjustment of T1, and aerial loading. However, matters are not too critical.

Transistors unsuitable for VHF cannot be used. Individual FETs of the same type will also differ somewhat, so that in some cases it might be necessary to omit C5, or change the value. Alternatively, a small pre-set (not over 10pF) may be fitted. It will be found that a very small capacitance is required.

Another cause of failing to obtain super-regeneration is too tight aerial coupling, and this is easily corrected by moving the aerial coupling winding a little further from the larger winding.

The RF choke can be a VHF type, or can consist of seventy five turns of 36swg enamelled wire, side by side on a ¼ inch diameter insulated rod. Touches of adhesive may hold the end turns only.

Short leads are essential, and assembly on a tag board is shown in Fig. 35B. The variable capacitor is fitted first, and it can later attach the finished board to a panel, by means of an extra nut.

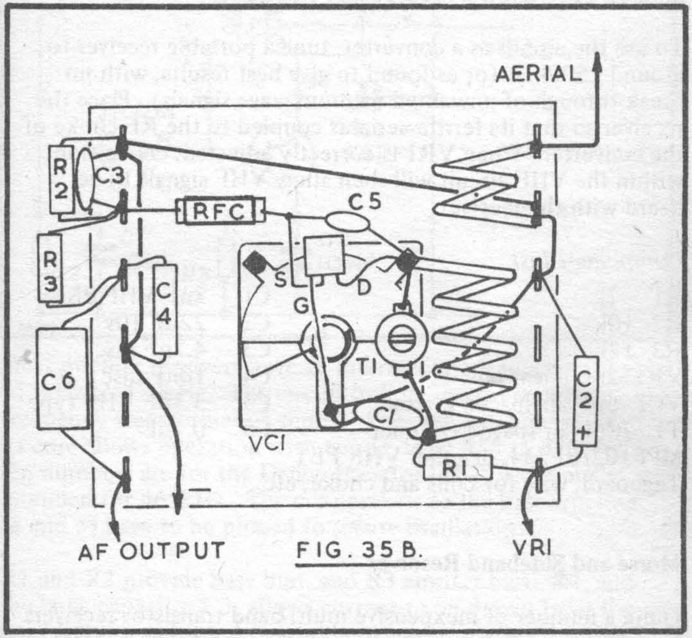

FIG. 35 B.

The larger coil has five turns of 18swg wire, and is ½ inch in diameter and ¾ inch long. The smaller coil has two turns. Wiring to VC1, T1, C5, the coil and C1 must be as short as can be arranged. Connections elsewhere are reasonably direct. A ceramic disc capacitor may be used at C1. T1 should be a low loss trimmer.

The board can be fitted in a case which will carry VR1, and large enough for the battery and on-off switch. The case can be used to support a telescopic aerial, some 30 inch or so when extended.

Frequencies tuned by VC1 can be raised by unscrewing T1, or by stretching the larger coil to separate the turns more, or by using a coil with fewer turns, or of smaller diameter. The receiver is most suitable for use over the 70MHz to 200MHz range.

For the single transistor alone, a sensitive pair of medium or high resistance phones will be needed. Audio ouptut from C6 can be taken to an amplifier, for more powerful reception with phones, or to allow a loudspeaker to be used.

To use the circuit as a converter, tune a portable receiver to around 1500kHz (or as found to give best results, with no break through of unwanted medium wave signals). Place the receiver so that its ferrite aerial is coupled to the RF choke of the converter. When VR1 is correctly adjusted, conversion within the VHF circuit will then allow VHF signals to be heard with the receiver.

Component List

R1	1k	C1	1nF VHF disc
R2	10k	C2	22μF 10v
R3	12k	C3	4.7nF disc
VR1	20k linear pot.	C4	10nF disc
VC1	5pF air-spaced variable	C5	3.3pF VHF type
T1	20pF air spaced trimmer	C5	0.1μF

MPF102/BF244, or other VHF FET
Tagboard, wire for coils and choke, etc.

Morse and Sideband Resolver

Quite a number of inexpensive multi-band transistor receivers have frequency coverage which includes one or more Amateur bands. That is, 1.8, 3.5, 7, 14, 21, or 28MHz. On these bands transmissions are often in Morse code or CW, or single sideband. CW or SSB can be resolved by the specialised Amateur receivers and transceivers, but not with the general purpose home receiver. To receive CW or SSB with the latter type of radio, it is necessary to add an oscillator. This either generates a heterodyne or beat note for code reception, or replaces the carrier which was eliminated from the single sideband transmission.

The circuit in Fig. 36 may be used for these purposes. It generates a carrier of about the same frequency as that used in the intermediate frequency stages of the receiver, exact adjustment of frequency being by VC1,

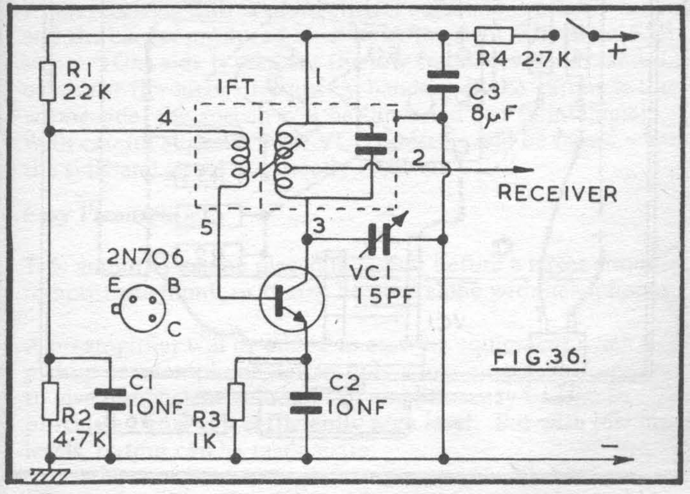

FIG.36.

Most popular receivers have an intermediate frequency of about 455kHz to 470kHz. The circuit in Fig. 36 uses an intermediate frequency transformer of similar type, and adjustment of its core allows operation with any IF from 455 to 470kHz. Pin numbers are for the Denco (Clacton) IFT13, normally intended for 465kHz. The connections to the base winding (4 and 5) have to be phased to secure oscillation.

R1 and R2 provide base bias, and R3 emitter bias. R4, and the large capacitor C3, allows current to be drawn from the same battery as is used to run the receiver, if wished. Alternatively, a separate 9v battery may be employed here.

Fig. 36B shows the practical layout of components, and wiring to them, using a perforated board. Connections and joints are kept close against the underside of the board. When wiring is finished, the board is mounted in a metal box, using ½ inch bolts with extra nuts for clearance.

VC1 can best have a dial with a central zero mark, this showing when the plates are half meshed.

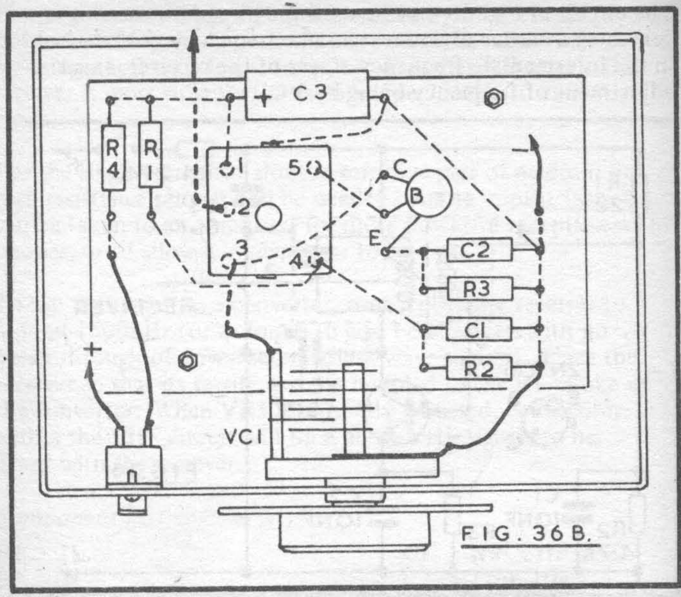

FIG. 36B.

An insulated lead is run from 2 on the IFT, and is placed near the receiver. This lead may be so positioned as to give best coupling into the IF stages of the receiver. If coupling is too weak, the signal from the CW/SSB Resolver will be inadequate. On the other hand, tight coupling will allow the carrier to swamp weak signals tuned in on the receiver. Fortunately the degree of coupling is not at all critical.

To set up the oscillator, first of all tune in any normal broadcast with the receiver. Set VC1 half closed and switch on. The core of the IFT should now be slowly adjusted until a loud heterodyne or beat note is heard accompanying the signal received. Set the core so that this note falls to about zero, or to a low pitch. It will then be possible to rotate VC1 either way from this central zero beat position, in each case producing a note which rises in frequency as VC1 is rotated.

To receive CW or Morse code, it is only necessary to tune in the code signal with the receiver, and then to adjust VC1 to

obtain a note of the most suitable pitch. VC1 can be rotated either way from its central or zero beat position, to obtain best freedom from interference.

When receiving SSB, a more critical adjustment of VC1 is needed and the carrier produced must be at the right side of the SSB signal. (One side is used for the low frequency bands, and the other for the higher frequency bands.) If the carrier is the wrong side, the speech will be "inverted" or unintelligible. With careful adjustment of VC1, a setting will be found where the sideband signal is correctly resolved.

Easy Preamplifier

This amplifier can be placed in circuit before a larger amplifier, to boost the input, or it may be used alone with headphones.

A preamplifier will be added to existing equipment when a pickup or microphone having only a low output has been found to give insufficient volume. An amplifier may be able to provide a signal at a sufficiently high level. But with low input levels, output can be inadequate.

The stage in Fig. 37 can provide considerable gain. Audio input is to C1, and R1 provides base current for the transistor. Amplified signals are obtained across R2, and pass to the main amplifier (or phones) via the isolating capacitor C2.

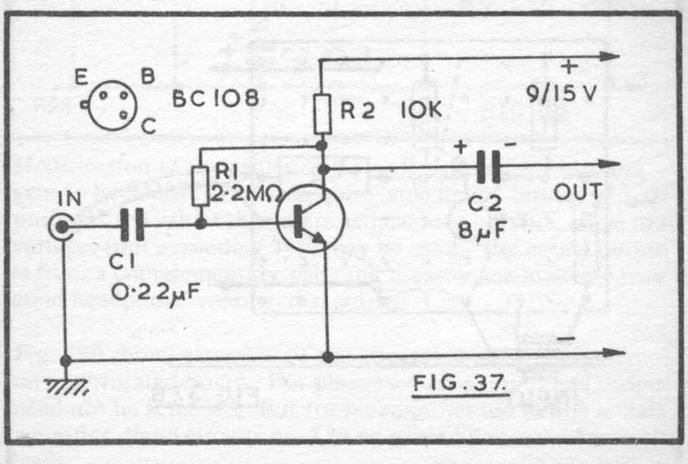

FIG. 37.

With a BC108 or similar transistor, the supply can be about 9v to 15v. Numerous NPN transistors, intended for preamplifier use and of low noise type, will perform well. Current may be from a separate battery. Or it may in many cases be drawn from the main amplifier. To avoid instability or similar troubles, the positive supply should be from a well decoupled and smoothed point in the main amplifier. If this is not available, connect a 220μF capacitor from positive to negative in Fig. 37, and draw the positive supply through a 1k resistor. Negative is returned to the negative line of the main amplifier.

Should the main amplifier have a positive earth line, then the transistor can be a PNP type, R1 and R2 being changed to suit.

Fig. 37B shows assembly on a perforated board. When the circuit provides input to a main amplifier, it should generally be in a metal case. Input is by a screened lead, and a screened lead is also fitted for the audio circuit to the main amplifier. The case should be earthed to the negative line. This can be by means of one of the bolts securing the board.

Screening to avoid hum or instability will generally prove to be unnecessary when using the stage to operate headphones. So two sockets, or a jack socket, will be more convenient for the phone circuit.

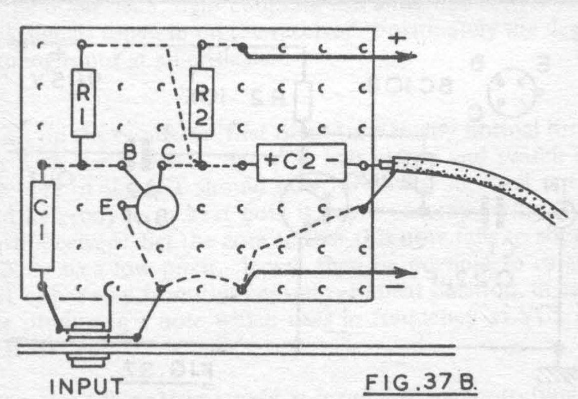

FIG.37B.

741 Preamplifier

Integrated circuits have several stages of amplification, formed as a single chip and enclosed in a small case. The 741 is an easily obtained audio IC which can be used in many circuits. The uA741CE or metal can 741 is convenient, as its leads may be spaced to suit any ready perforated board. The dual in line 741 is electrically similar.

Fig. 38 shows a preamplifier using the metal can 741. Input is via C1 to 3. As shown, 7 and 4 are positive and negative, while output is from 6. Negative feedback is applied to 2, via R4. C3 is an isolating capacitor.

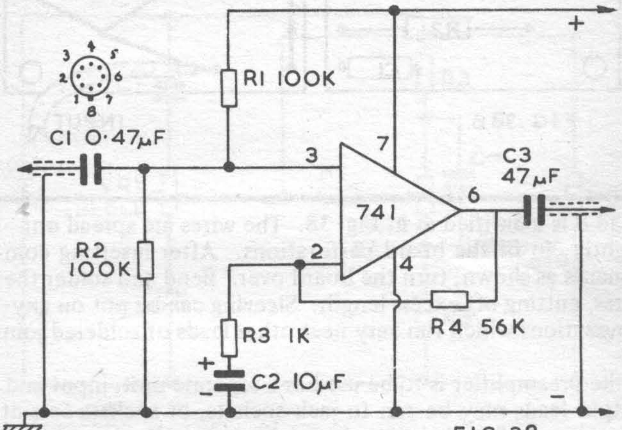

FIG.38.

Modification to the values of R3 or R4 allow feedback and gain to be adjusted between quite wide limits, but for general purposes the values shown are satisfactory. A wide range of voltages (not exceeding 36v) may be used. The actual output is from a complementary pair, and is easily adequate for very good headphone volume, if required.

Fig. 38B shows assembly of the integrated circuit amplifier on a perforated board. For phone working, input and output need not be screened. But for preamplifier use before a main amplifier, these circuits need to be made by means of screened leads.

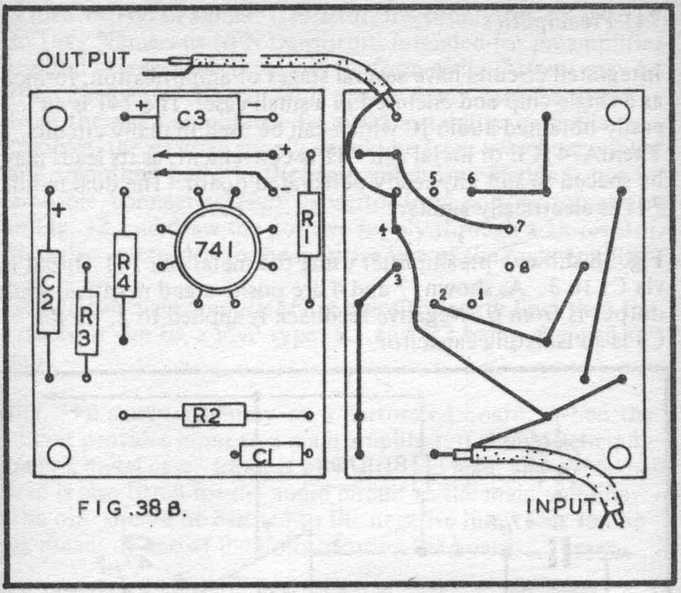

FIG.38B.

Lead 8 is identified as in Fig. 38. The wires are spread out slightly, to fit the board perforations. After inserting components as shown, turn the board over. Bend and solder the wires, cutting off excess length. Sleeving can be put on any connections which run very near other leads or soldered joints.

If the preamplifier is to be used as a separate unit, input and output leads may be run to jack sockets, or sockets to suit other items of equipment. A metal case is advised, as this will help to avoid any pick up of hum, or other troubles due to coupling to the board wiring. The board can be secured to the metal box or case with ½ inch 6ba bolts, and extra nuts to give a little clearance.

The amplifier may be run from its own battery; or from a supply used with a main amplifier. For the latter, it is wise to provide a decoupled point, in a similar manner to that described for the transistor preamplifier. This consists of a quite large capacitor (say 220μF to 470μF) from positive to negative across the preamplifier supply circuit, and a resistor of 1k to 2.2k or so from preamplifier positive to positive on the main amplifier.

Tone Control for Your Discs Etc.

The preamplifiers shown have a substantially flat response. That is, they do not purposely emphasize bass or treble, or have similar effects in modifying the audio frequencies present. In some circumstances it is necessary to boost treble or bass, or reduce the emphasis of these frequencies, in order to obtain the most satisfactory reproduction.

This may be achieved by making a fully adjustable tone control, and Fig. 39 is a suitable circuit. A volume control is also incorporated, as this has not been present in the preamplifiers.

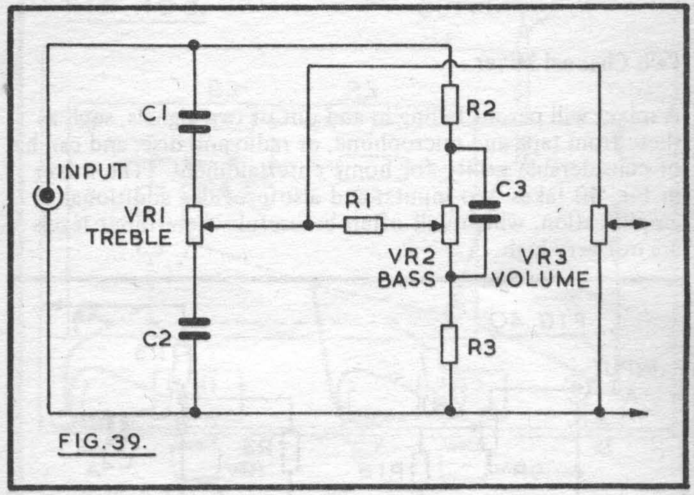

FIG.39.

The tone control is a passive circuit, needing no battery, and it can be placed between the source of audio signals (generally a microphone, pick-up, or radio tuner) and the amplifier or preamplifier. It may also be located between preamplifier and main amplifier, if wished.

Results are substantially flat with the treble control VR1 and base control VR2 in their central positions. Adjustment of VR1 and VR2 will provide separate boost or cut of treble, or boost or cut of bass frequencies. VR3 is the volume control.

The degree of control afforded by circuits of this type may be adjusted by alterations to the component values. In practice, it will be found that there is little point in having extreme levels of adjustment, which will never be required. The following values may be used: C1 and C2, 270pF; R1, 390k; R2 and R3, 100k; C3 680pF; VR1 and VR2, 2 megohm linear potentiometers; VR3, 2 megohm log potentiometer.

Components should be assembled in a metal case, for screening. The case is grounded to the earth line. Input and output sockets can be near the ends of the case, with VR1, VR2 and VR3 between. Components can be soldered directly to the potentiometer tags.

Two Channel Mixer

A mixer will permit fading in and out of two signals, such as these from tape and microphone, or radio and disc, and can be of considerable utility for home entertainment. The mixer in Fig. 40 takes two inputs, and also provides additional amplification, which will often be useful where input levels are not very high.

FIG.40.

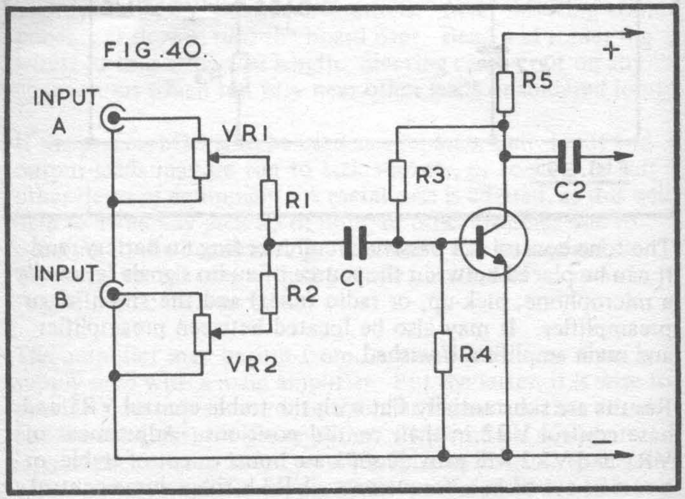

Input A has its individual volume control VR1, and Input B is similarly provided with VR2. Resistors R1 and R2 prevent short circuiting one input when the other potentiometer is set at zero. Coupling to the transistor base is by C1. R3 and R4 set base bias conditions, with R5 as the collector load and audio output from C2.

Component values are not very critical, though R3, R4 and R5 need to be selected to suit the transistor fitted. The mixer is connected and used in a similar way to that explained for the preamplifiers.

FIG.40B.

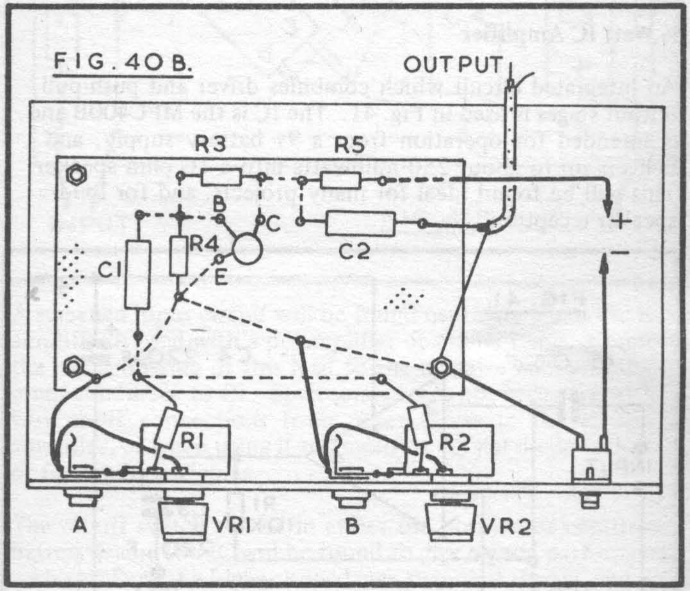

Fig. 40B is the layout, using a metal case or box. A screened output lead is provided. Input sockets A and B may be chosen to suit the type of plugs on existing equipment. The on-off switch is separate from the potentiometers, so that these can be left at pre-set levels if wished.

Component values may be chosen to suit low or high impedance inputs, if wished. Or the following may be used for general purposes:

Components List:

VR1, VR2, 1 megohm log potentiometers.

R1, R2, 200k	
C1 0.5μF	C2 0.5μF
R3 330k	BC109
R4 22k	
R5 10k	

This is when using a 9v supply.

¼ Watt IC Amplifier

An integrated circuit which combines driver and push-pull output stages is used in Fig. 41. The IC is the MFC400B and is intended for operation from a 9v battery supply, and delivers up to about 250 milliwatts into a 16 ohm speaker. This will be found ideal for many projects, and for loudspeaker reception.

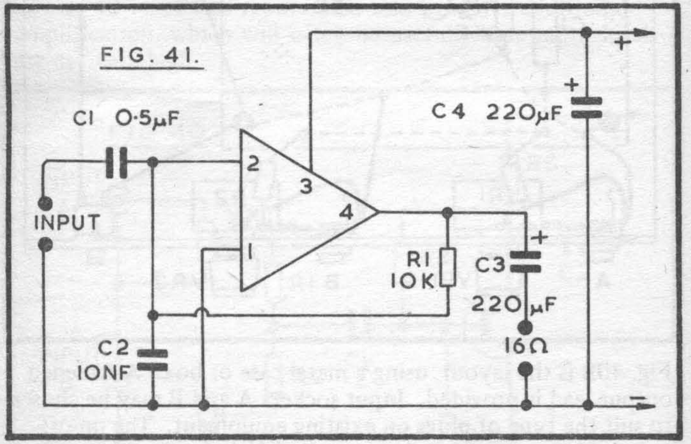
FIG. 41.

Audio input is to 2 of the IC, from the isolating capacitor C1. C2 is a by-pass capacitor to maintain stability. The IC has only four connections, 1 going to the negative line, and 3 to the positive supply. Audio output is from 4, and to the loudspeaker via C3. R1 provides negative feedback and bias to point 2.

The assembly of this amplifier on 0.15 inch matrix board is shown in Fig. 41B. Pin numbering is when looking at the top of the IC and the component side of the board.

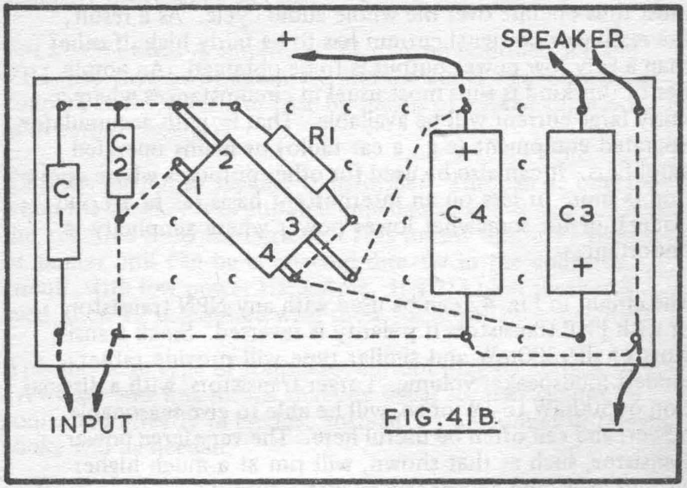
FIG.41B.

A screened input circuit will be found necessary when the IC amplifier is used with a preamplifier or similar items. Connect the outer brading of this lead to the negative line, and the inner conductor to C1. Such screening is not necessary with very short connections from other stages to the IC amplifier, or when using it to amplify a crystal diode radio, or for similar purposes.

The on-off switch can be in either the positive or negative battery lead. The IC will be found to give a good performance with speakers of a higher impedance than that shown, and up to 75 ohm may be used (with some small loss in output). Units of lower impedance than 16 ohms should not be fitted.

The small size of this IC and the few extra components needed make is very suitable for miniature apparatus, probably with a 2½ inch speaker. It will also operate a large speaker well (see "Boost Speaker") and at a volume level adequate for many purposes.

Class A Amplifier

A Class A amplifier uses a single transistor, Fig. 42, which must thus operate over the whole audio cycle. As a result, the resting or no-signal current has to be fairly high, if other than a very low power output is to be obtained. An amplifier of this kind is thus most usual in circumstances where a fairly large current will be available. That is, with accumulator operated equipment (e.g., a car radio) or mains operated amplifiers. It can also be used for other purposes where operation is more or less on an intermittent basis (as in a Loud Hailer), or for somewhat lower power where simplicity is important.

The circuit in Fig. 42 can be used with any NPN transistor, or with PNP transistors if polarity is reversed. Small transistors of the 300mW and similar type will provide rather modest loudspeaker volume. Larger transistors, with a dissipation of 500mW to 1W or so, will be able to give reasonable power, and can often be useful here. The very large power transistor, such as that shown, will run at a much higher current level, and deliver several watts output.

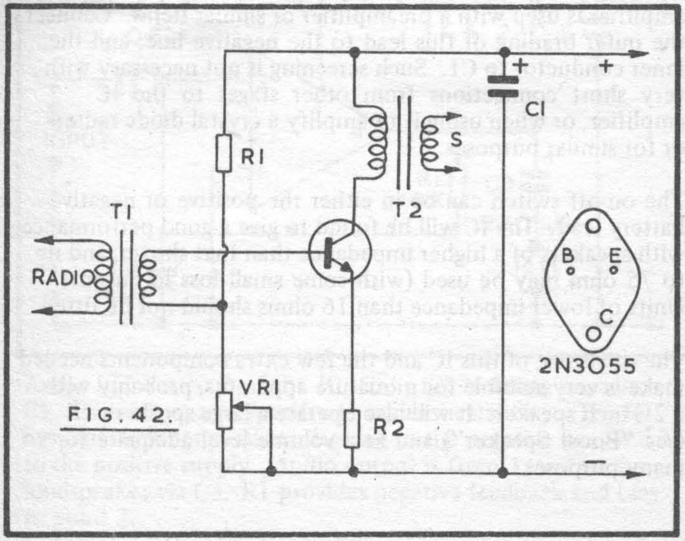

FIG. 42.

T1 is the coupling or driver transformer, which may receive audio signals from a radio receiver, smaller amplifier, or other equipment. A step down ratio of about 5:1 to 10:1 will generally prove to be suitable.

For smaller Class A amplifiers, T1 may be omitted. R1 and VR1 may then be of about 1k, and the audio signal can be coupled to the transistor base by means of a fixed capacitor, as in earlier circuits.

T2 is the output or speaker transformer. This can be omitted when the speaker is of suitable impedance, and is able to carry the required collector current. This means that a 75 ohm or similar unit can be connected directly in the collector circuit, with low power transistors. If a 2/3 ohm speaker is used, then T2 is required with these for impedance matching.

With the large type of transistor, current may be 500mA to 1A or so, and this means that the usual speaker cannot be connected directly in circuit. Instead, a transformer or tapped choke will be needed.

The circuit is often operated from 12v (a vehicle supply) but can be used with 9v, or more than 12v.

R1 and VR1 set bias conditions, and R2 limits peak collector current. For the 2N3055, R2 may be 1 ohm. R1 can be 100 ohm, and VR1 25 ohm. T2 can be 2:1 for a 3.5 ohm speaker. C1 is 3200μF. Begin with VR1 at minimum value, slowly increasing this until reproduction is satisfactory. (This will set collector current at about ½ ampere.)

With other transistors, adjustment can be made in the same way, by slowly increasing the base bias until best results are secured. Raising the value of VR1 in Fig. 42, or reducing the value of R1, will increase collector current, and thus raise the power which can be handled. However, current is limited by the power handling capacity of the transistor, so a meter should be put in one battery lead, and a check made that the power dissipation of the transistor is not exceeded. With small transistors, current can be limited to 50mA or so.

Fig. 42B shows assembly of the circuit in Fig. 42. The power transistor is mounted on an aluminium plate, which acts as a heat sink to carry away heat from the transistor itself. Note that this plate is common to the collector. Base and Emitter pins pass through clearance holes. Smooth off any burr round the fixing and clearance holes, so that the transistor can rest flat on the metal. Where it is awkward to have the plate or case common to the collector, the usual insulation set can be fitted. This consists of bushes for the fixing bolts, and a thin mica washer to go under the transistor, so that electrical isolation is obtained.

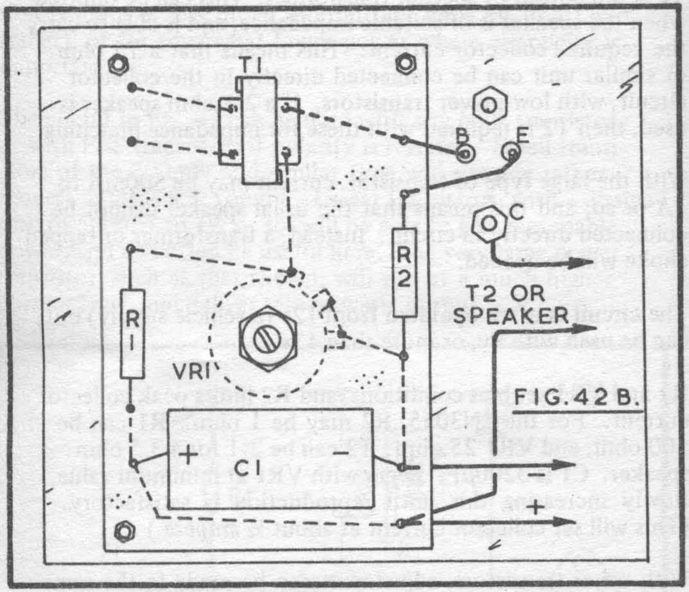

FIG. 42 B.

The insulated board carrying the other components is mounted by long bolts with extra nuts. A large Class A stage transistor of this type will grow very warm to the touch, if run at anything like maximum power rating. This does not indicate any defect.

Two Stage Amplifier

The circuit of a 2-stage amplifier is shown in Fig. 43, and this can be used for loudspeaker reproduction. Input to the driver stage is by the isolating capacitor C1. Operating conditions in this stage depend largely on the values of R2, R3, and the emitter bias resistor R5.

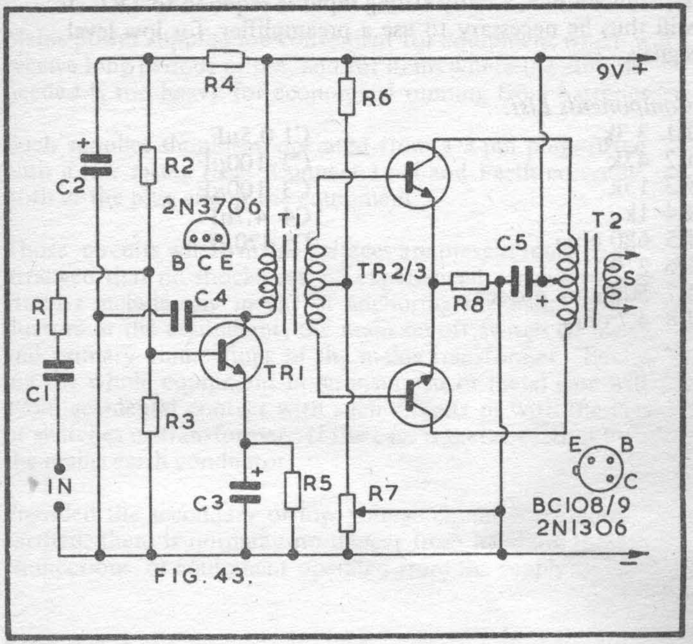

FIG. 43.

T1 is the driver transformer, with centre tapped secondary, so that TR2/3 may be operated in push-pull (see Fig. 12). T2 couples output to a 2/3 ohm speaker.

With the values shown, a BC109 will give very good results in the TR1 position, with 2 x BC108s for the output stage. R7 is set so that battery current is around 10mA, with no signal present. Current rises from 20mA to 60mA or so, according to volume, when the amplifier is driven.

A pair of 2N1306 transistors may be used for the output stage. Other transistors are also possible here, as operating conditions can be adjusted by R7.

The 2N3706 will operate in the TR1 position, though with lower gain than the BC109. Numerous other small audio transistors can also be used here.

For full volume, a fairly strong input is required to TR1. It will thus be necessary to use a preamplifier, for low level signals.

Components List:

R1 3.3k
R2 47k
R3 15k
R4 1k
R5 680 ohm
R6 2.7k
R7 500 ohm pre-set
R8 4.7 ohm

C1 0.5μF
C2 100μF
C3 100μF
C4 4.7nF
C5 470μF

SECTION 4
POWER SUPPLIES

Mains power supplies are convenient for equipment which will receive long periods of use, and for items where the current needed is too heavy for economical running from batteries.

Such supplies should be operated from a 3-pin plug, fitted with a low rating fuse. Connect L, N and Earth correctly, both at the plug and at the equipment.

Those circuits where mains voltages are present must be so arranged that no shocks can be experienced. Mains voltage circuits include any means of anchoring the mains conductors in the equipment, the main on-off switch (if used), and primary connections of the mains transformer. Enclosing the whole equipment in an insulated or metal case will avoid accidental contact with such circuits or with the tags of switches or transformer. If the case is metal, earth it by the mains earth conductor.

Provided the secondary or low voltage circuit is correctly earthed, there is normally no danger from handling these connections, or equipment operated from the supply.

Easy PSU

A useful power supply suitable for many purposes can use the circuit in Fig. 44. Mains L (live) and N (neutral) conductors go to the primary P of the transforer. A switch is placed in the L circuit.

The secondary S can delivery any wanted low voltage, and also give isolation from the mains.

Almost all equipment will need direct current. The four separate rectifiers, or an equivalent single device, produce pulsating DC. The reservoir capacitor C1 smoothes this. Resistor R1, with the second capacitor C2, gives additional smoothing. Output from the supply is then substantially smooth DC, as would be obtained from a battery.

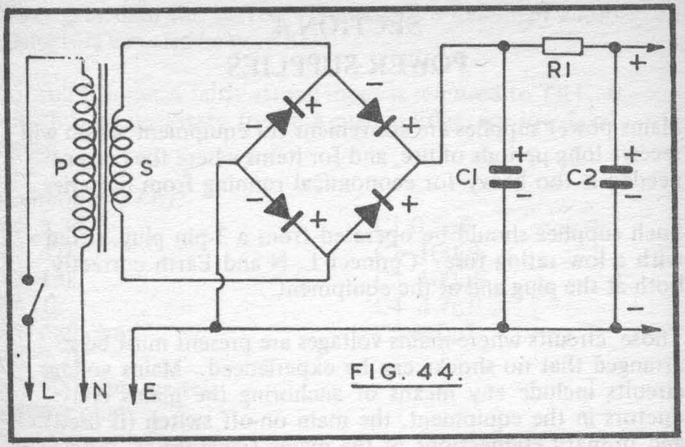

FIG. 44.

Sometimes R1 and C2 are omitted. Should hum accompany reproduction from an amplifier, this shows that smoothing is probably inadequate, and if so the cure is to add R1 and C2.

Any wanted voltage can be obtained by using a transformer with an appropriate secondary, with rectifiers and capacitors to suit.

C1 may charge up to approximately 1.4 times the RMS or usual voltage rating of the secondary S. Thus, if S were a 6.3v winding, about 8.8v would be expected across C1. But should R1 be present, a voltage drop will arise in it, the figure depending on the resistance value of R1, and current flowing.

For these reasons, the actual output may not be exactly the figure expected. This will often be unimportant, and such power supplies are largely used for all sorts of equipment.

For apparatus generally operated from a 9v or similar battery, a bell transformer with 3v, 5v and 8v output tappings is suitable. If R1 is 22 ohm, this will give around 9v at 100mA, with the 8v tap in use.

It is convenient to use 50v or 100v rectifiers, and the rating can be 1 ampere, with a 1 ampere transformer secondary. Four individual wire ended rectifiers can be mounted on a tag strip, and connected as in Fig. 44. A combined rectifier will have AC, positive, and negative tags (see Fig. 45) and can be fitted instead.

The capacitors need to be of large value, and have a voltage rating higher than the actual voltage which will be present. The 12v capacitors could be fitted in the circuit described. C1 can be 2500μF or larger, and C2 can be of similar value.

Easy Regulated PSU

With the circuit in Fig. 44, the exact voltage will depend on the current drawn, rising when current is low. To avoid this, some means of regulation may be employed.

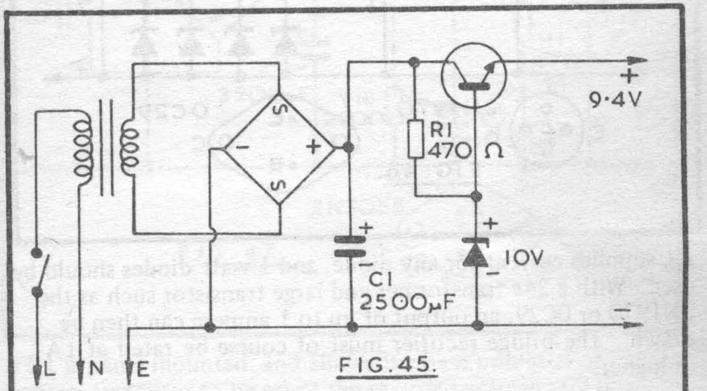

FIG.45.

Fig. 45 shows a typical regulation circuit, giving an output of approximately 9.4v. The output from the rectifier needs to be of somewhat higher voltage than required at the output, and a transformer with a 10v secondary is suitable. Current through R1 supplies the Zener diode, which holds the base of the transistor at virtually the same voltage for all current loads. Some voltage is lost in the transistor, so the output is nearly the same, for any current load. For up to about 200mA, a BFY51 or equivalent transistor is adequate, so this will do well for many purposes.

Multi-Voltage Regulated PSU

Fig. 46 is a somewhat similar circuit, but it has a switch, to select any one of the four Zener diodes. These may be chosen so that 6v, 9v, 12v and 15v will be available.

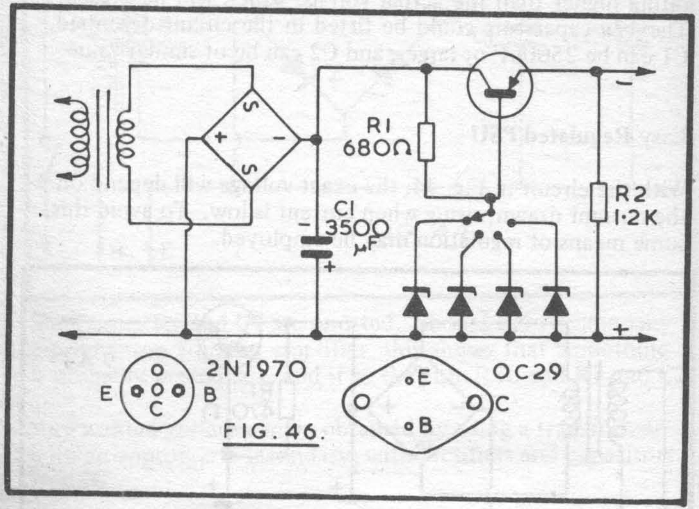

FIG. 46.

R1 supplies current for any diode, and 1 watt diodes should be used. With a 24v transformer and large transistor such as the 2N1970 or 0C29, an output of up to 1 ampere can then be drawn. The bridge rectifier must of course be rated at 1A or higher.

Note that this circuit has a PNP control transistor, so that polarity is obtained as shown. (Compare with the NPN transistor in the previous circuit.) However, either positive or negative can be grounded, and a separate earth terminal can be provided to allow this. A flying lead is placed between earth terminal and positive, or earth and negative, as required. The permanent negative earth connection shown in Fig. 45 must of course then be omitted.

More details of actual construction will be found later.

One Transistor Adjustable PSU

Fig. 47 is the circuit of a simple power supply unit which can be set to provide any output from zero to 12 volts. The transformer has a 17v secondary, rated at 1 ampere or higher. After rectification, DC is obtained from the reservoir capacitor C1.

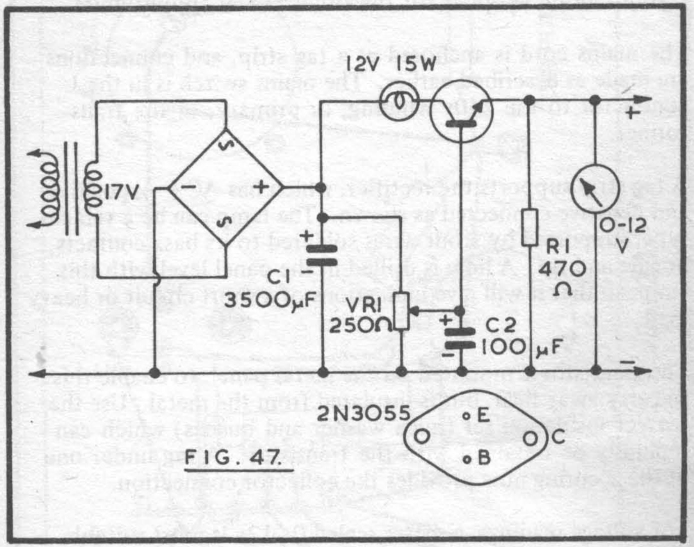

FIG. 47.

VR1 is panel mounted, and allows the base potential of the control transistor to be set at the required voltage. C2 provides "electronic smoothing" from the base circuit.

The 12 volt 15 watt bulb is for short circuit protection. With relatively low currents, little voltage is lost here. But if a short circuit arises in the output circuit, current is limited to a little over 1A by the bulb. This reduces the chances of any damage due to temporary shorts in leads or equipment.

R1 is a bleeder resistor, drawing some current even when no external load is connected. The meter is the usual high resistance type, scaled from zero to 12v. VR1 is adjusted until this instrument shows the wanted voltage.

The regulation of this supply is not so good as that of supplies using a number of transistors. But it will be found suitable for all ordinary purposes, and to supply power for circuits in this book, as only a small change in voltage arises, with changes to the load, or current drawn.

Fig. 47B shows practical construction of this PSU. A similar layout can be adopted for the other power supply units.

The mains cord is anchored at a tag strip, and connections are made as described earlier. The mains switch is in the L conductor to the 240v winding, or primary, of the transformer.

A tag strip supports the rectifier, which has AC (~), positive and negative connected as shown. The lamp can be a vehicle type, supported by stout wires soldered to its base contacts, or side and tip. A hole is drilled in the panel level with this lamp, so that it will give indications of a short circuit or heavy load.

The transistor is mounted on the metal panel, to enable this to carry away heat, but is insulated from the metal. Use the correct insulation set (mica washer and bushes) which can generally be obtained with the transistor. A tag under one of the securing nuts provides the collector connection.

For voltage readings, a meter scaled 0–12v is most suitable. Or a 1mA meter, with 12k 1% series resistor, may be used. If so, a new scale, marked 0–12, can be drawn for it.

A 3 watt linear potentiometer is used. Earth and negative are connected to the metal panel, for equipment with a negative ground line. The positive output terminal or socket is insulated from the metal.

The case can be metal, or wood. It should have a few ventilation holes. The transistor can be expected to run quite warm, at maximum current. Heating here rises as the output voltage is reduced by VR1, as in these circumstances more voltage is dropped in the transistor.

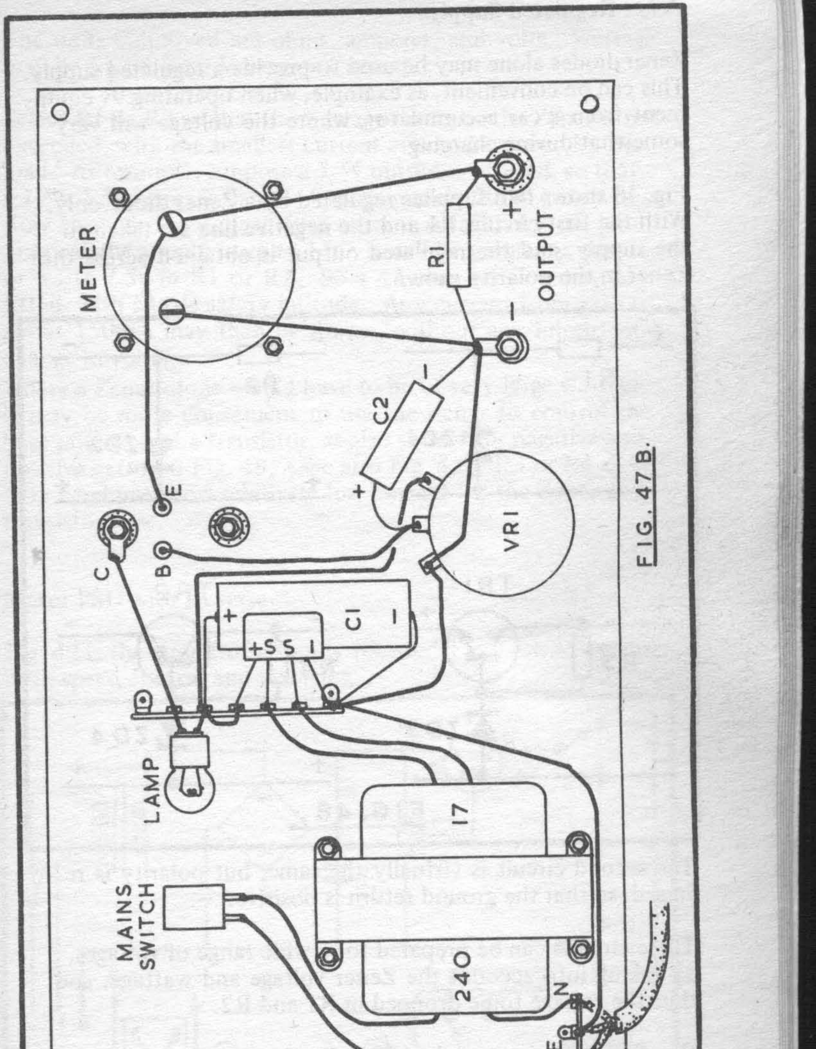

FIG. 47 B.

Zener Regulated Supply

Zener diodes alone may be used to provide a regulated supply. This can be convenient, as example, when operating 9v equipment from a car accumulator, where the voltage will vary somewhat during charging.

Fig. 48 shows two supplies regulated by a Zener diode only. With the first circuit, R1 and the negative line are taken to the supply, and the regulated output is obtained across the Zener in the polarity shown.

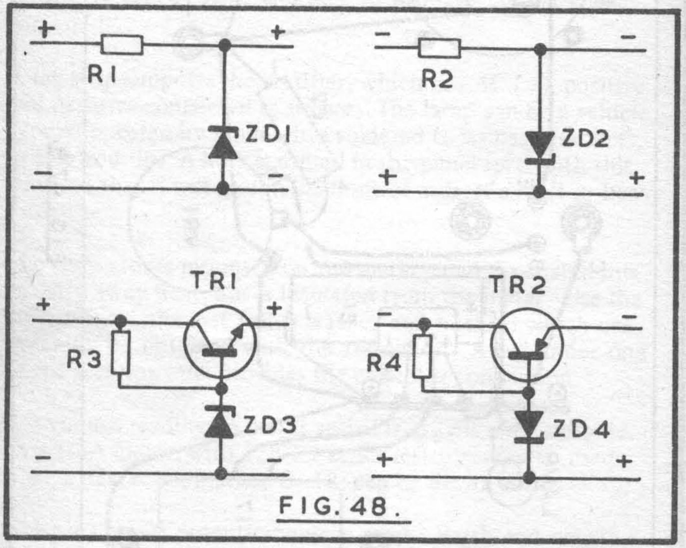

FIG. 48.

The second circuit is virtually the same, but polarity is reversed, so that the ground return is positive.

These circuits can be prepared for a wide range of voltages, by taking into account the Zener voltage and wattage, and thus the voltage to be dropped in R1 and R2.

For calculations, remember the following:
Voltage = Current in Amperes x Resistance in Ohms.
Current = Voltage divided by Resistance.
Resistance = Voltage divided by Current.

The units employed are ohms, amperes, and volts. Wattage may be found from Voltage x Current.

Select R1 or R2 so that the wattage rating of the Zener is not exceeded, with the smallest current ever to be drawn by the load. As example, suppose a 7.5v output is wanted, so that a 1½w 7.5v Zener is fitted. Approximately 200mA may thus flow through the Zener, if no current is taken by the load. Suppose the supply is up to 15v. It will be necessary to drop or waste 7.5v in R1 or R2. So a 42 ohm resistor may be fitted, with a little safety latitude. Any current from zero to about 150mA may then be drawn, without any important change in voltage.

Where a Zener diode would have to be of very large wattage, it may be more convenient to use the Zener to control the base potential of a transistor, as also shown for negative and positive earths in Fig. 48. (See also Fig. 45.) R3 or R4 then only need provide a relatively low current for the Zener and transistor base.

Motor PSU with Reverse

Fig. 49 is the circuit of a supply for model motors, and giving both speed control and reversing.

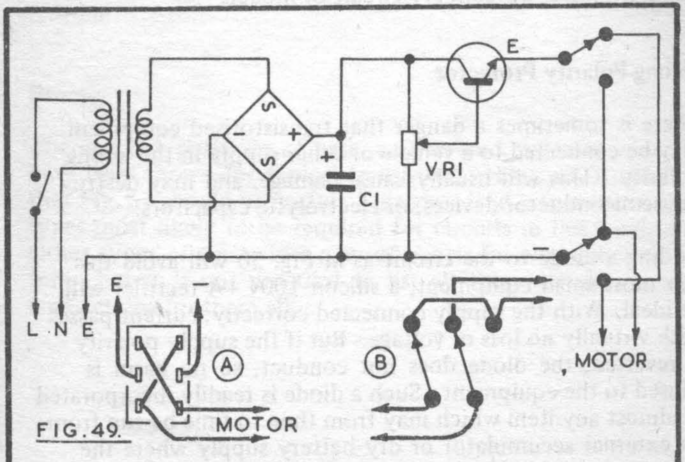

FIG. 49.

A transformer with a 12v secondary will be suitable for the usual model requiring up to 12v. The circuit is similar to that of the supply in Fig. 47. The potentiometer VR1 gives speed control. C1 can be around 2,000μF.

"A" is one type of ready-made switch, and it can be obtained with a central off position. "B" shows a home-built switch, similar to that described in Fig. 5. The switch must not short circuit the supply when changing over polarity, and one with central off is recommended.

Motors reversed in this way must be of the permanent magnet type, in which a change of polarity causes the armature to run in the opposite direction. Train and many other model motors are of this construction. Motors with a wound field will not be reversed by changing the polarity of supply.

A controller of this kind should be constructed as shown earlier, so that it is safe to use, and so that no joints in the mains circuits can be touched.

Earthing can only be omitted when a double-insulated transformer, of the type specifically made to avoid any need for earthing, has been used. The whole controller should be so constructed that, even if a fault develops, mains voltages cannot reach the secondary circuit or model.

Wrong Polarity Protector

There is sometimes a danger that transistorised equipment may be connected to a vehicle or other supply in the wrong polarity. This will usually cause damage, and may destroy the semiconductor devices, or electrolytic capacitors.

Adding a diode to the circuit as in Fig. 50 will avoid this. For most small equipment, a silicon 100v 1A rectifier will be ideal. With the supply connected correctly, current passes with virtually no loss of voltage. But if the supply polarity is reversed, the diode does not conduct, so no harm is caused to the equipment. Such a diode is readily incorporated in almost any item which may from time to time be run from an external accumulator or dry battery supply where the polarity is not known.

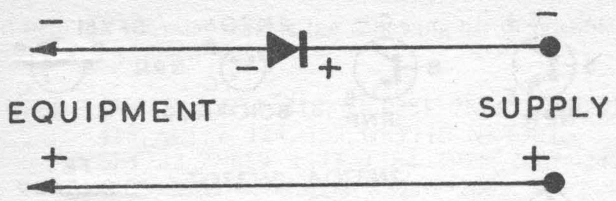

FIG. 50.

Component Tolerances

For circuits in this book, exact component values are generally not too important. Capacitor values such as 47nF (0.047μF) and 0.05μF may be regarded as interchangeable. In the same way, a 125μF capacitor may replace 100μF, or 200μF can be used instead of 220μF, and so on.

Resistors will generally be the standard 5% tolerance types. In some circuits exact resistor values are not at all critical. In others, and especially base, emitter and collector circuits, the values which provide best results can have much less latitude.

Bases

The lead-out positions of base, emitter and collector wires have been shown for typical transistors. However, it is felt that Fig. 51 should be useful. This gives lead information for types most likely to be required for circuits in this book. Other types of transistors may of course be used successfully, but it is not practical to list all these, or show connections for them all.

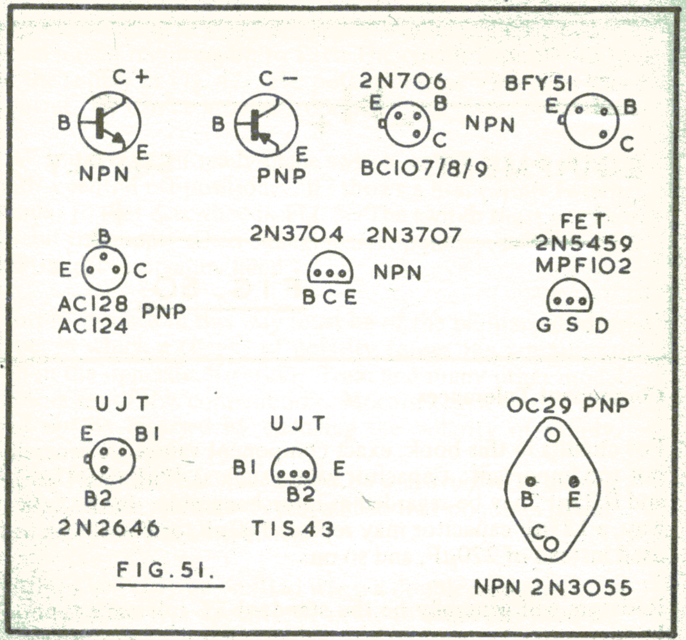
C +
B
E
NPN
C −
B
E
PNP
2N706
E
B
C
NPN
BC107/8/9
BFY51
E
B
C
B
E
C
AC128
AC124
PNP
2N3704
2N3707
B C E
NPN
FET
2N5459
MPF102
G S D
UJT
E
B1
B2
2N2646
UJT
B1
B2
E
TIS43
OC29 PNP
B
E
C
NPN 2N3055

FIG. 51.

EQUIVALENT SEMICONDUCTORS

To help our overseas readers the following list of possible equivalents is shown below: –

AC128	CV9778, OC74–318, NKT281, 2N467, 2SB222–415, AC117–124–153, GET110, NKT302, SFT232, 2N659–1373–1384–1926–2001–2431–4106, 2SB370, SK3004, RS276–2005.
ACY39	CV9475, NKT243, ACZ10, 2N652–2303–2801.
AF117	2N642, 2SA155, AF102–105–127–133–137, BFX48, BSW19–73, GM1213B, NKT603F, SFT354, SK3006, 2N2273–2635–3324–4916–5354, RS276–2003.
BC107	CV9780, AM251, BC147–167–207–317, MPS6566, SK3020–3122, ZTX107, TT107, RS276–2009.
BC108	CV10541, AM252, BC148–168–208–318, MPS6520, SK3020, ZTX108, TT108, RS276–2009.
BC109	CV10769–10806, AM253, BC149–169–209–319, MPS6521, SK3020, ZTX109, TT109, RS276–2009.
BC147	BC171–182/L–207–317, MPS6566, SK3020, RS276–2013,
BC148	BC172–183/L–208–318, MPS6520, SK3020, BS276–2009.
BC149	BC184/L–209–173–319, MPS6521, SK3020, RS276–2009.
BF200	BF162–173–225–271–311–314, SE5020/1/2/3/4, SK3018, 2N616–2708–3288–3693, RS276–2011.
BFY51	CV7723–7726–10827–11238, BFX51–68A, BFR19, BSX45/X, BSY46–85, SK3024, 40347, 2SC708, ME6102 2N2193/4–2218–2410–3053–3252–4046, RS276–2009.
BSY95A	CV7648–9379, BC108A, BSX51–73–91, BFX96, BSY20–58–62, ME9022, SK3122, 2SC321, 2N744–914–1708–2218–2369–5187, RS276–2011.
MPF102	E305, SK3116.
OA81	AA117–118–132–144, AAZ15, OA91–95–161, SD38, 1N38–56–476–618, 1S33.
OC26	CV8560, NKT452, 2N257–297–458–553–1021, 2SB3, AD130–131–138–140–149, OC28, SK3009, TF80, T1156–3027, 2N174–1073–2870–RS276–2006.

OC29	CV7083–8356, NKT402, 2N457, 2SB86–425, AD150, ASZ16, AUY21, OC28–35–36, SK3009, T13028, 2N174–1073–2870, RS276–2006.
OC44	CV5710–7003, OC170–171–410–613, 2N1303, 2SA15, AF101–117–126, ASY55, AC191, BC126, NKT211, SKT358, SK3005, 2N36–1191/2–1352–1373–1384, RS276–2003.
OC70	CV5457, OC303–602, NKT214, 2N279–1305, 2SB75, AC122–125–151–163, ACY27–34, ASY27, BC213–206A, BSX36, SK3003, 2N1193–1352–1375–2429, RS276–2004.
OC71	CV5712–7005–8344, OC70–304–604, NKT214, 2N280–1305, 2SB77, AC122–125–151–163, ACY35, BSX36, OC3041, SK3003, 2N1193–1352–1375–1384–2429, RS276–2004.
OC72	AC122/3–125/6–131–151 IV–152–162–191, BC126–213–206A, NKT211, SFT353, 2N282–1190–1352–1371–1384–1991–2431, CV5713–7006–8440, RS276–2005.
OC76	CV8314, OC307–6025, 2N284–1305, NKT212, 2SB89, AC128–131–151–152, ACY28, ASY58–76, BC126–213–260A, SK3004, 2N394–1191–1352–1371–1384–1991, RS276–2004.
ORP12	RS276–116, LDR03, RPY25.
1N4001	BY127, EM502, RS276–1101/1135, 1S100.
1N4002	BY127, EM502, RS276–1102/1136, 1S100.
2N1306	CV7350–10686, NKT736, 2N634–635–636–1891–1993–1994–1995, ASY26–29, BSX19–20, RS276–2001.
2N1613	CV7440–8843, BFY33–34–67, BSY44, 2N698–699–717–1711–1889–1890–1893, BFY41, BSX45/6, BSY53, MPS6530, 2SC708, 2N2101.

N.B.: Although equivalent semiconductors may have similar electrical properties, physical dimensions may be different and this must be born in mind if space is tight and for mounting details.

Remember polarities if replacing PNP with NPN types and vice versa. If in doubt always be advised by your dealer for suitable equivalent semiconductors.

Notes

by Peter J. Bertini, radioconnection@juno.com

Building An FM Crystal Receiver!

Our earlier crystal receiver projects proved popular, so here's another crystal receiver with an odd twist: it's intended to intercept stations on the 88- to 108-MHz FM band! With some modification, it'll also copy nearby VHF AM aviation band traffic, just the ticket for eavesdropping while watching the activity at your local airfield! FM crystal receiver projects aren't new; this project is based on material from an article that ran back in the 1950s, but with some modern refinements to improve performance. Most of the parts you'll need to build this project can be found in your junkbox, but I'll also provide some mail-order sources for those of you who don't have goodies hoarded away!

Monitoring FM signals is more challenging than tuning in your local AM stations with a simple crystal set. There are two reasons: AM signals are readily detected by a crystal-diode detector, and a simple longwire antenna will usually suffice in providing enough signal level so you can hear at least a few of the local stations. FM is a different animal. A crystal detector recovers both a DC and audio signal from an AM carrier, a simple diode detector will only produce a DC signal component when monitoring a FM signal. Since there are no AM sidebands to mix with the carrier in the detector, none of the audible information is recovered.

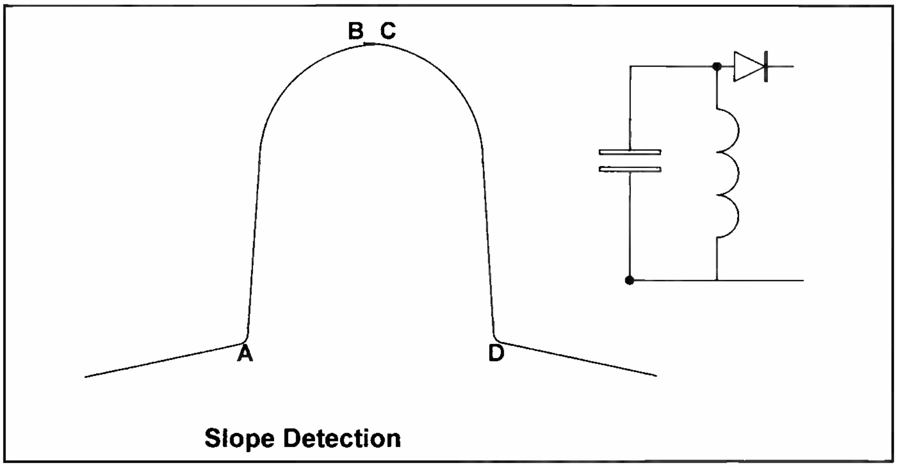

Figure 1. Here's a representative tuning curve for an LC tuned circuit as found in our FM Crystal Set. Slope detection occurs when the FM carrier straddles either side of the tuning curve between points A and B or C and D. An AM carrier would be centered between points B and C for best AM detection.

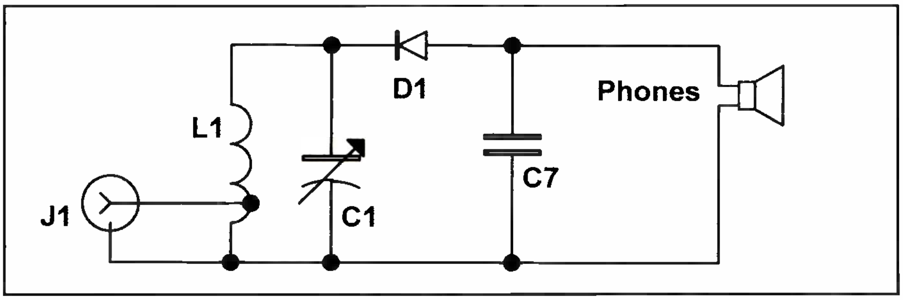

Figure 2. Here's the schematic for the simple and basic FM crystal receiver. FM performance is about 20 dB down from AM signals in the same tuning range. It makes a good companion at the local airfield to monitor AM aircraft and tower traffic in the 120-MHz region.

Slope Detection

So, how's it possible to make a working FM crystal radio? Good question. The answer lies in using a technique called "Slope Detection." FM is simply a carrier being shifted in frequency at an audible rate. The rate of shift corresponds to the frequency of the audio signal while the excursion, or carrier deviation, represents the audio signal level, or amplitude. U.S. FM broadcasters are allowed a deviation of ± 75 kHz, or 150 kHz total. This means we can hear FM signals on a simple crystal radio, providing the tuned circuit has enough Q for the recovered audio to be great enough to drive a pair of headphones. Whereas a crystal detector will detect AM signals without tuned circuits, the FM set relies on a very sharply tuned high-Q circuit to work. If we tune the FM station so it falls on either side of the slope, the modulated FM carrier will be converted to AM since its signal amplitude will vary depending on where it falls on the tuning slope at any given instant.

Many early 2-meter amateur FM repeater operators used AM transceivers, modified with simple external FM modulators, while using slope detection to hear the FM repeaters. Most of these early AM receivers were 15 or 30 kHz wide; the IF bandpass curves easily recovered the ± 5-kHz FM deviation! The trick was to tune to one side of the signal or the other, carefully centering the FM carrier on the center of the tuning skirt, not dead on at the peak! In short: The steeper the slope, the greater the recovered audio for a given deviation.

Figure 1 illustrates how slope detection works. The curve on the drawing represents a typical, but not exact, representation of how a tuned circuit responds as it is tuned across a carrier. Points A and D denote the bandwidth of the tuned LC circuit. The areas between points A and B and C and D represent the slopes of the tuning curve. Ideally, an AM signal would be tuned so its carrier is at the apex, or center, of the tuning curve. If we center the carrier frequency of the FM transmitter to the same spot, the carrier will swing over both slopes, and any recovered audio would be canceled. Instead, the "sweet spot" is where the FM carrier falls near the midpoint of the slope between A and B, or likewise on the slope of the curve between points C and D. The AM detector output will then swing in correspondence with the aural information carried on the FM carrier.

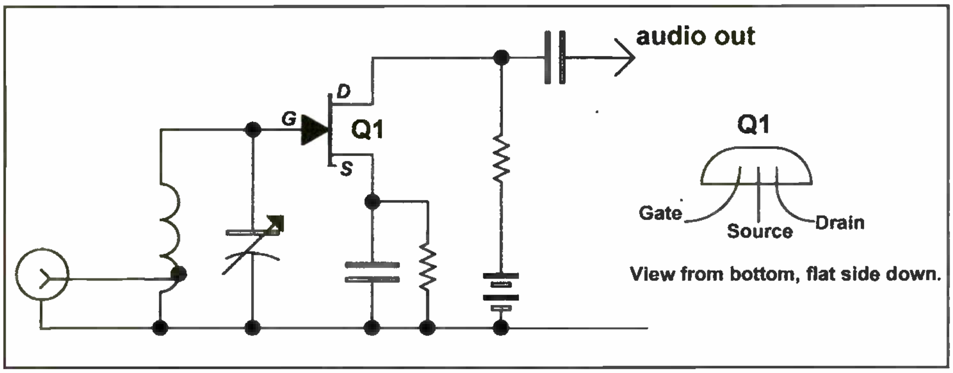

Figure 3. The infinite-impedance detector offers several advantages over a germanium diode detector. It has gain, doesn't load the tuned circuit, and it features better weak-signal detection.

Ideally, for maximum audio recovery, the FM deviation will "sweep" over the largest area of one side of the slope as is possible. It's hard to match the tuning curve to the FM deviation, and also any non-linearity in the slope along the rise or fall on the curve (a straight line, like a saw tooth, is ideal) will cause distortion. Too wide a slope and the FM deviation will fall on a very small area on the slope, giving very poor recovered audio. Too much deviation, where the carrier swings beyond the cutoff and into the opposing slope, yields large amounts of distortion on the recovered audio. In practice, these sets do work, but don't expect anything approaching high-fidelity reception!

The Simplest Receiver

There are several key elements needed for this project to work, most importantly, having a very strong local FM station or two or listen to. Most urban locations are so blessed. You'll also probably need a good outdoor FM or TV antenna to capture enough signal for the set to work, and it's a big plus if your system includes an inline VHF/FM preamplifier.

Let's get started by looking at **Figure 2**, the starting point for our project. It closely resembles the primitive FM crystal set projects shown in those early magazines. The heart of the receiver is the high-quality tuning capacitor and inductor. The tuning capacitor can be salvaged from an old FM tuner or converter, if need be. The tuning capacitor (C1 in the drawings) is a 35-pF capacitor with a built in 3:1 vernier dial reduction, and was salvaged from my junkbox. Any capacitor with a maximum capacity of 25 to 45 pF will work.

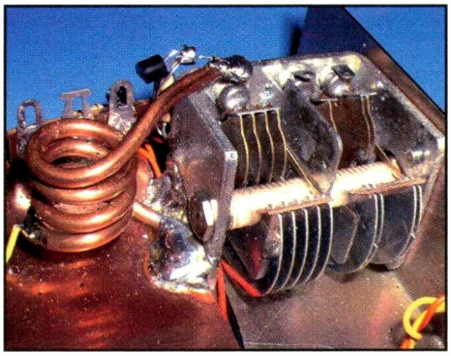

Photo A. The tuning capacitor and coil form the heart of the crystal set. The Q of these parts determines the tuning slope, and ultimately audio recovery on FM signals. The capacitor should be the best quality available. Often military surplus capacitors with ceramic insulation are available on the surplus market or at hamfests. If you can find one, use it. The tap point for the antenna is visible. Use a short lead between the tap and J1 connector.

The set tunes rather broadly, thus the vernier tuning isn't a necessity, but it is nice to have. One source for the cap is Fair Radio; its 3G-25 variable cap has a vernier drive and three sections with 4-25, 4-20, and 3-17 pF. Two of these sections could be paralleled to yield a larger tuning range. The inductor (L1 in all drawings) is wound from a length of #8 copper ground wire, available from any hardware store. You can substitute a smaller gauge if necessary. The circuit Q could be improved by silver-plating the coil—if you have a means to do so try it. Otherwise, I suggest polishing the copper to a sheen, followed by applying a coat of clear Krylon to prevent tarnishing. The coil consists of six turns; I used a half-inch wood dowel as a form to wind the coil. The antenna tap is made at one-half turn above ground on the coil. **Photo A** shows the tap point and also the tuning capacitor. **Table 1** lists the parts values and sources for all the components used in three versions of the receiver shown in this month's column.

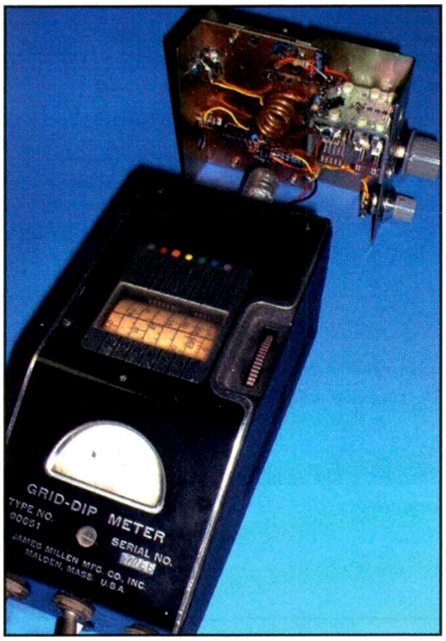

Photo B. Here's the Millen Grid Dip Meter in action! When tuned to the resonant frequency of the capacitor and coil, the meter on the Millen will dip to a lower value. Fifty or 60 years ago, every ham owned one of these useful pieces of test equipment. That was a time when hams built their own gear. Now using one has become an archaic art.

I know finding parts can be difficult. I listed a few resources in **Table 1**, including one for a basic kit, sans headphones, for building the basic FM crystal set. This five-dollar postpaid kit can serve as the nucleus for the more advanced versions shown here as well.

Checking The Tuning Range

A grid-dip meter is the easiest way to check the tuning range of your coil and tuning capacitor combination. A ham friend who is also a builder is a likely candidate to have one of these units in his workshop. **Photo B** shows my venerable James Miller *'dipper* being used to check that my set will tune the entire FM band. These devices are variable oscillators with calibrated tuning ranges. A set of companion plug-in coils allows covering from the broadcast band up through 300 MHz. When the *'dipper's* coil is held near a tuned circuit, and the grid dipper is tuned to that frequency, the grid current in the *'dipper* will drop (or "dip"), as shown on the *'dipper's* built-in meter.

Alternately, a signal generator would also do for checking the tuning range. Compressing or expanding the coil slightly will shift the range: squeezing the coil will allow lower frequency tuning, while

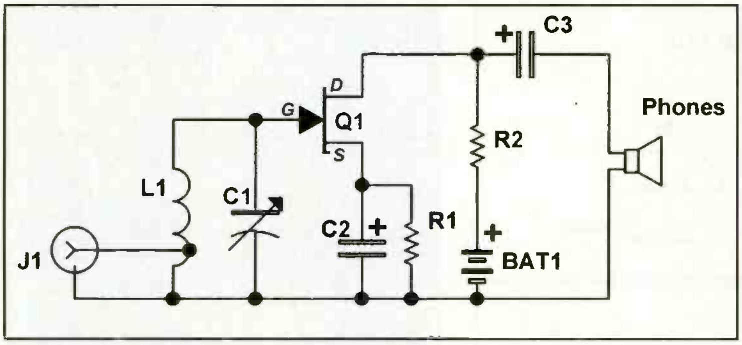

Figure 4. Adding the infinite-impedance detector to the basic FM crystal receiver might make the difference between hearing a station or not!

Photo C. The leads for the infinite-impedance detector should be kept short and direct for VHF. The J-FET leads and associated parts are mounted to the capacitor terminals and frame. →

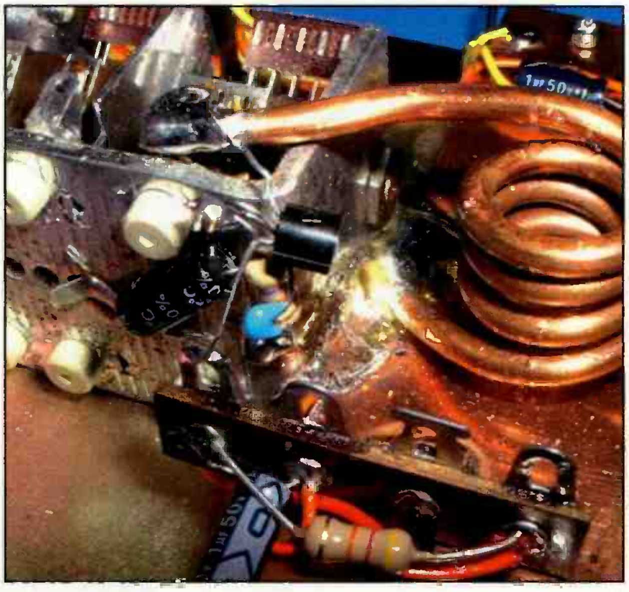

Table 1. Parts List And Recommended Suppliers

L1	#8 copper wire, see text
C1	variable capacitor, 5 to 50 pF max. with knob
D1	1N34A germanium diode, or equivalent
J1	RCA phono jack or suitable RF connector for antenna
C2	10 µfd @ 16 volts electrolytic
C3	10 µfd @ 16 volts electrolytic
C4	10 µfd @ 16 volts electrolytic
C5	0.1-mfd ceramic capacitor @ 16 volts
C6	220 to 470 µfd @ 16 volts electrolytic
C7	500-pF ceramic disc or silver mica capacitor
R1	47k-ohm 1/4-watt resistor
R2	47k-ohm 1/4-watt resistor
R3	10k-ohm variable potentiometer, audio taper/with knob
BAT1	9-volt transistor battery with battery clip
Phones	2000-ohm vintage headset or crystal earpiece
Speaker	8-ohm speaker in enclosure
Q1	J-FET, type J-310
U1	LM-386 DIP package 8-pin audio IC

Resources

- Dan's Small Parts, PO Box 3634, Missoula, MO 59806-3634, (406) 258-2782, www.danssmallpartsandkits.net

This is an excellent resource for all components (ICs, FETs, small parts, and variable caps). Dan's features low prices and is set up to deal with home experimenters. Dan's offers an online catalog and ordering system. Minimum order might apply, check with seller.

- Bill Turner, 1117 Pike Street, St. Charles, MO 63301, (636) 949-2210, www.dialcover.com

Bill offers a kit and copy of the original vintage magazine article for the FM Crystal receiver for $5 postpaid. No frills or headphones, just the basic parts. No minimum order.

- Fair Radio Sales, 2935 St. Johns Road, PO Box 1105, Lima, OH, (419) 227-6573, www.fairradio.com

Surplus stalwart Fair Radio has a nice three-section variable (catalog #3G-25) that will work well. Price is $6. They also offer some high-Q ceramic military versions, but these are more expensive. Minimum orders might apply, check with seller.

expanding the coil will increase the higher end of the tuning range. A 40-pF variable capacitor should tune from 80 MHz through at least 130 MHz into the aircraft band. At worst, you may have to wind a new coil with a turn added or removed before you're able to tune the desired range. Adding a full or partial turn will lower the frequency. **Photo B** shows my Millen dip meter in action.

Improving The Detector

Crystal diodes make poor detectors. Germanium diode detectors don't work well on weak signals; for a weak enough signal they act more like resistors than diodes. **Figure 3** shows a circuit for a neat device called an Infinite Impedance Detector. This detector offers some gain and doesn't load the tuned circuit as much as a diode detector would; hence the Q and selectivity are also improved. The J-FET gate adds some capacity to the tuned circuit, and this can reduce the high-frequency tuning by several MHz.

Photo C shows how the J-FET and associated parts for the infinite-impedance detector are mounted using short leads. The gate is tied directly to the tuning capacitor stator terminal, and the FET's *source* lead bypass caps and biasing resistor are tied directly to the capacitor frame for good VHF performance.

Figure 4 shows how the infinite-impedance detector is incorporated in the FM crystal receiver. Parts values are given in **Table 1**. The detector will drive a pair of vintage high-impedance headphones (2000 ohms or better) directly. I use a pair of military sound-powered headphones with a matching transformer. Adding the

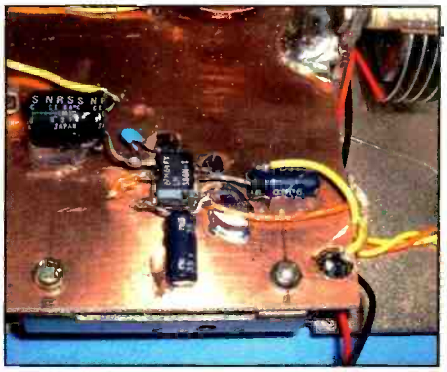

Photo D. The audio amp IC is mounted using "dead bug" construction techniques. The larger parts are glued to the pc board to alleviate stress on the IC leads.

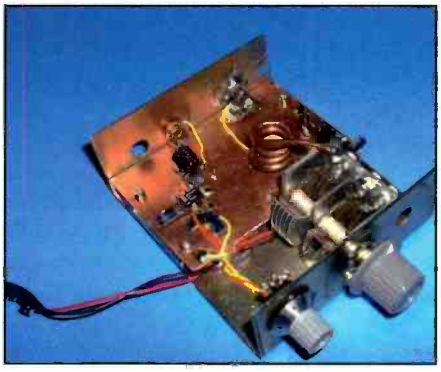

Photo E. Here's a top view of the completed receiver. Ugly, but it works—and it was intended as an experimental prototype to begin with.

infinite-impedance detector allowed me to hear some FM stations from my rural home when using some high-gain tower-mounted VHF antennas.

You might want to try this detector on your AM BCB crystal sets; it makes a big improvement!

Adding An Audio Amplifier

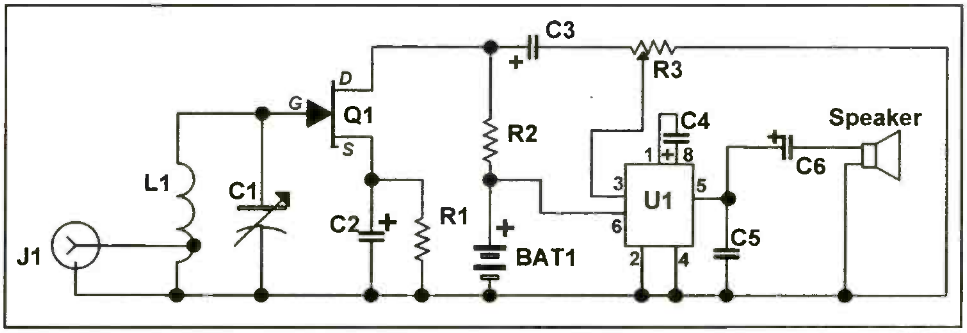

Figure 5. Adding the high-gain audio stage gives loudspeaker volume on strong stations.

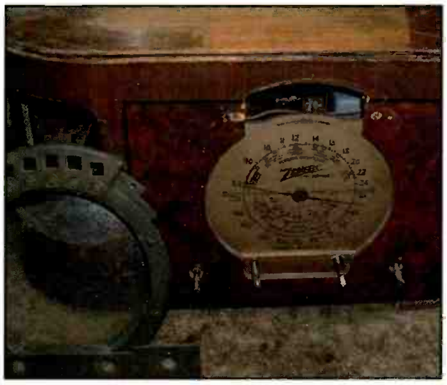

Photo F. This is my 6S321 Stars-and-Bars Zenith! Sorry for the mix up! As you can see, this cabinet needs a bit of work. I'd started scraping the old lacquer off the top when the photo was taken. Chemical strippers risk damaging the photo-etched lacquered patterns on the front, which I intend to restore and save.

Figure 5 shows the final phase of the FM crystal receiver. A small 8-pin LM-386 audio IC gives enough gain for loudspeaker level volume, or to drive more contemporary low-impedance headphones. I mounted the IC and supporting parts "dead bug" fashion on the pc board. If you're not familiar with this breadboarding technique, the parts are soldered directly to the copper side of an unetched pc board for ground, and are self-supporting or glued in place as needed. (ICs were sometimes mounted on their backs, looking like little dead bugs; hence the name.)

It's not very pretty, as shown in **Photo D**, but it works and is good for quickly prototyping circuits on the fly. I think some builders call this technique *Ugly Board* construction. Hot glue assisted in mounting and supporting the IC and larger components to the board. **Photo E** shows an overall view of the entire receiver. Again, it isn't pretty, since I considered this to be an experimental project. An off/on switch is needed to preserve battery life.

An Apology

Our opening photograph in the July 2004 issue purportedly showed the deplorable condition of my Zenith 6S321 cabinet. What accidentally got forwarded for publication was a photo of a nicely restored example from the webpages of Mike Urban's Internet site at www.urban-antiqueradio.com. I had saved his photo as an example to follow when restoring my Zenith. Hopefully **Photo F** is the correct photo. Anyway, visit Mike's website to see some nicely restored radios in his collection, and to check out a few he's offering for sale.

Well, that's it for this month. If you build one of these sets, I'd like to hear about it, and photos would be great! Let me hear about your other projects, as well. Until next month, keep those soldering irons warm and those letters coming in! ■

If you enjoy radio communications you'll love....

POPULAR COMMUNICATIONS

The World's most authoritative monthly magazine for Shortwave Listening and Scanner Monitoring. Read by more active listeners world-wide.

*Get fast home delivery of **Popular Communications** and save $30.93 a year over the newsstand price. Save even more on 2 or 3 year subs.*

☐ 1 year - 12 issues $28.95 (Save $30.93)

☐ 2 years - 24 issues $51.95 (Save $67.81)

☐ 3 years - 36 issues $74.95 (Save $104.69)

SAVE UP TO 58%

SUBSCRIBE TODAY!

Canada/Mexico–One year $38.95, two years $71.95, three years $104.95; Foreign Air Post–one year $48.95, two years $91.95, three years $134.95,

Payable in US dollars onlyNote: Allow 6-8 weeks for delivery of first issue.

Mail orders to:
Popular Communications, 25 Newbridge Rd, Hicksville, NY 11801
FAX 516-681-2926

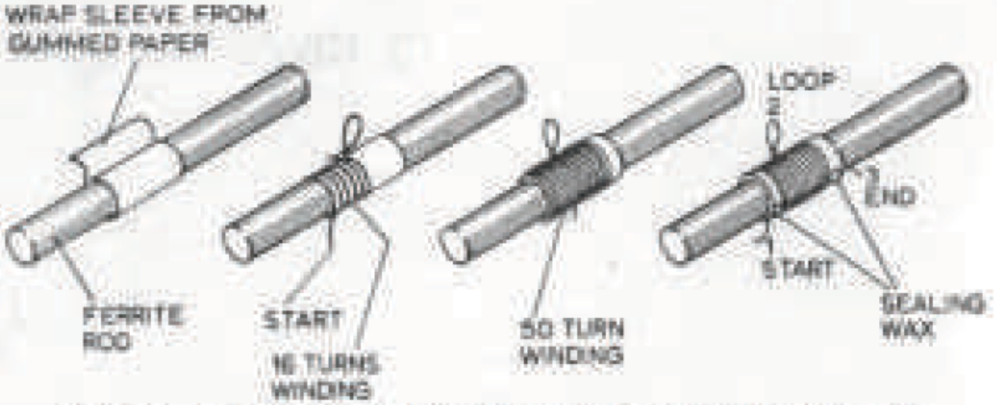

Fig. 3.1 Aerial tuning coil wound from 38-s.w.g. wire on a ⅜-in. diameter ferrite rod.

Fig. 3.1 details the construction of a suitable coil for a tuned circuit, matched to a ferrite rod of ⅜ in. diameter and about 4 in. long (see Chapter 4 for alternative coil windings on different rod sizes). If a rod is purchased longer than 4 in., it can be used as it is, or cut down by marking around with a file and then breaking off the surplus length.

Cut seven 1-in. lengths of gumstrip. Moisten one and wrap around the rod gummed side up. Now add about another half a dozen wrappings of similar length over the first, this time with gummed side *down* to form a reasonably rigid tube. Make sure that the paper tube is a *sliding* fit on the ferrite rod and leave to dry thoroughly (preferably removed from the rod so that it cannot become stuck to it).

When the paper tube is quite dry it should be rigid, when the coil windings can be applied. The wire to be used is 38-s.w.g. enamelled copper wire, the number 38 referring to the actual diameter of the wire according to the standard wire gauge (s.w.g.).*

Starting about ¼ in. in from one end of the paper tube, wind the wire carefully round the tube, with each turn tight against the one before it, until sixteen full turns have been completed. Then make a loop in the wire, as shown, and carry on winding, with succeeding turns touching, until fifty turns in all have been completed. The two loose ends of the coil (the start and finish) can be secured with a dab of sealing wax whilst the projecting loop can be twisted together (e.g. by putting a pencil through the loop and twisting up). Cut off the loop, leaving about ½ in. protruding from the main coil, bare the wire ends and solder together. This forms point 2 on the coil; the start is point 1, and the end point 3–see Fig. 3.1. It will be easy to remember these without marking since the loop or tapping point (2) comes much closer one end (1) than the other (3).

---

* American wire gauge (AWG) is about two sizes smaller than the British standard wire gauge (s.w.g.). For example, 38 s.w.g. is approximately the same as 36 AWG.

Cut a panel of *Paxolin sheet* to about the size shown in Fig. 3.2, using a hacksaw.* On this secure a *tag strip*, as shown, and drill a hole to mount a miniature or small-size 500 pF *variable capacitor.*

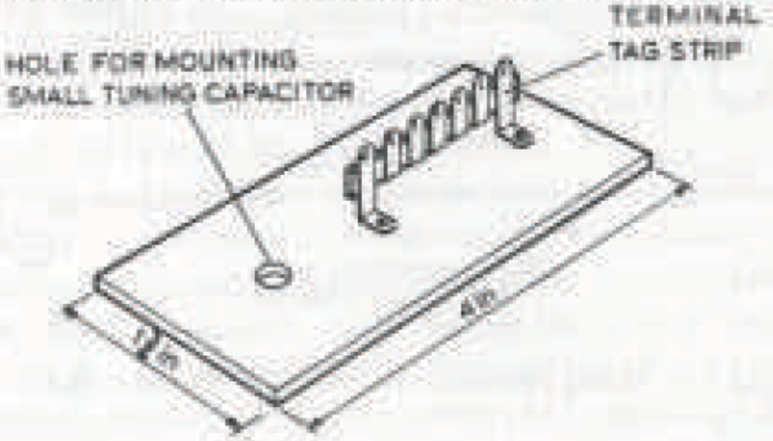

Fig. 3.2 Paxolin panel and tag strip for crystal set circuits.

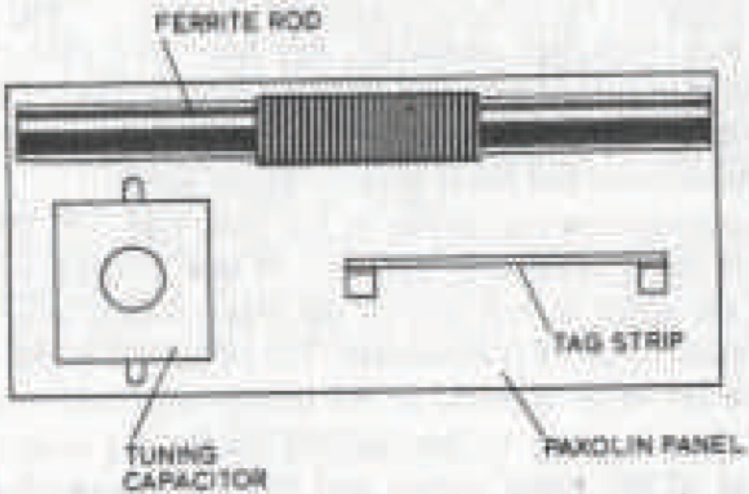

Fig. 3.3 Layout of components on Paxolin panel.

The aerial coil is then mounted on the panel as shown in Fig. 3.3, gluing the coil on to the Paxolin with two or three dabs of sealing wax, or some other suitable adhesive. *Note:* the ferrite rod must be free to slide in the paper tube for 'tuning' adjustments.

Virtually any miniature germanium or silicon diode will be suitable for the detector. Recommended types, which are readily available, are 1N34 and 1N914.

Earphones must be of *high-impedance* type, which need not be expensive to buy, and the higher the impedance the better the reception.

---

* A phenolic or fiberglass sheet may also be used.

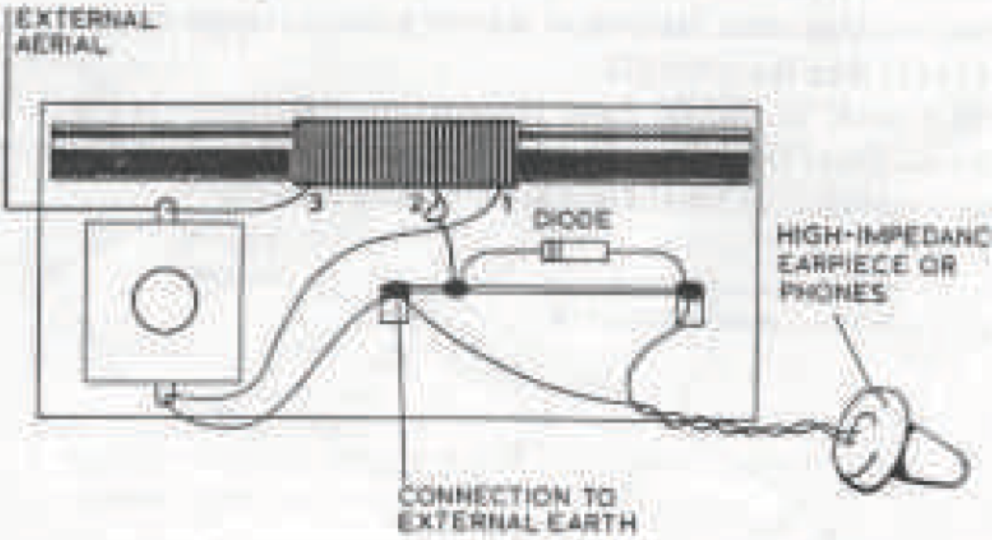
EXTERNAL AERIAL
3
2
1
DIODE
HIGH-IMPEDANCE EARPIECE OR PHONES
CONNECTION TO EXTERNAL EARTH

Fig. 3.4 Wiring connections to complete the crystal set.

Alternatively a deaf-aid type earpiece can be used; this will not give the same volume or quality of reproduction as headphones, but is a less expensive component. This should preferably be of the *high-impedance magnetic* type, with high sensitivity. Any high-impedance earpiece will suffice, but if of crystal type will require a resistor, connected across it, to complete the circuit. This will reduce the amount of current flowing through the earpiece and lower the strength of signal.

Wiring connections are shown in Fig. 3.4. End 3 of the aerial coil (the end of the 50-turn coil) connects to one terminal of the tuning capacitor, the aerial or 'hot' end of the tuned circuit, and the point to which an external aerial is connected. The other end of the coil (end 1) connects to the other terminal of the tuning capacitor, from which an additional wire is taken to the first tag on the tag strip. This is the 'earthy' end of the tuned circuit, and the point to which an external earth is connected. Leave plenty of slack wire between the coil and tuning capacitor.

The other connections are then as follows:

(i) Tapping point of the coil wire bared and connected to the second tag.
(ii) The diode also soldered to this same tag, and to any other free tag.
(iii) Headphone (or earpiece) leads to the 'earthing' tag, and to the 'free' tag to which the above diode has been connected. All connections should be made with *soldered* joints.

The wiring-up can be checked against the *circuit diagram* shown in Fig. 3.5 (ignoring the components shown with broken lines). The set should now be 'working'.

In areas of strong signal strength, no external aerial or earth connections should be necessary. Performance will, however, be improved in any area by attaching an aerial wire (which can be any thin wire, e.g. using the same wire as for the coil winding), of up to 160 feet. The longer the aerial the better the reception, provided it is led away from the receiver to as *high* a point as possible.

An earth connection may further improve aerial performance; by this we mean a connection to some conductor positively in contact with the ground (preferably buried), an excellent example being a metal water-pipe. Thus, if an earth connection is found to be necessary (or you want to try one to see how performance is affected), connect a wire from the 'earthing' tag on the receiver to a convenient water-pipe.

This question of obtaining a *good* aerial and earth is a most important one in areas of poor signal strength. Linking up to a television aerial is often a good plan, since TV aerials are also usually mounted as high as possible. If bare wire is used, it is also important that the upper (free) end of the aerial is not made fast to something which could produce an earth connection (e.g. a damp tree), or at least is suitably insulated from such a support. String is not an efficient insulator; that, too, can conduct when wet.

Quite good results are often obtained by using the springs of a bed as an aerial, in which case an earth connection is usually not necessary. Sometimes, too, when other attempts to yield a good signal strength in

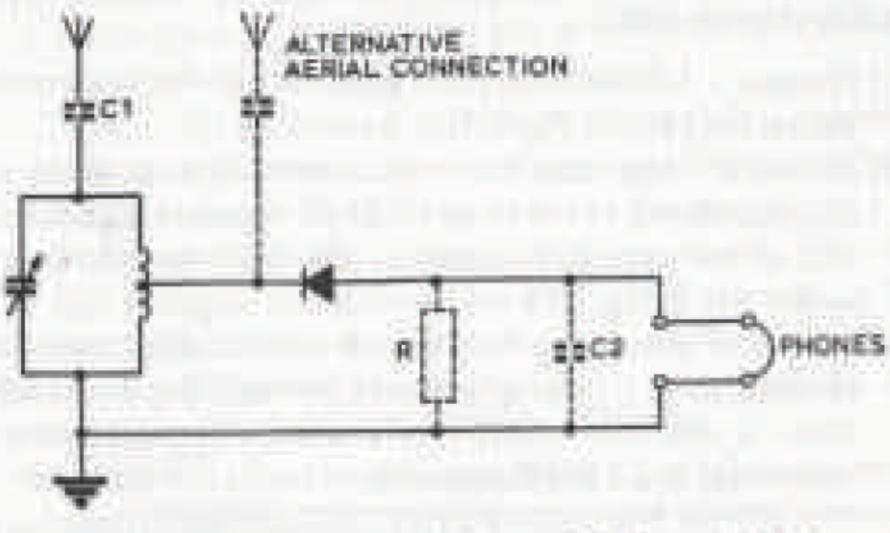

Fig. 3.5 *Circuit 1.* Circuit diagram of basic crystal set.

the aerial have failed, connecting the *aerial* side of the tuning coil to a *good earth* (a water-pipe) can produce better results, the normal earth connection being left off.

Tuning

The receiver is adjusted as follows. Turn the tuning capacitor to fully close the vanes, then open about half a turn on the spindle or knob. (If you are using a trimmer as a tuning capacitor, screw right down and then open half a turn.)

The tuning coil should now be slid up and down the ferrite rod (the coil leads were left a fairly long time to give the necessary freedom of movement) until BBC Radio 3 is heard.*

It may be necessary to slightly alter the adjustment of the variable condenser to tune in to this programme. Also, because of the ferrite-rod aerial the set will be directional, that is, the signal strength received will depend to some extent on the direction in which the aerial rod is pointing, so position the set to pick up the maximum volume.

Having established the best position of the tuning coil on the rod to receive BBC Radio 3, fix permanently with a dab of sealing wax. You should then find it possible to tune in to further stations by altering the setting of the variable capacitor—e.g. typically Radio 4 in about the middle of the capacitor travel and Radio 1 towards the other end.

Any reception you get will almost certainly be very weak and (unless you live close to a broadcast station) you can really feel satisfied if you get any station at all at audible strength. But it is surprising how, sometimes, even quite distant stations can be heard. Also you can often improve the reception and listening strength by quite simple modifications. Try these in order:

(i) Connect a 1,000-pF capacitor across the headphone (earpiece) connections (C2 in Fig. 3.5).

(ii) Instead of connecting the external aerial directly to the tuned circuit, connect one lead to a 220-pF capacitor, and the other end of the capacitor circuit to the 'hot' end of the tuned circuit (C1 in Fig. 3.5).

(iii) Instead of connecting the external aerial to the tuned circuit, connect to the tapping point of the coil (tag to which the diode is also connected). Try a direct connection, and also connecting in a 220-pF capacitor.

---

* Tune to an AM broadcast station near the lower end of the dial—close to 550 kHz.

(iv) Try connecting a 1.2-k resistor (or higher value) across the phone connection (R in Fig. 3.5). You may be using the wrong type of phones or earpiece, which do not provide a proper load or complete the circuit.

If there is a complete lack of response, check for faulty wiring-up. A more likely cause, however, is lack of an external aerial or earth connection in an area where these are strictly necessary for adequate reception; or an inefficient aerial (too short) or poor earth connection (bad electrical contact to a good earth point, or connection to a bad earth point).

Another possible cause of apparent failure may be too much outside noise entering the ear so that it is impossible to detect the very weak radio signal as it is being tuned in. Headphones are better than a single deaf-aid type of earpiece in this respect but, in any case, a really quiet room is virtually essential for initial setting up and tuning adjustments. Also, if your adjustment of the tuning control is too coarse, you may completely miss the setting for the station you are looking for, without realizing it.

Reception will also tend to vary with weather conditions. Some days it may be so poor that what was normally a strong station is hardly heard at all. The simple basic receiver has many limitations but, since it costs very little to construct and nothing at all to operate, this must be regarded as inevitable.

Providing you can hear something–even if too weak a signal to distinguish properly–you can certainly improve the performance of your basic set by further experimentation with tuned circuits (see Chapter 4) and/or the addition of amplification to the circuit. You can also try other types of basic crystal set, as described in the following projects.

Circuit 2 (Fig. 3.6) is identical to *Circuit 1* except that, instead of a diode, a transistor is used as a detector. Only two of the transistor leads are connected–the emitter (*e*) connection to the tapping part of the coil, and the base (*b*) to the 'earthy' end of the circuit. The collector lead of the transistor is ignored (bend it out of the way so that it cannot accidentally short out the other leads).

You can try almost any type of low-cost AF transistor; recommended types are OC42 2N370, 2N2925, 2N5088.

With the addition of two more components, Circuit 2 can be modified to work the transistor both as a detector and an amplifier, to give stronger signals through the headphones. Using the same transistor type

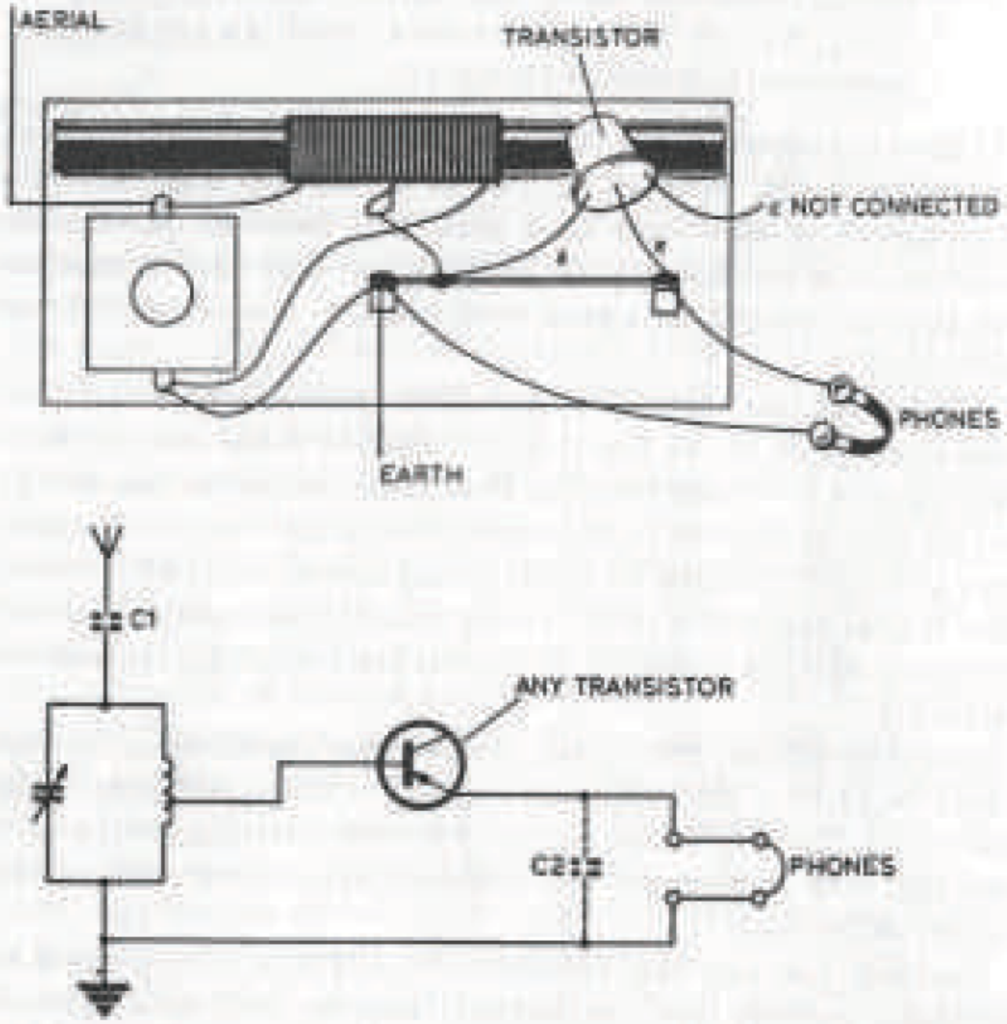

Fig. 3.6 *Circuit 2.* Crystal set using transistor instead of a diode.

as above (or near-equivalent) resistor R should be 15 k and capacitor C3 1 μF or higher.

This time a battery is also required to supply power for the transistor to work as an amplifier. This can be from 1.5 V up to 9 V. Remember the rules for polarity of connection: those shown in Fig. 3.7 are for a p-n-p transistor; an n-p-n transistor would need the battery connected the opposite way round. Battery polarity also affects the connections of capacitor C3 (if an electrolytic or polarized type).

Experiment further by trying the effect of using additional capacitors in the circuit, e.g. C1, 220 pF; C2, 0.001 μF (try other values as well); C4, 0.001 μF (try other values as well).

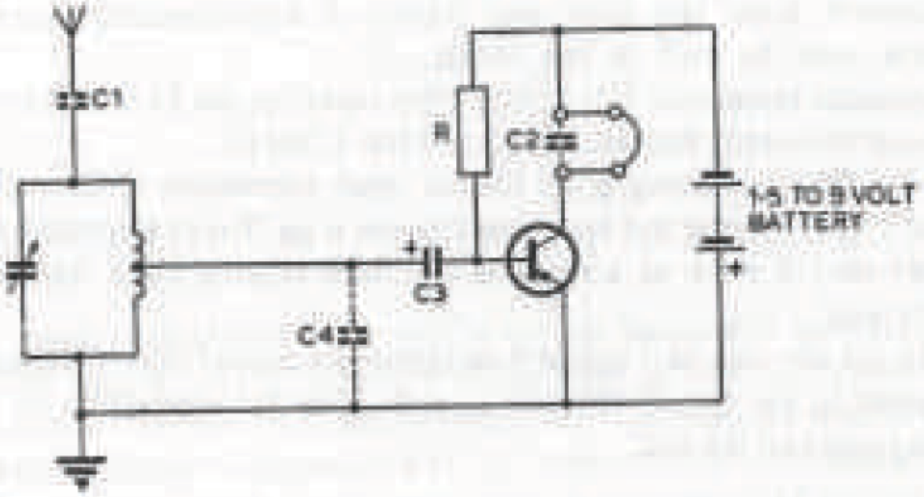
C1
R
C2
+
C3
C4
1·5 TO 9 VOLT
BATTERY
+

Fig. 3.7 *Circuit 3.* Crystal set with amplification.

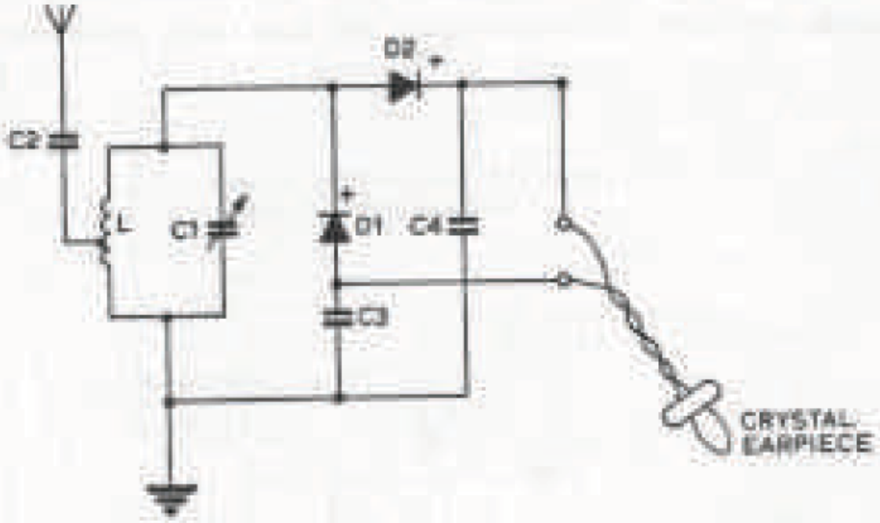
D2
+
C2
L
C1
+
D1
C4
C3
CRYSTAL
EARPIECE

Fig. 3.8 *Circuit 4.* Double diode crystal set.

This circuit incorporates 'voltage doubling' to improve the signal volume and should give better performance than a single diode circuit. L and C1 are the usual tuned circuit, but an unusual feature is that the *aerial* is connected to the coil tapping point. Any type of germanium diodes can be used (they should preferably be the same), making sure to connect them the right way round. A high-impedance *crystal* earpiece must be used in this circuit.

Capacitor values are: C2, 220 pF (this capacitor can be omitted–try with and without in the circuit); C3 and C4, 1,000 pF.

The optimum tapping point for the aerial connection to the coil L is best found by trial and error (see Chapter 4 on 'Tuned Circuits'), but the set should work at a nominal one-third tapping point from the 'earthy' end.

This set can also be tried with conventional tuned-circuit coupling–*i.e.* aerial to the top of the coil L, and diode D2 connection to the tapping point on the coil.

359

Creative SCIENCE & Research

Copyright 2002 - 2004 Creative Science & Research

PO BOX 557 New Albany, IN. 47151 USA

Information Big Brother & the Oil Companies do not want you to know!

Free Electricity

From the Phone Co.

Time is running out

LTF
230V330µF
40/105/56

www.FuellessPower.com ● **www.Fuelless.com**

Cover Page

Creative SCIENCE & Research

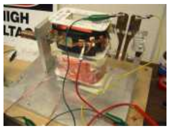

359

Copyright 2002 - 2004 Creative Science & Research

PO BOX 557 New Albany, IN. 47151 USA

Question: Is there enough useable energy from my phone line to light a 100 watt light bulb?

Answer: Yes, but only for a short time. If you are looking for a way to light up your home then build our fuelless engine.(see our web site at: www.fuellesspower.com or write us for a catalog.) This engine is powerful enough to run a 10 kw 120 vac 60 hz generator which can run your home.

You can still use your incoming phone line as a useful light source. You can get enough light to light up a room or two by using a diode and a large 100 volt capacitor bank to collect the 50 vdc. What you then want to do, is to use a high efficient lighting source, such as ultra bright white diodes, White neon lighting or flouresent bulbs. (great for emergancy lighting in any storm.)

The 50 vdc current you have collected will not light Neon or Flourescent bulbs until the voltage is much higher. 300 to 5,000 vac, You can buy a 15,000 volt neon transformer and pulse the 50 vdc into the primary using a wind up spring pulse switch (food timers) or build a low milliamp transistor type pulse generator, this will turn the DC to AC which you can then collect from the secondary out put of your transformer. You can also build our #363-a step up DC (No amperage loss! 100% eff. Using Photo caps) Or just for a quick fun experiment charge 20 - 50 volt capacitors, with each capacitor having about 4700 uf (micro farads), simply fully charge each capacitor by hand then connect them in series using Radio Shack allegator clips (wire), just like batteries, Negative to Positive, **Warning! Use rubber gloves when working with High Volatge, High Voltage can kill!**

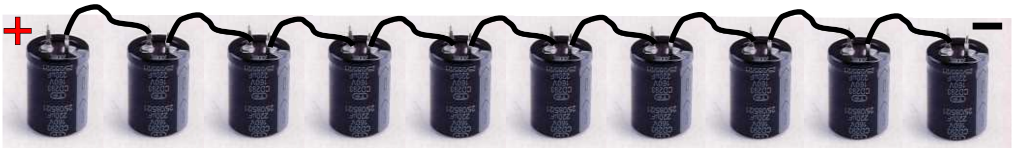

Once you are done with connecting them in series, use a small low amp (Miliamp) isolation transformer which you can buy from any electronic supplier or Radio Shack store. (Search for Electronic Suppliers on web search engine) connect the negative to the primary and the positive you will need to pulse the + on and off very quikly! the faster the pulse the better the output of energy! Connect the secondary leads from the isolation transformer to the ends of a neon or flourescent light bulb, as you pulse the positive to the primary of the transformer the bulb will light very brightly!

Input 12- 24 DC
on 120 VAC

15,000 v Neon Transformer

Or you can pulse this homemade Tesla Coil to light a bulb!

Copyright 2002 - 2004 Cr

Again, there are many ways to get free electricity from your phone lines. (also great for out door lighting.)

Updated 2004: The below method seems to be the best way so far to collect and store the 50 vdc energy. This method should be done at night or when the phone line is not in use.

To obtain 1 amp at 15 vdc we use a 10 volt capacitor to step down the 50 VDC that is coming from the lines. This capacitor is a very special capacitor, try to get the same type if they still sell them. I think it was made in Japan. I wish we had them in stock we would sell them. Photo caps are very special and might work as well.

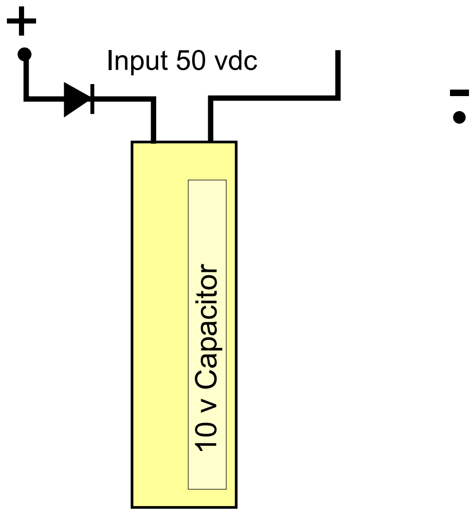

Electrolytic Capacitor
10 v x 1000 uf 105 C
Made by Elna part # 8643

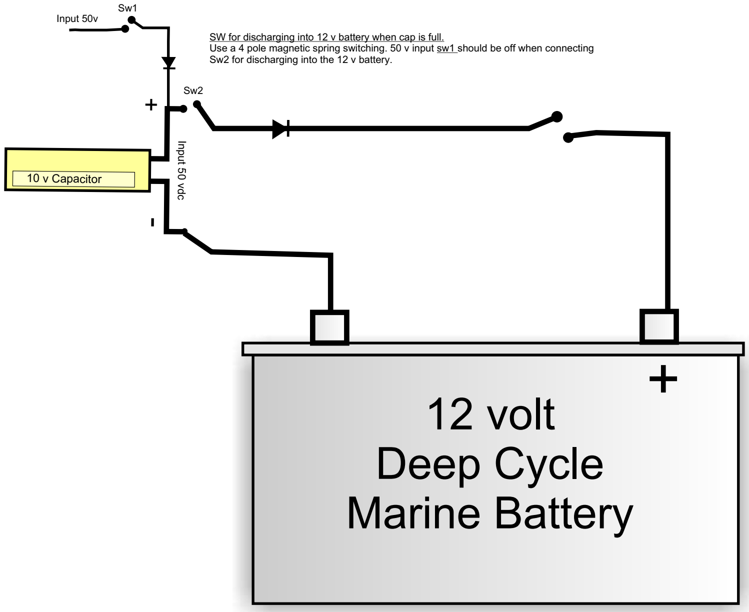

SW for discharging into 12 v battery when cap is full.
Use a 4 pole magnetic spring switching. 50 v input sw1 should be off when connecting Sw2 for discharging into the 12 v battery.

Power diode works best!

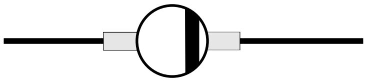

G1 G4g

Let cap charge every 4 to 6 seconds, you may want to use a spring wind up timer for switching or use transistors as dc pulser to control 4 pdt relays which all can be powered from free energy from the earth method, *(See our free energy from the earth plans #459 $29.95 plus $3.95 shipping.)* Or you can use small magnetic reed switches (about 10) placed and glued around a kitchen timer. Glue a small 1/4 diameter neodymium magnet to the rotating handle.

Copyright 2002 - 2004 Creative Science & Research

High Voltage Care

Be careful, high voltage or even low voltage with high amperage can kill you. Do not attempt to work on this project unless you are a trained professional in high and low - high amperage..

Remember to wear RUBBER GLOVES and RUBBER SOLE SHOES. Using rubber platex gloves from a dollar store or such, work good for up to 50 volts, any higher voltage you should wear a thicker rubber pair of gloves....

WARNING: discharge all capacitors before you touch and work on them or around them.. If you do not know what a electrolytic capacitor is then we suggest you buy a beginners book in electronics that you can get from from any Radio Shack or electronic store or book store.

You can also use the earth ground to gain more electrical energy into your capacitor by using one leg (the Negative) from the the telephone out put! A diode in the right direction of this leg will do wonders. it's that easy.

The following information is confidential and has copyright and patent pending: Capacitors store and discharge free energy! Radiant energy is stored in between the capacitors plates and the electrons rest on the metal plates themselves. Radiant energy is also created at discharge, but in reality is taking the energy from a very special place that we do not want to disclose at this time.

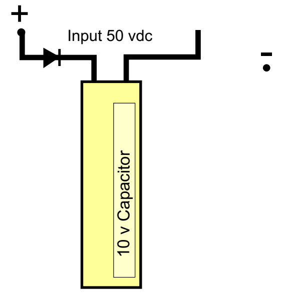

Electrolytic Capacitor
10 v x 1000 uf 105 C
Made by Elna part # 8643

Power diode works best!

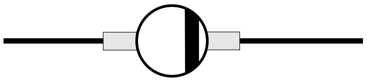

Copyright 2002 - 2004 Creative Science & Research

Part 2

We are going to show you several different methods that you can use to get free electricity from your phone line.

The amount of electricity will vary and depend on which set up you use. most of these set ups can be used at night time while your phone is not in use. using it this way will be the strongest.... by using these methods it will keep the phone 1/2 way off the hook ? (hard to explain)
One of the methods that we will explain is a bit crude....
If you should get an incoming phone call in the middle of the night or day it will charge the capacitors faster and with more amperage input, but with one disadvantage, the phone will not ring! But we have discovered a way past this, but with a slight amperage loss. By using what we invented called a ceramic permanent magnet capacitor blocker filter.. you will see the drawings on the following pages.

I'm sure there are much better ways to charge and gain more voltage and amperage from your phone line than what we have shown you here. but I am sure you will understand that our time is limited, it's a fun project, but we have been developing and researching much better technology than this! If you find a better way then please contact us so we can put it in our plans and share it to the world, we can mentionyour name as being the inventor if you like.

(Timer Switching 4 PDT relays)

The Electricity you will get can be constant when hooked up to multible capacitor banks (Differant caps than the one shown on page 2). Using one 9 volt battery capacitor timer switches, to fire each cap bank at it's proper time. or by using earth energy. (use 2 - 63 v x 4700 uf electrolytic capacitors)

We have actually lit a 100 watt 115 volt AC light Bulb with this set up. 2 caps connected in series. Charge each cap by hand, then connect together in series. But the main purpose is to use this system as a 12 volt deep cycle battery charger that will run a DC to AC 115 volt inverter to use as usable household energy to power a light bulb, an ac motor, TV, VCR etc...

Copyright 2002 - 2004 Creative Science & Research

You can Produce up to 3 amps every 30 seconds by charging a 35 volt x 4,700 uF electrolytic capacitor bank. Connecting large uf capacitors in parallel increases the amperage but slows down the charge time. You can also use them in series to step up the voltage to get 100 volts or 150 vdc if you use three fully charged caps connected in series. The Fuelless engine can be designed as a small 2 hp electric motor and made to run on .05 ma x 300 input and can run from the phone line using our Com /SW method, (See Our Plans) you would need six fully charged caps 50 to 63 v x 4700 uf caps, to run this type of High Effiecient free energy motor. the output will be greater than the input!

You can also use a DC to AC inverter (not the type to run your home)right from the phone wall outlet and then by using a step down transformer you can get 12 volts ac which you will then need to change to dc using a one way diode. The amperage will be greater and can charge a 12 volt marine battery.

Notice:

Be sure to be neat! Put all Components in a safe and convenient plastic box. Use PC board, make it professional looking don't be sloppy! You must keep it safe were Kids as well as some Adults will not get shocked or hurt.

IS IT LEGAL?

We don't Know?
When we first put it in our catalog we were confident that it was. Now we are not so sure? But, it's your phone and you are paying for it! Once it passes into your home it should be yours, as long as you hook it up right and it does not kick back up into the phone line and cause any harm to the phone companies generator lines or breaker boxes etc... that's why we use one way diodes. We have knocked out our phone lines for about an hour or so. when pulsing into the earth you sometimes get a kick back of more current and voltage than what is being collected or inputed.

Remember: it's your phone service! you pay for it every month. The electric you pay for is used to run your telephone, if you wish to use it for something else there should be no law against it.... WE HOPE!!!!????

Copyright 2002 - 2004 Creative Science & Research

OBJECTIVE;

All the methods described in these plans are ways to get cheap Free energy for your home, (#359 is not intended for your entire home).

There are millions of poor people all around the world who can't afford electricity but may have a phone, (this system is much better than using candles). It's fast and cheap to put together for emergancy low light conditions. If one can afford it, we recomend building our Fuelless engine to run a car altntr, or 10 kw home generator system and stop paying a high electric bill, our system can pay for itself. Someday I hope we will be allowed to manufacture our free energy motors then you can simply go down to your local hardware store and buy a free energy motor and generator to power your entire home, but that may not be soon. The only way the open public is going to be able to get a hold of a free energy electric motor is to build one for thereself. It's really not that hard! The cost to build our 2 hp (Fuelless) Free Energy Electric Motor is only about $195. You can use the 2 hp to run a 80 amp car alternator to charge a 12 volt marine deep cycle battery system connected to a 500 to 5000 watt inverter (Which you can buy or make, see our #579 plans $40) to run your lighting, fans, VCR, TV, Refrig, Microwave etc...

Collecting Free Energy from your phone line is also a fun science project! Your friends will love it. Most people do not even know they have 50 vdc running to there phone line, and when the phone rings it sends 3 times more amperage to your charging system. Set your phone up were it will ring itself every 2 minutes for one hour or until your batteries are fully charged. It's basically free! The phone company is not going to charge you extra for it.

Use 4 - 12 v deep cycle marine batteries, (car batteries will not work) and connect them in parralel to increase your amp hr. then connect your inverter or inverters to your fully charged battery bank and enjoy free lighting, free electricity etc... Notice: Do not let your batteries go under a 25% charge, the batteries will last years if you take care of them by not under charging or over charging them. Use a digital volt meter. If you need more amperage, again use our Fuelless Engine or our free energy from the earth.

Copyright 2002 - 2004 Creative Science & Research

Use a DC Volt meter to find the negative and the positive. WARNING! Do not ground the positive to the earth, it will knock out your phone line for about 1 hr!

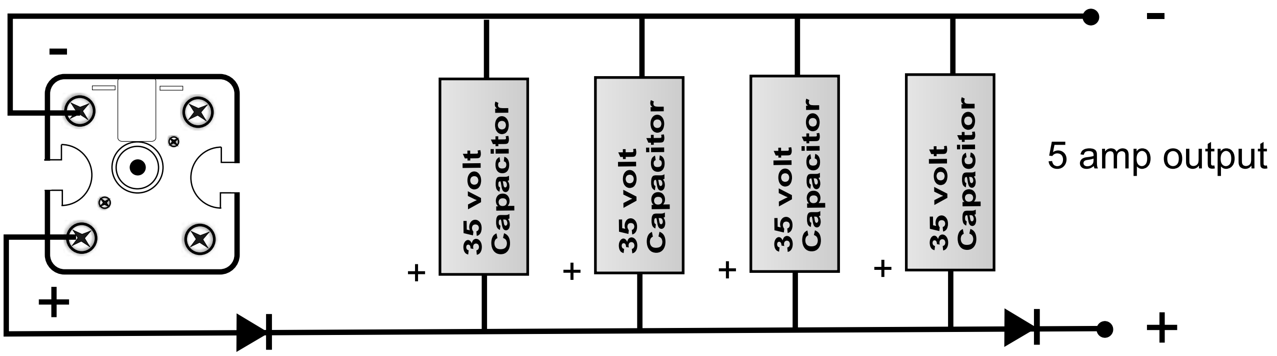
-
-
+
+
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
+
+
+
+
5 amp output

Using both the negative and the positive of the phone line seems to work best. Use 4 - 35 v x 4700 uf electrolytic capacitors. This can be used as a battery charger using the diode as shown or a com / sw switching method as seen in our Fuelless Engine plans or you can simply use a hand wind timer with a strong magnet glued or taped to the handle and using multiple reed switches. Or power a low volt, low milliamp 4 PDT relay using the earth power (see plans #459 $29.95) or our #396 homemade batteries $16.95

Many homes have 2 phone lines, you can use both lines to get 100 vdc with double the amperage output! Use 4 - 150 v x 4700 uf capacitors instead, you can then light a 100 watt bulb, or pulse the 100 volts into a 15,000 volt neon transformer to get a 14,000 vac out put to power several neon bulb tubes or flourescent bulbs.

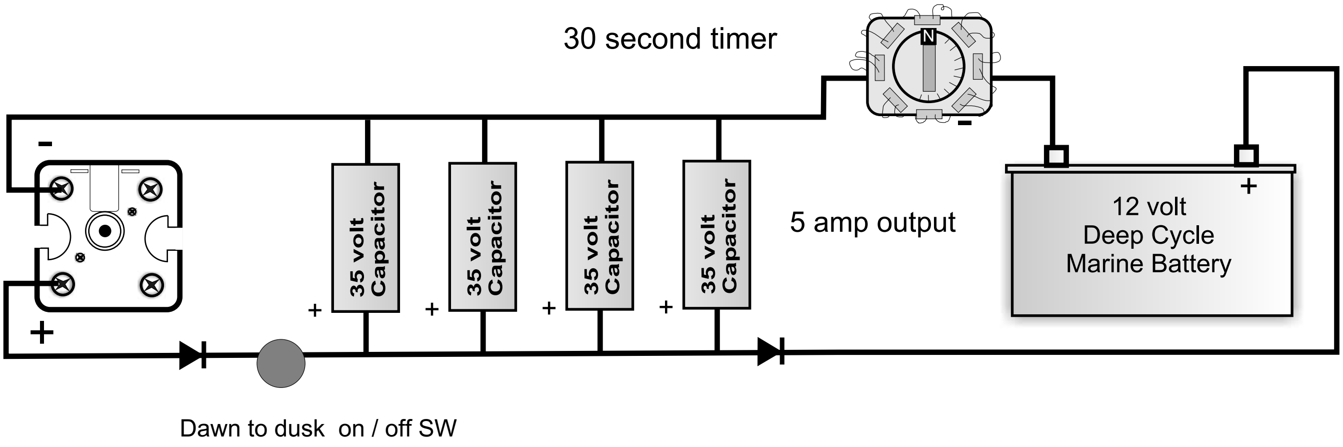
30 second timer
N
-
-
+
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
+
+
+
+
5 amp output
+
12 volt Deep Cycle Marine Battery
Dawn to dusk on / off SW

If you wish to build your own hand spring timer and relays you can buy material at **www.smallparts.com**

PO BOX 557 New Albany, IN. 47151 USA

Copyright 2002 - 2004 Creative Science & Research

Grounding to Earth

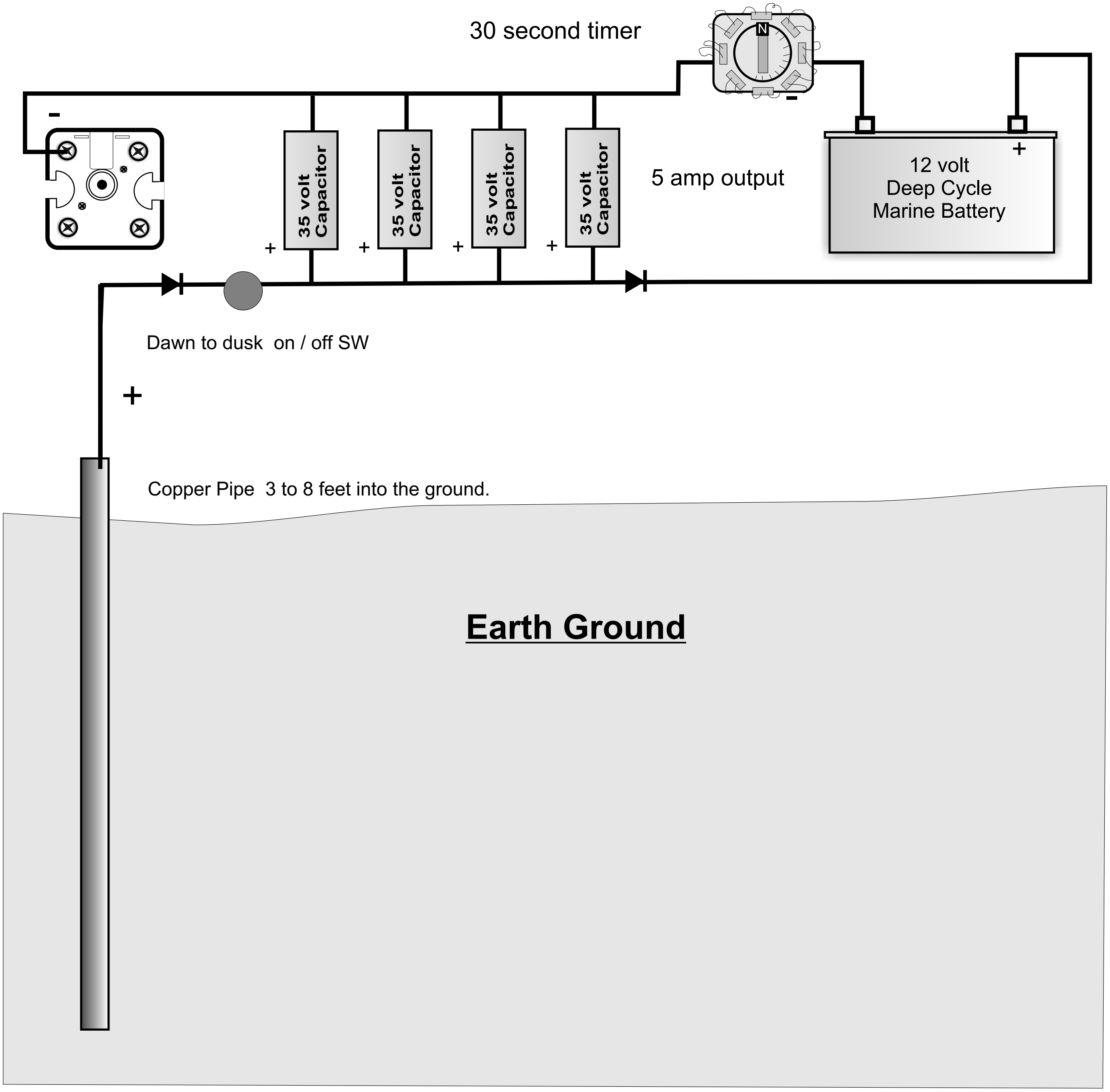
30 second timer
N
-
-
+
35 volt Capacitor
+
35 volt Capacitor
+
35 volt Capacitor
+
35 volt Capacitor
5 amp output
+
12 volt
Deep Cycle
Marine Battery
Dawn to dusk on / off SW
+
Copper Pipe 3 to 8 feet into the ground.
Earth Ground

If you wish to build your own hand spring timer and relays you can buy material at **www.smallparts.com**

Copyright 2002 - 2004 Creative Science & Research

Grounding to Earth

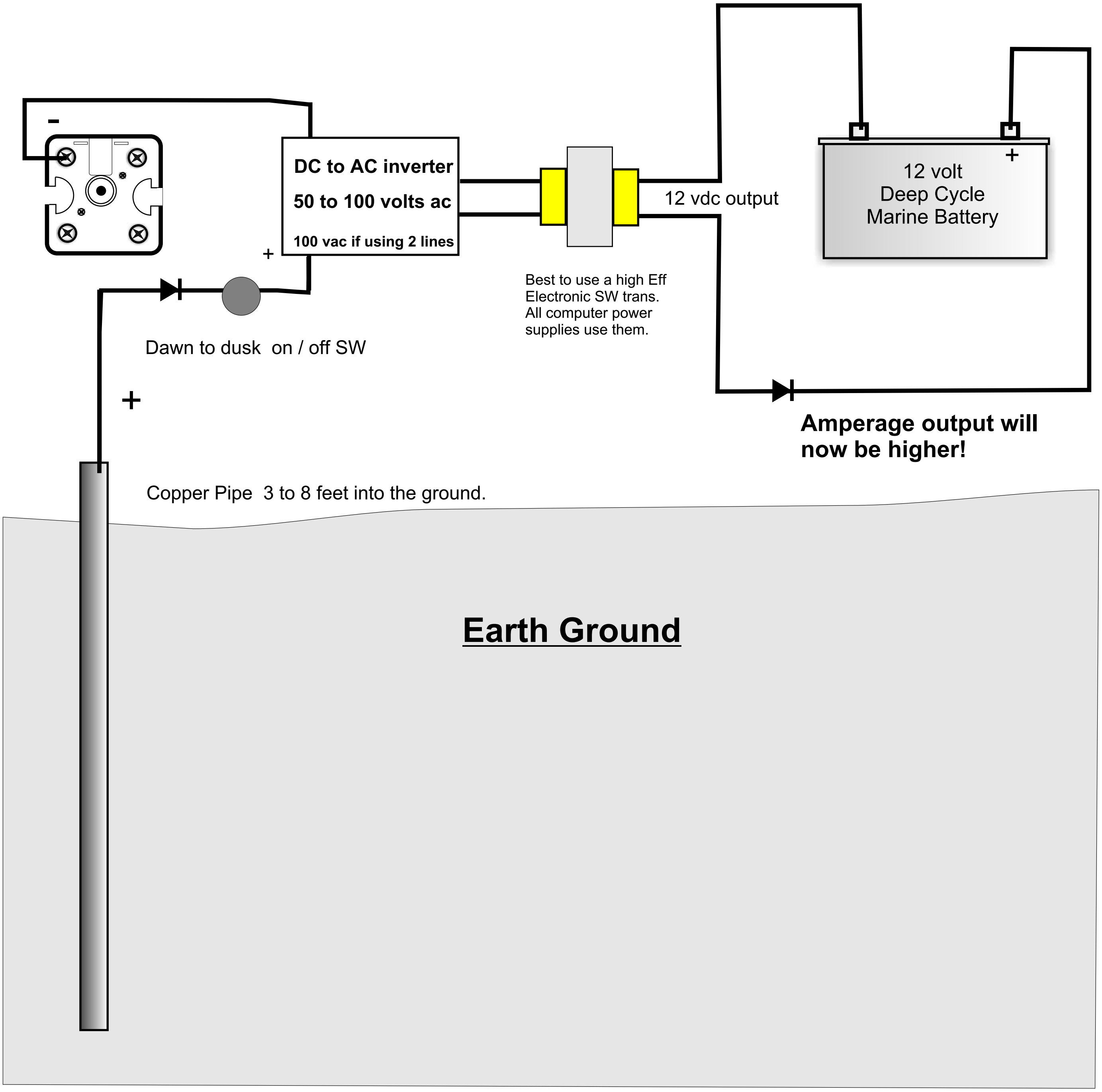

-
DC to AC inverter
50 to 100 volts ac
100 vac if using 2 lines
+
12 vdc output
12 volt
Deep Cycle
Marine Battery
+
Best to use a high Eff
Electronic SW trans.
All computer power
supplies use them.
Dawn to dusk on / off SW
+
Amperage output will
now be higher!
Copper Pipe 3 to 8 feet into the ground.
Earth Ground

If you wish to build your own hand spring timer and relays you can buy material at **www.smallparts.com**

Copyright 2002 - 2004 Creative Science & Research

A Simple but effective Inverter

A simple but very effective inverter that you can build if you know how to handle a soldering gun. Parts can be purchased at any radio shack store or electronic supply house. Search the web. When I say search the web, you would not believe how many people do not know what I am talking about so I will explain it. You can find just about anything you need on the web. If you have AOL.com then that is a search engine as well as Google.com, Yahoo.com etc.. They are all powerful search engines, you simply type in a word in the search the web white address bar, for example: Looking for capacitors, Type in capacitors or electronic parts and supplies. You can also buy beginners books in electronics from any Radio shack or on line store to quickly help you learn how to put this inverter together, some suppliers sell HV inverter kits.

The 50 vdc must be changed to ac voltage current or it will not work (step up voltage) into a step up transformer or diode / capacitor step up multiplier circuit.

High Voltage Power supply / DC to AC Inverter

Rated for 1.5 to + 9vdc if using 50 vdc or 100 find cap and transistors rated for that voltage or more, or you will burnout trans

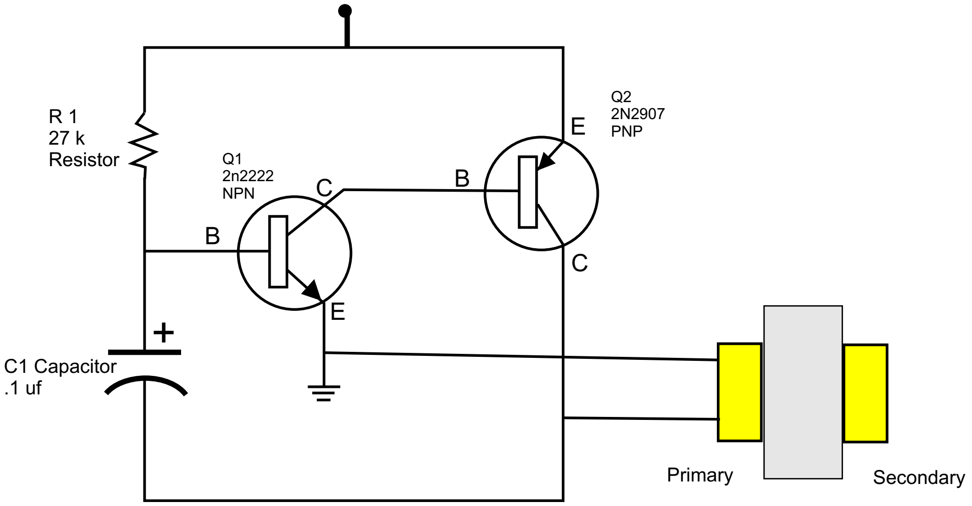

T1 = 6.3 v x 120v transformer This circuit will produce about 220 volts. Use a 1m resisitor for higher volatges.

We found this cicuit in an old electronics book, it is for low voltages but you can replace the parts with higher rated parts so they will work on higher amperages or higher voltages.

Catalog #27

Creative SCIENCE & Research

Copyright 2002 - 2004 Creative Science & Research

PO BOX 557 New Albany, IN. 47151 USA

Information Big Brother & the Oil Companies do not want you to know!

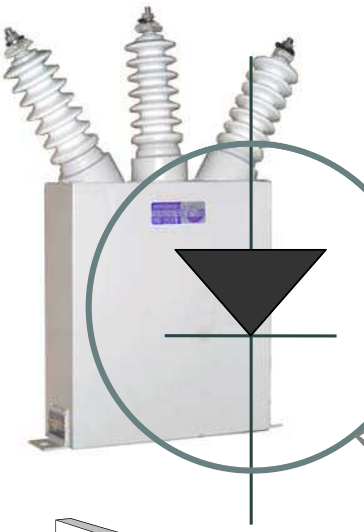

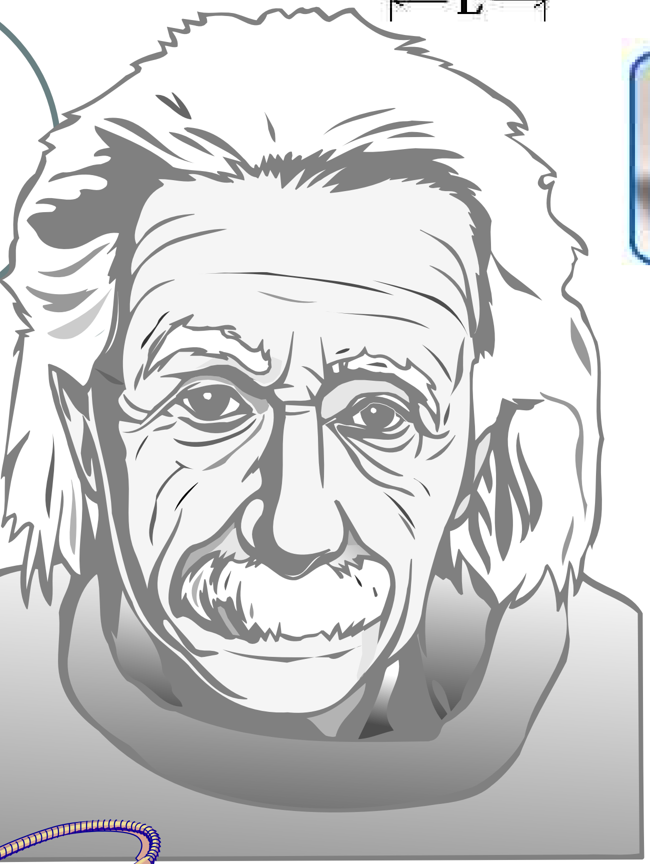

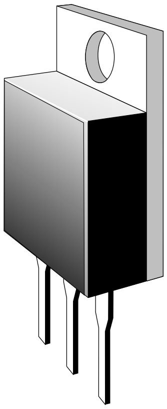

2004

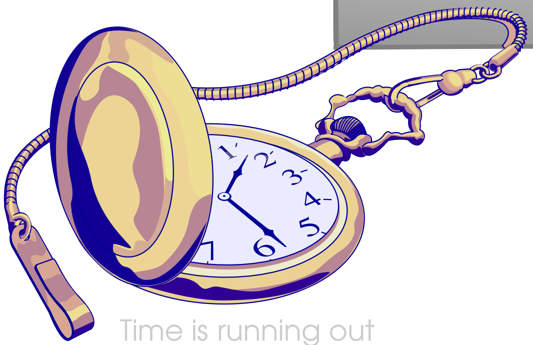

Time is running out

Search for:

FREE ENERGY ▼ Find

www.fuellesspower.com

Copyright 2002 - 2004 Creative Science & Research

INTRODUCTION

CREATIVE SCIENCE AND RESEARCH *"Since 1993"*

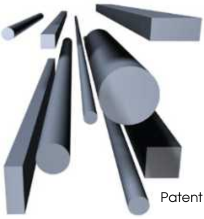

Patent Pending 2003

Free Energy using Carbon rod,
as well as special Molded Soft Iron Ductal Cores

The reason we made this **catalog** was to raise money for research and development and to also share to the World our findings and discoveries on **The Fuelless Engine.**
All of our research and findings on the topic of free energy have guided us to many other ideas and devices along the way which are included in this catalog, but **none can compare** to the **Fuelless Engine!** If the whole world uses this **free energy engine** we will all be able to free ourselves from the Gas, Oil, and Electric companies that now hold us prisoner. By the year 2020 every home should have their own power plant, and we will do everything in our power to make this happen.

This engine will run forever with a once a year maintenance check, and will provide **FREE ENERGY** for the rest of your life.

If you are a skeptic and find Free Energy and Perpetual motion a little hard to believe then consider this, The earth is FREE ENERGY in a perpetual motion state. Hydrogen and oxygen atoms are FREE ENERGY constantly being reused over and over. A coil of wire induced by a DC current and then turned off will produce **Free energy from a collapsing magnetic field**. HV Capacitor's play a very important roll as well..... Solar Cells are free energy devices but they have there limitations, they can only operate when sunlight is present, but our devices are far greater than solar cell technology and can operate 24 hrs a day 7 days a week.

We take what we do very seriously! We would not sell anything that we did not first test ourselves. I personally believe 100% in what we sell. **Our plans are Guaranteed to work or your money back!**

Free Energy is everywhere, don't be left in the dark.

Thank you
David Waggoner
David Waggoner
Owner / Research and Development

Creative Science & Research
PO BOX 557
New Albany, IN. 47151 USA

sales@fuellesspower.com

www.fuellesspower.com

www.fuelless.com

www.fuellessengine.com

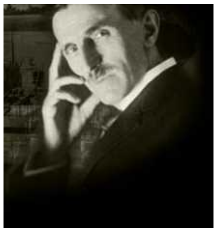

Nikola Tesla

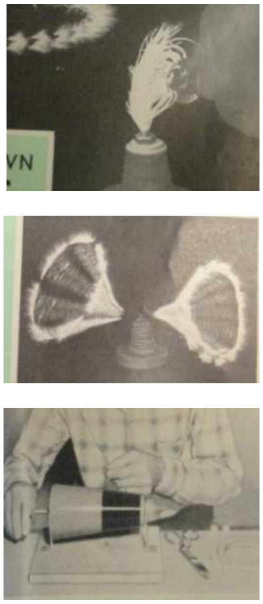

Tesla Coil

The Fuelless Engine

Copyright 2002 - 2004 Creative Science & Research

Free Energy Free Electricity for your home or car!

FREE POWER! For your home, garage or car!

The US Government nor the Oil Companies will not let us or anyone else manufacture these powerful engines right now, (They say it could destroy our economy?)
But if we can slowly introduce these devices to the world I think we can convert every home and automobile to run on free energy by the year 2020, I believe it could be done safely and would not hurt our economy at all. That's where you come in, We need backyard researchers just like yourself, to build our device(s), use them, and keep them confidential! We have made it as easy for you to build using our easy to follow plans.

2′ x 2′ x 1′ **350 Hp**

Many years of research went into designing this engine. There is no other engine like it. You can scale these plans up or down to build any size horsepower that you want..

THIS ENGINE CREATES IT'S OWN POWER, *It does not run on any type of gasoline, oil or any other type of combustible fuel. It has a high free energy output! Guys, if you do not get these plans now you will be kicking yourself in the pants tomorrow! This is the engine that every free energy researcher is talking about!*

EASY TO BUILD STEP BY STEP PLANS / BLUE PRINTS

MANY MAGAZINE ARTICLES HAVE BEEN WRITTEN ON OUR ENGINE! You may have read about us in the 1996 July Issue of Exotic Research magazine. Popular Science has done many articles on free energy in the early 1940's and 50's, I'm sure before BIG BROTHER asked them to stop!

Plans Only $16.95

1 - 50 hp Order # 362

GUARANTEED TO WORK OR YOUR MONEY BACK!

Video Only $29.95

Order # 361

50 to 350 hp plans only $40.00 Order#362rc

Creative Science & Research P.O. BOX 557 New Albany, IN. 47151
Once we teach you the secrete in building this engine You can then design and build this engine to any HP you desire.

Copyright 2002 - 2004 Creative Science & Research

All Our Plans are Now on CD!

FREE ENERGY MOTORS & GENERATORS

*** CONFIDENTIAL INFORMATION!**
For Selected Customers Only!

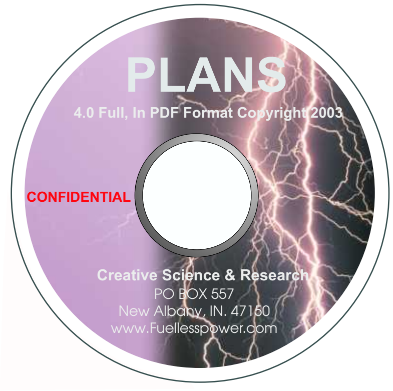
PLANS
4.0 Full, In PDF Format Copyright 2003
CONFIDENTIAL
Creative Science & Research
PO BOX 557
New Albany, IN. 47150
www.Fuellesspower.com

VISA
MasterCard
DISCOVER
AMERICAN EXPRESS
PLANS
PRIORITY MAIL

As requested by many of our customers!

* Information The Oil Companies and Big Brother Do Not Want You To Know!

We are now offering our entire web site and catalog of Free Energy Plans for only **$300**

You save over $200 (a $500.00 value)
Order # CD Package # 1

*** LOADED with information! Packed full of color photo's, More information! etc....**
Plans are sent to you on a CD Computer Disk, which you can then view or print from your computer.

PDF format for Adobe Reader Computer Program. If you do not have Adobe Reader 4.0
It is FREE for downloading at www.adobe.com

* Please add $3.95 for US Priority Shipping which includes online tracking

*** Guaranteed To Work Or Your Money Back!**

Other CD Packages Available: See our web site order form, Bottom page
www.FuellessPower.com

Please Send a Check or Money Order To:

Creative Science & Research
PO BOX 557 New Albany, IN. 47151 USA

www.fuellesspower.com or www.fuelless.com

NOTICE: Most of our plans can are now available by E-mail Download

Radio Receivers, *from crystal set to stereo*

author: Miomir Filipovic

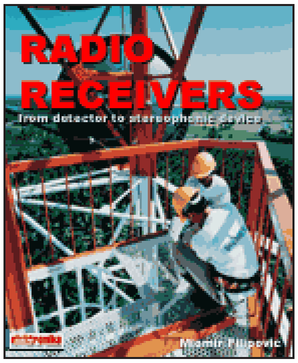

It is hard to imagine what would the modern world look like without the permanent exchange of huge quantity of information. It is being transferred by various means (newspapers, telephone, the Internet etc.), however, the fastest way of doing it, and sometimes the only one, is by radio, where transfer is being done by electromagnetic waves, traveling at the speed of light. This book covers the history and principles of radio transmission and an array of different radio receivers...

Chapter I INTRODUCTION

Chapter II PRINCIPLES OF RADIO TRANSMISSION

2.1. AM Transmitter
2.2. FM Transmitter
2.3. Wavebands

Chapter III DIRECT (TRF) RADIO RECEIVERS

3.1.The Simplest Radio Receiver
3.1.2. The Antenna
3.1.3. The Ground
3.1.4. Other Components
3.2. The Simplest Amplified Radio Receiver
3.3. Simple Radio Receiver with TDA7050 IC
3.4. Simple Radio Receiver with LM386 IC
3.5. Radio Receiver with Increased Sensitivity Audio Amplifier
3.6. Universal Audio Amplifier
3.7.Receiver with HF Amplifier
3.8. The Audion - Direct Receiver with Drain Detector
3.9.1. Reaction - Type Receiver
3.9.2. Direct SW Receiver for AM, AM-SSB & CW Signals
3.10. Miniature Receiver with ZN414 (ZN414Z) IC
3.11. Pocket Receiver with ZN414 & LM386 IC's
3.12. Miniature Receiver with ZN415E IC
3.13. Receiver with ZN415 & LM386 IC's
3.14. Mini Receiver with ZN415 & TDA7052 IC's
3.15. Direct (TRF) FM Receivers
3.15.1. The Simplest FM Receiver
3.15.2. The Simplest FM Receiver with Audio Amplifier
3.15.3. FM Receiver with One Transistor and Audio Amplifier
3.15.4. FM Receiver with (just) One Transistor
S

Chapter IV SUPERHETERODYNE RADIO RECEIVERS

4.1.Superheterodyne AM Receivers
4.1.1. The Simplest AM Superheterodyne Receiver (worldwide)
4.1.2. The Fully (not exactly 100%) superheterodyne AM Receiver No.1
4.1.3. The Fully (not exactly 100%) superheterodyne AM Receiver No.2
4.2. Superheterodyne FM Receivers
4.2.1. FM Receiver with TDA7000
4.2.2. FM Receivers with TDA7088T
4.2.2.1. Miniature FM Receiver
4.2.2.2. Stereophonic Receiver

Chapter V APPENDIX

5.1. Making PCB's
5.2. Computer - Aided Radio Receiver Control
5.3. Receivers with NE612 IC
5.3.1. Synchrodyne AM Receiver
5.3.2. AM Receiver with Synchro - Detector
5.3.3. Input Circuits for the Receivers with NE612
5.4. Universal Audio Receiver
5.5. Additional Circuits
5.5.1. Fine Tuning
5.5.2. Electronic Tuning
5.5.3. Signal Suppressing of Local Radio Transmitter(s)
5.5.4. Dual Tuning
5.5.5. Separation of Stages - Preventing the Oscillation
5.6. The Boxes
5.7. Bimboard, Protoboard
5.8. Universal PCB Plates
5.9. A Modern Oldtimer

CHAPTER 1 Introduction

It is hard to imagine what the modern world would look like without the constant exchange of a huge quantity of information. It is currently disseminated by various means such as newspapers, telephone and the Internet. However the fastest way, and sometimes the only way, is by radio. This is where the transfer is by electromagnetic waves, traveling at the speed of light. In radio communication, a radio transmitter comprises one side of the link and a radio receiver on the other. No conductor of any kind is needed between them, and that's how the expression **Wireless Link** came into being.
In the early days of radio engineering the terms **Wireless Telegraph** and **Wireless Telephone** were also used, but were quickly replaced with **Radio Communication**, or just **Radio**.

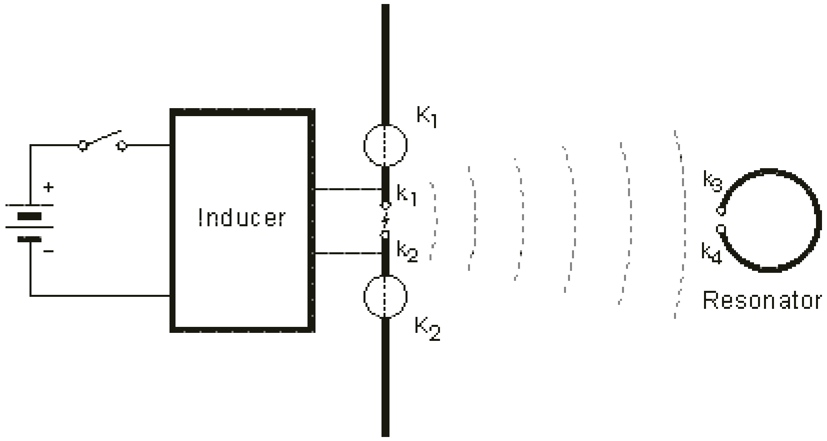

Pic.1.1. Electrical diagram of Hertz's transmitter / receiver

Radio communication is created by means of electromagnetic waves, of which the existence and features were theoretically described and predicted by James Maxwell, in 1864.
First experimental proof of this theory was given by Heinrich Hertz in 1888, ten years after Maxwell's death.
It was already known at that time that electric current exists in oscillatory circuits made of a capacitor of capacity **C** and coil of inductance **L**. It was Thomson, back in 1853 that determined the frequency of this arrangement to be:

$$f_r = \frac{1}{2\pi\sqrt{LC}}$$

Hertz used an oscillatory circuit with a capacitor made of two bowls, K1 and K2 (Pic. 1.1), and the "coil" was made of two straight conductors. The bowls could be moved along the conductors. In this way the capacitance of the circuit could be altered, and also its resonance frequency. With every interruption from the battery, a high voltage was produced at the output of the inductor, creating a spark between the narrow placed balls k1 and k2. According to Maxwell's theory, as long as there was a spark, i.e. alternating current in the circuitry, there was an electromagnetic field surrounding the conductors, spreading itself through the surrounding space. A few metres away from this device Hertz placed a bent conductor with metal balls k3, k4 placed on the ends, positioned very close to each other.
This also was an oscillatory circuit, called the resonator.

According to Maxwell's theory, voltage induced by the electromagnetic waves should be created in the resonator. Voltage existence would be shown by a spark between the balls k3 and k4.

And that's the way it was: Whenever there was a spark in the oscillator between the balls k1 and k2, a spark would also be produced by the resonator, between balls k3 and k4.

With various forms of the arrangement in Pic. 1.1, Hertz proved that electromagnetic waves behave as light since they could also be reflected and refracted.

It was also shown that light is of electromagnetic nature, as stated by Maxwell.

Hertz, however, did not believe in the practical value of his electromagnetic waves experiments. The range of the link was no further than a few meters. The transmitted signal was very weak, therefore the signal in the receiver had a very small amplitude and it wasn't possible to detect it at a greater distance. The possibility of amplifying the signal in the receiver did not exist at the time.

Besides the short range, another shortcoming of the link was noted: If another similar transmitter was working nearby, a receiver detected all the signals at the same time. It did not have the ability of isolation.

However crude and simple these experiments were at the time, they represented the birth of a new scientific branch - **Radio Engineering**.

The pioneers of radio were Popov and Marconi, but the place of honor belongs to Nikola Tesla, who demonstrated wireless broadcasting in 1893, at the Franklin Institute.

Pic.1.2 shows the arrangement of this broadcast system.

Tesla's idea was to produce electromagnetic waves by means of oscillatory circuits and transmit them over an antenna. A receiver would then receive the waves with another antenna and oscillatory circuit being in resonance with the oscillatory circuit of the transmitter. This represented the groundwork of today's radio communications.

In 1904 John Flemming created the diode, and in 1907 Lee De Forest invented the triode. That year can be considered the birth of electronics, with the triode being the first electronic component used in a circuit for signal amplification.

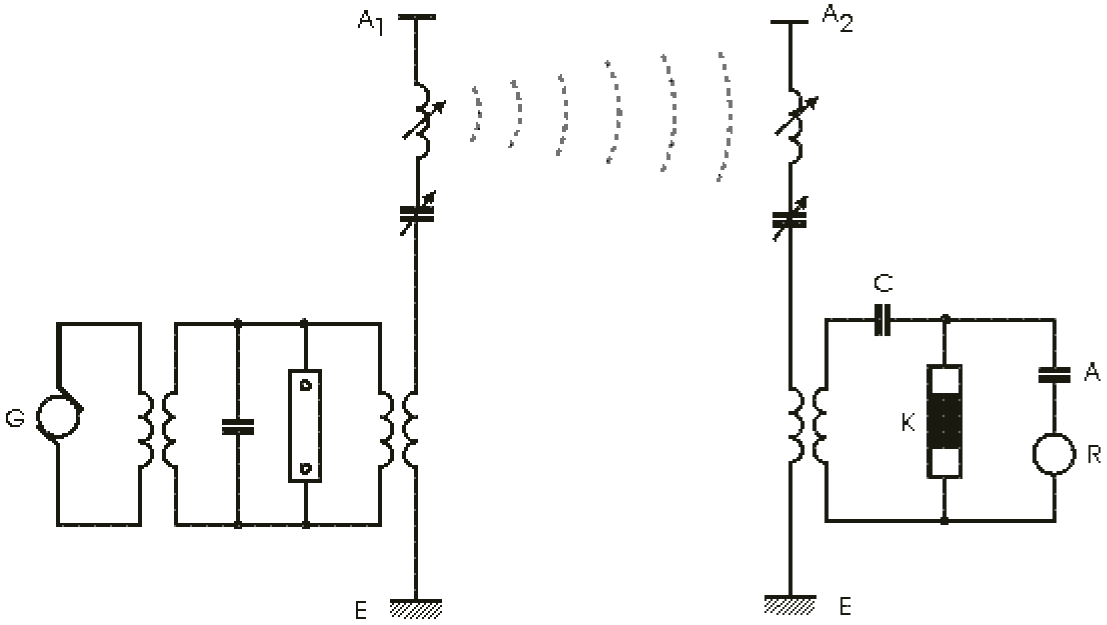

Pic. 1.2. Electrical diagram of Tesla's radio transmitter & receiver from 1893.

Rapid development of radio engineering over the ensuing years produced many innovations and after the First

World War a huge number of radio stations emerged.
At that time TRF (Tuned Radio Frequency) receivers were used. Compared to modern receivers they had both poor selectivity and sensitivity, but back then they fulfilled the demands. The number of radio stations was much less than today and their transmitting power was much smaller. The majority of listeners were satisfied with the reception of only local stations. However as the number of stations increased, as well as their transmitting power, the problem of selecting one station out of the jumble of stations, was becoming increasingly more difficult.

It was partially solved with an increase in the number of oscillatory circuits in the receiver and the introduction of positive feedback, but the true solution was the invention of the **superheterodyne receiver**. This was accomplished by Lewy (1917), and improved by E.H. Armstrong (1918).
An enormous impact on the world of radio was the invention of the transistor by Bardeen, Bretten & Schockley, in 1948. This reduced the size of the radio receiver and made truly portable sets a reality.
This was followed by the introduction of the integrated circuit, enabling the construction of devices that not only proved better in every way than those using values, but also new designs.
Radio amateurs' contribution to radio engineering should also be emphasized.
In the beginning, radio communication was being conducted in the LW and MW bands. But achieving long-distance reception required very powerful transmitters. The SW band was considered to be useless for radio broadcast on long distances and was given to radio amateurs.
The were banned from using LW and MW bands by commercial radio stations.
However, something unexpected happened: Amateurs were able to accomplish extremely long distance transmissions (thousands of kilometres), by using very low-power transmitters. This was later explained by the influence of the ionosphere layer, the existence of which was also predicted by Tesla.
Modern radio receivers differ greatly from the "classical" types, however the working principles are the same.

The only significant difference is in the way the receiver is tuned to a station. Classical devices used a variable capacitor, coil or varicap diode, with the frequency read from a scale with movable pointer. In modern devices, the adjustment is done with a frequency synthesizer controlled by a microprocessor and the reading is displayed on an optical readout.

The inclusion of a microprocessor enables any one of a large number of pre-tuned stations to be selected and displayed and the use of a remote control makes the receiver even more user friendly.

Chapter 2 Principles of radio transmission

2.1. AM Transmitter
2.2. FM Transmitter
2.3. Wavebands

Transfer of information (speech, music, image, computer data etc.) by radio can be presented in its simplest form with block - diagram as on Pic.2.1. That is a transmission realized by amplitude - modulated signal. Since, in our example, the information being transferred is the sound, the first step of such transmission is converting the sound into electrical signal, this being accomplished by a microphone. The low - frequency (LF) voltage at microphone output (Pic.2.1-a), that represents the electrical "image" of the sound being transferred, is being taken into the transmitter. There, under the effect of LF signal, the procedure called amplitude modulation is being carried out, and on its output high - frequency (HF) voltage is generated, its amplitude changing according to the current LF signal value. HF voltage creates HF current in the antenna, thus generating electromagnetic field around it. This field spreads through the ambient space, being symbolically shown on Pic.2.1 with dashed circles. Traveling at the speed of light (c=300 000 km/s), the electromagnetic field gets to the reception place, inducing the voltage in the reception antenna, as shown on Pic.2.1-c. This voltage has the same profile as the one on Pic.2.1-b, except it has much smaller amplitude. In the receiver, the amplification and detection are carried out first, resulting with the LF voltage on its output, that has the same profile as the one on Pic.2.1-a. This voltage is then transformed into sound by loudspeaker, that sound being exactly the same as the sound that acted upon the microphone. This, naturally, is the way it would be in ideal case. Back to reality, due to device imperfection as well as the influence of various disturbances, the sound being generated by the loudspeaker differs from the one that acts upon the microphone membrane. The block - diagram on Pic.2.1 (excluding the HF signal shape) is also applicable in case of radio transmission being carried out by frequency modulation. In that case frequency modulation is being carried out in the transmitter, under the effect of LF signal coming from the microphone, therefore HF signals on Pics.2.1-b and 2.1-c having constant amplitude, and their frequency being changed in accordance with the actual value of LF signal from the microphone. In fact, all types of radio transmission can be presented with Pic.2.1. First, the information being sent is always transformed into electrical signal through the appropriate converter. In telegraphy this converter is the pushbutton, in radiophony it's a microphone, in television engineering an image analysis cathode ray tube (CRT) etc. Then, with this "electrical image" of information, the modulation is being done. The modulated HF signal is being transferred into antenna and transmitted. On the reception place, the modulated signal from the reception antenna is being amplified and detected and then, again with the appropriate converter (pen recorder, loudspeaker, TV CRT etc.), the information is transformed back into its original form.

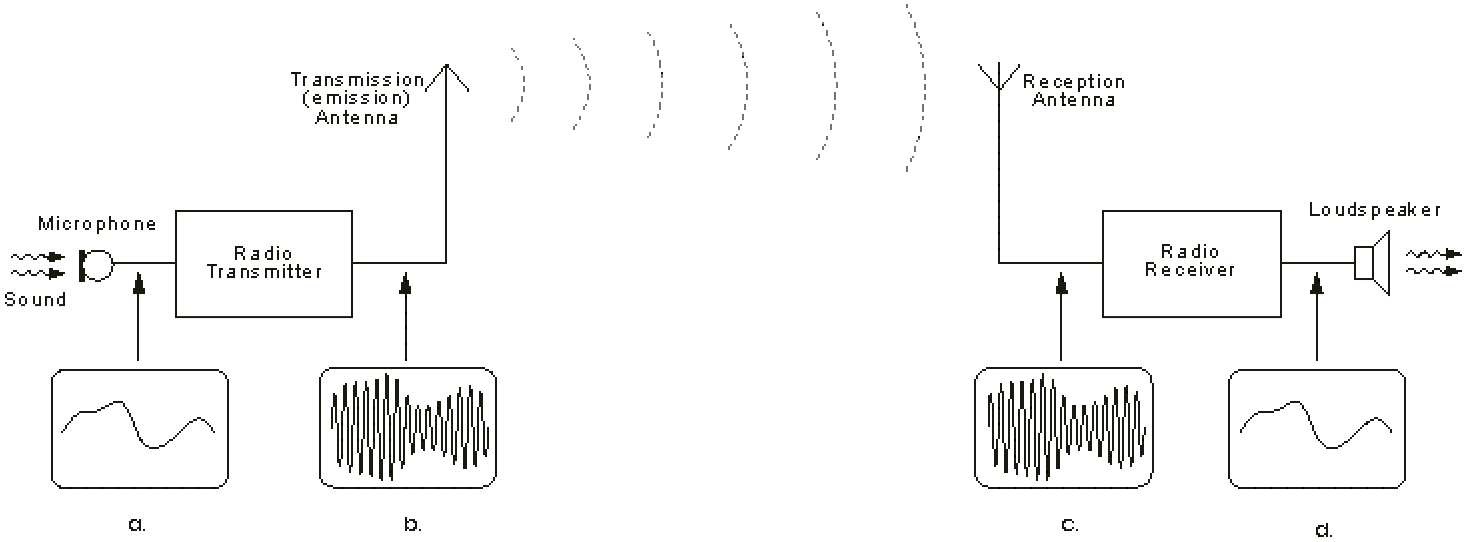

Pic.2.1. Radio Transmission Block-Diagram

2.1 AM Transmitter

In order to better understand the way the radio transmitter works, block - diagram of a simple AM (amplitude modulated) signal transmitter is shown on Pic.2.2. The amplitude modulation is being performed in a stage called the modulator. Two signals are entering it: high frequency signal called the carrier (or the signal carrier), being created into the HF oscillator and amplified in the HF amplifier to the required signal level, and the low frequency (modulating) signal coming from the microphone or some other LF signal source (cassette player, record player, CD player etc.), being amplified in the LF amplifier. On modulator's output the amplitude modulated signal UAM is acquired. This signal is then amplified in the power amplifier, and then led to the emission antenna.

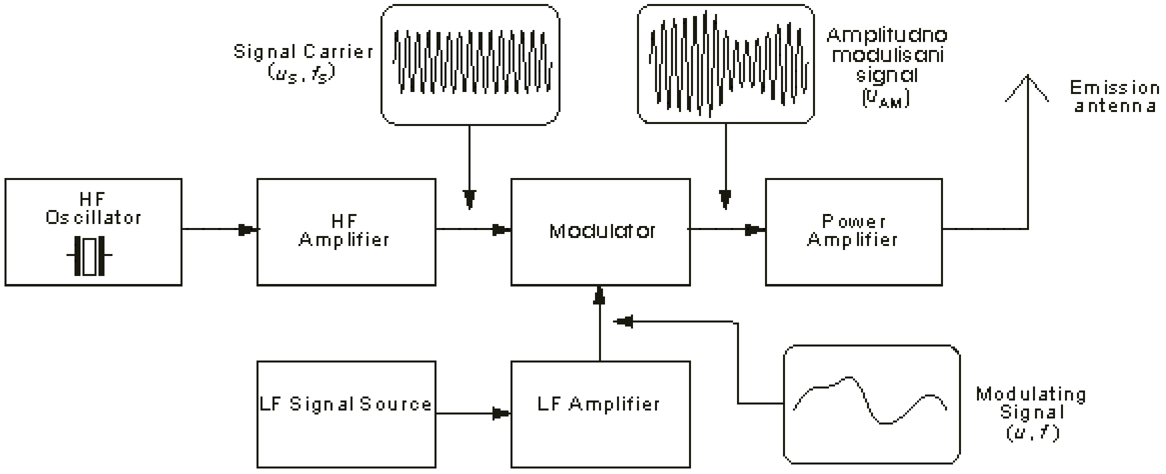

Pic. 2. 2. AM Transmitter Block Diagram

The shape and characteristics of the AM carrier, being taken from the HF amplifier into the modulator, are shown on Pic.2.3-a. As you can see, it is a HF voltage of constant amplitude US and frequency fS. On Pic.2.3-b the LF signal that appears at the input of the modulator at the moment t0 is shown. With this signal the modulation of the carrier's amplitude is being performed, therefore it is being called the modulating signal. The shape of the AM signal exiting the modulator is shown on Pic.2.3-c. From the point t0 this voltage has the same shape as that on Pic.2.3-a. From the moment t0 the amplitude of AM signal is being changed in accordance with the current value of the modulating signal, in such a way that the signal envelope (fictive line connecting the voltage peaks) has the same shape as the modulating signal.

Let's take a look at a practical example. Let the LF signal on Pic.2.3-b be, say, an electrical image of the tone being created by some musical instrument, and that the time gap between the points t0 and t2 is 1 ms. Suppose that carrier frequency is fS=1 MHz (approximately the frequency of radio Kladovo, exact value is 999 kHz). In that case, in period from t0 till t2 signals us on Pic.2.3-1 and uAM on 2.3-c should make a thousand oscillations and not just eighteen, as shown in the picture. Then It is clear that it isn't possible to draw a realistic picture, since all the lines would connect into a dark spot. The true picture of AM signal from this example is given on Pic.2.3-d. That is the picture that appears on screen of the oscilloscope, connected on the output of the modulator: light coloured lines representing the AM signal have interconnected, since they are thicker than the gap between them.

Block - diagram on Pic 2.2 is a simplified schematic of an AM transmitter. In reality there are some additional stages in professional transmitters that provide the necessary work stability, transmitter power supply, cooling for certain stages etc. For simple use, however, even simpler block diagrams exist, making the completion of an ordinary AM transmitter possible with just a few electronic components.

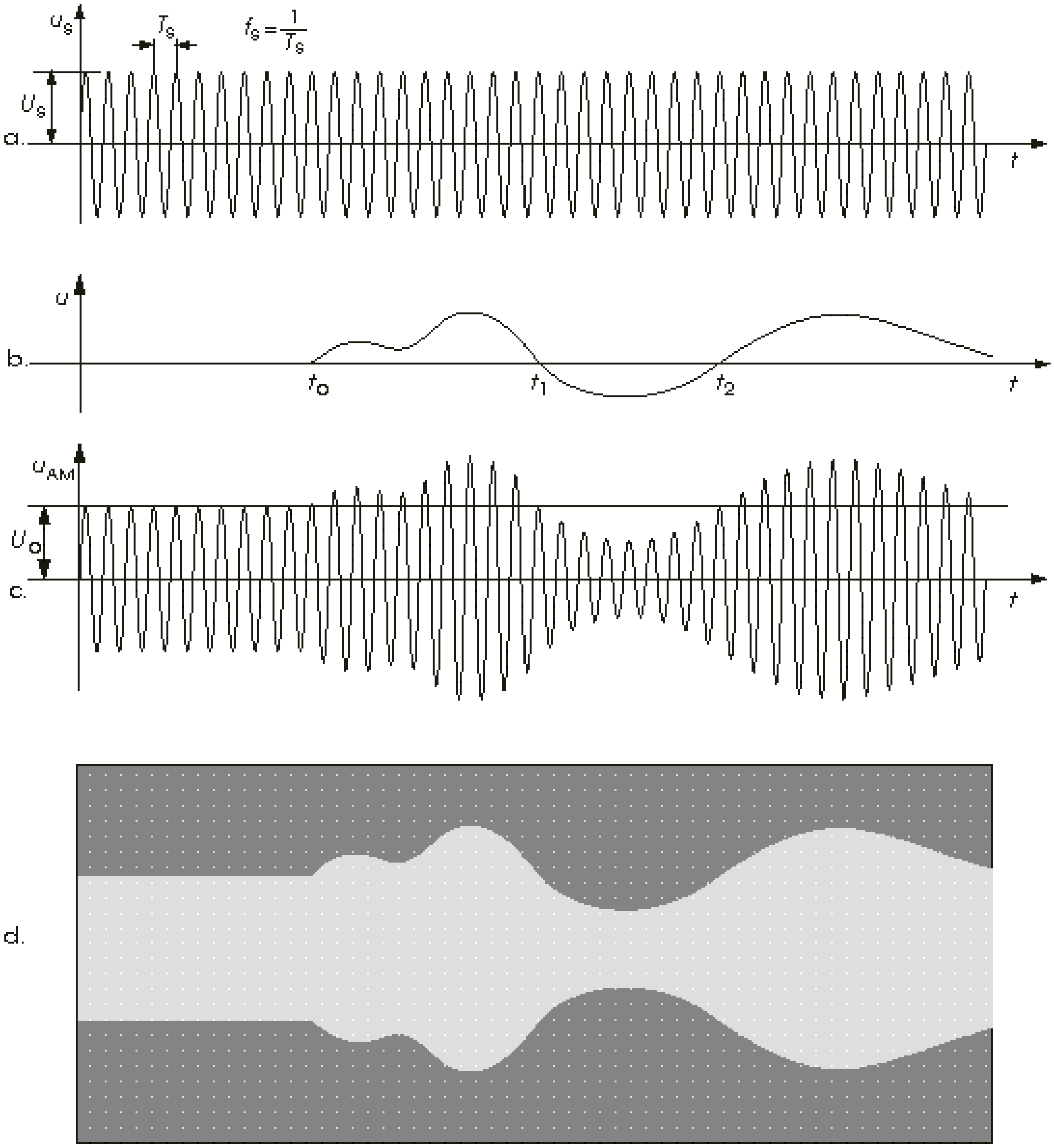

Pic.2.3. Voltage wave envelopes of an AM signal: a- The (signal) Carrier, b- The Modulating Signal (LF signal being transferred), c- AM (amplitude - modulated) signal, d- true look of the AM signal.

2.2. FM Transmitter

Block diagram of an FM (frequency modulated) transmitter is given on Pic.2.4. Information being transferred, i.e. the modulating signal, is a signal from some LF source. it is being amplified in LF amplifier and then led into the HF oscillator, where the carrier signal is being created. The carrier is a HF voltage of constant amplitude, whose frequency is, in the absence of modulating signal, equal to the transmitter's carrier frequency fS. In the oscillatory circuit of the HF oscillator a varicap (capacitive) diode is located. It is a diode whose capacitance depends upon the voltage between its ends, so when being exposed to LF voltage, its capacitance is changing in accordance with this voltage. Due to that frequency of the oscillator is also changing, i.e. the frequency modulation is being obtained. The FM signal from the HF oscillator is being proceeded to the power amplifier that provides the necessary output power of the transmission signal.

Voltage shapes in FM transmitter are given on Pic.2.5. Pic.2.5-a shows the LF modulating signal. The frequency modulation begins at moment t0 and the transmission frequency begins to change, as shown on Pic.2.5-b: Whilst current value of the LF signal is raising so is the trasmitter frequency, and when it is falling the frequency is also falling. As seen on Pic.2.5-c, the information (LF signal) is being implied in frequency change of the carrier.

The carrier frequencies of the radio difusion FM transmitters (that emmit the program for "broad audience") are placed in the waveband from 88 MHz til 108 MHz, the maximum frequency shift of the transmitter (during the modulation) being ±75 kHz. Because of that the FM signal should be drawn much "thicker", but it would result in a black-square-shaped picture.

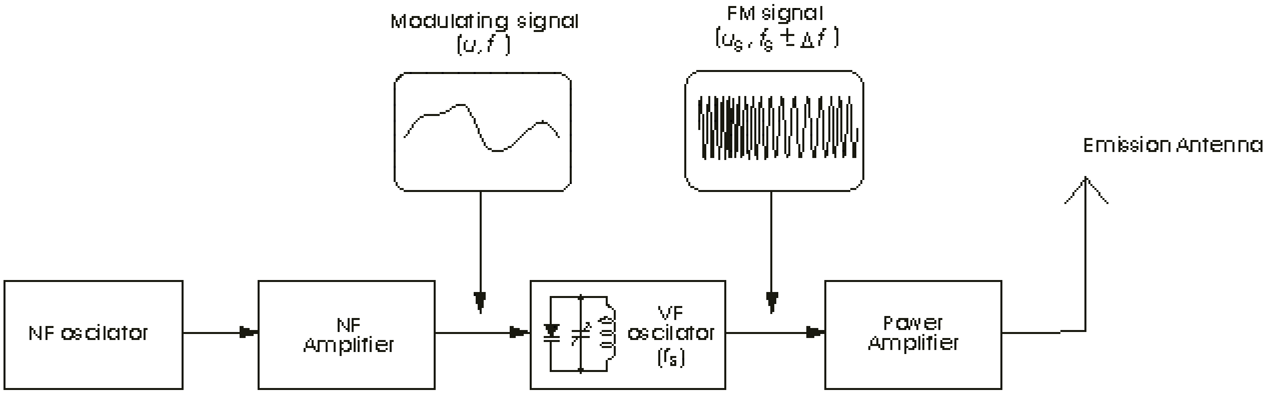
Modulating signal
(u, f)
FM signal
(u_s, f_s ± Δf)
Emission Antenna
NF oscilator
NF Amplifier
VF oscilator (f_s)
Power Amplifier

Picture 2.4. FM transmmiter Block diagram

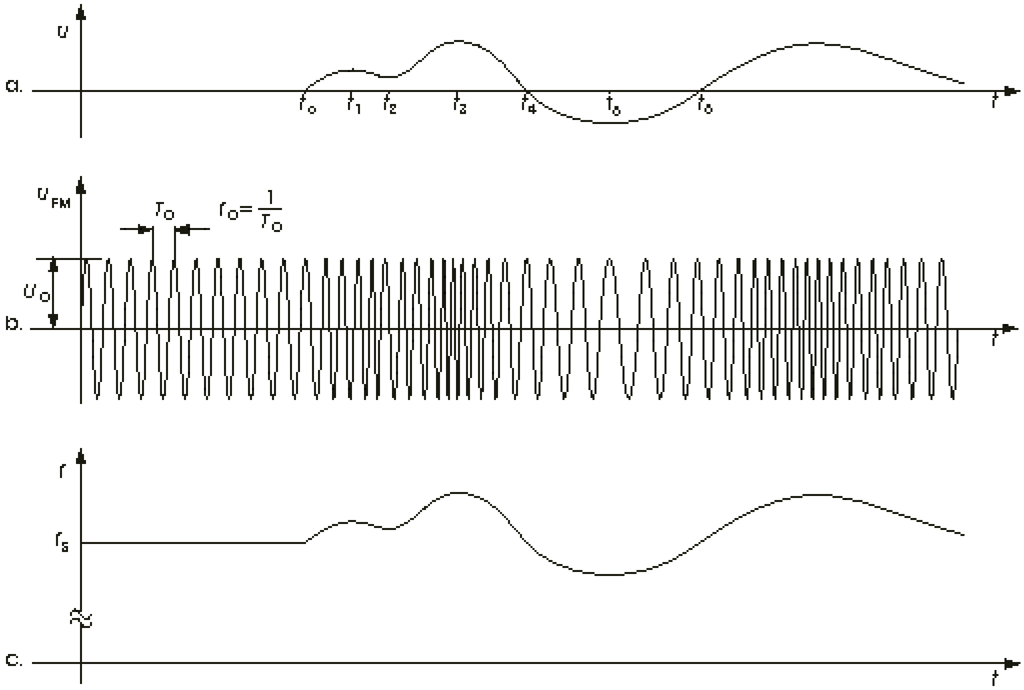
u
a.
t0
t1
t2
t3
t4
t5
t6
t
u_FM
T0
f0 = 1/T0
U0
b.
t
f
f_s
c.
t

2.3. Wavebands

While considering problems related to the realization of the long - distance radio links, significant differences between the electromagnetic waves of various frequencies must be kept in mind. For example, low frequency waves (below 500 kHz) can bend themselves following Earth's curvature, while the HF waves are moving in streamlines, just as light. Some waves can be reverberated from the ionosphere, others are passing through it etc. According to characteristics of their outspread, radio waves can be classified into several groups or ranges: long, mid, short and ultra-short. Limits between the wavebands are not precise, with the raise of their frequency the waves are gradually losing some features while gaining some others. This division is shown in Table 1.

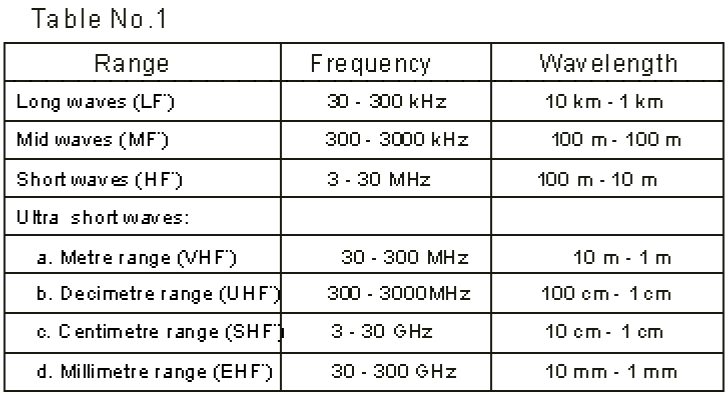

Table No.1

Range	Frequency	Wavelength
Long waves (LF*)	30 - 300 kHz	10 km - 1 km
Mid waves (MF*)	300 - 3000 kHz	100 m - 100 m
Short waves (HF*)	3 - 30 MHz	100 m - 10 m
Ultra short waves:		
a. Metre range (VHF*)	30 - 300 MHz	10 m - 1 m
b. Decimetre range (UHF*)	300 - 3000MHz	100 cm - 1 cm
c. Centimetre range (SHF*)	3 - 30 GHz	10 cm - 1 cm
d. Millimetre range (EHF*)	30 - 300 GHz	10 mm - 1 mm

* LF low frequencies, MF mid frequencies, HF high frequencies, VHF very high frequencies, UHF ultra high frequencies, SHF super high frequencies, EHF extra high frequencies. Waves with wavelength smaller than 30 cm are also called the microwaves.

In the third table column the wavelengths are given. Wavelength (ë) is distance that the wave passes moving at the speed of light (c=3*103 m/s), during the period that is equal to its oscillating period (T): ë=c*T. Having in mind that the wave frequency is f=1/T, one can easily get to the well known expression that gives the relation between the wavelength and the frequency:

$$\lambda = \frac{c}{f}$$

Using this formula one can calculate the wavelength knowing the frequency and vice versa. For example, wavelength of an FM transmitter emitting at f=100 MHz frequency is L=3*108/100*106=3 m. Similar to that, wavelength of Radio Belgrade 1 is L=439 m, which makes its frequency equal to f=3*108/439=684 kHz. Radio diffusion is being performed in certain parts of the wavebands given in Table 1, their boundary frequencies are (rounded values):

LW (long waves) 150 kHz (2km) 300 kHz (1 km)
MW (mid waves) 500 kHz (600 m) 1500 kHz (200 m)
SW (short waves) 6 MHz (50 m) 20 MHz (15 m)
FM (ultra short waves) 88 MHz (3.4 m) 108 MHz (2.78 m)

In LW, MW and SW the amplitude modulation is used, while in FM range it is the frequency modulation.

Here are the frequencies (in kHz) of some radio transmitters from the MW range, that can serve for tuning of the radio receivers being described in this issue: Timisoara 630, Belgrade1 684, Bucharest 855 .

3.1. The Simplest Radio Receiver

Each radio receiver must have a reception antenna. It is an electrical conductor, where voltages of various frequencies and amplitudes are being induced, under the influence of electromagnetic fields from various radio transmitters. Besides these voltages, those induced by EM fields that are created by various disturbance sources (such as electrical motors, various household appliances spark-plugs of an automobile and all other devices where electrical current is being switched on/off during work) are also present in the antenna, as well as those from fields originating from outer space or the Earth's atmosphere. Basic roles that a radio receiver has are:

a. To separate the signal (voltage) of the radio station that it is tuned at from the multitude of other voltages, whilst suppressing (weakening) all other signals as much as possible,
b. amplifies the extrapolated signal and take out information from it and
c. reproduces that information, i.e. restores it into its' original shape.

Even the simplest radio, the one we are discussing in this chapter, must be able to accomplish all these tasks. The electronic diagram of one such device is given on Pic.3.1. It is the famous (years ago) Detector Radio Receiver or shortly, Detector. The signal selection (separation) and voltage amplification are performed in the oscillatory circuit that is made of the capacitor C and coil L, the separation of information (speech or music) from the AM station signal in the detector that comprises the diode D, capacitor C2 and resistance of the headphones, and information restoring in the very headphones.

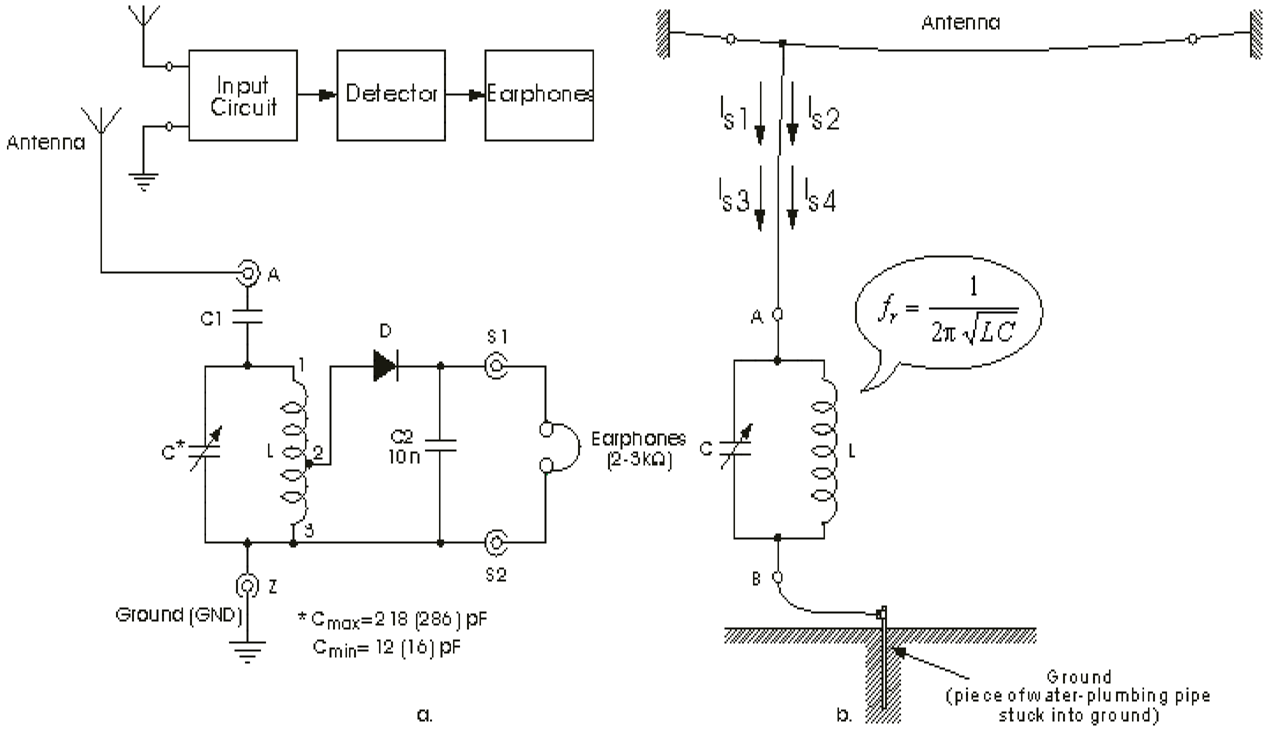

Pic.3.1- a-Block diagram and electronic diagram of a detector radio receiver, b-operating principle of input circuit

Main advantages of this device lie in its extreme simplicity and the fact that it requires no additional energy sources for its' operation. All the energy required it draws from the antenna, which therefore has to be at least a dosen metres long for proper operation. It is also useful to have a good ground. One can do without it but the reception with it is truly better, especially considering the distant and small-power transmitters.

3.1.1. Input Circuit

The capacitor that takes the signal from the antenna (so-called coupling capacitor) C1, variable capacitor C and coil L form the input circuit of the radio receiver. Its main role is to separate the signal of station the receiver is tuned at from multitude of voltages (having various frequencies and amplitudes) existing in the antenna, amplify that signal and turns it over to the detector.
In order to better understand the requests that are to be fulfilled during the practical realization of input circuit, it is necessary to know basic characteristics of circuit made of capacitor C and coil L. It is called 'The oscillatory circuit' and is shown on pic.3.2-a. The amount of its impedance (resistance to AC current) between points A and B, which is marked with , depends on the frequency, as it is shown on the diagram on pic.3.2-b. The most important characteristic of this circuit is its resonance frequency, being given by the Thomson's formula:

$$f_r = \frac{1}{2\pi\sqrt{LC}}$$

As one may notice, the resonance frequency depends on the capacitance of the capacitor C and inductivity of coil L, and changes if one of them change. In our receiver, a variable capacitor is used, that can change its capacitance from Cmax to Cmin, therefore changing the resonance frequency in boundaries from

$$f_d = \frac{1}{2\pi\sqrt{LC_{\max}}} \quad \text{to} \quad f_g = \frac{1}{2\pi\sqrt{LC_{\min}}}$$

The area between fd (lower boundary frequency) and fg (upper boundary frequency) is the reception area of the input circuit, as shown on pic.3.2-b. On this picture, carrier frequencies of four radio transmitters are being marked with fs1, fs2, fs3 and fs4. The resonance frequency of the oscillatory circuit is set (by means of C) to be equal to the carrier frequency of the second station: fs2. In that case, the impedance ZAB - frequency dependance is shown in continuous line. As one can see, the impedance ZAB for all received signals whose carriers have frequencies less than fs1 and greater than fs3 is less than 20 kOhms, while for the station that is tuned it is equal to 200 kOhms. Let us now imagine that the parallel oscillatory circuit is connected with the antenna and ground, as shown on pic.3.1-b. Imagine, also, that there are (only) four voltages in the antenna, that have the same amplitude and are created by four radio

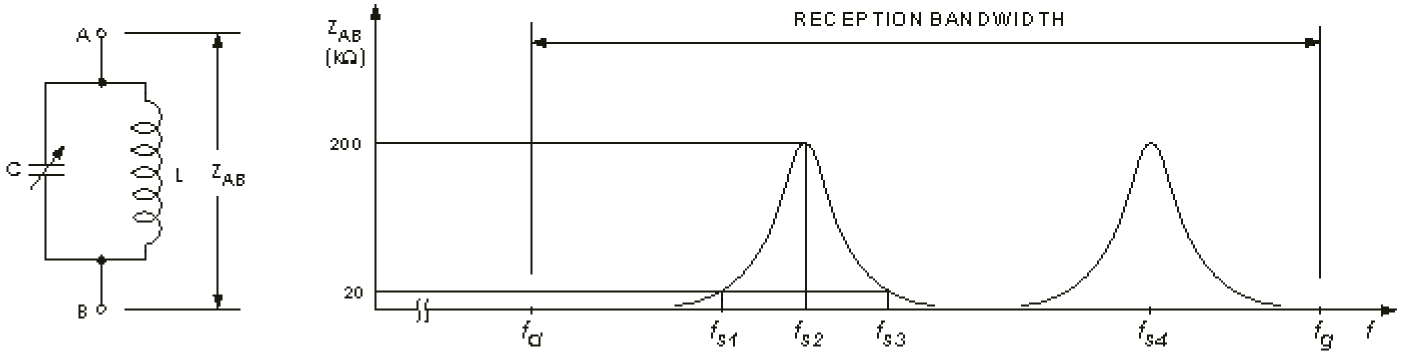

a. b.

Pic. 3.2 a-parallel oscillatory circuit, b-ipedance - frequency dependence of the parallel oscillatory circuit

transmitters, having carrier frequencies of fs1, fs2, fs2 and fs4. Since these voltages spread between the antenna and the ground, four currents will flow through the oscillatory circuit: Is1, Is2, Is3 and Is4. The voltages that are created by them in the oscillatory circuit, between points A and B, are equal, acc. to Ohm's Law, to the product of current and impedance: UAB=I*ZAB. Acc. to pic.3.2-b, for Is2, impedance of the circuit is ZAB=200 kOhms, and for currents Is1 and Is3 it is 10x smaller. That means that the voltage that is being created in the oscillatory circuit by the station that transmits on frequency fs2 will be ten times greater than the voltages being created by stations transmitting on frequencies fs1 and fs3. This is how selection of one station is performed, by means of the oscillatory circuit. Transition to some other station is performed by changing the capacitance of capacitor C, as long as the resonance frequency of the oscillatory circuit does not become equal to the carrier frequency of that station. If its frequency happens to be fs4 (acc. to pic.3.2-b), the impedance of the oscillatory circuit for that case is shown in dashed line, which causes that on the circuit output voltage of the station that transmits on frequency fs4 is acquired, while other stations' signals are suppressed.

At first glance, everything is just the way it should be: Parallel oscillatory circuit extrapolates one and suppresses all other stations. Unfortunately, the reality isn't so simple. First of all, radio transmitters operate with various output (emission) powers and on various geographic distances from the receiver, therefore making the voltages that their signals create in the reception antenna very different in amplitude. It is clear that stronger signals will "cover" the weak ones, thus disabling their reception. E.g. if radio transmitter that emits on the frequency fs1 is geographically much closer to our radio receiver that the transmitter operating on fs2, the voltage the former creates in the reception antenna can be even 200 times greater than the one created by the latter. The oscillatory circuit will do its job as previously described, but on its ends the voltage of the first transmitter will still be greater (20x) than that of the transmitter the receiver is tuned at, and normal reception won't be possible. There are also other problems whose solving will not be discussed herein, and readers that are interested in those can read a book "Radio Receivers", written by Momir Filipovic, issued by the National Textbook Publishing Company from Belgrade, Yugoslavia.

To conclude this chapter, we may say that the simplest radio receiver can cover only signals of the local and powerful radio transmitters.

3.1.2. The Antenna

Extremely important factor for good work of simple radio receivers is the outside antennaTHAT HAS to be long enough, and in which voltages induced by the radio transmitters will be high enough. At first sight, one can think of using instead some modest antenna made of a piece of wire, compensating that with supplying the receiver with amplifier strong enough to give the end result as if much better antenna have been used. That, of course, is not the case, since every amplifier creates noise that makes the reception worse, if not impossible. This fact is the cause for a radio-amateur saying that "Antenna is the best HF amplifier." The external (outside) antenna is being made of copper wire, thick enough to resist strong wind conditions. In the sense of mechanical strength, the best thing is to use the litz wire (cable), i.e. a cable made of huge number of thin threaded copper wires. There is no need to remove the wire isolation if it exists since it doesn't represent an obstacle for the electromagnetic waves. The length of antenna is being determined in accordance with the "TLTB" law (The Longer, The Better). The antenna that we have been using for testing the receivers described herein was 6 metres long (the length of the Radio Receivers Lab at "Tesla" highschool, where it was spread), but if you are in position, you should make it even longer (the author has a friend whose antenna is about 30 metres long). It should be moved away from the sources of electrical disturbances (public electricity cables, various household electrical devices, cars, electric motors etc.). Considering this, the best place for your antenna should be the building roof. The wire can be crossed between two chimneys (Pic.3.3), between a chimney and some pillar, between two purposefully built pillars, between two buildings, a building and a pole in the yard etc. You should, however, always keep in mind that the wire, however strong it may be, can snap during some big storm and, in case that happens, under NO CIRCUMSTANCES it should fall onto electrical network cables, telephone leads and similar, or to cause some other damage. if your roof cover is not covered with tin, the attic could also be a good place for the antenna. You can also stretch it between two stronger laths attached to two opposite window frames of your apartment. If you have no other options, you can put your antenna between the walls in your room.

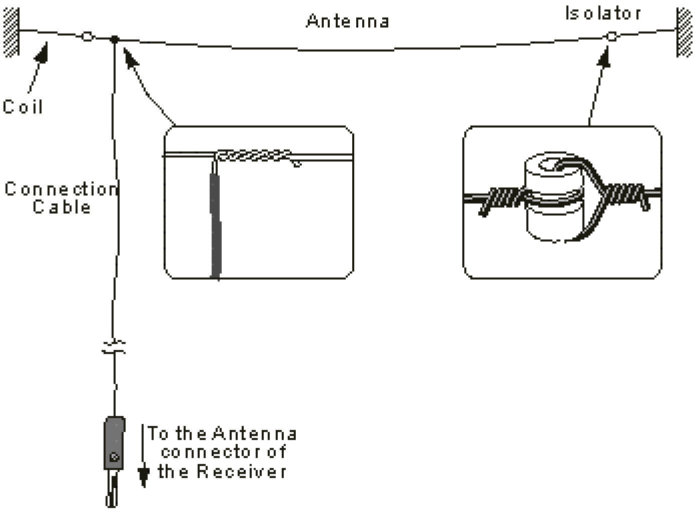

Pic.3.3. Outside Antenna

The antenna must be electrically isolated from the carriers being attached to. In amateur conditions, one can make the isolators of a piece of thick wall plastic pipe where, acc. to Pic.3.3, an indent should be made with the round rasp, in order for the wire not to slip away.

The antenna lead is another piece of wire which carries the signals from antenna into the receiver. It should be isolated and placed in such a way not to touch the walls, to be as far from metal parts as possible (gutters, city grounding etc.).

At the end of this chapter, let's just say that in mobile receivers ferrite antennas are being used, which we are going to talk about later.

3.1.3. The Ground

As all other sorts of ground, the ground for the radio receivers is being accomplished by connecting the receiver ground (point Z on Pic.3.1) to Earth over a coper wire. You can live without the ground but the reception is much better with it though, especially considering simple devices, such as one at the Pic.3.1-a. Water plumbing is an excellent ground (central heating pipes are not), but it is most often inappropriate for use. There is no housewife that would agree to have some dreadful wire stretched across the house, from bathroom to the living room! House electrical installation's ground is excellent, but it should be used under NO circumstances, since life-threatenning danger from electrical shock exists. If you live on a ground floor, and there's plain soil beneath your window you can make your own ground by sticking a piece of water plumbing in it, acc. to Pic.3.4. The pipe should be about 80 cm long, and on its end you should connect the receiver ground, attaching it with a metal ring and a screw with the nut.

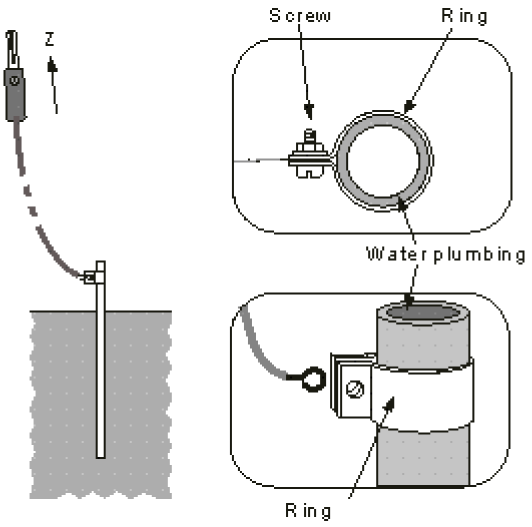

Pic.3.4. The Ground

3.1.4. Other Components

a. On Pic.3.1-a with letters A, Z, S1 and S2 the hubs where one can connect the antenna (A), ground (Z) and the headphones (S1 and S2) are labelled. Since the cabinet for our radio receiver(s) is being made of material that is the electrical isolator (plywood, plastics etc.), the simplest metal hubs can be used, although hubs with isolation plates (for metal plate mounting) can be found in shops more easily, but are significantly more expensive.
b. C1 capacitor is the so-called coupled capacitor, through it the signals from the antenna being led into the oscillatory circuit. Its capacitance depends upon the length of the antenna, and it lies within the limits of few pF (antenna longer than 10 m), up to few dozens pF (a couple of metres long antenna), the optimal value is to be found through experimenting. Every reception antenna behaves as a voltage generator, having its own internal resistance and capacitance. Antenna's resistance damps the oscillatory circuit and reduces its selectivity (which manifests as the "mixing" of stations) and sensitivity (which exerts as signal strength reduction), and antenna's capacitance reduces the reception bandwidth. More precise, antenna's capacitance reduces the upper bound frequency of the reception bandwith (Pic.3.2), making reception of the stations laying close to this frequency impossible. Both these features are undesirable and manifest themselves as

less as the capacitance C1 is smaller. On the other hand, the smaller the capacitance C1, the weaker the signal that goes through it from the antenna, the reception therefore getting weaker. As you can see, the compromise solution is a thing to go for, i.e. one must find the capacitance at which the signals from the antenna won't be much weakened while simultaneously keeping the selectivity and the bandwidth big enough. You can start with C1 being about 30 pF. Then,

using C, tune yourself to some radio stations you can receive. If all the stations that interest you are there, and the strongest one of them still does not jam the reception of other stations all's well. Try then with some bigger capacitance for the capacitor C1. The reception will be getting louder, so do continue increasing C1 as long as it is still possible, by changing C, to receive all the stations of your interest that can be heard in your place, without the interference of some strong or local station. If, however, reception of some nearby station isn't possible, smaller C1 should be tried out. In this manner the biggest capacitance for C1 should be found, that allows optimal reception both regarding selectivity and bandwidth. The simplest solution is using variable capacitor for C1, its capacitance ranging from few picofarads to few dozens pF, adjusting it to obtain optimal reception for each station individually. During this, whenever C1 is being changed, the receiver must be re-tuned to the station using C.

c. The coil is one of the components that cannot be bought, therefore it has to be manufactured. Its main property is the inductance L. As an example, we are going to take a look at how to build a coil for the MW receiver with band range from fd=540 kHz til fg=1620 kHz. The inductance is being calculated using the Thomson formula (being solved by L):

$$L = \frac{1}{4\pi^2 f_g^{\,2}\ (C_{\min} + C_x)}$$

Where Cx denotes the so-called parasite capacitance. It comprises the capacitance of the trimmer capacitor (its average value) that is connected parallel to the variable capacitor C, input capacitance of the next stage of the receiver (where the signal from the input circuit is being lead), antenna capacitance, coil capacitance and capacitance of the connections between the components of the input circuitry. The amount of this capacitance is not known in advance, therefore must be assumed. Taking that value, the coil inductance is calculated and the appropriate coil is made, together with the input circuit. The error being made with the assumption of the capacitance Cx is then compensated with the abovementioned trimmer capacitor. In all our projects this capacitor had minimum capacity Cmin=12 pF. We assumed Cx=15 pF, and therefore:

$$L = \frac{1}{40 \cdot 1{,}62^2 \cdot 10^{12} \cdot (12 \cdot 10^{-12} + 15 \cdot 10^{-12})} = 353\ \mu H$$

We made this coil, conented it with other components from Pic.3.1 and, after some experimenting and measurement, came upon the conclusion that its inductance should be somewhat smaller. We uncoiled a few reels, re-checked the bandwidth, then uncoiled some more, re-checked again, and after several tries came up with the solution. With variable capacitor that will be described in the following chapter, the abovementioned bandwidth is achieved with the coil of inductance L=330 ìH (microhenries).

The coil body i.e. the body where the coil is being reeled is a cylindrically shaped isolation material. For this purpose we have been using the carton core of the household aluminium foil package, its diameter being 3.2 cm. The number of bends required and wire diameter are calculated acc. to the formulas from Pic.3.5.

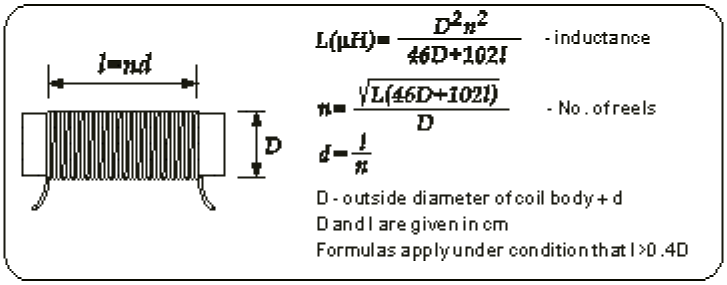

$$L(\mu H) = \frac{D^2 n^2}{46D + 102l}$$ - inductance

$$n = \frac{\sqrt{L(46D + 102l)}}{D}$$ - No. of reels

$$d = \frac{l}{n}$$

D - outside diameter of coil body + d
D and l are given in cm
Formulas apply under condition that $l > 0.4D$

Pic.3.5. Formulas to calculate the coil inductance and number of reels

In order to use these expressions coil length must be assumed first. If this length later proves to be incorrect because the wire is too thick or thin, new length is adopted and the calculation is repeated. Let us assume that coil length is l=4 cm. The number of reels and coil diameter are:

$$n = \frac{\sqrt{330 \cdot (46 \cdot 3{,}2 + 102 \cdot 4)}}{3{,}2} = 134$$

$$d = \frac{40\,mm}{134} = 0{,}298\,mm$$

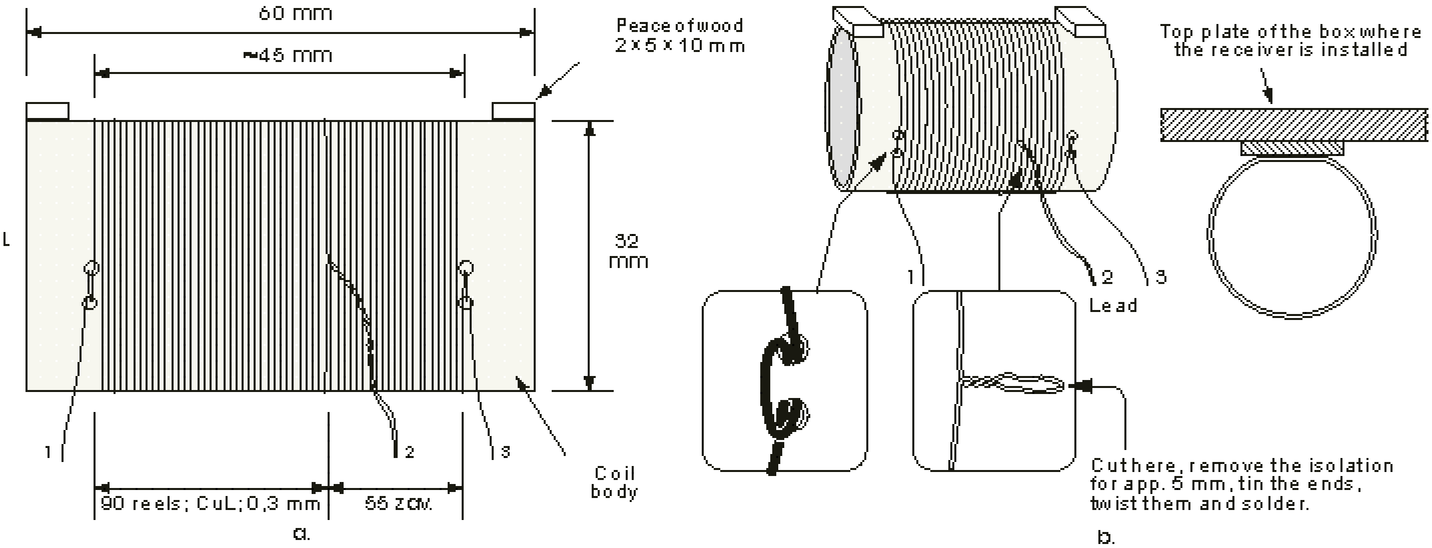
60 mm
~45 mm
Peace of wood
2 x 5 x 10 mm
L
32
mm
1
2
3
Coil
body
90 reels; CuL; 0,3 mm
55 zav.
a.
Top plate of the box where
the receiver is installed
1
2
3
Lead
Cut here, remove the isolation
for app. 5 mm, tin the ends,
twist them and solder.
b.

Pic.3.6. Coil of radio receiver's input circuit: a-dimensions, b-layout and details

Since there is no wire of such diameter, we adopt the closest existing value, d=0.3 mm. In that case the length l will be somewhat bigger, and so will be the number of reels. After a few iterations in calculus and later inductance check of the finished coil, we came upon the n=144 reels of lacquered copper wire (the mark for such wire is CuL), its diameter being d=0.3 mm.

This coil is shown on Pic.3.6. As you can see, two holes are made in coil body (with a bodkin) and through them the wire origin is being threaded twice. After that 90 reels are made, then a leg, then another 55 reels and finally the wire end is again threaded twice, through the other two holes. The leg is made by multiple twisting the wire. It is then cut, and from these new ends about 5 mm of isolation is removed, after which they are tinned, twisted around each other and finally, soldered (the easiest way to remove the isolation is by burning it with lighter, then carefully scrapping it with the pocket knife or similar). Two small pieces of wood are then glued onto coil's ends. When the coil is being mounted into the box, they are pasted onto its top panel, as shown in the rightmost part of Pic.3.6-b.

If you are using a coil of different diameter, you should keep in mind that the necessary inductance for the coil which measures more than 3.2 cm in diameter will be obtained with number of reels less than 144 and vice versa, if the coil body is less than 3.2 cm you will need more than 144 reels.
d. Variable capacitor C is hard to find in stores, therby we have been using in all our receivers the one that we took from a disused commercial pocket size MW radio receiver, the one shown on Pic.3.7. On Pic.3.7-a you can see it together with the reel with numbers that represent the frequencies, divided by 100, on which that receiver was able to be tuned at. On Pic.3.7-b you can see the front, side and rear views of this capacitor. Electrical diagram is given on Pic.3.7-c. As one may notice, there are actually two variable capacitors under the same cover, Co and Ca, and two trimmer capacitors connected parallel to them, Cto and Cta. The dashed line shows symbolically that the rotating plates of the variable capacitors are connected on a common shaft, so that by turning the reel their capacitances are being changed simultaneously. For our use, all four capacitors are parallel connected, by joining the legs O and A. The trimmers are set to minimal capacitance. In such way the variable capacitor is attained with capacitance ranging from Cmin=12 pF til Cmax=218 pF.

In commercial radios that can receive both stations from AM and FM ranges, variable capacitor shown on Pic.3.8 is being used. Four variable and four trimmer capacitors are placed under the same cover. If you wish to use capacitor like this in the receiver from Pic.3.1 (and in most of the receivers described in this book), you should then connect in parallel Cto, Co, Ca and Cta, after which you shall obtain a variable capacitor ranging from Cmin=16 pF til Cmax=286 pF. Other capacitors from this block are not being used.
In all input circuits (more about them soon to come), one end of the variable capacitor is always connected to the device ground. For capacitors shown on Pics.3.7 and 3.8 that is the middle leg, marked as G.
During the dismounting of the capacitor from the old radio, you should pay attention not to loose the screw for the reel attachment, and two screws for mounting the capacitor onto the PCB, since they are very hard to provide separately.

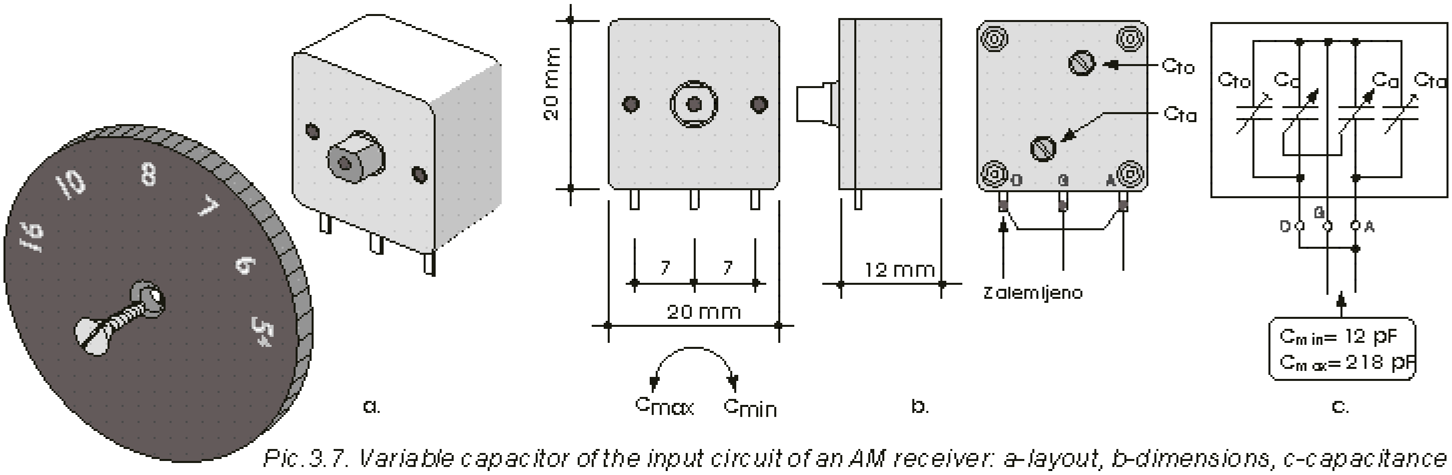
16
10
8
7
6
5
20 mm
7
7
20 mm
Cmax
Cmin
12 mm
Cto
Cta
O
G
A
Zalemljeno
Cto
Co
Ca
Cta
O
G
A
Cmin= 12 pF
Cmax= 218 pF
a.
b.
c.

Pic. 3.7. Variable capacitor of the input circuit of an AM receiver: a-layout, b-dimensions, c-capacitance

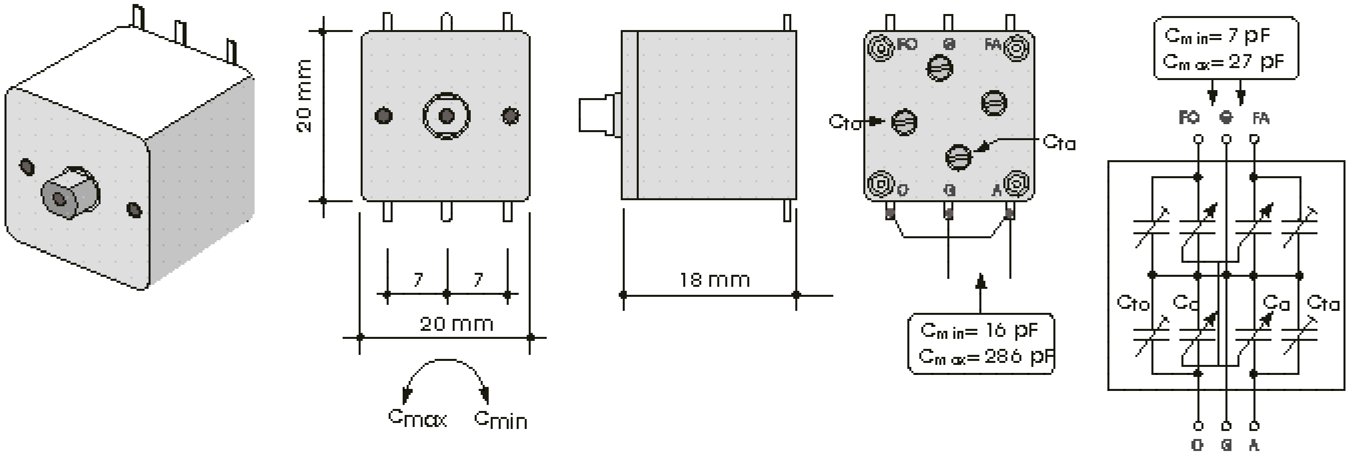
20 mm
7
7
20 mm
Cmax
Cmin
18 mm
FO
G
FA
Cto
Cta
O
G
A
Cmin= 16 pF
Cmax= 286 pF
Cmin= 7 pF
Cmax= 27 pF
FO
G
FA
Cto
Co
Ca
Cta
O
G
A

Pic. 3.8. Variable capacitor of the input circuit of an AM/FM receiver: a-layout, b-dimensions, c-capacitance

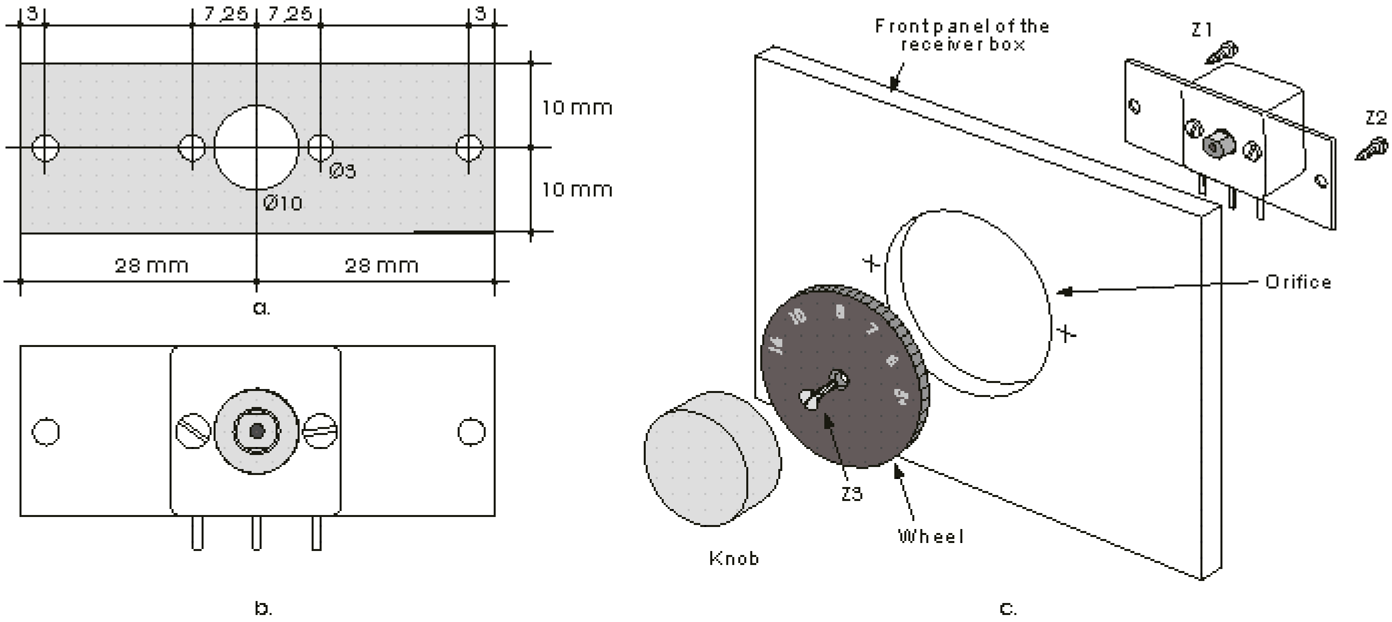

Pic.3.9. Variable capacitor assembly: a-plate dimensions, b-capacitor on plate, c-assemblage on the front receiver panel

If the reciever is being put into the box whose front pannel is made of isolating material not thicker than 1 mm, then one 10 mm hole should be drilled on it, followed by two 3 mm holes, as shown on Pic.3.9-a. Having thicker front panel does represent a problem, the shaft of the capacitor being too short to mounting the reel. In that case you will have to make the auxiliary plate about 1 mm thick as shown on Pic.3.9-1 and then mount the capacitor on it , acc. to Pic.3.9-b. On the front panel a round eye should be made, its diameter being a little bit bigger than the reel. The auxiliary plate with the capacitor should then be tightened onto this front panel with two small screws Z1 and Z2, and the reel on the capacitor shaft with the screw Z3 (While tightening this screw you should hold the reel with your other hand, and not the capacitor housing). Finally, a button made of a thick plywood can be glued to the reel. This is not necessary, but gives the device a more sophisticated looks.
Different variable capacitors than the ones described here can also be used, for example, an air variable capacitor described in the first issue of Practical Electronics. The important thing for it is to have a big max/min capacitance ratio, at least 15, i.e. Cmax/Cmin>15. While connecting the capacitor, care should be taken to connect its rotor with the ground (as on Pic.3.1), labeled Z, and its stator with the point 1 of the coil.
e. The diode D, capacitor C2 and headphones' resistance comprise the AM signal detector, also called the serial diode detector. When the AM signal of the station the receiver is tuned at is brought on its input, NF signal is obtained on the output, its shape being the same as the envelope of the AM signal. An example of this is given on Pic.3.10. When voltage uul is present on the input of the detector, the voltage uizl appears on its output. It is useful to notify that on the output, besides the LF voltage (speak, music etc.), DC voltage Uo is also present.
The detection diode D must be of low-power GERMANIUM type, such as AA112, AA116, AA121, 1N21, 1N34, 1N54, 1N78 etc.

Product of the capacitance C and resistance R (on Pic.3.1. R is the headphones resistance) should be approx. equal to 50 ìs (microseconds). That means that if you're using the bigger resistor (which is advisable, since the detector then damps less the oscillatory circuit), the capacitor should then be smaller. For example, if

R=500 kÙ then C=100 pF, and if R=10 kÙ, C=5 nF, etc.

f. The headphones are the electro acoustic convertor that transforms electrical signal into the sound. We have been using old fashioned electromagnetic headphones with 1.5 kÙ resistance that were serially connected, giving the total resistance of 3 kÙ. The receiver from Pic.3.1 will be working as better as headphones' resistance is bigger. if you're using the crystal headphones, parallel to them you should add a resistor of couple of hundreds of kiloOhms. There's a very simple way of testing the high resistance headphones: Hold one end of their cable between your fingers while rubbing the other over the surface of a big metal object, say, the radiator. If snapping can be heard from them they are, most likely, satisfactory.

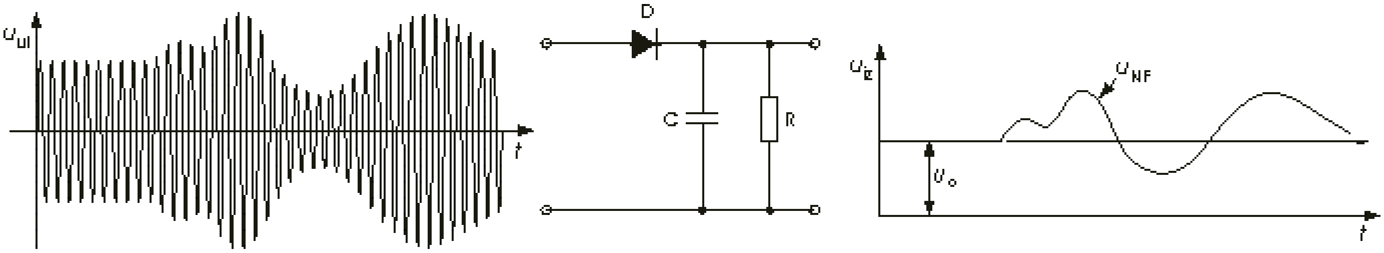

Pic.3.1. Input and output voltage at the serial diode detector

All the components of the receiver from Pic.3.1 should be placed into some kind of a box. That can be any box made of an isolation material (plastics, wood etc.), big enough to receive all the components. As an example, a receiver is shown in scale 1:1 on Pic.3.11, being placed in a box made of plywood. The top, bottom and side panes are made of plywood that is 5 10 mm thick. The front and the rear side are being made of some thinner material, that allows for simple mounting of the variable capacitor. One can notice straightaway that the box is at least twice as big as it could be. That has been done for the sake of better visibility, and for the box to be big enough to accept the devices that will be described later in this book.

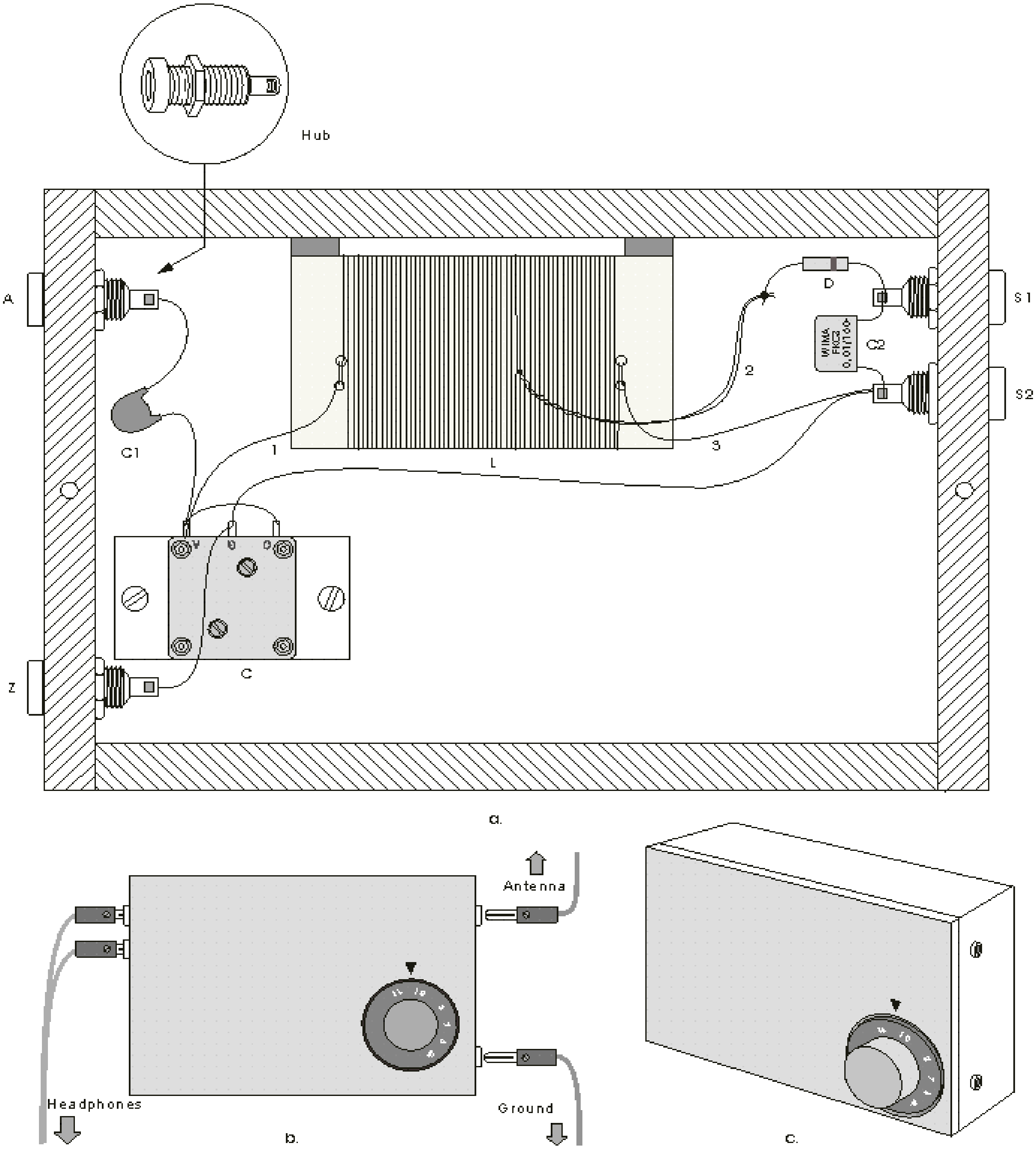

Pic.3.11. Practical realization of the Detector Receiver from pic.3.1-a: a-component layout in the box, b-front panel view, c-3D view

Component joining is being done by soldering. Performing soldering onto the hubs can represent a small problem. In order to complete this operation successfully, you should turn the box for the part of hub where cables are being soldered pointing upwards. Put a piece of the tinol wire through the eye on the hub, put the soldering iron top onto the eye from upwards and hold it like that a while, for the tinol to melt. Then add some more tinol, until the eye is completely covered with the solder. After that, push the wire end into the melted solder while holding it with pencers, until it hardens and cools itself a little. The pencers remove the heat and prevent its transfer onto the component that is being soldered.

3.2. The Simplest Amplified Radio Receiver

The most obvious shortcoming of the receiver described in the previous chapter is that it can perform the sound reproduction loud enough only in case when the programme from some local or very powerful radio transmitter is being received, which can create very strong signal in the reception antenna. The reception of signals from other transmitters is too weak. The only thing that can be done is either to increase the length of the antenna, which, of course, does have its limits, or to insert an amplifying stage into the receiver. The simplest way to perform the latter is to add a LF amplifying stage behind the detector in the detection receiver, Pic.3.1. The electrical diagram of one such receiver is given on Pic.3.12.
The electrical load in the detection stage are no longer headphones, but an ordinary resistor R1. An NF signal is obtained on its ends, which is then being lead into the LF amplifier with the transistor T, over the coupling capacitor C3. The electrical load in the collector circuit of the transistor are the headphones, which transform the amplified LF signal into sound.
The voltage negative feedback is being obtained with the capacitor C4, and the current negative feedback with the resistor R3. They enhance the characteristics of the amplifier (increase its stability, reduce distortion, widen the reception band), but they also reduce the amplification. The capacitor C5 prevents the AM signal carrier which, although very weak, also appears on the detector output, from entering the headphones. These 3 components however, can be omitted in most cases, for the sake of simplifying the device. C4 and C5 can be simply removed, and a piece of wire should be soldered instead of R3. Transistor operation point where the optimal reproduction (the biggest amplification, the smallest distortion) is being set by adjusting the resistence value of the resistor R2. The simplest way to do it is to connect the trimmer of couple of MÙ instead of the resistor, set the receiver on some station, then change the resistence until the optimal reception is being reached. The trimmer is then put out, its resistence measured, and a resistor of similar resistivity is then soldered on its place. The transistor T is any universal NPN - type.

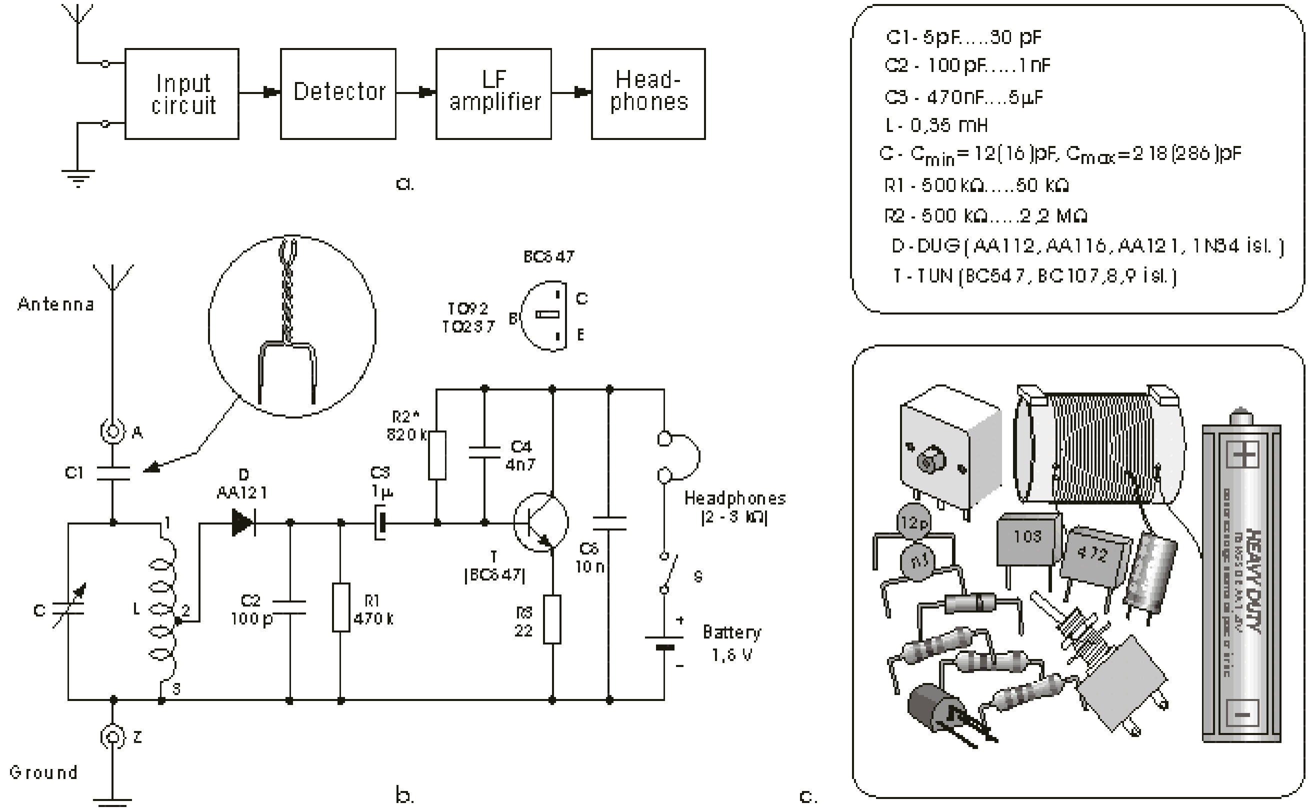

C1 - 5pF.....30 pF
C2 - 100 pF.....1 nF
C3 - 470nF....5μF
L - 0,35 mH
C - C_{min} = 12(16)pF, C_{max} = 218(286)pF
R1 - 500 kΩ.....50 kΩ
R2 - 500 kΩ.....2,2 MΩ
D - DUG (AA112, AA116, AA121, 1N34 isl.)
T - TUN (BC547, BC107,8,9 isl.)

Pic.3.12. Detection receiver with simple amplifier: a-block diagram, b-electronic diagram, c-components

Please note that in the case of a very long antenna, when C1 of very small capacitance is being used, a hand made trimmer capacitor can be used. It is made by twisting two isolated wires, and its capacitance is changed simply by removing some wire from the ends.
The component data is given on the electrical diagram and in table on the rightmost side of Pic.3.12. If you don't have a 100 pF capacitor (C2) you can put some of bigger capacitance, but you should then use smaller R1.
The PCB layout is given on Pic.3.13. On 3.13-a is a picture that should be copied with the thin alcohol marker onto well cleaned copper side of the pertinax plate. Etching is to be performed then, as well as drilling the 0.8 mm holes, in the way that has been described in detail in PE2 issue (Practical realization of electronic devices). The component layout is given on Pic.3.13-b. The printed circuit is also visible on this picture, and that can be achieved by using vitroplast plate instead of pertinax. The look of board upon completion is on 3.13-c. Before soldering the wires that connect the variable capacitor, battery, switch etc. with the board, put some tin on the wire ends, using the calofonium or the tinol wire. You should by all means do this, especially if you 're not using the wire that is not pre tinned, in order to avoid cold solders, very unpleasant surprises that are hard to detect and locate.
The layout of the entire receiver is given on Pic.3.14, in a scale 1:1. As in the previous receiver example, the coil is being glued to the top side of the box, over two small pieces of wood. The board is being tightened with a screw, that is screwed into a piece of wood being glued to the front end box plate. The battery is attached with a rubber band for the pieces of wood glued at the back plate of the box. We did this, however, only for the sake of having a clear and understandable drawing. Battery can be put inside the box in the same way as before, or some other way.

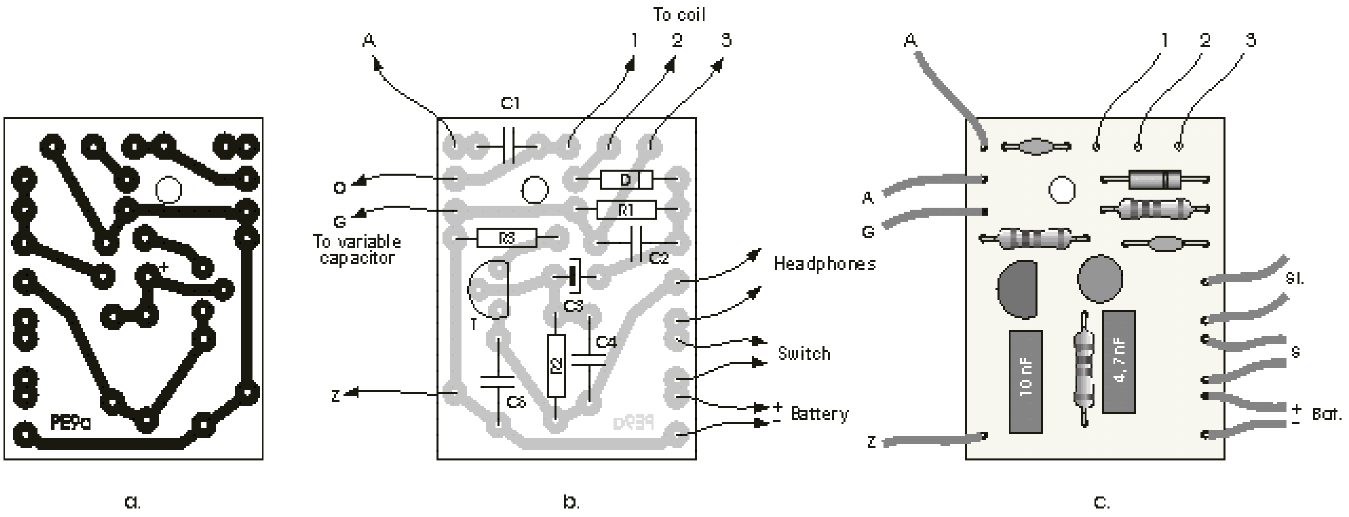

Pic.3.13. PCB for the receiver from pic.3.12: a-soldering side view, b-component side view, c-component layout

3.3. Simple Radio Receiver with TDA7050 IC

Readers that were able to browse through book 4 and, especially, book 5 of Practical Electronics edition, were able to convince themselves that there's a huge number of various audio amplifiers built with IC's, therefore amplifiers in discrete technique are practically no more being made. If we add on this the fact that integrated amplifiers outreach their discrete competition both by price and quality, it is then clear why we are going to use them in this book.

Electronic diagram of a simple radio receiver with LF stage built around the TDA7050 IC, where reproduction is being made through modern-type headphones, resistance being 32 or 64 Ù, is given on Pic.3.15. In book 5 of Practical Electronics you have acquainted yourself in more detail with this IC. Let us just repeat that it can be purchased both in 8-pin DIL package that we have been using, and in SO package, suitable for surface mount. Its label in the latter case is TDA7050T and it can also be used without any problems whatsoever. In that case changes on the PCB should be made considering that it is being mounted on the copper side of the board, and that drilling is now obsolete, since the pins are being soldered directly onto the copper foil. You can read the text that follows Pics.4.16 and 4.17 about soldering SMD components. Supply voltage for the IC is in range from 1.6 to 6 V. Idle current is 3 mA on 3 V supply voltage. Voltage amplification is 32 dB (40 x) on 6 V supply voltage and 32 Ù headphones resistance. Maximum output power is 150 mW, more than enough for the headphones amplifier device.

Regarding the Pic.3.15, both input and detection circuit described in two previous projects can be used, We have decided instead, to show you how to use the coil with multiple legs, since it offers more possibilities for experimenting in order to achieve optimal reception. The picture of such coil is given on Pic.3.16, the legs being made as previously described in this book. The first leg (numerated 6) is made after the 15-th quirk, the second (5) after 30-th, the third (4) after 45-th and the last one (2) after 55-th. Number of quirks between the legs isn't critical, you can have even more legs, being arranged more closely to each other. As seen on Pic.3.15, both the antenna and the detector are connected over these legs. Legs No. 4 and 5 have been used, but that is not a must. The closer the legwhere antenna is connected is to the ground (point Z), the less it damps the oscillatory circuit (therefore increasing the receiver's selectivity), and less it reduces the reception bandwidth. However, the signal that exits the antenna is then also smaller. The similar thing is with the leg where the diode (its anode end) is connected: The closer it is to the ground, the less the detector damps and tunes out the input circuit, but the signal being transferred to the detector is also smaller. It is clear now that a compromise must be made: Experimenting with various coil legs, those providing the optimal reception should be found.

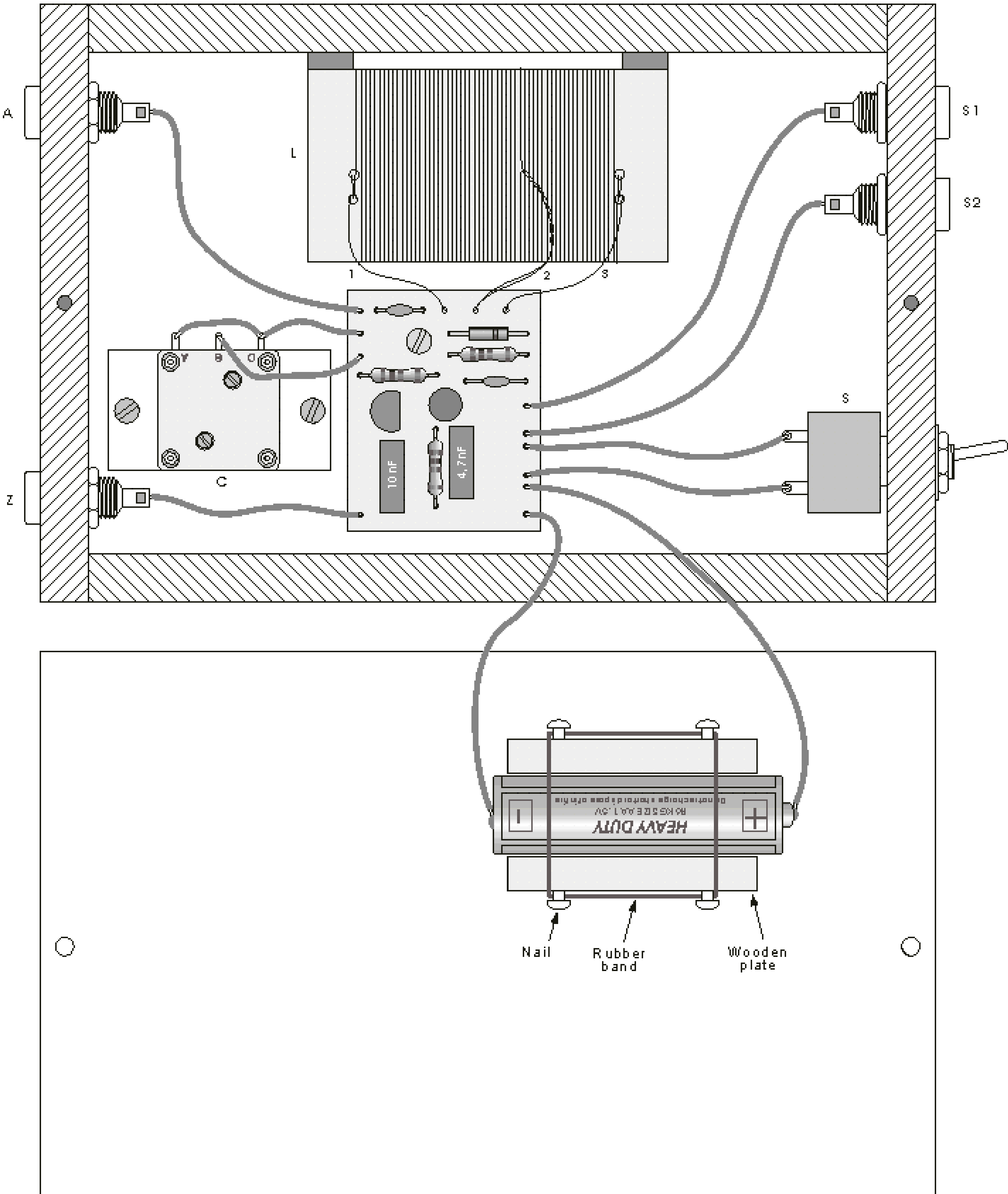
A
L
S1
S2
1
2
3
S
C
Z
10 nF
4,7nF
HEAVY DUTY
Nail
Rubber band
Wooden plate

Pic.3.14. Practical realization of the detector receiver from pic.3.12

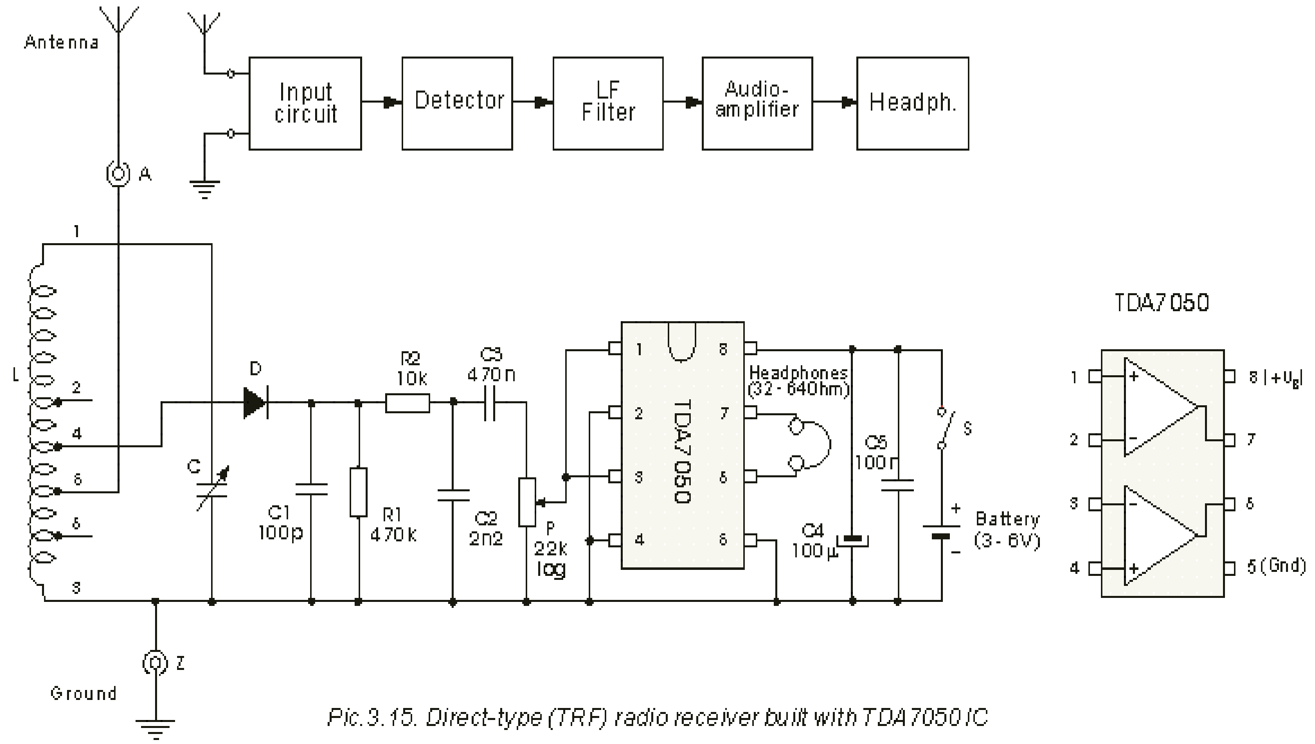
Antenna
Input circuit
Detector
LF Filter
Audio-amplifier
Headph.
A
1
L
2
4
6
6
8
C
D
C1
100p
R1
470k
R2
10k
C2
2n2
C3
470n
P
22k
log
TDA7050
1
2
3
4
8
7
6
5
Headphones
(32 - 640hm)
C5
100n
C4
100µ
S
Battery
(3 - 6V)
Z
Ground
TDA7050
1
2
3
4
8 (+UB)
7
6
5 (Gnd)

Pic.3.15. Direct-type (TRF) radio receiver built with TDA7050 IC

Let us finally add that the antenna can also be connected to point 1, over the coupling capacitor (C1 on Pic.3.1), and the detector on one of the coil legs, or you can connect both the antenna and the diode's anode on the same leg, to connect the anode to point 1, etc. You should try various combinations out, observing their influence on the reception. You should certainly re-tune the resonance of the oscillating circuit after every change, with the aid of the variable capacitor C.

The resistor R2 and capacitor C2 create an LF filter, whose role is to pass through at the next stage the LF signal being detected, preventing simultaneously the HF voltage do the same (this voltage originates from the AM signal carrier). This filter circuit affects the LF signal tone colour. If you don't like it, you should alter the capacitance C2 in order to change it.

PCB for this device is shown on Pic.3.17. The receiver can be put in a box just as on Pic.3.14. The only significant difference is that a 4.5 V battery pack must be used instead of 1.5 V battery, but there's plenty of room for it.

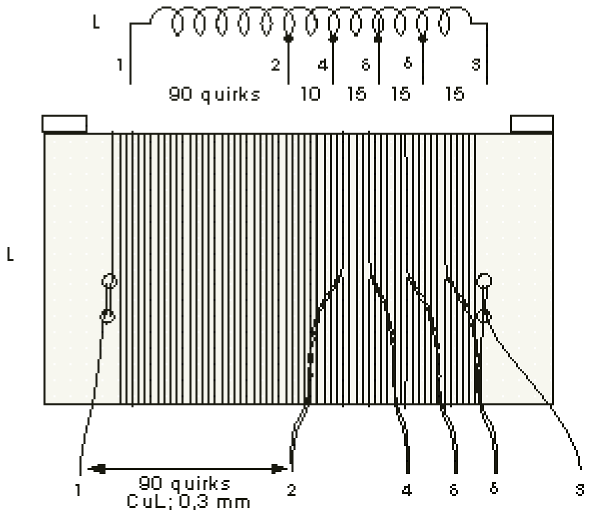
L
1
2
4
6
5
3
90 quirks
10
15
15
15
L
1
90 quirks
CuL; 0,3 mm
2
4
6
6
3

Pic.3.16. a-Multiple-leg coil for MW, for the receiver from pic.3.15

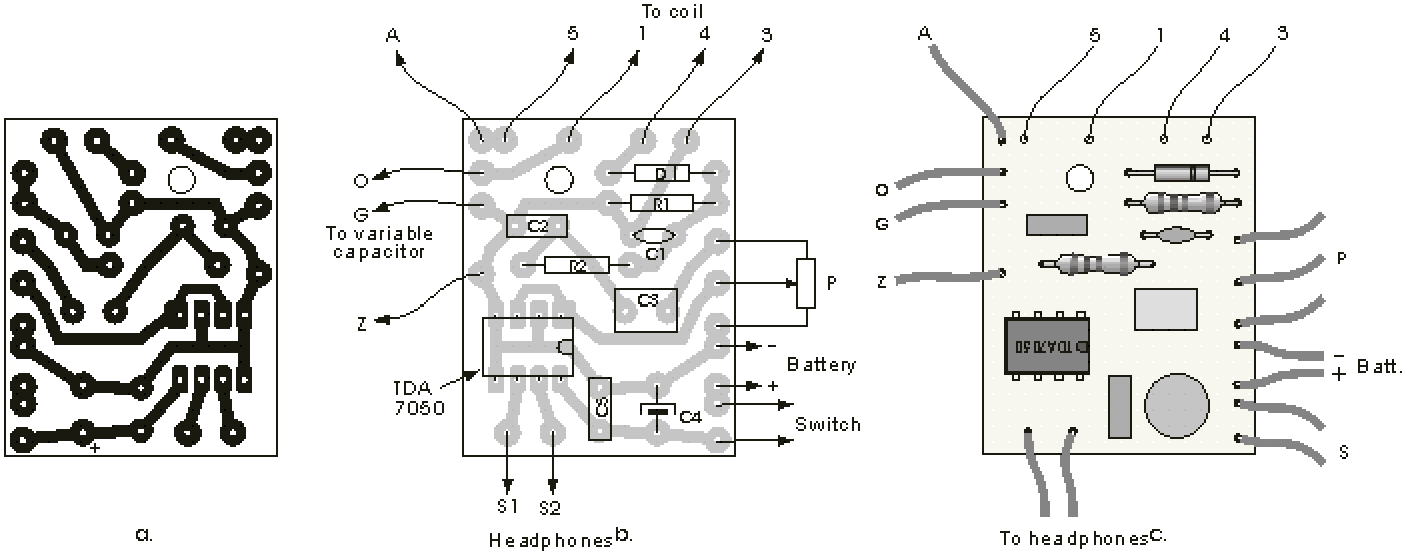
a.
A
5
To coil
1
4
3
O
G
To variable capacitor
Z
D
R1
C2
C1
R2
C3
P
TDA 7050
C4
Battery
-
+
Switch
S1
S2
Headphones
b.
A
5
1
4
3
O
G
Z
P
TDA7050
Batt.
-
+
S
To headphones
c.

Pic.3.17. PCB for the receiver from pic.3.15: a-soldering side view, b-component side view, c-photo-view

3.4. Simple Radio Receiver with LM386 IC

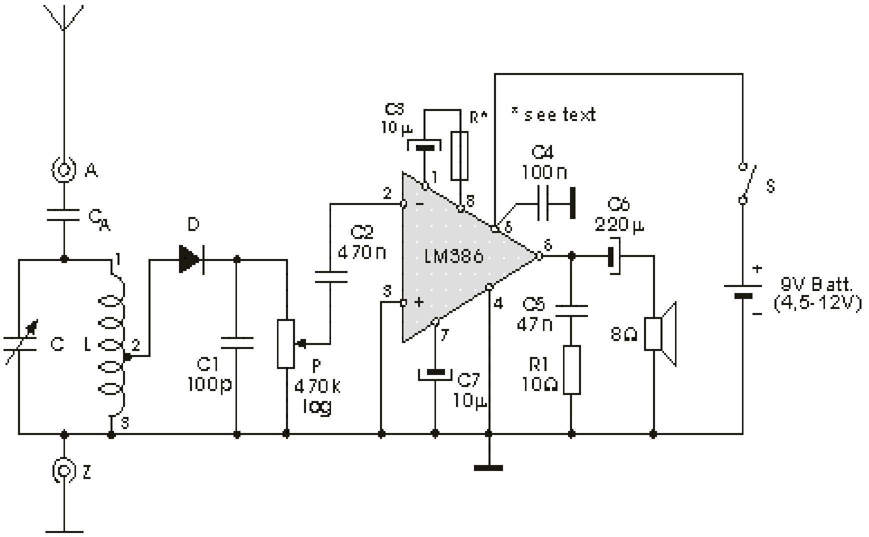

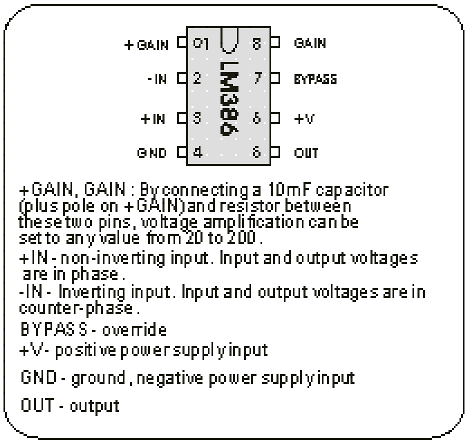

+GAIN, GAIN : By connecting a 10mF capacitor (plus pole on +GAIN) and resistor between these two pins, voltage amplification can be set to any value from 20 to 200.

+IN - non-inverting input. Input and output voltages are in phase.

-IN - Inverting input. Input and output voltages are in counter-phase.

BYPASS - override

+V- positive power supply input

GND - ground, negative power supply input

OUT - output

Pic.3.18 a-Electronic diagram of the detector (TRF) radio receiver with audio amplifier built around the LM386 IC

Listening the programme over the headphones has its advantages, but the true radio receiver is certainly the one that with the loudspeaker. In this project the simplest such receiver is being described, consisting of input circuit, diode detector and audio amplifier with LM386 IC, that allows the loudspeaker reproduction to be performed.

The author has firstly met with the LM386 circuit over twenty years ago, and has ever since been using it very successfully in various devices. The fact that this IC has survived at the market for such a long time is a considerable proof of its quality, however, its most significant advantage remaining extremely low price. The readers can find more information on this circuit in book 4 of PE. Purchasing this IC, one must have in mind that it is being manufactured in several versions, marked as LM386, LM386N-1, LM386N-3 and LM386N-4, that differ themselves by the supply voltage values and the output power. In case the supply voltage being no greater than 12 V, any of these can be used in this receiver.

Electronic diagram of the direct radio receiver that has a LM386 IC in its LF stage is given on Pic.3.18. The resistor in the detection circuit is the log type 470 kOhms potentiometer. The LF voltage is being led over its moving end and the coupling capacitor C2 onto the inverting input (leg No.2) of the LM386. Other, non inverting input (leg No.3) is connected to the ground. The output is on the leg No.5. On this leg the load (i.e. 8 Ohm impedance loudspeaker) is being placed, connected over the C6 capacitor to the ground. With smaller battery power supply voltages a 4 Ohm impedance loudspeaker can also be used. If there's nothing in between the legs 1 and 8, the voltage amplification of the IC is Au=20, in which case the capacitor C7 can be omitted. If, however, a 10 ìF electrolytic capacitor is connected between legs 1 and 8 (+ end on leg 1) as shown on Pic.3.18, the amplification is Au=200. Adding a resistor in line with the abovementioned capacitor, any amount of amplification between 20 and 200 can be achieved. This resistor is shown in dashed line on Pic.3.18 and is being marked with "*" sign.

The PCB and components layout for the receiver shown on Pic.3.18 are given on Pic.3.19. Miniature loudspeakers from the pocket-size radio receivers should be avoided, since their efficiency and sound quality are poor, especially in the low frequencies area. On Pic.3.19 a 1W loudspeaker is being shown, whose membrane is 8 cm in diameter. That isn't such a bad solution, but even better would be using the loudspeaker with greater power and membrane diameter (During the testing the 3 W , 8 Ohm loudspeaker has also been used, and the sound quality was much better than with the one that is shown on the picture). As you can see, the cables connecting the loudspeaker with the PCB are firmly twisted around each other. This is a must, especially for the cables being longer than a dozen centimetres. The same has to be done with cables that connect PCB with the battery and the main switch Z.

Significantly louder reception, with bandwith being narrower in the area of the lower bound frequency, can be accomplished with the coil that has smaller inductivity and the coupling capacitor CA with greater capacitance. For example, if you are interested in radio stations that emits the signal in the MW area from 750 kHz to 1700 kHz, you could use the coil that has been previously described, by using only its' part between the legs 1 and 2, as shown on Pic.3.20. For the antenna that is about 6 m long use the CA=33 pF, and the reception is going to get better. Try increasing the CA capacitance (47 pF, 100 pF, 200 pF), the reception is going to get louder and louder, but the reception bandwidth will be getting smaller and smaller, some stations will not be heard any more, others will start to "mix" etc. it is up to you to find an optimum that suits you best. If using very short antenna, you should omit CA, and connect the antenna directly to the oscillatory circuit (at pt. 1).

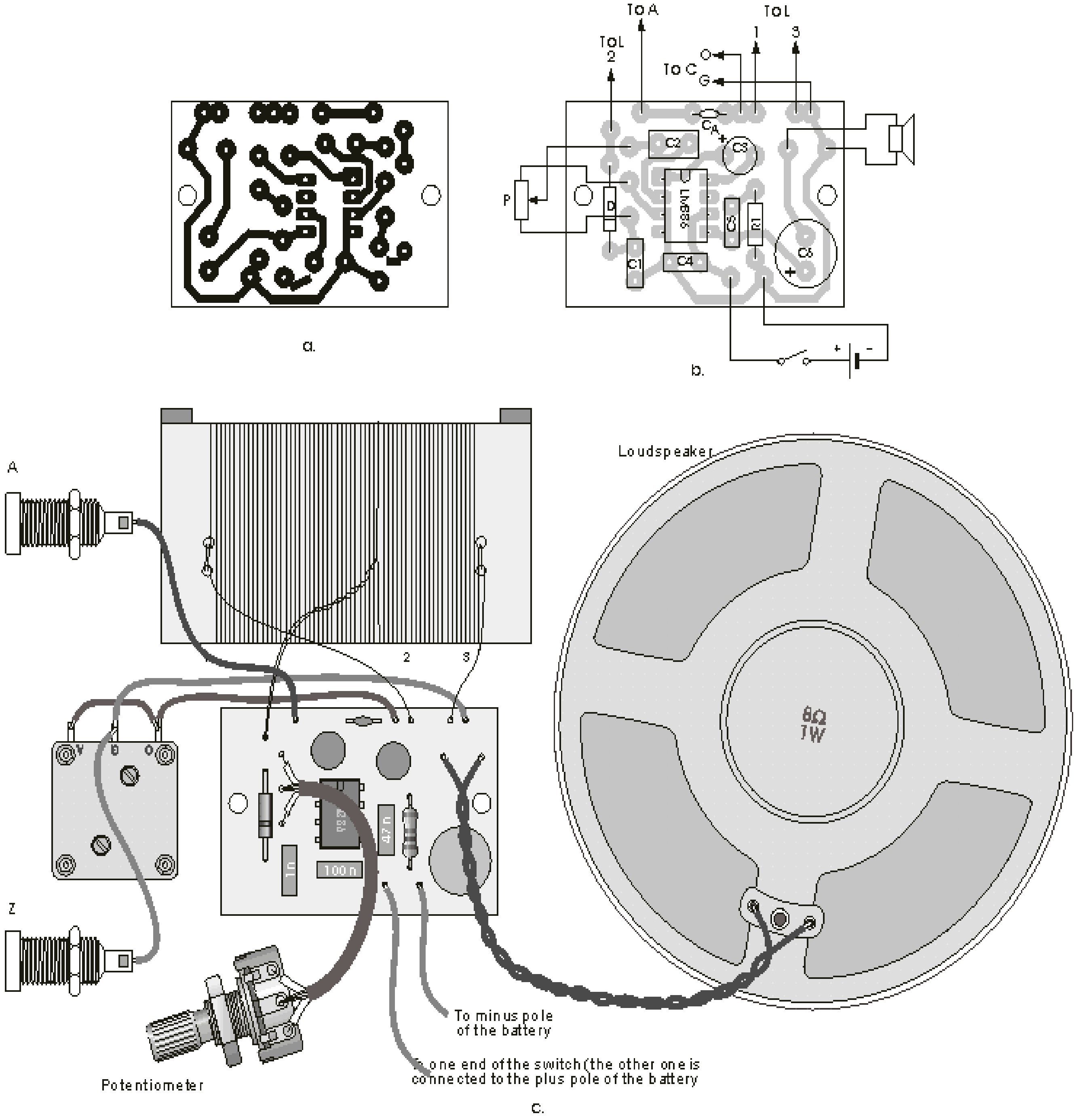
To A
To L
2
To C
O
G
To L
1
3
C2
CA
C3
P
D
LM386
C6
R1
C5
C1
C4
+
-
a.
b.
A
Loudspeaker
2
3
8Ω
1W
Z
1n
100 n
47 n
To minus pole
of the battery
To one end of the switch (the other one is
connected to the plus pole of the battery
Potentiometer
c.

Pic. 3.19. Practical realization of radio receiver from pic. 3.18: a-PCB, b-component layout, c-connecting the board with other components

3.5. Radio Receiver with Increased Sensitivity Audio Amplifier

If you cannot obtain the signal reception that is loud enough with receiver from the previous project, the sensitivity of the LF part of the device must be increased. The simplest way to manage this is adding a transistor pre-amplifier, as shown on pic.3.21. Please note that all the components except C1, C and L are located on the PCB, which allows you to use the input circuit from pics.3.15 and 3.20. The receiver is supplied from the 4.5 V battery but, if you intend to use an adaptor, the supply voltage can be bigger, its maximum value being from 12 V (for the LM386) till 18 V (for the LM386-4).

The optimal value for R3 resistor (obtaining the greatest amplification, minimal distortion etc.) depends upon the kind of transistor that is being used. The easiest way to finding it is by experimenting, with a few MOhms' trimmer resistor, in the way it has already been described in chapter 3.2.

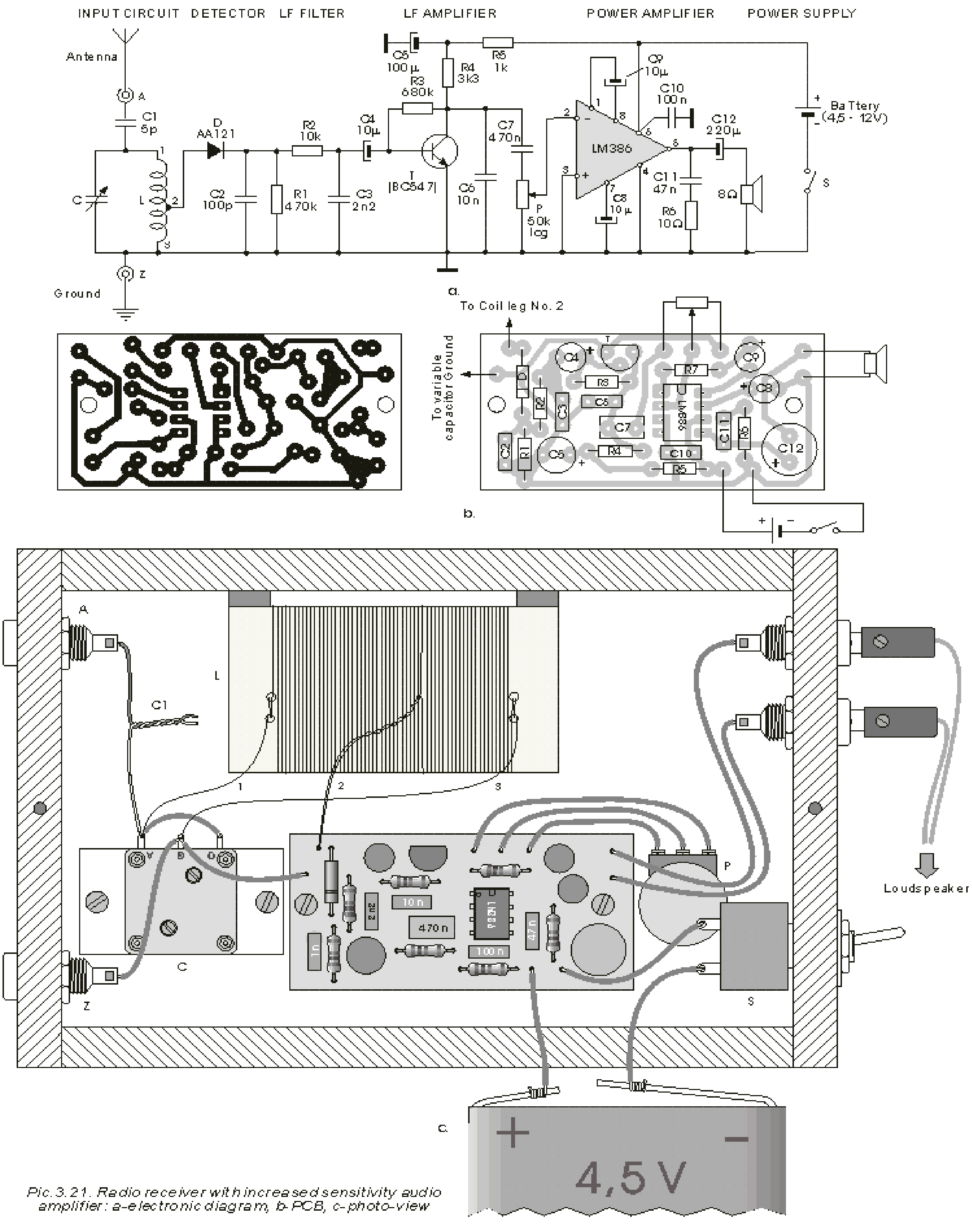
INPUT CIRCUIT
DETECTOR
LF FILTER
LF AMPLIFIER
POWER AMPLIFIER
POWER SUPPLY
Antenna
A
C1
5p
D
AA121
R2
10k
C4
10µ
C5
100µ
R3
680k
R4
3k3
R5
1k
C9
10µ
C10
100n
C12
220µ
Battery
(4,5 - 12V)
C
L
1
2
3
C2
100p
R1
470k
C3
2n2
T
BC547
C6
10n
C7
470n
P
50k
log
LM386
C8
10µ
C11
47n
R6
10Ω
8Ω
S
Z
Ground
a.
To Coil leg No. 2
To variable capacitor Ground
b.
Loudspeaker
c.
4,5 V

Pic.3.21. Radio receiver with increased sensitivity audio amplifier: a-electronic diagram, b-PCB, c-photo-view

3.6. Universal Audio Amplifier

In majority of the receivers described so far, as well as in many those that will be described further, an audio amplifier is being used. This makes it useful to build an universal amplifier that would be used during testing of all receivers. That can be the circuit from pic.3.21 (from C4 inclusive to the right), but it can be something else, too, e.g. the one from the pic.3.22. Two IC's are used in it: operational amplifier TLO71 as the pre-amplifier, and LM386 as the power amplifier. The TLO741 is exactly the same as the more famous 741, the only difference laying in fact that TLO71 has the FET on its input, and 741 - the bipolar transistor. Since inverting input is being used in the operational amplifier, its voltage amplification is given by Au=RP/R1, where RP stands for the potentiometer resistance from the sliding contact to its left end. When the slider is at its rightmost position, it is then RP=470 kOhms, yielding Au=470, and when it's at its leftmost position the RP=0, therefore Au=0. As you can see, moving the slider changes the amplification from 0 to 470. That's what theory says, the practice is, of course, somewhat different. When capacitors are being connected between the input and ground, maximum amplification gets lower than 470 but is still more than enough for our needs.

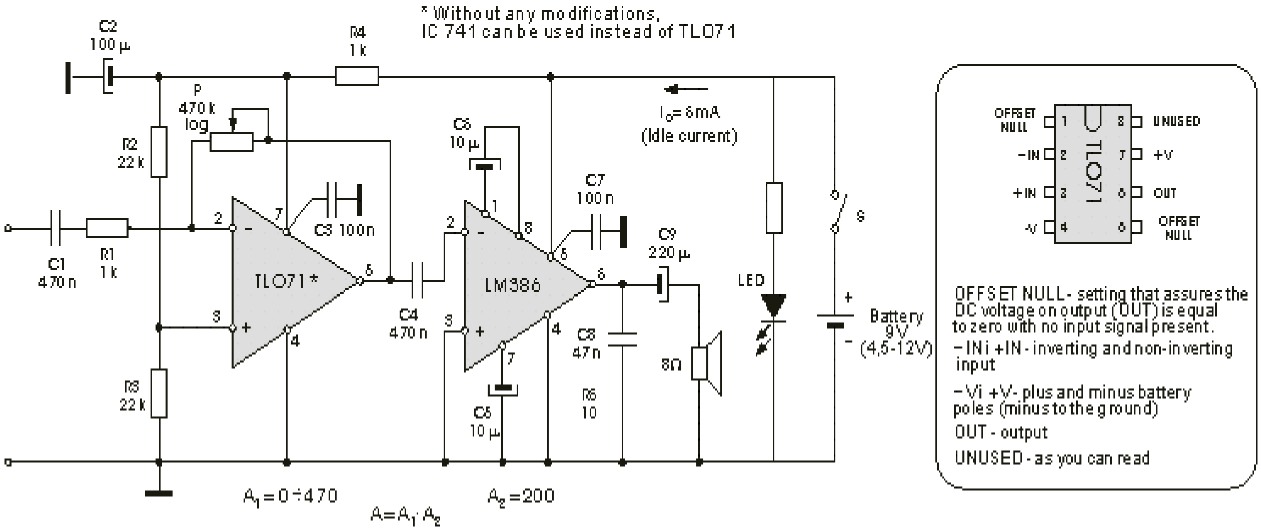
C2
100 µ
R4
1 k
P
470 k
log
R2
22 k
C6
10 µ
C7
100 n
I0 = 6mA
(Idle current)
C1
470 n
R1
1 k
TLO71*
C3 100n
LM386
C4
470 n
C9
220 µ
LED
S
Battery
9V
(4,5-12V)
C8
47 n
8Ω
R3
22 k
C5
10 µ
R6
10
A1 = 0 ÷ 470
A = A1 · A2
A2 = 200
OFFSET NULL
-IN
+IN
-V
UNUSED
+V
OUT
OFFSET NULL
TLO71

* Without any modifications, IC 741 can be used instead of TLO71

OFFSET NULL - setting that assures the DC voltage on output (OUT) is equal to zero with no input signal present.

-IN i +IN - inverting and non-inverting input

-V i +V - plus and minus battery poles (minus to the ground)

OUT - output

UNUSED - as you can read

Pic.3.22. Audio amplifier built with TLO71 and LM386

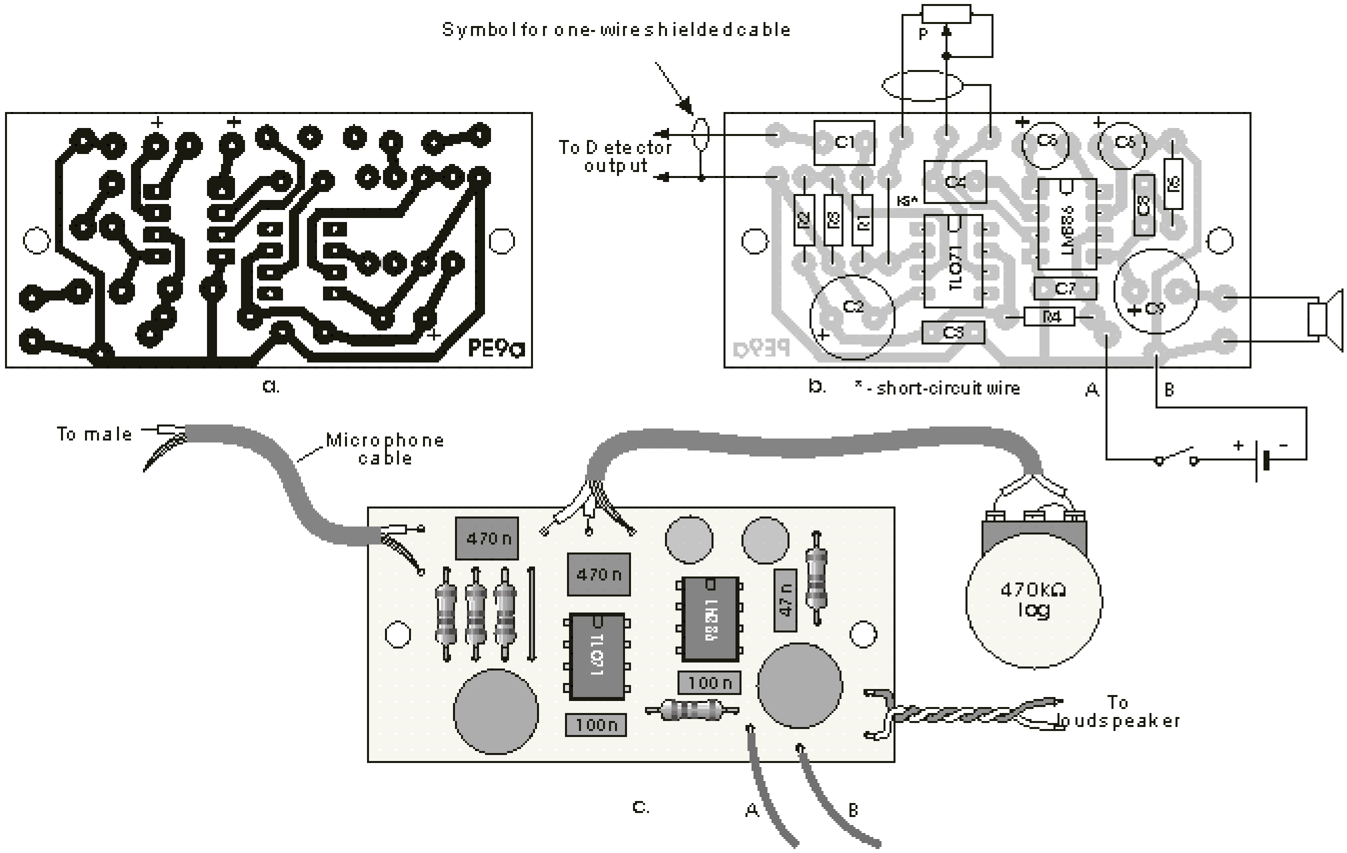
Symbol for one-wire shielded cable
To Detector output
P
C1
C6
C5
C4
KS*
R2
R3
R1
LM386
C8
R5
TL071
C7
C2
C9
R4
C3
PE9a
a.
b.
* - short-circuit wire
A
B
To male
Microphone cable
470n
470n
47n
100n
100n
470kΩ log
To loudspeaker
c.
A
B

Pic.3.23. PCB for the amplifier from pic.3.22: a-print side view, b-component side view, c-"photographic" view

* All the receivers that have been described so far, as well as all those to come, can be tested with this universal amplifier. That is done by taking the signal from the Detector output (through two wires as short as possible or through the microphone cable that can be significantly longer) to the amplifier input (between left end of C1 and Ground). If the antenna, input circuit (and other stages, if exist) are OK, you will be able to hear the tuned station program from the loudspeaker. The sound volume can be adjusted by moving the slider of the potentiometer P. PCB and component layout are given on pic.3.23.

* KS marks the short-circuit wire. It's a piece of wire which ends are taken through the holes on the plate and soldered. It can be omitted, but then the left end of potentiometer is to be soldered in such a way to be connected to the bottom end of R1.

* To connect the potentiometer with the PCB, as well as the male, a shielded cable should be used. The shield (outside conductor) is connected to the Ground. If cables are shorter than app. 10 cm, the ordinary wires can be used instead.

* The amplifier, the loudspeaker and the battery (or adapter) are put in the common box, on whose front plate the potentiometer, the switch and some simple plug (the chinch will do) are mounted. If the device is powered from the adapter, it is useful to also mount a LED. With these parts you have made yourself the so-called "active loudspeaker", which is a

very useful gadget for every electronics lover.

* Just before you start mounting and soldering the components, clean well the PCB copper conductors with a piece of wet sponge dipped in some cleaning powder. Components should be mounted and soldered in the following order: short-circuits, resistors, IC's, capacitors (all except C4), connecting wires and microphone cables (if used). The last component to be mounted is the C4 capacitor, but first you have to check whether the amplifier operates correctly.

Connect the loudspeaker, and the wire marked as "B" on pic.3.23-c connect to the minus battery pole. The other end of the wire marked as "A" should be connected to one probe of the multimeter, which is to be set to measure DC current. Contact the plus battery pole with second instrument probe, and it will measure the idle current of the amplifier, which should be app. 5 mA. If this is OK, touch the pin No.2 of the LM386 with finger, and if everything is fine you will hear the 50 Hz hum from the loudspeaker. The power amplifier is OK. You can now solder C4. Touch the left end of the wire that is connected to the C1 with your finger, the hum should be heard again from the loudspeaker, its loudness depending upon the position of the slider of the potentiometer P. (If you don't possess an instrument, connect the battery on the wires "A" and "B" and do the touching described. If you hear the hum, the amplifier is OK).

If the Idle current is zero or is significantly bigger than 5 mA, something is wrong. If it's zero, check out whether you have connected your instrument properly and whether it is functioning. Remove the instrument and connect the battery directly to the wires A and B. Touch the pin No.2 of LM386 with your finger, if you hear the hum - your instrument is malfunctioning. However, if the instrument is OK, you should check the DC voltages. Set your multimeter to measure DC voltages, connect one probe (the one marked with the ground sign) to the amplifier ground (minus battery pole) and with second probe touch the plus battery pole first, and then pin No.6 of LM386. These two voltages should be equal to the battery (supply) voltage. Contact then pin No.7 of the TLO71. This voltage should be slightly smaller than the supply voltage. Voltages at pin No.5 of LM386 and No.7 of the TLO71 should be app. equal to the one half of the supply (battery) voltage.

Try to locate the malfunction based on your measurements. Check out if there are some discontinued copper lines on the print, or if two adjacent lines are connected (by themselves or by a small piece of tin that you dropped from the iron during soldering), whether all junctions are the way they should be, if you happened to mount some electrolytic capacitors or IC's upside-down, etc.

When you make sure that everything is fine, un-solder the potentiometer and the loudspeaker and put them on the front box panel. Also mount the plug and the switch, the LED (if used) etc. Tighten the plate with two small wood screws to the side pane, placing the LM386 on top (above the TLO71). Two distant-rings are to be mounted between the plate and the wood, making the distance between them a few (say, 5) mm. (These rings can be cut from some small plastic pipe). The screws are first to be put through the plate holes, then the rings and then screwed in the panel. After all this you can connect the loudspeaker, the switch, the battery, etc.

3.7. Receiver with the HF Amplifier

In HF amplifier the signal coming from the radio station is being amplified in its original form. In our case, this means that AM signal is led at input of the HF amplifier, and on its output the same shaped signal is obtained, only with bigger amplitude. This device got its name because it is used to amplify HF signals, although more precise term for it is the Selective Voltage Amplifier (that's how it is called in professional books).
Electrical diagram of a direct receiver consisting of HF amplifier, diode detector and LF (audio) amplifier is presented on Pic.3.24. This receiver does not have a selective input circuit. It would be better that way, the selectivity of the receiver would be better, but "technical reasons" made us not to include it: the double variable capacitor (like the one in the first issue of PE) i.e. the one consisting of two equal variable capacitors connected onto the common shaft, is very hard to find these days. The active element of the HF amplifier from Pic.3.24 is the transistor BC557 that operates in the common base junction. The station signal that is being amplified is led onto the emitter (i.e. between the emitter and ground), and the amplified signal is taken from the collector (i.e. between it and ground). Working principle is similar to the one of the input circuit explained with the Pic. 3.1-b. To refresh your memory: we have been considering an example where four signals of equal amplitudes but different frequencies, were present in the antenna: fs1, fs2, fs3 and fs4. They were causing four different currents to flow through the LC oscillatory circuit: Is1, Is2, Is3, and Is4. All these currents were creating some voltage on the ends of the LC circuit, but the one caused by the current Is2 was significantly (about 20 times) bigger, due to the oscillatory circuit being set to its frequency. The parallel oscillatory circuit that is on Pic.3.24 as the collector .....is exactly the same as the the one on pic.3.1.

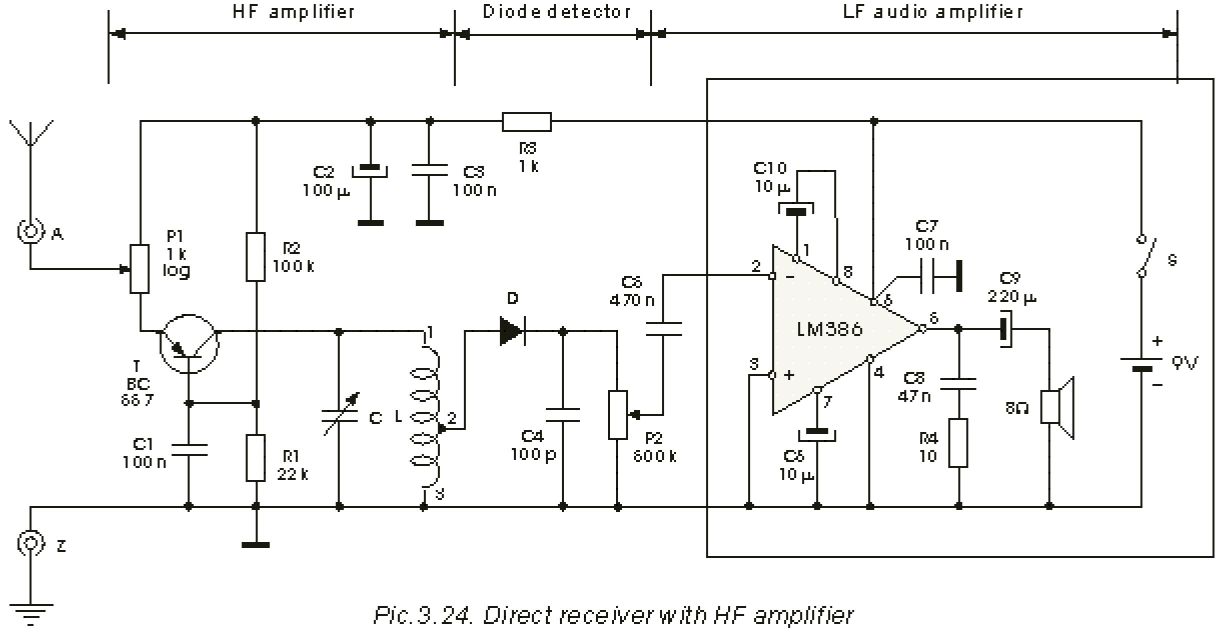

Pic.3.24. Direct receiver with HF amplifier

It also has the same role, therefore HF amplifier from pic.3.24 having the same selectivity as the input circuit in all the receivers described so far, with addition of extra amplification. This is being accomplished in the following manner: under the simultaneous load of 4 voltages that are coming to emitter from the antenna, their frequencies being fS1, fS2, fS3 and fS4, four currents flow simultaneously through the transistor. They share the same circuit: from positive battery pole, through P1, then transistor (in direction emitter-collector), over the LC circuit to the minus battery pole. All of them therefore flow simultaneously through the LC circuit as well. The resonance frequency of this circuit is set (by C) to be equal to the frequency of one of the currents and it acts upon it as a huge resistor (200 kOhms, as on pic.3.2-b). According to Ohm's Law, this current creates voltage on the oscillatory circuit. For other 3 currents the circuit acts as a resistor with much smaller resistance (less than 20 kOhms, as on pic.3.2-b) and they create much smaller voltage in the circuitry (10x smaller, as in our example). The important difference in operation of circuits from pics. 3.1 and 3.24 is that all currents are much smaller in the latter case (because of the amplifying effect of transistor), therefore the voltages on the LC circuit being much bigger.

* With P1 potentiometer the signal amplitude from antenna to the input of HF amplifier is regulated. If, on your device, you find the slider for all stations to be in rightmost position, put a resistor instead of potentiometer, and connect the antenna with emitter.

* As with all input circuits, when connecting the capacitor C care should be taken to connect the rotor to the ground (G-point on pic.3.7-a).

* The R3 resistor comprises with C2 and C3 capacitors the LF filter which prevents the feedback (that would lead to unstable operation) between the LF circuitry and HF amplifier. If the feedback still occurs, the R3 resistance should be increased.

* In LF part of the receiver the audio amplifier with LM386 IC is used. That is by no means necessary, any audio receiver will do.

* There is also a better variation of HF amplifier, with increased selectivity. Its electronic diagram is shown on pic.3.29-b.

3.8. The Audion - Direct Receiver with Drain Detector

During the experiments with this receiver, the Author had decided to name this chapter "The BEST Direct (TRF) Receiver", which he gave up on later, having in mind the old Latin saying: DE GUSTIBUS NON DISPUTANDUM EST (Tastes should not be discussed). It is, however, very hard to make something better with so little components. Anyway, the Author leaves to the readers of this chapter to name their own Best Receiver candidates, picking one of these described in this book, having in mind their own criteria for concepts of the beautiful, simple, cheap and useful. You can mail your voices to me on the address: ETŠ "Nikola Tesla" (Praktièna ELEKTRONIKA), Narodnog Fronta 31, Beograd, or by E-mail: tesla@drenik.net, Subject: Pe9.
On Pic.3.25 you can see the electrical diagram of this, however, anyway, nevertheless... Device. The signal of the tuned-in station is lead to the Gate of the BF256 transistor. Please notify that the signal is being taken from the upper end of the coil, and not from its leg, as it has been in previous projects. This is possible due to big input resistance of the FET (bigger than MegaOhm), compared to the one of the bipolar transistor (couple of kiloOhms). Why is this so important? Pic.3.26 shows the voltage - frequency curve of the parallel oscillatory circuit that is made of the coil L and capacitor C, while being tuned to the station whose frequency is fs2. In case the circuit is not loaded (the next stage of the device is not connected to it), this curve is shown in solid line and, as previously explained, the voltage

(measured at the ends of the osc. circuit) of the station with carrier frequency fs2 is significantly bigger than the voltages of the stations with frequencies fs1 and fs3, although all of them have the same size in the antenna. However, when the next stage, containing the bipolar transistor as, for example, the one on the Pic.3.12, is connected, is small input resistance is damping the circuit and the bandpass curve has

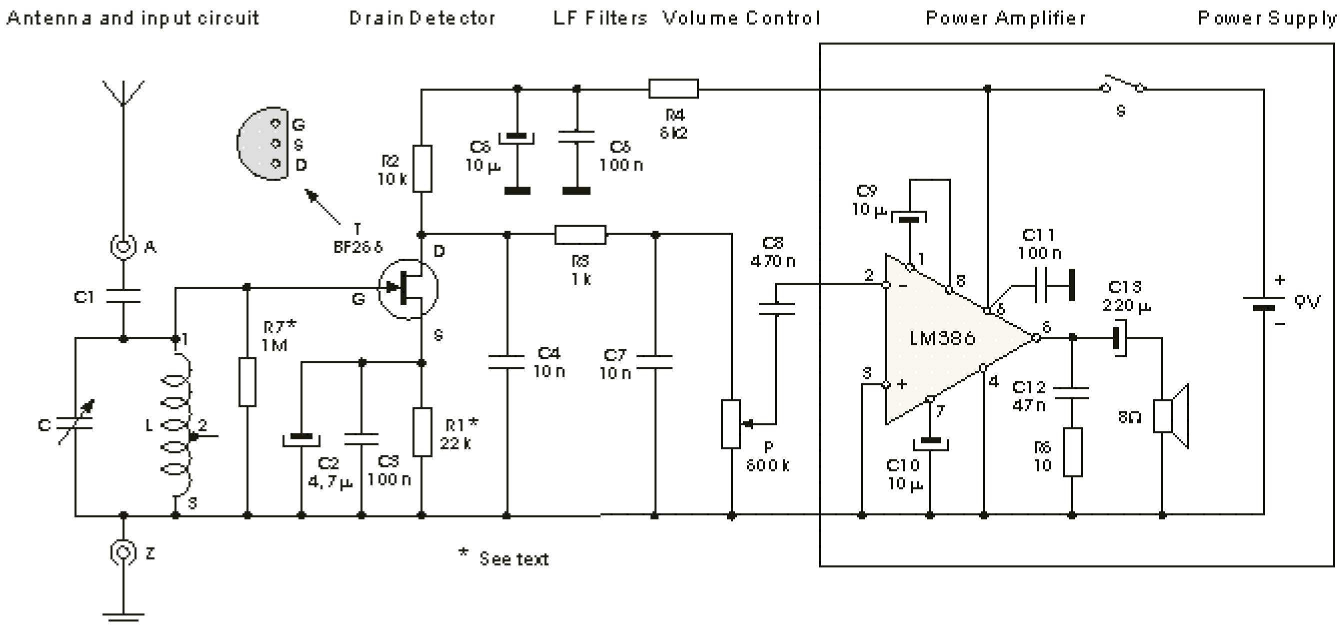
Antenna and input circuit
Drain Detector
LF Filters
Volume Control
Power Amplifier
Power Supply
A
C1
C
L
1
2
3
Z
G
S
D
T
BF256
R7*
1M
R2
10 k
C6
10 µ
C5
100 n
R4
6k2
C2
4,7 µ
C3
100 n
R1*
22 k
R3
1 k
C4
10 n
C7
10 n
C8
470 n
P
500 k
* See text
C9
10 µ
LM386
1
2
3
4
5
6
7
8
C11
100 n
C13
220 µ
C12
47 n
R5
10
C10
10 µ
8Ω
S
+
9V
−

Pic.3.25. Audion - Direct receiver with the Drain Detector

he shape shown with the dashed line, marked as Q2.When the FET is connected to the oscillatory circuit (as on Pic.3.25) there is practically no damping, and the bandpass curve remains as shown in solid line. This is, clearly, much better, since all the other station voltages are more suppressed (reduced), comparing to the voltage of the tuned station. Considering the curve marked as Q3, we will be discussing it more in context with the Pic.3.29-a.

It has been earlier noticed that the most important characteristic of a parallel oscillatory circuit being used in a receiver is its resonance frequency

$$f_r = \frac{1}{2\pi\sqrt{LC}}$$

Its second most important feature is the Goodness Factor of the receiver, which is most often being marked with the letter Q, and is therefore also known as the Q-Factor. A loaded circuit has smaller Q-factor than the non-loaded one, as shown on Pic.3.26, consequently being Q1>Q2.For example, the goodness factor of the circuit from Pic.3.6 is Q=95. Since it is not loaded, the oscillatory circuit on Pic.3.25 has smaller (narrower) bandpass and therefore better selectivity. Additionally, since the whole signal from the circuit is led in the next stage (instead partly, when getting the signal from the coil’s leg), the receiver has got bigger sensitivity (it’s capable of receiving weaker signals).

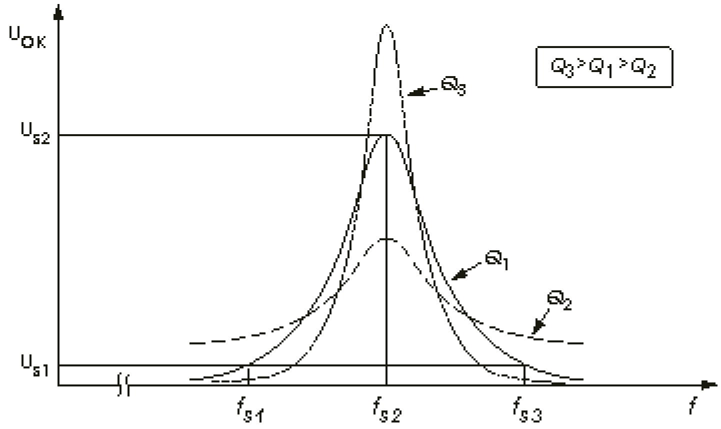

Pic. 3. 27. Bandpass curves of a parallel LC circuit

The FET, together with R1, R2, C2, C3 and C4 forms the so-called Drain Detector (its analogous circuit with bipolar transistors is the Collector Detector, and with the electronic tubes - the Anode Detector. The popular name for the anode detector was - the Audion). The LF signal being detected is received on the drain (D). It has the same shape as the LF signal obtained on the output of the diode detector, but is significantly bigger than it, since the drain detector also amplifies the signal. The LF signal is then led on the volume regulation potentiometer, over the filter that is used to suppress the remains of the HF signal carrier (R3 and C7). After that it goes to the audio - receiver.

* The R7 resistor can be omitted. If you do that you should switch off the battery (over the switch S) every time you are removing the coil from the circuitry (during experiments). If you fail to do so, the oscillation in the circuitry will occur, and loud hum will be heard from the loudspeaker.

* Since FETs have very different characteristics compared to each other, it might be necessary to change the value of R1 resistor. The simplest way to do it is to place a 50 kOhm variable resistor instead, tune the receiver to some radio station and then achieve the best possible reception by moving the slider. The resistor is then removed from the circuit, its resistance measured, and the (fixed value) resistor of similar resistivity soldered in the circuit. The same goes for R2.

* The filter (R3 and C7) used to suppress the remains of the signal carrier affects the colour of the tone of the LF signal. If you wish more bass tones you should increase the C7 capacitance. Similarly, if you wish more high pitch tones, C7 should then be decreased.

* The receiver will not start operating the very same moment the switch S is engaged. That is due to a fact that the FET doesn’t work under small supply voltages. Its supply voltage is the one on the C5 capacitor and the detector won’t work until C5 doesn’t fill. This is being achieved through R4. Since this resistance is quite big and so is the capacitance of C5, the filling time is rather long. If, however, you just need the receiver that will have a “late

start”, you should be increasing the capacitance and the resistance until reaching the desired delay time.

* This receiver works well also with the ferrite antenna. On Pic.3.27-a you can see the symbol for it, and on Pic.3.27-b its shape and dimensions are given. The simplest thing to do isusing the antenna from an old pocket radio, probably the same one you took the variable capacitor from. During the dismount, you should by no means cut the coil ends or shorten them later. Instead, you should carefully un-solder and unhook them from the PCB (the coil is made of the “litz wire”, consisting of a dozen very thin lacquer - isolated copper wires, wrapped together with the thread. If you cut this cable, you will find very hard to re-solder it, since it is difficult to remove the lacquer from all the wires without damaging (some of) them). Such an antenna, as seen on picture, has four ends. We shall be using the coil L, therefore only significant ends for us are those marked as 1 and 2. The end No.1 is easy to identify, it is the single one, which is not the case with the end No.2. To detect it in the group of three you will need an ohm-meter of some other conductivity tester, which you should connect to the end No.1, and then search for No.2 by touching those 3 remaining ends.

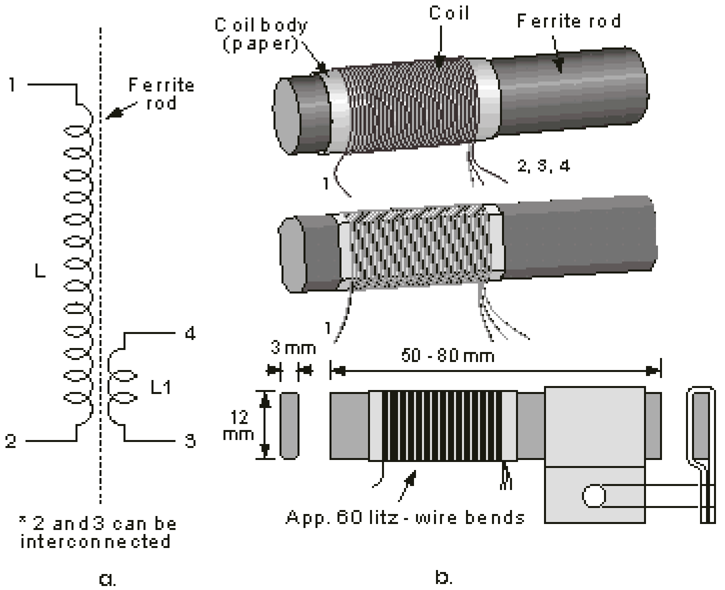

Pic.3.27. The Ferrite Antenna: a - symbol, b- appearance

In the lower right corner of Pic.3.27-b you can see means of attachment of such an antenna to the PCB. The envelope being shown is made of carton. It is wrapped around the ferrite core, and then glued to it. The screw is mounted through the hole and the antenna is then attached to the PCB. Instead of screw and nut, the antenna can also be glued onto the board. If the ferrite rod is longer than 6 cm, the envelopes should be mounted on both rod ends.

* C6 is the block-type capacitor that (together with R4) prevents the receiver from working in an unstable regime. If such a thing occurs, its capacitance should be increased. If that doesn’t help either, or if the detector is working improperly, you should try increasing R1

and R4.

* With this receiver the reception of the SW band stations can also be achieved. All that is to be done is making a new coil. For these purposes we utilized a piece of carton cylinder already used for building our coil L (described in previous projects). On it, we bended tightly 6 bends of 0.6 mm copper wire. The wire diameter isn’t critical, practically any can be used. This coil is shown on Pic.3.28-a, together with attachment plates. This coil should replace the coil L, as seen on Pic.3.25. With 6 m long antenna and the antenna capacitor C1=12 pF, the reception bandwidth should be app. from fd=7 MHz till fg=10 MHz. This can be changed by changing the number of bends on the coil and /or C1 capacitance. I such a way you can “take a peek” what’s happening in the civil area, what are the radio - amateurs doing, some professional links etc. You can even make a multiple - legged coil, such as on Pic.3.28-b (number of bends isn’t critical, it may be useful trying out some other values, too), and to be able to choose SW1, SW2, SW3 with a selector switch. Please do have in mind that the reception quality of the SW stations isn’t the same during the day. It is good in the afternoon hours, during the night and in the morning, the weakest reception quality is around noon. But, that isn’t all. It also depends on the season, solar activity etc. Anyway, you should see it for yourself.

There’s also a possibility of receiving professional stations working on considerably higher frequencies. The coil, which is then also an antenna, is given on Pic.3.28-c. It is made of stronger, thicker wire or an metal band, being circularly shaped and then attached to the wooden plate with two screws. With same screws cables connecting the antenna with the variable capacitor C are affixed. The antenna diameter varies from couple of cm till few dozens cm, the real value being found by experiments. It is a directive antenna, which means that the amount of voltage being induced in it depends also upon the direction where the waves are coming from. This gives you the opportunity of achieving the optimum reception of the desired station and simultaneous suppression of others by rotating the antenna. Similar antennas are used in radio - location (searching for whereabouts of an unknown radio transmitter).

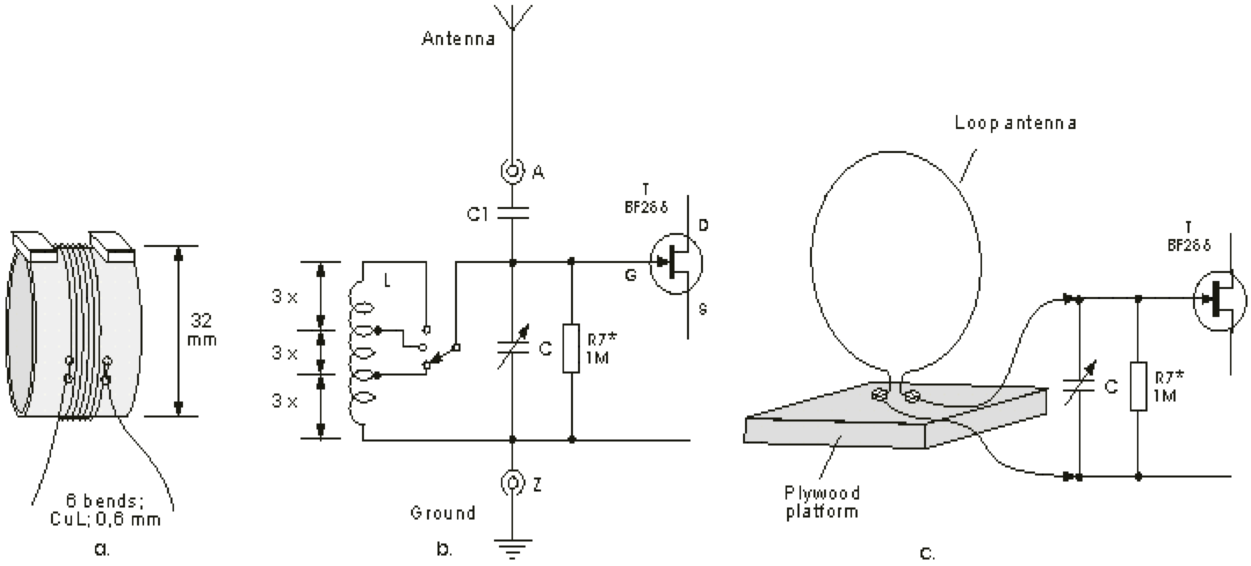

Pic.3.28 a - SW band coil; b - Input circuit with 3 SW bands; c - Coil for VHF & UHF bands

3.9-a Reaction - type Receiver

In the previous project, in the context of Pic.3.26, we have seen how important it its for an oscillatory circuit of the receiver to have as big Q- factor as possible. The bigger it is, the receiver gets more selective, i.e. capable to separate the signal of the station the oscillatory circuit is tuned at from the multitude of signals in the reception antenna, at the same time suppressing other signals. There is yet another important thing to be emphasized: the voltage in the oscillatory circuit is Q times bigger than the voltage that is being led to it from the antenna. If Q=95 and the voltage in the antenna is, say, 1 mV, the voltage exiting the circuit is then 95 mV. It is then clear that the Q- factor should be as big as possible. Some improvements can be made by using the silver - coated wire instead of the plain copper one, of using ceramic material for the coil body instead of carton etc. These improvements, however, are not significant. The solution for this was, eventually, found: It was the process known as the "regeneration", many scientists had been working on it simultaneously, A. Meissner being one of the more important ones. The complete process was patented in 1913 by the American radio - pioneer Edwin H. Armstrong, two months before his 23rd. Birthday. He discovered that the Q- factor could be extremely increased (even by couple of tens of times), if the signal from the inlet circuit would be amplified with the triode and then the small part of this amplified voltage adequately returned into the inlet circuit. The radio station signal could therefore be amplified to much greater extent, and the reception of very far stations became possible, the thing that wasn't even thought of, until then. From an average listener's point of view, the main mishap of this regenerative, or, here better known as reactive, receiver was the fact that optimal tuning required somewhat more skill and basic knowledge of the working principles. Much easier tuning was that of the superheterodyne receiver (invented also by Armstrong), which was reduced to turning only one reel, and it contributed, to a great extent, to its final victory. The battle between the two concepts lasted for almost half a century, but the reactive receiver did not retreat itself completely. Even nowadays an electrical diagram of some younger and more modern brother, or rather a grand-grandson of once famous reactive receiver, appears in some popular electronics magazine. One such diagram, where the triode is replaced by a MOSFET, is given on Pic.3.29-a.

Radio engineering enthusiasts know that, as soon as the signal from the input circuit is led on the gate (G1) and that the source (S) isn't connected to the ground but to the coil leg (point No.3), they have a diagram of the Hartley's oscillator in front of them and that connecting the source to the leg No.3, the so-called positive feedback (positive reaction)is achieved. But, the abovementioned labourers also know that the oscillator shall oscillate only if the leg No.3 is made on the right spot, and the P1, TP1 and TP2 variable resistors are set on the right values. Let us imagine that everything is OK: the coil leg is right where it should be, the slider of the variable resistor P1 is in the middle of the range, and trimmers TP1 and TP2 are set in such a manner that the oscillator is really oscillating. It behaves then as the generator, that creates the sinusoidally - shaped AC voltage, whose amplitude is a couple of volts. Theoretically speaking, the Q- factor is now infinitely big . All this shall produce a very strong whistling sound in the loudspeaker. The slider of P1 should now be carefully moved downwards. The transistor amplification is hereby reduced and the oscillating stops. The Q-factor is being reduced too, it is no longer infinite but is still very big. The receiver is now tuned on the desired station. If the station signal is weak, everything should be OK and the programme should be heard. The reception can then be made better, by carefully operating with P1's slider. If the whistling emerges again, the slider should be moved backwards until it stops. If we come upon some stronger station the whistling will start immediately, in which case the slider P2 should be carefully moved downwards until it stops and the station programme is heard, loud and clear. As one may notice, every time this receiver is being attuned to optimal receipt the whistling is being heard for a short while. This is why it has been named "The Whistler", here in Yugoslavia .

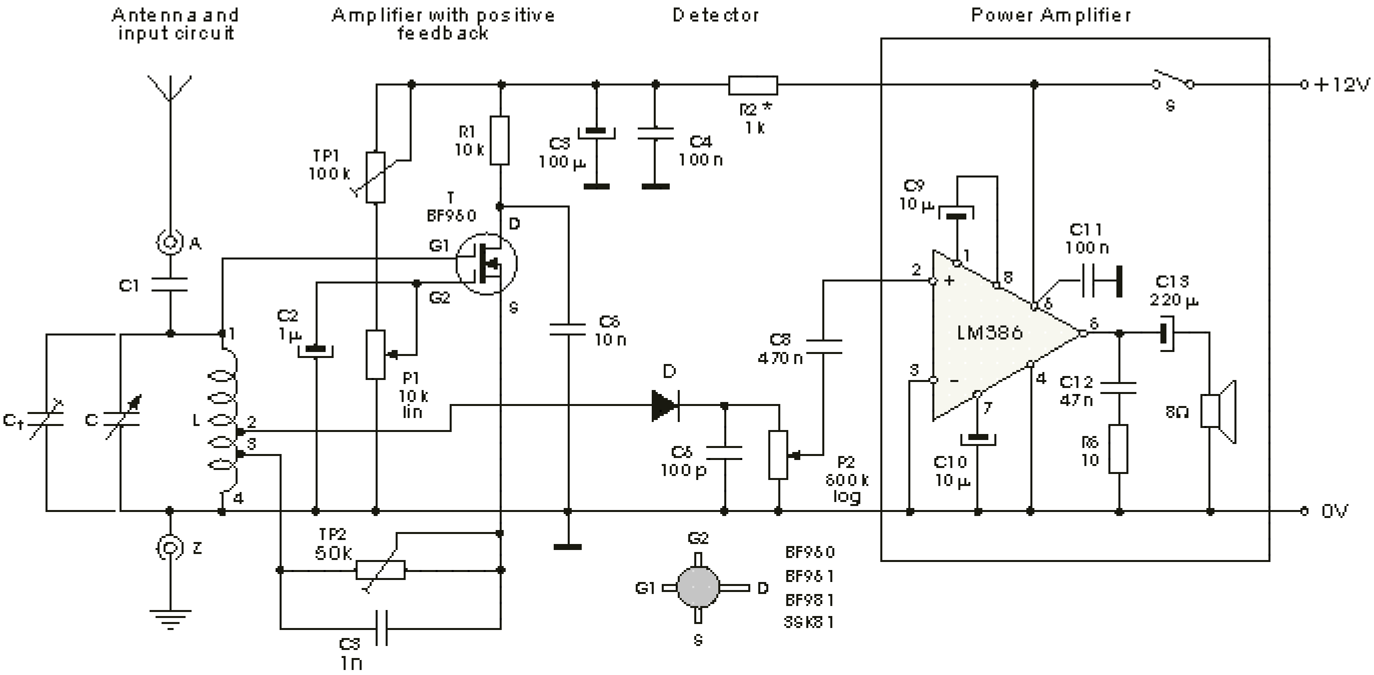

Pic. 3.29-a. Reactive Receiver with double - gate MOSFET.

The amount of the reaction (feedback, regeneration) is controlled with P1 potentiometer, which sets the magnitude of DC voltage on the gate G2 of the MOSFET, changing therewith the amount of its amplification. The range of this control is being determined with TP1 trimmer, G2 is connected to the AC signal’s ground over C2 simultaneously eliminating the noise coming from the potentiometer, and the FET is de-coupled from the supply line (and therefore all other stages of the device by the LF filter made of C3, C4 and R2.

The receiver is being tuned as follows: Put the slider of P2 in mid position and later, after tuning, you can set the volume as you wish. Set the P2 at minimum resistance (slider full down), and P2 on maximum (slider full right), connect the antenna and close the switch S. Start moving the slider P1 upwards, the reaction gets stronger and stronger, and you can hear the typical hiss or some radio programme from the loudspeaker. Move the variable capacitor C and tune the receiver to various stations. If the whistling starts, put the slider P1 back down.

Set the capacitor is minimum capacitance position (see Pic.3.7), put the slider P1 fully upwards and start carefully increasing the resistance TP2 until the whistling stops. Measure the TP2 and insert in the device the ordinary resistor of similar resistance. The TP1 trimmer should be set in such a way to have as big resistance as possible, keeping at the same time the reaction effective throughout the entire reception bandwidth of the receiver.

* During every station change (with C), a maximum amount of reaction should be set (with P1). Move the slider upwards until the oscillating occurs, then put it back down a little.

* While receiving very strong signal (local transmitter), an overload can occur. If that happens, you should insert a 1 MOhm potentiometer between the antenna (A) and the upper end of C1 capacitor; it should be connected as the rheostat (like TP1 and TP2), then you can set the optimum reception with slider.

* SW - band stations can also be received with this device, with a different coil. In this case it would be very useful to add a trimmer capacitor in parallel to the variable capacitor, being marked as Ct on Pic.3.29-a. With it the so-called “range yielding” can be done (the initial, approximate setting is done with C, and fine tuning between closely placed stations with Ct). It should be mounted on the front panel, as close to C as possible. Another type of capacitor can be used as Ct, see more about it in the Appendix. The SSB (Single Side Band) technique transmissions are also being placed in the SW band area. These signals cannot be received with the earlier described receivers, but they can with this one. In that case the slider P1 should be moved a bit more upwards, so that oscillating can occur. The reception becomes clear, before that it was unrecognizable.

* If the local radio station still corrupts the reception of other stations, you should insert the circuit that will suppress its signal. You can read more about it also in the Appendix.

3.9-b Direct SW Receiver for AM, AM-SSB & CW Signals

SSB stands for Single Side Band, which signifies the amplitude - modulated signal which gets its signal carrier and one sideband suppressed in the transmitter. Carrier suppressing gives huge savings in transmission power (the power necessary to accomplish the desired reach of the signal is significantly smaller than in the conventional - type transmitters), and cancellation of one sideband makes the signal have its spectrum two times narrower, allowing twice as many transmitters as usual to be placed into the same bandwidth.
CW is for Continuous Wave, which determines the radio link where the Morse Code is being transmitted by cutting the oscillator work in the transmitter.
SSB and CW signals are impossible to accomplish with the receivers that use the ordinary diode - type detector (earlier described AM receivers). The receipt can be done only by bringing another signal into the detector, the HF signal from the oscillator, known as the BFO (Beat Frequency Oscillator). Simpler solutions, however, do exist. These are the reaction - type receivers, i.e. receivers with positive feedback.

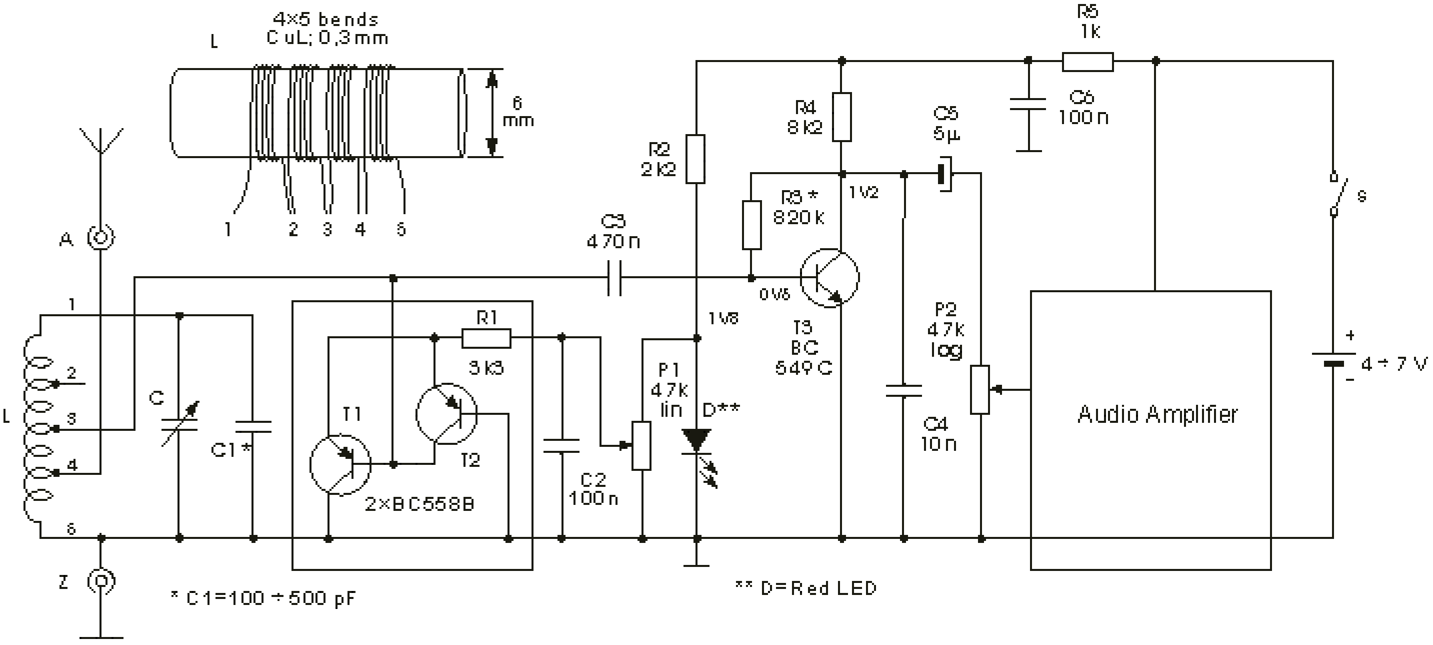

Pic. 3.29-b. AM, AM-SSB and CW Reactive SW - band Receiver

You have been able to see one of them in the previous project (3.29-a), and here we'll take a look at another one, which works so nice that we were sometimes having the impression it beats up much more sophisticated, modern supereterodyne receivers. Its electrical diagram is shown on Pic.3.29-b.
The coil L and capacitors C and C1 form a parallel oscillatory circuit whose role is to separate and amplify the signal of the tuned station, and to suppress all others. It doesn't entirely succeed in that, however. The reason for this is small Q- factor of the oscillatory circuit, being such because of big energy losses in the circuitry. There are many kinds (reasons) of these losses, but we can imagine in first approximation that there's a resistor RG in the circuit which represents these losses, its resistance being such that the oscillating current transforms itself into heat dissipation energy on it, its amount being the sum of all the (actual) losses in the circuit. We could, furthermore, solve the problem of these losses if connecting serially to RG a resistor RG', whose resistance would be negative and equal to the value of RG by its absolute value. The overall resistance would then be zero, there would be no energy losses and the Q- factor would become infinite. The oscillatory circuit would, together with the components that create this negative resistance, become an oscillator capable of receiving SSB and CW

signals.

We don’t really need an infinite Q- factor while receiving usual (conventional) AM RG by its absolute value. The resistances would not cancel each other completely, but the losses would be made very small, the Q- factor therefore becoming very big therefor increasing both the selectivity and sensitivity of the oscillatory circuit.

Transistors T1 and T2 constitute, together with resistor R3, a two-stage amplifier with strong positive feedback that has a negative dynamic input resistance. This negative resistance is connected between the leg No.3 on the coil and the ground, therefore superimposing itself with the resistance representing losses of the circuit. The quantity of this negative resistance depends on the amount of the DC current flowing through the transistors, which is being regulated by altering the DC voltage on the right end of the R3 resistor (by moving the slider of the P1 potentiometer).

The red LED D and the resistor R2 comprise a simple voltage stabilizer, obtaining 1.8 V of stabilized voltage on the P1. That means that the voltage on the right end of R3 shifts between 0 and 1.8 V while moving the slider of P1. The current flowing through the transistors thereat also changes, causing the voltage on the left end of R3 to vary between 0 and 0.6 V.

The signal of the station is being led from the leg No.3 of the coil into the collector-type detector made of T3, R3, R4 and C4. That is an AM signal detector that performs both signal detection and its amplification. Its name is the Audion. The LF signal is then, from the collector of T3, over the coupling capacitor C5, being led onto the sound volume potentiometer P2 and the audio amplifier. For the latter any of the earlier described devices can be used.

Tuning this receiver on the desired station requires both some knowledge and patience (that’s what finally “buried” this kind of receivers). Put the slider P1 in the upmost position. If strong whistling is heard that means the oscillating began. Move the slider carefully downwards until the oscillating stops. Then start slowly turning the rotor of the capacitor C until you come upon some station. If the whistling re-appears, move the slider of the potentiometer very little downwards, the whistling should stop and you should be able to hear the radio - station programme from the loudspeaker (loud and clear). For the next station tune yourself with C, then move the slider P1 upwards until the whistling appears, then put the slider back until it stops etc. All this may seem rather complex at first, but with a little practice and with two hands all will go quick and smooth.

The abovementioned method is for the signal reception of ordinary, broadcast stations. If you wish to receive the SSB and CW signals you should move the P1 slider upwards until the oscillating is achieved, so that articulate speech (SSB) or Morse code signs (CW) can be heard from the loudspeaker.

* The coil L is being made on the cylindrically - shaped body 6 mm in diameter, about 25 mm long. The plastic - made body taken from an old device is the best, like the one shown on Pic.5.14-b. The screw-shaped core allows the setup of the inductance, adjusting therewith the reception bandwidth of the device. If you cannot find such coil body, any plastic- or carton- made cylinder can be used instead. If you don’t have even that, then make yourself one. Cut the paper band to be 25 mm wide and about 150 mm long and reel it around the flat part of the 4 mm drill, adding every now and then some glue (UHU or similar). When the glue gets dry, remove the coil body off the drill.

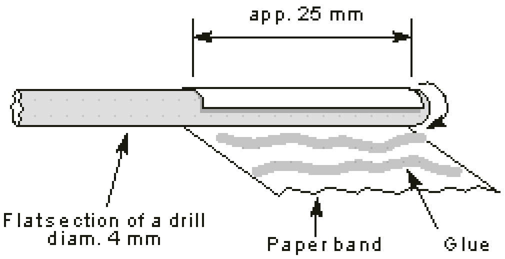

Pic.3.29-c. Spooling the coil body

The coil L has the total of 20 quirks of the lacquer - isolated copper wire, having 0.3 to 0.5 mm in diameter. A leg should be made on every fifth quirk. Latching of the wire ends (with small holes made in the coil body), as well as leg making (by making wire loops) can be done acc. to the instructions given with Pic.3.6. It can also be accomplished differently, as shown on Pic.3.29-b. First, 4 separate coils, each one made of 5 wire rings, are made side-by-side on the coil body. The starts and ends are fixed with scotch tape. The isolation is then removed from all coil ends, about 5 mm in length, after which they are tinned. On the PCB the legs are being soldered in pairs, the end of one coil with the beginning of the next (they are put together in the same hole on the PCB). For example, the end of the second and the beginning of the third coil should be connected on the same line where contact for the left end of C3 capacitor is, thus creating the leg No.3 of the coil. Putting two wires through one hole is not a very professional solution. The “real thing” are separate junctions, one for each wire, as shown on Pic.3.29-d-c.

* The feedback may happen to be not big enough, causing that there’s no oscillating even when the P1 slider is in the rightmost position. In that case, leg No.2 of the coil should be used instead of No.3. Switching between the legs can be done in many ways, the nicest (?) one given on Pic.3.29-d, made with factory-made contact pins and jumpers. On Pic.3.29-d-c you can see a detail of the PCB for the receiver shown on Pic.3.29-b. In the contacts marked as x, y and z (distance between them is 1/10 inch) the contact pins are soldered. The jumper is in position marked with dashed line, therefore making contacts x and y short-circuited. When it is moved in vertical position, the x and z contacts are in junction. In former case the coil leg No.3 is used, and in the latter it is No.2. In factory-made devices, these jumpers and contacts are used, together with appropriate connectors, to connect the PCB to the loudspeaker, power supply, variable capacitors, various switches etc.

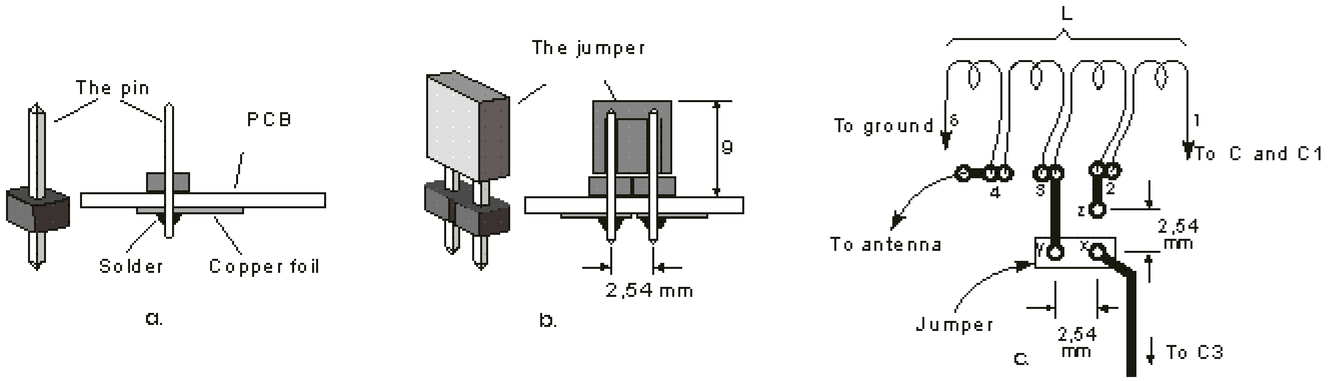

Pic.3.29-d. A- connector pin, b- pins with jumper, c- detail of the PCB from Pic.3.29-b

* Setting the collector-type detector circuit to optimum operation is done by changing the R3 resistance, until voltage on the collector of BC549C is 1.2 V.

* The antenna can be a piece of copper wire no longer than 50 cm, but with longer (few metres), external antenna, the results will be much better.

* This receiver is scheduled for the reception of SW stations from 6 MHz till 9 MHz, which is accomplished with C1 value of about 400 pF. The exact value for C1 is being determined experimentally and can be significantly different. Going down to the amateur range (about 3.75 MHz) is performed with bigger C1 capacitance.

3.10. Miniature Receiver with ZN414 (ZN414Z) IC

The receiver from Pic.3.25 can be made portable (without an external antenna) if ferrite antenna is used instead of the carton-body coil, right? Well, not exactly. The ferrite antenna is a directional - type antenna, meaning that it does matter where are the waves made by transmitter coming from. The voltage being induced in it under the effect of EM field is the biggest if the ferrite rod is in horizontal position, aimed in such a manner that its axis is perpendicular to the direction of spread of the waves (i.e. perpendicular to the direction towards the transmitter). If the rod is being turned (in horizontal plane), the voltage decreases and gets to minimum when the rod is turned for 90 degrees, i.e. When the rod axis is parallel to the direction pointing towards the transmitter. That means that when the receiver from Pic.3.25 (with the ferrite antenna) is tuned to some station, the reception can be additionally improved by simply rotating it. The directional feature of the antenna is useful here. This, however, can be bad if the receiver is used in motion, since the volume would be constantly changing, and the receiver should be constantly turned towards the transmitter. It would clearly prove itself useful to have an electronic circuit in the receiver that would be automatically changing the amplification of the receiver, in such a manner that it would be increased when the signal in the antenna weakens and vice versa. This would even be convenient in the stationary - type receivers, since it would provide the constant sound strength in the loudspeaker during the reception of signals of various power. Such circuit does exist in modern receivers, and is called the Automatic Amplification Regulation circuit (AAR).
AAR circuit could be added to some of the receivers described in previous projects. The mishap is that electrical diagrams would then become significantly more complex, which would reflect itself both on the price and the realizability. The solution, however, does exist, and is rather simple: it’s the ZN414 IC, being promoted as an “AM Receiver”. Its block diagram is given on the Pic.3.30. The signal from the input

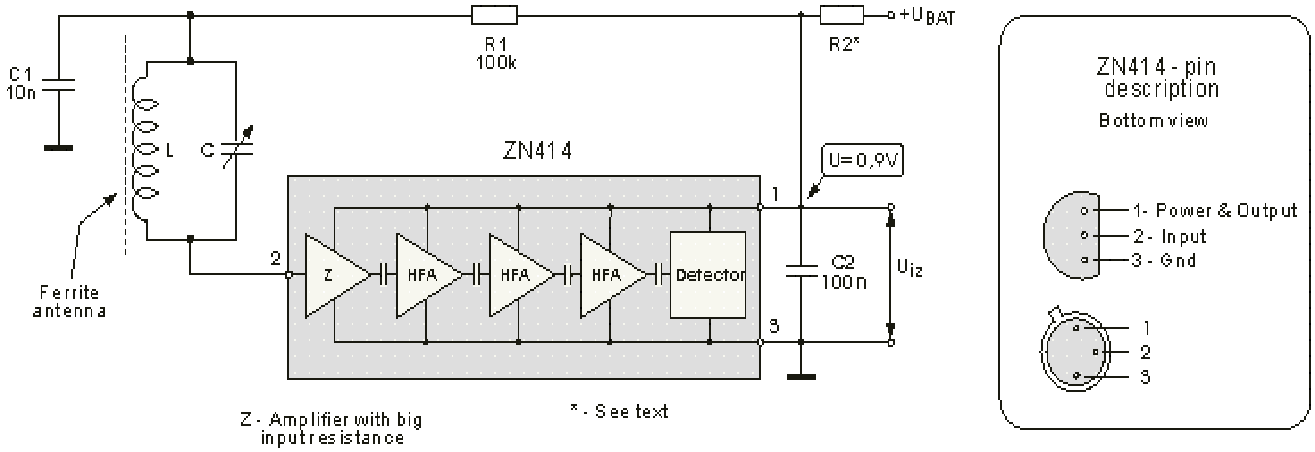

Z - Amplifier with big input resistance

* - See text

Pic.3.30. HF part of an AM radio receiver realized with the Zn414 IC

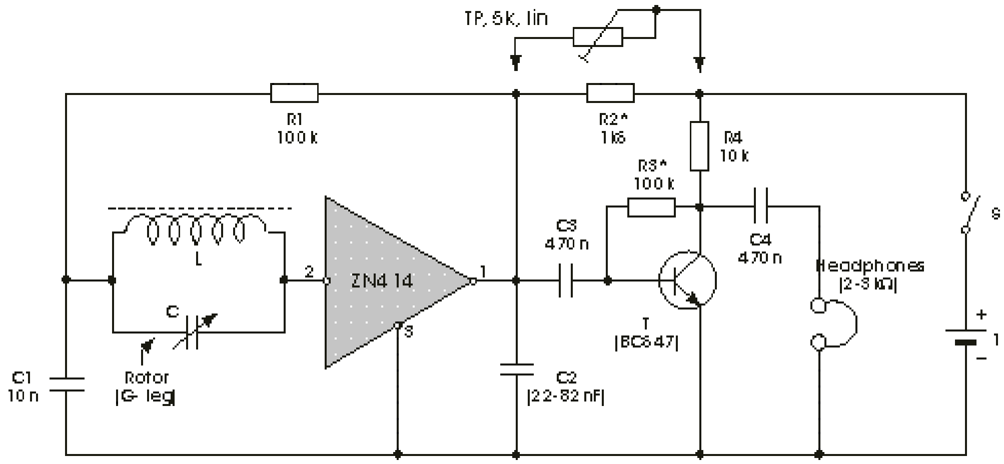

* The optimum value for the R3 resistor is being determined by the aid of the 2.2 mOhms trimmer, in the way that has been described in chapter 3.2.

Pic.3.31. AM radio receiver with ZN414 IC

circuit (with ferrite antenna) is led on pin 2 of the IC, which is the input of the amplifier (Z) that has very big input impedance (about 4 MOhms). This is very significant, since such amplifier does not load the oscillatory circuit and doesn't reduce its Q- factor. The signal is then being amplified through 3-stage HF amplifier (HFA) and demodulated in the detector, thereby exiting the IC as an LF signal (music, speech...). In the right part of the Pic.3.30 the pin description of the ZN414 is given. As you can see, it is placed in a housing usually used for small-power transistors, either plastic (on top, like BC547) or metal (bottom, like BC107).

One end of the oscillatory circuit is connected to the ground over the C1 capacitor (for AC voltages), since input into the amplifier Z is between pin 2 and the ground, where the input signal is to be brought.

The automatic amplification regulation (control) is achieved by returning the DC component of the detected signal from the output to the input of the IC, over the R1 resistor. This DC voltage is being created on R2 resistor. It is substantial for the correct operation of ZN414, its resistance must be such that DC voltage on the pin 1 (to the ground) when no station is being received, is equal to 0.9 V. While calculating the R2 one must have in mind that the idle current of the IC goes through it, its typical value being 0.3 mA, and maximum 0.5 mA (more data about ZN414 can be found in table on Pic.3.36).

The electrical diagram of a small portable MW radio receiver, reproduction being done over the headphones, is given on Pic.3.31. The LF signal is led from the ZN414 output over the coupling capacitor C3 to simple amplifier made with BC547 transistor (or similar), which we discussed about before. This can even be done without the LF amplifier. If you have high-resistance headphones that are sensitive enough connect them between the right end of C3 and the ground, and omit the transistor, R3 R4 and C4.

The 1.5V battery is being used in this device, therefore the necessary 0.6V voltage drop is done with R2=1.5 kOhms. You should, just in case, connect first a 5 kOhms trimmer instead of R2, put its slider in the mid position, turn on the receiver and set it on an empty place on the scale, where no signal can be heard. Connect the voltmeter between the pin1 and ground, and carefully move the slider until the instrument shows 0.9V. If you have no instrument, tune the receiver to some station and move the slider carefully until you reach an optimum receipt. Then turn the variable capacitor's knob across the entire scale, to ensure that receiver is working well throughout the entire reception range. If everything is OK turn off the receiver, disconnect the trimmer, measure its resistance and solder an appropriate resistor on the PCB. While experimenting with R2 please have in mind that its resistance should be in any case no less than 600 Ohms.

On Pic.3.32 the PCB, ferrite antenna and look of entire device are shown.
If you plan to make a different PCB, since the device works on high frequencies, you have to obey certain rules in order to have a reliable and stable operation:

a. The splitting capacitor C2 has to be mounted as close as possible to the pin 1 of the ZN414. Its capacitance affects both the amplification (which increases with increase of C2) and the limit frequency of the LF signal (which decreases with increase of C2), so the compromise has to be found. You may put for start C2=82 nF (or even 100 nF), and if the reproduction quality pleases you - everything is in order. You could try with smaller capacitance, the amplification will decrease but the reproduction will be better, etc.

b. All the connections, especially those near the ZN414, must be kept as short as possible.

c. The ferrite antenna and variable capacitor should be placed as far away from the battery, loudspeaker (if existing) and the cables connecting them to the PCB.

d. The rotor (G- leg) of the variable capacitor must be connected with the junction of R1 and C1.
Regarding the ferrite antenna, the best thing would be using some that is retrieved from some disused conventional receiver, more on this was told in the project No.3.8. If you can’t find one, or it is unsuitable for some reason, you can make it according to Pic.3.32-c. The length of the ferrite rod is 42 mm. If you have a longer rod, cut it down to size. This cannot be done with the saw, but a groove must be made with the rasp all around, after which the rod can be simply broken in two. The coil body is, again, made of paper tape that is spooled and glued onto the rod. Before you start with spooling, several pieces of 0.5 mm wire (3 on the picture) should be inserted between the rod and the paper. The coil has got 80 quirks of lacquer -isolated copper wire, its diameter being app. 0.2 mm. The beginning and the end of the coil are fixed with the scotch tape (the starting quirks are pressed on the coil body and fixed with several reels of 3 mm tape. The same is done with the ending ones). When the body is finished, the wires are removed. It is thus achieved that the coil body doesn’t lay firmly on the rod, which can now be moved side-to-side, changing thereby the inductance of the coil, so that its optimum value can be established. If you still decide not to use the ferrite antenna, you can use our coil from Pic.3.6. In that case, the leg should be kept “in air”, i.e. it is not used.

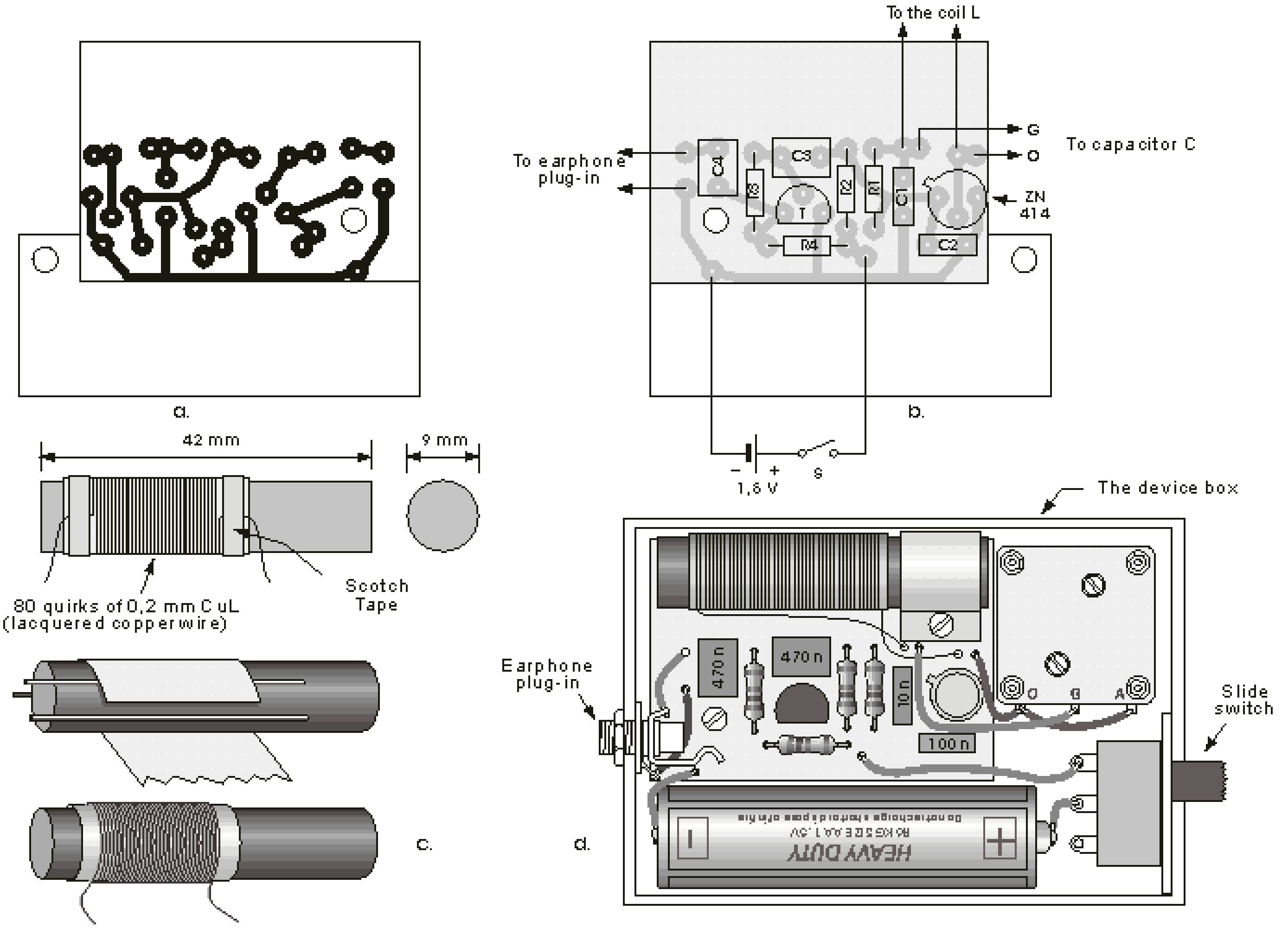
To the coil L
To earphone plug-in
G
O
To capacitor C
ZN 414
C4
C3
R3
T
R2
R1
C1
R4
C2
a.
b.
42 mm
9 mm
−
+
1,5 V
S
80 quirks of 0,2 mm CuL (lacquered copperwire)
Scotch Tape
The device box
Earphone plug-in
470 n
470 n
10 n
100 n
O
G
A
Slide switch
HEAVY DUTY
c.
d.

Pic.3.32. Practical implementation of the Radio Receiver from Pic.3.31-a: a- PCB (soldering side), b- PCB (component side), c- ferrite antenna making, d- complete device

* The battery can be connected to the PCB with two pieces of wire that are soldered to it. This solution is fine if you are skilled in soldering and can easily un-solder the old battery and attach the new one. But if you intend to give the receiver to someone, and he/she is not a soldering-lover, you'll have to find another solution. The simplest thing to do is take the battery housing from an old receiver, do the necessary adjustments and connect it with the PCB with two pieces of flexible (litz) wire. If you cannot do the former, make two battery platforms of brass, as shown on Pic.3.33 and solder them on two copper areas on PCB that are big enough to support them. If you accept this solution, your PCB must be bigger (The additional part is shown in dashed line, on Pic.3.32-a). The board now also contains the holes for the screws, which are fixing it onto the device box. On the platform that supports negative (-) battery pole, a small spring can be attached, to provide a good contact. If you don't have such a spring, bend the platforms inwards a little, to keep the battery firmly in place.
If you are using a power source whose voltage is greater than 1.5 V, the R2 resistance should be increased. The exact value for it is best to find as previously described, by using the 50 kOhms trimmer. Even better solution is using one of the circuits from the Pic.3.34. Which one should it be? The one on the Pic.3.34-d gives the best operating performance.

The setting is done with the TP trimmer. The slider is put in the lowest position, and then is slowly moved upwards until the voltage on the pin 1 doesn’t reach the level required. However, this circuit applies a big load onto the battery, surging from it the current I=(9V-3V)/680Ohm=8.8 mA.

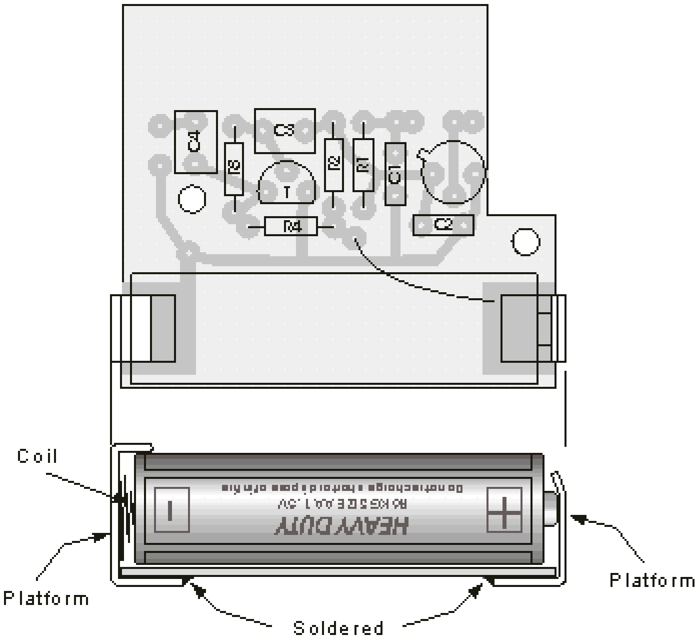
C4
C3
R3
T
R2
R1
C1
R4
C2
Coil
HEAVY DUTY
Platform
Platform
Soldered

Pic.3.33. Battery mounting on the PCB

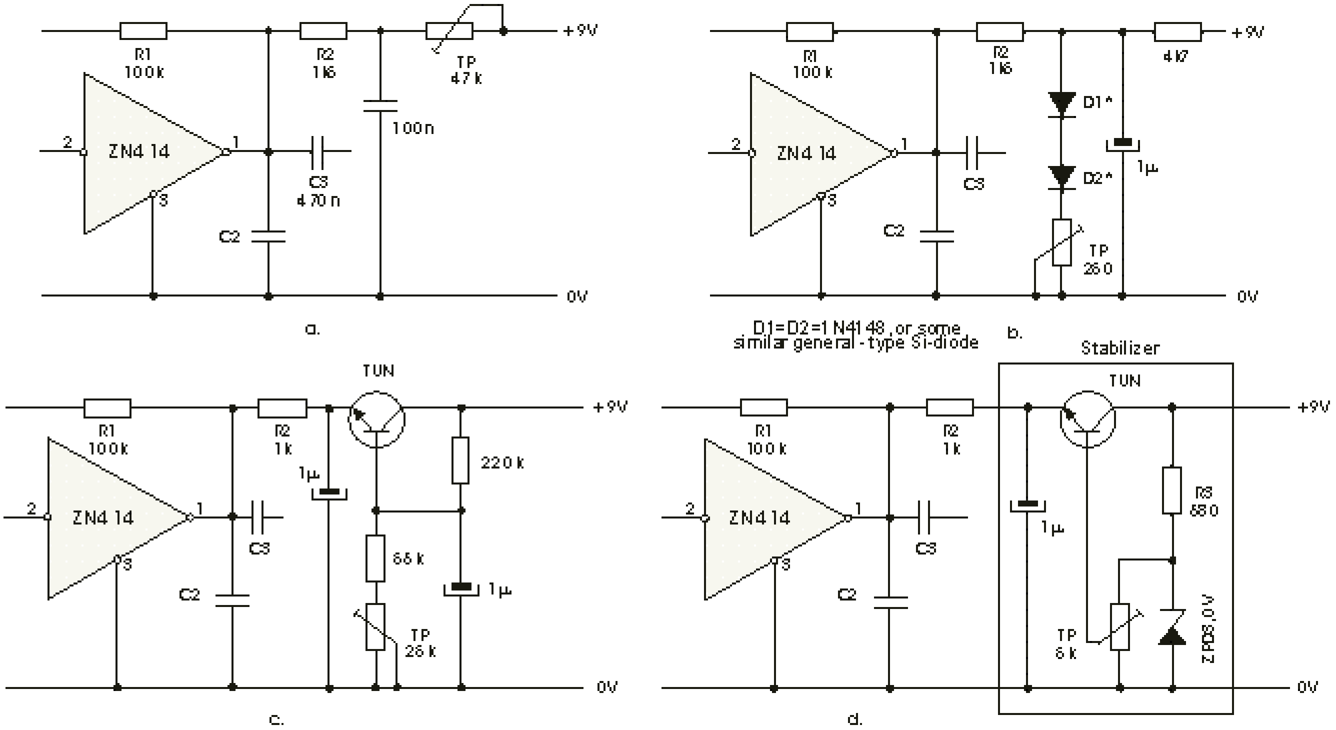
R1
100k
R2
1k6
TP
47k
+9V
2
ZN4 14
1
3
C3
470n
100n
C2
0V
a.
R1
100k
R2
1k6
4k7
D1*
D2*
1µ
C3
C2
TP
250
+9V
0V
D1=D2=1N4148, or some similar general-type Si-diode
b.
TUN
R1
100k
R2
1k
1µ
220k
68k
1µ
TP
25k
C3
C2
+9V
0V
c.
Stabilizer
TUN
R1
100k
R2
1k
1µ
R3
680
TP
5k
ZPD3,0V
C3
C2
+9V
0V
d.

Pic.3.34. Power supplying the ZN414 from the 9V battery: a- with the resistor divider, b- with a two-diode stabilizer, c- with the transistor voltage attenuator, d- with the stabilizer with a transistor and the Zener - diode

3.11. Pocket Receiver with ZN414 & LM386 IC's

The author was taking notes in his lab while testing the receivers being described herein, and used them later to write this book. One remark about the receiver whose electrical diagram is given on Pic.3.35 was: "Works EXCELLENT". The reader will probably ask himself: Isn't this the aforementioned "The Best Receiver"? No, it isn't, just remember: DE GUSTIBUS... If, however, you consider it to be "The One", please send your vote on E-mail: tesla@drenik.net;

This device is very similar to the one being described in the previous project. The most important difference is that the LF signal exiting the ZN414 does not go to the transistor amplifier but to the power amplifier built around the LM386, which was used in some previous projects. Any other audio amplifier can be used instead, e.g. those on Pics. 3.15, 3.21 and 3.22.

With this receiver, special care should be taken regarding the voltage on pin No.1. As you did in previous project, put the potentiometer's slider in mid-position, turn the receiver on and tune it to some station. Move the slider carefully, until you reach the optimum reception. Start changing the capacitance of the variable capacitor, covering its entire scope, to make sure that receiver works well in its entire operating range. If a problem occurs, re-position the slider again. When everything gets OK, turn the receiver off, disconnect the potentiometer, measure its resistance, and solder the resistor of such resistance on the board. The R2 resistance must be no less than 600 Ohms.

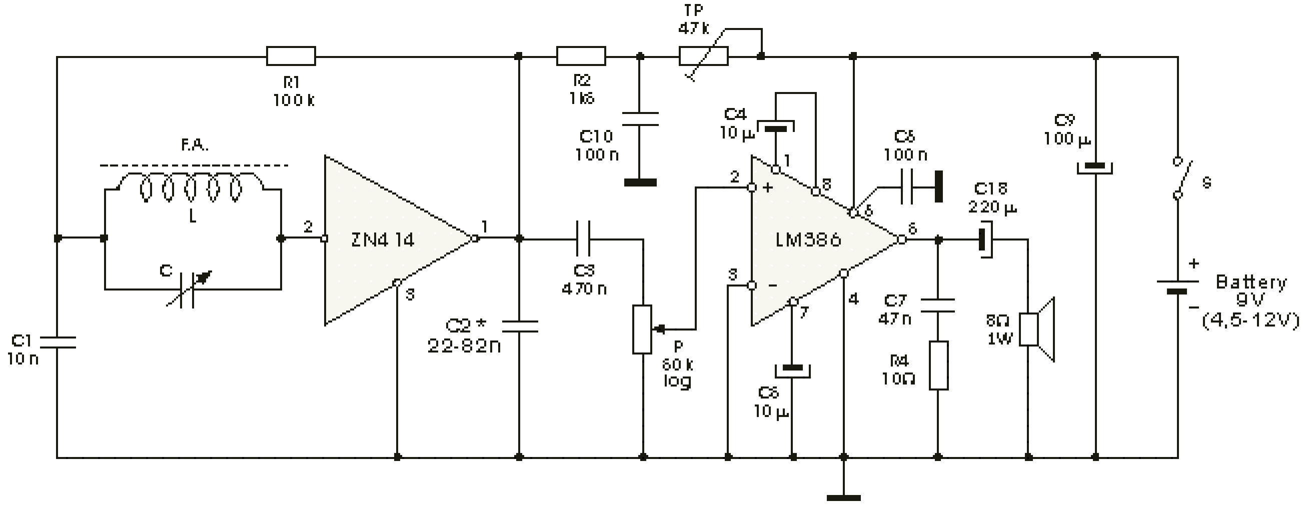

Pic.3.35. The AM Radio Receiver with ZN414 and LM386 IC's

3.12. Miniature Receiver with ZN415E (ZN416E) IC

The ZN415 IC, being promoted as “The World’s smallest radio” by its manufacturer (FERRANTI), is an improved version of ZN414. As seen on the Pic.3.36-a (IC’s block diagram), the 8-pin DIL package contains ZN414 IC with R1 and R2 resistors, an amplifier with T1 transistor and the separation stage performed with T2 transistor. All the necessary resistors are comprised in the IC. Someone will, perhaps, wonder why there are no capacitors integrated inside, so that the externally mounted parts would then be only the battery and the headphones? The answer is fairly simple: It isn’t yet possible to make the capacitors that would have enough big capacitance, in the monolithic IC’s manufacturing process. Regardless of this “problem”, the receiver built around the ZN415 is, really, extremely simple.
As described in project 3.10, the LF signal exits the ZN414. Through the coupling capacitor C3 it is led on the base of T1, and the amplified signal is obtained on its collector pin, and is taken directly on the base pin of T2. The remains of the HF signal are led to the ground over the C4 capacitor. The separation stage allows us to connect small - resistance (64 Ohms) headphones, which are mass - produced these days and are the easiest ones to obtain.
The 1.5 V battery is used for the power supply of the circuit, and is being connected between the pins 6 and 4. Having in mind the previous experience gained with the ZN414 IC, the author recommends to the readers to try improving the signal reception by adding the R2* resistor (helping themselves with linear 1-5 kOhms potentiometer), which is shown in dashed line on the picture.
In the upper end of the Pic.3.37 the table with basic data about ZN415 circuit is given. All these apply on the ZN414 as well, except for the last line. The smaller table given below it contains the data about the DC voltages on the pins of the IC. If something is wrong, the first thing to do is check these voltages and if they are (approximately) as those given in the table, the IC is OK and the error is therefore to be searched in the printings, the contacts, junctions, etc.
Everything that is said about the ferrite antenna in the previous projects applies here as well.

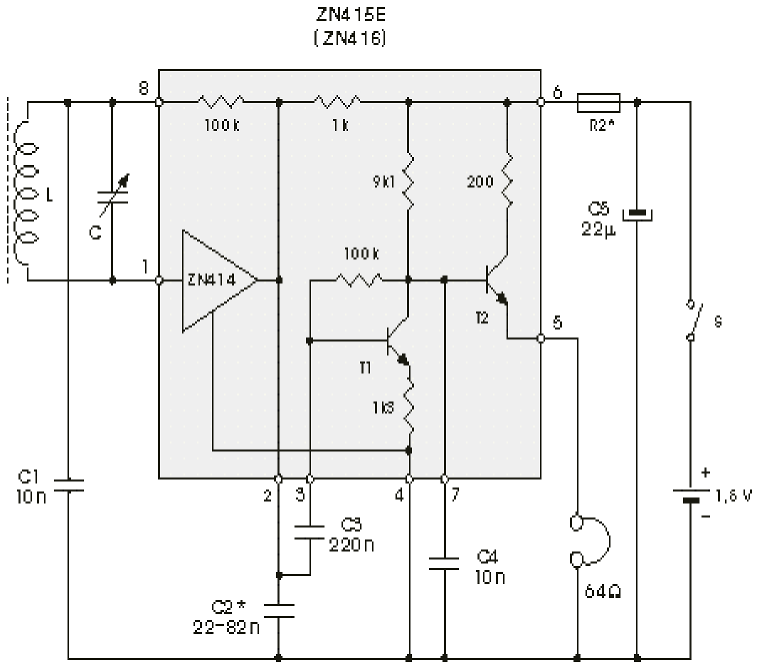

a.

Pic.3.36 Miniature AM receiver with ZN415E IC

The ZN415E (ZN414Z) features

Parameter	min.	tip.	max.	unit
Supply voltage	1,1	1,3	1,6	V
Current consumption		4 (0,3)	5 (0,5)	mA
Frequency range	150		3000	kHz
Input resistance		4		MOhm
HF power amplification		72		dB
LF voltage amplification		18		dB

* Data in the parenthesis are for ZN414Z

Pin	U(V)
1	0,75
2	1,1
3	0,7
4	0
5	0,2
6	1,5
7	0,9
8	0,75

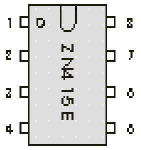

The DC voltages on ZN415E pins

b.

Pic.3.37. ZN... IC's data

3.13. AM Receiver with ZN415E & LM386 IC's

This FRT AM receiver combines the features of two magnificent IC's, ZN416E (HF amplifier, detector and the pre-amplifier) and LM386 (audio power amplifier). Add a power supply based on 7809 voltage stabilizer IC, use a loudspeaker with membrane diameter being about 15 cm, put all this in one of the boxes described in the Appendix, and you will have a new and improved silicone - based model of an old - fashioned radio receiver, built somewhat less than 100 years ago.

The electrical diagram of this device is given on Pic.3.38. The voltage reduction (and stabilization) from 9V to 1.5V, as needed by ZN415, is being done over the R2 resistor and D1 and D2 diodes, although any circuit given on Pic.3.34 can be used instead. Universal - type diodes 1N4148 are used here, but any other low-power silicon diodes will do. With 250 Ohms trimmer TP the fine - tuning of the voltage on the pin 6 is being done (the receiver is set to some station in the upper part of the reception bandwidth (app. 1500 kHz), and the slider of the TP is then used to obtain the optimum reception). After that, the receiver is switched off, the trimmer is disconnected, its resistance measured and an equivalent resistor soldered in the circuitry. On the PCB, this resistor is marked as R3.

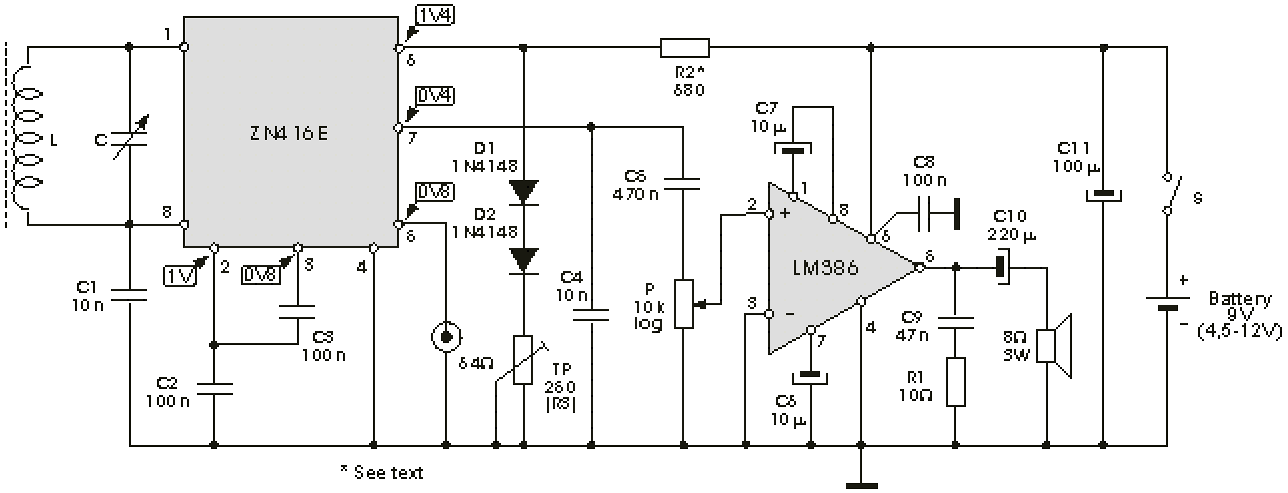
1V4
0V4
0V8
1V
0V8
1
2
3
4
6
7
5
8
ZN415E
L
C
C1
10 n
C2
100 n
C3
100 n
D1
1N4148
D2
1N4148
64Ω
TP
260
|RS|
C4
10 n
C5
470 n
P
10 k
log
R2*
680
C7
10 µ
LM386
+
-
1
2
3
4
5
6
7
8
C6
10 µ
C8
100 n
C9
47 n
R1
10Ω
C10
220 µ
8Ω
3W
C11
100 µ
S
+
-
Battery
9V
(4,5-12V)

* See text

Pic.3.38. AM radio receiver with ZN415E & LM386

The PCB, component layout and connections with other parts of the device ar shown on Pic.3.39. If the receiver is made for the household use, you should certainly put it into some bigger box where bigger loudspeaker (~15 cm) can be installed. The ferrite antenna should be mounted on the rear or side panel, as far from the loudspeaker as possible. All the conductors, especially those that connect the capacitor to the board, should be as short as possible. The wires that connect the loudspeaker and the adapter to the PCB have to be firmly twisted, as shown on picture for the loudspeaker. The potentiometer is to be connected with the shielded microphone cable, if it is more then 10 cm away from the board.

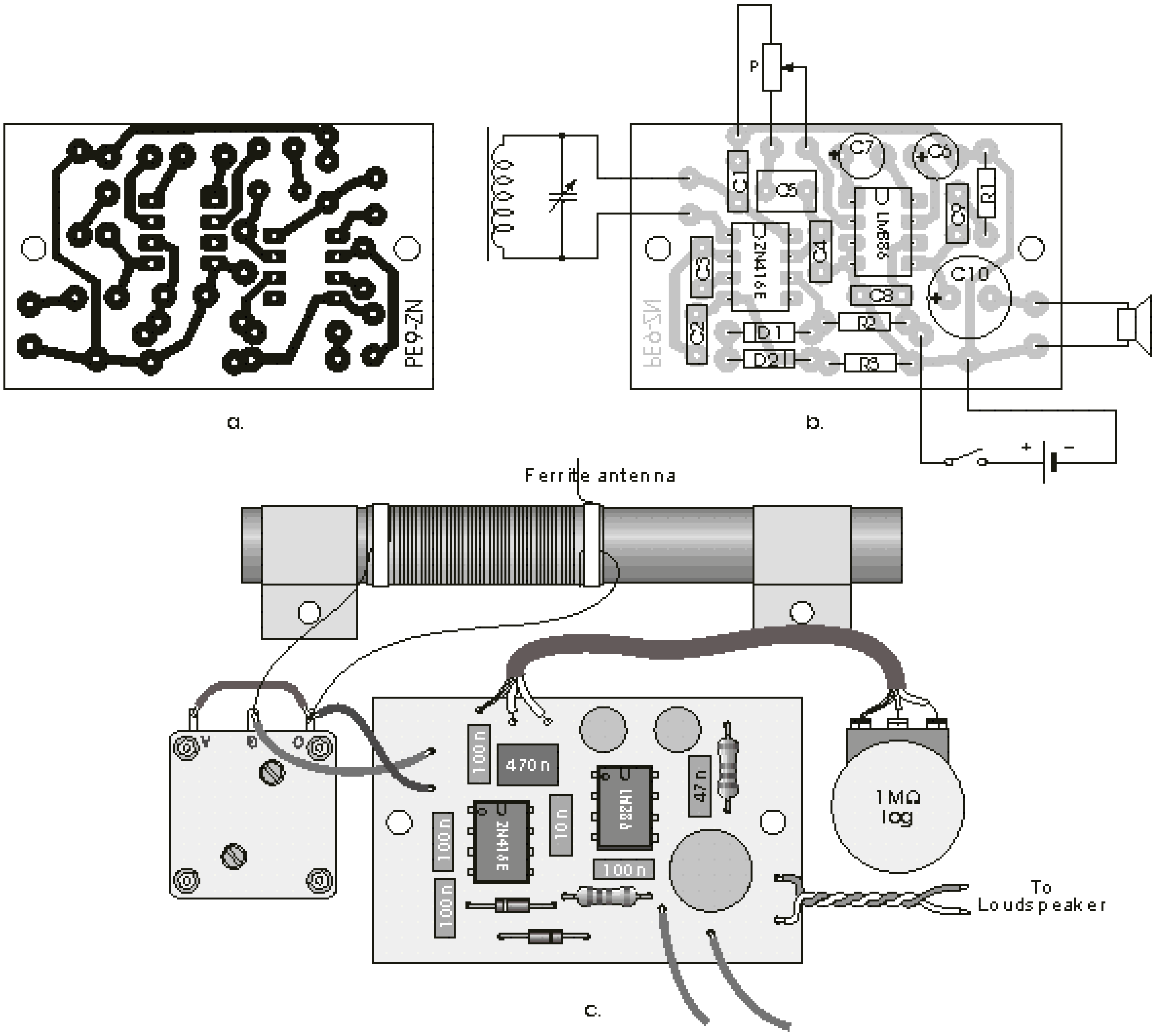
PE9-ZN
a.
P
C7
C6
C1
C5
R1
C9
LM386
2N416E
C4
C3
C10
C8
C2
D1
R2
D2
R3
PE9-ZN
b.
Ferrite antenna
100 n
470 n
47 n
10 n
LM386
2N416E
100 n
100 n
100 n
1MΩ log
To Loudspeaker
c.

Pic. 3.39. Practical implementation of the receiver from Pic. 3.38: a- PCB (solder side), b- PCB (component side), c- "Real" layout

A connector for 64-Ohm headphones can also be mounted on the box. It is then connected to the PCB via a microphone cable, connecting its "live" end directly on pin No.5, and the shield wire to the closest ground. Headphones can also be used instead of loudspeaker, in which case such connector must be chosen, that will switch off the loudspeaker when headphones are plugged in.

* Some more modern IC, other than LM386, can also be used in this device. The author recommends TDA7052A. Its main features are:

1. Extremely stable operation
2. Implemented output short-circuit protection
3. No cooler necessary
4. Small power consumption, IO=4 mA
5. No "clicking" during switching ON/OFF
6. Voltage amplification is 39 dB
7. Output power Piz=1.2 W (3x bigger than LM386)
8. The 220 mF capacitor isn't needed if the power rectifier that has an electrolytic capacitor on its output is used, and the cables connecting it with the device are short. In that case the

amplifier is practically made without the external components, if we exclude the (inevitable) 100 nF capacitor, which is in every way great advantage.

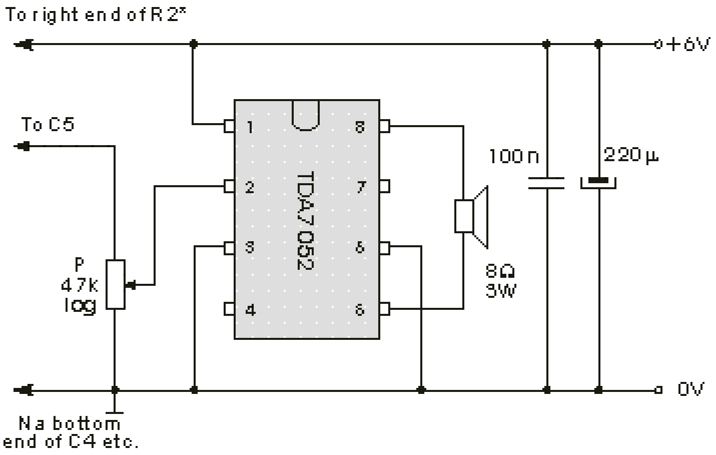

Pic.3.40. LF part of the receiver from pic.3.38.

3.14. Miniature receiver with ZN415E & TDA7052 IC's

The 9 V battery is not so popular nowadays as it used to be, probably for its quite big price/capacity ratio, so it is almost never used in the pocket-type receivers today. Much more frequently utilized these days are the cylindrically shaped 1.5 V AA type batteries. Electrical diagram of one pocket-type AM receiver that uses two of these for power supply is given on Pic.3.41. Two IC's, ZN415 and TDA7050 are being used in it. The DC supply voltage for TDA7050 must be exclusively 3 V, and should be no bigger than 4 V. The voltage being supplied from the batteries gets smaller during the use, but the IC operates properly as long as it doesn't drop below 1.6 V. Besides this nice feature, this IC has very small dimensions, very low idle current (4 mA), all the external components it requires are two capacitors (C8 and C9), which makes it ideal for usage in the mobile-type devices (radios, wakmans, etc.), the purpose that is originally designated to it by its manufacturer (Philips).
* Caution must be expressed with the power supply, since this two-batteries power supply cannot be simply replaced with some common-type 3V adaptor. If you wish to use such adaptor, the pin 6 of ZN414 must be connected, over the R2 resistor and one of the circuits from Pic.3.35, to the pin 8 of the TDA7050, and the adaptor is then to be connected between the pin 8 and the ground.
* The R2* resistor, shown in dashed line on the picture, should be added only if the receiver does not operate stable. Determining its exact value is explained in several previous projects.
* The 64 Ohm headphones are connected onto the output of the ZN415 (between pin 5 and Gnd), or instead of the loudspeaker (in which case the reproduction can be much louder).
* The PCB and components layout is shown on the Pic.3.42. Connecting other components is being done in the similar manner as in the previous project. In the left corner the batteries are shown, placed in some convenient holder taken from some old device. Between the plus pole of the left and the minus pole of the right battery the tin plate is inserted, with one of 3 cables connecting the supply to the PCB soldered to it.

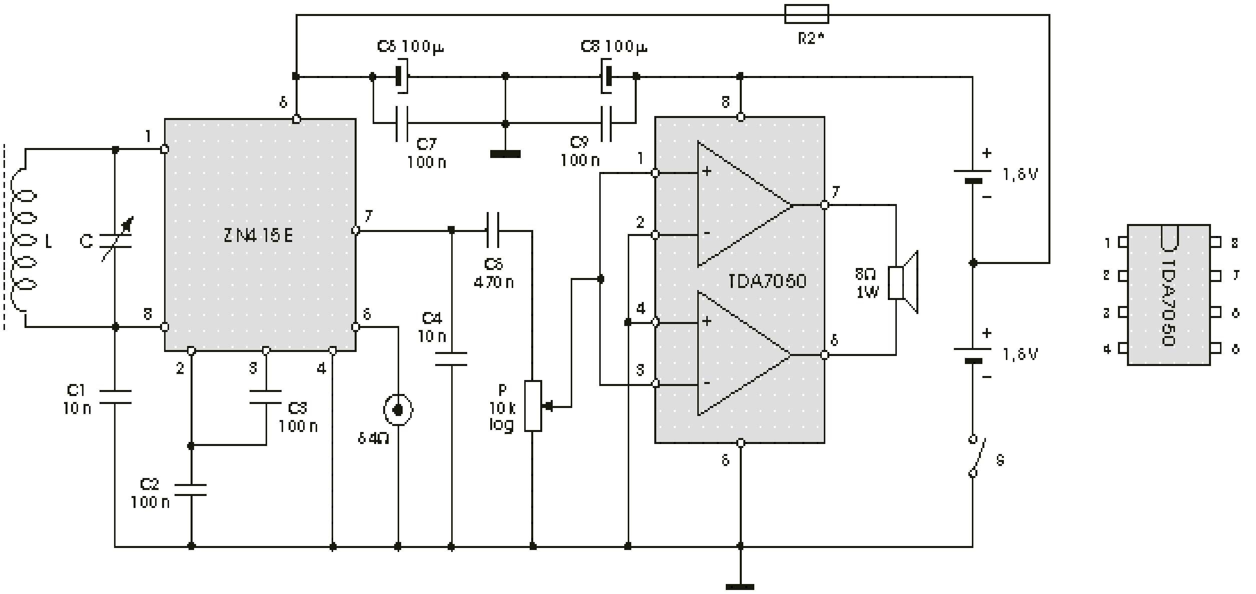

C5 100µ
C8 100µ
R2*
C7 100n
C9 100n
ZN415E
TDA7050
L
C
C1 10n
C2 100n
C3 100n
C6 470n
C4 10n
64Ω
P 10k log
8Ω 1W
1,5V
1,5V
S

Pic. 3.41. AM receiver with ZN415E i TDA7050

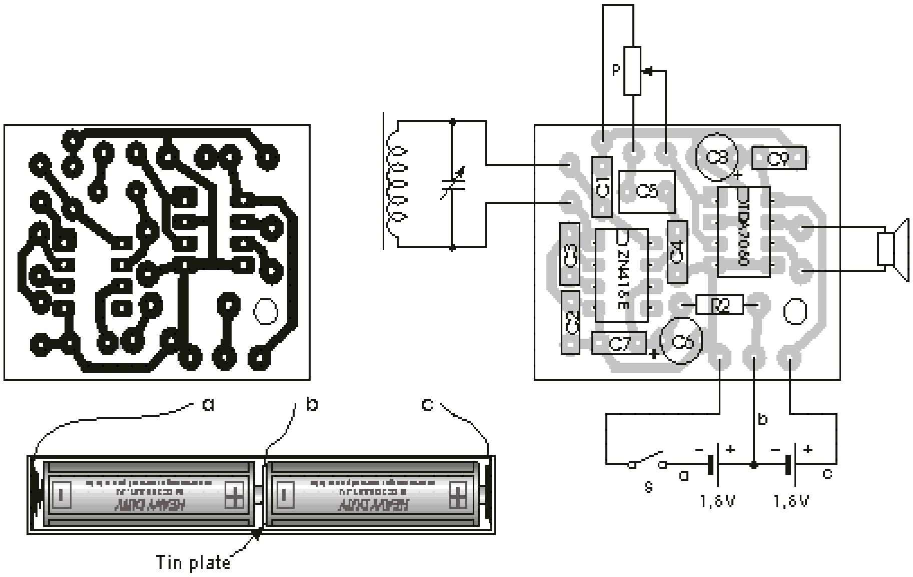

P
C1
C5
C8
C9
C3
C4
ZN415E
TDA7050
C2
R2
C7
C6
a
b
c
S
1,5V
1,5V
Tin plate

Pic. 3.42. The PCB for the device from pic. 3.41

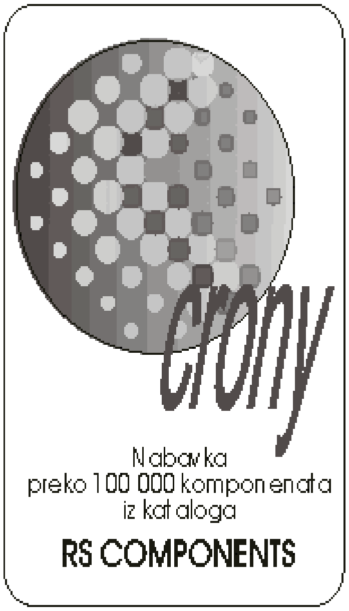

crony

Nabavka
preko 100 000 komponenata
iz kataloga

RS COMPONENTS

3.15. Direct (TRF) FM Receivers

Frequency modulation is used in radio broadcast in the bandwidth range from 88 MHz til 108 MHz. This range is being marked as “FM” on the band scales of the radio receivers, and the devices that are able to receive such signals are called the FM receivers.
Radio broadcast transmitters are using the amplitude modulation on LW, MW and SW bandwidths. According to international treaties, each of the transmitters has a 9 kHz wide broadcasting channel, therefore making maximum frequency of the information being transferred fNFmax=4.5 kHz, according to the characteristics of the AM signal. To put it

more simple, the highest frequency of the sound that can be heard from the loudspeaker of an AM receiver is 4.5 kHz, all above it will be simply truncated in the circuitry. Considering the speech itself, this isn't so important since the most important components are located below these 4.5 kHz (during the telephone transfer, all the components above 3.2 kHz are being cut, and nobody is complaining). Things stand different, however, for the transfer of music. Music has much more sound components, with their frequencies spreading up to 15 kHz, so truncating them above 4.5 kHz does deteriorate the transmission quality.

The radio-broadcast FM transmitter has a 250 kHz wide channel on its disposal, therefore allowing for the maximum frequency of the information (acc. to the characteristics of the FM signal) to be fNFmax=15 kHz. That means that music is being fully transferred and its quality is significantly better than in the case of the AM transfer. The FM transfer has some other advantages, perhaps the most significant of them being the possibility of eliminating various disturbances that are manifesting themselves as snapping, squeaking etc. The main disadvantage, however, is not the result of the frequency modulation itself, but rather of the fact that this method is being used on high frequencies, and that high-frequency electromagnetic waves behave themself as light, spreading themselves in straight line, not reflecting from the ionosphere etc. This is why obtaining this kind of radio-link requires optical visibility between the transmission and reception antennas, which is not the case for the links obtained on frequencies which are less than 40 MHz. In practical terms, it is possible to receive the SW signal from anywhere on Earth, whilst the range of an UHF link is limited to the horizon. Or, as Hamlet would say: "The quality or the range, that is the question!"

Can we have it both, somehow? Yes we can, and it is already being done, over the satellite links, using the same equipment as for the TV signal receipt and an audio amplifier connected to the audio output of the satellite receiver. For now, in the earthly conditions, those that are interested in the worldwide news will make and use the AM receivers, and music lovers will stick to the FM's. And what can those interested in both do? Well, they make AM-FM receivers :)

The direct-type (TRF) FM receivers have never been produced, the industry started right away with the superheterodynes, made acc. to the block diagram on Pic.4.6, which will later be discussed. In amateur life, however, the direct FM receivers do exist, having very simple electronic diagrams and being easy to manufacture. These receivers have very strong positive feedback, making the intermittent oscillations in it, and are therefore being called the super-reaction receivers.

3.15.1. The Simplest FM Receiver

On Pic.3.43 you can se the electronic circuit of an extremely simple direct FM receiver. The T2 transistor together with the R1 resistor, the coil L the variable capacitor C and internal capacitances of the T1 transistor, comprises the so-called Kolpitz oscillator. The resonance frequency of this oscillator is being set by C to correspond to the one of the station that we wish to hear (meaning it has to be altered between 88 and 108 MHz). The signal, i.e. the information being used in the transmitter to perform the modulation, is extracted on the R1 resistor, and being led from it to the high-resistance headphones, over the coupling capacitor C1.

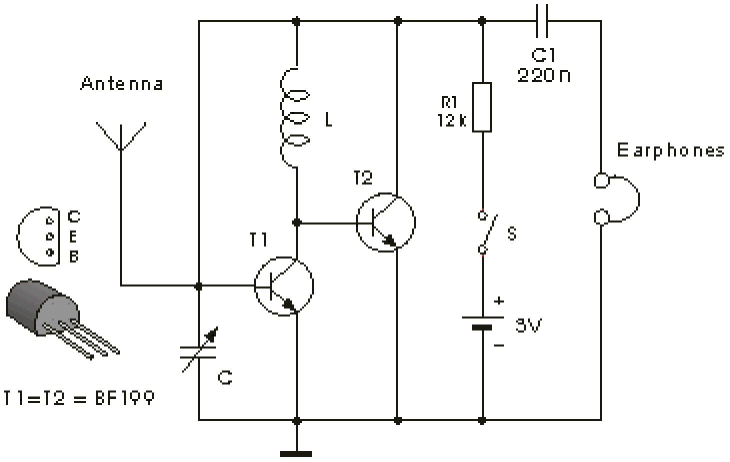

Pic.3.43. The simplest FM receiver

* The capacitance of the variable capacitor should be able to change from a couple of pF (Cmin) to app. 20 pF. During the testing off this device, we were using the capacitor from Pic.3.8. The legs marked as FO and G were used, the G leg being connected to the ground. When all the trimmers from the circuit on the Pic.3.8 are set to minimum capacitance, the capacitance between the FO and G legs should be adjustable between 7 and 27 pF.
* The coil L has 4 quirks of lacquer-isolated copper wire (CuL), bended to have a 4 mm internal diameter. The practical realization of this coil is explained in text connected with Pic.3.45. During the setup of the bandwidth, the inductance of the coil can be altered by changing the distance between the quirks. If the coil is stretched the inductance decreases, and vice versa. If this cannot give the desired results, new coil must be made.
* The telescopic antenna taken from a disused device can be used. If you can’t find one, you obtain very good results with a piece of isolated copper wire, about 60 cm long (the optimum length to be found experimentally).

3.15.2. The Simplest FM Receiver with Audio Amplifier

The radio-broadcast FM transmitters operate with output power that is much smaller than that of the AM transmitters. That is why the LF signal coming from the device on Pic.3.43 is rather small, urging the use of very sensitive headphones. They are much more expensive than the “ordinary” ones, making it better to use the cheap headphones in connection with audio amplifier. One such solution where TDA7050 IC is used is given on the Pic.3.44. The R3 resistor and capacitors C5 and C6 are to be added only if the operation of the device is unstable. There optimum values are to be found experimentally, starting with those shown in the picture.
For loudspeaker reproduction any of the previously described amplifiers can be used, e.g. that from Pic.3.21 (which we have been using, very successfully), or one of the devices described in P.E.4 and P.E.5. Since in these amplifiers a battery with voltage bigger than 3 V is used, using of R3 and C5 is obligatory. The R3 is counted from the formula

$$R_3 = \frac{U_{BAT} - 3}{0,235} \left[k\Omega\right]$$

where UBAT is battery voltage, and 0.235 mA is the current through R1, that supplies T1 and T2. E.g. if UBAT=9 V, it is then and the nearest existing resistor is used.

$$R_3 = \frac{9-3}{0{,}235} = 25{,}5\,k\Omega$$

Capacitors C5 and C6 comprise, together with R3, a pass-filter for very low frequencies, which is used to separate the HF and LF parts of the receiver.

The battery itself acts as a short-circuit for the AC currents. But when it ages its resistance increases and there is no more short-circuit. That is why C3 and C4 are added, to accomplish it.

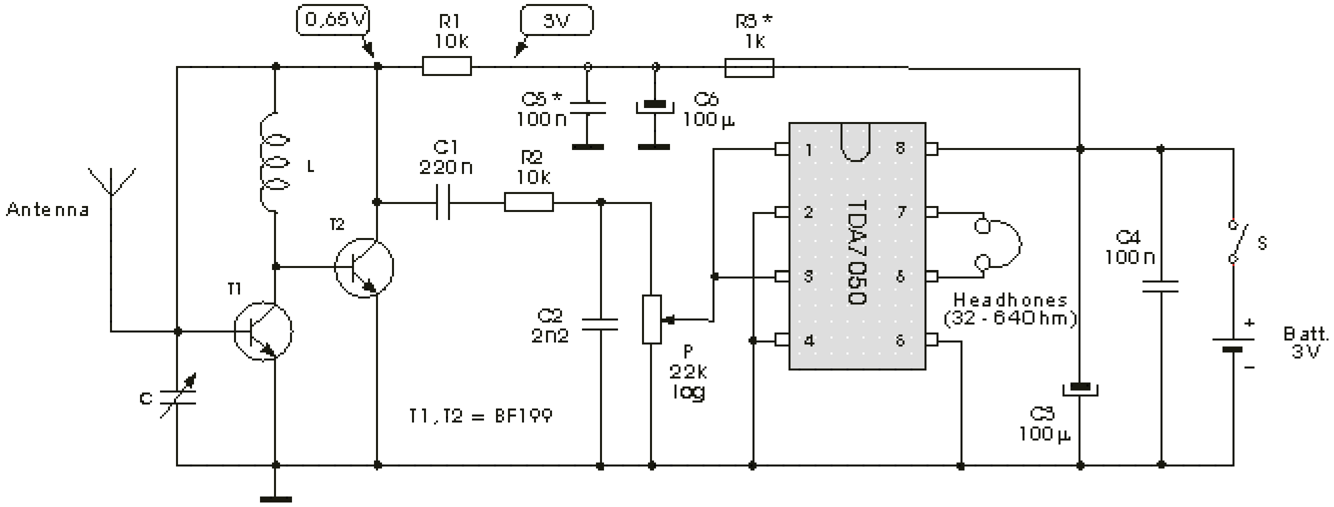
0,65V
R1
10k
3V
R3 *
1k
C5 *
100n
C6
100µ
C1
220n
R2
10k
L
Antenna
T2
T1
C
C2
2n2
P
22k
log
T1, T2 = BF199
TDA7050
1
2
3
4
5
6
7
8
Headhones
(32 - 64Ohm)
C4
100n
S
+
-
Batt.
3V
C3
100µ

Pic.3.44. The simplest FM receiver with audio amplifier

3.15.3. FM Receiver with one Transistor and Audio Amplifier

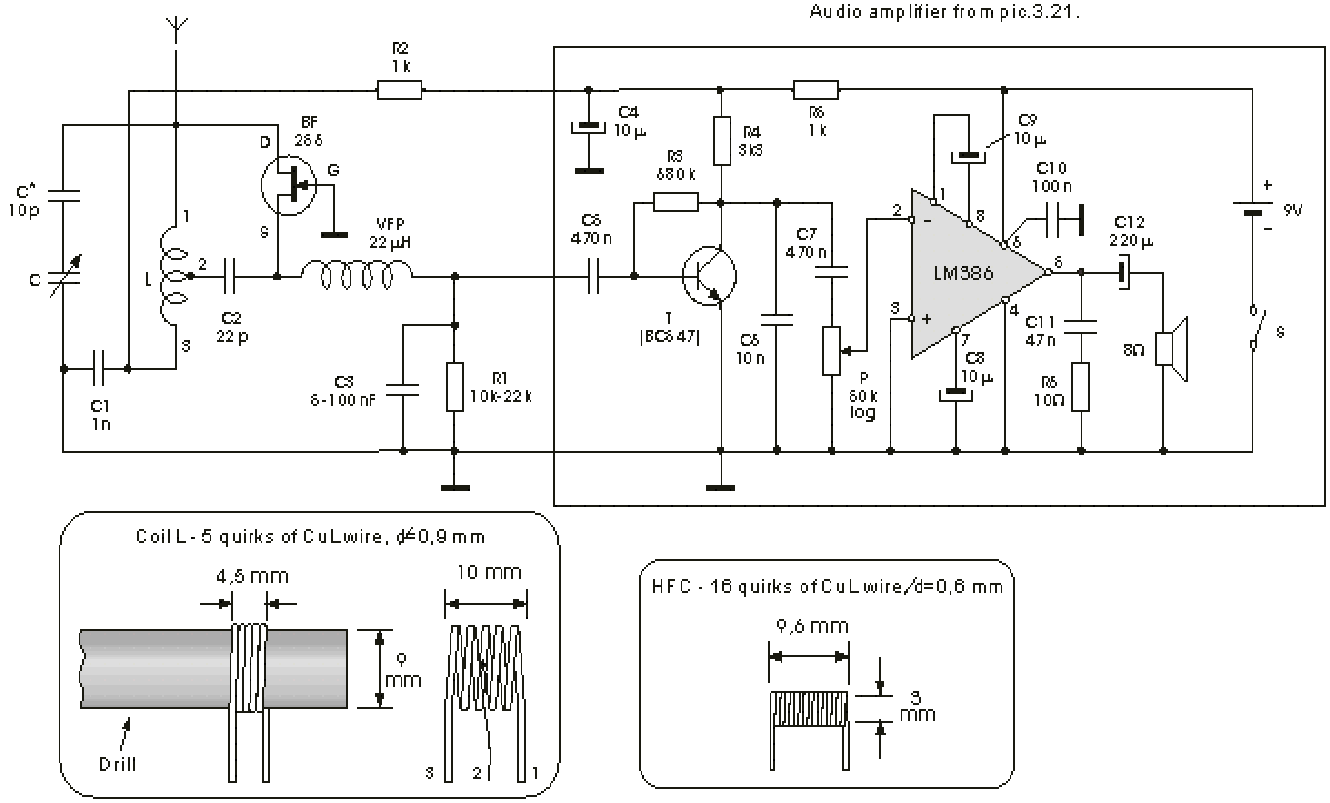
Audio amplifier from pic.3.21.
R2
1k
C*
10p
C
C1
1n
L
1
2
3
D
S
G
BF
256
C2
22p
VFP
22 µH
C3
6-100nF
R1
10k-22k
C4
10 µ
C5
470n
R3
680k
R4
3k3
T
(BC547)
C6
10n
C7
470n
P
50k
log
R6
1k
C9
10 µ
LM386
C8
10 µ
C10
100n
C11
47n
R5
10Ω
C12
220 µ
8Ω
9V
S
Coil L - 5 quirks of CuL wire, d=0,9 mm
4,5 mm
10 mm
9 mm
Drill
HFC - 16 quirks of CuL wire/d=0,6 mm
9,6 mm
3 mm

Pic.3.45. FM radio receiver with audio amplifier

We have made this receiver on the experimental plate, and it was playing for days in our lab. Its electronic diagram is given on Pic.3.46. Regretfully we had to disassemble it, since we needed the plate for one of the devices described later in this book. This, too, is a reaction-type receiver, where the BF256 transistor, coil L and capacitors C, C* and C2 form the Hartley oscillator. Its frequency is being adjusted by means of the variable capacitor C to be equal to the frequency of the station that we wish to listen to. The LF signal is being taken from the R1 resistor, and led into the audio amplifier.

* The coil L is self-supporting (doesn't have the body), made of 5 quirks of CuL wire, its diameter being from 0.8 to 1 mm. It is spooled on some cylindrical object (pencil, pen etc., the best thing is the round part of a 9 mm drill), in one layer, quirks put tight to each other, as shown in the left, framed part of the picture. When the coil is finished, it is taken off the cylinder and stretched a little, so that the quirks do not touch each other. Its final length should be about 10 mm. The mid coil leg, which is to be connected to the left end of the C3 capacitor, is made by taking off couple of millimetres of the lacquer from the wire, approximately in the middle of the coil. This place is then tinned and a piece of thin wire is soldered to it. The other end of this wire is soldered onto the PCB, on its place, to be connected to the left end of C2.

* For the variable capacitor C the one from the Pic.3.8 (legs FO and G, G goes to Gnd). If you are using some other capacitor, that has bigger capacitance, and you cannot achieve the reception of the full FM bandwidth (88 til 108 MHz), try changing the value of the C*. Its capacitance is to be determined experimentally, usually being about a dozen pF.

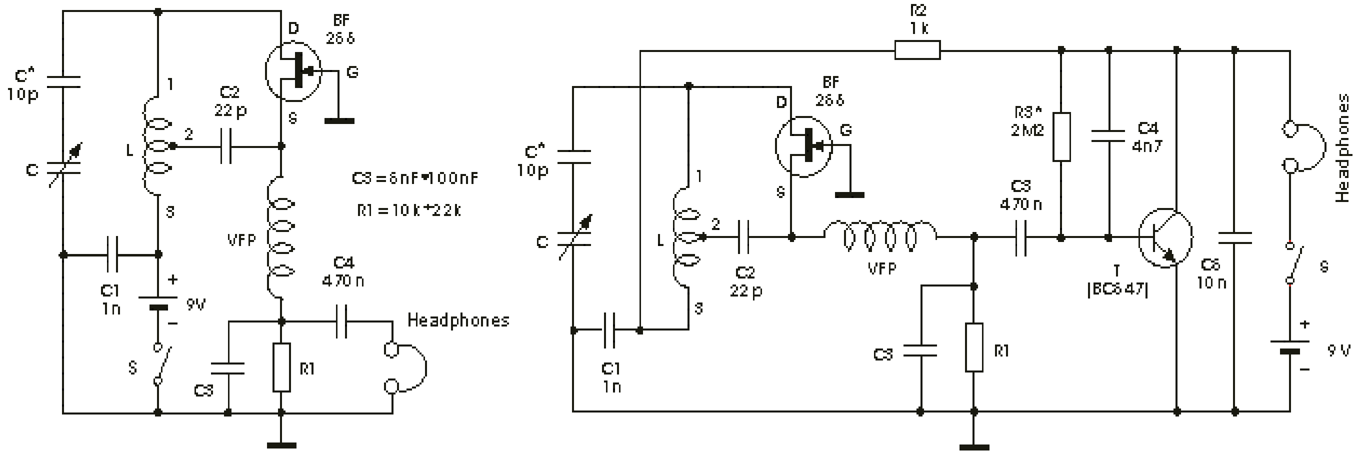
C*
10p
C
L
1
2
S
C2
22p
D
BF
256
G
S
VFP
C3 = 6nF÷100nF
R1 = 10k÷22k
C4
470n
Headphones
C1
1n
+
9V
−
S
C3
R1
R2
1k
BF
256
D
G
S
C*
10p
C
L
1
2
S
C2
22p
C1
1n
VFP
C3
470n
C3
R1
R3*
2M2
C4
4n7
T
(BC547)
C5
10n
Headphones
S
+
9V
−

Pic.3.46. FM receivers with headphones

* HFC is the high-frequency choke. Together with C2, it makes a filter that prevents the HF current to flow through the R1, simultaneously allowing for DC and LF current to go through. The muffler is, in fact, a coil that has 16 quirks of 0.6 mm CuL wire, spooled on a round part of a 3 mm drill.

* This receiver works well even without the external antenna. It can, of course, be connected to it, as shown in dashed line. Instead of antenna, a 50 mm piece of wire can also be used.

3.15.4. FM Receiver with (just) one Transistor

On the left side of the Pic.3.46 you can see the diagram of another very simple FM receiver, that has only one transistor as the active element. That is, as one can see, the HF part of the receiver from Pic.3.45, where the reproduction is being accomplished over the high-resistance headphones. But, as previously noticed, they are pretty expensive, therefore making it better to use the "regular" headphones and a simple amplifier, as shown on the right side of the Pic.3.46.

Chapter 4 Superheterodyne Radio Receivers

4.1.Superheterodyne AM Receivers
4.1.1. The Simplest AM Superheterodyne Receiver (worldwide)
4.1.2. The Fully (not exactly 100%) superheterodyne AM Receiver No.1
4.1.3. The Fully (not exactly 100%) superheterodyne AM Receiver No.2
4.2. Superheterodyne FM Receivers
4.2.1. FM Receiver with TDA7000
4.2.2. FM Receivers with TDA7088T
4.2.2.1. Miniature FM Receiver
4.2.2.2. Stereophonic Receiver

As far as the professional manufacturers are concerned, the direct (TRF) receiver has "played out his tune". After half-century struggle on the market, it has been replaced by the superheterodyne receiver, that was patented in 1918 by Edwin Armstrong. In that time, commercially speaking, its main advantage was its substantially easier tuning to the station. It requires only one button for this, comparing to the TRF receiver that needs two buttons to be intermittently adjusted for optimal reception, and also it requires much of the knowledge, skill and patience, which the average buyer does not have. The superheterodyne receiver is, however, also more complex than the TRF, and setting of its stages during its production requires some special instruments, that the average radio amateur does not possess. Nevertheless, it is not impossible to build such device in the amateur environment, and when the operating principles are known, the necessary adjustments can be done "by hearing".

4.1. Superheterodyne AM Receivers

On Pic.4.1 you can see the block diagram of a radio-broadcast superheterodyne receiver
The input circuit (UK) refines the signal of the tuned station from all the voltages created in the antenna (A) by various radio transmitters and sources of disturbances. In our example, it's an AM signal that has the carrier frequency fs, and is modulated by a single tone, as seen in the rectangle above its label. This signal is being led into the stage called the mixer. Another voltage is also led into it, the voltage from the local oscillator that has the frequency of f0, and a constant amplitude. Under the effect of these two signals, the phenomenon called the outbreak takes place in the mixer, and an AM signal appears on its output, its frequency being fm=455kHz. This signal is called the inter-frequency (IF) signal, and its frequency fm the interfrequency. The IF signal has the same envelope as the station signal entering the mixer. That means, that the information from the transmitter to the mixer is carried by the signal frequency fs, and in the mixer it is being assumed by a new carrier, that has the frequency fm. When transferring to another station, the user changes the capacitance of the variable capacitor C by turning the knob, setting up the resonance frequency of the input circuit to be equal to that station's one. Another variable capacitor, Co, is located on the same shaft as C, so its capacitance changes simultaneously to that of C. This capacitor is located in the local oscillator and that is how it gets the new oscillating frequency, having such value that the difference of the oscillator and station frequencies is again equal to the inter-frequency value.
Here's one numerical example. The interfrequency is being adopted by the constructor of the device, and it is mostly fm=455 kHz. When the receiver is set to the station that has the frequency of fm=684 kHz, the frequency of the local oscillator shall be fO=1139 kHz, therefore making there difference be
1139 kHz-684 kHz=455 kHz=fm.
When tuning to a station that operates on the frequency of fS=1008 kHz, the listener will change the capacitances of the two capacitors until the resonant frequency of the input circuit becomes fS=1008 kHz, and the oscillator frequency fO=1463 kHz, therefore yielding 1463 kHz-1008 kHz=455 kHz=fm.
If the receiver has more wavebands (LW, MW, SW1, SW2…) it is being constructed to have the same inter-frequency value for all of them.
What do we gain with this change of the carrier frequency? So far we haven't mentioned one very important thing, that is that the input circuit can never be selective enough, to extrapolate only the signal of the tuned station, from all the signals that exist in the antenna. On the output of this circuit, besides the station signal, also signals of strong and local transmitters are obtained, especially the signals from the neighbouring channels (their frequency being very close to the one of the tuned station). All these signals are receiving new signal carriers in the mixing stage, with their frequencies deviating fm as much as their carrying frequencies differ from fS. E.g., if the input circuit is set on the station whose frequency is 1008 kHz, another two signals from the neighbouring channels can also emerge on its exit.

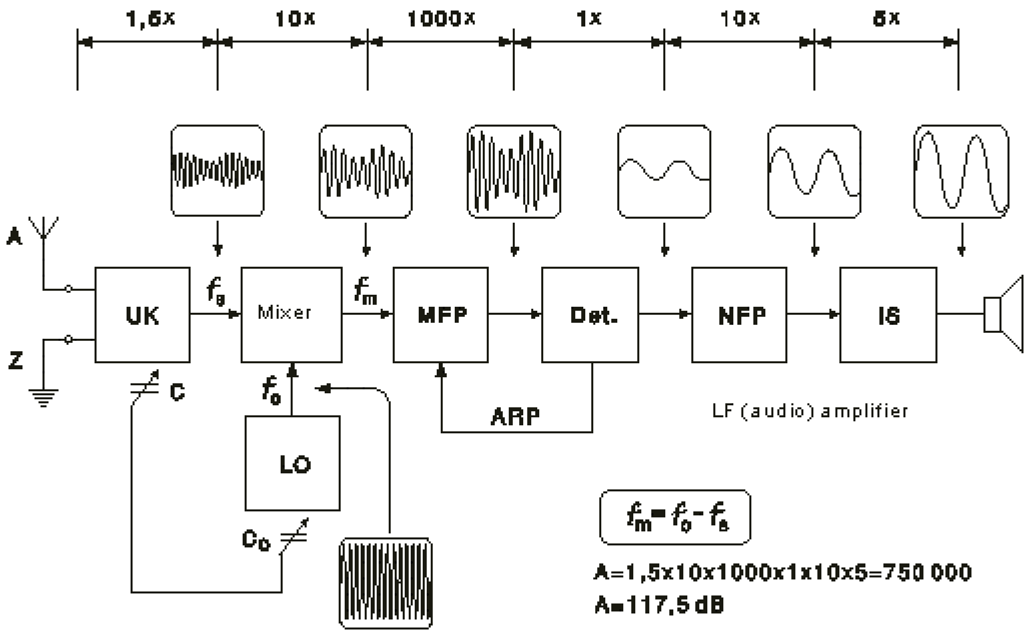

Pic.4.1. Block diagram of a superheterodyne AM receiver

Their frequencies would be 999 kHz and 1017 kHz. The ordinary TRF receiver would in this case be totally incapable of suppressing those signals, which is not the case with the superheterodyne receiver. These 3 signals are entering the mixer, which gets the 1463 kHz voltage from the oscillator. The outbreak occurs, and 3 AM signals are exiting the stage, their frequencies being 455 kHz, 464 kHz and 446 kHz. All 3 signals go to the IF amplifier (MFP), which has several amplifying stages with oscillatory circuits set to 455 kHz, making it very selective, so it amplifies only the 455 kHz signal and suppresses the others enough not to disturb the reception.

the signal exiting the IF amplifier is led onto the detector (Det.), the LF voltage amplifier (NFP) and the output stage (IS), the circuits we spoke about in the previous projects.

The ARP signifies the circuit that turns back the DC component of the detected signal into the IF amplifier, to obtain the automatic amplification regulation.

Above every block on the picture you can see the signal shape exiting that block, as seen on the oscilloscope, in case the modulation in the transmitter is done by the single, sinusoidally-shaped tone. The upper part of the picture contains the average voltage amplifications for each block, for the mass-production devices. Total voltage amplification, which is the ratio of the voltage on the loudspeaker to the voltage in the antenna is A=750000. The amplification in decibels is therefore: A(dB)=20logA=117.5

4.1.1. The Simplest Superheterodyne AM Receiver

The author presumes that most of the readers, especially those just entering the world of radio with this book, are somewhat scared by the block diagram from Pic.4.1. Their question probably is: Can an amateur build such a receiver? Yes, he can. The author has a friend that succeeded in this some 40 years ago, when all had been done with the electronic tubes, making the practical realization of a receiver much harder than it is today, with semiconductors (its radio amateur call sign is YT1FA, and those who doubt it may contact him). However, he was doing this in the premises of YU1EXY Radio Club, in the attic of the Electrotechnical Faculty in Belgrade, using the club (more less trophy) instruments and, more important, he had help of Sasa Piosijan, Radivoje Karakasevic and Kiro Stojcevski,

who knew all about the radios, especially Sasa.
The main problem in making a superheterodyne device is not the circuitry complexity but its setup, which requires lot of practical experience and some special instruments, that our readers probably don't possess. But they are much better than the TRF receivers, both regarding the sensitivity and selectivity, so we made simpler devices that are simple to set, with no instruments necessary than your ears. They are realized around the NE612 IC, whose pin description, block diagram and main features are given on Pics.4.2-a & b.
This IC comprises the critical stages of an AM superheterodyne receiver, the mixer and local oscillator. the station signal is led either on pin 1 or on pin 2 (or on both of them, in case of symmetrical coupling with the previous stage), and the IF signal is obtained on the pin 4 or 5 (or on both of them, in case of symmetrical coupling with the next stage). An oscillatory circuit, that determines the frequency of the local oscillator and the positive feedback circuit are connected between the pins No.6 and 7. Pin 3 is connected to Gnd, i.e. the minus pole of the DC supply voltage. Pin 8 receives a positive DC supply voltage which can, acc. to the table given on Pic.4.2, vary between 4.5 V to 8 V. The value of this voltage is not critical, but it is extremely important for normal operation of the receiver that this voltage is stable, therefore urging for it to be separately stabilized (with special care), as seen in some projects in this chapter and in the Appendix, that involve the NE612.
In the text that follows 3 simple superheterodyne receivers made with NE612 will be described.

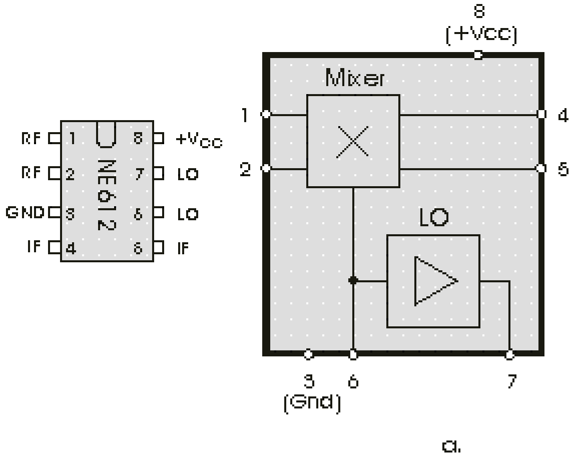
RF 1
RF 2
GND 3
IF 4
8 +Vcc
7 LO
6 LO
5 IF
NE612
8
(+Vcc)
Mixer
LO
1
2
4
5
3
6
7
(Gnd)
a.

NE612 IC Characteristics (NE602, SA612, SA602)

Parameter	Symbol	Range	Unit
Supply Voltage	V_{CC}	4,5 ··· 8	V
Supply Current	I_{CO}	2,4 ··· 3	mA
Input resistance (between 1 & 2 & Gnd)	R_{in}	1,5	kΩ
Output resistance (between 4 & 5 & Gnd)	R_{out}	1,5	kΩ
Mixer amplification	A_{mix}	14 ··· 17	dB
Maximum oscil-lator frequency	f_{omax}	200	MHz
Maximum mixer frequency	f_{Mmax}	500	MHz
Outside oscillator votage	U_o	200	mV

b.

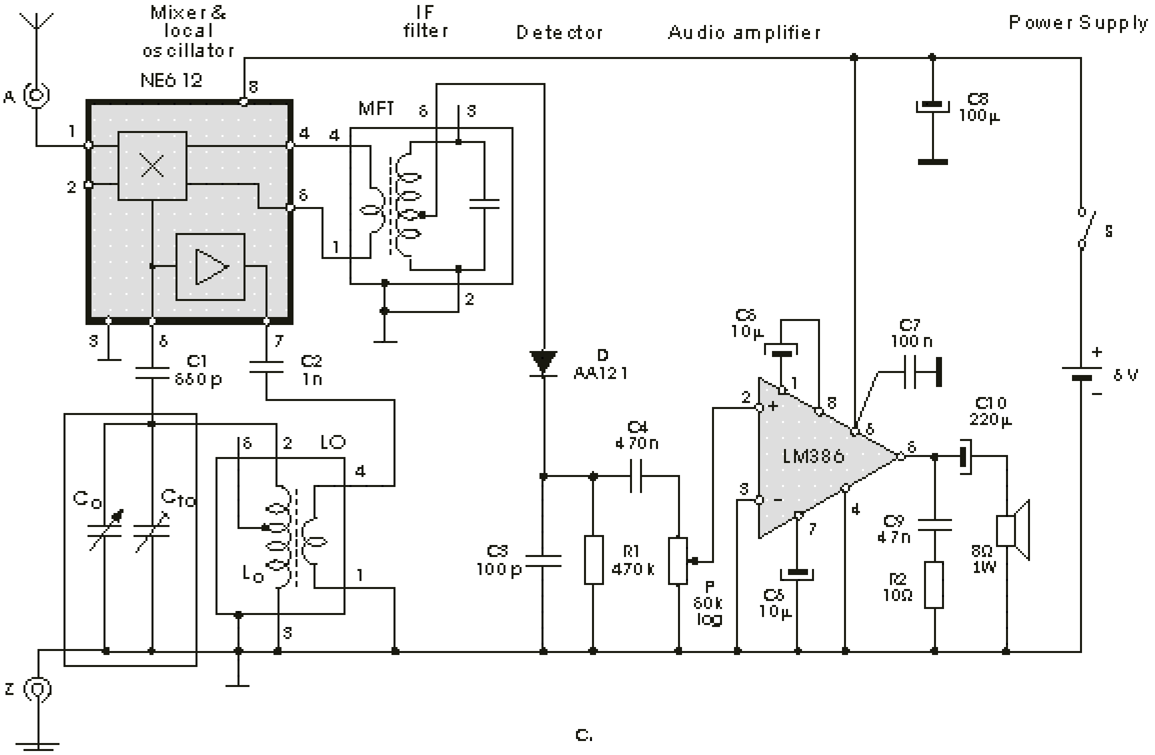
Mixer & local oscillator
IF filter
Detector
Audio amplifier
Power Supply
A
NE612
1
2
3
4
5
6
7
8
MFT
C1
560p
C2
1n
Co
Cto
Lo
LO
Z
D
AA121
C3
100p
R1
470k
C4
470n
P
50k
log
C5
10μ
C6
10μ
LM386
C7
100n
C8
100μ
C9
47n
R2
10Ω
C10
220μ
8Ω
1W
S
6V
c.

Pic.4.2. The simplest superheterodyne AM receiver: a-pin layout and block diagram of the NE612, b-IC technical data, c-electronic diagram of the receiver

More will be discussed in the chapter dedicated to NE612 IC, and the reader should pick one of these, or make the receiver that suits him best by combining these diagrams with earlier described HF amplifiers and input circuits.

The electronic diagram of the simplest superheterodyne AM receiver in the world, with reproduction over the loudspeaker, is shown on Pic.4.2-c. The device has got only one oscillatory circuit in the IF amplifier (being marked as MFT), whose frequency does not need to be set to some specific value (meaning the receiver will work OK even if its frequency is bigger or smaller than standard 455 kHz). Further simplification was done by omitting the

input circuit, therefore avoiding the problems with quite complex adjustments between the input circuit and the local oscillator. All these simplifications do have their price: this device is less sensitive and selective than the complete superheterodyne, and is also more prone to disturbances. Even so, it has better both the selectivity and sensitivity than the TRF.
Signals of all the stations are being led directly from the antenna onto the pin no.1, i.e. the mixer. On the other hand, the mixer also receives the HF voltage from the local oscillator, whose frequency is equal to the resonance frequency of the parallel oscillatory circuit made of CO, CtO, and LO. This frequency, if neglecting the parasite capacitances, is:

$$f_o = \frac{1}{2\pi\sqrt{L_o(C_o + C_{to})}}$$

On the mixer exit the signals from all the stations are obtained, but now they have new carrier frequencies, that are equal to the difference of the oscillator frequency and their original one. Nevertheless, only one of these signals will have the frequency that is equal to the resonance frequency of the MFT, and it will be the only one to appear on the ends of this oscillatory circuit. Here's a numerical example.

Let us assume that we have (only) 3 MW signals in the antenna, having the frequencies of fS1=711 kHz (Nis), fS2=855 kHz (Bucharest) and fS3=1008 kHz (Belgrade 2). The IF transformer frequency could be fm=455 kHz. If we set the frequency of our oscillator on fm=1166 kHz (with CO), the following signals, modulated by the radio stations' programs, will exit the mixer:

fm1=f0-fS1=1166-711=455 kHz,
fm2=f0-fS2=1166-855=311 kHz and
fm3=f0-fS3=1166-1008=158 kHz.

Since the oscillatory circuit on the mixer exit (MFT) is set to 455 kHz, we will have Radio Nis's signal from it, others will be suppressed. If we wish to hear Bucharest, the oscillator frequency should be set to 1310 kHz, and for Belgrade 2 1463 kHz. Of course that the listener doesn't need to know all these frequencies, he will just turn the knob of CO until hearing the desired station's broadcast.
The IF signal is led from the pin 3 to the detector with AA121 diode. The LF signal is taken from the R1 resistor and over the capacitor C4 it is led to the volume potentiometer P and the audio amplifier.

* the MFT is also being called the inter-frequency transformer. It is a special type component that is hard to find in the ordinary electronic shops, therefore the radio amateurs are usually obtaining them from disused factory-made devices. The IF transformer is shown on Pics.4.3-a,b,c & d. As you can see it on 4.3-a, the MFT is, in fact, a parallel oscillatory circuit with a leg on its coil. The coil body has a ferrite core (symbolically shown with single upward straight dashed line) that can be moved (with screwdriver), which allows for the setting of the resonance frequency of the circuit, being mostly fm=455 kHz. The same body contains another coil, with less quirks in it. Together with the bigger one it comprises the HF transformer that takes the signal from the oscillatory circuit into the next stage of the receiver. Both the coil and the capacitor C are placed in the square-shaped metal housing that measures 10x10x11 mm (Pic.4.3-b). From the bottom side of the

housing you can see 5 pins emerging from the plastic stopper, that link the MFT to the PCB, being connected inside the MFT as on Pic.4.3-a. Besides them, there are also two noses located on the bottom side, that are to be soldered and connected with the device ground. Japanese MFT's have the capacitor C placed in the cavity of the plastic stopper, as shown on Pic.4.3-c. The part of the core that can be moved with the screwdriver can be seen through the eye on the top side of the housing, Pic.4.3-d. This part is coloured in order to distinguish the MFT's between themselves, since there are usually at least 3 of them in an AM receiver. The colours are white, yellow and black (the coil of the local oscillator is also being placed in such housing, but is being painted in red, to distinguish it from the MFT).
Un-soldering the MFT isn't that simple and is to be performed very carefully. The iron is not to be kept leaned too long on the pins, since there's danger of melting the plastic stopper. All the tin from the pins and noses has to be removed first, by the aid of the iron and the vacuum pump (or a piece of wire stripped from the antenna coaxial cable). You can then safely remove the MFT from its original PCB.

* Pics.4.3-a, b, c & d almost fully apply for the oscillator coil as well (LO). The only difference is that LO doesn't have the capacitor C. looking from the outside, LO and MFT can be distinguished only by the marking colour, until they're lifted from the PCB.

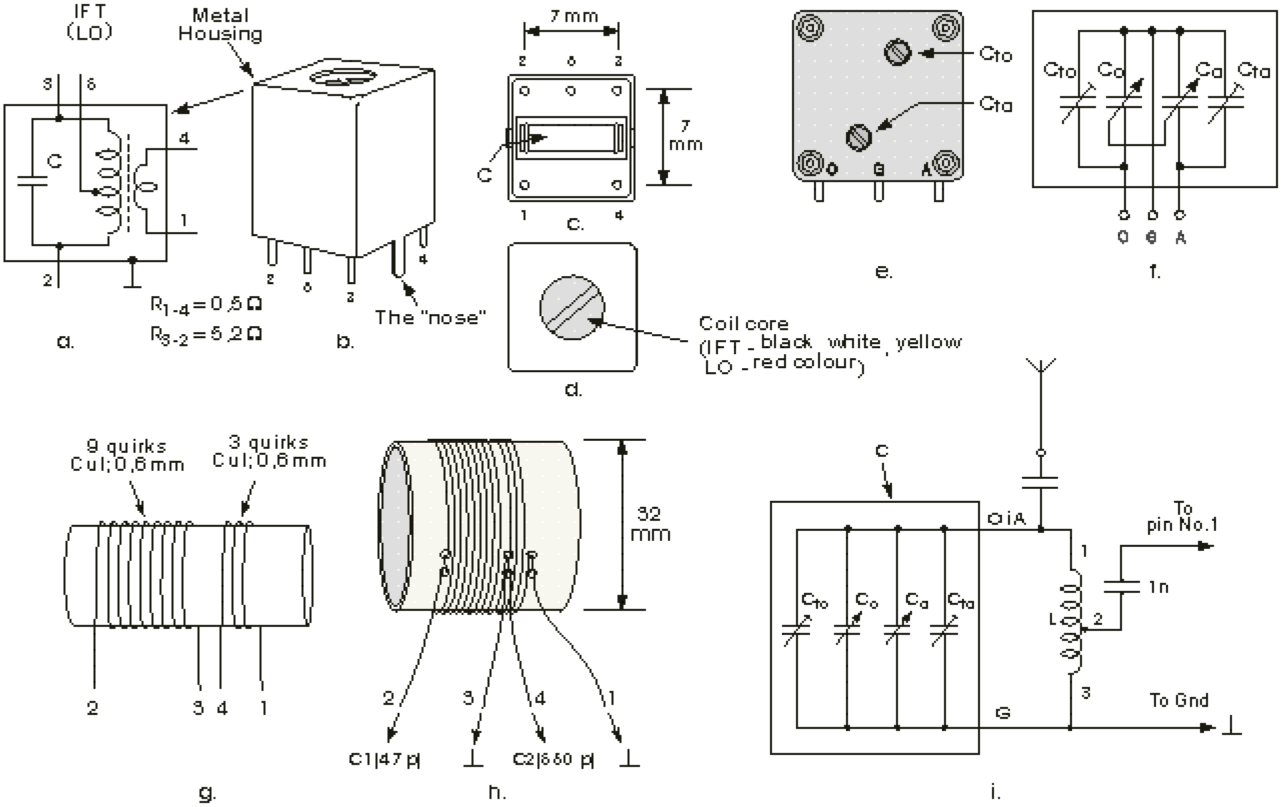
IFT
(LO)
Metal Housing
3
6
4
1
2
C
$R_{1-4}=0,5\Omega$
$R_{3-2}=5,2\Omega$
a.
2
6
3
4
The "nose"
b.
7mm
2 6 3
7 mm
C
1 4
c.
d.
Coil core
(IFT - black, white, yellow
LO - red colour)
Cto
Cta
O G A
e.
Cto Co Ca Cta
O G A
f.
9 quirks CuL;0,6mm
3 quirks CuL;0,6mm
2 3 4 1
g.
32 mm
2 3 4 1
C1|47 p|
C2|560 p|
h.
C
OiA
To pin No.1
1
1n
L
2
3
To Gnd
G
Cto Co Ca Cta
i.

Pic. 4.3. Components of the receiver from pic.4.2: a,b,c,d-IFT and LO; e,f-variable and trimmer capacitors of the local oscillator, g,h-SW oscillator coil; i-independent input circuit

LO's have red colour, while MFT's (IFT's) are white, black or yellow. During the PCB design, absolute care must be taken that pins 1 & 4, as well as 2 & 3, do not permute. If that would happen, the feedback would be negative (instead of positive) and the oscillator wouldn't function. However, if you conclude during the design phase that it would be more convenient to connect pin 4 to Gnd (instead of pin 1), do have in mind that it can be done only if you connect also pin 2 to Gnd (instead of pin 3).

* Fine tuning (if necessary) of the LO's and MFT's inductance values is done by adjusting the position of the ferrite core with screwdriver.

* With CO and CtO, variable capacitor and the trimmer capacitor in the oscillator are labelled. Acc. to Pic.4.3-e & f, which shows the capacitor we spoke about in the connection with Pic.3.7, the abovementioned capacitors are connected with the circuitry over the legs O and G (Ca and Cta are not used), with G connected to Gnd.

* The receiver from Pic.4.2 can be utilized for the reception of AM stations in the SW waveband. All there is to be done is to make a new oscillator coil, acc. to Pic.4.3-g & h. It is being made of 0.4 mm CuL wire (a thicker one can also be used), on the 32 mm diam. carton body, the same one used for making coils on Pics.3.6 & 3.28. Number of quirks on the picture is 9, but other combinations should also be tried, say, 12 quirks, or somewhat less than 9. The feedback coil has 3 quirks and is spooled along the oscillator coil (as shown on picture), or over it. If you have already accomplished the reception of SW stations with some of the previously described TRF devices, you will be surprised with much bigger selectivity of the receiver from Pic.4.2. in the evening hours you'll be able to perform the receipt of huge number of stations on the radio-broadcast, professional and amateur wavebands.
For the reception of SW stations smaller capacitances for C1 should also be tested, say, C1=33 pF and similar, since it affects the oscillator frequency.

* In the previous numerical example we saw that tuning is done by setting up the frequency of the local oscillator and that fm=455 kHz, Radio Nis will be heard when the oscillator frequency is fO=1166 kHz. The story is not over, though: What will happen if there is a station that operates on 1621 kHz? Mixing its signal with the voltage from the local oscillator the modified signal is made, its frequency being
1621 kHz-1166 kHz=455 kHz.
We now have two signals on the MFT. They both have the same carrier frequency (455 kHz), one of them is program of Radio Nis, and the other comprehends the program of the station transmitting on 1621 kHz. Both of them are being heard in the loudspeaker, the interference occurs. Speaking in expert language, the obstruction because of the symmetrical station occurred. That is a station whose frequency fSS iz greater than fm for the value of the oscillator frequency:
fSS=fO+fm
Suppressing the symmetrical station signal must be done before the mixing stage. In the radio-broadcast receivers this is being done over the input circuit, and in the professional devices, by input circuitry and the HF amplifier. If you have experienced disturbances while using the receiver from Pic.4.2 (mixing of stations or, more common, whistling or squeaking tone) try changing the MFT's oscillation frequency (by turning the ferrite coil), then re-tune the receiver.

* If the receiver from Pic.4.2 is power-supplied from the battery (or adaptor) whose voltage is over 6 V, a voltage stabilizer should be inserted in the plus (+) line of the power supply for NE612, as it was done with the receivers on Pics. 4.4, 5.7 and 5.9.
If you cannot receive the signal of some station transmitting on 1500 kHz, not even with the capacitor CO knob in the rightmost position, start reducing the CtO capacitance (turning the trimmer with screwdriver) until you hear the signal. similarly, if you can't hear some station you're fond of, that transmits on 500 kHz (e.g. Radio Budapest), try increasing the LO inductance (by turning the core towards inside with screwdriver). If this doesn't succeed, change a little the MFT frequency, then try again.

* The reception can be significantly improved if input circuit (UK) is added to the receiver. In order to avoid problems with attuning the UK and the LO, the UK with special variable capacitor can be used, as on Pic.4.3-i. It is "our" capacitor from Pic.3.7, with all the capacitors connected in parallel, and "our" coil from Pic.3.6. Station tuning is now being done with two buttons, which isn't "a job for everyone". The receiver is first roughly tuned to the station using these two buttons, and then the optimum reception is carefully searched.

* If you omit the amplifier with 386 IC on the Pic.4.2, and connect high-resistance headphones instead of R1, it is the truly the simplest superheterodyne receiver in the world.

4.1.2. The Fully (not exactly 100%) Superheterodyne AM Receiver No.1

Its electrical diagram is given on Pic.4.4. It is easily being noticed that this is the receiver from Pic.4.2 with inter-frequency (IF) amplifier with ZN415E added.
By adding ZN415 IC multiple enhancements are performed. Thanks to its huge input resistance, the MFT's oscillatory circuit is not choked, resulting in better selectivity. The sensitivity of the device is extremely increased since this IC has big amplification and the AAR (automatic amplification regulation) is also accomplished, making the usage of this device easier and more comfortable.
* It is very important to obtain the necessary value of the DC voltage in pin 6 of the ZN415 for its proper operation. Acc. to the table on Pic.3.36 it has to be about 1.3 V, and its setting is done via the TP1 trimmer. The receiver is set to some weaker station, the sound volume is made very low with potentiometer P, and the slider of the TP1 is carefully moved until the best reception is made. If that doesn't work, one should try changing the value of R5 resistor; this is to be done also if the supply voltage being used is other than 12 V. In case of voltage on the pin being much bigger than 1.3 V, and cannot be reduced on the trimmer, short-circuit one of the diodes.
* The voltage stabilizer with 78L06 isn't needed if the receiver is supplied from the 6 V battery.
* The receiver from Pic.4.2 needs input circuit to be 100% complete. That can be an independent input circuit from Pic.4.3-i, or input circuit and the HF amplifier that are described in the Appendix (Pic.5.10). If the former circuit is used, station tuning is being accomplished with 2 knobs, as explained in the previous chapter.

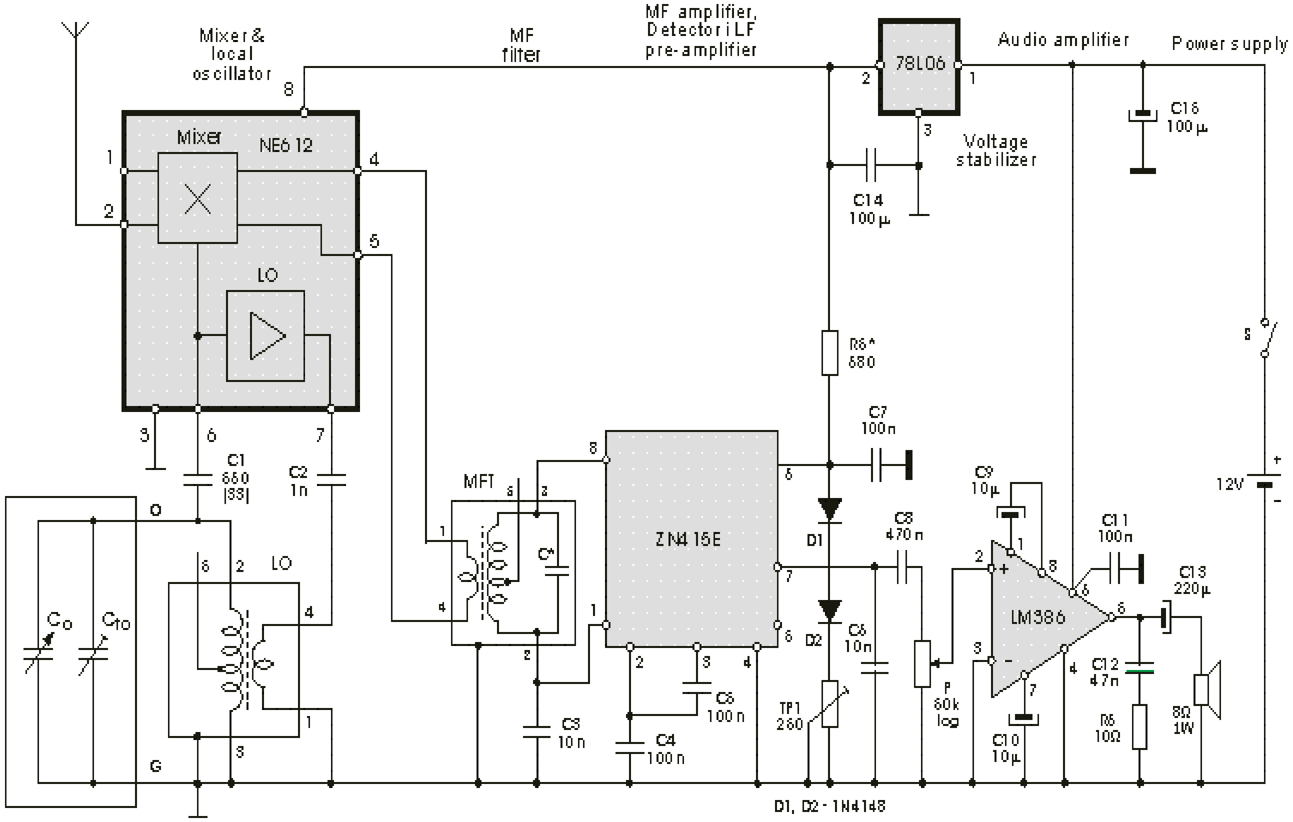
Mixer &
local
oscillator
MF
filter
MF amplifier,
Detector i LF
pre-amplifier
78L06
Audio amplifier
Power supply
Voltage
stabilizer
Mixer
NE612
LO
C16
100μ
C14
100μ
R6*
680
C7
100n
C1
680
(33)
C2
1n
O
Co
Cfo
LO
MFT
C*
ZN415E
D1
D2
C8
470n
C5
10n
C9
10μ
LM386
C11
100n
C13
220μ
C12
47n
R6
10Ω
8Ω
1W
12V
S
P
50k
log
TP1
250
C10
10μ
C3
10n
C6
100n
C4
100n
G
D1, D2 - 1N4148

Pic.4.4. Superheterodyne AM receiver with NE612, ZN415E & LM386

4.1.3. Fully (not exactly 100%) Superheterodyne AM Receiver No.2

All the receivers we made with NE612 IC were tested in our lab, except the one from the previous project, since we didn't have ZN415 "with us". We found, however, a ZN414 IC, so we tested the receiver from Pic.4.5 with it. The receiver was working great, from the amateur's point of view. He played us for long time, until we didn't require the board to test one of the receivers from previous projects afterwards, when we regretfully had to disassemble it.
* The diagram is very similar to that on Pic.4.4, so most of the things said about that receivers stands for this one, too.
* DC voltage setting on pin 1 of ZN414 is done with the trimmer TP. Its slider is put in mid position, the receiver is tuned to some weaker station close to the upper bound of the bandwidth. While making the reproduction very quiet (slider of P as low as possible), the trimmer slider is moved until reaching optimum reception. After that the trimmer is disconnected, its resistance measured and the ordinary resistor of similar value is put into circuit.
* The device operates nicely with the outside antenna made of a piece of wire measuring only half metres in length.
* The reception would certainly become better if an input circuit would be added, which we spoke about in the previous project.
* The receivers from pics. 4.4 and 4.5 can, with appropriate coils in the oscillator, accomplish the reception of AM stations from all the bandwidths from 70 kHz till 200 MHz.

4.2. Superheterodyne FM Receivers

The FM receivers being described in chapter 3.15 are the amateur solutions. These are extremely simple devices, that cannot perform the noiseless tuning, automatic oscillator frequency regulation and other features that ensure very high quality of the reproduction, being expected from an UHF FM receiver. The true solution is the superheterodyne FM receiver, whose block-diagram is given on Pic.4.6.

Station signals are taken from the dipole antenna and led through the appropriate cable into the input circuit (UK). Inside it, the signal selection is performed, of station whose frequency is fS, this signal is then amplified in the HF amplifier and led into the mixer. As in the case of earlier described AM receiver, the inter-frequency signal is obtained at the mixer output, whose carrier frequency is fm=10.7 MHz (this is the standard value, used in all radio-broadcast FM receivers). The IF signal is being amplified in the IF amplifier and led on the amplitude limiter (Ogr.). In this stage the signal whose amplitude exceeds certain level is being cut off, accomplishing with this the elimination of the parasite amplitude modulation, which is performed by various noise sources during the transmission (atmospheric charges, various electrical devices etc.), which significantly improves the signal quality. The signal then goes to the FM signal detector, where the information being modulated in the transmitter is extrapolated from the signal, followed by the LF part of the receiver. With AFC the circuit that performs the automatic frequency regulation of the local oscillator is labelled.

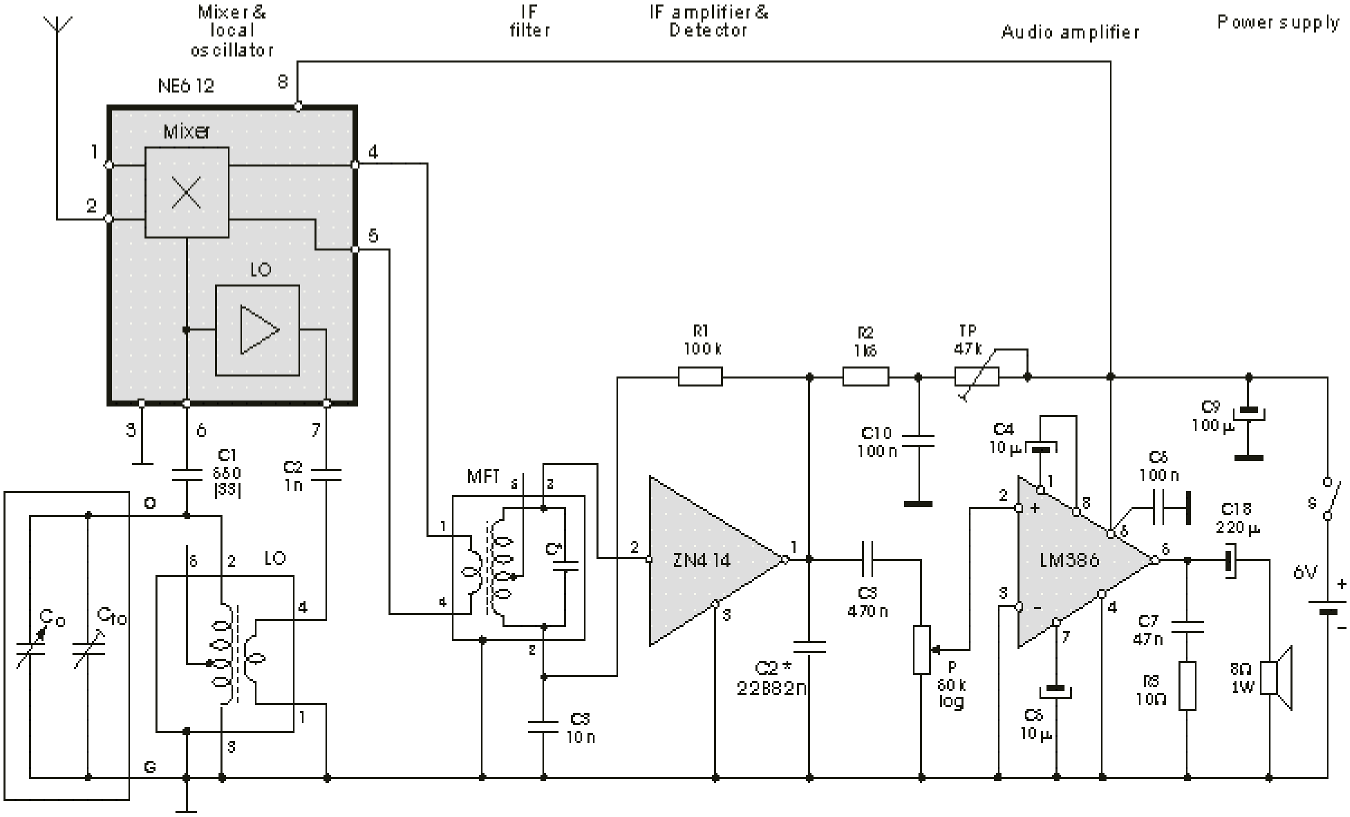
Mixer & local oscillator
IF filter
IF amplifier & Detector
Audio amplifier
Power supply
NE612
Mixer
LO
1
2
3
4
5
6
7
8
C1 560 (33)
C2 1n
O
G
Co
Cto
LO
MFT
C*
C3 10n
R1 100k
R2 1k6
TP 47k
C10 100n
ZN414
C2* 22B82n
C3 470n
P 50k log
C4 10µ
LM386
+
-
C6 10µ
C5 100n
C9 100µ
C18 220µ
C7 47n
R3 10Ω
8Ω 1W
S
6V
+
-

Pic.4.5. Superheterodyne AM receiver with NE612, ZN414 & LM386

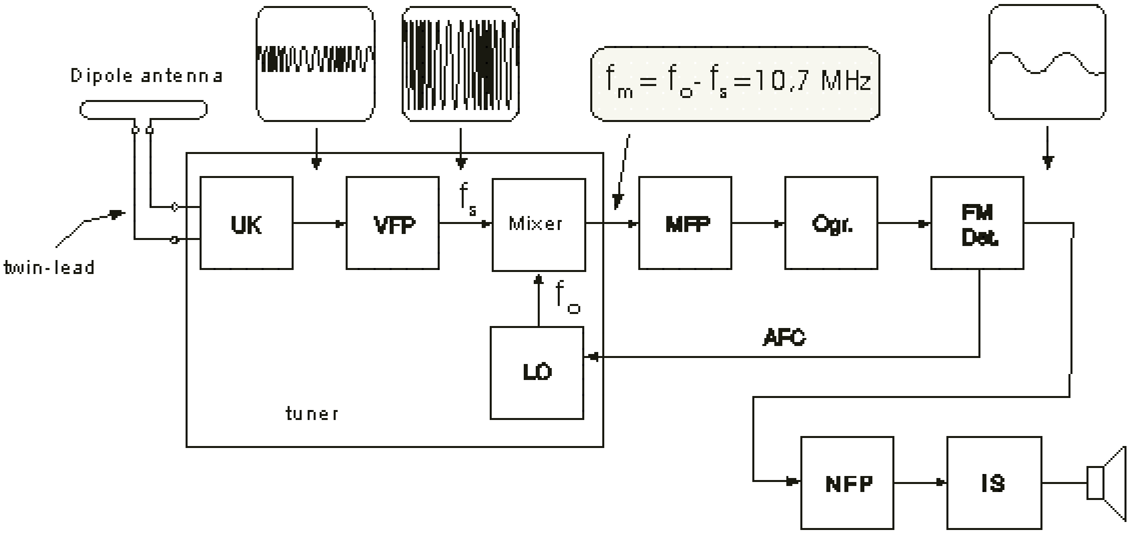

Pic.4.6. Block diagram of the monophonic superheterodyne FM receiver

4.2.1. FM Receiver with TDA7000

The face that FM receivers operate on pretty high frequencies makes their practical realization somewhat difficult, but most of the problems, as in many other amateur builds, originates from building the coils, except the self-bearing, small-inductance coils (without the coil body), which are easy to make, especially if there aren't many of them in the device and if no special instruments are required for setting up their proper inductance value. The coils used in this FM receiver are just like this, and there are only two of them, making the practical realization much easier.
The basic data about the famous Philips' IC used in this project, TDA7000, are given in the following table.

Supply Voltage	2,7 ÷ 10	V
Current Consumption (U_s=4,5V)	8	mA
Frequency Bandwidth	1,5 ÷ 100	MHz
Sensitivity	1,5	mV
LF signal on pin 2	75	mV
Package - DIL-18 ;SOT102-1	-	-

Electronic diagram of the HF part of the device (from antenna to the LF output) built with TDA7000 is shown on Pic.4.7. As one can see, it is a simple device, made with relatively small number of components. The IC contains all the stages of the superheterodine receiver: the mixer, the oscillator, the IF amplifier, the amplitude limiter, the FM detector and few others. More about them will be told in the next project which contains the description for a receiver with TDA7088T IC, which is the improved version of TDA7000.
The station signal is from the (telescopic) antenna led to the input circuit that consists of L2, C13, C12 and C14. It is a parallel oscillatory circuit damped with R3 resistor, which has the

reception bandwidth from 88 MHz till 108 MHz (it admits all the UHF signals on the pin 13, and weakens te signals outside the reception bandwidth). Inside the IC the signals are led into the mixer, where they are being given new carrier frequencies. The IF amplifier then follows, amplifying only one of those signals, the one whose frequency is equal to the inter-frequency, followed by the limiter, the FM detector, mute circuit and LF pre-amplifier. The output from the last stage is on the pin 2 (R2 is the collector load of the last transistor in the LF pre-amplifier). The oscillatory circuit of the local oscillator (L1, Cp, Cs, C and C5) is connected between pins 5 and 6.

Pic.4.8-a shows the PCB of the device from Pic.4.7, while Pic.4.8-b contains the component layout (on the PCB). The complete device can be seen on Pic.4.8-c. The variable capacitor from Pic.3.8 is used as the only variable capacitor here since the input circuit is aperiodic, the legs marked with FO and G. This capacitor serves us to tune the receiver to stations. In the LF part of the receiver, the amplifier made with LM386 from Pic.3.19 is utilized (the components left from the potentiometer are omitted).

* L1 and L2 are the self-bearing coils (without the core). They have few quirks and are made of relatively thick wire, therefore they don’t need a body of any kind, that is why they are called “self-bearing”. Their appearance is shown on Pic.4.9, and the calculus for them is done acc. to the table from Pic.3.5. They both have 6 quirks of the CuL wire, 0.6 mm in diameter, being spooled on the flat part of the 3 mm drill. In order to be able to solder the coil onto the PCB, the couple of mm of isolation has to be removed from the wire ends with sharp knife, and they have to be tinned afterwards. There must be a small gap between the adjacent quirks. The inductance of the coil is set by its shrinking (the inductance increases) or stretching (the inductance decreases). Stretching can be nicely done by inserting the screwdriver between the quirks and then turning it along the coil.

* The TDA7000 also contains the mute circuit (for noiseless tuning). It is being active when the S2 switch is open. Pocket-type receivers usually do not have S2 and R1 elements.

* The part of the receiver that requires biggest care during build is the oscillatory circuit of the local oscillator, which is connected between the pins 5 and 6. When changing the capacitance of C, its resonance frequency must change from 88 MHz (C=Cmax) till 108 MHz (C=Cmin). If that cannot be accomplished (not all the stations can be heard) some experimenting is required with capacitances of Cp and Cs. For start, you should omit the Cp. If the problem persists,

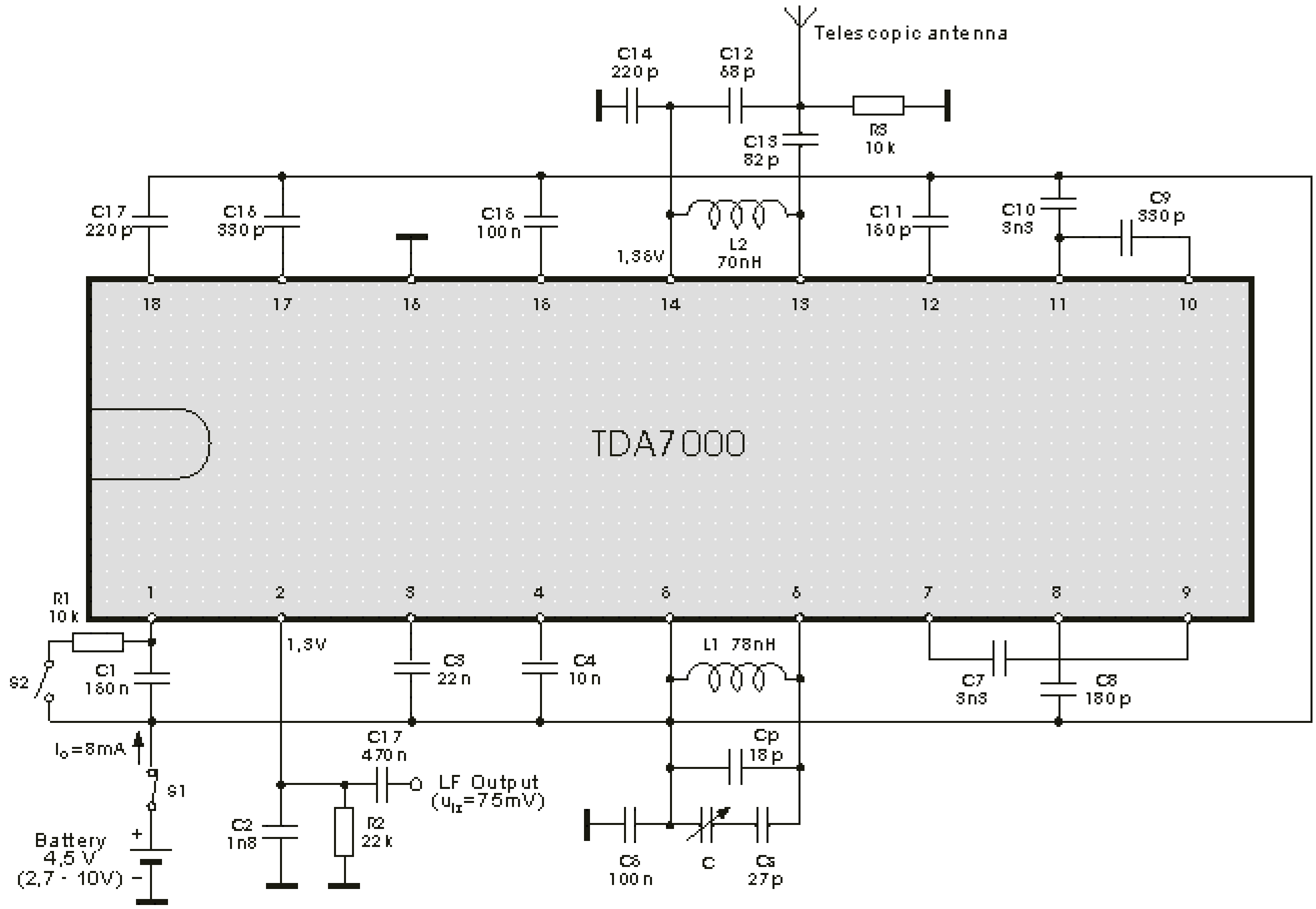
Telescopic antenna
C14
220 p
C12
68 p
C13
82 p
R3
10 k
C17
220 p
C16
330 p
C15
100 n
1,36V
L2
70nH
C11
160 p
C10
3n3
C9
330 p
18
17
16
15
14
13
12
11
10
TDA7000
1
2
3
4
5
6
7
8
9
R1
10 k
S2
C1
180 n
1,3V
C3
22 n
C4
10 n
L1 78nH
C7
3n3
C8
180 p
I_o=8mA
S1
C17
470 n
LF Output
(u_{iz}=75mV)
Cp
18 p
Battery
4,5 V
(2,7 - 10V)
+
-
C2
1n8
R2
22 k
C6
100 n
C
Cs
27 p

Pic. 4.7. FM radio receiver with TDA7000

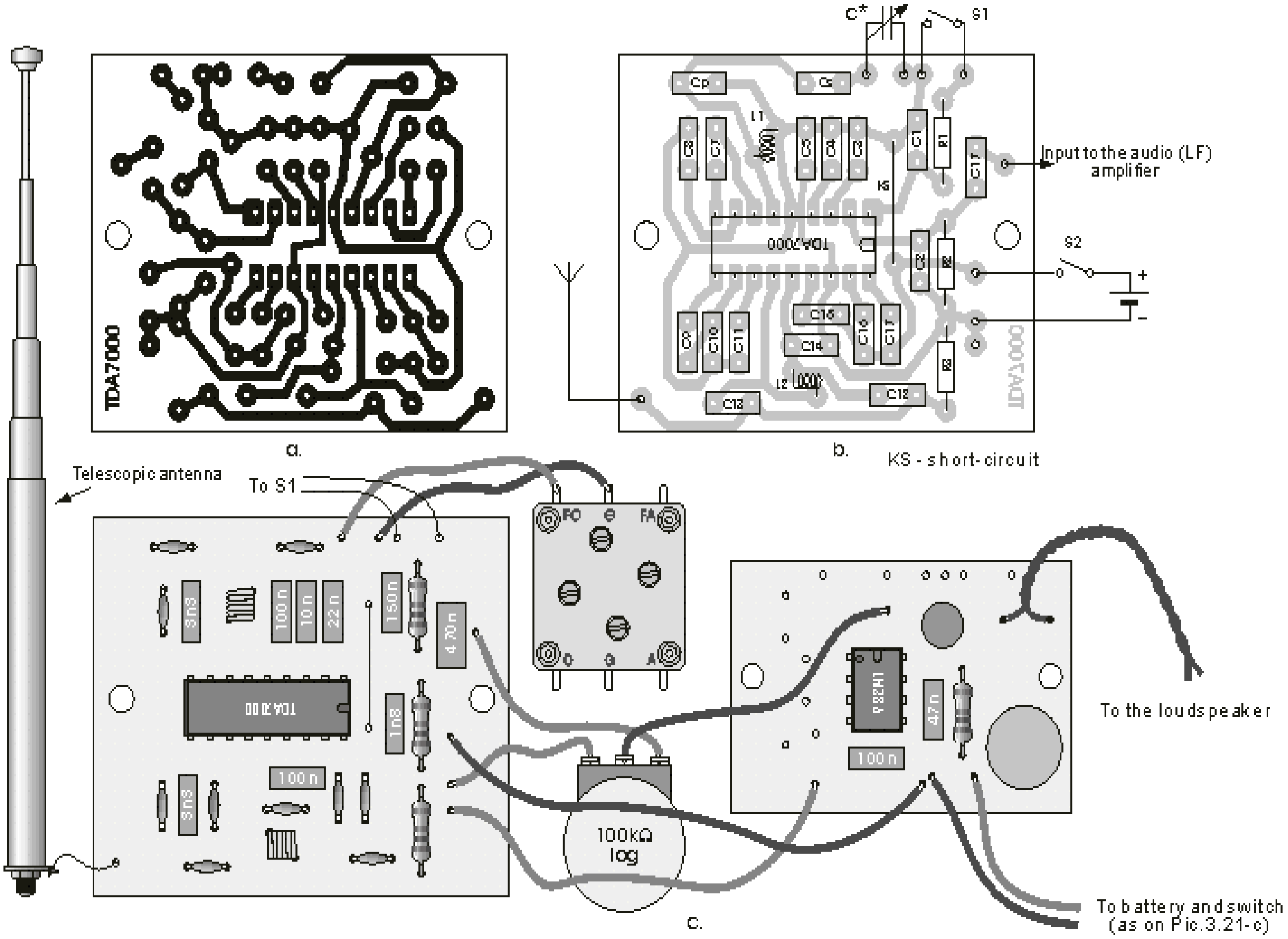
C*
S1
Cp
Cs
L1
KS
Input to the audio (LF) amplifier
TDA7000
S2
+
-
C16
C14
L2
C13
C12
a.
b.
KS - short-circuit
Telescopic antenna
To S1
FO
FA
A
100kΩ log
To the loudspeaker
LM386
To battery and switch (as on Pic.3.21-c)
c.

Pic.4.8 Practical realization of the receiver from Pic.4.7: a-PCB, b-component layout, c-complete device

capacitance of Cs should be reduced (to 15 pF, 10 pF etc.), or it should be short-circuited. You can also try compressing or stretching the L1 coil, etc. The setup of the oscillatory circuit is completed when with C=Cmax some station that operates on app. 88 MHz can be heard, and with C=Cmin the one that works on 108 MHz.

The input circuit setup (it is connected between pins 13 and 14), is performed by tuning the receiver to some mid-range station (about 98 MHz). Then, the best possible reception is searched, by changing capacitances C13 and C12 and inductance L2.

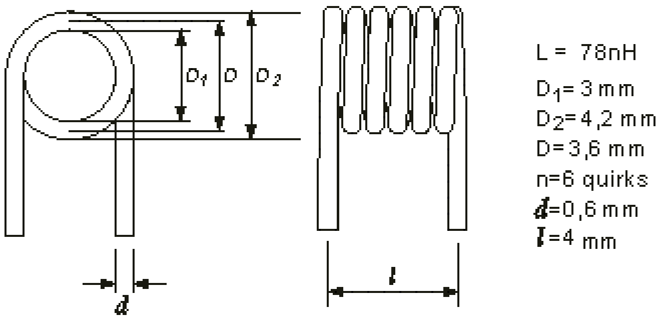
D1
D
D2
d
l
L = 78nH
D1= 3 mm
D2= 4,2 mm
D= 3,6 mm
n=6 quirks
d=0,6 mm
l=4 mm

Pic.4.9. Coils for the receiver from Pic.4.7

4.2.2. FM Receiver with TDA7088T IC

The receiver described in the last project has two IC’s, one variable capacitor, two small coils and fairly small few other components, so it can be put into some small box, by carefully placing the components. Further miniaturization can be accomplished by using the SMD components. These are the resistors, capacitors, transistors, IC’s and other components, whose dimensions are significantly smaller than these of “classical” components. They are mounted on the copper side of the PCB, therefore it isn’t necessary to drill the holes on the board. TDA7088T is also an SMD component. Its drawing is shown on Pic.4.10.
This IC is the successor of the famous TDA7000, i.e. it is an improved model of TDA7000, that allows to implement both monophonic and stereophonic FM receiver. The basic features of TDA7088T are given in the following table.

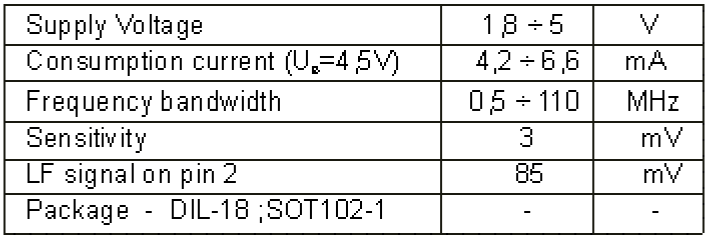

Supply Voltage	1,8 ÷ 5	V
Consumption current (U_B=4,5V)	4,2 ÷ 6,6	mA
Frequency bandwidth	0,5 ÷ 110	MHz
Sensitivity	3	mV
LF signal on pin 2	85	mV
Package - DIL-18 ;SOT102-1	-	-

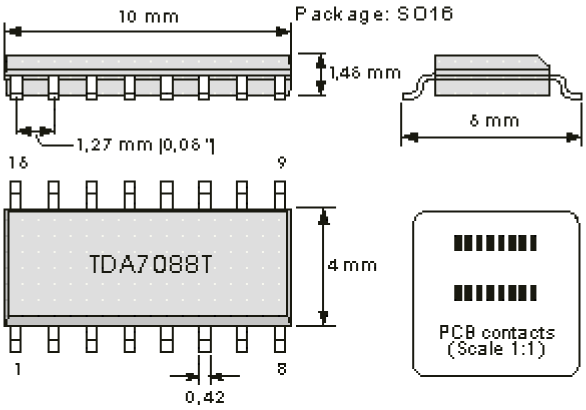

Pic.4.10. TDA7088T IC Layout

The electronic diagram of the HF part of the monophonic FM receiver made with TDA7088T IC is given on Pic.4.11. The IC contains all the parts of the classic superheterodyne receiver: the local oscillator, IF amplifier and FM detector, but also some other circuits that extend the possibilities and improve the features of this IC.
As far as practical use is concerned, the most significant novelty is the auto-tuning circuitry. No variable capacitor is necessary for tuning, as it was in all the previous projects, the BB910 varicap diode is used instead. Its capacitance is being changed by varying the DC voltage supplied to its anode over the 5k6 resistor. This is how the tuning is performed:
When the user press and releases the pushbutton marked with “RUN”, the positive voltage impulse is released to the S(et) input of the SEARCH TUNING circuit. The 100 nF capacitor then starts chargingl and the voltage on the pin 16 increases. This voltage is then

transferred, over the 5k6, to the anode of the BB910, causing its capacitance to decrease, which increases the frequency of the local oscillator (VCO). The VCO voltage is led into the mixer (MIXER) which also receives, over pin 11, the signals of all the other FM stations. The mixer outputs the FM signals whose frequencies are equal to the differences of the oscillator and the original station frequency. The only signal that can reach the demodulator (FM detector) is the one whose carrier frequency is equal to the inter-frequency, i.e. fm=73 kHz (selectivity is being accomplished by two active filters whose components are the capacitors connected to pins 6, 7, 8, 9 and 10). Therefore, the oscillator frequency increases until it gets the condition fO-fS=73 kHz is accomplished. When this happens, the charging of the capacitor is halted by the command that is sent into the SEARCH TUNING circuit by two detectors (diode-blocks) located in the MUTE circuit. The AFC (Automatic Frequency Control) circuit now gets its role and prevents the voltage on pin 16 to be changed, until the RUN button is pushed again (this voltage can vary from 0 V til 1.8 V during the tuning).

When the RESET button is pushed, the 100 nF capacitor is discharged, the voltage on pin 16 drops down to zero, and the receiver is set to the low end of the reception bandwidth, i.e. 88 MHz.

Let us get back to the mixer. On its output, the 73 kHz FM signal is obtained, and it is modulated by the programme of the first station that is found after the RUN button is pushed. This signal then passes the active filters, gets amplified in the IF amplifier (IF LIMITER) and passed onto the input of the demodulator. By connecting the demodulator exit, over the LOOP FILTER, the adder (+) and resistor, to the VCO, the so-called FFL (Frequency Feedback Loop) circuit is accomplished, reducing the deviations of the signal being received from ±75 kHz to ±15 kHz.

The LF (AF) signal is led from the demodulator, over the LOOP FILTER stage, the invertor (-1) and MUTE circuit onto the pin 2. The detectors (diode-blocks) control the operation of the MUTE circuit, preventing the LF (AF) signal to reach the output pin (2) until the tuning on the station that creates the signal in the antenna that is strong enough for quality reception is obtained.

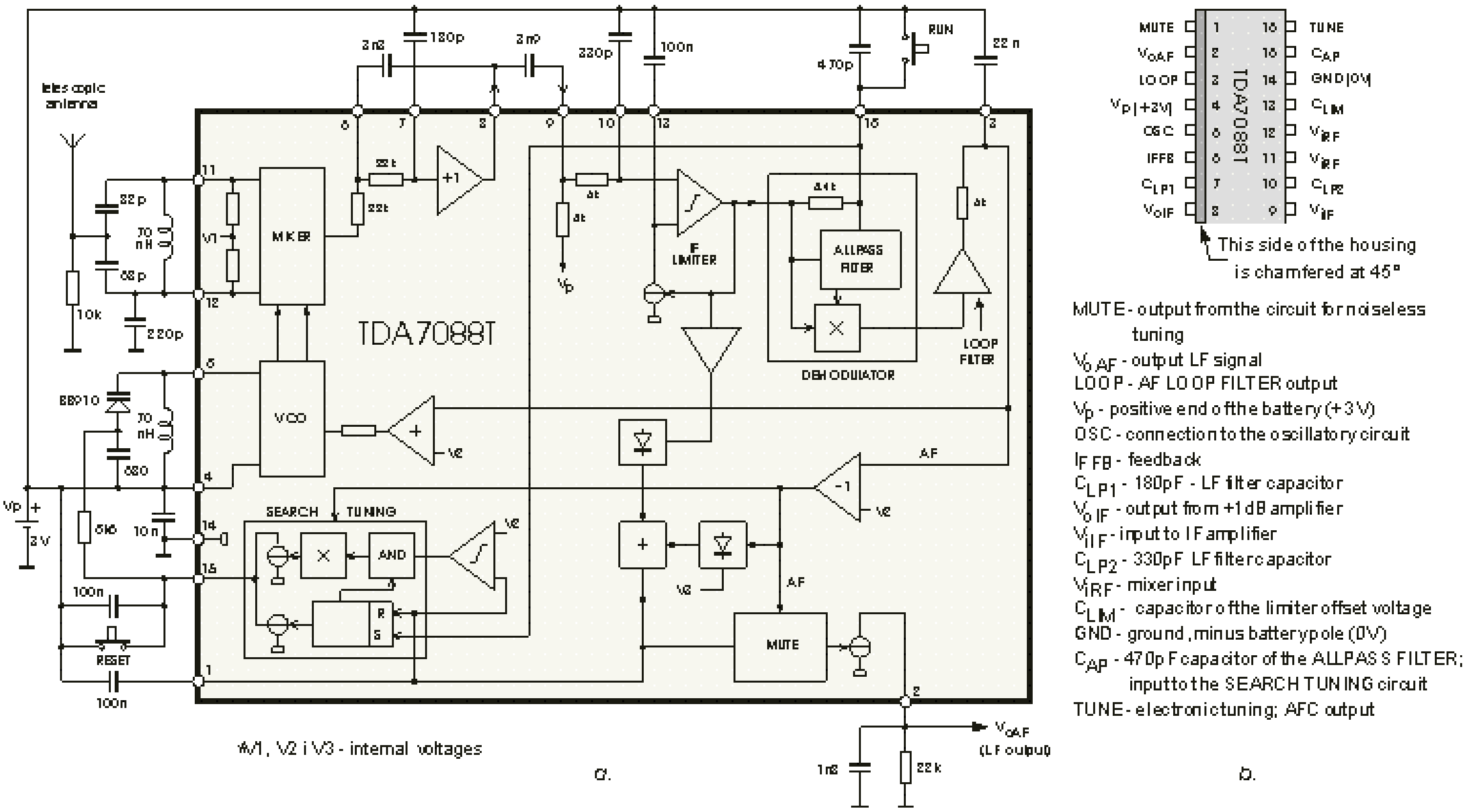
telescopic antenna
22p
70 nH
68p
10k
220p
11
12
V1
MIXER
3n3
180p
3n9
22k
22k
+1
6
7
8
9
10
13
5k
5k
Vp
220p
100n
470p
RUN
22n
16
3
IF LIMITER
44k
ALLPASS FILTER
DEMODULATOR
5k
LOOP FILTER
TDA7088T
5
BB910
70 nH
680
VCO
+
V2
4
Vp +
3V
5k6
10n
14
15
SEARCH
TUNING
AND
R
S
V2
100n
RESET
1
100n
AF
-1
V2
+
V2
AF
MUTE
2
V_{oAF}
(LF output)
1n8
22k
a.
MUTE
V_{oAF}
LOOP
V_P (+3V)
OSC
IFFB
C_{LP1}
V_{oIF}
1
2
3
4
5
6
7
8
16
15
14
13
12
11
10
9
TUNE
C_{AP}
GND (0V)
C_{LIM}
V_{iRF}
V_{iRF}
C_{LP2}
V_{iIF}
TDA7088T
This side of the housing is chamfered at 45°
b.

*V1, V2 i V3 - internal voltages

Pic.4.11. TDA7088T: a- block diagram, b-pin layout

MUTE - output from the circuit for noiseless tuning

$V_{o\,AF}$ - output LF signal

LOOP - AF LOOP FILTER output

V_P - positive end of the battery (+3V)

OSC - connection to the oscillatory circuit

$I_{F\,FB}$ - feedback

C_{LP1} - 180pF - LF filter capacitor

$V_{o\,IF}$ - output from +1dB amplifier

V_{iIF} - input to IF amplifier

C_{LP2} - 330pF LF filter capacitor

V_{iRF} - mixer input

C_{LIM} - capacitor of the limiter offset voltage

GND - ground, minus battery pole (0V)

C_{AP} - 470pF capacitor of the ALLPASS FILTER; input to the SEARCH TUNING circuit

TUNE - electronic tuning; AFC output

4.2.2.1 Mini FM Receiver

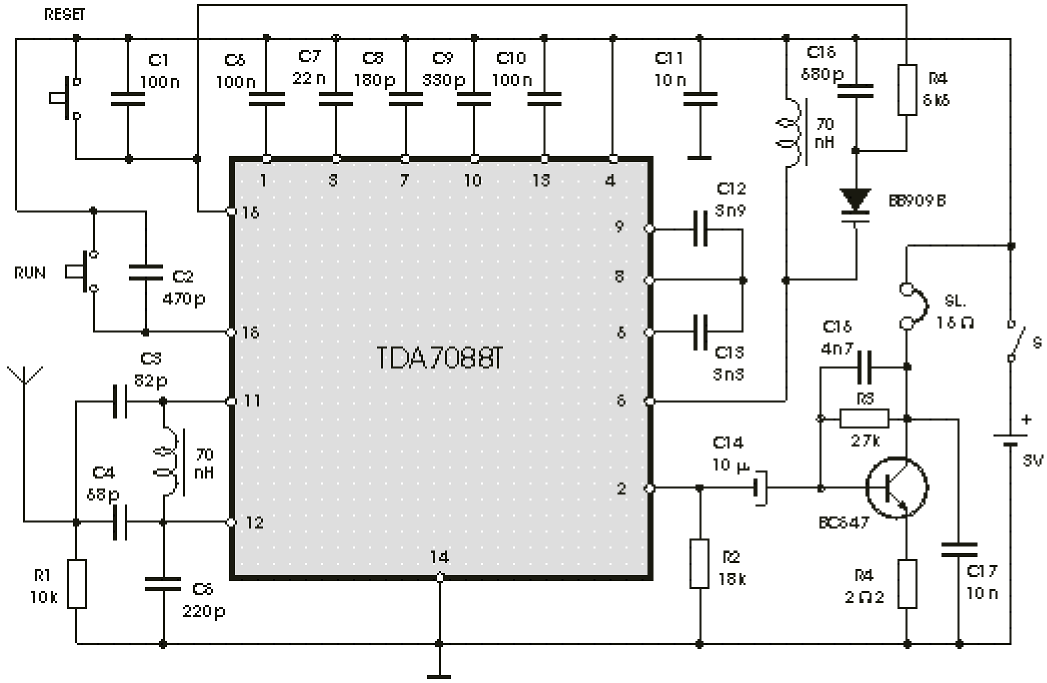
RESET
C1
100n
C5
100n
C7
22n
C8
180p
C9
330p
C10
100n
C11
10n
C16
680p
R4
6k8
70
nH
BB909B
1
3
7
10
13
4
15
C12
3n9
9
8
5
C13
3n3
6
RUN
C2
470p
16
TDA7088T
C3
82p
11
70
nH
C4
68p
12
R1
10k
C6
220p
14
2
C14
10 µ
R2
18k
C15
4n7
R3
27k
BC547
R4
2Ω2
C17
10n
SL.
16Ω
S
+
3V

Pic.4.12. Electronic diagram of the monophonic FM radio receiver with electronic tuning

The electronic diagram of the monophonic FM receiver made with TDA7088T is shown on Pic.4.12. If built with SMD components it can be placed in a matchbox, altogether with two button-type batteries. The operating principle of this device is given in the previous chapter. The only thing new is a very simple audio amplifier made with BC547 transistor, which is loaded by cheap 16-Ohm headphones. The telescopic antenna is used, as on Pic.4.8.
Small mishap of this receiver is that it has no indication of station tuning. This problem can be solved by adding a small voltmeter in parallel to the BB909, whose scale is graduated in MHz, as described in the Appendix. This solution is not appropriate for the miniature receiver, since the voltmeter that has the scale that is big enough takes too much space. It is in this case better using a manual tuning instead of automatic. Such solution is given on Pic.4.13.
The tuning is done via the variable capacitor C with numbers written on its button, similar to that on Pic.3.11. It is most simple to use numbers from 1 to 10. The variable capacitor is like the one on Pic.4.8. Some experimenting is to be done with capacitances of Cx and Cy, in order to cover the entire reception bandwidth, from 88 till 108 MHz.

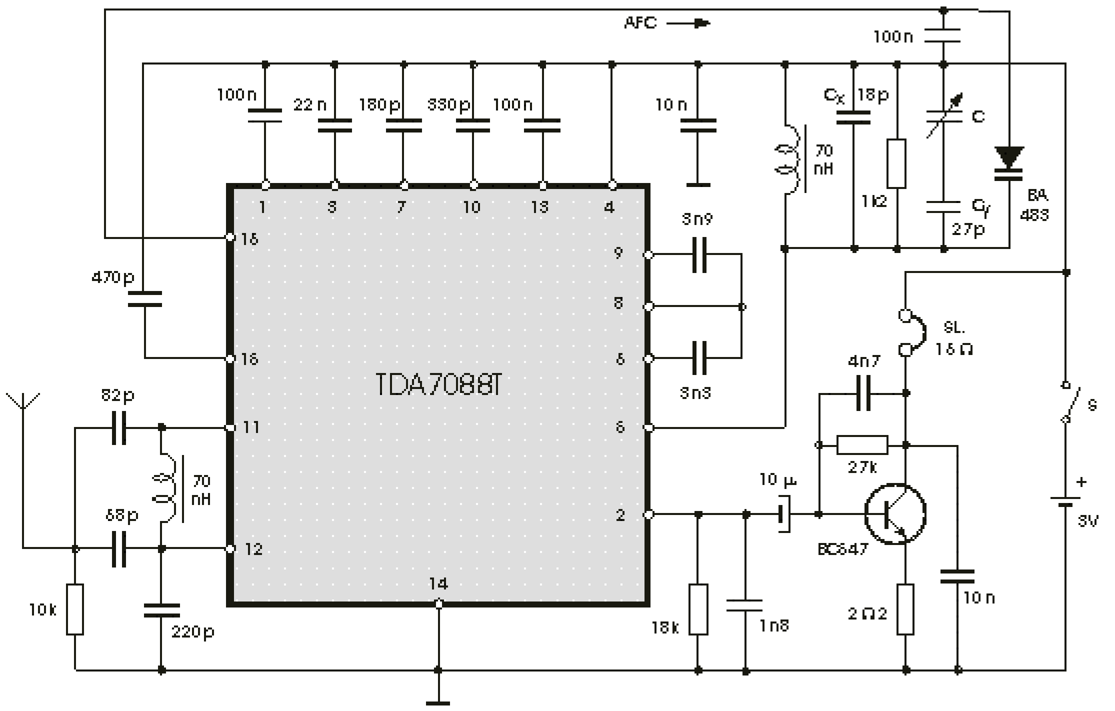
AFC
100n
100n
22n
180p
330p
100n
10n
Cx
18p
70 nH
1k2
C
Cf
27p
BA 483
1
3
7
10
13
4
15
16
11
12
9
8
5
6
2
14
3n9
3n3
TDA7088T
470p
82p
68p
70 nH
10k
220p
SL. 16 Ω
4n7
27k
10 µ
BC547
2Ω2
10n
18k
1n8
S
+
3V

Pic.4.13. Electronic diagram of the monophonic FM radio receiver with manual tuning

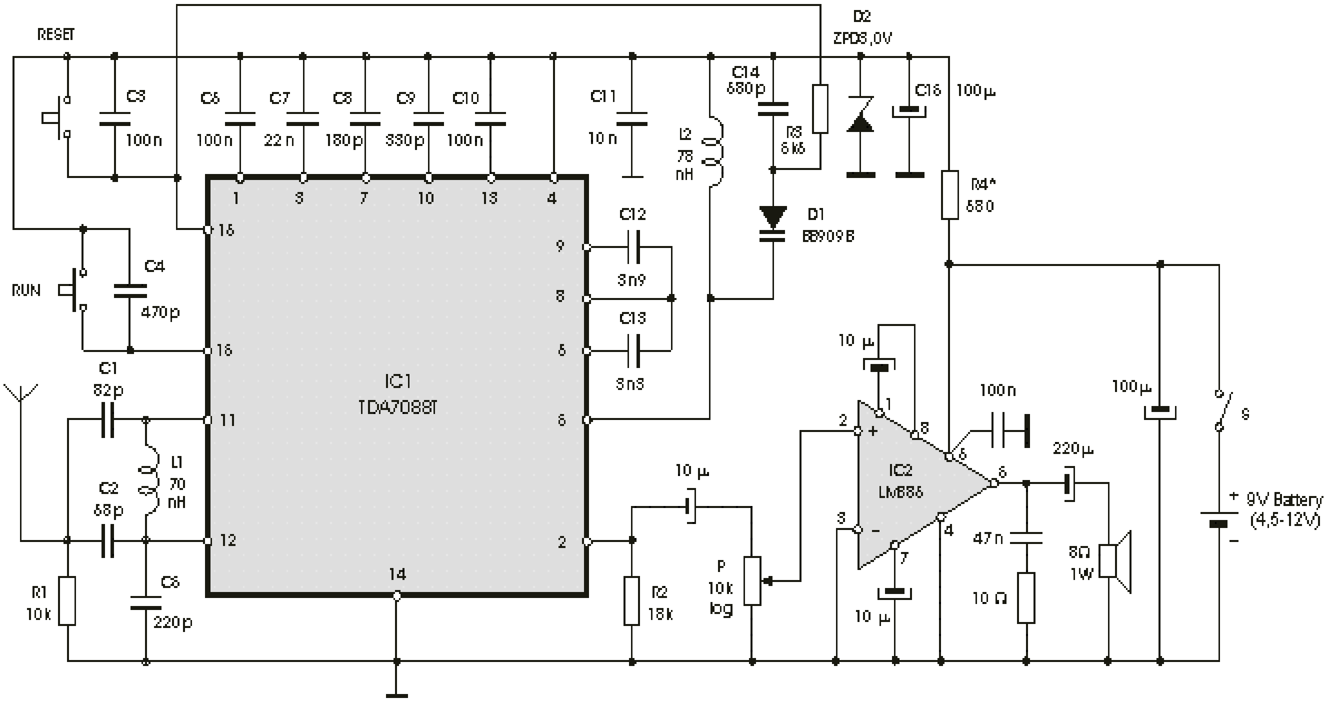
RESET
C3
100n
C5
100n
C7
22n
C8
180p
C9
330p
C10
100n
C11
10n
C14
680p
L2
78 nH
R3
5k6
D2
ZPD3,0V
C16
100µ
R4*
680
D1
BB909B
C12
3n9
C13
3n3
RUN
C4
470p
C1
82p
C2
68p
L1
70 nH
R1
10k
C6
220p
IC1
TDA7088T
1
3
7
10
13
4
15
16
11
12
9
8
5
6
2
14
10 µ
R2
18k
P
10k
log
IC2
LM386
10 µ
10 µ
100n
100µ
220µ
47n
10 Ω
8Ω
1W
S
+ 9V Battery
(4,5-12V)
-

Pic.4.14. Electronic diagram of the monophonic FM radio receiver with TDA7088T and LM386

The AFC (Automatic Frequency Control) of the local oscillator is accomplished with BA483 diode, obtaining that station's position on scale does not "walk" over the scale.
The complete radio receiver should still have a loudspeaker. Electronic diagram of such

receiver made with TDA7088T is given on Pic.4.14. As one can see, that is a receiver from Pic.4.12 with an audio receiver made with LM386 IC.
Maximum value of the DC supply voltage for the TDA7088T is 5V, therefore if using a 4.5 V battery the LM386 will work with reduced output power, the D2 diode and C15 capacitor should be omitted, and R4 should be short-circuited.

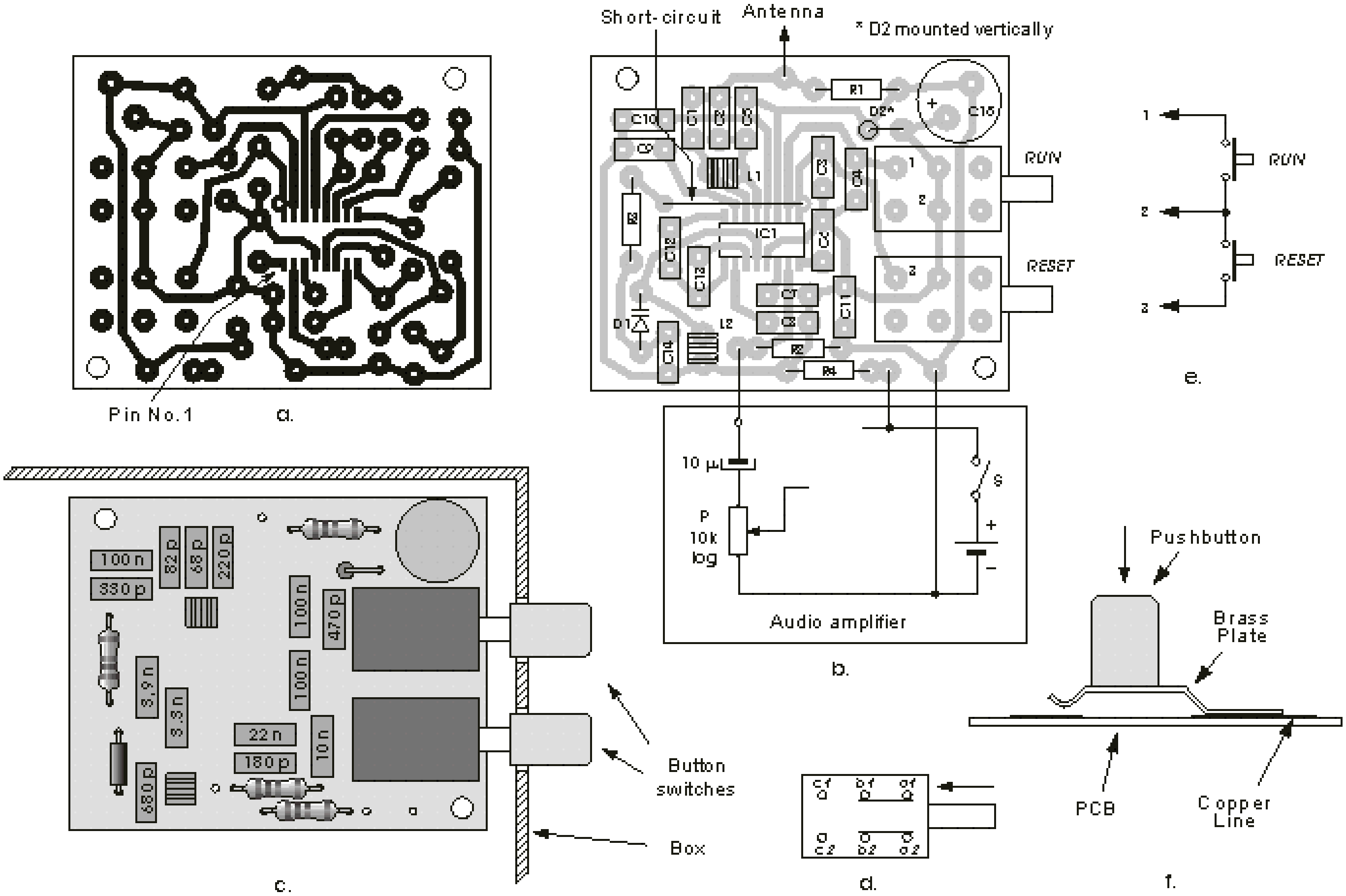
Pin No.1
a.
Short-circuit
Antenna
* D2 mounted vertically
C10
C9
C1
C2
C3
L1
R1
D2*
C15
RUN
RESET
R3
C6
C4
C5
IC1
C12
C13
C7
C8
C11
D1
L2
R2
R4
C14
10 µ
P
10k
log
S
Audio amplifier
b.
1
2
3
RUN
RESET
e.
100 n
330 p
82 p
68 p
220 p
100 n
470 p
100 n
3,9 n
3,3 n
22 n
180 p
10 n
680 p
Button switches
Box
c.
c1 b1 a1
a2 b2 c2
d.
Pushbutton
Brass Plate
PCB
Copper Line
f.

Pic.4.15. Practical realization of the receiver from Pic.4.14.: a- PCB copper side, b- PCB component side, c- board layout with components, d- switch for PCB mounting, e- connecting the switches that are mounted on the device box

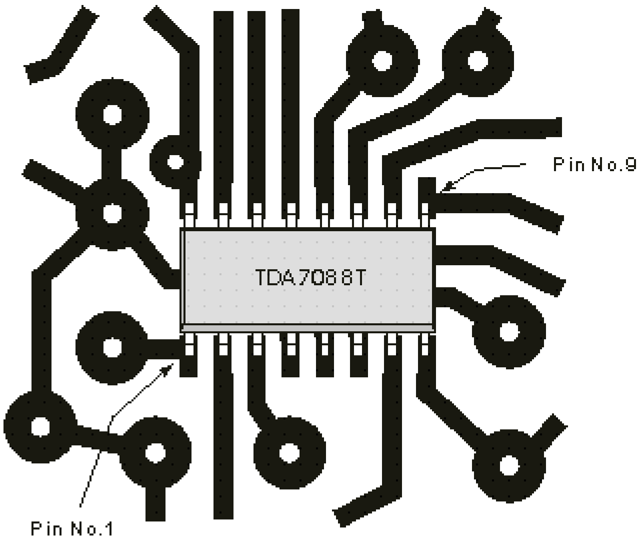

Pic.4.16. Position of the TDA7088T on the PCB

If higher voltage battery is used, the voltage stabilizer, comprised by the aforementioned components, has to be activated. D2 is a Zener diode with 3 V Zener voltage. The optimum value of R4 is found experimentally: in order to make the power consumption as low as possible it should have the resistance as big as possible, while simultaneously keeping the voltage on Pin 4 about 3 V and the device working well within the entire reception bandwidth (One should start with, say, R4=1.5 kOhm, and if the receiver operates well bigger resistance should be tried out, and if not smaller one, until the optimum value is found).

Pic.4.15. shows the PCB for the HF part of the receiver with TDA7088T, that is realized with ordinary components, instead of the SMD’s. Pic.4.15-a shows the board layout from the soldering (copper) side. All the components apart from TDA7088T are mounted on the opposite side of the board, their pins are put through the holes and soldered through the holes. The TDA is soldered on the copper side, directly onto the copper contacts. That is why it is being drawn in dashed line on pic.4.15-b, where the board layout on the component side is given.

* Pic.4.16-a shows 3x enlarged picture of the IC and the surrounding lines. The soldering procedure for SMD is as follows:
A thin tin layer is applied on the copper contacts where IC legs are to be soldered to. The firs legs to be soldered are the diagonally opposite ones, in this case No.1 and 9. A small cushion-shaped amount of tin (not profuse) is applied on the contacts where these pins are to be soldered (pic.4.17-a). The IC is placed in its position, with all the pins properly laid. Pin No.1 is pressed against the tin pillow with a top of a bodkin, with iron head simultaneously touching both the tin and the pin end. The tin gets melted, and the pin lies down on its place with the aid of the bodkin, and gets soldered.
It is now time to check out the positioning of the chip. If it needs to be corrected, the tin surrounding the pin No.1 is melted with iron tip and the chip position is quickly and carefully adjusted, in order not to overheat the pin. Soldering the pin No.9 is shown on Pic.4.17-b. First, the iron tip is simultaneously put on the top of the leg and the copper below it, so that

both of them are heated. After app. half a second, the iron is slightly removed from the leg but remains on the copper contact. The tip of the tinol wire is then approached to touch the iron, the pin top and the copper contact at the same time. The wire gets melted and adheres to the copper and the pin, so it has to be constantly moved downwards. When enough tin is applied, the tinol wire is removed first, then the iron, and the pin No.9 is soldered. Once again you have to check whether all the pins are properly placed, and then they too are soldered as it was just described. The solders are OK if they look app. as on Pic.4.17-c.

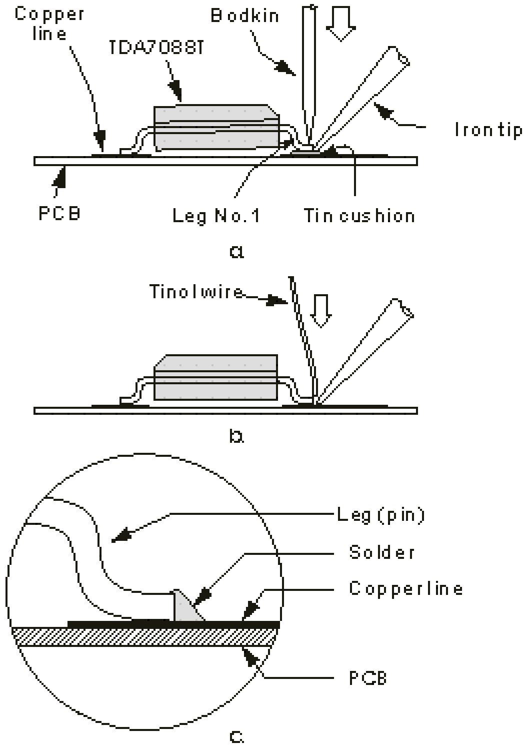

Pic.4.17. Soldering of the leg of TDA7088T: a-soldering the leg No.1, b-soldering the leg No.9, c-proper solder look

* Pic.4.15-b contains the PCB component side layout. The pushbuttons we used here are Siemens, type BO2AMAP-2. The common housing contains, as one can see, two button switches, one of which is being used by this device. Any other pushbutton switches can also be used. In that case small modifications on the PCB lines would probably be necessary. The board is mounted fairly close to the box edge, so that the switch shafts are passing through the panel, and that the buttons can be mounted on the outside. The panel-mount switches can also be used, in which case they are connected to the board by wires (pic.4.15-e).

* Any audio amplifier described so far can be used, e.g. the one with LM386, as on pic.4.8.

* Instead of the antenna, a 20 cm piece of wire can also be utilized.

4.2.2.2. Stereophonic Receiver Built with TDA7088T

Stereophonic radio broadcast is performed in the ultra short waveband, from 88 MHz till 108 MHz. All radio transmitters operating in this range are stereophonic, but their signal is designed so that monophonic receivers can also read it, performing the compatibility. The readers that wish to get acquainted in more details with the stereophonic broadcast basics can refer to the “Radio Receivers” textbook, for the IV grade of the Electrotechnical Highschool.

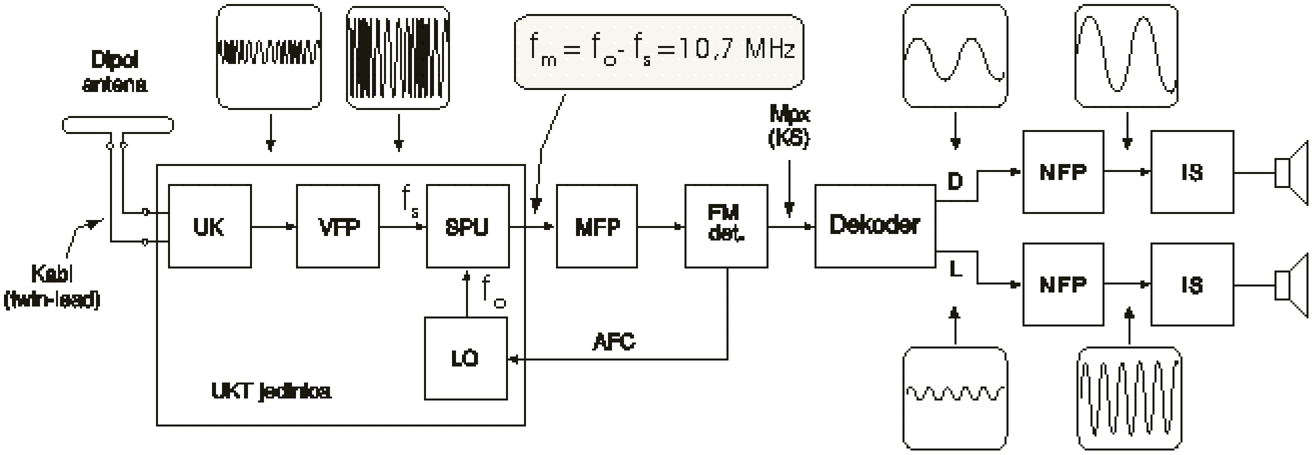

Pic.4.18. Block diagram of a stereophonic FM receiver

Making an introduction to this part, a operating principle of the stereophonic radio receiver shall be considered, its block diagram shown on pic.4.18. Comparing this diagram with the one of the monophonic receiver given on pic.4.6, one may notice that they are identical, up to the block called "The Decoder". It means that, as already described, exiting the FM detector the LF signal is obtained, i.e. the information that was used to perform the frequency modulation in the transmitter. However, this is not an ordinary LF signal, but the one, called the "composed" (KS) or "multiplexed" (Mpx) signal. Besides the full-scale LF signal used by the monophonic receiver, it also contains the so-called auxiliary signal which allows the separation of left (L) and right (R) channels in the stereophonic receiver. E.g. if a direct broadcast of some band music is performed, the left part of performers is being recorded with one microphone (the signal marked as L), whilst the right side is recorded with the other one (it’s a R signal). These two

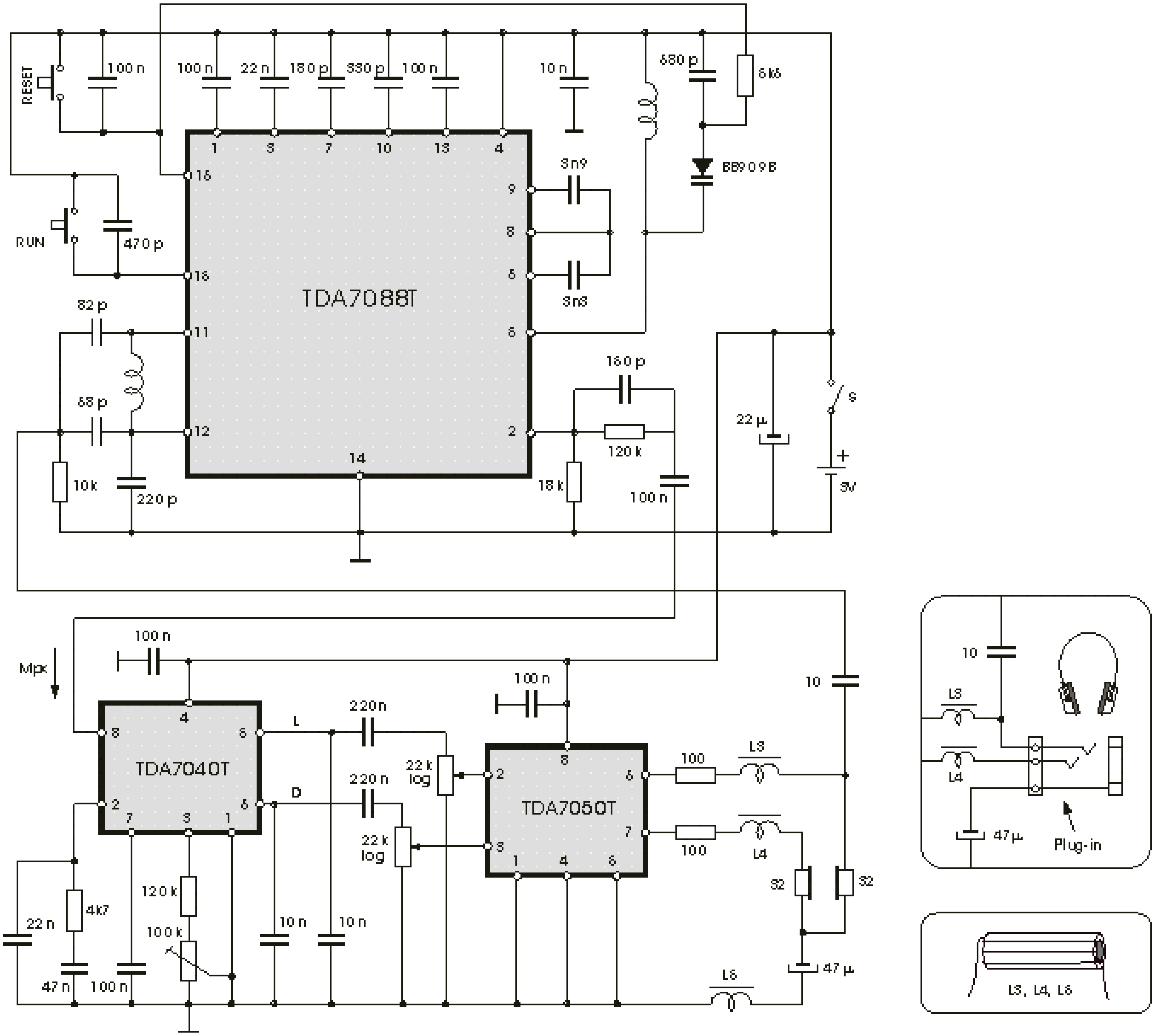
RESET
RUN
100 n
100 n
22 n
180 p
330 p
100 n
10 n
680 p
6k8
BB909B
470 p
3n9
3n3
TDA7088T
1
3
7
10
13
4
16
15
9
8
5
6
11
12
14
2
82 p
68 p
10k
220 p
160 p
120 k
18 k
100 n
22 µ
S
+
3V
Mpx
100 n
TDA7040T
L
D
220n
220 n
22 k log
22 k log
100 n
TDA7050T
100
100
L3
L4
S2
S2
10
47 µ
L5
22 n
4k7
47 n
100 n
120 k
100 k
10 n
10 n
10
L3
L4
47 µ
Plug-in
L3, L4, L5

Pic.4.19. Electronic diagram of the stereophonic radio receiver built with TDA7088T (HF deo), TDA7040T (decoder) i TDA7050T (audio-amplifier)

signals are being led in the FM transmitter in the stage called “the coder”. Exiting the coder we have the multiplexed signal Mpx which contains, in an indirect manner, both left (L) and right (R) signal. Frequency modulation of the transmitter is being performed with the Mpx signal. In the receiver, Mpx signal is obtained on the output of the FM Detector and is then led to the decoder. This stage plays a role complementary to the one of the coder in the transmitter, therefore two signals are exiting it, the L and D signal. They are being amplified over two identical audio amplifiers, then reproduced over two same loudspeakers. The listener can now hear the left half of the performers from the loudspeaker placed on its left, and the right half from the loudspeaker that is placed on its right. The performers that are situated in the middle of the orchestra are being equally reproduced from both loudspeakers, making an impression to the listener as if there's a third loudspeaker, located in the middle, between the left and right one. Based on all this, the listener has a picture about the layout of the performers in space, which significantly improves the total musical impression.

Electronic circuit of a portable stereophonic radio receiver with headphones reproduction, made with TDA7088T is shown on pic.4.19. It is a receiver whose practical realization was described in the previous project, with decoder with TDA7040T and dual audio amplifier with TDA7050T blocks added, the latter was discussed in PE5.

* L3, L4 and L5 are HF chokes that allow for the headphones cable to be used as a reception antenna. This is accomplished by connecting one of the headphones' contacts from the plug-in, over the 10 pF capacitor, to the point where, acc. to pic.4.14, the outside antenna is connected. The coils represent big resistance to the station signals, preventing them to "go to ground" over the 47 mF capacitor or over the TDA7050T output. Each coil has 3 quirks of the 0.2 mm CuL wire, threaded through ferrite pearls, as shown on detail in the right corner of the pic.4.19. If telescopic antenna is to be used, these coils should be omitted.

Chapter V Appendix

5.1. Making PCB's
5.2. Computer - Aided Radio Receiver Control
5.3. Receivers with NE612 IC
5.3.1. Synchrodyne AM Receiver
5.3.2. AM Receiver with Synchro - Detector
5.3.3. Input Circuits for the Receivers with NE612
5.4. Universal Audio Receiver
5.5. Additional Circuits
5.5.1. Fine Tuning
5.5.2. Electronic Tuning
5.5.3. Signal Suppressing of Local Radio Transmitter(s)
5.5.4. Dual Tuning
5.5.5. Separation of Stages - Preventing the Oscillation
5.6. The Boxes
5.7. Bimboard, Protoboard
5.8. Universal PCB Plates
5.9. A Modern Oldtimer

5.1. Making PCB's

Design and manufacturing of the PCB's has been explained in detail in the second issue of P.E. Here, we'll consider how to make a PCB whose drawing is already done. As an example, we'll take a drawing of the board of the receiver from pic.3.19, which measures 45 mm x 30 mm.

a. The PCB is being made of pertinax or vitroplast, i.e. a thin plate (about 1.5 mm) made of isolation material, which has a lean layer of copper put on one side. From the plate you buy in some electronic components' shop, a 45 mm x 30 mm piece should be cut. In amateur conditions, this means refracting. First, points A and B are marked on the non-copper side of the plate, acc. to pic.5.1-b. A ruler is placed over them and a groove is made by pressing with a screwdriver or a bodkin along it. Its depth should be about 0.5 mm (on picture it is shown in dashed line). When this is done, the plate is placed on the edge of the table, with copper facing downwards. With one hand the plate is pressed firmly to the table, and with the other, the piece that has to be refracted. And - it cracks just along the groove.

On the refracted piece, a new groove, measuring 30 mm from the edge, is made, and the procedure is repeated. In this way we finally have our 45 mm x 30 mm plate.

b. All of the copper has to be clean and shiny, since only in this case the etching and, later, soldering is performed quick and easy and well. If it seems to you that the copper you have just bought is clean enough, you're probably wrong. The plate must have spent some time in the shop, and the copper surface is certainly more-less corroded. The cleaning is most efficiently done with some abrasive powder (VIM or similar) which is otherwise used for cleaning of the cookers, bathrooms etc, but also the sodium bicarbonate, laundry detergent and even plain salt can well serve the purpose. Take a piece of cloth, wipe it with water, extract the water well and muss it to be ball-shaped. Dip it then in the powder, and scrub the copper until it "shines like the shiny sun". After that rinse the plate, and pay attention not to touch the copper with your fingers, since that will make it dirty again.

c. Put the plate, facing the copper up, beneath the sheet that contains the PCB layout, right under this drawing. In our example, that would be the one on the pic.5.2-a. With the pike of a bodkin the holes are made through the centers of all the contacts, and in the centers of two bigger holes that are placed sidewise, taking care not to move the plate. The bodkin has to be pressed firmly, in order to obtain good prods on the copper. When this is finished, the plate should look as on pic.5.2-b. i.e. it has to contain as much prods as there are contacts, plus two. If the drawing contains many contacts, the plate can be easily dislocated, and the procedure is to be done all over again. It is better practice then to make a copy of the picture, cut it out, and attach it to the plate with two pieces of scotch tape.

d. Drawing the contacts and lines on the plate is done with the acid-resistant marker paintstick. It can be recognized by its characteristic "alcohol smell", and is being sold in bookstores as a marker for "writing on glass". You can test it: write in the store (it will be later afterwards) something on the glass, piece of plastic and similar, wait for a couple of seconds, then try to wipe it out with your fingertip. If the paint remains - the marker is OK. Nevertheless, this test isn't 100% certain, it is much better to buy the marker in the electronic components store (you have to accent to the salesman that you need a marker for drawing lines on PCB's). With the tip of the marker draw a circle around every prod (except those two that are for bigger holes), measuring 2-3 mm in diameter. Move the marker slowly, in order to leave a thick layer of paint on the plate. Take care to leave a small copper isle around every hole. Then,

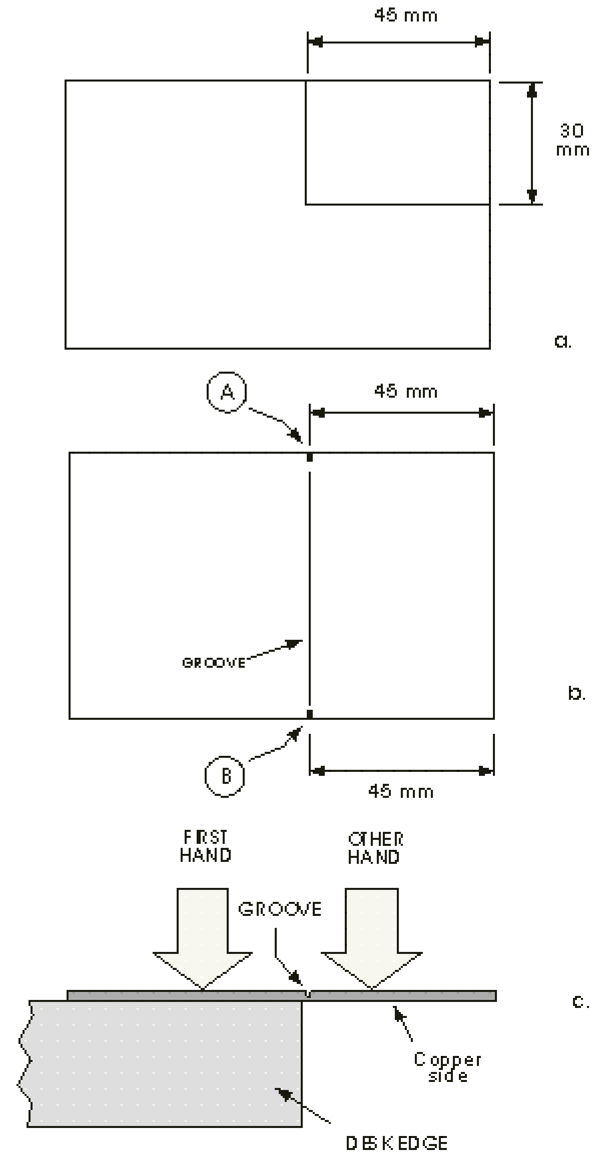

Pic.5.1. "Cutting" of the plate: a-pertinax plate, b-marking, c-cracking

you should, carefully and slowly, draw all the lines, by looking at the pic.5.2-a. They do not need to have the same shape as on the picture, especially they don't have to be that "chamfered". Line thickness should be about 1 mm, but that either is not obligatory, they can be somewhat thinner or a lot thicker (where applicable). The important thing is not to connect the nearby lines or contacts during the drawing, i.e. not to make junctions that do not exist on the drawing. If that happens anyhow, remove the paint surplus with a razor or a small, sharp screwdriver. Pic.5.2-c shows the beginning of drawing, several contacts and 3 lines are drawn. The drawing is finished when you have a pic.5.2-a on the copper foil.
e. Next step is etching, i.e. removing the copper that is not covered with marker paint. For this purpose, a mixture of hydrochloric acid (HCl), hydrogen peroxide (H_2O_2) and water (H_2O). Pure hydrochloric acid is not used, but its 35% solution, that is being sold as a household cleaning agent. Hydrogen peroxide is being sold in drug stores and cosmetic

stores. It is being sold as 30% solution, or even more diluted, 8-12%.
Hydrochloric acid and hydrogen peroxide are very aggressive media, especially for the eyes and skin, therefore care should be exercised when working with them. It would be the best for you to work with them in the bathroom, or some other place close to the running water supply. If some of these liquids spills on your skin, metal tool or clothing, wash them down with water immediately.

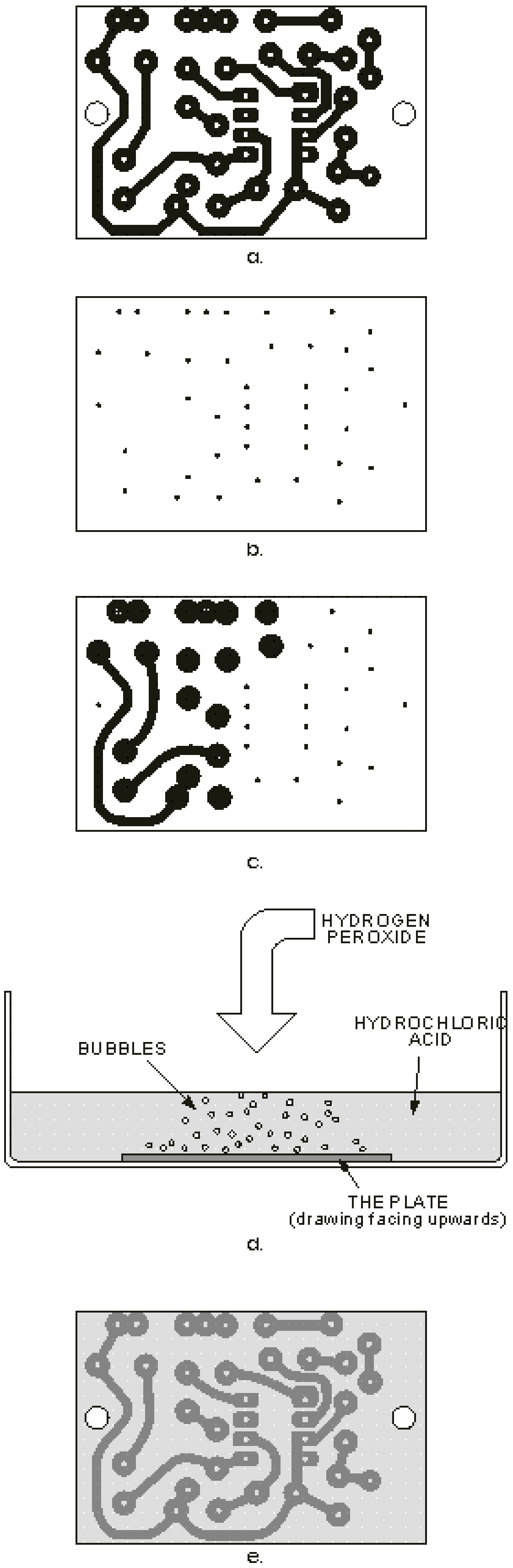
a.
b.
c.
d.
e.
HYDROGEN PEROXIDE
HYDROCHLORIC ACID
BUBBLES
THE PLATE (drawing facing upwards)

Pic.5.2. Making a PCB: a-drawing, b-plate with bodkin prods, c-drawing contacts and lines, d-etching in the hydrochloric acid while adding the hydrogen peroxide, e-finished plate

The etching mixture is being made directly before the etching, and is CERTAINLY being disposed of, right after the process. The plate is put at the bottom of a plastic, glass or porcelain dish, with copper facing upwards, and the acid is poured, in quantity enough to fully cover the plate (pic.5.2-d). Hydrogen peroxide is then added, being poured from the container directly over the plate. The amount of peroxide depends on its concentration, as well as on the concentration of the acid. So, put some peroxide, raise a little left end of the dish, then the right one, to allow a liquids to mix, and observe the plate. The mixture is transparent, and if the copper starts changing the colour after a dozen seconds - the etching has begun. During this process, the bubbles are formed in the mixture, in the amount somewhat more than in a glass of mineral water. If too little bubbles are present, add some more peroxide. Be careful, however, not to exaggerate, since if you happen to have too much bubbles, the mixture is going to heat up and the marker paint can be destroyed. From time to time, you should raise one end of a plate with a pointed wooden or plastic stick, in order to remove the old liquid from its surface, and allow for fresh mixture to take its place.

Etching is finished when there is no more uncovered copper on the plate. Raise one plate end with the stick, wait for the liquid to decant, take a plate with a laundry clip and wash it thoroughly in a jet of running water. You can then remove the paint by scrubbing, as previously described, with a wet cloth dipped in some powder. The copper contacts and lines will emerge on the plate.

f. If you were careful enough to leave a uncovered isle of copper in the centre of every contact, after etching this will be a small cavity, in the centre of the contact. Through these cavities, that will guide your drill, a 1 mm holes should be drilled (it is better if the holes are 0.8 mm in diameter, but such drills are harder to find, and a lot easier to break). Two holes for the fixing screws are usually about 3 mm in diameter. While drilling, a piece of thicker plywood or some flat hard-wood plank (beech, oak) should be put beneath the plate, and not a piece of polystyrene or something similar. Do not press the drill too hard, since the tool will be plucking tiny pieces of plastic on the other side of the plate.

5.2. Computer-Aided Radio Receiver Control

In Book 7/8 of Practical ELECTRONICS methods for simple control of various electrical devices by computer were discussed. Practical realization of various interface circuits and sensors was described, by which the computer is being connected with the outside world, so that it can turn on/off the heater, light, fan, TV set or some other electrical device at the desired moment, based on data comprising temperature, light intensity, humidity etc. In this chapter we shall present, in short terms, one of the projects from the aforementioned book, that deals with simple computer-aided turning on/off of the radio receiver at the desired moment.

The radio is connected with the computer via parallel port, the one where the printer is also being plugged. It is a 25-pin female connector, called Sub D-25, which is given on Pic.5.3. With appropriate programme, logical ones (voltage +3.6 V) and zeroes (0 V) can be sent to the outputs marked as DI-1, DI-2,...DI-8, that are located on the legs marked with numbers 2 to 9. Electrical devices that are being controlled are connected to these outputs over the interface circuit that is given on Pic.5.4. Two connected devices are shown on the picture, their maximum number is 8.

Pic.5.5-a shows the electronic diagram of an extremely simple interface circuit, which can serve to connect to computer the radio receiver, that can then be switched on or off at certain time, with adequate programme. The low-power transistor BC547 can be used for the consumers that use the current from the battery that is not greater than 100 mA. In case you have bigger power consumers, some stronger transistor or two transistors in Darlington junction can be used instead of BC547. The transistor bas is connected to the pin No.9 of the mail Sub D-25 connector over the R1 resistor, while the emitter and minus pole of the battery are connected to the pin No.25, i.e. to the computer ground. As long as there's a logical zero on the DI-8 output, the base voltage is zero and the transistor is locked and no current runs through it, therefore also through the consumer. When a logical one emerges on the DI-8, the transistor goes to the saturation regime, the voltage between collector and emitter becomes very small (practically zero) and the transistor behaves as if the collector and emitter are short-circuited. In that way almost the entire battery voltage is available on the receiver power supply input.

The PCB layout is shown on Pic.5.5-b: The component side is in the upper part, and the soldering side in lower part of the picture.

Pic.5.5-c shows how a small transistor radio receiver, powered by a 9 V battery, is connected to the parallel port, over the interface from pic.5.5-a. The plate is connected with clamps via the cables A and B, and with the battery over C and D.

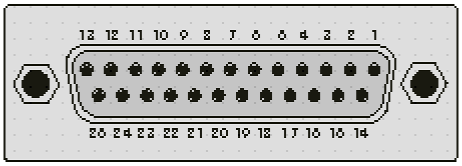
13 12 11 10 9 8 7 6 5 4 3 2 1
25 24 23 22 21 20 19 18 17 16 15 14

a - Sub-D-25 Connector
located on the computer housing back plate,
female type, front view

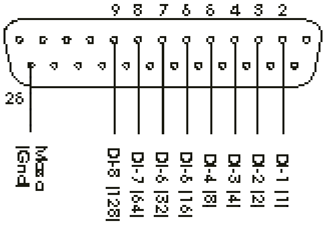
9 8 7 6 5 4 3 2
25
Masa (Gnd)
DI-8 (128)
DI-7 (64)
DI-6 (32)
DI-5 (16)
DI-4 (8)
DI-3 (4)
DI-2 (2)
DI-1 (1)

b- Pin marking on the Sub-D-25 Connector

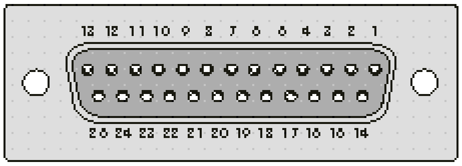
13 12 11 10 9 8 7 6 5 4 3 2 1
25 24 23 22 21 20 19 18 17 16 15 14

c-Sub-D-25 connector,
male type, watched from behind,
from the side where the
soldering legs are

Pic.5.3. Parallel Port

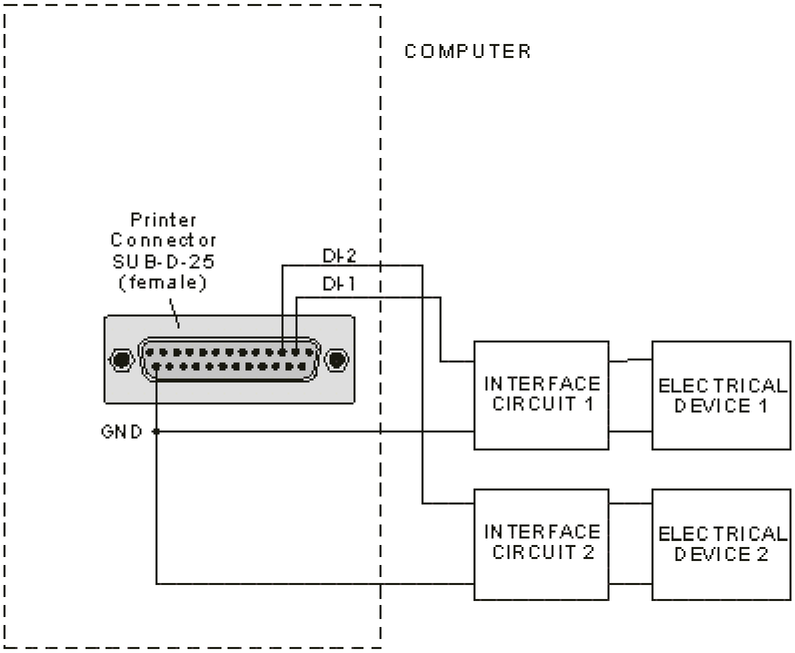
COMPUTER
Printer Connector SUB-D-25 (female)
D-2
D-1
GND
INTERFACE CIRCUIT 1
ELECTRICAL DEVICE 1
INTERFACE CIRCUIT 2
ELECTRICAL DEVICE 2

Pic.5.4. Connecting electrical devices onto the parallel port

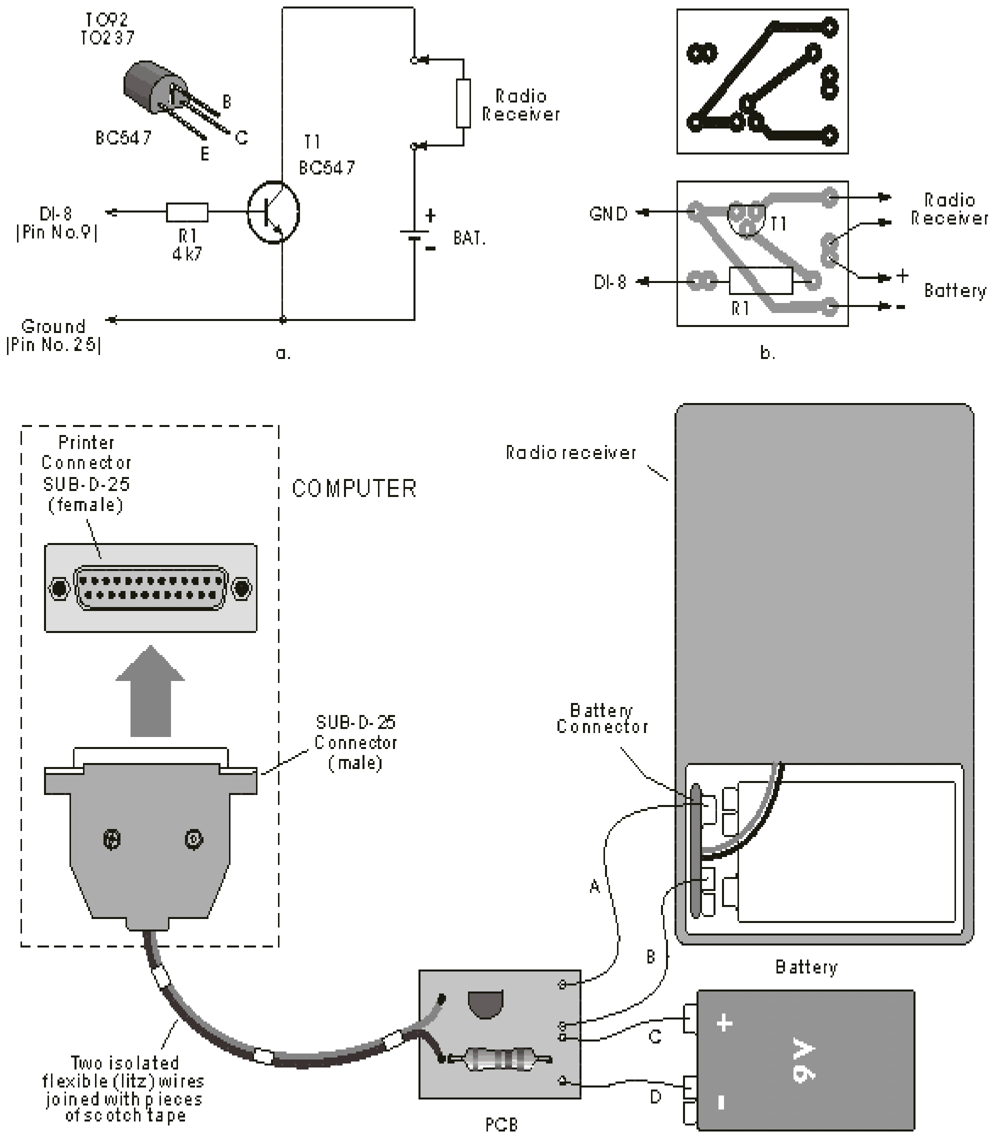
TO92
TO237
B
C
E
BC547
T1
BC547
Radio Receiver
+
-
BAT.
DI-8 (Pin No.9)
R1
4k7
Ground (Pin No. 25)
a.
GND
T1
DI-8
R1
Radio Receiver
+
-
Battery
b.
Printer Connector SUB-D-25 (female)
COMPUTER
Radio receiver
SUB-D-25 Connector (male)
Battery Connector
A
B
C
D
Battery
9V
+
-
Two isolated flexible (litz) wires joined with pieces of scotch tape
PCB

Pic.5.5. Connecting the Radio Receiver to the Parallel Port: a-Electronic diagram of the Interface Circuit, b-PCB, c-Connecting the Computer, Interface Circuit, Receiver and the Battery

5.3. NE612

5.3.1. Synchrodyne AM Receiver

If the author remembers well an article that he read in a professional magazine many years ago, the synchrodyne receiver is the ancestor of the superheterodyne receiver. Sometime at the beginning of the 20th century this device was called the Heterodyne receiver, and it was first constructed by Levvy. Armstrong improved it and gave new name to the new device, by adding the prefix SUPER to the old name.
The electronic diagram of this device is given on Pic.5.7. This receiver, as well as that on pic.4.2 has got the local oscillator with oscillatory circuit connected between pins 6 and 7. However, frequency of this oscillator is

not greater for the value of fm, but is in fact equal to the frequency of the station we wish to listen to: fm=fS. Because of that, the important design difference compared to the diagram from pic.4.2 is that on pic.5.7 capacitors Co and Cto are not used, but the capacitor C which is obtained by connecting the legs O and A, acc. to pic.3.7. Its capacitance can be changed from 12 pF till 218 pF, so that the oscillator frequency, in case of MW reception, goes between 500 kHz to 1500 kHz. The oscillator voltage is emanated in the mixer by the signals from all stations coming from the antenna. The result of emanation with signal of the station whose frequency is equal to the oscillator frequency is the LF signal (speech, music, Morse Code etc.) that serves for performing the modulation in the transmitter. This signal is obtained on pin No.4, from which it is then led, over the 1 ìF capacitor, to the volume potentiometer and audio amplifier. Products of mixing the oscillator voltage with other stations' signals are also obtained on that pin.

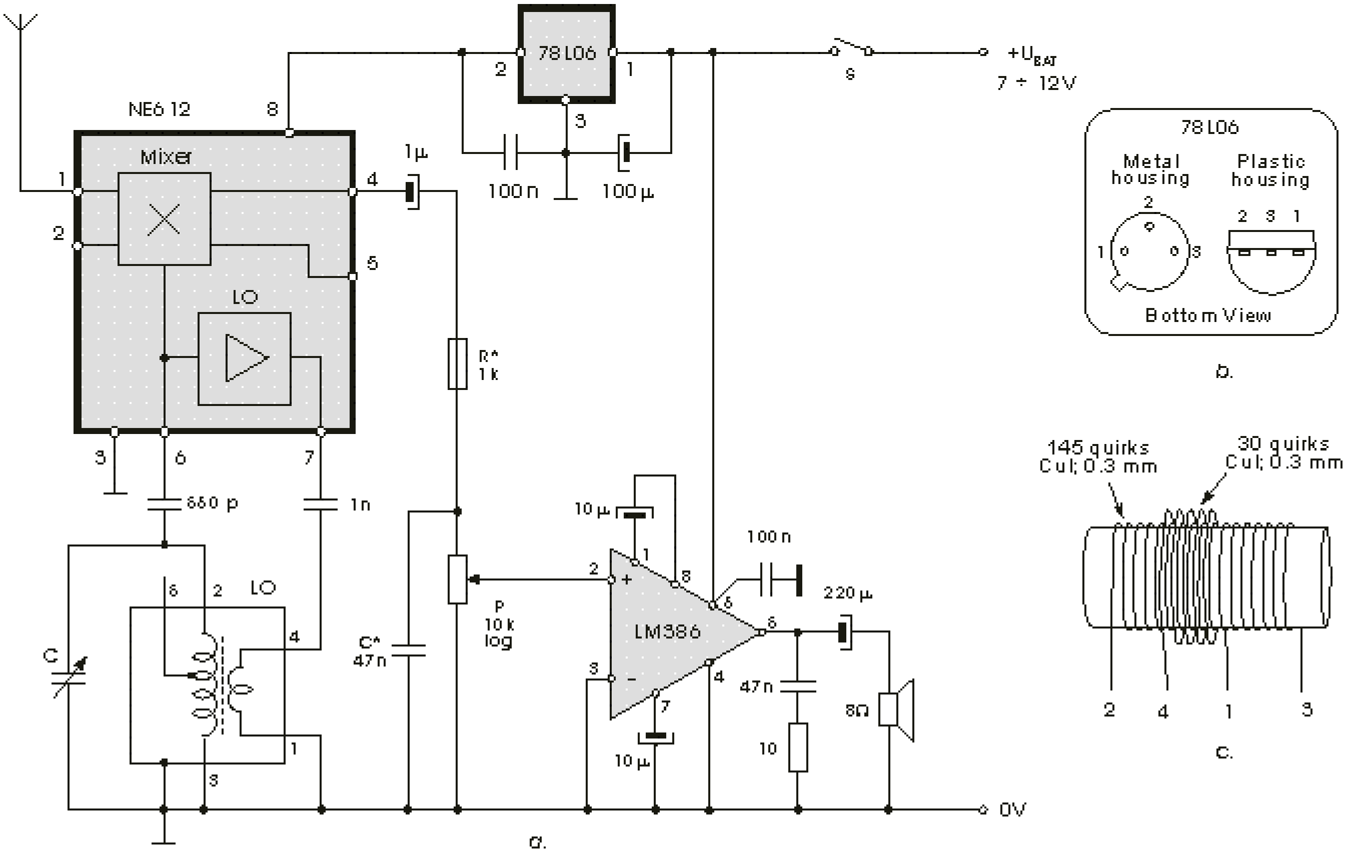
NE6 12
Mixer
LO
78L06
+U_BAT
7 ÷ 12V
S
100 n
100 µ
1µ
R* 1k
P 10 k log
C* 47n
660 p
1n
C
LO
LM386
10 µ
100 n
220 µ
47n
10
8Ω
0V
a.
78L06
Metal housing
Plastic housing
Bottom View
b.
145 quirks CuL 0.3 mm
30 quirks CuL 0.3 mm
c.

Pic.5.7. Synchrodyne AM Receiver: a-Electronic diagram, b-Pin layout for 78L06, c-Oscillator coil

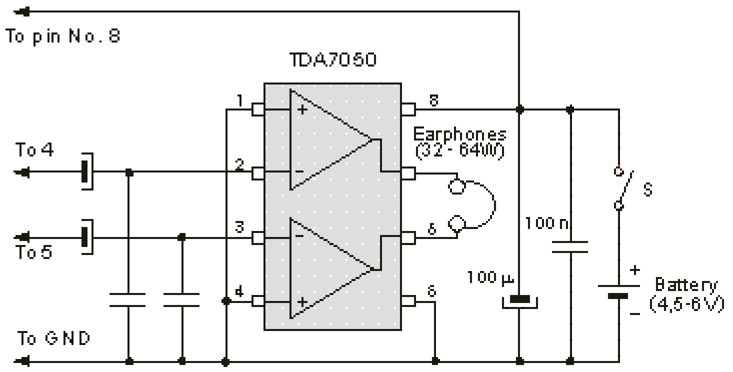

Pic.5.8. LF part of the Receiver from pic.5.7

They are being suppressed by the LF filter that comprises the R* resistor and C* capacitor. The device we were testing did not, however, contain R*. It is to be installed if some disturbances occur (whistling or similar), and its optimum value is to be found experimentally. If necessary, greater capacitance of C* is also to be tried out.

* As mentioned earlier, it is very important for the supply voltage of the NE612 to be stable. This values even more for the synchrodyne then the superheterodyne receiver. The voltage control is done by the stabilizer, made with 78L06 IC. It is being placed in the low-power transistor package, either metal (as for BC107) or plastic (as for BC547), and its maximum current is about 100 mA (pic.5.7-b). A simpler stabilizer, made with the Zener diode, can be used instead, as on pic.5.9.

* Instead of factory-made coil LO, the self-made one can also be used. The simplest solution is to use the one from pic.3.6, in which case the mid leg is not used. Over this coil, the feedback coil should be winded, acc. to pic.5.7-c (its ends are marked with 4 and 1). When connecting with capacitor C and pins 1 and 7 of NE612, care should be taken to join properly: coil ends 1 and 3 with ground, 2 with capacitors C and 560 pF, and 4 with 1 nF capacitor. It is, of course, possible to use smaller coil, wound on a smaller body, with more quirks of thinner wire. Its inductance should be about 350 mH, and the number of quirks required is to be found by testing. The feedback coil (4-1) has app. 3x fewer quirks than the oscillatory circuit coil (2-3).

* On the pin 5 of the NE612 the LF signal is also obtained. It is the same as the one on pin 4, but has a 180° phase shift compared to it (in simple words, while one signal increases, the other one decreases, and vice versa). That gives us the opportunity to use the dual audio amplifier in the LF part, that has two amplifiers, with inverting and non-inverting inputs. As shown on pic.5.8, the counter-phase LF signals from NE612 are led onto the same inputs. The output signal has 2x greater amplitude, therefore making the output power 4x greater than when only one input is used (as on pic.5.79).

5.3.2. AM Receiver with Synchro Detector

In previous project, the NE612 was in fact used as the AM signal detector. The LF signal exiting the mixer is product of the simultaneous (synchronous) action of the station signal and voltage from the local oscillator upon it. That is how the term “Synchro Detector” emerged. There’s also a possibility to use a station carrier instead of local oscillator’s voltage, so that the station signal gets beaten by itself, however strange this may sound.
Electronic diagram of one such device is given on pic.5.9.
The station signal, which the input circuit (C, L) is tuned at, is led to the regulating Gate of the BF960 MOSFET.

Under the effect of this voltage, the AC current that creates voltage drops on resistors R2 and R3 runs through the transistor. These two voltages, taken between the S and ground and D and ground, are mutually shifted in phase for 180°, and are being led over the coupling capacitors C2 and C3 to pins 1 and 2 of the NE612, i.e. on one input of the mixer. On the other mixer input the Drain signal is brought, over C4, and beating occurs in the mixer, the result of which is the LF signal on pin 4. This signal is, over C8, being led onto the volume regulation potentiometer and the audio amplifier.
* The unwanted (and parasite) products of mixing, that are manifested as whistling, squeaking etc. are being suppressed by the C7 capacitor. If the obstructions still exist, the capacitance of C7 is to be increased and/or the R* resistor added.
* The voltage stabilization of the DC voltage on pin 8 is performed by the ZPD6.2V Zener diode and resistor R5. A diode with smaller voltage is also possible to be used, say, 6.2 V and similar. If the supply voltage is less than 12 V, the resistance of R5 should be decreased.

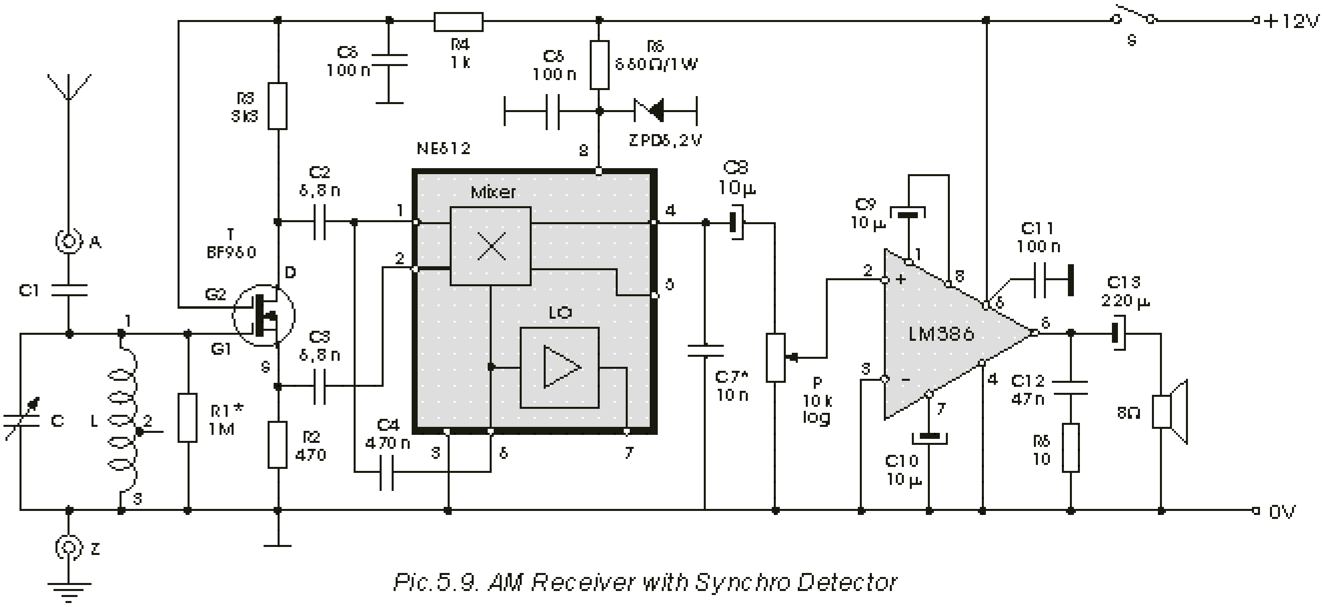

Pic.5.9. AM Receiver with Synchro Detector

5.3.3. Input Circuits for Receivers with the NE612 IC

All the receivers with NE612 that are described here work better, especially considering suppressing the noise in case of the symmetrical station, if the proper input circuitry is added to them. Pic.5.10 shows two examples of the MW receivers that use the ferrite antenna. In both cases, the antenna taken from an old commercial radio is being used.

5.4. The Universal Audio Amplifier

We already spoke about the universal amplifier in the text connected with pic.3.22. Pic.5.11 contains the diagram of another such device, where the transistor amplifier with BC547 is used as the pre-amplifier, instead of that with TLO71 IC. It can be used for practical check of all the earlier mentioned radio receivers. The LF signal is being taken from the detector in the HF part of the receiver to the hubs marked as In and Gnd (if the links aren’t too long the ordinary wires are used, otherwise - the microphone cable). On the third hub the DC voltage is outputted, which is used in some HF circuits for their operation (such as e.g. those on pics.3.24, 3.25, 3.29 etc.).
* The LED (and the appropriate resistor) are added if the amplifier is being supplied from the adapter connected to the household voltage installation. It can also be used if the amplifier is power supplied from the battery, but this is not recommendable, since its power consumption is fairly big, which significantly shortens the battery life.

* The amplifier can be put in a box of any kind, one of the possible solutions shown on pic.5.11-b.
* A very useful solution can be to place the adapter also in the box, with the ability to control its output voltage from few volts to 12 V. In that case, you have both the amplifier and adapter in the same box, which can be used for power supply and check-out of various electronic devices, and not just radio receivers.

5.5. Additional Circuitry

5.5.1. Fine Tuning

During the tuning of the receiver to some station at the SW band with the variable capacitor, a problem occurs. In simple terms, the station frequencies are too close to each other, so the capacitor's shaft should be turned for an extremely small angle in order to change station, which is practically impossible. It would certainly be useful if we could somehow stretch (a popular term for this) the part of the band near the frequency to which the receiver is tuned at. For the direct type (TRF) receivers that were described in the previous chapters, this can be accomplished if, acc. to pic.5.12, another variable capacitor (CR) is added in parallel to the variable capacitor at the input circuit. Its capacitance should vary at substantially smaller scale than that of C, meaning from a few pF til about 20 pF. The tuning is accomplished by setting the receiver, by means of C, approx. at the middle of the band we are interested in, then tuning by means of CR to some station in that area. E.g. if the stations we want to receive are located in the part of the SW band from 6.1 MHz till 6.2 MHz (it's a well-known 49-metre band), first we tune ourselves with C to approx. 6.15 MHz, and then we pick with CR some of the stations located in that area. The same applies for the famous Magic Band (at about 50 MHz).

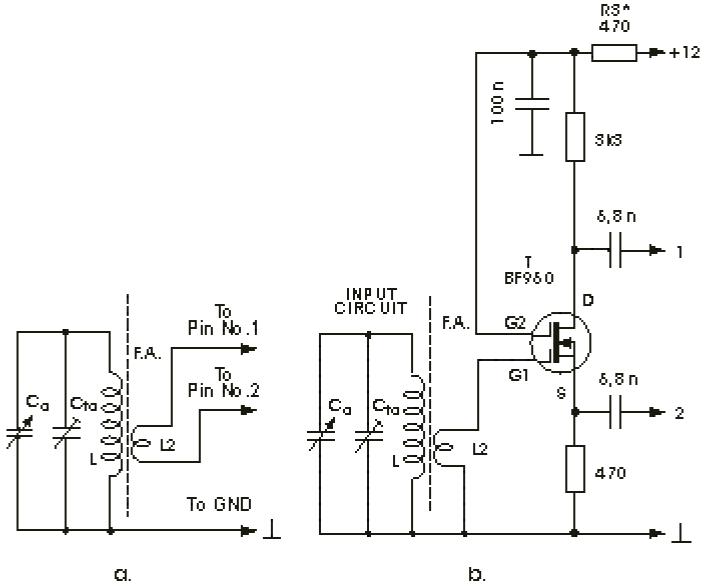

Pic.5.10. a-Input Circuit for the Receiver from pic 5.9,
b-Iput Circuit & HF Amplifier for the Receiver from pic.4.2

The CR capacitor is mounted close to C in order for their knobs to be near each other at the front plate.
* As CR, some air-type trimmer capacitor can be utilized, with adjustment knob mounted on its shaft. Also, one of the sections of the variable capacitor from pic.3.8 can be used, as shown on the right part of pic.5.12.
* The problem of the station "adjacency" at the SW band also exists at the superheterodyne receivers. It is being solved by adding the CR in parallel to the variable capacitor in the local oscillator circuit. The reason for this is that, at supereterodynes, the station is chosen over the local oscillator. The important thing for the

oscillator is to have the exact frequency, that is greater from the station frequency for the amount of the interfrequency. If the resonance frequency of the input circuit isn’t equal to the station frequency, it won’t significantly affect the reception. Because of all this, in the receiver on pic.4.5, CR is attached between the pins 2 and 3 of the LO circuit.

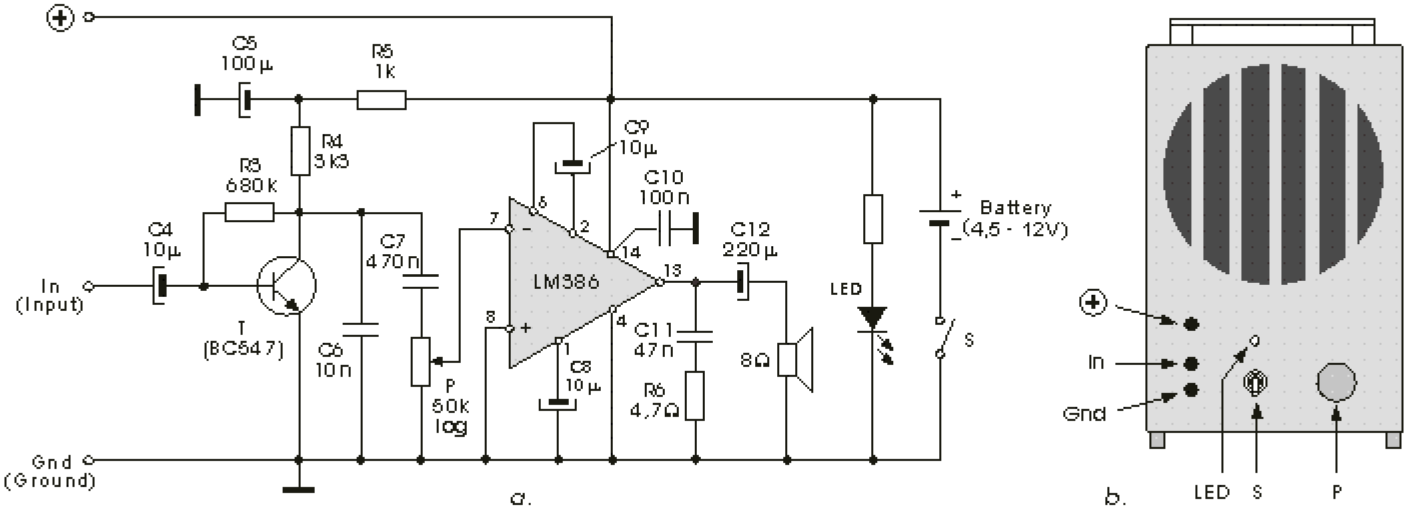
⊕
C5
100µ
R5
1k
R4
3k3
R3
680k
C4
10µ
In
(Input)
T
(BC547)
C6
10n
C7
470n
P
50k
log
C9
10µ
C10
100n
C12
220µ
LM386
7
6
2
14
13
8
4
1
C8
10µ
C11
47n
R6
4,7Ω
8Ω
LED
+
Battery
(4,5 - 12V)
S
Gnd
(Ground)
a.
⊕
In
Gnd
b.
LED
S
P

Pic.5.11. Universal Audio Receiver: a-Electronic Circuit, b-Box

5.5.2. Electronic Tuning

Instead of the capacitor CR, that was used for fine tuning in the previous project, a capacitive (varicap) diode can be used. It’s a special HF diode which is polarized by exposing it to DC voltage in order to be non-permeable (+ to the anode, - to cathode). By changing the voltage diode’s capacitance also changes, which allows for it to be utilized as variable capacitor. If, acc. to pic.5.13-a, the DC voltage between the cathode and anode (UAK) varies from U1 to U2, diode’s capacitance goes from Cmax till Cmin.
The electronic diagram for the electronic fine tuning circuitry is given on pic.5.13-b. Diode capacitance is changed by moving the slider of the P1 potentiometer. By means of trimmer TP the necessary Cmax is set, and when this is done TP can be replaced by an ordinary resistor. All the components are mounted on the PCB, together with other parts of the receiver, except the P1. It is mounted on the front panel, and connected to the PCB with 3 ordinary wires.
* The variable capacitors that were used for tuning in all the receivers described so far are solid, lasting, reliable

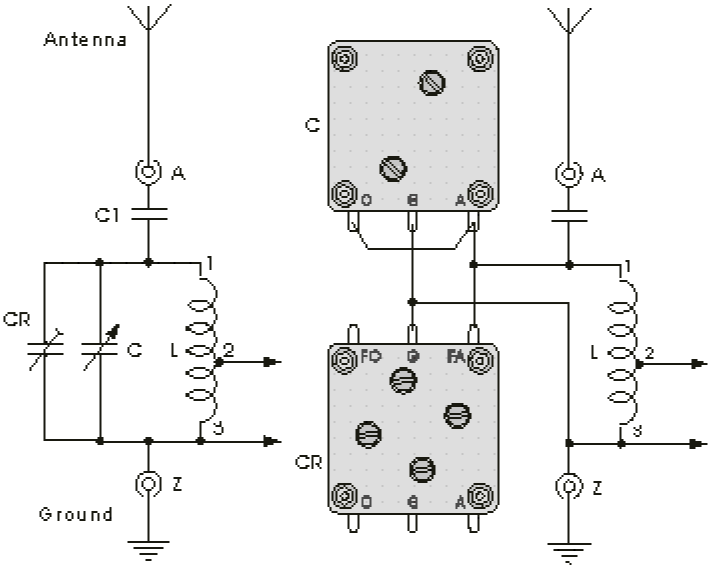

Pic.5.12. Fine Tuning
(stretching a part of the bandwidth)

components. Their mishap is they are hard to purchase, they are quite robust (compared to other device components), and their mounting isn't simple because the shaft for the knob must go through the front plate of the device box. That is why varicap diodes are also replacing them. With the diode that has Cmax/Cmin ratio that is big enough, say, Cmax/Cmin>15, the circuit form pic.5.13 can be used as the variable capacitor (C is simply omitted). In that case, some bigger knob with an arrow is mounted on the P1 handle, and numbers from 1 to 10 are written on the panel, as shown on pic.5.13. This scale allows the listeners to see what station is the receiver tuned at. Of course, for the MW band, the numbers as those on pic.3.7 can also be written.
* In case of SW band, the P2 potentiometer is added for fine tuning.
The optical indication of the tuning, with and knob with arrow is the simplest solution possible. More prettier one is using a small movable-coil instrument (V), such as those used as battery indicators in industrial devices, or for tuning indication and similar. The connecting is done acc. to the diagram on the left part of the pic.5.13-c. In series with the instrument, the TP potentiometer is attached. Its resistance depends on the maximum instrument current, and can be found experimentally. For start, you may use a 1 MOhm linear trimmer, with its slider at lowest position (so that its resistance is maximum). Put the P1 slider also at the lowest position. Turn on the receiver. Start moving the P1slider upwards, and observe the instrument needle. if it soon goes to the end, you'll have to take a trimmer with greater resistance or to add another resistor in series with it, so that when the P1 slider gets to its rightmost position, the needle goes somewhere around the middle of the full scale. If the needle, with P1 in topmost position, moves too little, you'll need a smaller resistance trimmer. When you succeed in having the needle in the middle of the scale with P1 in topmost position, start moving the TP slider until the needle reaches the end of scale. The circuit is well adjusted if the needle goes from zero to full scale while P1 slider is moved from bottommost to topmost position. The instrument can have any shape, but the most appropriate (and cheapest) is square, like the one on the picture.

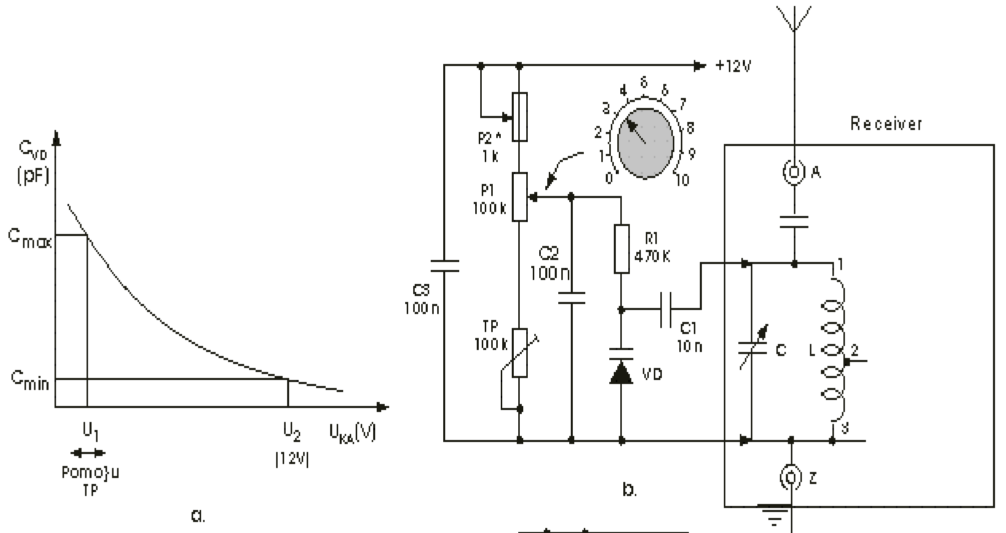
C_VD
(pF)
Cmax
Cmin
U1
U2
U_KA(V)
Pomo}u
TP
|12V|
a.
+12V
P2*
1k
P1
100k
0 1 2 3 4 5 6 7 8 9 10
C2
100n
R1
470K
C3
100n
TP
100k
C1
10n
VD
C
L
1
2
3
Receiver
A
Z
b.

Pic.5.13. Electronic Tuning: a-Diagram, b-Varicap Diode Characteristic, c-Receiver Scale

* The optical indication of station tuning at the receiver with electronic tuning, as those on pics.4.12, 4.14 and 4.19 is accomplished with the instrument that is connected in parallel to the varicap diode. the diagram is shown on the right part of pic.5.13-c. In the lower part of the picture the movable-coil instrument is shown.

Scales of these instruments are most often marked from 1 to 10, but, if you are skilful enough, you can write down numbers representing frequencies in MHz (e.g. 88...96....100.....108 MHz).

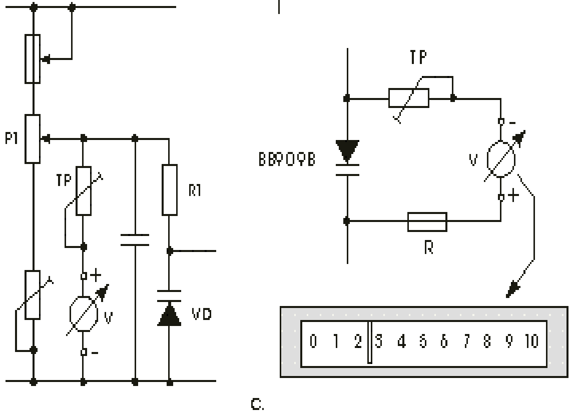
P1
TP
R1
V
VD
+
-
BB909B
TP
V
R
0 1 2 3 4 5 6 7 8 9 10
c.

5.5.3. Suppressing the Signal of the Local Transmitter

From all the signals in the reception antenna, the one that is created by the local transmitter is by far the strongest one, due to the fact that it is hundreds, sometimes even thousands times closer than other radio transmitters. That signal can be so strong that it can jam normal reception of other stations. In case of simpler receivers its programme is heard,more or less, in all the positions of the variable capacitor. The solution for this problem is the so-called seal circuit, which serves to weaken the signal of the local transmitter, so that it doesn’t interfere (but is still strong enough for normal reception, when the receiver is tuned at it).

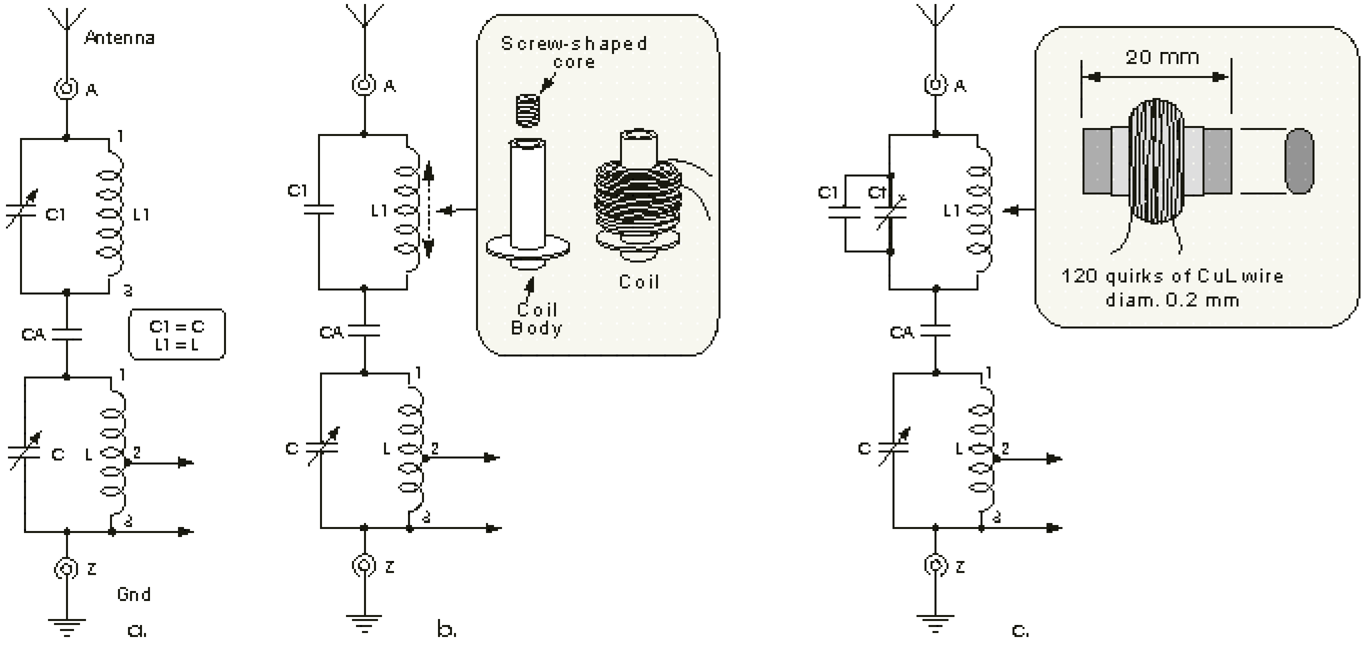

Pic.5.14. Suppressing the Signal of the Local Transmitter

The seal circuit is a parallel oscillatory circuit which comprises the coil L1 and capacitor C1, as shown on pic.5.14-a. By means of C1 the resonance frequency of the circuit is set so that it corresponds to the carrier frequency of the local station. On that frequency, this circuit behaves as a huge resistor (see pic.3.2-b) and decreases the current that is created by the local transmitter signal. For other signals it has very small resistance and practically has no effect on them. The setup is done by tuning the receiver on the local station, and the reception is weakened enough by turning the C1. If the decay is too strong, a resistor should be added in parallel to C1.

Using a variable capacitor in the seal circuit (pic.5.14-a) isn't an economical solution. It is much better, considering both economy and space, the solution given on pic.5.14-b. A block capacitor C1 and a variable inductance coil are used in the seal circuit. As shown on the framed part of picture, the coil is wound on the plastic body, with ferrite core. The number of quirks is found experimentally about couple of hundreds of quirks made with as thin copper wire as possible). The capacitance for C1 is also found experimentally (couple of hundreds of pF). The earlier mentioned IF transformer can also be used as a coil. With labelling acc. to pic.4.3-a, legs No.2 and 3 are used, the others are "hanging" (they are not soldered). C1 capacitance is also found experimentally. It is also possible to wind the coil on a piece of ferrite rod, as shown on pic.5.14-b, and setup to be done with trimmer Ct

5.5.4. Dual Tuning

The author of this book, as great radio techniqe lover (amateur, in French), owns great collection of over 150 pieces of various old-timer radio receivers. There is one among them that is over 60 years old, at which the tuning is being done by two knobs. With first one the

receiver is set roughly to the desired station, which is usually barely heard at that moment. The second knob is then turned until the optimum reception is achieved, which is significantly better than before, and in case of weak stations - extremely better.
The selectivity of simple receivers that were described in previous chapters can be significantly increased by using the aforementioned dual tuning. The electronic diagram is shown on pic.5.15-a. Another oscillatory circuit, made of L1 and C1 connected in series, is inserted between the antenna connector and input circuit of the receiver (it can be any of the earlier described AM receivers). As with the earlier mentioned parallel oscillatory circuit, the resonance frequency of the serial circuit is given by the Thompson pattern:

$$f_r = \frac{1}{2\pi\sqrt{LC}}$$

The serial oscillatory circuit has very small impedance (compared to the parallel circuit whose impedance is very big on the resonance frequency). The dependance of the impedance ("resistance") of the serial oscillatory circuit from the frequency is shown on the diagram on pic.5.15. As you can see, the serial circuit acts as a resistor of very small impedance only for the station that it's tuned at. For all other stations, it behaves as a huge resistor (impedance). All in all, from all the signals in the antenna, the biggest current, and therefore the biggest voltage on the input circuit is created by the transmitter that both serial and parallel oscillatory circuits are set to. The tuning is done as it has already been described, first with C (so-so), then with C1 (much better).

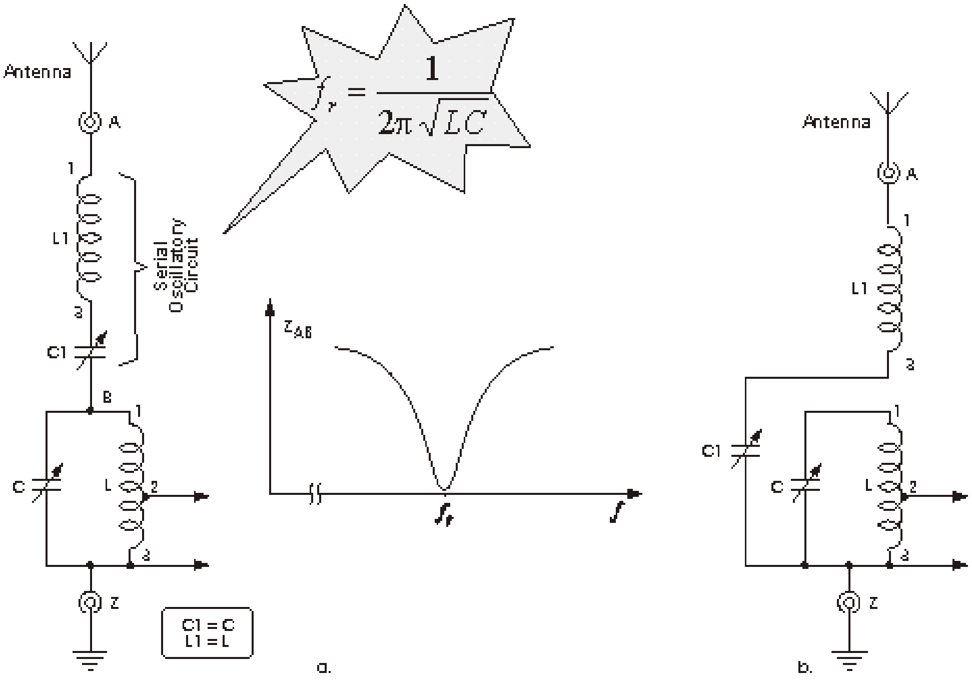

Pic.5.15. Dual Tuning: First by C (so-so), then by C1 (much better)

* Between the coils L1 and L a magnetic coupling should be prevented. This is accomplished by mounting the coils to be as far from each other as possible, and to position their axes mutually perpendicular.

* Greater experimenting opportunities with dual tuning provides the diagram on pic.5.15-b. Once again, it’s the serial resonance (in circuit L1, C1), and parallel resonance (in circuit L, C), that are being used. The coils are placed side-by-side, in order to generate magnetic coupling between them. The tuning is done as previously explained, but now we also have a possibility of changing the amount of magnetic coupling between the coils by moving them closer or farther, which affects the antenna’s influence on the L, C oscillatory circuit, therefore changing its selectivity and sensitivity.

5.5.5. Separation of Stages - Preventing the Oscillation

On of the significant problems that occur at devices that comprise more cascade-linked amplifying stages is the occurrence of the feedback over the conductors that connect those stages with the positive pole of the battery, or the power supply. By the way, the feedback is a phenomenon when part of the signal exiting an amplifier gets on its input. Under certain conditions, this feedback causes the oscillation of the stage, which in devices that have the loudspeaker on output, manifests itself as strong whistling, squeaking and similar.
On of the ways to prevent this feedback is given on pic.5.16, where a block-diagram of a radio receiver that has four amplifying stages with active components (transistors or IC’s) that require the battery supply is shown. Separation of stages for the AC current (preventing the feedback) is accomplished by the LF filters with resistors and capacitors. Resistors are from couple of hundreds of Ohms to 1 kOhm. Capacitances of C1 and C2 are from couple of tenths till couple of hundreds of nF, and of C3 from couple of hundreds of nF to about 100 mF. The stage PCBs should be designed in such way to make the contact where right end of the capacitor is soldered as close to the contact where the positive end of the power supply voltage is brought (e.g. on pic.5.9, the right contact for C6 should be as close as possible to the contact where pin 8 of NE612 is soldered).
In the devices supplied from the battery, the C5 capacitor, which has capacitance of couple of hundreds of micro Farads, serves to take the role of the battery when it gets emptied a little bit, and strong tones have to be reproduced at the loudspeaker (in simple terms, C5 acts as a small accumulator that helps the worn-out battery to give enough power to the power amplifier, when necessary. When its help isn’t needed, the capacitor is refilled). This capacitor is not needed when the receiver is supplied from the adaptor that already has an electrolytic capacitor on its output, and when the wires that connect the adaptor to the receiver are not longer than about 15 cm.

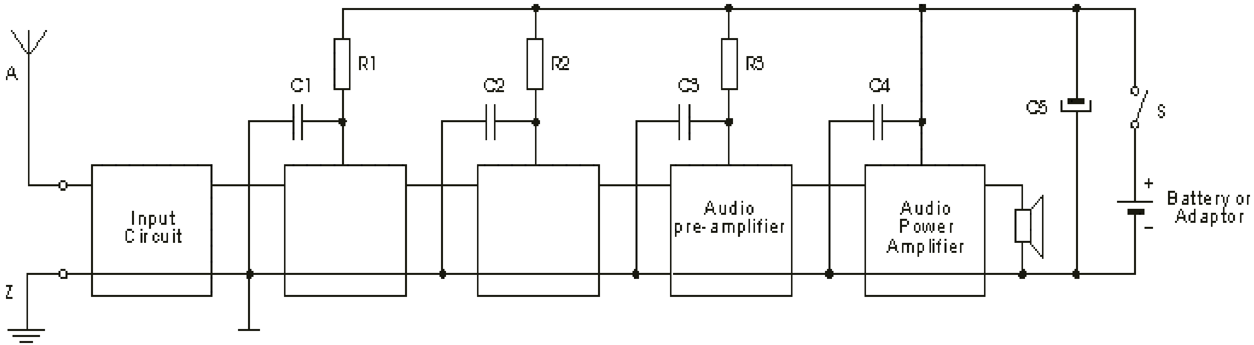

Pic.5.16. Separation of Amplifying Stages

5.6. The Boxes

For all lovers of the electronics, the box where their device is to be put is the famous “production weak link”. The finished boxes are either impossible to purchase, or they can be bought but their dimensions or shape is inappropriate, or they are too expensive, or... In cases like this one should be quick-witted enough to find some square-shaped box that is being used at household, or some packaging box or similar. That is how it’s done in the “whole white world”. Two years ago, in the famous electrotechnical magazine ETI TOP PROJECTS the article named “TIC TAC RADIO” was printed, where a receiver with ZN414 that is placed in the transparent plastic box of TIC TAC mints (In the abstract, it was written that making this device serves well as an excuse for buying candies, which is probably meant for the readers that are on a diet for aesthetic reasons).
However, the “finishing touch” is of great importance for everything. The majority of your friends will be more impressed by a lovely box where the receiver is placed, than the reproduction quality, type of modulation and other technical characteristics. And a nice, appropriate box cannot be bought, it is up to you to make it. It can be something as on pic.3.11 or similar. The idea can be also found in some catalogue of radio receivers’ manufacturers, or you can think of something of your own. As far as the author of these lines is concerned, he likes best the wooden boxes from the 20’s and 30’s of the previous century, from the times of the charleston, E. H. Armstrong and Al Capone. They looked something like those on the pic.5.17 and can serve you as an inspiration for your personal design.

The mid button is for the variable capacitor for station tuning, the right one is for the potentiometer for volume regulation. The button on the left can be a rotary switch for turning on/off (S). It can also be a tone regulation button, and for the reaction-type receivers it can be a button of the potentiometer that regulates the magnitude of the reaction. In the last two cases, the on/off switch (S) is located on the regulation potentiometer. The outside antenna and ground hubs are located at the rear panel of the box. The wires connecting the hubs with the PCB should be isolated, flexible and long enough to be able to open the panel and put it at upright position.
If the receiver is power supplied from the outside net, a green LED should also be added, as the power indicator. The good place for it is just above the variable capacitor’s button, instead of the triangle-shaped marker.
Pic.5.18 shows the parts for the first box from pic.5.17. For the front and rear side two pieces of 5 cm thick plywood, measuring 22 cm x 15 cm are needed; for the side panels, two pieces of 10 cm thick plywood, 15 cm x 9 cm, and for the bottom side - one piece of 10 cm thick plywood, measuring 13 cm x 9 cm. The best way to cut these parts is to be done by the carpenter on the special machine, since only then will they be of strictly rectangular shape, and bottom and side panels will have exactly the same width, which is very important during assembling. On the front side, the circle and the arc are drawn with the aid of the sector, and the cutting is done with the carving saw. The part that is cut from the back panel will serve as a closure. When it is cut it isn’t necessary to treat it with emery, since it will fit nicely in the hole on the rear panel even if it isn’t cut evenly. On the inner side of the rear panel two plywood lattices measuring about 2 cm x 13 cm should be nailed. Four wood screws will be screwed in them later (the holes are shown as four dots), which will serve to tighten the closure. Connecting of the pieces is done with the wood glue and small nails. Before you start hammering, it is very useful to drill a few holes for the nails in the front and rear panel with 1 mm drill. The nails are partially hammered into the panels, the edges are then covered with glue, and the nailing can then be done. When all this is finished, the box should look as the drawing at the right end art of the pic.5.18 *vertical stripes over the loudspeaker opening are not shown. They can be omitted, and you can nail in a few thin lattices, when the box is finished, as shown on the last drawing on the

pic.5.17). The semicircle part is made of 5 mm x 5 mm lattices, or similar, which are put side by side on the upper edges of the front and rear panels, that are covered with glue (the picture shows only one of these lattices). When the last one is fitted, the space between them is filled with "putty" that is made by mixing the fine wooden chips with the wood glue, with the aid of a steel plate. After that, the lattices are tightened to the panels' edges by two pieces of strong scotch tape, which are shown in dashed lines, and everything is left to dry well. When drying is, after about 10 hours, finished, all the edges and lattice parts that protrude are well flattened with emery. All the remaining holes are filled with the fast-drying putty, and everything is abraded once again, and the putty is applied again, and abraded again, etc., until the upper part is semicircle-shaped, all the sides smooth and the edges correct.

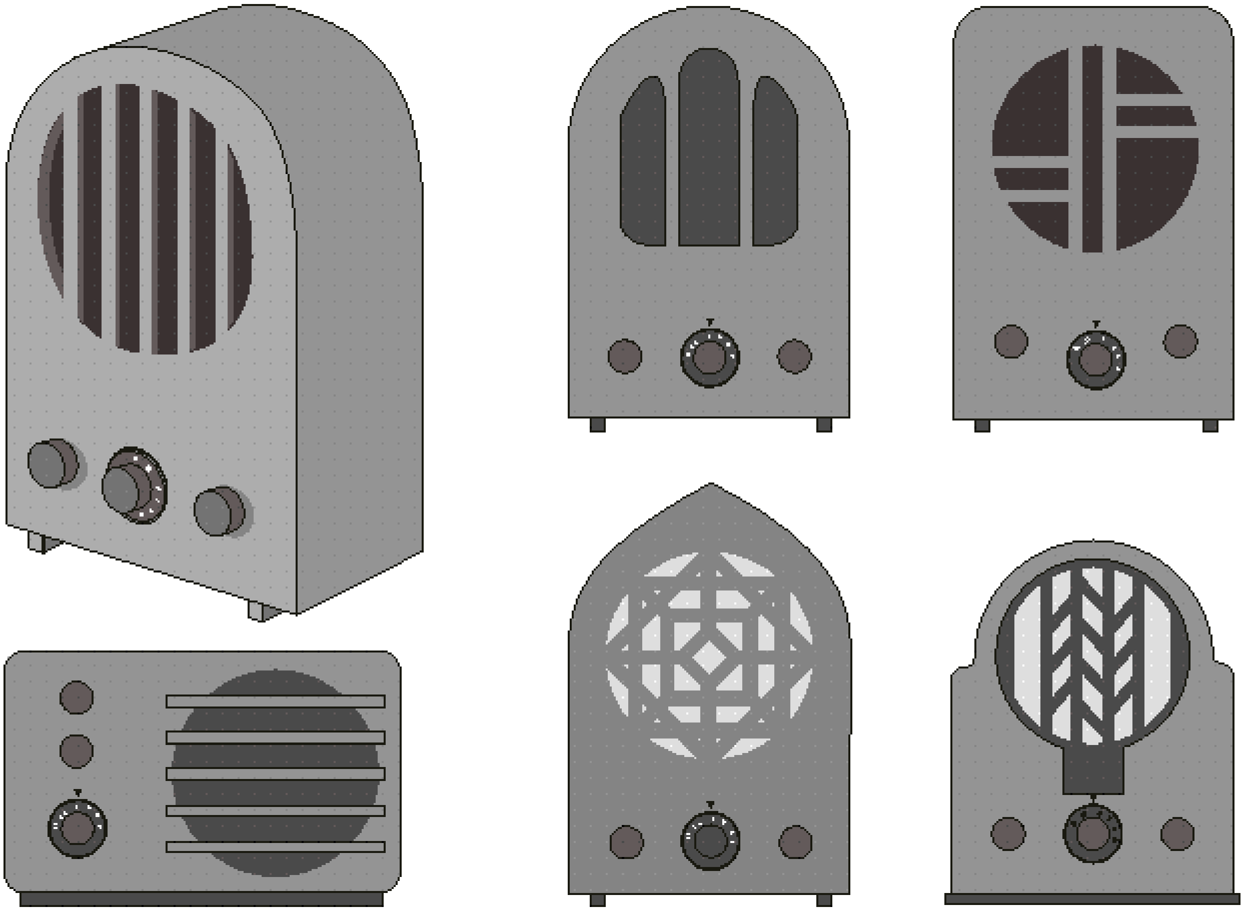

Pic.5.17. Radio Receiver Boxes

* Before the loudspeaker is attached with screws, a piece of decorating cloth should be placed between the panel and the loudspeaker, which will protect it and contribute to better looking box.

* Perhaps some of the readers will seem that there's a lot of exaggeration in previous lines, and even too much pedantry. There's a Latin proverb, that says: AGE QUOD AGIS - Do the

things you do, which, in our case, can be interpreted as: You should either make the box properly or not making it at all.

* This box is relatively small, it is predicted for the loudspeaker that is about 12 cm wide. If you have bigger loudspeaker, and it will certainly play both louder and better, you should make a bigger box. The dimensions calculation is done by dividing the diameter of the bigger opening, that will suit bigger loudspeaker, in centimetres, by 11, and all the measures on pic.5.18 are multiplied with the number attained. E.g. if the diameter for the new, bigger hole is 15 cm, new dimensions are obtained by multiplying the old ones by 1.36.

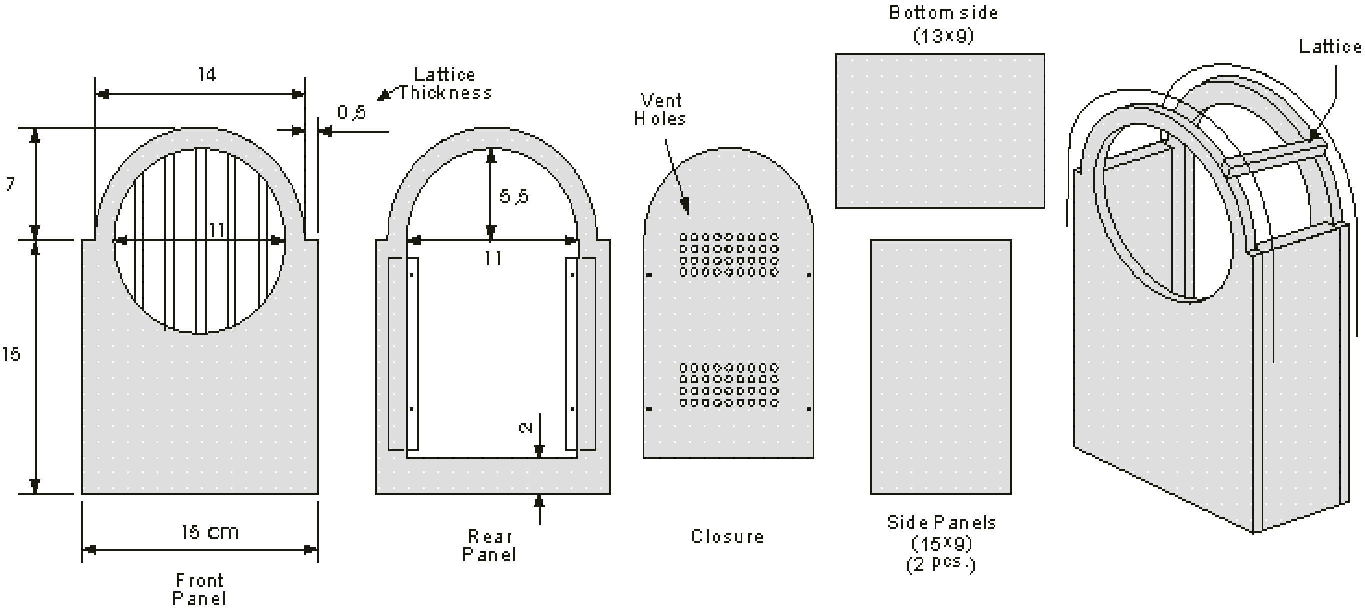

Pic.5.18. Parts of an old-fashioned Radio Receiver's Box

5.7. Bimboard, Protoboard . . .

The readers that have carefully studied all the radio receiver projects that are described here, have possibly noted that the author referred to experimenting either with values of some components or with entire circuits, and all that was in order to practically find the optimal solution. When small changes are discussed, such as finding the optimum value for some resistor, that can be done on the previously made PCB. In case of bigger changes that of course is not convenient, and sometimes it is almost impossible. For all kind of electrical diagrams check-outs as well as various experimenting with all electronic devices, including radios, it is the best to use a special experimental board, which can be purchased under various trade names: protoboard, bimboard, matador, steckplatine, steckboard etc. All of them have in common that component connecting is done without soldering, by simply inserting the legs into the small holes on the plate.
As an example, pic.5.19 shows a full-scale experimental board that can be purchased in one of the Belgrade electronic shops. It has 630 vertically aligned holes, connected internally in 126 groups of 5 holes each, and another 100 holes placed in two topmost lines, connected in two horizontal groups by 50 holes each. The connections between the holes are inside the board and cannot be seen, they are shown on picture in dashed lines. The two topmost

parts are used fo bring the supply voltage, and the battery or the adaptor is connected to them. One of them, most often the one that has minus pole connected to it (in all the devices described herein), also serves as the device Ground. The holes contain miniature metal hubs that are elastic, so when a leg is popped in, a reliable mechanical and electrical junction is accomplished. The distance between the adjacent holes is 2.54 mm (1/10 inch), which allows for connecting the vast majority of electronic components, which are being produced with inter-leg distance that is equal to a hole number multiplied with 2.54 mm (in the producers and sellers' catalogues the 2.54 mm distance is marked as R, which stands for raster, and the components that have their legs horizontally and vertically distanced to 2.54 mm multiplied by some whole number are said to have their legs in raster).

The necessary electrical junctions between the hole groups are accomplished with connecting wires that can be bought at Conrad, but are more often self-made from plastic-isolated 0.5 mm or 0.6 mm copper wire. These pieces vary in their length and can be bent as the biggest piece in the lower left part of the pic.3.19, although it is better and nicer to use regular pieces, shaped as the cyrillic letter P.

Pic.5.19 also shows an example of practical usage of one such board. The radio-receiver from pic.3.15 is made on it. As can be seen, the coil ends are stuck into the holes whose coordinates are j,37; j,39; i,45; i,47, the diode in holes i,39 and i,45, the pin No.1 of the IC in e,54, etc. With the connecting wires the legs No.1 and 3 are connected, the ones that are connected with the potentiometer slider, and legs No.2 and 4 are connected to the ground by means of 4 connecting wires (the minus pole of the battery), etc.

It is now clear that experimenting is done in a very simple manner. E.g. if you are interested how does a capacitance of C2 affect the tone colour in the headphones, all you should do is remove it and insert a capacitor of greater or smaller capacitance, etc.

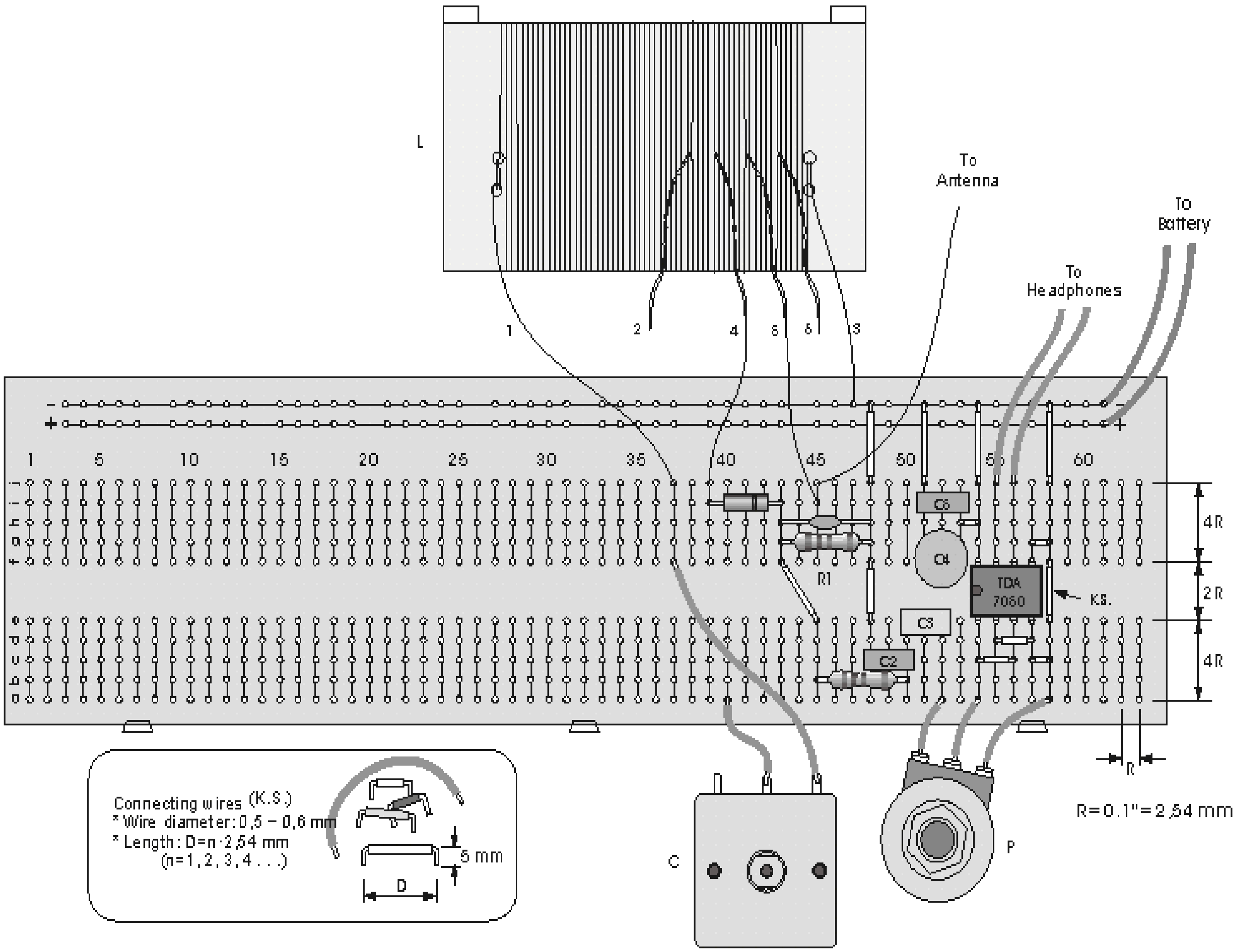
L
To Antenna
To Battery
To Headphones
1
2
4
6
5
3
-
+
1
5
10
15
20
25
30
35
40
45
50
55
60
j
i
h
g
f
e
d
c
b
a
R1
C6
C4
TDA 7060
K.S.
C3
C2
4R
2R
4R
R
R=0.1"=2,54 mm
C
P
Connecting wires (K.S.)
* Wire diameter: 0,5 – 0,6 mm
* Length: D=n·2,54 mm
(n=1, 2, 3, 4 . . .)
5 mm
D

Pic.5.19. The Receiver from pic.3.15 Mounted on the Protoboard

* The hubs on the board are elastic, so that conductors of various diameters can be easily inserted. No wires much thicker than 0.6 mm should be inserted, since the hubs will deform. The components whose legs are too thick as the variable capacitors, potentiometers, transformers and similar, are connected over pieces of wire that are soldered to them.

* It is useful for the connecting wires to be made with isolations of various colours, so that red ones could be used e.g. for connecting with the + battery pole, the black ones with Gnd, the yellow ones for the signal etc.

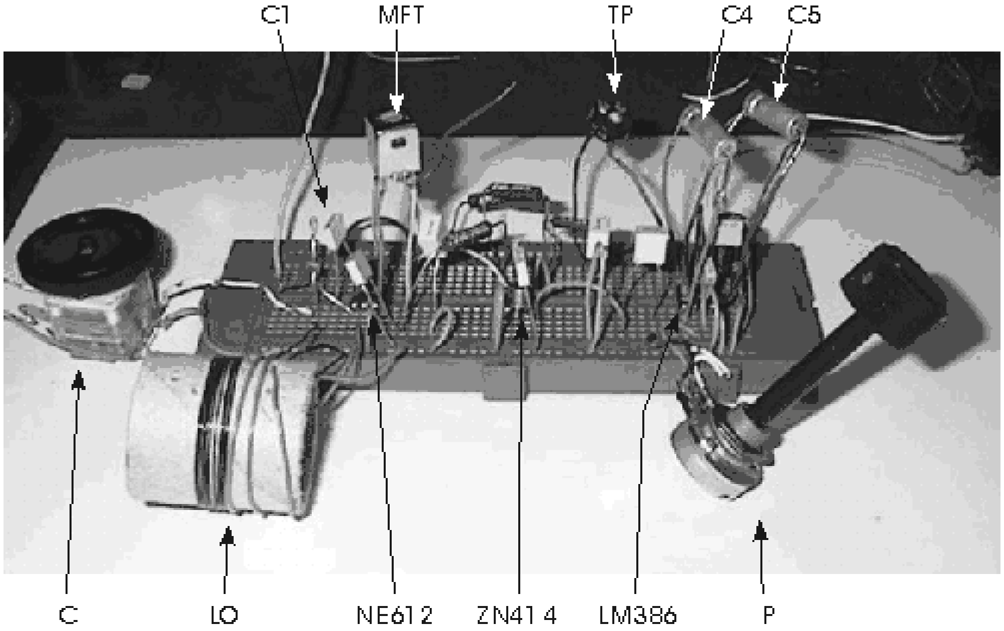

Pic.5.19-b The Receiver from pic.4.5, with the SW Coil, mounted on the Bimboard

5.8. Universal PCB

Practical realization of simple radio receivers, as well as other simple electronic devices, can be done in many ways, as it was discussed in PE No.2. One of those is shown also in this number. That is construction of the detector receiver from pic.3.11, where some of the components are mounted onto the box walls (the variable capacitor, coil and the hubs), while other (the diode and two block-capacitors) are placed between them. With some skills, and by the aid of few smaller nails nailed from the inside of the front panel, a more complex device could be made, say, that from pic.3.12. But this solution would start looking as “the dead cockroach technique”, which will be discussed in the “Funniest Electronics”. The real solution is the PCB that can be made from the drawings that are given, or those you will draw yourself, together with the instructions given in chapter 5.1.

There’s another option for practical realization. It is a universal PCB, that can be bought in the electronic components’ stores. There are more sorts of these PCB’s, and all of them have in common that the holes on them are drilled on the distance of 1/10 inch (R=2.54 mm).

One of the universal PCB’s is shown on pic.5.27. It consists from a huge number of round copper isles, with hole in the middle. The components are being soldered first (resistors, diodes, IC’s, capacitors...), and then the component pins are connected by pieces of isolated copper wire, on the soldering side.

As an example, pic.5.28 contains the photograph of the receiver from pic.3.21-a that is made with the universal PCB from pic.5.27. It can be placed into a box as on pic.3.21-c, except the box should be bigger, in order for the loudspeaker to.

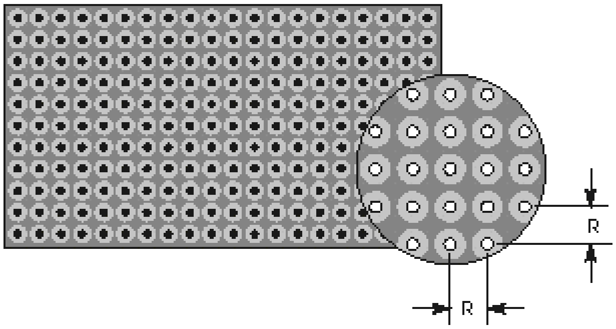

Pic.5.20 The Universal PCB

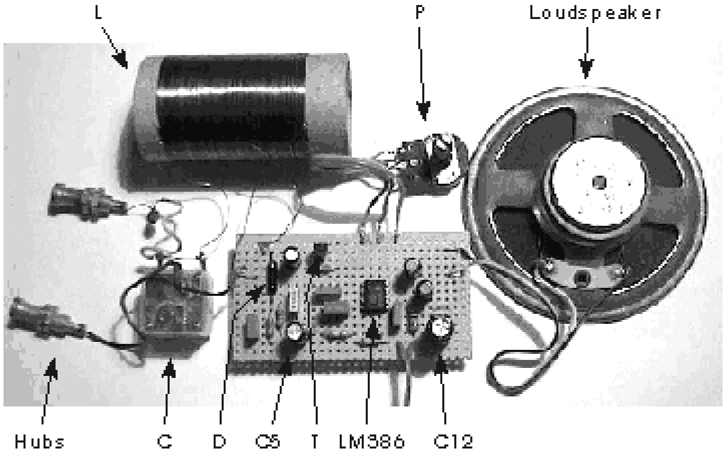

Pic.5.21 The Receiver from pic.3.21-a, mounted on the universal PCB

5.9. A Modern Oldtimer

The receiver on pic.5.22 is designed for the readers that wish to make a semi-conductor model of a complete direct radio receiver that was being produced many years ago, with electronic tubes. It had a total of 3 tubes, one of the contained the HF pentode (utilized in the HF amplifier) and the diode (used in detector), the other one had a triode (pre-amplifier) and powerful pentode (power amplifier), whilst the third one contained the duo-diode (the rectifier).

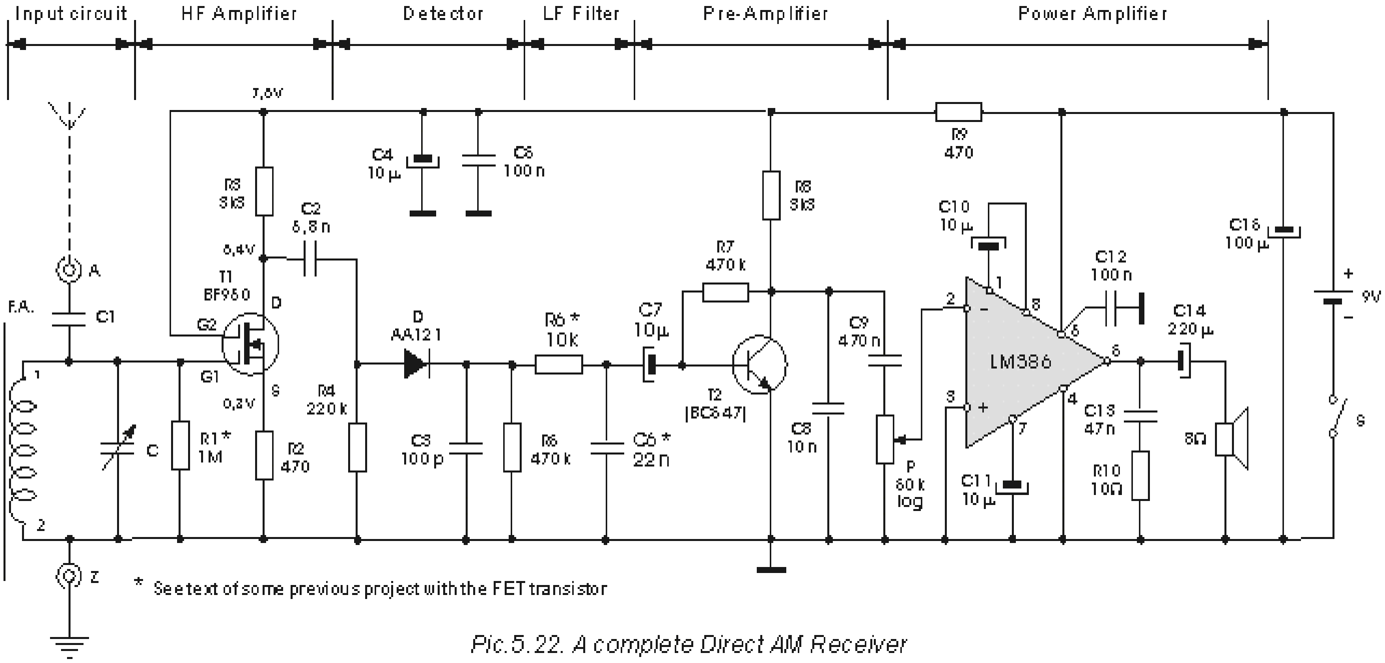
Input circuit
HF Amplifier
Detector
LF Filter
Pre-Amplifier
Power Amplifier
7,5V
A
F.A.
C1
1
2
C
R1*
1M
Z
R3
3k3
C2
6,8 n
6,4V
T1
BF960
D
G2
G1
S
0,3V
R2
470
R4
220 k
C4
10 µ
C5
100 n
D
AA121
C3
100 p
R5 *
10k
R6
470 k
C6 *
22 n
C7
10µ
R7
470 k
T2
[BC547]
R8
3k3
C8
10 n
C9
470 n
P
50 k
log
R9
470
C10
10 µ
1
2
-
8
3
+
7
LM386
6
5
4
C11
10 µ
C12
100 n
C13
47 n
R10
10Ω
C14
220 µ
8Ω
C15
100 µ
+
9V
-
S

* See text of some previous project with the FET transistor

Pic.5.22. A complete Direct AM Receiver

THE END

Visual inspection

What are we looking for during this inspection? In general terms we are inspecting the installation with regard to:

S afety
A ge
D eterioration
C orrosion
O verload
W ear and tear

An easy way to remember this is to use the acronym SADCOW.

Suitability and external influence should also be included. At this point it is a good idea to get from the client any documentation that is relevant to the installation. These documents could include:

- Plans
- Drawings
- Previous test results and certification
- Fuse charts

You should also make it clear that you will require access to all parts of the building and that the electricity supply will need to be turned off at some point. It is also a good idea to ask the client if they are aware of any alterations that have been carried out, as this information may be useful to you during inspection.

The visual inspection of any installation is as important as any testing that is carried out on an installation; if you are not familiar with the building it is also a good opportunity to find your way around first.

The first part of a visual inspection is to ensure that the system is safe to test and that you have enough information to be able to carry out the test safely. Generally, a good place to start would be the supply intake; this will give a reasonable indication of the age, type and size of the installation.

Things to look for at the supply intake before removal of any covers would be:

- The type of supply system – is it TT, TNS or TNCS?
- Is it old or modern?

- Are the conductors imperial or metric?
- What type of protection is there for the final circuits?
- Is documentation available for the original installation?
- Is the consumer's unit labelled correctly?
- Is the earthing conductor in place?
- What size is the earthing conductor?
- Is the earthing conductor green or green and yellow?
- Are all of the circuits in one consumer's unit or are there two or three units that need combining?
- Is there any evidence of equipotential bonding? Remember! It must start at the main earthing terminal.
- What size is the equipotential bonding? Is it large enough?
- Is there a residual current device (RCD)? If so has it a label attached? Is it a voltage or current operated type?
- Do the enclosures meet required IP codes? (Regulation 412-03-01)
- If alterations have been carried out is there documentation available for them, along with test results?
- Where alterations have been carried out since January 2005, has a warning notice been fitted on or near to the distribution board to indicate that new colours have been used? (Regulation 514-14-01)
- What size is the supply fuse? Is it large enough for the required load?
- Are the meter tails large enough?
- Are the seals broken on supply equipment? If they are it could indicate that the system has been tampered with since it was first installed and perhaps closer investigation is required.
- Have any alterations or additions been made?
- Would any alterations or additions affect the required disconnection time for the circuit concerned?

This list is not exhaustive and installation conditions may require more.

When the visual inspection of the supply intake area is complete, that is a good time to look around the building to make sure that there are no very obvious faults. All of this should be carried out before removal of any covers.

Things to look for:

- Are accessories fixed to the wall properly? Are they missing or damaged?
- Are the accessories old with wooden back plates?
- Are the socket outlets round pin or square? Is there a mixture of both?
- Have cables been installed in vulnerable situations?
- Have cables, enclosures and accessories been fixed securely?

- Have ceiling roses got perished flexes? (Particular attention should be given to the old braided and rubber type flexes.)
- Are any socket outlets likely to be used outside? If they are then they should be RCD protected. If they have been installed before the late 1990s, then it is not a requirement that they are, but an RCD should be listed as a recommendation.
- Are earthing clamps to BS 951 standards and correctly labelled?
- If gas, water is bonded using the same conductor, ensure that the conductor is continuous and not cut at the clamp.
- Is the supplementary bonding in place in bathroom? (See Figure 4 of the On-Site Guide.)
- Is the correct equipment for the correct zones in bath/shower room? (See 601 BS 7671)
- Has the bedroom had a shower installed? If so, are the socket outlets 3 metres from the shower and RCD protected?
- Is there any evidence of mutual detrimental influence; are there any cables fixed to water, gas or any other non-electrical services? (*The cables need to be far enough away to avoid damage if the non-electrical services are worked on.*)
- Are the cables of different voltage bands segregated? Low voltage, separated extra low voltage (SELV), telephone cables or television aerials should not be fixed together (*although they are permitted to cross*).

Whilst these items are being checked, look in any cupboards for sockets or lights. If your customer is uncomfortable with this it is vitally important that you document any areas that cannot be investigated in the extent and limitation section on the Periodic Inspection Report. During this purely visual part of the inspection you will gain some idea of the condition of the installation, and indeed any alterations which have been carried out by a qualified tradesman or by a cowboy/girl.

Clearly, if it is an old installation, an electrical installation certificate must be completed and some of the items listed above will apply. However, if it is a new installation, access to all areas must be secure; if this is not possible then the certificate should not be issued. Again, this list is not exhaustive but will not require removal of any fittings, etc.

Providing that you are happy that the installation is safe to tamper with, a more detailed visual inspection can be carried out and the dreaded but necessary form filling can be started.

Once again begin at the consumer unit. Before you start, this must be isolated. The Electricity at Work Regulations 1989 states that it is an offence to work live. Once you remove a cover you will be working live if you do not isolate it first. Having carried out the safe isolation procedure, remove the cover of the consumer unit.

- Your first impression will be important – has care been taken over the terminations of cables (*neat and not too much exposed conductor*)?
- Are all cables terminated and all connections tight (*no loose ends*)?
- Are there any signs of overheating?
- Is there a mixture of protective devices?
- Are there any rubber cables?
- Are there any damaged cables (*perished or cut*)?
- Have all circuits got Circuit Protective Conductors (CPCs)?
- Are all earthing conductors sleeved?
- On a photocopy of a Schedule of Test Results record circuits, protective devices and cable sizes.
- Look to see if the protective devices seem suitable for the size cables that they are protecting.
- Note any type D or 4 circuit breakers – these will require further investigation.
- Are all barriers in place?
- Have all of the circuit conductors been connected in sequence, with phase, neutral and CPC from circuit number 1 being in terminal number 1 – preferably the highest current nearest the main switch?
- Have any protective devices got multiple conductors in them, are they the correct size (*all the same*)?
- Is there only one set of tails or has another board been connected to the original board by joining at the terminals?

Having had a detailed look at the consumer unit, and with the installation still isolated, carry out a more detailed investigation of the rest of the installation.

It may be that you have agreed with your client that only 10% of the installation is to be inspected. This would mean 10% of each circuit. There would be little point in inspecting 10% of the circuits. If the period between inspections was 10 years it could be many years before a circuit was eventually inspected and the exercise would be pointless.

During your preliminary walk around, you will have identified any areas of immediate concern, and these must be addressed as your inspection progresses. There is no reason why you should not start your dead testing at this point, as you progress through your visual inspection.

On radial circuits this would be a good time to carry out CPC continuity, $R_1 + R_2$, insulation resistance and polarity tests as you work your way round. Start at circuit number 1 and work your way through the circuits one at a time.

But first what are you looking for? Let's look at a selection of circuits.

Shower circuit

- Is isolation provided, if so is it within prescribed zones?
- Has the correct size cable/protective device been selected?
- Is it bonded?
- Are connections tight?
- Has earth sleeving been fitted?
- Is the shower secure?
- Is there any evidence of water ingress?
- Is the shower in a bedroom?

Cooker circuit

- Is the switch within 2 metres of the cooker or hob?
- Has the cooker switch got a socket outlet? If so it requires a 0.4 second disconnection time.
- Green and yellow sleeving fitted.
- If it has a metal faceplate has it got an earth tail to the flush box?
- Is the cable the correct size for protective device?
- Are there any signs of overheating around the terminations?
- Is the cooker outlet too close to the sink? Building regulations require any outlets installed after January 2005 should be at least 300 mm from the sink.

Socket outlets

- Is there correct coordination between protective devices and conductors?
- Green and yellow sleeving fitted.
- Do any metal sockets have an earthing tail back to the socket box?
- Radial circuit not serving too large an area (*see Table 8A of the On-Site Guide*).
- Secure connections.
- Are cables throughout the circuit the same size?
- Are there any sockets outside? Are they waterproof? Are they 30 mA RCD protected?
- Are there any outlets in the bathroom? If there is, are they SELV?

- Are there socket outlets within 3 metres of a shower installed in a bedroom? If there is, are they 30 mA RCD protected?
- Will the protective device for the circuit provide 0.4 seconds disconnection time?

Fused connection units and other outlets

- As above but could be 5 second disconnection time.
- Does it supply fixed equipment in bathrooms? Are they in the correct zones?
- Do they require RCD protection? (*Regulation 601-09-02 and 601-09-03*)
- Permanently connected equipment must be protected locally by a plug or fused connection unit or comply with (*Regulation 476-03-04*).

Immersion heater circuits

- Is there correct coordination between the protective device and live conductors?
- Has the CPC been sleeved?
- Is the immersion the only equipment connected to this circuit? (*Any water heater with a capacity of 15 litres or more must have its own circuit*.) *On-Site Guide*, Appendix 8. Often you will find that the central heating controls are supplied through the immersion heater circuit.
- Is the immersion heater connected with heat resistant cord?
- The immersion heater switch should be a cord outlet type; not a socket outlet and plug.
- If the supplementary bonding for the bathroom is carried out in the cylinder cupboard, does the supplementary bonding include the immersion heater switch? (It should.)

Lighting circuits

- Is there correct coordination between the protective device and the live conductors?
- How many points are there on the circuit? A rating of 100 watts minimum must be allowed for each lighting outlet. Shaver points, clock points and bell transformers may be neglected for the purpose of load calculation. As a general rule, ten outlets per circuit is about right. Also remember that fluorescent fittings and discharge lamps are rated by their output, and the output must be multiplied by a factor of 1.8 if exact information is not available (*Table 1A of the On-Site Guide*).
- Are the switch returns colour identified at both ends?
- Have the switch drops got CPCs? If they have, are they sleeved with green and yellow?
- Are the CPCs correctly terminated?

- Are the switch boxes made of box wood or metal?
- Are ceiling roses suitable for the mass hanging from them?
- Only one flexible cord should come out of each ceiling rose unless they are designed for multiple cords.
- Light fittings in bathrooms must be suitable for the zones in which they are fitted.
- Circuits supplying luminaries fitted outside must have a 0.4 second disconnection time (*Regulation 471-08-03*).
- Is the phase conductor to ES lampholders connected to the centre pin? This does not apply to E14 and E27 lampholders (*Regulation 553-03-04*).

Three phase circuit/systems

These circuits should be inspected for the same defects that you could find in other circuits. In addition to this:

- Are warning labels fitted where the voltage will be higher than expected? For example, a lighting switch with two phases in it, or perhaps where sockets close to each other are on different phases.
- Are conductors in the correct sequence?
- Remember PFC should be double the phase to neutral fault current.

Occasionally other types of circuit will be found, but the same type of inspection should be carried out using common sense.

Always remember that the reason for this inspection is to ensure safety

Periodic testing

The level of testing will usually be far less for periodic testing than it is for initial verification; this is providing that previous test results are available. If they are not, then it will be necessary for the full survey and the complete range of tests to be carried out on the installation, to provide a comprehensive set of results.

The level of testing will depend largely on what the inspector discovers during the visual inspection, and the value of results obtained while carrying out sample testing. If any tests show significantly different results, then further testing may be required.

In some cases, up to 100% of the installation will need to be tested. Periodic testing can be dangerous, and due consideration should be given to safety.

Persons carrying out the testing must be competent and experienced in the type of installation being tested and the test instruments being used.

There is no set sequence for the testing which may be required for the completion of the periodic inspection report. The sequence and type of tests which are to be carried out are left to the person carrying out the test to decide upon. Where tests are required, the recommendations for these tests would be:

	Recommended tests
Continuity of protective conductors	Between the distribution board earth terminal and exposed conductive parts of current using equipment. Earth terminals of socket outlets (test to the fixing screw of outlet for convenience).
Continuity of bonding conductors	All main bonding and supplementary bonding conductors.
Ring circuit continuity	Only required where alterations or additions have been made to the ring circuit.
Insulation resistance	Only between live conductors joined and earthed. Or between live conductors with the functional switch open if testing lighting circuit.
Polarity	Live polarity tested at the origin of the installation. Socket outlets. At the end of radial circuits. Distribution boards.
Earth electrode resistance	Isolate installation and remove earthing conductor to avoid parallel paths.
Earth fault loop impedance	At the origin of the installation for Z_e. Distribution boards for the Z_e of that board. Socket outlets and at the end of radial circuits for Z_S.
Functional tests	RCD tests and manual operation of isolators, protective devices and switches.

Voltage drop in conductors

It is part of the inspection process to ensure that installed conductors have been correctly selected for current carrying capacity and voltage drop.
To check the suitability of the current carrying capacity it is simply a matter of looking at the installation method, and then checking on the current carrying capacity tables for the cable in Appendix 4 of BS 7671.

To ensure that the cable meets the voltage drop requirements is slightly more complex. A simple method is to measure the voltage at the origin of the circuit, and then measure the voltage at the end of the circuit with the load connected and switched on. The difference between the two measurements will be the volt drop.

If the first method is impractical, then a resistance test should be carried out between the phase and neutral of the circuit. This test is carried out using the same method as the $R_1 + R_2$ test although, instead of the test being between phase and CPC, it is between the phase and neutral for the circuit. Once the resistance $R_1 + R_n$ of the circuit has been measured it should be multiplied by the current that will flow in the circuit. This will give you the volt drop for the circuit.

Example

A circuit is wired in 2.5 mm^2 and is 25 metres in length. The current in the circuit is 18 amps.

The measured value of resistance is 0.37 Ω

$$\text{Voltage drop} = I \times R = V$$
$$18 \times 0.37 = 6.66 \text{ volts.}$$

This is the voltage drop for the circuit.

This page intentionally left blank

3

Testing of electrical installations

Safe isolation

It cannot be over-emphasized how important it is that isolation of electrical circuits is carried out in a set sequence, and that this sequence is repeated each time a circuit or complete installation is to be isolated.

If the same procedure is followed each time isolation is carried out, it will soon become a habit, which can only be a good thing as it may save your life.

It is vital that the correct test equipment is used for isolation and that it complies with the Health and Safety Executive document GS 38. This document gives guidance on the use of test equipment, particularly leads and probes.

The GS 38 document is not a statutory document but if the guidance given in the document is followed, it will normally be enough to comply with the Health and Safety at Work Act 1974, the Electricity at Work Regulations 1989 and any other statutory requirements that may apply. The items of equipment that should be available to persons carrying out the safe isolation procedure are:

- A proving unit

- An approval voltage indicator (left) and test Lamp (right)

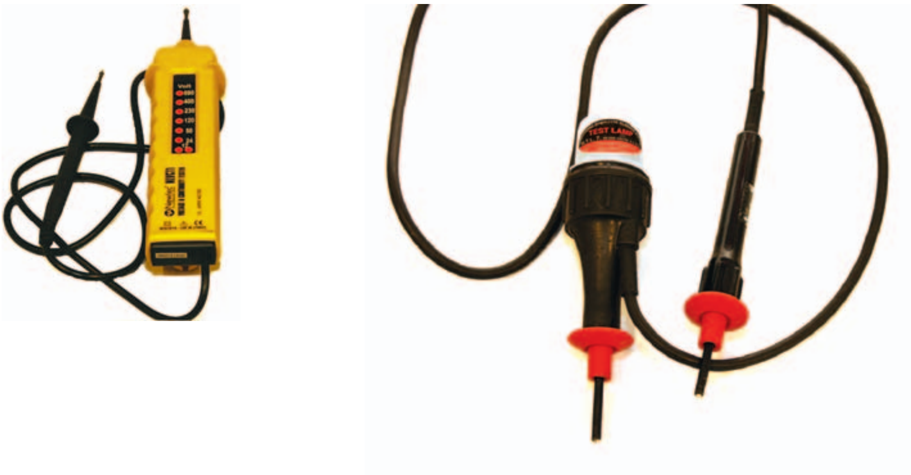

- Warning notices

DANGER
Do not switch on

- Locking devices

Another useful piece of equipment is:

- An R_1 and R_2 box. This will not only be useful for the safe isolation of socket outlets, it can also be used for ring circuit testing and the $R_1 + R_2$ testing of radial circuits incorporating a socket or socket outlets without having to remove them from the wall.

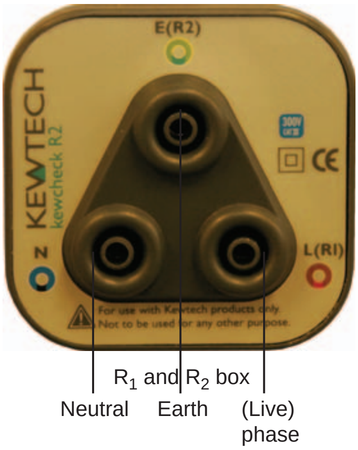
E(R2)
KEWTECH
kewcheck R2
N
L(RI)
For use with Kewtech products only
Not to be used for any other purpose.
R_1 and R_2 box
Neutral
Earth
(Live) phase

The leads should be:

- Flexible and long enough, but not too long.
- Insulated to suit the voltage at which they are to be used.
- Coloured where it is necessary to identify one lead from the other.
- Undamaged and sheathed to protect them against mechanical damage.

The probes should:

- Have a maximum of 4 mm exposed tip (preferably 2 mm).
- Be fused at 500 mA or have current limiting resistors.
- Have finger guards (*to stop fingers slipping on to live terminals*).
- Be colour identified.

Isolation procedure

It is very important to ensure that the circuit that you want to isolate is live before you start. To check this, a voltage indicator/test lamp or a piece of equipment that is already connected to the circuit should be used. If it appears that the circuit is already dead, you need to know why.

- Is somebody else working on it?
- Is the circuit faulty?
- Is it connected?
- Has there been a power cut?

You must make absolutely certain that you and you alone are in control of the circuit to be worked on. Providing the circuit is live you can proceed as follows:

STEP 1

Ensure voltage indicator/test lamp is working correctly.

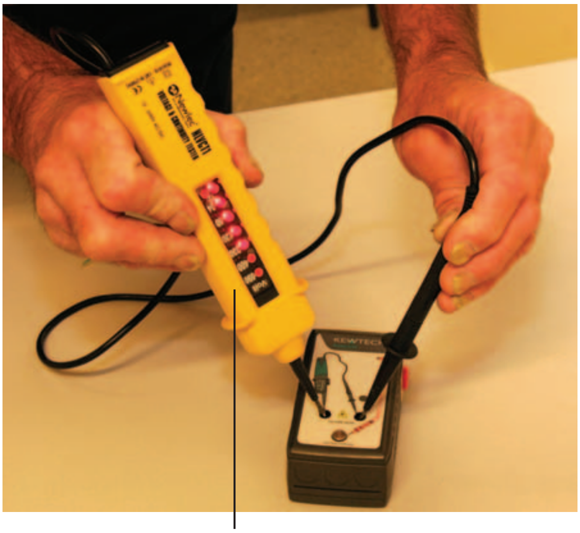
Voltage lights lit

STEP 2

Test between all live conductors and live conductors and earth.

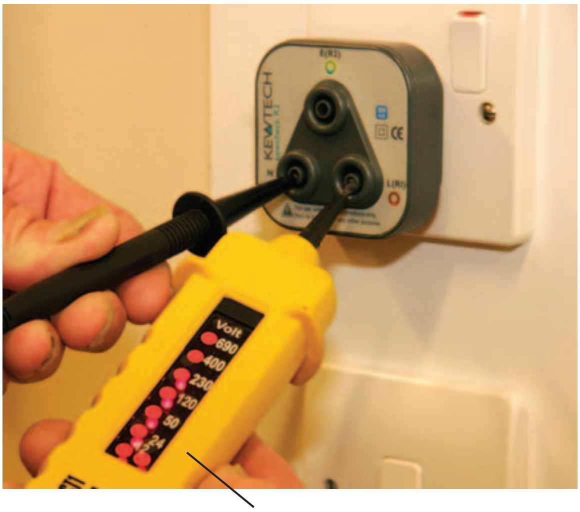

Voltage lights lit

STEP 3

Locate the point of isolation. Isolate and lock off.

Place warning notice (**DANGER ELECTRICIAN AT WORK**) at the point of isolation.

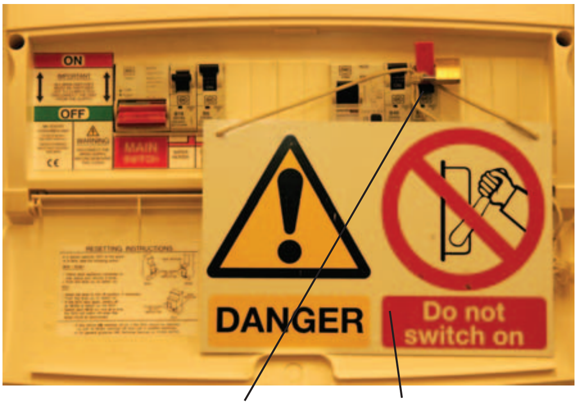
ON
OFF
MAIN SWITCH
DANGER
Do not switch on

Isolate and lock off

Place warning notice

STEP 4

Test circuit to prove that it is the correct circuit that you have isolated.

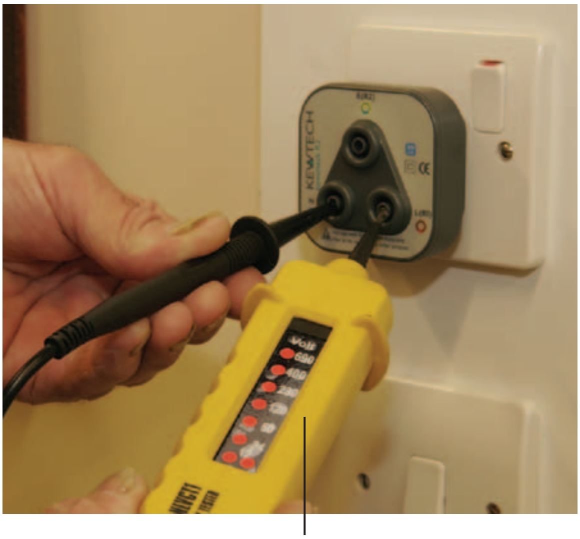
KEWTECH
Volt
690
400
230

No voltage lights lit

Be careful! Most test lamps will trip an RCD when testing between live and earth, it is better to use an approved voltage indicator to GS 38 as most of these do not trip RCDs

STEP 5

Check that the voltage indicator is working by testing it on a proving unit or a known live supply.

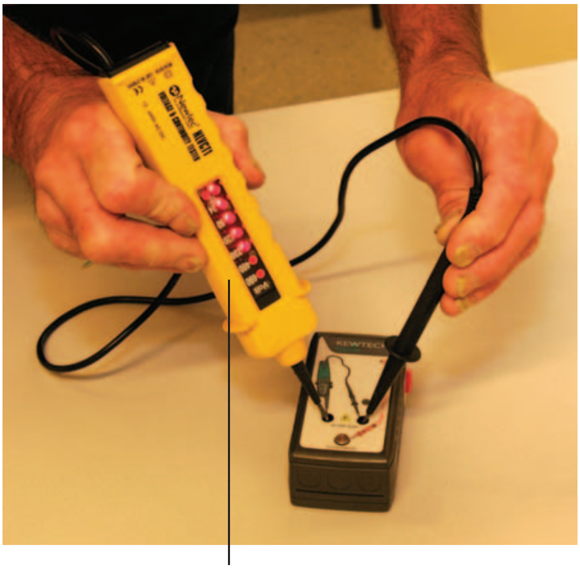

When carrying out the safe isolation procedure never assume anything, always follow the same procedure

It is now safe to begin work.

If the circuit which has been isolated is going to be disconnected at the consumer's unit or distribution board, **REMEMBER** the distribution board should also be isolated. The Electricity at Work Regulations 1989 do not permit live working.

Testing for continuity of protective conductors

Main equipotential bonding

This test is carried out to ensure that the equipotential bonding conductors are unbroken, and have a resistance low enough to ensure that, under fault conditions, a dangerous potential will not occur between earthed metalwork (*exposed conductive parts*) and other metalwork (*extraneous conductive parts*) in a building.

It is not the purpose of this test to ensure a good earth fault path but to ensure that, in the event of a fault, all exposed and extraneous conductive parts will be live at the same potential, hence EQUIPOTENTIAL bonding. In order to achieve this, it is recommended that the resistance of the bonding conductors does not exceed 0.05 Ω.

Table 54H of the *On-Site Guide* and Regulation 547-02-02 in BS 7671 cover the requirements of equipotential bonding. Table 10A of the *On-Site Guide* is also useful. Maximum lengths of copper bonding conductors before 0.05 Ω is exceeded.

Size mm^2	Length in metres
10	27
16	43
25	68
35	95

The test is carried out with a **Low Resistance Ohm meter** and often can only be carried out on the initial verification; this is because one end of the bonding conductor must be disconnected to avoid parallel paths. When disconnecting a bonding conductor, it is important that the installation is isolated from the supply. On larger installations it is often impossible to isolate the installation and, therefore, the conductor must remain in place. The instrument should be set on the lowest value of Ω possible.

STEP 1

Isolate supply (as safe isolation procedure)

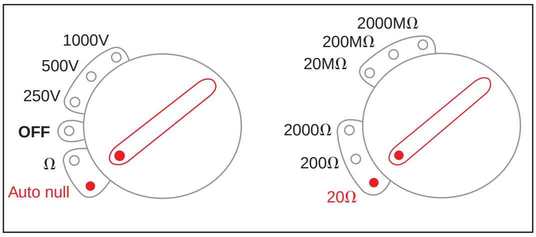
1000V
500V
250V
OFF
Ω
Auto null
2000MΩ
200MΩ
20MΩ
2000Ω
200Ω
20Ω

STEP 2

Disconnect one end of the conductor (*if possible, disconnect the conductor at the consumers unit, and test from the disconnected end and the metalwork close to the bonding conductor. This will test the integrity of the bonding clamp*).

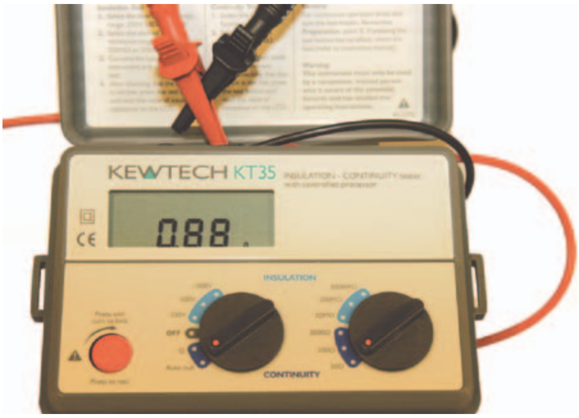
KEWTECH KT35
0.88.
INSULATION
CONTINUITY

Test leads

STEP 3

Measure the resistance of test leads or null leads (*these may be long as the only way that we can measure a bonding conductor is from end to end*).

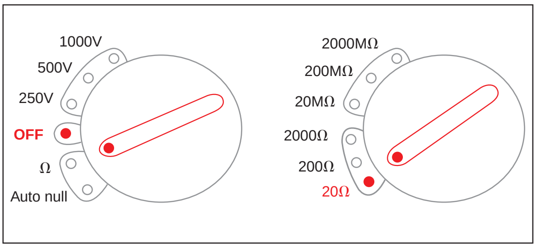
1000V
500V
250V
OFF
Ω
Auto null
2000MΩ
200MΩ
20MΩ
2000Ω
200Ω
20Ω

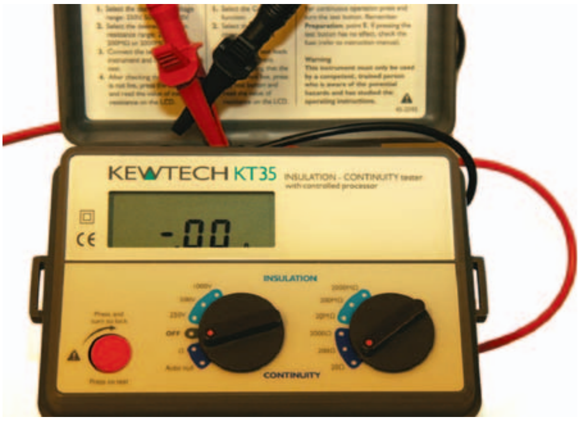
KEWTECH KT35
-0.0.
INSULATION
CONTINUITY

Nulled leads

STEP 4

Connect one test lead to the disconnected conductor at the consumer's unit.

Note: Safety notice removed for clarity.

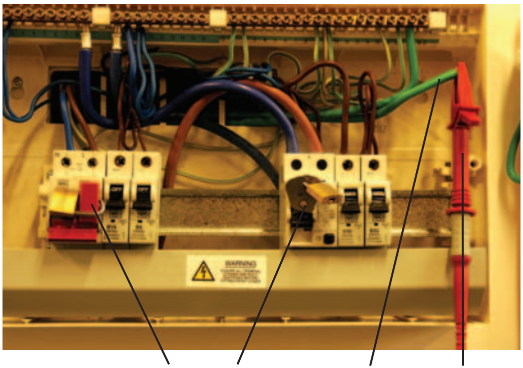
Isolated and locked off
Disconnected conductor
Test lead

STEP 5

Connect the other end of the test lead to the metalwork that has been bonded (connecting the lead to the metalwork and not the bonding clamp will prove the integrity of the clamp).

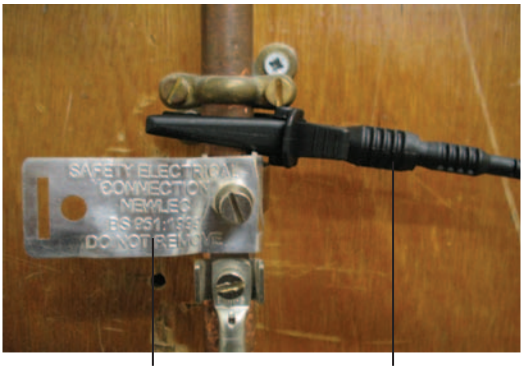
Bended metalwork
Other end of test lead

STEP 6

If the instrument is not nulled remember to subtract the resistance of the test leads from the total resistance. This will give you the resistance of the bonding conductor. If the meter you are using has been nulled, the reading shown will be the resistance of the conductor.

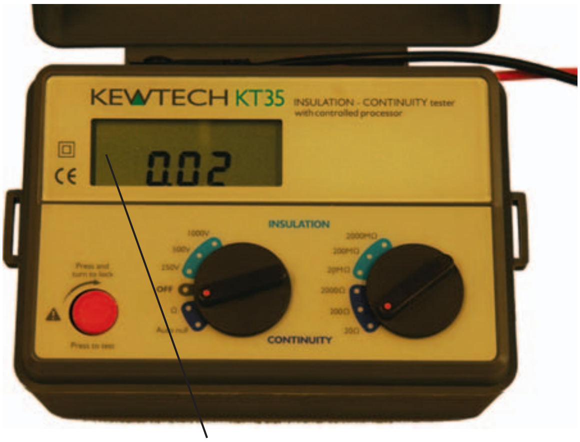
KEWTECH KT35
0.02
Very low value–less than 0.05A

STEP 7

Ensure that the bonding conductor is reconnected on completion of the test.

Whilst carrying out this test a visual inspection can be made to ensure that the correct type of BS 951 earth clamp, complete with label is present, and that the bonding conductor has not been cut if it is bonding more than one service.

If the installation cannot be isolated on a periodic inspection and test, it is still a good idea to carry out the test; the resistance should be a maximum of 0.05Ω as any parallel paths will make the resistance lower. If the resistance is greater than 0.05Ω the bonding should be reported as unsatisfactory and requires improvement.

In some instances the equipotential bonding conductor will be visible for its entire length; if this is the case, a visual inspection would be acceptable, although consideration must be given to its length.

For recording purposes on inspection and test certificates no value is required but verification of its size and suitability is.

Items to be bonded would include any incoming services, such as: water main, gas main, oil supply pipe, LPG supply pipe. Also included would be structural steel work, central heating system, air conditioning, and lightning conductors within an installation (*before bonding a lightning conductor it is advisable to seek advice from a specialist*).

This is not a concise list and consideration should be given to bonding any metalwork that could introduce a potential within a building.

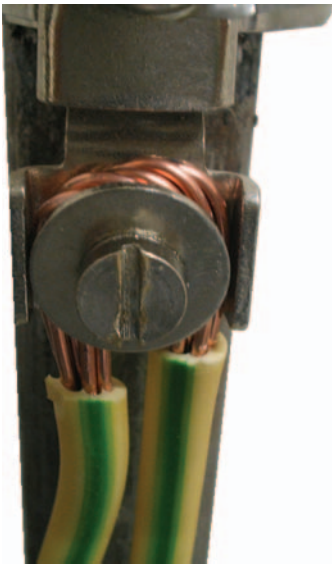

✓ Correct

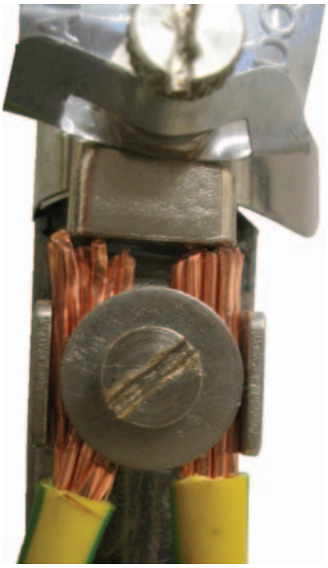

✗ Incorrect

Continuity of supplementary bonding

There are two general reasons for carrying out supplementary equipotential bonding.

Supplementary bonding 1

This is required when there is an increased risk of electric shock (Regulation 471-08-01).

BS 7671 states that supplementary bonding must be installed in bathrooms/shower rooms and swimming pools.

It should be remembered that, although Regulation 601-04-01 requires that we must bond the exposed and extraneous conductive parts in bathrooms, this Regulation applies only to zones 1, 2, and 3.

Due consideration must be given to other areas where there is an increased risk of electric shock. There is no specific requirement to carry out supplementary bonding in Kitchens. However, if it is thought by the installer that there is an increased risk of electric shock, there is no reason why bonding could not be carried out, it will do no harm providing it is carried out correctly.

On occasions it is often useful to carry out supplementary bonding, particularly under Kitchen sinks. This may not be for electrical reasons, more for visual purposes – bonding is not well understood by many people. A possible scenario might be where you may have travelled 20/30 miles to fit a Kitchen and completed everything to comply with the required regulations. A few days later, before you have been paid for the work you receive a phone call from your customer, informing you that his next door neighbour has spotted that you have not bonded the sink. Of course your customer will believe his neighbour is right and that you have forgotten something or tried to save a bit of money. The choice is now yours, do you try and convince your customer that his neighbour is wrong or do you travel back to the job to carry out the bonding to ensure payment? Perhaps for the sake of a couple of earth clamps and a short length of 4 mm^2, it would have been cheaper just to bond it in the first place.

Where complimentary supplementary bonding is used in this instance a test must be carried out to ensure that the resistance between exposed and extraneous conductive parts is in place and has a resistance of less than 0.05 Ω. The instrument to be used is a low resistance ohm meter.

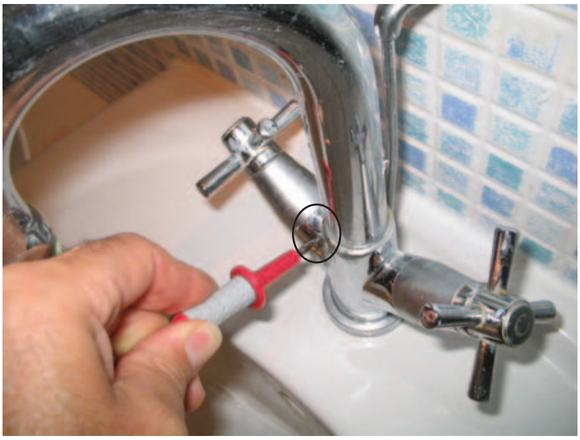

Probe on solid metal part of tap

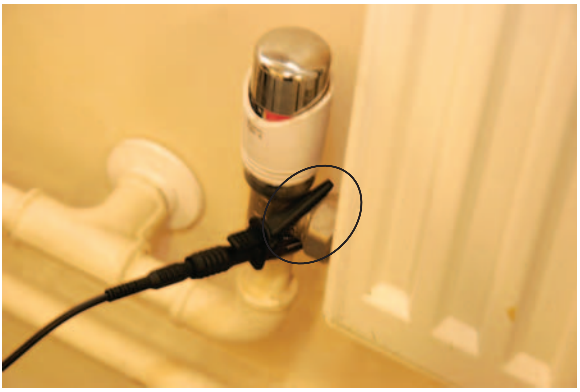

Probe on unpainted metal work

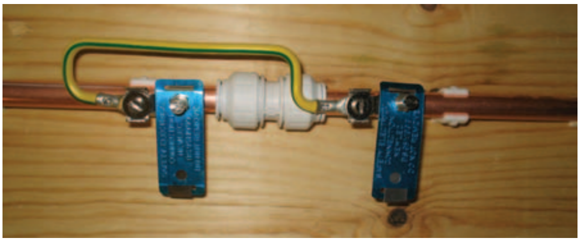

Bonding conductor

A visual check must be made to ensure that the correct earth clamps have been used and that they have the correct labels attached.

It is perfectly acceptable to use the pipe work and structural steelwork within the area as a bonding conductor, and bonding can be carried out adjacent to the area providing that the integrity of the pipe work/steelwork can be assured. An airing cupboard would be a good example of a suitable place to bond.

If it is necessary for the lighting point or electric shower in a bathroom (*remember if they are not within zones 1, 2, or 3 no bonding is required*), there is no reason why the bonding conductor could not be simply attached to a pipe within the roof space, which is bonded elsewhere and passes near the item which requires bonding. The correct bonding clamps and labels should always be used.

If using pipe work of plumbing and heating systems as bonding, continuity of the pipe work must be verified. The resistance values are the same as if copper cables were used. Tests must be made between exposed and extraneous conductive parts to ensure that the resistance does not exceed 0.05Ω. This is a simple test carried out using a low resistance ohm meter. A probe of one lead should be placed on one metal part and the probe of the other lead placed on an adjacent metal part. The resistance must be no greater than 0.05Ω.

Problems can arise if the pipe work is altered and plastic push fittings are used. Clearly these will not conduct and the bonding continuity could be compromised. If ever a plastic plumbing fitting is used on copper pipe work, consideration should be given to the installation of a bonding conductor installed across the fitting.

It is a common belief that water in pipe work will conduct; in fact, the current that would flow through water in a 15 mm diameter pipe which has a plastic joint in it is very small.

To find out just how much, I set up a simple controlled experiment, two short lengths of 15 mm pipe were joined using a 15 mm plastic push fit coupler. The pipe was then filled with water and the two ends of the joined pipe were connected to a 230 volt supply. The current flowing was measured to be 0.003 Amperes (3 mA). The current flow would increase if the water had central heating additives in it, but not considerably.

In bathrooms/shower rooms where the plumbing has been carried out using plastic pipe, the pipe work does not need supplementary bonding; however, it should be remembered that electrical appliances and any extraneous/exposed conductive parts within zones 1, 2 and 3 must still be bonded. Figure 4d and 4e in section 4 of the *On-Site Guide* are good places of reference for this.

For the sizing of supplementary bonding Table 10b in Appendix 10 of the *On-Site Guide* should be used. As a general rule, supplementary bonding within a bathroom should be 2.5 mm^2 if mechanically protected or 4 mm^2 if not. It is usually easier to use 4 mm^2 to save the trouble of mechanically protecting the bonding conductor (see Regulation 547-03 of BS 7671).

Supplementary bonding 2

Used where, due to various circumstances, the required disconnection time cannot be met by conventional methods, and where it is not desirable to use a residual current device. Supplementary equipotential bonding may be used as a means of compliance (Regulation 413-02-04 (i)).

Wherever supplementary bonding is used in this instance it must be tested to ensure that the resistance between two parts is less than the value R obtained by the following formulae:

$$R = \frac{50}{I_a}$$

R = maximum resistance of bonding conductor; 50 = the safe touch voltage (*this may be 25 volts in special locations*); I_a = fault current needed to operate the protective device within 5 seconds (Regulation 413-02-28).

If the circuit is protected by a Residual current device I_a may be substituted by $I\Delta_n$.

The fault current can be found in Appendix 3 in BS 7671 or by using the calculation:

$$I_a = \frac{U_{OC}}{Z_S}$$

U_{OC} = open circuit voltage of supply transformer; Z_S = maximum earth loop impedance for the protective device.

The test should be carried out between exposed and extraneous conductive parts, using a low resistance ohm meter with long leads to find out the resistance between them. If the resistance is higher than that required, then supplementary bonding is required.

Exposed conductive parts

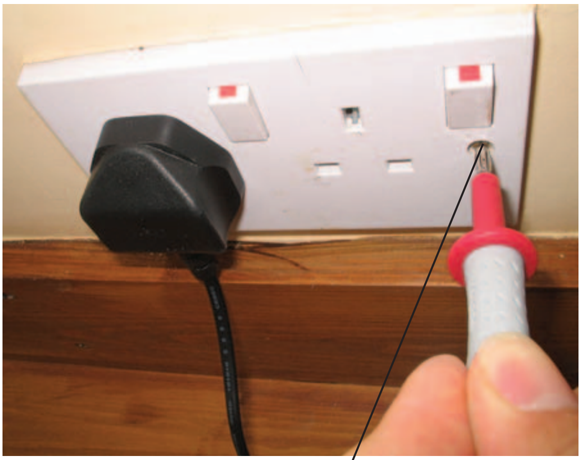

Screw fixings are earthed

Extraneous conductive parts

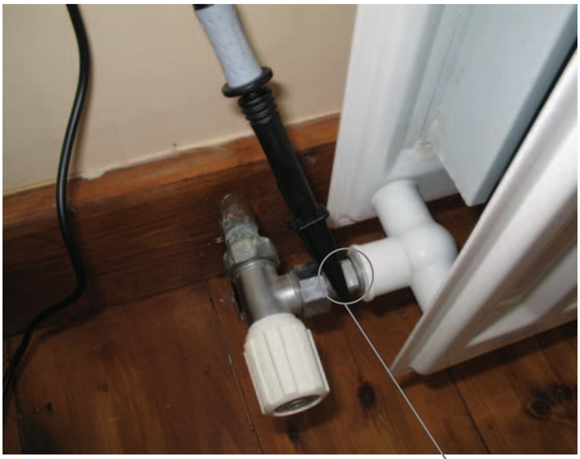

Unpainted metal work

Example 1

A circuit on a TN system is to be altered. The current carrying capacity and the volt drop of the cable are adequate for the load. However, the Z_S value is too high for the 20A BS 1361 protective device which is being used to protect the circuit.

The maximum resistance permissible between exposed and extraneous conductive parts must be calculated.

The first step is to find the current that would cause automatic disconnection of the supply.

$$I_a = \frac{U_{OC}}{Z_S}$$

$$I_a = \frac{240}{2.93} = 81.91\,amps$$

This value can also be found in Figure 3.1 of Appendix 3 of BS 7671 (it is rounded up to 82 amps).

The maximum permissible resistance between conductive parts can now be found by:

$$R = \frac{50}{I_a}$$

$$R = \frac{50}{82} = 0.6\Omega$$

Supplementary bonding is installed where there is a risk of simultaneous contact with any extraneous and exposed conductive parts. Its purpose is to ensure that the potential between any of these parts does not rise above a safe value. In most cases, this value is 50 volts, although some chapters in Part 6 of BS 7671 require a maximum potential of only 25 volts.

Determining if a part is extraneous, or just a piece of metal

A test should be made using an insulation resistance tester set on mΩ, supplying 500 volts.

Connect one test lead to the metal part and the other lead to a known earth. If the resistance value is 0.02 mΩ (20,000 Ω) or greater, no supplementary bonding is required. If less than 0.02 mΩ, supplementary bonding should be carried out.

If we use Ohm's law we can see how this works:

$$\frac{V}{R} = I: \qquad \frac{500}{20{,}000} = 0.025\text{ A}$$

This shows that a current of 25 mA would flow between the conductive parts; this would of course only be 0.012 amp if the fault was on a single phase 230 volt supply. This current is unlikely to give a fatal electrical shock.

The test must not be confused with a continuity test. It is important that an insulation resistance tester is used.

Continuity of circuit protective conductors

This test is carried out to ensure that the CPC of radial circuits are intact and connected throughout the circuit. The instrument used for this test is a **low resistance ohm meter** set on the lowest value possible.

This is a dead test and must be carried out on an isolated circuit.

Testing can be carried out using two methods.

Method one

STEP 1

Using a short lead with a crocodile clip on each end, bridge phase and CPC together at one end of the circuit (*it does not matter which end, although it is often easier to connect at the distribution board as this will certainly be one end of the circuit*).

The resistance of this lead plus the resistance of the test leads should be subtracted, or the instrument nulled before the $R_1 + R_2$ reading is recorded.

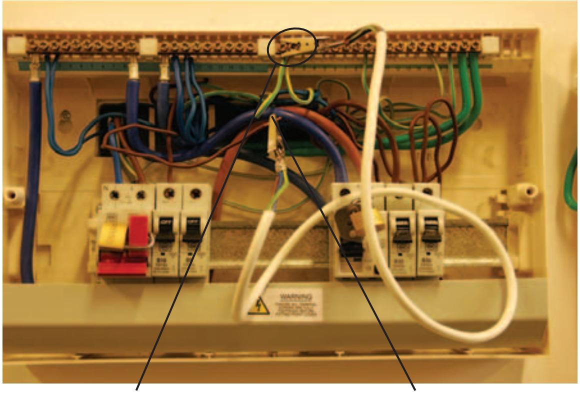
Connect to main earth terminal
Connect to phase conductor

STEP 2

At **each** point on the circuit test between phase and CPC.

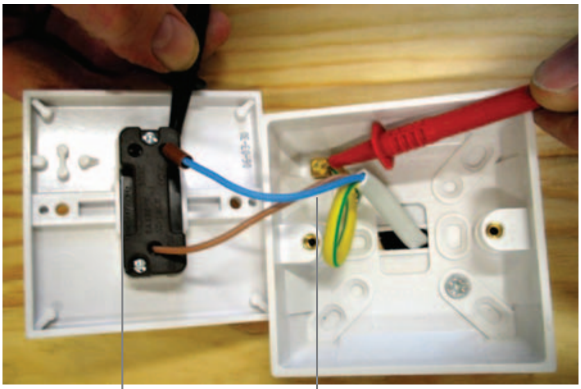
Terminal of switch return
Earthing terminal

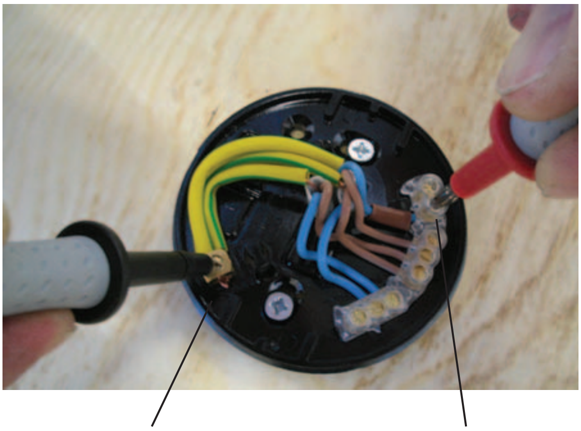
Earthing terminal
Switch return

Keep a note of the readings as you carry out the test, they should increase as you move further from the connected ends. The highest reading obtained should be at the furthest end of the circuit and will be the $R_1 + R_2$ of the circuit. This value should be recorded on the schedule of test results. If the highest reading is obtained at a point which is not the furthest from the circuit, further investigation should be carried out as it may indicate a loose connection (high resistance joint).

In some instances only the value of R_2 may be required. Where the phase conductor is the same size as the CPC the total measured resistance can be divided by 2 as the phase and CPC resistance will be the same. If the phase and CPC are of different sizes (this is usual in twin and earth thermoplastic cable) the R_2 value can be calculated using the following formula:

$$R_2 = R_1 + R_2 \times \frac{A\ phase}{A\ CPC + A\ phase}$$

R_2 = resistance of CPC in ohms; $R_1 + R_2$ = measured value of resistance in ohms; *A phase* = Area of phase conductor in mm^2; *A CPC* = Area of CPC mm^2.

Example 2

A radial circuit is wired in 2.5 mm^2 phase and 1.5 mm^2 CPC. The test resistance of $R_1 + R_2$ is 0.37 Ω.

To calculate the resistance of the CPC on its own:

$$R_2 = 0.37 \times \frac{2.5}{2.5 + 1.5}$$
$$R_2 = 0.37 \times 0.625 = 0.23\Omega$$

If the CPC is smaller than the phase conductor, the resistance of the CPC conductor will always be greater than the phase conductor as it has a smaller cross-sectional area.

Another method of determining R_2 is described in the ring circuit test.

Method two

This method will prove CPC continuity and is usually only used where the circuit is wired in steel enclosures where parallel paths to the CPC may be present and the $R_1 + R_2$ value would not be a true value. Or where the CPC resistance is required for use with Table 41C of the Wiring Regulations.

This method uses a long lead. One end is connected to the earth terminal of the distribution board, and the other connected to a **low resistance ohm meter**. The short lead of the ohm meter is then touched onto each fitting to ensure that it is connected to the CPC. The highest reading minus the resistance of the leads can be recorded as the R_2 reading.

If the furthest point of the circuit is known, and no parallel paths exist, the $R_1 + R_2$ reading can be carried out first using **Method one**, and then a test between earthed metal at each point can be made to ensure that the CPC is connected to each point on the circuit, using **Method two**. This method is particularly useful where there are a lot of enclosed metal fittings and dismantling them would be impractical.

Ring final circuit test

The purpose of this test is to ensure that:

- The cables form a complete ring.
- There are no interconnections.
- The polarity is correct on all socket outlets.

When this test is carried out correctly it also gives you the R_1 and R_2 value of the ring and identifies spurs.

Table 8A in the *On-Site Guide* provides information on final circuits for socket outlets. This table states that a ring circuit is to be wired in 2.5 mm^2 phase conductor and 1.5 mm^2 CPC as a minimum size. This type of circuit is an A1 ring and should be protected by a 30/32 amp overcurrent protective device.

Complete ring circuit

A test must be carried out on the conductors to verify that they form a complete loop. If it is found that they do not, overloading of the cables could occur. In installations where more than one ring circuit has been installed, it is possible for the ends of the ring to become muddled, resulting in the circuits being supplied through two protective devices.

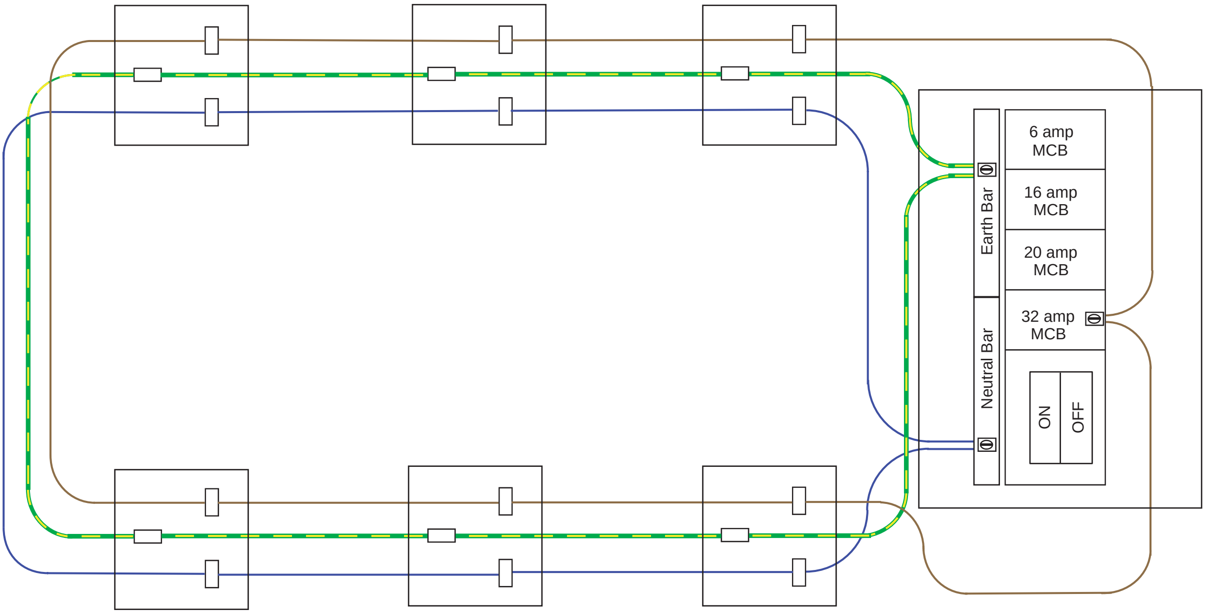

Complete ring circuit

The whole point of a ring circuit is that it can be wired in small CSA cables but carry a reasonably high current, this is because we have two 2.5 mm^2 cables wired in parallel (*Regulation 473-01-06*). If we look at Table 4D5A in BS 7671, the value of current that 2.5 mm^2 cable can carry is 20 amps in the worst type of conditions.

If we use two of these conductors in parallel, we will have a total current carrying capacity of 40 amps. As one of the jobs of the protective device is to protect the cable, this situation will be fine because the protective device is smaller than the total current carrying capacity of the cables in parallel.

Broken ring circuit

If, however, we found the ring to be broken, the protective device could not do its job as it is rated at 32amps and the cable is rated only at 20amps. Hence overloading!

Interconnections

Occasionally a situation will be found where there is a ring within a ring, in other words the ring is interconnected.

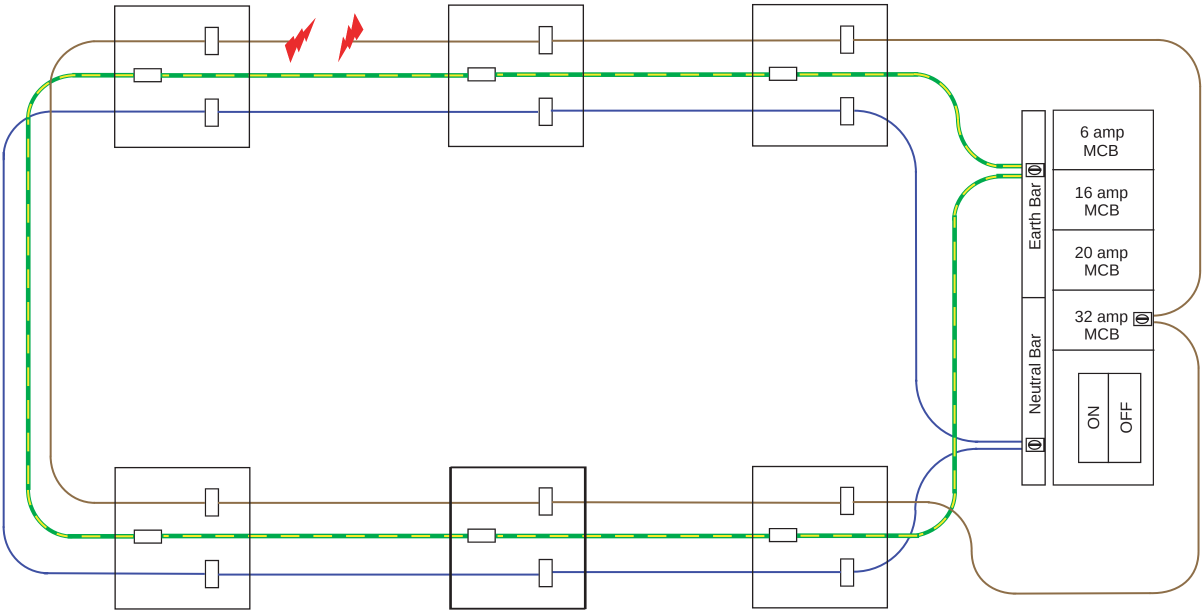
6 amp
MCB
16 amp
MCB
20 amp
MCB
32 amp
MCB
Earth Bar
Neutral Bar
ON
OFF

Broken ring circuit

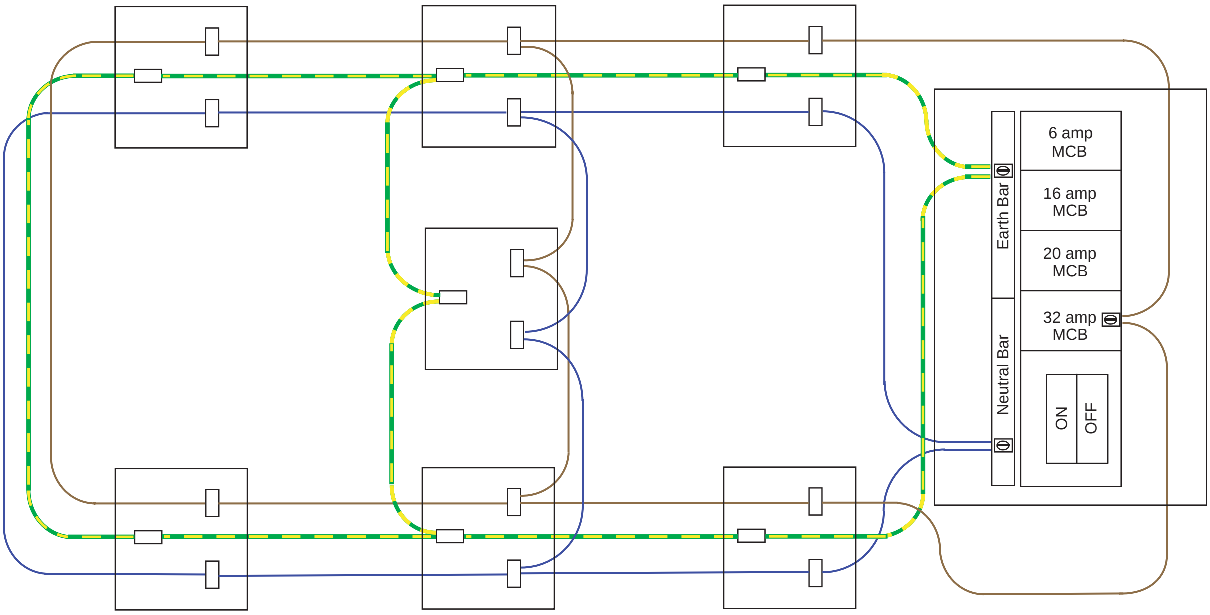
6 amp
MCB
16 amp
MCB
20 amp
MCB
32 amp
MCB
Earth Bar
Neutral Bar
ON
OFF

Interconnected ring circuit

This situation, as it is, will not present a danger. However, it will make it very difficult for a ring final circuit test to be carried out as, even if the correct ends of the ring are connected together, different values will be found at various points of the ring. If one loop is broken, a test at the consumer's unit will still show a complete ring. It will not be until further tests are performed that the interconnection/broken loop will be found.

Polarity

Each socket outlet must be checked to ensure that the conductors are connected into the correct terminals. Clearly if they are not, serious danger could occur when appliances are plugged in.

It could be that phase and neutral are the wrong polarity; the result of this is that the neutral would be switched in any piece of equipment with a single pole operating switch.

If the live conductors and CPC are connected with reverse polarity, then the case of any Class 1 equipment could become live and result in a fatal electric shock.

Performing the test

The instrument required is a low resistance ohm meter set on the lowest scale, typically 20 Ω. Be sure to zero the instrument or subtract the resistance of the leads each time you take a reading.

This is a dead test! Safe isolation must be carried out before working on this circuit

STEP 1

Isolate circuit to be tested.

STEP 2

Identify legs of ring.

STEP 3

Test between ends of phase conductor and note the resistance value.

Instrument set to Ω for whole test.

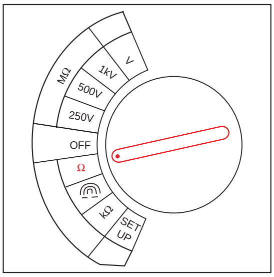
V
1kV
500V
250V
MΩ
OFF
Ω
kΩ
SET UP

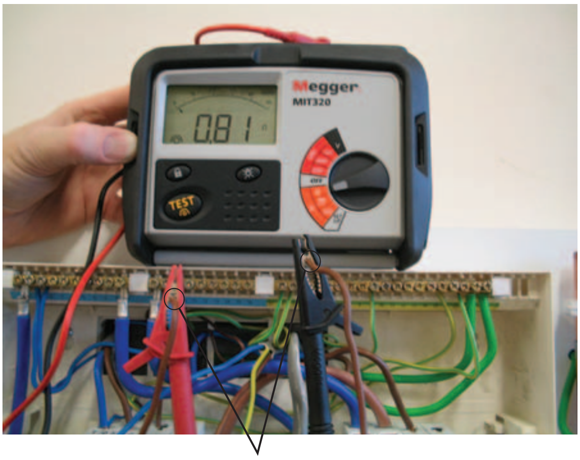
Megger
MIT320
0.81
TEST

Ends of phase conductor

STEP 4

Test between ends of neutral conductor. This value should be the same as the phase conductor resistance as the conductor must be the same size (*see Note 1*).

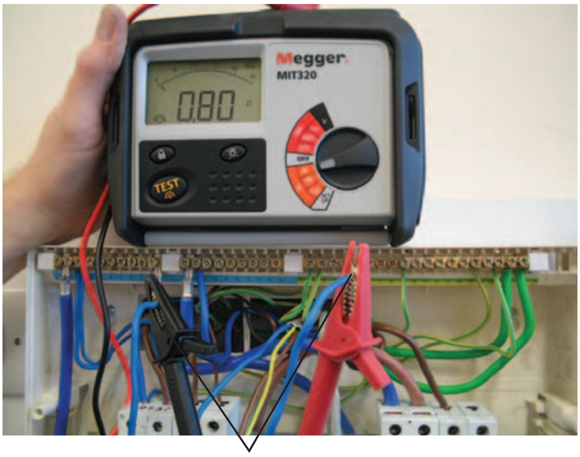
Megger
MIT320
0.80
TEST

Ends of neutral conductor

STEP 5

Test between the ends of the CPCs. If the conductor size is smaller than the live conductors (*as is usually the case when using twin and earth cable*), the resistance value will be higher (*see Note 2*); make a note of this reading.

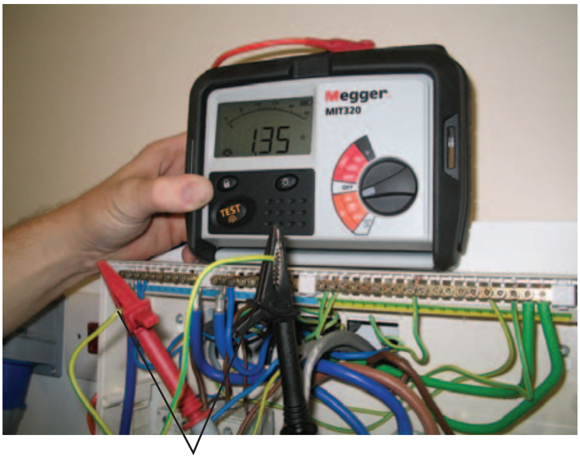
Megger
MIT320
1.35
TEST
Ends of CPCs

STEP 6

Join P of leg 1 to N of leg 2.

Test between N of leg 1 and P of leg 2. The measured resistance should be double that of the phase conductor.

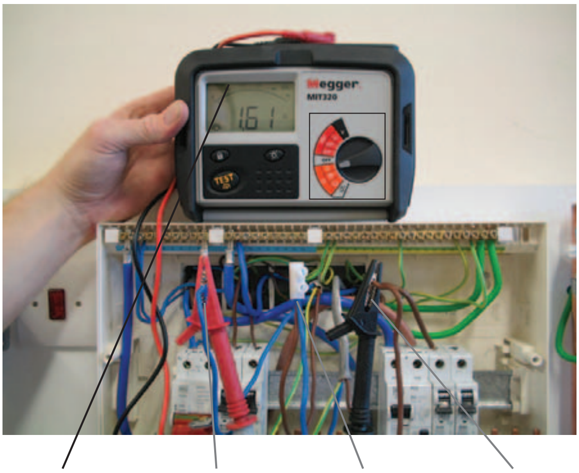
Megger
MIT320
1.61
TEST
Resistance double that of phase conductor
P2
P1 joined to N2
N1

STEP 7

Join N of leg 1 to P of leg 2 together (*leaving N2 and P1 joined*).

Test between joined ends.

The measured value should be $\frac{1}{4}$ of test between N of leg 1 and P of leg 2.

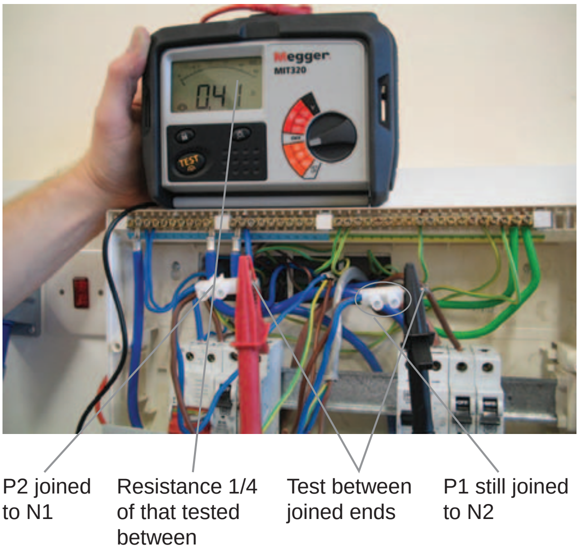
P2 joined to N1
Resistance 1/4 of that tested between N1 and P2
Test between joined ends
P1 still joined to N2

STEP 8

Leave the ends joined.

Test between P and N at each socket outlet, the resistance should be the same at each socket (*see Note 1*).

A higher reading should be investigated, although it will probably be a spur it should be checked as it may be a loose connection (*high resistance joint*).

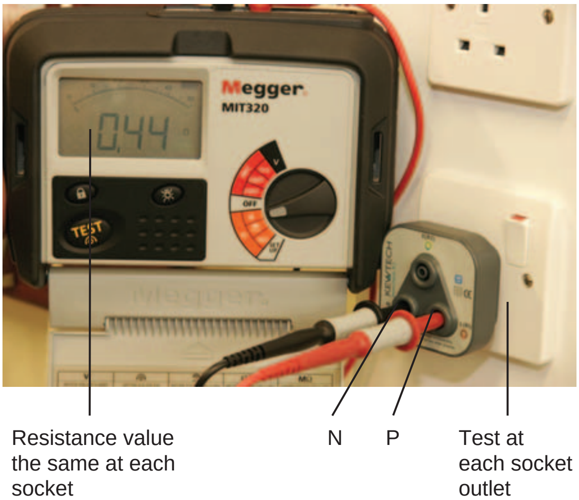
Resistance value the same at each socket
N
P
Test at each socket outlet

STEP 9

Disconnect the ends and repeat the test using phase and CPC conductors (*see Note 3*).

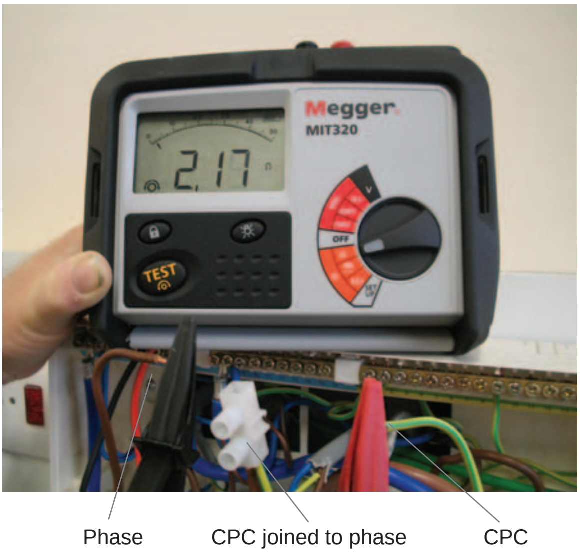
Megger
MIT320
2.17
TEST
OFF
Phase
CPC joined to phase
CPC

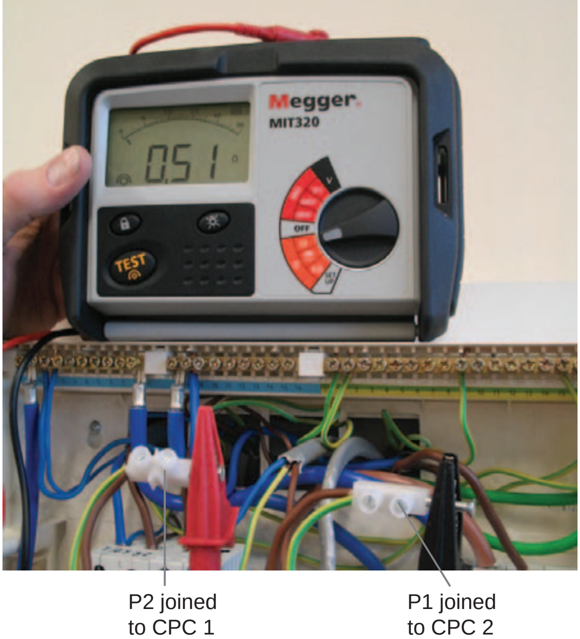
Megger
MIT320
0.51
TEST
OFF
P2 joined
to CPC 1
P1 joined
to CPC 2

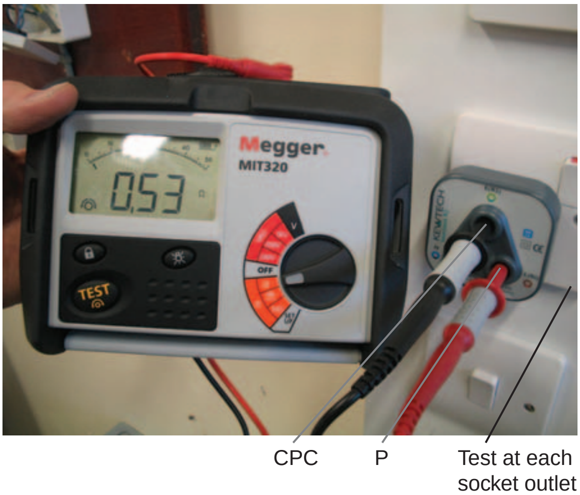
Megger
MIT320
0.53
TEST
OFF
KEWTECH
CPC
P
Test at each
socket outlet

The highest value (*which will be the spur*) will be the R_1 and R_2 value for this circuit.

Notes

1. If with ends connected (P1/N2 and P2/N1) a substantially different resistance value is measured at each socket outlet, check that the correct ends of ring are connected. A difference of 0.05 Ω higher or lower would be acceptable.
2. In a twin and earth cable the CPC will usually have a resistance of 1.67 times that of the phase conductor as it has a smaller cross-sectional area.
3. When phase and CPC conductors are not the same size a higher resistance value will be measured between Phase and CPC than Phase and neutral. It will also alter **slightly** as the measurement is taken around the ring, the resistance will be lower nearer the joined ends and will increase towards the centre of the ring. The centre socket of the ring will have the same resistance value as the test between the joined ends.
4. If the circuit is contained in steel conduit or trunking parallel paths may be present, this would result in much lower $R_1 + R_2$ resistance values.
5. Some certificates may require r_n to be documented. This is the resistance of the neutral loop measured from end to end.

Example 3

Let's use a 2.5/1.5 mm^2 twin and earth cable 22 metres long. If we look at Table 9A in the *On-Site Guide* we will see that the resistance of a copper 2.5 mm^2 conductor has a resistance of 7.41 mΩ per metre.

The resistance of the phase conductor will be:

$$\frac{7.41 \times 22}{1000} = 0.163\Omega$$

Divide the largest conductor by the smallest to find the ratio of the conductors (how much bigger is the larger conductor?).

$$\frac{2.5}{1.5} = 1.67$$

The 2.5 mm^2 conductor is 1.67 × larger than the 1.5 mm^2 conductor; therefore, it must have 1.67 × less resistance than the 1.5 mm^2 conductor.

If we now multiply the resistance of the phase conductor by 1.67:
$0.163 \times 1.67 = 0.27\,\Omega$ *this is the resistance of the* $1.5\,\text{mm}^2$ *conductor.*

We can check this by looking at Table 9A of the On-Site Guide once again, and we can see that the resistance of $1.5\,\text{mm}^2$ *copper is* $12.10\,\text{m}\Omega$ *per metre. Therefore, 22 metres of* $1.5\,\text{mm}^2$ *copper will be:*

$$\frac{22 \times 12.10}{1000} = 0.266\Omega$$

As a final check, if we look at Table 9A of the On-Site Guide for the resistance of a $2.5\,\text{mm}^2/1.5\,\text{mm}^2$ *cable, we will see that it has a resistance of* $19.51\,\text{m}\Omega$ *per metre, and that 22 metres of it will have a resistance of:*

$$\frac{22 \times 19.51}{1000} = 0.429\Omega$$

The resistance value of the $2.5\,\text{mm}^2$ *is* $0.163\,\Omega$*; and the resistance value of the* $1.5\,\text{mm}^2$ *is* $0.266\,\Omega$*.*

If we add them together: $0.163\Omega + 0.266\Omega = 0.429\Omega$*. Finally,* 0.429Ω *is the resistance of our* $2.5\,\text{mm}^2/1.5\,\text{mm}^2$ *measured as one cable.*

Insulation resistance test

This is a test that can be carried out on a complete installation or a single circuit, whichever is suitable or required. The test is necessary to find out if there is likely to be any leakage of current through the insulated parts of the installation. A leakage could occur for various reasons.

A good way to think of this test is to relate it to a pressure test – we know that voltage is the pressure where the current is located in a cable. On a low voltage circuit, the expected voltage would be around 230 V a.c. The voltage used in an insulation test on a 230 V circuit is 500 V, which is more than double the normal circuit voltage. Therefore, it can be seen as a pressure test similar to a plumber pressure testing the central heating pipes.

Low insulation resistance

Cable insulation could deteriorate through age. A low insulation resistance caused through age will often be found in installations where rubber-insulated cables have been used. Cables which are crushed under floor boards, clipped on edge, or worn thin where pulled through holes in joists next to other cables, can give a very low reading.

Low insulation resistance could be found if a building has been unused for a period of time, due to the installation being affected by dampness in the accessories. Low insulation resistance readings will also often be found where a building has been recently plastered. In theory, long lengths of cables or circuits in parallel could give low readings due to the amount of insulation (*the longer the circuit or the more circuits, the more insulation there will be for leakage to occur*).

The instrument used to carry out this test is an **insulation resistance tester**. To comply with the requirements of the Health and Safety Executive the instrument must be capable of delivering a current of 1 mA when a voltage of 500 V.d.c is applied to a resistance of 0.5 MΩ. Table 71A in BS 7671 gives the test voltages and minimum acceptable resistance values.

The values are shown here:

	Circuits between 0 V and 50 V a.c.	Circuits between 50 V a.c. and 500 V a.c.	Circuits between 500 V and 1000 V a.c.
Required test voltage	250 V d.c.	500 V d.c.	1000 V d.c.
Minimum acceptable	0.25 MΩ	0.5 MΩ	1 MΩ
In 17th Edition of the Wiring Regulations	(0.5 MΩ)	(1 MΩ)	(2 MΩ)

Domestic installations

Remember that testing should be carried out from the day the installation commences (Regulation 711-01-01).

Testing a whole installation

In new domestic installations it is often easier to carry out insulation resistance testing on the whole of the installation from the meter tails before they are connected to the supply. If this is the preferred choice the test should be carried out as follows:

- Safe isolation must be carried out before commencing this test.
- Inform any occupants of the building that testing is to be carried out.
- Ensure that all protective devices are in place and switched on.
- Remove all lamps from fittings where accessible.
- If the lamps are not accessible or if a luminaire with control gear (fluorescent) is connected, open the switch controlling the luminaire (*Note 6*).

The same applies to extra low voltage transformers.

- Where dimmer switches are fitted it is important that they are either removed and the switch wires joined, or that the switch is bypassed (*Note 7*).
- Any accessories with indicator lamps are switched off (*Note 8*).
- Passive infra red detectors (PIRs) are removed or bypassed (*Note 9*).
- All fixed equipment such as cookers, immersion heaters, boilers and television amplifiers are isolated.
- Shaver sockets are disconnected or isolated (*Note 9*).
- Items of portable equipment are unplugged.

Great care must be taken as, during this test, 500 V will be passed through any electrical equipment which is left connected. This could damage the equipment or, at the very least, cause low readings to be obtained during the test. Once all precautions have been taken proceed with the test as follows:

Notes

1. The control equipment within discharge lamps will cause very low readings. It is quite acceptable to isolate the fitting by turning off the switch. This is more desirable than disconnecting the fitting. After the test between live conductors is completed the control switch for the luminaire should be closed before carrying out the test between live conductors and CPC. This is to ensure that all live conductors are tested for insulation resistance to earth.
2. Most dimmer switches have electronic components in them and these could be damaged if 500 V were to be applied to them. It is important that wherever possible the dimmer switches are removed and the phase and switch return are joined together for the test.
3. Neon indicator lamps will be recognized as a load by the test instrument and will give a very low insulation value. All that is required is for the switch on the accessory to be turned off.
4. Passive infra red detectors will give very low readings and may be damaged by the test voltage. Either disconnect it or test between live conductors and earth only on circuits containing PIRs. The same applies to shaver sockets.

STEP 1 Set insulation resistance tester to 500 V.

Some instruments have settings for Meg ohms and some are self ranging. If yours requires setting then 200 MΩ or higher is the setting to use.

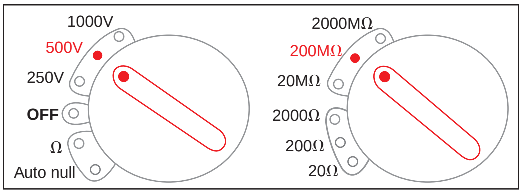
1000V
500V
250V
OFF
Ω
Auto null
2000MΩ
200MΩ
20MΩ
2000Ω
200Ω
20Ω

STEP 2

To ensure that the test results are accurate it is important to ensure correct operation of instrument and the integrity of the leads. Push the test button with the leads disconnected. The resistance shown on screen should be the highest that the instrument can measure.

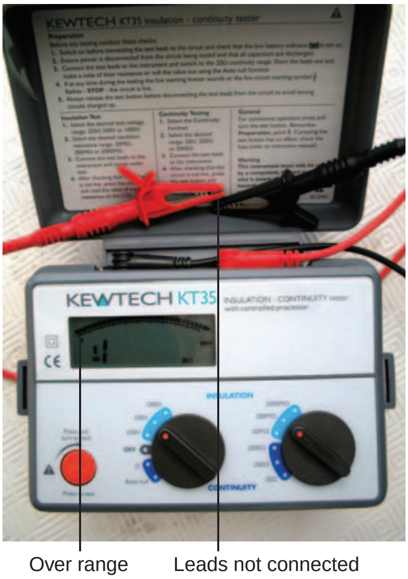
KEWTECH KT35
Over range
Leads not connected

STEP 3

Join leads and operate instrument again, the resistance shown on screen should be the lowest value possible (0.0 mΩ) in all cases.

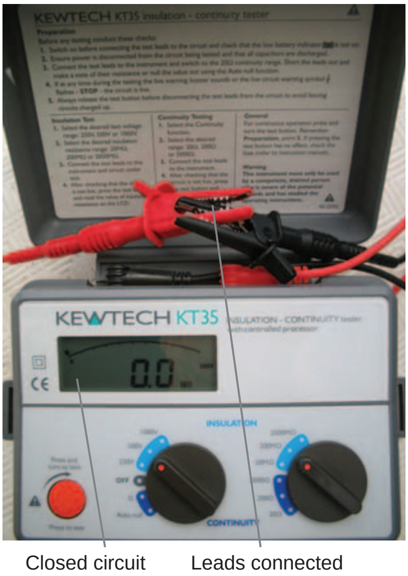
KEWTECH KT35
0.0
Closed circuit
Leads connected

STEP 4

When testing the whole installation from the disconnected tails, it is important that the main switch is in the on position and that the protective devices are in place. If they are circuit breakers they must be in the 'on' position.

Test between live conductors (tails) and operate any two-way and intermediate switching. This is to ensure that all switch wires and strappers are tested and that the switch returns have been correctly identified and connected (*no neutrals in the switches*).

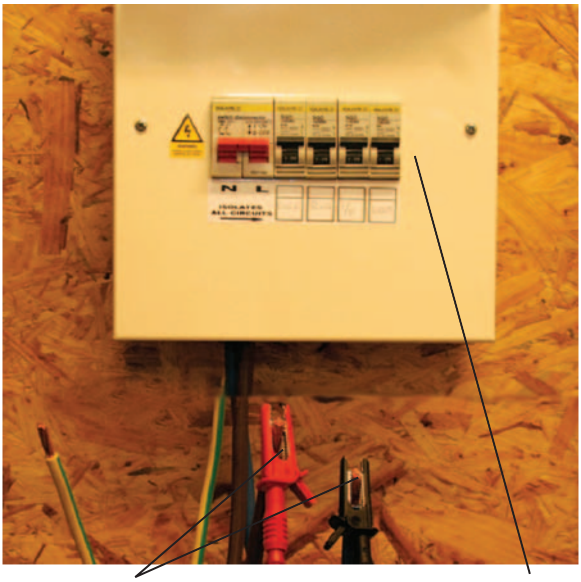

Connected to live conductors

Main switch and circuits on

STEP 5

Join live conductors (tails) together, connect IR tester leads, one on live conductors and the other on the earthing conductor, carry out the test and again operate all two-way and intermediate switching.

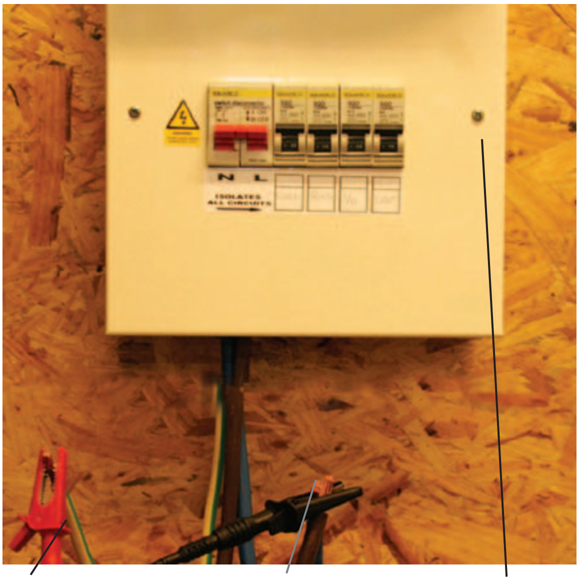

Earthing conductor

Live conductors joined together

Main switch and circuits on

Table 71A of BS 7671 give the acceptable insulation resistance as 0.5 MΩ. This is for a single circuit or a complete installation. *Guidance Note 3* in BS 7671 recommends that any circuit under 2 MΩ should be investigated, as such a low insulation resistance value could indicate a latent defect. Although these values will comply with the requirements of BS 7671, they are very low values for the majority of domestic circuits and installations.

Depending on the type of circuit it may be wise to carry out further investigation on the circuit with the lower reading. It would be very rare that an insulation resistance value of between 0.5 MΩ and 2 MΩ would be acceptable. There would have to be a very good reason for this, and it would need to be monitored to ensure that there was not continued deterioration.

A low value would possibly be acceptable if, for instance, a building had been empty and unused for a period of time, or perhaps an underground cable which had been unused for some time was to be reconnected to supply an outbuilding, or any other outdoor circuit. It could be that, after a period of continued use, the insulation resistance rises; only regular testing would show this.

Example 4

An installation consisting of five circuits is tested as a whole, and the insulation resistance value between live conductors and earth is found to be less than 2 MΩ.

The installation is split into individual circuits and each circuit is tested again, all circuits are found to be greater than 200 MΩ except a mineral insulated cable serving electrically opening gates 30 metres from the building. This circuit is found to have an insulation resistance value of 2 MΩ. In this instance, the value would be compared to previous results to see if any deterioration had occurred since the last test. In the likely event that previous results are not available, the result could be recorded for future reference.

If, however, the low value circuit was a new circuit wired in PVC thermoplastic cable, 6 metres long further investigation should be carried out as the insulation resistance for this circuit should be beyond the scale of the test instrument; in this case >200 MΩ.

When testing is carried out on any installations it is important that the results are thought about and not just recorded. As shown in the example above, an element of careful thought and sensible judgement has to be included in the testing process.

If a high insulation resistance value for a complete installation is measured, it is permissible to enter this value for all circuits on the schedule of test results.

Testing of individual circuits

If it is necessary to test individual circuits, the same process can be applied to new and existing circuits; and the same safety precautions must be taken.

Ensure that safe isolation of the circuit to be tested is carried out. It is preferable for safety reasons to isolate the whole of the distribution board if possible. If not, each circuit can be isolated individually by removing the fuse or turning off and locking off the circuit breaker.

It will be necessary to disconnect the neutral or CPC of the circuit to be tested. This is because the neutrals for all of the circuits will be connected to a common neutral bar and the CPCs for all circuits will be connected to a common earth bar.

If there was an insulation fault between any neutral and CPC in the system it would show up as a fault on all circuits, disconnection of the neutral or CPC will stop this happening.

Check that equipment vulnerable to testing, or any equipment that could produce a low insulation resistance reading is disconnected or isolated.

Carry out a test between live conductors. And then live conductors and earth.

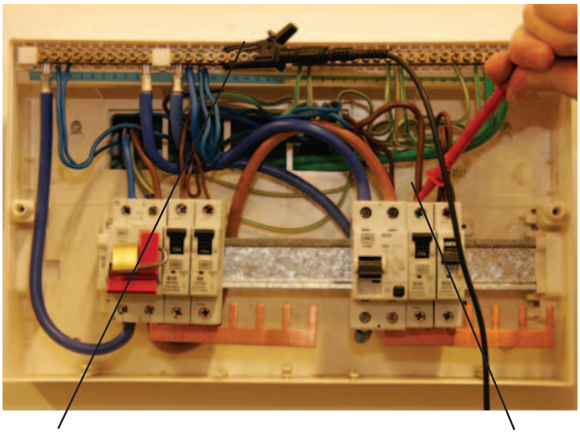

Between live conductors

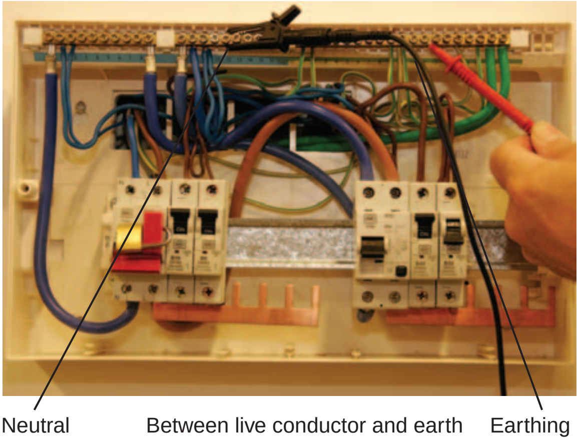

Between live conductor and earth

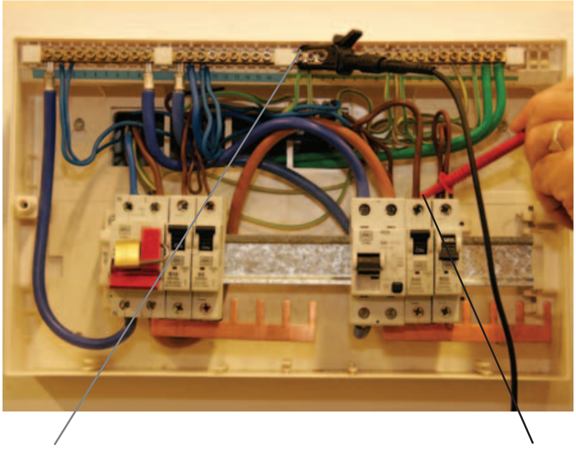
Earthing terminal
Between live conductor and earth
Phase

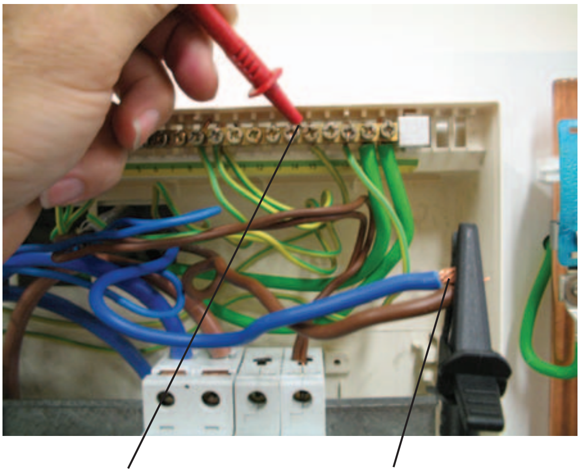
Earthing terminal
Live conductors

Where testing between live conductors and earth the live conductors can be joined and then tested to earth. Alternately, they can be tested separately, whichever is the easiest.

If, for some reason, there is a piece of equipment connected to the system that cannot be isolated from the circuit under test, do not carry out the test between live conductors – only test between live conductors and earth. This is to avoid poor readings and possible damage to equipment. This test should only be carried out on individual circuits, not whole installations, as it is important to test as much of the installation as possible.

Testing of 3 phase installations

When testing 3 phase and neutral (TP&N) installations or circuits the same safety precautions apply as when testing single phase installations or circuits.

The test for the whole installation can be carried out on the isolated side of the main switch. If this is the chosen method it is important to ensure that all of the protective devices are in the 'on' position. Safe isolation must be carried out before commencing this test.

STEP 1

The instrument should be set on 500 volts DC.

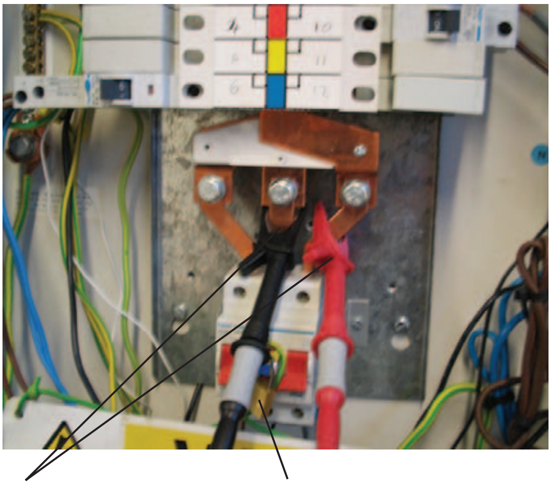
Phase L2 and L3
Locked off and isolated

STEP 2 Test between all phase conductors.

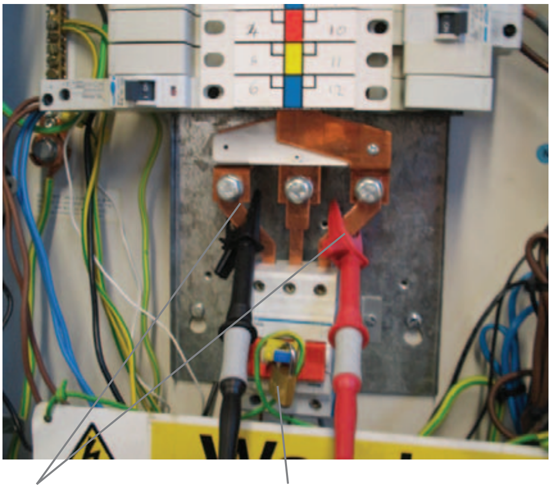
Phase L1 and L3
Locked off and isolated

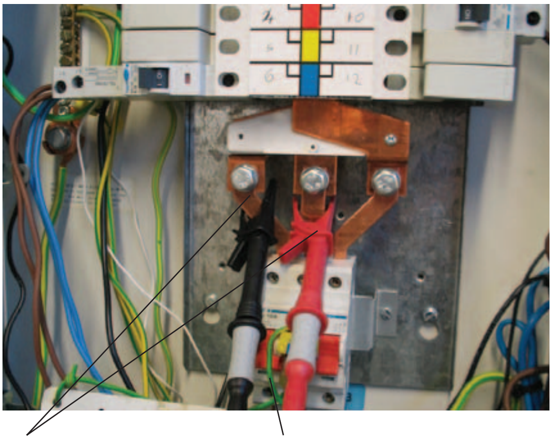
Phase L1 and L2
Locked off and isolated

STEP 3

Test between all phase conductors and neutral, the phase conductors can be joined together for this test to save time.

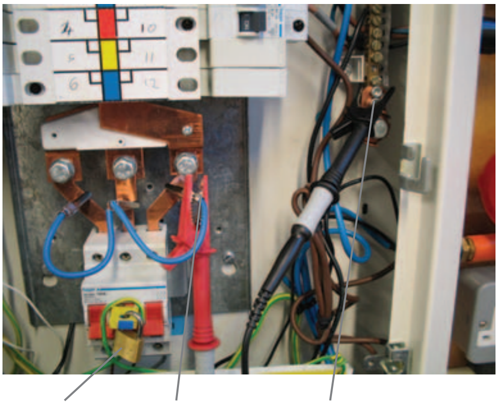
Locked off and isolated
All phases joined together
Neutral

STEP 4

Test between all phase conductors and earth. The phase and neutral conductors can be joined to save time.

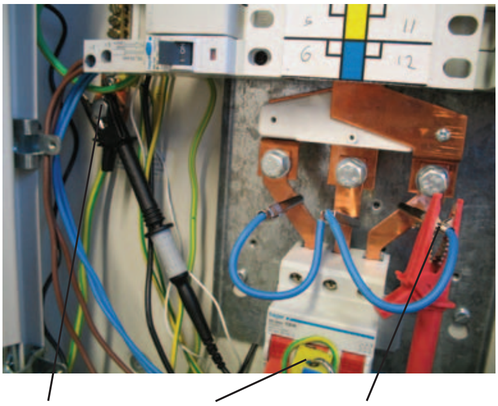
Earth
Locked off and isolated
All phases joined together

STEP 5

Test between neutral and earth.

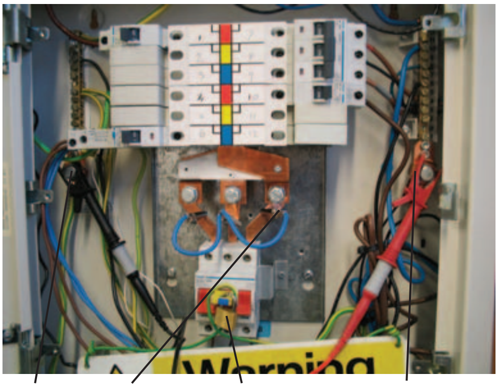
Earth
All phases joined together
Locked off and isolated
Neutral

STEP 6 The resistance values should be 2 MΩ or greater for each circuit and the whole installation must have an insulation resistance of greater than 0.5 MΩ.

STEP 7 If a circuit is found to have an insulation resistance of less than 2 MΩ it should be investigated.

STEP 8 Ensure that any links used for testing are removed prior to the switching on of the supply.

Be aware that if the live conductors are joined and then tested to earth, theoretically the insulation resistance value may be lower due to the conductors being in parallel.

Example 5

A three-phase sub-main is tested and the results are as follows:

L1 to earth is 130 MΩ
L2 to earth is 80 MΩ
L3 to earth is 50 MΩ
N to earth is 100 MΩ

If these conductors were now joined and tested to earth the value would be as given below.

Calculation is:

$$\frac{1}{R1} + \frac{1}{R2} + \frac{1}{R3} + \frac{1}{R4} = \frac{1}{Rt}$$

Put in the values:

$$\frac{1}{130} + \frac{1}{80} + \frac{1}{50} + \frac{1}{100} + \frac{1}{0.05} = 19.92\,M\Omega$$

Enter it this way into a calculator:

$$130X^{-1} + 80X^{-1} + 50X^{-1} = X^{-1} \text{ Answer is } 19.92\,M\Omega$$

This value is still acceptable but lower because the conductors are in parallel.

When completing a Schedule of Test Results, it is important that you document exactly which insulation tests have been carried out. If it is not possible to carry out a test between any conductors a note of this should be made in the remarks column of the schedule.

Testing of site applied insulation

This type of testing is only required where insulation is applied during the erection of equipment on site. It is not required to be carried out on site built assemblies where pre-tested equipment is built on site.

The test is carried out by wrapping metallic foil around all surfaces which has been insulated. Then joining all live conductors together and applying a voltage of 3750 V a.c. between the foil and the conductors for a duration of 60 seconds. The insulation is satisfactory if a breakdown of the insulation does not occur within this time. The instrument used to carry out this test is an applied voltage tester.

Further information on this type of test can be found in the IEE *Guidance Note 3* in BS 7671).

Polarity tests

This is a test carried out to ensure that:

- Protective devices are connected to the phase conductors of the circuits which they are protecting.
- Switches in circuits are in the phase conductor.
- ES lampholders have the centre pin connected to the phase conductor (except for E14 and E27 lampholders complying with BSEN 60238) (Regulation 553-03-04).
- Accessories such as fused connection units, cooker outlets and the like are correctly connected.

Remember! If you are working on a consumer unit or a distribution board, the whole board should be isolated.

In many cases this test can be carried out at the same time as the CPC continuity test. The instrument required is a low resistance ohm meter.

Ensure that the circuit that is to be tested has been isolated.

Polarity test on a radial circuit such as a cooker or immersion heater circuit

STEP 1

At the origin of the circuit, link the phase and CPC using a short lead with crocodile clips at each end. Whilst connecting the lead visually check that the phase conductor is colour identified and connected into the protective device.

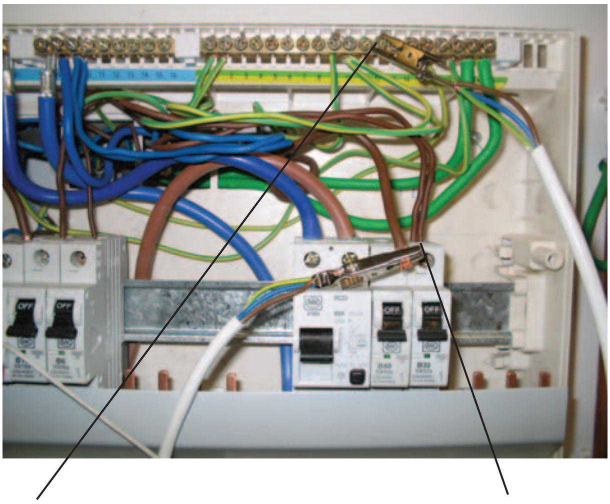
Earthing terminal
linked by lead
Phase

STEP 2

At the furthest point of the circuit remove the cooker or immersion heater switch, visually check that the phase conductor is identified and connected into the correct terminal.

Remember to null the leads of the instrument if the value is to be used for $R_1 + R_2$.

STEP 3

Connect the leads of the test instrument to the incoming phase and CPC terminals at the switch. Test to ensure continuity. The instrument should show a very low reading. This will also be the $R_1 + R_2$ for the circuit.

STEP 4

Remove the lead at the consumer unit and test again. The circuit this time should be open and the instrument reading will be high. This will prove that the correct circuit is being tested.

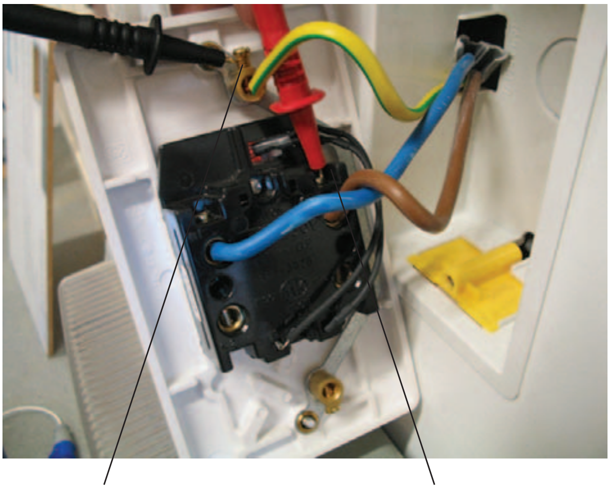
Test instrument probe at incoming CPC terminal of switch
Test instrument probe at incoming phase terminal of switch

Polarity test on a lighting circuit

STEP 1

At the origin of the circuit connect the phase and CPC, this can be done with a short lead with crocodile clips at each end.

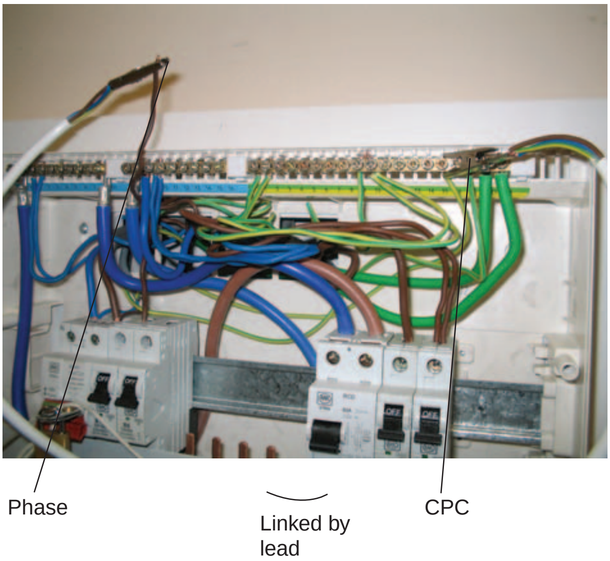
Phase
Linked by lead
CPC

STEP 2

At the ceiling rose or light fitting place the probes of the instrument on to the earthing terminal and the switched live.

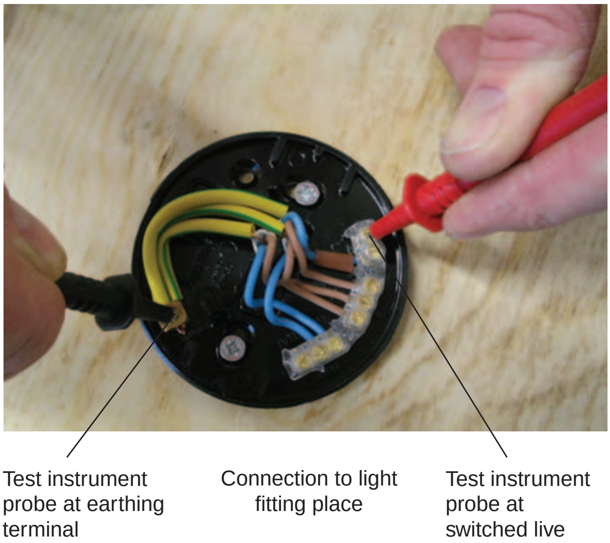
Test instrument probe at earthing terminal
Connection to light fitting place
Test instrument probe at switched live

STEP 3 Close the switch controlling the light and the instrument should read a very low resistance (this will also be the R_1 and R_2 reading for the circuit). When the switch is opened the instrument reading should be very high.

This test can also be carried out at the switch if required:

STEP 1

Place a link between the phase and CPC of the circuit.

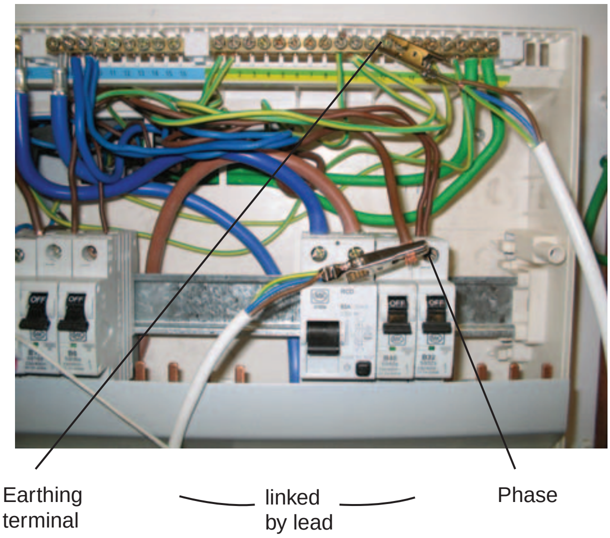
Earthing terminal
linked by lead
Phase

STEP 2

Place the probes of the test instrument on the earth terminal at the switch and the switch return terminal.

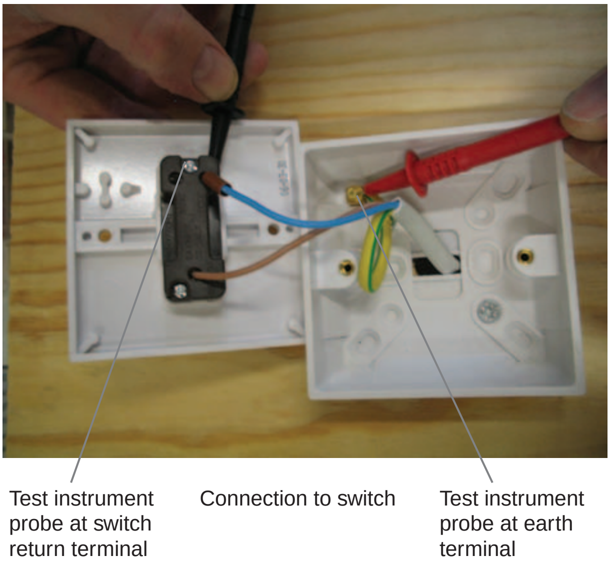
Test instrument probe at switch return terminal
Connection to switch
Test instrument probe at earth terminal

STEP 3

Close the switch and a low resistance reading should be shown on the instrument.

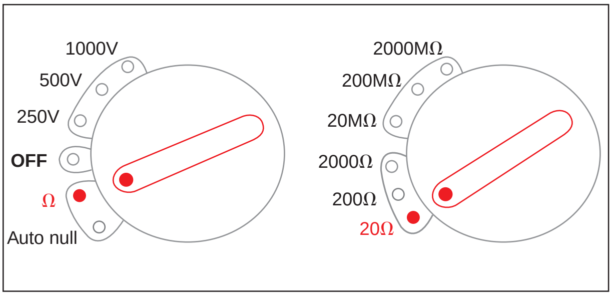
1000V
500V
250V
OFF
Ω
Auto null
2000MΩ
200MΩ
20MΩ
2000Ω
200Ω
20Ω

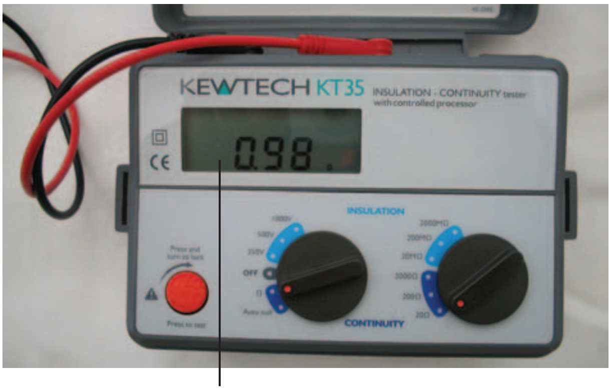
KEWTECH KT35
INSULATION · CONTINUITY tester
0.98
INSULATION
CONTINUITY

Low resistance reading

STEP 4

Open the switch and the instrument reading will show over range as the circuit should be open circuit.

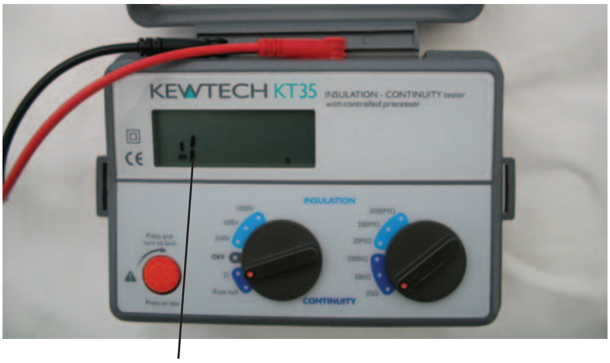
KEWTECH KT35
INSULATION · CONTINUITY tester
INSULATION
CONTINUITY

Over range

Live polarity test

This test is usually carried out at the origin of the installation before it is energized to ensure that the supply is being delivered to the installation at the correct polarity.

The instrument to be used is an approved voltage indicator or test lamp that complies with HSE document GS 38. It is acceptable for an earth loop impedance meter to be used as these instruments also show polarity.

Great care must be taken whilst carrying out this test as it is a live test.

STEP 1

Place the probes of the voltage indicator onto the phase and neutral terminal of the incoming supply at the main switch. The device should indicate a live supply.

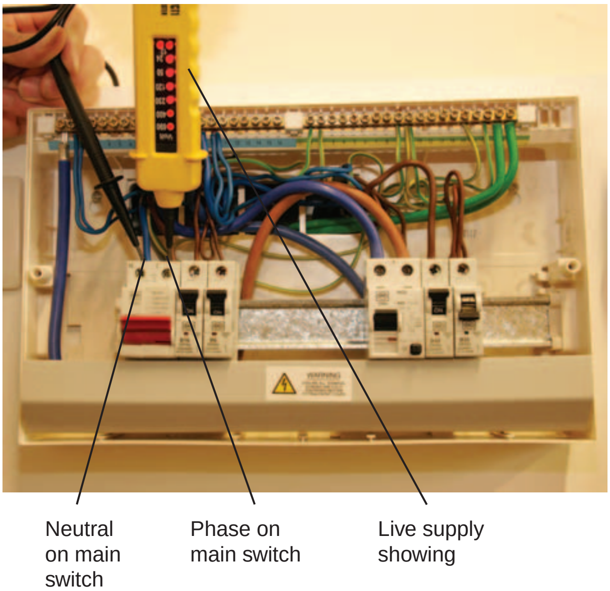
Neutral on main switch
Phase on main switch
Live supply showing

STEP 2

Place the probes of the voltage indicator onto the phase and earth terminal of incoming supply at the main switch. The device should indicate a live supply.

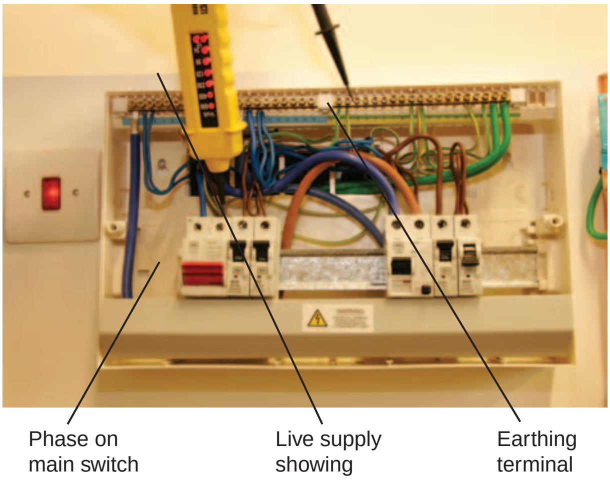
Phase on main switch
Live supply showing
Earthing terminal

STEP 3

Place the probes of the voltage indicator onto the earthing terminal and the neutral terminal at the main switch. The device should indicate no supply.

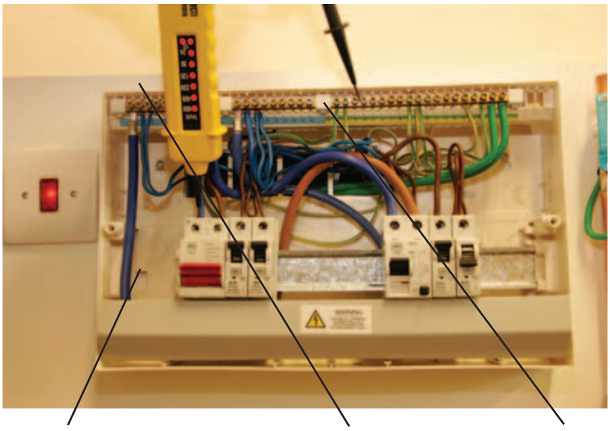
Neutral main switch
No supply showing
Earth terminal

Earth electrode testing

Earth fault loop impedance tester

For many installations the resistance of the earth electrode can be measured using an earth fault loop impedance test instrument. It is perfectly acceptable to use this type of instrument on a TT system where reasonably high resistance values could be expected.

The test is performed in exactly the same way as the external earth fault loop Z_e test.

STEP 1 Isolate the installation.

STEP 2 Ensure that the earthing conductor is correctly terminated at the earth electrode.

STEP 3 Disconnect the earthing conductor from the main earthing terminal.

STEP 4 Connect a lead of the earth fault loop meter to the disconnected earthing terminal.

STEP 5 Place the probe of the other lead on to the incoming phase conductor at the supply side of the main switch and carry out the test.

STEP 6 Record the result.

STEP 7 Reconnect the earthing conductor and leave the installation in safe working condition.

If a three lead test instrument is used, read the instrument instructions before carrying out the test. It may be that the leads must be connected on to the phase, neutral and earthing terminals, or possibly the neutral and earth lead of the test instrument, should be joined together.

Regulation 413-02-16 from BS 7671 tells us that the maximum permissible resistance value of the earth electrode must be no greater than 50/IΔn. This is to ensure that any exposed metalwork does not rise to a potential of greater than 50 volts.

50 is the maximum voltage
IΔn is the trip rating of the residual current device
Z_S is the earth fault loop impedance

If the rating of the device is 100 mA the calculation is:

$$\frac{50}{0.1} = 500\ \Omega$$

The maximum permissible value to comply with the regulations using this calculation is 500 Ω. Although this value would be deemed acceptable, it may not be reliable as it could rise to an unacceptable value if the soil dries out.

An acceptable value for RCDs up to 100 mA is stated as a maximum of 200 Ω. If the resistance is above this it should not be accepted under most circumstances. The installation of an additional or larger electrode may bring the resistance value down to an acceptable value.

The maximum calculated values for earth electrodes are:

Operating current of the RCD	Electrode resistance in ohms
100 mA	500
300 mA	160
500 mA	100

For special locations where the maximum touch voltage is 25 V the electrode resistance should be halved. Electrode tests should be carried out in the worst possible conditions. The worst condition for an earth electrode is when the soil is **dry**. Where lower values of earth electrode resistances are required, an earth electrode tester should be used.

Measurement using an earth electrode test instrument

This test requires the use of three electrodes: the earthing electrode under test, a current electrode and a potential electrode.

STEP 1 The earthing electrode (E) should be driven into the ground in the position that it is to be used. Attention should be paid to the length of the electrode which is in the ground.

STEP 2 The current electrode (C^2 spike) should be pushed into the ground at a distance of ten times the depth of the electrode under test away from it.

STEP 3 The potential electrode (P^2) should be pushed into the ground midway between E and C^2.

STEP 4 The leads of the test instrument should be connected to the appropriate electrodes.

STEP 5 Measure the value of resistance.

STEP 6 Move P^2 10% closer to C^2.

STEP 7 Measure the value of resistance.

STEP 8 Move P^2 back to 10% closer to E than the mid-point.

STEP 9 Measure the value of the resistance.

A calculation must now be carried out to find the percentage deviation of the resistance values.

Example 6

Three measurements are taken: 0.8 ohms, 0.86 ohms and 0.78 ohms. These must now be added together and an average value calculated.

Total value of three readings $= 2.44$

Find the average:

$$\frac{2.44}{3} = 0.81$$

The average value is 0.81 ohms.

Now find the highest difference between the average value and the measured values. In this case it will be $0.86 - 0.81 = 0.05$.

The percentage of this value to the average value must now be found:

$$\frac{0.05}{0.81} \times 100 = 6.1\%$$

This value is higher than 5% of the average value and it is not advisable to accept, it as a percentage deviation of greater than 5% is deemed to be inaccurate. To overcome this, the distance between the electrode under test (E) and the current spike (C^2) *should now be increased and the first three tests repeated to obtain a more accurate reading.*

If the required resistance value cannot be obtained by the use of a single electrode, additional electrodes may be added at a distance from the first electrode equal to its depth.

Earth fault loop impedance Z_e

This is a live test and great care must be taken ⚠

Z_e is a measurement of the external earth fault impedance (resistance) of the installation. In other words, it is the measured resistance of the supply transformer winding, the supply phase conductor, and the earth return path of the supply. It is measured in ohms.

Earth fault path for a TT system

This system uses the mass of earth for the fault return path.

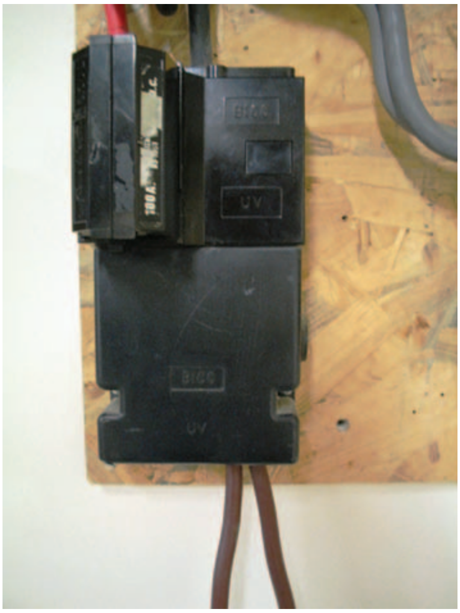

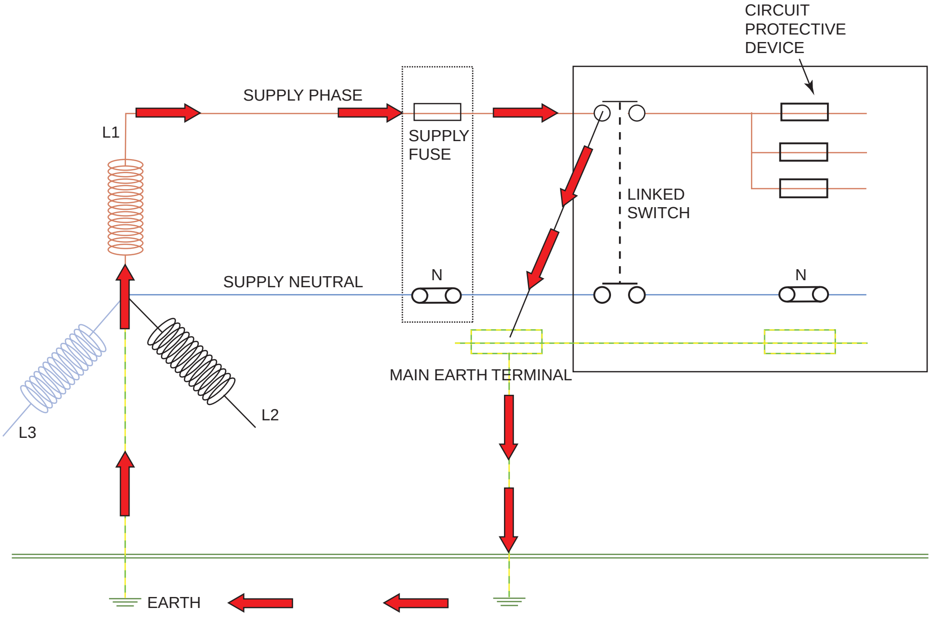
CIRCUIT
PROTECTIVE
DEVICE
SUPPLY PHASE
L1
SUPPLY
FUSE
LINKED
SWITCH
SUPPLY NEUTRAL
N
N
MAIN EARTH TERMINAL
L3
L2
EARTH

Earth fault path for a TNS system

This system uses the sheath of the supply cable for the earth fault return path.

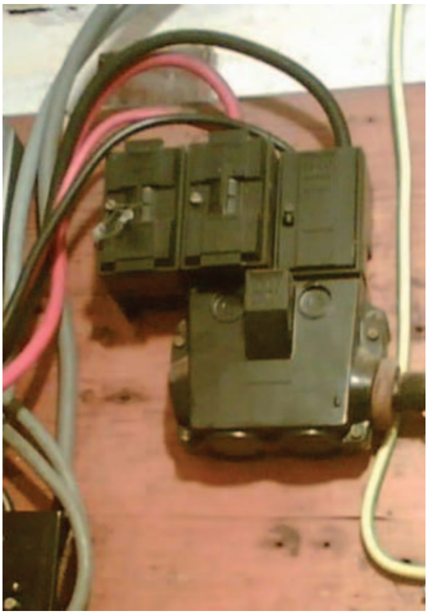

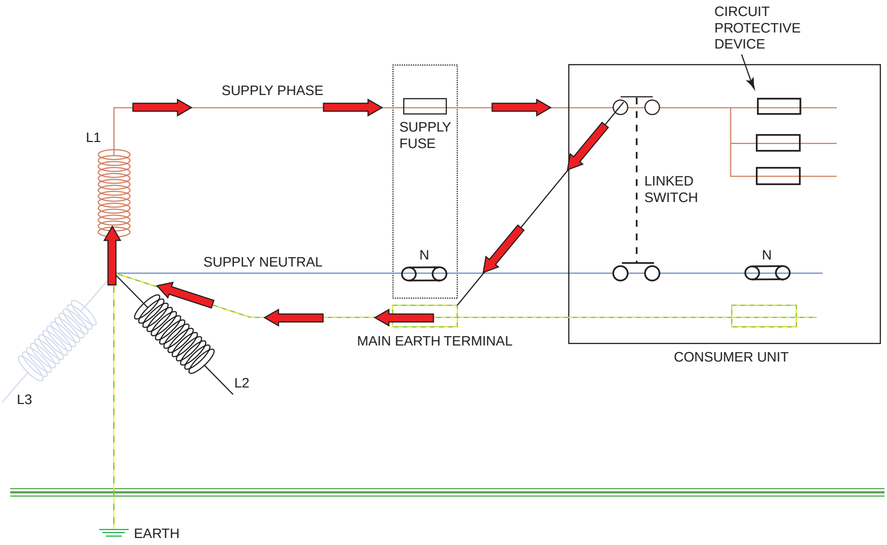
CIRCUIT
PROTECTIVE
DEVICE
SUPPLY PHASE
L1
SUPPLY
FUSE
LINKED
SWITCH
SUPPLY NEUTRAL
N
N
MAIN EARTH TERMINAL
CONSUMER UNIT
L2
L3
EARTH

Earth fault path for a TNCS system

This system uses the neutral (PEN) conductor of the supply for its earth fault return path.

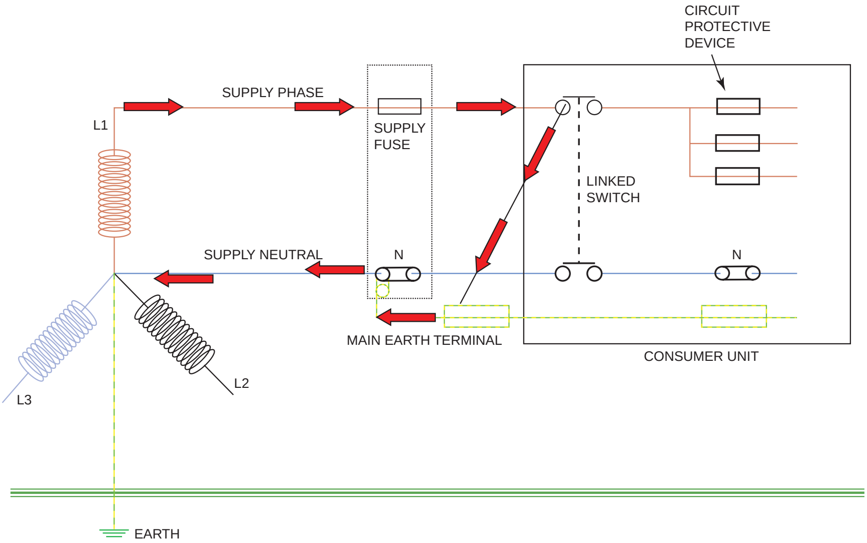
CIRCUIT
PROTECTIVE
DEVICE
SUPPLY PHASE
L1
SUPPLY
FUSE
LINKED
SWITCH
SUPPLY NEUTRAL
N
N
MAIN EARTH TERMINAL
CONSUMER UNIT
L2
L3
EARTH

To carry out this test correctly the installation should be isolated from the supply and the main earth disconnected from the main earth terminal (MET). This is to avoid the possibility of parallel paths through any earthed metalwork within the installation.

Often in industrial and commercial installations – where isolation may be impossible due to the building being in use – the only time that this test can be carried out with the main earthing conductor disconnected, is during the initial verification. It is important that the Z_e with the earthing conductor is disconnected during the initial verification as this will give a reference value for the life of the installation. If, during subsequent tests, the earthing conductor cannot be disconnected, a test can still be carried out but the parallel paths should give a lower impedance value. If a higher value is recorded it will indicate a deterioration of the supply earth.

The instrument used for this test is an earth fault loop impedance meter, and it is important that the person using the instrument has read and understood the operating instructions. There are many types of test instruments on the market and they all have their own characteristics.

Some instruments have three leads which must be connected to enable this test to be carried out correctly. Some instruments require that the leads are connected to the phase, neutral and earth of the circuit to be tested. Other instruments require the phase lead to be connected to the phase conductor and the earth and neutral leads to be connected to the earthing conductor of the circuit to be tested (*it is important to read the instructions of the instrument being used*).

If using a two-lead instrument it should be set on Z_e. One lead should be connected to the main earthing conductor and the other to the incoming phase on the supply side of the main switch.

Probes can be used for this providing they meet the requirements of GS 38, i.e.:

- Insulated
- Fused
- Finger guards
- Maximum of 4 mm exposed tips or retractable shrouds
- Long enough to carry out test safely
- Undamaged

Performing the test

This is a live test and care should be taken when carrying it out

STEP 1 Isolate the supply.

STEP 2 Disconnect the earthing conductor.

STEP 3 Set the instrument to loop test.

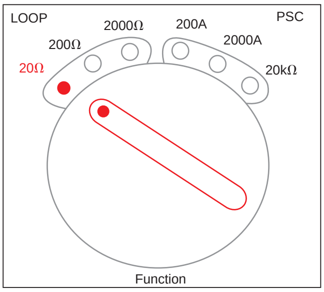
LOOP
PSC
2000Ω
200A
200Ω
2000A
20Ω
20kΩ
Function

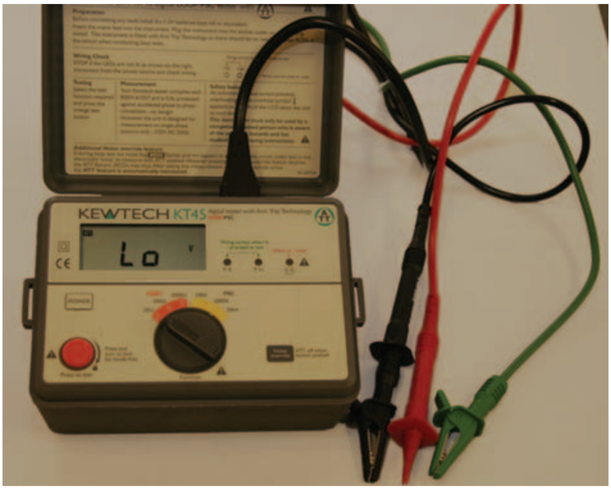
KEWTECH KT45
Lo

STEP 4 If you are using a two-lead instrument the leads should be connected as in the figures below. If you are using a three lead instrument, then the leads should be connected as shown here.

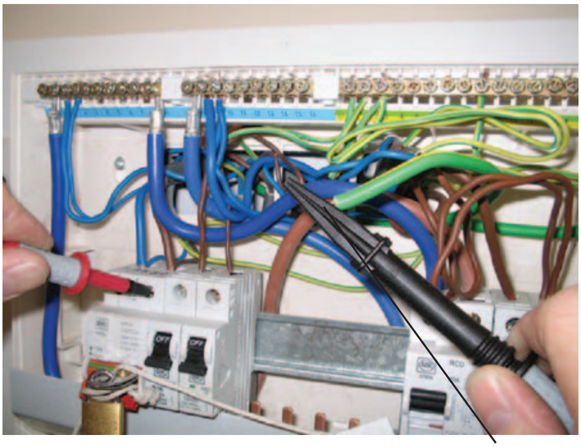

Connections for 2-lead instrument

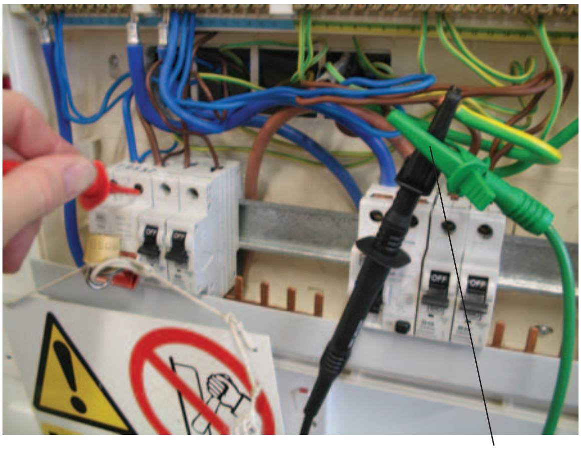

Connections for 3-lead instrument

STEP 5 The measurement obtained is Z_e and can be entered on the test certificates in the appropriate place.

STEP 6 Reconnect earthing conductor.

It is important that the instructions for the test instruments are read and fully understood before carrying out this test.

Care should be taken to reset time clocks, programmers, etc. when the supply is reinstated.

Circuit earth fault loop impedance

This is a live test and great care must be taken ⚠

Z_S is the value of the earth loop impedance (resistance) of a final circuit including the supply cable.

The earth fault loop (Z_S) path for a TT system.

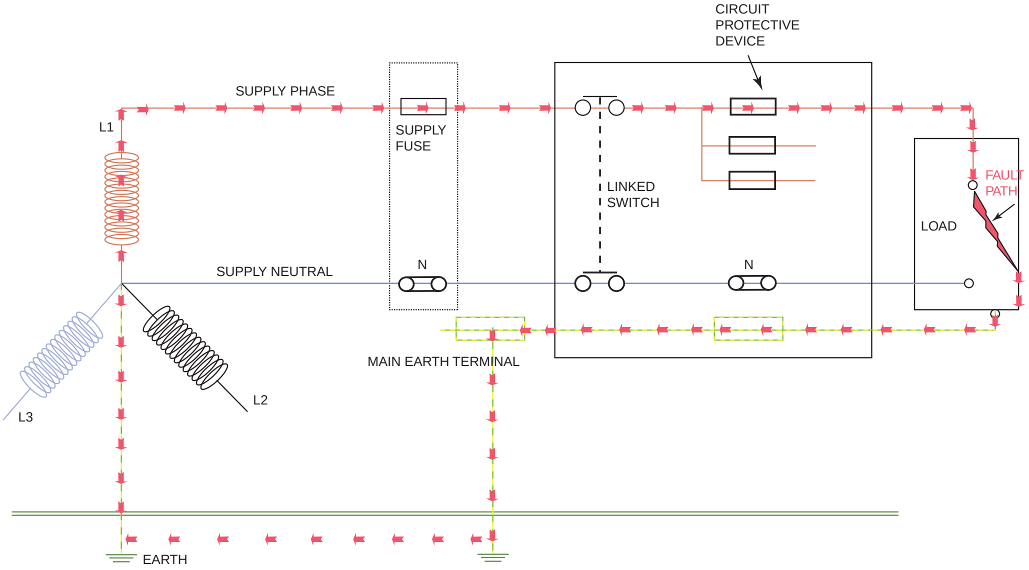
CIRCUIT
PROTECTIVE
DEVICE
SUPPLY PHASE
L1
SUPPLY
FUSE
LINKED
SWITCH
FAULT
PATH
LOAD
SUPPLY NEUTRAL
N
N
MAIN EARTH TERMINAL
L2
L3
EARTH

The earth fault loop (Z_S) path for a TNS system.

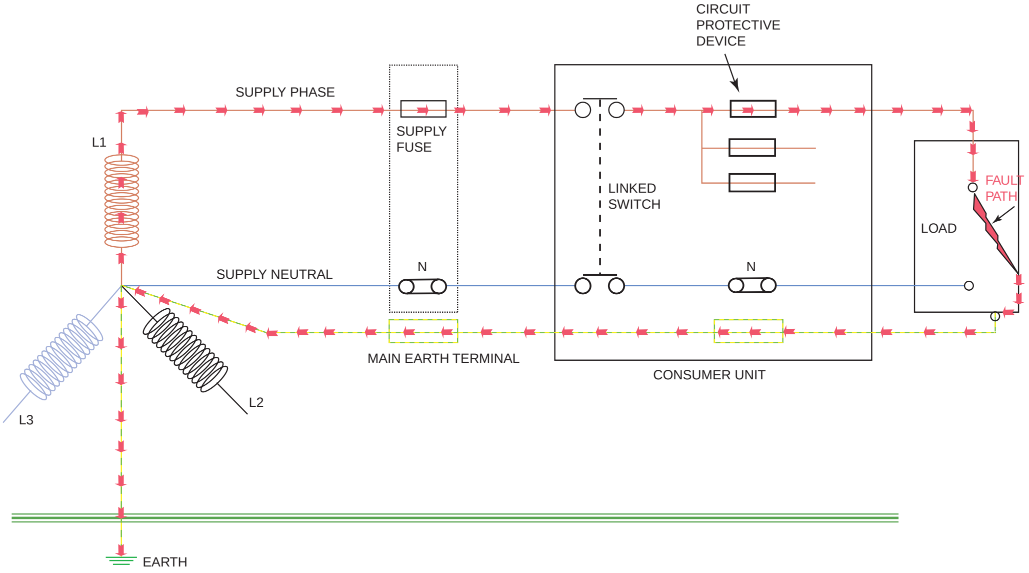
CIRCUIT
PROTECTIVE
DEVICE
SUPPLY PHASE
L1
SUPPLY
FUSE
LINKED
SWITCH
FAULT
PATH
LOAD
SUPPLY NEUTRAL
N
N
MAIN EARTH TERMINAL
CONSUMER UNIT
L2
L3
EARTH

The earth fault loop (Z_S) path for a TNCS system.

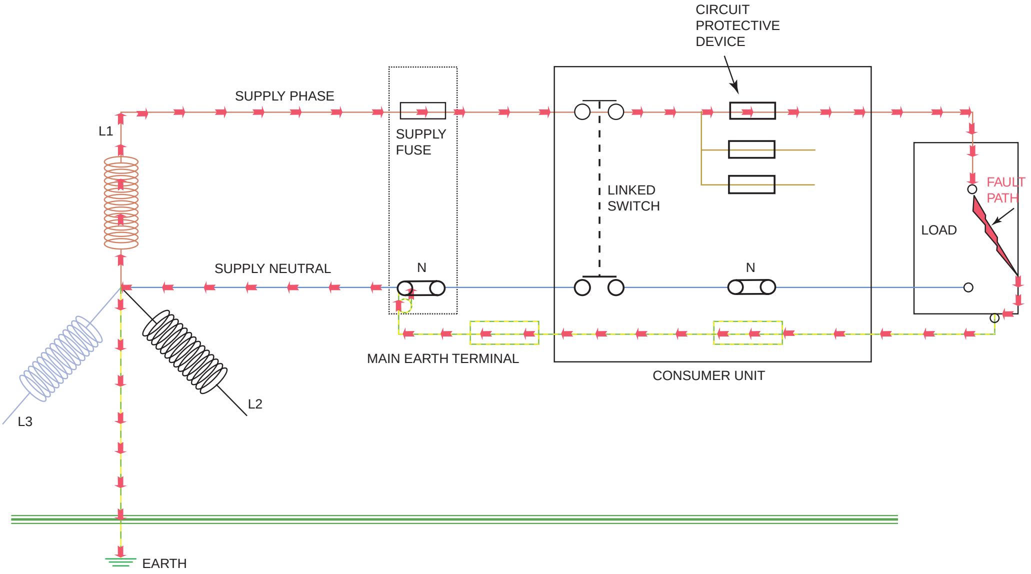
CIRCUIT PROTECTIVE DEVICE
SUPPLY PHASE
L1
SUPPLY FUSE
LINKED SWITCH
FAULT PATH
LOAD
SUPPLY NEUTRAL
N
N
MAIN EARTH TERMINAL
CONSUMER UNIT
L2
L3
EARTH

To obtain Z_S, the Z_e value should now be added to the $R_1 + R_2$ values that were obtained when carrying out the CPC continuity tests for each individual circuit.

The total value $Z_e + R_1 + R_2$ will be Z_S (earth loop impedance for the circuit).

This value (Z_S) should now be compared with the maximum values of Z_S given in BS 7671, Chapter 41, to verify that the protective device will operate in the correct time.

Unfortunately, it is not quite as simple as it seems. This is because the values Z_S have been measured when the conductors were at room temperature and the maximum Z_S values given in BS 7671 are at the conductor operating temperature of 70°C. This is the maximum temperature that the conductor could be operating at in a sound circuit. There are two methods.

Accurate test instruments must be used for these tests.

Method one

Measure the ambient temperature of the room and use the values from Table 9B in the *On-Site Guide* as dividers. (**DO NOT USE THEM AS MULTIPLIERS**)

This is because Table 9B is to correct the conductor resistances in Table 9A from 20°C to room temperature.

When they are used as dividers they will correct the cable from room temperature resistance, to the resistance that it would be at 20°C.

Example 7

$$\frac{R_1 + R_2}{\text{Temp factor}}$$

Measured $R_1 + R_2 = 0.84\Omega$ @ 25°C

Factor from Table 9b for 25°C = 1.02.

$$\frac{0.84}{1.02} = 0.82$$

Values at 20°C is 0.82 Ω.

Having corrected the measured values to 20°C the next step is to calculate what the resistance of the cable would be at its operating temperature.

As the resistance of copper changes by 2% for each 5°C the conductor resistance will rise by 20% if its temperature rises to 70°C.

Multiplying the resistance by 1.2 will increase its value by 20%, this value can now be added to Z_e *to give* Z_S.

Resistance value at operating temperature would be:

$$0.82\,\Omega \times 1.2 = 0.98\,\Omega$$

Example 8

The Z_e of an installation is 0.6 Ω. A circuit has been installed using twin and CPC 70°C thermoplastic (pvc) cable. The room temperature is 25°C and the measured $R_1 + R_2$ value is 0.48Ω. The circuit is protected by a BS EN 60898 16A type B device.

Correct the cable resistance to 20°C by using factor from Table 9B *On-Site Guide.*

$$\frac{0.48}{1.02} = 0.47\,\Omega$$

Adjust this value to conductor operating temperature by increasing it by 20%.

$$0.47 \times 1.2 = 0.56\,\Omega$$

Add this value to installation Z_e *to find* Z_S.

$$0.56 + 0.6 = 1.16\,\Omega$$

This value can now be compared directly with the maximum value Z_S *for a 16A type B protective device. This value is 3Ω and can be found in Table 41B2 in BS 7671.*

To comply with the regulations the actual value 1.16 Ω is acceptable as it is less than 3 Ω.

Method two

Measure the resistance of earth fault loop impedance at the furthest point of the circuit using the correct instrument (*remember the furthest point is the end of the circuit, not necessarily the furthest distance from the distribution board*). Record the value obtained onto the test result schedule.

This measurement cannot be compared directly with the values from BS 7671 because the operating temperature of the conductors and the ambient temperature of the room are unknown.

This method is useful for a periodic test where existing test results are available. If the measured value is higher than previous results it will indicate that there is a possible deterioration of the earth loop impedance of the circuit.

The usual method to check that the measured Z_S is acceptable is to use the rule of thumb method.

First look in the correct table in Chapter 41 of BS 7671 for the maximum permissible Z_S of the protective device for the circuit being tested.

Use three-quarters of this value and compare it with the measured value. Providing the measured value is the lowest the circuit will comply.

More information on this can be found in Chapter 5 (Protective devices).

Performing the tests

This is a live test and great care must be taken

A circuit incorporating a socket outlet on a ring or a radial

STEP 1 Use an earth fault loop impedance instrument. Set it onto 20Ω (unless you have a self ranging instrument).

STEP 2 Ensure all earthing and bonding is connected.

STEP 3 Plug in the instrument and record the reading.

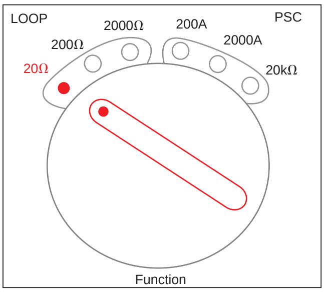
LOOP
2000Ω
200A
PSC
200Ω
2000A
20Ω
20kΩ
Function

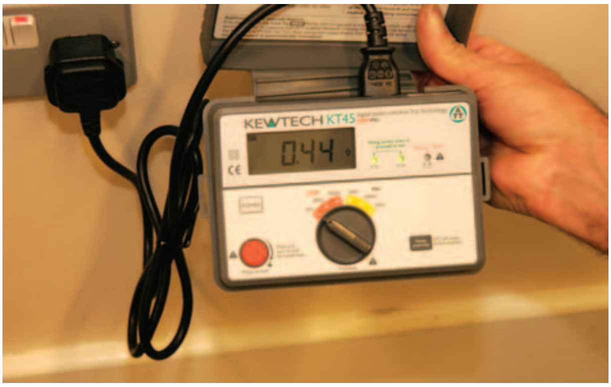
KEWTECH KT45
0.44

Performing the test on a radial circuit other than a socket outlet

STEP 1

Ensure earthing and bonding is connected.

STEP 2

Isolate circuit to be tested.

STEP 3

Remove accessory at the extremity of the circuit to be tested.

STEP 4

Use an earth fault loop impedance instrument with fly leads.

Place the leads on correct terminals. If you are using a two-lead instrument as connected here; if you are using a three-lead instrument as connected here (always read the instrument manufacturers instructions).

STEP 5

Energize the circuit.

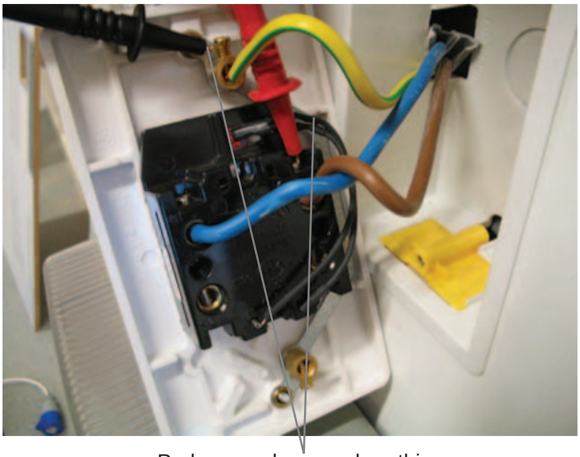

Probes on phase and earthing terminals for 2-lead instrument

Two-lead instrument as connected here

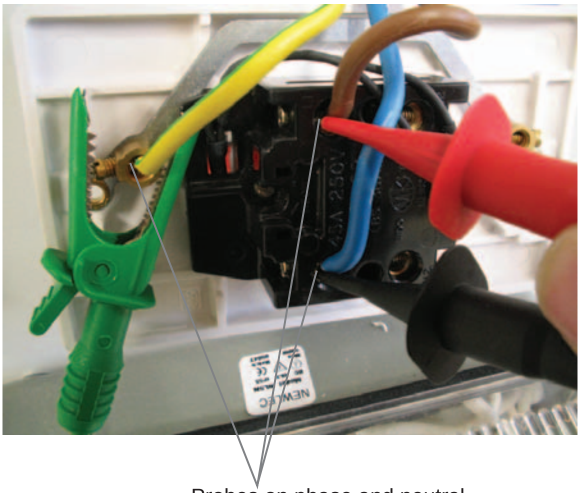

Probes on phase and neutral terminals, clip on earthing terminal for 3-lead instrument

Three-lead instrumented as connected here

STEP 6

Perform the test on a lighting circuit at the ceiling rose or at the switch.

If you are using a two-lead instrument, place the probes as shown on the left, this will also prove polarity if the switch is operated whilst carrying out the test (*this may be easier with two people*).

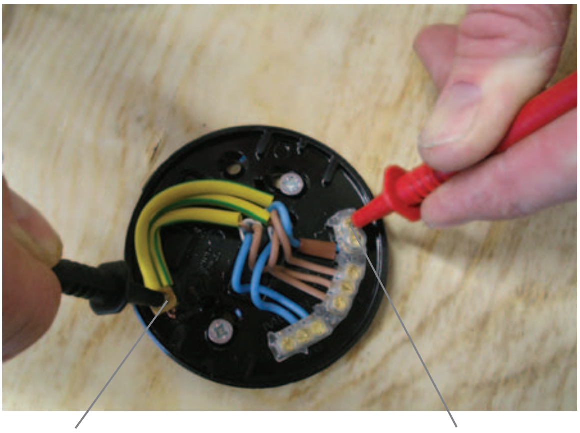
Test instrument probe at earthing terminal
Test instrument probe at switched line

Connection to light-fitting place for 2-lead instrument

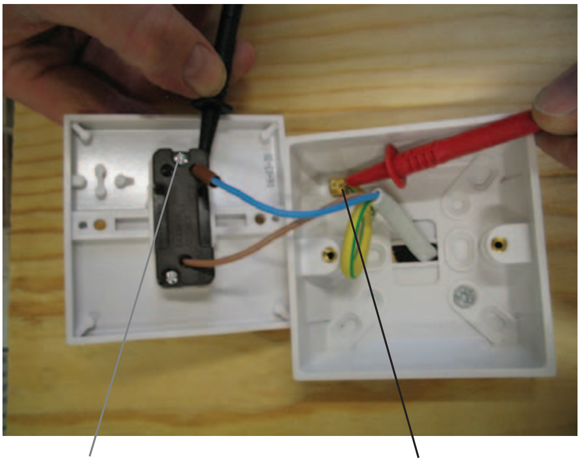
Test instrument probe at switch return terminal
Test instrument probe at earth terminal

Connection to switch for 2-lead instrument

If you are using a three-lead instrument then connect the probes as shown here (always read the instrument's instructions).

STEP 7 Isolate the circuit and remove the leads.

STEP 8 Replace accessory and energize circuit.

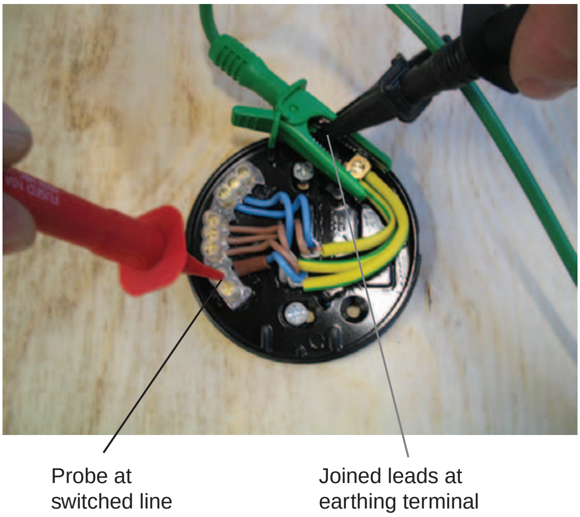
Probe at switched line
Joined leads at earthing terminal

Connection to light-fitting place for 3-lead instrument

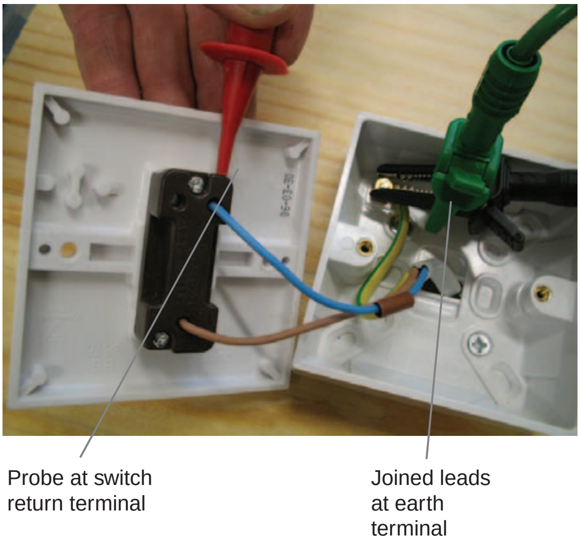
Probe at switch return terminal
Joined leads at earth terminal

Connection to switch for 3-lead instrument

Example 9

A ring final circuit is protected by a 30A BS3036 semi-enclosed rewirable fuse and the measured Z_S is 0.96 Ω.

As this is a ring final circuit, the disconnection time has to be 0.4 seconds. From Table 41B1 in BS 7671 the maximum Z_S *for a 30A rewirable fuse is 1.14 Ω.*

Three-quarters of this value must now be calculated, this can be achieved by multiplying it by 0.75.

$$1.14 \times 0.75 = 0.85\,\Omega$$

The measured value for the circuit must now be lower than the corrected value if it is to comply with BS 7671.

Measured value	*0.96 Ω*
Corrected value	*0.85 Ω*

The measured value of Z_S *is higher; therefore, the circuit will* ***not*** *comply.*

Option one is the preferred method because it will give an accurate value whereas the test in Option two will include parallel paths and because of this will often give lower readings.

Option one should always be used for an initial verification as the first reading will be used as a benchmark to be compared with results taken in future periodic tests.

If, on a periodic inspection using Option two, a higher test result is obtained than on the initial verification, this would indicate that the circuit is deteriorating and that further investigation would be required.

The methods described here must be fully understood by anyone who is intending to sit the City and Guilds 2391 exam on inspection and testing of electrical installations.

However! Providing that the ambient temperature is between 10°C and 20°C there is a much simpler way to check the results when you are actually working on site testing. The *On-Site Guide* and *Guidance Note 3* information contain tables which are already corrected for conductor operating current and ambient temperature. In the *On-Site Guide* these tables can be found in Appendix 2 (Table 2A to Table 2D). These tables can also be used where the CPC is of a different cross-sectional area than the live conductors.

These values are slightly more favourable than the rule-of-thumb method, as they are at approximately 80% of values given in BS 7671 and are perfectly acceptable (*remember, however, that the previous methods described must also be understood*).

Example 10

A circuit supplying a fixed load is protected by a 20A BS 3036 fuse and the measured Z_S is 2.32 Ω. The circuit CPC is 1.5 mm^2.

As this circuit is supplying a fixed load the maximum permitted disconnection time is 5 seconds.

The table that should be used is 2A(ii) from the On-Site Guide. Using this table it can be seen that for a 20A device protecting a circuit with a 1.5 mm^2 *CPC the maximum permissible* Z_S *is 3.2 Ω. As the measured* Z_S *is lower than the maximum permitted, the circuit complies with the Regulations.*

Example 11

A circuit supplying a cooker outlet which incorporates a socket outlet protected by a 45A BS 1361 fuse, has a measured Z_S of 0.4 Ω. The CPC is 4 mm^2.

As this circuit has a socket outlet on it, the disconnection time would be 0.4 seconds.

The table to use for this circuit is 2C(i) from the On-Site Guide and the maximum permitted Z_S *is 0.48 Ω. This circuit will comply.*

Example 12

A circuit supplying a lighting circuit protected by a 6A type C BS 60898 protective device has a measured Z_S *of 2.9 Ω.*

It should be remembered that miniature circuit breakers will operate at 0.1 seconds providing that the measured Z_S *is equal to or lower than the values given in the tables. We do not have to worry about 0.4 or 5 second disconnection times for these devices.*

The table to use for this example is 2D in the On-Site Guide, the maximum permitted Z_S *is 3.2 Ω. The measured value of 2.9 Ω is lower than the maximum permitted, therefore, this circuit would comply.*

Remember that the values in GN3 and the *On-Site Guide* are corrected for temperatures between 10°C and 20°C and no other calculation is required providing the ambient temperature is between these values.

If the ambient temperature is below 10°C or above 20°C then correction factors from Table 2E of the *On-Site Guide* must be used as follows.

Using Example 9, a circuit supplying a fixed load is protected by a 20A BS 3036 fuse and the measured Z_S is 2.32 Ω. The circuit CPC is 1.5 mm^2, and the **ambient temperature is 23°C.**

As this circuit is supplying a fixed load, the maximum permitted disconnection time is 5 seconds.

The table that should be used is 2A(ii) from the *On-Site Guide*. Using the table it can be seen that for a 20A device protecting a circuit with a 1.5 mm^2 CPC, the maximum permitted Z_S is 3.20 Ω.

Now the temperature has to be taken into account.

Using Table 2E from Appendix 2 of the *On-Site Guide* it can be seen that the nearest value to the temperature which was measured (23°) is 25°C (*always round up to be on the safe side*). The correction factor for 25°C is 1.06.

This value (1.06) is now used as a multiplier to the maximum permitted Z_S (3.20 Ω) to calculate the maximum Z_S for the circuit at 25°C

$$3.20 \times 1.06 = 3.39\,\Omega$$

This is the maximum measured Z_S permissible for the circuit at 25°C.

Prospective fault current test

This is a live test and great care must be taken ⚠

A prospective fault current test instrument is normally combined with an earth loop impedance test instrument, the measured value is normally shown in kA (kilo amps).

Regulations 434-02-01 and 713-12-01 require that the prospective short circuit current and the prospective earth fault current are determined. Once determined, the highest value must of course be recorded.

A prospective short circuit current is the maximum current that could flow between phase and neutral on a single-phase supply or between phase conductors on a three-phase supply.

A prospective earth fault current is the maximum current that could flow between live conductors and earth.

The higher of these values is known as prospective fault current.

The highest prospective fault current will be at the origin of the installation and must be measured as close to the meter position as possible, usually at the main switch for the installation. It is measured between phase and neutral.

This can be done by:

- Enquiry to the supplier
- Calculation
- Measurement

Enquiry

This is a matter of a phone call to the electricity supplier of the installation. They will tell you the maximum PFC. Usually this is a lot higher than the value will actually be, but if you use this value you will be on the safe side.

Calculation

The PFC can only be calculated on a TNCS system. This is because the neutral of the supply is used as a Protective earth and neutral (PEN) conductor.

When the earth fault loop impedance is measured, the value measured is in ohms. To convert this value to prospective short circuit current we must use the following equation:

$$\text{PSCC} = \frac{V}{Z_e} = I$$

It is important to remember that the open circuit voltage of the supply transformer is used U_{OC} 240 V (BS 7671, Appendix 3).

Example 13

Z_e is measured at 0.28 Ω

$$\frac{240}{0.28} = 857 \text{ amps}$$

A useful tip is that when you have measured Z_e on a TNCS system, set your instrument to PFC and repeat the test. This will give you the value for PFC and save you doing the calculation.

Measurement

This is carried out using a prospective fault current tester. As with all tests it is important that you have read the instructions for the instrument which you are going to use.

If you are using a two-lead instrument with leads and probes to GS 38:

STEP 1

Set instrument to PFC.

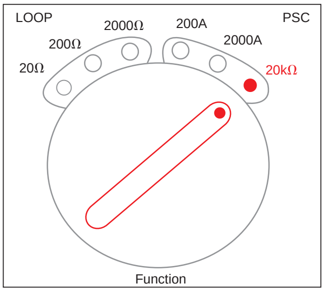
LOOP
PSC
2000Ω
200A
200Ω
2000A
20Ω
20kΩ
Function

KEWTECH KT45
0.24
POWER
LOOP
PSC
Function
Press to test

STEP 2

Place the probes on the phase and neutral terminals at supply side of the main switch.

STEP 3

Operate the test button and record the reading.

When carrying out the test using a three-lead instrument with leads to GS 38, it is important that the instrument instructions are read and fully understood before carrying out this test.

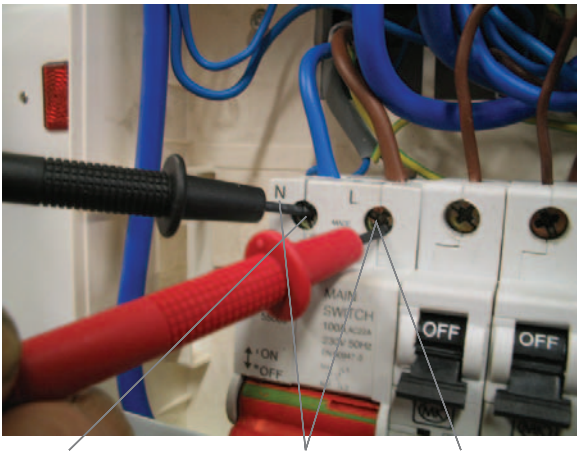

Probe at neutral at supply side of main switch

Insulated tips

Probe at phase at supply side of main switch

Connection for 2-lead instrument

STEP 1

Place the phase lead on the supply side of the main switch and the neutral and earthing probes/clips onto the earthing terminal.

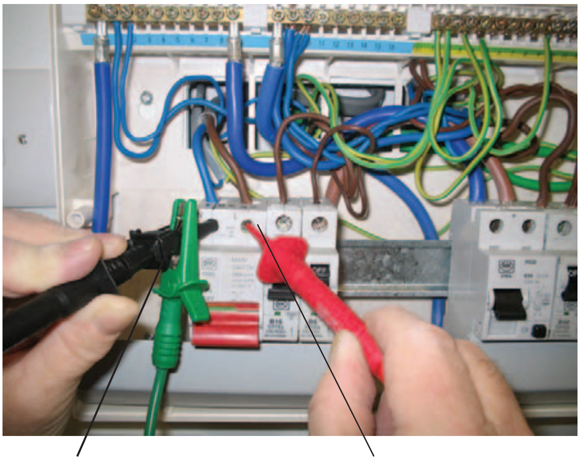

Joined leads at neutral at supply side of main switch

Probe at phase at supply side of main switch

Connection for 3-lead instrument

STEP 2 Operate the test button and record the reading.

If the supply system is a three-phase and neutral system then the highest current that could flow in it will be between phases. Some instruments will not be able to measure the high current that would flow under these circumstances.

Under these circumstances the measurement should be made between any phase and neutral at the main switch and the measured value should be **doubled**.

For your personal safety and the protection of your test equipment it is important to read and fully understand the instructions of your test instrument before commencing this test.

Some PSCC (protective short-circuit current) instruments give the measured value in ohms, not kA. If this is the case, a simple calculation, using ohms law is all that is required.

Example 14

Measured value is $0.08\,\Omega$.

Remember to use U_{OC} in this calculation (240 V)

$$PSCC = \frac{240}{0.08} = 3000\ A$$

It is important that the short circuit capacity of any protective devices fitted exceeds the maximum current that could flow at the point at which they are fitted.

When a measurement of PFC is taken as close to the supply intake as possible, and all protective devices fitted in the installation have a short circuit capacity that is higher than the measured value, then Regulation 432-02-01 will be satisfied.

In a large installation where sub mains are used to supply distribution boards it can be cost effective to measure the PFC at each board. The PFC will be smaller and could allow the use of a protective device with a lower short circuit rating. These will usually be less expensive.

Table 7.2A in the *On-Site Guide* gives rated short circuit capacities for devices. These values can also be obtained from manufacturer’s literature.

Examples	Rated short circuit capacity
Semi-enclosed BS 3036	1 kA to 4 kA depending on type
BS 1361 Type 1	16.5 kA
Type 2	33 kA
BS 88-2.1	50 kA at 415 volts
BS 88-6	16.5 kA at 240 volts
	80 kA at 415 volts

Circuit breakers to BS 3871 are marked with values M1 to M9 the number indicates the maximum value of kA that they are rated at.

Circuit breakers to BS EN 60898 and RCBOs to BS EN 61009 show two values in boxes, usually on the front of the device.

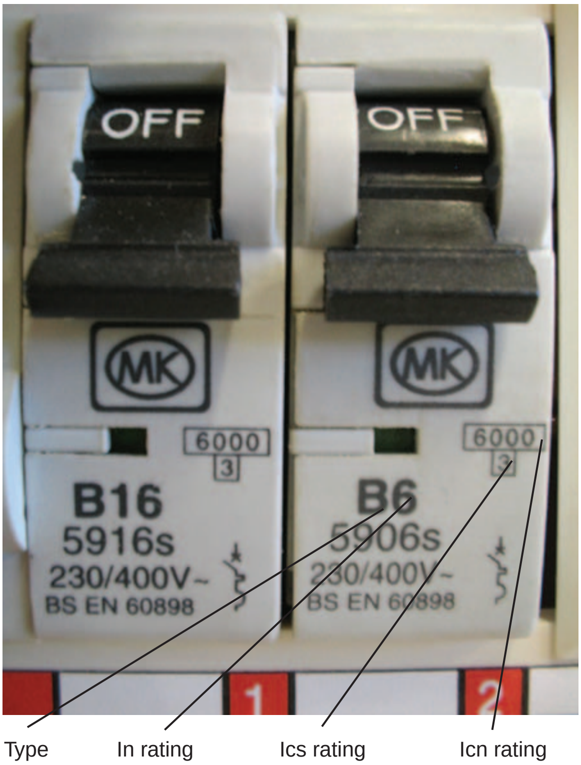
OFF
OFF
MK
MK
6000
3
6000
3
B16
5916s
230/400V~
BS EN 60898
B6
5906s
230/400V~
BS EN 60898
1
2
Type
In rating
Ics rating
Icn rating

Circuit breaches to BS EN 60898

The square box will indicate the maximum current that the device could interrupt and still be reset $\boxed{3}$. This is the Ics rating. The rectangular box will indicate the maximum current that the device can interrupt safely $\boxed{6000}$. This is the Icn rating.

If a value of fault current above the rated Isc rating of the device were to flow in the circuit, the device will no longer be serviceable and will have to be replaced. A value of fault current above the Icn rating would be very dangerous and possibly result in an explosion causing major damage to the distribution board/consumer's unit.

Functional testing

All equipment must be tested to ensure that it operates correctly. All switches, isolators and circuit breakers must be manually operated to ensure that they function correctly, also that they have been correctly installed and adjusted where adjustment is required.

Residual current device (RCD)

The instrument used for this test is an RCD tester, and it measures the time it takes for the RCD to interrupt the supply of current flowing through it. The value of measurement is either in seconds or milliseconds.

Before we get on to testing, let's consider what types of RCDs there are, what they are used for, and where they should be used.

Types of RCD

Voltage operated

Voltage operated earth leakage current breakers (ELCBs) are not uncommon in older installations. This type of device became obsolete in the early 1980s and must not be installed in a new installation or alteration as they are no longer recognized by BS 7671.

They are easily recognized as they have two earth connections, one for the earth electrode and the other for the installation earthing conductor. The major problem with voltage operated devices is that a parallel path in the system will probably stop it from operating.

These types of devices would normally have been used as earth fault protection in a TT system.

Although the Electrical Wiring Regulations BS 7671 cannot insist that all of these devices are changed, if you have to carry out work on a system which has one it must be replaced to enable certification to be carried out correctly. If, however, a voltage operated device is found while preparing a periodic inspection report, a recommendation that it should be replaced would be the correct way of dealing with it.

BS 4293 General purpose device

These RCDs are very common in installations although they ceased to be used in the early 1990s. They have been replaced by BS EN 61008-1, BS EN 61008-2-1 and BS EN 61008-2-2.

They are used as standalone devices or main switches fitted in consumers' units/distribution boards.

This type of device provides protection against earth fault current. They will commonly be found in TT systems 15 or more years old, although they may be found in TNS systems where greater protection was required.

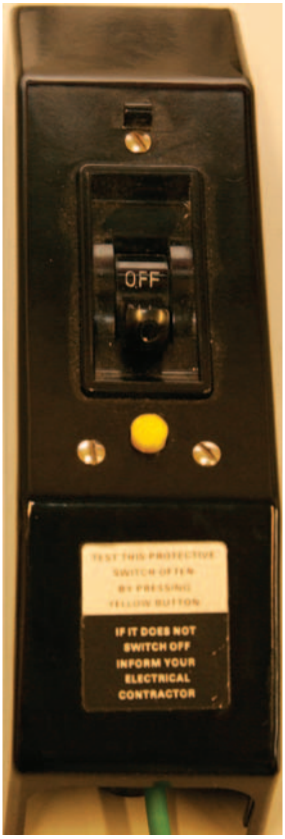
OFF
TEST THIS PROTECTIVE SWITCH OFTEN BY PRESSING YELLOW BUTTON
IF IT DOES NOT SWITCH OFF INFORM YOUR ELECTRICAL CONTRACTOR

A typical voltage – operated earth leakage current breaker (ELCB) – now obsolete

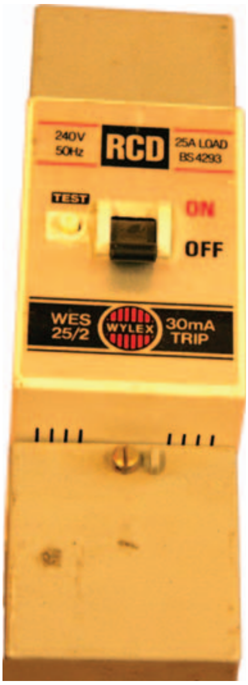
240V 50Hz
RCD
25A LOAD BS4293
TEST
ON
OFF
WES 25/2
WYLEX
30mA TRIP

A typical general purpose RCD to BS4293 – now obsolete

If this type of device is fitted to a TT system which is being extended or altered, it is quite safe to leave it in the system. If the system supplies socket outlets which could be used to supply portable equipment used outside, it must have a tripping current of no more than 30 mA. This includes socket outlets that could serve extension leads passed through open windows or doors.

The problem with using a low tripping current device as the main switch is that nuisance tripping could occur. The modern way of tackling this is explained later in this chapter.

A typical 2-phase RCCB to BS EN 61008-1

BS 4293 Type S

These are time delayed RCDs and are used to give good discrimination with other RCDs.

BS EN 61008-1 General purpose device

This is the current standard for a residual current circuit breaker (RCCB) and provides protection against earth fault current. These devices are generally used as main switches in consumers' units/distribution boards.

Three-phase devices are also very common.

BS 7288

This is the current standard for RCD-protected socket outlets and provides protection against earth fault currents. These socket outlets would be used in areas where there is an increased risk of electric shock, such as common areas of schools and colleges. It is also a requirement that any socket outlet used for portable equipment outdoors must have supplementary protection provided by an RCD (*Regulation 471-16-02*). Where the socket outlets are sited outside, waterproof BS 7288 outlets are used to IP 56.

A typical 3-phase RCCB to BS EN 61008-1

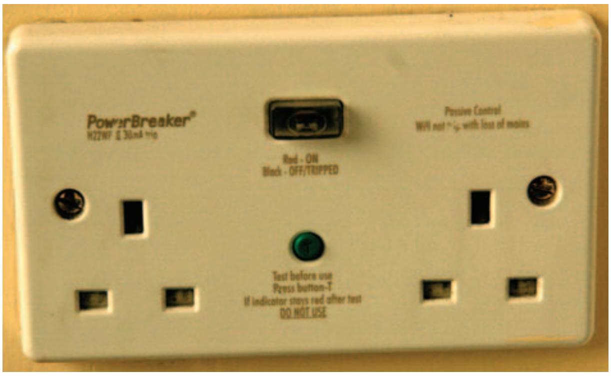

BS EN 61009-1

This is the standard for a residual current circuit breaker with overload (RCCBO) protection.

These devices are generally used to provide single circuits with earth fault protection, overload protection and short circuit protection. They are fitted in place of miniature circuit breakers and the correct type should be used (types B, C or D).

BS EN 61008-1 Type S

These are time delayed RCDs and are used to give good discrimination with other RCDs.

Section 3 of the *On-Site Guide* gives good examples of how these devices should be used within an installation.

RCDs and supply systems

TT system

If the installation is a TT system and it is possible for any of the sockets to supply portable equipment outdoors or RCD protection is required for other

A typical RCCBO to BS EN 61009-1

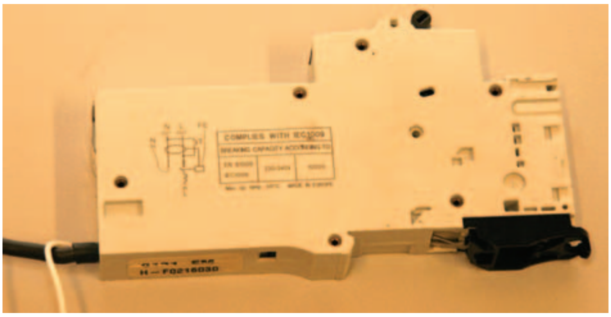

Side view of a typical RCCBC to BS EN 61009-1 reasons (*protection for fixed equipment in zones 1-3 in bathrooms, for instance*), there are various options available. While all will be safe, they will vary in cost.

Option 1 Use a 100 mA S-Type RCD (*BS EN 610081*) as the main switch and RCD protection for all circuits. Then use a 30 mA RCBO (*BS EN 61009*) as a circuit protective device for the circuit which requires protection. In this case, the 100 mA RCD must be labelled 'Main Switch'.

Option 2 Use a split board with 100 mA S-Type RCD (*BS EN 61008*) as a main switch and a 30 mA RCD (*BS EN61008*). All circuits require supplementary protection. This method is useful if more than one circuit requires 30 mA protection.

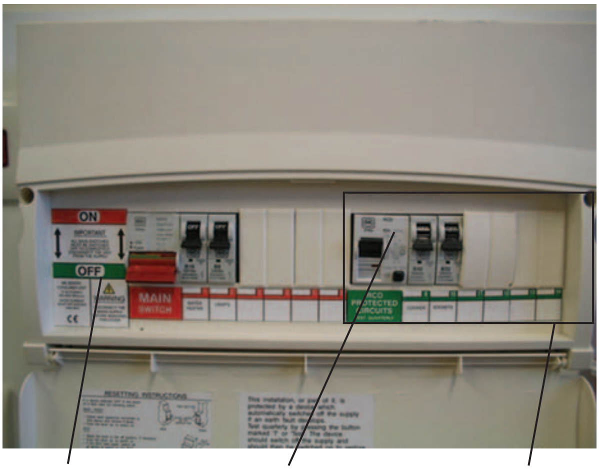

100 mA S-type RCD (BS EN 61008) as main switch

30 mA RCD (BS EN 61008) protection

Only this part is RCD – protected

Option 3 Use a split board with a 100 mA RCD for fixed equipment and lighting with a 30 mA RCD for the circuits requiring supplementary bonding. This option would require a separate main switch.

Option 4 Another method would be to use a consumer unit with a main switch to BSEN 60947-3 and RCBOs to BS EN 610091 as protective devices for all circuits. This option is perfectly satisfactory but can work out a little expensive!

TNS and TNCS systems

If the supply system was TNS or TNCS, Options 1 and 2 could be used. But instead of using a 100 mA S-Type RCD as a main switch, it could be substituted for a main switch to BS EN 60947-3.

Option 4 will remain pretty much the same but BS EN 60898 devices could be used where RCD protection is not required.

Testing of RCDs

Remember that these are live tests and care should be taken whilst carrying them out

The instrument to be used to carry out this test is an RCD tester, with leads to comply with GS 38.

Before carrying out the tests ensure that all loads are removed; failure to do this may result in the readings being inaccurate.

Voltage operated (ELCBs)

No test required as they should now be replaced.

BS 4293 RCDs

If this type of RCD is found on TT systems or other systems where there is a high value of earth fault impedance (Z_e), the RCD tester should be plugged into the nearest socket or connected as close as possible to the RCD. The tester should then be set at the rated tripping current of the RCD ($I\Delta n$); for example, at 30 mA (*be careful and do not mistake the tripping current for the current rating of the device*).

STEP 1

The test instrument must then be set at 50% of the tripping current (15 mA).

STEP 2

Push the test button of the instrument, the RCD should not trip.

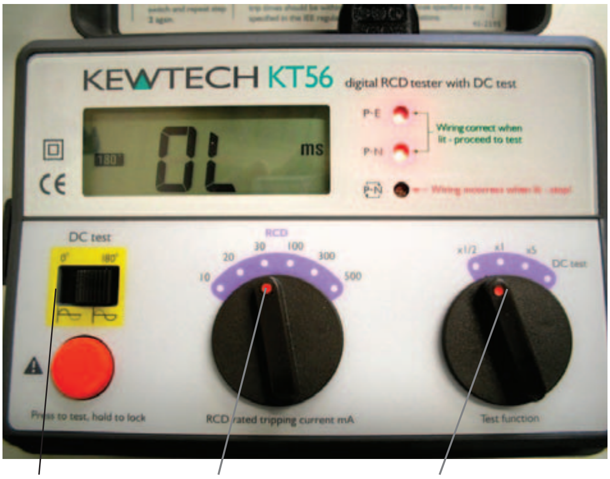
KEWTECH KT56 digital RCD tester with DC test
0L
ms
P-E
P-N
Wiring correct when lit - proceed to test
DC test
RCD
Press to test, hold to lock
RCD rated tripping current mA
Test function
Switch on 180°
Set to rated tripping current of the RCD
Set at 50% of the tripping current

STEP 3

The test instrument will have a switch on it which will enable the instrument to test the other side of the waveform 0°~180°. This switch must be moved to the opposite side and the test repeated. Again the RCD should not trip.

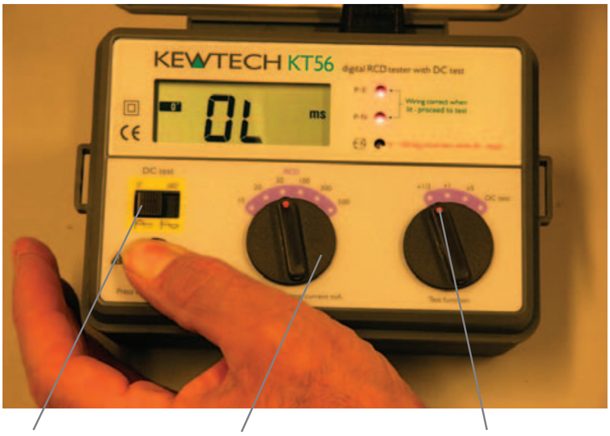
KEWTECH KT56 digital RCD tester with DC test
0L
ms
Switch on 0°
Set to rated tripping current of the RCD
Set at 50% of the tripping current

If, while testing an RCD it trips during the 50% test, do not automatically assume that the RCD is at fault.

Consider the possibility that there is a small earth leakage on the circuit or system. Switch all circuits off and test RCD on the load side at 50% using fly leads. If it still trips, then the RCD should be replaced.

If the RDC does not trip, then turn each circuit on one at a time, carrying out a 50% test each time a circuit has been turned on. When the RCD does trip, switch off all circuits except the last one which was switched on. Test again. If the RCD trips carry out an insulation test on this circuit as it probably has a low insulation resistance. If, however, the RCD does not trip it could be an accumulation of earth leakage from several circuits and they should all be tested for insulation resistance.

STEP 4

Now set the test current to the rated tripping current (30 mA)

STEP 5

Push the test button, and the RCD should trip within 200 milliseconds.

STEP 6

Reset the RCD.

STEP 7

Move the waveform switch to the opposite side, and repeat the test. Again it must trip within 200 milliseconds.

STEP 8

Reset the RCD and the slowest time in which it trips should be entered on to the test result schedule.

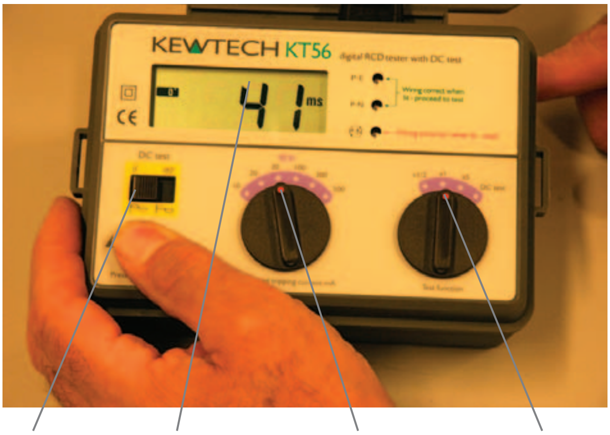

Switch on 0°

Tripping time <200 ms

Set to rated tripping current of the RCD

Set to 100% of the tripping current

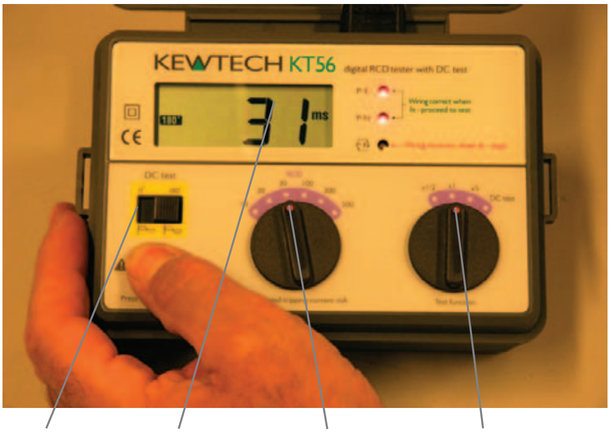

Switch on 180°

Tripping time <200 ms

Set to rated tripping current of the RCD

Set to 100% of the tripping current

If the RCD is used to provide supplementary protection against direct contact, or for socket outlets used to supply portable equipment outdoors – fixed equipment in zone 1 or other current using equipment in zone 3 of a bathroom, the following test should be performed.

STEP 9

Set the test current to 5 times the rated tripping current (150 mA).

STEP 10

Push the test button and the RCD should trip within 40 milliseconds.

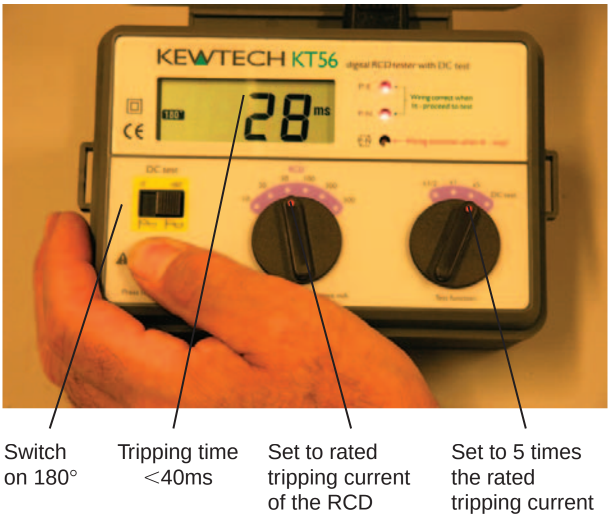
KEWTECH KT56
180°
28 ms
Switch on 180°
Tripping time <40ms
Set to rated tripping current of the RCD
Set to 5 times the rated tripping current

STEP 11

Move the waveform switch to the opposite side, and repeat the test. Again it must trip within 40 milliseconds (five times faster than the times 1 test).

After completion of the instrument tests

STEP 12

Push integral test button on RCD to verify that the mechanical parts are working correctly.

STEP 13

Ensure that a label is in place to inform the user of the necessity to use the test button quarterly.

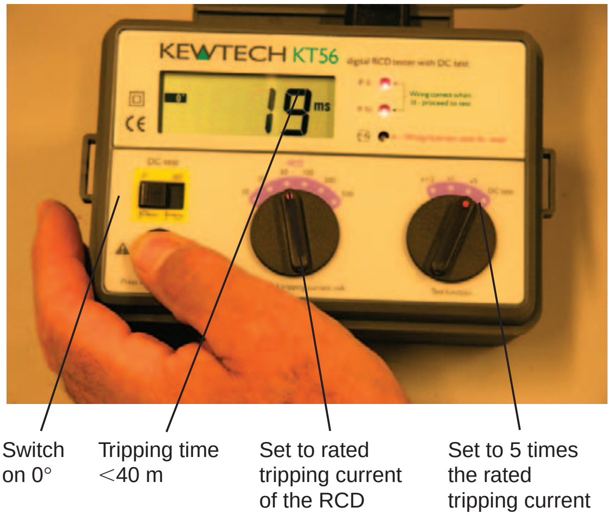
KEWTECH KT56
0°
18 ms
Switch on 0°
Tripping time <40 m
Set to rated tripping current of the RCD
Set to 5 times the rated tripping current

BS EN 610081

These devices should be tested in exactly the same manner as BS 4293 using the same test instrument. However, the difference is that, when carrying out the 100% test, the tripping time is increased to 300 milliseconds.

The 5 times test must only be carried out on RCDs with trip ratings (IΔn) up to 30 mA.

BS 4293 Type S

This device has a built in time delay. The simple way to think about this is that it does not recognize a fault for 200 milliseconds, and they must trip within 200 milliseconds after that.

STEP 1 Plug in or connect an RCD as close as possible to the RCD to be tested.

STEP 2 Set the instrument on the trip current of the RCD and ensure that it is set for 'S' type.

STEP 3 Test at 50% and the device should not trip.

STEP 4 Repeat the test on the opposite waveform.

STEP 5 Set test instrument on 100% and carry out test. The RCD should trip within 400 milliseconds (200 *ms time delay and 200 ms fault*).

STEP 6 Repeat on the opposite wave form.

The slowest operating time at the 100% test should be recorded, as should the fact that it is an 'S' type.

Always ensure that it is safe to carry out these tests. Remember to remove any loads, ensure that the disconnecting of the supply due to the test will not effect any equipment or cause damage. If any people are within the building ensure that they are aware of testing being carried out, and that a loss of supply is likely.

BS EN 61008 Type S

This device has a time delay of 200 milliseconds and a tripping time of 300 milliseconds, making a maximum tripping time of 500 milliseconds.

The test should be carried out as the BS 4293 Type S but remember the different tripping time.

BS 7288 RCD protected socket

This device should be tested the same as for a BS 4293 and the tripping times are the same.

Consideration should be given to whether the socket will supply portable equipment outdoors. If it can it should be tested at 5 times its rating.

BS EN 61009 RCBOs

These devices should be tested as for BS 4239 RCDs but the disconnection times are:

- 50% test on both sides of waveform, no trip.
- 100% test on both sides of waveform; must trip within 300 milliseconds.
- If used as supplementary protection, the 5 times test must also be carried out; it must trip within 40 milliseconds.

5

Safety in electrical testing

Correct selection of protective devices

(***While protective devices are mentioned throughout this book, this chapter brings all of the information together for reference***)

When carrying out an inspection and test on any electrical installation it is important to ensure that the correct size and type of device has been installed.

To do this we must have a good knowledge of the selection of protective devices and the type of circuits that they are protecting.

Why are they installed?

Protective devices are installed to protect the cable of the circuit from damage – this could be caused by overload, overcurrent and fault current.

The definition for overload given in Part 2 of BS 7671:2004 is: *Overcurrent occurring in a circuit which is electrically sound*. This is when the circuit is installed correctly and the equipment connected to it is drawing too much current. For instance:

An electric motor connected to the circuit is used on too heavy a load, leading to an overload of the circuit. Provided that the correct size of protective device was installed, the device will operate and interrupt the supply preventing the cable from overloading.

If additional luminaries were installed on an existing circuit which was already fully loaded, the protective device should operate and protect the cable of the circuit.

Overcurrent is a current flow in a circuit which is greater than the rated current carrying capacity of the cables. This would normally be due to a fault on the circuit or incorrect cable selection. For example:

If a 20 amp cable protected by a 32 amp circuit breaker was loaded by 25 A, the cable would overheat and the device would continue to allow current to flow – this could damage the cable.

A fault current is a current which is flowing in a circuit due to a fault. For example:

A nail is driven through a cable causing an earth fault or a short circuit fault. This would cause a very high current to flow through the circuit, which must be interrupted before the conductors reach a temperature that could damage the insulation or even the conductors.

So what are we looking for with regard to protective devices during an inspection?

What type of device is it? Is it a fuse or circuit breaker?

A fuse has an element which melts when too much current is passed through it, whether by overload or fault current.

Fuses in common use are:

BS 3036 semi-rewirable fuse
BS 88 cartridge fuse
BS 1361 cartridge fuse

A circuit breaker is really two devices in one unit. The overload part of the device is a thermal bi metal strip, which heats up when a current of a higher value than the nominal current rating (I_n) of the device passes through it.

Also incorporated within the device is a magnetic trip, which operates and causes the device to trip when a fault current flows through it. For the device to operate correctly it must operate within 0.1 seconds. The current which has to flow to operate the device in the required time has the symbol (I_a).

Circuit breakers in common use are:

- ☞ BS 3871 types 1, 2 and 3
- ☞ BS EN types B, C and D (*A is not used, this is to avoid confusion with Amps*)

Is the device being used for protection against indirect contact?

In most instances this will be the case.

What type of circuit is the device protecting, is it supplying fixed equipment only, or could it supply handheld equipment?

If the circuit supplies fixed equipment only, the device must operate on fault current within 5 seconds. If it supplies socket outlets it must operate within 0.4 seconds (s*ee BS 7671, Regulations 413-02-08 and 413-02-09*).

When using circuit breakers to BS 3871 and BS EN 60898 these times can be disregarded. Providing the correct Z_S values are met, they will operate in 0.1 seconds or less.

If it is a circuit breaker is it the correct type?

Table 7.2b of the *On-Site Guide* provides a good reference for this.

Types 1 and B should be used on circuits having only resistive loads (*have you ever plugged in your 110 v site transformer and found that it operated the circuit breaker? If you have it will be because it was a type 1 or B*).

Types 2, C and 3 should be used for inductive loads such as fluorescent lighting, small electric motors and other circuits, where surges could occur.

Types 4 and D should be used on circuits supplying large transformers or any circuits where high inrush currents could occur.

Will the device be able to safely interrupt the prospective fault current which could flow in the event of a fault?

Table 7.2A or the manufacturer's literature will provide information on the rated short circuit capacity of protective devices.

Is the device correctly coordinated with the load and the cable?

Correct coordination is:

> Current carrying capacity of the cable under its installed conditions must be equal to or greater than the rated current of the protective device (I_z).

> The rated current carrying capacity of the protective device (I_n) must be equal to or greater than the design current of the load (I_b).

In short, $I_z > I_n > I_b$ (*Appendix 4. item 4 BS 7671 or Appendix 6 On-Site Guide*).

Additional information regarding circuit breakers

Overload current

The symbol for the current required to cause a protective device to operate within the required time on overload is (I_2).

Circuit breakers with nominal ratings up to 60 amps must operate within 1 hour at 1.45 × their nominal rating (I_n).

Circuit breakers with nominal ratings above 60 amps must operate within 2 hours at 1.45 × its rating (I_n).

At 2.55 times the nominal rating (I_n), circuit breakers up to 32 amps must operate within 1 minute; and circuit breakers above 32 amps, must operate within 2 minutes.

They must not trip within 1 hour at up to 1.13 their nominal rating (I_n).

Maximum earth fault loop impedance values (Z_S) for circuit breakers

These values can be found in Table 41B2 in BS 7671, they can also be found in Sections 604 and 605 of BS 7671.

Because these devices are required to operate within 0.1 of a second, they will satisfy the requirements of BS 7671 with regard to disconnection times in all areas. Therefore, the Z_S values for these devices are the same wherever they are to be used (this only applies to circuit breakers) even in special locations where the disconnection time must be 0.2 seconds.

Calculation of the Maximum Z_S of circuit breakers

It is often useful to be able to calculate the maximum Z_S value for circuit breakers without the use of tables. This is quite a simple process for BS 3871 and BS EN 60898 devices. Let's use a 20A BS EN 60898 device as an example:

Table 7.2b shows that a type B device must operate within a window of 3 to 5 times its rating. As electricians we always look at the worst case scenario. Therefore, we must assume that the device will not operate until a current equal to 5 times its rating flows through it (I_a). For a 20 A type B device this will be 5 × 20 A = 100 A.

If we use a supply voltage of 240 volts – which is the assumed open circuit voltage (U_{OC}) of the supply (Appendix 3, BS 7671) – Ohm's law can be used to calculate the maximum Z_S.

$$Z_S = \frac{U_{OC}}{I_a}$$

$$Z_S = \frac{240}{100} = 2.4\ \Omega$$

And to check this, we look in Table 41B2 of BS 7671, we will see that the value Z_S for a 20 A type B device is 2.4 Ω.

Now let's use the same procedure for a 20 A type C device. Table 7.2b shows us that a type C device must operate at a maximum of 10 times its rating (I_n).

$$10 \times 20 = 200\ \text{A}$$

$$\frac{240}{200} = 1.2\ \Omega$$

If we check again in Table 41B2, we will see that the maximum Z_S for a 20 A type C device is 1.2 Ω.

A type C circuit breaker with a nominal operating current (I_n) must operate at a maximum of 20 times its rating:

$$20 \times 20 = 400\ \text{A}$$

$$\frac{240}{400} = 0.6\ \Omega$$

Again, if we check in Table 41B2 we will see that the Z_S value is 0.6 Ω.

We can see that the maximum Z_S values for a type C are 50% of the Z_S value of a type B device, and that the Z_S value for a type D are 50% of the Z_S value of a type C device.

Maximum earth loop impedance (Z_S) for fuses

Fuses have to operate at 0.4 or 5 seconds depending on the type of circuit which they are protecting. To find the maximum permissible Z_S value for a fuse, the current curves in Appendix 3 of BS 7671 will have to be looked at.

Example

Find the maximum permissible Z_S for a BS 1361 fuse with a rating of 20 amps with a required disconnection time of 0.4 seconds.

Look in Appendix 3, Figure 3.1. The left-hand side of the grid represents the disconnection time.

From the bottom left-hand corner, follow the line upwards until the horizontal line representing 0.4 seconds is found. Now follow the horizontal line across to the right until the bold line for a 20 amp fuse is found, from where the horizontal line touches the bold line move vertically down the page until you meet the bottom line of the grid. The bottom line represents the automatic operating current for the fuse. It can be seen that the current required is around 130 amps.

The table in the right-hand top corner of the page will show this value to be 135 amps.

The calculation is:

$$Z_S = \frac{U_{OC}}{I_a}$$

$$Z_S = \frac{240}{135}$$

$$Z_S = 1.777\ \Omega$$

Rounded up, this is 1.77 Ω. As a check, look in BS 7671, part 4, chapter 1, Table 41B1 and you will see that the Z_S for a BS 1361 20 amp fuse is 1.77 Ω.

Do not forget to correct this value for the conductor operating temperature and ambient temperature, if required.

This calculation can be used for any type of protective device. But remember that the disconnection time for a circuit breaker will always need to be 0.1 seconds, and that for some special locations it will be 0.2 seconds.

Comparing maximum Z_S and measured Z_S

Unfortunately, we cannot compare this value directly to any measured Z_S values that we have. This is because the values given in BS 7671 for Z_S are for when the circuit conductors are at their operating temperature (generally 70°C).

However, we can use a simple calculation which is called the rule-of-thumb (Section 2.7.14 of *Guidance note 3*). This calculation will allow us to compare our measured values with the values from BS 7671.

The values given in the Table below are the worst case values. In these types of calculations we must always use the worst case values to ensure a safe installation.

From this table below, it can be seen that types D, 3 and 4 will have very low maximum permitted Z_S values. This will often result in the use of an RCD.

Circuit breaker type		Worst case tripping current	Typical uses
BS EN 60898	B	5 times its rating	General purpose with very small surge currents (i.e. small amounts needed for fluorescent lighting). Mainly domestic
	C	10 times its rating	Inductive loads. Generally commercial or industrial where higher switching surges would be found (i.e. large amounts of fluorescent lighting or motor starting)
	D	20 times its rating	Only to be used where equipment with very high inrush currents would be found
BS 3871	1	4 times its rating	As for type B
	2	7 times its rating	As for type C
	3	10 times its rating	As for type C but slightly higher inrush currents
	4	50 times its rating	As for type D

Example

Let's assume that we have a circuit protected by a 32A BS EN 60898 type B device. The measured value of Z_S is 0.98 Ω.

Following the procedure described previously:

$$5 \times 32 = 160 \text{ A}$$

$$\frac{240}{160} = 1.5\ \Omega$$

The maximum Z_S *at 70°C for the circuit is 1.5 Ω.*

To find $\frac{3}{4}$ *of 1.5 we can multiply it by 0.75.*

$$1.5 \times 0.75 = 1.125\,\Omega$$

So, 1.125 Ω now becomes our maximum value, and we can compare our measured value directly to it without having to consider the ambient temperature or the conductor operating temperature. Our measured value must be less than the corrected maximum value; in this case it is, and the 32 amp type B device would be safe to use.

This type of calculation must be understood by any student studying for the City and Guilds 2391 inspecting and testing course

Test equipment

It is important that the test equipment you choose is suitable for your needs. Some electricians prefer to use individual items of equipment for each test; others like to use multi-function instruments.

Any test instruments used for testing in areas such as petrol filling stations, or areas where there are banks of storage batteries, etc. In fact, anywhere there is a risk of explosion, must be intrinsically safe for the purpose.

Most electricians are aware that electrical test instruments must comply with BS standards. However, students studying for Part P and City and Guilds 2391 exams, an understanding of the basic operational requirement for the most common types of test instrument is very important.

Whichever instrument you choose it must be suitable for the use to which it is to be put and be manufactured to the required British standard. It is also vital that you fully understand how to operate it before you start testing.

Instruments required

Low resistance ohm meter

Used to measure the resistance and verify the continuity of conductors. This instrument must produce a test voltage of between 4 and 24 volts and a current of not less than 200 mA. The range required is 0.02 to 2 ohms although most modern instruments are self ranging and will measure higher values if required.

Insulation resistance tester

Used to measure the insulation resistance between live conductors and live conductors and earth. This test is a pressure test of conductor insulation. The instrument must deliver a current of 1 mA on a resistance of 0.5 MΩ. At 250 volts d.c. for extra low voltage circuits, 500 volts d.c. for low voltage circuits up to 500 volts a.c. and 1000 volts d.c. for circuits between 500 and 1000 volts d.c.

This instrument is sometimes called a high resistance tester as it measures values in megohms.

Earth fault loop impedance tester

This instrument allows a current of up to 25 amps to flow around the earth fault loop path. It measures the current flow and by doing so can calculate the resistance of the earth fault loop path. The values given are in ohms.

As a current of 25 amps will trip RCDs and some smaller circuit breakers, it is useful to have an instrument that can carry out low current testing where required. Use of this type of instrument will avoid the tripping of devices during testing.

Prospective short circuit current test instrument

This instrument measures the current that would flow between live conductors in the event of a short circuit. It is usually incorporated in the earth loop impedance tester and normally gives a value in kA. Some instruments give the value in ohms which then needs to be converted to amps by using Ohm's law (*use 240 volts*).

Most instruments will measure the value between phase and neutral and not between phases; to find the value between phases it is simply a matter of doubling the phase to neutral value.

Earth electrode resistance tester

This is normally a battery operated 3 or 4 terminal instrument with a current and potential spike. The values given are in ohms and the instrument instructions should be fully understood before using it. The instrument would be used where low and very accurate earth electrode resistance values are required such as for generators or transformers.

Residual current device tester

This instrument measures the tripping times of RCDs in seconds or milliseconds.

Phase rotation

This instrument is used to ensure the correct phase rotation of 3 phase supplies.

Calibration of test instruments

To carry out any kind of test properly your instruments have to be accurate. If they were not then the whole point of carrying out the test would be lost. While it is not a requirement to have instruments calibrated on an annual basis, a record must be kept to show that the instruments are regularly checked for accuracy.

Instrument accuracy can be tested using various methods. For an earth loop impedance tester all that is required is a dedicated socket outlet. Use your earth fault loop impedance instrument to measure the value of the socket outlet. This value can then be used as a reference to test the accuracy of the instrument at a later date. You can also test any other earth fault loop instrument on the dedicated outlet to check its accuracy. The loop impedance values of the socket outlet should not change.

For an insulation resistance tester, or a low resistance ohm meter, the accuracy can be checked quite simply by using various values of resistors. The instruments could even be checked against values given by another instrument. If the values given are not the same when testing against another instrument, this will indicate that one of the instruments is inaccurate and further investigation using resistors should be undertaken.

An RCD test instrument accuracy is a little more difficult to check and often the best way is to check it against another instrument. However, if you do check it in this way do not expect exactly the same values as the trip time could be slightly different each time you test it due to the instrument increasing slightly in temperature.

It is also possible to purchase a calibration instrument which will check the accuracy of all electrical test instruments. When it is found that an instrument is not accurate, then it must be returned to the manufacturer or specialist for re-calibration. This is not a job that can normally be carried out by the owner of the instrument.

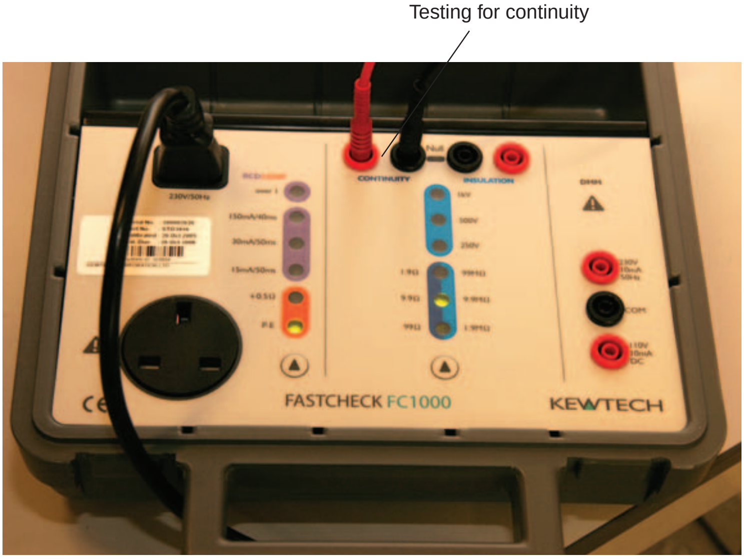
Testing for continuity
CONTINUITY
INSULATION
FASTCHECK FC1000
KEWTECH

Calibration instrument

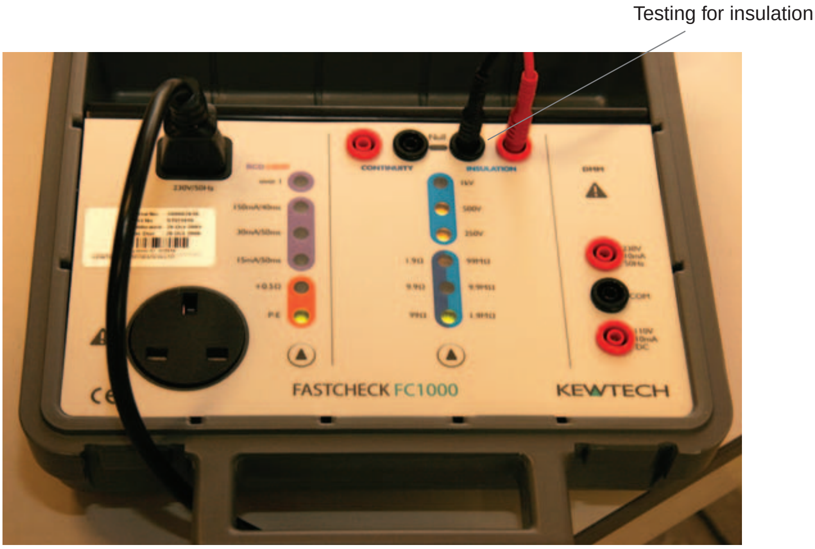
Testing for insulation
CONTINUITY
INSULATION
FASTCHECK FC1000
KEWTECH

Calibration instrument

If, for any reason, your instrument does require re-calibration, it should be returned to the instrument manufacturer or a calibration specialist.

Recordkeeping for accuracy testing of the instrument is quite a simple but important process. A record showing the instrument model, serial number and the date of the test along with the recorded values is all that is required and will satisfy most regulatory bodies. Records can be kept in a ledger, on a computer. Alternatively, calibration registers can be purchased to make life a little easier.

Electric shock

An electric shock is caused by current flowing through a body. A very small amount – between 50 and 80 mA is considered to be lethal to most human beings, although this would of course depend on the person's health and other circumstances. In livestock the lethal current would be considerably less.

The electrical regulations are set out to provide for the safety of persons and livestock. An electric shock is one risk of injury; others are:

- Excessive temperatures likely to cause burns, fire and other injurious effects.
- Mechanical movement of electrically actuated equipment, in so far as such injury is intended to be prevented by electrical emergency switching or by switching for mechanical maintenance of non-electrical parts of such equipment.
- Explosion.

This can be found in Chapter 13 of BS 7671 (Regulation 130-01-01).

Regulation 130-02-01 tells us that persons and livestock shall be protected so far as is reasonably practical against dangers that may arise from contact with live parts of the installation. This protection can be achieved by one of the following methods.

For direct or indirect contact:

- Preventing current passing through the body of any person or livestock.
- Limiting the current which can pass through a body to a value lower than the shock current.

For indirect contact:

- Automatic disconnection of the supply in a determined time on the occurrence of a fault likely to cause a current to flow through a body in contact with exposed conductive parts, where the value of that current is equal to or greater than the shock current.

The regulations quoted mention **Direct** and **Indirect** contact.

Direct contact is electric shock received by touching a known live part which is intended to be live.

Indirect contact is electric shock received when touching exposed conductive parts made live due to a fault.

Protection can be achieved by various methods which can be found in Chapters 41 and 47 of BS 7671. The most common methods used for protection against direct contact within electrical installations are: the use of insulation, and enclosures.

Protection from electric shock from indirect contact can be by many methods. The most common method used within a normal electrical installation is by the use of Earthed Equipotential Bonding and the Automatic Disconnection of the supply (EEBADs). Class 2 equipment (double insulated) and electrical separation (shaver socket) are also very common methods.

In a single phase system, current flow is achieved by creating a difference in potential.

If we were to fill a tank with water and raise the tank a metre or so, then connect a pipe with a tap on one end of it to the tank, when we open the tap the water will flow from the tank to the open end of the pipe. This is because there is no pressure outside of the pipe – the higher we raise the tank the greater the pressure of water and therefore the greater the flow of water.

Current flow is very similar to this. If we think of voltage as pressure, then to get current to flow we have to find a way of creating a difference in pressure. This pressure in an electrical circuit is called **potential difference** and it is achieved in a single phase system by pegging the star point of the supply transformer to earth. The potential of earth is known to be at 0 volts.

If we place a load between a known voltage and earth, the current will flow from the higher voltage through the load to earth. If we increase the voltage, then more current will flow, just as more water would flow if we increased the height of the water tank.

The problem we have with electricity is that if we use our body to provide the current with a path to earth it will use it, and possibly electrocute us at the same time.

Current will not flow unless it has somewhere to flow to – that is, from a high pressure to a lower pressure, possibly zero volts but not always. It is also possible in some instances to get different voltages in an installation, particularly during a fault where volt drops may occur due to loose connections, high resistance joints and different sizes of conductors. We must also remember that during a fault it will not only be the conductors that are live, but any metalwork connected to the earthing and bonding system, either directly or indirectly. It is highly likely that a shock by indirect contact could be received between pipe work at different voltages.

In any installation, protection must be in place to prevent electric shock. The protection we use against direct contact is self-explanatory and we can only prevent unintentional touching of live parts – if a person is intent on touching a live conductor, we can only make it difficult for them, not impossible.

Protection against indirect contact is a different problem altogether and we can achieve it by different methods. Firstly, if there is a fault to earth all of the metal work connected to the earthing system, whether directly or indirectly, would become live. In the first instance we need to ensure that enough current will flow through the protective device to earth to operate the protective device very quickly. This is achieved by selecting the correct type of protective device, and ensuring that the earth fault loop path has a low enough impedance to allow enough current to flow and operate the device in the required time. On its own this is not enough and that is where the equipotential and supplementary bonding is used. The basic principle is that, if one piece of metal work becomes live, any other parts that could introduce a potential (voltage) difference also become live at the same potential. If everything within the building is at the same potential, current cannot possibly flow from one part to another via a person or livestock.

Testing transformers

It is a requirement to test isolation and SELV transformers to ensure the user's safety. It is also useful to be able to test them to ensure that they are working correctly.

Step up or down double wound transformer

Use a low reading ohm meter to test between to primary (*cable that connects to the main supply*) side. The resistance should be quite high – this will of course depend on the size of the transformer. It may be that the resistance is so high that a multi meter set on its highest resistance value will have to be used. If this is the case, then set the instrument to the highest value possible and turn it down until a reading is given. If the winding is open circuit then the transformer is faulty. Repeat this test on the secondary winding.

Now join the ends of the primary winding together and join the ends of the secondary winding together. Use an insulation resistance meter set on 500 V d.c. to test between the joined ends. Then test between the joined ends and earth. The maximum insulation value permissible in both cases is 0.5 MΩ. If the resistance is less, then the transformer is faulty.

Isolation transformer

Carry out the test in the same manner as the double wound transformer and the values for insulation resistance are the same, 0.5 MΩ.

Separated Extra Low Voltage transformers (SELV)

These transformers are tested using the same procedure as for the step up or down transformer. The insulation resistance test values are different for this test. If the SELV circuits from the secondary side of the transformer are being tested, then the test voltage must be 250 V d.c. and the maximum resistance value is 0.25 MΩ – although this would be considered a very low value and any value below 5 MΩ must be investigated.

The test voltage is increased to 500 V d.c. for a test between the actual transformer windings. The minimum insulation resistance value is 0.5 MΩ although any value below 5 Ω must be investigated.

Testing a 3 phase induction motor

There are many types of 3 phase motors but by far the most common is the induction motor. It is quite useful to be able to test them for serviceability.

Before carrying out electrical tests it is a good idea to ensure that the rotor turns freely. This may involve disconnecting any mechanical loads. The rotor should rotate easily and you should not be able to hear any rumbling from the motor bearings. Next, if the motor has a fan on the outside of it, check that it is clear of any debris which may have been sucked in to it. Also check that any air vents into the motor are not blocked.

Generally, if the motor windings are burnt out there will be an unmistakable smell of burnt varnish. However, it is still a good idea to test the windings as the smell could be from the motor being overloaded. Three phase motors are made up of three separate windings – in the terminal box there will be six terminals as each motor winding will have two ends. The ends of the motor windings will usually be identified as W1, W2; U1, U2; or V1, V2. The first part of the test is carried out using a low resistance ohm meter. Test each winding end to end (W1 to W2, U1 to U2 and V1 to V2). The resistance of each winding should be approximately the same and the resistance value will depend on the size of the motor. If the resistance values are different, then the motor will not be electrically balanced and it should be sent for rewinding. If resistance values are the same, then the next test is carried out using an insulation resistance tester. Join W1 and W2 together, U1 and U2 together and V1 and V2 together. Carry out an insulation resistance test between the joined ends, i.e. W to U then W to V and then between U and V. Then repeat the test between joined ends and the case, or the earthing terminal of the motor (*these tests can be in any order to suit you*). Providing the insulation resistance is 2 MΩ or greater then the motor is fine. If the insulation resistance is above 0.5 MΩ this could be due to dampness and it is often a good idea to run the motor for a while before carrying out the insulation test again as the motor may dry out with use.

To reconnect the motor windings in star, join W2, U2 and V2 together and connect the 3 phase motor supply to W1, U1 and V1. If the motor rotates in the wrong direction, swap two of the phases of the motor supply.

To reconnect the motor windings in delta, join W1 to U2, U1 to V2 and V1 to W2 and then connect the 3 phase motor supply one to each of the joined ends. If the motor rotates in the wrong direction, swap two phases of the motor supply.

BLACK+DECKER™

The Best DIY Series from
the Brand You Trust

ADVANCED HOME WIRING

Updated 4th Edition

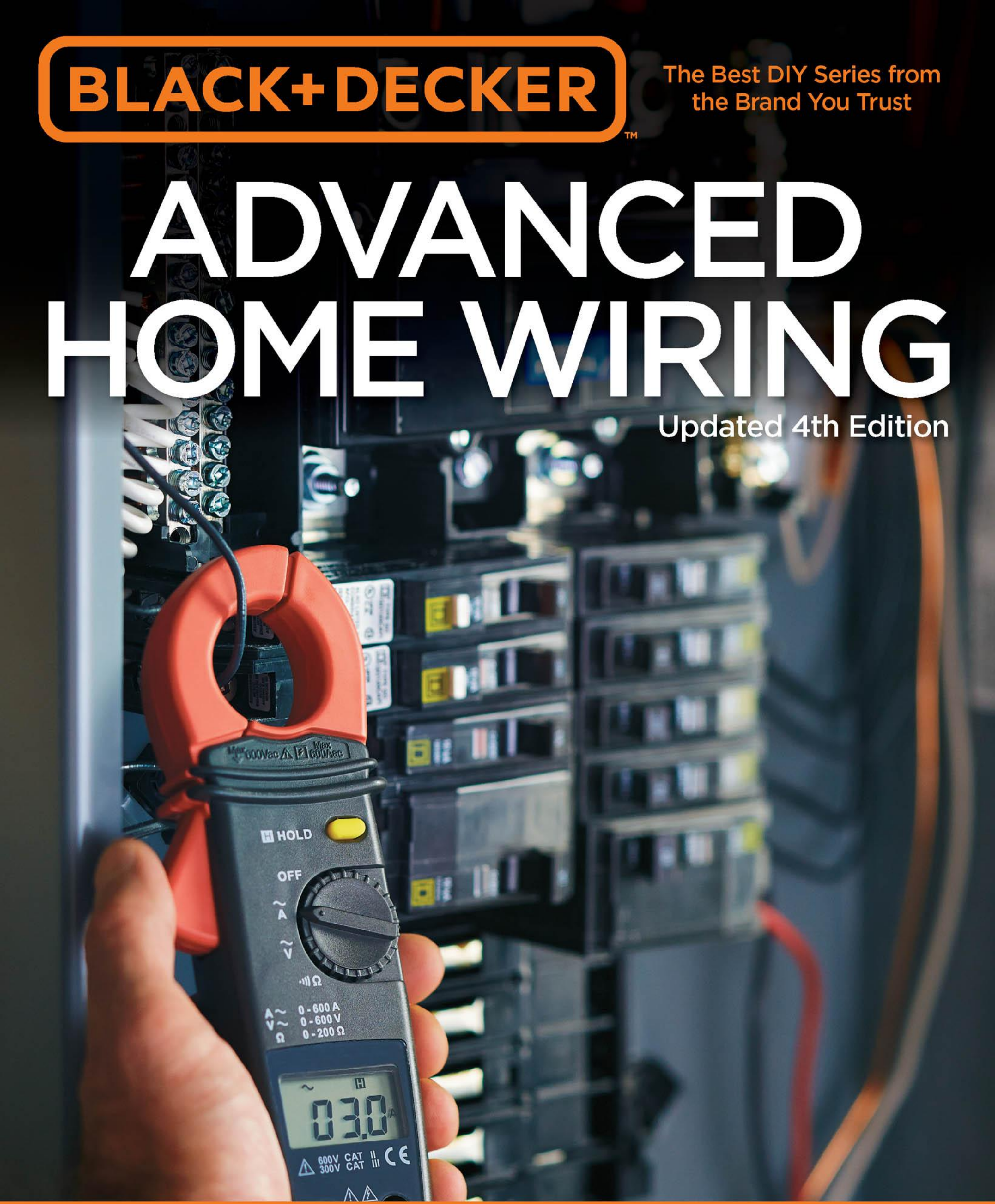
HOLD
OFF
0-600 A
0-600 V
0-200 Ω
03.0

DC Circuits • Transfer Switches
Panel Upgrades • Circuit Maps • More

Introduction

Experienced home electricians understand the need to keep up with changes in the world of wiring. Wiring projects, and more advanced projects in particular, almost always require a permit from your municipality and typically an on-site inspection or two as well. If you aren't up-to-date with wiring codes, there is a likelihood that your project will not pass inspection. But beyond the practicality of passing inspections, the codes that govern wiring practices are updated for good reason: they improve safety. And when you're talking about your own home and family, that's worth paying attention to.

This newest edition of BLACK+DECKER *Advanced Home Wiring* has been reviewed and revised to reflect the many changes to wiring code published in the 2014 edition of the National Electrical Code (NEC), which is updated every three years. And this time around there were an unusually high number of changes. They include the requirement that an available neutral wire be present in every switch box, expansion of the types of circuits that require AFCI protection, a new prohibition on tying into a garage receptacle, and more. Several of these changes are reflected in the updated information you'll find here.

Almost all of the advanced wiring projects featured in this book involve new circuitry, panel upgrades, or troubleshooting with diagnostic equipment. Among the high-level projects: making a direct-current, solar-electric circuit; upgrading the grounding and bonding on your new 200-amp or larger home circuit; installing an automatic transfer switch for your backup power supply; wiring a room addition; and using a multimeter to precisely locate an open neutral in a home circuit.

Because the projects found in this book are advanced in nature, do not attempt any of them unless you are confident in your abilities. Consult a professional electrician if you have any concerns—in many cases your best solution might be to do some of the work yourself, such as pulling new sheathed cable through walls, and to have the electrical contractor do the other work, such as making the connections. But do keep in mind that home wiring can be a fun and fascinating pursuit, and successfully accomplishing a major project is personally gratifying and can also save you substantial amounts of money.

Wiring Safety

Safety should be the primary concern of anyone working with electricity. Although most household electrical repairs are simple and straightforward, always use caution and good judgment when working with electrical wiring or devices. Common sense can prevent accidents.

The basic rule of electrical safety is: Always turn off power to the area or device you are working on. At the main service panel, remove the fuse or shut off the circuit breaker that controls the circuit you are servicing. Then check to make sure the power is off by testing for power with a voltage tester. *Tip: Test a live circuit with the voltage tester to verify that it is working before you rely on it.* Restore power only when the repair or replacement project is complete.

Follow the safety tips shown on these pages. Never attempt an electrical project beyond your skill or confidence level.

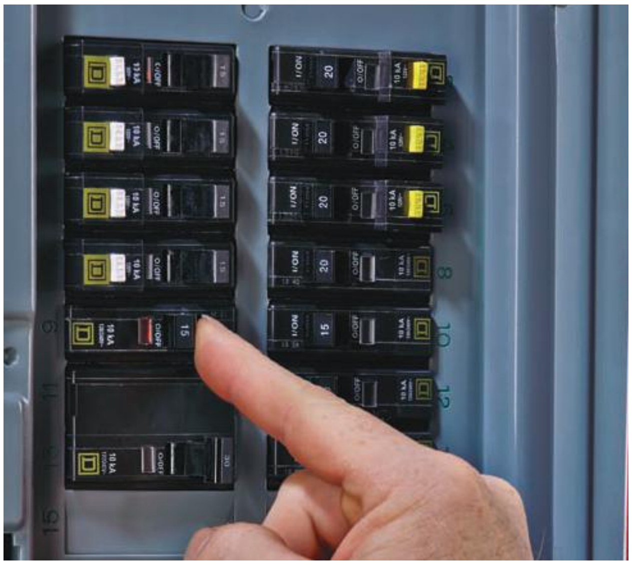

Shut power OFF at the main service panel or the main fuse box before beginning any work.

Create a circuit index and affix it to the inside of the door to your main service panel. Update it as needed.

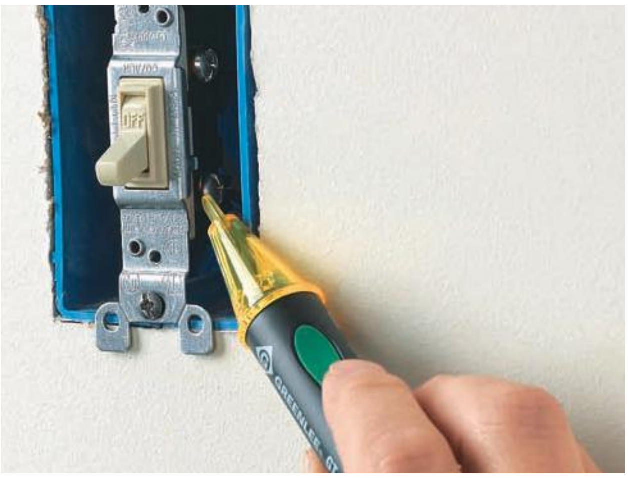

Confirm power is OFF by testing at the outlet, switch, or fixture with a voltage tester.

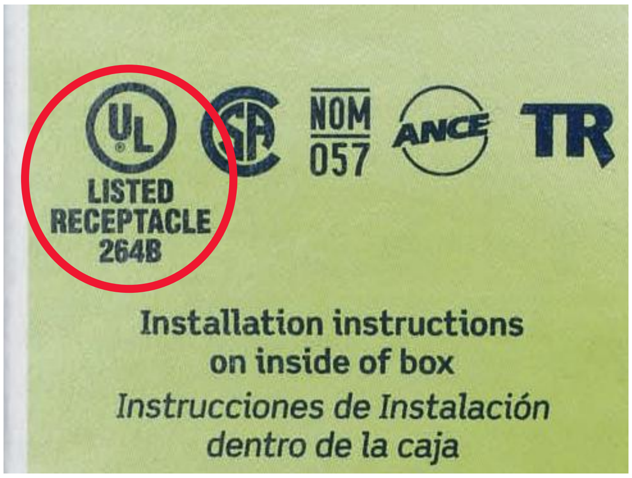

Use only UL-approved electrical parts or devices. These devices have been tested for safety by Underwriters Laboratories.

Wear rubber-soled shoes while working on electrical projects. On damp floors, stand on a rubber mat or dry wooden boards.

Use fiberglass or wood ladders when making routine household repairs near the service mast.

Extension cords are for temporary use only. Cords must be rated for the intended usage.

Electracraft
plug
fuses
32-583
TIME DELAY FUSE
20

Breakers and fuses must be compatible with the panel manufacturer and match the circuit capacity.

Never alter the prongs of a plug to fit a receptacle. If possible, install a new grounded receptacle.

Do not penetrate walls or ceilings without first shutting off electrical power to the circuits that may be hidden.

1st
proofs for reference only

Planning Your Project

Careful planning of a wiring project ensures you will have plenty of power for present and future needs. Whether you are adding circuits in a room addition, wiring a remodeled kitchen, or adding an outdoor circuit, consider all possible ways the space might be used, and plan for enough electrical service to meet peak needs.

For example, when wiring a room addition, remember that the way a room is used can change. In a room used as a spare bedroom, a single 15-amp circuit provides plenty of power, but if you ever choose to convert the same room to a family recreation space, you will need additional circuits.

When wiring a remodeled kitchen, it is a good idea to install circuits for an electric oven and countertop range, even if you do not have these electric appliances. Installing these circuits now makes it easy to convert from gas to electric appliances at a later date.

A large wiring project adds a considerable load to your main electrical service. In about 25 percent of all homes, some type of service upgrade is needed before new wiring can be installed. For example, many homeowners will need to replace an older 60-amp electrical service with a new service rated for 100 amps or more. This is a job for a licensed electrician but is well worth the investment. In other cases, the existing main service provides adequate power, but the main circuit breaker panel is too full to hold any new circuit breakers. In this case it is necessary to install a circuit breaker subpanel to provide room for hooking up added circuits. Installing a subpanel is a job most homeowners can do themselves (see pages 65 to 67).

This chapter gives an easy five-step method for determining your electrical needs and planning new circuits.

Five Steps for Planning a Wiring Project

1
200
MAIN
PRINCIPAL

Examine your main service panel (see page 10). The amp rating of the electrical service and the size of the circuit breaker panel will help you determine if a service upgrade is needed.

2
NFPA 70
2014
nec
National Electrical Code
International Electrical Code Series
NFPA

Learn about codes (see pages 11 to 13). The National Electrical Code (NEC), and local electrical codes and building codes, provide guidelines for determining how much power and how many circuits your home needs. Your local electrical inspector can tell you which regulations apply to your job.

1st
proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 8
10/30/14 10:06 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 8

3

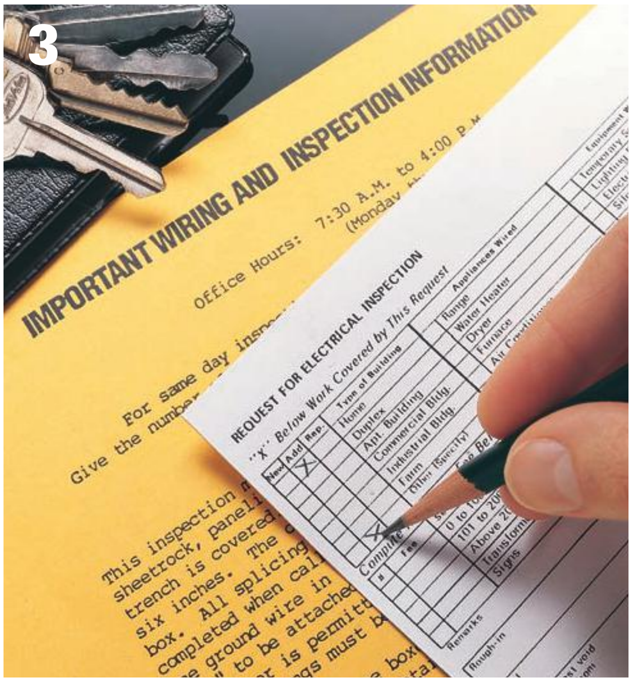
IMPORTANT WIRING AND INSPECTION INFORMATION
Office Hours: 7:30 A.M. to 4:00 P.M.
REQUEST FOR ELECTRICAL INSPECTION
"X" Below Work Covered by This Request

Prepare for inspections (see pages 14 to 15). Remember that your work must be reviewed by your local electrical inspector. When planning your wiring project, always follow the inspector's guidelines for quality workmanship.

4

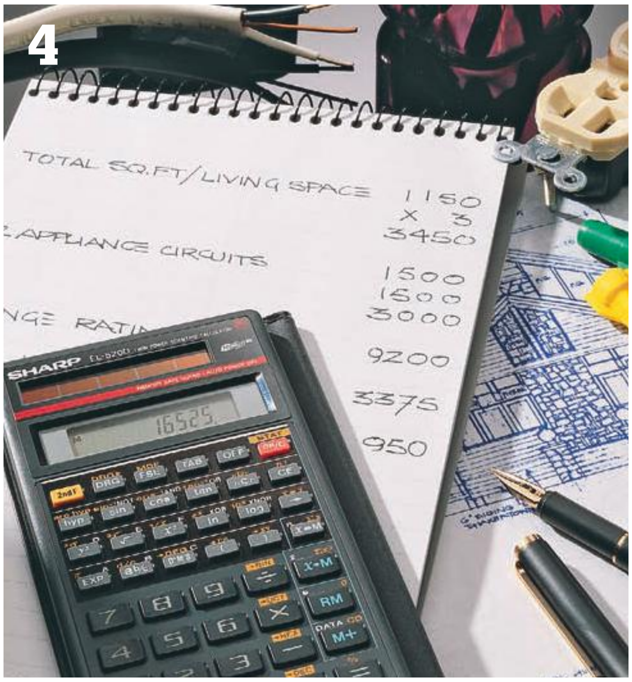
TOTAL SQ.FT/LIVING SPACE
1150
x 3
3450
APPLIANCE CIRCUITS
1500
1500
3000
9200
3375
950
SHARP

Evaluate electrical loads (see pages 16 to 19). New circuits put an added load on your electrical service. Make sure that the total load of the existing wiring and the planned new circuits does not exceed the service capacity or the capacity of the panel.

5

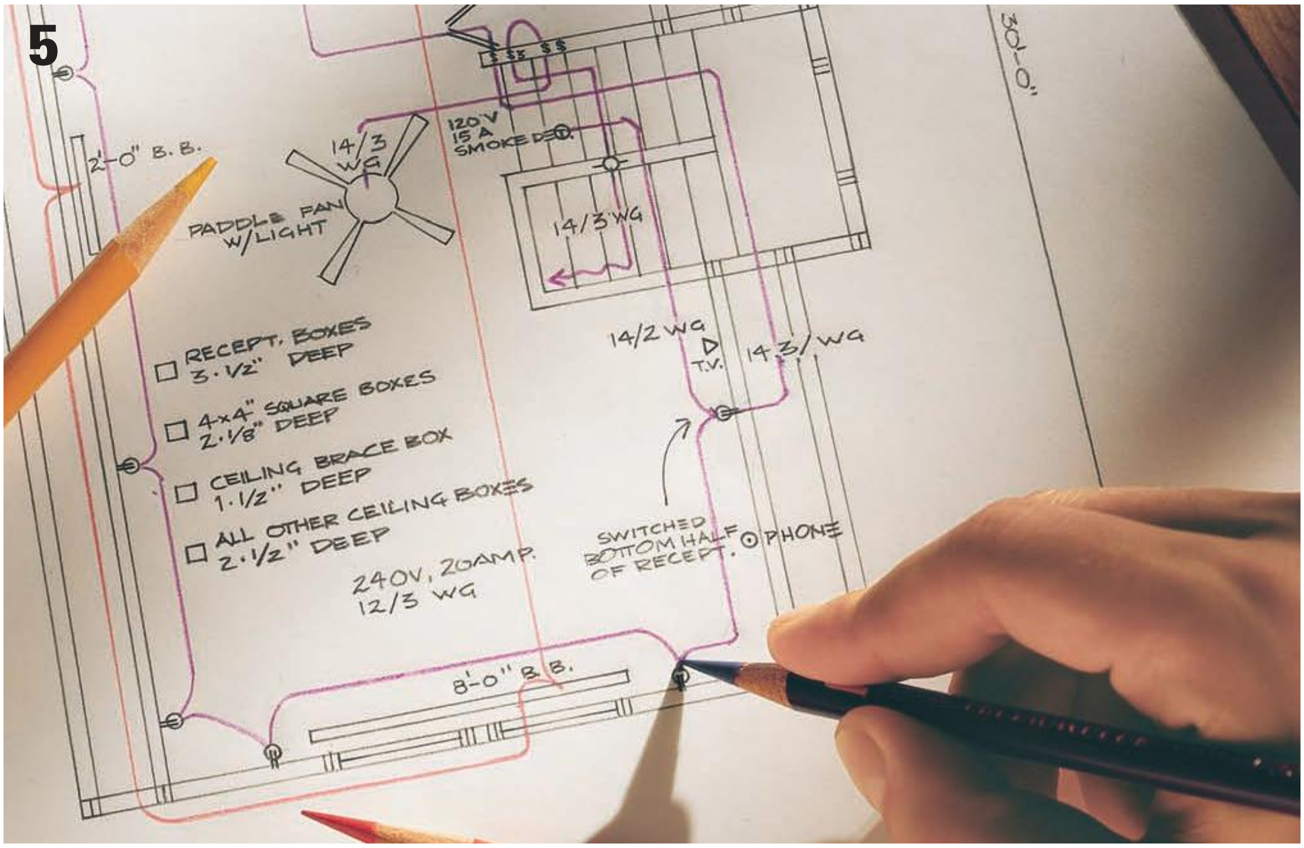
2'-0" B.B.
14/3 WG
PADDLE FAN W/LIGHT
120 V 15 A SMOKE DET.
14/3 WG
30'-0"
14/2 WG
T.V.
14/3/WG
☐ RECEPT. BOXES 3·1/2" DEEP
☐ 4×4" SQUARE BOXES 2·1/8" DEEP
☐ CEILING BRACE BOX 1·1/2" DEEP
☐ ALL OTHER CEILING BOXES 2·1/2" DEEP
240V, 20AMP. 12/3 WG
SWITCHED BOTTOM HALF OF RECEPT.
PHONE
8'-0" B.B.

Draw a wiring diagram and get a permit (see pages 20 to 21). This wiring plan will help you organize your work.

1st proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 9
10/10/14 9:06 AM
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 9
Edition
Page: 8
Pro-Vision C M Y K Text

Examine Your Main Service Panel

The first step in planning a new wiring project is to look in your main circuit breaker panel and find the size of the service by reading the amperage rating on the main circuit breaker. As you plan new circuits and evaluate electrical loads, knowing the size of the main service helps you determine if you need a service upgrade.

Also look for open circuit breaker slots in the panel. The number of open slots will determine if you need to add a circuit breaker subpanel.

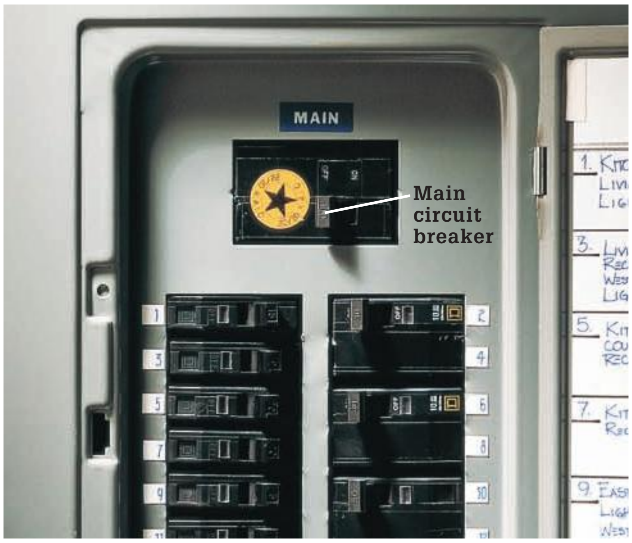

Find the service size by opening the main service panel and reading the amp rating printed on the main circuit breaker. In most cases, 100-amp service provides enough power to handle the added loads of projects such as the ones shown in this book. A service rated for 60 amps or less should be upgraded. Note: In some homes the main circuit breaker is located in a separate box.

Older service panels use fuses instead of circuit breakers. Have an electrician replace this type of panel with a circuit breaker panel that provides enough power and enough open breaker slots for the new circuits you are planning.

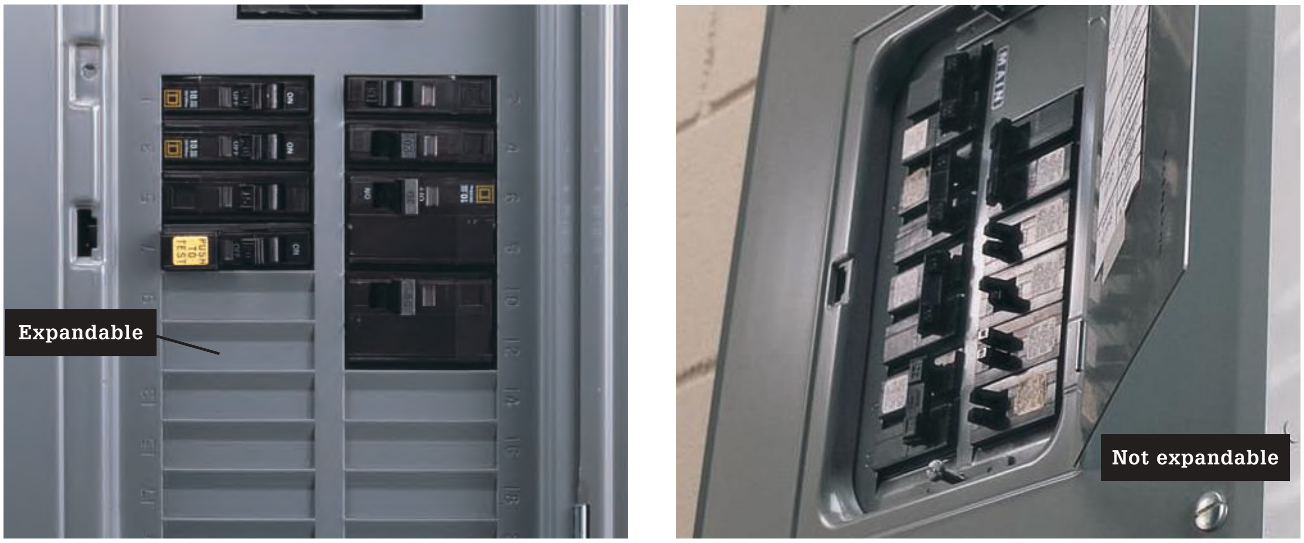

Look for open circuit breaker slots in the main circuit breaker panel or in a circuit breaker subpanel, if your home already has one. You will need one open slot for each 120-volt circuit you plan to install and two slots for each 240-volt circuit. If your main circuit breaker panel has no open breaker slots, install a subpanel (see pages 65 to 67) to provide room for connecting new circuits.

1st proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 10
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 9:06 AM
Page: 10
Pro-Vis

Learn About Codes

To ensure public safety, your community requires that you get a permit to install new wiring and have the work reviewed by an inspector. Electrical inspectors use the National Electrical Code (NEC) as the primary authority for evaluating wiring, but they also follow the local building code and electrical code standards.

Most communities use a version of the NEC that is not the most current version. Also, many communities make amendments to the NEC, and these amendments may affect your work.

As you begin planning new circuits, call or visit your local electrical inspector and discuss the project with him or her. The inspector can tell you which of the national and local code requirements apply to your job and may give you a packet of information summarizing these regulations. Later, when you apply to the inspector for a work permit, he or she will expect you to understand the local guidelines as well as a few basic NEC requirements.

The NEC is a set of standards that provides minimum safety requirements for wiring installations. It is revised every three years. The national code requirements for the projects shown in this book are thoroughly explained on the following pages. For more information, you can find copies of the current NEC, as well as a number of excellent handbooks based on the NEC, at libraries and bookstores.

In addition to being the final authority of code requirements, inspectors are electrical professionals with years of experience. Although they have busy schedules, most inspectors are happy to answer questions and help you design well-planned circuits.

Basic Electrical Code Requirements

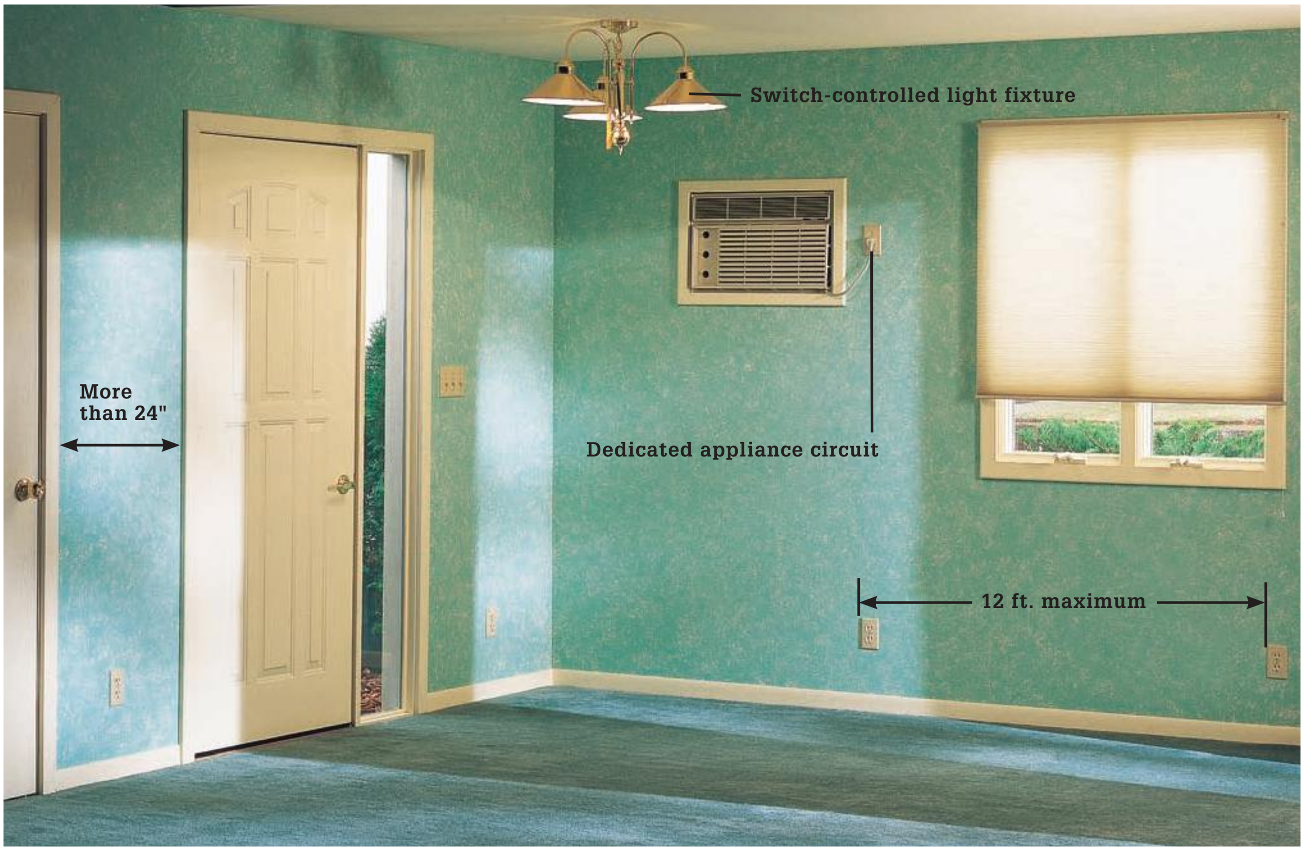

Electrical code requirements for living areas: Living areas need at least one 15-amp or 20-amp basic lighting/receptacle circuit for each 600 sq. ft. of living space and should have a dedicated circuit for each type of permanent appliance, such as an air conditioner, or a group of baseboard heaters. Receptacles on basic lighting/receptacle circuits should be spaced no more than 12 ft. apart. Many electricians and electrical inspectors recommend even closer spacing. Any wall more than 24" wide also needs a receptacle. Every room should have a wall switch at the point of entry to control either a ceiling or wall-mounted light or plug-in lamp. Kitchens and bathrooms must have a ceiling or wall-mounted light fixture.

1st proofs for reference only
Edition
age: 10
41940 - Advanced Home Wiring_001-063.indd 11
Pro-Vision C M Y K Text
10/10/14 9:06 AM
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 11

Highlights of the National Electrical Code ▸

BY MATERIAL

Panels

- Maintain a minimum 30" wide by 36" deep of clearance in front of the panel.
- Match the amperage rating of the circuit when replacing fuses.
- Use handle ties on all 240-volt breakers and on 120-volt breakers protecting multi-wire branch circuits.
- Close all unused panel openings.
- Label each fuse and breaker clearly on the panel.

Electrical Boxes

- Use boxes that are large enough to accommodate the number of wires and devices in the box.
- Install all junction boxes so they remain accessible.
- Leave no gaps greater than ⅛" between wallboard and the front of electrical boxes.
- Place receptacle boxes flush with combustible surfaces.
- Leave a minimum of 3" of usable cable or wire extending past the front of the electrical box.

Wires & Cables

- Use wires that are large enough for the amperage rating of the circuit.
- Drill holes at least 2" from the edges of joists. Do not attach cables to the bottom edge of joists.
- Do not run cables diagonally between framing members.
- Use nail plates to protect cable that is run through holes drilled or cut into studs less than 1¼" from the front edge of a stud.
- Do not crimp cables sharply.
- Contain spliced wires or connections entirely in a plastic or metal electrical box.
- Use wire connectors to join wires.
- Use staples to fasten cables within 8" of an electrical box and every 54" along its run.
- Leave a minimum ¼" (maximum 1") of sheathing where cables enter an electrical box.
- Clamp cables and wires to electrical boxes with approved clamps. No clamp is necessary for one-gang plastic boxes if cables are stapled within 8".
- Connect only a single wire to a single screw terminal. Use pigtails to join more than one wire to a screw terminal.

Switches

- Use a switch-controlled receptacle in rooms without a built-in light fixture operated by a wall switch.
- Use three-way switches at the top and bottom on stairways with six risers or more.
- Use switches with grounding screws with plastic electrical boxes.
- Locate all wall switches within easy reach of the room entrance and not behind the door.
- Install a neutral wire in switch boxes.
- Use black or red wires to supply power to switched devices.

Receptacles

- Install receptacles on all walls 24" wide or greater.
- Install receptacles so a 6-ft. cord can be plugged in from any point along a wall or every 12 ft. along a wall.
- Include receptacles in any hallway that is 10 ft. long or longer.
- Use three-slot, grounded receptacles for all 15- or 20-amp, 120-volt branch circuits.
- Include a switch-controlled receptacle in rooms without a built-in light fixture operated by a wall switch.
- Install GFCI-protected circuits in bathrooms, kitchens, garages, crawl spaces, unfinished basements, and outdoor receptacle locations.

Light Fixtures

- Use mounting straps that are anchored to the electrical boxes to mount ceiling fixtures.
- Keep non-IC-rated recessed light fixtures 3" from insulation and ½" from combustibles.
- Include at least one switch-operated lighting outlet in every room.

Grounding

- Ground receptacles by connecting receptacle grounding screws to the circuit grounding wires.
- Use switches with grounding screws whenever possible. Always ground switches installed in plastic electrical boxes and all switches in kitchens, bathrooms, and basements.

1st
proofs for reference only

41940 - Advanced Home Wiring_001-063.indd 12
Pro-Vision C M Y K Text
10/10/14 9:08 AM
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 12
Pro-Vis

BY ROOM

Kitchens/Dining Rooms

- Install at least two 20-amp small appliance receptacle circuits.
- Install dedicated 15-amp, 120-volt circuits for dishwashers and food disposals (required by many local codes).
- Install GFCI protection for all countertop receptacles; receptacles behind fixed appliances do not need to be GFCI protected.
- Position receptacles for appliances that will be installed within cabinets, such as microwaves or food disposals, according to the manufacturer's instructions.
- Include receptacles on all counters wider than 12".
- Space receptacles a maximum of 48" apart above countertops and closer together in areas where many appliances will be used.
- Locate receptacles on the wall above the countertop not more than 20" above the countertop.
- Install at least one receptacle not more than 12" below the countertop on islands and peninsulas that are 12" × 24" or greater.
- Do not connect lights to the small appliance receptacle circuits.
- Install at least one wall or ceiling-mounted light fixture.

Bathrooms

- Install a separate 20-amp GFCI-protected circuit only for bathroom receptacles.
- Ground switches in bathrooms.
- Install at least one receptacle not more than 36" from each sink.
- Install at least one ceiling- or wall-mounted light fixture.

Utility/Laundry Rooms

- Install a separate 20-amp circuit for a washing machine.
- Install approved conduit for wiring in unfinished rooms.
- Use GFCI-protected circuits for 120-volt receptacles within 6 feet from a sink (including the washing machine receptacle).

Living, Entertainment, Bedrooms

- Install at least one 15- or 20-amp lighting/receptacle circuit for each 600 sq. ft. of living space.
- Install a dedicated circuit for each permanent appliance, such as an air conditioner or group of electric baseboard heaters.
- Use electrical boxes listed and labeled to support ceiling fans.
- Include receptacles on walls 24" wide or more.
- Space receptacles on walls in living and sleeping rooms a maximum of 12 ft. apart.
- Check with your local electrical inspector about requirements for installing smoke and carbon monoxide alarms during remodeling.

Outdoors

- Check for underground utilities before digging.
- Use UF cable or other wiring approved for wet locations for outdoor wiring.
- Run cable and wires in schedule 80 PVC plastic and other approved conduit, as required by local code.
- Install in-use rated weatherproof receptacle covers.
- Bury cables and wires run in conduit at least 18" deep; cable not in conduit must be buried at least 24" deep.
- Use weatherproof electrical boxes with watertight covers.
- Install GFCI-protected circuits for receptacles.
- Support boxes that are not attached to a building and that contain switches or receptacles using at least two pieces of conduit. Secure the conduit not more than 18 feet from the box. Locate the box at least 12" above the ground.

Stairs/Hallways

- Use three-way switches at the top and bottom on stairways with six risers or more.
- Include receptacles in any hallway that is 10 ft. long or longer.
- Position stairway lights so each step and landing is illuminated.

1st proofs for reference only

Edition
age: 12
41940 - Advanced Home Wiring_001-063.indd 13
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 9:08 AM
Page: 13

Prepare for Inspections

Electrical inspectors who issue the work permit for your wiring project will also visit your home to review the work. Make sure to allow time for these inspections as you plan the project. For most projects, inspectors make two visits.

The first inspection, called the rough-in, is done after the cables are run between the boxes but before the insulation, wallboard, switches, and fixtures are installed. The second inspection, called the final, is done after the walls and ceilings are finished and all electrical connections are made.

When preparing for the rough-in inspection, make sure the area is neat. Sweep up sawdust and clean up any pieces of scrap wire or cable insulation. Before inspecting the boxes and cables, inspectors will check to make sure all plumbing and other mechanical work is completed. Some electrical inspectors will ask to see your building and plumbing permits.

At the final inspection, inspectors check random boxes to make sure the wire connections are correct. If they see good workmanship at the selected boxes, the inspection will be over quickly. However, if they spot a problem, inspectors may choose to inspect every connection.

Inspectors have busy schedules, so it is a good idea to arrange for an inspection several days in advance. In addition to basic compliance with code, inspectors expect your work to meet their own standards for quality. When you apply for a work permit, make sure you understand what the inspectors will look for during inspections.

You cannot put new circuits into use legally until an inspector approves them at the final inspection. If you have planned carefully and done your work well, electrical inspections are routine visits that give you confidence in your own skills.

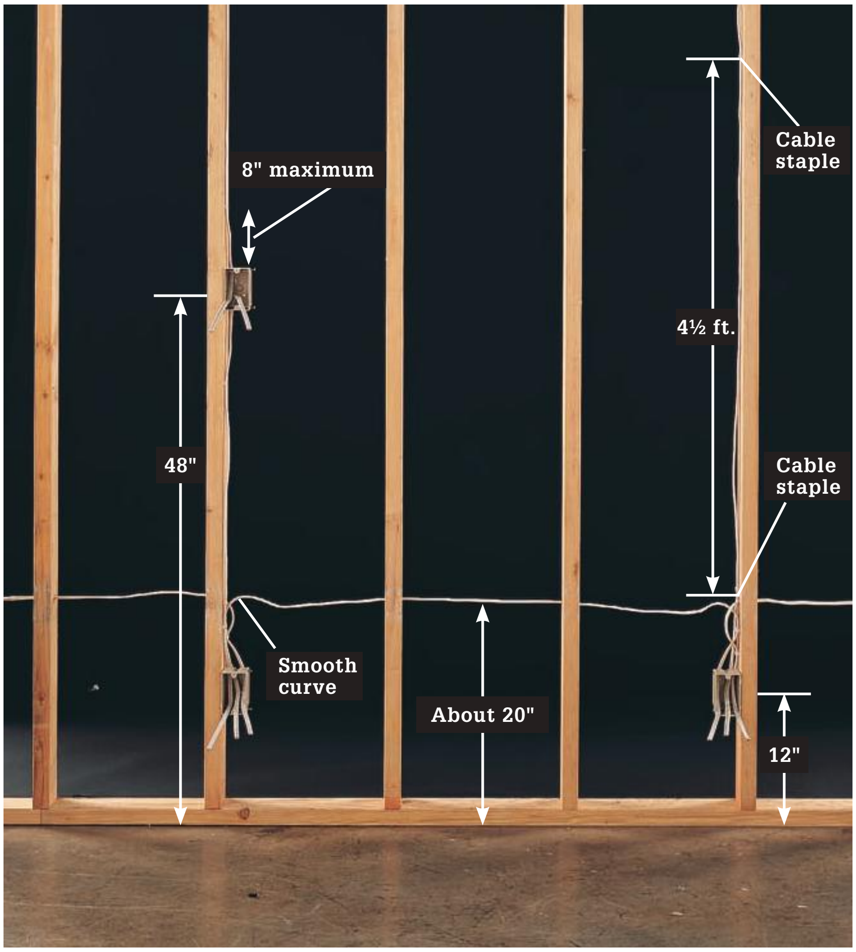

Inspectors may measure to see that electrical boxes are mounted at consistent heights. Height may not be dictated by code, but consistency is a sign of good workmanship. Measured from the center of the boxes, receptacles in living areas typically are located 12" above the finished floor and switches at 48". For special circumstances, inspectors allow you to alter these measurements. For example, you can install switches at 36" above the floor in a child's bedroom, or set receptacles at 24" to make them more convenient for someone using a wheelchair.

Inspectors will check cables to see that they are anchored by cable staples driven within 8" of each box and every 4½ ft. thereafter when they run along studs. When bending cables, form the wire in a smooth curve. Do not crimp cables sharply or install them diagonally between framing members. Some inspectors specify that cables running between receptacle boxes should be about 20" above the floor.

What Inspectors Look for

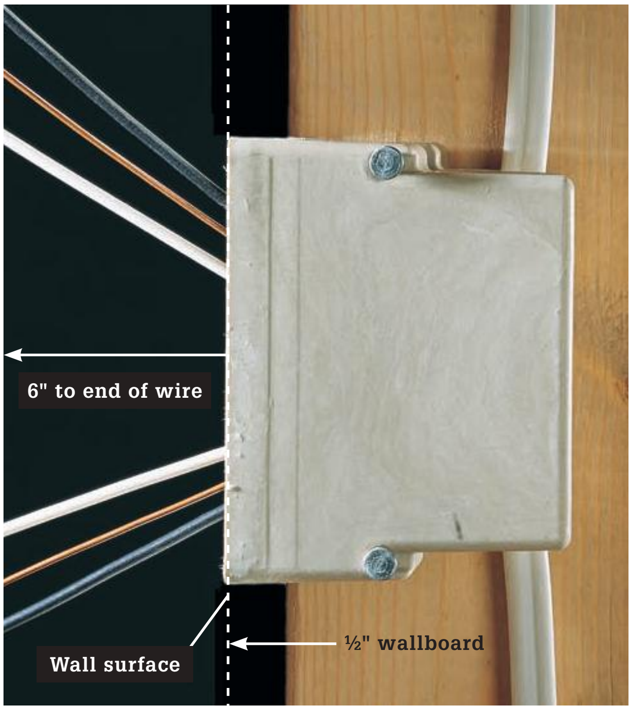
6" to end of wire
Wall surface
½" wallboard

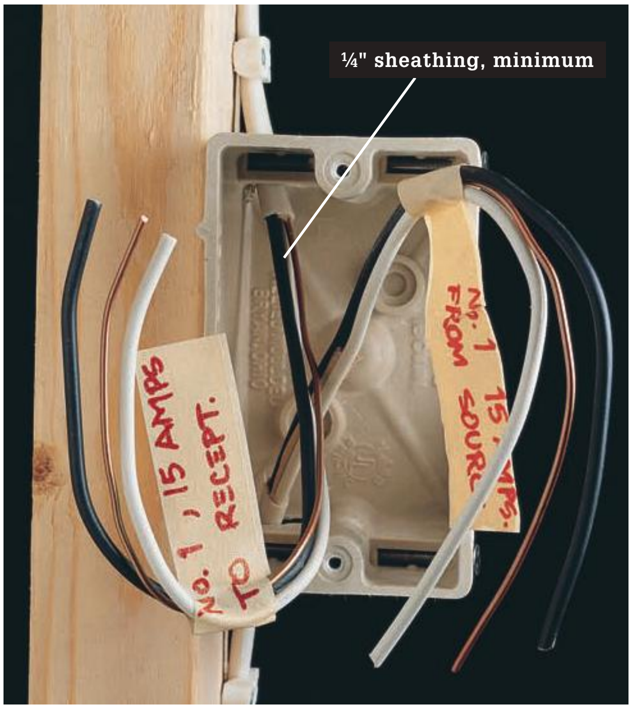
¼" sheathing, minimum
No. 1, 15 AMPS TO RECEPT.
No. 1 15 AMPS FROM SOURC

Electrical box faces should extend past the front of framing members so the boxes will be flush with finished walls (left). Inspectors will check to see that all boxes are large enough for the wires they contain. Cables should be cut and stripped back so that at least 3" of usable length extends past the front of the box and so that at least ¼" of sheathing reaches into the box (right). Label all cables to show which circuits they serve: inspectors recognize this as a mark of careful work. The labels also simplify the final hookups after the wallboard is installed.

Is your Receptacle Spacing Correct?

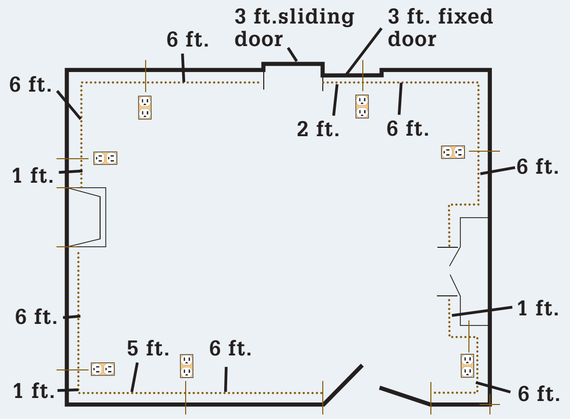
3 ft.sliding door
3 ft. fixed door
6 ft.
6 ft.
2 ft.
6 ft.
1 ft.
6 ft.
6 ft.
1 ft.
5 ft.
6 ft.
1 ft.
6 ft.

Example of receptacle spacing requirements in a typical room. Measure receptacle spacing distance along the wall line. Install receptacles along partial height walls and along balcony guards in lofts and similar areas.

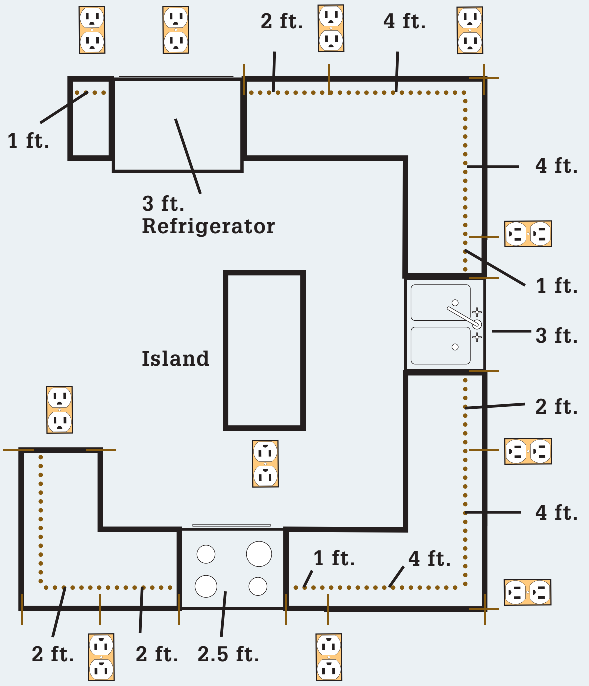
2 ft.
4 ft.
1 ft.
3 ft.
Refrigerator
4 ft.
1 ft.
3 ft.
Island
2 ft.
4 ft.
1 ft.
4 ft.
2 ft.
2 ft.
2.5 ft.

Example of countertop receptacle spacing in a typical kitchen (right).

1st proofs for reference only
Edition
age: 14
41940 - Advanced Home Wiring_001-063.indd 15
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 9:08 AM
Page: 15

Evaluate Electrical Loads

Before drawing a plan and applying for a work permit, make sure your home's electrical service provides enough power to handle the added load of the new circuits. In a safe wiring system, the current drawn by fixtures and appliances never exceeds the main service capacity.

To evaluate electrical loads, use whatever evaluation method is recommended by your electrical inspector. Include the load for all existing wiring as well as that for proposed new wiring when making your evaluation.

Most of the light fixtures and plug-in appliances in your home are evaluated as part of general allowances for basic lighting/receptacle circuits and small-appliance circuits. However, appliances that are permanently installed require their own dedicated circuits. The electrical loads for these appliances are added in separately when evaluating wiring.

If your evaluation shows that the load exceeds the main service capacity, you must have an electrician upgrade the main service before you can install new wiring. An electrical service upgrade is a worthwhile investment that improves the value of your home and provides plenty of power for present and future wiring projects.

Amperage ▸

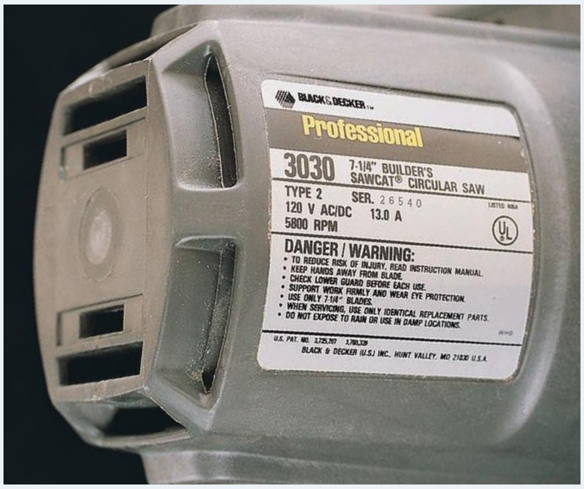

Amperage rating can be used to find the wattage of an appliance. Multiply the amperage by the voltage of the circuit. For example, a 13-amp, 120-volt circular saw is rated for 1,560 watts.

AMPS × VOLTS	TOTAL CAPACITY	SAFE CAPACITY
15 A × 120 V =	1,800 watts	1,440 watts
20 A × 120 V =	2,400 watts	1,920 watts
25 A × 120 V =	3,000 watts	2,400 watts
30 A × 120 V =	3,600 watts	2,880 watts
20 A × 240 V =	4,800 watts	3,840 watts
30 A × 240 V =	7,200 watts	5,760 watts

1st proofs for reference only

41940 - Advanced Home Wiring_001-063.indd 16
10/10/14 9:08 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 16
Pro-Vis

Calculating Loads

Add 1,500 watts for each small appliance circuit required by the local electrical code. In most communities, three such circuits are required—two in the kitchen and one for the laundry—for a total of 4,500 watts. No further calculations are needed for appliances that plug into small-appliance or basic lighting/receptacle circuits.

S.A.49022
KILOWATTS
240 VOLTS
12.3
LIS
923
Nameplate

If the nameplate gives the rating in kilowatts, find the watts by multiplying kilowatts times 1,000. If an appliance lists only amps, find watts by multiplying the amps times the voltage—either 120 or 240 volts.

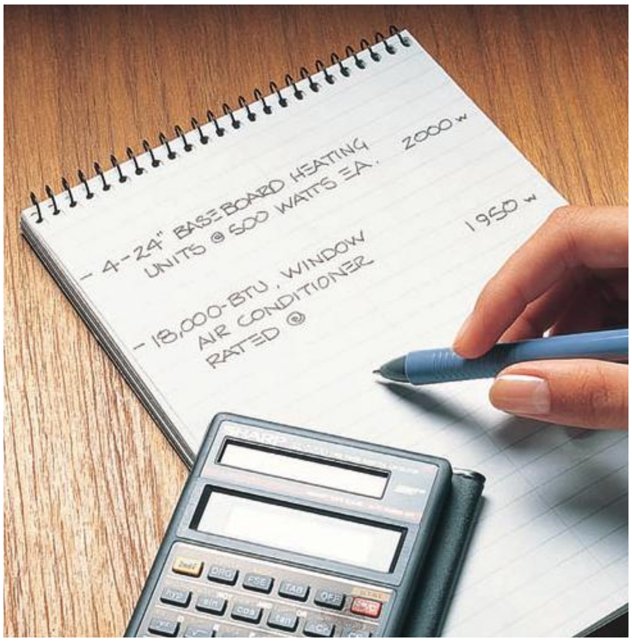
- 4-24" BASEBOARD HEATING UNITS @ 500 WATTS EA.
2000 w
- 18,000-BTU, WINDOW AIR CONDITIONER RATED @
1950 w

Air-conditioning and heating appliances are not used at the same time, so figure in only the larger of these two numbers when evaluating your home's electrical load.

Fixed Devices ▸

Do not connect one or more fixed devices that in total exceed 50 percent of a multiple outlet branch circuit's amperage rating. Fixed devices do not include light fixtures. This means that that all fixed devices (such as a permanently wired disposal or hot water circulating pump) on a multiple outlet branch circuit may not exceed 7.5 amps (about 900 watts) on a 15-amp multiple outlet branch circuit and may not exceed 10 amps (about 1,200 watts) on a 20-amp multiple outlet branch circuit.

Locating Wattage

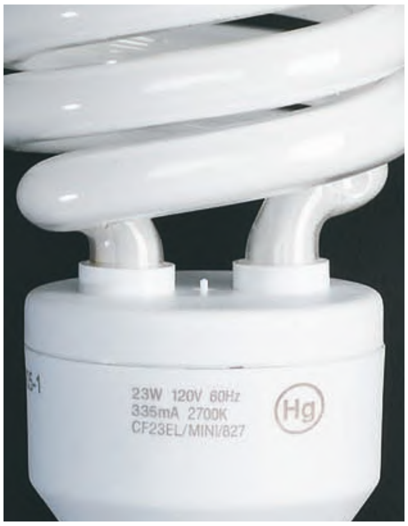

Light bulb wattage ratings are printed on the top of the bulb. If a light fixture has more than one bulb, remember to add the wattages of all the bulbs to find the total wattage of the fixture.

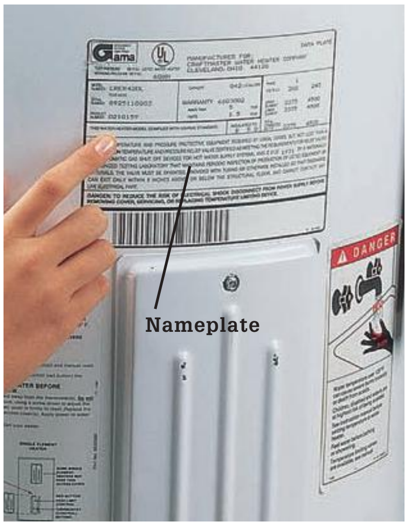

Electric water heaters are permanent appliances that require their own dedicated 30-amp, 240-volt circuits. Most water heaters are rated between 3,500 and 4,500 watts. If the nameplate lists several wattage ratings, use the one labeled “Total Connected Wattage” when figuring electrical loads.

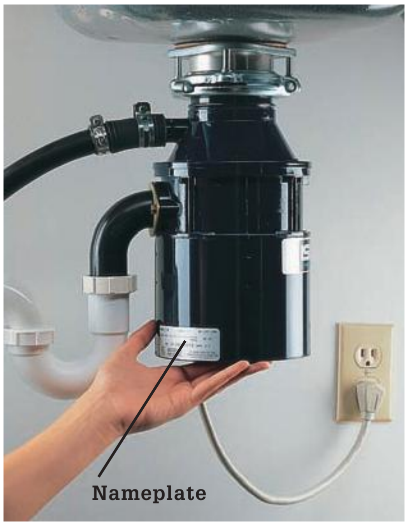

Food disposers are considered permanent appliances and may require their own dedicated 15-amp, 120-volt circuits. Most disposers are rated between 500 and 900 watts.

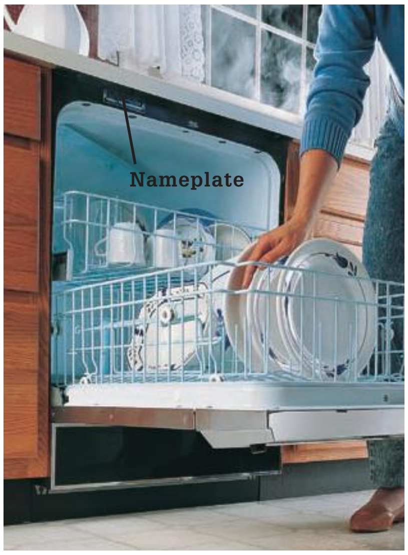

Dishwashers installed permanently under a countertop may need a dedicated 15-amp, 120-volt circuits. Dishwasher ratings are usually between 1,000 and 1,500 watts. Portable dishwashers are regarded as part of small appliance circuits and are not added in when figuring loads.

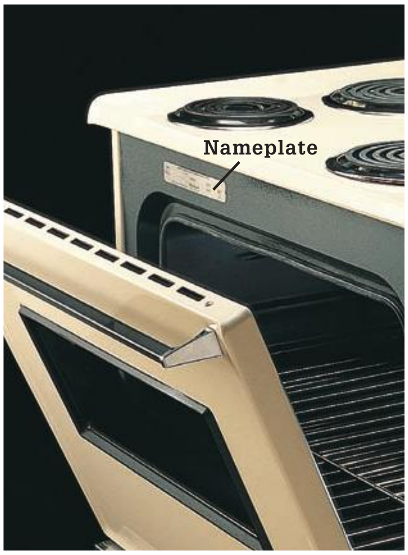

Electric ranges can be rated for as little as 3,000 watts or as much as 12,000 watts. They require dedicated 120/240-volt circuits. Find the exact wattage rating by reading the nameplate found inside the oven door or on the back of the unit.

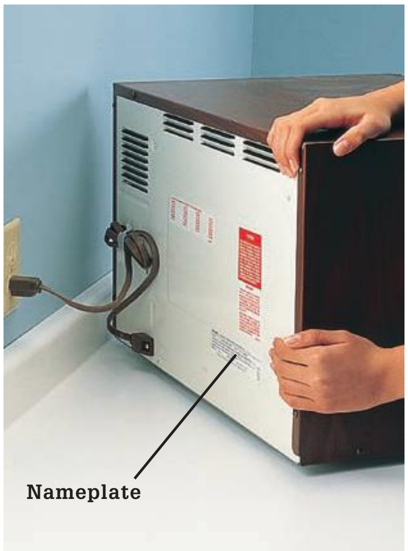

Microwave ovens are regarded as permanent appliances. Add in its wattage rating when calculating loads. The nameplate is found on the back of the cabinet or inside the front door. Most microwave ovens are rated between 500 and 1,200 watts.

1st proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 18
10/10/14 9:06 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 18
Pro-Vis

Freezers are permanent appliances that may need a dedicated 15- or 20-amp, 120-volt circuits. Freezer ratings are usually between 240 and 480 watts. But combination refrigerator-freezers rated for 1,000 watts or less are plugged into small appliance circuits and do not need their own dedicated circuits. The nameplate for a freezer is found inside the door or on the back of the unit, just below the door seal.

Electric clothes dryers are permanent appliances that need dedicated 30-amp, 120/240-volt circuits. The wattage rating is printed on the nameplate inside the dryer door. Use 5,000 watts as a minimum, regardless of the printed rating. Washing machines and gas-heat clothes dryers with electric tumbler motors do not need dedicated circuits. They plug into the 20-amp small-appliance circuit in the laundry room.

Forced-air furnaces and heat pump air handlers have electric fans and are considered permanent appliances. They require dedicated 15-amp, 120-volt circuits. Include the fan wattage rating, printed on a nameplate inside the control panel, when figuring wattage loads for heating. You should also include the wattage rating for heat pump backup heating coils.

A central air conditioner requires a dedicated 240-volt circuit. Estimate its wattage rating by adding the numbers labeled RLA and FLA on the air conditioner's metal plate. Multiply the RLA+FLA by 240.

Window air conditioners, both 120-volt and 240-volt types, are permanent appliances that require dedicated circuits. The wattage rating, which can range from 500 to 2,000 watts, is found on the nameplate located inside the front grill. Make sure to include all window air conditioners in your evaluation.

Electric room heaters that are permanently installed require a dedicated circuit and must be figured into the load calculations. Use the maximum wattage rating printed inside the cover. In general, 240-volt baseboard-type heaters are rated for 180 to 250 watts for each linear foot.

1st proofs for reference only
Edition
age: 18
41940 - Advanced Home Wiring_001-063.indd 19
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 9:06 AM
Page: 19

1st
proofs for reference only

GFCI & AFCI Breakers

Understanding the difference between GFCI (ground-fault circuit interrupter) and AFCI (arc fault circuit interrupter) is tricky for most homeowners. Essentially it comes down to this: Arc-fault interrupters keep your house from burning down; ground-fault interrupters keep people from being electrocuted.

The National Electric Code (NEC) requires that an AFCI breaker be installed on most branch circuits that supply outlets or fixtures in newly constructed homes. The NEC also requires adding AFCI protection to these circuits when you add new circuits and modify or extend existing circuits. They're a prudent precaution in any home, especially if it has older wiring. AFCI breakers will not interfere with the operation of GFCI receptacles, so it is safe to install an AFCI breaker on a circuit that contains GFCI receptacles. For a discussion on codes that concern AFCI and GFCI breakers see pages 120 to 121.

GROUND-FAULT CIRCUIT-INTERRUPTERS

A GFCI is an important safety device that disconnects a circuit in the event of a ground fault (when current takes a path other than the neutral back to the panel).

On new construction, GFCI protection is required for receptacles in these locations: kitchen counter tops, bathrooms, garages, unfinished basements, crawlspaces, outdoors, within six feet of sinks, and in unfinished accessory buildings such as storage and work sheds. In general it is a good practice to protect all receptacle and fixture locations that could encounter damp or wet circumstances.

Tools & Materials ▸

Insulated screwdriver
Circuit tester
Combination tool
AFCI or GFCI breaker

ARC-FAULT CIRCUIT INTERRUPTERS

AFCIs detect arcing (sparks) that can cause fires between and along damaged wires. AFCI protection is required for 15- and 20-amp, 120-volt circuits that serve living rooms, family rooms, dens, parlors, libraries, dining rooms, bedrooms, sun rooms, kitchens, laundry areas, closets, and hallways. AFCI protection is not required for circuits serving bathrooms, garages, the exterior of the home, appliances such as furnaces and air handlers.

The easiest way to provide AFCI protection for a circuit is to install an AFCI circuit breaker labeled as a "combination" device in the electrical panel. The 2014 NEC allows several alternate methods of providing AFCI protection, but you should consult an electrician before using these alternate methods. You should install combination AFCI circuit breakers when installing new circuits that require AFCI protection. You should install either combination AFCI circuit breakers or AFCI receptacles when you modify, replace, or extend an existing circuit that requires AFCI protection.

AFCI breakers (left) are similar in appearance to GFCI breakers (right), but they function differently. AFCI breakers trip when they sense an arc fault. GFCI breakers trip when they sense fault between the hot wire and the ground.

An AFCI-protected receptacle

41940 - Advanced Home Wiring_001-063.indd 48
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/30/14 10:06 AM
Page: 48
Pro-Vis

How to Install an AFCI or GFCI Breaker

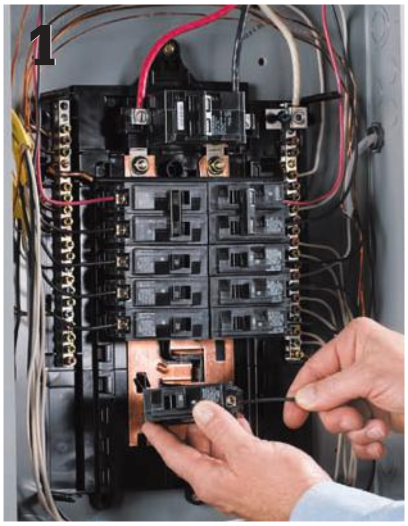

Locate the breaker for the circuit you'd like to protect. Turn off the main circuit breaker. Remove the cover from the panel, and test to ensure that power is off. Remove the breaker you want to replace from the panel. Remove the black wire from the LOAD terminal of the breaker.

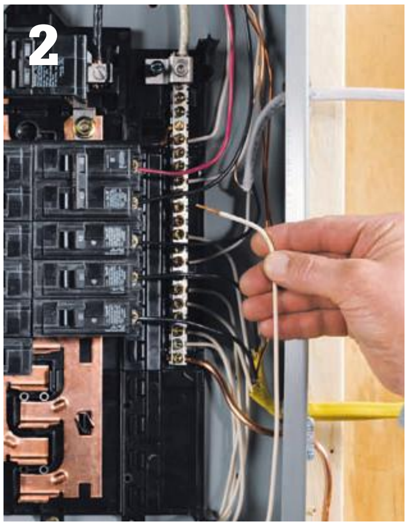

Find the white wire on the circuit you want to protect, and remove it from the neutral bus bar.

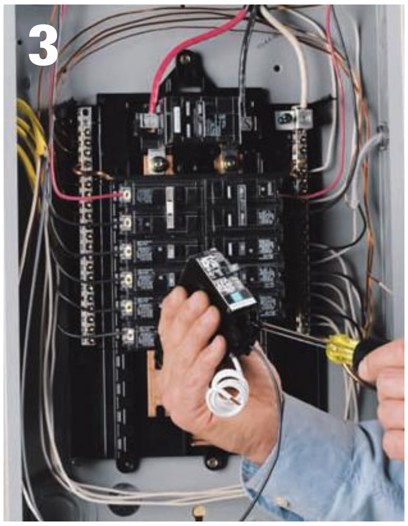

Flip the handle of the new AFCI or GFCI breaker to OFF. Loosen both of the breaker's terminal screws. Connect the white circuit wire to the breaker terminal labeled PANEL NEUTRAL. Connect the black circuit wire to the breaker terminal labeled LOAD POWER.

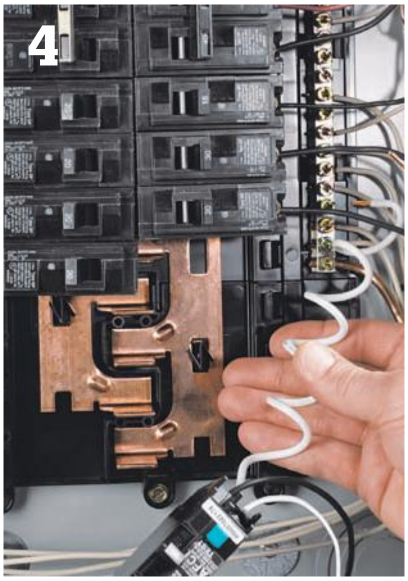

Connect the new breaker's coiled white wire to the neutral bus bar on the service panel.

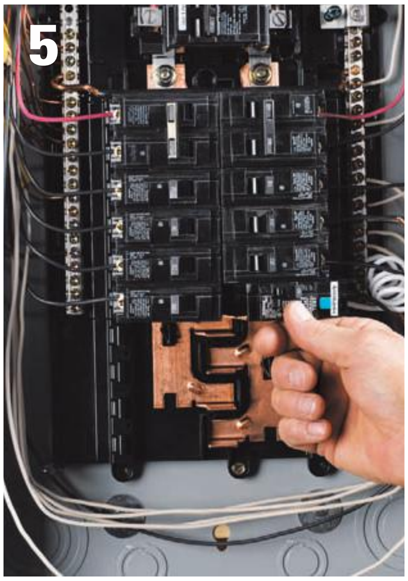

Make sure all the connections are tight. Snap the new breaker into the bus bar.

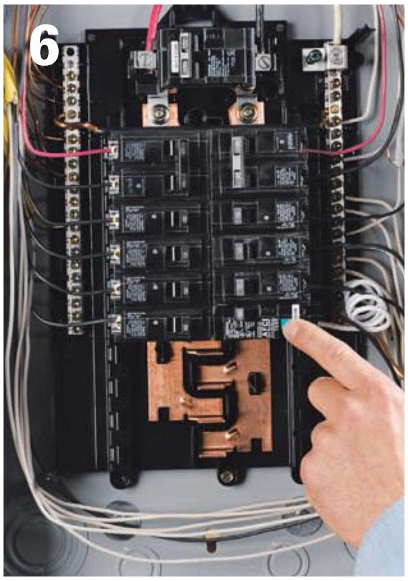

Turn the main breaker on. Turn off and unplug all fixtures and appliances on the AFCI or GFCI breaker circuit. Turn the AFCI or GFCI breaker on. Press the test button. If the breaker is wired correctly, the breaker trips open. If it doesn't trip, check all connections or consult an electrician. Replace the panel cover.

1st proofs for reference only
Edition
age: 48
41940 - Advanced Home Wiring_001-063.indd 49
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 9:06 AM
Page: 49

Replacing a Service Panel

Only a generation ago, fuse boxes were commonplace. But as our demands for power increased, homeowners replaced the 60-amp boxes with larger, safer, and more reliable circuit breaker panels. Typical new homes were built with perfectly adequate 100-amp load centers. But today, as average home size has risen to more than 2,500 sq. ft. and the number of home electronics has risen exponentially, 100 amps is often inadequate service. As a result, many homeowners have upgraded to 200-amp service, and new single-family homes often include 250 amps or even 400 amp service.

Upgrading your electrical service panel from 100 amps to 200 amps is an ambitious project that requires a lot of forethought. The first step is to obtain a permit. When you are ready to begin, you will need to have your utility company disconnect your house from electrical service at the transformer that feeds your house. When you schedule this, talk to your utility company about the size of your service drop or lateral. That may need to be upgraded too. Not only does this involve working them into your schedule, it means you will have no power during the project. You can rent a portable generator to provide a circuit or two, or you can run a couple extension cords from a friendly neighbor. But unless you are a very fast worker, plan on being without power for at least one to two days while the project is in process.

Also check with your utility company to make sure you know what equipment is theirs and what belongs to you. In most cases, the electric meter and everything on the street side belongs to the power company, and the meter base and everything on the house side is yours. Be aware that if you tamper with the sealed meter in any way, you likely will be fined. Utility companies will not re-energize your system without approval from your inspecting agency.

Upgrading a service panel is a major project. Do not hesitate to call for help at any point if you're unsure what to do.

Tools & Materials ▸

- 200-amp load center (service panel)
- 200-amp bypass meter base
- Circuit breakers (AFCI if required by local code)
- Schedule 80 or RMC conduit and fittings
- Weatherhead
- Service cable
- Circuit wires
- Plywood backer board
- Screwdrivers
- Drill/driver
- Tape
- Allen wrench
- Circuit tester
- Multimeter

Modern homeowners consume more power than our forebears, and it is often necessary to upgrade the electrical service to keep pace. While homeowners are not allowed to make the final electrical service connections, removing the old panel and installing the new panel and meter base yourself can save you hundreds or even thousands of dollars.

1st proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 50
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/30/14 10:06 AM
Page: 50

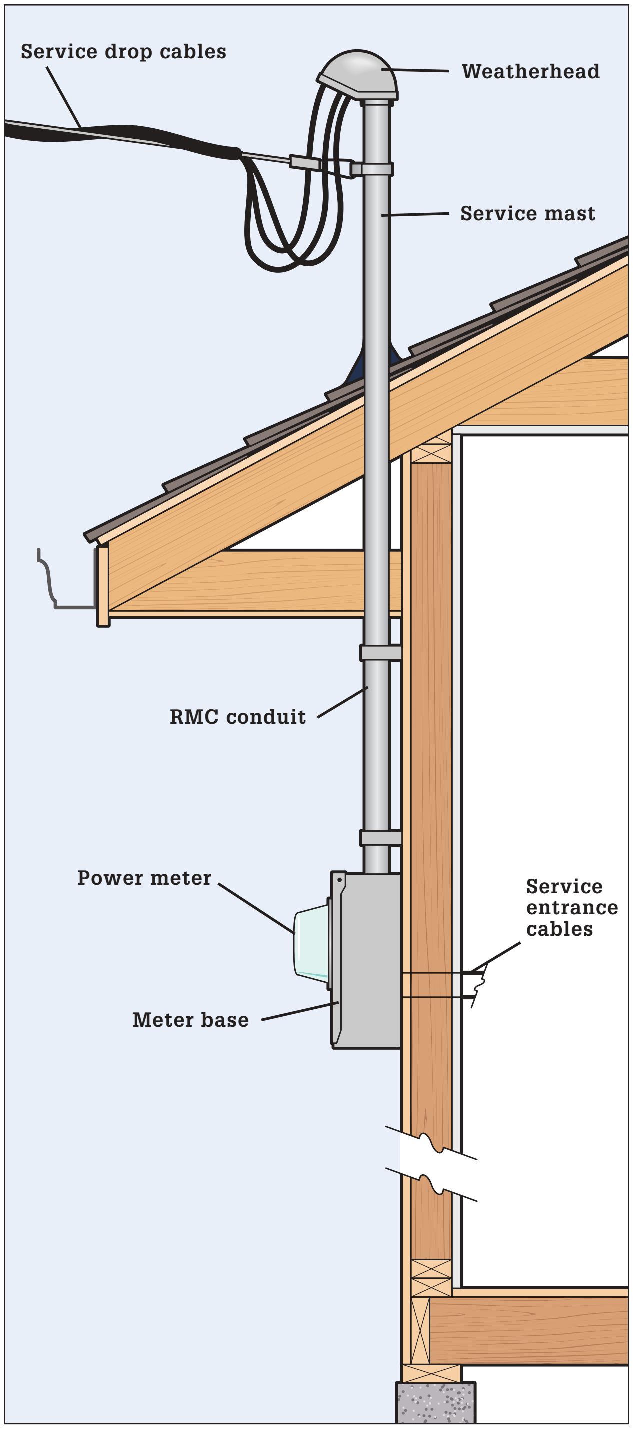
Service drop cables
Weatherhead
Service mast
RMC conduit
Power meter
Service entrance cables
Meter base

Aboveground service drop. In this common configuration, the service cables from the closest transformer (called the service drop) connect to service entrance wires near the weatherhead. This connection is called the service point and is where your property usually begins. The service entrance wires from the weatherhead are routed to a power meter that's owned by your utility company but is housed in a base that's considered your property. From the meter the service entrance wires enter your house through the wall and are routed to the main service panel, where they are connected to the main circuit breaker.

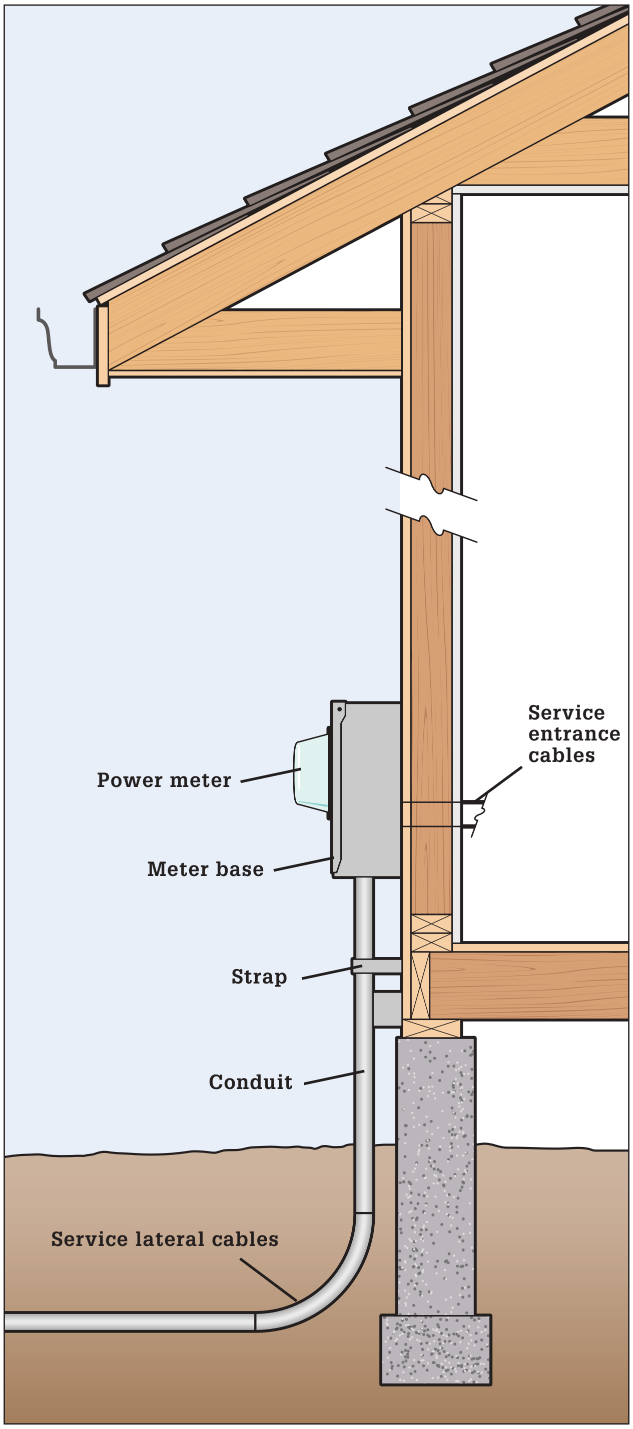
Service entrance cables
Power meter
Meter base
Strap
Conduit
Service lateral cables

Underground service lateral. Increasingly, homebuilders are choosing to have power supplied to their new homes underground instead of an overhead service drop. Running the cables in the ground eliminates problems with power outages caused by ice accumulation or fallen trees, but it entails a completely different set of cable and conduit requirements. For the homeowner, however, the differences are minimal, because the hookups are identical once the power service reaches the meter.

1st proofs for reference only

Edition
age: 50
41940 - Advanced Home Wiring_001-063.indd 51
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 9:06 AM
Page: 51

Locating Your New Panel ▸

Local codes dictate where the main service panel may be placed relative to other parts of your home. Although the codes vary (and always take precedence), national codes stipulate that a service panel (or any other distribution panel) may not be located near flammable materials, in a bathroom, clothes closet or other area designated for storage, above stairway steps, or directly above a workbench or other permanent work station or appliance. The panel also can't be located in a crawl space. If you are installing a new service entry hookup, there are many regulations regarding height of the service drop and the meter. Contact your local inspections office for specific regulations.

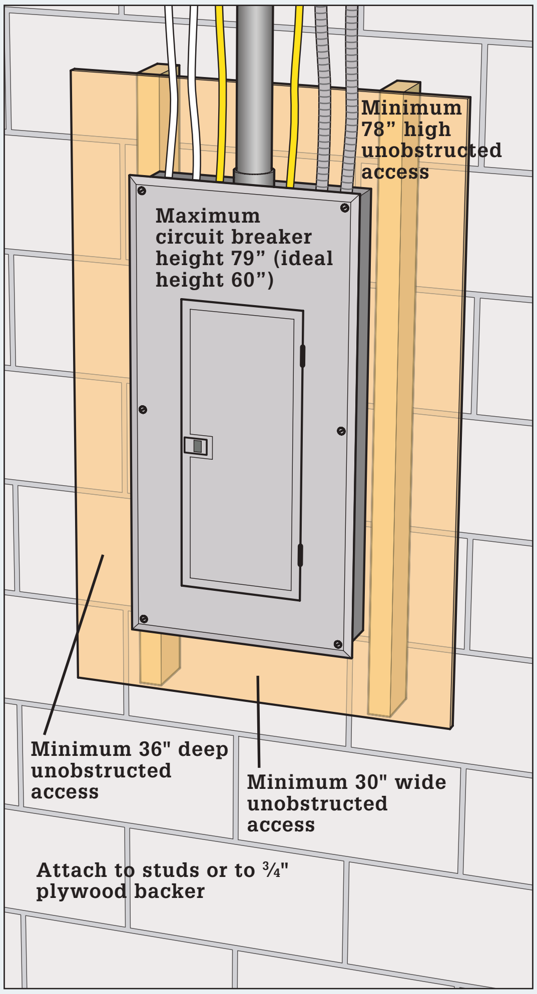
Minimum 78" high unobstructed access
Maximum circuit breaker height 79" (ideal height 60")
Minimum 36" deep unobstructed access
Minimum 30" wide unobstructed access
Attach to studs or to ¾" plywood backer

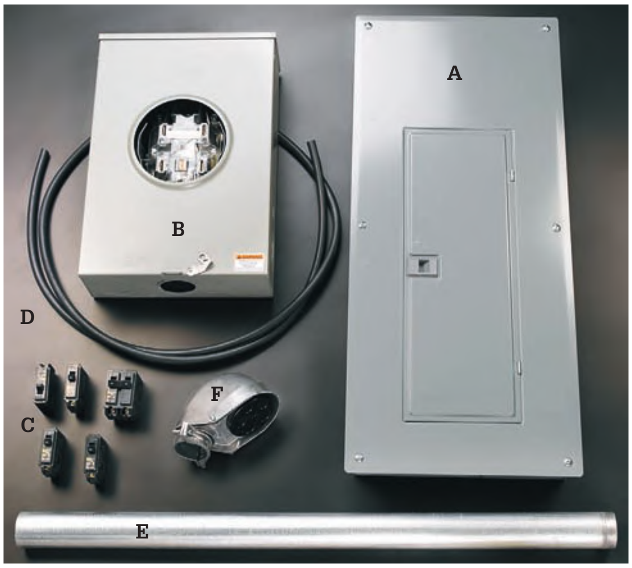

All the equipment you'll need to upgrade your main panel is sold at most larger building centers. It includes (A) a new 200-amp panel; (B) a 200-amp bypass meter base (also called a socket); (C) individual circuit breakers (if your new panel is the same brand as your old one you may be able to reuse the old breakers); (D) new, THW, THHW, THWN-2. RHW, RHW-2, XHHW (2/0 copper seen here); (E) 2" dia. rigid metallic conduit; (F) weatherhead shroud for mast.

Meter
Shutoff switch

The main circuit breaker (called the service equipment) may need to be located outside next to the electric meter if your main panel is too far away from the point where the service cable enters your house. The maximum distance allowed varies widely, from as little as 3 ft. to more than 10 ft. Wiring the service cable through the shutoff has the effect of transforming your main panel into a subpanel, which will impact how the neutral and ground wires are attached (see Subpanels, pages 64 to 65).

1st proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 52
10/10/14 9:06 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 52
Pro-Vis

How to Replace a Main Panel

Shut off power to the house at the transformer. This must be done by a technician who is certified by your utility company. Also have the utility worker remove the old meter from the base. It is against the law for a homeowner to break the seal on the meter.

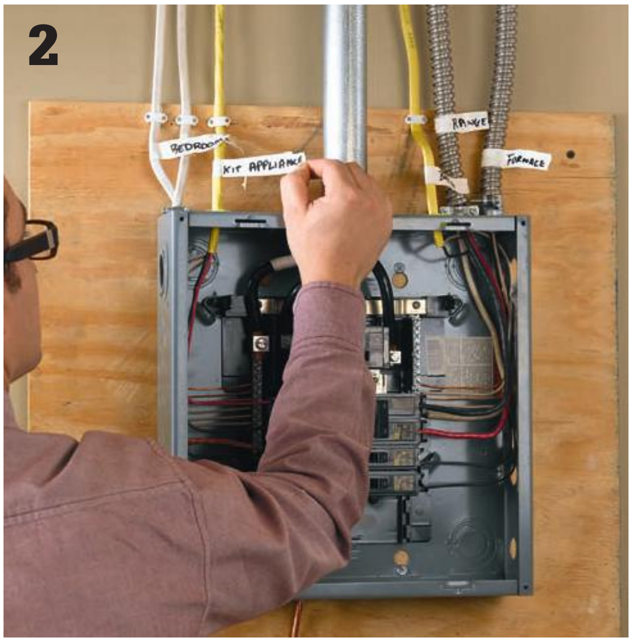

Label all incoming circuit wires before disconnecting them. Labels should be written clearly on tape that is attached to the cables outside of the existing service panel. Test the circuits before starting to make sure they are labeled correctly.

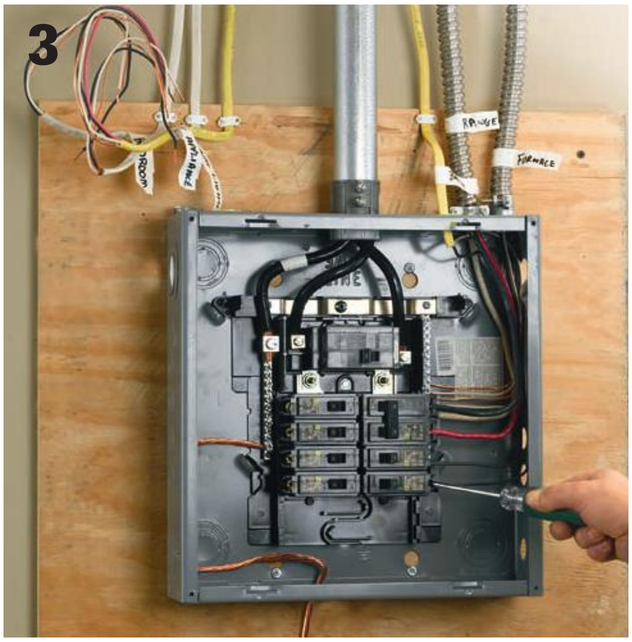

Disconnect incoming circuit wires from breakers, grounding bar, and neutral bus bar. Also disconnect cable clamps at the knockouts on the panel box. Retract all circuit wires from the service panel and coil it up neatly, with the labels clearly visible.

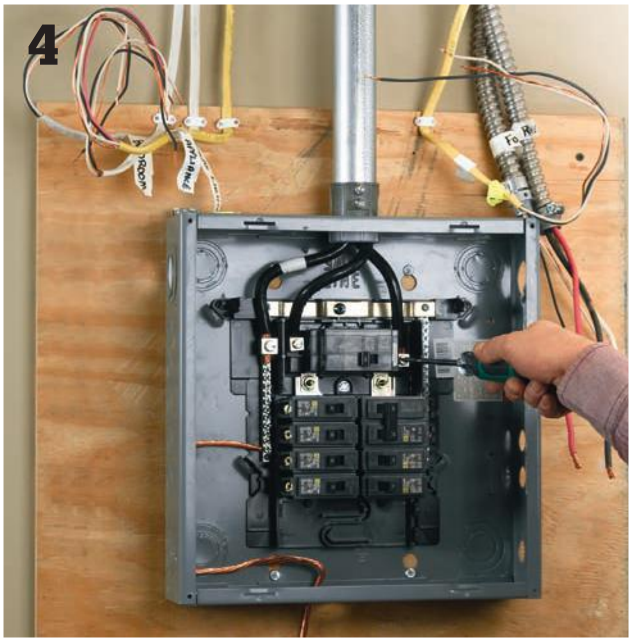

Unscrew the lugs securing the service entry cables at the top of the panel. For 240-volt service you will find two heavy-gauge SE cables, probably with black sheathing. Each cable carries 120 volts of electricity. A neutral service cable, usually of smaller gauge than the SE cables, will be attached to the neutral bus bar. This cable returns current to the source.

(continued)

1st
proofs for reference only

Edition
age: 52
41940 - Advanced Home Wiring_001-063.indd 53
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 9:06 AM
Page: 53

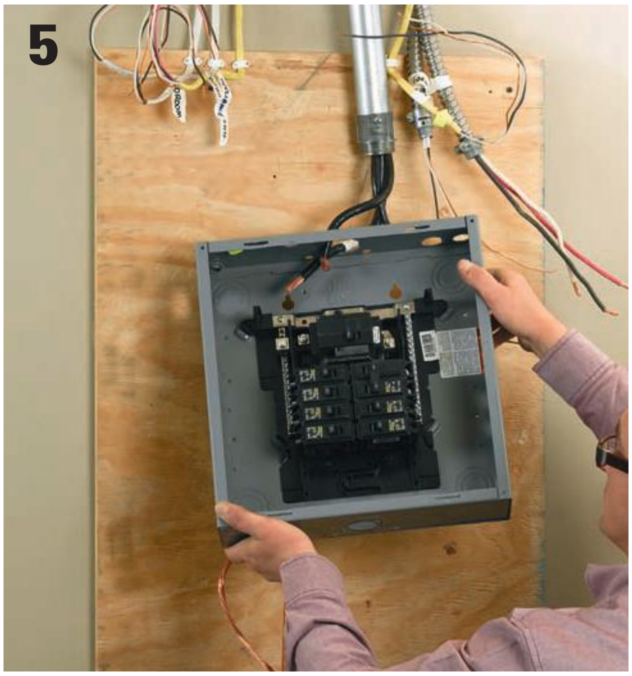

Remove the old service panel box. Boxes are rated for a maximum current capacity; and if you are upgrading, the components in the old box will be undersized for the new service levels. The new box will have a greater number of circuit slots as well.

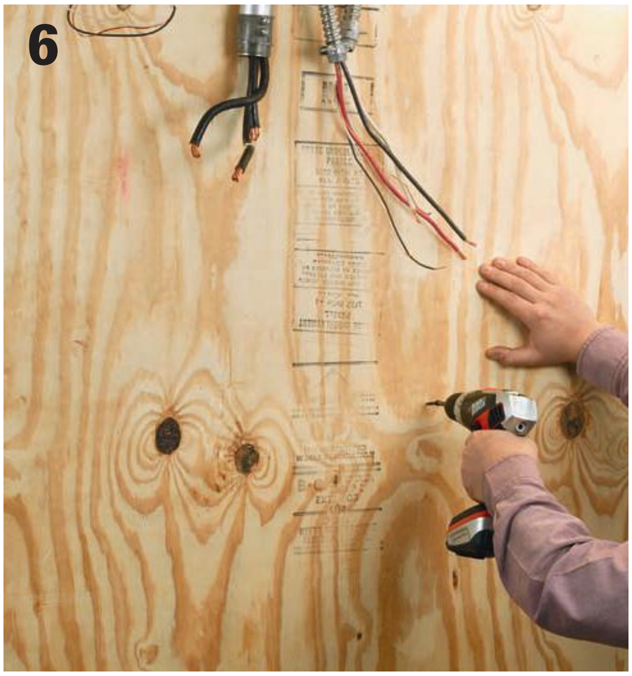

Replace the old panel backer board with a larger board in the installation area (see sidebar, page 52). A piece of ¾" plywood is typical. Make sure the board is well secured at wall framing members.

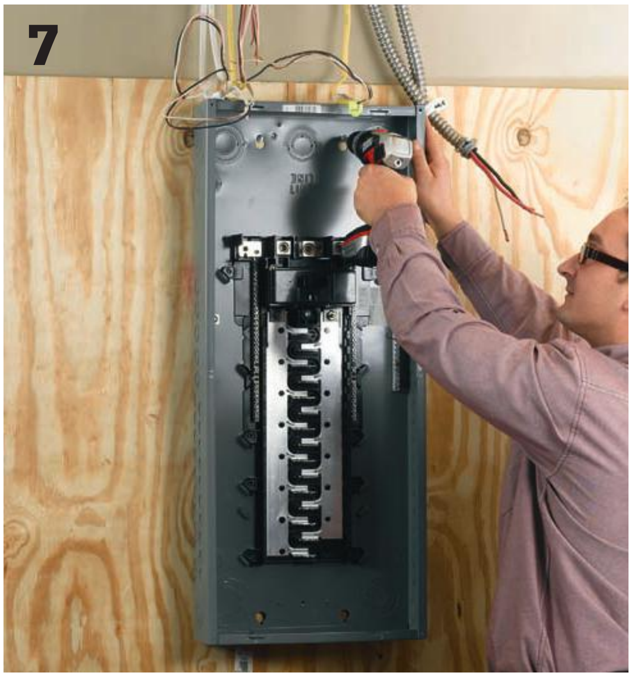

Attach the new service panel box to the backer board, making sure that at least two screws are driven through the backer and into wall studs. Drill clearance holes in the back of the box at stud locations if necessary. Use roundhead screws that do not have tapered shanks so the screwhead seats flat against the panel.

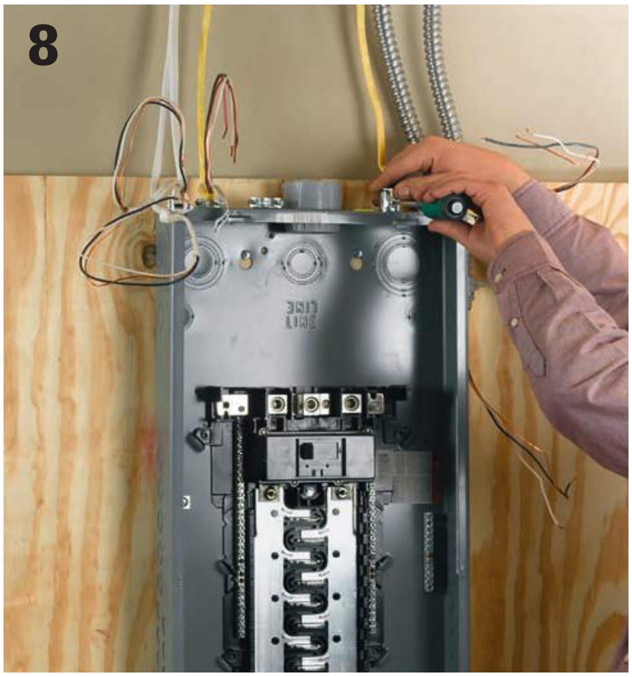

Attach properly sized cable clamps to the box at the knockout holes. Install one cable per knockout in this type of installation and plan carefully to avoid removing knockouts that you do not need to remove (if you do make a mistake, you can fill the knockout hole with a plug).

1st
proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 54
10/10/14 9:08 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 54

Splicing in the Box ▸

Some wiring codes allow you to make splices inside the panel box if the circuit wire is too short. Use the correct wire cap and wind electrical tape over the conductors where they enter the cap. If your municipality does not allow splices in the panel box, you'll have to rectify a short cable by splicing it in a junction box before it reaches the panel and then replacing the cable with a longer section for the end of the run. Make sure each circuit line has at least 12" of slack.

9

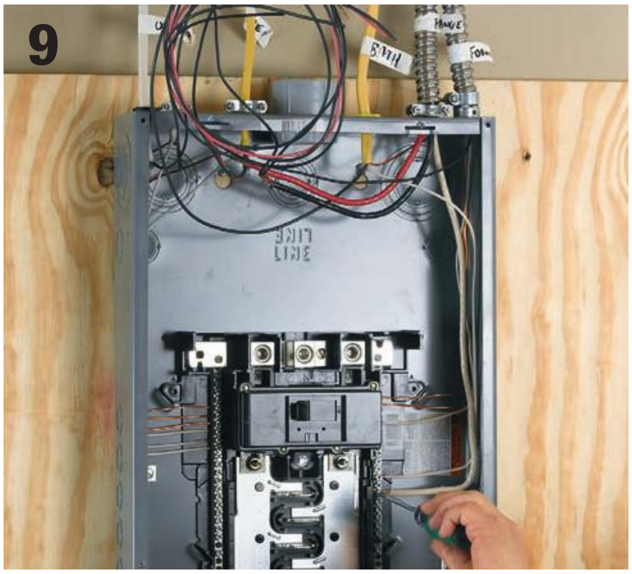

Attach the white neutral from each circuit cable to the neutral bus bar. Most panels have a preinstalled neutral bus bar, but in some cases you may need to purchase the bar separately and attach it to the panel back. The panel should also have a separate grounding bar that you also may need to purchase separately. Attach the grounds as well.

10

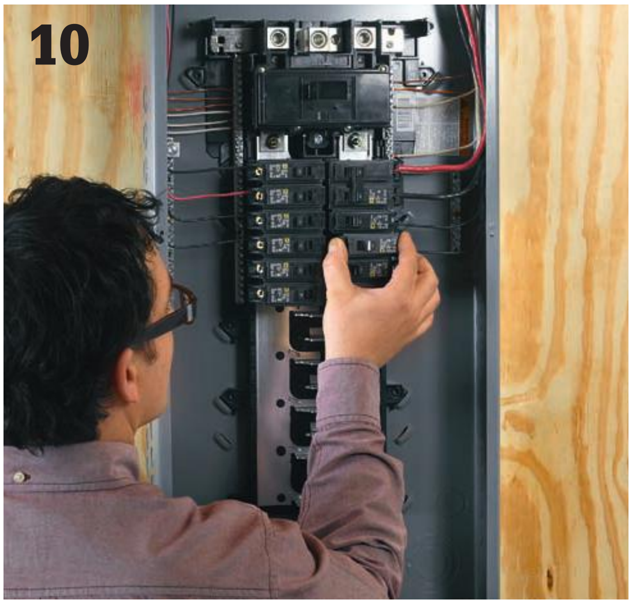

Attach the hot lead wire to the terminal on the circuit breaker, and then snap the breaker into an empty slot. When loading slots, start at the top of the panel and work your way downward. It is important that you balance the circuits as you go to equalize the amperage. For example, do not install all the 15-amp circuits on one side and all the 20-amp circuits on the other.

11

Create an accurate circuit index and affix it to the inside of the service panel door. List all loads that are on the circuit as well as the amperage. Once you have restored power to the new service panel (see step 18), test out each circuit to make sure you don't have any surprises. With the main breakers on, shut off all individual circuit breakers, and then flip each one on by itself. Walk through your house and test every switch and receptacle to confirm the loads on that circuit.

(continued)

1st proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 55
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 9:06 AM
Page: 55
Edition
age: 54

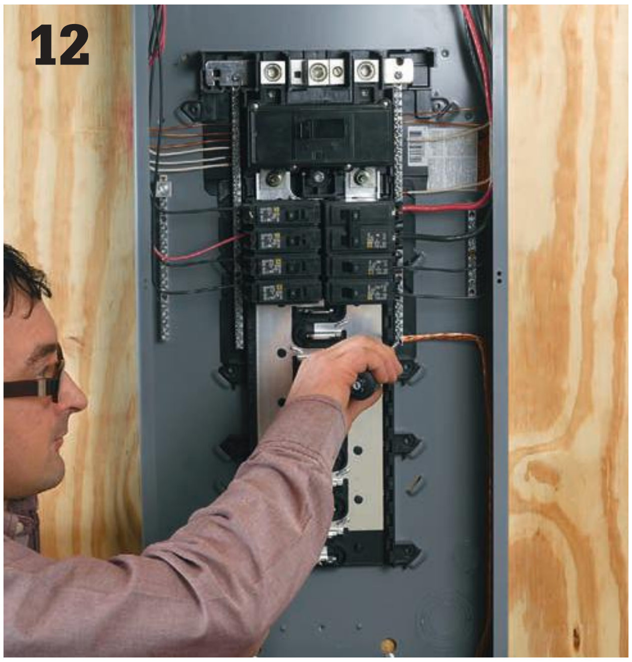

Install grounding conductors (see pages 58 to 59). Local codes are very specific about how the grounding and bonding needs to be accomplished. For example, some require multiple rods driven at least 6 ft. apart. Discuss your grounding requirements thoroughly with your inspector or an electrician before making your plan.

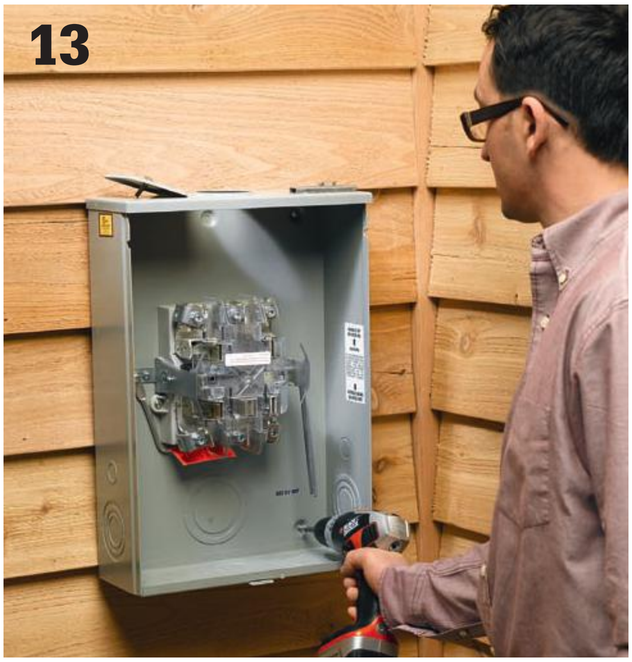

Replace the old meter base (have the utility company remove the meter when they shut off power to the house, step 1). Remove the old meter base, also called a socket, and install a new base that's rated for the amperage of your new power service. Here, a 200-amp bypass meter base is being installed.

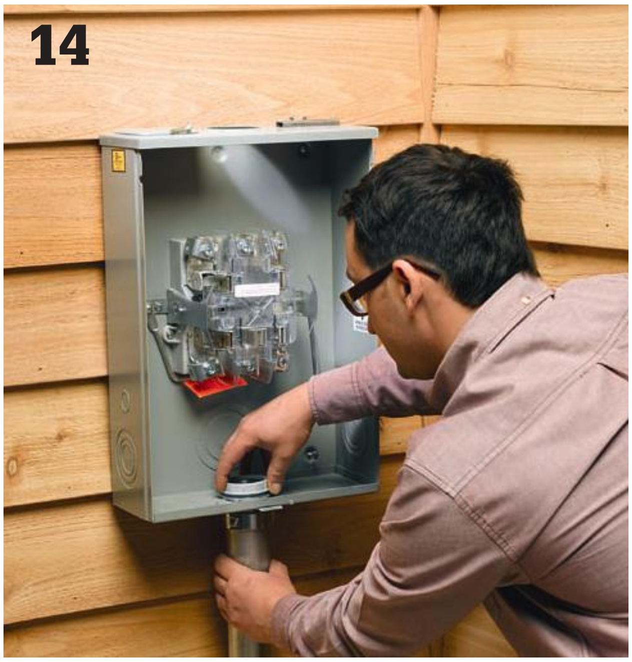

Update the conduit that runs from your house to the bottom of the meter base. This should be 2" rigid conduit in good repair. Attach the conduit to the base and wall with the correct fittings. Rigid metal conduit is a good option, but Schedule 80 PVC is probably the best choice for housing the service entrance wires.

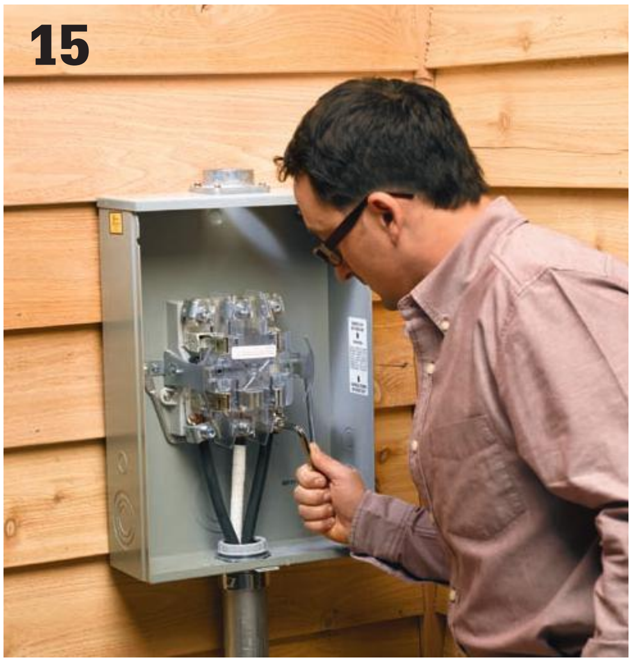

Install new service entrance wires. Each wire carries 120 volts from the meter to the service wire lugs at the top of your service panel. Code is very specific about how these connections are made. In most cases, you'll need to tighten the terminal nuts with a specific amount of torque that requires a torque wrench to measure. Also attach the sheathed neutral wire to the neutral/grounding lug.

1st proofs for reference only

41940 - Advanced Home Wiring_001-063.indd 56
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 9:06 AM
Page: 56
Pro-Vis

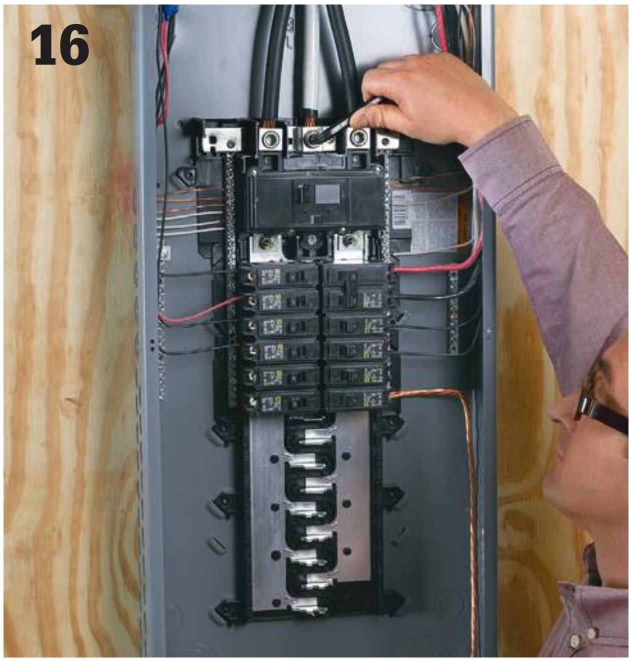

Attach the SE wires to the lugs connected to the main breakers at the top of your service entry panel. Do not remove too much insulation on the wires—leaving the wires exposed is a safety hazard. The neutral service entry wire is attached either directly to the neutral bus bar or to a metal bridge that is connected to the neutral bonding bus bar. Install the green grounding screw provided with the panel.

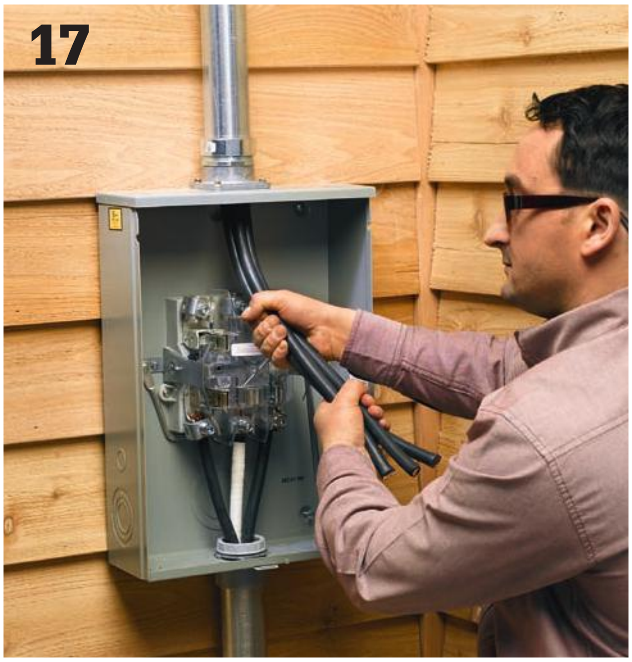

Install service entrance wires from the meter to the weatherhead, where the connections to the service drop wires are made. Only an agent for your public utility company may make the hookup at the weatherhead.

Tall Mast, Short Roof ▸

The service drop must occur at least 10 ft. above ground level, and as much as 14 ft. in some cases. Occasionally, this means that you must run the conduit for the service mast up through the eave of your roof and seal the roof penetration with a boot.

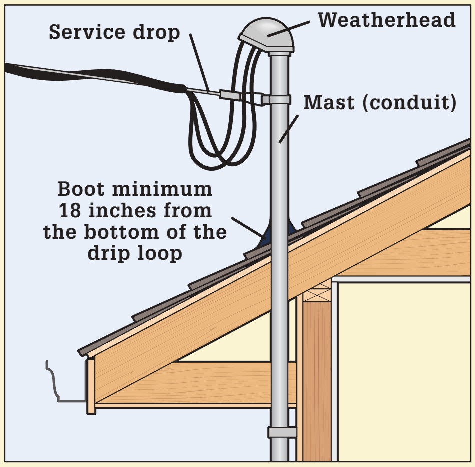

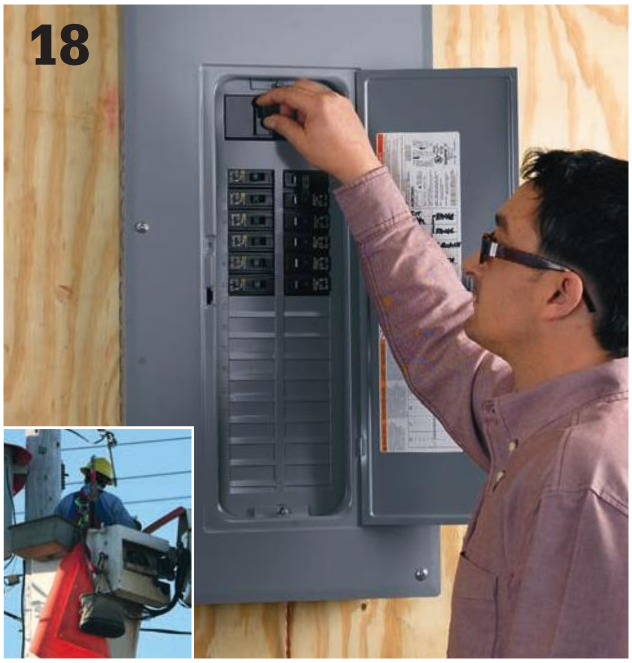

Have the panel and all connections inspected and approved by your local building department, and then contact the public utility company to make the connections at the power drop. Once the connections are made, turn the main breakers on and test all circuits.

1st proofs for reference only

Grounding & Bonding a Wiring System

All home electrical systems must be bonded and grounded according to code standards. This entails two tasks: the metal water and gas pipes must be connected electrically to create a continuous low resistance path back to the main electrical panel; and the main electrical panel must be grounded to a grounding electrode such as a ground rod or rods driven into the earth near the foundation of your house. Although the piping system is bonded to the ground through your main electrical service panel, the panel grounding and the piping bonding are unrelated when it comes to function. The grounding wire that runs from your electrical panel to grounding electrode helps even out voltage increases that often occur because of lightning and other causes. The wires that bond your metal piping are preventative, and they only become important in the unlikely event that an electrical conductor energizes the pipe. In that case, correct bonding of the piping system will ensure that the current does not remain in the system, where it could electrocute anyone who touches a part of the system, such as a faucet handle. Bonding is done relatively efficiently at the water heater, as the gas piping and water piping generally there.

Gas pipe in older homes is usually steel or copper. The bonding connection point for these pipes can be at any accessible location, such as at the water heater or at the gas meter. Gas pipe in some new homes is a flexible material called corrugated stainless steel tubing (CSST). The bonding point for CSST must be at the first piece of steel or copper pipe where the gas service enters the home. This is because lightning can blow holes in CSST, causing a gas leak.

Tools & Materials ▸

Hammer
Straight edge screwdriver
Drill
½" drill bit
A length of ground wire
Some wire staples
3 pipe ground clamps
Eye and ear protection
Work gloves
Grounding rods
5-lb. maul
Caulk

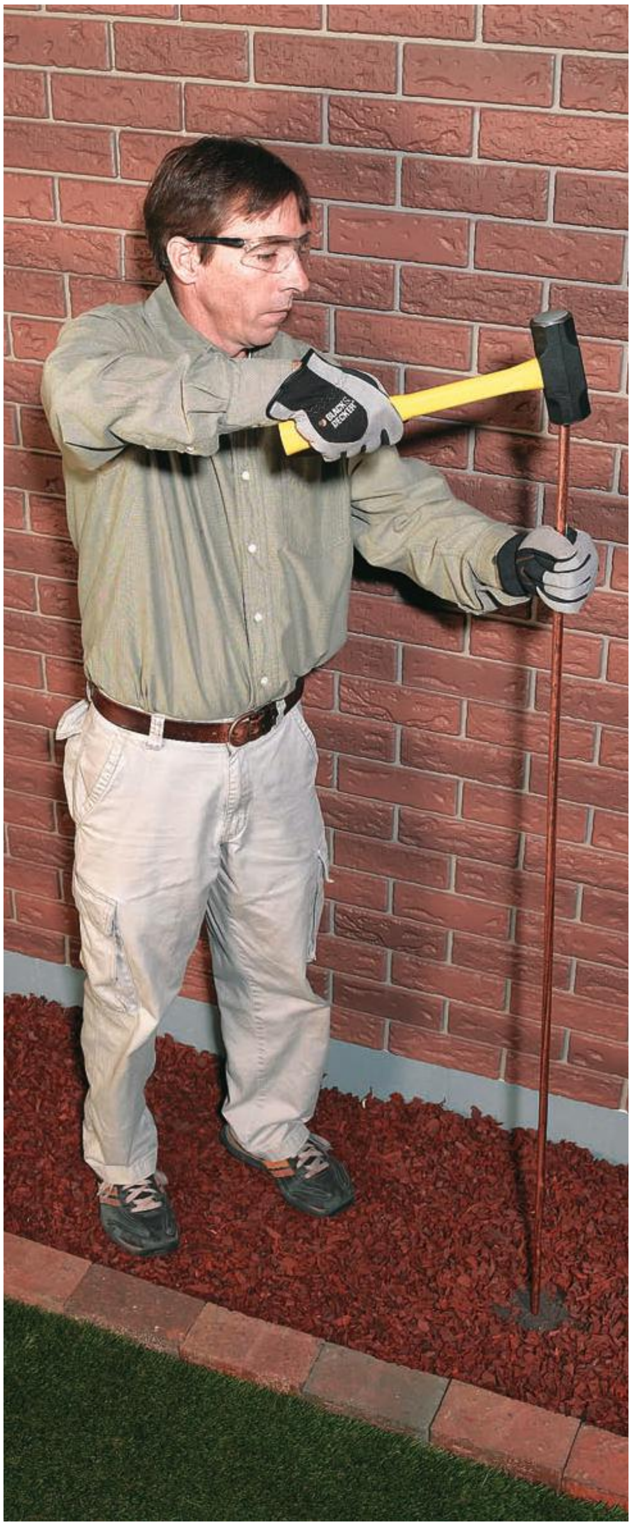

A pair of 8-ft.-long metal ground rods are driven into the earth next to your house to provide a path to ground for your home wiring system.

1st proofs for reference only

41940 - Advanced Home Wiring_001-063.indd 58
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/30/14 10:06 AM
Page: 58
Pro-Vis

How to Bond Metallic Piping

1

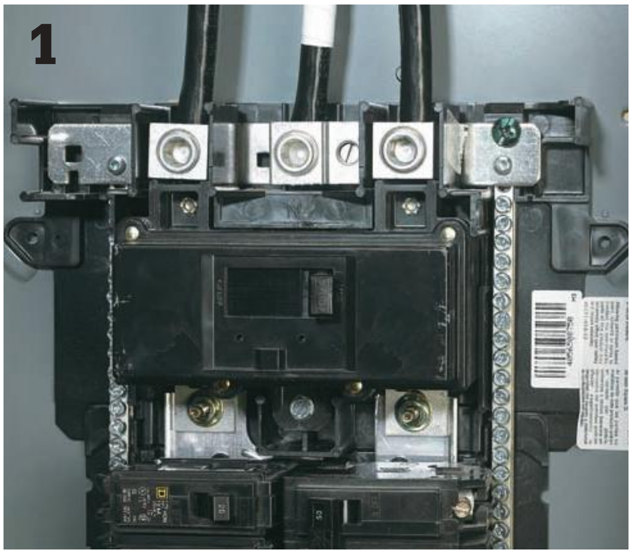

Determine the amperage rating of your electrical service by looking at your main breakers. The system amperage (usually 100 or 200 amps) determines the required gauge of the bonding wire you need. #4 copper wire is sufficient for service not exceeding 200 amps. Smaller, less expensive copper wire is allowed for services between 100 and 175 amps. Check with your electrical inspector if you want to use wire smaller than #4.

2

Run the bonding wire from a point near your water heater (a convenient spot if you have a gas-fueled water heater) to an exit point where the wire can be bonded to the grounding wire that leads to the exterior grounding electrodes. This is frequently done at the service panel. Run this wire as you would any other cable, leaving approximately 6 to 8 ft. of wire at the water heater. If you are running this wire through the ceiling joists, drill a ½" hole as close to the center as possible to not weaken the joist. Staple the wire every 2 ft. if running it parallel to the joists.

3

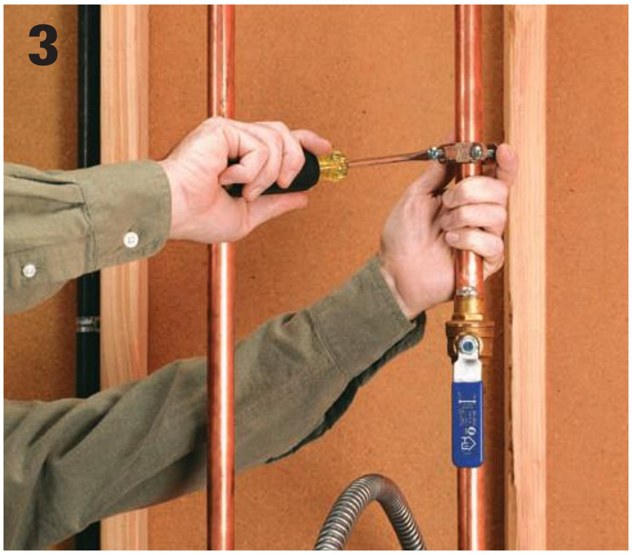

Install pipe ground clamps on each pipe (hot water supply, cold water supply, gas), roughly a foot above the water heater. Do not install clamps near a union or elbow because the tightening of the clamps could break or weaken soldered joints. Also make sure the pipes are free and clear of any paint, rust, or any other contaminant that may inhibit a good clean connection. Do not overtighten the clamps. Use clamps that are compatible with the pipe so that corrosion will not occur. Use copper or brass clamps on copper pipe. Use brass or steel clamps on steel pipe.

4

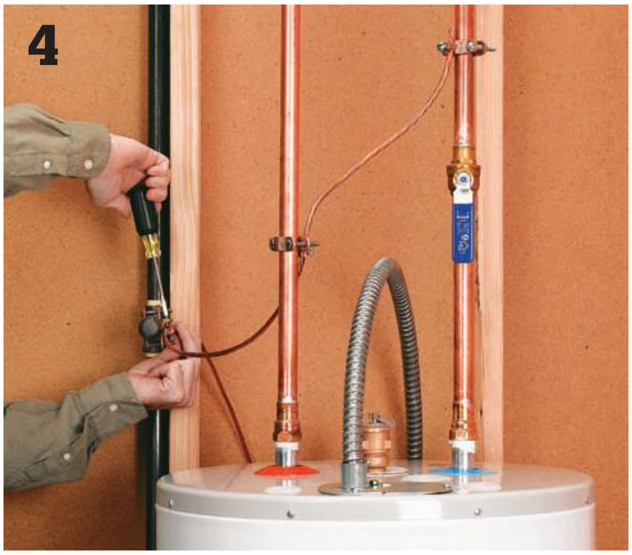

Route the ground wire through each clamp wire hole and then tighten the clamps onto the wire. Do not cut or splice the wire: The same wire should run through all clamps.

(continued)

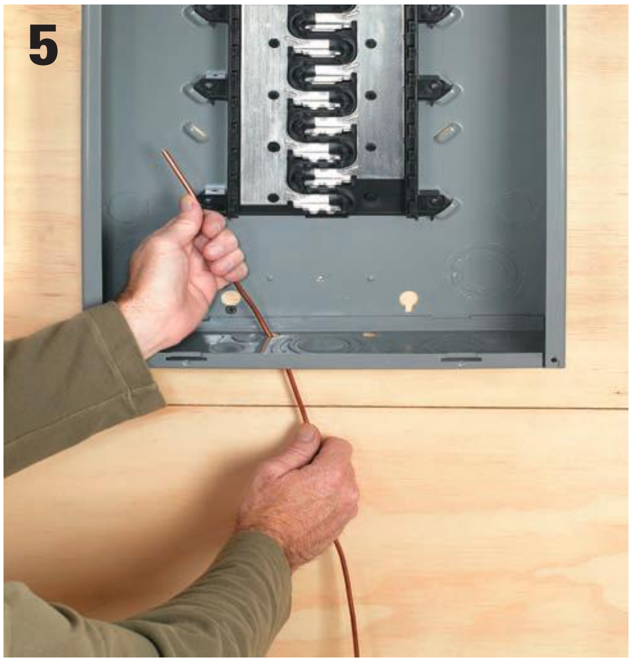

At the panel, turn off the main breaker. Open the cover by removing the screws, and set the cover aside. Route the ground wire through a small ⅜" hole provided towards the rear of the panel on the top or bottom. You will usually have to knock the plug out of this hole by placing a screwdriver on it from the outside and tapping with a hammer. Make sure the ground wire will not come into contact with the bus bars in the middle of the panel or any of the load terminals on the breakers.

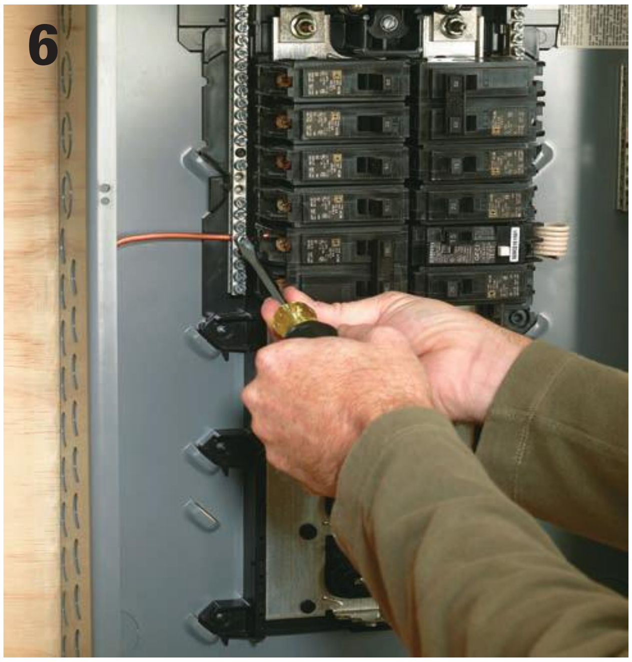

Locate an open hole on your ground and neutral bus and insert the ground wire. These holes are large enough to accommodate up to a #4 awg wire, but it may be difficult at times. If you're having trouble pushing the wire in, trim a little wire off the end and try with a clean cut piece. Secure the set screw at the lug. Replace the panel cover and turn the main breaker back on.

Tips for Grounding the Main Service Panel

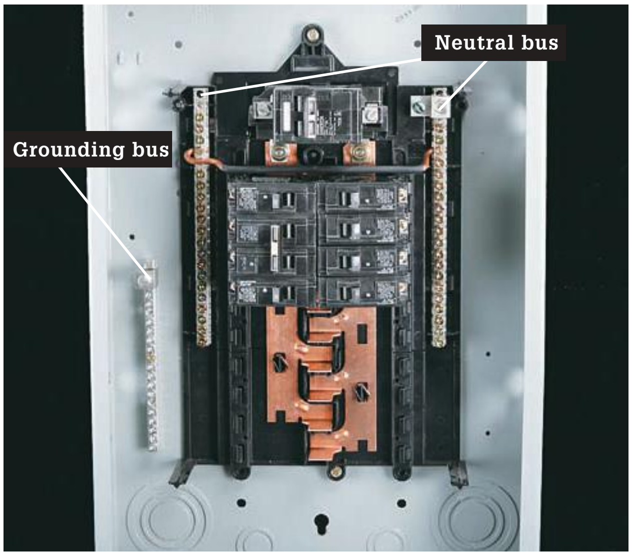

The neutral and grounding wires should not be connected to the same bus in most subpanels. The grounding bus should be bonded to the subpanel cabinet. The neutral bus should not be bonded to the subpanel cabinet.

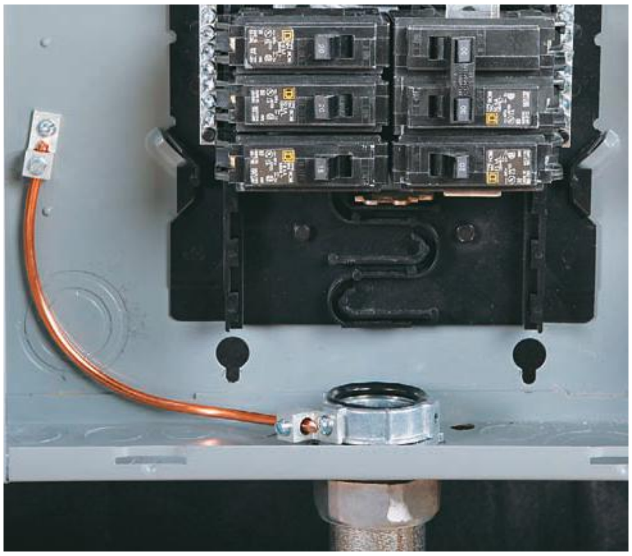

Metallic conduit must be physically and electrically connected to panel cabinets. A bonding bushing may be required in some cases, where not all of a knockout is removed.

1st
proofs for reference only
41940 - Advanced Home Wiring_001-063.indd 60
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 9:08 AM
Page: 60
Pro-Vis

Ground Rod Installation

The ground rod is an essential part of the grounding system. Its primary function is to create a path to ground for electrical current, such as lightning, line surges, and unintentional contact with high voltage lines. If you upgrade your electrical service you likely will need to upgrade your grounding wire and rods to meet code.

Note: Different municipalities have different requirements for grounding, so be sure to check with the AHJ (Authority Having Jurisdiction) first before attempting to do this yourself.

Call before you dig! Make sure the area where you will be installing the ground rods is free and clear from any underground utilities.

Exercise Your Breakers ▸

Your breakers (including the main) should be "exercised" once a year to ensure proper mechanical function. Simply turn them off and then back on. A convenient time to perform the exercise is at daylight savings time, when you'll need to reset all of your clocks anyway.

How to Install a Grounding Electrode System

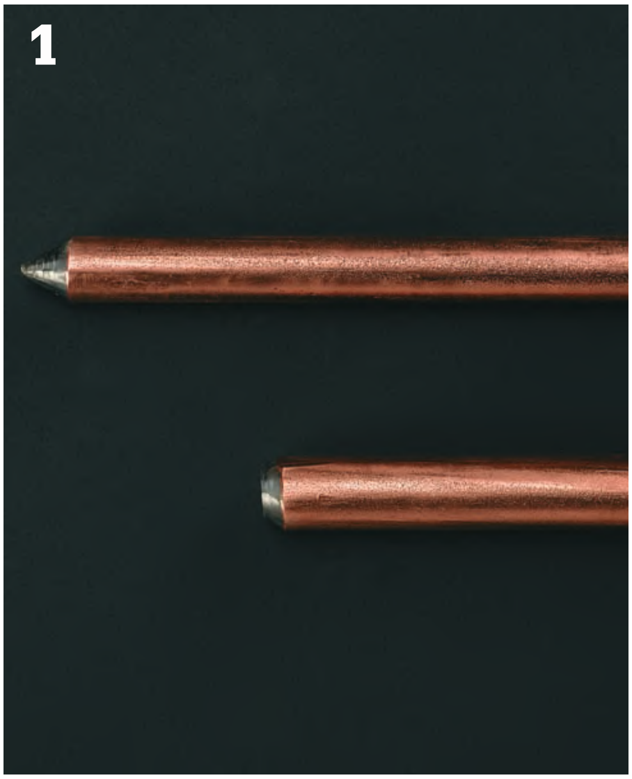

Begin by purchasing two copper-coated steel ground rods 5⁄8" diameter by 8' long. Grounding rods have a driving point on one end and a striking face on the other end.

2
BLACK & DECKER

Drill a 5⁄16" hole in the rim joist of your house, as close as practical to the main service panel to the outside of the house above the ground level at least 6".

(continued)

1st proofs for reference only
Edition
age: 60
41940 - Advanced Home Wiring_001-063.indd 61
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 9:06 AM
Page: 61

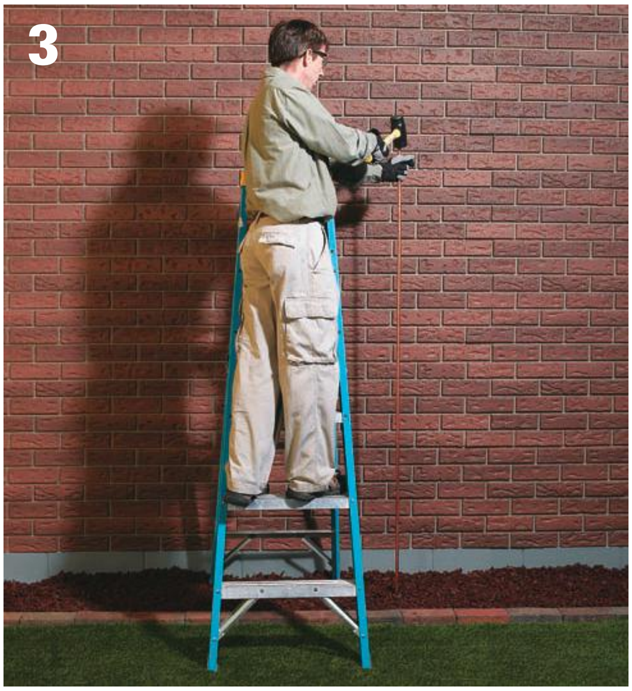

About a foot from the foundation of the house, pound one ground rod into the earth with a five-pound maul. If you encounter a rock or other obstruction, you can pound the ground rod at an angle as long as it does not exceed 45°. Drive until only 3" or 4" of the rod is above ground. Measure at least 6 ft. from the first ground rod and pound in another one.

Run uninsulated #4 copper wire from the ground bus in your main service panel through the hole in the rim joist and to the exterior of the house, leaving enough wire to connect the two ground rods together.

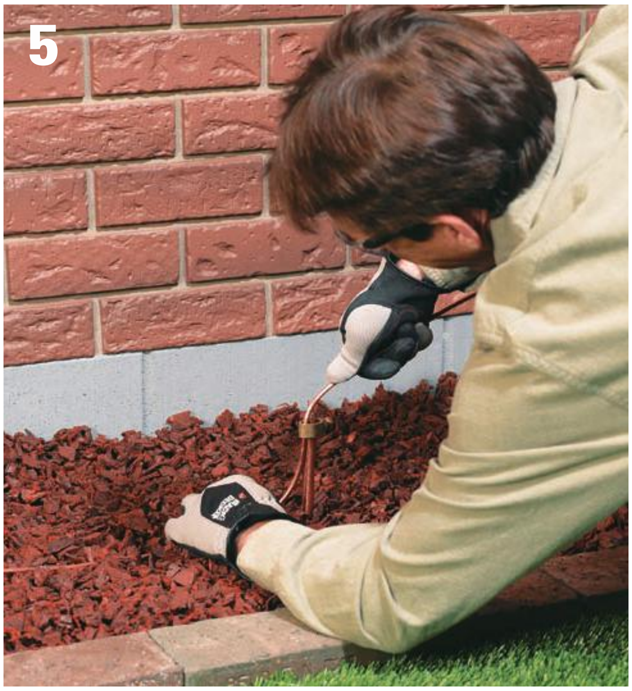

Using a brass clamp commonly referred to as an acorn, connect the wire to the first ground rod, pulling the wire taut so no slack exists. Continue pulling the wire to reach the second grounding rod, creating a continuous connection.

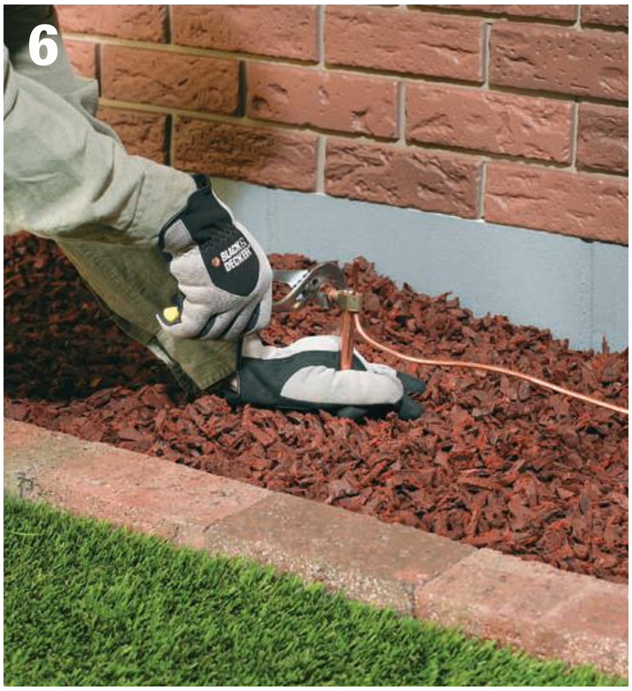

Connect the second ground rod with another acorn to the uncut grounding wire previously pulled through the first acorn. Trim the excess wire.

1st
proofs for reference only

41940 - Advanced Home Wiring_001-063.indd 62
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 9:06 AM
Page: 62
Pro-Vis

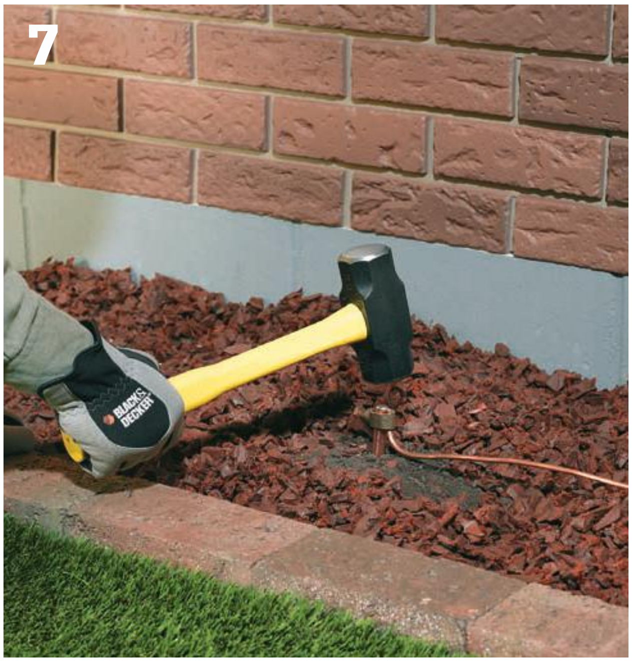

Dig out a few inches around each rod to create clearance for the five-pound maul. Creating a shallow trench beneath the grounding wire between the rods is also a good idea. Drive each rod with the maul until the top of the rod is a few inches below grade.

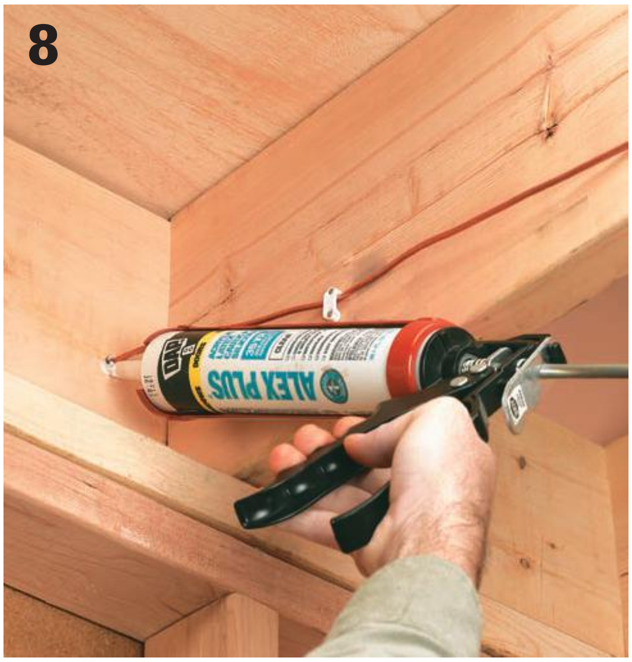

Inject caulk into the hole in the rim joist on both the interior and exterior side.

Tips for Grounding

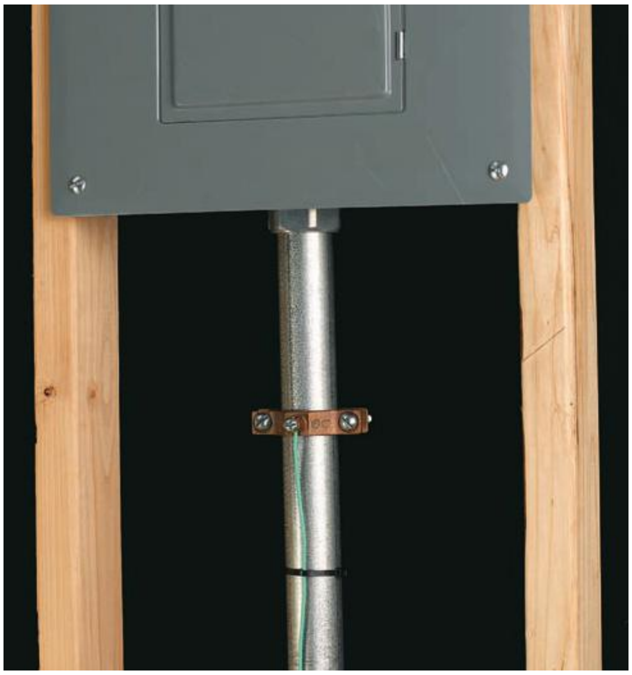

A listed metal strap may be used to ground indoor communication wires such as telephone and cable TV if an intersystem bonding terminal is not available.

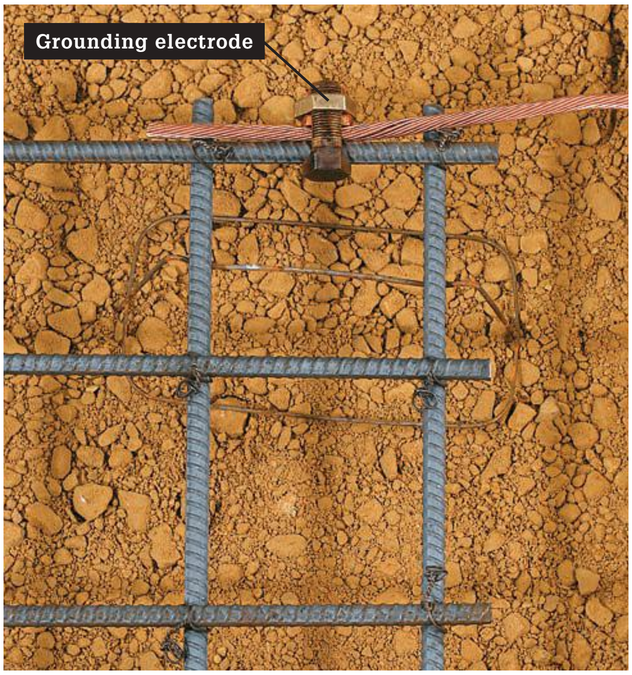

A piece of reinforcing bar encased in a concrete footing is a common grounding electrode in new construction. Called an ufer, the electrode must be No. 4 or larger rebar and at least 20 ft. long. (Shown prior to pouring concrete.)

1st
proofs for reference only
Edition
age: 62
41940 - Advanced Home Wiring_001-063.indd 63
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 9:06 AM
Page: 63

Installing a Subpanel

Install circuit breaker subpanels if the main circuit breaker panel does not have enough open breaker slots for the new circuits you are planning. Subpanels serve as a second distribution center for connecting circuits. They receive power from a double-pole circuit breaker you install in the main circuit breaker panel.

If the main service panel is so full that there is no room for the double-pole subpanel breaker, you can reconnect some of the existing 120-volt circuits to special slimline breakers (photos below).

Plan your subpanel installation carefully, making sure your electrical service supplies enough power to support the extra load of the new subpanel circuits. Assuming your main service is adequate, consider installing an oversized subpanel breaker in the main panel to provide enough extra amps to meet the needs of future wiring projects.

Also consider the physical size of the subpanel, and choose one that has enough extra slots to hold circuits you may want to install later. The smallest panels have room for up to six single-pole breakers (or three double-pole breakers), while the largest models can hold 20 single-pole breakers or more.

Subpanels often are mounted near the main circuit breaker panel. Or, for convenience, they can be installed close to the areas they serve, such as in a new room addition or a garage. In a finished room, a subpanel can be painted or housed in a decorative cabinet so it is less of a visual distraction. If it is covered, make sure the subpanel is easily accessible and clearly identified.

Tools & Materials ▸

Hammer
Screwdriver
Circuit tester
Cable ripper
Combination tool
Screws
Cable clamps
Three-wire NM cable
Cable staples
Double-pole circuit breaker
Circuit breaker subpanel
Slimline circuit breakers

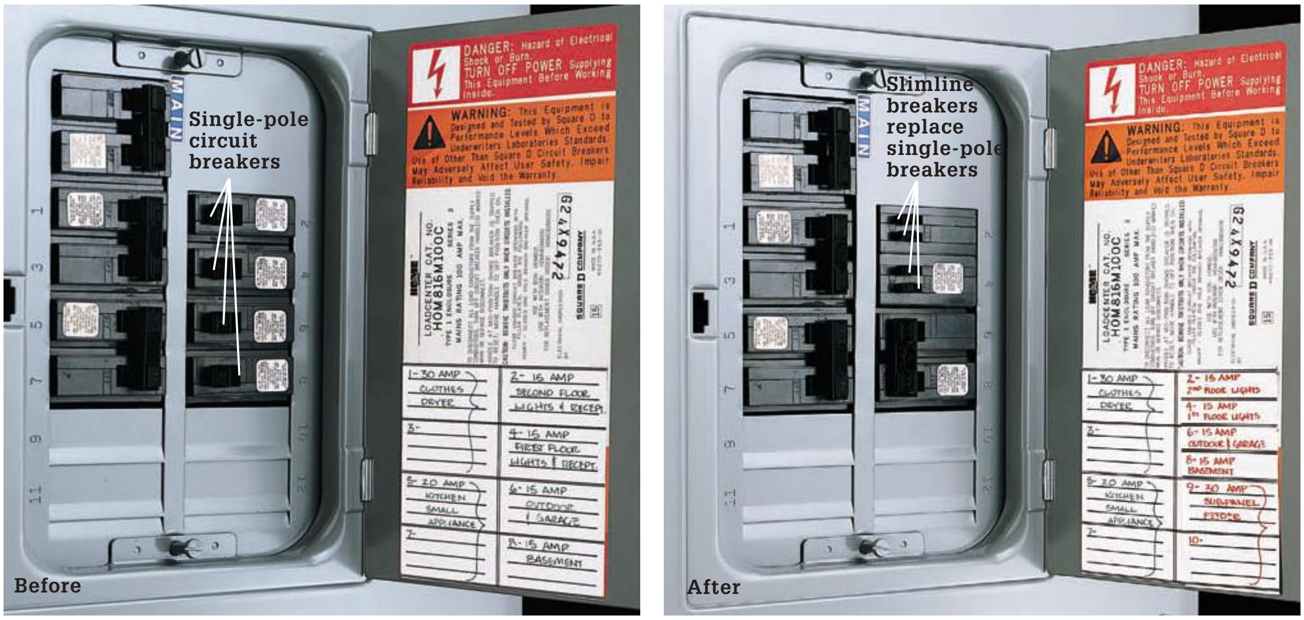
Single-pole circuit breakers
Before
Slimline breakers replace single-pole breakers
After

To conserve space in a service panel, you can replace single-pole breakers with slimline breakers. Slimline breakers take up half the space of standard breakers, allowing you to fit two circuits into one single slot on the service panel. In the service panel shown above, four single-pole 120-volt breakers were replaced with slimline breakers to provide the double opening needed for a 30-amp, 240-volt subpanel feeder breaker. Use slimline breakers (if your municipality allows them) with the same amp rating as the standard single-pole breakers you are removing, and make sure they are approved for use in your panel. If your municipality and panel allow slimline breakers, there may be restrictions on the quantity and location where they may be installed on the panel.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 64
Pro-Vision C M Y K Text
10/10/14 12:05 AM
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 64

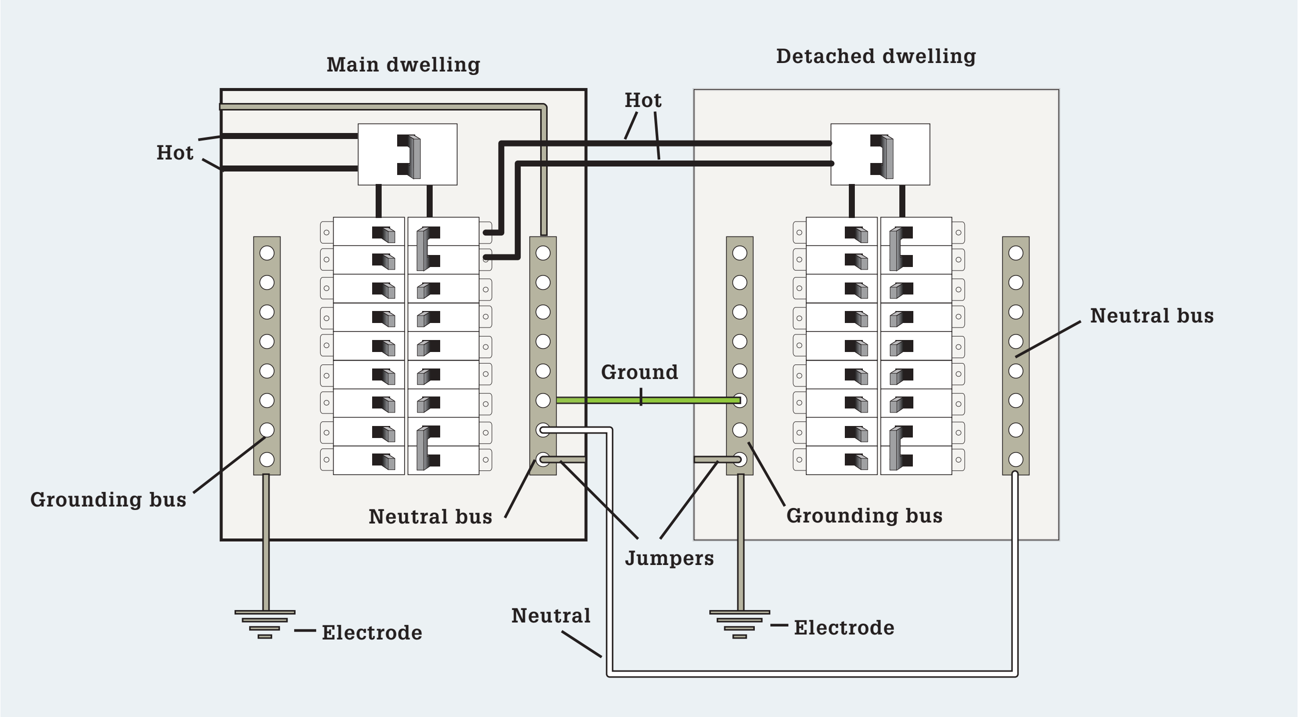
Main dwelling
Detached dwelling
Hot
Hot
Neutral bus
Ground
Grounding bus
Neutral bus
Grounding bus
Jumpers
Electrode
Neutral
Electrode

Wiring diagram for wiring a feeder from the main service panel to a subpanel in a separate building.

How to Install a Subpanel

1

2

3

Subpanels are subject to the same installation and clearance rules as service panels. The subpanel can be mounted to the sides of studs or to plywood attached between two studs. The panel shown here extends ½" past the face of studs so it will be flush with the finished wall surface.

Open a knockout in the subpanel using a screwdriver and hammer. Run the feeder cable from the main circuit breaker panel to the subpanel, leaving about 2 ft. of excess cable at each end.

Attach a cable clamp to the knockout in the subpanel. Insert the cable into the subpanel, and then anchor it to framing members within 8" of each panel and every 54" thereafter.

(continued)

1st
proofs for reference only

Edition
age: 64

41940 - Advanced Home Wiring_064-128.indd 65
Pro-Vision C M Y K Text

Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 1:42 PM
Page: 65

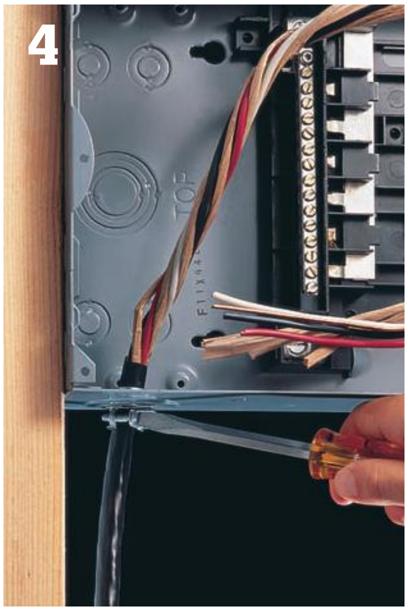

Strip away outer sheathing from the feeder cable using a cable ripper. Leave at least ¼" of sheathing extending into the subpanel. Tighten the cable clamp screws so the cable is held securely, but not so tightly that the wire sheathing is crushed.

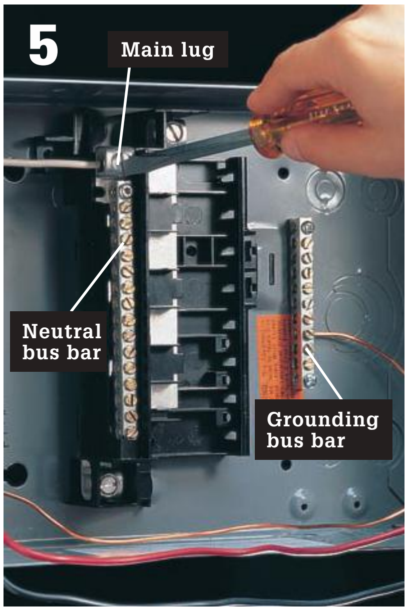

Strip ½" of insulation from the white neutral feeder wire, and attach it to the main lug on the subpanel neutral bus bar. Connect the grounding wire to a setscrew terminal on the grounding bus bar. Fold excess wire around the inside edge of the subpanel.

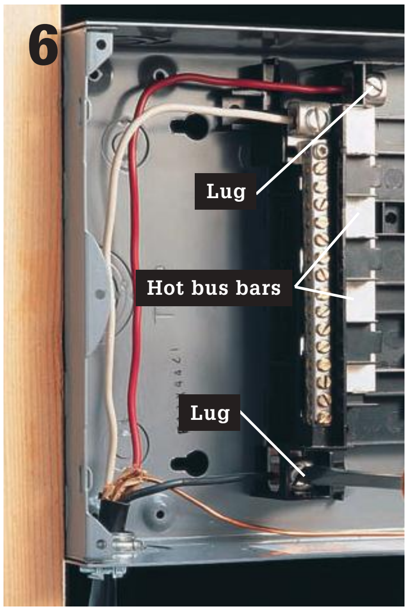

Strip away ½" of insulation from the red and the black feeder wires. Attach one wire to the main lug on each of the hot bus bars. Fold excess wire around the inside edge of the subpanel.

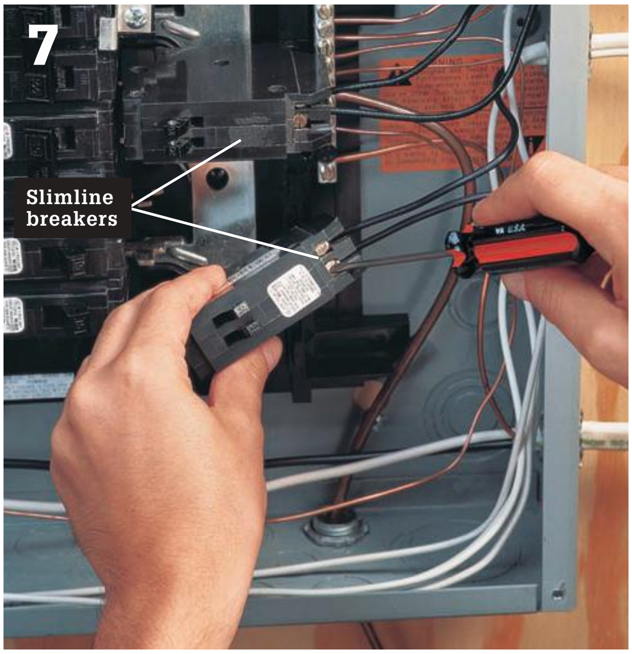

At the main circuit breaker panel, shut off the main circuit breaker, and then remove the coverplate and test for power. If necessary, make room for the double-pole feeder breaker by removing single-pole breakers and reconnecting the wires to slimline circuit breakers. Open a knockout for the feeder cable using a hammer and screwdriver. Note: some panels do not allow slimline breakers and some restrict where slimline breakers can be installed. Read the instructions on the panel cover.

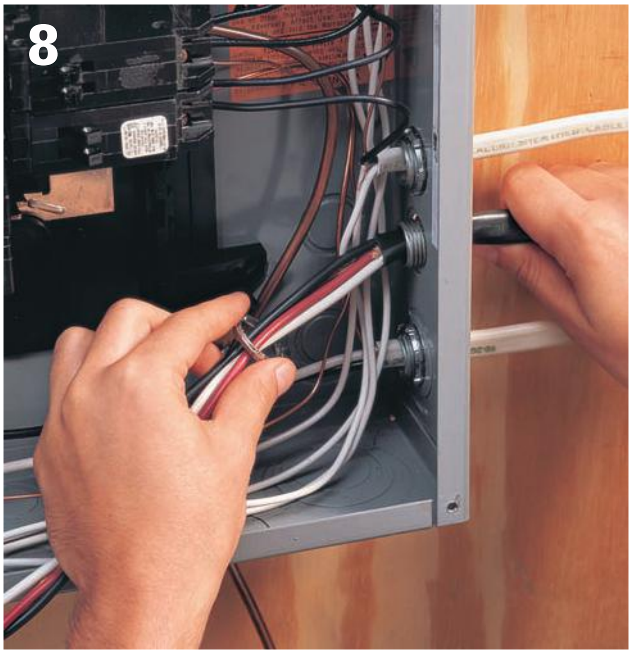

Strip away the outer sheathing from the feeder cable so that at least ¼" of sheathing will reach into the main service panel. Attach a cable clamp to the cable, and then insert the cable into the knockout, and anchor it by threading a locknut onto the clamp. Tighten the locknut by driving a screwdriver against the lugs. Tighten the clamp screws so the cable is held securely, but not so tightly that the cable sheathing is crushed.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 66
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 66

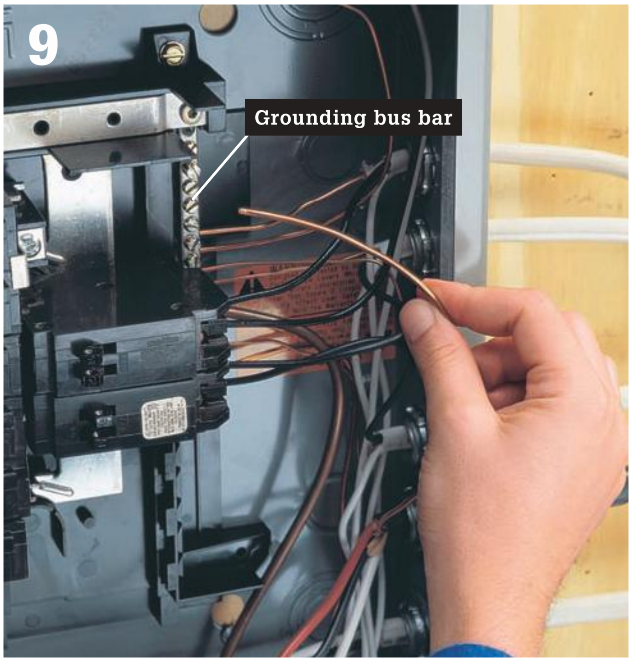

Bend the bare copper wire from the feeder cable around the inside edge of the main circuit breaker panel, and connect it to one of the setscrew terminals on the grounding bus bar.

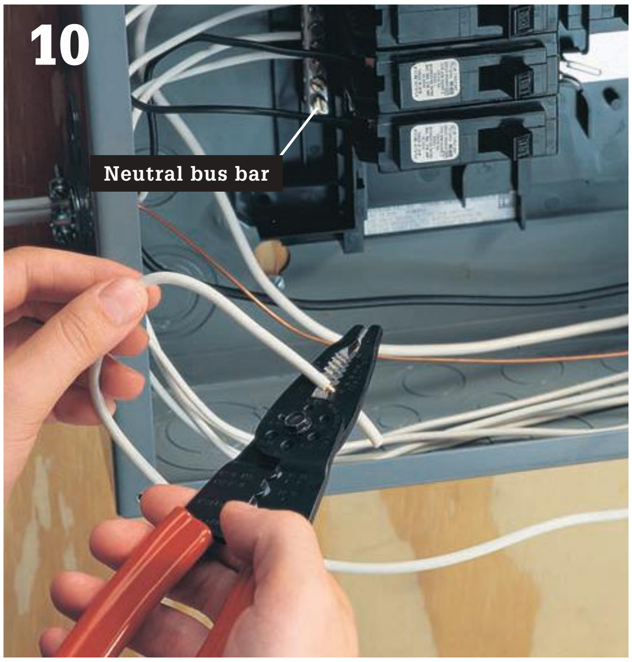

Strip away ½" of insulation from the white feeder wire. Attach the wire to one of the setscrew terminals on the neutral bus bar. Fold excess wire around the inside edge of the service panel.

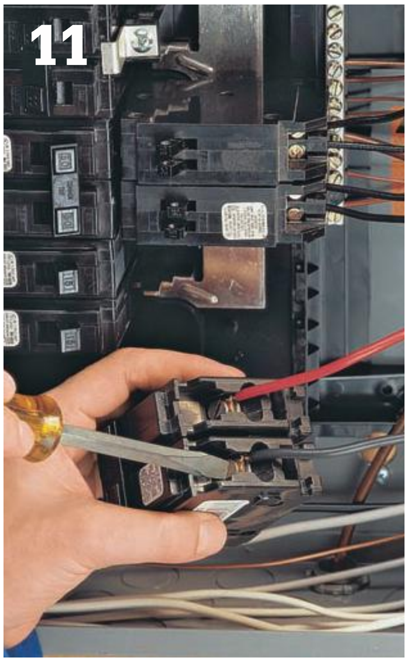

Strip ½" of insulation from the red and the black feeder wires. Attach one wire to each of the setscrew terminals on the double-pole feeder breaker. *Note: If your subpanel arrived with a preinstalled grounding screw in the panel back, remove and discard it.*

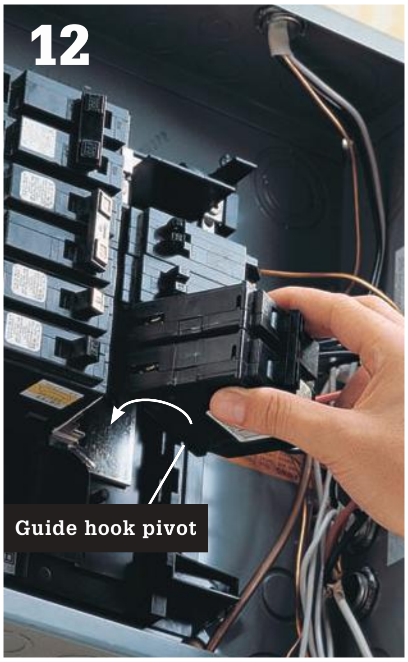

Hook the end of the feeder circuit breaker over the guide hooks on the panel, and then push the other end forward until the breaker snaps onto the hot bus bars (follow manufacturer's directions). Fold excess wire around the inside edge of the circuit breaker panel.

If necessary, open two tabs where the double-pole feeder breaker will fit, and then reattach the cover plate. Label the feeder breaker on the circuit index. Turn the main breaker on, but leave the feeder breaker off until all subpanel circuits have been connected and inspected.

1st proofs for reference only

Edition
age: 66

41940 - Advanced Home Wiring_064-128.indd 67

Pro-Vision C M Y K Text

Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 67

Baseboard Heaters

Baseboard heaters are a popular way to provide additional heating for an existing room or primary heat to a converted attic or basement.

Heaters are generally wired on a dedicated 240-volt circuit controlled by a thermostat. Several heaters can be wired in parallel and controlled by a single thermostat (see circuit map 15, page 39).

Baseboard heaters are generally surface-mounted without boxes, so in a remodeling situation, you only need to run cables before installing wallboard. Be sure to mark cable locations on the floor before installing drywall. Retrofit installations are also not difficult. You can remove existing baseboard and run new cable in the space behind. Baseboard heaters (and other heating equipment) get very hot and can ignite nearby combustible materials. Maintain the manufacturers recommended distance between the heater and materials such as curtains, blinds, and wood.

Tools & Materials ▸

- Drill/driver
- Wire stripper
- Cable ripper
- Wallboard saw
- Baseboard heater or heaters
- 240-thermostat (in-heater or in-wall)
- 12/2 NM cable
- Electrical tape
- Basic wiring supplies

Baseboard heaters can provide primary or supplemental heat for existing rooms or additions. Install heaters with clear space between the heater and the floor.

Baseboard Thermostats

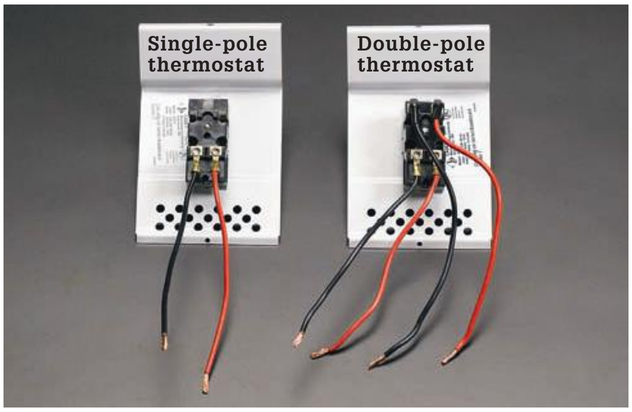

Single-pole and double-pole thermostats work in a similar manner, but double-pole models are safer. The single-pole model will open the circuit (causing shutoff) in only one leg of the power service. Double-pole models have two sets of wires to open both legs, lessening the chance that a person servicing the heater will contact a live wire.

In-heater thermostat

Wall-mount thermostat
60
70
80
50
90
OFF
LUX

In-heater and wall-mount are the two types of baseboard thermostats you can choose from. If you are installing multiple heaters, a single wall-mount thermostat is more convenient. Individual in-heater thermostats give you more zone control, which can result in energy savings.

How Much Heater Do You Need?

If you don't mind doing a little math, determining how many lineal feet of baseboard heater a room requires is not hard.

1. Measure the area of the room in square feet (length × width): ________
2. Multiply the area by 10 to get the baseline minimum wattage: ________
3. Add 5% for each newer window or 10% for each older window: ________
4. Add 10% for each exterior wall in the room: ________
5. Add 10% for each exterior door: ________
6. Add 10% if the space below is not insulated: ________
7. Add 20% if the space above is not well insulated: ________
8. Add 10% if ceiling is more than 8 ft. high: ________
9. Total of the baseline wattage plus all additions: ________
10. Divide this number by 250 (the wattage produced per foot of standard baseboard heater): ________
11. Round up to a whole number. This is the minimum number of feet of heater you need. ________

Note: It is much better to have more feet of heater than is required than fewer. Having more footage of heater does not consume more energy; it does allow the heaters to work more efficiently.

Planning Tips for Baseboard Heaters

- Baseboard heaters require a dedicated circuit. A 20-amp, 240-volt circuit of 12-gauge copper wire will power up to 16 ft. of heater.
- Do not install a heater beneath a wall receptacle. Cords hanging down from the receptacle are a fire hazard.
- Do not mount heaters directly on the floor. You should maintain at least 1" of clear space between the baseboard heater and the floor covering.
- Installing heaters directly beneath windows is a good practice.
- Locate wall thermostats on interior walls only, and do not install directly above a heat source.

1st
proofs for reference only
Edition
age: 68
41940 - Advanced Home Wiring_064-128.indd 69
10/10/14 12:05 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 69

How to Install a 240-Volt Baseboard Heater

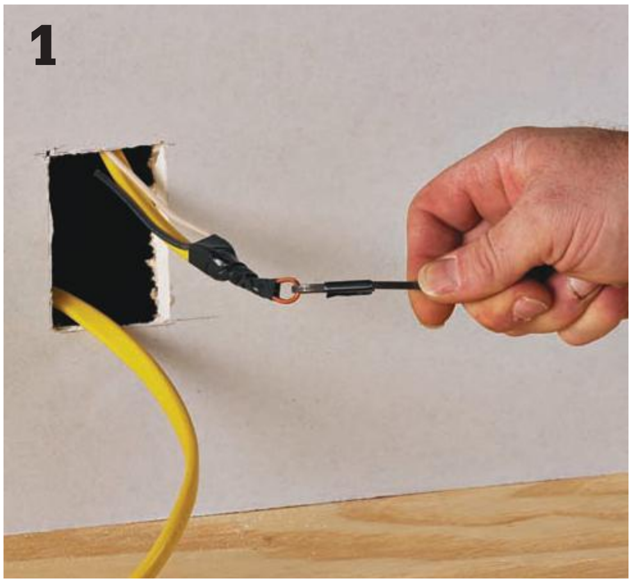

At the heater locations, cut a small hole in the drywall 3" to 4" above the floor. Pull 12/2 NM cables through the first hole: one from the thermostat, the other to the next heater. Pull all the cables for subsequent heaters. Middle-of-run heaters will have two cables, while end-of-run heaters have only one cable. (See also circuit map 15, page 39.)

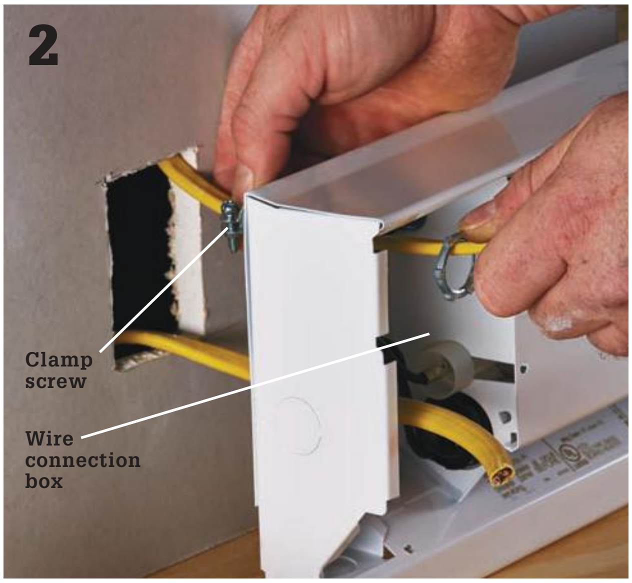

Remove the cover on the wire connection box. Open a knockout for each cable that will enter the box, and then feed the cables through the cable clamps and into the wire connection box. Attach the clamps to the wire connection box, and tighten the clamp screws until the cables are gripped firmly.

Anchor the heater against wall about 1" off floor by driving flathead screws through the back of the housing and into studs. Strip away cable sheathing so at least ½" of sheathing extends into the heater. Strip ¾" of insulation from each wire using a combination tool.

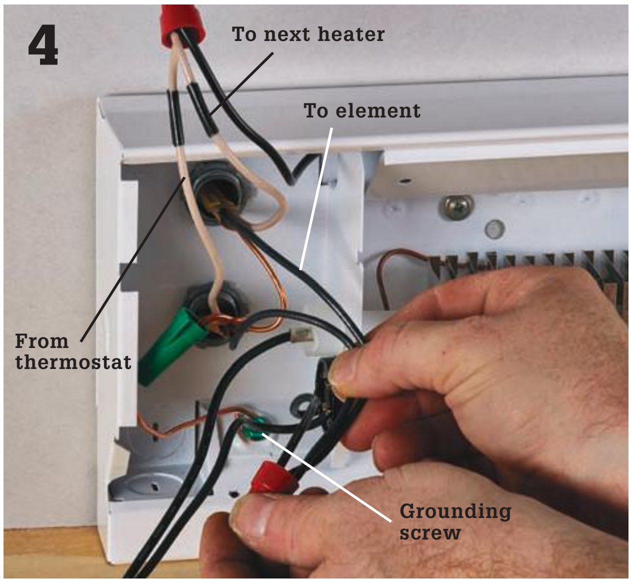

Make connections to the heating element if the power wires are coming from a thermostat or another heater controlled by a thermostat. See the next page for other wiring schemes. Connect the white circuit wires to one of the wire leads on the heater. Tag white wires with black tape to indicate they are hot. Connect the black circuit wires to the other wire lead. Connect a grounding pigtail to the green grounding screw in the box, and then join all grounding wires with a wire connector. Reattach the cover.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 70
10/10/14 12:05 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 70

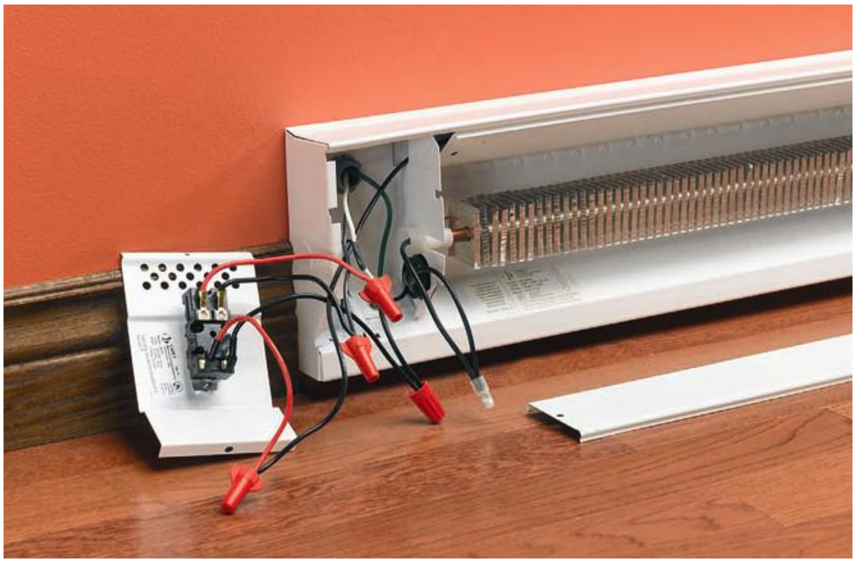

One heater with end-cap thermostat. Run both power leads (black plus tagged neutral) into the connection box at either end of the heater. If installing a single-pole thermostat, connect one power lead to one thermostat wire and connect the other thermostat wire, to one of the heater leads. Connect the other hot LINE wire to the other heater lead. If you are installing a double-pole thermostat, make connections with both legs of the power supply.

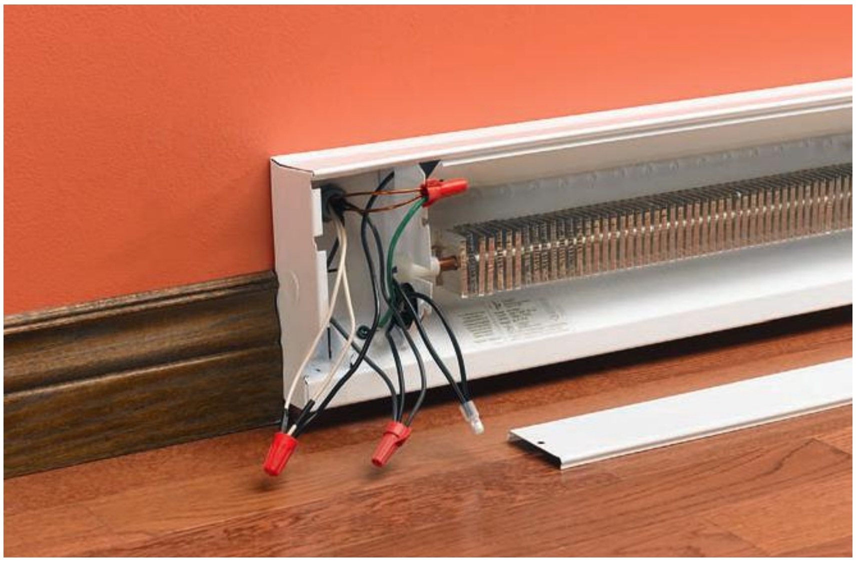

Multiple heaters. At the first heater, join both hot wires from the thermostat to the wires leading to the second heater in line. Be sure to tag all white neutrals hot. Twist copper ground wires together and pigtail them to the grounding screw in the baseboard heater junction box. This parallel wiring configuration ensures that power flow will not be interrupted to the downstream heaters if an upstream heater fails.

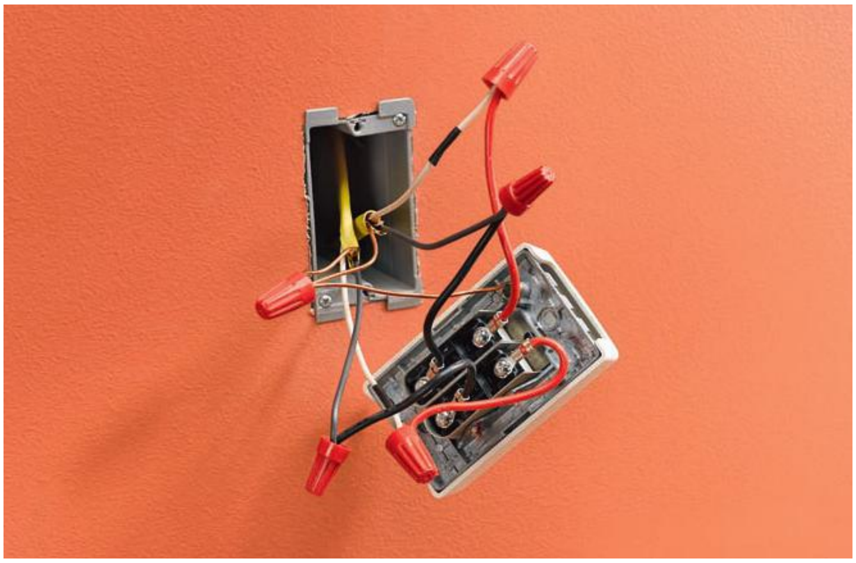

Wall-mounted thermostat. If installing a wall-mounted thermostat, the power leads should enter the thermostat first and then be wired to the individual heaters singly or in series. Hookups at the heater are made as shown in step 4. Be sure to tag the white neutral as hot in the thermostat box as well as in the heater box.

Wall Heaters

Installing a wall heater is an easy way to provide supplemental heat to a converted attic or basement without expanding an existing HVAC system.

Wall heaters are easy to install during a remodel (most have a separate can assembly that you attach to the framing before the drywall is installed). They can also be retrofitted.

Most models available at home centers use 120-volt current (shown below), but 240-volt models are also available.

Tools & Materials ▸

Drywall saw
Drill
Fish tape
Combination tool
Screwdrivers
12/2 NM cable
Wire connectors
Wall heater
Thermostat
(optional)
Wallboard saw

Wall heaters are an easy-to-install way to provide supplemental heat. Some models have built-in thermostats, while others can be controlled by a remote thermostat.

How to Install a Wall Heater in a Finished Wall

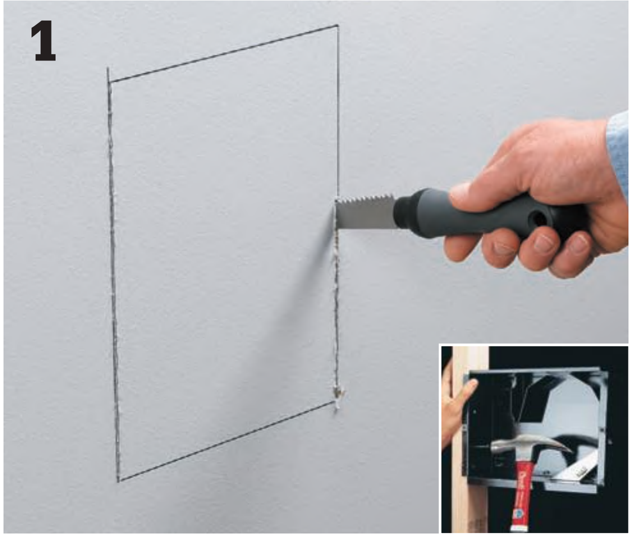

Make an opening in the wall for the heater. Use a stud finder to locate a stud in the area where you want to install the heater. Mark the opening for the heater according to the manufacturer's guidelines so that one side of the heater sits flush with a stud. Pay attention to clearance requirements. Cut the opening with a wallboard saw. If the wall is open, install the heater can before hanging drywall (inset).

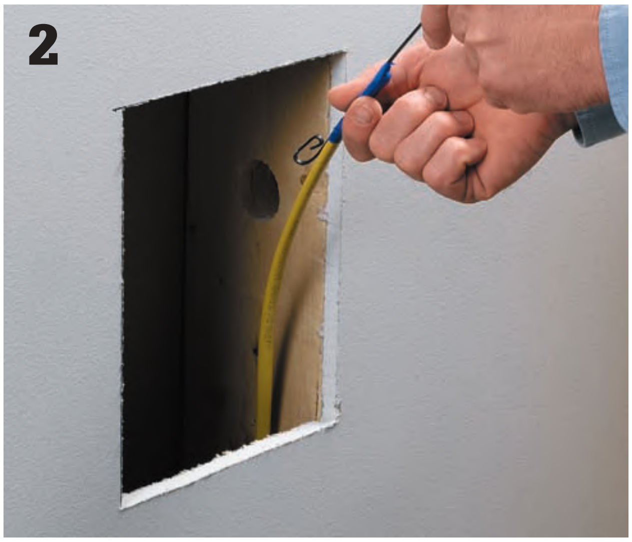

Turn power off, and test for power. Pull 12/2 NM cable from the main panel to the wall opening. If the heater is controlled by a separate thermostat, pull cable to the thermostat, and then run another cable from the thermostat to the heater location.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 72
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 72
Pro-Vis

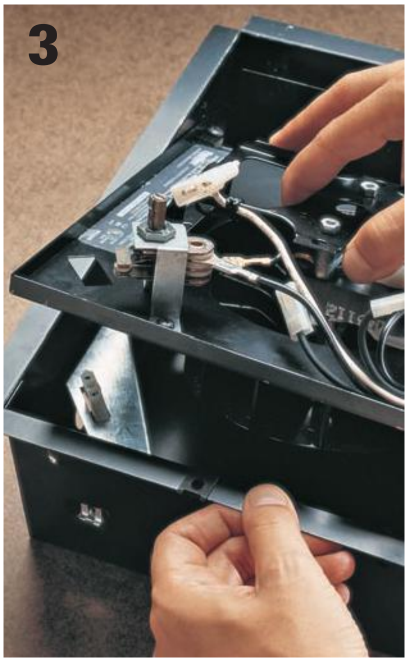
3

Disconnect and remove the motor unit from the heater can. Remove a knockout from the can, and route the cable into the can.

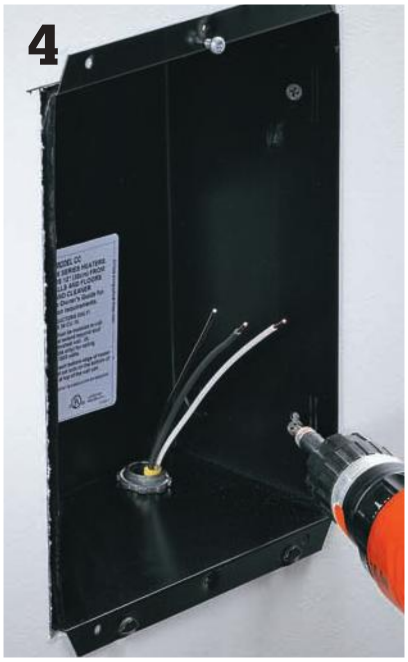
4

Install the can in the opening. Secure the cable with a clamp, leaving 8" to 12" of cable exposed. Attach the can to the framing as directed by the manufacturer.

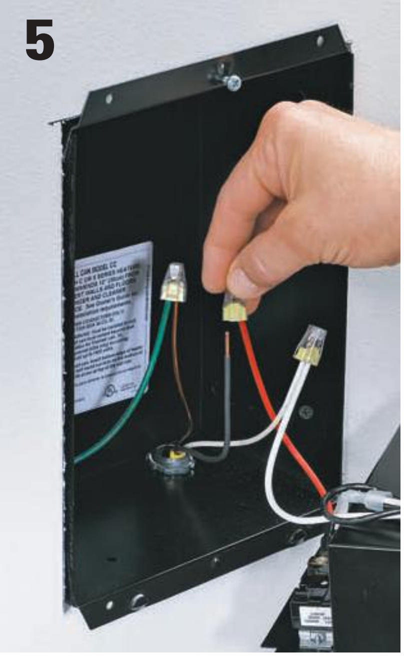
5

Wire the heater. Connect the black circuit wire to one of the black heater leads. Connect the white circuit wire to the other lead. Connect the grounds.

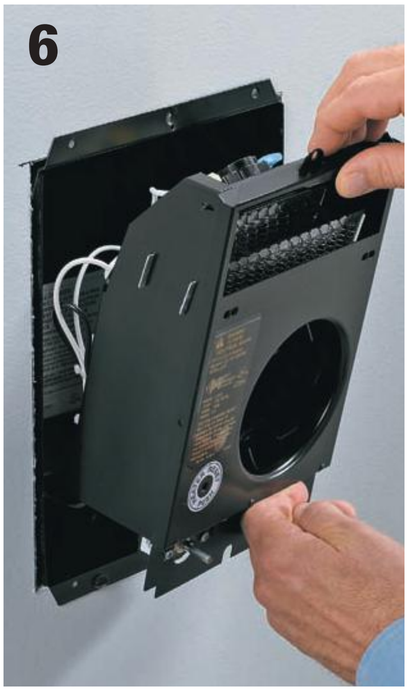
6

Secure the heater unit in the can as directed by the manufacturer. Reconnect the motor if necessary. Attach the grill and thermostat knob as directed. Connect the new circuit breaker at the main panel.

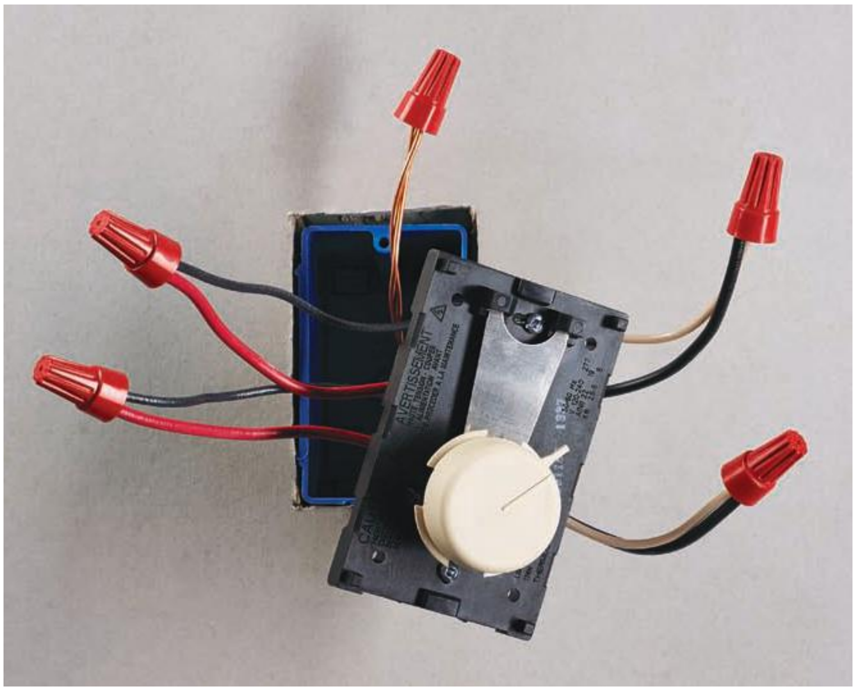

Variation: Connect a thermostat to control a wall heater. Some wall heaters do not use built-in thermostats. Install a thermostat in the heater circuit before the wall heater. Connect the black and the white wires coming from the main panel to the red leads on the thermostat. Connect the wires going to the heater to the black leads on the thermostat. Connect the grounds.

1st proofs for reference only
Edition
age: 72
41940 - Advanced Home Wiring_064-128.indd 73
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 73

Backup Power Supply

Installing a backup generator is an invaluable way to prepare your family for emergencies. The simplest backup power system is a portable gas-powered generator and an extension cord or two. A big benefit of this approach is that you can run a refrigerator and a few worklights during a power outage with a tool that can also be transported to remote job sites or on camping trips when it's not doing emergency backup duty. This is also the least expensive way to provide some backup power for your home. You can purchase a generator at most home centers and be up and running in a matter of hours. If you take this approach, it is critically important that you make certain any loads being run by your generator are disconnected from the utility power source.

The next step up is to incorporate a manual transfer switch for your portable generator. Transfer switches are permanently hardwired to your service panel. They are mounted on either the interior or the exterior of your house between the generator and the service panel. You provide a power feed from the generator into the switch. The switch is wired to selected essential circuits in your house, allowing you to power lights, furnace blowers, and other loads that can't easily be run with an extension cord. But perhaps the most important job a transfer switch performs is to disconnect the utility power. If the inactive utility power line is attached to the service panel, "backfeed" of power from your generator to the utility line can occur when the generator kicks in. This condition could be fatal to line workers who are trying to restore power. The potential for backfeed is the main reason many municipalities insist that only a licensed electrician hook up a transfer switch. Using a transfer switch not installed by a professional may also void the warranty of the switch and the generator.

Automatic transfer switches turn on the generator and switch off the utility supply when they detect a significant drop in line voltage. They may be installed with portable generators, provided the generator is equipped with an electric starter.

Large standby generators that resemble central air conditioners are the top of the line in backup power supply systems. Often fueled by home natural gas lines or propane tanks that offer a bottomless fuel source, standby generators are made in sizes with as much as 20 to 40 kilowatts of output—enough to supply all of the power needs of a 5,000-sq.-ft. home.

Generators have a range of uses. Large hard-wired models can provide instant emergency power for a whole house. Smaller models (below) are convenient for occasional short-term backup as well as job sites or camping trips.

HONDA
EU3000is
HONDA

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 74
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 74

Options for Backup Generators

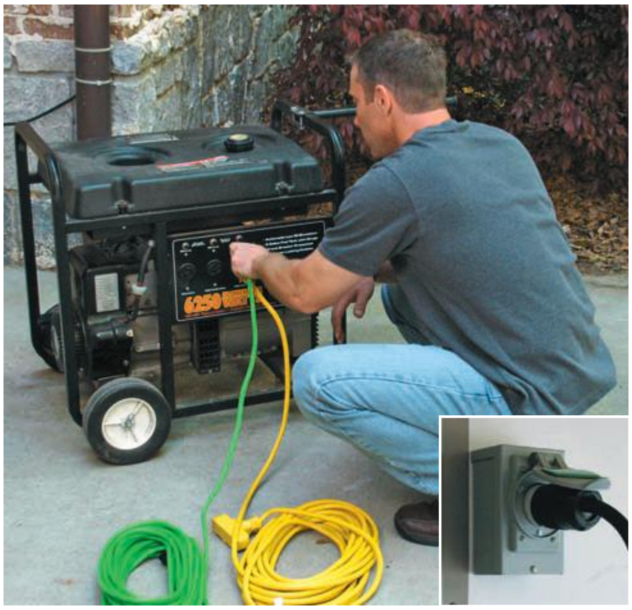

A 2,000- to 5,000-watt gas-powered generator and a few extension cords can power lamps and an appliance or two during shorter-term power outages. Appliances must not be connected to household wiring and the generator simultaneously. Never plug a generator into an outlet. Never operate a generator indoors. Run extension cords through a garage door.

A permanent transfer switch patches electricity from a large portable generator through to selected household circuits via an inlet at your service panel (inset), allowing you to power hardwired fixtures and appliances with the generator.

GUARDIAN
QUIETSOURCE

For full, on-demand backup service, install a large standby generator wired through to an automatic transfer panel. In the event of a power outage, the household system instantly switches to the generator.

1st
proofs for reference only

Edition
age: 74
41940 - Advanced Home Wiring_064-128.indd 75
Pro-Vision C M Y K Text
10/10/14 12:05 AM
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 75

A Typical Backup System

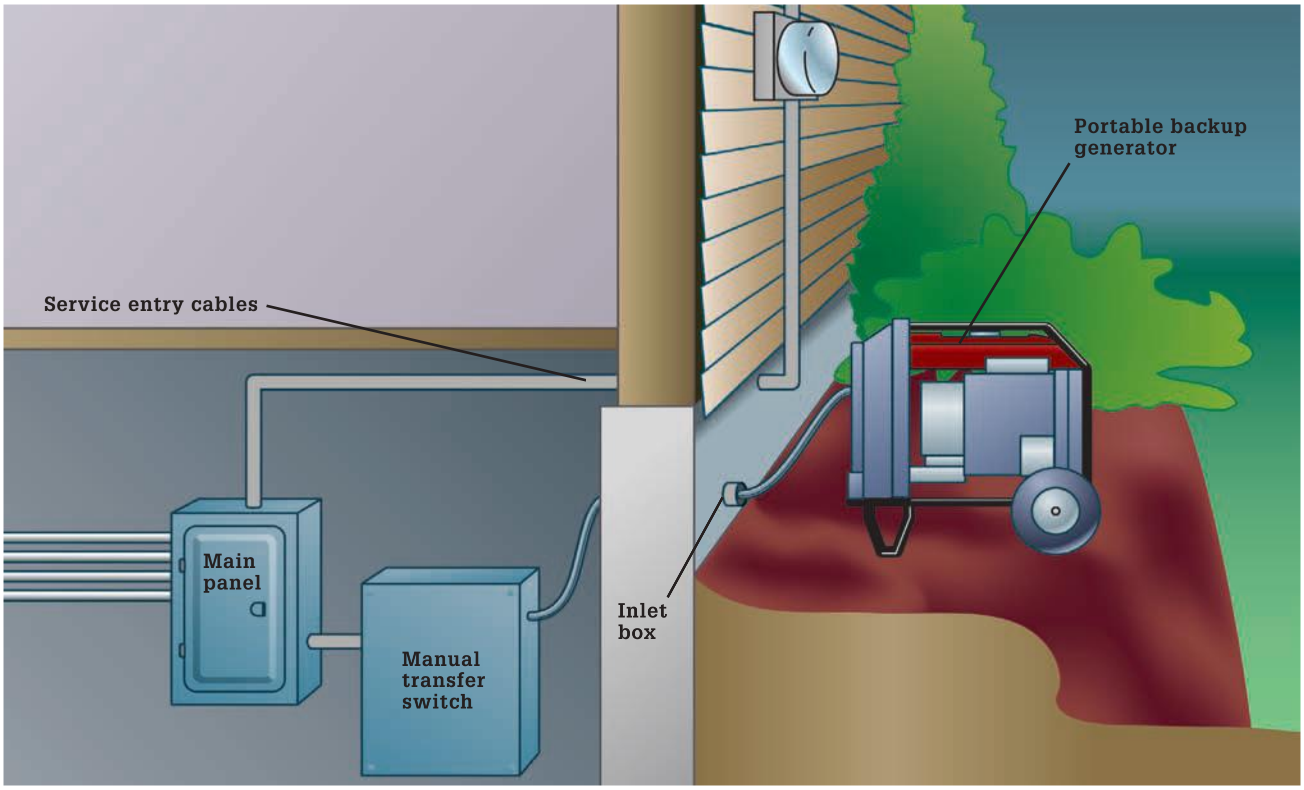

Backup generators supply power to a manual transfer switch, which disconnects the house from the main service wires and routes power from the generator through selected household circuits.

Choosing a Generator

Choosing a generator for your home's needs requires a few calculations. The chart below gives an estimate of the size of generator typically recommended for a house of a certain size. You can get a more accurate number by adding up the power consumption (the watts) of all the circuits or devices to be powered by a generator. It's also important to keep in mind that, for most electrical appliances, the amount of power required at the moment you flip the ON switch is greater than the number of watts required to keep the device running. For instance, though an air conditioner may run on 15,000 watts of power, it will require a surge of 30,000 watts at startup (the power range required to operate an appliance is usually listed somewhere on the device itself). These two numbers are called run watts and surge watts. Generators are typically sold according to run watts (a 5,000-watt generator can sustain 5,000 watts). They are also rated for a certain number of surge watts (a 5,000-watt generator may be able to produce a surge of 10,000 watts). If the surge watts aren't listed, ask, or check the manual. Some generators can't develop many more surge watts than run watts; others can produce twice as much surge as run wattage.

It's not necessary to buy a generator large enough to match the surge potential of all your circuits (you won't be turning everything on simultaneously), but surge watts should factor in your purchasing decision. If you will be operating the generator at or near capacity, it is also a wise practice to stagger startups for appliances.

SIZE OF HOUSE (IN SQUARE FEET)	RECOMMENDED GENERATOR SIZE (IN KILOWATTS)
Up to 2,700	5–11
2,701–3,700	14–16
3,701–4,700	20
4,701–7,000	42–47

Types of Transfer Switches

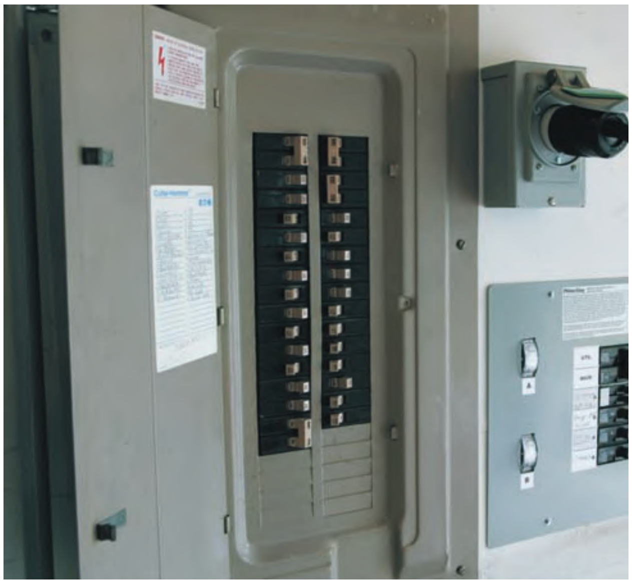

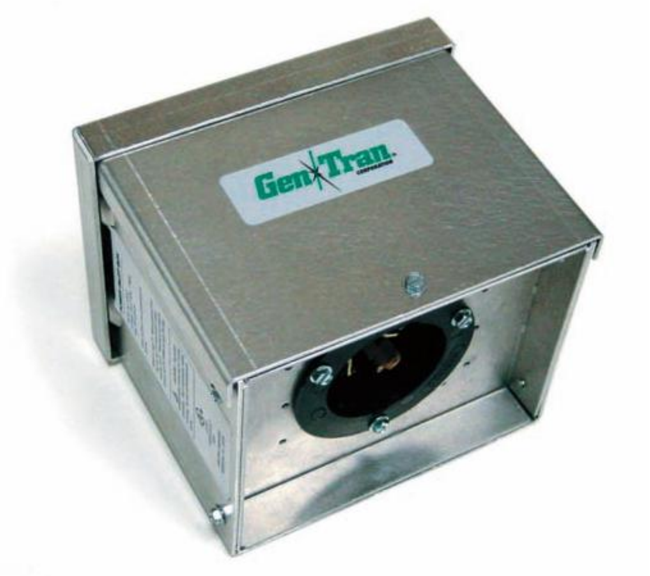

Cord-connected transfer switches (shown above) are hard-wired to the service panel (in some cases they're installed after the service panel and operate only selected circuits). These switches contain a male receptacle for a power supply cord connected to the generator. Automatic transfer switches (not shown) detect voltage drop-off in the main power line and switch over to the emergency power source.

When using a cord-connected switch, consider mounting an inlet box to the exterior wall. This will allow you to connect a generator without running a cord into the house.

Generator Tips ▸

HONDA
AC OUT120V
120/240V

If you'll need to run sensitive electronics such as computers or home theater equipment, look for a generator with power inverter technology that dispenses "clean power" with a stable sine wave pattern.

A generator that will output 240-volt service is required to run most central air conditioners. If your generator has variable output (120/240) make sure the switch is set to the correct output voltage.

Running & Maintaining a Backup System

Even with a fully automatic standby generator system fueled by natural gas or propane, you will need to conduct some regular maintenance and testing to make sure all systems are ready in the event of power loss. If you're depending on a portable generator and extension cords or a standby generator with a manual transfer switch, you'll also need to know the correct sequence of steps to follow in a power emergency. Switches and panels also need to be tested on a regular basis, as directed in your owner's manual. And be sure that all switches (both interior and exterior) are housed in an approved enclosure box.

Pull-cord starter

Smaller portable generators often use pull-cords instead of electric starters.

ANATOMY OF A PORTABLE BACKUP GENERATOR

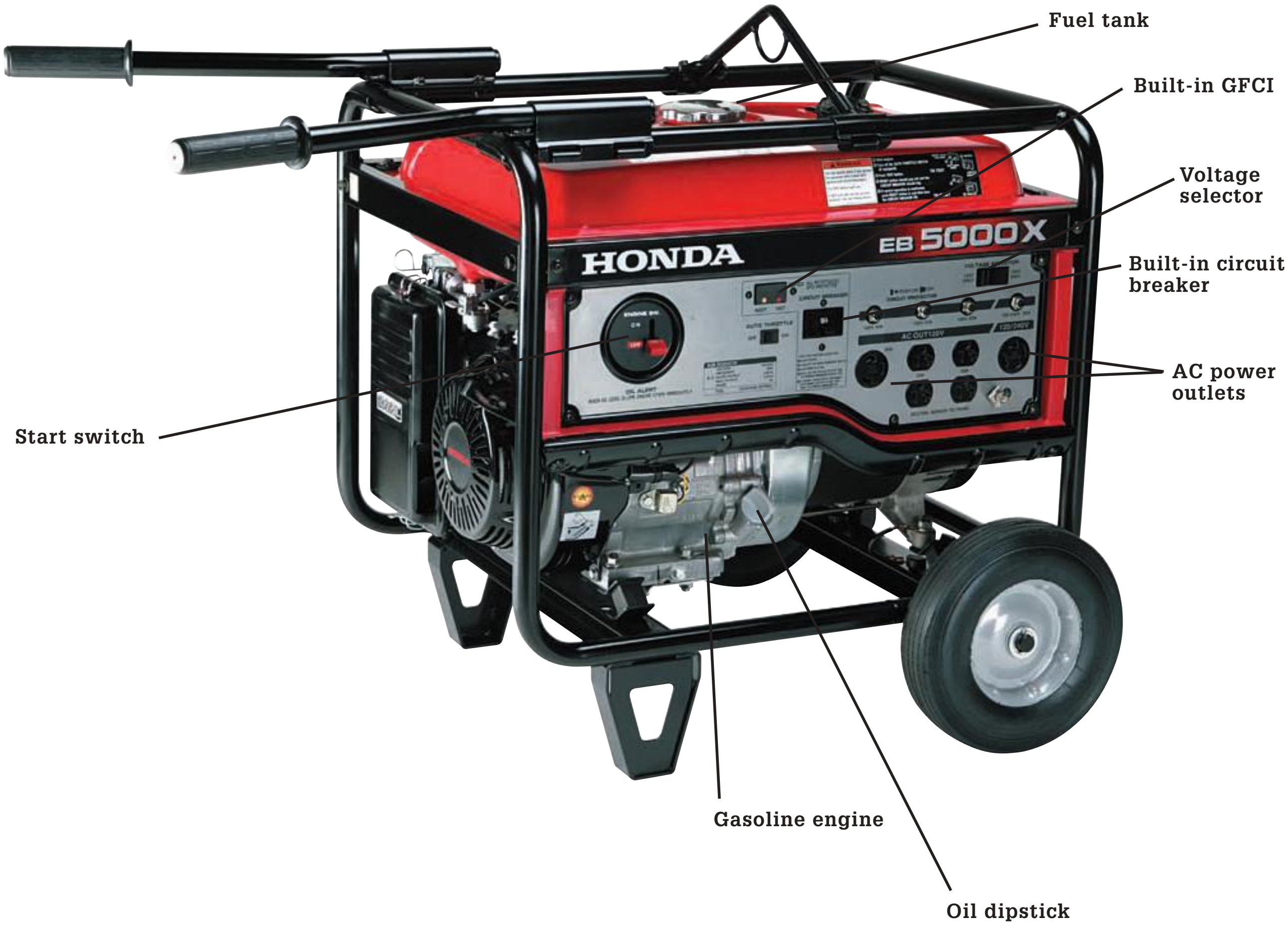
Fuel tank
Built-in GFCI
Voltage selector
Built-in circuit breaker
AC power outlets
Start switch
Gasoline engine
Oil dipstick
HONDA
EB 5000X

Portable generators use small gasoline engines to generate power. A built-in electronics panel sets current to AC or DC and the correct voltage. Most models will also include a built-in circuit breaker to protect the generator from damage in the event it is connected to too many loads. Better models include features like built-in GFCI protection. Larger portable generators may also feature electric starter motors and batteries for push-button starts.

Operating a Manual System During an Outage

Plug the generator in at the inlet box. Make sure the other end of the generator's outlet cord is plugged into the appropriate outlet on the generator (120-volt or 120/240-volt AC) and the generator is switched to the appropriate voltage setting.

Start the generator with the pull-cord or electric starter (if your generator has one). Let the generator run for several minutes before flipping the transfer switch.

Flip the manual transfer switch. Begin turning on loads one at a time by flipping breakers on, starting with the ones that power essential equipment. Do not overload the generator or the switch, and do not run the generator at or near full capacity for more than 30 minutes at a time.

Maintaining & Operating an Automatic Standby Generator

If you choose to spend the money and install a dedicated standby generator of 10,000 watts or more and operate it through an automatic transfer switch or panel, you won't need to lift a hand when your utility power goes out. The system kicks in by itself. However, you should follow the manufacturer's instructions for testing the system, changing the oil, and running the motor periodically.

Installing a Transfer Switch

A transfer switch is installed next to the main service panel to override the normal electrical service with power from a backup generator during a power outage. Manual transfer switches require an operator to change the power source, while automatic switches detect the loss of power, start the back-up generator, and switch over to the backup power feed. Because the amount of electricity created by a backup generator is not adequate to power all of the electrical circuits in your house, you'll need to designate a few selected circuits to get backup current (see page 81).

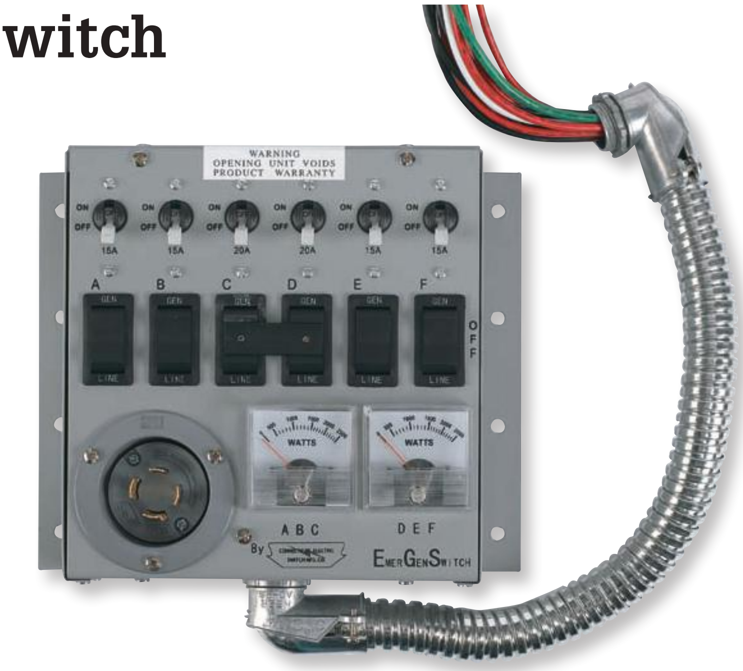

A manual transfer switch connects emergency circuits in your main panel to a standby generator.

Tools & Materials ▸

Circuit tester
Drill/driver
Screwdrivers
Hammer
Wire cutters
Cable ripper
Wire strippers
Level
Manual transfer switch
Screws
Wire connectors (yellow)
Standby power generator

One flip of a switch reassigns the power source for each critical circuit so your backup generator can keep your refrigerator, freezer, and important lights running during an outage of utility power.

1st proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 80
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 12:05 AM
Page: 80
Pro-Vis

Selecting Backup Circuits ▸

Before you purchase a backup generator, determine which loads you will want to power from your generator in the event of a power loss. Generally you will want to power your refrigerator, freezer, and maybe a few lights. Add up the running wattage ratings of the appliances you will power up to determine how large your backup generator needs to be. Because the startup wattage of many appliances is higher than the running wattage, avoid starting all circuits at the same time—it can cause an overload situation with your generator. Here are some approximate running wattage guidelines (see page 132 to 137 for more information on calculating electrical loads):

- Refrigerator: 750 watts
- Forced air furnace: 1,100 to 1,500 watts
- Incandescent lights: 60 watts per bulb (CFL and LED lights use less wattage)
- Sump pump: 800 to 1,000 watts
- Garage door opener: 550 to 1,100 watts
- Television: 300 watts

Add the wattage values of all the loads you want to power, and multiply the sum by 1.25. This will give you the minimum wattage your generator must produce. Portable standby generators typically output 5,000 to 7,500 watts. Most larger, stationary generators can output 10,000 to 20,000 watts (10 to 20 kilowatts).

How to Install a Manual Transfer Switch

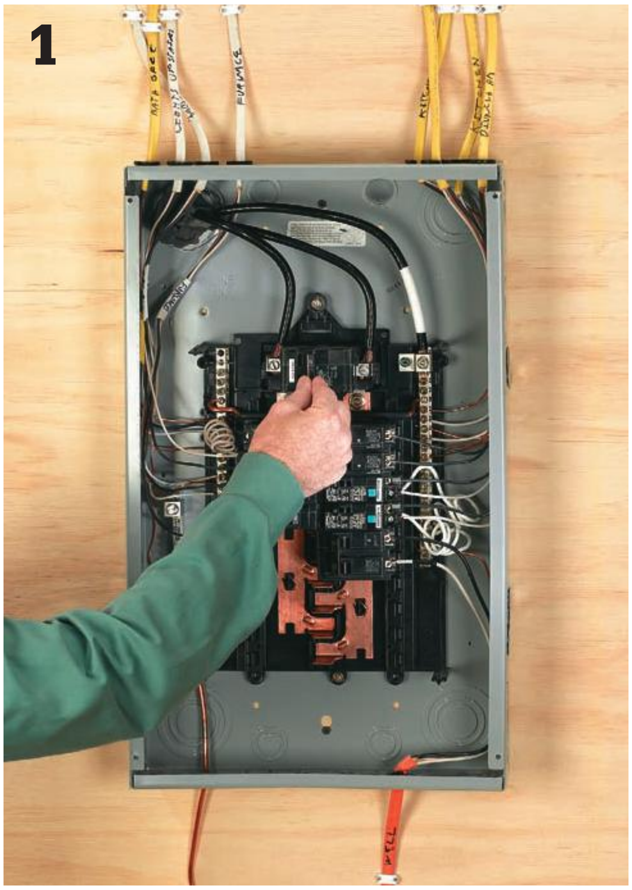

Turn off the main power breaker in your electrical service panel. CAUTION: The terminals where power enters the main breakers will still be energized.

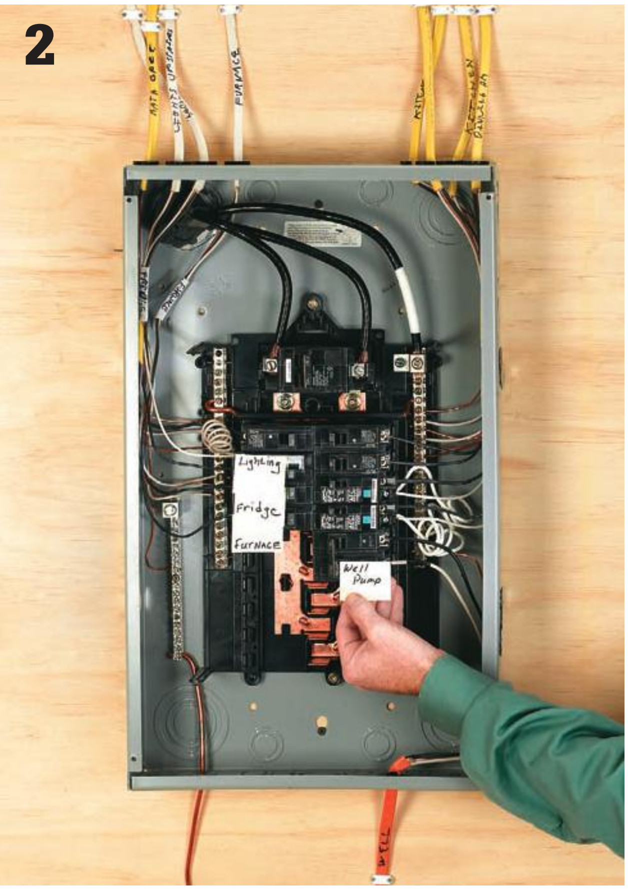

Determine which household circuits are critical for emergency usage during a power outage. Typically this will include the refrigerator, freezer, furnace, and at least one light or small appliance circuit.

(continued)

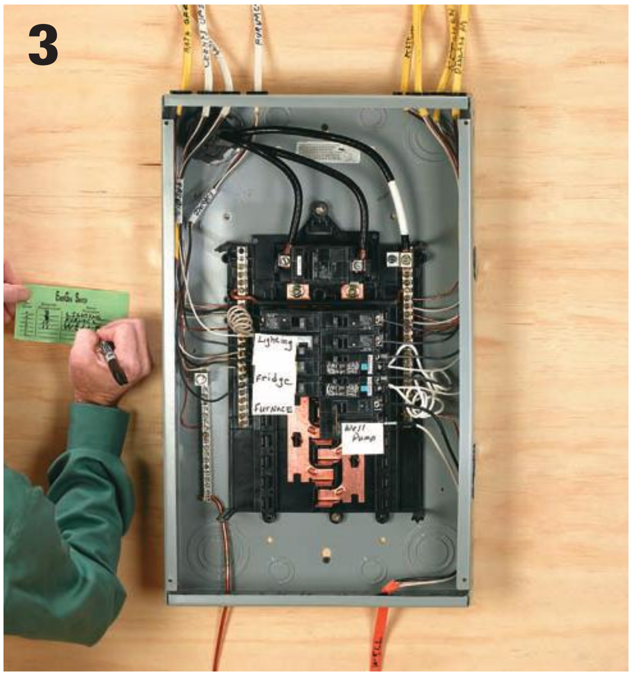

Match your critical circuits with circuit inlet on your prewired transfer switch. Try to balance the load as best you can in the transfer switch: For example, if your refrigerator is on the leftmost switch circuit, connect your freezer to the circuit farthest to the right. Double-pole (240-volt) circuits will require two 120-volt circuit connections. Also make sure that 15-amp and 20-amp circuits are not mismatched with one another.

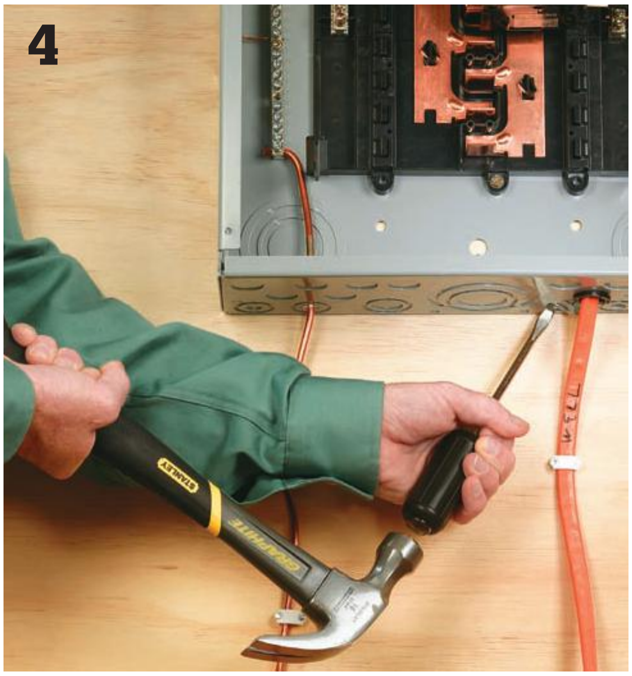

Select and remove a knockout at the bottom of the main service panel box. Make sure to choose a knockout that is sized to match the connector on the flexible conduit coming from the transfer switch.

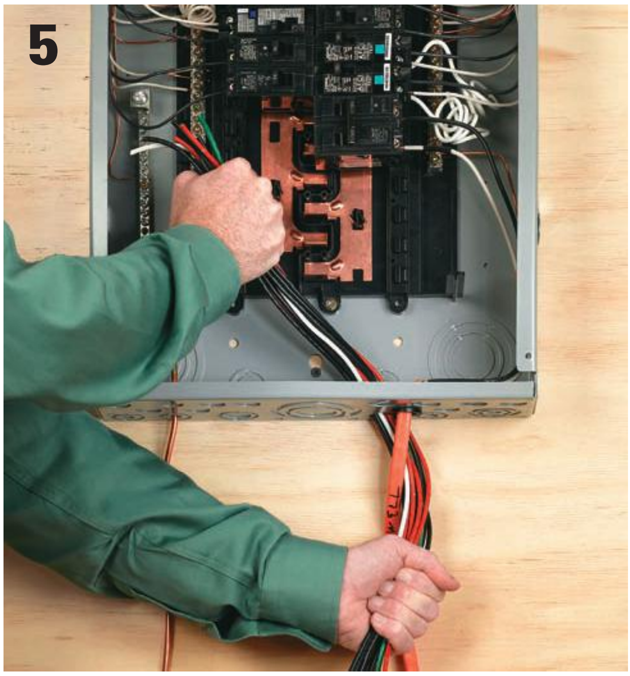

Feed the wires from the transfer switch into the knockout hole, taking care not to damage the insulation. You will note that each wire is labeled according to which circuit in the switch box it feeds.

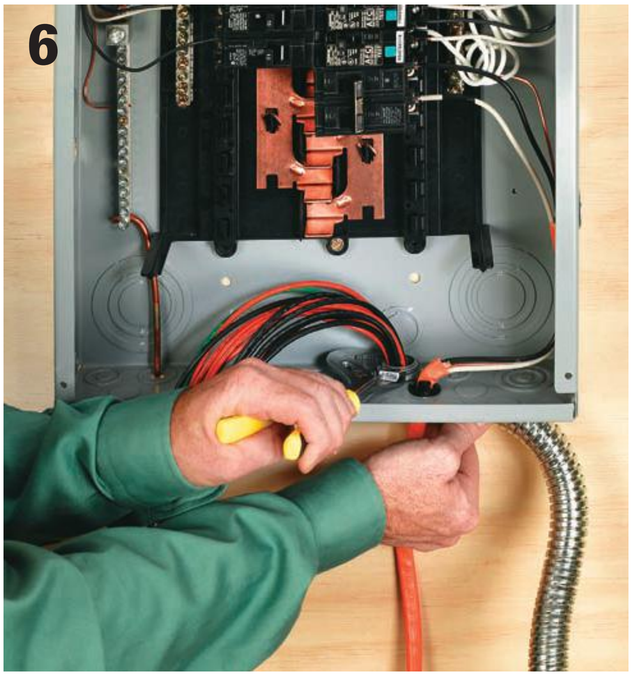

Secure the flexible conduit from the switch box to the main service panel using a locknut and a bushing where required.

1st proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 82
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 82
Pro-Vis

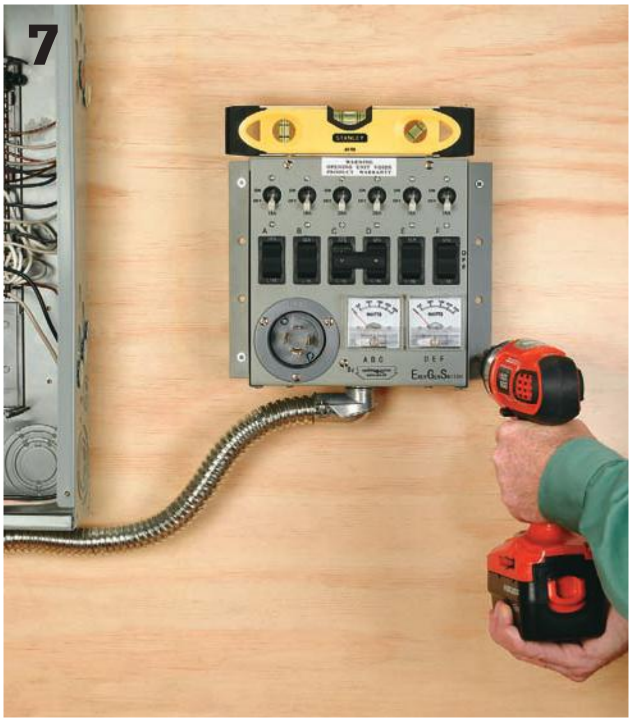

Attach the transfer switch box to the wall so the closer edge is about 18" away from the center of the main service panel. Use whichever connectors make sense for your wall type.

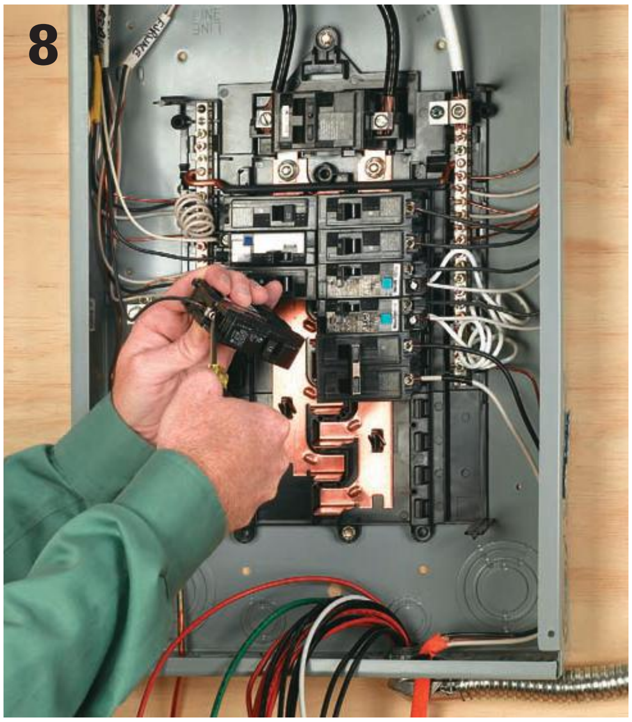

Remove the breaker for the first critical circuit from the main service panel box, and disconnect the hot wire lead from the lug on the breaker.

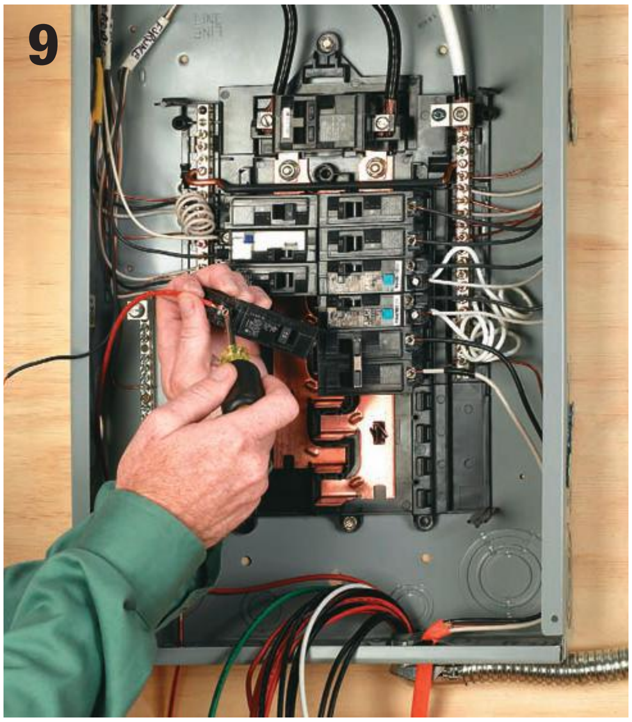

Locate the red wire for the switch box circuit that corresponds to the circuit you've disconnected. Attach the red wire to the breaker you've just removed, and then reinstall the breaker.

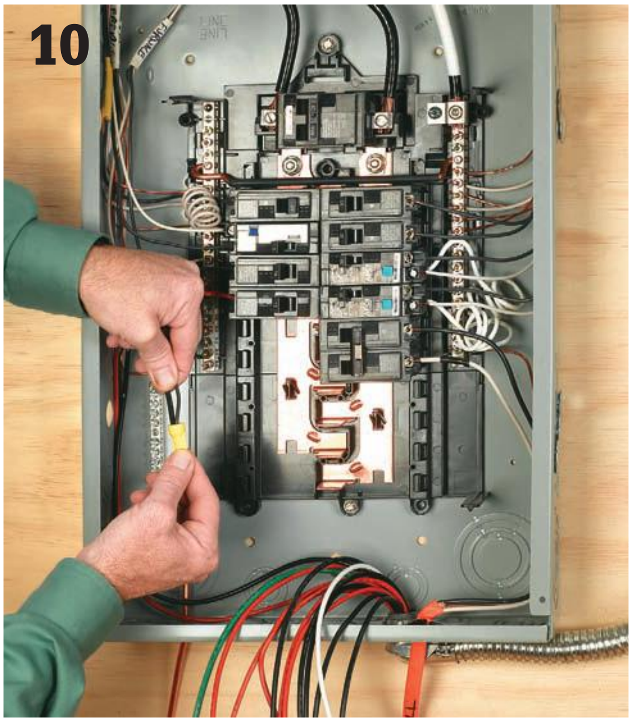

Locate the black wire from the same transfer switch circuit, and twist it together with the old feed wire, using a yellow wire connector. Tuck the wires neatly out of the way at the edges of the box. Proceed to the next circuit, and repeat the process.

(continued)

1st
proofs for reference only

Edition
age: 82

41940 - Advanced Home Wiring_064-128.indd 83
Pro-Vision C M Y K Text

Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 83

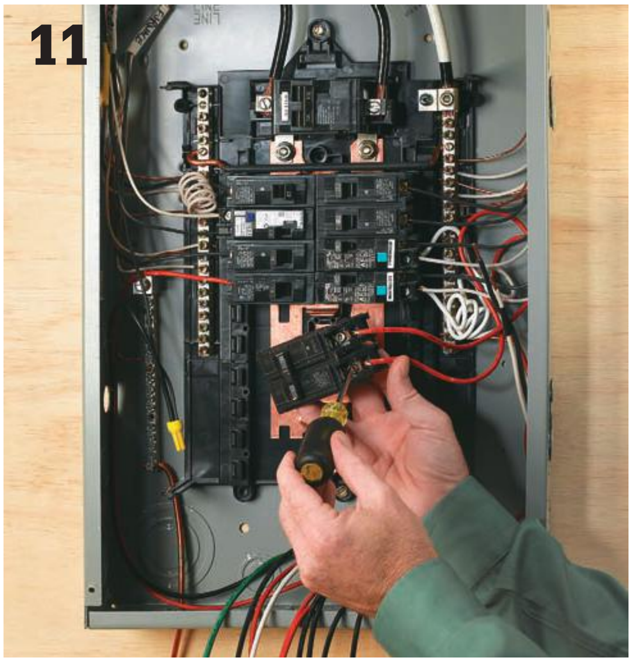

If any of your critical circuits are 240-volt circuits, attach the red leads from the two transfer switch circuits to the double-pole breaker. The two circuits originating in the transfer switch should be next to one another, and their switches should be connected with a handle tie. If you have no 240-volt circuits you may remove the preattached handle tie and use the circuits individually.

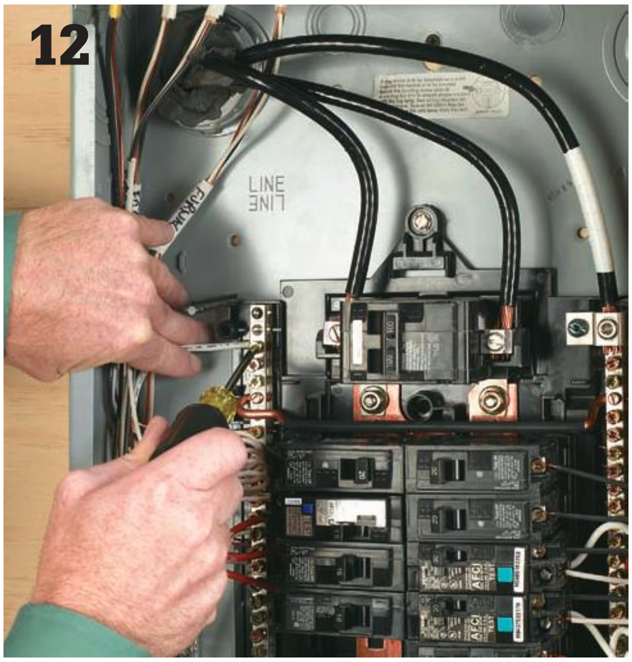

Once you have made all circuit connections, attach the white neutral wire from the transfer switch to an opening in the neutral bus bar of the main service panel.

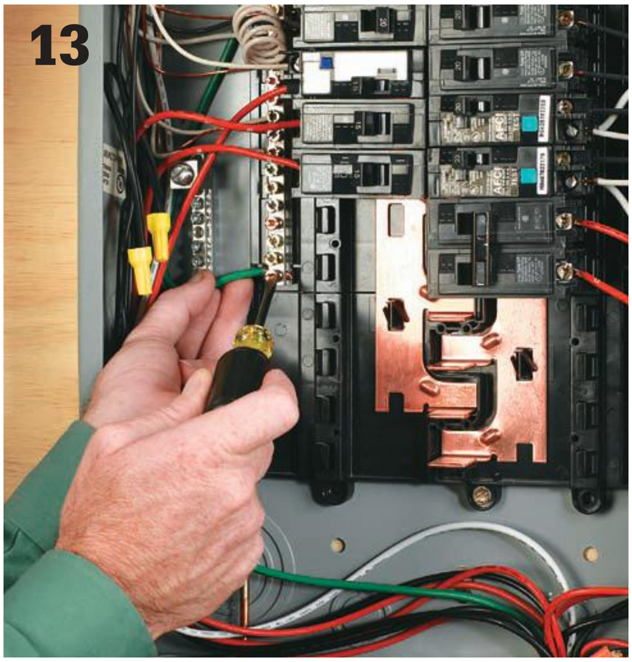

Attach the green ground wire from the transfer switch to an open port on the grounding bar in your main service panel. This should complete the installation of the transfer switch. Replace the cover on the service panel box, and make sure to fill in the circuit map on your switch box.

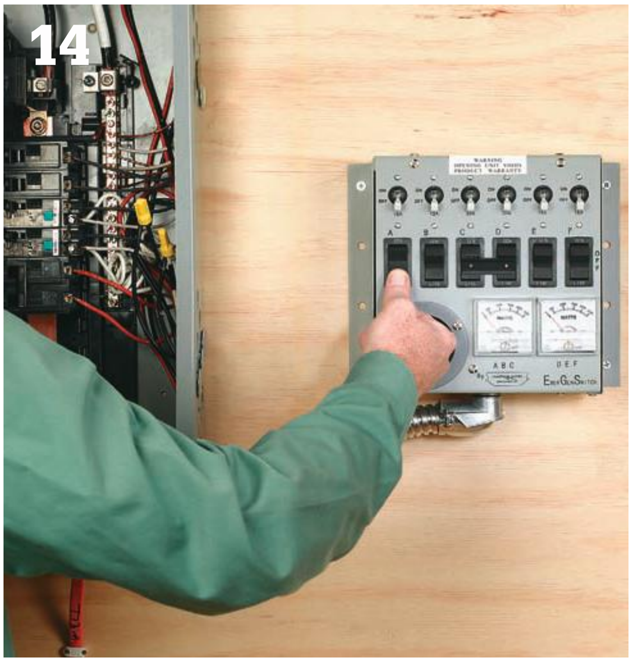

Begin testing the transfer switch by making sure all of the switches on it are set to the LINE setting. The power should still be OFF at the main panel breakers.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 84
Pro-Vision C M Y K Text
10/10/14 12:05 AM
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 84
Pro-Vis

Standby Generators ▸

KOHLER. POWER SYSTEMS

Make sure your standby generator is operating properly and has been installed professionally. See page 76 for information on choosing a generator that is sized appropriately for your needs.

15

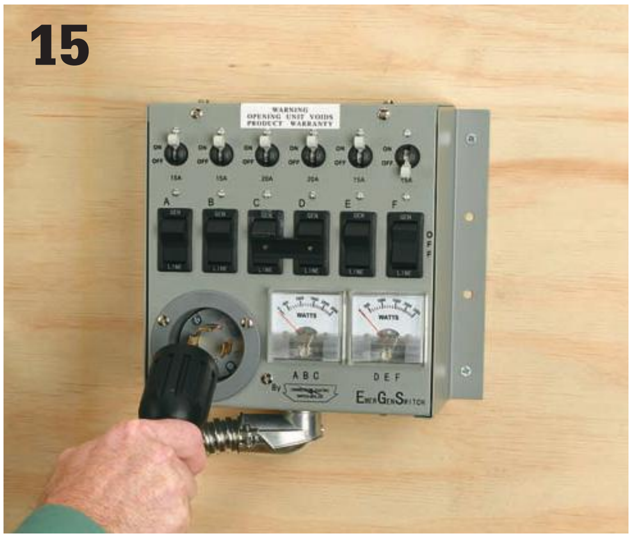

Before turning your generator on, attach the power cord from the generator to the switch box. Never attach or detach a generator cord with the generator running. Turn your standby power generator on, and let it run for a minute or two.

16

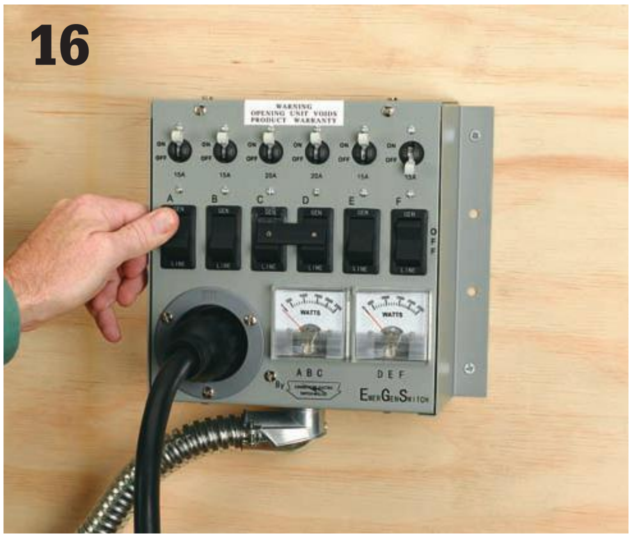

Flip each circuit switch on the transfer switch box to GEN, one at a time. Try to maintain balance by moving back and forth from circuits on the left and right side. Do not turn all circuits on at the same time. Observe the onboard wattage meters as you engage each circuit, and try to keep the wattage levels in balance. When you have completed testing the switch, turn the switches back to LINE, and then shut off your generator.

1st proofs for reference only

Edition
age: 84
41940 - Advanced Home Wiring_064-128.indd 85 10/10/14 12:05 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 85

Outbuildings

Nothing improves the convenience and usefulness of an outbuilding more than electrifying it. Running a new underground circuit from your house to an outbuilding lets you add receptacles and light fixtures both inside the outbuilding and on its exterior. If you run power to an outbuilding, you are required to install at least one receptacle.

Adding one or two 120-volt circuits is not complicated, but every aspect of the project is strictly governed by local building codes. Therefore, once you've mapped out the job and have a good idea of what's involved, visit your local building department to discuss your plans and obtain a permit for the work.

This project demonstrates standard techniques for running a circuit cable from the house exterior to a shed, plus the wiring and installation of devices inside the shed. To add a new breaker and make the final circuit connections to your home's main service panel, see page 52. If you run power to an outbuilding, you are required to install at least one receptacle.

First, determine how much current you will need. For basic electrical needs, such as powering a standard light fixture and small appliances or power tools, a 120-volt, 15-amp circuit should be sufficient. A small workshop may require one or two 120-volt, 20-amp circuits. If you need any 240-volt circuits or more than two 120-volt, 20-amp circuits, you will need to install at least a 60-amp subpanel with appropriate feeder wires. Installing a subpanel in an outbuilding is similar to installing one inside your home, but there are some important differences.

You may use #14 copper wire for one 120-volt, 15-amp circuit or #12 copper wire for one 120-volt, 20-amp circuit. Use #10 copper wire for two 120-volt, 20 amp circuits. Also, if the shed is more than 150 ft. away from the house, you may need heavier-gauge cable to account for voltage drop.

Most importantly, don't forget to call before you dig. Have all utility and service lines on your property marked even before you make serious project plans. This is critical for your safety of course, and it may affect where you can run the circuit cable.

Adding an electrical circuit to an outbuilding such as this shed greatly expands the activities the building will support and is also a great benefit for home security.

1st proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 86

Pro-Vision C M Y K Text

Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 86

Pro-Vis

Tools & Materials ▸

- Spray paint
- Trenching shovel (4" wide blade)
- 4" metal junction box
- Metal L-fittings (2) and conduit nipple for conduit
- Wood screws
- Conduit with watertight threaded and compression fittings
- Wrenches
- Hacksaw
- 90° sweeps for conduit (2)
- Plastic conduit bushings (2)
- Pipe straps
- Silicone caulk and caulk gun
- Double-gang boxes, metal (2)
- One exterior receptacle box (with cover)
- Single-pole switches (2)
- Interior ceiling light fixture and metal fixture box
- Exterior motion-detector fixture and plastic fixture box
- EMT metal conduit and fittings for inside the shed
- Utility knife
- UF two-wire cable (12 gauge)
- THNN wire (12 gauge)
- 20-amp GFCI-protected circuit breaker
- Wire stripper
- Pliers
- Screwdrivers
- Wire connectors
- Hand tamper
- Schedule 80 conduit
- Eye protection

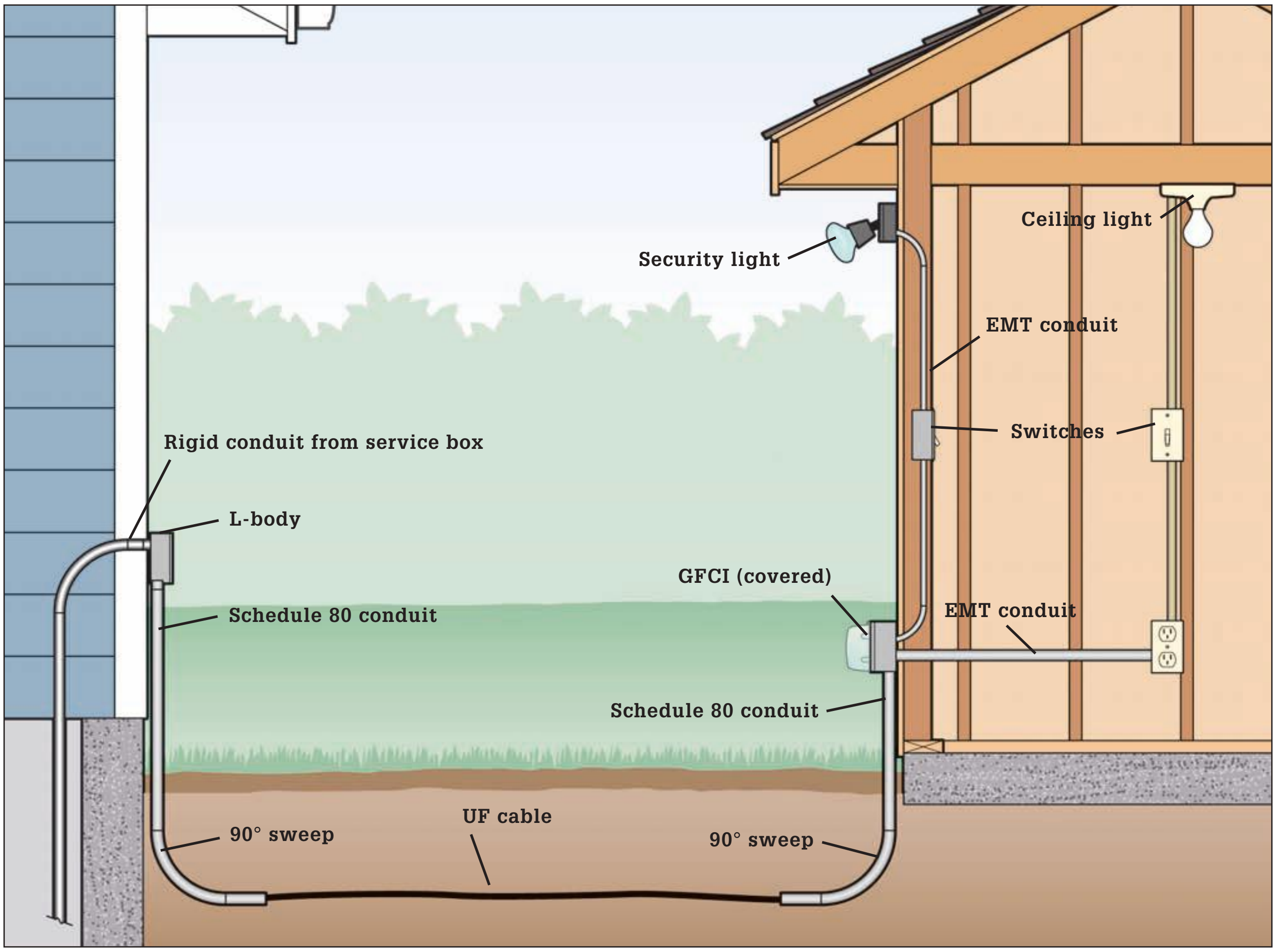
Ceiling light
Security light
EMT conduit
Switches
Rigid conduit from service box
L-body
GFCI (covered)
Schedule 80 conduit
EMT conduit
Schedule 80 conduit
UF cable
90° sweep
90° sweep

A basic outdoor circuit starts with a waterproof fitting at the house wall connected to a junction box inside. The underground circuit cable—rated UF (underground feeder)—runs in a 24"-deep trench and is protected from exposure at both ends by metal or PVC conduit. Inside the shed, standard NM cable runs through metal conduit to protect it from damage (not necessary if you will be adding interior wallcoverings). All receptacles in the shed must be GFCI protected.

How to Wire an Outbuilding

Identify the circuit's exit point at the house and entry point at the shed and mark them. Mark the path of the trench between the exit and entry points using spray paint. Make the route as direct as possible. Dig the trench to the depth required by local code (24") using a narrow trenching shovel.

From outside, drill a hole through the exterior wall and the rim joist at the exit point for the cable (you'll probably need to install a bit extender or an extra-long bit in your drill). Make the hole just large enough to accommodate the L-body conduit fitting and conduit nipple.

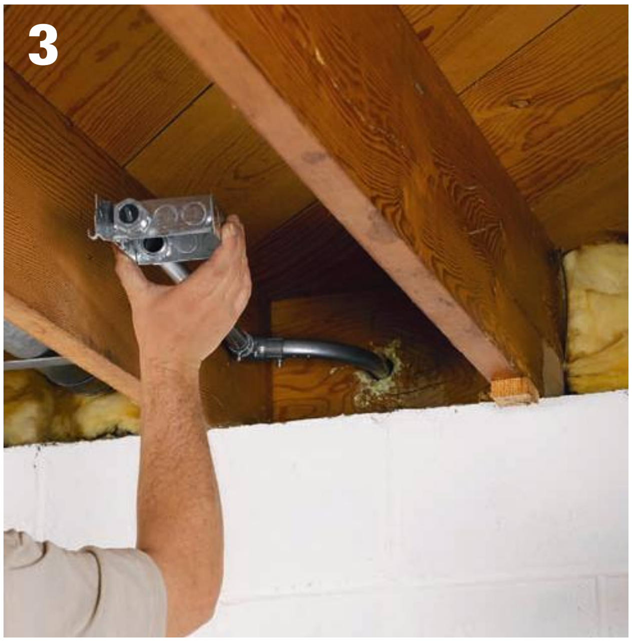

Assemble the conduit and junction box fittings that will penetrate the wall. Here, we attached a 12" piece of ¾" IMC (intermediate metallic conduit) and a sweep to a metal junction box with a compression fitting and then inserted the conduit into the hole drilled in the rim joist. The junction box is attached to the floor joist.

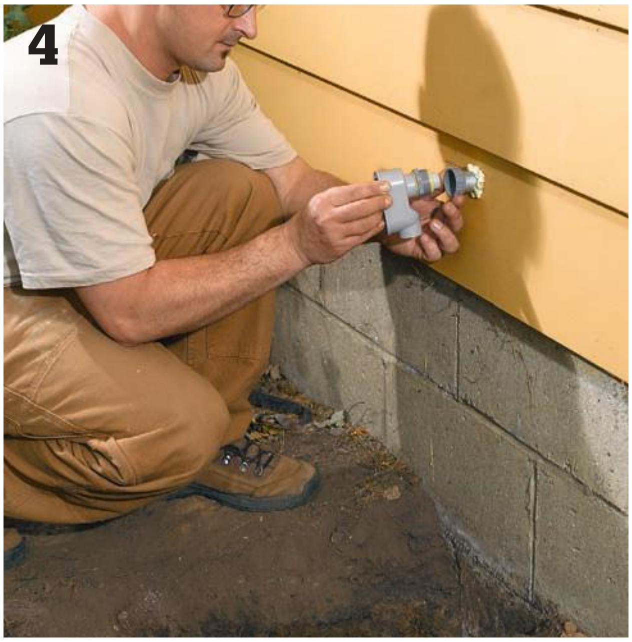

From outside, seal the hole around the conduit with expandable spray foam or caulk, and then attach the free end of the conduit to the back of a waterproof L-body fitting. Mount the L-body fitting to the house exterior with the open end facing downward.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 88
10/10/14 12:05 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 88

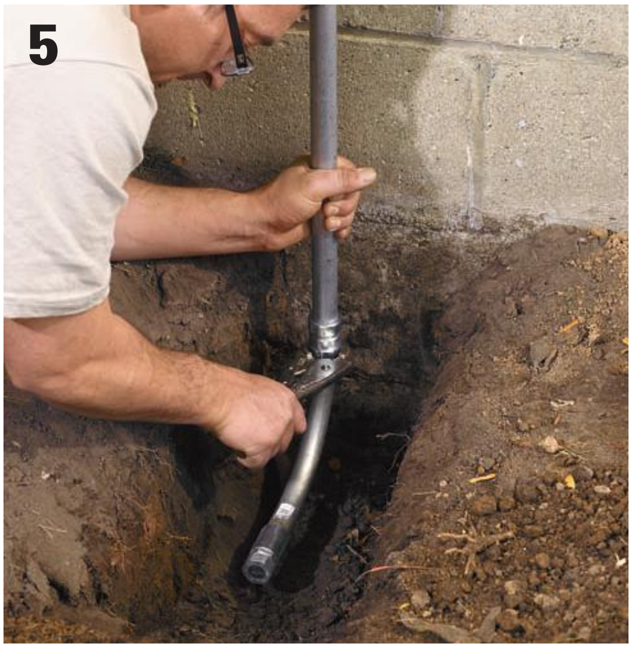

Cut a length of IMC to extend from the L-fitting down into the trench using a hacksaw. Deburr the cut edges of the conduit. Secure the conduit to the L-fitting, and then attach a 90° sweep to the bottom end of the conduit using compression fittings. Add a bushing to the end of the sweep to protect the circuit cable. Anchor the conduit to the wall with a corrosion-resistant pipe strap.

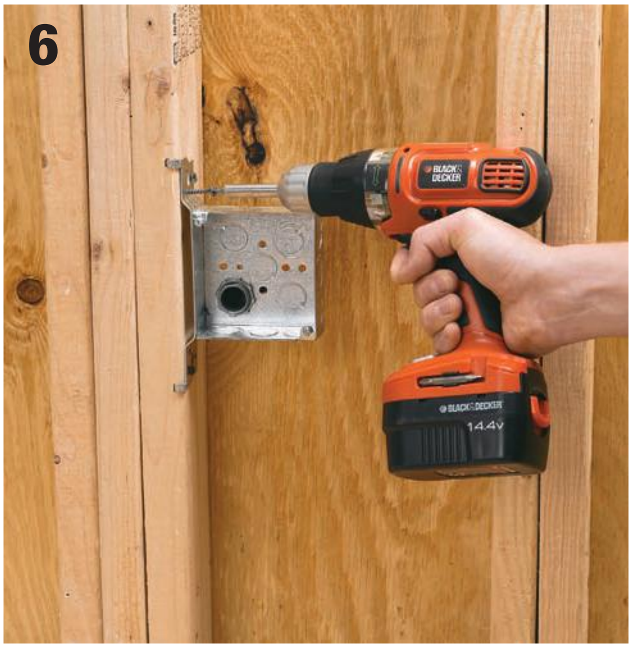

Inside the shed, drill a ¾" dia. hole in the shed wall. On the interior of the shed, mount a junction box with a knock-out removed to allow the cable to enter through the hole. On the exterior side directly above the end of the UF trench, mount an exterior-rated receptacle box with cover. The plan (and your plan may differ) is to bring power into the shed through the hole in the wall behind the exterior receptacle.

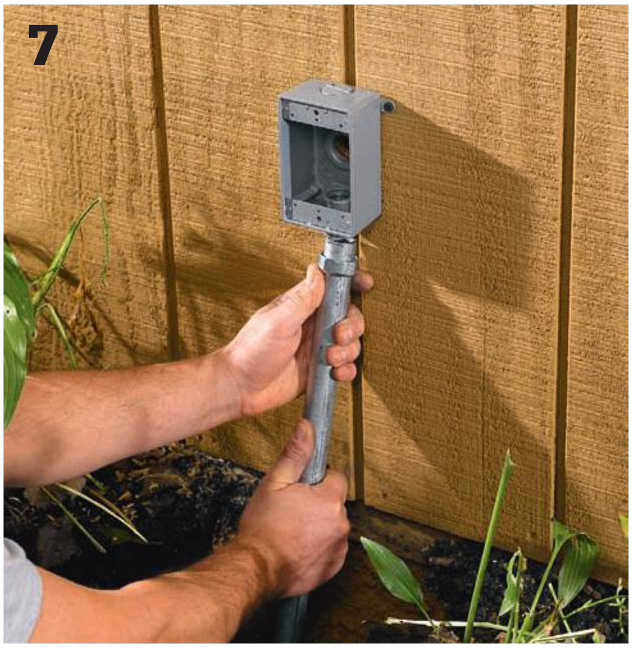

Run conduit from the exterior box down into the trench. Fasten the conduit to the building with a strap. Add a 90° sweep and bushing, as before. Secure the conduit to the box with an offset fitting. Anchor the conduit with pipe straps, and seal the entry hole with caulk.

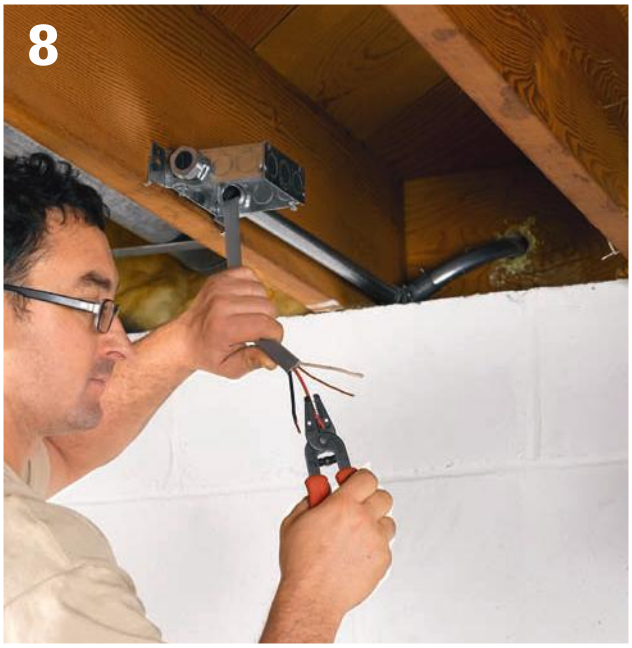

Run UF cable from the house to the outbuilding. Feed one end of the UF circuit cable up through the sweep and conduit and into the L-fitting at the house (the back or side of the fitting is removable to facilitate cabling). Run the cable through the wall and into the junction box, leaving at least 12" of extra cable at the end.

(continued)

1st proofs for reference only
Edition
age: 88
41940 - Advanced Home Wiring_064-128.indd 89
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 12:05 AM
Page: 89

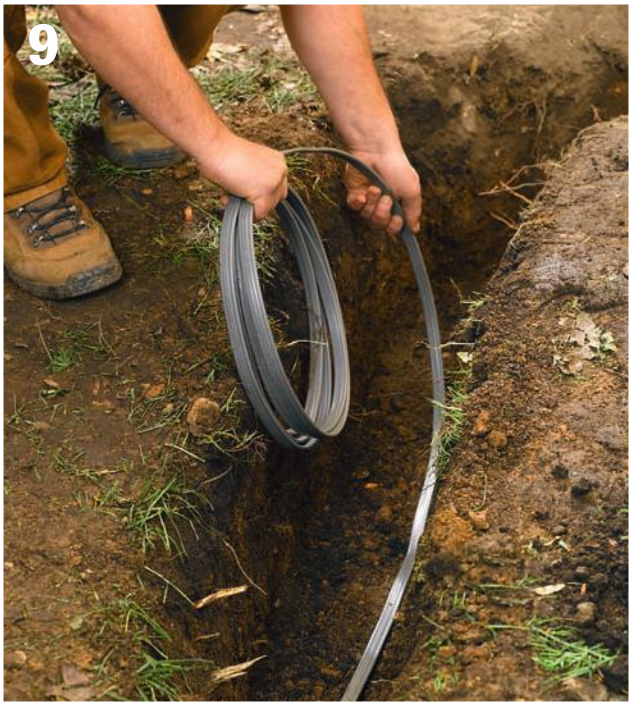

Lay the UF cable into the trench, making sure it is not twisted and will not contact any sharp objects. Roll out the cable, and then feed the other end of the cable up through the conduit and into the receptacle box in the shed, leaving 12" of slack.

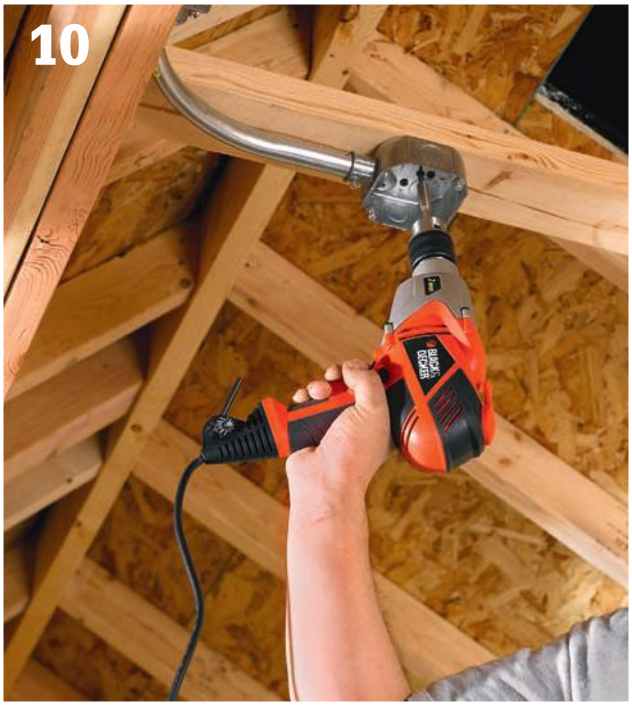

Inside the outbuilding, install the remaining boxes for the other switches, receptacles, and lights. With the exception of plastic receptacle boxes for exterior exposure, use metal boxes if you will be connecting the boxes with metal conduit.

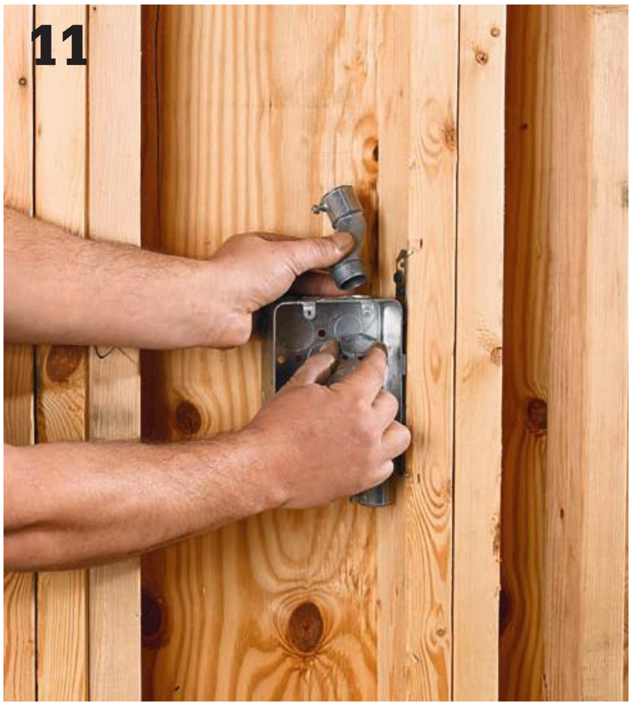

Connect the electrical boxes with conduit and fittings. Inside the outbuilding, you may use inexpensive EMT to connect receptacle, switch, and fixture boxes. Once you've planned your circuit routes, start by attaching couplings to all of the boxes.

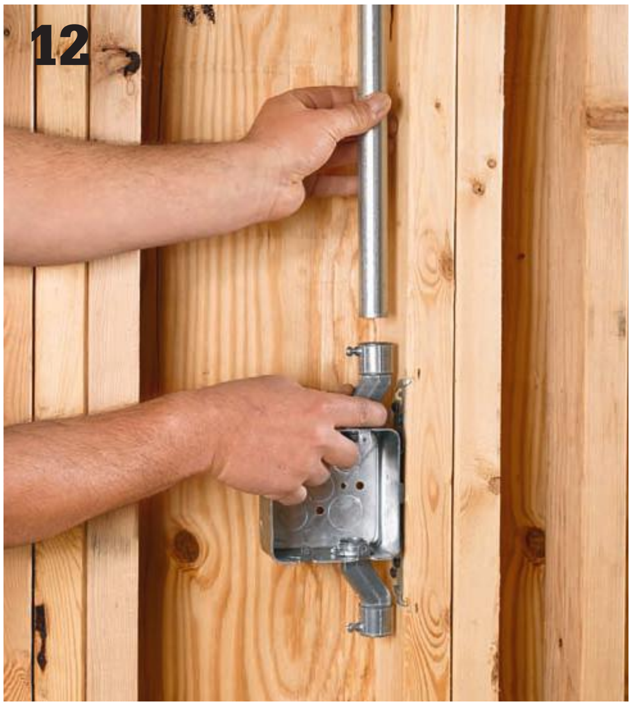

Cut a length of conduit to fit between the coupling and the next box or fitting in the run. If necessary, drill holes for the conduit through the centers of the wall studs. Attach the conduit to the fitting that you attached to the first box.

1st
proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 90
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 90
Pro-Vis

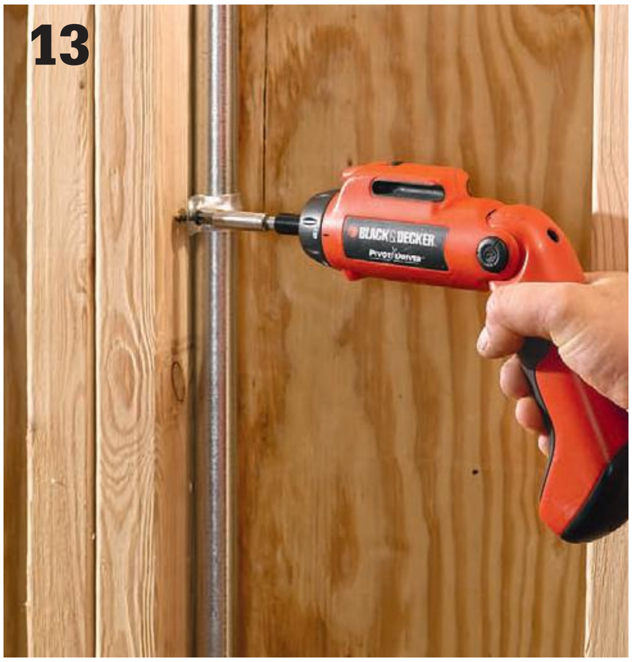

If you are surface-mounting the conduit or running it up or down next to wall studs, secure it with straps no more than 3 ft. apart. Use elbow fittings for 90° turns and setscrew couplings for joining straight lengths as needed. Make holes through the wall studs only as large as necessary to feed the conduit through.

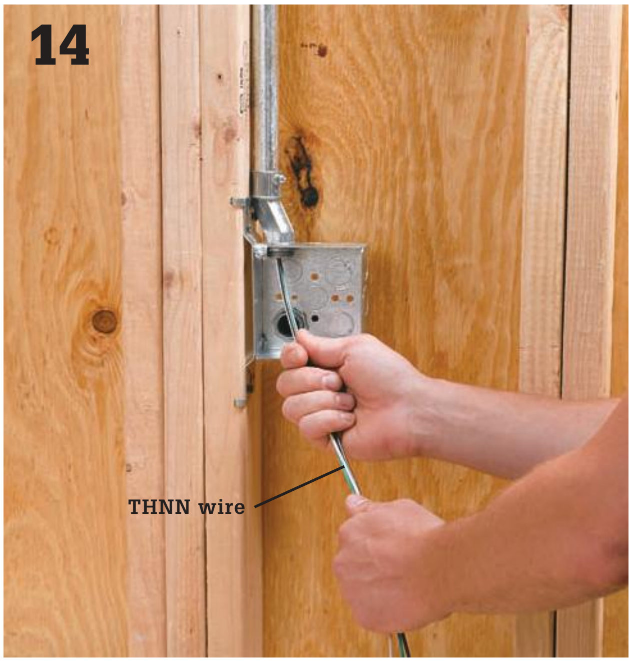

Measure to find how much wire you'll need for each run, and cut pieces that are a foot or two longer. Before making L-turns with the conduit, feed the wire through the first conduit run.

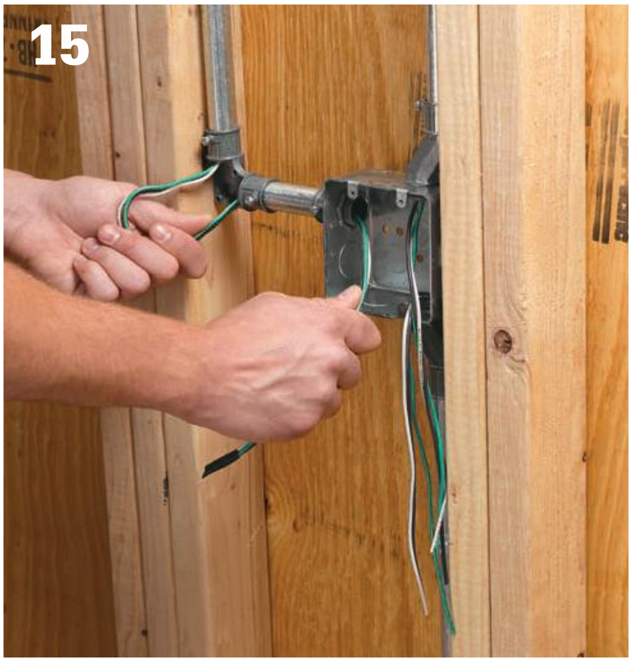

Feed the other ends of the wires into the next box or fitting in line. It is much easier to feed wire into 45° and 90° elbows if they have not been attached to the conduit yet. Continue feeding wire into the conduit and fitting until you have reached the next box in line.

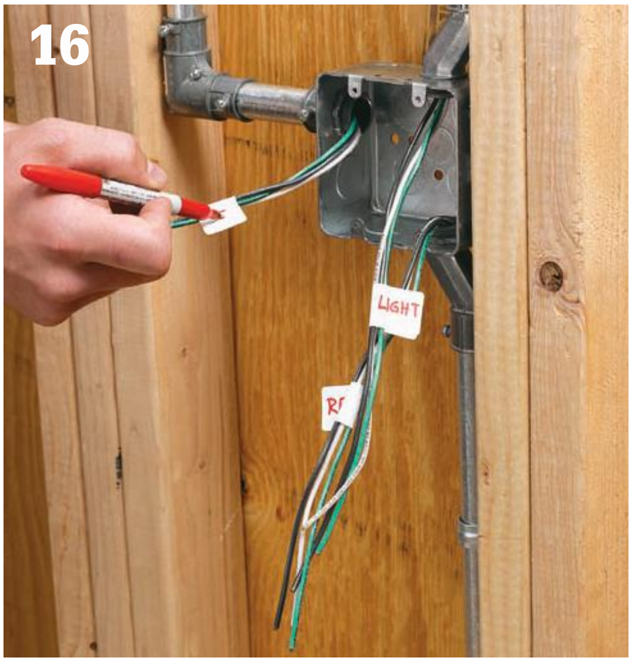

Once you've reached the next box in line, coil the ends of the wires and repeat the process with new wire for the next run. Keep working until all of the wire is run and all of the conduit and fittings are installed and secured. If you are running multiple feed wires into a single box, write the origin or destination on a piece of masking tape and stick it to each wire end.

(continued)

1st proofs for reference only

Edition
age: 90
41940 - Advanced Home Wiring_064-128.indd 91
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 91

17

Note: Your Code may require an in-use rated receptacle box cover.

Make the wiring connections at the receptacles. Strip ¾" of insulation from the circuit wires using a wire stripper. Connect the white (neutral) wire and black (hot) wire of the UF cable to the LINE screw terminals on the receptacle. Connect the white (neutral) and black (hot) wires from the NM cable to the LOAD terminals. Pigtail the bare copper ground wires and connect them to the receptacle ground terminal and the metal box. Install the receptacle and cover plate.

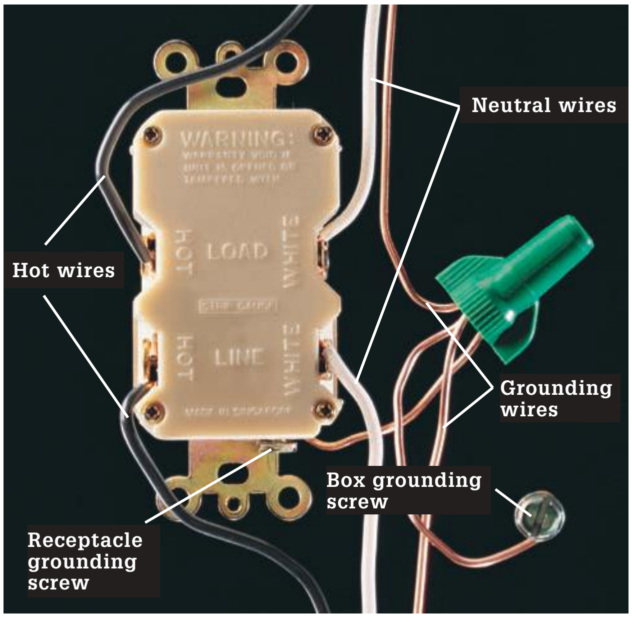

Variation: Installing a GFCI-protected breaker for the new circuit at the main service panel is the best way to protect the circuit and allows you to use regular receptacles in the building, but an alternative that is allowed in many areas is to run the service into a GFCI-protected receptacle and then wire the other devices on the circuit in series. If you use this approach, only the initial receptacle needs to be a GFCI receptacle.

18

Continue installing receptacles in the circuit run, and then run service from the last receptacle to the switch box for the light fixture or fixtures. (If you anticipate a lot of load on the circuit, you should probably run a separate circuit for the lights). Twist the white neutral leads and grounding leads together and cap them. Attach the black wires to the appropriate switches. Install the switches and cover plate.

19

Install the light fixtures. For this shed, we installed a caged ceiling light inside the shed and a motion-detector security light on the exterior side.

1st
proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 92
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 12:05 AM
Page: 92
Pro-Vis

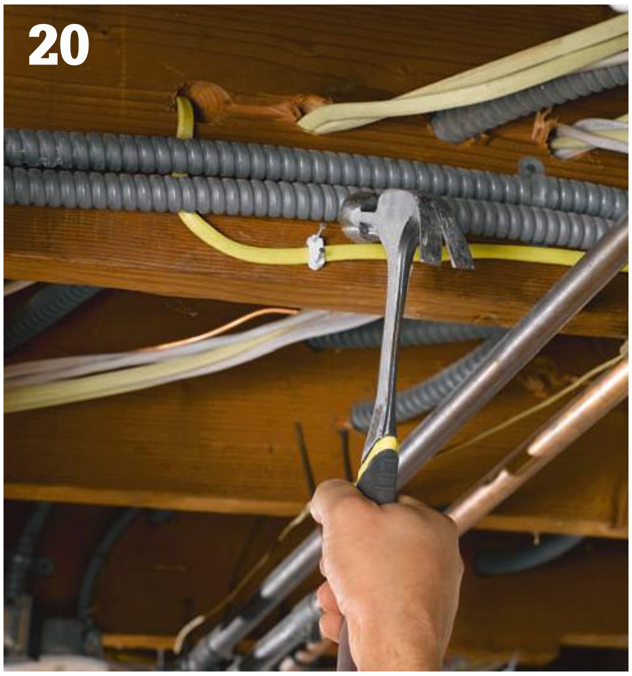

Run NM cable from the electrical box in the house at the start of the new circuit to the main service panel. Use cable staples if you are running the cable in floor joist cavities. If the cable is mounted to the bottom of the floor joists or will be exposed, run it through conduit.

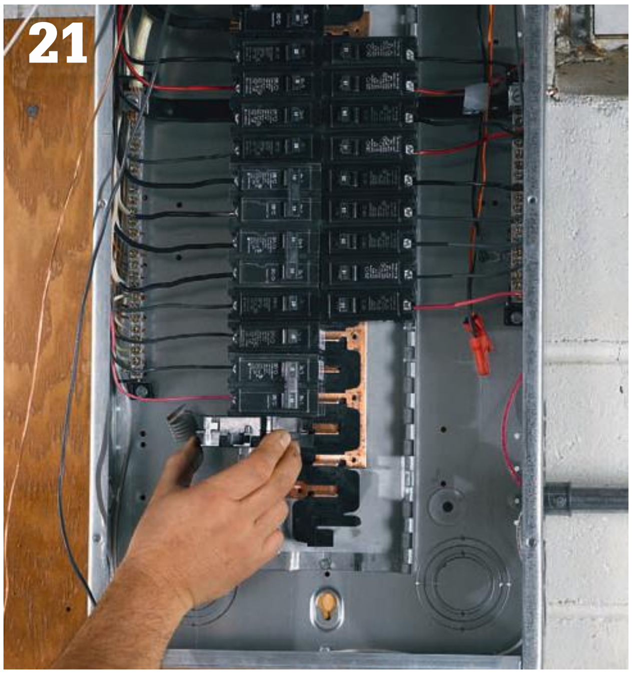

At the service panel, feed the NM cable in through a cable clamp. Arrange for your final electrical inspection before you install the breaker. Then attach the wires to a new circuit breaker, and install the breaker in an empty slot. Label the new circuit on the circuit map.

Turn on the new circuit, and test all of the receptacles and fixtures. Depress the Test button and then the Reset button if you installed a GFCI receptacle. If any of the fixtures or receptacles is not getting power, check the connections first, and then test the receptacle or switch for continuity with a multimeter. Backfill the trench.

Standalone Solar Lighting System

A self-contained electrical circuit with dedicated loads, usually 12-volt light fixtures, is one of the most useful solar amenities you can install. A standalone system is not tied into your power grid, which greatly reduces the danger of installing the components yourself. Plus, the fact that your light fixtures are independent of the main power source means that even during a power outage you will have functioning emergency and security lights.

Installing a single solar-powered circuit is relatively simple, but don't take the dangers for granted. Your work will require permits and inspections in most jurisdictions, and you can't expect to pass if the work is not done to the exact specifications required.

Solar panels that convert the sun's energy into electricity are called photovoltaic (PV) panels, and they produce direct current (DC) power. PV solar panel systems can be small and designed to accomplish a specific task, or they can be large enough to provide power or supplementary power to an entire house. Before you make the leap into a large system, it's a good idea to familiarize yourself with the mechanics of solar power. The small system demonstrated in this project is relatively simple and is a great first step into the world of solar. The fact that the collector, battery, and lights are a standalone system makes this a very easy project to accomplish. By contrast, installing panels that provide direct supplementary power through your main electrical service panel is a difficult wiring job that should be done by professional electricians only.

This 60-watt solar panel is mounted on a garage roof and powers a self-contained home security lighting system. Not only does this save energy costs, it keeps the security lights working even during power outages.

1st proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 94
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 94
Pro-Vis

Schematic Diagram for an Off-the-Grid Solar Lighting System

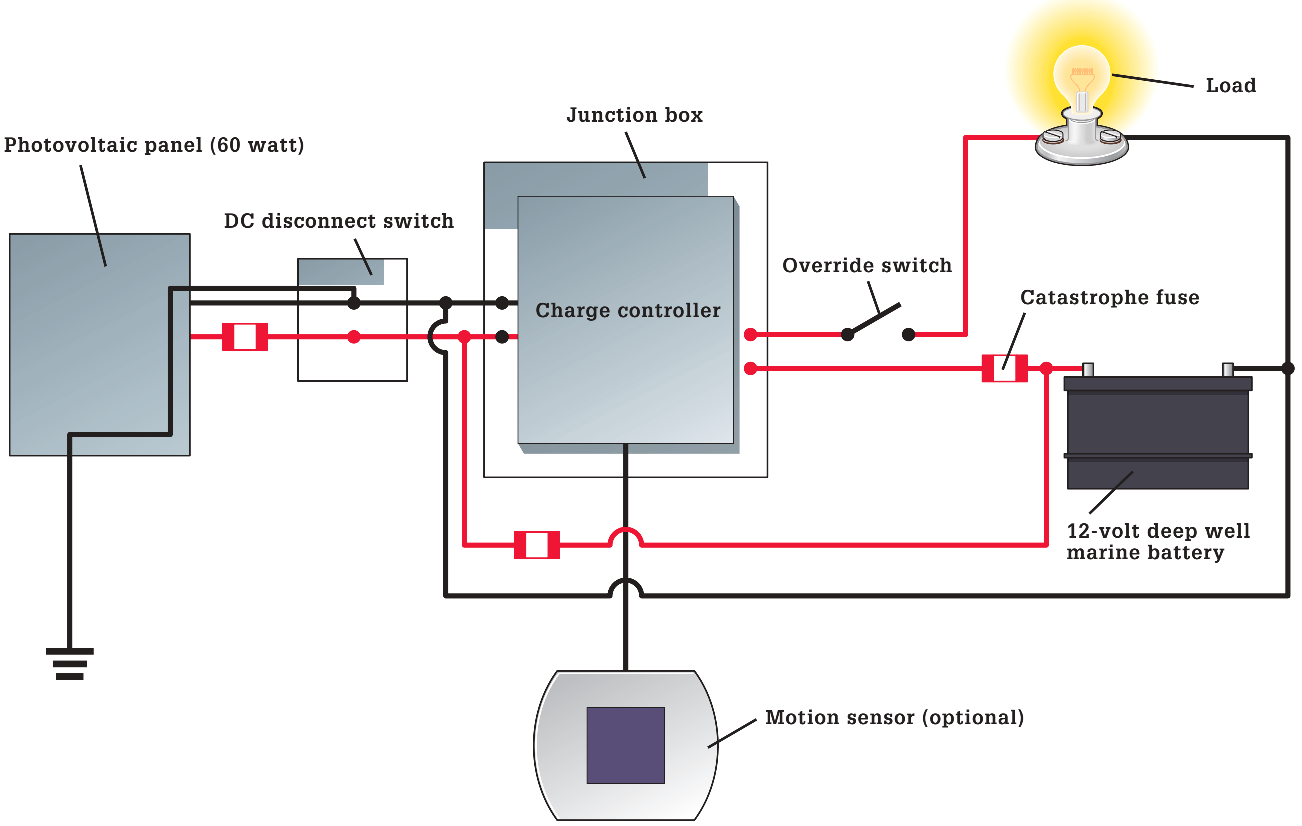
Load
Junction box
Photovoltaic panel (60 watt)
DC disconnect switch
Charge controller
Override switch
Catastrophe fuse
12-volt deep well marine battery
Motion sensor (optional)

Tools & Materials

Tape measure
Drill/driver with bits
Caulk gun
Crimping tool
Wiring tools
Metal-cutting saw
Photovoltaic panel (50 to 80 watts)
Charge controller
Catastrophe fuse
Battery sized for 3 day autonomy
Battery case
Battery cables
12-volt LED lights including motion-sensor light
Additional 12-volt light fixtures as desired
20 ft. Unistrut 1⅞" thick U-channel (See Resources, page 125)
(4) 45° Unistrut connectors
(2) 90° Unistrut angle brackets
(4) Unistrut hold down clamps
(12) ⅜" spring nuts
(12) ⅜"-dia. × 1"-long hex-head bolts with washers
DC-rated disconnect or double throw snap switch
6" length of ½"-dia. liquid-tight flexible metallic conduit
(2) ½" liquid tight connectors
(2) Lay-in grounding lugs
(2) Insulated terminal bars to accept one 2-gauge wire and 4 12-gauge wires
(2) Cord cap connectors for ½"-dia. cable
½" ground rod and clamp
Copper wire (6, 12-gauge)
Green ground screws
½" Flexible metallic conduit or Greenfield
½" Greenfield connectors
(4) 11⁄16" junction boxes with covers
(4) square boxes with covers
PVC 6"× 6" junction box with cover
14/2 UF wire
¼" × 20 nuts and bolts with lock washers
Roof flashing boot
Roof cement
Silicon caulk
Eye protection

1st proofs for reference only
Edition
age: 94
41940 - Advanced Home Wiring_064-128.indd 95
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 95

Mounting PV Panels

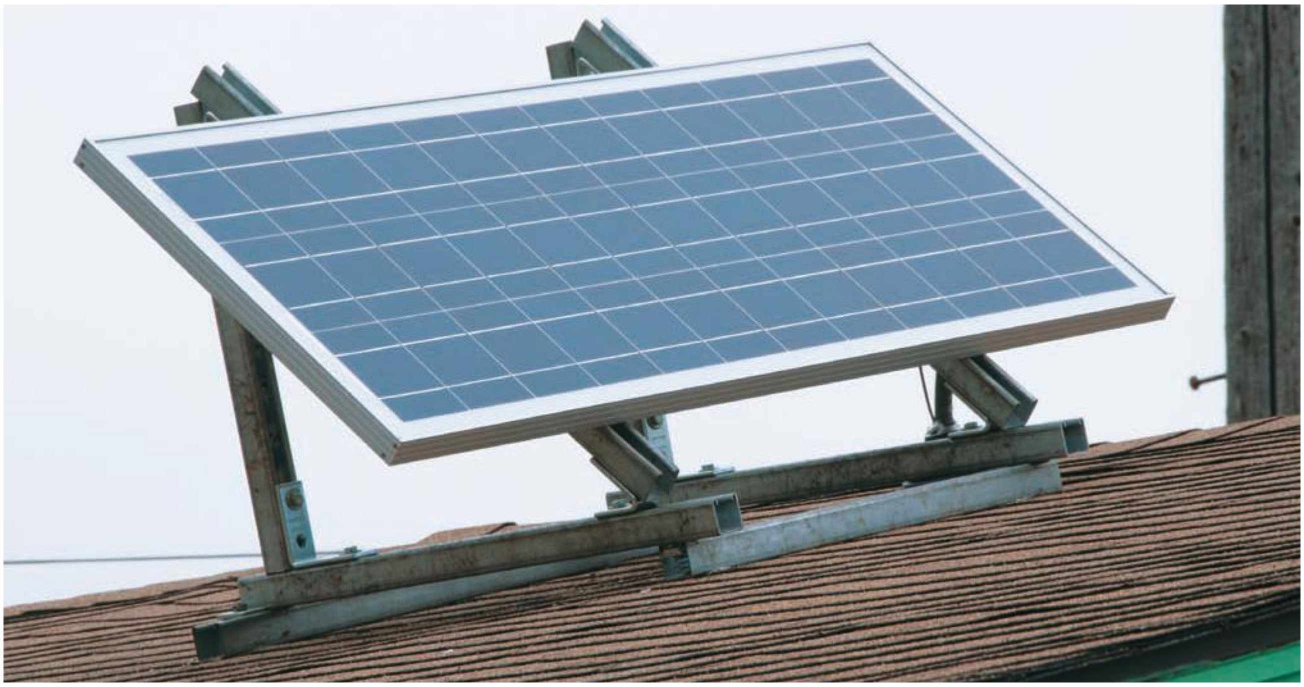

The mounting stand for the PV panel is constructed from metal U-channel (a product called Unistrut is seen here. See Resources page 125) and pre-bent fasteners. Position the solar panel where it will receive the greatest amount of sunlight for the longest period of time each day—typically the south-facing side of a roof or wall. For a circuit with a battery reserve that powers two to four 12-volt lights, a collection panel rated between 40 and 80 watts of output should suffice. These panels can range from $200 to $600 in price, depending on the output and the overall quality.

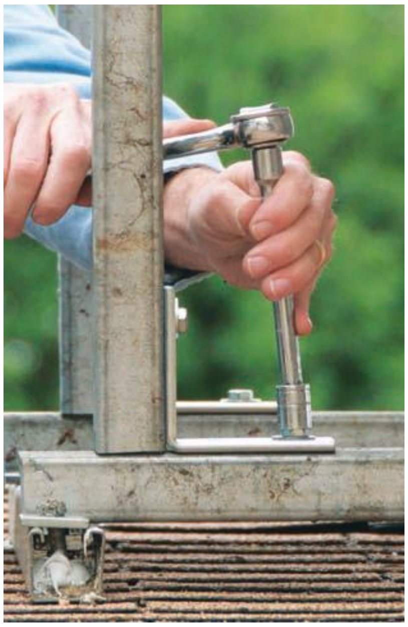

The stand components are held together with bolts and spring-loaded fasteners. The 45° and 90° connectors are manufactured specifically for use with this Unistrut system.

Connections for the feed wires that carry current from the collector are made inside an electrical box mounted on the back of the collector panel.

An EPDM rubber boot seals off the opening where the PVC conduit carrying the feed wires penetrates the roof.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 96
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 96
Pro-Vis

How to Wire a DC Lighting Circuit

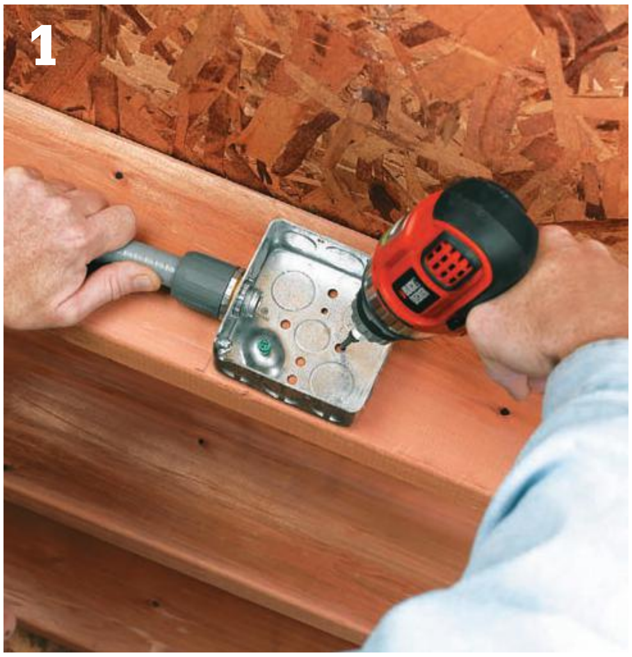

Mount a junction box inside the building where the conduit and wiring enter from the power source. Secure the box to the conduit with appropriate connectors. Run two #14 awg wires through the conduit and connect them to the positive and negative terminals on the panel (see previous page).

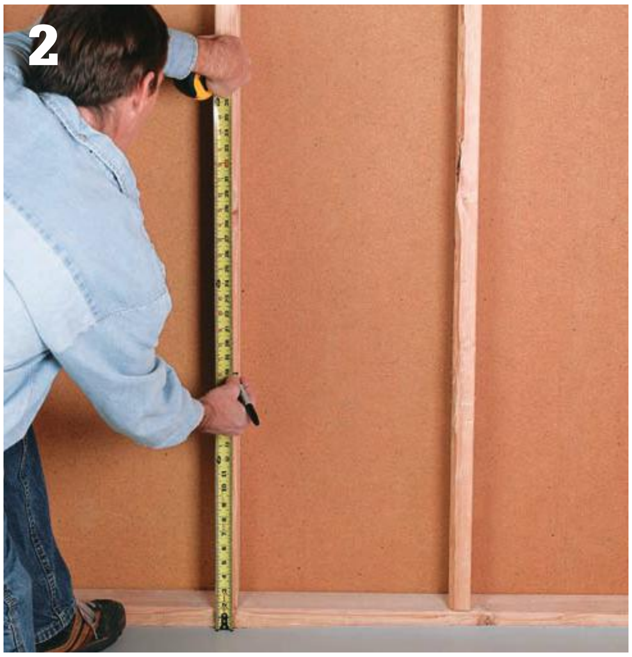

Plan the system layout. Determine the placement of the battery, and then decide where you will position the charge controller and DC disconnect. The battery should be placed at least 18" off the floor, in a well-ventilated area where it won't be agitated by everyday activity. Mark locations directly on the wall.

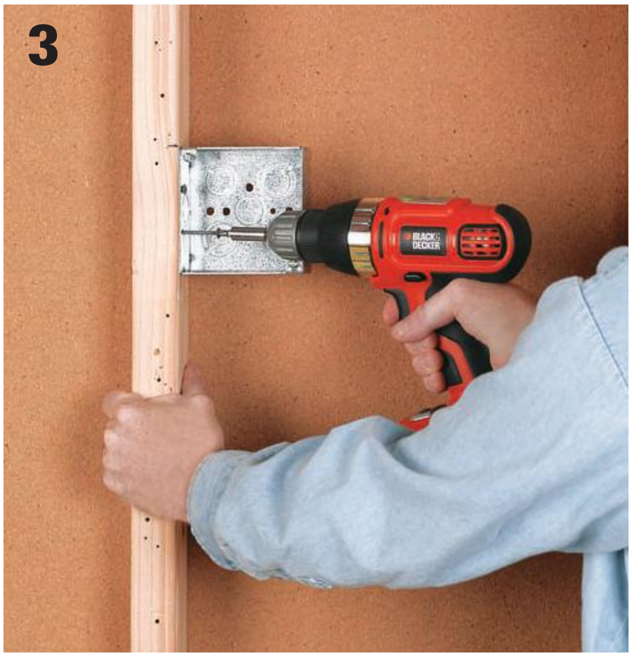

Attach a junction box for enclosing the DC disconnect, which is a heavy-duty switch, to a wall stud near the battery and charge controller location. Use a metal single-gang box with mounting flanges.

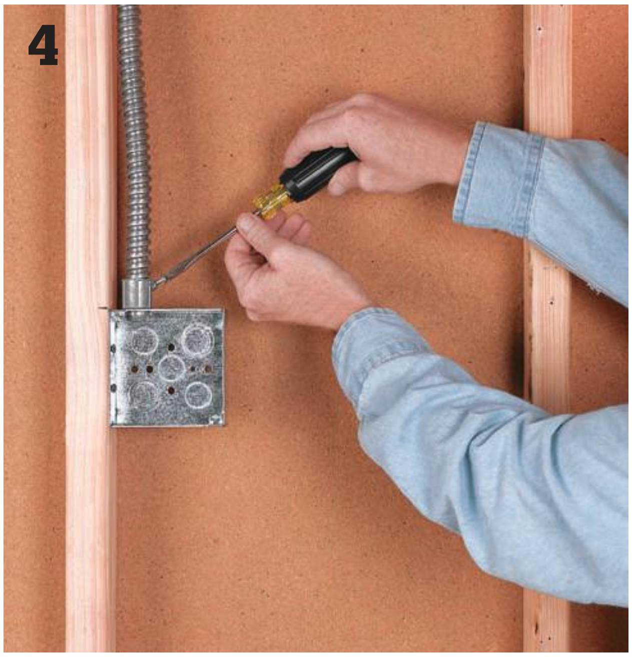

Run flexible metal conduit from the entry point at the power source to the junction box for the DC disconnect box. Use hangers rated for flexible conduit.

(continued)

1st
proofs for reference only

Edition
age: 96
41940 - Advanced Home Wiring_064-128.indd 97
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 12:05 AM
Page: 97

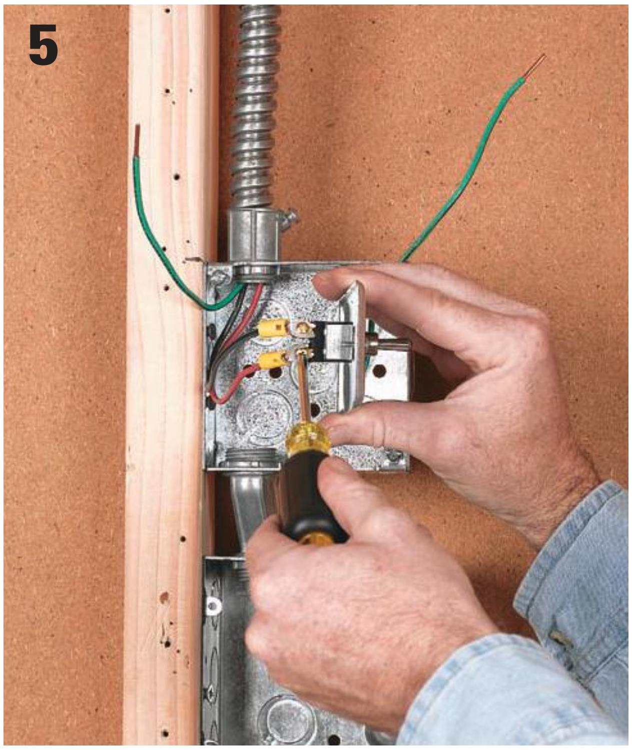

5

Attach the DC disconnect switch to the wire leads from the power source.

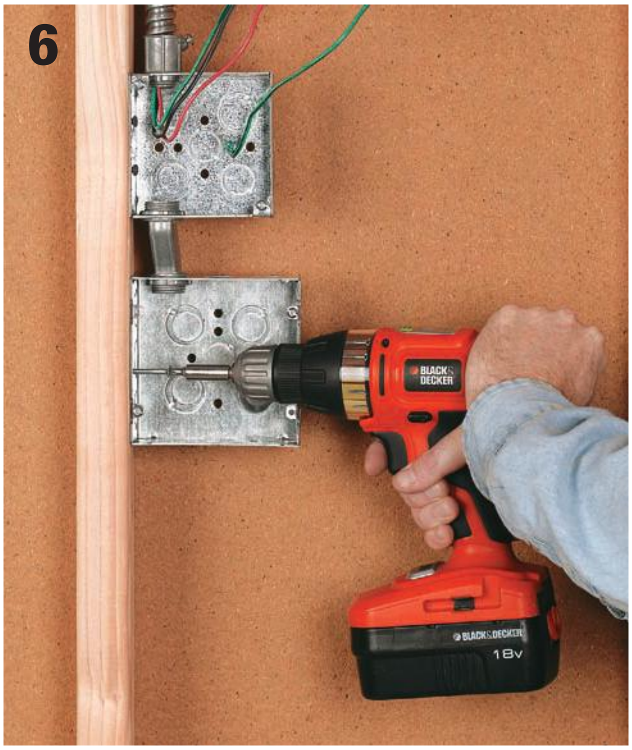

6

Attach a double gang metal junction box to the building's frame beneath the DC disconnect box to enclose the charge controller.

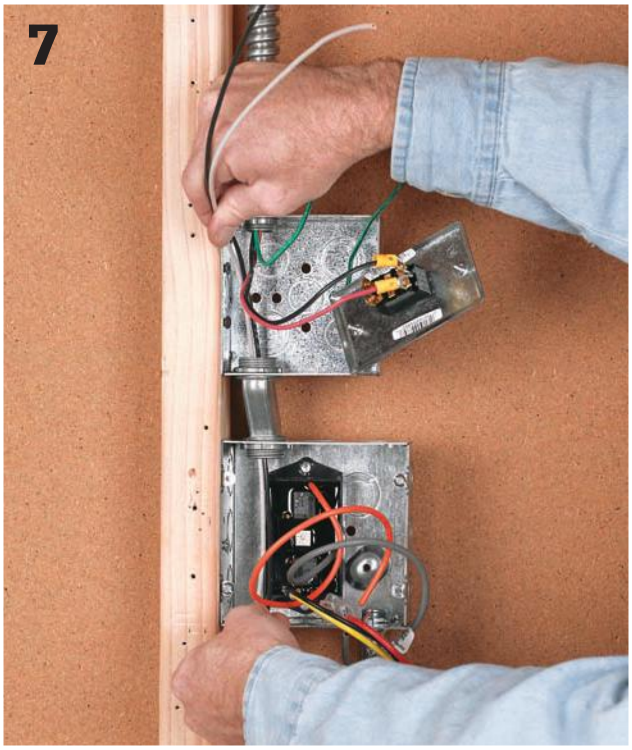

7

Install the charge controller inside the box. Run flexible conduit with connectors and conductors from the disconnect box and to the charge controller box.

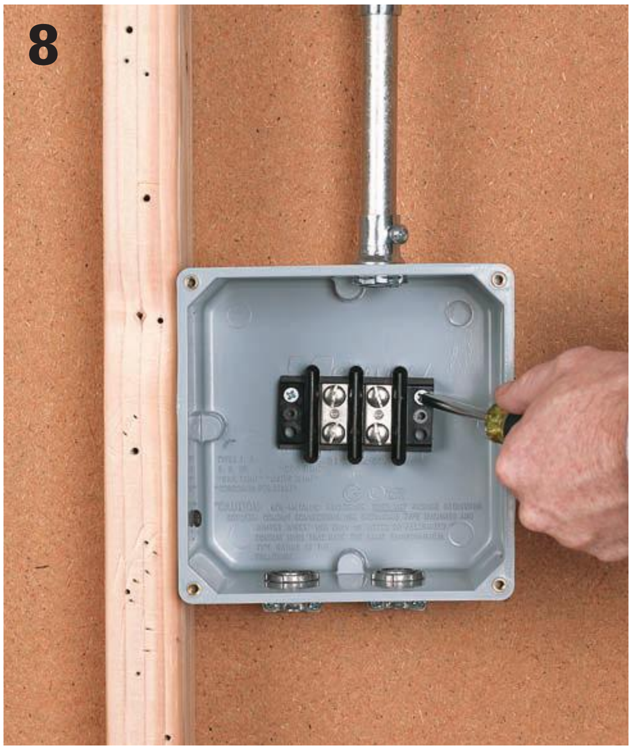

8

Mount a PVC junction box for the battery controller about 2 ft. above the battery location, and install two insulated terminal bars within the box.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 98
10/10/14 12:05 AM
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 98
Pro-Vis

Build a support shelf for the battery using 2 × 4s. The shelf should be at least 18" above ground. Set the battery on the shelf in a sturdy plastic case.

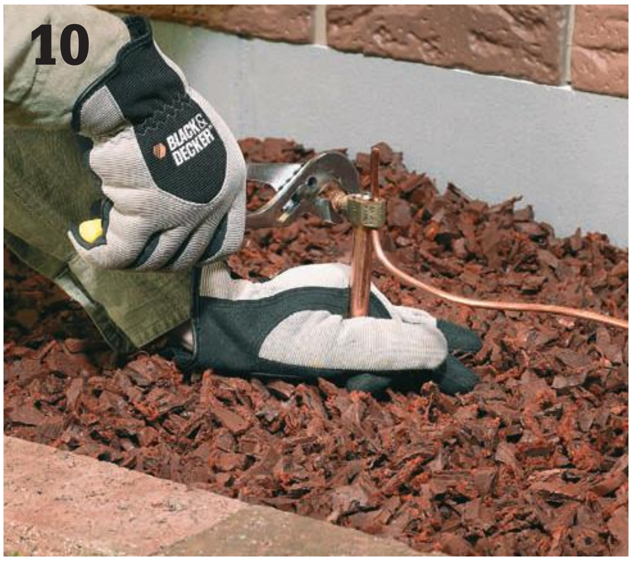

Set up grounding protection. Pound an 8-ft. long, ½"-dia. ground rod into the ground outside the building, about 1 ft. from the wall on the opposite side of the charge controller. Leave about 2" of the rod sticking out of the ground. Attach a ground rod clamp to the top of the rod. Drill a 5⁄16" hole through the garage wall (underneath a shake or siding piece) and run the #6-gauge THWN wire to the ground rod. This ground will facilitate lightning protection. See pages 58 to 63 for more information on grounding the system.

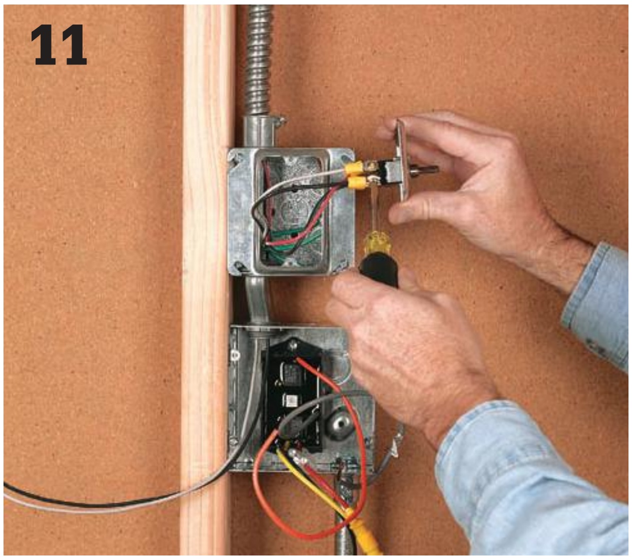

Wire the DC disconnect. Attach the two #14-gauge wires to the two terminals labeled "line" on the top of the DC disconnect switch.

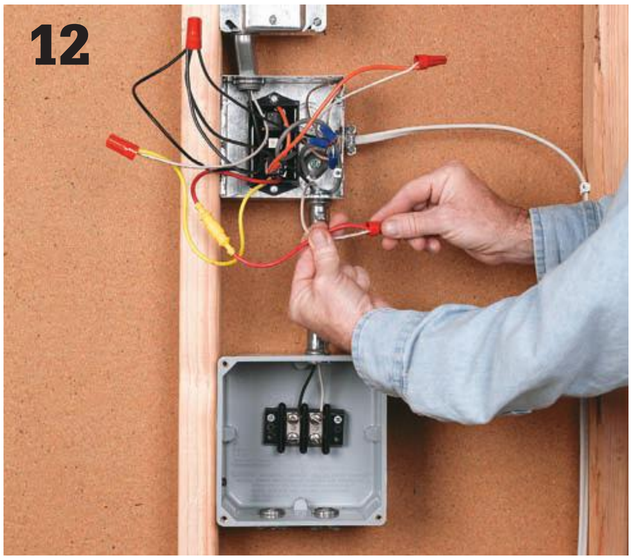

Wire the charge controller. Route two more #14-gauge wires from the bottom of the DC disconnect terminals into the 4" × 11⁄16" junction box and connect to the "Solar Panel In" terminals on the charge controller. The black wire should connect to the negative terminal in the PVC box and the red to the positive lead on the charge controller. Finish wiring of the charge controller according to the line diagram provided with the type of controller purchased. Generally the load wires connect to the orange lead, and the red wire gets tied to the battery through a fuse.

(continued)

1st proofs for reference only

OPTION: Attach a motion sensor. Some charge controllers come equipped with a motion sensor to maximize the efficiency of your lighting system—these are especially effective when used with security lighting. The motion sensor is typically mounted to a bell box outside and wired directly to the charge controller with an 18-gauge × 3-conductor insulated cable. A system like this can support up to three motion sensors. Follow the manufacturer's directions for installing and wiring the motion sensor.

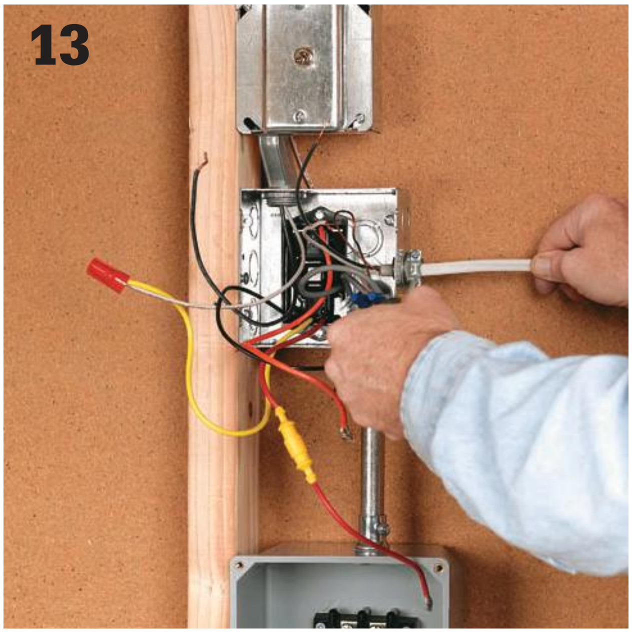

Run wiring to the loads (exterior DC lighting fixtures in this case) from the charge controller. DC light fixtures (12-volt) with LED bulbs can be purchased at marine and RV stores if you can't find them in your home center or electrical supply store.

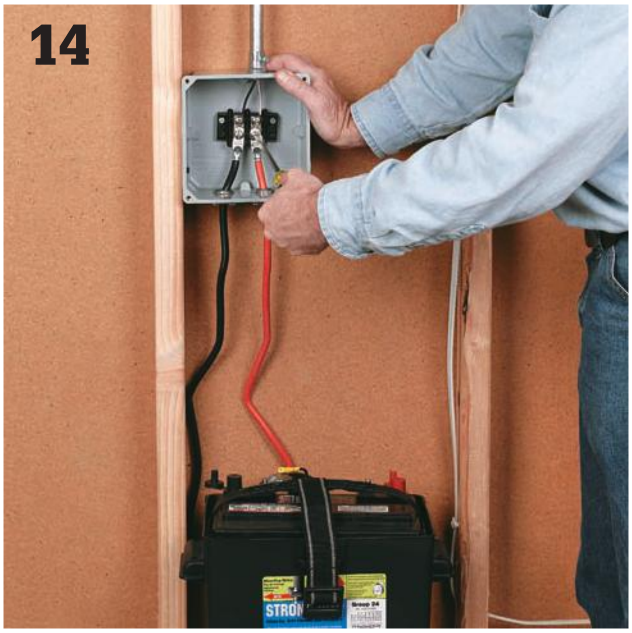

Install the battery. Here, a deep-cell 12-volt marine battery is used. First, cut and strip each of the two battery cables at one end and install into the battery control junction box through cord cap connectors. Terminate these wires on two separate, firmly mounted insulated terminal blocks.

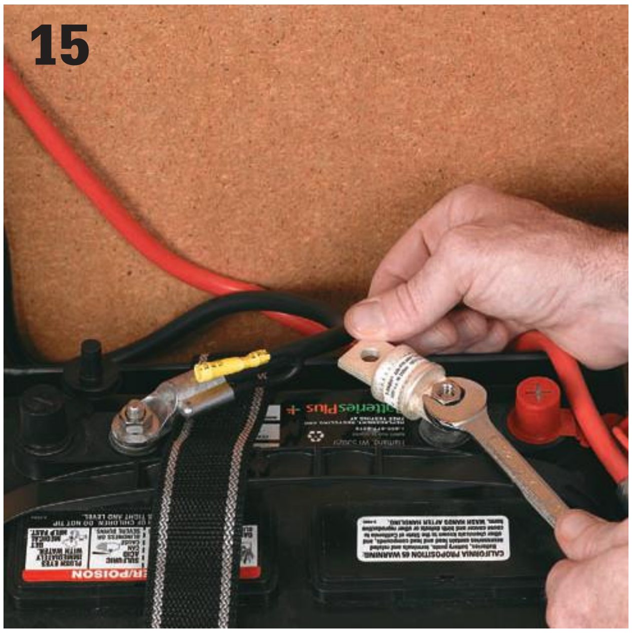

Install the catastrophe fuse onto the positive terminal using nuts and bolts provided with the battery cables. Connect the battery cables to the battery while paying close attention to the polarity (red to positive and black to negative). Make sure all connections have been made and double checked.

1st
proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 100
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 100
Pro-Vis

Troubleshooting & Repairs

Running new circuits and hooking up new fixtures are fairly predictable projects when it comes to estimating time and expense. This is less true with repairing problems in your system and fixtures. In some cases, a repair is as simple as opening an electrical box, spotting a loose wire connector and remaking the connection. But there are also times when fixing a dead circuit or device is a highly frustrating proposition. Such cases are almost always caused by tricky diagnostic challenges. Wires are hidden behind walls and there very often are no visual clues to system breakdowns. So essentially, minimizing repair frustration boils down to learning to deploy logical, systematic diagnostics. Educated troubleshooting, you could say.

In this project you'll learn how to use the most important diagnostic tool in any electrician's toolkit: the multimeter. These handy devices come in a dizzying array of types and qualities, but for diagnostic purposes they are used to take readings for current (amperage), voltage and continuity (whether an electrical path is open or closed). Once you learn the basics of operating a multimeter, you can enlist it in a logical, deductive manner to track down the source of a wiring problem. Once located, correcting the problem is usually very simple.

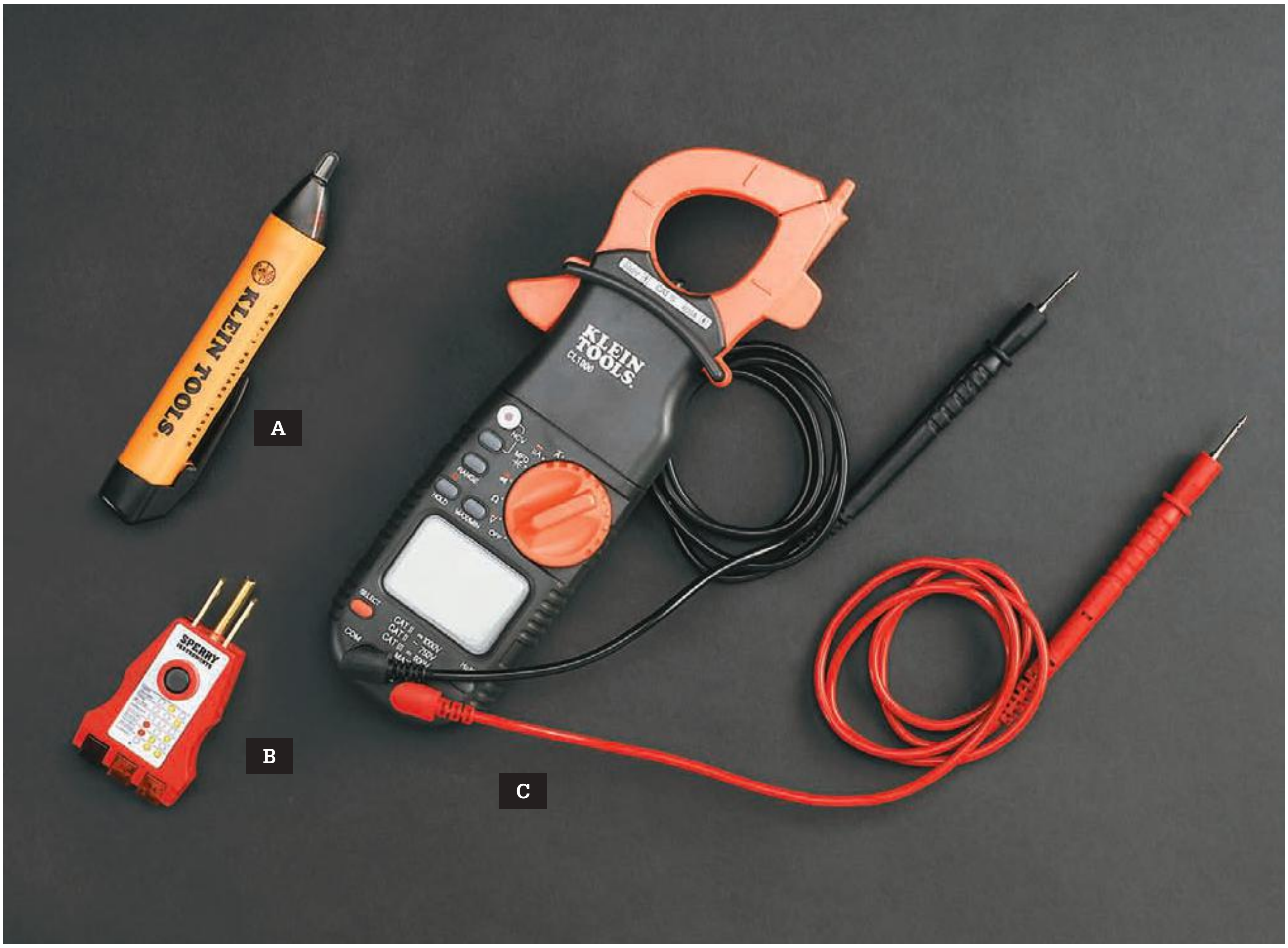

Diagnostic tools for home wiring use include: Touchless circuit tester (A) to safely check wires for current and confirm that circuits are dead; Plug-in tester (B) to check receptacles for correct polarity, grounding and circuit protection; Multimeter (C) to measure AC/DC voltage, AC/DC current, resistance, capacitance, frequency and duty cycle (model shown is an auto-ranging digital multimeter with clamp-on jaws that measure through sheathing and wire insulation).

1st proofs for reference only

Multimeters

Multimeters are nearly indispensible diagnostic tools for doing intermediate to advanced level electrical work (as well as automotive and electronics repair). They are used to measure voltage, current (amperage) and a few other conditions such as continuity, capacitance and frequency. For your home electrical system, by far the most used feature of a multimeter is testing voltage and current, although there are occasions where testing for resistance is needed. Among professional electricians, the most common and widely used multimeters have a clamp-on ammeter that measures current through the wire insulation so you don't have to disconnect the circuit and expose bare wire. Most clamp-on multimeters also are fitted with insertible probes with which you can measure voltage and continuity in the traditional way. An example of a clamp-on multimeter can be seen on the next page. Among homeowners, however, the most common multimeters these days are digital, auto-ranging tools that use probes or alligator clamps at the ends of wire leads for diagnostic work. Older multimeters that do not have autoranging capability must be pre-set to estimated calibration levels before use. Non-digital multimeters or ammeters usually have a dial gauge that gives readouts. These tools are somewhat more difficult to use and are less precise. Considering that digital, autoranging multimeters can be found for just a few dollars (the top of the line models cost over $100) there is really no good reason not to replace your old device with one that resembles the tools seen on these pages.

Time to Replace that Neon Tester ▸

Neon circuit testers are inexpensive and easy to use (if the light glows the circuit is hot), but they are less sensitive than multimeters and can be unsafe. In some cases, neon testers won't detect the presence of lower voltage in a circuit. This can lead you to believe that a circuit is shut off when it is not—a dangerous mistake. The small probes on a neon circuit tester also force you to get too close to live terminals and wires. For the most reliable readings, buy and learn to use a multimeter. At the very least, switch to a touchless tester like the one on page 107, Step 1.

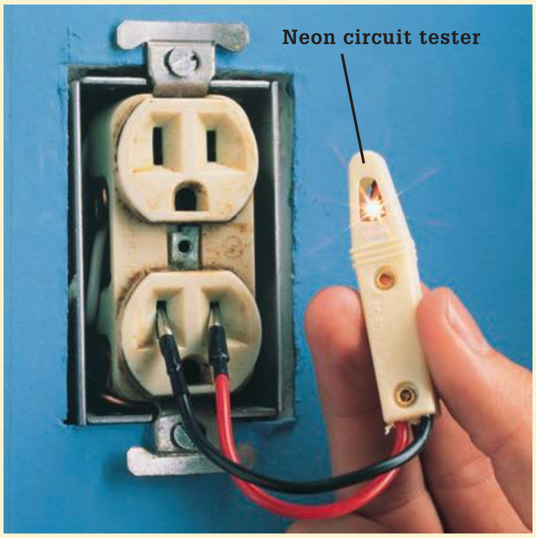
Neon circuit tester

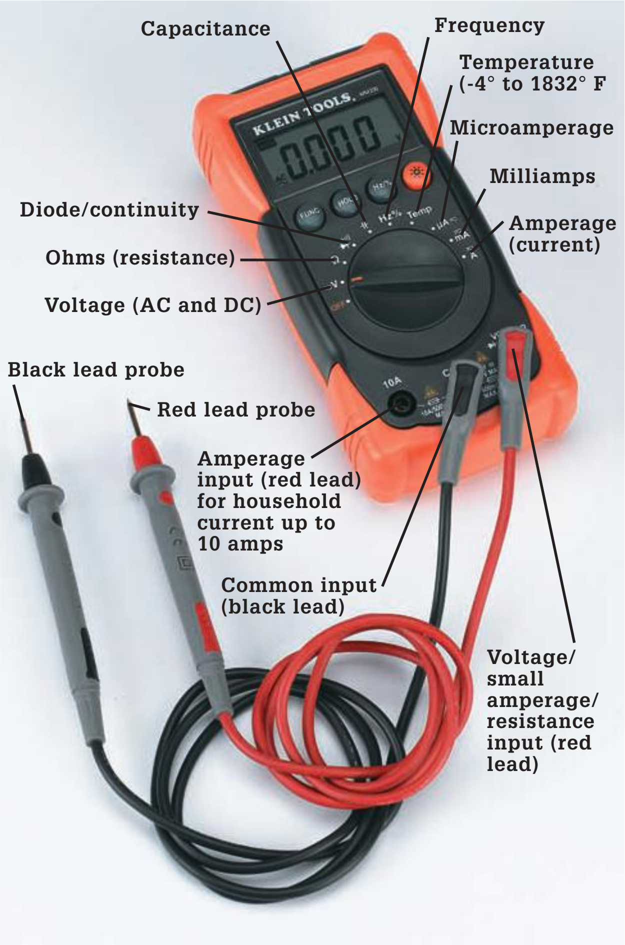
Capacitance
Frequency
Temperature (-4° to 1832° F
Microamperage
Milliamps
Amperage (current)
Diode/continuity
Ohms (resistance)
Voltage (AC and DC)
Black lead probe
Red lead probe
Amperage input (red lead) for household current up to 10 amps
Common input (black lead)
Voltage/ small amperage/ resistance input (red lead)
KLEIN TOOLS

A digital, autoranging multimeter must be adjusted to the proper setting for the reading you want to take. The probe leads also must be inserted into the correct inlet at the bottom of the tool. Inserting the red lead into the incorrect inlet can cause the tool to trip an internal fuse. Study your owner's manual carefully before using any tool.

How to Measure Current

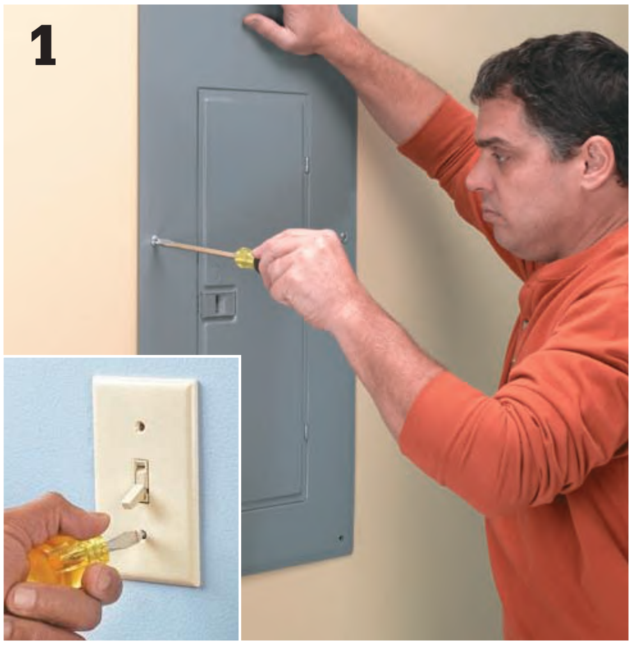

Create access to the wires you need to test. In most cases this requires that you remove the cover to an electrical service panel or an electrical box (inset).

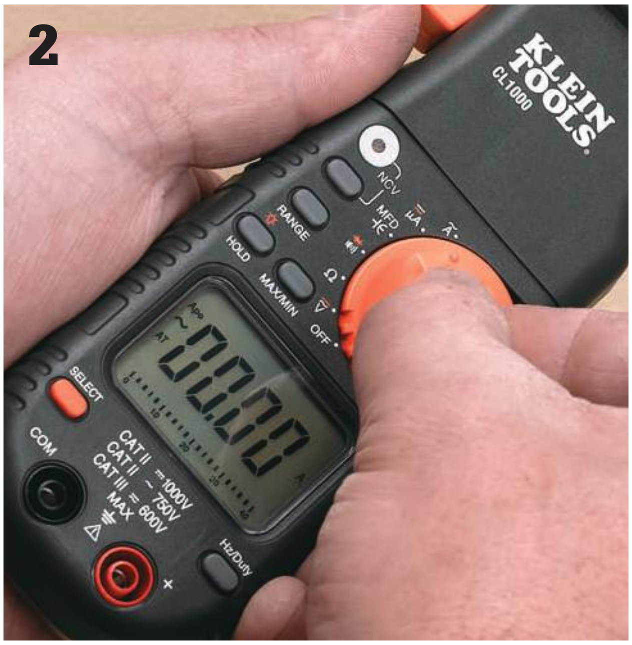

Set the multimeter to test for amperage (current is measured in amperes or amps). On some multimeters you need to select between amperage settings that are above or below 40 amps. Use the rated amperage of the circuit as a guide (amperage is printed on the circuit breaker switch).

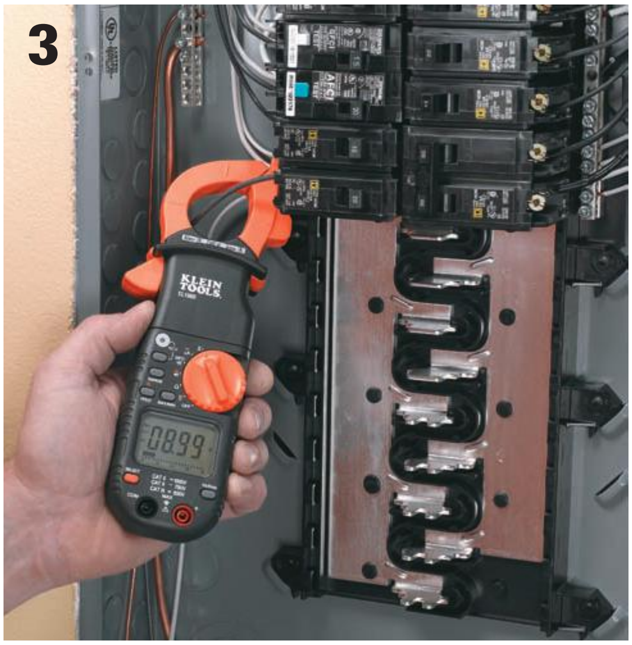

Clamp the jaws of a clamp-on multimeter onto the conductor or one of the conductors (if more than one) leading to a circuit breaker. If you are using a non-clamping multimeter, touch one probe to the screw terminal where the hot lead is attached to the breaker and touch the other probe to the metal panel box. The readout on your meter is the amount of current flowing in that circuit.

Taking Measurements at a Receptacle ▸

You may use a multimeter to measure for voltage at a wall receptacle. Regardless of whether the outlet is in service, if it is live you will get a voltage reading in the approximate range of the receptacle rating—here, 120 volts. To detect live current, measured in amps, the receptacle must be in use, with an appliance drawing from it. Taking an amperage reading in such an instance will only yield the amount of current being drawn, which is a factor of the appliance, not the circuit capacity.

1st proofs for reference only

Edition
ge: 102
41940 - Advanced Home Wiring_064-128.indd 103
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 103

How to Measure Voltage

1
KLEIN TOOLS

To measure voltage using the multimeter, you will have to use the two probes provided with the multimeter and have access to a live terminal or slot as well as a grounded terminal or slot. If your meter has probe holders at the top, snap the probes into them. They are like extra hands.

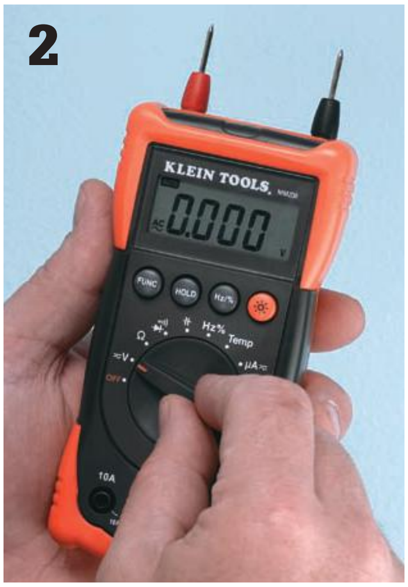

Turn the multimeter to the VAC setting to measure AC voltage that is found in your house. Set the multimeter to VDC if measuring DC voltage, such as in a car or a battery-fed device. On some multimeters, like the one above, you select "V" for voltage then change between AC and DC with the"FUNC" button.

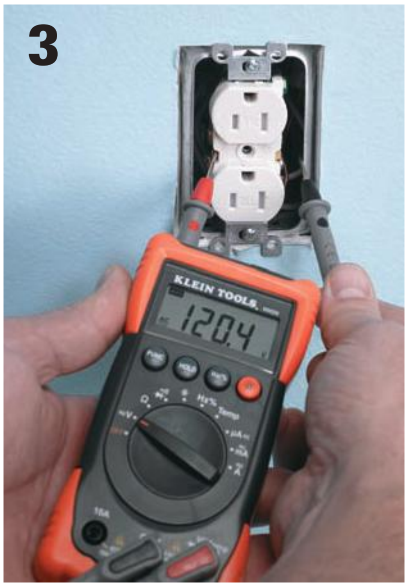

To measure the AC voltage, place one probe on a grounded surface, such as the metallic junction box or the bare ground wire. Place the other probe on the hot screw terminal or into the receptacle slot associated with the hot wire. The voltage readout should be in the range of 120 volts, plus or minus 5 volts (usually 120 volts in a residence in the US).

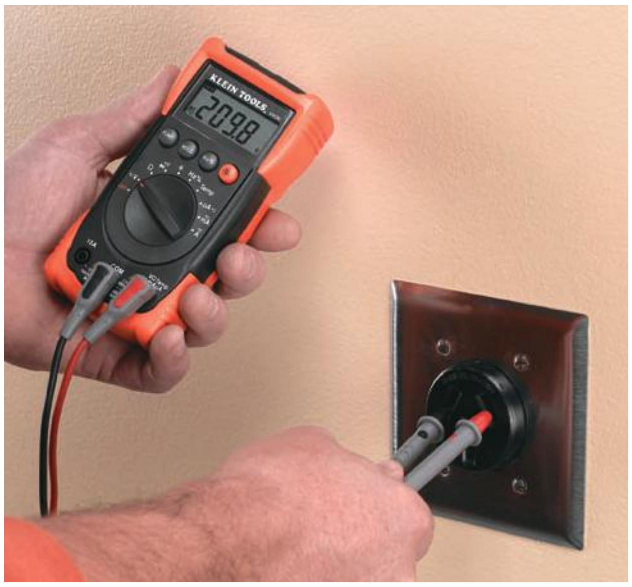

240 VOLTS. You can also measure voltage across the two hot leads to determine if you have 240 volts. This can be done at your range receptacle, dryer receptacle, or any other 240-volt receptacle. Place one probe in one of the small slots and the other probe in the other slot directly across from it. The voltage should read 240 volts, plus or minus 5 volts.

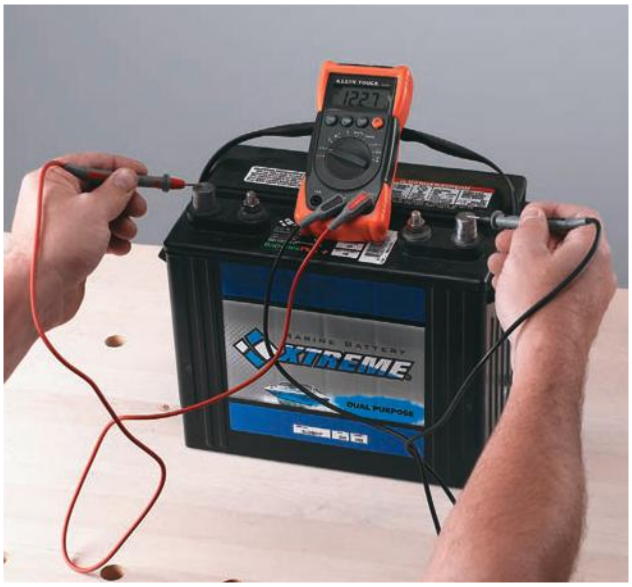

DC Voltage. When testing DC voltage, such as in a car battery, you can measure exactly the same way as for AC as long as the meter is set to the DC function. For more accurate results, test the voltage while the battery is in use.

1st
proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 104
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 104
Pro-Vis

How to Test for Continuity

Continuity is a condition in a circuit where the conductors form an unbroken pathway through which current may flow. When measuring for continuity, always make sure there is no power present on the circuit you are testing or damage may occur to the meter. You can also measure the resistance in this mode as well.

1

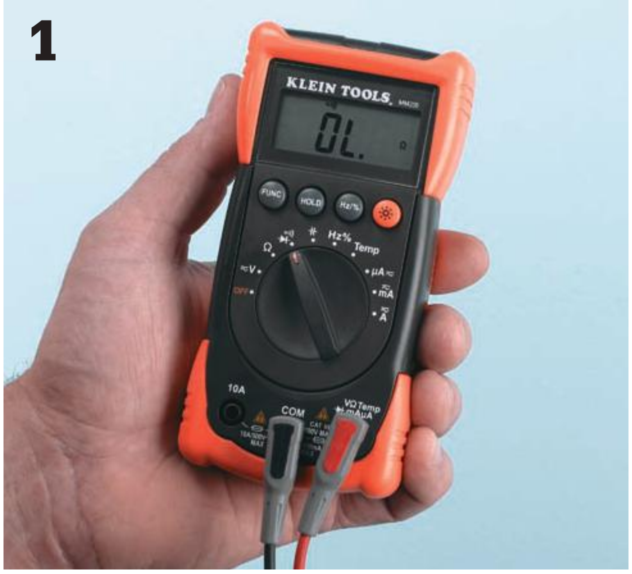

The setting for continuity is an “audible” or diode symbol display on the dial. Select this setting.

2

KLEIN TOOLS
0000
10A
COM

Verify that the continuity tester is functional by touching the two probes together. You should hear an alert sound and/or see a reading of zero ohms (Ohms is a value of the resistance to current flow).

3

KLEIN TOOLS
0000

To test a circuit, touch one probe to one of the wires on a given circuit and the other to the second wire of the circuit. If you hear an audible sound or read a value of resistance other than zero, you have a complete or unbroken path for current to flow.

How to Test a 3-way Switch ▸

Remove the switch from the circuit and place one of the probes onto the common terminal and the other probe onto one of the other two terminals used for the traveler wires. If the meter indicates infinity ohms or there is no sound, flip the switch and if it is in working order the meter should read zero ohms or emit an audible sound. It should only work in one direction or the other, not both.

Troubleshooting an Open Neutral

An open neutral is an electrical problem where the circuit is broken on the return path wire or neutral (white wire). When this situation occurs anything plugged into or connected to this circuit can experience low or high voltage, which could damage voltage sensitive electronics such as a computer or flat screen TV. The lights will be dim or not work at all, depending upon where the problem lies within the circuit.

Possible Symptoms of an Open Neutral ▸

1. When a whole circuit does not work and the breaker associated with that circuit is operating normally.
2. When the neutral or white wire registers as a hot wire by using a non contact voltage tester when the circuit breaker is on. This most likely indicates a problem between the main service panel and the utility transformer. The condition should be readable on other receptacles as well.
3. If you register a voltage lower than 110 volts between the hot and neutral.
4. When the incandescent lights work, but are very dim.
5. When the fluorescent lights are barely lit and are flickering.
6. Discoloration of the wires or exposed copper turning green under the wirenut holding the neutral wires together.

1

KLEIN TOOLS
0899

2

3

KLEIN TOOLS
092.4

6

1st proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 106
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 106

How to Troubleshoot an Open Neutral

1
KLEIN TOOLS

Verify which lights and receptacles are on the circuit by turning the breaker off and by checking for power with a non-contact voltage tester. It is helpful to either draw a map of the house or place some tape on every affected opening.

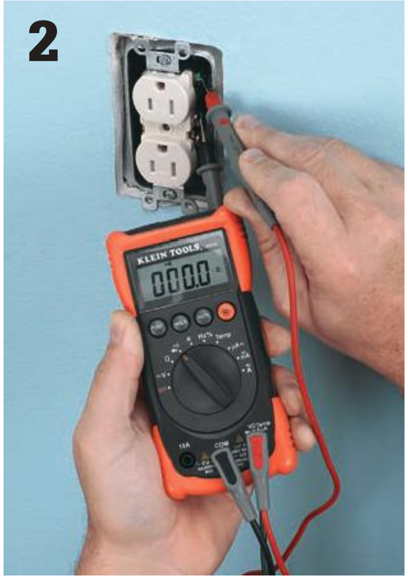

Start at the outlet nearest to the panel. With the breaker off and using a multimeter, check for continuity between the neutral (white wire) and the ground (bare or green wire). These two wires land at the same point electrically in your electrical panel. If there is an indication of continuity between these two wires, the neutral and ground connections are sound and you should proceed to the next outlet as you move away from the panel.

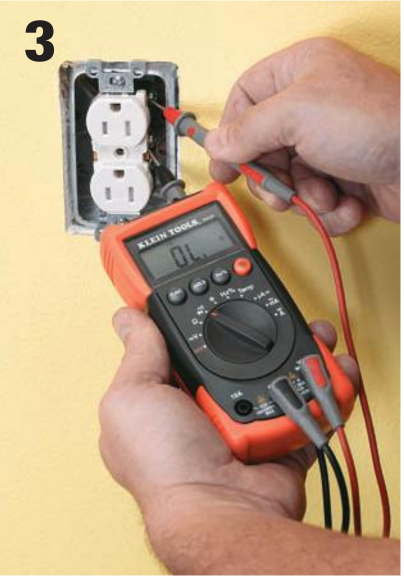

When you encounter a point at which you read infinity ohms or there is no continuity between the neutral and ground wires, the problem lies within the connections in that box or the box just upstream (toward) the panel from the one you are checking. Sometimes you will see evidence of arcing on the wire cap containing the connection which may include discoloration, or a blackish char near the copper.

When you have found the problem connection, remove the wire cap and, if it is possible, cut the damaged portions of the wires off and restrip the wires to expose new copper. Line the wire ends up and twist on a new wire cap.

5
KLEIN TOOLS
120.3

Turn the circuit breaker back on and verify the proper voltage is present at your receptacles by measuring with a multimeter.

Troubleshooting a Short Circuit

Short circuits are a direct connection between the hot or power wire (black or red) and to either the neutral (white) or ground (bare) wire. This connection between the two will cause your circuit breaker or fuse to blow, which should interrupt power to the affected circuit.

Short circuits are a common problem and can usually be solved by taking the following steps. The idea behind electrical troubleshooting is to simplify the circuit by checking it at certain points, in order to narrow down the problem point by process of elimination. Generally, the problem is that there is a bare ground wire touching a hot terminal within a switch or an outlet box. There will usually be a black scorched mark or some sign of an electrical arc where the problem lies.

How to Troubleshoot a Short Circuit

1

Turn the power off at the affected breaker and verify with a non-contact voltage tester that there is no power present. Unplug everything from the receptacles and turn the lights off on the circuit that is affected.

2

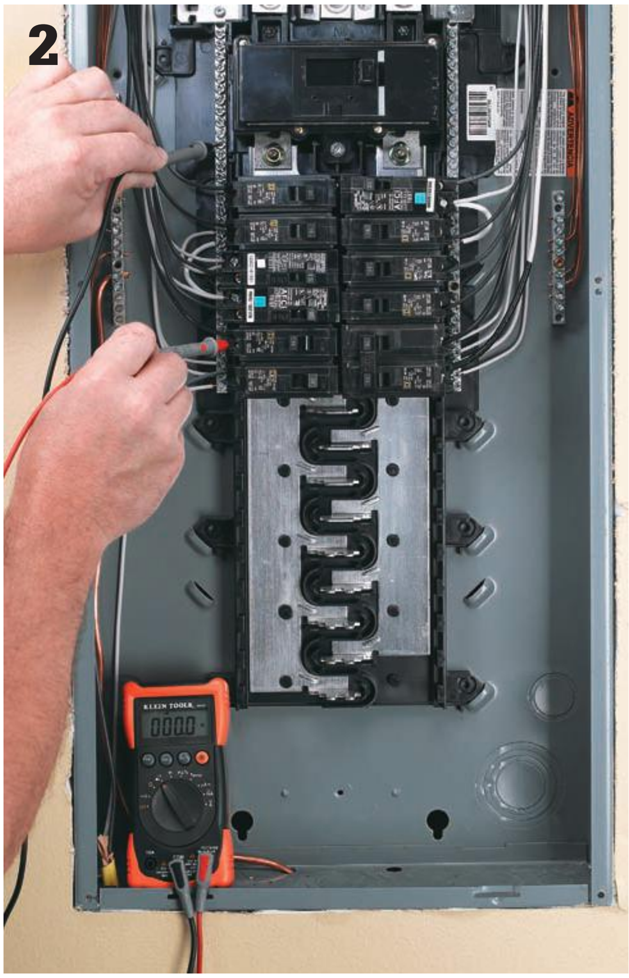

Using a multimeter set to the ohms or continuity setting, check the wires at the panel. Touch one of the probes to the hot or black wire and the other probe to the ground or bare wire. If the meter rings or indicates a low resistance value, you have a direct short to ground. If the meter does not ring or indicates a high resistance value the circuit is clear. If the meter does not ring, start by turning the switches on one-by-one and re-testing to verify the resistance value. If the meter indicates a low resistance value or a short circuit, the problem is downstream from the switch or within the light fixture itself.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 108
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 108

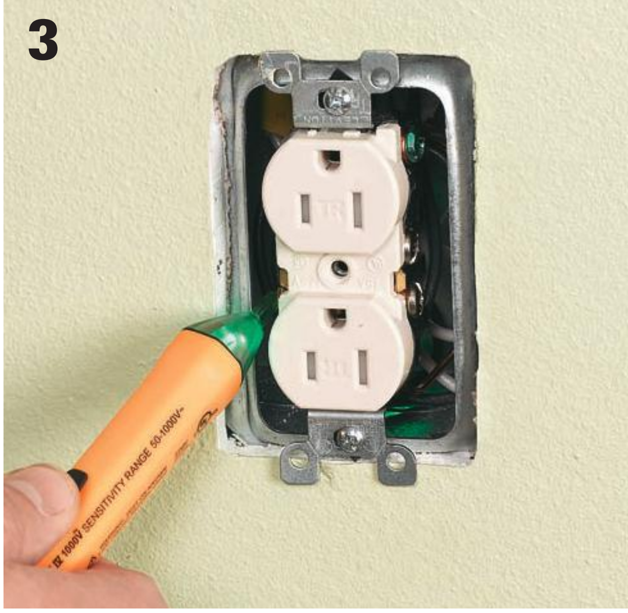
3

If the meter consistently rings or indicates a low resistive value, you will need to find the electrical box that contains the affected circuit. Choose a box that is convenient to open and preferably in the middle of the run, such as a receptacle. Verify there is no power present by touching all of the wires within the box with a non-contact voltage tester.

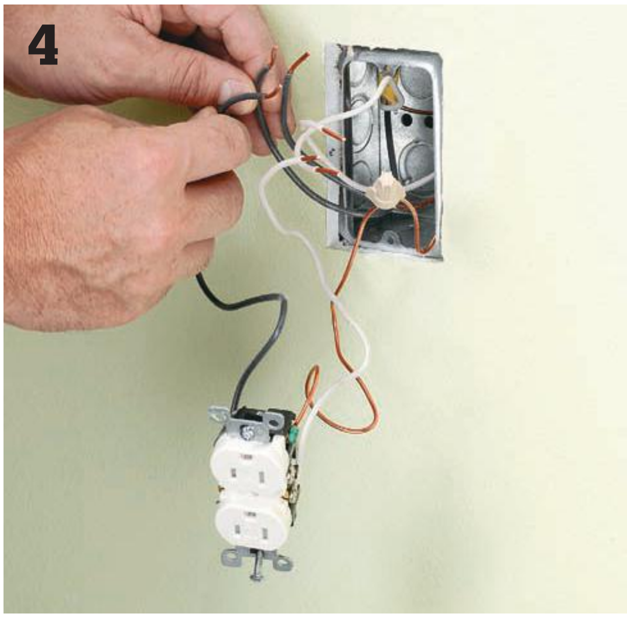
4

If the box you have chosen is in fact in the middle of the run, it will contain at least two cables. Remove the receptacle from the two cables and separate all of the wires.

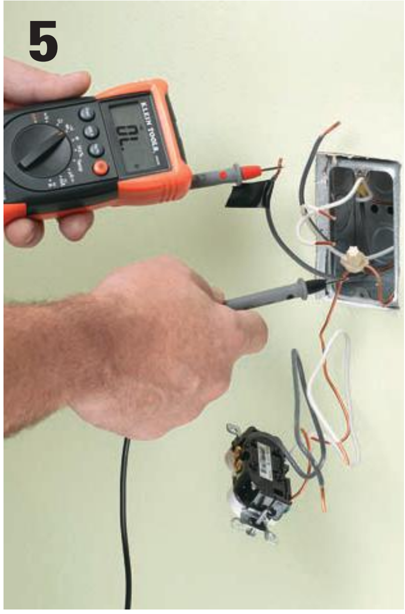
5

Check the resistance between the black and the ground on both sets of cables. One of the cables should cause the continuity alert to ring and the other should not. Mark the affected one with a piece of black tape and place wire caps over the exposed ends of the black wires.

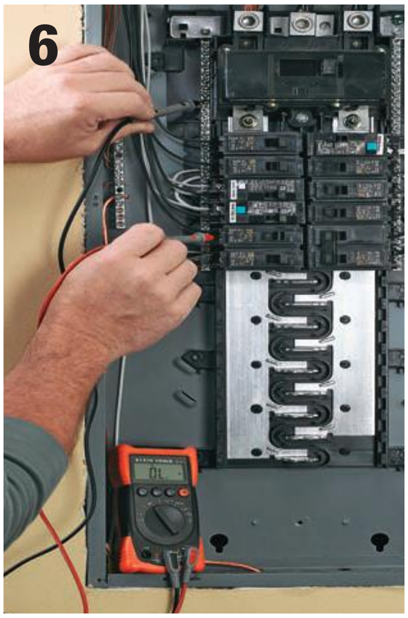
6

Check the wires at the panel to see if the short has cleared. If the short is clear, the problem lies down stream from the opened box and it is now safe to turn the breaker back on to help eliminate further problem points. If the short is still present, the problem lies between the opened electrical box and the panel.

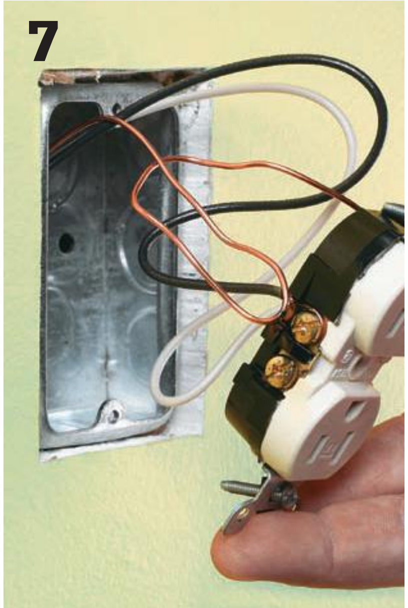
7

Choose another box in the middle of the affected circuit, there by narrowing down the possible problem areas until the short circuit can be positively identified and corrected. When you have discovered the short circuit, verify the wires are still in good shape and repair the connection.

1st
proofs for reference only

Types of Wall Switches

Wall switches are available in three general types. To re-connect or replace a switch, it is important to identify its type.

Single-pole switches are used to control a set of lights from one location. Three-way switches are used to control a set of lights from two different locations and are always installed in pairs. Four-way switches are used in combination with a pair of three-way switches to control a set of lights from three or more locations.

Identify switch types by counting the screw terminals. Single-pole switches have two screw terminals, three-way switches have three screw terminals, and four-way switches have four. Most switches include a grounding screw terminal, which is identified by its green color.

When replacing a switch, choose a new switch that has the same number of screw terminals as the old one. The location of the screws on the switch body varies depending on the manufacturer, but these differences will not affect the switch operation.

Whenever possible, connect switches using the screw terminals rather than push-in fittings. Some specialty switches (pages 118 to 119) have wire leads instead of screw terminals. They are connected to circuit wires with wire connectors.

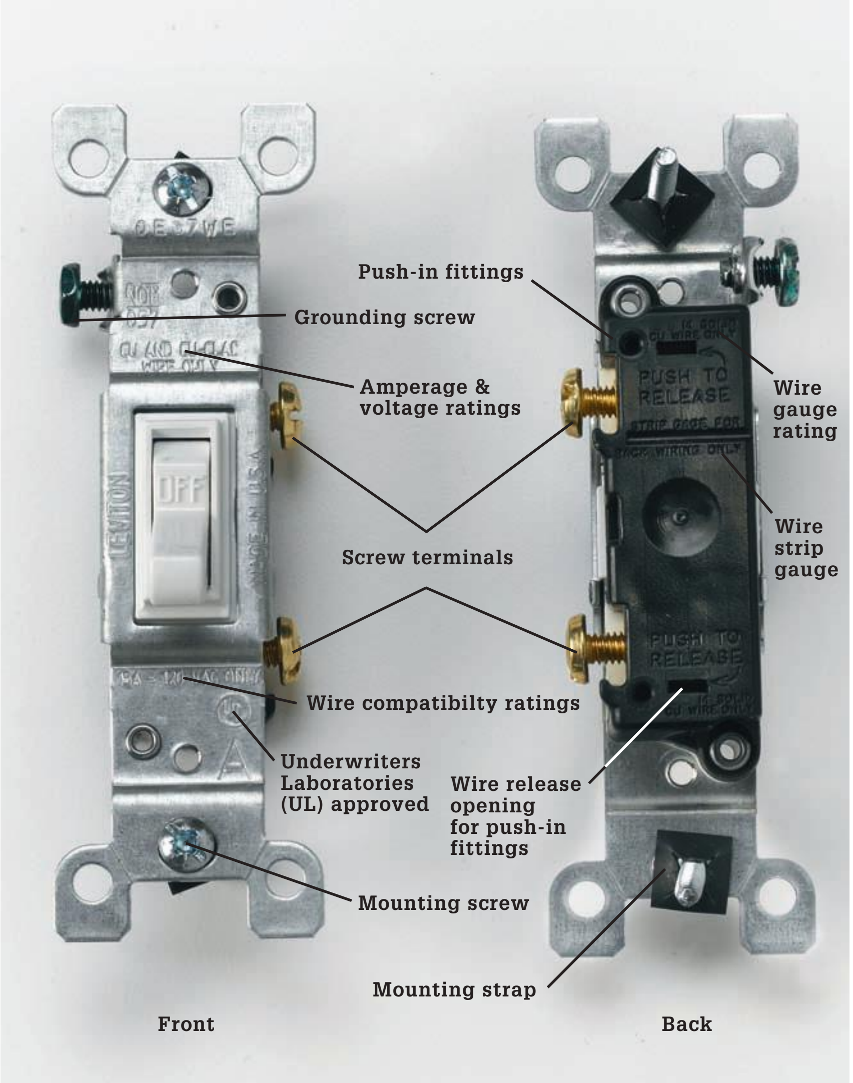
Push-in fittings
Grounding screw
Amperage & voltage ratings
Wire gauge rating
Screw terminals
Wire strip gauge
Wire compatibilty ratings
Underwriters Laboratories (UL) approved
Wire release opening for push-in fittings
Mounting screw
Mounting strap
Front
Back
OFF
LEVITON
MADE IN U.S.A.
PUSH TO RELEASE

A wall switch is connected to circuit wires with screw terminals or with push-in fittings on the back of the switch. A switch may have a stamped strip gauge that indicates how much insulation must be stripped from the circuit wires to make the connections.

The switch body is attached to a metal mounting strap that allows it to be mounted in an electrical box. Several rating stamps are found on the strap and on the back of the switch. The abbreviation UL or UND. LAB. INC. LIST means that the switch meets the safety standards of the Underwriters Laboratories. Switches also are stamped with maximum voltage and amperage ratings. Standard wall switches are rated 15A or 125V. Voltage ratings of 110, 120, and 125 are considered to be identical for purposes of identification.

For standard wall switch installations, choose a switch that has a wire gauge rating of #12 or #14. For wire systems with solid-core copper wiring, use only switches marked COPPER, CU, or CO/ALR. For aluminum wiring, use only switches marked CO/ALR. Note that while CO/ALR switches and receptacles are approved by the National Electrical Code for use with aluminum wiring, the Consumer Products Safety Commission does not recommend using these. Switches and receptacles marked AL/CU can no longer be used with aluminum wiring, according to the National Electrical Code.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 110
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 110
Pro-Vis

Single-Pole Wall Switches

A single-pole switch is the most common type of wall switch. It has ON-OFF markings on the switch lever and is used to control a set of lights, an appliance, or a receptacle from a single location. A single-pole switch has two screw terminals and a grounding screw. When installing a single-pole switch, check to make sure the ON marking shows when the switch lever is in the up position.

In a correctly wired single-pole switch, a hot circuit wire is attached to each screw terminal. However, the color and number of wires inside the switch box will vary, depending on the location of the switch along the electrical circuit.

If two cables enter the box, then the switch lies in the middle of the circuit. In this installation, both of the hot wires attached to the switch are black.

If only one cable enters the box, then the switch lies at the end of the circuit. In this installation (sometimes called a switch loop), one of the hot wires is black, but the other hot wire usually is white. A white hot wire should be coded with black tape or paint.

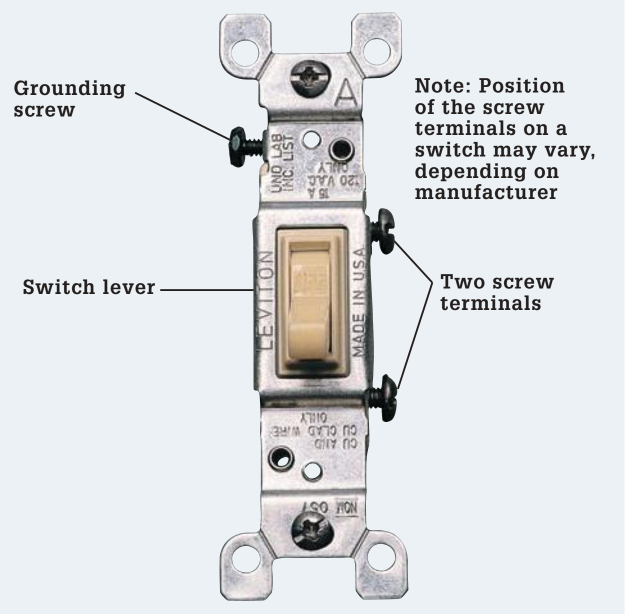
Grounding screw
Note: Position of the screw terminals on a switch may vary, depending on manufacturer
Switch lever
Two screw terminals

A single-pole switch is essentially an interruption in the black power supply wire that is opened or closed with the toggle. Single-pole switches are the simplest of all home wiring switches.

Typical Single-Pole Switch Installations

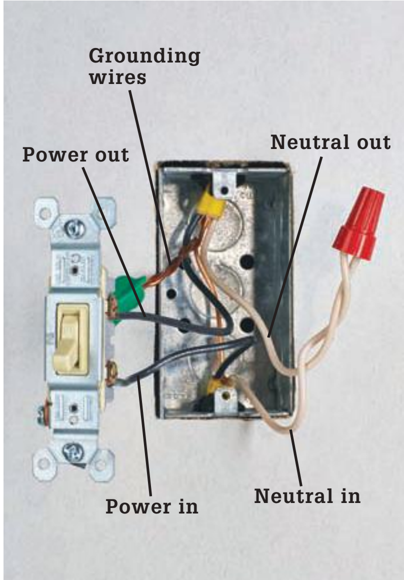
Grounding wires
Power out
Neutral out
Power in
Neutral in

Two cables enter the box when a switch is located in the middle of a circuit. Each cable has a white and a black insulated wire, plus a bare copper grounding wire. The black wires are hot and are connected to the screw terminals on the switch. The white wires are neutral and are joined together with a wire connector. Grounding wires are pigtailed to the switch.

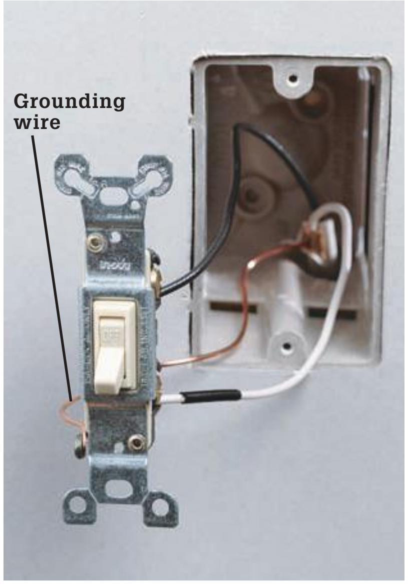
Grounding wire

Old method: One cable enters the box when a switch is located at the end of a circuit. In this installation, both of the insulated wires are hot. The white wire should be labeled with black tape or paint to identify it as a hot wire. The grounding wire is connected to the switch grounding screw.

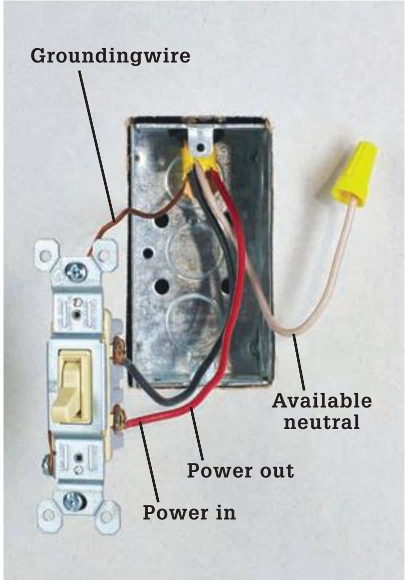
Groundingwire
Available neutral
Power out
Power in

Code change: In new switch wiring, the white wire should not supply current to the switched device and a separate neutral wire should be available in the switch box.

Three-Way Wall Switches

Three-way switches have three screw terminals and do not have ON-OFF markings. Three-way switches are always installed in pairs and are used to control a set of lights from two locations.

One of the screw terminals on a three-way switch is darker than the others. This screw is the common screw terminal. The position of the common screw terminal on the switch body may vary, depending on the manufacturer. Before disconnecting a three-way switch, always label the wire that is connected to the common screw terminal. It must be reconnected to the common screw terminal on the new switch.

The two lighter-colored screw terminals on a three-way switch are called the traveler screw terminals. The traveler terminals are interchangeable, so there is no need to label the wires attached to them.

Because three-way switches are installed in pairs, it sometimes is difficult to determine which of the switches is causing a problem. The switch that receives greater use is more likely to fail, but you may need to inspect both switches to find the source of the problem.

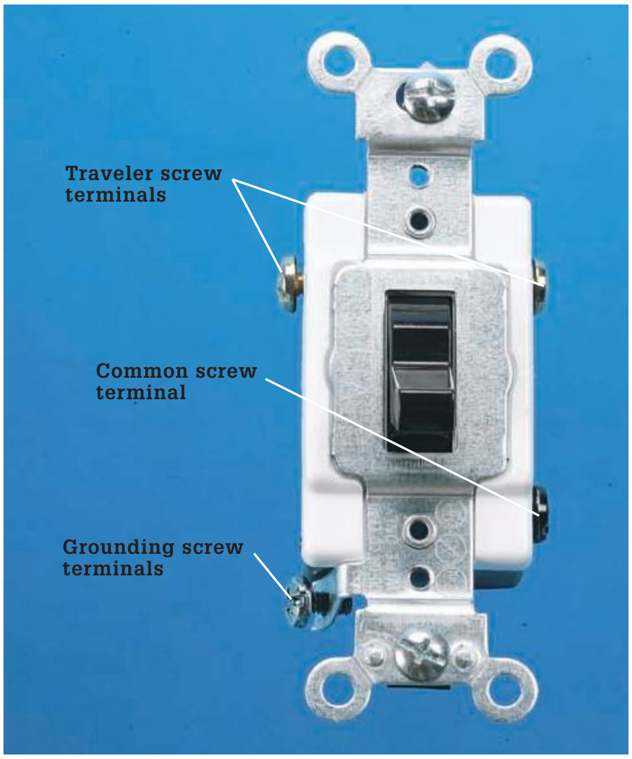

Typical Three-Way Switch Installation

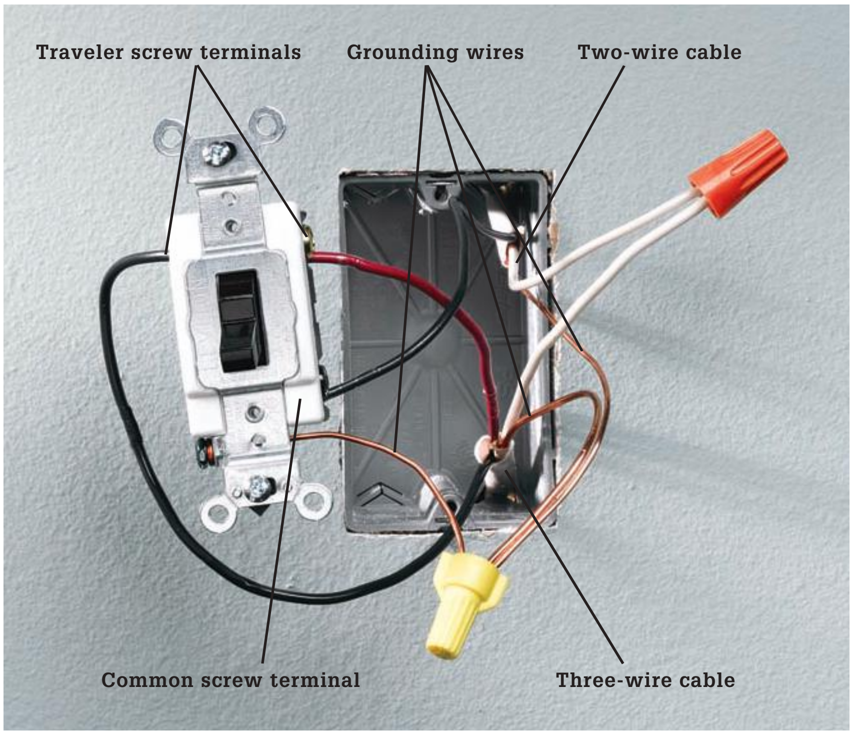

Two cables enter the box: one cable has two wires, plus a bare copper grounding wire; the other cable has three wires, plus a ground. The black wire from the two-wire cable is connected to the dark common screw terminal. The red and black wires from the three-wire cable are connected to the traveler screw terminals. The white neutral wires are joined together with a wire connector, and the grounding wires are pigtailed to the grounded metal box.

1st proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 112
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 112

How to Replace a Three-Way Wall Switch

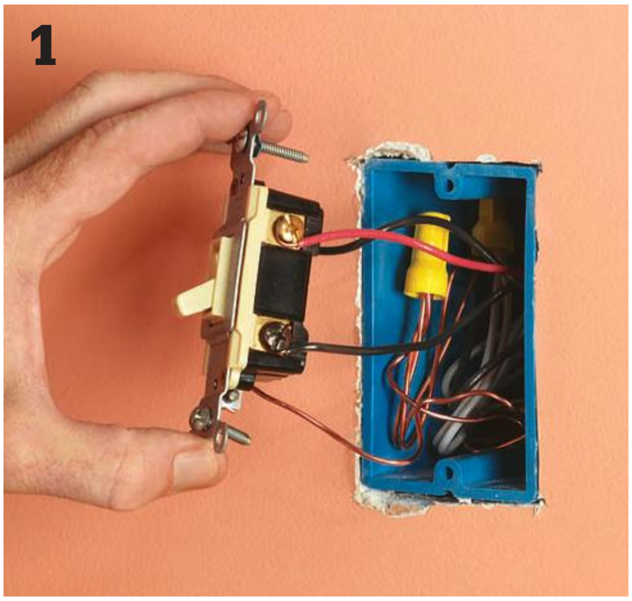

Turn off the power to the switch at the panel, and then remove the switch cover plate and mounting screws. Holding the mounting strap carefully, pull the switch from the box. Be careful not to touch the bare wires or screw terminals until they have been tested for power. *Note: If you are installing a new switch circuit, you must provide a neutral conductor at the switch.*

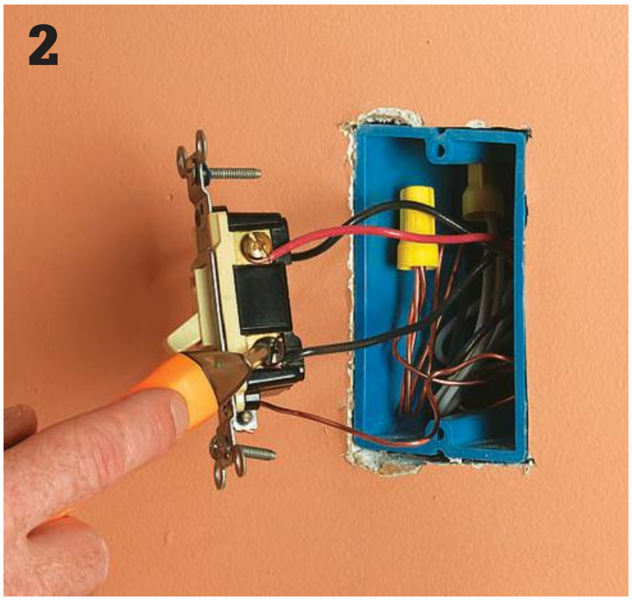

Test for power by touching one probe of the circuit tester to the grounded metal box or to the bare copper grounding wire and touching the other probe to each screw terminal. Tester should not glow. If it does, there is still power entering the box. Return to the panel, and turn off the correct circuit.

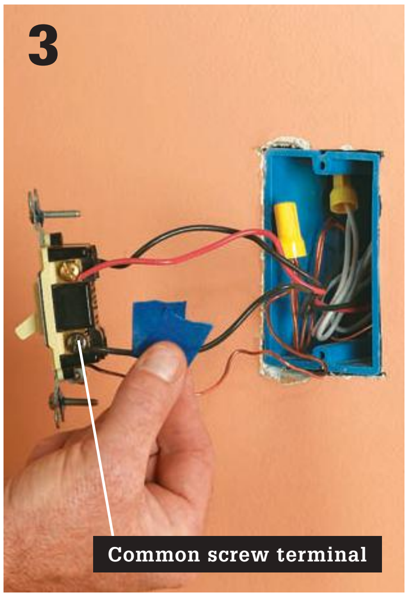

Locate the dark common screw terminal, and use masking tape to label the "common" wire attached to it. Disconnect wires and remove switch. Test the switch for continuity. If it tests faulty, buy a replacement. Inspect wires for nicks and scratches. If necessary, clip damaged wires and strip them.

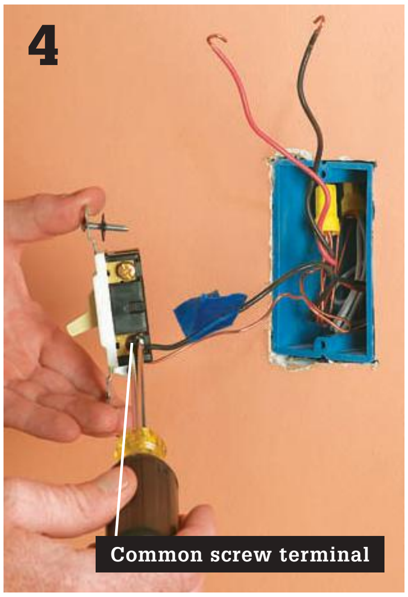

Connect the common wire to the dark common screw terminal on the switch. On most three-way switches, the common screw terminal is black. Or it may be labeled with the word COMMON stamped on the back of the switch. Reconnect the grounding screw, and connect it to the circuit grounding wires with a pigtail.

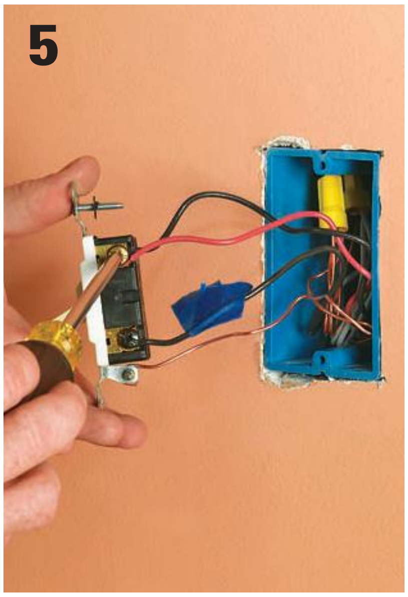

Connect the remaining two circuit wires to the screw terminals. These wires are interchangeable and can be connected to either screw terminal. Carefully tuck the wires into the box. Remount the switch, and attach the cover plate. Turn on the power at the panel.

1st proofs for reference only
Edition
ge: 112
41940 - Advanced Home Wiring_064-128.indd 113
Pro-Vision C M Y K Text
10/10/14 12:05 AM
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 113

Four-Way Wall Switches

Four-way switches have four screw terminals and do not have ON-OFF markings. Four-way switches are always installed between a pair of three-way switches. This switch combination makes it possible to control a set of lights from three or more locations. Four-way switches are common in homes where large rooms contain multiple living areas, such as a kitchen opening into a dining room. Switch problems in a four-way installation can be caused by loose connections or worn parts in a four-way switch or in one of the three-way switches (facing page).

In a typical installation, there will be a pair of three-wire cables that enter the box for the four-way switch. With most switches, the white and red wires from one cable should be attached to the bottom or top pair of screw terminals, and the white and red wires from the other cable should be attached to the remaining pair of screw terminals. However, not all switches are configured the same way, and wiring configurations in the box may vary, so always study the wiring diagram that comes with the switch.

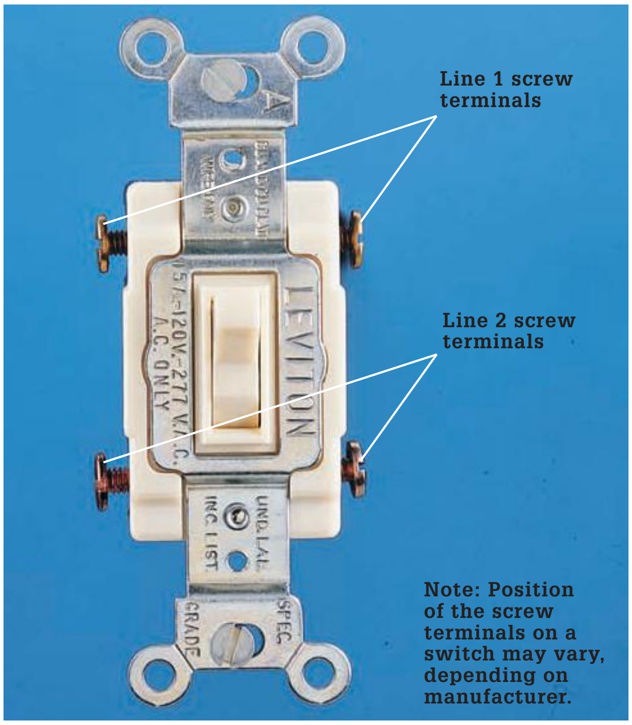
Line 1 screw terminals
Line 2 screw terminals
LEVITON
Note: Position of the screw terminals on a switch may vary, depending on manufacturer.

Typical Four-Way Switch Installation

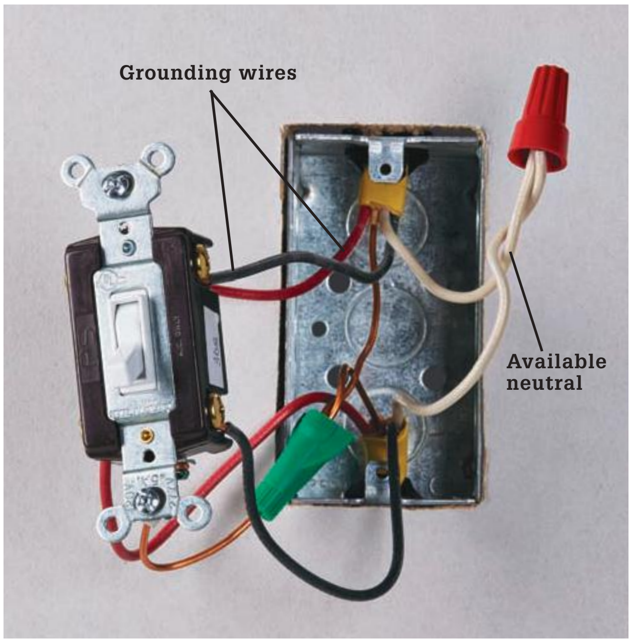

Four wires are connected to a four-way switch. The red and white wires from one cable are attached to the top pair of screw terminals, while the red and white wires from the other cable are attached to the bottom screw terminals. In new switch wiring, the white wire should not supply current to the switched device, and a separate neutral wire should be available in the switch box.

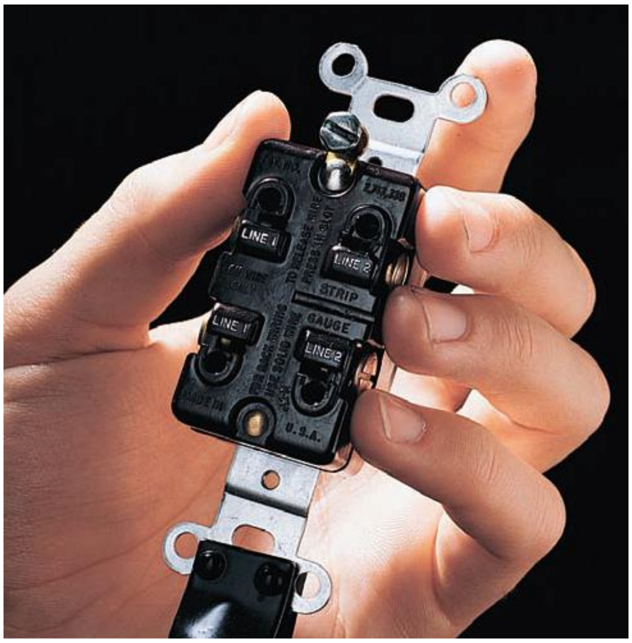

Switch variation: Some four-way switches have a wiring guide stamped on the back to help simplify installation. For the switch shown above, one pair of color-matched circuit wires will be connected to the screw terminals marked LINE 1, while the other pair of wires will be attached to the screw terminals marked LINE 2.

How to Replace a Four-Way Wall Switch

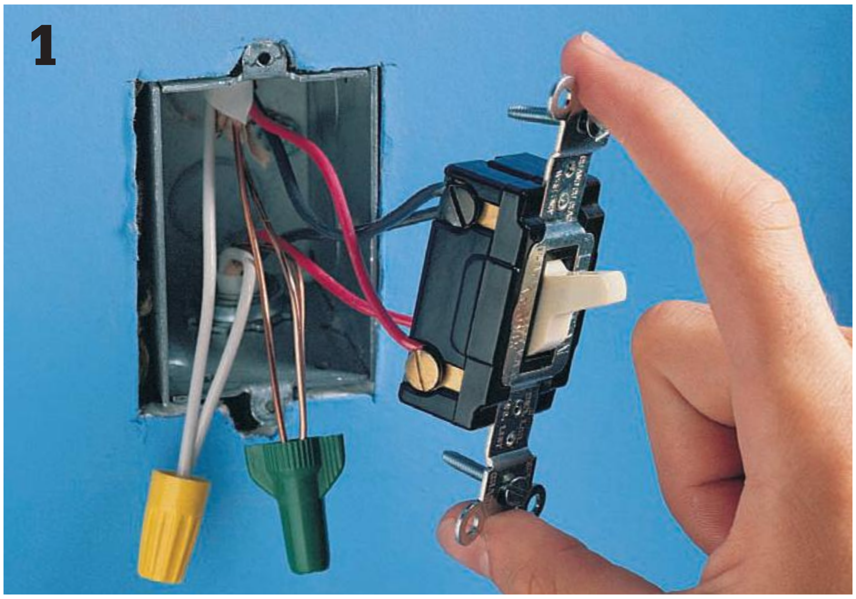

Turn off the power to the switch at the panel, and then remove the switch cover plate and mounting screws. Holding the mounting strap carefully, pull the switch from the box. Be careful not to touch any bare wires or screw terminals until they have been tested for power. Test for power by touching one probe of the neon circuit tester to the grounded metal box or bare copper grounding wire and touching the other probe to each of the screw terminals. The tester should not glow. If it does, there is still power entering the box. Return to the panel, and turn off the correct circuit.

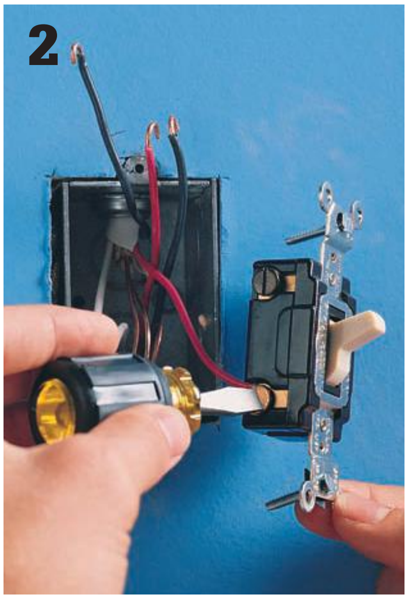

Disconnect the wires and inspect them for nicks and scratches. If necessary, clip damaged wires and strip them. Test the switch for continuity. Buy a replacement if the switch tests faulty.

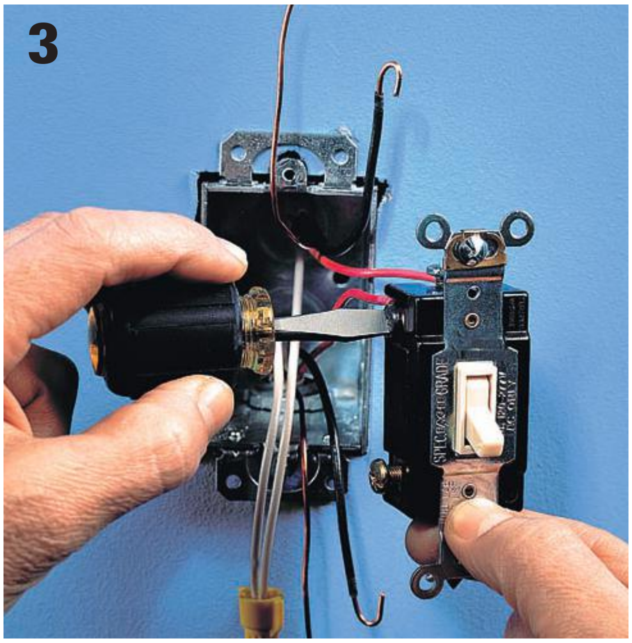

Connect two wires from one incoming cable to the top set of screw terminals.

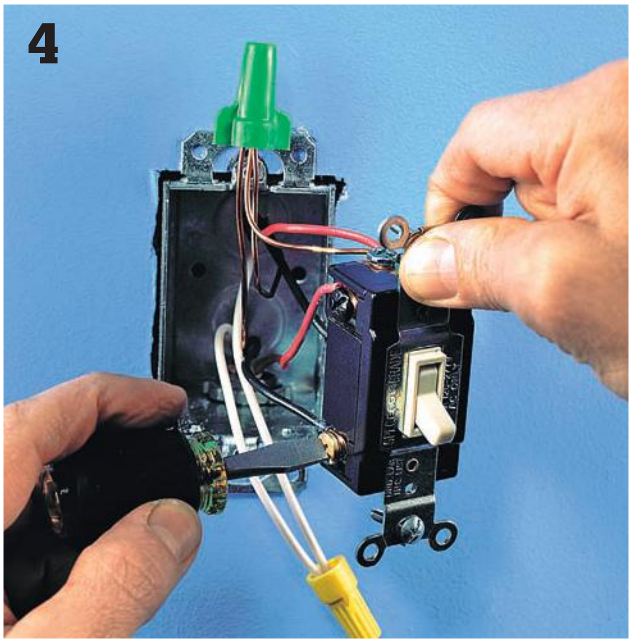

Attach remaining wires to the other set of screw terminals. Pigtail the grounding wires to the grounding screw. Carefully tuck the wires inside the switch box, and then remount the switch and cover plate. Turn on power at the panel.

1st proofs for reference only

Edition
ge: 114
41940 - Advanced Home Wiring_064-128.indd 115
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 12:05 AM
Page: 115

Double Switches

A double switch has two switch levers in a single housing. It is used to control two light fixtures or appliances from the same switch box.

In most installations, both halves of the switch are powered by the same circuit. In these single-circuit installations, three wires are connected to the double switch. One wire, called the feed wire (which is hot), supplies power to both halves of the switch. The other wires, called the switch leg, carry power out to the individual light fixtures or appliances.

In rare installations, each half of the switch is powered by a separate circuit. In these separate-circuit installations, four wires are connected to the switch, and the metal connecting tab joining two of the screw terminals is removed (see photo below).

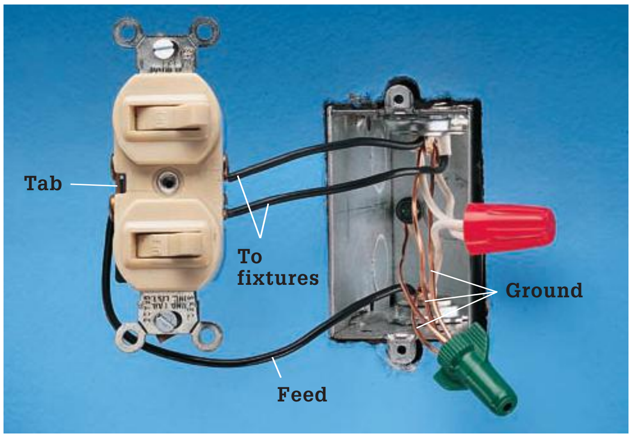

Single-circuit wiring: Three black wires are attached to the switch. The black feed wire bringing power into the box is connected to the side of the switch that has a connecting tab. The wires carrying power out to the light fixtures or appliances are connected to the side of the switch that does not have a connecting tab. The white neutral wires are connected together with a wire connector.

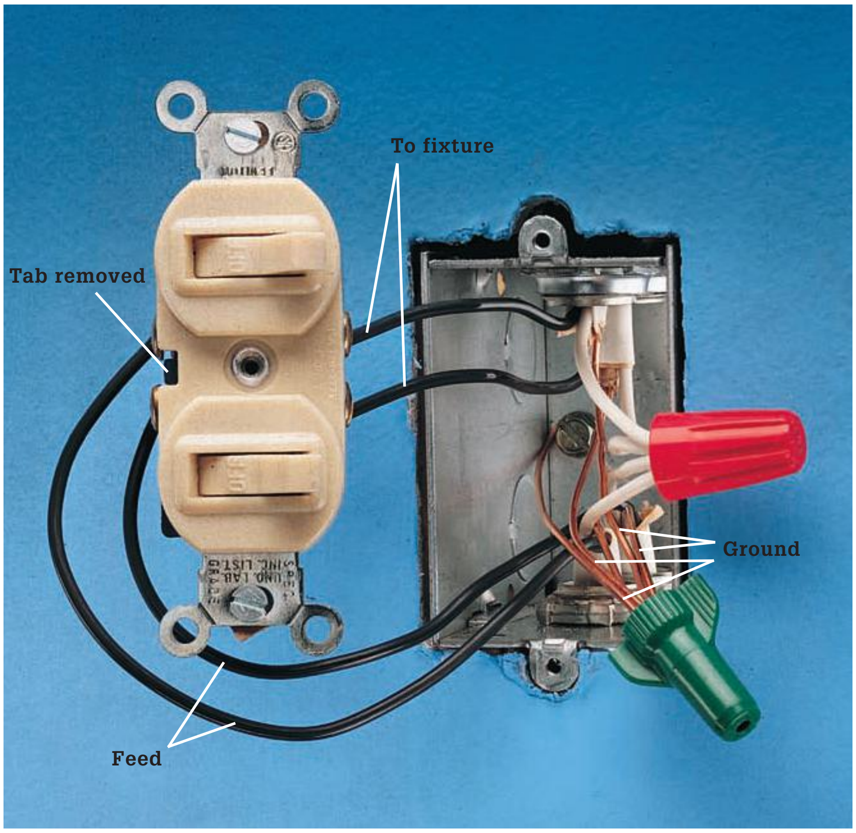

Separate-circuit wiring: Four black wires are attached to the switch. Feed wires from the power source are attached to the side of the switch that has a connecting tab, and the connecting tab is removed (photo, right). Wires carrying power from the switch to light fixtures or appliances are connected to the side of the switch that does not have a connecting tab. White neutral wires are connected together with a wire connector.

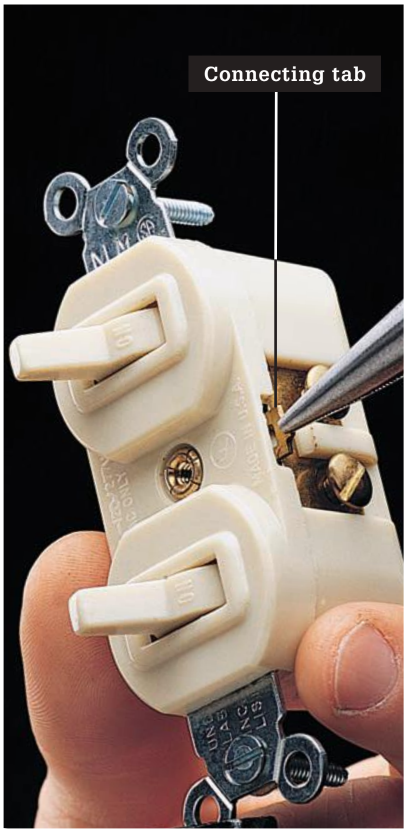

Remove the connecting tab on a double switch when wired in a separate-circuit installation. The tab can be removed with needlenose pliers or a screwdriver.

1st proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 116
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
10/10/14 12:05 AM
Page: 116
Pro-Vis

Pilot-Light Switches

A pilot-light switch has a built-in bulb that glows when power flows through the switch to a light fixture or appliance. Pilot-light switches often are installed for convenience if a light fixture or appliance cannot be seen from the switch location. Basement lights, garage lights, and attic exhaust fans frequently are controlled by pilot-light switches.

A pilot-light switch requires a neutral wire connection. A switch box that contains a single two-wire cable has only hot wires and cannot be fitted with a pilot-light switch.

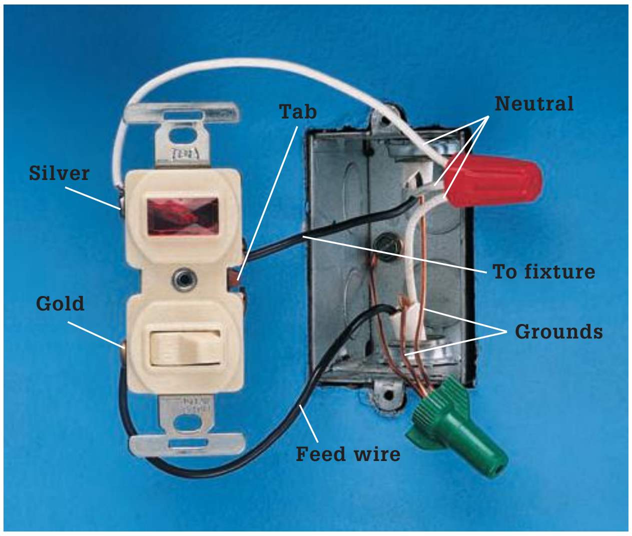

Pilot-light switch wiring: Three wires are connected to the switch. One black wire is the feed wire that brings power into the box. It is connected to the brass (gold) screw terminal on the side of the switch that does not have a connecting tab. The white neutral wires are pigtailed to the silver screw terminal. The black wire carrying power out to a light fixture or appliance is connected to the screw terminal on the side of the switch that has a connecting tab.

Switch/Receptacles

A switch/receptacle combines a grounded receptacle with a single-pole wall switch. In a room that does not have enough wall receptacles, electrical service can be improved by replacing a single-pole switch with a switch/receptacle.

A switch/receptacle requires a neutral wire connection. A switch box that contains a single two-wire cable has only hot wires and cannot be fitted with a switch/receptacle.

A switch/receptacle can be installed in one of two ways. In the most common installations, the receptacle is hot even when the switch is off (photo, right).

In rare installations, a switch/receptacle is wired so the receptacle is hot only when the switch is on. In this installation, the hot wires are reversed, so that the feed wire is attached to the brass screw terminal on the side of the switch that does not have a connecting tab.

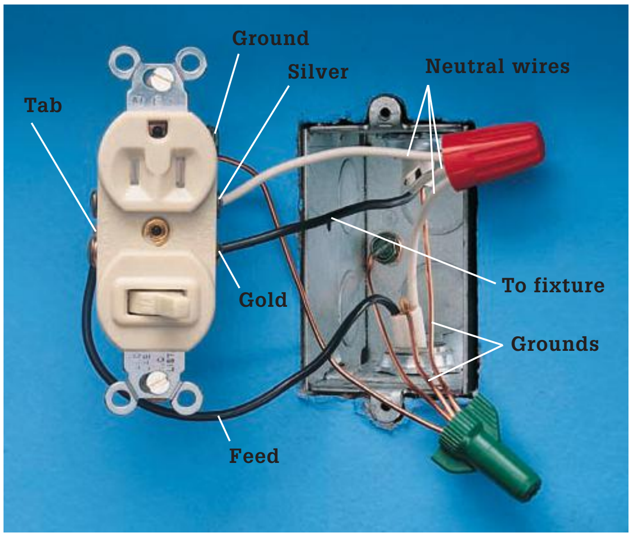

Switch/receptacle wiring: Three wires are connected to the switch/receptacle. One of the hot wires is the feed wire that brings power into the box. It is connected to the side of the switch that has a connecting tab. The other hot wire carries power out to the light fixture or appliance. It is connected to the brass screw terminal on the side that does not have a connecting tab. The white neutral wire is pigtailed to the silver screw terminal. The grounding wires must be pigtailed to the green grounding screw on the switch/receptacle and to the grounded metal box.

1st proofs for reference only

Edition
ge: 116
41940 - Advanced Home Wiring_064-128.indd 117
Pro-Vision C M Y K Text
10/10/14 12:05 AM
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition
Page: 117

Specialty Switches

Your house may have several types of specialty switches. Dimmer switches (pages 96 to 97) are used frequently to control light intensity in dining and recreation areas. Timer switches and time-delay switches (below) are used to control light fixtures and exhaust fans automatically. Electronic switches provide added convenience and home security, and they are easy to install. Electronic switches are durable, and they rarely need replacement.

Most specialty switches have preattached wire leads instead of screw terminals and are connected to circuit wires with wire connectors. Some motor-driven timer switches require a neutral wire connection and cannot be installed in switch boxes that have only one cable with two hot wires. It is precisely due to the rise in popularity of "smart" switches that the NEC Code was changed in 2014 to require an available neutral wire in newly-installed switch boxes.

If a specialty switch is not operating correctly, you may be able to test it with a continuity tester. Timer switches and time-delay switches can be tested for continuity, but dimmer switches cannot be tested. With electronic switches, the manual switch can be tested for continuity, but the automatic features cannot be tested.

Timer Switches

Countdown timer switches can be set to turn lights or fans on and off automatically once each day. They are commonly used to control outdoor light fixtures.

Timer switches have three preattached wire leads. The black wire lead is connected to the hot feed wire that brings power into the box, and the red lead is connected to the wire carrying power out to the light fixture. The remaining wire lead is the neutral lead. It must be connected to any neutral circuit wires. A switch box that contains only one cable has no neutral wires, so it cannot be fitted with a timer switch.

After a power failure, the dial on a timer switch must be reset to the proper time.

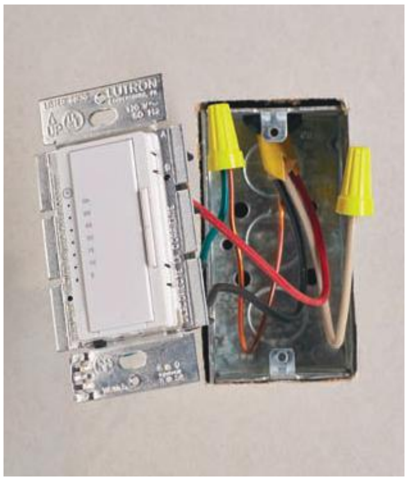

Countdown timer switch. This rocker-type switch gives you the option to easily program the switch to shut off after a specified time: from 5 to 60 minutes. Garage lights or basement lights are good applications: anywhere you want the light to stay on long enough to allow you to exit, but not to stay on indefinitely. These switches often are used to control vent fans.

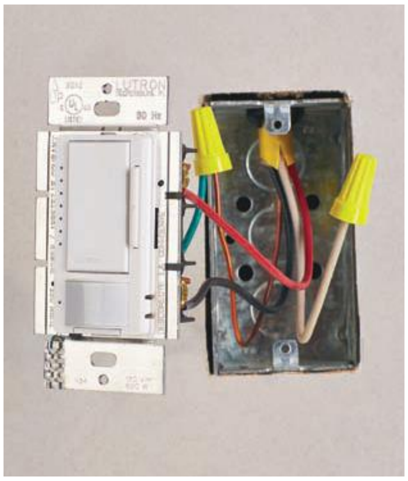

Occupancy sensor. Many smart switches incorporate a motion detector that will switch the lights on if they sense movement in the room and will also shut them off when no movement is detected for a period of time. The model shown above also has a dimmer function for further energy savings.

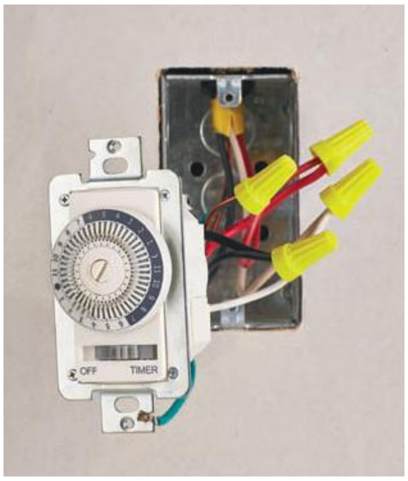

Programmable timer switch. A dial-type timer allows you to program the switch to turn on for specific time periods at designated times of day within a 24-hour cycle. Security lights, space heaters, towel warmers, and radiant floors are typical applications.

1st
proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 118
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 118
Pro-Vis

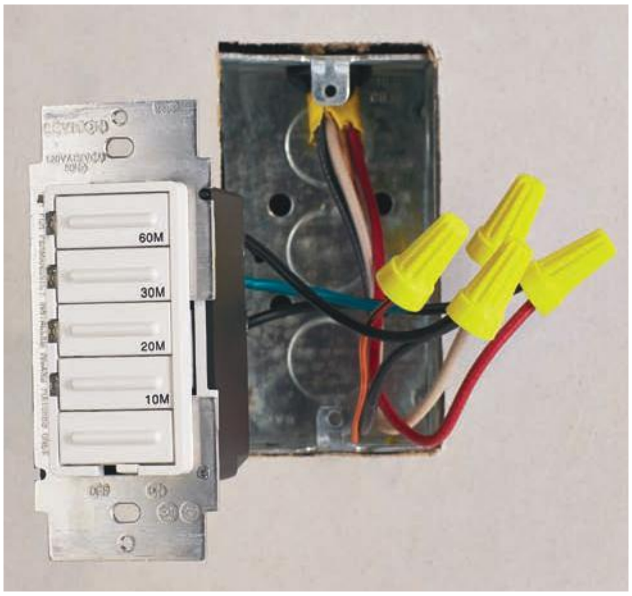

Preset timer switch. This lets you turn on lights, heat lamps, and other loads for a designated amount of time (10 to 60 minutes) with one easy push of a button. The green LED at the bottom of this unit provides a readout of how much time is left before the switch shuts off. The model shown is not compatible with fluorescent ballasts.

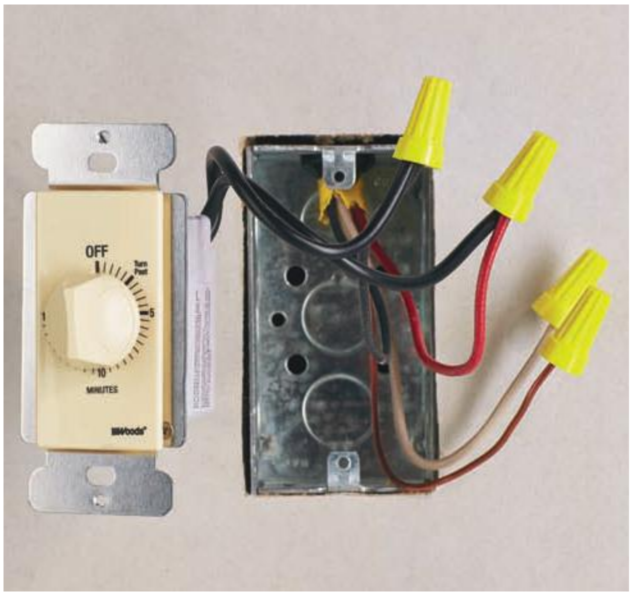

Spring-wound timer switch. A relatively simple device, this timer switch functions exactly like a kitchen timer, employing a hand-turned dial to and spring mechanism to shut the switch off in increments up to 15 minutes.

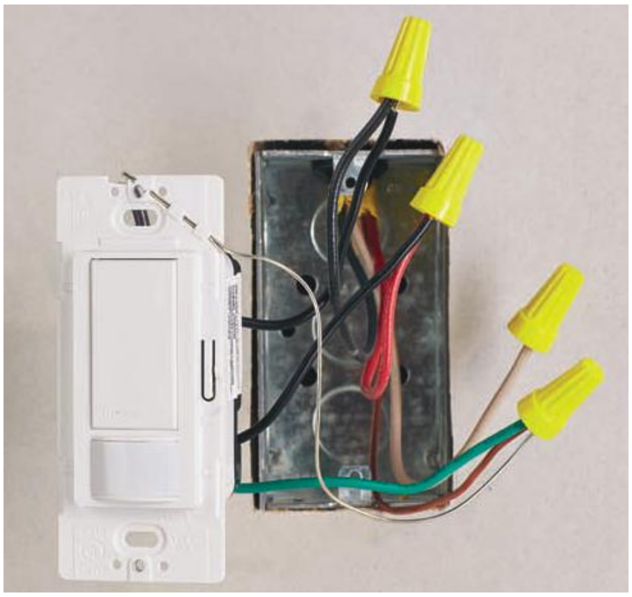

Daylight sensor switch. This switch automatically turns on when light levels drop below a proscribed level. It can also be programmed as an occupancy sensor to shut off when the room is vacant and turn on when the room is entered.

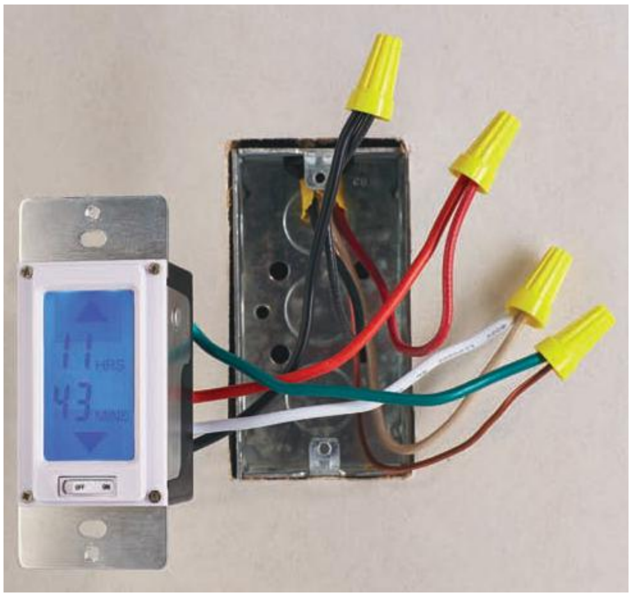

Backlit countdown timer. This digital switch lets you program lights or other devices to stay on for up to 24 hours and then shut off automatically. The backlit, LED readout gives a countdown, in minutes, of the amount of time left in the "on" cycle. Up and down buttons let you raise or lower the remaining time easily, and a manual override button will shut off the switch until it is turned back on.

Ground-fault (GFCI) & Arc-fault (AFCI) Protection

GROUND-FAULT LOCATION REQUIREMENTS

1. Kitchen receptacles. Install ground-fault circuit interrupt (GFCI) protection on all 120-volt receptacles that serve kitchen countertops. This does not include receptacles under the kitchen sink receptacles located on kitchen walls that do not serve the countertop and receptacles that are not within six feet of a sink.
2. Kitchen. Install ground-fault circuit interrupt (GFCI) protection on the outlets that supply dishwashing machines.
3. Bathroom receptacles. Install ground-fault circuit interrupt (GFCI) protection on all 120-volt receptacles located in bathrooms. This applies to all receptacles regardless of where they are located in the bathroom and includes receptacles located at countertops, inside cabinets, and along bathroom walls. This also applies to bathtubs and shower stalls that are not located in a bathroom. Install ground-fault circuit interrupt (GFCI) protection on all circuits serving electrically heated floors in bathrooms, kitchens, and around whirlpool tubs, spas, and hot tubs.
4. Garage and Accessory Building receptacles. Install ground-fault circuit interrupt (GFCI) protection on all 120-volt receptacles located in garages and grade-level areas of unfinished accessory buildings.
5. Exterior receptacles. Install ground-fault circuit interrupt (GFCI) protection on all 120-volt receptacles located outdoors. This does not apply to receptacles that are dedicated for deicing equipment and are located under the eaves. This applies to holiday lighting receptacles located under the eaves.
6. Basement receptacles. Install ground-fault circuit interrupt (GFCI) protection on all 120-volt receptacles located in unfinished basements. An unfinished basement is not intended as habitable space and is limited to storage and work space.
7. Crawl space receptacles. Install ground-fault circuit interrupt (GFCI) protection on all 120-volt receptacles located in crawl spaces. Receptacles in crawl spaces are not required unless equipment requiring service is located there.
8. Sink receptacles. Install ground-fault circuit interrupt (GFCI) protection on all 120-volt receptacles that are located within six feet of the outside edge of a sink. This includes wall, floor, and countertop receptacles.
9. Boathouse receptacles. Install ground-fault circuit interrupt (GFCI) protection on all 120-volt receptacles located in boathouses.

Ground-fault receptacles and circuit breakers detect unwanted current running between an energized wire and a grounded wire.

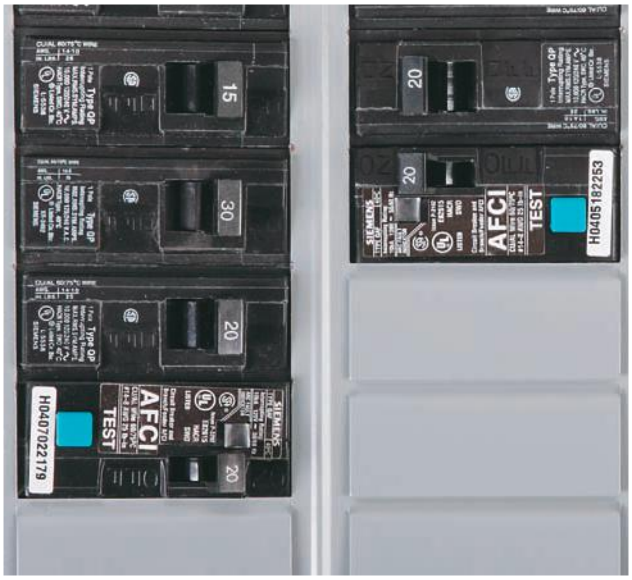

A combination ARC-fault circuit breaker detects sparking (arcing) faults along damaged energized wires and detects these faults between wires. A branch ARC-fault circuit breaker only detects arcing faults between wires.

10. Spas, tubs, and other circuits requiring ground-fault protection. Install ground-fault circuit interrupt (GFCI) protection on all circuits serving spa tubs, whirlpool tubs, hot tubs, and similar equipment. Refer to the general codes for more information about receptacles serving these components.
11. Install GFCI circuit breakers and receptacles so that they are readily accessible.

ARC-FAULT LOCATION REQUIREMENTS

1. Install a combination type or an outlet (receptacle) type arc-fault circuit interrupter (AFCI) on all 15- and 20-amp, 120-volt branch circuits serving sleeping, family, dining, living, sun, and recreation rooms, kitchens, laundry areas, and parlors, libraries, dens, hallways, closets, and similar rooms and areas. This means that 15-and 20-amp, 120-volt branch circuits serving most interior spaces in a home are required to have AFCI protection. Note that garages, basements, utility and mechanical rooms, and exterior branch circuits are not included in this list, although local building officials may include these areas by interpretation.
2. You may provide AFCI protection for the entire branch circuit by installing a combination-type AFCI circuit breaker in the electrical panel where the branch circuit originates.
3. You may provide AFCI protection to a branch circuit using several different combinations of branch-circuit type AFCI circuit breakers and branch-circuit type AFCI receptacles. Refer to general codes or your local building inspector for details about these alternate methods.
4. Provide AFCI for branch circuits that are modified, replaced, or extended. You may use either of the following methods: (a) install a combination-type AFCI circuit breaker in the electrical panel where the branch circuit originates, or (b) install a branch-circuit type AFCI receptacle at the first receptacle in the existing branch circuit.
5. Install AFCI circuit breakers and receptacles so that they are readily accessible.

Receptacles for whirlpool tubs must be GFCI protected.

Junction Boxes, Device Boxes & Enclosures

All electrical boxes are available in different depths. A box must be deep enough so a switch or receptacle can be removed or installed easily without crimping and damaging the circuit wires. Replace an undersized box with a larger box using the Electrical Box Fill Chart (see page 124) as a guide. **The NEC also says that all electrical boxes must remain accessible. Never cover an electrical box with drywall, paneling, or wall coverings.**

NONMETALLIC BOX INSTALLATION

1. Use nonmetallic boxes only with NM type cable or with nonmetallic conduit or tubing. You may use nonmetallic boxes with metallic conduit or tubing if you maintain the electrical continuity of the metallic conduit or tubing by installing a bonding jumper through the box. In many situations it is easier to use a metallic box with metallic conduit or tubing.
2. Extend NM cable sheathing at least ¼ inch into a nonmetallic box knockout opening.
3. Secure NM cable, conduit, and tubing to each box. You may secure NM cable with cable clamps inside the box or with compression tabs provided where the cable enters the box. You do not need to secure NM cable to a standard single-gang box (2¼ by 4 inches) mounted in a wall or ceiling if you fasten the cable not more than eight inches from the box and if the sheathing enters the box at least ¼ inch. Measure the eight inches along the length of the sheathing, not from the outside of the box.

LIGHT FIXTURE BOX INSTALLATION

1. Use boxes designed for mounting light fixtures if a light fixture is to be mounted to the box. These boxes are usually four-inch round or octagonal.
2. You may use other boxes to mount light fixtures on walls if the fixture weighs less than 6 pounds and if the fixture is secured to the box using at least #6 screws.
3. Support light fixtures weighing at least 50 pounds independently from the light fixture box. You may use the light fixture box to support light fixtures weighing less than 50 pounds. Note that ceiling fans are not light fixtures.

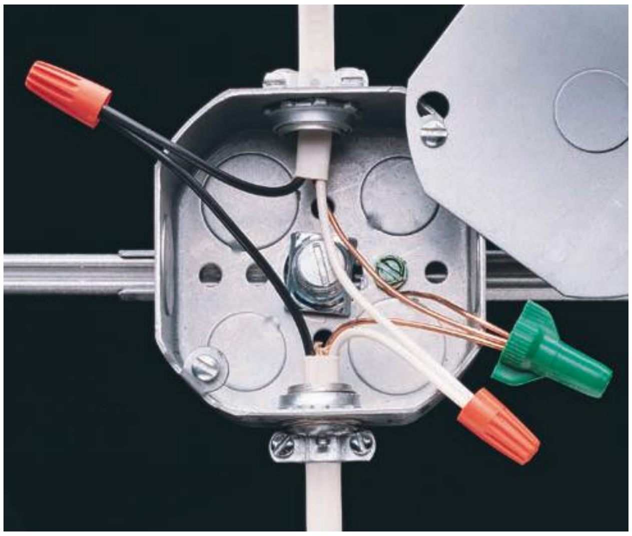

Box shape is directly related to function, as electrical fixtures are created to fit on boxes of a particular shape. Octagonal and round boxes generally are designed for ceiling mounting, while square and rectangular boxes are sized for single-pole, duplex, and other standard switch and receptacle sizes.

Do not support heavy light fixtures using only the light fixture electrical box. The eye hook supporting this chandelier is driven into the same ceiling joist to which the electrical box is mounted.

1st
proofs for reference only
41940 - Advanced Home Wiring_064-128.indd 122
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 122
Pro-Vis

BOX CONTENTS LIMITATIONS

1. Limit the number of wires, devices (such as switches and receptacles), and fittings in a box. This limitation is primarily based on the heat generated by the wires and devices in the box. The actual size of the box relative to its contents is a secondary consideration.
2. Use the cubic inch volume printed on the box or provided in the box manufacturer's instructions to determine box volume. Do not attempt to measure the box volume. Do not estimate box volume from the volume of similar size boxes. You will probably not get the same volume as provided by the manufacturer.
3. Use table "Wire Volume Unit" to determine the volume units required by wires, devices, and fittings in a box.

BOX INSTALLATION TOLERANCES

1. Install boxes in non-combustible material, such as masonry, so that the front edge is not more than ¼ inch from the finished surface.
2. Install boxes in walls and ceilings made of wood or other combustible material so that the front edge is flush with the finished surface or projects from the finished surface.
3. Cut openings for boxes in drywall and plaster so that the opening is not more than ⅛ inch from the perimeter of the box.

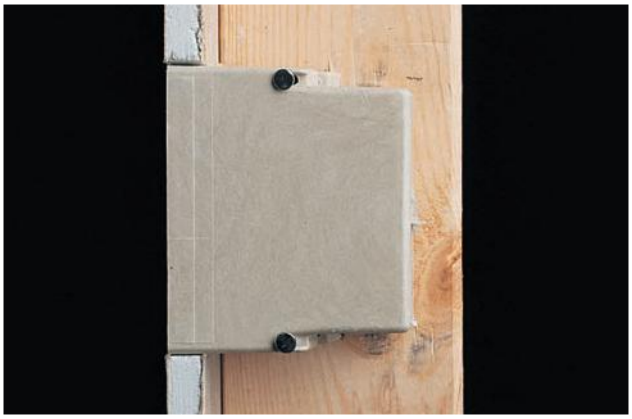

Boxes must be installed so the front edges are flush with the finished wall surface, and the gap between the box and the wall covering is not more than ⅛".

Wire Volume Unit ▸

WIRE SIZE (AWG)	WIRE VOLUME
14	2.00 in.^{3}
12	2.25 in.^{3}
10	2.50 in.^{3}
8	3.00 in.^{3}
6	5.00 in.^{3}

Volume Units ▸

Calculate the volume units required by wires, devices, and fittings based on the following definitions:

Volume units for current-carrying wires. Allow one volume unit for each individual hot (ungrounded) and neutral (grounded) wire in the box. Use Table 47 to determine the volume units of common wire sizes. Example: two pieces of #14/2 NM are in a box. Each piece of this cable contains one hot (ungrounded) and one neutral (grounded) wire and one grounding wire. From table "Wire Volume Unit", each #14 wire uses 2.00 cubic inches in the box. The total volume units required by the hot (ungrounded) and neutral (grounded) wires is eight cubic inches.

Volume units for devices. Allow two volume units for each device (switch or receptacle) in the box. Base the volume units on the largest hot (ungrounded) or neutral (grounded) wire in the box. Example: NM cable size #14 and #12 are in a box. From Table 47, #14 wire uses 2.00 cubic inches and #12 wire uses 2.25 cubic inches. Allow 4.5 cubic inches volume units (2 × 2.25 cubic inches) for each switch or receptacle in the box based on the volume of the larger #12 NM cable.

Volume units for grounding wires. Allow one volume unit for all grounding wires in the box. Base the volume unit on the largest hot (ungrounded) or neutral (grounded) wire in the box.

Volume units for clamps. Allow one volume unit for all internal cable clamps in the box, if any. Base the volume unit on the largest hot (ungrounded) or neutral (grounded) wire in the box.

Volume units for fittings. Allow one volume unit for all fittings in the box, if any. Base the volume unit on the largest hot (ungrounded) or neutral (grounded) wire in the box.

Electrical Box Fill Chart ▸

BOX SIZE AND SHAPE	MAXIMUM NUMBER OF VOLUME UNITS PERMITTED (SEE NOTES BELOW)			
(If volume not labeled by manufacturer)	14 AWG	12 AWG	10 AWG	8 AWG
JUNCTION BOXES				
4 × 1¼" R or O	6	5	5	4
4 × 1½" R or O	7	6	6	5
4 × 2⅛" R or O	10	9	8	7
4 × 1¼" S	9	8	7	6
4 × 1½" S	10	9	8	7
4 × 2⅛" S	15	13	12	10
4¹¹⁄₁₆ × 1¼" S	12	11	10	8
4¹¹⁄₁₆ × 1½" S	14	13	11	9
4¹¹⁄₁₆ × 2⅛" S	21	18	16	14
DEVICE BOXES				
3 × 2 × 1½"	3	3	3	2
3 × 2 × 2"	5	4	4	3
3 × 2 × 2¼"	5	4	4	3
3 × 2 × 2½"	6	5	5	4
3 × 2 × 2¾"	7	6	6	4
3 × 2 × 3½"	9	8	7	6
4 × 2⅛ × 1½"	5	4	4	3
4 × 2⅛ × 1⅞"	6	5	5	4
4 × 2⅛ × 2⅛"	7	6	5	4

Notes:

- R = Round; O = Octagonal; S = Square or rectangular
- Each hot or neutral wire entering the box is counted as one volume unit
- Grounding wires are counted as one volume unit in total—do not count each one individually.
- Raceway fittings and external cable clamps do not count. Internal cable connectors and straps count as one volume unit.
- Devices (switches and receptacles mainly) each count as two volume units.
- When calculating total volume units, any non-wire components should be assigned the gauge of the largest wire in the box.
- For wire gauges not shown here, contact your local electrical inspections office.

BOX SUPPORT IN WALLS, CEILINGS & FLOORS

1. Provide support for boxes that rigidly and securely fasten them in place. You may use nails or screws to support these boxes.
2. Protect screws inside boxes so that the threads will not damage the wires.
3. Wood braces used to support boxes must be at least one by two inches.
4. Use "cut-in" or "old work" retrofit boxes only if they have approved clamps or anchors that are identified for the location where they are installed.

DAMP LOCATIONS

1. Install a receptacle box cover that is weatherproof when the cover is closed and a plug is not inserted into a receptacle located in a damp location. This applies to 15-amp and 20-amp receptacles. A damp area is protected from direct contact with water. Refer to the definition of damp location. You may use a receptacle cover suitable for wet locations in a damp location.
2. Install a watertight seal between a flush-mounted receptacle and its faceplate. This will require a gasket or sealant between the finished surface (such as stucco, brick, or siding) and the faceplate.

WET LOCATIONS

1. Install a receptacle box cover that is weatherproof when the cover is closed on any receptacle located in a wet location. This applies to 15-amp and 20-amp receptacles in any indoor or outdoor wet location. This applies regardless of whether or not a plug is inserted into the receptacle.
2. Install a watertight seal between a flush-mounted receptacle and its faceplate. This will require a gasket or sealant between the finished surface (such as stucco, brick, or siding) and the faceplate.

1st
proofs for reference only

41940 - Advanced Home Wiring_064-128.indd 124
Pro-Vision C M Y K Text
Job: 41940 Title: CSP - Advanced Home Wiring-4th Edition 10/10/14 12:05 AM
Page: 124
Pro-Vis

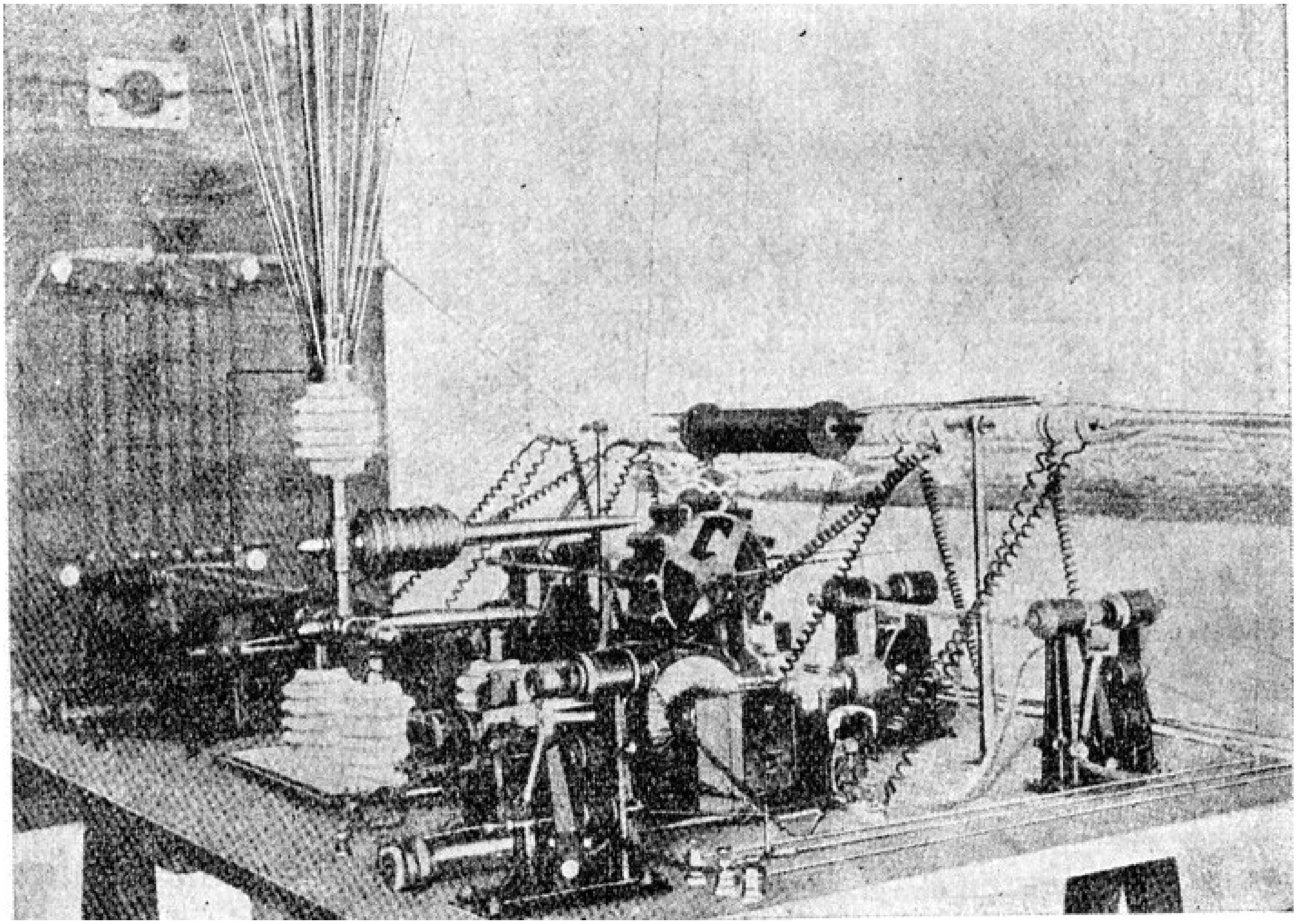

Shared by L. Cozma (Thanks !) --

The Guillot device generated about 2.5-3 Kilowatts with antenna height of ~ 20 meters. Power depends on the total collector surface and height of the vertical antenna. The apparatus in the photo produced ~300 watts with a collector 2 meters tall.

"Practical Utilization of Atmospheric Electricity"

by

Gillbert Darida

in *The Invention Encyclopedia,* pp. 204-207 (1930, Geo. Constantinescu, Ed.)

"The Earth has its own negative electricity, in the soil, and the atmosphere which surrounds the Earth is charged with positive electricity. The electric potential (the voltage) increases with the altitude, so we can say that the electric force is proportional to the atmospheric altitude (after Franklin, Quetelet, Lord Kelvin, Mascart, Joubert and other scientists).

The recent observations demonstrate that the air at 6000-7000 meters in altitude, is very highly charged with positive electricity, which could be explained by the friction between the external photosphere and the upper atmosphere of Earth, which rotates at a speed of more than 100,000 Kilometer per hour.

In that way, the Earth works like an electrostatic generator with electric charging by influence --- the upper atmosphere is positively charged by influence and the Earth crust obtains the negative polarity.

Between the two environments, the air and the soil, and inside the low atmosphere, in conditions of good weather, there are sbout 800 positive ions and also, 680 negative ion (and electrons) in just one square-centimeter of normal ionised air.

The Earth behaves like an huge electric armature negatively charged, which repels the electrons and attracts the positive ions. That positive ions' attraction determinates an electric current, also called "convective current". That's like an invisible continous bombardment, subject to daily and seasonal variations, which could be aproximated at **3 x 10 ^ -16 Amperes per square-centimeter,** and that is a total value of 1500 Amperes for the entire surface of the Earth.
The question is --- how does this current always maintains the same direction ? We can suppose that the natural radioactive emissions of the soil is responsable for this. We also know that radioactive emissions of the Earth works usually near the soil surface, and that explains the ionisation phenomenona inside caves.

The ionisation of the low atmosphere could also be the effect of the radioactive emissions of the Earth, especially when the X radiation works. Also, the Earth atmosphere is ionisated by the external radiation proceeding from the Sun and from the space environment, especially the action of ultraviolet radiation and the electron fascicles emitted by the Sun surface, at the temperature of 6000 deg. Celsius.

The capture of atmosphere electricity has been used in France, with aerial cables mounted on the Mont Blanc, and also in Germany --- with conductive cables carried by the captive baloons.

The atmosphere electricity collect system invented by eng. Jules Guillot is most ingenious and it relies on "the **electric siphon**" [*m.n. --- the discharging devices or spark-gap used today from the protection of aerial electric cables against the atmospheric electricity ; Jules Guillot has thought to recuperate and to use that electricity*].

His method consisted in the direct "pumping" of the atmosphere electricity using a collecting device which had two antennae and several collecting rods.

One antenna is **vertical** and it has a lot of **rods** scattered like an opened **fan**, with the tips against the **zenith**, for collecting the **negative** electricity which comes from the air ; the **horizontal** antenna is orientated against the **South** and its role is to collect the **positive** electricity.

The air electricity seems to have the double-polarity, as we can see at two electrised clouds when between them apears the electric discharge. The air could have different electric charging and the ionisation processus of atmosphere is very heterogenous. The inventor Guillot used **two separated and insulated armatures with the positive armature against the South** (more precisely, against the **Equator**...) and the **negative** armature against the **zenith**.

We can see in the scheme :

1,- the device P is the "**collector apparatus**" having **15-20 meters height** ; it also has a lot of **rods** which are scattered **fan**-like, all this on top of an **iron pillar** ; and also an **horizontal antenna** S oriented against the **South** ; the two antennae are disposed at **90 deg. angle** forming the two armatures where the electrical tension must apear. As the **height** of the pillar **increases**, the **voltage** increases more and more ;

2,- a system of **lightning-rods** (p1, p2, p3) for protection against discharges of atmospheric electricity during storms;

3,- the regulator system (**voltage regulator**, etc) R ;

4,- the electric **resistors** (Cn) for absorbing the parasite currents ;

5,- the "electric siphons" which have the role of making a magnetic field as a good environment for the transport of the ionized fluid, absorbing the electricity from the air ; the E1 and E2 are their regulator devices ;

6,- an **auxilliary continous current power supply** which send the current against an special **excitation coil**, used as starting device ; it works only when the system is starting.
This collector device works very good for lighting and heating. The device made by Guillot had **20 meters height** on the total surface of **3 square-meters**. Also, J.Guillot used and electrical transformer for the industrial utilization of this "collector" --- as power supply for industrial electric engines."

French Patent # 551,882

Apparatus for Capture of Electric Currents in the Atmosphere

16 April 1923

Considering the terrestrial globe like the inductor of a dynamo, where the extraterrestrial ether is the inductor of electric currents circulating in the atmosphere; the apparatus of the invention described here selects through the device described below two perfectly distinct currents and eliminates others.

The device includes:

1, The air sensor; 2, a series of lightning rods; 3, vacuum; 4, an array of resistances; 5, vacuum; 6, witnesses.

1. An aerial sensor mounted on a pole is composed of **magnetic steel** fixed and sealed by a porcelain **insulator**, and surrounded a base by a **bronze ring** which are screwed **32 points of soft iron**, all forming antenna a pole. level beside it and isolated from the first, forming the 2nd pole, is fixed a sharp point of magnetic steel fixed in the horizontal position and directed towards the south and **movable to an angle of 45** is sealed in a porcelain insulator. This point is also circled a **ring of copper** notches figure. 1.

2. A series of 6 lightning surge arrestors at corners between each pole and the earth and the various gauges.

3. A first **regulator** form of a device similar absolutely similar to the aerial amtenna, but the two poles are superposed and opposed to the vertical point, and a **copper disk connects to ground**. Opposite the horizontal point, a **ring of tin** to which are welded tubes alternatively **16 tubes** composed each composed of **copper-lead** and **iron-lead** . The ring connects to the **ground** as shown in figure. 2.

A second regulator form of **automatic breakers** balanced on 3 poles and two poles also equally balanced.

4. A panel of **resistances** composed of **wire mesh** in bunches from different sections of **glass tubes** containing **copper** dust, **coal** and flowers of **sulfur**.

Vacuum cleaners consists of a **wooden box** on each pole, containing a **porcelain vase** in which **layers** isolated with **mica** is made up of **mercury, tin, coal, copper** and **sulfur**, all contaied in a **copper tube**.

6. The witnesses are comprised of ordinary incandescent lamps.

Resume

By the point at the zenith and the point to the south we channel two currents forming the two poles. We also protect from lightning. It regularizes the flow by a regulator and similar devices by a controlling each devise of adequate strength of the current harmful nature of these devices do not have loads. The refined current is conducted by ordinary copper wires**.**

French Patent # 565,395

Combined Apparatus for Capture of Atmospheric Electric Currents with Immediate Implementation

25 January 1924

We know that earth with its constitution, its rotation and movement in space, provides the electricity in the atmosphere. The electrical currents escape into space or largely accumulate towards the equator, as a result of the greater periphery of the globe.

In the atmosphere, there are two clearly distinct poles, i.e., clouds that can be electrically positive or negative; everything in nature shows this, otherwise it would be difficult to explain the lightning that occurs between the clouds and which are none other than contrary cloud electrical charges discharged by too close proximity.

The invention relates to a set of devices capable of capturing the atmospheric electricity.

So far, all searches made for this purpose has been to capture the atmosphere, i.e., that brought together into one system to capture two poles and opposes effect has been to destroy or cancel each other leaving the amount as the difference of the strongest to weakest.

It is therefore easy to understand why, considering the two poles of air as about equal, it is almost impossible to measure some potential with the land, the highest on the lowest remaining which can be positive or negative, copper being the strongest of one or other of these polarities.

It is on this basis that all research until now have been made, and that is why we must consider that all the time, between the two poles captured in the atmosphere and not in any mixture of systems uptake, it is possible to obtain tension and intensity, a considerable and measurable power captured separately between the two poles and remaining isolated from one another until their utilization .

In accordance with this invention, this collection is made by attraction of primary electrical currents in teh atmosphere, either positively or negatively charged, by an antenna that has two points absolutely isolated from one another, where one points to the zenith to attract negative electricity, and the other turned south, toward the equator to draws in the masses grouped in this area.

This shows and proves the existence of atmospheric electrical currents, and that it is possible to capture them with specially designed equipment primarily acting as regulators, as the persistent difficulty encountered so far in realizing their capture lies with large variations of tension in which the current atmosphere is present in space, for each of the poles, and an overload could inevitably fatal.

The patent in its present form of presentation does not cover the regulators or devices that are anticipated in the system, because they can be designed in different ways and give the same result, but covers the application and grouping of equipment operating as automatic relays with variable influence to limit current tension to that chosen for utilization.

These devices have a role associated with an overflow reservoir outside carrying excess liquid; these regulators will divert the excess voltage flows to ground.

We can still absorb these regulators having filters because they have the sole purpose of diverting the post being used , currents that are not yet known but likely that we will call for simplicity, abnormal currents. [sic...]

The patent also covers systems for regulating currents, applied to each of the poles because it is recognized that the shape and tension of the currents of the two poles are not equivalent.

The method of capture is by special antenna has directed two points in the atmosphere, as will be stated after this, with Boot ecoulement prior to the current atmosphere is also of great importance.

Finally, the current atmospheric reception is adequate even with the installation of a post, which can be expected at any location without the need for special altitude, as has been attempted in previous efforts.

We refer to the attached drawings:.

Figure 1 represents in elevation and side view an antenna that only has two points, to overcome an elevation ;

Figure 2 is the end of one of the peaks, the horizontal:

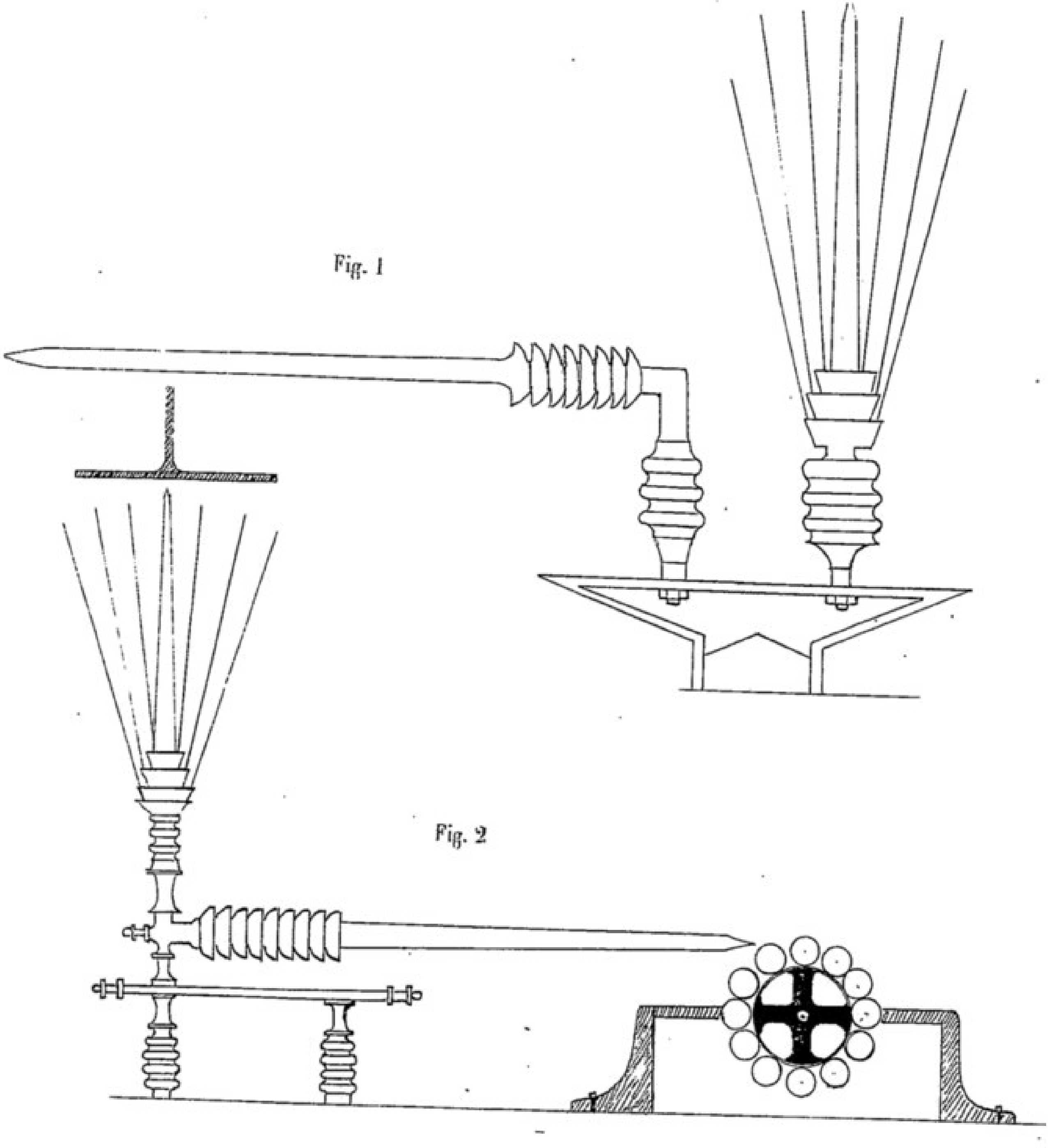

Fig. 1

Fig. 2

Figure 3 shows schematically the table on which are mounted various devices or accessories that act as regulators of current tensions.

Fig.3.

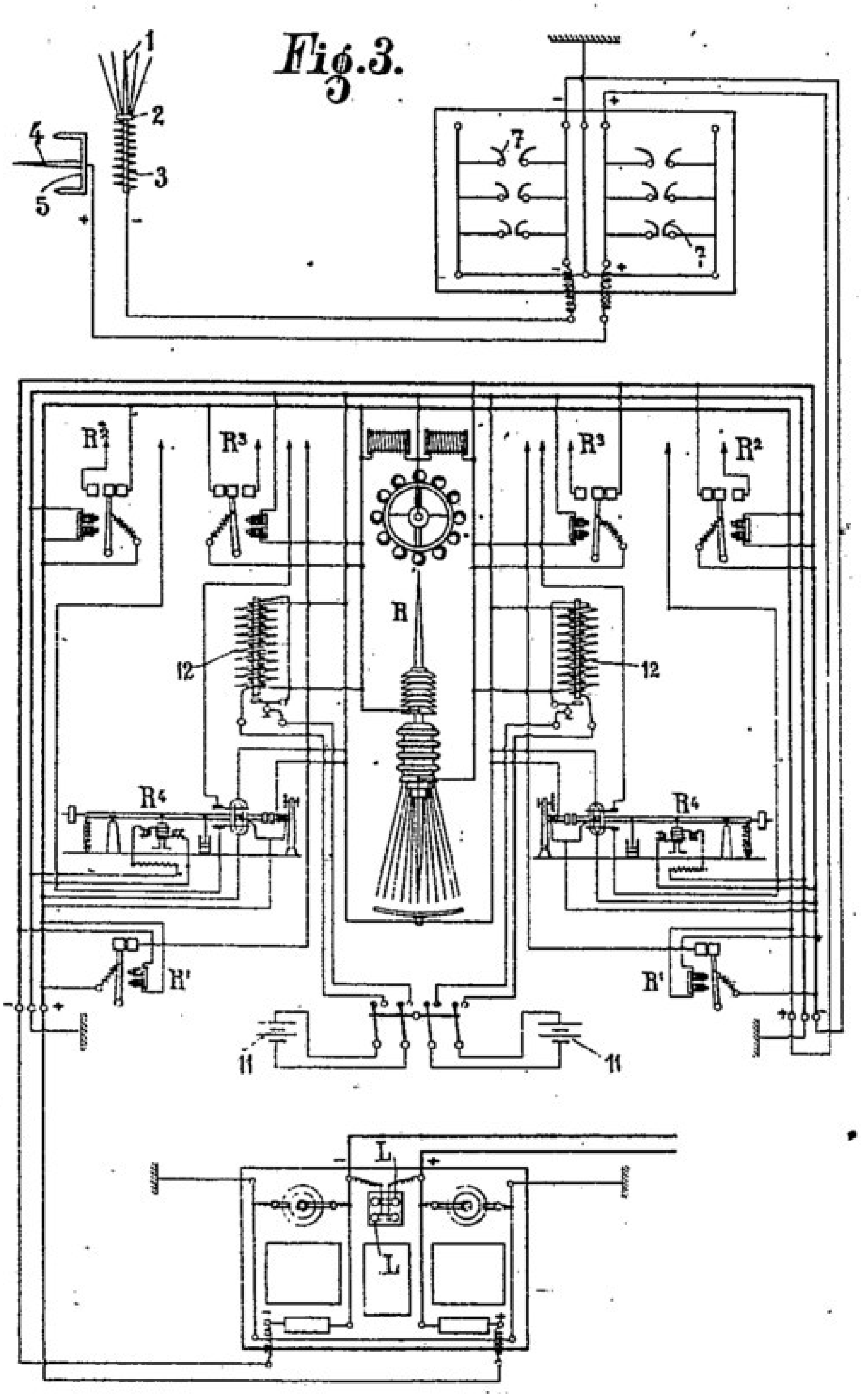
1
2
3
4
5
+
-
7
7
R^2
R^3
R^3
R^2
R
12
12
R^4
R^4
R^1
R^1
11
11
L
L

Figure 4 is a scheme of surge arrestors, and

Figure 5 a table of resistances.

Fig. 4.

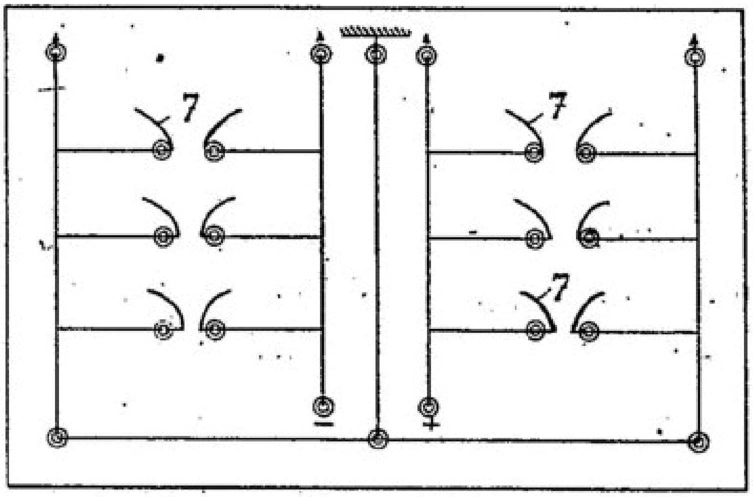

Fig. 5.

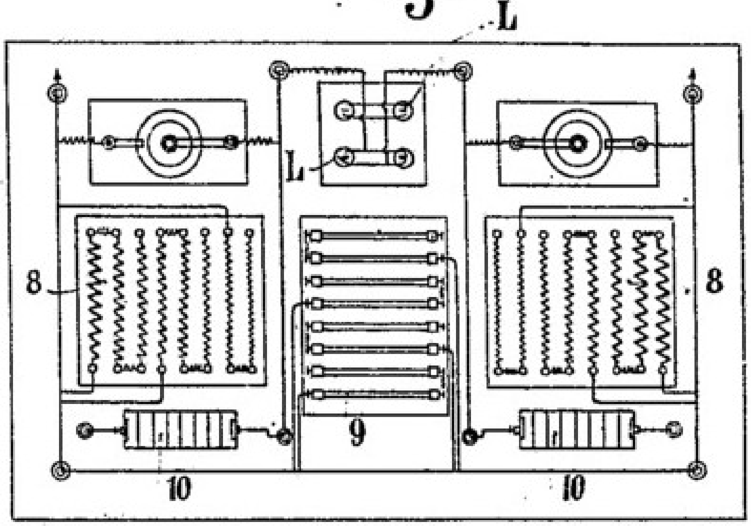

The necessary arrangements to capture atmospheric current form is as follows:

1. The **antenna** in figure 1 is formed with two peaks: the 1 is **steel**, supported by a **bronze ring** 2, surrounded by a **spiral form** 3; this point is oriented for attracting the **zenith** layers of **negative electricity** and the other edge 4 contains **copper coils** 6, and is oriented **south** to the **equator** to draw in the masses grouped in this area. To use this antenna it suffices to be raised on a mast or **pole** to a height a little higher than that of surrounding houses.

2. On the table **lightning arrestors**, contained 4, forms of 7 **copper horns** used to absorb large atmospheric discharges by certain time, as it would be dangerous to allow movement in the devices regulators.

3. On a table of **two** symmetrical systems **regulators**, Figure 3, one for each pole, to obtain and allow a regularization of large excesses in the current tensions, by automatically absorbing surges before passing over the surfacea chosen for the desired use. These regulators are represented on the drawing by the references R, R1, R2, R3, R4.

4. In the **resistance** panel in communication with regulators, these resistances are formed as **spiral coils or screens** 8, **tubes 9 of glass filled with fine coal dust , aluminum powder, sulfur and fine copper dust**. Finally, **two** special 10 **processors** and constitutes **layers** of metal forming an **absorbant** for harmful or abnormal currents.

5. As a source of energy formed **auxiliary batteries** 11 and **Ruhmkorff coils** 12 to get into the antenna sufficient attraction by means of an adjustment of intensity, which once it begins, allows the constant flow of atmospheric current in the system.

In considering Figure 3, which represents a sort of an installation scheme of the device, the reception of atmospheric electrical current will, as set out above, producing a priming of these currents. This boot is made in launching the system which is double to answer each of the poles, the current batteries 11 through 12 and reels going to the antenna.

This boot will continue until the disposal of electrical current atmosphere is evident by the one or more table lamps L of resistance shown in Figure 5.

Once the lights illuminate it is advisable to charge the batteries to boot. Natural flow and constant currents will be captured by the regulators , Figure 3; they will regularize in tension and in amperage and then directed to use in table lamps.

It has been said above that regulators R, etc., are intended and designed to automatically eliminate earth surge currents capture and abnormal currents that cannot yet be level but that these regulators could be replaced by devices with the same function.

There is therefore no need to describe in detail and especially the claims. We only specify that these regulators must be proper protection for the purposes described above.

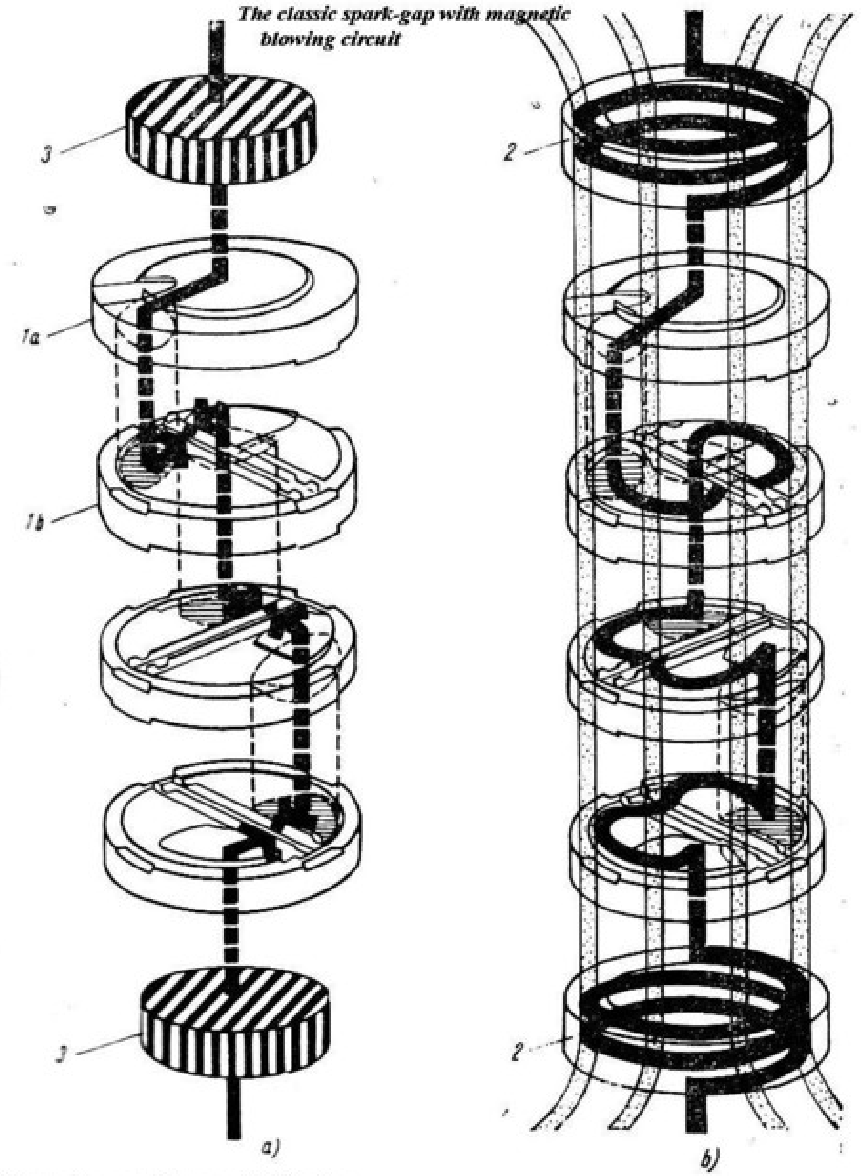

The classic spark-gap with magnetic blowing circuit

The spark-gap with magnetic blowing :
a,- the initial stage ; the impulse current send in soil through the shunt-resistences 3, because the coils 2 have an impedance too big for the impulse of 10/20 µs ;
the spark-gap 1 has two elements disc-shaped 1a and 1b, the inferior disc have one electrode and the upper disc the other electrode of the spark-gap circuit.
b,- the second stage ; the accompanying current discharge - after the impulse current ;
this accompanying current pass through the coils 2 which produce the magnetic field with the role to extend the electric arc into the slit between insulator discs 1a and 1b ; the thickness of this slit is smaller and smaller in accord with the extension of the electric arc, in this manner producing the magnifing of its electric resistance and finally, the electric arc extinction.

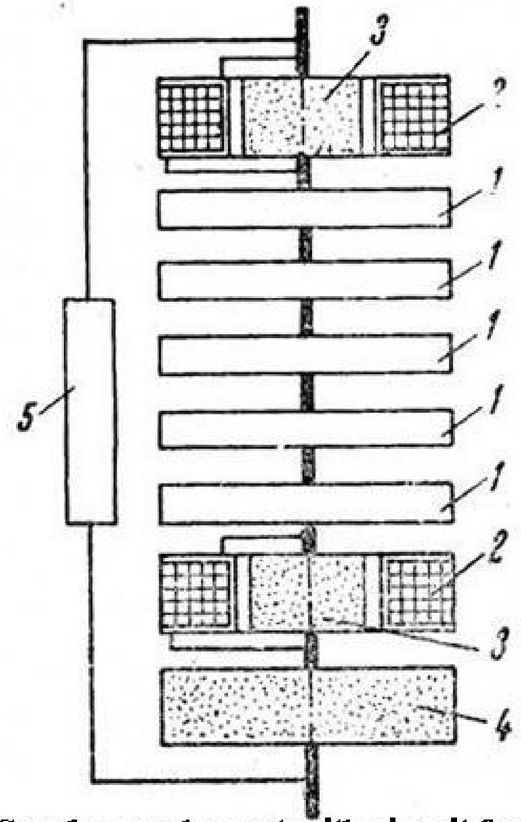

1,- the spark-gap individual devices which work in the manner to extend the electric arc of the accompanying electric current ;
2,- the coils which produce the magnetic field for the magnetic blowing ;
3,- the electric shunt-resistances in parallel connection on the blowing coils, with the role of reducing the current inside this coils
4,- the disc with unlinear resistance ;
5,- electrical resistance for the homogenous distribution of the electric field in all the elements of the spark-gap.

In Jules Guillot scheme, that kind of device could be used for leading the current.

Spark-gap element with circuit for the magnetic blowing of the electric fascicle

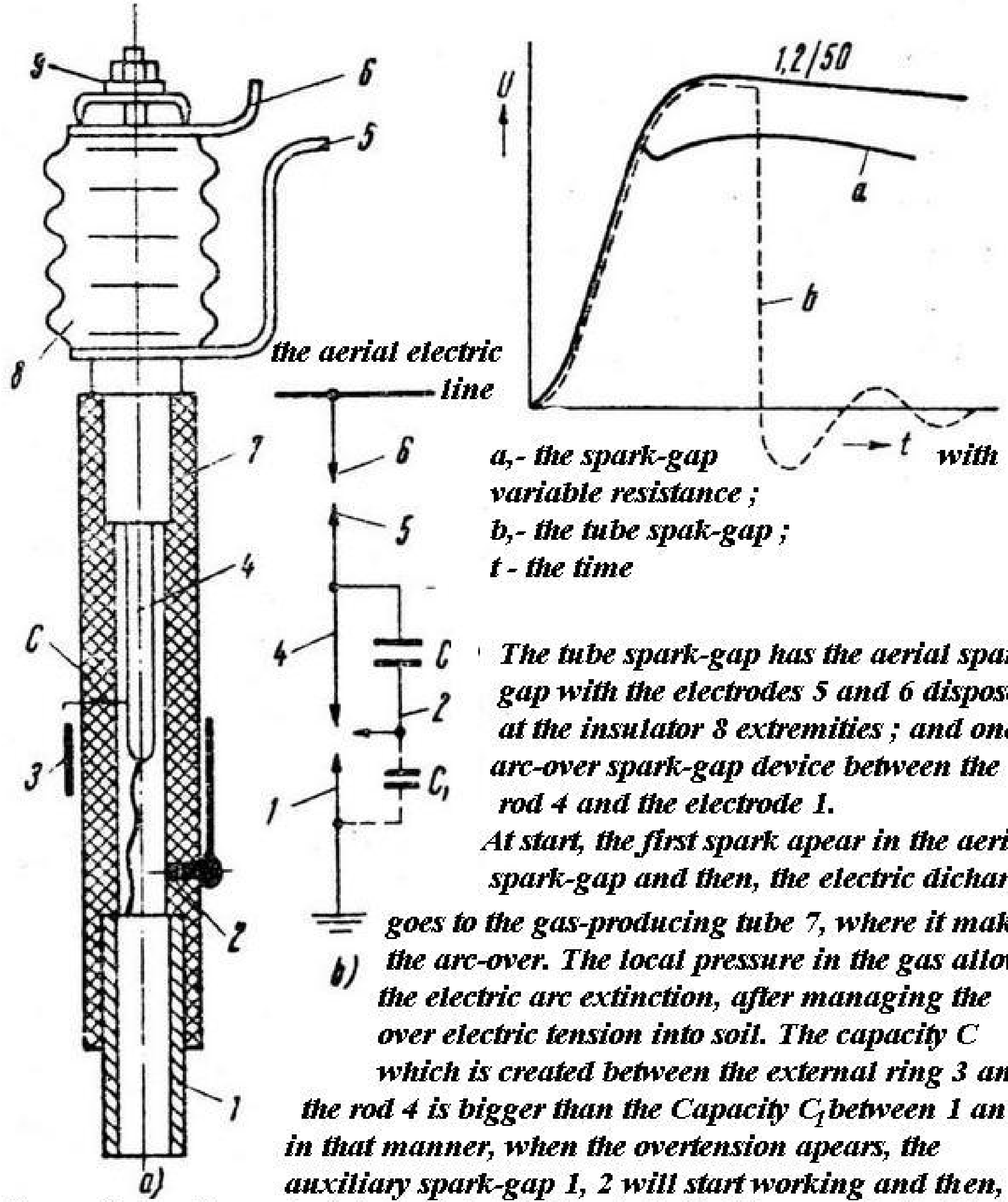

a,- the spark-gap with variable resistance ;
b,- the tube spak-gap ;
t - the time

The tube spark-gap has the aerial spark gap with the electrodes 5 and 6 disposed at the insulator 8 extremities ; and one arc-over spark-gap device between the rod 4 and the electrode 1. At start, the first spark apear in the aerial spark-gap and then, the electric dicharge goes to the gas-producing tube 7, where it makes the arc-over. The local pressure in the gas allow the electric arc extinction, after managing the over electric tension into soil. The capacity C which is created between the external ring 3 and the rod 4 is bigger than the Capacity C_1 between 1 an 2 in that manner, when the overtension apears, the auxiliary spark-gap 1, 2 will start working and then, the main spark-gap 1, 4. In this circuit, the electrode 9 is connected at the aerial electric line and the electrode 1 at the Earth.

In the image above, it is an classic tube spark-gap used for the protection of aerial electric cables. The atmospheric electricity is discharged in soil and lost... The Guillot's inventions used that kind of devices for leading and absorbing of the atmospheric electricity.

DARIDA, Gillbert : "Practical Utilization of Atmospheric Electricity"; (in *The Invention Encyclopedia,* pp. 204-207 ; 1930, Geo. Constantinescu, Ed.) --

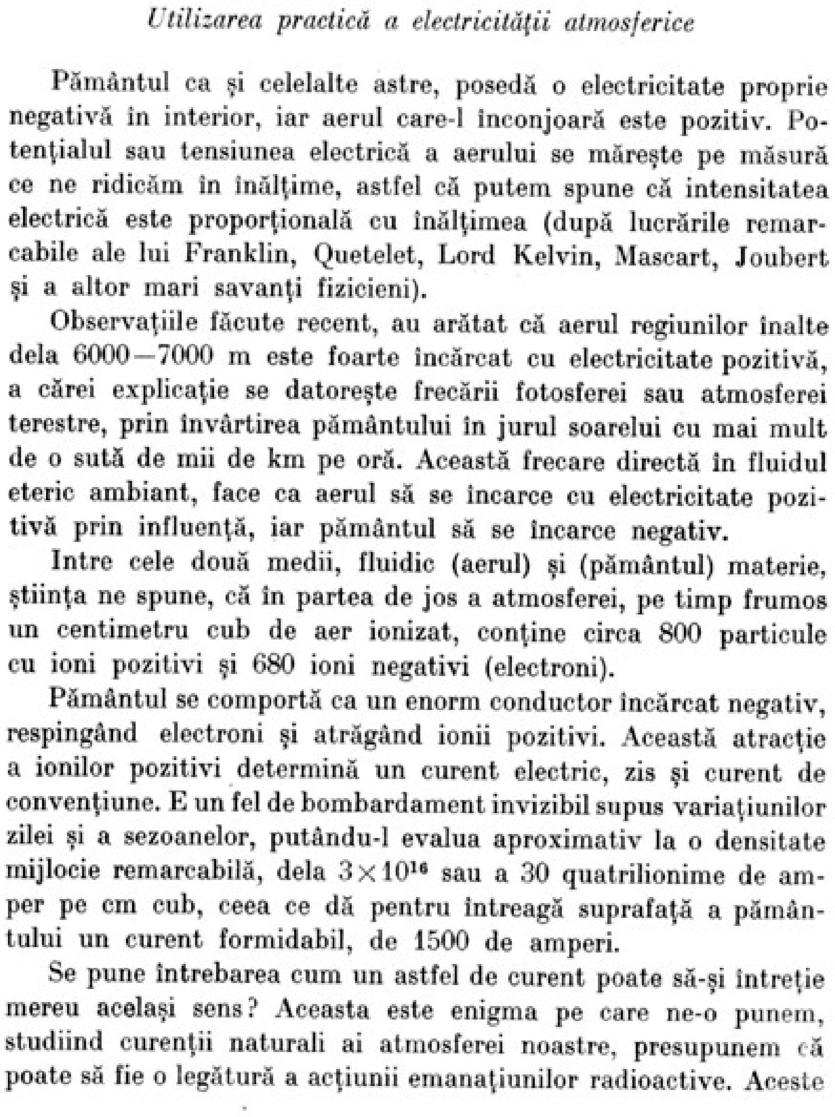

Utilizarea practică a electricităţii atmosferice

Pământul ca şi celelalte astre, posedă o electricitate proprie negativă în interior, iar aerul care-l înconjoară este pozitiv. Potenţialul sau tensiunea electrică a aerului se măreşte pe măsură ce ne ridicăm în înălţime, astfel că putem spune că intensitatea electrică este proporţională cu înălţimea (după lucrările remarcabile ale lui Franklin, Quetelet, Lord Kelvin, Mascart, Joubert şi a altor mari savanţi fizicieni).

Observaţiile făcute recent, au arătat că aerul regiunilor înalte dela 6000—7000 m este foarte încărcat cu electricitate pozitivă, a cărei explicaţie se datoreşte frecării fotosferei sau atmosferei terestre, prin învârtirea pământului în jurul soarelui cu mai mult de o sută de mii de km pe oră. Această frecare directă în fluidul eteric ambiant, face ca aerul să se încarce cu electricitate pozitivă prin influenţă, iar pământul să se încarce negativ.

Intre cele două medii, fluidic (aerul) şi (pământul) materie, ştiinţa ne spune, că în partea de jos a atmosferei, pe timp frumos un centimetru cub de aer ionizat, conţine circa 800 particule cu ioni pozitivi şi 680 ioni negativi (electroni).

Pământul se comportă ca un enorm conductor încărcat negativ, respingând electroni şi atrăgând ionii pozitivi. Această atracţie a ionilor pozitivi determină un curent electric, zis şi curent de convenţiune. E un fel de bombardament invizibil supus variaţiunilor zilei şi a sezoanelor, putându-l evalua aproximativ la o densitate mijlocie remarcabilă, dela 3×10^{16} sau a 30 quatrilionime de amper pe cm cub, ceea ce dă pentru întreagă suprafaţă a pământului un curent formidabil, de 1500 de amperi.

Se pune întrebarea cum un astfel de curent poate să-şi întreţie mereu acelaşi sens? Aceasta este enigma pe care ne-o punem, studiind curenţii naturali ai atmosferei noastre, presupunem că poate să fie o legătură a acţiunii emanaţiunilor radioactive. Aceste emanaţiuni fiind din gaze radioactive grele, se găsesc de obiceiu în mare abundenţă aproape de pământ ceea ce ne explică marea ionizare observată în general în grote şi caverne.

Ionizarea atmosferică poate de asemenea fi produsă în parte, de acţiunea razelor X, foarte pătrunzătoare şi de diferite substanţe radioactive închise în subsolul pământului. De asemenea poate să mai fie, ţinând seama de acţiunea «Calei Lactee» şi de ionizarea luminei solare, foarte bogată în radiaţiuni ultraviolete

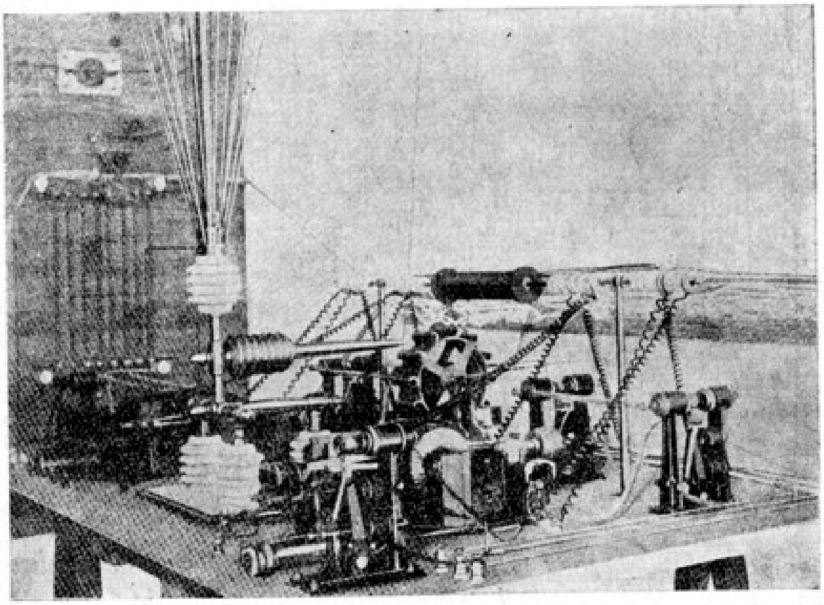

Fig. 108

precum şi a electronilor emişi direct din radierea căldurii solare de circa 6000 de grade.

Captarea electricităţii atmosferice a fost utilizată în Franţa prin cabluri aeriene montate pe Mont Blank, iar în Germania prin cabluri susţinute de baloane captive.

Sunt mai multe sisteme, vom da pe scurt pe cel mai ingenios procedeu, inventat de ing. Jules Guillot, bazat pe sifonul electric. Procedeul consistă de a pompa direct din atmosferă, electricitate

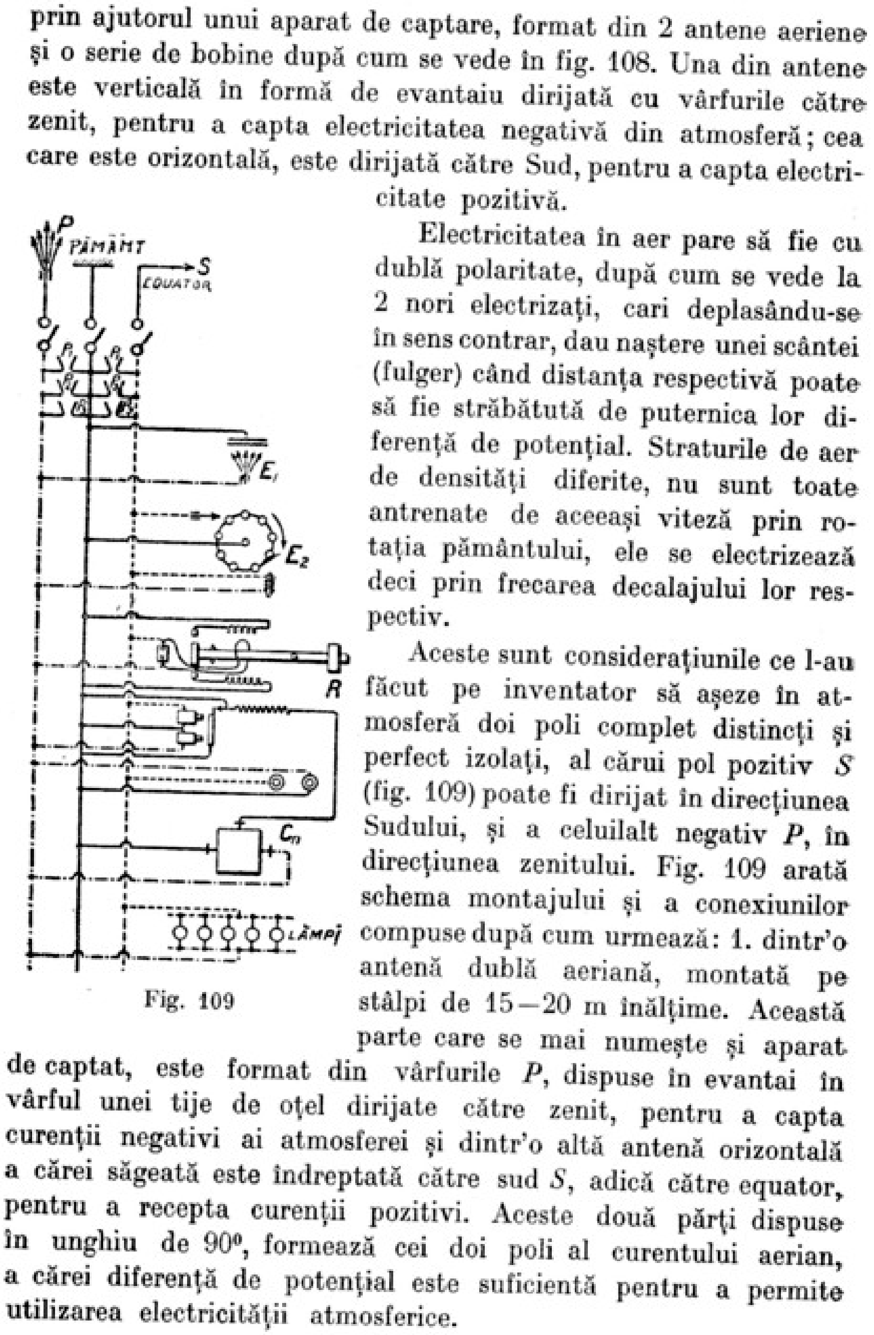
P
PĂMÂNT
S
EQUATOR
E₁
E₂
R
Cₙ
LĂMPI

prin ajutorul unui aparat de captare, format din 2 antene aeriene şi o serie de bobine după cum se vede în fig. 108. Una din antene este verticală în formă de evantaiu dirijată cu vârfurile către zenit, pentru a capta electricitatea negativă din atmosferă; cea care este orizontală, este dirijată către Sud, pentru a capta electricitate pozitivă.

Electricitatea în aer pare să fie cu dublă polaritate, după cum se vede la 2 nori electrizaţi, cari deplasându-se în sens contrar, dau naştere unei scântei (fulger) când distanţa respectivă poate să fie străbătută de puternica lor diferenţă de potenţial. Straturile de aer de densităţi diferite, nu sunt toate antrenate de aceeaşi viteză prin rotaţia pământului, ele se electrizează deci prin frecarea decalajului lor respectiv.

Fig. 109

Aceste sunt consideraţiunile ce l-au făcut pe inventator să aşeze în atmosferă doi poli complet distincţi şi perfect izolaţi, al cărui pol pozitiv *S* (fig. 109) poate fi dirijat în direcţiunea Sudului, şi a celuilalt negativ ***P***, în direcţiunea zenitului. Fig. 109 arată schema montajului şi a conexiunilor compuse după cum urmează: 1. dintr'o antenă dublă aeriană, montată pe stâlpi de 15—20 m înălţime. Această parte care se mai numeşte şi aparat de captat, este format din vârfurile *P*, dispuse în evantai în vârful unei tije de oţel dirijate către zenit, pentru a capta curenţii negativi ai atmosferei şi dintr'o altă antenă orizontală a cărei săgeată este îndreptată către sud *S*, adică către equator, pentru a recepta curenţii pozitivi. Aceste două părţi dispuse în unghiu de 90°, formează cei doi poli al curentului aerian, a cărei diferenţă de potenţial este suficientă pentru a permite utilizarea electricităţii atmosferice.

French Patent # 551,882

Appareil Capteur de Courants Electriques dans L'Atmosphere

No. 674,427. Patented May 21, 1901.

A. PALENCSÁR.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

(Application filed July 10, 1900.)

(No Model.) 3 Sheets—Sheet 1.

Fig:1.

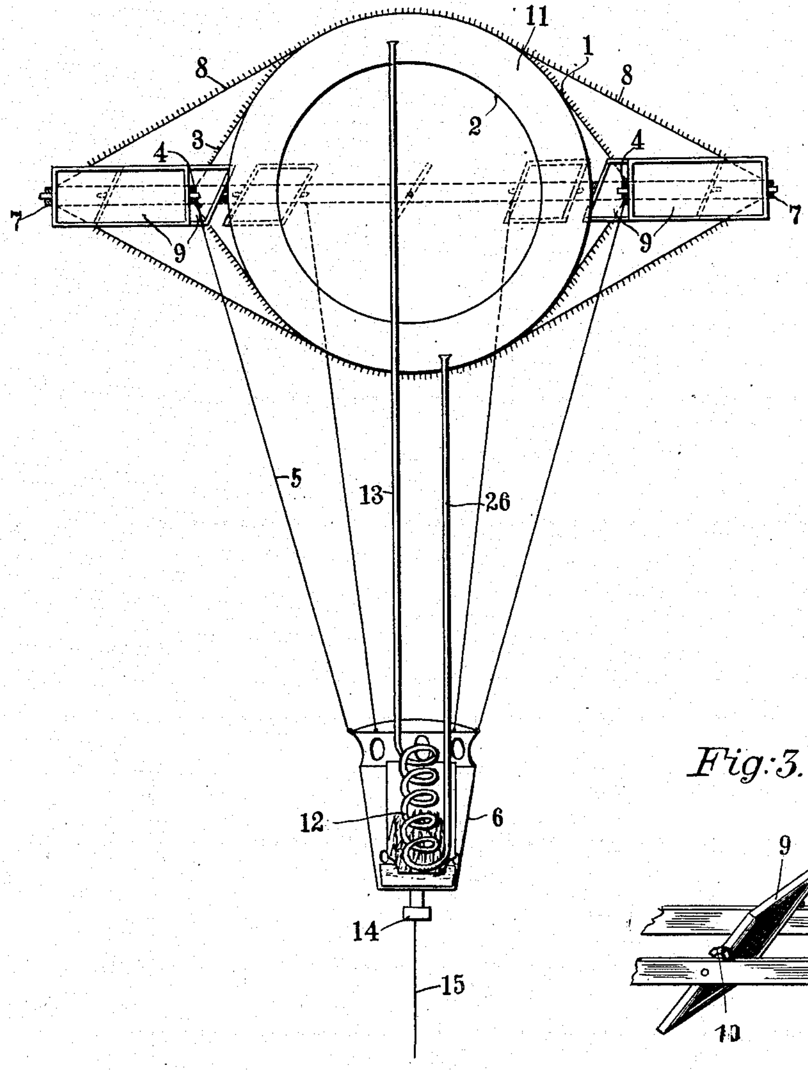

11
1
8
8
2
3
4
4
7
7
9
9
9
9
5
13
26
12
6
14
15

Fig:3.

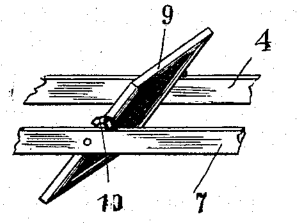

WITNESSES:
Ella L. Giles
Otto Munk

INVENTOR
Andor Palencsár
BY Richards
ATTORNEYS

THE NORRIS PETERS CO., PHOTO-LITHO., WASHINGTON, D. C.

No. 674,427. Patented May 21, 1901.

A. PALENCSÁR.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

(Application filed July 10, 1900.)

(No Model.) 3 Sheets—Sheet 2.

Fig: 2.

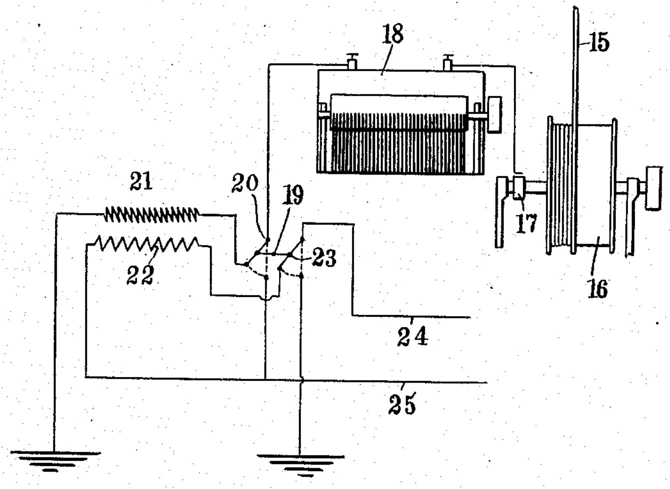

WITNESSES:
Ella L. Giles
Otto Munk

INVENTOR
Andor Palencsár
BY
ATTORNEYS

THE NORRIS PETERS CO., PHOTO-LITHO., WASHINGTON, D. C.

No. 674,427. Patented May 21, 1901.

A. PALENCSÁR.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

(Application filed July 10, 1900.)

(No Model.) 3 Sheets—Sheet 3.

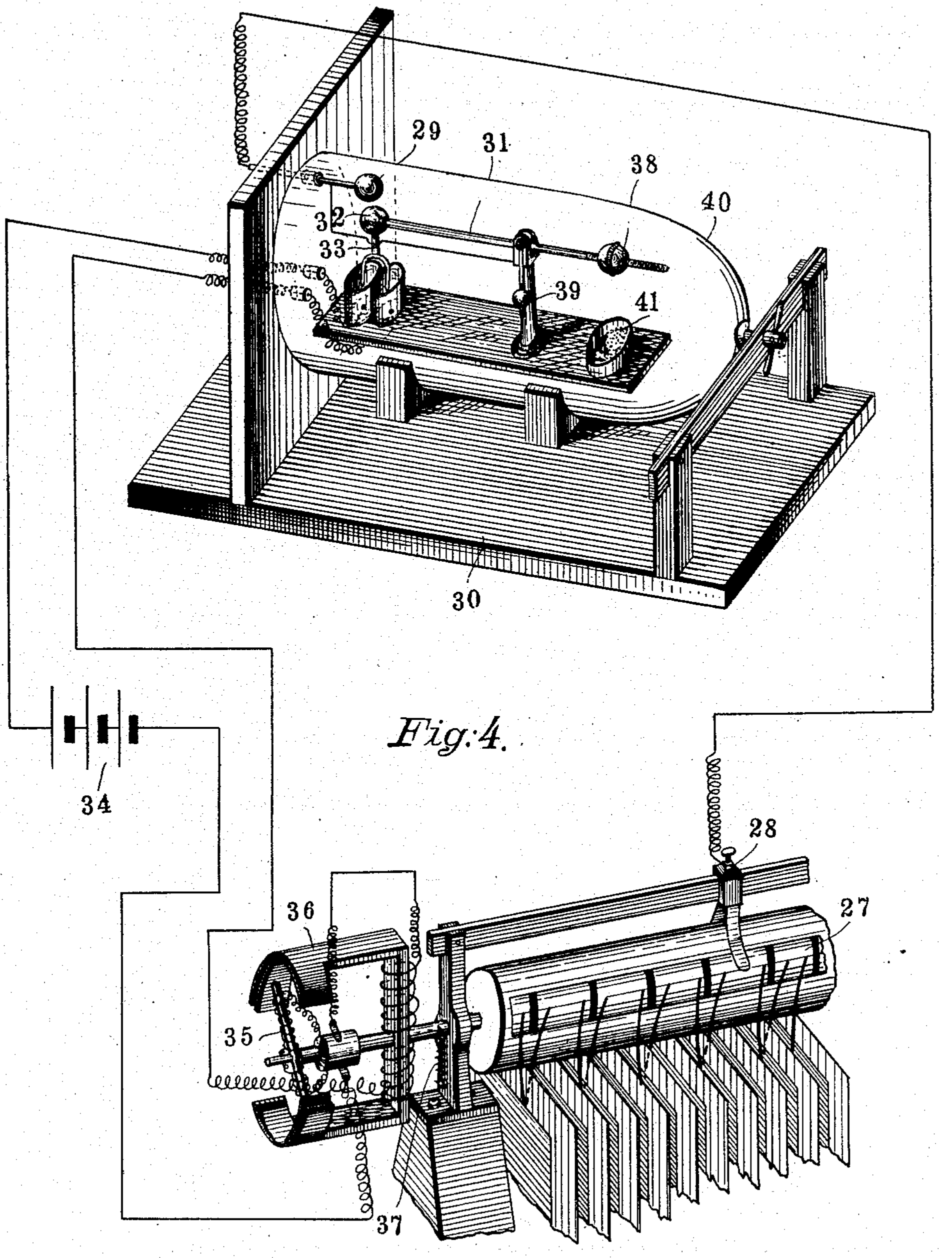
29
31
38
40
32
33
39
41
30
34
28
27
36
35
37

Fig. 4.

WITNESSES:
Ella L. Giles
Otto Munk

INVENTOR
Andor Palencsár
BY
Richards & Co
ATTORNEYS

THE NORRIS PETERS CO., PHOTO-LITHO., WASHINGTON, D. C.

UNITED STATES PATENT OFFICE.

ANDOR PALENCSÁR, OF BUDA-PESTH, AUSTRIA-HUNGARY.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

SPECIFICATION forming part of Letters Patent No. 674,427, dated May 21, 1901.

Application filed July 10, 1900. Serial No. 23,102. (No model.)

To all whom it may concern:

Be it known that I, ANDOR PALENCSÁR, a subject of the King of Hungary, residing at Buda-Pesth, Austria-Hungary, have invented a certain new and useful Apparatus for Collecting Atmospherical Electricity, of which the following is a full, clear, and exact specification.

The present invention concerns an apparatus for collecting and driving or conducting atmospherical electricity which renders it practicable to obtain material quantities of the same in serviceable form.

The experiments made hitherto have been limited to deviating or conducting the electricity by simple lightning-rods without making allowance for the progress of the theory of atmospherical electricity. It is obvious that only extremely small quantities of electricity can be collected in this manner, and, moreover, it was obtained in a form which entirely precluded its utilization. It is evident that a source of electricity of irregular yield can only be utilized by means of accumulators and for charging accumulators only a current of constant potential can be employed. The potential of electricity derived or deviated by means of a lightning-rod varies within wide limits, and, moreover, it is so high that it cannot be used at all for charging accumulators. All these drawbacks are remedied by the present invention, by which the atmospheric electricity is obtained in larger quantities with as low and constant a potential as may be desired.

The idea underlying the invention is based on the modern theory of atmospheric electricity, according to which it is produced by the condensation of steam or aqueous vapors, and that the increase of potential is effected by the concentration of the small drops of water into larger ones, as the proportion of the surface of the drops to the volume of same is materially reduced thereby. According to this theory the water-droplets floating in the layers of air are considered as vehicles or carriers of the electricity, and a rational system of the deviation of the atmospheric electricity must derive it from the water-drops. This is attained by the present invention in the following manner: A collecting-body of as large a surface as possible is provided with sharp needles. This is moved in the higher layers of air, while being continually heated by a suitable heating device. Owing to the heat the water-drops immediately ambient to the collecting-body will be evaporated, their capacity is gradually reduced, while the potential of the charge grows until it reaches an infinite height with infinite smallness of the drops. It is readily apparent that the whole charge of the drops as soon as they have been evaporated will have passed to the collector or collecting-body, from which it can be conducted. For the purpose of replacing the evaporated drops, which have been deprived of their charge by new-charged drops, the collector is moved in relation to the ambient air.

One way of carrying out the invention is represented in diagram in the accompanying drawings, in which—

Figure 1 is a longitudinal section of the collecting-balloon. Fig. 2 shows the deviating and conversion device, and Fig. 3 a detail. Fig. 4 is a diagrammatic perspective view of a form of automatic regulator for the rheostatic machine.

The apparatus consists of a balloon having two walls and covered with a light wire net, preferably of aluminium wire, said net being studded with needles. Besides, the balloon carries the net 3, on which the ring 4, made of a solid but light material, (wood, cane, &c.,) is fixed. This ring carries the basket 6 by means of the cords or ropes 5. On a level with the ring 4 is the ring 7, which is kept spread by the blades or wings 9, which are journaled in a manner to rotate easily. The ring 7 is kept fixedly in position by the cords or ropes 8. The blades or wings consist of a frame covered with light material, and their rotation in either direction is limited by the stops or ledges 10. The blades or wings form advantageously an angle of sixty to seventy degrees with the vertical line.

All wire nets, ropes, rings, and blades or wings may be covered with small metal needles which are electrically connected with each other.

From the space 11, formed between the double walls of the balloon, the pipe 26 leads from the lowest point of the same to the serpentine 12, arranged in the basket 6, which in turn is connected with the pipe 13, ending in the upper part of the space 11 between the two walls 1 2. The serpentine is heated by means of a suitable source of heat, whereby a warm current of gas or air circulates continually between the double walls of the balloon.

Under the basket the ball-bearing 14 is arranged in electrical connection with the wire nets, and its stud is electrically connected with the carefully-insulated light though sufficiently strong cable 15.

On the earth's surface is a winch 16, Fig. 2, by means of which the balloon can be made to ascend or descend as soon as the interior space of the balloon is filled with illuminating gas or hydrogen.

The end of the cable-core is soldered to a collector arranged upon and insulated from the axle of the winch, and the electricity is conducted from this collector by means of a sliding contact. The collecting of the electricity takes place by moving the balloon continuously up and down by means of the winch. In this movement the balloon is turned by means of the wings or blades 9, which are adapted to turn the balloon always in the same direction whether ascending or descending, as in the change from ascending into descending of the balloon, or vice versa, the blades are turned over by the aerial resistance, and thus impart the rotating motion to the balloon in the same direction. In order to avoid torsion of the cable, the ball-bearing 14 is provided. This up-and-down motion and rotating of the balloon accomplishes the purpose of bringing it into contact with as many water particles floating in the air as possible.

As the electricity conducted from the collector-wing 17 possesses a much too high and varying potential for making its direct application practicable, and as it is usual with an irregular source of electricity to first charge accumulators and to further utilize the easily-regulated current of the same only, it becomes necessary to seek to maintain the electricity conducted from the collector 17 for the charging of the accumulators at a constant potential and convert the potential to a much lower one; but as we deal in this case with a direct current ordinary converters cannot be used for this purpose. Moreover, the electricity possesses in this case a much too high potential, so that with the employment of ordinary converters the greatest part of the collected electricity would be lost again. The only practical method for this purpose is the converting by means of the Planté rheostatic machine, by which this high-voltage electricity can be transformed almost without any loss whatever. Thus the question of conversion would be solved, and only the question of maintenance of constant potential remains. This is obtained in the following manner: The rheostatic machine 18 or only a part of the plates is connected with an electrometer of any construction, the movable part of which closes a contact which eventually actuates an electromagnet which effects the switching or reversion of the rheostatic machine. After the switching of the rheostatic machine it is discharged, the potential falls to zero, (*nil*,) and the electrometer resumes its initial position, whereby the current of the electromagnet which effects the reversion is interrupted and the plates of the rheostatic machine are reconnected to potential. The machine is then ready for renewed charge and is again discharged when the determined potential is reached. This action is continually repeated as long as the apparatus is in operation.

A form of the automatic regulator for the rheostatic machine is represented in Fig. 4. 27 is the contact-cylinder of the rheostatic machine, on which, for clearness sake, only the contacts for the charging position of the condenser-plates are shown, while the contacts for the discharging position, which come into action after the contact-cylinder has been turned, are omitted. The coating of the condenser-plates is connected in electrical circuit with the stationary ball 29 and the movable ball 32 of the charge-meter 30. When the charge of the rheostatic machine rises, the ball 32, arranged on the one extremity of the lever 31, is repelled and at a certain stroke actuated by means of the fork 33, which is fastened on the ball 32, and by dipping into a mercury-cup it closes the circuit of the source of current 34. This current passes through the windings of the anchor or armature 35, fixed on the shaft of the contact-cylinder, and it also passes through the electromagnet 36. Thereby the contact-cylinder is turned by a certain angle and the rheostatic machine is thus reversed. Then if the tension decreases by discharge and the repulsion of the balls 29 32 declines so far that the contact is interrupted at 33 the spring 37 turns the contact-cylinder into its normal position and the rheostatic machine is again switched to tension. The regulation of the electrometer is effected by the adjustable weights 38 39. When larger quantities of electricity shall be derived or deviated, two rheostatic machines may operate alternately so, that while the one is being discharged the other can be charged. It is readily apparent that if the capacity of the rheostatic machine is not changed the switching or reversion by the electrometer will always take place at the same potential of the rheostatic machine, and as the number of plates, and hence the proportion of conversion, remains the same the current impulses derived or deviated from the rheostatic machine will also have the same potential.

The irregularities of the source of electricity change the interval of time in which the charges follow each other; but as long as the potential remains constant this has no injurious effect on the charge of the accumulators. The current derived, or deviated from the rheostatic machine can be further transformed by an ordinary converter 21 22, and whenever it is sufficiently constant it can be utilized without the intervention of the accumulators. The converters can be connected or disconnected by means of the double switch 19 20 23.

24 and 25 are the conducting-wires, which run either directly to the place of consumption or to an accumulator-battery.

Having now described my invention, what I claim as new, and desire to secure by Letters Patent, is—

1. An apparatus for collecting atmospheric electricity comprising a collecting-body adapted to be kept in motion, heating means for said body, and a rheostatic machine and a converter connected with the said body electrically, substantially as described.

2. An apparatus for collecting atmospheric electricity for storage comprising a collecting-body adapted to be kept in motion in the ambient air, heating means for said body, a conductor leading from said body, a rheostatic machine connected with the said conductor, an electrometer connected with said rheostatic machine electrically, a contact controlled by the electrometer, and an electromagnet controlled by said contact, said electromagnet controlling the reversing of the rheostatic machine, substantially as described.

3. In combination, the balloon-like collecting-body, having the collecting-points, means for heating the interior space of said body, a conductor leading from the balloon-like body, and electrical devices for receiving the current therefrom, substantially as described.

4. In combination, the balloon-like collecting-body, means for turning the same constantly in one direction in both ascending and descending, conducting means leading from the balloon and electrical devices for receiving the current therefrom, substantially as described.

In witness whereof I have hereunto set my hand in presence of two witnesses.

ANDOR PALENCSÁR.

Witnesses:
EUGÈNE HARRÄNJO,
PAUL BÖLESKEY.

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 1

Fig. 1.

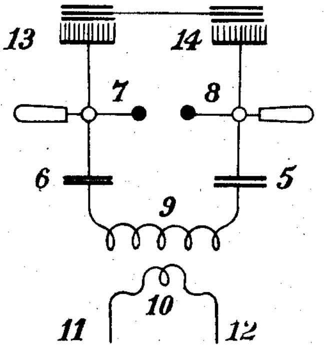
13
14
7
8
6
5
9
10
11
12

Fig. 2.

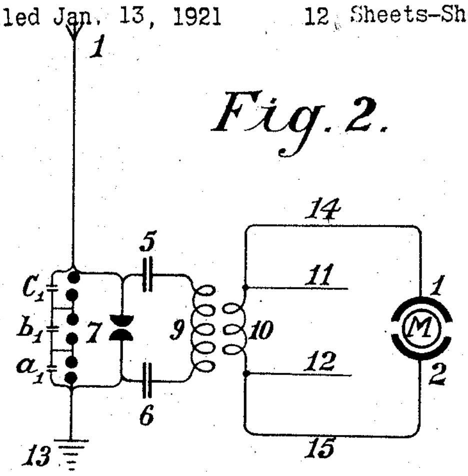
1
14
5
C_1
b_1
7
9
10
11
a_1
12
6
13
15
M
1
2

Fig. 3.

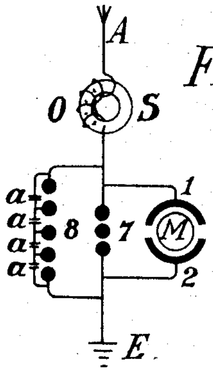
A
O
S
a
a
a
a
8
7
1
M
2
E.

Fig. 4.

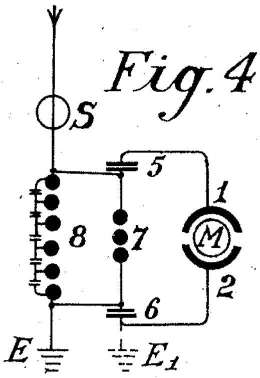
S
5
1
8
7
M
2
6
E
E_1

Fig. 5.

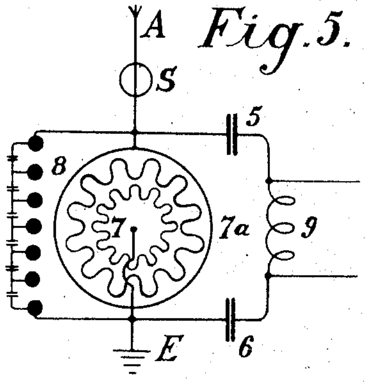
A
S
5
8
7
7a
9
E
6

Fig 6

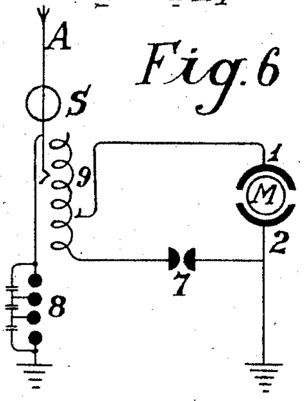

Inventor
Hermann Plauson

by Knight Bros
attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 2

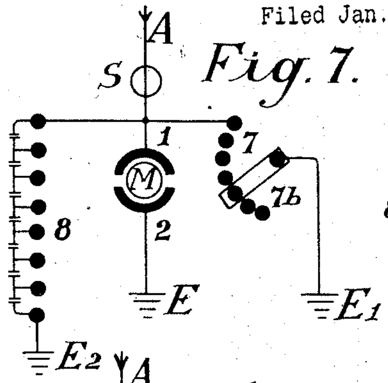
A
S
1
M
2
7
7b
8
E
E1
E2

Fig. 7.

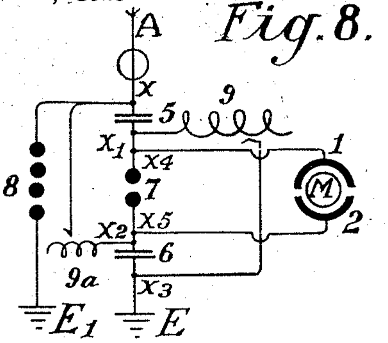
A
X
5
9
1
X1
X4
M
7
8
X5
2
X2
6
9a
X3
E1
E

Fig. 8.

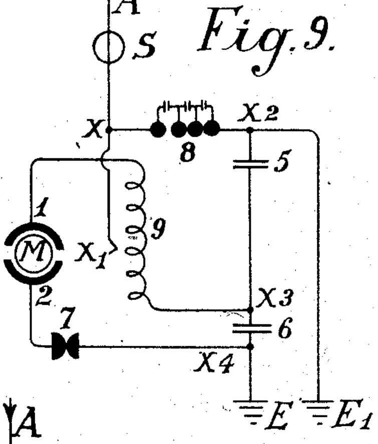
A
S
X
8
X2
5
1
M
X1
9
2
X3
7
6
X4
E
E1

Fig. 9.

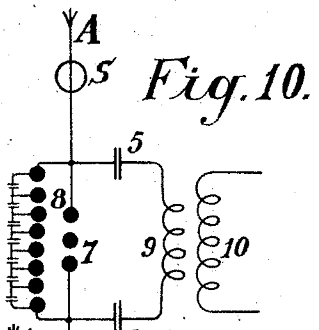
A
S
5
8
7
9
10
6
E

Fig. 10.

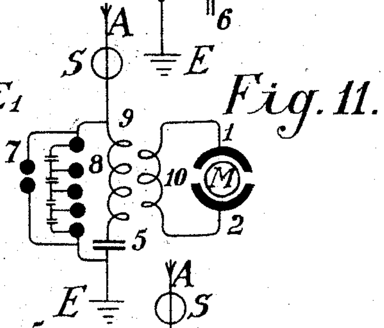
A
S
9
1
7
8
10
M
5
2
E

Fig. 11.

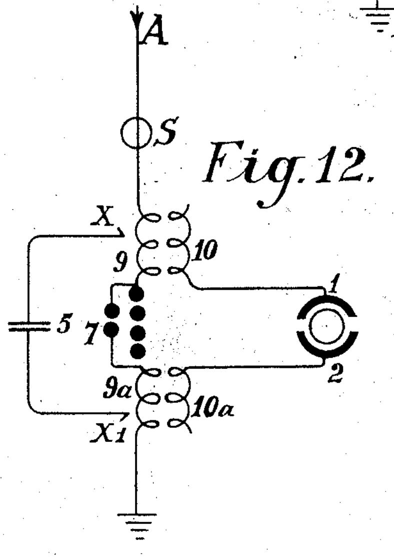
A
S
X
10
9
1
5
7
2
9a
10a
X1

Fig. 12.

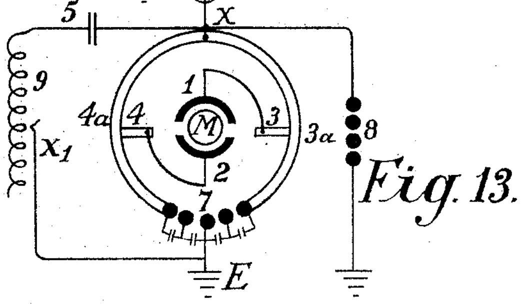
A
S
X
5
9
1
4a
4
M
3
3a
8
X1
2
7
E

Fig. 13.

Inventor

Normann Plauson

by Knight Bros attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 3

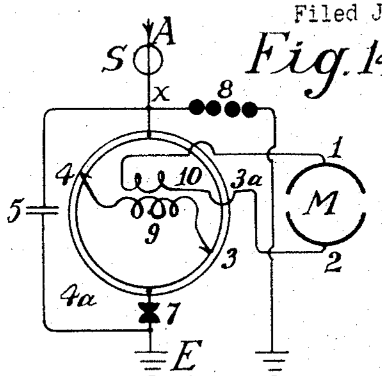
Fig. 14.

A
S
x
8
1
M
2
10
3a
4
5
9
3
4a
7
E

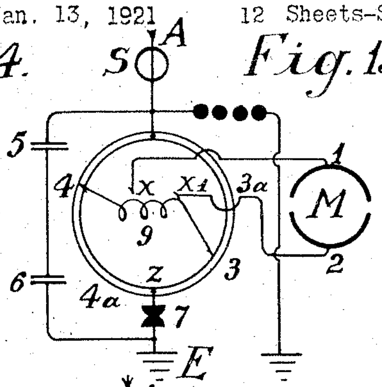
Fig. 15.

A
S
5
1
M
2
4
x
x1
3a
9
3
z
6
4a
7
E

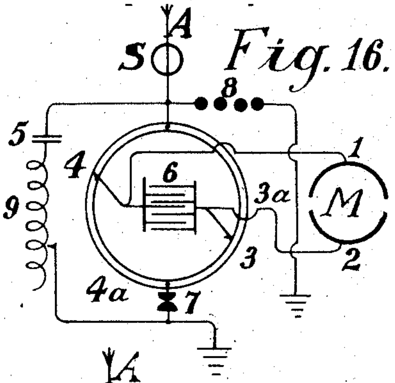
Fig. 16.

A
S
8
5
4
6
1
M
3a
9
2
3
4a
7

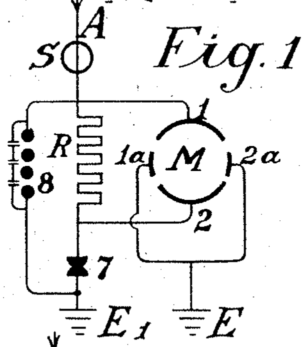
Fig. 17.

A
S
R
1
1a
M
2a
8
2
7
E1
E

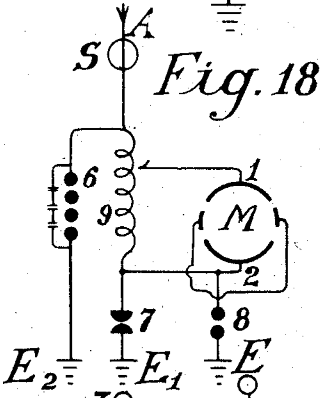
Fig. 18.

A
S
6
9
1
M
2
7
8
E2
E1
E

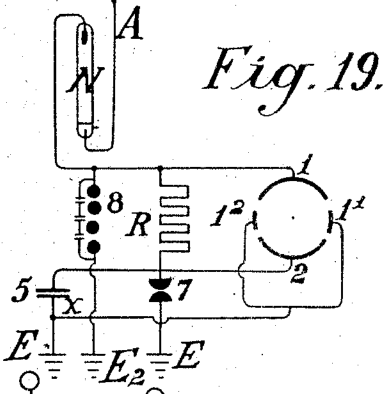
Fig. 19.

A
N
1
8
R
1²
1¹
2
5
x
7
E
E2
E

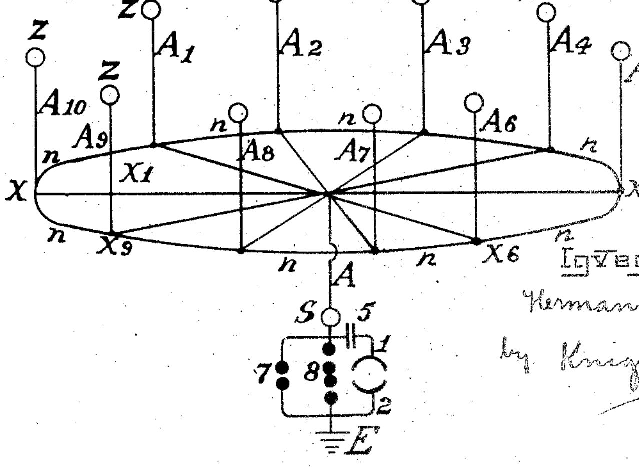
Fig. 20.

z
A1
A2
A3
A4
A5
z
A10
z
A9
n
A8
n
A7
A6
n
n
X
X1
X5
n
X9
n
A
n
X6
S
5
1
7
8
2
E

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 4

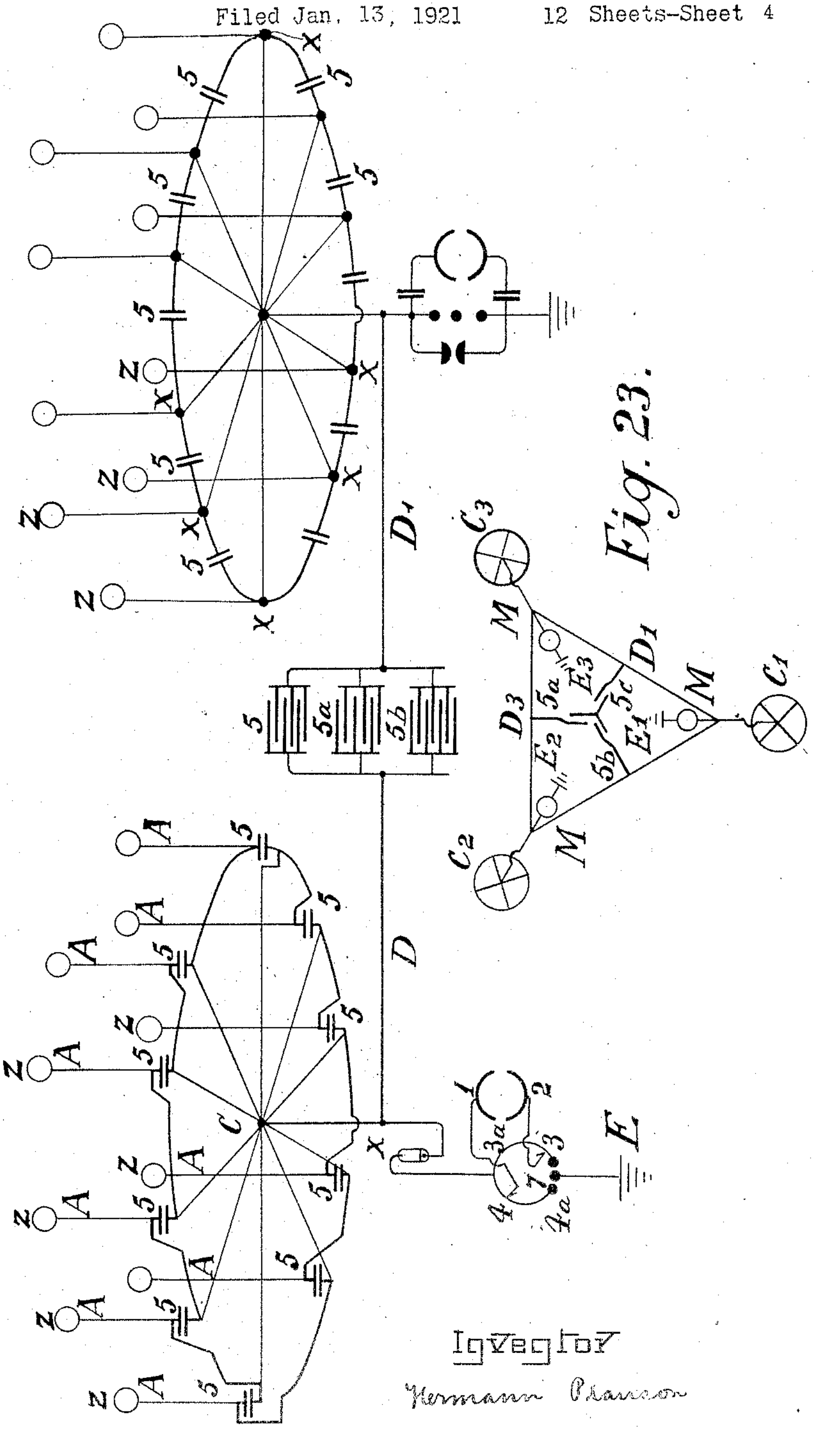
5
5a
5b
5c
X
Z
A
C
D
D1
D3
E
E1
E2
E3
M
C1
C2
C3
1
2
3
3a
4
4a
7

Fig. 22.

Fig. 21.

Fig. 23.

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 5

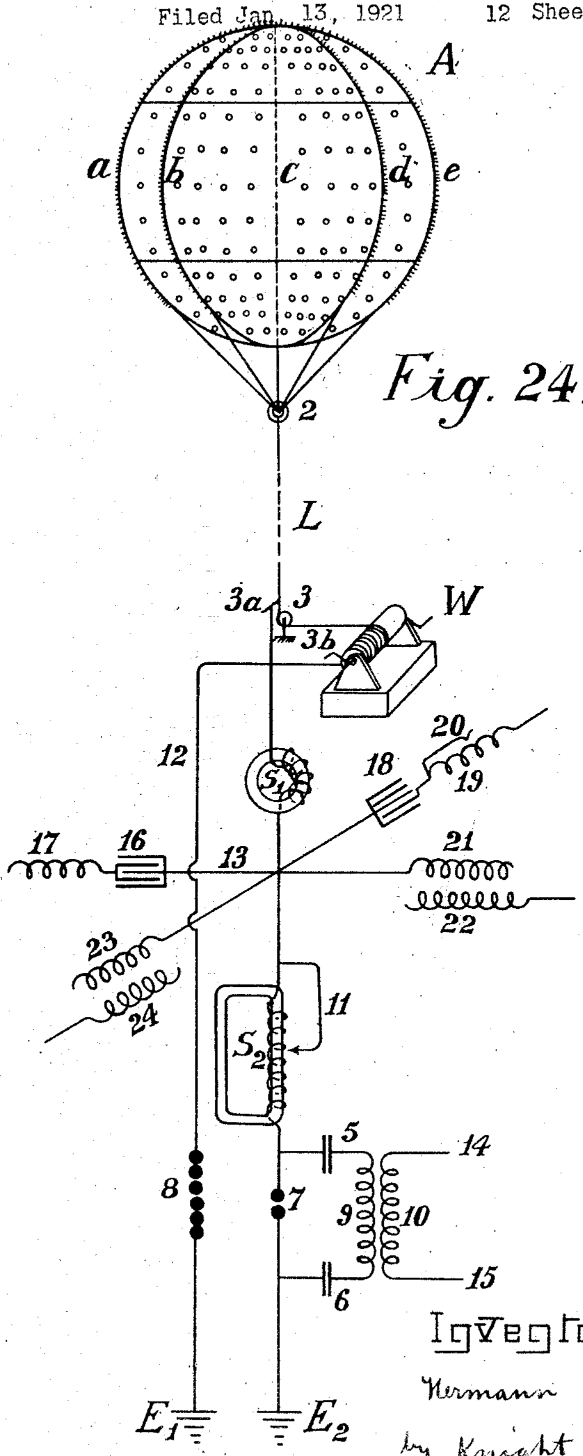
A
a
b
c
d
e
2
L
3a
3
3b
W
12
20
19
18
S_1
17
16
13
21
22
23
24
11
S_2
5
14
8
7
9
10
6
15
E_1
E_2

Fig. 24.

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 6

Fig. 25.

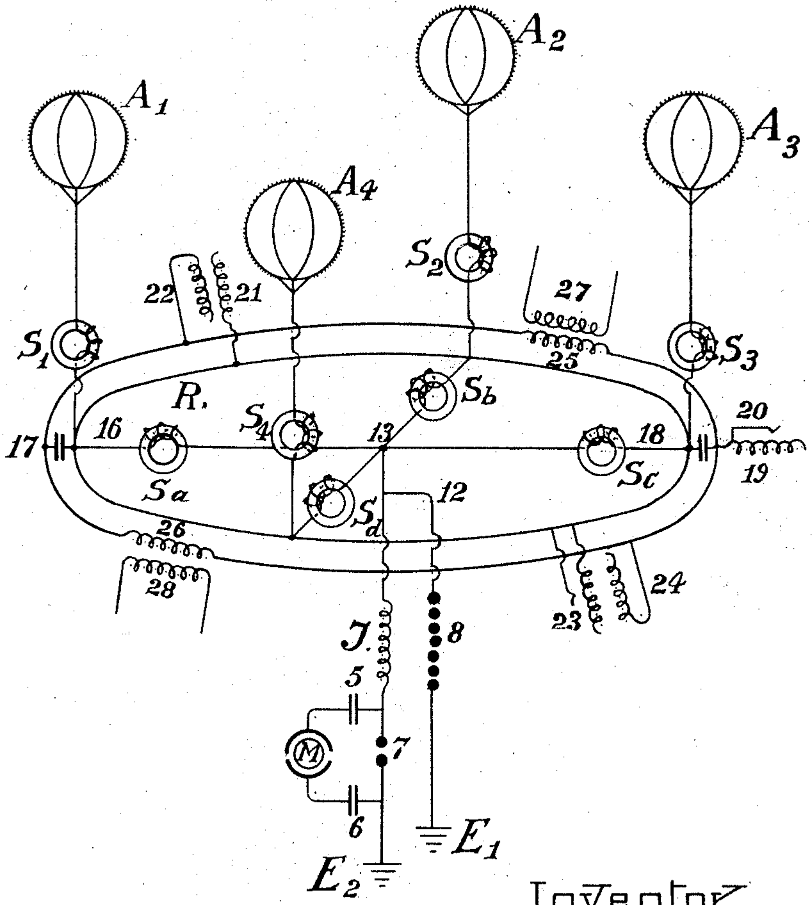

A_2
A_1
A_3
A_4
S_2
27
22
21
S_1
25
S_3
R.
S_b
S_4
20
17
16
13
18
19
S_a
S_c
12
S_d
26
28
24
23
J.
8
5
M
7
6
E_1
E_2

Inventor

Hermann Plauson

by

Knight Bros

attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 7

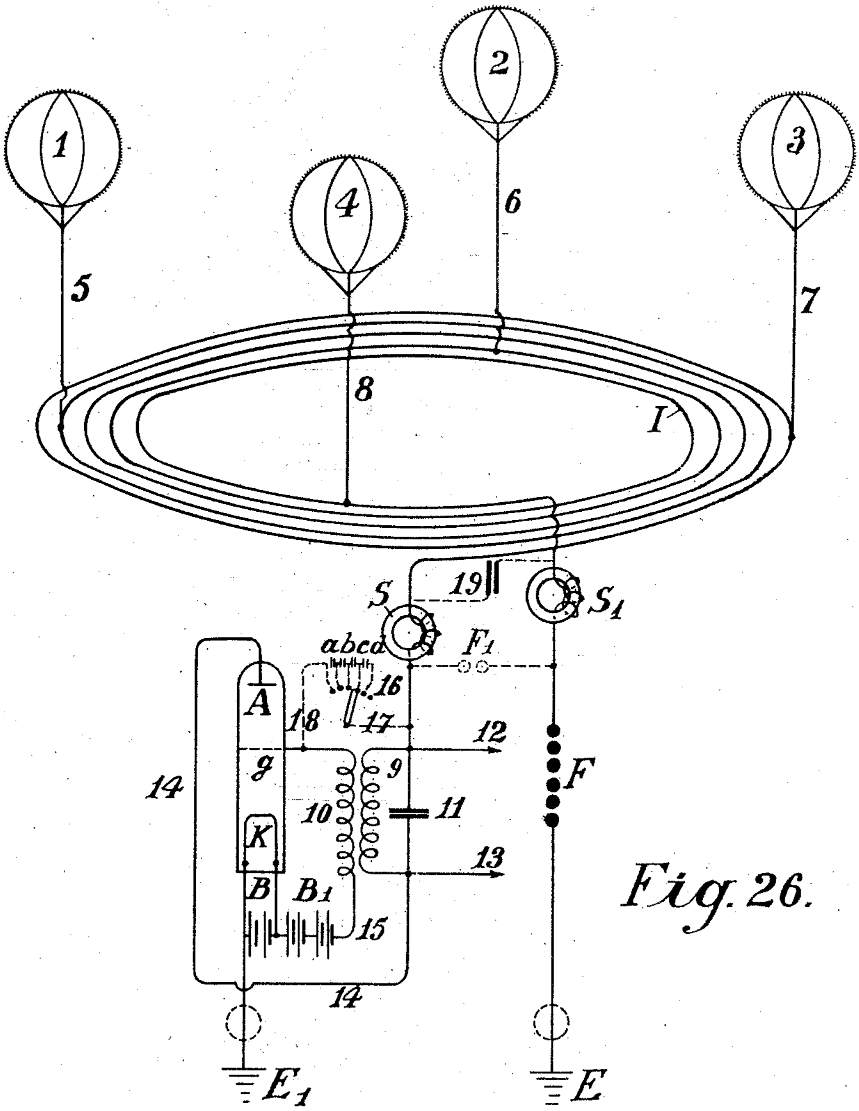
1
2
3
4
5
6
7
8
I
19
S
S1
abcd
F1
16
A
18
17
12
9
g
14
10
11
F
K
13
B
B1
15
14
E1
E

Fig. 26.

Inventor
Hermann Plauson
by
Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921. 12 Sheets-Sheet 8

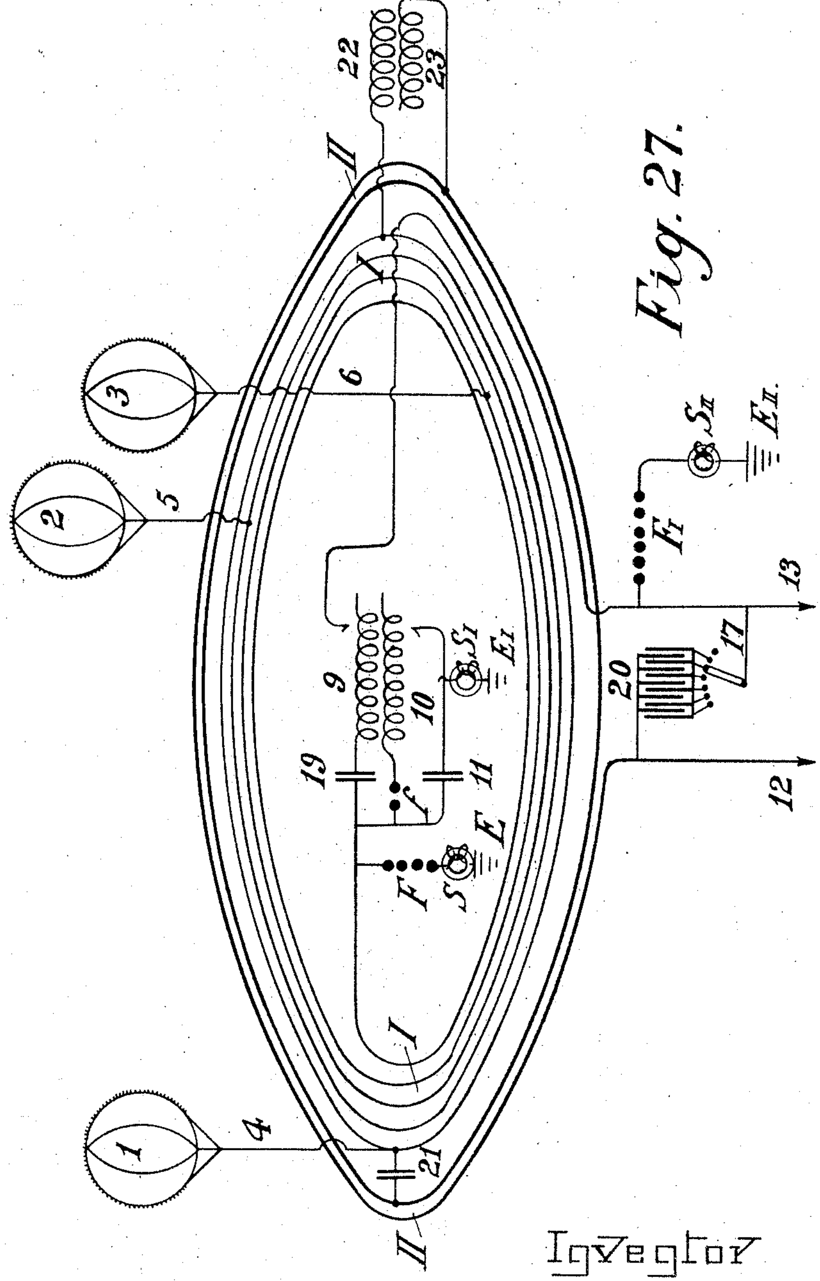
22
23
II
I
6
3
2
5
1
4
21
II
I
9
10
S_I
E_I
19
11
F
E
S
F
20
17
13
12
F_I
S_{II}
E_{II}

Fig. 27.

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 9

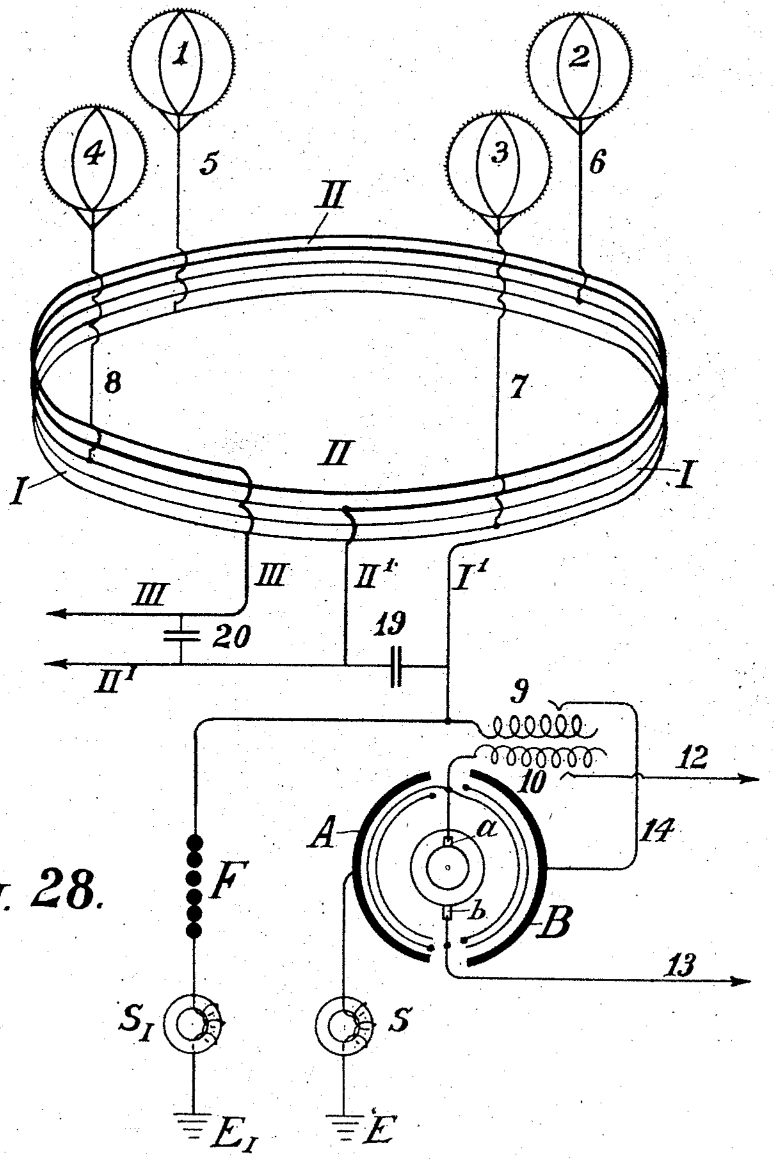
1
2
4
3
5
6
II
8
7
II
I
I
III
II¹
I¹
III
20
19
II¹
9
12
10
A
a
14
B
b
F
13
S_I
S
E_I
E

Fig. 28.

Inventor
Hermann Plauson
by
Knight Bros
attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921. 12 Sheets-Sheet 10

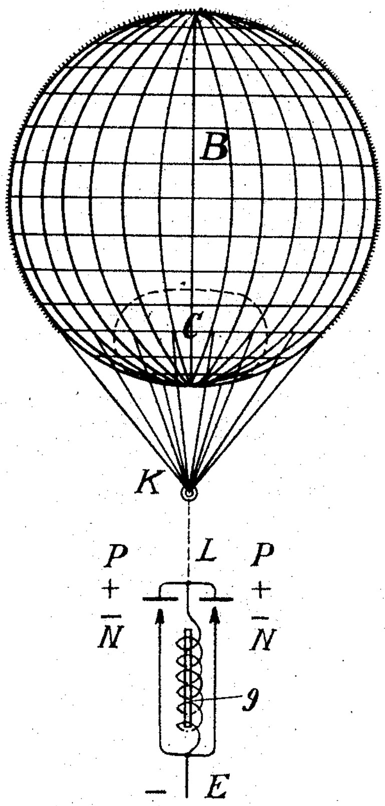
B
C
K
P
L
P
+
+
N
N
9
−
E

Fig. 29.

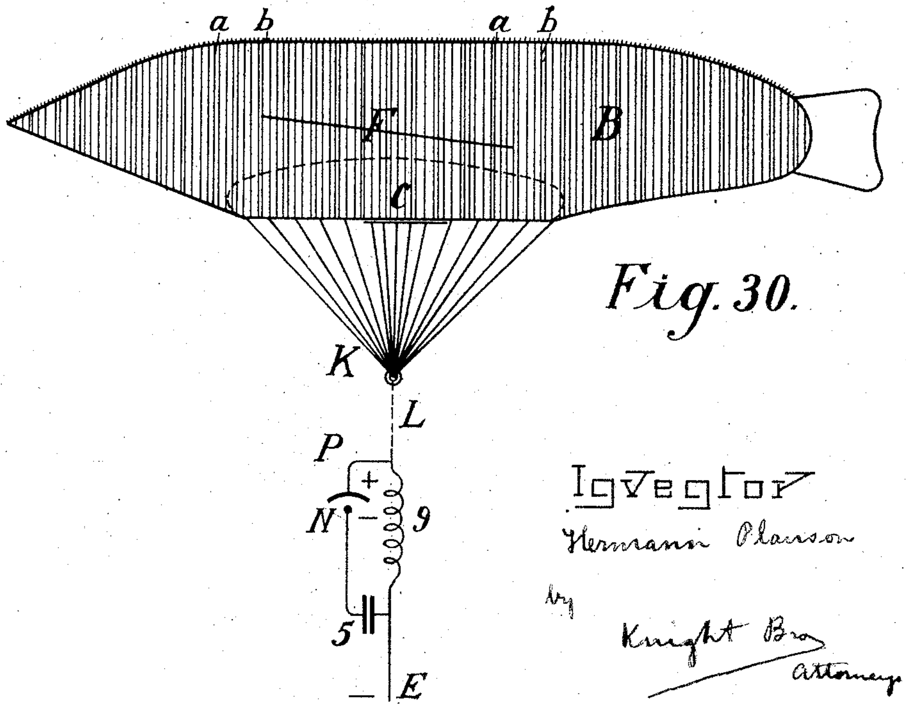
a
b
a
b
F
B
C
K
L
P
+
N
−
9
5
−
E

Fig. 30.

Inventor
Hermann Plauson
by
Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 11

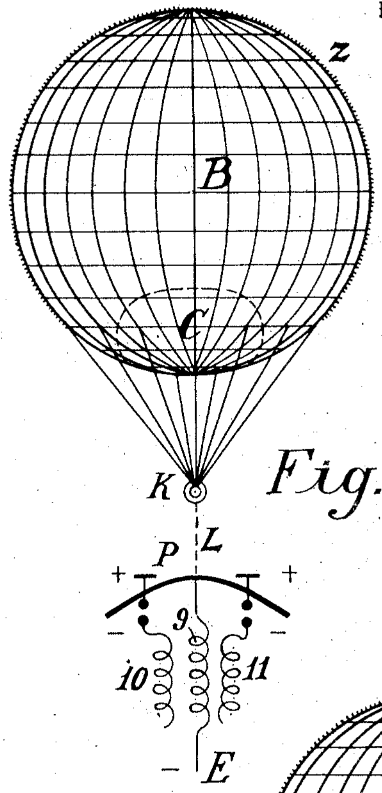
z
B
C
K
P
L
+
+
–
–
9
10
11
–
E

Fig. 31.

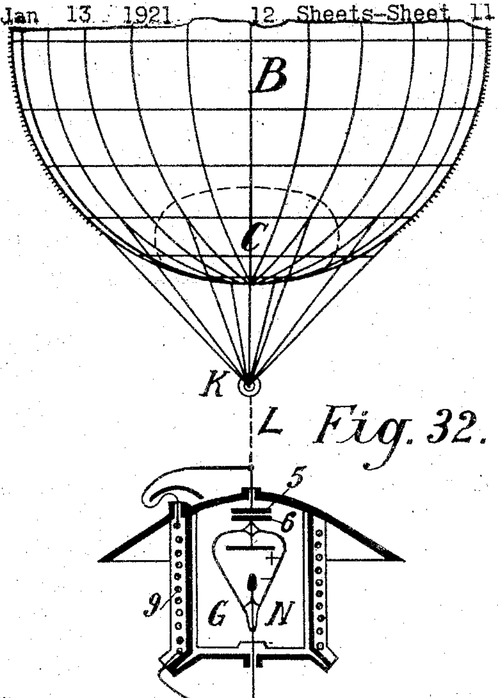
z
B
C
K
L
5
6
+
–
9
G
N
–
E

Fig. 32.

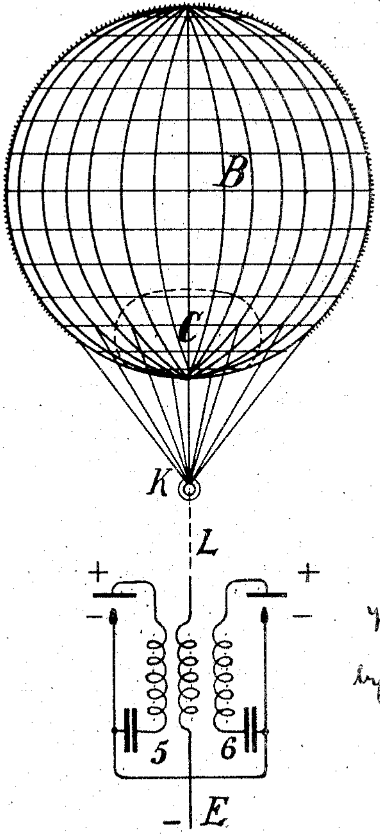

Fig. 33.

Inventor
Hermann Plauson
by Knight Bros
attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 12

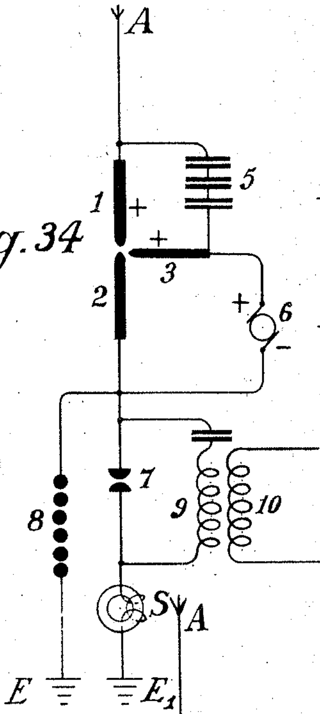
A
1
+
5
+
3
2
+
6
-
7
8
9
10
S
E
E1

Fig. 34

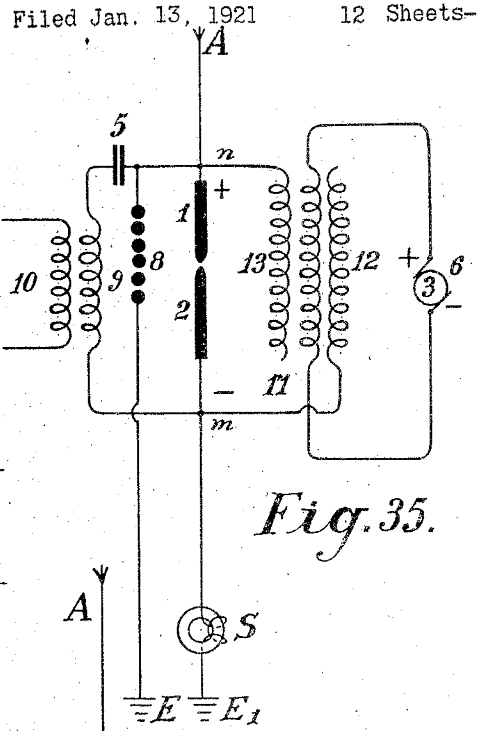
A
5
n
+
1
8
9
10
13
12
+
3
6
-
2
-
11
m
S
E
E1

Fig. 35.

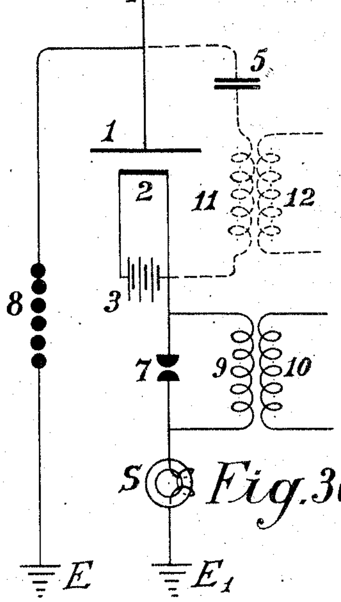
A
5
1
2
11
12
8
3
7
9
10
S
E
E1

Fig. 36.

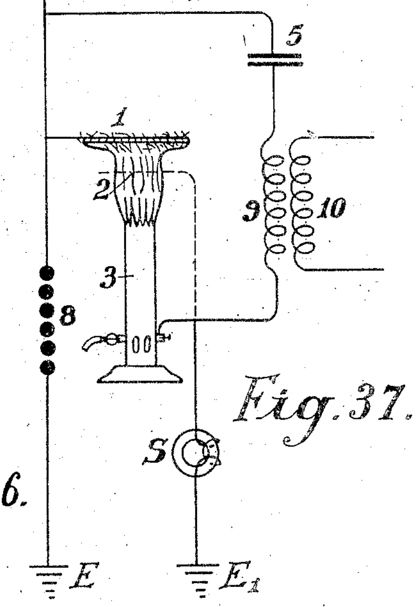
A
5
1
2
3
9
10
8
S
E
E1

Fig. 37.

Inventor

Hermann Plauson

by Knight Bros

Attorneys

Patented June 9, 1925. **1,540,998**

UNITED STATES PATENT OFFICE.

HERMANN PLAUSON, OF HAMBURG, GERMANY.

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY.

Application filed January 13, 1921. Serial No. 437,107.

To all whom it may concern:

Be it known that I, HERMANN PLAUSON, Esthonian subject, residing in Hamburg, Germany, have invented certain new and useful Improvements in the Conversion of Atmospheric Electric Energy, of which the following is a specification.

Methods of obtaining atmospheric electricity by means of metallic nettings set with spikes which are held by means of ordinary or anchored kite balloons made of fabrics and filled with hydrogen, are in theory already known. Atmospheric electricity obtained in this way has been suggested to be used in the form of direct current for the charging of accumulators. This knowledge however is at present only theoretical as the conversion in practice has hitherto been a failure. No means are known of protecting the apparatus from destruction by lightning. The balloons used for collecting the charge must also be made of very large size in order to be able to support the weight of the metallic netting and the heavy cable connections.

Instead of using heavy metallic netting as collectors attached to single air balloons of non-conducting materials which are liable to be torn and are permeable to the gas, it is proposed to use metallic balloon collectors which have the following important advantages—

(*a*) The metallic cases are impenetrable to helium and hydrogen; they also represent large metallic weather-proof collecting surfaces.

(*b*) Radio active means and the like may be easily applied internally or externally; whereby the ionization is considerably increased and therewith also the quantity of atmospheric electricity capable of being collected.

(*c*) Such balloon collectors of light metal do not require to be of large size as they have to carry only their own moderate weight, and that of the conducting cable or wire.

(*d*) The entire system therefore offers little surface for the action of storm and wind and is resistant and stable.

(*e*) Each balloon can be easily raised and lowered by means of a winch so that all repairs, recharging and the like can be carried out without danger during the operation.

It is further proposed to use a collecting aerial network of several separate collectors spread out in the air above the earth, which collectors are interconnected by electrical conductors.

According to this invention charges of atmospheric electricity are not directly converted into mechanical energy, and this forms the main difference from previous inventions, but the static electricity which runs to earth through aerial conductors in the form of direct current of very high voltage and low current strength is converted into electro-dynamic energy in the form of high frequency vibrations. Many advantages are thereby obtained and all disadvantages avoided.

The very high voltage of static electricity of a low current strength can be converted by this invention to voltages more suitable for technical purposes and of greater current strength. By the use of closed oscillatory circuits it is possible to obtain electromagnetic waves of various amplitude and thereby to increase the degree of resonance of such current. Such resonance allows various values of inductance to be chosen whereby again the governing of the starting and stopping of machines driven thereby by simply tuning the resonance between coils of the machine and the transformer circuit forming the resonance can easily be obtained. Further, such currents have the property of being directly available for various uses, even without employing them for driving motors, of which there may be particularly mentioned, lighting, production of heat and use in electro-chemistry.

Further, with such currents a series of apparatus may be fed without direct current supply through conductors and also the electro-magnetic high frequency currents may be converted by means of special motors adapted for electro-magnetic oscillations into mechanical energy, or finally converted by special machines into alternating current of low frequency or even into direct current of high potential.

The invention is more particularly described with reference to the accompanying diagrams in which:—

Figure 1 is an explanatory figure.

Figure 2 is a diagrammatic view of the simplest form.

Figure 3 shows a method of converting atmospheric electrical energy for use with motors.

Figure 4 is a diagram showing the use of protective means.

Figure 5 is a diagram of an arrangement for converting large current strengths.

Figure 6 is a diagram of an arrangement including controlling means.

Figure 7 shows means whereby the spark gap length can be adjusted.

Figure 8 shows a unipolar connection for the motor.

Figure 9 shows a weak coupled system suitable for use with small power motors.

Figures 10, 11 and 12 show modified arrangements.

Figure 13 shows a form of inductive coupling for the motor circuit.

Figure 14 is a modified form of Figure 13 with inductive coupling.

Figure 15 is an arrangement with non-inductive motor.

Figure 16 is an arrangement with coupling by condenser.

Figures 17, 18 and 19 are diagrams of further modifications.

Figure 20 shows a simple form in which the aerial network is combined with special collectors.

Figure 21 shows diagrammatically an arrangement suitable for collecting large quantities of energy.

Figure 22 is a modified arrangement having two rings of collectors.

Figure 23 shows the connections for three rings of collectors.

Figure 24 shows a collecting balloon and diagram of its connection of condenser batteries.

Figures 25 and 26 show modified collector balloon arrangements.

Figure 27 shows a second method of connecting conductor for the balloon aerials.

Figure 28 shows an auto-transformer method of connection.

Figure 29 shows the simplest form of construction with incandescent cathode.

Figure 30 shows a form with cigar shaped balloon.

Figure 31 is a modified arrangement.

Figure 32 shows a form with cathode and electrode enclosed in a vacuum chamber.

Figure 33 is a modified form of Figure 32.

Figure 34 shows an arc light collector.

Figure 35 shows such an arrangement for alternating current.

Figure 36 shows an incandescent collector with Nernst lamp.

Figure 37 shows a form with a gas flame.

Figure 1 illustrates a simple diagram for converting static electricity into dynamic energy of a high number of oscillations. For the sake of clearness in the drawings an influence machine is assumed to be employed and not an aerial antenna. 13 and 14 are combs for collecting the static electricity of the influence machine. 7 and 8 are spark discharging electrodes, 6 and 5 condensers, 9 an inductive primary coil, 10 secondary coil, 11 and 12 ends of conductors of the secondary coil 10. When the disc of the static influence machine is rotated by mechanical means, the combs collect the electric charges one the positive and the other the negative, and charge the condensers 5 and 6 until such a high potential is formed across the spark gap 7—8, that the spark gap is jumped. As the spark gap 7—8 forms a closed circuit with condensers 6 and 5, and inductive resistance 9, as is well known, waves of high frequency electromagnetic oscillations will pass in this circuit.

The high frequency of the oscillations produced in the primary circuit induces waves of the same periodicity in the secondary circuit. Thus in the primary circuit electromagnetic oscillations are formed by the passage of the spark over the spark gap and these waves are maintained by fresh charges of static electricity.

By suitably selecting the ratio between the number of the coils in the primary and secondary circuits with regard to a correct application of the co-efficients of resonance (capacity, inductance, and resistance) the high voltage of the primary circuit may be suitably converted into low voltage and high current strength.

When the oscillatory discharges in the primary circuit becomes weaker or entirely cease, the condensers are charged again by the static electricity until the accumulated charge again breaks down the spark gap. All this is repeated as long as electricity is produced by the static machine by employing mechanical energy.

An elementary form of the invention is shown in Figure 2 in which two spark gaps in parallel are used one of which may be termed the working gap 7 in Figure 2, whilst the second serves as a safety device for excess voltage and consists of a larger number of spark gaps than the working section, which gaps are arranged in series and are bridged by very small capacities as is illustrated in a_1, b_1, c_1, Figure 2 which allow of uniform sparking in the safety section.

In Figure 2 A is the aerial antenna for collecting charges of atmospheric electricity, 13 is the earth connection of the second part of the spark gap, 5 and 6 are condensers, 9 a primary coil. Now when through the aerial A the positive atmospheric electricity seeks to combine with the negative charge to earth, this is prevented by (the air gap between) the spark gaps. The resistance of the spark gap 7 is, as shown in the drawings, lower than that of the other safety section which consists of three spark gaps connected in series, and consequently a three times greater air resistance is offered by the latter.

So long therefore, as the resistance of the spark gap 7 is not overloaded, so that the other spark gaps have an equal resistance with it the discharges take place only over spark gap 7. Should however the voltage be increased by any influences so that it might be dangerous for charging the condensers 5 and 6 or for the coil insulation 9 and 10 in consequence of break down, by a correct regulation of this spark gap the second spark gap can discharge free from inductive effects direct to earth without endangering the machine.

Without this second spark gap, arranged in parallel having a higher resistance than the working spark gap it is impossible to collect and render available large quantities of electrical energy.

The action of this closed oscillation circuit consisting of spark gap 7, two condensers 5 and 6, primary coil 9, and also secondary coil 10 is exactly the same as the one described in Figure 1 with the arrangement of the static induction machine with the only difference that here the second spark gap is provided. The electromagnetic high frequency alternating current obtained can be tapped off from the conductors 11 and 12 for lighting and heating purposes. Special kinds of motors adapted for working with these peculiar electrical charges may be connected at 14 and 15 which can work with static electricity charges or with high frequency oscillations.

In addition to the use of spark gaps in parallel a second measure of security is also necessary for taking off the current. This precaution consists according to this invention, in the introduction of and method of connecting certain protective electromagnets or choking coils in the aerial circuit as shown by S in Figure 3.

A single electromagnet only having a core of the thinnest possible separate laminations is connected with the aerial.

In the case of high voltages in the aerial network or at places where there are frequent thunder storms, several such magnets may however be connected in series.

In the case of large units or plants several electromagnets can be employed in parallel or in series parallel.

The windings of these electromagnets may be simply connected in series with the aerials. In this case the winding preferably consists of several thin parallel wires, which make up together, the necessary section.

The winding may be made of primary and secondary windings in the form of a transformer. The primary winding will be then connected in series with the aerial network, and the secondary winding more or less short-circuited over a regulating resistance or an induction coil. In the latter case it is possible to regulate to a certain extent the effect of the choking coils. In the further description of the connecting and constructional diagrams the aerial electromagnet choke coil is indicated by a simple ring S.

Figure 3 shows the simplest way of converting atmospheric electricity into electromagnetic wave energy by the use of special motors adapted for high oscillatory currents or static charges of electrical energy. Recent improvements in motors for working with static charges and motors working by resonance, that is to say, having groups of tuned electromagnetic cooperating circuits render this possible but such do not form part of the present invention.

A motor adapted to operate with static charges will for the sake of simplicity be diagrammatically indicated by two semicircles 1 and 2 and the rotor of the motor by a ring M. (Figure 3.) A is a vertical aerial or aerial network. S the safety choke or electromagnet with coil O as may be seen is connected with the aerial A. Adjacent the electromagnet S the aerial conductor is divided into three circuits, the circuit 8 giving the safety spark gap, the circuit 7 with the working spark gap, and then a circuit including the stator terminal 1, the rotor and stator terminal 2 at which a connection is made to the earth wire. The two spark gaps are also connected metallically with the earth wire. The method of working these diagrams is as follows:

The positive atmospheric electric charge collected tends to combine with the negative electricity (or earth electricity) connected with the earth wire. It travels along the aerial A through the electromagnet S without being checked as it flows in the same direction as the direct current. Further, its progress is arrested by two sparks gaps placed in the way and the stator condenser surfaces. The stator condenser surfaces are charged until the charge is greater than the resistance of the spark gap 7, whereupon a spark springs over the spark gap 7 and an oscillatory charge is obtained as by means of the motor M, stator surfaces 1 and 2, and spark gap 7, a closed oscillation circuit is obtained for producing the electromagnetic oscillations. The motor here forms the capacity and the necessary inductance and resistance, which, as is well known, are necessary for converting static electricity into electromagnetic wave energy.

The discharges formed are converted into mechanical energy in special motors and can not reach the aerial network by reason of the electromagnet or choke. If, however, when a spark springs over the spark gap 7 a greater quantity of atmospheric electricity tends to flow to earth, a counter voltage is induced in the electromagnet, which is greater the more rapidly and strongly the flow of current direct to the earth is. By the formation of this opposing voltage a sufficiently high resistance is offered to the flow of atmospheric electricity direct to earth to prevent a short circuit with the earth.

The circuit containing spark gap 8 having a different wave length which is not in resonance with the natural frequency of the motor, does not endanger the motor and serves as security against excess voltage, which, as practical experiments have shown, may still arise in certain cases, but can be conducted direct to earth through this spark gap.

In the diagram illustrated in Figure 4 the spark gap 7 is shunted across condensers 5 and 6 from the motor M. This construction affords mainly a better insulation of the motor against excess voltage and a uniform excitation through the spark gap 7.

In Figure 5 a diagram is illustrated for transforming large current strengths which may be employed direct without motors, for example, for lighting or heating purposes. The main difference is that here the spark gap consists of a star shaped disc 7 which can rotate on its own axis and is rotated by a motor opposite similarly fitted electrodes 7^a. When separate points of stars face one another, discharges take place, thus forming an oscillation circuit over condensers 5 and 6 and inductance 9 for oscillatory discharges. It is evident that a motor may also be directly connected to the ends of the spiral 9.

The construction of the diagram shown in Figure 6 permits of the oscillation circuit of the motor being connected with an induction coil. Here a regulating inductive resistance is introduced for counter-acting excess voltages in the motor. By cutting the separate coils 9 (coupled inductively to the aerial) in or out the inductive action on the motor may be more or less increased or variable aerial action may be exerted on the oscillation circuit.

In Figure 7 the oscillation circuit is closed through the earth (E and E_1). The spark gap 7 may be prolonged or shortened by more or fewer spark gaps being successively connected by means of a contact arm 7^b.

Diagram 8 shows a unipolar connection of the motor with the aerial network. Here two oscillation circuits are closed through the same motor. The first oscillation circuit passes from aerial A through electromagnet S, point x, inductance 9^a to the earth condenser 6 and further, over spark gap 7 to the aerial condenser 5 and back to x. The second oscillation circuit starts from the aerial condenser 5 at the point x^1 over the inductance 9 to the earth condenser 6 at the point x^3 and through the condenser 6 over the spark gap 7 back to x^1. The motor itself is inserted between the two points of the spark gap 7. From this arrangement slightly damped oscillation wave currents are produced.

In the diagram illustrated in Figure 9 a loosely coupled system of connections is illustrated which is assumed to be for small motors for measuring purposes. A indicates the aerial conductor, S the electromagnet in the aerial conductor, 9 the inductance, 7 the spark gap, 5 and 6 condensers, E the earth, M the motor, and 1 and 2 stator connections of the motor. The motor is directly metallically connected with the oscillation circuit.

In Figure 10 a purely inductive coupling is employed for the motor circuit. The motor is connected with the secondary wire 10 as may be seen in Figure 11 in a somewhat modified diagram connection. The same applies to the diagram of Figure 12.

The diagrams hitherto described preferably allow of motors of small and medium strength to be operated. For large aggregates, however, they are too inconvenient as the construction of two or more oscillation circuits for large amounts of energy is difficult; the governing is still more difficult and the danger in switching on or off is greater.

A means of overcoming such difficulties is shown in Figure 13. The oscillation circuit here runs starting from the point x over condenser 5, variable inductance 9, spark gap 7 and the two segments (3^a and 4^a) forming arms of a Wheatstone bridge, back to x. If the motor is connected by brushes 3 and 4 transversely to the two arms of the bridge as shown in the drawings, electromagnetic oscillations of equal sign are induced in the stator surfaces 1 and 2 and the motor does not revolve. If however, the brushes 3 and 4 are moved in common with the conducting wires 1 and 2 which connect the brushes with the stator poles a certain alteration or displacement of the polarity is obtained and the motor commences to revolve.

The maximum action will result if one brush 3 comes on the central sparking contact 7 and the other brush 4 on the part x. They are however, usually in practice not brought on to the central contact 7 but only held in the path of the bridge segments 4^a and 3^a in order not to connect the spark gaps with the motor oscillation circuit.

As however, the entire oscillation energy can thereby not act on the motor it is better to carry out the same system according to the diagram 14. The diagram 14 differs from the foregoing only by the motor not being directly metallically connected with the segments of the commutator, but only a primary coil 9 which induces in a secondary coil 10, current which feeds the motor M and takes the place of the rotor. By this arrangement a good transforming action is obtained, a loose coupling and also an oscillation circuit without a spark gap.

In Figure 15 the motor is not purely inductively as in 14, but directly metallically branched off from the primary coil (at x and x^1) after the principle of the auto-transformer.

In Figure 16 instead of an inductance a condenser 6 is in similar manner, and for the same object inserted between the segments 3^a and 4^a. This has the advantage that the segments 3^a and 4^a need not be made of solid metal but may consist of spiral coils whereby a more exact regulation is possible and further motors of high inductance may be employed.

The arrangements of Figures 17, 18 and 19 may be employed for use with resonance and particularly with induction condenser motors; between the large stator induction condenser surfaces, small reversing pole condensers are connected, which, as may be seen from Figures 17, 18 and 19 are led together to earth. Such reversing poles have the advantage that with large quantities of electrical energy the spark formation between the separate oscillation circuits ceases.

Figure 19 shows a further method which prevents electromagnetic oscillations of high number of alternations formed in the oscillation circuit striking back to the aerial conductor. It is based on the well known principle that a mercury lamp, one electrode of which is formed of mercury, the other of solid metal such as steel allows an electric charge to pass in only one direction from the mercury to the steel and not vice versa. The mercury electrode of the vacuum tube N is therefore connected with the aerial conductor and the steel electrode with the oscillation circuit. From this it results that charges can pass only from the aerial through the vacuum tube to the oscillation circuit, but not vice versa. Oscillations which are formed on being transformed in the oscillation circuit cannot pass to the aerial conductor.

In practice these vacuum tubes must be connected behind an electromagnet as the latter alone affords no protection against the danger of lightning.

As regards the use of spark gaps, all arrangements as used for wireless telegraphy may be used. Of course the spark gaps in large machines must have a sufficiently large surface. In very large stations they are cooled in liquid carbonic acid or better still in liquid nitrogen or hydrogen; in most cases the cooling may also take place by means of liquefied low homologues of the metal series or by means of hydrocarbons the freezing point of which lies at between −90° C. and −40° C. The spark gap casing must also be insulated and be of sufficient strength to be able to resist any pressure which may arise. Any undesirable excess super-pressure which may be formed must be automatically let off. I have employed with very good results mercury electrodes which were frozen in liquid carbonic acid, the cooling being maintained during the operation from the outside through the walls.

Figure 20 is one of the simplest forms of construction of an aerial network in combination with collectors, transformers and the like illustrated diagrammatically. E is here the earth wire, 8 the safety spark gap, 7 the working spark gap, 1 and 2 the stator surfaces of the motor, 5 a condenser battery, S the protective magnet which is connected with the coil in the aerial conductor, A^1 to A^{10} aerial antennæ with collecting balloons, N horizontal collecting or connecting wires from which, to the centre a number of connections run.

The actual collectors consist of metal sheaths preferably made of an aluminium magnesium alloy, and are filled with hydrogen or helium and are attached to copper plated steel wires. The size of the balloon is selected so that the actual weight of the balloon and the weight of the conducting wire is supported thereby. On the top of the balloon aluminium spikes, made and gilded in a special manner hereinafter described, are arranged in order to produce a conductor action. Small quantities of radium preparations, more particularly polonium-ionium or mesothorium preparations considerably increase the ionization, and therewith the action of these collectors.

In addition to metal balloons, fabric balloons which are superficially metal coated according to Schoop's metal spraying process, may however also be employed. A metallic surface may also be produced by lacquering with metallic bronzes, preferably according to Schoop's spraying process or lacquering with metallic bronze powders in two electrical series of widely different metals, because thereby the collecting effect is considerably increased.

Instead of the ordinary round balloons, elongated cigar shaped ones may be employed. In order also to utilize the frictional energy of the wind, patches or strips of non-conducting substances which produce electricity by friction, may be attached to the metallized balloon surfaces. The wind will impart a portion of its energy in the form of frictional electricity, to the balloon casing, and thereby the collecting effect is substantially increased.

In practice however, very high towers (up to 300 metres is fully admissible) may be employed as antennæ. In these towers copper tubes rise freely further above the top of the tower. A gas lamp secured against the wind is then lit at the point of the copper tube and a netting is secured to the copper tube over the flame of this lamp to form a collector. The gas is conveyed through the interior of the tube up to the summit. The copper tube must be absolutely protected from moisture at the place at which it enters the tower and also rain must be prevented running down the walls of the tower which might lead to a bad catastrophe. This is done by bell shaped enlargements which expand downwards, being arranged in the tower in the form of high voltage insulators of Siamese pagodas.

Special attention must be devoted to the foundations of such towers. They must be well insulated from the ground, which may be obtained by first embedding a layer of concrete in a box form to a sufficient depth in the ground and inserting in this an asphalt lining and then glass bricks cast about 1 or 2 metres in thickness. Over this in turn there is a ferro-concrete layer in which alone the metal foot of the tube is secured. This concrete block must be at least 2 metres from the ground and be fully protected at the sides by a wooden covering, from moisture. In the lower part of the tower a wood or glass house for the large condenser batteries or for the motors may be constructed. In order to lead the earth connection to the ground water, a well insulated pit constructed of vitreous bricks, must be provided. Several such towers are erected at equal distances apart and connected with a horizontal conductor. The horizontal connecting wires may either run directly from tower to tower or be carried on bell shaped insulators similar to those in use for high voltage conductors. The width of the network may be of any suitable size and the connection of the motors can take place at any suitable places.

In order to collect large quantities of electricity with few aerials it is well to provide the aerial conductor with batteries of condensers as shown in two methods of construction in Figures 21 and 22. In Figure 21 the batteries of condensers 5 are connected on the one hand with the aerial electricity collectors Z by the aerial conductor A, and on the other hand interconnected in series with an annular conductor from which horizontal conductors run to the connecting points C to which the earth wire is connected.

Figure 22 shows a similar arrangement. Should two such series of antennæ rings be shown by a voltmeter to have a large difference of potential (for example, one in the mountains and one in the plain) or even of different polarity these differences may be compensated for by connecting sufficiently large condenser batteries ($5, 5^a, 5^b$) by means of Maji star conductors D and D^1. In Figure 23 a connection of three such rings of collectors to form a triangle with a central condenser battery is illustrated.

The condenser batteries of such large installations must be embedded in liquefied gases or in liquids freezing at very low temperatures. In such cases a portion of the atmospheric energy must be employed for liquefying these gases. It is also preferable to employ pressure. By this means the condenser surfaces may be diminished, and still allow for large quantities of energy to be stored, secure against breakdown. For smaller installations the immersing of the condensers in well insulated oil or the like, suffices. Solid substances on the other hand cannot be employed as insulators.

The arrangement in the diagrams hitherto described was always such that the condenser batteries were connected with both poles directly to the aerial conductors. An improved diagram of the connections for obtaining atmospheric electricity for the condenser batteries has however, been found to be very advantageous, this arrangement consists in that they are connected by only one pole (unipolar) to the collecting network. Such a method of arrangement is very important, as by means of it a constant current and an increase of the normal working pressure or voltage is obtained. If for example a collecting balloon aerial which is allowed to rise to a height of 300 metres, shows 40,000 volts above earth voltage, in practice it has been found that the working voltage (with a withdrawal of the power according to the method hereinbefore described by means of oscillating spark gaps and the like) is only about 400 volts. If however, the capacity of the condenser surfaces be increased, which capacity in the above mentioned case was equal to that of the collecting surface of the balloon aerials, to double the amount, by connecting the condenser batteries with only one pole, the voltage rises under an equal withdrawal of current up to and beyond 500 volts. This can only be ascribed to the favourable action of the connecting method.

In addition to this substantial improvement it has also been found preferable to insert double inductances with electromagnets and to place the capacities preferably between two such electromagnets. It has also been found that the useful action of such condensers can be further increased if an induction coil be connected as inductive resistance to the unconnected pole of the condenser, or still better if the condenser itself be made as an induction condenser. Such a condenser may be compared with a spring which when compressed carries in itself accumulated force, which it again gives off when released. In charging, a charge with reversed sign is formed at the other free condenser pole, and if through the spark gap a short circuit results, the accumulated energy is again given back since now new quantities of energy are induced at the condenser pole connected with the conductor network, which in fact charges with opposite signs to that at the free condenser pole. The new induced charges have of course the same sign as the collector network. The whole voltage energy in the aerial is thereby however increased. In the same space of time larger quantities of energy are accumulated than is the case without such inserted condenser batteries.

In Figures 24 and 25 two different diagrams of connections are more exactly illustrated, Figure 24 shows a collecting balloon and the diagram of the connections to earth. Figure 25 four collecting balloons and the parallel connection of the condenser batteries belonging thereto.

A is the collecting balloon made of an aluminium magnesium alloy (electron metal, magnalium) of a specific gravity of 1.8 and a thickness of plate 0.1 to 0.2 mm. Inside there are eight strong vertical ribs of T shaped section about 10 to 20 mm. in height and about 3 mm. in thickness with the projecting part directed inwards (indicated by *a*, *b*, *c*, *d* and so forth); they are riveted together to form a firm skeleton and are stiffened in a horizontal direction by two cross ribs. The ribs are further connected with one another internally and transversely by means of thin steel wires, whereby the balloon obtains great power of resistance and elasticity. Rolled plates of 0.1 to 0.2 mm. in thickness made of magnalium alloy are then either soldered or riveted on this skeleton so that a fully metallic casing with smooth external surface is obtained. Well silvered or coppered aluminium plated steel wires run from each rib to the fastening ring 2. Further, the coppered steel hawser L preferably twisted out of separate thin wires (shown in dotted lines in Figure 24) and which must be long enough to allow the balloon to rise in the desired height, leads to a metal roller or pulley 3 and from thence to a winch W, well insulated from the earth. By means of this winch, the balloon, which is filled with hydrogen, or helium, can be allowed to rise to a suitable height (300 to 5,000 metres) and brought to the ground for recharging or repairs.

The actual current is taken directly through a friction contact from the metal roller 3 or from the wire, or even from the winch or simultaneously from all three by means of brushes (3, 3^a and 3^b). Beyond the brushes the conductor is divided, the paths being:—firstly over 12 to the safety spark gap 8, from thence to the earth conductor E^1, and secondly over electromagnet S^1, point 13, to a second loose electromagnet having an adjustable coil S^2, then to the spark gap 7 and to the second earth conductor E^2. The actual working circuit is formed through the spark gap 7, condensers 5 and 6, and through the primary coil 9; here the static electricity formed by oscillatory discharges is accumulated and converted into high frequency electromagnetic oscillations. Between the electromagnets S^1 and S^2 at the crossing point 13, four condenser batteries are introduced which are only indicated diagrammatically in the drawings each by one condenser. Two of these batteries (16 and 18) are made as plate condensers and prolonged by regulating induction coils or spirals 17 and 19 while the two others (21 and 23) are induction condensers. As may be seen from the drawings each of the four condenser batteries 16, 18, 21, 23 is connected only by one pole to the aerial or to the collector conductor. The second poles 17, 19, 22, 24 are open. In the case of plate condensers having no inductive resistance an induction coil is inserted. The object of such a spiral or coil is the displacement of phase of the induction current by ¼ periods, whilst the charging current of the condenser poles which lie free in the air, works back to the collector aerial. The consequence of this is that in discharges in the collector aerial the back inductive action of the free poles allows a higher voltage to be maintained in the aerial collecting conductor than would otherwise be the case. It has also been found that such a back action has an extremely favourable effect on the wear of the contacts. Of course the inductive effect may be regulated at will within the limits of the size of the induction coil, the length of the coil in action being adjustable by means of wire connection without induction (see Fig. 24, No. 20).

S^1 and S^2 may also be provided with such regulating devices in the case of S^2 (illustrated by 11). If excess voltage be formed it is conducted to earth through the wire 12 and spark gap 8 or through any other suitable apparatus, since this formation would be dangerous for the other apparatus.

The action of these condenser batteries has already been hereinbefore described.

The small circles on the collector balloon indicate places at which zinc amalgam or gold amalgam or other photoelectric acting metals in the form of small patches in extremely thin layers (.01 to .05 mm. in thickness) are applied to the balloon casing of light metal. Such metallic patches may also be applied to the entire balloon as well as in greater thickness to the conducting network. The capacity of the collector is thereby considerably strengthened at the surface. The greatest possible effect in collecting may be obtained by polonium amalgams and the like. On the surface of the collector balloon metal points or spikes are also fixed along the ribs, which spikes serve particularly for collecting the collector charge. Since it is well known that the resistance of the spikes is less the sharper the spike is, for this purpose it is therefore extremely important to employ as sharp spikes as possible. Experiments made as regards these have shown that the formation of the body of the spike or point also plays a large part. for example, spikes made of bars or rollers with smooth surfaces, have a many times greater point resistance as collector accumulator spikes than those with rough surfaces. Various kinds of spike bodies have been experimented with for the collector balloons hereinbefore mentioned. The best results were given by spikes which were made in the following way. Fine points made of steel, copper, nickel, or copper and nickel alloys, were fastened together in bundles and then placed as anode with the points in a suitable electrolyte (preferably in hydrochloric acid or muriate of iron solutions) and so treated with weak current at 2 to 3 volts pressure. After 2 to 3 hours according to the thickness of the spikes or pins the points become extremely sharp and the bodies of the spikes have a rough surface. The bundle can then be removed and the acid washed off with water. The spikes are then placed as cathode in a bath consisting of solution of gold, platinum, iridium, paladium or wolfram salts or their compounds and coated at the cathode galvanically with a thin layer of precious metal, which must however be sufficiently firm to protect them from atmospheric oxidation.

Such spikes act at a 20 fold lower voltage almost as well as the best and finest points made by mechanical means. Still better results are obtained if polonium or radium salts are added to the galvanic bath when forming the protective layer or coating. Such pins have a low resistance at their points and even at one volt and still lower pressures have an excellent collector action.

In Figure 24 the three unconnected poles are not connected with one another in parallel. That is quite possible in practice without altering the principle of the free pole. It is also preferable to interconnect in parallel to a common collector network, a series of collecting aerials.

Figure 25 shows a diagram for such an installation. A^1, A^2, A^3, A^4 are four metal collector balloons with gold or platinum coated spikes which are electrolytically made in the presence of polonium emanations or radium salts, which spikes or needles are connected over four electro-magnets S^1, S^2, S^3, S^4, through an annular conductor R. From this annular conductor four wires run over four further electromagnets S^a, S^b, S^c, S^d, to the connecting point 13. There the conductor is divided, one branch passing over 12 and the safety spark gap 8 to the earth at E^1, the other over inductive resistance J and working spark gap 7 to the earth at E^2. The working circuit, consisting of the condenser 5 and 6 and a resonance motor or a condenser motor M, such as hereinbefore described, is connected in proximity round the sparking gap section 7.

Instead of directly connecting the condenser motor of course the primary circuit for high frequency oscillatory current may also be inserted.

The condenser batteries are connected by one pole to the annular conductor R and can be either inductionless (16 and 18) or made as induction condensers as shown by 21 and 23. The free poles of the inductionless condensers are indicated by 17 and 19, those of the induction condensers by 22 and 24. As may be seen from the drawings all these poles 17, 22, 19, 24 may be interconnected in parallel through a second annular conductor without any fear that thereby the principle of the free pole connection will be injured. In addition to the advantages already set forth the parallel connection also allows of an equalization of the working pressure in the entire collector network. Suitably constructed and calculated induction coils 25 and 26 may also be inserted in the annular conductor of the free poles, by means of which a circuit may be formed in the secondary coils 27 and 28 which allows current produced in this annular conductor by fluctuations of the charges or the like appearances to be measured or otherwise utilized.

According to what has been hereinbefore stated separate collector balloons may be connected at equidistant stations distributed over the entire country, either connected directly with one another metallically or by means of intermediate suitably connected condenser batteries through high voltage conductors insulated from earth. The static electricity is converted through a spark gap into dynamic energy of a high number of oscillations and may in such form be coupled as a source of energy by means of a suitable method of connecting, various precautions being observed and with special regulations. The wires leading from the collector balloons have hitherto been connected through an annular conductor without this endless connection, which can be regarded as an endless induction coil, being able to exert any action on the whole conductor system.

It has now been found that if the network conductor connecting the aerial collector balloons with one another is not made as a simple annular conductor, but preferably short circuited in the form of coils over a condenser battery or spark gap or through thermionic tubes or valves or audions, then the total collecting network exhibits quite new properties. The collection of atmospheric electricity is thereby not only increased but an alternating field may be easily produced in the collector network. Further, the atmospheric electrical forces showing themselves in the higher regions may also be directly obtained by induction. In Figures 26 and 28 a form of construction is shown on the basis of which the further foundations of the method will be more particularly explained.

In Figure 26 1, 2, 3, 4 are metal collector balloons, 5, 6, 7, 8 their metallic aerial conductors and I the actual collector network. This consists of five coils and is mounted on high voltage insulators in the air, on high voltage masts (or with a suitable construction of cable embedded in the earth). One coil has a diameter of 1 to 100 km. or more. S and S^1 are two protective electromagnets, F the second safety section against excess voltage, E its earth conductor and E^1 the earth conductor of the working section. When an absorption of static atmospheric electricity is effected through the four balloon collectors, the current in order to reach the earth connection E^1 must flow spirally through the collector network over the electromagnet S, primary induction coil 9, conductor 14, anode A of the audion tube, incandescent cathode K, as the way over the electromagnet and safety spark gap F offers considerably greater resistance. Owing to the fact that the accumulated current flows in one direction, an electromagnetic alternating field is produced in the interior of the collector network coil, whereby the whole free electrons are directed more or less into the interior of the coil. An increased ionization of the atmosphere is therefore produced. In consequence of this the points mounted on the collector balloon show a considerably reduced resistance and therefore increased static charges between the points on the balloon and the surrounding atmosphere are produced. The result of this is a considerably increased collector effect.

A second effect which could not be obtained otherwise is obtained by the electromagnetic alternating field which running parallel to the earth surface, acts more or less with a diminishing or increasing effect on the earth magnetic field, whereby in the case of fluctuations in the current a return induction current of reversed sign is always produced in the collector coil by earth magnetism. Now if, however, a constantly pulsating continuous alternating field is produced as stated in the above collector network I, an alternating current of the same periodicity is produced also in the collecting network coil. As the same alternating field is further transmitted to the aerial balloon, the resistance of its points is thereby considerably reduced, whilst the collector action is considerably increased. A further advantage is that positive electrons which collect on the metal surfaces during the conversion into dynamic current produce a so-called drop of potential of the collector area. As an alternating field is present, the negative ions surrounding the collector surfaces, when discharge of the collector surfaces takes place produce by the law of induction, an induction of reversed sign on the collector surface and so forth (that is to say again a positive charge). In addition to the advantages hereinbefore set forth, the construction of connecting conductors in coil form when of sufficiently large diameter, allows of a utilization of energy arising in higher regions also in the simplest way. As is well known electric discharges frequently take place at very great elevations which may be observed, such as St. Elmo's fires or northern lights. These energy quantities have not been able to be utilized up to now. By this invention all these kinds of energy, as they are of an electromagnetic nature and the direction of the axis of the collector coils stands at right angles to the earth's surface, can be more or less absorbed in the same way as a receiver in wireless telegraphy absorbs waves coming from a far distance. With a large diameter of the spiral it is possible to connect large surfaces and thereby to take up also large quantities of energy.

It is well known that large wireless stations in the summer months, and also in the tropics are very frequently unable to receive the signals in consequence of interruptions which are caused by atmospheric electricity, and this takes place with vertical coils of only 40 to 100 m. diameter. If on the contrary horizontal coils of 1 to 100 km. diameter be employed very strong currents may be obtained through discharges which are constantly taking place in the atmosphere. Particularly in the tropics or still better in the polar regions where the northern lights are constantly present, large quantities of energy may probably be obtained in this way. A coil with several windings should act the best. In similar manner any alteration of the earth magnetism should act inductively on such a coil.

It is not at all unlikely that earthquakes and spots on the sun will also produce an induction in such collector coils of sufficient size. In similar manner this collector conductor will react on earth currents more particularly when they are near the surface of the earth or even embedded in the earth. By combining the previous kind of current collectors so far as they are adapted for the improved system with the improved possibilities of obtaining current the quantities of free natural energy which are to be obtained in the form of electricity are considerably increased.

In order to produce in the improved collector coil uniform current oscillations of an undamped nature so-called audion high vacuum or thermionic tubes of suitable connection are employed instead of the previously known spark gaps (Fig. 26, Nos. 9–18). The main aerial current flows through electromagnet S (which in the case of a high number of alternations is not connected here but in the earth conductor E^1) and may be conveyed over the primary coils in the induction winding through wire 14 to the anode A of the high vacuum grid tube. Parallel with the induction resistance 9 a regulating capacity of suitable size, such as condenser 11 is inserted. In the lower part of the vacuum grid tube is arranged the incandescent filament or the cathode K which is fed through a battery B. From the battery B two branches run, one to the earth conductor E^1 and the other through battery B^1 and secondary coil 10 to the grid anode *g* in the vacuum tube. By the method of connections shown in dotted lines, a desired voltage at the grid electrode *g* may also be produced through the wire 17 which is branched off from the main current conductor through switches 16 and some small condensers (*a*, *b*, *c*, *d*) connected in series, and conductor 18, without the battery B^1 being required.

The action of the entire system is somewhat as follows:—

On the connecting conductor of the aerial collector network being short circuited to earth, the condenser pole 11 is charged and slightly damped oscillations are formed in the short circuited existing oscillation circuit formed of the condenser 11 and self inductance 9. In consequence of the coupling through coil 10, fluctuations of voltage take place in the grid circuit 15 with the same frequency, which, fluctuations in turn influence the strength of the electrode current passing through the high vacuum amplyfying tube and thus produce current fluctuations of the same frequency in the anode circuit. A permanent supply of energy to the oscillation circuits 9 and 10 consequently takes place, until a condition of balance is set up, in which the consumed oscillation energy is equal to that absorbed. Thereby constant undamped oscillations are now produced in the oscillation circuits 9–11.

For regular working of such oscillation producers high vacuum strengthening tubes are necessary and it is also necessary that the grid and anode voltages shall have a phase difference of 180° so that if the grid is negatively charged, then the anode is positively charged and vice versa. This necessary difference of phase may be obtained by most varied connections, for example, by placing the oscillation circuit in the grid circuit or by separating the oscillation circuit and inductive coupling from the anodes and the grid circuit and so forth.

A second important factor in this way of converting static atmospheric electricity into undamped oscillations is that care must be taken that the grid and anode voltages have a certain relation to one another; the latter may be obtained by altering the coupling and a suitable selection of the self induction in the grid circuit, or as shown by dotted lines 18, 17, 16 by means of a larger or smaller number of condensers of suitable size connected in series; in this case the battery B^1 may be omitted. With a suitable selection of the grid potential a glow discharge takes place between the grid *g* and the anode A, and accordingly at the grid there is a cathode drop and a dark space is formed. The size of this cathode drop is influenced by the ions which are emitted in the lower space in consequence of shock ionization of the incandescent cathodes K and pass through the grid in the upper space. On the other hand the number of the ions passing through the grid is dependent on the voltage between the grid and the cathode. Thus if the grid voltage undergos periodic fluctuations (as in the present case) the amount of the cathode drop at the grid fluctuates and consequently the internal resistance of the tube correspondingly fluctuates, so that when a back coupling of the feed circuit with the grid circuit takes place, the necessary means are afforded for producing undamped oscillations and of taking current, according to requirements from the collecting conductor.

The frequency of the undamped oscillations produced is with a suitably loose coupling equal to the self frequency of the oscillation circuits 9 and 10. By a suitable selection of the self induction of the coil 9 and capacity 11 it is possible to extend from frequencies which produce electromagnetic oscillations of only a few metres wave length down to the lowest practical alternating current frequency. For large installations a suitable number of frequency producing tubes in the form of the well known high vacuum transmission tubes of .5 to 2 kw. in size may be connected in parallel so that in this respect no difficulty exists.

The use of such tubes for producing undamped oscillations, and also the construction and method of inserting such transmission tubes in an accumulator or dynamo circuit is known and also that such oscillation producing tubes only work well at voltages of 1,000 up to 4,000 volts, so that on the contrary their use at lower voltages is considerably more difficult. By the use of high voltage static electricity this method of producing undamped oscillations as compared with that through spark gaps must be regarded as an ideal solution particularly for small installations of outputs of from 1 to 100 kw.

By the application of safety spark gaps, with interpolation of electro-magnets, not only is short circuiting avoided but also the taking up of current is regulated. Oscillation producers inserted in the above way form a constantly acting electromagnetic alternating field in the collector coil, whereby as already stated, a considerable accumulating effect takes place. The withdrawal wire or working wire is connected at 12 and 13, but current may be taken by means of a secondary coil which is firmly or movably mounted in any suitable way inside the large collector coil, i. e. in its electromagnetic alternating field, so long as the direction of its axis runs parallel with that of the main current collecting coil.

In producing undamped oscillations of a high frequency (50,000 per second and more) in the oscillation circuits 9 and 11, electromagnets S and S^1 must be inserted if the high frequency oscillations are not to penetrate the collector coil, between the oscillation producers and the collector coil. In all other cases they are connected shortly before the earthing (as in Figs. 27 and 28).

In Figure 27 a second method of construction of the connecting conductor of the balloon aerials is illustrated in the form of a coil. The main difference consists in that in addition to the connecting conductor I another annular conductor II is inserted parallel to the former on the high voltage masts in the air (or embedded as a cable in the earth) but both in the form of a coil. The connecting wire of the balloon aerials is indicated as a primary conductor and also as a current producing network; the other is the consumption network and is not in unipolar connection with the current producing network.

In Figure 27 the current producing network I is shown with three balloon collectors 1, 2, 3 and aerial conductors 4, 5, 6; it is short circuited through condenser 19 and inductance 9. The oscillation forming circuit consists in this diagram of spark gap *f*, inductance 10, and condenser 11; the earth wire E, is connected to earth over electromagnet S^1. F is the safety spark gap which is also connected to earth through a second electromagnet S at E. On connecting up the condenser circuit 11 this is charged over the spark gap *f* whereby an oscillatory discharge is formed. This discharging current acts through inductance 10 on the inductively coupled secondary 9, whereby in the producing network a modification of the potential of the condenser 19 is produced. The consequence of this is that oscillations arise in the coil shaped producer network. These oscillations induce a current in the secondary circuit II, which has a smaller number of windings and a less resistance, the voltage of which, according to the proportion of the number of windings and of the ohmic resistance, is considerably lower whilst the current strength is greater.

In order to convert the current thus obtained into current of an undamped character, and to tune its wave lengths, a sufficiently large regulatable capacity 20 is inserted between the ends 12 and 13 of the secondary conductor II. Here also current may be taken without an earth conductor, but it is advisable to insert a safety spark gap E^1 and to connect this with the earth over an electromagnet S^2.

The producer network may be connected with the working network II over an inductionless condenser 21 or over an induction condenser 22, 23. In this case the secondary conductor is unipolarly connected with the energy conductor.

In Figure 28 the connecting conductor between the separate accumulator balloons is carried out according to the autotransformer principle. The collecting coil connects four aerial balloons 1, 2, 3, 4, the windings of which are not made side by side but one above the other. In Figure 28 the collector coil I is shown with a thin line, the metallically connected prolongation coils II with a thick line. Between the ends I^1 and II^1 of the energy network I a regulating capacity 19 is inserted. The wire I^1 is connected with the output wire and with the spark gap F.

As transformer of the atmospheric electricity an arrangement is employed which consists in using rotary pairs of condensers in which the one stator surface B is connected with the main current, whilst the other A is connected with the earth pole. Between these pairs of short circuited condensers are caused to rotate from which the converted current can be taken by means of two collector rings and brushes, in the form of an alternating current, the frequency of which is dependent on the number of balloons and the revolutions of the rotor. As the alternating current formed in the rotor can act, in this improved method of connection described in this invention, through coils 10 on the inductance 9, an increase or diminution of the feed current in I can be obtained according to the direction of the current by back induction. Current oscillations of uniform rhythm thereby result in the coil shaped windings of the producer net-work.

As the ends of this conductor are short circuited through the regulatable condenser 19 these rhythms produce short circuited undamped oscillations in the energy conductor, the periodicity and wave lengths of which oscillations can be adjusted according to desire by altering the capacity 19 to a given wave length and therewith also to a given frequency. These currents may also be employed in this form directly as working current through the conductors II^1 and III. By inserting the condenser 20 a connection between these conductors may also be made, whereby harmonic oscillations of desired wave length are formed. By this means quite new effects as regards current distribution are obtained. The withdrawal of current can even take place without direct wire connection if, at a suitable point in the interior of the producing network (quite immaterially whether this has a diameter of 1 or 100 km.) a coil tuned to these wave lengths and of the desired capacity is firmly or movably mounted in the aerial conductor in such a way that its axial direction is in parallel with that of the collector coil. In this case a current is induced in the producing network, the size of which is dependent on the total capacity and resistance and also on the periodicity employed. A possibility is thereby afforded in future, of taking energy from the producer network by wireless means. As thereby in addition to atmospheric electricity also magnetic earth currents and the energy from the higher atmosphere (at least partially) may be simultaneously obtained, this last system for collecting the atmospheric energy is of particular importance for the future.

Of course everywhere instead of spark gaps suitable grid vacuum tubes may be employed as producers for undamped oscillations. The separate coils of the producer net-work with large diameters may be connected with one another through separate conductors all in parallel or all in series or in groups in series. By regulating the number of oscillations and also the extent of the voltage more or less large collector coils of this kind may be employed. The coils may also be divided spirally over the entire section. The coils may be carried out in annular form or also in triangular, quadrangular, hexagonal or octagonal form.

Of course wires may be carried from a suitable place to the centre or also laterally which serve the current waves as guides. This is necessary when the currents have to be conducted over mountains and valleys and so forth. In all these cases the current must be converted into a current of suitable periodicity.

As already hereinbefore mentioned separate collecting balloons may be directly metallically interconnected at equidistant stations distributed over the entire country or may be connected by interpolation of suitable condenser batteries by means of high voltage conductors. The static electricity is converted through a spark gap into dynamic energy of a high number of oscillations, and could then in such form, with a suitable arrangement of the connections, observing various measures of precaution, be employed as source of energy after separate or special regulation.

According to this invention in order to increase the collecting effect of the balloon in the aerial collector conductor or in the earth wire, radiating collectors are employed. These consist either of incandescent metal or oxide electrodes in the form of vacuum grid tubes, or electric arcs (mercury and the like electrodes) Nernst lamps, or finally flames of various kinds may be simply connected with the respective conductor.

It is well known that energy can be drawn off from a cathode consisting of an incandescent body opposite an anode charged with positive electricity (vacuum grid tube). Hitherto however, a cathode was always first directly placed opposite an anode, and secondly the system always consisted of a closed circuit.

Now if we dispense with the ordinary ideas in forming light or flame arcs in which a cathode must always stand directly opposite an anode, and if we place an incandescent cathode opposite an anode charged to a high potential or another body freely floating in the air, or regard the incandescent cathode only as a source of unipolar discharge (which represent group and point discharges in electro-static machines similar to unipolar discharges), it may be ascertained that incandescent cathodes and less perfectly all incandescent radiators, flames and the like admit of relatively large current densities and allow large quantities of electric energy to radiate into the open space in the form of electron streams as transmitters.

The object of this invention is as described below, if such incandescent oxide electrodes or other incandescent radiators or flames are not freely suspended in space but connected metallically with the earth so that they can be charged with negative terrestrial electricity, these radiators possess the property of absorbing the free positive electrical charges contained in the air space surrounding them (that is to say of collecting them and conducting them to earth). They can therefore, serve as collectors and have, in comparison to the action of the spikes, or points, a very large radius of action R; the effective capacity of these collectors is much greater than the geometrical capacity (R_0–) calculated in an electro-static sense.

Now as our earth is surrounded as is well known, with an electro-static field and the difference of potential

$$\frac{\delta V}{\delta h}$$

of the earth field according to the latest investigations, is in summer about 60 to 100 volts and in winter 300 to 500 volts per metre of difference in height (δh), a simple calculation gives the result that when such a radiation collector or flame collector is arranged for example on the ground, and a second one is mounted vertically over it at a distance of 2,000 metres and both are connected by a conducting cable, there is a difference of potential in summer of about 2,000,000 volts and in winter even of 6,000,000 volts and more.

According to Stefan Boltzmann's law of radiation, the quantity of energy which an incandescent surface (temperature T) of 1 sq. cm. radiates in a unit of time into the open air (temperature T_o) is expressed by the following formula:

$$S = \sigma\,(T^4 - T^4_o)\ \text{watt/sq. cm.}$$

and the universal radiation constant σ is according to the latest researches of Ferry (Annales de Chimie et de physique 17 page 267 (1909)) equal to 6.30×10^{-12} watt/sq. cm.

Now if an incandescent surface of 1 sq. cm. shows, as compared with the surrounding space a periodic fall of potential δV it radiates (independent of the current direction, that is to say of the sign) in accordance with the above formula, for example at a temperature of 3725° C. an energy of 1.6 kw. per sq. cm. per second. As for the radiation the same value can be calculated for the collection of energy, but reversed. Now as carbon electrodes at the temperature of the electric arc support on the current basis a current density up to from 60 to 65 amperes per sq. cm. no difficulties will result in this direction in employing radiating collectors as accumulators.

If the earth be regarded as a cosmically insulated condenser in the sense of geometrical electro-statics *x* there results from the geometric (compare Ewald Rasch, "das elektrische Bogenlicht" (The electric arc light) page 169) capacity of the earth according to Chwolson:

For negative charging 1.3×10^6 Coulomb

For negative potential $V = 10 \times 10^8$ volts.

From this there results however, $EJT \cong 24.7 \times 10^{24}$ watt/Sec. Now if it is desired to make a theoretic short circuit through an earthed flame collector this would represent an electric total work of about 79,500 10^{10} kilowatt years. As the earth must be regarded as a rotating mechanism which is thermo-dynamically, electromagnetically, and also kinematically coupled with the sun and stars system by cosmic radiations and gravitation a diminution of the electric energy of the earth field is not to be feared. The energies which the incandescent collectors would withdraw from the earth field can only cause by the withdrawal of motor work a lowering of the earth temperature (temperature $T_E = 300$) and reduce this to that of the world space ($T = O$) by using the entire energy. This is however not the case as the earth does not represent a cosmically entirely insulated system. On the contrary there is conveyed ~~to the~~ same according to the recent value corrected by Ferry for the solar constants through the radiation from the sun an energy of 18,500 $\times 10^{10}$ kw. Accordingly any lowering of the earth temperature (T_E) without a simultaneous lowering of the sun's temperature (T_S) would contradict Stefan Boltzmann's law of radiation.

$$S = \sigma\,(T_S^4 - T_E^4).$$

From this it must be concluded that if the earth temperature (T_E) sinks, the total radiation S absorbed by the earth increases, and further also that the secular speed of cooling of the earth is directly dependent on that of the sun and the other radiators cosmically coupled with the sun and is connected most closely with these.

The incandescent radiation collectors may, according to this invention, be employed for collecting atmospheric electricity if they (1) are charged with the negative earth electricity (that is to say when they are directly connected by means of a metallic conductor with the earth) and (2) if large capacities (metal surfaces) charged with electricity are mounted opposite them as positive poles in the air. This is regarded as the main feature of the present invention as without these inventive ideas it would not be possible to collect with an incandescent collector, sufficiently large quantities of the electrical charges contained in the atmosphere as technology requires; the radius of action of the flame collectors would also be too small, especially if it be considered that the very small surface density (energy density) (σ about $= 2 \times 7 \cdot 10^9$ St. E. per sq. cm.) does not allow of large quantities of charge being absorbed from the atmosphere.

x) Calculated according to Poisson's calculation;

$\Delta V = -4\pi\delta$; as here the alteration of the potential or potential gradients only takes place in the direction of the normal, this calculation assumes the simple form

$$\delta = \frac{1}{4\pi} \times \frac{\delta^2 V}{\delta n^2}$$

It has indeed already been proposed to employ flame collectors for collecting atmospheric electricity and it is known that their collecting effect is substantially greater opposite the points. It is however, not known that the quantities of current which could hitherto be obtained are too small for technical purposes. According to my experiments the reason for this is to be found in the too small capacities of the collector conductor poles. If such flame or radiating collectors have no or only small positive surfaces, their radius of action for large technical purposes is too small. If the incandescent collectors be constantly kept in movement in the air they may collect more according to the speed of the movement, but this is again not capable of being carried out in practice.

By this invention the collector effect is considerably increased by a body charged with a positive potential and of the best possible capacity being also held floating (without direct earth connection) opposite such an incandescent collector which is held floating in the air at a desired height. If for example, a collecting balloon of sheet metal or of metalized balloon fabric be caused to mount to 300 up to 3,000 metres in the air and as positive pole it is brought opposite such a radiating collector connected by a conductor to earth, quite different results are obtained.

The metallic balloon shell (with a large surface) is charged to a high potential by the atmospheric electricity. This potential is greater the higher the collecting balloon is above the incandescent collector. The positive electricity acts concentratedly on the anode floating in the air as it is attracted through the radiation shock ionization, proceeding from the incandescent cathode. The consequence of this is that the radius of action of the incandescent cathode collector is considerably increased and thereby also the collecting effect of the collecting balloon surface. Further the large capacity of the anode floating in the air plays therefore an important part because it allows of the taking of large charges, and thereby a more uniform current is obtained even when there is a large consumption: this cannot be the case with small surfaces.

In the present case the metallic collecting balloon is a positive anode floating in the air and the end of the earth conductor of this balloon serves as positive pole surface opposite the surface of the radiating incandescent cathode, which in turn is charged with negative earth electricity being conductingly connected to earth.

The process may be carried out by two such contacts (negative incandescent cathode and anode end of a capacity floating in the air) a condenser and an inductive resistance being switched on in parallel, whereby simultaneously undamped oscillations may be formed.

In very large installations it is advisable to connect two such radiating collectors in series. Thus an arc light incandescent cathode may be placed below on the open ground and an incandescent cathode which is heated by special electro-magnetic currents be located high in the air. Of course for this the special vacuum Liebig tubes with or without grids may also be employed. An ordinary arc lamp with oxide electrodes may be introduced on the ground and the positive pole is not directly connected with the collecting balloon, but through the upper incandescent cathode or over a condenser. The method of connecting the incandescent cathode floating in the air may be seen in Figs. 29–33.

B is the air balloon, K a Cardan ring (connection with the hawser) C the balloon, L a good conducting cable, P a positive pole, N negative incandescent cathode, and E earth conductor.

Fig. 29 represents the simplest form of construction. If electric oscillations are produced below on the ground by means of a carbon arc lamp or in other suitable way a considerably greater electric resistance is opposed to that in the direct way by inserting an electrical inductive resistance 9. Consequently between P and N a voltage is formed, and as, over N and P only an inductionless ohmic resistance is present, a spark will spring over so long as the separate induction co-efficients and the like are correctly calculated. The consequence of this is that the oxide electrode (carbon or the like) is rendered incandescent and then shows as incandescent cathode an increased collecting effect. The positive poles must be substantially larger than the negative in order that they may not also become incandescent. As they are further connected with the large balloon area which has a large capacity and is charged at high voltage, an incandescent body which is held floating in the air and a positive pole which can collect large capacities is thereby obtained in the simplest way. The incandescent cathode is first caused to become incandescent by means of separate energy produced on the earth, and then maintained by the energy collected from the atmosphere.

Fig. 30 only shows the difference that instead of a round balloon a cigar shaped one (of metal or metalized fabric) may be employed and also a condenser 5 is inserted between the incandescent cathode and the earth conductor so that a short circuited oscillation circuit over P. N 5 and 9 is obtained. This has the advantage that quite small quantities of electricity cause the cathode to become incandescent and much larger cathode bodies may be rendered incandescent.

In this form of construction both the incandescent cathode and also the positive electrode may be enclosed in a vacuum chamber as may be seen in Fig. 32. A cable L is carried well insulated through the cover of a vessel and ends in a condenser disc 5. The cover is arched in order to keep off the rain. The vessel is entirely or partially made of magnetic metal and well insulated inside and outside. Opposite the disc 5 another disc 6 and on this again a metallic positive pole of the vacuum tube *g* with the incandescent cathode (oxide electrode) N is arranged. The negative electrode is on the one hand connected with the earth conductor E, and on the other hand with the inductive resistance 9 which is also connected with the cable L with the positive pole and wound round the vessel in coils. The action is exactly the same as that in Fig. 29 only instead of an open incandescent cathode one enclosed in vacuo is employed. As in such collectors only small bodies can be brought to incandescence in large installations a plurality of such vacuum tubes must be inserted in proximity to one another. According to the previous constructions Figs. 31 and 33 are quite self evident without further explanations.

Figs. 34–37 represent further diagrams of connections over radiating and flame collectors, and in fact, how they are to be arranged on the ground.

Fig. 34 shows an arc light collector with oxide electrodes for direct current and its connection; Fig. 35 a similar one for alternating current, Fig. 36 an incandescent collector with a Nernst lamp and Fig. 37 a similar one with a gas flame.

The positive pole 1 of the radiating collectors is always directly connected to the aerial collecting conductor A. In Fig. 34 this is further connected over the condenser battery 5 with a second positive electrode 3. The direct current dynamo *b* produces current which flows over between the electrodes 3 and 2 as an arc light. On the formation of an arc the negative incandescent electrode 2 absorbs electricity from the positive poles standing opposite it and highly charged with atmospheric electricity and conveys the same to the working circuit. The spark gap 7, inductive resistance 9 and induction coil 10 are like the ones previously described. The protective electromagnet S guards the installation against earth circuiting, the safety spark gap 8 from excess voltage or overcharging.

In Fig. 35 the connection is so far altered that the alternating current dynamo feeds the exciting coil 11 of the induction condenser. 12 is its negative and 13 its positive pole; if the coil 3 on the magnet core of the dynamo is correctly calculated and the periodicity of the alternating current is sufficiently high an arc light can be formed between the two poles 1 and 2. As the cathode 2 is connected with the negatively charged earth, and therefore always acts as a negative pole, a form of rectification of the alternating current produced by the dynamo 3 is obtained, the second half of the period is always suppressed. The working circuit may be carried out in the same way as in Fig. 34; the working spark gap 7 may however be dispensed with, and instead thereof between the points *n* and *m* a condenser 5 and an induction resistance 9 may be inserted from which the current is taken inductively.

Fig. 36 represents a form of construction similar to Fig. 34 only that here instead of an arc lamp a Nernst incandescent body is employed. The Nernst lamp is fed through the battery 3. The working section is connected with the negative pole, the safety spark gap with the + poles. The working spark gap 7 may also be dispensed with and the current for it taken at 12 over the oscillation circuit 5, 11 (shown in dotted lines).

Flame collectors (Fig. 37) may also be employed according to this invention. The wire network 1 is connected with the aerial collector conductor A and the burner with the earth. At the upper end of the latter, long points are provided which project into the flame. The positive electrode is connected with the negative over a condenser 5 and the induction coil 9 with the earth.

The novelty in this invention is firstly, the use of incandescent cathodes opposite positive poles which are connected with large metallic capacities as automatic collecting surfaces, (2) the connection of the incandescent cathodes with the earth whereby, in addition to the electricity conveyed to them from the battery or machine which causes the incandescing, also the negative charge of the earth potential is conveyed, and (3) the connection of the positive and negative poles of the radiating collectors over a condenser circuit alone or with the introduction of a suitable inductive resistance, whereby simultaneously an oscillatory oscillation circuit may be obtained. The collecting effect is by these methods quite considerably increased.

I declare that what I claim is:—

1. An electrical energy generating system, comprising a conducting surface for static charges, means to support same at a distance above the earth, a conductor leading to the earth level, a spark gap associated with said conductor to convert electrostatic charges into electromagnetic high frequency oscillations means to supply said electromagnetic energy to a net work, and a spark gap of greatly increased relative resistance in parallel therewith.

2. An electrical energy generating system

comprising a conductor, means to support same above the earth level, an inductance therein, a spark gap associated with said conductor, a second spark gap of much higher relative resistance in parallel therewith and an energy receiving circuit coupled with the spark gap of lesser resistance.

3. An electrical energy generating system comprising a collecting surface, means to support same above the earth level, a conductor connecting said collecting surface with the earth level, a choke in said conductor, an electromagnetic resistance converting electrostatic energy to electromagnetic energy, a safety higher resistance in parallel therewith and a net work coupled with the conversion resistance of lesser value.

4. An electrical energy generating system comprising electric conductors spaced above the earth to form electromagnetic oscillating circuits, conductors connecting to earth level, electrostatic to electromagnetic energy conversion means therein, a safety high electrostatic resistance in parallel therewith and means to alter the electromagnetic characteristic of the circuits.

5. An electrical energy generating system comprising in combination a static collecting surface arranged above the earth, conductors connecting to earth level, a pair of spark gaps in parallel of different electrostatic resistance, a utilization net work shunted across the spark gap of lesser resistance and an electromagnetic choke in said conductors.

6. An electrical energy generating system comprising an open circuit energy collecting aerial, a pair of sparking gaps in parallel of widely different resistance, connected thereto and a closed electric oscillation circuit in shunt across the gap of lesser resistance.

7. An electrical energy generating system comprising an open circuit energy collecting aerial, a pair of sparking gaps in parallel of widely different resistance connected thereto, a closed electric oscillation circuit in shunt across the gap of lesser resistance, a plurality of electrostatic collecting surfaces, means to connect said collecting surfaces in parallel in groups and means to connect said groups symmetrically with said aerial.

In witness whereof, I have hereunto signed my name this 30 day of Dec., 1920, in the presence of two subscribing witnesses.

HERMANN PLAUSON.

Witnesses:
H. F. ARMSTRONG,
W. H. BEESTON.

US 20150102676A1

(19) **United States**
(12) **Patent Application Publication**
Dinwiddie et al.

(10) **Pub. No.: US 2015/0102676 A1**
(43) **Pub. Date: Apr. 16, 2015**

(54) **METHOD AND APPARATUS FOR EXTRACTING AND CONVEYING ELECTRICAL ENERGY FROM THE EARTH'S IONOSPHERE CAVITY**

(71) Applicant: **Earth Energies, Inc.**, Johns Creek, GA (US)

(72) Inventors: **John Dinwiddie**, Cary, NC (US); **Terry L. Wright**, Suwanee, GA (US)

(21) Appl. No.: **14/509,846**

(22) Filed: **Oct. 8, 2014**

Related U.S. Application Data

(60) Provisional application No. 61/889,894, filed on Oct. 11, 2013.

Publication Classification

(51) **Int. Cl.**
H01F 38/14 (2006.01)
(52) **U.S. Cl.**
CPC ***H01F 38/14*** (2013.01); *H01F 2038/146* (2013.01)

(57) **ABSTRACT**

The system and apparatus of one or more embodiments of the present invention extracts, conditions, and conveys electric power from the earth ionosphere cavity through the integrated and collaborative operation of the system and apparatus consisting of a capacitively-coupled insulated elevated terminal (coupled capacitor upper plate), an evacuated spark gap, an integrated step-down transformer and resonant capacitor, and a ground terminal (coupled capacitor lower plate). The architecture governing exemplary embodiments of the system and apparatus of the present invention emulates the natural architecture governing the interaction of living trees with the electrical energy resident in the earth ionosphere cavity. The implementation of such exemplary embodiments of the present invention utilizes standard and customized components appropriate for their function within the system and apparatus.

The extracted electric power from the earth ionosphere cavity manifests in the form of broadband electromagnetic waves in the 0 to 200 Hz frequency range. Electric energy from these waves produce magnetic fields inside the integrated step-down transformer which are added together by the primary coil of the integrated step-down transformer, and delivered as 60 Hz 120/240 VAC to an electric load via the secondary coil of the integrated step-down transformer resonating at 60 Hz due to the influence of the integrated resonant capacitor.

The system and apparatus of exemplary embodiments of the present invention extracts, conditions, and conveys ionosphere cavity-resident electric power in alternating current and direct current realms.

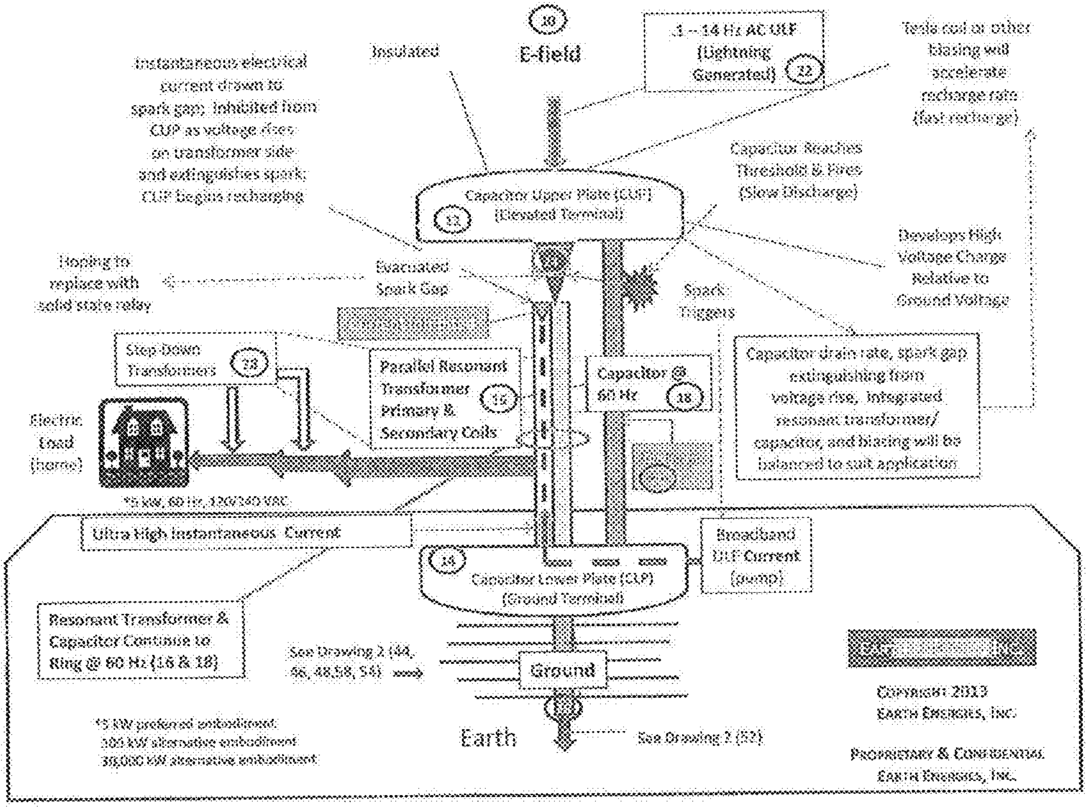
10
E-field
Insulated
.1 – 14 Hz AC ULF (Lightning Generated) 22
Tesla coil or other biasing will accelerate recharge rate (fast recharge)
Instantaneous electrical current drawn to spark gap; Inhibited from CUP as voltage rises on transformer side and extinguishes spark; CUP begins recharging
Capacitor Reaches Threshold & Fires (Slow Discharge)
Capacitor Upper Plate (CUP) (Elevated Terminal) 12
Develops High Voltage Charge Relative to Ground Voltage
Hoping to replace with solid state relay
Evacuated Spark Gap
Spark Triggers
Step-Down Transformers 28
Parallel Resonant Transformer Primary & Secondary Coils 16
Capacitor @ 60 Hz 18
Capacitor drain rate, spark gap extinguishing from voltage rise, integrated resonant transformer/capacitor, and biasing will be balanced to suit application
Electric Load (home)
*5 kW, 60 Hz, 120/240 VAC
Ultra High Instantaneous Current
14
Capacitor Lower Plate (CLP) (Ground Terminal)
Broadband ULF Current (pump)
Resonant Transformer & Capacitor Continue to Ring @ 60 Hz (16 & 18)
See Drawing 2 (44, 46, 48,58, 54)
Ground
Earth
See Drawing 2 (52)
*5 kW preferred embodiment
100 kW alternative embodiment
35,000 kW alternative embodiment
COPYRIGHT 2013
EARTH ENERGIES, INC.
PROPRIETARY & CONFIDENTIAL
EARTH ENERGIES, INC.

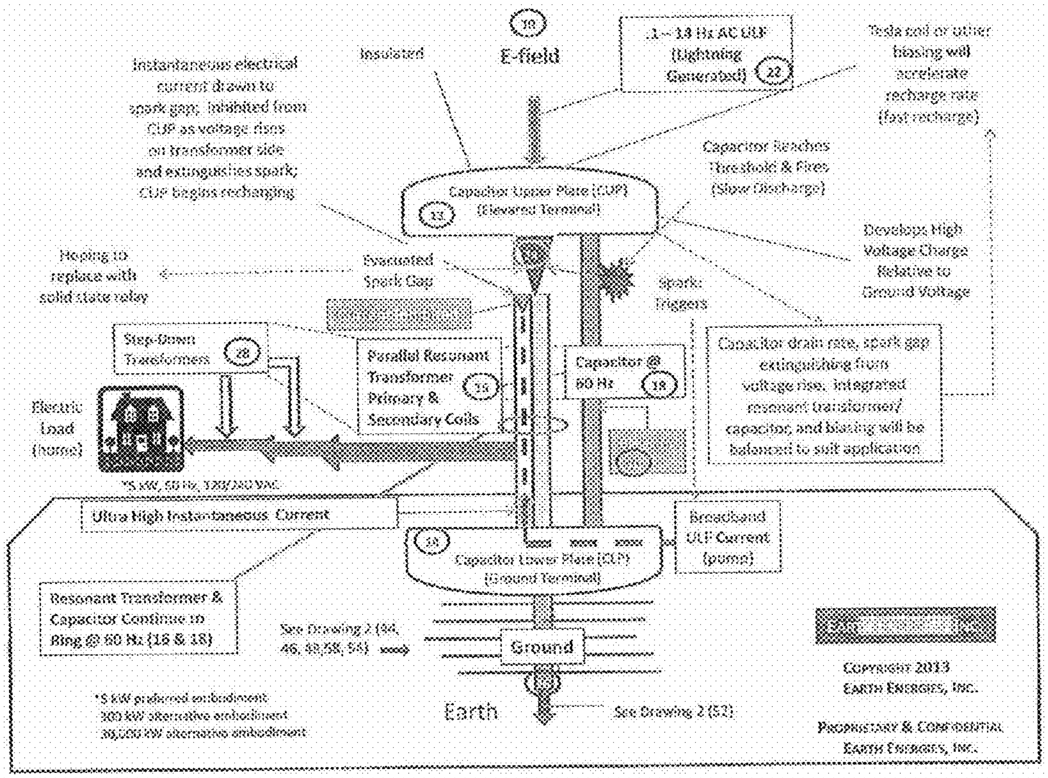
10
E-field
Insulated
.1 ~ 13 Hz AC ULF (Lightning Generated) 22
Tesla coil or other biasing will accelerate recharge rate (fast recharge)
Instantaneous electrical current drawn to spark gap; inhibited from CUP as voltage rises on transformer side and extinguishes spark; CUP begins recharging
Capacitor Upper Plate (CUP) (Elevated Terminal) 12
Capacitor Reaches Threshold & Fires (Slow Discharge)
Develops High Voltage Charge Relative to Ground Voltage
Hoping to replace with solid state relay
Evacuated Spark Gap
Spark Triggers
Step-Down Transformers 28
Parallel Resonant Transformer Primary & Secondary Coils 26
Capacitor @ 60 Hz 18
Capacitor drain rate, spark gap extinguishing from voltage rise, integrated resonant transformer/ capacitor, and biasing will be balanced to suit application
Electric Load (home)
*5 kW, 60 Hz, 120/240 VAC
Ultra High Instantaneous Current
16
Capacitor Lower Plate (CLP) (Ground Terminal)
Broadband ULF Current (pump)
Resonant Transformer & Capacitor Continue to Ring @ 60 Hz (16 & 18)
See Drawing 2 (44, 46, 48,58, 54)
Ground
Earth
See Drawing 2 (52)
*5 kW preferred embodiment
300 kW alternative embodiment
35,000 kW alternative embodiment
Copyright 2013
Earth Energies, Inc.
Proprietary & Confidential
Earth Energies, Inc.

FIG. 1

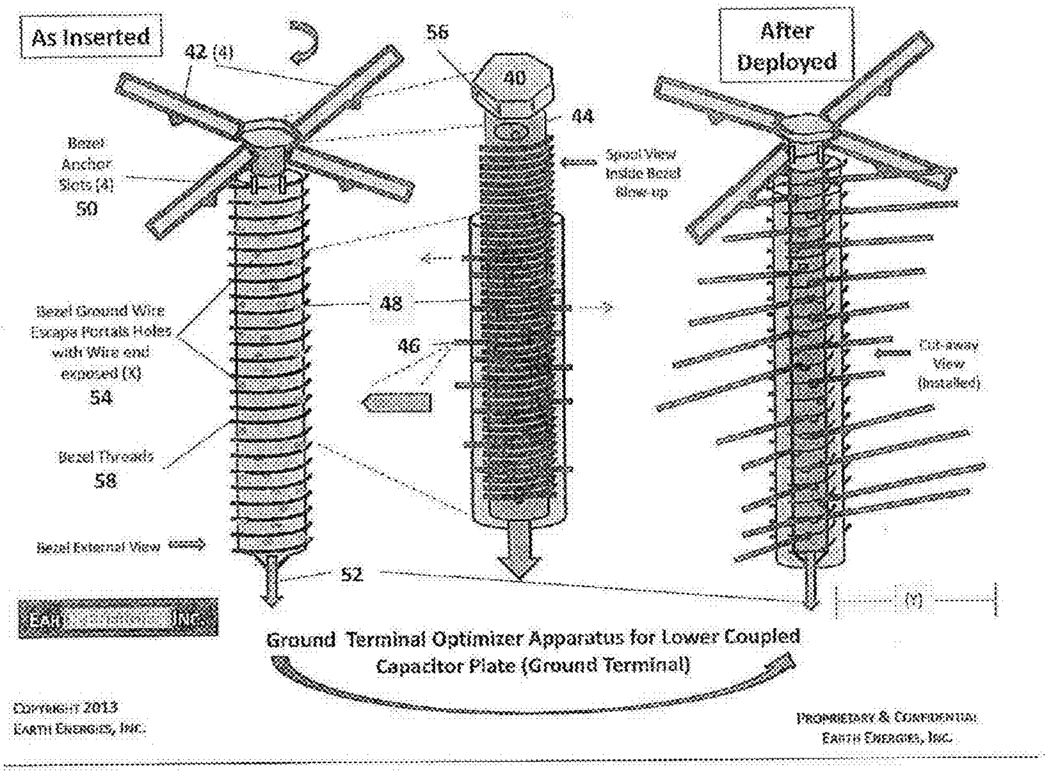
As Inserted
42 (4)
56
40
44
Spool View
Inside Bezel
Blow-up
Bezel
Anchor
Slots (4)
50
48
46
Bezel Ground Wire
Escape Portals Holes
with Wire end
exposed (X)
54
Bezel Threads
58
Bezel External View
52
After
Deployed
Cut-away
View
(Installed)
(Y)
Ground Terminal Optimizer Apparatus for Lower Coupled
Capacitor Plate (Ground Terminal)
Copyright 2013
Earth Energies, Inc.
Proprietary & Confidential
Earth Energies, Inc.

FIG. 2

METHOD AND APPARATUS FOR EXTRACTING AND CONVEYING ELECTRICAL ENERGY FROM THE EARTH'S IONOSPHERE CAVITY

[0001] This application claims benefit of U.S. Provisional Application No. 61/889894, filed 11 Oct. 2013, titled, "Method and Apparatus for Extracting and Conveying Electrical Energy From the Earth's Ionosphere Cavity," the disclosure of which is incorporated herein by reference in its entirety.

BACKGROUND OF THE INVENTION

[0002] A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by any-one of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright whatsoever.

[0003] The invention is an apparatus and methodology for extracting electric power manifesting as a broadband collection of electric field oscillations in the 0 to 200 Hz frequency range within the Earth's ionosphere cavity, through the use of an elevated biased hemispherical or other curved surface capacitor coupling the ionosphere cavity and Earth, and then conveying this electrical energy at the proper frequency, amperage, and voltage to existing homes and buildings where it is delivered for consumption over existing home or building wiring.

DESCRIPTION OF THE RELATED ART

[0004] The current state of the art in this field is focused on the bending of radio frequency waves in the Earth's ionosphere cavity for the purpose of distorting (through reflection) the resident cavity's magnetic field to physically direct the stimulating electrical energy to targeted locations within the Earth's ionosphere cavity and/or coordinates on the Earth's surface at specific energy levels appropriate to specific applications. The regenerative feedback was done by Sutton's active antenna.

[0005] Nikola Tesla attempted to transmit wireless power in the "natural medium"; however, Tesla bad no description of a device to receive natural oscillations as described in the present disclosure.

[0006] There is no known prior art in the field of electric energy extraction from the ionosphere cavity and conveyance to other terminals within the cavity for delivery of usable electric power.

SUMMARY OF THE INVENTION

[0007] In one exemplary embodiment of the present invention, a properly-scaled electric power extraction/conveyance apparatus is installed at a typical consumer residential home to deliver 60 Hz 120/240 VAC electric power, sufficient to convey the equivalent of 5 kW of continuous electricity, to the home's existing electric service entry component to power electrical appliances in the home over existing home electric wiring.

[0008] In another exemplary embodiment of the present invention, a properly-scaled electric power extraction/conveyance apparatus is installed at an appropriate location within an industrial park or neighborhood to deliver 60 Hz 120/240 V AC electric power, sufficient to convey the equivalent of 300 kW of continuous electricity, to the service entry components of industrial buildings or neighborhood homes to power electrical devices and appliances over existing building or home electric wiring.

[0009] Alternative exemplary embodiments of the present invention will extract direct current electric power from the Earth ionosphere cavity, generated by the earth's rotating magnetosphere, and convey in the form of conditioned direct current to direct current load sources, or convert to alternating current and convey in the form of conditioned alternating current to alternating current load sources.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] A disclosure embodiments of the present invention, is set forth more particularly in the remainder of the specification, which makes reference to the accompanying figures, in which:

[0011] FIG. **1** provides an example of an electric power extraction and conveyance apparatus for one residential-scale embodiment of the present invention. Certain additional example embodiments of the present invention utilize similar fundamental technologies, but differ in the scale (size) of the components and characteristics of the electricity conveyed for consumption; and

[0012] FIG. **2** is a schematic illustration of ground terminals for use with embodiments of the present invention.

DETAILED DESCRIPTION

[0013] Applicant has discovered that trees extract and use electric power from the Earth ionosphere cavity. The present invention emulates the architecture of electric power extraction and usage methodology inherent in the natural operation of living trees, but uses standard and customized components, common in the field, for its implementation. The natural electric power extraction and use of electric power by trees leverages the shape and physical composition of trees as instances of large electrical capacitors coupled to the ionosphere cavity. The canopy of trees (upper limbs and leaf structures) emulates the upper plates of a coupled capacitor, the trunks of trees emulate a combined dielectric and resonant transformer/capacitor, and the root systems of trees emulate ground terminals (lower coupled capacitor plates). The extraction of electric power from the ionosphere cavity by trees is in the form of ultra-low frequency broadband waves (oscillations) in the frequency range of 0.1 to 14 Hertz (ULF). Inherent tree physical composition and control structure (i.e., DNA/RNA) govern the use of the extracted electric power to accomplish tree growth, seasonal metabolic adjustments, and reproductive processes. Any excess extracted electric power by trees is returned to the ionosphere cavity. The present invention is an exemplary instance of this same natural architecture using standard and customized components that facilitate extraction and conveyance of ionosphere cavity-resident broadband electric power at useful scales sufficient to provide continuous electrical power for electrical devices used by people.

[0014] One preferred exemplary embodiment of the present invention is in the form shown in Drawing **1** above and will deliver continuous 60 Hz 120/240 V AC electric power to consumer residences (and other structures with similar electricity conditioning and consumption requirements) equivalent to electric power capacity of 5 kilowatts.

[0015] A secondary exemplary embodiment of the present invention is also in the form shown in FIG. **1**, appropriately scaled for larger electrical power demands and related conveyance capacity, and will deliver continuous 60 Hz 120/240 V AC electric power to larger buildings or electrical distribution facilities for commercial office parks with electric power capacity of 300 kilowatts.

[0016] A third exemplary embodiment of the present invention is also in the form shown in FIG. **1**, appropriately scaled for larger electrical power capacity, and will deliver continuous 60 Hz 120/240 VAC electric power to utility substations with electric power capacity of 30,000 kilowatts (or higher).

[0017] The following system, power sources, and component descriptions provide additional details of the present invention in various embodiments.

[0018] Earth's ionosphere cavity (**10**) is a reservoir of continuously-generated electrical energy (E-Field) in both direct current (DC) and alternating current (AC) realms. The DC electric energy is created by the constant rotation of the Earth's magnetosphere exposed to the solar wind. Earth's perpetually moving magnetic fields create the DC electric power in the ionosphere cavity E-Field per the equations in James Clerk Maxwell's fundamental theory of electromagnetism. AC electrical energy in the ionosphere cavity is generated from the persistent E-Field disturbances (i. e., oscillations) caused by lightning discharges. The present invention includes the direct extraction of AC electric power from the ionosphere cavity, and the direct extraction of DC electric power from the ionosphere cavity which is converted to AC electric power before conveyance. Optional operational configurations of the present invention will also enable conveyance of extracted DC electric power for potential scenarios where the conveyance of conditioned DC electricity is required.

[0019] Electric-field oscillations (**22**) are induced by lightning and encapsulated in the earth-ionosphere cavity, creating a continuous source of renewable electrical energy. The frequency of lightning occurrences in the ionosphere cavity averages thirty (30) strikes per second, with each strike representing approximately one (1) terawatt (TW) of instantaneous energy. Assuming an average energy dissipation rate for each lightning strike of fifteen (15) microseconds (μsec), i. e., the time involved for one (1) terawatt of instantaneous energy to dissipate to zero (0) energy, these lightning-induced electric field oscillations are sufficient to provide 4,500 times the entire electric energy consumption on planet Earth at the projected consumption rate for the year 2100 (55.3 TWh annually). Lightning-induced E-Field oscillations vary between 0.1 and 14 Hertz with wavelengths exceeding the practicality of an electromagnetic antenna for collecting the energy. The present invention utilizes capacitive coupling to overcome this obstacle.

[0020] An insulated upper capacitor plate (terminal) (12) elevated above the influence of the ground floor collects E-Field oscillations from the ionosphere cavity. This insulated elevated terminal, capacitively-coupled to the E-Field bounded by the earth-ionosphere cavity, is arranged in a surface of large radii of curvature, supporting maximum surface connection to the E-Field without leaking the high voltage charge applied to the elevated terminal. The capacitance and resistance of the elevated terminal are responsive to the reception of broadband electric-field frequencies in the 0 to 200 hertz range. The elevated terminal is subject to high voltage alternating current biasing used to collect electric-field oscillations generated by lightning impulses in the earth-ionosphere cavity.

[0021] An evacuated spark gap (**24**) connected to the elevated terminal (**12**) prevents electrical discharge occurring between the elevated terminal and the ionosphere cavity. As the high voltage alternating current approaches peak voltage the spark gap energizes and a large voltage is applied to the step-down transformer primary coil (**16**). The high voltage pumps high current from the ground terminal (**14**, **26**) converting the electric-field to a magnetic field within the transformer (**16**). The step-down transformer is connected to a 60 hertz resonant capacitor (**18**), the secondary side of the transformer resonating at 60 hertz will supply a 60 hertz 120/240 VAC power to the load. Filter circuits will condition the power. The evacuated spark gap function may be accomplished by solid state circuitry in the implementation of the present invention.

[0022] The resonant transformer (**16**) is integrated with a parallel resonant capacitor across the transformer where the integrated combination resonates at 60 hertz supplying power to the electricity consuming structure (home or business). Various step-down transformers (**28**) are utilized to condition the conveyed electricity to accommodate the requirements of the receiving station for various embodiments of the present invention.

[0023] Ground (**26**) is the source of electric current for the present invention. A ground terminal, connected to the soil, is arranged in a surface of large radii of curvature to permit instantaneous sourcing of high currents from earth ground without temperature and voltage rise inhibiting the ground terminal collection of current. The capacitance and resistance of the ground terminal to earth is minimized to promote on demand current sourcing. Drawing **2** illustrates an exemplary apparatus of the present invention that optimizes ground connection quality by maximizing the penetration extent and resulting surface area of the ground terminal's extension into the soil.

[0024] Step-down transformers (**28**) are scaled to transform the electrical current and voltage into structured and conditioned electricity that is compatible with the electric power needs of the load sources for various embodiments of the present invention.

[0025] A mechanical nut (**40**) solidly attached to shaft **44** to facilitate extension, when turned, of large gauge copper wire extrusions (**46**).

[0026] A downward pointed anchor flat shaft (**42**) attached to the bezel **50** via a T-slot in bezel.

[0027] A copper (or other highly-conductive metal) spool base (**44**) sourcing "X" independently attached large gauge pointed copper wires (**46**) of length "Y", where "X" is equal to the number of escape portals in bezel (**50**), and "Y" is the length determined to yield optimum surface area of wires of type **46** for high quality connectivity to soil. The spool can freely rotate inside the bezel (**48**) and emulates the primary root structure of a living tree.

[0028] Sharpened large gauge (reinforced) copper wires (**46**), quantity "X", of length "Y" emulating the root structure of living trees.

[0029] Threaded bezel (spool-housing sleeve) (**48**) with drilled escape portal holes with end of copper wires (**46**) protruding through holes ½ inch.

[0030] Bezel anchor slots (**50**) of twice the height of anchor flat shafts (**42**) and T-slotted to accommodate anchor flat shafts.

[0031] Extended pointed end (**52**) of inner spool base shaft.

[0032] Escape portals (**54**) in bezel to allow for large gauge copper wire (**46**) extension into soil.

[0033] Connector port (**56**) for connection to ground terminal (**14**) in Drawing **1**.

[0034] Bezel threads (**58**).

[0035] Various embodiments of the present invention will include one or more of the following characteristics in various combinations. These descriptions are provided for purposes of example. Not all of the following characteristics will be required for any given exemplary embodiment.

[0036] Characteristic 1, The form (shape), physical composition, size, insulation, location altitude, discharge capacitive threshold and discharge rate, and the strength of AC or DC biasing applied to the (CUP) upper coupled capacitor plate (**12**: elevated terminal) all form an integrated whole, the purpose of which is to extract and temporarily store DC or AC electric power from the Earth's ionosphere cavity (**10**) at high voltage, and to trigger the energizing of the evacuated spark gap (**24**).

[0037] Characteristic 2, The elevated terminal (**12**) is capacitively coupled to the E-Field (**10**) bounded by the earth ionosphere cavity, and arranged in a surface of large radii of curvature, supporting maximum surface connection to the E-Field without leaking the high voltage charge applied to the elevated terminal (**12**).

[0038] Characteristic 3, The capacitance and resistance of the elevated terminal (**12**) are responsive to the reception of broadband electric-field frequencies in the 0 to 200 hertz range.

[0039] Characteristic 4, The elevated terminal (**12**) is subject to high voltage alternating current biasing used to collect E-Field oscillations generated by lightning impulses in the earth-ionosphere cavity.

[0040] Characteristic 5, An alternative instantiation of elevated terminal (**12**) is subject to high voltage direct current biasing used to collect E-Field direct current generated by the Earth's rotating magnetosphere.

[0041] Characteristic 6, The evacuated spark gap (**24**) may be in the form of a cohesive physical spark ignition or an operational emulation of this function through solid state circuitry.

[0042] Characteristic 7, The evacuated spark gap (**24**) connected to the elevated terminal (**12**) prevents electrical discharge occurring between the elevated terminal and the ionosphere cavity.

[0043] Characteristic 8, The high voltage alternating current from the E-Field, induced into the elevated terminal (**12**), approaches peak voltage and causes the evacuated spark gap (**24**) to energize and a large voltage is applied to the step-down transformer primary coil (**16**).

[0044] Characteristic 9, The high voltage alternating current impulse from the elevated terminal (**12**), which causes the evacuated spark gap (**24**) to energize, manifests as alternating current on broadband frequencies between 0.1 and 200 Hz.

[0045] Characteristic 10, In an alternative instantiation of the elevated terminal (**12**) the stored high voltage direct current, which causes the evacuated spark gap (**24**) to energize, manifests as direct current on broadband frequencies between 0 and 200 Hz.

[0046] Characteristic 11, The high voltage applied to the step-down transformer primary coil (**16**) pumps high current from the ground terminal (**14**), converting the electric-field to a magnetic field within the transformer (**16**).

[0047] Characteristic 12, The high current pumped from the ground terminal (**14**), caused by high voltage applied to the step-down transformer primary coil (**16**), will attempt to normalize the field and extinguish the evacuated spark gap (**24**).

[0048] Characteristic 13, The transformer's (**16**) magnetic field expands in response to the high current electric-field inrush from the ground terminal (**14**).

[0049] Characteristic 14, The primary winding inside the step-down transformer (**16**) will add the energy from each broadband frequency together into a single magnetic field, where it will become a 60 Hz frequency at the secondary winding of the transformer (**16**).

[0050] Characteristic 15, The step-down transformer (**16**) is connected to a 60 Hertz resonant capacitor (**18**).

[0051] Characteristic 16, The secondary (coil) side of the step-down transformer (**16**), resonating at 60 hertz, will supply a 60 hertz 120/240 VAC power to the electric load.

[0052] Characteristic 17, Filter circuits within the step-down transformer will condition the power according to the electric power needs of the electric load.

[0053] Characteristic 18, Electric-field lines exist (within the Earth's ionosphere cavity) that are horizontal to the surface of the Earth and extend to heights above 30,000 feet, with voltage gradients typically 100 Volts per vertical meter.

[0054] Characteristic 19, Lightning impulses create global broadband oscillations in the Earth ionosphere cavity electric-field between 0.1 and 14 Hertz.

[0055] Characteristic 20, Rotation of the Earth's magnetosphere creates global electric field oscillations in the Earth ionosphere cavity that manifest as electromagnetic waves in the 0 to 200 Hertz frequency range.

[0056] Characteristic 21, Global broadband oscillations in the Earth ionosphere cavity electric-field can be used for wireless energy transmission without harm to people, plants, or trees.

[0057] Characteristic 22, A maximum quality ground connection is achieved by expanding the surface area of the ground connection (**48** and **44**) through extension of high-conductor large gauge copper wires (**46**) horizontally through bezel (**48**) escape portals (**54**) along depth of source spool (**44**) and bezel (**48**).

[0058] Characteristic 23, The technique described in Characteristic 22 effectively expands the diameter of the ground source spool (**44**) with regard to quality of ground connection.

[0059] Characteristic 24, Bezel threads (**58**) expand surface area of bezel (**48**) touching earth, enhancing ground connection quality.

[0060] Characteristic 25, Bezel threads (**58**) enhance ground connection quality of bezel (**48**), source spool (**44**), and high conductor large gauge copper wires (**46**) through rotational insertion and resulting friction hold v. other (direct drive) insertion methods which can result in corruption of ground hole integrity caused by imprecise angle and/or vibration during insertion.

[0061] While one or more embodiments of the present invention have been described above, it should be understood that any and all equivalent realizations of the present invention are included within the scope and spirit thereof. Thus, the embodiments presented herein are by way of example only and are not intended as limitations of the present invention. Therefore, it is contemplated that any and all such embodiments are included in the present invention.

What is claimed is:

1. A power receiver for extracting electrical energy from the earth's electric field, said power receiver comprising:

a resonant transformer connected between a ground terminal disposed below the surface of the earth and an elevated terminal;

impulse generator for generating and applying a high voltage electrical impulse to a primary winding of the resonant transformer to induce current flow from the ground terminal through the primary winding of the transformer; and

a power conversion circuit connected to a secondary winding of the resonant transformer to convert electrical current flowing through the secondary winding to a desired form.

2. The power receiver of claim **1** wherein a resonant frequency of the resonant transformer is below 200 Hz.

3. The power receiver of claim **1** wherein the elevated terminal comprises an upper capacitive plate coupled to the earth's ionosphere cavity.

4. The power receiver of claim **3** wherein the impulse generator comprises:

the upper capacitive plate;

a pair of electrodes separated by a spark gap, said electrodes connected between the upper capacitive plate and the primary winding of the resonant transformer and configured to generate a spark when a voltage difference between the electrodes reaches a predetermined level.

5. The power receiver of claim **1** further comprising a lower capacitive plate coupled to the ground terminal.

6. The power receiver of claim **1** wherein the resonant transformer comprises a capacitor connected in parallel with a primary winding of the resonant transformer.

* * * * *

Creative SCIENCE & Research

459

Information Big Brother & the Oil Companies do not want you to know!

Copyright 2002 - 2004 Creative Science & Research

PO BOX 557 New Albany, IN. 47151 USA

Free Energy *From The* EARTH

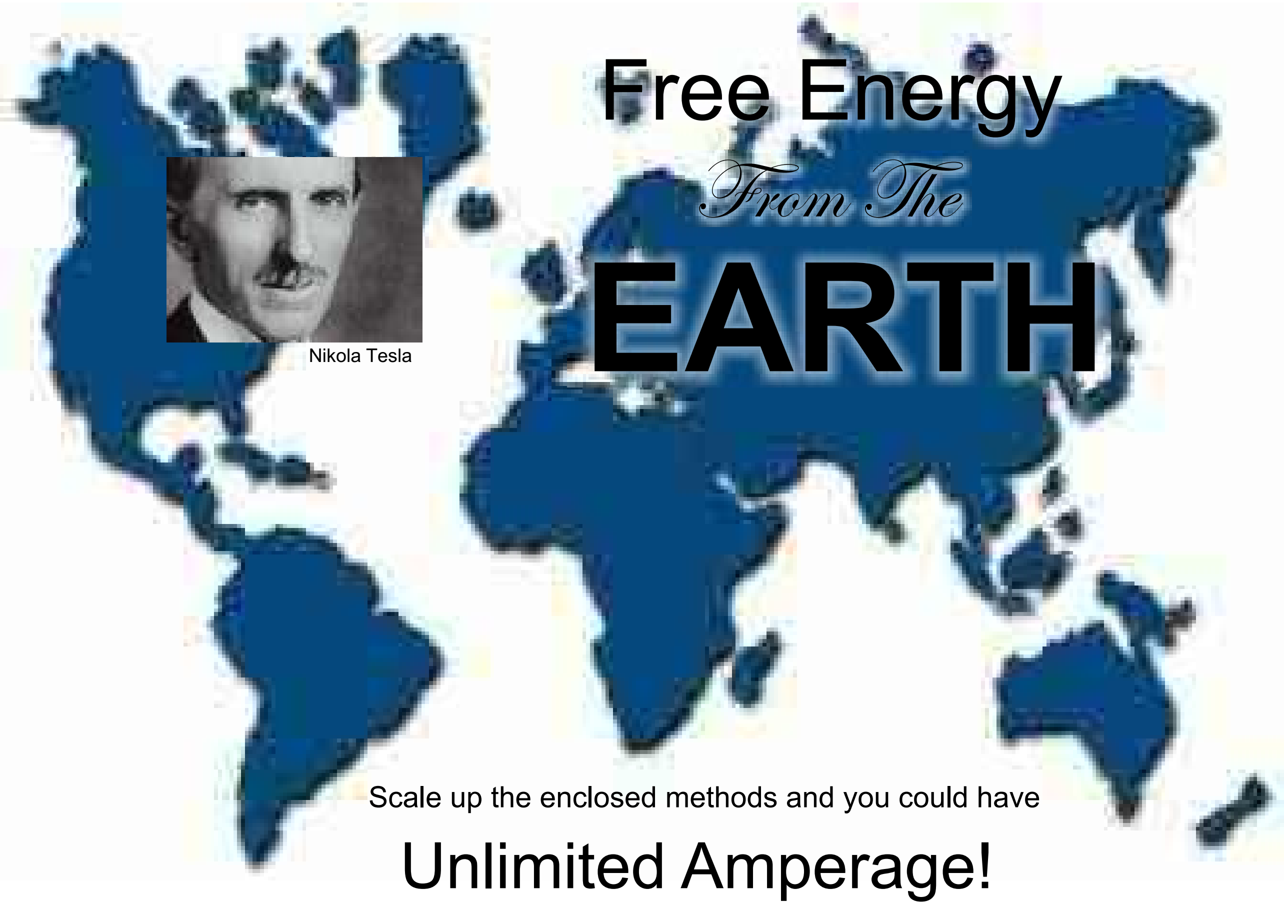

Nikola Tesla

Scale up the enclosed methods and you could have

Unlimited Amperage!

By Unlimited we mean you can keep adding more earth cells to your earth power station as you can afford it year by year adding more and more amp power.

Aluminum

Copper

There are many ways to pick up on the free energy in the earth, Just a few are described in these plans. These devices if made right will pick up earth energy that many call ether or static currents as well as acting as high efficient earth battery's. Ether travels through plastic wood etc... Study and research these plans closely there is more to these plans than what you may think. You can learn to get enough power to run your home!

Copyright 2002 - 2004 Creative Science & Research

Learning The basic's of our invention

Experiment #1 How to assemble a 12 vdc system

1. You will have 12- 4" 3/4" copper piping, or If you did not purchase the kit then cut your copper to this size. KIT: take each piece of copper and tape the tops with 1" masking tape.

2. Now using RED LACQUER spray paint (Which you can buy at any hardware store), Paint the outside of the copper all around. (DO NOT PAINT THE INSIDE.) let air dry for 24 hrs. Now remove 1" masking tape.

3. Take the copper pipe outside to the area that you are going to use. use a hammer and hammer each one Into the ground up to the 1" copper exposed top. do not let the exposed copper touch the earth in the final placement. Dirt should be moist. Not real dry.

4. Now remove each pipe and place a thick piece of plastic on bottom of copper pipe to cap it. LET IT BE LOSE, DO NOT TAPE IT. YOU WANT RAIN WATER TO BE ABLE TO ESCAPE FROM EACH PIPE, Do not let exposed outside copper touch the earth.

5. Now hammer each pipe back Into the same holes. place each 2" piece of zinc rod (or a long bolt, most bolts are zinc plated) into the center of each pipe. The zinc Is the NEGATIVE and the copper is the + POSITIVE.

6. Now connect them in series to get 12 volts, use clip wire or solder them. If you are going to leave It outside for a long time then make sure connections are soldered and then painted to protect them from the weather. The rain will replenish your earth batteries. Top get more amaperage with this type simply add more copper and zinc rod cells USE 3' PIECES OF COPPER PIPE INSTEAD OF 4" PIPE. THEN CONNECT EVERY ROW OF 12 PIPES IN PARALLEL.. ADD AS MANY ROWS AS YOU NEED UNTIL YOU GET THE DESIRED AMPERAGE YOU WANT. **(THINK BIG!)** This can be a very powerful free energy system for your home or other. Not only will you get an earth battery effect, but you will also be collecting earth currents from ether /static and under ground radio waves as well.

7. The deeper you go with the copper pipe, the more amperage and voltage you can get.

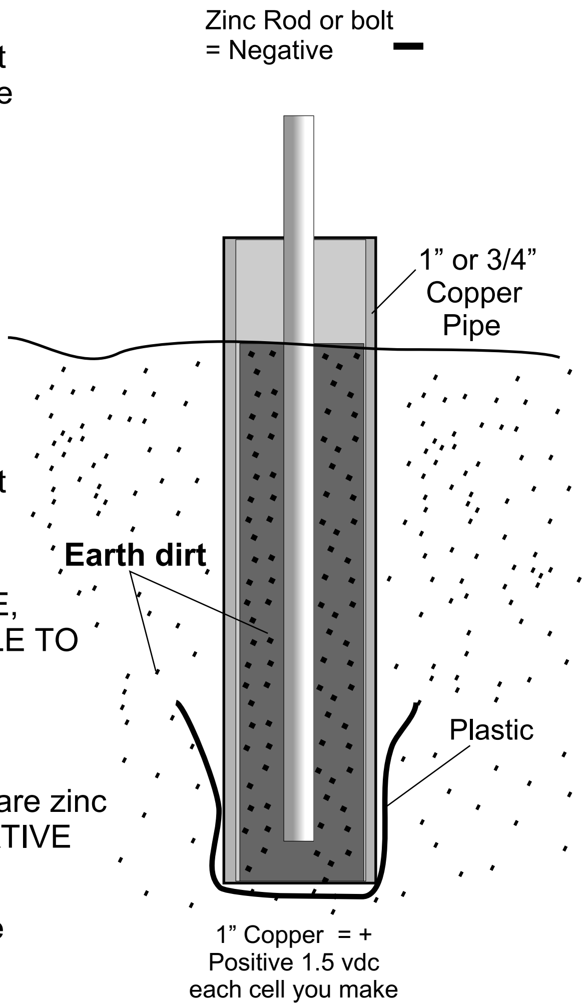

This represents one cell. A qty of 12 to 14 will give you 12 - 14 volts dc when connected in series.

Cut 12 - 4" L x 3/4" copper pipe

Copyright 2002 - 2004 Creative Science & Research

Learning The basic’s of our invention

Experiment #2 How to assemble a 12 vdc system

This is just an example to go by and is not the best way but is low in efficiency! See our capacitor earth cells, coil type or flat sheet type. They produce far more amperage and voltages than this type.

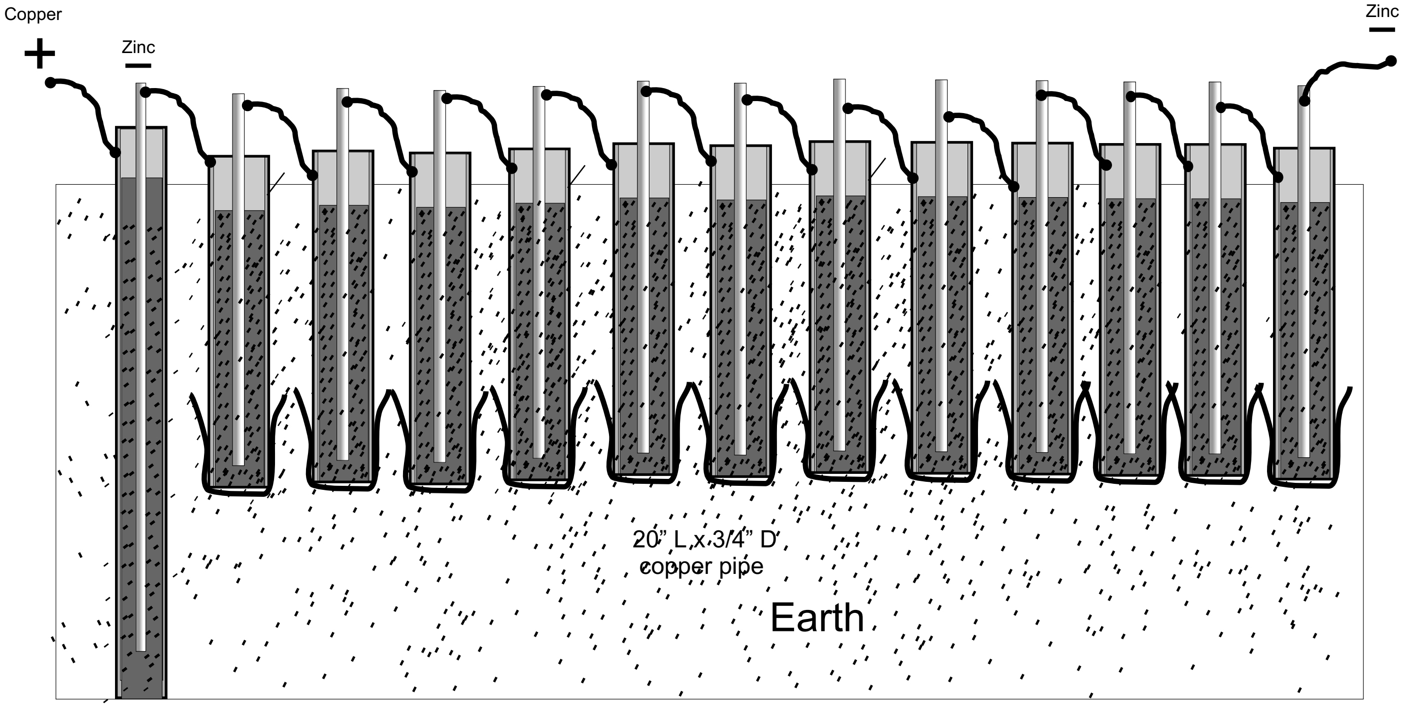

= 1 - 12vdc array

NOTICE: Do not paint the 10 foot copper cell. Keep all copper exposed to the earth. The more exposed copper the better. For higher amperage output, use a zinc or aluminum rod 10/16” Diameter which will leave a space of more than 1/16” to the copper.

Advantages of earth powered batteries

1. Free Energy
2. Long Life
3. The collection of aether energy.
4. Cells are also replenished by outside weather. Rain as well as lightning charges.

Copyright 2002 - 2004 Creative Science & Research

Learning The basic's of our invention

You can purchase Zinc rods or zinc plated bolts at any hardware store, most all nuts and bolts they sell are zinc coated and will work just fine. They also carry aluminum sheeting. Most hardware stores can order you copper sheeting as well.

Experiment #3 How to assemble a 12 vdc system

This is a simple way to produce more amperage but is not practical, we are only showing you this to educate you. The closer the zinc rod is to the copper the more amperage you will get as well as going deeper into the earth. If you do decide to build this basic set up. It is suggested that you solder all leads very well and dig deep enough into the earth to hide the tops of each cell. By covering the tops with dirt the cells can short out so what you want to do is to paint the exposed tops with lacquer or enamel paint then cover with plastic, cover with dirt and allow both the negative and the positive leads to stick out of the ground. But like I said this is not practical, what your goal should be is to make a very strong uf earth capacitor / battery, This will allow you to pick up and collect earth radiant waves of energy as well as lighting energy during storms, the energy you can collect in a strong lighting storm will amaze you! You must use an antenna on the zinc rod or copper rod. **WARNING! Be careful when collecting lighting it can kill you, be sure the positive and negative wires are not coming into your home. During a storm it is advised to have the lead wires going into a large uf cap bank via a one way diode on the positive lead to the positive of the cap or battery. The earth battery's can be dangerous at times and must be completely buried in the earth so a child, adult or animal can not touch or get to. Our earth capacitor batteries can store and hold thousands of volts. So be careful. Use rubber gloves and clothing to protect yourself. We are not responsible for anything in these plans you build at your own risk.**

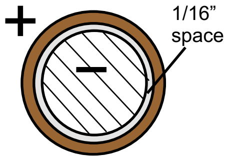

Example of a perfect Zinc rod copper pipe cell

Top view of a 12 volt dc array high amp system

Connect the series arrays in parallel to each other as shown.
Or for a high voltage system connect all in series! Pulse the dc into a 10x step up transformer, 15,000 v Neon transf.

Zinc rod can be replaced by using rolled up aluminum sheeting or Zinc sheeting, Use 100% cotton cloth wrapped around the rolled sheeting or rod to give a small space in between the aluminum sheet and copper.

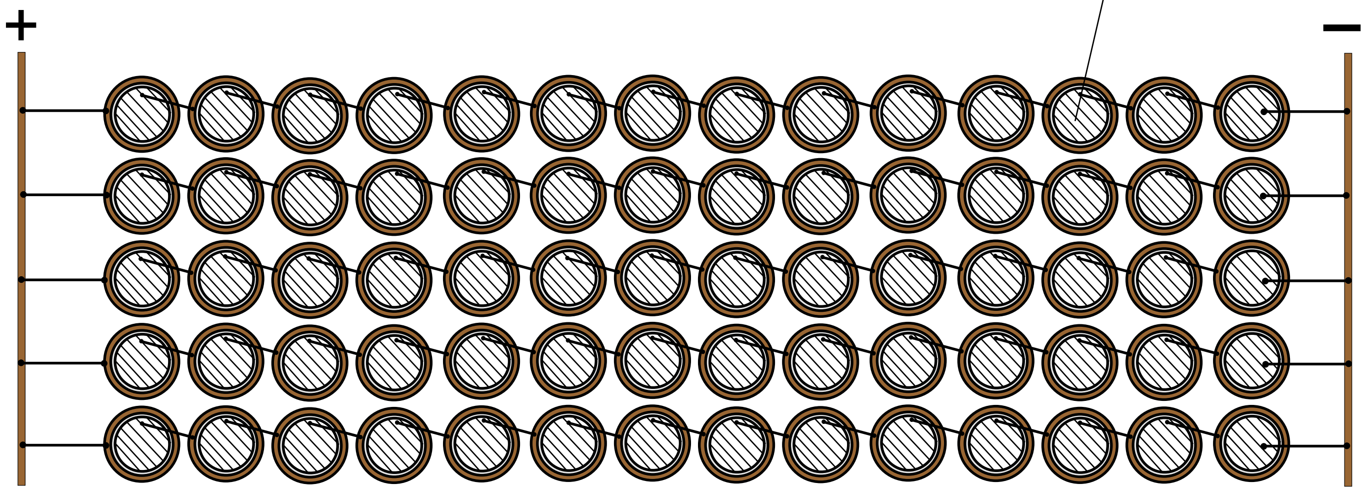

Copyright 2002 - 2004 Creative Science & Research

Learning The basic’s of our invention

Experiment #4 The 6 foot spacing method (old way)

There are many US Patents that were Issued back in the 1800's, one of them was issued to a Mr. Deckman. Deckman found that if you take several small pieces of Zinc and Carbon rod and stick them into the ground next to each other and connect them in series (much like a battery) you get very liittle gain at all.

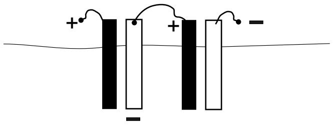

But if you put them 6 feet apart you will get a gain In voltage and they will not cancel one another out. so by doing this you can put them In series to increase your voltage and your volt amps. The theory is that there Is some sort of natural energy VORTEX that takes up about that much space for each unit or cell.

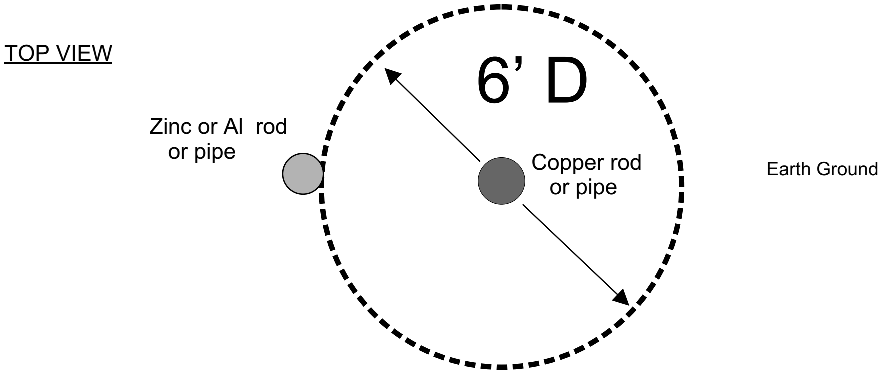

By using this method it will take a lot of ground space, ground space that many people just do not have with the exception of farmers. If you do have the space then think BIG! dig deep and have as many cells as you can get up to 2 to 4 acres. There are much better ways - vs - the old method as you will see. but by using Deckmans old way or our new methods, you can get as much free energy as you want with voltages or amperages as high as you want or desire. The higher the desired amperage the more costly it can get. We are always working to better our inventions to keep the cost down.

Copyright 2002 - 2004 Creative Science & Research

Learning The basic’s of our invention

By using our new system you can get as much voltage or amperage as you want. Just size It to meet your needs or your budget.

How to get AC current for house hold use.

You can convert the incoming dc current to ac by using a home made inverter (which we sell the plans to, **5 kw inverter plans $40 order #579**) or you can purchase a 12 v to 115 vac inverter from a store or solar panel supply house.. If you are planing to make a 115 VAC system this Is the only way you can do it. You can not get AC current from the ground. You must use our system to charge a deep cycle 12 v marine battery, and then connect our 5,000 watt 120 vdc to ac inverter to the battery to run your home. (we do not sell kits or inverters assembled, we only sell the plans at this time.)

More on the basic pipe method

Please note: the dirt level inside each pipe should be up to 1/4" from the top. each pipe should be tightly fliied with dirt. Which will automatically happen when you hammer the pipe into the ground the first time.

Each copper pipe (or cell) should be spaced 1 Inch to 1 1/2 inch apart. Doing this will give you more amperage and voltage.

How to dig a simple deep hole for 8' pipe: There are many ways in which you can do this. you can rent a ground drill bit and rig it up with a electric 2 hp motor or you can use 1" copper pipe.

First: prepare the ground with water to make it moist but not muddy, 2nd: hammer a 1" x 5' copper pipe into the moist ground about 5 inches deep/ then remove it from the ground and take a stick or a water hose to remove the dirt from inside the pipe. Then just repeat the process over and over until you get the desired depth that you want. you can get about 10 feet or more using this method.

WARNINGI BE CAREFUL OF ELECTIC POWER LINES WHEN YOU PLACE OR PULL THE COPPER PIPE FROM THE GROUND, IT WILL KILL YOU IF YOU TOUCH A POWER LINE WITH THE END OF THE COPPER PIPE.

We are dealing with a new type of energy as well as conventional energy < Ether energy can go through rubber/ plastic. Iron, etc....

What is ether or Radiant energy?

The understanding comes from a very deep study of etheric or static forces, this energy that we are getting out of the ground is what some call LIGHT ETHER, when you separate the light ether from the warm ether you can get these superconductive events. It appears like electricity but it has no heat events related to It.

The point is that the best polarity's of things to get a differential in the light ether is to have a silvered coloured metal and then something that is totally absorbitive/ Like black Carbon.

So now you have something that's the perfect reflector and something that is the perfect absorber, and once you understand what your pulling out of the ground (and that it Is not classic electricity as we get out of a normal type of chemical battery), then your going to be farther down the road to understanding how to get much more power out of this stuff.....

There are researchers now that are getting 1 /2 volt to 2 volts dc on a "Quote" volt meter, but it will shock you through an inch of rubber shoes this is not normal electricity!

Q: Is this energy coming from the sun and being absorbed In the earth or from the center of the earth?

A: Yes, but the primary source of light ether is the sun light, yes the earth does have an economy of light ether which does follow the sun around. BUT IT IS ALSO IN THE GROUND! AND ITS IN THE AIR!

Q: Did NokolaTesla know about this in the 1900's?

A: Yes, if you read into his writings he was always saying things like, there is more energy moving than just the electricity in this stuff.

Example: If you take a large Tesla coil and fire it up, and you drive it with a van de graph generator and you put the output of that into a Tesla coil. you can throw sparks across the room! What these electrostatic machines are gathering is light ether.

Copyright 2002 - 2004 Creative Science & Research

A Very Strange and New Discovery!

While working outside on another experiment I discovered a very strange generator effect, much to my surprise, I found that by using 2 aluminum 8” disks connected to my DC volt meter and by placing both disks in a vertical position while touching the bare earth dirt, and moving one of the disks and allowing the other to be stationary I get a dc voltage positive output! This effect works with any size or shape of aluminum, we used 8” diameter disk x 1/8” thick, we found that size to work best. Also by stacking them like a capacitor you will find that the current output is far greater, For Example: see figure 1, Disk A would have 4 more disks connected in parallel of the same size separated by paper die electrics. The same goes for Disk B .

First find a spot any where on the ground where there is no grass (best when wet). Please note, it's not just creating an acid battery effect, you are also getting a good contact with the ground as a conductor to pick up ether energy.

Now place the disks side by side, attach your DC voltmeter probes onto the disks, one negative and one positive. DC meter should be on the lowest setting, now place both disks on the ground very softly and move **disk B** and **do not move disk A** (positive.) You will notice you are getting a small amount of voltage and amperage. Now Move **disk A** and do not move **disk B. THE POLARITY CHANGES**.

The rule here is; The disk that moves is always positive. since we can get a polarity change just by moving the disks at opposite times. You can make a very simple mechanical device to get AC current from on top of the earth. Again, this is not just a chemical reaction we have done many test to prove this. You can also get a small voltage reading in dry dirt. The reason for the smaller voltage in using dry dirt is because a low conductance resistance is taking place here, keeping the system from being in contact with the earth and collecting the earth **ether energy.**

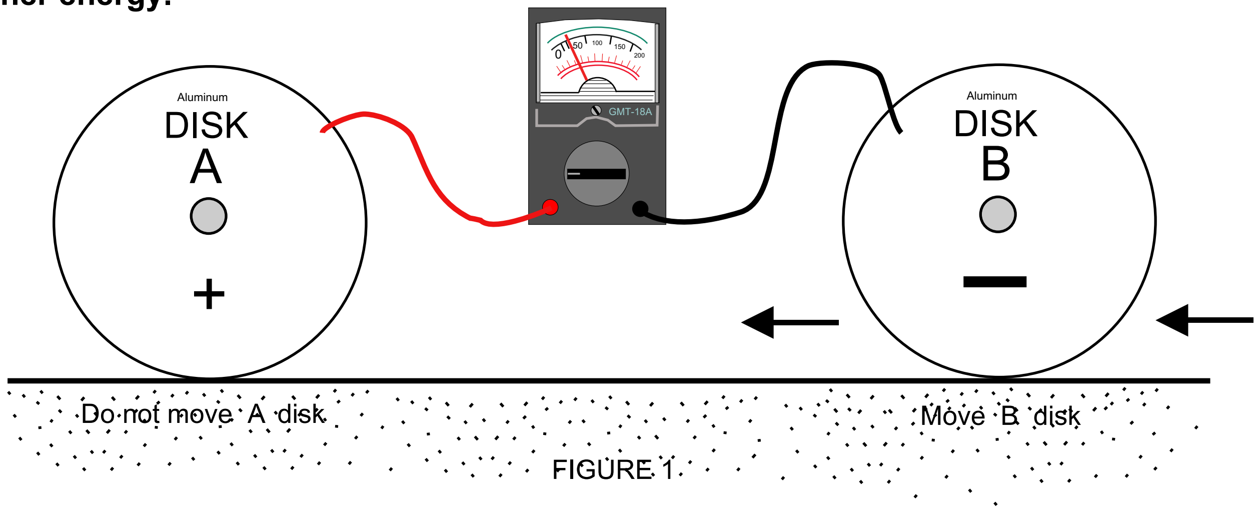

FIGURE 1

Copyright 2002 - 2004 Creative Science & Research

Capacitor Sheet Method

This method is far better than using pipes or rods. By using copper and zinc or aluminum foil sheets you will get much more amperage out of your system!

The amperage you will be collecting will be coming from 3 different sources,

1. **The acid in the ground and water**
2. **Energy that is being transmitted from the earth itself**
3. **Energy that is being transmitted from the sky and space.**

This may all seem unbelievable but it's true and is a fact! The more plates that you add the more power you will get! The copper sheet(s) is the positive, it must be facing down toward the earth. The aluminum or zinc sheet(s) is the negative and must be facing toward the sky! Between each sheet place 1- cotton or course screen printing mesh or other plastic course mesh cloth material. you can use a hole puncher to punch exact holes in each sheet, use 2 holes per sheet copper and aluminum. Use a 8 1/2" x 11" paper puncher one that is adjustable. next use a 3/4 wood base and drill hole for wood dowel rods to fit to the holes you punched in the sheets, you will have a total of 4 wood dowel rods, 2 on the right for the copper sheets and 2 on the left for the aluminum sheets to fit over, cloth die electric can be done the same. You then need to assemble dry and use another 3/4" wood board for top holding plate, which will fit over the wood dowel rods on top, sandwhich and press the assembly together and tape in place with duct tape, drill 2 long bolt holding holes on bottom wood base and top wood base then attach long bolts and nuts and tighten down. Remove tape. and attach wires, water the entire capacitor assembly and bury in the ground.

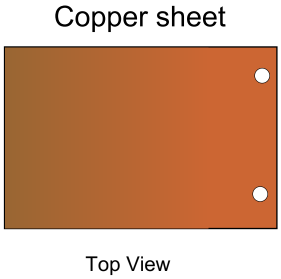
Copper sheet
Top View

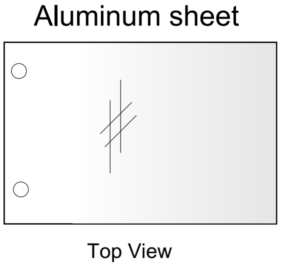
Aluminum sheet
Top View

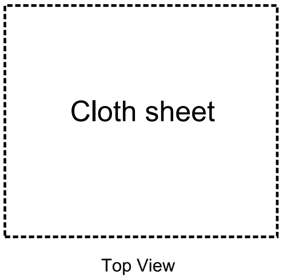
Cloth sheet
Top View

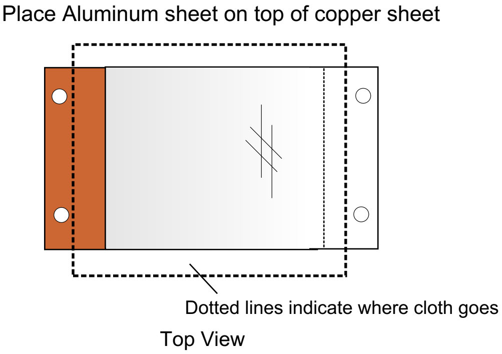
Place Aluminum sheet on top of copper sheet
Dotted lines indicate where cloth goes
Top View

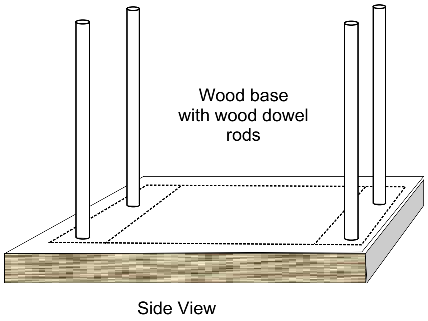
Wood base with wood dowel rods
Side View

Copyright 2002 - 2004 Creative Science & Research

Capacitor Sheet Method

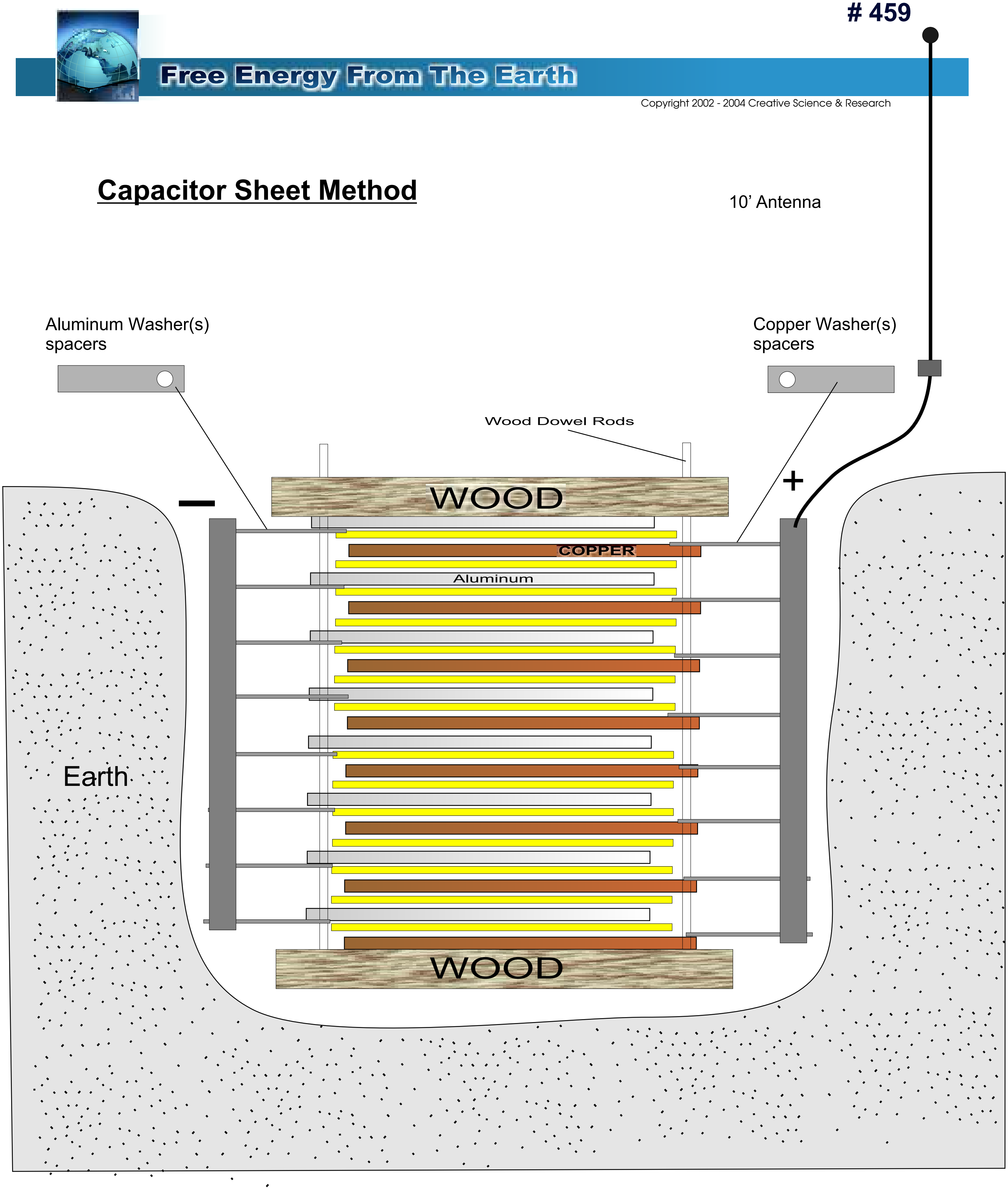
10’ Antenna
Aluminum Washer(s) spacers
Copper Washer(s) spacers
Wood Dowel Rods
WOOD
COPPER
Aluminum
–
+
Earth
WOOD

Copyright 2002 - 2004 Creative Science & Research

Capacitor Sheet Method

Again, the more plate sheets of metal you add (or cells) the more power you will get! Remember it's all about capacitance. You are collecting more than just simple battery type power. The copper plate sheet is the Positive + and the Aluminum sheet is the negative. There are many ways in which one can build this type of earth energy collector, below is simply an example of another shape and form. In this type we use water holes. water holes must all be in exact spots.

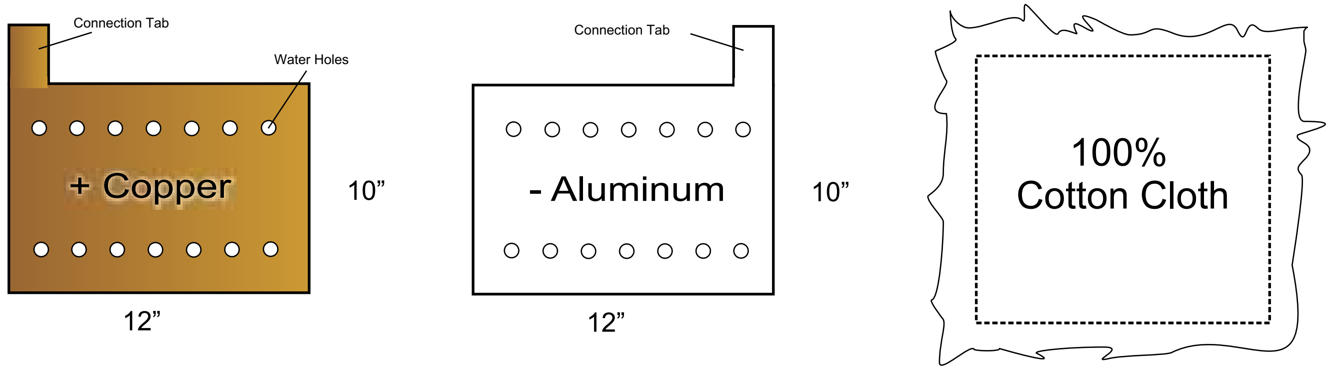
Connection Tab
Water Holes
+ Copper
10"
12"
Connection Tab
- Aluminum
10"
12"
100%
Cotton Cloth

Figure #1 Top View of separate pieces

Figure #1 Side View

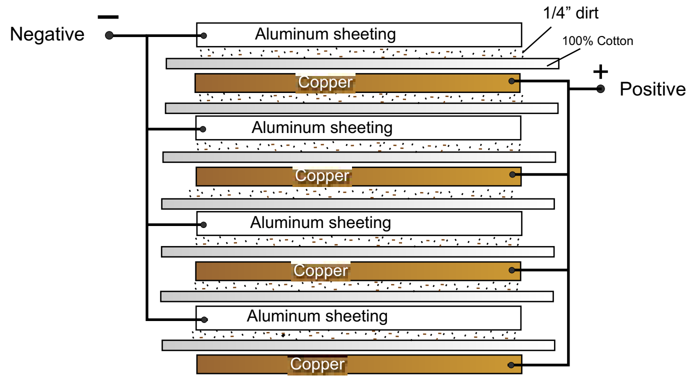
Negative
-
Aluminum sheeting
1/4" dirt
100% Cotton
Copper
+
Positive
Aluminum sheeting
Copper
Aluminum sheeting
Copper
Aluminum sheeting
Copper

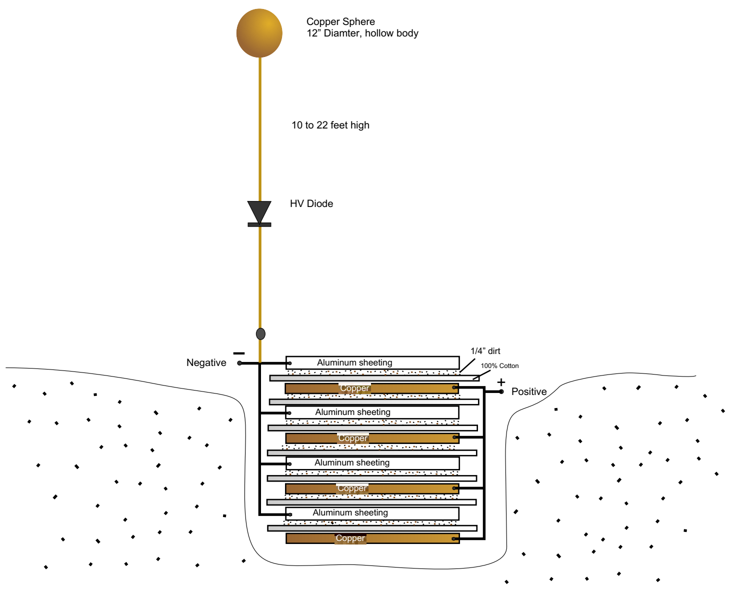
Copper Sphere
12" Diamter, hollow body
10 to 22 feet high
HV Diode
Negative
-
Aluminum sheeting
1/4" dirt
100% Cotton
Copper
+
Positive
Aluminum sheeting
Copper
Aluminum sheeting
Copper
Aluminum sheeting
Copper

Copyright 2002 - 2004 Creative Science & Research

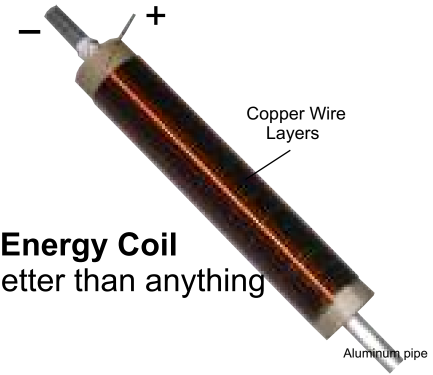

OUR MORE ADVANCED EARTH CAPACITOR!

(EECC) Earth Energy Coil Capacitor!

To collect more ether earth energy atoms, build the following **Earth Energy Coil Capacitor**, **Designed & Created by Creative Science & Research.** This is far better than anything we have found yet! Much more power and is easier to build.

Start with our EECC Model #1, Take a 5/16” zinc rod or aluminum rod, cut to 7.5” long. Now use #27 bare copper wire, secure it in a drill press (which you will then turn by hand). Cover zinc rod with paper one layer, use very tiny bits of tape to secure. Now apply the copper wire over the paper, remember to use copper that is not coated! Tape one end of the copper wire to the end of the zinc or aluminum rod and begin to slowly wind. Each wind must be perfectly fitted side by side, paper should be 5.5” long, total length of winds should be 4” long centered on paper. Once you are done with the first layer, apply another piece of paper to the top layer of the first copper layer, do not use to much tape to secure paper in place! Now begin your 2nd layer, repeat this same process until you have about 10 layers, (the more layers the better!) Copper must not touch the zinc or aluminum. When done hot glue or epoxy a small end piece of the copper wire so it will not come unwound. This is one complete cell, keep in mind this is a small test version, to get more power you have to THINK BIG! Build larger cells, using larger size copper. Then place them in salt water to test, 2 cells can be connected in series, after that you must connect them by use of capacitors, diodes and electronic or hand wound spring switches. If you build 20 of these and try to connect them all in series without a cap/ diode switching method, the cells will cancel each other out. Keep in mind you want water to get to each layer as well as the zinc or aluminum. You can also use Aluminum or zinc pipe.

Now concerning the paper Die Electric, Paper will not last very long, test and find a water absorbing plastic die electric,(mesh type). We suggest that you use Plastic screen mesh that you can buy at any hardware store, if all you are interested in is collecting small voltages then use one layer in between each layer of copper, if you are interested in collecting higher voltages such as lightning charges then you must use more layers of plastic screen mesh to thicken the die electric so the higher voltages will not short out your system. Paper may not last longer than 1 year, Plastic screen mesh or polyester screen printing mesh will last for many years.

EECC Model #2 Same as #1, but different. In this design we use #27 copper COATED wire, instead of bare wire. since the copper wire is coated and not bare, each layer must be sanded, that is, the top portion of the copper coated layer must be sanded off so it will be exposed to the water and acid, a die electric should still be used in this case. This type makes for a very high efficient earth cap!

EECC Model #3 Same as #1 but reversed! You will use aluminum wire and a copper rod or pipe. Use a 3/4” copper pipe and wind the die electric paper or plastic mesh and Aluminum bare wire around the copper pipe as explained in Model #1

Copyright 2002 - 2004 Creative Science & Research

OUR MORE ADVANCED EARTH CAPACITOR!

(EECC) Earth Energy Coil Capacitor!

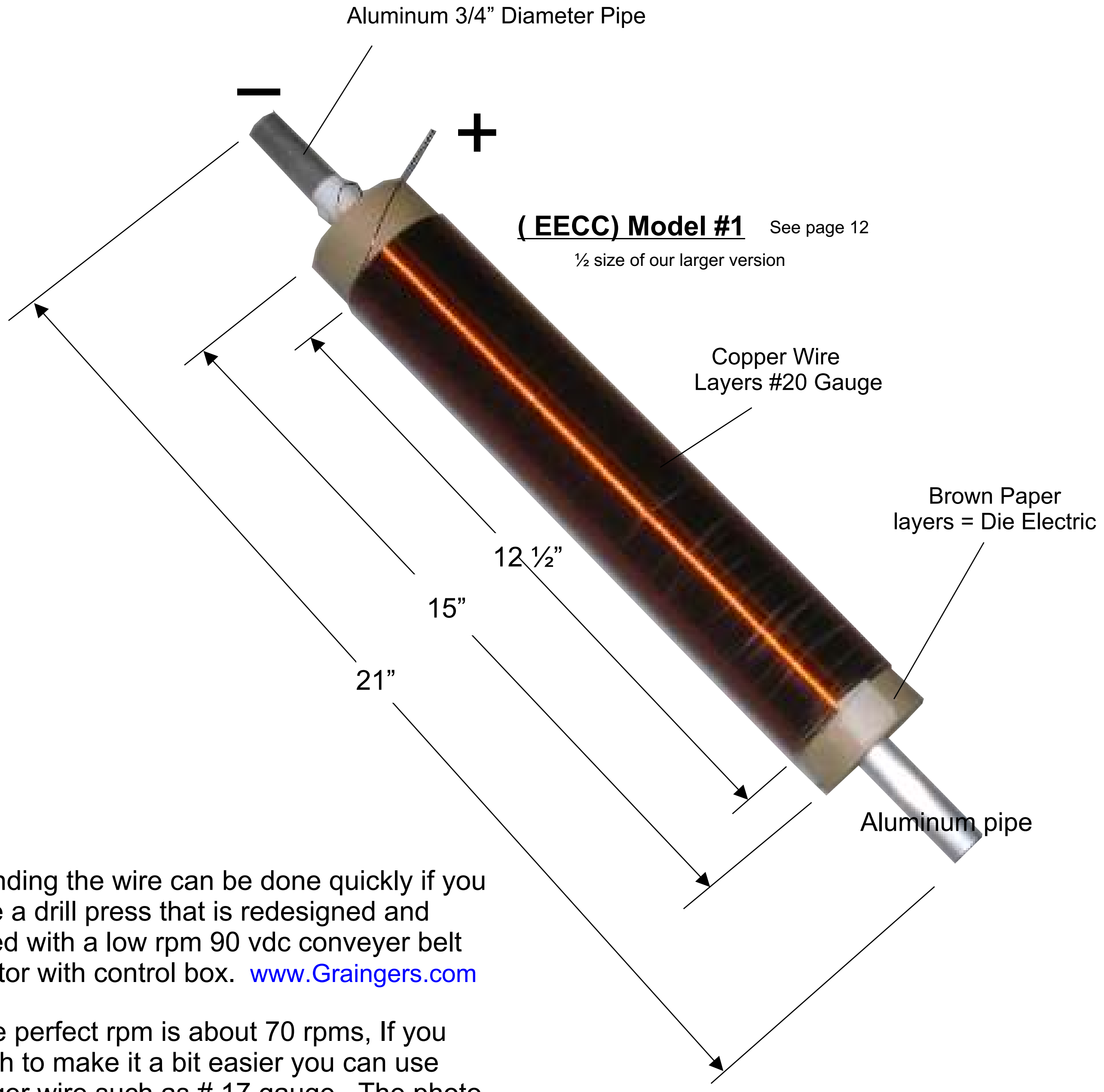

(EECC) Model #1 See page 12

½ size of our larger version

Aluminum 3/4” Diameter Pipe
–
+
Copper Wire
Layers #20 Gauge
Brown Paper
layers = Die Electric
12 ½”
15”
21”
Aluminum pipe

Winding the wire can be done quickly if you use a drill press that is redesigned and fitted with a low rpm 90 vdc conveyer belt motor with control box. www.Graingers.com

The perfect rpm is about 70 rpms, If you wish to make it a bit easier you can use larger wire such as # 17 gauge. The photo is an example of a mid size cell and can be very powerful! This is ½ size of a large cell.

Copyright 2002 - 2004 Creative Science & Research

OUR MORE ADVANCED EARTH CAPACITOR!

(EECC) Earth Energy Coil Capacitor!

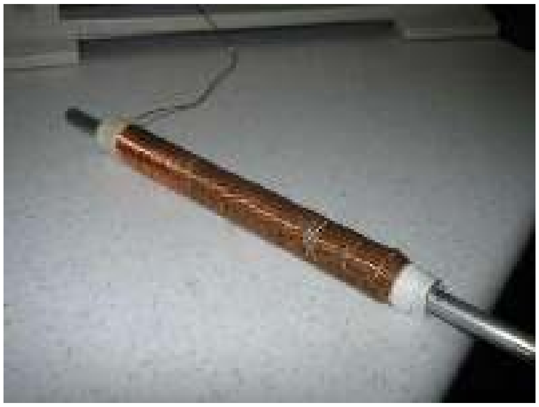

Our small test model #1, we used #22 bare copper wire wrapped around a zinc rod, each layer had a paper die electric, tested in tap water. 5 layers. Output = 89 vdc x 3.5 ma, this is low. more layers will = more capacitance which ='s more amperage and voltage.

Small Drill press turned into a Coil Winder

The more layers, the more free earth energy you can collect!

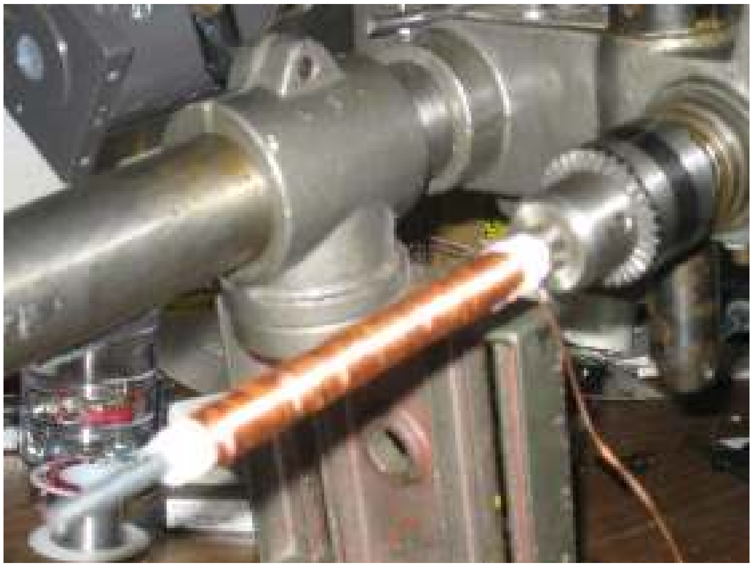

Small Drill press turned into a Coil Winder. We turned this by hand. Larger coils is best to use low rpm motor at about 70 rpms, In this picture you see that we wound each wind side by side, this makes for a more high efficient earth cap. You can wild wind if you like if you use smaller wire like #31 to #34.

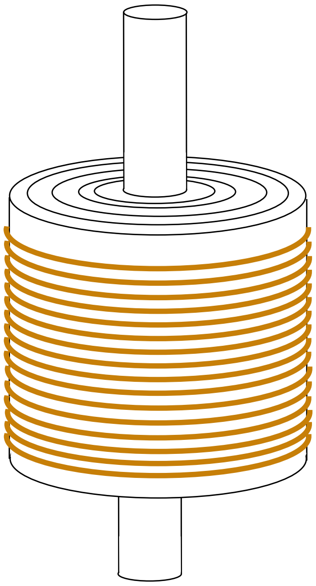

Side View
Ideal size for experimenting

Copyright 2002 - 2004 Creative Science & Research

(EECC) Earth Energy Coil Capacitor!

Rolled Earth Capacitor

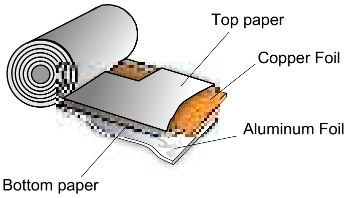
Top paper
Copper Foil
Aluminum Foil
Bottom paper

MODEL #4 Rolled Earth Capacitor

Copper foil is placed between 2 pieces of paper strips and Aluminum is placed on bottom. Use 400 foot strips 2" wide. This method is very hard to do. But is very high eff.

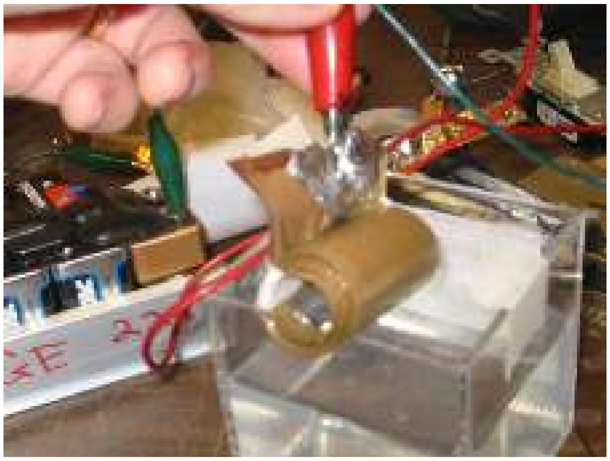

We used salt water to test rolled caps before placing them in the ground.

Copyright 2002 - 2004 Creative Science & Research

(EECC) Earth Energy Coil Capacitor!

MODEL #4 Rolled Earth Capacitor

Rolled Earth Capacitor

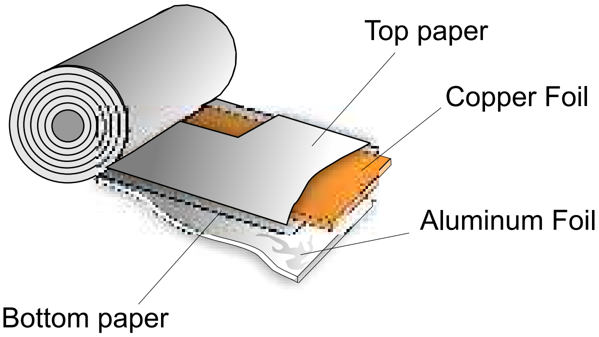

Again rolling this type of earth capacitor is very hard, Everything must be kept straight and inline with one another, Manufacturers of rolled electrolytic capacitors use computers and machinery. I am sure a simple home made roller with pin registration could be designed very easily. This type of earth battery is very high efficient. Paper is not a good die electric for high voltages! If you are wanting to collect earth and storm HV Currents you must use a thicker die electric. Metals must be spaced properly for storing high voltage.

Copyright 2002 - 2004 Creative Science & Research

(EECC) Earth Energy Coil Capacitor!

Rolling a capacitor is the best way but not the easiest,
Which makes this pipe coil model much more desirable to build

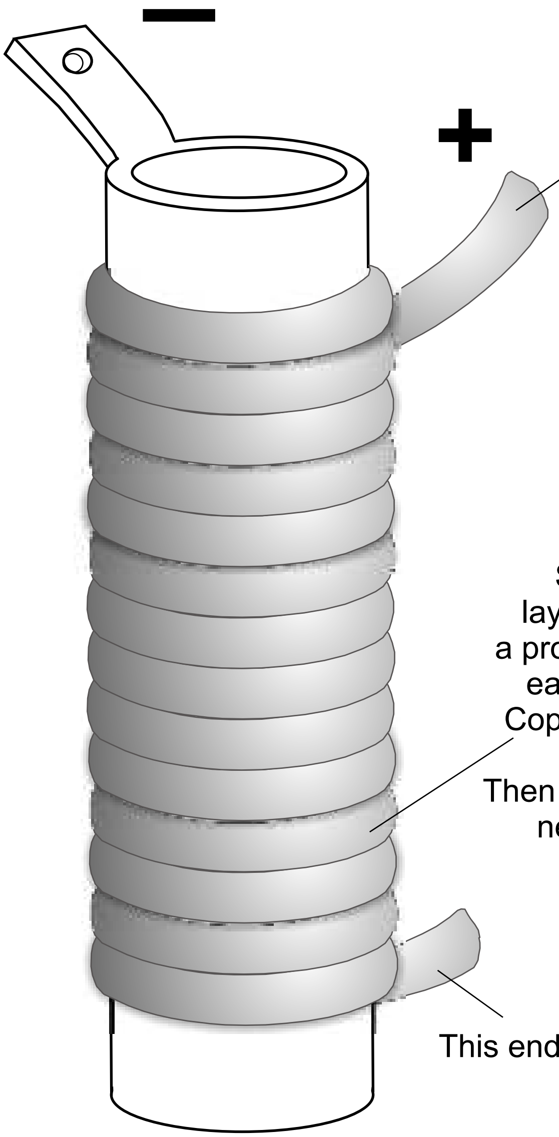

Sand or scratch off top layer coating, this will leave a protective coating in between each wire and underneath. Copper should not touch itself or the aluminum pipe. Then cover with paper and begin next layer of copper wire.

Symbols

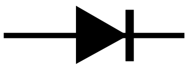

= One way DIODE

Photo of Diode

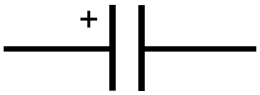

Capacitor (Electrolytic)

Photo of one type of Capacitor

Copyright 2002 - 2004 Creative Science & Research

CONNECTING THE CELLS IN SERIES

How to connect the cells in series to increase the output voltage without allowing the cells to cancel each other out. Figure **A** The Copper pipe and zinc rod method without the use of plastic on bottom of pipe. **Figure B** is the coil earth cap method.

Figure A

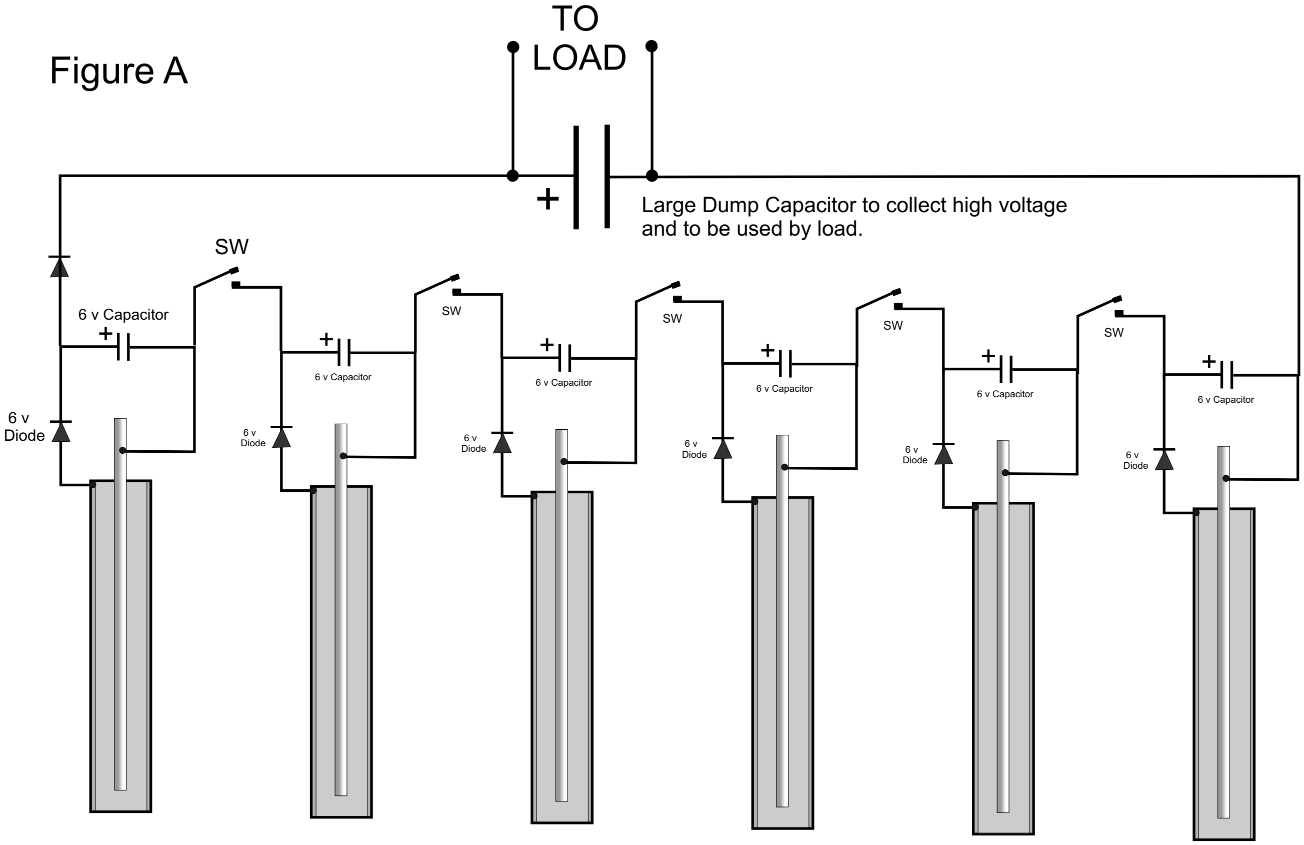
TO LOAD
+
Large Dump Capacitor to collect high voltage and to be used by load.
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode

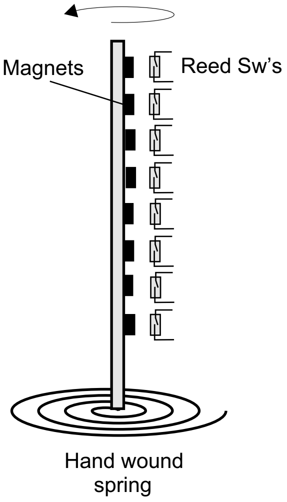
Magnets
Reed Sw's
Hand wound spring

Use electrolytic capacitors to store the charge coming from each cell, then connect them in series using Commutator Sw's connected to a small rotor shaft, connected to a milliamp solar cell hobby motor or build a small fuelless engine electric motor to run on and off switch's. You can also build a very low milliamp electronic switching circuit and use the earth energy to power it. Or another more simple way would be to use a hand spring type device (timer type) glue small magnets on a long rotating rotor arm connected to the inner spring. Then place small reed relay magnetic switches, on each timing magnet ON position. should have multi reed sw's, as the unit turns if gives time for the caps to charge and then comes around and connects caps all in series which in turn discharges into main cap dump.

Free Energy From The Earth

Copyright 2002 - 2004 Creative Science & Research

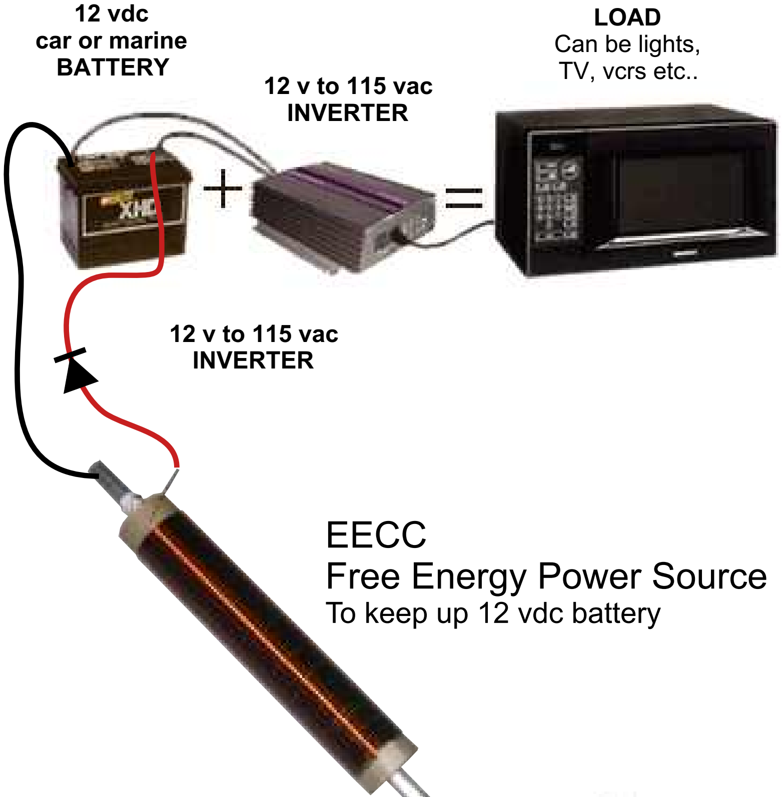
12 vdc
car or marine
BATTERY
12 v to 115 vac
INVERTER
LOAD
Can be lights,
TV, vcrs etc..
+
=
12 v to 115 vac
INVERTER

EECC
Free Energy Power Source
To keep up 12 vdc battery

Copper

Symbols

= One way DIODE

Photo of Diode

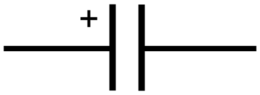

Capacitor (Electrolytic)

Photo of one type
of Capacitor

Copyright 2002 - 2004 Creative Science & Research

Suppliers List

Free News

Google™

Copper Foil Search

See www.google.com

1. Http://www.delphiglass.com/

   Use there search engine to find copper foil tape

2. http://www.glassmart.com/foil_menu.htm
3. http://www.globalsources.com

Search Google.com for:

1. Copper flashing
2. Inverters
3. Batteries
4. Copper coated wire
5. Copper wire
6. Electronic parts and supplies
7. Diodes
8. Capacitors
9. Free energy from the earth
10. Zinc rods

Copyright 2002 - 2004 Creative Science & Research

EECC update

Step One

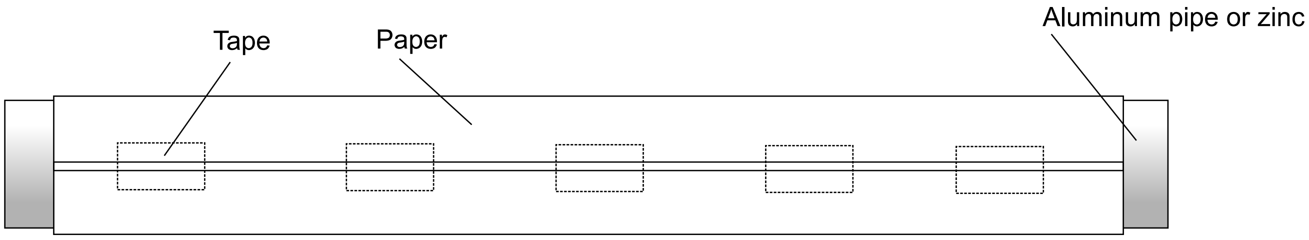
Tape
Paper
Aluminum pipe or zinc

Step two

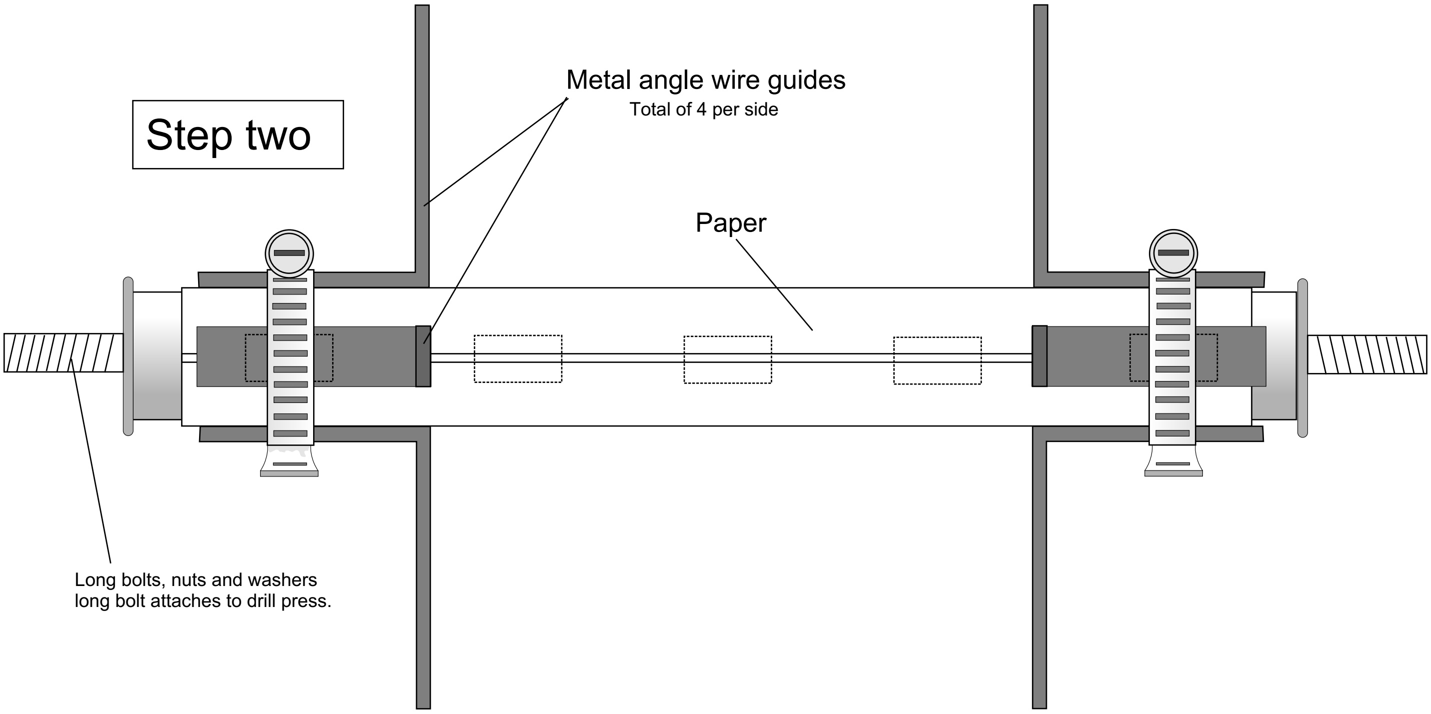
Metal angle wire guides
Total of 4 per side
Paper
Long bolts, nuts and washers
long bolt attaches to drill press.

Apply first layer of wire, remove, metal wire guides, apply more paper, then repeat over and over again until you have about a 5 to 6 inch diameter roll of wire and paper. Makes a very powerful earth capacitor!

Copyright 2002 - 2004 Creative Science & Research

EECC update

These are just a few extra idea’s that we threw together. I hope they are helpful to you. Let us know how you like these plans. Send us photo’s of what you have done we love to hear from our customers. If you got any new free energy idea’s let us know we can research them for you. Many customers do not have the extra research time that we have.

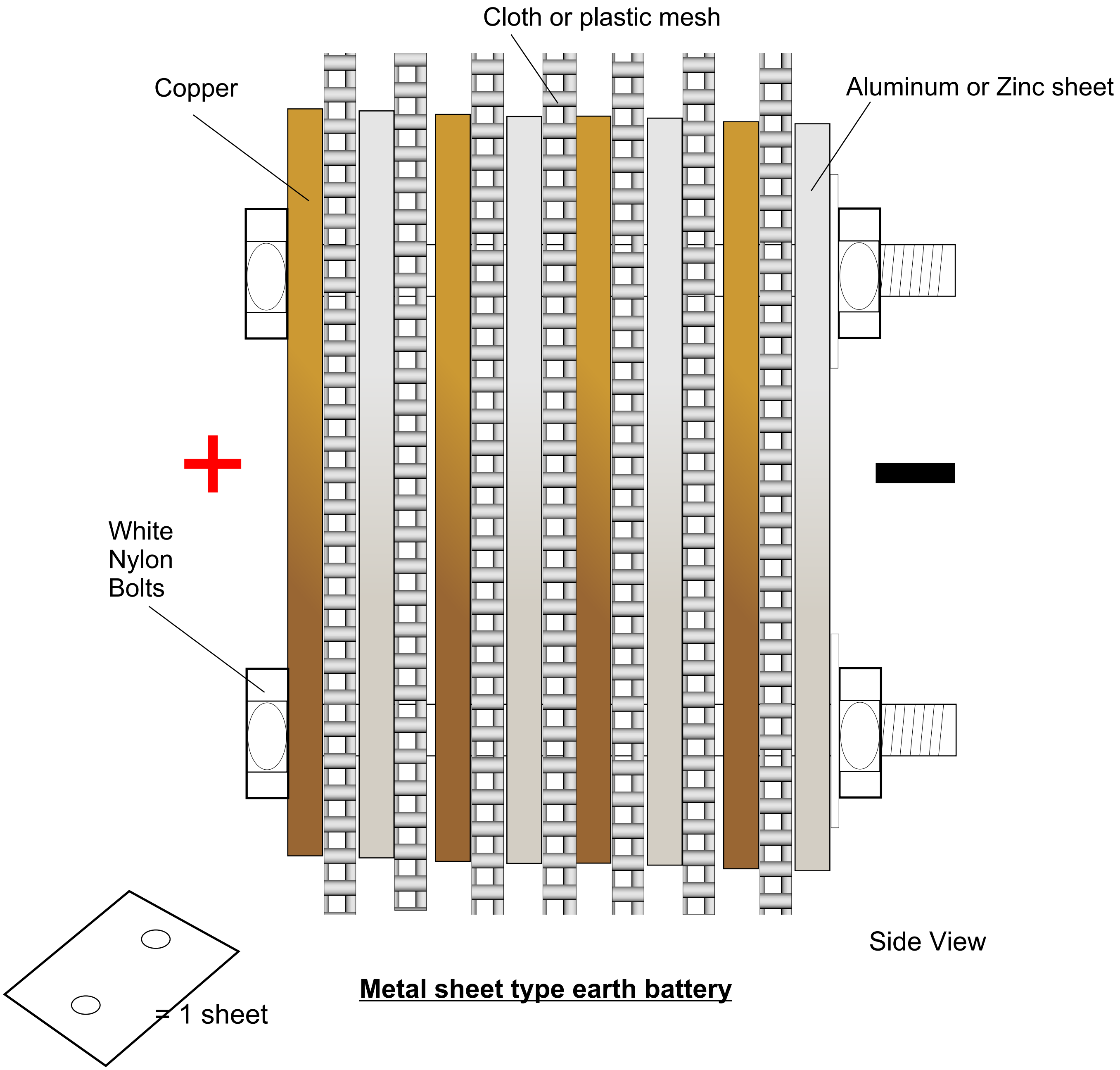
Cloth or plastic mesh
Copper
Aluminum or Zinc sheet
+
White Nylon Bolts
−
Side View
= 1 sheet

Metal sheet type earth battery

Fighting the pollution by public radio broadcast antennas

Un texte en français :	La pollution radio par les émetteurs FM
Een tekst in het Nederlands:	De radiovervuiling door FM-stralen
A petition:	www.petition-electrosmog.be

The electromagnetic pollution has now reached such a level that not only hypersensitive persons are being harmed. The debate most often focuses on cell phones, their relay antennas, Wi-Fi routers and similar technologies. It is often pointed out that domestic appliances like power saving lamps and cheap electronic devices, are harmful too. Recently, I came to the conclusion that I too had problems with electromagnetic radiations. Yet after a few months investigating, I was surprised to find out that far out the main cause of my problems was nothing I had heard about. A 100 MHz FM public radio broadcast antenna is located atop a local hospital and burns the whole city with an astounding force of 2 V/m (pike value).

I found other persons that are harmed. For example people who develop severe headaches or tiredness after a few hours working in a given place. They have no problems in other places, that are less exposed to the FM radio waves. It seems to me that people who live in strongly exposed habitations often develop psychiatric problems. My guess is that probably several thousand persons in the city are severely affected. Maybe tens of thousands are affected but without getting obviously sick.

Most puzzling is the fact that this hammering by radio waves is completely useless. It wouldn't change anything to the quality of the reception by the listeners in the city, if the force of the radio waves was divided by 100 or preferably 1,000. The purpose of those antennas is to emit far away in the countryside. That's why they are located atop hills or tall buildings and they have a huge power. But, due to their rudimentary construction, quite much of that power is wasted and hits the nearby city.

It is legal to hit the ground with such force. Whether I measure in rooms or in the open, it is at worst a little below the norms. Hence the problem is with the norms. They keep being too high, despite the efforts of many scholars and health organizations. The situation is catastrophic. If the people did understand the harm that is done to their children, revolts would break open.

Below is a picture of a cheap calculator that I transformed. It contains no more battery nor solar power cell. Instead it draws its energy from a 1 meter long antenna. Anyplace in the city center, in direct view of the emitter, the power of the radio waves is strong enough to feed the device. I made this because nobody was frightened by the figures I was reporting, except some friends that are into electronics. Everybody knows that a car can kill even at 30 km/h speed but almost nobody has any knowledge about radio waves figures. When I showed the calculator, everybody understood and some whitened. (The calculator also works in some classrooms where I tried it out. The emitter was visible through the windows... Should I warn the parents of the students that failed their exams?)

Powered by 000webhost

This didn't help anyway, so I invented another device: the "snake". The picture below shows six of them, an early version.

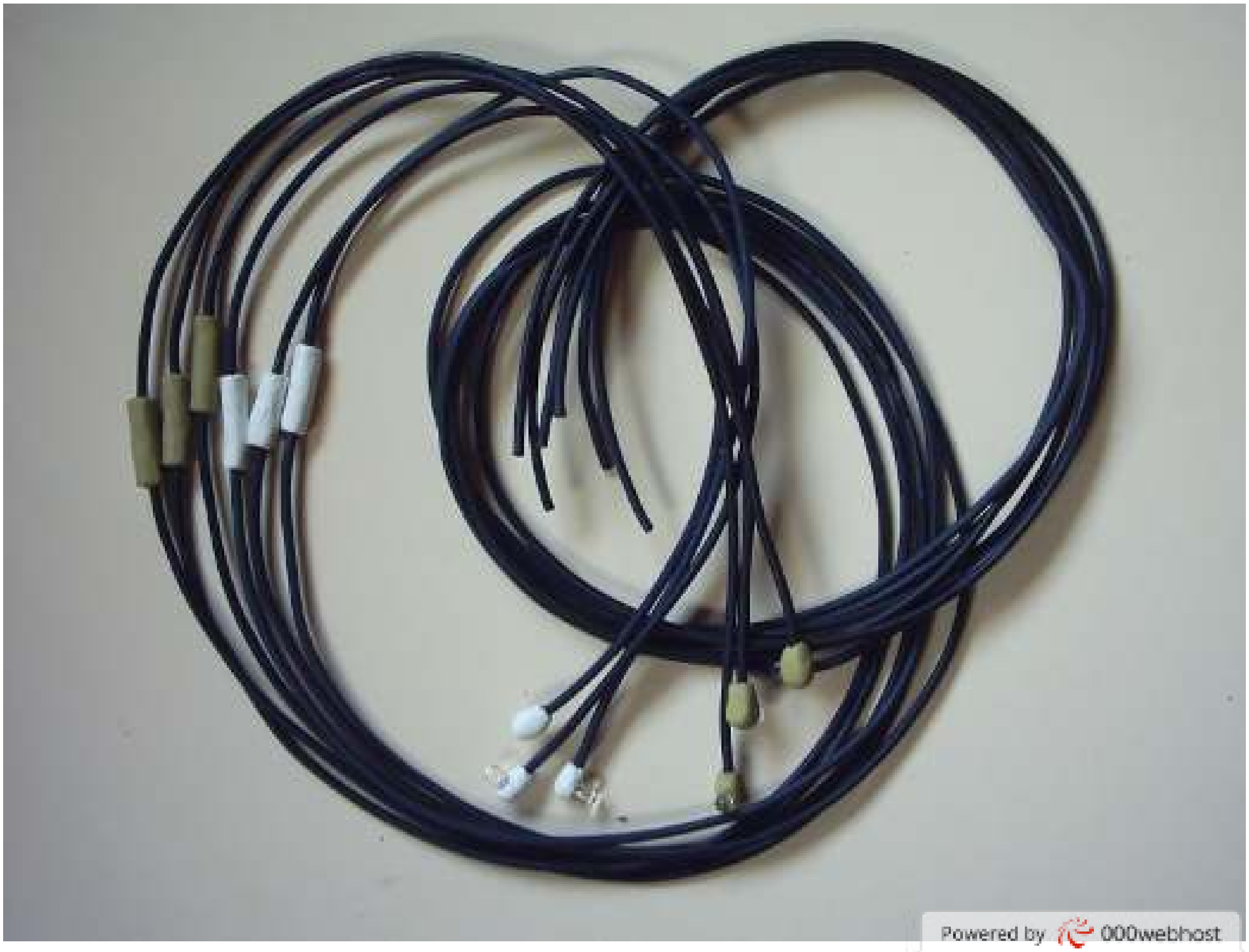

Powered by 000webhost

You just hang the snake in the flux of radio waves and the LED lamp at one end will bright up, clearly visible in daytime and enough to light and read a text at night.

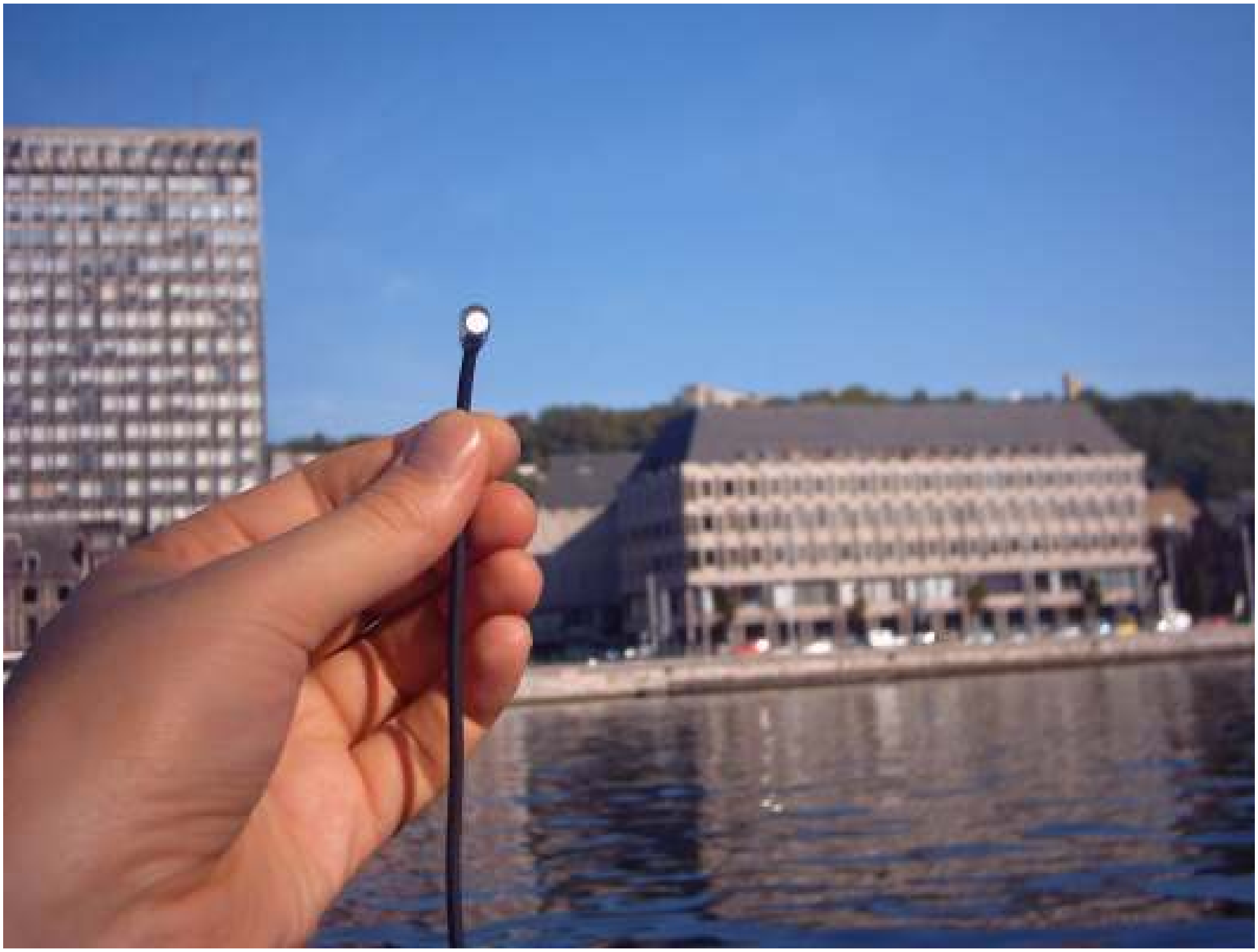

The snake is less impressive than the calculator but it is much cheaper and easier to make and it can be sent in an envelope. I made tens of them and sent them to politicians, newspapers, universities... I gave some to local people, together with a user guide on a sheet of paper. Just like the calculator, the snake contains no battery. The lamp lights up solely on the power of the radio waves. The picture below shows a later version, that can be rolled up in an envelope that fits the conditions to be sent with only one stamp.

Powered by 000webhost

However mad it is to send such power on a city, nothing has been done yet against this. It's like an 19th century chemical plant sending its pollutants all around. Some complain but others argue that closing the factory would mean people laid off. Some then reply that being employed to feed children that die from ugly sicknesses may be nonsense... And so on, except for the fact that the chemical factory indeed had a purpose while the strength of the radio waves we're talking about, has no serious purpose at all.

Something common to all heavy sources of electromagnetic pollution is that it wouldn't cost much to make them harmless. For example, high tension power lines can be made harmless by feeding them DC current instead of AC current. This requires some more heavy electronics at the input and the output but the price of it is a detail in the whole. I studied the schematics of power saving lamps and simply noticed that they were conceived by people who do not understand the propagation of radio waves. Adding the necessary components to make the lamps radio-silent would almost not increase their prices. Enforcing proper norms will not change our way of life, it will just stop to favor the sociopathics amongst the industrials and lobbyists.

So, what can I do to help people understand the situation? They need to appropriate the scientific knowledge involved. Most assume that the necessary studies will be made by labs and then the governments will adapt the laws accordingly... That would be OK if it wasn't for the trillions of money involved. A minister in Belgium tried to prevent sugar dispensers in schools, she immediately got a made-up scandal on her shoulders. All she ultimately managed to get was to have some health advice notices glued on the sacred dispensers...

Flows of observations are now available, showing the effects of radio waves that match the norms: brains of rats destroyed, people unable to work, eggs that don't hatch, quarreling birds, increased risk of cancer, cows aging abnormally... I myself had seeds germinate and saw them die quicker when closer to a power saving lamp. Their strategy to cope with this is simple: "OK, there does exist scientific evidence of some danger... but there also does exist evidence that there is no danger at all! So let's balance the whole..." The norms are lowered but kept above what cell phone operators need for the current systems. The fact in itself of making a balance between scientific evidence is nonsense. Either a study proves a given danger and it can be verified, or bust. But anyway, what are those studies that tend to prove that there would be no danger? Lots of them are simply payed for by the cell phone industry... Yet some are perfectly serious, for example those concluding that radiations within the norms will not heat the brain or the body abnormally. This is perfectly true. There is no problem of scientific fairness involved with these studies. But, is a bullet harmless because it only very slightly heats your bones? A more sophisticated approach is to pick out studies that have borderline conclusions: "there may be a problem but it is not clear..." The study will be assumed to conclude that there is no problem at all, while it just was an inconclusive experiment like many are in science.

Powered by 000webhost

(When I'm exposed, I tend to compulsively perform simple tasks. What if a study examines if being exposed hampers the ability to perform that kind of tasks?)

Then, there is the problem that I'm going to make people sick. Indeed, electromagnetic hypersensitivity has a psychological component. Once people understand that they are victim to the radio waves and what symptoms the radio waves cause, they can overreact and become very sick when they start feeling the symptoms or just when they know there is a strong radiation force around. The radio waves induce a physiological stress, that hampers the proper functioning of the brain. When you add a psychological stress, due to the awareness of the problem, the summation of the two multiplies... Therapies for hypersensitive persons imply to dislearn them to overreact. Some friends of mine just refuse to start talking about electromagnetic sensitivity because they don't want to embark in the sickness. They do so with everything, meanwhile they eat very good food and have lots of pleasure in life. It works... with people with a strong health and a tad of selfishness.

When you are in a cold wind and you feel a pain in the neck, you know it is due to the wind and you take cover. Almost nobody can do this with the symptoms of an exposition to strong radio waves. You feel tired, some brain fog, like if your head was in a clamp, a headache, or simply you make stupid errors... You will blame yourself or the location, possibly the noisy people around... while in fact you are mainly undergoing electromagnetic sensitivity. The home of a friend is strongly exposed. After about twenty minutes I can nor more find my words and speak correctly, unless I wrap my head in aluminum foil. There is a fair chance that a while ago, I would have blamed my friend for being annoying (he is) to the point of making me speechless. There also is a fair chance that my friend is annoying because he hangs in that electromagnetic pollution everyday. Since he lives there, he became an alcoholic, lost his job, manages to keep no girlfriend... (Everything I mention here can have other causes. Quite often, such causes just add up. For example, the person blaming the noisy people around, would maybe have felt nothing if the noisy people weren't there. Like you don't feel your bad knee when you carry no bag.)

The impact of the electromagnetic pollution is severe. It is a far too heavy, and pointless, contribution to our burdens and losses. But how do you establish, scientifically, be it as an amateur, that there indeed is a problem? My initial approach can be an illustration of this. As my home is uninhabitable, I spend most of the day working in local libraries. I noticed that I could work quite well at some places and not at all at other places. Then I saw an article taped on a board, about the effects of cell phone and Wi-Fi radiations. Okay... I made a spreadsheet, with a row for each location and a note from 0 to 10 according to my ability to work there. Although I knew it was a weak approach, I then filled in each row the strength of the closest Wi-Fi router, as listed by my laptop. There was no correlation... but this proved not at all that the Wi-Fi was innocent. Weeks later I made a nonsensical but working little assembly that made the variability of the radio waves hearable. This time I got a correlation. Mild, but obvious. The ability to work is horizontally while the intensity of the radio signals is vertically:

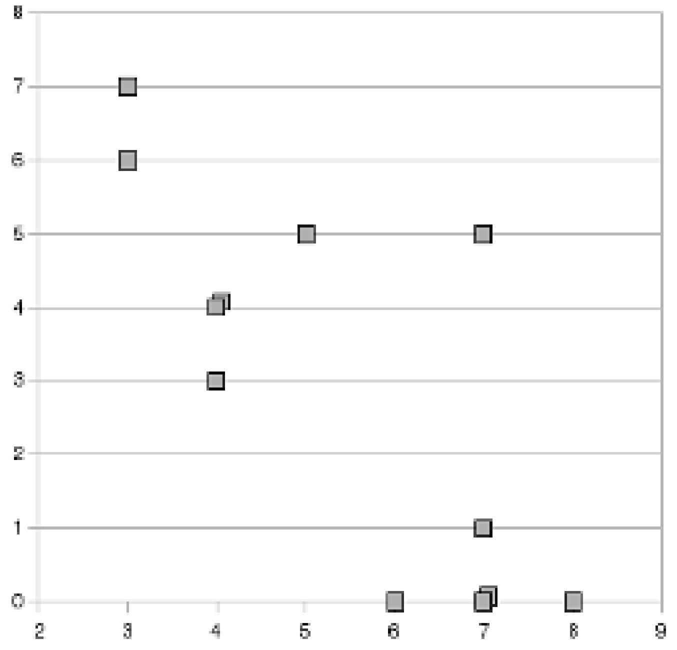

It quickly became obvious that I was measuring no strong signal in the GHz bands. I had no doubt that there were cell phone and Wi-Fi emissions coming from all directions and echoing all around but their strength was obviously

Powered by 000webhost

weak. The problem was with lower frequencies. I accused the fluorescent lamps for a while, till I finally understood that there was only one strong source, hitting me everywhere; those 100 MHz public radio broadcasts. The picture below shows the final correlation. Only one dot disagrees with the theory:

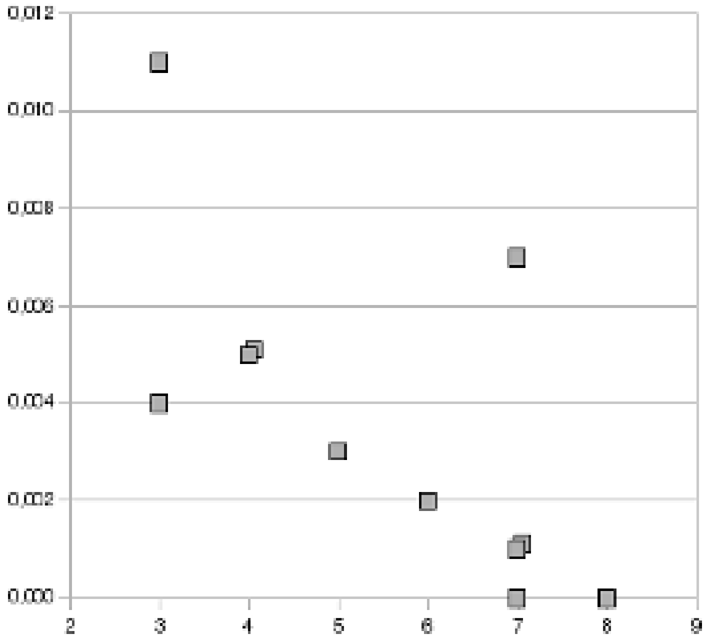

I was heavily disturbed by the radio waves, simply unable to work, long before I started puzzling on the subject. This is a strong proof that the matter is not just psychological. Reciprocally, once I understood the psychological component, I tried to use it positively, to counterbalance the effects of the electrosmog. I did get some results but I was never able to work correctly in the most exposed places. It is by no means "just psychological". It is useless "to be willing to make a small effort". I did efforts that frightened other people. One thing that does clearly help is eating a lot of sugar (Organic jam on rice bread...) (by the way, is that why sugar dispensers are placed everywhere kids are supposed to think and learn, to counterbalance the effects of the nearby cell phone relay antennas?) Alcohol does also help and I'm afraid this is why people seem to become alcoholics in exposed habitations.

Also, it does sometimes happen that I feel like undergoing strong radio waves but when I pop out my measuring device, I measure nothing serious. I have a natural tendency to be knocked-out... The radio waves are far out my heaviest problem but they are not always responsible.

I started studying other sources of radio pollution. I had an obvious benefit of removing every power saving lamp in my home. I put ferrite cores on most power chords. On one power chord that I sometimes use to feed my laptop inside my shielded tent, I had to put ten ferrites till I felt no more direct effect. Something very interesting is that till then I had no problem when using my cell phone. You guessed correctly: I started having problems, to the point that I never more hold it against my ear. I always use an earplug linked to the cell phone by a wire with two ferrite cores. I suppose that since I started avoiding the exposure to strong radio waves (which significantly improved my life and ability to work), my brains are no more constantly knocked out. So they get knocked out if I hold my cell phone against my head. And I feel it. This is one more synergy between the public broadcasts emitters and the cell phone industry: one hides the other away. 'Want to prove that inhabitants close to a relay antenna are no worse than those further away? Do the experiment in a city that is strongly exposed to public broadcasts...

So: buy a measuring tool or have a friend assemble one for you. And start comparing your problems and those of the people around you, with the level of electromagnetic pollution. If you can, make large-scale experiments with tens of persons, under advice of somebody with knowledge in statistics (to build an expressive statistic yet completely false, is *easy*). We need many such experiments, till it becomes unbearable for local interlocutors to negate the problem and not to dim the emissions. Step by step, we need to conquer our world back. Some enterprises and some countries have already set their norms at responsible levels.

One very important parameter is the rate at which the problems build up and then disappear, when exposed. Some people claim that they instantly feel when their cell phone is going to ring. I never could verify this. A friend claims that his cat reacts when he is going to get a text message on his cell phone. I tried several times to see the cat change

Powered by 000webhost

behavior while making my cell phone emit close to him but I got nothing. So I suppose that the cat reacts to the tone of the cell phone and my friend brain-confuses the timing. (I don't pretend nobody can feel a text message arriving, I'm just saying that I never could verify.) What I'm sure about is that my problems need minutes to build up, most often some ten or twenty minutes, the problem possibly becoming really painful after an hour. Once the problem built up, it needs much more time to diminish, maybe four times more. I often had to sleep a whole night on it. One day I "burned" my brains by holding a very polluting power saving lamp about 3 decimeters away from my head. I only noticed the problem after half an hour. But then it was horrible, like if a flame had been sweeping inside my head. Really just like the ache of a burn wound. I had a strong nausea and headache, that calmed down only hours later. I identified a few persons whose problems are obviously due to the electromagnetic pollution, in all cases the same applies: it needs time to build up. This means you cannot simply walk in and out of an exposed zone and claim you feel no difference. Also, you cannot get into a safe zone and claim you feel no better. If you were already poisoned, you will stay that way a long while. The only sure approach is to stay really long in a safe zone, preferably spend the night there, then go working in a polluted area. Try to be aware of what happens to you. The next day go working in a safe zone and try to see the difference. (This is a goldmine to build inconclusive studies.)

Everybody is sensitive to electromagnetic waves. The question is about the level of the sensibility. That's why people who get seriously sick with the current levels of exposition are called "hypersensitive". They are more sensitive than the average, sometimes with very painful and debilitating consequences.

There is a whole ladder of symptoms. This is roughly the ladder for me, exposed to the 100 MHz FM broadcasts:

0.01	V/m	No problem
0.1	V/m	No more able to concentrate and work
1	V/m	head in a clamp, maybe difficulties to talk
10	V/m	Tried it once for 2 hours, had to stay the whole next day in bed with an awful headache.

The serious problems begin for me at 0.1 V/m. Yet I know some people that wouldn't be called hypersensitive and that get problems similar to mine at about 0.3 V/m. A friend spent with me the two hours under the 10 V/m radiation and claimed he endured no problem... but he made surprising errors on the way back. It is very difficult for me, in my city, to find locations where the force is around 0.01 V/m. Values between 0.1 and 0.3 are common. It is easy to find building with rooms reaching 1 V/m. The force is quite predictable, according to the location and the orientation of the building, how much it emerges out of other buildings... So, because many people are being affected and because the radiation is reasonably easy to measure or to estimate, making statistics linking common health problems with the level of exposition is quite doable.

Hypersensitive persons would be 1% of the population. The need to protect this minority is enough to stop those powerful broadcasts from reaching the ground. I don't understand why I have to add that everybody is being harmed, be it to a lesser degree. I met official people whose job is to help hypersensitive persons. They've seen their pain and anguish by themselves... but they feel clueless as to how to stop the emissions. It's a very strange situation, when even the people close to the steer and convinced of a problem, can't do anything to stop it.

It has been claimed that the hypersensitivity of some persons is caused by an accumulation of toxic metals or organic pollutants. In such cases, antioxidants can be of immediate help and detoxifications would greatly help. (Note that common medical analysis like looking for mercury and the like in samples of blood, urine or hair, will reveal nothing even under severe intoxication, because blood and hair renew constantly. Only a sudden intoxication can be revealed by such analysis. The toxics slowly accumulated in standing body cells like the brain... Best method seem to be to take a well-tuned quantity of chelating molecules and then perform an analysis of the blood or urine.)

Cell phone and Wi-Fi radiations are different, for several reasons: they don't spread the same way in the body, they are pulsated and their frequency make they will not target the same macromolecules in the body. It would seem that you need 3 times more force of FM broadcast waves than cell phone or Wi-Fi waves, to have a same global impact...

I was frightened when I understood that under those 2 V/m force of radiation, an electric current in the order of 1 mA oscillates through the body. Such a current at a high frequency of 100 MHz will interact with the body quite differently than a DC or low frequency AC current. You cannot make direct comparisons... Anyway, 100 mA of DC current can be enough to kill a person by electrocution. 1 mA of low frequency current is enough to feel the current as a tickle. What are the consequences of such a current on the chemical reactions in the liver, on the production of

Powered by 000webhost

hormones by glands everywhere in the body? What if the person being exposed is a pregnant woman, with the baby just at the worst place in the flux of current? The length of an "antenna" matters much to allow it to build up current. Teenagers have the "optimal" length for 100 MHz radio waves... One thing I'm sure about is that if the government decided that from now on the school kids will be latched electrodes on the body and a current of 1 mA send through several hours a day, whatever the frequency, the parents will not accept. But this *is* happening, through radio waves, and I could feel the neuroleptic-like effect of it. (By the way, this implies that those experiments that show brains of rats destroyed by cell phones waves within the norms, apply only to a lesser degree to human brains, because of the difference in size. Rats are close to the optimal length to be harmed by cell phone waves... The optimal wavelength to dammage teenagers is that of those 100 MHz public broadcasts...)

The radiations harm me more inside buildings than outside. I may get more sick under 0.2 V/m in a room than under 2 V/m directly exposed to an antenna. I don't know exactly why. Here are four possibilities:

- The roofs atop buildings act like prisms and send the waves towards the inhabitants beneath. Radio waves coming from above would be more harmful than when coming from aside.
- The radio waves resonate and echo inside the building and build up in metallic structures. This creates halos of proximity that will inject more power inside nearby living bodies. For example, while simply wrapping aluminum foil around my head, it is obvious that some designs of shielding will make me even more sick than using no aluminum at all. The same way, when some parts of the shielding of my tent wear off and stop to be conducting, I can get serious problems till I replace or complement the faulty area.
- Buildings contain their own sources of electromagnetic pollution. They add their effects to those of the pollution from the outside...
- Close to metallic structures, the electromagnetic field can become rotating. This because waves with equivalent strength come from different directions with different phase shifts.

Again, the four possibilities mentioned above are potentially very suitable to build experiments that demonstrate there is no problem. Perform the experiment in an anechoic place with the waves coming from aside and no other radiation than the one from the experiment... Just the opposite from real world situations but you can claim to be serious at avoiding parasitic variables...

If you want to find the FM emitters somewhere, fmscan.org is the reference I used till now.

This is the schematic of the "snake":

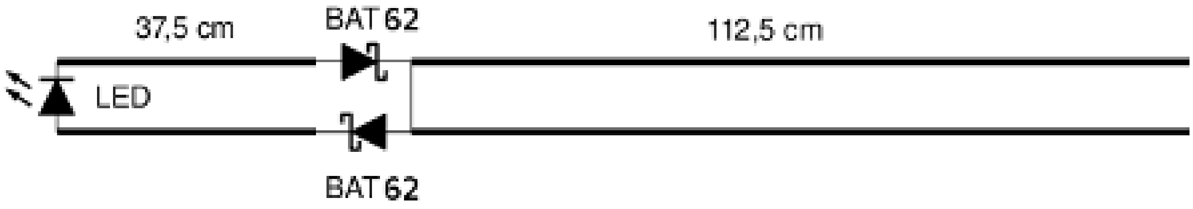

BAT62 detection diodes are no more produced. **BAT15** diodes work fine but they wear out; after a few months they stop functioning. **SMS7630** diodes are great but very little and mechanically fragile. **MMSD701T1G** diodes are sturdy and powerful; an excellent choice for a beginner. Such SMD diodes do also work for cell phone frequencies, which allows to test out a snake with a calling cell phone pushed against it. But any detection diodes that can manage 100 MHz will do.

The LED I'm currently using is the **L-7113SEC-H** . It lights up with a low tension and a very low current (the bluer a LED, the more tension it needs). Its color is red yet close to orange hence it is easily seen by the human eye (the eye is most sensitive to green, yellow and orange). The beam is quite narrow so when the LED is directed towards somebody's eyes it will appear quite bright.

For the lengths of 37.5 and 112.5 centimeters, any electric wire with two copper conductors will do. Audio signal wire is a practical solution. Use the shielding as one of the two conductors. The lengths of the two segments must not be precise. What matters is that the total length of the snake be 1.5 meters. Do not hesitate to try out if a little

Powered by 000webhost

longer or shorter snake gives better results.

A schematic of my current probes, that I connect to a standard multimeter, measuring Volts DC. The measure displayed by the multimeter must be multiplied by 10. When using a 200.0 mV scale, just read while forgetting the dot:

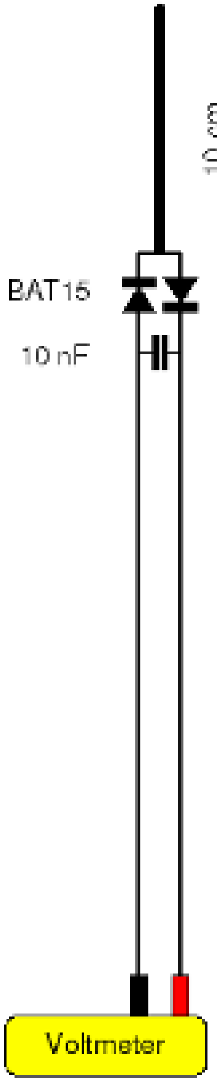

This is a picture of a device that tries to pack the whole in a neat gadget. The electronic circuit is a €7 digital Volts display with a scale of 200.0 mV. I soldered away the jumper that makes the dot be displayed. The antenna is only 7 cm long, so an RMS value is displayed (for roughly steady radio waves). The whole cost about €12. Hold the antenna perpendicular to the direction the waves come from and try to hold the device away from conducting objects like your body or metallic objects. You cannot closely trust such simple device yet it is very handy and adequate to compare the radiation level in different places:

Powered by 000webhost

Powered by 000webhost

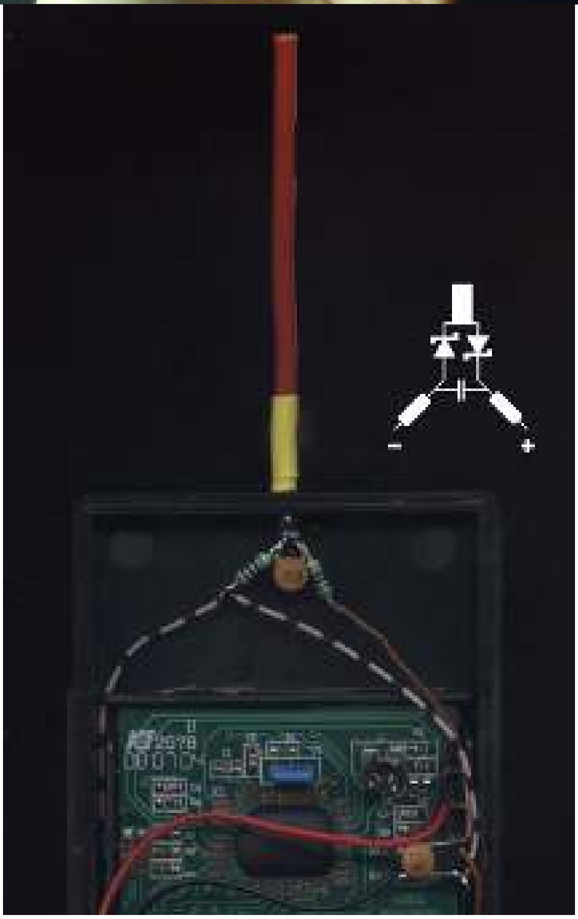

Powered by 000webhost

Those probes and device, quite easily allow to know the direction some 100 MHz radiation come from. Hold the thing at some decimeters from your body, roughly at the height of the belly. Your body and the device form a directional antenna... When the device is towards the emitter, it will display up to twice the value, while when it is hidden from the emitter by your body, the measure will be very low.

If you measure something strong, yet by turning in different directions the device always shows approximately the same value, then you are probably undergoing low-frequency radio waves like short wave broadcasts.

If the emitter is close, like a power saving lamp or a wrongly build power supply, converge towards it simply using the fact that the closer you get, the stronger the measure.

Such measuring tools are handy to find out power chords and appliances that pollute, yet then the measure displayed doesn't mean much. Once you are really close to an emitter, the measuring device and the emitter interact and the power flowing through the measuring device can be tremendous even though the device is only weakly polluting. You are in the "halo of proximity" of the device... Best example is most quality laptops, which emit almost no radio waves. I never could notice a problem while using my laptop when it is plugged to nothing... Yet if I hold a measuring tool against the screen I get frightening data... A few decimeters further, where my head can be, the measuring tool tells there is nothing... So, the high figures close to the laptop indeed tell that strong currents flow through it, but competent people made things such that those currents harm nobody...

If you want a global measure, away from the effects of your body, hold the device above your head. Or place it on some plastic or cardboard box and walk some distance away.

The windows are a common entry point of radio waves and electric wires can be awfull ducts, yet in some cases I noticed that metallic structures unrelated with electricity, like the copper tubes of water heating radiators, were the main ducts of radio pollution towards a room.

Below is a device I assembled to focus on 1.8 GHz cell phone radio waves. Close to the nearby cell phone relay antenna, in direct view of it, I could not measure much more than 0.1 V/m (pike value). While this is still too much, it is by no means comparable to the frightening 2 V/m intensity of the FM broadcasts. I also went measuring in places where people complain about problems similar to mine. They thought the reason was the nearby cell phone relay antennas. But I could measure no significant intensity of cell phone radio waves at all, while the force of the FM broadcasts was above 0.2 V/m...

Powered by 000webhost

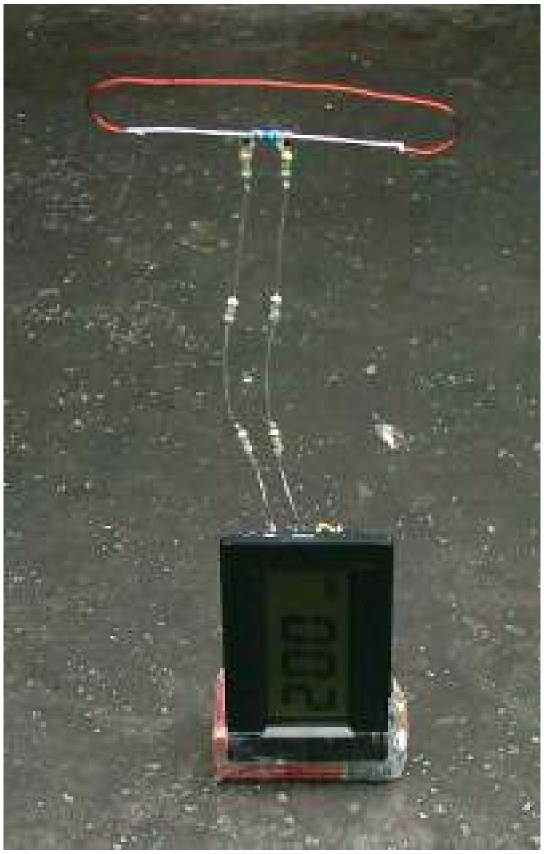

After about a year experimenting, the most simple protection I could come up with against 100 MHz broadcasts is a headband of aluminum, much the shape and size of a sweatband. I hide it inside a bonnet. I have no idea as to how and why it helps, aside from the general fact that any conducting structure will change something to the propagation of high frequency currents. Don't make it too high, a 3 to 4 centimeters height is OK. To try this out, you can just fold in a circle a 1 meter long wrap of aluminum and compress it to a flattened ring that fits your head. If you intent to re-use it, then you must strengthen it with lengths of tape before (inside) and after (outside) the ring is constituted. Make sure the tape does not hamper a perfect electric contact all around the ring. This protection is not perfect but it really helps.
When the field strength is above 0.1 V/m, I need a complementary vertical band in order to be able to work.
This also is a cheap way to make statistics. Compare days with the headband to days without the headband. In order to be scientific, you must ask somebody else to put the headband inside your bonnet every morning, at random. Every morning, the other person writes down if there was a conducting headband inside the bonnet or not. Every evening, you write a comment down to how your day was (headache? brain fog? errors? doomed day?) Then after a while you compare the two. At worst, if you have to do it alone, buy two identical bonnets, push a plastic headband inside one and an aluminum headband inside the other, and spend each day without knowing which one you picked. At the end of the day, write your appreciation down before checking and writing down which bonnet you picked.

Powered by 000webhost

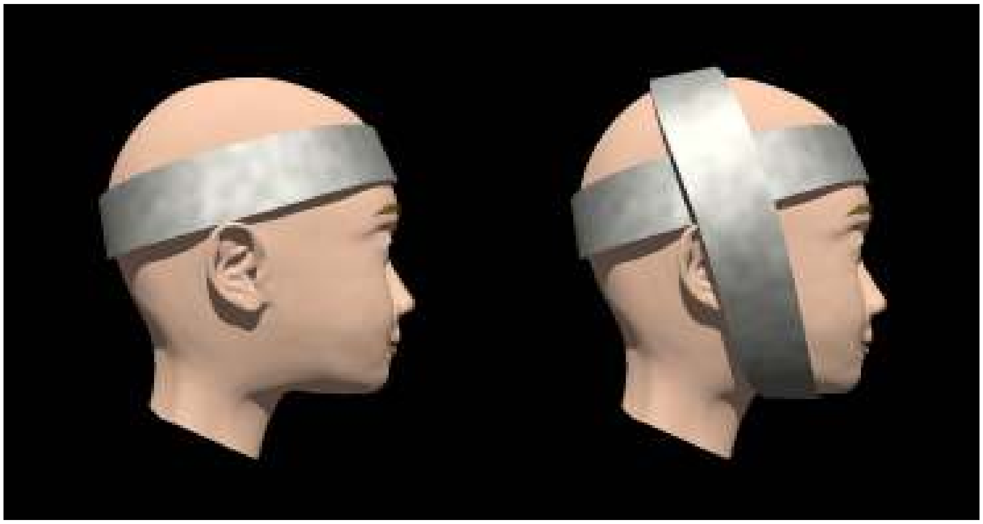

A friend who is a physician points out that in many administrations, be they public or private, the employees are being persecuted. (That often happens when the work done by the administration is mostly useless and only serves to justify the social position and the salaries of a few people.) A statistic about the health impact of radio waves in such administrations could be of few significance, because the people have all kinds of health problems due to the harassment. The addition to this by the radio waves may be comparatively negligible. (Yet another way to build a statistic that shows few or no impact of the radio waves: perform it in a place where people bow under humilations...)

I was quite frightened to discover that the symptoms caused by radio waves, as I see them around me, are very close to those of "radiation poisoning". This is part of nuclear medicine. When a victim is exposed to a short and strong burst of radioactivity or any other ionizing radiation, its body will quickly react like if it had been poisoned. A "low" poisoning by radiations yields headaches, tiredness, difficulties to think... while stronger poisoning can lead to the person die in a few days like if it was heavily burn. The exposure to ionizing radiations is measured in "Gray". 1 Gy leads to a low poisoning. So I computed out how much energy a person gets from the local FM emitter, like the person I know that get sick at their desk. The result is 0.01 Gy for 8 hours of exposition, which is quite close.

Radio waves can induce all the typical problems caused by that ionisinig radiations but the energy levels involved can be quite different. For example, ionizing radiations cause burns without increasing the temperature of the body, while radio waves can only burn if they are powerful enough to heat the body. Radio waves can induce cancer but again the energy required is ways higher, like spending ten years using a cell phone against the head, several hours a day.

Body cells can be attacked by all kinds of means. Ionizing radiations lead to direct damage of the cell membrane. Radio waves disturb the molecular cell gates operate. The cell will always react the same way, as for poisoning: it will close the cell gates, to try to shun itself from the outside world and from the harm. This has many consequences. The cell is less able to fulfill its purpose inside the body. It uses resources to try to protect its internal parts. It will accumulate waste. That's why people get tired, less efficient or even sick. When the aggression made to the cell is deemed too strong, the cell will auto-destruct. That's the way radiations and chemicals can be used against cancer, making the cancer cell suicide.

The cells can be helped to better recover from the state of shock. Also they can be made to go less likely into the state of shock. Medication and health advice can be helpful for this.

A link has been proposed between autism and radio pollution. The growth of the human brain is very complex, with cells traveling from one side of the brain towards the other to find their place. This involves a tremendous amount of communication and coordination between the cells. What if those cells are too often put into a state of shock?

These are two scientific publications about the link between autism and radio waves :

http://www.ncbi.nlm.nih.gov/pubmed/16530334

http://www.ncbi.nlm.nih.gov/pubmed/14962625

Powered by 000webhost

Publications about how radio waves disturb cells :

http://www.ncbi.nlm.nih.gov/pubmed/10860806

http://www.ncbi.nlm.nih.gov/pubmed/12379225

Links

http://www.teslabel.be
http://www.clag.be
http://www.001.be.cx
http://www.criirem.org
http://www.next-up.org
http://www.robindestoits.org
http://www.beperkdestraling.org
http://www.stopumts.nl

A petition: www.petition-electrosmog.be

Eric Brasseur - October 3 2009 till October 16 2015

Powered by 000webhost

rexresearch.com

Roy J. MEYERS

Absorber

How To Patent Your Idea
3 Easy Steps To Patent & Make Money Get A Free Patent & Invention Kit
www.Patent.InventionHome.com

Superior X-Ray Solutions
Performance, Reliability & Speed from the world leader in SCD
www.Bruker-AXS.com

Do You Have an Idea?
We Can Help Turn Your Idea Into A Real Product. Get Free Information!
NewProductHelp.com

Buckyballs Magnetic Toy
Get the Gadget on Everyone's List. Best prices on this year's hot toy!
www.getbuckyballs.com

Car Door Magnets: 50% Off
Magnetic Car Signs $4.99. No Mins. Full-Color. Upload Photos & Logos!
www.Vistaprint.com

Power Your Own Home
Reduce Utility Bills by 50% Lower C02 Emissions by 1/3
www.ClearEdgePower.com

Ads by Google

Renewable Clean Energy
Compact Electricity & Heat Unit For Homes & Small Businesses
www.ClearEdgePower.com

Do You Have an Idea?
We Can Help Turn Your Idea Into A Real Product. Get Free Information!
NewProductHelp.com

Patent Attorney/Engineer
Over 36 years experience. Staff available 24/7 by phone.
www.invention.net

C. Edholm: "Picks Power from Air" (*Technology World Magazine*, 1912 ?)
Dr L. Hirshberg: "Electricity from Air New Great Discovery (*Modern Electrics*, 1914 ?)
Roy J. Meyers: British Patent # 1098 (1913) -- Improvements in and Relating to Apparatus for Producing Electricity (PDF Format, requires Adobe Reader)
R. J. Meyers: British Patent # 1098 (1913) --- Transcription w/ enlarged figures
Notes & Comments
R. J. Meyers: British Patent # 1098 (1913) --- JPG version

***Technology World Magazine*, p. 279-281 (Year unknown, apparently circa 1912; another article about Meyers appeared in the November 1912 issue of *Electronic World*)**

"Picks Power from the Air"

by
Charlton Lawrence Edholm

A remarkable scene took place in the legislature of Arizona this spring when the lawmakers enthusiastically voted for the parole of a certain convict in the State penitentiary, granting him a leave of absence for 30 days and by means of private contributions raising a fund to defray his expenses to Washington DC and return.

The prisoner, Roy J. Meyers, is serving a 3-1/2 year sentence, but in spite of the fact that he bears the stigma of a convicted lawbreaker, he has demonstrated that a convict can be a useful member of society. During his imprisonment he perfected an electrical device of such original character as to arouse feelings of wonder and skepticism until experts had seen it in actual operation. It is a device to draw electricity from the atmosphere for light and power, and the 30-day parole was granted in order that the inventor might protect his rights through the patent office at Washington.

With the acquiescence of the legislature, Governor Hunt granted the parole and the prisoner was allowed to go free without any guard or any assurance but his word of honor that he would return. Two days before the period had elapsed, Meyer again presented himself before the governor, having accomplished his mission, and then returned to the penitentiary at Florence, where he continues to serve his sentence.

This, in brief, is the picturesque story which has called attention of the civilized world to a newly discovered electrical genius, and to another feature of the case which is of equal importance and human interest; namely, the enlightened policy pursued by our youngest State in its treatment of convicts...

Before entering the prison, Meyer had already applied for various patents, among them one for an improved trolley wheel head which prevents the trolley wheel from jumping the wire. Meyers had a conference with Superintendent Sims and Parole Clerk Sanders, and it was to these gentlemen that the inventor first explained the principles of his new device for securing electrical energy from the air. The officials were willing to give the man the opportunity to develop his plan and a little wooden building outside the walls was turned over to Meyers and was fitted up as a workshop and a laboratory. The first demonstration of the new apparatus was made shortly thereafter, the electricity drawn from the atmosphere being used to spark the gas engines of the pump house, and although the device was crude yet it did the work, and removed the doubts of his friends. Furtter development of the "absorber" followed, and his second model was constructed, and developed 8 volts. The machine came to the attention of the remarkable woman who brought his name before the legislature.

This was Miss Kate Barnard, State Commissioner of Charities and Corrections of Oklahoma, who was a guest of Mr. Sims, while studying prison conditions. She saw the machine at work, became familiar with the facts of Meyers' case, and was impressed by his rather blunt and unaffected personality, for Meyers has nothing of the polish or glibness of the poseur. He is a simple, earnest student of mechanical problems and not the sort of man to make a sentimental appeal for sympathy because of any grace of person or manner. Therefore it was the value of Meyers' invention, together with his essential integrity (in spite of his lapse) which so strongly impressed Miss Barnard that when she appeared before the Arizona legislature not long afterwards, addressing that body on the need of enlightened legislature along the line of her own work, she told the story of Roy Meyers and his epoch-making invention.

So, early in May, Meyers set out for Washington, unaccompanied.

In his own words: "When I arrived in Washington and laid my plans before the patent office experts, they merely smiled and told me that I would have to build a model and demonstrate my claims --- that it seemed strange that I, unknown as I am in the electrical world, should have accomplished the things for which Edison, Tesla and other experts have been striving for years.

"They could grasp the meaning of my drawings nor the explanation I tried to make to them. There was little time to spare, as I had only 20 days left of my leave, but I set to work in a few days was able to take a crude model around to the patent office to make a demonstration.

"Arriving at the patent office I telephoned to a friend who had been so kind as to introduce me and aid me in reaching the proper officials. The absorber was hoisted on two short poles and made to work. While they were as yet unable to understand the principles involved and hardly willing to believe their eyes, they were forced to admit that I had something new and different, and they told me that there would be no further objection; that I might file my application without further delay.

"I hope to construct my first large machine right here in Phoenix. I feel grateful; to Governor Hunt and others for what they have done for me and to the help they have given in securing protection I might not otherwise have had, and I am desirous of demonstrating this gratitude. I am going back to Florence today to resume the serving of my sentence, which will expire in 10 months. Then, here in Phoenix, I will begin the work of making my machines."

While there are some details of the device which the inventor refuses to make public, yet there are many general features that may be explained. It is planned that the machine, to be set up in Phoenix, will generate sufficient power to light the city, and will consist of a 200 foot tower upon which is placed the "absorber". The latter consists of a series of magnetized steel plates set in a circle (the manner of preparing them is kept secret) and this mechanism attracts the electricity from the atmosphere. This is carried by wires to a transformer in the engine house below and thence is applied to produce either power or light after the usual manner.

In an authorized statement Meyers says: "The flow of electricity is constant. When it emerges into the transformer it is in the form of a direct current. It will absorb the electricity day and night and will work whenever the wireless will work. I can put up a plant to supply such a building as the Adams Hotel for about $1500, and one of the principal items of the expense is the cost of the towers, the wires, the magnetizing of one set of plates, which is part of the secret of the treatment which makes it respond to the accumulations of the atmosphere.

"For use in the case of an electrical storm I have made what I call a modified form of circuit breaker, such as is commonly used as a lightning arrester on telegraph lines. In case of a storm the accumulator would suddenly become overcharged, possibly, and as the electricity would not of itself flow back into the air, the result might be disastrous. So I send it down into the ground, whenever the voltage rises above a certain amount."

Roy J. Meyers & the "Absorber"

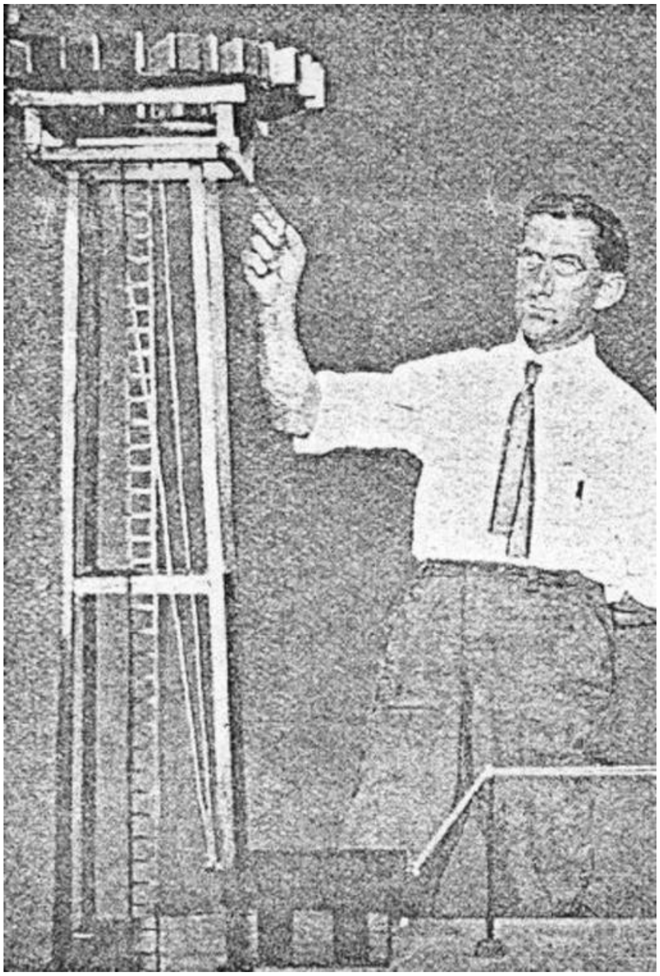

Modern Electronics (1914 ?)

Electricity From Air New Great Discovery

by

Dr Leonard Keene Hirshberg

Working quietly in the heart of Baltimore for weeks on an invention which some critics say will revolutionize the method of converting electricity to practical use has been Roy J. Meyers, who like Benjamin Franklin, extracts the electric current from the air.

Mr Meyers invention was made last summer while he was confined in the penitentiary at Florence, Arizona. His first finished apparatus was made in Baltimore.

A practical, unlettered electrician, Mr Meyers, while in Arizona, was arrested on a comparatively minor charge and sent to the penitentiary. There he was placed in charge of the prison electrical plant, and there he says he made his discovery that the current which the civilized world is beginning to use most extensively for light and power could be transformed from the atmosphere without the aid of moving machines or batteries.

Miss Kate Barnard, Commissioner of Charities and Corrections, of Oklahoma, hearing of Meyers' invention and of his desire to have it patented, appeared before the Arizona Legislature to make an appeal in behalf of the young convict. As a result a special bill was passed which granted Meyers a month's leave of absence on parole. He went unaccompanied to Washington, filed his patent applications and returned to the penitentiary. Since then he has been indefinitely paroled.

He came to Baltimore as the place where he could easily obtain the mechanical parts needed to make a more nearly perfect machine than the crude model he has fashioned in the penitentiary workshop, and is making his headquarters here while working on his invention. With him is W.E. Chenot, who has been his assistant in assembling and testing the machine and who says that he has bought Meyers' patent rights for Germany.

They have proved beyond doubt that the invention is practical and that when finally brought to a state of perfection it will introduce a new epoch in the industrial use of electricity. By Westinghouse meters they tested the strength of the current gathered from the air, and with the use of only two of the four rectifying transformers the voltmeter recorded four and one-half volts, and the ammeter, which had the capacity of recording 75 amperes, was broken by the force of the current.

The machine itself is simple. It is in reality a transformer, which is familiar to anyone knowing anything at all about electricity in its practical uses. On a high tripod, which resembles somewhat the framework of a windmill tower, is the transformer, which Mr Meyers calls his 'absorber'. It is made up of an iron core, wrapped with copper wire. The secret of the invention is the manner in which the disks composing this 'absorber' are magnetized, and this secret Meyers says he found by accident while at work in prison.

What the machine, when finally perfected, will do is yet to be seen. Its inventor claims that it will greatly reduce the cost of making electricity. No batteries of any kind are needed, he says, and not a part of the machine turns upon the other. It is as durable, apparently, as an electric light pole. One of these machines, says Meyers, when perfected may be placed on a vehicle and transform enough electricity to give motive power, be that vehicle a locomotive or an automobile. He declared it can be placed on a building to furnish electric lights or power, and that the only wear will be upon the machinery which its current runs.

Meyers is 34 years old and he gained his knowledge of electricity by working in shops along the Pacific Coast. The depths of the mysteries of electricity he has not explored, but he is certain that he has found the means of absorbing it from the air and of converting it to the use of mankind.

British (GB) Patent # 191301098

Improvements in and Relating to Apparatus for Producing Electricity.

1-14-1914

Roy Jerome Meyers

Classification: - international: H05F7/00; H05F7/00; - european: H05F7/00
Application number: GBD191301098 19130114
Priority number(s): GBT191301098 19130114

Abstract ~ Vapour apparatus, arrangements of. - A rectifier for use with apparatus for producing electricity from the earth consists of mercury- vapour lamps constructed and arranged as shown in Fig. 4. Each lamp comprises two wires 6<1>, 7<1> wound around a steel tube 15 surrounding a mercury tube 11 preferably of copper. The coil 6<1> is connected between the electrode 14 and the terminal 18, and the coil 7<1> between the terminals 19, 5. The coils 6<1>, 7<1> are preferably composed of soft iron. Reference has been directed by the Comp- troller to Specifications 16,709/87, 14,033/99, and 5457/11, [all in Class 53, Galvanic batteries], and 15,412/06.

British Patent # 1098

(January 14, 1913)

Improvements in and Relating to Apparatus for Producing Electricity

Roy Jerome Meyers

This invention relates to improvements in apparatus for the production of electrical currents, and the primary object in view is the production of a commercially serviceable electrical current without the employment of mechanical or chemical action. To this end the invention comprises means for producing what I believe to be dynamic electricity from the earth and its ambient elements.

I am, of course aware that it has been proposed to obtain static charges from upper strata of the atmosphere, but such charges are recognized as of widely variant potential and have thus far proved of no practical commercial value, and the present invention is distinguished from all such apparatus as has heretofore been employed for attracting static charges by the fact that this improved apparatus is not designed or employed to produce or generate irregular, fluctuating or other electrical charges which lack constancy, but on the other hand I have by actual test been able to produce from a very small apparatus at comparatively low elevation, say about 50 or 60 feet above the earth's surface, a substantially constant current at a commercially usable voltage and amperage. This current I ascertained by repeated tests is capable of being readily increased by additions of the unit elements in the apparatus hereinafter set forth, and I am convinced from the constancy of the current obtained and its comparatively low potential that the current is dynamic and not static, although, of course, it is not impossible that certain static discharges occur and, in fact, I have found occasion to provide against the damage which might result from such discharge by the provision of lightning arresters and cut-out apparatus which assist in rendering the obtained current stable by eliminating sudden fluctuations which sometimes occur during conditions of high humidity from what I consider static discharges. The nature of my invention is obviously such that I have been unable to establish authoritatively all of the principles involved, and some of the theories herein expressed may possibly prove erroneous, but I do know and am able to demonstrate that the apparatus which I have discovered does produce, generate, or otherwise acquire a difference of potential representing a current amperage above stated, or varied therefrom at the will of the operator according to the uses which the current is to be subjected.

The invention comprises generically means for producing electrical currents of serviceable potential substantially without the employment of mechanical or chemical action, and in this connection I have been able to observe no chemical action whatever on the parts utilized although deterioration may possibly occur in some of the parts, but so far as I am able to determine such deterioration does not add to the current supply but is merely incidental to the effect of climatic action.

The invention more specifically comprises the employment of a magnet or magnets and a co-operating element, such as zinc disposed adjacent to the magnet or magnets and connected in such manner and arranged relative to the earth so as to produce current, my observation being that current is produced only when such magnets have their poles facing substantially to the north and south and the zincs are disposed substantially along the magnets.

The invention also comprehends other details of construction, combinations and arrangements of parts as will hereinafter be fully set forth and claimed.

In the accompanying drawings:

Figure 1 is a top plan view of an apparatus embodying the features of the present invention, the arrow accompanying the figure indicating substantially the geographical north, parts of the figure being diagrammatic for condensing the showing.

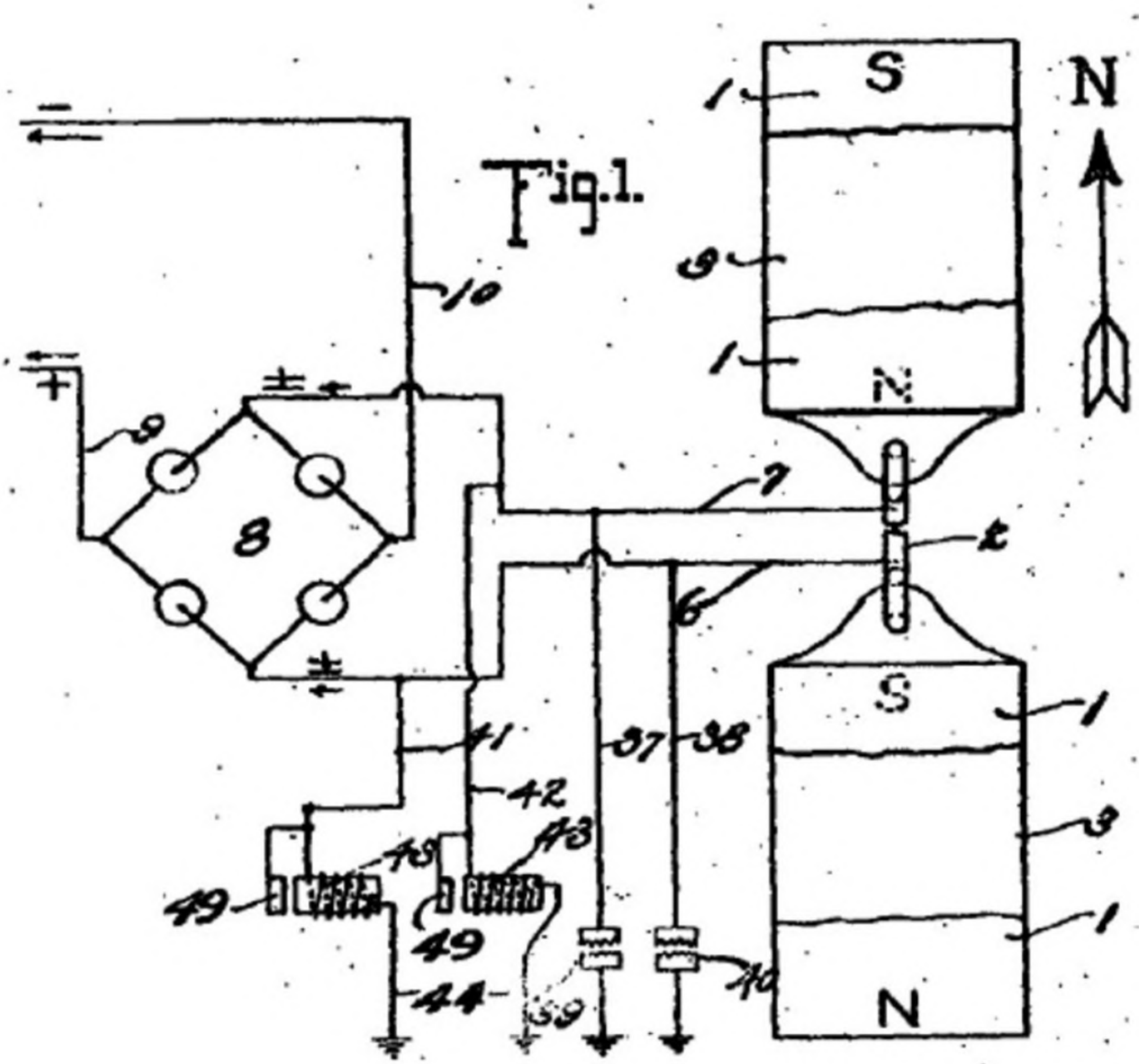

Figure 2 is a view is side elevation of the parts seen in plan in Figure 1.

Fig. 2.

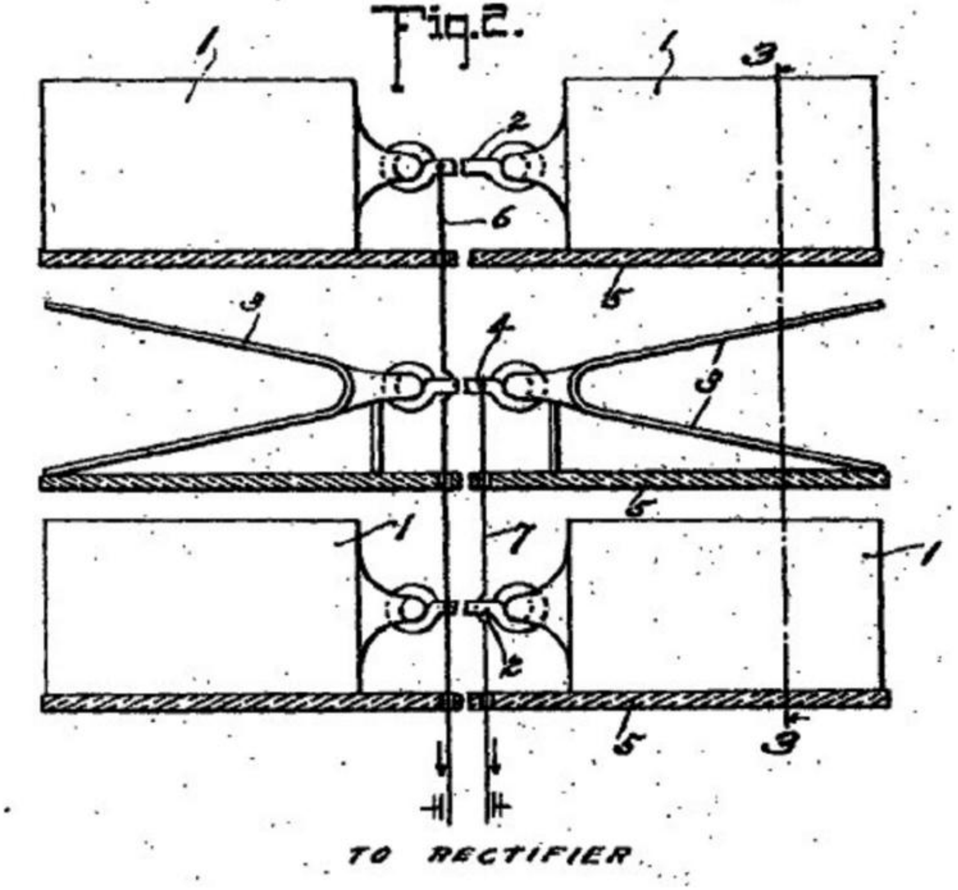
1
1
3
2
6
5
9
4
9
6
1
7
1
2
5
3
TO RECTIFIER

Figure 3 is a vertical section taken on the plane indicated by the line 3-3 of Figure 2 and looking in the direction indicated by the arrow.

Fig. 5.

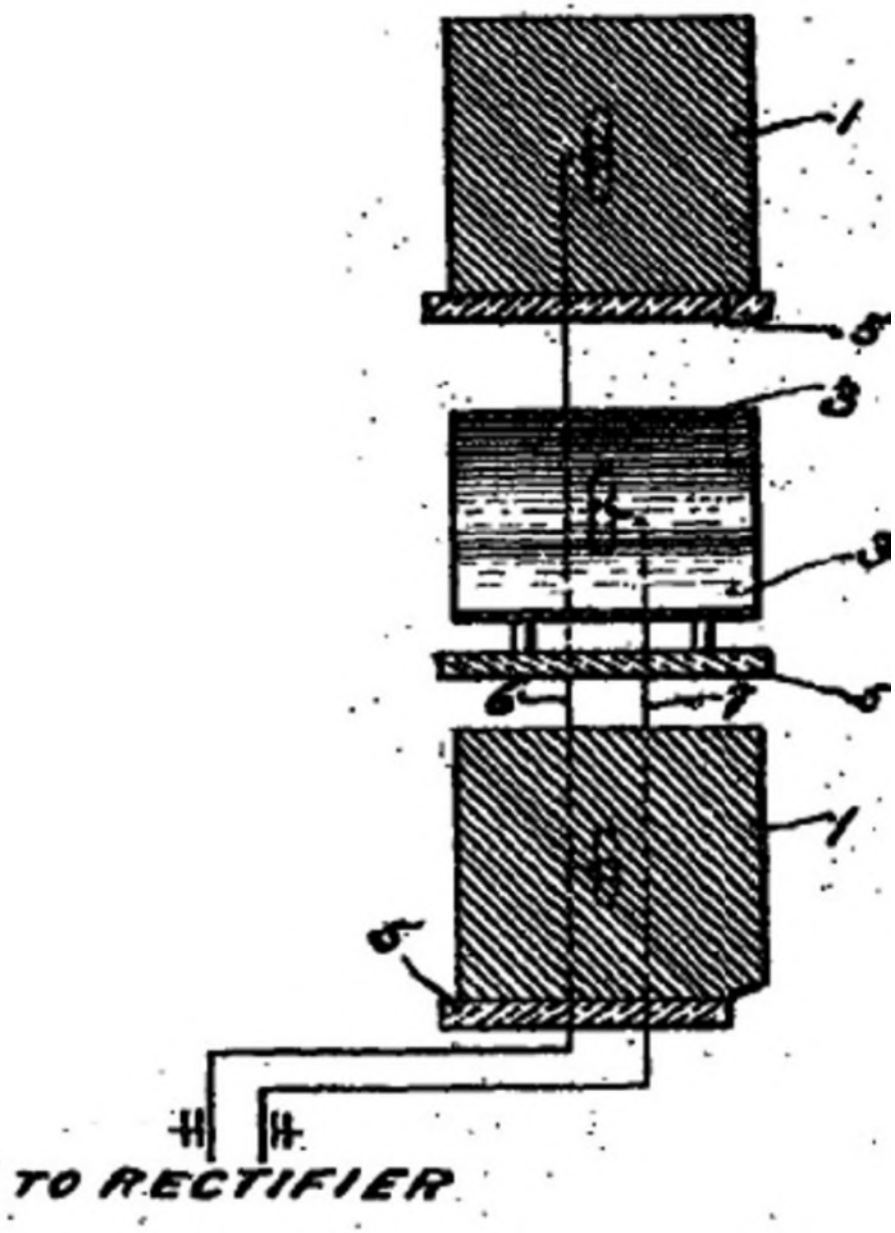
1
4
5
3
6
3
5
6
7
1
4
5
TO RECTIFIER

Figure 4 is a detail view partly in elevation and partly in section showing the detail connections of the converter and intensifier.

Fig 4

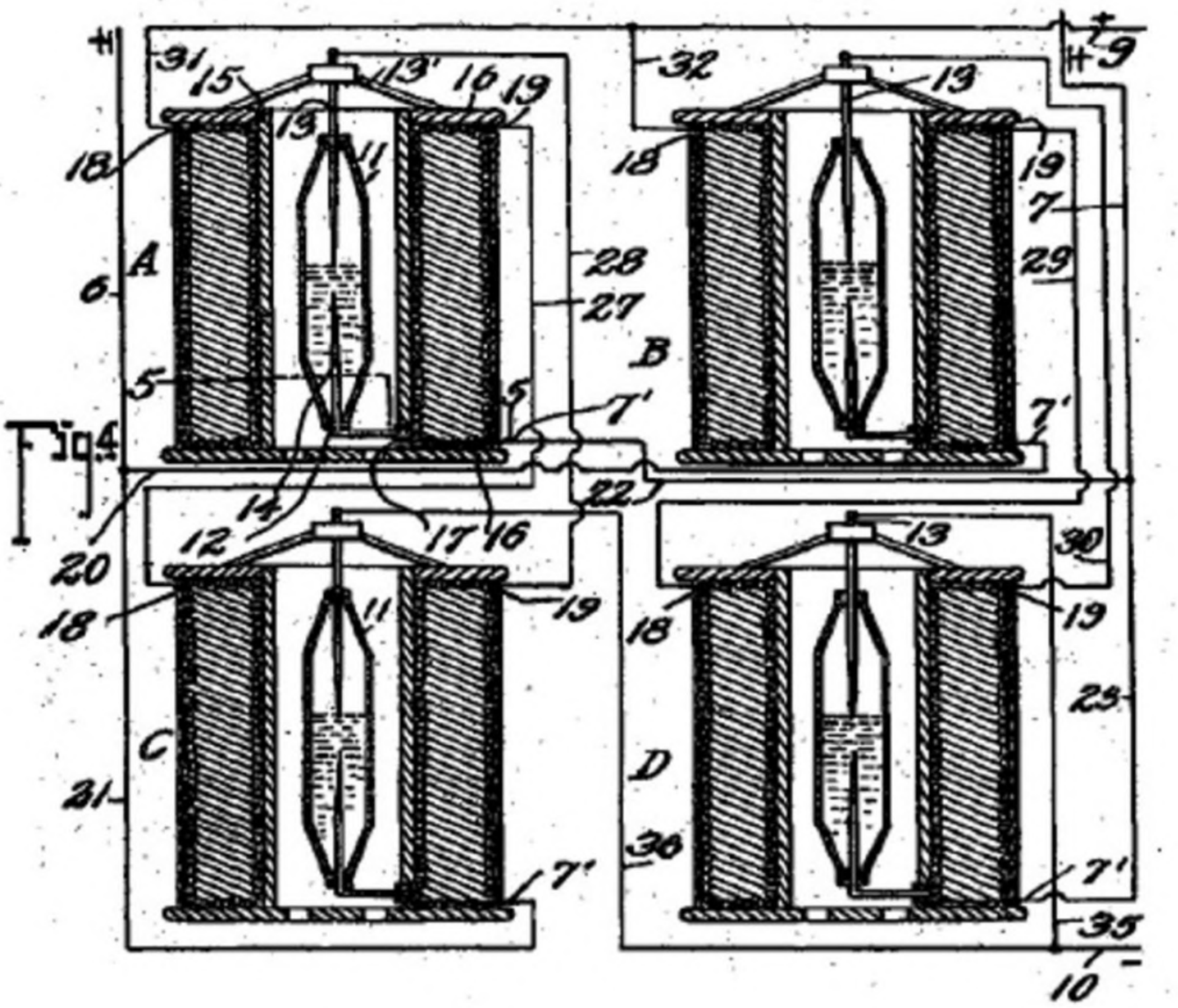
+
31
15
13'
16
19
32
13
+
9
18
13
11
18
19
7
A
28
29
6
27
B
5
5
7'
7'
20
14
17
16
22
12
13
30
18
11
19
18
19
C
23
21
D
36
7'
7'
35
10
-

Figure 5 is a transverse section taken on the planes indicated by line 5-5 of Figure 4 and looking downwardly.

Fig 5.

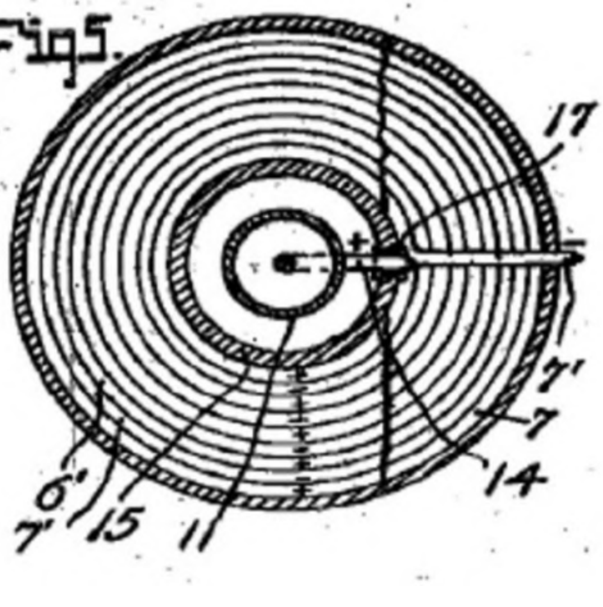
17
+
-
7'
7
14
6'
7'
15
11

Figure 6 is an enlarged detail fragmentary section illustrating the parts at the juncture of the conductors and one of the intensifiers.

Fig.6

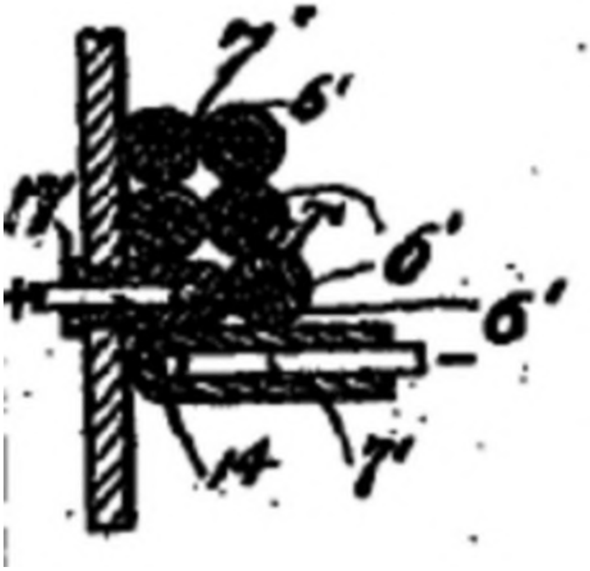

Figure 7 is an enlarged detail view partly in elevation and partly in section of one of the automatic cut-outs and

Fig.7.

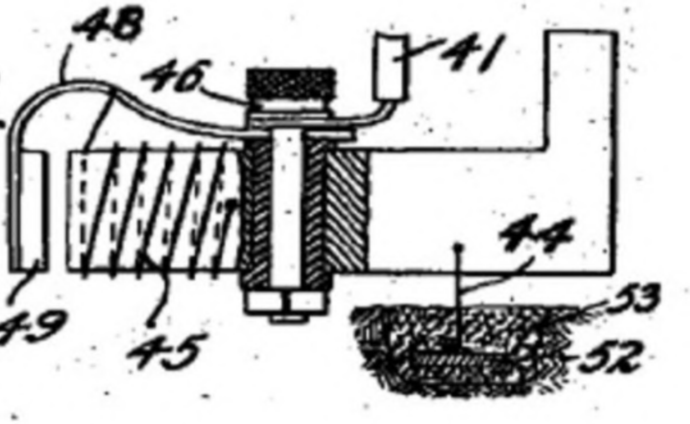

Figure 8 is a diagrammatic view of one of the simplest forms of embodiment of the invention.

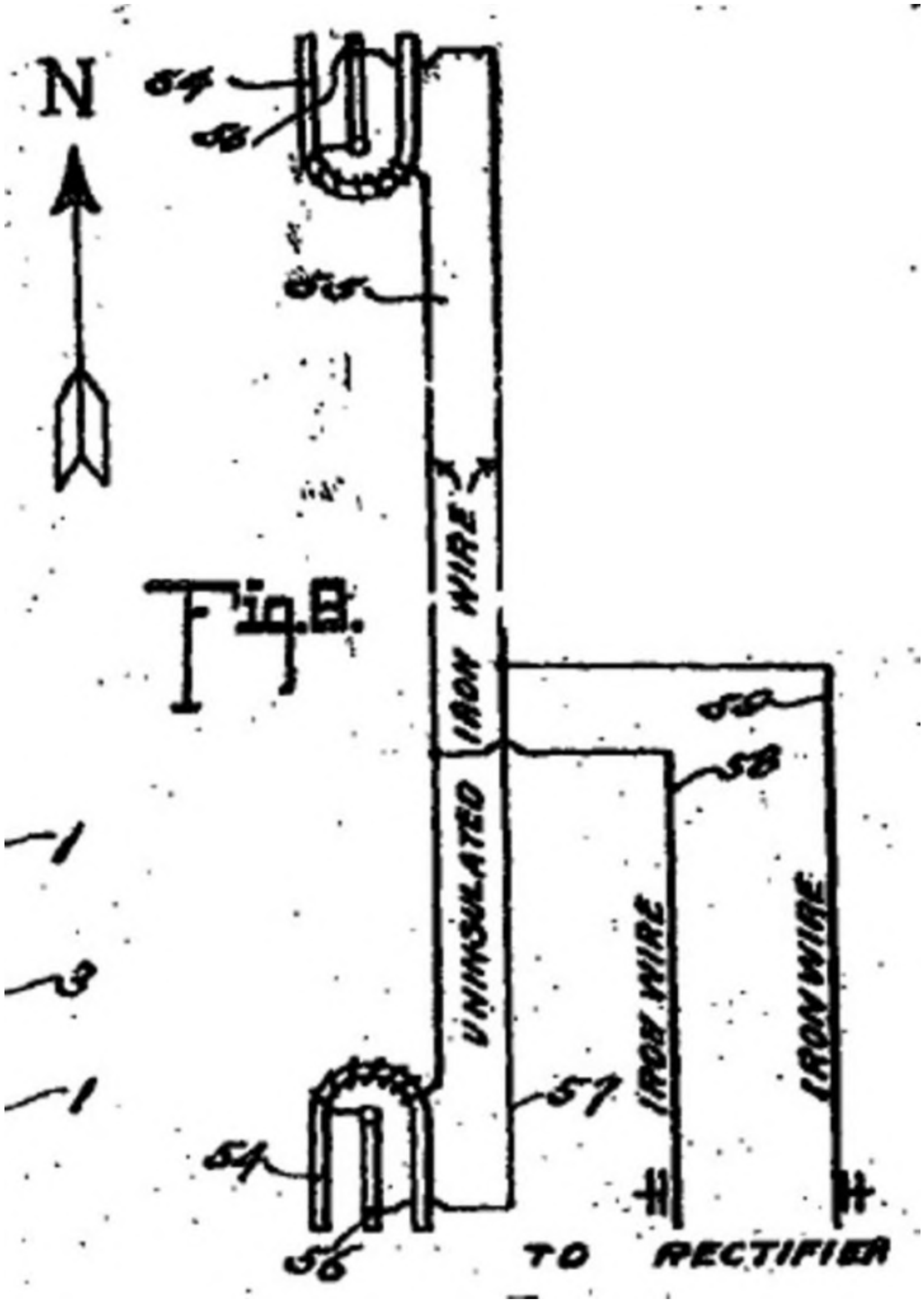

Fig.8.

Referring to the drawing by numerals, 1,1 indicates magnets connected by a magnetic substance 2, preferably an iron wire. The magnets 1 are arranged in pairs, one pair being spaced beneath the other, and interposed between the magnets are zinc plates 3,3 connected by an iron wire conductor 4. Suitable insulating supports 5 are arranged for sustaining the respective magnets 1 and plates 3,3. Each plate 3 is preferably bent substantially into V form, as clearly seen in Figure 1, and the V1 of one of the plates opens or faces toward the north and the V of the other plate to the South. I have determined by experimentation that it is essential that the plates 3 be disposed substantially north and south with their flat faces approximately parallel to the adjacent faces of the co-operating magnets, although by experience I have not discovered any material difference in the current obtained when the plates are disposed slightly to one side of north and south, as for instance when the plates are disposed slightly to one side of north and south, as for instance when disposed in the line of the magnetic polarity of the earth. The same is true with respect to the magnets 1, the said magnets being disposed substantially north and south for operative purposes, although I find that it is immaterial whether the north pole of one of the magnets is disposed to the north and the south pole to the south, or vice versa, and it is my conviction from experience that it is essential to have the magnets of each pair connected by magnetic material so that the magnets substantially become one with a pole exposed to the north and a pole exposed to the north. In Figure 1, I have indicated in full lines by the letters 8 and N the respective polarities of the magnets 1, and have indicated in dotted lines the other pole of those magnets when the connection 2 is severed. I have found that the magnets and zinc plates operate to produce, whether by collection or generation I am not certain, electrical currents when disposed substantially north and south, but when disposed substantially east and west no such currents are produced. I also find that the question of elevation is by no means vital, but it is true that more efficient results are obtained by placing the zincs and magnets on elevated supports. I furthermore find from tests that it is possible to obtain currents from the apparatus with the zincs and magnets disposed in a building or otherwise enclosed, although more efficient results are obtained by having the said elements arranged in the open.

While in Figures 1, 2, and 3, I have shown the magnets and the zinc plates as superimposed, it will be apparent, as hereinafter fully set forth, that these elements may be juxtaposed in horizontal planes, and substantially the same results will be secured. Furthermore, the magnets 1 with the interposed zincs 3, as shown in Figures 1, 2 and 3 merely represent a unit which may be repeated either horizontally or vertically for increasing the current supply, and when the unit is repeated the zinc plates are arranged alternating with the magnets throughout the entire series as hereinafter indicated.

A conductor 6 is connected in multiple with the conductors 2 and a conductor 7 is connected with conductor 4, the conductor 6 extending to one terminal of a rectifier which I have indicated by the general reference character 8, and the conductor 7 extending to the other terminal of said rectifier. The rectifier as seen in diagram in Figure 1 may assume any of several well known embodiments of the electrical valve type and may consist of four asymmetric cells or Cooper-Hewitt mercury vapor lamps connected as indicated in Figure 1 for permitting communication of the positive impulses from the conductor 6 only to the line conductor 9 and the negative impulses from conductor 6 on only to the line conductor 10. The current from this rectifier may be delivered through the conductors 9 and 10 to any suitable source for consumption.

While the said rectifier 8 may consist of any of the known types, as above outlined, it preferably consists of a specially constructed rectifier which also has the capacity of intensifying the current and comprises specifically the elements shown in detail in Figures 4, 5, and 6 wherein I have disclosed the detail wiring of the rectifier when composed of four of the rectifying and intensify in elements instead of asymmetric cells or simple mercury vapor valves. As each of these structures is an exact embodiment of all the others, one only will be described, and the description will apply to all. The rectifying element of each construction consists of a mercury tube 11 which is preferably formed of glass or other suitable material, and comprises a cylinder having its end portions tapered and each terminating in an insulating plug or stopper 12. Through the upper stopper 12 is extended the electrode 13 which extends well into the tube and preferably substantially one-half the length thereof to a point adjacent the inner end of an opposing electrode 14 which latter electrode extends thence downwardly through the insulation 12 at the lower end of the tube. The tube 11 is supplied with mercury and is adapted to operate on the principle of the mercury vapor lamp, serving to rectify current by checking back impulses of one sign and permitting passage of impulses of the other. To avoid the necessity for utilizing a starter, as is common with the lamp type of electrical valve, the supply of mercury within the tube may be sufficient to contact with the lower end of the electrode 13 when current is not being supplied, so that as soon as current is passed from one electrode to the other sufficiently for volatilizing that portion of the mercury immediately adjacent the lower end of electrode 13, the structure begins its operation as a rectifier. The tube 11 is surrounded by a tube 15 which is preferably spaced from tube 11 sufficiently for allowing atmospheric or other cooling circulation to pass the tube 11. In some instances, it may be desirable to cool the tube 11 by a surrounding body of liquid, as hereinafter indicated. The tube 15 may be of insulating material but I find efficient results attained by the employment of a steel tube, and fixed to the ends of the of the tube are insulating disks 16, 16 forming a spool on which are wound twin wires 6^1 and 7^1, the wire 6^1 being connected at the inner helix of the coil with the outer end of the electrode 14, the lower portion of said electrode being extended to one side of the tube 11 and passed through an insulating sleeve 17 extending through the tube 15, and at its outer end merging into the adjacent end of the wire 6^1. The wire 7^1 extends directly from the outer portion of the spool through the several helices to a point adjacent the juncture of the electrode 14 with wire 6^1 and thence extends in mechanical parallelism with the wire throughout the coil, the wire 6^1 ending in a terminal 18 and the wire 7^1 ending in a terminal 19. For the sake of convenience of description and of tracing the circuits, each of the apparatus just above described and herein known as an intensifier and rectifier will be mentioned as A, B, C and D, respectively. Conductor 6 is formed with branches 20 and 21 and conductor 7 is formed with similar branches 22 and 23. Branch 20 from conductor 6 connects with conductor 7^1 of intensifier B and branch 21 of conductor 6 connects with the conductor 7^1 of intensifier C, while branch 22 of conductor 7 of intensifier C, while branch 22 of conductor 7 connects with conductor 7^1 of intensifier D. A conductor 27 is connected with terminal 19 of intensifier A and extends to and is connected with the terminal 18 of intensifier C, and a conductor 7 connects with conductor 7^1 of intensifier D. A conductor 27 is connected with terminal 19 of intensifier A. and extends to and is connected with terminal 18 of intensifier C, and a conductor 28 is connected with the terminal 19 of intensifier C and extends from the terminal 19 of intensifier B to the terminal 18 of intensifier D to electrode 13 of intensifier B. Each electrode 13 is supported on a spider 13^1 resting on the upper disk 16 of the respective intensifier. Conductors 31 and 32 are connected with the terminals 18 of intensifiers A and B and are united to form the positive line wire 9 which co-operates with the negative line wire 10 and extends to any suitable point of consumption. The line wire 10 is provided with branches 35 and 36 extending to the electrodes 13 of intensifiers C and D for completing the negative side of the circuit.

Thus it will be seen that alternating currents produced in the wires 6 and 7 will be rectified and delivered in the form of a direct current through the line wires 9 and 10, and I find by experiment that the wires 6 and 7 should be of iron, preferably soft, and may of course be insulated, the other wiring not specified as iron being of copper or other suitable material.

In carrying out the operation as stated, the circuits may be traced as follows: A positive impulse starting at the zincs 3 is directed along conductor 7 to branch 23 to conductor 7^1 and the winding of the rectifier of intensifier B through said rectifier to the conductor 6^1, through the winding thereof to the contact 18, conductor 32 and to the line wire 9. The next or negative impulse directed along conductor 7 cannot find its way along branch 23 and the circuit just above traced because it cannot pass across the rectifier of intensifier B but instead the negative impulse passes along conductor 22 to conductor 7 of intensifier A and the winding thereof to the contact 19 and to conductor 27 to contact 18 of intensifier C, to the winding of the wire 6^1 thereof to the electrode 14 through the rectifier to the of the electrode 13 and conductor of intensifier A, electrode 14 thereof and conductor 6^1 to contact 18 and wire 31 to line wire 9. Obviously the positive impulse cannot pass along the wire 20 because of its inverse approach to the rectifier of intensifier B. The next impulse or negative impulse delivered to conductor 6 cannot pass along conductor 21 because of its connection with electrode 13 of the rectifier of intensifier A, but instead passes along conductor 20 to the wire 7^1 and its winding forming part of intensifier B to the contact 19 and conductor 29 to contact 18 and the winding of wire 6^1 of intensifier D to the electrode 14 and through the rectifier to the electrode 13 and conductor 35 to line wire 10. Thus the current is rectified and all positive impulses directed along one line and all negative impulses along the other lie s that the potential difference between the two lines will be maximum for the given current of the alternating circuit. It is, of course, apparent that a less number of intensifiers with their accompanying rectifier elements may be employed with a sacrifice of the impulses which are checked back from a lack of ability to pass the respective rectifier elements, and in fact I have secured efficient results by the use of a single intensifier with its rectifier elements, as hereinafter set forth.

Grounding conductors 37 and 38 are connected respectively with the conductors 6 and 7 and are provided with the ordinary lightning arresters 39 and 40 respectively for protecting the circuit against high tension static charges.

Conductors 41 and 42 are connected respectively with the conductors 6 and 7 and each connects with an automatic cutout 43 which is grounded as at 4. Each of said automatic cutouts is exactly like the other and one of the same is shown in detail in Figure 7 and comprises the inductive resistance 45 provided with an insulated binding post 46 wit which the respective conductor 6 or 7 is connected, said post also supporting a spring 48 which sustains an armature 49 adjacent to the core of the resistance 45. The helix of resistance 45 is connected preferably through the spring to the binding post at one end and at the other end is grounded on the core of the resistance, the said core being grounded by ground conductor 44 which extends to the metallic plate 52 embedded in moist carbon or other inductive material buried in the earth. Each of the conductors 41, 42 and 44 is of iron, and in this connection I wish it understood that where I state the specific substance I am able to verify the accuracy of the statement by the results of tests which I have made, but of course I wish to include along with such substances al equivalents, as for instance, where iron is mentioned its byproducts, such as steel, and its equivalents such as nickel and other magnetic substances are intended to be comprehended. The cutout apparatus seen in detail in Figure 7 is employed particularly for insuring against high tension currents, it being obvious from the structure shown that when potential rises beyond the limit established by the tension of the spring sustaining the armature 40, the armature will be moved to a position contacting with the core of the cutout device and thereby directly close the ground connection for line wire 41 with conductor 44, eliminating the resistance of winding 45 and allowing the high tension current to be discharged to the ground. Immediately upon such discharge the winding 45 losing its current will allow the core to become demagnetized and release the armature 49 whereby the ground connection is substantially broken leaving only the connection through the winding 45 the resistance of which is sufficient for insuring against loss of low tension current.

In Figure 8 I have illustrated an apparatus which though apparently primitive in construction and arrangement comprehends the first successful embodiment which I produced in the course of discovery of the present invention, and it will be observed that the essential features of the invention are therein disclosed. The structure delineated in said figure consists of horseshoe magnets 54, 55, one facing north and the other south, that is, each opening in the respective directions indicated and the two being connected by an iron wire 55 which is uninsulated and wrapped about the respective magnets each end portion of the wire 55 being extended from the respective magnets to and connected with, as by being soldered to, a zinc plate 56, there being a plate 56 for each magnet and each plate being arranged longitudinally substantially parallel with the legs of the magnet and with the faces of the plate exposed toward the respective legs of the magnet, the plate being thus arranged endwise toward the north and south. An iron wire 57 connects the plates 56, the ends of the wire being preferably connected adjacent the outer ends of the plates but from experiment I find that the wire may be connected at practically any point to the plate. Lead wires 58 and 59 are connected respectively with the wires 55 and 57 and supply an alternating current at a comparatively low tension, and to control such current the wires 58 and 59 may be extended to a rectifier or combined rectifier and intensifier, as above set forth.

The tests which I have found successful with the apparatus seen in Figure 8 were carried out by the employment first of horseshoe magnets approximately 4 inches in length, the bar comprising the horseshoe being about one inch square, the zincs being dimensioned proportionately and from this apparatus with the employment of a single intensifier and rectifier, as above stated, I was able to obtain a constant current of 8 volts.

It should be obvious that the magnets forming one of the electrodes of this apparatus may be permanent or may be electromagnets, or a combination of the two.

While the magnets mentioned throughout the above may be formed of any magnetic substance, I find the best results obtained by the employment of the nickel chrome steel.

While the successful operation of the various devices which I have constructed embodying the present invention have not enabled me to arrive definitely and positively at fixed conclusion relative to the principles and theories of operation and the source from which current is supplied, I wish it to be understood that I consider myself as the first inventor of the general type hereinbefore described capable of producing commercially serviceable electricity, for which reason my claims hereinafter appended contemplate that I may utilize a wide range of equivalents so far as concerns details of construction suggested as preferably employed.

The current which I am able to obtain is dynamic in the sense that it is not static and its production is accomplished without chemical or mechanical action either incident to the actual chemical or mechanical motion or incident to changing caloric conditions so that the elimination of necessity for the use of chemical or mechanical action is to be considered as including the elimination of the necessity for the use of heat or varying degrees thereof.

Having now particularly described and ascertained the nature of my said invention, and in what manner the same is to be performed, I declare that what I claim is: --- [Claims not included here]

NOTES & COMMENTS

From the Article in *Tech. World Mag*.:

1. First demo model was powerful enough to spark a gas engine.
2. Second model developed 8 volts.
3. Demo model at Patent Office was elevated on short poles.
4. The model planned to power Phoenix AZ would be elevated 200 feet.
5. The Absorber "consists of a series of magnetized steel plates set in a circle (the manner of preparing them is kept secret)".
6. "[T]he magnetizing of one set of plates... is part of the secret of the treatment which makes it respond to the accumulations of the atmosphere".

From British Patent # 1098 (1913):

"I have been able... to produce from a very small apparatus at comparatively low elevation, say about 50 or 60 feet above the earth's surface, a substantially constant current at a commercially useable voltage and amperage".

"This current... is capable of being readily increased by additions of the unit elements in the apparatus".

Fig. 1 and Fig. 2 show the magnet poles are connected N-S by a thick iron rod (thick compared to the lines used for wires in the drawings).

No angle is specified for the V-shaped zinc plates. The article (but not the patent) states that the plates are magnetized (obviously not zinc). Zinc-galvanized steel? Will a thin film of Zn work? Or, powdered Zn in a binder (more surface area)? Or, zinc-galvanized iron wire in a coil?

The Palmer Craig device (www.rexresearch.com/craig/craig.htm) is powered by the terrestrial magnetic field, and employs a thin film of bismuth to capture the energy as diamagnetism. Perhaps this can be integrated with Meyers' device.

Figure 8 (the demonstration of principle) show uninsulated iron wire being used to connect the plates and magnets. The wire is wound around the Bloch wall area of the horseshoe magnets. Perhaps Coler-type windings around the poles could be used here (See: www.rexresearch.com/coler/coler.htm). Coler used copper plates as "condensers" in his device. Could copper plates be used for the Meyers device? Perhaps flat (Tesla non-inductive) coils could be integrated here.

"It is essential that the plates 3 be disposed substantially N and S with their flat faces approximately parallel to the adjacent faces of the co-operating magnets....

"I find that it is immaterial whether the N pole of one of the magnets is disposed to the N and the S pole to the S, or vice versa".

"[T]he magnets and zinc plates... produce electrical currents when disposed... N and S, but when disposed... E and W no such currents are produced".

"[E]levation is by no means vital, but... more efficient results are obtained by placing the zincs and magnets on elevated supports".

"The elements may be disposed in horizontal planes [or vertically]...".

The "zinc plate 56... [is] arranged longitudinally substantially parallel with the legs of the magnet and with the faces of the plate exposed toward the respective legs of the magnet, the plate being thus arranged endwise toward the north and south".

The first model used "horseshoe magnets approximately 4 inches in length, the bar comprising the horseshoe being about one inch square, the zincs being dimensioned proportionately and from this apparatus with the employment of a single intensifier and rectifier, as above stated, I was able to obtain a constant current of 8 volts... [T]he magnets... may be permanent or may be electromagnets, or a combination of the two... I find the best results obtained by... nickel chrome steel".

Comments & Questions:

The rectifier is described as a preferred embodiment, but other designs also work. The Ed Gray capacitor design comes to mind (www.rexresearch.com/evgray/1gray.htm). The Tate Ambient Power Module also might apply (www.rexresearch/tate/tate.htm).

Would non-ferrous magnets work? Is there a frequency involved (oscilloscope tests)? Coler found that ferromagnetism has a resonant frequency about 180 KHz. Can the components be made adjustable for RLC-resonance?

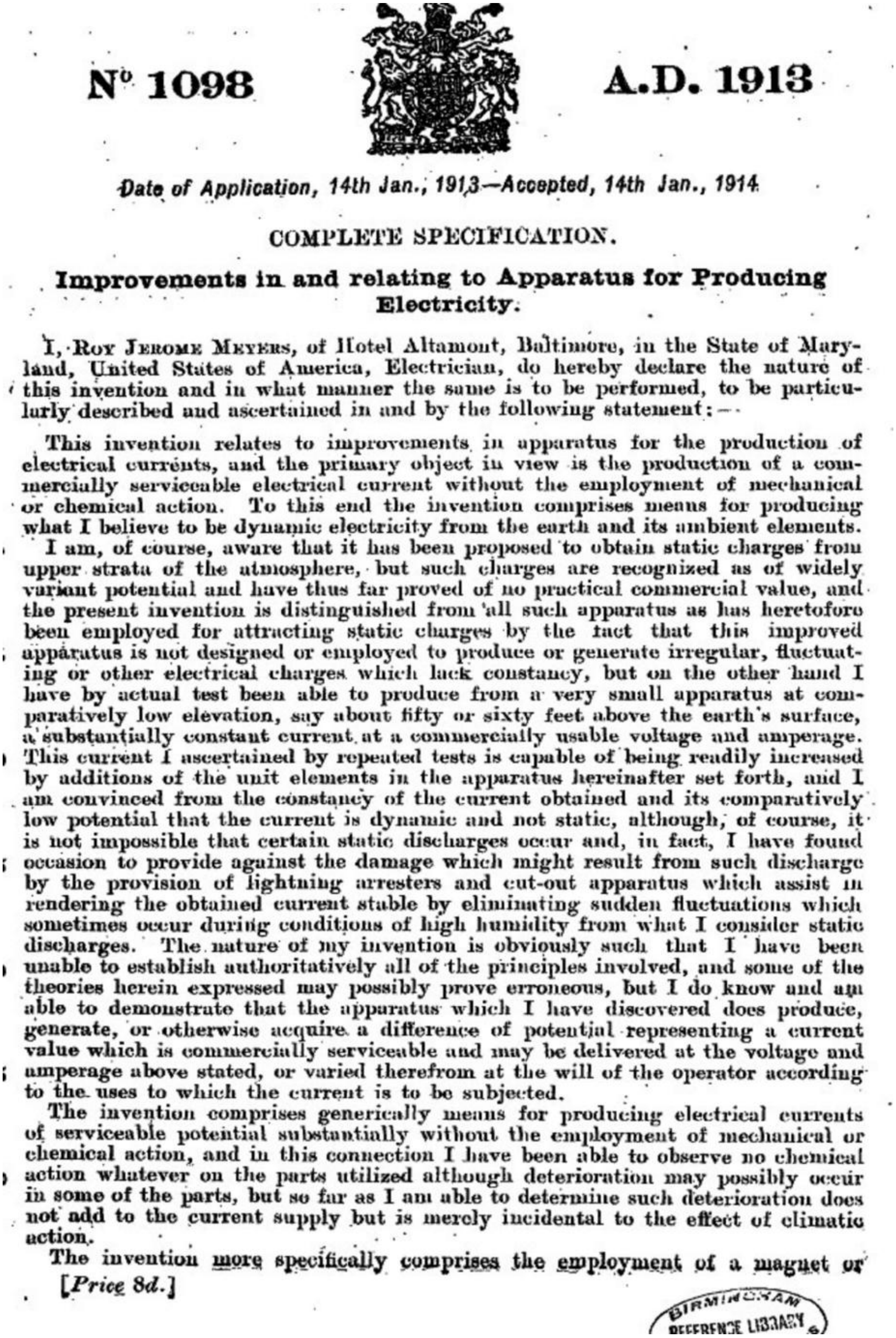

Nº 1098 **A.D. 1913**

Date of Application, 14th Jan., 1913—Accepted, 14th Jan., 1914

COMPLETE SPECIFICATION.

Improvements in and relating to Apparatus for Producing Electricity.

I, ROY JEROME MEYERS, of Hotel Altamont, Baltimore, in the State of Maryland, United States of America, Electrician, do hereby declare the nature of this invention and in what manner the same is to be performed, to be particularly described and ascertained in and by the following statement:—

This invention relates to improvements in apparatus for the production of electrical currents, and the primary object in view is the production of a commercially serviceable electrical current without the employment of mechanical or chemical action. To this end the invention comprises means for producing what I believe to be dynamic electricity from the earth and its ambient elements. I am, of course, aware that it has been proposed to obtain static charges from upper strata of the atmosphere, but such charges are recognized as of widely variant potential and have thus far proved of no practical commercial value, and the present invention is distinguished from all such apparatus as has heretofore been employed for attracting static charges by the fact that this improved apparatus is not designed or employed to produce or generate irregular, fluctuating or other electrical charges which lack constancy, but on the other hand I have by actual test been able to produce from a very small apparatus at comparatively low elevation, say about fifty or sixty feet above the earth's surface, a substantially constant current at a commercially usable voltage and amperage. This current I ascertained by repeated tests is capable of being readily increased by additions of the unit elements in the apparatus hereinafter set forth, and I am convinced from the constancy of the current obtained and its comparatively low potential that the current is dynamic and not static, although, of course, it is not impossible that certain static discharges occur and, in fact, I have found occasion to provide against the damage which might result from such discharge by the provision of lightning arresters and cut-out apparatus which assist in rendering the obtained current stable by eliminating sudden fluctuations which sometimes occur during conditions of high humidity from what I consider static discharges. The nature of my invention is obviously such that I have been unable to establish authoritatively all of the principles involved, and some of the theories herein expressed may possibly prove erroneous, but I do know and am able to demonstrate that the apparatus which I have discovered does produce, generate, or otherwise acquire a difference of potential representing a current value which is commercially serviceable and may be delivered at the voltage and amperage above stated, or varied therefrom at the will of the operator according to the uses to which the current is to be subjected.

The invention comprises generically means for producing electrical currents of serviceable potential substantially without the employment of mechanical or chemical action, and in this connection I have been able to observe no chemical action whatever on the parts utilized although deterioration may possibly occur in some of the parts, but so far as I am able to determine such deterioration does not add to the current supply but is merely incidental to the effect of climatic action.

The invention more specifically comprises the employment of a magnet or

[*Price 8d.*]

BIRMINGHAM REFERENCE LIBRARY

magnets and a co-operating element, such as zinc disposed adjacent to the magnet or magnets and connected in such manner and arranged relative to the earth so as to produce current, my observation being that current is produced only when such magnets have their poles facing substantially to the north and south and the zincs are disposed substantially along the magnets.

The invention also comprehends other details of construction, combinations and arrangements of parts as will hereinafter be fully set forth and claimed.

In the accompanying drawings:

Figure 1 is a top plan view of an apparatus embodying the features of the present invention, the arrow accompanying the figure indicating substantially the geographical north, parts of the figure being diagrammatic for condensing the showing.

Figure 2 is a view in side elevation of the parts seen in plan in Figure 1.

Figure 3 is a vertical section taken on the plane indicated by the line 3—3 of Figure 2 and looking in the direction indicated by the arrow.

Figure 4 is a detail view partly in elevation and partly in section showing the detail connections of the converter and intensifier.

Figure 5 is a transverse section taken on the planes indicated by line 5—5 of Figure 4 and looking downwardly.

Figure 6 is an enlarged detail fragmentary section illustrating the parts at the juncture of the conductors and one of the intensifiers.

Figure 7 is an enlarged detail view partly in elevation and partly in section of one of the automatic cut-outs and

Figure 8 is a diagrammatic view of one of the simplest forms of embodiments of the invention.

Referring to the drawing by numerals, 1, 1 indicates magnets connected by a magnetic substance 2, preferably an iron wire. The magnets 1 are arranged in pairs, one pair being spaced beneath the other, and interposed between the magnets are zinc plates 3, 3 connected by an iron wire conductor 4. Suitable insulating supports 5 are arranged for sustaining the respective magnets 1 and plates 3, 3. Each plate 3 is preferably bent substantially into **V** form, as clearly seen in Figure 1, and the **V** of one of the plates opens or faces toward the north and the **V** of the other plate to the south. I have determined by experimentation that it is essential that the plates 3 be disposed substantially north and south with their flat faces approximately parallel to the adjacent faces of the co-operating magnets, although by experience I have not discovered any material difference in the current obtained when the plates are disposed slightly to one side of north and south, as for instance when disposed in the line of the magnetic polarity of the earth. The same is true with respect to the magnets 1, the said magnets being disposed substantially north and south for operative purposes, although I find that it is immaterial whether the north pole of one of the magnets is disposed to the north and the south pole to the south, or *vice versa*, and it is my conviction from experience that it is essential to have the magnets of each pair connected by magnetic material so that the magnets substantially become one with a pole exposed to the north and a pole exposed to the south. In Figure 1, I have indicated in full lines by the letters S and N the respective polarities of the magnets 1, and have indicated in dotted lines the other poles of those magnets when the connection 2 is severed. I have found that the magnets and zinc plates operate to produce, whether by collection or generation I am not certain, electrical currents when disposed substantially north and south, but when disposed substantially east and west no such currents are produced. I also find that the question of elevation is by no means vital, but it is true that more efficient results are obtained by placing the zincs and magnets on elevated supports. I furthermore find from tests that it is possible to obtain currents from the apparatus with the zincs and magnets disposed in a building or otherwise enclosed, although more efficient results are obtained by having the said elements arranged in the open.

While in Figures 1, 2 and 3, I have shown the magnets and the zinc plates as superimposed, it will be apparent, as hereinafter fully set forth, that these elements may be juxtaposed in horizontal planes, and substantially the same results will be secured. Furthermore, the magnets 1 with the interposed zincs 3, as shown in Figures 1, 2 and 3 merely represent a unit which may be repeated either horizontally or vertically for increasing the current supply, and when the unit is repeated the zinc plates are arranged alternating with the magnets throughout the entire series as hereinafter indicated.

A conductor 6 is connected in multiple with the conductors 2 and a conductor 7 is connected with conductor 4, the conductor 6 extending to one terminal of a rectifier which I have indicated by the general reference character 8, and the conductor 7 extending to the other terminal of said rectifier. The rectifier as seen in diagram in Figure 1 may assume any of several well known embodiments of the electrical valve type and may consist of four assymmetric cells or Cooper-Hewitt mercury vapor lamps connected as indicated in Figure 1 for permitting communication of the positive impulses from conductor 7 only to the line conductor 9 and the negative impulses from the conductor 7 only to the line conductor 10, and permitting the positive impulses from the conductor 6 only to the line conductor 9 and the negative impulses from conductor 6 only to the line conductor 10. The current from this rectifier may be delivered through the conductors 9 and 10 to any suitable source for consumption.

While the said rectifier 8 may consist of any of the known types, as above outlined, it preferably consists of a specially constructed rectifier which also has the capacity of intensifying the current and comprises specifically the elements shown in detail in Figures 4, 5 and 6 wherein I have disclosed the detail wiring of the rectifier when composed of four of the rectifying and intensifying elements instead of assymmetric cells or simple mercury vapor valves. As each of these structures is an exact embodiment of all the others, one only will be described, and the description will apply to all. The rectifying element of each construction consists of a mercury tube 11 which is preferably formed of copper but may be formed of glass or other suitable material, and comprises a cylinder having its end portions tapered and each terminating in an insulating plug or stopper 12. Through the upper stopper 12 is extended the electrode 13 which extends well into the tube and preferably substantially one-half the length thereof to a point adjacent the inner end of an opposing electrode 14 which latter electrode extends thence downwardly through the insulation 12 at the lower end of the tube. The tube 11 is supplied with mercury and is adapted to operate on the principle of the mercury vapor lamp, serving to rectify current by checking back impulses of one sign and permitting passage of impulses of the other. To avoid the necessity for utilizing a starter, as is common with the lamp type of electrical valve, the supply of mercury within the tube may be sufficient to contact with the lower end of the electrode 13 when current is not being supplied, so that as soon as current is passed from one electrode to the other sufficiently for volatilizing that portion of the mercury immediately adjacent the lower end of electrode 13, the structure begins its operation as a rectifier. The tube 11 is surrounded by a tube 15 which is preferably spaced from tube 11 sufficiently for allowing atmospheric or other cooling circulation to pass the tube 11. In some instances, it may be desirable to cool the tube 11 by a surrounding body of liquid, as hereinafter indicated. The tube 15 may be of insulating material but I find efficient results attained by the employment of a steel tube, and fixed to the ends of the tube are insulating disks 16, 16 forming a spool on which are wound twin wires 6^1 and 7^1, the wire 6^1 being connected at the inner helix of the coil with the outer end of the electrode 14, the lower portion of said electrode being extended to one side of the tube 11 and passed through an insulating sleeve 17 extending through the tube 15, and at its outer end merging into the adjacent end of the wire 6^1. The wire 7^1 extends directly from the outer portion of the spool through the several helices to a point adjacent the juncture of the electrode 14 with wire 6^1 and thence extends in mechanical parallelism with the wire throughout the coil, the wire 6^1 ending in a terminal 18 and the wire 7^1 ending in a terminal 19. For the sake of convenience of description and of tracing the circuits, each of the apparatus just above described and herein known as an intensifier and rectifier will be mentioned as A, B, C and D, respectively. Conductor 6 is formed with branches 20 and 21 and conductor 7 is formed with similar branches 22 and 23. Branch 20 from conductor 6 connects with conductor 7^1 of intensifier B and branch 21 of conductor 6 connects with the conductor 7^1 of intensifier C, while branch 22 of conductor 7 connects with conductor 7^1 of intensifier A and branch 23 of conductor 7 connects with conductor 7^1 of intensifier D. A conductor 27 is connected with terminal 19 of intensifier A and extends to and is connected with terminal 18 of intensifier C, and a conductor 28 is connected with the terminal 19 of intensifier C and extends to and is connected with electrode 13 of intensifier A. A conductor 29 extends from the terminal 19 of intensifier B to the terminal 18 of intensifier D and a conductor 30 extends from the terminal 19 of intensifier D to electrode 13 of intensifier B. Each electrode 13 is supported on a spider 13^1 resting on the upper disk 16 of the respective intensifier. Conductors 31 and 32 are connected with the terminals 18 of intensifiers A and B and are united to form the positive line wire 9 which co-operates with the negative line wire 10 and extends to any suitable point of consumption. The line wire 10 is provided with branches 35 and 36 extending to the electrodes 13 of intensifiers C and D for completing the negative side of the circuit.

Thus it will be seen that alternating currents produced in the wires 6 and 7 will be rectified and delivered in the form of a direct current through the line wires 9 and 10, and I find by experiment that the wires 6 and 7 should be of iron, preferably soft, and may of course be insulated, the other wiring not specified as iron being of copper or other suitable material.

In carrying out the operation as stated, the circuits may be traced as follows: A positive impulse starting at the zincs 3 is directed along conductor 7 to branch 23 to conductor 7^1 and the winding of intensifier D to terminal 19 through conductor 30 to electrode 13 of the rectifier of intensifier B through said rectifier to the conductor 6^1, through the winding thereof to the contact 18, conductor 32 and to the line wire 9. The next or negative impulse directed along conductor 7 cannot find its way along branch 23 and the circuit just above traced because it cannot pass across the rectifier of intensifier B but instead the negative impulse passes along conductor 22 to conductor 7 of intensifier A and the winding thereof to the contact 19 to conductor 27 to contact 18 of intensifier C, to the winding of the wire 6^1 thereof to the electrode 14 through the rectifier to the electrode 13 and conductor 36 to the line wire 10. A positive impulse delivered to wire 6 passes along the said wire to the branch 21 to the conductor 7^1 of intensifier C and the winding thereof to the contact 19, conductor 28, electrode 13 of the rectifier of intensifier A, electrode 14 thereof and conductor 6^1 to contact 18 and wire 31 to line wire 9. Obviously the positive impulse cannot pass along the wire 20 because of its inverse approach to the rectifier of intensifier B. The next impulse or negative impulse delivered to conductor 6 cannot pass along conductor 21 because of its connection with electrode 13 of the rectifier of intensifier A, but instead passes along conductor 20 to the wire 7^1 and its winding forming part of intensifier B to the contact 19 and conductor 29 to contact 18 and the winding of wire 6^1 of intensifier D to the electrode 14 and through the rectifier to the electrode 13 and conductor 35 to line wire 10. Thus the current is rectified and all positive impulses directed along one line and all negative impulses along the other line so that the potential difference between the two lines will be maximum for the given current of the alternating circuit. It is, of course, apparent that a less number of intensifiers with their accompanying rectifier elements may be employed with a sacrifice of the impulses which are checked back from a lack of ability to pass the respective rectifier elements, and in fact I have secured efficient results by the use of a single intensifier with its rectifier elements, as hereinafter set forth.

Grounding conductors 37 and 38 are connected respectively with the conductors 6 and 7 and are provided with the ordinary lightning arresters 39 and 40 respectively for protecting the circuit against high tension static charges.

Conductors 41 and 42 are connected respectively with the conductors 6 and 7 and each connects with an automatic cut-out 43 which is grounded as at 44. Each of said automatic cut-outs is exactly like the other and one of the same is shown in detail in Figure 7 and comprises the inductive resistance 45 provided with an insulated binding post 46 with which the respective conductor 6 or 7 is connected, said post also supporting a spring 48 which sustains an armature 49 adjacent the core of the resistance 45. The helix of resistance 45 is connected preferably through the spring to the binding post at one end and at the other end is grounded on the core of the resistance, the said core being grounded by ground conductor 44 which extends to the metallic plate 52 imbedded in moist carbon or other inductive material 53 buried in the earth. Each of the conductors 41, 42 and 44 is of iron, and in this connection I wish it understood that where I state the specific substance I am able to verify the accuracy of the statement by the results of tests which I have made, but of course I wish to include along with such substance all equivalents, as for instance, where iron is mentioned its by-products, such as steel, and its equivalents such as nickel and other magnetic substances are intended to be comprehended. The cut-out apparatus seen in detail in Figure 7 is employed particularly for insuring against high tension currents it being obvious from the structure shown that when potential rises beyond the limit established by the tension of the spring sustaining the armature 49, the armature will be moved to a position contacting with the core of the cut-out device and thereby directly close the ground connection for line wire 41 with conductor 44, eliminating the resistance of winding 45 and allowing the high tension current to be discharged to the ground. Immediately upon such discharge the winding 45 losing its current will allow the core to become demagnetized and release the armature 49 whereby the ground connection is substantially broken leaving only the connection through the winding 45 the resistance of which is sufficient for insuring against loss of low tension current.

In Figure 8 I have illustrated an apparatus which though apparently primitive in construction and arrangement comprehends the first successful embodiment which I produced in the course of the discovery of the present invention, and it will be observed that the essential features of the invention are therein disclosed. The structure delineated in said figure consists of horse shoe magnets 54, 54, one facing north and the other south, that is, each opening in the respective directions indicated and the two being connected by an iron wire 55 which is uninsulated and wrapped about the respective magnets at or adjacent the neutral zone thereof, and the wire 55 is preferably soldered to the respective magnets each end portion of the wire 55 being extended from the respective magnet to and connected with, as by being soldered to, a zinc plate 56, there being a plate 56 for each magnet and each plate being arranged longitudinally substantially parallel with the legs of the magnet and with the faces of the plate exposed toward the respective legs of the magnet, the plate being thus arranged endwise toward the north and south. An iron wire 57 connects the plates 56, the ends of the wire being preferably connected adjacent the outer ends of the plates but from experiment I find that the wire may be connected at practically any point to the plate. Lead wires 58 and 59 are connected respectively with the wires 55 and 57 and supply an alternating current ordinarily at a comparatively low tension, and to control such current the wires 58 and 59 may be extended to a rectifier or combined rectifier and intensifier, as above set forth.

The tests which I have found successful with the apparatus seen in Figure 8 were carried out by the employment first of horse shoe magnets approximately four inches in length, the bar comprising the horse shoe being about one inch square, the zincs being dimensioned proportionately and from this apparatus with the employment of a single intensifier and rectifier, as above stated, I was able to obtain a constant current of 8 volts.

It should be obvious that the magnets forming one of the electrodes of this apparatus may be permanent or may be electro-magnets, or a combination of the two.

While the magnets mentioned throughout the above may be formed of any magnetic substance, I find the best results obtained by the employment of the nickel chrome steel.

While the successful operation of the various devices which I have constructed embodying the present invention have not enabled me to arrive definitely and positively at fixed conclusions relative to the principles and theories of operation and the source from which current is supplied, I wish it to be understood that I consider myself as the first inventor of an apparatus of the general type hereinbefore described capable of producing commercially serviceable electricity, for which reason my claims hereinafter appended contemplate that I may utilize a wide range of equivalents so far as concerns details of construction suggested as preferably employed.

The current which I am able to obtain is dynamic in the sense that it is not static and its production is accomplished without chemical or mechanical action either incident to the actual chemical or mechanical motion or incident to changing calorific conditions so that the elimination of necessity for the use of chemical or mechanical action is to be considered as including the elimination of the necessity for the use of heat or varying degrees thereof.

Having now particularly described and ascertained the nature of my said invention, and in what manner the same is to be performed, I declare that what I claim is:—

1. Means for producing dynamic electricity without mechanical or chemical action.

2. Electricity producing means as claimed in Claim 1, characterized in that said means is capable of producing low tension current.

3. Electricity producing means as claimed in Claim 1, characterized in that said means comprises a magnet, and means co-operating with the latter.

4. Means as set forth in Claim 3, characterized in that the parts are stationary.

5. Means as claimed in Claim 3, characterized in that the magnet is disposed substantially north and south.

6. Means as claimed in Claims 1, 3 and 5, wherein the means coacting with the magnet is a metal and said parts are disposed substantially north and south.

7. Means as claimed in Claims 1, 3 and 6, in which the metal is zinc, and combined with conductors connected with the co-operating parts to deliver current therefrom.

8. Means as set forth in Claim 1, characterized in that said means comprises spaced magnets with an adjacent co-operating zinc, and an iron wire connecting the magnets, and means to deliver electrical current from said parts.

9. Means as set forth in Claim 8, characterized in that the current delivery means comprises an intensifier and rectifier.

10. Means as set forth in Claim 8, characterized in that the current delivery means comprises wires connected with the magnets and zinc, an electrical valve connected with one of said wires, and independent helices surrounding said valve, one of said helices being connected with the terminal of the valve at the opposite side from the contact of said wire and the other of said helices being connected with the other wire, the electrical valve comprising a mercury vapor rectifier.

11. The process of producing electricity comprising exposing a magnet and

co-operating means disposed substantially north and south and taking off current therefrom.

12. The process of producing electricity comprising disposing stationary elements in co-operative relation in respect to each other and with respect to the earth for giving off relatively low tension electrical current, substantially without chemical action.

13. Electricity producing means constructed and operating substantially as described with reference to the accompanying drawings.

Dated this 13th day of January, 1913.

W. P. THOMPSON & Co.,
6, Lord Street, Liverpool, and at
Bradford & London,
Agents for the Applicant.

Reference has been directed, in pursuance of Section 7, Sub-section 4, of the Patents and Designs Act, 1907, to Specifications No. 16,709 of 1887, No. 14,033 of 1899, No. 15,412 of 1906, and No. 5457 of 1911.

Redhill: Printed for His Majesty's Stationery Office, by Love & Malcomson, Ltd.—1914.

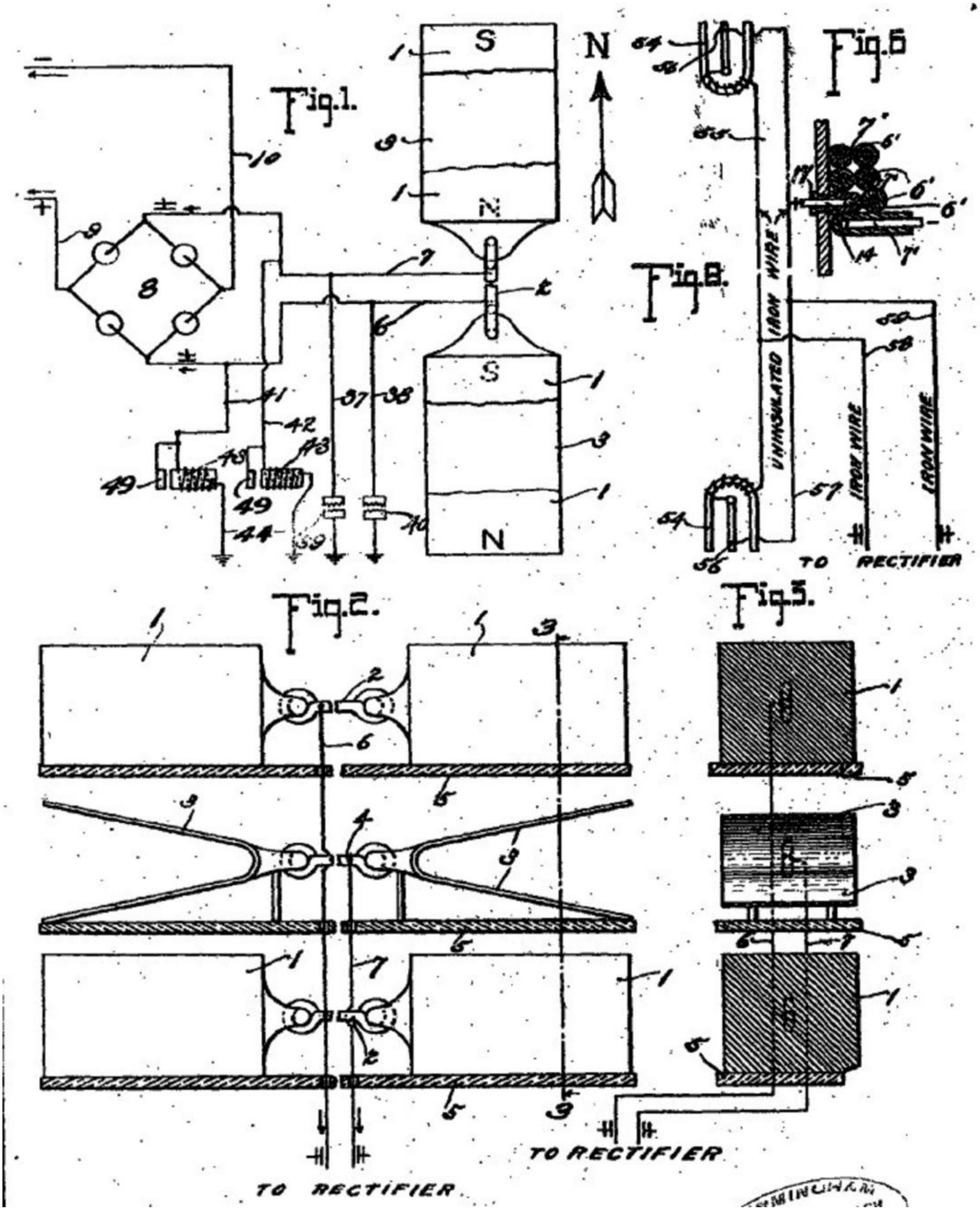
Fig.1.
S
N
t
Fig.6
Fig.8.
UNINSULATED IRON WIRE
IRON WIRE
TO RECTIFIER
Fig.2.
Fig.3.
TO RECTIFIER

Fig 7.

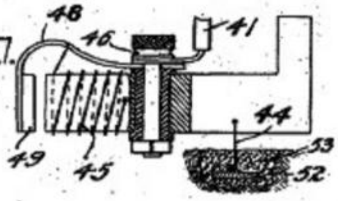
48
46
41
44
53
52
49
45

Fig 4.

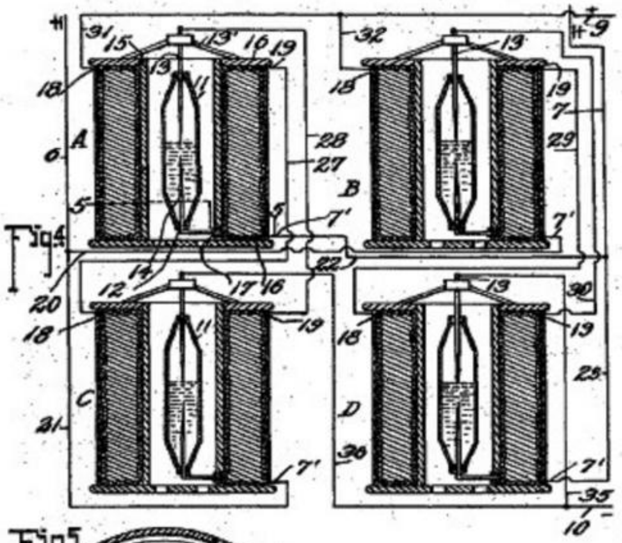
+1
31
15
13'
16
19
32
13
+9
18
A
6
11
28
27
B
5
7'
19
7
29
12
14
17
16
22
20
18
11
19
13
30
C
21
D
36
7'
25
35
10

Fig 5.

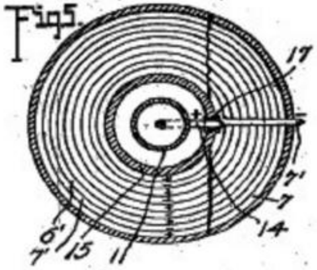
17
7'
7
14
6'
7
15
11

Your Support Maintains this Service. Please Invest in the mind-boggling Rex Research Website CD -- Always The Best Investment ... Everything on this Website & More (the Bonus Files CD) -- Only $13, Postpaid Anywhere ! Secure Transaction via PayPal: All Major Credit Cards Accepted:

Buy Now

No Credit ? Don't like PayPal ? Send a Check/Money Order to:

Rex Research, PO Box 19250, Jean, NV 89019 USA

---

rexresearch.com

Roy J. Meyers Absorber, GB Patent 191301098.

This device was brought to my attention by Grumage. Here is my take on why it worked.

In its simplest form the device consists of two horseshoe magnets, some zinc plates and some uninsulated iron wire. The iron wire is wound around the curved ends of the horseshoe magnet and connected to the zinc plates as shown in figure 1. The system has to be aligned with the earth's magnetic field.

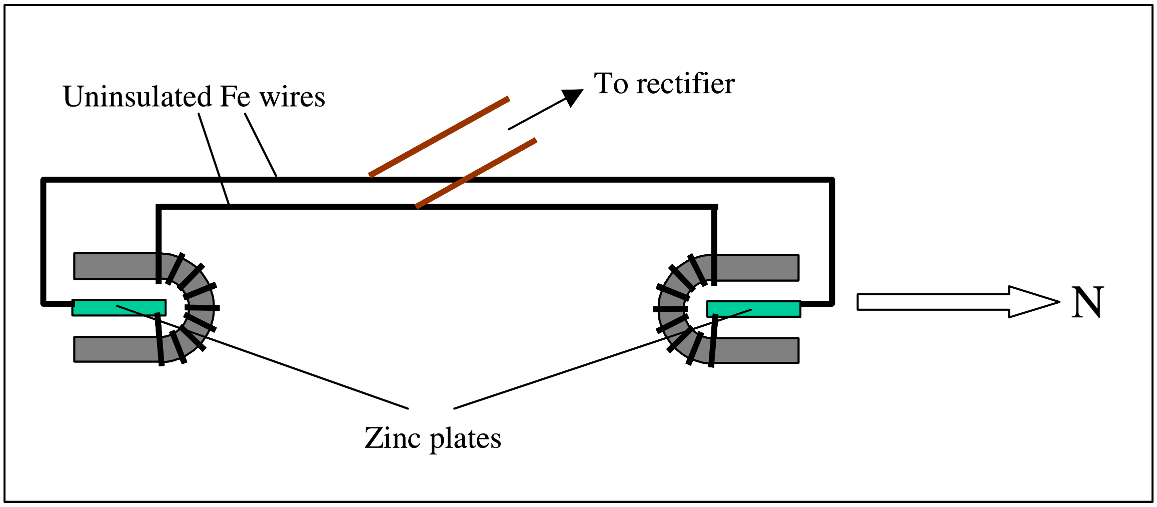

Figure 1. Simple Scheme reported to supply alternating current at low voltage

At first sight this seems incapable of providing alternating voltage but here is a possible explanation. If you follow the magnetization axis of each horseshoe it is seen that the earth's field supplies a magnetic gradient around the curved section, hence spin polarized conduction electrons therein will be dragged along thus creating a tiny potential difference between the two legs of each magnet. The iron wires wound over that curved end do nothing except create contact points at the ends of each curvature. We can discount any galvanic potential between iron wire and magnet since the magnets used were of magnetized steel. However we cannot ignore the galvanic potential between the zinc and the iron connection, or indeed the solder used for the connection. Although the two connections to the zinc would normally result in a zero overall potential, there exists a difference in the magnetic field at each end of the zinc, which could result in a significant (but low) voltage present. This magneto-Seebeck effect is a fairly new phenomenon and would not have been known in 1913 when the patent application was made. Also unknown at that time was the presence of spin-polarized conduction electrons.

This all predicts the presence of DC voltage, so how could this become AC? Well the magneto-Seebeck effect can result in a positive or a negative potential, and it also changes with temperature. In fact there can be a crossover temperature where it changes from positive to negative. Hence the DC current flowing around the closed loop of iron wire could heat the thin zinc plates to take it through that crossover temperature, whereupon the current decreases, the zinc cools down then the whole process repeats itself. This would result in a very low frequency AC voltage being observed.

US 20140152016A1

(19) **United States**

(12) **Patent Application Publication**

JENNINGS

(10) **Pub. No.: US 2014/0152016 A1**

(43) **Pub. Date: Jun. 5, 2014**

(54) **ATMOSPHERIC TRANSDUCTION SYSTEM**

(71) Applicant: **JAMES EDWARD JENNINGS**, SUPERIOR, CO (US)

(72) Inventor: **JAMES EDWARD JENNINGS**, SUPERIOR, CO (US)

(21) Appl. No.: **13/692,121**

(22) Filed: **Dec. 3, 2012**

Publication Classification

(51) **Int. Cl.**

H02N 11/00	(2006.01)
F03D 9/00	(2006.01)

(52) **U.S. Cl.**

CPC ***H02N 11/002*** (2013.01); ***F03D 9/003*** (2013.01)

USPC **290/55**; 310/308

(57) **ABSTRACT**

Atmospheric Transduction System including a Power Frequency broadcast station, a receiver, and a network. The Power Frequency broadcast station includes a transmitter and a computer server. The receiver is in communication with the Power Frequency broadcast transmitter and also includes a user interface for receiving user input commands comprising a request for information from the Power Frequency broadcast station. The receiver is configured to establish a two-way communication path between the receiver and the Power Frequency broadcast transmitter. The network is in communication with the transducer, controller and the receiver for exchanging information therebetween. In response to oscillation translation and/or rotation of the electronic transducer, portions of forces induced by the mass are transferred to the piezoelectric elements. Electrical energy output by these piezoelectric elements is received in a power controller and can be applied to the battery as self charging. The piezoelectric transducer includes a conductive rotor and bearings, at least one of them incorporating a vibrator of mechanical oscillation, having a piezoelectric transducer converting mechanical vibrations into electric power.

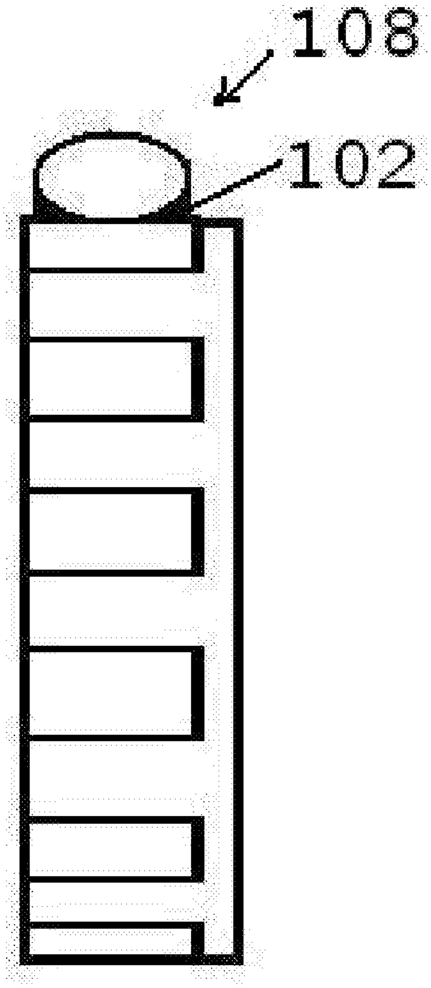

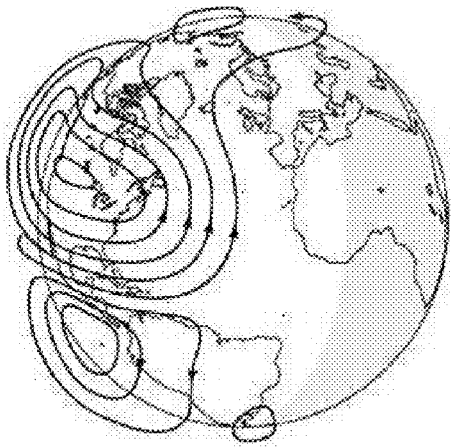

FIG. 1

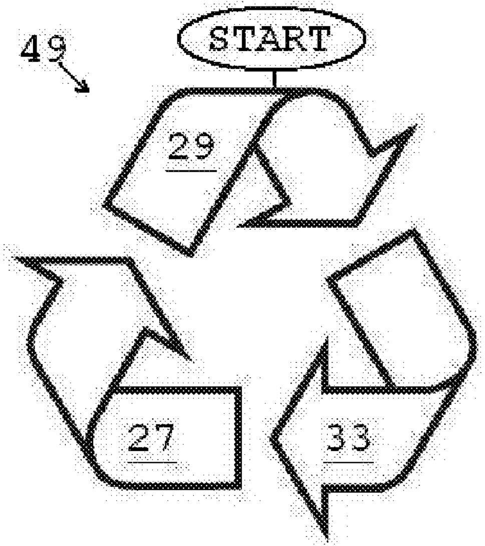
49
START
29
27
33

FIG. 2

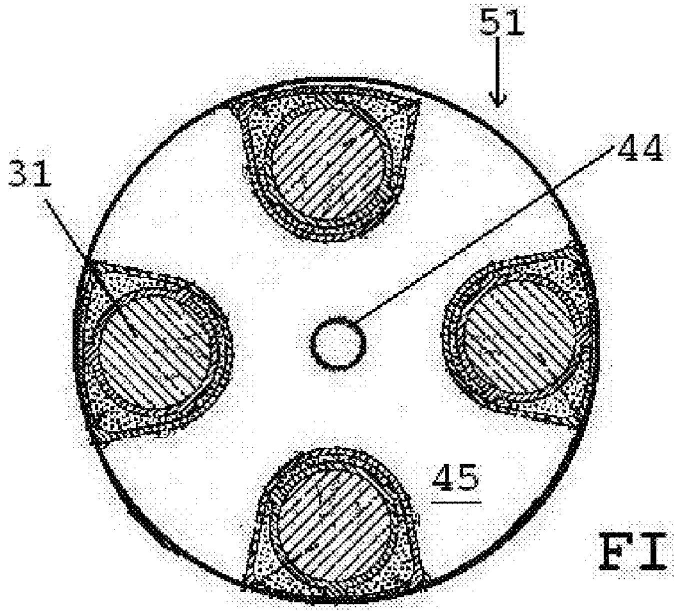

FIG. 3

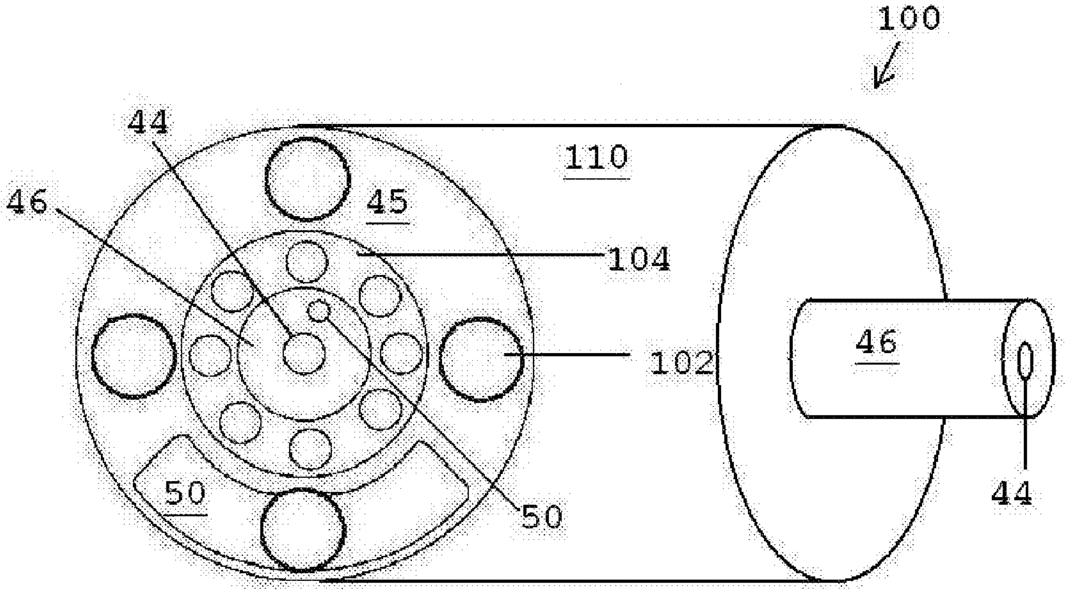
100
44
110
46
45
104
46
102
50
50
44

FIG. 4

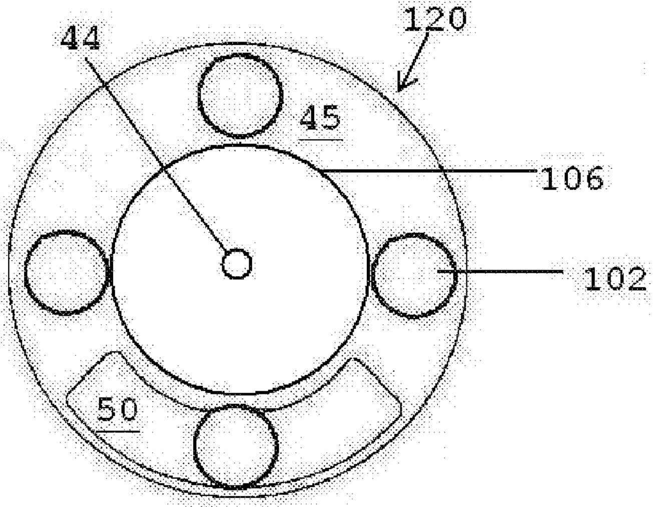
120
44
45
106
102
50

FIG. 5

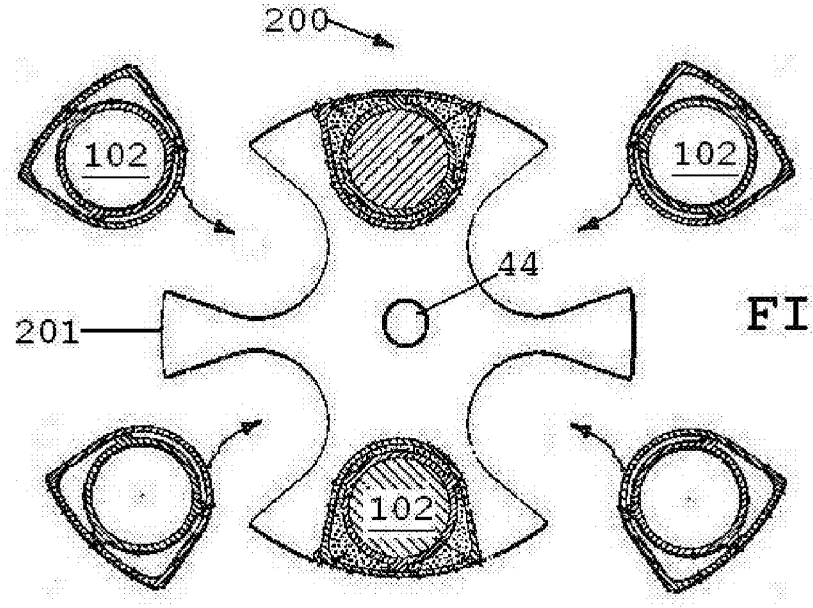
200
102
102
44
201
102

FIG. 6

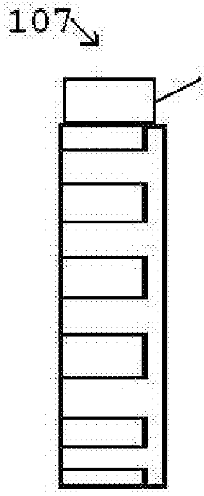

FIG. 7

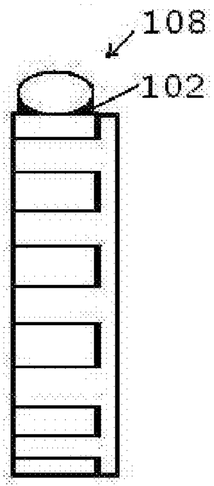

FIG. 8

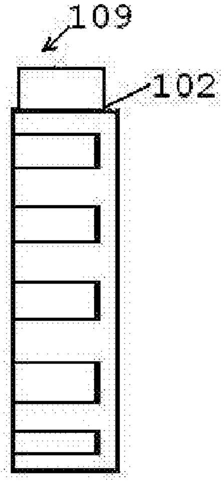

FIG. 9

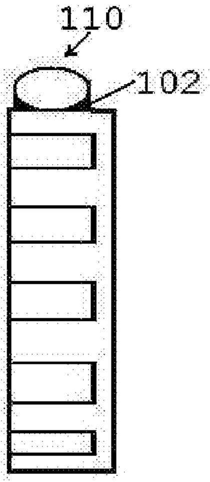

FIG. 10

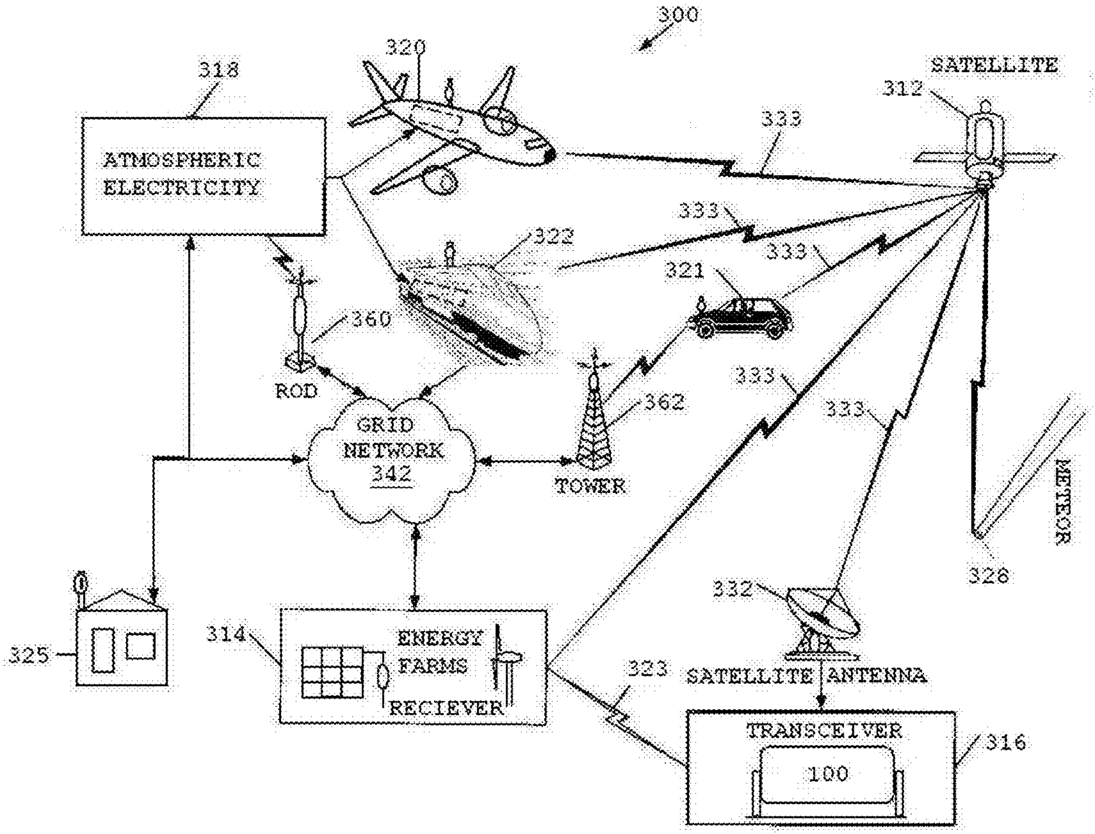
ATMOSPHERIC ELECTRICITY
SATELLITE
ROD
GRID NETWORK 342
TOWER
METEOR
ENERGY FARMS
RECIEVER
SATELLITE ANTENNA
TRANSCEIVER

FIG.11

FIG.12

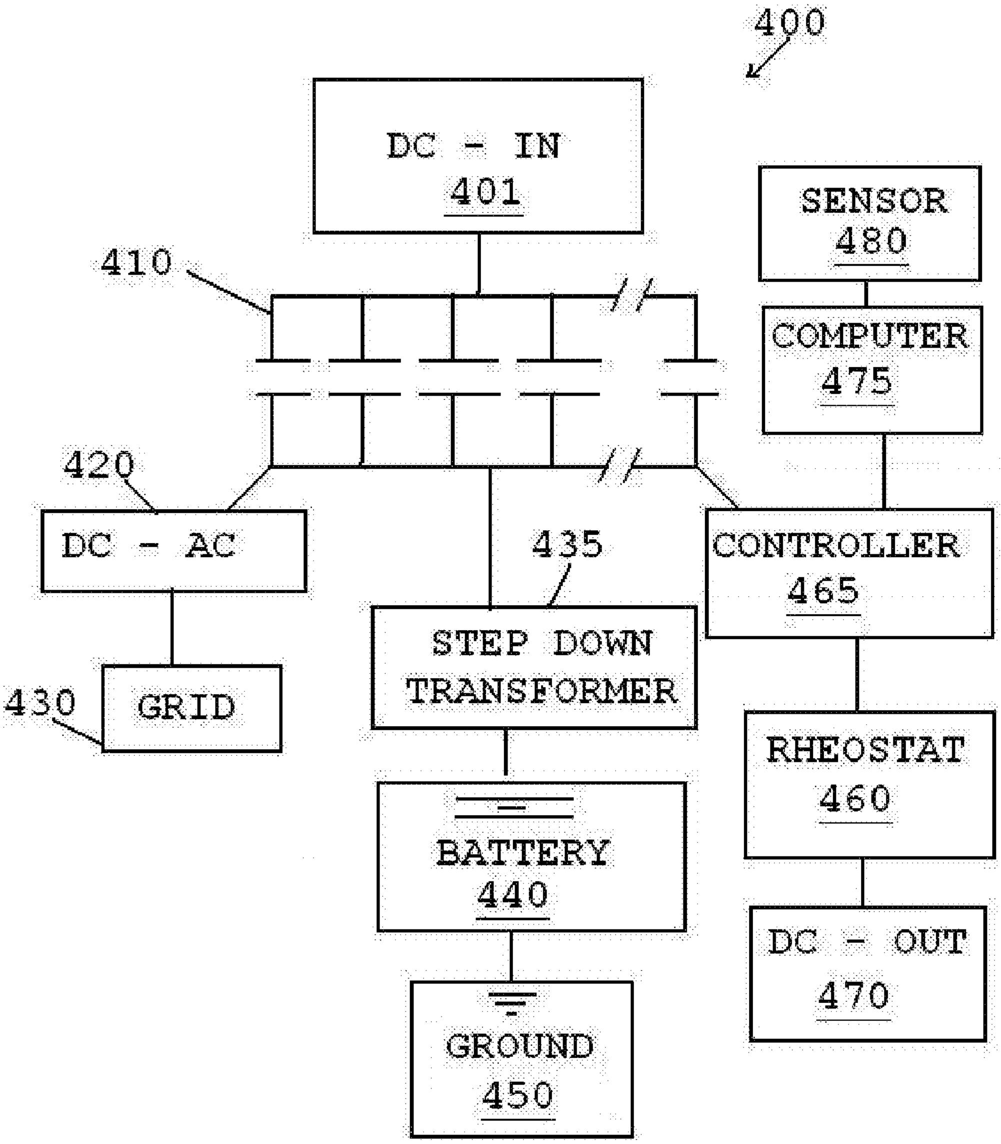
400
DC - IN
401
410
SENSOR
480
COMPUTER
475
420
DC - AC
435
CONTROLLER
465
STEP DOWN
TRANSFORMER
430
GRID
RHEOSTAT
460
BATTERY
440
DC - OUT
470
GROUND
450

FIG. 13

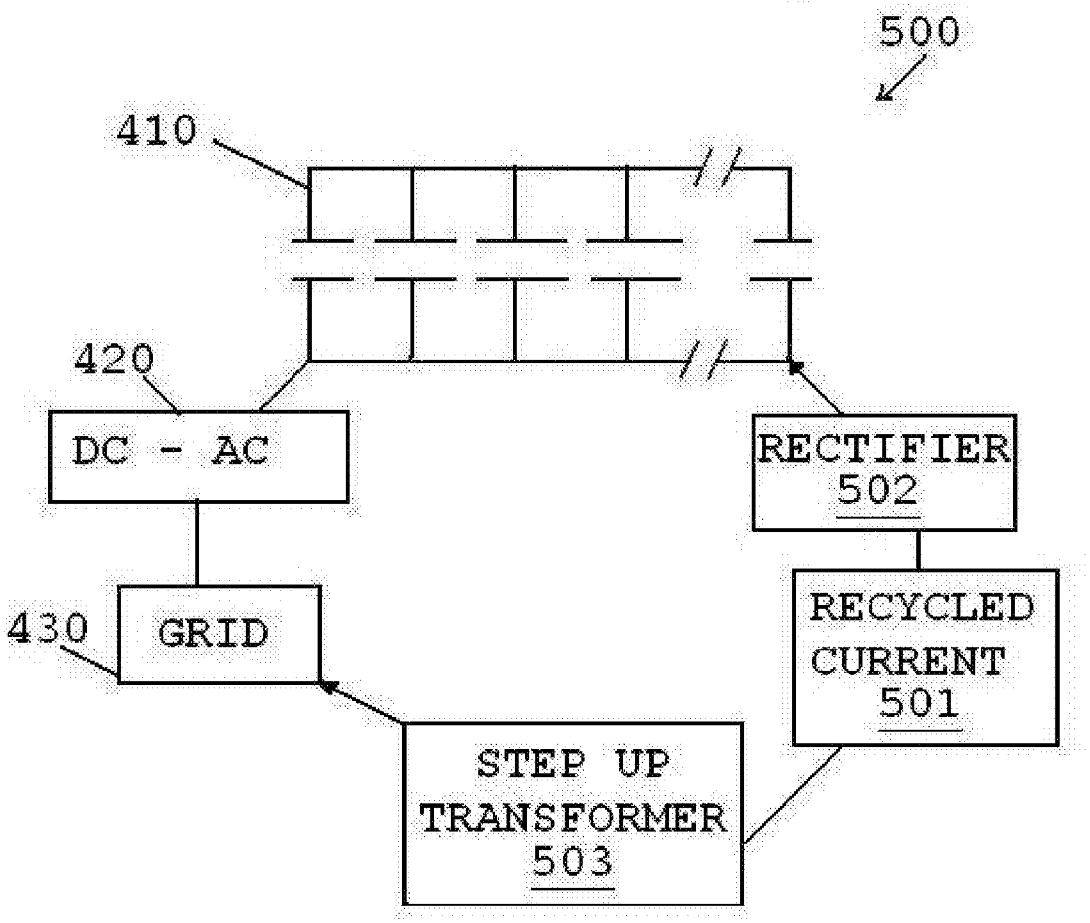
500
410
420
DC - AC
430
GRID
RECTIFIER
502
RECYCLED
CURRENT
501
STEP UP
TRANSFORMER
503

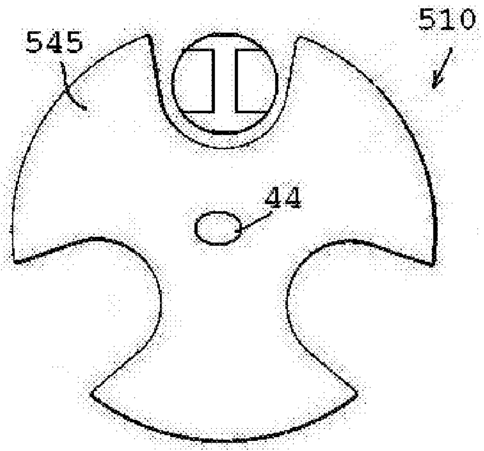

FIG.14

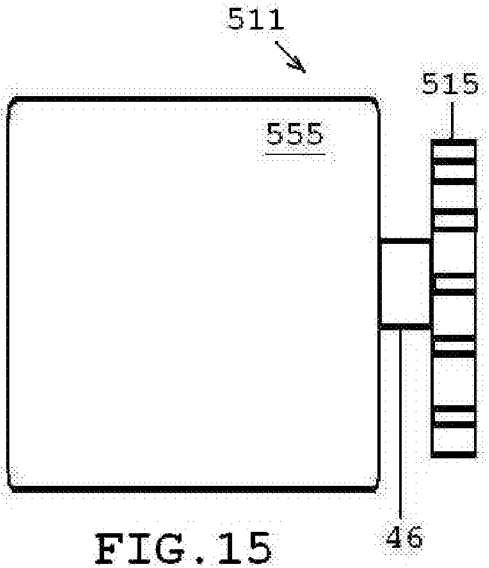
511
515
555
46

FIG.15

FIG.16

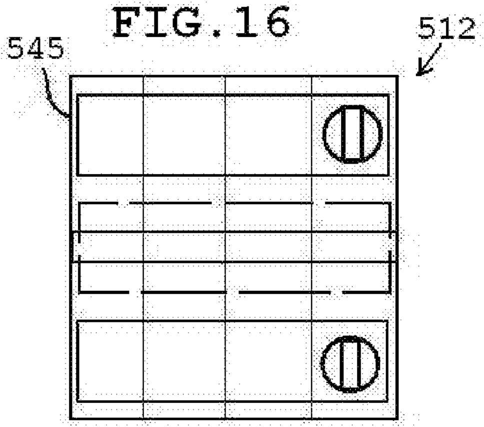

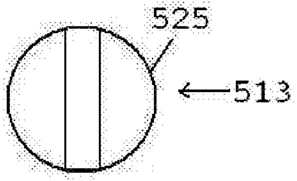

FIG.17

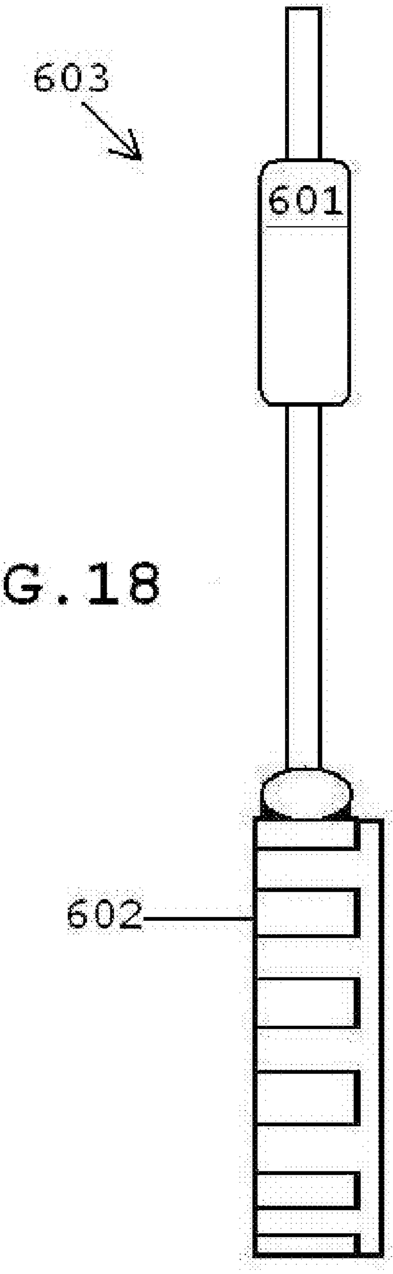

FIG.18

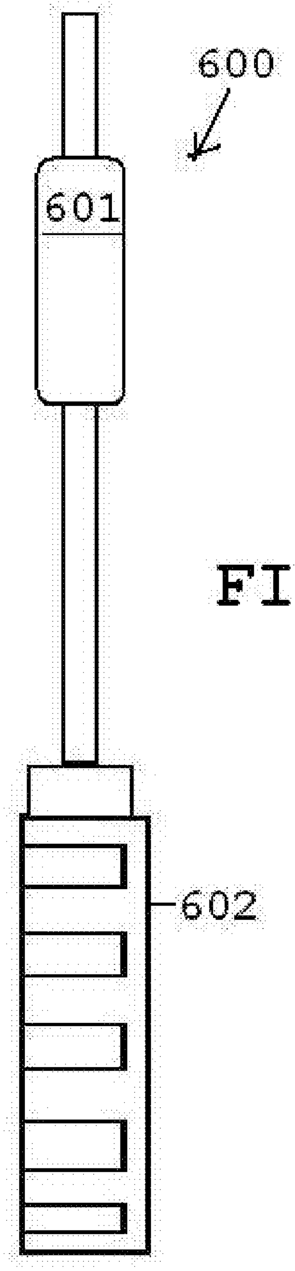

FIG.19

FIG. 20

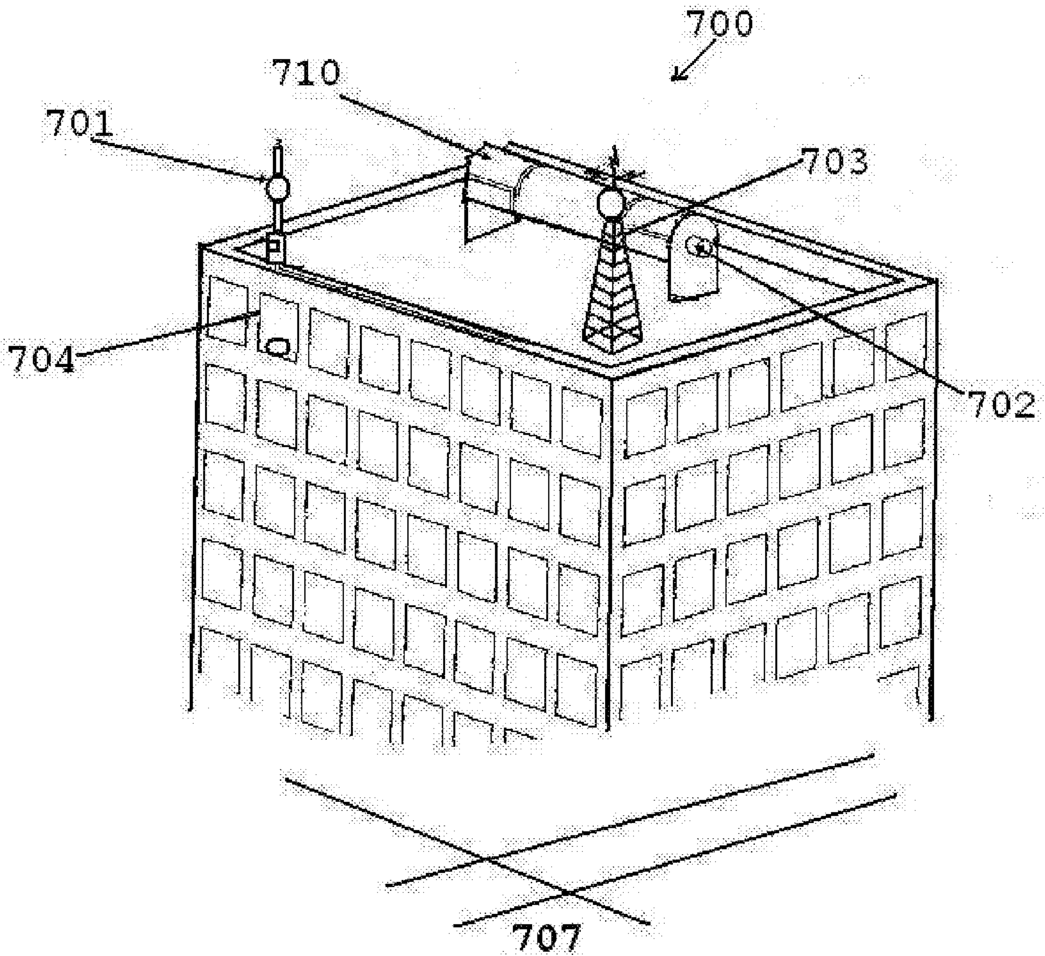
700
710
701
703
704
702
707

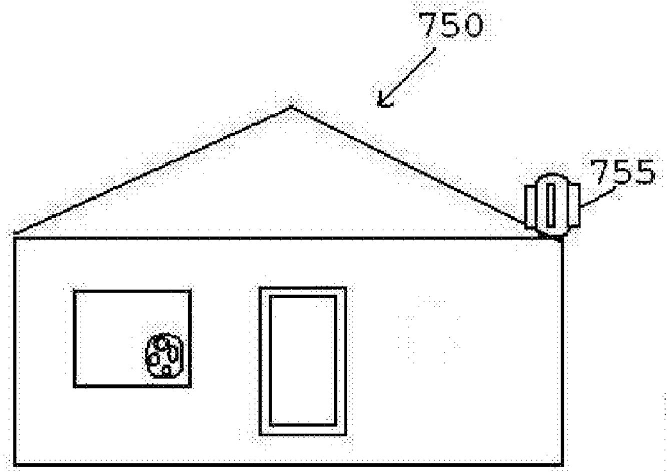

FIG.21

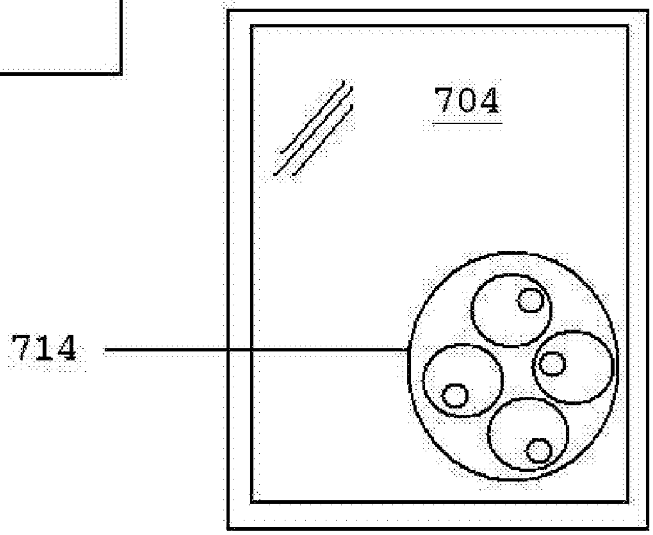

FIG.22

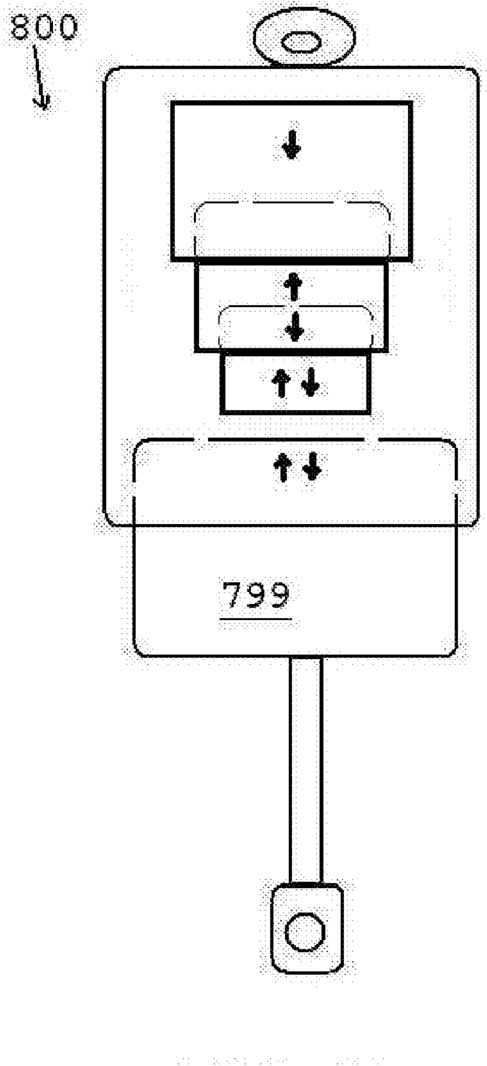

FIG.23

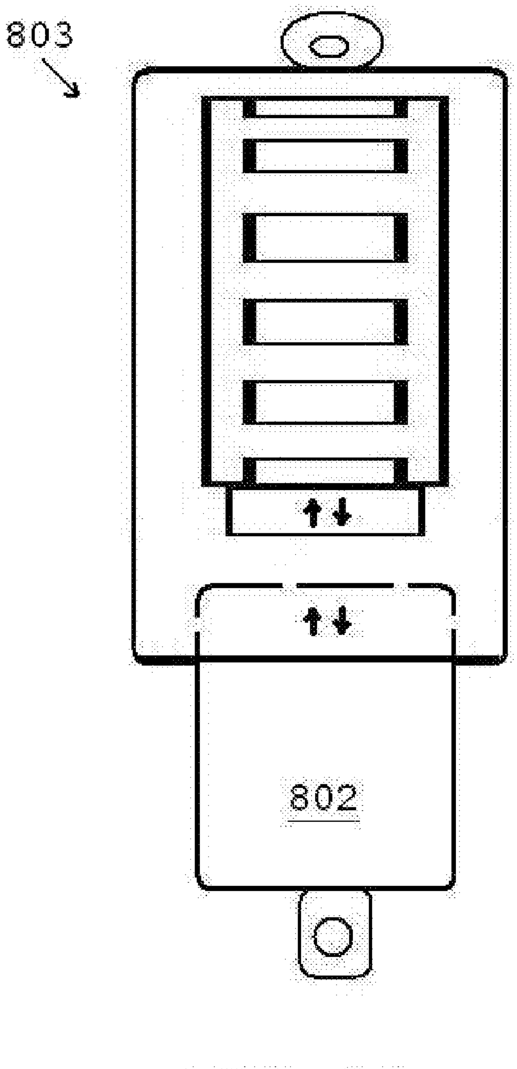

FIG.24

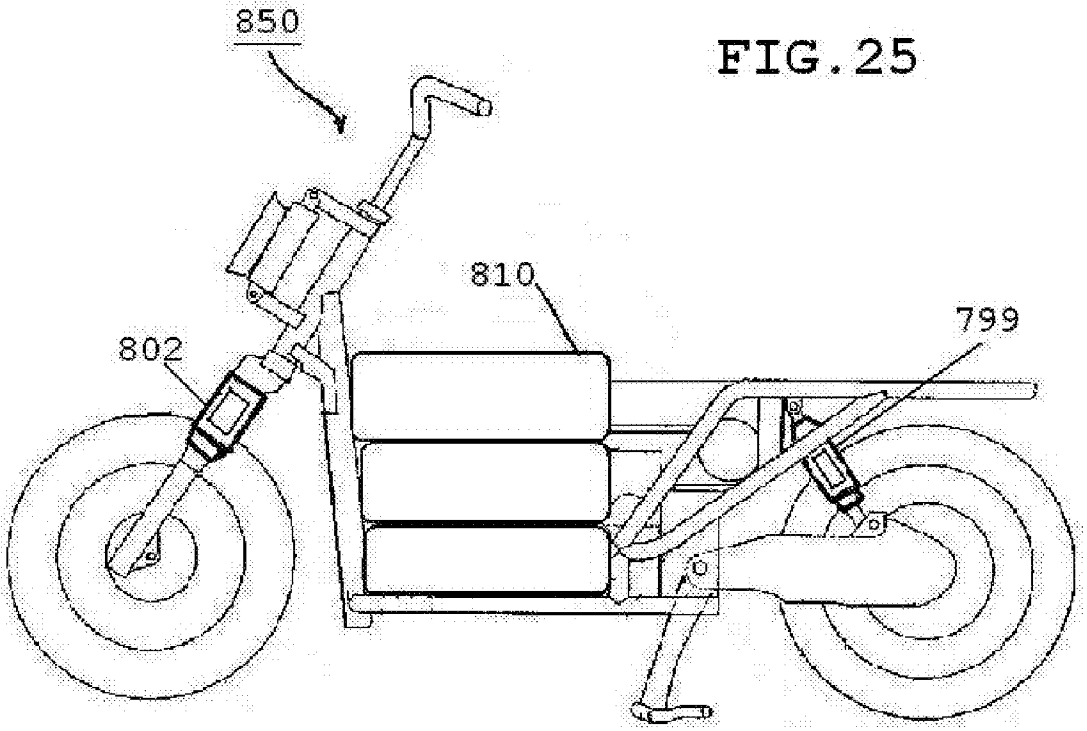
850
810
799
802

FIG.25

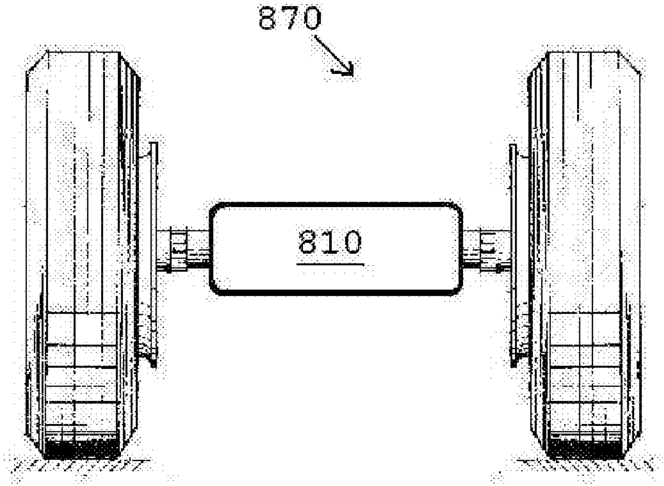

FIG.26

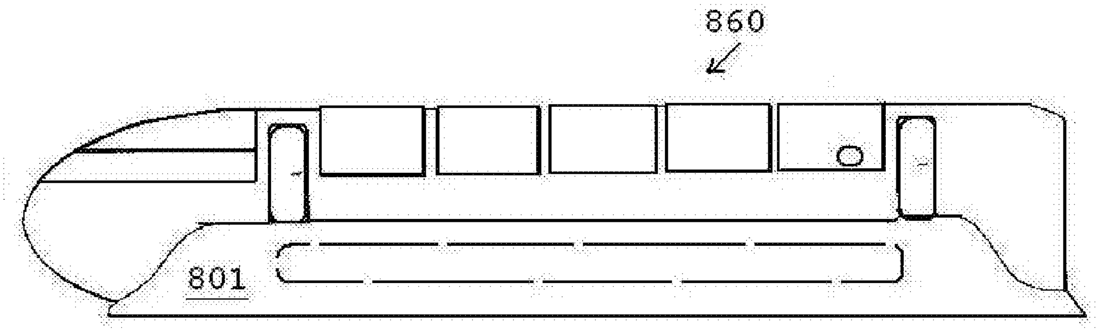

FIG. 27

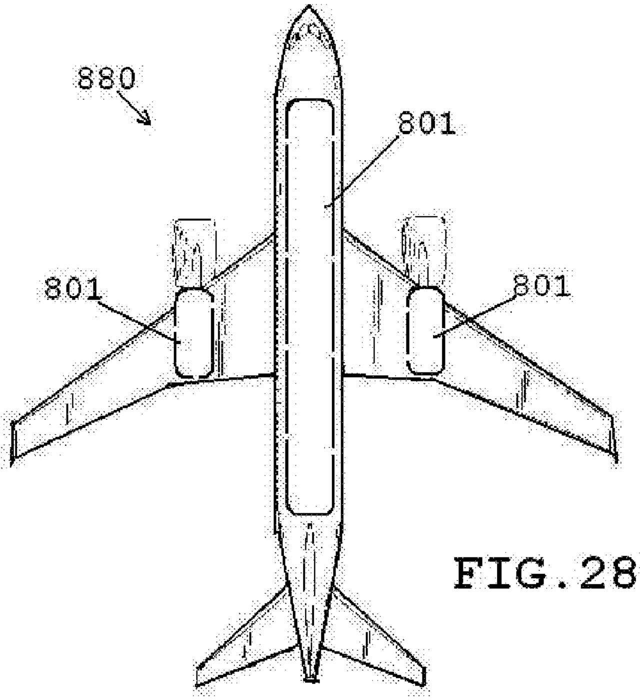
880
801
801
801

FIG. 28

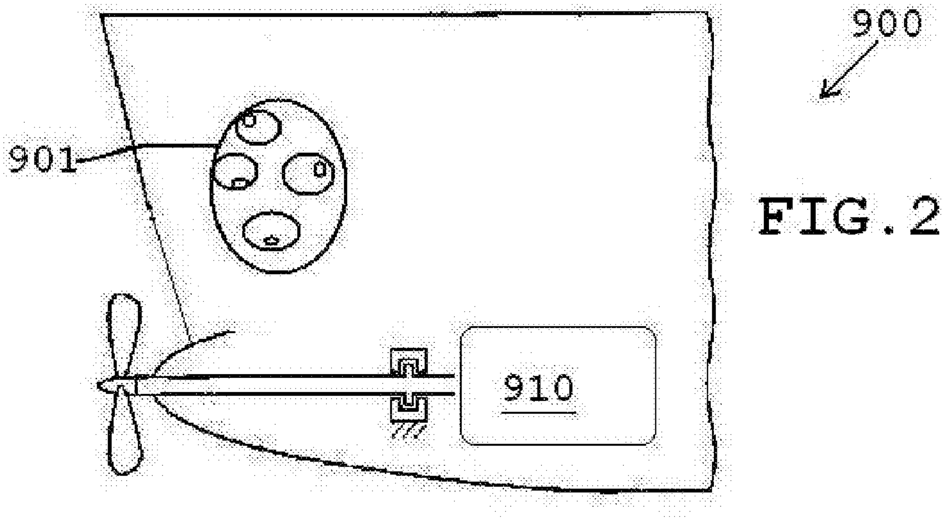

FIG.29

FIG.30

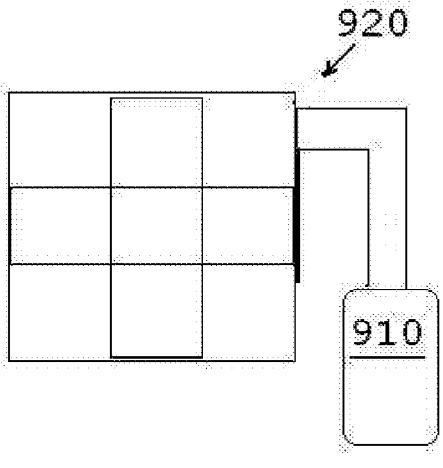

FIG.31

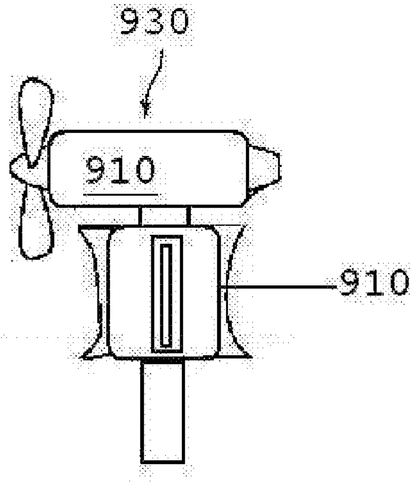

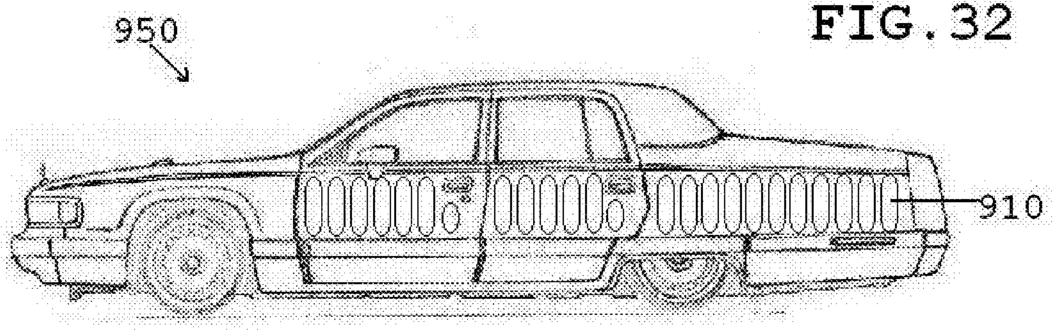

FIG. 32

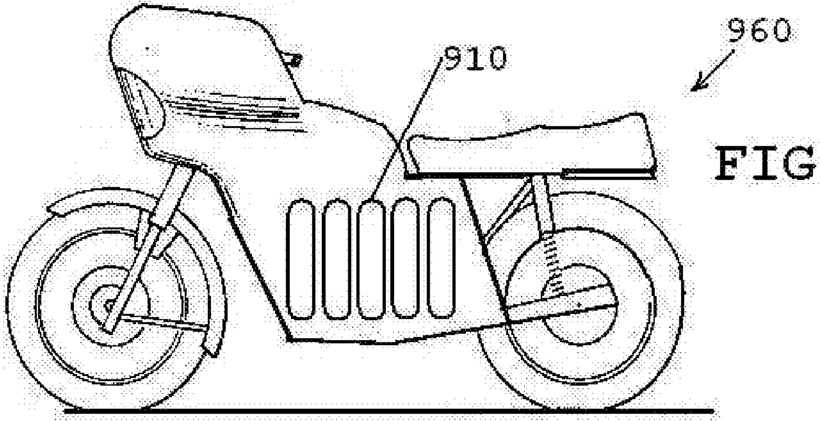

FIG. 33

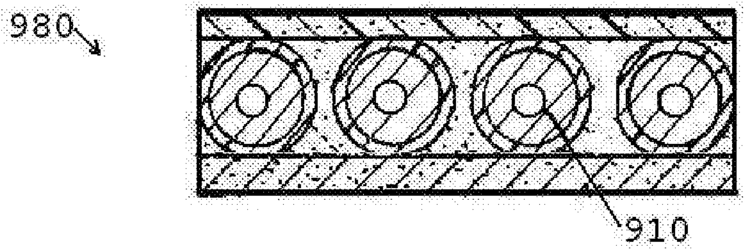

FIG. 34

ATMOSPHERIC TRANSDUCTION SYSTEM

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] NONE

BACKGROUND OF THE INVENTION

Field of the Invention

[0002] Self-propelled travel is a type of recreational adventure travel using human powered transport. This includes non-motorized machines such as a bicycle or skateboard. It is in contrast to traveling in a powered vehicle (such as an automobile) as in that case it is the vehicle which powers itself. Self-propelled travel is used to travel short distances or even for much longer distances such as bicycle touring. Self propelled describes something that moves, progresses or acts on its own power without needing outside help. Leonardo da Vinci's 1478 Self-Propelled Car: It was more than 500 years ago, however—sometime around the year 1478 to be more or less specific—when Leonardo drew out his plans for the world's first self-propelled vehicle. Experts originally believed two leaf springs, the simplest form of the spring typically used for automotive suspensions, somehow powered the vehicle. Closer inspection eventually revealed the power came from bigger, coiled springs located in tambours, cylindrical drum-like casings, inside the car's frame. The machine works like a robot or a wind-up toy simply by rotating the wheels opposite of their intended direction, which winds up the springs inside and gives it power. Self propulsion (of a vehicle) provided with its own source of tractive power rather than requiring an external means of propulsion.

[0003] In the middle of the 18th century, Benjamin Franklin's experiments showed that electrical phenomena of the atmosphere were not fundamentally different from those produced in the laboratory. By 1749, Franklin observed lightning to possess almost all the properties observable in electrical machines.

[0004] In July 1750, Franklin hypothesized that electricity could be taken from clouds via a tall metal aerial with a sharp point. Before Franklin could carry out his experiment, in 1752 Thomas-Francois Dalibard erected a 40-foot (12 m) iron rod at Marly-la-Ville, near Paris, drawing sparks from a passing cloud. With ground-insulated aerials, an experimenter could bring a grounded lead with an insulated wax handle close to the aerial, and observe a spark discharge from the aerial to the grounding wire. In May 1752, Dalibard affirmed that Franklin's theory was correct.

Piezoelectric Motor

[0005] A piezoelectric motor or piezo motor is a type of electric motor based upon the change in shape of a piezoelectric material when an electric field is applied. Piezoelectric motors make use of the converse piezoelectric effect whereby the material produces acoustic or ultrasonic vibrations in order to produce a linear or rotary motion. In one mechanism, the elongation in a single plane is used to make a series stretches and position holds, similar to the way a caterpillar moves. A transducer is a device that converts one form of energy to another. Energy types include (but are not limited to) electrical, mechanical, electromagnetic (including light), chemical, acoustic or thermal energy. While the term transducer commonly implies the use of a sensor/detector, any device which converts energy can be considered a transducer. Transducers are widely used in measuring instruments. Piezoelectric materials can also be used to harvest low levels of mechanical energy into electrical energy suitable for powering wireless sensors, low power microprocessors or charging batteries. Rotary Uses include rotating machines such as fans, turbines, drills, the wheels on electric cars, locomotives and conveyor belts. Also, in many vibrating or oscillating machines, an electric motor spins an unbalanced mass, causing the motor (and its mounting structure) to vibrate.

Atmospheric Electricity

[0006] There is always free electricity in the air and in the clouds, which acts by induction on the earth and electromagnetic devices. Experiments have shown that there is always free electricity in the atmosphere, which is sometimes negative and sometimes positive, but most generally positive, and the intensity of this free electricity is greater in the middle of the day than at morning or night and is greater in winter than in summer. In fine weather, the potential increases with altitude at about 30 volts per foot (100 V/m).

Atmospheric Layers

[0007] The electrical conductivity of the atmosphere increases exponentially with altitude. The amplitudes of the electric and magnetic components depend on season, latitude, and height above the sea level. The greater the altitude the more atmospheric electricity abounds. The exosphere is the uppermost layer of the atmosphere and is estimated to be 500 km to 1000 km above the Earth's surface, and its upper boundary at about 10,000 km. The thermosphere (upper atmosphere) is the layer of the Earth's atmosphere directly above the mesosphere and directly below the exosphere. Within this layer, ultraviolet radiation causes ionization. Theories that have been proposed to explain the phenomenon of the polar aurora, but it has been demonstrated by experiments that it is due to currents of positive electricity passing from the higher regions of the atmosphere to the earth.

[0008] The mesosphere (middle atmosphere) is the layer of the Earth's atmosphere that is directly above the stratosphere and directly below the thermosphere. The mesosphere is located about 50-80/85 km above Earth's surface. The stratosphere (middle atmosphere) is a layer of Earth's atmosphere that is stratified in temperature and is situated between about 10 km and 50 km altitude above the surface at moderate latitudes, while at the poles it starts at about 8 km altitude. The stratosphere sits directly above the troposphere and directly below the mesosphere. The troposphere (lower atmosphere) is the densest layer of the atmosphere.

[0009] The planetary boundary layer (PBL), also known as the atmospheric boundary layer (ABL), is the lowest part of the atmosphere and its behavior is directly influenced by its contact with the planetary surface. It is also known as the "exchange layer". (see also: p-n junction.)

[0010] There is a potential gradient at ground level ("Atmosphere ground layer") and this vertical field corresponds to the negative charge in and near the Earth's surface. The negative potential gradient falls rapidly as altitude increases from the ground. Most of this potential gradient is in the first few kilometers. The positive potential gradient rises rapidly as altitude increases from the ground. Volta, over two centuries before the 21st century, discovered with some degree of exactitude that the proportions of the ordinates of the curve or gradient of electric potential increased as the distance from the earth increases, and, more recently, Engel has provided data to calculate the increase (Image to the right).

Drum-Type Generator

[0011] A drum-type homopolar generator has a magnetic field (B) that radiates radially from the center of the drum and induces voltage (V) down the length of the drum. A conducting drum spun from above in the field of a “loudspeaker” type of magnet that has one pole in the center of the drum and the other pole surrounding the drum could use conducting ball bearings at the top and bottom of the drum to pick up the generated current.

Astrophysical Unipolar Inductors

[0012] Unipolar inductors occur in astrophysics where a conductor rotates through a magnetic field, for example, the movement of the highly conductive plasma in a cosmic body’s ionosphere through its magnetic field. In their book, Cosmical Electrodynamics, Hannes Alfvén and Carl-Gunne Fälthammar write:

[0013] “Since cosmical clouds of ionized gas are generally magnetized, their motion produces induced electric fields [. . .] For example the motion of the magnetized interplanetary plasma produces electric fields that are essential for the production of aurora and magnetic storms” [. . .]

[0014] “ . . . the rotation of a conductor in a magnetic field produces an electric field in the system at rest.

[0015] This phenomenon is well known from laboratory experiments and is usually called ‘homopolar’ or ‘unipolar’ induction.

The Faraday Disc

[0016] The homopolar generator was developed first by Michael Faraday during his experiments in 1831. It is frequently called the Faraday disc in his honor. It was the beginning of modern dynamos—that is, electrical generators which operate using a magnetic field. It was very inefficient and was not used as a practical power source, but it showed the possibility of generating electric power using magnetism, and led the way for commutated direct current dynamos and then alternating current alternators.

Boeing 737-800

[0017] The Boeing Fuel Cell Demonstrator Airplane has a Proton Exchange Membrane (PEM) fuel cell/lithium-ion battery hybrid system to power an electric motor, which is coupled to a conventional propeller. The fuel cell provides all power for the cruise phase of flight. During takeoff and climb, the flight segment that requires the most power, the system draws on lightweight lithium-ion batteries.

[0018] The demonstrator aircraft is a Dimona motor glider, built by Diamond Aircraft Industries of Austria, which also carried out structural modifications to the aircraft. With a wing span of 16.3 meters (53.5 feet), the airplane will be able to cruise at approximately 100 kilometers per hour (62 miles per hour) on power from the fuel cell.

[0019] Nikola Tesla explored the wireless transmission of energy through his work with radio and microwaves and his creation of the Tesla coil and the magnifying transmitter. In 1898, Tesla demonstrated his radio-controlled boat, which he was able to control remotely. In the 1930s, Tesla claimed to have invented a particle beam weapon, or, as some called it, a “peace ray.” The device was, in theory, capable of generating an intense, targeted beam of energy and sending it across great distances to demolish warplanes, foreign armies, or anything else you’d rather didn’t exist.

[0020] “Roy J. Meyers, British Patent Number 1098”

[0021] This invention relates to improvements in apparatus for the production of electrical currents, and the primary object in view is the production of a commercially serviceable electrical current without the employment of mechanical or chemical action. To this end the invention comprises means for producing what I believe to be dynamic electricity from the earth and its ambient elements.

[0022] Edward Leedskalnin: Magnetic Current—

[0023] The Perpetual Motion Holder is primarily a teaching device but it has many functions including an electromagnet, this is easy enough to see; it is a generator—spin a magnet between the coils it will generate electricity; it functions as a transformer; it demonstrates how permanent magnets are made, and is a holder of perpetual motion.

Strategic Defense Initiative

[0024] The Strategic Defense Initiative (SDI) was proposed to use ground and space-based systems to protect the United States from attack by strategic nuclear ballistic missiles.

Description of the Related Art

[0025] The present invention relates to a ball bearing assembly structure, an electromagnetic clutch having the ball bearing assembly structure, and a gas compressor equipped with the electromagnetic clutch.

[0026] When operating the gas compressor, the electromagnet of the electromagnetic clutch is energized to attract or adsorb the follower armature plate to an end surface of the prime-mover pulley and join the prime-mover pulley and the rotor shaft, thereby rotating the rotor shaft.

[0027] The ball bearing of the electromagnetic clutch conventionally has used one having an even number of balls per row. Generally, the ball bearing causes vibration and noise due to rotation. In the case of the ball bearing rotating while undergoing a radial load due to a tension of the belt, vibration and noise considerably occur. Particularly when other vibration and noise levels are lowered during engine idling, the vibration and noise of the ball bearing transmitted to the vehicular compartment is not negligible.

[0028] The inventor has conducted various experiments and discovered that the one factor of high vibration and noise level is an even number of balls of the ball bearing. In the ball bearing having an even number of balls per one row, the balls are in a facing relation to have linear-symmetry arrangement between the inner race and the outer race. The deformation and vibration at a regular particular frequency is caused in the inner and outer races. It is to be considered that the vibration as a source also increases noise.

[0029] The present invention relates generally and in various embodiments to piezoelectric mechanical systems. More specifically, the present invention relates generally and in various embodiments to atmospheric oscillation transducer apparatuses, systems, and methods.

[0030] Although various implementations of the present invention, among many, may be described herein with reference to the specific illustrative embodiments related to particular applications, those skilled in the art will understand that the invention is not in any way intended to be limited to such embodiments and/or applications. Those having ordinary skill in the art and reference to the description of the embodiments herein will recognize additional modifications, applications, and other embodiments falling within the scope of the claimed invention and additional fields in which the present invention may be practiced.

[0031] Digital Radio (also known as Satellite Radio or Satellite Digital Audio Radio Service (SDARS)) is a subscriber-based digital radio service that is broadcast via satellites. Digital radio service provides compact-disc (CD) quality programming that may be digitally transmitted via one or more satellites and/or space stations to one or more Earth-based (terrestrial) digital radio stations, receivers, and/or repeaters. In satellite-based direct-broadcast radio services, digitally-encoded audio program material may be broadcast to terrestrial fixed or mobile digital radio receivers. Fixed receivers may include, for example, stand alone digital radio receivers or digital radio receivers connected via computer networks, including for example, the Internet. Mobile receivers may include, for example, digital radio receivers located in automobiles, aircrafts, watercrafts, and the like.

[0032] Satellite-based digital audio radio services such as SDARS, for example, may be broadcast to one or more digital radio receivers either directly from an orbiting satellite, or indirectly from one or more repeater stations. Such repeater stations may be useful where the digital radio receiver is located in a shielded location or where there is no direct line of sight between the radio and the satellite. In other digital audio radio services systems, the audio programs also may be transmitted in digital form by one or more space stations directly to fixed, mobile, and/or portable radio stations. Such systems may comprise, for example, orbiting satellites, complementary repeating terrestrial transmitters, telemetry, tracking, and control facilities.

[0033] Combinations of mechanical devices U.S. Pat. Nos. 4,019,073, 6,615,968 and atmospheric system interaction are disclosed in U.S. Pat. Nos. 1,119,732, 787,412, 6,902,513 to Nikola Tesla; 28,793 to Charles Vion; and U.S. Pat. No. 1,540,998 to Herman Plauson. Lastly, U.S. Pat. No. 8,102,078 and US2008/0009240. Agnoff discloses an Oscillating watch winder in U.S. Pat. No. 6,543,929, Jennings further discloses an oscillating smart device in application No. 13,572,679.

[0034] As illustrated by a large body of prior art, including the above-noted patents, and a large number of commercial devices, efforts are continuously being made in an attempt to improve helmets, headsets and their methods of fabrication. Nothing in the prior art, however, suggests the present inventive combination of materials and method steps as herein described and claimed. The present invention achieves its purposes, objects and advantages over the prior art through a new, useful and unobvious combination of components and method steps which improve safety, comfort and noise abatement performance.

[0035] Therefore, it is an object of this invention to provide Effectively, the provision of energy such that it meets the needs of the present without compromising the ability of future generations to meet their own needs . . . . Sustainable Energy has two key components: renewable energy and energy efficiency.

[0036] It is still a further objection of this invention to promote Dynamic harmony between equitable availability of energy-intensive goods and services to all people and the preservation of the earth for future generations." And, "the solution will lie in finding sustainable energy sources and more efficient means of converting and utilizing energy.

[0037] It is a further object of the present invention to produce Green Power Energy: is energy that can be extracted, generated, and/or consumed without any significant negative impact to the environment, green power; as electricity produced from solar, wind, geothermal, biogas, biomass, and low-impact small hydroelectric sources.

[0038] Thus, there is a need for a clean energy system that uses atmospheric electricity.

Prior Art

[0039] Quartz crystals have been in regular use for many years to give an accurate frequency for all radio transmitters, radio receivers and computers. Their accuracy comes from an amazing set of coincidences: Quartz—which is silicon dioxide like most sand—is unaffected by most solvents and remains crystalline to hundreds of degrees Fahrenheit. The property that makes it an electronic miracle is the fact that, when compressed or bent, it generates a charge or voltage on its surface. This is a fairly common phenomenon called the Piezoelectric effect. In the same way, if a voltage is applied, quartz will bend or change its shape very slightly.

[0040] If a bell were shaped by grinding a single crystal of quartz, it would ring for minutes after being tapped. Almost no energy is lost in the material. A quartz bell—if shaped in the right direction to the crystalline axis—will have an oscillating voltage on its surface, and the rate of oscillation is unaffected by temperature. If the surface voltage on the crystal is picked off with plated electrodes and amplified by a transistor or integrated circuit, it can be re-applied to the bell to keep it ringing.

[0041] The electronics of the watch initially amplifies noise at the crystal frequency. This builds or regenerates into oscillation—it starts the crystal ringing. The output of the watch crystal oscillator is then converted to pulses suitable for the digital circuits.

Polymers

[0042] Polyvinylidene fluoride (PVDF): PVDF exhibits piezoelectricity several times greater than quartz. Unlike ceramics, where the crystal structure of the material creates the piezoelectric effect, in polymers the intertwined long-chain molecules attract and repel each other when an electric field is applied.

Near Space

[0043] Solar particles become trapped within the Earth's magnetic field and form radiation belts. The Van Allen radiation belt is a torus of energetic charged particles (i.e. a plasma) around Earth, trapped by Earth's magnetic field.

[0044] At elevations above the clouds, atmospheric electricity forms a continuous and distinct element (called the electrosphere) in which the Earth is surrounded. The electrosphere layer (from tens of kilometers above the surface of the earth to the ionosphere) has a high electrical conductivity and is essentially at a constant electric potential. The ionosphere is the inner edge of the magnetosphere and is the part of the atmosphere that is ionized by solar radiation. (Photoionisation is a physical process in which a photon is incident on an atom, ion or molecule, resulting in the ejection of one or more electrons.)

Advantages/Disadvantages

[0045] Energy in electronic elements: Electric potential energy, or electrostatic potential energy, is a potential energy (measured in joules) that results from conservative Coulomb forces and is associated with the configuration of a particular set of point charges within a defined system. The term "electric potential energy" is used to describe the potential energy in systems with time-variant electric fields, while the term "electrostatic potential energy" is used to describe the potential energy in systems with time-invariant electric fields.

[0046] Capacitance is the ability of a body to store an electrical charge. Any body or structure that is capable of being charged, either with static electricity or by an electric current, exhibits capacitance. A common form of energy storage device is a parallel-plate capacitor. In a parallel plate capacitor, capacitance is directly proportional to the surface area of the conductor plates and inversely proportional to the separation distance between the plates. If the charges on the plates are +q and −q, and V gives the voltage between the plates, then the capacitance C is given by

$$C=q/V.$$

[0047] The capacitance is a function only of the physical dimensions (geometry) of the conductors and the permittivity of the dielectric. It is independent of the potential difference between the conductors and the total charge on them.

[0048] Piezoelectricity is the combined effect of the electrical behavior of the material:

$$D=\in E$$

where D is the electric charge density displacement (electric displacement), $\in$ is permittivity and E is electric field strength, and

$$\text{Hooke's Law: } S=sT$$

where S is strain, s is compliance and T is stress.

[0049] Polyvinylidene fluoride, or polyvinylidene difluoride (PVDF) is a highly non-reactive and pure thermoplastic fluoropolymer produced by the polymerization of vinylidene difluoride. PVDF is a specialty plastic material in the fluoropolymer family; it is used generally in applications requiring the highest purity, strength, and resistance to solvents, acids, bases and heat and low smoke generation during a fire event. Compared to other fluoropolymers, it has an easier melt process because of its relatively low melting point of around 177° C.

[0050] It has a low density (1.78) and low cost compared to the other fluoropolymers. It is available as piping products, sheet, tubing, films, plate and an insulator for premium wire. It can be injected, molded or welded and is commonly used in the chemical, semiconductor, medical and defense industries, as well as in lithium ion batteries. It is also available as a crosslinked closed cell foam, used increasingly in aviation and aerospace. PVDF has a glass transition temperature (Tg) of about −35° C. and is typically 50-60% crystalline. To give the material its piezoelectric properties, it is mechanically stretched to orient the molecular chains and then poled under tension. PVDF exists in several forms: alpha (TGTG'), beta (TTTT), and gamma (TTTGTTTG') phases, depending on the chain conformations as trans (T) or gauche (G) linkages. When poled, PVDF is a ferroelectric polymer, exhibiting efficient piezoelectric and pyroelectric properties. These characteristics make it useful in sensor and battery applications. Thin films of PVDF are used in some newer thermal camera sensors.

[0051] Copolymers: Copolymers of PVDF are also used in piezoelectric and electrostrictive applications. One of the most commonly-used copolymers is P(VDF-trifluoroethylene), usually available in ratios of about 50:50 wt % and 65:35 wt % (equivalent to about 56:44 mol % and 70:30 mol %). Another one is P(VDF-tetrafluoroethylene). They improve the piezoelectric response by improving the crystallinity of the material.

[0052] While the copolymers' unit structures are less polar than that of pure PVDF, the copolymers typically have a much higher crystallinity. This results in a larger piezoelectric response: d33 values for P(VDF-TrFE) have been recorded to be as high as −38 pC/N versus −33 pC/N in pure PVDF.

[0053] Applications:

[0054] The piezoelectric properties of PVDF are used to advantage to manufacture tactile sensor arrays, inexpensive strain gauges and lightweight audio transducers. Piezoelectric panels made of PVDF are used on the Venetia Burney Student Dust Counter, a scientific instrument of the New Horizons space probe that measures dust density in the outer solar system. PVDF is the standard binder material used in the production of composite electrodes for lithium ion batteries. 1-2% weight solution of PVDF dissolved in N-Methyl-2-pyrrolidone (NMP) is mixed with an active lithium storage material such as graphite, silicon, tin, LiCoO2, LiMn2O4, or LiFePO4 and a conductive additive such as carbon black or carbon nanofibers. This slurry is cast onto a metallic current collector and the NMP is evaporated to form a composite or paste electrode. PVDF is used because it is chemically inert over the potential range used and does not react with the electrolyte or lithium. Piezoelectric elements can be used in laser mirror alignment, where their ability to move a large mass (the mirror mount) over microscopic distances is exploited to electronically align some laser mirrors. By precisely controlling the distance between mirrors, the laser electronics can accurately maintain optical conditions inside the laser cavity to optimize the beam output. Piezoelectric sensors especially are used with high frequency sound in ultrasonic transducers for medical imaging and also industrial nondestructive testing (NDT).

[0055] For many sensing techniques, the sensor can act as both a sensor and an actuator—often the term transducer is preferred when the device acts in this dual capacity, but most piezo devices have this property of reversibility whether it is used or not. Ultrasonic transducers, for example, can inject ultrasound waves into the body, receive the returned wave, and convert it to an electrical signal (a voltage). Most medical ultrasound transducers are piezoelectric.

[0056] Advantageously,

[0057] Sustainable energy is the sustainable provision of energy that meets the needs of the present without compromising the ability of future generations to meet their needs. Technologies that promote sustainable energy include renewable energy sources, such as hydroelectricity, solar energy, wind energy, wave power, geothermal energy, and tidal power, and also technologies designed to improve energy efficiency.

[0058] This sequence of oscillations causes the rotor within the watch to spin rapidly thereby winding the watch in a manner closely simulating the spinning of the rotor that occurs during normal winding of the watch when the watch is worn by a user. Due to the forces that are exerted, the rotor spins around the watch shaft during the oscillations, as opposed to the partial rotation observed in prior art mechanisms. Therefore, the time required to wind the watch, and the energy required, is substantially reduced. Moreover, since the rotor is spinning about the shaft, as opposed to merely being held in a downward position while the watch is rotated, winding more closely approximating the design mechanism is achieved, thereby putting less wear on the watch.

[0059] This invention relates to satellite communications systems using multiple spot beams from a geosynchronous earth orbit satellite to provide selective coverage of the continental United States and, more particularly, relates to a system having a satellite receiving hub in every spot beam that allows for asynchronous communications between each hub and the satellite for maximizing frequency re-use.

[0060] These purposes, objects and advantages should be construed as merely illustrative of some of the more prominent features and applications of the present invention. Many other beneficial results can be obtained by applying the disclosed invention in a different manner or by modifying the invention within the scope of the disclosure. Accordingly, other purposes, objects and advantages as well as a fuller understanding of the invention may be had by referring to the summary herein mentioned and detailed description describing the preferred embodiments of the invention, in addition to the scope of the invention, as defined by the claims, taken in conjunction with the accompanying drawings.

SUMMARY OF THE INVENTION

[0061] In one general respect, an embodiment of the present invention is directed to a system. The system includes a Power Frequency broadcast station, a receiver, and a network. The Power Frequency broadcast station includes a transmitter and a server. The receiver is in communication with the Power Frequency broadcast transmitter and also includes a user interface for receiving user input commands comprising a request for information from the Power Frequency broadcast station.

[0062] The receiver is configured to establish a two-way communication path between the receiver and the Power Frequency broadcast transmitter. The network is in communication with the server and the receiver for exchanging information therebetween. The request for information is provided to the server via the network and the server is configured to receive the request and transmit a response message to the receiver in accordance with the request.

Continuously Outboard Recharged Electric Vehicle (COREY)

[0063] Given suitable infrastructure, permissions and vehicles, BEVs (battery electric vehicles) can be recharged while the user drives. The BEV establishes contact with an electrified rail, plate or overhead wires on the highway via an attached conducting wheel or other similar mechanism (see Conduit current collection). The BEV's batteries are recharged by this process—on the highway—and can then be used normally on other roads until the battery is discharged. Some of battery-electric locomotives used for maintenance trains on the London Underground are capable of this mode of operation. Power is picked up from the electrified rails where possible, switching to battery power where the electricity supply is disconnected.

[0064] The present invention is directed to overcome the disadvantages of the prior art. The invention is a mechanism which taps into the naturally occurring static electricity in the atmosphere. Whereas heretofore, the attempt to garner electricity from the atmosphere has focused exclusively on capturing lightning, the present invention syphons off the static electricity which is generated from any agitated air and voids lightning.

[0065] Lightning is only the final discharge of the static electricity, whether that lightning is intra-cloud lightning, cloud-to-ground lightning, or inter-cloud lightning. Other types of final discharges are known as heat lightning, summer lightning, sheet lightning, ribbon lightning, silent lightning, ball lightning, bead lightning, elves, jets, and sprites. Well before these discharges are observed, as the atmosphere becomes agitated by wind or thermal, static electricity is being generated. The present invention recognizes that this static electricity is being formed and creates a mechanism to capture it.

[0066] In the preferred embodiment, a sensor array is used to monitor the activities both at the base unit (such as electrical flow within the conductor) and in the surrounding locale. A sensor monitoring the electrical flow (i.e. voltage and/or current) within the conductor is used to monitor the electrical activity within the conductor.

[0067] In the preferred embodiment, a lightning sensor monitors for lightning activity within the locale. As noted earlier, the electrical characteristic of lightning is so extreme that ideally this discharge is avoided as it might damage the mechanism of this invention. The sensor array is utilized by a controller, such as microprocessor, programmed to operate the mechanism as outlined herein.

[0068] The controller operates the winch motor to extend or withdraw the conductive line and by extension the altitude of the balloon. The controller is programmed to operate the winch by monitoring the electrical characteristics of the conductor and adjusting the balloon's altitude to maintain these characteristics within the conductor within a preset range.

[0069] This preset range is established either in the base programming of the controller or is established by an operator of the system. As example, by controlling the amount of current being withdrawn from the atmosphere, the mechanism operates within a safe range and also provides a relatively stable current flow from which a variety of activities can take place (such as DC-AC conversion).

[0070] The controller also utilizes the lightning sensor to protect the mechanism from a lightning strike. Should lightning be detected within a pre-determined range (as established by the software or defined by an operator), then the balloon is pulled down to minimize the risk of damage from a lightning strike.

[0071] Another aspect of the invention relates to the electrical system which accepts the fluctuating atmospheric charge and changes it into an acceptable configuration for either the desired load or for the existing power grid.

[0072] Power grids in the United States operate with a frequency of 60 hertz in an alternating current arrangement. While this basic configuration seems to be universally accepted, the voltage within the grid varies dramatically, such as 15 kv, 34 kv, 69 kv, and even 112 kv.

[0073] Each atmospheric generator is placed proximate to or within easy access to a specific grid; this establishes the required electrical output configuration (i.e. that which is accepted by the power grid). As example, one of the atmospheric electrical collector units as described above collects the atmospheric electrical power as direct current and then supplies the appropriate power grid a specific flow (as example, AC, 60 hertz, at 69 kv).

[0074] The difficulty lies the fact that the DC current being garnered from the atmosphere varies depending on the actual agitation being generated in the atmosphere. This means that the source of DC current is fluctuating.

[0075] The present invention uses a monitoring system which checks the input DC voltage. Depending on the actual voltage being received, the appropriate converter is connected to the input DC voltage so that the desired output is obtained.

[0076] As example, suppose the DC input voltage is 1500 volts, the monitoring system, sensing this input, closes the switch connecting the DC voltage to a converter which accepts DC voltage in the range of 1000-2000 volts which then delivers an AC, 60 hertz 69 kv signal to the power grid. If the DC input voltage increases to 2100 volts, then the monitoring system opens the switch to the first converter (1000-2000 volts) and closes the switch to a second converter (such as 2000-4000 volts) to deliver the desired output of AC (60 hertz, 69 kv) for the power grid.

[0077] In this manner, regardless of the fluctuating input DC voltage, the electrical grid is supplied with a fully configured electrical input conforming to the needs of that specific electrical grid.

[0078] Another aspect of the present invention is the use of a tower or permanent structure instead of an aircraft. In this embodiment, the building or tower is electrically isolated from the ground and a rod (similar to a lightning rod) is extended into the atmosphere. The rod collects the atmospheric charge which is conveyed via an electrical conduit (ideally insulated) where the collected DC charge is reconfigured to meets the need of the locale.

[0079] In this context, for one embodiment of the invention, a tower is placed onto the top of a building. The tower is electrically isolated from the building using such mechanisms well known to those of ordinary skill in the art such as rubber mats. A rod ideally extends from the top of the tower to facilitate the collection of the DC electrical energy.

[0080] Piezoelectricity is a material property that is manifested when voltage is produced by applying mechanical forces, and vice versa, the effect has been described as direct and converse. Piezoelectricity has been described as coupling between a quasi-static electric field and dynamic mechanical motion.

[0081] A piezoelectricity converter mechanism such as described above, is connected to the tower to flow the DC electricity to a converter which modifies the DC current for the specific application. In one application, the DC current is converted to the electrical needs of the building, thereby providing at least some of the electrical requirements of the building itself.

[0082] As noted earlier, the dynamic converter system of the present invention allows a power generator to address a variable voltage in an efficient manner. This makes the dynamic converter system ideal for a variety of alternative energy sources such as the above described atmospheric electrical generator and other alternative energy sources such as wind and wave powered systems. In these systems, the energy being generated must be converted to a proper electrical configuration for a identified load. This may be a particular motor or connection to the power grid which act as a load to the power generating mechanism.

[0083] For these energy generating systems, the converter assembly of this invention utilizes multiple converters. Each converter is configured to accept a unique range of voltages and from these voltages, create the desired electrical configuration. By using multiple converters, a full range is available, from a minimum voltage input to a maximum voltage input.

[0084] The present invention can include systems and methods for integrating sensors for tracking atmospheric transducer system performance metrics into media devices and accessories therefor, thereby reducing or eliminating the need for additional independent monitoring devices. In one embodiment of the present invention.

[0085] It also is known to provide such transducers with connectors to allow their rechargeable batteries to be charged. In some cases, the connector is a Universal Serial Bus (USB) connector, allowing the transducer to be charged by plugging it into the USB port of a computer, grid circuit or other device.

[0086] These and other objects and advantages of the invention will appear more clearly from the following description in which the preferred embodiment of the invention has been set forth in conjunction with the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0087] FIG. **1** Prior Art depicts perspective view of Electric currents created in sunward ionosphere;

[0088] FIG. **2** is a box flow chart of the propulsion cycle systems present invention;

[0089] FIG. **3** is an top view of the piezoelectric disk cylinder orb of the present invention;

[0090] FIG. **4** is a longitudinal sectional view showing an embodiment of this invention;

[0091] FIG. **5** is a top view of the (ATS) slip rotor piezoelectric chamber of the present invention;

[0092] FIG. **6** is an alternate top view of the piezoelectric cylinder orb of the present invention;

[0093] FIG. **7-10** elevation views of spine piezo stacks embodiments of the present invention;

[0094] FIG. **11** illustrates one embodiment of a power service (ATS) system architecture;

[0095] FIG. **12** is a block diagram of the ATS charge schematics systems of the present invention;

[0096] FIG. **13** is a block diagram of an ATS charge and recycle schematics of the present invention;

[0097] FIG. **14** is a view of the piezoelectric ball race cylinder of the present invention;

[0098] FIG. **15** is a view of the piezoelectric housing and gear of the present invention;

[0099] FIG. **16** is a view of a multiple piezoelectric ball race cylinder of the present invention;

[0100] FIG. **17** is another embodiment of the piezoelectric stack ball of the present invention;

[0101] FIGS. **18** & **19** are antenna rod transmit device embodiments of the (ATS) present invention;

[0102] FIG. **20** illustrates substation embodiment of the power service (ATS) system architecture;

[0103] FIG. **21** illustrates a home embodiment of the power service (ATS) system architecture;

[0104] FIG. **22** illustrates a window embodiment of the power service (ATS) system architecture;

[0105] FIG. **23** is (ATS) shock device in accordance with an embodiment of the present invention;

[0106] FIG. **24** is (ATS) shock device in accordance with an embodiment of the present invention;

[0107] FIG. **25** is (ATS) motorcycle device embodiment of the present invention;

[0108] FIG. **26** is (ATS) axle device in accordance with an embodiment of the present invention;

[0109] FIG. **27** is (ATS) train device in accordance with an embodiment of the present invention;

[0110] FIG. **28** is (ATS) plane device in accordance with an embodiment of the present invention;

[0111] FIG. **29** is (ATS) boat device in accordance with an embodiment of the present invention;

[0112] FIG. **30** is (ATS) solar device in accordance with an embodiment of the present invention;

[0113] FIG. **31** a (ATS) turbine device in accordance with an embodiment of the present invention;

[0114] FIG. **32** is (ATS) auto body panel device embodiment of the present invention;

[0115] FIG. **33** is (ATS) motorcycle body fairing panel device in of the present invention; and

[0116] FIG. **34** is (ATS) cross section view of body embed panel device of the present invention.

DETAILED DESCRIPTION

[0117] This sequence of oscillations causes the conductive rotor within piezoelectric molded housing device to spin rapidly thereby winding the mechanism in a manner closely simulating the spinning of the conductive rotor that occurs during normal electric activity when the device is activated. Due to the forces that are exerted, the conductive rotor spins around the piezoelectric cylinder device shaft during the oscillations, as opposed to the partial rotation observed in prior art mechanisms. Therefore, the time required to charge the Atmospheric Transduction System (ATS) device, and the energy required, is substantially reduced. Moreover, since the rotor is spinning about the shaft, as opposed to merely being held in a downward position while the ATS device is rotated, recharging more closely approximating the design mechanism is achieved, thereby putting less wear on the ATS chamber device Innovative Piezoelectric housing and ball bearing, coupler and book spine stacks. There relies the notion of negative ground electricity and positive aerial electricity which is in abundance. This invention substantiates land vehicles recycle recharge by reverse oscillation. Aerial vehicles recycle recharge by forward oscillation in accumulation of environmental positive and negative electricity. These aforementioned activities are integral or synchronous with frequency.

[0118] Prior Art FIG. **1**

[0119] Electric currents created in sunward ionosphere. FIG. **2** is a recycle box flow chart **49** of a self propulsion unit consisting of a battery **27**, engine **29** and piezoelectric transducer **33** unit. FIG. **3** is a multiple disk **31** load part **51** about a 360 degree cylinder **45** and shaft **44**. FIG. **4** is a piezoelectric molded device **100** housing **110** containing a ball bearing race **104**, disk stacks capacitor **102** and conductive rotor **46**. FIG. **5** is a piezoelectric chamber **120** configuration including a counterweights **50**, shaft **44**, conductive slip rotor **106**. FIG. **6** is an alternate part **200** chamber cylinder **201** embodiment containing piezoelectric spine disk stack capacitors **102**. FIG. **7-10** are variations of spine **107**,**108**,**109**,**110** case piezoelectric stack capacitor **102** with plate and ball heads. FIG. **11** Various embodiments of the present invention, among others, will now be described with reference to the accompanying drawings. Accordingly, FIG. **11** illustrates an embodiment of a Atmospheric Transduction (ATS) System **300** architecture. The system **300** may include, for example, a satellite broadcast station **318** that transmits signals **333** containing frequency content from a geostationary satellite **312** by way of airplane antenna **326**. In turn, the satellite **312** transmits line-of-sight (LOS) signals **333** to one or more ATS terrestrial frequency power receiver farms **314**. The system **300** also may include one or more terrestrial repeaters **316** which receive and retransmit the satellite signals **333** as repeater signals **323** to facilitate reliable reception in geographic areas where LOS reception from the satellite **312** is obstructed by tall buildings, hills, tunnels, and other similar impediments to the signals **333**. The ATS receivers **314** maybe designed to receive one or more signals **333** from the satellite **312** and/or from the terrestrial repeater transceiver **316**. In operation, such ATS receivers **314** may receive both the satellite signals **333** and the repeater signals **323**. The receivers **314** also may be located in mobile environments **320**, **321**, **322** which include, but are not limited to, land vehicles **321**, **322**, aircraft **320**, watercraft **900**, and handheld devices, among others. The receivers **314** also may be fixed in stationary units for residential use (e.g., home **325**,**750** entertainment, etc.) or commercial **314**, **328**,**360** use (e.g., business **314**, office **700**, security **328**, etc.). The power frequency broadcast station **318** also may be in communication with a grid network **342**. Two-way communication between the ATS receivers **314** and the power frequency broadcast station **318** may occur via the network **342**. Furthermore, information feedback from the power frequency broadcast station **318** may be transmitted to the ATS receiver **314** both by way of the network **342** as well as via the satellite **312**. Information also may be transmitted to the power broadcast station **318** wirelessly via a wireless network **342**,**707** by way of transducer tower **360**.

[0120] Further disclosed in FIG. **12** and FIG. **13** (in block diagrams **400**, **500**) are electrical schematics for handling the static charge from the atmosphere. By maintaining the voltage being collected in a prescribed range, an electrical conversion system is easily designed. While FIGS. **12**, and **13** illustrate some electrical configurations, those of ordinary skill in the art readily recognize a variety of other configurations which will serve the same function.

[0121] Referencing FIG. **12**, Direct Current In (DC IN) **401** is buffered by a gang of capacitors **410** before being communicated to a DC/AC converter **420**. The DC/AC converter **420** converts the direct current into a an alternating current suitable for placement over an existing electrical grid **430** such as normally found from a power-plant. FIG. **12** Also illustrates an electrical arrangement suitable for use in charging a battery **440**. DC IN **401** is buffered by capacitor **410** bank before entering into a step down transformer **435**. Step down transformer **435** reduces the voltage so that the voltage can safely be introduced into battery **440** which is connected to ground **450** at the battery's other pole. In FIG. **12**, DC IN **401** is fed into an adjustable rheostat **460** which is controlled by the controller **465** so that the DC OUT **470** falls within a computer **475** monitored and sensor **480** specified range. FIG. **13** Hypothetically, unused energy may be recycled current **501** and/or recaptured by reversing the oscillated spin rotation of devices **51**, **100**, **120** with the use of a rectifier **502** and Step up transformer **503**, returned to grid **430** capacitor **410**. This theory lends itself to the concept of positive and negative frequency. FIG. **14-17** Self charging propulsion embodiment of the invention where Da Vinci's ball race **510**,**512** is combined with a disk cylinder **45** and **513** stack balls **525**, conductive rotor **46**, **511** tooth gear **515** and piezoelectric molded housing **555**. FIGS. **18** & **19** Improvement structures Franklin's lightening rod **600**,**603** and a molded vibration transducer **601** quasi replicating Vion's tubes and piezoelectric spine stacks capacitor **602** improving Tesla's Atmospheric transmit device. FIG. **20** is an atmospheric receiving building sub station **700** where energy is consumed and excess rendered to the grid **707** by conductive rotor **702** transducer **710**, tower transducer **703**, antenna rod **701** and transducer windows **704**.

[0122] FIG. **12** and FIG. **2** flowchart illustration also includes battery **440**. Battery **440** may provide electrical power to components of ATS devices within FIG. **11**. Charging circuitry may also be provided to charge battery **440** when an external power supply is connected to an ATS device **100**. FIG. **14**-**17** eliminates one or more steps by presenting a self charge retaining transducer **510**-**513** may be configured with an accelerator sensor **480** controller **465** and gears **515**, provide reciprocal power incorporated within piezoelectric molded and ceramic housing **555** along with stack balls **525** and cylinder **545**. This assembly more resembles a motor by characteristics given power with application.

[0123] In operation, as illustrated in FIGS. **2**-**20** and FIGS. **21**-**34**, is periodically energized by movement to rotate Orb in either a clockwise or counterclockwise direction. The length of time or activity is energized, and the length of time between the period when the capacitor **31**, **102** battery is energized, will depend on the particular ATS device design. As the Orb rotates, the outer end of the disk moves along a 360° circular pathway to push against with forward and rearwardly spinning. Upon engagement of the Orb, ATS device disk is rotated until carried to the apex or top of the circular pathway. Upon reaching the apex, the gravitational and vibrational force or counterweight **50** promotes additional oscillation. ATS device movement rapidly rotate on Orb at a rotational speed greater than the speed of rotation of Orb. Counterweight **50** is then carried beyond the bottom or lowest point of the pathway by its momentum to a point near the apex on the opposite side of the pathway. The cycle is repeated through multiple increasing oscillations of the ATS device until counterweight **50** stops at the bottom position, or until once again engages to again move counterweight to the top of its circular pathway.

[0124] FIG. **2** is a flow chart showing generation of energy using a rotor according to one or more of the above-described embodiments. First, battery **27** starts the engine **29** and/or mobile transducer **33** is oscillated. In response to this acceleration, forces are imposed on one or more rotation piezoelectric devices. In response to those forces, the piezoelectric devices output electrical energy, which energy is extracted at a power controller **465**. The power controller **465** sensor **480** then makes this energy available to recharge a capacitor **410** battery **440** and/or to electronic components of the mobile terminal. Although FIG. **2** shows a serial flow of events, it is to be appreciated that the events of blocks **33**, **27** and **29** occur substantially instantaneously upon acceleration of the mobile terminal.

Preferred Alternate Embodiments

[0125] The present invention (ATS) device in accordance with an embodiment of the present invention overcomes the foregoing problem in the conventional art and provides an electro energy vibration and alternative to gas, oil or fossil fuel consumption in FIG. **21** homes **750** transducer **755**, FIG. **28** airplane **880** transducer **801**, FIG. **27** train **860** transducer **801**, FIG. **32** auto **950** transducer **910**, FIG. **25** and FIG. **33** motorcycles **960**,**850** transducers **799**, **802**, **810**, **910** and FIG. **29** boat **900** transducers **901**,**910**.

[0126] In order to solve the foregoing problems in the conventional art, the present invention provides an electro transducer having a ball bearing assembly which is compatible with FIG. **26** axle shaft **44** transducer **810** wheel **870** assembly. In order to solve the foregoing problems in the conventional art, the present invention provides an electro transducer having an FIG. **34** housing panel **980** transducer **910** assembly which is compatible with an exterior body assembly. Thereby, an Atmospheric transducer device may be shock **800** integrated as within the FIG. **23** coupler **802**, and FIG. **24** ladder **799**, (vertical friction) piezoelectric absorbers assemblies. In order to solve the foregoing problems in the conventional art, the present invention provides an electro transducer having a capable flat assembly which is compatible with FIG. **22** window **704** disk **31** plate transducer **714**, FIG. **30** solar panel **920** transducer **910**, FIG. **31** wind turbine **930** transducer **910** assembly. This invention provided FIG. **3** piezoelectric plate disks **31**, FIG. **7**-**10** spine piezoelectric disks stacks **107**-**110** capacitor **102**, FIG. **17** piezoelectric stack balls **525** capacitor, the arrangement of balls **525** will not be in a facing relation. The deformation in the inner and outer races during rotation of the ball bearing while undergoing a radial load is made irregular and complicated. The spine eliminates the deformation and vibration level increased by a combination of ceramic and molded piezoelectric materials at a regular predetermined frequency thereby multiplying the level of vibration and noise reduction.

[0127] Certain modifications and improvements will occur to those skilled in the art upon a reading of the foregoing description. It should be understood that all such modifications and improvements have been deleted herein for the sake of conciseness and readability but are properly within the scope of the following claims.

[0128] Thus it is seen that an atmospheric transducer device may be integrated and/or provided. It will be understood that the foregoing is only illustrative of the principles of the invention, and that various modifications can be made by those skilled in the art without departing from the scope and spirit of the invention, and the present invention is limited only by the claims that follow.

1-**3**. (canceled)

4. An atmospheric transduction system comprising:

means for recycle recharge by oscillation and frequency in accumulation of environmental positive and negative electricity, maintaining the voltage being collected in a prescribed range, providing an electrical conversion broadcast network; and

means for collecting the atmospheric electrical power as direct current and then supplies the appropriate power grid, transceiver and capacitates a charge; and

means for self charging propulsion provides motor characteristics, and frequency engine.

5. The atmospheric transduction system in accordance with claim **1**, wherein said means for recycle recharge by oscillation and frequency in accumulation of environmental positive and negative electricity, maintaining the voltage being collected in a prescribed range, an electrical conversion broadcast network comprises piezoelectric transducer molded device(s), rotor, stack ball bearing, coupler, book spine stacks and antenna rod stacks;

wherein said means for collecting the atmospheric electrical power as direct current and then supplies the appropriate power grid and capacitates a charge comprises a piezoelectric network, piezoelectric grid, piezoelectric spine stack antenna, piezoelectric transmitters, piezoelectric receivers, piezoelectric devices, piezoelectric cylinders and orbs, power frequency broadcast;

wherein said means for charging providing motor characteristics, said power frequency engine comprises a sensor accelerator, rotor/gear, battery/capacitor, antenna, controller, for propulsion.

6. An atmospheric transduction system comprising:

a piezoelectric transducer molded device(s), rotor, spine ball bearing, coupler and book spine stack transducers, for recycle recharge by oscillation and frequency in accumulation of environmental positive and negative electricity, maintaining the voltage being collected in a prescribed range, providing an electrical conversion broadcast server network; and

a piezoelectric network, piezoelectric grid, piezoelectric spine stack antenna, piezoelectric transmitters, piezoelectric receivers, piezoelectric devices, piezoelectric cylinders and disc orbs, stack rods, power frequency broadcast, for collecting the atmospheric electrical power as direct current and then supplies the appropriate power grid, transceiver and capacitates a charge; and

a sensor accelerator, rotor/gear, battery/capacitor, controller, antenna, means for self charge propulsion providing motor characteristics, thereby an engine of power frequency.

* * * * *

(No Model.)

M. W. DEWEY.

METHOD OF UTILIZING NATURAL ELECTRIC ENERGY.

No. 414,943. Patented Nov. 12, 1889.

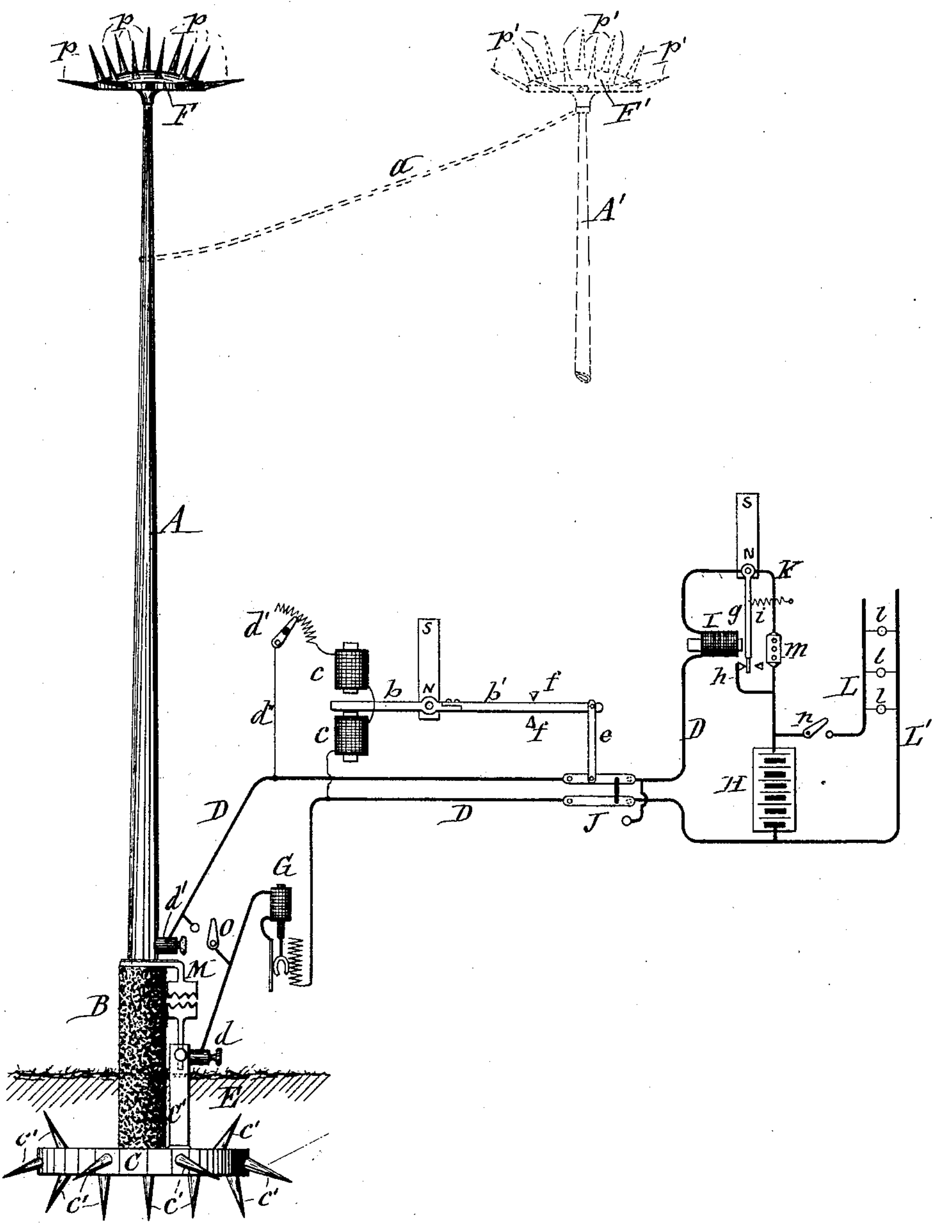

p
p
p
F
p'
p'
p'
F'
a
A'
A
d'
c
c
b
N
S
b'
f
f
e
d
D
D
J
S
N
K
I
g
i
m
h
l
l
l
L
n
L'
D
H
G
d'
o
M
B
d
E
c'
c'
C
c'
c'
c'
c'

WITNESSES:

C. L. Bendixon

A. F. Walz.

Mark W. Dewey INVENTOR

BY

Duell, Laass & Duell

ATTORNEYS

N. PETERS, Photo-Lithographer, Washington, D. C.

UNITED STATES PATENT OFFICE.

MARK W. DEWEY, OF SYRACUSE, NEW YORK, ASSIGNOR TO THE DEWEY CORPORATION, OF SAME PLACE.

METHOD OF UTILIZING NATURAL ELECTRIC ENERGY.

SPECIFICATION forming part of Letters Patent No. 414,943, dated November 12, 1889.

Application filed May 1, 1889. Serial No. 309,171. (No model.)

To all whom it may concern:

Be it known that I, MARK W. DEWEY, of Syracuse, in the county of Onondaga, in the State of New York, have invented new and useful Improvements in the Method of Utilizing Natural Electric Energy, of which the following, taken in connection with the accompanying drawings, is a full, clear, and exact description.

The object of this invention is to provide a method whereby natural electrical energy—such as the so-called "atmospheric electrical energy," or electrical energy that may be derived from the difference of potential existing between two points, one being the earth and the other the atmosphere at an elevation above the earth—may be collected or utilized for the various uses to which electricity is applied.

It has been found that the presence of electricity in the upper regions of the atmosphere is not confined to thunder-clouds, but can be detected at all times and often in great quantities in different conditions of the atmosphere. In fine weather this electricity is mostly positive; but in showery or stormy weather negative electricity is as frequently met with as positive, and it is in such weather that the indications of electricity, whether positive or negative, are usually the strongest. It has also been found that as we proceed farther from the earth's surface, whether upward from a level plane thereof or horizontally from an elevation, the potential of points in the air becomes more and more different from that of the earth, the difference being, in a broad sense, simply proportional to the distance; hence we can infer that there is electricity residing on the surface of the earth, the density of which at any moment in the locality of observation is measured by the difference of potential found to exist between the earth and a given point in the air near it. The results of observations show that the variations of the electricity residing in the atmosphere is the main cause of the variations of the electricity on the surface of the earth. A charged cloud or body of air induces electricity of the opposite kind to its own on the parts of the earth's surface over which it passes and produces such variations. The difference of potential in increasing the distance from the earth is due to electricity induced on the surface of the earth by opposite electricity in the air overhead, and the air being a non-conductor the electricities are unable to combine. As electrical density is greater on projecting parts of a surface than on those which are plane or concave, stronger indications are obtained on hills than in valleys, if the collecting apparatus be at the same distance from the ground in both cases. The average difference of potential is greater in the winter than in the summer. Little or no effects can be obtained within inclosures or under trees, as they tend to screen the apparatus.

Inasmuch as electricity travels in preference through the best conductors, it follows that if a path of low resistance is formed (such as the erection of a metal pole) to a sufficient elevation above the earth the electricities in the atmosphere and that on the surface of the earth will tend to combine and travel through the said path in a current or currents, and if this pole terminates in a metal point or a number of such points the earth and clouds exchange their opposite electricities without a disruptive discharge—as the lightning—but in a slow and gradual way through convection. Besides supplying the top of the pole with points, a metal plate to which the said points are fixed attracts the opposite electricity in the atmosphere. In order to obtain a greater and increased effect, a large metal plate is buried at the foot of the pole and electrically connected therewith, and is provided with points or branches extending in different directions in the ground. The plate and branches may be surrounded by metallic refuse, coke, or other good conducting substance. The metallic points on the top of the pole should be sharp, and preferably of copper, and may be platinized, gilded, or galvanized to prevent corrosion. It having been ascertained by practical experiments that either a flame or dropping water at an elevation above the surface of the earth produces convection of electricity, it is obvious that such means may be employed in place of the points hereinbefore referred to. As the electricity in the atmosphere is sometimes positive and other times negative, the direction of the currents is not always the same—that is, the atmospheric electrical energy is composed of a current of an alternating character, flowing in one direction on an average about as much of the time as in the opposite direction, but the length of the current in a certain direction, or the lengths of time between reversals or changes in the direction of its flow, is greatly varied. This and also the varied strength of the currents have prevented the utilization of atmospheric electrical energy in commercial quantities for the various purposes for which electricity is generally employed. In order to utilize such a current or currents, they should be transformed into a continuous direct current of uniform strength. The apparatus for accomplishing the transformation of atmospheric electricity into a direct current of uniform strength is susceptible of being greatly modified. The preferred form of apparatus, however, for carrying the invention practically into effect I will now proceed to describe to show that the method is capable of actual performance. Said apparatus is illustrated in the diagram accompanying this specification.

Referring specifically to said diagram, A represents a metallic pole, which is shown bare, but may be enveloped in suitable insulating material, if desirable. B is the base of the aforesaid pole, which base is of insulating material set in the ground E. C is a large metal plate beneath the said base, and has points or branches *c′* extending therefrom in different directions in the ground. C′ is a metallic post extending from the said plate above the surface of the ground and having a terminal *d* of the circuit D. On the top of said pole is mounted a metallic cap F, consisting, preferably, of a convex disk provided with sharp iron or copper points *p*, which project in all directions from the same. As before mentioned, the said points may be plated with a suitable metal that forms a good conductor and prevents corrosion. Similar caps F′ may be placed on other insulated poles, as A′, in the vicinity, and connected with the main pole A by an electric conductor *a*, for increasing the effect. The pole A′ may be of wood, and the pole A may also be of the same material if provided with a metal conductor within or on the outside, extending from the cap to the other terminal *d′* of the circuit D. The said circuit D leads from the terminal *d* through an automatic variable resistance G, thence to one of the poles of a secondary or storage battery H, and from the other pole of said battery through an automatic current-regulator to the terminal *d′*. An automatic current-reverser or pole-changer J is located in the said circuit for reversing the current whenever there is a change in its direction, so that it may be rectified or straightened during transit and caused to travel at all times whether its direction is toward or from the earth in one and the same direction through a portion of the circuit containing the secondary battery. The reversals are accomplished automatically by means of a pivoted polarized armature *b*, located between two electro-magnets *c c*, having their coils included in a shunt-circuit between the leads of the circuit D. The magnets are wound so that a north pole will be presented to the armature on one side and a south pole on the other. When the current is flowing in a certain direction, the said polarized armature will be repelled by one magnet and attracted by the other, and thereby moved to one side. When the current changes its direction through the magnets, the poles of the said magnets are reversed and the armature is both repelled and attracted to the other side. An arm *b′*, of diamagnetic material, is fixed to and extends from the armature and is moved by the same. Between a movable end of the arm and the said current-reverser is a pivoted connection or link *e*, by which the motion of the arm is conveyed to the reverser. Stops *f* are provided for limiting the movement of the said arm, and an adjustable resistance *d′* is included in the shunt *d″* to regulate the current through the same.

The current-regulator hereinbefore referred to prevents short-circuiting or the rapid discharge of the secondary battery into the air and ground when the strength of said battery-current becomes greater than that passing to the battery. Said regulator is composed of an electro-magnet I, having its coil in the circuit D. A pivoted polarized armature *g* is connected at its pivot to one terminal of the coil of said magnet, and when the current is flowing to the battery said armature is attracted by the magnet I and held in contact with the stop *h*, to which the terminal of the secondary battery is connected. In the aforesaid condition a free or low-resistance path for the current is provided to the battery; but when the battery-current exceeds the charging-current the magnet-poles are reversed and the armature is repelled by the magnet, and the free path of the circuit is broken between the armature *g* and stop *h*. In order to maintain the armature in the latter position until the charging-current has been increased in strength above that of the battery-current or discharging-current, and so that the said regulator will automatically operate, a shunt path K of high resistance is provided around the armature *g* and stop *h*. The high resistance of the said shunt is obtained by including a rheostat *m*. This shunt path or circuit K, with the resistance, permits a small but sufficient amount of current to flow through the magnet I to hold the armature away from the stop *h* until the current is reversed, and then move the armature back to stop *h* to close the free or low-resistance path. A spring *i* is provided to assist the movement of the armature from the magnet when it is repelled by the same.

The electric current may be directly conducted to translating devices—such as lamps or electromotors; but, as hereinbefore stated, said current is preferably employed to charge one or more cells of the secondary battery H, and this battery stores or accumulates the electrical energy and supplies the said translating devices. To illustrate the latter feature, leads or wires L and L′ are extended from the poles or electrodes of said battery, and translating devices, in the shape of incandescent lamps *l*, are connected with the said wires in multiple arc. In the wire L is a common circuit maker and breaker *n* to close and open the circuit to the lamps. The said battery may be charged in series or parallel.

The automatic variable resistance G maintains the current flowing through the circuit D approximately uniform by increasing the resistance therein upon an increase of strength. Said resistance is not absolutely necessary and may be dispensed with.

The lightning-arrester M is to short-circuit a very heavy current to prevent the same from passing through the other parts of the apparatus and injuring it. A low-resistance shunt O, with a circuit maker and breaker therein, is connected between the leads of the circuit D, near the terminals *d* and *d′*, to completely short-circuit the apparatus when desired.

Having described my invention, what I claim as new, and desire to secure by Letters Patent, is—

1. The method of utilizing atmospheric electrical energy, consisting in conducting the electric current or currents between the earth and a point in the atmosphere at an elevation above the earth, rectifying or straightening the currents during transit, and storing or accumulating the electrical energy of said rectified current or currents.

2. As preliminary steps in the method of utilizing atmospheric electrical energy, conducting the electric current or currents through a path of low resistance between the earth and the atmosphere at an elevation above the earth, maintaining an approximately uniform strength of current and rectifying or straightening the same during transit.

3. As preliminary steps in the method of utilizing atmospheric electrical energy, conducting the electric current or currents through a path of low resistance between the earth and the atmosphere at an elevation above the earth, and rectifying or straightening said currents during transit.

4. The method of utilizing atmospheric electrical energy or deriving energy from the difference of electrical potential existing between the earth and a point or points in the atmosphere at an elevation above the earth, consisting in conducting the current or currents between the two points through a path of low resistance, rectifying or straightening the said currents in a portion of the path during transit, and storing or accumulating the electrical energy thereof.

5. The method of utilizing atmospheric electrical energy or deriving energy from the difference of electrical potential existing between the earth and a point or points in the atmosphere at an elevation above the earth, consisting in conducting the current or currents between the two points through a path of low resistance, maintaining an approximately uniform strength of current, rectifying or straightening the said currents in a portion of the path during transit, and storing or accumulating the electrical energy.

6. The method of utilizing atmospheric electrical energy composed of a current of a varied alternating character, consisting in conducting the electric current or currents through a path of low resistance between the earth and the atmosphere at an elevation above the earth, and then accumulating the electrical energy of said current or currents while flowing in both directions in one or more cells of a secondary or storage battery, as described.

In testimony whereof I have hereunto signed my name this 29th day of April, 1889.

MARK W. DEWEY. [L. S.]

Witnesses:
C. H. DUELL,
C. L. BENDIXON.

W. I. PENNOCK.

APPARATUS FOR COLLECTING ELECTRICAL ENERGY.

APPLICATION FILED JAN. 4, 1911.

1,014,719. Patented Jan. 16, 1912.

2 SHEETS—SHEET 1.

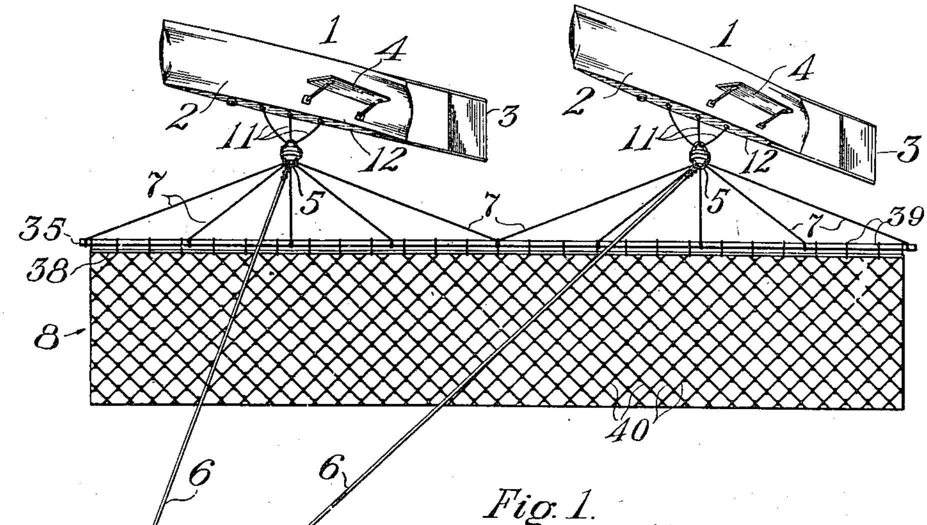
1
4
2
3
11
12
5
7
35
39
38
8
40
6

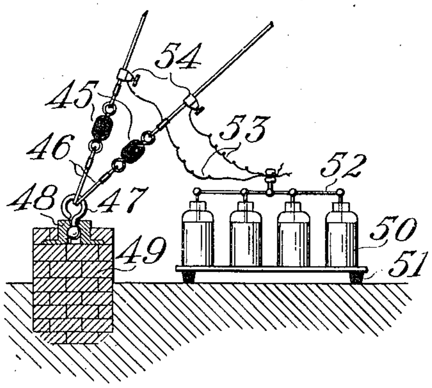
54
45
53
46
52
48
47
50
49
51

Fig. 1.

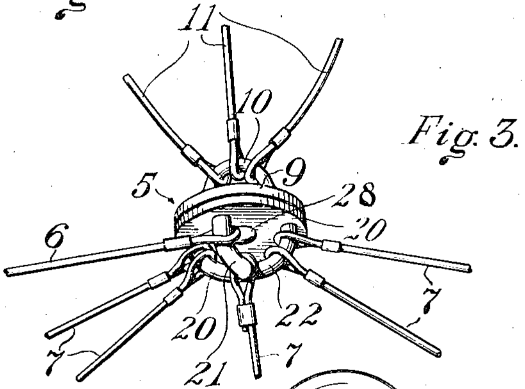
11
10
9
5
28
20
6
7
22
20
21

Fig. 3.

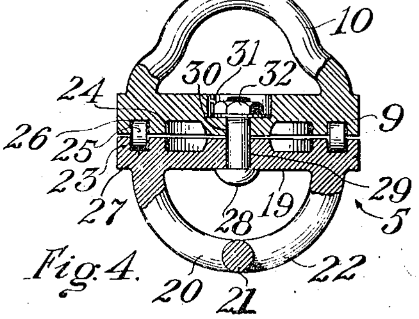
10
31
30
32
24
9
26
25
23
29
27
19
28
5
22
20
21

Fig. 4.

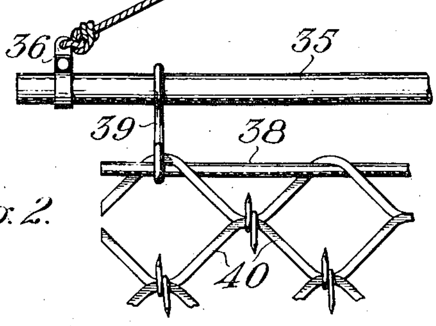
7
36
35
39
38
40

Fig. 2.

INVENTOR

Walter I. Pennock.

WITNESSES

Elizabeth L. Smith

J. Stuart Freeman.

BY

ATTORNEY

W. I. PENNOCK.

APPARATUS FOR COLLECTING ELECTRICAL ENERGY.

APPLICATION FILED JAN. 4, 1911.

1,014,719. Patented Jan. 16, 1912.

2 SHEETS—SHEET 2.

Fig. 5.

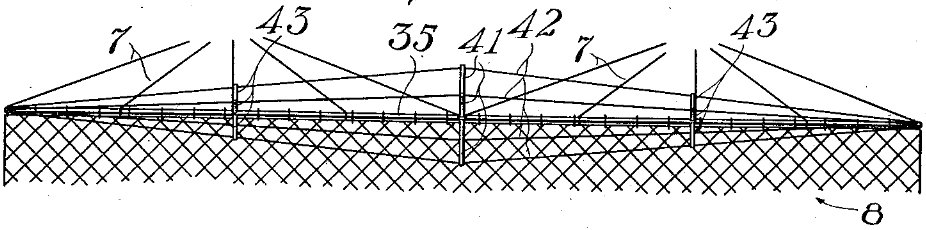
7
43
35
41
42
7
43
8

Fig. 6.

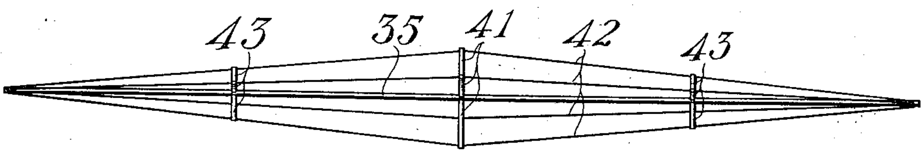
43
35
41
42
43

Fig. 7.

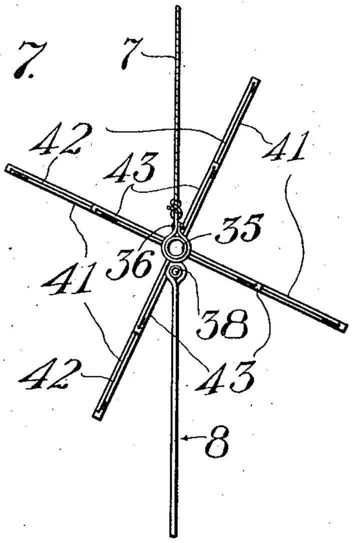
7
42
43
41
35
36
41
38
42
43
8

INVENTOR

Walter I. Pennock.

WITNESSES

Elizabeth L. Smith,

J. Stuart Freeman.

BY

ATTORNEY

UNITED STATES PATENT OFFICE.

WALTER I. PENNOCK, OF PHILADELPHIA, PENNSYLVANIA.

APPARATUS FOR COLLECTING ELECTRICAL ENERGY.

1,014,719. Specification of Letters Patent. Patented Jan. 16, 1912.

Application filed January 4, 1911. Serial No. 600,777.

To all whom it may concern:

Be it known that I, Walter I. Pennock, a citizen of the United States, residing at Philadelphia, county of Philadelphia, and State of Pennsylvania, have invented certain new and useful Improvements in Apparatus for Collecting Electrical Energy, of which the following is a full, clear, and exact disclosure.

The present invention relates to an improved means for collecting the charges of electricity from the upper atmosphere and more particularly to that form which consists in one or more captive balloons from which is suspended a suitable form of metallic conductor.

The principal objects of the device are: to provide a collector for atmospheric and static electricity, which when in operative position will present a large surface to currents of air, but which will offer comparatively little resistance thereto, to provide a collector of such material and construction as will be more efficient in its operation than any previously constructed for a similar purpose, to provide a means for maintaining such a collector suspended in the air and at right angles to opposing currents thereof, and to provide a suitable anchorage for holding said means captive.

With these principal objects in view, the present invention consists in further advantages which are brought out in the following specification and accompanying drawings, in both of which like numerals refer to like parts, and in which drawings—

Figure 1 is a perspective view of the complete device in operative position, Fig. 2 is an enlarged detail view of the wire mesh and the manner in which it is attached to the supporting balloons, Fig. 3 is a detail of the manner of securing the collector-supporting and anchor cables to the balloons, Fig. 4 is an enlarged cross section of the swivel connection shown in Fig. 3, Fig. 5 is top plan view of the reinforcing braces on the screen, Fig. 6 is an elevation of the same and Fig. 7 is an end view of the structure shown in Figs. 5 and 6.

Referring to the drawings, in Fig. 1 thereof, a plurality of balloons 1 of any suitable type is shown, each of which embodies hollow metallic elongated gas tanks 2, extending from the rear of which are single, rigidly affixed rudders 3, while on the sides of the tanks are secured stationary lifting planes 4.

To the bottom and slightly to the rear of the center of the tanks 2 is secured a suitable swivel 5, by which the anchor ropes 6 and the suspension ropes 7 for the metallic conductor 8 are secured to the balloons 1. A suitable form of swivel joint is illustrated in Figs. 3 and 4, but any type can be used that embodies the essential features shown therein.

The swivel joint illustrated consists in the base plate 9 having a looped portion 10 integral therewith and projecting from the upper face thereof. Secured to the loop 10 is a set of three light electrically conductive supporting ropes or cables 11 which extend upwardly and are secured at intervals to the bottom 12 of the balloon above.

The lower or revoluble member 19 of the swivel joint preferably comprises three upwardly directed curved arms, 20, 21, and 22, respectively, forming at their junction a T-shape as shown, said arms at their upper extremities being integral with the plate 23. The member 19 is revoluble below and concentric with the plate 9, and the two members are lightened in weight by opposed concentric grooves as shown at 24. Contact between said members is made through the roller bearings 25, which are carried in the opposed concentric grooves 26 and 27 of the respective upper and lower plates. Furthermore, said plates are maintained in coöperative relation with each other by means of a bolt 28 passing through centrally drilled holes 29 and 30 in the respective lower and upper plates, the drilled hole 30 opening upwardly into an enlarged recess 31, in which is sunken the nut 32 on the bolt 28. Two of the arms 20 and 22 of the lower revoluble member 19 extend in diametrical alinement, while from the central point thereof extends the third arm 21 at right angles thereto, and upward to the plate 23.

Suspended below the plurality of balloons is a hollow rod 35, of any suitable material, connected at regular intervals such as at points 36, by metallic ropes 7, to the alined arms 20 and 22 of the swivel joint 5. Below and parallel to the rod 35 is a similar but smaller rod 38 suspended therefrom by means of suitable couplings 39. From the rod 38 hangs a wire mesh 40 of any suitable form, such as can be extended over a considerable area and which will offer comparatively little resistance to the passage of air therethrough. The metal in such a screen is preferably rough, sharp or jagged and a convenient form embodying these characteristics is expanded sheet metal such as is used for lathing and for reinforcing concrete construction. Instead of the single hollow rod 35 alone, applicant contemplates the use of reinforcing means to prevent the buckling of the rod in a stiff wind, such, for instance, as the arrangement of wire bracing shown in Figs. 5, 6 and 7 in which cross rods 41, secured to the rod 35 at the center thereof are attached to the ends thereof by means of stretched wires 42. Between the rods 41 and the ends of the rod 35 are secured sets of cross rods 43 which support the wires 42 and strengthen the rod 35 at as many points as they may be placed.

For the purpose of maintaining the balloons and apparatus suspended therefrom captive, the light metallic cables 6 terminating downwardly in insulators 45 are employed. These insulators in turn are connected by means of short sections of rope 46 to the eye 47 of a suitable swivel 48, embedded in the anchorage 49.

When the balloons with the metallic screen suspended therefrom are allowed to rise into one of the higher altitudes, the entire apparatus being of metallic construction and uninsulated will become energized by contact with the surrounding natural charges of electricity. From the above description it is evident then that, while the screen 8, on account of its great extent, will be the greatest collecting agent, it will be seen that the balloons themselves and the suspending wires will also coöperate as one large collector, since no parts of which are insulated from any of the neighboring parts thereof. Consequently, when the apparatus described has reached an altitude or strata of the atmosphere abounding in static charges of electricity, an amount of the said charges proportionate to the surface area of the metal exposed will collect upon the apparatus as a whole and will be conducted downwardly toward the earth by means of the various anchor ropes 6, but will not pass into the ground on account of the interposition of the insulators 45.

To use the electrical charges thus acquired, a plurality of Leyden jars 50, or other suitable collectors are supported above the surface of the earth and insulated therefrom by any suitable means as represented by the blocks 51. Either the inner or outer conducting surfaces of the jars may be connected together and energized by the accumulated charge. In the present instance, the inner surfaces of the accumulators are shown to be connected, and the connecting means 52 is in turn connected to the ropes 6 by means of wires 53. These wires are secured to the ropes mentioned by means of suitable binding posts 54.

In the device described the anchor ropes 6 are of substantially the same length, and when the apparatus is raised to the desired altitude and is being blown by the currents of air, the balloons are turned by means of the vanes 3 to parallel relation with each other, and furthermore, from the manner in which the device is held captive and the collecting net 40 is suspended from the balloons, it is obvious that said net will at all times readily swing into a plane substantially perpendicular to any current of air acting upon the balloons above. The purpose of the swivel joint shown in Figs. 3 and 4 is principally for allowing the balloons to readily aline themselves with any new direction of the wind before the apparatus, including the suspended screen, can swing about the swivel 48, and said joint will also prevent the twisting of the ropes 7, when any rapid shifting of the air currents may occur.

While applicant has shown a set of Leyden jars as the accumulators in the accompanying drawings, it is obvious that any other suitable form may be used, and furthermore, that although not illustrated, any suitable apparatus may be run thereby, such as for instance, wireless telegraphic instruments.

Furthermore, although but one embodiment of the invention has been described, it is to be understood that various modifications may be made therein, and in fact several are contemplated by applicant that are of such structure as fall well within the scope of the appended claim.

Having thus described my invention, what I claim and desire to protect by Letters Patent of the United States, is:

A collector for charges of electricity, comprising a plurality of supporting means, a metallic gauze sustained thereby to lie in a plane and substantially equidistant from each of said supporting means, anchoring means emanating from a common point to each of said supporting means, means operative to maintain said supporting means in their normal positions, and means between each of said supporting means and said gauze to permit each of said supporting means to readily and independently aline itself to accord with any alteration in the direction of opposing air currents.

In witness whereof I have hereunto set my hand this 28 day of December, A. D. 1910.

WALTER I. PENNOCK.

Witnesses:
Mildred S. Temple,
E. Eugenia Pennock.

US007439712B2

(12) **United States Patent**
McCowen

(10) **Patent No.:** **US 7,439,712 B2**
(45) **Date of Patent:** **Oct. 21, 2008**

(54) **ENERGY COLLECTION**

(76) Inventor: **Clint McCowen**, 1902 Rue La Fontaine, Navarre, FL (US) 32566

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 509 days.

(21) Appl. No.: **11/358,264**

(22) Filed: **Feb. 21, 2006**

(65) **Prior Publication Data**

US 2007/0195481 A1 Aug. 23, 2007

(51) **Int. Cl.**
H02N 1/00 (2006.01)

(52) **U.S. Cl.** .... **322/2 A**; 307/149; 361/225; 361/117; 174/2; 320/101

(58) **Field of Classification Search** .... 361/220, 361/117, 212, 233, 221, 222, 225; 322/2 A, 322/2 R; 307/149; 174/2, 3; 320/101
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

674,427 A * 5/1901 Palencsar 307/149
911,260 A * 2/1909 Pennock 307/149
1,014,719 A * 1/1912 Pennock 307/149
4,104,696 A 8/1978 Cochran, Jr.
5,047,892 A 9/1991 Sakata et al.
6,735,830 B1 5/2004 Merciel
6,974,110 B2 12/2005 Grandics
2003/0107869 A1 6/2003 Pan et al.
2004/0160711 A1 8/2004 Stumberger
2004/0212945 A1 10/2004 Sprenger et al.

* cited by examiner

Primary Examiner—Mark T Le
(74) *Attorney, Agent, or Firm*—Thomas, Kayden, Horstemeyer & Risley, LLP

(57) **ABSTRACT**

An energy collection system may collect and use the energy generated by an electric field. Collection fibers are suspended from a support wire system supported by poles. The support wire system is electrically connected to a load by a connecting wire. The collection fibers may be made of any conducting material, but carbon and graphite are preferred. Diodes may be used to restrict the backflow or loss of energy.

28 Claims, 10 Drawing Sheets

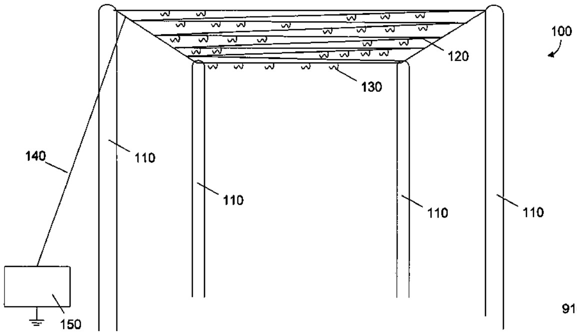
100
120
130
140
110
110
110
110
150

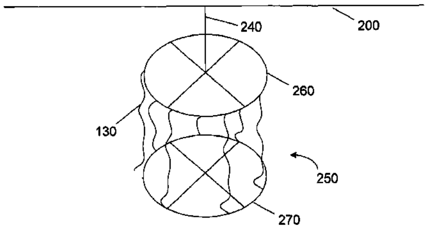
240
200
260
130
250
270

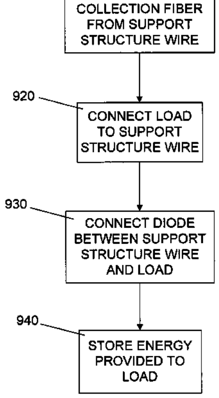
910
SUSPEND COLLECTION FIBER FROM SUPPORT STRUCTURE WIRE
920
CONNECT LOAD TO SUPPORT STRUCTURE WIRE
930
CONNECT DIODE BETWEEN SUPPORT STRUCTURE WIRE AND LOAD
940
STORE ENERGY PROVIDED TO LOAD

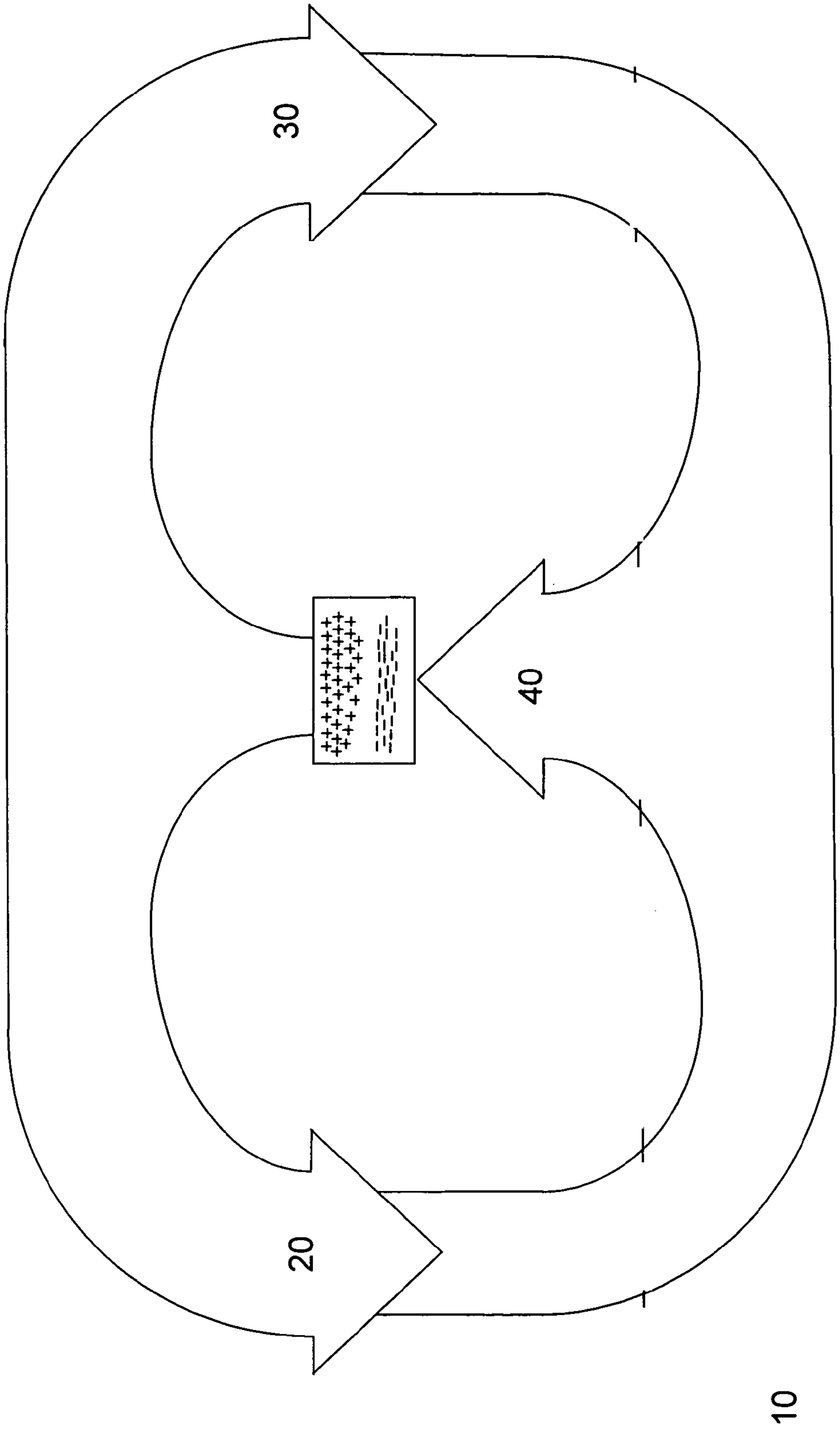

FIGURE 1

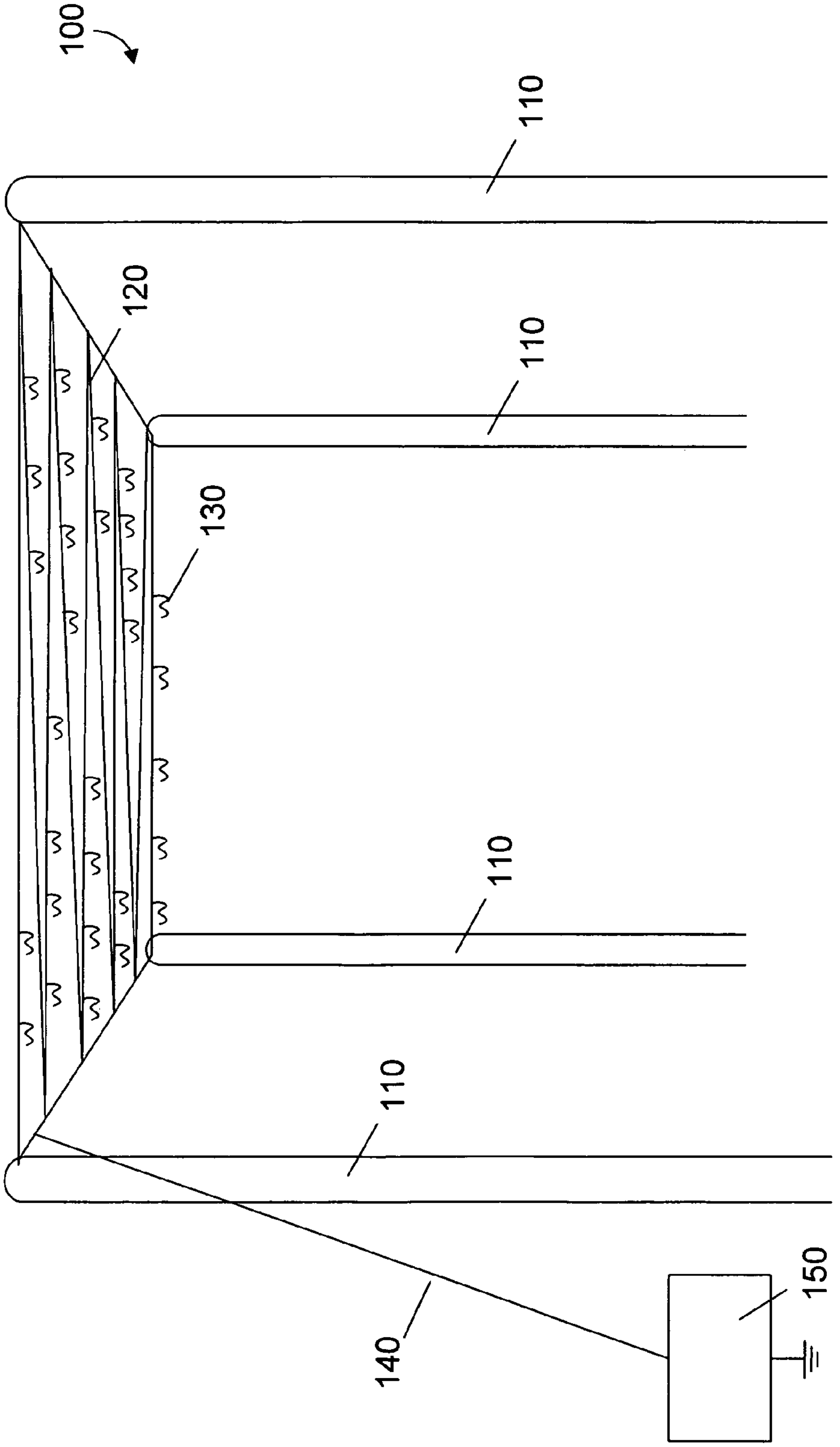
100
110
110
120
130
110
110
140
150

FIGURE 2

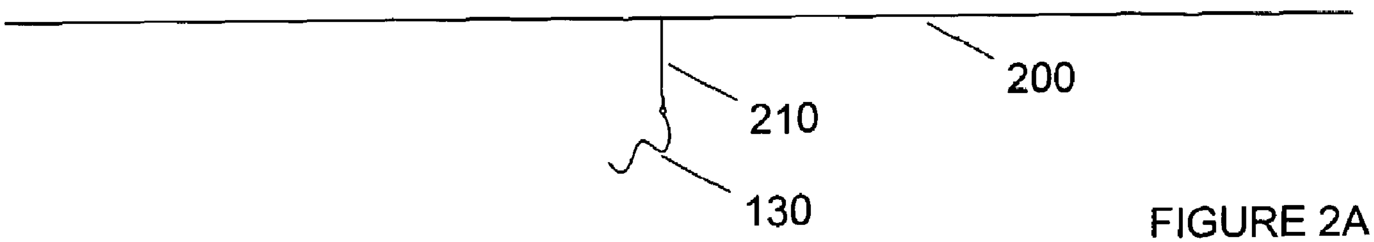

FIGURE 2A

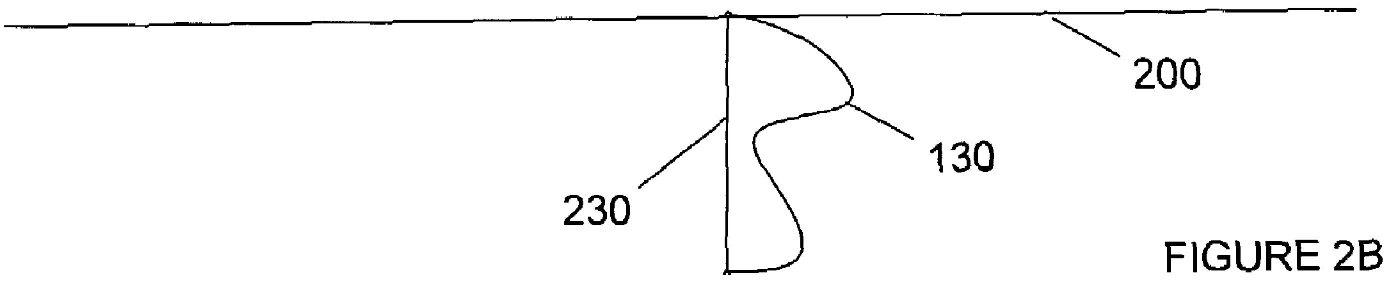

FIGURE 2B

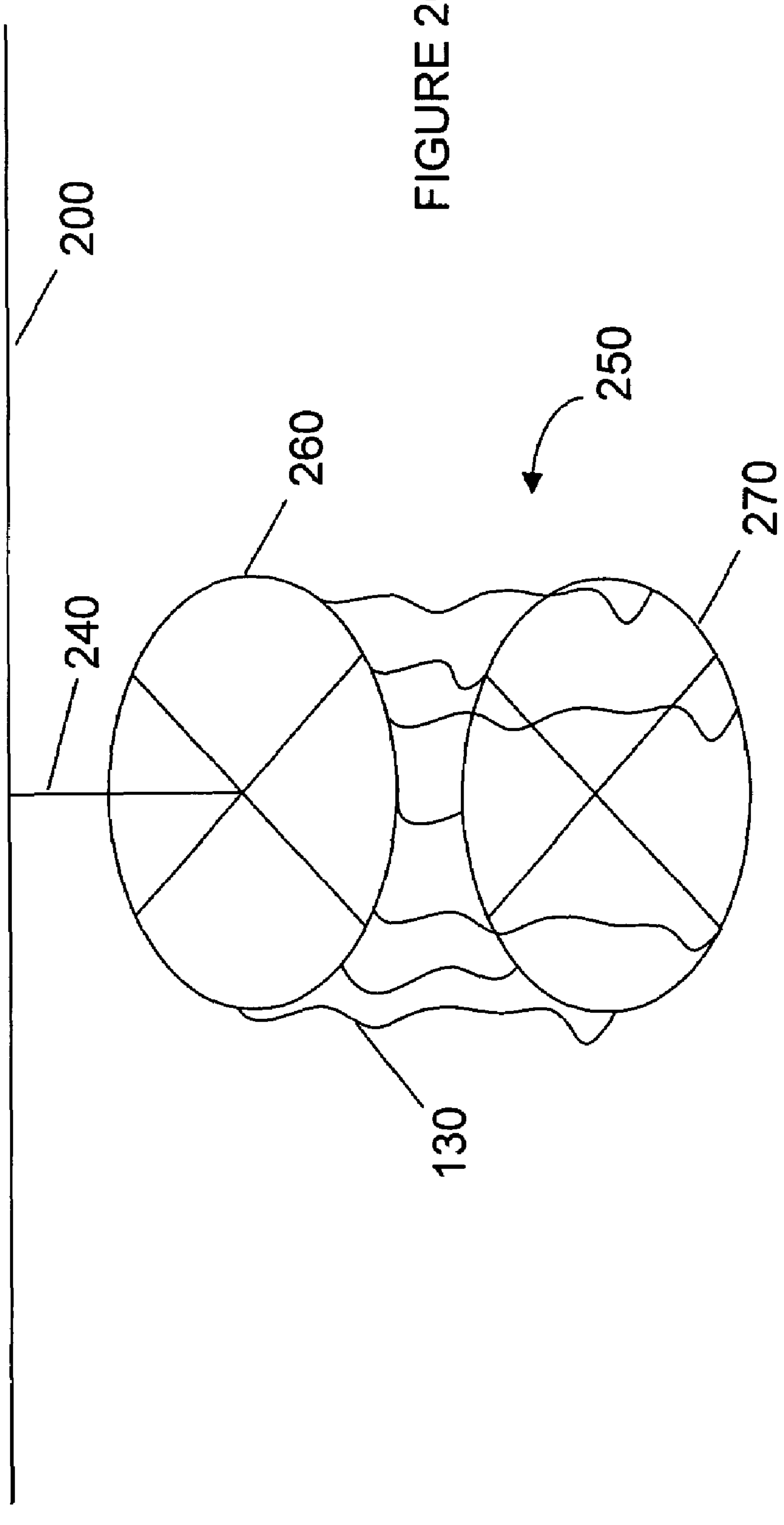
200
250
260
270
240
130

FIGURE 2C

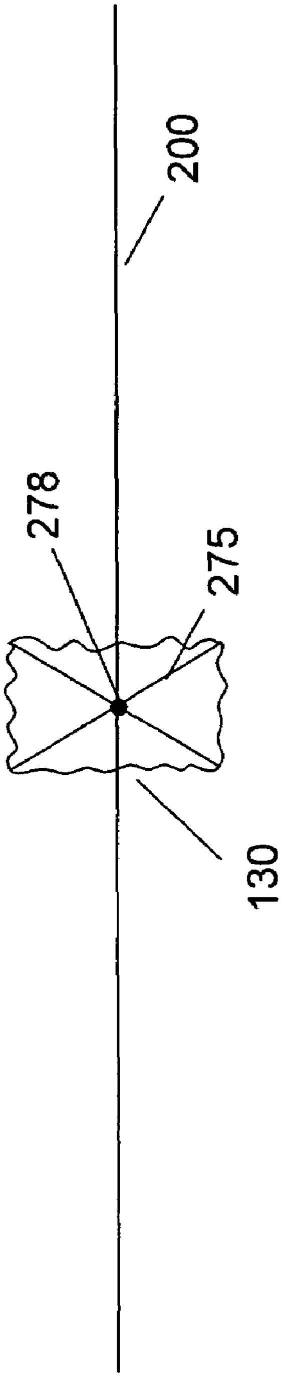
200
278
275
130

FIGURE 2D

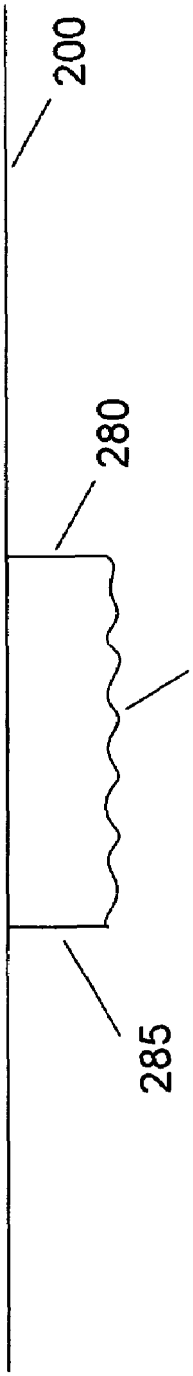
200
280
130
285

FIGURE 2E

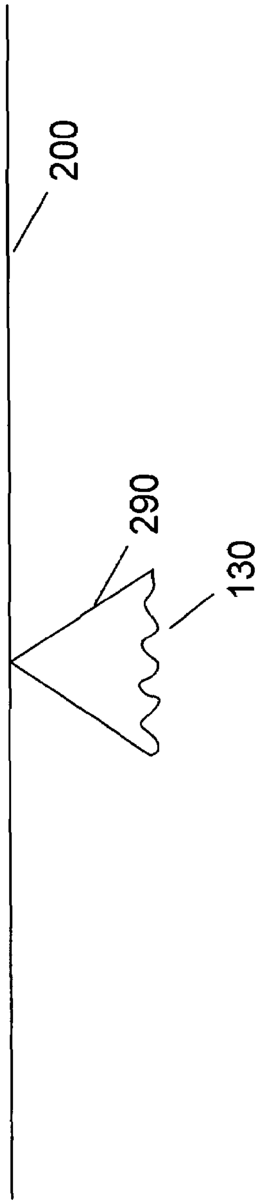

FIGURE 2F

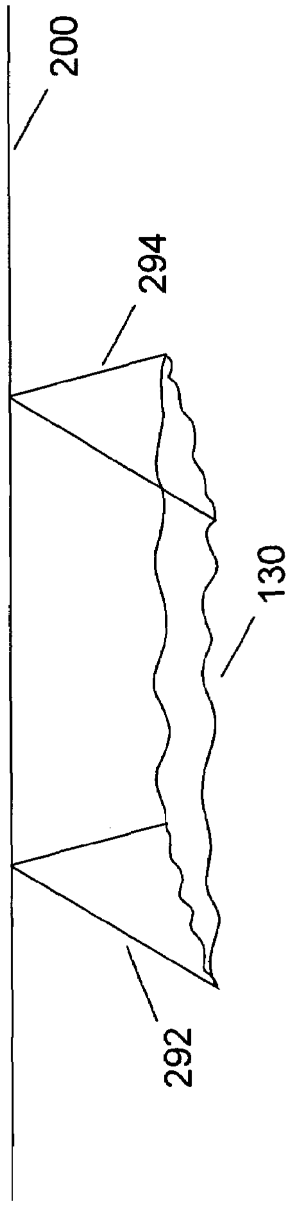
200
294
130
292

FIGURE 2G

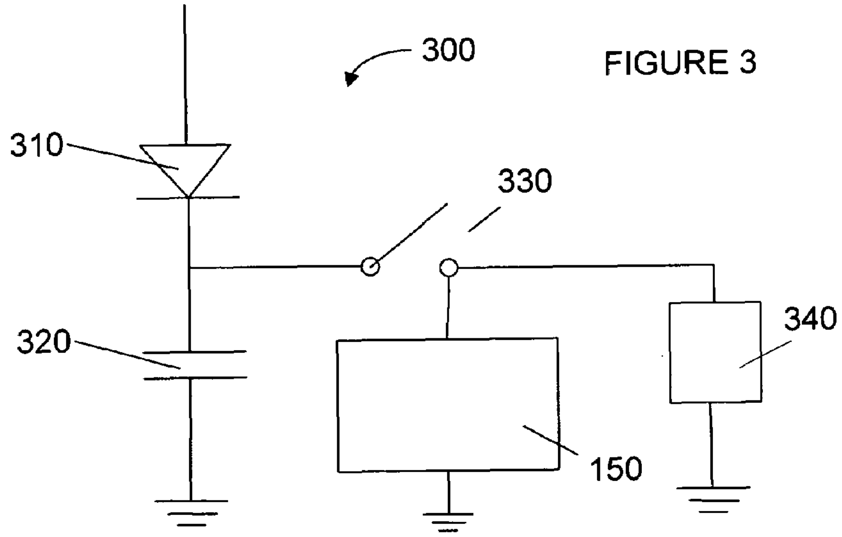
300
310
330
320
340
150

FIGURE 3

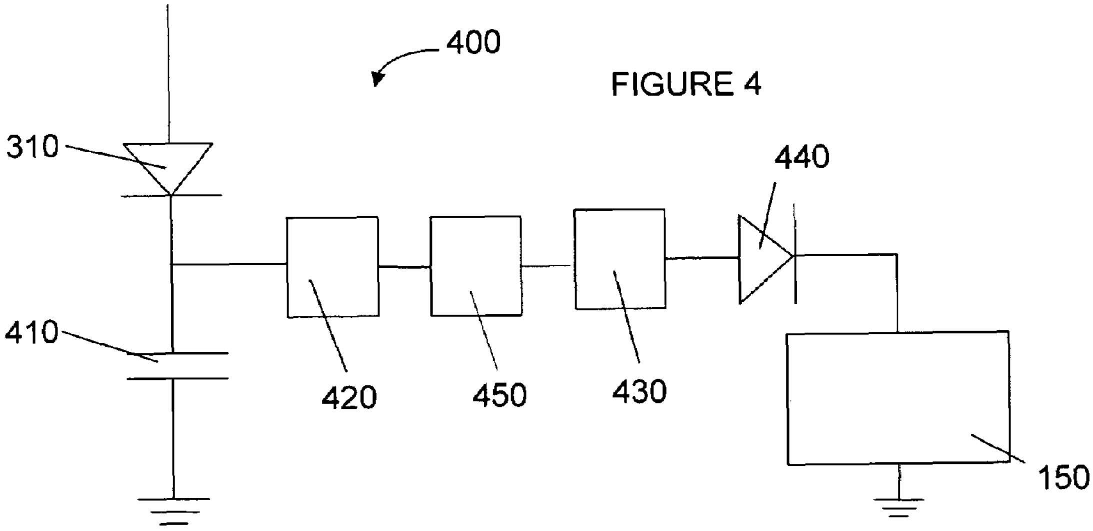
400
310
440
410
420
450
430
150

FIGURE 4

FIGURE 5

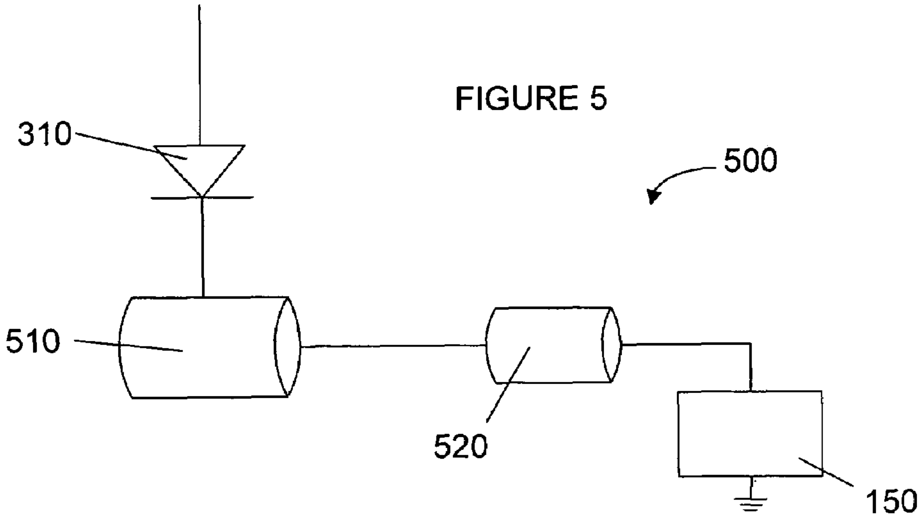
310
500
510
520
150

FIGURE 6

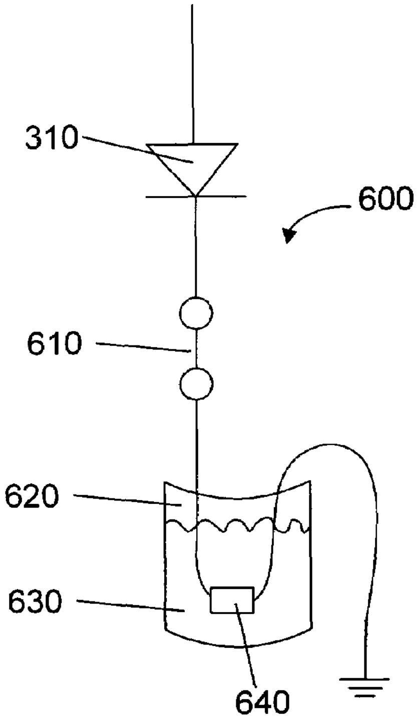
310
600
610
620
630
640

FIGURE 7

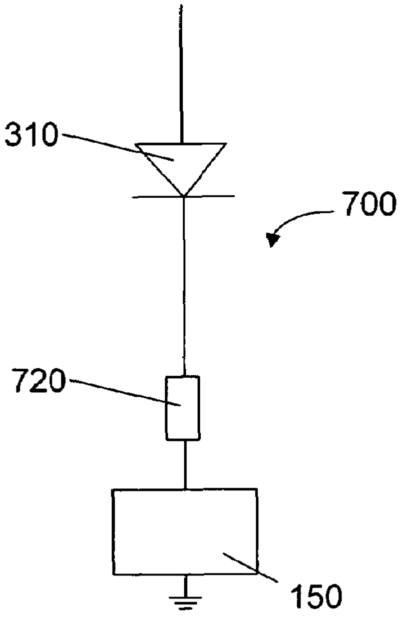
310
700
720
150

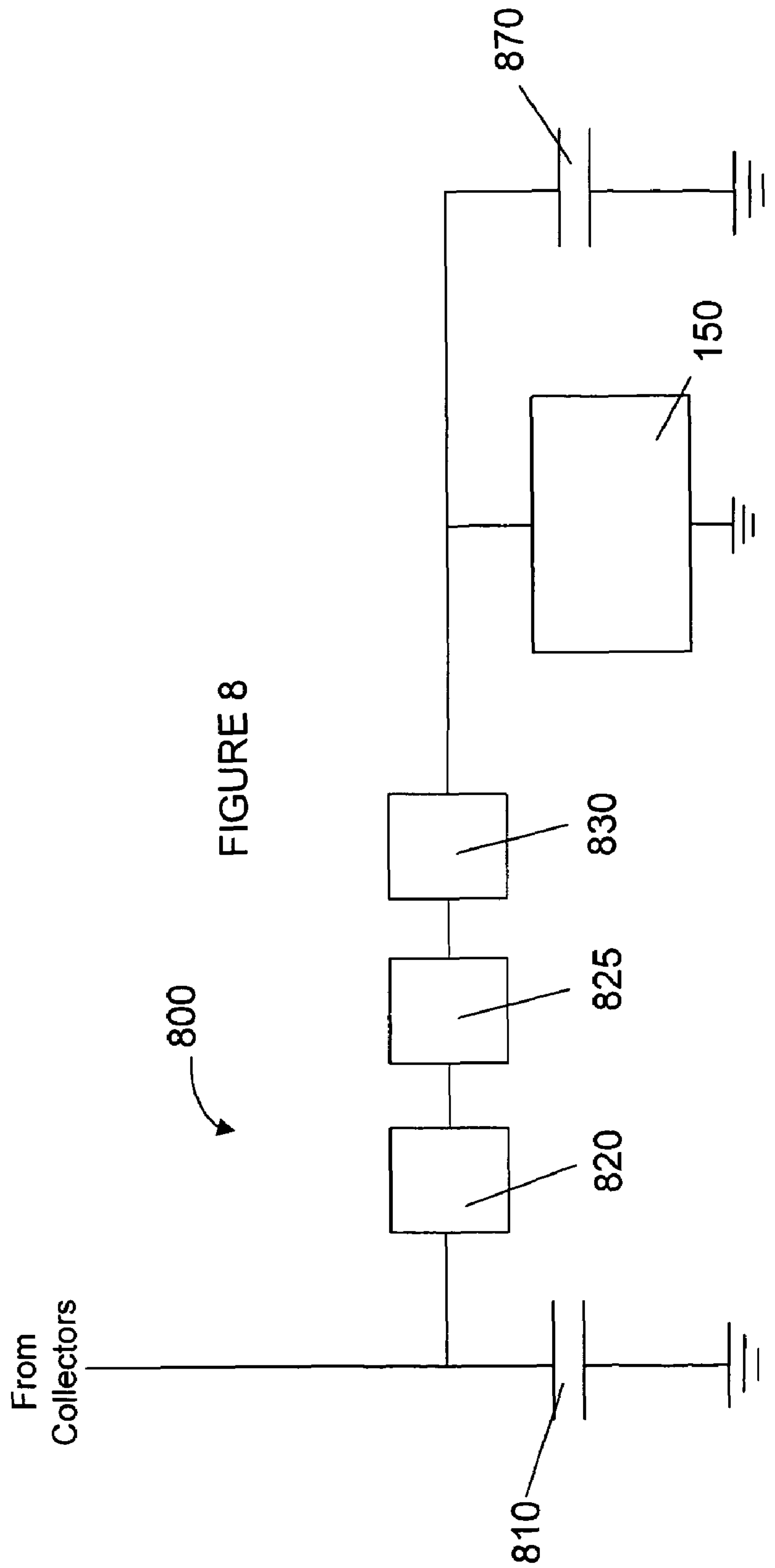
From Collectors
810
820
825
830
800
150
870

FIGURE 8

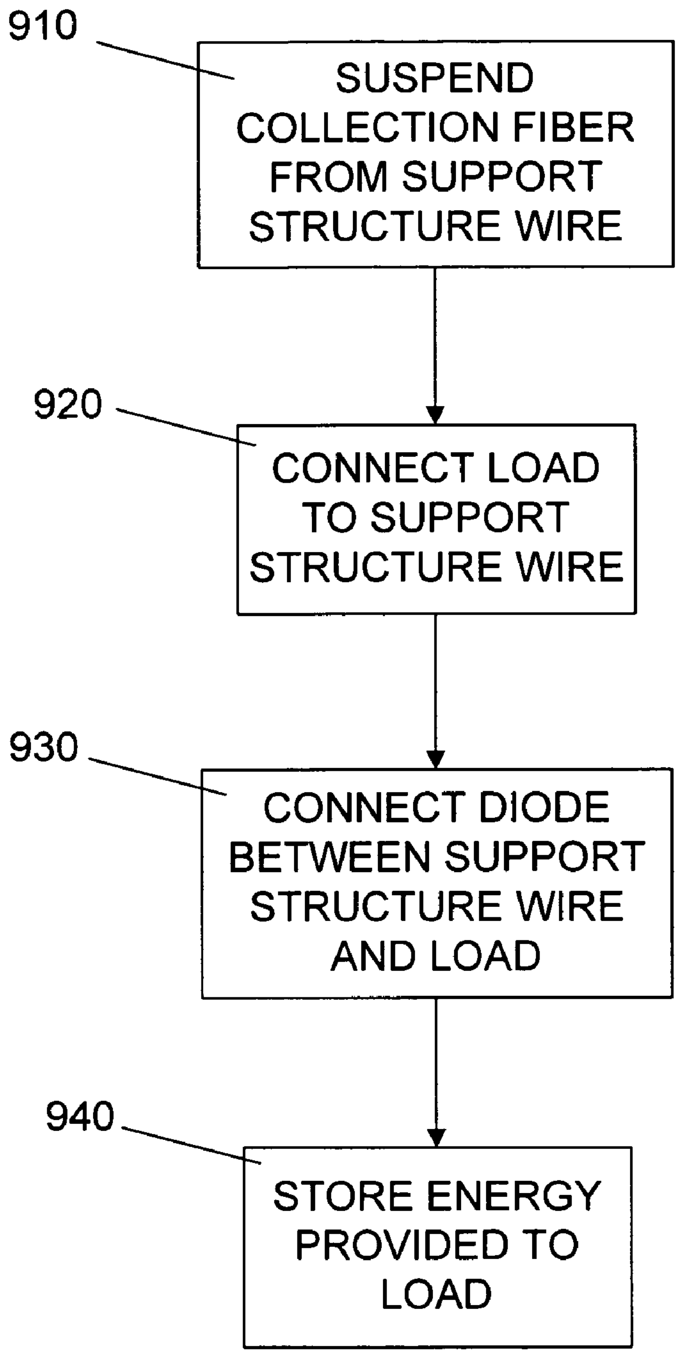
910
SUSPEND COLLECTION FIBER FROM SUPPORT STRUCTURE WIRE
920
CONNECT LOAD TO SUPPORT STRUCTURE WIRE
930
CONNECT DIODE BETWEEN SUPPORT STRUCTURE WIRE AND LOAD
940
STORE ENERGY PROVIDED TO LOAD

FIGURE 9

ENERGY COLLECTION

TECHNICAL FIELD

The present disclosure is generally related to energy and, more particularly, is related to systems and methods for collecting energy.

BACKGROUND

The concept of fair weather electricity deals with the electric field and the electric current in the atmosphere propagated by the conductivity of the air. Clear, calm air carries an electrical current, which is the return path for thousands of lightening storms simultaneously occurring at any given moment around the earth. For simplicity, this energy may be referred to as static electricity or static energy. FIG. **1** illustrates a weather circuit for returning the current from lightning, for example, back to ground **10**. Weather currents **20**, **30** return the cloud to ground current **40**.

In a lightening storm, an electrical charge is built up, and electrons arc across a gas, ionizing it and producing the lightening flash. As one of ordinary skill in the art understands, the complete circuit requires a return path for the lightening flash. The atmosphere is the return path for the circuit. The electric field due to the atmospheric return path is relatively weak at any given point because the energy of thousands of electrical storms across the planet are diffused over the atmosphere of the entire Earth during both fair and stormy weather. Other contributing factors to electric current being present in the atmosphere may include cosmic rays penetrating and interacting with the earth's atmosphere, and also the migration of ions, as well as other effects yet to be fully studied.

Some of the ionization in the lower atmosphere is caused by airborne radioactive substances, primarily radon. In most places of the world, ions are formed at a rate of 5-10 pairs per cubic centimeter per second at sea level. With increasing altitude, cosmic radiation causes the ion production rate to increase. In areas with high radon exhalation from the soil (or building materials), the rate may be much higher.

Alpha-active materials are primarily responsible for the atmospheric ionization. Each alpha particle (for instance, from a decaying radon atom) will, over its range of some centimeters, create approximately 150,000-200,000 ion pairs.

While there is a large amount of usable energy available in the atmosphere, a method or apparatus for efficiently collecting that energy has not been forthcoming. Therefore, a heretofore unaddressed need exists in the industry to address the aforementioned deficiencies and inadequacies.

SUMMARY

Embodiments of the present disclosure provide systems and methods for collecting energy. Briefly described in architecture, one embodiment of the system, among others, can be implemented by a support structure wire elevated above a ground level, at least one collection fiber electrically connected to the support structure wire; a load electrically connected to the support structure wire; and a diode electrically connected between the load and at least one collection fiber.

Embodiments of the present disclosure can also be viewed as providing methods for collecting energy. In this regard, one embodiment of such a method, among others, can be broadly summarized by the following steps: suspending at least one collection fiber from a support structure wire elevated above ground level, the fiber electrically connected to the support structure wire; providing a load with an electrical connection to the support structure wire to draw current; and providing a diode electrically connected between the collection fiber and the load.

Other systems, methods, features, and advantages of the present disclosure will be or become apparent to one with skill in the art upon examination of the following drawings and detailed description. It is intended that all such additional systems, methods, features, and advantages be included within this description, be within the scope of the present disclosure, and be protected by the accompanying claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Many aspects of the disclosure can be better understood with reference to the following drawings. The components in the drawings are not necessarily to scale, emphasis instead being placed upon clearly illustrating the principles of the present disclosure. Moreover, in the drawings, like reference numerals designate corresponding parts throughout the several views.

FIG. **1** is a circuit diagram of a weather energy circuit.

FIG. **2** is a perspective view of an exemplary embodiment of many energy collectors elevated above ground by a structure.

FIG. **2**A is a side view of an energy collection fiber suspended from a support wire.

FIG. **2**B is a side view of an exemplary embodiment of an energy collection fiber suspended from a support wire and with an additional support member.

FIG. **2**C is a perspective view of a support structure for multiple energy collection fibers.

FIG. **2**D is a side view of an exemplary embodiment of a support structure for multiple energy collection fibers.

FIG. **2**E is a side view of a support structure for an energy collection fiber.

FIG. **2**F is a side view of an exemplary embodiment of a support structure for an energy collection fiber.

FIG. **2**G is a side view of a support structure for multiple energy collection fibers.

FIG. **3** is a circuit diagram of an exemplary embodiment of a circuit for the collection of energy.

FIG. **4** is a circuit diagram of an exemplary embodiment of a circuit for the collection of energy.

FIG. **5** is a circuit diagram of an exemplary embodiment of an energy collection circuit for powering a generator and motor.

FIG. **6** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy and using it for the production of hydrogen and oxygen.

FIG. **7** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy, and using it for driving a fuel cell.

FIG. **8** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy.

FIG. **9** is a flow diagram of an exemplary embodiment of collecting energy with a collection fiber.

DETAILED DESCRIPTION

Electric charges on conductors reside entirely on the external surface of the conductors, and tend to concentrate more around sharp points and edges than on flat surfaces. Therefore, an electric field received by a sharp conductive point may be much stronger than a field received by the same charge residing on a large smooth conductive shell. An exemplary embodiment of this disclosure takes advantage of this property, among others, to collect and use the energy generated by an electric field in the atmosphere. Referring to collection system **100** presented in FIG. **2**, at least one collection device **130** may be suspended from a support wire system **120** supported by poles **110**. Collection device **130** may comprise a diode or a collection fiber individually, or the combination of a diode and a collection fiber. Support wire system **120** may be electrically connected to load **150** by connecting wire **140**. Supporting wire system **120** may be any shape or pattern. Also, conducting wire **140** may be one wire or multiple wires. The collection device **130** in the form of a fiber may comprise any conducting or non-conducting material, including carbon, graphite, Teflon, and metal. An exemplary embodiment utilizes carbon or graphite fibers for static electricity collection. Support wire system **120** and connecting wire **140** can be made of any conducting material, including aluminum or steel, but most notably, copper. Teflon may be added to said conductor as well, such as non-limiting examples of a Teflon impregnated wire, a wire with a Teflon coating, or Teflon strips hanging from a wire. Conducting wire **120**, **140**, and **200** may be bare wire, or coated with insulation as a non-limiting example. Wires **120** and **140** are a means of transporting the energy collected by collection device **130**.

An exemplary embodiment of the collection fibers as collection device **130** includes graphite or carbon fibers. Graphite and carbon fibers, at a microscopic level, can have hundreds of thousands of points. Atmospheric electricity may be attracted to these points. If atmospheric electricity can follow two paths where one is a flat surface and the other is a pointy, conductive surface, the electrical charge will be attracted to the pointy, conductive surface. Generally, the more points that are present, the higher energy that can be gathered. Therefore, carbon, or graphite fibers are examples that demonstrate exemplary collection ability.

In at least one exemplary embodiment, the height of support wire **120** may be an important factor. The higher that collection device **130** is from ground, the larger the voltage potential between collection device **130** and electrical ground. The electric field may be more than **100** volts per meter under some conditions. When support wire **120** is suspended in the air at a particular altitude, wire **120** will itself collect a very small charge from ambient voltage. When collection device **130** is connected to support wire **120**, collection device **130** becomes energized and transfers the energy to support wire **120**.

A diode, not shown in FIG. **2**, may be connected in several positions in collection system **100**. A diode is a component that restricts the direction of movement of charge carriers. It allows an electric current to flow in one direction, but essentially blocks it in the opposite direction. A diode can be thought of as the electrical version of a check valve. The diode may be used to prevent the collected energy from discharging into the atmosphere through the collection fiber embodiment of collection device **130**. An exemplary embodiment of the collection device comprises the diode with no collection fiber. A preferred embodiment, however, includes a diode at the connection point of a collection fiber to support system **120** such that the diode is elevated above ground. Multiple diodes may be used between collection device **130** and load **150**. Additionally, in an embodiment with multiple fibers, the diodes restricts energy that may be collected through one fiber from escaping through another fiber.

Collection device **130** may be connected and arranged in relation to support wire system **120** by many means. Some non-limiting examples are provided in FIGS. **2A-2G** using a collection fiber embodiment. FIG. **2A** presents support wire **200** with connecting member **210** for collection device **130**.

Connection member **210** may be any conducting material allowing for the flow of electricity from connection device **130** to support wire **200**. Then, as shown in FIG. **2**, the support wire **200** of support system **120** may be electrically connected through conducting wire **140** to load **150**. A plurality of diodes may be placed at any position on the support structure wire. A preferred embodiment places a diode at an elevated position at the connection point between a collection fiber embodiment of collection device **130** and connection member **210**.

Likewise, FIG. **2B** shows collection fiber **130** electrically connected to support wire **200** and also connected to support member **230**. Support member **230** may be connected to collection fiber **130** on either side. Support member **230** holds the fiber steady on both ends instead of letting it move freely. Support member **230** may be conducting or non-conducting. A plurality of diodes may be placed at any position on the support structure wire. A preferred embodiment places a diode at elevated position at the connection point between collection fiber **130** and support wire **200** or between fiber **130**, support member **230**, and support wire **200**.

FIG. **2C** presents multiple collection fibers in a squirrel cage arrangement with top and bottom support members. Support structure **250** may be connected to support structure wire **200** by support member **240**. Structure **250** has a top **260** and a bottom **270** and each of the multiple collection fibers **130** are connected on one end to top **260** and on the other end to bottom **270**. A plurality of diodes may be placed at any position on support structure **250**. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support structure wire **200**.

FIG. **2D** presents another exemplary embodiment of a support structure with support members **275** in an x-shape connected to support structure wire **200** at intersection **278** with collection fibers **130** connected between ends of support members **275**. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support wire **200**.

FIG. **2E** provides another exemplary embodiment for supporting collection fiber **130**. Collection fiber **130** may be connected on one side to support member **285**, which may be connected to support structure wire **200** in a first location and on the other side to support member **280**, which may be connected to support structure wire **200** in a second location on support structure wire **200**. The first and second locations may be the same location, or they may be different locations, even on different support wires. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places one or more diodes at elevated positions at the connection point(s) between collection fiber **130** and support wire **200**.

FIG. **2F** presents another exemplary embodiment of a support for a collection fiber. Two support members **290** may support either side of a collection fiber and are connected to support wire **200** in a single point. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support wire **200**.

FIG. **2G** provides two supports as provided in FIG. **2F** such that at least two support members **292**, **294** may be connected to support structure wire **200** in multiple locations and collection fibers **130** may be connected between each end of the support structures. Collection fibers **130** may be connected between each end of a single support structure and between multiple support structures. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places one or more diodes at elevated positions at the connection point(s) between collection fiber **130** and support structure wire **200**.

FIG. **3** provides a schematic diagram of storing circuit **300** for storing energy collected by one or more collection devices (**130** from FIG. **2**). Load **150** induces current flow. Diode **310** may be electrically connected in series between one or more collection devices (**130** from FIG. **2**) and load **150**. A plurality of diodes may be placed at any position in the circuit. Switch **330** may be electrically connected between load **150** and at least one collection device (**130** from FIG. **2**) to connect and disconnect the load.

Capacitor **320** may be connected in parallel to the switch **330** and load **150** to store energy when switch **330** is open for delivery to load **150** when switch **330** is closed. Rectifier **340** may be electrically connected in parallel to load **150**, between the receiving end of switch **330** and ground. Rectifier **340** may be a full-wave or a half-wave rectifier. Rectifier **340** may include a diode electrically connected in parallel to load **150**, between the receiving end of switch **330** and ground. The direction of the diode of rectifier **340** is optional.

In an exemplary embodiment provided in FIG. **4**, storage circuit **400** stores energy from one or more collection devices (**130** from FIG. **2**) by charging capacitor **410**. If charging capacitor **410** is not used, then the connection to ground shown at capacitor **410** is eliminated. A plurality of diodes may be placed at any position in the circuit. Diode **310** may be electrically connected in series between one or more collection devices (**130** from FIG. **2**) and load **150**. Diode **440** may be placed in series with load **150**. The voltage from capacitor **410** can be used to charge spark gap **420** when it reaches sufficient voltage. Spark gap **420** may comprise one or more spark gaps in parallel. Non-limiting examples of spark gap **420** include mercury-reed switches and mercury-wetted reed switches. When spark gap **420** arcs, energy will arc from one end of the spark gap **420** to the receiving end of the spark gap **420**. The output of spark gap **420** may be electrically connected in series to rectifier **450**. Rectifier **450** may be a full-wave or a half-wave rectifier. Rectifier **450** may include a diode electrically connected in parallel to transformer **430** and load **150**, between the receiving end of spark gap **420** and ground.

The direction of the diode of rectifier **450** is optional. The output of rectifier **450** is connected to transformer **430** to drive load **150**.

FIG. **5** presents motor driver circuit **500**. One or more collection devices (**130** from FIG. **2**) are electrically connected to static electricity motor **510**, which powers generator **520** to drive load **150**. A plurality of diodes may be placed at any position in the circuit. Motor **510** may also be directly connected to load **150** to drive it directly.

FIG. **6** demonstrates a circuit **600** for producing hydrogen. A plurality of diodes may be placed at any position in the circuit. One or more collection devices (**130** from FIG. **2**) are electrically connected to primary spark gap **610**, which may be connected to secondary spark gap **640**. Non-limiting examples of spark gaps **610**, **640** include mercury-reed switches and mercury-wetted reed switches. Secondary spark gap **640** may be immersed in water **630** within container **620**. When secondary spark gap **640** immersed in water **630** is energized, spark gap **640** may produce bubbles of hydrogen and oxygen, which may be collected to be used as fuel.

FIG. **7** presents circuit **700** for driving a fuel cell. A plurality of diodes may be placed at any position in the circuit. Collection devices (**130** from FIG. **2**) provide energy to fuel cell **720** which drives load **150**. Fuel cell **720** may produce hydrogen and oxygen.

FIG. **8** presents exemplary circuit **800** for the collection of energy. Storage circuit **800** stores energy from one or more collection devices (**130** from FIG. **2**) by charging capacitor **810**. If charging capacitor **810** is not used, then the connection to ground shown at capacitor **810** is eliminated. A plurality of diodes may be placed at any position in the circuit. The voltage from capacitor **810** can be used to charge spark gap **820** when it reaches sufficient voltage. Spark gap **820** may comprise one or more spark gaps in parallel or in series. Non-limiting examples of spark gap **820** include mercury-reed switches and mercury-wetted reed switches. When spark gap **820** arcs, energy will arc from one end of spark gap **820** to the receiving end of spark gap **820**. The output of spark gap **820** may be electrically connected in series to rectifier **825**. Rectifier **825** may be a full-wave or a half-wave rectifier. Rectifier **825** may include a diode electrically connected in parallel to inductor **830** and load **150**, between the receiving end of spark gap **820** and ground. The direction of the diode of rectifier **825** is optional. The output of rectifier **825** is connected to inductor **830**. Inductor **830** may be a fixed value inductor or a variable inductor. Capacitor **870** may be placed in parallel with load **150**.

FIG. **9** presents a flow diagram of a method for collecting energy. In block **910**, one or more collection devices may be suspended from a support structure wire. In block **920**, a load may be electrically connected to the support structure wire to draw current. In block **930** a diode may be electrically connected between the support structure wire and the electrical connection to the load. In block **940**, energy provided to the load may be stored or otherwise utilized.

Any process descriptions or blocks in flow charts should be understood as representing modules, segments, or portions of code which include one or more executable instructions for implementing specific logical functions or steps in the process, and alternate implementations are included within the scope of the preferred embodiment of the present disclosure in which functions may be executed out of order from that shown or discussed, including substantially concurrently or in reverse order, depending on the functionality involved, as would be understood by those reasonably skilled in the art of the present disclosure.

It should be emphasized that the above-described embodiments of the present disclosure, particularly, any "preferred" embodiments, are merely possible examples of implementations, merely set forth for a clear understanding of the principles of the disclosure. Many variations and modifications may be made to the above-described embodiment(s) of the disclosure without departing substantially from the spirit and principles of the disclosure. All such modifications and variations are intended to be included herein within the scope of this disclosure and the present disclosure and protected by the following claims.

Therefore, at least the following is claimed:

1. A method of collecting energy comprising:

suspending at least one collection device from a plurality of support structure wires elevated above a ground level, the collection device including a carbon or graphite fiber that collects electrical enemy from the air and is electrically connected to the support structure wire; and

providing a load with an electrical connection to the plurality of support structure wires to draw current.

2. The method of claim **1**, wherein the collection device comprises a diode.

3. The method of claim **1**, wherein the collection device comprises a diode electrically connected between the fiber and the load.

4. The method of claim **1**, further comprising storing energy provided to the load.

5. The method of claim **4**, wherein storing energy provided to the load comprises storing energy in a capacitor or an inductor.

6. The method of claim **2**, wherein the diode is elevated relative to the ground level.

7. A system of energy collection comprising:

a plurality of support structure wires elevated above ground level; at least one collection device including at least one carbon or graphite fiber that collects electrical enemy from the air and is electrically connected to the plurality of support structure wires; and

a load electrically connected to the plurality of support structure wires.

8. The system of claim **7**, wherein the collection device comprises a diode.

9. The system of claim **7**, wherein the collection device comprises a diode electrically connected between the load and the fiber.

10. The system of claim **8**, wherein the diode is elevated relative to the ground level.

11. The system of claim **7**, further comprising a diode electrically connected between the at least one collection device and the plurality of support structure wires.

12. The system of claim **7**, wherein the fiber comprises a first end and a second opposing end, and wherein the system further comprises a support structure connected to both ends of the fiber.

13. The system of claim **7**, comprising:

said at least one fiber including a plurality of collection fibers; a support frame having a top side and a bottom side; and a conducting connecting wire between the plurality of support structure wires and the support frame, wherein one end of each of a plurality of collection fibers is connected to the top side of the support frame, and the opposing end of each of a plurality of collection fibers is connected to the bottom side of the support frame.

14. The system of claim **7**, further comprising a rigid structure, the rigid structure comprising multiple supports extending outward from a single point of the support structure, wherein each end of the fiber is connected to an end of the multiple supports.

15. The system of claim **7**, further comprising a rigid structure, the rigid structure comprising multiple supports extending outward from multiple points of the support structure, wherein each end of the fiber is connected to an end of the multiple supports.

16. The system of claim **14**, further comprising a plurality of rigid structures wherein said at least one fiber includes collection fibers connected between a plurality of the rigid structures.

17. The system of claim **7**, further comprising:

a switch connected in series between the plurality of support structure wires and the load; and

a capacitor connected in parallel with the switch and the load.

18. The system of claim **17**, wherein the switch comprises an interrupter connected between the plurality of support structure wires and the load.

19. The system of claim **18**, wherein the interrupter comprises one of a fluorescent tube, a neon bulb, an AC light, or a spark gap.

20. The system of claim **18**, further comprising a transformer connected between the interrupter and the load.

21. The system of claim **7**, further comprising:

a motor for providing power the motor connected between the plurality of support structure wires and the load; and

a generator powered by the motor.

22. The system of claim **7**, wherein the load comprises a spark gap in a container of fluid, and the load is used to produce a chemical reaction.

23. The system of claim **22**, wherein the fluid comprises water and the chemical reaction comprises the production of hydrogen and oxygen.

24. The system of claim **7**, further comprising a fuel cell between the plurality of support structure wires and the load.

25. The system of claim **7**, wherein the load comprises a fuel cell.

26. The system of claim **24**, wherein the fuel cell produces hydrogen and oxygen.

27. The system of claim **24**, further comprising a diode connected between the plurality of support structure wires and the fuel cell.

28. A system of collecting energy comprising:

means for suspending carbon or graphite collection fibers, the means elevated above a ground level, the collection fibers collecting electrical enemy from the air and electrically connected to the means for suspending collection fibers;

means for inducing current flow, the means for inducing current flow electrically connected to the means for suspending collection fibers; and

means for restricting the backflow of charge carriers, the means for restricting the backflow of charge carriers electrically connected between the collection fibers and the means for inducing current flow.

* * * * *

US008686575B2

(12) **United States Patent**
McCowen

(10) **Patent No.: US 8,686,575 B2**
(45) **Date of Patent: Apr. 1, 2014**

(54) **ENERGY COLLECTION**

(75) Inventor: **Clint McCowen**, Navarre, FL (US)

(73) Assignee: **Ion Power Group, LLC**, Navarre, FL (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **12/255,130**

(22) Filed: **Oct. 21, 2008**

(65) **Prior Publication Data**

US 2009/0040680 A1 Feb. 12, 2009

Related U.S. Application Data

(63) Continuation of application No. 11/358,264, filed on Feb. 21, 2006, now Pat. No. 7,439,712.

(51) **Int. Cl.**

F02B 63/04	(2006.01)
F03G 7/08	(2006.01)
H02K 7/18	(2006.01)
H02N 1/00	(2006.01)

(52) **U.S. Cl.**
USPC .... **290/1 R**; 322/2 A

(58) **Field of Classification Search**
USPC .... 290/1 R; 322/2 A; 126/660; 60/641.8, 60/641.11, 641.12, 659, 698; 244/158.2
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

674,427	A		5/1901	Palencsar	307/149
911,260	A		2/1909	Pennock	307/149
1,014,719	A		1/1912	Pennock	307/149
2,473,819	A	*	6/1949	Pittman	315/36
3,532,959	A	*	10/1970	Erickson	320/166
3,780,722	A	*	12/1973	Swet	126/680
3,946,227	A		3/1976	Bingham	250/281
4,079,225	A	*	3/1978	Warner	219/110
4,092,250	A		5/1978	Sano et al.	210/500.29
4,104,696	A		8/1978	Cochran, Jr.	361/229
4,146,800	A	*	3/1979	Gregory et al.	290/44
4,180,056	A	*	12/1979	Schnabel et al.	126/660
4,201,197	A	*	5/1980	Dismer	126/600
4,206,396	A	*	6/1980	Marks	322/2 A
4,224,496	A	*	9/1980	Riordan et al.	219/110
4,307,936	A	*	12/1981	Ochiai	385/120
4,314,192	A	*	2/1982	Caro	322/2 R
4,346,478	A	*	8/1982	Sichling	398/107
4,425,905	A	*	1/1984	Mori	126/578
4,433,248	A	*	2/1984	Marks	290/44
RE31,678	E	*	9/1984	Ochiai	385/115

(Continued)

FOREIGN PATENT DOCUMENTS

WO	2007098341	8/2007

OTHER PUBLICATIONS

"Emissivity Values for Common Materials & Non-Metals", Raytek—A Fluke Company, http://www.raytek.com/Raytek/en-r0/IREducation/EmissivityNonMetals.htm, 1999.*

Primary Examiner — Pedro J Cuevas

(74) *Attorney, Agent, or Firm* — Benjamin A. Balser; Balser & Grell IP Law

(57) **ABSTRACT**

An energy collection system may collect and use the energy generated by an electric field. Collection fibers are suspended from a support wire system supported by poles. The support wire system is electrically connected to a load by a connecting wire. The collection fibers may be made of any conducting material, but carbon and graphite are preferred. Diodes may be used to restrict the backflow or loss of energy.

25 Claims, 10 Drawing Sheets

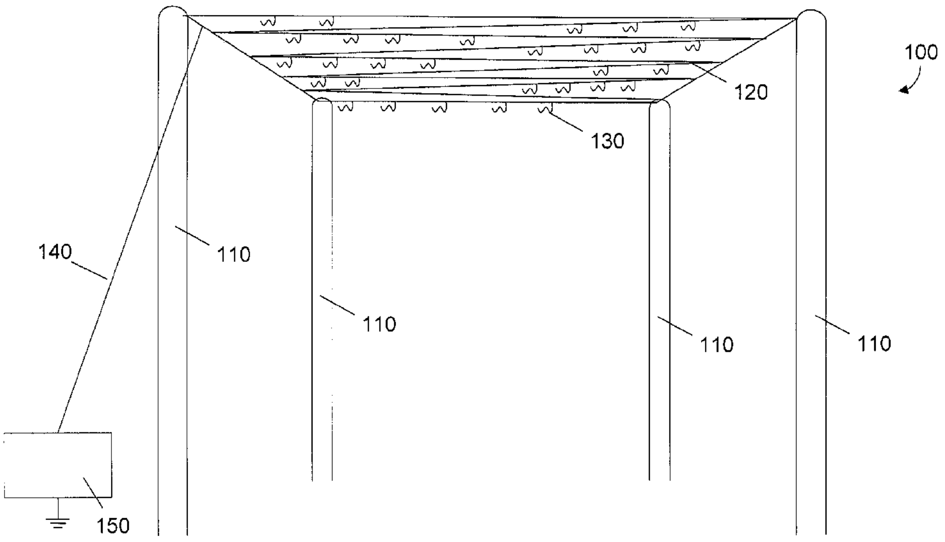

(56) References Cited

U.S. PATENT DOCUMENTS

4,483,311	A *	11/1984	Whitaker	126/602
4,489,269	A *	12/1984	Edling et al.	322/2 R
4,512,335	A *	4/1985	Mori	126/678
4,653,223	A *	3/1987	Mori	47/17
4,653,472	A *	3/1987	Mori	126/683
4,676,226	A *	6/1987	Mori	126/680
4,809,675	A *	3/1989	Mori	126/578
4,852,454	A *	8/1989	Batchelder	89/1.11
4,943,125	A *	7/1990	Laundre' et al.	385/35
4,988,159	A *	1/1991	Turner et al.	385/33
5,047,892	A	9/1991	Sakata et al.	361/231
5,114,101	A *	5/1992	Stern et al.	244/172.8
5,145,257	A *	9/1992	Bryant et al.	374/131
5,379,103	A *	1/1995	Zigler	356/73
5,851,309	A *	12/1998	Kousa	136/248
5,942,806	A *	8/1999	Veliadis	290/1 R
6,025,591	A	2/2000	Taylor et al.	250/292
6,038,363	A *	3/2000	Slater et al.	385/147
6,116,544	A *	9/2000	Forward et al.	244/158.2
6,173,922	B1 *	1/2001	Hoyt et al.	244/158.2
6,226,440	B1 *	5/2001	Lyons	385/119
6,419,191	B1 *	7/2002	Hoyt et al.	244/158.2
6,425,391	B1 *	7/2002	Davoren et al.	126/683
6,735,830	B1	5/2004	Merciel	250/324
6,765,212	B2 *	7/2004	Goetz et al.	250/339.11
6,846,447	B2 *	1/2005	Mbachu et al.	264/406
6,853,447	B2 *	2/2005	Goetz	356/237.1
6,894,772	B2 *	5/2005	Goetz et al.	356/237.1
6,925,852	B2 *	8/2005	Susko	73/23.2
6,974,110	B2	12/2005	Grandics	244/171.5
7,109,597	B1 *	9/2006	Bose	290/43
7,128,867	B2 *	10/2006	Mbachu et al.	264/406
7,231,809	B2 *	6/2007	Susko	73/31.05
7,259,475	B2 *	8/2007	Hong et al.	307/64
7,439,712	B2 *	10/2008	McCowen	322/2 A
7,449,668	B2 *	11/2008	Schutten et al.	250/214 SW
7,478,712	B2 *	1/2009	McCowen	191/45 R
7,532,819	B1 *	5/2009	Triebes et al.	398/121
2002/0026933	A1 *	3/2002	Gottlieb	126/628
2002/0108892	A1 *	8/2002	Goetz et al.	209/576
2002/0109094	A1 *	8/2002	Goetz et al.	250/339.11
2002/0109835	A1 *	8/2002	Goetz	356/237.1
2003/0107869	A1	6/2003	Pan et al.	361/306.2
2003/0125630	A1 *	7/2003	Furnish	600/461
2003/0193319	A1 *	10/2003	Wood et al.	322/2 A
2004/0011925	A1 *	1/2004	Grandics	244/62
2004/0083793	A1 *	5/2004	Susko	73/31.05
2004/0094853	A1 *	5/2004	Mbachu et al.	264/40.1
2004/0160711	A1	8/2004	Stumberger	361/20
2005/0101024	A1 *	5/2005	Mbachu et al.	436/85
2005/0136311	A1 *	6/2005	Ueda et al.	429/30
2005/0270525	A1 *	12/2005	Susko	356/326
2007/0107765	A1 *	5/2007	Schutten et al.	136/243
2007/0195481	A1 *	8/2007	McCowen	361/233
2007/0218323	A1 *	9/2007	Sudo et al.	429/13
2007/0223171	A1	9/2007	Guy et al.	361/126
2007/0273206	A1 *	11/2007	McCowen	307/9.1
2007/0297892	A1 *	12/2007	Kildegaard	415/1
2008/0145714	A1 *	6/2008	Kagami	429/13
2009/0107842	A1	4/2009	Park et al.	204/554
2009/0114495	A1 *	5/2009	McCowen	191/45 R
2009/0216292	A1 *	8/2009	Pless et al.	607/33
2010/0090562	A1 *	4/2010	McCowen	310/309
2010/0090563	A1 *	4/2010	McCowen	310/309
2012/0164555	A1 *	6/2012	Yamazaki et al.	429/482
2012/0299559	A1 *	11/2012	McCowen	322/2 A
2013/0276776	A1 *	10/2013	Te Kamp et al.	126/660

* cited by examiner

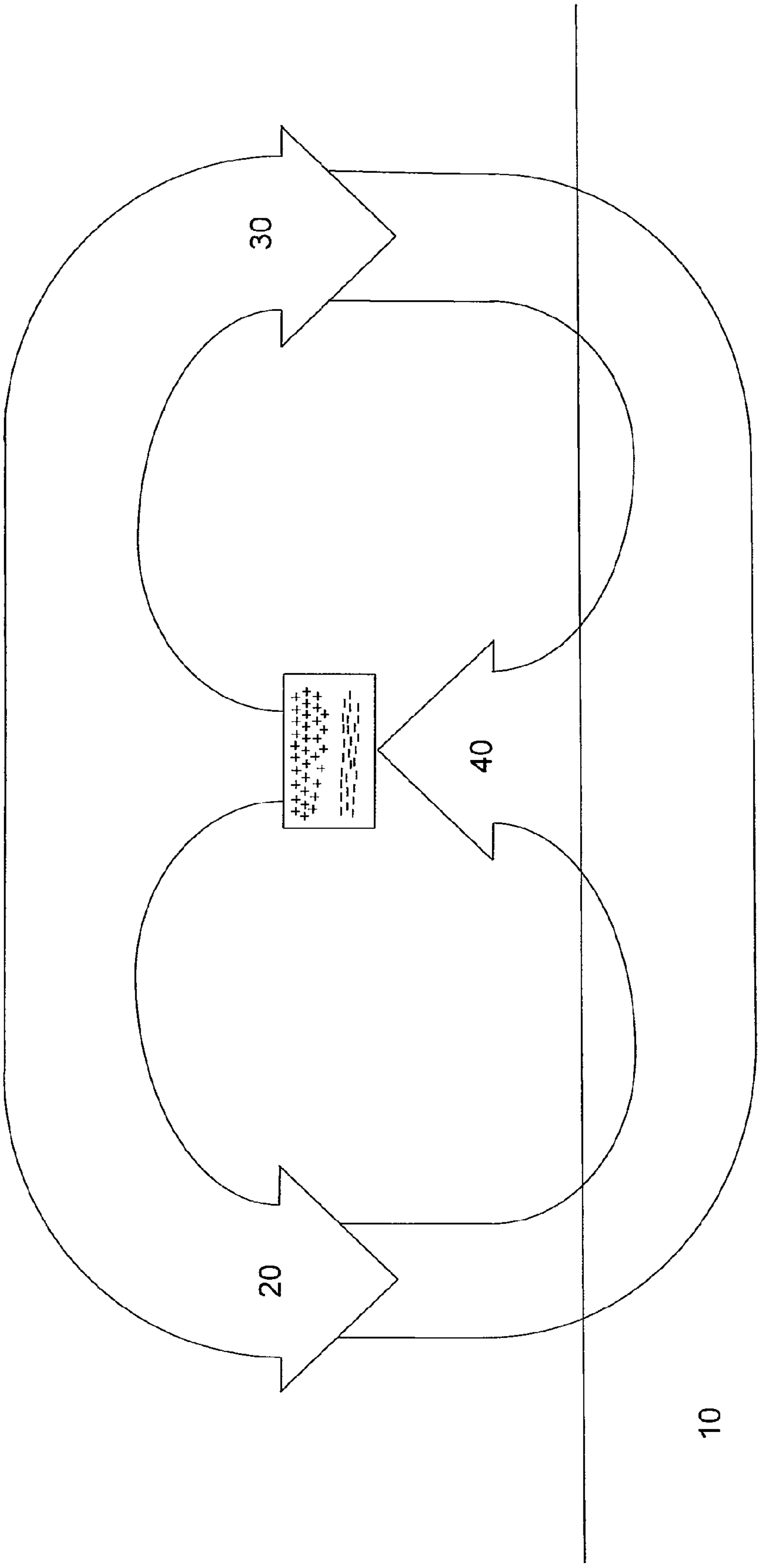

FIGURE 1

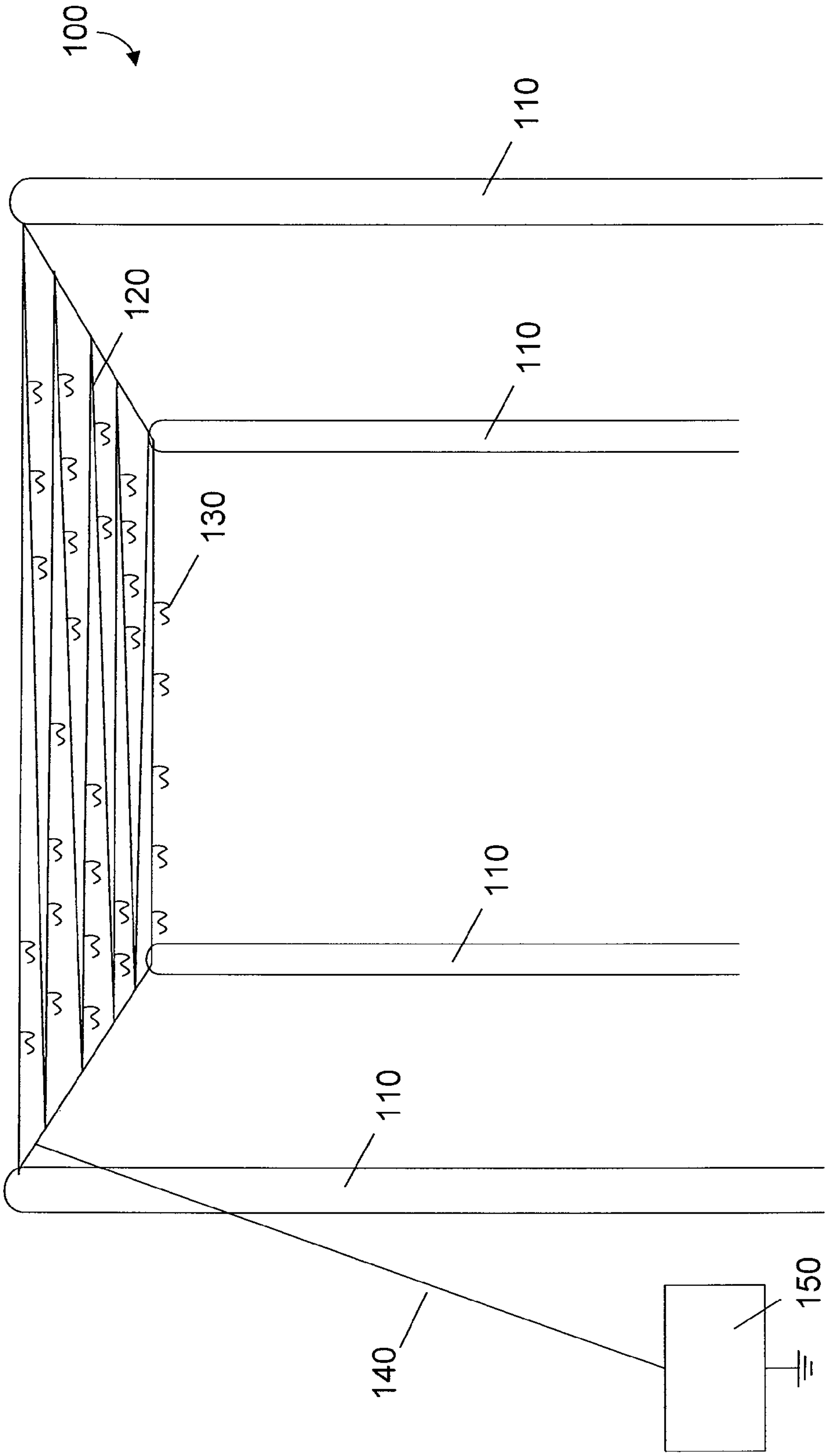
100
110
110
120
130
110
110
140
150

FIGURE 2

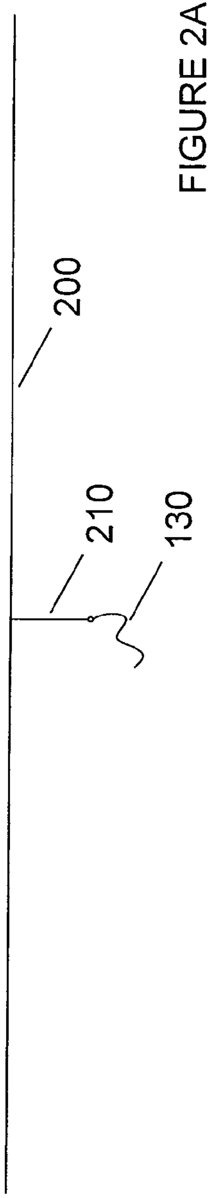

FIGURE 2A

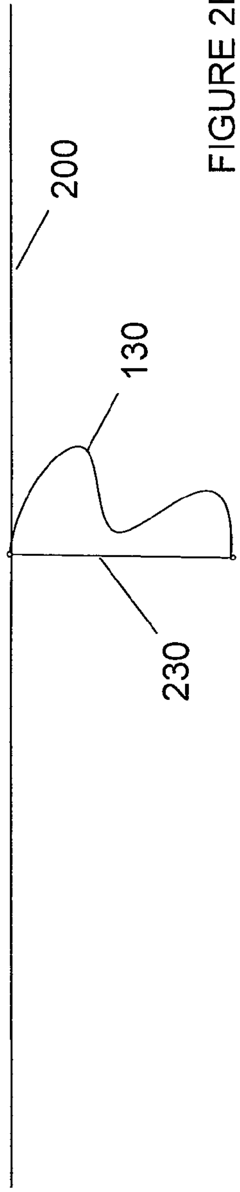

FIGURE 2B

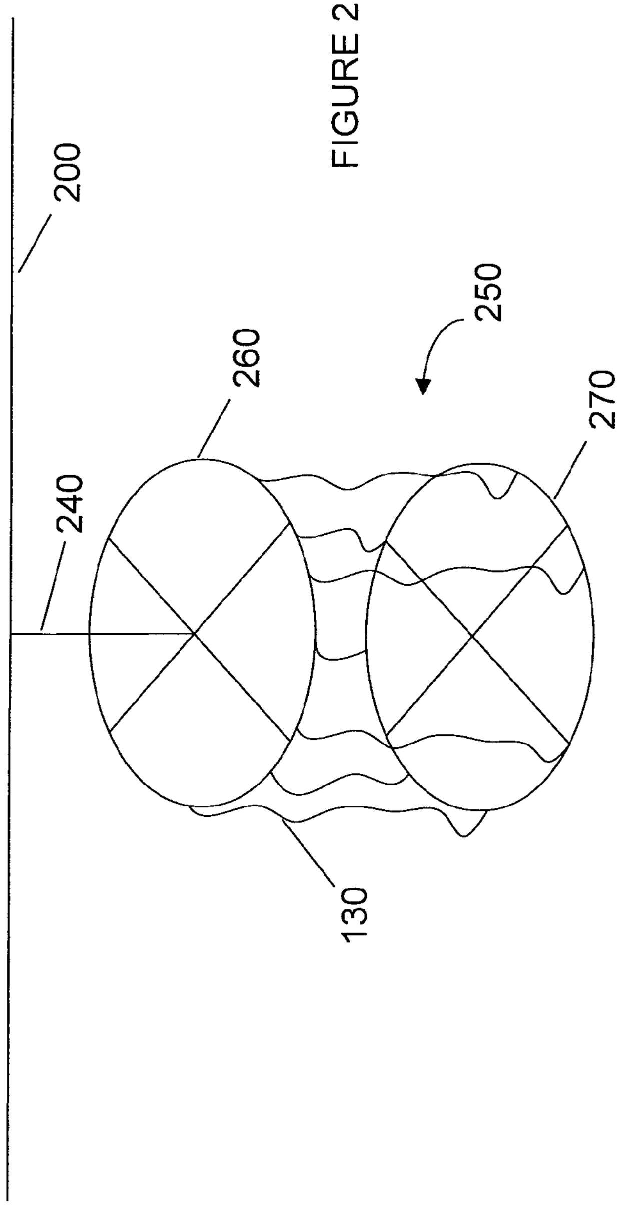
200
240
260
250
270
130

FIGURE 2C

FIGURE 2D

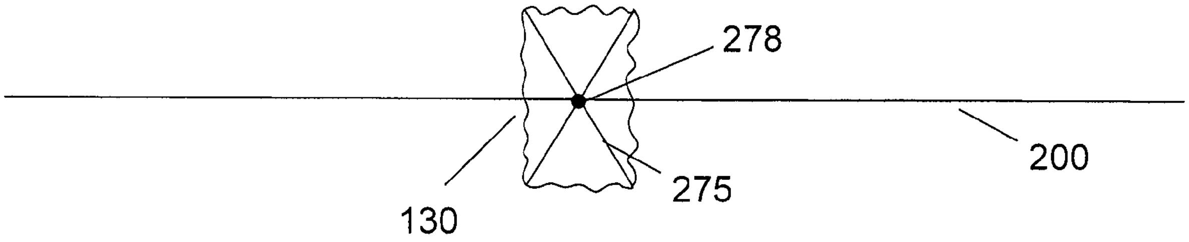
278
200
275
130

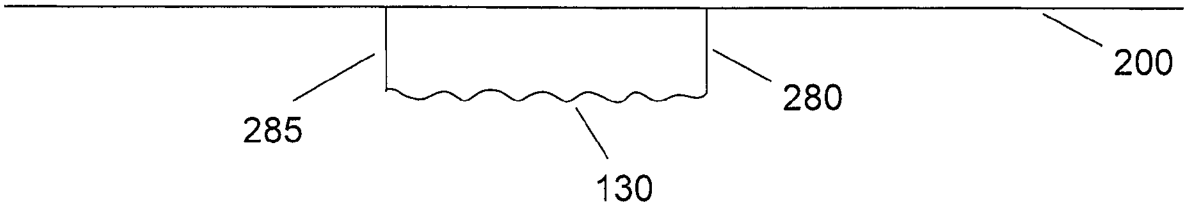
200
280
285
130

FIGURE 2E

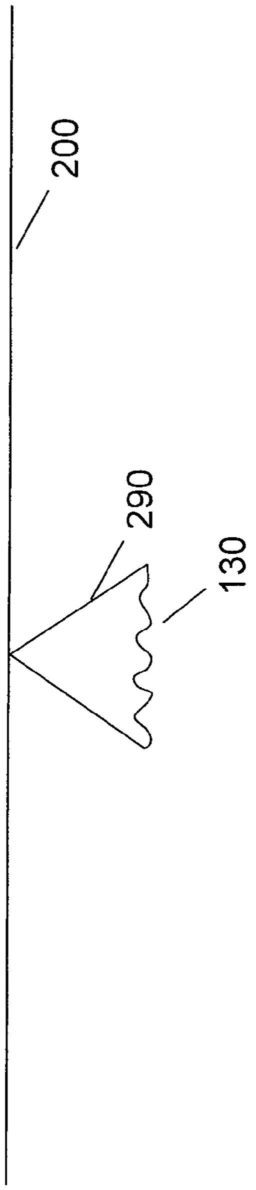

FIGURE 2F

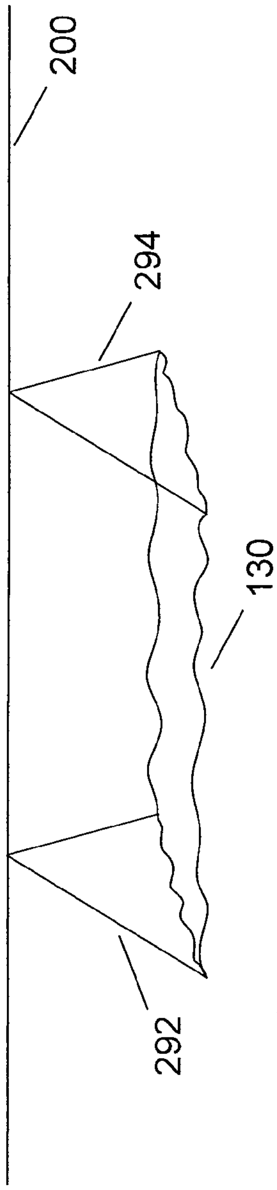
200
294
130
292

FIGURE 2G

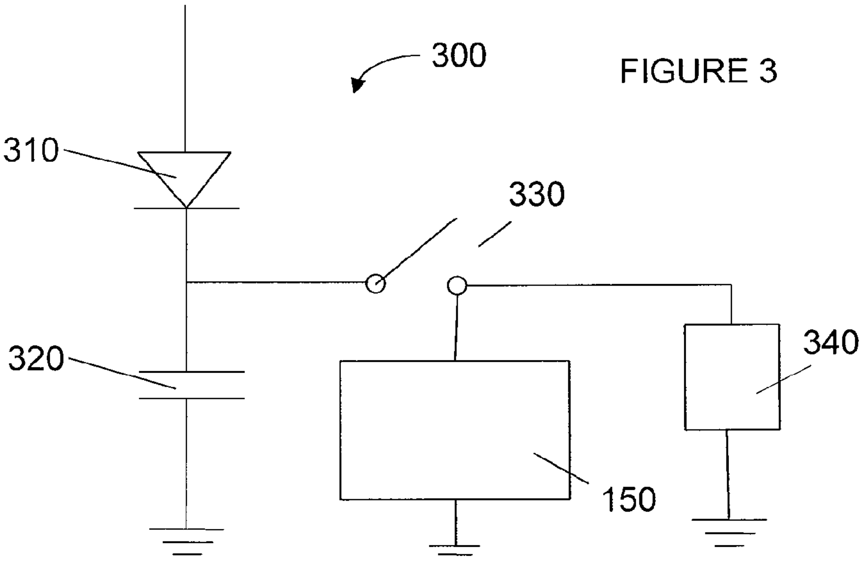
FIGURE 3
300
310
330
320
340
150

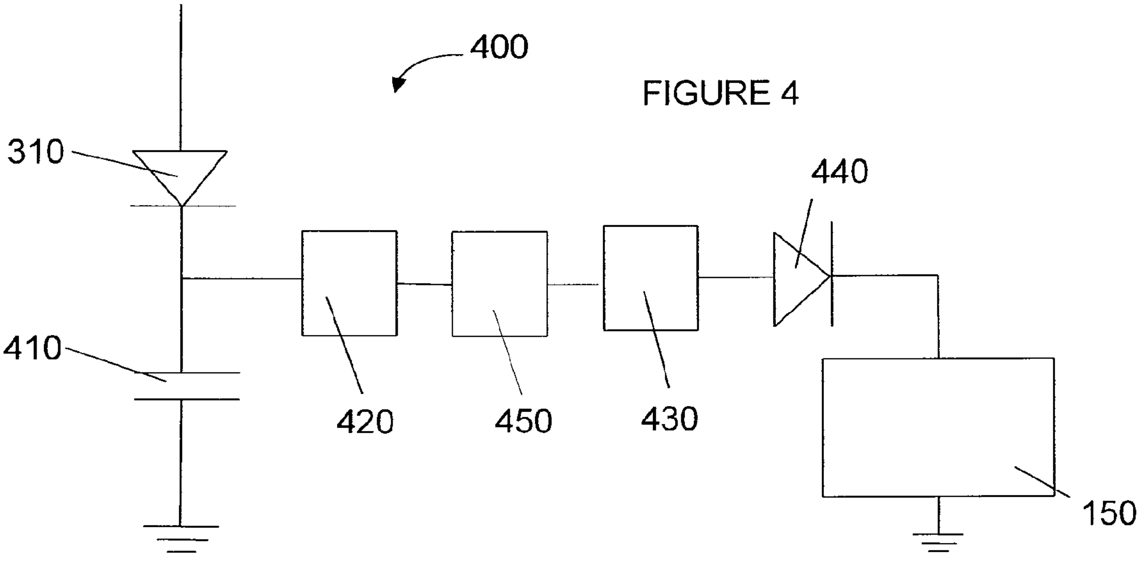
FIGURE 4
400
310
440
410
420
450
430
150

FIGURE 5

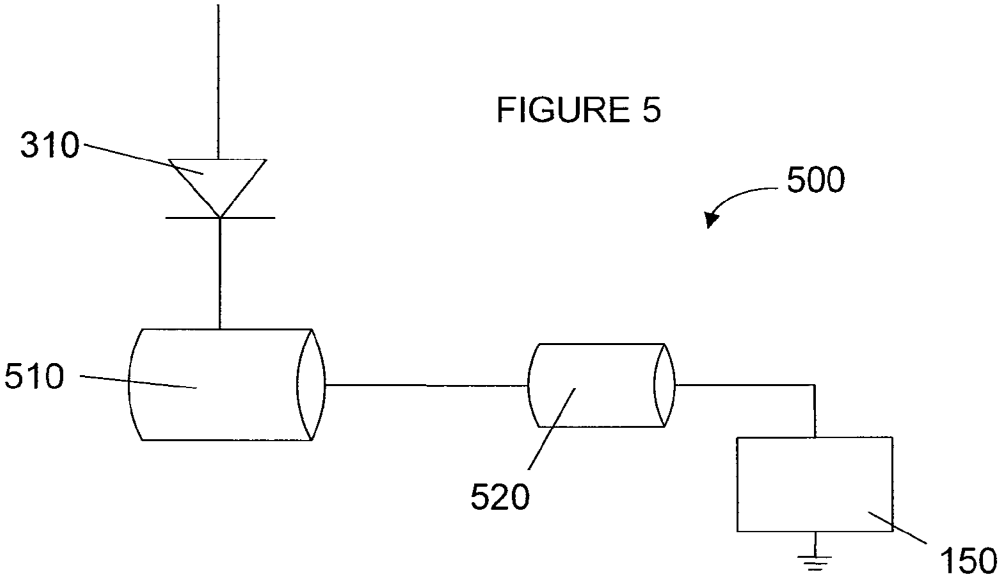
310
500
510
520
150

FIGURE 6

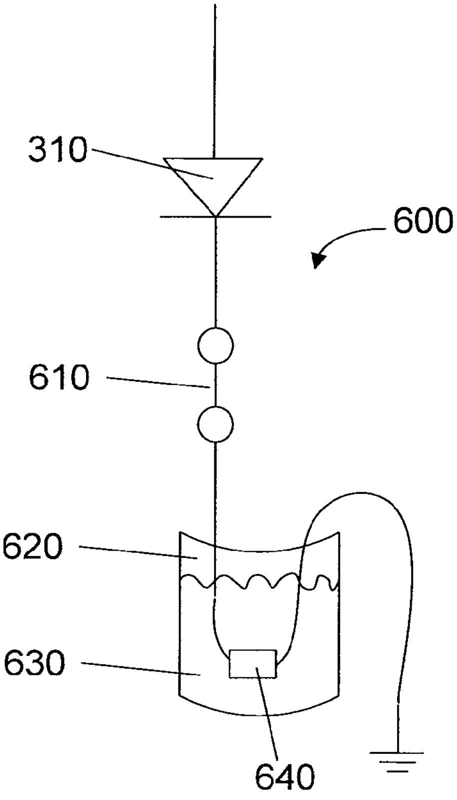
310
600
610
620
630
640

FIGURE 7

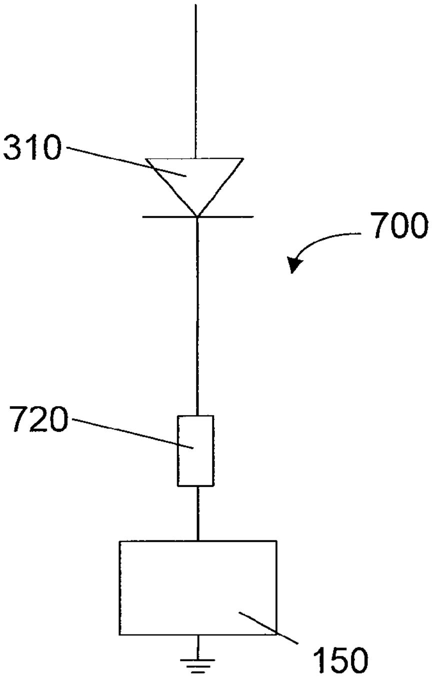
310
700
720
150

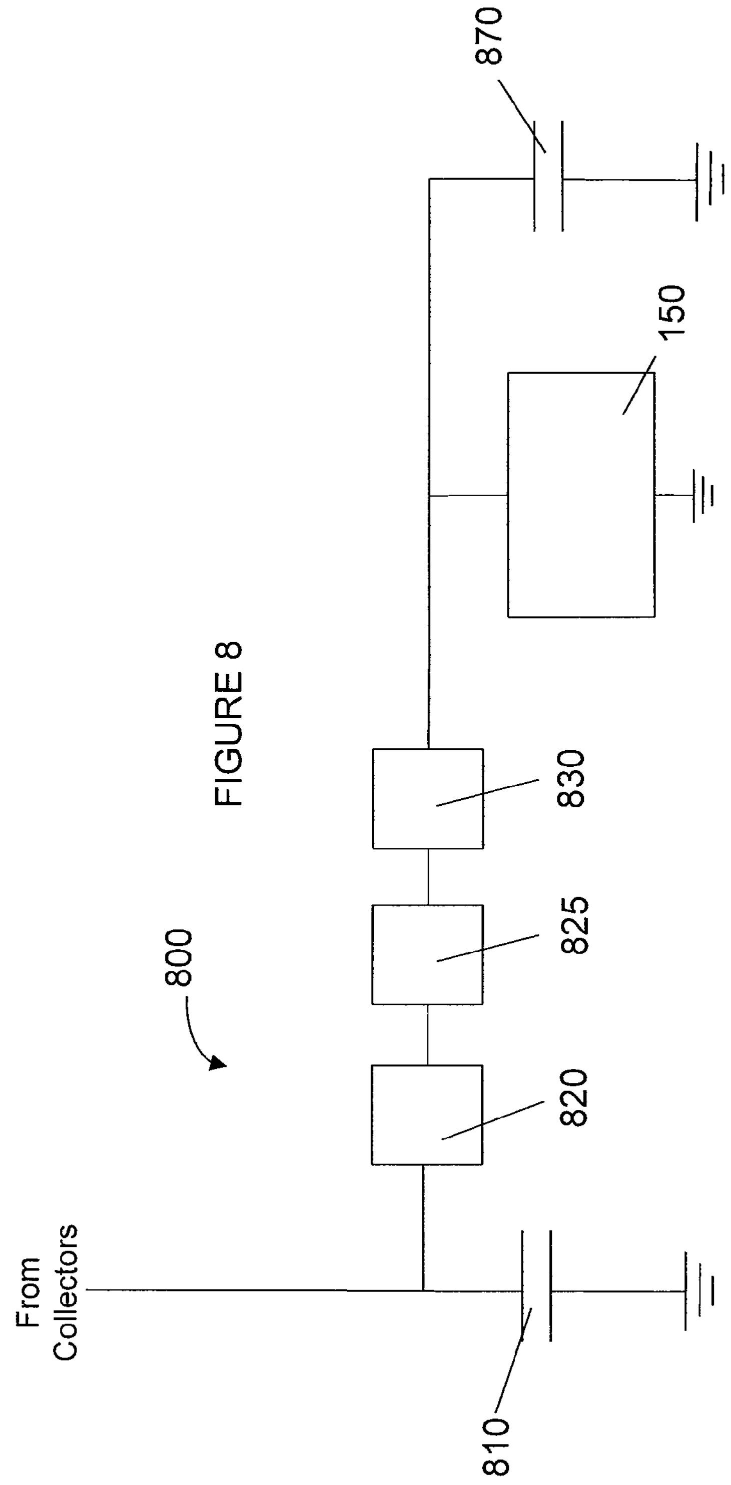

FIGURE 8

From Collectors
810
820
825
830
800
150
870

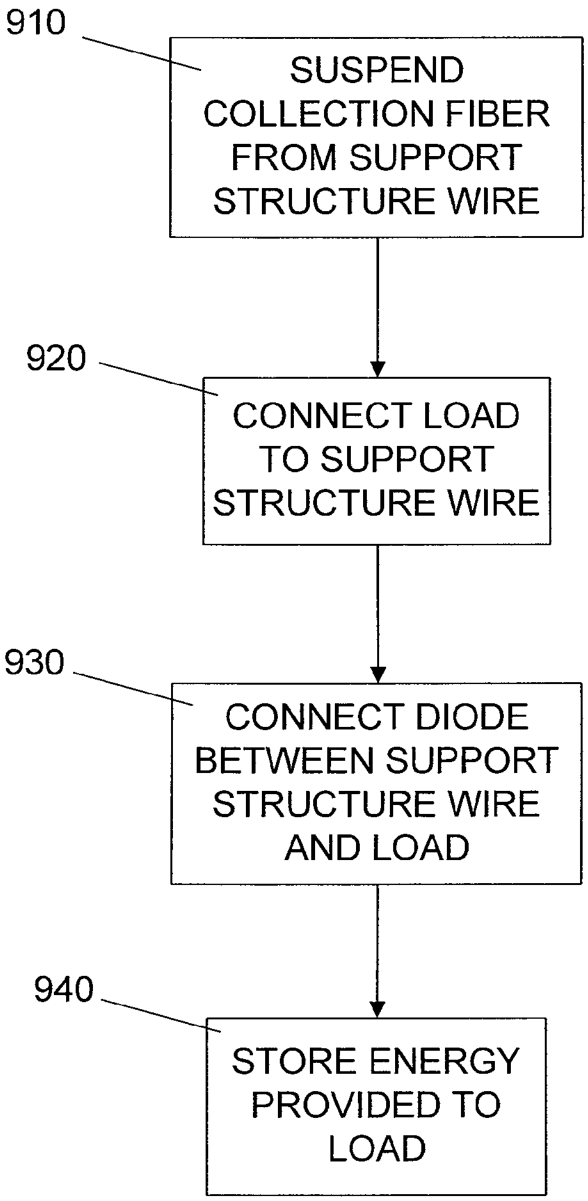
910
SUSPEND COLLECTION FIBER FROM SUPPORT STRUCTURE WIRE
920
CONNECT LOAD TO SUPPORT STRUCTURE WIRE
930
CONNECT DIODE BETWEEN SUPPORT STRUCTURE WIRE AND LOAD
940
STORE ENERGY PROVIDED TO LOAD

FIGURE 9

ENERGY COLLECTION

CROSS REFERENCE TO RELATED APPLICATION

This application is a continuation application of U.S. patent application Ser. No. 11/358,264, filed on Feb. 21, 2006, which is incorporated by reference herein.

TECHNICAL FIELD

The present disclosure is generally related to energy and, more particularly, is related to systems and methods for collecting energy.

BACKGROUND

The concept of fair weather electricity deals with the electric field and the electric current in the atmosphere propagated by the conductivity of the air. Clear, calm air carries an electrical current, which is the return path for thousands of lightening storms simultaneously occurring at any given moment around the earth. For simplicity, this energy may be referred to as static electricity or static energy. FIG. **1** illustrates a weather circuit for returning the current from lightning, for example, back to ground **10**. Weather currents **20**, **30** return the cloud to ground current **40**.

In a lightening storm, an electrical charge is built up, and electrons arc across a gas, ionizing it and producing the lightening flash. As one of ordinary skill in the art understands, the complete circuit requires a return path for the lightening flash. The atmosphere is the return path for the circuit. The electric field due to the atmospheric return path is relatively weak at any given point because the energy of thousands of electrical storms across the planet are diffused over the atmosphere of the entire Earth during both fair and stormy weather. Other contributing factors to electric current being present in the atmosphere may include cosmic rays penetrating and interacting with the earth's atmosphere, and also the migration of ions, as well as other effects yet to be fully studied.

Some of the ionization in the lower atmosphere is caused by airborne radioactive substances, primarily radon. In most places of the world, ions are formed at a rate of 5-10 pairs per cubic centimeter per second at sea level. With increasing altitude, cosmic radiation causes the ion production rate to increase. In areas with high radon exhalation from the soil (or building materials), the rate may be much higher.

Alpha-active materials are primarily responsible for the atmospheric ionization. Each alpha particle (for instance, from a decaying radon atom) will, over its range of some centimeters, create approximately 150,000-200,000 ion pairs.

While there is a large amount of usable energy available in the atmosphere, a method or apparatus for efficiently collecting that energy has not been forthcoming. Therefore, a heretofore unaddressed need exists in the industry to address the aforementioned deficiencies and inadequacies.

SUMMARY

Embodiments of the present disclosure provide systems and methods for collecting energy. Briefly described in architecture, one embodiment of the system, among others, can be implemented by a support structure wire elevated above a ground level, at least one collection fiber electrically connected to the support structure wire; a load electrically connected to the support structure wire; and a diode electrically connected between the load and at least one collection fiber.

Embodiments of the present disclosure can also be viewed as providing methods for collecting energy. In this regard, one embodiment of such a method, among others, can be broadly summarized by the following steps: suspending at least one collection fiber from a support structure wire elevated above ground level, the fiber electrically connected to the support structure wire; providing a load with an electrical connection to the support structure wire to draw current; and providing a diode electrically connected between the collection fiber and the load.

Other systems, methods, features, and advantages of the present disclosure will be or become apparent to one with skill in the art upon examination of the following drawings and detailed description. It is intended that all such additional systems, methods, features, and advantages be included within this description, be within the scope of the present disclosure, and be protected by the accompanying claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Many aspects of the disclosure can be better understood with reference to the following drawings. The components in the drawings are not necessarily to scale, emphasis instead being placed upon clearly illustrating the principles of the present disclosure. Moreover, in the drawings, like reference numerals designate corresponding parts throughout the several views.

FIG. **1** is a circuit diagram of a weather energy circuit.

FIG. **2** is a perspective view of an exemplary embodiment of many energy collectors elevated above ground by a structure.

FIG. **2A** is a side view of an energy collection fiber suspended from a support wire.

FIG. **2B** is a side view of an exemplary embodiment of an energy collection fiber suspended from a support wire and with an additional support member.

FIG. **2C** is a perspective view of a support structure for multiple energy collection fibers.

FIG. **2D** is a side view of an exemplary embodiment of a support structure for multiple energy collection fibers.

FIG. **2E** is a side view of a support structure for an energy collection fiber.

FIG. **2F** is a side view of an exemplary embodiment of a support structure for an energy collection fiber.

FIG. **2G** is a side view of a support structure for multiple energy collection fibers.

FIG. **3** is a circuit diagram of an exemplary embodiment of a circuit for the collection of energy.

FIG. **4** is a circuit diagram of an exemplary embodiment of a circuit for the collection of energy.

FIG. **5** is a circuit diagram of an exemplary embodiment of an energy collection circuit for powering a generator and motor.

FIG. **6** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy and using it for the production of hydrogen and oxygen.

FIG. **7** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy, and using it for driving a fuel cell.

FIG. **8** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy.

FIG. **9** is a flow diagram of an exemplary embodiment of collecting energy with a collection fiber.

DETAILED DESCRIPTION

Electric charges on conductors reside entirely on the external surface of the conductors, and tend to concentrate more around sharp points and edges than on flat surfaces. Therefore, an electric field received by a sharp conductive point may be much stronger than a field received by the same charge residing on a large smooth conductive shell. An exemplary embodiment of this disclosure takes advantage of this property, among others, to collect and use the energy generated by an electric field in the atmosphere. Referring to collection system **100** presented in FIG. **2**, at least one collection device **130** may be suspended from a support wire system **120** supported by poles **110**. Collection device **130** may comprise a diode or a collection fiber individually, or the combination of a diode and a collection fiber. Support wire system **120** may be electrically connected to load **150** by connecting wire **140**. Supporting wire system **120** may be any shape or pattern. Also, conducting wire **140** may be one wire or multiple wires. The collection device **130** in the form of a fiber may comprise any conducting or non-conducting material, including carbon, graphite, Teflon, and metal. An exemplary embodiment utilizes carbon or graphite fibers for static electricity collection. Support wire system **120** and connecting wire **140** can be made of any conducting material, including aluminum or steel, but most notably, copper. Teflon may be added to said conductor as well, such as non-limiting examples of a Teflon impregnated wire, a wire with a Teflon coating, or Teflon strips hanging from a wire. Conducting wire **120**, **140**, and **200** may be bare wire, or coated with insulation as a non-limiting example. Wires **120** and **140** are a means of transporting the energy collected by collection device **130**.

An exemplary embodiment of the collection fibers as collection device **130** includes graphite or carbon fibers. Graphite and carbon fibers, at a microscopic level, can have hundreds of thousands of points. Atmospheric electricity may be attracted to these points. If atmospheric electricity can follow two paths where one is a flat surface and the other is a pointy, conductive surface, the electrical charge will be attracted to the pointy, conductive surface. Generally, the more points that are present, the higher energy that can be gathered. Therefore, carbon, or graphite fibers are examples that demonstrate exemplary collection ability.

In at least one exemplary embodiment, the height of support wire **120** may be an important factor. The higher that collection device **130** is from ground, the larger the voltage potential between collection device **130** and electrical ground. The electric field may be more than 100 volts per meter under some conditions. When support wire **120** is suspended in the air at a particular altitude, wire **120** will itself collect a very small charge from ambient voltage. When collection device **130** is connected to support wire **120**, collection device **130** becomes energized and transfers the energy to support wire **120**.

A diode, not shown in FIG. **2**, may be connected in several positions in collection system **100**. A diode is a component that restricts the direction of movement of charge carriers. It allows an electric current to flow in one direction, but essentially blocks it in the opposite direction. A diode can be thought of as the electrical version of a check valve. The diode may be used to prevent the collected energy from discharging into the atmosphere through the collection fiber embodiment of collection device **130**. An exemplary embodiment of the collection device comprises the diode with no collection fiber. A preferred embodiment, however, includes a diode at the connection point of a collection fiber to support system **120** such that the diode is elevated above ground. Multiple diodes may be used between collection device **130** and load **150**. Additionally, in an embodiment with multiple fibers, the diodes restricts energy that may be collected through one fiber from escaping through another fiber.

Collection device **130** may be connected and arranged in relation to support wire system **120** by many means. Some non-limiting examples are provided in FIGS. **2A**-**2G** using a collection fiber embodiment. FIG. **2A** presents support wire **200** with connecting member **210** for collection device **130**. Connection member **210** may be any conducting material allowing for the flow of electricity from connection device **130** to support wire **200**. Then, as shown in FIG. **2**, the support wire **200** of support system **120** may be electrically connected through conducting wire **140** to load **150**. A plurality of diodes may be placed at any position on the support structure wire. A preferred embodiment places a diode at an elevated position at the connection point between a collection fiber embodiment of collection device **130** and connection member **210**.

Likewise, FIG. **2B** shows collection fiber **130** electrically connected to support wire **200** and also connected to support member **230**. Support member **230** may be connected to collection fiber **130** on either side. Support member **230** holds the fiber steady on both ends instead of letting it move freely. Support member **230** may be conducting or non-conducting. A plurality of diodes may be placed at any position on the support structure wire. A preferred embodiment places a diode at elevated position at the connection point between collection fiber **130** and support wire **200** or between fiber **130**, support member **230**, and support wire **200**.

FIG. **2C** presents multiple collection fibers in a squirrel cage arrangement with top and bottom support members. Support structure **250** may be connected to support structure wire **200** by support member **240**. Structure **250** has a top **260** and a bottom **270** and each of the multiple collection fibers **130** are connected on one end to top **260** and on the other end to bottom **270**. A plurality of diodes may be placed at any position on support structure **250**. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support structure wire **200**.

FIG. **2D** presents another exemplary embodiment of a support structure with support members **275** in an x-shape connected to support structure wire **200** at intersection **278** with collection fibers **130** connected between ends of support members **275**. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support wire **200**.

FIG. **2E** provides another exemplary embodiment for supporting collection fiber **130**. Collection fiber **130** may be connected on one side to support member **285**, which may be connected to support structure wire **200** in a first location and on the other side to support member **280**, which may be connected to support structure wire **200** in a second location on support structure wire **200**. The first and second locations may be the same location, or they may be different locations, even on different support wires. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places one or more diodes at elevated positions at the connection point(s) between collection fiber **130** and support wire **200**.

FIG. **2F** presents another exemplary embodiment of a support for a collection fiber. Two support members **290** may support either side of a collection fiber and are connected to support wire **200** in a single point. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support wire **200**.

FIG. **2G** provides two supports as provided in FIG. **2F** such that at least two support members **292**, **294** may be connected to support structure wire **200** in multiple locations and collection fibers **130** may be connected between each end of the support structures. Collection fibers **130** may be connected between each end of a single support structure and between multiple support structures. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places one or more diodes at elevated positions at the connection point(s) between collection fiber **130** and support structure wire **200**.

FIG. **3** provides a schematic diagram of storing circuit **300** for storing energy collected by one or more collection devices (**130** from FIG. **2**). Load **150** induces current flow. Diode **310** may be electrically connected in series between one or more collection devices (**130** from FIG. **2**) and load **150**. A plurality of diodes may be placed at any position in the circuit. Switch **330** may be electrically connected between load **150** and at least one collection device (**130** from FIG. **2**) to connect and disconnect the load. Capacitor **320** may be connected in parallel to the switch **330** and load **150** to store energy when switch **330** is open for delivery to load **150** when switch **330** is closed. Rectifier **340** may be electrically connected in parallel to load **150**, between the receiving end of switch **330** and ground. Rectifier **340** may be a full-wave or a half-wave rectifier. Rectifier **340** may include a diode electrically connected in parallel to load **150**, between the receiving end of switch **330** and ground. The direction of the diode of rectifier **340** is optional.

In an exemplary embodiment provided in FIG. **4**, storage circuit **400** stores energy from one or more collection devices (**130** from FIG. **2**) by charging capacitor **410**. If charging capacitor **410** is not used, then the connection to ground shown at capacitor **410** is eliminated. A plurality of diodes may be placed at any position in the circuit. Diode **310** may be electrically connected in series between one or more collection devices (**130** from FIG. **2**) and load **150**. Diode **440** may be placed in series with load **150**. The voltage from capacitor **410** can be used to charge spark gap **420** when it reaches sufficient voltage. Spark gap **420** may comprise one or more spark gaps in parallel. Non-limiting examples of spark gap **420** include mercury-reed switches and mercury-wetted reed switches. When spark gap **420** arcs, energy will arc from one end of the spark gap **420** to the receiving end of the spark gap **420**. The output of spark gap **420** may be electrically connected in series to rectifier **450**. Rectifier **450** may be a full-wave or a half-wave rectifier. Rectifier **450** may include a diode electrically connected in parallel to transformer **430** and load **150**, between the receiving end of spark gap **420** and ground. The direction of the diode of rectifier **450** is optional. The output of rectifier **450** is connected to transformer **430** to drive load **150**.

FIG. **5** presents motor driver circuit **500**. One or more collection devices (**130** from FIG. **2**) are electrically connected to static electricity motor **510**, which powers generator **520** to drive load **150**. A plurality of diodes may be placed at any position in the circuit. Motor **510** may also be directly connected to load **150** to drive it directly.

FIG. **6** demonstrates a circuit **600** for producing hydrogen. A plurality of diodes may be placed at any position in the circuit. One or more collection devices (**130** from FIG. **2**) are electrically connected to primary spark gap **610**, which may be connected to secondary spark gap **640**. Non-limiting examples of spark gaps **610**, **640** include mercury-reed switches and mercury-wetted reed switches. Secondary spark gap **640** may be immersed in water **630** within container **620**. When secondary spark gap **640** immersed in water **630** is energized, spark gap **640** may produce bubbles of hydrogen and oxygen, which may be collected to be used as fuel.

FIG. **7** presents circuit **700** for driving a fuel cell. A plurality of diodes may be placed at any position in the circuit. Collection devices (**130** from FIG. **2**) provide energy to fuel cell **720** which drives load **150**. Fuel cell **720** may produce hydrogen and oxygen.

FIG. **8** presents exemplary circuit **800** for the collection of energy. Storage circuit **800** stores energy from one or more collection devices (**130** from FIG. **2**) by charging capacitor **810**. If charging capacitor **810** is not used, then the connection to ground shown at capacitor **810** is eliminated. A plurality of diodes may be placed at any position in the circuit. The voltage from capacitor **810** can be used to charge spark gap **820** when it reaches sufficient voltage. Spark gap **820** may comprise one or more spark gaps in parallel or in series. Non-limiting examples of spark gap **820** include mercury-reed switches and mercury-wetted reed switches. When spark gap **820** arcs, energy will arc from one end of spark gap **820** to the receiving end of spark gap **820**. The output of spark gap **820** may be electrically connected in series to rectifier **825**. Rectifier **825** may be a full-wave or a half-wave rectifier. Rectifier **825** may include a diode electrically connected in parallel to inductor **830** and load **150**, between the receiving end of spark gap **820** and ground. The direction of the diode of rectifier **825** is optional. The output of rectifier **825** is connected to inductor **830**. Inductor **830** may be a fixed value inductor or a variable inductor. Capacitor **870** may be placed in parallel with load **150**.

FIG. **9** presents a flow diagram of a method for collecting energy. In block **910**, one or more collection devices may be suspended from a support structure wire. In block **920**, a load may be electrically connected to the support structure wire to draw current. In block **930** a diode may be electrically connected between the support structure wire and the electrical connection to the load. In block **940**, energy provided to the load may be stored or otherwise utilized.

Any process descriptions or blocks in flow charts should be understood as representing modules, segments, or portions of code which include one or more executable instructions for implementing specific logical functions or steps in the process, and alternate implementations are included within the scope of the preferred embodiment of the present disclosure in which functions may be executed out of order from that shown or discussed, including substantially concurrently or in reverse order, depending on the functionality involved, as would be understood by those reasonably skilled in the art of the present disclosure.

It should be emphasized that the above-described embodiments of the present disclosure, particularly, any "preferred" embodiments, are merely possible examples of implementations, merely set forth for a clear understanding of the principles of the disclosure. Many variations and modifications may be made to the above-described embodiment(s) of the disclosure without departing substantially from the spirit and principles of the disclosure. All such modifications and variations are intended to be included herein within the scope of this disclosure and the present disclosure and protected by the following claims.

Therefore, at least the following is claimed:

1. A method of collecting energy comprising:

suspending at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the at least one collection device electrically connected to the support structure;

providing a load with an electrical connection to the at least one collection device to passively draw current; and powering a fuel cell between the support structure and the load, the powering performed with energy collected from the at least one collection device.

2. The method of claim **1**, wherein the collection device comprises a diode.

3. The method of claim **1**, wherein the collection device comprises a collection fiber.

4. The method of claim **1**, wherein the collection device comprises a diode and a collection fiber and the diode is electrically connected between the collection fiber and the load.

5. The method of claim **1**, further comprising storing energy provided to the load.

6. The method of claim **5**, wherein storing energy provided to the load comprises storing energy in a capacitor or an inductor.

7. The method of claim **3**, wherein the collection fiber comprises carbon fiber or graphite fiber.

8. A system of energy collection comprising:

a support structure;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the collection device electrically connected to and suspended from the support structure, the at least one collection device configured to passively draw current;

a load electrically connected to the at least one collection device; and

a diode electrically connected between the at least one collection device and the support structure.

9. The system of claim **8**, wherein the collection device comprises a diode.

10. The system of claim **8**, wherein the collection device comprises a collection fiber.

11. The system of claim **8**, wherein the collection device comprises a collection fiber and a diode electrically connected between the load and the collection fiber.

12. The system of claim **9**, wherein the diode is elevated relative to the ground level.

13. The system of claim **10**, wherein the collection fiber comprises a carbon fiber or a graphite fiber.

14. The system of claim **8**, further comprising:

a switch connected in series between the at least one collection device and the load; and

a capacitor connected in parallel with the switch and the load.

15. The system of claim **14**, wherein the switch comprises an interrupter connected between the at least one collection device and the load, and wherein the interrupter comprises at least one of a fluorescent tube, a neon bulb, an AC light, and a spark gap.

16. The system of claim **15**, further comprising a transformer connected between the interrupter and the load.

17. The system of claim **8**, further comprising:

a motor for providing power, the motor connected between the at least one collection device and the load; and

a generator powered by the motor.

18. The system of claim **8**, further comprising a fuel cell between the support structure and the load.

19. The system of claim **18**, wherein the fuel cell produces hydrogen and oxygen.

20. A system of collecting energy comprising:

means for suspending at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the at least one collection device electrically connected to the means for suspending;

means for passively inducing current flow, the means for inducing current flow electrically connected to the means for suspending; and

means for restricting the backflow of charge carriers, the means for restricting the backflow of charge carriers electrically connected between the at least one collection device and the means for inducing current flow.

21. A system of energy collection comprising:

a support structure;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the collection device electrically connected to and suspended from the support structure, the at least one collection device configured to passively draw current;

a load electrically connected to the at least one collection device; and

a switch connected in series between the at least one collection device and the load, the switch comprising an interrupter connected between the at least one collection device and the load, and wherein the interrupter comprises at least one of a fluorescent tube, a neon bulb, an AC light, and a spark gap.

22. The system of claim **21**, further comprising a transformer connected between the interrupter and the load.

23. A system of energy collection comprising:

a support structure;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the at least one collection device electrically connected to and suspended from the support structure, the collection device configured to passively draw current;

a load electrically connected to the at least one collection device;

a motor for providing power, the motor connected between the at least one collection device and the load; and

a generator powered by the motor.

24. A system of energy collection comprising:

a support structure;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the at least one collection device electrically connected to and suspended from the support structure the at least one collection device configured to passively draw current;

a load electrically connected to the at least one collection device; and

a fuel cell between the support structure and the load.

25. The system of claim **24**, wherein the fuel cell produces hydrogen and oxygen.

* * * * *

US008810049B2

(12) **United States Patent**
McCowen

(10) **Patent No.: US 8,810,049 B2**
(45) **Date of Patent: Aug. 19, 2014**

(54) **ENERGY COLLECTION**

(75) Inventor: **Clint McCowen**, Navarre, FL (US)

(73) Assignee: **Ion Power Group, LLC**, Navarre, FL (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 88 days.

(21) Appl. No.: **13/569,133**

(22) Filed: **Aug. 7, 2012**

(65) **Prior Publication Data**

US 2012/0299559 A1 Nov. 29, 2012

Related U.S. Application Data

(63) Continuation-in-part of application No. 12/255,130, filed on Oct. 21, 2008, now Pat. No. 8,686,575, which is a continuation of application No. 11/358,264, filed on Feb. 21, 2006, now Pat. No. 7,439,712.

(51) **Int. Cl.**

F02B 63/04	(2006.01)
F03G 7/08	(2006.01)
H02K 7/18	(2006.01)
F03D 9/00	(2006.01)
H02P 9/04	(2006.01)

(52) **U.S. Cl.**
USPC .... **290/1 R**; 290/44; 290/55; 310/300; 310/310; 416/95

(58) **Field of Classification Search**
CPC Y02E 10/721; Y02E 10/70; Y02E 10/722; B32B 2262/101; C01B 31/00
USPC .... 290/1 R, 44, 55
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

674,427	A	5/1901	Palencsar	307/149
911,260	A	2/1909	Pennock	307/149
1,014,719	A	1/1912	Pennock	307/149
2,473,819	A	6/1949	Pittman	315/36
3,532,959	A	10/1970	Erickson	320/166
3,780,722	A	12/1973	Swet	126/680
3,946,227	A	3/1976	Bingham	250/281
4,079,225	A	3/1978	Warner	219/110
4,092,250	A	5/1978	Sano et al.	210/500.29
4,104,696	A	8/1978	Cochran, Jr.	361/229
4,146,800	A *	3/1979	Gregory et al.	290/44
4,180,056	A	12/1979	Schnabel et al.	126/660
4,201,197	A *	5/1980	Dismer	126/600
4,206,396	A *	6/1980	Marks	322/2 A
4,224,496	A	9/1980	Riordan et al.	219/110
4,307,936	A	12/1981	Ochiai	385/120

(Continued)

FOREIGN PATENT DOCUMENTS

WO	2007098341	8/2007

OTHER PUBLICATIONS

U.S. Appl. No. 12/637,720, Dec. 14, 2009, McCowen.

Primary Examiner — Pedro J Cuevas

(74) *Attorney, Agent, or Firm* — Benjamin A. Balser; Balser & Grell IP Law

(57) **ABSTRACT**

An energy collection system may collect and use the energy generated by an electric field. Collection fibers are suspended from a support wire system supported by poles. The support wire system is electrically connected to a load by a connecting wire. The collection fibers may be made of any conducting material, but carbon and graphite are preferred. Diodes may be used to restrict the backflow or loss of energy.

20 Claims, 11 Drawing Sheets

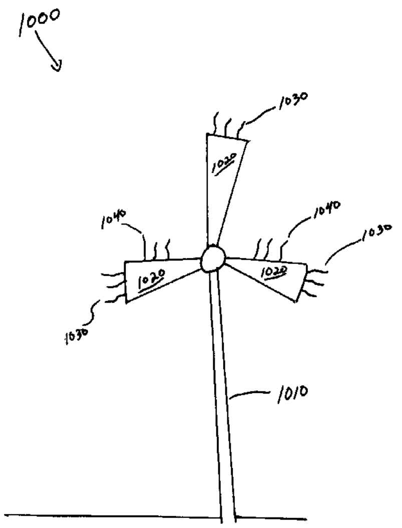

(56) References Cited

U.S. PATENT DOCUMENTS

4,314,192	A	2/1982	Caro	322/2 R
4,346,478	A	8/1982	Sichling	398/107
4,425,905	A	1/1984	Mori	126/578
4,433,248	A	2/1984	Marks	290/44
RE31,678	E	9/1984	Ochiai	385/115
4,483,311	A	11/1984	Whitaker	126/602
4,489,269	A	12/1984	Edling et al.	322/2 R
4,512,335	A	4/1985	Mori	126/678
4,653,223	A	3/1987	Mori	47/17
4,653,472	A	3/1987	Mori	126/683
4,676,226	A	6/1987	Mori	126/680
4,809,675	A	3/1989	Mori	126/578
4,852,454	A	8/1989	Batchelder	89/1.11
4,943,125	A	7/1990	Laundre’ et al.	385/35
4,988,159	A	1/1991	Turner et al.	385/33
5,047,892	A	9/1991	Sakata et al.	361/231
5,114,101	A	5/1992	Stern et al.	244/172.8
5,145,257	A	9/1992	Bryant et al.	374/131
5,379,103	A	1/1995	Zigler	356/73
5,851,309	A	12/1998	Kousa	136/248
5,942,806	A	8/1999	Veliadis	290/1 R
6,025,591	A	2/2000	Taylor et al.	250/292
6,038,363	A	3/2000	Slater et al.	385/147
6,116,544	A	9/2000	Forward et al.	244/158.2
6,173,922	B1	1/2001	Hoyt et al.	244/158.2
6,226,440	B1	5/2001	Lyons	385/119
6,419,191	B1	7/2002	Hoyt et al.	244/158.2
6,425,391	B1	7/2002	Davoren et al.	126/683
6,457,943	B1 *	10/2002	Olsen et al.	416/230
6,735,830	B1	5/2004	Merciel	250/324
6,765,212	B2	7/2004	Faus et al.	250/339.11
6,846,447	B2	1/2005	Mbachu et al.	264/406
6,853,447	B2	2/2005	Goetz	356/237.1
6,894,772	B2	5/2005	Goetz et al.	356/237.1
6,925,852	B2	8/2005	Susko	73/23.2
6,974,110	B2	12/2005	Grandics	244/171.5
7,109,597	B1	9/2006	Bose	290/43
7,128,867	B2	10/2006	Mbachu et al.	264/406
7,231,809	B2	6/2007	Susko	73/31.05
7,259,475	B2	8/2007	Hong et al.	307/64
7,303,373	B2 *	12/2007	Viertl	416/1
7,439,712	B2 *	10/2008	McCowen	322/2 A
7,449,668	B2	11/2008	Schutten et al.	250/214 SW
7,478,712	B2 *	1/2009	McCowen	191/45 R
7,532,819	B1	5/2009	Triebes et al.	398/121
7,637,462	B2 *	12/2009	Pal	244/199.2
8,115,333	B2 *	2/2012	Parsche et al.	290/55
8,118,559	B2 *	2/2012	Llorente Gonzalez et al.	416/226
8,342,805	B2 *	1/2013	Mendez Hernandez et al.	416/146 R
8,519,596	B1 *	8/2013	Kim et al.	310/310
8,629,570	B1 *	1/2014	Kamenov	290/44
8,662,853	B2 *	3/2014	Vasudeva et al.	416/226
8,686,575	B2 *	4/2014	McCowen	290/1 R
2002/0026933	A1	3/2002	Gottlieb	126/628
2002/0108892	A1	8/2002	Goetz et al.	209/576
2002/0109094	A1	8/2002	Goetz et al.	250/339.11
2002/0109835	A1	8/2002	Goetz	356/237.1
2003/0125630	A1	7/2003	Furnish	600/461
2003/0193319	A1	10/2003	Wood et al.	322/2 A
2004/0011925	A1	1/2004	Grandics	244/62
2004/0083793	A1	5/2004	Susko	73/31.05
2004/0094853	A1	5/2004	Mbachu et al.	264/40.1
2004/0160711	A1	8/2004	Stumberger	361/20
2004/0212945	A1	10/2004	Sprenger et al.	361/212
2005/0101024	A1	5/2005	Mbachu et al.	436/85
2005/0136311	A1	6/2005	Ueda et al.	700/297
2005/0270525	A1	12/2005	Susko	356/326
2006/0051207	A1 *	3/2006	Becerra et al.	416/5
2007/0098551	A1 *	5/2007	Viertl	416/61
2007/0107765	A1	5/2007	Schutten et al.	136/243
2007/0195481	A1 *	8/2007	McCowen	361/233
2007/0217918	A1 *	9/2007	Baker et al.	416/227 R
2007/0218323	A1	9/2007	Sudo et al.	429/413
2007/0223171	A1	9/2007	Guy et al.	361/126
2007/0252047	A1 *	11/2007	Pal	244/35 R
2007/0273206	A1 *	11/2007	McCowen	307/9.1
2007/0297892	A1	12/2007	Kildegaard	415/1
2008/0145714	A1	6/2008	Kagami	429/413
2009/0040680	A1	2/2009	McCowen	361/233
2009/0107842	A1	4/2009	Park et al.	204/554
2009/0114495	A1	5/2009	McCowen	191/45 R
2009/0216292	A1	8/2009	Pless et al.	607/33
2010/0090562	A1	4/2010	McCowen	310/309
2010/0329881	A1 *	12/2010	Mendez Hernandez et al.	416/230
2012/0134826	A1 *	5/2012	Arocena De La Rua et al.	416/146 R
2012/0312918	A1 *	12/2012	Heppe	244/30
2013/0028739	A1 *	1/2013	Erichsen et al.	416/146 R
2013/0032671	A1 *	2/2013	Giles et al.	244/134 R
2013/0037650	A1 *	2/2013	Heppe	244/2
2013/0136598	A1 *	5/2013	Loewe et al.	416/95
2013/0170992	A1 *	7/2013	Loewe et al.	416/95
2013/0334824	A1 *	12/2013	Freda	290/55
2014/0084748	A1 *	3/2014	Wang et al.	310/300
2014/0086748	A1 *	3/2014	Peltola et al.	416/95
2014/0099208	A1 *	4/2014	Janiuk	416/210 R

* cited by examiner

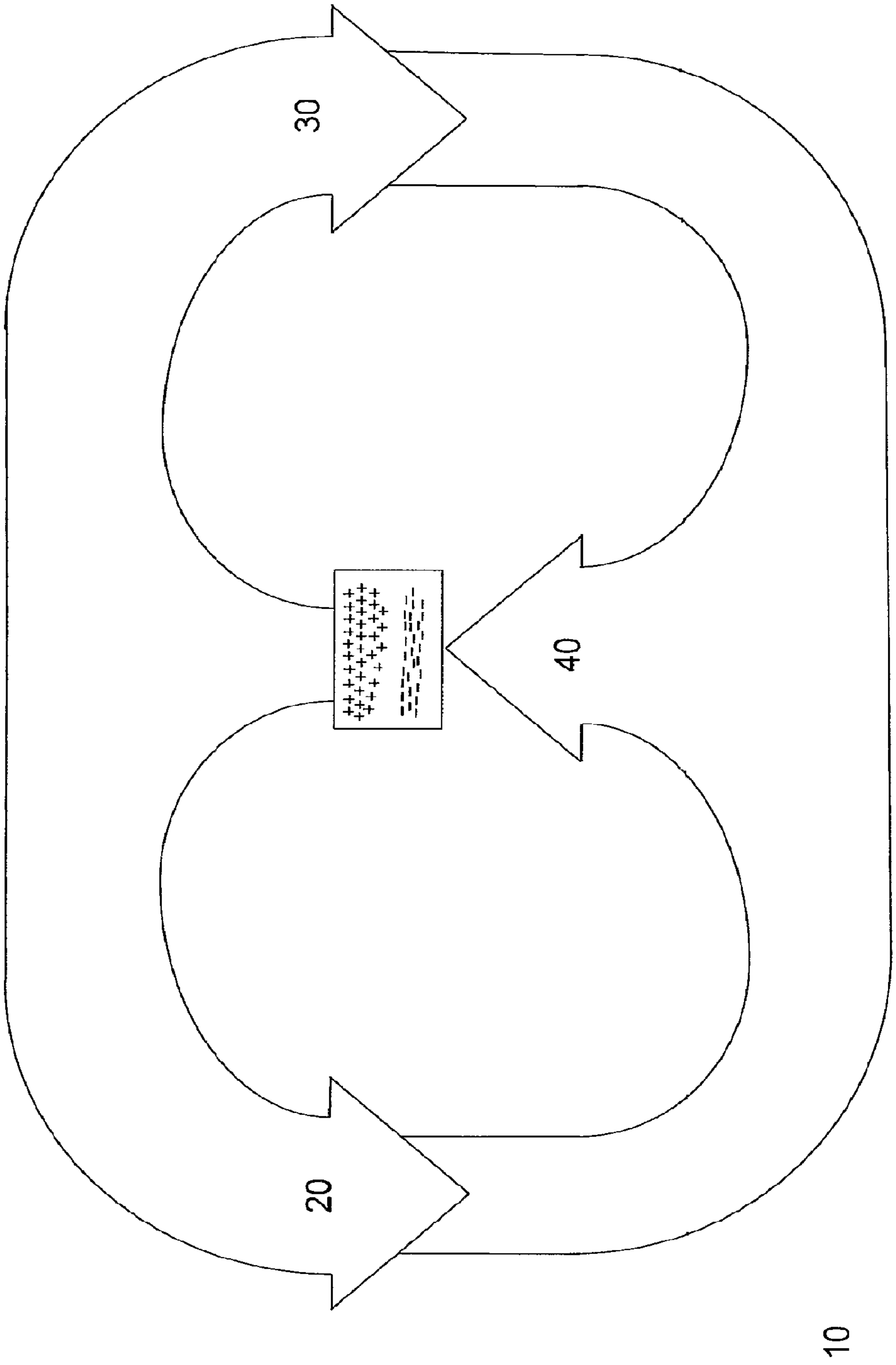

FIGURE 1

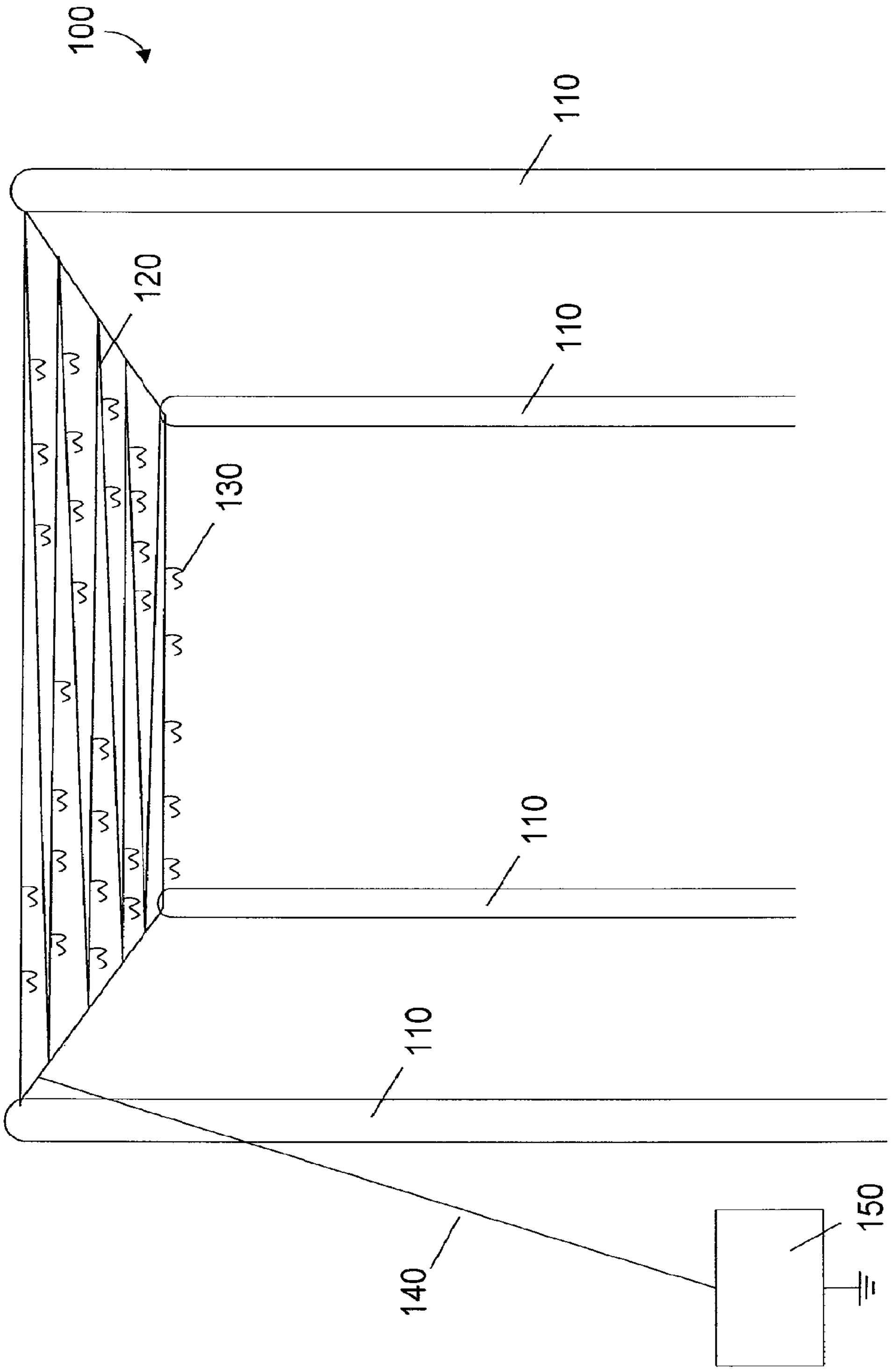
100
110
110
120
130
110
110
140
150

FIGURE 2

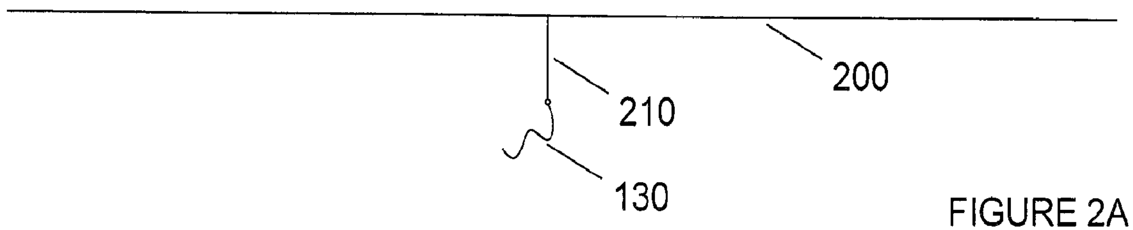

FIGURE 2A

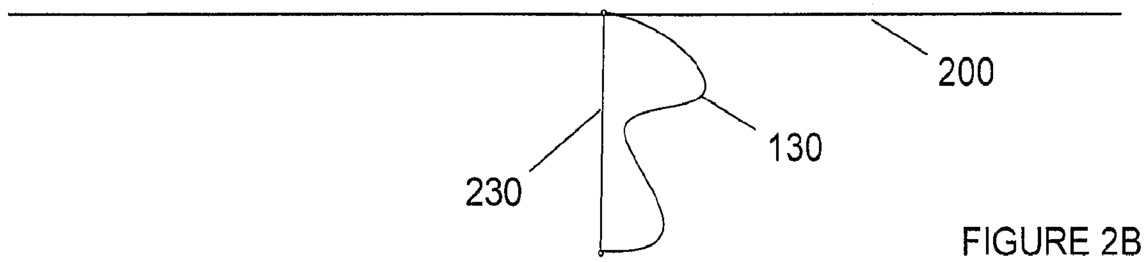

FIGURE 2B

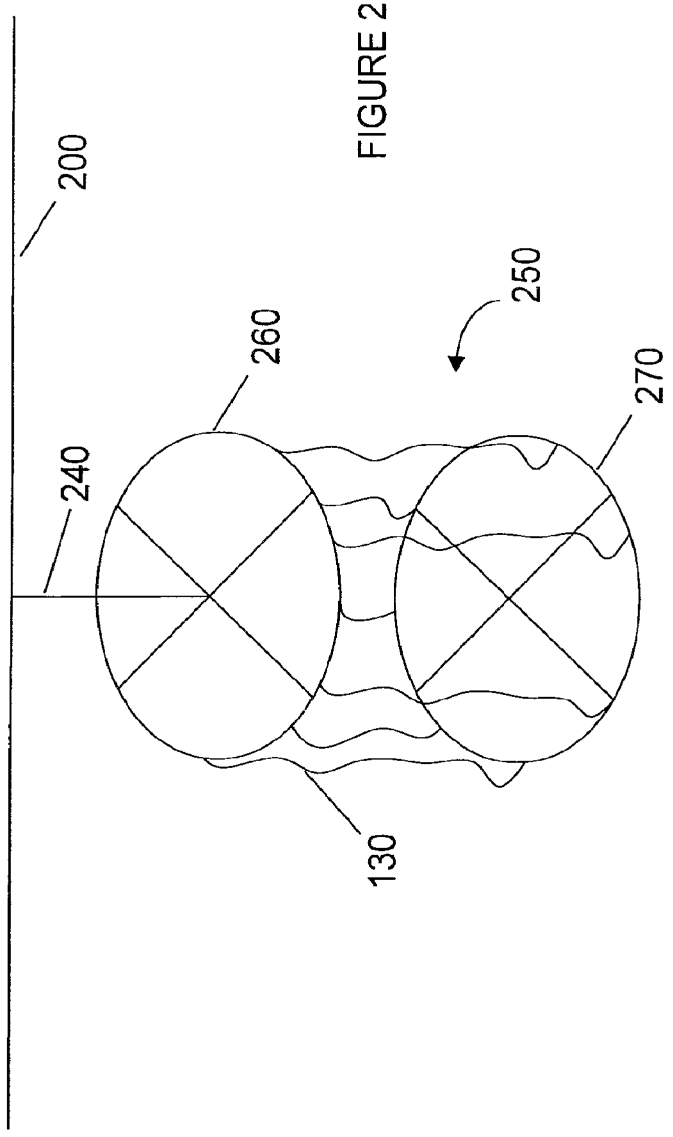
200
250
260
270
240
130

FIGURE 2C

FIGURE 2D

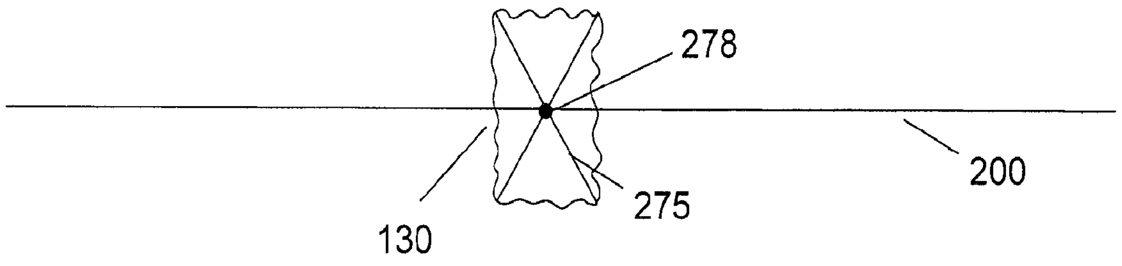
278
200
275
130

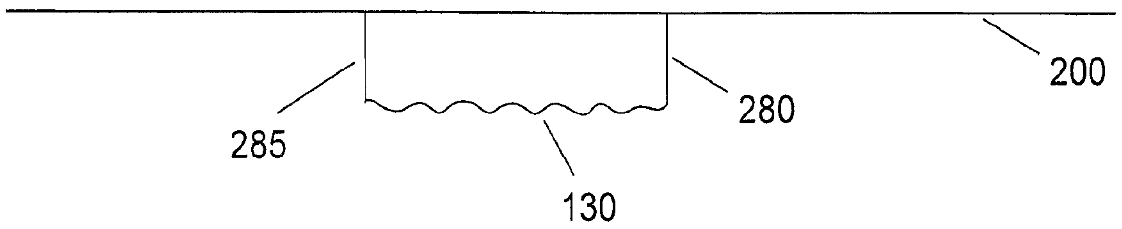
200
285
280
130

FIGURE 2E

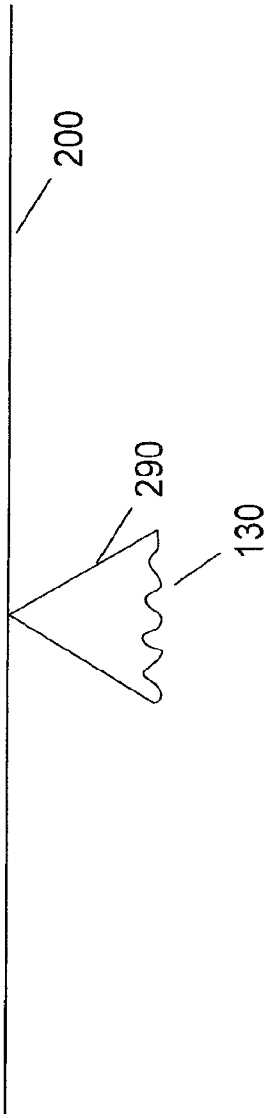

FIGURE 2F

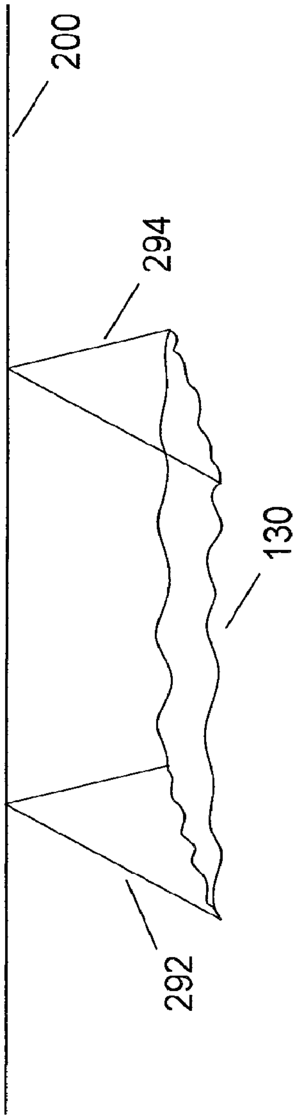
200
294
130
292

FIGURE 2G

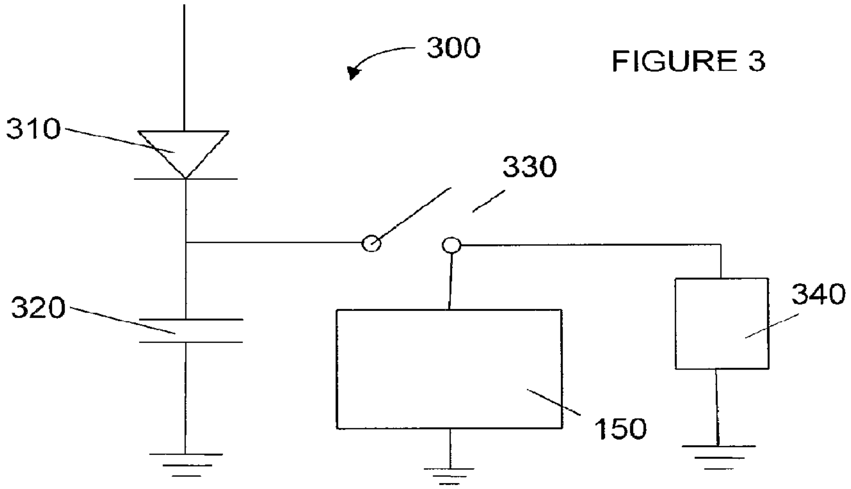
FIGURE 3
300
310
330
320
340
150

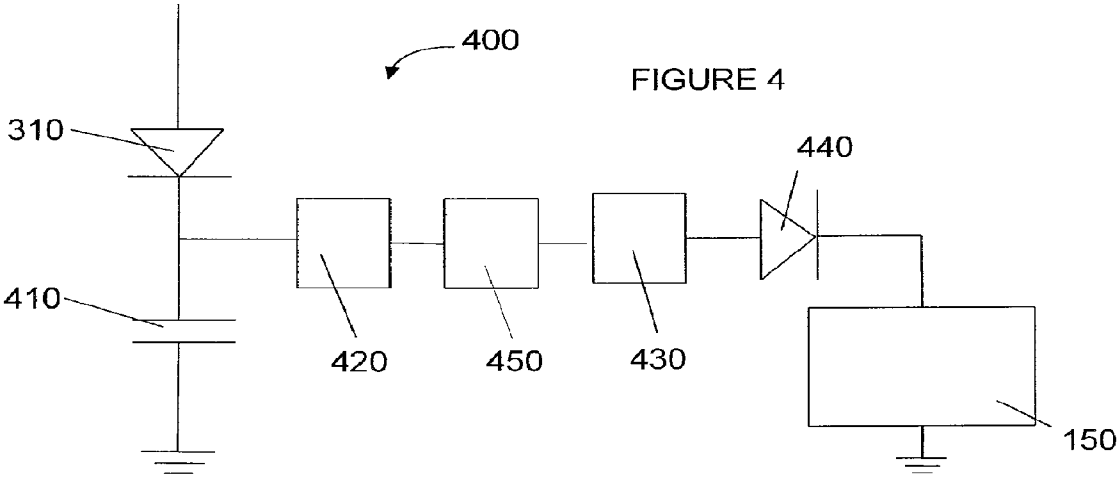
FIGURE 4
400
310
440
410
420
450
430
150

FIGURE 5

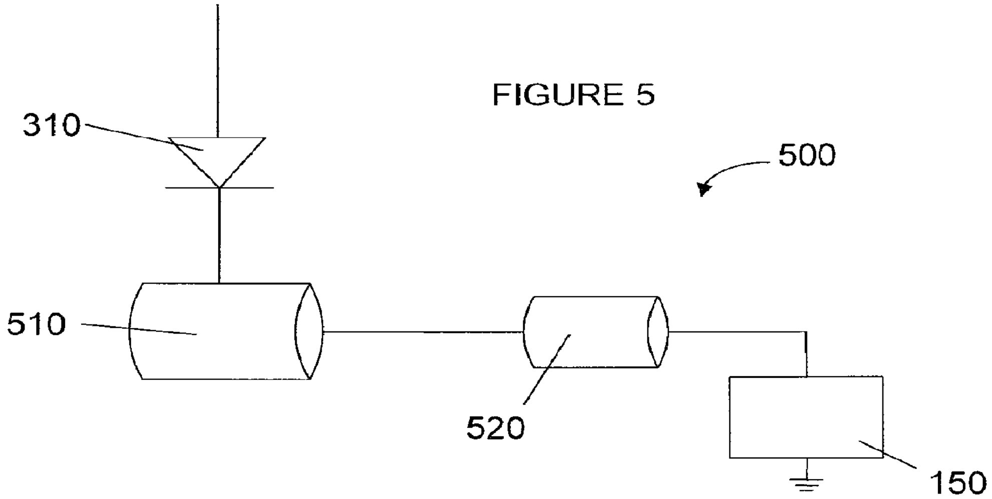
310
500
510
520
150

FIGURE 6

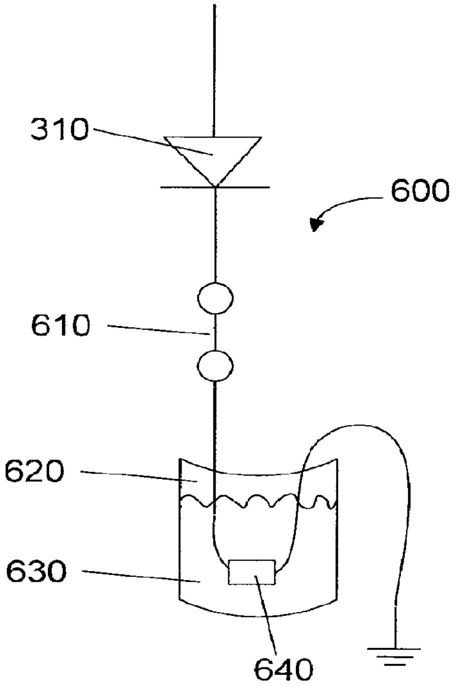
310
600
610
620
630
640

FIGURE 7

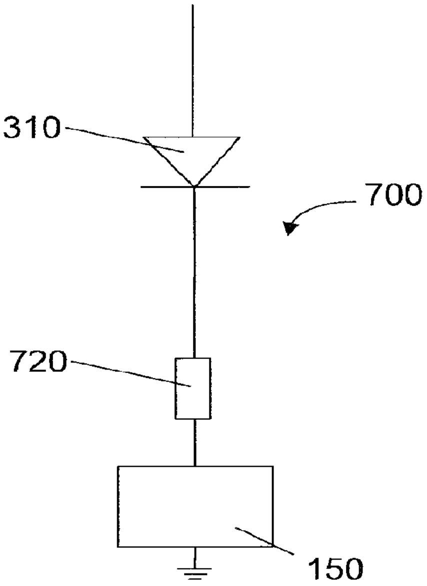
310
700
720
150

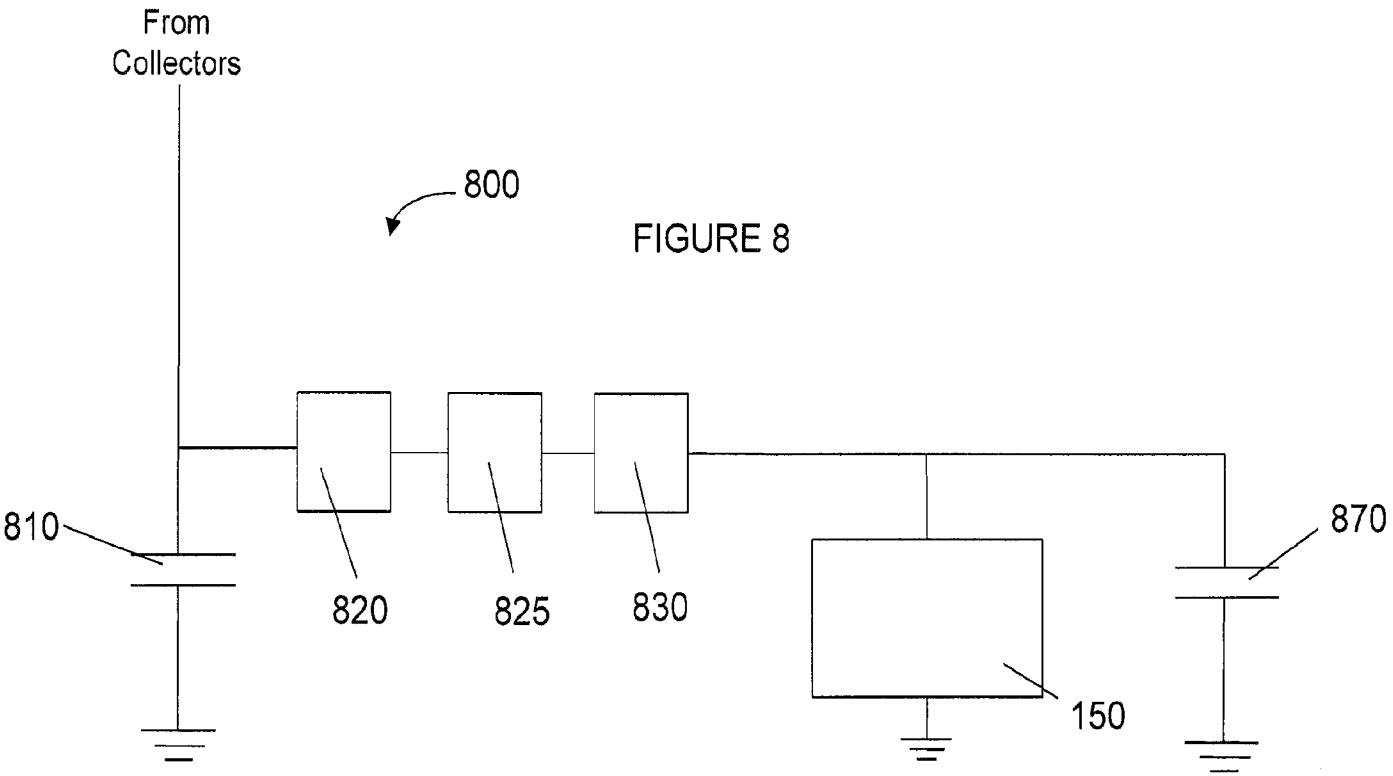
From
Collectors
800
810
820
825
830
150
870

FIGURE 8

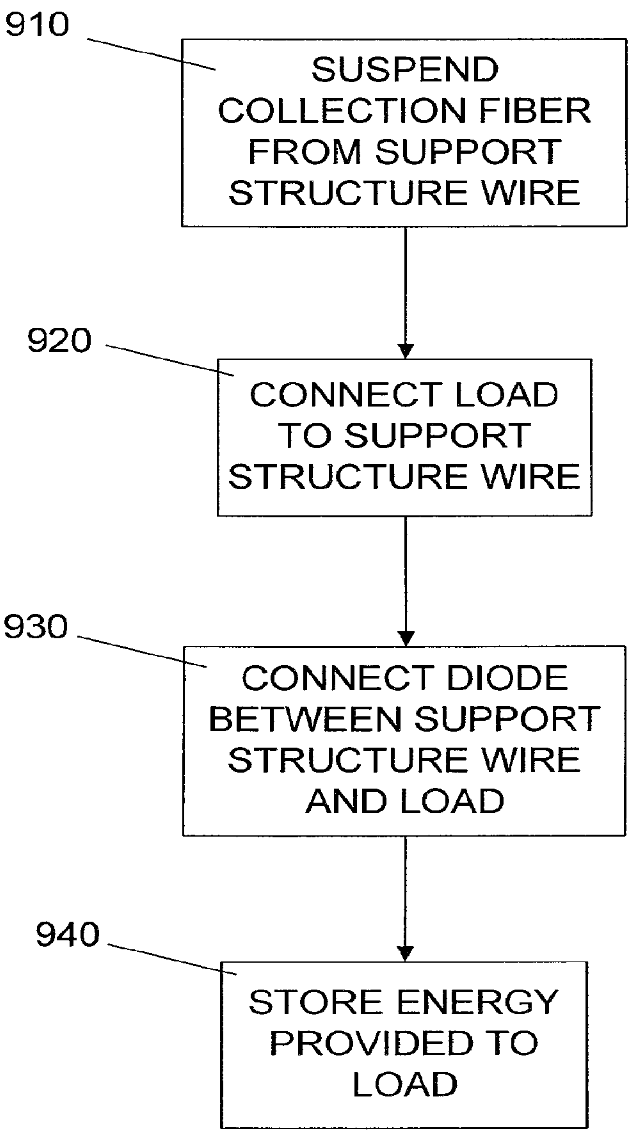
910
SUSPEND COLLECTION FIBER FROM SUPPORT STRUCTURE WIRE
920
CONNECT LOAD TO SUPPORT STRUCTURE WIRE
930
CONNECT DIODE BETWEEN SUPPORT STRUCTURE WIRE AND LOAD
940
STORE ENERGY PROVIDED TO LOAD

FIGURE 9

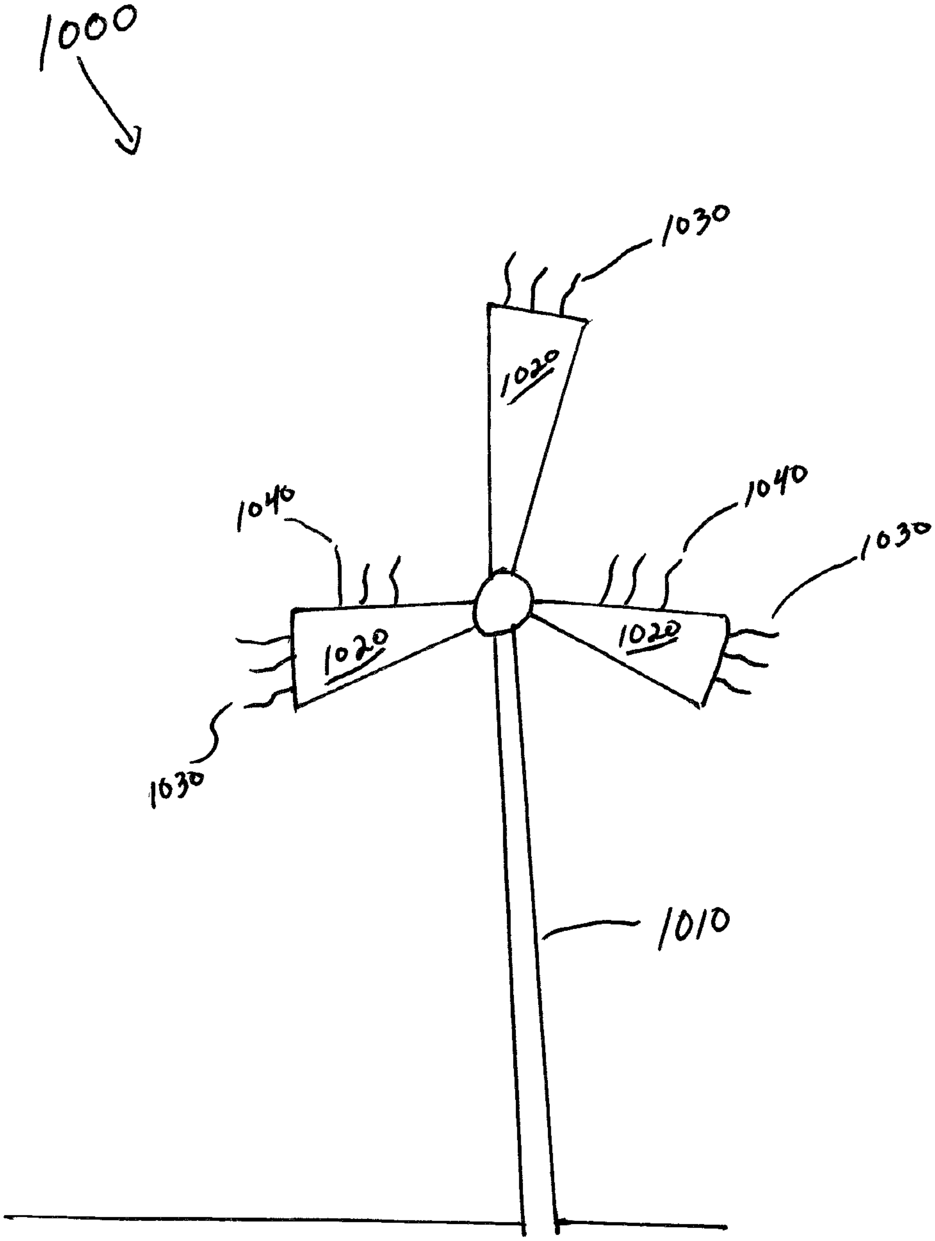
1000
1030
1020
1040
1040
1030
1020
1020
1030
1010

FIG. 10

ENERGY COLLECTION

CROSS REFERENCE TO RELATED APPLICATIONS

This application is a continuation in part application of U.S. patent application Ser. No. 12/255,130 filed on Oct. 21, 2008, which is a continuation application of U.S. patent application Ser. No. 11/358,264, filed on Feb. 21, 2006, which are both incorporated by reference herein.

TECHNICAL FIELD

The present disclosure is generally related to energy and, more particularly, is related to systems and methods for collecting energy.

BACKGROUND

The concept of fair weather electricity deals with the electric field and the electric current in the atmosphere propagated by the conductivity of the air. Clear, calm air carries an electrical current, which is the return path for thousands of lightening storms simultaneously occurring at any given moment around the earth. For simplicity, this energy may be referred to as static electricity or static energy. FIG. **1** illustrates a weather circuit for returning the current from lightning, for example, back to ground **10**. Weather currents **20**, **30** return the cloud to ground current **40**.

In a lightening storm, an electrical charge is built up, and electrons arc across a gas, ionizing it and producing the lightening flash. As one of ordinary skill in the art understands, the complete circuit requires a return path for the lightening flash. The atmosphere is the return path for the circuit. The electric field due to the atmospheric return path is relatively weak at any given point because the energy of thousands of electrical storms across the planet are diffused over the atmosphere of the entire Earth during both fair and stormy weather. Other contributing factors to electric current being present in the atmosphere may include cosmic rays penetrating and interacting with the earth's atmosphere, and also the migration of ions, as well as other effects yet to be fully studied.

Some of the ionization in the lower atmosphere is caused by airborne radioactive substances, primarily radon. In most places of the world, ions are formed at a rate of 5-10 pairs per cubic centimeter per second at sea level. With increasing altitude, cosmic radiation causes the ion production rate to increase. In areas with high radon exhalation from the soil (or building materials), the rate may be much higher.

Alpha-active materials are primarily responsible for the atmospheric ionization. Each alpha particle (for instance, from a decaying radon atom) will, over its range of some centimeters, create approximately 150,000-200,000 ion pairs.

While there is a large amount of usable energy available in the atmosphere, a method or apparatus for efficiently collecting that energy has not been forthcoming. Therefore, a heretofore unaddressed need exists in the industry to address the aforementioned deficiencies and inadequacies.

SUMMARY

Embodiments of the present disclosure provide systems and methods for collecting energy. Briefly described in architecture, one embodiment of the system, among others, can be implemented by a support structure wire elevated above a ground level, at least one collection fiber electrically connected to the support structure wire; a load electrically connected to the support structure wire; and a diode electrically connected between the load and at least one collection fiber.

Embodiments of the present disclosure can also be viewed as providing methods for collecting energy. In this regard, one embodiment of such a method, among others, can be broadly summarized by the following steps: suspending at least one collection fiber from a support structure wire elevated above ground level, the fiber electrically connected to the support structure wire; providing a load with an electrical connection to the support structure wire to draw current; and providing a diode electrically connected between the collection fiber and the load.

Other systems, methods, features, and advantages of the present disclosure will be or become apparent to one with skill in the art upon examination of the following drawings and detailed description. It is intended that all such additional systems, methods, features, and advantages be included within this description, be within the scope of the present disclosure, and be protected by the accompanying claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Many aspects of the disclosure can be better understood with reference to the following drawings. The components in the drawings are not necessarily to scale, emphasis instead being placed upon clearly illustrating the principles of the present disclosure. Moreover, in the drawings, like reference numerals designate corresponding parts throughout the several views.

FIG. **1** is a circuit diagram of a weather energy circuit.

FIG. **2** is a perspective view of an exemplary embodiment of many energy collectors elevated above ground by a structure.

FIG. **2**A is a side view of an energy collection fiber suspended from a support wire.

FIG. **2**B is a side view of an exemplary embodiment of an energy collection fiber suspended from a support wire and with an additional support member.

FIG. **2**C is a perspective view of a support structure for multiple energy collection fibers.

FIG. **2**D is a side view of an exemplary embodiment of a support structure for multiple energy collection fibers.

FIG. **2**E is a side view of a support structure for an energy collection fiber.

FIG. **2**F is a side view of an exemplary embodiment of a support structure for an energy collection fiber.

FIG. **2**G is a side view of a support structure for multiple energy collection fibers.

FIG. **3** is a circuit diagram of an exemplary embodiment of a circuit for the collection of energy.

FIG. **4** is a circuit diagram of an exemplary embodiment of a circuit for the collection of energy.

FIG. **5** is a circuit diagram of an exemplary embodiment of an energy collection circuit for powering a generator and motor.

FIG. **6** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy and using it for the production of hydrogen and oxygen.

FIG. **7** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy, and using it for driving a fuel cell.

FIG. **8** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy.

FIG. **9** is a flow diagram of an exemplary embodiment of collecting energy with a collection fiber.

FIG. **10** is a system diagram of an exemplary embodiment of a windmill with energy collectors.

DETAILED DESCRIPTION

Electric charges on conductors reside entirely on the external surface of the conductors, and tend to concentrate more around sharp points and edges than on flat surfaces. Therefore, an electric field received by a sharp conductive point may be much stronger than a field received by the same charge residing on a large smooth conductive shell. An exemplary embodiment of this disclosure takes advantage of this property, among others, to collect and use the energy generated by an electric field in the atmosphere. Referring to collection system **100** presented in FIG. **2**, at least one collection device **130** may be suspended from a support wire system **120** supported by poles **110**. Collection device **130** may comprise a diode or a collection fiber individually, or the combination of a diode and a collection fiber. Support wire system **120** may be electrically connected to load **150** by connecting wire **140**. Supporting wire system **120** may be any shape or pattern. Also, conducting wire **140** may be one wire or multiple wires. The collection device **130** in the form of a fiber may comprise any conducting or non-conducting material, including carbon, graphite, Teflon, and metal. An exemplary embodiment utilizes carbon or graphite fibers for static electricity collection. Support wire system **120** and connecting wire **140** can be made of any conducting material, including aluminum or steel, but most notably, copper. Teflon may be added to said conductor as well, such as non-limiting examples of a Teflon impregnated wire, a wire with a Teflon coating, or Teflon strips hanging from a wire. Conducting wire **120**, **140**, and **200** may be bare wire, or coated with insulation as a non-limiting example. Wires **120** and **140** are a means of transporting the energy collected by collection device **130**.

An exemplary embodiment of the collection fibers as collection device **130** includes graphite or carbon fibers. Graphite and carbon fibers, at a microscopic level, can have hundreds of thousands of points. Atmospheric electricity may be attracted to these points. If atmospheric electricity can follow two paths where one is a flat surface and the other is a pointy, conductive surface, the electrical charge will be attracted to the pointy, conductive surface. Generally, the more points that are present, the higher energy that can be gathered. Therefore, carbon, or graphite fibers are examples that demonstrate exemplary collection ability.

In at least one exemplary embodiment, the height of support wire **120** may be an important factor. The higher that collection device **130** is from ground, the larger the voltage potential between collection device **130** and electrical ground. The electric field may be more than 100 volts per meter under some conditions. When support wire **120** is suspended in the air at a particular altitude, wire **120** will itself collect a very small charge from ambient voltage. When collection device **130** is connected to support wire **120**, collection device **130** becomes energized and transfers the energy to support wire **120**.

A diode, not shown in FIG. **2**, may be connected in several positions in collection system **100**. A diode is a component that restricts the direction of movement of charge carriers. It allows an electric current to flow in one direction, but essentially blocks it in the opposite direction. A diode can be thought of as the electrical version of a check valve. The diode may be used to prevent the collected energy from discharging into the atmosphere through the collection fiber embodiment of collection device **130**. An exemplary embodiment of the collection device comprises the diode with no collection fiber. A preferred embodiment, however, includes a diode at the connection point of a collection fiber to support system **120** such that the diode is elevated above ground. Multiple diodes may be used between collection device **130** and load **150**. Additionally, in an embodiment with multiple fibers, the diodes restricts energy that may be collected through one fiber from escaping through another fiber.

Collection device **130** may be connected and arranged in relation to support wire system **120** by many means. Some non-limiting examples are provided in FIGS. **2**A-**2**G using a collection fiber embodiment. FIG. **2**A presents support wire **200** with connecting member **210** for collection device **130**. Connection member **210** may be any conducting material allowing for the flow of electricity from connection device **130** to support wire **200**. Then, as shown in FIG. **2**, the support wire **200** of support system **120** may be electrically connected through conducting wire **140** to load **150**. A plurality of diodes may be placed at any position on the support structure wire. A preferred embodiment places a diode at an elevated position at the connection point between a collection fiber embodiment of collection device **130** and connection member **210**.

Likewise, FIG. **2**B shows collection fiber **130** electrically connected to support wire **200** and also connected to support member **230**. Support member **230** may be connected to collection fiber **130** on either side. Support member **230** holds the fiber steady on both ends instead of letting it move freely. Support member **230** may be conducting or non-conducting. A plurality of diodes may be placed at any position on the support structure wire. A preferred embodiment places a diode at elevated position at the connection point between collection fiber **130** and support wire **200** or between fiber **130**, support member **230**, and support wire **200**.

FIG. **2**C presents multiple collection fibers in a squirrel cage arrangement with top and bottom support members. Support structure **250** may be connected to support structure wire **200** by support member **240**. Structure **250** has a top **260** and a bottom **270** and each of the multiple collection fibers **130** are connected on one end to top **260** and on the other end to bottom **270**. A plurality of diodes may be placed at any position on support structure **250**. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support structure wire **200**.

FIG. **2**D presents another exemplary embodiment of a support structure with support members **275** in an x-shape connected to support structure wire **200** at intersection **278** with collection fibers **130** connected between ends of support members **275**. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support wire **200**.

FIG. **2**E provides another exemplary embodiment for supporting collection fiber **130**. Collection fiber **130** may be connected on one side to support member **285**, which may be connected to support structure wire **200** in a first location and on the other side to support member **280**, which may be connected to support structure wire **200** in a second location on support structure wire **200**. The first and second locations may be the same location, or they may be different locations, even on different support wires. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places one or more diodes at elevated positions at the connection point(s) between collection fiber **130** and support wire **200**.

FIG. **2**F presents another exemplary embodiment of a support for a collection fiber. Two support members **290** may support either side of a collection fiber and are connected to support wire **200** in a single point. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support wire **200**.

FIG. **2**G provides two supports as provided in FIG. **2**F such that at least two support members **292**, **294** may be connected to support structure wire **200** in multiple locations and collection fibers **130** may be connected between each end of the support structures. Collection fibers **130** may be connected between each end of a single support structure and between multiple support structures. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places one or more diodes at elevated positions at the connection point(s) between collection fiber **130** and support structure wire **200**.

FIG. **3** provides a schematic diagram of storing circuit **300** for storing energy collected by one or more collection devices (**130** from FIG. **2**). Load **150** induces current flow. Diode **310** may be electrically connected in series between one or more collection devices (**130** from FIG. **2**) and load **150**. A plurality of diodes may be placed at any position in the circuit. Switch **330** may be electrically connected between load **150** and at least one collection device (**130** from FIG. **2**) to connect and disconnect the load. Capacitor **320** maybe connected in parallel to the switch **330** and load **150** to store energy when switch **330** is open for delivery to load **150** when switch **330** is closed. Rectifier **340** may be electrically connected in parallel to load **150**, between the receiving end of switch **330** and ground. Rectifier **340** may be a full-wave or a half-wave rectifier. Rectifier **340** may include a diode electrically connected in parallel to load **150**, between the receiving end of switch **330** and ground. The direction of the diode of rectifier **340** is optional.

In an exemplary embodiment provided in FIG. **4**, storage circuit **400** stores energy from one or more collection devices (**130** from FIG. **2**) by charging capacitor **410**. If charging capacitor **410** is not used, then the connection to ground shown at capacitor **410** is eliminated. A plurality of diodes may be placed at any position in the circuit. Diode **310** may be electrically connected in series between one or more collection devices (**130** from FIG. **2**) and load **150**. Diode **440** may be placed in series with load **150**. The voltage from capacitor **410** can be used to charge spark gap **420** when it reaches sufficient voltage. Spark gap **420** may comprise one or more spark gaps in parallel. Non-limiting examples of spark gap **420** include mercury-reed switches and mercury-wetted reed switches. When spark gap **420** arcs, energy will arc from one end of the spark gap **420** to the receiving end of the spark gap **420**. The output of spark gap **420** may be electrically connected in series to rectifier **450**. Rectifier **450** may be a full-wave or a half-wave rectifier. Rectifier **450** may include a diode electrically connected in parallel to transformer **430** and load **150**, between the receiving end of spark gap **420** and ground. The direction of the diode of rectifier **450** is optional. The output of rectifier **450** is connected to transformer **430** to drive load **150**.

FIG. **5** presents motor driver circuit **500**. One or more collection devices (**130** from FIG. **2**) are electrically connected to static electricity motor **510**, which powers generator **520** to drive load **150**. A plurality of diodes may be placed at any position in the circuit. Motor **510** may also be directly connected to load **150** to drive it directly.

FIG. **6** demonstrates a circuit **600** for producing hydrogen. A plurality of diodes maybe placed at any position in the circuit. One or more collection devices (**130** from FIG. **2**) are electrically connected to primary spark gap **610**, which may be connected to secondary spark gap **640**. Non-limiting examples of spark gaps **610**, **640** include mercury-reed switches and mercury-wetted reed switches. Secondary spark gap **640** may be immersed in water **630** within container **620**. When secondary spark gap **640** immersed in water **630** is energized, spark gap **640** may produce bubbles of hydrogen and oxygen, which may be collected to be used as fuel.

FIG. **7** presents circuit **700** for driving a fuel cell. A plurality of diodes may be placed at any position in the circuit. Collection devices (**130** from FIG. **2**) provide energy to fuel cell **720** which drives load **150**. Fuel cell **720** may produce hydrogen and oxygen.

FIG. **8** presents exemplary circuit **800** for the collection of energy. Storage circuit **800** stores energy from one or more collection devices (**130** from FIG. **2**) by charging capacitor **810**. If charging capacitor **810** is not used, then the connection to ground shown at capacitor **810** is eliminated. A plurality of diodes may be placed at any position in the circuit. The voltage from capacitor **810** can be used to charge spark gap **820** when it reaches sufficient voltage. Spark gap **820** may comprise one or more spark gaps in parallel or in series. Non-limiting examples of spark gap **820** include mercury-reed switches and mercury-wetted reed switches. When spark gap **820** arcs, energy will arc from one end of spark gap **820** to the receiving end of spark gap **820**. The output of spark gap **820** may be electrically connected in series to rectifier **825**. Rectifier **825** may be a full-wave or a half-wave rectifier. Rectifier **825** may include a diode electrically connected in parallel to inductor **830** and load **150**, between the receiving end of spark gap **820** and ground. The direction of the diode of rectifier **825** is optional. The output of rectifier **825** is connected to inductor **830**. Inductor **830** may be a fixed value inductor or a variable inductor. Capacitor **870** may be placed in parallel with load **150**.

FIG. **9** presents a flow diagram of a method for collecting energy. In block **910**, one or more collection devices may be suspended from a support structure wire. In block **920**, a load may be electrically connected to the support structure wire to draw current. In block **930** a diode may be electrically connected between the support structure wire and the electrical connection to the load. In block **940**, energy provided to the load may be stored or otherwise utilized.

A windmill is an engine powered by the energy of wind to produce alternative forms of energy. They may, for example, be implemented as small tower mounted wind engines used to pump water on farms. The modern wind power machines used for generating electricity are more properly called wind turbines. Common applications of windmills are grain milling, water pumping, threshing, and saw mills. Over the ages, windmills have evolved into more sophisticated and efficient wind-powered water pumps and electric power generators. In an example embodiment, as provided in FIG. **10**, windmill tower **1000** of suitable height and/or propeller **1020** of windmill tower **1000** may be equipped with energy collecting fibers **1030**, **1040**. Collecting fibers **1030**, **1040** may turn windmill **1000** into a power producing asset even when there is not enough wind to turn propellers **1020**. During periods when there is enough wind to turn propellers **1020**, collecting fibers **1030**, **1040** may supplement/boost the amount of energy the windmill produces.

Windmill **1000**, properly equipped with ion collectors **1030**, **1040**, such as non-limiting example carbon fibers, can produce electricity: 1) by virtue of providing altitude to the carbon fiber to harvest ions, and 2) while the propeller is turning, by virtue of wind blowing over the carbon fiber producing electricity, among other reasons, via the triboelectric effect (however, it is also possible for the triboelectric effect to occur, producing electricity, in winds too weak to turn the propeller).

There are at least two ways that energy collectors may be employed on or in a windmill propeller to harvest energy. Propellers **1020** may be equipped with energy collectors **1030**, **1040** attached to, or supported by, propeller **1020** with wires (or metal embedded in, or on propeller **1020**) electrically connecting energy collectors **1030**, **1040** to a load or power conversion circuit. There may be a requirement to electrically isolate energy collectors **1030**, **1040**, which is added to propeller **1020**, from electrical ground, so that the energy collected does not short to ground through propeller **1020** itself or through support tower **1010**, but rather is conveyed to the load or power conversion circuit. Energy collectors may be connected to the end of propellers **1020** such as collectors **1030**. Alternatively, energy collectors may be connected to the sides of propellers **1020** such as collectors **1040**.

Alternatively, propeller **1020** may be constructed of carbon fiber or other suitable material, with wires (or the structural metal supporting propeller **1020** may be used) electrically connecting to a load or power conversion circuit. In the case of propeller **1020** itself being constructed of carbon fiber, for example, the fiber may be 'rough finished' in selected areas so that the fiber is "fuzzy." For example, small portions of it may protrude into the air as a means of enhancing collection efficiency. The fuzzy parts of collectors **1030**, **1040** may do much of the collecting. There may be a requirement to electrically isolate carbon fiber propeller **1020** from electrical ground, so that the energy it collects does not short to ground through metal support tower **1010**, but rather is conveyed to the load or power conversion circuit. Diodes may be implemented within the circuit to prevent the backflow of energy, although diodes may not be necessary in some applications.

In an alternative embodiment, windmill **1000** may be used as a base on which to secure an even higher extension tower to support the energy collectors and/or horizontal supports extending out from tower **1010** to support the energy collectors. Electrical energy may be generated via ion collection due to altitude and also when a breeze or wind blows over the collectors supported by tower **1010**.

Any process descriptions or blocks in flow charts should be understood as representing modules, segments, or portions of code which include one or more executable instructions for implementing specific logical functions or steps in the process, and alternate implementations are included within the scope of the preferred embodiment of the present disclosure in which functions may be executed out of order from that shown or discussed, including substantially concurrently or in reverse order, depending on the functionality involved, as would be understood by those reasonably skilled in the art of the present disclosure.

It should be emphasized that the above-described embodiments of the present disclosure, particularly, any "preferred" embodiments, are merely possible examples of implementations, merely set forth for a clear understanding of the principles of the disclosure. Many variations and modifications may be made to the above-described embodiment(s) of the disclosure without departing substantially from the spirit and principles of the disclosure. All such modifications and variations are intended to be included herein within the scope of this disclosure and the present disclosure and protected by the following claims.

Therefore, at least the following is claimed:

1. A method of collecting energy comprising:

suspending at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment from a windmill, the at least one collection device electrically connected to the windmill; and

providing a load with an electrical connection to the at least one collection device to draw current.

2. The method of claim **1**, wherein the collection device collects energy by triboelectric effect.

3. The method of claim **1**, wherein the collection device comprises a collection fiber.

4. The method of claim **1**, wherein the collection device comprises a diode and a collection fiber and the diode is electrically connected between the collection fiber and the load.

5. The method of claim **1**, further comprising storing energy provided to the load.

6. The method of claim **5**, wherein storing energy provided to the load comprises storing energy in a capacitor or an inductor.

7. The method of claim **3**, wherein the collection fiber comprises carbon fiber or graphite fiber.

8. A system of energy collection comprising:

a windmill;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment electrically connected to the windmill; and

a load electrically connected to the at least one collection device.

9. The system of claim **8**, wherein the collection device collects energy by triboelectric effect.

10. The system of claim **8**, wherein the collection device comprises a collection fiber.

11. The system of claim **8**, wherein the collection device comprises a collection fiber and a diode electrically connected between the load and the collection fiber.

12. The system of claim **9**, wherein the diode is elevated relative to the ground level.

13. The system of claim **10**, wherein the collection fiber comprises a carbon fiber or a graphite fiber.

14. The system of claim **8**, further comprising a diode electrically connected between the at least one collection device and the support structure.

15. The system of claim **8**, further comprising: a switch connected in series between the at least one collection device and the load; and a capacitor connected in parallel with the switch and the load.

16. The system of claim **15**, wherein the switch comprises an interrupter connected between the plurality at least one collection device, and wherein the interrupter comprises at least one of a fluorescent tube, a neon bulb, an AC light, and a spark gap.

17. The system of claim **8**, further comprising: a motor for providing power, the motor connected between the at least one collection device and the load; and a generator powered by the motor.

18. The system of claim **8**, further comprising a fuel cell between the support structure and the load.

19. The system of claim **18**, wherein the fuel cell produces hydrogen and oxygen.

20. A system of collecting energy comprising:

means for suspending at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment from a windmill structure, the at least one collection device electrically connected to the means for suspending;

means for inducing current flow, the means for inducing current flow electrically connected to the means for suspending; and

means for restricting the backflow of charge carriers, the means for restricting the backflow of charge carriers electrically connected between the at least one collection device and the means for inducing current flow.

* * * * *

US 20120286623A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2012/0286623 A1**

Ogram (43) **Pub. Date: Nov. 15, 2012**

(54) **ATMOSPHERIC ENERGY COLLECTION**

(75) Inventor: **Mark E. Ogram**, Tucson, AZ (US)

(73) Assignee: **Sefe, Inc.**, Tempe, AZ (US)

(21) Appl. No.: **13/103,963**

(22) Filed: **May 9, 2011**

Publication Classification

(51) **Int. Cl.**
H02N 3/00 (2006.01)

(52) **U.S. Cl.** **310/309**

(57) **ABSTRACT**

The subject matter described herein is an atmospheric energy collector. The atmospheric energy collector includes of a windsock arrangement that has a large up-wind opening on one side and that tapers from the larger up-wind opening on the one side to a small down-wind opening on the other side. The up-wind side is secured to a tether such that an electrically conducting material (e.g. metal) included in construction of the atmospheric energy collector is connected to the tether. The windsock arrangement is extended outwards by wind and the like atmospheric conditions such that the electrically conducting material collects the atmospheric energy and transfers the collected energy to the tether.

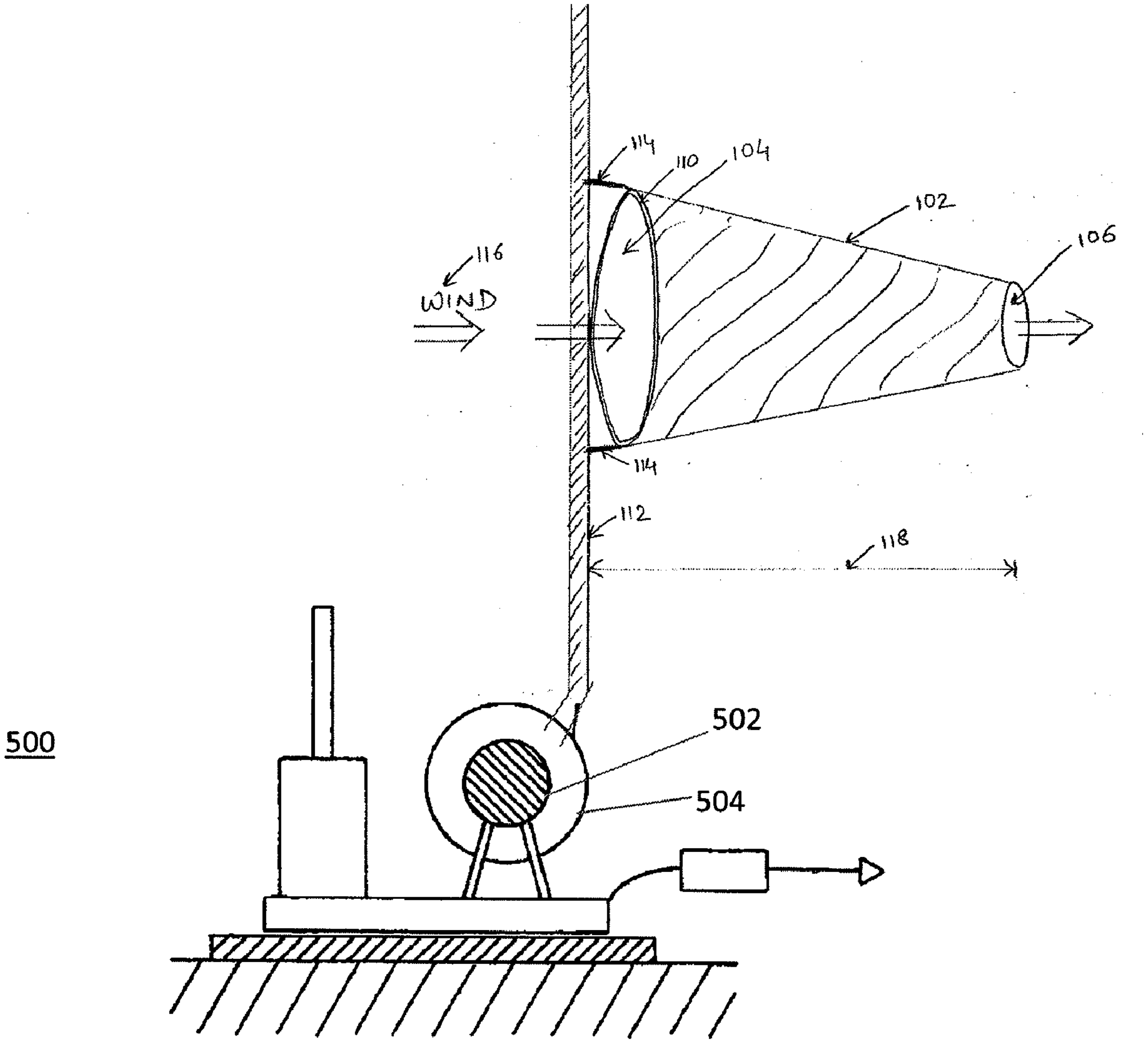

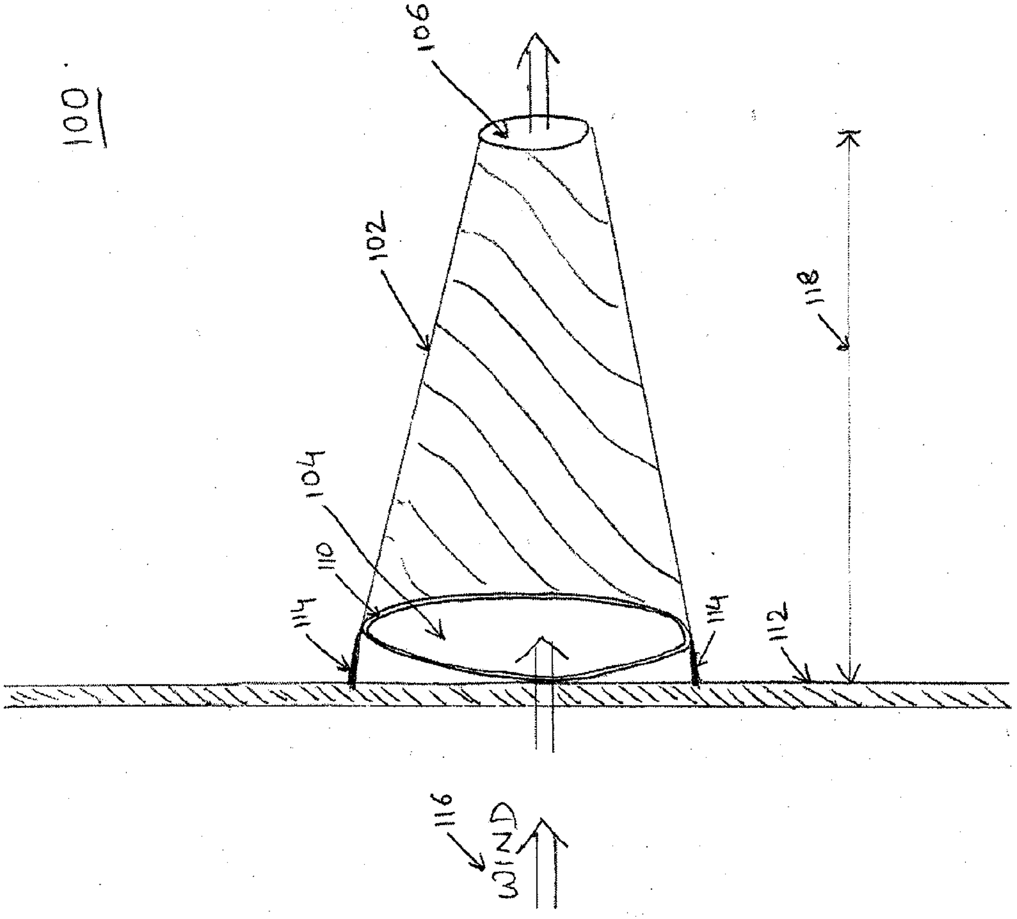
100
106
102
118
104
110
114
114
112
116
WIND

FIG. 1A

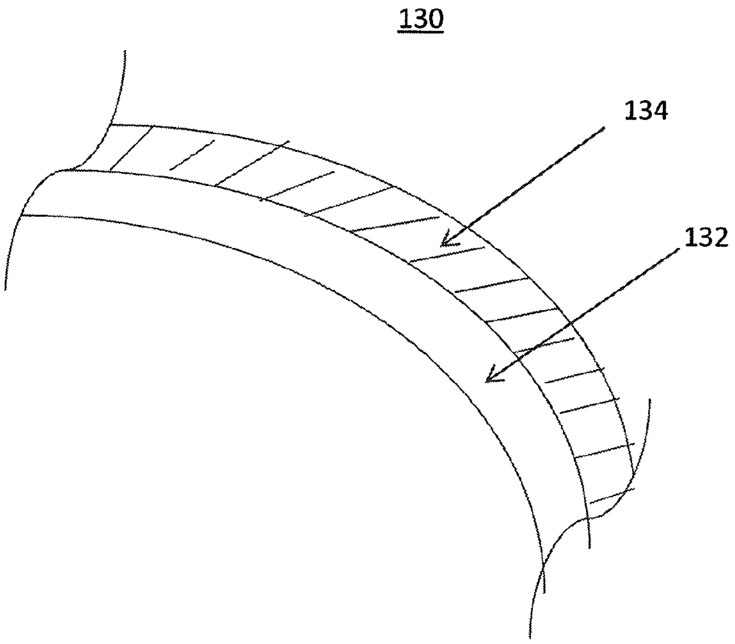

FIG. 1B

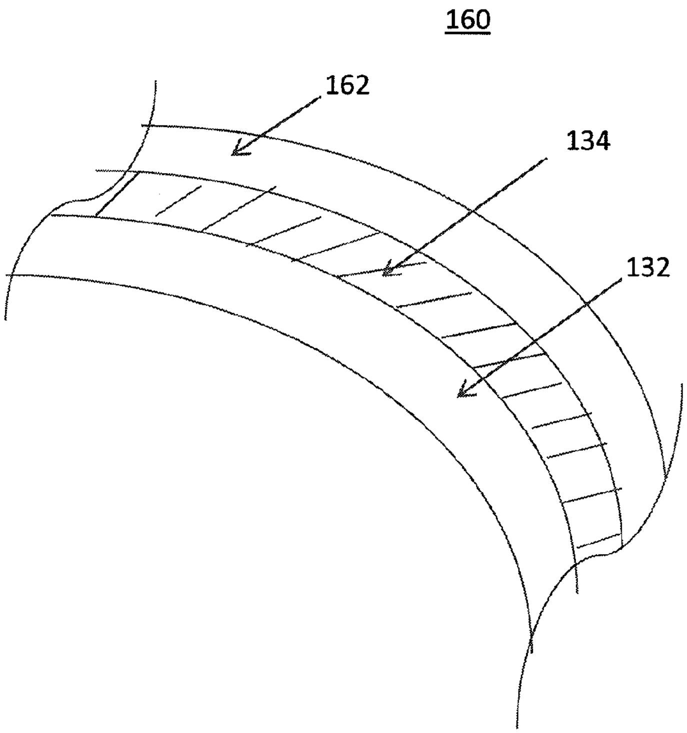
160
162
134
132

FIG. 1C

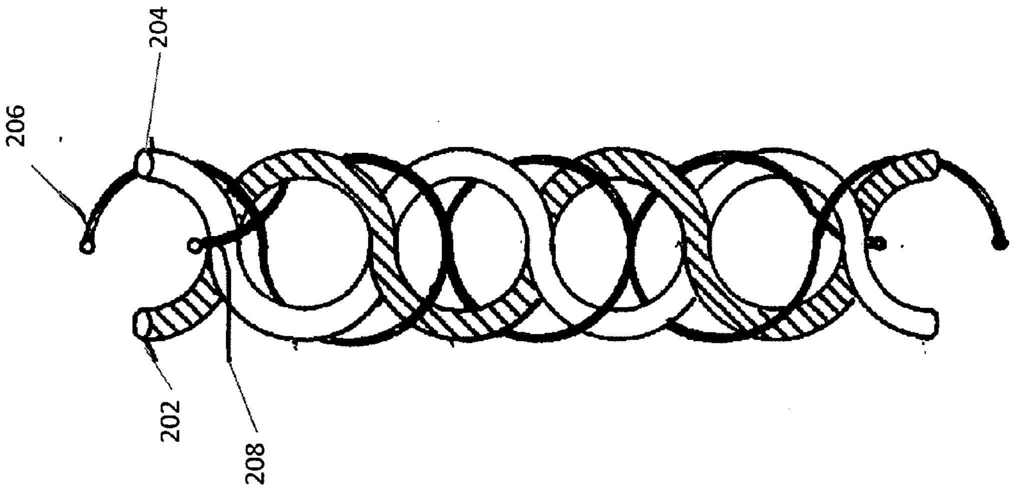
206
204
202
208

FIG. 2

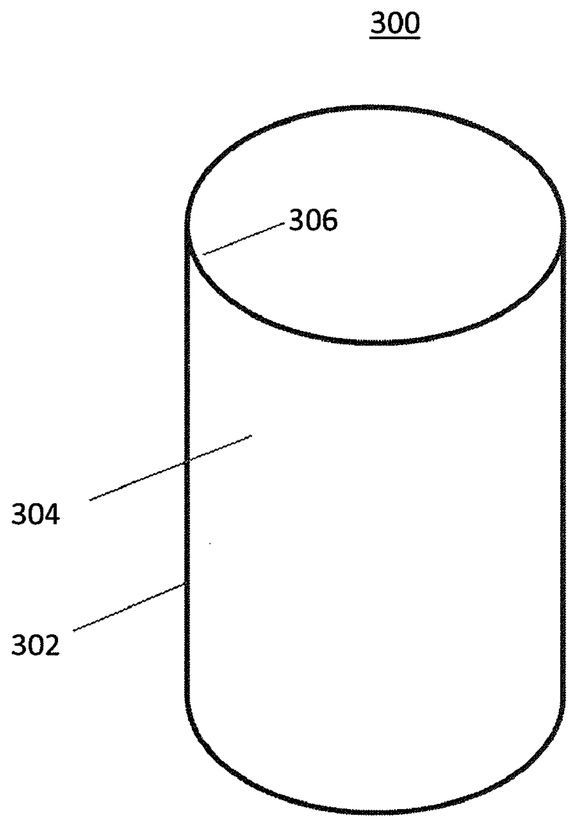
300
306
304
302

FIG. 3A

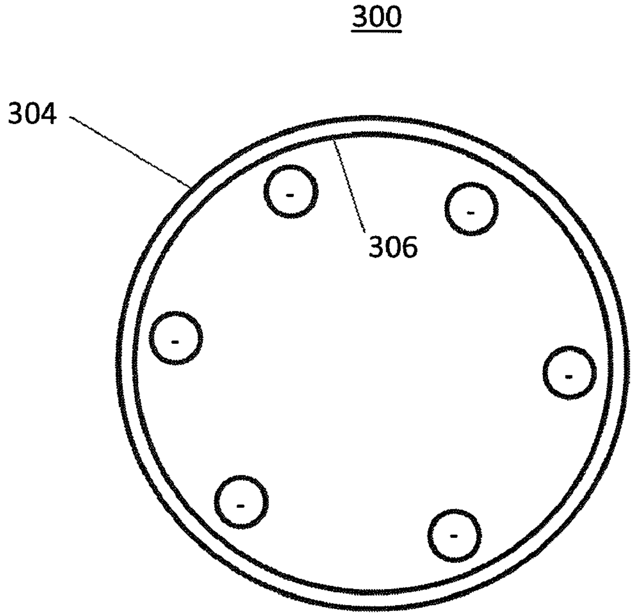

FIG. 3B

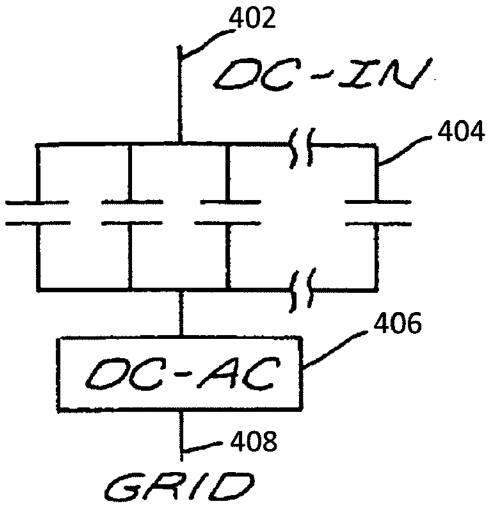
402
DC-IN
404
406
DC-AC
408
GRID

FIG. 4A

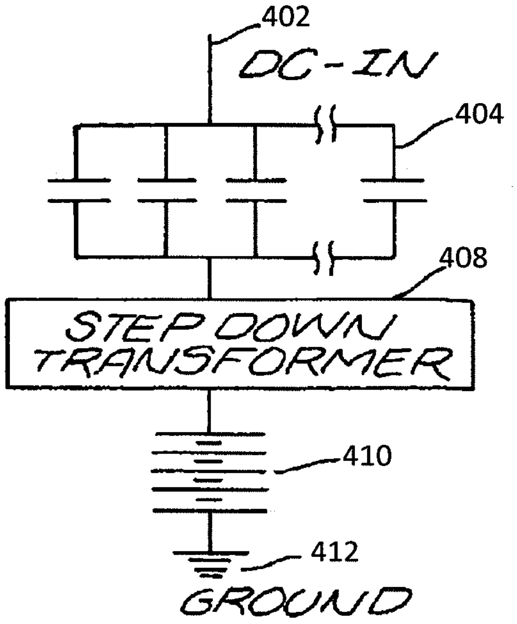
402
DC-IN
404
408
STEP DOWN TRANSFORMER
410
412
GROUND

FIG. 4B

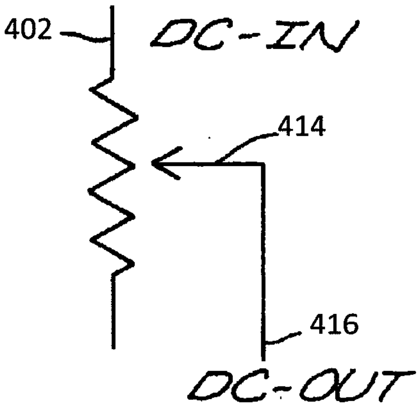
402
DC-IN
414
416
DC-OUT

FIG. 4C

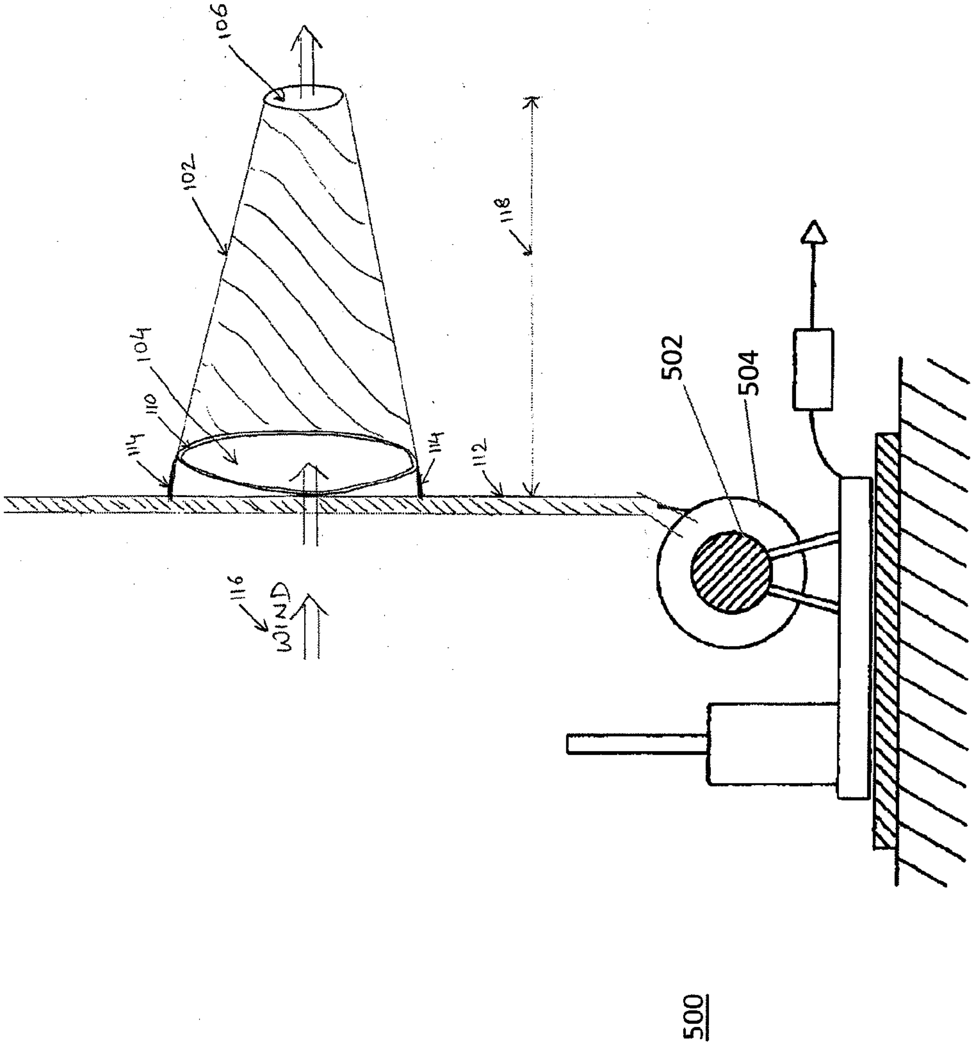
106
102
118
104
110
114
114
112
116
WIND
502
504
500

FIG. 5

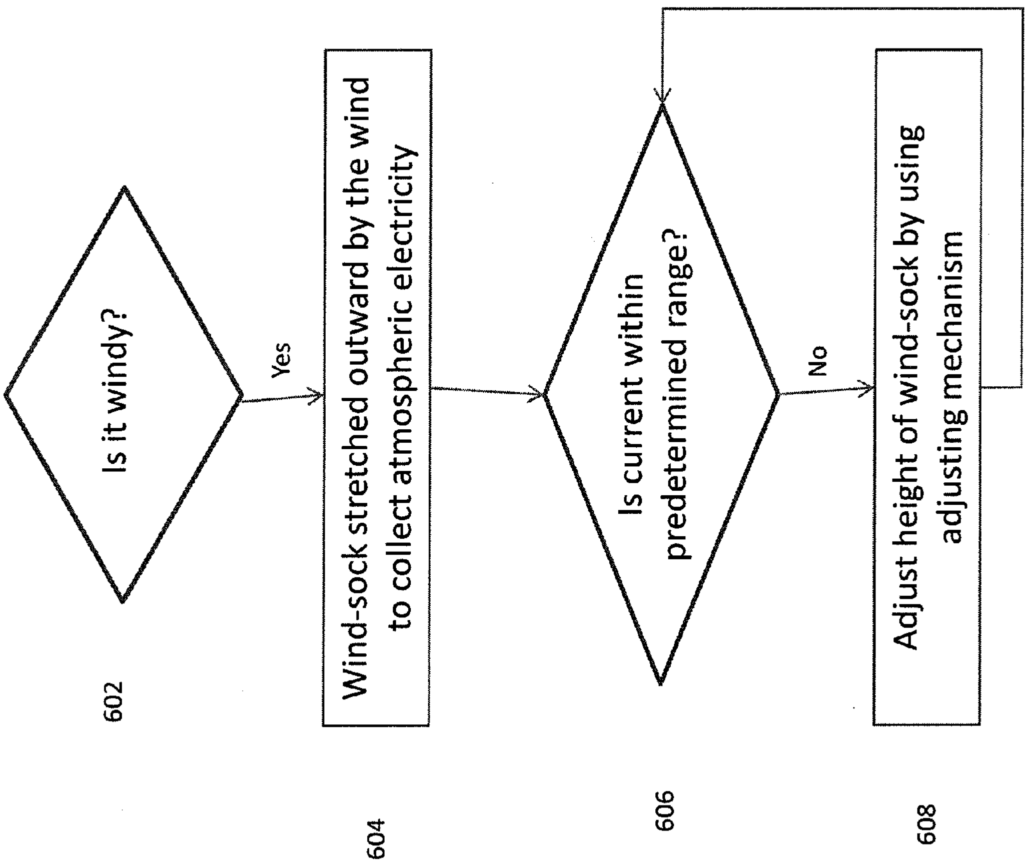
602
Is it windy?
Yes
604
Wind-sock stretched outward by the wind to collect atmospheric electricity
606
Is current within predetermined range?
No
608
Adjust height of wind-sock by using adjusting mechanism

FIG. 6

ATMOSPHERIC ENERGY COLLECTION

TECHNICAL FIELD

[0001] The subject matter described herein relates to a light-weight, extendible electricity collecting windsock arrangement comprising an enhanced collection surface for atmospheric electrical energy collection.

BACKGROUND

[0002] The atmosphere above the earth is known to include electrical charge. The earth's surface is negatively charged, while the air above it is positively charged. All atmospheric effects are a result of an interplay between these two huge areas of opposite electrical energy. The potential difference between the positively charged atmosphere and the negatively charged earth surface causes atmospheric electrical energy to be developed. Some of this atmospheric electrical energy may be experienced through thunderstorms.

[0003] Even though there is a huge amount of atmospheric electrical energy in the atmosphere, collection of the atmospheric electrical energy still remains a problem. Due to high voltage and low current conditions, there is a need for large collection surfaces for maximizing the collection of atmospheric electrical energy. Note that the terms atmospheric electricity, atmospheric energy, and atmospheric electrical energy have been used interchangeably in this specification.

[0004] The problem of low current is compounded by a heavy weight of the collectors of atmospheric electrical energy. The heavy weight of the collectors increases payload requirements of a lift mechanism, which may be used to lift the collectors to an appropriate height or altitude so that the collection of atmospheric electrical energy is maximized. Thus, a need exists for large yet light-weight collectors such that payload requirements for the lift mechanism remain reasonable.

SUMMARY

[0005] The subject matter described herein relates to an atmospheric energy collector. The atmospheric energy collector includes an electricity collecting windsock arrangement that has a large up-wind opening on one side and that tapers from the larger up-wind opening on the one side to a small down-wind opening on the other side. The up-wind side is secured to a tether such that an electrically conducting material (e.g. metal) included in construction of the atmospheric energy collector is connected to the tether. The electricity collecting windsock arrangement is extended outwards by wind and the like atmospheric conditions such that the electrically conducting material collects the atmospheric energy and transfers the collected energy to the tether. Further, the electricity collecting windsock includes light-weight collectors such that payload requirements for the lift mechanism remain reasonable (i.e. remain within a predetermined value).

[0006] In one aspect, an electricity collection apparatus includes a tether and a windsock. The windsock is formed of an insulation material with a metal deposited on at least a portion of the insulation material. The metal of the windsock is electrically connected to the tether. The windsock extends in a direction of ambient wind to provide a surface area based on the ambient wind, and the surface area provides the metal to collect electrical energy from the ambient wind, the collected electrical energy being transferred to the tether.

[0007] In another aspect, an electricity collection apparatus includes a windsock formed of an insulation material the windsock to extend in a direction of ambient wind to provide a surface area based on the ambient wind. The apparatus further includes one or more electrical conductors provided on at least a portion of the insulation material to collect electrical energy from the ambient wind. The apparatus further includes an electrically conductive tether connected with windsock to anchor the windsock in the direction of the ambient wind, the electrically conductive tether being electrically connected to the one or more electrical conductors provided on at least a portion of the insulation material to transfer the collected electrical energy from the one or more electrical conductors to an electrical storage.

[0008] The subject matter described herein provides many advantages. For example, large collection surfaces allow a maximized collection of atmospheric energy. Moreover, the light-weight collectors allow payload requirements for the lift mechanism to remain reasonable (i.e. remain within a predetermined value).

[0009] The details of one or more variations of the subject matter described herein are set forth in the accompanying drawings and the description below. Other features and advantages of the subject matter described herein will be apparent from the description and drawings, and from the claims.

DESCRIPTION OF DRAWINGS

[0010] The accompanying drawings, which are incorporated in and constitute a part of this specification, show certain aspects of the subject matter disclosed herein and, together with the description, help explain some of the principles associated with the disclosed implementations. In the drawings,

[0011] FIG. **1**A illustrates extendible electricity collecting windsock arrangement that is used for collection of atmospheric electrical energy in accordance with implementations of the current subject matter;

[0012] FIG. **1**B illustrates a cross-section of one implementation of the electrically conducting windsock in accordance with implementations of the current subject matter;

[0013] FIG. **1**C illustrates a cross-section of another implementation of the electricity collecting windsock in accordance with implementations of the current subject matter;

[0014] FIG. **2** illustrates an exemplary conductive line in accordance with implementations of the current subject matter;

[0015] FIG. **3**A illustrates a front view of a conductive line in accordance with implementations of the current subject matter;

[0016] FIG. **3**B illustrates a top view of the conductive line shown in FIG. **3** in accordance with implementations of the current subject matter;

[0017] FIGS. **4**A, **4**B, and **4**C illustrate electrical schematics for handling the static charge from the atmosphere in accordance with implementations of the current subject matter;

[0018] FIG. **5** illustrates an adjusting apparatus (winch motor and spool) that enables adjustment of location/height of the electricity collecting windsock arrangement in accordance with implementations of the current subject matter; and

[0019] FIG. **6** is a process flow diagram illustrating aspects of a method in accordance with implementations of the current subject matter.

[0020] When practical, similar reference numbers denote similar structures, features, or elements.

DETAILED DESCRIPTION

[0021] To address these and potentially other issues with currently available solutions, one or more implementations of the current subject matter provide methods, systems, articles or manufacture, and the like that can, among other possible advantages, provide an energy collector formed as a windsock and having an enhanced collection surface for atmospheric electrical energy collection. In preferred implementations, an electricity collecting windsock is formed as a flexible cylinder, a truncated cone, or a cone with a metal surface, and is mounted to a mast or other tether so as to capture ambient wind which extends the windsock outward from the mast, in turn forming the largest possible surface area for the wind conditions to enable the metal surface to collect atmospheric amperage. The collected amperage is transferred from the metal surface of the electricity collecting windsock to the mast or tether, and eventually to an electricity storage device. Although implementations in which windsock arrangements having one windsock each are discussed below, those or ordinary skill in the art understand that a windsock arrangement may include two or more windsocks for enhanced atmospheric electricity collection.

[0022] FIG. **1**A illustrates an implementation of electricity collecting windsock arrangement **100** that is used to collect atmospheric electrical energy. The electricity collecting windsock arrangement **100** includes an extendible electricity collecting windsock **102**. The extendible electricity collecting windsock **102** is made of a material that has a high tensile strength and provides electrical insulation. This electrically insulating material **134** (discussed below with respect to FIG. **1**B) may be a plastic or a flexible glass, such as polyester like polyethylene terephthalate (PET) or biaxially-oriented polyethylene terephthalate (BoPET). BoPET is known in the industry by different trade names, some of which are Mylar, Melinex and Hostaphan.

[0023] As is described later with respect to FIGS. **1**B and **1**C, which illustrate cross-sections of implementations of electricity collecting windsock **102**, a thin coat of an electrically conducting material **132**, **162** is deposited on the electrically insulating material **134**. The deposited electrically conducting material **132**, **162** is a good conductor of electricity, and helps electricity collecting windsock **102** to tap atmospheric electrical energy. The electrically conducting material may be any one of a metal or a suitable alloy, such as one of or a combination of one or more of gold, silver, copper, aluminum, and the like.

[0024] The electricity collecting windsock **102** has a large up-wind opening up-wind opening **104** on one side. The electricity collecting windsock **102** tapers from the large up-wind opening **104** on the one side to a small down-wind opening **106** on the other side of the electricity collecting windsock **102**. The up-wind opening **104** has a metal loop **110** on the circumference of the up-wind opening **104**. The metal loop **110** is attached to a tether **112** using an attachment mechanism **114**. The attachment mechanism **114** may include a soldering mechanism. In some implementations, other attachment mechanisms **114** are also known to be used, such as nut and bolt mechanism, threading mechanism, gluing mechanism, and the like. The electricity collecting windsock **102** is extended outwards by wind **116** and/or like atmospheric conditions such that the electrically conducting material (e.g. metal) collects the atmospheric electrical energy and transfers the collected energy to the tether **112**.

[0025] The tether **112** includes a conductive line that is isolated from the ground. The conductive line is used to transfer the collected atmospheric electrical energy from the electricity collecting windsock **102** to an electricity storage device (not shown) where the collected atmospheric charge may be gathered for later or simultaneous use.

[0026] When wind **116** strikes against the electricity collecting windsock arrangement **100** in the direction illustrated, the electricity collecting windsock **102** extends outward, thus allowing the electrically conducting material **132** (discussed below with respect to FIG. **1**B) to collect maximum possible atmospheric electricity. In one implementation, the outward extension **118** is fifty feet or more. When there is no wind, the electricity collecting windsock **102** may be in a relatively compressed state such that the length of the electricity collecting windsock **102** is less than the outward extension **118** caused by windy conditions or like atmospheric conditions. The collected atmospheric electrical energy is communicated to the tether **112**.

[0027] FIG. **1**B illustrates a cross-section **130** of one implementation of the electrically conducting windsock **102**. The cross-section **130** illustrates the electrically conducting material **132** deposited on the inner surface of the electrically insulating material **134**. The electricity collecting windsock **102** includes the electrically conducting material **132** and the electrically insulating material **134** such that the electrically conducting material **132** forms the inner surface of the electricity collecting windsock **102**. When the wind **116** strikes the electricity collecting windsock arrangement, the electricity collecting windsock **102** extends outward, thus allowing the electrically conducting material **132** to collect maximum possible atmospheric electricity. The electrically conducting material **132** may include multiple collectors that are spaced apart such that the collection surface is greatly enhanced. The multiple collectors may be spaced apart in patterns, such as spaced apart lines, squares, rectangles, circles, and the like, or any other pattern that may maximize the collection of the atmospheric electrical energy. As noted above, the electrically insulating material **134** may be a plastic or a flexible glass, such as polyester like polyethylene terephthalate (PET) or biaxially-oriented polyethylene terephthalate (BoPET). The electrically conducting material **132** may be a metal or alloy, such as one of or a combination of one or more of gold, silver, copper, aluminum, and the like.

[0028] FIG. **1**C illustrates a cross-section **160** of another implementation of the electricity collecting windsock **102**. The electrically conducting material (e.g. metal deposit) **132**, **162** may be placed on both sides of the electrically insulating material **134** such that the collection ability is increased even further. Multiple collectors are implemented and are spaced apart on electrically conducting material **132**, **162** such that the collection surface is greatly enhanced. The multiple collectors may form one or more patterns, as noted above with respect to FIG. **1**B. Furthermore, the electrically insulating material **134** (e.g. mylar) used in the electricity collecting windsock **102** is light, thereby minimizing the overall weight of the electricity collecting windsock **102** and reducing the weight for payload calculations. The electrically conducting materials **132**, **162** may be an electrically conducting metal or alloy, such as one of or a combination of at least one of gold, copper, aluminum, silver, and the like. The electrically conducting materials **132**, **162** may be made of the same metal/alloy or a different metal/alloy.

[0029] The electrically conducting materials **132**, **162** may be attached to the electrically insulating material **134** by a gluing mechanism. In some implementations, other attachment mechanisms are known to be implemented, such as paint coating mechanism, a threading mechanism, a nut and bolt mechanism, soldering mechanism, and the like. The attachment mechanism between the electrically conducting material **132** and the electrically insulating material **134** may or may not be the same as the attachment mechanism between the electrically conducting material **162** and the electrically insulating material **134**.

[0030] FIG. **2** illustrates an exemplary conductive line in accordance with one implementation. This type of conductive line is commonly referred to as poly-wire. The conductive line consists of multiple interwoven strands of plastic **202** and **204** woven into a cord or rope arrangement having intertwined therein exposed metal wires **206** and **208**. Although FIG. **2** illustrates two plastic strands and two metal wires, any number of possible combinations of plastic strands and metal wires is possible. The exposed metal wires **206** and **208** attract the atmospheric static charge and transmit the charge down to the electricity storage device (not shown).

[0031] FIG. **3**A illustrates a front view of a conductive line **300** in accordance with another implementation. FIG. **3**B illustrates a top view of the conductive line **300** illustrated in FIG. **3**. The conductive line **300** creates an ionized pathway for the flow of the static charges from the atmosphere to the electricity storage device via the electricity collecting windsock arrangement **100**. This conductive line utilizes a tube **302** having an outer layer **304** of PET Film (Biaxially-oriented polyethylene terephtalate polyester film). The tube **302** provides exceptionally high tensile strength and is chemically and dimensionally stable. In one implementation, the tube **302** may have an ideal diameter of between two and three inches. An interior metal coating **306** provides an initial conduit for the flow of static charge. The static charge through the metal may force the tube **302** to expand due to the repulsion experienced by like charges. Further, the flow of electricity causes the interior of the tube **302** to become ionized to provide an additional pathway for the atmospheric static charges to the electricity storage device (not shown).

[0032] FIGS. **4**A, **4**B, and **4**C illustrate electrical schematics for handling the static charge from the atmosphere. By maintaining the voltage being collected in a prescribed range, an electrical conversion system is easily designed. While FIGS. **4**A, **4**B, and **4**C illustrate some electrical configurations, those of ordinary skill in the art readily recognize a variety of other configurations that may serve the same function.

[0033] Referencing FIG. **4**A, Direct Current In (DC IN) **402** is buffered by a gang of capacitors **404** before being communicated to a DC/AC converter **406**. The DC/AC converter converts the direct current into alternating current suitable for placement over an existing electrical grid **408** such as normally found from a power-plant. Those of ordinary skill in the art readily recognize a variety of DC/AC converters that may work in this capacity.

[0034] FIG. **4**B illustrates an electrical arrangement suitable for use in charging a battery. DC IN **402** is buffered by capacitor bank **404** before entering into a step down transformer **408**. Step down transformer **408** reduces the voltage so that the voltage can safely be introduced into battery **410**, which is connected to ground **412** at the battery's other pole. Those of ordinary skill in the art readily recognize a variety of batteries that may work in this capacity.

[0035] In FIG. **4**C, DC IN **402** is fed into an adjustable rheostat **414**, which is controlled by the controller so that the DC OUT **416** falls within a specified range.

[0036] FIG. **5** illustrates an adjusting apparatus **500** that enables adjustment of height of the electricity collecting windsock arrangement **100** in accordance with one implementation. As per one implementation, the height of the electricity collecting windsock **102** can be adjusted using a winch motor **502** and a spool **504**. The winch motor **502** and the spool **504** can release or withdraw the tether **112** to adjust the height of the electricity collecting windsock **102** that is connected to the tether **112**. In one implementation, this release and the withdrawal may be performed manually. In another implementation, this release and the withdrawal may be performed automatically based on information obtained from a sensor system that measures atmospheric electrical energy being collected per unit time. If the sensor indicates that a current flow is diminishing, then the tether **112** is released/extended from the spool **504** to increase the altitude of the electricity collecting windsock **102** such that more static charge from the atmosphere is gathered. If the sensor indicates that collected atmospheric electrical energy exceeds a preset level per unit time, the tether **112** is withdrawn onto the spool **504** to decrease the static charge being collected from the atmosphere.

[0037] FIG. **6** is a process flow diagram illustrating aspects of a method consistent with implementations of the current subject matter. At step **602**, the functioning of the electricity collecting windsock arrangement **100** depends on whether atmospheric conditions are windy. If the atmospheric conditions are windy, the wind **116** stretches/extends outward the electricity collecting windsock **102**—Step **604**. Next, it is determined whether the atmospheric electrical energy (atmospheric current) that is collected is in a predetermined range—Step **606**. If the collected atmospheric electrical energy is in the predetermined range, the flow goes back to step **604**. If the current is not in the predetermined range, the height of the electricity collecting windsock **102** may be adjusted using adjusting apparatus **500**, such as the winch motor **502** and spool **504**. One skilled in the art understands that other implementations may include any other adjusting apparatus **500**, such as a pulley, a wheel mechanism, or the like.

[0038] The implementations set forth in the foregoing description do not represent all implementations consistent with the subject matter described herein. Instead, they are merely some examples consistent with aspects related to the described subject matter. Although a few variations have been described in detail herein, other modifications or additions are possible. In particular, further features and/or variations can be provided in addition to those set forth herein. For example, the implementations described above can be directed to various combinations and sub-combinations of the disclosed features and/or combinations and sub-combinations of one or more features further to those disclosed herein. In addition, the logic flows depicted in the accompanying figures and/or described herein do not necessarily require the particular order shown, or sequential order, to achieve desirable results. The scope of the following claims may include other implementations or embodiments.

What is claimed is:

1. An electricity collection apparatus comprising:

a tether; and

a windsock formed of an insulation material with a metal deposited on at least a portion of the insulation material, the metal of the windsock being electrically connected to the tether, the windsock to extend in a direction of ambient wind to provide a surface area based on the ambient wind, the surface area providing the metal to collect electrical energy from the ambient wind, the collected electrical energy being transferred to the tether.

2. The apparatus in accordance with claim **1**, wherein the metal is deposited on at least one of opposite sides of the insulation material.

3. The apparatus in accordance with claim **2**, wherein the metal is a metal selected from a group of metals comprising gold, silver, copper and aluminum.

4. The apparatus in accordance with claim **1**, wherein the insulation material is a polyester film.

5. The apparatus in accordance with claim **1**, wherein the transferred electrical energy at the tether is further transferred from the tether to an electricity storage device.

6. The apparatus in accordance with claim **1**, wherein the metal comprises conductors that are light-weight thereby allowing payload requirements for adjusting the windsock to an optimum location to be within a predetermined value, the optimum location comprising a height of the windsock from ground level.

7. The apparatus in accordance with claim **6**, wherein the deposited metal conductors form a pattern that maximizes the collection of the electrical energy at the optimum location.

8. An electricity collection apparatus comprising:

a windsock formed of an insulation material having a metal deposited on at least a portion of the insulation material, the windsock to extend in a direction of ambient wind to provide a surface area based on the ambient wind, the surface area providing the metal to collect electrical energy from the ambient wind; and

an electrically conductive tether connected with windsock to anchor the windsock in the direction of the ambient wind, the electrically conductive tether being electrically connected to the metal of the insulation material, the collected electrical energy being transferred to the tether.

9. The apparatus in accordance with claim **8**, wherein the metal is deposited on at least one of opposite sides of the insulation material.

10. The apparatus in accordance with claim **9**, wherein the metal is a metal selected from a group of metals comprising gold, silver, copper and aluminum.

11. The apparatus in accordance with claim **8**, wherein the insulation material is a polyester film.

12. The apparatus in accordance with claim **8**, wherein the transferred electrical energy at the tether is further transferred from the tether to an electricity storage device.

13. The apparatus in accordance with claim **8**, wherein the metal comprises conductors that are light-weight thereby allowing payload requirements for adjusting the windsock to an optimum location to be within a predetermined value, the optimum location comprising a height of the windsock from ground level.

14. The apparatus in accordance with claim **13**, wherein the metal conductors form a pattern that maximizes the collection of the electrical energy at the optimum location.

15. An electricity collection apparatus comprising:

a windsock formed of an insulation material the windsock to extend in a direction of ambient wind to provide a surface area based on the ambient wind;

one or more electrical conductors provided on at least a portion of the insulation material to collect electrical energy from the ambient wind; and

an electrically conductive tether connected with windsock to anchor the windsock in the direction of the ambient wind, the electrically conductive tether being electrically connected to the one or more electrical conductors provided on at least a portion of the insulation material to transfer the collected electrical energy from the one or more electrical conductors to an electrical storage.

* * * * *

US 20170117714A1

(19) **United States**

(12) **Patent Application Publication**

Dinwiddie et al.

(10) **Pub. No.: US 2017/0117714 A1**

(43) **Pub. Date: Apr. 27, 2017**

(54) **POWER RECEIVER FOR EXTRACTING POWER FROM ELECTRIC FIELD ENERGY IN THE EARTH**

(71) Applicant: **Earth Energies, Inc.**, Johns Creek, GA (US)

(72) Inventors: **John Dinwiddie**, Cary, NC (US); **Terry L. Wright**, Suwanee, GA (US); **David R. Ames**, Johns Creek, GA (US)

(21) Appl. No.: **15/397,281**

(22) Filed: **Jan. 3, 2017**

Related U.S. Application Data

(63) Continuation-in-part of application No. 14/509,772, filed on Oct. 8, 2014, now Pat. No. 9,564,268.

(60) Provisional application No. 61/889,894, filed on Oct. 11, 2013.

Publication Classification

(51) **Int. Cl.**

H02J 4/00	(2006.01)
H02J 7/35	(2006.01)
H02J 3/38	(2006.01)

(52) **U.S. Cl.**

CPC ***H02J 4/00*** (2013.01); ***H02J 3/385*** (2013.01); ***H02J 7/35*** (2013.01)

(57) **ABSTRACT**

A resonant transformer connected between a ground terminal and elevated terminal draws current from the earth's electric field through a primary winding of the transformer. An impulse generator applies a high voltage impulse to the primary winding of the resonant transformer to cause current to flow from the ground terminal through the primary winding. The flow of current through the primary winding of the resonant transformer induces a current in the secondary winding, which may be converted and filtered to a usable form, e.g. 60 Hz AC or DC.

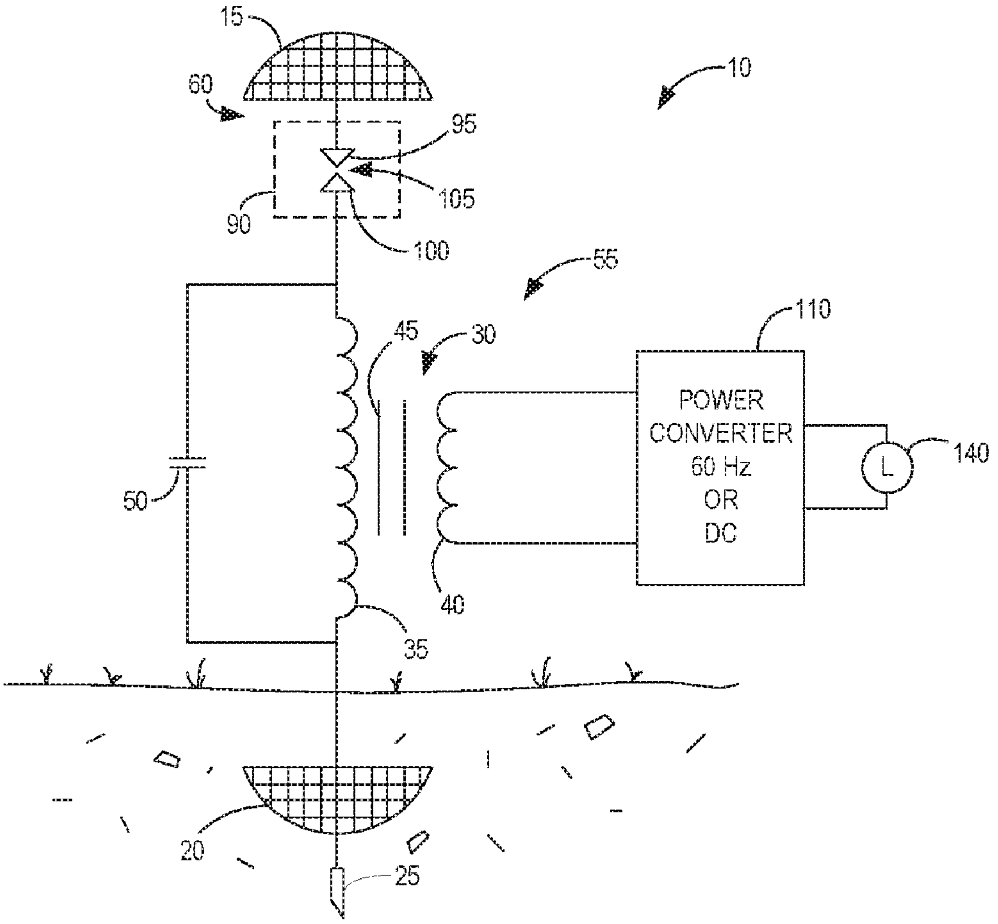

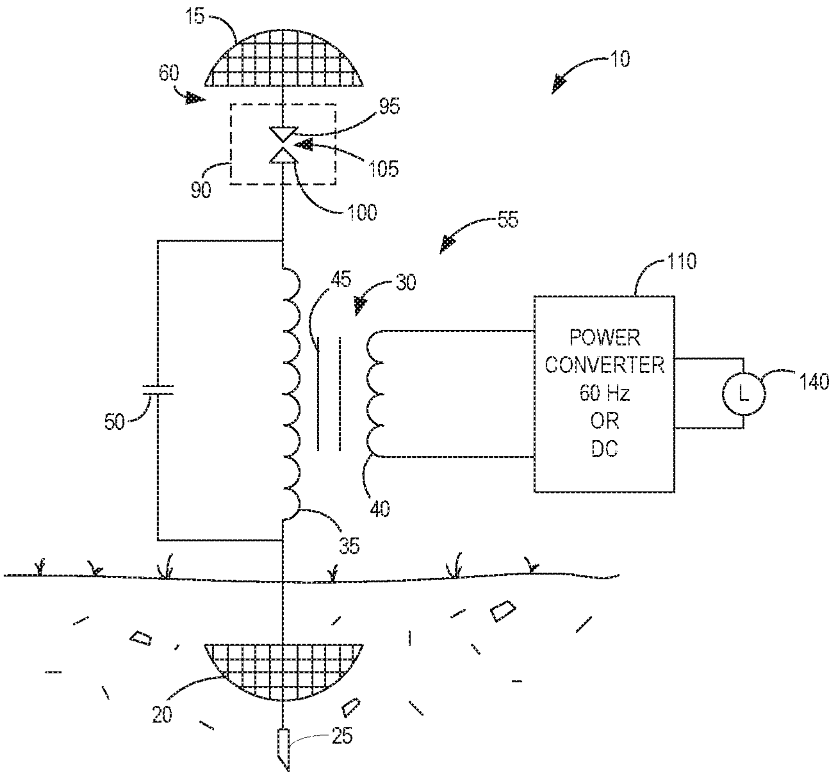
POWER CONVERTER 60 Hz OR DC
L

FIG 1

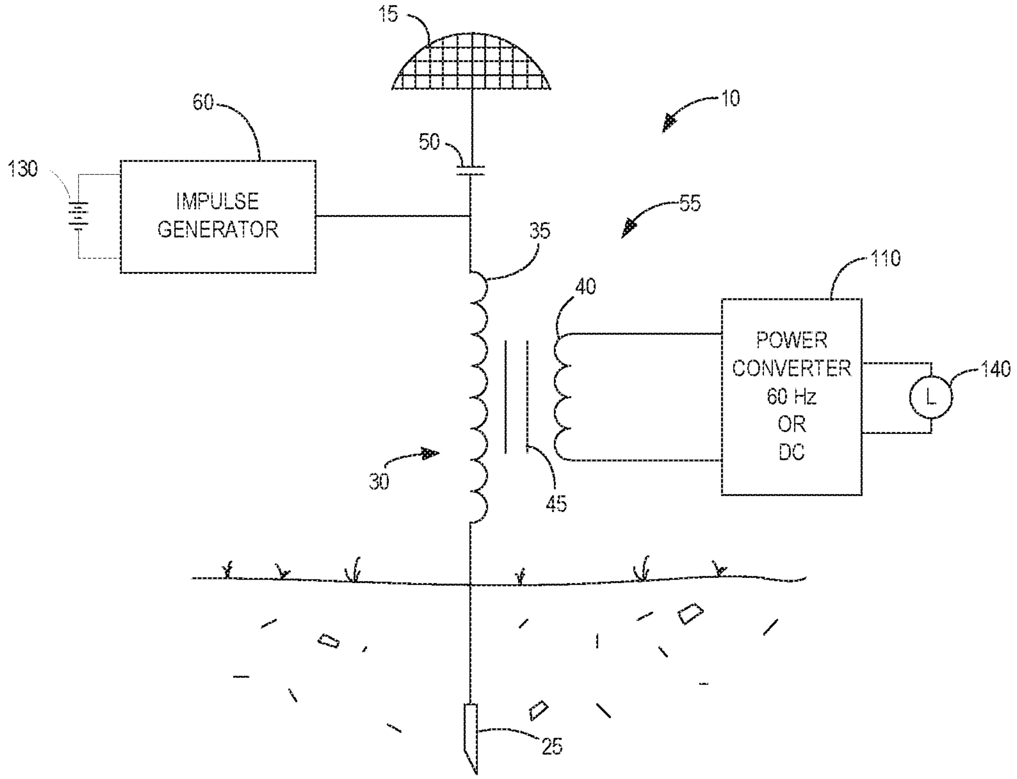
IMPULSE GENERATOR
POWER CONVERTER 60 Hz OR DC
L

FIG 2

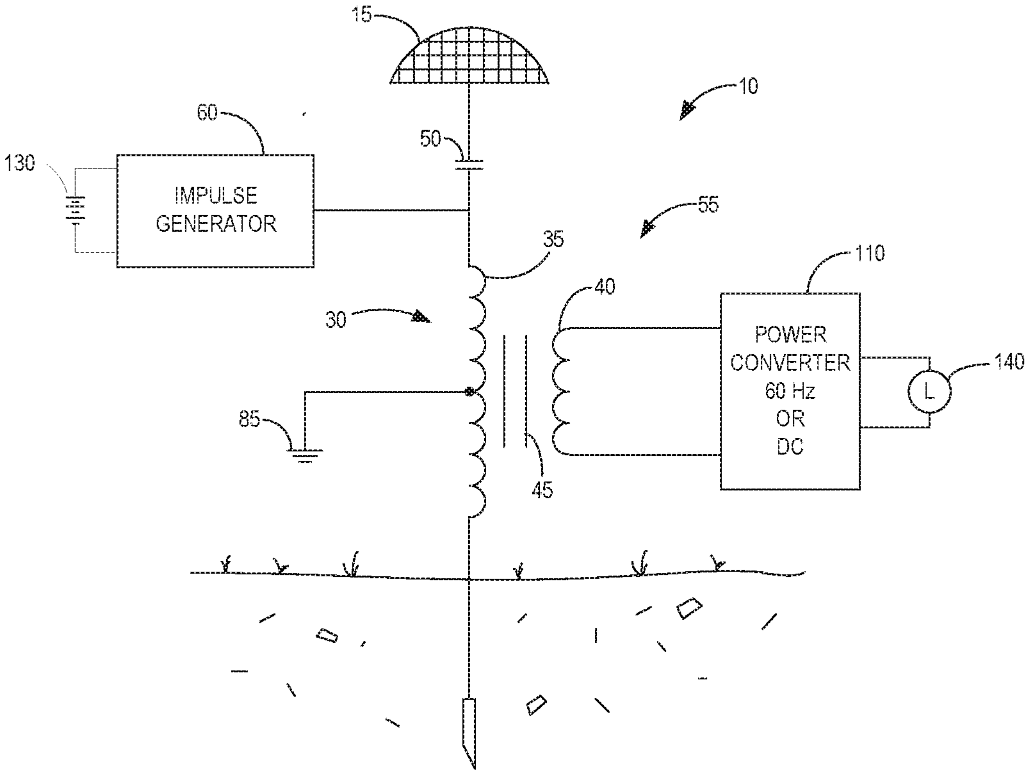
IMPULSE GENERATOR
POWER CONVERTER 60 Hz OR DC
L

FIG 3

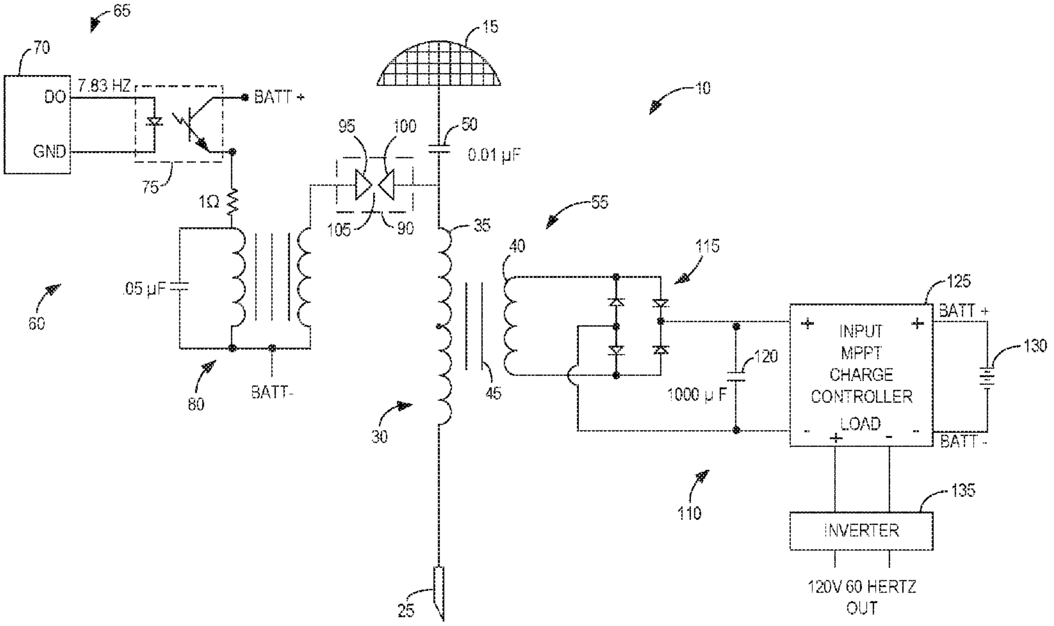
65
70
DO
GND
7.83 HZ
BATT +
75
1Ω
.05 µF
60
80
BATT-
15
10
50
0.01 µF
95
100
105
90
35
55
40
115
30
45
25
120
1000 µ F
125
INPUT
MPPT
CHARGE
CONTROLLER
LOAD
+
-
BATT +
BATT -
130
110
135
INVERTER
120V 60 HERTZ
OUT

FIG 4

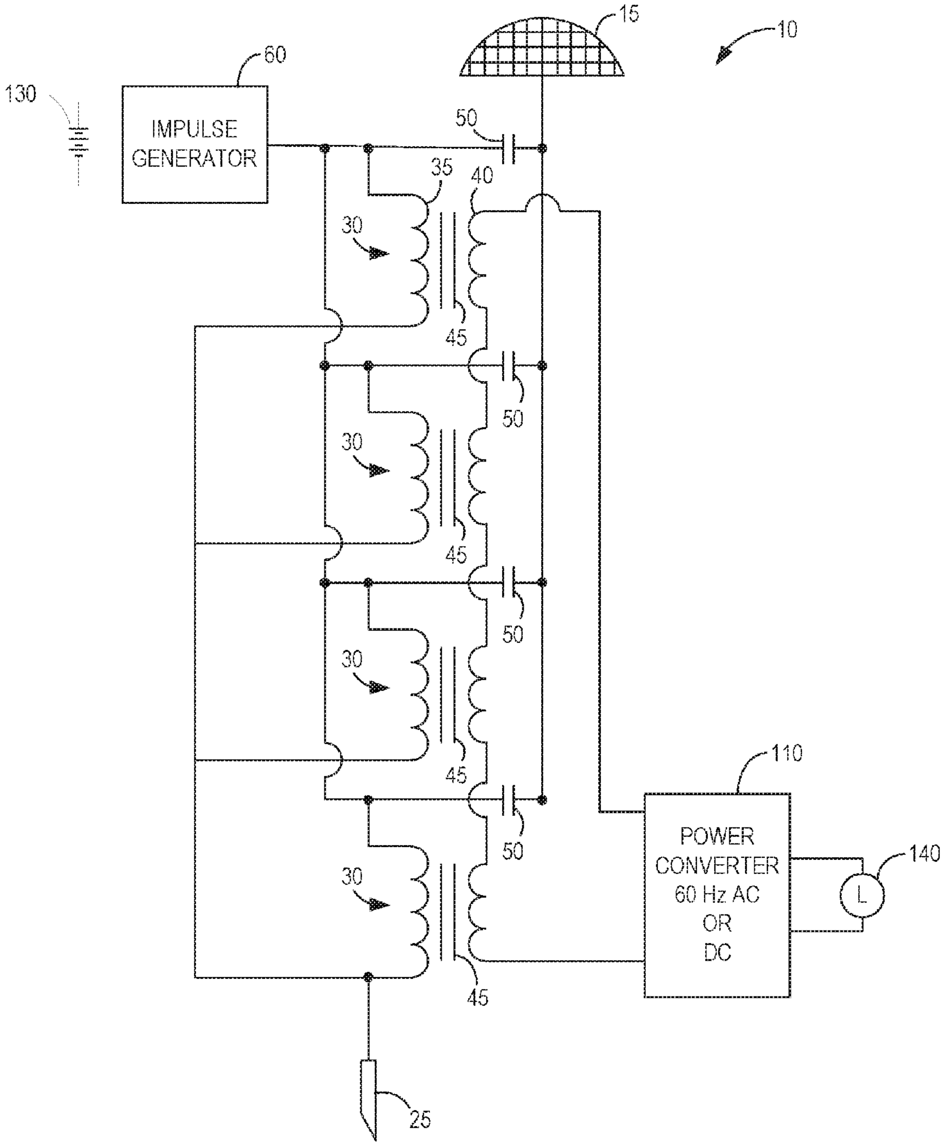
IMPULSE GENERATOR
POWER CONVERTER 60 Hz AC OR DC
L

FIG 5

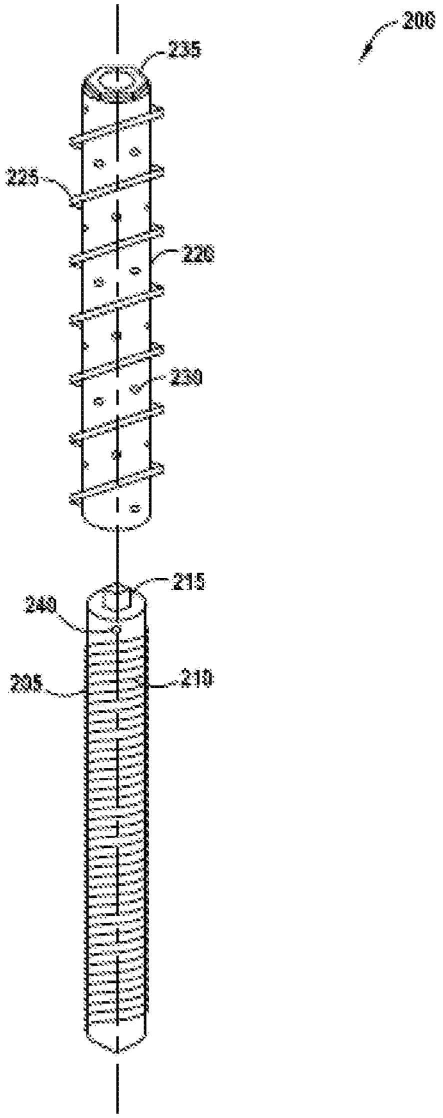
200
235
225
220
230
215
240
205
210

FIG 6A

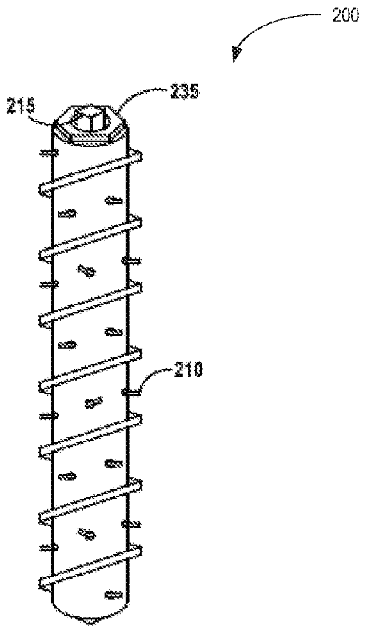
200
215
235
210

FIG 6B

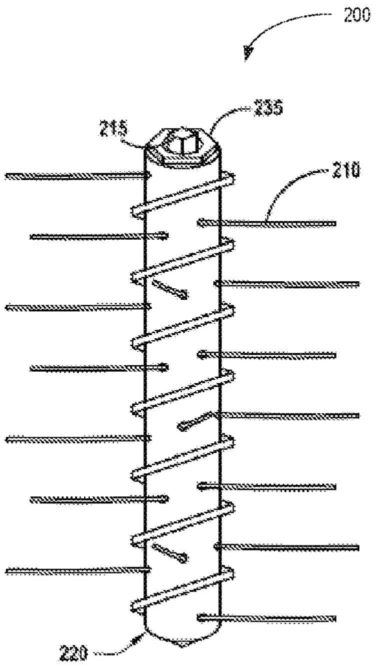
200
215
235
210
220

FIG 6C

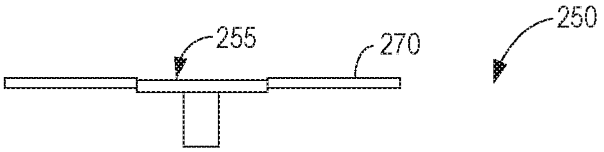

FIG 7A

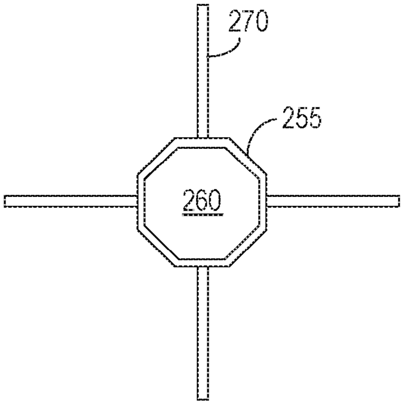

FIG 7B

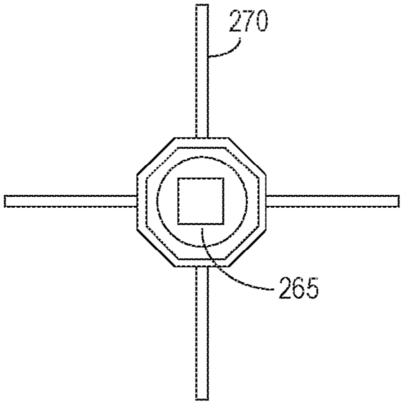

FIG 7C

POWER RECEIVER FOR EXTRACTING POWER FROM ELECTRIC FIELD ENERGY IN THE EARTH

[0001] This application is a continuation-in-part of prior U.S. application Ser. No. 14/509772 filed 8 Oct. 2014 which claims the benefit of U.S. Provisional Application No. 61/889894 filed 11 Oct. 2013, the disclosures of all of which are incorporated by reference herein in their entirety.

TECHNICAL FIELD

[0002] The present invention relates generally to renewable energy, and more particularly to methods and apparatus for extracting energy from subsurface electrical fields beneath the earth's surface.

BACKGROUND

[0003] The earth and the ionosphere cavity may be viewed as a global electric circuit. Electrical currents are constantly flowing within the earth and its atmosphere. Within the earth, the majority of the earth's energy is carried by extremely low frequency (ELF) and ultralow frequency (ULF) waves in the 0-200 Hz frequency range. The earth's rotating magnetic field and positive lightning are two energy sources that sustain the ELF/ULF waves within the earth and the atmosphere.

[0004] A great deal of research has been devoted to studying the electric field present in the earth's ionosphere cavity. Joseph M. Crawley, the "Fair Weather Atmosphere as a Power Source", Proceedings ESA Annual Meeting on Electrostatics 2011; O. Jefimenko, "Operation of Electric Motors from Atmospheric Electric Field," American Journal of Physics, Vol. 39, Pgs. 776-779, 1971; M. L. Breuer, "Usability of Tapping Atmospheric Charge as a Power Source," Renewable Energy, Vol. 28, Pgs. 1121-1127, 2003. Numerous attempts have been made in the past to extract electrical energy from the earth's atmosphere. For example, U.S. Pat. No. 1,540,998 to Plauson describes a system for converting atmospheric electrical energy into usable power. These past attempts have been successful in producing only small amounts of power from the electrical field in the earth's ionosphere cavity. The modest success of these experiments compared to results from other renewable energy sources, such as solar and wind, has tempered further research and prevented widespread use of the electric field in the ionosphere cavity as an energy source.

SUMMARY

[0005] The present invention relates to a power receiver for extracting power from electric fields beneath the earth's surface. In embodiments of the present disclosure, a resonant transformer connected to a ground terminal draws current from the earth's electric field through the primary winding of the transformer. Current flow through the resonant transformer is induced by applying a high voltage impulse to the primary winding. The flow of current through the primary winding of the resonant transformer induces a current in the secondary winding, which may be converted and filtered to a usable form, e.g. 60 Hz AC or DC.

[0006] In some embodiments of the power receiver, the resonant frequency of the resonant transformers is below 200 Hz.

[0007] In some embodiments of the power receiver, the resonant transformer comprises a ferro-resonant transformer.

[0008] In some embodiments, the power receiver further comprises an elevated terminal.

[0009] In some embodiments of the power receiver, the primary winding of the resonant transformer is connected between the ground terminal and elevated terminal.

[0010] In some embodiments of the power receiver, the elevated terminal comprises an upper capacitive plate coupled to the earth's ionosphere cavity.

[0011] In some embodiments of the power receiver, the impulse generator comprises the upper capacitive plate and a spark gap connected between the upper capacitive plate and the primary winding of the resonant transformer. The spark gap comprises a pair of electrodes separated by a gap and configured to generate a spark when a voltage difference between the electrodes reaches a predetermined level.

[0012] In some embodiments of the power receiver, the impulse generator comprises a pulse generator for generating low voltage pulses, a step-up transformer for converting the low voltage pulses provided by the pulse generator to high voltage impulses, and a spark gap connected between the step-up transformer and the primary winding of the resonant transformer to generate a spark responsive to the high voltage impulses from the step-up transformer.

[0013] In some embodiments of the power receiver, the impulse generator comprises a pulse generator for generating low voltage pulses, and a step-up transformer connected to the primary winding of the resonant transformer for converting the low voltage pulses provided by the pulse generator to high voltage impulses.

[0014] In some embodiments of the power receiver, the impulse generator comprises a solid state spark generator.

[0015] In some embodiments of the power receiver, the resonant transformer includes a capacitor connected in parallel with the primary winding.

[0016] In some embodiments of the power receiver, the resonant transformer includes a capacitor connected in series with the primary winding between the impulse generator and the elevated terminal.

[0017] In some embodiments, the power receiver comprises multiple resonant transformers having primary windings connected in parallel between the ground terminal and the elevated terminal.

[0018] In some embodiments of the power receiver, the resonant transformers have different resonant frequencies.

[0019] In some embodiments of the power receiver, the resonant frequencies of the resonant transformers are all below 200 Hz.

[0020] In some embodiments of the power receiver, the resonant frequencies of the resonant transformers are matched to respective Schumann resonances.

[0021] Another embodiment of the power receiver comprises a resonant circuit connected to a ground terminal disposed below the surface of the earth, an impulse generator for generating and applying a high voltage electrical impulse to the resonant circuit to induce current flow from the ground terminal through the resonant circuit, and a power conversion circuit connected to the resonant circuit to convert electrical current flowing through the resonant circuit to a desired form. The resonant circuit has a resonant frequency below 200 Hertz.

[0022] In some embodiments of the power receiver, the resonant circuit comprises a resonant transformer having a primary winding, a secondary winding, and resonant capacitor connected in series with the primary winding.

[0023] In some embodiments of the power receiver, the resonant circuit comprises multiple resonant transformers having primary windings connected in parallel to the ground terminal.

[0024] In some embodiments of the power receiver, the resonant transformers have different resonant frequencies.

[0025] In some embodiments of the power receiver, the resonant frequencies of the resonant transformers are all below 200 Hz.

[0026] In some embodiments of the power receiver, the resonant frequencies of the resonant transformers are matched to respective Schumann resonances.

[0027] Other embodiments of the disclosure comprise a ground terminal for a power receiver. In one embodiment, the ground terminal comprises a ground shaft configured for insertion beneath the surface of the earth, a hollow cylinder surrounding the ground shaft and having a plurality of openings, and a plurality of ground wires connected at one end to the ground shaft. The ground wires are wound around the ground shaft and have free ends protruding through respective openings in the hollow shaft so that rotation of the ground shaft relative to the hollow cylinder causes the ground wires to extend radially into the earth.

[0028] Other embodiments of the disclosure comprise methods of extracting power from the earth. In one embodiment, the method comprises applying a high voltage impulse to resonant circuit coupled to a ground terminal disposed beneath the surface of the earth to initiate resonance in the resonant circuit and induce the flow of current from the ground terminal to the resonant circuit, and converting the current flowing from the ground terminal into the resonant circuit into a useful form.

[0029] In some embodiments of the method, the resonant circuit comprises a resonant transformer including a primary winding coupled to the ground terminal and a second winding coupled to a power converter, and applying a high voltage impulse to resonant circuit comprises applying a high voltage impulse to the primary winding of the resonant transformer.

[0030] In some embodiments of the method, applying a high voltage impulse to the primary winding of the resonant transformer comprises applying an impulse in the range to 10,000 to 40,000 volts to primary winding of the transformer.

BRIEF DESCRIPTION OF THE DRAWINGS

[0031] FIG. **1** illustrates a first exemplary embodiment of a power receiver.

[0032] FIG. **2** illustrates a second exemplary embodiment of a power receiver.

[0033] FIG. **3** illustrates a third exemplary embodiment of a power receiver.

[0034] FIG. **4** illustrates a fourth exemplary embodiment of a power receiver.

[0035] FIG. **5** illustrates a fifth exemplary embodiment of a power receiver.

[0036] FIG. **6**A is an exploded perspective view of an exemplary ground antenna array for the power receiver.

[0037] FIG. **6**B is a perspective view of an assembled ground antenna array before being deployed.

[0038] FIG. **6**C is a perspective view of an assembled ground antenna array after being deployed.

[0039] FIG. **7**A is a side view of an insertion tool for installing the ground antenna array.

[0040] FIG. **7**B is a top view of the insertion tool for installing the ground antenna array.

[0041] FIG. **7**C is a bottom view of the insertion tool for installing the ground antenna array.

DETAILED DESCRIPTION

[0042] Referring now to the drawings, a power receiver for extracting energy from the earth's electric field are illustrated and indicated generally by the numeral **10**. Various embodiments of the power receiver **10** are described and similar reference numbers are used throughout the description to indicate similar components.

[0043] The power receiver **10** converts energy in the ELF/ULF waves to useful form, e.g. 60 Hz AC or DC. The power receiver **10** is essentially a resonance circuit that resonates at the natural resonance frequencies in the earth's electric field. These resonance frequencies, known as Schumann resonance frequencies, occur at 7.83 Hz, 14.3 Hz, 20.8 Hz, 27.3 Hz, and 33.8 Hz. A high voltage impulse initiates resonance within the power receiver **10**. In the resonant mode, the impedance of the power receiver **10** is reduced to near zero thus inducing ground currents to flow into the power receiver **10** where the ground currents are converted to useful form.

[0044] FIG. **1** illustrates a first embodiment of the power receiver **10**. The power receiver **10** comprises a resonant transformer **30** connected between an elevated terminal **15** and ground terminal **25**. In this embodiment, the elevated terminal **15** is capacitively coupled to electric fields within the earth's ionosphere cavity and functions as an upper capacitive plate. A lower capacitive plate **20** is connected to the ground terminal **25** beneath the surface of the earth.

[0045] The resonant transformer **30** comprises a primary winding **35**, secondary winding **40**, ferromagnetic core **45**, and capacitor **50**. One end of the primary winding **35** is connected to the lower capacitive plate **20** and ground terminal **25**. The opposite end of the primary winding **35** is connected via a spark gap **90** to the elevated terminal **15**. The capacitor **50** is connected in parallel with the primary winding **35** of the resonant transformer **30** to form an LC circuit **55** with a resonance frequency range of between about 0.1 and 200 Hz. In a preferred embodiment, the resonant transformer has a Q of about 10 or greater and resonance frequency in the range of about 0.1-200 Hertz. For example, the resonant transformer **30** may have a resonance frequency of about 7.83 Hz, the fundamental Schumann resonance frequency. The secondary winding **40** of the resonant transformer **30** is connected to a power converter **110** as will be hereinafter described in greater detail. The power converter **110** converts the energy extracted from the earth's electric field by the power receiver **10** into a usable form for driving a load **140**.

[0046] The elevated terminal/upper capacitive plate **15** comprises an insulated, dish-shaped plate with a large radius of curvature. The capacitance and resistance of the elevated terminal is chosen for receiving broadband electric field frequencies in the 0-200 Hz range. The upper capacitive plate **15** is sized to maximize to the extent practical coupling with the electric field in the earth's ionosphere cavity.

[0047] The lower capacitive plate **20** is also a dish-shaped plate with a large radius of curvature. One function of the lower capacitive plate **20** is to collect charge from the earth's ground currents and provide an instantaneous source of current as hereinafter described. The capacitance and resistance of the lower capacitive plate **20** is selected to promote the flow of current from the ground with minimal losses.

[0048] The spark gap **90** connected between the elevated terminal **15** and resonant transformer **30** comprises a pair of electrodes **95**, **100** separated by an evacuated air gap **105**. Electrode **95** is connected to the upper capacitive plate **15**. Electrode **100** is connected to the resonant transformer **30**. The spark gap **105** prevents electrical discharge from the upper capacitive plate **15** to the earth's ionosphere cavity. The spark gap **90** in combination with the elevated terminal **15** function as an impulse generator that applies a high voltage impulse of about 10,000-40,000 volts to the primary winding **35** to initiate resonance in the transformer **30**.

[0049] In operation, the capacitive coupling of the upper capacitive plate **15** induces a high voltage operating current in the upper capacitive plate **15**. The upper capacitive plate is connected to a first electrode **95** to the spark gap **90**. When the voltage difference between the electrodes **95** and **100** reaches a threshold, a spark forms across the electrodes **95**, **100** and a high voltage impulse is applied to the primary winding **35** of the resonant transformer **30**. This high voltage impulse initiates resonance within the transformer **30**.

[0050] In resonant mode, the impedance of the resonance transformer is reduced to nearly zero allowing current to flow from the capacitive plate **20** and ground terminal **25** through the primary winding **35** of the transformer **30**, which in turn induces current in the secondary winding **40**. Power converter **110** converts the current flowing through the secondary winding **40** into a usable form for driving a load **140**. The transformer **30** will continue to resonate for a short period of time. By providing high voltage impulses to the primary winding **35** of the resonant transformer **30** at periodic intervals, it is possible to maintain a continuous flow of current from the earth into the resonant transformer **30**, thus producing a continuous supply of power.

[0051] FIG. **2** discloses a second embodiment of the primary receiver **10**. This embodiment includes a resonant transformer **30** connected between an elevated terminal **15** and ground terminal **25**. The resonant transformer **30** comprises a primary winding **35**, secondary winding **40**, ferromagnetic core **45** and a high voltage capacitor **50**. One end of the primary winding **35** is connected to the ground terminal **25**. The opposite end of the primary winding **35** is connected to the elevated terminal **15**. The capacitor **50** has a capacitance of about 0.01 micro-farads. In contrast to the previous embodiment, capacitor **50** is connected in series with the primary winding **35** and elevated terminal **15** and forms a LC circuit **55** with a Q of about 10 or greater and a resonance frequency in the range of about 0.1 to 200 Hz. In a preferred embodiment, the resonance frequency of the transformer **30** is 7.83 Hz, the fundamental Schumann resonance frequency. An impulse generator **60** is connected between the primary winding **35** of the resonant transformer **30** and the series capacitor **50** and applies a high voltage impulse to the primary winding **35** of the resonant transformer **30**. A battery **130** or other external power source supplies power to the impulse generator **60**. As previously described, the high voltage impulse applied by the impulse generator **60** initiates resonance within the resonant transformer **30** inducing current flow from the ground terminal **25** into the primary winding **35** of the resonant transformer **30**. The flow of current from the ground terminal **25** through the primary winding **35** induces current in the secondary winding **40**. Power converter **110** converts the electrical energy in the current flowing through the primary winding **40** into a usable form.

[0052] In contrast to the first embodiment, it is not required to capacitively couple the elevated terminal **15** in the second embodiment to the earth's ionosphere cavity. Rather, the elevated terminal **15** in this embodiment provides lightning protection and dissipates some of the energy flowing into the power receiver **10** to the earth's ionosphere cavity. Also, in contrast to the first embodiment, the capacitor **50** is connected in series between the primary winding **35** of the transformer **30** and the elevated terminal **15**. Those skilled in the art will appreciate that the capacitor **50** could also be connected in parallel rather than series with the primary winding **35** as shown in FIG. **1**. Another difference is that the impulse generator **60** has an external power source. The amount of energy generated by the power receiver **10**, however, is far greater than the energy needed to generate high voltage impulses. The first embodiment does not require an external power source to generate high voltage impulses.

[0053] FIG. **3** illustrates a third embodiment of the power receiver **10**. This embodiment is essentially the same as the embodiment shown in FIG. **2**. The main difference is that a center tap of the primary winding **35** in the resonant transformer **30** is connected to an electrical ground **85**. It should be appreciated that the electrical ground **85** may be different than the earth ground. When the center tap of the resonant transformer **30** is grounded at a distance away from the ground terminal **25** (e.g. 50 ft to 100 ft), the power receiver **10** becomes a transmitter via the ground loop formed.

[0054] FIG. **4** illustrates the power receiver **10** of FIG. **2** in greater detail. The power receiver includes a resonant transformer **30** connected between a ground terminal **25** and an elevated terminal **15**. The ground terminal **25** may comprise a ⅝-inch×8-foot copper ground rod, such as the ERICO 615880UPC. The elevated terminal **15** may comprise a 90% copper mesh formed into a hemisphere with a radius of about 9 inches. The elevated terminal **15** may be elevated at a height of approximately 6 feet above the ground.

[0055] The resonant transformer **30** includes a primary winding **35**, secondary winding **40**, ferromagnetic core **45** and series capacitor **50** configured as previously described. The resonant transformer **30** may have a Q of about 10 and a resonance frequency in the range of about 0.1 to 200 Hz. The resonant transformer **30** may be made using an Allanson transformer (part #1530BP120R) connected in series with a 0.01 micro-farad capacitor, such as the Condensor Products high voltage capacitor (part #TC 103-17-125). The resonant transformer **30** is used in a step-down configuration. The center tap of the resonant transformer **30** may optionally be connected to a ground.

[0056] An impulse generator **60** is connected between the primary winding **35** of the resonant transformer **30** and the series capacitor **50** and applies a high voltage impulse in the range of about 10,000 to 40,000 volts to the primary winding of the transformer **30**. A battery **130** or other external power source supplies power to the impulse generator **60**. The power converter **110** connects to the secondary winding **40**

of the resonant transformer **30** for converting current in the secondary winding of the transformer to a useful form.

[0057] The impulse generator **60** comprises a pulse generator **65** for generating low voltage pulses, a step-up transformer **80** for converting the low voltage pulses from the pulse generator **65** to high voltage pulses, and a spark gap **90** for generating sparks responsive to the high voltage pulses from the step-up transformer **80**.

[0058] The pulse generator **65** comprises a square wave generator **70**, such as a Sinometer VC2002 function signal generator, and solid state relay **75**. The square wave generator **70** generates a digital pulse stream. In one embodiment, the digital pulse stream generates a square waveform with a frequency of about 7.83 Hz. The frequency of the digital pulse stream is selected to match the resonance frequency of the transformer **30**, though such is not necessarily required. The pulse stream output from the square wave generator **70** is applied to the solid state relay **75**. The solid state relay **75** is connected between a battery or other power source and a first winding of the step-up transformer **80**. The battery may comprise a 12 V, 7.0 A/H sealed lead acid battery, such as the ELB 1270A by Lithonia Lighting. The solid state relay **75** functions as a switch that is activated responsive to the waveform from the square wave generator **70** to provide a continuous stream of low voltage pulses from the battery to the first winding of the step-up transformer **80**. A 1 ohm resistor is connected between the solid state relay **75** and step-up transformer **80**.

[0059] The step-up transformer **80** may comprise a Transco 15 kV, 30 mA neon sign transformer (part #S15612). The step-up transformer **80** converts the low voltage pulses from the pulse generator **65** to high voltage pulses that are applied to the spark gap **90**. The step-up transformer has a 0.5 micro-farad capacitor connected in parallel with the primary winding of the step-up transformer **80**. The step-up transformer produces pulses at the output of about 30,000 to 40,000 volts.

[0060] The spark gap **90** comprises a pair of electrodes **95**, **100** separated by an air gap **105**. A suitable spark gap electrode pair is the Information Unlimited SPARK05 ¼-inch×1-inch tungsten electrodes. As previously described, when the voltage potential between the electrodes **95**, **100** reaches a threshold, a spark forms between the electrodes **95**, **100** and supplies a nearly instantaneous, high voltage impulse to the primary winding **35** of the resonant transformer **30**. This high voltage impulse initiates resonance in the resonant transformer **30** inducing current flow from the ground terminal **25** through the primary winding **35** of the resonant transformer **30**.

[0061] The power converter **110** comprises a bridge rectifier **115**, filter capacitor **120**, charge controller **125**, and inverter **135**. A suitable rectifier is the Micro Commercial Components 10 amp, 1000 volt bridge rectifier (Part #GBJL 1010). The bridge rectifier **115** converts the AC current flowing through the secondary winding **40** of the resonant transformer to a DC current. A filter capacitor **120** removes unwanted frequencies from the DC current. A suitable capacitor **120** is Cornell Dubilier 1000uF 450VDC capacitor (part #383LX102M450N082). The filter capacitor **120** has a capacitance of about 1000 micro-farads. The DC current is input to the charge controller **125**. The charge controller **125** may, for example, comprise a maximum power point tracking (MPPT) charge controller, such as a Tracer 4215 BN MPPT Solar Charge Controller, which is commonly used in solar power generating systems. The charge controller **125** applies a small amount of energy to a battery **30** to charge the battery **130**. As previously noted, the battery **130** serves as a power source for the impulse generator **60**. The remaining current is supplied to an inverter **135**, which converts the DC current to an AC current with a desired voltage and frequency, e.g., 120 volts/60 Hz AC. A suitable inverter **135** is the 1500 W Pure Sine power inverter (AIMS) (part #PWRI1500125).The power converter **110** as shown in FIG. **4** may be utilized in the embodiment shown in FIGS. **1**, **2** and **3**.

[0062] FIG. **5** illustrates a power receiver **10** according to another embodiment. The power receiver **10** comprises a plurality of resonant transformers **30** connected between a ground terminal **25** and elevated terminal **15**. Each of the resonant transformers **30** comprises a primary winding **35**, secondary winding **40**, ferromagnetic core **45** and series capacitor **50**. The primary windings **35** of the resonant transformers **30** are connected in parallel. The secondary windings **40** are connected in series with the power converter **110**. An impulse generator **60** applies a high voltage impulse to the primary windings **35** of the resonant transformers **30**. A battery **130** or other external power source supplies power to the impulse generator **60**. The power converter **110** converts the current in the power converter circuit to a usable form for driving a load **140**.

[0063] In one embodiment, each of the resonant transformers **30** shown in FIG. **5** is configured to have a different resonant frequency. In one embodiment, the resonant transformers **30** are configured to resonate at frequencies of 7.83 Hz, 14.8 Hz, 20.3 Hz and 26.8 Hz respectively. Additional resonant transformers **30** could be added to operate at other resonance frequencies.

[0064] FIGS. **6A-6C** illustrate a high quality ground antenna array **200** which may be used as a ground terminal **25**. The ground antenna array **200** comprises a generally cylindrical ground shaft **205** disposed with a hollow cylinder **220** and a plurality of reinforced, heavy gauge ground wires **210** attached at one end to the ground shaft **205**. The ground shaft **205** and ground wires **210** should be highly conductive and have low resistance to supply current from the ground to the power receiver **10**. In one embodiment, the ground wires **210** may be copper or other highly-conductive metal. The end of the ground shaft may be pointed to facilitate insertion into the earth. A connection port on the ground shaft **220** is provided to electrically connect the ground antenna array **220** to the resonant transformer **30**.

[0065] The hollow cylinder **220** has external threads **25** to facilitate insertion into the ground. A rotator nut **235** is fixedly secured to the top end of the hollow shaft **220**. A square shaft **215** protrudes from the top end of the ground shaft **205** into the opening in the rotator nut **235**. FIG. **6B**. A tool **250**, shown in FIG. **7**, engages with the rotator nut **235** and square shaft **215** during insertion of the ground antennas array **200** into the ground as will be hereinafter described.

[0066] The insertion tool **250** is shown in FIG. **7**. The insertion tool **250** includes a tool body **255** having a first socket **260** on one side to fit the rotator nut **235** on the hollow cylinder **220** and a second socket **265** on the other side to fit the square shaft **215** on the ground shaft **205**. Arms **270** extend from the outer periphery of the tool body **255** for manually or mechanically turning the insertion tool **250**.

[0067] Before the antenna array **200** is deployed, the ground wires **210** are wound around the ground shaft **205** with the free ends protruding slightly from respective openings **230** in the hollow cylinder **220** to a distance not to exceed one half (½) the depth of the external threads **225** on the hollow cylinder **220**. FIG. **6**B illustrates the ground antenna array **200** before deployment. FIG. **6**C illustrates the ground antenna array in a deployed configuration.

[0068] Installation of the ground antenna array **200** is performed in two stages. In the first stage, a hole slightly smaller in diameter than the threads **235** of the hollow cylinder **220** is drilled into the Earth to a depth matching the length of the hollow cylinder **220** or slightly longer. The hole is filled with water and the water is allowed to soak into the soil. After the ground is softened, the hollow cylinder **220** is rotated using the insertion tool **250** to insert the ground antenna array **200** into the ground. The first socket **260** of the insertion tool **250** is engaged with the rotator nut **230** and the insertion tool **250** is turned by hand or a mechanized rotating shaft fitted and attached to the tool arms **270** to thread the ground assembly into the hole. During the initial insertion of the ground antenna array **200**, the ground shaft **205** is fixed to the hollow shaft **220** and rotates with the hollow shaft. The hollow cylinder **220** is rotated until it reaches the full depth of the hole.

[0069] Once the ground antenna array **200** has been fully inserted into the earth, the insertion tool **250** is flipped over and the second socket **265** of the insertion tool **250** is engaged with the square shaft **215**. The insertion tool **250** is turned by hand or a mechanized rotating shaft fitted and attached to the tool arms **270** to rotate the ground shaft **205**. During the second phase, the ground shaft **205** rotates freely inside the hollow cylinder **220**. Rotation of the ground shaft **205** causes the reinforced ground wires **210** to extend radially into the earth. The ground shaft **220** is rotated until the ground wires are fully extended. The ends of the ground wires may be sharpened to aid in the extension of the ground wires during the second phase.

[0070] After the ground antenna array **200** is deployed, a connection cable **280** is attached to a connection port **240** on the ground shaft **220** to electrically connect the ground antenna array **220** to the resonant transformer **30** in the power receiver **10**.

What is claimed is:

1. A power receiver for extracting electrical energy from the earth's electric field, said power receiver comprising:

a resonant transformer connected to a ground terminal disposed below the surface of the earth;

an impulse generator for generating and applying a high voltage electrical impulse to a primary winding of the resonant transformer to induce current flow from the ground terminal through the primary winding of the transformer; and

a power conversion circuit connected to a secondary winding of the resonant transformer to convert electrical current flowing through the secondary winding to a desired form.

2. The power receiver of claim **1** wherein a resonant frequency of the resonant transformers is below 200 Hz.

3. The power receiver of claim **1** wherein the resonant transformer comprises a ferro-resonant transformer.

4. The power receiver of claim **1** further comprising an elevated terminal, and wherein the primary winding of the resonant transformer is connected between the ground terminal and elevated terminal.

5. The power receiver of claim **1** wherein the impulse generator comprises:

a pulse generator for generating low voltage pulses;

a step-up transformer for converting the low voltage pulses provided by the pulse generator to high voltage impulses;

a spark gap connected between the step-up transformer and the primary winding of the resonant transformer to generate a spark responsive to the high voltage impulses from the step-up transformer.

6. The power receiver of claim **1** wherein the impulse generator comprises a solid state spark generator.

7. The power receiver of claim **1** wherein said resonant transformer includes a capacitor connected in parallel with the primary winding.

8. The power receiver of claim **1** wherein said resonant transformer includes a capacitor connected in series with the primary winding between the impulse generator and the elevated terminal.

9. The power receiver of claim **1** comprising multiple resonant transformers having primary windings connected in parallel between the ground terminal and the elevated terminal.

10. The power receiver of claim **9** wherein the resonant transformers have different resonant frequencies.

11. The power receiver of claim **10** wherein the resonant frequencies of the resonant transformers are matched to respective Schumann resonances.

12. A power receiver for extracting electrical energy from the earth's electric field, said power receiver comprising:

a resonant circuit connected to a ground terminal disposed below the surface of the earth, said resonant circuit having a resonant frequency below 200 Hertz

an impulse generator for generating and applying a high voltage electrical impulse to the resonant circuit to induce current flow from the ground terminal through the resonant circuit; and

a power conversion circuit connected to the resonant circuit to convert electrical current flowing through the resonant circuit to a desired form.

13. The power receiver of claim **12** wherein the impulse generator comprises:

a pulse generator for generating low voltage pulses;

a step-up transformer for converting the low voltage pulses provided by the pulse generator to high voltage impulses;

a spark gap connected between the step-up transformer and resonant circuit to generate a spark responsive to the high voltage impulses from the step-up transformer.

14. The power receiver of claim **12** wherein the resonant circuit comprises a resonant transformer having a primary winding, a secondary winding, and resonant capacitor connected in parallel with the primary winding.

15. The power receiver of claim **12** wherein the resonant circuit comprises a resonant transformer having a primary winding, a secondary winding, and resonant capacitor connected in series with the primary winding.

16. The power receiver of claim **12** wherein the resonant circuit comprises multiple resonant transformers having primary windings connected in parallel to the ground terminal.

17. The power receiver of claim **16** wherein the resonant transformers have different resonant frequencies.

18. The power receiver of claim **17** wherein the resonant frequencies of the resonant transformers are matched to respective Schumann resonances.

* * * * *

Capacitors we used are electrolytic rated at 400 volts x 47 uF put in series to equal 6,000 volts, the diodes we used were silicon 1000 volt 2 amp placed in series to equal 6,000 volts......... Ground was connected to laboratory wall out let ground.

If you decide to try Tesla's experiment by pumping DC into the ground be careful, I tried this and it does work but is very dangerous to you or your neighbors. If someone is taking a shower or using water they can get killed or shocked. do this experiment far away from humans and animals. you can get far more energy out than you put in. I will not tell you much more because it is such a dangerous experiment.

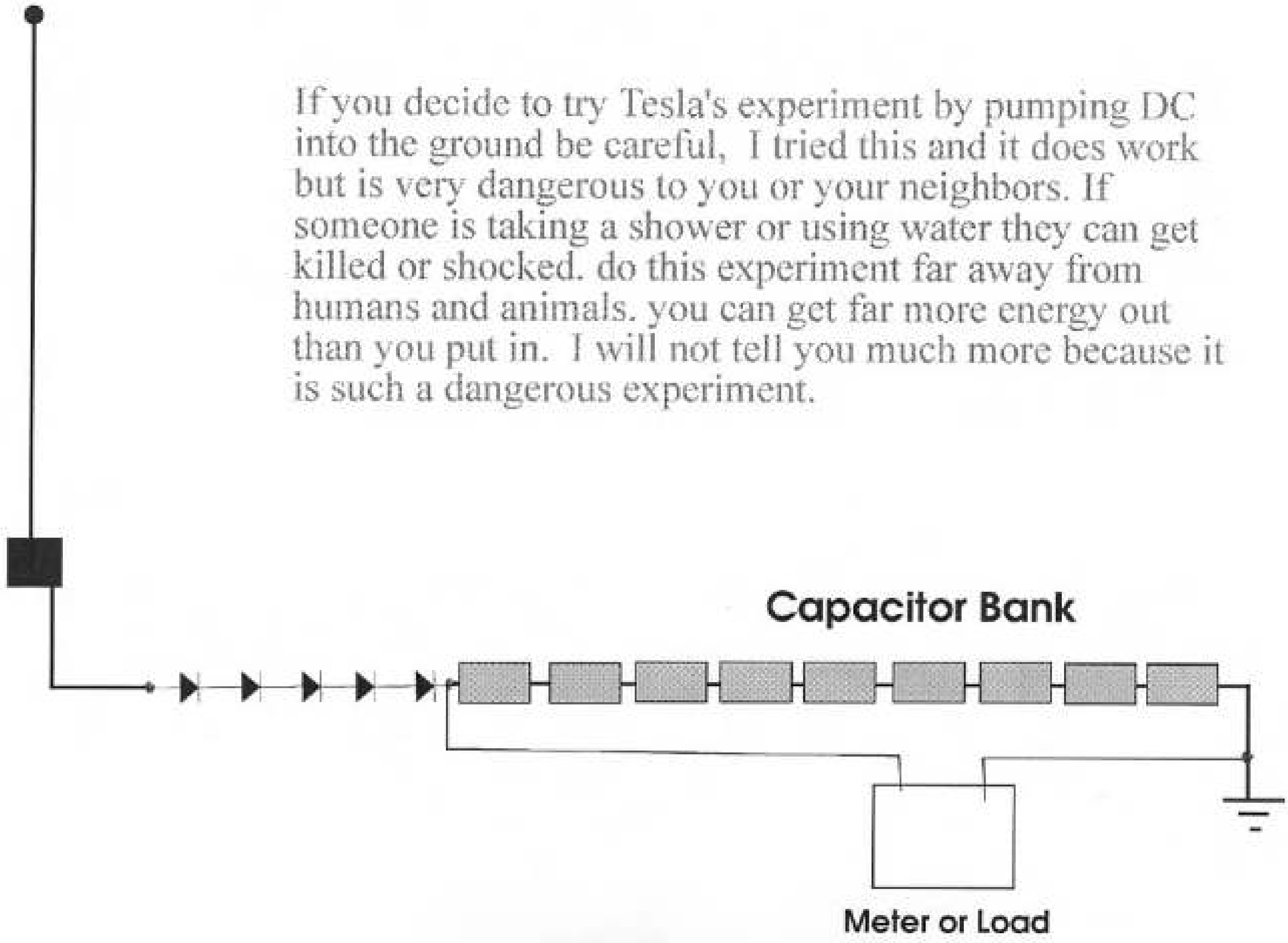

Energy Projects

Electrodynamic Tether Photos

Deployment of the Tether Satellite

Boom Deploy

Initial Deploy

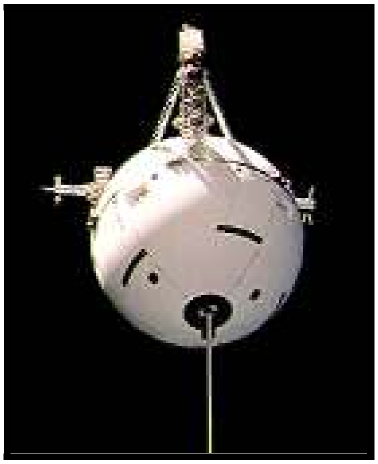

Satellite Deployed

Closeup of boom showing tether

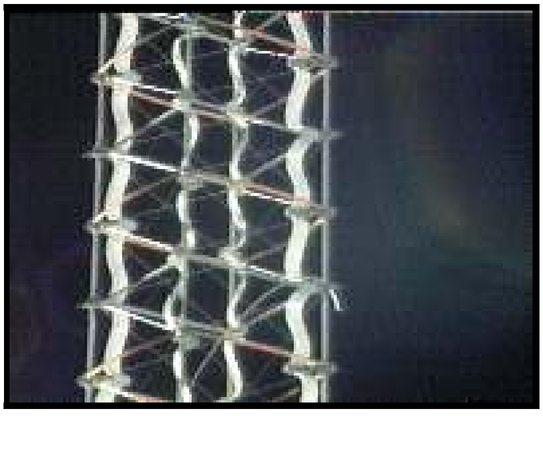

Close up of end of boom

Tether Break

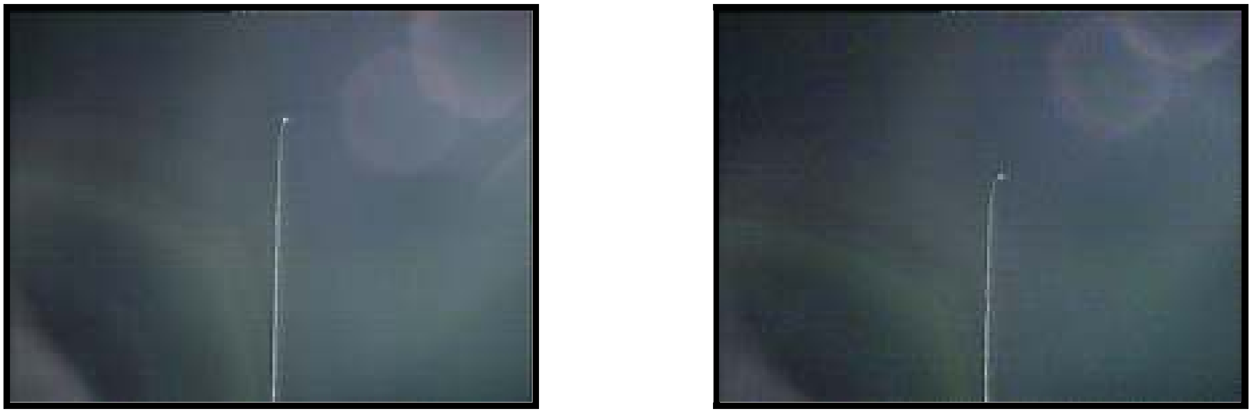

The tether deployed to one mile

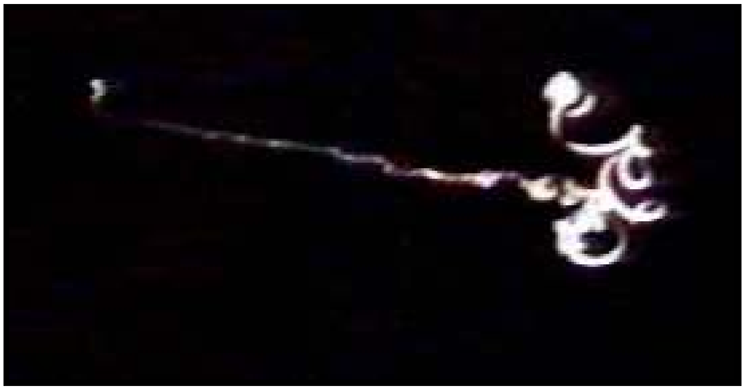

Shortly after the breakreak

The tether coils as it moves away from the Shuttle

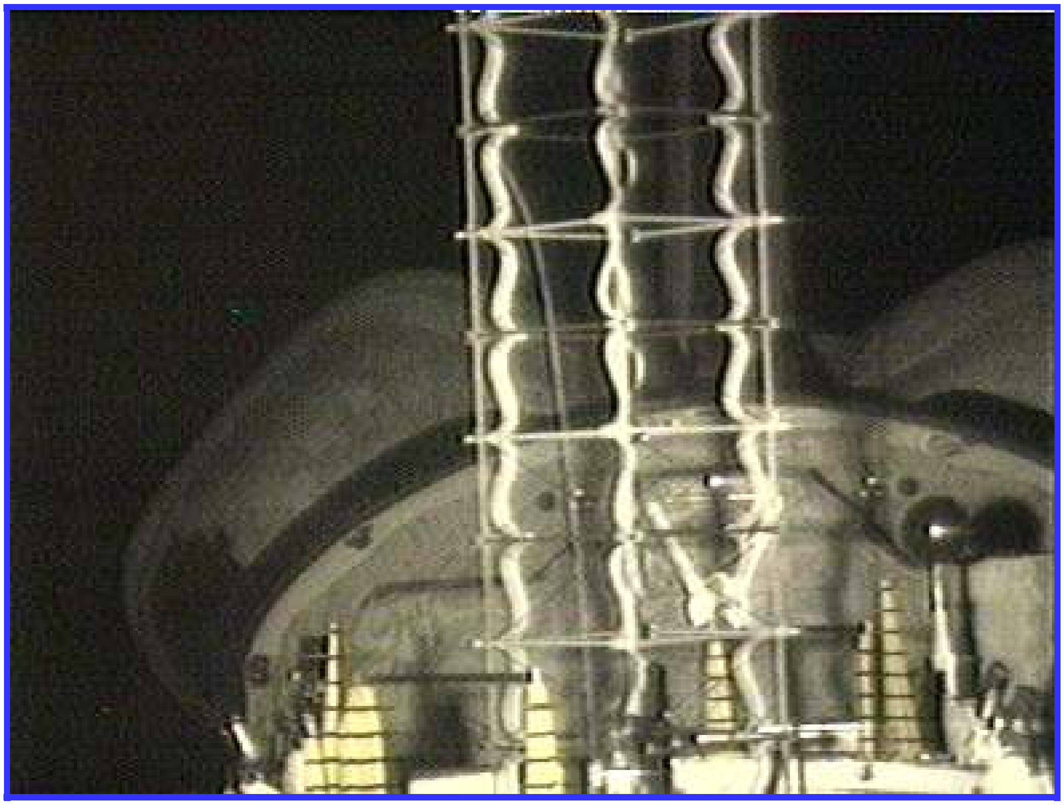

The base of the tether boom. The area looks more like a Tesla Laboratory than the space shuttle hanger bay.

After the Break

The following three images are those displayed for the public showing the white specks that appearall around the tether after it has broken free. NASA calls these "dust and debris" particles, though Mission control in the audio portion of the videos state that they see a "lot of stuff swimming around" and that the tether is wider than expected.

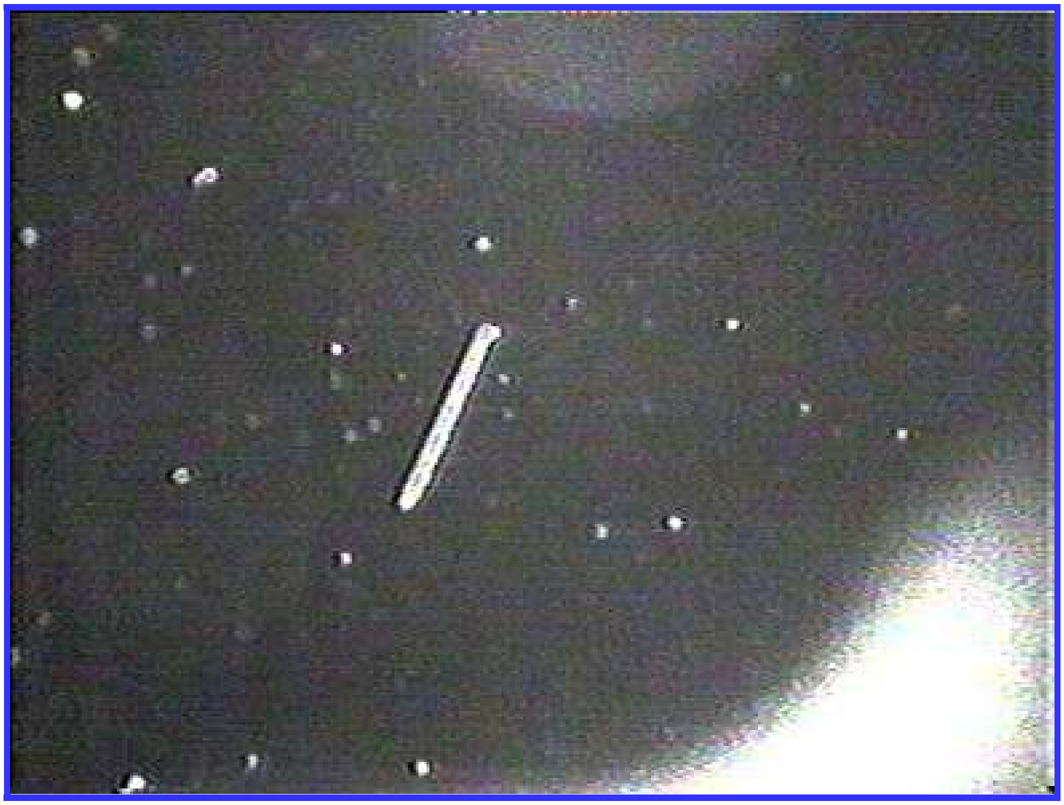

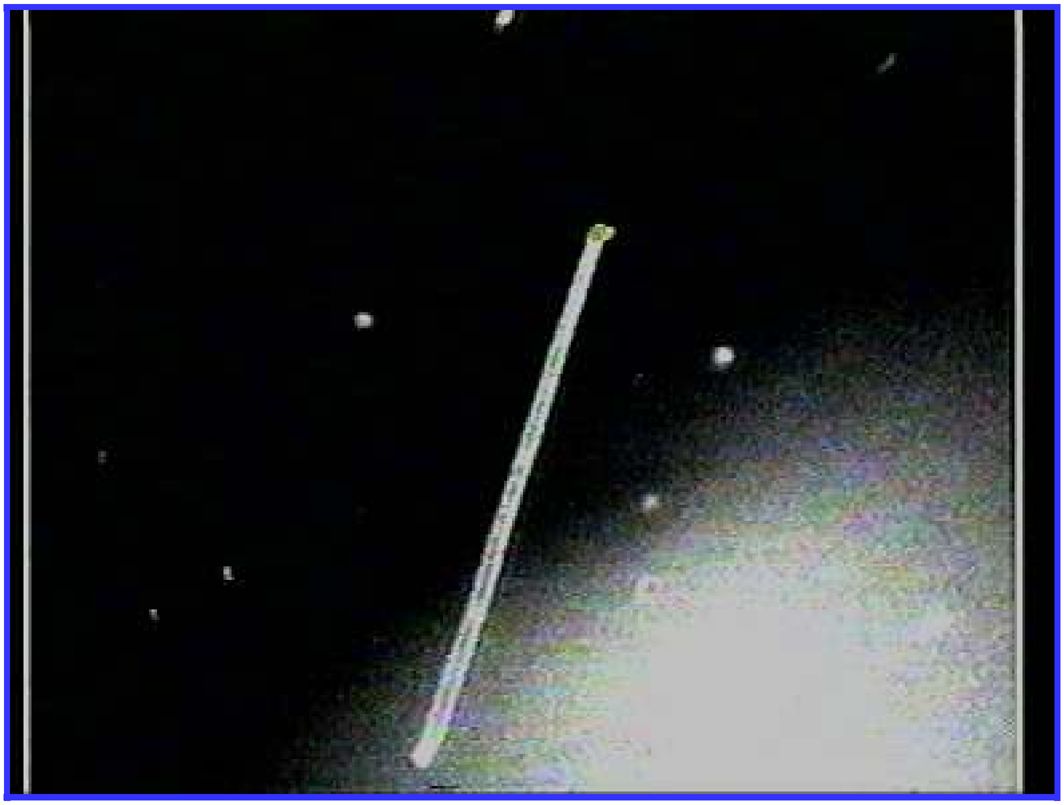

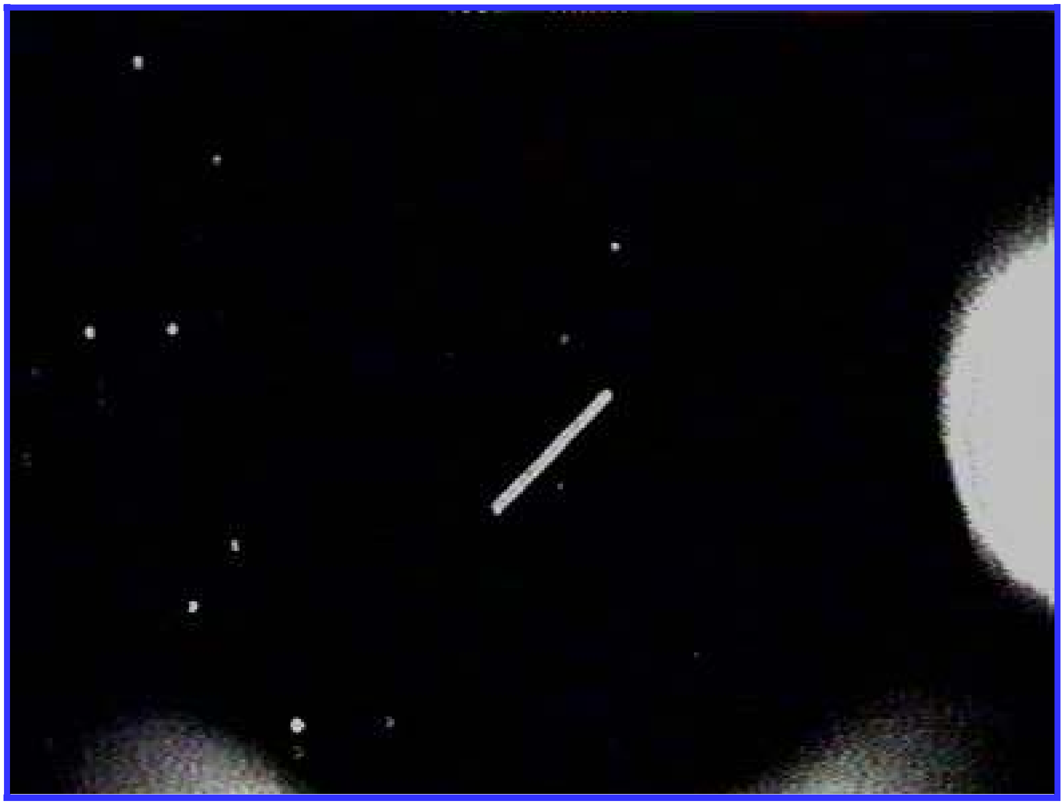

If you compare these images to the ones we captured from a hard to find high resolution copy of the video on this page you will see a remarkale difference. The video is a must see.

Video Clip
STS-75 'The Tether Incident' Feb 1996
Editors Note: This video has been removed several times by Youtube. Currently it is available again. If it does get removed again, email me and I can send you a copy as we have it on disk now. Contact me at webmaster
This one is from the hand held Infrared Camera and is not as clear as the other one was STS-75 Alternate Copy

Comments:

What is most interesting to note is that most of the video clips above are taken from this video, yet NASA does not show the video itself. The copy on Youtube is a transmission intercept copy. However the fact that the NASA info site DOES use those clips gives credibility to the video that it is indeed a NASA film - Zorgon

Tethered Satellite System (TSS-1R)
Source of Clips: NASA STS-75 Mission Pictures

Kennedy Space Center STS-75 Mission Fact Sheet

"Critters" from STS-75

The Report After the Fact

NASA/TP—2003-21228
Low Earth Orbit Spacecraft Charging Design Guidelines

Below are a few keynote excerpts from the PDF file. The entire 367 page document is available below for those with technically savy who want to read the entire paper...

Excerpt 1:
In the case of the TSS-1R tether, its 20 km length produced a maximum of about 3500 V potential between its most positive and negative ends, since it wasn?t oriented perfectly perpendicular to the velocity vector and the Earth?s magnetic field. A satellite at its upper end collected electrons, and an electron gun at the lower end emitted electrons to complete the circuit. When the electron gun was not in operation, a large resistance prevented the Shuttle from being biased thousands of volts negative of its surrounding plasma. However, there remained a large voltage between the tether lower end and the Shuttle orbiter. This enormous bias eventually led to a continuous arc on the tether (see The Continuous Arc, section 4.2.3.1 below), which broke, freeing the satellite and ending the experiment. During the arc, the satellite collected over 1 Amp of electron current to keep the arc going. Probe theory (Cohen et al, 19870010625 N) is usually used to calculate the total current collected by a wire with distributed potentials. However, before the break, TSS-1R demonstrated that a satellite at a high positive potential could collect an anomalously large electron current. See Zhang, et al (20000110580), Stone and Raitt (19990084046 and 20000025437), and Stone, et al (19980202347)

Excerpt 2:
Sustained arcs (continuous arcs) - These are the events that have been attributed with the destruction of on-orbit solar arrays. Generally, the process begins with a fast transient (a so-called trigger arc). Under some conditions, the transient develops into an arc that is fed directly by the entire array, effectively becoming a short-circuit. Such events invariably involve large quantities of energy and can be severely damaging to cells, interconnects or power traces.

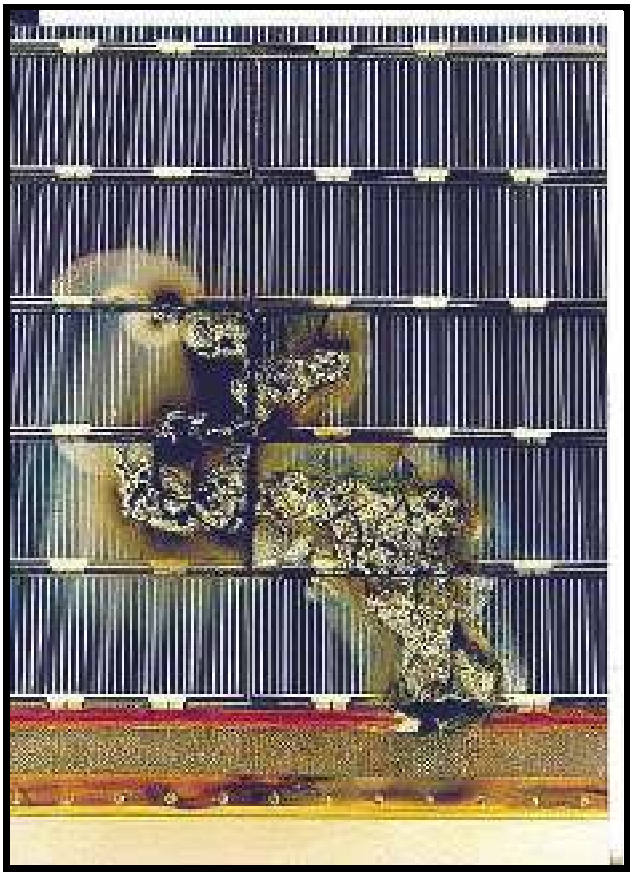

Excerpt 3:

When the structure or array capacitance electrically connected to the arc site is sufficiently large, the initial transient arcs themselves can be large enough to produce significant damage. In Figure 9, we see an anodized aluminum plate that has undergone repeated arcing in the laboratory with the ISS structure capacitance attached. Its thermal properties have been completely destroyed, along with most of the insulating surface layer of aluminum oxide.

Excerpt 4:

The most famous sustained arc event of all led to the breakage of the TSS-1R electrodynamic tether, and the loss of the attached satellite. The image below shows the burned, frayed and broken tether end still attached to the Shuttle after the break. **Incidentally, the tether continued arcing long after it and its satellite were drifting free, until finally it went into night conditions where the electron density was insufficient to sustain the arc.** Noel Sargent (2002) has investigated whether the TSS-1R arc was seen to disrupt Shuttle communications. Although he has found no record of disturbed communications during the event, for most of the time the arc was shielded by metallic structures from the communications antennas, and when the tether broke, the arc was many meters from the receiving antennas. It remains to be seen whether sustained arcs produce radio noise severe enough to be a communications problem.

Comments:

This is official confirmation that the tether continued to produce plasma energy long after it broke free, accounting for the "fluorescent bulb" glowing effect viewed after. We believe this concentrated collection of plasma energy is what attracted the "swarm" of "critters" to a "feeding frenzy"

To get the full PDF file you can download LEO_Charging_Guidelines_v1.3.1.zip

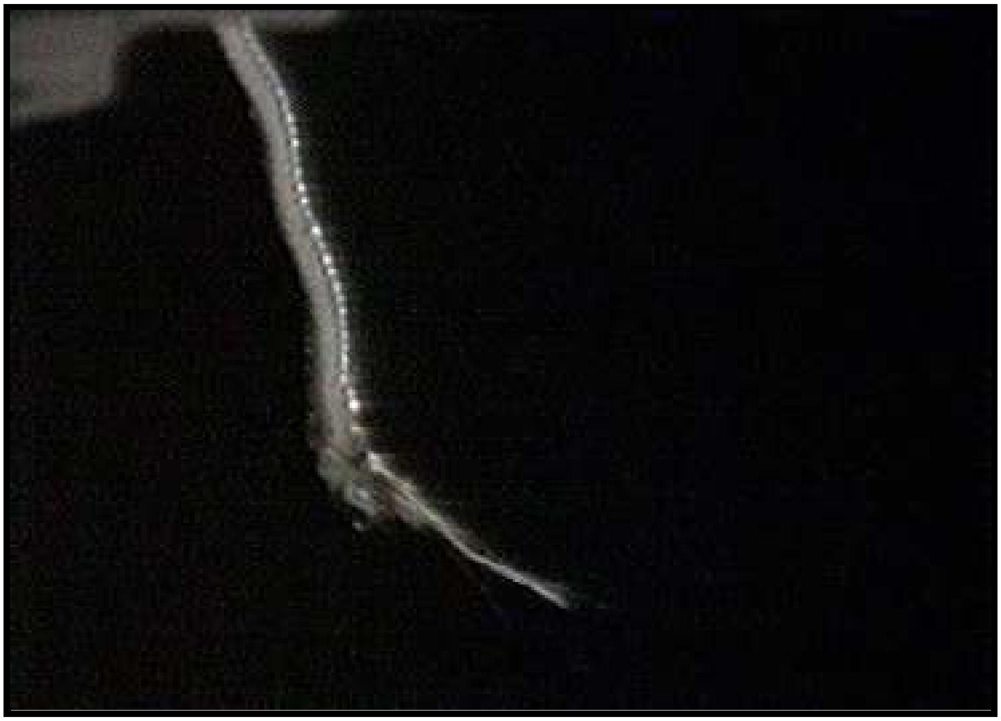

Tether Optical Phenomena Experiment (TOP)

Using a hand-held camera system with image intensifiers and special filters, the TOP investigation will provide visual data that may allow scientists to answer a variety of questions concerning tether dynamics and optical effects generated by TSS-1R. In particular, this experiment will examine the high-voltage plasma sheath surrounding the satellite...

In one mode of operation, the current developed in the Tethered Satellite System is closed by using electron accelerators to return electrons to the plasma surrounding the orbiter. The interaction between these electron beams and the plasma is not well understood...

Associate Investigator: Stephen Mende, Lockheed Martin

SOURCE

The Report After the Fact

...

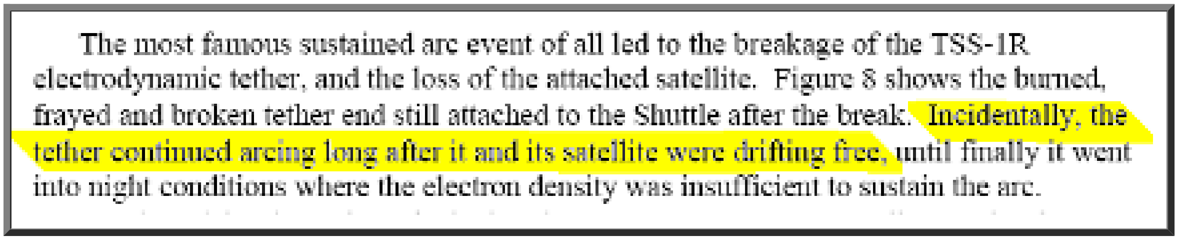

The most famous sustained arc event of all led to the breakage of the TSS-1R electrodynamic tether, and the loss of the attached satellite. Figure 8 shows the burned, frayed and broken tether end still attached to the Shuttle after the break. Incidentally, the tether continued arcing long after it and its satellite were drifting free, until finally it went into night conditions where the electron density was insufficient to sustain the arc.

.....

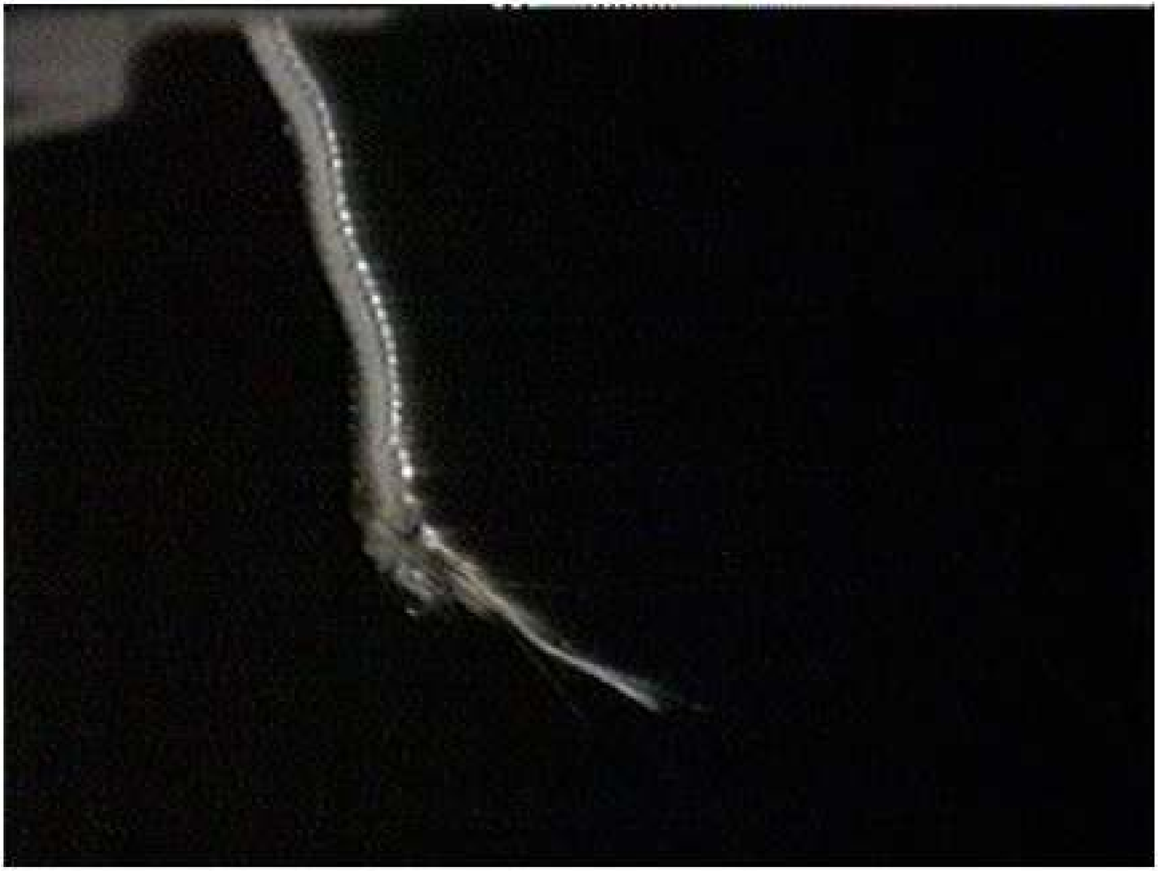

Figure 8 – The end of the remaining TSS-1R tether

…

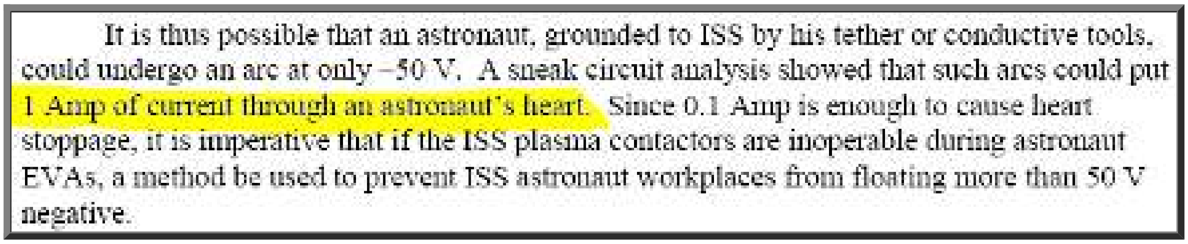

It is thus possible that an astronaut, grounded to ISS by his tether or conductive tools, could undergo an arc at only −50 V. A sneak circuit analysis showed that such arcs could put 1 Amp of current through an astronaut's heart. Since 0.1 Amp is enough to cause heart stoppage, it is imperative that if the ISS plasma contactors are inoperable during astronaut EVAs, a method be used to prevent ISS astronaut workplaces from floating more than 50 V negative.

…

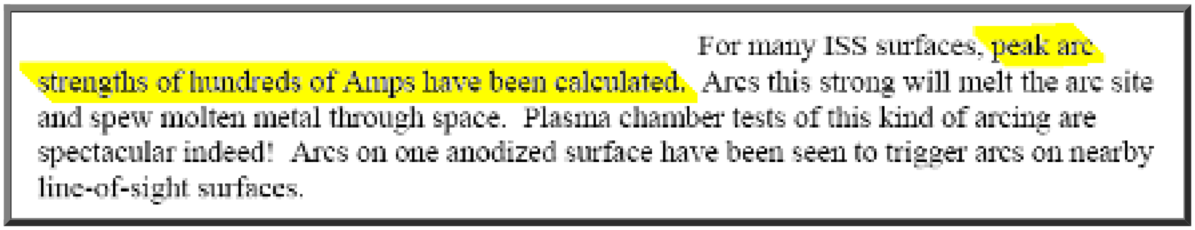

For many ISS surfaces, peak arc strengths of hundreds of Amps have been calculated. Arcs this strong will melt the arc site and spew molten metal through space. Plasma chamber tests of this kind of arcing are spectacular indeed! Arcs on one anodized surface have been seen to trigger arcs on nearby line-of-sight surfaces.

…

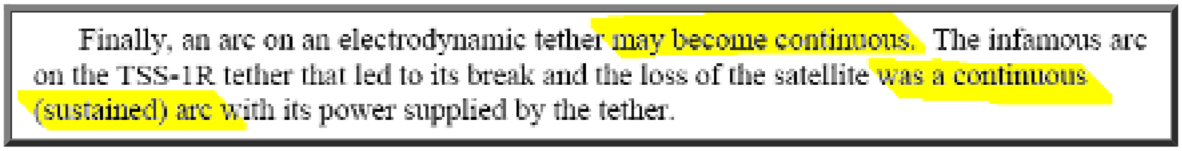

Finally, an arc on an electrodynamic tether may become continuous. The infamous arc on the TSS-1R tether that led to its break and the loss of the satellite was a continuous (sustained) arc with its power supplied by the tether.

…

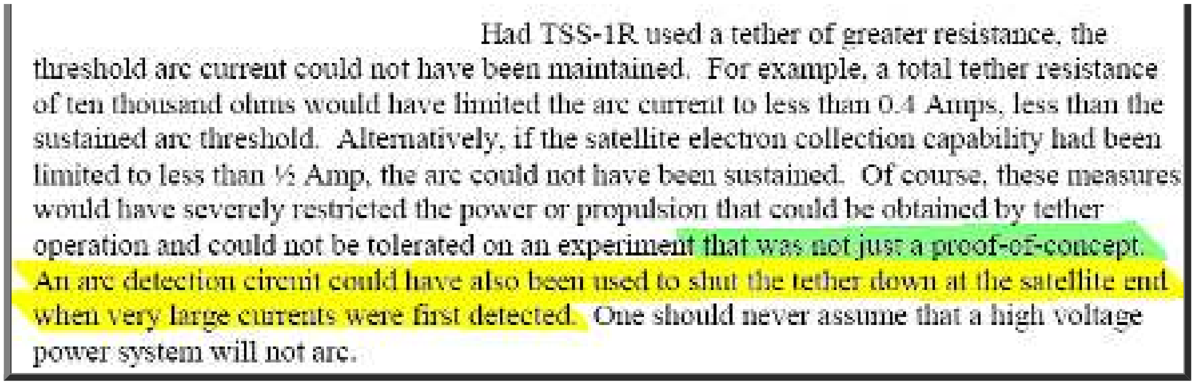

Had TSS-1R used a tether of greater resistance, the threshold arc current could not have been maintained. For example, a total tether resistance of ten thousand ohms would have limited the arc current to less than 0.4 Amps, less than the sustained arc threshold. Alternatively, if the satellite electron collection capability had been limited to less than ½ Amp, the arc could not have been sustained. Of course, these measures would have severely restricted the power or propulsion that could be obtained by tether operation and could not be tolerated on an experiment that was not just a proof-of-concept. An arc detection circuit could have also been used to shut the tether down at the satellite end when very large currents were first detected. One should never assume that a high voltage power system will not arc.

...

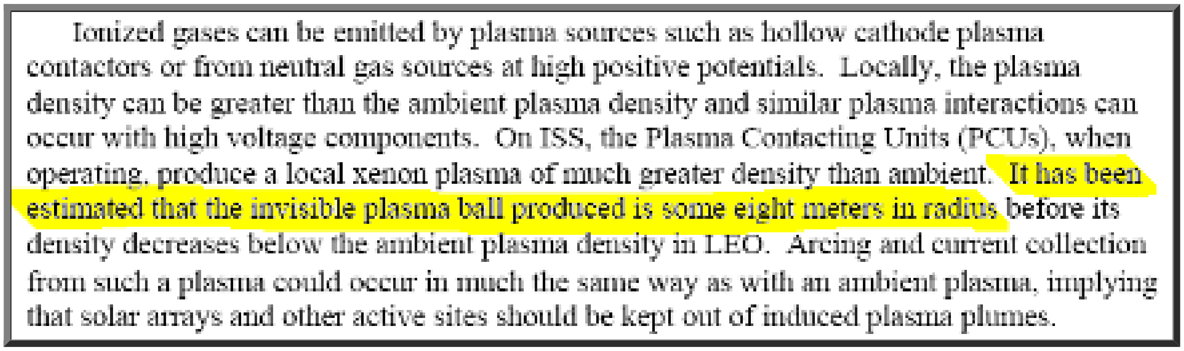

Ionized gases can be emitted by plasma sources such as hollow cathode plasma contactors or from neutral gas sources at high positive potentials. Locally, the plasma density can be greater than the ambient plasma density and similar plasma interactions can occur with high voltage components. On ISS, the Plasma Contacting Units (PCUs), when operating, produce a local xenon plasma of much greater density than ambient. It has been estimated that the invisible plasma ball produced is some eight meters in radius before its density decreases below the ambient plasma density in LEO. Arcing and current collection from such a plasma could occur in much the same way as with an ambient plasma, implying that solar arrays and other active sites should be kept out of induced plasma plumes.

...

Electrodynamic Tethers ~ Menu
Critters ~ Critters 01 ~ Critters 02 ~ Critters 03

Copyright Notice

All pictures and photos on this page, unless otherwise noted, have been gathered from pubic domain sources or are available under GNU License. Further documentation is available on our Copyright Page and our Legal Department.

Article Quotes all have links to their appropriate source and are noted. All other text not so noted is available under the terms of the GNU Free Documentation License

All information on this page is used only for non profit educational purposes. All reasonable attempts have been made that no credits are missed, but with a work of this size omissions may occur. If you see anything of yours that has not been properly credited or wish to have removed please contact the Webmaster

Menu

Webpages © 2001-2006
Blue Knight Productions

STS 75 TETHER INCIDENT

Tether Deployed

Tether Satellite being deployed from the Shuttle Bay of STS-75

219506148 DAY:056

Tether Activated

Tether Satellite begins to glow as it gathers electricity from the Ionosphere

213809089 DAY:056

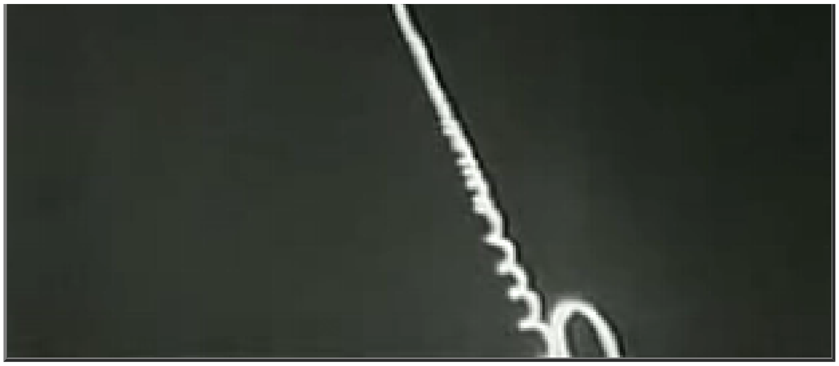

Tether Breaks Free

Tether Satellite cable fries the connection and breaks free... still producing energy. The glow increases and the cable gets wider. (More on this later)

-

Tether Broke Free

Tether Satellite worth 100 million has overloaded the circuits within minutes of being activated because the NASA Scientists under estimated the potential. The tether in just a few moments of operation produced 10 times the anticipated power and they forgot to install a circuit breaker.

-

NOTES:
Add "Critter" Portion
Add Video
Add FLV Viewer

TETHER BROKEN FREE AND DRIFTING AWAY

Source: Original Image NASA Cassini/ Huygens N00008771.jpg

~ MENU ~

US006116544A

United States Patent [19]

Forward et al.

[11] **Patent Number: 6,116,544**

[45] **Date of Patent: Sep. 12, 2000**

[54] **ELECTRODYNAMIC TETHER AND METHOD OF USE**

[75] Inventors: **Robert L. Forward**, Houston, Tex.; **Robert P. Hoyt**, Seattle, Wash.

[73] Assignee: **Tethers Unlimited, Inc.**, Seattle, Wash.

[21] Appl. No.: **08/929,271**

[22] Filed: **Sep. 12, 1997**

[51] **Int. Cl.**^{7} **B64G 1/00**

[52] **U.S. Cl.** .... **244/158 R**; 244/166; 244/172

[58] **Field of Search** .... 244/158 R, 164, 244/166, 172

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,097,010	6/1978	Colombo et al.	244/158 R
4,824,051	4/1989	Engelking	244/158 R
4,923,151	5/1990	Roberts	244/158 R

OTHER PUBLICATIONS

"Tethers Open New Space Options" by Ivan Bekey, *Astronautics & Aeronautics*, 1983.

Primary Examiner—J. Woodrow Eldred
Attorney, Agent, or Firm—Arthur M. Dula

[57] **ABSTRACT**

The present invention comprises an electrodynamic tether structure and a method of use. The structure of the tether taught by the present invention is a short, wide, interconnected multiwire (compared to the long, narrow single wires of the prior art) conductive tether whose area maximizes electrodynamic drag while simultaneously minimizing the Area-Time-Product swept by the tether during its operating life. The preferred tether length is two kilometers to five kilometers. The preferred tether mass is one percent to five percent of the spacecraft mass. The method of operation comprises orienting the tether structure at an angle to the local vertical to maximize electrodynamic drag on the host spacecraft and minimize tether instability. The angle of 35.26 degrees is preferred.

26 Claims, 12 Drawing Sheets

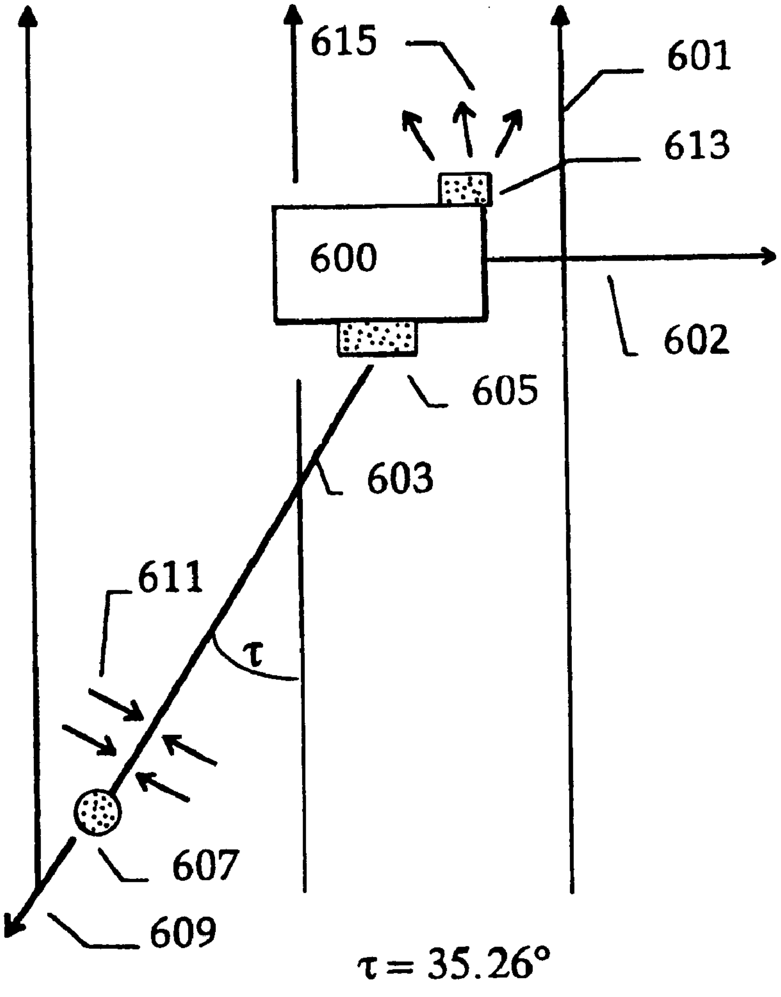

τ = 35.26°

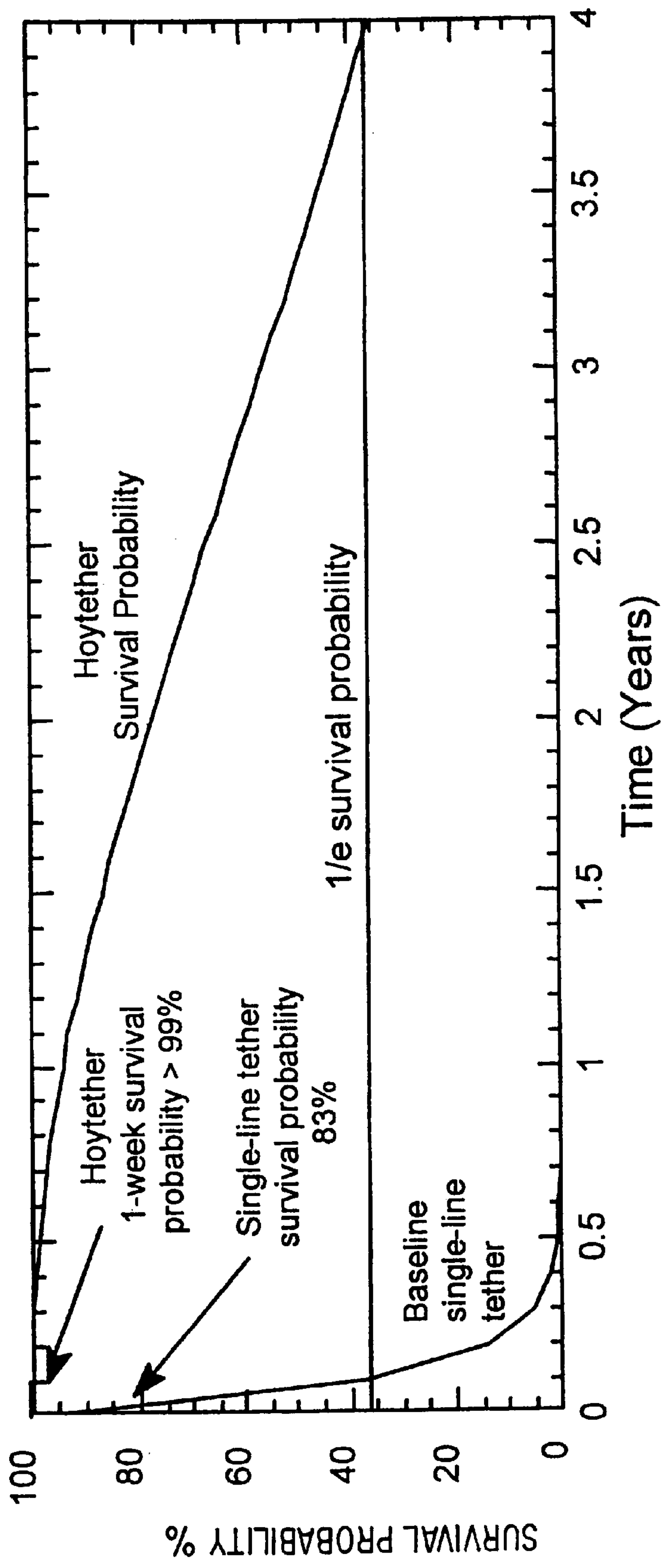
Hoytether
Survival Probability
1/e survival probability
Hoytether
1-week survival
probability > 99%
Single-line tether
survival probability
83%
Baseline
single-line
tether
SURVIVAL PROBABILITY %
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time (Years)

FIG. 1a

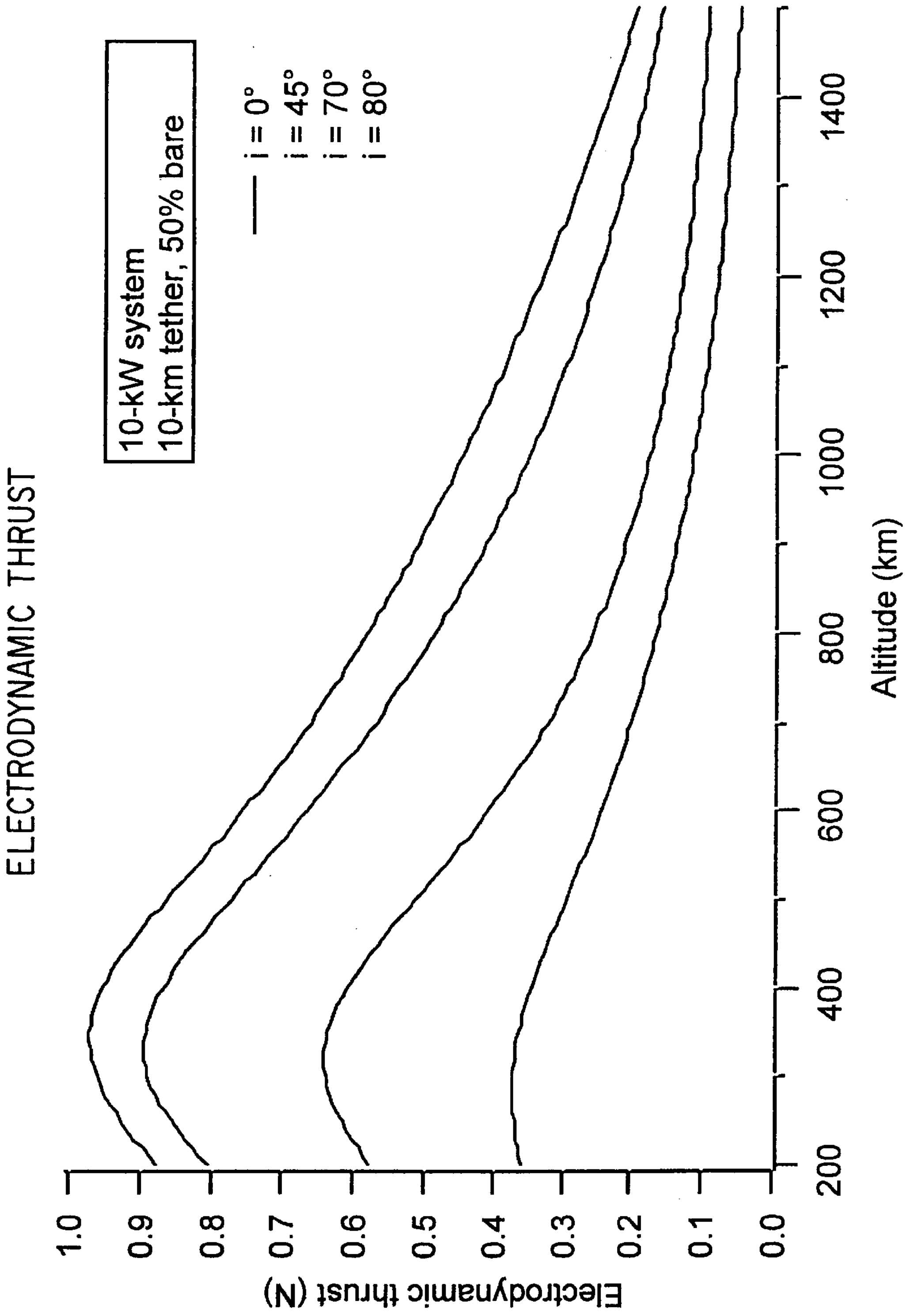
ELECTRODYNAMIC THRUST
10-kW system
10-km tether, 50% bare
i = 0°
i = 45°
i = 70°
i = 80°
Electrodynamic thrust (N)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
200
400
600
800
1000
1200
1400
Altitude (km)

FIG. 1b

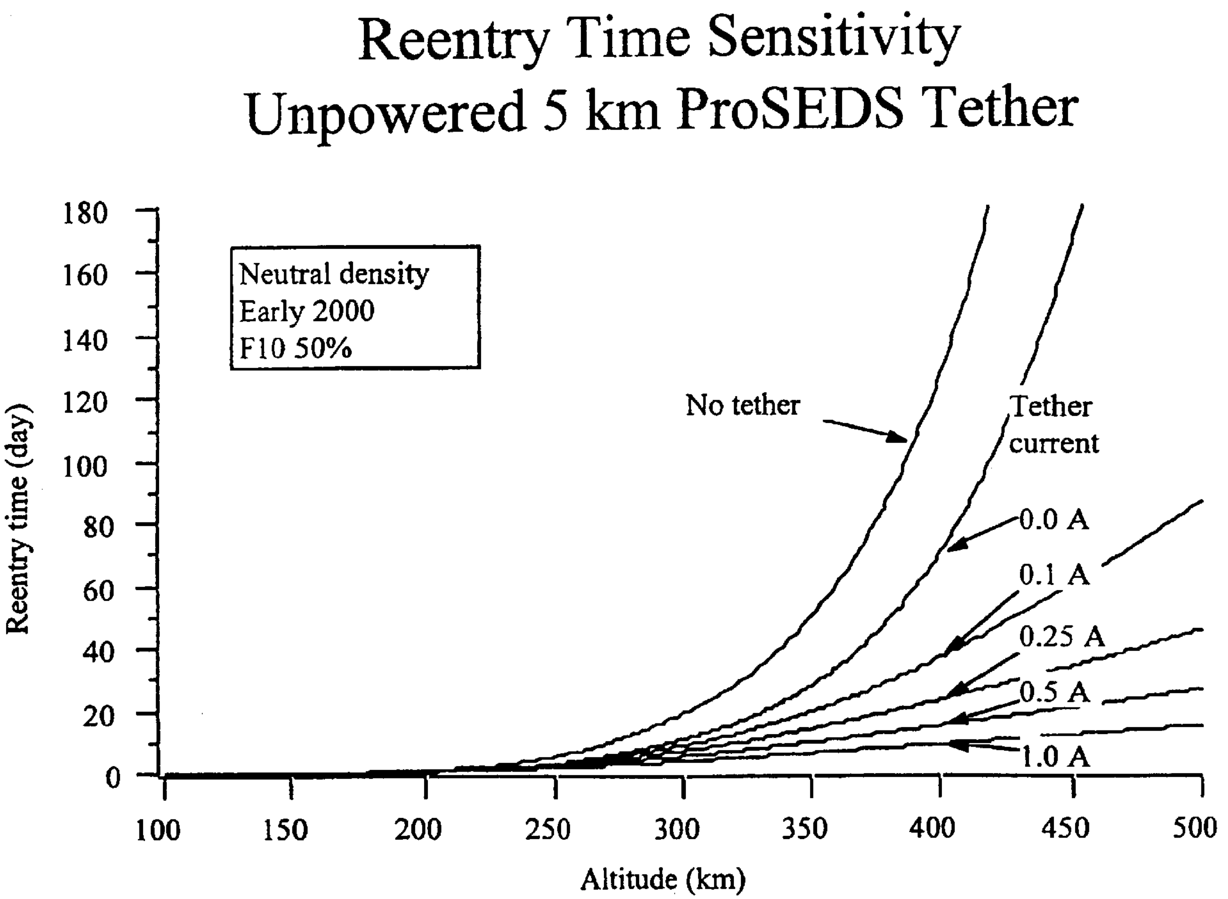

Enrico Lorenzini, SAO, NASA ProSEDS Mission (July 1997),
NASA/MSFC Grant NAG8-1303

FIG. 1c

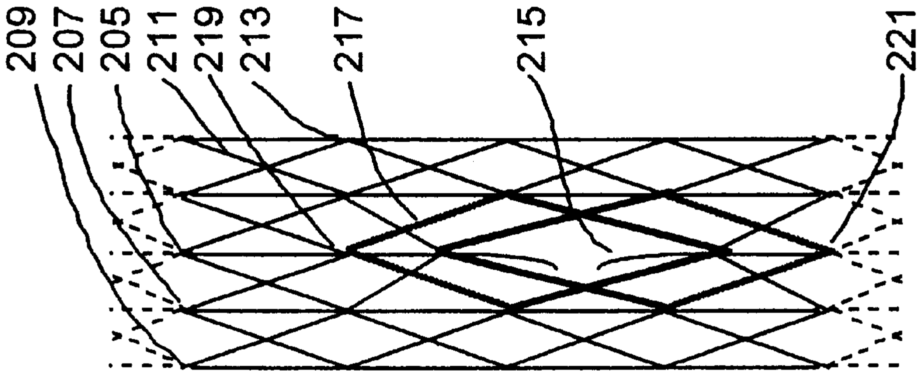
209
207
205
211
219
213
217
215
221

FIG. 2c

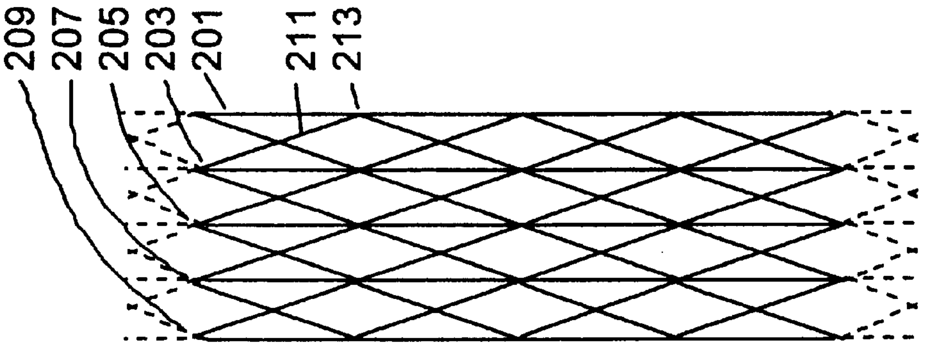
209
207
205
203
201
211
213

FIG. 2b

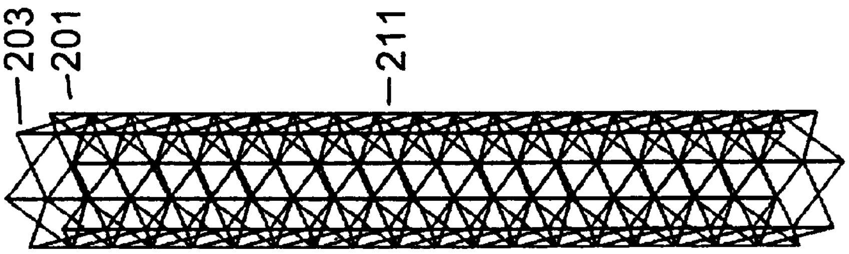

FIG. 2a

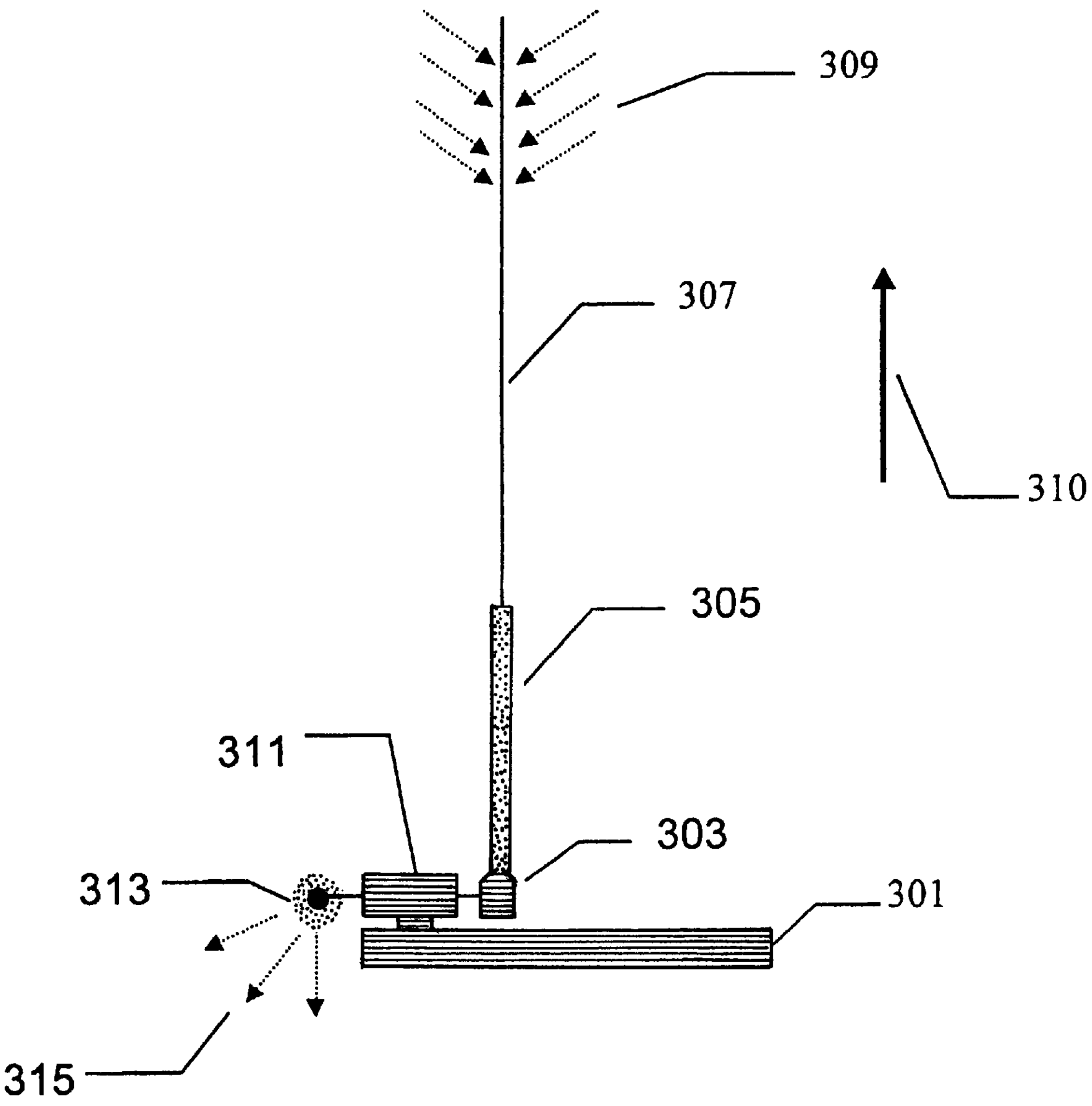
309
307
310
305
311
303
313
301
315

FIG. 3

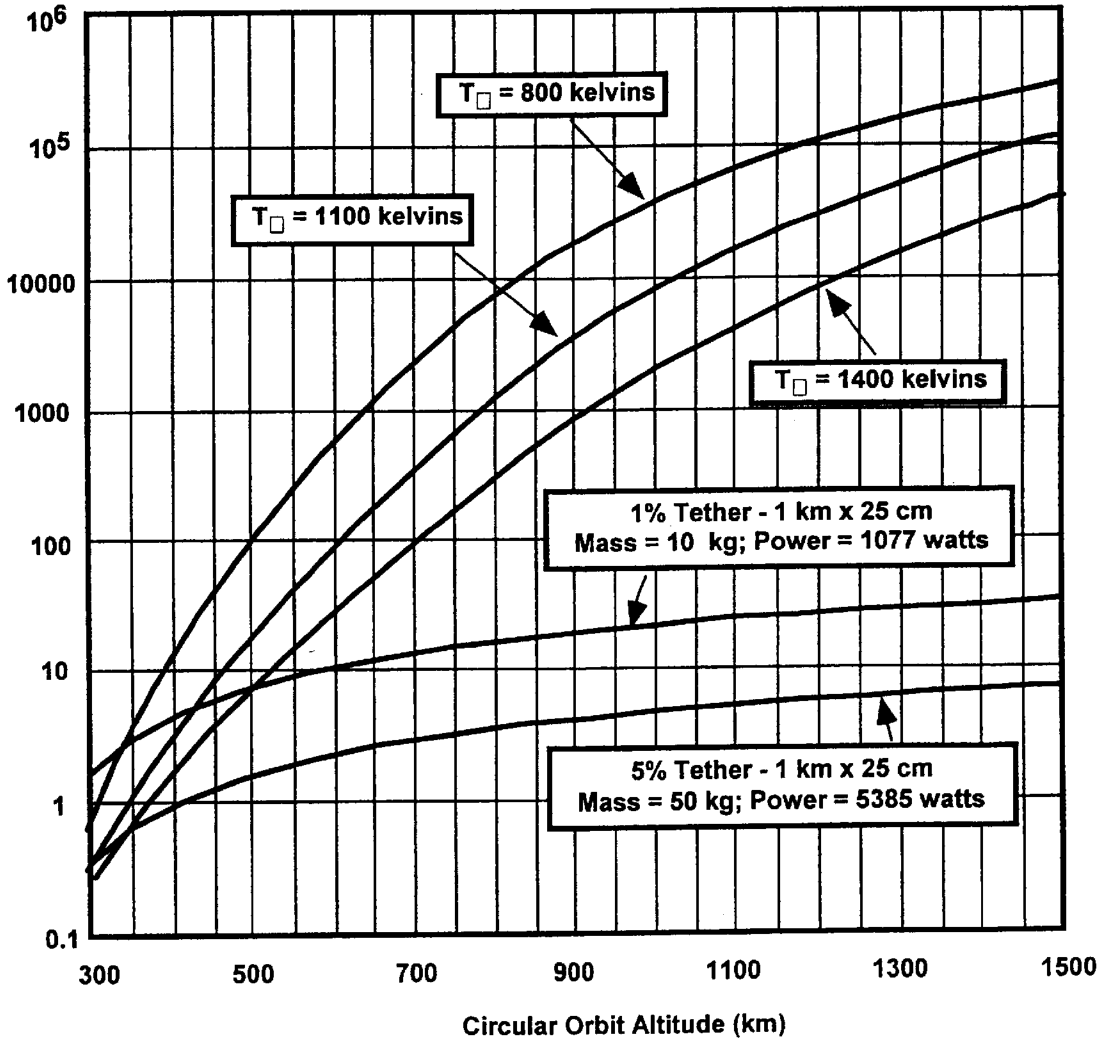

Area-Time Product for Mean and Extremes of Exospheric Temperature

(m = 1000 kg, Cd = 2.0 ; 1% Tether, 1 km x 25 cm; power = 1077 watts)

(5% Tether, 1 km x 25 cm; power =5385 watts)

FIG. 4

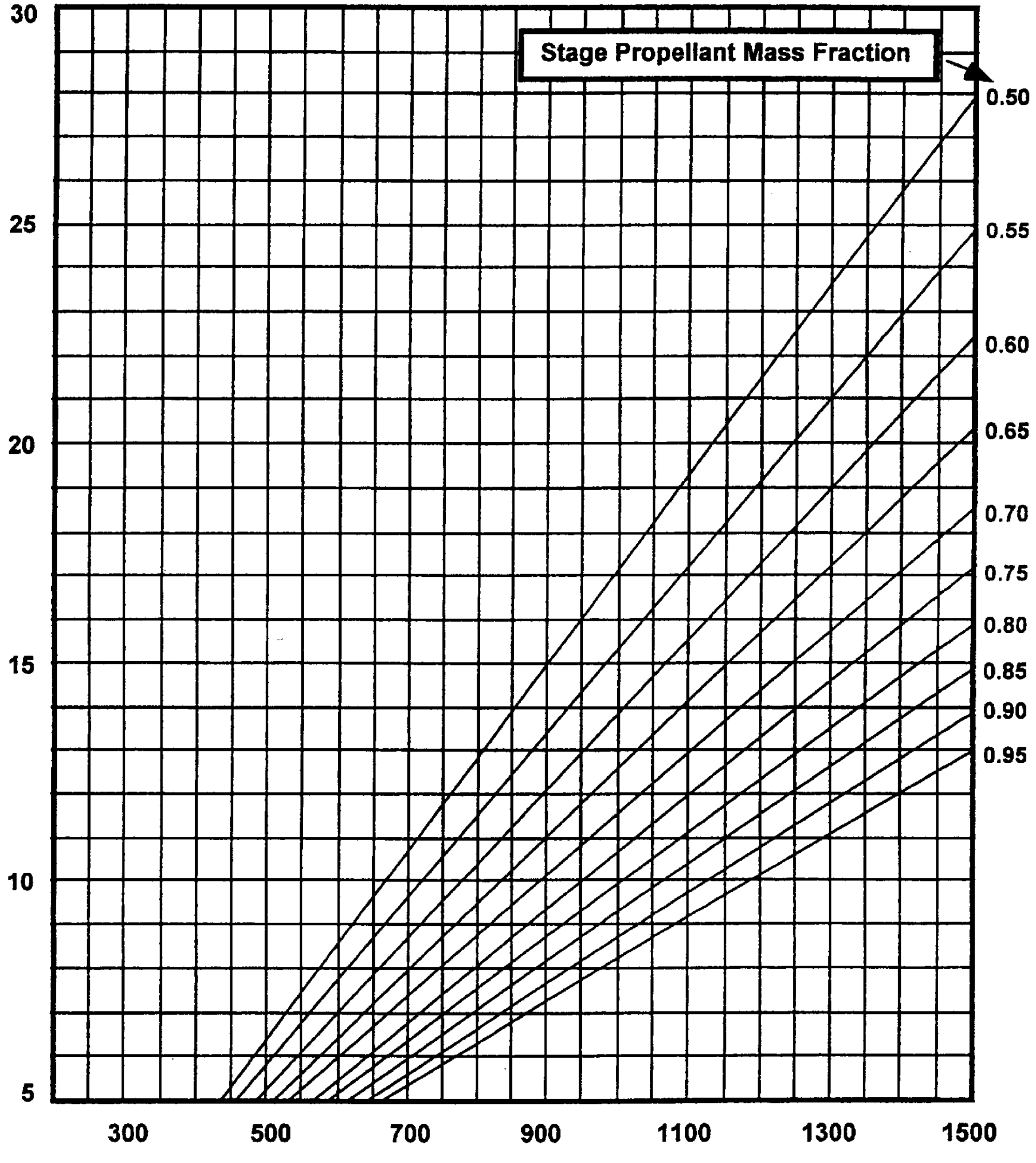

Conventional Solid Rocket Motor System

Percent Increase in Orbital Mass vs. Altitude and I' (Isp = 288 sec)

FIG. 5

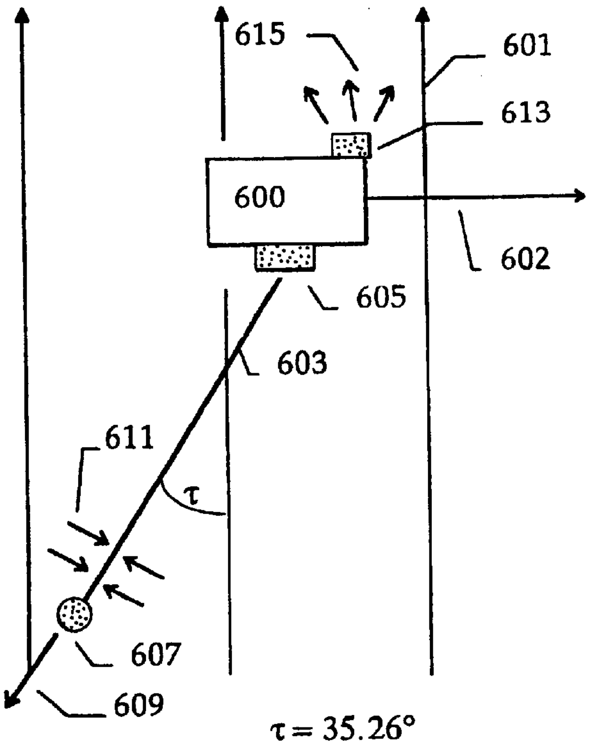
τ
τ = 35.26°

FIG. 6

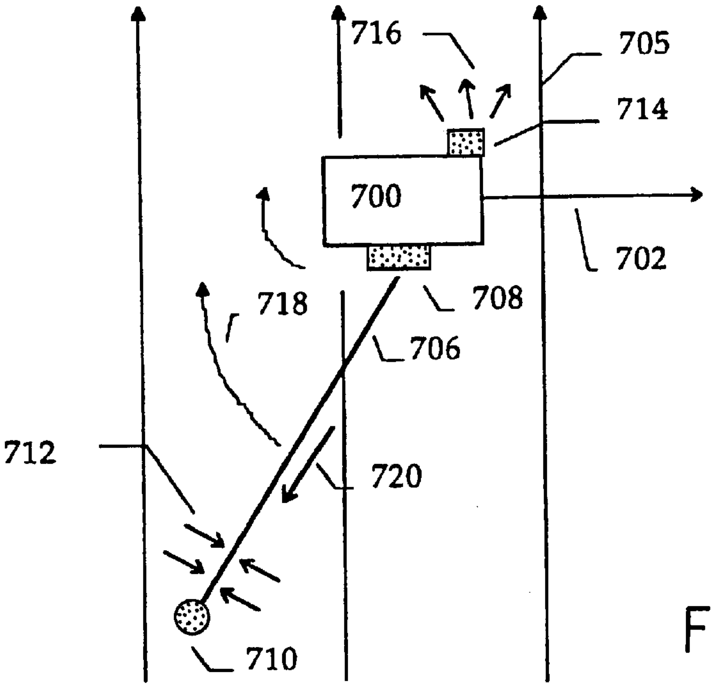

FIG. 7

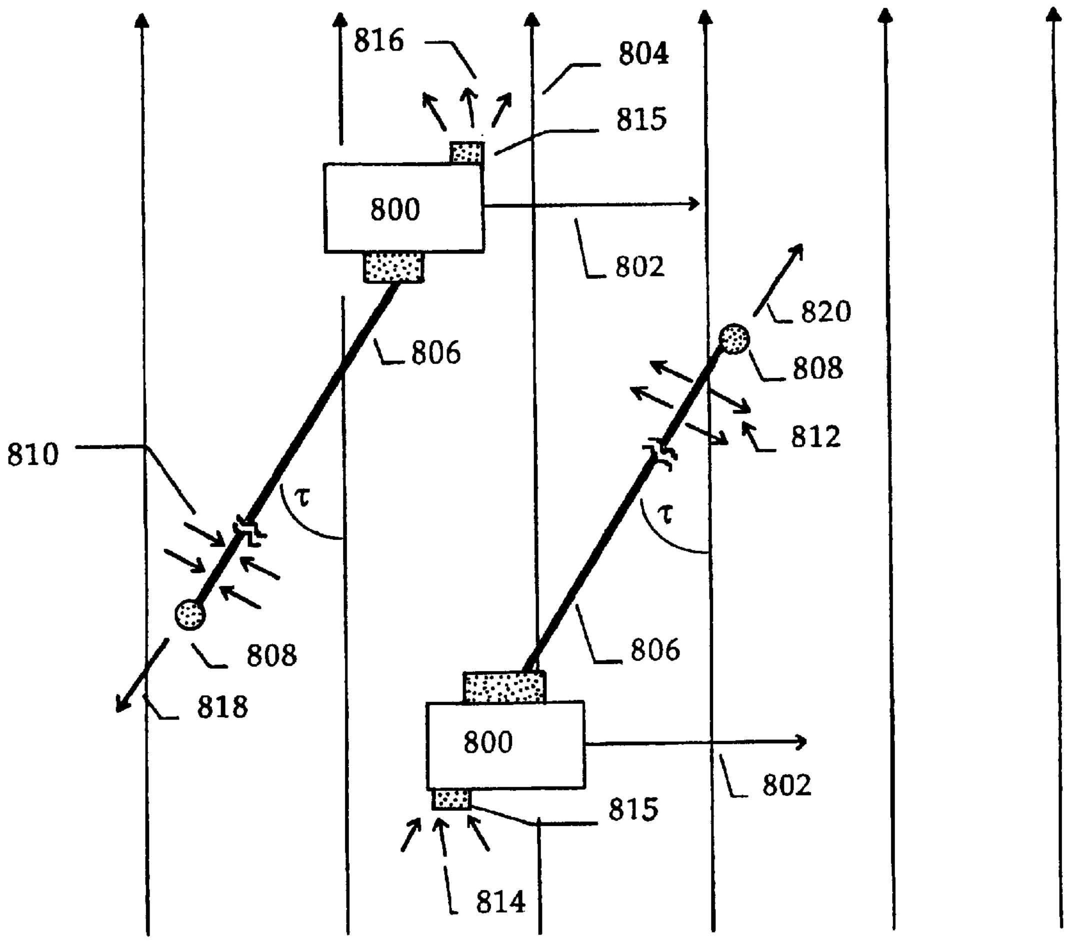
τ
τ

FIG.8

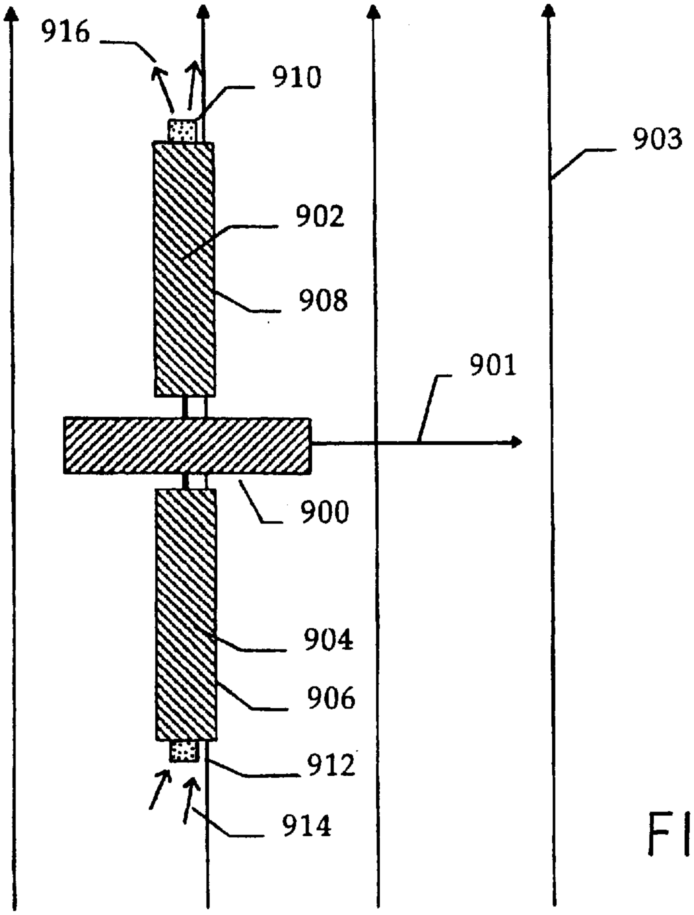

FIG. 9

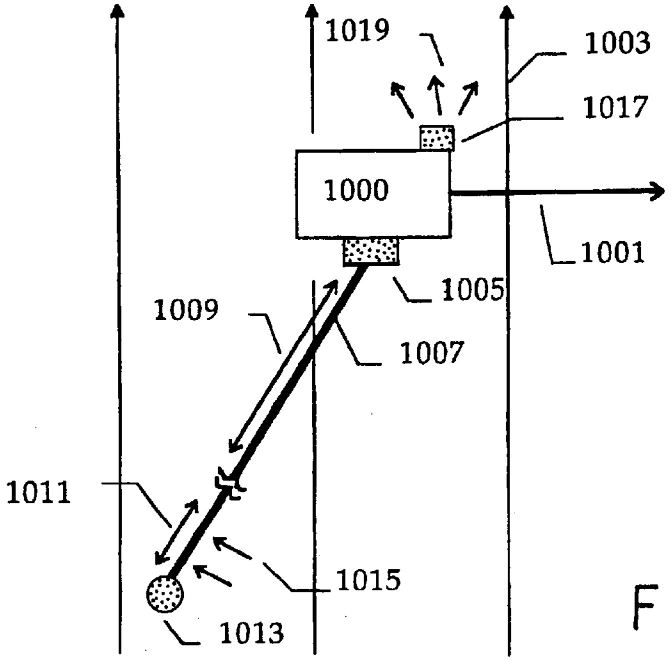

FIG. 10

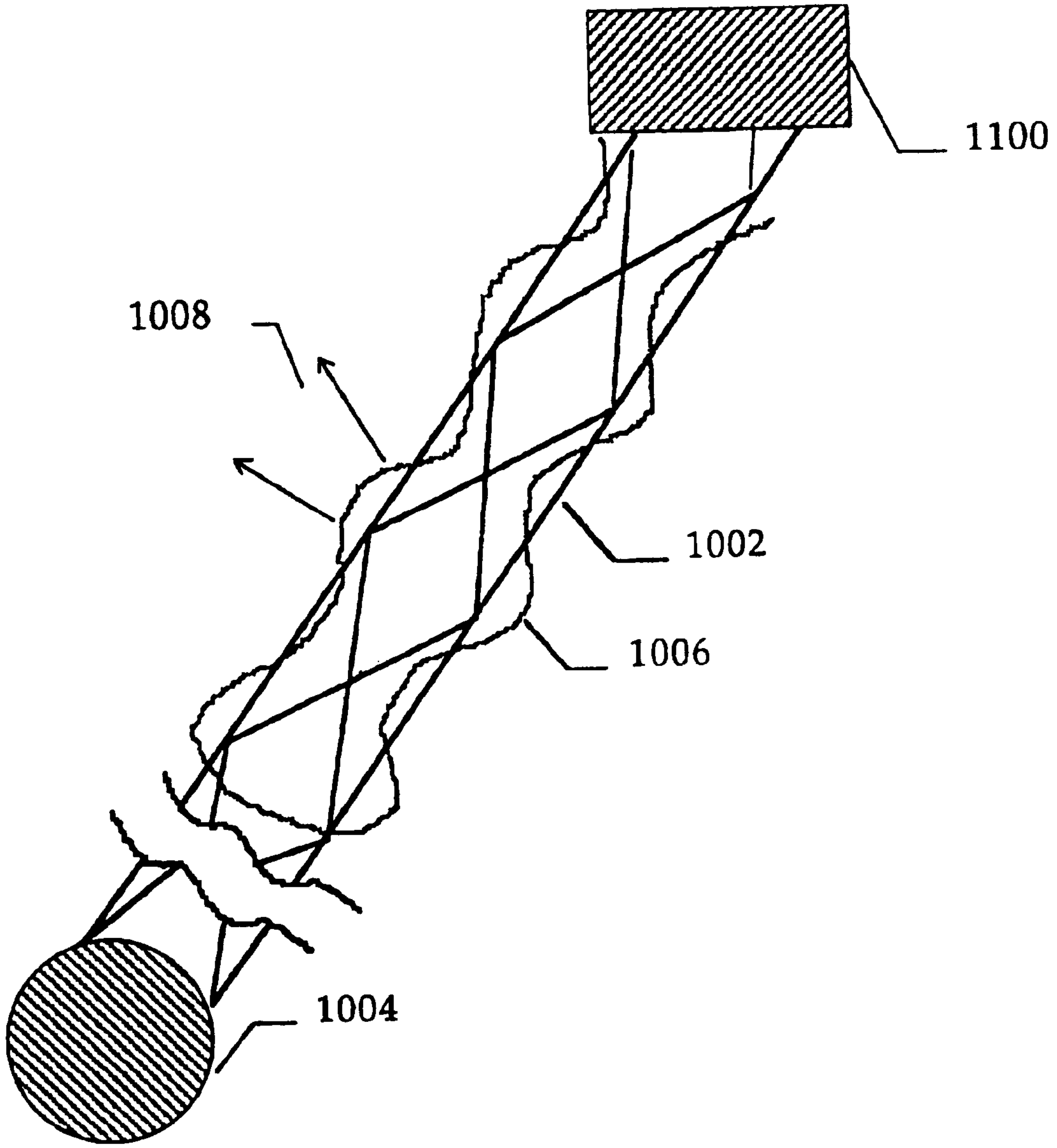
1100
1008
1002
1006
1004

FIG. 11

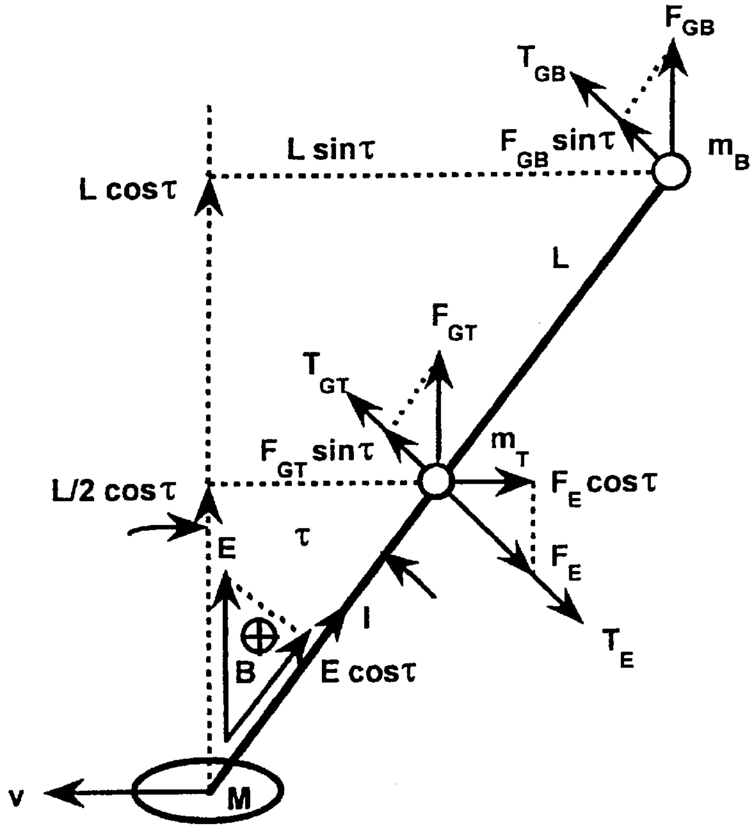
F_GB
T_GB
F_GB sinτ
m_B
L sinτ
L cosτ
L
F_GT
T_GT
F_GT sinτ
m_T
F_E cosτ
L/2 cosτ
τ
E
F_E
T_E
I
B
E cosτ
v
M

FORCES AND TORQUES ON AN UPWARD DEPLOYED CONDUCTING TETHER DUE TO THE MOTION OF THE HOST SPACECRAFT THROUGH THE MAGNETIC FIELD OF THE EARTH.

FIG. 12

ELECTRODYNAMIC TETHER AND METHOD OF USE

TECHNICAL FIELD

This invention relates generally to apparatus and methods useful for changing the state vector of a space object when the space object is moving relative to a magnetic field. More specifically, the present invention relates to an apparatus and method of using a conducting tether to produce an electrodynamic force to deorbit a satellite from its orbit around a celestial body, such as the Earth, which has an associated magnetic field.

RESERVATION OF RIGHTS

This application is subject to certain rights of the U.S. government as a result of contracts between the U.S. government and the inventors.

BACKGROUND ART

The present invention has its principal utility in outer space, primarily for deorbiting satellites at the end of their useful life to mitigate the harm and reduce the liability created by the proliferation of space debris. In order to obtain a better understanding of the present invention it is helpful to understand the prior art of space tethers, especially tether dynamics and tether electrodynamics. The present invention may be more readily understood through a review of the experimental prior art and a mathematical analysis of electrodynamic space tethers.

Prior Art Tethers:

A tether was originally a rope or chain used to fasten an animal so that it grazed only within certain limits. Tethers have been used for decades in space to attach astronauts to their spacecraft.

In 1974 Professor Guiseppe Colombo, holder of the Galileo chair of astronomy at the University of Padua in Italy, proposed using a long tether to support a satellite from an orbiting platform. U.S. Pat. No. 4,097,010, which issued to Professor Colombo and Mario Grossi on Jun. 27, 1978, teaches a satellite connected by means of a long tether to a powered spacecraft. Colombo actively pursued the design of a tethered satellite system.

Several NASA experiments, such as the two Small Expendable Deployer System (SEDS 1 & 2) experiments and the Plasma Motor Generator (PMG) experiment used tethers in space. SEDS used a nonconducting tether. The PMG used a 500-meter conducting tether. The Tethered Satellite System flights in 1992 and 1996 (TSS-1 & 1R) used a 20,000-meter conducting tether.

On the TSS-1 mission the tether deployed only 260 meters (853 feet) before the deployer failed. On the TSS-1R the tether was deployed 19,500 meters. In the SEDS-2 flight, a 0.8-mm diameter, 20,000-meter long braided single-line tether was deployed to study tether dynamics and lifetime. Orbital debris or a meteoroid severed this tether in less than four days.

In the TSS-1R flight, the conducting single-line tether was severed after five hours of deployment. This failure was caused by an electric arc produced by the 3,500 volts of electric potential generated by the conductive tether's movement through the Earth's magnetic field.

The Tether Physics and Survivability (TiPS) satellite consists of two end masses connected by a 4,000-meter long non-conducting tether. This satellite was deployed on Jun. 20, 1996 at an altitude of 1,022 kilometers (552 nautical miles). Its tether is an outer layer of Spectra™ 1000 braid over a core of acrylic yarn. The yarn will "puff" its outer braid to two millimeters to "give it a larger cross section to improve its resistance to debris and small micrometeoroids", according to the National Reconnaissance Office (NRO), which is a sponsor of the TiPS mission. As of Jun. 21, 1997 the TiPS tether had survived one year.

References:

1. Joseph A. Carroll, "SEDS Deployer Design and Flight Performance", paper WSEDSA-1 at the 4th International Conference on Tethers in Space, Washington, D.C., April 1995.
2. James E. McCoy, et. al. "Plasma Motor-Generator (PMG) Flight Experiment Results", pp.57–84, Proceedings of the 4th International conference on Tethers in Space, Washington, D.C., April 1995.
3. W. John Raitt, et. al. "The NASA.ASI-TSS-1 Mission, Summary of Results and Reflight Plans, pp. 107–118, Proceedings of the 4th International conference on Tethers in Space, Washington, D.C., April 1995.
4. Joseph C. Anselmo, "NRO Orbiting Spacecraft Studies Tether Survivability", Aviation Week, page 24, Jul. 1, 1996.

These experiments all used single line tethers.

The following reference is illustrative of the current state of the art in space tethers: Paul A. Penzo and Paul W. Ammann. Tethers in Space Handbook—Second Edition. NASA Office of Space Flight, NASA Headquarters, Washington, D.C. 20546. See also the hundreds of references in the 33 page bibliography at the end of the handbook.

The "Hoytether"™, an Improved, High-Reliability Tether:

In 1991, one of the present inventors, Robert Hoyt, invented a lightweight net-like structure that provides many redundant load-bearing paths. A number of primary load bearing lines running the length of the structure are connected periodically by diagonal secondary lines. The disclosed embodiment of this invention has the secondary lines firmly connected by knots to the primary lines. The secondary lines are connected only to the primary lines. At either end of the disclosed structure, a support ring enforces the cylindrical spacing between the primary lines. The secondary lines are designed with a small amount of slack. These secondary lines are only put under load if a primary line fails. This specific tether structure was disclosed to the public in 1992 (Forward, R. L., "Failsafe Multistrand Tether Structures for Space Propulsion", AIAA paper 92–3214, 28th Joint Propulsion Conference, Nashville, Tenn., 1992 (hereinafter "1992 Hoytether structure"). This structure was named a "Hoytether". The term "Hoytether" is used throughout the remainder of this specification for this type of tether structure.

The present invention uses an improved Hoytether, which was invented by the same inventors as the present invention. This improved Hoytether is the subject of a copending PCT application. The Hoytether is discussed briefly in this specification to aid understanding of the present invention.

The 1992 Hoytether design teaches that the normally slack secondary lines have half the cross-section (0.707 the diameter) of the primary lines. There are twice as many secondary lines as primary lines, thus the mass of the secondary lines is equal to the mass of the primary lines. In an undamaged Hoytether, the primary lines carry the entire load, while none of the secondary lines are under load.

While the survival probability of a single-line tether decreases exponentially with time, the Hoytether can maintain a high, i.e. greater than 99 percent, survival probability for periods of months or years (forward and Hoyt, "Failsafe Multiline Hoythther Lifetimes", Paper AIAA 95–2890, 31st Joint Propulsion Conference, July 1995).

References:

1. Robert L. Forward, Failsafe Multistrand Tethers for Space Propulsion, Forward Unlimited, P.O. Box 2783, Malibu, Calif. 90265, July 1992, Final Report on NASA Contract NAS8-39318 SBIR 91-1 Phase I.
2. Robert L. Forward and Robert P. Hoyt, Failsafe Multistrand Tether SEDS Technology Demonstration, Final Report on NAS8-40545 with NASA/MSFC (Jun. 14, 1995).
3. Robert L. Forward and Robert P. Hoyt, "High Strength-to-Weight Tapered Hoytether for LEO to GEO Payload Transfer" Final Report on contract number NAS8-40690 with NASA/MSFC (Jul. 10, 1996).

The Hoytether is essentially a tri-axial net structure, with 'primary' lines running along the length of the tether and two sets of 'secondary' lines connecting these primaries diagonally. They can be made by hand and connected with knots as is taught by the 1992 Hoytether structure. Because knotted connections severely limit the strength of a structure, it is desirable to use a knotless fabrication technique to achieve interconnections that have strengths approaching the limits of the constituent material. As these tethers may be many kilometers long; fast and inexpensive mechanical methods are required for their practical fabrication.

Hoytethers may be made by mechanical braiding, i.e. three-dimensional braiding, such as 3-D rotation braiding using braiding machines such as those developed by the Herzog Company in Germany (August Herzog Maschinenfabrik GmbH & Co., Postfach 2260.26012, Oldenburg, Germany. The specialized loom developed by the Nichimo Company of Japan (Nichimo Company Ltd., 2-6-2 Ohtemachi, Chiyoda-Ku, Tokyo, Japan) is used to produce "Ultracross" knotless fishing nets in which the individual strands are braided as a 4-braid line, and the strands are interbraided where they cross. This produces netting that has slipless interconnections that are very strong, approaching the maximum capability of the fiber. Such a loom could, with some modifications, produce the present invention's structure. Only two such machines exist, one in Japan, the other in Washington State. Unfortunately neither can work with the small line diameters needed to practice the preferred embodiment of the present invention. See generally, Ko, F. K., "Braiding", in *Engineered Materials Handbook, Vol.* 1., *Composites*. ASM International, Metals Park, Ohio, 1957. Pp. 519–528.

The most common 3-dimensional braiding machines are 4-step braiders based upon the designs of Maistre (German Patent P230-16986, issued 1973) and Forentine (U.S. Pat. No. 4,312,261, issued 1982). Braiding is accomplished by using pneumatics or solenoids to push the parts of the braiding machine to the proper positions. This is a slow process and making a Hoytether kilometers long with these machines would be very time consuming and expensive. The composites division of Albany International (Albany International Research Company, 777 West Street, Mansfield, Mass.) also produces a 3-D braiding machine. This machine uses modular braiding components that arc assembled breadboard fashion on a large wall.

Although braiding is the preferred technique, alternate fabrication methods such as Raschel knitting and crocheting can be used successfully. Multikilometer long Hoyththers are presently being produced for the inventors by the vendors Culzean Fabrics and Flemings Textiles using an electronically controlled crochet machine produced by Comez in Italy.

Space Tether Systems:

The prior art teaches the use of tethers in space applications. U.S. Pat. No. 5,163,641, issued on Apr. 9, 1990 to Yasaka, teaches the use of a powered spacecraft connected by a tether to a satellite. This tether is disconnected to change the state vector of the satellite. The state of the art of energy and momentum transfer using space tethers is discussed in Ivan Beckey's article "Tethering, a new Technique for Payload Deployment", Aerospace America, March 1997, at pages 36–40. Beckey concludes, "Tethers can perform the same functions as propulsive upper stages of direct payload injection, but at lower weight and cost per pound." U.S. Pat. No. 4,923,151, issued Mar. 1, 1988 to Roberts, Wilknson and Webster, teaches a tether power generator for earth orbiting satellites. U.S. Pat. No. 4,580,747, issued Mar. 15, 1983 to Pearson, teaches use of a long tether extending downward into the atmosphere from a satellite. The state vector of the satellite is changed by forces acting on a lifting body connected to the end of the tether. U.S. Pat. No. 4,824,051, issued Jan. 12, 1987 to Engelking, teaches passing an electric current through a conductive tether attached to a satellite to provide propulsive force to alter the orbit of the satellite. U.S. Pat. No. 5,082,211, issued Jan. 21, 1992 to Werka, teaches use of a tether to deorbit space debris. U.S. Pat. No. 4,727,373, issued Mar. 31, 1986 to Hoover, teaches an orbiting stereo imaging radar system having two spacecraft in synchronous parallel orbits connected by a tether.

Tether Dynamics:

In order to understand the forces that cause a tethered satellite to move upward and away from an orbiting satellite, for example, it is first necessary to explain briefly how a satellite remains in orbit. An orbiting satellite is acted on by the force of gravity which pulls it toward Earth, and by a centrifugal force, which pushes it away from Earth. The centrifugal force" (actually inertia) results from the motion of the satellite around its circular orbit. This is the same force that one can experience by swinging a ball around on the end of a string. A satellite is maintained in its orbit when it travels at the natural speed for its altitude and, as a result, the centrifugal force is equal to the gravitational force.

At the typical orbital altitude of 250 kilometers for a low-Earth orbit satellite, for example, a speed of approximately 7.6-km per second is required to create sufficient centrifugal force to balance gravitational attraction on the satellite. If the altitude is changed, the two opposing forces will no longer be in balance unless the satellite also changes its speed. A higher orbital altitude requires a slightly lower speed so the satellite will take longer to complete an orbit. Because of this, if two free-flying satellites are in orbits at different altitudes, the lower satellite will circle the Earth in less time than the satellite in the higher orbit.

If two satellites, at different altitudes, are connected to each other by a tether, they are forced to travel around their orbits together—in the same period of time, which is longer than the natural period of the lower satellite but shorter than that of the upper satellite. The lower satellite will, therefore, slow down below the natural speed for its orbit and will tend to fall to a lower orbit because the centrifugal force will now be less than the gravitational attraction of the Earth. An upward force in the tether that makes up the difference between centrifugal and gravitational forces holds it in place, however.

Correspondingly, the upper satellite will be accelerated above its natural orbiting speed (increasing its centrifugal force above the gravitational attraction) and will tend to move to a higher orbit. It, too, is held in place by an additional force (downward) in the tether. In other words, the net force downward on the lower satellite is balanced, through the tether, by the net force upward on the upper satellite. The effect of unbalanced forces on the two satellites is, therefore, to create tension in the tether. During the TSS-1 & 1R experiments, the inertia of the tethered satellite causes the satellite to rise above the orbiter as the tether is reeled out. Very close to the orbiter, there is little difference in the two orbits, and the tension force is insufficient to overcome friction in the deployer mechanism; therefore, until the satellite reaches a separation of approximately 1000-meters, the tension is augmented by small tether-aligned thrusters on the satellite. Beyond this point, the tension in the tether is the only force required.

By experimenting with a ball hung on a piece of elastic cord (a paddleball, for example) it is possible to simulate all the different types of oscillations that are possible on a space-based tether system. The elastic cord, representing the tether, may compress and stretch, causing the ball to bounce up and down (longitudinal oscillation). It also may move in a circular (skip-rope) motion or may develop wave-like motions (transverse oscillations). Even if the string itself remains straight, it is possible to get the ball swinging back and forth about its attachment point on the paddle like a child on a swing rope (pendulous motion).

Each type of motion occurs with a particular frequency, which depends on the length and tension of the tether. When the frequencies are different, the motions do not interact; however, at some tether lengths, the frequencies of two or more types of oscillation can become very close. At this point, energy can be transferred from one typc of motion to another, a phenomenon known as resonance. For instance, the transverse oscillations in the tether may cause the satellite to swing back and forth in pendulous motion.

Many different factors may cause oscillations; the movements of the satellite or Shuttle are but two of these. For an electrodynamic tether, the skip-rope and pendulous oscillations are of particular interest. If a current is passed through a tether, the current will interact with Earth's magnetic field, resulting in a force that may produce skip-rope and pendulous oscillations. Because it is necessary to maintain control of the satellite, much study has gone into identifying the different types of possible motions and the methods used to control them.

One way to control the magnitude of those motions that cause a change in tension or transverse motion at the end of the tether is to have an end mass connected to the Hoytether that maintains a controlled tension on the tether, working much like a spring-loaded 'dog leash'. This may be as simple as a coiled spring, or as complex as an active control system that measures the tension and transverse forces on the tether and adjusts the applied tension according to a local or remotely operating algorithm.

Electrodynamic Effects of Conductive Tethers:

Electric potential is generated across a conductive tether by its motion through the Earth's magnetic field. Electromagnetic forces acting on a conductive tether in orbit can make the tether system behave like an electric motor or generator, thereby exerting a useful force to alter the state vector of the tether and any mass attached to it.

Electrodynamic tether propulsion is unlike most other types of space propulsion in use or being developed for space application today—there is no hot gas expelled to provide thrust. Instead, the environment of near-Earth space is being utilized to propel a spacecraft or upper stage via electrodynamic interactions.

A charged particle moving in a magnetic field experiences a force that is perpendicular to its direction of motion and the direction of the field. When a current flows through a long, conducting tether the electrons flowing through the tether experience this force due to the fact that they are moving along the wire in the presence of Earth's magnetic field. This force is transferred to the tether and to whatever the tether is attached (like a spacecraft, satellite, space station or upper stage). It can be an orbit-raising thrust force or orbit-lowering drag force, depending upon the direction of current flow. Operation in one mode allows boost from LEO to higher orbit while reversing the current flow provides negative thrust for deboost. The principle is much the same for an electric motor; reverse its operation and it acts as a generator. The current passing through the tether is returned through the ionosphere to complete the current loop with collection and emission occurring on opposite ends of the tether.

The PMG experiment demonstrated that a conducting tether can be used as both a motor and a generator. The TSS experiments, especially TSS-1R showed that very large voltages (about 3500 volts) can be generated by a sufficiently long tether.

Uses of an electrodynamic tether as an orbit raising and lower propulsion system has many advantages over competing systems:

a. It is nearly propellantless. Most other systems expel hot gases and require extensive resupply. To emit current, the electrodynamic tether propulsion system may use plasma contactors developed as a part of the International Space Station Program. These contactors consume less than 20 kg of xenon gas per year with a 50% duty cycle. The electrodynamic tether propulsion system can also use field emitter arrays, which emit electrons without the use of any gas.

b. It can change both altitude and inclination. The Earth's magnetic field is non-uniform and can therefore provide both in- and out-of-plane forces for inclination changes as well as altitude changes. This is of particular interest to payloads requiring polar orbits in that they can be launched on a small launch vehicle into a lower inclination orbit and have it raised in space by the proper phasing of current through the tether.

A demonstration of the propulsive capabilities of electrodynamic tethers was recently approved for a flight test in 1999. The Propulsive SEDS or ProSEDS mission, will fly as a secondary payload on a Delta II launch vehicle and deploy a 5-km conducting tether using the existing SEDS deployer concept. The ProSEDS experiment will be followed by the Electrodynamic Tether Upper Stage (EDTUS) experiment that will demonstrated the use of electrodynamic forces to change both the altitude and inclination of the experimental spacecraft. FIGS. **1**B and **1**C show the calculated electrodynamic thrust at several inclinations and the reentry time sensitivity of the ProSEDS tether, respectively.

One application for long-life conducting electrodynamic tethers is as a "Terminator Tether™" for removing from orbit unwanted Earth orbiting spacecraft at the end of their useful lives. When the mission of the satellite is completed, the Terminator Tether™, weighing a small fraction of the mass of the satellite, would be deployed. The electrodynamic interaction of the conducting tether with the Earth's magnetic field will induce current flow in the tether conductor. The resulting energy loss from the heat generated by the current flowing through the ohmic resistance in the conducting tether will remove energy from the spacecraft, eventually causing it to deorbit, thus reducing the amount of orbital space debris that must be coped with in the future.

In the following analysis, it is shown that the amount of energy loss generated by an electrodynamic tether is essentially independent of its length or area, and instead is primarily proportional to the tether mass and the physical properties of the conductor metal chosen. In the typical example calculated, a 1000-kg spacecraft can be deorbited from a 1000-km high Earth orbit by a 10-kg mass tether in a month, while a 1-kg tether can deorbit a 1000-kg spacecraft in less than a year.

To the knowledge of the inventors, Joseph P. Loftus of NASA/JSC first proposed the general concept of using an electrodynamic tether to deorbit spent satellites. (Joseph P. Loftus <JLoftus@ems.jsc.nasa.gov>, personal communication via email to Robert Forward, Monday Jun. 10, 1996 15:50:10.) In order to show that the Loftus deorbit concept was not obvious to those skilled in the art of electrodynamic tethers, Forward contacted the leading expert, Joseph Carroll, of Chula Vista, Calif., who built and participated in the flight test of the PMG. After being told of the Loftus concept in a telephone conversation, his reply in an Email message dated Aug. 5, 1996, was "such a system would be feasible . . . by it is still not obvious to me that it would be useful . . . "

Loftus was considering the use of electrodynamic drag from a conducting tether to achieve this goal of bringing the unwanted spacecraft down from its high orbit (where atmospheric drag is negligible) to a 200-km orbit, where atmospheric drag would rapidly finish off the task of removing the unwanted spacecraft from orbit. The tether Loftus was considering was a single-line, conducting tether, typically 1-mm in diameter, 1-km long, and, if made of aluminum, 2-kg in mass. He would include means at the ends of the tether to contact the ambient space plasma around the Earth to complete the current loop.

Unfortunately it is probable that space impactors would sever the 1-mm diameter, 1-km long single-line tether proposed by loftus within a 1/e lifetime of four months. This would produce orbital debris rather than removing it. The motivation for this work is the NASA Safety Standard NSS 1740.14 "(Guidelines and Assessment Procedures for Limiting Orbital Debris." The relevant portion of the Standard starts on page 6-3: General Policy Objective-Postmission Disposal of Space Structures. Item 6-1: "Disposal for final mission orbits passing through LEO: A spacecraft or upper stage with perigee altitude below 2000 km in its final orbit will be disposed of by one of three methods." The method of interest is the atmospheric reentry option, Option a: "Leave the structure in an orbit in which, using conservative projections for solar activity, atmospheric drag will limit the lifetime to no longer than 25 years after completion of mission. If drag enhancement devices are to be used to reduce the orbit lifetime, it should be demonstrated that such devices will significantly reduce the area-time product of the system or will not cause the spacecraft or large debris to fragment if a collision occurs while the system is decaying from orbit."

The NASA standard applies only to NASA spacecraft and even then only to completely new spacecraft designs. New versions of existing designs are to make a "best effort" to meet the standard, but will not be required to change their design to do so. The Department of Defense has adopted the NASA standard with the same provisos. An Interagency Group report has recommended that the NASA standard be taken as a starting point for a national standard. It is NASA's recommendation to the Interagency Group that the safety requirement be phased in only as spacefaring nations reach consensus internationally, which is being done through the International Debris Coordination Working Group whose members are Russia, China, Japan, ESA, UK, India, France, Italy, and the US.

Thus, although the NASA Safety Standard in its present form is not the "Law", the existence of the standard means that some time in the future a similar requirement may be imposed on all spacecraft. This could result in major growth in future space tether business, with a sale to every non-geostationary spacecraft being "mandated" by government safety regulations, somewhat as the sale of seat belts and airbags for every car are mandated.

In fact, three of the companies planning to set up "constellations" of low to medium orbit communications: Teledesic, Iridium and Odessey have committed their companies to abide by the spirit of NASA Safety Standard 1740.14 by using one means or another to deorbit their spacecraft before they reach end of life.

Problems with Prior Art Tethers:

All electrodynamic tether designs proposed by the prior art teach that the tether should be operated at a right angle to the magnetic field through which the tether is moving. This is a problem because the electrodynamic force acting on the tether causes the tether to align itself with the magnetic field force lines. To overcome this problem the prior art teaches the use of a large ballast mass attached to the end of the tether and/or use of a very long (tens to hundreds of kilometers) tether. The large ballast mass is expensive to take to orbit because it replaces useful payload. The long tether sweeps a larger Area-Time-Product during its useful life and thus is more likely to impact other space objects, either debris or another spacecraft.

Another problem common to all proposed prior art tethers is tether instability. If the tether produces a large electrodynamic drag force, which is desirable because a large drag force will cause the satellite to deorbit quickly, then the tether will be dynamically unstable. This instability can cause the tether to lose its effectiveness, act uncontrollably and even wrap around the satellite or otherwise malfunction. Experts skilled in the art of tether design have opined that this dynamic instability is inherently unavoidable in any electrodynamic tether system. The prior art solution, such as that presently being used in the ProSEDS experiment, has been to use a large ballast mass to increase the stabilizing gravity-gradient force and/or to limit the electrodynamic drag of the tether to less that the maximum that could be produced. In the ProSEDS experiment, the conducting electrodynamic tether is five kilometers long. To insure stability, it will be augmented by a 20–35 kilometer long non-conducting tether, which to further have stability will have a 40 kilogram ballast end mass.

Yet another problem of all proposed prior art electrodynamic tether systems is how to radiate away the energy produced by the tether's operation. A satellite moving at an orbital velocity of 18,000 miles per hour has a kinetic energy of over 10,000 calories per gram. To put this amount of energy in an understandable perspective, it may be noted that when nitroglycerine explodes it produces about 1,500 calories per gram. Prior art designs of electrodynamic drag tethers teach the use of the electrical energy generated by the tether to charge batteries or operate electronics, with the excess energy being converted into heat by a resistive load. This excess heat must be radiated to the space environment or it will melt the resistive load. Thus the resistive load, and/or its associated radiator structures, must be massive and replace useful payload.

DISCLOSURE OF THE INVENTION

The present invention comprises an electrodynamic tether structure and a method of use. The principal industrial utility of the present invention is to deorbit satellites in Earth orbit at the end of their useful life. This embodiment of the present invention is sometimes referred to in this specification as a "Terminator Tether™" because it terminates the orbital lifetime of the host spacecraft. The structure of the tether taught by the present invention is a short, wide, interconnected-multiwire (compared to the long single wires of the prior art) conductive Hoytether whose area maximizes electrodynamic drag while simultaneously minimizing the Area-Time-Product swept by the tether during its operating life. The preferred tether length is two to five kilometers. The preferred tether mass is one to five percent (1%–5%) of the spacecraft mass. The method of operation comprises orienting the tether structure at a 35.26-degree trailing angle to the local vertical to maximize electrodynamic force on the tether while avoiding tether instability and allowing use of a small tether end mass.

The present invention also teaches that the satellite-tether system may be rotated around its common center of mass to centrifugally produce tension force in the tether structure to oppose forces causing tether instability. The angle of the conductive tether structure of the present invention with respect to the velocity vector of the host spacecraft may be controlled by the method of the present invention so it interacts with the encountered magnetic field to induce a maximum current flow in the tether. This produces maximum electrodynamic drag. All or a portion of this electric power may be stored and then controllably applied to the conductive tether to produce an induced electrodynamic force. This induced electrodynamic force may by used to enhance the drag force, to rotate the tether-satellite system and/or to provide satellite propulsion, i.e. to change the state vector of the satellite for any useful purpose, e.g. to avoid collision or to change the host spacecraft's orbit to an orbit more favorable for more rapid deorbiting.

The present invention also teaches a tether structure that also functions as a thermal radiator and/or plasma contactor. An embodiment of the present invention using conducting elements of the satellite, e.g. the solar arrays, as electrodynamic tether structures is also disclosed.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention may be better understood with reference to the following drawings:

FIG. **1***a* is a graph comparing the survival probability of a single line tether as taught by the prior art to the Hoytether used by the present invention.

FIG. **1***b* is a graph showing the level of electrodynamic thrust for a 10 kilometer, 10 kilowatt tether at various altitudes and inclinations.

FIG. **1***c* is a graph showing reentry time sensitivity of a five kilometer ProSEDS tether for several conditions of current flow in the electrodynamic tether.

FIGS. **2***a*, **2***b* and **2***c* show the Hoytether structure used by the preferred embodiment of the present invention.

FIG. **2***d* shows the Hoytape structure used by the preferred embodiment of the present invention.

FIG. **3** is a block electrical diagram of an electrodynamic tether system.

FIG. **4** is a graph showing a graph of the Area-Time product for three separate levels of exospheric temperature (for neutral drag calculations) and for two separate assumptions regarding tether mass and power drawn from the ambient plasma (for the Terminator Tether™ calculations).

FIG. **5** is the percent additional mass required in orbit to drop the perigee of a circular orbit at altitude, at, to a value of 200 kilometers using standard chemical rocket as taught by the prior art for deorbiting unwanted spacecraft.

FIG. **6** is a deorbit tether system shown operating with its electrodynamic tether at a trailing 35.26-degree angle to the encountered magnetic field, as is taught by the preferred embodiment of the present invention.

FIG. **7** shows an embodiment of the present invention in which the electrodynamic tether—satellite system is rotated about its center of mass to avoid tether instability.

FIG. **8** shows an embodiment of the present invention wherein the tether is powered to provide induced electrodynamic force.

FIG. **9** shows an embodiment of the present invention wherein the solar power system structure of the satellite is used as an electrodynamic tether.

FIG. **10** shows the use of a portion of the tether structure as a thermal radiator and plasma contactor.

FIG. **11** shows the braiding of an ohmic resistive load into the tether structure.

FIG. **12** is a force diagram showing the forces and torques on an upward deployed conducting tether due to the motion of the host spacecraft through the magnetic field of the Earth.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The principal industrial utility of the present invention is the deorbiting of satellites in Earth orbit, which necessarily can happen only in outer space. Until the present invention is reduced to actual practice by use with satellites in outer space and a body of practical experience is thereby obtained, the present invention can only be constructively reduced to practice, such as by this written specification and its associated drawings, diagrams and graphs and by reference to electric measurements made on conducting tethers in space by the PMG and TSS experiments. The inventors provide herein a detailed discussion of the theory of the present invention to help those skilled in the art of aerospace and tether engineering to understand the present invention and to make and use the best embodiment of the present invention known to the inventors at the time this specification was prepared, without undue experimentation.

Although the present invention is discussed in this specification in its preferred embodiment as a means of deorbiting satellites, it must be emphasized that discussion of this specific use in no way limits the broad scope of the present invention. The present invention can be used to change the direction and speed, i.e. the state vector, of any space object providing that object is in motion relative to any magnetic field. This magnetic field may be that of the Earth or of any other celestial body, for example Jupiter or the sun.

In this specification the material forming the structure of the electrodynamic tether is specified as being a conductor. For example, a metal, such as copper or aluminum wire could be used. Likewise, nonmetallic conductor, such as carbon nanotubes, or a conductive polymer could form the conductive structure of the tether.

FIG. **1***a* shows the survival curve for a Hoytether as used by the present invention vs. the survival curve of a prior art single line tether. The survival probability curve of the Hoytether as a function of time does not conform to the well known "1/e decay" shape of a single-line tether. The Hoytether can maintain a high level of survival probability, about 99 percent, until it nears its 'lifetime'. Its survival probability then drops rapidly to zero. A detailed mathematical analysis of the difference between Hoytether and single tether survival probabilities is given in Appendix E "Small Impactor Survival Probabilities of Hoytethers" and Appendix F "Large Orbital Debris Survival Probabilities of Hoytethers", both contained in the Final Report of NASA Contract NAS8-40545. The resulting cut probability with time for the Hoytether has a "bingo curve" shapes. In a bingo game, at least five numbers must be called before anyone can win, and usually many numbers have to be called before one of the bingo cards gets five in a row. In the Hoytether at least four cuts must happen at the same level before any failure occurs, and many cuts have to be made before any one of the levels has all four lines cut. The bingo curve has the property that the probability of survival stays very high for periods short compared to the lifetime. The probability of survival is greater than 99% for periods shorter than 10% of the lifetime. This is much better performance than the 1/e curve of a single line tether, where the probability of survival is only 90% at 10% of the 1/e lifetime.

FIG. **1***b* shows the calculated level of electrodynamic thrust, in Newtons, produced by a ten kilowatt, ten kilometer single conductor tether at altitudes between 200 and 1400 kilometers and at orbital inclinations between zero degrees and eighty degrees.

FIG. **1***c* is a graph showing the reentry time sensitivity calculated for a five kilometer ProSEDS single conductor tether at altitudes between 200 and 500 kilometers with current flow in the tether between on flow and one ampere.

FIGS. **1***b* and **1***c* are the work of Enrico Lorenzini of the Smithsonian Astrophysical Observatory from his study "Performance Evaluation of the Electrodynamic Tether Tug", NASA/MSFC Grant NAG8-1303 (Nov. 12, 1966). They show that experts in the field believe that the present invention will have industrial utility, as the current flow of one ampere in FIG. **1***c* decreases the deorbit time for the calculated case from more than 180 days to less than 20 days.

FIG. **2***a* is an isometric drawing showing the generally cylindrical structure of the Hoytether. In FIG. **2***a*, primary lines **201** and **203** are shown connected via secondary lines **211**. In FIG. **2***b*, which is a schematic illustration of the structure shown in FIG. **2***a*, primary lines **201, 203, 205, 207** and **209** are connected, each to their respective adjacent primary lines, by a plurality of secondary lines **221**. These connections, for example as shown at interconnection **213**, are made by knotless, slipless interconnections, such as Soutache braiding of twisted connections.

FIG. **2***c* shows the Hoytether structure of FIG. **2***b* but with a severed primary line **205** at breakpoint **215**. The second level of secondary lines **217**, shown as bold lines in FIG. **2***c*, redistribute the load from the severed primary line **215** at point **217** back to line **215** at points **219** and **221**, above and below the break, respectively.

For the electrodynamic tether, a preferred embodiment would be to unroll the tubular Hoytetherr™ into a flat Hoytape™ as shown in FIG. **2***d*.

FIG. **3** generally shows the arrangement of functional elements of an electrodynamic tether system. Of course such a system must have a means of attachment to the spacecraft, a deployer, the tether itself, some form of space plasma cathode contactor, some form of space plasma anode contactor, a power dissipation system, and a communication and control system. Each of these individual elements are known in some form to those skilled in the art of aerospace engineering. Thus the present inventors do not feel it is necessary to describe them in detail in this specification.

In FIG. **3**, a space object **301**, i.e. a satellite in Earth orbit, or any other space object either natural or man made, is physically connected to the tether system. The tether system comprises a deployer **303** from which a conductive Hoytether **305** having a bare segment **307** extends upward from space object **301**. The positively biased anode end **309** of tether **305** collects electrons from the ionosphere as space object **301** moves in direction **310** across the Earth's magnetic field. These electrons flow through the conductive structure of the Hoytether to the power system interface **311**, where it supplies power to an associated load, not shown. The electrons then flow to the negatively biased cathode **313** where electrons are ejected into the space plasma **315**, thus completing the electric circuit.

In order to allow those skilled in the art to better appreciate the broad scope of the present invention, the inventors will now provide an example using specific tether and system values to compare the Terminator Tether™ to prior art means, ie a rocket, for deorbiting a satellite.

Terminator Tethers™:

When an Earth orbiting conducting space tether moves through the magnetic field of the Earth, an electric field is generated in the tether that is proportional to the velocity of the conductor, the magnetic field strength of the Earth, and the angle between the conducting tether and the magnetic field lines. From data obtained during the various electrodynamic experiments that have been conducted in space to date, such as the PMG, TSS-1, and TSS-1R experiments, a typical value of the generated emf per unit length of tether of E=100–200 Volts/kilometer can be assumed. The electric potential V developed at the ends of a tether of length L is then V=EL. For a tether of length L=10 km, the electric potential developed is V=1000 V. For calibration, the 20-km long TSS-1R tether, at the moment of failure, was developing a potential of 3500 volts, which is 175 volts/km.

The mass of a conducting tether of length L, cross-sectional area A, and densin, d is given by m=dLA. Typical values for the density are d=2700 kg per cubic meter for aluminum and 8900-kg per cubic meter for copper. For a typical aluminum tether of mass m=10 kg and length L=10 km, the cross-sectional area will be A=0.37 mm squared. If this were a solid-wire single-line tether, the diameter of the wire would be D=0.69 mm (21.5 gauge). If this were an 18 line tubular Hoytether, the diameter of the lines in the Hoytether would be D=0.16 mm (34 gauge).

The resistance of a conducting tether of length L and cross-sectional area A is given by R=rL/A, where r is the resistivity of the conductor in nano-ohm-meters (nΩ-m). Typical values for the resistivity are r=27.4nΩ-m for aluminum and 17.0 -nΩ-m for copper. For the m=10 kg aluminum tether of length L=10 km and cross-sectional area A=0.37 mm squared at 20 degrees C, the end-to-end resistance is 750 Ω. This value of resistance is the essentially the same whether the conductor area is concentrated into a single-line tether or divided up into many lines as in a multiline Hoytether.

The current I generated in the conducting tether by the electric potential V between the ends of the tether applied across the tether resistance R is I=V/R. For the 10 kg mass aluminum tether of length L=10 km, electric potential V=EL=1000 volts, and resistance R=750 Ω the current is I=V/R=1.33 Amps. Currents near these values were measured in the TSS-1R experiment at the time of failure. At the time of failure of the TSS-1R tether, none of the plasma contactors on the Space Shuttle or the Italian Satellite were operating. Thus the current was being collected by the conductive surface area of the Space Shuttle and the Italian

Satellite at the two ends of the tether. The Space Shuttle area is quite large, so it was not the limiting factor in current collection. The diameter of the Italian Satellite was 1.6 m, which would give it an effective plasma contact area of about 8 square meters. This shows that, if the plasma contact area of the ends of a conductive space tether can be make large enough, then ampere level currents can be extracted from the ambient space plasma without the use of space plasma contactors.

The power dissipated as ohmic heating in the tether is given by P=IE. For an aluminum tether of mass m=10 kg, resistivity r=27 nΩ-m, and density d=2700 kg/cubic meter, subjected to an electric field of E=100 V/m, the power dissipated in the ohmic losses of the conductor is P=IV= 1330 Watts. For a 1 kg mass tether, it would be still a considerable 133 Watts of dissipation.

There will no doubt be additional dissipation of energy in plasma ohmic losses, plasma wave generation, and plasma ion acceleration, but the ohmic losses in the conducting tether alone are sufficient for the task of deorbiting an unwanted spacecraft massing 100 to 1000 times more than the tether.

The decay time of a metric ton spacecraft moving from a 1000 km altitude orbit to a 200 km altitude orbit with an energy difference of dU=3.3GJ, when its energy is being dissipated at a power of P=1330 W by an aluminum tether massing just 10 kg, or 1% the mass of the spacecraft, is about one month. This is a remarkably short time, and indicates that the concept of using a conductive tether to deorbit a spacecraft is indeed feasible. If the aluminum tether massed only 1 kg, or 1/1000th the mass of the spacecraft it was deorbiting, then the decay time would rise to 10 months, still a reasonable value.

In reality, of course, the actual decay time will be longer than this. If the electrodynamic drag force is very large, and becomes larger than the gravity gradient forces pulling on the ends of the tether (which force is proportional to the mass of the tether), then the tether will tend to align itself along the magnetic field lines instead of across them, and the drag force will decrease because of the small angle between the conductor length and the magnetic field lines. The tether will then settle into an angle determined by the balance between these two forces.

Is the Terminator Tether™ theoretically a better means than atmospheric drag or a rocket engine for deorbiting satellites at the end of their useful life?

In the formula for da/dt (change in altitude per change in time), using the assumptions of near-circular spiral trajectories, the Area-Time-Product (Z), the criteria by which NASA judges compliance with Safety Standard 1740.14, is given by:

$$Z = A\int dt = -\frac{m}{C_D}\int \frac{da}{\rho(a)\sqrt{\mu a}},$$

where ρ(a) is simply the density as a function of semi-major axis. Thus, for a static atmosphere, the problem of area-time product is reduced to quadrature. Note that the area-time product depends only on the density profile and the ratio of spacecraft mass to drag-coefficient and is linearly related to that ratio.

The basic altitude-dependence of the atmospheric density, represented here by ρ(a) can be expressed in a low-order power series of ln(density)=f(ln(altitude)), so that it will be possible to develop a representation of the log—log relationship between altitude and density. The three static levels of atmospheric density are modeled as 5th order polynomials in the natural log of the altitude and are representative of well known atmosphere models, for example, Jacchia, L. G., "Thermospheric Temperature, Density, and Composition: New Model," SAO Special Report 375, March 1977. 2. Anon., Marshall Space Flight Center, "Long-Range Statistical Solar Activity Estimation," Atmospheric Sciences Division, 1989.) The values of exospheric temperature are reasonable (about 2-sigma) high and low solar activity values of 1400 and 800 kelvins respectively. The mean value is taken to be 1100 kevins.

FIG. **4** shows a graph of the Area-Time product for three separate levels of exospheric temperature (for neutral drag calculations) and for two separate assumptions regarding tether mass and power drawn from the ambient plasma (for the Terminator Tether™ calculations). The density profiles for the three values of exospheric temperature represent the extremes and mean values to be expected during the next several decades. The exospheric temperature goes through a cycle of about an 11 year period with maximum about 1400 kelvins and minimum about 800 kelvins.

The tether calculations include an inherent assumption that the Terminator Tether™ transfers all the energy extracted from the ambient space plasma into drag that decreases the orbital energy of the spacecraft at a rate given by the power drawn by the tether. Thus, the change in energy for a spacecraft of mass m at an initial altitude h is given by the expression:

$$\Delta E = -m\left\{\frac{\mu}{2(R_e+h)} - \frac{\mu}{2(R_e+250)}\right\},$$

and the time required to effect this change in energy of the mass, m, is just

$$\Delta t = \Delta E/P,$$

where P is the power drawn by the tether from the ambient plasma charge.

The curves for 1% and 5% tethers do not include the effects of atmospheric drag because the electrodynamic drag is orders of magnitude greater than the neutral atmospheric drag. At altitudes greater than about 700 km, the electrodynamic drag is 200 to 3000 times greater than the neutral drag forces.

It is clear, from FIG. **4** that the Terminator Tether™ concept is far superior to neutral drag in removing spacecraft from orbit, no matter how much additional area is added to the passive spacecraft to increase the atmospheric drag. The analyst should note that FIG. **4** is proportional to mass; that is, the Area-Time-Product values should be multiplied by the mass of the spacecraft and divided by 1000 kg. This is true for the neutral drag and Terminator Tether™ curves alike. The power levels assumed for the 1% and 5% tethers are only 80% of their theoretical values. This is done to provide a 20% margin on the power available from the ambient plasma and electrodynamic gradient.

Conventional rocket mechanisms can remove spacecraft from orbit, but this mechanism is apparently not viable from mass considerations when these factors are compared with the capabilities of the Terminator Tether™. A satellite owner may decide to satisfy the NASA safety requirements by adding a small solid motor and the associated hardware, software, sensors and structure to make the package independent, to be used to deorbit the spacecraft in case the main spacecraft power, attitude, or propulsion system fails.

The requirements of such a rocket motor system are more stringent than those attributed to ordinary spacecraft. The rocket motor deorbit system must operate when some or all other systems of the spacecraft have failed. These more stringent requirements are balanced by lesser requirements of performance. The backup system must simply deorbit the spacecraft; it does not have to perform all the other duties of the spacecraft. But the backup system must know when to fire under all kinds of anomalous situations, including tumbling, offset of center of mass (because of loss of parts due to collisions), and lack of knowledge of the orbital position.

FIG. **5** shows the percent additional mass required by the rocket motor system to drop the perigee of a circular orbit at some initial perigee altitude to a value of 200 km. The atmospheric drag at this 200 km perigee altitude will remove any spacecraft (in the range considered) from orbit in a few revolutions. The contours of constant stage propellant mass fraction, λ, range from low values of 0.5 through reasonable values of 0.65 to 0.7, up to the values associated with the best solid motors ($\lambda\nabla 0.93$) without adding any extra hardware to the emergency stage. An effective, independent stage to provide a retro deltaV of from 50 to 325 m/s will almost certainly have a λ of the order of 0.6 to 0.75. If the emergency stage is required to perform its own attitude determination, the stage propellant mass fraction may be as low as 0.55 or 0.50. Note that the additional mass, shown in FIG. **5**, must also be lofted to orbit in the first place, to provide the mass on orbit for the originally intended service. The stage propellant mass fraction, λ, is a key stage performance parameter that describes the ratio of the mass of propellant to the mass of the stage. That is $\lambda=m_p/(m_p+m_l)$, where m_p is the mass of the propellant and m_l is the mass of everything else in the stage (not including the payload or any stages above or below the stage being considered). Typical independent small rocket stages have values of λ from 0.60 to 0.75. The curves of stage propellant mass fraction are shown as straight lines (a minor approximation) and are included only from 0.5 to 0.95, the extremes of reasonable design practice.

These calculations that show, beyond any reasonable doubt, that the Terminator Tether™ concept is far superior to conventional mechanisms such as drag enhancement devices or small rocket deorbit propulsion systems. The superiority is measured in terms of Area-Time-Product, NASA's measure of the likelihood of collision with other spacecraft in the path of the descending spent member of a constellation. Tether calculations were made using conservative assumptions that the power extractable from the ambient plasma and electrodynamic gradient is only 80% of the theoretical power available to a perfect tether crossing the magnetic field lines at a right angle, i.e. normally.

The following analysis is presented by the inventors to help those skilled in the art to better understand the present invention when it is used as a Terminator Tether™ for removing from orbit unwanted non-geostationary Earth-orbiting spacecraft at the end their useful lives. The primary result of that analysis is that the electrical power P in the tether that is converted into heat by the resistance of the tether and radiated away into space will remove energy from the spacecraft, causing it to rapidly deorbit, thus reducing the amount of orbital space debris that must be coped with in outer space. Tether electric power P is given by:

$$P=(vB)^2m/2rd$$

where m is the mass of the conducting tether, r and d are the resistivity and density of the conducting material, and v is the velocity of the spacecraft's motion through the Earth's magnetic field B. For a m=10 kg tether of aluminum with resistivity of r=27.4 nΩ-m and density d=2700 kg/m^3, moving at a velocity v=7037 m/s relative to the Earth's horizontal magnetic field B=26.5 μT, the power dissipated is P=2350 W! This energy loss in the form of heat must necessarily come out of the kinetic energy of the host spacecraft. For a typical example, a 1000 kg spacecraft in a 1000 km high orbit subjected to an energy loss of 2350 J/s from a 10 kg tether (1% the mass of the host spacecraft) will be deorbited in a few weeks.

Power levels of the magnitude estimated in the previous paragraph have been measured in a real orbital space experiment, the TSS-1R mission carried out on the Shuttle Orbiter in 1995. In that experiment, a large Italian spacecraft, 1.6 m in diameter, was deployed upward from the Shuttle Orbiter at the end of a conducting copper wire tether covered with electrical insulation. As the tether was slowly deployed upwards, a series of measurements were made of the open circuit voltage induced in the tether by its motion through the Earth's magnetic field. The voltage between the end of the tether and the Orbiter ground varied from zero volts at the start to 3500 V when the amount of tether deployed approached its maximum length of 20 km. Periodically, the end of the tether was connected either to one of two different electron guns, which supplied contact to the surrounding space plasma, or to the Orbiter ground. The bare surfaces of the Shuttle Orbiter proved to be a surprisingly good plasma contactor via a combination of ton collection and secondary electron emission. The current flow through the tether was deliberately limited by control circuits and the current capacity of the electron guns to about 0.5 amperes, but power levels of 1800 Watts were reached.

The tether was intended to have a fully deployed length of 20 km, but at a deployed length of 19.5 km, when about 3500 V was being induced at the end of the tether inside the Orbiter reel mechanism, a flaw in the insulation allowed an electrical spark to jump in an uncontrolled manner from the tether to the Orbiter ground. With no control circuits to keep the current level down to 0.5 amperes, the current flow jumped to 1.1 amperes, and the total power generated was P=3850 Watts Most of this energy went into the electrical arc, which burned through the tether, causing it to break and halting the experiment. This experiment showed that large areas of bare conducting material, such as that provided by the 8 square meter area of the Italian spacecraft at one end of the tether and the very large surface area of the Shuttle Orbiter spacecraft at the other end of the tether, can collect and emit amperes of current, while thousands of volts of potential can be generated by sufficiently long tethers moving at orbital speeds.

Thus, both theory and experimental data collected in the space environment indicate that significant amounts of electrodynamic drag force can be obtained from a low mass conducting tether attached to a host spacecraft, provided the ends of the conductor can exchange sufficient numbers of electrons with the surrounding space plasma.

Experimental data from the TSS-1R data also produced the amazing result that the efficiency of a bare metal surface in "contacting" the space plasma is many times better than the standard theory would predict. The 8 square meters of bare surface area of the Italian spacecraft were sufficient to collect the 1.1 A of electron current. This amount of area is easily replicated by a few hundred meters of bare wire, considering that the effective collection diameter around the wire is the Debye length, which is a few centimeters at the typical values for space plasma density and temperature.

Because of this result, that a bare wire can easily collect electrons, Les Johnson, Nobie Stone, Chris Rupp, and others at NASA Marshall Space Flight Center have formed a team, which includes the present inventors, which is embarked on a new flight experiment. The experiment is scheduled for a piggy-back flight on a Delta II launch of an AF Global Positioning Satellite in early 2000. The goal of the experiment is to demonstrate that electrodynamic drag from a wire moving at orbital speeds through the Earth's magnetic field will create a large enough electrodynamic drag force to deorbit the Delta II second stage, whose mass is greater than 1000 kilograms, in a few weeks. This is essentially a demonstration of the Loftus electrodynamic drag deorbit concept and the first step in the development of a Terminator Tether™.

The ProSEDS (Propulsion Small Expendable-tether Deployer System) mission will use a 5 km long copper wire conductive tether massing 18 kg connected to a 20–35 km long nonconducting tether, which is in turn connected to a 25–40 kg ballast mass. The total of 25–40 km of tether length and the 25–40 kg ballast mass on the end will provide enough gravity gradient force to keep the tether aligned near the zenith, so that the direction of the current in the tether is at right angles to both the direction of the spacecraft motion in the nominal EW direction and the Earth's near-equatorial magnetic field in the nominal NS direction.

An important feature of the ProSEDS experiment is that it is designed to be completely self-powered. It uses a battery to initiate deployment and to power up the plasma contactor, but once current is flowing through the tether, some of the power is tapped off and used to recharge the battery. The battery, in turn, powers the current control electronics, the telemetry, system, and the plasma contactor. The ProSEDS mission will not be designed to allow ground control changes in operation, primarily because of the increase in complexity and cost associated with that option.

The present invention is the use of a small, low-mass deployer/controller package containing a large collecting area, short length, multiline space tether, such as a Hoytape mesh made of aluminum wire, as a "Terminator Tether™" for a constellation spacecraft. The Terminator Tether™ would be deployed when the host spacecraft is no longer working or no longer wanted. The electrodynamic drag from the Terminator Tether™ would rapidly remove the unwanted spacecraft from the constellation and a few weeks later complete the deorbit of the host spacecraft from space by burnup in the upper atmosphere of the Earth. For a Terminator Tether™ to be of maximum usefulness for constellation spacecraft, it would be desirable to minmize the mass and the length of the tether. A lower added mass means more mass for revenue producing transponders, while a shorter tether length means a lower collision cross-section Area-Time Product during deorbit.

FIG. **6** shows a spacecraft **600** having a state vector **602** that causes the spacecraft to move across magnetic field lines **601**. Spacecraft **600** has a conductive tether **603** attached to it by tether deployer and control system **605**. The outer end of tether **603** is attached to tether end mass **607**. Spacecraft **600** also has an attached electron emitter means **613**.

In FIG. **6**, spacecraft **600** has a velocity and direction defined by state vector **602**. As spacecraft **600** moves it causes conductive tether **603** to cut magnetic field lines **601**. As this happens the distal end of the conductive tether collects electrons, shown as arrows **611**, from the space plasma. Electrons **611** move through the conductive tether **603**, are passed through a resistive load, not shown, in control system **605**, and are emitted back into the space plasma as electrons **615** from electron emitter **613**. This produces an electrodynamic drag in the direction shown by arrow **609**.

As is discussed in detail in the electrodynamic drag analysis below, especially in that sections of the analysis titled "Optimization of Tether Angle", the maximum electrodynamic drag of the tether is achieved when the angle, shown as the Greek letter τ in FIG. **6**, is 35.26 degrees, trailing, to the direction of the magnetic field lines. This angle may be maintained by a feedback control circuit that maximizes the electrodynamic drag of the tether system, as is discussed in detail below. The tether structure **603** is a Hoytether having a length of about 2 to 5 kilometers. The use of the 35.26 degree angle to the magnetic field reduces the tether instability, allowing the use of a short tether which will have a small Area-Time-Product. This will minmize the possibility that the tether will impact another space object during its use.

In FIG. **7**, spacecraft **700** has a state vector **702** across magnetic field lines **704**. Conductive tether **706** is attached to spacecraft **700** by tether deployer and control system **708**. The distal end of tether **706** is attached to a tether end mass **710**. Spacecraft **700** has an electron emitter means **714**.

Functionally, electrons **712** flow from the space plasma into conductive tether **706**, are passed through a resistive load in control unit **708** and are emitted into the space plasma by electron emitter **714** as free electrons **716**. As is discussed in detail in the analysis below, this produces an electrodynamic drag on satellite **700**, which causes it to deorbit.

In FIG. **7**, spacecraft **700** is rotating with an angular velocity indicated by arrow **718**. This rotation causes a centrifugal force **720** to place tension on conductive tether **706**. This force **720** places tension on the tether **706** in a direction that counters tether instability. The Tether **706** may be at any angle to the magnetic field while the rotation of the tether-satellite system is producing this useful tension. FIG. **7** shows the angle to be 35.26 degrees, which is optimum, but any angle will work.

FIGS. **8***a* and **8***b* show a powered Terminator Tether™. In FIG. **8***a* the spacecraft **800** has a state vector **802** that causes it and its associated conductive tether **806** to move across magnetic field **804**. Tether **806** has an end mass **808** that is proximate a plasma contactor **810**. Just as was discussed in the embodiment of the present invention shown in FIG. **6** and FIG. **7**, an electric current moves through the circuit formed by the space plasma, the plasma contactor **810**, the conductive tether **806**, and the electron emitter **815**, thence back into the space plasma as electrons **816**. This creates an electrodynamic drag on the tether in the direction shown by arrow **818**.

In FIG. **8***b*, similar numbers indicate similar structures. In FIG. **8***b*, however, an electric power source, not shown, in spacecraft **800** provides electrons to conductive tether **806**. The electrons are emitted from the end **812** of the tether **806**. Electrons are collected to from the space plasma **814** by contactor **815**, thus completing the electric circuit. The result is an electrodynamic force on tether **806**, and therefore on spacecraft **800**, in the direction show by arrow **820**.

Spacecraft **800** in FIG. **8***b* may be rotated as is taught in the embodiment of the invention shown in FIG. **7**, above, to allow the control system to time the application of electric current to the tether **806** to cause force **820** to be exert force in any desired direction along the circle made by the rotating tether. This allows the state vector of the spacecraft to be modified to raise or lower the spacecraft's orbital altitude, or to otherwise usefully change its orbital elements.

FIG. **9** show an embodiment of the present invention wherein the conductive structure of a satellite is used as the tether. In FIG. **9**, a spacecraft **900** which has a state vector

901 and is moving across magnetic field **903**, has two attached solar panels, **902** and **904**. Solar panels **902** and **904** have conductive aluminum frames **908** and **906**, respectively. Frame **906** is electrically and mechanically connected to plasma contactor **912**. Frame **908** is electrically and mechanically connected to electron emitter **910**. As spacecraft **900** moves across magnetic field **903**, electrons from the space plasma **914** flow through contactor **912** and frame **906** to the body of spacecraft **900**, where the current flows through a resistive load, not shown, and then into frame **908** and electron emitter **910** to the space plasma **916**, thus completing the circuit. This creates an electrodynamic drag on spacecraft **900**. If the power output from solar panels **902** and **904** was directed to flow into the frames, **906** and **908**, then the electrodynamic interaction with the space plasma would cause a useful force to be impressed on spacecraft **900**, as was discussed in connection with FIG. **8** above. The spacecraft may be rotated and/or oriented to a desired angle with the magnetic field lines, as was also discussed above.

FIG. **10** shows an embodiment of the present invention wherein the plasma contactor and the resistive load are structurally part of the conductive tether. In FIG. **10** spacecraft **1000** has a state vector **1001** that causes it to pass through magnetic field **1003**. Spacecraft **1000** has a tether control and deployment system **1005** that is connected to a conductive tether **1007** having an end mass **1013**. A contactor portion **1015** of the conductive tether **1007** near end mass **1013** is adapted to contact the space plasma so as to receive or emit electrons. This may be the bare wire strands of a wide Hoytether, or it may be any other plasma contactor means that can be conveniently made electrically part of the Hoytether structure, such as points or metal fuzz. A second resistive portion **1009** of tether structure **1007** is adapted to be a resistive load. This may be done by making a portion of the tether structure from a material, such as nichrome steel wire, that has suitable electrical resistance The large surface area of the Hoytether provides an excellent radiator structure into the three degree Kelvin radiation sink of outer space.

FIG. **11** shows an alternative embodiment of the load resistance/radiator structure of the present invention. In FIG. **11**, a biline Hoytether **1002** is attached to its control and deployer system **1100** on a spacecraft, not shown, and to a tether end mass **1004**. A resistive load, **1006** is woven into the Hoytether structure, whereby this load **1006** and the section of the Hoytether **1002** into which it is woven act together as a thermal radiator allowing photons **1008** to radiate into the thermal sink of outer space.

The following detailed analysis of the present invention will help those skilled in the art to better understand and use the present invention.

Electrodynamic Drag Analysis:

The following is an optimization analysis of the electrodynamic drag produced by a conducting tether deployed from a host spacecraft that it is to deorbit. The force and torque diagram used in the analysis is shown in FIG. **12**, which shows the forces and torques on an upward deployed conducting tether due to the motion of the host spacecraft through the magnetic field of the Earth.

Physical Constants and Assumptions:

The analysis that follows will use the following physical constants and assumptions:

Newton's gravitational constant $G=6.67\times10^{-11}$ $m^3/kg\text{-}s^2$.

Mass of Earth $M_e=5.976\times10^{24}$ kg.

Radius of Earth near equator $R_e=6378$ km.

Assumed host spacecraft altitude h=622 km.

Assumed host spacecraft orbital radius a=7000 km.

Host spacecraft orbital velocity $w=\omega a=[GM_e/a\ ]^{1/2}=7546$ m/s.

Vertical gravity gradient at spacecraft $2\Gamma=2GM_e/a^3=2.32\times10^{-6}$ s^{-2}.

Centrifugal gradient at spacecraft $\omega^2=GM_e/a^3=1.16\times10^{-6}$ s^{-2}.

Combined gradient at spacecraft $3\Gamma=3GM_e/a^3=3\omega^2=3.49\times10^{-6}$ s^{-2}.

Magnetic field of Earth (tilted dipole approximation):

Magnitude $B_o=35$ μT=0.35 gauss.

Angle between magnetic pole and spin pole 11.5°.

Field components at spacecraft with altitude a and angle β between radius vector of spacecraft and magnetic equatorial plane.

Horizontal $B_H=B_o(R_e/a)^3\cos\beta=26.5$ μT $\cos\beta$.

Vertical $B_V=2B_o(R_e/a)^3\sin\beta=52.9$ μT $\sin\beta$.

Electrodynamic Tether Constants and Assumptions:

The electrodynamic tether is assumed to be made of a conducting metal, and have a length L, density d, resistivity r, and cross-sectional area A that is constant along the length of the tether. If the tether is a single round wire of diameter D, then the cross-sectional area is $A=\pi D^2/4$. Because of the micrometeorite and space debris hazard, however, it is likely the tether will be made up of redundantly interconnected multiple lines whose individual cross-sectional areas add up to A. Given these assumptions, the tether mass is then $m_T=dLA$, while the end-to-end tether resistance is $R_T=rL/A=rdL^2/m_T$.

Specific Conductivity Parameter:

The choice of the metal conductor to be used in a space tether is determined by a combination of low resistivity (high conductivity) and low density, with cost, strength, and melting point as secondary considerations for certain applications. Copper has a resistivity r=17.0 nΩ-m, a density d=8933 kg/m, and a "specific conductivity" of 1/rd=6,585 m^2/Ω-kg. Aluminum has a resistivity r=27.4 nΩ-m, which is significantly greater than that of copper, but it has a much lower density of d=2700 kg/m^3. As a result, aluminum's "specific conductivity" of 1/rd=13,500 m^2/Ω-kg is twice the conductivity per unit mass of copper. Silver, because of its higher density and higher cost, is not competitive as an electrodynamic space tether even though its resistivity of 16.1 nΩ-m is slightly better than that of copper. An alternate candidate material would be beryllium, with a resistivity r=32.5 nΩ-m, density d=1850 kg/m^3, and a "specific conductivity" of 1/rd=16,630 m^2/Ω-kg, slightly better than that of the much cheaper aluminum. Beryllium also has a higher melting point at 1551 K than aluminum at 933 K, so some of its alloys may be a preferred material for some electrodynamic applications despite its higher materials cost. Unfortunately, despite decades of metallurgical research by the nuclear power industry, highly ductile alloys of beryllium have not been found, so it is difficult to pull beryllium into wire. As a result, because of its high specific conductivity, low cost, and ready availability in ductile wire form, it will be assumed for this analysis that the electrodynamic tether will be made of aluminum wire.

Typical Resistance Values:

To be economically competitive, the mass of the tether needs to be a small fraction of the mass of the host spacecraft it is required to deorbit. Since a typical constellation spacecraft has a mass of about 1000 kg, a typical Terminator Tether™ with a mass that is 2% of the host spacecraft mass would consist of a deployer/controller package with a mass $m_D=10$ kg, containing an aluminum tether with a mass $m_T=10$ kg with a volume of $LA=m_T/d=3.70\times10^{-3}$ m^3. If this 10 kg of aluminum were formed into a tether with a length of L=2 km and a cross-sectional area of A=1.85 mm^2, then the end-to-end resistance of the tether would be $R=rL/A=rdL^2/m_T=29.6\ \Omega$ A longer tether would have a proportionately smaller cross-sectional area and a higher resistance; for example, a 5 km long tether with the same mass would have a resistance of 185 Ω.

Orbit Inclination Assumption:

In order to make the remainder of the analysis mathematically tractable, it will be assumed that the orbit of the host spacecraft is circular and above the magnetic equator, so that the angle between the radius vector to the spacecraft and the magnetic equatorial plane is β=0 degrees. In this orbit, the velocity v of the spacecraft with respect to the magnetic field is the orbital velocity w less the rotational velocity of the magnetic field at the orbital radius a due to the once per day rotation of the Earth, or:

$$v=W-2\pi a/1\text{day}=(7546-509)\text{m/s}=7037\text{m/s}$$

This equation also shows why electrodynamic drag will not be useful for removing geostationary spacecraft from orbit. At the geostationary orbital radius of a=42,200 km, the relative velocity of the spacecraft and the rotating magnetic field of the Earth is zero.

With this assumption of an orbit above the magnetic equator, the vertical component of the Earth's magnetic field is zero and the horizontal component of the Earth's magnetic field is at right angles to both the local vertical and the direction of motion of the spacecraft. The total magnetic field seen by the host spacecraft and its Terminator Tether™ is then orthogonally horizontal and has the magnitude:

$$B=B_H=B_o[R_e/a]^3\cos\beta=26.5\,\mu\text{T}.$$

Where for simplicity the subscript H will be dropped for the remainder of this analysis.

This orbit, with its inclination of 11.5° with respect to the spin equator, will not stay in the plane of the magnetic equator, but will vary ±11.5° above and below it as the Earth rotates, causing a variation in magnetic field strength of ±0.5 μT, as well as a slight variation in angle. This 2% variation is negligible in terms of the other uncertainties in this analysis. It is not until the orbital inclination of the host spacecraft orbit reaches 60°, where cos 60°=0.5, that there is a significant drop in the expected magnitude of the electrodynamic drag forces calculated in this analysis.

Electromagnetic Drag Effects in Polar Orbit:

In many medium Earth orbit communication satellite constellations, there are a significant number of spacecraft at high inclinations and in nearly polar orbits. The high inclination spacecraft, with inclinations between 60° and 78.5°, will all have orbits that stay between the magnetic poles. Although the amount of electrodynamic drag will be significantly less than that experienced by spacecraft with orbits at lower inclinations, the direction of the induced electric fields in the tether will always be in the proper direction. If the tether is vertically upward, the outer tip of the tether will be positively charged and the bare wire in the tether will pull electrons out of the surrounding space plasma, while the electron emitter at the host spacecraft end will eject the electrons back into the space plasma to complete the circuit.

For spacecraft in near polar orbits with inclinations between 78.5° and 90°, however, there will be much more drastic variations. First of all, for a spacecraft in a 90° orbit that happens to pass directly over the magnetic poles, there is no horizontal component of the magnetic field when it is passing over one of the magnetic poles, so no voltage is generated in the tether during that part of the orbit, while the horizontal component of the magnetic field near the magnetic equator, while strong in magnitude, is along the spacecraft velocity vector, so no voltage is generated in the tether in that part of the orbit either. As a result, there will be negligible electrodynamic drag experienced by the host spacecraft during that particular orbit. The Earth is rotating, however, and the magnetic pole is rotating with it. A few orbits later, the host spacecraft will be passing over the Earth's spin pole at a point where the horizontal component of the magnetic field is exactly at right angles to the direction of motion of the spacecraft, so the full voltage is generated in the tether. When passing over the spin pole, the spacecraft radius vector is at an angle of 11.5° from the magnetic pole and at an angle β=78.5° away from the magnetic equatorial plane. With these assumptions, the magnitude of the horizontal component of the magnetic field at the Earth's spin pole is a respectable:

$$B=B_o[R_e/a]^3\cos 78.5°=5.28\,\mu\text{T}$$

or 20% of the maximum value experiences by spacecraft orbiting above the magnetic equator. This value will drop slightly as the orbit continues, then build back up as the spacecraft passes over the opposite pole. As the Earth continues to rotate, bringing the magnetic pole again under the orbit, and the interaction of the tether with the horizontal component of the magnetic field again drops to zero. The average coupling of a tether to the Earth's magnetic field over all polar trajectories has been calculated to be 12.3 percent.

A problem experienced only by spacecraft with orbit inclinations greater than 78.5°:

The spacecraft will no longer be traveling from west to east with respect to the magnetic field axis, but will have a retrograde motion as it moves through the magnetic field. As a result, the voltage generated by the motion of the tether through the Earth's magnetic field will switch direction. The outer tip will be negatively charged and will attempt to collect ions, which is a much less efficient process than collecting electrons.

There are a number of solutions to this problem. The first is to increase the mass and length of the tether supplied to a spacecraft assigned to a polar orbit, so that higher voltages, currents, and drag are generated during the limited times the spacecraft is passing over the poles in the right direction. The second is to supply a tether with plasma contactors at both ends that can emit electrons from either end, allowing the current to flow either way, depending upon which direction the spacecraft is passing around the magnetic pole. The third is to utilize the first number of passes to torque the orbit of the spacecraft until the orbit inclination has been shifted below 60°, then turn on the electrodynamic drag full time to deorbit the spacecraft from this more favorable orbit inclination. This orbit torqueing maneuver is accomplished by activating the electrodynamic drag mechanism only when the magnetic field orientation is such that a strong out-of-plane component of force is created. This orbit torqueing maneuver can be augmented by switching to a propulsion mode, where power saved in the batteries during the drag force mode is pumped back into the tether when the magnetic field is in the opposite direction, applying electrodynamic propulsion to torque the orbit even further. With the tether at a large and stable trailing orientation, the coupling to the magnetic field can be significant to a vertical tether.

In summary, spacecraft in near-polar orbits might take longer to bring down, and might have to utilize specially designed Terminator Tethers™ that might cost and mass more than the simpler Terminator Tethers™ usable in lower inclination orbits, but a Terminator Tether™ can still remove a spacecraft from a polar orbit when desired.

The inventors now present a detailed analysis and optimization of a typical Terminator Tether™ attached to a typical host spacecraft in a typical low inclination orbit.

Electromotive Generation of Voltage and Current in the Tether:

When an object is moved at a velocity v through a magnetic field B, an electric field is generated in the frame of reference of the moving object given by:

$$E = v \times B = vB$$

where the magnetic field B of the Earth, being mostly tangent to the Earth's surface in the north-south direction, is at right angles to the velocity vector v of the spacecraft, assumed to be orbiting in a generally west-east direction. The direction of the electric field E will be at right angles to both v and B, or along the local vertical. It should be noted that this electric field exists in the moving frame of reference because a moving magnetic field creates an electric field. No object actually has to be there, but if it is, then the relative motion of the magnetic field of the Earth will not only apply magnetic forces to whatever material the object is made out of, but electric forces too.

Note also that the velocity used in this equation is the relative velocity between the object and the magnetic field. Because the Earth's magnetic field rotates with the Earth, the motion of the magnetic field must be subtracted from the orbital velocity of the object to obtain the relative velocity.

Voltage generation in a Conducting Tether:

If the moving object is a long conducting wire of length L, the electric field E, generated in the wire produces a voltage V between the opposite ends of the wire given by:

$$V = E \bullet L = EL \cos \tau = vBL \cos \tau$$

where τ is the angle between the length vector L of the tether and the electric field vector E, assumed to be in the vertical direction at right angles to the velocity vector v in the plane of FIG. **12** and the magnetic field vector B out of the plane of FIG. **12**. A typical value for the voltage level built up in a vertically oriented tether with length L=5 km and $\tau=0$, moving at a velocity of v=7037 m/s through the Earth's horizontal magnetic field of strength 26.5 μT, is 932 V, or 0.186 V/m. Spacecraft in higher inclination orbits would experience somewhat smaller electric fields. For calibration, the voltage measured between the ends of the 19.5 km long TSS-1R tether was 3500 V, or 0.175 V/m.

Contacting the Space Plasma:

Although a voltage will build up between the ends of the conducting tether, no current will flow unless the circuit is completed. The circuit cannot be completed with another wire, for it too will have a similar voltage generated in it by the moving magnetic field. Fortunately, empty space is not empty, and in near-Earth regions not too distant from the Earth's atmosphere there exists highly electrically conductive space plasma, kept partially ionized by radiation from the Sun. The electron and ion density varies from the dark to light side, with altitude, with season, with sunspot cycle, with contamination level, etc. but it typically varies from 10^{11} to 10^{13} electrons/cc.

Fortunately, it has been found that if a bare conductive surface such as the spherical Italian spacecraft in the TSS-1R experiment, or the long bare wire to be used in the ProSEDS experiment, or a multiline conducting wire Hoytape mesh, is charged to a few hundred volts, the bare conductor will readily pull electrons out of the space plasma. So all that is needed to complete a connection to the space plasma at the positively charged (upper) end of the tether is a sufficient large area of uninsulated conductor.

It is more difficult to eject electrons from a wire or to collect positive ions from the space plasma. Although it is conceivable that a very large area at the other end of the tether could collect enough ions to complete the circuit as was demonstrated during the TSS-1R mission when the Shuttle orbiter was found to be an adequate plasma contactor for over an amp of current, the present method chosen is to use an electron emitter of some sort, either a hot cathode, a plasma cathode or contactor, a field-emission device, or something similar. Once provisions have been made at both ends of the tether to allow the flow of electrons out of one end of the tether and into the other end, and the altitude of the host spacecraft is not too high, then there will be sufficient conductivity in the space plasma surrounding the host spacecraft to allow current to flow through the tether.

Current Flow In The Tether:

The amount of current I flowing through the tether depends upon the total resistance R in the circuit. This resistance will consist of three components, the effective resistance of the plasma, the resistance of the tether, calculated earlier as $R_T=rL/A=rdL^2/m_T$, and a control resistor R_C, which will be varied as needed to optimize the Terminator Tether™ performance. There will also need to be a parasitic load on the current in the form of a charging device to charge some batteries. The batteries in turn will be used to power the control and communication circuits, and drive the electron emitting devices at the negative end of the tether. A well-designed Terminator Tether™ will thus be completely self-powered, except for an initial charge in the batteries to provide electrical power for the deployment and startup procedure. For simplicity of analysis it will be assumed that this battery charging load, which absorbs power like a resistor, but which stores it and uses it later instead of dissipating it immediately as heat, is included in the control resistor R_c. Normally, this load would act as a "base resistance" below which the control resistor could not be lowered, but since the charging circuit can be turned off when desired, and operations continued without interruption using the power stored in the batteries, it will be assumed that the control resistor can be taken to near zero value in those circumstances where the space plasma conductivity is low, or the magnetic field is in the wrong orientation and the voltage being generated in the tether is not large. Under these circumstances, lowering the control resistor to near zero allows a much higher current to flow for a given generated voltage, thus increasing the power being dissipated in the tether and maintaining a high level of electrodynamic drag on the host spacecraft.

A properly designed Terminator Tether™ will have plenty of bare metal area for electron collection at the positive end, while the electron emitters at the other end are efficient in terms of emitting large electron currents at low voltages and therefore low powers, while at the same time the mass, length, and area of the conducting tether have been made such that the resistance of the tether is moderately high. Under these conditions, the effective resistance of the space plasma will be much less than the design resistance of the tether. To make the mathematics more tractable it will be assumed that the plasma resistance is can be neglected and that the voltage available to drive the tether is the full voltage generated by the moving magnetic field. Although voltage will be needed to power the electron emitter, whatever its form, it will be assumed that the required voltage and power will be supplied by the batteries. The batteries in turn will be powered by the battery charging circuit, whose effective resistance is included in the control resistor (except for those short periods when the plasma or magnetic field interactions are weak).

Given all these assumptions concerning the total resistance in the circuit, the current I flowing through the tether is then given by:

$$I=V/R=vBL\cos\tau/(R_C+rdL^2/m_T)$$

A typical current level can be estimated by assuming that the control resistance R_C should be about the same size as the tether resistance R_T, which for an aluminum tether with resistivity r=27.4 nΩ-m, density d=2700 kg/m^3, length L=5 km, and mass m_T=10 kg, is R_T=185 Ω, giving a total resistance, including the control resistance R_C, of R=370 Ω. Thus the current flowing though a tether of length L=5 km, generating a voltage of V=932 V, would be I=V/R=2.52 A.

Power Dissipation Analysis:

The induced current I flowing through the tether resistance R_T will generate heat in the wire, which will be radiated into space and lost. That radiated energy must come from somewhere. There is also power taken out of the current flow by the battery charging circuit and ultimately dissipated. In order to conform to energy conservation laws, this power and energy must come from a decrease in the total kinetic and potential energy of the host spacecraft, causing it to deorbit. Set forth below is a detailed, optimized force analysis which will calculate the drag force on the host spacecraft. In this section, by using some simplifying assumptions, it is possible to produce a general argument based on energy conservation laws, which will illustrate the broad scope of the present invention.

For this energy conservation analysis it will be assumed that the plasma resistance is small compared with the tether resistance, and that the control resistance is equal to the tether resistance:

$$R_C=R_T=rL/A=rdL^2/m_T$$

Assume that the ballast mass at the end of the tether is a large piece of the defunct host spacecraft, such as a solar panel, antenna, or battery pack, so that the gradient force is large and the tether is always oriented along the vertical so that the angle τ=0°. Also assume the spacecraft is in an orbit above the magnetic equator and thus is moving at right angles to the Earth's magnetic field, so that the angle between the orbital velocity vector and the magnetic field vector is 90°, and the electric field vector is in the vertical direction, aligned perfectly with the tether length vector. Under these ideal conditions, the voltage between the ends of the tether is given by:

$$V=E\bullet L=(v\times B)\bullet L=vBL$$

The current in the tether is then just:

$$I=V/(R_C+R_T)=vBm_T/2rdL$$

The power dissipated as ohmic heating in the tether is then given by:

$$P=IV=(vB)^2m_T/2rd$$

This equation shows the interesting result that the power dissipated does not depend upon the length or the area of the tether (within reason), but only the mass of the tether m_T, the resistivity r and density d of the tether material, and the velocity v of the conducting tether through the Earth's magnetic field B.

For a typical aluminum tether of mass m_T=10 kg, the power dissipated is an impressive P=2350 W. Even if only a small fraction of this power is dissipated in a real system, the kinetic energy of the host spacecraft would suffer a significant energy loss with time.

Orbital Energy Extraction Analysis:

The total energy U of a spacecraft of mass M in orbit around the Earth consists of two components, its positive orbital kinetic energy and its negative gravitational potential energy. If a circular orbit of altitude h and radius a=R_e+h is assumed, then the total energy of the spacecraft moving at a velocity w=$(GM_e/a)^{1/2}$ is:

$$U=+Mw^2/2-GMM_e/a=-GMM_e/2a$$

The system will be dissipating energy in the tether to decrease the energy U of the unwanted spacecraft from its relatively low negative value in a high Earth orbit to a greater negative value in a lower Earth orbit. To give a specific example, an unwanted spacecraft with a mass of M=1000 kg (a metric ton) in an orbit with an altitude of h=1000 km and a high orbital radius of a(H)=R_e+h=7371 km, then its (negative) total energy is:

$$U(H)=-GMM_e/2a(H)=-27.1\text{ GJ or }-27.1\text{ MJ/kg}$$

It is desired to lower the spacecraft to an orbit with an altitude just outside the atmosphere of h=200 km or a lower orbital radius of a(L)=R_e+h=6571 km, where atmospheric drag will finish the job. The total (negative) energy of a metric ton spacecraft at 200 km altitude is:

$$U(L)=-GMM_e/2a(L)=-30.4\text{ GJ}$$

Thus, the amount of energy needed to be dissipated in order to move the 1000 kg spacecraft from a 1000 km orbit with a total energy U(H)=−27.1 GJ down to a 200 km orbit with energy U(L)=−30.4 CJ is:

$$dU=U(H)-U(L)=3.3\text{ GJ.}$$

The decay time T of a metric ton spacecraft moving from a 1000 km altitude orbit to a 200 km altitude orbit with an energy difference of dU=3.3 GJ, when its energy is being dissipated at a power of P=2350 W by a typical aluminum tether massing just 10 kg or 1% the mass of the spacecraft, is found to be T=dU/P=16 days.

The actual decay time will be longer than this. If the electrodynamic drag force is very large, and becomes larger than the gravity gradient forces pulling on the ends of the tether, then the tether will tend to align itself along the magnetic field lines instead of across them, and the drag force will decrease because of the small angle between the conductor length and the magnetic field lines. The tether current will need to be controlled until the angle of the tether settles into an angle determined by the balance between these two forces. These electrodynamic torque vs. gradient torque dynamic interaction effects are discussed in the next section.

Force and Torque Balance Analysis:

It is now possible to calculate the forces and torques on the tether. It is also possible to use the fact that the electrodynamic and gravity forces and torques on the tether must balance each other out to calculate and teach some optimum values for some of the Terminator Tether™ parameters.

Electrodynamic Force and Torque:

As discussed above, both theory and experiment show that: provided the conducting tether is moved rapidly through the Earth's magnetic field in order to generate a voltage across it, and provided good contact is made with the space plasma, the conducting tether will have a current flowing through it. When a wire (moving or not) carrying a current I is embedded in a magnetic field B, there will be an electrodynatnic force F_E generated on each element of the wire. The electrodynamic force will be at right angles to the magnetic field vector and the length vector of the wire element, with a magnitude given by:

$$F_E=J\times B=ILB$$

where B is the horizontal component of the magnetic field, which is perpendicular to the assumed magnetic equator orbit plane, while the tether length vector L is assumed to lie in the orbit plane. If, as will be the case most of the time, the electron current is leaving the space plasma and entering the tether along the length of the tether, then IL needs to be replaced with an integral of the current along the length of the tether.

Note that the electrodynamic force amplitude or direction is not directly dependent on the motion of the tether through the Earth's magnetic field. The electrodynamic force would be the same if the tether were not moving and the current was being supplied by a battery. Since the current I is a function of the orbital velocity, however, there is a secondary dependence of the electrodynamic force magnitude on the orbital velocity, but not its direction. The electrodynamic force is always at right angles to the conductor, and stays at right angles to the conductor as the angle τ varies, as shown in FIG. **12**.

Assuming that the electrodynamic drag force is applied uniformly along the length of the tether, it is possible to make the simplifying assumption that the integrated force is effectively applied at right angles to the center of mass of the tether at the point L/2 as shown in FIG. **12**. The electrodynamic torque on the tether is then:

$$T_E=F_E L/2=IBL^2/2=vB^2L^3\cos\tau/2(R_C+rdL^2/m_T)=vB^2L^3\cos\tau/2R$$

Gravity Gradient Forces and Torques:

When a tether and its ballast end mass are deployed from a host spacecraft, the gravity gradient force field of the Earth, combined with the orbital centrifugal gradient force field, will cause the tether to deploy either up or down from the host spacecraft. The direction desired depends on which end of the tether is connected to the electron emitter. Normally, the electron emitter will be on the end attached to the host spacecraft, in which case the desired direction of deployment will be upward so that the induced voltage in the tether will produce an excess of electrons at the electron emitter end of the tether. The desired upward direction is chosen by having the deployer eject the ballast mass in the upward direction. Once the ballast mass has been started in that direction, the centrifugal force due to the orbital motion of the ballast mass will cause the ballast mass to continue to accelerate in the upward direction until it is brought to a halt by the full deployment of the tether.

If there were no electrodynamic or atmospheric drag, the equilibrium direction of the tether would be exactly along the vertical, since the combined gradient field is a maximum in that direction. Because a significant amount of electromagnetic drag is expected, the actual angle of the tether with respect to the local vertical will be at some angle τ, lagging behind the spacecraft motion in the plane of the orbit, as shown in FIG. **12**. In the following analysis shows that there is an optimum angle for τ that produces the largest electrodynamic drag force on the host spacecraft without producing tether instability.

The combined vertical gravity gradient and centrifugal gradient field 3Γ acting on the ballast mass m_B at the end of the tether of length L will produce a gradient force F_{GB} given by:

$$F_{GB}=-3\Gamma m_B L\cos\tau$$

The strength of the force depends not only on the ballast mass m_B and the strength of the gradient field 3Γ, but also the component along the radial direction of the distance of the ballast mass from the center of mass, which is L cos τ. This force acts in the vertical direction along the radius vector leading from the ballast mass away from the center of the Earth. The amount of gradient force is not large. For a ballast mass of $m_B=10$ kg and a tether length of L=5 km, the gradient force is about 0.175 N.

As can be seen in FIG. **12**, the component of this gradient force that is at right angles to the tether, given by $F_{GB}\sin\tau$, will produce a torque T_{GB} on the tether that tends to restore the tether toward the vertical, lessening the angle τ.

$$T_{GB}=LF_{GB}\sin\tau=-3\Gamma m_B L^2\sin\tau\cos\tau$$

The tether mass m_T also contributes to the gradient force and torque. If it is assumed that the tether has a uniform cross section, then it is possible to replace the distributed mass of the tether with an equivalent point mass m_T placed at the center of mass of the tether, which is the point L/2 along the tether, and a distance L/2 cos τ in the radial direction. The gradient force due to the tether mass is then:

$$F_{GT}=-3/2\Gamma m_T L\cos\tau$$

While the gradient torque is:

$$T_{GT}=-L/2F_{GT}\sin\tau=-3/4\Gamma m_B L^2\sin\tau\cos\tau$$

The total gradient torque attempting to restore the tether to its vertical orientation is then:

$$T_G=T_{GB}+T_{GT}=3\Gamma(m_B+m_T/4)L^2\sin\tau\cos\tau$$

It is important to notice the variation of the total gravity gradient torque as the tether angle τ is varied. Since the gradient force is always in the radial or vertical direction, there is no torque on the tether when the tether is vertical, as is the case when there are no aerodynamic or electromagnetic drag forces. Once the drag forces become important and start to apply torque to the tether, increasing the tether angle τ, those drag torques causing an increase in tether angle τ will be opposed by a rising gradient torque which will attempt to decrease the tether angle. For small tether angles, the gradient torque increases first linearly with τ, then as sin τ, since cos τ is near unity for small τ, then at τ=45° the gradient torque reaches its maximum, where sin τ=cos τ=0.707 and sin τ cos τ=0.50. When this angle is reached, the tether is a point of catastrophic instability, for if there is a further increase in the electrodynamic drag force, causing the angle τ to become greater than 45°, the gradient torque, instead of growing stronger to counteract the increased drag torque, will become weaker and the tether angle will go rapidly to 90°.

To restore control to the tether angle if the instability occurs, it will be necessary to turn off the electrodynamic drag forces by shutting off the current flow through the tether. The τ=90° position for the tether and ballast mass is a gravitationally unstable orientation. After a time, slight fluctuations in the gravity field will allow the gradient force to slowly take over and restore the tether to the vertical orientation, which, unless it can be controlled in some way, is equally likely to be up or down. It would therefore be desirable to maintain control of the tether angle so as to avoid the tether angle getting into the region of instability.

Torque Balance on the Tether:

The angle τ of the tether is determined by a balance between the electrodynamic torque T_E attempting to increase the angle τ and the gradient torque T_G attempting to decrease the angle τ. Balance is achieved when the two torques are equal:

$$T_E=T_G=T_{GB}+T_{GT}$$

or:

$$vB^2L^3\cos\tau/2R=3_{\Gamma}(m_B+m_T/4)L^2\sin\tau\cos\tau$$

Simplifing and rearranging yields an equation giving us the angle τ at which torque balance occurs:

$$\tau=\arcsin[vB^2L/6_{\Gamma}(m_B+m_T/4)R]$$

The physical interpretation of this equation is that the maximum electrodynamic force that can be sustained on the tether is limited by the gradient force on the tether. It is possible to increase the electrodynamic force by decreasing the total resistance R of the tether (for example, by using a tether with a larger mass and a lower tether resistance), but if the tether resistance is too low, then the quantity in brackets becomes greater than unity and this equation has no solution, indicating that the tether has reached an angle where instability sets in.

To maintain control of the tether angle, it will be necessary to vary the control resistance of the tether to compensate for variations in magnetic field strength and direction, plasma density (which affects the plasma resistance), and other factors, and thereby maintain the tether at an intermediate angle where both the electrodynamic and gradient forces are at an appreciable level and balance each other.

Optimization of Tether Angle:

At first glance, it might seem that the optimum angle for the tether would be 45°, since at that angle the gradient torque is largest and therefore can counteract a larger electrodynamic drag force. However, since the 45° angle is the point where instability sets in, it is desirable to set the tether angle at some value below 45°. The optimum angle is that which maximizes the horizontal or drag component of the electrodynamic force. This optimum angle τ is derived from the equation for the horizontal component of the electrodynamic force, or the electrodynamic drag force, since it opposes the host spacecraft motion:

$$F_{ED}=F_E\cos\tau=6[m_B+m_T/4]\Gamma L\sin\tau\cos^2\tau$$

By setting the partial derivative$\square F_{ED}/\square\tau=0$ and solving, it is possible to calculate that the optimum angle for the tether that gives the maximum electrodynamic drag force F_{ED}, while still keeping the tether torques balanced and under control, is τ=arctan(0.707)=35.26°. With this angle selected and maintained by varying the control resistor R_C to compensate for variations in plasma contact resistance and variatins in the strength and direction of the Earth's magnetic field B seen at the spacecraft, the tether experience the maximum stable value for the electromagnetic drag force of:

$$F_{ED}(\max,\tau=35.26^\circ)=2.31[m_B+m_T/4]\Gamma L$$

There are many ways to provide the sensing information needed to provide the feedback signals to the control resistor, but the simplest is to merely measure the drag acceleration on the host spacecraft with a set of accelerometers, and maximiize the force in the direction opposite to the host spacecraft motion. Another method would be to measure the current in the tether, and knowing the tether resistance and the amount of control resistance, calculate the power being extracted and maximize that value. Alternate methods would be to use an optical position sensor or GPS receivers at both ends of the tether to measure the angle of the tether or the position of the ballast mass with respect to the host spacecraft.

To make an estimate of the magnitude of drag force attainable, assume that a typical aluminum tether with a length of L=5 km, a tether mass of m_T=10 kg and a ballast mass of m_B=10 kg is being used. Then the maximum gradient force limited electrodynamic drag force is F_{ED}=0.167 N. The power being dissipated in the canoical tether at the angle τ, with $R_C=R_T=rdL^2/m_T=185\ \Omega$ is given by:

$$P=IR=[vBL\cos\tau]^2(R_C+rdL^2m_T)$$

which for an angle τ=35.26° results in a power dissipation of 1570 W. Thus, the thrust level per unit power obtained by the typical Terminator Tether™ would be about 0.106 N/kW. This number is comparable to the value of 0.148 N/kW estimated for the much heavier and longer TSS-1R tether.

Reconciliation of Energy and Force Analysis:

By a force and torque balance analysis an optimum angle for the tether has been discovered at which it is possible to obtain a maximum in the drag component of the electrodynamic force. There is also an additional component of the electrodynamic force, the component in the vertical direction, $F_{EV=FE}$ sin(τ), which is downward for an upwardly deployed tether. This component of force combines with the gravity force of the Earth to effectively allow the host spacecraft to orbit a little faster than normal for that orbital altitude. It does not contribute to the deorbiting of the host spacecraft. But since this vertical component of force is created by current running through the tether, and that current is creating heat and dissipating energy as it passes through the tether, there might be some concern that the force and torque balance analysis above does not conform to the law of conservation of energy.

The inventors will now show, in a very general manner, that no matter what the tether angle, the electrical power being dissipated in the tether is exactly equal to the power being lost by the slowing of the host spacecraft.

The "deorbit power" P_D that must be removed from a spacecraft moving at a velocity v when that motion is opposed by a drag force F_{ED} is:

$$P_D=v\bullet F_{ED}=vF_E\cos\tau=vILB\cos\tau=I[vBL\cos\tau]$$

But, since the voltage V induced across the tether of length L and tilt angle τ moving at a velocity v through a horizontal magnetic field B is

$$V=E\bullet L=EL\cos\tau=vBL\cos\tau$$

Therefore:

$$P_D=IV=P_E$$

And the deorbit power P_D extracted from the slowing of the spacecraft by the drag component of the electrodynamic force is always exactly equal to the electrical power P_E being dissipated as heat in the tether circuits, independent of the tether angle.

Optimization of Tether Mass Distribution:

If it is not possible to use a piece of the host spacecraft as ballast mass, then the mass of the ballast must be included in the Terminator Tether™ mass. It would be desirable to minimize the total Terminator Tether™ mass, since every kilogram saved means that another kilogram's worth of revenue-producing transponders can be added to the working payload of the host spacecraft. Given a total mass for the Terminator Tether™ and the mass of the deployer/controller unit, it is possible to optimize the mass distribution between the ballast mass and the tether mass to obtain a minimum total Terminator Tether™ mass. A well-designed Terminator Tether™ will also have most, if not all, of the deployer mass incorporated into the ballast mass.

Assume that the total Terminator Tether™ mass consists of the deployer/controller mass m_D, the tether mass m_T, and the ballast mass m_B, with $m_B > m_T > m_D$. Of the three mass components in the Terminator Tether™, two of them affect the electrodynamic drag performance. If it is assumed that the ballast mass is a factor X larger than the tether mass, or $m_B = Xm_T$, the maxinum drag force that can be obtained is now:

$$F_{ED}(\max)=2.31m_T(X+0.25)\Gamma L$$

If it is assumed that the control resistor has been adjusted so that this maximum value for the electrodynamic drag force is maintained as the motion of the spacecraft along its orbit moves the spacecraft into regions with different magnetic field strengths and plasma densities, then for the angle $\tau=35.26°$, $\cos(35.26°)=0.817$, and the maximum electrodynamic drag force in terms of the electrodynamic parameters will be:

$$0.817vB^2L^2/(R_C+rdL^2/m_T)=F_{ED}(\max)=2.31m_T(X+0.25)\Gamma L$$

or canceling out terms and rearranging:

$$(X+0.25)=0.353vB^2L/\Gamma m_T(R_C+rdL^2/m_T)$$

It is now possible to make the further simplifying assumption that to maintain control of the tether, the control resistor needs to be roughly the same size as the tether resistance or $R_C \approx rdL^2/m_T$. Using this assumption produces an equation for the optimum value for the ratio X of the ballast mass to the tether mass:

$$\begin{aligned} X = vB^2/\Gamma rdL - 0.25 &= -0.25 + \frac{vB^2}{12\sin\gamma\cos\gamma\Gamma rdL} \\ &= -0.25\frac{\Lambda}{L} \end{aligned}$$

Where Λ is the "effective electrodynamic length".

If the tether is short, so that L is small, this equation indicates that the ballast mass must be increased to increase the gradient force, since it is the gradient force which determines the upper limit to the amount of electrodynamic drag force that can be generated without losing control of the tether.

Assuming the usual values for the Earth's magnetic and gradient fields, the velocity of the host spacecraft through the magnetic field, and the resistivity and density of aluminum, this equation becomes:

$$X=10.9/L-0.25$$

where L is in kilometers.

For a typical tether of length L=5 km, X=1.93. If the total mass for the Terminator Tether™ is 20 kg, or 2% of the host spacecraft mass, and the mass of the deployer is 10 kg, then the remaining 10 kg should be distributed so that 6.59 kg is in the ballast mass while 3.41 kg is in tether mass.

Optimization of Tether Length:

If the ballast mass consists of a piece of the host spacecraft, then the mass of the ballast does not have to be considered in the optimization of the Tether system. In this case, the optimum distribution of the Tether mass is to put as much mass as possible into the tether, with a minimum in the deployer/controller package. Under this assumption, the optimum length of the tether is determined by the desire to keep the Area-Time Product of the host spacecraft plus Tether to a minimum.

In a long tether of length L, the effective "collision" cross-sectional area is not the area of the tether, but the larger area produced by multiplying the length of the tether by sum of the width of the tether plus the width of the "target" spacecraft that would be damaged by the tether hitting it. Thus, to decrease the "area" portion of the Area-Time Product, it would be desirable to shorten the tether, making it thicker, and perhaps slightly wider, at the same time, thus keeping the same tether mass and electron collection area, and thereby maintaining the drag force and keeping the decay time at the same level. The result will be a shortening of the Area-Time Product.

If the tether is shortened too much, however, the smaller voltage generated across the shorter tether will leave less voltage margin for the operation of the plasma contactors and the battery recharge system. Since the maximum voltage that can be generated is about 200 V/km, and typical plasma contactors and battery, chargers take 25–100 V to operate, a minimum length for a Tether would be roughly 2 km. Host spacecraft operating in polar orbits, where the conditions for voltage generation are not as good, may require a tether length of 5 km or more.

Since many watts of heat power will be dissipated in the control resistor, means must usually be provided to radiate the heat away into space. One low-mass method of accomplishing this is to make the control resistor (or resistors) in the form of a long, electrically insulated, high resistivity wire similar to those used in electric blankets, but designed to operate at a higher temperature, and during the fabrication of the electrodynamic tether, incorporate the high resistivity wire (or wires) into the end of the tether closest to the control circuit by weaving or braiding the insulated high resistivity wire in with the uninsulated aluminum wire of the electrodynamic tether. The surface area of a long wire is very large, so in this extended wire configuration, exposed to the space all around it, the hot control resistor wire can self-radiate its power into cold space without incurring the mass penalty of a separate radiating surface.

Implementation:

The basic optimum structure for a Terminator Tether™ would be one of the many types of Hoytethers. A multiline (6–12 primary line) Hoytape™ will provide the largest contact area with the plasma, since both sides of the tape would be able to pass current to the plasma. If the spacing between the primary lines is chosen to be larger than twice the Debye length of the plasma, then the effective current collection area per unit length of the Hoytape is proportional to the width of the Hoytape mesh, not the diameter of the wires in the mesh. Thus, a Hoytape not only provides an assured longer life for the Terminator Tether™, but very short lengths will also provide a very large current collection areas.

The deployer for the tether can deploy the Terminator Tether™ either down or up or both. The deployer can stay attached to the spacecraft as was done in the SEDS missions, or perhaps a better alternative would have the deployer ejected from the spacecraft, with one end of the tether still attached to the spacecraft, reeling out tether as it leaves. The empty deployer would then act as a ballast mass at the end of the Terminator Tether™, which would improve the performance.

The "Remora Remover"

In addition to attaching Terminator Tethers to spacecraft before launch, it is possible to consider a missile-like application. This "Remora Remover" missile would use a Terminator Tether™ carried by a seeker missile similar to the small "hit-to-kill" missiles developed by the Space Defense Initiative Office that deployed a net loaded with oriented metal rods. The Remora Remover missile would hunt down a spacecraft that needs to be removed from space, but instead of hitting the spacecraft, the missile would be programmed to rendezvous with the spacecraft and attach itself to the host spacecraft using a hooked net, harpoon, or adhesive "sucker". The Remora Remover missile would then deploy the Terminator Tether™, which would bring down both the spacecraft and the missile.

The inventors have presented two analyses, backed up with data obtained from space flight experiments. One analysis was based on generalized energy conservation laws while the other analysis use force and torque balance arguments. Both analyses show that optimized aluminum wire tethers 2 to 5 km in length and massing just 1% to 5% of the mass of the host spacecraft can deorbit the host spacecraft in the order of a few weeks, thus mitigating the long-lived orbital debris hazard created by a constellation spacecraft after their end-of-life.

Power Augmented Operation of the Tether:

It is well known in the aerospace literature that if a spacecraft has a power supply and a conducting tether hanging from it, and the spacecraft is in an orbit or trajectory that takes it near a body, such as the Earth or Jupiter, which has both a magnetic field and a region of moderate density electron-ion plasma, and electrical current is pumped from the power supply through the tether and back through the plasma, that the current flowing through the tether will interact with the magnetic field of the body, producing forces on the tether and thence on the spacecraft, changing its orbit or trajectory. These forces can be used to increase or decrease the spacecraft altitude and/or inclination. The amount of altitude or inclination change is proportional to the ratio of the power available from the power supply divided by the mass of the spacecraft and varies with altitude. The unpowered Terminator Tether™ is a simple example of such a system, in which only altitude decrease is possible, although small amounts of both inclination increase or decrease are possible at the same time. Some typical references to this prior art are: Les Johnson, "Propulsive Small Expendable Deployer System Mission (ProSEDS)", Proceedings of the OAST 8th Advanced Propulsion Workshop, JPL, Pasadena, Calif., 20–22 May 1997; and Les Johnson, Joe Carroll, Robert D. Estes, Enrico Lorenzini, Brian Gilchrist, Manuel Martinez-Sanchez, Juan Sanmartin, and Irwin Vas, "Electrodynamic Tethers for Reboost of the International Space Station and Spacecraft Propulsion, AIAA Paper 1996.

Since the Terminator Tether™ of the invention is normally associated with a satellite that has a power supply in the form of a solar panel combined with a storage battery, the Terminator Tether™ can be operated in the "powered propulsion" mode if desired. Such a mode would be useful when attempting to avoid a collision between the Terminator Tether™ and another spacecraft. In such cases, the power from the solar panel, augmented by the power stored in the battery, can be used to temporarily operate the Terminator Tether™ in the propulsion mode instead of the normal drag mode. This would enable to Terminator Tether™ to raise its altitude and/or increase or decrease its inclination to avoid a collision.

In addition, the Terminator Tether™ could be operated in the "power-augmented drag" mode. In this mode, there will be a current induced in the tether by the motion of the spacecraft through the magnetic field, which will cause a drag force on the spacecraft. This induced current would then be augmented by additional current in the same direction generated by the solar panel and the storage batteries in the power supply. The increased current will cause an increased drag force. If there is a significant vertical component of the magnetic field there will also be an increased force normal to the orbital plane, which will cause an increased rate of change of inclination.

When the Terminator Tether™ is operated in the "powered propulsion" or "power-augmented drag" mode, the conditions for stable optimum operation will be essentially the same as in the "unpowered drag" mode. The reason the conditions stay the same is that the maximum tether force than can be effectively utilized is limited by the strength of the gradient force, which has not changed, since the length and mass of the tether and the ballast mass has not changed for these different modes of operation. As a result the optimum operational angle for the tether is still 35.26 degrees, and the optimum ratio of ballast mass to tether mass for different tether lengths and tether material is still determined by the analysis of the invention as is set forth above.

In the analysis given above, where the optimum tether angle was found, the tether was assumed to be in the plane of the orbit and lagging behind the spacecraft motion, since in the "unpowered drag" mode, that would be its normal position, although if the component of the magnetic field along the orbit were high, there would be some tilt of the tether out of the plane of the orbit, producing a force tending to change the inclination of the orbit. In either of the "powered" modes of operation of the Terminator Tether™, however, the tilt of the tether could be forward toward the motion of the spacecraft or strongly tilted to one side or the other of the orbital plane. In all these cases, if the electrodynamic force is allowed to become too large, and the tether angle exceeds 45 degrees, the restoring force of the gradient field will drop off and the tether angle will go unstable. As a result, the optimum angle for the tether that will give the maximum stable force, whether it is a maximum drag force, and maximum propulsion force, or a maximum inclination change force, will be 35.26 degrees, and the optimum ratio for the mass of the tether versus the mass of the ballast will be the same as for the unpowered drag case.

The analysis that gave the optimum angle as 35.26 degrees assumed that the tether would be oriented like a rigid rotor. In reality, variations in forces along the tether will probably cause the tether to hang in a slightly curved shape, where the optimum angle may not be exactly 35.26 degrees from the local vertical. The optimum angle of the tether as it leaves the host spacecraft in such a case is where the drag force is largest.

The present invention is discussed in this disclosure in terms of its space applications as a Terminator Tether™ useful for the removal of unwanted satellites. It should be understood, however, that the present invention is useful in any application where a space object can use a conductive tether to produce electrodynamic force through interaction with a magnetic field. The invention, therefore, should be limited not by this description, but only by the appended claims and their equivalents.

What is claimed is:

1. A method of operating a conductive tether attached to a space object, said conductive tether being oriented at an angle to magnetic field lines of an external magnetic field and said conductive tether moving across the magnetic field lines of said external magnetic field at said angle to producing an electric current in said conducting tether and a resulting electrodynamic force acting on said tether and said attached space object comprising,

Controlling the amount of electric current flowing in said conducting tether to vary the angle between the conductive tether the field lines of the external magnetic field and the electrodynamic force acting on said tether and attached space object.

2. A method as in claim **1** wherein said angle is less than 45 degrees.

3. A method as in claim **1** wherein said angle is 35.26 degrees.

4. A method as in claim **1** wherein said angle is controlled to average 35.26 degrees over time.

5. A method as in claim **1** wherein said angle is controlled to average 35.26 degrees over the length of the tether.

6. A method as in claim **1** wherein the space object and the tether connected to it are rotated about their center of mass while the conductive tether interacts with the external magnetic field.

7. A method as in claim **1** wherein the angle of the conductive tether to the external magnetic field is controlled by sensing a measurable parameter of the space object/conductive tether system.

8. A method as in claim **7** wherein the parameter is current flow in the conductive tether.

9. A method as in claim **7** wherein the parameter is the electrical power dissipated in the conductive tether and the current control system.

10. A method as in claim **7** wherein the parameter is the relative position of the space object and the position of the far end tip of the tether.

11. A method as in claim **7** wherein the parameter is the acceleration of the space object.

12. A method as in claim **6** wherein the parameter is the state vector of the space object as measured by the Global Positioning System.

13. A method as in claim **1** including the step of applying electric power to the conductive tether to change the state vector of the space object.

14. A method of operating a conductive tether attached to a space object, said conductive tether being oriented at an angle to the local vertical between said space object and a celestial body having a magnetic field having magnetic field lines external to said conductive tether; said conductive tether moving across said magnetic field lines of said external magnetic field at an angle to the local vertical to produce an electric current in said conducting tether and a resulting electrodynamic force acting on said tether and said attached space object comprising,

Controlling the amount of electric current flowing in said conducting tether to vary the angle between the conductive tether the local vertical between said space object and said celestial body to vary the electrodynamic force acting on said tether and attached space object.

15. A method as in claim **14** wherein said angle is less than 45 degrees.

16. A method as in claim **14** wherein said angle is controlled to average 35.26 degrees over time.

17. A method as in claim **14** wherein the space object and the tether connected to it are rotated about their center of mass while the conductive tether interacts with the external magnetic field.

18. A method as in claim **14** wherein the angle of the conductive tether to the external magnetic field is controlled by sensing a measurable parameter of the space object/conductive tether system.

19. A method as in claim **18** wherein the parameter is current flow in the conductive tether.

20. A method as in claim **18** wherein the parameter is the acceleration of the space object caused by the electrodynamic force acting on the space object.

21. A method as in claim **18** wherein the parameter is the change in the state vector of the space object as measured by the Global Positioning System.

22. A method as in claim **14** including the step of applying electric power to the conductive tether to change the state vector of the space object.

23. A method as in claim **14** wherein the celestial body is the Earth and the magnetic field is the Earth's magnetic field.

24. A method as in claim **14** wherein the celestial body is the sun and the magnetic field is the Sun's magnetic field.

25. A method as in claim **14** wherein the magnetic field is the vector sum of the magnetic fields from all celestial bodies in the solar system.

26. A method as in claim **14** wherein the conductive tether is a Hoytether.

* * * * *

US008253639B2

(12) **United States Patent**
Cohen

(10) **Patent No.: US 8,253,639 B2**
(45) **Date of Patent: Aug. 28, 2012**

(54) **WIDEBAND ELECTROMAGNETIC CLOAKING SYSTEMS**

(76) Inventor: **Nathan Cohen**, Belmont, MA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 498 days.

(21) Appl. No.: **12/547,104**

(22) Filed: **Aug. 25, 2009**

(65) **Prior Publication Data**

US 2010/0156556 A1 Jun. 24, 2010

Related U.S. Application Data

(60) Provisional application No. 61/189,966, filed on Aug. 25, 2008.

(51) **Int. Cl.**
H01Q 15/02 (2006.01)
H01Q 19/06 (2006.01)

(52) **U.S. Cl.** **343/753**; 343/909

(58) **Field of Classification Search** 343/753, 343/909, 910; 333/135
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,471,224	A *	11/1995	Barkeshli	343/909
5,714,961	A *	2/1998	Kot et al.	343/769
6,473,048	B1 *	10/2002	Diaz	343/753

* cited by examiner

Primary Examiner — Michael C Wimer

(74) *Attorney, Agent, or Firm* — McDermott Will & Emery LLP

(57) **ABSTRACT**

Arrangement of resonators in an aperiodic configurations are described, which can be used for electromagnetic cloaking of objects. The overall assembly of resonators, as structures, do not all repeat periodically and at least some of the resonators are spaced such that their phase centers are separated by more than a wavelength. The arrangements can include resonators of several different sizes and/or geometries arranged so that each size or geometry corresponds to a moderate or high "Q" response that resonates within a specific frequency range, and that arrangement within that specific grouping of akin elements is periodic in the overall structure. The relative spacing and arrangement of groupings can be defined by self similarity and origin symmetry.

14 Claims, 4 Drawing Sheets

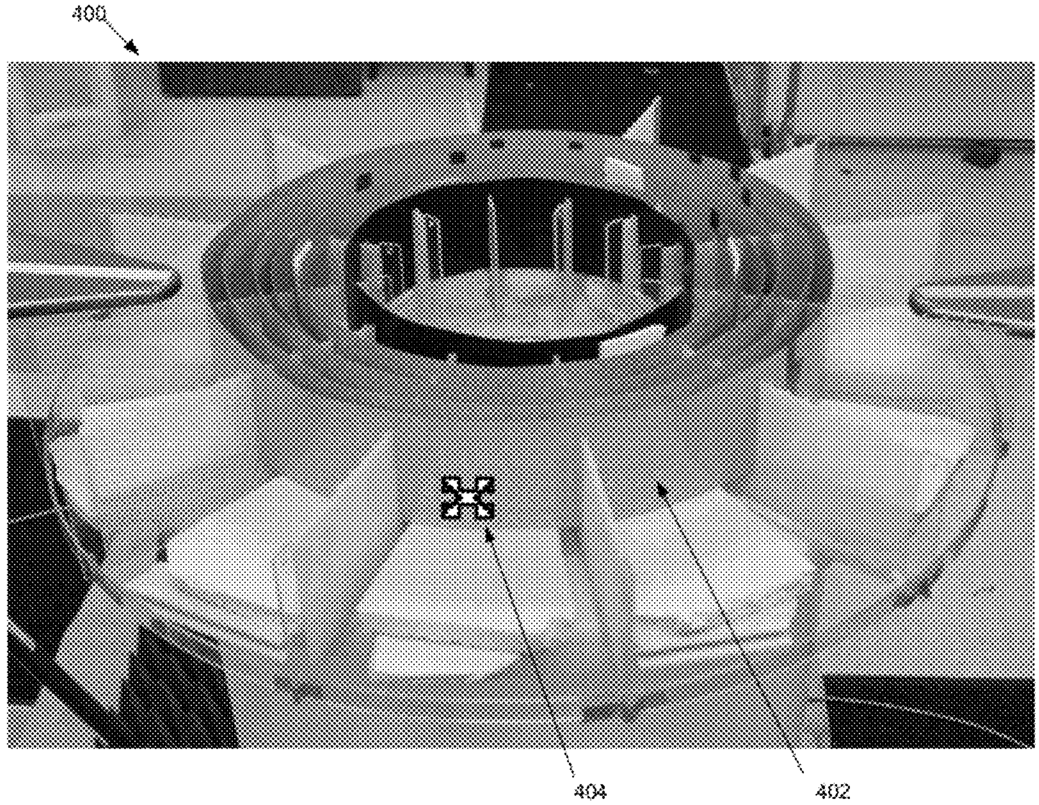

FIG. 1

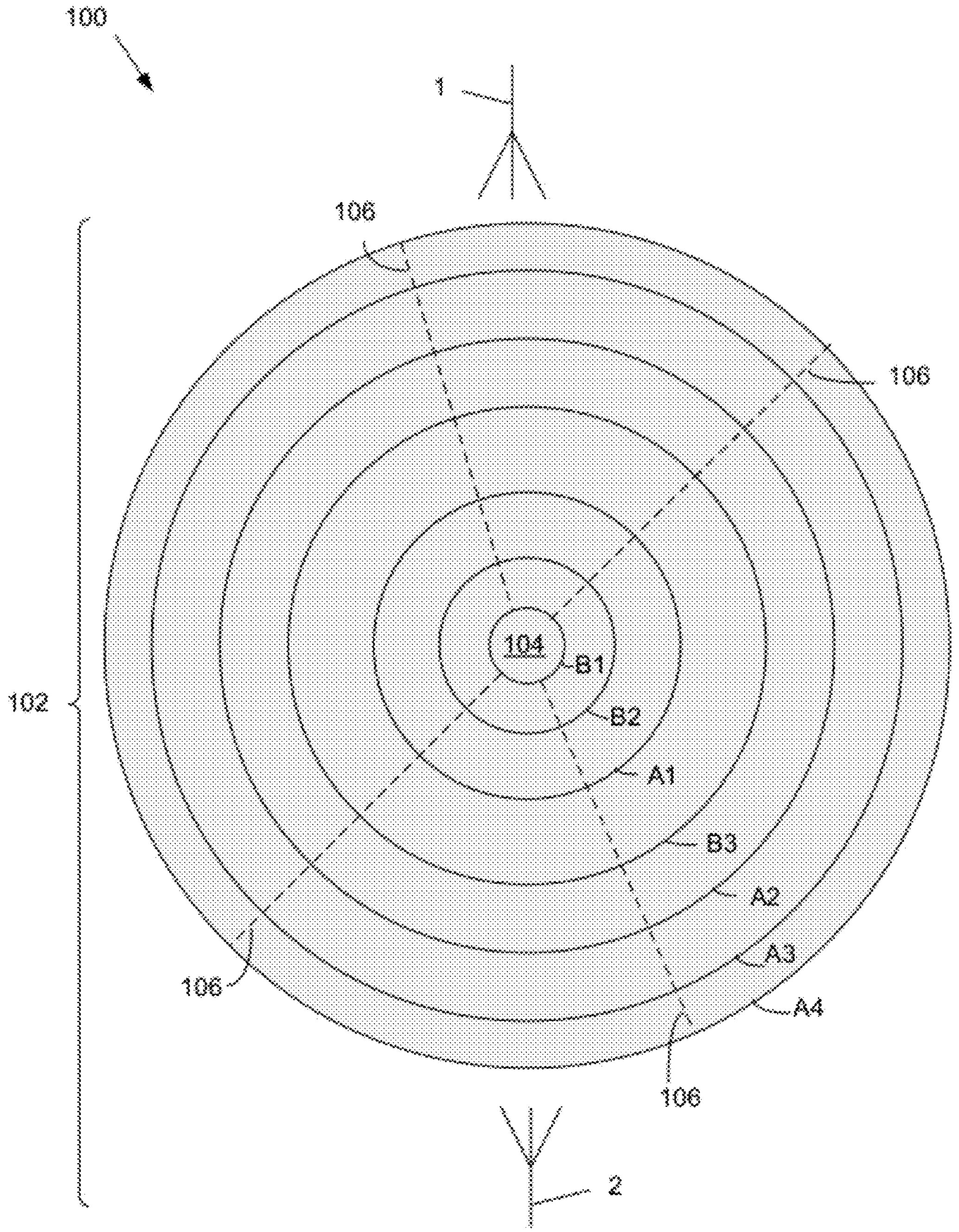
100
1
106
106
102
104
B1
B2
A1
B3
A2
A3
A4
106
106
2

FIG. 2

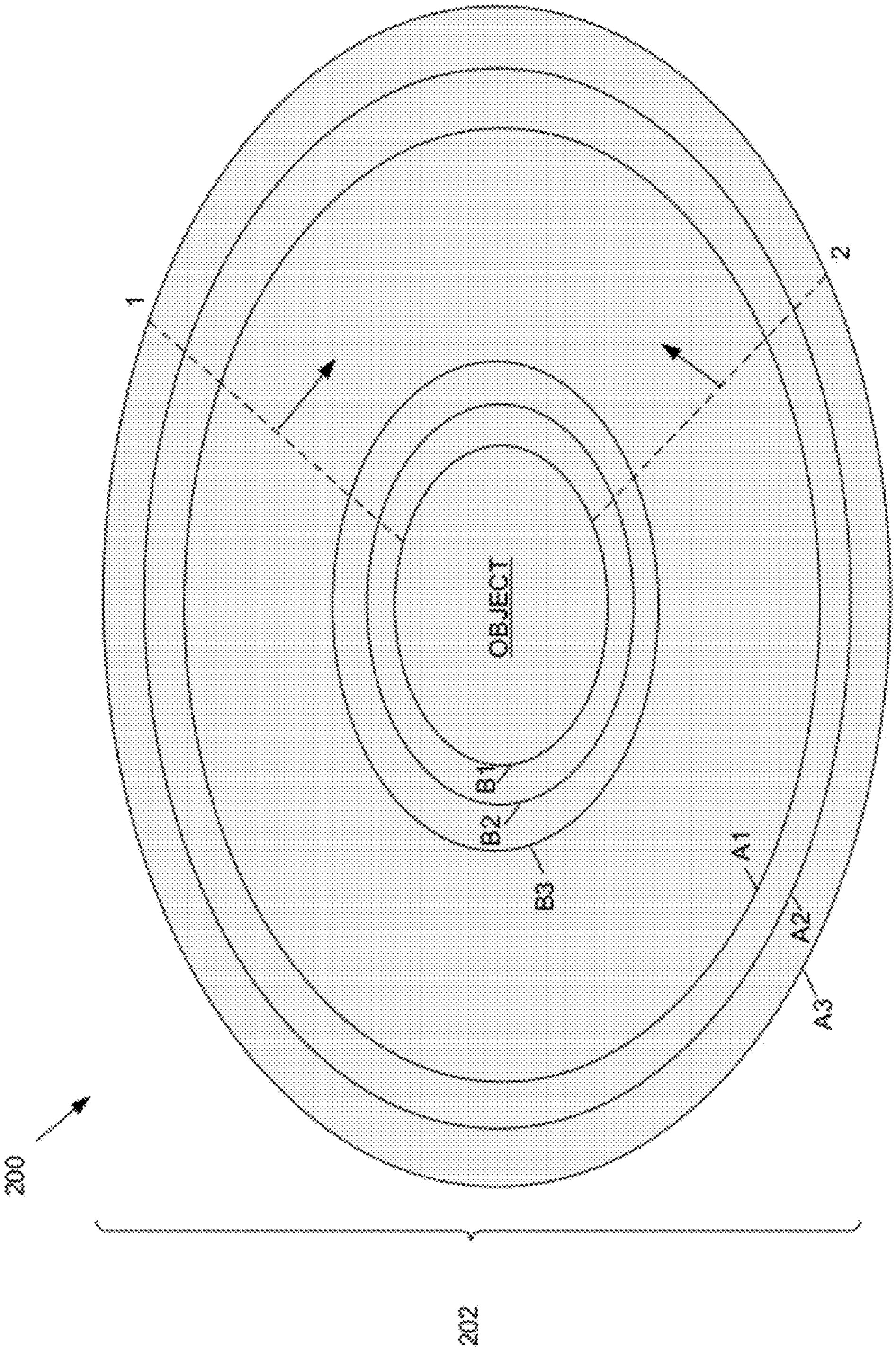
200
202
OBJECT
1
2
B1
B2
B3
A1
A2
A3

FIG. 3

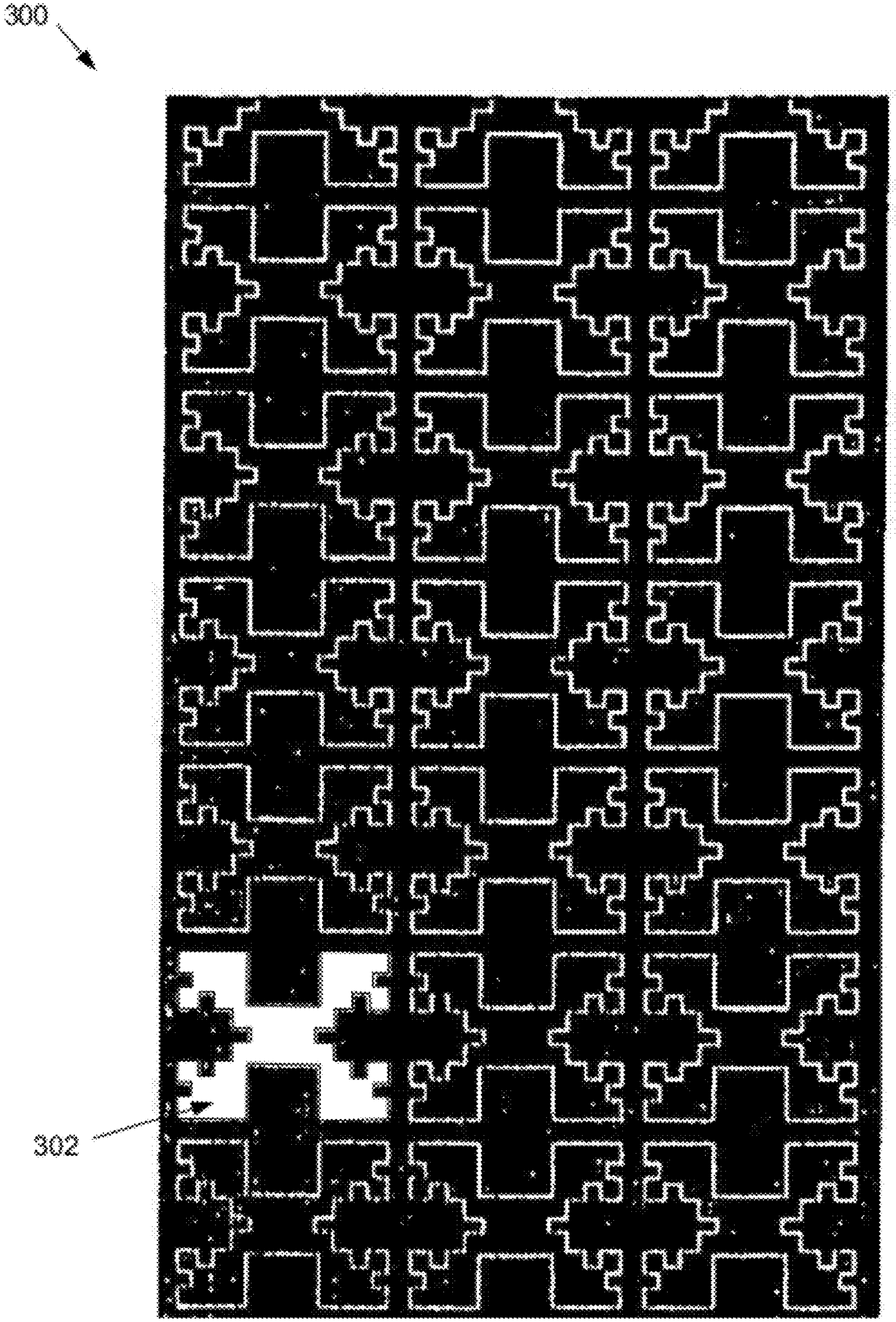

FIG. 4

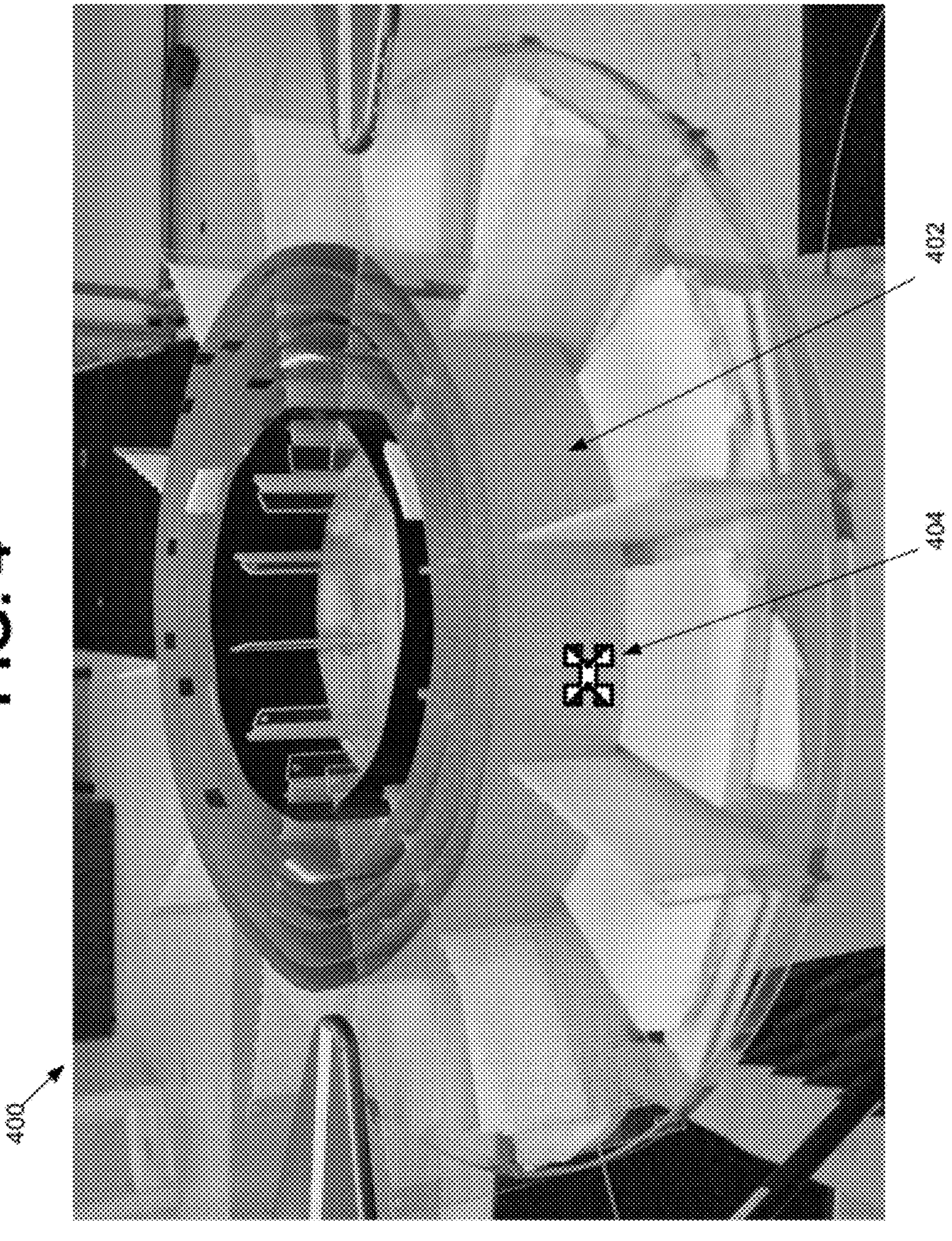

WIDEBAND ELECTROMAGNETIC CLOAKING SYSTEMS

RELATED APPLICATION

This application claims priority to U.S. Provisional Patent Application No. 61/189,966, filed 25 Aug. 2008 and entitled "Method and Apparatus for Wideband Electromagnetic Cloaking, Negative Refractive Index Lensing and Metamaterial Applications," the entire contents of which are incorporated herein by reference.

BACKGROUND

Much time and effort has been devoted to the quest for so-called invisibility machines. Beyond science fiction, however, there has been little if any real progress toward this goal.

Materials with negative permittivity and permeability leading to negative index of refraction were theorized by Russian noted physicist Victor Veselago in his seminal paper in *Soviet Physics* USPEKHI, 10, 509 (1968). Since that time, metamaterials have been developed that produce negative index of refraction, subject to various constraints. Such materials are artificially engineered micro/nanostructures that, at given frequencies, show negative permeability and permittivity. Metamaterials have been shown to produce narrow band, e.g., typically less than 5%, response such as bent-back lensing. Such metamaterials produce such a negative-index effect by utilizing a closely-spaced periodic lattice of resonators, such as split-ring resonators, that all resonate. Previous metamaterials provide a negative index of refraction when a sub-wavelength spacing is used for the resonators.

In the microwave regime, certain techniques have been developed to utilize radiation-absorbing materials or coatings to reduce the radar cross section of airborne missiles and vehicles. While such absorbing materials can provide an effective reduction in radar cross section, these results are largely limited to small ranges of electromagnetic radiation.

SUMMARY

Embodiments of the present disclosure can provide techniques, including systems and/or methods, for cloaking objects at certain wavelengths/frequencies or over certain wavelength/frequency ranges (bands). Such techniques can provide an effective electromagnetic lens and/or lensing effect for certain wavelengths/frequencies or over certain wavelength/frequency ranges (bands).

The effects produced by such techniques can include cloaking or so-called invisibility of the object(s) at the noted wavelengths or bands. Representative frequencies of operation can include, but are not limited to, those over a range of 500 MHz to 1.3 GHz, though others may of course be realized. Operation at other frequencies, including for example those of visible light, infrared, ultraviolet, and as well as microwave EM radiation, e.g., K, Ka, X-bands, etc. may be realized, e.g., by appropriate scaling of dimensions and selection of shape of the resonator elements.

Exemplary embodiments of the present disclosure can include a novel arrangement of resonators in an aperiodic configuration or lattice. The overall assembly of resonators, as structures, do not all repeat periodically and at least some of the resonators are spaced such that their phase centers are separated by more than a wavelength. The arrangements can include resonators of several different sizes and/or geometries arranged so that each size or geometry ("grouping") corresponds to a moderate or high "Q" (that is moderate or low bandwidth) response that resonates within a specific frequency range, and that arrangement within that specific grouping of akin elements is periodic in the overall structure—even though the structure as a whole is not an entirely periodic arrangement of resonators. The relative spacing and arrangement of groupings (at least one for each specific frequency range) can be defined by self similarity and origin symmetry, where the "origin" arises at the center of a structure (or part of the structure) individually designed to have the wideband metamaterial property.

For exemplary embodiments, fractal resonators can be used for the resonators in such structures because of their control of passbands, and smaller sizes compared to non-fractal based resonators. Their benefit arises from a size standpoint because they can be used to shrink the resonator (s), while control of passbands can reduce or eliminates issues of harmonic passbands that would resonate at frequencies not desired.

It should be understood that other embodiments of wideband electromagnetic resonator or cloaking systems and methods according to the present disclosure will become readily apparent to those skilled in the art from the following detailed description, wherein exemplary embodiments are shown and described by way of illustration. The systems and methods of the present disclosure are capable of other and different embodiments, and details of such are capable of modification in various other respects. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

BRIEF DESCRIPTION OF THE DRAWINGS

Aspects of the disclosure may be more fully understood from the following description when read together with the accompanying drawings, which are to be regarded as illustrative in nature, and not as limiting. The drawings are not necessarily to scale, emphasis instead being placed on the principles of the disclosure. In the drawings:

FIG. **1** depicts a diagrammatic plan view of a resonator cloaking system utilizing a number of cylindrical shells, in accordance with exemplary embodiments of the present disclosure;

FIG. **2** depicts a diagrammatic plan view of a resonator cloaking system utilizing a number of shells having an elliptical cross-section, in accordance with an alternate embodiment of the present disclosure;

FIG. **3** depicts an exemplary embodiment of a portion of shell that includes repeated conductive traces that are configured in a fractal-like shape; and

FIG. **4** depicts a perspective view (photograph) of an exemplary embodiment of the present disclosure.

While certain embodiments depicted in the drawings, one skilled in the art will appreciate that the embodiments depicted are illustrative and that variations of those shown, as well as other embodiments described herein, may be envisioned and practiced within the scope of the present disclosure.

DETAILED DESCRIPTION

The present disclosure is directed to novel arrangements of resonators useful for obscuring or hiding objects at given bands of electromagnetic radiation. Embodiments of the present disclosure can provide techniques, including systems and/or methods, for hiding or obscuring objects at certain wavelengths/frequencies or over certain wavelength/frequency ranges or bands. Such techniques can provide an effective electromagnetic lens and/or lensing effect for certain wavelengths/frequencies or over certain wavelength/frequency ranges or bands. The effects produced by such techniques can include cloaking or so-called invisibility of the object(s) at the noted wavelengths or bands.

Representative frequencies of operation can include, but are not limited to, those over a range of about 500 MHz to about 1.3 GHz, though others may of course be realized. Operation at other frequencies, including for example those of visible light, infrared, ultraviolet, and as well as microwave EM radiation, e.g., K, Ka, X-bands, etc. may be realized, e.g., by appropriate scaling of dimensions and selection of shape of the resonator elements.

Embodiments of the present disclosure include arrangement of resonators or resonant structures in aperiodic configurations or lattices. The overall assembly of resonator structures can include nested or concentric shells, that each include repeated patterns of resonant structures. The resonant structures can be configured as a close-packed arrangement of electrically conductive material. The resonant structures can be located on the surface of a circuit board.

The overall assemblies, as structures, do not all repeat periodically and at least some of the resonators are spaced such that their phase centers are separated by more than a wavelength. The arrangements can include resonators of several different sizes and/or geometries arranged so that each size or geometry ("grouping") corresponds to a moderate or high quality-factor "Q" response (that is, one allowing for a moderate or low bandwidth) that resonates within a specific frequency range, and that arrangement within that specific grouping of like elements is periodic in the overall structure—even though the structure as a whole is not an entirely periodic arrangement of resonators. The relative spacing and arrangement of groupings (at least one for each specific frequency range) can be defined by self similarity and origin symmetry, where the "origin" arises at the center of a structure (or part of the structure) individually designed to have the wideband metamaterial property.

For exemplary embodiments, fractal resonators can be used for the resonators because of their control of passbands, and smaller sizes. A main benefit of such resonators arises from a size standpoint because they can be used to shrink the resonator(s), while control of passbands can reduce/mitigate or eliminate issues of harmonic passbands that would resonate at frequencies not desired.

Exemplary embodiments of a resonator system for use at microwave (or nearby) frequencies can be built from belts of circuit boards festooned with resonators. These belts can function to slip the microwaves around an object located within the belts, so the object is effectively invisible and "see thru" at the microwave frequencies. Belts, or shells, having similar closed-packed arrangements for operation at a first passband can be positioned within a wavelength of one another, e.g., $\frac{1}{10}\lambda$, $\frac{1}{8}\lambda$, $\frac{1}{4}\lambda$, $\frac{1}{2}\lambda$, etc.

An observer can observe an original image or signal, without it being blocked by the cloaked object. Using no power, the fractal cloak can replicates the original signal (that is, the signal before blocking) with great fidelity. Exemplary embodiments can function over a bandwidth from about 500 MHz to approximately 1500 MHz (1.5 GHz), providing 3:1 bandwidth; operation within or near such can frequencies can provide other bandwidths as well, such as 1:1 up to 2:1 and up to about 3:1.

FIG. **1** depicts a diagrammatic plan view of a cloaking system **100** and RF testing set up in accordance with exemplary embodiments of the present disclosure. As shown in FIG. **1**, a number of concentric shells (or bands) **102** are placed on a platform (parallel to the plane of the drawing). The shells include a flexible substrate (e.g., polyimide with or without composite reinforcement) with conductive traces (e.g., copper, silver, etc.) in fractal shapes or outlines. The shells **102** surround an object to be cloaked (shown as **104** in FIG. **1**). A transmitting antenna **1** and a receiving antenna **2** are configured at different sides of the system **100**, for verifying efficacy of the cloaking system **100** and recording results. The shells **102** can be held in place by radial supports **106** (while only four are shown, **12** were used in the exemplary embodiment indicated).

The shells indicated in FIG. **1** are of two types, one set (A**1**-A**4**) configured for optimal operation over a first wavelength/frequency range, and another set (B**1**-B**3**) configured for optimal operation over a second wavelength/frequency range. (The numbering of the shells is of course arbitrary and can be reordered, e.g., reversed.)

For an exemplary embodiment of system **100**, the outer set of shells (A**1**-A**4**, with A**1** being the innermost and A**4** the outmost) had a height of about 3 to 4 inches (e.g., 3.5 inches) and the inner set of shells had a height of about 1 inch less (e.g., about 2.5 to 3 inches). The spacing between the shells with a larger fractal shape (A**1**-A**4**) was about 2.4 cm while the spacing between shells of smaller fractal generator shapes (B**1**-B**3**) was about 2.15 cm (along a radial direction). In a preferred embodiment, shell A**4** was placed between shell B**2** and B**3** as shown. The resonators formed on each shell by the fractal shapes can be configured so as to be closely coupled (e.g., by capacitive coupling) and can serve to propagate a plasmonic wave.

It will be appreciated that while, two types of shells and a given number of shells per set are indicated in FIG. **1**, the number of shell types and number of shells for each set can be selected as desired, and may be optimized for different applications, e.g., wavelength/frequency bands.

FIG. **2** depicts a diagrammatic plan view of a cloaking system (or electrical resonator system) according to an alternate embodiment in which the individual shells have an elliptical cross section. As shown in FIG. **2**, a system **200** for cloaking can include a number of concentric shells (or bands) **202**. These shells can be held in place with respect to one another by suitable fixing means, e.g., they can be placed on a platform (parallel to the plane of the drawing) and/or held with a frame. The shells **202** can include a flexible substrate (e.g., polyimide with or without composite reinforcement) with a close-packed arrangement of electrically conductive material formed on the first surface. As stated previously for FIG. **1**, the closed-packed arrangement can include a number of self-similar electrical resonator shapes. The resonator shapes can be made from conductive traces (e.g., copper, silver, gold, silver-based ink, etc.) having a desired shape, e.g., fractal shape, split-ring shape, and the like. The shells **202** can surround an object to be cloaked, as indicated in FIG. **2**.

As indicated in FIG. **2** (by dashed lines **1** and **2** and arrows), the various shells themselves do not have to form closed surfaces. Rather, one or more shells can form open surfaces. This can allow for preferential cloaking of the object in one direction or over a given angle (solid angle). Moreover, while dashed lines **1** and **2** are shown intersecting shells B**1**-B**3** and A**1**-A**3** of system **200**, one or more shells of each group of shells (B**1**-B**3** and A**1**-A**3**) can be closed while others are open.

With further regard to FIGS. **1**-**2**, it should be appreciated that the cross-sections shown for each shell can represent closed geometric shapes, e.g., spherical and ellipsoidal shells.

As indicated previously, each shell of a cloaking system can include multiple resonators. The resonators can be repeated patterns of conductive traces. These conductive traces can be closed geometric shapes, e.g., rings, loops, closed fractals, etc. The resonator(s) can being self similar to at least second iteration. The resonators can include split-ring shapes, for some embodiments. The resonant structures are not required to be closed shapes, however, and open shapes can be used for such.

In exemplary embodiments, the closed loops can be configured as a fractals or fractal-based shapes, e.g., as depicted by **302** in FIG. **3** for an exemplary embodiment of a shell **300**, or **402** in FIG. **4**. The dimensions and type of fractal shape can be the same for each shell type but can vary between shell types. This variation (e.g., scaling of the same fractal shape) can afford increased bandwidth for the cloaking characteristics of the system (e.g., system **100** of FIG. **1**) This can lead to periodicity of the fractal shapes of common shell types but aperiodicity between the fractal shapes of different shell types.

Examples of suitable fractal shapes (for use for shells and/or a scatting object) can include, but are not limited to, fractal shapes described in one or more of the following patents, owned by the assignee of the present disclosure, the entire contents of all of which are incorporated herein by reference: U.S. Pat. Nos. 6,452,553; 6,104,349; 6,140,975; 7,145,513; 7,256,751; 6,127,977; 6,476,766; 7,019,695; 7,215,290; 6,445,352; 7,126,537; 7,190,318; 6,985,122; 7,345,642; and, U.S. Pat. No. 7,456,799.

Other suitable fractal shape for the resonant structures can include any of the following: a Koch fractal, a Minkowski fractal, a Cantor fractal, a torn square fractal, a Mandelbrot, a Caley tree fractal, a monkey's swing fractal, a Sierpinski gasket, and a Julia fractal, a contour set fractal, a Sierpinski triangle fractal, a Menger sponge fractal, a dragon curve fractal, a space-filling curve fractal, a Koch curve fractal, an Iypanov fractal, and a Kleinian group fractal.

FIG. **3** depicts an exemplary embodiment of a shell **300** (only a portion is shown) that includes repeated conductive traces that are configured in a fractal shape **302** (the individual closed traces). For the exemplary embodiment shown, each resonator shape **302** is about 1 cm on a side. Such resonator could, e.g., be used for the fractal shapes of shells B**1**-B**3** of FIG. **1**, in which case similar fractal shapes of larger size (e.g., about 1.5 cm on a side) could be used for shells A**1**-A**4**. The conductive trace is preferably made of copper. While exemplary fractal shapes are shown in FIG. **3**, the present disclosure is not limited to such and any other suitable fractal shapes (including generator motifs) may be used in accordance with the present disclosure.

It will be appreciated that the resonant structures of the shells may be formed or made by any suitable techniques and with any suitable materials. For example, semiconductors with desired doping levels and dopants may be used as conductive materials. Suitable metals or metal containing compounds may be used. Suitable techniques may be used to place conductors on/in a shell, including, but no limited to, printing techniques, photolithography techniques, etching techniques, and the like.

It will also be appreciated that the shells may be made of any suitable material(s). Printed circuit board materials may be used. Flexible circuit board materials are preferred. Other material may, however, be used for the shells and the shells themselves can be made of non-continuous elements, e.g., a frame or framework. For example, various plastics may be used.

FIG. **4** depicts a perspective view (photograph) of an exemplary embodiment of a cloak system **400** according to the present disclosure. As shown, the system includes a number of resonator shells **402** each having a close-packed arrangement of electrically conductive material (self-similar resonators **404**) formed on one surface. Two different shell configurations are shown, four larger shells and two smaller shells. The smaller shells included close-packed arrangements of resonator structures in which each resonator shape (as shown by **302** in FIG. **3**) was about 1 cm on a side. Similar fractal shapes of larger size (e.g., about 1.5 cm on a side) were used for the larger shells.

In FIG. **4**, a transmitting (source) antenna and a receiving antenna are shown as triangular shapes on the left and right, respectively (though functionally of each could of course be interchanged for the other). Twelve radially arrayed spacers are shown in FIG. **4**. The system **400** is shown supported on a Nalgene tank and Delrin platform and Delrin supports (radial supports) RF absorbers were placed in the immediate vicinity of the set up; further RF tripods (e.g., available from ETS) were used; all such materials were substantially transparent at the RF frequencies investigated/used. The cloak system **400** consists of six belts of fractal metamaterial (i.e., fractal-resonant structures shown in FIG. **3**) on flexible Taconic EF35 (low loss) circuit board. The belts are shown surround a scattering ring (object). The arrangement is supported by RF transparent plastics in a comb support. The entire system **400** was shown to be easily built up and broken down within a minute or two. The scale in FIG. **4** is about 0.7 meters across. The height of each shell can of course be selected as desired depending on the situation/application.

While embodiments are shown and described herein as having shells in the shape of concentric rings (circular cylinders), shells can take other shapes in other embodiments. For example, one or more shells could have a generally spherical shape (with minor deviations for structural support). In an exemplary embodiment, the shells could form a nested arrangement of such spherical shapes, around an object to be shielded (at the targeted/selected frequencies/wavelengths). Shell cross-sections of angular shapes, e.g., triangular, hexagonal, while not preferred, may be used.

One skilled in the art will appreciate that embodiments and/or portions of embodiments of the present disclosure can be implemented in/with computer-readable storage media (e.g., hardware, software, firmware, or any combinations of such), and can be distributed and/or practiced over one or more networks. Steps or operations (or portions of such) as described herein, including processing functions to derive, learn, or calculate formula and/or mathematical models utilized and/or produced by the embodiments of the present disclosure, can be processed by one or more suitable processors, e.g., central processing units ("CPUs") implementing suitable code/instructions in any suitable language (machine dependent on machine independent).

While certain embodiments and/or aspects have been described herein, it will be understood by one skilled in the art that the methods, systems, and apparatus of the present disclosure may be embodied in other specific forms without departing from the spirit thereof.

For example, while certain wavelengths/frequencies of operation have been described, these are merely representative and other wavelength/frequencies may be utilized or achieved within the scope of the present disclosure.

Furthermore, while certain preferred fractal generator shapes have been described others may be used within the scope of the present disclosure. Accordingly, the embodiments described herein are to be considered in all respects as illustrative of the present disclosure and not restrictive.

What is claimed is:

1. An electrical resonator system, comprising:

a plurality of concentric electrical resonator shells, each shell including a substrate having first and second surfaces and a close-packed arrangement of electrically conductive material formed on the first surface, wherein the closed-packed arrangement comprises a plurality of self-similar electrical resonator shapes and is configured to operate at a desired passband of electromagnetic radiation;

wherein the close-packed arrangements of at least two of the electrical resonator shells are different in size and/or shape; and

wherein a resonator in the close-packed arrangement comprises a second order or higher fractal.

2. The system of claim **1**, wherein said passband is about 2:1.

3. The system of claim **2**, wherein said passband is about 3:1.

4. The system of claim **1**, wherein the electrical system is configured and arranged so that radiation incident on the system from a given direction has an intensity on a point-by-point basis such at each respective antipodal point, relative to an object placed at the center of the system, the radiation has the same or similar intensity.

5. The system of claim **1**, wherein the system is configured and arranged so that radiation incident on the system from a direction in cylindrical coordinates has the same or similar intensity at the antipodal point after having traversed the system.

6. The system of claim **1**, wherein the plurality of shells comprises a first pair of shells having similar closed-packed arrangements for operation at a first passband, wherein the two shells are positioned within ⅛λ of one another.

7. The system of claim **6**, wherein the plurality of shells comprises a second pair of shells having similar closed-packed arrangements for operation at a second frequency band, wherein the two shells are positioned within ⅛λ of one another.

8. The system of claim **1**, wherein the plurality of shells are hemispherical.

9. The system of claim **1**, wherein the plurality of shells are cylindrical.

10. The system of claim **1**, wherein the plurality of shells are spherical.

11. The system of claim **1**, wherein said fractal is selected from the group consisting of a Koch fractal, a Minkowski fractal, a Cantor fractal, a torn square fractal, a Mandelbrot, a Caley tree fractal, a monkey's swing fractal, a Sierpinski gasket, and a Julia fractal.

12. The system of claim **1**, wherein the fractal is selected from the group consisting of a contour set fractal, a Sierpinski triangle fractal, a Menger sponge fractal, a dragon curve fractal, a space-filling curve fractal, a Koch curve fractal, a Lypanov fractal, and a Kleinian group fractal.

13. The system of claim **1**, wherein the plurality of concentric electrical resonator shells are configured and arranged for operation at K band, Ka band, or X-band.

14. The system of claim **1**, wherein the system is operational over a bandwidth from about 500 MHz to about 1500 MHz.

* * * * *

United States Patent [19]
Eastlund

[11] **Patent Number:** **4,686,605**
[45] **Date of Patent:** **Aug. 11, 1987**

[54] METHOD AND APPARATUS FOR ALTERING A REGION IN THE EARTH'S ATMOSPHERE, IONOSPHERE, AND/OR MAGNETOSPHERE

[75] Inventor: **Bernard J. Eastlund,** Spring, Tex.

[73] Assignee: **APTI, Inc.,** Los Angeles, Calif.

[21] Appl. No.: **690,333**

[22] Filed: **Jan. 10, 1985**

[51] **Int. Cl.**^{4} **H05B 6/64;** H05C 3/00; H05H 1/46

[52] **U.S. Cl.** .... **361/231;** 89/1.11; 380/59; 244/158 R

[58] **Field of Search** .... 361/230, 231; 244/158 R; 376/100; 89/1.11; 380/59

[56] **References Cited**

PUBLICATIONS

Liberty Magazine, (2/35) p. 7 N. Tesla.

New York Times (9/22/40) Section 2, p. 7 W. L. Laurence.

New York Times (12/8/15) p. 8 Col. 3.

Primary Examiner—Salvatore Cangialosi
Attorney, Agent, or Firm—Roderick W. MacDonald

[57] **ABSTRACT**

A method and apparatus for altering at least one selected region which normally exists above the earth's surface. The region is excited by electron cyclotron resonance heating to thereby increase its charged particle density. In one embodiment, circularly polarized electromagnetic radiation is transmitted upward in a direction substantially parallel to and along a field line which extends through the region of plasma to be altered. The radiation is transmitted at a frequency which excites electron cyclotron resonance to heat and accelerate the charged particles. This increase in energy can cause ionization of neutral particles which are then absorbed as part of the region thereby increasing the charged particle density of the region.

15 Claims, 5 Drawing Figures

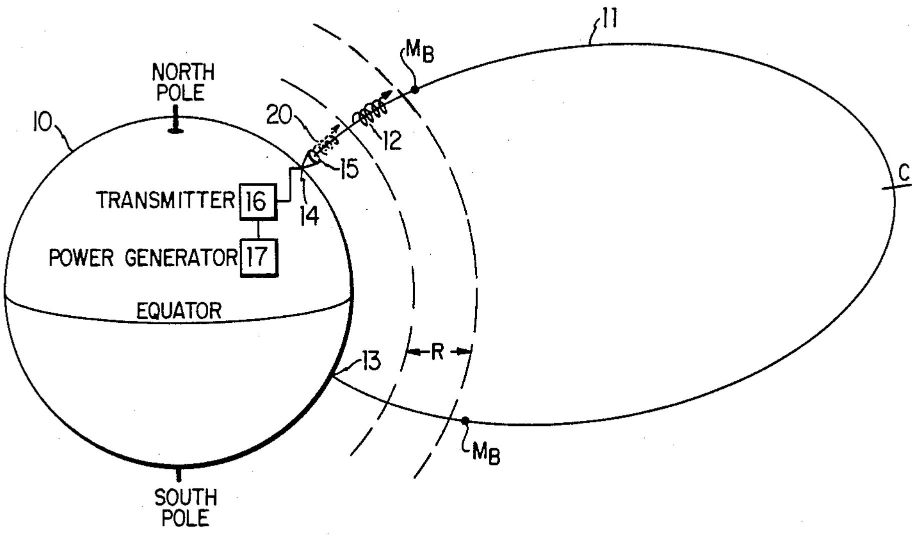

FIG. 1

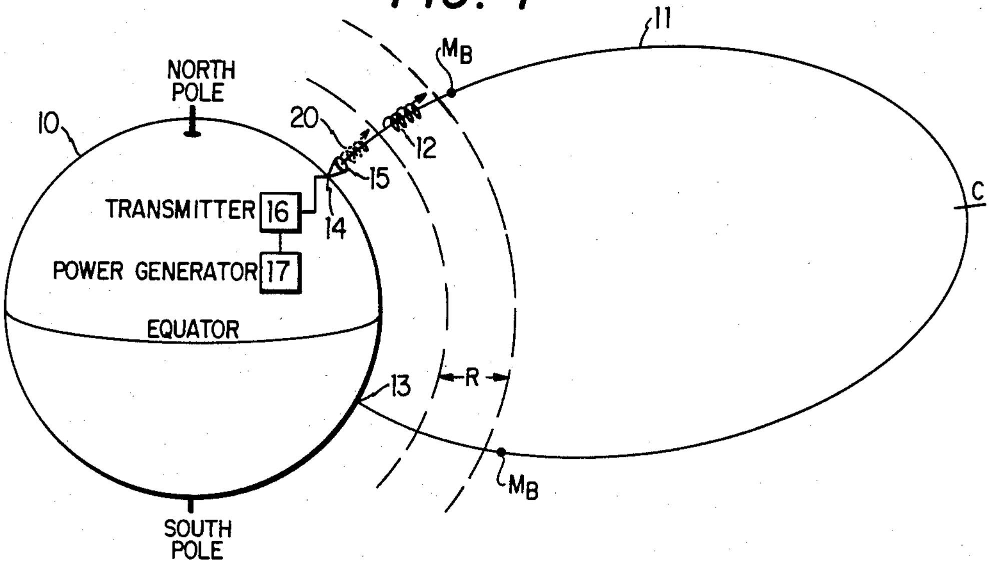
11
MB
NORTH POLE
10
20
12
15
TRANSMITTER 16
14
POWER GENERATOR 17
C
EQUATOR
13
R
MB
SOUTH POLE

FIG. 2

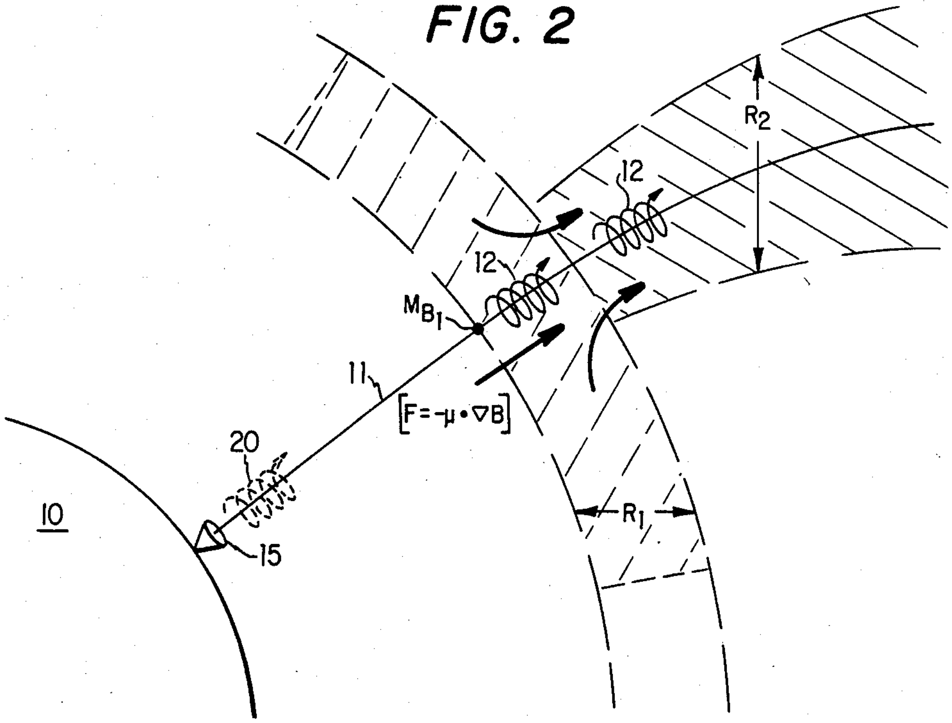
R2
MB1
[F=-μ•∇B]
R1

FIG. 3

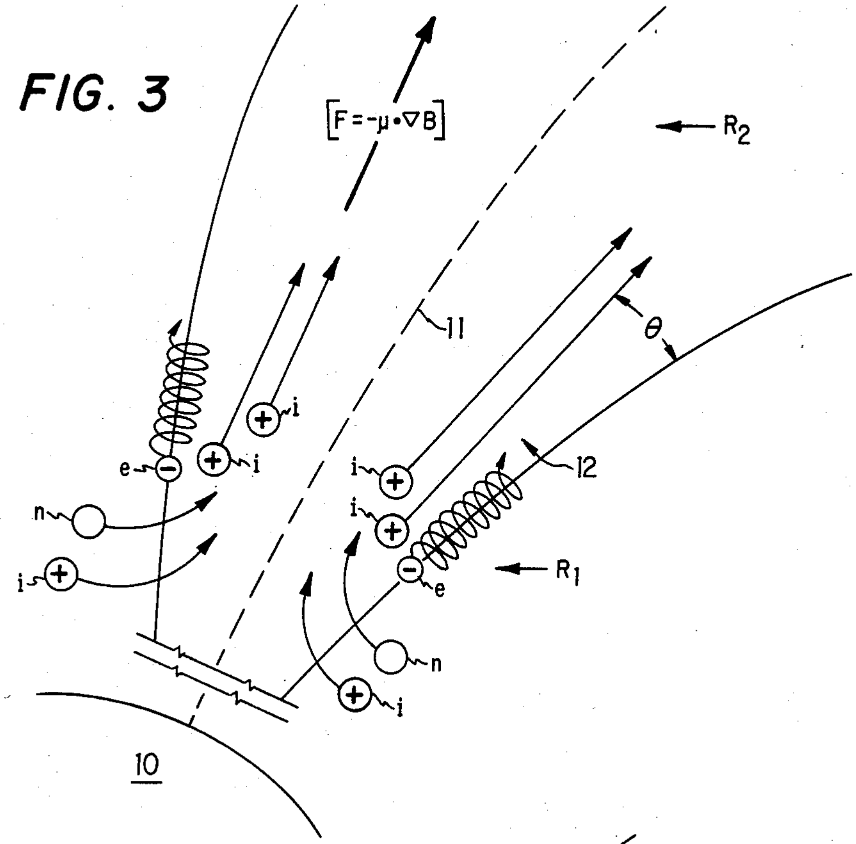
[F = -μ•∇B]
R2
11
θ
12
R1
e
i
n
10

FIG. 4

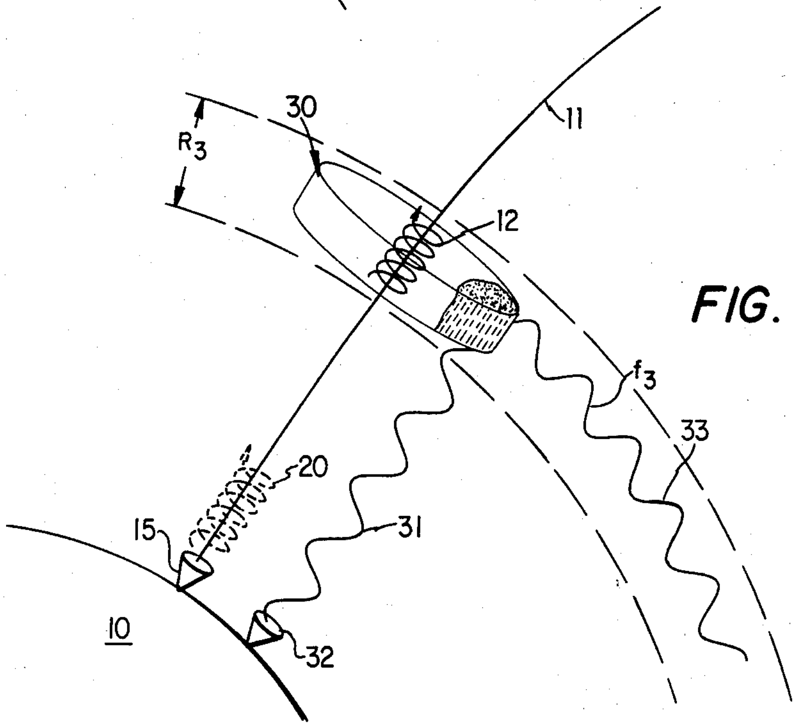
R3
30
11
12
f3
33
20
31
15
32
10

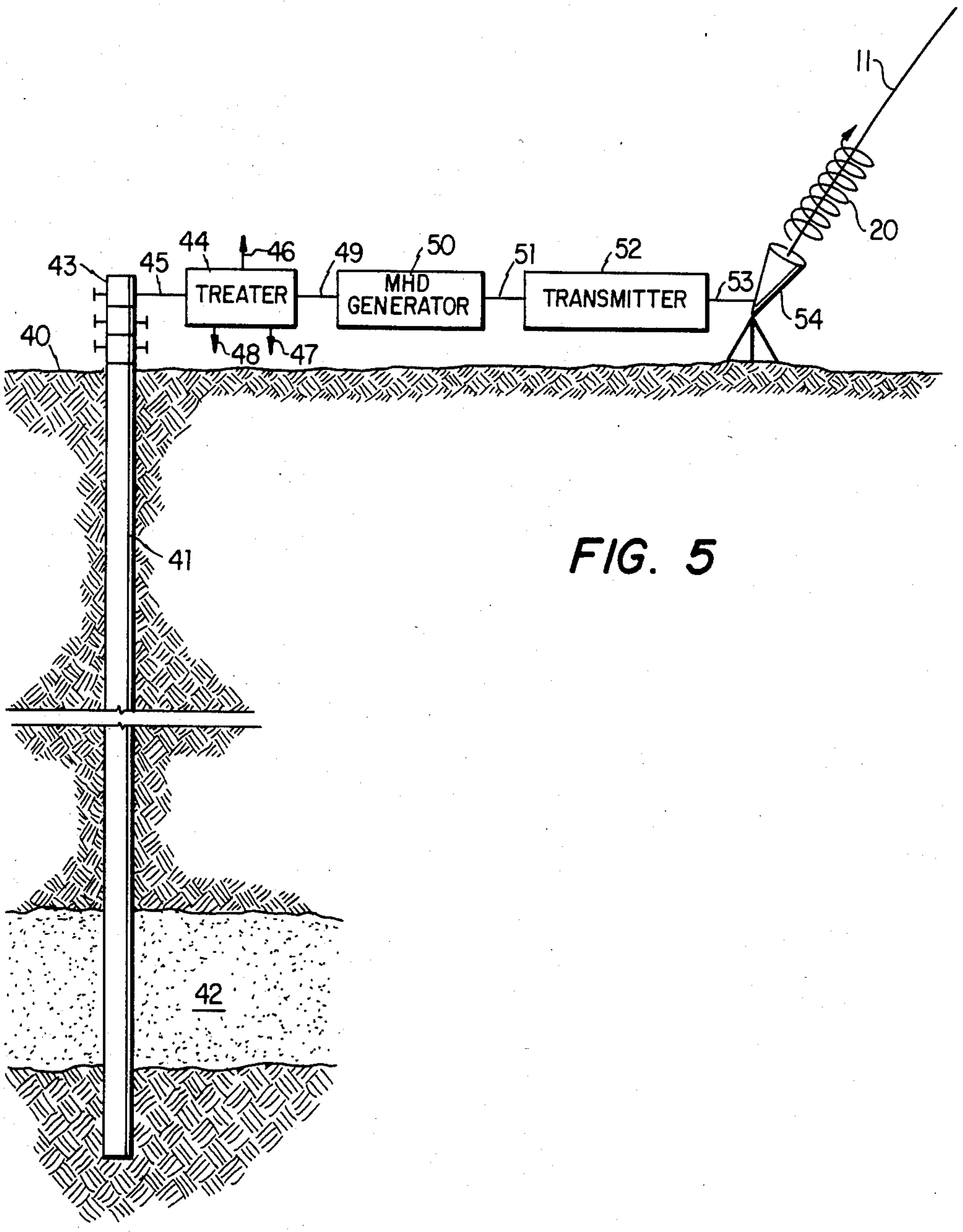
TREATER
MHD GENERATOR
TRANSMITTER

FIG. 5

METHOD AND APPARATUS FOR ALTERING A REGION IN THE EARTH'S ATMOSPHERE, IONOSPHERE, AND/OR MAGNETOSPHERE

DESCRIPTION

1. Technical Field

This invention relates to a method and apparatus for altering at least one selected region normally existing above the earth's surface and more particularly relates to a method and apparatus for altering said at least one region by initially transmitting electromagnetic radiation from the earth's surface essentially parallel to and along naturally-occurring, divergent magnetic field lines which extend from the earth's surface through the region or regions to be altered.

2. Background Art

In the late 1950's, it was discovered that naturally-occuring belts exist at high altitudes above the earth's surface, and it is now established that these belts result from charged electrons and ions becoming trapped along the magnetic lines of force (field lines) of the earth's essentially dipole magnetic field. The trapped electrons and ions are confined along the field lines between two magnetic mirrors which exist at spaced apart points along those field lines. The trapped electrons and ions move in helical paths around their particular field lines and "bounce" back and forth between the magnetic mirrors. These trapped electrons and ions can oscillate along the field lines for long periods of time.

In the past several years, substantial effort has been made to understand and explain the phenomena involved in belts of trapped electrons and ions, and to explore possible ways to control and use these phenomena for beneficial purposes. For example, in the late 1950's and early 1960's both the United States and U.S.S.R. detonated a series of nuclear devices of various yields to generate large numbers of charged particles at various altitudes, e.g., 200 kilometers (km) or greater. This was done in order to establish and study artifical belts of trapped electrons and ions. These experiments established that at least some of the extraneous electrons and ions from the detonated devices did become trapped along field lines in the earth's magnetosphere to form artificial belts which were stable for prolonged periods of time. For a discussion of these experiments see "The Radiation Belt and Magnetosphere", W. N. Hess, Blaisdell Publishing Co., 1968, pps. 155 et sec.

Other proposals which have been advanced for altering existing belts of trapped electrons and ions and/or establishing similar artificial belts include injecting charged particles from a satellite carrying a payload of radioactive beta-decay material or alpha emitters; and injecting charged particles from a satellite-borne electron accelerator. Still another approach is described in U.S. Pat. No. 4,042,196 wherein a low energy ionized gas, e.g., hydrogen, is released from a synchronous orbiting satellite near the apex of a radiation belt which is naturally-occurring in the earth's magnetosphere to produce a substantial increase in energetic particle precipitation and, under certain conditions, produce a limit in the number of particles that can be stably trapped. This precipitation effect arises from an enhancement of the whistler-mode and ion-cyclotron mode interactions that result from the ionized gas or "cold plasma" injection.

It has also been proposed to release large clouds of barium in the magnetosphere so that photoionization will increase the cold plasma density, thereby producing electron precipitation through enhanced whistler-mode interactions.

However, in all of the above-mentioned approaches, the mechanisms involved in triggering the change in the trapped particle phenomena must be actually positioned within the affected zone, e.g., the magnetosphere, before they can be actuated to effect the desired change.

The earth's ionosphere is not considered to be a "trapped" belt since there are few trapped particles therein. The term "trapped" herein refers to situations where the force of gravity on the trapped particles is balanced by magnetic forces rather than hydrostatic or collisional forces. The charged electrons and ions in the ionosphere also follow helical paths around magnetic field lines within the ionosphere but are not trapped between mirrors, as in the case of the trapped belts in the magnetosphere, since the gravitational force on the particles is balanced by collisional or hydrostatic forces.

In recent years, a number of experiments have actually been carried out to modify the ionosphere in some controlled manner to investigate the possibility of a beneficial result. For detailed discussions of these operations see the following papers: (1) Ionospheric Modification Theory; G. Meltz and F. W. Perkins; (2) The Platteville High Power Facility; Carrol et al.; (3) Arecibo Heating Experiments; W. E. Gordon and H. C. Carlson, Jr.; and (4) Ionospheric Heating by Powerful Radio Waves; Meltz et al., all published in Radio Science, Vol. 9, No. 11, November, 1974, at pages 885–888; 889–894; 1041–1047; and 1049–1063, respectively, all of which are incorporated herein by reference. In such experiments, certain regions of the ionosphere are heated to change the electron density and temperature within these regions. This is accomplished by transmitting from earth-based antennae high frequency electromagnetic radiation at a substantial angle to, not parallel to, the ionosphere's magnetic field to heat the ionospheric particles primarily by ohmic heating. The electron temperature of the ionosphere has been raised by hundreds of degrees in these experiments, and electrons with several electron volts of energy have been produced in numbers sufficient to enhance airglow. Electron concentrations have been reduced by a few percent, due to expansion of the plasma as a result of increased temperature.

In the Elmo Bumpy Torus (EBT), a controlled fusion device at the Oak Ridge National Laboratory, all heating is provided by microwaves at the electron cyclotron resonance interaction. A ring of hot electrons is formed at the earth's surface in the magnetic mirror by a combination of electron cyclotron resonance and stochastic heating. In the EBT, the ring electrons are produced with an average "temperature" of 250 kilo electron volts or kev (2.5×10^9K) and a plasma beta between 0.1 and 0.4; see, "A Theoretical Study of Electron—Cyclotron Absorption in Elmo Bumpy Torus", Batchelor and Goldfinger, Nuclear Fusion, Vol. 20, No. 4 (1980) pps. 403–418.

Electron cyclotron resonance heating has been used in experiments on the earth's surface to produce and accelerate plasmas in a diverging magnetic field. Kosmahl et al. showed that power was transferred from the electromagnetic waves and that a fully ionized plasma was accelerated with a divergence angle of roughly 13 degrees. Optimum neutral gas density was 1.7×10^{14} per cubic centimeter; see, "Plasma Acceleration with Microwaves Near Cyclotron Resonance", Kosmahl et al., Journal of Applied Physics, Vol. 38, No. 12, Nov., 1967, pps. 4576–4582.

DISCLOSURE OF THE INVENTION

The present invention provides a method and apparatus for altering at least one selected region which normally exists above the earth's surface. The region is excited by electron cyclotron resonance heating of electrons which are already present and/or artifically created in the region to thereby increase the charged particle energy and ultimately the density of the region.

In one embodiment this is done by transmitting circularly polarized electromagnetic radiation from the earth's surface at or near the location where a naturally-occurring dipole magnetic field (force) line intersects the earth's surface. Right hand circular polarization is used in the northern hemisphere and left hand circular polarization is used in the southern hemisphere. The radiation is deliberately transmitted at the outset in a direction substantially parallel to and along a field line which extends upwardly through the region to be altered. The radiation is transmitted at a frequency which is based on the gyrofrequency of the charged particles and which, when applied to the at least one region, excites electron cyclotron resonance within the region or regions to heat and accelerate the charged particles in their respective helical paths around and along the field line. Sufficient energy is employed to cause ionization of neutral particles (molecules of oxygen, nitrogen and the like, particulates, etc.) which then become a part of the region thereby increasing the charged particle density of the region. This effect can further be enhanced by providing artificial particles, e.g., electrons, ions, etc., directly into the region to be affected from a rocket, satellite, or the like to supplement the particles in the naturally-occurring plasma. These artificial particles are also ionized by the transmitted electromagnetic radiation thereby increasing charged particle density of the resulting plasma in the region.

In another embodiment of the invention, electron cyclotron resonance heating is carried out in the selected region or regions at sufficient power levels to allow a plasma present in the region to generate a mirror force which forces the charged electrons of the altered plasma upward along the force line to an altitude which is higher than the original altitude. In this case the relevant mirror points are at the base of the altered region or regions. The charged electrons drag ions with them as well as other particles that may be present. Sufficient power, e.g., 10^{15} joules, can be applied so that the altered plasma can be trapped on the field line between mirror points and will oscillate in space for prolonged periods of time. By this embodiment, a plume of altered plasma can be established at selected locations for communication modification or other purposes.

In another embodiment, this invention is used to alter at least one selected region of plasma in the ionosphere to establish a defined layer of plasma having an increased charged particle density. Once this layer is established, and while maintaining the transmission of the main beam of circularly polarized electromagnetic radiation, the main beam is modulated and/or at least one second different, modulated electromagnetic radiation beam is transmitted from at least one separate source at a different frequency which will be absorbed in the plasma layer. The amplitude of the frequency of the main beam and/or the second beam or beams is modulated in resonance with at least one known oscillation mode in the selected region or regions to excite the known oscillation mode to propagate a known frequency wave or waves throughout the ionosphere.

BRIEF DESCRIPTION OF THE DRAWINGS

The actual construction, operation, and apparent advantages of this invention will be better understood by referring to the drawings in which like numerals identify like parts and in which:

FIG. **1** is a simplified schematical view of the earth (not to scale) with a magnetic field (force) line along which the present invention is carried out;

FIG. **2** is one embodiment within the present invention in which a selected region of plasma is raised to a higher altitude;

FIG. **3** is a simplified, idealized representation of a physical phenomenon involved in the present invention; and

FIG. **4** is a schematic view of another embodiment within the present invention.

FIG. **5** is a schematic view of an apparatus embodiment within this invention.

BEST MODES FOR CARRYING OUT THE INVENTION

The earth's magnetic field is somewhat analogous to a dipole bar magnet. As such, the earth's magnetic field contains numerous divergent field or force lines, each line intersecting the earth's surface at points on opposite sides of the Equator. The field lines which intersect the earth's surface near the poles have apexes which lie at the furthest points in the earth's magnetosphere while those closest to the Equator have apexes which reach only the lower portion of the magnetosphere.

At various altitudes above the earth's surface, e.g., in both the ionosphere and the magnetosphere, plasma is naturally present along these field lines. This plasma consists of equal numbers of positively and negatively charged particles (i.e., electrons and ions) which are guided by the field line. It is well established that a charged particle in a magnetic field gyrates about field lines, the center of gyration at any instance being called the "guiding center" of the particle. As the gyrating particle moves along a field line in a uniform field, it will follow a helical path about its guiding center, hence linear motion, and will remain on the field line. Electrons and ions both follow helical paths around a field line but rotate in opposite directions. The frequencies at which the electrons and ions rotate about the field line are called gyromagnetic frequencies or cyclotron frequencies because they are identical with the expression for the angular frequencies of gyration of particles in a cyclotron. The cyclotron frequency of ions in a given magnetic field is less than that of electrons, in inverse proportion to their masses.

If the particles which form the plasma along the earth's field lines continued to move with a constant pitch angle, often designated "alpha", they would soon impact on the earth's surface. Pitch angle alpha is defined as the angle between the direction of the earth's magnetic field and the velocity (V) of the particle. However, in converging force fields, the pitch angle does change in such a way as to allow the particle to turn around and avoid impact. Consider a particle moving along a field line down toward the earth. It moves into a region of increasing magnetic field strength and therefore sine alpha increases. But sine alpha can only increase to 1.0, at which point, the particle turns around and starts moving up along the field line, and alpha decreases. The point at which the particle turns around is called the mirror point, and there alpha equals ninety degrees. This process is repeated at the other end of the field line where the same magnetic field strength value B, namely Bm, exists. The particle again turns around and this is called the "conjugate point" of the original mirror point. The particle is therefore trapped and bounces between the two magnetic mirrors. The particle can continue oscillating in space in this manner for long periods of time. The actual place where a particle will mirror can be calculated from the following:

$$\sin^2 \mathrm{alpha}_o = B_o/B_m \qquad (1)$$

wherein:

alpha_o=equatorial pitch angle of particle

B_o=equatorial field strength on a particular field line

B_m=field strength at the mirror point

Recent discoveries have established that there are substantial regions of naturally trapped particles in space which are commonly called "trapped radiation belts". These belts occur at altitudes greater than about 500 km and accordingly lie in the magnetosphere and mostly above the ionosphere.

The ionosphere, while it may overlap some of the trapped-particle belts, is a region in which hydrostatic forces govern its particle distribution in the gravitational field. Particle motion within the ionosphere is governed by both hydrodynamic and electrodynamic forces. While there are few trapped particles in the ionosphere, nevertheless, plasma is present along field lines in the ionosphere. The charged particles which form this plasma move between collisions with other particles along similar helical paths around the field lines and although a particular particle may diffuse downward into the earth's lower atmosphere or lose energy and diverge from its original field line due to collisions with other particles, these charged particles are normally replaced by other available charged particles or by particles that are ionized by collision with said particle. The electron density (N_e) of the plasma will vary with the actual conditions and locations involved. Also, neutral particles, ions, and electrons are present in proximity to the field lines.

The production of enhanced ionization will also alter the distribution of atomic and molecular constituents of the atmosphere, most notably through increased atomic nitrogen concentration. The upper atmosphere is normally rich in atomic oxygen (the dominant atmospheric constituent above 200 km altitude), but atomic nitrogen is normally relatively rare. This can be expected to manifest itself in increased airglow, among other effects.

As known in plasma physics, the characteristics of a plasma can be altered by adding energy to the charged particles or by ionizing or exciting additional particles to increase the density of the plasma. One way to do this is by heating the plasma which can be accomplished in different ways, e.g., ohmic, magnetic compression, shock waves, magnetic pumping, electron cyclotron resonance, and the like.

Since electron cyclotron resonance heating is involved in the present invention, a brief discussion of

same is in order. Increasing the energy of electrons in a plasma by invoking electron cyclotron resonance heating, is based on a principle similar to that utilized to accelerate charged particles in a cyclotron. If a plasma is confined by a static axial magnetic field of strength B, the charged particles will gyrate about the lines of force with a frequency given, in hertz, as $f_g = 1.54 \times 10^3 B/A$, where: B=magnetic field strength in gauss, and A=mass number of the ion.

Suppose a time-varying field of this frequency is superimposed on the static field B confining the plasma, by passage of a radiofrequency current through a coil which is concentric with that producing the axial field, then in each half-cycle of their rotation about the field lines, the charged particles acquire energy from the oscillating electric field associated with the radio frequency. For example, if B is 10,000 gauss, the frequency of the field which is in resonance with protons in a plasma is 15.4 megahertz.

As applied to electrons, electron cyclotron resonance heating requires an oscillating field having a definite frequency determined by the strength of the confining field. The radio-frequency radiation produces time-varying fields (electric and magnetic), and the electric field accelerates the charged particle. The energized electrons share their energy with ions and neutrals by undergoing collisions with these particles, thereby effectively raising the temperature of the electrons, ions, and neutrals. The apportionment of energy among these species is determined by collision frequencies. For a more detailed understanding of the physics involved, see "Controlled Thermonuclear Reactions", Glasstone and Lovberg, D. Van Nostrand Company, Inc., Princeton, N.J., 1960 and "The Radiation Belt and Magnetosphere", Hess, Blaisdell Publishing Company, 1968, both of which are incorporated herein by reference.

Referring now to the drawings, the present invention provides a method and apparatus for altering at least one region of plasma which lies along a field line, particularly when it passes through the ionosphere and/or magnetosphere. FIG. **1** is a simplified illustration of the earth **10** and one of its dipole magnetic force or field lines **11**. As will be understood, line **11** may be any one of the numerous naturally existing field lines and the actual geographical locations **13** and **14** of line **11** will be chosen based on a particular operation to be carried out. The actual locations at which field lines intersect the earth's surface is documented and is readily ascertainable by those skilled in the art.

Line **11** passes through region R which lies at an altitude above the earth's surface. A wide range of altitudes are useful given the power that can be employed by the practice of this invention. The electron cyclotron resonance heating effect can be made to act on electrons anywhere above the surface of the earth. These electrons may be already present in the atmosphere, ionosphere, and/or magnetosphere of the earth, or can be artificially generated by a variety of means such as x-ray beams, charged particle beams, lasers, the plasma sheath surrounding an object such as a missile or meteor, and the like. Further, artificial particles, e.g., electrons, ions, etc., can be injected directly into region R from an earth-launched rocket or orbiting satellite carrying, for example, a payload of radioactive beta-decay material; alpha emitters; an electron accelerator; and/or ionized gases such as hydrogen; see U.S. Pat. No. 4,042,196. The altitude can be greater than about 50 km if desired, e.g., can be from about 50 km to about 800 km, and, accordingly may lie in either the ionosphere or the magnetosphere or both. As explained above, plasma will be present along line **11** within region R and is represented by the helical line **12**. Plasma **12** is comprised of charged particles (i.e., electrons and ions) which rotate about opposing helical paths along line **11**.

Antenna **15** is positioned as close as is practical to the location **14** where line **11** intersects the earth's surface. Antenna **15** may be of any known construction for high directionality, for example, a phased array, beam spread angle (θ) type. See "The MST Radar at Poker Flat, Alaska", Radio Science, Vol. 15, No. 2, Mar.-Apr. 1980, pps. 213–223, which is incorporated herein by reference. Antenna **15** is coupled to transmitter **16** which generates a beam of high frequency electromagnetic radiation at a wide range of discrete frequencies, e.g., from about 20 to about 1800 kilohertz (kHz).

Transmitter **16** is powered by power generator means **17** which is preferably comprised of one or more large, commercial electrical generators. Some embodiments of the present invention require large amounts of power, e.g., up to 10^9 to 10^{11} watts, in continuous wave or pulsed power. Generation of the needed power is within the state of the art. Although the electrical generators necessary for the practice of the invention can be powered in any known manner, for example, by nuclear reactors, hydroelectric facilities, hydrocarbon fuels, and the like, this invention, because of its very large power requirement in certain applications, is particularly adapted for use with certain types of fuel sources which naturally occur at strategic geographical locations around the earth. For example, large reserves of hydrocarbons (oil and natural gas) exist in Alaska and Canada. In northern Alaska, particularly the North Slope region, large reserves are currently readily available. Alaska and northern Canada also are ideally located geographically as to magnetic latitudes. Alaska provides easy access to magnetic field lines that are especially suited to the practice of this invention, since many field lines which extend to desirable altitudes for this invention intersect the earth in Alaska. Thus, in Alaska, there is a unique combination of large, accessible fuel sources at desirable field line intersections. Further, a particularly desirable fuel source for the generation of very large amounts of electricity is present in Alaska in abundance, this source being natural gas. The presence of very large amounts of clean-burning natural gas in Alaskan latitudes, particularly on the North Slope, and the availability of magnetohydrodynamic (MHD), gas turbine, fuel cell, electrogasdynamic (EGD) electric generators which operate very efficiently with natural gas provide an ideal power source for the unprecedented power requirements of certain of the applications of this invention. For a more detailed discussion of the various means for generating electricity from hydrocarbon fuels, see "Electrical Aspects of Combustion", Lawton and Weinberg, Clarendon Press, 1969. For example, it is possible to generate the electricity directly at the high frequency needed to drive the antenna system. To do this, typically the velocity of flow of the combustion gases (v), past magnetic field perturbation of dimension d (in the case of MHD), follow the rule:

$$v=df$$

where f is the frequency at which electricity is generated. Thus, if $v=1.78\times10^6$ cm/sec and d=1 cm then electricity would be generated at a frequency of 1.78 mHz.

Put another way, in Alaska, the right type of fuel (natural gas) is naturally present in large amounts and at just the right magnetic latitudes for the most efficient practice of this invention, a truly unique combination of circumstances. Desirable magnetic latitudes for the practice of this invention interest the earth's surface both northerly and southerly of the equator, particularly desirable latitudes being those, both northerly and southerly, which correspond in magnitude with the magnetic latitudes that encompass Alaska.

Referring now to FIG. 2 a first ambodiment is illustrated where a selected region R_1 of plasma **12** is altered by electron cyclotron resonance heating to accelerate the electrons of plasma **12**, which are following helical paths along field line **11**.

To accomplish this result, electromagnetic radiation is transmitted at the outset, essentially parallel to line **11** via antenna **15** as right hand circularly polarized radiation wave **20**. Wave **20** has a frequency which will excite electron cyclotron resonance with plasma **12** at its initial or original altitude. This frequency will vary depending on the electron cyclotron resonance of region R_1 which, in turn, can be determined from available data based on the altitudes of region R_1, the particular field line **11** being used, the strength of the earth's magnetic field, etc. Frequencies of from about 20 to about 7200 kHz, preferably from about 20 to about 1800 kHz can be employed. Also, for any given application, there will be a threshhold (minimum power level) which is needed to produce the desired result. The minimum power level is a function of the level of plasma production and movement required, taking into consideration any loss processes that may be dominant in a particular plasma or propagation path.

As electron cyclotron resonance is established in plasma **12**, energy is transferred from the electromagnetic radiation **20** into plasma **12** to heat and accelerate the electrons therein and, subsequently, ions and neutral particles. As this process continues, neutral particles which are present within R_1 are ionized and absorbed into plasma **12** and this increases the electron and ion densities of plasma **12**. As the electron energy is raised to values of about 1 kilo electron volt (kev), the generated mirror force (explained below) will direct the excited plasma **12** upward along line **11** to form a plume R_2 at an altitude higher than that of R_1.

Plasma acceleration results from the force on an electron produced by a nonuniform static magnetic field ($\overline{B}$). The force, called the mirror force, is given by

$$F=-\mu\nabla B \quad (2)$$

where μ is the electron magnetic moment and $\nabla\overline{B}$ is the gradient of the magnetic field, μ being further defined as:

$$W_\perp/B=mV_\perp^2/2B$$

where $W_\perp$ is the kinetic energy in the direction perpendicular to that of the magnetic field lines and B is the magnetic field strength at the line of force on which the guiding center of the particle is located. The force as represented by equation (2) is the force which is responsible for a particle obeying equation (1).

Since the magnetic field is divergent in region R_1, it can be shown that the plasma will move upwardly from the heating region as shown in FIG. **1** and further it can be shown that

$$\tfrac{1}{2}M_eV_{e\perp}{}^2(x) \approx \tfrac{1}{2}M_eV_{e\perp}{}^2(Y) + \tfrac{1}{2}M_iV_{i\parallel}{}^2(Y) \qquad (3)$$

where the left hand side is the initial electron transverse kinetic energy; the first term on the right is the transverse electron kinetic energy at some point (Y) in the expanded field region, while the final term is the ion kinetic energy parallel to B at point (Y). This last term is what constitutes the desired ion flow. It is produced by an electrostatic field set up by electrons which are accelerated according to Equation (2) in the divergent field region and pulls ions along with them. Equation (3) ignores electron kinetic energy parallel to B because $V_{e\parallel} \approx V_{i\parallel}$, so the bulk of parallel kinetic energy resides in the ions because of their greater masses. For example, if an electromagnetic energy flux of from about 1 to about 10 watts per square centimeter is applied to region R, whose altitude is 115 km, a plasma having a density (N_e) of 10^{12} per cubic centimeter will be generated and moved upward to region R_2 which has an altitude of about 1000 km. The movement of electrons in the plasma is due to the mirror force while the ions are moved by ambipolar diffusion (which results from the electrostatic field). This effectively "lifts" a layer of plasma **12** from the ionosphere and/or magnetosphere to a higher elevation R_2. The total energy required to create a plasma with a base area of 3 square kilometers and a height of 1000 km is about 3×10^{13} joules.

FIG. **3** is an idealized representation of movement of plasma **12** upon excitation by electron cyclotron resonance within the earth's divergent force field. Electrons (e) are accelerated to velocities required to generate the necessary mirror force to cause their upward movement. At the same time neutral particles (n) which are present along line **11** in region $R_1$are ionized and become part of plasma **12**. As electrons (e) move upward along line **11**, they drag ions (i) and neutrals (n) with them but at an angle θ of about 13 degrees to field line **11**. Also, any particulates that may be present in region R_1, will be swept upwardly with the plasma. As the charged particles of plasma **12** move upward, other particles such as neutrals within or below R_1, move in to replace the upwardly moving particles. These neutrals, under some conditions, can drag with them charged particles.

For example, as a plasma moves upward, other particles at the same altitude as the plasma move horizontally into the region to replace the rising plasma and to form new plasma. The kinetic energy developed by said other particles as they move horizontally is, for example, on the same order of magnitude as the total zonal kinetic energy of stratospheric winds known to exist.

Referring again to FIG. **2**, plasma **12** in region R_1 is moved upward along field line **11**. The plasma **12** will then form a plume (cross-hatched area in FIG. **2**) which will be relatively stable for prolonged periods of time. The exact period of time will vary widely and be determined by gravitational forces and a combination of radiative and diffusive loss terms. In the previous detailed example, the calculations were based on forming a plume by producing 0+energies of 2 ev/particle. About 10 ev per particle would be required to expand plasma **12** to apex point C (FIG. **1**). There at least some of the particles of plasma **12** will be trapped and will oscillate between mirror points along field line **11**. This oscillation will then allow additional heating of the trapped plasma **12** by stochastic heating which is associated with trapped and oscillating particles. See "A New Mechanism for Accelerating Electrons in the Outer Ionosphere" by R. A. Helliwell and T. F. Bell, Journal of Geophysical Research, Vol. 65, No. 6, June, 1960. This is preferably carried out at an altitude of at least 500 km.

The plasma of the typical example might be employed to modify or disrupt microwave transmissions of satellites. If less than total black-out of transmission is desired (e.g., scrambling by phase shifting digital signals), the density of the plasma (N_e) need only be at least about 10^6 per cubic centimeter for a plasma orginating at an altitude of from about 250 to about 400 km and accordingly less energy (i.e., electromagnetic radiation), e.g., 10^8 joules need be provided. Likewise, if the density $N_e$is on the order of 10^8, a properly positioned plume will provide a reflecting surface for VHF waves and can be used to enhance, interfere with, or otherwise modify communication transmissions. It can be seen from the foregoing that by appropriate application of various aspects of this invention at strategic locations and with adequate power sources, a means and method is provided to cause interference with or even total disruption of communications over a very large portion of the earth. This invention could be employed to disrupt not only land based communications, both civilian and military, but also airborne communications and sea communications (both surface and subsurface). This would have significant military implications, particularly as a barrier to or confusing factor for hostile missiles or airplanes. The belt or belts of enhanced ionization produced by the method and apparatus of this invention, particularly if set up over Northern Alaska and Canada, could be employed as an early warning device, as well as a communications disruption medium. Further, the simple ability to produce such a situation in a practical time period can by itself be a deterring force to hostile action. The ideal combination of suitable field lines intersecting the earth's surface at the point where substantial fuel sources are available for generation of very large quantitities of electromagnetic power, such as the North Slope of Alaska, provides the wherewithal to accomplish the foregoing in a practical time period, e.g., strategic requirements could necessitate achieving the desired altered regions in time periods of two minutes or less and this is achievable with this invention, especially when the combination of natural gas and magnetohydrodynamic, gas turbine, fuel cell and/or EGD electric generators are employed at the point where the useful field lines intersect the earth's surface. One feature of this invention which satisfies a basic requirement of a weapon system, i.e., continuous checking of operability, is that small amounts of power can be generated for operability checking purposes. Further, in the exploitation of this invention, since the main electromagnetic beam which generates the enhanced ionized belt of this invention can be modulated itself and/or one or more additional electromagnetic radiation waves can be impinged on the ionized region formed by this invention as will be described in greater detail herein after with respect to FIG. **4**, a substantial amount of randomly modulated signals of very large power magnitude can be generated in a highly nonlinear mode. This can cause confusion of or interference with or even complete disruption of guidance systems employed by even the most sophisticated of airplanes and missiles. The ability to employ and transmit over very wide areas of the earth a plurality of electromagnetic waves of varying frequencies and to change same at will in a random manner, provides a unique ability to interfere with all modes of communications, land, sea, and/or air, at the same time. Because of the unique juxtaposition of usable fuel source at the point where desirable field lines intersect the earth's surface, such wide ranging and complete communication interference can be achieved in a resonably short period of time. Because of the mirroring phenomenon discussed hereinabove, it can also be prolonged for substantial time periods so that it would not be a mere transient effect that could simply be waited out by an opposing force. Thus, this invention provides the ability to put unprecedented amounts of power in the earth's atmosphere at strategic locations and to maintain the power injection level, particularly if random pulsing is employed, in a manner far more precise and better controlled than heretofore accomplished by the prior art, particularly by the detonation of nuclear devices of various yeilds at various altitudes. Where the prior art approaches yielded merely transitory effects, the unique combination of fuel and desirable field lines at the point where the fuel occurs allows the establishment of, compared to prior art approaches, precisely controlled and long-lasting effects which cannot, practically speaking, simply be waited out. Further, by knowing the frequencies of the various electromagnetic beams employed in the practice of this invention, it is possible not only to interfere with third party communications but to take advantage of one or more such beams to carry out a communications network even though the rest of the world's communications are disrupted. Put another way, what is used to disrupt another's communications can be employed by one knowledgeable of this invention as a communications network at the same time. In addition, once one's own communication network is established, the far-reaching extent of the effects of this invention could be employed to pick up communication signals of other for intelligence purposes. Thus, it can be seen that the disrupting effects achievable by this invention can be employed to benefit by the party who is practicing this invention since knowledge of the various electromagnetic waves being employed and how they will vary in frequency and magnitude can be used to an advantage for positive communication and eavesdropping purposes at the same time. However, this invention is not limited to locations where the fuel source naturally exists or where desirable field lines naturally intersect the earth's surface. For example, fuel, particularly hydrocarbon fuel, can be transported by pipeline and the like to the location where the invention is to be practiced.

FIG. 4 illustrates another embodiment wherein a selected region of plasma R_3which lies within the earth's ionosphere is altered to increase the density thereof whereby a relatively stable layer **30** of relatively dense plasma is maintained within region R_3. Electromagnetic radiation is transmitted at the outset essentially parallel to field line **11** via antenna **15** as a right hand circularly polarized wave and at a frequency (e.g., 1.78 megahertz when the magnetic field at the desired altitude is 0.66 gauss) capable of exciting electron cyclotron resonance in plasma **12** at the particular altitude of plasma **12**. This causes heating of the particles (electrons, ions, neutrals, and particulates) and ionization of the uncharged particles adjacent line **11**, all of which are absorbed into plasma **12** to increase the density thereof. The power transmitted, e.g., 2×10^6 watts for up to 2 minutes heating time, is less than that required to generate the mirror force F required to move plasma **12** upward as in the previous embodiment.

While continuing to transmit electromagnetic radiation **20** from antenna **15**, a second electromagnetic radiation beam **31**, which is at a defined frequency different from the radiation from antenna **15**, is transmitted from one or more second sources via antenna **32** into layer **30** and is absorbed into a portion of layer **30** (cross-hatched area in FIG. 4). The electromagnetic radiation wave from antenna **32** is amplitude modulated to match a known mode of oscillation f_3 in layer **30**. This creates a resonance in layer **30** which excites a new plasma wave **33** which also has a frequency of f_3 and which then propogates through the ionosphere. Wave **33** can be used to improve or disrupt communications or both depending on what is desired in a particular application. Of course, more than one new wave **33** can be generated and the various new waves can be modulated at will and in a highly nonlinear fashion.

FIG. 5 shows apparatus useful in this invention, particularly when those applications of this invention are employed which require extremely large amounts of power. In FIG. 5 there is shown the earth's surface **40** with a well **41** extending downwardly thereinto until it penetrates hydrocarbon producing reservoir **42**. Hydrocarbon reservoir **42** produces natural gas alone or in combination with crude oil. Hydrocarbons are produced from reservoir **42** through well **41** and wellhead **43** to a treating system **44** by way of pipe **45**. In treater **44**, desirable liquids such as crude oil and gas condensates are separated and recovered by way of pipe **46** while undesirable gases and liquids such as water, H_2S, and the like are separated by way of pipe **47**. Desirable gases such as carbon dioxide are separated by way of pipe **48**, and the remaining natural gas stream is removed from treater **44** by way of pipe **49** for storage in conventional tankage means (not shown) for future use and/or use in an electrical generator such as a magnetohydrodynamic, gas turbine, fuel cell or EGD generator **50**. Any desired number and combination of different types of electric generators can be employed in the practice of this invention. The natural gas is burned in generator **50** to produce substantial quantities of electricity which is then stored and/or passed by way of wire **51** to a transmitter **52** which generates the electromagnetic radiation to be used in the method of this invention. The electromagnetic radiation is then passed by way of wire **53** to antenna **54** which is located at or near the end of field line **11**. Antenna **54** sends circularly polarized radiation wave **20** upwards along field line **11** to carry out the various methods of this invention as described hereinabove.

Of course, the fuel source need not be used in its naturally-occurring state but could first be converted to another second energy source form such as hydrogen, hydrazine and the like, and electricity then generated from said second energy source form.

It can be seen from the foregoing that when desirable field line **11** intersects earth's surface **40** at or near a large naturally-occurring hydrocarbon source **42**, exceedingly large amounts of power can be very efficiently produced and transmitted in the direction of field lines. This is particularly so when the fuel source is natural gas and magnetohydrodynamic generators are employed. Further, this can all be accomplished in a relatively small physical area when there is the unique coincidence of fuel source **42** and desirable field line **11**. Of course, only one set of equipment is shown in FIG. **5** for sake of simplicity. For a large hydrocarbon reservoir **42**, a plurality of wells **41** can be employed to feed one or more storage means and/or treaters and as large a number of generators **55** as needed to power one or more transmitters **52** and one or more antennas **54**. Since all of the apparatus **44** through **54** can be employed and used essentially at the sight where naturally-occurring fuel source **42** is located, all the necessary electromagnetic radiation **20** is generated essentially at the same location as fuel source **42**. This provides for a maximum amount of usable electromagnetic radiation **20** since there are no significant storage or transportation losses to be incurred. In other words, the apparatus is brought to the sight of the fuel source where desirable field line **11** intersects the earth's surface **40** on or near the geographical location of fuel source **42**, fuel source **42** being at a desirable magnetic latitude for the practice of this invention, for example, Alaska.

The generation of electricity by motion of a conducting fluid through a magnetic field, i.e., magnetohydrodynamics (MHD), provides a method of electric power generation without moving mechanical parts and when the conducting fluid is a plasma formed by combustion of a fuel such as natural gas, an idealized combination of apparatus is realized since the very clean-burning natural gas forms the conducting plasma in an efficient manner and the thus formed plasma, when passed through a magnetic field, generates electricity in a very efficient manner. Thus, the use of fuel source **42** to generate a plasma by combustion thereof for the generation of electricity essentially at the site of occurrence of the fuel source is unique and ideal when high power levels are required and desirable field lines **11** intersect the earth's surface **40** at or near the site of fuel source **42**. A particular advantage for MHD generators is that they can be made to generate large amounts of power with a small volume, light weight device. For example, a 1000 megawatt MHD generator can be construed using superconducting magnets to weigh roughly 42,000 pounds and can be readily air lifted.

This invention has a phenomenal variety of possible ramifications and potential future developments. As alluded to earlier, missile or aircraft destruction, deflection, or confusion could result, particularly when relativistic particles are employed. Also, large regions of the atmosphere could be lifted to an unexpectedly high altitude so that missiles encounter unexpected and unplanned drag forces with resultant destruction or deflection of same. Weather modification is possible by, for example, altering upper atmosphere wind patterns or altering solar absorption patterns by constructing one or more plumes of atmospheric particles which will act as a lens or focusing device. Also as alluded to earlier, molecular modifications of the atmosphere can take place so that positive environmental effects can be achieved. Besides actually changing the molecular composition of an atmospheric region, a particular molecule or molecules can be chosen for increased presence. For example, ozone, nitrogen, etc. concentrations in the atmosphere could be artificially increased. Similarly, environmental enhancement could be achieved by causing the breakup of various chemical entities such as carbon dioxide, carbon monoxide, nitrous oxides, and the like. Transportation of entities can also be realized when advantage is taken of the drag effects caused by regions of the atmosphere moving up along diverging field lines. Small micron sized particles can be then transported, and, under certain circumstances and with the availability of sufficient energy, larger particles or objects could be similarly affected. Particles with desired characteristics such as tackiness, reflectivity, absorptivity, etc., can be transported for specific purposes or effects. For example, a plume of tacky particles could be established to increase the drag on a missile or satellite passing therethrough. Even plumes of plasma having substantially less charged particle density than described above will produce drag effects on missiles which will affect a lightweight (dummy) missile in a manner substantially different than a heavy (live) missile and this affect can be used to distinguish between the two types of missiles. A moving plume could also serve as a means for supplying a space station or for focusing vast amount of sunlight on selected portions of the earth. Surveys of global scope could also be realized because the earth's natural magnetic field could be significantly altered in a controlled manner by plasma beta effects resulting in, for example, improved magnetotelluric surveys. Electromagnetic pulse defenses are also possible. The earth's magnetic field could be decreased or disrupted at appropriate altitudes to modify or eliminate the magnetic field in high Compton electron generation (e.g., from high altitude nuclear bursts) regions. High intensity, well controlled electrical fields can be provided in selected locations for various purposes. For example, the plasma sheath surrounding a missile or satellite could be used as a trigger for activating such a high intensity field to destroy the missile or satellite. Further, irregularities can be created in the ionosphere which will interfere with the normal operation of various types of radar, e.g., synthetic aperture radar. The present invention can also be used to create artificial belts of trapped particles which in turn can be studied to determine the stability of such parties. Still further, plumes in accordance with the present invention can be formed to simulate and/or perform the same functions as performed by the detonation of a "heave" type nuclear device without actually having to detonate such a device. Thus it can be seen that the ramifications are numerous, far-reaching, and exceedingly varied in usefulness.

I claim:

1. A method for altering at least one region normally existing above the earth's surface with electromagnetic radiation using naturally-occurring and diverging magnetic field lines of the earth comprising transmitting first electromagnetic radiation at a frequency between 20 and 7200 kHz from the earth's surface, said transmitting being conducted essentially at the outset of transmission substantially parallel to and along at least one of said field lines, adjusting the frequency of said first radiation to a value which will excite electron cyclotron resonance at an initial elevation at least 50 km above the earth's surface, whereby in the region in which said electron cyclotron resonance takes place heating, further ionization, and movement of both charged and neutral particles is effected, said cyclotron resonance excitation of said region is continued until the electron concentration of said region reaches a value of at least 10^6 per cubic centimeter and has an ion energy of at least 2 ev.

2. The method of claim **1** including the step of providing artificial particles in said at least one region which are excited by said electron cyclotron resonance.

3. The method of claim **2** wherein said artificial particles are provided by injecting same into said at least one region from an orbiting satellite.

4. The method of claim **1** wherein said threshold excitation of electron cyclotron resonance is about 1 watt per cubic centimeter and is sufficient to cause movement of a plasma region along said diverging magnetic field lines to an altitude higher than the altitude at which said excitation was initiated.

5. The method of claim **4** wherein said rising plasma region pulls with it a substantial portion of neutral particles of the atmosphere which exist in or near said plasma region.

6. The method of claim **1** wherein there is provided at least one separate source of second electromagnetic radiation, said second radiation having at least one frequency different from said first radiation, impinging said at least one second radiation on said region while said region is undergoing electron cyclotron resonance excitation caused by said first radiation.

7. The method of claim **6** wherein said second radiation has a frequency which is absorbed by said region.

8. The method of claim **6** wherein said region is plasma in the ionosphere and said second radiation excites plasma waves within said ionosphere.

9. The method of claim **8** wherein said electron concentration reaches a value of at least 10^{12} per cubic centimeter.

10. The method of claim **8** wherein said excitation of electron cyclotron resonance is initially carried out within the ionosphere and is continued for a time sufficient to allow said region to rise above said ionosphere.

11. The method of claim **1** wherein said excitation of electron cyclotron resonance is carried out above about 500 kilometers and for a time of from 0.1 to 1200 seconds such that multiple heating of said plasma region is achieved by means of stochastic heating in the magnetosphere.

12. The method of claim **1** wherein said first electromagnetic radiation is right hand circularly polarized in the northern hemisphere and left hand circularly polarized in the southern hemisphere.

13. The method of claim **1** wherein said electromagnetic radiation is generated at the site of a naturally-occurring hydrocarbon fuel source, said fuel source being located in at least one of northerly or southerly magnetic latitudes.

14. The method of claim **13** wherein said fuel source is natural gas and electricity for generating said electromagnetic radiation is obtained by burning said natural gas in at least one of magnetohydrodynamic, gas turbine, fuel cell, and EGD electric generators located at the site where said natural gas naturally occurs in the earth.

15. The method of claim **14** wherein said site of natural gas is within the magnetic latitudes that encompass Alaska.

* * * * *

Nov. 12, 1957

L. R. CRUMP

POWERING ELECTRICAL DEVICES WITH ENERGY
ABSTRACTED FROM THE ATMOSPHERE

Filed March 12, 1954

2,813,242

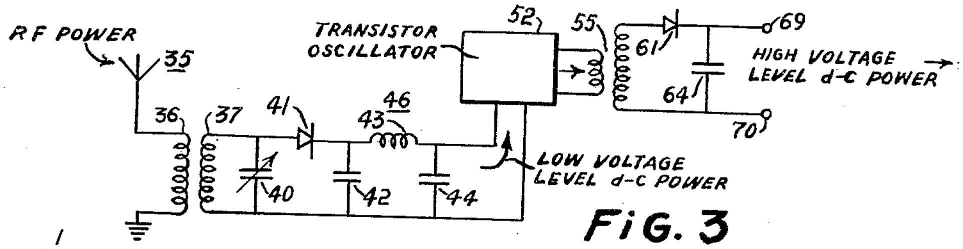
RF POWER
35
36
37
41
43
46
40
42
44
TRANSISTOR OSCILLATOR
52
55
61
64
69
70
HIGH VOLTAGE LEVEL d-c POWER
LOW VOLTAGE LEVEL d-c POWER

FiG. 3

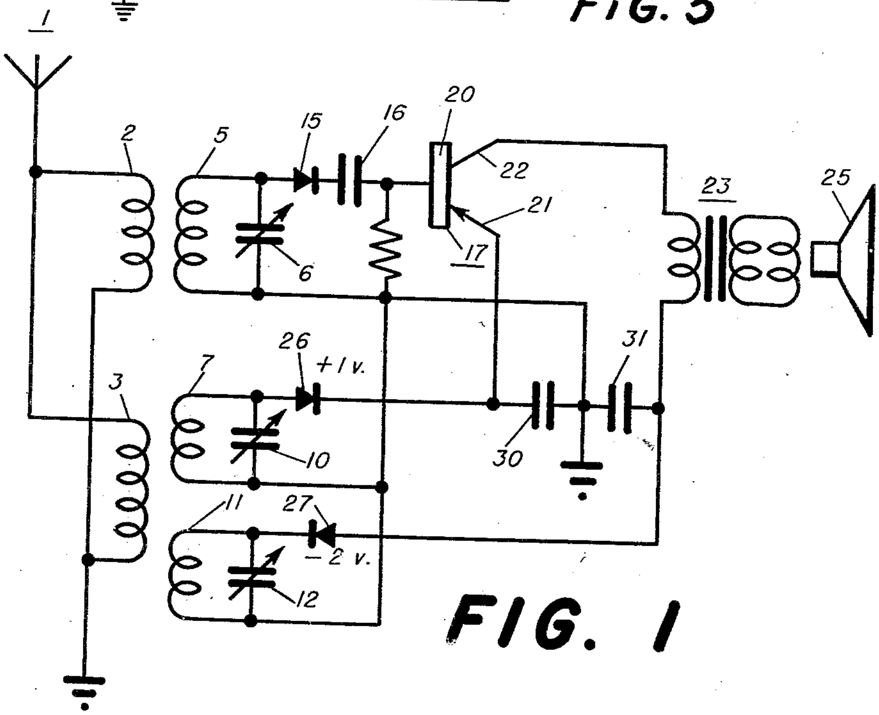
1
2
3
5
6
7
10
11
12
15
16
17
20
21
22
23
25
26
+1 v.
27
-2 v.
30
31

FIG. 1

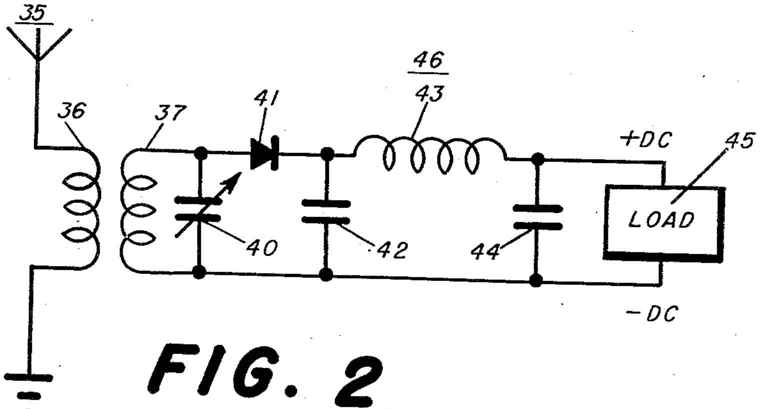
35
36
37
41
46
43
40
42
44
+DC
45
LOAD
-DC

FIG. 2

INVENTOR
Lloyd R. Crump

BY W.E.Thibodeau+ A.W. Dew

ATTORNEYS

United States Patent Office

2,813,242

Patented Nov. 12, 1957

2,813,242

POWERING ELECTRICAL DEVICES WITH ENERGY ABSTRACTED FROM THE ATMOSPHERE

Lloyd R. Crump, Silver Spring, Md.

Application March 12, 1954, Serial No. 415,986

1 Claim. (Cl. 321—2)

(Granted under Title 35, U. S. Code (1952), sec. 266)

The invention described herein may be manufactured and used by or for the Government for governmental purposes without payment to me of any royalty thereon.

This invention relates to the convenient and economical provision of power for the operation of electronic circuits and devices using transistors, and of other electrical devices having modest power requirements.

A great advantage of transistors, and a major reason for their enthusiastic reception since their introduction a few years ago, is the fact that they will operate satisfactorily with very low supply voltages and currents. One milliwatt or even less is sufficient to power a transistor in many applications. Various batteries have been developed to provide, in a minimum of space, the relatively minute amounts of power needed by transistors.

My invention provides methods and means that permit transistor circuits, and also other low-powered electrical devices, to be economically and conveniently operated without any batteries whatever, and indeed without any power supply whatever as power supplies are ordinarily conceived.

The invention centers around my discovery that it is practicable to construct operative transistor circuits that are able to abstract from the atmosphere sufficient electromagnetic energy to provide all necessary supply voltages and currents for their own operation. Circuits and devices powered according to my invention will operate indefinitely without any local power source whatever.

I have successfully constructed and demonstrated such circuits. For example, I have constructed a batteryless transistor radio receiver on which I have listened to either nearby or distant broadcast stations as desired, using either headphones or a loudspeaker; this receiver has been powered entirely by electromagnetic energy abstracted from the atmosphere.

From the successful operation of this receiver, and from other experimental work, it becomes clear that, by the methods and means of the invention, a great variety of practical and useful transistor circuits can be powered entirely by energy abstracted from the atmosphere.

Furthermore, as will become apparent below, my invention is applicable to the powering of other electrical devices requiring relatively small amounts of power.

An object of the present invention is to provide methods and means for powering transistor circuits entirely from radiofrequency energy abstracted from the atmosphere.

Another object is to provide methods and means for powering remote radio receivers, low-powered radio transmitters, and other low-powered electrical devices, with energy received by radio from a master station, so that no local power supplies are needed by the devices and so that the powering or non-powering of the remote device is under the control of the master station.

A further object is to provide methods and means for powering transistor circuits and other low-powered electrical devices with radiofrequency energy received from one or more remote radio transmitters.

Other objects, aspects, uses, and advantages of the invention will become apparent from the following description and from the drawing.

Figure 1 is a schematic diagram of a transistor radio receiver in which all necessary power is supplied by energy abstracted from the atmosphere in accordance with the invention.

Figure 2 is a schematic diagram showing a general application of the invention to provide direct-current power to a load.

Figure 3 is a schematic diagram of a system for obtaining a high energy D.-C. source at a high voltage level using energy abstracted from the atmosphere.

Referring to Figure 1, a receiving antenna **1** is connected to antenna coupling coils **2** and **3**, the other ends of which are connected to ground. A parallel resonant circuit consisting of coil **5** and capacitor **6** is coupled to coil **2**. A second parallel resonant circuit consisting of coil **7** and capacitor **10** is coupled to coil **3**. A third parallel resonant circuit consisting of coil **11** and capacitor **12** is also coupled to coil **3**.

Coil **5** and capacitor **6** are tuned to the frequency of a radio transmitter from which it is desired to receive information—for instance, an amplitude-modulated standard broadcast station. The signal received from this transmitter need not be strong. The signal is detected by diode **15** to obtain an audio-frequency information signal. This audio signal is coupled through a capacitor **16** and is amplified by a circuit that includes a transistor **17** having a base **20**, an emitter **21**, and a collector **22**. The amplified audio output of the transistor is coupled through an audio transformer **23** to an electroacoustical transducer, preferably a permanent-magnet dynamic loudspeaker **25** as shown.

The novelty of the invention lies largely in the method and means by which the necessary direct-current power is supplied to the emitter and collector circuits of transistor **17**. This method and means will now be described.

Coil **7** and capacitor **10**, and also coil **11** and capacitor **12**, are tuned to receive radio signals of relatively high strength. It does not matter whether these signals contain information. These power signals are rectified by diodes **26** and **27** to provide direct-current power that is filtered by capacitors **30** and **31**. The D.-C. power thus obtained is utilized to power the transistor **17**.

In the circuit shown, two tuned circuits (coil **7** and capacitor **10**, and coil **11** and capacitor **12**) are tuned to power signals and the D.-C. voltages obtained from each are connected in series. The tuned power circuits may be tuned to the same or different power signals. Under certain circumstances it may be desirable to use more than two tuned power circuits and to tune them to more than two power signals; in this way power can be obtained from several signals and combined. On the other hand, if a strong power signal is available, a single tuned power circuit may suffice to give the needed D.-C. power.

Even weak information signals can be received successfully. A plurality of transistor amplifier stages can be used if desired, or other circuits such as superheterodyne circuits can be used. It is merely necessary that a sufficiently strong power signal or signals be available to provide the small amount of D.-C. needed to power the transistors.

If the information signal happens to be strong, it can be used as the power signal; all of the tuned circuits (coil **5** and capacitor **6**, coil **7** and capacitor **10**, coil **11** and capacitor **12**) are tuned to the information signal.

Engineers who have observed my invention in operation have been surprised at the unexpectedly good results obtained, even with readily available power signals of quite moderate strength. For instance, sufficient power for satisfactory operation of a loudspeaker at low volumes is readily obtained from a 5-kilowatt standard broadcast station 5 miles away, using only an indoor antenna to pick up the power signal as well as information signals. In typical operation under these conditions a D.-C. voltage of about 2.5 to 3 volts is obtained between the emitter and the collector, at a current of about 250 microamperes; D.-C. power input to the transistor is thus of the order of 0.5 to 1 milliwatt. So far as I am aware, no one has ever before discovered and demonstrated the practicability of this method of powering a radio receiver.

Because existing broadcast stations within a radius of a number of miles provide adequate power signals, the invention is readily practicable with existing power signals in almost any location in or near any city in the United States.

Although I have described a transistor radio receiver powered by my invention, it will be readily apparent that the invention is applicable to the powering of any transistor circuit using one or a number of transistors, and to the powering of other devices requiring relatively small amounts of power. For instance, sensitive electromechanical, electrochemical, or electrothermal devices can be operated by the method of the invention.

Referring to Figure 2, which shows a more general embodiment of my invention, an antenna **35** picks up radiofrequency energy from the atmosphere. This energy flows through coil **36**, which is coupled to a tuned circuit consisting of coil **37** and capacitor **40**. The radiofrequency voltage across capacitor **40** is rectified by diode **41** and filtered by a low-pass filter **46** consisting of capacitors **42** and **44** and choke coil **43**. The resulting D.-C. voltage is applied to a load **45**.

In the practice of my invention, larger amounts of power can be obtained for short periods of time by storing received energy in a suitable energy storage device. Stored energy may then be withdrawn at intervals at a more rapid rate than that at which it was received and put into the storage device. In this way the invention can be used to provide short pulses of relatively very high electrical energy. This result can be readily obtained by charging a relatively large capacitor with direct current and then discharging the capacitor rapidly into a load when desired. This rapid discharge can be initiated automatically when the voltage across the capacitor reaches a certain level, or it can be initiated when a transistor radio receiver receives a certain information signal.

Higher voltages can be obtained with the invention by means of well known devices for raising D.-C. voltages as shown in Figure 3. The D.-C. voltage output from the capacitor **44** can be used to power a low frequency transistor oscillator **52** whose A.-C. output is raised to a higher voltage level by the transformer **55**. This relatively high A.-C. voltage can then be rectified by a diode **61** and fed to a capacitor **64** to provide a high energy D.-C. source at a relatively high voltage level at the terminals **69** and **70**. If desired, energy can now be withdrawn from the capacitor **64** at intervals in short pulses of high energy at a high voltage level. Pulsed radio transmission is one of the possible uses for this form of the invention. Other uses would be to provide a single relatively powerful pulse needed to actuate an electrothermal or electromechanical device.

As has been indicated above, in many locations and particularly anywhere in or near most American cities, power signals normally present in the atmosphere are readily available for the easy and convenient practice of the invention. However, the invention also has important applications in systems in which the necessary power signal is generated and transmitted specifically for the operation of the particular system. Such systems can, for example, comprise a master station transmitting all the power that is needed for hundreds or thousands of fixed or mobile transistor receivers or other remote devices over a range of many miles. This eliminates the need for hundreds or thousands, as the case may be, of local power supplies. At the same time, such a system has the advantage that all of the remote devices can be simultaneously activated or deactivated at the will of the master station, simply by starting or stopping the transmission of the power signal. In such systems it will often be advantageous to use power signals of frequencies sufficiently high to permit the use of resonant receiving antennas of small physical dimensions for signal pickup at the remote devices. In addition to the power signal, the master station may transmit an information signal on the same or a different carrier.

Certain types of devices powered entirely by received radio waves are of course well known. The well-known "crystal set" of the early days of radio, which used a diode rectifier to demodulate an amplitude-modulated radiofrequency signal, is an outstanding example of such a device. My invention is readily distinguishable from such prior devices, however. In typical prior devices a modulated radiofrequency signal is applied to a diode to obtain unidirectional half-wave pulses whose amplitudes vary with modulation. These pulses are integrated by means of a capacitor to obtain a unidirectional signal the amplitude of which follows the audiofrequency modulation envelope. If the radiofrequency signal is received with sufficient strength the audio signal may have sufficient power to operate headphones or similar utilization device without power amplification; but the signal is utilized for its information content, rather than to supply non-information-containing power.

My invention, on the other hand, entails the utilization of received radiofrequency energy to supply power to at least one pair of circuit points (across capacitor **31** in Fig. 1, for example), such circuit points requiring power solely for its power content and not for any information or modulation it may contain. In other words, my invention entails the utilization of radiofrequency energy to supply power that would otherwise have to be supplied by batteries, generator, or other local power source.

It will be apparent that the embodiments shown are only exemplary and that various modifications can be made in construction and arrangement within the scope of the invention as defined in the appended claim.

I claim:

An electrical device for obtaining a high energy D.-C. source at a high voltage level using energy abstracted from the atmosphere, said device comprising in combination: resonant means for receiving radio waves, first rectifier means for converting said radio waves into first direct current energy, first capacitor means for storing said first direct current energy, an oscillator powered by said direct current energy, said oscillator producing an A.-C. output, transformer means for raising said A.-C. output to an increased voltage level, second rectifier means for converting the A.-C. output of increased voltage level from said transformer into second direct current energy, and second capacitor means for storing said second direct current energy.

References Cited in the file of this patent

UNITED STATES PATENTS

1,540,998	Plauson	June 9, 1925
2,056,824	Cawley	Oct. 6, 1936
2,553,734	Adler	May 22, 1951
2,569,687	Pan	Oct. 2, 1951
2,647,957	Mallinckrodt	Aug. 4, 1953

OTHER REFERENCES

Turner: "A crystal receiver with transistor amplifier," Radio and Television News, January 1950.

H. C. VION.

Electric Apparatus.

No. 28,793. Patented June 19, 1860.

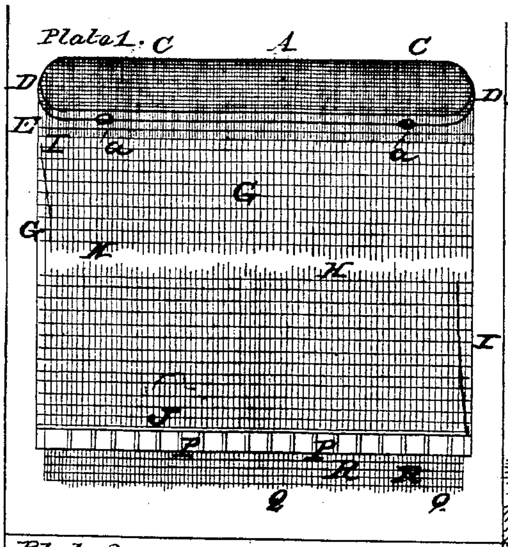
Plate 1.
C
A
C
D
D
E
a
a
G
G
K
K
I
J
P
P
R
R
Q
Q

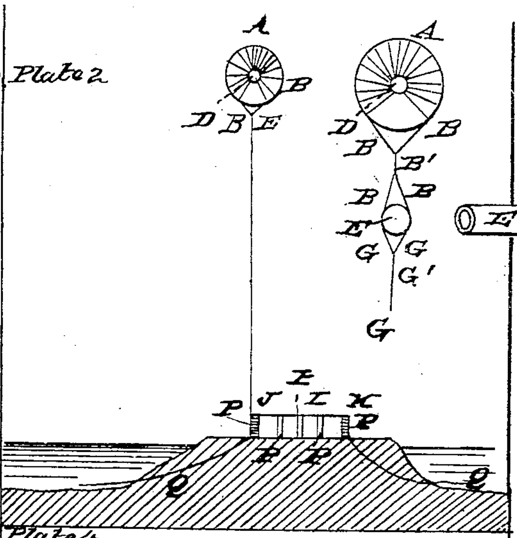
Plate 2
A
B
D
B
E
A
D
B
B
B'
B
B
E
G
G
G'
G
E
I
P
J
I
K
P
P
P
Q
Q

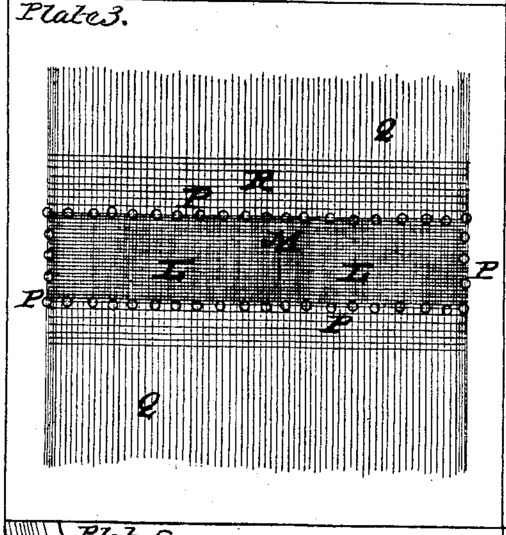
Plate 3.
Q
R
P
M
L
L
P
P
P
Q

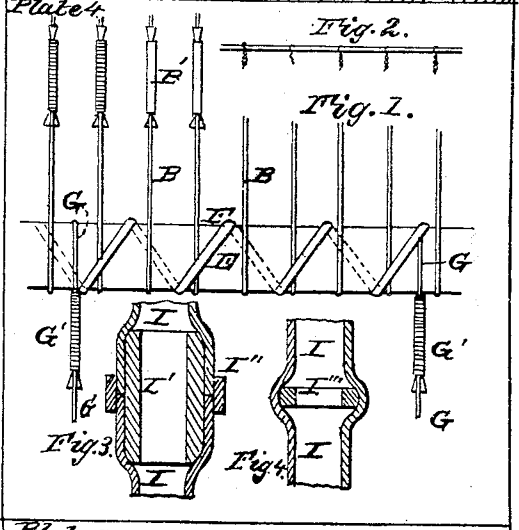
Plate 4.
Fig. 2.
E'
Fig. 1.
B
B
G
E
D
G
G'
G'
I
I'
I''
I
I'''
6
G
Fig. 3.
I
Fig. 4.
I

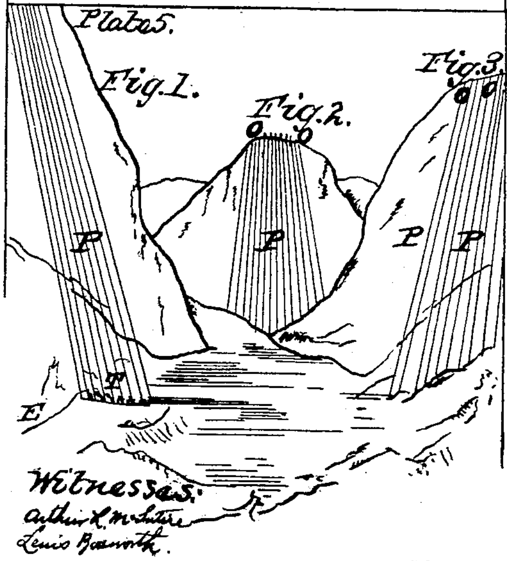
Plate 5.
Fig. 1.
Fig. 2.
Fig. 3.
P
P
P
P
P
E

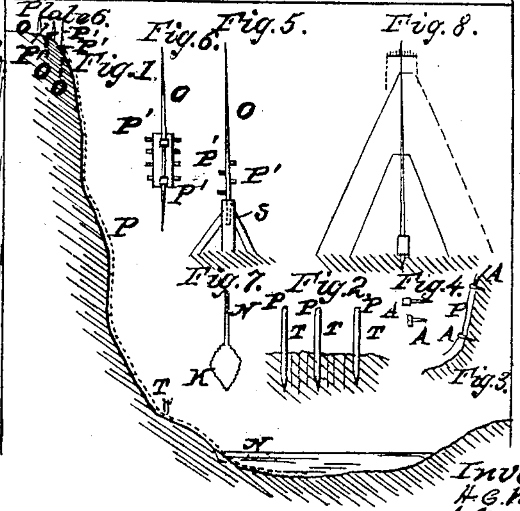
Plate 6.
Fig. 1.
Fig. 6.
Fig. 5.
Fig. 8.
O
O
P'
P'
P'
S
P
Fig. 7.
Fig. 2.
Fig. 4.
N
P
P
T
T
T
A
A
A
P
K
T
N
Fig. 3.

Witnesses:
Arthur L. McIntire
Lewis Bosworth

Inventor:
H. C. Vion
by Munn & Co.
Attorneys.

N. PETERS, Photo-Lithographer, Washington, D. C.

UNITED STATES PATENT OFFICE.

H. CHARLES VION, OF PARIS, FRANCE.

IMPROVED METHOD OF UTILIZING ATMOSPHERIC ELECTRICITY.

Specification forming part of Letters Patent No. **28,793**, dated June 19, 1860.

To all whom it may concern:

Be it known that I, HIPPOLYTE CHARLES VION, of Paris, in the Empire of France, engineer, have invented a new Mode of Obtaining Atmospheric Electricity and Terrestrial Electricity and its Industrial Applications; and I do hereby declare that the following is a full, clear, and exact description of the same, reference being had to the accompanying drawings, making part of this specification, in which—

Plate 1 represents a front elevation; Plate 2, a side elevation of an aerostat, in combination with certain other devices for obtaining atmospheric and terrestrial electricity; Plate 3, a plan of the device for obtaining terrestrial electricity; Plate 4, detached portions of the apparatus. Plate 5 shows a front view, and Plate 6 a vertical section and various details of the apparatus when applied in mountainous regions, the use of an aerostat being dispensed with.

The object of my invention is to form an electric pile of great power by using the positive electricity contained in the atmosphere, and the negative electricity contained in the earth, so as to make the electricity therein contained available for industrial purposes.

To enable others skilled in the art to make and use my invention, I will proceed to describe its construction and operation.

A is aerostat of a tubular form closed at both ends. It is made of suitable material, so as to be impervious to air. Its dimensions are such as to give it great ascensional power when filled with gas.

a a are valves in the surface of the aerostat, to be opened when the inflation of the aerostat should become too great.

A long india-rubber tube, I, communicates with the interior of the aerostat, being attached to the latter, near one end of its ends D, as seen in Plate 1, while the lower end of the tube I is attached to a gasometer, (not shown in the drawings.) The gasometer is to be fed with hydrogen gas, produced by the action of the pile itself, the negative wires of which (afterward to be described) enter a body of water at or near the base of the apparatus and decompose the water so as to produce the hydrogen gas. As the aerostat is supposed to be at a considerable height the tube I must be of corresponding length, and is constructed of a number of tubes, short wooden tubes I′ being inserted where the tubes I are joined, and a fastening-ring, I″, being slipped over each of the joints, as seen in Fig. 3 of Plate 4. At certain distances the tube I is fastened to the net-work of the positive wires (afterward to be described) in order to secure the tube against the action of the wind, and at each of these fastening-places a washer, I‴, is inserted in the tube in order that the tube shall not be compressed by the wire or other means employed to fasten it to the net-work of positive wires. (See Fig. 4, Plate 4.)

The aerostat is surrounded with a net-work of wires, one layer of the wires, C, being parallel with the axis of the aerostat and fastened to rings D at both ends of the aerostat, and the other layer of wires, B, extending partially around the aerostat at right angles to the wires C. One end of each of the wires B extends around an iron tube, E, some distance below the aerostat and meets the other end of it between the tube E and the aerostat. The two ends are fastened together by a ligature, B′. (See Fig. 1, Plate 4.) Each end of the tube E terminates into a ball, *e*. The wires B are fastened to the surface of the tube E by means of a helical wire, F, wound around the tube and across the wires B, as seen in Fig. 1, Plate 4. The upper ends of long vertical wires G are also wound around cylinder E, each wire G between two of the wires B, and the ends secured by a ligature, G′, as seen in Fig. 1, Plate 4. The helical wire F is also wound across the wires G, so as to keep them in their places on tube E.

The vertical wires G, which are to be the conductors of the positive electricity of the atmosphere, must be of a length proportionate to the desired efficacy of the electric pile, and the size and ascensional power of the aerostat must, of course, be adequate to sustain the weight of and keep suspended the wires G, (a weight still further increased by the horizontal cross-wires H, with which the vertical wires G are interlaced, in order to form a net-work not liable to be deranged by the action of the winds or similar influences.) The two outside wires, G, are stronger than the rest of them, and their lower ends are fastened to dynamometers of any suitable construction. These dynamometers are attached to the ends or a massive iron cylinder, J, and they serve to indicate the tension in the outside wires, G, and the corresponding ascensional power of the aerostat. According to the reading of these dynamometers the aerostat has to be supplied (through tube I) with more or less gas. The lower end of each of the wires G is wound around the cylinder J, and secured by ligatures similar to those above described. The wires G are all insulated (by a coating of gutta-percha or similar substance,) except where they are in contact with tube E and with cylinder J, and a similar insulating coating is laid on cylinder J, after the lower ends of the wires G have been fastened to it.

Another cylinder, K, similar to J, is placed at some distance from and parallel to cylinder J. It is connected with cylinder J by wires L, wound around both cylinders and interlaced with cross-wires M. The wire-work L M and cylinders are insulated (in a manner already described) against outside influences, so that the only electric communication between the two cylinders will be through the wires L. The two cylinders are placed upon insulated columns P. The cylinder K may be used as a substitute for cylinder J, and vice versa, whenever repairs become necessary. Insulated branch wire or wires are attached to the cylinders J K and wires L, so as to conduct the positive electricity obtained from the atmosphere by means of the above-described apparatus to wherever it is desired for industrial purposes. The insulated wires Q (interlaced with cross-wires R) are placed on the ground underneath and parallel to the positive wires L. Both ends of each of the wires Q are sunk into the earth or submerged in water, and fastened to a metallic plate coated with a metal not subject to oxidation. These wires Q are the conductors for the negative electricity of the earth, and a branch wire or wires attached to the wires Q serve to transmit the negative terrestrial electricity to wherever it is wanted for industrial or other purposes.

By uniting to the ends of the positive and the negative branch wire or wires a powerful electric current will be obtained, one pole of which is the atmosphere and the other the earth, and may be applied to any suitable useful purpose.

I will now proceed to describe the modification of the above-described apparatus when to be applied in mountainous countries.

P represents the positive electric copper or other metal wires coated over with an insulating substance. The upper ends of each of the positive wires is soldered to a prompter, O, at P′, Figs. 1, 5, and 6, Plate 6. The lower portion of each of the positive-wires is secured to an insulator, T, Fig. 2, Sheet 6. The positive wires are held above the ground by joints A, Figs. 3 and 4, Sheet 6, projecting from the soil at suitable distances from each other. The wires P are intended to follow the inequalities of the ground on which they are laid.

The prompters O, Figs. 5 and 6, Sheet 6, are iron rods sharpened to a point and silvered or coppered at their upper ends. The lower part of the prompter is fastened into a pole, S, covered with tar, which isolates the prompter and holds it in a firm position. A large metallic plate may be soldered to each prompter, as shown in Fig. 6. The positive wires may be soldered to the rod of each prompter or to the plate which is fastened thereto.

One or more branch lines, E, are soldered up to the positive wires to transmit the positive atmospheric electricity for which the wires P are the conductors to any desirable point.

N are negative iron or other metal conductors coated with an insulating substance. The upper ends of these wires rest on the ground near the positive insulators. The lower ends of these wires are soldered to a metallic plate or plates, V, Fig. 7, Plate 6, coated with a metal not subject to oxidation. The negative wires are sunk into the ground at very great depth or into wells, rivers, or into the sea. The negative electric branch wires are attached to the negative conductors N in the same manner as the positive branch wires are to the positive conductors. The branch wires and the soldering are coated over with an insulating substance. They are intended to carry the negative terrestial electricity to any desired point.

What I claim as my invention, and desire to secure by Letters Patent, is—

1. The peculiar arrangement of means herein specified, whereby I am enabled to use the positive electricity contained in the atmosphere and the negative electricity contained in the earth, and thus form an electric pile of considerable power and make the electricity therein contained available for industrial purposes, as set forth.

2. The combination of an aerostat and vertical wire-work with a tube, I, for admitting gas into the aerostat, in the manner and for the purposes above set forth.

CHARLES VION.

Witnesses:
Harry W. Spencer,
A. Guion, Jr.

W. I. PENNOCK.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

APPLICATION FILED JUNE 26, 1907.

911,260. Patented Feb. 2, 1909.

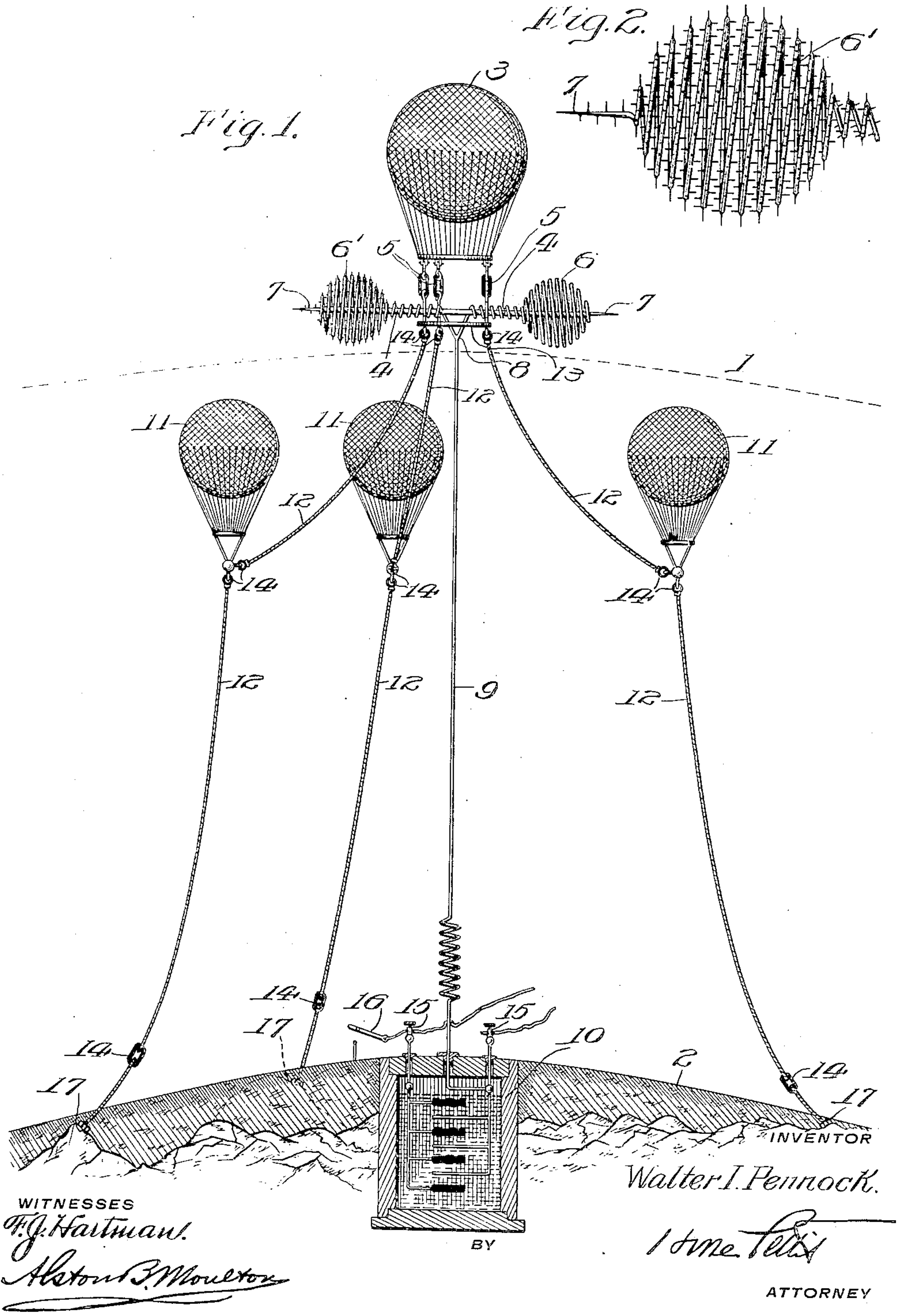
Fig. 2.
6'
7
Fig. 1.
3
5
4
6'
5
6
7
7
14
14
4
8
13
1
12
11
11
11
12
12
14
14
14
12
12
9
12
14
16
15
15
10
17
14
17
2
14
17

INVENTOR

Walter I. Pennock.

WITNESSES

F. J. Hartman.

Alston B. Moulton

BY

ATTORNEY

UNITED STATES PATENT OFFICE.

WALTER I. PENNOCK, OF PHILADELPHIA, PENNSYLVANIA.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

No. 911,260. Specification of Letters Patent. Patented Feb. 2, 1909.

Application filed June 26, 1907. Serial No. 380,907.

To all whom it may concern:

Be it known that I, Walter I. Pennock, a citizen of the United States, residing at Philadelphia, in the county of Philadelphia and State of Pennsylvania, have invented certain new and useful Improvements in Apparatus for Collecting Atmospheric Electricity, of which the following is a specification.

My invention relates to a method of collecting electricity from a strata laden with electricity at high altitudes in the atmosphere, through the medium of a wire cable suspended from one or more balloons and in conveying this electricity to the earth's surface.

The object of my invention is to provide a conveyance of the electro motive force to be found in the upper strata of the atmosphere to the earth's surface; where it may be utilized for commercial and other purposes.

A further object of my invention is to provide a device or mechanism by which a suitable collector for the electrical energy in the upper strata of the earth's atmosphere may be elevated in the said strata and by which the said electric energy may be transmitted to and collected at the earth's surface, from which point it may be conducted to any place where it is desired to use the same.

A further object of my invention is to support and anchor said device in any desired position.

A further object of my invention is to provide an improved form of collector through which the energy of the said upper strata of the earth's atmosphere may be collected and transmitted for various purposes to the earth's surface.

Other objects of my invention will appear in the specification and claims below.

For a further full, clear and complete disclosure of my invention, reference may be had to the following description and accompanying drawings, in which like reference characters refer to corresponding parts.

Figure 1 is an elevational view of one form or embodiment of my invention and Fig. 2 is a detailed view of one form of my improved collector.

The passage of the electrical current to the earth under ordinary conditions is prevented by an obstruction afforded by the dense lower strata of the atmosphere, which is a bad conductor of electricity, as shown by the electrodes of an ordinary static machine. When the electrodes are placed close together, the atmosphere is seen to carry the current across from one electrode to another, but if placed far apart the current is obstructed by the intervening air. The dense lower strata of the atmosphere affords one of the best non-conductors of electricity, as shown in the conduction of the electric current by telegraph, or trolley wires on the earth's surface, where only a small quantity of the electric current escapes through the atmosphere; while rarefied atmosphere affords a good conducting media, as shown by the vacuum tube. The upper strata of the atmosphere being rare in proportion to the altitude, and being a good conductor of electricity while the lower strata of the atmosphere being dense and affording a non-conducting media for the electric current, thus causes an obstruction to the electric current, or power in its passage to the earth's surface from the electric strata of the atmosphere under ordinary conditions. When, however, the cumulus clouds of a thunder storm pass over the surface of the earth, these clouds being of very great height, the moisture in said clouds forms a better conductor of electricity than does the dry air, with the result that the electricity in the upper strata of the atmosphere breaks through the said cloud as a streak of lightning and in that form reaches even to the earth's surface, while the thin or shallow strata of clouds, observed in the so-called "settled rain" storm, do not extend upward to a sufficient height to form a conducting medium for the electricity from the electric strata to the earth's surface. For this reason there is usually no lightning during the said latter variety of rain storm.

By means of my invention, I have provided a mechanism for collecting the electrical energy or power created by nature and stored in the upper strata of rarefied air of the earth's atmosphere and have provided a conductor for said electric energy to the earth's surface.

Referring now to the drawings 1 indicates what may be called the lower limits or boundaries of the strata of electric energy above the surface 2 of the earth.

3 indicates a balloon which is elevated to a high altitude until it enters the said stra-

tum. The said balloon 3 carries a bar 4 of wood or any other suitable non-conducting material suspended by insulating links 5 or any other suitable form of insulation from the balloon 3. Upon either end of the said bar 4 I mount collectors, one form of which I have illustrated as spheres of coiled wire 6—6′ the outer ends 7 of which terminate in sharp points. The inner turns of the said spheres 6 and 6′ are wrapped around the wooden bar 4 and united as at 8 to a conductor of electric energy, preferably a large copper wire 9. This wire or conductor 9 extends to the earth's surface and may have its end suitably attached to an electric accumulator or other piece of electrical apparatus. I have illustrated one form of my invention in which the conductor 9 is connected to one pole of a storage battery 10 on the earth's surface.

The spiral spheres 6, 6′ are preferably provided with a polished metallic surface to form a good conductor of electricity and the material of said spheres should also be of such a character that it will not rust or corrode or tarnish. A polished copper wire or a copper wire plated with platinum or gold or a solid platinum or gold wire may be used for this purpose, inasmuch as these materials are least affected by moisture and the atmosphere. The said spheres or collectors may be made of smooth wire as shown in the sphere designated by the numeral 6 or of barbed wire, as is shown at 6′ and illustrated on a larger scale in Fig. 2 of the drawings. The latter form is preferable inasmuch as it provides a large number of points through which the electricity may flow to the wire from the surrounding air in the said upper strata of the earth's atmosphere.

In order that the supporting balloon 3 may be held in a relatively fixed position, it should be suitably anchored to the earth's surface. Inasmuch as the balloon 3 must be elevated to a very high position, the weight of the anchoring cables forms an important consideration, and if desired or necessary one or more supplemental balloons 11, 11, 11 may be attached to each of the anchoring cables 12, 12, 12 as illustrated in Fig. 1, to relieve the balloon 3 of such weight as would prevent it from ascending into the said electrical strata. In order that the electricity from the said upper strata of the earth's atmosphere may not be conducted down the anchoring cables 12, 12, 12, I attach them to the supporting balloon 3 and to the supplemental balloons 11 and to the earth's surface through suitable insulating devices 14.

In the form of my invention illustrated in Fig. 1, below the bar 4, I suspend a ring 13 of any suitable material from the balloon 3, and attach the anchoring cables 12 thereto by means of insulating rings 14, 14, 14. I may also provide additional insulating rings 14 between that portion of the anchoring cables 12 between the balloon 3 and the supplemental balloons 11, and also between that part of the cable 12 between the supplemental balloons 12 and the earth's surface. I may also provide, near the earth's surface and at the lower end of the anchoring cables 12, similar insulating links 14. While I have described links as forming a convenient form of insulating device for the purposes above set forth, I do not wish to be construed as being limited to the same, inasmuch as any suitable non-conducting connection may be used in place of the links 14.

The terminals 15, 15 of the storage battery 10 may be connected to any piece of electric apparatus which it is desired to run or operate and if desirable one of the poles of the battery is adapted to be connected by the switch 16 with the earth's surface. The lower ends of the anchoring cables 12 are securely anchored to the earth's surface as at 17.

With the apparatus arranged and connected in the manner illustrated in Fig. 1, the electric energy in the high strata of the earth's atmosphere passes to and through the conductive spiral spheres 6 or 6′ to the conductor 9 and is suitably stored or used at the earth's surface, while the balloons 11 support a part of the weight of the anchoring cables 12, and permit the balloon 3 to ascend as high as is possible, or necessary for it to enter the said electrical strata of the earth's atmosphere. By arranging the anchorage 17 of the cables 12 symmetrically or in any other position than in a straight line, the balloon 3 may be held in a substantially fixed position with relation to the earth.

While I have illustrated in the drawings and have described in the specification a form of apparatus in which my invention may be carried out, it is obvious that the drawings are more or less diagrammatic drawings, that is to say, that the proportions of the various parts are not necessarily those which would operate to the best advantage, inasmuch as certain portions have been shown as greatly enlarged in the drawings for the sake of clearness, and that it is likely that more than one supplemental balloon would be requird for each cable in order to support the weight of the same, and to relieve the supporting balloon 3 of such weight, as would prevent it from ascending into the high electrical strata of the earth's atmosphere, but such changes in form, proportion and arrangement I regard as being fully within the aim and scope of my invention, so long as such forms or modifications fall within the scope of the appended claims. It is also to be understood that the storage battery or accumulator which I have shown as being connected to my collector, is only a type of apparatus which can be operated by the current collected by the spheres 6, 6′ and transmitted to the earth through the wire 9, and when I use the word "accumulator", I mean any piece of useful apparatus which is operated by the current transmitted thereto through the wire or conductor 9.

Having thus described my invention, what I claim and desire to protect by Letters Patent of the United States is:

1. The combination with an electrical collector comprising a bar of non-conducting material, and an open spherical conductor carried by said bar, of means to support said collector in the high electrical strata of the earth's atmosphere.

2. The combination with a balloon, of an electrical collector supported thereby and insulated therefrom, comprising a bar of non-conducting material, and a conductor wound spirally around said bar.

3. The combination with an electrical collector comprising a non-conducting bar, and a conductive wire wound thereon to form an open substantially spherical body, and means to support said collector in the high electrical strata of the earth's atmosphere.

4. The combination with a balloon, of an electrical collector carried thereby and comprising a non-conducting bar, and a polished wire wound spirally thereon to form an open substantially spherical body.

5. The combination with an electrical collector comprising a non-conducting bar, and a wire wound spirally thereon to form an open substantially spherical body, said wire being provided with pointed conducting projections, and means to support said collector in the high electrical strata of the earth's atmosphere.

6. The combination with a balloon, of an electrical collector carried thereby comprising a non-conducting bar, and a conducting wire wound spirally thereon to form open substantially spherical bodies upon the opposite ends thereof.

7. The combination with a balloon, of an electrical collector carried thereby comprising a non-conducting bar, a conducting wire wound spirally thereon to form open substantially spherical bodies upon the opposite ends thereof, an electrical accumulator, and an electrical connection between the said collector and said accumulator.

8. The combination with a balloon, and means to anchor said balloon, of an electrical collector supported by said balloon and insulated therefrom, an electrical accumulator, and a conductor connecting said collector and said accumulator.

9. The combination of an electrical collector, means to support said collector at a high elevation from the earth's surface and within the electrical strata of the earth's atmosphere, an electrical accumulator at the earth's surface, an electrical connection between said collector and said accumulator, and means to insulate said supporting means from said collector and from the earth.

In testimony whereof, I have hereunto set my hand this 25th day of June, 1907.

WALTER I. PENNOCK.

Witnesses:
HUGH F. QUINN,
WM. G. GLENN.

W. I. PENNOCK.

APPARATUS FOR COLLECTING ELECTRICAL ENERGY.

APPLICATION FILED JAN. 4, 1911.

1,014,719. Patented Jan. 16, 1912.

2 SHEETS—SHEET 1.

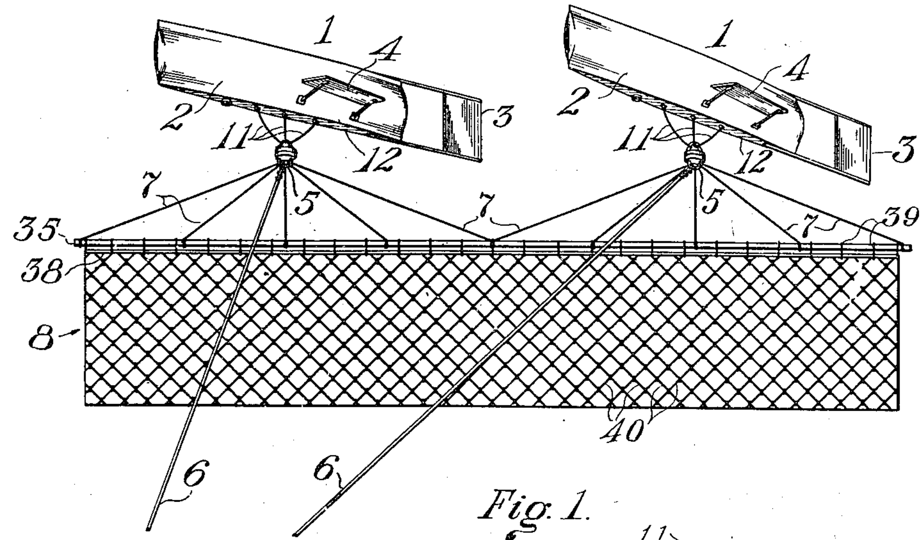

1
2
3
4
11
12
5
7
35
38
8
39
40
6

Fig. 1.

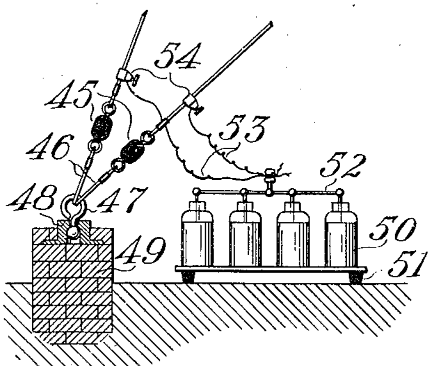

54
45
53
46
52
48
47
50
49
51

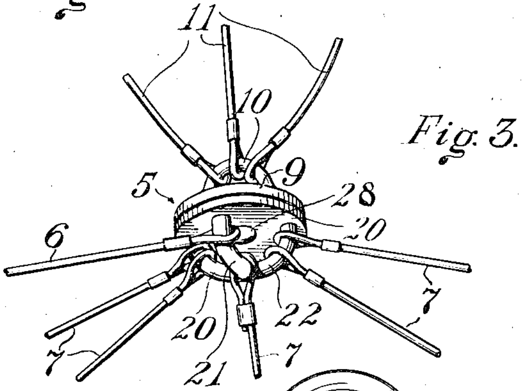

11
10
9
5
28
20
6
7
22
20
21

Fig. 3.

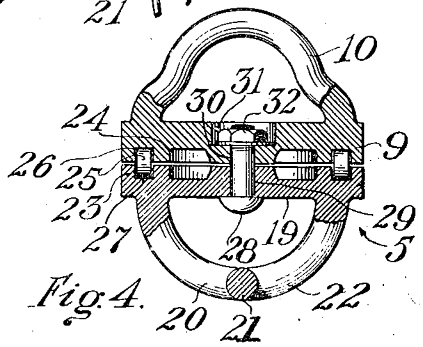

10
31
30
32
24
9
26
25
23
27
29
19
28
5
22
20
21

Fig. 4.

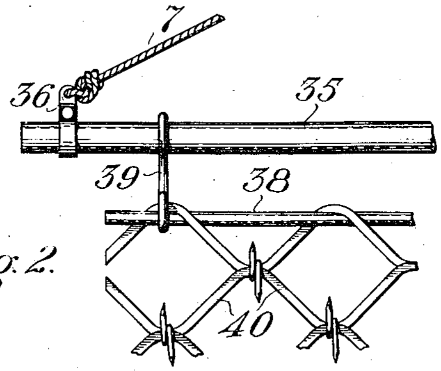

7
36
35
39
38
40

Fig. 2.

WITNESSES

Elizabeth L. Smith

J. Stuart Freeman.

INVENTOR

Walter I. Pennock.

BY

ATTORNEY

W. I. PENNOCK.

APPARATUS FOR COLLECTING ELECTRICAL ENERGY.

APPLICATION FILED JAN. 4, 1911.

1,014,719. Patented Jan. 16, 1912.

2 SHEETS—SHEET 2.

Fig. 5.

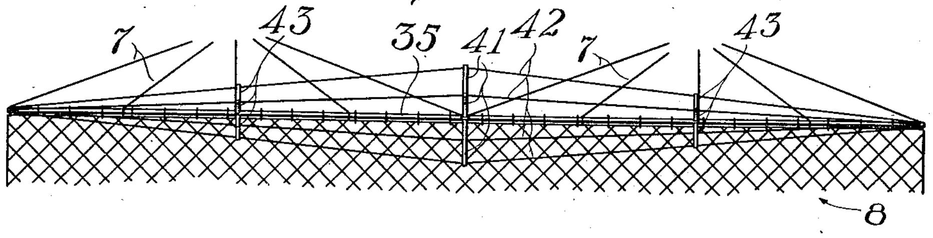
7
43
35
41
42
7
43
8

Fig. 6.

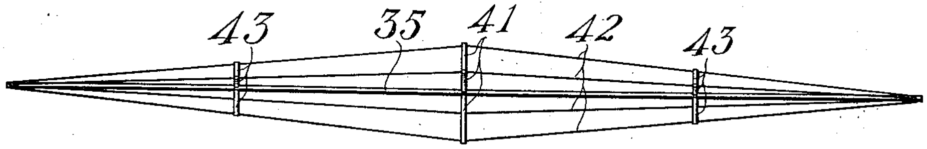
43
35
41
42
43

Fig. 7.

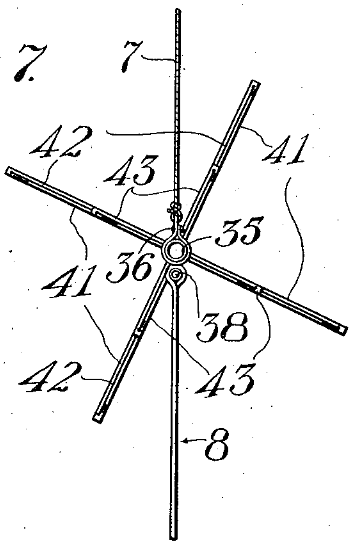
7
41
42
43
35
36
41
38
42
43
8

WITNESSES

Elizabeth L. Smith,

J. Stuart Freeman.

INVENTOR

Walter I. Pennock.

BY

ATTORNEY

UNITED STATES PATENT OFFICE.

WALTER I. PENNOCK, OF PHILADELPHIA, PENNSYLVANIA.

APPARATUS FOR COLLECTING ELECTRICAL ENERGY.

1,014,719. Specification of Letters Patent. Patented Jan. 16, 1912.

Application filed January 4, 1911. Serial No. 600,777.

To all whom it may concern:

Be it known that I, Walter I. Pennock, a citizen of the United States, residing at Philadelphia, county of Philadelphia, and State of Pennsylvania, have invented certain new and useful Improvements in Apparatus for Collecting Electrical Energy, of which the following is a full, clear, and exact disclosure.

The present invention relates to an improved means for collecting the charges of electricity from the upper atmosphere and more particularly to that form which consists in one or more captive balloons from which is suspended a suitable form of metallic conductor.

The principal objects of the device are: to provide a collector for atmospheric and static electricity, which when in operative position will present a large surface to currents of air, but which will offer comparatively little resistance thereto, to provide a collector of such material and construction as will be more efficient in its operation than any previously constructed for a similar purpose, to provide a means for maintaining such a collector suspended in the air and at right angles to opposing currents thereof, and to provide a suitable anchorage for holding said means captive.

With these principal objects in view, the present invention consists in further advantages which are brought out in the following specification and accompanying drawings, in both of which like numerals refer to like parts, and in which drawings—

Figure 1 is a perspective view of the complete device in operative position, Fig. 2 is an enlarged detail view of the wire mesh and the manner in which it is attached to the supporting balloons, Fig. 3 is a detail of the manner of securing the collector-supporting and anchor cables to the balloons, Fig. 4 is an enlarged cross section of the swivel connection shown in Fig. 3, Fig. 5 is top plan view of the reinforcing braces on the screen, Fig. 6 is an elevation of the same and Fig. 7 is an end view of the structure shown in Figs. 5 and 6.

Referring to the drawings, in Fig. 1 thereof, a plurality of balloons 1 of any suitable type is shown, each of which embodies hollow metallic elongated gas tanks 2, extending from the rear of which are single, rigidly affixed rudders 3, while on the sides of the tanks are secured stationary lifting planes 4.

To the bottom and slightly to the rear of the center of the tanks 2 is secured a suitable swivel 5, by which the anchor ropes 6 and the suspension ropes 7 for the metallic conductor 8 are secured to the balloons 1. A suitable form of swivel joint is illustrated in Figs. 3 and 4, but any type can be used that embodies the essential features shown therein.

The swivel joint illustrated consists in the base plate 9 having a looped portion 10 integral therewith and projecting from the upper face thereof. Secured to the loop 10 is a set of three light electrically conductive supporting ropes or cables 11 which extend upwardly and are secured at intervals to the bottom 12 of the balloon above.

The lower or revoluble member 19 of the swivel joint preferably comprises three upwardly directed curved arms, 20, 21, and 22, respectively, forming at their junction a T-shape as shown, said arms at their upper extremities being integral with the plate 23. The member 19 is revoluble below and concentric with the plate 9, and the two members are lightened in weight by opposed concentric grooves as shown at 24. Contact between said members is made through the roller bearings 25, which are carried in the opposed concentric grooves 26 and 27 of the respective upper and lower plates. Furthermore, said plates are maintained in coöperative relation with each other by means of a bolt 28 passing through centrally drilled holes 29 and 30 in the respective lower and upper plates, the drilled hole 30 opening upwardly into an enlarged recess 31, in which is sunken the nut 32 on the bolt 28. Two of the arms 20 and 22 of the lower revoluble member 19 extend in diametrical alinement, while from the central point thereof extends the third arm 21 at right angles thereto, and upward to the plate 23.

Suspended below the plurality of balloons is a hollow rod 35, of any suitable material, connected at regular intervals such as at points 36, by metallic ropes 7, to the alined arms 20 and 22 of the swivel joint 5. Below and parallel to the rod 35 is a similar but smaller rod 38 suspended therefrom by means of suitable couplings 39. From the rod 38 hangs a wire mesh 40 of any suitable form, such as can be extended over a considerable area and which will offer comparatively little resistance to the passage of air therethrough. The metal in such a screen is preferably rough, sharp or jagged and a convenient form embodying these characteristics is expanded sheet metal such as is used for lathing and for reinforcing concrete construction. Instead of the single hollow rod 35 alone, applicant contemplates the use of reinforcing means to prevent the buckling of the rod in a stiff wind, such, for instance, as the arrangement of wire bracing shown in Figs. 5, 6 and 7 in which cross rods 41, secured to the rod 35 at the center thereof are attached to the ends thereof by means of stretched wires 42. Between the rods 41 and the ends of the rod 35 are secured sets of cross rods 43 which support the wires 42 and strengthen the rod 35 at as many points as they may be placed.

For the purpose of maintaining the balloons and apparatus suspended therefrom captive, the light metallic cables 6 terminating downwardly in insulators 45 are employed. These insulators in turn are connected by means of short sections of rope 46 to the eye 47 of a suitable swivel 48, embedded in the anchorage 49.

When the balloons with the metallic screen suspended therefrom are allowed to rise into one of the higher altitudes, the entire apparatus being of metallic construction and uninsulated will become energized by contact with the surrounding natural charges of electricity. From the above description it is evident then that, while the screen 8, on account of its great extent, will be the greatest collecting agent, it will be seen that the balloons themselves and the suspending wires will also coöperate as one large collector, since no parts of which are insulated from any of the neighboring parts thereof. Consequently, when the apparatus described has reached an altitude or strata of the atmosphere abounding in static charges of electricity, an amount of the said charges proportionate to the surface area of the metal exposed will collect upon the apparatus as a whole and will be conducted downwardly toward the earth by means of the various anchor ropes 6, but will not pass into the ground on account of the interposition of the insulators 45.

To use the electrical charges thus acquired, a plurality of Leyden jars 50, or other suitable collectors are supported above the surface of the earth and insulated therefrom by any suitable means as represented by the blocks 51. Either the inner or outer conducting surfaces of the jars may be connected together and energized by the accumulated charge. In the present instance, the inner surfaces of the accumulators are shown to be connected, and the connecting means 52 is in turn connected to the ropes 6 by means of wires 53. These wires are secured to the ropes mentioned by means of suitable binding posts 54.

In the device described the anchor ropes 6 are of substantially the same length, and when the apparatus is raised to the desired altitude and is being blown by the currents of air, the balloons are turned by means of the vanes 3 to parallel relation with each other, and furthermore, from the manner in which the device is held captive and the collecting net 40 is suspended from the balloons, it is obvious that said net will at all times readily swing into a plane substantially perpendicular to any current of air acting upon the balloons above. The purpose of the swivel joint shown in Figs. 3 and 4 is principally for allowing the balloons to readily aline themselves with any new direction of the wind before the apparatus, including the suspended screen, can swing about the swivel 48, and said joint will also prevent the twisting of the ropes 7, when any rapid shifting of the air currents may occur.

While applicant has shown a set of Leyden jars as the accumulators in the accompanying drawings, it is obvious that any other suitable form may be used, and furthermore, that although not illustrated, any suitable apparatus may be run thereby, such as for instance, wireless telegraphic instruments.

Furthermore, although but one embodiment of the invention has been described, it is to be understood that various modifications may be made therein, and in fact several are contemplated by applicant that are of such structure as fall well within the scope of the appended claim.

Having thus described my invention, what I claim and desire to protect by Letters Patent of the United States, is:

A collector for charges of electricity, comprising a plurality of supporting means, a metallic gauze sustained thereby to lie in a plane and substantially equidistant from each of said supporting means, anchoring means emanating from a common point to each of said supporting means, means operative to maintain said supporting means in their normal positions, and means between each of said supporting means and said gauze to permit each of said supporting means to readily and independently aline itself to accord with any alteration in the direction of opposing air currents.

In witness whereof I have hereunto set my hand this 28 day of December, A. D. 1910.

WALTER I. PENNOCK.

Witnesses:
MILDRED S. TEMPLE,
E. EUGENIA PENNOCK.

N. TESLA.

APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.

APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.

1,119,732. Patented Dec. 1, 1914.

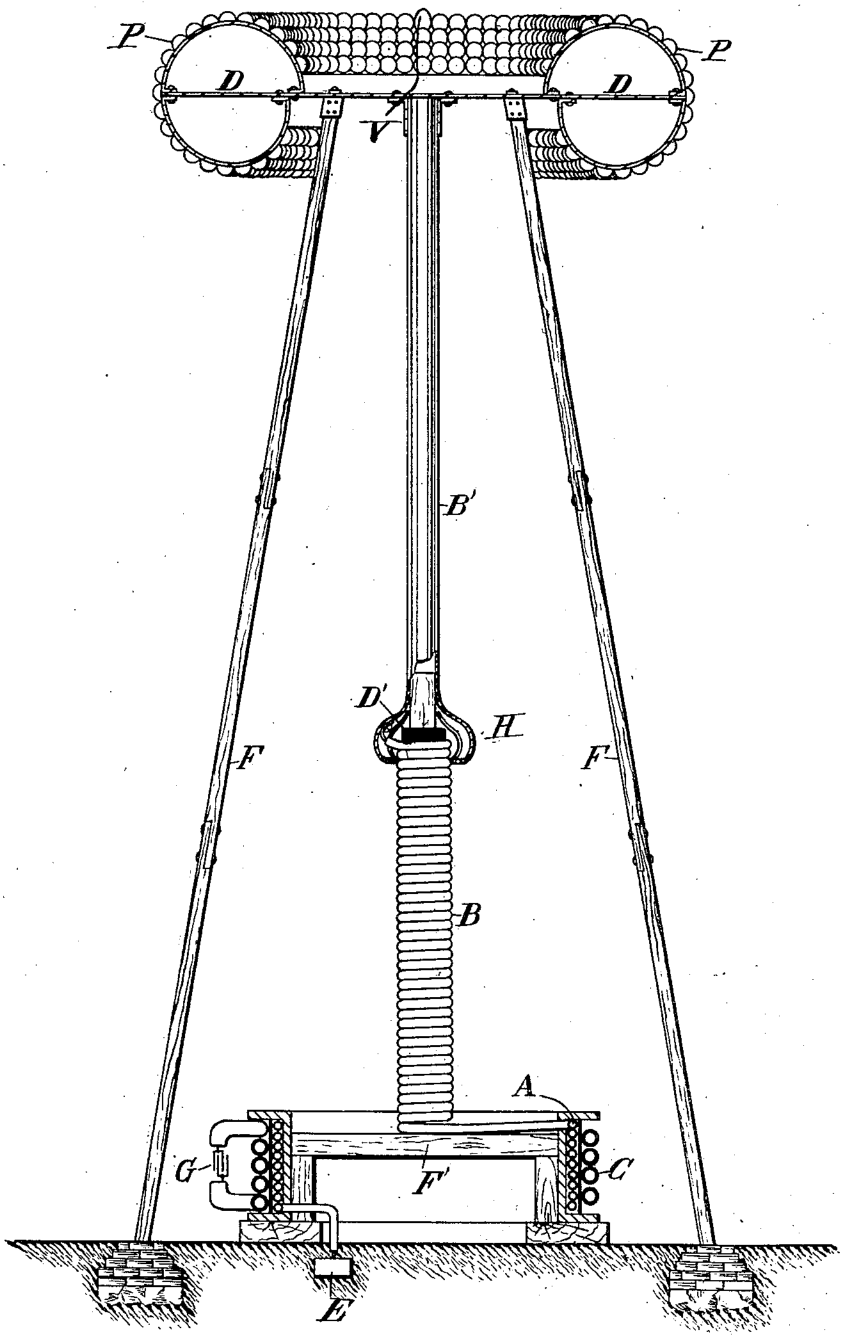

WITNESSES:

N. Lawson Dyer

Benjamin Miller

Nikola Tesla, INVENTOR,

BY Kerr, Page & Cooper,

his ATTORNEYS.

UNITED STATES PATENT OFFICE.

NIKOLA TESLA, OF NEW YORK, N. Y.

APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.

1,119,732. Specification of Letters Patent. **Patented Dec. 1, 1914.**

Application filed January 18, 1902, Serial No. 90,245. Renewed May 4, 1907. Serial No. 371,817.

To all whom it may concern:

Be it known that I, Nikola Tesla, a citizen of the United States, residing in the borough of Manhattan, in the city, county, and State of New York, have invented certain new and useful Improvements in Apparatus for Transmitting Electrical Energy, of which the following is a specification, reference being had to the drawing accompanying and forming a part of the same.

In endeavoring to adapt currents or discharges of very high tension to various valuable uses, as the distribution of energy through wires from central plants to distant places of consumption, or the transmission of powerful disturbances to great distances, through the natural or non-artificial media, I have encountered difficulties in confining considerable amounts of electricity to the conductors and preventing its leakage over their supports, or its escape into the ambient air, which always takes place when the electric surface density reaches a certain value.

The intensity of the effect of a transmitting circuit with a free or elevated terminal is proportionate to the quantity of electricity displaced, which is determined by the product of the capacity of the circuit, the pressure, and the frequency of the currents employed. To produce an electrical movement of the required magnitude it is desirable to charge the terminal as highly as possible, for while a great quantity of electricity may also be displaced by a large capacity charged to low pressure, there are disadvantages met with in many cases when the former is made too large. The chief of these are due to the fact that an increase of the capacity entails a lowering of the frequency of the impulses or discharges and a diminution of the energy of vibration. This will be understood when it is borne in mind, that a circuit with a large capacity behaves as a slackspring, whereas one with a small capacity acts like a stiff spring, vibrating more vigorously. Therefore, in order to attain the highest possible frequency, which for certain purposes is advantageous and, apart from that, to develop the greatest energy in such a transmitting circuit, I employ a terminal of relatively small capacity, which I charge to as high a pressure as practicable. To accomplish this result I have found it imperative to so construct the elevated conductor, that its outer surface, on which the electrical charge chiefly accumulates, has itself a large radius of curvature, or is composed of separate elements which, irrespective of their own radius of curvature, are arranged in close proximity to each other and so, that the outside ideal surface enveloping them is of a large radius. Evidently, the smaller the radius of curvature the greater, for a given electric displacement, will be the surface-density and, consequently, the lower the limiting pressure to which the terminal may be charged without electricity escaping into the air. Such a terminal I secure to an insulating support entering more or less into its interior, and I likewise connect the circuit to it inside or, generally, at points where the electric density is small. This plan of constructing and supporting a highly charged conductor I have found to be of great practical importance, and it may be usefully applied in many ways.

Referring to the accompanying drawing, the figure is a view in elevation and part section of an improved free terminal and circuit of large surface with supporting structure and generating apparatus.

The terminal D consists of a suitably shaped metallic frame, in this case a ring of nearly circular cross section, which is covered with half spherical metal plates P P, thus constituting a very large conducting surface, smooth on all places where the electric charge principally accumulates. The frame is carried by a strong platform expressly provided for safety appliances, instruments of observation, etc., which in turn rests on insulating supports F F. These should penetrate far into the hollow space formed by the terminal, and if the electric density at the points where they are bolted to the frame is still considerable, they may be specially protected by conducting hoods as H.

A part of the improvements which form the subject of this specification, the transmitting circuit, in its general features, is identical with that described and claimed in my original Patents Nos. 645,576 and 649,621. The circuit comprises a coil A which is in close inductive relation with a primary C, and one end of which is connected to a ground-plate E, while its other end is led through a separate self-induction coil B and a metallic cylinder B′ to the terminal D.

The connection to the latter should always be made at, or near the center, in order to secure a symmetrical distribution of the current, as otherwise, when the frequency is very high and the flow of large volume, the performance of the apparatus might be impaired. The primary C may be excited in any desired manner, from a suitable source of currents G, which may be an alternator or condenser, the important requirement being that the resonant condition is established, that is to say, that the terminal D is charged to the maximum pressure developed in the circuit, as I have specified in my original patents before referred to. The adjustments should be made with particular care when the transmitter is one of great power, not only on account of economy, but also in order to avoid danger. I have shown that it is practicable to produce in a resonating circuit as E A B B′ D immense electrical activities, measured by tens and even hundreds of thousands of horse-power, and in such a case, if the points of maximum pressure should be shifted below the terminal D, along coil B, a ball of fire might break out and destroy the support F or anything else in the way. For the better appreciation of the nature of this danger it should be stated, that the destructive action may take place with inconceivable violence. This will cease to be surprising when it is borne in mind, that the entire energy accumulated in the excited circuit, instead of requiring, as under normal working conditions, one quarter of the period or more for its transformation from static to kinetic form, may spend itself in an incomparably smaller interval of time, at a rate of many millions of horse power. The accident is apt to occur when, the transmitting circuit being strongly excited, the impressed oscillations upon it are caused, in any manner more or less sudden, to be more rapid than the free oscillations. It is therefore advisable to begin the adjustments with feeble and somewhat slower impressed oscillations, strengthening and quickening them gradually, until the apparatus has been brought under perfect control. To increase the safety, I provide on a convenient place, preferably on terminal D, one or more elements or plates either of somewhat smaller radius of curvature or protruding more or less beyond the others (in which case they may be of larger radius of curvature) so that, should the pressure rise to a value, beyond which it is not desired to go, the powerful discharge may dart out there and lose itself harmlessly in the air. Such a plate, performing a function similar to that of a safety valve on a high pressure reservoir, is indicated at V.

Still further extending the principles underlying my invention, special reference is made to coil B and conductor B′. The latter is in the form of a cylinder with smooth or polished surface of a radius much larger than that of the half spherical elements P P, and widens out at the bottom into a hood H, which should be slotted to avoid loss by eddy currents and the purpose of which will be clear from the foregoing. The coil B is wound on a frame or drum D^1 of insulating material, with its turns close together. I have discovered that when so wound the effect of the small radius of curvature of the wire itself is overcome and the coil behaves as a conductor of large radius of curvature, corresponding to that of the drum. This feature is of considerable practical importance and is applicable not only in this special instance, but generally. For example, such plates at P P of terminal D, though preferably of large radius of curvature, need not be necessarily so, for provided only that the individual plates or elements of a high potential conductor or terminal are arranged in proximity to each other and with their outer boundaries along an ideal symmetrical enveloping surface of a large radius of curvature, the advantages of the invention will be more or less fully realized. The lower end of the coil B—which, if desired, may be extended up to the terminal D—should be somewhat below the uppermost turn of coil A. This, I find, lessens the tendency of the charge to break out from the wire connecting both and to pass along the support F′.

Having described my invention, I claim:

1. As a means for producing great electrical activities a resonant circuit having its outer conducting boundaries, which are charged to a high potential, arranged in surfaces of large radii of curvature so as to prevent leakage of the oscillating charge, substantially as set forth.

2. In apparatus for the transmission of electrical energy a circuit connected to ground and to an elevated terminal and having its outer conducting boundaries, which are subject to high tension, arranged in surfaces of large radii of curvature substantially as, and for the purpose described.

3. In a plant for the transmission of electrical energy without wires, in combination with a primary or exciting circuit a secondary connected to ground and to an elevated terminal and having its outer conducting boundaries, which are charged to a high potential, arranged in surfaces of large radii of curvature for the purpose of preventing leakage and loss of energy, substantially as set forth.

4. As a means for transmitting electrical energy to a distance through the natural media a grounded resonant circuit, comprising a part upon which oscillations are impressed and another for raising the ten-

sion, having its outer conducting boundaries on which a high tension charge accumulates arranged in surfaces of large radii of curvature, substantially as described.

5. The means for producing excessive electric potentials consisting of a primary exciting circuit and a resonant secondary having its outer conducting elements which are subject to high tension arranged in proximity to each other and in surfaces of large radii of curvature so as to prevent leakage of the charge and attendant lowering of potential, substantially as described.

6. A circuit comprising a part upon which oscillations are impressed and another part for raising the tension by resonance, the latter part being supported on places of low electric density and having its outermost conducting boundaries arranged in surfaces of large radii of curvature, as set forth.

7. In apparatus for the transmission of electrical energy without wires a grounded circuit the outer conducting elements of which have a great aggregate area and are arranged in surfaces of large radii of curvature so as to permit the storing of a high charge at a small electric density and prevent loss through leakage, substantially as described.

8. A wireless transmitter comprising in combination a source of oscillations as a condenser, a primary exciting circuit and a secondary grounded and elevated conductor the outer conducting boundaries of which are in proximity to each other and arranged in surfaces of large radii of curvature, substantially as described.

9. In apparatus for the transmission of electrical energy without wires an elevated conductor or antenna having its outer high potential conducting or capacity elements arranged in proximity to each other and in surfaces of large radii of curvature so as to overcome the effect of the small radius of curvature of the individual elements and leakage of the charge, as set forth.

10. A grounded resonant transmitting circuit having its outer conducting boundaries arranged in surfaces of large radii of curvature in combination with an elevated terminal of great surface supported at points of low electric density, substantially as described.

NIKOLA TESLA.

Witnesses:
M. Lamson Dyer,
Richard Donovan.

Nov. 12, 1957

L. R. CRUMP

2,813,242

POWERING ELECTRICAL DEVICES WITH ENERGY
ABSTRACTED FROM THE ATMOSPHERE
Filed March 12, 1954

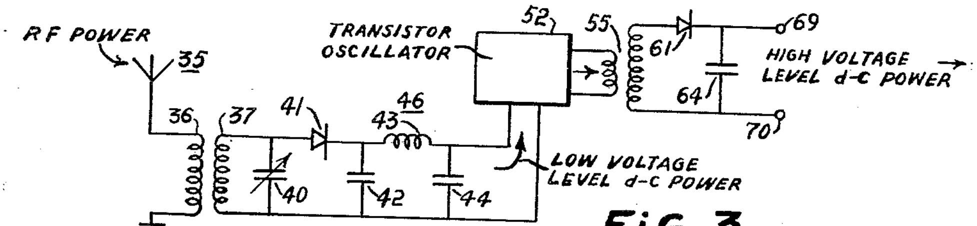
RF POWER
35
36
37
41
46
43
40
42
44
TRANSISTOR OSCILLATOR
52
55
61
64
69
70
HIGH VOLTAGE LEVEL d-c POWER
LOW VOLTAGE LEVEL d-c POWER

FiG. 3

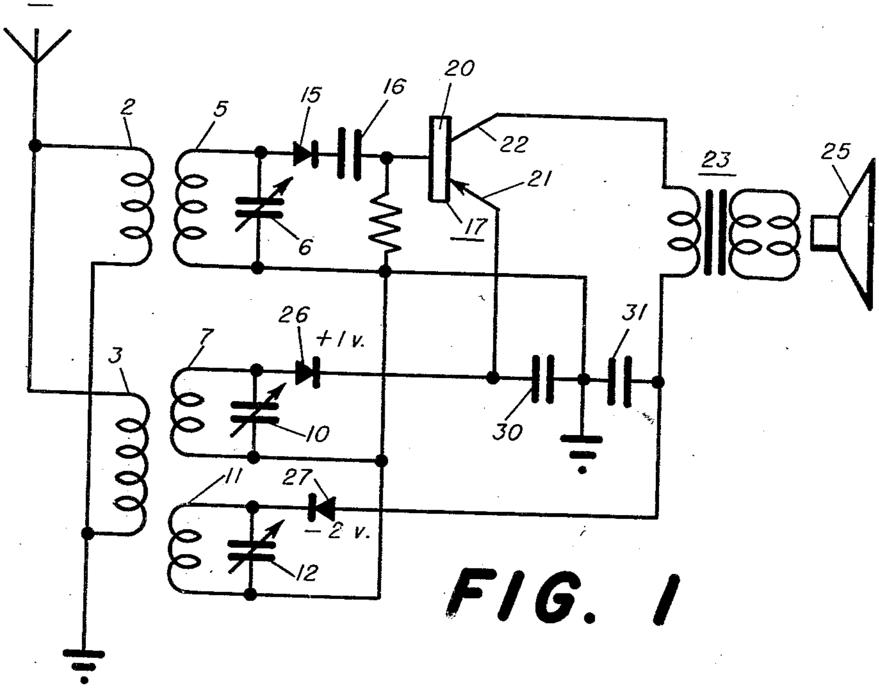
1
2
5
15
16
20
22
21
17
6
23
25
26
+1 v.
7
3
10
30
31
11
27
-2 v.
12

FIG. 1

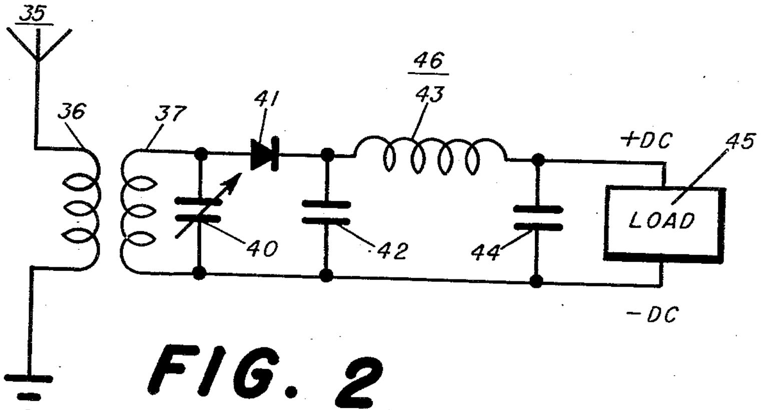
35
36
37
41
46
43
+DC
45
LOAD
40
42
44
-DC

FIG. 2

INVENTOR
Lloyd R. Crump

BY W. E. Thibodeau + A. W. Dew

ATTORNEYS

2,813,242
Patented Nov. 12, 1957

2,813,242

POWERING ELECTRICAL DEVICES WITH ENERGY ABSTRACTED FROM THE ATMOSPHERE

Lloyd R. Crump, Silver Spring, Md.

Application March 12, 1954, Serial No. 415,986
1 Claim. (Cl. 321—2)
(Granted under Title 35, U. S. Code (1952), sec. 266)

The invention described herein may be manufactured and used by or for the Government for governmental purposes without payment to me of any royalty thereon.

This invention relates to the convenient and economical provision of power for the operation of electronic circuits and devices using transistors, and of other electrical devices having modest power requirements.

A great advantage of transistors, and a major reason for their enthusiastic reception since their introduction a few years ago, is the fact that they will operate satisfactorily with very low supply voltages and currents. One milliwatt or even less is sufficient to power a transistor in many applications. Various batteries have been developed to provide, in a minimum of space, the relatively minute amounts of power needed by transistors.

My invention provides methods and means that permit transistor circuits, and also other low-powered electrical devices, to be economically and conveniently operated without any batteries whatever, and indeed without any power supply whatever as power supplies are ordinarily conceived.

The invention centers around my discovery that it is practicable to construct operative transistor circuits that are able to abstract from the atmosphere sufficient electromagnetic energy to provide all necessary supply voltages and currents for their own operation. Circuits and devices powered according to my invention will operate indefinitely without any local power source whatever.

I have successfully constructed and demonstrated such circuits. For example, I have constructed a batteryless transistor radio receiver on which I have listened to either nearby or distant broadcast stations as desired, using either headphones or a loudspeaker; this receiver has been powered entirely by electromagnetic energy abstracted from the atmosphere.

From the successful operation of this receiver, and from other experimental work, it becomes clear that, by the methods and means of the invention, a great variety of practical and useful transistor circuits can be powered entirely by energy abstracted from the atmosphere.

Furthermore, as will become apparent below, my invention is applicable to the powering of other electrical devices requiring relatively small amounts of power.

An object of the present invention is to provide methods and means for powering transistor circuits entirely from radiofrequency energy abstracted from the atmosphere.

Another object is to provide methods and means for powering remote radio receivers, low-powered radio transmitters, and other low-powered electrical devices, with energy received by radio from a master station, so that no local power supplies are needed by the devices and so that the powering or non-powering of the remote device is under the control of the master station.

A further object is to provide methods and means for powering transistor circuits and other low-powered electrical devices with radiofrequency energy received from one or more remote radio transmitters.

Other objects, aspects, uses, and advantages of the invention will become apparent from the following description and from the drawing.

Figure 1 is a schematic diagram of a transistor radio receiver in which all necessary power is supplied by energy abstracted from the atmosphere in accordance with the invention.

Figure 2 is a schematic diagram showing a general application of the invention to provide direct-current power to a load.

Figure 3 is a schematic diagram of a system for obtaining a high energy D.-C. source at a high voltage level using energy abstracted from the atmosphere.

Referring to Figure 1, a receiving antenna **1** is connected to antenna coupling coils **2** and **3**, the other ends of which are connected to ground. A parallel resonant circuit consisting of coil **5** and capacitor **6** is coupled to coil **2**. A second parallel resonant circuit consisting of coil **7** and capacitor **10** is coupled to coil **3**. A third parallel resonant circuit consisting of coil **11** and capacitor **12** is also coupled to coil **3**.

Coil **5** and capacitor **6** are tuned to the frequency of a radio transmitter from which it is desired to receive information—for instance, an amplitude-modulated standard broadcast station. The signal received from this transmitter need not be strong. The signal is detected by diode **15** to obtain an audio-frequency information signal. This audio signal is coupled through a capacitor **16** and is amplified by a circuit that includes a transistor **17** having a base **20**, an emitter **21**, and a collector **22**. The amplified audio output of the transistor is coupled through an audio transformer **23** to an electroacoustical transducer, preferably a permanent-magnet dynamic loudspeaker **25** as shown.

The novelty of the invention lies largely in the method and means by which the necessary direct-current power is supplied to the emitter and collector circuits of transistor **17**. This method and means will now be described.

Coil **7** and capacitor **10**, and also coil **11** and capacitor **12**, are tuned to receive radio signals of relatively high strength. It does not matter whether these signals contain information. These power signals are rectified by diodes **26** and **27** to provide direct-current power that is filtered by capacitors **30** and **31**. The D.-C. power thus obtained is utilized to power the transistor **17**.

In the circuit shown, two tuned circuits (coil **7** and capacitor **10**, and coil **11** and capacitor **12**) are tuned to power signals and the D.-C. voltages obtained from each are connected in series. The tuned power circuits may be tuned to the same or different power signals. Under certain circumstances it may be desirable to use more than two tuned power circuits and to tune them to more than two power signals; in this way power can be obtained from several signals and combined. On the other hand, if a strong power signal is available, a single tuned power circuit may suffice to give the needed D.-C. power.

Even weak information signals can be received successfully. A plurality of transistor amplifier stages can be used if desired, or other circuits such as superheterodyne circuits can be used. It is merely necessary that a sufficiently strong power signal or signals be available to provide the small amount of D.-C. needed to power the transistors.

If the information signal happens to be strong, it can be used as the power signal; all of the tuned circuits (coil **5** and capacitor **6**, coil **7** and capacitor **10**, coil **11** and capacitor **12**) are tuned to the information signal.

Engineers who have observed my invention in operation have been surprised at the unexpectedly good results obtained, even with readily available power signals of quite moderate strength. For instance, sufficient power for satisfactory operation of a loudspeaker at low volumes is readily obtained from a 5-kilowatt standard broadcast station 5 miles away, using only an indoor antenna to pick up the power signal as well as information signals. In typical operation under these conditions a D.-C. voltage of about 2.5 to 3 volts is obtained between the emitter and the collector, at a current of about 250 microamperes; D.-C. power input to the transistor is thus of the order of 0.5 to 1 milliwatt. So far as I am aware, no one has ever before discovered and demonstrated the practicability of this method of powering a radio receiver.

Because existing broadcast stations within a radius of a number of miles provide adequate power signals, the invention is readily practicable with existing power signals in almost any location in or near any city in the United States.

Although I have described a transistor radio receiver powered by my invention, it will be readily apparent that the invention is applicable to the powering of any transistor circuit using one or a number of transistors, and to the powering of other devices requiring relatively small amounts of power. For instance, sensitive electromechanical, electrochemical, or electrothermal devices can be operated by the method of the invention.

Referring to Figure 2, which shows a more general embodiment of my invention, an antenna **35** picks up radiofrequency energy from the atmosphere. This energy flows through coil **36**, which is coupled to a tuned circuit consisting of coil **37** and capacitor **40**. The radiofrequency voltage across capacitor **40** is rectified by diode **41** and filtered by a low-pass filter **46** consisting of capacitors **42** and **44** and choke coil **43**. The resulting D.-C. voltage is applied to a load **45**.

In the practice of my invention, larger amounts of power can be obtained for short periods of time by storing received energy in a suitable energy storage device. Stored energy may then be withdrawn at intervals at a more rapid rate than that at which it was received and put into the storage device. In this way the invention can be used to provide short pulses of relatively very high electrical energy. This result can be readily obtained by charging a relatively large capacitor with direct current and then discharging the capacitor rapidly into a load when desired. This rapid discharge can be initiated automatically when the voltage across the capacitor reaches a certain level, or it can be initiated when a transistor radio receiver receives a certain information signal.

Higher voltages can be obtained with the invention by means of well known devices for raising D.-C. voltages as shown in Figure 3. The D.-C. voltage output from the capacitor **44** can be used to power a low frequency transistor oscillator **52** whose A.-C. output is raised to a higher voltage level by the transformer **55**. This relatively high A.-C. voltage can then be rectified by a diode **61** and fed to a capacitor **64** to provide a high energy D.-C. source at a relatively high voltage level at the terminals **69** and **70**. If desired, energy can now be withdrawn from the capacitor **64** at intervals in short pulses of high energy at a high voltage level. Pulsed radio transmission is one of the possible uses for this form of the invention. Other uses would be to provide a single relatively powerful pulse needed to actuate an electrothermal or electromechanical device.

As has been indicated above, in many locations and particularly anywhere in or near most American cities, power signals normally present in the atmosphere are readily available for the easy and convenient practice of the invention. However, the invention also has important applications in systems in which the necessary power signal is generated and transmitted specifically for the operation of the particular system. Such systems can, for example, comprise a master station transmitting all the power that is needed for hundreds or thousands of fixed or mobile transistor receivers or other remote devices over

a range of many miles. This eliminates the need for hundreds or thousands, as the case may be, of local power supplies. At the same time, such a system has the advantage that all of the remote devices can be simultaneously activated or deactivated at the will of the master station, simply by starting or stopping the transmission of the power signal. In such systems it will often be advantageous to use power signals of frequencies sufficiently high to permit the use of resonant receiving antennas of small physical dimensions for signal pickup at the remote devices. In addition to the power signal, the master station may transmit an information signal on the same or a different carrier.

Certain types of devices powered entirely by received radio waves are of course well known. The well-known "crystal set" of the early days of radio, which used a diode rectifier to demodulate an amplitude-modulated radiofrequency signal, is an outstanding example of such a device. My invention is readily distinguishable from such prior devices, however. In typical prior devices a modulated radiofrequency signal is applied to a diode to obtain unidirectional half-wave pulses whose amplitudes vary with modulation. These pulses are integrated by means of a capacitor to obtain a unidirectional signal the amplitude of which follows the audiofrequency modulation envelope. If the radiofrequency signal is received with sufficient strength the audio signal may have sufficient power to operate headphones or similar utilization device without power amplification; but the signal is utilized for its information content, rather than to supply non-information-containing power.

My invention, on the other hand, entails the utilization of received radiofrequency energy to supply power to at least one pair of circuit points (across capacitor **31** in Fig. 1, for example), such circuit points requiring power solely for its power content and not for any information or modulation it may contain. In other words, my invention entails the utilization of radiofrequency energy to supply power that would otherwise have to be supplied by batteries, generator, or other local power source.

It will be apparent that the embodiments shown are only exemplary and that various modifications can be made in construction and arrangement within the scope of the invention as defined in the appended claim.

I claim:

An electrical device for obtaining a high energy D.-C. source at a high voltage level using energy abstracted from the atmosphere, said device comprising in combination: resonant means for receiving radio waves, first rectifier means for converting said radio waves into first direct current energy, first capacitor means for storing said first direct current energy, an oscillator powered by said direct current energy, said oscillator producing an A.-C. output, transformer means for raising said A.-C. output to an increased voltage level, second rectifier means for converting the A.-C. output of increased voltage level from said transformer into second direct current energy, and second capacitor means for storing said second direct current energy.

References Cited in the file of this patent

UNITED STATES PATENTS

1,540,998	Plauson	June 9, 1925
2,056,824	Cawley	Oct. 6, 1936
2,553,734	Adler	May 22, 1951
2,569,687	Pan	Oct. 2, 1951
2,647,957	Mallinckrodt	Aug. 4, 1953

OTHER REFERENCES

Turner: "A crystal receiver with transistor amplifier," Radio and Television News, January 1950.

FIP8102 XR 3,273,066

SEARCH ROOM

4

X5264R

Sept. 13, 1966 **L. H. RUHNKE** SD **3,273,066**

APPARATUS FOR DETECTING CHANGES IN THE ATMOSPHERIC ELECTRIC FIELD

Filed Dec. 20, 1963 2 Sheets-Sheet 1

Fig. 1

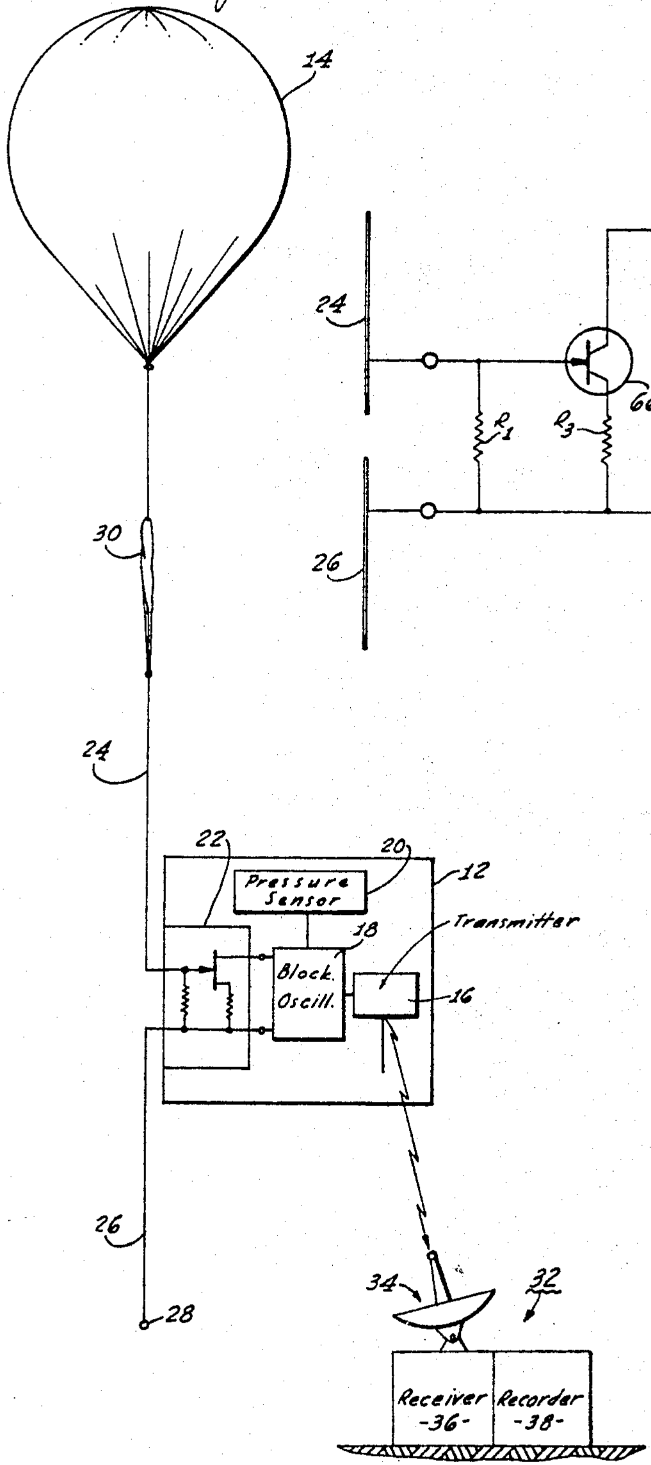
14
30
24
22
20
12
Pressure Sensor
18
Transmitter
Block. Oscill.
16
26
28
34
32
Receiver -36-
Recorder -38-

Fig. 3

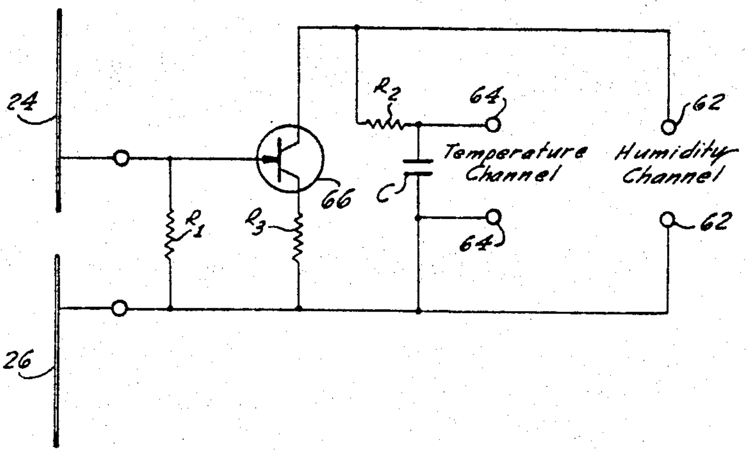
24
26
R_1
R_2
R_3
64
64
62
62
Temperature Channel
Humidity Channel
C
66

INVENTOR:
Lothar H. Ruhnke

By Alan C. Rose
Attorney

Sept. 13, 1966 L. H. RUHNKE 3,273,066

APPARATUS FOR DETECTING CHANGES IN
THE ATMOSPHERIC ELECTRIC FIELD

Filed Dec. 20, 1963 2 Sheets-Sheet 2

Fig. 2

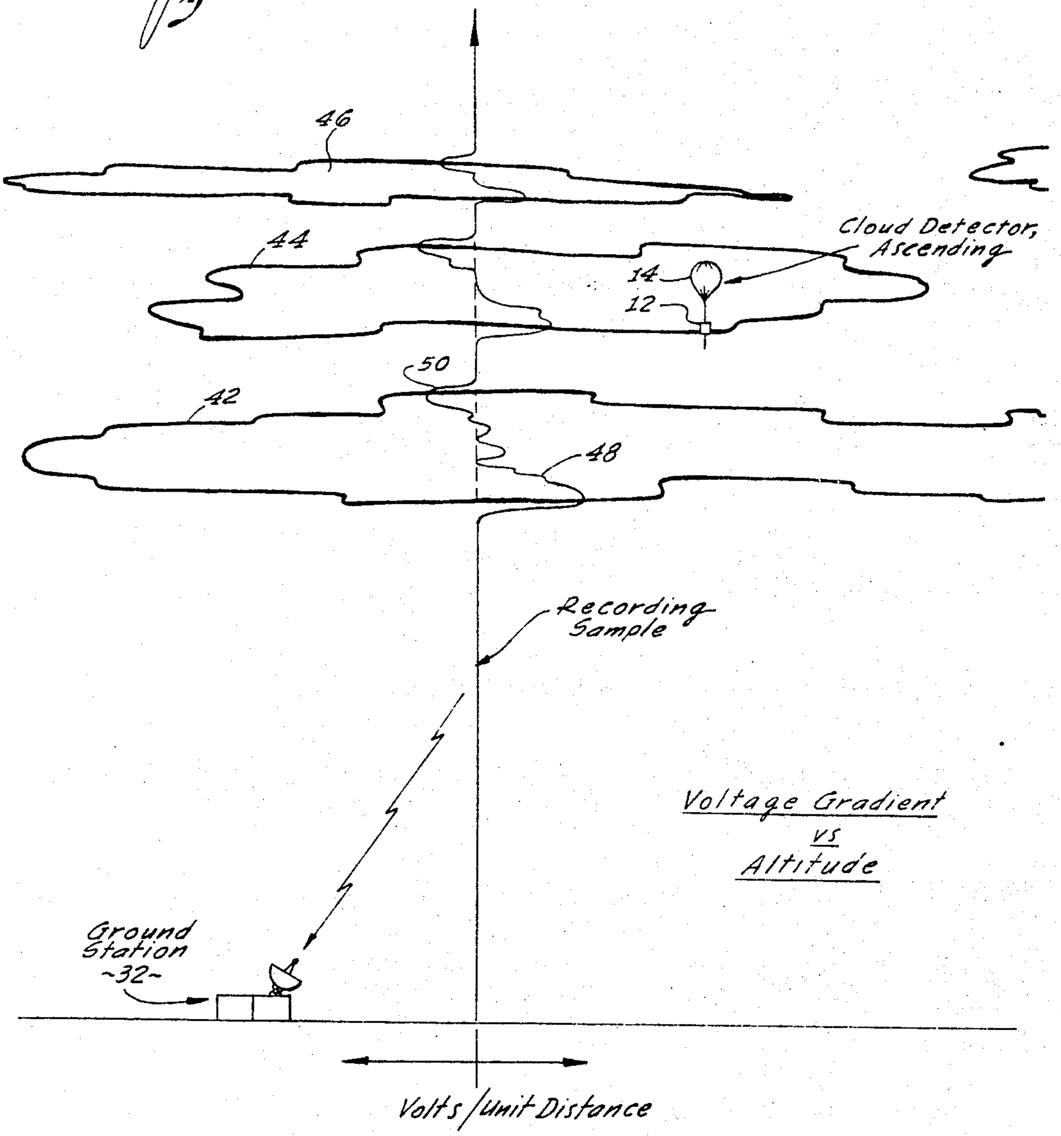
46
44
Cloud Detector, Ascending
14
12
50
42
48
Recording Sample
Voltage Gradient vs Altitude
Ground Station ~32~
Volts/Unit Distance

INVENTOR:
Lothar H. Ruhnke

By Alan C. Rose
Attorney

United States Patent Office

3,273,066

Patented Sept. 13, 1966

3,273,066
APPARATUS FOR DETECTING CHANGES IN THE ATMOSPHERIC ELECTRIC FIELD

Lothar H. Ruhnke, Minneapolis, Minn., assignor to Litton Systems, Inc., Beverly Hills, Calif.

Filed Dec. 20, 1963, Ser. No. 332,186
4 Claims. (Cl. 325—113)

This invention relates to a cloud detector or an apparatus for sensing electrostatic conditions in the earth's atmosphere.

Cloud detection apparatus which is presently employed is usually of the "ceilometer" type which merely measures the height of the lower surface of the clouds to indicate the ceiling conditions. One object of the present invention is to determine the location of both the top and the bottom of one or more cloud formations of the atmosphere.

While electrostatic sensors of one type or another have been proposed heretofore for use in the atmosphere, they have been so complex and expensive that they have not enjoyed widespread use. In addition, cloud radars have recently been developed, but their usefulness is limited by failures when different zones of the atmosphere contain droplets of different sizes, as discussed more fully below. Accordingly, it is another object of the present invention to simplify and reduce the cost of cloud detectors, while maintaining reliability, compatibility with existing equipment, and adequate sensitivity for all practical purposes.

In accordance with an illustrative embodiment of the present invention, a conventional U.S. Weather Bureau or U.S. Army meteoroligical radiosonde transmitter may be provided with a modulator which is controlled by a field effect transistor. The input to the field effect transistor is a dipole antenna. Each of the arms of the dipole antenna may be five to fifty feet in length, for example, and one of the dipole elements may trail the radiosonde package while the other element may extend forward in the direction of movement of the unit. When a balloon is employed to carry the cloud detector unit aloft, a weight may be attached to one of the antenna elements so it hangs below the instrument package, while the other antenna element can extend along, or even form the flexible connection between the radiosonde unit and the balloon. A receiving unit is provided to pick up the modulated signals from the radiosonde apparatus, demodulate these signals and record them. While it is contemplated that the receiver may normally be situated on the ground and receive signals from a rising radiosonde unit, the radiosonde unit could well be employed by an aircraft to determine conditions below a cloud layer. Under these conditions the receiver would be mounted on the plane and the radiosonde unit could be parachuted to the ground.

In accordance with a feature of the present invention, therefore, a system for detecting electrostatic field conditions in the atmosphere includes a transmitter, arrangements for moving the transmitter through the atmosphere, an antenna extending a substantial distance from the transmitter, and a field effect transistor connected to receive electrostatic input signals from said antenna and to modulate the output of the transmitter in accordance with voltages detected by the antenna.

In accordance with one aspect of the present invention, simplicity of circuitry and stability of operation are achieved by the use of dynamic circuitry which measures the instantaneous difference in potential between two antenna elements. Thus, a high input resistor may be connected between the input antenna elements. With this type of arrangement the instrument package must be continuously moved through the atmosphere in order to give meaningful results. Under these conditions, the recorded output indicates changes in electric fields. A constant electric field, corresponding to a linear change in potential produces no substantial output signal. It has been determined that a balloon velocity of about five meters per second will produce good results with a dipole antenna shunted by a 1,000 megohm resistor and connected to a field effect transistor as discussed in detail in the present specification. With lower velocities, a higher input impedance or greater amplification would be required, with higher velocities, such as rocket propulsion or the like, permitting even lower input impedances and simpler input circuits.

Accordingly, it is a feature of the invention that apparatus is provided for moving a radiosonde unit through the atmosphere at a predetermined rate, that two element input antenna preferably of the dipole type is provided, and that the antenna is connected to the radiosonde by simple dynamic amplification circuitry, which requires continuous movement of the apparatus to give useful output information.

The novel features which are believed to be characteristic of the invention both as to its organization and method of operation, together with further objects and advantages thereof, will be better understood from the following description considered in conjunction with the accompanying drawings in which a typical embodiment of the invention is illustrated by way of example. It is to be expressly understood, however, that the drawings are for the purpose of illustration and description only and are not intended as a definition of the limits of the invention.

In the drawing:

FIG. 1 is a schematic block diagram of a radiosonde equipment for sensing changes in the electric field of the atmosphere, and thus detecting clouds, in accordance with the present invention;

FIG. 2 is a diagrammatic showing of the record obtained under typical cloud conditions; and

FIG. 3 is a circuit diagram indicating a typical arrangement of connections of a dipole antenna, a field effect transistor, and the input terminals of a standard meteorological radiosonde apparatus.

With reference to the drawings, FIG. 1 shows a radiosonde package **12** supported by a balloon **14**. The radiosonde unit includes a transmitter **16**, blocking oscillator **18**, pressure sensor **20**, and a field effect transistor unit **22**. A dipole antenna having an upper element **24** and a lower element **26** is connected to the input of the field effect transistor unit **22**. A small weight **28** is secured to the lower end of the antenna element **26** to hold it in the extended position. Similarly, the upper antenna element **24** extends in the other direction, and it may either form the support for the radiosonde unit **12** or may be closely associated with the supporting cable. A parachute **30** is provided for safe return of the instrument package to ground.

A ground or receiving station **32** may include an antenna **34**, receiver and demodulator **36**, and a recording apparatus **38**.

FIG. 2 shows a typical recording of the electrostatic field obtained as the ballon **14** and instrument package **12** rise through the successive layers of clouds **42**, **44**, and **46**, along the center line of the drawing. It is known that the electrostatic potential in the atmosphere increases from ground potential to plus 300,000 volts at elevations of about 30 miles. It is also known that the atmospheric electric field, which is the gradient of the potential, is substantially increased within clouds. As the package enters the lower edge of a cloud, the sudden increase in electric field causes a pulse of current to flow through resistor R_1 in one direction, and upon leaving the cloud, the sudden reduction in electric field causes an oppositely directed current pulse. This produces the positive swing shown at **48** in FIG. 2 as the unit enters the cloud and the negative peak **50** as the apparatus is rising through the upper surface of the cloud stratus **42**. A similar pattern is observed as the unit goes through the higher cloud formations **44** and **46**. With a knowledge of the speed of ascent of the balloon and instrument package, the location of the upper and lower surfaces of successive cloud layers may be determined by the present apparatus. Accordingly, despite the presence of a relatively low cloud layer which completely obscures other cloud formations, ground personnel may readily determine the meterological conditions relating to clouds above this low ceiling.

Returning to FIG. 1, the radiosonde equipment represented by blocks **16**, **18** and **20** is well known. The meterological radiosonde system which will be taken as typical is designated by the military by the numbers AN/AMT 4B. The Army technical manual which describes this radiosonde system is identified as TM 11–2432A–TO 31M4–2AMT–11, dated June 26, 1958. The radiosonde transmitter operates at a frequency of 1680 megacycles per second. This transmitter is modulated by a blocking oscillator **18**, the frequency of the blocking oscillator being controlled by the meteorological conditions which are detected. Normally, a thermistor for measuring the temperature and a humidity sensor are alternately switched to the input terminals of the blocking oscillator **18**. In the present case, the cloud detector circuitry is connected to both of these terminals. The barometric pressure sensor **20** may also be retained in the unit, and it may be connected to control the frequency of the blocking oscillator **18** by a switching action as described below.

With reference to FIG. 3, the humidity channel input terminals **62** and the temperature channel input terminals **64** of the standard radiosonde unit are shown at the right-hand side of the circuit. Atmospheric electric signals received by the dipole antenna elements **24** and **26** are applied to the field effect transistor **66** and are then coupled by a suitable circuit to the input terminals **62** and **64** of the radiosonde. While other field effect transistors may be employed, type C 610 of Crystallonics, Inc., was found to be suitable from a sensitivity and reliability standpoint for the present purposes. The input to the field effect transistor is shunted by a 1,000 megohm resistor R_1 to stabilize the input characteristic. Suitable biasing of the transistor **66** is provided by resistor R_3. Because the impedance of the transistor **66** is substantially lower than that of the temperature and humidity sensors which are normally employed, a 110,000 ohm resistor in the blocking oscillator was removed and a capacitor C was connected in parallel to the input terminals. The capacitor was selected to produce a modulation frequency of approximately 100 cycles per second in the absence of an input signal. Positive and negative voltages will then raise or lower this frequency.

In the radiosonde apparatus, signals from the pressure sensor **20** are normally employed to switch from the temperature channel terminals **64** to the humidity channel terminals **62**. This switch is well known in the art as a baroswitch. The barometric pressure switching action provides an indication of altitude, so it is useful to preserve this information. Accordingly, the resistor R_2 is employed between the transistor **66** and one terminal **64** to reduce the output signal slightly, thus changing the modulation level and preserving the pressure and thus altitude information.

With regard to the length of the dipole elements **24** and **26**, the size of resistor R_1 and the amplification of the field effect transistor, these are fixed by several factors. First, the antenna elements should be large enough to provide a substantial input signal, but should be relatively small as compared with normal cloud thicknesses. Lengths of 5 to 50 feet for each dipole would be of the proper order of magnitude. The resistor R_1 was chosen to reduce the input impedance between dipoles **24** and **26** below the point where fog and dampness and the resultant shunt impedance will introduce instability into the transmitted signals. With these parameters, and a speed of ascent of the order of 5 meters per second, it was determined that a single field effect transistor provided the necessary amplification for operation with the standard radiosonde unit.

As mentioned generally above, cloud radars are subject to failure under certain atmospheric conditions. Thus, the radar return is proportional to the sixth power of the droplet or particle size, while the human eye and optical systems respond to the second power of the droplet size. With the radar return proportional to such a high power of the droplet diameter, if a radar is adjusted to pick up a cloud having large particles, other clouds having small size particles are not detected. The present electrostatic system, however, responds in a manner which is nearly the same as the optical response, and therefore does not fail under atmospheric conditions where different size droplets or particles are present in successive clouds.

It is to be understood that the above described arrangements are illustrative of the application of the principles of the invention. Numerous other arrangements within the scope of the invention may be devised by those skilled in the art. Thus, by way of example and not of limitation, a somewhat different shape of antenna could be employed to pick up electrostatic field variations and apply them to the field effect transistor, an inherently stabilized transistor may be used, and the instrument package may be carried by rocket or other suitable propulsion arrangements than the balloon shown in the drawings. Accordingly, it is evident that various changes may be made in the present invention without departing from the spirit of the invention as defined in the present claims.

What is claimed is:

1. An apparatus for sensing electrostatic conditions in the atmosphere comprising:

a radiosonde apparatus including means for modulating the output frequency thereof, said means including an R-C circuit;

a pair of probes positioned in respectively different locations in the atmosphere with respect to said radiosonde;

a field effect transistor having its drain and gate terminals connected across said pair of probes and having its source and drain terminals connected in series with the resistance of said R-C circuit to modulate said radiosonde; and

a stabilizing high resistance connected across the two probes and across said gate and drain terminals of said transistor.

2. In an apparatus for sensing atmospheric electrostatic field conditions;

a balloon;

an instrument package connected to and spaced from said balloon, said instrument package including a transmitter, means including an R-C circuit for modulating said transmitter, and a field effect transistor having source and drain terminals connected in series with the resistance in said R-C circuit to modulate said transmitter;

antenna means including an antenna and resistance means connected across said antenna, said resistance means also being connected across the gate and the drain of said field effect transistor, said antenna extending upward from said package toward said balloon and downwardly from said package for applying signals to said field effect transistor in response to said field conditions to render said modulating means responsive to said field conditions.

3. An apparatus for sensing changes of the atmospheric electric field at cloud boundaries, which comprises:

a blocking oscillator-type modulated radiosonde including fixed resistance means and capacitance means connected in parallel with said fixed resistance means, said capacitance means being effective to discharge through said resistance means to control the modulation of said radiosonde;

dipole antenna means responsive to changes of the atmospheric electric field for producing a signal indicative of said changes;

resistance means connected across said dipole antenna means; and

a field effect transistor having source and drain terminals connected in series with said fixed resistance means and a gate terminal connected to said dipole antenna means for reception of said signal, said field effect transistor being responsive to said signal for varying the rate of discharge of said capacitance means to cause the blocking oscillator to modulate said radiosonde in accordance with the second changes in the atmospheric electric field.

4. An apparatus for sensing changes of the atmospheric electric field at cloud boundaries, which comprises:

radiosonde means including a blocking oscillator having fixed resistance means and capacitance means connected in parallel with said fixed resistance means, said capacitance means being effective to discharge through said resistance means to control the frequency of oscillation of said blocking oscillator;

dipole antenna means responsive to changes of the atmospheric electric field for producing a signal indicative of said changes;

resistance means connected across said dipole antenna; and

a field effect transistor having source and drain terminals connected in series with said fixed resistance means and a gate terminal connected to said dipole antenna means for reception of said signal, said field effect transistor being responsive to said signal for varying the rate of discharge of said capacitance means to control the frequency of oscillation of said blocking oscillator in accordance with the sensed changes in the atmospheric electric field.

References Cited by the Examiner

UNITED STATES PATENTS

2,463,527	3/1949	Dunmore ----------	325—113
3,038,154	6/1962	Zworykin et al. --	325—113 XR
3,056,888	10/1962	Atalla -------------	307—88
3,062,043	11/1962	Marsh et al. -------	340—193.2

DAVID G. REDINBAUGH, *Primary Examiner.*

JOHN W. CALDWELL, *Examiner.*

United States Patent [19]
Phillips, Sr.

[11] **Patent Number:** **4,685,047**
[45] **Date of Patent:** **Aug. 4, 1987**

[54] APPARATUS FOR CONVERTING RADIO FREQUENCY ENERGY TO DIRECT CURRENT

[76] Inventor: **Raymond P. Phillips, Sr.,** 1200 S. Frances, Terrell, Tex. 75160

[21] Appl. No.: **886,138**

[22] Filed: **Jul. 16, 1986**

[51] **Int. Cl.**^{4} **H02M 7/06**
[52] **U.S. Cl.** **363/126;** 307/151
[58] **Field of Search** 363/125, 126; 307/151

[56] **References Cited**

U.S. PATENT DOCUMENTS

649,621	11/1901	Tesla	455/39
685,957	5/1902	Tesla	455/619
3,434,678	3/1969	Brown et al.	307/151
3,445,767	5/1969	Beard	324/78
3,678,365	8/1972	Fremiot	307/151
3,723,891	3/1973	Whitely	307/246
3,760,257	9/1973	Fletcher et al.	307/151
3,784,845	1/1974	Haas	307/519
3,792,290	2/1974	Brocker	307/522
3,944,935	3/1976	Plant	307/519
4,079,268	3/1978	Fletcher et al.	307/151
4,079,325	3/1978	Mawhinney et al.	307/261
4,085,377	4/1978	Turski et al.	307/519
4,195,237	3/1980	Meslener	307/261
4,297,557	10/1981	Tyler et al.	219/10.55 B
4,360,741	11/1982	Fitzsimmons et al.	307/151
4,360,865	11/1982	Yasumura et al.	363/126
4,393,441	8/1983	Enge	363/126

Primary Examiner—Patrick R. Salce
Assistant Examiner—Jeffrey Sterrett
Attorney, Agent, or Firm—Millen & White

[57] ABSTRACT

Apparatus for and methods of converting radio frequency energy into direct current for generating electric power includes a dipolar antenna for receiving radio frequency energy and a circuit connected thereto for converting the radio frequency energy to direct current. The circuit has a positive output line connected to one pole of the antenna and a negative output line connected to the other pole of the antenna. A positive transmitting diode is in the positive output line and a negative transmitting diode is in the negative output line. First and second bus lines and a pair of tuned circuits of opposite polarity couple the positive output line and negative line to the bus line with one of the bus lines being connected to ground. Each tuned circuit includes a first bridging line connecting the positive output line to the first and second ground lines and a second bridging line connecting the negative output line to the first and second ground lines. Each bridging line has a diode therein oriented at a polarity which is reverse with respect to the input diode. The bridging lines of each tuned circuit are connected to one another by an inductance and have capacitors disposed between the diode and the bus lines. A direct current device is connected to the positive line of the circuit.

11 Claims, 1 Drawing Figure

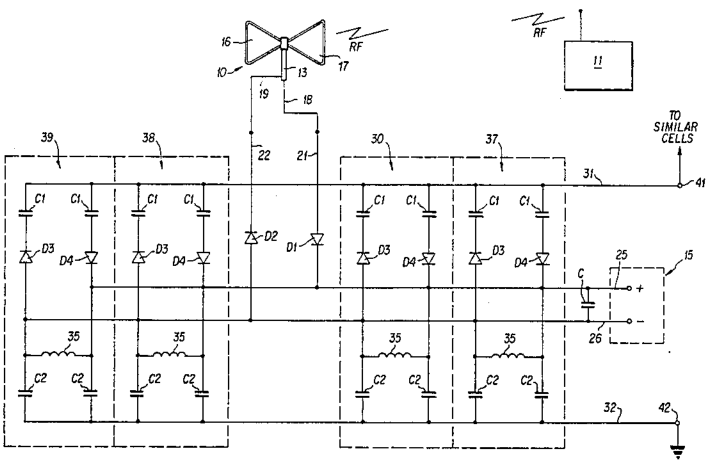

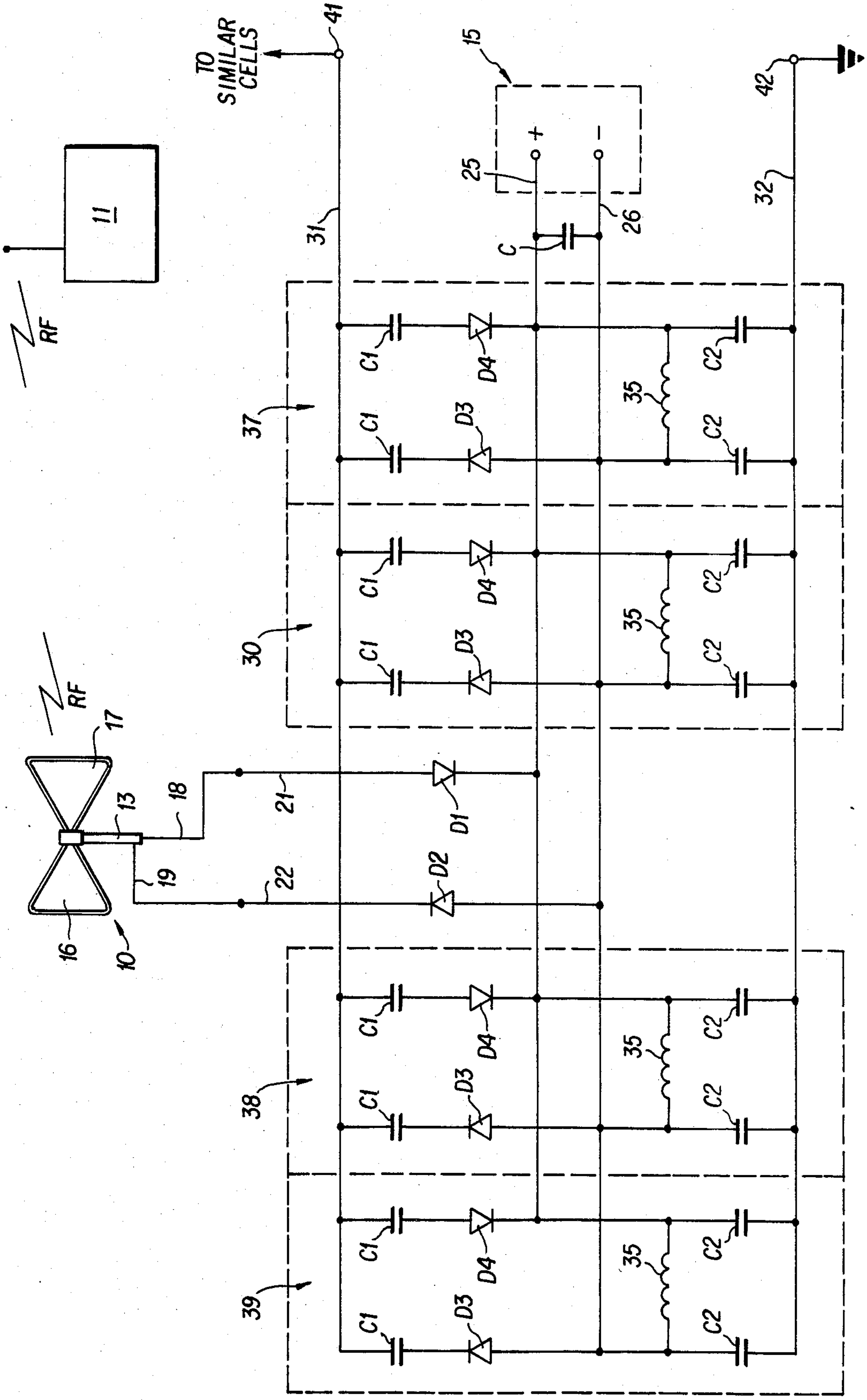
TO SIMILAR CELLS
+
-
C
RF
C1
C2
D3
D4
RF
D1
D2

APPARATUS FOR CONVERTING RADIO FREQUENCY ENERGY TO DIRECT CURRENT

BACKGROUND OF THE INVENTION

1. Field of the Invention

The instant invention relates to methods of an apparatus for converting radio frequency energy to direct current; more particularly, the instant invention relates to methods and apparatus for converting radio frequency energy to direct current wherein the direct current is of sufficient magnitude to power devices such as battery charges and electric motors without the use of amplification.

2. Technical Considerations in Prior Art

There has long been interest in technology directed to transmitting electrical energy over a distance without utilizing conductors, such as wire conductors. Development of such a technology has enormous potential. This was first recognized by Nikola Tesla who in 1899 constructed a 200 foot Tesla coil rated at 300 kilowatts and 150 kilocycles. Tesla hoped to set up standing waves of electrical energy around the whole surface of the earth, so that receiving antennas set at optimum points could tap the power when needed. Tesla was able to light hundreds of lamps at a distance of about 40 kilometers with his device without utilizing electrical conductors. The scheme has generally remained a scientific curiosity but has provided the initial groundwork for current developments wherein attempts are being made to transmit power by microwaves. However, power transmitted by microwaves is envisioned in the form of a beam of very high intensity which is focused from a microwave generator to a receiving antenna. This technology is envisioned as being used for many types of purposes such as transmitting microwave energy collected from gigantic solar power satellites and "star wars"-type weapons systems. However, the focused microwave beam is not suitable for many applications in that the beam must be directed toward a receiving antenna and cannot be transmitted through most objects, including living objects, without destroying the objects.

The instant invention relies on converting energy from standing waves which are emitted from radio frequency antennas in the RF range rather than the microwave range. Of particular interest are very low frequencies which are not used in communications and are available for transmitting power. Also of interest with respect to the instant invention are the low frequency waves emitted by the earth due to pulsation thereof caused by its magnetic field. These low frequency standing "earth" waves can be picked up by receivers tuned thereto.

SUMMARY OF THE INVENTION

It is an object of the invention to provide new and improved methods of an apparatus for converting radio frequency currents to direct current for practical uses other than communications, wherein the direct current energy converted from the radio frequency input energy does not require amplification.

Upon further study of the specification and appended claims, further objects and advantages of this invention will become apparent to those skilled in the art.

The instant invention contemplates an RF antenna for receiving radio waves. The RF antenna connected to a circuit configured in accordance with the principles of the instant invention to convert the RF signals to direct current. The radio frequency signals received by the antenna are transmitted to first and second leads, with one lead being rectified to transmit positive voltage and the other lead being rectified to transmit negative voltage. The positive voltage lead being connected directly to a positive output line and the negative voltage lead being connected directly to a negative output line. The positive output line is connected to a pair of bus lines through a first pair of capacitors, while the negative output line is connected to the pair of bus lines by a second pair of capacitors. Disposed between the first bus line and the positive output line is a reverse diode of negative polarity, while disposed between the negative output line and first bus line is a reverse diode of positive polarity. The positive and negative output lines are connected to one another through an inductance which is in parallel with the capacitors of the first and second pair connected between the second bus line and the positive and negative output lines.

In accordance with one embodiment of the invention the afore-described circuit is duplicated for each positive and negative output line. In accordance with another embodiment of the invention, the afore-described circuitry is coupled to additional circuits identically configured in order to increase the direct current output of the arrangement.

In accordance with a further configuration of the invention, the antenna utilized is a dipolar antenna of aluminum wire arranged in a "butterfly" configuration.

The instant invention further contemplates the method of utilizing the afore-described elements so as to generate direct current having sufficient power to perform tasks such as charging batteries, lighting lamps and powering direct current electric motors without the use of amplifiers.

BRIEF DESCRIPTION OF THE DRAWING

Various other objects, features and attendant advantages of the present invention will be more fully appreciated as the same becomes better understood when considered in connection with the accompanying drawing, in which like reference characters designate the same or similar parts throughout the several views, and wherein:

There is shown a diagram of a circuit in accordance with the instant invention in combination with a driven device and a dipolar antenna which receives radio frequency waves which are converted to DC current for powering the driven device.

DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring now to the drawing, there is shown a dipolar antenna, designated generally by the numeral **10**, which receives radio frequency waves from an RF transmitter **11**. The radio frequency waves are transmitted to a radio frequency to direct current converting circuit, designated generally by the numeral **12**, through a coaxial cable **13** and DC current from the circuit **12** is used to power an output device, designated generally by the numeral **15**, which may for example by a battery charger DC motor or lighting device. The circuit **12** has no other power inputs other than the radio frequency energy transmitted thereto by the antenna **10** and therefore includes no amplifiers for amplifying the RF energy.

The source of radio frequencies convertible to direct current by the circuit shown may include sources of high frequency, low frequency (LF), very low frequency (VLF) and extremely low frequency (ELF) radio waves as well as seismic vibration of the earth's magnetic fields.

Preferably, the dipolar antenna **10** is formed of two triangular loops of aluminum wire **16** and **17**, one of which is connected to the annular conductor **18** of the coaxial cable **13** and the other of which is connected to the center conductor **19** of the coaxial cable. The size of the bipolar antenna **10** is dependent on the particular application to which it is put. In accordance with one embodiment of the invention, the antenna **10** is approximately 12 inches in width and 18 inches in length. Such an antenna is used to receive five watt energy, such as that generated by a walkie-talkie or citizen-band radio.

The annular conductor **18** of the coaxial cable **13** is connected to a positive lead **21** of the circuit **12**, while the center conductor **19** of the coaxial cable is connected to a negative lead **22** of the circuit. A positive transmitting diode D1 is disposed between the lead **21** and the remainder of the circuit **12** while a negative transmitting diode D2 is disposed between the lead **21** to a positive output line **25** while the negative diode D2 is connected to a negative output line **26**. Accordingly, the positive voltages with respect to ground are produced on output line **25** and negative voltages with respect to ground are produced on output line **26**.

In order to provide a DC output of sufficient power, a plurality of inductance-capacitance, RF, tuned circuits **30**, each forming a positive cell, or a negative cell, are utilized for connecting the positive output line **25** and negative output line **26** to first and second bus lines **31** and **32**, respectively. Bus line **32** is connected to ground while bus line **32** can be connected to circuits similar to circuit **12**. The positive output line **25** is connected by a first bridging line **33** to the first and second bus lines **31** and **32** while the negative output line **26** is connected by a second bridging line **34** to the first and second bus lines. The bridging line **33** has capacitors C1 and C2 disposed between the positive output line **25** and the first and second bus lines **31** and **32**, while the bridging line **34** also has capacitors C1 and C2 dispensed between the negative output line **26** and the first and second bus lines **31** and **32**. Connected between the bridging lines **33** and **34**, is an inductor **35** which serves as an RF choke, while disposed between the positive output line **25** and the capacitor C1 there is a negative polarity diode D4 referred to herein as a bridging diode and disposed between the negative output line **26** and capacitor C1 in line **34** there is a positive polarity diode D3 referred to herein as a bringing diode. As is seen of the drawing, the RF tuned circuit cell **30** is repeated a plurality of times. In the specific example shown, the circuit **12** has separate cells **30**, **37**, **38** and **39**. The cells **30** and **38** are of opposite polarity and balance one another while the cells **37** and **39** of opposite polarity and also balance one another. In order for the system to function, a pair of opposite polarized cells must be utilized. The particular number of cells **30** and the value of the components thereof are determined by the configuration of the dipolar antenna **10** and the power and frequency of the RF transmitter **11**.

The radio frequency to direct current conversion circuit **12** may itself be connected to a duplicate circuit via pins **41** so as to provide additional direct current output on lines similar to positive output line **25** and negative output line **26** the output lines may be connected together in order to boost the total output of the system.

An operative embodiment of the invention utilizes the following elements:

Diodes D1, D2, D3 and D4—Germanium Diodes, Archer **1** N34A,Catalog #1123.
Inductor 35-47 Milli henry R. F. Choke
Capacitors C1 and C2-0.47 Pico Farads at 200 volts
Coaxial Cable **13**—50 ohms
Dipolar Antenna **10**—aluminum wire triangular loops approximately 12 inches by 18 inches.

From the foregoing description, one skilled in the art can easily ascertain the essential characteristics of this invention, and without departing from the spirit and scope thereof, can make various changes and modifications of the invention to adapt it to various usages and conditions.

What is claimed is:

1. Apparatus for converting radio frequency energy into direct current for generating electric power, the apparatus comprising:

a dipolar input for receiving radio frequency energy;

a positive output line connected to one pole of the dipolar input and a negative output line connected to the other pole of the dipolar input;

a positive transmitting input diode means in the positive output line and a negative transmitting input diode means in the negative output line;

first and second bus lines and a pair of tuned circuits of opposite polarity coupling the positive output line and negative output line to the bus lines, one of the bus lines being connected to ground;

each tuned circuit including a first bridging line connecting the positive output line to the first and second bus lines and a second bridging line connecting the negative output line to the first and second bus lines, each bridging line having a bridging diode means therein oriented at a polarity which is reverse with respect to the transmitting input diode means of the respective output line; the bridging lines of each tuned circuit being connected to one another by an inductance and having capacitors disposed between the bridging diode means thereon and the bus lines, and

a direct current device connected to the bridging lines of the tuned circuit.

2. The apparatus of claim **1**, wherein there are a plurality of similarly configured tuned circuits connected between the output lines and the bus lines.

3. The apparatus of claim **2**, wherein there are a plurality of radio frequency-to-direct current conversion circuits connected to one another to provide a direct current power array tuned to a specific radio frequency such as a high frequency source, low frequency source, very low frequency source, extremely low frequency source, or source created by seismic vibrations of the earth's magnetic field.

4. The apparatus of claim **3** wherein the dipolar input is a dipolar antenna.

5. The apparatus of claim **1**, wherein the device connected to the output lines is a direct current motor.

6. The apparatus of claim **1**, wherein the device connected to the output line is an illuminating device.

7. The apparatus of claim **1**, wherein the device connected to the output lines is a battery charger.

8. The apparatus of claim **1**, wherein the device connected to the output lines is a DC-to-AC inverter.

9. The apparatus of claim **1**, wherein there are a plurality of radio frequency to direct current converter circuits connected to one another to form an array of circuits tuned to a particular radio frequency.

10. The apparatus of claim **1** wherein the dipolar input is a dipolar antenna.

11. The apparatus of claim **10**, wherein the dipolar antenna utilizes aluminum wire arranged in pair of triangular loops.

* * * * *

US007855476B2

(12) **United States Patent**
Ogram

(10) **Patent No.: US 7,855,476 B2**
(45) **Date of Patent: Dec. 21, 2010**

(54) **ATMOSPHERIC ELECTRICAL GENERATOR**

(76) Inventor: **Mark Ellery Ogram**, 780 S. Freeman Rd., Tucson, AZ (US) 85748

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 254 days.

(21) Appl. No.: **12/218,297**

(22) Filed: **Jul. 14, 2008**

(65) **Prior Publication Data**

US 2010/0007218 A1 Jan. 14, 2010

(51) **Int. Cl.**
H02G 11/00 (2006.01)

(52) **U.S. Cl.** **307/145**; 307/149; 361/218; 361/230

(58) **Field of Classification Search** 307/149, 307/145; 244/30, 31; 174/2; 361/212, 215–218, 361/230, 231
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

674,427	A *	5/1901	Palencsar	307/149
911,260	A *	2/1909	Pennock	307/149
1,014,719	A *	1/1912	Pennock	307/149
1,540,998	A *	6/1925	Plauson	310/309
3,121,196	A *	2/1964	Kasemir	324/464
4,234,167	A *	11/1980	Lane	254/291
4,486,669	A *	12/1984	Pugh	290/44
4,842,219	A *	6/1989	Jakubowski et al.	244/31
5,203,542	A	4/1993	Coley et al.	
6,254,034	B1 *	7/2001	Carpenter	244/153 R
6,829,911	B2	12/2004	Davis et al.	
6,961,662	B2	11/2005	Murphy	
7,016,785	B2	3/2006	Makela et al.	
7,200,418	B2	4/2007	Kaikuranta et al.	
7,330,366	B2	2/2008	Lee et al.	
7,375,492	B2	5/2008	Calhoon et al.	
7,378,181	B2	5/2008	Skinlo	
7,388,350	B1	6/2008	Wright	
7,394,671	B2	7/2008	Fukumoto et al.	
7,397,220	B2	7/2008	Uchida et al.	

* cited by examiner

Primary Examiner—Fritz M Fleming
(74) *Attorney, Agent, or Firm*—Mark Ogram

(57) **ABSTRACT**

A mechanism to tap an electrical source which uses an aircraft (preferably a lighter than air balloon) tethered by a conductive line. The conductive line is extended/withdrawn by a winch motor to adjust the altitude of the aircraft. The conductive line is isolated from the ground and an electrical conductor is connected to the conductive line and to an electrical load. In this manner, static electricity generated in the atmosphere is gathered for use.

15 Claims, 4 Drawing Sheets

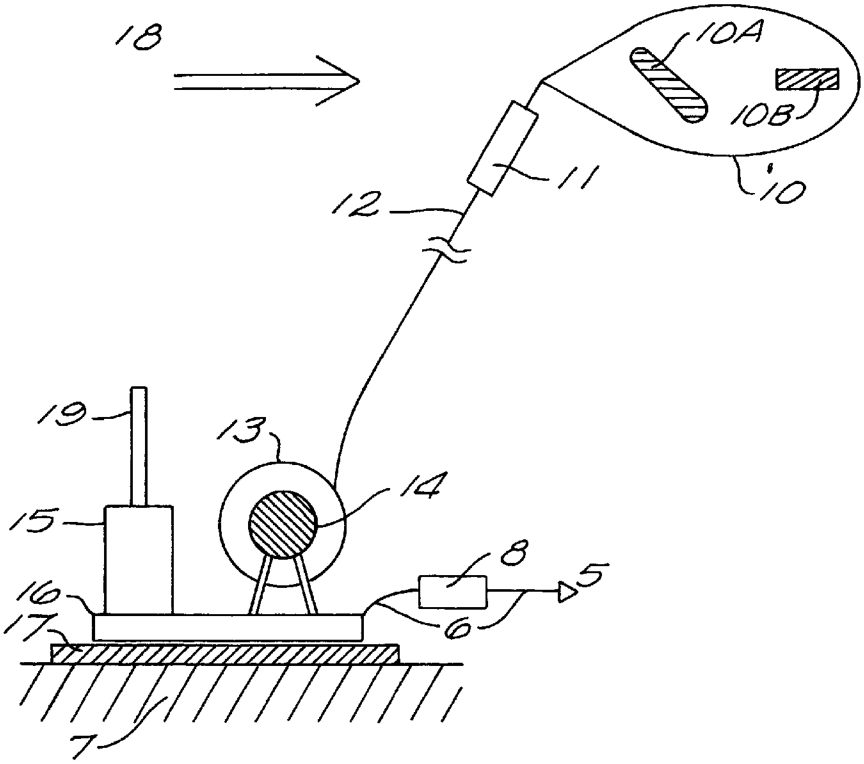

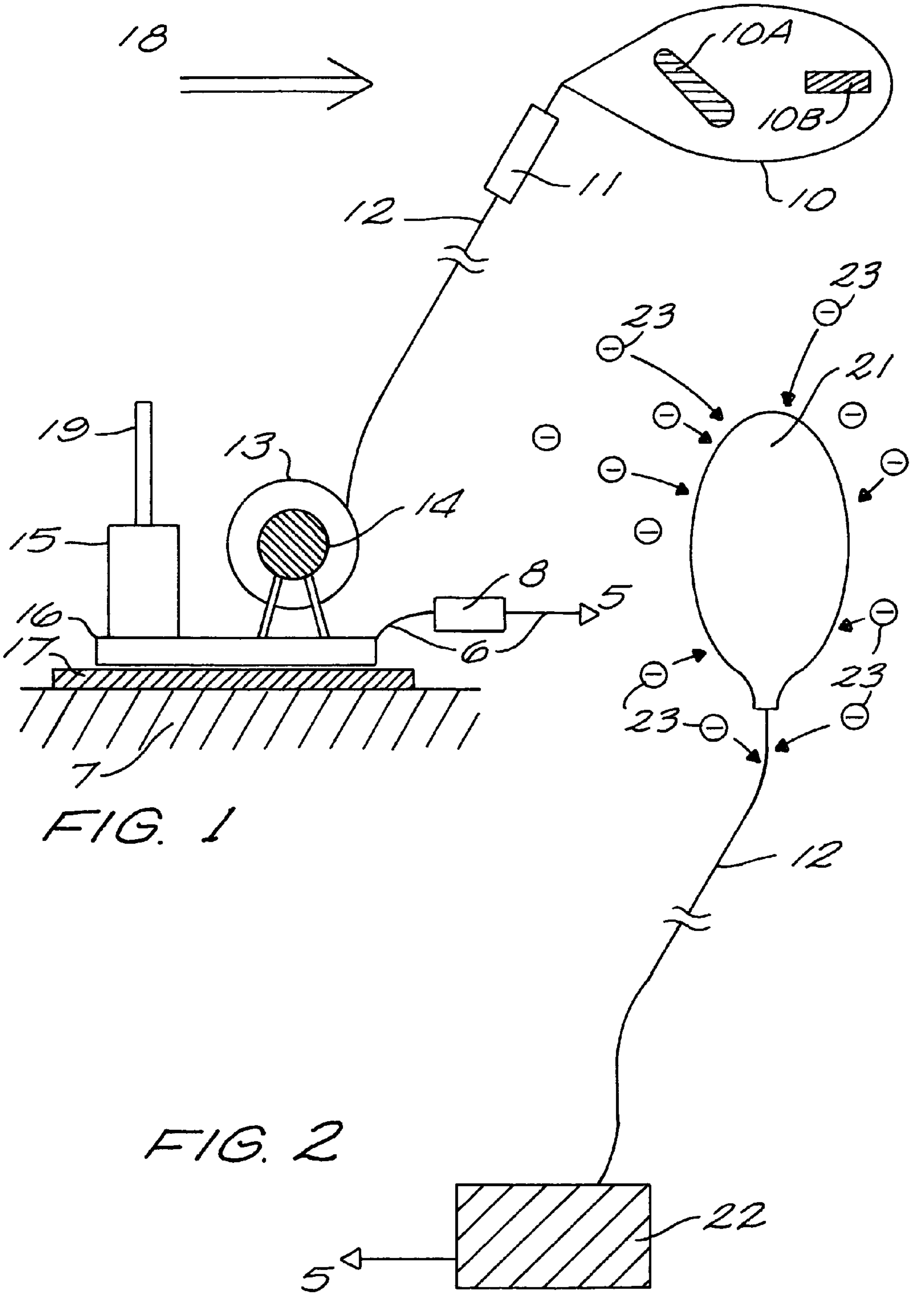
18
10A
10B
10
11
12
19
13
14
15
8
5
16
17
6
7
23
23
21
23
23
12
22
5

FIG. 1

FIG. 2

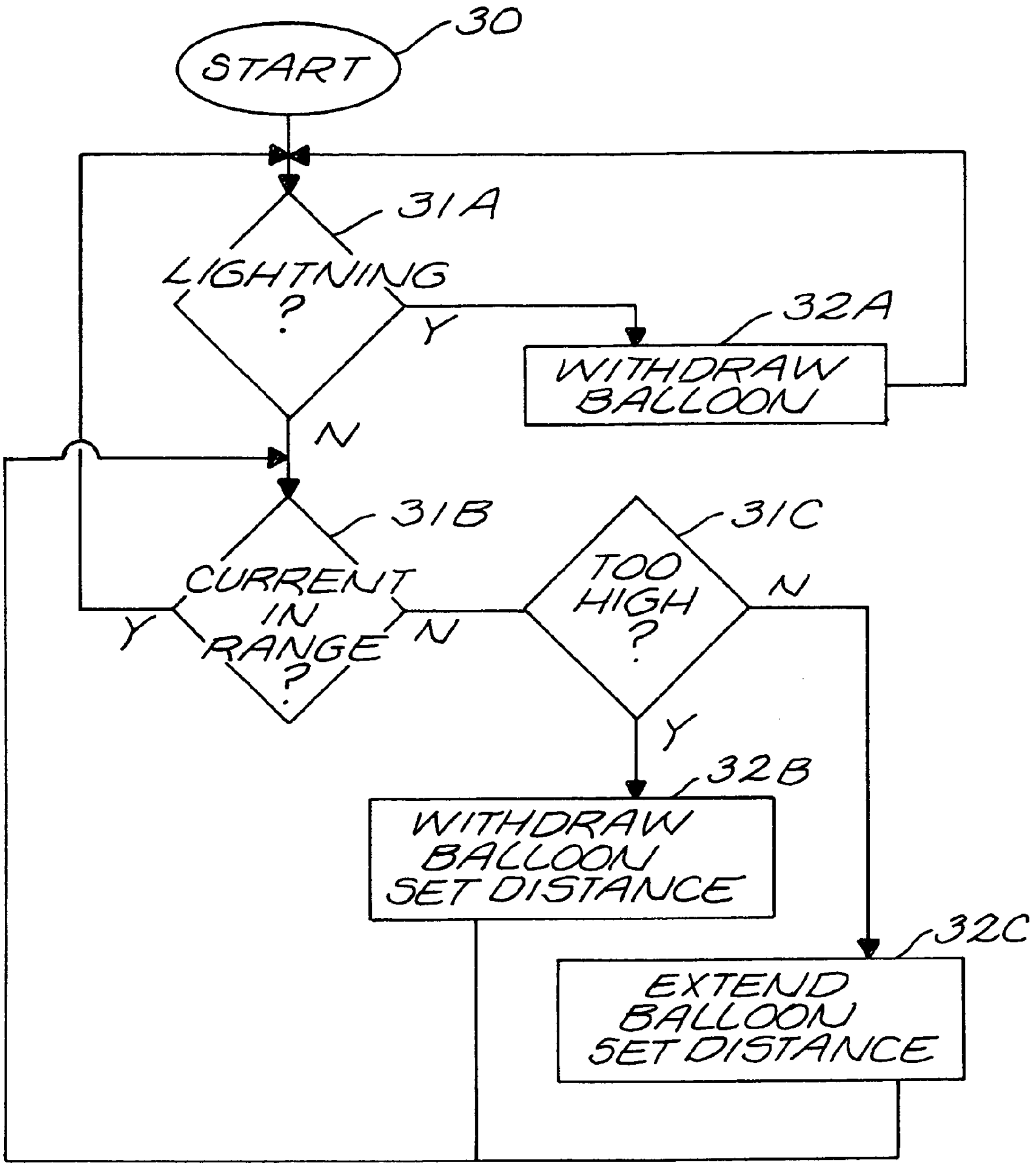
30
START
31A
LIGHTNING ?
Y
32A
WITHDRAW BALLOON
N
31B
CURRENT IN RANGE ?
Y
N
31C
TOO HIGH ?
N
Y
32B
WITHDRAW BALLOON SET DISTANCE
32C
EXTEND BALLOON SET DISTANCE

FIG. 3

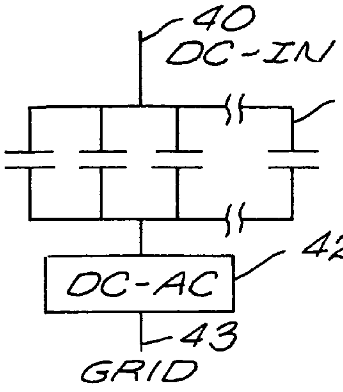
40
DC-IN
41
42
DC-AC
43
GRID

FIG. 4A

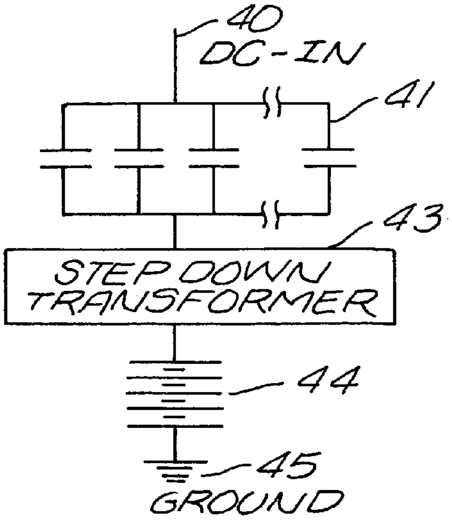
40
DC-IN
41
43
STEP DOWN TRANSFORMER
44
45
GROUND

FIG. 4B

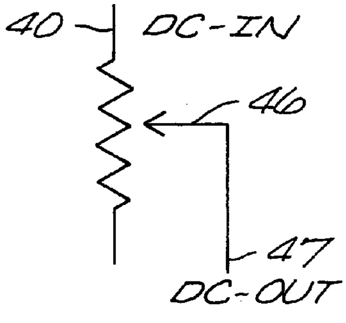
40
DC-IN
46
47
DC-OUT

FIG. 4C

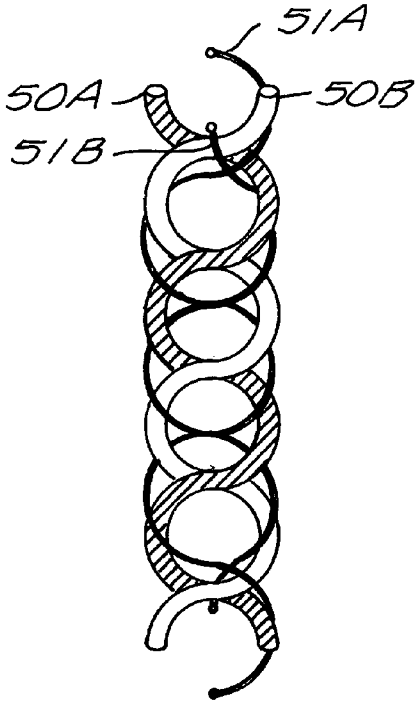
51A
50A
50B
51B

FIG. 5

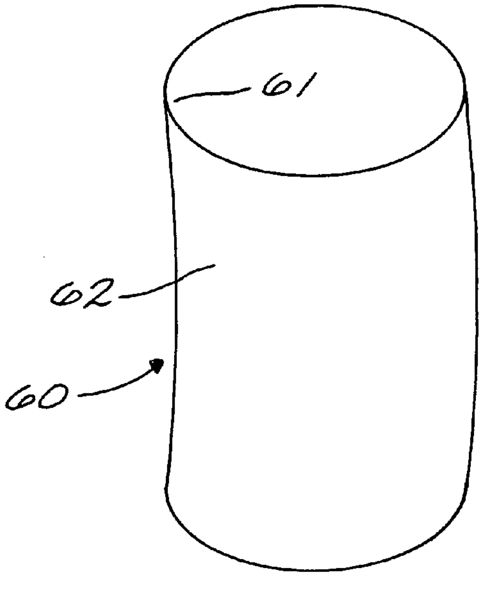

FIG. 6A

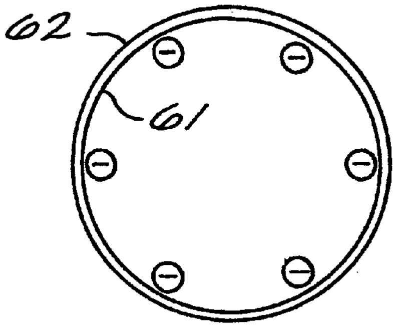

FIG. 6B

REPLACEMENT

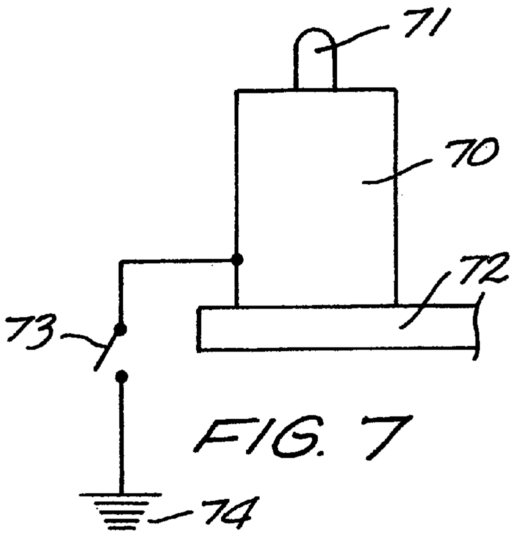
71
70
72
73
74

FIG. 7

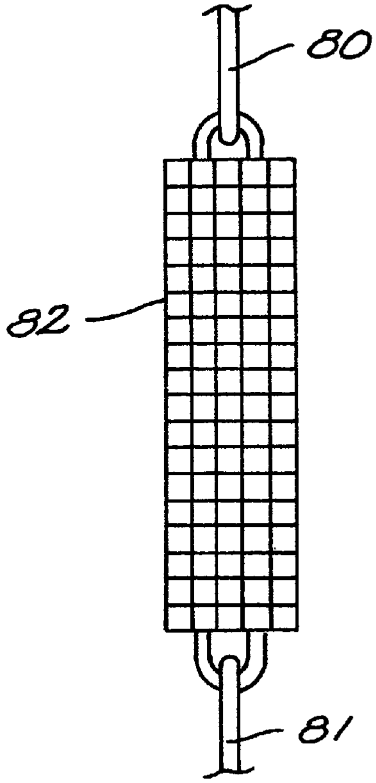

FIG. 8A

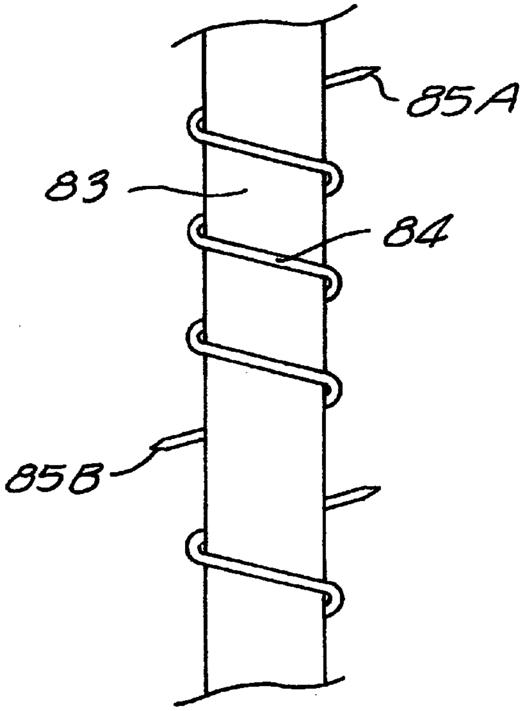
85A
83
84
85B

FIG. 8B

ATMOSPHERIC ELECTRICAL GENERATOR

BACKGROUND OF THE INVENTION

This invention relates to the production of electrical energy and more particularly to the production of electrical energy from the atmosphere.

Everyone is familiar with Benjamin Franklin's kite experiment of 1752. Using a kite whose string had become wet, negative charges from the passing clouds flowed into the string, down to the suspended key, and then into a Leyden jar via a thin metal wire. Franklin was protected by a dry silk string; but, when Franklin's knuckle came too close to the key, he received a strong shock. Fortunately, Benjamin Franklin was not killed, others who tried this same experiment were not so lucky.

Since then, the formation of lightning has remained something of a mystery. Lightning bolts are triggered when a negatively charged cloud base induces a positive charge from the ground, thereby forming a "pathway" for the discharge of the collected electrical energy.

Lightning travels up to 60,000 miles per hour with a flash that is brighter than ten million 100-watt lightbulbs. This wattage is as much power as is produced by all of the electricity plants in the United States and with a voltage of up to 300 million volts.

It is this very fact, the power within lightning is immense, that has prevented any successful collection of the electrical energy from lightning. The electricity in lightning is far too extreme for current technology to harness.

While lightning has attracted a energy starved industrial world, no one has developed any technique to harness this naturally occurring electrical source.

It is clear there is a continuing need for an electrical source other than carbon-based fuels and that the naturally occurring electricity in the atmosphere is being ignored.

SUMMARY OF THE INVENTION

The invention is a mechanism which taps into the naturally occurring static electricity in the atmosphere. Whereas heretofore, the attempt to garner electricity from the atmosphere has focused exclusively on capturing lightning, the present invention syphons off the static electricity which is generated from any agitated air and avoids lightning.

Lightning is only the final discharge of the static electricity, whether that lightning is intra-cloud lightning, cloud-to-ground lightning, or inter-cloud lightning. Other types of final discharges are known as heat lightning, summer lightning, sheet lightning, ribbon lightning, silent lightning, ball lightning, bead lightning, elves, jets, and sprites. Well before these discharges are observed, as the atmosphere becomes agitated by wind or thermal activity, static electricity is being generated.

The present invention recognizes that this static electricity is being formed and creates a mechanism to capture it.

The mechanism of this invention utilizes an aircraft such as a lighter than air balloon. While the preferred embodiment uses a foil balloon, a variety of other aircraft are obvious to those of ordinary skill in the art, including, but not limited to: gliders, rubber balloons (such as weather balloons), biaxially-oriented polyethylene terephthalate polyester film balloons, and latex balloons.

Within this discussion, the balloon is referenced, but, the invention is not intended to be limited solely to balloons.

The balloon is sent aloft and is tethered by a conductive line. In this context, the conductive line may be any obvious to those of ordinary skill in the art. For the preferred embodiment, the conductive line is a generically referred to as a "poly-rope" and is commercially available through a variety of sources. A suitable conductive line is described in U.S. Pat. No. 5,203,542, entitled "Apparatus for Improved Electric Fence Wire Construction for use with Intensive Grazing" issued Apr. 20, 1993, to Coley, et al. and incorporated hereinto by reference.

The conductive line is played out of a winch to control the altitude of the balloon. The motor controlling the winch is able to reverse direction to both extend and withdraw the conductive line which is wrapped around a spool on the winch. The winch/spool combination are part of a base unit.

In some embodiments of the invention, the spool is constructed of rubber so as to insulate the conductive line from the winch assembly. In this embodiment, only the conductive line is charged by the atmospheric static electricity while the winch remains neutral.

In yet another embodiment, the winch/spool are part of a base unit which is itself isolated from the ground by an insulator. In this embodiment, the entire base unit is charged by the atmospheric static electricity.

A conductor, such as an insulated wire, is electrically connected to the conductive line. In one embodiment, where the conductive line is electrically isolated from the spool and winch motor, the conductor is connected to the conductive line. In the embodiment where the conductive line is electrically connected to the base unit, then the conductor is connected anywhere on a metallic base unit.

The other end of the conductor is connected to a load. The load in this case can be any of a variety of electrical loads well known to those of ordinary skill in art, including, but not limited to a motor, a battery system, or the electrical grid for the system.

In the preferred embodiment, a sensor array is used to monitor the activities both at the base unit (such as electrical flow within the conductor) and in the surrounding locale.

A sensor monitoring the electrical flow (i.e. voltage and/or current) within the conductor is used to monitor the electrical activity within the conductor.

In the preferred embodiment, a lightning sensor monitors for lightning activity within the locale. As noted earlier, the electrical characteristic of lightning is so extreme that ideally this discharge is avoided as it might damage the mechanism of this invention.

The sensor array is utilized by a controller, such as microprocessor, programmed to operate the mechanism as outlined herein.

The controller operates the winch motor to extend or withdraw the conductive line and by extension the altitude of the balloon. The controller is programmed to operate the winch by monitoring the electrical characteristics of the conductor and adjusting the balloon's altitude to maintain these characteristics within the conductor within a preset range.

This preset range is established either in the base programming of the controller or is established by an operator of the system.

As example, by controlling the amount of current being withdrawn from the atmosphere, the mechanism operates within a safe range and also provides a relatively stable current flow from which a variety of activities can take place (such as DC-AC conversion).

The controller also utilizes the lightning sensor to protect the mechanism from a lightning strike. Should lightning be detected within a pre-determined range (as established by the software or defined by an operator), then the balloon is pulled down to minimize the risk of damage from a lightning strike.

The invention, together with various embodiments thereof will be more fully explained by the following description of the accompanying drawings.

DRAWINGS IN BRIEF

FIG. **1** diagrams the preferred embodiment of the invention.

FIG. **2** illustrates the collection of the negative charged particles in the atmosphere.

FIG. **3** is a flow-chart of the operation of the controller for the preferred embodiment of the invention.

FIGS. **4**A, **4**B, and **4**C are electrical schematics for handling the static charge from the atmosphere.

FIG. **5** illustrates a conductive line used in the preferred embodiment of the invention.

FIGS. **6**A and **6**B illustrate an alternative conductive line creating an ionized pathway for the flow of the static charges from the atmosphere.

FIG. **7** illustrates the controller of an alternative embodiment and the associated safety devices.

FIGS. **8**A and **8**B illustrate two embodiments of enhanced electrical collection leads.

DRAWINGS IN DETAIL

FIG. **1** diagrams the preferred embodiment of the invention.

Balloon **10** is an aircraft which, in this illustration, is a lighter than air balloon. Wings **10**A, extending from the body of balloon **10**, provide additional lift in air flow **18**. Tail **10**B helps to stabilize balloon **10**.

Balloon **10** is tethered to the ground via conductive line **12**. As noted earlier, a variety of configurations and materials are available to serve as conductive line **12**. In this illustration, a poly-wire is used Poly-wire is commercially available through a variety of vendors, including, but not limited to: Jeffers Livestock and Zareba Systems, Inc, of Ellendale, Minn.

In this embodiment, located proximate to balloon **10**, is an electrical collection enhancement lead **11** which assists in the collection of the static electrical charge in the atmosphere. Electrical collection enhancement lead **11** is configured to attract the static charge and conduct the charge into the conductive line **12**.

The electricity flows down the conductive line into spool **13**, where the conductive line **12** is collected and either withdrawn or dispensed through operation of winch motor **14**.

Winch motor **14** and spool **13** are mounted onto base unit **16** which is electrically isolated from ground **7** using insulator **17**. Note, in this embodiment of the invention, when electricity is being collected from the atmosphere, the entire base unit **16** becomes charged. In another embodiment of the invention, spool **13** is constructed of rubber, thereby preventing base unit **16** from becoming charged, thereby restricting the charging from the atmosphere to only conductive line **12**.

In this embodiment, conductor **6** is connected to base unit **16** (since the entire base unit **16** is charged and the base unit is metallic) to communicate the electrical current to load **5**. Conductor **6** is ideally an insulated wire.

The electrical current through conductor **6** is measured using sensor **8**.

In the alternative embodiment discussed above, where only the conductive line **12** is charged, then conductor **6** is connected to conductive line **12**.

Controller **15**, located in this embodiment on base unit **16**, operates winch motor **14** in response to signals from sensor **8** (measuring the current being discharged to load **5**) to maintain the current flow within a pre-defined range. As the current flow diminishes, then the conductive line **12** extended from spool **13** to increase the altitude of balloon **10** to that more static charge from the atmosphere is gathered; as the current flow falls exceeds a preset level, conductive line **12** is withdrawn onto spool **13** to decrease the static charge being collected from the atmosphere.

The range of current flow through conductor **6** is ideally set by the program, although some embodiments of the invention permit an operator to establish this range of operation.

In an alternative embodiment, the sensor monitoring conductor **6** monitors the voltage therein.

In the preferred embodiment of the invention, controller **15** is also equipped with a lightning sensor **19**. In this embodiment, when lightning is sensed within a preset range, then substantially all of conductive line **12** is wound onto spool **13** to pull balloon **10** near the ground and protect the entire mechanism from being damaged from a lightning discharge.

In the preferred embodiment, the "safe" distance form lightning is set in the programming of controller **15** and is ideally two miles; other embodiments permit the operator to "safe" distance.

There are a variety of lightning sensors well known to those of ordinary skill in the art, including, but not limited to those described in: U.S. Pat. No. 7,016,785, entitled "Lightning Detection" issued to Makela, et al. on Mar. 21, 2006; U.S. Pat. No. 6,829,911, entitled "Lightning Detection and Prediction Alarm Device" issued to Jones, et al. on Dec. 7, 2004; U.S. Pat. No. 7,200,418, entitled "Detection of Lightning" issued to Karikuranta, et al. on Apr. 3, 2007; and U.S. Pat. No. 6,961,662, entitled "Systems and Methods for Spectral Corrected Lightning Detection" issued to Murphy on Nov. 1, 2005; all of which are incorporated hereinto by reference.

In another embodiment of the invention, controller **15** is not located on base unit **16**, rather it is remote and communicates its control signals to winch motor **14** using radio waves.

FIG. **2** illustrates the collection of the negative charged particles in the atmosphere.

Static charges **23** are generated in the atmosphere by agitated air. These static charges are often collected at the bottom of clouds, but exist in other environments as well.

Balloon **21** is extended into this strata of static charges **23** which are then attracted to conductive line **12** to flow to base unit **22** and then onto load **5**.

By increasing or decreasing the altitude of balloon **21** (defined by the length of the extended conductive line **12**), conductive line **12** is selectively exposed to varying densities and levels of the static charge strata, and by extension, the current flow or voltage is increased or decreased.

FIG. **3** is a flow-chart of the operation of the controller for the preferred embodiment of the invention.

Once the program starts **30**, the lightning sensor is checked to determine if lightning has occurred within the unsafe range **31**A, if it has, then the balloon is lowered **32**A, and the program continues monitoring the status of lightning until no lightning is detected.

When the lightning status is acceptable, then the current within the conductor is checked to see if the current is within the prescribed range **31**B. If the current is acceptable (within range) the program returns to check the lightning status **31**A; otherwise a determination is made to see if the current is above the prescribed range **31**C.

If the current is above the prescribed range, then the altitude of the balloon is withdrawn a set amount **32**B (ideally twenty-five feet) and the program loops back to see if the current is within range **31**B.

If the current is below the prescribed range, then the altitude of the balloon is extended a set amount **32**C (ideally twenty-five feet) and the program loops back to see if the current is within range **31**B.

In this manner of feed-back and minor adjustments in the altitude of the balloon, the current is maintained within a prescribed range which can be handled by the downstream electrical system.

As noted earlier, some embodiments of the invention monitor the voltage instead of the current.

FIGS. **4**A, **4**B, and **4**C are electrical schematics for handling the static charge from the atmosphere.

By maintaining the voltage being collected in a prescribed range, an electrical conversion system is easily designed. While FIGS. **4**A, **4**B, and **4**C illustrate some electrical configurations, those of ordinary skill in the art readily recognize a variety of other configurations which will serve the same function.

Referencing FIG. **4**A, Direct Current In (DC IN) **40** is buffered by a gang of capacitors **41** before being communicated to a DC/AC converter **42**. The DC/AC converter converts the direct current into a an alternating current suitable for placement over an existing electrical grid **43** such as normally found from a power-plant.

Those of ordinary skill in the art readily recognize a variety of DC/AC converters, including, but not limited to: U.S. Pat. No. 7,394,671, entitled "Controller IC, DC-AC Conversion Apparatus, and parallel running system of DC-AC Conversion Apparatuses" issued to Fukumoto, et al. on Jul. 1, 2008; and, U.S. Pat. No. 7,330,366, entitled "DC-AC Converter" issued to Lee, et al. on Feb. 12, 2008; all of which are incorporated hereinto by reference.

FIG. **4**B illustrates an electrical arrangement suitable for use in charging a battery. DC IN **40** is buffered by capacitor bank **41** before entering into a step down transformer **43**. Step down transformer **43** reduces the voltage so that the voltage can safely be introduced into battery **44** which is connected to ground **45** at the battery's other pole.

Those of ordinary skill in the art readily recognize a variety of batteries which will work in this capacity, including, but not limited to those described in: U.S. Pat. No. 7,378,181, entitled "Electric Storage Battery Construction and Manufacture" issued to Skinlo on May 27, 2008; U.S. Pat. No. 7,388,350, entitled "Battery with Electronic Compartment" issued to Wright on Jun. 17, 2008; U.S. Pat. No. 7,397,220, entitled "Connection Member and Battery Pack" issued to Uchida, et al. on Jul. 8, 2008; and, U.S. Pat. No. 7,375,492, entitled "Inductively Charged Battery Pack" issued to Calhoon, et al. on May 20, 2008; all of which are incorporated hereinto by reference.

In FIG. **4**C, DC IN **40** is fed into an adjustable rheostat **46** which is controlled by the controller so that the DC OUT **47** falls within a specified range.

FIG. **5** illustrates a conductive line used in the preferred embodiment of the invention.

This type of conductive line is commonly called poly-wire and consists of multiple interwoven strands of plastic **50**A and **50**B woven into a cord or rope arrangement having intertwined therein exposed metal wires **51**A and **51**B. While this illustration shows two plastic strands and two metal wires, any number of possible combinations is possible.

The exposed metal wires **51**A and **51**B attract the atmospheric static charge and transmit the charge down to the base unit (not shown).

FIGS. **6**A and **6**B illustrate an alternative conductive line creating an ionized pathway for the flow of the static charges from the atmosphere.

This conductive line utilizes a tube **60** having an outer layer **62** of PET Film (Biaxially-oriented polyethylene terephtalate polyester film) which provides exceptionally high tensile strength and is chemically and dimensionally stable. The tube has an ideal diameter of between two and three inches.

An interior metal coating **61** provides an initial conduit for the flow of static charge. The static charge through the metal forces the tube to expand due to the repulsion experienced by like charges. Further, the flow of electricity causes the interior of the tube **60** to become ionized to provide an additional pathway for the atmospheric static charges to the base unit (not shown).

Because outer layer **62** provides a gas barrier, the resulting ionization is not dissipated by air currents, thereby providing a highly stable pathway.

FIG. **7** illustrates the controller of an alternative embodiment and the associated safety devices.

In this embodiment, controller box **70**, resting on insulating pad **72**, is in communication with the sensors as described above. Using the input from these sensors, when there is flow of electricity through the base unit, warning flashing light **71** is illuminated. To electrically neutralize the mechanism, switch **73** is activated to pass any existing current into the ground **74**.

FIGS. **8**A and **8**B illustrate two embodiments of enhanced electrical collection leads.

Referencing FIG. **8**A, enhanced electrical collection lead **82** is a wire mesh which is in electrical communication with conductive line **81** and balloon **80**. Because of the significant amount of metal exposed by enhanced electrical collection lead **82**, more static electricity from the atmosphere is drawn to the collection lead **82**, and then down conductive line **81** to the base unit (not shown).

Conductive lead **82** is positioned proximate to balloon **80**.

In FIG. **8**B, poly-wire **83** has enhanced electrical collection leads **84** wrapped therearound. Collection leads **84** have pointed ends **85**A and **85**B which have a propensity to attract more electricity than rounded ends do.

It is clear from the foregoing that the present invention captures an entirely new source of electrical energy.

What is claimed is:

1. A mechanism to tap an electrical source comprising:

a) a lighter than air balloon suspended in the atmosphere;

b) a base unit having a spool of conductive line on a winch motor, one end of said conductive line secured to said lighter than air balloon, a portion of said conductive line collecting electricity in the atmosphere, said winch motor capable of selectively extending or withdrawing said conductive line from said spool;

c) an insulator electrically isolating said conductive line from ground;

d) a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load being powered by collected electricity from said conductive line;

e) an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor; and,

f) a controller receiving said electrical flow indicia and selectively operating said winch motor such that said electrical flow indicia remains within a selected operating range.

2. The mechanism to tap an electrical source according to claim **1**, wherein said selected operating range is established by an operator.

3. The mechanism to tap an electrical source according to claim **2**, further including a warning light activated when said electrical flow indicia is non-zero.

4. The mechanism to tap an electrical source according to claim **1**,

a) further including a lightning sensor generating a lightning presence indicia indicative of lightning within a prescribed range, said presence indicia being communicated to said controller; and,

b) wherein said control mechanism, in response to said lightning presence indicia, operates said winch motor to withdraw substantially all of said conductive line onto said spool.

5. The mechanism to tap an electrical source according to claim **4**, wherein said prescribed range is established by an operator.

6. The mechanism to tap an electrical source according to claim **1**, further including an electrical collection enhancement lead in electrical contact with a first end of said conductive material, said electrical collection enhancement lead configured to attract static electricity.

7. The mechanism to tap an electrical source according to claim **6**, wherein said electrical collection enhancement lead is positioned proximate to said lighter than air balloon.

8. The mechanism to tap an electrical source according to claim **7**, wherein said electrical collection enhancement lead includes at least two pointed electrical conductors.

9. The mechanism to tap an electrical source according to claim **1**, wherein said insulator electrically isolates said conductive line from said winch motor.

10. The mechanism to tap an electrical source according to claim **1**, wherein said insulator electrically isolates said base unit from ground.

11. A mechanism comprising:

a) an airborne aircraft having a conductive line secured to a winch capable of extending or withdrawing said conductive line from a spool, when said airborne aircraft is aloft, said conductive line collecting electricity from the atmosphere;

b) a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load such that collected electricity from said conductive line powers said load;

c) an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor; and,

d) a controller receiving said electrical flow indicia adjusting an altitude of said airborne aircraft via said winch such that said electrical flow indicia remains within a selected operating range.

12. The mechanism to tap an electrical source according to claim **11**,

a) further including a lightning sensor communicating with said controller; and,

b) wherein said controller, in response to selected signals from said lightning sensor withdraws substantially all of said conductive line.

13. A mechanism to tap an electrical source comprising:

a) a lighter than air balloon suspended in the atmosphere;

b) a base unit having a spool of conductive line on a winch motor, one end of said conductive line suspended by said balloon, a portion of said conductive line collecting electricity from the atmosphere, said winch motor capable of adjusting an altitude of said lighter than air balloon by selectively extending or withdrawing said conductive line from said spool;

c) a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load being powered by said collected electricity from said conductive line;

d) a sensor array having,

  1) an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor, and,

  2) a lightning sensor monitoring existence of proximate lightning; and,

e) a controller responsive to said electrical flow indicia from said sensor array to selectively operate said winch motor.

14. The mechanism to tap an electrical source according to claim **13**,

a) further including a warning light; and,

b) wherein said controller activates said warning light when said electrical flow in said conductor is non-zero.

15. The mechanism to tap an electrical source according to claim **14**, further including an electrical collection enhancement lead in electrical contact with said conductive material, said electrical collection enhancement lead configured to attract static electricity and positioned proximate to said lighter than air balloon.

* * * * *

US009160156B2

(12) United States Patent
Allen et al.

(10) **Patent No.: US 9,160,156 B2**
(45) **Date of Patent: Oct. 13, 2015**

(54) **SYSTEM FOR HARVESTING ATMOSPHERIC ELECTRICITY**

(75) Inventors: **Edward H. Allen**, Lancaster, CA (US); **Sergey Macheret**, Palmdale, CA (US)

(73) Assignee: **Lockheed Martin Corporation**, Bethesda, MD (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 357 days.

(21) Appl. No.: **13/276,259**

(22) Filed: **Oct. 18, 2011**

(65) **Prior Publication Data**

US 2013/0093261 A1 Apr. 18, 2013

(51) **Int. Cl.**
H05F 7/00 (2006.01)
H02G 13/00 (2006.01)
B64D 45/02 (2006.01)

(52) **U.S. Cl.**
CPC ***H02G 13/20*** (2013.01); ***B64D 45/02*** (2013.01); *H05F 7/00* (2013.01)

(58) **Field of Classification Search**
CPC H02G 13/20; B64D 45/02; H05F 7/00
USPC 361/231, 230
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

3,719,829 A * 3/1973 Vaill 307/149
4,017,767 A * 4/1977 Ball 361/117
4,950,978 A * 8/1990 Govaert 324/72
5,726,855 A * 3/1998 Mourou et al. 361/213
8,045,314 B2 * 10/2011 Ibok 361/231
2014/0231679 A1 * 8/2014 Kremeyer et al. 250/505.1

OTHER PUBLICATIONS

Tzzortzakis et al. "Concatenation of plasma filaments created in air by femtosecond laser infrared laser pulses", 2003, Applied Physics B, vol. 767, pp. 609-612.*

Zhou et al. , Revival of femtosecond laser plasma filaments in air by a nanosecond laser, Jun. 24, 2009, Optic Express, vol. 17, No. 14, pp. 1 -7.*

Assion et al., "Control of Chemical Reactions by Feedback-Optimized Phase-Shaped Femtosecond Laser Pulses", Science, Oct. 30, 1998, pp. 919-922, vol. 282.

* cited by examiner

Primary Examiner — Zeev V Kitov
(74) *Attorney, Agent, or Firm* — Baker Botts L.L.P.

(57) **ABSTRACT**

Systems, methods and apparatus for harvesting atmospheric electricity are provided. The system includes a laser configured to form a plasma filament and a collector configured to collect electricity flowing along the plasma filament. The plasma filament comprises an electrically conducting plasma filament. Atmospheric electricity may be collected by having the plasma filament form at least a part of a conducting path: (1) between ground and a cloud, (2) between differently charged regions of the same cloud, (3) between differently charged regions of different clouds, and (4) between different regions of atmosphere, where there is a vertical voltage gradient. When the plasma filament is not long enough to form the entire conducting path, a lightning may be triggered to complete the conducting path needed to collect atmospheric electricity.

21 Claims, 6 Drawing Sheets

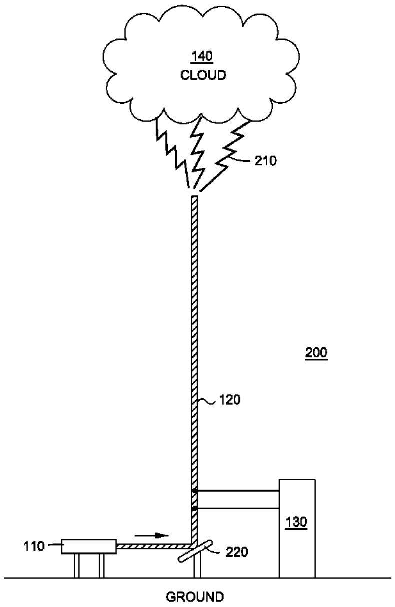

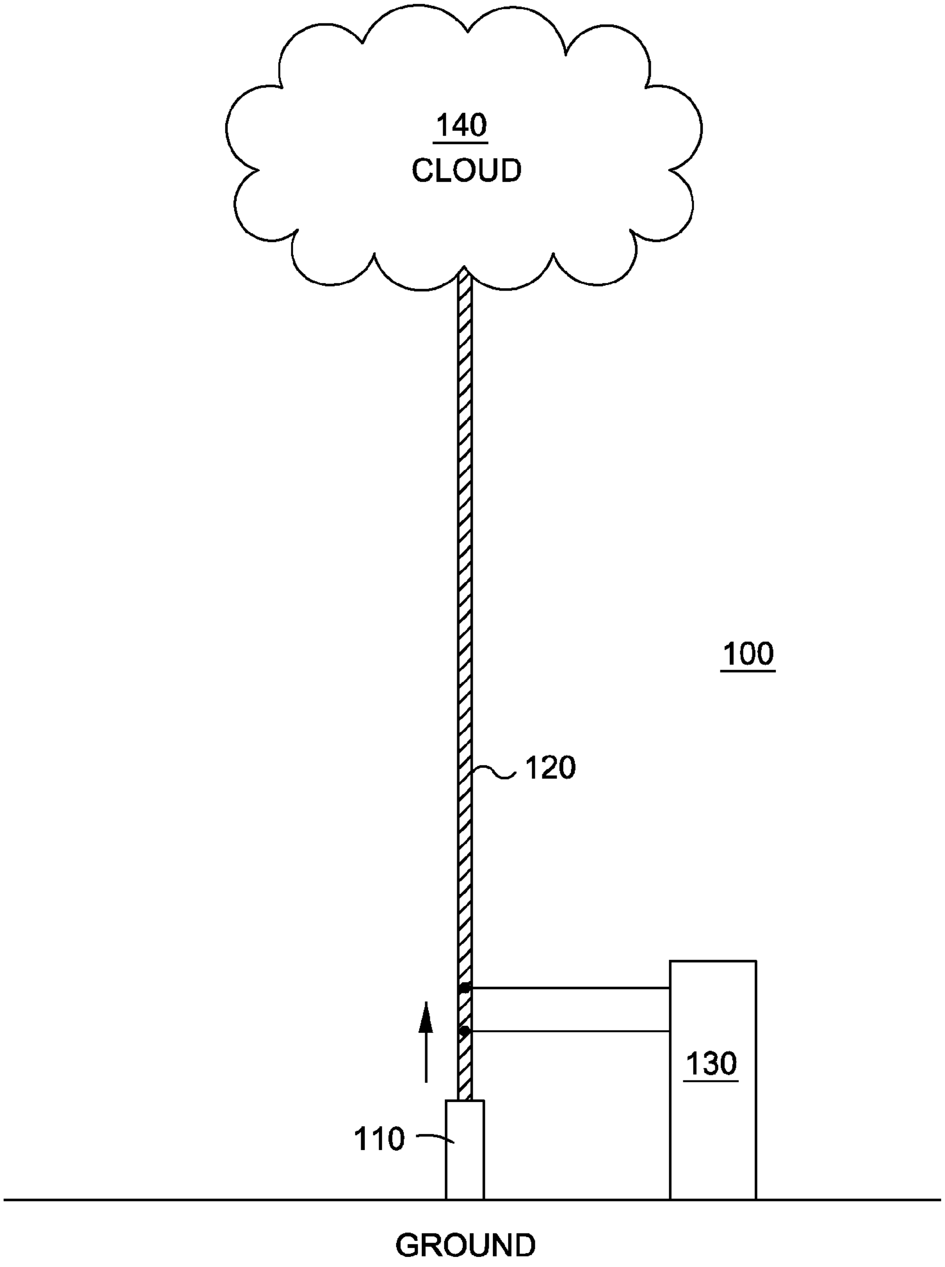
140
CLOUD
100
120
130
110
GROUND

FIG. 1

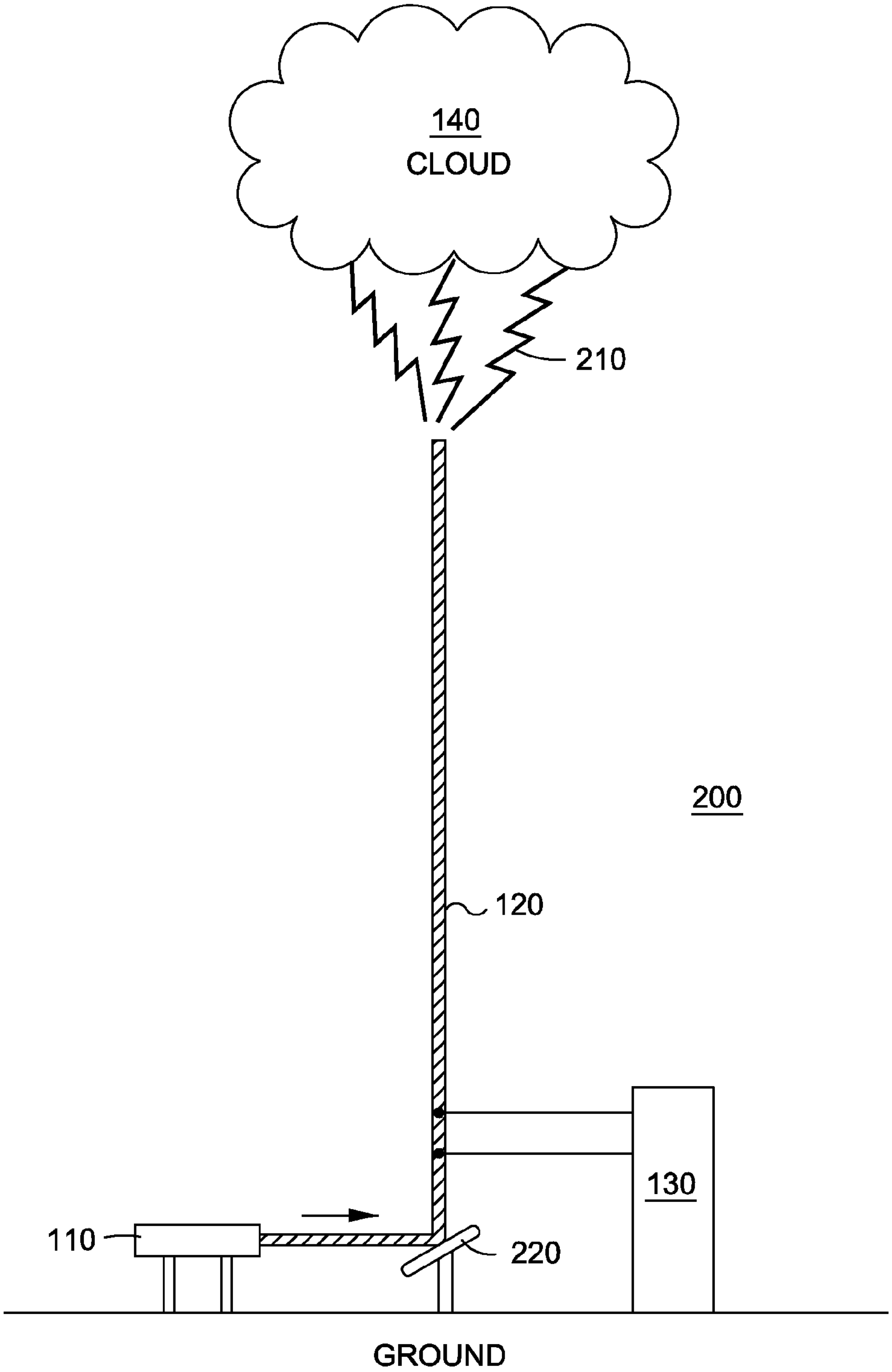
140
CLOUD
210
200
120
130
110
220
GROUND

FIG. 2

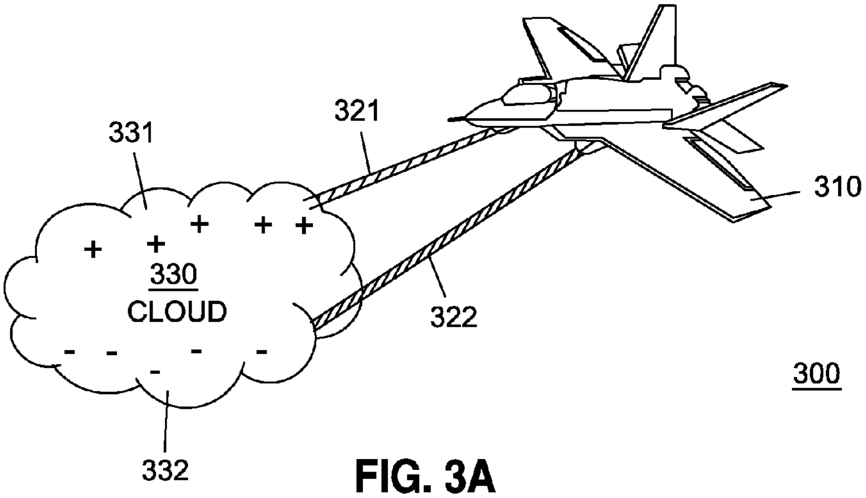
+ + + + + +
CLOUD
- - - - - -

FIG. 3A

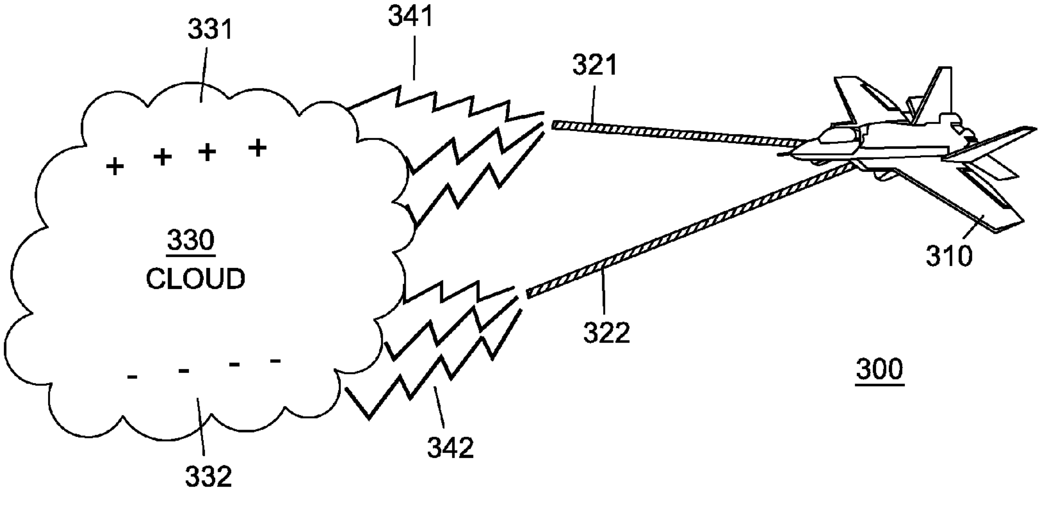
+ + + +
CLOUD
- - - -

FIG. 3B

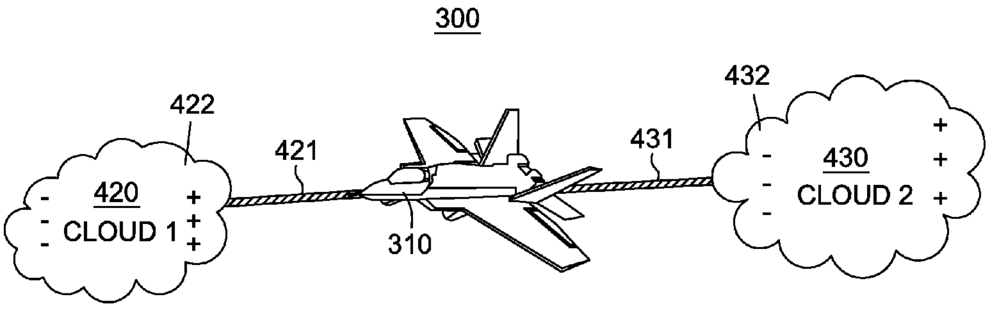
300
422
421
420
CLOUD 1
- - -
+ + +
310
431
432
430
CLOUD 2
- - -
+ + +

FIG. 4

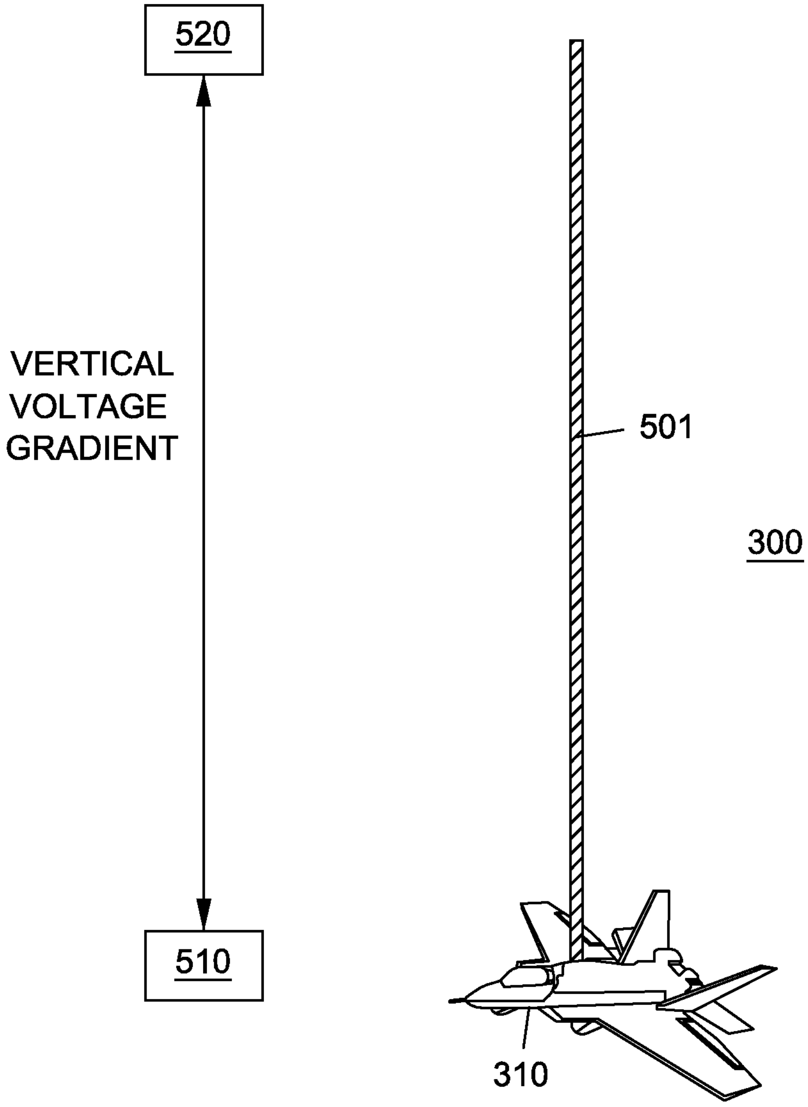
520
VERTICAL
VOLTAGE
GRADIENT
510
501
300
310

FIG. 5

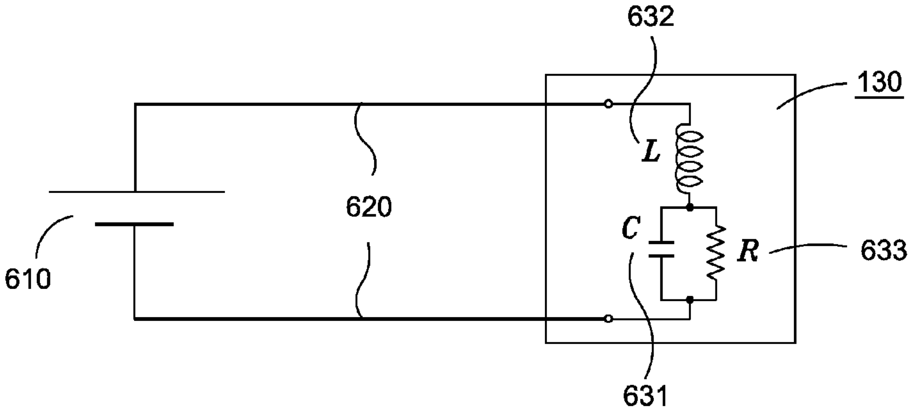
632
130
L
620
C
R
633
610
631

FIG. 6

SYSTEM FOR HARVESTING ATMOSPHERIC ELECTRICITY

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

Not applicable.

FIELD OF THE INVENTION

The subject technology relates in general to a system for harvesting atmospheric electricity, and more particularly to a system that uses a laser to form an electrically conducting plasma filament for harvesting atmospheric electricity.

BACKGROUND

It is well known that large quantities of electrical energy are present in the atmosphere and in lightning. Lightning was one of the first forms of electricity harnessed in the modern age by Benjamin Franklin in his famous kite experiment.

A lightning discharge typically contains on the order of 10^{10} Joules of energy. It has been estimated that the total electrical power of lightning across the Earth is of the order of 10^{12} watts. Collection of electrical energy from clouds associated with lightning is a problem that needs to be resolved.

Lightning is but a small part of the total electrical activity of the atmosphere. When a local build up of charge above the Earth exceeds the local breakdown potential of the atmosphere a lightning discharge occurs. However, there is a continual invisible flow of charge from Ionosphere to Earth occurring day and night over the entire surface of the globe, which exceeds the global lightning power output by many times. The source of this flow of charge from the atmosphere to Earth is the Sun. Radiation from the Sun helps form most of the ions found in the Ionosphere, a highly charged region above the atmosphere, through ionization of atmospheric molecules. Solar radiation, particularly in the UV and soft x-ray bands, consist of photons whose energies are well suited to ionization. The Sun also emits continuously a solar wind of positively charged particles. These are captured by the Earth's magnetic field and further contribute to the Ionosphere. The positively charged region in the Ionosphere in turn induces (by electrostatic induction) a negative charge on the surface of the Earth. The Earth becomes in effect an enormous spherical capacitor. A potential gradient or electric field is thus established between the two "plates" of this capacitor, the Ionosphere (or Electrosphere) and the surface of the Earth. While the upper strata of the atmosphere conduct electricity reasonably well, the lower levels act as an insulator or dielectric. Near the surface of the Earth, this electrostatic potential gradient is on the order of about 100 Vm^{-1} in summer, rising to 300 Vm^{-1} in winter. This flow of charge can be tapped and directed to provide useable electrical power. This source of atmospheric electricity has the following advantages: (1) Simple and robust technology; (2) Low cost technology—much cheaper than photovoltaics or wind turbines; (3) Available day and night in all weather conditions—in fact, more power is produced at night than during the day; and (4) Available at any point on the Earth's surface. Collection of this source of atmospheric electricity is another problem that needs to be resolved.

SUMMARY

According to various aspects of the subject technology, a system for harvesting atmospheric electricity is provided that solves some or all of the foregoing problems. In some aspects, the system may be used to harvest atmospheric electricity arising from an electrostatic potential gradient: (1) between ground and a cloud, (2) between differently charged regions of the same cloud, (3) between differently charged regions of different clouds, and (4) between different regions of atmosphere, where there is a vertical voltage gradient. In other aspects, the system may be used to harvest atmospheric electricity arising from an electrostatic potential gradient between the Ionosphere (or Electrosphere) and the surface of the Earth.

According to various aspects of the subject technology, a system for harvesting atmospheric electricity is provided. The system comprises a laser configured to form a plasma filament. The system also comprises a collector configured to collect electricity flowing along the plasma filament.

According to various aspects of the subject technology, a method for harvesting atmospheric electricity is provided. The method comprises forming a plasma filament with a laser. The method also comprises collecting electricity flowing along the plasma filament.

According to various aspects of the subject technology, an apparatus for harvesting atmospheric electricity is provided. The apparatus comprises means for forming a plasma filament with a laser. The system also comprises means for collecting electricity flowing along the plasma filament.

It is understood that other configurations of the subject technology will become readily apparent to those skilled in the art from the following detailed description, wherein various configurations of the subject technology are shown and described by way of illustration. As will be realized, the subject technology is capable of other and different configurations and its several details are capable of modification in various other respects, all without departing from the scope of the subject technology. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are included to provide further understanding of the subject technology and are incorporated in and constitute a part of this specification, illustrate aspects of the subject technology and together with the description serve to explain the principles of the subject technology.

FIG. **1** illustrates a first example of a system for harvesting atmospheric electricity. The system comprises a laser configured to form a plasma filament and a collector configured to collect electricity flowing along the plasma filament. The plasma filament forms a conducting path between ground and a cloud.

FIG. **2** illustrates a second example of a system for harvesting atmospheric electricity. A plasma filament and a lightning together form a conducting path between ground and a cloud. A condensing mirror directs the beam of laser towards the cloud.

FIG. **3**A illustrates a third example of a system for harvesting atmospheric electricity. A laser system situated on an airplane forms two plasma filaments that together complete a conducting path between a first region of a cloud and a second region of the cloud. The first region of the cloud and the second region of the cloud carry opposite electric charges. A collector configured to collect electricity flowing along the two plasma filaments is also situated on the same airplane as the laser system.

FIG. **3**B illustrates a fourth example of a system for harvesting atmospheric electricity. It is similar to FIG. **3**A, except two lightnings complete the conducting path between a first region of a cloud and a second region of the cloud (i.e., a first lightning electrically connects the first region of the cloud to one of the plasma filaments and a second lightning electrically connects the second region of the cloud to the other plasma filament).

FIG. **4** illustrates a fifth example of a system for harvesting atmospheric electricity. A laser system situated on an airplane forms two plasma filaments that together complete a conducting path between a first cloud and a second cloud. The region of the first cloud that is connected to the first plasma filament and the region of the second cloud that is connected to the second plasma filament carry opposite electric charges.

FIG. **5** illustrates a sixth example of a system for harvesting atmospheric electricity. A laser situated on an airplane forms a plasma filament that completes a conducting path between a first region of atmosphere and a second region of atmosphere. A collector situated on the same airplane as the laser can collect electricity flowing along the plasma filament, because a vertical voltage gradient exists between the first region of atmosphere and the second region of atmosphere.

FIG. **6** illustrates an example of a collector configured to collect electricity flowing along the plasma filament. The collector may be a charge collection circuit characterized by capacitance, inductance, and resistance. The electricity may be collected by charging a capacitor.

DETAILED DESCRIPTION

The detailed description set forth below is intended as a description of various configurations of the subject technology and is not intended to represent the only configurations in which the subject technology may be practiced. The appended drawings are incorporated herein and constitute a part of the detailed description. The detailed description includes specific details for the purpose of providing a thorough understanding of the subject technology. However, it will be apparent to those skilled in the art that the subject technology may be practiced without these specific details. In some instances, well-known structures and components are shown in block diagram form in order to avoid obscuring the concepts of the subject technology. Like components are labeled with identical element numbers for ease of understanding.

FIG. **1** illustrates a first example of a system **100** for harvesting atmospheric electricity. The system **100** comprises a laser **110** configured to form a plasma filament **120** and a collector **130** configured to collect electricity flowing along the plasma filament **120**. The plasma filament **120** forms a conducting path between ground and a cloud **140**.

The plasma filament **120** comprises an electrically conducting plasma filament. The plasma comprises of electrons and positive ions. The electrically conducting plasma filament may be formed in atmosphere using an ultra short pulse laser (USPL). The USPL forms a plasma filament by raising the conductivity of a very small diameter channel (~100 microns or less) for a very brief time. The propagation of high power (~10^{14} W/cm^2) pulses from the USPL is accompanied by filamentation—self-channeling of femtosecond laser pulses in stable high-intensity light filaments with ~100 microns diameters (so that the Kerr effect focusing balances defocusing due to plasma formation). This filamentation keeps the beam virtually free of diffraction divergence. The long plasma strings have high electron densities (10^{16}/cm^3) which support high conductivities. The ionization of air at these intensities results from multi-photon processes. Filamentation has been produced and detected at altitudes as high as 13-20 km. Almost no beam scattering occurs in filamentation phenomenon, and virtually all the pulse energy is invested in ionization of air molecules. With the typical electron density and channel diameter, a 160 mJ pulse is sufficient to generate a 1 km long channel.

Once initiated, plasma filaments cannot go on forever, and typically only propagate a kilometer or so before energy depletion and exhaustion. However, by using two coaxial USPL projection optics having different time-lensing parameters, a near and far filament could be initiated in a "daisy-chain" to extend the total length of the conductive channel. The onset of filamentation in the far filament is time-lensed to coincide with the exhaustion range of the near filament. Additionally, using negative chirp, femtosecond laser pulses can propagate almost without loss until different frequency components of the wave packet, propagating at different speeds due to dispersion in air, come together at a certain desired location so that the laser intensity there exceeds the filamentation threshold, resulting in long ionized filaments created at a distance of up to several miles from the laser source. In other words, negative chirp is used to essentially focus a beam of the laser at a desired distance to form the plasma filament there. Accordingly, this technique may be used to move the starting point and to extend the length of the plasma filament.

The size, electrical conductivity, and lifetime of the plasma filaments are enhanced by an auxiliary radio-frequency electromagnetic waves or microwaves. The plasma filament may act like an antenna to absorb the auxiliary radio-frequency (RF) electromagnetic waves or microwaves.

The pulsed collection of electric charge and energy is repeated with an optimal (possibly very high) repetition rate of the pulsed laser and of the auxiliary RF waves or microwaves. The energy collected per one pulse should exceed the energy spent on creating and enhancing the plasma filaments, so that the net energy collected is positive. The optimal pulse repetition rate (PRR) depends upon atmospheric conditions which can vary widely over even short time durations. As a general rule, one tries to tune the repetition rate to the plasma relaxation time (the ion-electron recombination time) so that the conductivity of the filament remain as high as is optimal given the atmospheric conditions at the time. A feedback control system that varies the PRR as well as the rest of the pulse-shaping parameters may be employed to optimize the net power production of the system. For more information regarding pulse shaping and tailoring, which is the central theme of the field called "quantum control", please see A. Assion, T. Baumert, M. Bergt, T. Brixner, B. Kiefer, V. Seyfried, M. Strehle, G. Gerber (30 October). "Control of Chemical Reactions by Feedback-Optimized Phase-Shaped Femtosecond Laser Pulses". Science 282 (5390): 919-922. doi: 10.1126/science.282.5390.919. PMID 9794756.

FIG. **2** illustrates a second example of a system **200** for harvesting atmospheric electricity. A plasma filament **120** and a lightning **210** together form a conducting path between ground and a cloud **140**. A condensing mirror **220** directs the beam of laser **110** towards the cloud **140**. A collector **130** is configured to collect electricity flowing along the plasma filament **120**.

In the case of electricity collection from the cloud **140**, the initial and RF-enhanced plasma filaments can be relatively short, not reaching all the way from ground to the cloud **140**. The relatively short plasma filament can be optimized so that it would trigger propagation of the so-called leader (as in natural lightning), which would effectively use the energy and potential of the cloud to eventually create an electrically conducting path between ground and the cloud **140**. The final conducting path would comprise of a plasma filament **120** and a lightning **210**.

It is not shown in the figures, but alternatively both the laser **110** and the collector **130** may be placed on a mobile device, which moves together with the cloud. One such mobile device may be a vehicle moving on the ground, but another such mobile device may be an airplane flying in the air. In the case of an airplane flying in the air, plasma filaments may be formed by the laser to reach both ground and the cloud at the same time. In the event the plasma filament directed at ground is not long enough to reach ground, a lightning may also be formed to complete the conducting path to ground.

FIG. **3**A illustrates a third example of a system **300** for harvesting atmospheric electricity. A laser system situated on an airplane **310** forms two plasma filaments (**321**, **322**) that together complete a conducting path between a first region **331** of a cloud **330** and a second region **332** of the cloud **330**. The first region **331** of the cloud **330** and the second region **332** of the cloud **330** carry opposite electric charges. In the example shown here, the first region **331** of the cloud **330** carries positive electric charges, while the second region **332** of the cloud **330** carries negative electric charges. A collector configured to collect electricity flowing along the two plasma filaments is also situated on the same airplane **310** as the laser. Positive and negative charges are naturally separated within the cloud, so this example makes use of the electrostatic potential gradient between different regions of the same cloud.

FIG. **3**B illustrates a fourth example of the system **300** for harvesting atmospheric electricity. It is similar to FIG. **3**A, except two lightnings (**341**, **342**) complete the conducting path between a first region **331** of a cloud **330** and a second region **332** of the cloud **330** (i.e., a first lightning **341** electrically connects the first region **331** of the cloud **330** to one of the plasma filaments **321** and a second lightning **342** electrically connects the second region **332** of the cloud **330** to the other plasma filaments **322**). The lightning formation is again because the initial and RF-enhanced plasma filaments can be relatively short, not reaching all the way from the collector situated on airplane **310** to the cloud **330**. These relatively short plasma filaments can be optimized so that they would trigger propagation of the so-called leader (as in natural lightning), which would eventually create an electrically conducting path between the two regions of cloud **330** via a connection through airplane **310**. The final conducting path would comprise of two plasma filaments (**321**, **322**) and two lightnings (**341**, **342**). Of course, one of the lightnings may not be formed, if one of the plasma filament can reach all the way from airplane **310** to cloud **330**.

FIG. **4** illustrates a fifth example of the system **300** for harvesting atmospheric electricity. A laser system situated on an airplane **310** forms two plasma filaments (**421**, **431**) that together complete a conducting path between a first cloud **420** and a second cloud **430**. The region **422** of the first cloud **420** that is connected to the first plasma filament **421** and the region **432** of the second cloud **430** that is connected to the second plasma filament **431** carry opposite electric charges. In the example shown here, region **422** of cloud **420** carries positive electric charges, while region **432** of cloud **430** carry negative electric charges. Positive and negative charges are naturally separated within the cloud, so this example makes use of the electrostatic potential gradient between different regions of different clouds. Once again, in case one of the plasma filament cannot reach all the way from airplane **310** to any one of the clouds, then lightning may be formed to complete an electrically conducting path from airplane **310** to any one of the clouds. But these embodiments are not shown in the figures.

FIG. **5** illustrates a sixth example of the system **300** for harvesting atmospheric electricity. A laser system situated on an airplane **310** forms a plasma filament **501** that completes a conducting path between a first region **510** of atmosphere and a second region **520** of atmosphere. A collector situated on the same airplane **310** as the laser system can collect electricity flowing along the plasma filament **501**, because a vertical voltage gradient exists between the first region **510** of atmosphere and the second region **520** of atmosphere. Because no clouds are needed for this embodiment, atmospheric electricity collection is possible under fair weather conditions (relying on the naturally existing vertical voltage gradient in the atmosphere) as well as under foul weather conditions.

FIG. **6** illustrates an example of a collector **130** configured to collect electricity flowing along a plasma filament **620**. The collector may be a charge collection circuit characterized by capacitance **631**, inductance **632**, and resistance **633**. The electricity may be collected by charging a capacitor associated with capacitance **631**. Battery **610** represents these naturally existing voltage sources: (1) between ground and a cloud, (2) between differently charged regions of the same cloud, (3) between differently charged regions of different clouds, and (4) between different regions of atmosphere, where there is a vertical voltage gradient. The plasma filament **620** can, for the purpose of analysis and optimization of the charge collection process, be viewed as electromagnetic transmission lines characterized by their effective capacitance, inductance, and resistance per unit length. Accordingly, the parameters of the charge collection circuit and of the effective transmission line representing the plasma filaments are optimized, so as to maximize the net charge and energy collected (minus the energy spent on generating and enhancing the plasma filaments).

The foregoing description is provided to enable a person skilled in the art to practice the various configurations described herein. While the subject technology has been particularly described with reference to the various figures and configurations, it should be understood that these are for illustration purposes only and should not be taken as limiting the scope of the subject technology.

There may be many other ways to implement the subject technology. Various functions and elements described herein may be partitioned differently from those shown without departing from the scope of the subject technology. Various modifications to these configurations will be readily apparent to those skilled in the art, and generic principles defined herein may be applied to other configurations. Thus, many changes and modifications may be made to the subject technology, by one having ordinary skill in the art, without departing from the scope of the subject technology.

It is understood that the specific order or hierarchy of steps in the processes disclosed is an illustration of exemplary approaches. Based upon design preferences, it is understood that the specific order or hierarchy of steps in the processes may be rearranged. Some of the steps may be performed simultaneously. The accompanying method claims present elements of the various steps in a sample order, and are not meant to be limited to the specific order or hierarchy presented.

A phrase such as "an aspect" does not imply that such aspect is essential to the subject technology or that such aspect applies to all configurations of the subject technology. A disclosure relating to an aspect may apply to all configurations, or one or more configurations. An aspect may provide one or more examples of the disclosure. A phrase such as an "aspect" may refer to one or more aspects and vice versa. A phrase such as an "embodiment" does not imply that such embodiment is essential to the subject technology or that such embodiment applies to all configurations of the subject technology. A disclosure relating to an embodiment may apply to all embodiments, or one or more embodiments. An embodiment may provide one or more examples of the disclosure. A phrase such an "embodiment" may refer to one or more embodiments and vice versa. A phrase such as a "configuration" does not imply that such configuration is essential to the subject technology or that such configuration applies to all configurations of the subject technology. A disclosure relating to a configuration may apply to all configurations, or one or more configurations. A configuration may provide one or more examples of the disclosure. A phrase such as a "configuration" may refer to one or more configurations and vice versa.

Furthermore, to the extent that the term "include," "have," or the like is used in the description or the claims, such term is intended to be inclusive in a manner similar to the term "comprise" as "comprise" is interpreted when employed as a transitional word in a claim.

The word "exemplary" is used herein to mean "serving as an example, instance, or illustration." Any embodiment described herein as "exemplary" is not necessarily to be construed as preferred or advantageous over other embodiments.

A reference to an element in the singular is not intended to mean "one and only one" unless specifically stated, but rather "one or more." The term "some" refers to one or more. Underlined and/or italicized headings and subheadings are used for convenience only, do not limit the subject technology, and are not referred to in connection with the interpretation of the description of the subject technology. All structural and functional equivalents to the elements of the various configurations described throughout this disclosure that are known or later come to be known to those of ordinary skill in the art are expressly incorporated herein by reference and intended to be encompassed by the subject technology. Moreover, nothing disclosed herein is intended to be dedicated to the public regardless of whether such disclosure is explicitly recited in the above description.

What is claimed is the following:

1. A system for harvesting atmospheric electricity, comprising:

a laser configured to form a plasma filament comprising at least a first filament and a second filament, wherein onset of formation of the second filament coincides with an exhaustion range of the first filament;

an auxiliary source of electromagnetic energy operable to enhance the plasma filament by applying electromagnetic energy to the plasma filament; and

a collector configured to collect electricity flowing along the plasma filament.

2. The system of claim **1**, wherein the plasma filament comprises an electrically conducting plasma filament.

3. The system of claim **1**, wherein the laser is a pulsed laser.

4. The system of claim **3**, wherein the pulsed laser is pulsed at an optimal repetition rate.

5. The system of claim **1**, wherein the electromagnetic energy comprises at least one of radio-frequency energy or microwave energy.

6. The system of claim **1**, wherein the plasma filament forms at least a part of a conducting path between ground and a cloud.

7. The system of claim **6**, wherein at least one lightning forms another part of the conducting path between ground and the cloud.

8. The system of claim **1**, wherein the plasma filament forms at least a part of a conducting path between a first region of a cloud and a second region of the cloud.

9. The system of claim **8**, wherein at least one lightning forms another part of the conducting path between the first region of the cloud and the second region of the cloud.

10. The system of claim **1**, wherein:

the laser is further configured to form a second plasma filament,

the plasma filament is connected to a region of a first cloud, and

the second plasma filament is connected to a region of a second cloud.

11. The system of claim **1**, wherein the plasma filament completes a conducting path between a first region of atmosphere and a second region of atmosphere.

12. The system of claim **11**, wherein a vertical voltage gradient exists between the first region of atmosphere and the second region of atmosphere.

13. The system of claim **1**, wherein the collector comprises a charge collection circuit.

14. The system of claim **1**, wherein the laser is situated on ground.

15. The system of claim **1**, wherein the laser is situated on a mobile device.

16. The system of claim **15**, wherein the mobile device is an airplane.

17. The system of claim **1**, wherein negative chirp is used to move the starting point of formation of the plasma filament to a desired location.

18. A method for harvesting atmospheric electricity, the method comprising:

forming a plasma filament with a laser, wherein the plasma filament comprises at least a first filament and a second filament, and wherein onset of formation of the second filament coincides with an exhaustion range of the first filament;

applying electromagnetic energy to the plasma filament to enhance the plasma filament; and

collecting electricity flowing along the plasma filament.

19. An apparatus for harvesting atmospheric electricity, the apparatus comprising:

means for forming a plasma filament with a laser, wherein the plasma filament comprises at least a first filament and a second filament, and wherein onset of formation of the second filament coincides with an exhaustion range of the first filament;

means for applying electromagnetic energy to the plasma filament to enhance the plasma filament; and

means for collecting electricity flowing along the plasma filament.

20. The system of claim **1**, wherein the laser is an ultra short pulse laser (USPL).

21. The system of claim **1**, wherein the laser comprises two coaxial USPL projection optics having different time-lensing parameters.

* * * * *

US005590031A

United States Patent [19]

Mead, Jr. et al.

[11] **Patent Number: 5,590,031**

[45] **Date of Patent: Dec. 31, 1996**

[54] SYSTEM FOR CONVERTING ELECTROMAGNETIC RADIATION ENERGY TO ELECTRICAL ENERGY

[76] Inventors: **Franklin B. Mead, Jr.**, 44536 Avenida Del Sol, Lancaster, Calif. 93535; **Jack Nachamkin**, 12314 Teri Dr., Poway, Calif. 92064

[21] Appl. No.: **281,271**

[22] Filed: **Jul. 27, 1994**

[51] **Int. Cl.**^{6} **H02M 1/00**

[52] **U.S. Cl.** **363/8**; 363/178; 342/6

[58] **Field of Search** 363/8, 13, 178; 342/6, 61, 73, 173, 175

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,882,503	5/1975	Gamara	343/100 R
4,725,847	2/1988	Poirier	343/840
5,008,677	4/1991	Trigon et al.	342/17

Primary Examiner—Peter S. Wong
Assistant Examiner—Adolf Berhane
Attorney, Agent, or Firm—Chris Papageorge

[57] **ABSTRACT**

A system is disclosed for converting high frequency zero point electromagnetic radiation energy to electrical energy. The system includes a pair of dielectric structures which are positioned proximal to each other and which receive incident zero point electromagnetic radiation. The volumetric sizes of the structures are selected so that they resonate at a frequency of the incident radiation. The volumetric sizes of the structures are also slightly different so that the secondary radiation emitted therefrom at resonance interfere with each other producing a beat frequency radiation which is at a much lower frequency than that of the incident radiation and which is amenable to conversion to electrical energy. An antenna receives the beat frequency radiation. The beat frequency radiation from the antenna is transmitted to a converter via a conductor or waveguide and converted to electrical energy having a desired voltage and waveform.

14 Claims, 8 Drawing Sheets

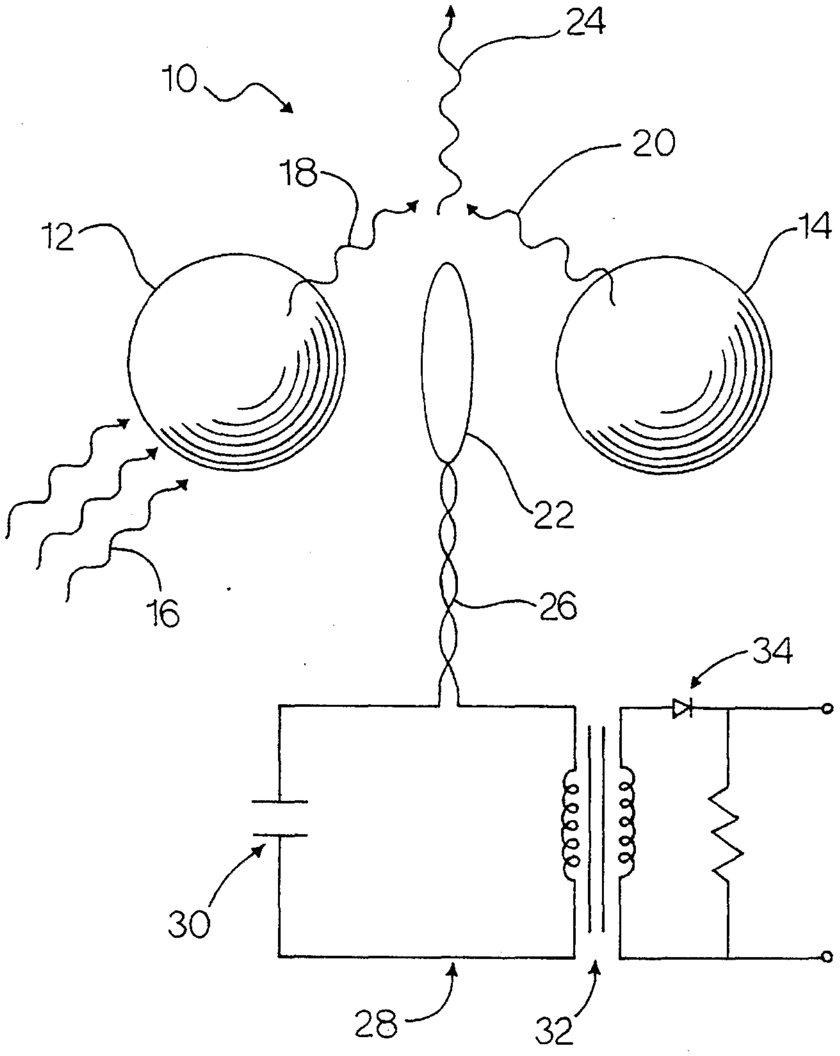

FIG. 1

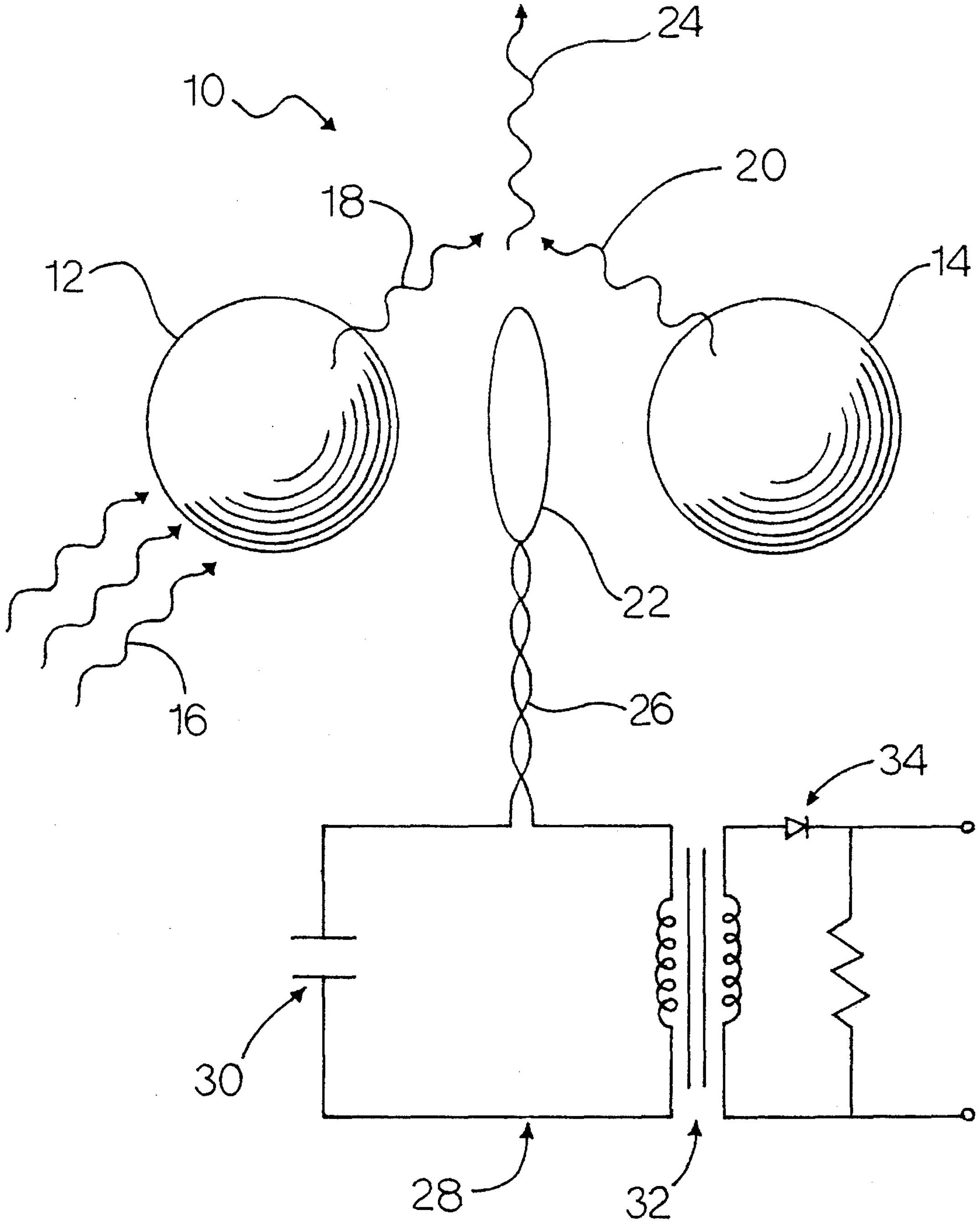
24
10
20
18
12
14
22
16
26
34
30
28
32

FIG. 2

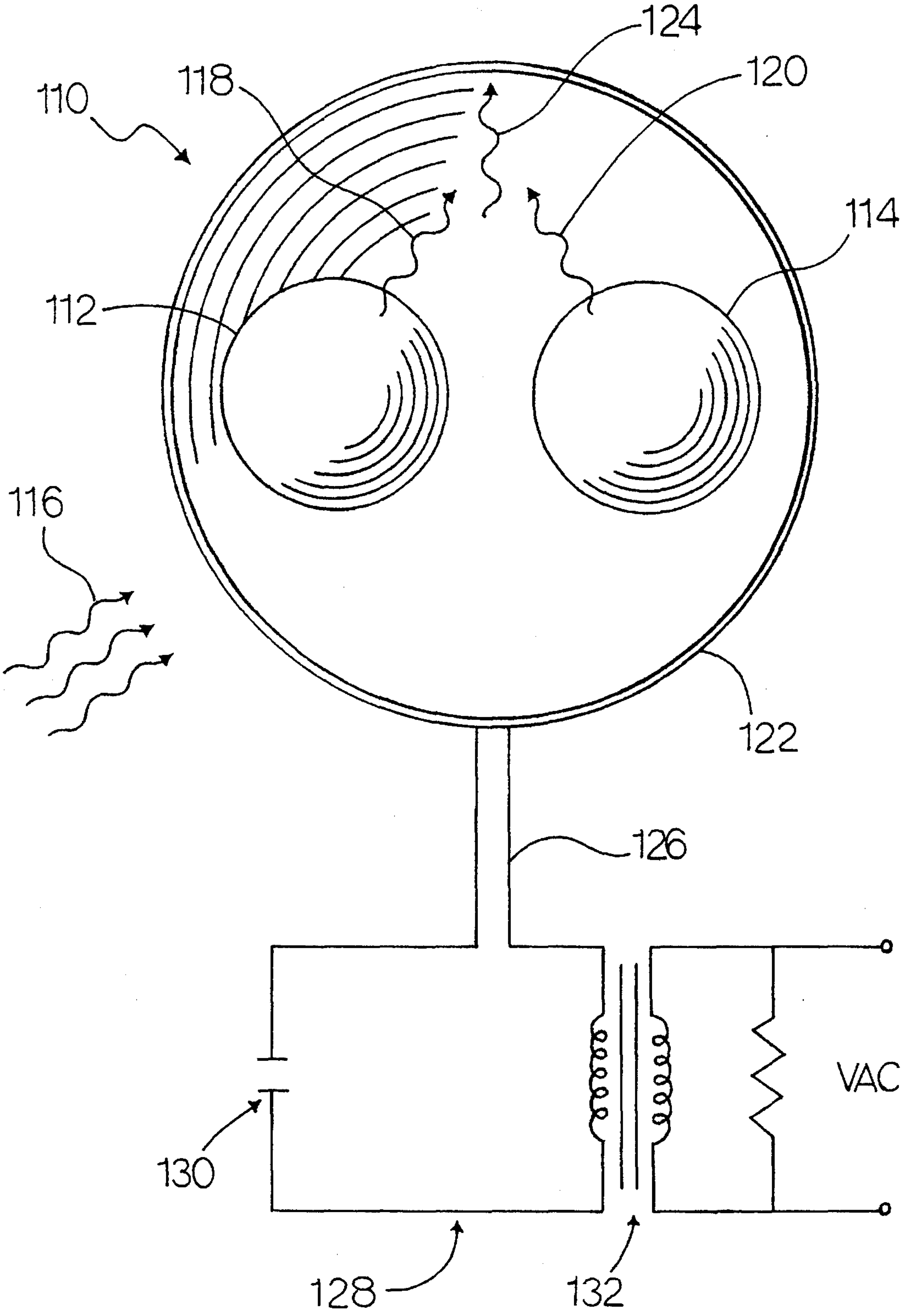
124
110
118
120
114
112
116
122
126
VAC
130
128
132

FIG. 3

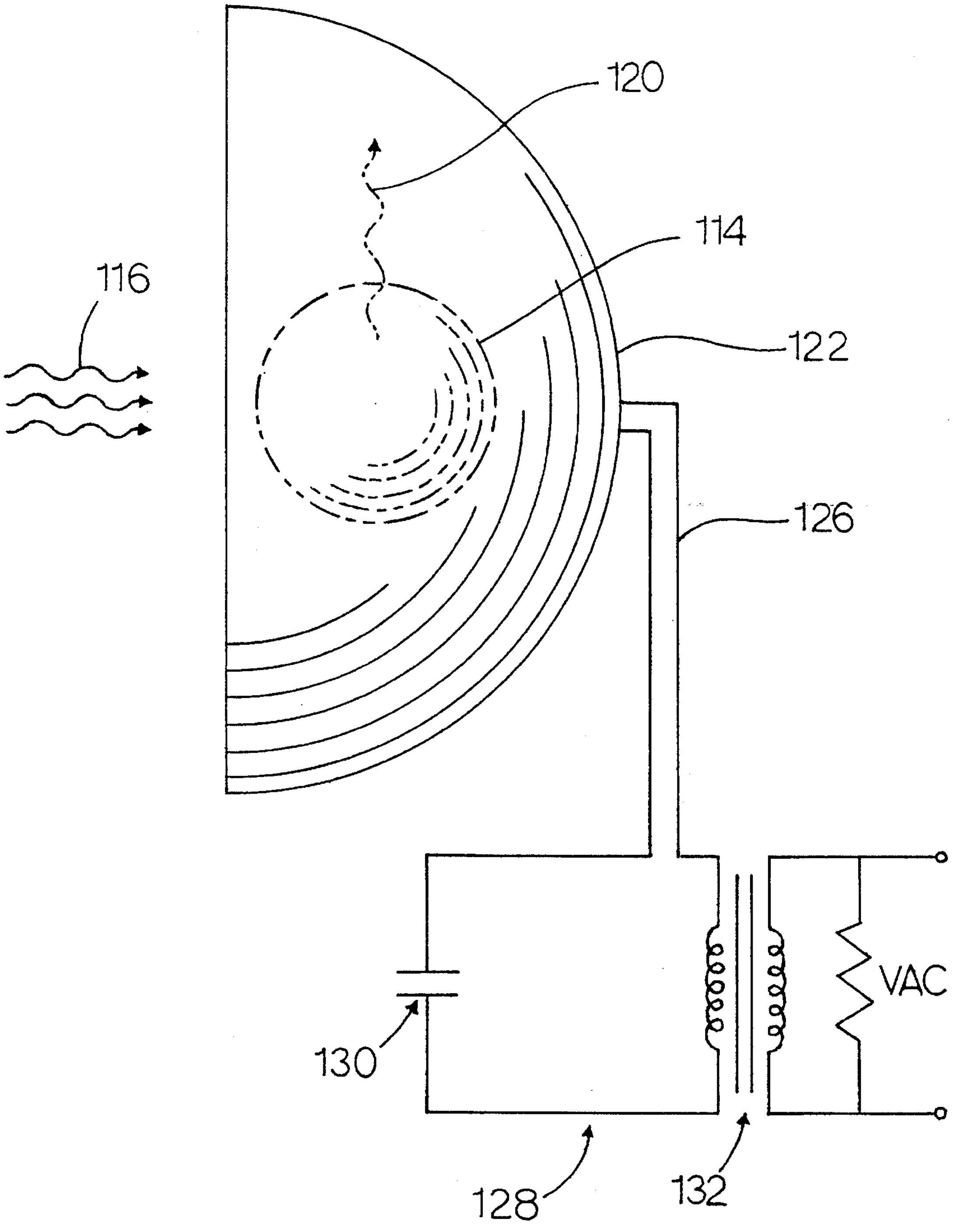
120
114
116
122
126
VAC
130
128
132

FIG. 4

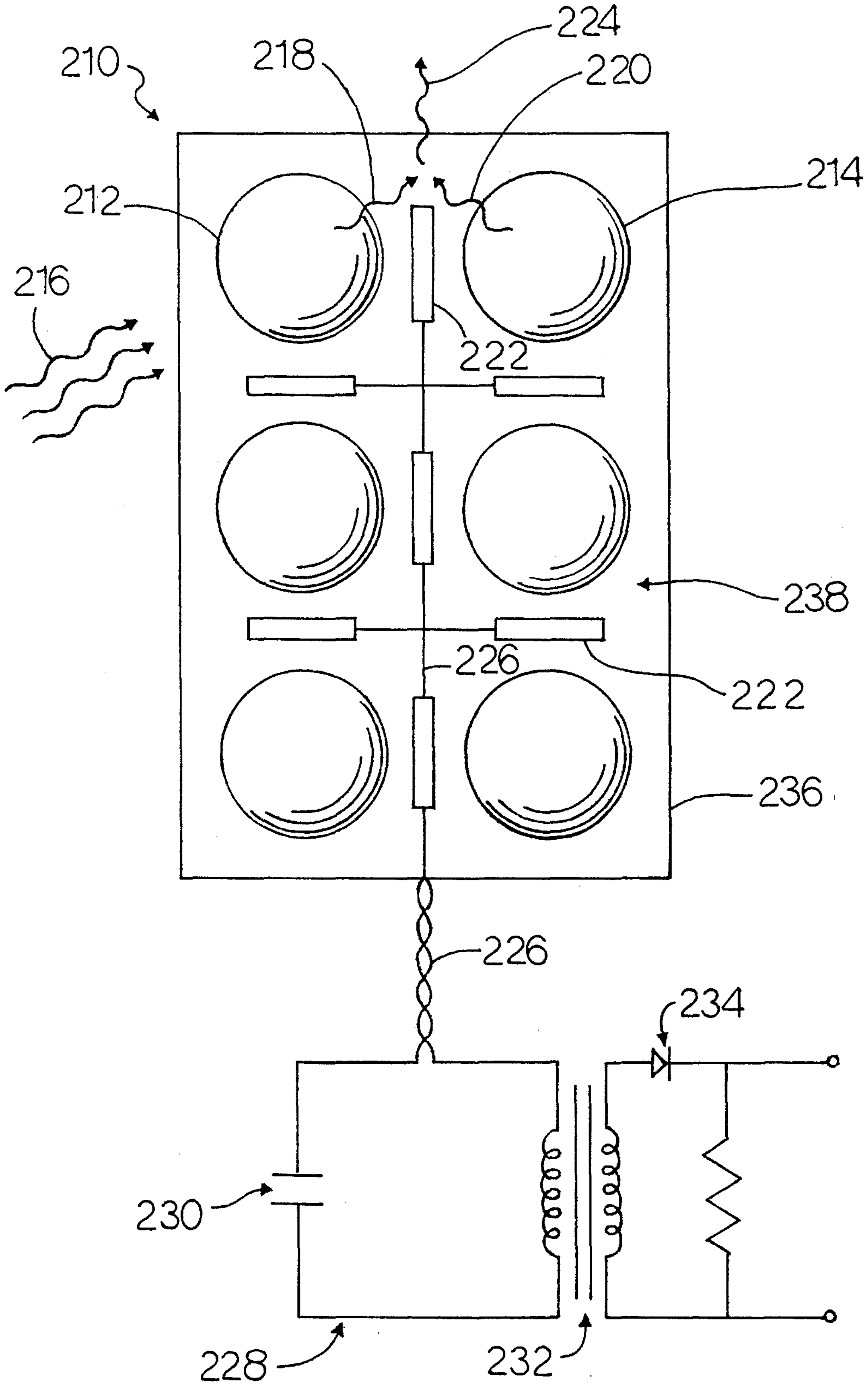

FIG. 5

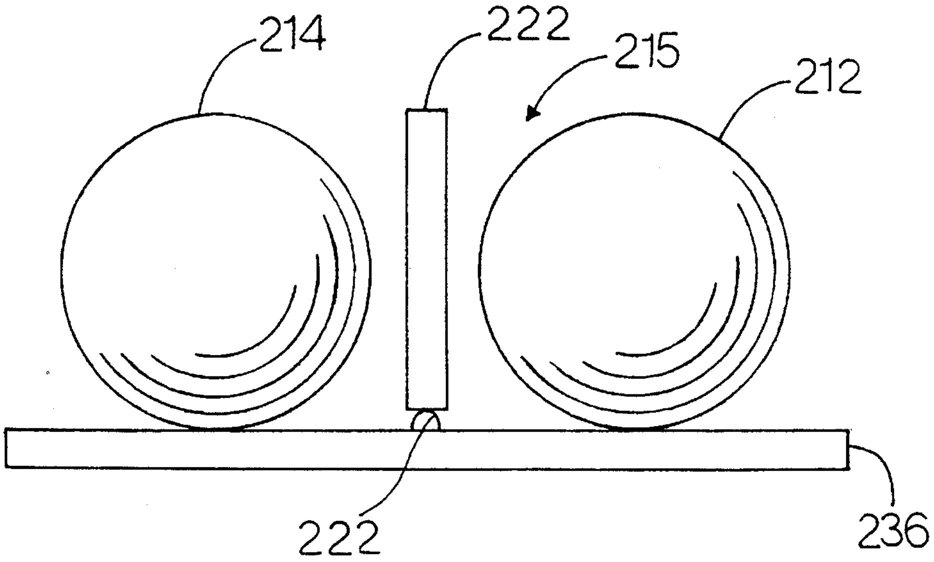
214
222
215
212
222
236

FIG. 6

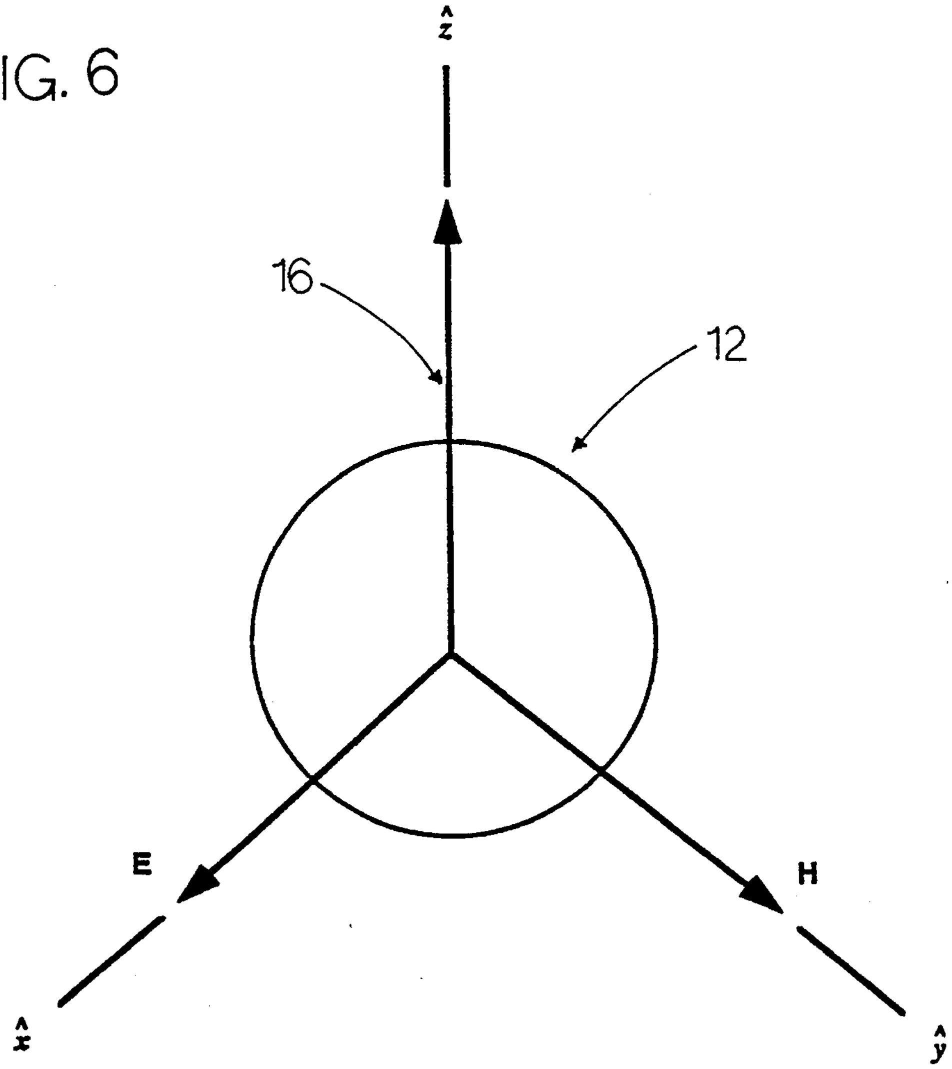
$\hat{z}$
16
12
E
H
$\hat{x}$
$\hat{y}$

FIG. 7

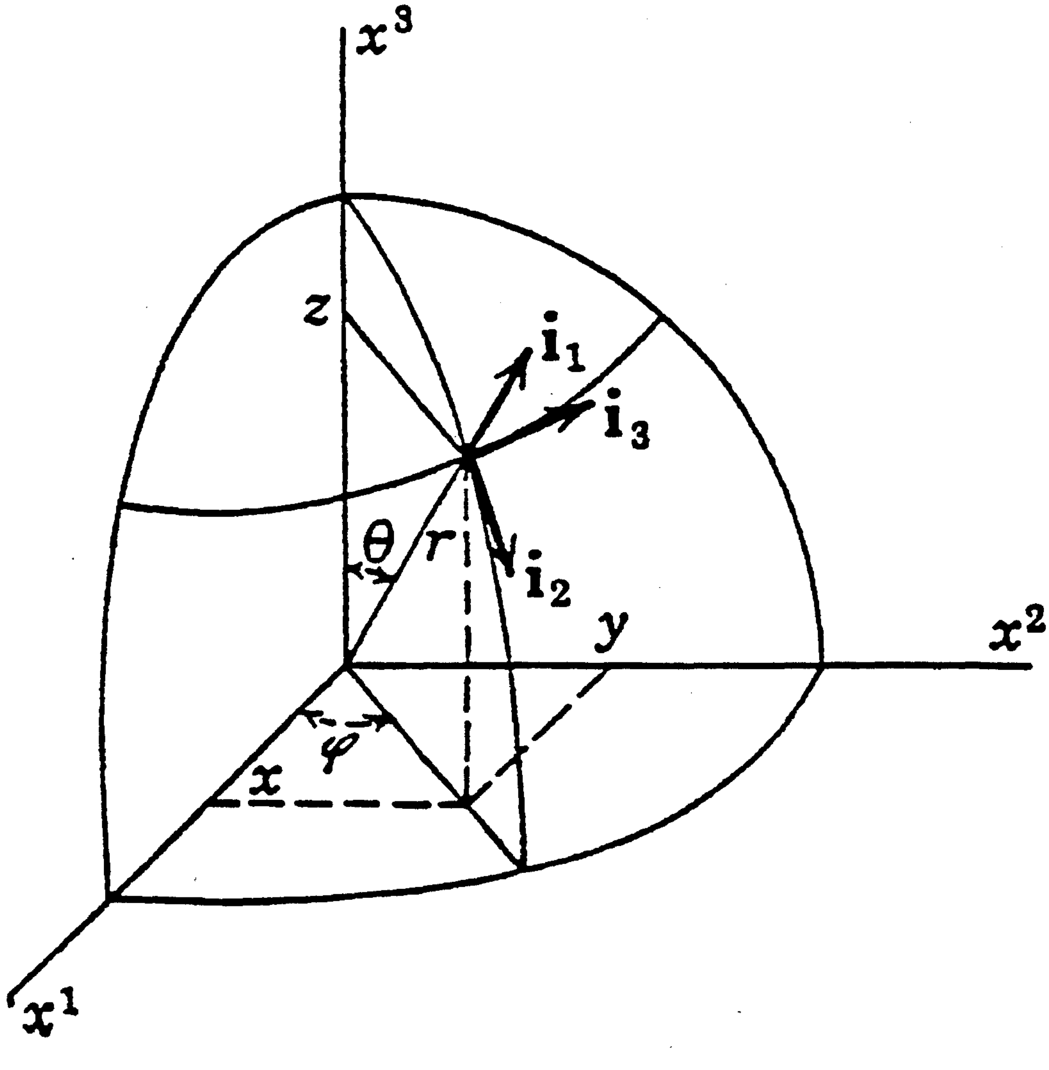
x^3
z
$\mathbf{i}_1$
$\mathbf{i}_3$
θ
r
$\mathbf{i}_2$
y
x^2
φ
x
x^1

FIG. 8

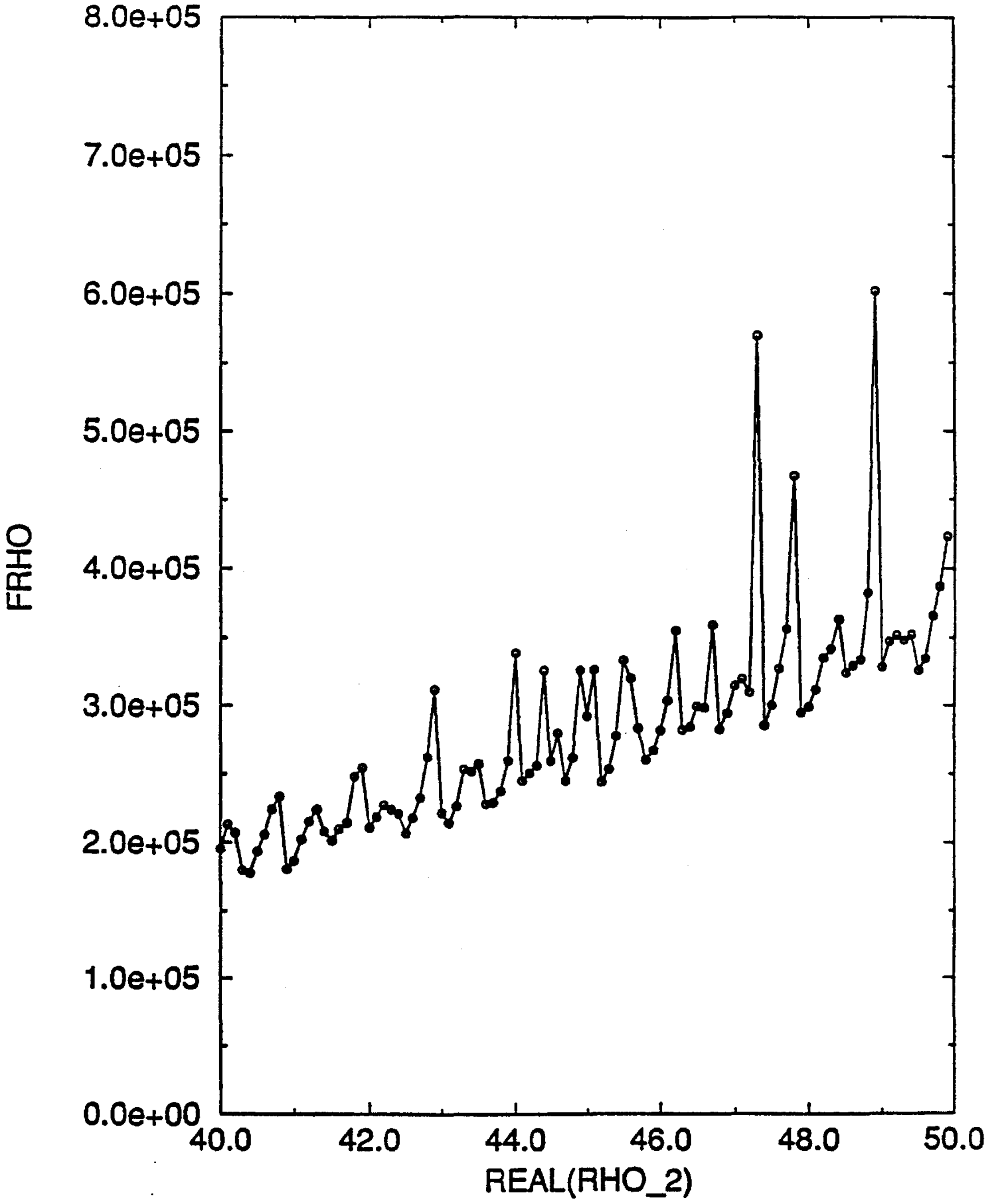
8.0e+05
7.0e+05
6.0e+05
5.0e+05
4.0e+05
3.0e+05
2.0e+05
1.0e+05
0.0e+00
FRHO
40.0
42.0
44.0
46.0
48.0
50.0
REAL(RHO_2)

FIG.9

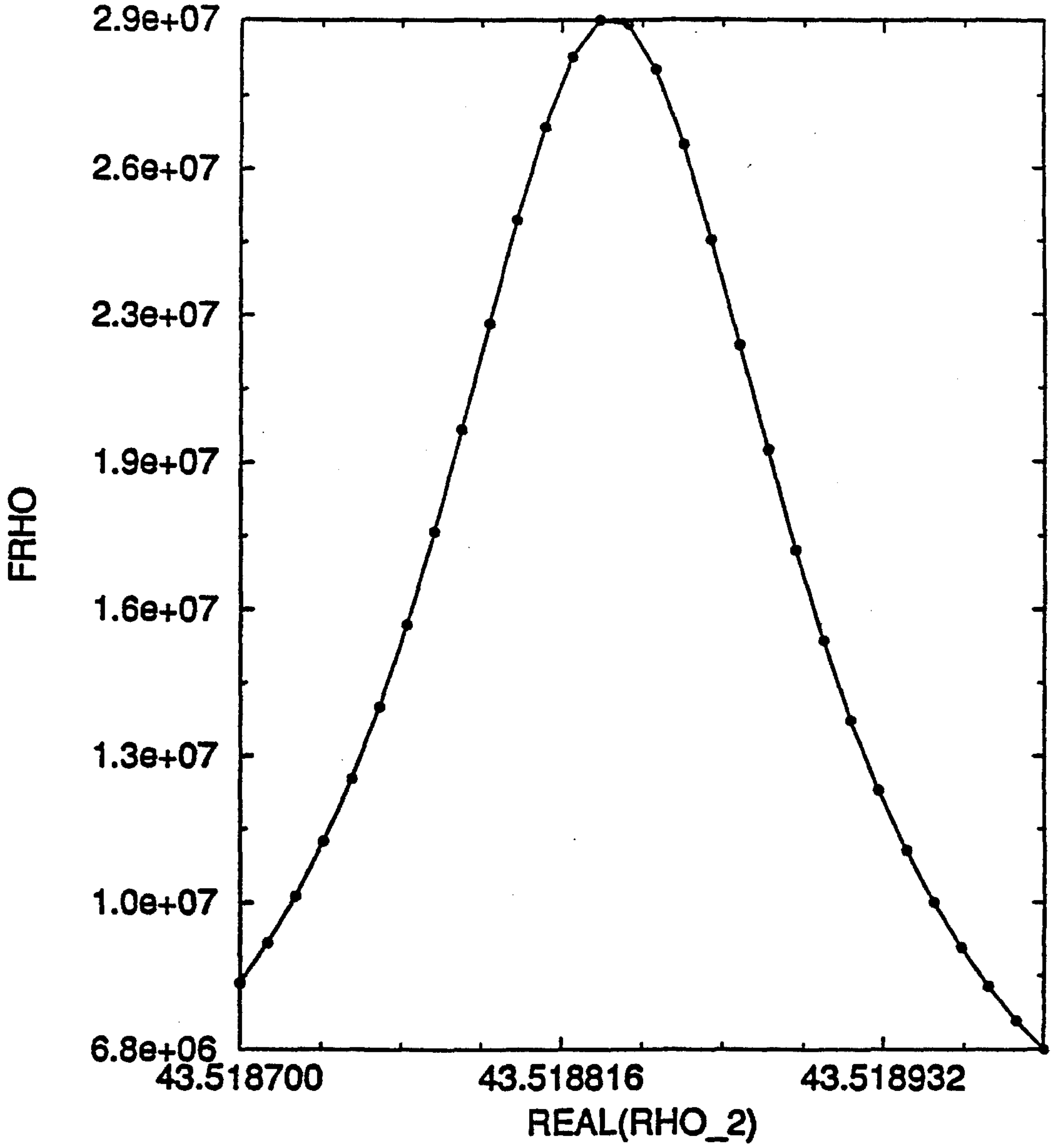
2.9e+07
2.6e+07
2.3e+07
1.9e+07
1.6e+07
1.3e+07
1.0e+07
6.8e+06
FRHO
43.518700
43.518816
43.518932
REAL(RHO_2)

SYSTEM FOR CONVERTING ELECTROMAGNETIC RADIATION ENERGY TO ELECTRICAL ENERGY

BACKGROUND OF THE INVENTION

The invention relates generally to conversion of electromagnetic radiation energy to electrical energy, and, more particularly, to conversion of high frequency bandwidths of the spectrum of a type of radiation known as zero point electromagnetic radiation to electrical energy.

The existence of zero point electromagnetic radiation was discovered in 1958 by the Dutch physicist M. J. Sparnaay. Mr. Sparnaay continued the experiments carried out by Hendrik B. G. Casimir in 1948 which showed the existence of a force between two uncharged parallel plates which arose from electromagnetic radiation surrounding the plates in a vacuum. Mr. Sparnaay discovered that the forces acting on the plates arose from not only thermal radiation but also from another type of radiation now known as classical electromagnetic zero point radiation. Mr. Sparnaay determined that not only did the zero point electromagnetic radiation exist in a vacuum but also that it persisted even at a temperature of absolute zero. Because it exists in a vacuum, zero point radiation is homogeneous and isotropic as well as ubiquitous. In addition, since zero point radiation is also invariant with respect to Lorentz transformation, the zero point radiation spectrum has the characteristic that the intensity of the radiation at any frequency is proportional to the cube of that frequency. Consequently, the intensity of the radiation increases without limit as the frequency increases resulting in an infinite energy density for the radiation spectrum. With the introduction of the zero point radiation into the classical electron theory, a vacuum at a temperature of absolute zero is no longer considered empty of all electromagnetic fields. Instead, the vacuum is now considered as filled with randomly fluctuating fields having the zero point radiation spectrum. The special characteristics of the zero point radiation which are that it has a virtually infinite energy density and that it is ubiquitous (even present in outer space) make it very desirable as an energy source. However, because high energy densities exist at very high radiation frequencies and because conventional methods are only able to convert or extract energy effectively or efficiently only at lower frequencies at which zero point radiation has relatively low energy densities, effectively tapping this energy source has been believed to be unavailable using conventional techniques for converting electromagnetic energy to electrical or other forms of easily useable energy. Consequently, zero point electromagnetic radiation energy which may potentially be used to power interplanetary craft as well as provide for society's other needs has remained unharnessed.

There are many types of prior art systems which use a plurality of antennas to receive electromagnetic radiation and provide an electrical output therefrom. An example of such a prior art system is disclosed in U.S. Pat. No. 3,882,503 to Gamara. The Gamara system has two antenna structures which work in tandem and which oscillate by means of a motor operatively attached thereto in order to modulate the radiation reflected from the antenna surfaces. The reflecting surfaces of the antennas are also separated by a distance equal to a quarter wavelength of the incident radiation. However, the Gamara system does not convert the incident radiation to electrical current for the purpose of converting the incident electromagnetic radiation to another form of readily useable energy. In addition, the relatively large size of the Gamara system components make it unable to resonate at and modulate very high frequency radiation.

What is therefore needed is a system which is capable of converting high frequency electromagnetic radiation energy into another form of energy which can be more readily used to provide power for transportation, heating, cooling as well as various other needs of society. What is also needed is such a system which may be used to provide energy from any location on earth or in space.

SUMMARY OF THE INVENTION

It is a principle object of the present invention to provide a system for converting electromagnetic radiation energy to electrical energy.

It is another object of the present invention to provide a system for converting electromagnetic radiation energy having a high frequency to electrical energy.

It is another object of the present invention to provide a system for converting zero point electromagnetic radiation energy to electrical energy.

It is another object of the present invention to provide a system for converting electromagnetic radiation energy to electrical energy which may used to provide such energy from any desired location on earth or in space.

It is another object of the present invention to provide a system for converting electromagnetic radiation energy to electrical energy having a desired waveform and voltage.

It is an object of the present invention to provide a miniaturized system for converting electromagnetic radiation energy to electrical energy in order to enhance effective utilization of high energy densities of the electromagnetic radiation.

It is an object of the present invention to provide a system for converting electromagnetic radiation energy to electrical energy which is simple in construction for cost effectiveness and reliability of operation.

Essentially, the system of the present invention utilizes a pair of structures for receiving incident electromagnetic radiation which may be propagating through a vacuum or any other medium in which the receiving structures may be suitably located. The system of the present invention is specifically designed to convert the energy of zero point electromagnetic radiation; however, it may also be used to convert the energy of other types of electromagnetic radiation. The receiving structures are preferably composed of dielectric material in order to diffract and scatter the incident electromagnetic radiation. In addition, the receiving structures are of a volumetric size selected to enable the structures to resonate at a high frequency of the incident electromagnetic radiation based on the parameters of frequency of the incident radiation and propagation characteristics of the medium and of the receiving structures. Since zero point radiation has the characteristic that its energy density increases as its frequency increases, greater amounts of electromagnetic energy are available at higher frequencies. Consequently, the size of the structures are preferably miniaturized in order to produce greater amounts of energy from a system located within a space or area of a given size. In this regard, the smaller the size of the receiving structures, the greater the amount of energy that can be produced by the system of the present invention.

At resonance, electromagnetically induced material deformations of the receiving structures produce secondary fields of electromagnetic energy therefrom which may have evanescent energy densities several times that of the incident radiation. The structures are of different sizes so that the secondary fields arising therefrom are of different frequencies. The difference in volumetric size is very small so that interference between the two emitted radiation fields, and the receiving structures at the two different frequencies produces a beat frequency radiation which has a much lower frequency than the incident radiation. The beat frequency radiation preferably is at a frequency which is sufficiently low that it may be relatively easily converted to useable electrical energy. In contrast, the incident zero point radiation has its desirable high energy densities at frequencies which are so high that conventional systems for converting the radiation to electrical energy either cannot effectively or efficiently so convert the radiation energy or simply cannot be used to convert the radiation energy for other reasons.

The system of the present invention also includes an antenna which receives the beat frequency radiation. The antenna may be a conventional metallic antenna such as a loop or dipole type of antenna or a rf cavity structure which partially encloses the receiving structures. The antenna feeds the radiation energy to an electrical conductor (in the case of a conventional dipole or comparable type of antenna) or to a waveguide (in the case of a rf cavity structure). The conductor or waveguide feeds the electrical current (in the case of the electrical conductor) or the electromagnetic radiation (in the case of the waveguide) to a converter which converts the received energy to useful electrical energy. The converter preferably includes a tuning circuit or comparable device so that it can effectively receive the beat frequency radiation. The converter may include a transformer to convert the energy to electrical current having a desired voltage. In addition, the converter may also include a rectifier to convert the energy to electrical current having a desired waveform.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** is a plan view of the receiving structures and antenna of a first embodiment of the system of the present invention with a schematic view of the conductor and converter thereof and also showing the incident primary and emitted secondary electromagnetic radiation.

FIG. **2** is a front view of the receiving structures, antenna and waveguide of a second embodiment of the system of the present invention with a schematic view of the converter thereof and also showing the incident primary and emitted secondary electromagnetic radiation.

FIG. **3** is a perspective view of the receiving structures, antenna and waveguide of the second embodiment shown in FIG. **2** with a schematic view of the converter thereof and also showing the incident primary and emitted secondary electromagnetic radiation.

FIG. **4** is a front view of the substrate and a plurality of pairs of the receiving structures and a plurality of antennas of a third embodiment of the system of the present invention with a schematic view of the conductor and converter thereof and also showing the incident primary and emitted secondary electromagnetic radiation.

FIG. **5** is a top view of some of the components of the third embodiment of the system of the present invention showing two of the plurality of pairs of receiving structures and two of the plurality of antennas mounted on the substrate.

FIG. **6** is a diagram of a receiving structure of the system of the present invention showing an incident electromagnetic plane wave impinging on the receiving structure and illustrating the directions of the electric and magnetic field vectors thereof.

FIG. **7** is a diagram of a spherical coordinate system as used in the formulas utilized in the system of the present invention.

FIG. **8** is a graph showing an imaginary ρ parameter plotted against a real ρ parameter illustrating the values thereof at resonance as well as values thereof at other than resonance.

FIG. **9** is a graph showing a portion of the graphical representation shown in FIG. **8** illustrating the real and imaginary ρ values at or near a single resonance.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring to the drawings, a first embodiment of the present invention is generally designated by the numeral **10**. The system **10** includes a first and second means for receiving **12** and **14** incident electromagnetic radiation **16**. The means for receiving **12** and **14** are preferably a pair of spherical structures **12** and **14** which are preferably composed of a dielectric material. Alternatively, the spheres **12** and **14** may be cubical structures or any other suitable shape. The spheres **12** and **14** may be mounted on a suitable foundation by any suitable mounting means (not shown), or spheres **12** and **14** may be suspended from a suitable foundation by any suitable suspension means (not shown). The spheres **12** and **14** are preferably composed of a dielectric material. The dielectric spheres **12** and **14** scatter and concentrate electromagnetic waves. At very sharpely defined frequencies, the spheres **12** and **14** will have resonances wherein the internal energy densities can be five orders of mangitude larger than the energy density of the incident electromagnetic field driving the spheres **12** and **14**. At resonance, the electromagnetic stresses, equivalent to pressures proportional to the energy density, can cause material deformation of the spheres **12** and **14** which produce a secondary electromagnetic field. The spheres **12** and **14** are preferably positioned proximal to each other, as shown in FIG. **1**. Although the proximity of the spheres to each other will adversely affect the resonances, the very high "Q"s of the isolated-sphere resonances results in such adverse affect being relatively small. However, the proximity of the spheres **12** and **14** allows the spheres to interact electromechanically which increases the magnitude of the secondary radiation emitted therefrom.

The electromagnetic radiation incident upon the spheres **12** and **14** which drives the spheres to resonance is preferably zero point radiation **16**. However, other types of electromagnetic radiation may also be used to drive the spheres **12** and **14**, if desired.

The effect of a dielectric sphere such as **12** or **14** on an incident electromagnetic radiation such as a plane wave thereof is shown in FIG. **6**. The plane wave propagates in the z axis direction and is diffracted by the sphere **12** resulting in scattering thereof. This scattering is commonly known as Mie scattering. The incident radiation wave has an electric vector component which is linearly polarized in the x axis direction and a magnetic vector component which is linearly polarized in the y axis direction.

An electromagnetic wave incident upon a structure produces a forced oscillation of free and bound charges in synch with the primary electromagnetic field of the incident electromagnetic wave. The movements of the charges produce a secondary electromagnetic field both inside and outside the structure. The secondary electromagnetic radiation comprising this secondary electromagnetic field is shown in FIG. **1** and designated by the numerals **18** and **20**. An antenna which is shown simply as a loop antenna but may also be a dipole or any other suitable type of antenna is also shown in FIG. **1** and designated by the numeral **22**. The nonlinear mutual interactions of the spheres produces interference between the secondary electromagnetic radiation **18** and **20** produces a beat frequency radiation **24** which is preferably at a much lower frequency than the primary radiation **16**. It is this beat frequency radiation **24** which is desired for conversion into electrical energy because it preferably is within the frequency range of rf radiation which may be converted into electrical energy by generally conventional systems. Thus, the radiation **24** received by the antenna **22** is fed via an electrical conductor **26** to a means for converting the beat frequency radiation **24** to electrical energy. This means for converting is designated by the numeral **28** and preferably includes a tuning capacitor **30** and a transformer **32** and a rectifier (preferably a diode) **34**. Instead of including the capacitor **30**, transformer **32** and rectifier **34**, the converter **28** may alternatively include an rf receiver of any suitable type.

The resultant field at any point is the vector sum of the primary and secondary fields. For the equations that follow, the structure receiving the incident plane wave is a sphere of radius a having a propagation constant k_1 positioned in an infinite, homogeneous medium having a propagation constant k_2. The incident plane wave propagates in the z axis direction and is as shown in FIG. **6**. The spherical coordinate system used for the vector spherical wave functions is shown in FIG. **7**. Expansion of the incident field provides:

$$E_i = E_o e^{-i\omega t} \sum_{n=1}^{\infty} i^n \frac{2n+1}{n(n+1)} \left(m^{(1)}_{01n} - i n^{(1)}_{eln}\right)$$

$$H_i = -\frac{k_2}{\omega\mu_2} E_o e^{-i\omega t} \sum_{n=1}^{\infty} i^n \frac{2n+1}{n(n+1)} \left(m^{(1)}_{eln} + i n^{(1)}_{01n}\right)$$

where E is the electric field and H is the magnetic field; and

$$m^{(1)}_{\substack{01n\\eln}} = \pm\frac{1}{\sin\theta} j_n(k_2R)P_n{}^1(\cos\theta) \begin{matrix}\cos\\ \sin\end{matrix} \phi i_2 - j_n(k_2R)\frac{\partial p_n{}^1}{\partial\theta} \begin{matrix}\sin\\ \cos\end{matrix} \phi i_3$$

$$n^{(1)}_{\substack{0ln\\eln}} = \frac{n(n+1)}{k_2R} j_n(k_2R)P_n{}^1(\cos\theta) \begin{matrix}\sin\\ \cos\end{matrix} \phi i_1 + \frac{1}{k_2R}[k_2Rj_n(k_2R)]' \times$$

$$\frac{\partial P_n{}^1}{\partial\theta} \begin{matrix}\sin\\ \cos\end{matrix} \phi i_2 \pm \frac{1}{k_2R\sin\theta}[k_2Rj_n(k_2R)]'P_n{}^1(\cos\theta) \begin{matrix}\cos\\ \sin\end{matrix} \phi i_3.$$

The electric and magnetic fields of the incident wave transmitted into the sphere i.e., R<a, can be similarly expanded:

$$E_t = E_0 e^{-i\omega t} \sum_{n=1}^{\infty} i^n \frac{2n+1}{n(n+1)} \left(a_n^{tm}\,{}^{(1)}_{01n} - i b_n^{tn}\,{}^{(1)}_{eln} \right)$$

$$H_t = \frac{k_2}{\phi\mu_1} E_0 e^{-i\omega t} \sum_{n=1}^{\infty} i^m \frac{2n+1}{n(n+1)} \left(b_n^{tm}\,{}^{(1)}_{eln} - i a_n^{tn}\,{}^{(1)}_{01n} \right)$$

If $j_n(k_2R)$ is replaced by $h_n^{(1)}(k_2R)$ in the previous equations, the functions $m^{(1)}$ and $n^{(1)}$ become $m^{(3)}$ and $n^{(3)}$. The outgoing fields i.e., R>a, are represented by:

$$E_R = E_0 e^{-i\omega t} \sum_{n=1}^{\infty} i^n \frac{2n+1}{n(n+1)} \left(a_n^{rm}\,{}^{(3)}_{01n} - i b_n^{tn}\,{}^{(3)}_{eln} \right)$$

$$H_r = \frac{k_2}{\phi\mu_1} E_0 e^{-i\omega t} \sum_{n=1}^{\infty} i^m \frac{2n+1}{n(n+1)} \left(b_n^{rm}\,{}^{(3)}_{eln} - i a_n^{tn}\,{}^{(3)}_{01n} \right)$$

where H_r represents the resultant wave in the medium surrounding the sphere. At resonance, the values of ρ at resonance require that the $a_n{}^t$ and $b_n{}^t$ coefficients be infinite. In order to determine these values of $a_n{}^t$ and $b_n{}^t$, the boundary conditions at the sphere radius are needed. Since there must be continuity of the E and H values at the surface, the following equations are used:

$$i_1\times(E_i+E_r)=i_1\times E_t \text{ and}$$

$$i_1\times(H_i+H_r)=i_1\times H_t$$

which lead to two pairs of inhomogeneous equations:

$$a_n{}^t j_n(N\rho)-a_n{}^r h_n{}^{(1)}(\rho)=j_n(\rho)$$

$$\mu_2 a_n{}^t[N\rho j_n(N\rho)]'-\mu_1 a_n{}^r[\rho h_n{}^{(1)}(\rho)]'=\mu_1[\rho j_n(\rho)]' \text{ and}$$

$$\mu_2 N b_n{}^t j_n(N\rho)-\mu_1 b_n{}^r h_n{}^{(1)}(\rho)=\mu_1 j_n(\rho)$$

$$b_n{}^t[N\rho j_n(N\rho)]'-N b_n{}^r[\rho h_n{}^{(1)}(\rho)]'=N[\rho j_n(\rho)]'$$

where $k_1=Nk_2$, $\rho=k_2a$, $k_1a=N\rho$. Spherical Bessel functions of the first kind are denoted by j_n, while those of the third kind are denoted by $h_n{}^{(1)}$. The resulting equations are:

$$a_n{}^t = \frac{\mu_1 j_n(\rho)[\rho h_n^{(1)}(\rho)]' - \mu_1 h_n^{(1)}\rho)[\rho j_n(\rho)]'}{\mu_1 j_n(N\rho)[\rho h_n^{(1)}(\rho)]' - \mu_2 h_n^{(1)}(\rho)[N\rho j_n(N\rho)]'}$$

and

$$b_n{}^t = \frac{\mu_1 N j_n(\rho)[\rho h_n^{(1)}(\rho)]' - \mu_1 N h_n^{(1)}\rho)[\rho j_n(\rho)]'}{\mu_2 N^2 j_n(N\rho)[\rho h_n^{(1)}(\rho)]' - \mu_1 h_n^{(1)}(\rho)[N\rho j_n(N\rho)]'}$$

At a resonance, the denominator of either $a_n{}^t$ or $b_n{}^t$ will be zero. Thus, ρ values are found using the above equations that correspond to a resonant combination of angular frequency (ω) and radius (a) for a given sphere material and given surrounding medium. In determining such values of ρ, the following equations are also specifically used:

$$\rho = ak_2 = a\omega\sqrt{\epsilon_2\mu_2} \text{ and}$$

$$\rho_1 = (k_1/k_2)\rho$$

where ρ_1 corresponds to the sphere material. An iterative method is preferably used to find the desired values of ρ at resonance. In calculating ρ utilizing the above equations for purposes of example, it was assumed that $\mu_1=\mu_2=\mu_0=4\pi\times 10^{-7}$ and $\epsilon_2=\epsilon_0=8.85419\times 10^{-12}$.

One major root of ρ which was found has a value of:

Real (ρ)=+66.39752607619131

Imaginary (ρ)=−0.6347867071968998.

These particular values are not shown in FIG. **8**. However, other values of ρ found using the equations set forth herein are shown in FIG. **8**. The peaks in FIG. **8** are the resonances. One of these resonances shown in FIG. **8** is shown in detail in FIG. **9**. These resonance values are shown for purposes of example. Other resonances also exist which have not been determined; thus, not all possible resonance values are shown in FIGS. **8** and **9**.

Calculation of these values also allows the determination of a possible am combination which would have these root values. For ρ, ϵ (epsilon)=ϵ_0 and $\mu=\mu_0$, and

$$\rho = a\omega\sqrt{\epsilon_0\mu_0} = a\omega/c.$$

Expressed in SI units, the speed of light $c=2.99792458\times10^{14}$ m/s. If an a value of 10^{-6} m is assumed for the examples shown herein, then:

$$\omega=\rho c/a\approx1.9919\times10^{16}-i1.9044\times10^{14}\text{ radians/s.}$$

This is an example of the angular frequency required within the impingent EM radiation in order to create a resonant situation. Examples of other resonances were indicated, and these are shown in FIG. **8**. No complex-frequency plane waves exist. Therefore, the calculations were made by considering only the real portion of the above root and setting the imaginary portion equal to zero. However, upon doing this, the iterative calculation procedure becomes insensitive to any root in the vicinity of the root's real portion. In the iterative calculation procedure, initially a range of ρ values is input into the equations. These ρ values are in the neighborhood of the prospective root. A range of ρ values is subsequently studied to find any imaginary ρ i.e., $f\rho$ (a function of ρ), peaks in that range. Next, once a peak has been chosen, the function order n giving the dominant $f\rho$ is determined. This also gives a clue as to whether the peak is due to a magnetic resonance (a_n approaches infinity) or an electrical resonance (b_n approaches infinity). A large number of Newton-Raphson iterations is preferably performed in order to converge upon a root ρ value.

FIGS. **2** and **3** show a second embodiment of the present invention generally designated by the numeral **110**. Embodiment **110** is essentially the same as embodiment **10** except that the antenna is a rf cavity structure **122** which feeds the received beat frequency radiation **124** to a waveguide **126**. Embodiment **110** also preferably includes two spheres **112** and **114** which receive the primary incident electromagnetic radiation **116** and emit the secondary electromagnetic radiation **118** and **120**. As with the spheres **18** and **20** of embodiment **10**, spheres **118** and **120** are preferably composed of a dielectric material. Embodiment **110** also includes converter **128**, capacitor **130**, transformer **132** and rectifier **134** which are essentially identical to the correspondingly numbered elements of embodiment **10**. Therefore, a description of these components of embodiment **110** will not be repeated in order to promote brevity. In addition, the same equations and method of calculation set forth above with regard to embodiment **10** also apply to embodiment. Therefore, their description will not be repeated in order to promote brevity.

FIGS. **4** and **5** show a third embodiment of the present invention generally designated by numeral **210**. Embodiment **210** is essentially identical to the first embodiment **10** except that the embodiment **210** includes a plurality of pairs **215** of receiving means (spheres) **212** and **214** mounted on a substrate **236**. The spheres **212** and **214** are thus in the form of an array **238**. The pairs **215** of the array **238** are preferably positioned proximal to each other in order to maximize the amount of energy extracted from a particular area or space of a given size. Since, as set forth hereinabove, the energy density of the zero point radiation increases as the frequency of the radiation increases, it is desirable that the spheres resonate at as high a bandwidth of frequencies as possible. Because the spheres **212** and **214** must be small in direct proportion to the wavelength of the high frequencies of the incident electromagnetic radiation **216** at which resonance is desirably obtained, the spheres **212** and **214** are preferably microscopic in size. Current lithographic techniques are capable of manufacturing such microscopically small spheres mounted on a suitable substrate thereby providing a suitably miniaturized system **210**. A miniaturized system enhances the energy output capability of the system by enabling it to resonate at higher frequencies at which there are correspondingly higher energy densities. Consequently, utilization of array **238** in the system **210** enhances the maximum amount of electrical energy provided by the system **210**.

Lithographic techniques may be more amenable to manufacturing microscopically small receiving structures **212** and **214** which may be disc shaped, semispherical or have another shape other than as shown in FIGS. **4** and **5**. Consequently, the receiving means **212** and **214** may accordingly have such alternative shapes rather than the spherical shape shown in FIGS. **4** and **5**. In addition, a large number of small spheres may be manufactured by bulk chemical reactions. Packing a volume with such spheres in close proximity could enhance the output of energy.

Embodiment **210** also includes a plurality of antennas **222** positioned preferably between the spheres **212** and **214** which receive the beat frequency radiation **224** produced by the interference between the secondary radiation **218** and **220**. The antennas **222** are shown as loop antennas **222** but may be any other suitable type of antennas as well.

Embodiment **210** has a plurality of electrical conductors **226** which preferably include traces mounted on the substrate **236** which occupies a finite volume. The electrical conductors **226** feed the electrical output from the antennas **222** to a suitable converter **228** which preferably includes tuning capacitor **230**, transformer **232** and rectifier **234**, as with embodiments **10** and **110**. Except as set forth above, the components of embodiment **210** are identical to embodiment **10** so the detailed description of these components will not be repeated in order to promote brevity. In addition, the same equations and method of calculation set forth above for embodiment **10** also apply to embodiment **210**. Therefore, the description of these equations and method of calculation will not be repeated in order to promote brevity.

Accordingly, there has been provided, in accordance with the invention, a system which converts high frequency zero point electromagnetic radiation into electrical energy effectively and efficiently and thus fully satisfies the objectives set forth above. It is to be understood that all terms used herein are descriptive rather than limiting. Although the invention has been specifically described with regard to the specific embodiments set forth herein, many alternative embodiments, modifications and variations will be apparent to those skilled in the art in light of the disclosure set forth herein. Accordingly, it is intended to include all such alternatives, embodiments, modifications and variations that fall within the spirit and scope of the invention as set forth in the claims hereinbelow.

What is claimed is:

1. A system for converting incident electromagnetic radiation energy to electrical energy, comprising:

a first means for receiving incident primary electromagnetic radiation, said means for receiving producing emitted secondary electromagnetic radiation at a first frequency, said first means for receiving having a first volumetric size selected to resonate at a frequency within the frequency spectrum of the incident primary electromagnetic radiation in order to produce the secondary electromagnetic radiation at the first frequency at an enhanced energy density;

a second means for receiving the incident primary electromagnetic radiation, said means for receiving producing emitted secondary electromagnetic radiation at a second frequency, the secondary radiation at the first frequency and the secondary radiation at the second frequency interfering to produce secondary radiation at a lower frequency than that of the incident primary radiation, said second means for receiving having a second volumetric size selected to resonate at a frequency within the frequency spectrum of the incident primary electromagnetic radiation in order to produce the emitted secondary electromagnetic radiation at the second frequency at an enhanced energy density;

an antenna for receiving the emitted secondary electromagnetic radiation at the lower frequency, said antenna providing an electrical output responsive to the secondary electromagnetic radiation received;

a converter electrically connected to said antenna for receiving electrical current output from said antenna and converting the electrical current output to electrical current having a desired voltage and waveform.

2. The system of claim **1** wherein:

said first means for receiving is composed of a dielectric material; and

said second means for receiving is composed of a dielectric material.

3. The system of claim **1** wherein:

said first means for receiving is spherical; and

said second means for receiving is spherical.

4. A system for for converting incident zero point electromagnetic radiation energy to electrical energy, comprising:

a first means for receiving incident primary zero point electromagnetic radiation, said means for receiving producing emitted secondary electromagnetic radiation at a first frequency;

a second means for receiving the incident primary zero point electromagnetic radiation, said means for receiving producing emitted secondary electromagnetic radiation at a second frequency, the secondary radiation at the first frequency and the secondary radiation at the second frequency interfering to produce secondary radiation at a beat frequency which is lower than that of the incident primary radiation;

an antenna for receiving the emitted secondary electromagnetic radiation at the lower frequency, said antenna providing an electrical output responsive to the secondary electromagnetic radiation received;

means for transmitting the emitted secondary electromagnetic radiation at the beat frequency from said antenna, said means for transmitting connected to said antenna;

a converter connected to said means for transmitting for receiving the emitted secondary electromagnetic radiation at the beat frequency from said antenna and converting the same to electrical current having a desired voltage and waveform.

5. The system of claim **4** wherein:

said first means for receiving has a first volumetric spherical size selected to resonate in response to the incident primary electromagnetic radiation in order to produce the secondary electromagnetic radiation at the first frequency at an enhanced energy density; and

said second means for receiving has a second volumetric spherical size selected to resonate in response to the incident primary electromagnetic radiation in order to

produce the emitted secondary electromagnetic radiation at the second frequency at an enhanced energy density, said first and second volumetric sizes selected based on parameters of propagation constant of said first and second means for receiving, propagation constant of medium in which said first and second means for receiving are located and frequency of the incident primary electromagnetic radiation.

6. The system of claim **5** wherein the first and second volumetric sizes are selected by utilizing the formulas:

$$a_n{}^t=\frac{\mu_1 j_n(\rho)[\rho h_n^{(1)}(\rho)]'-\mu_1 h_n^{(1)}\rho)[\rho j_n(\rho)]'}{\mu_1 j_n(N\rho)[\rho h_n^{(1)}(\rho)]'-\mu_2 h_n^{(1)}(\rho)[N\rho j_n(N\rho)]'}$$

$$b_n{}^t=\frac{\mu_1 N j_n(\rho)[\rho h_n^{(1)}(\rho)]'-\mu_1 N h_n^{(1)}\rho)[\rho j_n(\rho)]'}{\mu_2 N^2 j_n(N\rho)[\rho h_n^{(1)}(\rho)]'-\mu_1 h_n^{(1)}(\rho)[N\rho j_n(N\rho)]'}$$

$$\rho=a\omega\sqrt{\epsilon_2\mu_2}$$

wherein at a resonance, the denominator of either equation for $a_n{}^t$ or $b_n{}^t$ will be approximately zero and wherein k_1=propagation constant of the means for receiving, k_2=propagation constant of medium through which the incident electromagnetic radiation propagates, a is the radius of either means for receiving, $N=k_1/k_2$, $\rho=k_2a$, $k_1a=N\rho$, $a_n{}^t$=magnitude of oscillations of the electric field of the nth order, $b_n{}^t$=magnitude of oscillations of the magnetic field of the nth order, ω=angular frequency of the incident electromagnetic radiation, ϵ is the permittivity of the material or medium and μ is the permeability of the material or medium.

7. The system of claim **6** wherein the radius of the first means for receiving is different from the radius of the second means for receiving, difference between the radius of said first means for receiving and the radius of said second means for receiving selected so that the beat frequency resulting from the difference is a frequency which facilitates conversion of the beat frequency electromagnetic radiation to electrical energy.

8. The system of claim **4** wherein:

said first means for receiving is composed of a dielectric material; and

said second means for receiving is composed of a dielectric material.

9. The system of claim **4** wherein:

said first means for receiving is spherical; and

said second means for receiving is spherical.

10. The system of claim **4** wherein said antenna is positioned generally between said first and second means for receiving.

11. The system of claim **4** wherein said antenna is a loop antenna.

12. The system of claim **4** wherein said antenna is a generally concave shell partially enclosing said first and second means for receiving.

13. The system of claim **4** wherein said means for transmitting is a waveguide.

14. A system for for converting incident zero point electromagnetic radiation energy to electrical energy, comprising:

a substrate;

a plurality of pairs of first means for receiving incident primary zero point electromagnetic radiation and second means for receiving incident primary zero point electromagnetic radiation, said plurality of pairs of means for receiving mounted on said substrate, said first means for receiving producing emitted secondary electromagnetic radiation at a first frequency, said second means for receiving the incident primary zero point electromagnetic radiation producing emitted secondary electromagnetic radiation at a second frequency, the secondary radiation at the first frequency and the secondary radiation at the second frequency interfering to produce secondary radiation at a beat frequency which is lower than that of the incident primary radiation, said first means for receiving having a first volumetric size selected to resonate in response to the incident primary electromagnetic radiation in order to produce the secondary electromagnetic radiation at the first frequency at an enhanced energy density, and said second means for receiving having a second volumetric size selected to resonate in response to the incident primary electromagnetic radiation in order to produce the emitted secondary electromagnetic radiation at the second frequency at an enhanced energy density, said first and second volumetric sizes selected based on parameters of propagation constant of said first and second means for receiving, propagation constant of medium in which said first and second means for receiving are located and frequency of the incident primary electromagnetic radiation, said first and second volumetric sizes being different from each other;

a plurality of antennas for receiving the emitted secondary electromagnetic radiation at the lower frequency, said antenna providing an output responsive to the secondary electromagnetic radiation received, said plurality of antennas mounted on said substrate, each of said plurality of antennas receiving the emitted secondary electromagnetic radiation of one of said pairs of first and second means for receiving;

means for transmitting the emitted secondary electromagnetic radiation at the beat frequency from said antenna, said means for transmitting connected to said plurality of antennas;

a converter connected to said means for transmitting for receiving the emitted secondary electromagnetic radiation at the beat frequency from said antenna and converting the same to electrical current having a desired voltage and waveform.

* * * * *

"The machines I made out there did the work", Mr Perrigo said, "but they were bulky and bunglesome and got out of order easily.

"Other things than getting rid of bulk bothered me, too. My early machines were affected by passing air currents. Power would increase when I fanned the machine or when a person walked past it, and decrease when the atmosphere was calm.

"The machine I have now gives a steady flow of current, whether in the basement, 14,000 feet in the air, in a motor car or on a fast moving train. The copper pegs in the wooden block do that."

The Completed Machine ~

The Perrigo consists of only four parts: two lead plates, a wooden block and a coil of copper wire. The coil that Mr Perrigo says will deliver 500 horsepower is 10 inches across, 4 inches high, of solid copper. Fine copper wire is wound smoothly about neat rows of copper stays, hardly larger than a toothpick, but longer, as long as the coil is high. The wire is connected in many devious ways. In those connections is the secret of the mysterious power.

The lead plates for this size Perrigo are a foot square. On each plate are 100 spring coils of copper wire, spaced in rows, an inch apart. The plates appear identical, but are different in the way the connections of hair-like copper wire run from one coil to another.

The wooden block is a foot square, also, and an inch thick. One hundred copper plugs run through the block, spaced just as the coils on the lead plates are spaced. Each plug is a bundle of copper stays, making a contact.

That is all there is to the Perrigo, so far as anyone can see. Complete, the 500 horsepower size weighs 87 and one-half pounds. There are no moving parts.

A Perrigo to operate an ordinary size motor car need be no larger than a 1-pound coffee can", Mr Perrigo said. "A size to provide all the current needed to heat and light a 5-room house will go into the can. The different sizes can be made to furnish any desired voltage, and either direct or alternating current, by a slight change in the wiring."

The Inventor Explains It ~

Trying to avoid technical terms, difficult business for an engineer, Mr Perrigo explained his invention this way:

"The device is a generator as truly as the power-driven rotary generator in any power plant. Those generators don't actually 'make' electricity. They condense it from the air. So does the Perrigo. But it does it through the system of wiring, rather than revolutions through a magnetic field. I get my starting point from the air by breaking up the ether waves. The coils on the lead plates do that. I know why they do. It's the way they are connected, one from the other. That's my secret.

"They do break up the ether waves, gathering electricity and conducting it into the big coil underneath. That's the generator. Its size and the way it is wired determines the voltage, the horsepower. Outgoing wires from this coil take the 'juice' where you want it and it is there when you want it."

Mr Perrigo has great plans for his electric 'chore boy' and great faith in it.

"It will replace every other source of power, light and heat", he predicts. "It means the doom of the steam engine, the end to coal mining, to the cutting of timber for fuel. It means petroleum will be used only for lubrication. It means smokeless, sootless cities. It means chimneys will disappear from housetops. It means cheap power for the farmer, the reclamation of much country that cannot be irrigated now because power is not available."

What Other Persons Think ~

Mr Perrigo is able to impart his faith to others, too. The enthusiasm of persons who have seen the device work is second only to that of Mr Perrigo himself.

S.W. Fries, an electrical engineer, and district sales manager here for the Economy Fuse and Manufacturing Company, saw the Perrigo first about three months ago.

"When I heard about the machine through Dr McDowell, I told him it couldn't be done", Mr Fries said the other day. "I've been converted. I don't know how it works, but it does. Its possibilities are too big to grasp. Its use will mean a new age in industry. I believe Mr Perrigo will be the most widely known inventor in the world as soon as his device comes into general use, and he will become one of the world's most wealthy men, just from returns which already seem assured."

"Mr Perrigo gets enough electricity from somewhere to knock him unconscious", Dr O.W. Butler (3700 Benton Blvd) pointed out. "I've been called to his house many times in the last four years to revive him, and once I carried him out of his basement. He has worked at his experiments as long as four days and nights without sleep --- worked until his health is broken and his constitution is a bundle of jagged nerves."

"How are you going to manufacture your machine and get it on the market?" Mr Perrigo was asked when he asserted there was no stock for sale and he was seeking no financial aid.

"Responsible backers are furnishing all the money I need for experiments, models for the patent office and other work I'm doing now", he answered.

"As soon as one final amendment to my patent application is approved I'll be ready to permit motor car manufacturers to make the Perrigo in their own plants, charging them a small royalty on each machine. They will be eager for it when they see what it is. This will provide funds enough in a short time for my associates and myself to being to manufacture the machines for home use.

"We don't expect to sell the Perrigo. We will lease them on the same plan the telephone companies use for their machines, charging a monthly rent, probably about $3 for a 5-room cottage size. That's cheap enough, isn't it, for all heat and light and power?

"I've always said I never would sell out to any big corporation. My invention is for the benefit of the poor amn. Even on that basis I'll get more money out of it than I can ever use."

The Householder's Point of View ~

Mr Perrigo explained that it will not be necessary for the householder who desires the Perrigo installed to buy an expensive electric furnace, electric range or any special equipment.

"A gas range can be wired through the pipes which now carry gas", he said. "It will be necessary only to replace the gas burners with electric heating plates and install snap switches where the gas valves are now. A furnace can be fitted in the same way, by removing the grate. No change will be necessary with the lights. The new machine will be installed where the meter is now. That's all."

"When one man has a machine, won't it be possible for his neighbors to come in, see how it works, and manufacture their own?" Mr Perrigo was asked.

"No", he answered, "Each one will be sealed, just as the electric meter is sealed. To break the seal will put it out of order and the subscriber will have to call for a 'trouble' man. Anyway, if a man would take one apart he couldn't put it back together again without my drawings and blueprints. That's my secret and I'll keep it."

A Demonstration ~

When a visitor expressed a desire to see a machine actually produce light or power or heat, Mr Perrigo acquiesced. He went to the basement and returned almost immediately, bearing a boxlike affair, mounted on a little platform. A small electric motor, light sockets and switches were on the platform. The top of the box was glass. Through it Mr Perrigo pointed out parts of the machine inside.

Mr Perrigo fastened the loose ends of two wires that extended from the box to the connection posts of the motor and pushed a switch button. The motor started at once. The inventor said he had not changed the machine or even opened the box, which was closed with screws since he made it five years ago.

Kansas City Star (March 27, 1922) ~

"Can't Use His Invention"

Patent Laws Prevent Demonstration, H.E. Perrigo Tells Inquirer ~

Miami, OK --- To *The Star*: "In your paper of January 15 there appeared a wonderful story relating to the invention of Harry E. Perrigo, an electrical engineer of Kansas City, of a device to generate electrical energy.

"One got the impression from reading the story that Mr Perrigo's device was a demonstrated success, that it had been patented, and that he was practically ready to permit its manufacture as a source of power.

"It was a whale of a story and interested me, for it seemed if it were true that his device was destined to have as far-reaching influence on the human race in the future as the grain binder and the gas engine had had in the past.

"I wrote to various publications devoted to mechanics and electricity, seeking further information. None knew anything about it. One said no such device had been patented, another that a vast amount of research and experimentation had been done in an effort to develop such a device and that the only result had been failure.

"I would like to know if Mr Perrigo's device has been patented and if he is prepared at the present time to demonstrate it in a convincing way."

The story of Mr Perrigo's invention was printed in the Star as the record of an ambitious and interesting enterprise. The apparent success of the device was vouched for by persons who had seen the machine in operation and were convinced it produced the results Mr Perrigo claimed for it.

Questioned recently as to new developments, Mr Perrigo said there could be no new mechanical developments.

"The machine has been developed to a state as near perfect as I can make it, for two years", he said. "I am waiting for the patent office to take final action. My applications have been approved and investigation has shown no conflicting patents on record. The rest is simply a matter of routine work in the patent office. As soon as the final patents are granted I'll be ready to manufacture the Perrigo.

"I can't give a demonstration without going to considerable trouble to set the machine up. Of can't keep it set up, nor even use it in my home, because technically that would amount to 'commercial use' and would interfere with the granting of patent rights."

---

(1) Harry E. Perrigo, the inventor of a "free energy" device which he believes will revolutionize all industry ~

HARRY E PERRIGO, THE INVENTOR OF A "FREE ENERGY" DEVICE WHICH HE BELIEVES WILL REVOLUTIONIZE ALL INDUSTRY

(2) Here, Perrigo says, is his secret: the 100 little spring coils of copper wire break up the aether waves and conduct the electricity they gather into the big generating coil. There are two of these plates, identical in appearance but different in wiring.

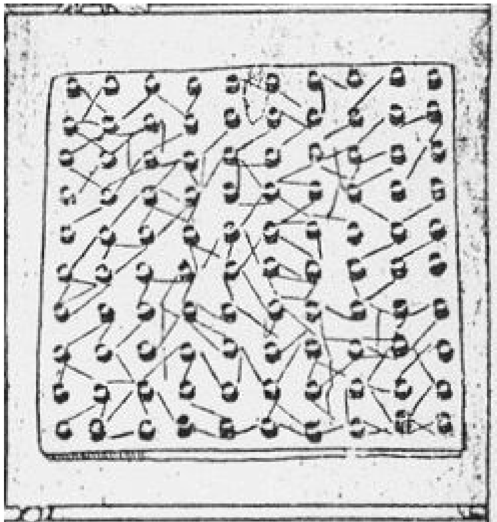

HERE, MR. PERRIGO SAYS, IS HIS SECRET: THE ONE HUNDRED LITTLE SPRING COILS OF COPPER WIRE BREAK UP THE ETHER WAVES AND CONDUCT THE ELECTRICITY THEY GATHER INTO THE BIG GENERATING COIL. THERE ARE TWO OF THESE PLATES, IDENTICAL IN APPEARANCE, BUT DIFFERENT IN WIRING

(3) The 100 little bundles of copper stays form the plugs driven through the wooden block are soldered over the ends, making them appear smooth.

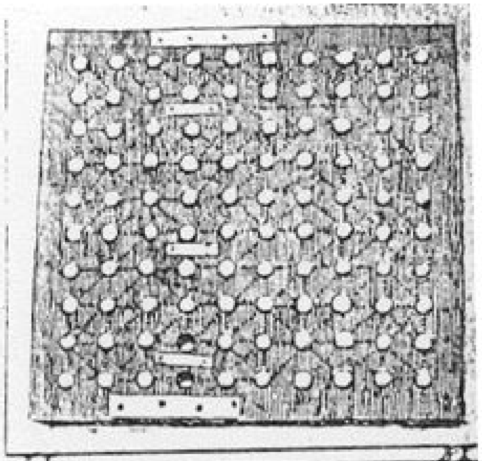

THE ONE HUNDRED LITTLE BUNDLES OF COPPER STAYS, WHICH FORM THE PLUGS DRIVEN THROUGH THE WOODEN BLOCK ARE SOLDERED OVER THE ENDS, MAKING THEM APPEAR SMOOTH

(4) Mr. Perrigo's first machine, as he remembers it, was a crude device made from two embroidery hoops, a table leaf, bread box and other materials picked up around the house.

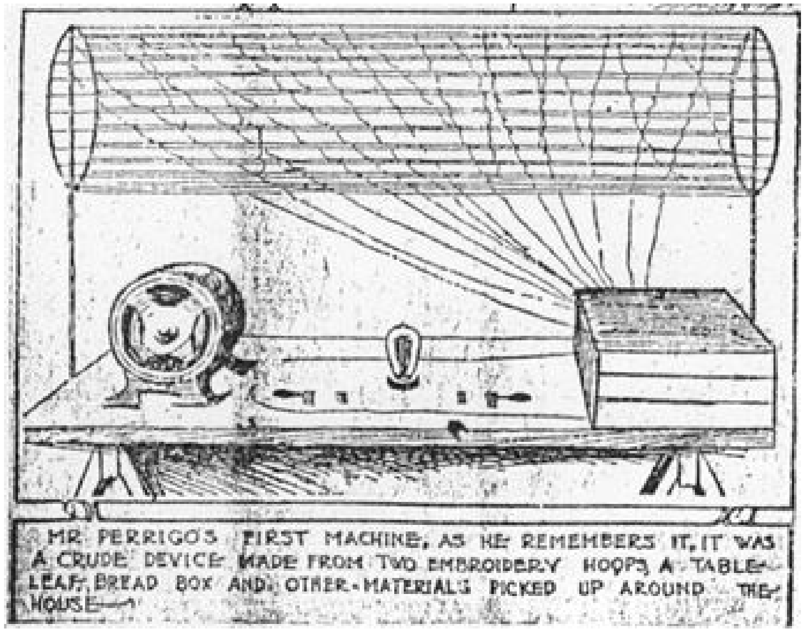
MR. PERRIGO'S FIRST MACHINE, AS HE REMEMBERS IT, IT WAS A CRUDE DEVICE MADE FROM TWO EMBROIDERY HOOPS, A TABLE-LEAF, BREAD BOX AND OTHER-MATERIALS PICKED UP AROUND THE HOUSE

(5) The top of the big coil, showing the ends of the copper stays and the windings of wire. The space separating the two parts of the coil are the "field gap", Mr. Perrigo says.

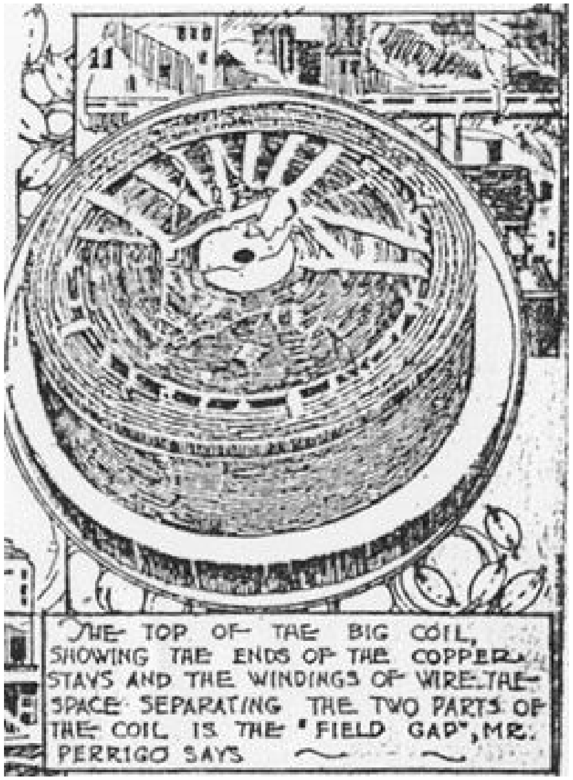
THE TOP OF THE BIG COIL,
SHOWING THE ENDS OF THE COPPER
STAYS AND THE WINDINGS OF WIRE. THE
SPACE SEPARATING THE TWO PARTS OF
THE COIL IS THE "FIELD GAP", MR.
PERRIGO SAYS

(6) At left, a closeup of one of the 100 plugs in the wooden block.

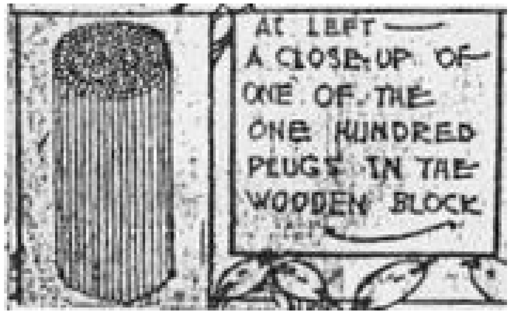
AT LEFT —
A CLOSE-UP OF
ONE OF THE
ONE HUNDRED
PLUGS IN THE
WOODEN BLOCK

(7) The complete Perrigo, without its wooden case. The drawing was made from a machine Mr. Perrigo made by hand which, he says, has developed more than 500 horsepower. The commercial machine will be an exact copy of this.

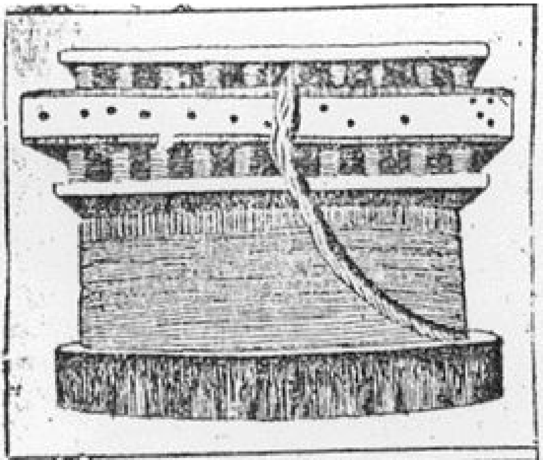

THE COMPLETE PERRIGO, WITHOUT ITS WOODEN CASE. THE DRAWING WAS MADE FROM A MACHINE MR. PERRIGO MADE BY HAND, WHICH, HE SAYS HAS DEVELOPED MORE THAN FIVE HUNDRED HORSEPOWER. THE COMMERCIAL MACHINE WILL BE AN EXACT COPY OF THIS

(8) This is the machine the writer saw in operation. It was built by Mr. Perrigo 5 years ago for demonstration in Washington, when the inventor successfully opposed a war measure which would have prevented for a period of 17 years the granting of his patents on any "free energy" device. The operation of this machine, on the speaker's table in the lower House, is recorded in the ***Congressional Record*, December 15, 1917, pages 369-383.** [Actually, pp. 357-372]

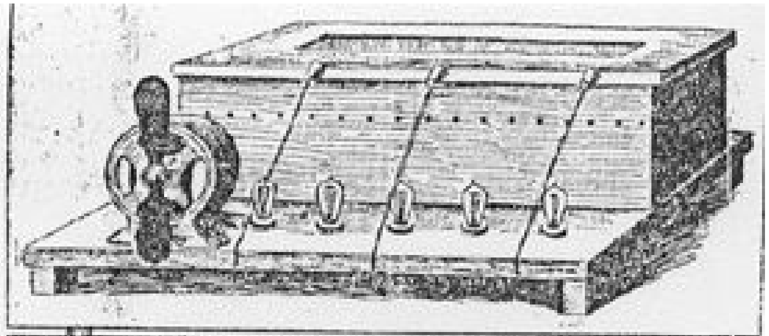

THIS IS THE MACHINE THE WRITER SAW IN OPERATION. IT WAS BUILT BY H.A. PERRIGO FIVE YEARS AGO, FOR DEMONSTRATION IN WASHINGTON, WHEN THE INVENTOR SUCCESSFULLY OPPOSED A WAR MEASURE WHICH WOULD HAVE PREVENTED FOR A PERIOD OF SEVENTEEN YEARS, THE GRANTING OF HIS PATENTS ON ANY "FREE ENERGY" DEVICE. THE OPERATION OF THIS MACHINE ON THE SPEAKERS' TABLE IN THE LOWER HOUSE IS RECORDED IN THE CONGRESSIONAL RECORD DECEMBER 15TH, 1917, PAGES 369 – 83

Kansas City Journal (August 7, 1927), page 1 ~

"Power Drawn From Air Propels Auto Over K.C. Streets"

Inventor Claims Tiny Motor Will Drive Plane Around World, Doing Away With Transportation of Fuel

A motor car was driven 40 miles an hour in Kansas City yesterday on power drawn from the air.

A device making this feat possible was demontrated after years of rebuffs and failures by its inventors, Harry Perrigo, 1116 Bennington Avenue, in the presence of Col, Paul Henderson, Chicago, vice president of the National Air Transport Inc., air mail contractors.

"It is the greatest invention since the stone age", was Col. Henderson's comment.

Lou E. Holland, president of the Chamber of Commerce, who saw the demonstration with Col. Henderson, was enthusiastic over the invention and said he believed it would have a great effect on public utilities if it can ever be brought into general use.

Coils Gather Power ~

The device consists of a plate 14 inches square which, by a multiple arrangement of connected copper coils, it was explained, attracts electric current from the air.

The electricity thus trapped passes through a generator and to a regular direct current motor, which was connected to the transmission shaft of the car.

Mr Perrigo has been working on the device more than 10 years. Three years ago he suffered a nervous breakdown and has been back at his workshop only two months.

While photographers focused their cameras, Mr Perrigo and his wife entered his roadster in which the device was installed. He turned a switch and the motor hummed.

The inventor sat silent, listening. Then he threw in the clutch gently. The car moved away with only the sound of the cogs, whining softly as they meshed. He threw it into high and sped away.

At the corner of 10th Street and Bennington Avenue, near his home, he stopped the machine while movie cameramen ground their machines. Then he backed it and turned around, stopping in the middle of the street.

Later Mr Perrigo took the car out for a spin. It breezed along at 40 miles an hour.

"100 Just As Easy" ~

"I could make it go 100 miles an hour just as easily", the inventor remarked, "if it were not for the danger of hurting someone."

The total weight of the motor, generator and controls is about 86 pounds, according to the inventor. In the rear of the car are reduction coils, because the device is five times too powerful for the work of pulling the car, Mr Perrigo said.

The proper motor for a Reo car will not weigh more than 30 pounds", he said. "The motor to run a Chevrolet will weigh not more than 10 pounds. An airplane motor would weigh around 50 pounds, and with that motor the air traveler could go around the world without waiting."

Col. Henderson walked around the machine, inspecting it.

"What it would mean to my airplanes", he said, "--- no weight for motors, no space for gas, no motor trouble."

Like other spectators, he had difficulty in believing what he saw.

The work of bringing the invention to its present state of development has been the story of a lone man working with the unknown, usually meeting with discouragement.

Congress Wouldn't Believe ~

Five years ago, the inventor took one of his electric motors to Washington, set the boxlike contrivance on the speaker's stand in Congress and ran a motor and five lights with it. They called it an infernal machine.

Even now, when he is seeking patents for his invention, the patent office refuses to believe the principles he advances are possible.

"The machine has been taken up in an airplane 10,000 feet and it operated the same as if it were on the ground", Mr Perrigo said. "It has been tried out in hot places and cold places."

He pictures it as power for every purpose --- the farm, the factory, the office, doing away with coal, gas, cost of water power and the cumbersome weight of engines.

The models of the machine now are being kept in bank vaults.

Mr Holland is convinced, after two weeks investigation, that the invention is genuine.

Kansas City Journal (August 8, 1927), p. 3 ~

"Perrigo Dreams Of Aiding Humanity With Ether Wave Machine"

Inventor Declares Generator, When Patented, Shall Not Be "Hogged" or "Shelved", But World Will Reap Benefits Of Cheap Power.

Homes lighted and meals cooked for a cost of $5 per month; motor cars operating with a quart of lubricating oil about once every six months and a little grease in the rear axles and yet damaging no existing corporations by the working of these revolutions.

These are the dreams of H.E. Perrigo, 1116 Bennington Avenue, inventor of the Perrigo ether wave generator.

Lou E. Holland, president of the Chamber of Commerce and Col. Paul Henderson, general manager of the National Air Transport, Inc., are convinced his invention is practical.

Mr Holland, Mr Henderson and newspaper men Saturday witnessed a demonstration in which Mr Perrigo drove a motor car on electric power generated from the air.

"It is too early yet to talk to manufacturing the invention", Mr Holland said yesterday. "The patent rights must be perfected in both the United States and foreign countries. Also, it must be made clear that nothing will be done to demoralize present power manufacturing."

Mr Perrigo's invention is the result of more than 12 years labor on the part of the inventor.

"Twelve years ago I conceived the idea that the generation of power from ether waves was possible, while I was employed ion the power plant at Pee Dee, NC, where some wires not connected with anything that was 'live' seemed to be generating power", Mr Perrigo said yesterday.

"After more than a year of experimenting I finally produced light in a bulb about the size of those used in flash lights. Now in my laboratory I light three 300 watt globes from a wire no larger than the thread which holds the button on a man's overcoat.

"Three times in my experiments I was knocked unconscious because I did not know how much power it would generate.

"The invention has been tested under all conditions. It has been strapped to the running board of a locomotive running between Kansas City and Chicago, it has been tested in an airplane at a height of 12,000 feet. It will generate power anywhere that air circulates.

"With the invention every home will have its own power plant and all the electricity needed for any purpose can be provided at a roughly estimated cost of $5 per month.

"The Kansas City Public Service company, for instance, could afford to install these plants because they would eliminate the enormous overhead which is the principal expense of such companies."

Both Mr Perrigo and Mr Holland made it plain that the invention will not be permitted to upset the electrical world. No one will be given exclusive rights of its use, they said.

"Any person or any manufacturer can use the invention who will pay the royalty fees", Mr Perrigo explained. "No one will be permitted to 'hog' the invention and no one can purchase it and shelve it.

"I have no doubt that it will bring me more money than Mrs Perrigo and I will need to supply our simple wants and I want humanity to benefit.

"I want to see rural sections which have remained barren because it cost too much to pump the water to irrigate them, spring into bloom. I want to see the smallest farm house in the most isolated places with its electric lights and stoves; in other words, I want humanity to benefit."

"How about the gasoline industry if motor cars are to be run without fuel?" he was asked.

"The supply of petroleum is limited", he answered. "There are 30 other purposes to which petroleum products can be turned, so the oil producers will not suffer."

"Can you explain your invention so that the non-technically trained can understand it?"

"That would be difficult, especially without divulging information that is now in the patent office; besides, the generation of power always is difficult to explain.

"Electric generators have been used in power plants for a great many years. We all know what you have to do to make a generator and we all know that electricity is its product but we don't know very much about what really is going on inside that generator."

While the patents for his device still are pending, Mr Perrigo is keeping his working models in the vault of a downtown bank to prevent the possibility of their being copied.

Kansas City Star (August 7, 1927) ~

"Electricity Power from Air?"

An invention that will revolutionize radically all power producing and power using machinery of the entire world, if in practical use it equals the miracle of its demonstration, was given a showing here yesterday to newspaper men and Lou E. Holland, president of the Chamber of Commerce.

It is a device to collect electrical energy from the ether and convert it into a powerful current of electricity of a type dissimilar in many ways from the direct or alternating current now known to electricians.

So far as could be determined even by the most skeptical of those witnessing the demonstration yesterday, the device is exactly what Harry E. Perrigo, the inventor, asserts it to be -- a method for collecting natural electricity from the ether, in unlimited quantity and without cost.

More Power Than Needed ~

One feature of the demonstration was the operation of a motorcar by power from a small model of the invention. The engine had been removed from the car and an electric motor substituted. The device supplied power in such quantity that it had to be reduced and yet propelled the car with speed and ease.

A detailed examination of the car showed the absence of any possible form of power except the inventor’s small device and it is of such an open type that one may see clear through the mass of wires and coils.

Other demonstrations were given with other models of the device, with the machine and electric lights and motors held in the hands of spectators, yet the device produced current to do any electrical task assigned to it.

Col. Paul E. Henderson, general manager of the National Air Transport, Inc., was there with Mr Holland, being his guest for the day. He took an active part in the testing and enthusiastically was declaiming the invention as revolutionary.

Cheap As Similar Motor ~

To operate a motor car would require a model weighing about 20 pounds, Mr Perrigo computes, but he has given no thought the probable cost of building it. However, he estimates the cost at no more than an electric motor of the same size.

A device of the size of a coffee can would light and heat an average home, he declares, cutting off forever all fuel and lighting bills.

And one can go on and dream of an electrified world with free power for all industries and operations, increased yields of foods from dry areas that cold be irrigated with this free energy, the passing of wood and coal and oil as fuels.

At first thought, not one of the persons seeing yesterday's demonstration could give much credence to the inventor's declaration that his invention was a way to obtain unlimited electrical power from the ether without any cost.

There was not a doubter left as to the success of the demonstration, but the witnesses could not in any sense qualify as electrical engineers.

Electricity Always In Ether ~

As near as a layman can understand, Mr Perrigo's theory is the revolution of the earth sets up a form of electric currents that are forever present in the ether. His theory is to capture those electrical impulses in very much to same way that a radio antenna picks up the programs broadcast from WDAF. Instead of a machine to turn the radio impulses into music, Mr Perrigo has a machine to turn the ether's electrical store into controlled power. He declares it is really no more mysterious than the fact that an electric dynamo picks electricity out of the air, although the dynamo must have a power to revolve it while his device sits perfectly still and seemingly produces many fold more electricity than a dynamo of the same bulk.

Demonstrating the different nature of this electricity, Mr Perrigo showed how high voltage could be transmitted over hair-size wires and light a series of electric lamps although a sufficient power of the well known electricity to light those lamps would have melted the small wires immediately.

Clearer Light Than Usual ~

And it imparted an unusual glow to the electric lamps, giving them a clear brilliancy with none of the effervescent haze that surrounds the wires in a lamp when lighted with ordinary electricity.

Mr Perrigo has spent years on his invention. Years ago he conceived the idea that there was an unlimited source of electricity in the air that could be harnessed with a collector. He has not reached the present measure of success without a row of hard knocks along the way.

In whatever neighborhood he has lived since moving here 15 years ago, it has been common knowledge that Mr Perrigo was the frequent victim of electric shocks that often came near proving fatal. He has been revived by pulmotors time and again.

The first model of his device was constructed with makeshift material. A leaf from Mrs Perrigo's dining room table was the basis, with the bread box used, sheets torn up and the strips shellacked to be used for insulation. But as junky as it was, it operated a small motor, and that original model is still preserved and used a s part of his demonstration.

Sickness Delayed His Work ~

In 1922 he got the device to such a point of perfection that he went before Congress and defeated a war-time measure that proposed to give 17-year rights to another person for a blanket patent on all free energy devices.

Then three years of sickness came, and the invention progressed slowly. Mrs Perrigo is not electrically inclined, but she has carried forward the experiments under his direction and has always been his assistant.

Mr Perrigo is not offering stock for sale and says he has made all necessary arrangements for financing the device when it is ready for manufacture.

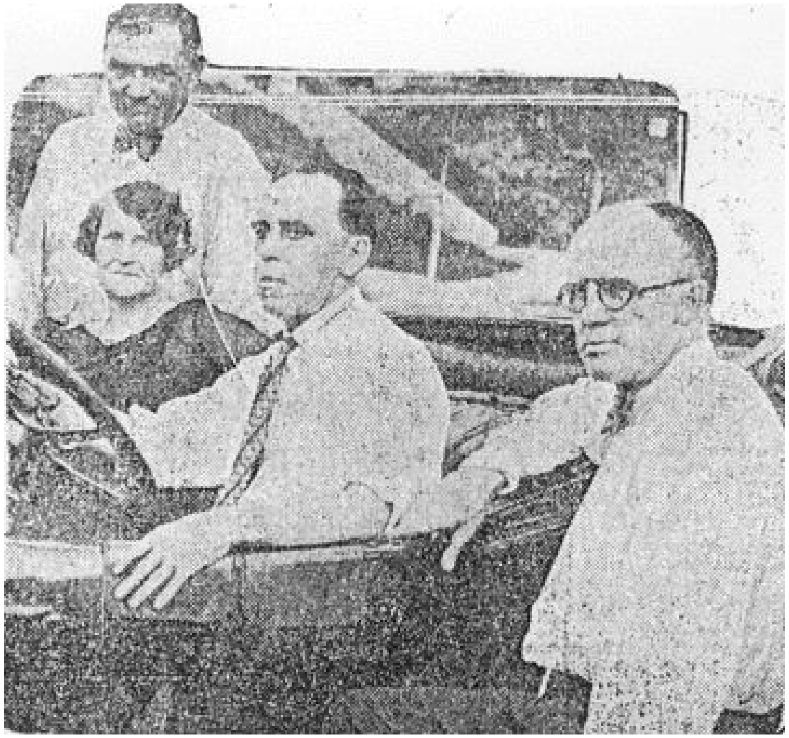

Photo Caption: In a semi-public demonstration here yesterday of an invention said by its inventor to be able to collect electricity from the ether, in unlimited quantity and without cost, a motor car was propelled with perfect success, then an examination of the car allowed to show the absence of engine, storage battery or other usual form of power. The upper photograph shows under the motorcar hood, engine missing and an electric motor in its place. The equipment above the motor is a series of resistance coils, the collection device used being too powerful for the motor. The middle photograph shows the collector placed on the floor by the driver, a makeshift arrangement for the test. Below is Mr Perrigo, with his wife, seated in the test car. Lou E. Holland, president of the Chamber of Commerce, is standing on the near side of the car. Col. Paul E. Henderson, general manager of the National Air Transport, Inc., stands on the other side. They were among those attending the demonstration.

Kansas City Times (August 8, 1927) ~

"Electricity from the Air"

H.E. Perrigo, a nervous, red-haired little electrician, stood last nigh among a confused display of strange devices in is basement workshop at 1116 Bennington Avenue. He was still a bit haggard from an illness of three years, but is energy, dynamic as that which his devices "pick out of the air", was unabated. Mr. Perrigo had just returned from a conference with Lou E. Holland, president of the Chamber of Commerce.

"There are no batteries in this room", he said. "There are no light and power wires. There is nothing but these unconnected accumulators you see before you."

On an old kitchen chair stood an object about one and one-half feet square, several inches thick. It appeared to consist of two parallel metal squares, separated and held together by numerous pegs woven around which was a maze of thin copper wires. It stood on edge, resting against the back of a chair. On the seat of the chair was a round metallic object, resembling, at a glance, a huge spring, a foot and a half in diameter, from an alarm clock. In a mechanical sense it was in no way similar. But it might as well have been an alarm clock spring, for all Mr. Perrigo would tell of its construction. The first object was the collector, the second the generator.

No Moving Parts, No Wheels ~

Neither had any moving parts. They have no wheels. They are immobile, simply an arrangement of wires. On the to edge of the flat box was a switch. Thence ran two wires, connected with forty-five 100-watt standard electric light bulbs.

Mr. Perrigo pulled the switch. There was no arc as the contact was made. The 45 bulbs flashed brightly and burned with a steady white glow. There was never a flicker. He turned 44 lights off, leaving one. It did not flicker as the 44 lights went off and on.

"This little device, the Perrigo Electric Accumulator", said Mr. Perrigo, will light 8,000 bulbs as easily as it lights one. I can build one of any size, to produce the results that any amount of dynamo electricity will produce."

He picked up a little narrow box, in the top of which was a pane of glass, through which one might see more pegs, more intricate wiring. He connected a small electric fan to the poles, and the fan whizzed.

Tests Everywhere But Under The Sea ~

"There are skeptics yet", he said. "Some think I am picking up leaking electricity. This free energy device has been tested at an altitude of 12,000 feet, on the sea, on deserts, everywhere except in a submarine. Everywhere it runs smoothly, without fluctuation".

In his Reo roadster he had a large motor mounted, the one he demonstrated Saturday to Lou E. Holland, president of the Chamber of Commerce, and Col. Paul E. Henderson, general manager of the National Air Transport, Inc. Mr. Perrigo crawled into the seat, threw a switch, and a big all-speed motor, standing in the place of the gasoline engine, roared.

"Rheostat control", said Mr. Perrigo. "It will run at all speeds. I do not need clutches".

There were two "mystery boxes" in the car, one of which picked the electricity out of the ether, the inventor explained, another which regulated its intensity.

"It will be on the market in the near future", Mr. Perrigo said. "I cannot say more at this time. It will be manufactured by a Kansas City company.

"We do not intend to harm the men who produce electrical current by other means. Perhaps they will manufacture and distribute the little units. The can do it gradually, so there will not be a sudden, destructive revolution. That is bad economy. There is to be no monopoly. But

<

Maximize client uptime. Level up your business.
Try SolarWinds® Backup LocalSpeedVault for the fastest possible data recovery.
solarwinds msp
START YOUR FREE TRIAL!

Maximize client uptime. Level up your business.
Try SolarWinds® Backup LocalSpeedVault for the fastest possible data recovery.
solarwinds msp
START YOUR FREE TRIAL!

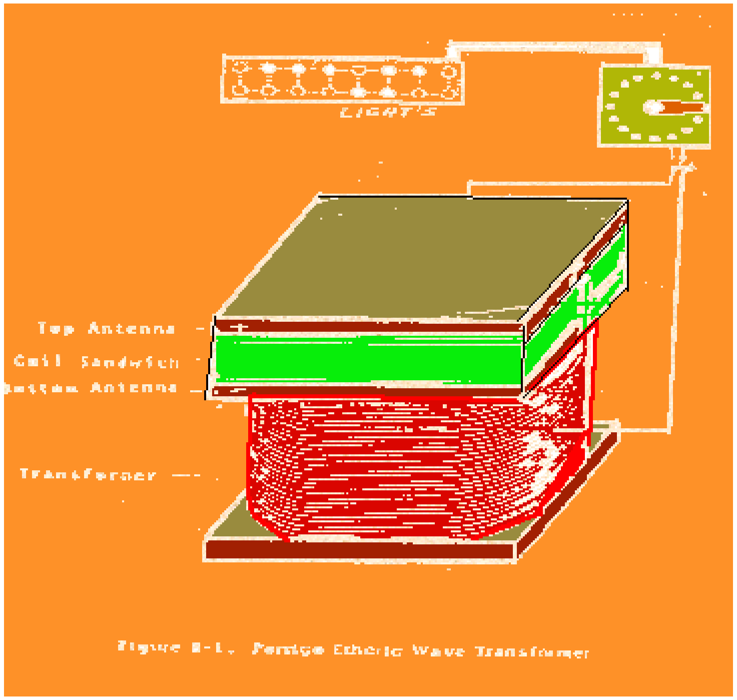

Figure 8-1. Perrigo Etheric Wave Transformer

METHOD AND APPARRATUS FOR ACCUMULATING ELECTRIC ENERGY AND TRANSFORMING ETHER ELECTRIC ENERGY

Drawings by Geoff Egel ,Content Author unknown

In 1926. Harry E. Perrico claimed to have discovered a method to tap the energy from atmospherics. He even had a car which he converted to run on electricity. that was generated in his ether wave accumulator".

He applied for a patent December 31 1925 with serial number 78,715 being assigned. Perrigo's patent application Is considered a "classic" In patent law and Is listed under the classification of Perpetual Motion Machines and other Impossible Inventions.

Despite the Patent Office's attitude towards Perrico's invention, there were a number of reputable people who claimed to have witnessed his device In action producing useable electric power.

The existence of electromagnetic radiation, the modem term, or electric waves In the ether as it used to be called was known by Hertz, a research scientist who discovered the photoelectric effect

In 1887. Atmosphierics, an electrical disturbance in the atmosphere, were known to produce noises In the early radio telegraph stations, some being strong enough to drown out the received signal.

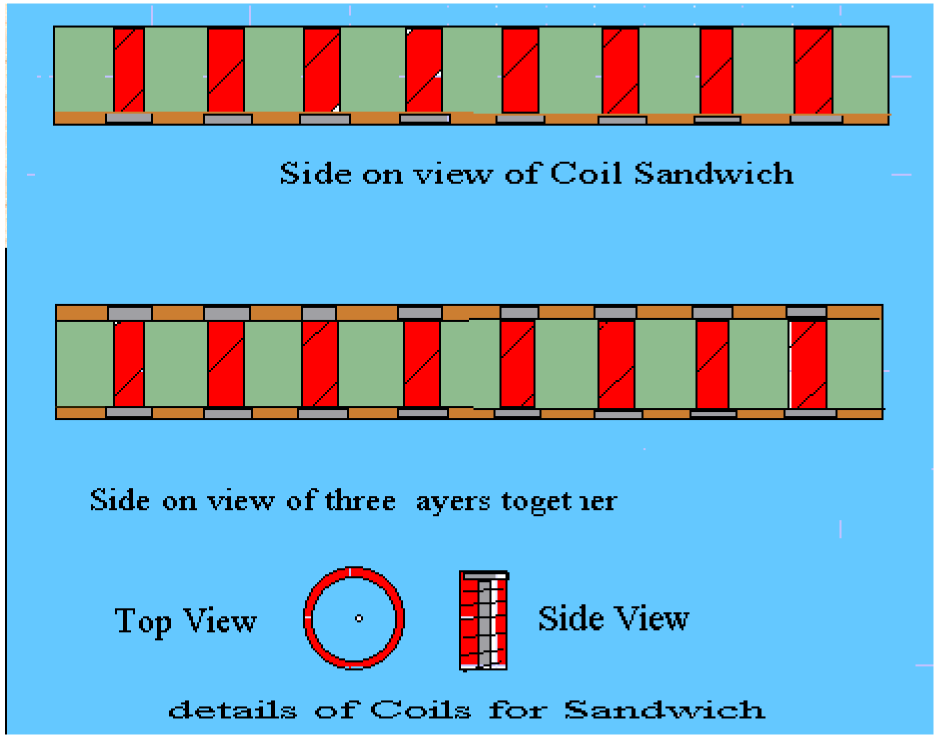
Side on view of Coil Sandwich
Side on view of three ayers toget ıer
Top View
Side View
details of Coils for Sandwich

> Perrigo deduced that here was a possible source of electrical power.

All that was needed was a method of transforming the existing radiation Into useable energy. Hie claimed to have developed a mechanism to intercept and collect from the general ether field electric wave energy and to transform it into useable electromotive force.

The basic method he used was an antenna arrangement which collected and resisted the Incoming energy raising It to a high enough current level where it could be run through a special electrical transformer to further intensify the available power.

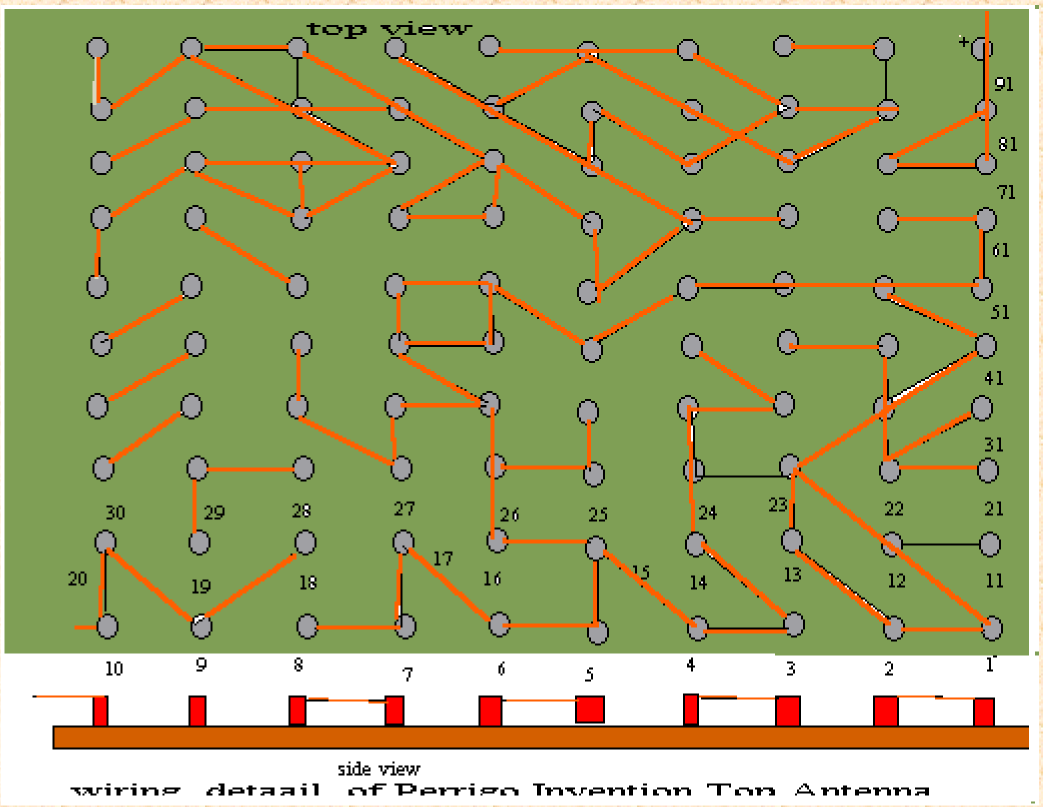
top view
+
91
81
71
61
51
41
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
side view
wiring detaail of Perrigo Invention Top Antenna

Perrigo's antenna was derived from his experiments with various wire shapes, sizes and arrangements. One of his more successful attempts was to partially pound 100 roofing nails into a board a 10 x 10 array and wrap very fine wire around each nail,making it a small electromagnet.

Then by trial and error approach he connected the ends of the electromagnets to other nails in such a way that there was a maximum voltage between the wire and the nail.. His patent application mirrored this electrical connection scheme in a more refined electromechanical approach.

Two accumulator plates were made with 100 round protruding knobs in a square 10 X 10 array

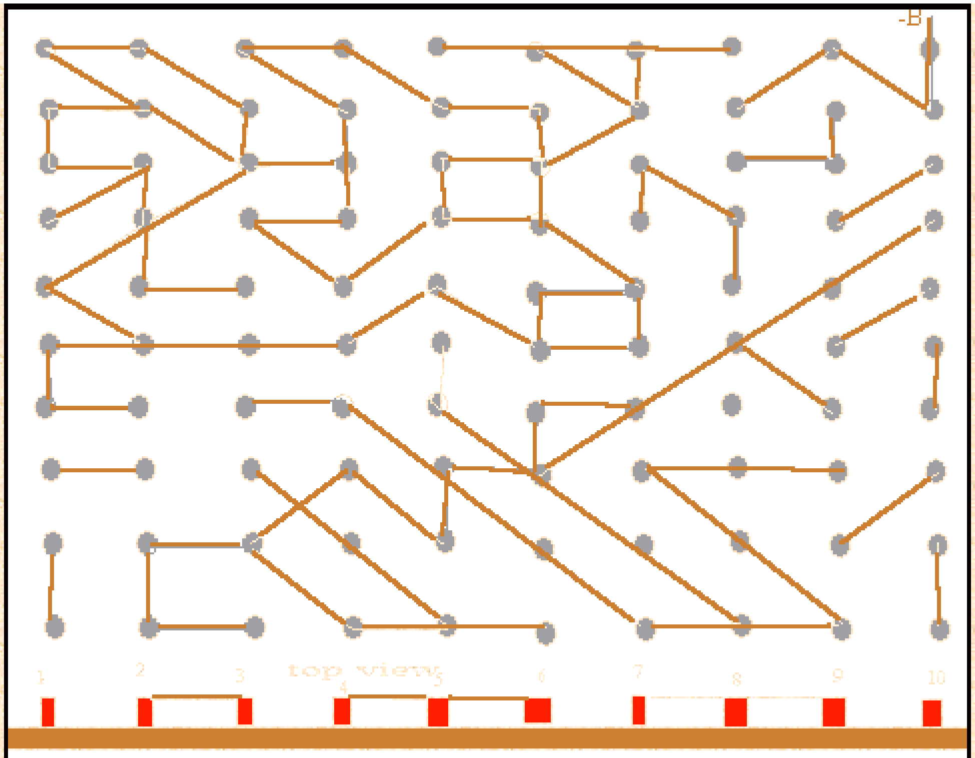

igure 6-3 wiring detail of Perrigo Invention Bottom Antenna

The accumulator plates were then sandwiched together with an insulator material between them. The insulator had 100 holes matching the protrusions on the plates. Placed in each hole was a special coil wrapped around a bundle of wire . Once the accumulator plates were sandwiched ,a measureable electrical voltage existed between points A and B on plates one and two respectively.

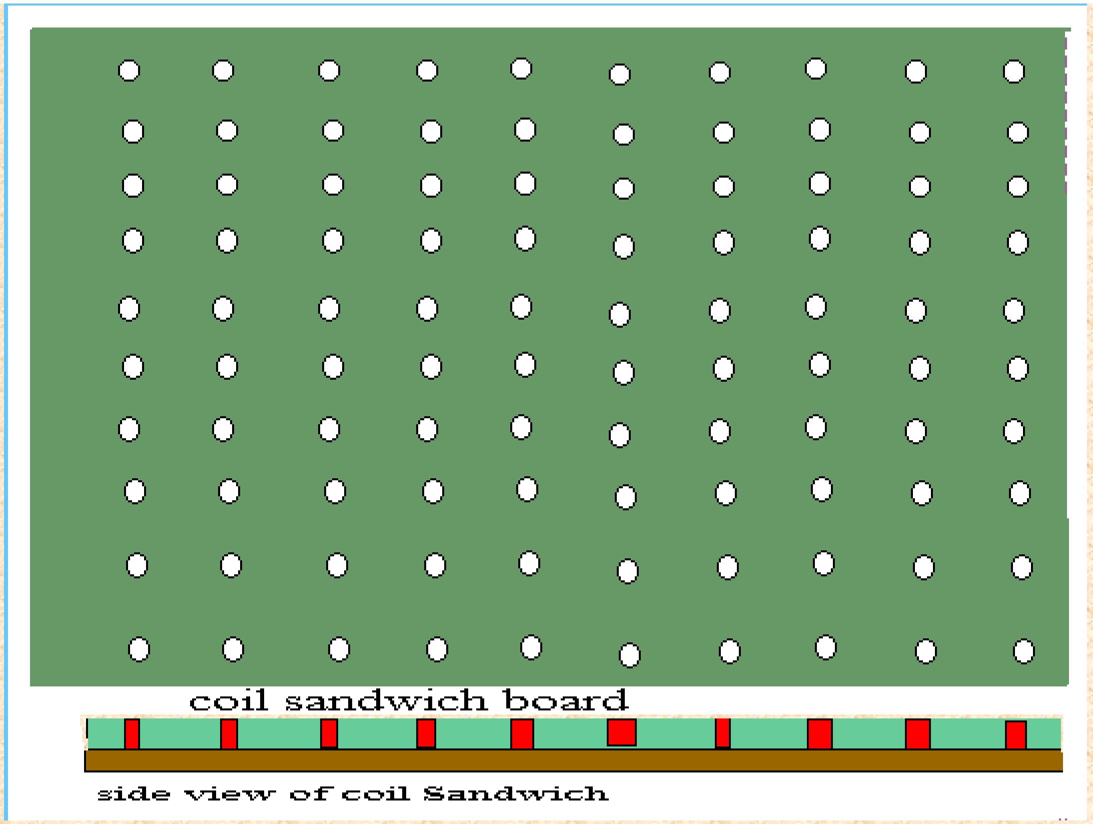

A very complicated transformer was attached to these two points.

The plates were set on top of the transformer and Perrigo claimed this arrangement enhanced the energy accumulation process.

I have no idea of what materials the plate or protrusions were made The patent drawing would lead one to believe they are the same material.

It could be a metal or a nonconductor. such as wood or a combimation. The protrusions were connected by the same wiring scheme

Previously mentioned for the roofing nail model. The connections were different for the two plates.

(See diagrams for the plates connections

egel@main.murray.net.au

Geoff Egel 18 Sturt Street, Loxton ,5333, South Australia ,Australia

---

Join the Energy 21 newsgroup and mailing list today

FREEBIES

Latest Additions

---

nexar
FULL HD
Drive Stress-Free
Buy & Get A Full Rebate

US008102082B2

(12) **United States Patent**
Ogram

(10) **Patent No.: US 8,102,082 B2**
(45) **Date of Patent: *Jan. 24, 2012**

(54) **ATMOSPHERIC STATIC ELECTRICITY COLLECTOR**

(75) Inventor: **Mark Ellery Ogram**, Tucson, AZ (US)

(73) Assignee: **Sefe, Inc.**, Scottsdale, AZ (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 295 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **12/321,306**

(22) Filed: **Jan. 16, 2009**

(65) **Prior Publication Data**

US 2010/0008011 A1 Jan. 14, 2010

Related U.S. Application Data

(63) Continuation-in-part of application No. 12/218,297, filed on Jul. 14, 2008, now Pat. No. 7,855,476.

(51) **Int. Cl.**
H02G 11/00 (2006.01)

(52) **U.S. Cl.** **307/145**; 307/149; 307/151; 361/218; 361/230

(58) **Field of Classification Search** 361/212, 361/215, 216–218, 230, 231; 307/145, 149, 307/151; 174/2; 244/30, 31
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

674,427	A *	5/1901	Palencsar	307/149
1,014,719	A *	1/1912	Pennock	307/149
1,540,998	A *	6/1925	Plauson	310/309
4,486,669	A *	12/1984	Pugh	290/44

* cited by examiner

Primary Examiner — Fritz M Fleming
(74) *Attorney, Agent, or Firm* — Mintz Levin Cohn Ferris Glovsky and Popeo, P.C.

(57) **ABSTRACT**

An antenna for the collection of atmospheric static electricity in which an electrically conductive hub is suspended from a balloon or blimp via a tether. The hub is either solid or uses a spoke/arm arrangement. A number of rods extend from the hub enhance the collection of atmospheric static electricity. The collected atmospheric electricity is conducted from the rods to an electrical connection where the electricity is conducted to earth via a conductive line.

16 Claims, 7 Drawing Sheets

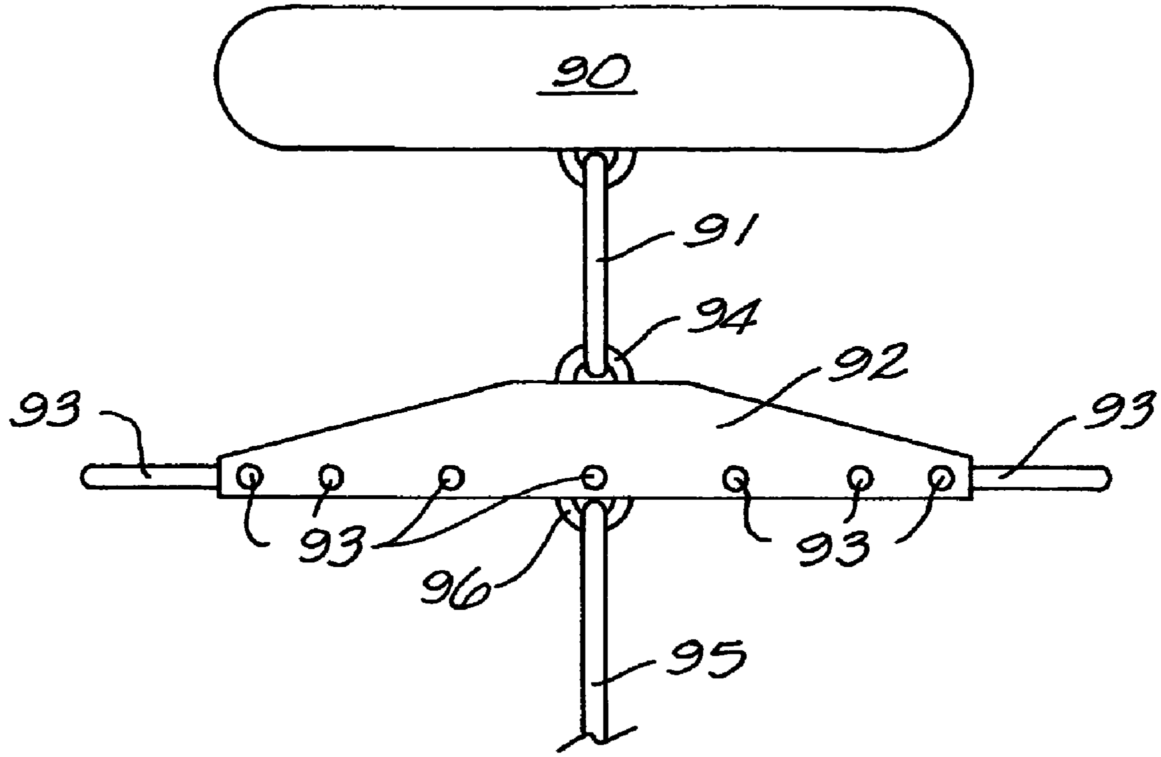

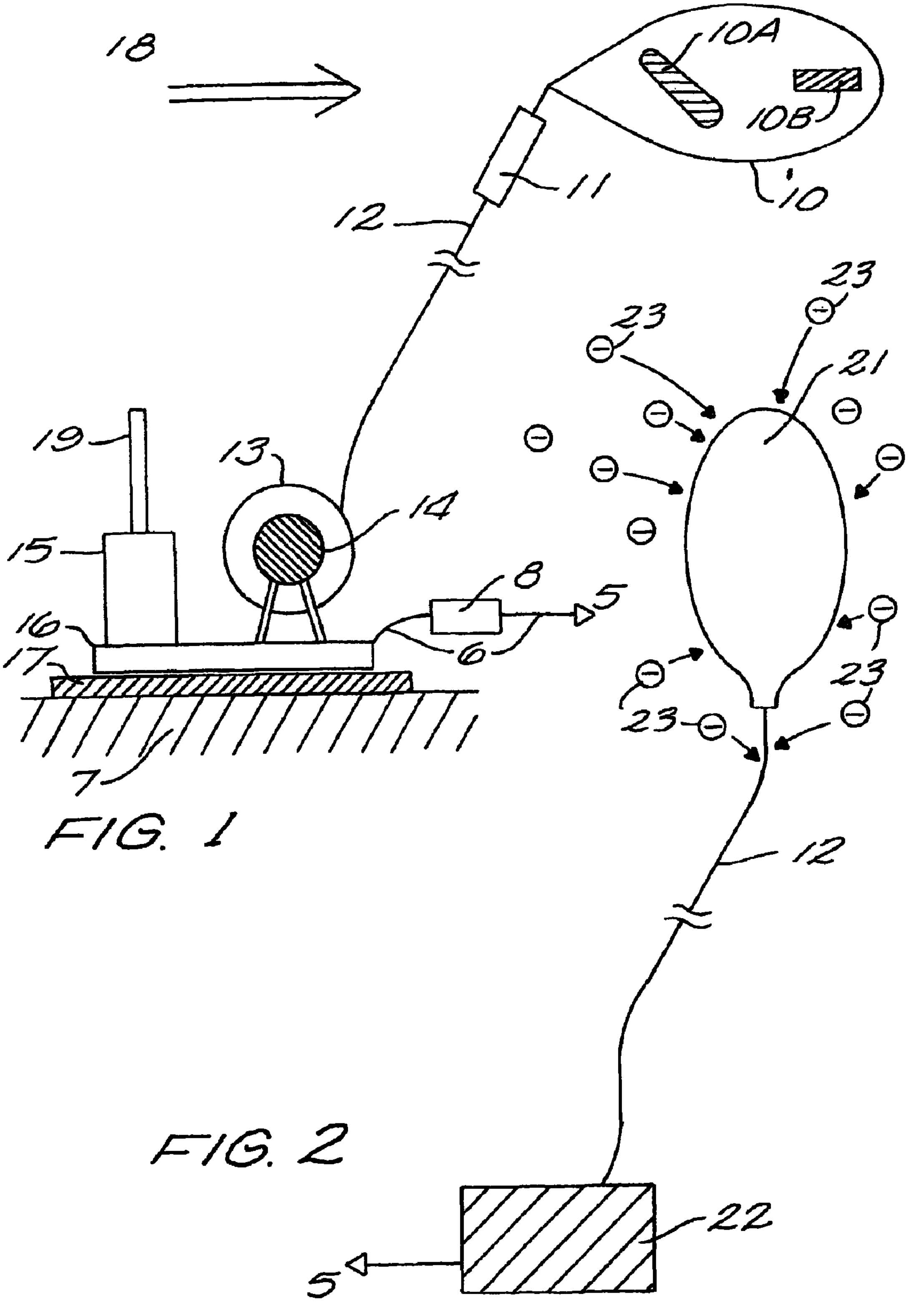
18
10A
10B
10
11
12
19
13
14
15
8
5
16
6
17
7
23
23
21
23
23
23
23
12
22
5

FIG. 1

FIG. 2

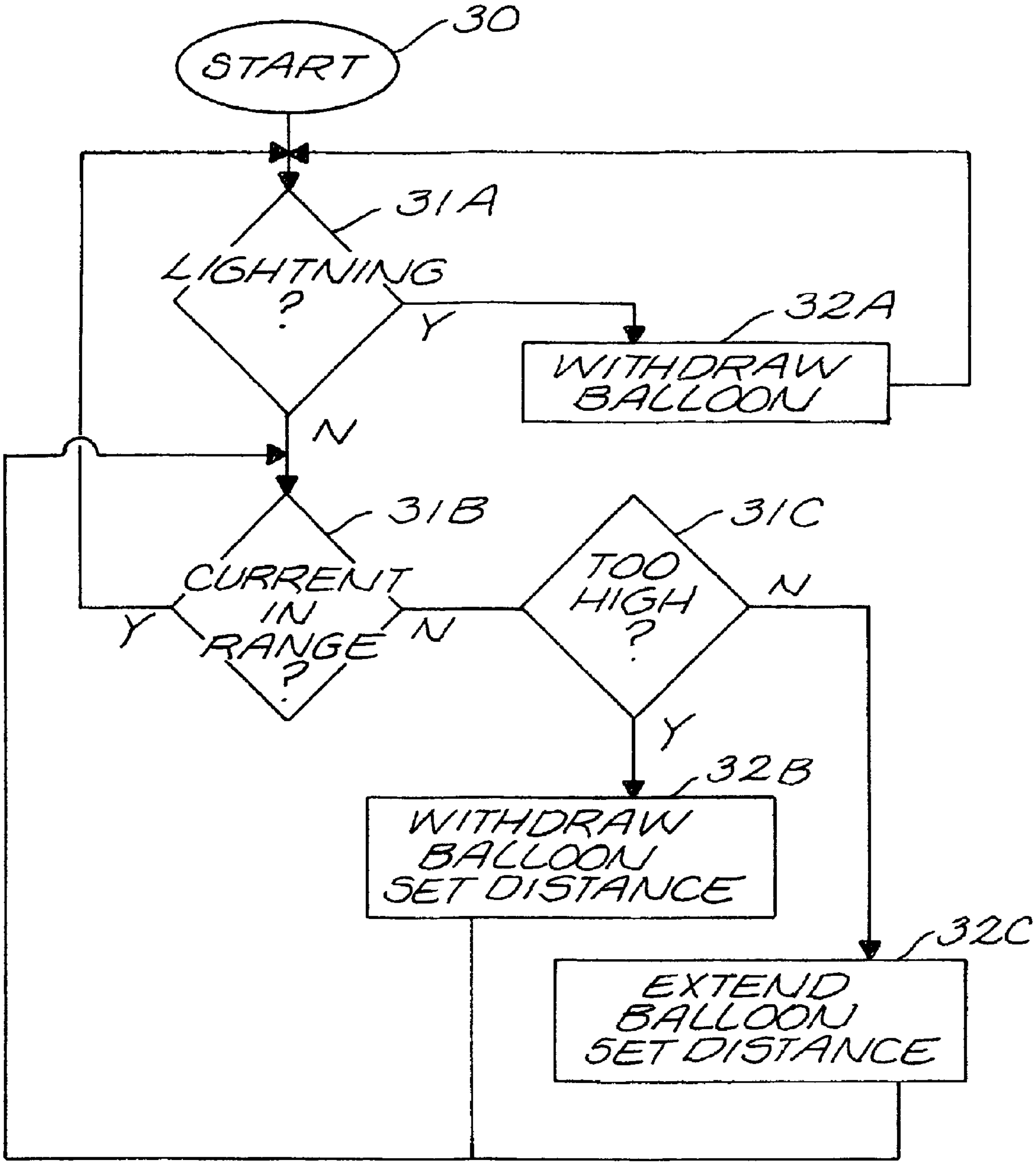
30
START
31A
LIGHTNING ?
Y
32A
WITHDRAW BALLOON
N
31B
CURRENT IN RANGE ?
Y
N
31C
TOO HIGH ?
N
Y
32B
WITHDRAW BALLOON SET DISTANCE
32C
EXTEND BALLOON SET DISTANCE

FIG. 3

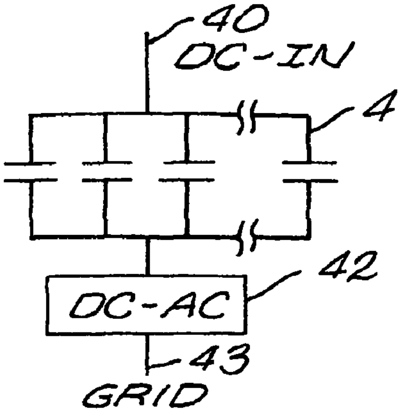
40
DC-IN
41
42
DC-AC
43
GRID

FIG. 4A

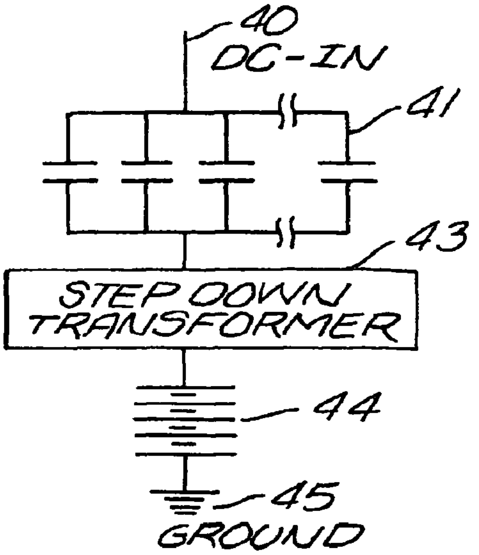
40
DC-IN
41
43
STEP DOWN TRANSFORMER
44
45
GROUND

FIG. 4B

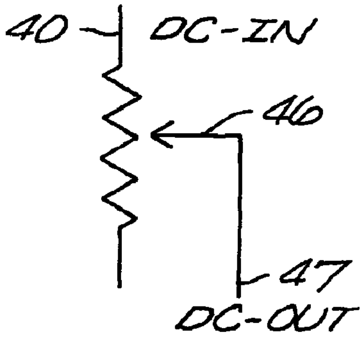
40
DC-IN
46
47
DC-OUT

FIG. 4C

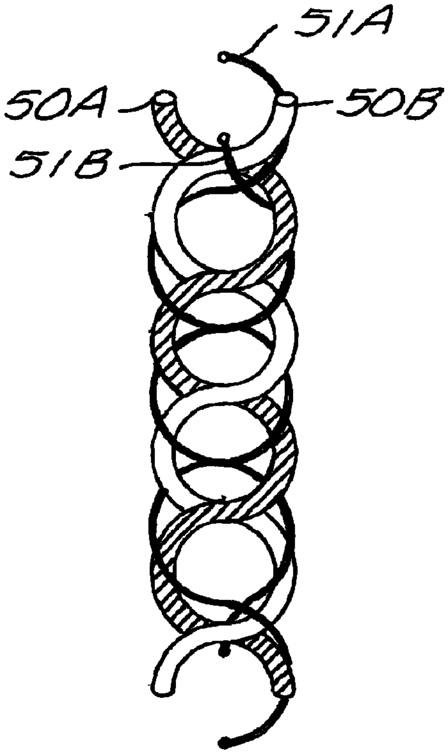
51A
50A
50B
51B

FIG. 5

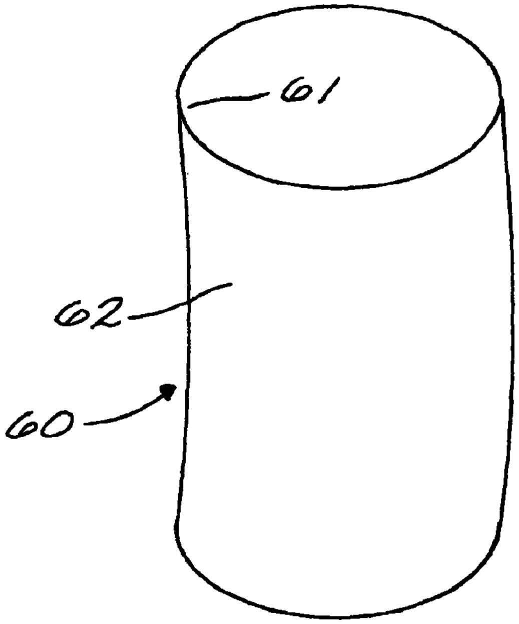

FIG. 6A

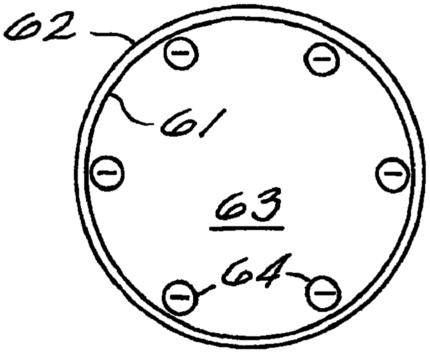

FIG. 6B

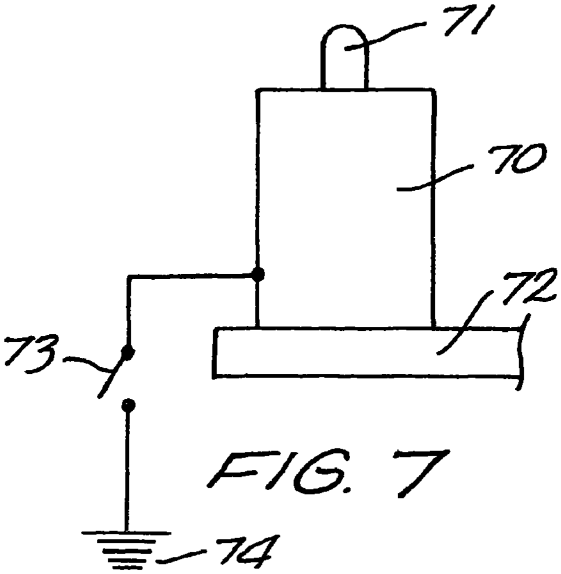
71
70
72
73
74

FIG. 7

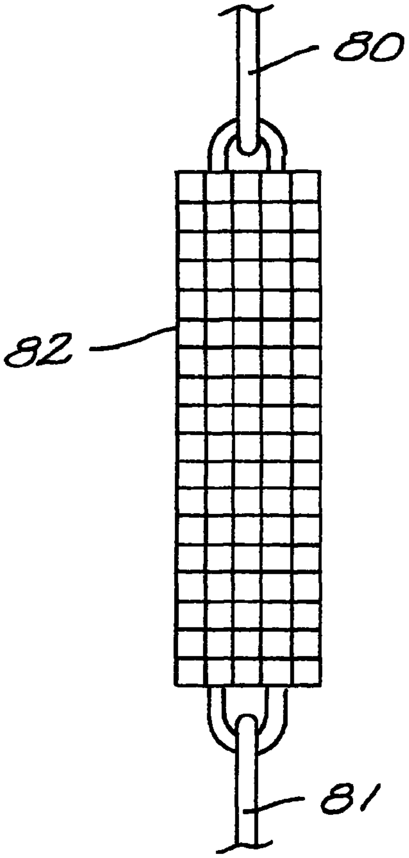

FIG. 8A

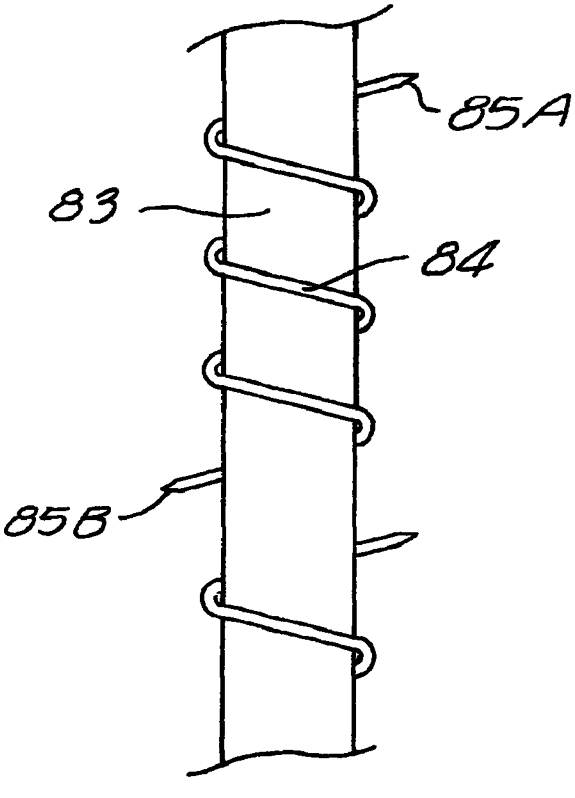
85A
83
84
85B

FIG. 8B

FIG. 9A

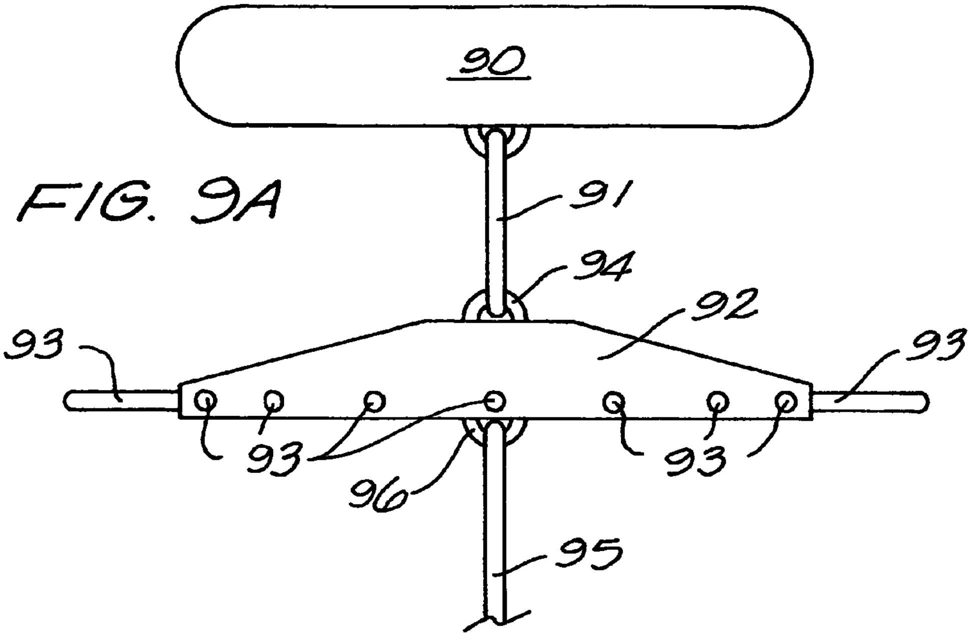
90
91
94
92
93
93
93
93
96
95

FIG. 9B

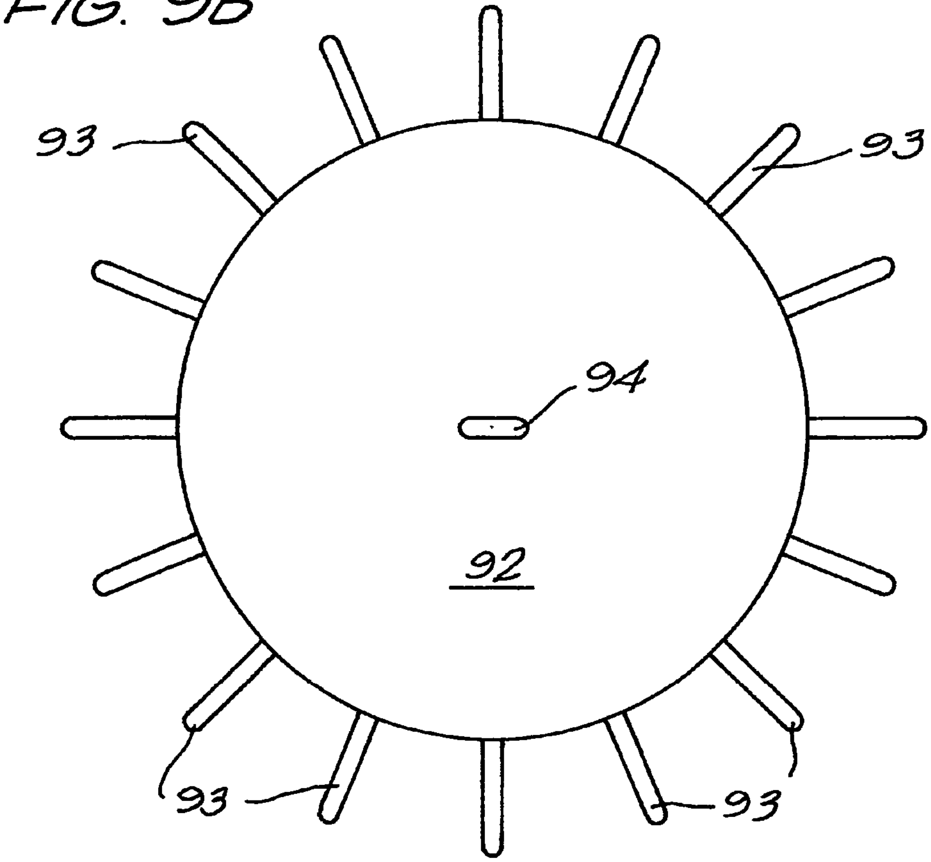
93
93
94
92
93
93

FIG. 10

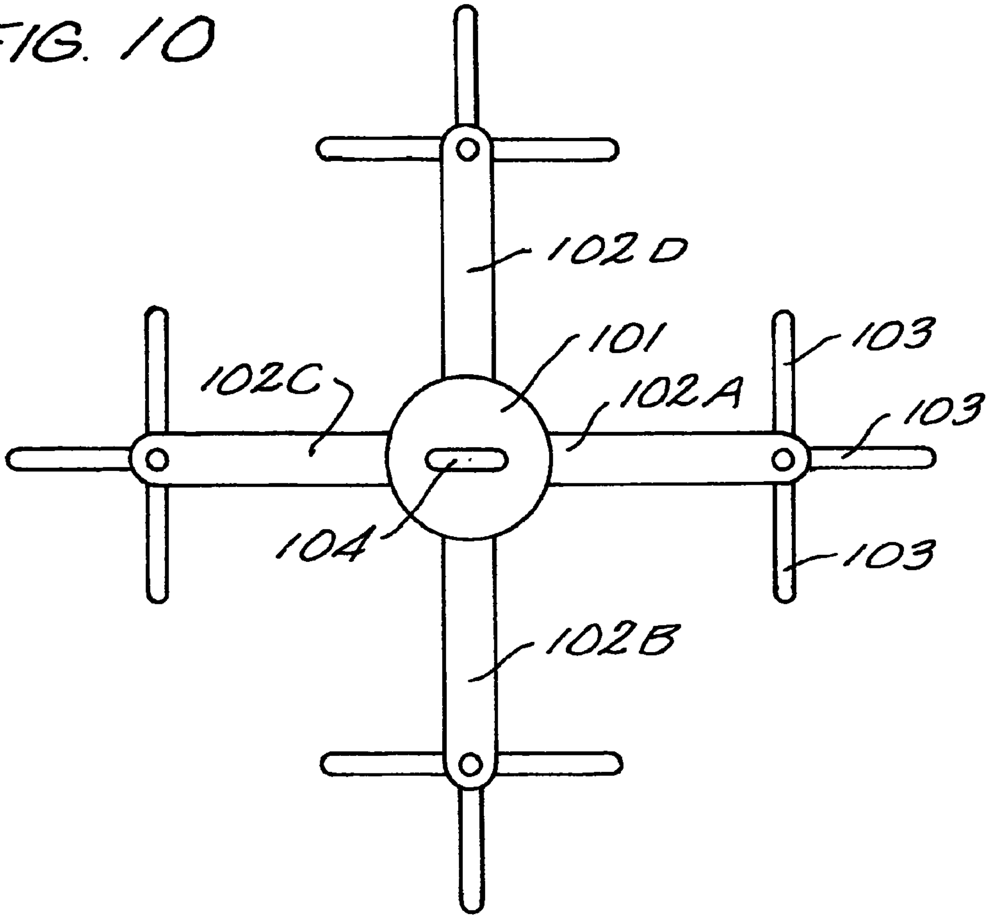
102D
101
103
102C
102A
103
104
103
102B

FIG. 11

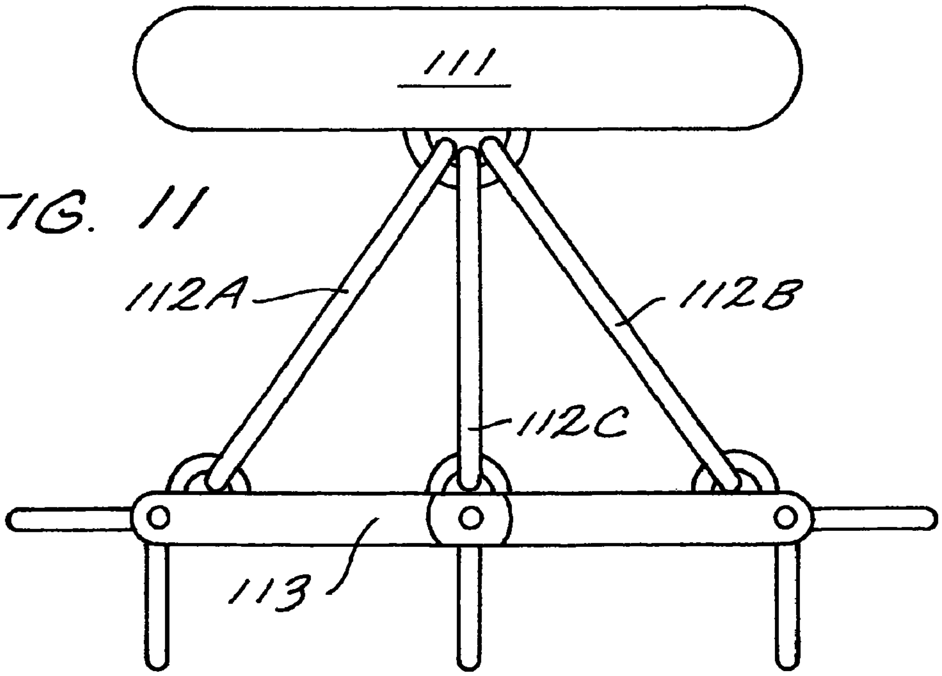
111
112A
112B
112C
113

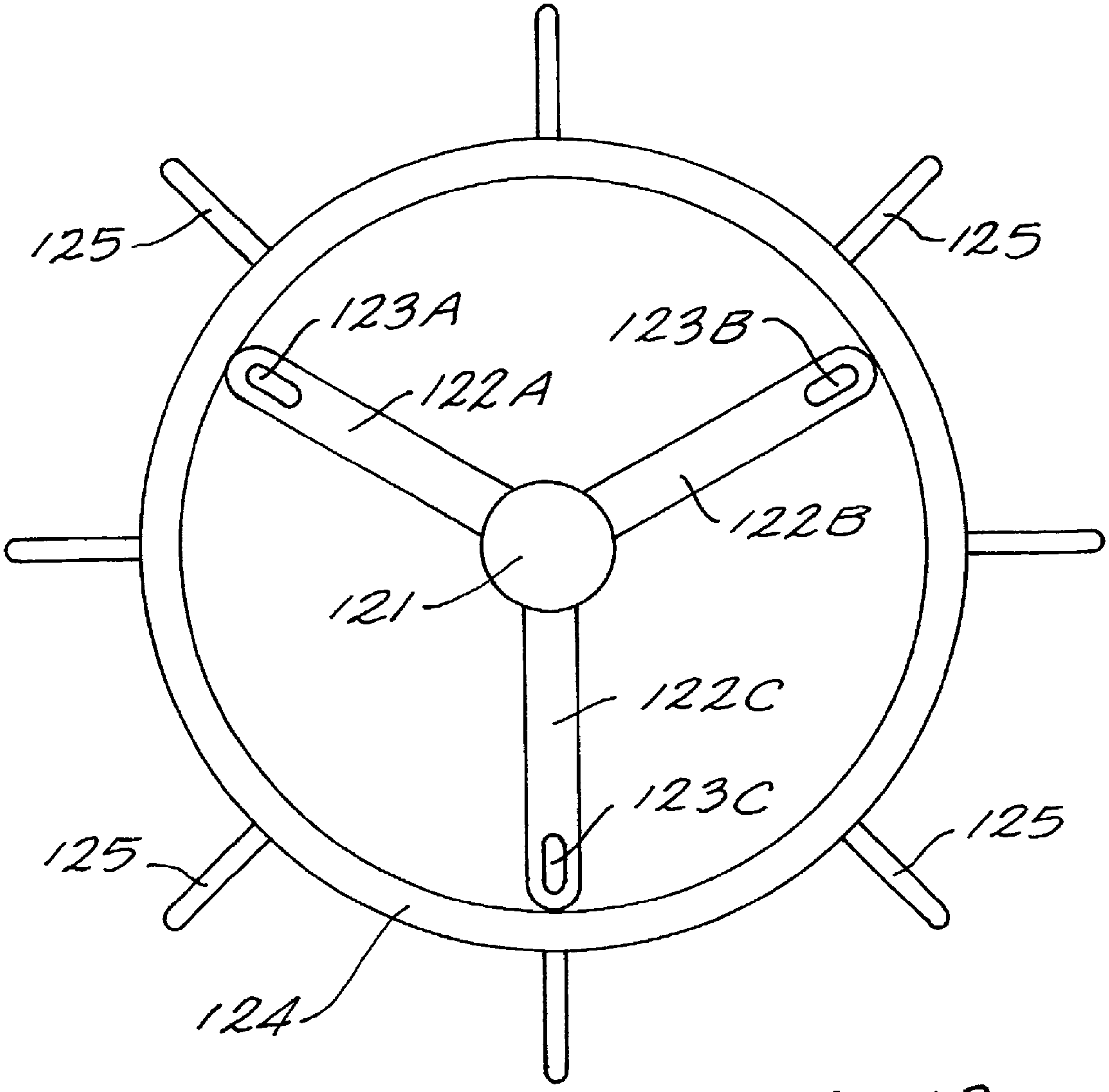
125
125
123A
123B
122A
122B
121
122C
123C
125
125
124

FIG. 12

ATMOSPHERIC STATIC ELECTRICITY COLLECTOR

This is a continuation in part of U.S. patent application Ser. No. 12/218,297, filed on Jul. 14, 2008, now U.S. Pat. No. 7,855,476 and entitled, "Atmospheric Electrical Generator".

BACKGROUND OF THE INVENTION

This invention relates to the production of electrical energy and more particularly to the collection of static electricity from the atmosphere.

Everyone is familiar with Benjamin Franklin's kite experiment of 1752. Using a kite whose string had become wet, negative charges from the passing clouds flowed into the string, down to the suspended key, and then into a Leyden jar via a thin metal wire. Franklin was protected by a dry silk string; but, when Franklin's knuckle came too close to the key, he received a strong shock. Fortunately, Benjamin Franklin was not killed, others who tried this same experiment were not so lucky.

Since then, the formation of lightning has remained something of a mystery. Lightning bolts are triggered when a negatively charged cloud base induces a positive charge from the ground, thereby forming a "pathway" for the discharge of the collected electrical energy.

Lightning travels up to 60,000 miles per hour with a flash that is brighter than ten million 100-watt lightbulbs. This wattage is as much power as is produced by all of the electricity plants in the United States and with a voltage of up to 300 million volts.

It is this very fact, the power within lightning is immense, that has prevented any successful collection of the electrical energy from lightning. The electricity in lightning is far too extreme for current technology to harness.

While lightning has attracted a energy starved industrial world, no one has developed any technique to harness this naturally occurring electrical source.

It is clear there is a continuing need for an electrical source other than carbon-based fuels and that the naturally occurring electricity in the atmosphere is being ignored.

SUMMARY OF THE INVENTION

The invention is a mechanism which taps into the naturally occurring static electricity in the atmosphere. Whereas heretofore, the attempt to garner electricity from the atmosphere has focused exclusively on capturing lightning, the present invention syphons off the static electricity which is generated from any agitated air and voids lightning.

Lightning is only the final discharge of the static electricity, whether that lightning is intra-cloud lightning, cloud-to-ground lightning, or inter-cloud lightning. Other types of final discharges are known as heat lightning, summer lightning, sheet lightning, ribbon lightning, silent lightning, ball lightning, bead lightning, elves, jets, and sprites. Well before these discharges are observed, as the atmosphere becomes agitated by wind or thermal, static electricity is being generated.

The present invention recognizes that this static electricity is being formed and creates a mechanism to capture it.

The mechanism of this invention utilizes an aircraft such as a lighter than air balloon. While the preferred embodiment uses a foil balloon, a variety of other aircraft are obvious to those of ordinary skill in the art, including, but not limited to: gliders, rubber balloons (such as weather balloons), biaxially-oriented polyethylene terephthalate polyester film balloons, and latex balloons.

Within this discussion, the balloon is referenced, but, the invention is not intended to be limited solely to balloons.

The balloon is sent aloft and is tethered by a conductive line. In this context, the conductive line may be any obvious to those of ordinary skill in the art. For the preferred embodiment, the conductive line is a generically referred to as a "poly-rope" and is commercially available through a variety of sources. A suitable conductive line is described in U.S. Pat. No. 5,203,542, entitled "Apparatus for Improved Electric Fence Wire Construction for use with Intensive Grazing" issued Apr. 20, 1993, to Coley, et al. and incorporated hereinto by reference.

The conductive line is played out of a winch to control the altitude of the balloon. The motor controlling the winch is able to reverse direction to both extend and withdraw the conductive line which is wrapped around a spool on the winch. The winch/spool combination are part of a base unit.

In some embodiments of the invention, the spool is constructed of rubber so as to insulate the conductive line from the winch assembly. In this embodiment, only the conductive line is charged by the atmospheric static electricity while the winch remains neutral.

In yet another embodiment, the winch/spool are part of a base unit which is itself isolated from the ground by an insulator. In this embodiment, the entire base unit is charged by the atmospheric static electricity.

A conductor, such as an insulated wire, is electrically connected to the conductive line. In one embodiment, where the conductive line is electrically isolated from the spool and winch motor, the conductor is connected to the conductive line. In the embodiment where the conductive line is electrically connected to the base unit, then the conductor is connected anywhere on a metalic base unit.

The other end of the conductor is connected to a load. The load in this case can be any of a variety of electrical loads well known to those of ordinary skill in art, including, but not limited to a motor, a battery system, or the electrical grid for the system.

In the preferred embodiment, a sensor array is used to monitor the activities both at the base unit (such as electrical flow within the conductor) and in the surrounding locale.

A sensor monitoring the electrical flow (i.e. voltage and/or current) within the conductor is used to monitor the electrical activity within the conductor.

In the preferred embodiment, a lightning sensor monitors for lightning activity within the locale. As noted earlier, the electrical characteristic of lightning is so extreme that ideally this discharge is avoided as it might damage the mechanism of this invention.

The sensor array is utilized by a controller, such as microprocessor, programmed to operate the mechanism as outlined herein.

The controller operates the winch motor to extend or withdraw the conductive line and by extension the altitude of the balloon. The controller is programmed to operate the winch by monitoring the electrical characteristics of the conductor and adjusting the balloon's altitude to maintain these characteristics within the conductor within a preset range.

This preset range is established either in the base programming of the controller or is established by an operator of the system.

As example, by controlling the amount of current being withdrawn from the atmosphere, the mechanism operates within a safe range and also provides a relatively stable current flow from which a variety of activities can take place (such as DC-AC conversion).

The controller also utilizes the lightning sensor to protect the mechanism from a lightning strike. Should lightning be detected within a pre-determined range (as established by the software or defined by an operator), then the balloon is pulled down to minimize the risk of damage from a lightning strike.

An aspect of the present invention is the use of an antenna which are used to collect the atmospheric static electricity. The antenna is shaped as a hub which is suspended from the blimp/balloon. The hub is ideally spoked shaped although an alternative embodiment uses a solid hub.

A number of rods extend from the hub so as collect atmospheric static electricity These rods are ideally rounded at the ends to enhance the attraction of the atmospheric static electricity.

The collected atmospheric electricity is conducted from the rods to an electrical connection on the hub where the electricity is conducted to a power plant on earth such as described above.

The invention, together with various embodiments thereof will be more fully explained by the following description of the accompanying drawings.

DRAWINGS IN BRIEF

FIG. **1** diagrams the preferred embodiment of the invention.

FIG. **2** illustrates the collection of the negative charged particles in the atmosphere.

FIG. **3** is a flow-chart of the operation of the controller for the preferred embodiment of the invention.

FIGS. **4**A, **4**B, and **4**C are electrical schematics for handling the static charge from the atmosphere.

FIG. **5** illustrates a conductive line used in the preferred embodiment of the invention.

FIGS. **6**A and **6**B illustrate an alternative conductive line creating an ionized pathway for the flow of the static charges from the atmosphere.

FIG. **7** illustrates the controller of an alternative embodiment and the associated safety devices.

FIGS. **8**A and **8**B illustrate two embodiments of enhanced electrical collection leads.

FIGS. **9**A and **9**B are side views and top views of an embodiment of the antenna used to collected atmospheric electricity.

FIG. **10** is a top view of an alternative embodiment of the antenna of this invention.

FIG. **11** is a side view of yet another alternative embodiment of the antenna used to collect atmospheric electricity.

FIG. **12** is the preferred embodiment of the antenna of this invention.

DRAWINGS IN DETAIL

FIG. **1** diagrams the preferred embodiment of the invention.

Balloon **10** is an aircraft which, in this illustration, is a lighter than air balloon. Wings **10**A, extending from the body of balloon **10**, provide additional lift in air flow **18**. Tail **10**B helps to stabilize balloon **10**.

Balloon **10** is tethered to the ground via conductive line **12**. As noted earlier, a variety of configurations and materials are available to serve as conductive line **12**. In this illustration, a poly-wire is used. Poly-wire is commercially available through a variety of vendors, including, but not limited to: Jeffers Livestock and Sareba Systems, Inc. of Ellendale, Minn.

In this embodiment, located proximate to balloon **10**, is an electrical collection enhancement lead **11** which assists in the collection of the static electrical charge in the atmosphere. Electrical collection enhancement lead **11** is configured to attract the static charge and conduct the charge into the conductive line **12**.

The electricity flows down the conductive line into spool **13**, where the conductive line **12** is collected and either withdrawn or dispensed through operation of winch motor **14**.

Winch motor **14** and spool **13** are mounted onto base unit **16** which is electrically isolated from ground **7** using insulator **17**. Note, in this embodiment of the invention, when electricity is being collected from the atmosphere, the entire base unit **16** becomes charged. In another embodiment of the invention, spool **13** is constructed of rubber, thereby preventing base unit **16** from becoming charged, thereby restricting the charging from the atmosphere to only conductive line **12**.

In this embodiment, conductor **6** is connected to base unit **16** (since the entire base unit **16** is charged and the base unit is metalic) to communicate the electrical current to load **5**. Conductor **6** is ideally an insulated wire.

The electrical current through conductor **6** is measured using sensor **8**.

In the alternative embodiment discussed above, where only the conductive line **12** is charged, then conductor **6** is connected to conductive line **12**.

Controller **15**, located in this embodiment on base unit **16**, operates winch motor **14** in response to signals from sensor **8** (measuring the current being discharged to load **5**) to maintain the current flow within a pre-defined range. As the current flow diminishes, then the conductive line **12** extended from spool **13** to increase the altitude of balloon **10** to that more static charge from the atmosphere is gathered; as the current flow falls exceeds a preset level, conductive line **12** is withdrawn onto spool **13** to decrease the static charge being collected from the atmosphere.

The range of current flow through conductor **6** is ideally set by the program, although some embodiments of the invention permit an operator to establish this range of operation.

In an alternative embodiment, the sensor monitoring conductor **6** monitors the voltage therein.

In the preferred embodiment of the invention, controller **15** is also equipped with a lightning sensor **19**. In this embodiment, when lightning is sensed within a preset range, then substantially all of conductive line **12** is wound onto spool **13** to pull balloon **10** near the ground and protect the entire mechanism from being damaged from a lightning discharge.

In the preferred embodiment, the “safe” distance form lightning is set in the programming of controller **15** and is ideally two miles; other embodiments permit the operator to “safe” distance.

There are a variety of lightning sensors well known to those of ordinary skill in the art, including, but not limited to those described in: U.S. Pat. No. 7,016,785, entitled “Lightning Detection” issued to Makela, et al. on Mac. 21, 2006; U.S. Pat. No. 6,829,911, entitled “Lightning Detection and Prediction Alarm Device” issued to Jones, et al. on Dec. 7, 2004; U.S. Pat. No. 7,200,418, entitled “Detection of Lightning” issued to Karikuranta, et al. on Apr. 3, 2007; and U.S. Pat. No. 6,961,662, entitled “Systems and Methods for Spectral Corrected Lightning Detection” issued to Murphy on Nov. 1, 2005; all of which are incorporated hereinto by reference.

In another embodiment of the invention, controller **15** is not located on base unit **16**, rather it is remote and communicates its control signals to winch motor **14** using radio waves.

FIG. **2** illustrates the collection of the negative charged particles in the atmosphere.

Static charges **23** are generated in the atmosphere by agitated air. These static charges are often collected at the bottom of clouds, but exist in other environments as well.

Balloon **21** is extended into this strata of static charges **23** which are then attracted to conductive line **12** to flow to base unit **22** and then onto load **5**.

By increasing or decreasing the altitude of balloon **21** (defined by the length of the extended conductive line **12**), conductive line **12** is selectively exposed to varying densities and levels of the static charge strata, and by extension, the current flow or voltage is increased or decreased.

FIG. **3** is a flowchart of the operation of the controller for the preferred embodiment of the invention.

Once the program starts **30**, the lightning sensor is checked to determine if lightning has occurred within the unsafe range **31**A, if it has, then the balloon is lowered **32**A, and the program continues monitoring the status of lightning until no lightning is detected.

When the lightning status is acceptable, then the current within the conductor is checked to see if the current is within the prescribed range **31**B. If the current is acceptable (within range) the program returns to check the lightning status **31**A; otherwise a determination is made to see if the current is above the prescribed range **31**C.

If the current is above the prescribed range, then the altitude of the balloon is withdrawn a set amount **32**B (ideally twenty-five feet) and the program loops back to see if the current is within range **31**B.

If the current is below the prescribed range, then the altitude of the balloon is extended a set amount **32**C (ideally twenty-five feet) and the program loops back to see if the current is within range **31**B.

In this manner of feed-back and minor adjustments in the altitude of the balloon, the current is maintained within a prescribed range which can be handled by the downstream electrical system.

As noted earlier, some embodiments of the invention monitor the voltage instead of the current.

FIGS. **4**A, **4**B, and **4**C are electrical schematics for handling the static charge from the atmosphere.

By maintaining the voltage being collected in a prescribed range, an electrical conversion system is easily designed. While FIGS. **4**A, **4**B, and **4**C illustrate some electrical configurations, those of ordinary skill in the art readily recognize a variety of other configurations which will serve the same function.

Referencing FIG. **4**A, Direct Current In (DC IN) **40** is buffered by a gang of capacitors **41** before being communicated to a DC/AC converter **42**. The DC/AC converter converts the direct current into a an alternating current suitable for placement over an existing electrical grid **43** such as normally found from a power-plant.

Those of ordinary skill in the art readily recognize a variety of DC/AC converters, including, but not limited to: U.S. Pat. No. 7,394,671, entitled "Controller IC, DC-AC Conversion Apparatus, and parallel running system of DC-AC Conversion Apparatuses" issued to Fukumoto, et al. on Jul. 1, 2008; and, U.S. Pat. No. 7,330,366, entitled "DC-AC Converter" issued to Lee, et al. on Feb. 12, 2008; all of which are incorporated hereinto by reference.

FIG. **4**B illustrates an electrical arrangement suitable for use in charging a battery. DC IN **40** is buffered by capacitor bank **41** before entering into a step down transformer **43**. Step down transformer **43** reduces the voltage so that the voltage can safely be introduced into battery **44** which is connected to ground **45** at the battery's other pole.

Those of ordinary skill in the art readily recognize a variety of batteries which will work in this capacity, including, but not limited to those described in: U.S. Pat. No. 7,378,181, entitled "Electric Storage Battery Construction and Manufacture" issued to Skinlo on May 27, 2008; U.S. Pat. No. 7,388,350, entitled "Battery with Electronic Compartment" issued to Wright on Jun. 17, 2008; U.S. Pat. No. 7,397,220, entitled "Connection Member and Battery Pack" issued to Uchida, et al. on Jul. 8, 2008; and, U.S. Pat. No. 7,375,492, entitled "Inductively Charged Battery Pack" issued to Calhoon, et al. on May 20, 2008; all of which are incorporated hereinto by reference.

In FIG. **4**C, DC IN **40** is fed into an adjustable rheostat **46** which is controlled by the controller so that the DC OUT **47** falls within a specified range.

FIG. **5** illustrates a conductive line used in the preferred embodiment of the invention.

This type of conductive line is commonly called poly-wire and consists of multiple interwoven strands of plastic **50**A and **50**B woven into a cord or rope arrangement having intertwined therein exposed metal wires **51**A and **51**B. While this illustration shows two plastic strands and two metal wires, any number of possible combinations is possible.

The exposed metal wires **51**A and **51**B attract the atmospheric static charge and transmit the charge down to the base unit (not shown).

FIGS. **6**A and **6**B illustrate an alternative conductive line creating an ionized pathway for the flow of the static charges from the atmosphere.

This conductive line utilizes a tube **60** having an outer layer **62** of PET Film (Biaxially-oriented polyethylene terephtalate polyester film) which provides exceptionally high tensile strength and is chemically and dimensionally stable. The tube has an ideal diameter of between two and three inches.

An interior metal coating **61** provides an initial conduit for the flow of static charge. The static charge through the metal forces the tube to expand due to the repulsion experienced by like charges. Further, the flow of electricity causes the interior of the tube **60** to become ionized to provide an additional pathway for the atmospheric static charges to the base unit (not shown).

Because outer layer **62** provides a gas barrier, the resulting ionization is not dissipated by air currents, thereby providing a highly stable pathway.

FIG. **7** illustrates the controller of an alternative embodiment and the associated safety devices.

In this embodiment, controller box **70**, resting on insulating pad **72**, is in communication with the sensors as described above. Using the input from these sensors, when there is flow of electricity through the base unit, warning flashing light **71** is illuminated. To electrically neutralize the mechanism, switch **73** is activated to pass any existing current into the ground **74**.

FIGS. **8**A and **8**B illustrate two embodiments of enhanced electrical collection leads.

Referencing FIG. **8**A, enhanced electrical collection lead **82** is a wire mesh which is in electrical communication with conductive line **81** and balloon **80**. Because of the significant amount of metal exposed by enhanced electrical collection lead **82**, more static electricity from the atmosphere is drawn to the collection lead **82**, and then down conductive line **81** to the base unit (not shown).

Conductive lead **82** is positioned proximate to balloon **80**.

In FIG. **8**B, poly-wire **83** has enhanced electrical collection leads **84** wrapped therearound. Collection leads **84** have pointed ends **85**A and **85**B which have a propensity to attract more electricity than rounded ends do.

FIGS. **9**A and **9**B are side views and top views of an embodiment of the antenna used to collected atmospheric electricity.

Blimp **90** supports antenna **92** via tether **91** which is attached to antenna **92** by connection **94**. In this embodiment, antenna **92** is made of electrically conductive material and is ideally light in weight to lessen the payload requirements on blimp **90**.

Encircling antenna **92** and extending therefrom are rods **93** which are rounded at their distal ends so as to enhance attraction of the atmospheric static electricity. The static electricity is communicated to connection **96** and then to electrical line **95** which communicates the electricity to the ground based station (not shown) as described above.

FIG. **10** is a top view of an alternative embodiment of the antenna of this invention.

In this embodiment of the antenna a central hub **101** has the blimp connection **104** (capable of receiving the tether to the blimp) secured thereto. In this illustration, four arms **102**A, **102**B, **102**C, and **102**D extend from hub **101**. Hub **101**, and arms **102**A, **102**B, **102**C, and **102**D are all electrically conductive in this embodiment.

At the end of each arm are rods **103** which are used to enhance the collection of the atmospheric static electricity.

Although this embodiment illustrates four arms, the invention is not intended to be limited to four arms, rather, any number of arms may be used and the number of rods extending from the distal ends of the arms also varies.

FIG. **11** is a side view of yet another alternative embodiment of the antenna used to collect atmospheric electricity.

In this embodiment, several tethers **112**A, **112**B, and **112**C, are used to secure the antenna **113** to the blimp **111**. This arrangement of several tethers provides heightened stability of the antenna by reducing the affects wind will have on the antenna.

FIG. **12** is the preferred embodiment of the antenna of this invention.

In this embodiment of the antenna, arms **122**A, **122**B, and **122**C extend from a central hum and are electrically connected to rim **124**. Tether connectors **123**A, **123**B, and **123**C, are used to secure the antenna to the blimp or balloon.

Rods **125** extend from rim **124** to increase the collection of the static charges in the atmosphere.

It is clear from the foregoing that the present invention captures an entirely new source of electrical energy.

What is claimed is:

1. A system for collection of atmospheric static electricity comprising:

a balloon;

a hub that is electrically conductive and connected to the balloon by a tether, the hub having:

at least three electrically conductive arms, a first end of each of said conductive arms electrically connected to said hub; and

a plurality of rods, a first end of each of said rods being electrically connected to a second end of at least one of said at least three arms;

a base unit having a spool of conductive line on a winch motor, one end of said conductive line secured to said hub, a portion of said conductive line and said hub collecting electricity in the atmosphere, said winch motor selectively extending or withdrawing said conductive line from said spool;

a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load being powered by collected electricity from said conductive line;

an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor; and

a controller receiving said electrical flow indicia and selectively operating said winch motor such that said electrical flow indicia remains within a selected operating range.

2. The system according to claim **1**, wherein each of said plurality of rods has a rounded second end.

3. The system according to claim **2**, wherein at least two rods are attached to the second end of each of said at least three arms.

4. The system according to claim **1**, further comprising an electrically conductive rim connected to a second end of said electrically conductive arms.

5. The system according to claim **4**, wherein at least a portion of said plurality of electrically conductive arms are connected substantially at right angles to an exterior of said electrically conductive rim.

6. A system for collection of atmospheric static electricity comprising:

a radial arrangement of at least three electrically conductive arms;

a plurality of rods, each of said rods being electrically connected to at least one of said at least three arms; and

a balloon tether connected between a balloon and said radial arrangement of at least three electrically conductive arms;

a base unit having a spool of conductive line on a winch motor, one end of said conductive line secured to said radial arrangement of at least three electrically conductive arms, a portion of said conductive line and said radial arrangement of at least three electrically conductive arms collecting electricity in the atmosphere, said winch motor selectively extending or withdrawing said conductive line from said spool;

a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load being powered by collected electricity from said conductive line;

an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor; and

a controller receiving said electrical flow indicia and selectively operating said winch motor such that said electrical flow indicia remains within a selected operating range.

7. The system according to claim **6**, wherein each of said plurality of rods has a rounded end distal from said electrically conductive arm.

8. The system according to claim **7**, further including a balloon tether connection secured to the first end of said at least three arms.

9. The system according to claim **8**, further including an electrical connection secured to the first end of said at least three arms.

10. The system according to claim **9**, wherein said at least two rods attached to said at least three arms is five rods.

11. The system according to claim **9**, further including an electrically conductive rim connected to a second end of said electrically conductive arms.

12. The system according to claim **11**, wherein said plurality of arms are electrically connected to said electrically conductive rim.

13. The system according to claim **12**, wherein at least a portion of said plurality of rods are connected to an exterior of said electrically conductive rim substantially at right angles thereto.

14. A system for the collection of atmospheric static electricity comprising:

a generally circular electrically conductive body;

a balloon tether connected to a first side of said generally circular electrically conductive body and to a balloon; and

an electrical connection connected to a second side of said generally circular electrically conductive body;

a base unit having a spool of conductive line on a winch motor, one end of said conductive line secured to said electrical connection, a portion of said conductive line and said generally circular electrically conductive body collecting electricity in the atmosphere, said winch motor selectively extending or withdrawing said conductive line from said spool;

a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load being powered by collected electricity from said conductive line;

an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor; and

a controller receiving said electrical flow indicia and selectively operating said winch motor such that said electrical flow indicia remains within a selected operating range.

15. The system according to claim **14**, wherein said generally circular electrically conductive body comprises:

an electrically conductive hub;

at least three electrically conductive arms, each of said arms connected at a first end to said hub; and

an electrically conductive rim connected to the second end of each of said at least three arms.

16. The system according to claim **15**, further comprising a plurality of electrically conductive rods, a first end of each of said rods connected to said rim and extending therefrom substantially at rights angles thereto.

* * * * *

The Lakhovsky Multiple Wave Oscillator Handbook

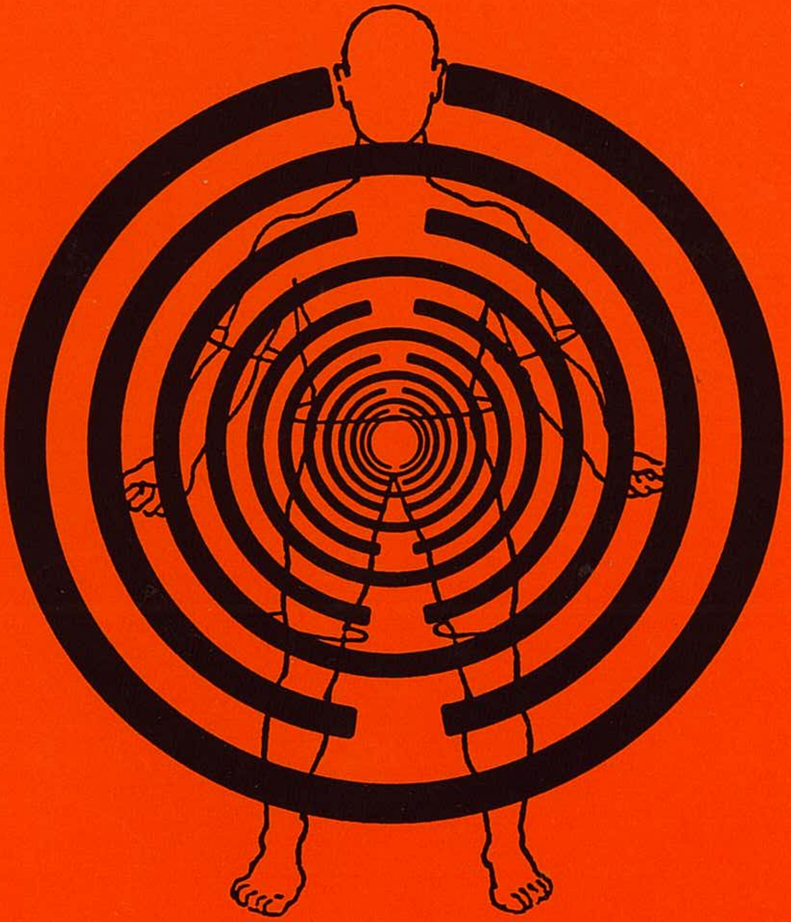

Compiled and Edited by

Thomas J. Brown

The Lakhovsky Multiple Wave Oscillator Handbook

Comprising the
Borderland Sciences Research Foundation
Lakhovsky Multiple Wave Oscillator
&
Radio-Cellular Oscillator
Research Files

Compiled and Edited by
Thomas J. Brown
Executive Director of
Borderland Sciences Research Foundation

Published May 31, 1992
in the Light of further research,
in the interests of BSRF Associates by:

BORDERLAND SCIENCES RESEARCH FOUNDATION

P.O. BOX 429 ★ GARBERVILLE, CA 95440-0429 ★ U.S.A.

©1992 Borderland Sciences Research Foundation

All rights reserved. No part of this pubication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior permission of the publisher except for brief passages quoted in reviews.

ISBN 0-945685-03-3

Library of Congress Catalog Number 92-71664

Third Revised Edition, Published May 31, 1992 by:

Borderland Sciences Research Foundation
Garberville, California 95440-0429

10 9 8 7 6 5 4 3 2 1

DISCLAIMER

This book is a research publication of Borderland Sciences Research Foundation. It consists of our MWO and RCO files which have been put into some semblance of order to preserve the records of many people with new and different ideas on how to study the energetic states of living systems and how these systems may be influenced by externally applied energies, and also to allow for further experimentation in this field by competent researchers. Read this book thouroughly before undertaking any experiments becasue you will find many different opinions, some conflicting, as to how to proceed. High frequency electrical apparatus can cause serious shock injury and should only be handled by knowledgable, competent people. All experiments are conducted solely at your own risk. We disclaim any responsibility for the use or misuse of any information in this book. While this book touches on the subject of human health, all information should be considered as anecdotal and should not in any way be construed as medical advice. We are not qualified to give medical advice and the use of the experimental equipment and research techniques described in this book may provide a false sense of security wherein one may neglect professional medical care. Use your brains. Thanks.

Tom Brown
Editor/Publisher

Table of Contents

PART ONE -- Lakhovsky's Writings and Work

Radiations and Waves – Source of Our Life by Georges Lakhovsky 5-6
Influence of Cosmic Waves on the Oscillation of Living Cells by Georges Lakhovsky 7-8
The Radio-Cellular Oscillator and its Applications by Georges Lakhovsky 9-10
Attempt at Treatment of Experimental Plant Cancer by Georges Lakhovsky.................... 11-13
The Lakhovsky Radio-Cellular Oscillator by Jorge Resines 14-19
Apparatus With Circuits Oscillating Under Multiple Wave Lengths by Georges Lakhovsky, U.S. Patent 1,962,565 20-21
Tube for Producing Multiple Wave Lengths by G. Lakhovsky, U.S. Patent 2,351,055 22-25
Lakhovsky's Vibratory Circle 26-29

PART TWO -- The BSRF MWO Research Reports

The Russian Lakhovsky Rejuvenation Machine by Bob Beck, 1963 30-35
Transistorized MWO by G.W. Steele 36-37
The Lakhovsky Multi-Wave Oscillator Power Supply by Klark Kent 38-45
Excerpts from **The Waves That Heal** by Mark Clement with commentary by Riley Crabb ... 46-52
Curing Cancerous Plant With Ultra Radio Frequencies by Georges Lakhovsky, 1926 52-53
Excerpts from **The Journal of Borderland Research, 1964-1971** 54-83
BSRF MWO 1988 Update 84-94
The Spiral Coil Multi-Wave Oscillator by Louis A. Schad 95-103
Electrostatic MWO by Peter Lindemann 104-105
The Sun-Ether Disc of Oscar Korschelt..106-107
View From the Borderlands by Tom Brown, 1988-90 108-111
Excerpts from **The Journal of Borderland Research,** 1990-1992 111-122

PART THREE -- Reference Materials

High Frequency Oscillators for Electro-Therapeutic Purposes by Nikola Tesla.............123-133
The Magnetic-Arc Instrument of S.S. Knight by Mark Gallert 134
Frequency-Independent Antennas by John D. Dyson 135-137
Short Waves in Medicine: **Medical Aspects of Ultra Short Waves,** C.H. West; **Ultra Short Waves in Medicine,** Dr. F. Noack; **Human Beings as Antennas,** Dr. I. Schliephake.........137-142
Distributed Series Capacity by C.P. Steinmetz 143-144

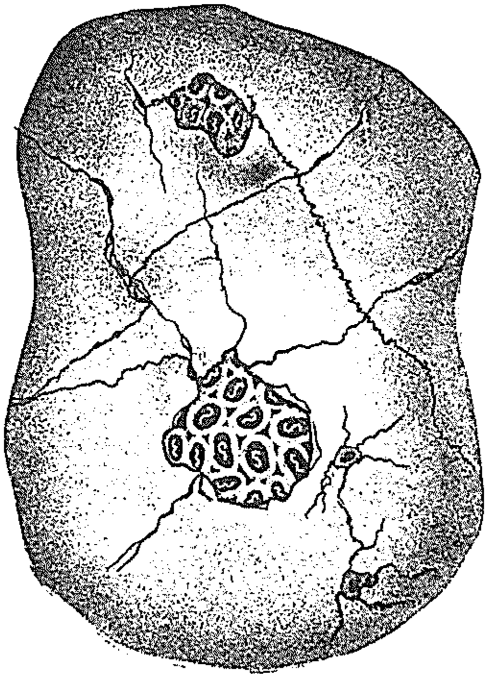

A dead cell. This diagram shows a cell where the chromosomes have stopped oscillating, and for this reason have ceased to live. But the chondromes, visible through the opening, continue to live an independent life. They are surrounded by a tough membrane and are multiplying so as to burst open the outside of the cell, as a baby chick, when it comes to maturity, breaks through the egg shell to come out.

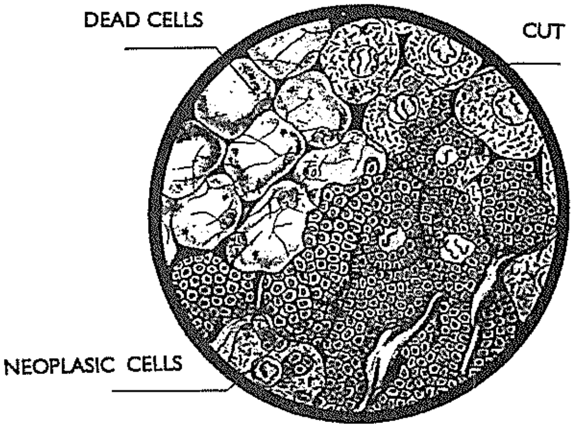

A group of dead cells, but not yet destroyed, in which the neoplasic cells are forming. One can see the transformation in the neoplasic tissue.

Photograph taken during the treatment of Mrs S..., who sits between the MWO antennas at Calvary. Behind the patient is Dr. Marquer, head physician of the hospital, to the left is Dr Z..., assistant, three voluntary Nurses, & to the right is Georges Lakhovsky.

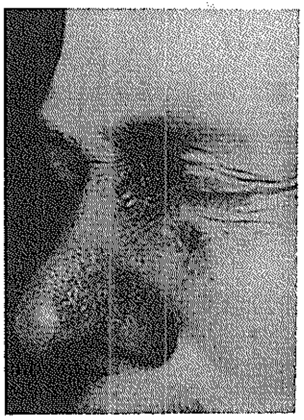

At left, Madame C... *Before Treatment.* Rodent Ulcer in inner angle of left eye.

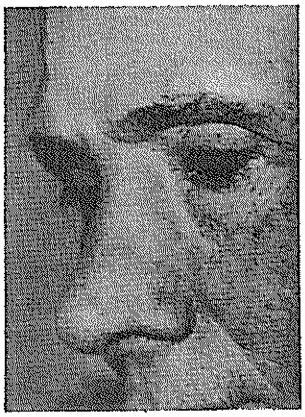

At right, Madame C... *After Treatment.*

Excerpts from:

RADIATIONS AND WAVES

Source of Our Life

by Georges Lakhovsky, 1941

THEORY OF CELLULAR OSCILLATION

Before presenting the therapeutic results obtained with my multiple short wave oscillator and presented at the International Congress of Short Waves in Vienna (July, 1937) it may be useful to give a rapid survey of my theory of cellular oscillation which I have developed fully in a number of books.*

All living cells are composed of two essential elements; the nucleus and the protoplasm in which it is bathed. This nucleus is itself composed of many tubular filaments: the chromosomes. In addition, hundreds of much smaller filaments or chondromes are present in the cytoplasm.

Chromosomes and chondromes are sheathed in an insulating substance (cholesterine, resin, fat, plastrin, etc.) and contain a liquid-like serum with the same mineral content as sea water, and consequently a conductor of electricity. Thus, these filaments constitute ultramicroscopic oscillating circuits capable of oscillating electrically over a wide scale of very short wavelengths.

I have demonstrated in my works that these cellular oscillating circuits, chromosomes and chondromes, vibrate electrically under the stimulus of electro-magnetic waves: cosmic, atmospheric and telluric.

Now, many internal and external influences may upset the oscillating equilibrium of these cells. For instance, a variation or change in the field of cosmic, telluric or atmospheric waves, a demineralization of the organic matter constituting the cellular substance, traumas causing the destruction by shock of the protoplasm or the nucleus.

I have shown in my books, **The Secret of Life** and especially in **The Earth and Ourselves**, that every living cell draws its oscillatory energy from the field of secondary radiations resulting from the ionization of the geological substances of the earth by cosmic radiations.

But certain natural radiations are particularly toxic, especially those originating in geologically-induced geopathic zones. Many cancer cases have been attributed to these toxic radiations which has been proven experimentally, notably in Germany by Dr. Rambeau of Marburg. Therefore, earth radiations sometimes cause disturbance of the cellular oscillatory equilibrium of the organism.

Under these varied circumstances cellular oscillation may cease; the cell is then dead. But within the dead cell, the chondromes sometimes continue to oscillate electrically on their own natural frequencies. Fortunately, this phenomenon occurs rarely, or all humanity would already have perished from cancer.

The chondromes then envelope themselves in a membrane and continue to oscillate and multiply independently of the cell. They may then become neoplasic cells.

THE RADIO CELLULAR OSCILLATOR

To re-establish this equilibrium, I thought of creating, in 1923, a constant compensating field of very short radiations (2 to 10 metre wavelengths) to neutralize the action of the disturbing rays, and give the living cell the necessary stimulation for a return to its normal oscillation.

To this end, in 1923, I constructed my short-wave oscillator, using two triode tubes for very short waves made especially for this apparatus. I tried several cross leakages for this machine using one or more tubes and then multiple triodes with a tube containing oscillatory circuits within the bulb. Finally I adopted the oscillator with symmetrical cross leakage comprising two triodes. The oscillating circuits formed a single spiral, branched between the two grilles and the two anodes. It was fed directly by alternating current from the local supply circuit.

With this short-wave apparatus I was able to cure plants inoculated with cancer. For six years at the Saltpetriere I observed and checked the effects of these short waves.

Using very low power, from 10 to 12 watts, and a limited duration of treatment, I succeeded in curing cancer in human beings, but also had to record some failures.

The news of the success of my experiments became widespread. In many countries, as early as 1928, they began to build short-wave oscillators of considerable power producing thermal effects.

But here was great danger that the chromosomes and

*** The Secret of Life; Contribution to the Etiology of Cancer; The Earth and Ourselves; Cellular Oscillation; Nature and Her Wonders.**

chondromes, which are barely a ten-thousandth or twenty-thousandth of a millimeter in thickness, might not survive under a high frequency current. They offer much resistance, even to a low current which is sufficient to dissolve and destroy them.

It is simple to prove this by bringing a small bulb of from 2 to 5 volts with a filament of several hundredths of a millimeter, inserted between metal rods forming antennas, within the radius of a short wave transmitter. The bulb will light up and sometimes burn out, if it is brought too near the apparatus.

Moreover, the chondromes and chromosomes of all living cells, which are infinitely finer than the filaments of the bulbs, are sensitive centers of thermal phenomena, which may provoke their fusion. Undoubtedly this method is effective in killing microbes in the organism and in neoplasic cells. But it can also destroy millions of cells of healthy tissue in every irradiation.

THE MULTIPLE WAVE OSCILLATOR

On the other hand, I thought it possible to obtain better results by administering an oscillatory shock to all the cells in the organism simultaneously. Such a shock, very brief, produced by dampened or weakened electrostatic waves, does not provoke thermal and prolonged effects and involves no risk of burning living cells.

I therefore sought to produce an artificial oscillatory shock causing a periodic oscillation of the weak or dead cells.

At first glance this problem seems physically insoluble as there are approximately 200 quintillion cells in our bodies, each oscillating on its own natural wavelength. Theoretically, therefore, we would have to have a different wave length for each cell, so that every cell in the organism would oscillate in resonance on its own wavelength.

After much research I was able to construct an apparatus creating an electrostatic field covering all frequencies from 3 metres to the infra-red, so that every cell can find its natural frequency and vibrate in resonance.

We know that in physics, a circuit fed by damped or weakened high frequency currents creates many harmonics. Consequently, I conceived an oscillator of multiple wavelengths with a broad scale in which every organ, every gland, every tissue, every nerve, could find its natural frequency.

To obtain this result I set up a transmitter composed of a series of circular concentric oscillating circuits linked by a silk cord but not contiguous. These circuits are stimulated by damped high frequency currents from a spark gap. Thus each circuit of the transmitter vibrates not only on its natural frequency, but also on numerous harmonics.

Thus, I built an oscillator with all the basic wavelengths from 10 centimetres to 400 metres, that is, all frequencies from 750,000 to 3,000,000,000 cycles per second. But each circuit also emits many harmonics, which, with their basic waves, their interferences and their effluvia can reach the scale of infra-red and even that of visible light (1 to 3000 trillion vibrations per second.)

Since all the cells as well as the chondromes oscillate precisely at frequencies in that range, they can therefore find, in the output of such an oscillator, the frequencies which cause them to vibrate in resonance.

You know the results I obtained with continuous very short waves using triode tubes at a distance, with no contact electrode....

As early as 1931, I began using this multiple wave oscillator in various Paris hospitals: the Saint Louis, Val-de-Grace, Calvary, Necker, etc... Among the many cures with this treatment, I will mention especially those of various cases of cancer which X-ray and radium treatments failed to improve. These patients, cured six years ago, have had no recurrence and are in perfect health at this date. In all pathogenic cases this treatment gives very good results. As it does not attack the microbes directly, it does not destroy live tissue, but reinforces the vitality of the organism by accelerating cellular oscillation. It is therefore the reinforced organism that successfully resists the microbes and all pathogenic causes.

So, while X-rays and radium destroy microbes, neoplasic cells and healthy tissue at the same time -- which accounts for the serious accidents which occur during and after such treatments -- high frequency radiations (short waves) applied at a distance and without thermal effect cure diseases of all kinds, even those of the prostate to a considerable degree.

Whatever the pathogenic cause, the multiple wave oscillator reproduced the frequencies necessary to re-establish the cellular oscillatory equilibrium.

In general, it is sufficient to seat the patient, or to have him stand in the radius of the apparatus, before the transmitter. The duration of the treatment is usually from five to ten minutes, every other day. These figures are purely arbitrary, since these radiations reinforcing cellular oscillation do not produce organic disturbances, whatever the duration of the treatment may be.

Despite the many cases successfully treated, almost without exception, it must be understood that my oscillator cannot cure all types of cancer in all its stages of development. In many cases, when the cancerous tumor has already destroyed important blood vessels, my apparatus is powerless to rebuild that tissue before the fatal haemorrhage.

INFLUENCE OF COSMIC WAVES ON THE OSCILLATION OF LIVING CELLS

Part 3, Chapter 1, From La Science et le Bonheur, by Georges Lakhovsky, Paris 1930

Translated by Brigitte Donvez, Edited by Tom Brown

© 1991 B.S.R.F.

This part of my study is devoted to the trials, experimentation and applications of the theory of the oscillation of living beings.

I have assembled here short technical reports about the processes and methods employed, as well as a collection of all the papers read at the French Sciences Academy and at various French and foreign Scientific Societies about oscillatory therapeutics and all its domains of application.

We intend first to examine successively the influence of cosmic waves on the oscillation of living cells, taking as a concrete example the influence of solar activity upon the quality of wine and vintage. Then we will examine the construction of an electromagnetic short-wave generator, such as the radio-cellular-oscillator for the treatment of animal and plant organisms.

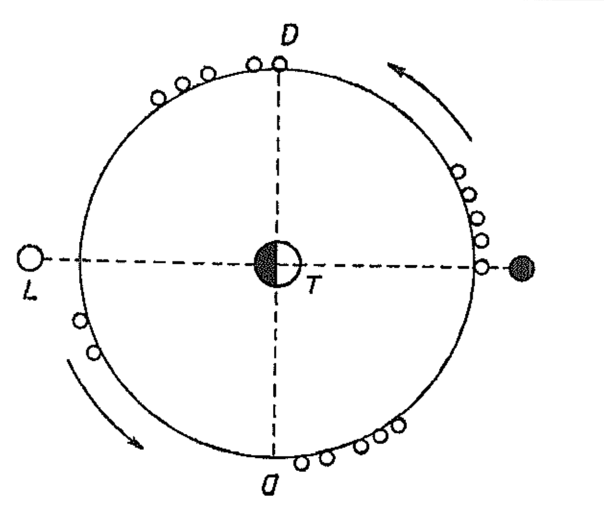

Fig. I -- Distribution of the highest intensity of electromagnetic waves observed by Mr. Vincent, as a function of the moon phases.

We will later consider the applications of the radio-cellular-oscillator, particularly in the experimental treatment of plant cancer, diathermy and the artificial production of fever.

Concurrently, we will introduce the application of open oscillating circuits to the treatment of diseases in various living beings, in the absence of any other oscillations but those produced by the natural cosmic and atmospheric waves. We will also review in this chapter the observations of scientists and practitioners who clinically applied my oscillating circuits.

Finally we will take up the very latest applications of oscillating circuits to water and liquid sterilization, and to the activation of the germinative power of plants.

Obviously, the number of these applications, presently limited to certain domains, tends to rapidly expand and touch upon the various problems presented by therapeutics, biology, agriculture and zoology.

I have discussed elsewhere the role played by cosmic radiations in the development of living beings, due to the very oscillatory nature of the cell. Now, as a consequence of the earth's rotation and the stars' reciprocal motion, the cosmic and atmospheric radiation field is not constant, but rather variable. Living beings are subsequently under the influence of these variable cosmic radiations, whose nature either weakens or strengthens the cell's oscillatory balance.

I have mentioned elsewhere[1] that the lunar radiation interferes with the waves of radioelectric stations. Every week, the recurrence of moon phases enables us to obtain highest and lowest intensities in the reception of electromagnetic waves (see Fig. 1).

On the other hand, the sun does not only send light, heat and actinic (ultra-violet) radiations, but also electromagnetic waves, especially during the eruptive periods of its prominences (sunspots). The influence of the sun's electromagnetic radiation becomes obvious if we consider the variations in the earth's magnetic field and the spinning of a compass' magnetic needle during magnetic storms, under the influence of the telluric currents induced in the ground. The ionization produced by the sun's radiation also generates atmospherics and statics which disturb radioelectric receptions.

Astrophysicists, notably those of the Meudon Observatoire under the direction of Mr. Deslandres, have been able to establish a correlation between the periodicity of cataclysms on earth (tidal waves, tides, earthquakes) and that of the recurrence of sun-spots in the earth plane (about 27 days).

We can observe that the graphs which render geophysical, electric (ionization, gas conductivity, aurora borealis—northern lights), magnetic (variations of the earth field, compass spinning) and electromagnetic (disturbance in wave propagation, atmospherics and telluric parasites) phenomena, as a function of time, are very appreciably parallel to each other and to the curve of solar activity. An 11½ year periodicity appears on these graphs, confirmed by numerous, previous observations in the domain of meteorology. In particular certain tropical rainfall patterns show a 33 to 35 year periodicity, three times longer than the previous one.

I researched the extent to which cosmic radiations which condition physical and meteorological phenomena, also influence biological phenomena and the development of living beings.

As early as 1801, while cosmic radiations were not even suspected, Hershel pointed out: "the scarcity of vegetation when the sun has no spots". Starvations recur every eleven years in India with alternances of heat and rain. From 1901 to 1909, the canon Th. Moreux noticed that the wheat crop in France and in the world more or less followed the solar activity, so did arthritic conditions, and the nervousness of students and teachers.

I compared the graphs of the solar activity, as recorded at the Meudon Observatory, with the statistics of wine production in the Bourgogne and Beaujolais regions, and I noticed that the outstanding vintages coincided precisely with the year of recrudescent solar activity.

On this topic, Professor d'Arsonval agreed to present in my name at the Academie des Sciences (Academy of Sciences) on March 28, 1927, an original report entitled Influence of Cosmic Waves On The Oscillations Of Living Cells which is entirely reproduced here.

---

[1] L'univérsion, p.127, Gauthiers Villars publishers

INFLUENCE OF COSMIC WAVES ON THE OSCILLATION OF LIVING CELLS

(Paper by Mr. Georges Lakhovsky, presented on March 28, 1927 to the Academie des Sciences (Academy of Science) by Professor d'Arsonval.)

In my book The Origin of Life, Radiation and Living Beings, that Professor D'Arsonval did me the honor to present to the Academie des Sciences (Academy of Science), at the sitting of the 15th of February 1926, I exposed my theory about the influence of (cosmic) penetrating radiations on living beings. As a matter of fact, I demonstrated that the nucleus of each living cell, having the form of a tubular filament made of a dielectric matter filled with a conductive substance, could be compared to an oscillating circuit, equipped with self-inductances, electric resistance and capacity. The living cells can thus oscillate at very high frequencies under the influence of cosmic radiation emitted by stars.

I made certain to verify the exactness of this theory by studying the influence of cosmic radiations (sun-spots, comets, interferences of stars' radiations, etc...) on biological matter.

My observations were based on the curves recorded by the astrophysicists of the Meudon Observatoire, curves which revealed since 1845 the activities of the sun-spots, the frequency of magnetic disturbances, and the frequency of aurora borealis -- northern lights.

These three curves are appreciably parallel. I endeavored to study the existing correlation between these cosmic radiations, on the one hand, and the development of vital activity in the plant and animal domains, on the other. As for any given individual, periods of fatigue and illness alternate with periods of health, in the same way there is, for fruit and crops in general, for each species, years of good quality and years of rather mediocre quality.

As to what concerns wine, according to the documentation established by the Chambers of Commerce of the Bordeaux and Bourgogne regions, I observed that, from 1845 until 1915, the outstanding vintages corresponded exactly to the highest activity of the sun-spots, as shown by the curves on Figure 2.

The results for red Bordeaux wines are:

Maximum of 1848: outstanding vintages of 1847 and 1848;
Maximum of 1858: outstanding vintages of 1857 and 1858;
Maximum of 1869: outstanding vintages of 1869 and 1870;
From 1880 to 1889: period of phylloxera.
Maximum of 1893: rather good vintages of 1890 and 1893;
Maximum of 1906: good vintages of 1906 and 1907.

In addition, let's point out that the wine of 1811 was called the wine of the comet, and we could ascribe its excellent quality to this comet's radiation. It is the same with white Bordeaux and Bourgogne wines.

I am sure that there are many other biological phenomena which recur periodically after a few years or a few centuries as a function of the relative revolution of the earth around the stars. I am convinced that all the scourges which ravaged humanity during the Middle-Ages: plague, leprosy, cholera, etc... and disappeared thereafter without traces in other regions will reappear some day, when the earth shall once more occupy the same position as then in relation to the stars, producing thus by interference with the cosmic waves the same oscillatory imbalances of the cells, which at that time, had produced these same diseases.

Fortunately till then, thanks to the oscillating circuits and the new methods that scientific discoveries will put to our disposal without delay, we'll be in a position to fight victoriously against these illnesses provoked by the interference of the stars' radiations.

Figure 2
RED BORDEAUX WINES

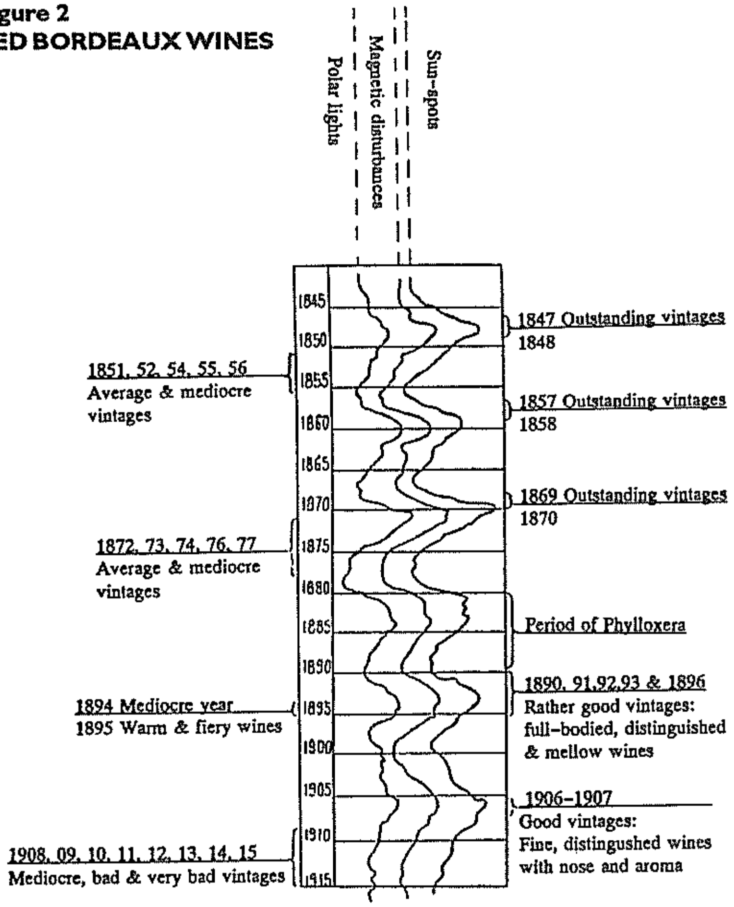

Figure 2 -- Graph showing the coincidence of outstanding vintages with the intensity of the waves emitted by the sun and corresponding to the variation in intensity: 1) of sun-spots, 2) of earth magnetic disturbances, 3) of polar lights. The graph has been established for red Bordeaux wines; the greatest intensities correspond to the outstanding vintages, the lesser intensities correspond to the years of mediocre wines. We can establish a similar graph for the other vintages, notably Bourgogne and Beaujolais wines.

THE RADIO-CELLULAR-OSCILLATOR AND ITS APPLICATIONS

by Georges Lakhovsky

Part 3, Chapter 2, from La Science et le Bonheur, Paris 1930

I demonstrated in the last chapter, how, according to my cellular oscillation theory, alteration in organic tissues came from a variation in the intensity of the cosmic waves' radiations and from these waves' frequency, thus causing the oscillation of the cells.

To reestablish the balance of these natural astral fields, the radiation of which we can no more directly influence, than a meteorologist can influence the weather, I thought of using a local oscillator, generating waves and creating an artificial and auxiliary magnetic field.

Incidentally, I should add that this method is analogous to the methods used in radiotelegraphy, known as heterodyne and super-heterodyne methods. These methods use the interferences produced between the waves received and the waves generated by the local generator at the required frequency and intensity.

About 40 years ago (1890) Professor D'Arsonval advocated the treatment of certain illnesses by high frequency electrical currents, that were at the time produced in waves strongly damped by means of a Ruhmkorff coil and a spark gap (arrester).

My radio-cellular-oscillator is a generator of continuous waves of very high frequencies, the first application of triode lamps to biological research and pathological treatments.

We know the difficulty of obtaining very short waves, because of the parasitic capacitance present in electronic lamps due to the electrodes themselves and, within the bulbs, to tube sockets and connective wires.

In 1924, when I studied the possible construction of a triode lamp oscillator for very short waves, shorter than 2 meters, the radioelectrical technique, in France and abroad, was still at the stage of long-waves and the length of the so-called short-waves was in fact from 200 to 500 meters.

Therefore, I had to do my own research in view of studying a type of lamp appropriate to the generation of very short waves. Thus, I have been led to design lamps with multiple electrodes enclosed in the same bulb. These multiple lamps have been patented in France[1] and in Germany[2].

Later, I intentionally abandoned the idea of working this patent and preferred to give it up to the public domain, in a scientific aim, so that no patent claim could oppose disinterested researches undertaken for the radioelectric prophylaxis of diseases, especially cancer.

However, if the French industry did not engage in the construction of these multiple lamps, the problem was reconsidered in Germany three years later and partly resolved as the double or triple "multi-valve" by Dr. Siegmund Loewe[3].

To reduce connections to a minimum, as well as the parasitic capacity between electrodes, to achieve the generation of waves of less than 1.5 meters while avoiding low efficiency, deformation, or distortion of the wave which would jeopardize the purity of the transmission and particularly alter the modulation, I proposed in my patent to enclose the oscillating circuit and its connections within the bulb. This invention, which cannot conveniently be applied to long waves is, on the contrary, easy to carry out for very short wave circuits, of which the bobbins are reduced to several turns of coil, even to a fraction of a turn, and of which the condensers have a very small capacity (Fig. 3).

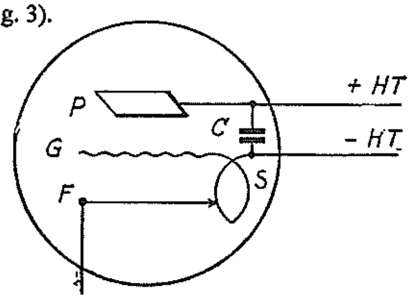

Fig. 3 -- Triode generator of very short waves, after the original drawing of Lakhovsky's patent.

Practically, the oscillating circuit only includes a turn of several centimeters in diameter. The only connections getting out of the tube are the feeders, bringing in the plate voltage (anode voltage) and the heating current (thermo-current).

These tubes can generate oscillations and transmit them by induction to outer circuits (aerials, loops, antennas).

Beside the practical simplification, the improvement of the efficiency comes from the fact that all electronic elements work with the same degree of vacuum. This is what I observed when I grouped within one same glass tubing the bulbs of a 4 triode lamp device (Fig. 4). The equal distribution of the vacuum among the various electrodes increases the purity and stabilizes the functioning.

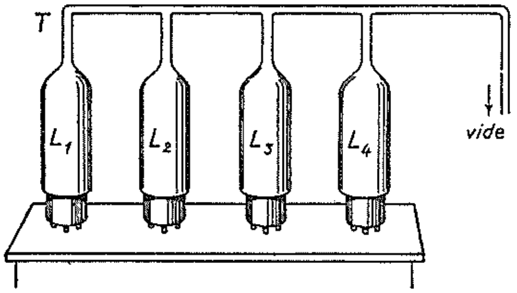

Fig. 4-- Glass tube T helps maintain the same vacuum pressure inside the four triodes of a classic resonance receiver. Author's realization.

According to this principle, I indicated the possibility of constructing not only a certain number of simple or multiple electrical cell oscillators but also all kinds of triodes or electronic lamps contained in a same bulb. Figure 5, quoted from my patent, shows the diagram of a symmetrical oscillator, with double electrodes and filaments connected in parallel. This oscillator is completely enclosed in a same casing.

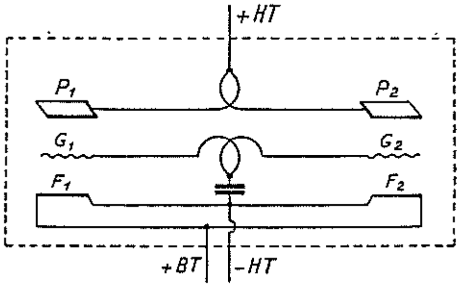

Fig. 5 -- According to the original drawing of the author's patent. Multiple triode including inside the bulb the connections of a symmetrical short-wave transmitter, with filaments set up in parallel.

---

[1] G. Lakhvosky. French patent #601,155 of October 18, 1924

[2] G. Lakvhosky. German patent #427,596 of May 26, 1925

[3] Patent application in France on May 7, 1927

Figure 6, also taken from my patent, presents a double, symmetrical oscillator of which the filaments are connected in series. A possible tuning condenser is anticipated, which would be placed in the same vacuumed bulb as the electrodes and turns of the circuit.

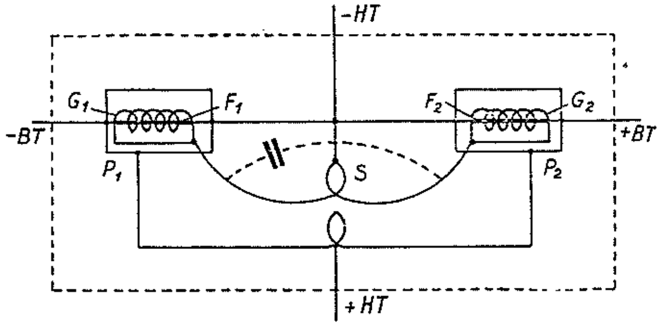

Fig. 6 – Multiple triode symmetrically connected with filaments in series, according to the original drawing of the author's patent.

Figure 7 shows the case of a double oscillator with a unique filament.

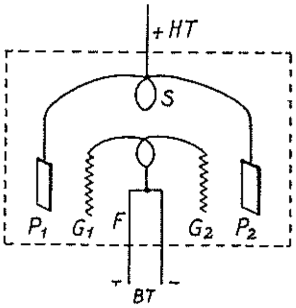

Fig. 7 – Double triode multiple lamp with symmetrical connections and a single filament, according to the original drawing of the author's patent.

For long-waves, moreover, I indicated the possibility of putting together the circuits outside the bulb.

These various constructions are extremely general. We know that for all practical purposes Dr. Loewe has only introduced resistances and capacitances inside his multiple lamps, reducing therefore its use to low and high frequencies amplifications of long waves. These lamps' efficiency is bound to be rather mediocre for the following reasons: great connection resistances, strong plate voltage (about 160 to 200 volts), and impossibility to stabilize the circuits and cause a resonance.

On the contrary, I always anticipated inner couplings through bobbins and capacitances ensuring the greatest efficiency. The resonance is obtained through a variable condenser inside the lamp and commanded from outside without any mechanical element. For example, if you use the gravity field, it suffices to place a counterweight on the sliding condenser and to direct the lamp in space. Or you can use a magnetic field produced by a magnet or an electromagnet. (Fig. 8).

I indicated the means of building an oscillator of good efficiency for ultra-short waves, with the possibility to tune it into a whole range of frequencies.

To resolve this problem on a practical level, before people started to produce multiple lamps, I used a symmetrical oscillator including two triode lamps. It is the radio-cellular-oscillator which I used in 1924, at the wards of the Saltpetriere Hospital in Paris, to treat and heal geraniums inoculated with experimental plant cancer.

Fig. 8 – Variable sliding condenser M, balanced with a counterweight W within a triode lamp, according to the original drawing in author's patent.

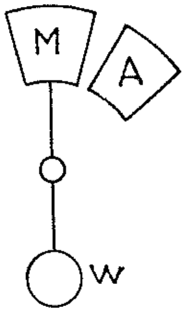

The simplified diagram of the radio-cellular-oscillator is indicated in Figure 9 (Mesny set-up).

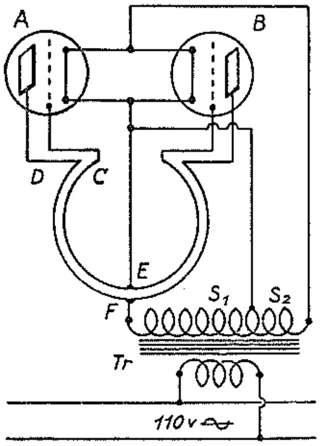

Fig. 9 – Simplified diagram of the Lakhovsky radio-cellular-oscillator.

Plates, grids and filaments parallel connected. The grids and plates circuits all comprise one turn. These two turns are variably coupled between them. Plates and filaments are directly supplied with alternating current of the mains by means of the appropriate transformer. In these conditions, the transmission of waves starts as soon as the plug is connected onto the mains. By using a tuning condenser, one can produce the entire range of waves of a length comprised between 2 and 10 meters. The supply of alternating current creates a great number of harmonics with which the oscillations of the various cells can enter into resonance. The fundamental wave length can moreover be regulated according to the nature of the cells to be treated.

The biological role of the radio-cellular-oscillator is to reestablish the oscillatory equilibrium of the living cell, threatened by any alteration, and in particular by neighboring microbes. When it resumes its normal vigor, thanks to the auxiliary radiation of the radio-cellular-oscillator, the cell can destroy the microbe itself by damping down its pernicious radiations.

The fundamental experiment I presented at the Saltpetriere Hospital demonstrates the functioning of the radio-cellular-oscillator. It dealt with cancerous plants inoculated according to Erwin Smith's method.

This fundamental experiment was the subject, on the 26th of July 1924, of a report to the Societe de Biologie (Society of Biology) and it is reproduced here in extenso, from the society's newsletter.

ATTEMPT AT TREATMENT OF EXPERIMENTAL PLANT CANCER

Report to the Society of Biology

by A. Gosset, A. Gutmann, G. Lakhovsky, J. Magrou

We know that we can produce tumors comparable to animal cancer (Erwin F. Smith)[1] on various plants, by inoculating them with the Bacterium tumefaciens. Using this method[2], one of us obtained experimentally a great number of tumors. These tumors have an indefinite growth; they might partially necrotize but they only die completely when the entire plant or at least the branch infected by the tumor has succumbed to the cachexia. Even when surgically removed, these tumors continued to proliferate.

We intend to study, in this present report, the influence of magnetic waves of great frequencies, obtained by means of a device built by one of us, for therapeutic applications and following his theoretical views[3]: the Georges Lakhovsky Radio-Cellular-Oscillator. This device generates oscillations of a λ wave length = about 2 meters, and corresponds to 150 millions (150,000,000) vibrations per second.

A first plant (Pelargonium zonatum) has been experimented with a month after its inoculation with Bacterium tumefaciens; it showed at that point several, small, white tumors, of the approximate volume of a cherry kernel. The plant has been exposed twice to the radiation, with a 24-hour break and during three hours each time.

In the days following the treatment, the tumor continued to grow rapidly, like the check sample tumors, forming a great plurilobate mass. About sixteen days later, after the first treatment day, the tumor suddenly started to necrotize. A few days later (about 15 days) the necrosis was complete; the lobes of the tumor, retracted and completely dried out, separated from the stalk bearing them, through furrows of elmination, and the tumor could be taken off by the slightest traction. The necrosis effect of the radiation was extremely elective, limited only to the cancerous tissues which it followed into the very depth and root of the tumor; the healthy parts, stalk and leaves remained unhurt and retained all their vigor.

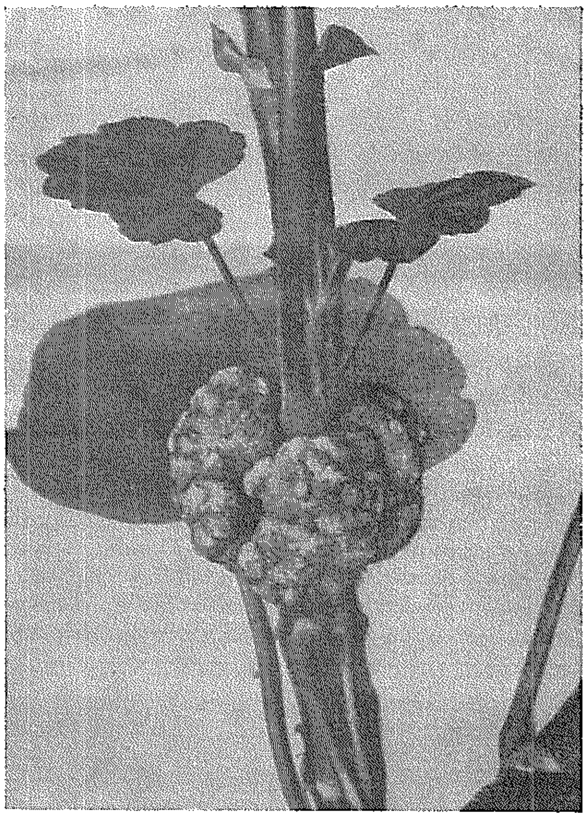

Fig. 10 -- Sight of the tumor of an untreated witness check sample. -- Subject: Pelargonium zonatum inoculated April 10, 1924. We see the growth of the cancerous tumor on the stalk. *(Clinique Chirurgicale de la Saltpetriere).*

A second Pelargonium was treated in the like manner, but the time of exposure to the radiations was lengthened (eleven 3-hour sessions); 16 days after the first session, the tumor it carried started to necrotize and several days later it dried out completely. As in the first case, the healthy part remained unhurt.

With a third Pelargonium which underwent nine hours of radiation (three 3-hour sessions) the necrosis of the tumor's lobes followed the same path.

Sixteen control Pelargoniums remained untreated. All bore tumors in full activity which often became enormous.

As a summary, we may say that the Pelargonium which became cancerous after inoculation of the bacterium tumefaciens and for which a surgical intervention could not prevent recurrence seemed to heal under the influence of the magnetic waves indicated above. *(Clinique Chirurgicale de la Saltpetriere.)*

A very clear conclusion can be drawn from this experiment and its counterpart.

On the one side, we observe that, on the control subjects, inoculated with the *Bacterium tumefaciens* and left untreated, enormous tumors grew which by absorbing the plants' vital energy lead them to a quick and certain death.

On the other hand, the plants treated by the radio-cellular-oscillator, and moreover taken at random from among the inoculated subjects, have been healed quickly and acquired a considerable vitality, as is proven by their growth, their continual development and abundant flowering.

We can see it in the photograph (Figure 12), taken one year after the healing of the geranium.

My theory of cellular oscillation, that I had moreover proposed before I realized this experiment, enables us to understand cancer pathology and to direct its therapy onto a new path.

Cancer, mostly reaching older people, comes from the ageing of tissues as a result of the cells and blood decay. This impairing of the cells results from a modification of the

---

[1] Erwin F. Smith, *An Introduction to Bacterial Diseases of Plants*, 1st vol., Philadelphia and London 1920.

[2] J. Magrou, *Revue de pathologie comparee*, March 1924. Two other reports will be soon published in the *Revue de pathologie vegetale et d'entolmologie agricole* and in the *Annales de l'Institut Pasteur.*

[3] Georges Lakhovsky, *Radio-revue*, November 1923 and Conference to *L'Ecole Superieure des P.T.T.*, June 2nd, 1924.

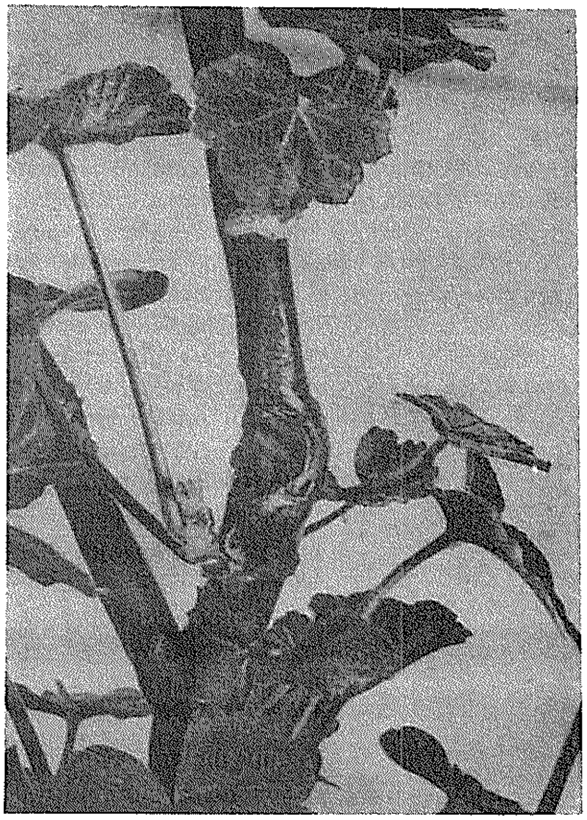

Fig. 11 Sight of the scar of the treated subject. Subject: Pelargonium zonatum, inoculated April 10, 1924 with the Bacterium tumefaciens, treated from May 24, to June 14, 1924, during eleven three-hour sessions with the Lakhovsky oscillator with antennas, photographed after it healed on July 24, 1924. *(Clinique Chirurgicale de la Saltpetriere).*

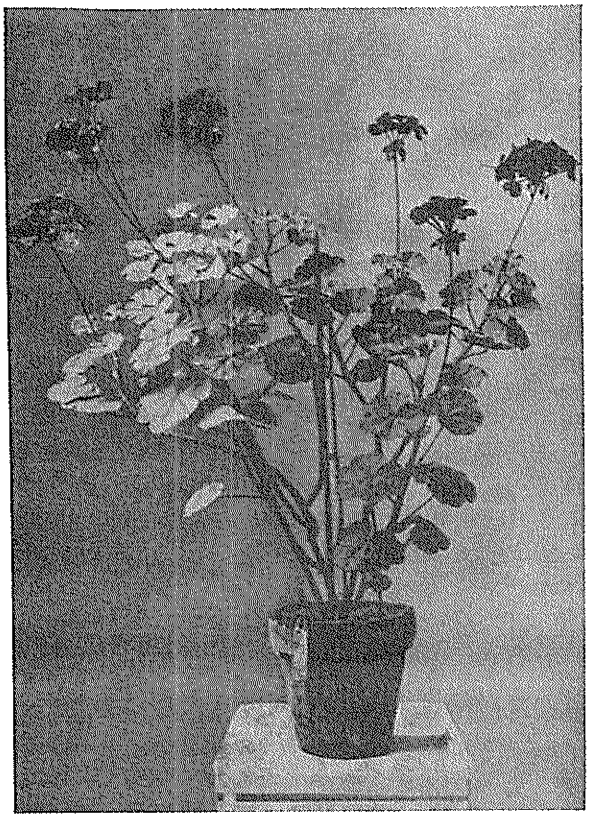

Figure 12 Sight of the treated subject after healing. -- This pelargonium is no other than the one of Figure 10, treated by the Lakhovsky oscillator, on May 24, 1924, healed on the 4th of June, 1924, and photographed in July 1925. As you can see, this plant is in perfect health and abundantly flowered. On the other hand, the sixteen inoculated but untreated check samples died long ago.

chemical composition due to the contribution of foreign bodies which appear as residues of the organism. Such are, for instance, blood haematoblasts, rich in iron and phosphorus, produced by residues of fibrin, leucocytes, and hematites. Each cell destruction brings with it a small mass of cholesterol, an insulating substance forming the armature of the nuclear filament, which later clogs the nucleus. Therefore older people first experience a decrease in the number of active cells and of the white and red corpuscles, which is reflected in the regression of the blood count, and on the other hand, an increase of the quantity of mineral (iron, phosphorus, haematoblasts) and organic (cholesterol) residues. These chemical modifications of the cell result in the variation of the cellular oscillation both in intensity, as a consequence of the changes of electric conductivity and resistance, and in the radiation frequency, because of the changes of capacitance and inductance.

The decayed cell appears like a pendulum with a pendulum-bob one could have lengthened, shortened or damaged. This oscillatory disorder provokes, because of the excess of cosmic waves, a too rapid cellular division, which results in the formation of neoplasms, or cells vibrating at a different frequency from the regular frequency of the tissue.

For long years, it has been wondered whether cancer is contagious and hereditary, or whether it is caused by a microbe. Histological analyses seem to invalidate the existence of a microbe. On the other hand, inoculations of cancer only succeed, and moreover very rarely, with impaired tissues composed of abnormal cells, such as certain moles and this would seen to ruin the hypothesis of contagion and heredity.

In fact, I believe that the true cancer microbe is the neoplastic cell which forces the neighboring cells into oscillating at its own frequency, thus transforming by electromagnetic induction, the healthy cells into neoplasms. To prevent cancer, is to prevent this neoplastic induction and to give back the healthy cell enough vigor to reestablish its oscillatory equilibrium.

I was pleased to observe that, apart from the skepticism of a few, my experiments have been taken up again with great success in France and abroad by various scientists and practitioners who understood the future of oscillation therapeutics.

It seems appropriate to quote the works pursued in Germany by Professor Esau at the Technisch-Psychaliches Institut of Iena.

Although directed toward paths sometimes very different from mine, these works present a very clear analogy with those I have conducted from the use of a very short wave oscillator similar to my radio-cellular-oscillator.

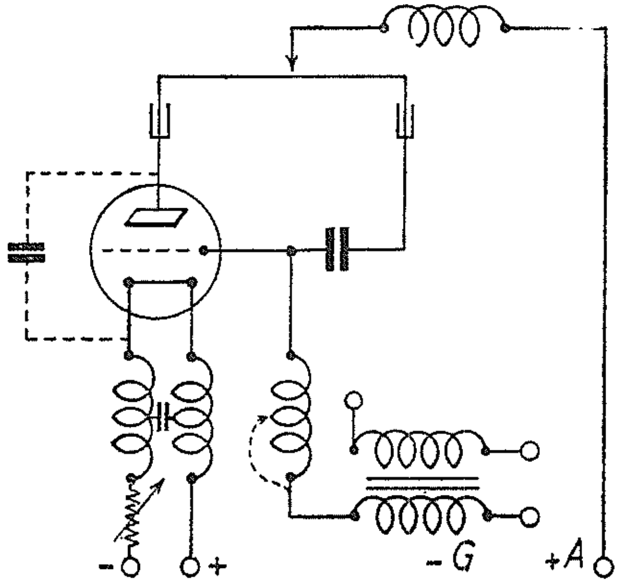

Fig. 13 -- Simplified diagram of Dr. Esau's very short wave oscillator.

The transmitter-receivers used by Professor Esau do not essentially differ from the devices I advocated. I indicated moreover that connections were irrelevant as long as they produced the desired very short waves. Professor Esau's connections, which were described in the technical magazines of Dr. Ernest Busse, assistant at the Ieana Institute[4], are characterized by the use of a single triode lamp, functioning in direct current. The transmitter's filament (Fig. 13) is supplied through a filter preventing high frequency back currents. The coilings are simple fractions of turns, as in the radio-cellular-oscillator, which only includes about half a turn for each oscillating circuit.

A test receiver (Fig. 14) is provided with square loop circuits on which variable plugs can be installed. Another device is provided with a circular turn, of which a radial cursor allows connection to the necessary position.

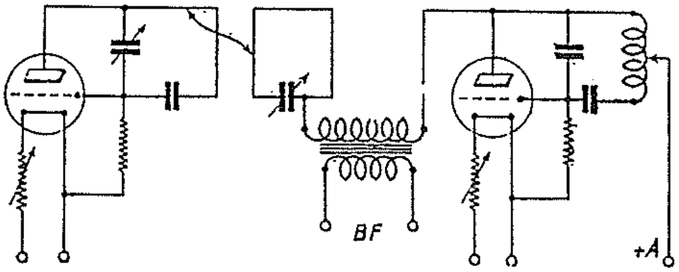

Fig. 14 -- Simplified diagram of Dr. Esau's very short wave receiver.

The works of Professor Esau in the biological domain are rather recent, for during the Radio-electric Symposium in Bremen (August 1928)[5], he stated that he had been doing research, for only a year, on the physiological effects of *ultra-short* waves, or, according to his definition, very high frequency waves comprised between a length of 1.50 and 3 meters, of which I indicated the use in 1924 in my report addressed to the Societe de Biologie, reproduced here. He indicated then that since July 1928, he successfully constructed a very short wave receiver.

About this subject Professor Esau pointed to the fact that the main difficulties came from the electronic lamps, and that he had to use foreign lamps, particularly French lamps, which are more appropriate for very short waves techniques. I indicated earlier that, 4 years before, I had studied and had these special short-wave lamps built, which were designed for my radio-cellular-oscillator.

I shall also point out that a very short transmitter has recently been used by Dr. Frantz Seidel from Vienna, for the destruction of bacteria, particularly in the sterilization of milk.[6] This subject will be investigated at a later time.

---

[4] Dr. Ernest Busse, *Die Erzeugung ultrakurzer Wellen.* (*Radio-Welt*, No. 49, 1928, p. 77. Vienna.)

[5] Oscar Gadamer. *The ultra-short waves and their field of application.* (*Der Deutsche Rundfunk*, August 31, 1928, t. VI, p. 240, Berlin.)

[6] *Les ondes ultra-courte contre les bacteries.* (*Radio-Welt*, January 27, 1929, No. 4, p. 108, Vienna.)

THE LAKHOVSKY RADIO-CELLULAR OSCILLATOR

a synopsis by Jorge Resines

What you are about to read is my resume of the book La Science et le Bonheur, Longevite et Immortalite par les vibrations -- *Science and Good Health, Longevity and Immortality Through Vibrations* -- by Georges Lakhovsky. More specifically, this synopsis encompasses Part III, chapters 1 and 2, being pages from 137-171 of said work, and whatever I translate from it will be forwarded by the words "Lakhovsky said:".

Mr. Lakhovsky begins by indicating the influence of heavenly bodies upon all facets of everyday life, from their influence upon the cells up to how magnetic compasses are affected by magnetic storms originating in the Sun (he refers the readership to his book L'Universion, published by Gauthier-Villars of Paris, France). He says that the French Astronomer Mr. Deslandres with his collaborators at the Observatory of Meudon have established a cycle of about 27 days for sunspots affecting both tides and earthquakes. This lapse is part of another larger cycle of 11½ years (or about) that most usually forms in turn part of another, larger, cycle of 33 to 35 years.

Mr. Lakhovsky also discovered, as indicated by the small circles at the lunar orbit of Figure 1, that lunar radiations provoke interferences with the waves broadcasted from radioelectric stations at said marked positions.

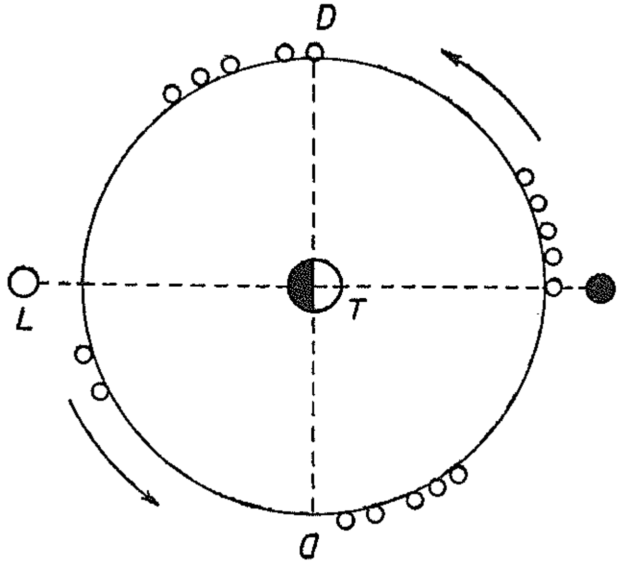

Figure 1

T = Earth
L1 = Full Moon
L2 = Waning Moon
L3 = New (Dark) Moon
L4 = Waxing Moon

Carrying his researches a step further, helped also by the people at the Observatory of Meudon, he correlated the highs-and-lows of three different factors (to wit: a. Sunspot activity, b. Perturbations of Terrestrial Magnetism, and c. Polar Aurora) with the type of wines collected at different French regions. For a single zone and type of wine, his findings are shown in Figure 2 (next page). You will easily notice how the highs of the three factors are in coincidence with the best collections of wines, with a single exception: the 1880-1889 period when the grape-wines were massively attacked at France by the "Phylloxera" plague. Though Mr. Lakhovsky does not explain the reason for this adverse period (when the opposite was expected), he does voice his opinion that cellular and biological phenomena is periodically affected by the same type of radiations (and the generated-sicknesses they produce), proposing at the same time to destroy the scourges by overriding the "astrally-produced radiations" (his words) with this "Radio-Cellular Oscillator."

About the basic concept behind his devices, Lakhovsky said: "To re-establish the balance of these heavenly fields from nature, upon which variations we are able to act directly, unlike Meteorology which cannot make weather change, I will resort to a local oscillating-generator of waves creating an auxiliary artificial electromagnetic field.

"By the way, I must add this method is analogous to that employed in radiotelegraphy and known under the names of heterodyning and superheterodyning. These procedures employ the interferences generated between the received and generated waves by a local generator, at the necessary frequency and intensity."

Mr. Lakhovsky states that his inspiration came from Professor Arsene D'Arsonval, who in 1890 had put forth the opinion that different sicknesses could be eliminated by means of high-frequency currents. He continues on, explaining that by about 1924 he studied the possibility of generating waves under 2 meters in length, but that the electronics of said period considered "short" only those waves from 200 to 500 meters.

To overcome such a hindrance, Lakhovsky developed a vacuum tube with multiple-electrodes and received two patents for it -- French Patent No. 601,155 granted in October 18, 1924 and German Patent No. 427,596 granted in May 26th, 1925 -- afterwards another scientist of that time, Dr. Siegmund Loewe, improved on his ideas and demanded a French Patent on May 07, 1927 (this Patent is not indicated as either granted or not; further investigation is needed here).

Figure 3

+HT = Positive High Voltage
-HT = Negative High Voltage
G = Grid
P = Plate
C = Condenser
S = Loop of Coil
F = Filament

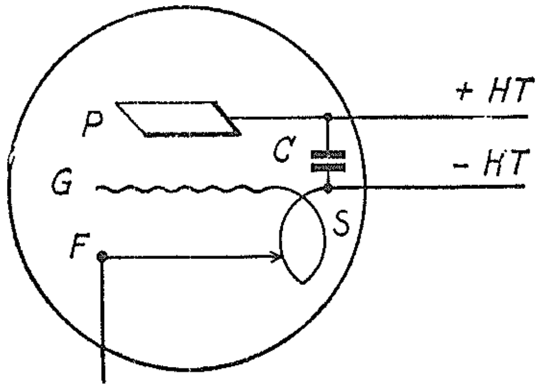

Figure 2 -- RED BORDEAUX WINES

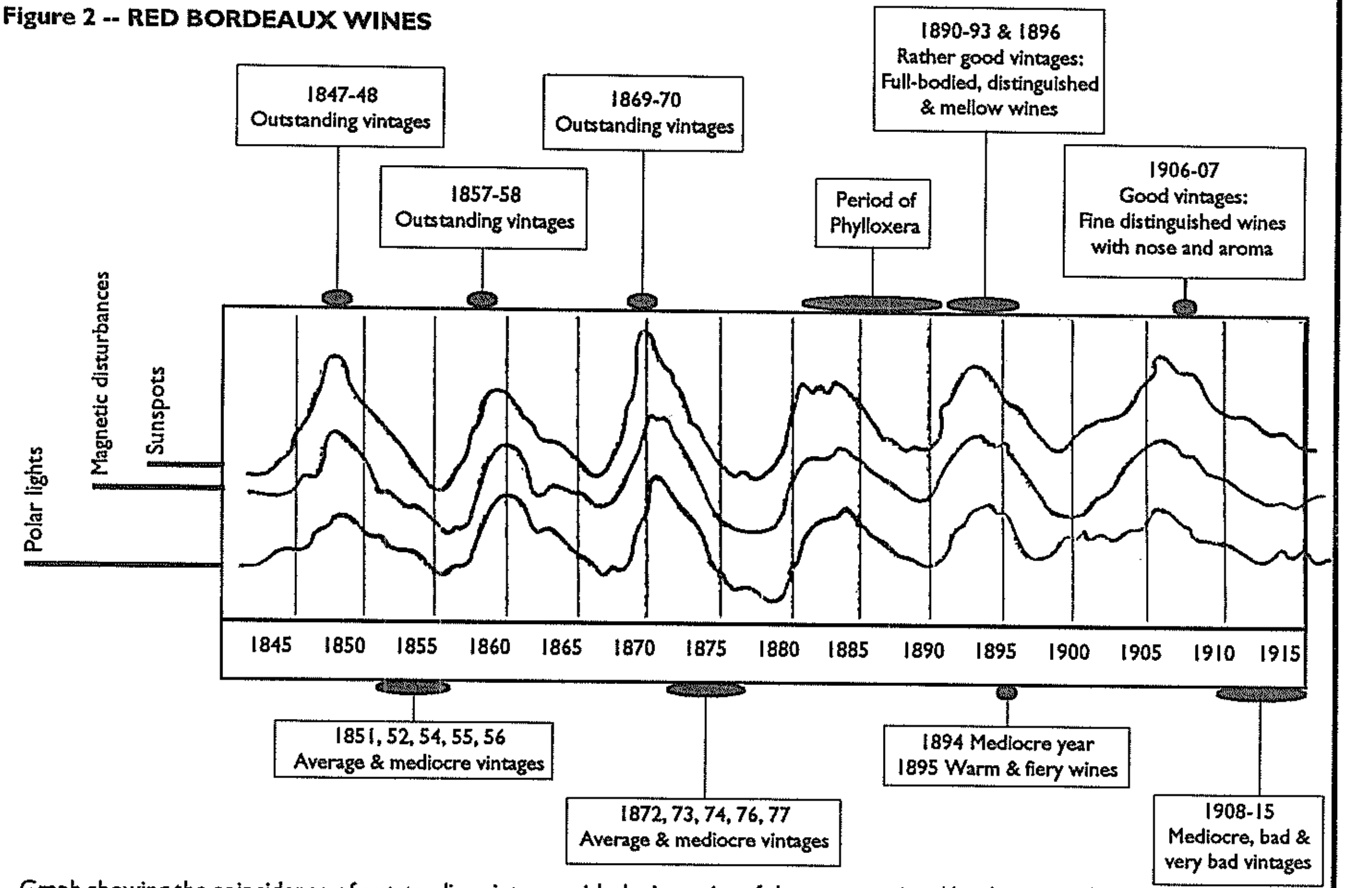

Graph showing the coincidence of outstanding vintages with the intensity of the waves emitted by the sun and corresponding to the variation in the intensity: 1) of sun-spots, 2) of earth magnetic disturbances, 3) of polar lights. The graph has been established for red Bordeaux wines; the greatest intensities correspond to the outstanding vintages, the lesser intensities correspond to the years of mediocre wines. We can establish a similar graph for the other vintages, notably Bourgogne and Beaujolais wines.

To produce waves "under 1½ meters" (his words), Mr. Lakhovsky enclosed within a vacuum tube's glass vessel different elements, as is seen in Figure 3. About this new type of tube, Lakhovsky said: "The oscillating circuit does not involve but a loop with a diameter of several centimeters. The only connections going out of the tube are the supply-wires, leading the plate-voltage and the heating current. These tubes can generate oscillations and transmit them to external circuits (square-loop antennas) by induction."

I will stop now the synopsis to introduce a personal comment: It is not clear on the illustration of Figure 3, WHY the filament is connected -- in a controllable fashion as the schematic indicates -- to the loop "S"; neither do I see clearly indicated whether the tube is one of direct or indirect heating and which type of vacuum tube -- existing at the time, with its specifications of that era -- served as the foundation to develop the tube of Figure 3, and how this was made in both constituents, procedure of manufacture, etc.

Neither is it indicated if there was a constitutional modification involving a change of material(s) for the new type of tube Mr. Lakhovsky first employed. I do not want to spoil anybody's fun in experimentation, I am just suggesting caution... Let's go on with Mr. Lakhovsky.

Another feature employed to ensure a high-quality when in use by Mr. Lakhovsky was to put a group of electronic tubes, as illustrated in Figure 4, directly over these lines, in the same vacuum pump. As all the tubes were connected in series, when the apparatus was made to work it produced the same degree of vacuum within them all.

Figure 4

L_1, L_2, L_3, L_4 = Identical Vacuum Tubes
T = joining tube

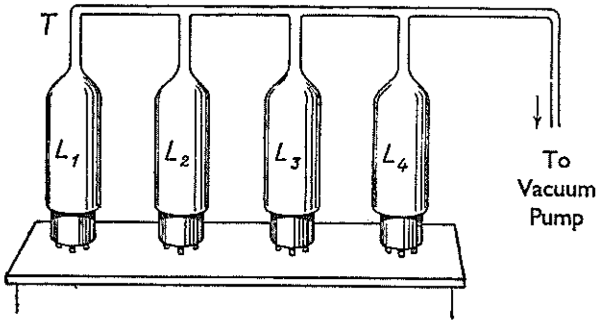

A further development of his idea, "extracted from his patent" he claims without indicating which one (French or German) this is, is the multiple triode illustrated in Figure 5. From its analysis we notice that Mr. Lakhovsky adhered to electronic theory, which says that electricity is a flow of electrons, and in a semiconducting device as the usual vacuum tube there is only a single direction of flow, from the "Cathode" (or negative pole) to the "Anode" (or positive pole). In the multiple triode of Figure 5, the filaments are connected in parallel.

Figure 5

+BT = Positive Filament Voltage
+HT = Positive High Voltage
-HT = Negative High Voltage
P_1 = Plate Tube No. 1
G_1 = Grid Tube No. 1
F_1 = Filament Tube No. 1
P_2, G_2, F_2 = Same for Tube #2
C = Condenser

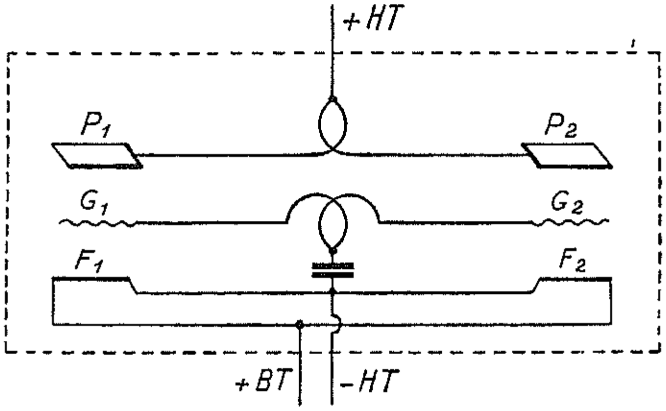

A further improvement of the device is seen in Figure 6, of which Lakhovsky says: "This illustration, also taken from my patent, shows a double and symmetrical oscillator where the filaments are mounted in series. It is also provided for, eventually, to place within a condenser so located as the electrodes and the loops of the oscillating circuit, within the same evacuated glass-vessel".

Figure 6

P_1 = Plate Tube No. 1
G_1 = Grid Tube No. 1
F_1 = Filament Tube No. 1
P_2, G_2, F_2, = Same for Tube #2
C = Condenser
-HT = Negative High Voltage
+HT = Positive High Voltage
-BT = Negative Filament Voltage
+BT = Positive Filament Voltage

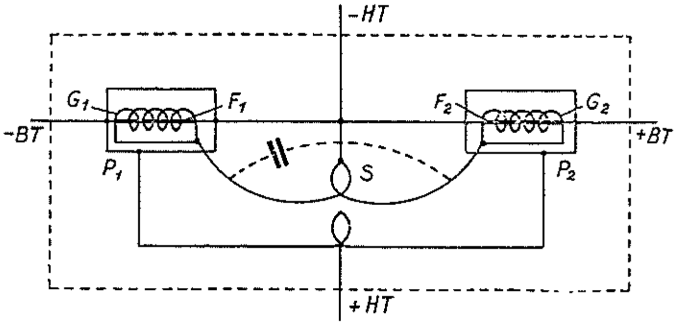

Figure 7 shows us a double-oscillator with a single filament, that is the two lamps of Figure 6 within a single vessel and with just one heat-source.

Figure 7

+HT = Positive High Voltage
+/- BT = Positive and Negative Filament Voltage
F = Filament
S = Loop of Coil
G_1 = Grid No. 1
P_1 = Plate No. 1
G_2 = Grid No. 2
P_2 = Plate No. 2

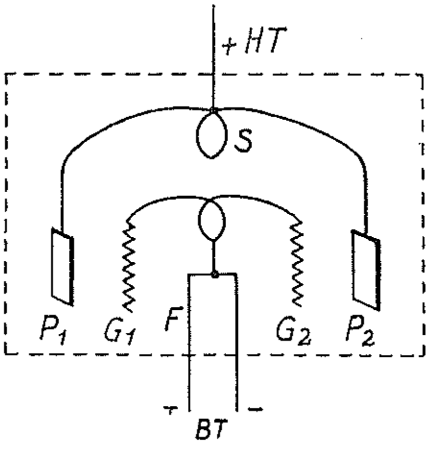

I remind the readership that all the models of vacuum tubes shown so far are for short waves under 2 meters. Lakhovsky indicated that to obtain long waves it sufficed to place the same elements outside a vacuum tube.

After saying that everything so far shown were just "extremely general things" Lakhovsky said: "One can, almost, say that Dr. Loewe just introduced resistances and capacities within his multiple-tubes, where performance is just reduced to amplifying low and high frequencies upon long waves. The performance of these tubes is very mediocre because of the following reasons: a) Heavy linking resistances, b) Strong plate-voltages (about 160 to 200 volts), c) impossibility of coupling circuits and thus producing a resonance."

Figure 8

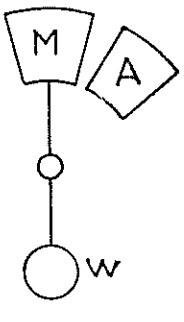

M = Mobile blade of condenser
A = Fixed blade of condenser
W = Weight to move blade "M"
H = Turning point for the combination of "M" with "W"

Here, in the former comment, is an apparent contradiction: Lakhovsky criticized Dr. Loewe for including resistances and capacities within his multiple-tubes, while at the

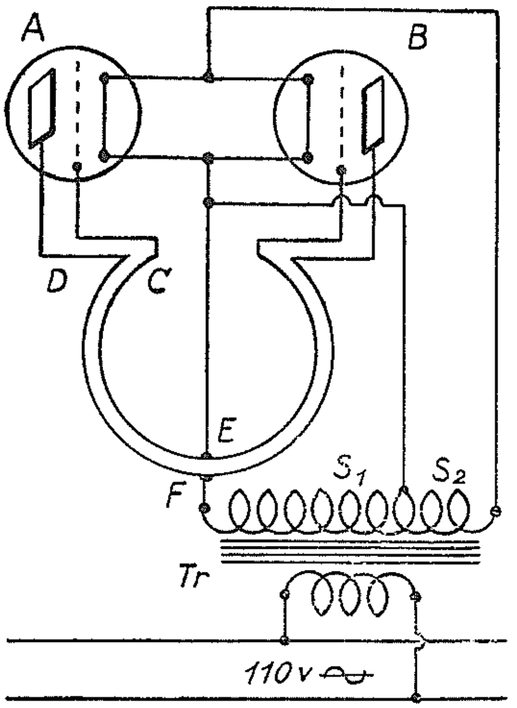

Figure 9
A = a Lakhovsky Triode
B = another Lakhovsky Triode
C = Loop associated to Grids
D = Loop associated to Plates
E = Connection to Filament
F = Connection to High Voltage of Transformer secondary tap S_1
S_1 = High Voltage tap of AC Transformer secondary
S_2 = Filament Voltage tap of AC Transformer secondary
Tr= AC step-up Tranformer with a tapped secondary winding

same time he did the same! But Lakhovsky's capacitors are different to the common types.

What he did was to "separate" a variable condenser's blades and to "counterweight" the moving blades, as indicated in Figure 8, so that the blades would move within the fixed armature of the condenser as the whole tube was rotated and weight "W" responded to gravity! He does not indicate, thus making it an object of personal experimentation on the part of the readership, if there was just one or more of these special variable condensers.

Now, in order to produce a wide range of frequencies with only two of his special triodes, Lakhovsky built the "Radio-Cellular Oscillator" that is shown in schematic form in Figure 9, which he used at the Clinic of Salpetriere in 1924 to treat and heal Geranium plants which he experimentally inoculated with the disease of cancer; it was mounted, he says, according to the "Montage Mesny," thus indicating that one man called Mesny originated the setup at that time (this is another source for research, try finding in old books from either the '20s or '30s anything made by a Mr. Mesny).

This is what, on the device, Lakhovsky says: "Plates, grids and filaments are connected in parallel. The circuits of grid and plate have each a single loop. These two loops are coupled variably between themselves. Plates and filaments are fed directly from the Alternating Current of the place by means of the appropriated transformer. Within these conditions, it is produced the emission of waves because it is branched the intake of current upon the electrified section.

"Using, accordingly, a condenser it is possible to produce the whole spectrum of waves between 2 and 10 meters of wavelength. The supply from AC produces a large number of harmonics upon which can enter into resonance the oscillations of different cells. Therefore, the fundamental wavelength can be regulated according to the nature of those cells to treat.

"The biological role of the Radio-Cellular Oscillator is to re-establish the oscillatory equilibrium in the living cell, under threat by any alteration, particularly by a close microbe. When recovering its normal strength thanks to the auxiliary irradiation from the Radio-Cellular Oscillator, the cell can all-by-itself destroy the microbe by counter-acting its deleterious radiation."

Mr. Lakhovsky goes on indicating that he established the aforementioned concept as the foundation of his device by first inoculating several plants of *Pelargonium zonatum* with the *Bacterium tumefaciens* in order to generate tumors similar to those of animals in plants (his source was: An Introduction to Bacterial Diseases of Plants by Erwin F. Smith, London and Philadelphia, 1920). He was helped with inoculating the plants by Mr. A. Gosset, Mr. A. Gutmann and Mr. J. Magrou, the latter being the one who had more articles published on the subject of plant-inoculation (see the Revue de Pathol. Vegetale et d'entomologie agricole and within that the Annales de L'Institut Pasteur, by this author).

He then applied his Radio-Cellular Oscillator, using waves of about 2 meters, which he equated with a vibration of 150,000,000 cycles per common-second. He broadcasted the oscillator's output upon the infected plants and got them all back to health after a given time. He had already explained how his oscillator worked to his fellow Frenchmen in an article (Radio-Revue, November 1923) and a conference (given at the Ecole Superieur des Postes, Telegraphes and Telephones

Figure 10

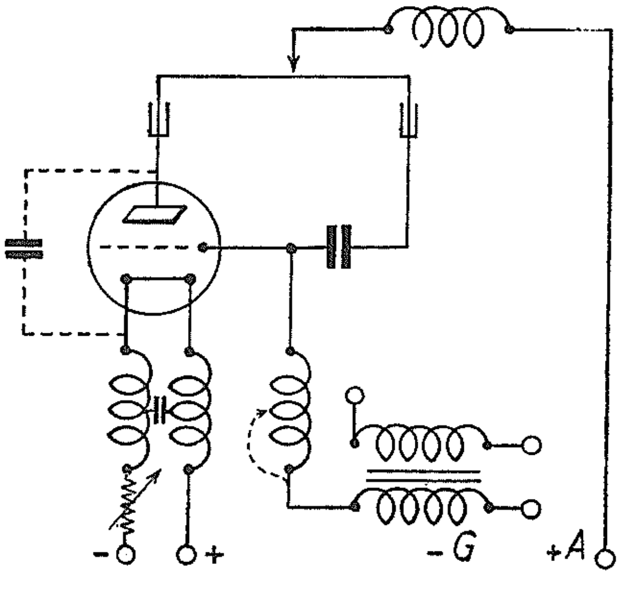

on June 2nd, 1924).

After detailing how the experiment in plant-healing progressed and how the plants were healed, he goes on to explain that -- in his conception -- cancer is provoked by an "aging" of the cells that makes them both reproduce at a slower rate and to evacuate a higher amount of wastes. This, in combination with radiations coming from space at different times, produces an oscillatory disequilibrium in the cells resulting in cancer of different types. Mr. Lakhovsky rejects, on the ground of histological examinations he performed, both the "microbial" theory of cancer and its hereditary transmission.

If you have followed all of the writing up to this point, carefully noting the concepts involved in Lakhovsky's procedures, you will remember that:

(1) His Radio-Cellular Oscillator created an auxiliary electromagnetic field capable of restoring balance to the diseased cell(s) and body.

(2) This was done by means of very short waves (during the period he experimented) of about 2 meters and 150,000,000 cycles per second.

(3) To achieve the necessary shortness of wave he introduced different elements within an evacuated glass-vessel, thus creating a different type of vacuum tube.

(4) He purposely indicated that external elements created long waves that did NOT comply with what he needed for therapy.

(5) He included, within his special tubes, a variable condenser that could be tuned to the desired capacity by moving the tube in space so that the internal variable capacity could generate the coupling needed in therapy.

If all of this has been, so far, an essential group of conditions necessary to produce the short waves of about 2 meters, then why -- as seen in Figure 9 -- does he place two electrified loops OUTSIDE the vacuum tubes of his special design????

In compliance with this "loops outside" line, he goes on quoting the works by one of many foreign scientists who, he claims, have followed his line of research:

Lakhovsky says: "It is convenient to quote the works made by Professor Esau at Jena's Technisch-Physicalishes Institut (Germany). If they are directed by ways somewhat different to mine, his works show the similitude of using an oscillator of very short waves similar to my Radio-Cellular Oscillator. Professor Esau's emitters and receivers do not differ essentially from the apparatus I champion. I have already indicated that the arrangements of constituents are indifferent, because they produce the very short waves needed.

"The setups by Professor Esau -- described in the Institute's technical magazines by Dr. Ernest Busse (see Die Erzeugung Ultakurzer Wellen -- *Research on Ultrashort Waves* -- in Radio Welt, No. 49, page 77, Vienna, 1928), secretary to the Institute of Jena, are characteristic because they employ a single triode working with DC.

"The emitter's filament -- as seen in Figure 10 -- is fed by a filter preventing the return of high-frequency currents. Windings are made with simple fractions of turns, as in the Radio-Cellular Oscillator, which does not involve but a half-turn per oscillating circuit.

"An experimental receiver (see Figure 11) is supplied with a circuit possessing a single square-loop, upon which can be made variable taps. Another apparatus is possessed of a single circular-loop upon which a radial cursor allows to take just the necessary fraction of circuit.

"Works by Professor Esau in the realm of Biology are very recent, because at the August 1928 Radioelectric Con-

Figure 11

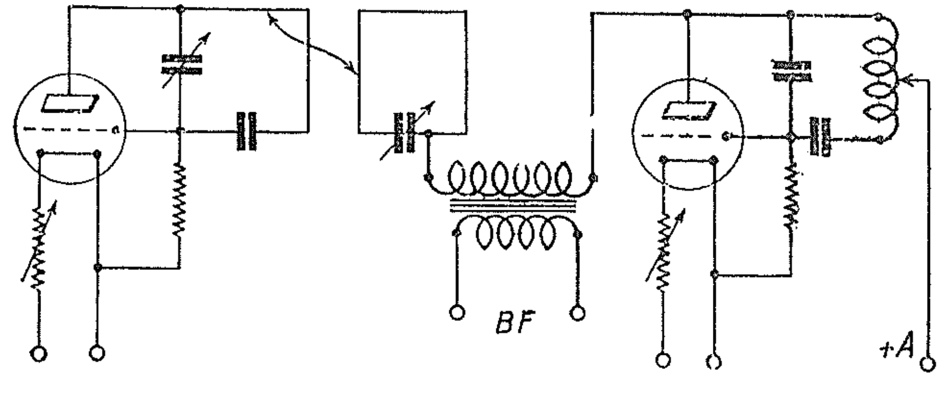

gress of Bremen (see Les ondes ultra-courtes et leur champ d'application -- *Ultrashort-waves and Their Field of Application* -- by Oscar Gadamer in Der Deutsche Rundfunk -- *German Radio* -- of August 31st, 1928, volume VI, page 240, Berlin) this author states that after about a year he investigated the physiological effects of ultrashort-waves that is, according to his definition, waves of very high frequencies between 1½ and 3 meters in length, of which I have indicated how I employ them in a 1924 communication to the Society of Biology (Note: It is on the inoculation and healing of the *Pelargonium zonatum*, as explained on page four of this writing). Prof. Esau then indicated that, since July 1928, he built a receiver of very short waves.

"About this subject, Professor Esau remarked that the main difficulties come from electronic tubes, for he was forced to use foreign tubes, particularly French ones, that are better fitted to the technology of very short waves. I must remind (the readership) that four years before him I studied and had built these special lamps for short waves, to be used in my Radio-Cellular Oscillator.

"I will also point out that an emitter of very short waves has recently been employed by Dr. Franz Seidel of Vienna for destroying bacteria, particularly to sterilize milk (see Les ondes ultra-courtes contre les bacteries -- *Ultra-short Waves Against Bacteria* -- in Radio Welt -- *Radio World*, January 27th, 1929, No. 4, page 108, Vienna)."

We have two definitely contradictory positions in the same chapter dealing with the same type of device:

(1st) In the paragraph translated under Figure 3 of this paper, Lakhovksy emphasized that the elements to produce ultrashort waves (under 2 meters) must be inside the tube. Figures 3 to 8 are dedicated to showing specific constituents for this special type of tube. Though he only speaks on Triodes, there is no reason to suppose he did not experiment with Tetrode and Pentode tubes (plus others of his own design not usually covered in publicly available technical literature).

(2nd) The formerly-indicated conditions change radically when he speaks of his experimental device employed to eliminate tumors affecting inoculated plants, this being the device of Figure 9 of this paper. Notice how the emitting elements in this apparatus are outside the vacuum tube; furthermore Mr. Lakhovsky claims that other, similar, devices developed by Professor Esau at Jena's Technisch-Physicalysches Institut are based in his "internal" design when the opposite is quite true!

(3rd) In order to make a better synopsis, I went back to read my copy of The Secret of Life, a translation into Spanish by M. Aquilar (Madrid, 1929). Therein are explained the basics of the Multi-Wave Oscillator but the Radio-Cellular Oscillator is NOT shown. It is reproduced in the 1924 communication to the Society of Biology. Neither source mentioned here reproduced the special vacuum tubes, with internal constituents, as explained in La Science et le Bonheur, the subject of this synopsis.

(4th) In Mark Clement's The Waves that Heal, the Multi-Wave Oscillator is explained as: "The apparatus consists of a transmitter and a receiving resonator, both arranged so as to set up an electromagnetic field in their immediate vicinity." The Multi-Wave Oscillator seems, in my opinion, derived from the "experimental receiver" of Figure 11 plus those of Heinrich Hertz (as I have indicated in my Some Free Energy Devices, a BSRF publication, quoting directly from The Secret of Life and introducing complementary information from other sources), rather than from the multiple-tubes shown in Figures 3 through 8 of this paper.

(5th) Considering everything explained from (1st) to (4th), it is my opinion Mr. Lakhovsky produced several types of apparati, along different lines of research, while pretending in his books that only a single-line of investigation was being pursued.

What are the reasons for doing this? There are many:

(*) Trying to disorient his fellow scientists by covering up his actual intentions and actions, so that the actual secret of the apparati rested with him.

(*) Trying to foil any attempt by the "organised medical profession" to steal his "secret" -- whatever this could actually be.

(*) Having discovered something that he wanted to keep for himself in the course of his researches. This is, in my opinion, the most likely possibility, for Mr. Lakhovsky's researches were well-received in Europe, South America and Asia before WWII began. Even more, he was outspokenly backed by men of science and Academies of great reputation and his enemies were very few (outside of the USA, of course....).

I do not mean to slander him and his researches, I just want to indicate the necessity for greater and deeper research into what he did from all possible sources of information.

VERY IMPORTANT NOTE

In order to facilitate the understanding of the illustrations, I have taken a few liberties to aid the non-French speaking associates:

(1) All the "lateral letterings" from Figure 3 to Figure 9 are of my doing, they did not exist in the original text.

(2) I changed into English, keeping the true meaning in translation, the writing in Figure 2.

(3) I made some additional lettering in some illustrations for greater understanding: L1 to L4 in Figure 1; I added the "C" in Figure 5, the same "C" was added in Figure 6; Finally, I added the "H" in Figure 8 for greater clarity.

(4) In French the letters "HT" are the abbreviation of "Haute Tension" that means in English "High Voltage"; in Spanish we also use the word "Tension" in electricity meaning "Voltage," as the French do. The letters "BT" stand for "Basse Tension" which I translate not as "Low Tension" but as "Filament Voltage" for this is the actual meaning.

UNITED STATES PATENT OFFICE

1,962,565

APPARATUS WITH CIRCUITS OSCILLATING UNDER MULTIPLE WAVE LENGTHS

Application November. 13, 1931, Serial No. 574,907

In France, May 2, 1931

7 Claims. (Cl. 250-23)

Fig. 1.

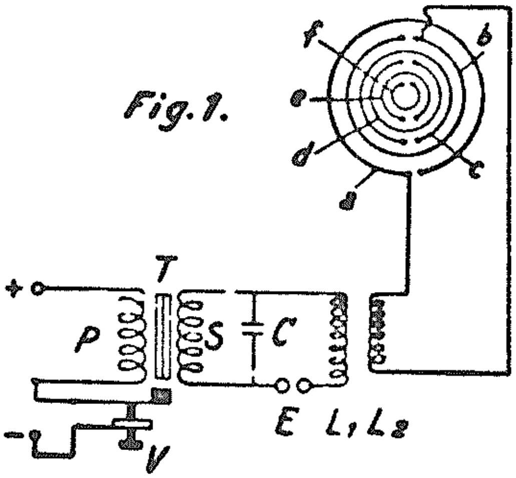

Fig. 2.

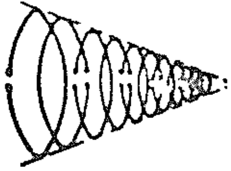

Fig. 3.

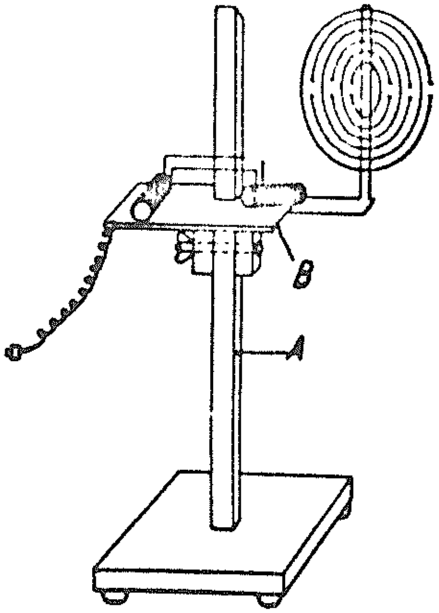

The present invention has for its object an apparatus capable of sending out simultaneously different wave lengths so that among these waves or their harmonics there is found almost always one or several capable of producing the best effect sought for any application whatever.

There is used for that purpose the property of open circuits having well determined wave length if they are excited by electric impulses emanating from discharges of any source.

The radiating part of the system, producing waves of various lengths, is therefore constituted by a certain number of concentric open rings of different diameters, which are maintained insulated by any appropriate means. These rings may or may not be terminated by small spheres forming capacities.

These rings may be arranged either in the same plane or in different planes so as to form any surfaces, such as portions of cones, spheres, paraboloids, ellipsoids, etc..., these surfaces being of a nature to concentrate the waves in a determined direction in the manner of a reflector, for example.

These rings may be fed by any producing device whatever of high frequency, for example, a combination comprising a trembler coil (or any other transformer) and a high frequency circuit provided with a self-induction coil and a capacity.

One of several points of this high frequency circuit is or are connected to the end of one or several of the rings forming the radiating apparatus. The other rings not connected are excited by induction.

The complete apparatus (coil, high frequency circuits, rings, etc....), may be mounted in a box or upon a small board capable of sliding along a guide or rocking so as to be able to be placed in the most convenient position.

In the accompanying drawing, by way of example, not limitative, there is shown:

Fig. 1 a diagram of the system.

Fig. 2 the position of the rings in the form of a cone.

Fig. 3 the apparatus mounted upon a foot or base.

In Fig. 1 there is seen at T the transformer or coil (P the primary, S the secondary of this coil), V the screw of the trembler, C the condenser the discharge of which across the spark gap E produces oscillations in the self-induction coil L1, which in its turn acts by induction upon the self-induction coil L1; this self-induction coil L2 is connected on one side to the first exterior ring a, on the other side to the following ring b. It thus forms an electrostatic field of high frequency which induces oscillations in the other rings c, d, e, f.

As a modification, L2 may be suppressed and the rings a and b may be connected each to one of the ends of L1.

One of the ends, or even any point of the self-induction coils L1 or L2 may likewise be connected to any one of the rings.

Fig. 2 shows the arrangement of the rings in the form of a cone. The open circuits a, b, c....f can be constituted by solid conductors. They can also be tubular circuits, this being advantageous, since, for the same bulk, the radiating surface is larger. Finally, the area of the cross section of each of these various circuits can vary from one circuit to the other for instance, this area can decrease from the periphery (ring a) to the center (ring f) of the radiating system.

Fig. 3 shows the apparatus mounted upon a foot A along which it can slide, in order to be placed at any desired height and upon a hinge B which allows it to be oriented upon the horizon.

I claim:--

1. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters, the split ends being spaced from each other and terminating with small spheres.

2. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters and terminated by small spheres forming capacities.

3. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters and arranged in one and the same plane the split ends being spaced from each other and terminating with small spheres, in staggered position as regards adjacent rings.

4. An apparatus of the type described comprising a base, a plate slidably mounted on said base, a support pivoted to the said plate, a series of open and insulated, high frequency radiating circuits mounted on the pivoted support, and a high frequency producing device for the excitation of the said circuits.

5. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits, said circuits being open, insulated, and having dimensions different from each other.

6. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits arranged in the same plane, said circuits being open, insulated, and having dimensions different from each other.

7. An apparatus of the type described comprising a transformer, a high frequency energizing circuit fed by the said transformer and provided with a self-induction coil and a capacity, a series of open and insulated high-frequency radiating circuits having dimensions different from each other, and inductive connection between two of these radiating circuits and respectively each of the poles of the outgoing element of the high frequency energizing circuit.

G. Lakhovsky
INVENTOR
By: Marks & Clerk
Attys.

June 13, 1944. G. LAKHOVSKY 2,351,055

TUBE FOR PRODUCING MULTIPLE WAVE LENGTHS

Filed Nov. 21, 1941

Fig. 1.

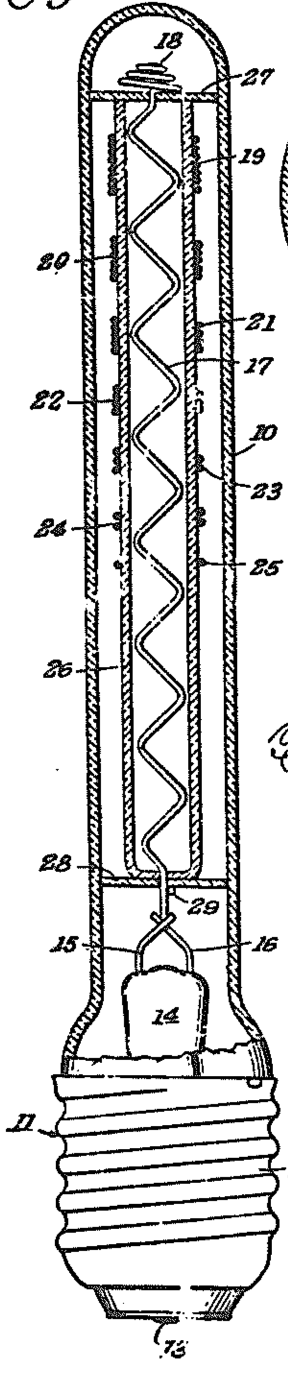
18
27
19
20
21
17
22
10
23
24
25
26
28
29
15
16
14
11
12
75

Fig. 2.

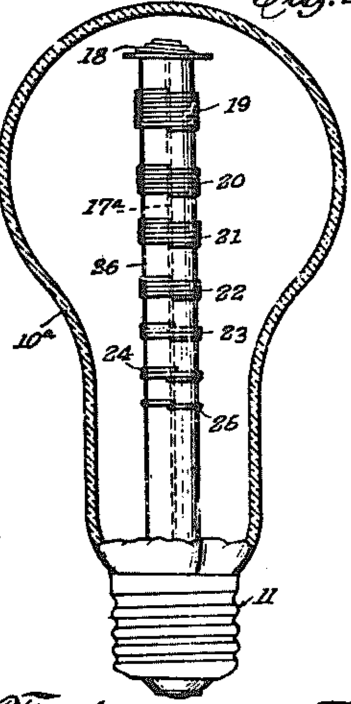
18
19
20
17a
21
26
22
10a
23
24
25
11

Fig. 3.

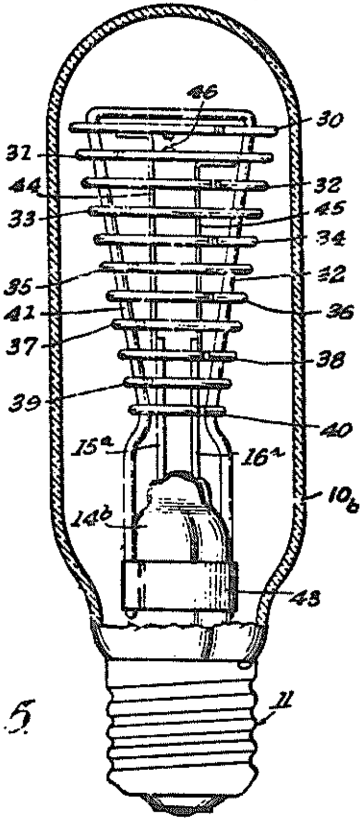
46
30
31
44
32
33
45
34
35
32
41
36
37
38
39
40
15a
16a
10b
14b
43
11

Fig. 4.

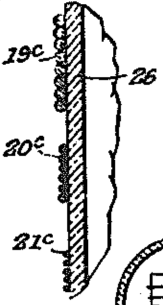
19c
26
20c
21c

Fig. 5.

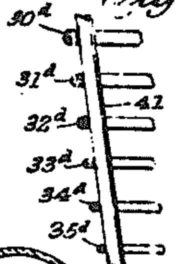
30d
31d
41
32d
33d
34d
35d

Fig. 6.

Fig. 7.

Fig. 8.

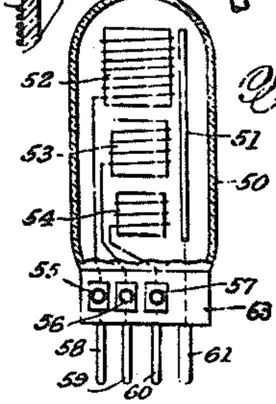
52
51
53
50
54
55
57
63
56
58
61
59
60

Fig. 9.

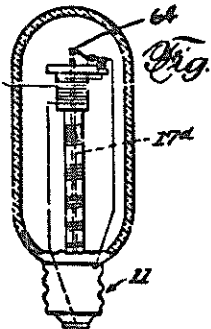

INVENTOR
GEORGE LAKHOVSKY

BY Leon M. Strauss
ATTY.

Patented June 13, 1944

2,351,055

UNITED STATES PATENT OFFICE

2,351,055

TUBE FOR PRODUCING MULTIPLE WAVE LENGTHS

George Lakhovsky, New York, N. Y.; Anne-Marie Louise Lakhovsky, administratrix of said George Lakhovsky, deceased, assignor to Henry S. Blum, Chicago, Ill.

Application November 21, 1941, Serial No. 420,006

11 Claims. (Cl. 250—33)

This invention relates generally to electric devices excited by electric impulses and more particularly to multiple wave length conducting and/or producing means. This invention has for its primary object the provision of such means disposed within a vacuum tube or a tube containing rare gas or gases.

My work over a period of more than 20 years has led me to the belief that there is great need for means conducting and/or producing multiple wave lengths which may be employed in many industrial and medical fields. Suitable apparatus for producing waves of various lengths is disclosed in my U. S. Letters Patent No. 1,962,565, dated June 12, 1934. Embodiments of this apparatus have been in use the world over. Because of the need for simplification in such apparatus so that it may be handy and easily used by persons lacking skill in electrical matters, I have concluded that by providing one unit which incorporates oscillation generating means and another in which the wave lengths are produced, the apparatus may be more universally used in the treatment of cellular life in therapy, in the aging or treating of liquids and in other industrial fields.

The present invention deals with those units, preferably combined with one another and constructed as a portable device and as stated above, has for its major object the provision of a unitary multiple wave length producing and/or conducting device in the form of a vacuum tube.

The present invention further contemplates the incorporation in such a tube, of means for varying the effect of wave lengths, or selectively employing the means therefor enclosed in or forming part of said tube.

This invention further seeks to provide a tube of the indicated type incorporating means for generating oscillations.

Another object of the invention is to provide a single unit which contains an oscillation generator, which produces waves of definite frequency, and further inductance means each permitting emanation of wave lengths of various values.

The structural features of my invention also form a material part of this disclosure, the objects and advantages being attained in structures such as shown in the accompanying drawing, which exemplifies the invention. The following specification, based on said drawing, more clearly points out the purposes and advantages of my invention.

In the drawing:

Fig. 1 is a vertical, partial sectional, partial elevational view of a tube incorporating features of my invention.

Fig. 2 is a similar view of an alternate form thereof.

Fig. 3 illustrates in a similar manner a modified form of my invention.

Fig. 4 is a fragmentary sectional view of a modification made in accordance with the invention.

Fig. 5 is a similar view of another modification.

Figs. 6 and 7 are sectional views illustrating types of wires used in either of the forms of my invention as shown in Figs. 1, 2 and 3.

Fig. 8 is a partial sectional, partial elevational view illustrating a tube incorporating a plurality of grids and selective capacity means for varying the effects of said grids.

Fig. 9 is a similar view of a tube, such as shown in Figs. 1 or 2 and incorporating oscillating or vibrating means.

In that form of my invention shown in Fig. 1, there is provided a sealed preferably transparent envelope **10** made of glass or like material affixed in the usual manner to a screw or Edison plug **11**. The latter is conventionally provided with a shell terminal **12** and a central terminal **13**. The support or seal **14** of the tube serves to hold wires such as **15** and **16** which connect to the terminals **12** and **13**.

Connected to the wires **15** and **16** is a primary inductance member in the form of a loose coil **17** which may be disposed axially of the envelope **10**, as shown. The upper end of said member **17** forms a tight wound coil **18** which serves to radiate the strongest induced wave lengths.

Surrounding the primary inductance member **17**, there are arranged a plurality of separate coils of wire each being of different length and capacity. Thus, I provide a coil **19** having greatest capacity and successive coils **20**, **21**, **22**, **23**, **24**, and **25**, each having less capacity than the preceding coil. These coils **19** to **25** form secondary inductance means each permitting emanation of a wave length of different magnitude and value than the other.

I may support the secondary coils in several ways. As shown, I prefer to provide a dielectric member such as a glass sleeve **26** surrounding the primary inductance wire member **17** and to mount the secondary coils **19** to **25** on the outer surface of said sleeve, leaving the coil **18** exposed at the top thereof. In order to effectively support and centralize the glass sleeve **26**, I provide spacing washers **27** and **28** of mica or the like, top and bottom, and provide a supporting crosspiece **29** on the lower portion of the wire **17**.

The tube above described may be employed as indicated for the treatment of cellular structures by connecting it to any one of the rings in the electrostatic field of high frequency, such as shown in said U. S. Letters Patent No. 1,962,565, each of the coils 18 to 25 producing by induction a different wave length, as can be understood.

In the form of my invention shown in Fig. 2, the envelope 10*a* is somewhat differently shaped, and the wire 17*a* straight instead of arranged as an open coil. In other respects, the structure follows that described with reference to Fig. 1.

As shown in Fig. 3, there may be arranged a plurality of open rings 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, and 40 in frustro-conical form, for instance, with the largest of said rings at the top and the smallest one adjacent the bottom. I have shown said rings as supported, for instance, by glass posts 41 and 42, said posts being supported on the seal 14*b*, as by a band 43. I provide extensions for the terminal wires 15*a* and 16*a* and connect said extensions 44 and 45 to the upper ring 30 and the next lower ring 31, respectively.

The rings 30 and 31 may become the primary inductance members, whereas the remaining rings constitute the secondary inductance members as before set forth. Electric current conducted to the suitably spaced rings 30, 31 will cause a spark 48 to be drawn between the rings 30 and 31 providing oscillations emanating from electrical discharges therebetween and causing the radiation of waves of various lengths within the envelope 10*b* and affecting the remaining rings.

In Figs. 1 and 2, I have shown the secondary coils as formed of wires having uniform cross-section. Fig. 4 illustrates how such wires may be made progressively smaller in the succeeding coils 19*c*, 20*c*, and 21*c*, etc. Similarly, I have shown the rings of Fig. 3 made progressively smaller in Fig. 5, said rings being designated by numerals 30*d*, 31*d*, 32*d*, 33*d*, 34*d* and 35*d*, etc. In such various manners, the wave length producing effect of the secondary coils or rings may be arrived at.

The secondary coils or rings may be round or somewhat flattened wire as shown in Fig. 6 or tubular as in Fig. 7, offering a yet greater flexibility of design.

The devices of Figs. 1 and 2 above described may be incorporated in the system shown in said U. S. Letters Patent by connecting one of the rings *a*, *b*, *c*, *d*, *e*, or *f* with plug 11 of Fig. 1, since either terminal 15, 16 is in contact with primary inductance member 17. To this end a conductor (not shown) connects such a ring with said plug or socket.

In Fig. 8, I have shown a modification of the invention which comprises a vacuum tube 50 containing a plate 51 and a plurality of different wave lengths producing grids 52, 53, and 54. I connect each grid with a variable capacity device, such as 55, 56, and 57 carried by said tube 50 and preferably by its base support 63, whereby the grids may be selectively connected into an operating circuit (not shown) by means of the respective prongs 58, 59, and 60, and the plate prong 61. Heating or other exciting means may be incorporated in the tube in a well-known manner. This combination tube 50 is designed to replace a plurality of known tubes each operating with different and variable wave length effect, mere exterior manipulation at said base support brings about change of the capacities 55, 56, and 57 and cutting in one or more of the grids 52, 53, and 54.

The effect of the spark as produced in the form of the invention depicted in Fig. 3 may be further utilized in Fig. 9 in which I also incorporate a vibrator 62 which produces electrical oscillations by a spark intermittently generated between the end of the primary inductance member 17*d* and vibrator armature 64. In other respects the structure follows that of Fig. 1.

It may be noted that any type of spark producing or oscillation generating means may be substituted for the vibrator shown. The tubes shown in Figs. 3 and 9 may be used in the manner described for Fig. 1 by connecting one terminal of the base as set forth to produce an induction effect. These tubes may be also connected across both terminals with a suitable potential to obtain the spark effect.

From the foregoing it may be seen that I have provided tubes in various forms for the purpose of generating and conducting multiple wave lengths. Other forms of the invention may be produced within the spirit and scope of the invention as claimed.

Having thus described my invention, what I claim as new and desire to be secured by Letters Patent, is:

1. A device of the character described comprising a vacuum tube having an envelope and base, primary inductance means, and a plurality of separate secondary inductance means of different length for radiating waves of various lengths extending above said primary inductance means, all said inductance means being arranged within said envelope.

2. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member axially arranged in said envelope, and a plurality of secondary inductance members spaced apart and independent from each other each member surrounding a respective portion of said primary inductance member and within the field of influence thereof.

3. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member axially arranged in said envelope, and a plurality of secondary inductance members each surrounding the primary inductance member and within the field of influence thereof, each of said secondary inductance members having a different wave length effect than the others.

4. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member comprising a pair of split rings in inductive relation, and a plurality of secondary inductance members each comprising split rings.

5. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member comprising a pair of split rings in inductive relation, and a plurality of secondary inductance members each comprising split rings, said latter rings being arranged in different planes and being of different length to produce a different wave length effect.

6. In a device of the character described, means for producing multiple wave lengths comprising a wire member, enclosing dielectric means for said member, and a plurality of separate wire coils of different lengths on said dielectric means and in the field of inductance of said wire member.

7. In a device of the character described,

means for producing multiple wave lengths comprising a wire member, enclosing dielectric means for said member, and a plurality of wire coils on said dielectric means and in the field of inductance of said wire member, each coil extending over said wire member and having a different number of convolutions than the others.

8. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings.

9. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings, the plurality of rings being of smaller diameter than the mentioned two rings and each progressively smaller than the others.

10. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings, the plurality of rings being of smaller diameter than the mentioned two rings and each progressively smaller than the others, all of the rings being arranged in frusto-conical disposition.

11. In a vacuum tube having a primary inductance member and secondary inductance members, a vibrator arranged in the field of influence of said primary inductance member and secondary inductance members, a vibrator arranged in the field of influence of said primary inductance member for affecting the induction of said secondary members.

GEORGE LAKHOVSKY.

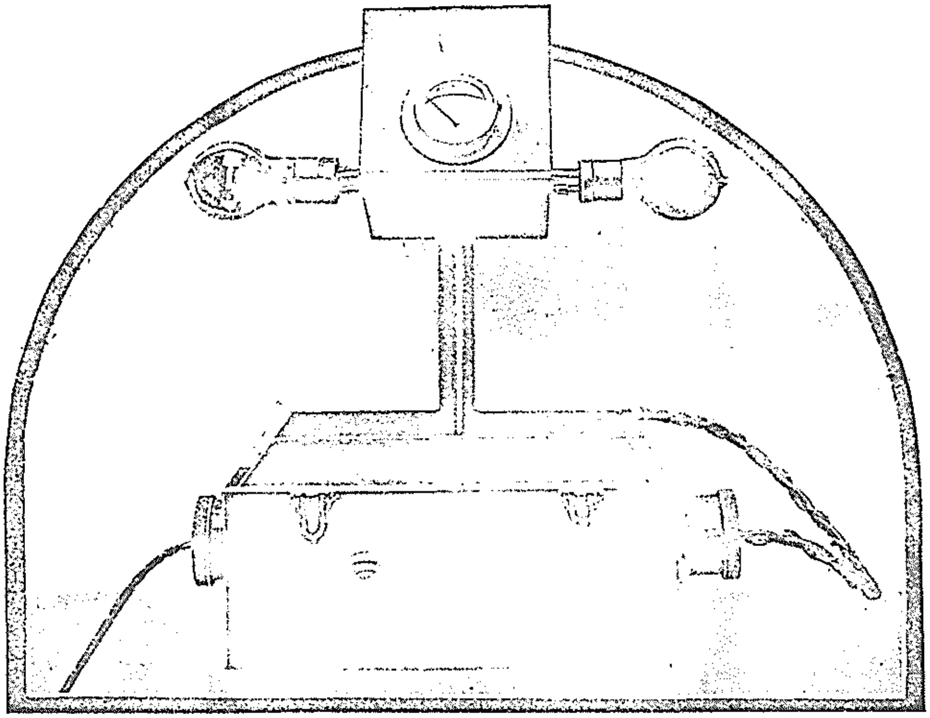

The two-meter oscillator used by the author in his experiments with the supply box containing the necessary controls and transformers.

Picture from *Radio News* for February 1925

Curing Cancer with Ultra Radio Frequencies

by George Lakhovsky

LAKHOVSKY'S VIBRATORY CIRCLE

Author Unknown, Source: BSRF files
Translated from the German by Bodo Capeller, 1988

"We owe to magic all progress in physics and chemistry."
(Jules Lermina, 1915)

The Paris professor with the Russian name Georges Lakhovsky claims to capture, with so called open vibratory circuits (circuits oscillants), cosmic waves and pass them on to humans (neck antenna) and plants. Healing of people has been verified by French and Italian medical practitioners, healing of Cancer injected plants has been proven through photographic records.

The vibratory circles are manufactured from heavy copper wire with a diameter of 5-10 millimeters. Those wires 5-8 millimeters in diameter had a more pronounced effect than the much thinner variety. The spiral (of the open ended circle), which can be single or double, is guided around the plant and is supposed to be insulated against its support by "Ebonit" (1).

Instead of copper wire medical practitioner Dr. Ernst Busse (Garmisch, Germany) recommends the use of thin copper tubing: "because of the short wave length of cosmic vibration only the surface effect (skin effect) is effective (that means electricity is only present on the surface of a conductor, not in the middle) which requires the use of maximum surface area. Besides this the tubing is easier to work with and has less weight. (Private Correspondence 1/10/1953)

Helmut Friedensdorff (Berlin, Germany) did report already in 1932, from himself as well as others, on the effectiveness of an open vibratory circuit against lumbago among other ailments. He inserted a simple isolated copper wire into surgical tubing and positioned it at the problem spot in a manner that both wire ends did not touch (2).

Canon Smythe of Great Britain did proceed in a similar manner: "The only tools required are a one meter length of flexible electrical wire and a sidereal pendulum. The insulation is removed at both ends for a length of 2 centimeters, the wire is stripped and the wire strands are removed except a single strand. Then the wire is laid out on a table and the pendulum is positioned over the wire ends. It will rotate over the positive end in a clockwise motion, over the negative end in reverse. Then one proceeds to make a knot at the negative end for identification purposes. It should be positioned on the left side of the body or in the vicinity of the leg, the bare wire pointing to the left. The wire may be installed over the clothing and can be braided or fastened with safety pins, the ends should point into the proper direction.

"Arthritis in one toe should be treated by using the wire like a stocking holder wound around the leg where the knotted end will point to the left. The author makes no claims for the healing arthritis but reports a lessening of pain.

"With severe headaches the wire is wound just above the eye level around the head so that the negative end rests over the left temple.

"Severe coughing is treated with one wire around the throat, another one around the abdomen."

A just started cold is removed by our associate in a few hours with two wires, one with the positive connected to the negative end of the next wire joined as a 2 meter long wire. "I wound the wire over the shirt under a sweater around the chest until I reached under the arm pits and made sure that the negative end is on left side of the chest pointing to the left, the positive end pointing the opposite way."

Smythe published this in a paper many years ago called "The Other Half of Medicine". He is supported by the research of a recently deceased friend, Dr. Martin of Airaines, France who had been promoted to the medical faculty in Paris in 1932 with a dissertation "Diagnostic Radiesthesique en medecine veterinaire" under the tutelage of professor Nathering--without an examination! In 1935 a Dr. Girard presented a similar thesis before the University at Lyon.

Medical practitioner Arnold Mannlicher (1880-1958) from Salzburg carried in his rear pocket an open vibratory circuit which also acted as a cell detector.

Vibratory circuits proliferated after the announcements by Lakhovsky's discovery as "throat antennas, broadcast jewelry and healing chains" in the market place. I cannot say at the moment if they were of the open or closed kind. Apparently between the two kinds exists a fundamental difference in their effectiveness, open circuits reduce the tension within the body, closed circuits raise it. A closed broadcast chain can easily be transformed into an open one when it is opened and is closed with an insulator inbetween (rubber band, silk thread).

The open circuit is nothing new at all. Dr. Beck, former head of the municipal children's clinic in Bayreuth says: "the open bronze rings of the Hallstadt area from grave sites can be understood as protection against radiation effects very similar to the rings worn by the Negro tribes around the arms, legs and throat." The ancient Germans wore "broadcast jewelry"--in order to coin a modern word--an open solar spiral. Dr. Heermann has revived this. He writes: "invigorating liver and intestines the solar wheel acts like a plated iron spiral with the windings running from the center clockwise."

Smaller--sown into green material--is my usual aid in stomach ulcers and swollen glands. Underneath his illustration (not shown) the researcher explains: "a spiral ending growth rays to the abdomen" (3).

The never tired Dr. Busse since 1954 has also tried out open rings beside vibratory chains and belts. Earlier he replaced acupuncture needles with metal foil band aids and states now in a letter dated 4/19/1954: "instead of metal foil one may arrive at the same effect by wearing open copper rings on certain fingers and joints. When you put a ring that is cut open or manufactured from heavy copper wire on the corresponding finger an immediate change of pulse at the radial pulsation can be observed".

One of his associates who experimented with these open finger rings was surprised by the achieved results.

I am very much indebted to this versatile medical practitioner.

Because copper rings are a tuner (strengthening, invigorating) it was required to find a second metal which reduces organic functions. It turned out to be aluminum. On closed rings it was already reported in the chapter "Amulets". As far as the application of aluminum is concerned the welder Herbert Bauer (born 1921) from Freiburg, Germany has manufactured many years ago the so called "Zelemente".

According to the description it was recently manufactured from a three metal alloy which reminds of the "seven metal (electrum magicum)" of Hermetic fame. We are dealing here with closed chains that are supposed to protect from dangerous radiations of technical and cosmic nature and are therefore named radiation protective antennas. They look flat and are meant to be moistened in the morning and evening.

Whatever the effect is of the vibratory circuit on the world of plants has been shown in 1905 when sick elm trees were healed on a Hamburg street.

The Vienna engineer Lambert Binder (born 1905) caused, during a growth experiment in 1944 on orange trees, accelerated growth by surrounding the tree with an isolated wire spool.

R.E. Esspiau tried out in 1954 the following experiment with the Lakhovsky antenna arrangement. He connected the first on a three stalk head of cabbage to free standing antenna, the second to ground the third one to water which was replaced every 24 hours. In less than a week the first buds began to form, within four weeks other stalks were formed and developed leaves while a control plant withered away (4).

PHOTO TEXT:

Illustration A: This illustration shows a Lakhovsky experiment on a Cancer infected geranium plant. The cancerous growth is clearly visible. Pelargonium zonatum (Geranium) inoculated with Bacterium tumefaciens on April 10th, 1924 and photographed June 6th, 1924. It will be observed that the stem of the plant bears a massive tumour.

Illustration B: An astounding discovery by Prof. Lakhovsky: a free standing antenna ring manufactured from copper that surrounds the geranium caused the plant to drop the cancerous growth (resting on the rim of the container) and made the plant whole again. Photographed June 1925.

Illustration C: The test plant three years later. The antenna ring has completely healed the plant and protected it from other damage. To the side are two unprotected similar control plants of same age that became victims of the same disease. Photographed on March 13, 1928.

Picture credits: These three illustrations are designated tables II, V, and VII in Lakhovsky's Secret of Life.

BIBLIOGRAPHY

(1) Lakhovsky, Georges: Das Geheimnis des Lebens (Secret of Life), Kosmische Wellen und vitale Schwingungen, Munich 1931. Schroedter, Willy: "Der Drahtring um die Geranie, Frankfurter Ilustrierte, #15, 1932, page 356.
(2) Frensdorff, Helmut: Einfache Mittel gegen Hexenshuss und Gliederschmerzen, "Die Zukunft", #31, Berlin 1932.
(3) Heermann, Alex, M.D.: Neues von Strahlen, Stroemen und Wellen, Bad Ausee 1935, page 13.
(4) "La Radiesthesie pour Tous", December 1954, Paris. Schroedter, Willy: Kosmische Strahlen und vitale Schwingungen, "Mensch & Schicksal", #5, 5/15/1953, Villach, page 7

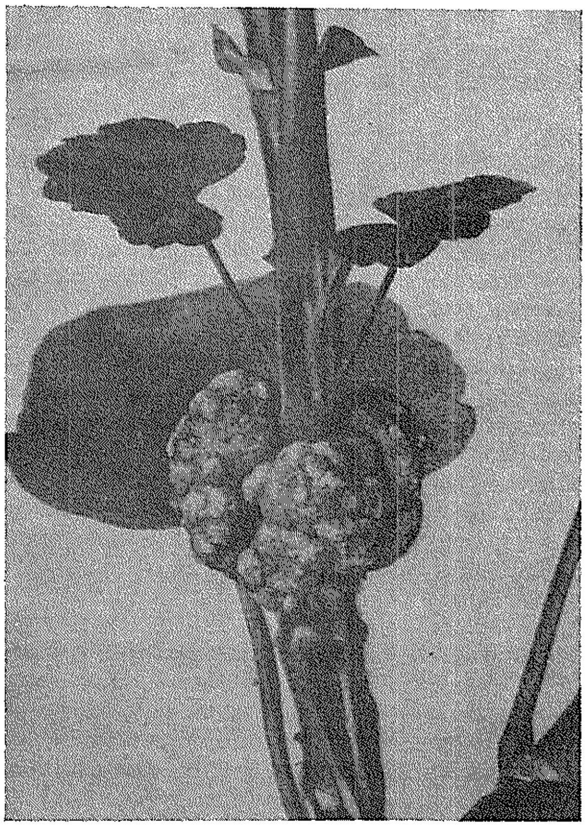

Illustration A

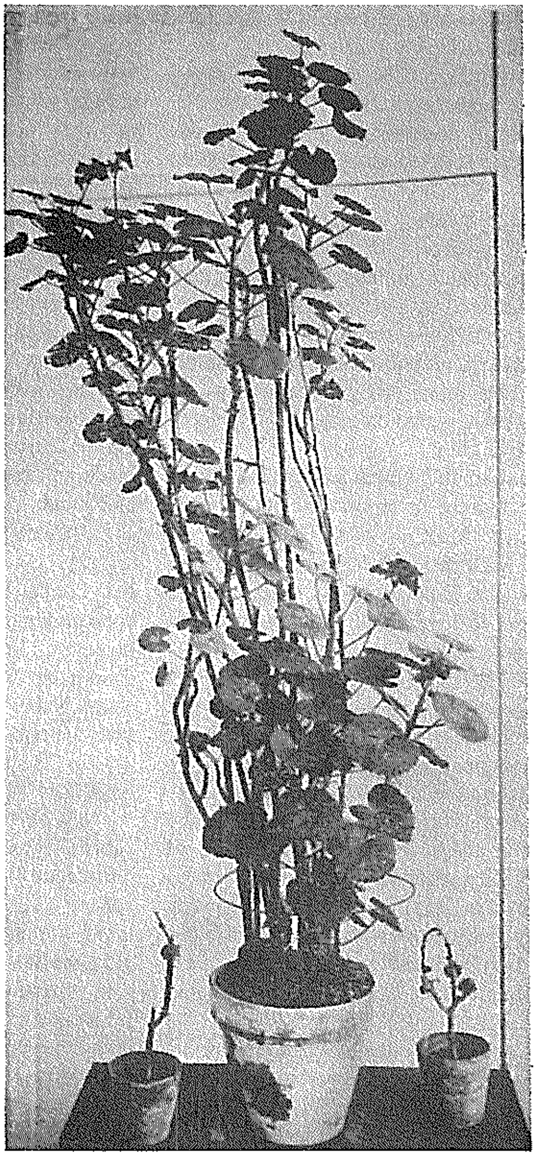

Illustration C

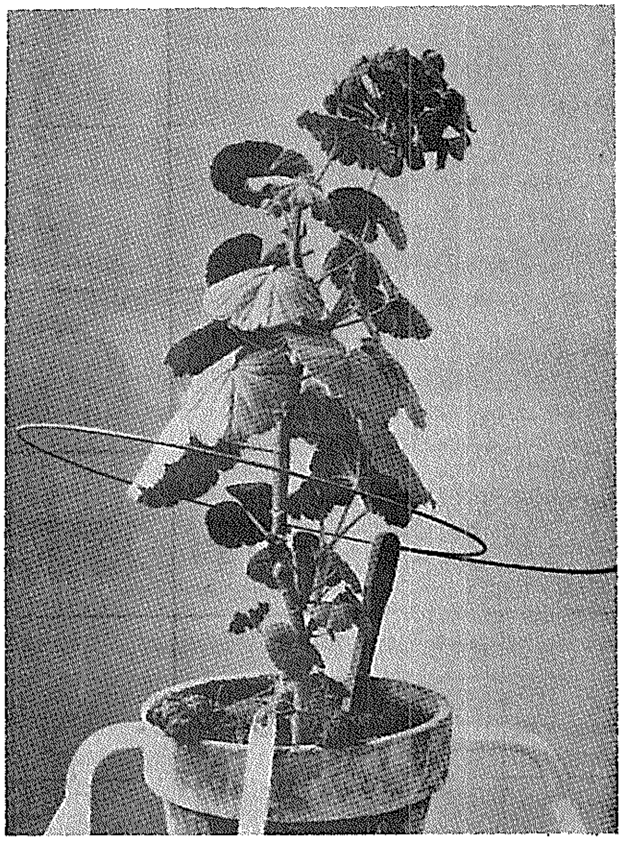

Illustration B

THE RUSSIAN LAKHOVSKY REJUVENATION MACHINE

by Bob Beck

from the November 1963 Journal of Borderland Research

In September 1963, Dr. L.O. Anderson, Anderson Research Center Director and BSR Associate, phoned me about a "marvelous electronic machine" purported to "cure" almost any human ailment. I've looked into dozens of these "modalities" every year; so with extreme skepticism I gathered instruments and took off to check another possible hoax.

I found a heavy spark-gap Tesla Coil inside a four-foot cabinet. This console was rigged to circular "antennas" and was partially shielded inside a home-built Faraday cage in the living room of Dr. Anderson's friend. It hadn't functioned in over a decade, and had been given up by some twenty-odd medical instrument repairmen and engineers as unfathomable.

The machine, which had no nameplate or other identification, turned out to be a Lakhovsky "Multiple Wave Oscillator", or MWO. Fortunately the present owner had found patents and two books describing its operation. We were able to reconstruct the circuits and after several hours had it operating perfectly. The machine was not a hoax.

Lakhovsky did his research prior to world War II and died in 1942. No infringement can be involved in duplicating Lakhovsky's work, since his U.S. Patent #1,962,565 expired in 1951 and is in public domain. I have found from the book **The Waves that Heal** by Mark Clements that MWOs are extensively used in hospitals in France, Germany and Italy, and although tested in a New York Hospital with fantastic success, MWOs remain practically unknown in the United States.

MDs testing Multiple Wave Oscillators reported up to 98% cures of terminal cancer, arthritis and other "hopeless" conditions. Case histories were complete with before-and-after photos. Often rejuvenation was demonstrated, not only in human beings but in plants and animals as well!

So, on Sept. 18, 1963, I designed and built a modern MWO based on Nikola Tesla and Lakhovsky's original work and patents. The transistorized instrument fits into a brief case and weighs under 14 lbs. Experiments with the MWO on some eighteen volunteers have demonstrated almost unbelievable results. At this writing, Oct 20th, every single subject, including myself, reports fantastic gains.

My mother, age 72, who has suffered from hypertension and glaucoma for many years, measured absolutely normal blood pressure the morning after her first treatment. And the eye tension, tested by her MD the week following, had dropped to a lower point than in the previous two years. Another woman with a long-standing back condition says she was permanently relieved in one 15 minute exposure. An asthmatic with chronic bronchial complications claims he is now symptom-free after two exposures and her husband with a bad dermatological condition of seven years history appears to be very definitely improved. Perhaps the most dramatic is Mr. N.V.C. who had broken his leg skiing last December. Because of an old polio complication the broken leg had never healed and flesh surrounding the fracture was in a very poor, waxy condition. After one exposure he phoned to say he had regained feeling in the area. The next day we made an "after" photo slide of nearly normal tissue.

A friend suffering allergies of long standing says she can now eat anything and her MD admits he as startled by her youthful appearance and new vitality. The Lakhovsky literature illustrated many such rejuvenations as an unexpected side effect of cancer cases. Subject's wrinkles disappeared along with the tumors!

THEORY OF OPERATION

Lakhovsky pointed out that all cells capable of reproduction contain in their nuclei "filaments" of highly conductive material surrounded by insulating media. This filament, which may be the RNA-DNA complex, is always in the form of a spiral or helix -- in other words, a coil. Therefore, each will react as a tuned circuit if its resonant frequency can be approximated by an external oscillating coil.

Lakhovsky didn't carry this to it's conclusion; however, I postulate that by exciting the nuclei with electromagnetic energy a "charge" can be induced by the long established principle of electromagnetic induction. This demonstratively raises the energy level and perhaps the vitality of every cell in the field simultaneously. Since each cell is an individual, and of slightly different physical dimensions, the exciting wavelengths must be multiple, and must span a broad frequency spectrum. Diathermy machines, limited to crystal-controlled single frequencies in the 27.255 MC region, can do nothing but heat the tissue; and yet this approach, abandoned by Lakhovsky in the 1930s, can still be found in "modern" doctor's offices!

The Electromotive Force (EMF) produced by the MWO and induced in the cell nucleus, can raise the cell's metabolic rate by Electrolysis, and perhaps jog the RNA-DNA "memory" and reproductive capabilities to their level at an earlier, younger age, thus the rejuvenation. Even more subtle changes might be postulated, such as magnetic "progression" effects as evidenced by heavy water in magnetic fields.

By contrast, any chemical taken into one's system by mouth or by injection -- such as vitamins, food or medicine -- must first be absorbed in proper solution and ratio with unknown tract elements, then be transported by the blood to affected tissue only to be possibly rejected by the target areas.

Perhaps in cancer the EMF induced by the MWO raises the vitality and memory of marginal cells to normal reproduction levels. In the case of other disease, perhaps a cell's given higher energy levels can more readily throw off affliction.

A WAVE FOR EVERY BODY CELL

The Multiple Wave Oscillator described here radiates a bandwidth of radio frequency (RF) energy from the audio frequencies up beyond microwave frequencies. By actual measurement with standard field strength meters, this vast bandwidth of frequencies and harmonics can be shown. In fact, a bluish glow of "brush discharge" surrounds the antenna when operating. A fluorescent lamp held anywhere within several feet of the subject glows brilliantly. Within this multiple-wave range of frequencies, every cell in the body can find its ONE resonant frequency and absorb energy at its own natural wavelength.

Obviously the Electrostatic energy cannot penetrate the body. This is known as the "skin effect". However, the Electromagnetic component of the energy can and does permeate and will induce an EMF in each cell. It is precisely this energy to which Lakhovsky attributes his almost miraculous "cures".

A word of warning, although no deleterious side effects have ever been reported for human reactions, the MWO has input power exceeding 35 watts. It radiates RF interference over a bandpass of 15 Kilocycles to well beyond 250 Megacycles. The Federal Communications Commission has stiff penalties for unlicensed transmitters and radio and TV interference -- not to mention making all your neighbors very unhappy. We operate our machine in a shielded room, and the battery input prevents RF interference from feeding back into the power line.

The deluxe MWO diagrammed here can be built by any intelligent 16-year old with readily available electronic parts. I do not plan to build, manufacture, sell or otherwise exploit this device, as my interest is purely in research. Several Associates given this circuit have successfully constructed MWOs. We would certainly like to hear from any BSR Associates who undertake this project, but please do not communicate until after building the unit. I cannot correspond over details which any intelligent person would be able to thresh out with his or her own technical friends.

At one stage we had planned constructing a much higher-powered unit but the results with the "portable" are so gratifying that Model Two was shelved. Extensive tests are scheduled with plants and animals, and every effort is being made to remove any element of "suggestion" from results. A physician friend is cooperating with the project by sending a few "problem" patients for test. It is much too soon, however, to make a finished statement concerning the MWO's potential. No cures can be claimed even if obviously demonstrated, since our laws are controlled by medical unions forbidding the use of the word except by MDs. But I have satisfied myself with nearly twenty miracles in these few weeks of tests that the MWO might well be the most promising device to come to light in many years.

I have one request, that no one building this device ever be tempted to charge a fee, donation, or reimbursement of any kind whatsoever for its use. I feel most strongly that any device potentially helpful to suffering humanity should be available completely free of charge to anyone needing it.

CONSTRUCTION TIPS

My MWO is built on an 11 x 11½ inch plastic panel. Components extend 3½ in. below this, so the unit fits nicely into a small attache case. The 6 volt battery was a Sonotone lifetime Nickel Cadmium unit; however, five war-surplus NiCads (1.2 v each) wired in series would be less expensive.

My first unit used a transistorized driver for the Tesla Coil but the Ford coil works equally well, is far cheaper, but has the disadvantage of making a buzzing noise. The Ford Model T coils may still be available at Pep Boys stores in California, or can possibly still be purchased from J.C. Whitney & Co., 1917 Archer Ave., Chicago, IL 60616. The spark gap was made from salvaged tungsten relay contact points about 1/4 in. in diameter. (Chevrolet Six timer points, 1950, do very well). The relay spring arm can be tuned by a Nylon 6-32 screw adjusting its natural spring tension. The gap must be carefully tuned for maximum output for every individual, since the subject's absorption seems dependent on his body fluids and weight. *Adjust for maximum sparking at random times and intervals between several of the inner antenna rings.*

The automatic timer, zero to 15 minutes, the meter, the internal battery charger, the RFC circuits, etc. are all simply conveniences and may be left out of the circuit without in any way subtracting from results. By not building in these deluxe features, the entire MWO will cost much less to build. Use plastic instead of metal screws around the high voltage coils. High voltage capacitors can be tracked down at surplus stores or well stocked electronics parts sources. The Tesla Coil must be hand wound as it is not available commercially.

We have determined experimentally that a workable pair of antennas, radiator and resonator, can consist of 16 concentric rings, ranging from an outside diameter of 10-3/4 inches down to 1/2 inch. These were constructed by cementing a one-foot square of

BECK ADAPTATION OF LAKHOVSKY MULTIPLE-WAVE OSCILLATOR, U.S. Patent #1,962,565

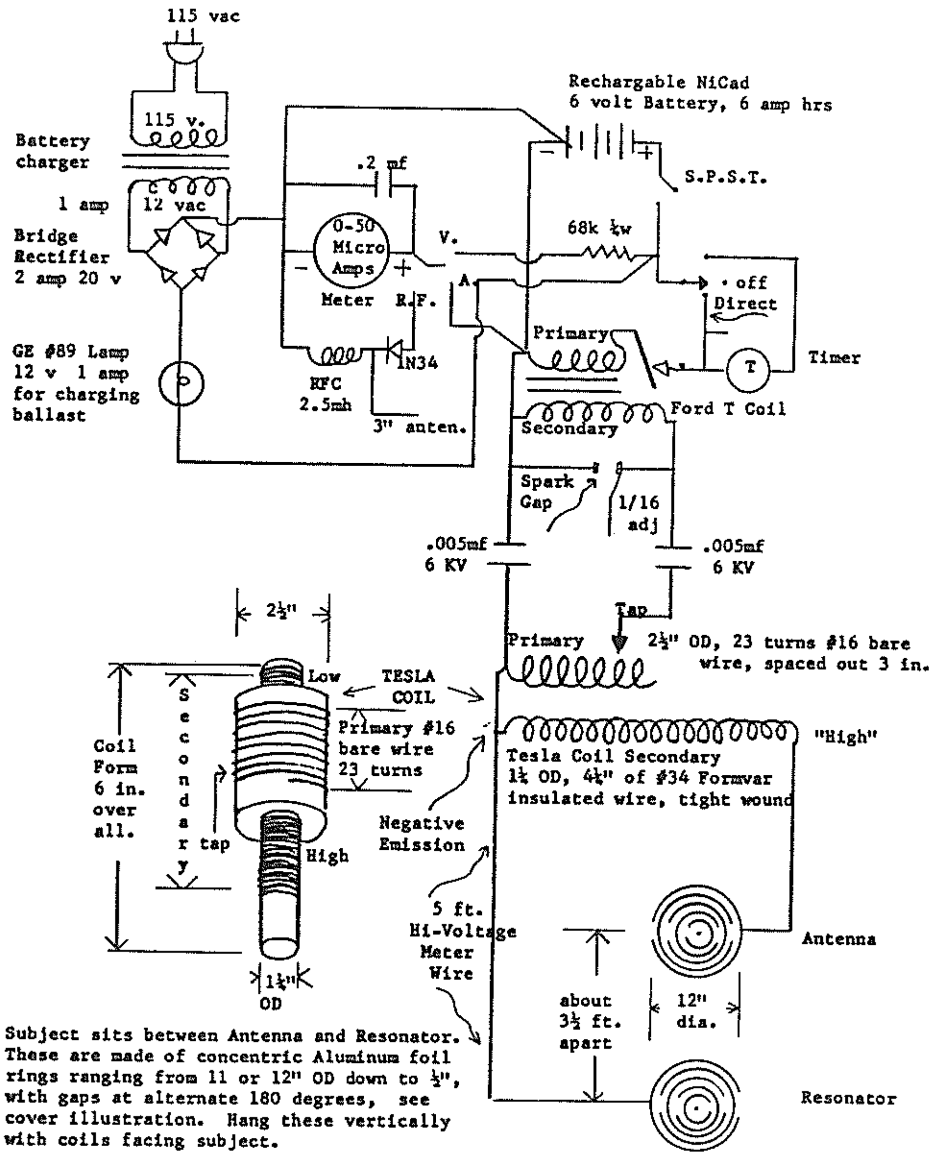
115 vac
115 v.
Battery charger
12 vac
1 amp
Bridge Rectifier 2 amp 20 v
GE #89 Lamp 12 v 1 amp for charging ballast
.2 mf
0-50 Micro Amps
−
+
Meter
R.F.
V.
A.
RFC 2.5mh
1N34
3" anten.
Rechargable NiCad 6 volt Battery, 6 amp hrs
−
+
S.P.S.T.
68k ¼w
• off
Direct
T
Timer
Primary
Ford T Coil
Secondary
Spark Gap
1/16 adj
.005mf 6 KV
.005mf 6 KV
Tap
Primary
2½" OD, 23 turns #16 bare wire, spaced out 3 in.
TESLA COIL
"High"
Tesla Coil Secondary 1¼ OD, 4¼" of #34 Formvar insulated wire, tight wound
2½"
Low
Secondary
Primary #16 bare wire 23 turns
Coil Form 6 in. over all.
tap
High
1¼" OD
Negative Emission
5 ft. Hi-Voltage Meter Wire
Antenna
about 3½ ft. apart
12" dia.
Resonator

Subject sits between Antenna and Resonator. These are made of concentric Aluminum foil rings ranging from 11 or 12" OD down to ½", with gaps at alternate 180 degrees, see cover illustration. Hang these vertically with coils facing subject.

Original cover providing template for aluminum MWO antenna.

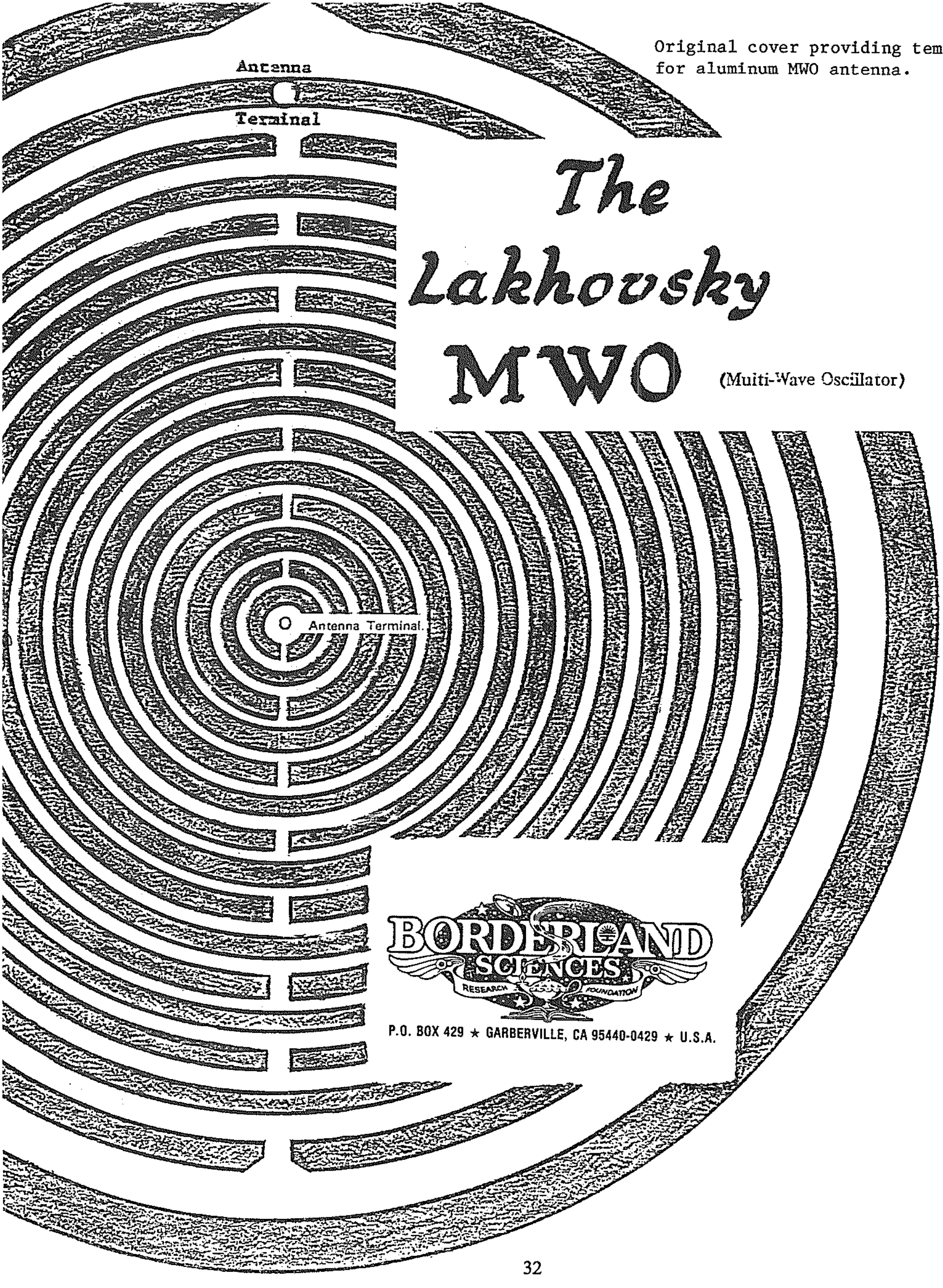

The Lakhovsky MWO

(Multi-Wave Oscillator)

BORDERLAND SCIENCES RESEARCH FOUNDATION

P.O. BOX 429 ★ GARBERVILLE, CA 95440-0429 ★ U.S.A.

Antenna Terminal
Antenna Terminal.

Reynolds kitchen aluminum foil to an 11-inch square of thin fiberglass. The outline of each set of circles was carefully scribed with a sharpened compass, razor or Scripto blade. By scoring or cutting the aluminum foil was cut through to the plastic backing and the waste area strips pulled away before the cement hardened. Excess cement was removed with a solvent. The gaps were then cut with the razor. Note that gaps are alternated in position each 180°, as per drawing on original cover page. This was drawn from an actual antenna and you can set your compass distances for each ring from it. Keep gaps between inner rings narrow to encourage random arcing between rings. Trim the corners from the gaps of the three or four outside rings to discourage fixed arcing at these places. Also be sure to clean away any bits of foil between rings. These also will cause fixed arcing.

In use, the ring second from the outside is connected to the high side of the Tesla secondary. (Refer to "antenna terminal" on original cover page illustration) This is the radiator coil. The arcing from this ring to the outside ring causes the "shock excitation" of the fourteen other antenna elements. Since each radiator is of different diameter, the array spans a broad, multiple-wave segment of the Radio Frequency (RF) spectrum. The other antenna element, called the Resonator, is connected with a five-foot, high-voltage meter wire to the ground side of the Tesla coil secondary.

PUTTING THE MWO TO USE

The antenna and resonator are suspended from non-conductive supports about three feet apart, such as wooden chairs with no nails or screws. The subject of the experiment, preferably with all metal removed from their person, including belt buckles, zippers, hooks, fasteners, coins, hairpins, earrings, rings, etc, is seated on a wooden or plastic rest between the antenna and resonator (hung vertically, the foil rings toward the subject. Unless the Resonator is powered by a separate Tesla Coil it will not arc. Metal on the subject will produce eddy currents in the field of radio frequencies, reducing their effectiveness.) Experiments should last from nine to over 15 minutes. A four-day interval between experiments is recommended. Four such experiments over 16 days should suffice to demonstrate anything.

Warning, two-inch sparks can be drawn from antenna; so keep subject at least three inches away. If touched a mild and harmless sting can be felt. HF effects can be demonstrated in a darkened room by holding a fluorescent lamp anywhere near the subject. The RF energy will cause the lamp to light although no wires are touching it. Field strength measurements read several watts of power from the audio frequency spectrum to beyond 250 megacycles.

The illustration to the right is from Clement's **Waves That Heal**, Lakhovsky's illustration of a living cell. "In the centre is a twisted filament which, possessing self-inductance and capacity, constitutes an oscillating circuit. The similarity to a short-wave circuit is manifest; the filament shown here oscillates like a coil having a very small number of spirals." He quotes Lakhovsky: "Every living being emits radiations", and adds: "the living cell can thus play the part of a transmitter or receiver of exceedingly short radio-electric waves which give rise to high frequency currents in the circuits of the cellular nucleus."

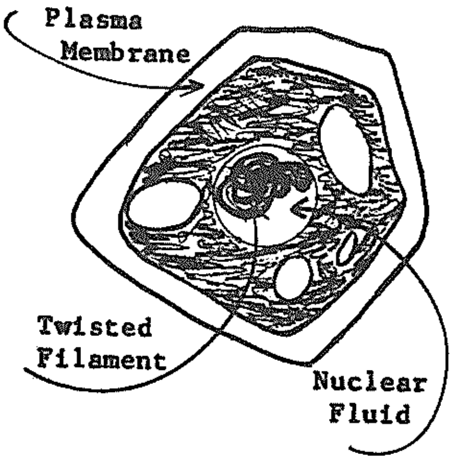

THE FOLLOWING PHOTOS were kindly furnished by Borderland Sciences Research Associates from their own experimental work with the MWO during the years of 1964 and 1965.

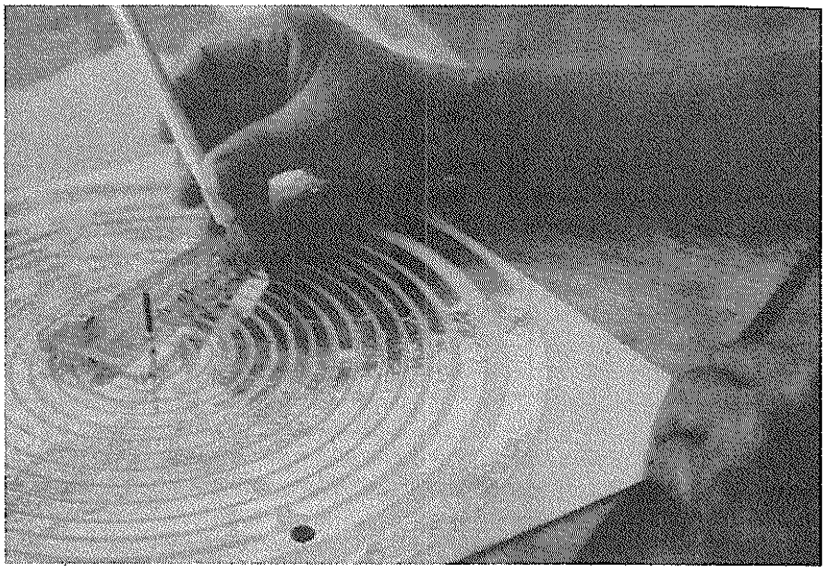

THE PRINTED-CIRCUIT ANTENNA foil is glued to an 11-inch square of 1/16th inch thick fiberglass. Spray one surface of the fiberglass with Hobby Shop spray glue and stretch smooth-surfaced kitchen aluminum foil across it. Press the foil on firmly from the center out, smoothing out repeatedly with a soft cloth. Any remaining air bubbles can be released through a small pin hole. The circles are more easily cut if you make a jig with a 16-inch square of plywood with a headless nail for a center post. Using the antenna pictured on the front cover as a pattern, mark and notch an old ruler or stick where each circular cut is to be made. This marked guide ruler is mounted over the antenna on the plywood jig as shown at right. The foil-covered surface is then turned clockwise with the left hand against the Scripto blade held firmly in the successive notches by the right hand. Be careful to cut through the foil. If you don't, in lifting the cutouts afterward, you may also lift or tear the rings. If a metal grommet is placed in the terminal hole in the second ring this will prevent the high-frequency arc from burning the foil and fiberglass during operation. Copper foil can be used for the printed-circuit antenna rings but deposits build up on it from arcing, wherever it occurs. There is no deposit buildup on aluminum foil.

THE RADIANT ENERGY of the MWO is obvious in this three-minute time exposure during operation in a darkened room. Note the brush discharge of etheric energy all along the edge of the outside ring, going three or four inches. Look at the random arcing between the antenna rings, down to the eighth ring! This antenna received full power by being mounted on the MWO case, directly at the positive end of the Tesla Coil.

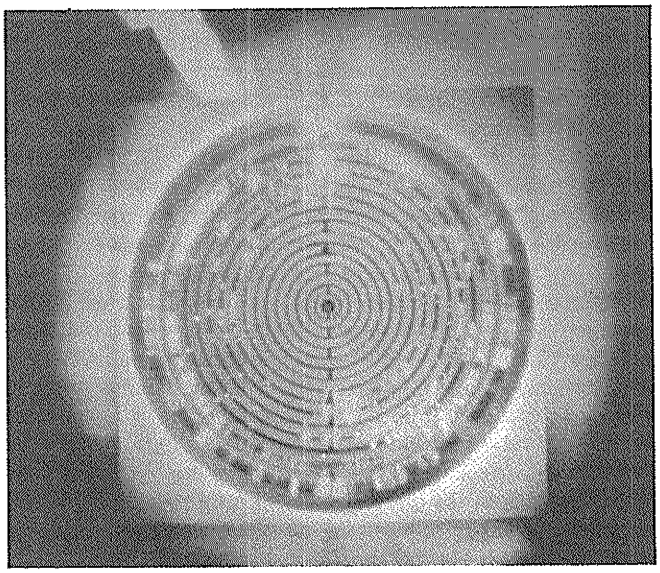

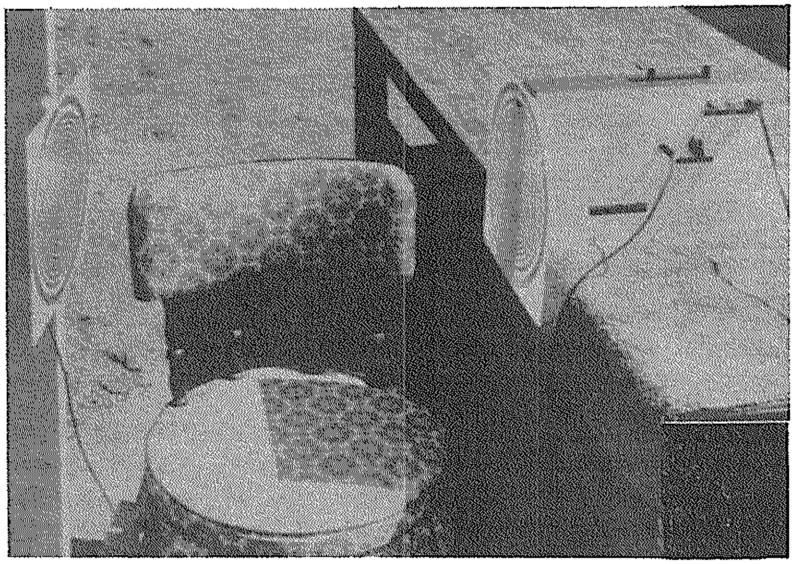

AT LEFT IS A TYPICAL EXPERIMENTAL SETUP, with a chair for "guinea pig" between antenna and resonator. Beck's original equipment called for an 18-inch lead wire from the positive end of the Tesla coil to the antenna. It was found that there was too much loss of power along this wire. The later development shown above has the MWO set upright, with the antenna mounted directly against the positive end of the Tesla Coil in the case.

THE INTERIOR HARDWARE LAYOUT is shown at right in this photo of another MWO, neatly packaged for an attache case. A. The five Ni-Cad batteries, 1-1/4 volts each at full power. B. Timer. C. Tesla Coil. D. 115VAC transformer (for house current operation if desired). E. Spark-gap using modified Chevy Six points. F. Model T Ford Coil. G. Battery Charger.

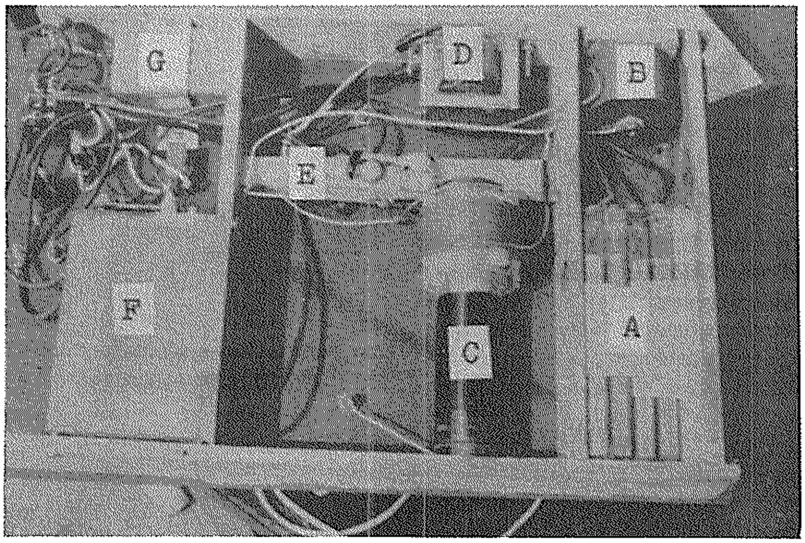

RECTIFYING THE FORD COIL OUTPUT using a 1B3GT or 1G3GT high-voltage tube.

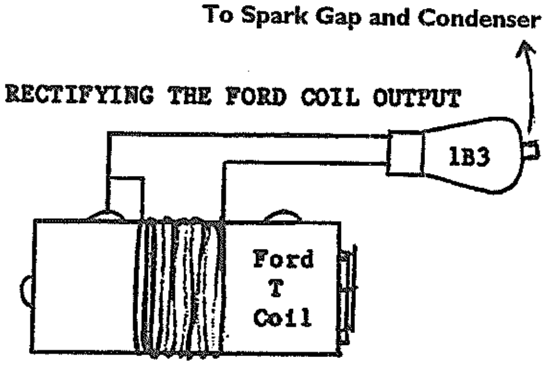

Wind 15 or more turns of fairly heavy insulated copper wire around the middle of the Model T Ford Coil. One lead from the winding goes to the Hi-voltage output of the Coil, and to pin 2 or 7 of the 1B3 tube. The other lead goes to the other pin, as only two pins of the tube, 2 and 7, are used in this hookup. The souped-up or rectified voltage is drawn off the top of the tube to the spark gap and condenser.

MODIFIED VERSION OF THE BECK TESLA COIL

115 volt house current into the black leads of the 6 volt transformer, one green lead to the T Coil negative terminal, the other to an SCR 45 amp Diode to make direct current to the Plus terminal of the T Coil. A 12-volt current to the T Coil heats it up and will not increase the spark. Radio Shack's 6.3 volt, 3 amp transformer does very well for a power supply.

Wrap 16 turns of insulated hookup wire around the middle of the T Coil and feed to pins 2 and 7 of the 1B3 rectifier tube. Feed the plus output of the T Coil into pin 2 or 7 of the 1B3 also.

Hi-voltage from the top of 1B3 goes to Spark plug. Widen gap of plug to 1/8 inch or more for maximum output of Secondary coil to Antenna.

One 10,000 WVDC (Working Volts Direct Current) commercial-type Capacitor, on the plus side, will do very well in this circuit. Capacity can be from .0015 to .005 microfarads.

Beck's tunable Tesla Primary called for 20 or more turns of bare wire for his movable contact but experience shows that 16 turns of insulated #19 wire for a Primary and adjust one turn more or less for maximum spark at Plus end Secondary coil. This is 4-1/4 inches of #34 insulated wire, or 4-3/4 inches of #32 insulated wire, for 420 turns more or less on 1-1/4 inch coil form 6 in. long. The Primary is wound on a 2 or 2½ in. coil form. There should be heavy insulation between Primary and Secondary to prevent power-loss arcing between Coils.

REMEMBER, the goal is a thin blue, or blue-white spark at least 3/4 inch long at the plus end of the Secondary. When fed into the center hole of the Antenna it should light up like a Christmas tree, with arcing between every ring. Feeding the coil output into the center hole gives more power to the shorter wave lengths. We get plenty of the longer wave-lengths from other electrical apparatus and from power lines.

This coil would be a good high-frequency power supply for Kirlian Photography.

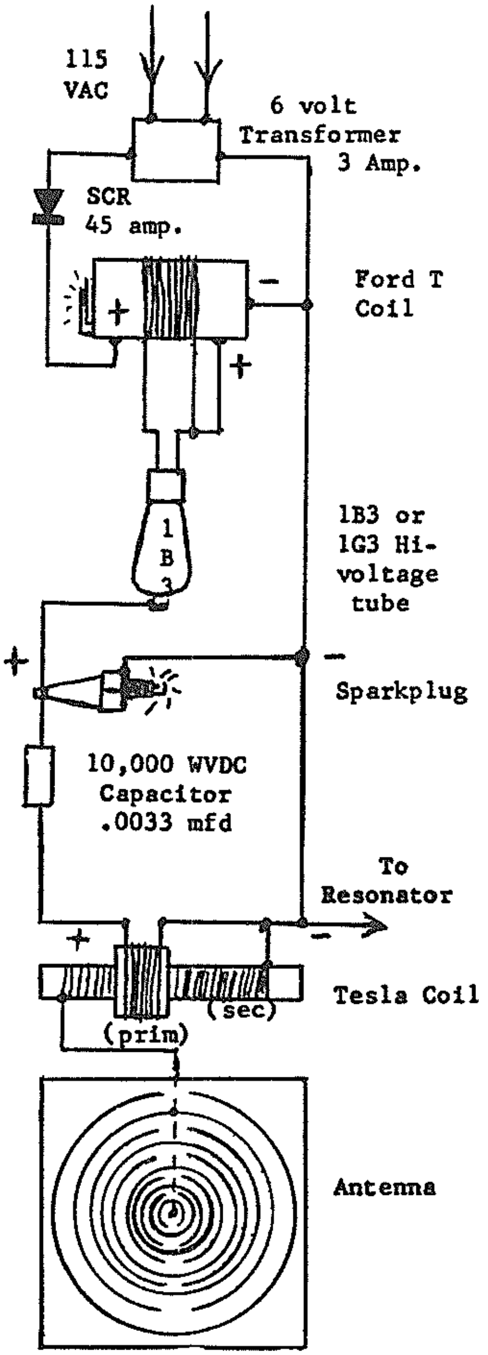

THE TRANSISTORIZED MWO

By Gerald W. Steele

Here is tentative information on the solid-state MWO. I say tentative since this is the 3rd version I've built. The first used an automobile spark coil for T2 but I found this undesirable since it lacked higher frequency harmonics. Even worse, the lower frequency ones could be harmful to living systems.

Version 2 was with a small black and white TV flyback transformer running at a slow (100 pulses per sec.) pulse rate. I eventually sped it up but could never get the required 1" spark from the Tesla coil.

This version uses a color TV flyback and a higher power discharge circuit. Nominal lowest frequency is 1.5 Khz though it could go as low as 270 Hz.

All diodes, SCR, and transistor can be substituted with universal replacement types available at electric part outlets.

As you can see, electronically it is simple. The power metering circuitry in the original MWO can also be used here if desired. I've had enough trouble maximizing the output of this device to worry about monitoring the power.

This device is excellent for use in Kirlian photography several ways: A modified circuit could be used for A, for a wide range of pulse rate settings.

There are two possible voltage sources for Kirlian use, each with unique characteristics. First is the output directly from T2 with or without the Tesla coil circuitry connected. Second is the Tesla coil output itself.

Other modifications include rectifying T2's output. The old MWO manual recommended the use of a tube rectifier out there are solid state types available.

Be sure it is a color TV high voltage rectifier. Connect as shown:

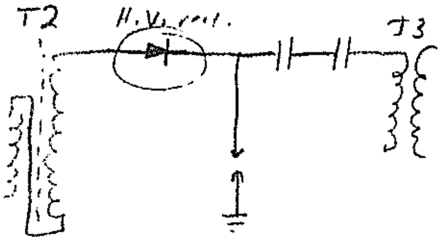

One question may arise in the builders mind: the need to heatsink SCR. No need. Not much heat generated.

Be sure R3 (or combination of resistors forming it) get ventilation. They get hot.

Always short C1 just after turning off power to work on the device. Remove short before reapplying power. The device is perfectly safe if you exercise care. R6 is a safety feature, it discharges C1 when turned off - but still short C1 when servicing, a standard safety procedure.

I am making constant improvements. The alternate values given for R3, C5 may give greater power at a higher pulse rate. I have yet to try them since 225 watt resistors are hard to find and apply. Besides it seems to work quite well as it stands. The only complaint one might have is the cost of construction principally due to the flyback transformer T2. But scavenging could help here.

Enclosing the spark gap in a plastic box could reduce its audible noise level. Ventilation of the spark gap gives more consistent operation. If the spark gap can be physically constructed above R3 the heat will produce air currents giving natural ventilation.

As I said, I am making constant improvements. Its too early to finalize it, but perhaps other associates may be interested in constructing and improving it. The lack of interrupter contacts gives a high level of reliability and consistent output.

Its too early in use for me to comment on its effects except that I think it is quite positive.

CIRCUIT DESCRIPTION:

The circuit consists of: low voltage oscillator, high voltage capacitive discharge generator, and Tesla coil.

The low voltage oscillator consists of Q1 a unijunction transistor operating as a relaxation oscillator. Frequency is variable from 270 Hz to 26 Khz with most power generated at 1500 Hz. VR1 controls this frequency. The DC Power for this circuit comes from a voltage doubler on the power transformer low voltage windings. The low voltage DC value was chosen to be as low as possible with consistent firing of the SCR.

The high DC voltage is rectified and filtered in C1. R1 is surge protection. The high DC voltage charges C5 through R3 until Q1 fires SCR1 whereby C5 is discharged through the primary of flyback transformer T2. D5 protects the SCR from high reverse voltage transients.

The secondary of T2 feeds the adjustable shunt spark gap and then through the Tesla coil resonating capacitors. Two capacitors are used here due to voltage requirements though one 15Kv capacitor could be used. From this point on it is all standard MWO fare.

Transformer T2 is expensive although any used color TV flyback could be tried. In all experimentation or adjustment the goal is maximum spark output at the Tesla coil secondary.

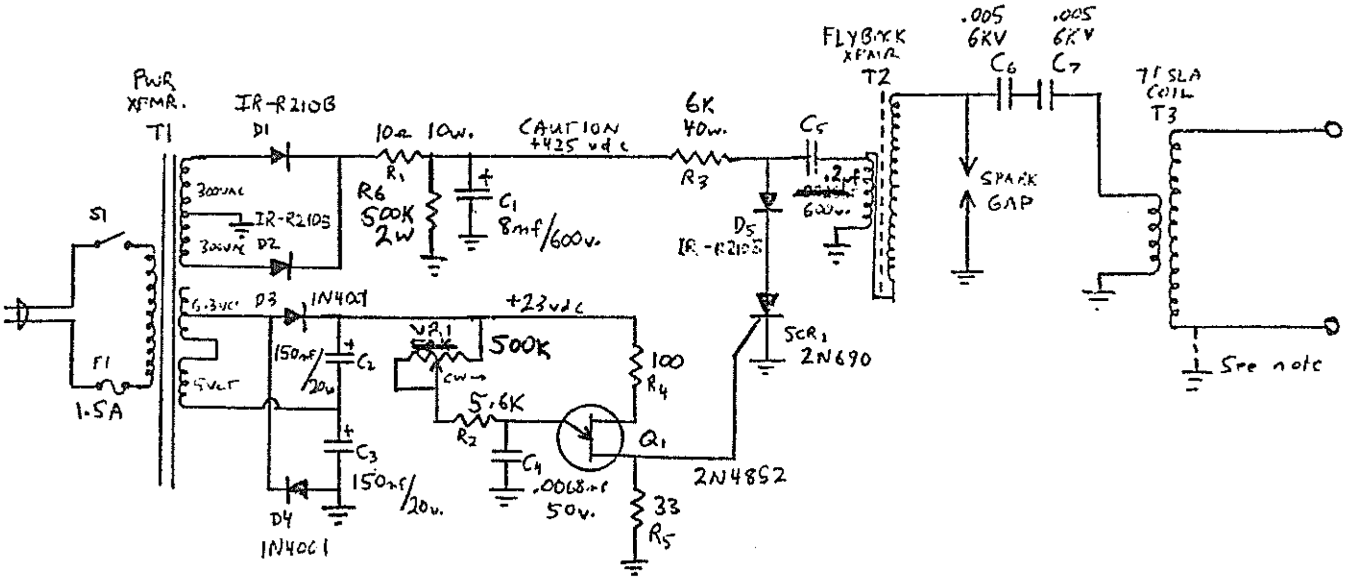
PWR XFMR. T1
IR-R210B D1
10Ω 10w. R1
CAUTION +425 vdc
6K 40w. R3
R6 500K 2W
C1 8mf/600v.
300vac
IR-R210B
300vac D2
S1
F1 1.5A
6.3vac D3 1N4001
5vac
150mf/20v C2
C3 150mf/20v.
D4 1N4001
+23vdc
VR1 500K
cw
5.6K R2
C4 .0068mf 50v.
100 R4
Q1 2N4852
33 R5
C5 .2mf 600v.
D5 IR-R210B
SCR1 2N690
FLYBACK XFMR. T2
.005 6KV C6
.005 6KV C7
SPARK GAP
TESLA COIL T3
See note

Note: Optional ground for Tesla coil secondary, if it reduces output do not use.
R3 made from four 25K 10w resistors in parallel. VR1 linear taper. C1 may or may not be polarized. C6, C7 could be single .0025/15KV capacitor. SCR1, Q1 and all diodes can be standard replacement equivelants.
MFR Type: T1 - TRIAD. T2 - Stancor HO 6246, Thordarson fly.
Alternate values: R3 - 1K ohm, 225 watts; C5 - 1 microfarad, 600v.

THE LAKHOVSKY MULTI-WAVE POWER SUPPLY
by KLARK KENT

Many of us have read the Borderland Sciences Research Foundation (BSRF) publication "The Lakhovsky MWO" in which electrical wizard Bob Beck astounds us with tales of wonder. He tells how he was informed of the location of an old Lakhovsky Multi-Wave Oscillator that had been declared "unfathomable". He was able to reconstruct the circuits and after several hours the machine worked perfectly!

Mr. Beck and fellow experimenter Ed Skilling designed and built a modern version of the MWO based upon Nikola Tesla's and Lakhovsky's original works and patents. Georges Lakhovsky, a herald of the New Age, was no doubt a skilled experimenter and investigator of the Life Energy. This is evident by his practical application of Nikola Tesla's famous coil.

Super Science has updated both Lakhovksy's original and Bob Beck's adaptation of the Multi-Wave Oscillator. Looking at the schematic diagram in Bob Beck's article, we will eliminate everything but the 12 volt battery (or 12 volt power supply), switch, Model T Ford Coil, high voltage capacitors, and the Tesla Coil. A large high voltage diode is added to rectify the output of the Model T Ford coil which leaves us with a stripped down but thoroughly functional MWO power supply.

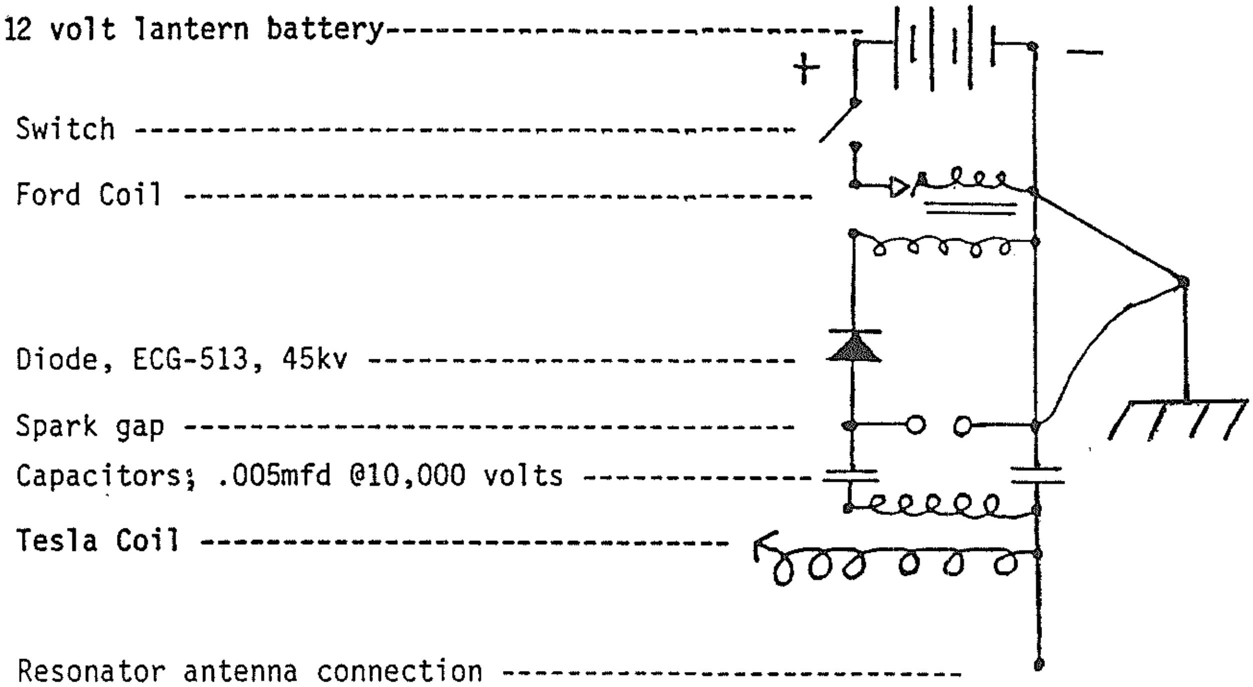

Before proceeding further, let's remember that the MWO is an experimental device and not for any healing or medical purposes! All experimenters should be in at least reasonably good health when building or experimenting with this high voltage, high frequency Tesla Coil Multi-Wave Oscillator power supply.

LET'S BUILD!

To begin our MWO Tesla Coil power supply project, we can round up the needed parts and components. Basic parts list:

-12 volt power supply; this can be a 12 volt lantern battery available from Radio Shack or any plug in power supply rated from approx 2 to 7 amperes.

-Model T Ford Coil; these old time induction coils are still available from J.C. Whitney Co, 1917 Archer Ave, PO Box 8410, Chicago, IL 60680. The MWO power supply will work adequately with the Model T Ford coil but a transistorized driver for a modern "hot rod" auto ignition coil will produce a Tesla Coil power supply with far more power!

-Switch; the on/off switch can be of any type. There is a switch already present on the plug in 12 volt power supply. I prefer to add a second switch of the key type, which prevents unauthorized use of the unit and a piece of vinyl tubing can be fitted over the metal key to create an insulated switch (this is important when turning the unit off).

-Diode, ECG-513 at 45 KV. The diode rectifies the output from the Model T Ford coil or the auto ignition coil by taking out the stray ac signals. This diode is a high voltage device rated at 45 Kilovolts (45 KV) peak inverse voltage.

-Spark Gap; an ordinary spark plug works adequately in this application. Try wrapping the spark plug in a large piece of vinyl tubing to insulate the spark gap. For a better spark gap, use a set of tungsten points. Chevy 6 cylinder 1965 model year works well to fabricate an adjustable gap.

-Capacitors; .005 mfd and rated at least at 10 KV (20 KV is better). We buy our capacitors or make them ourselves. I prefer to buy them. Check with your local electronic supply house to special order these high voltage capacitors; they will not be cheap... An alternative is to scrounge the local surplus stores to fabricate high voltage capacitors from several capacitors of a lower rating. It is easy to round up some 2500 wvdc capacitors with a 5000 volt test rating and .005 mfd capacitance. Wire the capacitors together in a series/parallel arrangement to arrive at the desired value. Series connections add the voltage rating and lessen the capacitance by division of the total number of capacitors in the series. Thus four capacitors rated at 5000 volts wired together in series yield one 20,000 volt capacitor. If the capacitance was originally .005 mfd each, now it will be .00125 mfd for the series. Thus we need to connect four of the series capacitors together in a parallel arrangement to create one 20 KV capacitor at .005 mfd. Parallel connections have no effect on the voltage rating but parallel connections add the capacitance of each individual capacitor.

CAUTION! Capacitors store electrical energy and can shock even when unit is turned off. Be sure to discharge the capacitors before touching them...

TESLA COIL

We make the Tesla Coil ourselves using the basic parameters worked out by Bob Beck in his original MWO adaptation from Georges Lakhovsky. The Super Science Tesla Coil is a close coupled, liquid filled coil that produces a 3" spark discharge when driven by a Darlington transistor and "hot rod" auto ignition coil setup.
For making the Tesla Coil, I like to use pvc pipe for the secondary coil form and acrylic tube for the primary coil form. The acrylic is transparent which permits visual inspection of the interior of the coil to observe any corona leakage or malfunction.

For the secondary winding: cut a 1 1/4" outside diameter piece of pvc pipe seven inches long. For the primary winding: cut a 2 1/2" outside diameter piece of clear acrylic tubing (1/4" wall thickness) seven inches long. Wind the secondary coil form with #32 gauge bare enameled copper wire so that the turns are tightly packed to create a winding approx six inches long with a 1/2" gap at either end. Wind the primary coil form with #14 gauge stranded and insulated copper wire. Wind exactly 18 turns and tightly pack them together. The vinyl insulation on the #14 gauge stranded wire will provide the correct spacing between turns.

Place the secondary winding inside the primary winding and use Super glue to hold the windings in place. For the terminals of the windings, use ordinary banana jacks to create a modular coil with a standard banana plug connector system. There will be tremendous electrical stress between the primary and secondary windings. We could reduce this dialetric stress by using a pancake style primary winding, but the pancake style primary takes up a lot of room. To solve the dialetric breakdown between the two windings, we insulate the space with liquid mineral oil. The liquid mineral oil is self healing should a dialetric breakdown occur. Solid forms of insulation such as paraffin wax are useless once a short circuit path has been created between the two windings.

To use the mineral oil insulation system, finish the coil first. Then drill a small hole in one end and squirt the oil in using a squeeze bottle. Then seal the hole using a piece of acrylic and Super glue. A word about Super Glue! This stuff is quite remarkable and is sold in hobby shops that specialize in flying, gas powered, model airplanes. Use accelerator to fast dry the glue. The brand name of the Super glue is "Gap Filler" and the accelerator is "Insta Set" both made by Bob Smith Industries.

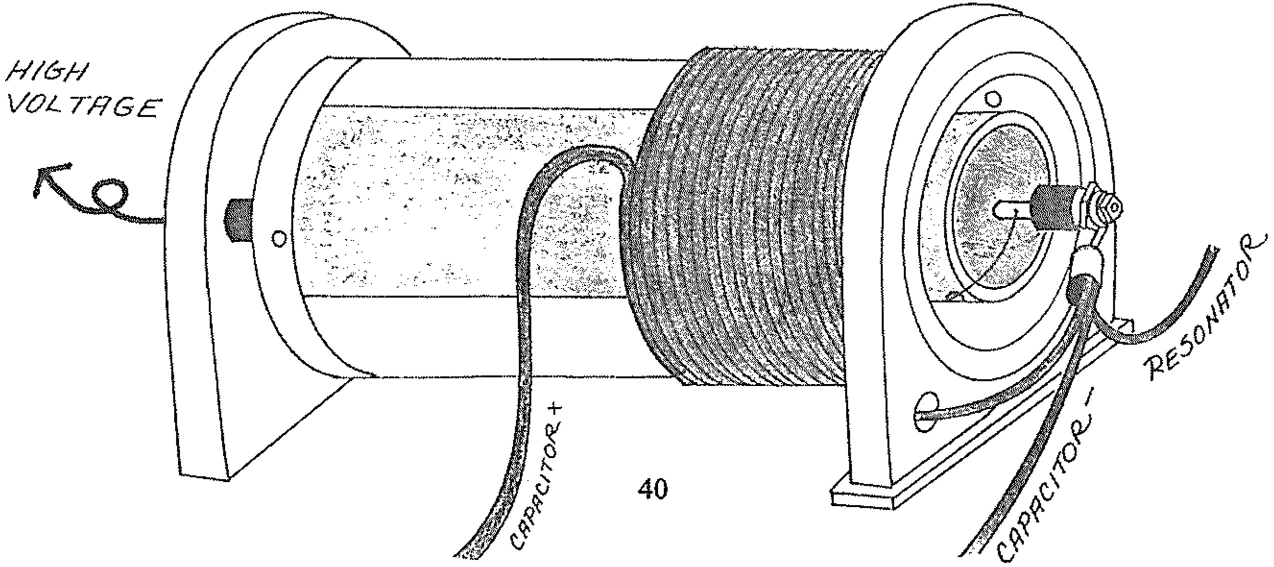

SPARK GAP

Even though an ordinary spark plug will work for a spark gap, we can do much better with an adjustable spark gap made from a set of tungsten auto ignition points. Chevy six cylinder, vintage 1960's are fine. To fabricate the adjustable spark gap, cut the base plate of the point assembly into two parts; make the cut near the spot where the two tungsten contacts normally meet. You should now have two base sections, with a contact point on each section. The reason we cut the base plate is to prevent the high voltage from jumping from the connections to the base and bypassing the breaker points. File down any sharp points where the base was cut. Mount the two tungsten points in such a way that the contact points barely touch each other and the base sections are as far apart as possible. Mount the spark gap assembly on a small piece of acrylic sheet. Wood is not suitable to mount the spark gap because the high voltage current will pass right through the wood base, possibly setting it on fire! Use a large nylon bolt to mechanically adjust the points. Mount the nylon bolt (Toilet seat bolts work well) in the side of the briefcase so that it butts against the spring loaded contact arm; this acts as a spark gap adjuster.

The advantage with an adjustable spark gap is that we can tune the gap while the MWO is in operation. The closer the two tungsten points are to each other, the faster the frequency of the spark jumping the gap. The further apart, the slower the frequency of the spark jumping the gap. Typically the Violet Ray Bulb Assembly calls for a fast frequency spark gap setting; while the High Power Radiator Antenna or Golden Ratio Antenna call for a wider gap spark setting to produce more power.

One side of the spark gap points will be at "ground" potential; while the other side will be at "high voltage" potential. Leave a metal discharge point exposed on the high voltage side of the spark gap assembly. We do this so that the entire MWO power supply may be discharged with a test wire running from the Ground Bus to the high voltage point. This provides a quick way to discharge the capacitors before troubleshooting the unit. Also, cover the spark gap assembly with a visual block made from translucent acrylic. This visual screen will block the ultraviolet light emitted by the spark jumping the gap. Leave the sides of the spark gap assembly open to allow circulation of air around the points; this will "quench" the spark gap and prevent the jumping spark from turning into a constant arc. Also, mount the cooling fan air intake near or directly above the spark gap assembly; this will serve to further quench the gap for better, more bouncy performance...

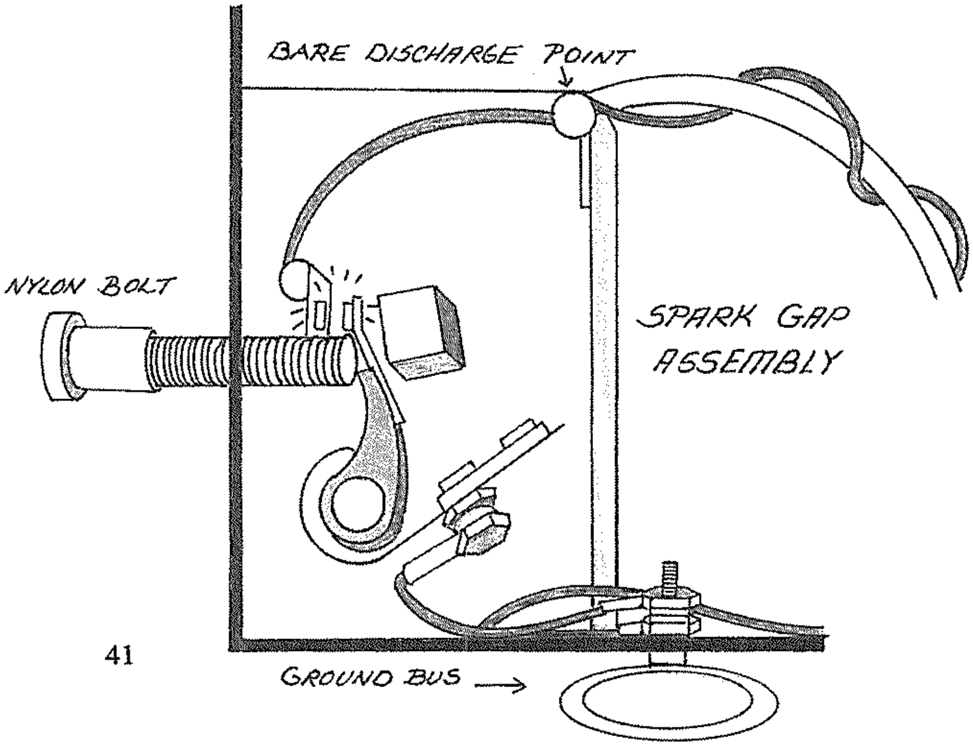

555 TIMER AND TRANSISTORIZED DRIVER

Even though the Model T Ford coil works adequately to power the MWO Tesla Coil power supply, we can really do better with a modern "hot rod" auto ignition coil. Mallory makes an aftermarket coil for drag racing that puts out approx 45 KV. The "MSD Blaster" is another aftermarket auto ignition coil that really gets the job done. Even an ordinary UC-15 replacement coil with its 30 KV rating will outperform the Model T Ford coil.

However there is one consideration to using an auto ignition coil as the primary high voltage source. Since we are working with 12 volts direct current, we need to interrupt or chop the 12 volts going into the auto coil to enable it to function as a transformer. The Model T Ford coil had a built in interrupter for this purpose. We will make a much better interrupter from a 555 timer chip and a high power Darlington transistor. The 555 timer chip will chop or interrupt the 12 volts going into the auto ignition coil from 130 times per second to 1300 times per second while the Darlington transistor handles the power. With the Model T Ford coil we were stuck with the fixed frequency of the interrupter (approx 200 Hz). With the 555 timer system, we can tune the frequency of the primary high voltage from 130 Hz to 1300 Hz. This really makes a difference when we try to get maximum performance from our Tesla Coil when it is driving the MWO Golden Ratio Radiator Antenna or driving the Violet Ray Bulb Assembly for direct stimulation.

Below is the schematic for the 555 Timer/Transistor Assembly. It is shown hooked up to the auto ignition coil. You can hardwire this circuit using a Radio Shack #276-162 project board. I use a large heatsink to mount the Darlington transistor (RCA 8766) and to mount the 50K ohm tuning potentiometer. When the MWO power supply is housed in a briefcase, the shaft of the 50K ohm potentiometer can extend through a hole in the side of the case. Go ahead & slip a short piece of vinyl tubing over the potentiometer shaft both to extend the shaft and to insulate it from high voltage.

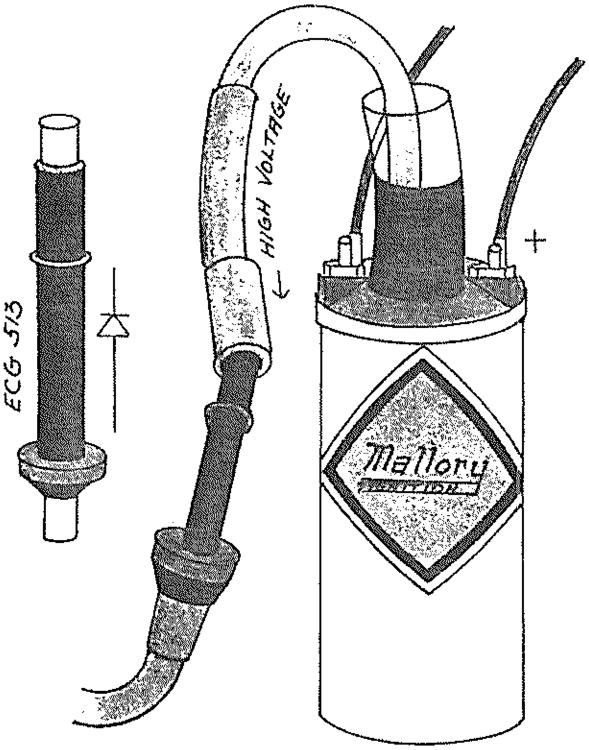

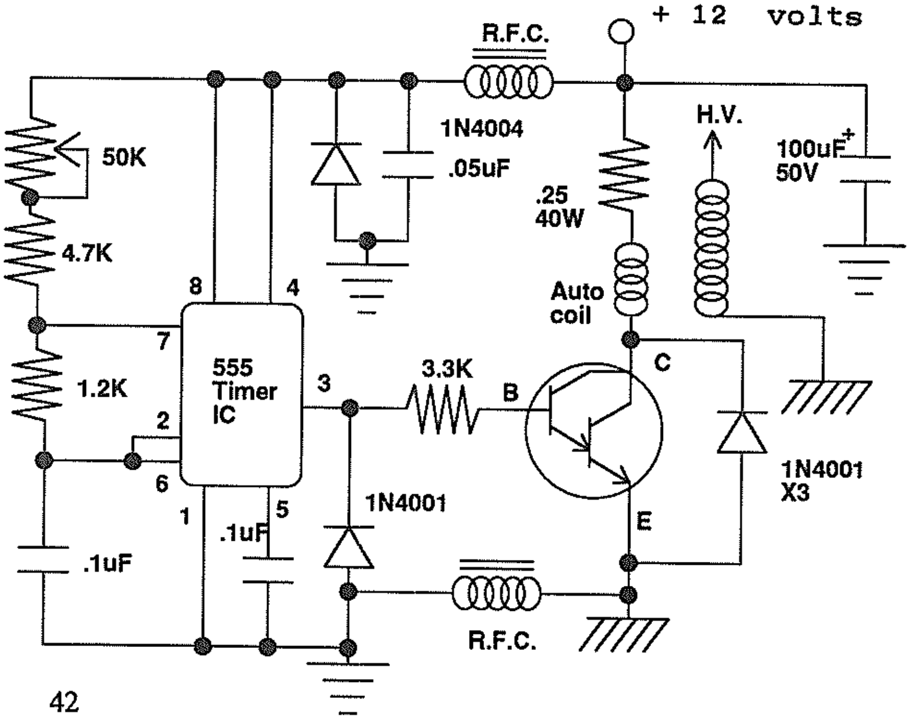

GROUND BUS

An external Ground Bus is important because it lets us reference the high voltage to the zero potential of the Earth itself. Without a true Earth ground, the negative side of the circuit can pick up a substantial voltage by induction from the high voltage potentials in the circuit. This condition is known as a "floating ground" and could result in the experimenter getting shocked. The shock itself is harmless, but the surprise that an accidental shock causes may lead the experimenter to jump back and bump his head! The Ground Bus may be connected by a long wire to a cold water pipe or to a steel fence post or to a piece of rebar driven into the ground. The Ground Bus is a handy reference point when discharging the circuit before troubleshooting.

SAFETY

Try to follow the "one hand rule" when actively adjusting the live high voltage circuit. This means keep one hand in your pocket or one hand behind your back. If an accidental shock should occur, then the electrical current does not travel across the heart when the experimenter follows the "one hand rule". Another safety procedure is to use a 12 volt lantern battery to power the circuit while it is still under construction. Once all the problems have been solved, then a more powerful 12 volt supply can be used.

TESTING THE CIRCUIT

Let the circuit run in darkness and try to observe any corona leaking from exposed high voltage points. There are two good and easy ways to stop high voltage leaks. One method is ordinary vinyl tubing: wrap wires carrying high voltage in vinyl tubing to absolutely stop them from leaking power robbing corona. Auto parts store sell expensive spark plug wire which will withstand the high voltages we are dealing with. If you use spark plug wire, be sure that it is metal wire, not "radio suppressor" wire (it's carbon fiber and impossible to solder or connect to).

Another good method of insulating high voltage is 100% silicone clear caulk! With your caulking gun, apply some 100% silicone clear caulking to exposed wires, exposed high voltage points and other places where you can observe leaking high voltage corona.

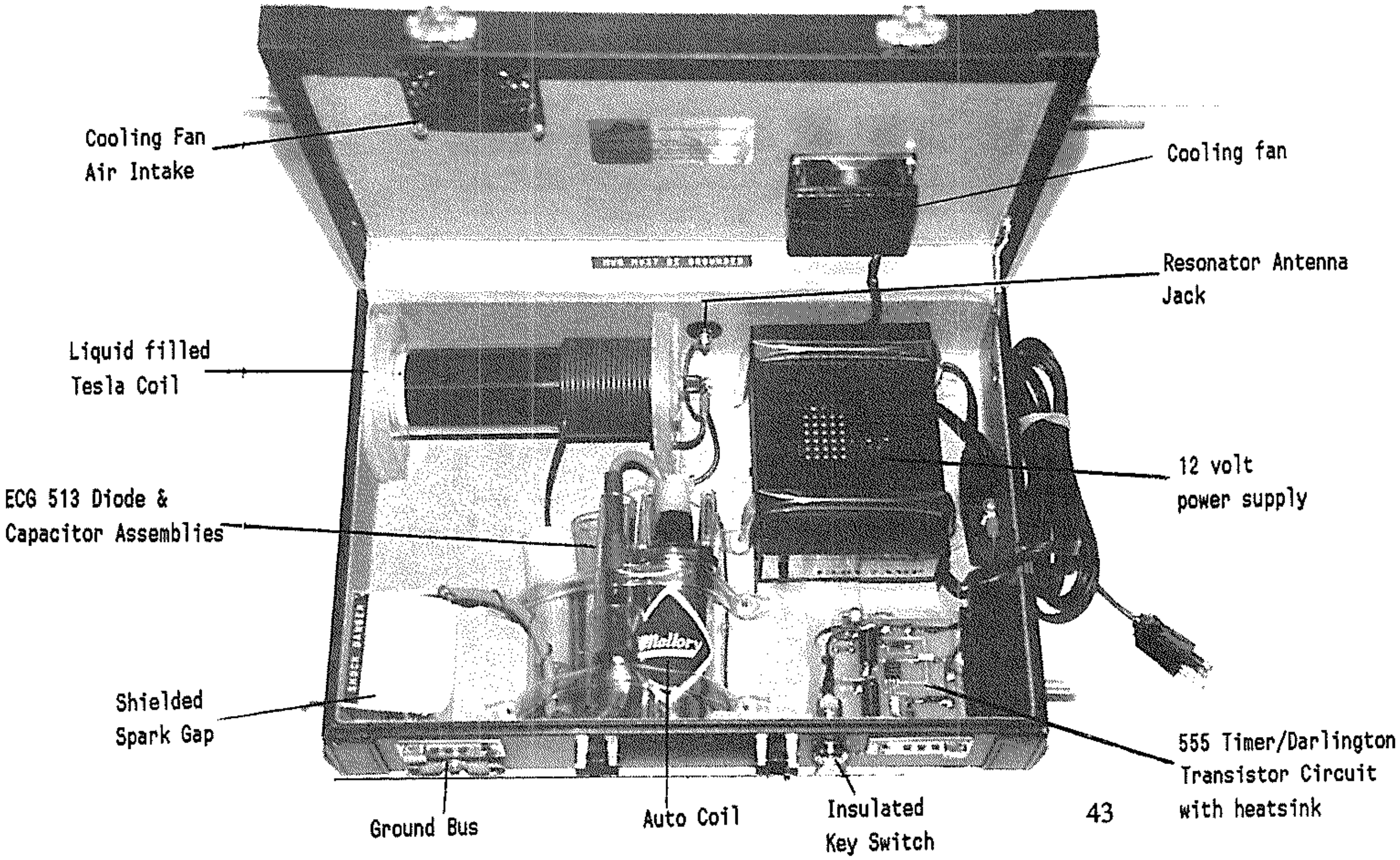
Cooling Fan
Air Intake
Cooling fan
Resonator Antenna
Jack
Liquid filled
Tesla Coil
12 volt
power supply
ECG 513 Diode &
Capacitor Assemblies
Shielded
Spark Gap
555 Timer/Darlington
Transistor Circuit
with heatsink
Ground Bus
Auto Coil
Insulated
Key Switch

LAKHOVSKY MULTI-WAVE OSCILLATOR.

HOW IT WORKS:
The MWO sets up a high voltage oscillating electromagnetic field between the two antennas. Experimental subject sits or stands in this field and every cell in the body is exposed to the resonant effect of the multiple waves. Just as one guitar string will sound when exposed to its resonant tone; so will each cell vibrate to its own natural frequency!

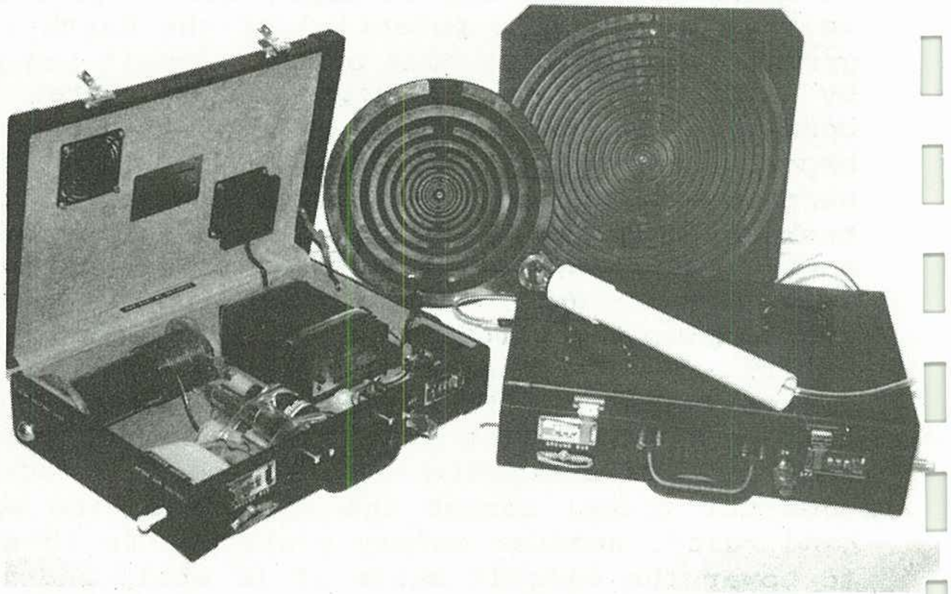

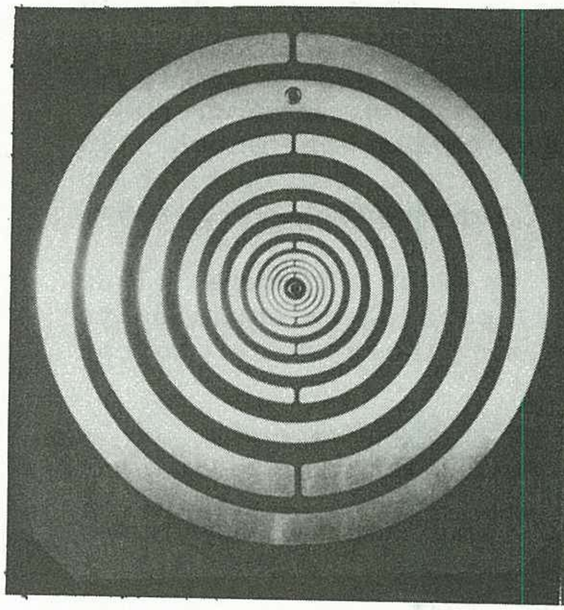

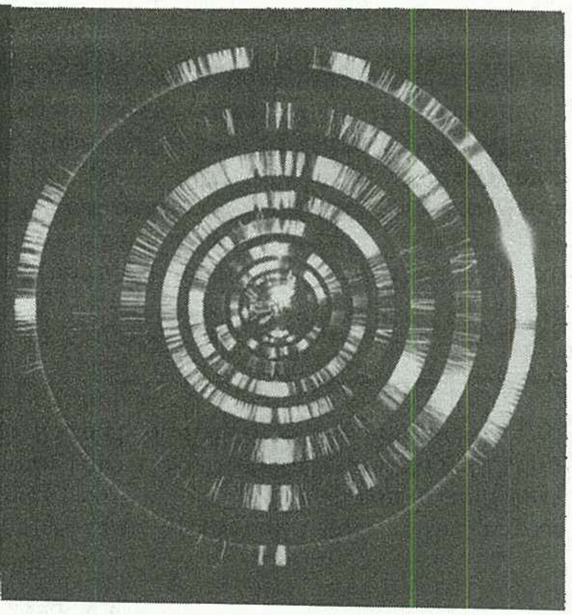

In this time lapse photo, we can see the dramatic effect of the high voltage excitation of each ring of the MWO antenna! Each ring radiates at its characteristic wavelength and frequency to cover a range from approx 200 Hz to 14 Giga Hz!

Modular construction works well for the Tesla Coil in the Super Science MWO power supply. This high voltage Tesla Coil can produce a 3" spark discharge from the smooth HV Discharge Terminal! The unit is a versatile experimenter's kit that can be used for Kirlian photography, or to create a plasma of Argon gas in the Violet Ray Bulb Assembly.

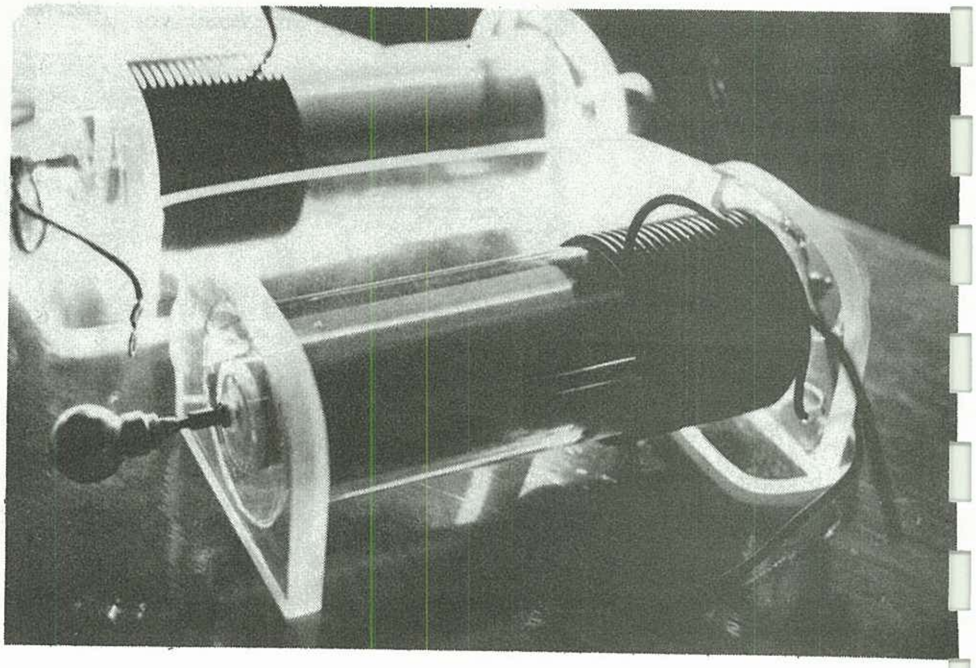

SUPER SCIENCE
P.O. Box 392
Dayton, Ohio 45409
(513) 298-7116

ANTENNA HOOKUP

Experimental subject is placed within the oscillating electrostatic field between the two antennas which are approx 4 to 6 feet apart. All metal such as heavy belt buckles is removed from the body. Clothes are optional. Sessions should last approx 20 minutes; skip at least one day between sessions to allow the body a chance to detoxify.

With the Golden Ratio Radiator Antenna, the use of the second Resonator is not required due to the unique "Scalar" properties of this design.

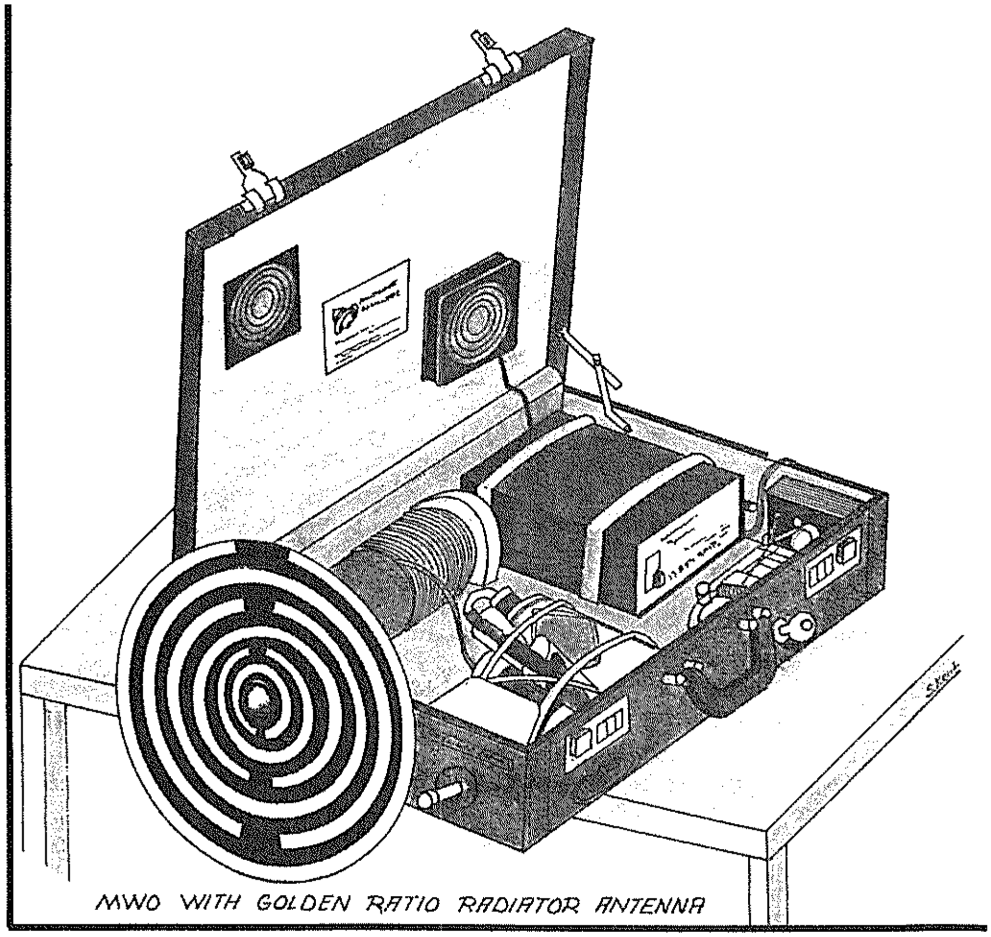

MWO WITH GOLDEN RATIO RADIATOR ANTENNA

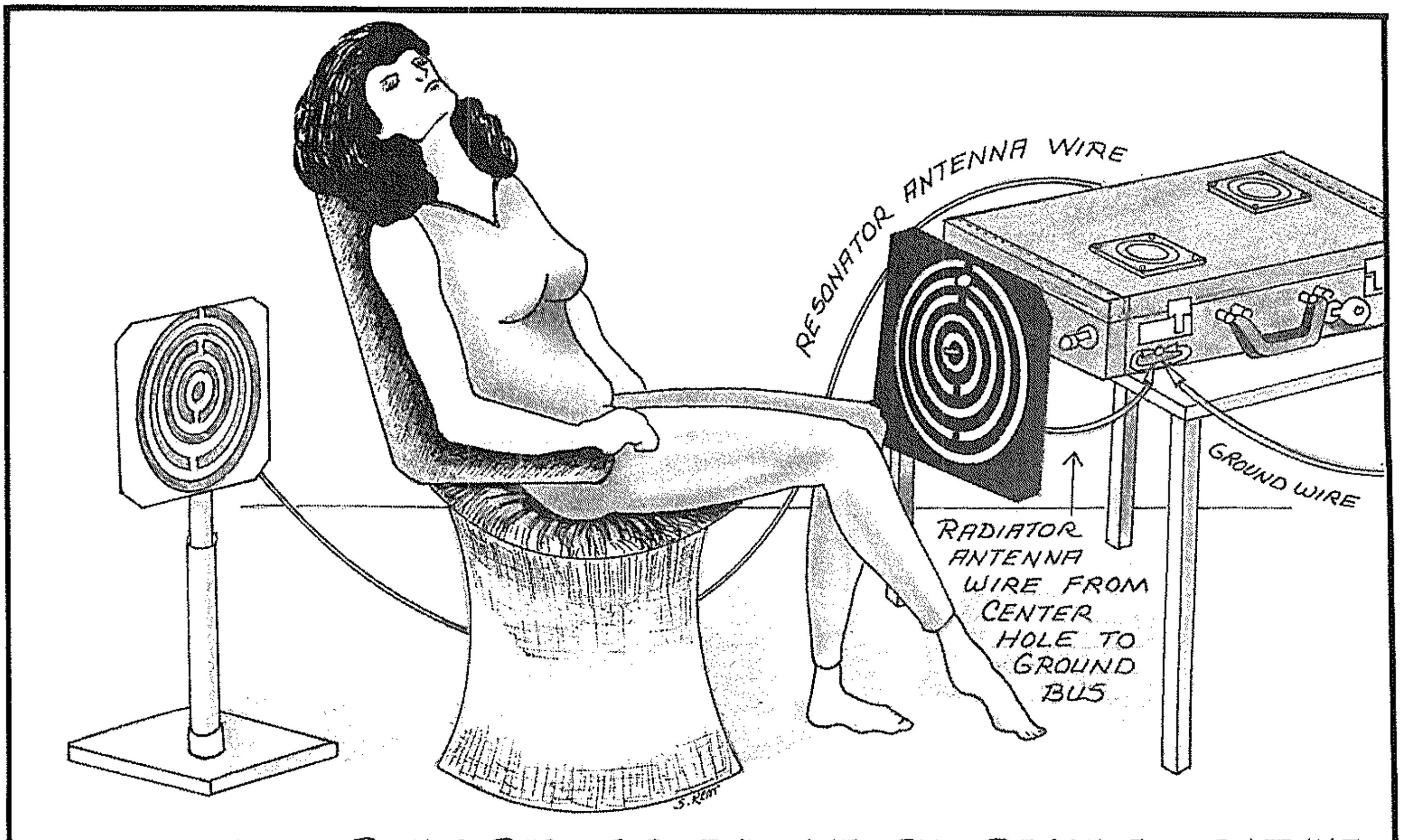
RESONATOR ANTENNA WIRE
GROUND WIRE
RADIATOR ANTENNA WIRE FROM CENTER HOLE TO GROUND BUS

MWO WITH HIGH POWER RADIATOR ANTENNA AND RESONATOR ANTENNA

From the December 1963 Journal of Borderland Research
Excerpts from:

THE WAVES THAT HEAL

The New Science of Radiobiology

by Mark Clement

edited with commentary by Riley Hansard Crabb

Bob Beck has loaned us a copy of the late Mark Clement's brief, illustrated review of Georges Lakhovsky's Multiple Wave Oscillator showing pictures of antennas, before-and-after skin cancers apparently cured by the Oscillator, and of cancerous plants returned to normal health by two means: 1. Radio-Cellular Oscillator, and 2. Single Wire Oscillating Circuits.

"On the day Hitler's mechanized hordes drove into Prague a book was published in London entitled **The Secret of Life** by Georges Lakhovsky, a Russian-born engineer resident of Paris. The book had previously appeared in French, German, Italian and Spanish translations and Lakhovsky's theories were well known on the Continent long before the English Translation was published.

"The evil designs of the Nazis soon gave rise to depressing repercussions in all spheres of human activity and Lakhovsky's remarkable work was left virtually unnoticed in this country (England) by the medical profession and the public alike. A few discerning critics, however, expressed their views with appreciative understanding. The Hospital said 'Lakhovsky's theories are so interesting from the point of view of future developments that it will repay any keen hospital officer to give them his close attention.' Science Forum said 'The admirable translation of Lakhovsky's **The Secret of Life** will be heartily welcomed in progressive circles. The general theory is one which physicists, biologists and medical men cannot afford to ignore.' But these were voices in the wilderness and the field for experimental investigation was left open to a few amateurs devoid of scientific training and lacking the necessary resources.

"The impartial investigation of claims of cure of various diseases by un-orthodox methods demands high qualities of mind and character possessed by very few men indeed. Scientific truth is no respecter of persons or of vested interests, and Lakhovsky's spectacular results in the treatment of plants and animals and human beings were inevitably bound to antagonize orthodox practitioners and exponents of popular medical cults.

"On the Continent Lakhovsky's work attracted a great deal of attention in scientific circles, particularly in Germany and Italy. Italian investigators were among the first to study Lakhovsky's theories and put them to the test in laboratories and clinics. It was, of course, to be expected that the new science of Radiobiology should make a special appeal to the intellectual heirs of Galvani, Volta and Marconi. It is somewhat depressing to observe that in England, the country of Faraday and Clerk Maxwell, Lakhovsky's theories have not yet received the attention they deserve.

REGENERATING OSCILLATING CIRCUITS

"Lakhovsky has been criticized by physicists ignorant of biology and by biologists ignorant of physics! These self-appointed censors of knowledge soon found they were confronted with a redoubtable opponent who fully realized the value of experimental evidence. Its publication resulted in causing annoyance to the custodians of infallible doctrines who made up with carping verbiage what they lacked in clarity of vision. But Lakhovsky's achievements with his oscillating circuits supported by astonishing photographs of regenerated tissues in plants and human beings could not be exploded by mere verbal fireworks, and as evidence accumulated the skepticism of hostile critics relapsed into morose silence.

"The indisputable fact remains that Lakhovsky was the first experimenter to make use of high frequency electro-magnetic waves in the domain of biology. Thus out of the application of radio-electricity to biology his work developed and gradually established the foundations of the new science of Radiobiology. The first international Congress of Radiobiology was held in Venice in 1934. A leading authority on electro-therapy, the late Dr. E.P. Cumberbatch, wrote: 'Although it had been frequently observed that the short Hertzian waves could produce heat at a distance from the transmitter the first scientific investigation from a biological point of view was made by Lakhovsky and his colleagues who published a paper in 1924 on the effects of very short waves on cancer in plants...'

"When France was occupied by German troops, Lakhovsky, being a prominent anti-Nazi, decided to leave Paris and made his way to New York where he died in 1942 at the age of 73.

"In the United States Lakhovsky's theories were put to the practical test with typical American speed. Noticing that patients treated with Lakhovsky's Multiple Wave Oscillator showed signs of rejuvenating appearance, the possibility of exploiting this side-effect proved irresistible to an enterprising 'beautician'. A film was made featuring cases after treatment with Lakhovsky's MWO, which proved to be both interesting and convincing.

"In New York Lakhovsky was approached by several hospitals with a view to testing his apparatus experimentally. The remarkable results obtained . . . in a large New York City hospital, and also by a prominent Brooklyn urologist are fully set out in the text.

"In this country (England), owing mainly to the inauspicious time at which Lakhovsky's major work was published, very little interest has been shown on the part of those best qualified to judge its merits. The medical profession, whose conservatism is the most formidable barrier to progress, has been notoriously slack in investigating the new radio-electrical methods of treating disease as originated by Lakhovsky...."

Mark Clement's observation on the formidable conservatism of the medical profession in 1949 is still true today. Effective electronic diagnosis and treatment of disease is still a closed door here in America as far as the AMA is concerned. Those brave MDs who have tried to keep medical practice abreast of developments in electronic science have been discouraged, if not actually driven out of practice. Dr Abrams is a good example, with his Oscilloclast, developed and proven effective in San Francisco about the time Lakhovsky was developing the MWO in Paris in the 1920s. Radical innovations in any line of endeavor are usually thought up and proven by rank outsiders. Georges Lakhovsky was no doctor, but an electrical engineer!

Doctors, with their own pharmacies and with massive investments in the drug industry, are not likely to show much interest in radical new ways for treating human ailments, especially if the new way is a threat to profitable practice; and, as long as Americans think they can buy health in a pill or a shot in the fanny, doctors will continue to treat them with poisonous drugs at top prices. This is good business.

BORDERLAND RESEARCH

But for those of us who would like to keep our bodies as uncontaminated -- and as healthy -- as possible, Lakhovsky's theories

It is highly significant that about the only rays and ray machines approved for general use by the American Medical Association are the destructive ones of X-rays and atomic radiation for the mis-treatment of cancer.

"Since Lakhovsky's oscillating circuits were first used for the treatment of cancerous geraniums, an experiment fully described in communications to the French Academy of Science, the same method has been successfully applied in Continental hospitals and nursing homes."

Photos of the successful treatment of cancerous geranium plants are included in Clement's book, along with the pictures of the disappearing skin cancers in human beings. So Clement describes the major features of Lakhovsky's work in this first chapter, the simple, one-coil Oscillating Circuit powered by cosmic rays, and the electric-powered Multiple Wave Oscillator. There are photos of the Multiple Wave Oscillator, showing clearly the unique circular antennas. These were made of metal tubing, separated and suspended with silk thread. The simpler, aluminum-foil printed circuit in Bob Beck's adaptation is an easy way to experiment with the idea.

THE OSCILLATING CIRCUIT

"The diagram on this page shows where the oscillating circuits should be worn, i.e. round the neck, waist, and above the elbows and knees. The effect is intensified if more than one circuit is worn. Generally speaking, the neck and waist are the favoured sites and likely to bring about the best results.

"In the light of Lakhovsky's theories the fight between the living organism and microbes is fundamentally a 'war of radiations'. If the radiations of the microbe win the cell ceases to oscillate and death is the ultimate result. If, on the other hand, radiations of the cell gain the ascendant the microbe is killed and health is preserved.

THE SECRET OF LIFE (AT THE FORM LEVEL)

"Broadly speaking, health is equivalent to oscillatory equilibrium of living cells whereas disease is characterized by oscillatory dis-equilibrium. This general principle has given rise to a vast number of experiments covering the whole field of biology. Plants, animals and human beings have been successfully treated by the application of Lakhovsky's oscillating circuits and the reader is referred to his major work The Secret of Life for a full discussion of this important subject.

Those of you who have been studying the Kabala with Mrs. Crabb and me must have been struck immediately by the metaphysical soundness of Lakhovsky's basic theory expressed above. One wonders if this Russian escapee from Communism, and from Naziism, wasn't indeed a student of the Western Mystery Tradition, perhaps even an Initiate! The Tree of Life is the groundplan of our Tradition. On the Tree we show the descent and the ascent of power, oscillating back and forth from left to right and right to left across the Middle Pillar, or balance point.

Now for the final paragraph of Chapter One. "During recent years observations on the part of several investigators have established the fact that most animals, including insects and birds, emit radiations while they are also sensitive to the influence of external electromagnetic waves. These radiations emitted by all living beings cover an indefinite range (infinite range?) and are characterized by a multitude of different wave lengths. In the immense range of existing radiations we can only perceive the luminous octave (visible light) but we know that whole gamuts of radiations exist beyond the narrow limits of the visible spectrum. The limitations of our senses prevent us from perceiving radiations of living beings and this sensory incapacity also excludes from the field of direct awareness the vast gamut of electromagnetic waves traversing our atmosphere. Yet all these radiations and waves exist and affect all forms of organic life in various ways."

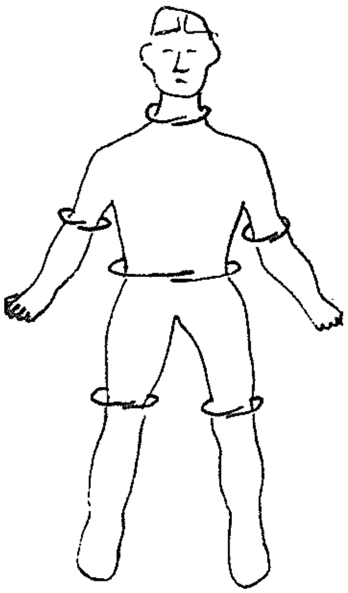

Single wire oscillating circuits around subject's neck, arms, waist and knees.

DR. RAUL ARAUJO'S REPORT

In Chapter Two Mark Clement gives "Medical Reports on Lakhovsky's Oscillating Circuits." We'll give only one, by Dr. Araujo, Physician at the Prophylactic Institute, Montevideo Uruguay. "In 1929, Dr. Araujo was suffering from a long-standing infection of the larynx which his eminent colleagues had failed to cure. While in Paris he met Lakhovsky who advised him to wear one of his oscillating circuits in the form of a collar. Although feeling rather skeptical about this unorthodox method of treatment Dr. Araujo wore the oscillating circuit for several months with the result that he fully recovered and has had no relapse since. This personal experience led him to treat his own patients with Lakhovsky's oscillating circuits. Having treated over 300 patients suffering from a wide variety of diseases Dr. Araujo succeeded in curing or greatly improving cases of rheumatism, asthma, anemia, neuralgia, gastric ulcer, colitis, glandular disturbances, sterility, genito-urinary affections, general weakness and even cancer. Dr. Araujo observed that in cases in which Lakhovsky's oscillating circuits appeared to give no definite results there was invariably a latent infection which required special treatment such as syphilis, cholecystitus, sinusitis, pyelitis, dental sepsis, etc...A full account of Dr. Araujo's Report appeared in a monograph entitled 'Cosmic Rays and Oscillating Circuits' by Dr. P. Rigaux, Paris, 1932."

CHAPTER THREE, THE MULTIPLE WAVE OSCILLATOR

Most of this has been well explained by Bob Beck in his article in the November Journal, though we might quote again in Clement's own words: "After many experiments Lakhovsky succeeded in constructing an apparatus generating an electrostatic field in which all frequencies from 3 metres to the infra-red region could be produced. Hence in this field every cell could find its own frequency in resonance... Thus a type of oscillator was obtained giving all fundamental wave lengths from 10 cms. to 400 metres, corresponding to frequencies of 750,000 to 3 milliards per second...

"Since 1931 the Multiple Wave Oscillator has been used in most European countries and in America for the treatment of many diseases including cancer. No contra-indications for the use of the apparatus nor any harmful effects on patients have ever been reported. This is in striking contrast with short-wave therapy in general, X-rays and radium whose application in the hands of experts has not infrequently been followed by the most serious consequences..."

Then follow several pages of photos of before-and-after cancer cures and medical reports from doctors. This one is typical, from Dr. Nicholas Gentile, Chief Radiologist of the Clinic for Incurables of the Apostolic Ladies of the Sacred Heart of Jesus, offer a way out of the clutches of the medical trust. There is little likelihood that the AMA will support research in this area of electronic science for years to come. There is every probability that there will be murderous opposition to any attempt to exploit the MWO on a national scale. This means that you will have to build and test your own equipment, or make friends with some electronic technician -- engineer, TV or radio repairman -- and prevail upon him or her to build it for you. This is the active kind of borderland research which requires that you get up out of your easy chair and get to work! If you live in an isolated spot, take a mail order course in basic radio construction. If you live near a high school that has an adult education program, sign up for the evening class in radio theory and bench work. I did at Barstow in 1958, so can you. If you want to be a citizen of the New Age you should make some effort to keep up to date, even if the sick industry doesn't!

What are Lakhovsky's general theories? Mark Clement sums them up neatly in Chapter One of **The Waves That Heal**:

"The fundamental principle of Lakhovsky's scientific system may be summed up in the axiom 'Every living being emits radiation'. Guided by this principle Lakhovsky was able to explain such diverse phenomena as instinct in animals, migration in birds, health, disease, and, in general, all the manifestations of organic life.

"According to Lakhovsky the nucleus of a living cell may be compared to an electrical oscillating circuit. This nucleus consists of tubular filaments and chromosomes made up of insulating material and filled with a conducting fluid containing all the mineral salts found in sea water. These filaments constitute minute oscillating circuits endowed with capacity and inductance and capable of oscillating according to a specific frequency. They are comparable to the circuits, coils and windings of radio receivers."

I believe Clement's comparison of a body cell with a radio receiver is only half-true. The cell can also be compared with a radio transmitter when it is releasing energy. We are both radio-receptive and radio-active! Lakhovsky's theory should be acceptable to any occult scientist. The push-pull idea, the positive and the negative, the Law of Opposites, is basic to all forms and to all life at this level of existence. By inductance Clement means that radio-electricity can be induced or put into every cell of your body. By capacity he means that each body cell can hold a charge of radio-electricity, just as the battery of your car can carry a charge of electricity. In the case of your body, the charge is put in through invisible but very real radio waves. These can travel unimpeded through any solid object!

"The living cell can thus play the part of a transmitter or receiver of exceedingly short radio-electric waves which give rise to high frequency currents in the circuits of the cellular nucleus. Now the vibration of an oscillating circuit is maintained by radiated energy. But where does it come from.

COSMIC RAYS

"Scientists have established the existence of highly penetrating waves to which, by reason of their universality, they have given the name of cosmic rays. These rays, generated by the interference of astral radiations, are capable of traversing dense metals and soils to a great depth. Lakhovsky has demonstrated the cellular oscillation of living organisms is maintained by cosmic radiation. According to him disease may be said to result from oscillatory dis-equilibrium due to either a deficiency or an excess of cellular radiation. An important factor in this connection is the nature of the soil. A reinforcement of cosmic rays takes place on soils which are good conductors of electricity such as clay, ferruginous and carboniferous soils.

RESONATE IN HARMONY WITH THE UNIVERSE

"To avoid the harmful effects of loss of oscillatory equilibrium it is necessary to ensure the regularization of the cosmic field round the living organism by filtration. Lakhovsky's own explanation of the problem of filtration is as follows: 'The terrestrial atmosphere is known to be impregnated by a vast range of electro-magnetic oscillations of all wave lengths and intensities as a result of unceasing electrical discharges (thunder, etc.). Besides that we know that electric motors and other electrical machines create in the atmosphere a field of auxiliary waves. Moreover, during the past 25 years the earth has been flooded with radio-telegraphic, radio-telephonic and other waves. Under such conditions it is easy to understand why any oscillating circuit of any form or dimensions is likely to find in this vast field of waves its own resonance with which it can vibrate harmoniously. It is therefore obvious that to make any oscillating circuit oscillate there is no need to have recourse to a local generator such as the Radio Cellular Oscillator used by Lakhovsky in 1924 when he first treated and cured geraniums inoculated with cancer. The oscillating circuit utilized by Lakhovsky in his second series of experiments -- a circuit having no artificial source of excitation -- proved equally effective in curing geraniums rendered cancerous by inoculation."

Yes, in 1924 Lakhovsky 'graduated' from electronics to radionics! He discovered and proved to his own satisfaction that in some ways a simple, one-turn coil of wire or fibre was as effective as the electric powered Multiple Wave Oscillator. The simple Oscillating Circuit, as Lakhovsky called it, has two advantages over the MWO. It doesn't have to be plugged in to a battery or 110 volt home power supply, and it broadcasts no interference with radio and TV sets. It acts as a step down transformer for life-positive cosmic radiation.

Clement says that "it must be particularly stressed that only Circuits made in strict accordance with Lakhovsky's directions should be used, otherwise the results will prove to be entirely negative." Unfortunately we do not have Lakhovsky's directions, and Clement certainly doesn't give them. You are still free to investigate this type of research and many have. If you want to experiment in this area then get yourself some copper wire (solid or stranded, research should be done on both) in various lengths, (8 inches is one wave length which has worked well on tomato plants) and make them into a circle with the ends overlapping, but with the metal ends insulated or separated from each other. Lengths can be determined through radiesthetic means (that is, by dowsing) or perhaps in geometric relationship with the structure of the plant. Considering the radionic and radiesthetic viewpoint geometric shapes would also have an influence and can be experimented with. Always use control samples in your experiments to see what works and what doesn't.

To get back to Clement's text: "Through the action of this constant field of radio-electric waves which emanates from the surrounding atmosphere the oscillating circuit is made to resonate like an instrument in tune with certain sounds. Thus the oscillating circuit creates a field which canalizes and filters, so to speak, the cosmic waves necessary for cellular oscillation.

THE DANGER OF BURNING X-RAYS

"Furthermore, this filtration is a general phenomenon. It has been discovered, through the investigations of Lakhovsky, that radiations such as those associated with light, ultra-violet rays, X-rays, etc., have the property of making cosmic rays act, as it were, to the advantage or disadvantage of the cells. Now the field created by an oscillating circuit is constant and emits a steady and sustained force. It invariably acts in a favorable manner on the oscillation of the body cells, facilitating by the filtration of cosmic rays, regular cellular division, a fact which explains the immunity imparted to the cells and their consequent capacity of resistance to diseases and microbes.

Rome, published in the medical review Medicina Nuova XXVI, N.5., 1935, Rome: "**Effects On The Nervous System.** Generally speaking I have observed a marked analgesic action following the application of the Multiple Wave Oscillator in all affections associated with pain, particularly in the classical cases typical of such affections. For example, a patient suffering from cephalagia, resulting from violent cranial trauma, for a duration of 10 years, was subjected to a great variety of treatments, but all in vain. The patient was definitely cured after 2 months' treatment with the MWO. Improvement was manifest after only 3 applications."

Dr. Gentile also reported the effects of the MWO on the Central Nervous System, on Chronic Inflammations, on Metabolism, on Disorders of the Cardio-Vascular System, and on miscellaneous cases. He says he adopted the "technique of localized irradiation. Generally speaking, treatment is given every 4 days for a duration of 5-15 minutes per session. This is the technique which has given me the best results after having experimented with many forms of application. For children, old people and debilitated women, a relatively weaker application is indicated. If results are not apparent after 7 or 8 applications, I continue until 15-30 applications are given at the rate of a single treatment once a week. I have never observed any ill-effects from the MWO treatment, whatever the number of applications was in any given case. Moreover, the general condition of patients is always improved, even if only temporarily, so much so that the patients themselves sometimes insist on having further treatments. . ."

This chapter on the MWO concludes with four pages of reports of notable improvement or cure by American doctors. Chapter Five is concerned with the remarkable results obtained by using the Multiple Wave Oscillator on animals. Chapter Six contains Lakhovsky's prophecy that some time in the future diagnosis of disease would be obtained by television! This is something which Dr. Ruth Drown's Radio-Vision and Homo-Vibra Ray instruments have already accomplished.

THE NOVEMBER CONVENTION TESTING PROGRAM

There were two Multiple Wave Oscillators available. These were generously furnished by Bob Beck and Ed Skilling, and were set up in the shielded seance room. Only short exposures of from 2 to 5 minutes were given, during the lunch and dinner hours on Saturday and Sunday. In that time perhaps 100 people volunteered for tests. Until a follow up inquiry is made, no statistical conclusion can be made, or even begun!

A few individual cases reported to your Director are worthy of note, including my own. By mid-morning on Saturday a spinal injury from my old, hockey-playing days was acting up. Doctors have called it a spasm in the sacro-iliac area. Ed Skilling invited me to sit in the rays from the MWO for a couple of minutes while the seance room was being readied for testing. Within a minute the spasm relaxed completely and the pain was gone. This was not a permanent change, of course, but the analgesic effect was certainly noticeable.

Mrs. Fred Rannenberg, financial secretary of Harmony Grove and a BSR Associate, told me she had had a sharp pain in the area of her left hip for over two weeks at the time of the Convention. One 3-minute exposure to the rays of the MWO removed this completely. When I attended the Grove Board meeting a week later she said the pain had still not returned.

One Associate did report a negative reaction to the MWO. She attributed it to arsenic poisoning. She had a severe case of this earlier in the year. When the machine was turned on she felt a severe pain through her right side, as though someone had jabbed a knife into her. It could be that some of the lead poison was still in her system and the stirring of the cells in the mid-section produced this reaction. It was endurable, however, and she sat out the exposure without a murmur. The pain persisted that night and she had to put cold compresses on her side to get to sleep. The next morning she felt fine and was eager to take another exposure to the multiple waves. This time there was no negative reaction.

The most startling subject of all was the woman who had been awake all Saturday night with a migraine headache, an old and recurring ailment. By Sunday morning, Nov. 10th, she was sick and nauseated. It was suggested that before taking her to a doctor or hospital she be given a short exposure to MWO rays. Within minutes all pain was gone and she returned to her cottage to lay down and sleep like a baby.

THE PSYCHEDELIC EFFECT

Ed Skilling is the only I know who has stayed in the MWO field or rays for longer than 15 minutes. While testing and adjusting his equipment at his microphone factory several weeks ago he deliberately stayed between the antennas for an hour. At the end of that time he was beginning to feel lightheaded and dreamy he thought it a good idea to get out before losing consciousness altogether.

An even more spectacular possibility is the Time-Machine effect in connection with rejuvenation. If, as Bob Beck suggested in his article, the MWO tends to take the body cell back down the time-track to a more youthful, more vital period, what would happen to a person who stayed in the field of the machine for four hours? Would his vibratory rate be speeded up to where he would disappear completely? Like the hero of H.G. Wells' "Time Machine"? Anybody want to try it? It seems, from hints George Van Tassel has been dropping over the years, that his Integratron at Giant Rock, California might do something like this -- if and when he gets it working! As I recall, both George and the Canadian Saucerian, the late W.B. Smith, spoke of the four dimensions of electricity: 1. is the electricity moving in one direction along a copper wire, 2. is the magnetic field around the wire, 3. is the static charge created by the first two effects, and 4. is Time -- or the "tempic field" as Smith called it. The Time Field equates with Space, where consciousness is everywhere present. Seems to me Lakhovsky's Multiple Wave Oscillator is a gadget which gives us the opportunity -- for the first time -- of approaching the Time Field with scientifically acceptable equipment, operating in known frequencies. Remember, that with its circular antennas it operates on the vortex principle. Each individual antenna in the group is a segment of a 60 degree cone. The 1/2 inch gaps, at alternate 180 degrees from each other, conform to the alternating principle and give a damped or pulsed wave.

THE SPACE AGE EFFECT

Copies of our November 1963 **Journal** have gone all over the world. Each day's mail brings new orders. We are confident that Lakhovsky's invention, far ahead of its time in the 1920s and 30s, has been given a new lease on life and will be pushed to its ultimate development on every continent. There will be vigorous opposition from the right-wing conservatives and reactionaries. Their resistance will furnish the platform from which electronic "medicine" will eventually sweep the world.

I see the day when a device, similar to the MWO, will be a standard part of every home bathroom or cleansing center. You will consider it as necessary and as routine to take an electronic showerbath, as to take a water bath. Home from work, tired, dirty and discouraged, you strip, flip on the MWO switch, and while shaving, doing your hair or whatever, re-balance your whole system with a two or three minutes exposure to the pulsing waves from the oscillator as they flood the whole room. For the sale of the house of the future it will be as important for the ad to say "shielded bath with MWO" as it will to say "tile bath with radiant heat."

Mr. Skilling told me he was tired and

"dragged out" after processing dozens of volunteer "guinea pigs" late Saturday afternoon. Before shutting down the equipment he sat in the nail-less wooden chair between the antennas and gave himself a three minute shot of pulsed waves. Refreshed and revived he looked forward to a good dinner and the full evening program.

I don't know of a better Christmas gift we could have given the Associates than the idea and the plans for the Lakhovsky MWO, researched and so generously offered us by Bob Beck in the true borderland science tradition. Now it is up to you to carry it forward in your own individual way. We welcome your comments, observations and research data as you develop equipment and a testing program.

AND NOTHING HAPPENED!

"No doubt you are deluged with correspondence about your articles on the Lakhovsky Rejuvenation Machine. Since you said it is possible for any intelligent 16-year-old to build one that is exactly that which I have had to rely on. Fortunately, my son is fairly well informed on electronics but obviously no expert at his age, 16. He followed your diagram but instead of battery power he used an electric train transformer, which he assures me, is equally efficient. He also found he had to make one change in the hookup of the Tesla Coil. I have taken four exposures and the results! Nil!"

Maybe this lady expected to be rejuvenated back to her early twenties. As far as we know, no such claims have ever been made for the Lakhovsky gadgets -- though it would be interesting if it happened! As for power supply to the Model T Ford coil, any 6, 8, 10 or 12 volt battery or transformer should do, depending on what you have available and where you intend to do your experimenting. We understand the old Model T magneto delivered as high as 18 volts to the four coils. One handy source of 6 or 12 volt electricity is the cigarette lighter on the dash of the car. Plug in there for power when you want to expose yourself while driving or out in the country.

"I had been fighting a virus for several days and the day after the first exposure I felt a great deal better, but the second day I was right back where I had been. But that was the pattern the virus had been taking. One day I'd feel fairly normal, the next miserable again. So it is difficult to know if the Lakhovsky machine could be credited with the temporary improvement."

That's right, ma'am, some people show little or no reaction to the mytogenic waves put out by the Lakhovsky instrument. The results are unpredictable. Perhaps that's why the research was pushed no further in the 1930s by authorities here in the United States. You, like most people, think that health can be bought in a package. Lakhovsky didn't agree with this erroneous conclusion. His theory was that health is a condition of balance of all the forces of the body. Ill health is a condition of imbalance caused by your misbehavior. If you continue to misbehave, any force applied from the outside, whether it be mytogenic waves, manipulation, drugs, medicine or surgery, will only bring temporary relief. Pain is an insistent teacher, urging you to change your mind; but if you can't change your mind, how can you change your body?

"One possible reaction to the exposure I can report is that I took the first one just before going to bed at night. It took me three hours to get to sleep. Several things puzzle me about the machine we have constructed. First of all, although we have experimented with a Tesla coil of the size your diagram calls for, as well as one almost twice the size and strength, we cannot get a two-inch spark from the antenna or resonator, 1/2 to 3/4 of an inch is the absolute maximum."

The mention of the two-inch spark was Mr. Beck's error. 1/2 to 3/4 inch arcing is all this circuitry will produce and that is adequate to affect notable changes in many forms of cell life. More power might produce faster changes, and much harm! Be reasonable, a living cell is so small as to be invisible without a microscope. Only minute amounts of energy are necessary to affect a cell, to stimulate it or to kill it.

"Second, under no conditions do we get the 'bluish glow of brush discharge' from the antenna that you mention; although we have no trouble lighting up a two-foot fluorescent lamp."

Operate your Oscillator in total darkness and you should see some brush discharge at some points on the antenna. True this redesigned equipment puts out nothing like the five-inch brush discharge visible in the photo of some of the original Lakhovsky equipment in Mark Clement's book, "Waves That Heal", and Lakhovsky's antenna appears to be at least two feet across.

But power is not the answer to balanced cell life! Dr. Abrams discovered this with his early electronic experiments in San Francisco in the 1920s. It is the quality and the variety of the pulsed waves that counts! In any event weaker equipment can be compensated by longer exposure, and this may be more beneficial to living matter.

"And thirdly, since we have not constructed any outer shield for the machine, such as a Faraday cage of copper screen, I have been monitoring the operation on our radios and television. On the radio, one notes an annoying and loud buzz, but on T.V. except for a couple of light horizontal lines and a very slight hum, there is no serious effect. So I doubt that we're bothering our neighbors very much. In any case, we try to use the oscillator very early or very late in the day, when few people are using their radios or TV. I have sent for Clement's book, 'Waves That Heal,' and Lakhovsky's 'The Secret of Life', not only to better understand what we are trying to work with but also in the anxious hope that I will find out why our machine does not have the affect it apparently should have."

Let me assure you, good lady, there is no single cure-all for man's ills. The clue to better health is in the profound statement of the Savior of our race: "As a man thinketh in his heart, so is he." Pain is still the great teacher in the world and will be for a long, long time to come. I don't know what your physical condition is and it's none of my business; that's between you and your doctor. If you have a long history of frustration and thwarted ambition, going back to childhood and even into past lives, this mass of energy of crystallized thought-forms, is not going to be broken up and dissipated overnight. If you have consciously and willfully flouted Cosmic laws, or have turned them to your own selfish ends, the consequences of this misbehavior will have to be worked out at this level. The occult scientist calls this Fixed Karma. In this connection, one of my favorite teachers, Dion Fortune, reminds us: "What cannot be cured must be endured."

If the Lakhovsky oscillator does have an analgesic affect -- as it appears to have had in some cases -- it has this advantage over pain-relieving drugs; it does not leave a residue of poison in our already over-loaded liver and kidneys!

WORD OF A SIMILAR INSTRUMENT UNDER DEVELOPMENT

We have a second letter from this same lady Associate, also worthy of note in connection with the Lakhovsky Multi-Wave Oscillator.

"First off, when my brother received the machine we had built for him, using the battery indicated in the schematic, he felt the power source might be too weak. It so happens he has a business contact with one of the largest electronic instrument companies in the area where he lives. He decided to go straight to the president and seek his advice. With the whole idea placed before him this gentleman said, 'This is very interesting. I want you to meet one of our physicists who is working on a very similar instrument.' A meeting with the physicist revealed two interesting pieces of information: first, the physicist had never heard of Lakhovsky or his machine; second as soon as this new instrument was completed, arrangements had been made to test it in a large nearby hospital! Is this latter not a cheering bit of news?"

Both cheering and discouraging. The basic principle discovered by Lakhovsky is certainly available to the mental probing of other sincere researchers, so let us keep it alive by all means. But barring a revolution in the ultra-conservative leadership of the American Medical Association, no electronic developments will be allowed to threaten the high-profit position of the drug industry.

"In their discussion of the power source, the physicist suggested my brother use a wet-cell car battery, #50, 39 plates, 70 A.H. We are using one too and it makes a much better spark from the antenna, although it still is no greater than 1/2 to 3/4 inch. It does seem stronger. Of course the machine is not so readily portable with such a heavy battery. My brother had only one exposure at this writing but he reports a noticeable improvement in the arthritic pain bothering him in his right arm for months. On the other hand, the report on the use of my own machine is somewhat less encouraging. The only positive results concern my husband. After his first ten-minute exposure, within a few hours, a painful bursur in his shoulder cleared up. He has been subject to off-again, on-again attacks of bursitis there for some 15 years; for which he had taken medication, short wave, and even X-ray treatments, when the attacks were severe. He has since taken another exposure, but neither this nor the first one seemed to help an itching-skin allergy which has bothered him for some years.

NOT THE FOUNTAIN OF YOUTH SHE EXPECTED!

"As for myself, I have taken at least a half-dozen exposures and cannot in all honesty detect the slightest difference from them. I'm not sure if it is a matter of 'suggestion' or a desire to prove the capabilities of the machine. On the contrary, I notice such strange contradictions as the fact that although I took an exposure yesterday morning, in the evening the fourth and fifth fingers on my right hand were painful, from some kind of arthritis or rheumatism. I had never had anything like this before. Does it not seem odd that it should develop on the day of the exposure?"

Not at all! Any positive reaction to exposure to the mytogenic rays of the Lakhovsky oscillator indicates that it is working! Mr. Lakhovsky's theory is that the waves or oscillations excite the cell, causing it to release mor energy, also to shake off accumulated poisons and/or tensions. If your system is loaded down with tensions from daily living in America, if your system is loaded down with poisons from processed foods and contaminated air and water, you can expect painful reactions when the junk breaks free. I believe that is what sickness is the body struggling to get rid of poison. For you, things may get worse before they get better!

HOW OFTEN AND HOW LONG?

"With the mention of two successive exposures in two days I get to what is the most important question in the use of the MWO. No one has explained why exposures are suggested at four-day intervals. Is it harmful or useless to take them more frequently? When one is uncomfortable or in pain, one is anxious to rid oneself of the discomfort as quickly as possible. And to have to wait four more days for additional improvement seems unnecessary if there is no harm in the exposures. I have exposed myself at more frequent intervals and still have observed no reactions?"

The suggested exposure rate of fifteen minutes every four days came from Dr. Gentile's extensive researches with Lakhovsky equipment in Rome in the 1930s. This is a typical professional approach to treatment of disease by doctors. From long experience with drugs and other proven forms of therapy, they know that it takes time for the human organism to adjust itself to radical changes -- without harming the organism.

Whether or not this holds true for the operation and effect of the Lakhovsky machine remains to be proven and you'll have to be your own guinea pig. It's your machine; it's your body; it's your life. You can take an exposure every hour on the hour if you like; stay in all day. Who's going to stop you? This is borderland research and you are a borderland scientist, looking for new knowledge which may be of benefit to yourself and the human race, judging from the last lines of your letter.

"I am so excited and hopeful of the efficacy of this MWO machine that I do not want to leave any avenue unexplored that may possibly help in understanding its use and making it freely available to anybody I can possibly help. May I take this opportunity to tell you how tremendously enthusiastic and interested I am in BSRA, not only what you are trying to do but what you are accomplishing, too. From what you imply of the financial situation, it seems to be largely a labor of love and service on your part, for which you deserve great praise and the heartfelt thanks of every BSRA member and others whose lives they touch. Please accept this very, very small donation of the enclosed check, as a symbol of my appreciation."

The donation is most welcome and will be put toward the expenses of carrying on this work, also the words of appreciation. The lady was also most curious about the Lakhovsky circuits, those single bands of wire which are placed around the waist, or knees, or ankles, wrist or neck, and seem to have a subtle but very real affect on the wearers. We have some information on the Lakhovsky Circuits or Belts from Associate A.E. Powell, a retired lieutenant colonel of the British Indian Army:

THE OSCILLATING CIRCUIT

"The Belts can be made by almost anyone, for a few cents. At most hardware stores, for a few cents per foot, one can buy twin electric cable insulated with white rubber, used for lamps, etc. Get the smaller size 1/4 inch wide, having 42 strands of copper wire. Slit an end with a knife and the two wires easily tear apart.

"For the button, get white sheet plastic about 1/8th inch thick, saw it into pieces about 1½ by 5/8th inches, and drill near each end a hole large enough to take two wires. Round the corners of the plastic and bevel edges. Thread the wires, from opposite directions, up through one hold and down through the other, leaving about half an inch projecting beyond each end of the button. Where a circuit can be slipped off the body without unbuttoning, you may overlap the ends of the wire belt an inch or two, bind with fine string and smear with Duco cement for strength.

"Parallelism and overlap of the ends are important. This builds up capacitance, produces inductance and generates an alternating current. Circuits may be worn, as many as you please, round chest, waist, neck, arms, wrists, legs, ankles, neck or head continuously. Water will do them no harm.

"Gladly will I supply Circuits, in any quantity, with buttons or bound ends. To avoid trouble with the Food and Drug people, I make no charge. Donations, however, are acceptable, and go intact into my Fund for the Sick. For long this has been doing fine work and I'm sure you will approve. For reasons stated I make no claims whatsoever for the Circuits, and no guarantee or promise of any kind. If you want to give them a trial, that is wholly your affair." Col. A.E. Powell (Col. Powell died in March 1969).

MWO USED TWENTY YEARS AGO

"The recent article in the BSRA Journal pertaining to the Lakhovsky Oscillator is of special interest to me. First, I am an engineer, Physician and Surgeon, and a Scientific Consultant for the U.S. Government. (Technical Intelligence). I am not in any manner snooping for the government. This is completely personal. We built several oscillators in 1944. These were used in my sanatorium in Dallas where we had some startling results. As we did not have the proper facilities to research the instruments adequately, we became more and more conservative. Many patients would show reactions, resembling the old fashioned Flu, after a few days following exposure to the oscillator. I realized we were using too much power, so cut the level several times with good results; but we did not completely eliminate the side effects. In 1946 I was sent to Germany for the Government to explore Electro-biology and Biophysics in its application to electro-medical equipment. I felt it wise to destroy the machines so injudicious use would not be indulged. The possibilities have intrigued me ever since. It looks as though you have solved most if not all of the problems involved. Hence my renewed interest. The excess power upset the enzyme complex of the body, or as some felt, distorted the double electric layers of charge carried on the cells, hence allowing for the penetration of virii directly into the cells (virii operate within the cells -- bacteria operate between the cells). Cells as well as organs function in a force field so mild oscillating currents not only work through resonance but seem to regenerate these force fields. Here is where application to cancer can be optimistically explored."

Speak for your own state, Doc! Here in California the medical association has decreed that only four forms of treatment for cancer are legal; drugs, surgery, burning rays and death.

"Proof of energy flow from the fingers and thumb is here and must be recognized. We have demonstrated it by its effect on Yeast cells, viewed under the microscope (Girwich research). Whether it comes from the conscious rather than the subconscious mind is harder to prove, but I am impelled to believe it does. This may interest you. Years ago we experimented with crystallization. Precipitated water vapor to snowflakes in a chamber with electrodes planted in two sides. When we threw a current into the chamber at a given frequency the flakes were all the same. Changing the wave length changed the crystallization pattern. Each frequency had its own pattern. Here is the explanation of energy absorption by the chemical used, and the control of the crystalline patterns by the nature of the energy flowing."

Naturopaths are in 100% disagreement with the doctor's above explanation of the side affects of his use of the MWO on his patients. If I understand the theory correctly, according the natural view of colds, flu, virus infections, etc., these are attempts by the human organism to throw off mucous, to eliminate intolerable accumulations of poison from the system. Exposure to the field created by the MWO tends to raise the general tone of the whole system by stimulation. The raised vibrations release an unusual amount of mucous and other junk into the elimination channels, and until this is eliminated one would have a cold or flu and the usual aches and pains which generally go with such a condition.

Another doctor stopped by here not long ago to tell of his sad experience in trying to use spark-gap electronic equipment in his medical practice "for those suffering patients not helped by drugs." This powerful, high-frequency machine produced a two-inch spark! Every time he turned it on, everyone in town seemed to know it and the phone would ring. Even a Faraday cage didn't shield the output of this monster. Finally he was called before the county medical board to explain what he was doing. The board presented a staggering combination of fatuous conceit and stubborn ignorance. "They didn't want to know anything about the affects of radio waves on the body and didn't care!"

CURING CANCEROUS PLANTS WITH ULTRA RADIO FREQUENCIES

by Georges Lakhovsky

From Radio News February 1925

"The new application of short wavelength oscillations described in the following article" wrote the Editor of Radio News, "is one of the most important ever found, and we are pleased to present to our readers this article by Mr. Georges Lakhovsky, the well known French scientist. The experiments described were carried out in collaboration with several doctors and scientists of high standing and it seems from results obtained, that the very high frequency treatments will play an important role in the future. We shall publish in a later issue another article on this new system together with the effects of ultra radio frequencies in the treatment of diseases in human beings."

Since November 1923 I have published in various technical and radio publications, several papers in which I explained my theory that the instinct or special feeling, which permits birds to direct themselves in space, is only the results of the emission and reception of rays by living beings. While developing this theory, I explained how thoroughly I was convinced that science will discover, some day, not only the nature of microbes by the radiation which they produce, but also a method of killing disease bacilli within the human body by means of the proper radiations.

The researches I have made by means of special apparatus have shown such results, that I believe my theory is correct. This theory is that life is born from radiations, kept going by radiation and suppressed by any accident producing the destruction of the oscillatory equilibrium, especially by the radiations of certain microbes, which suppress those of the weaker cells.

Before going any further in our reasoning, it is necessary, in order to present the facts to the uninitiated reader, to imagine what oscillations really are. The motion of a pendulum will be used for this explanation. When a pendulum is displaced from the position of equilibrium, it moves back and forth producing what are known as ochrone oscillations, until the energy stored is entirely exhausted. By means of a motor, a spring, or an electro-magnet, it is possible to keep the motion of the pendulum of constant amplitude, producing undamped oscillations. If, on the contrary, the source of power is removed, the oscillations die down and it is necessary not only to re-apply the power sustaining the oscillations, but also to furnish additional energy to start the pendulum in motion. This oscillation of a pendulum reproduces exactly what happens in the cells of a living being.

Our organs are composed of cells formed of protoplasm containing various mineral matters and acids, such as iron, chloride, phosphorous, etc. It is by the combination of these elements that the cells detect outside waves and vibrate continuously at a very high frequency, probably higher than the period of X-rays or over all other vibrations known and measured today (1925). The amplitude of cell oscillations must reach a certain value, in order that the organism be strong enough to repulse the destructive vibrations from certain microbes.

COSMIC RAYS AND CELL LIFE

The astrophysicians are actually carrying out experiments of great interest on the existence of vibrations, which have been called penetration rays and of which the frequency is higher than that of X-rays and of the alpha, beta and gamma rays of radium. Such rays, according to the theory, are produced by the earth itself and some others come from outer space. Some accurate measurements have proved the correctness of this theory. Therefore, it is quite permissible to believe that these penetration rays, or at least some of them, produce the vibratory motion of living cells and consequently their life.

For instance, let us suppose a cell vibrates at a certain frequency and a microbe vibrates at a different frequency; the microbe begins to fight the cell, and sickness is started. If the cell cannot repel the stronger vibrations and if the amplitude of its own vibration decreases, the microbe gains and its vibrations in decreasing and stopping those of the cells, bring dangerous sickness or death. If, on the contrary, the living cell is started vibrating with the proper amplitude by inside or outside causes, the oscillatory attack is repulsed. Such is my theory.

The problem is somewhat similar to the situation in which a rescuer finds himself when, coming to help a friend in a dangerous situation, finds himself fighting hand to hand against strong aggressors. The rescuer does not dare to fire his gun; fearing to harm his friend mixed up with the aggressors in the melee. Similarly, microbes and healthy cells are exposed to electric or radio-active action, which could be used to destroy the unwanted rays and it is difficult to suppress them without harming or killing at the same time, the cells which are to be treated. In fact, since Pasteur, scientists have been constantly searching for means of destroying microbes. The great difficulty with all methods found was that in destroying the bacillae, the cell was attacked too. The experience gained in cancer and tuberculosis treated with radium, X-ray, or ultra-violet rays, shows how difficult is the work of the investigators.

A RADICALLY NEW METHOD

The remedy in my opinion is not to kill the microbes in contact with the healthy cells, but to reinforce the oscillations of the cell either directly by reinforcing the radio activity of the blood or in producing on the cells a direct action by means of the proper rays. During January 1924 I began to build, according to this theory, and with the purpose of therapeutic applications, an apparatus which I have called Radio-Cellular-Oscillator, with the firm belief that the cells vibrating at extremely short wave-lengths would find their own in the Hertzian waves, which have the properties of producing extremely short harmonics. The cell with very weak vibrations, when placed in the field of multiple radiations, finds its own frequency and starts again to oscillate normally through the phenomenon of resonance. This type of vibration which I propose to use, is harmless, unlike those of X-rays and radium. Their application, therefore, does not present any danger to the operator.

I exposed in front of my apparatus, during long periods, a certain number of microbes in culture, which developed themselves normally. I, myself, have never felt the effect of these ultra radio frequencies, although I remained for a great many days near the apparatus, during the treatment applied to the living cells. It is only when two living beings such as a cell and a microbe, are in contact, that the rays produced by the Radio-Cellular-Oscillator have any direct effect upon cellular structure.

WORK IN A FRENCH HOSPITAL

The experiments which I carried out at the Salpetriere Hospital in Paris, in the service and with the collaboration of Prof. Gosset, were made with plants inoculated with cancer, and the results were described in a paper presented on July 26, last year, before the Biological Society. The text of this paper follows.

"One knows that it is possible to produce by inoculation of Bacterium tumefaciens in plants tumors similar to those of cancer in animals. One of us obtained experimentally by this method, a great number of tumors. These had various degrees of development. Some of them dry up partially, but do not die entirely until the entire plant or at least the limb bearing the tumor dies. Even removed by surgical methods, these grew again on a sick limb.

"We propose to describe in this paper, the action of electromagnetic waves of very high frequency obtained by means of the Radio-Cellular-Oscillator of Georges Lakhovsky. This apparatus produces wave-lengths of the order of two meters and less, corresponding to 150 million cycles per second. A first plant was submitted to the effect of radiation one month after being inoculated with cancer; at this time small tumors of the size of a cherry stone were visible upon it. This plant was submitted to the rays twice, for three hours each time. During the following days, the tumors continued to grow rapidly in the same way as those on plants, which had not been submitted to the effect of radiations. However, 16 days after the first treatment, the tumors began to shrink and dry up. A few days later the tumors were entirely dried up and could be very easily detached from the limb of the plant by merely touching them. The drying action of the radio frequency radiations is selective and affects only the sick part of the plant. Even the inside sick tissues were destroyed, although they were next to healthy cells in the center of the limb, showing that the radiations had not affected the healthy parts.

"Another plant was treated in the same way, except that it was exposed 11 times for three hours each time, to the radiations of the oscillator. Sixteen days after the first exposure the tumors, which were rather large as shown in one of the photographs, began to shrink and dry up and were easily detached from the limb exactly as in the first case. Again in this case, the healthy parts of the plant were not affected in the least. A third plant exposed to the radiations for nine hours, that is, three treatments of three hours each, was cured in the same manner as the two others. Sixteen plants also inoculated with cancer, were left without treatment. They have tumors in full activity, several of which are very large. These experiments show conclusively that plants inoculated with cancer can be treated and cured by means of the ultra radio frequency vibrations, whereas surgical treatment fails.

"Such are the results of my researches with plants. At the present time, similar experiments are being carried out with animals and it seems that the effect on cancerous animals is the same as on cancerous plants.

"I am highly pleased to present my theory and the results of my work in a scientific review of the United States, this great country, which has always been in the lead in the fight against this horrible sickness, cancer, and whose sympathy and help I would greatly appreciate.

"In conclusion I wish to call the attention of the reader to the fact that I have obtained very conclusive results not only with a wave length of two meters, but with longer and shorter wave-lengths. The main thing is to produce the greatest number of harmonics possible."

* * *

This February 1925 article from Radio News magazine shows how long some of Lakhovsky's theories have been known in America. This article was furnished to BSRA by an Associate who had met the Russian-French scientist in New York. He claims also to have visited several times the New York hospital which set up a research program on the Multi-Wave Oscillator and saw the Lakhovsky-type equipment. No information on the machine or its effects could be elicited from the hospital authorities and in the early 1950s the shielded room was dismantled and equipment destroyed.

We feel that Lakhovsky's theories and early researches should be followed up. It should be publicly proven that high-frequency waves, in the proper concentrations, do affect living organisms in a beneficial way, or that Lakhovsky be declared an idle dreamer. Since research of this kind falls within the purposes of Borderland Sciences Research Foundation, Inc., members are encouraged to build their own equipment and, in a suitably shielded or isolated environment, carry on their own research program. If you cannot obtain the necessary materials, BSRF may be able to put you in touch with another borderlander who can be of help. Until the medical authorities come forth with official approval of this radical approach to the state of imbalance or un-health attributed to disease, no one but a fool would go into commercial production of this kind of equipment in America today. It is now said to take a minimum of $300,000 in research money and three years of testing to develop medically acceptable data.

HOW DR. JAMISON WAS TRAPPED

In a letter dated Aug. 29, 1967, Dr. Jamison enclosed this news clip from the **Modesto Bee**: "Turlock – Veterinarian Stanton L. Jamison yesterday was fined $249 in Turlock Judicial District Court after being convicted of three charges stemming from his use of an electronic machine to treat human patients for cancer. A jury of nine women and three men returned a verdict of guilty on one count of practicing medicine without a license and two counts of advertising a device represented to have effect on cancer, stomach ulcers or bone diseases. The decision followed a two day trial in which Jamison defended himself. . .

"They (Jamison and his assistant, Martha Ann Davis) were arrested last December following a lengthy investigation by Turlock police officers and agents of the State Bureau of Food and Drug Inspection. The investigation began when Police Chief John Viarengo requested assistance from the bureau after learning 'the cancer-curing machine' was being operated here.

"During the investigation, Food and Drug agents Willis Worley and Harvey Lee arranged several meetings with Jamison on the pretense they were interested in purchasing machines similar to his. Jamison reportedly organized a church as a front for his operation. The facility, located at 200 West Main St., was known as the Life Aquarium Center. Worley and Lee became bishops of the church as did deputy District Attorney Frank Damrell, Jr., who prosecuted the case. At the time of the investigation, Damrell was employed in the State Attorney General's office.

"Worley said Jamison's intention was to open other treatment centers throughout the state, using his bishops to set up facilities under protection of the church. 'This was the most fantastic investigation I've ever been involved in,' said the agent this morning.

(Yes, it is fantastic to think anyone would want to give or donate healing services free when there are billions to be made on the four kinds of cancer treatment authorized by the California Medical Association: Drugs, Surgery, Radiation and Death. RHC.)

"Worley points out the manufacturer of Jamison's machine was convicted of grand theft two weeks ago in Los Angeles after selling a similar machine to a Food and Drug agent under false pretenses. (This was probably Ralph Bergstresser, Burbank, Calif., who was manufacturing and selling the Bob Beck MWO for $250.)

"After Jamison's conviction yesterday, Damrell called for imposition of a prison term for the veterinarian. If he had received the maximum penalty, he could have been sentenced to 18 months in jail and fined $1,500.

"We felt very strongly on this," said Worley, "because Jamison could easily become a menace to many innocent victims throughout the state."

MWO IN USE AT INDIAN CLINIC

"I am continuing my experiments with the MWO and my experience is that in practically all cases there is an aggravation of the existing symptoms, and previous symptoms are brought to the surface. Hemorrhage was seen in another case and treatment was stopped forthwith. This is one problem which has to be tackled before MWO can be used with confidence. I think MWO treatment should not be given to tubercular cases. Also in cases where there is a hemorrhagic tendency this would do more harm than good.

"Neuritis seems to respond nicely to this treatment. Sciatica is definitely cured by this. Some cases of Rheumatism and Arthritis are benefited, responding at first but the improvement does not last much longer. I cannot understand the reason for this eccentric behavior. I think there must be something with the frequencies."

Don't blame it on the frequencies, Doc, blame it on the patients who stubbornly clings to their fixed ideas and rigid unyielding mental patterns. We have no evidence that the radio frequencies of the MWO reach into the mental world. Each person's aura is loaded with "signatures", national, family and religious concepts hammered into him over the years. His creative energies, flowing freely down from the higher worlds, channel through these deeply cut "grooves" in mind and emotions. The health or illth of the physical body indicate whether the person's ideas are in tune or out of tune with Creation. As a man thinketh in his heart, so is he.

"Can you please let me know what are the ranges of frequencies these machines have? Have they been measured and are they the same in all instruments? Is there any difference in the models prepared in the USA and France? Which model was tried by Lakhovsky in the New York hospital? What are the actual frequencies used by him?"

The only frequency readings we have on an MWO were those submitted by Bob Beck originally, from the output of his own machine. "It radiates RF interference over a bandpass of 15 Kilocycles to well beyond 250 Megacycles. This blocks AM, FM and Television reception for hundreds of yards" But so does my electric razor, and I've been using one for years. I was told by a sailor that electric razors are very effective as "scramblers" for knocking out the electronic brains of guided missiles, and have been so used as defensive equipment during missile exercises on aircraft carriers at sea. No one has furnished us with RF output figures from original Lakhovsky equipment, either at the New York hospital or from Europe. If anyone has these figures I hope they will be kind enough to forward them to you. Lakhovsky's available technical record is very slim.

"An Associate writes 'A very good psychic in Florida gave the following, in a sitting devoted to the MWO. It is not anywhere near developed to its potential. It has unlimited amplifications to be made. The use of the MWO can change the auric field, thus spiritually it can change the individual. This is done with the work of the seven rays of life. The MWO should be used in conjunction with the radionics instrument.' What about this radionics instrument? I would appreciate it if you can throw some light on this."

Sorry Doc, the punitive hand of the American Medical Association has driven radionics builders and users underground here in the United States. No one but a fool like me would stand up on the firing line. England isn't quite the police state America is and you can get information on radionics and radionics equipment from the Delawarre Laboratories, in England, and from Bruce Copen, in Sussex, England. Radionics experts have developed rates for every organ in the body. Equilibrating those rates with specific radio frequencies offers an interesting field for borderland research.

"I have tried both the Vitic and the Eeman Screens. I am still continuing my experiments with the Eeman Screens. Only a couple of days back I gave a set of screens to a patient who is having an encyclopedia of complaints. The MWO did not give her much improvement hence the change. I have one set of Vitic with me and whenever I feel a bit out of sorts I sit for ten minutes and get back my vigor. I am using a very strong pair of magnets and a carbon rod 1½ inch in diameter. After I received the MWO my interest in the other two flagged. I think I will have to cultivate them again."

Dr. A.K. Bhattacharya

From the March 1964 Journal of Borderland Research

ARMY MEDICAL RESEARCH SUPPORTS LAKHOVSKY THEORY!

That's our headline, and it's derived from news clips loaned to us by an Associate. One especially is from the New York "Times" for Friday, Aug. 19, 1960 by John Osmundsen.

The heading says that Radio Microwaves Produce Molecular Changes in Human Beings. This is merely a repetition of what Georges Lakhovsky proved with his pioneering work in Paris hospitals in the 1920s!

"New and possibly important biological effects of radio and radar waves were reported at a three-day scientific meeting that concluded here yesterday. One scientist told of producing 'a profound molecular change' in human gamma globulin with specific wave lengths of radio waves in the high frequency to very high frequency, or VHF range.

"Alterations so created in the molecules that constitute an important part of the body's defense apparatus were accompanied by at least a fourfold increase in their biological activity, the scientist said. 'The changes can be produced under widely varying conditions of voltage, power, pulse width, and pulse repetition rate, provided the frequency is suitable,' according to Lieut. Col. Sven A. Bach of the Army Medical Research Laboratory in Fort Knox, Ky."

If Col. Bach had been working from Lakhovsky's original theories and equipment, he wasn't saying. This is obvious in reporter Osmundsen's next paragraph.

"This is believed to be the first report of a proven frequency-specific effect of radio waves on human chemistry. Dr. Bach's report appeared to cause more comment at the conference than almost any other, for the effect of radio frequencies on human gamma globulin in the test tube were clearly not a result of heating.

It was suggested by Dr. Bach that the increased biological activity of the molecules as measured against a rabbit serum sensitized to them might have resulted from an unfolding of their structure, thus exposing more active spots."

YUP, THEY MIGHT EVEN CURE CANCER

Later in 1960 an electronics magazine revealed even more about the proceedings of this Fourth Annual Tri-Service Microwave Conference held in New York. Our news clip is only half a page which reveals neither the magazine nor the writer, but about Dr. Bach's claim of a four-fold increase in the production of gamma globulin, he says:

"What does this mean to you? Gamma globulin plays an important part in the body's defenses against disease. Speed up its activity fourfold and it may be four times as effective a disease fighter. Some day your doctor may give you 'a shot of microwave' to keep you well.

Dr. Joe Howland of the University of Rochester found that dogs which had been exposed to microwaves were able to withstand heavy doses of X-rays far better than dogs not so treated. No specific application for this technique is now known, but such discoveries frequently lead to undreamed-of advances.

"And finally, some evidence has been gathered to indicate that certain kinds of cells, under the right conditions, will simply disintegrate -- fly to pieces -- in the presence of the right radiation. But surrounding cells are unaffected. Could such a selective beam ever be used to destroy cancer cells while leaving healthy tissue unaffected? Scientists don't yet know."

The writer didn't say which scientists he is speaking for, certainly not for Georges Lakhovsky, who died in that very same New York City 17 years before the Microwave Conference. Before-and-after photos of skin cancer cures effected by the Lakhovsky Multi-Wave Oscillator in Paris hospitals in the 1930s show that the Russian inventor did know just how selective radio microwaves can be. These pictures are reproduced in Lakhovsky's book, "The Secret of Life".

To continue with the magazine article: "said Col. George M. Knauf, one of the foremost experts on the biological effects of microwaves, speaking of these recent advances, 'Our work has progressed sufficiently for us to be fairly certain that the beneficial uses to which this energy may be put will greatly outweigh any harmful results it may produce in men.' All this of course does not mean we should forget that uncontrolled microwave energy can be dangerous. But like fire or W-rays, microwaves can, when used intelligently and with understanding and respect for the hazards involved, be among man's most valuable tools."

MULTI-WAVE OSCILLATOR USE IN A NEW YORK HOSPITAL

Georges Lakhovsky crossed the Atlantic to New York City in 1941, there to observe the effects of his MWO under controlled, technical conditions in a New York Hospital. A brief report of the positive and beneficial results of MWO treatments is appended to the English translation of his book, "The Secret of Life", by Mark Clement, on page 201. These treatments were given from July 1st to Aug. 21st and the name of the hospital and of each patient are withheld for ethical reasons.

Patient	Illness	Treatments	Results
X.	Arthritis, both knees	14	Good improvement
O.P.	Endocervicitis. Abdominal pain	6	Condition good
F.T.	Periarthritis of shoulder	12	Marked improvement after first treatment, condition good
M.M.	Osteoarthritis	7	Marked improvement after two treatments, stiffness diminished
M.K.	General Chronic Arthritis	11	Slight improvement
M.O.	Arthritis of ankle	3	Results good
A.E.	Post-operative fracture	6	Results good
C.V.	Arthritis, both knees	14	Marked improvement
M.D.	General arthritis	8	Marked improvement
J.H.	Epicondylitis of elbow	7	Results good
D.G.	Chronic arthritis	10	No cough, no soreness, marked improvement
S.L.	General arthritis	10	Marked improvement
B.M.	Arthritis and circulatory disturbances	6	Improving
M.L.	Arthritis of ankle and knee	7	Improvement, no recurrence
M.B.	Arthritis of shoulder	7	Marked improvement
M.O.	Tenosynovitis, hand	6	Improvement
S.S.	Arthritis, both knees	7	Improvement

S.C.N.	Congenital hip dislocation	3	Marked improvement
R.L.	Fracture, both ulnae. Synovitis, both knees	5	No pain

Additional Cases

Case 1. G.D., female, aged 25: Patient complained of pain in the chest and muscles of the right arm and back, and constant cough. She did not respond to medical treatment and massage. Treatment with Lakhovsky Multiple Wave Oscillator was instituted. After the fourth treatment coughing stopped, pain in the chest became infrequent, and the tired feeling in the back disappeared. At the end of 12 treatments the patient no longer complained of previous symptoms.

Case 2. M.M., female, aged 51: Patient complained of pain and stiffness in the right hip. X-ray examination disclosed a marked narrowing of this articulation associated with large spurs at the margin of the acetablum which appeared to ankylose the joint. There were proliferative changes in the upper margin of the femoral head. Diathermy and massage failed to relieve the pain and stiffness. After 10 treatments with Lakhovsky's M.W.O. the stiffness markedly reduced.

Case 3. F.T., female, aged 38: For two months the patient complained of pain in both shoulders. Abduction of the right humerus was limited to 45 degrees. A diagnosis of periarthritis was made and treatment with Lakhovsky's M.W.O. was started. After three treatments the patient said she felt much better. The improvement continued. After a series of 11 treatments the patient did not return. Apparently permanent relief had been obtained.

Case 4. A.B., male, aged 60: Patient gave history of pain in both knees beginning in 1940. Diathermy and massage for one year gave only slight relief. Treatment with Lakhovsky's M.W.O. was instituted. After one treatment patient stated that he felt better. After four treatments patient felt so much better that he took only one treatment in two weeks.

Case 5. M.M., female, aged 42: Patient complained of pain in lower left back and hip. X-ray examination of the lumbar spine and pelvis showed calcification at the anterior margins of the inter-vertebral discs between the ninth and tenth and the tenth and eleventh vertebrae. There was also a slight scoliosis of the lumbar spine. Treatment with Lakhovsky's M.W.O. was instituted. After the first treatment patient felt better. After the fifth treatment improvement was very marked. Patient did not return after nine treatments.

Case 6. S.N., female, aged 59: Patient complained of pain in the right shoulder and inability to extend the arm more than 5-10 degrees, accompanied by severe pain. X-ray examination disclosed irregular calcification in the region of the greater tuberosity of the right humerus. Treatment with Lakhovsky's M.W.O. was instituted. After one treatment patient said she felt much better. The improvement continued steadily until 10 treatments were completed.

Case 7. C.P., female, aged 38: Gynecological examination showed uterus slightly enlarged and hard. Right para-metrium tender. Diagnosis --Parametritis. Treatment with Lakhovsky's M.W.O. was instituted on July 17, 1941. After four treatments patient felt better, and after 12 treatments patient was found to be much improved. Examination on Oct. 7th showed the adnexa and parametrium to be painless. Patient was considered cured.

REPORT OF CASES TREATED WITH LAKHOVSKY'S MULTIPLE WAVE OSCILLATOR BY A PROMINENT BROOKLYN UROLOGIST

The following results were obtained by a prominent Brooklyn urologist whose name must be omitted for ethical reasons. He is a Fellow of the American College of Surgeons and enjoys a great reputation as a skilled specialist. He treated hundreds of patients with Lakhovsky's MWO but only a few typical cases can be given here.

H.G.	Acute urinary retention due to enlarged prostate	2	Retention ceased
R.B.	Bleeding from uterine fibroids of two years standing	3	Bleeding ceased
S.M.	Mild multiple sclerosis with halting gate	6	Marked improvement, was able to run on several occasions
R.R.	Cellulitis of nose with pain	2	Cleared completely after two treatments of ten minutes each in one day.
B.T.	Acute urinary retention due to enlarged prostate	2	Retention ceased.
S.B.	Bleeding from cancer of the bladder	3	Bleeding checked. The tumor cleared up nicely and was reduced in size when seen again on cystoscopy three weeks later.
J.G.	Chronic aczema of Axilla	2	Cleared entirely.
L.G.	Severe pains form duodenal ulcer	2	Greatly improved.

In addition to the above cases this Brooklyn specialist treated six cases of enlarged prostate with nocturnal frequency (three to five times per night). After two treatments with Lakhovsky's M.W.O. frequency was reduced "to only once a night or did not occur at all..."

The above American reports were not included in the first edition of Georges Lakhovsky's "The Secret of Life", published in 1935, but were added by the English translator, Mark Clement, to later editions. We are quoting from the 1963 edition published by Health Science Press, Sussex, England. Health Research, (P.O.Box 70 Mokelumne Hill, California 95245) offers this same book.

This early (1941) research with the Lakhovsky equipment seems to nave been ignored by the American medical authorities, who are still concentrating on cure by surgery or medicine. If you want to discover the effects of weak radio waves on living organisms you'll have to dig for yourself. Our BSRA brochure, "The Lakhovsky Multi-Wave Oscillator", will give you some hints about setting up your own research program, and is available from headquarters.

'DOCTORS SAY RADIO TREATMENT IS EFFECTIVE AGAINST ARTHRITIS'

Apparently some doctors didn't ignore the research work done by Georges Lakhovsky. Take this news item from the New York "Post" for Feb. 8, 1962, with an Associated Press dateline from Washington, D.C. "A device theoretically capable of bouncing radio signals off the moon has yielded highly effective results in the treatment of stubborn cases of arthritis and certain other ills, a group of doctors reported today. The machine generates high-frequency, short-wave-length radio waves and emits them as pulses lasting only 1/17,000th of a second each. The pulsed currents are directed into a patient's body, with a maximum achievable penetration of eight inches.

"Use of the technique in treating arthritis, bone infection after hip operations, inflammation of the female pelvis due to gonorrhea and other conditions, and in speeding the healing of surgical wounds was described in reports by Dr. Euclid M. Smith of Hot Springs, Ark.; Dr. Dana Street of the University of Arkansas; Dr. Bruce Camern of Baylor University; Dr. Marshall Lobell of Harlem Hospital, N.Y.; and Dr. Solon N. Blackberg of Chicago. The reports were made last night at a private conference to which several Congressmen and Public Health Service doctors were invited. Blackberg later told reporters the researchers would like to enlist government support for expanded research by themselves and their investigators.

"Smith, in his report on arthritis, said the technique had been employed with "highly "effective results in 99 out of a series of 100 cases treated by him. Declaring that some of the patients treated had arthritis for up to 25 years before receiving radioactive treatment. Smith told reporters that "marked improvement" was achieved in all cases in from three to six months.

PUBLIC RECOGNITION OF RADIOWAVE TREATMENT

This sensational advance in the treatment of human ailments should have been welcomed from coast to coast in every newspaper. Dr. Blackberg and his distinguished fellow scientists from Baylor, Arkansas and Harlem should have received public acclaim from President Kennedy's Dept. of Health and Welfare and from the American Medical Association. But only a thunderous silence followed that early 1962 news item ion the private meeting in Washington -- until the Aug. 24, 1963 edition of the Saturday evening Post! In this national magazine, radiowave treatment of arthritis was branded medical quackery in the article "Hucksters of Pain". There Drs. Smith, Street, Camern, Lobell and Blackberg had their answer to "government support for expanded research". Fortunately, not every American is going to take this kind of libel without striking back at the mouthpieces of the vested interests who lend themselves so readily to control of the mass mind. Read this news clip from the Los Angeles "Times" for Nov. 1, 1963.

"SATURDAY EVENING POST SUED FOR $7.5 MILLION"

"New York -- The Saturday Evening Post, still smarting from one of the largest libel awards in history, was hit Thursday with another libel suit demanding $7.5 million.

"The suit was filed in Federal District Court here by the Diapulse Corp. of America, manufacturer of a machine that employs high-frequency waves in the treatment of infections and inflammations, including arthritis. The company said it was defamed in an article, "The Hucksters of Pain", which appeared in the Aug. 24 edition of the magazine. The article dealt with quack remedies for arthritis. The suit seeks $2.5 million in compensatory damages and $5 million in exemplary damages."

From the June 1964 Journal of Borderland Research

LAKHOVSKY OSCILLATING CIRCUITS

"About eight years ago I sent to the Health Science Press, Sussex, England for a little book, "Quick and Free Healing", and also "The Waves That Heal" and found that there are healing waves in the air we breathe. I am a sensitive with the pendulum. The book told to take an insulated copper wire and find the positive and negative ends with the pendulum, so I would know which end to put to the left side of my body --positive to the right and the ends overlapping. If you can't keep them in place, wrap an elastic band around them. For me, the pendulum swings clockwise for negative and back and forth for positive. First use of the wire came when I skinned my shin with a five-inch gash. The bleeding made me sick and I had to lie down. I thought of my wire circuit, got up and put it on just above the knee. It always had taken a long time for sores to heal on my body but this healed in three days time and showed only a narrow line about the width of small yarn. This convinced me that the oscillating circuit wasn't foolish.

"I loaned my little books to a high school student that was taking science. I thought they would interest him but he made all manner of fun of me, for believing such a thing. None of my family are interested in the good things that I have learned. I have been a member of BSRA since 1951 and think your journal is the best magazine I have ever read. Anyone wanting to learn more about the Lakhovsky circuits can write to Col. A.E. Powell. (Now deceased)

"Forty years ago I was taking treatments from an osteopathic doctor who sometimes diagnosed with the pendulum. When I became pregnant he told me he was going to discover the sex of my unborn child. He took his gold band ring off, tied a string to it and held it over me. It didn't seem to have any definite swing, maybe around once and back and forth. 'That's strange,' he said, 'I've never missed yet!' When the baby came there were two, a boy and a girl! I didn't ask him then where he learned about the pendulum or what it meant. Years afterward I sent to New York for one and found that I could ask questions and get yes or no answers that satisfied me. A medium told me that the pendulum is spirit controlled and mine was controlled by a young girl who was rather mischievous. If you want to put my experiences in the BSRF Journal it is all right with me. I got my education in the little red schoolhouse. I am 73 years old and still learning. They say you never get too old to learn and I find it is very true." Mrs. Earl Holfinger, Piqua, Ohio

A "HOPPED UP MWO" IS DANGEROUS!

"It is far too early to draw final conclusions regarding my experiences with the higher powered Multi-Wave Oscillator, a word of warning is in order. I powered mine with a 250 VA, 10 KV oil burner transformer. This is DYNAMITE and could, in my opinion, be fatal if too much power is used and I didn't miss it by much. My equipment sparks the antenna down to the 4th ring and the resonator to the second, with occasionally a week spark to the third. 10 hours after the first high-power exposure I felt like running instead of walking (age 72) and walked all around town and back, about six miles and no tiredness, whereas two miles would tire me before. That first exposure was about ten minutes for me and four for the wife (age 75). Immediate reaction for me was a moderately strong sensation of tingling from head to foot for about three hours before I could get back to sleep. None of this for the wife. Then, as aforementioned, I felt on top of the world. 24 hours after that, however, the wife and I developed the grandaddy of all colds and flu symptoms that we'd ever had. I hadn't had any real cold or flu in 15 years. This one dredged such an amount of mucus and toxic waste as I'd never seen before in my life. It was rough. It lasted two days before it wore off.

"This all had me guessing until your brochure arrived, with the experience of your Engineer-Physician Associate. He told that many of his patients developed flu symptoms after MWO exposure. His increasing conservatism was well founded. I decided to find out if the doctor's hypothesis of viral penetration would hold up. If so, a second high-powered treatment of equal strength and duration should produce much the same flu symptoms. But, having wound another coil, I used more power than the first time, sparking antenna and resonator down to the fourth and fifth rings. Results: no more flu, only a very slight tingling sensation, no feeling of exhilaration as before. Five days later I took another exposure, of seven minutes, with a little less power. This one did it. No tingle, no exhilaration, only a feeling of rapidly increasing depression, no sleep that night, a rather severe pain in the chest and heart region, extremely weak pulse, hardly pep enough to life an arm, and a nervous shaking all over. For a while I thought, this is it. It took five days for this to wear off and it isn't all gone yet. One night I lay awake until three a.m. figuring out how to pep up the MWO still more. Out of the blue, a voice clear and distinct and very emphatic: 'This is dangerous!' Again very emphatically, "Yes!" Normally I am not the least psychic, clairvoyant or clairaudient; so I choose to heed this as a warning from a discarnate guide or guardian."

Let this Associate's experience be a warning to you, if you feel the urge to build a hopped-up Oscillator, to transform yourself overnight! Bob Beck's original article cautions against this and his original equipment falls well within the safety tolerances set by the Air Force and the telephone companies for personnel working around radar and micro-wave broadcasting and relay equipment. Most of you taking part in this research program have built your equipment only in the last two or three months. It would be wise to stick closely to the original design for a year before altering it for further research.

LAKHOVSKY'S ORIGINAL THEORIES ARE SOUND!

With one or two exceptions all reports in to date show beyond question that Georges Lakhovsky's theories are sound: Radio waves do effect living organisms. In certain concentrations these waves appear to benefit cell growth, even to stimulate it. This BSR Associates have proven to their own satisfaction in only six months from the publication of our first article. Now it is up to others to exploit this knowledge while we turn our attention to other projects.

*** *** *** ***

From the July-August 1964 Journal of Borderland Research

LAKHOVKSY OSCILLATING CIRCUITS

"The danger of strong reaction to a hopped up MWO, as told in the June Journal, reminds me of my experience with the single circuits. Months ago I wrote to Col. A.E. Powell and he sent me single circuits to wear around waist, neck etc. This I did for two months or so, daily. I'd been under the weather enough to hope for benefit. Instead, I had three successive colds -- something I never, never had in my life before, even during the worst of my catarrh years. It never once dawned on me it could have been the Lakhovsky oscillating circuits; now I wonder?"

A California Associate

Yes, it had been the experience of others that the apparently powerless Oscillating circuit, of only one wire looped around the body, can produce a strong reaction on living tissue. A San Fransisco Associate told us during our May trip north that she received a wire loop from Co. Powell and wore it to bed that night, and forgot about it. Early in the morning she awoke with such a violent pain through her middle she thought she'd die! She lay there and stuck it out until the pain left and went back to sleep. In the morning she awoke refreshed and ready to go, with the oscillating circuit still around her middle. It began to dawn on her that this was the cause of the pain, but it also had affected the healing of a long-suffered condition. At the time we were there, weeks later, the condition had not returned. Certainly no permanent cure of any condition can be expected unless the original cause in mind, emotions or spirit is changed, and if a person insists on indulging his vices, use of oscillating circuits may only cause needless pain!

*** *** *** ***

From the OCTOBER 1964 Journal of Borderland Research

HE PREFERS VITIC TO MWO

"Personally, while feeling the passage of the high tension flow from the antenna of the Multi-Wave Oscillator, I found no beneficial effect. On the other hand, the Associate with whom I obtained the MWO realized definite improvement. He has kept up the use and told me lately that he is feeling steadily better and lays the improvement to the use of the outfit. Some years ago I had frequent conferences (thru a medium) with one who was said to be Madame Blavatsky. During the course of these talks she dismissed the subject of electricity as of little value in comparison with a study of magnetism. That in the study of magnetism lay the secrets of many valuable matters. I am led to refer to this because I recently obtained relief, and apparent cure, from a troublesome irritability of the bladder by the use of a magnet. Other treatments had been of no value. I made no report of this until the fact was definite. The use of the magnet was experimental and came from the suggestions of a Chicago University specialist in the subject, relayed to me via various Biomagnetic reports and some correspondence."

Associate A.N. Onymous, Dixie, U.S.A.

This letter from an elderly Associate was dated Sept. 3, 1964. It certainly provides an interesting and stimulating commentary on the material in the preceding pages of the Journal. Interestingly enough, we've had many reports of successful and continued relief of irritating bladder and prostate conditions from exposure to the MWO. In the Vitic device we now have a borderland gadget that is far less troublesome to build and to operate. How effective it is remains to be proven.

THOSE LAKHOVSKY CIRCUITS AT WORK

"About the Lakhovsky wire circuits, there is no other way to account for the consistently maintained energy I receive. I wear one all the time as do several others I know, all of whom report amazing results. One man past 80 finds it peps him up too much, so he wears it intermittently."

Howard D. Clark, Yucca Valley, Calif.

*** *** *** ***

From Jan-Feb 1965 Journal of Borderland Research

HE WORKED WITH GEORGES LAKHOVSKY

"A friend of mine sent me your pamphlet describing the Lakhovsky Multiple Wave Oscillator. I was the first man that bought this instrument in New York in 1940, on the advice of a physician who saw it at a medical congress in Vienna. My young son had developed a face lesion. It was diagnosed as tubercular. I was having a joint office at that time with a progressive MD. I was practicing dentistry. Her success with the MWO was phenomenal, but when she began to treat cancer cases the American Medical Association got after her and me and revoked our licenses. I had translated some of Lakhovsky's works into English from the French. Professor Lakhovsky came often to my New York office. He told me the Pope was using one of his instruments. The success obtained in the hospitals was excellent but the reports were suppressed. The good professor died of a broken heart. Our case was a regular frameup by the AMA, the lawyers and the courts. I was foolish enough to fight the case and spend lots of money, not knowing that the AMA controls even the courts and judges. I still have some copies of the booklet and a photo of my Lakhovsky instrument. If I can locate them I'll send them to you. I sold my instrument to a doctor in Texas and left New York state in 1945. I am a licensed Naturopath in Florida and Chiropractor in California. Am enclosing $ for membership. I belonged to BSRA several years ago."

Dr. N.S. Hanoka, Miami Beach, Florida

*** *** *** ***

From March 1965 Journal Borderland Research

"THE SECRET OF LIFE"

"I loaned my copy of Lakhovsky's 'The Secret of Life' to my chiropractor, also your Journal. He had one build and I've been taking treatments from him on the Multi-Wave Oscillator. After two treatments I got an awful cold and I have never blown so much bloody mucous out of my nose. Also, after each treatment, it would leave me weak. I would have to lie down several times a day. There always was a reaction but my bad sinus condition has cleared up considerable. I also had trouble with my liver and there were so many things that I hadn't been able to eat, for 15 years or more. Now I can eat them and enjoy them. The MWO is certainly worth having and I am improving."

Mrs. A.N. Onymous, Somewhere, Indiana.

*** *** *** ***

From the October 1966 Journal of Borderland Research

HITTING THE HEALING TRAIL WITH THE MWO

"It would be appreciated if you would put an announcement in the Journal that Dr. S.L. Jamison, D.V.M., N.D. (Gr.Br.) announces the opening of a practice of naturopathy in Santa Cruz, California.

"The climate has changed greatly from the old burning-of-books days by the FDA. We cannot all stand back and wait for "George" to do something. I have the advantage of a lot of training in a legitimate field of allopathic medicine. Only a veterinarian would be able to make such a move and escape retribution for it. A regular M.D. would be branded a traitor to his fellows.

"I have moved to Santa Cruz to set up practice as a naturopath, using the M.W.O. unit, spiritual healing and diet -- along with naturopathy measures for the healing of old injuries and chronic degenerative diseases. There will be no charge if there are no results. Since my fees will have no legal status, this will be made clear to the people using the service so they will be under no compulsion to pay, even if results are satisfactory!

"There are no incurable diseases but there are incurable patients!

"With the MWO we have not gotten uniform results classified by the disease the patient had. About two out of ten people do not respond to the MWO unit at all. You had just as well breathe on them. I strongly suspect that these people should consult an allopathic physician, M.D., as they would be helped better by conventional medicine. You might say our cases were selected, as people came to us in the last stages of sheer desperation -- conventional methods (including chiropractic) having failed.

"Some of these (not many) also do not respond to the MWO. They usually respond to spiritual healing methods when this happens. This I have not figured out good as yet. Once in awhile an arthritic will claim to have been worsened by the unit. This occurs usually in an arthritic with arthritis in a number of joints. One treatment seems to break the poison loose in their body, and they fall like the wrath of hell about two days after. Get them in immediately and run them again, and everything is alright.

"To sum it all up, I have got so I do not care what disease a person has got as long as it is not an acute infectious disease, or a recent stroke case. I have been afraid to try it on either of these as all we would need is for somebody to fall over dead while on the unit and we would be wearing striped suits or breathing HCN in a gasper chamber. . . The unit regenerates the whole body." S.L. Jamison, D.V.M.

*** *** *** ***

From the July-Aug 1967 Journal of Borderland Research

A TRANSISTORIZED MWO

"I would like to compare operational notes with some one who has built one of these MWOs because I question some of the statements made in the original article. I'll just give a brief description of mine and some of the things I ended up doing. I started with aluminum elements but had trouble with the gluing process so ended up with using printed circuit boards so my elements (antennas) are copper. I have good arcing between first four rings. Can make 'em all arc if I bring point of wooden pencil near the smallest, but pencil must be held certain way or I'll get knocked on my haunches.

"I ended up with a 9 in. Tesla coil. Tried 7, 6 and 5 but ended up with nine. The trouble here was with leakage between primary and secondary. It was terrific until I took means to reduce to a minimum. The transistorized drive I use is partially a hangover from the days of when I used transistorized ignition in my one vehicle. The ignition coil was modified to increase the step-up ratio and since the coil was mounted originally in oil, it was a 'slippery as an eel' project. The driver uses 5 transistors plus diode and operates on a source power supply which can supply up to 20 volts d.c. The drive seems to function best at about 15 - 16 volts.

"Now for a question. If Mr. Beck, as he indicated in his original article, 'reconstructed the circuits' of an original MWO in a certain man's house, why didn't you people publish a diagram and dimensional data on that original piece of equipment? That would be a project a man could sink his teeth in. Instead this pee-wee MWO which you now say is not much good unless everything arcs. You mean arcing all the way down to the smallest ring?????? When you get your up-dated MWO data published, let me know as I would like a copy. R.A. Rieck, Rochester, Minn.

Arcing between the first four rings, from the outside in on the MWO antenna, is about as good as any we saw when we were researching the instrument in 1964; so it sounds as though your transistorized MWO -- the first we've heard of --is doing very well and should be putting out enough energy to affect living cells. You any wish you had cut aluminum antenna rings after all. For some reason, the arcing between copper rings builds up deposits which must be cleaned off periodically. Bob Beck didn't offer us a schematic of the original Lakhovsky equipment he saw. He said it was housed in a 300 lb. cabinet. The antennas were of circular-formed pipe rings. The outside ring was two feet in diameter. There was a large, separate Tesla Coil driving each antenna. This is a professional model designed to impress the patient. It also probably puts out enough static to antagonize the neighbors for blocks around. We believe Bob Beck deserves all credit for grasping the principle of multiple wave propagation developed by Lakhovsky during radio's infancy in the 1920s and redesigning it into a simpler, smaller device for present-day borderland research.

LATER COMMENT ON THE MWO BY MR. BECK

"The key factor is that the rings MUST arc profusely all around the circumference of the outside and THIRD ring simultaneously. There is no effect, either on a calibrated Hewlett-Packard field strength meter, or physiologically, at the target frequencies if this arcing does not shock-excite the rings. And yet I have seen at least three machines that did not have sufficient voltage from the Tesla Coil to even drive the outside ring, let alone the third element! It would be a shame to invalidate the device because of this. Some of the machines I've seen drive the Model T Coil from an A.C. transformer! The obvious fallacy here is that the coil's vibrator is a resonant device supplying the make-and-break voltages (and magnetic field) to the T coil primary at the designed frequency. Since the transformer is on and off 120 times a second, the chances of the T coil points 'breaking' at the precise instant that the 60-cycle house current is at a POSITIVE peak is about one in fifty. And even if it fires 'sometimes', the T coil is modulated with the 120 cycle line (twice the 60 cycle frequency, since each cycle has two peaks, both a positive and a negative).

"Try it at 300 mc. and 10 kmc. on a field-strength meter some time. You'll see what I mean. The Tesla Coil, incidentally, is just a convenient supply of free electrons. The cascade of electrons or 'high voltage' breaks down the air's resistance (by arcing) and then drives the antenna rings. But if the drive is interrupted, the rings may not absorb and resonate properly. At any rate, it is difficult to kid the Hewlett-Packard meter, and I'd like to see the original design followed, at least for the first year."

AN IMPORTANT MODIFICATION

"The greatest modification to date was the discovery of a box of war surplus very high voltage rectifiers in a bin at a Burbank electronics store. The 30,000 to 60,000 volt selenium stack rectifiers.....eliminates the necessity for a tuneable gap, and roughly quadruples the output of the Tesla Coil, also lowers battery drain to half."

BEWARE THE PUNITIVE HAND OF THE AMA

Turlock, Calif. "Daily Journal", Dec. 2, 1966: "Dr. Stanton L. Jamison, formerly in charge of the state Poultry Pathological Lab here, was arrested yesterday for practicing medicine without a license. The veterinarian is charged with using an electronic device -- with high voltage sparks crackling and jumping about on either side of a patient --with the claim it was beneficial in cancer and bone diseases.

"After posting $1100 bail, Dr. Jamison defended his machine. He said parts for the machine cost less than $50. If someone followed up on the apparatus, he claimed, it would be a 'blessing to humanity'. Police Chief John Viarengo, Lt. Bill Ladd and investigators for the California Bureau of Food and Drug Inspection took Dr. Jamison into custody in the offices of his Life Aquarian Center in the Mercantile Building. Lt. Ladd said a sign reading 'Universal Life Church, Inc.' was prominently displayed in the front office. He said four or five agents had been 'ordained' as ministers in the church by telling Dr. Jamison they wanted to become preachers. Soon their official ordination certificates arrived in the mail, said Lt. Ladd.

"Police described the apparatus as similar to an electric chair, with four-inch metal rings looking like handcuffs hanging over doors to the room. The theory was to determine polarity of a patient with a pendulum and then proceed to change negative polarity to positive with the 70,000 volt device. Chief Viarengo said patients were being treated for a variety of diseases, including stomach ulcers and cancer. Investigation into Dr. Jamison's operations started in May, when police began receiving complaints from relatives of patients. No money was charged for treatment, according to Lt. Ladd, but patients were expected to make donations to the church. Dr. Jamison headed the state poultry lab here for a decade. He will be arraigned on the charges in Turlock Judicial Court next Friday morning.

"In a statement released following his arrest, Dr. Jamison said the 'multiple wave oscillator' picked up by police resembles many electronic gadgets used, usually ineffectively, in the past for treatment of many complaints. However, he insisted, his oscillator appears to be the exception that proves the rule.

"'The unit suffers from a sort of 'guilt by association' with other electronic gadgets to the point where serious doubt would be the reaction of any trained scientist on first observing it,' he admitted. 'I know that this was my reaction.' However, he said he suffered continuous pain following a 1962 automobile accident and, willing to try anything, tried the unit. It helped him, he insisted, and has helped many others. He said the unit had been tested and checked by qualified medical men in Europe in the late 1930s. 'Some progressive healer will follow up my lead in continuing to develop the electronic unit,' he said."

*** *** *** ***

From the November 1968 Journal of Borderland Research

AN MWO PHENOMENON

"Please find enclosed $ for another year's subscription to your wonderful Journal. It has been very enlightening, and I always look forward to receiving it. I have taken treatments on the MWO for a long time and have been puzzled. Soon after I started to take treatments, the hair stopped growing under my arms. I have wondered about this and if anyone else has had the same reactions?"

Mrs. E.H. Piqua, Ohio

This is new to us! If any of you researchers have experienced or observed a similar phenomenon, we'd be glad to hear about it and share it with the rest through the pages of the Journal. The Tesla Coil problem has been solved with the discovery of a good commercial coil made by Master Appliances, Marion Indiana. This coil will drive one antenna very well.

*** *** *** ***

From the Jan-Feb 1971 Journal of Borderland Research

THE PASSING OF CYRIL SCOTT

"Cyril Scott, born Sept. 27. 1879 in England, died first week of 1971, so two friends here inform me. I'm always happy to see one who writes so intelligently regarding diet and health, vindicates his ideas by living that long. Not many do! BUT, I read in our local paper yesterday of the death of a former neighbor who died on 11th January at 93, and she knew nothing regarding diet, positive thinking, etc. So, where are we? I can't ever make 2 & 2 come to four, woe is me.

"Two small books come from Muriel Archdale, England, by Cyril Scott, 'Safe and Saner Remedies', and 'Health, Diet and Common Sense'. On page 162 of the latter he mentions the George Lakhovsky book, 'The Secret of Life', memory of which caused Mr. Scott to suggest using a copper wire around his middle to a man who was too breathless to walk far or uphill. The man wore the Lakhovsky loop for twelve hours, had a copious and offensive bowel movement, then on following day and another uphill walk, the man was able to keep up with Scott! I found my copper flex soon after but it is awkward to keep in place to make sure the negative end is to the left. Cyril Scott gives the homeopathic remedies needed, particularly emphasizes potassium for cancer."

A.F., Santa Cruz, California

COPPER BRACELET CURES "TENNIS ELBOW"

News Item from Crosby Golf Tournament at Pebble Beach, California a year ago. 1/24/70: "Bert Yancey one had to give up golf because of a nervous breakdown. He also came close to quitting the professional tour last year because of a 'tennis elbow'. But he didn't mainly because he found a remedy endorsed by Australian tennis pros John Newcombe and Tony Roche. The so-called remedy is a $7.50 copper bracelet, otherwise known as 'The Absorber' or a 'Voodoo Bracelet".

"Yancey wears one on his right wrist to correct an arthritic condition in his elbow and he has worn it for the past five tournaments, including the $150,000 Bing Crosby National Pro-Am, which he leads by two strokes at 137 going into today's third round.

"'I don't know why it works, but it does,' said Yancey, a 31-year old former West Point cadet who finished third in the 1967 and 1968 Masters and won the Atlanta Classic last year. He said a tennis pro in Napa Valley suggested last summer that he wear the bracelet and got confirmation from Newcombe and Roche, who have suffered from similar ailments, that it would work.

"Yancey strained his right elbow two summers ago lifting a concrete birdbath in the backyard of his home in Tallahassee, Fla. He took a pain-killing drug that enabled him to compete in the U.S. Open at Rochester (where he finished third), but he was warned that the drug w 'dangerous...that it gives you ulcers'. 'I never got ulcers, but I felt a pull in my stomach. That's when I quit using the drug.'

"He said he hasn't needed the pills anyway, since wearing the bracelet. The therapeutic effect of the copper, he said, has to do with chemical displacement within the body. 'I majored in chemistry while taking pre-med courses in college and I think I know enough about copper deficiency in the body to believe that, somehow, the bracelet can help me.'

"Somebody then pressed Yancey for an answer about what makes the bracelet such an effective remedy. 'I don't know for sure,' he drawled, 'maybe it's all psychological.'"

YANCEY WINS THE CROSBY OPEN

Sports writers are always looking for a new angle for stories and Bert Yancey withstood a spectacular charge by Jack Nicklaus on the final round, with a 65, and came in with a hot round of his own a 69, to win the tournament and a cash prize of $25,000. Yancey and his Voodoo Bracelet were headlined all over the country. The LA "Times" called it a "Copper Caper" in its Jan. 26, 1970 story.

As was to be expected, the medical authorities in desperation had to prepare an immediate counter blast of propaganda to keep the public hypnotized. The sales of medical snakeoil went down as arthritis sufferers flocked to buy non-medical copper bracelets! The "Times" and other metropolitan newspapers willingly cooperated. After all, patent medicine advertising is a big source of revenue, and who knows how much profitable pharmaceutical stock is owned by the publishers of the papers?

The "Times medical writer, Harry Nelson, dutifully wrote: "The Arthritis Foundation, a legitimate group interested in getting better treatment for arthritis sufferers (There is a contradiction here, Harry, for medical authorities claim there is no cure for arthritis!) refers to copper bracelets as 'an age-old fraud which has no scientific basis whatsoever'. Yancey is the latest sports figure to get taken by the 'copper caper', which the attorney general of New York last summer called 'one of the oldest and most vicious swindles'."

These opinions by the Arthritic Foundation and the attorney general are of course; utterly worthless. They are not based on the results of intelligent scientific research, but are based on the profit motive of the medical trust. The electrical nature of living cells was established beyond any shadow oaf doubt two generations ago, by the sound theories and research of Dr. Abrams in San Fransisco and Georges Lakhovsky in Paris. Thousands of supporting tests since then have proven conclusively that body cells have polarity, and that in an unbalanced or unpolarized condition, disease or injury, the subtle oscillations of a wire coil or loop will restore polarity. Sad to say, there is nothing of this in the training of medical students today and it is left to Borderland science to carry the torch. The honest, straightforward testimony of public figures like Bert Yancey, Tony Roche and John Newcombe is ridiculed by the witch doctors of authority for medicine.

JUST KEEP TAKING YOUR EXPENSIVE PILLS!

Under orders from above Harry Nelson interviewed Dr. John Calabro, the arthritis authority at University of California in Los Angeles, "He points out that the bracelets may seem to work because arthritis can improve spontaneously and it is easy for the person to associate his improvement with the bracelet.

"'I tell patients it's OK to wear the bracelet providing they do all the other things I ask them to do,' said Dr. Calabro." Including, of course, the purchase of poisonous pain killers at fancy prices. How much scientific integrity does Dr. Calabro have when he and his university accepts fat research contracts from the medical trust? Is it any wonder that today's Aquarian Age students are publicly questioning the morals and ethics of their professors? They are told that the goal of a university is the search for truth -- at least that is what the Chancellors say in their public speeches. But when the student gets to the laboratory he finds that the search is for profitable products for industry -- or help maintain a well-established profit position.

LAKHOVSKY BELTS, LOOPS or OPEN CIRCUITS

"Thank you for the clarification of the Lakhovsky belts described by the late Col. Powell in his publication. I tried what you suggested and after some thought I finally understood what he meant and I believe the belts should look like this:

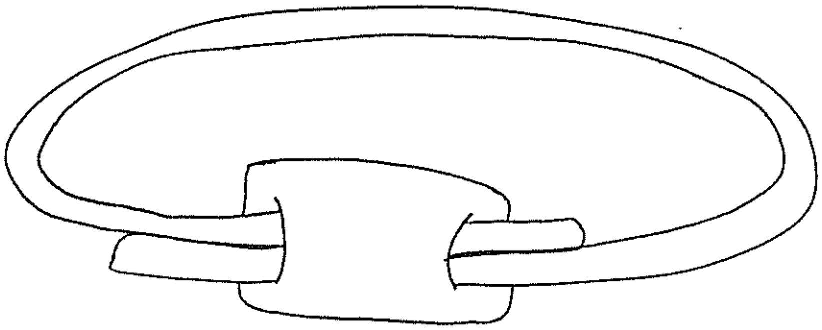

You might want to try experimenting with flat ribbon cable for these belts or coils. It can be bought at most electronic stores and it comes in different widths or colors. I am wondering also, is it really necessary to separate the lamp cord and use only half of it? Could it not be used as it is, with two strands of copper wire, insulated, side by side?" G.B.F., San Fernando, California

Yes, in our one-loop coils of white insulated lamp cord, we now use it just as it comes from the store, without separating the two strands. But you plastic 'buckle' or 'button' shown above may allow too much slippage; so the ones we make for ourselves and for research-minded Associates have separate entrance and exit holes for each end of the insulated wire. The white insulated wire seems preferable from the hygiene angle, seeing that it is usually worn next to the skin.

*** *** *** ***

From the March-April 1971 Journal of Borderland Research

OUR LOOP-THE LOOP PROJECT

"It looks to me like the Lakhovsky loops work homeopathically. It's not the copper element, per se, but the loop the loops it makes, throwing back into the body forces which the body is throwing off. It's an automatic homeopathic prescription, each body being its own pharmacist!

"The Yogi does it when he goes into suspended animation, re-breathing his own air. Urine therapy, drinking one's own urine, as described by Armstrong, is homeopathic. An ostrich putting his head in the sand breathes his own air. Breathing into a paper bag and re-breathing the contents is a specific cure for old age pains and hiccoughs. Dianetics was mental homeopathy, i.e., the original book but not no more. Dianetics is re-thinking one's own thoughts. You imbibe your own aberrations, attaching like with like. Homeopathy is like a feedback system. Gods ate their own offspring. It's Isis picking up the pieces. The smaller the dose the more potent the cure. Etc. Etc.

"Send me two kits of three experimental loops. Maybe I can fatten up on them."
P.F., Florence, Arizona

HE FEELS LIKE A NEW PERSON

"I read that part in the Journal about the copper wire loop; so today I did loop one around me and sure enough the oscillating motion calmed me down. It sure was a sensational feeling because lately at times it got so bad I though I would spin off, even though I do the five Rites nightly. On occasion I also experience a terrific pain in my left side which forces me to sit down, just a great pressure especially after standing over the stove preparing meals for maybe three to four hours at a time. The customers just couldn't understand how I stood up under pressure of serving so many people. Seems like mass production.

"So with this new item I feel like a new person! And I'm going to have you send me one of your experimental kits, to see how it compares to what I hooked up on myself. Of course I just taped the wire on and this is only about six hours since I put it on." J.M.S., Charleston, South Carolina

A DEDICATED BORDERLANDER

"Enclosed is a check for membership renewal and for five sets of experimental loops. I want to compare the effects of the Lakhovsky coil with the Japanese-made magnetic health bands which I sell to my tennis customers. I am very active in tennis playing and teaching. The results with this band are very good."

A.M., Miami, Florida

BEYOND THE PHYSICAL

An Associate writes that she had success in "curing" and arthritic elbow by wearing a copper bracelet on her wrist. Then the pesky condition showed up in her knee! From there is moved to another part of her body. Then it began to dawn on her that their Higher Self was trying to tell her something, that the swollen and painful joints were an outward manifestation of an inner mento-emotional condition of resentment and bitterness over an unhappy marriage. With the cooperation of an understanding doctor and more orthodox heat treatments, etc., and a more positive and cheerful attitude on her part, the condition was finally eliminated. It took months to cleanse the system of the accumulated poisons, physical, emotional and mental.

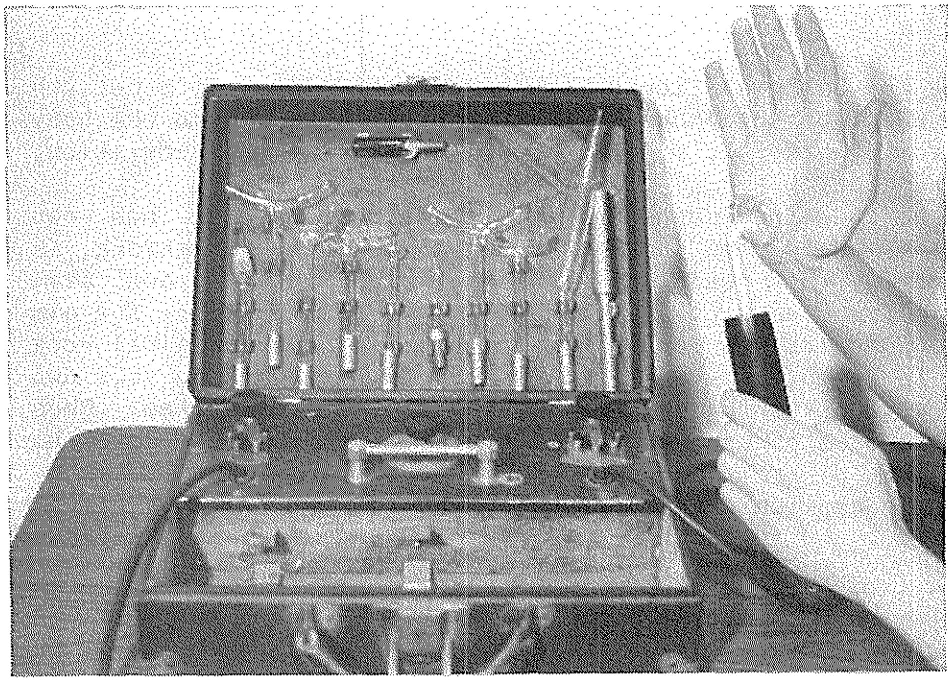

1920's era Violet Ray Massager. From the BSRF Collection.

From the Jan-Feb 1966 Journal of Borderland Research

THE LAKHOVSKY MWO IN INDIA

"The MWO is giving wonderful service and days are passing and I am gaining some new experiences which have not been recorded in the book. I will appreciate if you kindly clear some of my doubts. In practically all the cases I am finding aggravation after three or four sittings. Do you think that it is due to a sudden oscillatory shock that is given to the patient as it so happens when Homeopathic medicine is given? I feel that it should have at least been recorded by some one. One more question. What is the reason for metals being removed from the person of the patient? In case they are not removed will there be any adverse effect? This is very vital because here in Bengal there is a custom that the married ladies must wear one Iron bangle. They will not remove it under any circumstances. What is the remedy for this?

"I find that the appetite improves in practically all cases. Rheumatism and Arthritis cases are definitely benefitted. In one case a Tumour which is there since about 20 years has definitely gone down with only 4 sittings. I am watching the case carefully and if the Tumour disappears it will be a great achievement for the MWO. Debility is markedly reduced in nearly all cases. In a recent case of Spondylitis the MWO has done wonderful work. The flexibility was 8-1/2" from the floor which has come to 6" after 7 treatments and all the pains have practically disappeared. Due to the New Moon and Eclipse there is an aggravation but I hope it will gradually go. Asthma cases have not responded so well but I am not yet disappointed with the results. It is as yet too early for me to assess the capabilities of the MWO on various diseases. As you say I also believe that this instrument will work on all persons irrespective of their complaints but still there must be some specific organs on which there is more influence of this.

"Something which I forgot to mention previously, can you tell me whether you have seen Hemorrhage as a result of MWO treatment? In at least 3 cases I have seen bleeding start where there was no previous history of bleeding. What is it due to? In one case there was profuse bleeding after 2 sittings and that has made me a bit nervous. Just now a patient reported that he was getting blood in the sputum even though there is nothing in the lungs. In the third case blood appeared in the urine. Please give a thought to this as this is a very serious complication if it is caused by the MWO." Dr. A.K. Bhattacharya, West Bengal, India.

The above is a portion of a letter to Associate Al Goeke, Redmond, Washington, which Al passed on to me for comment. I thought the rest of you would be pleased to know that this East Indian doctor has gotten very encouraging results from the very beginning in the use of the Multi-Wave Oscillator. If by aggravation the doctor means a sudden release of mucous, heavy cold or flu, this is a sure indication that the MWO is working and affecting a release of accumulated poisons in the body. This is touched on in our brochure, in the comment from a midwestern doctor which built and used such equipment in the 1940's. When he began producing mild flue -- as he called it -- in his patients, he got scared off! The hemorrhaging by the doctor's East Indian patients is more serious but we believe it is still a result of the same thing, a sudden releasing of accumulated poisons through the shedding of bad blood. We know one oriental who had a heavy nose bleed after the first exposure to the MWO field. Apparently orientals, some of them at least, are more sensitive to the soft radio wavers than Caucasians living in America; so the doctor's prescribed treatments should probably be somewhat shorter in duration and spaced further apart.

Yet I well remember the radio technicians in Honolulu, Hawaii. I was an announcer and producer there for several years. Some of them were of Chinese and Japanese ancestry. When repairs were needed on the radio towers, these men wouldn't hesitate to jump on, climb up and work for hours while the station was putting out a steady five to ten thousand watts of power. I don't recall any of them suffering any ill effects such as internal bleeding. Yet their tools would get so hot because of the strong field that they would have to wear gloves in handling them! It was these eddy currents caused by an inert piece of metal in the pulsing field that Bob Beck was concerned about. There isn't enough power in the MWO to cause heating in bits of metal about the body, if one can't strip down for an exposure. The piece of metal, say a zipper or a ring, might reduce the efficiency of the field in that area but we didn't worry about this when testing the equipment we had for a short while in 1964. I don't think Dr. Bhattacharya should either. The machine seems to produce results anyhow.

*** *** *** ***

From the March 1966 Journal of Borderland Research

BASIC NEW AGE PHYSICS

by Trevor Constable

(Publication of Dr. Bhattacharya's letter in the Journal, page 20, Jan-Feb 1966 issue, elicited the following significant comments by Mr. Constable.)

A word on Dr A.K. Bhattacharya's reports concerning the Lakhovsky Multi-Wave Oscillator he is using successfully at his West Bengal, India clinic. The results are indistinguishable from those systematically recorded and published over a long period by the late Dr. Wilhelm Reich, M.D., concerning his "orgone therapy". Reich was the discoverer of Primary Energy -- pre-material and mass-free -- which he named Orgone.

The improvement of the appetite, the reversal of rheumatism and arthritis even to structural changes, the rapid reduction of tumors and the hemorrhage effects are all to be found in Reich's massive clinical literature on orgone therapy. Dr Bhattacharya's reports, plus others any my own experiences with the MWO serve to confirm my previous evaluation of this unit as published in the Journal over two years ago. In essence, the MWO is a device for orgone therapy. It achieves its startling results because it is a generator of primary energy, or at least a focus for primary energy.

Any person with a modest extension of physical vision who will observe an operating MWO under conditions of indirect, rather dim fluorescent lighting, will see for himself that primary energy from the earth's atmosphere (the Orgone of Reich) condenses prolifically between the coils of the unit. The concentration is many times that which normally surrounds a living organism.

The basic New Age physics involved, for which we are indebted to Reich, are as follows: Wherever secondary energy, i.e. energy obtained from matter in some way, is released, primary energy tends to concentrate to extinguish this energy. Secondary energy is life-negative, i.e. inimical to the living, a fact which is illustrated, in its ultimate extension, in the lethality of atomic energy. Where so-called "soft" radio waves are involved, the antagonism of primary energy to their emission is evidently considerably diminished over the violent effects that accompany radioactive substances in the presence of primary energy accumulators.

In the case of the MWO, the tremendous band of frequencies covered results in an extreme concentration of primary energy, which also has its frequencies and correspondences (the Rates of radionics) with the secondary, electromagnetic spectrum. That this effect is achieved with minimal secondary emission is the beauty of the Lakhovsky unit.

TWO-FOLD FUNCTION OF THE MWO

In my view, the dramatic results obtained are due to a two-fold function of the unit. First, there is the undeniable concentration of primary energy around the body of the patient, particularly potent where a limb or other readily accessible member may be exposed. Secondly, there is the reaction of the specific biological energy of the individual organism, from within, to the stimulus of the secondary waves, which it rushes to combat.

Under the laws of primary energy first discovered by Reich, the patient, as a living organism, is a stronger vital system than the MWO. Accordingly the patient attracts and withdraws the charge of primary energy produced by the device. This is in direct contradiction to formal electrical laws, and as such is typical of the criteria that rule primary energy. The continued emissions of the Lakhovsky unit provide a continuous concentration of primary energy on which the patient draws. Just how long this process should be allowed to continue, per treatment, is where we run afoul of allopathic (MD) thinking.

The acquisition of a strong primary energy charge by the ailing organism results in a reversal, or at least in the reduction of life-negative, i.e. diseased, processes. Reich found with his orgone accumulators that the reduction of tumors was not as great a problem as the auto-infection resulting from the overloading of the excretory apparatus. The debris and bacilli from the diminishing tumors must be eliminated, and unfortunately, because the excretory apparatus is frequently involved in the general putrefaction of which the tumor is the end result, this was the major problem in managing such cases, in Reich's experience.

The medical lectures of Rudolf Steiner, Ph.D., contain one of the master keys to successful therapy, particularly appropriate to the use of the MWO. Steiner points out that what the ailing organism needs is the significant, vital impulse in the right direction, then careful nursing and management as it recovers out of its own forces. This was also the view of another genius, the late Dr Ruth Drown, who knew from experience how important the first impulse back to health is for the sick individual.

Hence, it seems that the old allopathic urge to convulse the patient with therapy, to repeat the massive dose again and again, is something that should be resisted in using the MWO, at least where tumors are being dealt with. The idea of applying the MWO in daily doses may stand in need of modification. We are all to some degree creatures of gross, non-vital thinking that stands behind our allopathic medical civilization, and we will need to carefully train and think ourselves out of the absurdity that the massive dose is the sine qua non of therapeutic success.

THE TRUE-BLUE ROOT CAUSE OF DIS-EASE

As far as unexpected hemorrhaging from the lungs and similar manifestations are concerned, I think the problem is essentially that diagnosis is so inept and inaccurate that the physician does not have a true blueprint of his patient. After my years of association with Dr Drown I am sharply aware of the fact that nobody just has one ailment or problem, but that we are all bearers of a veritable plexus of malfunctions which usually proceed from one primary and obscure root cause. What happens with the MWO is that the injection of primary energy into an ailing organism results in multiple manifestations of the healing impulse, and not just a direct effect upon the object of our attention --which may be a tumor, ulcer or other extreme symptom.

All the effects produced by exposure to MWO units are identical to those experienced in orgone accumulator. The healing agency in both cases is primary energy, and I hold that this is so no matter what plausible arguments might be advanced to reconcile the MWO with formal theories and thus win the neurotically-desired approval of physicists and doctors. It perhaps seems incredible that one can obtain the same effects produced by the MWO by sitting in an orgone accumulator, with absolutely no electronic apparatus connected to is, but it happens to be so. Try it and see!

MWO experimenters owe it to themselves to study Reich's work in full. As long as they wander in these fields, they may as well build into themselves the realization that electromagnetic energy has an inevitable and unavoidable and inescapable concomitant -- primary energy, life-positive, mass-free energy whose criteria and laws are the exact opposite of the energy accepted by formal physics. The best testimony to Reich's tremendous contribution to human advancement is that he died in a U.S. Federal Prison. As all Borderlanders know, innocent bumblers die in bed in their own homes, untroubled by the status quo.

*** *** *** ***

From the April 1966 Journal of Borderland Research

MWO SURGERY AT SHASTRI VILLAGE AND RELATED MATTERS

by Al Goeke

I believe you and the Associates would like to hear the latest comments from Dr. A.K. Bhattacharya, Naihati, West Bengal, India on the Beck adaptation of the Lakhovsky Multi-Wave Oscillator he is using in his clinic there.

"The other day I had an unique experience with the MWO. I think this should be published. I was treating a lady for rheumatism. After four sittings, once a week, she had a pain under the tongue and there was some inflammation with pus coming out. I asked her to take a further sitting. After about five minutes all of a sudden the thing burst inside the mouth and five pieces of stone came out. When I took her medical history I found that she had an operation two years before and a bit of Salivary Calculi was extracted, but the surgeon had told her that some still remained behind. The MWO treatment eliminated the foreign matter which would not have come out without an operation. This then paves the way for research with Gall Stones and Kidney Stones. I would like to know if anyone else had an experience like mine.

"This time I can give you some reports about the special attachment you sent. A patient was having much pain in the knee and there was swelling. I held the Bulb against the knee for five minutes, and after that it was found that the swelling had reduced and the pain also was much less. In an asthma case I applied the Bulb on the chest for ten minutes. The report is the patient had no recurring attack for two days. Yesterday I was having pain in my thigh so I applied the Bulb. In five minutes I got wonderful relief. Gradually I am getting the hang of it and I think I will be using it more and more."

I, too, am using the Argon more and more, and the more I experiment with it, the greater I think it is, also the MWO. I have had two test cases of young men who, after five minutes in the MWO, said they could feel something happening inside their heads. Both cases had a history of head injuries. I take it that the MWO reactivates a dormant, abnormal condition that perhaps one day would cause trouble. Another test case was a mental patient in Texas. The use of the MWO was part of the test program which brought about a marked improvement in only nine days.

Some time ago a BSRA associate spent an evening with me and told of an interesting borderland gadget. People came from miles around to sit in its healing radiations. I built one for experimentation and found the results good. A Texas Associate says that it works! The original was set in the ground, but I wanted it in the house where it would be more convenient, and figured on running a ground wire outside; but this is not necessary. I have found that it corrects polarity of anyone sitting in the same room. The one I built radiates out 25 ft. One length of 6" stovepipe is used, filled with alternate 2" layers of crushed granite, and of charcoal. For a quick charge of energy, touch your foot on wire at lower end and hand on wire or rod sticking out of the upper end. I have checked the walls and ceilings or rooms (with pendulum) and found them negative. After a few minutes with this borderland gadget in a room, the walls and ceilings check positive! Here is something that should be shared with Associates.

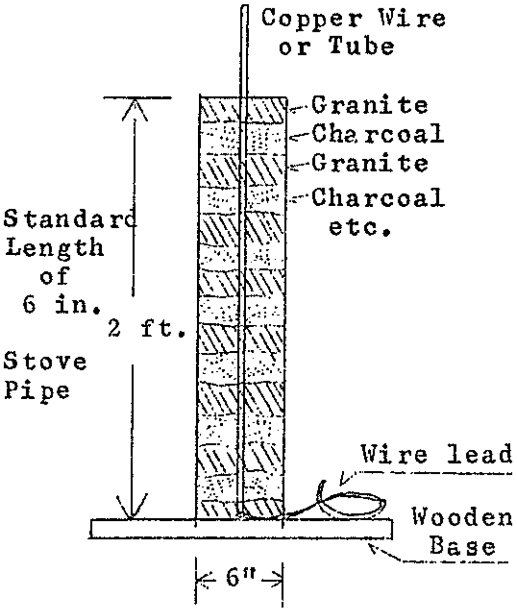

ETHERIC CONDENSER

(Here is our drawing from the rough sketch in Goeke's letter, of the borderland gadget explained above. This is just another of many simple devices which concentrate or condense etheric energies of the higher physical sublevels into a field or vortex of force available or useful in beneficial amounts to a living organism within that field. The drawing is not to scale. The dimensions are not critical. It is the relation of the parts that is important. You can consider the upper end of the copper wire as positive and the lower end coming out of the base as negative, but with the alternating layers of crushed granite and charcoal there must be an alternating or pulsating effect. Remember, charcoal is carbon, so this Etheric Condenser is a Vitic device, but not nearly so concentrated as the field between the Alnico magnets of Vitic. Iron filings sifted in the crushed granite would probably increase the power of this thing. This gadget would augment the power of your Eeman Screens, if you are using them, by adding extra etheric vitality to the circuit. Add long wire leads to the upper and lower ends of the Condenser, so you can hold one in each hand while holding the handles of the Eeman Screen leads. Which Condenser lead should go to each hand? I dunno! This is borderland research and you'll have to experiment on yourself, and be you own authority! This is one of the first principles of the Aquarian Age into which we are moving at breakneck speed. Now to continue Al Goeke's article. RHC)

GERMICIDAL EFFECT OF ULTRA-VIOLET LIGHT

The use of the Argon bulb (General Electric AR-1) as a pain-killer and infection reducer is not a new idea. Nikola Tesla used a gas-filled tube or bulb, excited by high-frequency current to rejuvenate himself every day in his laboratory. 40 years ago the RenuLife Electric Co. of Detroit was making and selling such a treating device to doctors all over the country. The offered a choice selection of vacuum tubes for getting into the body to release rejuvenating, purifying ultra-violet rays. I don't know that their tubes were filled with Argon gas, as is the current General Electric bulb. Maybe this is why the GE bulb is so effective. There is a greater concentration of ultra-violet light in the most beneficial wave-lengths, from pure, excited Argon gas.

We visited one chiropractor on a trip last year who used an MWO machine on his patients, and also uses the Tesla Coil of the MWO to drive the Argon bulb. General Electric designed this little gas-filled bulb for use on the regular 115 volt house current. As a very dimly burning, ultra-violet night light it is a good spook-chaser and can be left on all the time because it uses only pennies worth of electricity in a month! The little bulb has a curved metal plate inside it, as shown here, on which you can see a faint, ultra-violet luminescence when it is turned on, in a standard light socket.

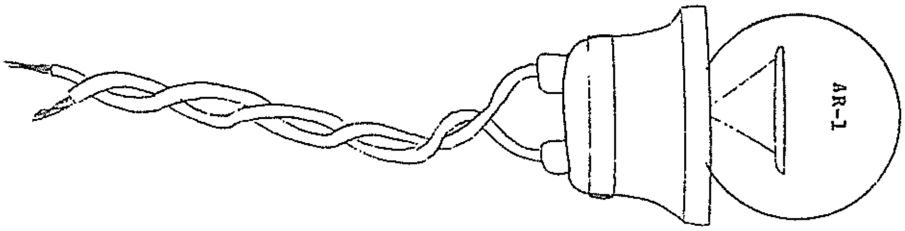

BUT, when you put a hundred thousand volts through it from the MWO Tesla Coil, that Argon gas gets real excited and throws off powerful ultra-violet energy which can penetrate into the body and destroy infection and break up tension, inflammation. Showers of sparks jump from the surface of the clear glass to the skin unless the bulb is held so it touches. Sparks will also jump into your hand unless the socket mounting is well insulated. The mounting or socket shown here is the standard rubber base with heavy, six-inch leads for outdoor lighting. Our local hardware store sells the socket for 39 cents. The Argon bulb needs only one connection -- to the antenna lead of your MWO. The black and white leads from the socket can be twisted together and soldered to a female Banana Plug. This will take the male connection from the Antenna lead. The twisted socket leads can be wrapped with heavy, insulating tape if any of the ultra-high voltage gets through to tickle your hand.

The chiropractor who makes use of this in his work asks his patients if they have any infected teeth, or a toothache. One five-minute treatment over an aching tooth, rubbing the Argon bulb on the cheek over the infected area, stopped a toothache for us. It also relieved a throat, sore from too much lecturing.

General Electric also makes an identical Neon bulb. This gas throws off an orange radiation, which would probably give off more of a surface, heating effect because of the longer wave-length. You won't find these special bulbs in your local stores, probably, but will have to order them through some wholesale dealer, say an electric contractor and builder who has the GE catalog of special bulbs. A compact, professionally built Tesla Coil can be bought from Edmond Scientific Co, Barrington, New Jersey. It doesn't put out as much energy as the MWO coil designed by Bob Beck, but it will cause the GE Argon bulb to emit considerably more ultra-violet radiation than the 115 volt house current.

From the MAY-JUNE 1971 Journal of Borderland Research

A HIGH-POWERED MULTI-WAVE OSCILLATOR

A HIGH-POWER MULTI-WAVE OSCILLATOR

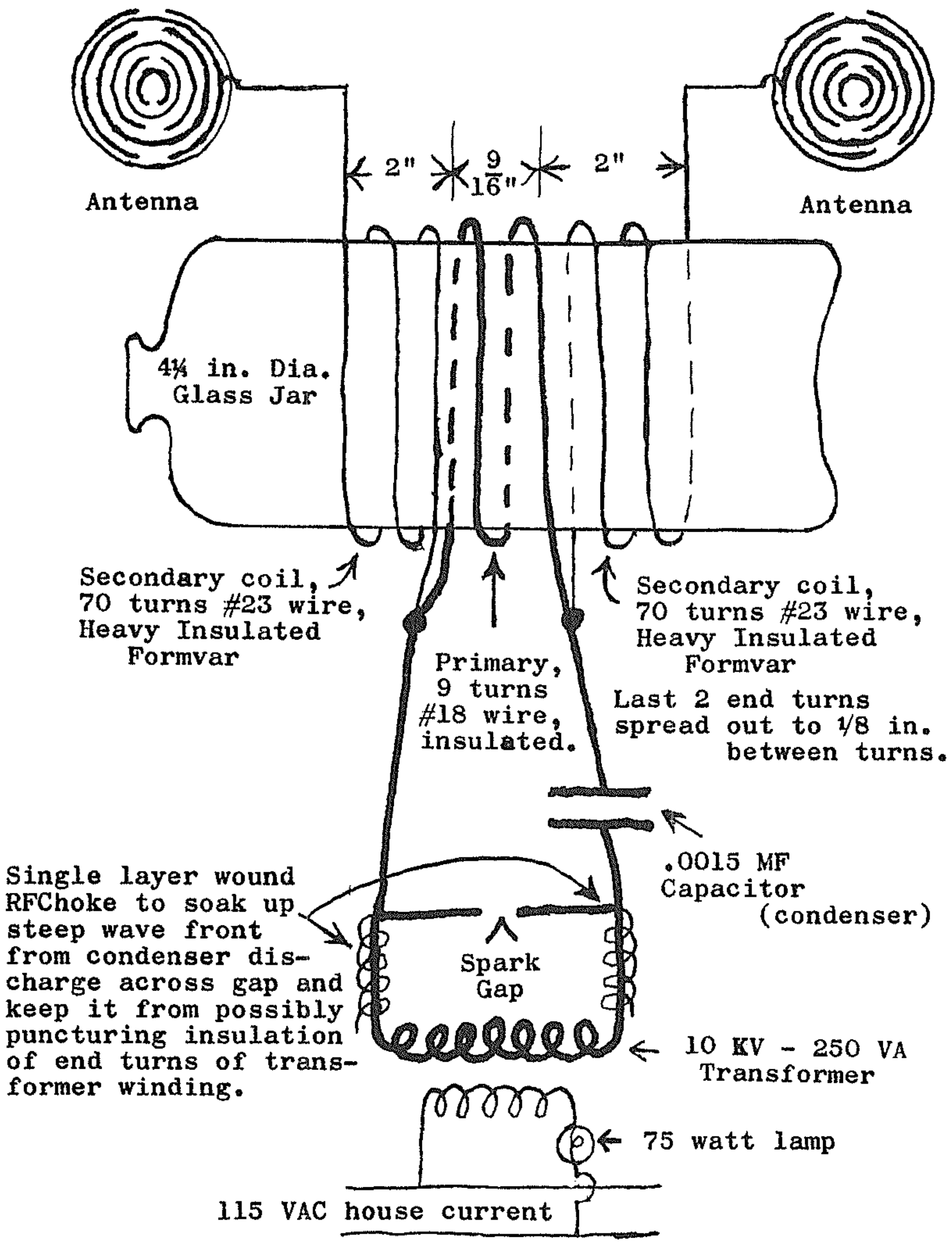
Antenna
2"
$\frac{9}{16}$"
2"
Antenna
4¼ in. Dia. Glass Jar
Secondary coil, 70 turns #23 wire, Heavy Insulated Formvar
Primary, 9 turns #18 wire, insulated.
Secondary coil, 70 turns #23 wire, Heavy Insulated Formvar
Last 2 end turns spread out to 1/8 in. between turns.
.0015 MF Capacitor (condenser)
Single layer wound RFChoke to soak up steep wave front from condenser discharge across gap and keep it from possibly puncturing insulation of end turns of transformer winding.
Spark Gap
10 KV - 250 VA Transformer
75 watt lamp
115 VAC house current

"Your brochure article stating that Georges Lakhovsky's original MWO had a separate Tesla coil driving each antenna is most interesting. This is in line with my experience here in Manila. I find that when livening up both antennas to full voltage, to spark each antenna down to the 3rd ring with an occasional slop-over to the 4th ring, it is much more effective than when keeping one antenna (or resonator) down to near ground potential; and it is least effective when actually grounding this antenna -- as Mr. Beck does --through a 1/10 MF capacitor!

"One five minute treatment a week is enough for me and the equivalent of two 15 minute treatments with only one antenna up to full voltage. However, I drive both antennas up to full voltage with only one Tesla coil. It has rather tightly coupled primary and secondary windings, wound as an auto-transformer, on a 4-1/4 in. diameter glass jar, as per enclosed sketch.

"The present spark gap is 1/8 inch, a fixed, constant gap, needs no turning with variable distances between antennas, subject or no subject. The sparking is much the same and the gap is not at all critical, 1/32nd, plus or minus, works just as well. Gap electrodes are 1/4 inch tungsten. The present primary circuit capacitor is .0015 Micro Farads, but this also is not at all critical, can be varied up to .0025 and still give much the same sparking and effect on subject.

"I tried using #28 Formvar wire, heavy insulated, 70 turns each on the secondaries, to shorten coil and get tighter coupling, primary to secondary, but coil was too short and flashed over from end to end.

"Just plug it in and use it, no tuning manipulation of any kind required. The input of 115 VAC to the 10,000 volt oil burner ignition transformer is only 17 VA through the 75 watt lamp as a limiting resistor. Never use more than a 100 watt lamp. This is all I can take. More power becomes rapidly depressing. And the full 250 VA out put of the transformer I once used, with a 5/8 inch spark gap, was nearly fatal. It took me four months to get over the effects of it. Pass this info on to your Associates."

J. Gilbert E. Wright

Associates can benefit from this dedicated borderlander's experience, which confirms Lakhovsky's warning that healthy cells can be over-stimulated and killed by powerful high-frequency waves -- as well by X-ray and radium waves. It may be easier to locate Neon Sign transformers of 5,000 to 7,500 volts output as power supplies for this MWO, at a cost of $15 to $20. In this case the wattage of the "limiting resistor" electric light could be increased to 150 or 200 watts, or eliminated altogether, if it is necessary to get enough power to cause random arcing down to the third and fourth rings on both antennae. The good doctor says nothing about shielding but care should be taken to prevent radio interference from this powerful spark.

"No data is given in the drawing (May-June 71 Journal) on the size of the antennas. I would suggest that the outer ring be about 18 inches in diameter with a 3 inch gap. Alternate down through the rings separating them about 2 inches apart and held together with plastic electrical tape. Wire should be #10 bare. The secondary leads connects to the second ring from the largest diameter outside ring."

George Van Tassel, Giant Rock, California

An even more spectacular possibility is the Time-Machine effect in connection with rejuvenation. If, as Bob Beck suggested in his article, the MWO tends to take the body cell back down the time-track to a more youthful, more vital period, what would happen to a person who stayed in the field of the machine for four hours? Would his vibratory rate be speeded up to where he would disappear completely? Like the hero of H.G. Wells' 'Time Machine'? Anybody want to try it? It seems, from hints George Van Tassel has been dropping over the years, that his Integratron at Giant Rock, California might do something like this -- if and when he gets it working! As I recall, both George and the Canadian Saucerian, the late W.B. Smith, spoke of the four dimensions of electricity: 1. is the electricity moving in one direction along a copper wire, 2. is the magnetic field around the wire, 3. is the static charge created by the first two effects, and 4. is Time -- or the 'tempic field' as Smith called it. The Time Field equates with Space, where consciousness is everywhere present. Seems to me Lakhovksy's Multiple Wave Oscillator is a gadget which gives us the opportunity -- for the first time -- of approaching the Time Field with scientifically acceptable equipment, operating in known frequencies.

BIOPHYSICIST CONFIRMS LAKHOVSKY'S THEORIES!

LA "Times", March 28, 1971: "The 'magnetic resonance' of the nuclei of atoms (in living cells) has been used to tell the difference between cancer and normal tissues. This resonance is the reaction of the nuclei to electromagnetic energy --first its absorption and then its emission. The process can be timed.

"Raymond Damadian, a biophysicist at the State University of New York's Manhattan Medical Center, said his achievement may open the way to major advances in the diagnosis of cancer in humans... Like other attempts to find a way of making early diagnosis of cancer, Damadian's research began with a hunt for significant differences between cancer and normal cells.

"He relied on a law of physics on which nuclear magnetic resonance is based -- that the atoms in each element in the periodic table absorb and then emit electromagnetic energy on an individual wavelength that is different from that absorbed and emitted by the atoms of any other element. In his system, radio waves are aimed at atoms in strong magnetic fields. If the wavelength is the right one for that atom, the atom will absorb the energy and then emit it, returning to a 'resting' state afterwards.

"The reactions differ according to the kind of molecule the various atoms are incorporated in. They also differ when very subtle changes have taken place in the molecules. The measure Damadian uses is what is called the 'relaxation time' of hydrogen atoms in water molecules within the (body) cells... He found that it takes longer for protons (nuclei) in cancerous tissue to 'relax' than it does for protons in normal tissue... The relaxation time of protons in normal liver tissue, as one example, was .25 of a second, compared with .86 of a second for cancerous tissue.

"Damadian's findings confirm the theory proposed by several scientists that water molecules in a normal cell are held in a fairly well organized structure by electromagnetic attraction exerted by the molecules in the cell. But, it was theorized, when a cell begins the uncontrolled growth of cancer, the structure begins to collapse because of the cell's increasing disorganization.

"Damadian thinks the device for detecting cancer in humans (his experiments were conducted on mice and mouse tissue) that may come out of his work would be a radio frequency coil to emit the electromagnetic waves to be aimed at the atoms in the water molecules under study. The device also would include a magnet to create the field required for measurement.

"The coil would be wrapped around the patient while the magnet would be moved back and forth over his body. The reading of how long it takes the atomic nuclei to relax would be made and then computers would do the matching up and make the diagnosis."

Interesting, isn't it, how the 1970s researches of biophysicist Damadian parallel the pioneering work of electrical engineer Georges Lakhovsky in the 1920s? The goals are quite different, of course, in that Damadian only wants to prove the existence of cancer so the victim can go on to be drugged, operated or radiated with destructive X-rays or Cobalt rays, at fancy prices, a procedure inspired by the soulless monsters who guide the American Medical Association from the Lower Astral plane.

The irony of Damadian's proposed diagnostic technique is that it combines two electro-magnetic healing procedures. If it is used on sick people, sooner or later someone is going to be healed of cancer while being diagnosed! Can you imagine the disagreeable shock this would cause in orthodox medical circles? Extensive surgery would have to be cancelled. The sale of drugs would be lost. Under these circumstances it isn't likely that his diagnostic technique will ever receive the approval of the AMA.

So, borderland science will have to continue to carry the torch of research on electro-magnetics in this area, as George Van Tassel is doing with his Integratron up at Giant Rock, California. In his newsletter "Proceedings" he gives more details of his research program there. We learn that in addition to creating a field of soft, high frequency radio waves, and a magnetic field, the Integratron will also bathe the subject in an atmosphere charged with negative ions.

HOPPING UP THE HIGHER FREQUENCIES

"There is one thing you may wish to try on the MWO. Close the center hole in the driven antenna with a washer and screw. Connect this to the driven ring (second ring from the outside) with a 500 mmf condenser, 10,000 volt rating. This increases the power of the higher frequencies. I think you will be pleased with the results. Mount the condenser at the back of the antenna. Mankind has been saturated with the lower frequencies for years; 60 cycles and their first few harmonics are almost everywhere in more than adequate power. Radio, TV, radar, etc. also saturate our world but the amount of power available to the average person is very small. In 1942 I used to cure my sinus trouble in minutes by getting close to the transmitter section of a radar transmitter. Few get that close, though, and it is a good thing as this power can kill. I have seen birds and larger animals killed instantly. It can happen to humans, too." H.B., Melbourne, Florida

SEND A FEW BUTTONS

"Received the Lakhovsky loops and am pleased to feel the response to the body. I wear my belt only at night. I gave one to a friend who was having a problem with her leg and anticipating hospital treatment for same and the response was also to the good... So all I want now is for you to send me a few buttons so I can make some more loops myself and help to lessen some misery around here, for someone. I tried to have some buttons made here, but no success." J.M.S., Charleston, S.C.

The non-slip loop buttons we make are of 1/16 inch thick, white fiber glass plastic.

MORE POWER NEEDED

"Enclosed is $. Please send me the magnet assembly with carbon rod (March-April 71 Journal). Have not had the hoped for results with the loops. Possibly the assembly, which can be used more directly over the research area, will be more effective." M.S., Ellensburg, Washington

If we've proven one thing in our borderland research, there is no single cure for any or all of mankind's ailments. Injury and disease are outward manifestations of an invisible cause, which may involve three or four levels of consciousness and extend back two or three lives. You Director has found in over thirty years of counselling people on their borderland problems, half of them don't really want to get well, they just want to be free of pain so they can go on indulging their vices. And by the time the hidden vice surfaces as a disease in some target organ of the body, an unhealthy trend has been established in the cell life there. It usually requires a strong jolt or shock to reverse this trend, such as a powerful magnetic field, a drug, chiropractic, even surgery -- anything to break the negative pattern which created the condition. One of the simplest and most effective methods, and cheapest, is fasting for a week or two. This gives the system a chance to purge itself of all filth.

OPEN ENDED COPPER BRACELETS

"Your article on copper bracelets sent me to the typewriter! It is quite well known in England and has figured on TV as anklets for racehorses! Open-ended copper bracelets sold freely in shops for some years here, but not recently. However, I have a copper bracelet from Rhodesia which I wear always, never take it off, and of course no arthritis or rheumatism. But I'd like your experimental loops for throat, waist and wrist.

J.G., London, England

THE PRO AND THE CON

"Thanks for the wire rings. Sorry to report no favorable results. My husband refused to try. I wore them several nights but seemed to feel even worse afterward. It was a good experiment. I guess my trouble is psychosomatic."

Mrs. A.H., Enid, Oklahoma

"I want to thank you for sending me the kit of Lakhovsky coils. I especially appreciate your taking the time to prepare a larger-than-usual coil for my chest. I guess I'm, too young to have any really serious physical ailments -- except for insomnia! And the coils eliminated that the first night I wore them to bed!"

J.T.D, Venice, California

*** *** *** ***

From the September-October 1971 Journal of Borderland Research

UPGRADING CELLULAR ACTIVITY WITH ELECTRO-MAGNETISM

By Aaron H. Steinberg, Ph.D.

For too many years now the scientific knowledge of electro-magnetics and its positive effects on cell life has been collecting dust and cobwebs on the shelves of ignorance, brought on by those special interests who fear the Truth. We are now seeing a rebirth, if you will, with scientists both professional and amateur, who are reviewing the work of their predecessors and making improvements in the mechanical application of magnetic fields.

In 1969 I visited the USSR to see for myself what and how they were utilizing Electro-magnetism. In several country homes for retirees I was most pleasantly surprised to see electro-magnetism used daily on most of the old folks. I should say young folks since may who claimed to be over 100 years young, looked and acted like most people at 50! The main object of the E-M was to reverse the ageing process by altering the cellular structure. I was most anxious to purchase a piece of the equipment, but alas no sale.

Seek and you shall find has always been one of my dominant characteristics. So when I learned that Japan was truly advanced in this field, having been using such equipment since 1936, I decided to go there in 1970. I sought out the one company that had been most productive in this field. They were most kind and cooperative. I was introduced to several scientists who had done much useful research. Their claims were backed up by scientific facts and I was convinced that they had a well constructed instrument, simple to use by anyone so desirous. They named their equipment The Magnetizer.

I returned to the U.S. with a Magnetizer and immediately began a research program. I can report that the results obtained to date have been most favorable, after nine months of experimentation. The magnetic flux generated by the coils can be measured accurately over any part of the body, thus determining those areas where the flux is penetrating as well as those where it is weak.

Chemicals and pollution are surely contributing to an unhealthy alteration in our cellular activity and human magnetic filed. It will be many years, if at all, before this mess can be rectified and controlled. But in the meantime, those who are aware and see the danger signals, will not sit by and wait for the clean up, which may never happen. Instead they will seek to upgrade their cellular activity so that the body's natural resistance may be at peak ohms. There is some indication that the aura force centers known as chakras or vortexes are stimulated by electro-magnetism.

From the Japanese scientists I learned the following about the effects of magnetic flux. It is different from ordinary electric current, which only flows along the surface of matter. It is different from X-rays, which do not penetrate the bones. The magnetic flux of ultra-long wave generated by the Magnetizer penetrates deeply in muscles, fat and bones and has an intensive effect on the nerves.

Magnetic flux never causes unpleasant sensations in the body, such as pain or shock, but instead produces comfortable, warm sensations. These sensations are also known as Joule's heat, which strengthens the function of the cell, corrects spasms and inflammations. When magnetic flux passes through tissues, a secondary electric current called the eddy current is created around the magnetic lines of force in the tissue cells, which ionized the protoplasm and rejuvenates the tissues as a result of activating metabolism. Furthermore, magnetic flux, in the process of penetrating the tissues, works to increase hormone secretions. These maintain youth by providing energy as a result of normalizing function of the internal organs.

Flux strongly stimulates magnetic substances in the blood, like iron. Accordingly, the hemoglobin in the blood vessels moves actively, accompanying the lymph circulation, when the Magnetizer is turned on. The therapeutic effect is not singular but collective, thus eliminating constitutional weakness.

*** *** *** ***

From the May-June 1971 Journal of Borderland Research Clips, Quotes & Comments

OSCILLATING LOOPS AND COILS

"I have a necklace which is definitely an oscillating circuit. I'm sure you have seen them. They are a wire around the neck instead of a chain. The only thing is that they come with a hook arrangement to clasp them. I cut the ends off, and rounded them to make a real oscillating circuit. Then, too, I got a pair of earrings which are simply coiled wires, like this:

One of these days maybe I'll have a headache (this is rare), and I will try either or both of these.
Mrs. C.M., Watertown, Wisconsin

REMEDIES, HOMEOPATHIC AND ELECTRO-MAGNETIC

"You might spread the word around that the Copen Vibro-Potentizer actually can make any Homeopathic remedy in any potency from 1X to 10M.

"Since the medical-industrial complex is trying to destroy Homeopathy and its pharmacists, those who know of it and are doing research on its remedies would do well to get such a machine. It can also make a remedy from a radionic rate. I made Skull-Staphylococcus 10M and reduced an intractable sinus infection. The possibilities are fantastic and open up a new field in healing. One can make color remedies and aura remedies with greater power than radionic treatment of the same."

From his own research this Associate is suggesting that homeopathic remedies, being closer to physical matter than radionic treatment rates, may bring physical body changes more effectively. A hundred years ago in Philadelphia, Dr. Pancoast was curing practically every known human disease with colored light, either blue or red. But he (she?) also prepared color-treated dosages of oil or water for internal use, to back up or augment the shining of colored light on the outside of the body.

"May I remind you that the Multi-Wave Oscillator requires an antenna without any edges, to give efficient radiation. Lakhovsky's original had round bars or tubing with spherical ends. Edges lose most of the radiation. In addition the largest ring should be based on that fundamental resonance of the human organism, 3.66 meters as discovered by Bell Laboratories. This could be a quarter-wave antenna length of .91 meters or a circle of pi diameter.

"However, this should be doubled so that the human organism resonance is the first octave. Individual resonances vary slightly above and below 3.66 meters. By giving lower octave and with the variations supplied by the spark gap all wavelengths would be reached. Therefore, the outer ring should be 1.82/pi meters.

"The relative size of the outer rings could be determined by examining one of Lakhovsky's original antennas. The outer ring should be fed current at both ends, quarter wave style, I believe (to a H.V. Coil) not at the middle of the second ring as the Beck antenna does. The frequency of human resonance, 3.66 meters or approximately 84.6 megacycles is the key to radionics. [By my calculations, using the formula wavelength in meters (3.66 meters) equals (=) the speed of light in meters per second (300,000,000 m/s) divided by the frequency in cycles per second, I figure the frequency to be 81.97 megacycles. T.B.] The human sample or witness -- blood spot, etc. radiates around that frequency, varying according to the disease and the person, I believe. The homeopathic remedy also has a resonant frequency and alters the human frequency for better (or worse). In addition there are harmonics and possibly sub-harmonics.

"The resistance rates discovered by Dr. Abrams, followed by Guyon Richards, are still valid even though this work was done 50 years ago. The rate for Aqua Marina (sea water) put on a Copen V.P. produces a very effective remedy good for almost everyone. This rate comes from Richards' book 'Chain of Life', 1934."
S.M.S., Chicago, Illinois

We welcome this Associate's observations on the MWO antenna, radionics rates and homeopathic remedies.
Riley Hansard Crabb, editor JBR 1959-1985

BSRF MWO Update - Issued May 31, 1988

Compiled by Tom Brown from the BSRF Files
In consultation with Eric Dollard & other BSRF Associates
(Contains the complete 1986-87 MWO Update materials)

As there has been a continuing strong interest in the Lakhovsky Multi-Wave Oscillator we felt the need to compile this information into the present book. This documentation is the result of a wide variety of analysis and experimentation with the MWO and covers many years of active research.

There have been many comments made concerning the inherent healing effects of the MWO, but as it cannot be stated enough, the device is merely a tool of research, of which one may draw their own conclusions as to whether or not the device actually has any healing properties. It may be interesting to note that in a personal conversation with Dr. Ita Wegman, Rudolf Steiner stated that electricity, magnetism, and the force upon which nuclear phenomena is based, are all corrupted ethers.

We have found through studying Goethe's indications on the phenomenon of Light and Color in conjunction with Eric Dollard's high frequency researches that electricity is a reflection of the Light Ether. This matter is being discussed further in The Journal of Borderland Research, so it is sufficient for present reference to say that Light has its two poles, red-yellow and blue-violet. The red-yellow pole in nature is longitudinal (direct rays from sun) and the blue-violet pole is a progressive type of transverse wave (rounded sky, Reich's KRW wave/blue orgone). Electricity has its two poles, electro-magnetism: red-yellow-hot, (retarded transverse); and dielectricity: blue-violet-cold (longitudinal). Hence, there are experimental indications that electricity really is a reflection of the Light Ether.

In light of this trail of thought we quote with permission from a 1986 private communication with Trevor James Constable: "The MWO I regard as a false path. The cellular stimulation theories sound and seem good because they are mechanistic, and therefore sympathetic with the mechanistic weltanschaung. With the MWO you can avoid the etheric--or so it seems--and therefore those who feel adrift in the etheric ocean cling to things like this. Healing is one hundred percent an etheric problem. Any dysfunction or hyperfunction not caused by direct trauma has to be tackled via the ether body, and even the direct traumas respond best to etheric correction, such as getting the displaced double back into occlusion through the use of arnica compresses, etc., etc. In between the MWO antennas, there is a general marshaling or focus of light and chemical ethers. Its quite evident even with a minor extension of vision. This is the healing agency, not the sparks. The ether rushes to suppress the EM activity, and because it is low power, the ether accumulates around the spirals and in between them. Any stricken area placed in such a high concentration of formative energy, is going to respond with a return towards normal. BUT, why go to all that trouble when a simple orgone blanket--a layered structure that can be directly applied to a traumatized or dysfunctional area--does a much more efficient, controllable job without that infernal juice. In my time with Ruth Drown, I heard it over and over again, and I pass it on to you: keep that 115V juice away from the organism. The post material EM energy is life negative, but no one seems to feel that they can get results unless they INJECT the alien EM activity into the stricken carcass. It's pure ignorance really. My own experience with orgone accumulators and blankets leaves me with no other choice but to regard the MWO as superfluous. A good rule for any experimenter: If it gets simpler, you are on the right track. Remember always Steiner's dictum: "the spirit of man is always healthy!.....it is the obstruction or dysfunction in the lower sheaths that distorts the manifestation of the spirit and personality via the body."

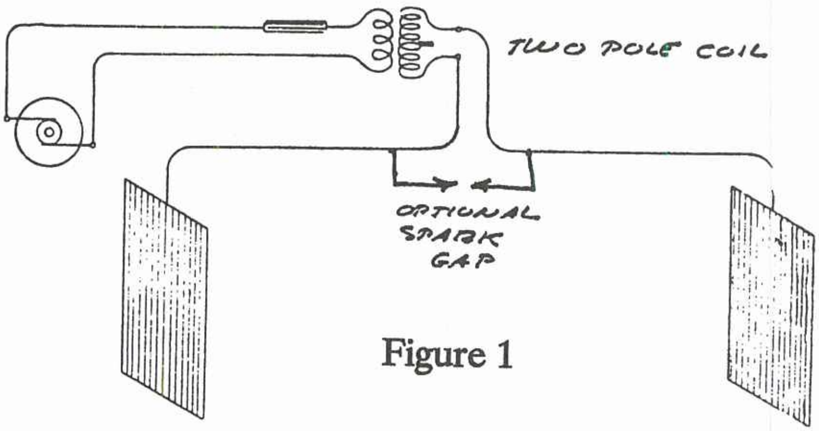
TWO POLE COIL
OPTIONAL SPARK GAP

Figure 1

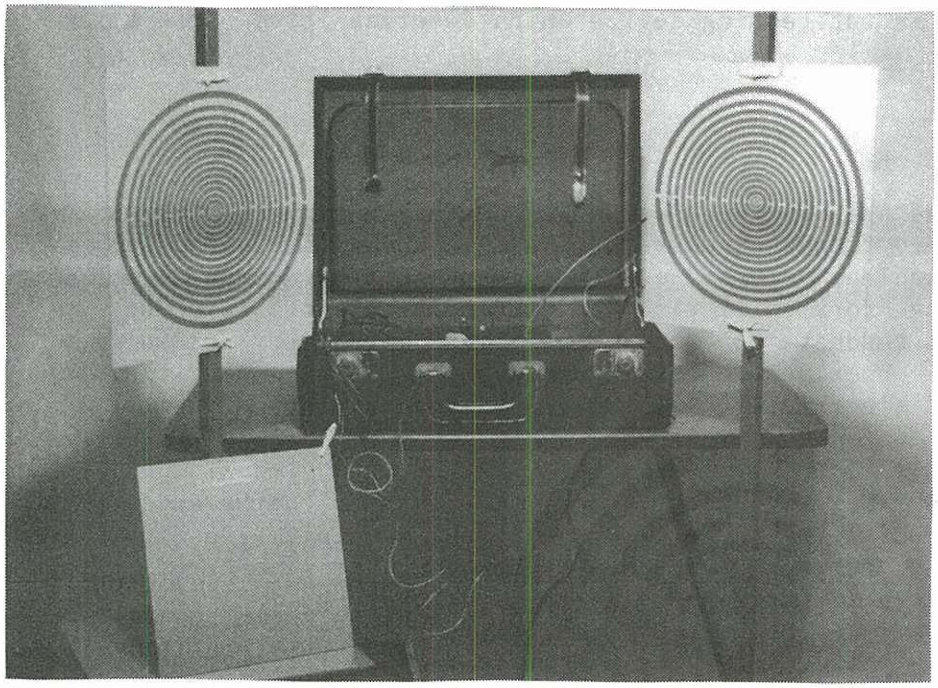

Photo 1
"Beck" style MWO from
the BSRF collection
(note: foot plate at lower
left goes to ground)

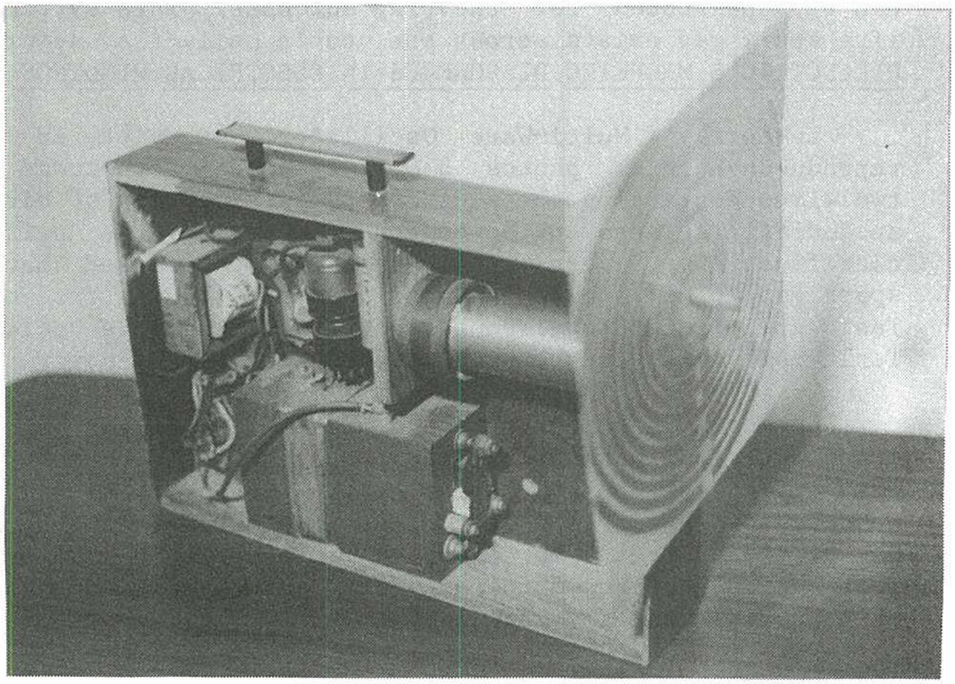

Photo 2
Single antenna version
from the BSRF collection
(Resonator antenna can
be attached at ground)

Not much more need be said in that vein except that of a reinforcement of Trevor's comment. Our present society is not on the path of Nature so therefore some people feel the need for a device such as this. Wilhelm Reich discovered how to concentrate the living energy of the universe, which he termed the Orgone. The Yogic sciences teach us how to accumulate this energy without any outside apparatus, so therefore the simpler and more passive a device is, the closer it resembles Nature.

In the following pages you will find some of the results of MWO experimentation and consideration by BSRF and Associates. We hope the information proves useful, but again it is for research purposes only. We do not recommend the MWO for medical use because we are not qualified to give medical advice, nor do we want to create any false hope where one may neglect professional health care. We do fully support a persons right to investigate and research with devices such as the MWO. No one has all the answers and the MWO is an interesting device which deserves true scientific verification or rejection.

THE BASICS

The general theory behind the MWO is to excite a concentric ring antenna which produces a wide spectrum of radio frequency waves. In this theory it is understood that the RNA-DNA coil in the nucleus of every cell has a resonant frequency within this range. Just as a tuning fork will ring when an identical fork is struck in close proximity, the cells will resonate to their individual frequency pulled from the frequency ocean of the MWO. Most people don't consider that there are different types of waves and two poles of electricity, and these distinction have not generally been noted.

The antennas on the currently popular "Beck" MWO system, serve as simple capacitor plates which transmit dielectric waves out of the Tesla Coil (see figure 1 and photos 1 and 2). The second MWO plate in such a system must be connected to a second Tesla coil or to a double pole (half wave) Tesla coil (May-June 1971 JBR), not a single pole (quarter wave) Tesla coil as per the Bob Beck/Klark Kent style MWO. Action in the MWO antenna is neutralized by the fact that capacitor currents flow radially inward and little electron current flows around the loops. Copper plates can be substituted for the ring antennas. This system operates 1000 times stronger if a spark gap exists across the double coil. A description of the action is in DIELECTRIC & MAGNETIC DISCHARGES IN ELECTRICAL WINDINGS by Eric Dollard from BSRF.

Lakhovsky's Multi-Wave Oscillator, as pictured in THE SECRET OF LIFE and reproduced here as photos 3 & 4, created a structured electro-magnetic field fed by two wires contacting the ends of the outside ring of his antennas. This is obviously an end-fire antenna so the component reaching the person is a novel energy needing clarification. Several researchers have expressed that Lakhovsky used a quenched spark gap which creates damped waves allowing free oscillation in the antenna. A Tesla coil is not required since the Multiple Wave Oscillator IS THE COIL. However a Tesla coil still can be used as the high voltage power supply.

Lakhovsky's patents will cover both types of units so it seems as though he experimented with both, but the pictures in THE SECRET OF LIFE are an indicator as to what he actually used. After having gone over the patents I'm beginning to think that Lakhovsky buried his secret to mislead copycats. That is the purpose for a patent, to protect the inventor. The Tube MWO Patent 2,351,055 gives some interesting insights into the overall picture of Lakhovsky's work and study of it will provide many research leads.

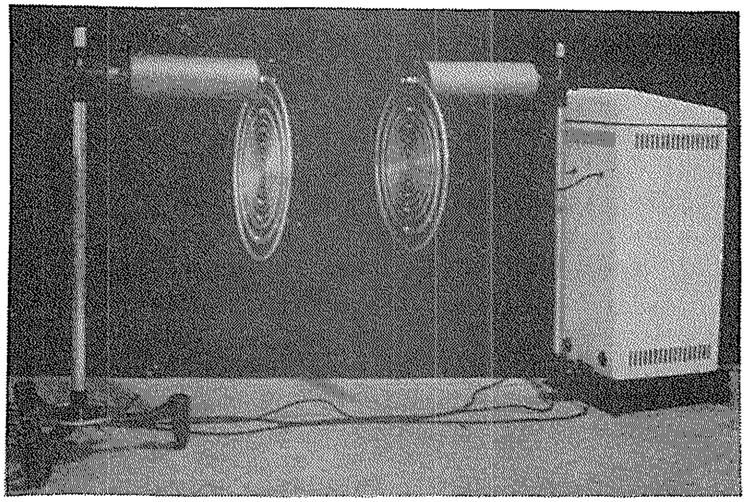

Photo 3
Lakhovsky's Multiple Wave Oscillator

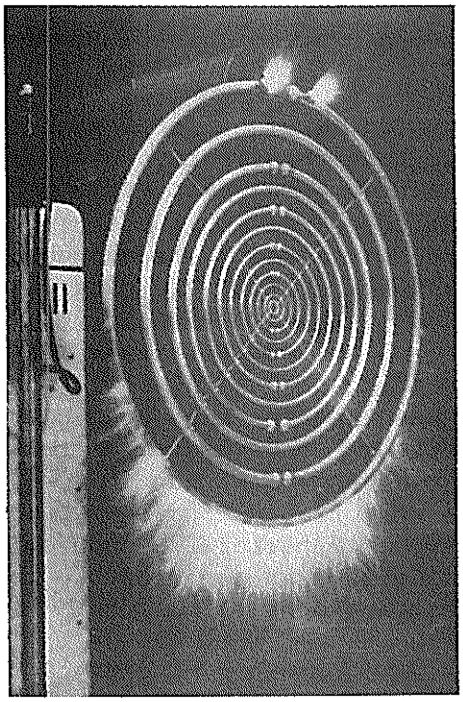

Photo 4
Lakhovsky's MWO. Close-up of transmitter in action, showing *effluve* (electric brush). Note this is end-fire antenna. Regeneration may have come from novel component at 90° to antenna plane.

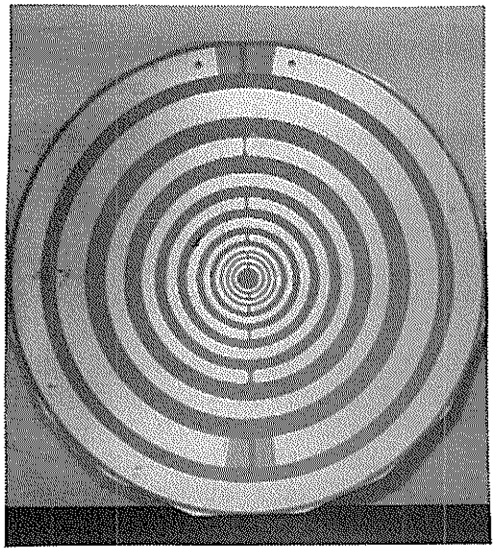

Photo 5
Yagi MWO antenna

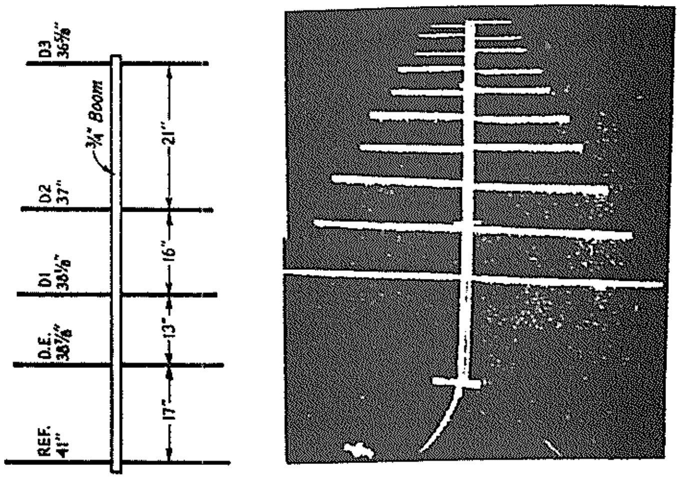

D3 36⅝"
¾" Boom
21"
D2 37"
16"
D1 38⅛"
13"
D.E. 38⅞"
17"
REF. 41"

Yagi antenna

Log periodic antenna

Figure 2

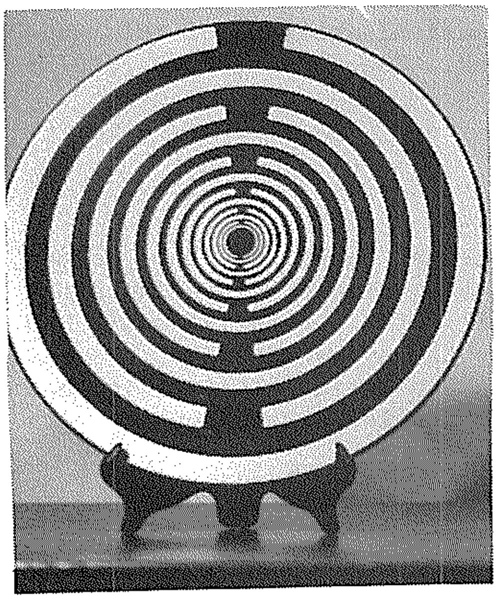

Photo 6
Golden Ratio – Log Periodic
MWO antenna

LOG PERIODIC ANTENNAS

Our research indicates that Lakhovsky worked with log periodic antennas and may have been the first to use them. The antenna will create a harmonious structure in the ether. The Beck style MWO produces a dielectric field 90 degrees out of phase with this, and utilizes the dielectric breakdown (arcing) of the field. This is basically a radio frequency noise generator and the most popular derivation of Lakhovsky's first MWO patent.

Scaling the rings non-uniformly using correction factors is not needed according to log periodic antenna theory (see figure 2 and photos 5 & 6). Lakhovsky states that any high frequency radiating circuit will drive the antennas. We are striving for waveforms mirroring the harmonic laws of nature. As an example the old Yagi antenna has non-uniform distribution of the elements. Log periodic design uses the uniform distribution of the elements such as is found everywhere in nature. This presents superior MWO performance. Lakhovsky's Speaker and Microphone Patent indicates that the log periodic design can be applied to sound as well as radio frequency energy.

LAKHOVSKY'S MWO

Figures 3 & 4 are a preliminary attempt by Eric Dollard to accurately reproduce the MWO system as pictured in SECRET OF LIFE. Figure 3 is a simplified version for a single antenna, figure 4 is a double antenna system. Photo 7 is an experimental setup utilizing this design. We feel that Lakhovsky was working with quenched spark gaps which produce damped waves in the form of the antenna. This type of circuit is needed to power log periodic antennas to create their structure in the ether harmonious with the Golden Ratio structure of the human body and life in general.

This information is not definitive, but is the result of experimental and theoretical considerations. We present this so that Borderland researchers can carry on with new input. We are not saying that the "Beck/Kent" models do not work. There are many claims that they do, but based on Tesla's theories, rather than Lakhovsky's. Figure 1 unit works off of longitudinal dielectric waves (displacement current). Figs 3 & 4 work off of Transverse electromagnetic waves (Hertzian) and residual scalar waves.

Figure 5 is a low power driver based on the Steinmetz lightning wave, a reproduction of the wave naturally occurring in the atmosphere. Values given need to be experimented on for optimum pulse formation. A mathematical analysis of this wave by Steinmetz is included as an addendum to this book. See DISTRIBUTED SERIES CAPACITY.

GOLDEN RATIO ANTENNA

Eric Dollard has designed a log periodic antenna based on the Golden Proportion, as found in living systems such as plants and animals. These antennas are two sided, the first side looking somewhat similar to currently used MWO antennas, the second side being capacitive loading (as Lakhovsky mentions, he used spheres on the ends of his antenna rings) based on the Golden Ratio. This design can be seen in photo 8. These antennas are gold plated to provide direct contact with the ether. Their actual size is 12" in diameter.

Figure 3

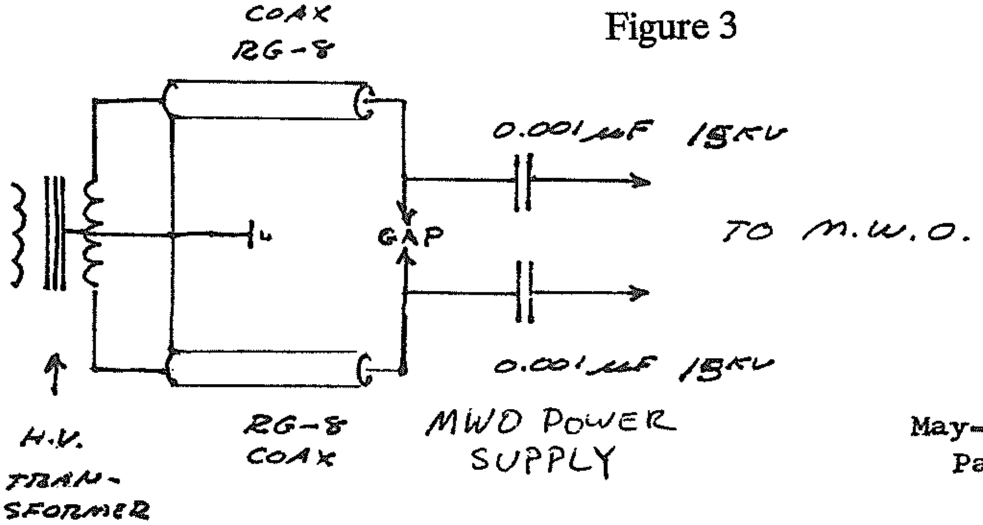
COAX
RG-8
0.001µF 15KV
GAP
TO M.W.O.
0.001µF 15KV
H.V.
TRANSFORMER
RG-8
COAX
MWO POWER
SUPPLY

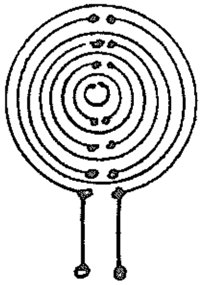

May-June 1986 JBR,
Page 33

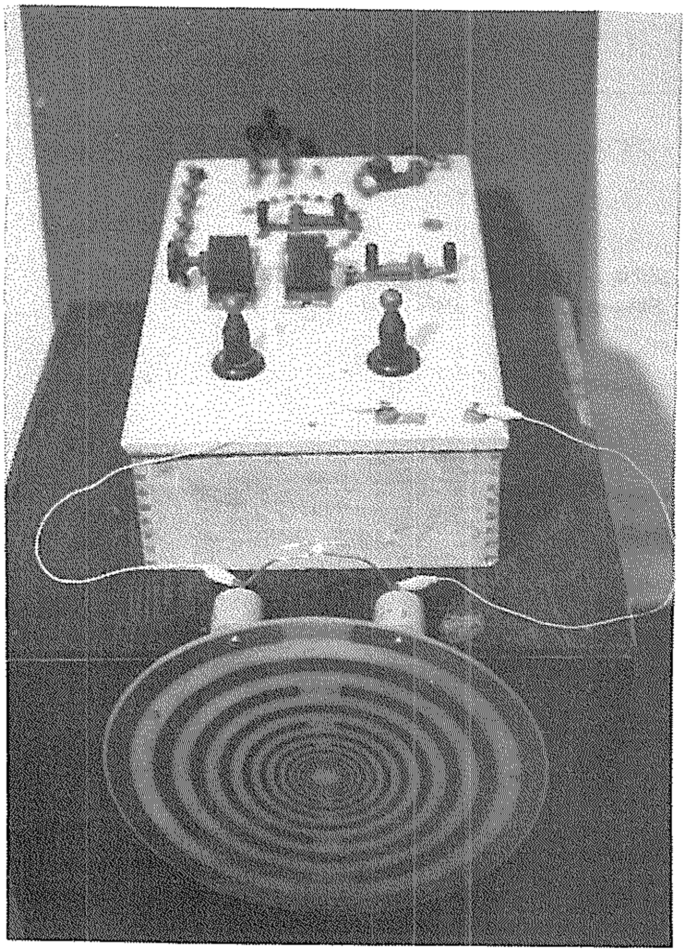

Photo 7
Golden Ratio antenna driven by 1919 era double-phase marble-top medical Tesla Coil. Mid-frequency D'Arsonval currents are utilized in this arrangement. Setup is as per Figure 3 (above), and is our understanding of the arrangement as utilized in THE SECRET OF LIFE. Coil donated to BSRF by John Crane from laboratory of Royal R. Rife.

Action of Lakhovsky setup

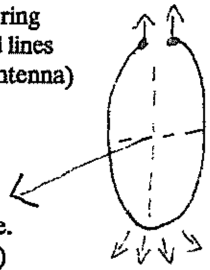
Electrical
connection
Outer antenna ring
(crossed hatched lines
indicate plane of antenna)
Novel component that
may regenerate cell tissue.
(at 90° to antenna plane)
E.M. radiation
(end fire antenna, *effluve* not
necessary during regular operation)

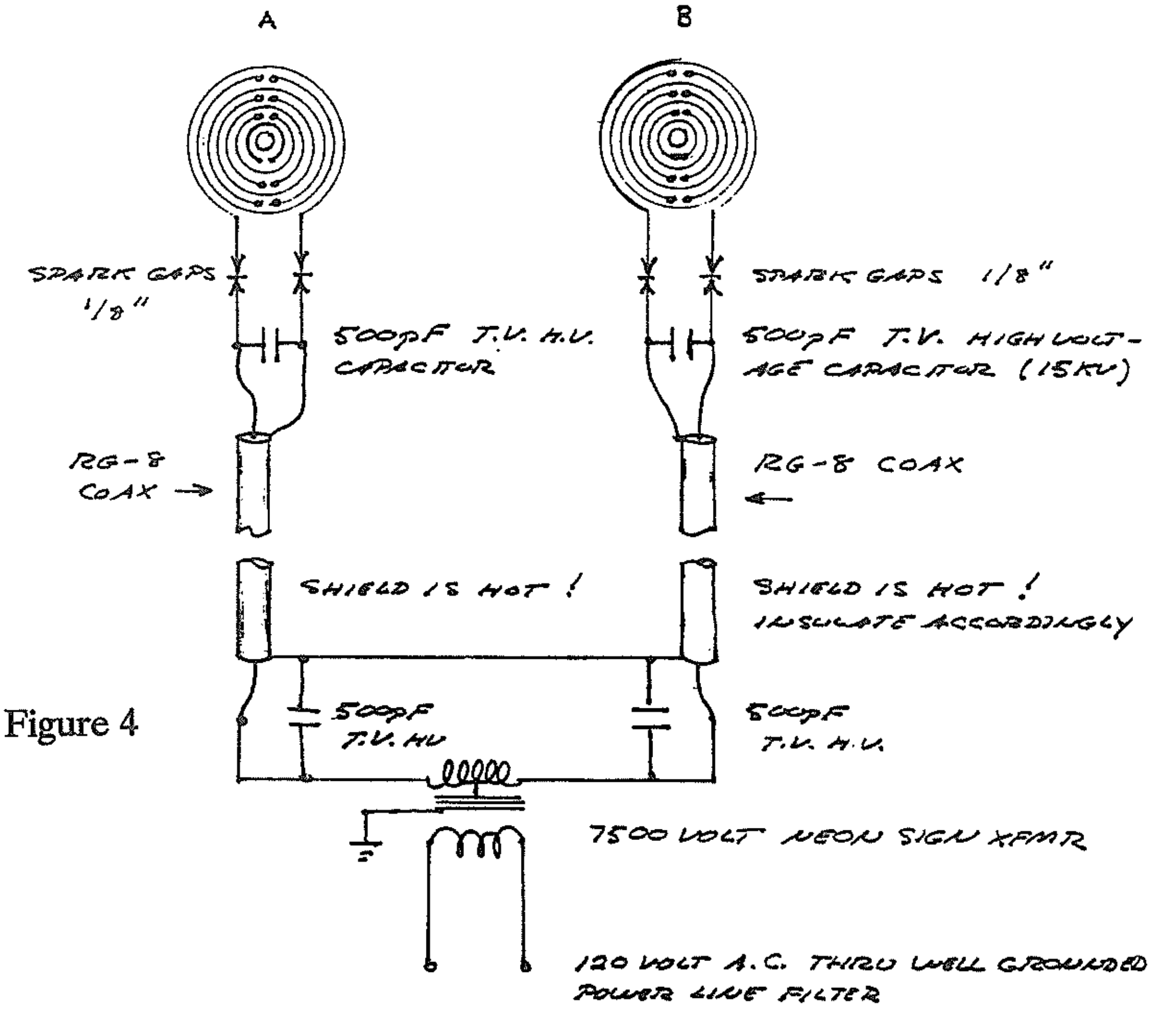
A
B
SPARK GAPS 1/8"
SPARK GAPS 1/8"
500pF T.V. H.V. CAPACITOR
500pF T.V. HIGH VOLTAGE CAPACITOR (15KV)
RG-8 COAX →
RG-8 COAX ←
SHIELD IS HOT !
SHIELD IS HOT ! INSULATE ACCORDINGLY
500pF T.V. HV
500pF T.V. H.V.
7500 VOLT NEON SIGN XFMR
120 VOLT A.C. THRU WELL GROUNDED POWER LINE FILTER

Figure 4

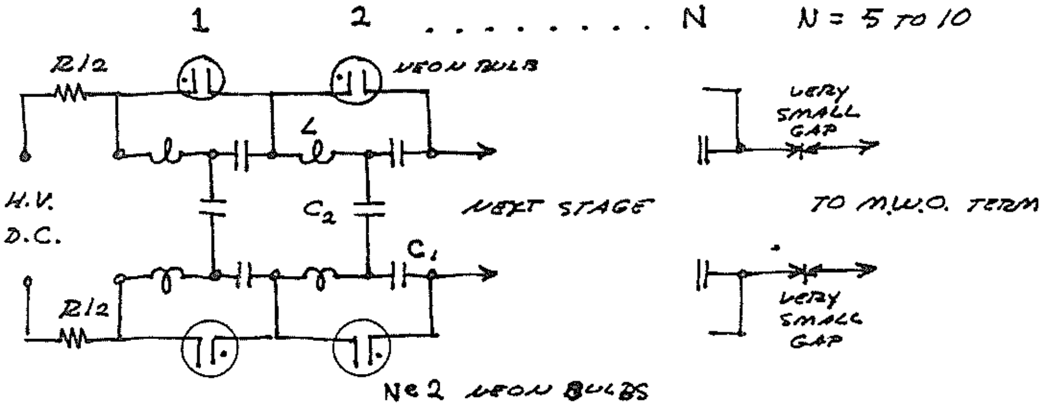
1
2
N
N = 5 TO 10
R/2
NEON BULB
L
H.V.
D.C.
C₂
NEXT STAGE
C₁
R/2
VERY SMALL GAP
TO M.W.O. TERM
VERY SMALL GAP
Ne2 NEON BULBS

Figure 5

D.C. INPUT = 240V / STAGE ALL C_1 RATED 600V

R = 100KΩ / STAGE 1 WATT / STAGE

L = 1 µH / STAGE C_1 = 100pF × № OF STAGES

C_2 = 0.02 µF / STAGE 600V × № OF STAGES

The term Golden Mean, Ratio or Section refers to the mathematics of life, which is a logarithmic function. This ratio can be found everywhere in nature. It is the Sacred Geometry used in the architecture of ancient Greek and Roman temples, the Great Pyramid, etc. It has also been concluded by certain Orgonomists that "the Golden Ratio is a basic mathematical property of the orgone energy" discovered by Dr. Wilhelm Reich. (Journal of Orgonomy, V.8, N.2, Rosenblum, THE GOLDEN SECTION). It is evident from Lakhovsky's MWO Tube patent that he was aware of the Log Periodic design.

The MWO has usually been understood as having to be driven by a high-frequency source such as Tesla Coils and neon transformers generate. However, those who have looked into Lakhovsky's work know that he used simple wire circles to cure plants of cancers, and to make plants grow better in general. The Golden Ratio Antenna is a series of Lakhovsky rings because the rear conjugate strips on the antenna close the loops and in effect make Lakhovsky oscillating rings which are a complete circuit for focusing the ethers. The log periodic (following a logarithmic pattern) antenna design carries this effect further and preliminary indications are that the log periodic design creates a structure in the ether even without power, in may be an orgone accumulator of sorts. Futher experimentation in this vein is being undertaken.

Log periodic antennae have the interesting property of creating a virtual (non existent but functioning) antenna extending about 1-1/2 times beyond the actual. It is also frequency independent and is an excellent antenna for Amateur Radio communications.

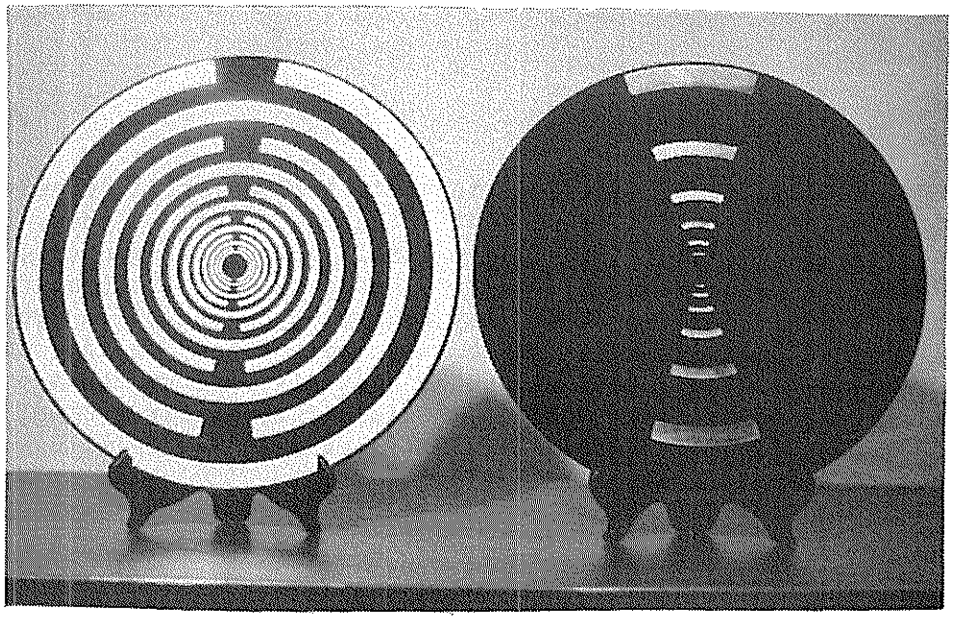

Photo 8
Golden Ratio antennas designed for MWO use by Eric Dollard. Forward side similar to Lakhovsky antenna. Reverse side provides capacitance as did metal spheres at ends of original Lakhovsky antennas. (see Photo 4)

The ethers harmonize to Golden Section Mathematics, (as demonstrated by Trevor Constable in his PRIMARY ENERGY WEATHER ENGINEERING ON THE HIGH SEAS video), so it can be understood that a living organism is a discharge of energy into the Golden Section space. We have observed the discharge of sprout energy into living sprouts when one of these antennas is directed at them. We found that it makes the sprouts grow better. Then we found that simply the pattern makes the sprouts grow better. So here we present you with the Golden Ratio Pattern that you can photocopy and use for experiments. This pattern may be copied as many times as you like for experimentation. We have reports that this pattern keeps the scale out of water heaters and fish tanks.

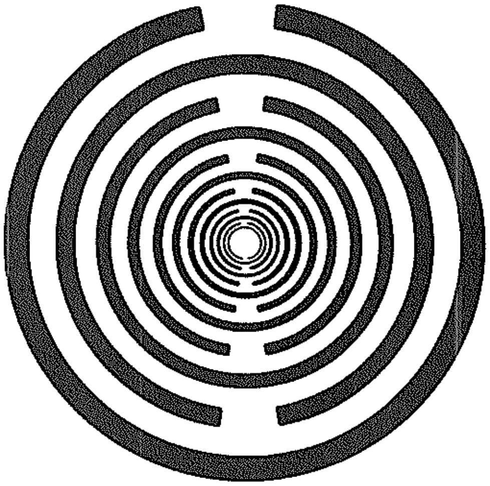

© 1986 BSRF

Log Periodic -- Golden Ratio design for research into the subtle vibrations from patterns.

In 1986 Borderland produced a limited amount of Golden Ratio Antennas in copper and gold. These were precision engineered Log Periodic antennas, and were quickly absorbed into the research network. They can be reproduced if sufficient interest is generated through this book. There is an appendix on log periodic antennas in this book for those who wish to pursue this interesting subject directly related to MWO research.

BUT - DO THEY WORK?

For all the material in this book we still haven't personally seen any solid documentation on the results of either type of MWO that could be re-verified as true to any investigator. I've seen listings and testimonials, but even the photos in THE SECRET OF LIFE look doctored to enhance the cancers (tar can be seen on the wound and the hair!). What was Lakhovsky really up to and which type of device works better? If anyone reading this knows any more than we do then please drop a line. In Lakhovsky's Tube patent he states that his instruments are in use the world over. It is important to check out one of his originals first hand, if any readers have access to such please drop Borderland Sciences a line. We'll make the information available for research. Thanks.

CONCLUSIONS

There are no conclusions at this point in time. Borderland has been publishing MWO information for 25 years now. This book marks the beginning of a new phase of research and development. Perhaps in 1995 after people do some serious research into the subject of the Multi-Wave Oscillator can some conclusions be drawn.

It is important to keep in mind during these researches that just because a frog leg will jump when electricity is applied it doesn't necessarily mean that electricity and life are the same thing. If we labor under this (quite common) delusion we may as well hope for our television sets to "heal" us of our ills. Rather, as astute phenomenological observers, we see that there is a relationship between electricity and life and in the understanding of that relationship we can come to a clearer comprehension of our place in this manifest universe.

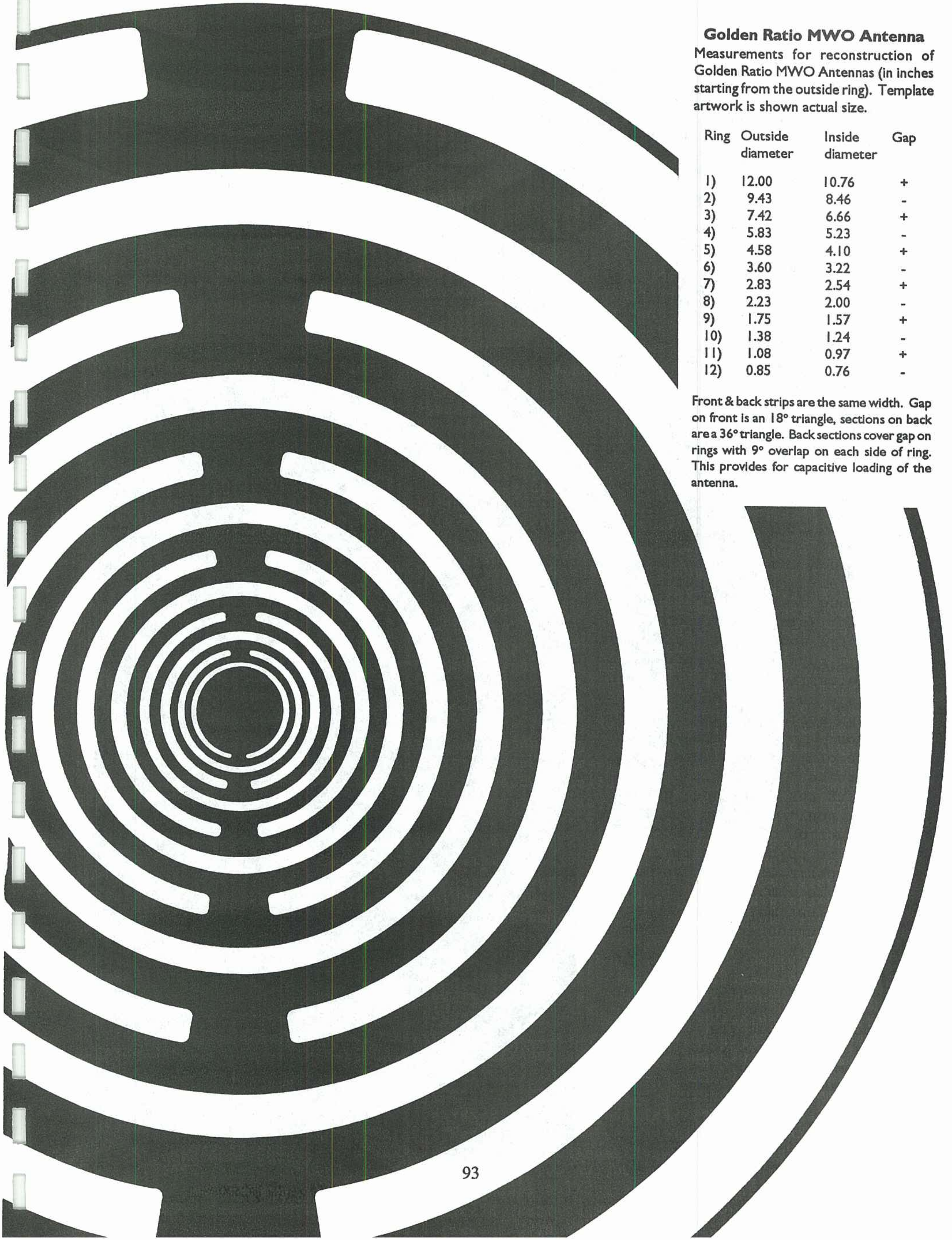

Golden Ratio MWO Antenna

Measurements for reconstruction of Golden Ratio MWO Antennas (in inches starting from the outside ring). Template artwork is shown actual size.

Ring	Outside diameter	Inside diameter	Gap
1)	12.00	10.76	+
2)	9.43	8.46	-
3)	7.42	6.66	+
4)	5.83	5.23	-
5)	4.58	4.10	+
6)	3.60	3.22	-
7)	2.83	2.54	+
8)	2.23	2.00	-
9)	1.75	1.57	+
10)	1.38	1.24	-
11)	1.08	0.97	+
12)	0.85	0.76	-

Front & back strips are the same width. Gap on front is an 18° triangle, sections on back are a 36° triangle. Back sections cover gap on rings with 9° overlap on each side of ring. This provides for capacitive loading of the antenna.

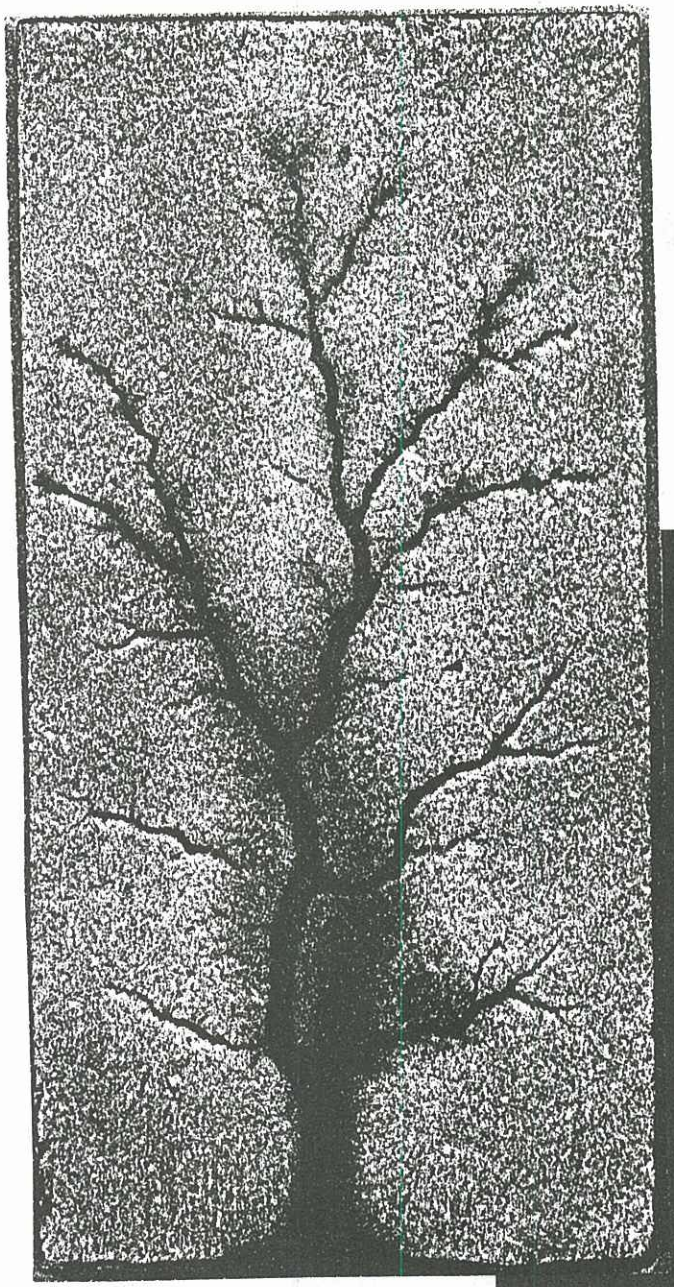

At left is a Golden Ratio pattern burned into particle board using Tesla Coil pictured below, seen here discharging into free space. Scientist Eric Dollard, builder of this coil, has deduced that the Golden Ratio is a property of space itself. The proportions are decided by the amount and type of discharge.

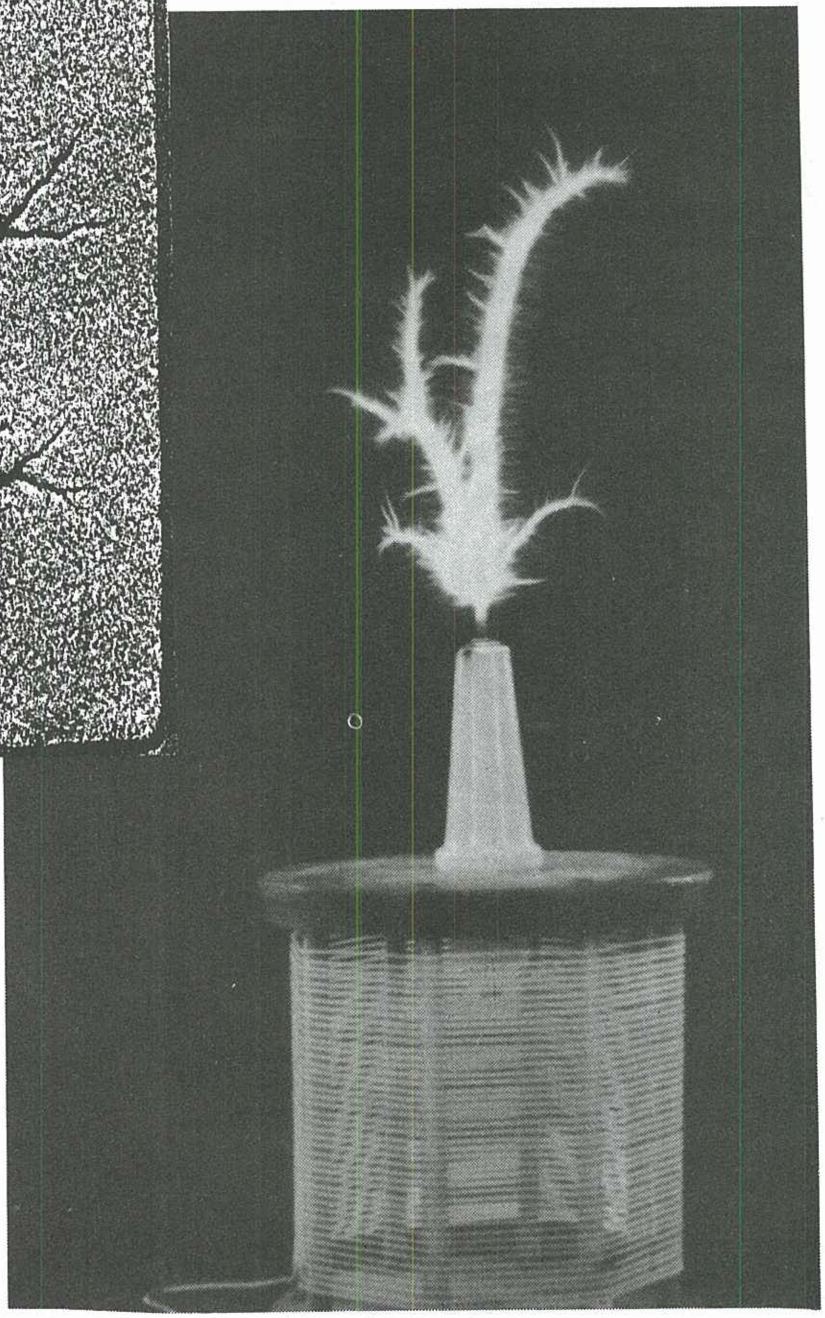

THE SPIRAL COIL MULTI-WAVE OSCILLATOR

Louis A. Schad

DISCLAIMER

Let me make it perfectly clear, I present this information for information and research purposes only and I make no suggestions or recommendations as to its use for healing or medical purposes whatsoever. If you are sick see your doctor. I present to you only what I have learned in my research and observations. I am not a medical doctor and the information I give should be used for education and research only.

To preface this, I have been interested in and studied and collected articles on the unusual sciences for over 30 years and have built many unusual devices. My attention was directed to the Lakhovsky M.W.O. by an article given to me by a friend. I read it several times and studied it and then put it away, (punched holes in it and put it in a notebook). I don't know why -- I can't tell you -- but I couldn't get it out of my mind, so I got it out again and read it some more -- this time going through it sentence by sentence digesting it -- now I was hooked!

I tried to gather all the information I could on the M.W.O., and as usual it wasn't easy -- I had to dig. Bottom line, I started to build Bob Beck's high power multi-wave oscillator -- I still haven't finished it, but I will. The more I studied the M.W.O. (being an average person -- not an electrical engineer, but a mechanical engineer) it seemed to me, complicated, difficult, and had to be adjusted for each person and circumstance. Looking at it now I see that isn't all true -- it was me. Back to square one -- I search for some way simpler, and out of it came the spiral coil M.W.O.

I am a recent member of Borderland Sciences and I have read many fine articles published in this Journal. Many of the articles are very high-tech and delve into the how, where, when and why, and that is how it should be if one is to research. Questions have to be asked and answers have to be found. I am not a professional writer and it isn't one of my greatest talents, so I ask that you have patience with me and I will do my best to tell you about my spiral coil M.W.O.

When I was trying to develop this spiral coil M.W.O. all I wanted was results and I didn't really care how or why it worked, my goal was results, and in doing so I tried many things that I wouldn't have normally tried had I planned each step and tried to fight my way through with pure brain power. In retrospect, I can see that I approached this project with a more light hearted attitude and I have enjoyed developing and building this M.W.O. But at times I did have problems to work out.

I experimented by winding coils, and coils, and coils. I got to be the electric supply house's best copper wire customer, or at least it seemed that way. After boxes and boxes of copper coils, I was ready to throw in the towel and give up -- but having the handicap of being stubborn like I am, I decided to wind one more coil, so I went out to the garage and wouldn't you know it, no more wire. I looked around and all I could find was some old solid strand 16 gauge house wire. It was too light and flimsy for what I wanted, so in desperation I decided to take 2 pieces and twist them together to make 1 heavy piece of wire. Well -- bingo -- I made a coil and got a tingle, that did it. After winding more coils and more coils and experimenting I got to where I am now, so I present to you the Spiral Coil Multi-Wave-Oscillator.

Warning -- this unit is deceiving (it just sits there and looks innocent like it couldn't do much), in fact some of my friends upon first seeing it have laughed at it, and commented you have to be kidding. Another friend of mine asked me to help him build one -- which I did -- and he was lying on his bed in his shorts with a sheet over him and had the unit about 2 foot above the bed and he fell asleep for about 3 hours (he had an infection from a wound in his leg). When he woke up he had a white blister on him about the size of a nickel -- he has done this twice. I warn you, use it with care, it is more powerful than it seems. I think he percolated the infection up out of his leg, as he is fine now.

Now, I know someone is going to say -- I made mine from an old coat hanger and hung it off the coiling with a piece of string, and I'll say -- gee that's nice, do what you wish -- but if you want it to work and work properly build it as it is, use it, experiment with it, and then change it as you wish.

Warning -- persons who have or are diabetic should not use this unit, as it may change the blood sugar level in some people. Diabetes is one of the few things that a M.W.O. will not correct or help. Lakhovsky states this in his book **THE SECRET OF LIFE**. I think I know why but I won't get into it now.

NOTE -- The spiral coils of this unit work all the time -- that is -- they never stop -- they work with or without the vibrator. Who said there is no free energy? If you put it over a bed with an electric blanket, it will charge the blanket and you will probably have difficulty sleeping. If you do not swing it far enough away from the bed at night it will cause unusual dreaming.

As near as I can tell this coil broadcasts at about 120° angle, although its peak energy is centered straight down from the small coil.

Question: is it patented or am I going to patent it? Answer: NO -- as far as I'm concerned it's free to everybody and that's as it should be.

The advantage of using a spiral is that one should get all the frequencies up to the maximum diameter of the outer diameter (O.D.) of the spiral. If you draw a spiral on a paper and place one point of a compass at the center, you can choose any diameter you wish up to the coil O.D. diameter and the scribing diameter will fall some place on the spiral.

This unit can be operated in 6 modes:

1. The coil only.
2. The coil plus the vibrator.
3. The coil plus the vibrator plus the coil shocker.
4. The coil plus the vibrator plus constant coil charge.
5. The coil (no vibrator) plus the coil shocker.
6. The coil (no vibrator) plus the constant coil charge.

So as one can see there is plenty to experiment with here.

NOTE -- I also tried connecting a ion generator to mine -- that really blows me away -- try it if you have one.

If you are experimenting with magnets -- be careful if you use them with this unit as some people get bad reactions. I personally have used magnetic water and this M.W.O. and have had no bad reactions.

Warning -- although I have -- I suggest you do not use this on the head or above the neck.

Warning -- if you have a heart problem do not use this unit as in some people it tends to reduce blood pressure and lower the pulse or heart rate.

Well -- now that I've scared the hell out of you, if you still want to build it I'll tell you the good side. In most all people who have used it or tried it they get a tingle or a feeling of heat (light). Most people get a boost of energy when they use it and it seems to help circulation. I am not sure why. I don't know if it tends to dilate the blood vessels of not. One man that used it can now wiggle his toes and he hadn't been able to do that for about 3 to 4 years.

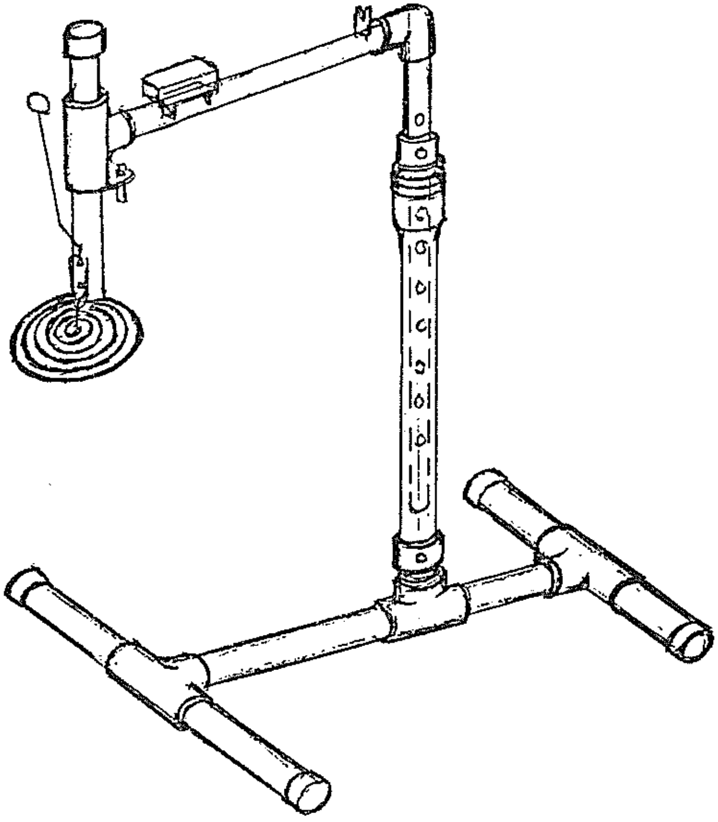

SPIRAL COIL MULTIPLE WAVE OSCILLATOR

I know someone will ask what type or kind and amount of energy is emitted by these coils and the answer is I do not know, but I do know that it works. I will leave that to someone who has the proper instruments and the know how to check it, so there is still a lot of work to be done here if someone wishes to do it. Now, before someone starts climbing my ladder, let me say that I give the credit to Lakhovsky -- what I did was to modify, update and expand upon the base of some of his work. You do not have to be gifted in the knowledge of the sciences to build this M.W.O. and do not need special equipment or tools. However, there are little tricks that I have learned that make it easier to build that I will give in another article. The body or base of this unit is made of P.V.C. pipe, also the vertical and swing arm, the vibrator base and also the drop tube that hangs the coils. We want a unit that is low in cost, tough and durable, light in weight, corrosion resistant, portable, easy to fabricate (and repair if necessary) and looks nice, so we will build our unit out of P.V.C. pipe.

ALL COILS ON THIS M.W.O. MUST BE RIGHT HAND COILS, IN RELATION TO THE UNIT AND EACH OTHER. If they are not you will get negative energy. (I'm not speaking of polarity here.)

I use a fish tank air pump (modified) -- (Challenger 1) about $8.50. This provides the oscillatory motion to the coils. I'm not in love with it and would like to find something better -- something of higher frequency, but it works. If anyone has any suggestions on this I would like to hear from them.

I chose to design the unit so it can access any part of the body freely without a lot of excess adjustments. The body when lying down is in one plane or height (approximately), so I built the unit so it would extend over a bed or table, added a swing arm, and a vertical adjustment (holes in the vertical tube arm and a wood dowel or pin stop) -- that seemed simple enough, remember we're trying not to get complicated here.

If you are going to build this M.W.O. I strongly suggest that you read Lakhovsky's book, THE SECRET OF LIFE. When I finally decided to try to build an alternative type of M.W.O., I read everything I could get my hands on about the subject. I read carefully and took notes of key points, functions etc., then I put my books away and studied my notes -- the basics.

M.W.O. -- MULTI-WAVE OSCILLATOR

MULTI -- meaning more than one or many.

WAVE -- energy (given off)

OSCILLATOR -- I looked this up in Webster's -- and simply stated it says -- motion deflected from center to one side -- back to center -- deflected to the opposite side -- back to center, OR -- Vibration.

Now, Lakhovsky in his book stated he made an oscillator of 1 loop or coil of wire errata: This page was inadvertantly left out of the book.

electrical supply house. If you cannot get bare wire, you will have to get plastic coated and strip the coating with a sharp jack knife, taking care not to nick the wire, or yourself.

2. You need a piece of aluminum bar 3/16 x 1-1/2 inces, 20 to 24 inches long.
3. You need a medium size pair of pliers.
4. Soldering iron and acid core solder.
5. Small coil of 19 gauge coated (shellacked) copper wire.

Now – we have 17-1/2 ft. of 10 gauge copper wire, lay it out on a flat clean surface and straighten it out taking the curve out of it and the kinks and bends. Next we need a secure and strong place to fasten the ends of the wire to – like a trailer hitch or bumper of a car or truck etc. After you have found a place to secure your wire, fasten both ends of the wire to it as shown in Fig. 1.

Slowly pull the center of the wire so that you have 2 pieces of wire, put your aluminum bar in the loop and pull tight. With a pair of pliers close the wire down on the bar tight so that it passes between the screws. Pull hard on the wire with the bar – it won't break. Start twisting as shown in Fig. 1 – always keep tension on the wire – do not let go of the bar. (COUNTER CLOCKWISE). Twist the wire, stopping at time, give the wire a couple of good yanks. NO, it won't break. Twist the wire until you have a good tight twist on it. When you think you have twisted it enough – twist it a little more. When you get it where you like it – SLOWLY back the pressure off of the bar.

Remove 1 screw on the bar and slide the loop out or off of the bar. Unfasten the wire where you secured it and cut the spiral wire as close as possible at secured end. Lay the twisted wire down flat. Measure 3" down from the top of the loop as shown in Fig. 2 and mark it with a pencil or pen. Hard or tight bend the wires as shown in Fig 2, 90 degrees. Grip the twisted wire at the bend point with a pair of pliers, and start forming the spiral. Continue forming 1 or 2 loops. Remove the pliers, grip the wire and hand form the spiral as shown in the drawing #2. When you have formed about 4 loops of the spiral, stop and apply light heat to the coil with a hair dryer or similar device to relieve some of the stress in the wire, so that the loops that you have formed will stay where you have bent them. Repeat the operations and continue forming until you have your spiral.

DO NOT BEND END OR TIP IN YET.

Put your spiral in the freezer of your refrigerator for 30 to 45 minutes, and freeze it. This will tend to stabilize the copper some and relieve some more of the stress.

In Fig. 4 we can see that if we bend a piece of metal the inside of the curve is compressed and the outside is stretched, this stress is what we are trying to get rid of. About now some of you will be thinking – GAD – he even tells us how to hold the hammer. I am sure that some of your who are pro's at metal forming and fabrication can build this thing – WHAM-BANG, but please have patience, not all are gifted in some of these skills.

When you have formed your spiral and it is shaped as you want it, hold the loop of the spiral with your thumb and point fingers at eye level height and sight across the spiral, it should hang so that all coils are level and even with each other, so that it is flat – if not form it so it is. Put the coil in the freezer again, repeat freezing operation several times, this tends to age harden the coil some.

With a very hot soldering iron solder the end of the coil, wash it in warm water, dry it, and bend end in as on drawing #2. After you bend the tip in, bend it down a little bit so the bent in tip clears the crossed coil by about 1/4 inch.

Drill 2 small holes in the spiral at points marked on Fig. 6, taking care not to break the drill, use light pressure. Cut 1 piece of coated coil wire (19 ga.) about 4 1/2 feet long. Push one end of wire (about 1-1/2 inches) up through bottom of hole in spiral, and wrap wind (twisted) loops of wire to second hole. Straight wrap 1 loop at hole area, this locks the standing wire. Pull standing wires tight with pliers and cut standing wires to about 1-1/4 inches.

Refreeze spiral and warm to room temperature. Sand or remove shellac from standing wires down about 3/4 of an inch, this makes electrical contact for the coil shocker. You have finished the main spiral coil. You have just done the hardest part, the rest of it is easy. Save your aluminum bar; you may want to build other types and shapes of coils and spirals to experiment with. I have wound and tried 2 wire coils, 3 wire coils, cone shaped spirals, step spirals, spirals with wires attached and hanging down with soldered tips etc. There is no end – you can experiment to your heart's delight, it's only limited by your imagination.

In closing, I would like to tell you a story – a true story about a close relative who gave me permission to tell of his experience with my M.W.O., but prefers to remain anonymous. I will not elaborate, but just state the facts as told to me by him and his nurse.

This relative took a trip by car to the southwest part of the U.S. with some friends. Before he left he called me and said that he would come to California and would stop to visit with me. This was in the late summer of 1990. After arriving at his destination in the southwest, the weather turned very hot for the whole western U.S. Due to this extreme hot spell, he phoned me and said he was returning home and would not come to California.

When I talked to him on the phone before he left, I told him I had built a Multi Wave Oscillator for him and I would give it to him when he came to see me, and he could take it back with him. He did not make it to Calif. but returned to his home in the midwest. After returning home he became ill and went to see his doctor. After being checked by his doctor, he had tests, and the doctors discovered that he had cancer of the large intestine. When I phoned him and found that he had cancer, I asked him if he wanted to try my M.W.O. I told him that I could not guarantee that it would help him – but I said – what do you have to lose. He said yes, he wanted to try it, so I sent it to him.

When he received the M.W.O., it was assembled from instructions I sent him. He used it twice a day for approximately 6 weeks. He had additional medical problems, he had diabetes and congestive heart failure plus the cancer plus a back problem. I had made the M.W.O. for him in hopes that it might help his back. When he went back in 6 weeks to see a cancer specialist doctor they gave him extensive tests and could not find any traces of

of a given diameter that overlapped. Also, I have read that it isn't the intensity or power of the unit that is desired, but the quality of the signal and also the number of frequencies, signals or waves, as the power can be offset by a longer exposure time. It seemed obvious that a spiral should produce all or most of the waves or signals desired and that it would be better than one loop with a limited signal effect. If Lakhovsky could cure cancer in plants and people with one loop it is logical that a spiral should be even better. OSCILLATION -- how about inducing it artificially -- vibrate or agitate the coil or coils.

This unit produces it's peak energy at 12 o'clock at night, but it can be used during the daylight hours too; it's energy diminishes some toward mid-day. Lakhovsky stated this and he is right. I've got to give it to him, he did his homework. Exposure rates or time of treatment with an electronic double screen M.W.O. is said to be 10 to 15 minutes every 3 to 4 days. Exposure time for a spiral coil M.W.O. (rule of thumb) is about 20 to 25 minutes or 45 minutes maximum per day, or less, depending on what is treated.

Lakhovsky also stated that he treated plants, animals and people with belts, collars, bracelets etc., consisting of a single loop of copper wire. From what I can gather, if a subject was wearing a Lakhovsky coil they never took it off.

My spiral coil M.W.O. is half and half -- part 1 loop + of wire, and part electronic. If a subject got 1 exposure every 3 to 4 days with a conventional M.W.O., it means that the subject's system was hit hard or shocked, and it needs time to adjust. A spiral coil M.W.O. requires a longer or more often exposure and this can be an advantage as it gives us more control of the research subject's requirements. The negative side is that the subject has to spend more time more often.

There is one unusual thing that I and others have experienced with this unit. When you first use it the sensation of heat or tingling is light and you require longer exposure time, like about 25 minutes before you feel that you have had enough. After you use it for about 1 to 2 weeks you will find that after 10 to 15 minutes you feel that you have had enough. Simply stated Lakhovsky said that a M.W.O. activates or excites the cells. It's kind of like pushing a car, it takes a lot of energy to get it started, but once it's rolling it's easier to push. Maybe the body just becomes more sensitized to the energy after the M.W.O. is used for a while. In Lakhovsky's THE SECRET OF LIFE on page 75, "The circuit is subjected to any kind of electric or magnetic shock it is then said that it vibrates according to its natural period." So, now what do we have?

1. A spiral coil (Providing Multiple Frequencies)
2. Vibrator (artificial oscillation).
3. Electric current (D.C.) (shock to the coil) provided by a small transformer limited by bulbs turned on and off by a Christmas light blinker or winker.

Now I said previously that most of the unit is made of P.V.C. schedule 40 pipe (white). P.V.C. pipe being what it is has the tendency to sag if placed in a horizontal position, fastened at one end with weight suspended at the other end. We can correct this by inserting a piece of 1 inch diameter .050 aluminum tubing inside the P.V.C. and gluing it with an adhesive or simply by drilling a 1/8 inch hole through both pieces and inserting a #8 sheet metal screw. Remember also I said there were little tricks you can use to build it as given above, so before you get too anxious to build, wait till I can write the HOW TO BUILD article, it may make your job easier and you will not have to change or rebuild the second time. I am trying to save some of you all the mistakes I made.

The heart of this M.W.O. is the large main spiral coil. This coil is in simple terms an antenna. It is a dual antenna, that is, 2 in 1 as it is a receiving antenna and also a broadcasting antenna. The antenna can be shaped in many forms and still retain the right hand spiral. Do we wind it tight toward the center and increase the spacing as we go out? This configuration I have found best, but other configurations can be formed to achieve different effects. I have experimented with different shapes and some have given unusual results. There is still a lot of work that could be done here also. I wound a 3 wire coil that was very effective but I had problems twisting the copper wire and lost several of them. I would suggest that the first coil you make be a 2 wire twist coil. If you want to build the M.W.O. complete you will have a total of 4 coils.

1 large main spiral coil.
1 small spiral coil (that is mounted under the large coil).
2 straight tube coils.

We use these coils because they all have independent functions. We are going to squeeze it and milk it for all we can, to get everything we can out of it. The main spiral coil produces most of the waves or energy, however, due to its design and nature the inside of the spiral cannot be wound tight to create small loops in the coil, so we will add another smaller spiral coil under the large main-spiral, this will give us the higher waves or frequencies we desire. We drill 2 small holes in the main spiral and tight wind a piece of 19 gauge coated coil wire around the twisted copper wire, and pull the wire up through the holes, leaving 2 standing 19 gauge wires. We sand the ends of these and fasten spring clips that hold the tube coils. The other end of the tube coil wire is connected to a banana plug -- into a receptacle, we connect our coil shock to these.

I know there are some of you out there who are like I am, (impatient) and are curious and would like to try this to see if this guy is telling the truth or if he is just full of it. I am aware that the Journal is published every 2 months and maybe some of you don't want to wait 2 months, so for those of you who are like I am -- let me tell you how to build the main spiral coil. This one coil in and of itself won't knock you over -- but you can get started, and you can play with it until I can write the HOW TO BUILD article. This coil will emit energy by itself if suspended horizontally. Fasten it to something, a piece of plastic tube or wood dowel etc. Try it on your feet as the feet seem to be sensitive to these energies. Well -- here we go.

HOW TO BUILD THE MAIN SPIRAL COIL

You will need to purchase a few things.

1. You need 17-1/2 feet of 10 gauge (solid) bare copper wire (cost about 14 cents per foot). You can get this at any

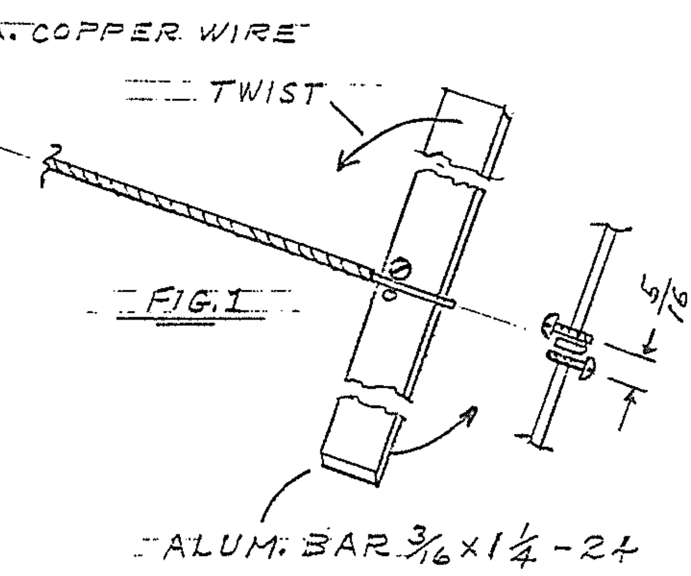
10 GA. COPPER WIRE
SECURE
TWIST
5/16
ALUM. BAR 3/16 × 1 1/4 - 24

FIG. 1

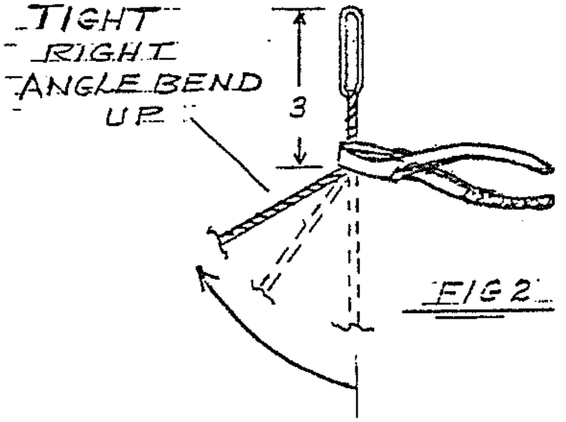
TIGHT
RIGHT
ANGLE BEND
UP
3

FIG 2

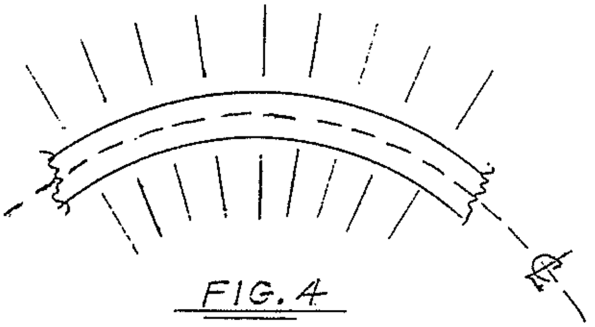

FIG. 4

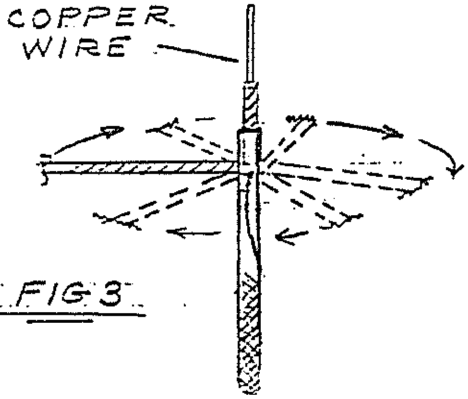
COPPER
WIRE

FIG 3

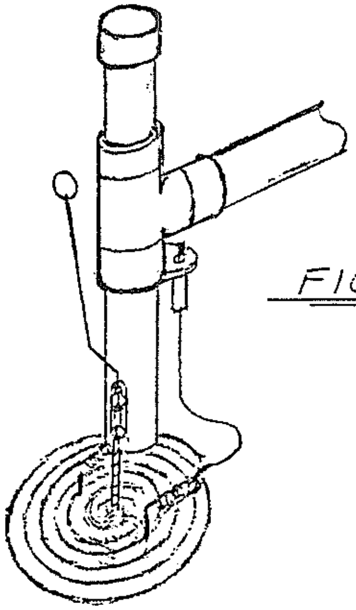

FIG. 5

Spiral Coil Multiple Wave
Oscillator Antenna
Construction Drawings
Figures 1-5

cancer. After his tests he phoned me and said the doctors could find NO traces of cancer whatsoever. I asked him -- did you take any medications whatsoever for your cancer? -- he said NO. I asked him -- did you take any radiation treatments for you cancer? -- he said NO. I asked him -- did you take anything at all for your cancer? -- he said NO. To this date he has not had any return of his cancer.

Now, where does this leave us? He either had spontaneous remission or he was helped by the M.W.O. As I understand and have been told, some forms of cancer in the upper stomach are inoperable. My relative is the 3rd person who seemingly has had remission from cancer from using my M.W.O. I do not like to use the word cured when speaking of cancer, and I say seemingly as this is all in the experimental stages of research.

Then there is Al Mills who built and used a copy of my M.W.O. He had prostate cancer; he's fine now, but we can't count him as he has taken medications for it. I have heard that doctors have said that prostate cancer is very difficult to treat. Then there also is Simon Lucas who built my M.W.O. and loaned it to a lady who has LUPUS (a cutaneous disease due to the tubercle bacillus). He said she is feeling and doing much better now. I also built a M.W.O. for my mother who lives in So. Sioux City, Nebraska, across the river from Sioux City, Iowa. I sent it to her -- she could not assemble it so I flew back to do it for her. Several years back she was leaving a market and caught her foot in a disposed plastic bag on the sidewalk, fell and broke her hip and leg. The doctors repaired it by wrapping screen around the bone and added a few screws and pins here and there. She was OK for several years and then she got to the point where the pain was so bad she could walk only a short distance. She has used the M.W.O. for about 2 months. She called me the other day and said she took the bus all over town, did shopping, came back, took her pull cart and did her food shopping -- 6 blocks up and 6 blocks back -- and then walked up to the restaurant and got her dinner, all in 1 day and she said she feels fine. The doctors gave her the choice of pain shots or cutting off her leg -- (some choice). NOTE -- She does not have a coil shocker on her unit.

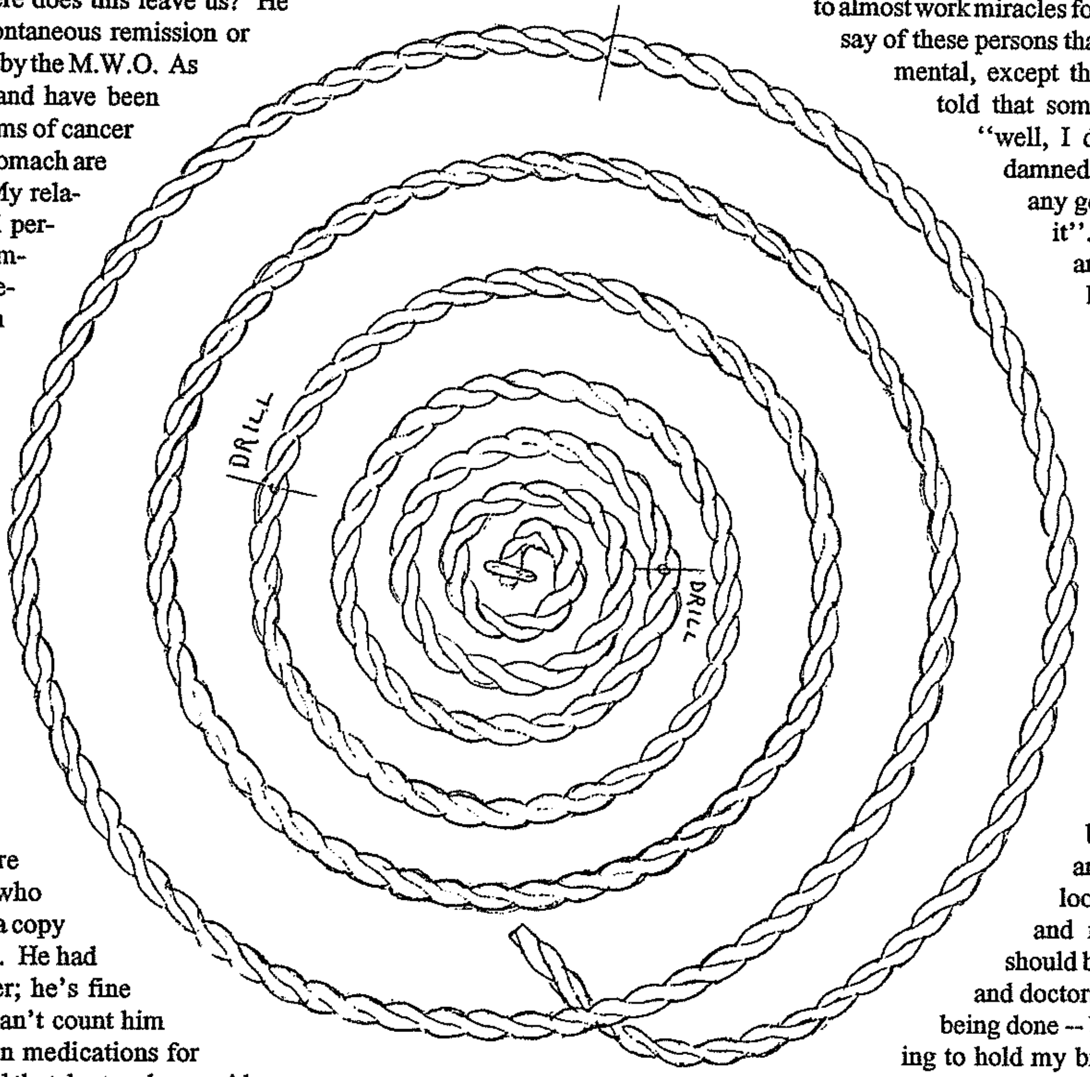

FIGURE 6

In all honesty I have to say that my spiral coil M.W.O. does not affect all people the same. I have had no negative or bad reports to date from anyone using it with good common sense and discretion. There are a few people that it does not seem to help to any great degree. There are also some people that it seems to almost work miracles for them. I would say of these persons that maybe it was mental, except that I have been told that some have said -- "well, I don't think the damned thing will do any good, but I'll try it". I could go on and on but we'll leave it at that.

If you decide to build my M.W.O. I would welcome any serious and honest experimental input as this unit still has many areas that need to be explored. Sometimes I back way off and just stop and look at all this -- and realize that it should be the scientists and doctors doing what is being done -- but I'm not going to hold my breath until that happens. I guess the only way it's going to get done is for people like you and me getting off our backsides and getting it done. If enough will work at it and add a little bit, it'll get done. I sometimes wonder what the end-run potential of my M.W.O. is. I love the doctors (God Bless them!), we need them, but I wish they were more open minded and that they were not so profit orientated and stigmatized by their organizations. I think these ideas I present here will help.

THE SPIRAL COIL M.W.O.

PART II - HOW TO BUILD

by Lou Schad

In Part One I explained how the spiral coil M.W.O. evolved and how I came to build it and notes on things I have learned about it and its use. I also explained with drawings how to build the main spiral coil. For those of you who wish to build this unit, I have added additional drawings and notes so you can build the rest of the coils, base, etc.

Before I begin I would like to address an issue that has bothered me for some time, it pertains to my M.W.O. and maybe other M.W.O's in general. It is a situation or function that has been ignored by many or all but a few. In my article in Part I, I said all coils on this unit must be right hand coils. What I am referring to is what I call RIGHT HAND SPIN. In my study of Lakhovsky's and others work, I have not been able to find any information pertaining to direction of spin, although he gave very precise directions -- how collars, belts, bracelets etc. were to be made. Did he know and was it his secret? Was he not telling? Lakhovsky stated that his M.W.O. stimulated the twisted filament or nucleus of the cell, and the cell oscillated according to its natural frequency. A weak oscillating cell would increase its oscillations back to normal. The question is, do certain oscillations cause the cell to spin right hand or left hand, or clockwise or counter-clockwise, if the frequencies spin right hand or left hand? Some might say -- what difference does it make. The people who use or practice the Five Tibetan Rites always spin right -- never left. Sir Lawrence Bragg (Nobel Prize winner) discovered that healthy people have "RIGHT HAND" spinning blood. And physicist Dr. E. Mersmann discovered that people exposed to electronic smog have "LEFT HAND" spinning blood, the same as people who have cancer -- they also have "LEFT HAND" spin blood. Maybe if the earth spun in the opposite direction, left hand spin blood would be normal. So, is left hand spin or right hand spin important? Yes, I think so. I have purposely made and used left hand coils and they made me feel sick and ill. There is an instrument that has been built that emits weak pulsed electromagnetic waves that causes the blood to spin right hand in people who have left hand spin blood, however if the unit is removed the blood will go back to abnormal left hand spin.

If you put your hand in a bucket of water you can spin it to the right or to the left -- is this polarity? Does magnetic polarity have an effect on matter causing right or left spin? Yes, it seems to. But how about induced spin direction on a non-magnetic or non-electrical energy or wave? If I hold a magnetic compass near the spiral coil of my M.W.O., I get nothing -- no needle movement. If I hold a pocket radio under the coil with vibrator and tune it off station, I get static on A.M. from about 80 on down, none above 90 and up, but I don't think that static is coming from the vibrator. Warning: I damaged my radio by using if for testing. At first I thought it was the battery, but a new battery did not correct it. Sometimes when I turn it on now it doesn't want to come on, by flipping the switch on and off several times it then works. I had no problem with this radio before I used it for testing.

Whatever the answers are, I would suggest you use right hand coils. Look very carefully on pages 113 and 114, these are right hand coils in relation to the subject, (ref. The Secret of Life). If you build the spiral coil M.W.O. you will have the basic unit. You can experiment and try all types of coils and spirals, as I said in Part I, it is only limited by your imagination. I find it fun and interesting to try unusual and weird coils, it's not expensive and you only have to remove a couple of screws, no big deal.

I have been thinking of building a cheap copy of Lakhovsky's coil of tube and wire, also a foil coil of equal size, to test if the coils when vibrated emit equal energies etc., and what the differences would be. It may answer the question why some of the foil coils don't seem to produce the results as good as some would like. In Lakhovsky's book there are pictures of, and he speaks of, the oscillator coils of Hertz. There is a striking resemblance of Lakhovsky's coils on his oscillator antenna to Hertz's coil. Could he have been influenced by or borrowed the Hertz oscillator coil for his unit -- did he expand upon the base of Hertz's work for his antenna coil? (interesting thought)

Enough digression -- back to our spiral coil building project. If you look at drawing #3 you will see that it is self descriptive, giving diameters, lengths etc. We make most all of our unit out of schedule 40 PVC plastic pipe. Take enough care to make your cuts square and clean, debur the ends of the tubes so they will slide into the glued slip fittings correctly. Now, here comes the stinky part. If you want your unit to look nice, you will have to clean the tube, (after cutting), taking off all the dirt, the ink lettering etc. I use a laundry tub, a small amount of water, laundry detergent, Comet, and some wet or dry #200 sandpaper and a lot of elbow grease.

Warning: PVC glue sets fast, in a matter of seconds. I put glue on the female fittings only -- none on the male side, slide it together firmly and whack it with a short piece of 2 x 4. Glue part #7 to part #8 and drill a 5/16 hole through fitting -- this is so you can tighten or remove the threaded fitting by inserting a phillips screw driver or rod in the hole. This makes your unit portable.

Glue pieces 1-1-5-4-4, also 3-3-5-4-4.

Now glue part 2 to part 6, also part 3 to part 6.

Thread assembly 7-8 (large vertical tube) into assemble 2-6-3.

We will slide assembly 11-5-44 onto tube #2, but we will not glue it. Place this on a flat table, and adjust tube 8 so it is vertical.

When you get it where you like it, glue assembly 33-5-44 to loose tube 3, let this dry for 20 minutes.

Now remove assembly 11-5-44, apply glue to fitting #5 and glue it to tube #2 on your flat table. You now have the base finished.

Glue the rest of the tubes and fittings together on your table. NOTE: Do not glue parts #19 and #20, we leave these loose and just set them in place, as we may add something additional here later.

NOTE: Before you glue tubes #11 and #15, cut some one inch diameter .050 aluminum tube full length of tubes and slide them inside of plastic tubes, this eliminates sag. Drill a 1/8 inch hole through plastic tube and aluminum tube, screw a #8 sheet metal screw through both tubes. Tube #11 has holes in it so we can adjust the height, drill these 2 inches apart, 5/16 inch holes. We insert a 1/4 inch wood dowel through fitting #10 and through tube #11.

We drill 2 holes in tube #21 and thread them 10-24, we fasten our spiral coils and antenna to these. The first hole is drilled 5/16 of an inch up from the bottom edge of the tube and the second hole is drilled 3/4 of an inch from the first hole.

When you buy your 1-1/2 PVC tube, get a piece with a flared end, as we set a reducer fitting in it to make a sleeve for the #11 one inch tube. To keep tube #11 guided at the bottom of tube #8, drill 3 holes at the bottom of tube #11 equally spaced around the diameter and insert 3 #8 sheet metal screws.

Parts #9 and #10 will have to be filed out or machined out so the one inch tube will slide through them.

Part #16, the vibrator base, is made of flat stock PVC. If you cannot buy it from a plastic supply house, cut a couple pieces of 1-1/2 PVC tube about 3-1/2 inches long, cut them again length wise, so you have half tube pieces. Place the half tube pieces between two - 1/8 inch aluminum plates and heat in a stove oven with a brick or weight on top. Let them cool and the plastic will come out flat. Layout and mark parts on flat sheet and cut out.

Glue all the pieces of vibrator base together and fit to tube #15 and glue it tight. A wood match works wood to lay glue in joints.

Make and fit your metal brace from aluminum or non-corrosive metal and fasted with two - #8 sheet metal screws. Cut out part #23, fit it to T - #22, drill two holes for banana receptacles and glue securely, also cord holder #14.

Make your small spiral coil from .0625 low carbon stainless steel heli-arc welding rod -- 36 inches long, do not twist wire, use as is. You can get this from any welding supply; get 4 or 5 pieces. Also make antenna from same wire. To make small spiral use the full length or rod. Start by forming an eye on one end for the 10-24 screw. Move down 3 1/4 inches and form spiral same as large spiral coil, also bending end tip in. Put an offset kink in the stem so it will pass through the main coil.

The tube coils are made from #19 shellac insulated coil wire, 12 to 14 turns. I tight bend a 90 degree bend, so I have a 1 inch long tail. I place this parallel on a pencil or wood dowel and tight wind the coil. Grip the wire tail and dowel with a pair of pliers. Remove the dowel and freeze for 30 minutes, warm to room temperature. Grip the ends of the coil with your thumb and middle finger nails and stretch the coil to the desired length, refreeze for 30 minutes. Bend long tail of coil 90 degrees and fit to banana plug. Trim short end to 3/4 inch, clean and solder pinch clip to coil. The pinch clip is made from light gauge hard stainless or brass. Blank size is 7/32 x 1-1/4.

After you have everything glued and finished put a coat of auto polish on PVC tubes, as this keeps them clean and they will not finger mark or get dirty. Also put a coat of vaseline on the threaded fittings so they can be removed easily as they tend to stick.

If you are, or want to be, the serious experimenter type, buy a note book and keep records and notes. Keeping a log with dates, exposure times, etc. can be a very useful tool to compare results of different coils etc. I would welcome hearing from you on the use of the spiral coil M.W.O., i.e. your success and problems. NOTE: Don't expect results in 2 or 3 exposures, as this unit is more gentle and may require a longer time of weeks or several months to get permanent results. To this date I have not had a single report from anyone for over-exposure if used with common discretion. I have used belts, collars, wire loops etc. with the M.W.O., this seems to multiply the energy. I also have used a spin magnetic unit with a reticular energy filter before and after M.W.O. exposure. This seems to work well for joint problems.

Please Note: Lou Schad does not manufacture or sell this MWO. He is giving his knowledge and experience as a gift to the world. He may be written c/- BSRF.

LEGEND

USE P.V.C. (WHITE) SCHEDULE 40

KEY

L=LENGTH D=DIAMETER C=CUT
T=TUBE T.F.=THREADED FITTING

PART #	**DESCRIPTION** (in inches)
1.	T-D=1-1/2, CL=10
2.	T-D=1-1/2, CL=15-1/2
3.	T-D=1-1/2, CL=8-1/4
4.	CAP-D=1-1/2 (4)
5.	TEE-D=1-1/2 (2)
6.	TEE-D=1-1/2 THD'D
7.	NIPPLE-D=1-1/2 THD'D
8.	T-D=1-1/2, CL=28-1/2
9.	REDUCER - D=1-1/2 TO 1
10.	REDUCER - D=1-1/2 TO 1 INV.
11.	T-D=1, CL=28-1/2
12.	T.F.-D=1 NIPPLE
13.	ELBOW - D=1 - T.F.
14.	CORD HOLDER (ELEC.)
15.	T-D=1, CL=20-1/2
16.	VIBRATOR BASE
17.	METAL BRACE
18.	TEE-D=1
19.	T-D=1, CL=3-1/4
20.	CAP - D=1
21.	T-D=1, CL=5-1/2
22.	ANTENNA
23.	F.T.G. BANANA RECP.

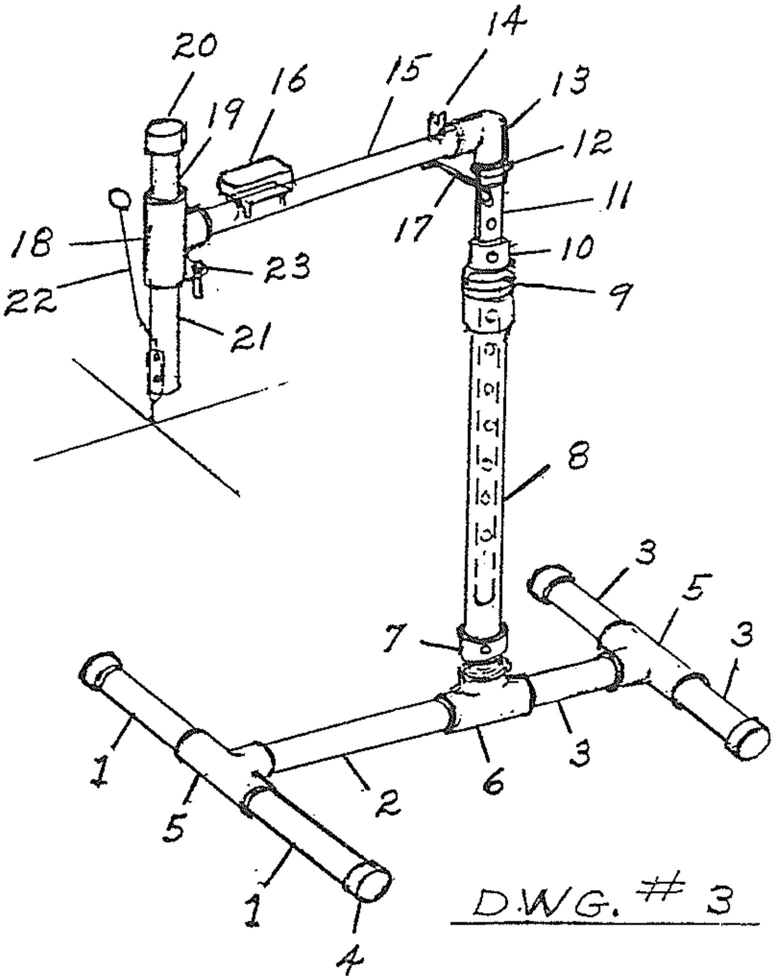

D.W.G. # 3

COIL SHOCKER FOR SCHAD SPIRAL COIL M.W.O.

This unit provides a D.C. shock to the spiral coil. It should be built in a wood box, or grounded metal box for safety sake as we use 115V A.C. power source. All the parts can be purchased from an electrical supply, except auto bulb and socket. Get these from an auto parts supply house.

The Leviton dimmer is not necessary, but it gives more control. I use a ceramic base for 40 watt bulb. A smaller transformer can be used, but I find the 3A-12V best. It is nice to add a handle to the box. I dropped mine once. Put ventilation holes in the box, and rubber bumper feet on the bottom.

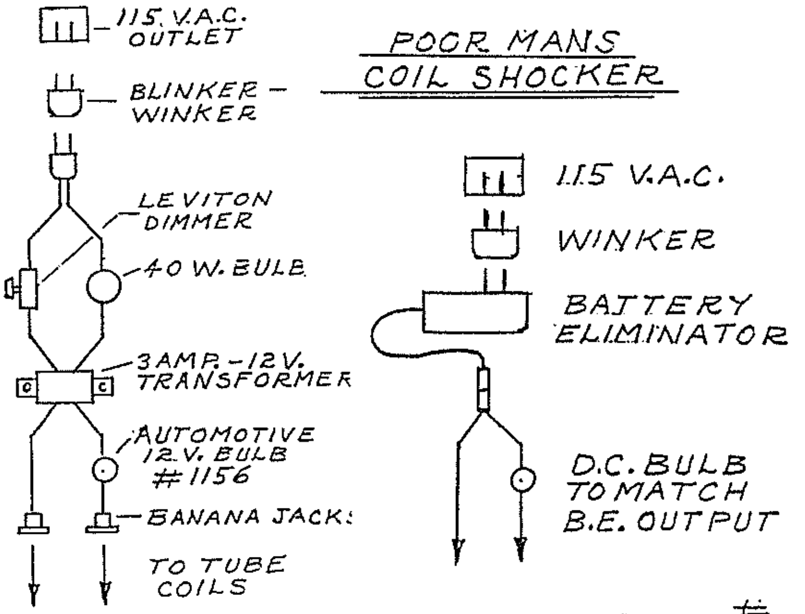
115 V.A.C. OUTLET
BLINKER – WINKER
LEVITON DIMMER
40 W. BULB
3 AMP. – 12 V. TRANSFORMER
AUTOMOTIVE 12 V. BULB #1156
BANANA JACKS
TO TUBE COILS
POOR MANS COIL SHOCKER
115 V.A.C.
WINKER
BATTERY ELIMINATOR
D.C. BULB TO MATCH B.E. OUTPUT

DWG. # 4

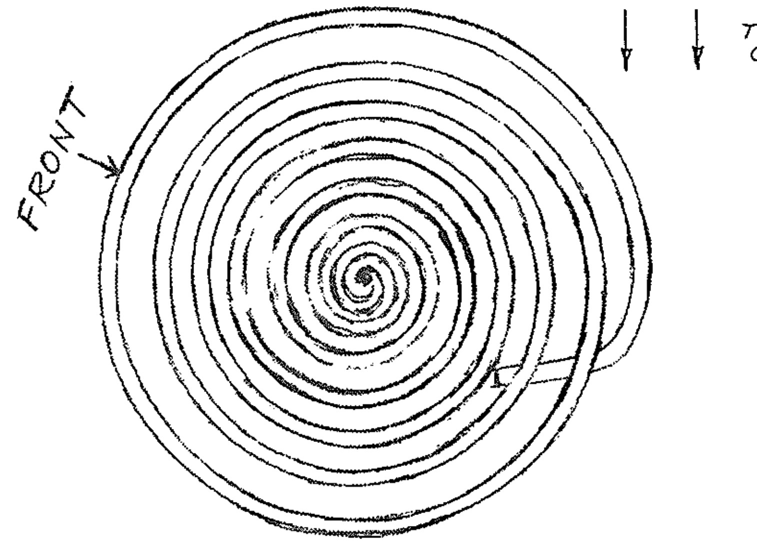

TRACING OF SMALL SPIRAL

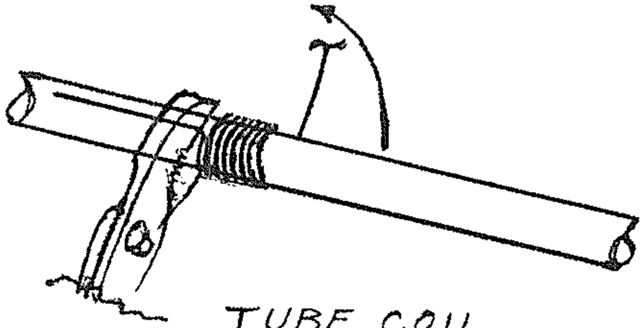

TUBE COIL
#19 COIL WIRE 2 REQ.

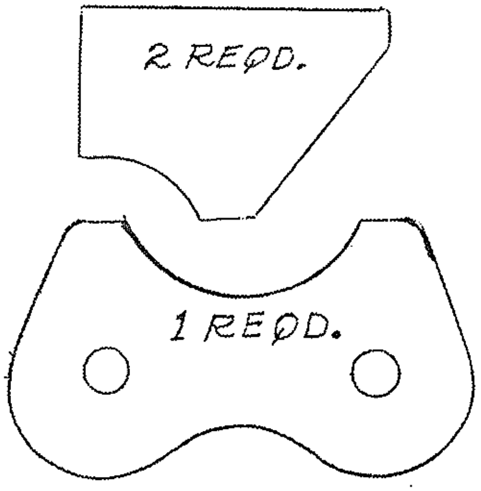
2 REQD.
1 REQD.

DWG. # 5

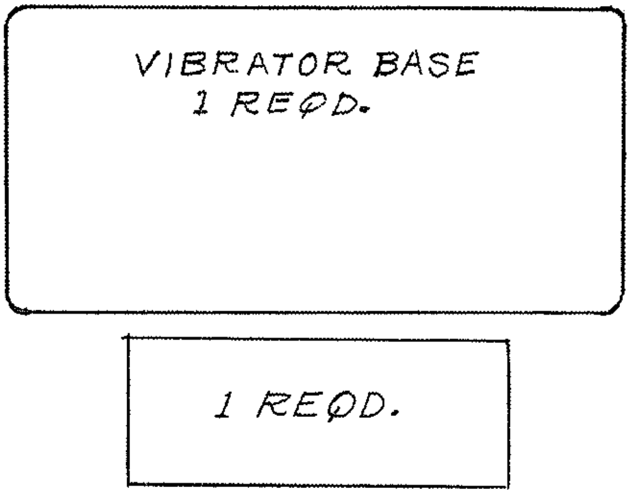
VIBRATOR BASE 1 REQD.
1 REQD.

SPIRAL COIL M.W.O. UPDATE

Lou Schad

This is an update on the work I have done on the spiral coil MWO. I have built and tested several new coils and made changes in design of shapes and twist of the wire and different size of wires etc. I have experimented on plants and myself for the past several months.

In my article in the Journal of Borderland Research of March-April 1991, page 15, figure 1, I stated, to twist the wire counter clockwise, as I was told by someone of authority that this was best. I now find after experimenting and testing that this is not so. I now twist the wire clockwise, causing a left hand spin of the wire when viewed -- or counter clockwise. If the wire is twisted in this manner it more than doubles the energy and the plants and I both seem to like it better. If the wire is twisted as stated it will produce energy not only in a spiral shape but also in single loop rings (overlapped) and also in just a straight length of twisted wire.

I made a coil like a target or a series of one loop rings of twisted wire of different diameters and wire sizes, from about 1 inch to 8 inch and overlapped them about ¾ inch and wired them to a X shaped piece of thin plywood, and it produces more energy than I need. I can stay under it only about 7 to 10 minutes. If I stay under it longer than this, I get over exposure or too much energy in my body, and it makes my feet and hands tingle.

I made a spiral coil from 1/8 inch copper tubing, twisting it like a solid wire, and it produces more energy than a like spiral made from solid wire. I do not know if the energy travels on the inside of the tube and the outside also, but the energy produced is much more than is required for normal use.

I also have been experimenting with new and different ways to pump or project the energy. In my original design I used a vibrator to increase the output. I now find that with the new coils (twisted as stated) that they develop as much or more energy without the vibrator as the old coils did with the vibrator. It seems that the field or energy pattern given off by a straight (twisted) wire is different than those of a single loop (twisted wire) shape or spiral shape.

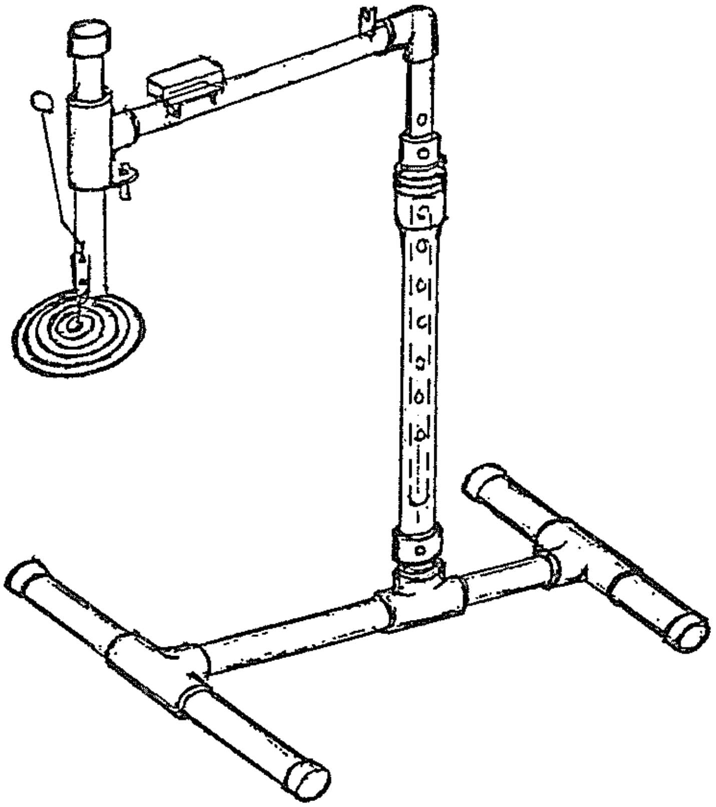

Spiral Coil Multiple Wave Oscillator

I am building a new coil that uses both and looks like a spider with 8 legs of different length. It has 2 single loops (overlapped). One loop is 2-3/8 in diameter and the other is 4-1/8 in diameter. The small loop is centered in the large loop and in the same plane. I made a wood jig with nails for pins to hold them in place and soldered twisted straight lengths of wire pointing from the center outward. I will bend the tips of the straight twisted wires down 90° producing a shape like a spider. I'll let you know if it is different from the spiral coil shape in energy output.

It now becomes obvious that a spiral is not the only configuration that will generate or collect and broadcast energy of unknown origin. If other shapes or designs will also produce energy, is their output different? If so, what are the differences? Are the wave lengths different, are there different patterns to the waves or energy? How does it affect the living organism?

When a spiral or coil is placed in front of a reticular energy field it produces unusual results. Coils can be combined with many other forms of energy, magnetics, light, audio, vibration etc.

In my article I said you can design and build coils of all shapes and types, as it is only limited by your imagination. It is my personal belief that someday someone will design and develop a coil of a nature that will be highly beneficial to humanity, we need your input, ideas and help.

If a coil of twisted wire was built in a target configuration of a series of overlapping loops so as to be similar in design to a Lakhovsky coil, and was powered by a Beck MWO output, what would be the results? If a twisted wire coil will produce and project energy, connecting it to the field of a Beck type unit should be interesting. Are there any takers?

ELECTROSTATIC M.W.O.

INTRODUCTION TO A NEW PHASE OF RESEARCH

by Peter A. Lindemann

Lakhovsky's multiple wave oscillator has always intrigued me. I tried to build my first one in 1976 based on Bob Beck's schematics. These early experiments failed miserably, but I remained interested. In 1987, I had the opportunity to see one of Bob Beck's MWOs at a meeting of the Los Angeles Psychotronics Association. I am a healer and a sensitive by nature, and I was looking forward to finally seeing the mythical machine in action. Bob had said that is was a very high power model and he was right about that.

The unit was of the design that has been popular since the 60s; a high voltage spark-gap transmitter driving an antenna of concentric rings. Bob turned it on and ran it for about three minutes. It made a loud zapping noise as hundred of sparks jumped from ring to ring on the antenna.

I was standing in the back of the room, about 30 feet from Bob's MWO. As soon as he turned it on I felt a blast of irritation hit my aura. I could handle the etheric frying of my energy field for only about thirty seconds before I had to leave the room. When I came back in, everyone's energy had been fried. The meeting quickly digressed into a break, with everyone talking to their neighbor. It was quite a sensation!

In thirty seconds I had become convinced of two things. First, this MWO was extremely damaging to the subtle energy field and therefore must be incapable of promoting systemic healing. And second, this MWO must be unrelated to what Georges Lakhovsky was reporting in his book Secret of Life.

Since I drew those conclusions back in 1987, I have been confronted with numerous testimonials of people gaining benefit from this style of MWO. While I do not doubt the accuracy of these stories, I must admit, I do not understand by what process the benefit is derived.

Here is where the MWO languished in my mind until Lou Schad's articles appeared in the March/April and May/June 1991 issues of this journal. Intuitively I felt Lou was on to something very important. On a visit to Borderland Headquarters, I had a chance to evaluate a unit Lou had given to Tom. The energy coming from the antenna was gentle but very penetrating. I was excited. I went home with the intention of building one of Lou's units to experiment with.

After considering Lou's design carefully, I decided to make a few changes and by late May 1991 I had my first unit. As it turned out, my unit did not look at all like Lou's, but it did embody the fundamental principles that he innovated. These included a spiral main antenna that was charged with static electricity and physically vibrated. Lou's antenna was complicated to build. I decided to use an antenna design that I had been developing uses for for three years, a spiral based on the phi ratio that I could have printed right onto a circuit board. That solved the problem of antenna fabrication. Another thing I changed was the method to induce the physical vibrations. Lou had used an aquarium pump. It worked well but was limited to only one rate of vibration. I decided that the rate of vibration might be important to look at so my unit was made with a variable rate of vibration from 2½ vibrations per second up to about 14 vibrations per second. The last change I made was the vertical antenna. Not knowing the height I needed for the optimum effect, I simply used a retractable replacement antenna from Radio Shack, the kind you usually see on portable radios. This gave me a way to vary the electrostatic charge on the spiral, by raising or lowering the antenna.

By early June 1991 I had made an arrangement with a local acupuncturist in town to run some clinical trials with the unit. This medical professional has asked to remain anonymous and so for the purposes of this article, I will refer to this person as "John". John is a gifted sensitive and an advanced bioenergetic diagnostician using equipment from Germany called the VegaTest. John also has had assistance in this evaluation by another gifted sensitive who can see auras in vivid color and detail. I will refer to this second person as "Pam". As of this writing in early November, their evaluation has been on-going for 4½ months or since mid-June. The results are preliminary. No cures are claimed. However, all indications point to systemic detoxification, energizing and rejuvenation, just what you would expect to see from a broadly stimulated natural healing process.

The VegaTest procedure is very sophisticated. It uses a bio-feedback mechanism to establish baselines for toxicity level, systemic stress and specific organ stress as well as the presence and location of specific pathogens and types of toxic substances. Toxicity level is measured on a scale of 0 to 21, where 21 is the most toxic. A "Biological Score" is developed from the toxicity values that come up and can have from 1 to 6 numbers ranging from 1 to 21, 6 numbers being the most toxic. An "Ideal Score" is determined on a scale from 1 to 15 where normal for adults on first testing range from 7 to 11. Stress on the immune system and endocrine system is measured on a scale from 0 to 5.

CASE A is a 21-year old female who has never been treated with other remedies. The patient was VegaTested, then received one 20 minute session with the MWO and then was retested. (See box insert) These scores suggest that even one session can temporarily reduce stress levels significantly and promote detoxification.

SCORES BEFORE SESSION	SCORES AFTER SESSION
Ideal Score: 8	Ideal Score: 6
Biological Score: 19-16-14-12-9	Biological Score: 12-10-8
5 scores with high numbers	Immune System Stress: 1
is considered very toxic.	Endocrine System Stress: 0
Immune System Stress: 3	
Endocrine System Stress: 5	

CASE B is a 72-year old male with extreme intraocular pressure. The patient received only three sessions with the MWO. The pressure within the eyes lessened with each session. Pam reported that before the MWO sessions, the region of the eyes and head was completely dark and devoid of life force. After the first session, the darkness cleared about 40%. With the second session, the darkness cleared about 80% and with the third session, the energy around the eyes was completely vibrant.

Without going into a lot of specific case histories, or naming disease conditions treated, John and Pam are consistently seeing the following:

1. The MWO expands the energy field around the body 3 to 6 feet.
2. It functions to reduce accumulated stress on many levels.
3. It clears the energetic imprints of illnesses.
4. It fills in breaks and cracks in the aura.
5. It builds vitality at a very deep level and combats fatigue.
6. It promotes the drainage of stale and stagnant life energy down the arms and legs and out the hands and feet. Several times

Pam has seen dark yellow or dark green energy drain out of acupuncture points or organs, to be replaced by cleaner and clearer colors.

7. Everyone experienced improvement, regardless of their condition.

METHOD OF USE FOR THE MWO, (developed by John and Pam during this testing period) is as follows:

1. A person using the MWO should sit in a comfortable chair with their feet flat on the floor and their hands apart, laying on their thighs, palms up is best. The chair should be about 6 feet from the MWO. During the session they should periodically rotate their wrists and ankles to help promote the draining.
2. The MWO should be started at slow vibration rates (2-3) and slowly moved to higher vibration rates (5-9) and then at the end, moved to slower ones again. Sessions can be from 10 to 25 minutes.
3. After the MWO is turned off, the person should remain seated for 10-15 minutes to continue to drain while the energy field slowly shrinks back from its highly expanded state.
4. Rooms where the MWO is used may need to be cleared after its use due to the stale energies that drain out. Sun light, fresh air, toning, sage or other means may need to be employed to keep the MWO session room clear.

THE BSRF MWO PROJECT continues to develop, as this and other articles in this issue clearly show. A special thanks to Lou Schad for his brilliant and original departure from the errors of the past and showing us the true Borderland spirit in action.

One final note. We at Borderland are constantly railing against people who invoke the name of some famous researcher, like Rife, Tesla or Lakhovsky, and then sell the public a piece of equipment unrelated to the original work. This is no better than bait-and-switch hucksterism and we deplore it. Therefore, let us state for the record from the outset, that this new Electrostatic MWO is not what Georges Lakhovsky was doing. Neither is it a direct attempt to duplicate his work. Lakhovsky claimed to have had his units in use around the world, but after years of research we have found no physical evidence of this. He never published detailed schematics of any complete devices. This new MWO is a modern attempt to address the same problems that Lakhovsky was grappling with. We call it a multiple wave oscillator because it is, clearly, a device that produces a wide spectrum of oscillatory emissions simultaneously. Lakhovsky pioneered this line of research. This unit is not a "Lakhovsky MWO" but, strictly speaking, no other unit sold in the last 30 years has been either.

As this MWO is part of an ongoing research project, a limited number of these units will be produced and made available to the public.

The new BSRF Electrostatic MWO is for investigational purposes only. It is not a substitute for professional medical care and is not indicated as a specific cure for any disease condition. Any such use goes directly against the intent of the manufacturer. The performance characteristics of this device have not been established.

MULTIPLE WAVE OSCILLATOR

Borderland Sciences Research Foundation is proud to announce that their new Electrostatic Multiple Wave Oscillator is now available for investigational purposes.

TOTALLY NEW APPROACH

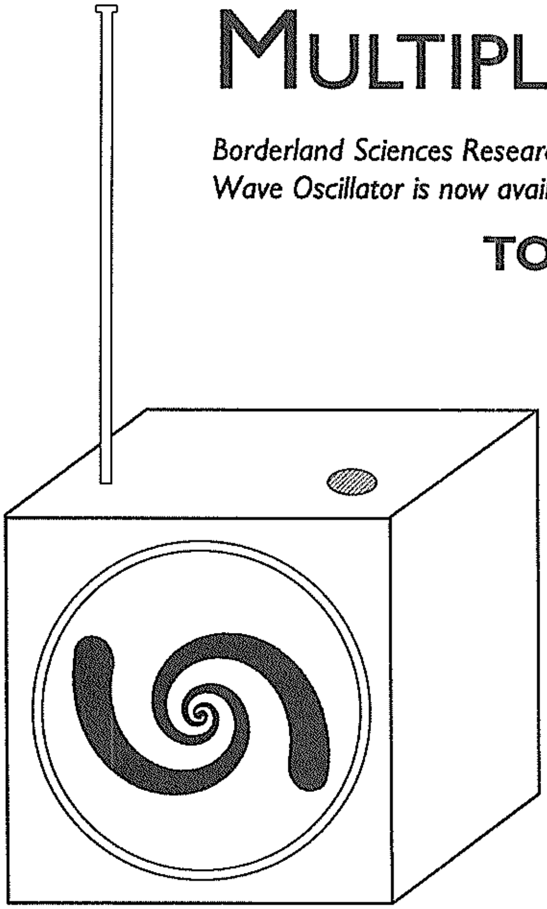

This unique Multiple Wave Oscillator uses a specially designed spiral coil that is both physically vibrated and simultaneously charged with a safe quantity of static potential, unlike designs using high voltage discharges between concentric rings. The result is a wide-spectrum wavefront of electromagnetic and etheric vibrations. It also provides a means to introduce a wide variety of other subtle influences into the broadcast much like a substance biocircuit. The Electrostatic MWO from Borderland Sciences is a versatile investigational tool.

The MWO has a wide range of vibratory rates and charge intensities. It is a one square foot cube with a Golden Mean Spiral MWO antenna on the front and a retractable antenna on the top for tuning. Powered with standard AC. Constructed using no toxic materials.

ELECTROSTATIC MWO

The Electrostatic MWO is a high quality device produced in limited quantity for our research network. Interested researchers may inquire with BSRF as to current availability of units.

SUN-ETHER DISC
By Albert Zock

The German inventor Oscar Korschelt based his cosmic ether accumulator, German Patent #69340, dated July 14, 1891, on the teaching of Wilhelm Weber not realizing that it incorporated the implosion principle as well. Weber professed that all particles of the ether, having an electrical charge, are circling around solid objects without touching them, the molecular particles being negative and the ether particles being positive, according to the law of attraction and repulsion. Since a solid body attracts ether particles, Korschelt searched for a way to collect, condense and to rectify them.

On a wooden disc with a hole in its center, Korschelt fastened spirals of copper wire, on both sides, one wound clockwise and the other, as a mirror image, counter clockwise, and connected the ends together through the hole in the center. His assumption was that the collected energy would be condensed in the center and then be beamed off at a right angle to the disc. He termed the side facing the light positive and the other negative. To increase the condensation and lengthen the antenna, Korschelt wound small coils like key rings and assembled them like links in a chain. He was convinced this device would give additional life energy if the beam was directed at the neck of a person in the height where the 18 cranial nerves are leaving the skull, (medulla oblongata). Such a spiral resembles a Nornen coil, which is stretched out and is used to eliminate undesirable emanations.

An application of a Korschelt coil seems to have its merits. In one case a relative of an elderly man suffering an acute pneumonia, being already unconscious and according to his physician, beyond hope, hung a Sun-Ether Disc over his bed. From this moment, his condition improved rapidly. Fourteen days later, he was almost normal again. In another case, a Sun-Ether Disc hangs over the door in the waiting room of a Naturopath. Every one walking in is amazed over its warmth, and the flowers are flourishing like in a green house, even though it is not heated and faces north. When asked for the reason, the owner points smilingly to the disc and gladly explains how it works and how to build one.

On a plywood disc with a diameter of 50 cm (20-3/4 in.) are two coils, one on each side, wound in a mirror image, the front one counter-clockwise and the other clockwise, almost on top of each other. Both are connected through the hole in the center which has a diameter of 10 mm (6/8 in.). The space between each winding is 15 mm (5/8 in.). Aluminum wire is best. To make the key rings, take a 4 mm (3/16 in.) thick wire, fasten it into an electric drill and wind the armature wire around it tightly, by allowing the rotating wire in the drill to do the spinning. Make sure to wind the links for the front counter-clockwise and for the back clockwise. Then separate every fourth winding and cut it to get the links shaped like key rings. Now attach the rings onto one another like a chain. Each link will lengthen the antenna and work like an oscillator. Fasten the chain onto the board, using nails or screws, but to be accurate, give the chain a light tension so it will not slip off. Twist the ends in the hole together, or better, solder them. The counter-clockwise winding for the front applies to people living in the northern hemisphere; in the southern hemisphere it might need to be reversed.

In Korschelt's time, some made "beam-handbags" carried on a shoulder strap. The material had to be framed to it could not move, then the spiral was embroidered and fasted to the inside of the bag, the whole bag measuring about 20 x 20 cm (8 x 8 in.) and the windings being 1 cm (3/8 in.) apart. Such bags were more powerful than the disc.

From Mar-Apr 1988 Journal of Borderland Research

ETHER RADIATING APPARATUS

An Apparatus For Therapeutic Use Without Direct Or Indirect Suggestion
Oscar Korschelt of Leipzig. German Patent #69340, July 14, 1891
Translated from the German by Bodo Capeller, 1986

This apparatus utilizes the living forces of the ethers for healing purposes. The electro dynamic theory of Wilhelm Weber concerning the structure of smallest parts results in the insight that all ether is identical with Webers electric particles, that all solids rotate around it at a certain distance because of attraction of positive to negative ether particles in relation to solids. Through a corresponding arrangement of solids it ought to be possible to condense the diffuse ether and redirect it. A similar arrangement has been utilized in the following described apparatus which has been named by the inventor "Ether Radiating Apparatus".

This apparatus can be used for the purpose of increasing the intensity of the power of living ether. The difference between several versions of this apparatus is that of shape of materials used. In order to get increased power electric current may be used. Since the same appears to consist of positive and negative particles that try to find equilibrium, the amount of liberated ether particles within the wiring of the apparatus and the diameter of the wire formed ether vortex will increase with the spun off ether particles.

In drawings 1 through 5 an ether radiating device is shown. Within the apparatus the arrangement of the metal wiring is formed in a spiral that will cause ether particles to spin in a predetermined direction so that the living force can be utilized in many ways.

The version in drawing #3 shows the wire is formed in spiral configuration on both sides of plate E and is connected in the center. Through this arrangement the ether particles will accumulate in the center of the disk and be thrown forward. In order to achieve a longer path for the ether vortex one may utilize a chain instead of a simple wire which consists of horizontal displaced links (drawing #5). The disk is intended to cause physiological changes in humans (see drawing #1).

Patent request: An apparatus for therapeutic purposes with direct or indirect suggestion consisting of a non conductive disk (E), free standing or installed in a tube, where both sides of the disk contain spiral wound wire with the side toward the light exposed as the collector and the opposite the radiating portion.

Fig. 1

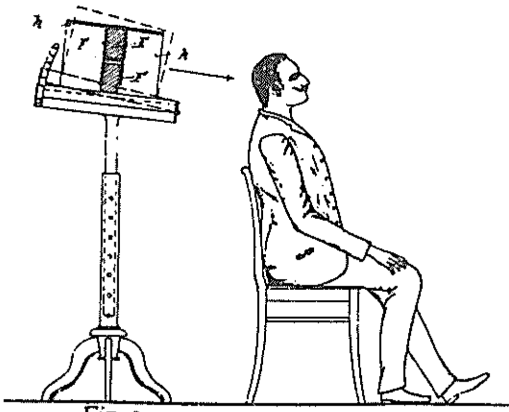

Fig. 2.

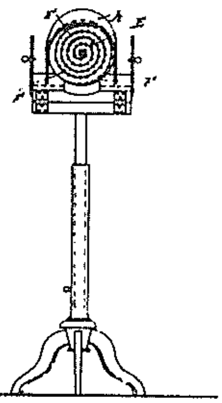

Fig. 3.

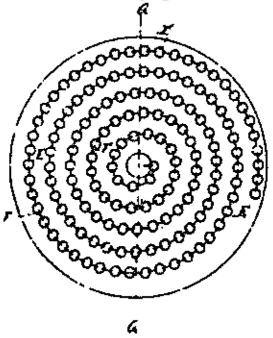

Fig. 4.

Schnitt a b, Fig. 3)

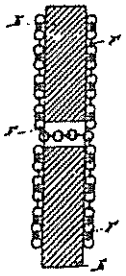

Fig. 5.

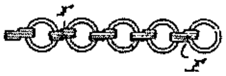

Fig. 6.

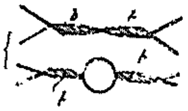

March-April 1988 JBR,
VIEW FROM THE BORDERLANDS
by Tom Brown

NATURAL ELECTRICITY

Ernst Lehrs, writing in **MAN OR MATTER** about early views by Volta and Galvani on electric fish, provides insight into the nature of animal electricity. "The electric organ of such a fish consists of many thousands of little piles, each made up of a very great number of plates of two different kinds, arranged in alternating layers. The two kinds differ in substance: in one case the plate is made from a material similar to that present in the nervous system of animals: in the other the resemblance is to a substance present in the muscular system, though only when the muscles are in a state of decay. In this way the two opposing systems of the animal body seem to be brought here into direct contact, repeated many thousands of times.

"In the electric fishes, accordingly, sensation and will are brought into a peculiar interrelation. For as a result of the spatial fixation and the state of deterioration of its bodily foundation the animal's will-pole is related to its bodily foundation in a manner which otherwise obtains only between the nervous system and the psychological processes co-ordinated with it. These fishes have the capacity to send out force-currents which produce in other animals and in man 'concussion of the limbs', or in extreme cases paralysis and even death. When contemplated with pure observation this phenomenon tells us that in cases of minor strength, this force causes processes of a volitional character in bodies affected by it. Through stronger effects the organic system which serves the will of the creature involved passes into disuse, either for a time or for good. By describing the process in this way we realize that electricity appears here as metamorphosed animal will, which takes this peculiar form because part of the animals' volitional system is assimilated to its sensory system in an exceptional manner.

"It is known today that what nature reveals so strikingly in the case of the electric fish is nothing but the manifestation of a principle at work in the bodies of all beings endowed with sensation and volition -- in corporeal terms, with the duality of a nervous and a muscular system -- and therefore at work also in the human body. Observation has shown that the activities of these two systems in man and animal are accompanied by the occurrence of different electric potential in different parts of the body. Manifold in detail though the relevant observations are, they all confirm a definite correlation of negative electricity with the nerve-and-sense pole (sensation) and of positive electricity with the blood-and-muscle pole (volition)." From **MAN OR MATTER**, © 1958 Ernst Lehrs, Rudolf Steiner Press, London.

Thus we see that natural electricity is formed by processes which can be mirrored in non-living matter. But if we can have access to the primary form of the energy why pursue the secondary, and seek out that "magic frequency" that will cure AIDS, cancer, and assorted body clogs that go under fancy medical names? The answer is that everyone has their individual path to travel. Spiritual healing sometimes takes a long time, while an electronic device may get the person back on the track of their day-to-day life. The spiritual lesson of the sickness is buried for the future. Electricity is the subsensible force, but it is a real force with the ability to alter the functioning of the human body.

November-December 1989 JBR,
VIEW FROM THE BORDERLANDS
by Tom Brown

MWO NOTES

The work of Georges Lakhovsky continues to be of strong interest. I have been investigating the Lakhovsky MWO for about four years now and my views have consistently changed with new information. BSRF's MWO Handbook is getting quite popular! However, my views have changed somewhat since doing that book, and eventually a new edition will have to be put out to keep up with the new data.

What was Lakhovsky really up to? Well, his basic experiments were with copper wires which were circled around plants. Lakhovsky felt that these rings were activated by cosmic radiation. Lakhovsky was extremely interested in subtle energies, such as those that guided birds on their migrations. The copper rings, when placed around plants, would heal cancers in the plants. THE SECRET OF LIFE, Lakhovsky's best known book in English shows pictures of these experiments. Lakhovsky then moved to bring these healing energies to humans.

Lakhovsky also intuited the DNA spiral coil in the center of every cell long before anyone "discovered" it. He felt that a properly generated multiple wave electrical field would make the coil in each cell resonate at its proper frequency, just as a tuning fork resonates when its frequency is struck on a nearby similar fork.

Today's conceptions of a Multiple Wave Oscillator (MWO) are quite different from what Lakhovsky was really doing. Lakhovsky tried to reproduce the healing energies of the cosmos by electrically energizing coils to produce energy fields which apparently regenerated aberrant cells along the lines of his theories. Lakhovsky first did this by producing an antenna, comprised of concentric copper tubes, the outer one being energized with two wires from his power source. Our impression of his antennas is that they were log periodic, that is logarithmic in nature (Lakhovsky doesn't come right out and say this, but his patents are definitely worded to cover such an arrangement). New data coming in (but not yet verified) tells us that these copper tubes were filled with one or more of the noble gasses (also called rare, or inert gasses).

As anyone could guess, these antennas would be somewhat hard to manufacture. Lakhovsky then miniaturized the antenna and inserted it in a gas bulb. His US patent #2,351,055 says vacuum bulb, but also covers low pressure gasses. There are two main keys here. One is the log periodic antenna. All higher forms of life are based on logarithmic patterns, from ferns to humans. Lakhovsky was creating a structure in the ether which would bring cells into proper functioning. The other is the vacuum / low pressure gas effect. For those who wish to understand the potent effect of the vacuum on life then the reading of The Nature Of Substance by Rudolf Hauschka is required. According to Walter Russell the noble gasses are the octave markers in the harmonic spiral of the elements, which also exists in a logarithmic pattern.

I recently visited with Dr. Norman McVea of Stinson Beach, California. Norman has several high frequency gas tubes built trying to understand what Lakhovsky and Royal R. Rife were up to. Norman has a MWO built of 12 glass concentric rings with various noble gasses inserted in them. When we checked it out it was running at 60 cycles, pulsed at 7.8 cycles. My recommendation to Norman was to run the unit with a radio-frequency oscillator which can be tuned from 1 to 30 megacycles, and modulated with a variety of sources for research. Norman has produced the most unique MWO I've seen in our modern times, and he is certainly on the right track. He is also working to reproduce Rife's high-frequency gas bulbs.

The key to Royal R. Rife's frequency instrument is also the use of the noble gasses energized at high frequency (possibly 4.5 megacycles), which were then modulated with specific musical tones. Lakhovsky's units were energized similarly, but used the frequency-independent log periodic antennas, giving a broad band of treatment, rather than the specifics of Rife's units. In many ways these two pioneers were working with the same keys.

WHO DUNNIT?

Rife and Lakhovsky have another similarity, that is, that people today are building equipment obliquely related to their work and offering them as original, or "improvements." This needs serious examination. The last two years has brought on an underground cult of so called "Rife Frequency Generators." There are hordes of these so-called "Rife" units being sold which are merely fancy variations of TENS (Transcutaneous Electrical Nerve Stimulation) devices. The conception for these units, usually producing square audio waves, comes from John Crane who knew and worked with Rife in his later years. John has affidavits from people who swear that they were healed from cancers and such by his square wave units in the early 1960's. It is extremely important to note that Rife's work in the 1930's through the 1950's was with high frequency gas tubes and all his documentation covers those units. History is rapidly becoming blurred in the race to cash in on people's poor health.

Modern purveyors of the so-called "Rife" equipment are making several errors: The first is that they claim that Rife developed the use of the square wave generators with electrodes. That is incorrect, it was John Crane (which makes Crane a pioneer of TENS research!). The second is that they claim that Crane's equipment was junk and would never work. Maybe the equipment wasn't of the highest quality (we have an original unit in our collection!) but Crane does have a lot of signed affidavits, several of them in our files, claiming that the units worked on specific diseases. In general, people are being misled in order to sell them equipment they probably don't need. That is a bad reflection on the field of electrotherapeutic research.

Perhaps these modern "Rife" (TENS) units have some merit in killing certain types of pathogens. I've spoken with many people who were completely dissatisfied with their so-called "Rife" units, and I've spoken with two people who have claimed that they have killed infections such as foot rot and herpes. Maybe they do, but for all of the thousands of devices being put out where is the documentation? Where is the proof? I have seen Crane's affidavits of the early 1960's, but have seen no medical studies from anyone, not even basic notes!!! There is obviously a very important bit of work to do in researching the effects of pulsed DC on the human body. TENS units kill pain, that has been clinically documented. The claims of the so-called "Rife" units are just that -- claims, *and unsubstantiated to boot.* Until any substantial research work is done in this area I would recommend that people do not follow the mob hysteria which shrewd hucksters have instigated over Royal R. Rife's work.

Lakhovsky's story is similar. All the units being built today are based on Bob Beck's early 1960's interpretation of Lakhovsky's first patent (#1,962,565). After four years of looking at these devices and reading Lakhovsky's patents and books I've come to the conclusion that Bob was misled by the patent. From an electrical engineer's viewpoint Bob's units make sense. From the standpoint of understanding subtle energies within and surrounding the body the units are merely RF noise generators which have questionable effects on the body, and most likely interfere with natural energetic processes (chi flow, etc.). The Beck MWO is a variation of one of Tesla's electrotherapeutic devices, the variation being that Tesla went to great pains to make the antennas spark-proof, and Beck's antennas spark away. Lakhovsky's units didn't spark either!

What makes this more interesting is an article which appeared some months ago in the **Health Freedom News**, titled Who Dunnit? by Mike Brown. Mike gives a weak case for his view that Lakhovsky was inept, that his earlier units didn't work, and who needed Tesla to fix them up and make them work. Perhaps Mike didn't notice that Lakhovsky's SECRET OF LIFE, which he sells to promote his units (in the ad following the article), is based on the units he claims in the article didn't work! Mike refers readers to Tesla's article High Frequency Oscillators For Electro-Therapeutic And Other Purposes (**Lectures, Patents, Articles** by Nikola Tesla, Beograd, 1956), but fails to notice that Tesla's article contradicts what he is doing with his units. I wonder how that article got past the editors?

CORRECT DIRECTIONS

Lakhovsky was a brilliant scientist, highly regarded on the European continent. His name comes up in many books, most recently I found him referenced in Lilly Kolisko's **AGRICULTURE OF TOMORROW**. If he and Tesla met, I'm sure it was for more than to fix poorly engineered units. There are three types of units that I feel should be built for research into this area: One is Tesla's oscillator circuit. The higher quality Beck MWO units, such as those built by Klark Kent, can be converted fairly easily. Use the unit's power supply as is, but for antennas use insulated copper plates (or spheres, as Tesla preferred for maximum effect) coated with heavy dielectric material such as beeswax. Eliminate the sparking! A necessary design change for future units will be to change the ¼ wave coil to ½ wave.

The second is a reproduction of Lakhovsky's Tube MWO. It will take miniaturizing the Golden Ratio antennas that Eric Dollard designed for the MWO and inserting them into low-pressure rare gas bulbs, which will then be driven at high frequencies. It is my estimation that this is the best way to pursue Lakhovsky's high frequency researches. The mathematics for the antenna, along with further reference materials on log periodic antennas, are in The Lakhovsky MWO Handbook.

The third is a reproduction of Rife's original type of equipment being a rare gas bulb driven at high frequencies, modulated with musical tones. This will take RF broadcast equipment and a rare gas tube. No electrodes!

It is time that researchers in this area begin to go in the correct direction. Bob Beck made a noble effort in the early 1960's with his conception of the MWO, and gave people something to work with for 25 years. That phase is definitely over and a new one beginning.

The "Rife generator" craze should be stopped now before it takes people another 25 years to find they've been misled. Serious research into the effects of pulsed DC should be undertaken by those with the wherewithal.

And keep it uppermost in your mind that electricity is a sub-sensible force. In the borderland sciences one should look to the supersensible, such as Dr. Wilhelm Reich's orgone energy (which, by the way, reacts to electricity in such a way that makes it detrimental to the human body), and the human vital energy such as is utilized in Biocircuit research (as begun by L.E. Eeman).

January-February 1990 JBR,
VIEW FROM THE BORDERLANDS
by Tom Brown

THE WONDROUS ELECTRICAL FORCES

Electrical research is a primary area of borderland research. While the term "occult" puts many people off, it should be realized that electricity is an occult force whose effects we only see when equipment is properly engineered. Four to five hundred years ago one would probably be burned at the stake if it would have been possible for them to demonstrate electrical devices.

Electricity is considered by some to be a degraded form of the Light Ether, that the etheric force gets bound up in matter in a way that it can be used. But being bound up in matter it provides mankind with the option of using it properly or improperly. Many articles on the detrimental effects of electric fields have been published, and more studies are no doubt underway on this important topic. There are also some records of electro-therapeutic devices that used this force for healing. Tesla envisioned a worldwide electrical power system and did the basic testing to show himself that it was viable, and needed but to be built. Tesla was a man of great intellect and was a reader of Goethe and other great minds. One would feel that he was in some way familiar with Goethe's *Farbenlehre* (color theory). The type of electricity being sent around our globe on wires is of a different, lower form that that of Tesla's transmission of electricity without wires. It is quite possible that Tesla's electricity was of a beneficial nature. Tesla did research into electrotherapeutics and was no doubt aware of both beneficial and detrimental effects. With the power system he envisioned on line we may not have ended up with the electrical problem we are just beginning to see the "tip of the iceberg" of.

This point needs clarification and elaboration beyond what most people now imagine. We periodically mention certain aspects of Dr. Lilly Kolisko's experiments with the crystallization of metallic salts rising *contra* gravity. These experiments were done to establish whether or not the planets had any direct connection to the metals as alchemical lore supposed.

The connections she investigated are Gold-Sun, Silver-Moon, Mercury-Mercury, Copper-Venus, Iron-Mars, Tin-Jupiter, Lead-Saturn. For example Dr. Kolisko would take a gold chloride solution (1%) and allow a portion of it in a laboratory dish to *climb* a segment of blotter paper. As it dried the fine details of crystalline form could be seen. After many experiments the characteristic form of each metal's salt could be comprehended. It was scientifically established that the actions of cosmic bodies would have direct effects upon the crystallizing actions in the lab. Gold chloride would not crystallize properly during several solar eclipses that Dr. Kolisko reported on (1927, 1947). A darkening effect was noticed on these slides.

Just recently I came across this in the recollections of Dr. Rudolf Hauschka, founder of Wala Pharmaceuticals: "In the course of researches into the delicate changes in matter which are revealed in the capillary-dynamic method, an alarming symptom came to the fore. At one of the Easter conferences of the Anthroposophical Doctors' Working Group in Comburg in 1964, Frau Kolisko, the past-master of the capillary-dynamic method, spoke on 'Gold and the Signs of the Times.' In earlier years she had demonstrated by means of the *Steigbild* (literally rising pictures, also known as chromatograms) the connections between the metals and the planets, and in particular the effect on gold of an eclipse of the sun. During an eclipse of the sun the gold *Steigbild* which normally glows with beautiful colors, shows a darkening and the emergence of grey formations. Frau Kolisko now showed a series of gold *Steigbild* produced in the last years almost hourly and at all times of the day and of the year. Without exception all of them showed the same sinister dark formations which had appeared earlier, in the first half of the century, only at the time of a sun eclipse. She closed her lecture with a question to the doctors, namely whether, considering the change now in gold, they could still be answerable for its use as a medicine. There was no doubt about it: since the middle of the century gold had become different – it was sick. Or, to express it differently, the connection between the spheres of the sun which stream down and the metal which they permeate is disturbed."

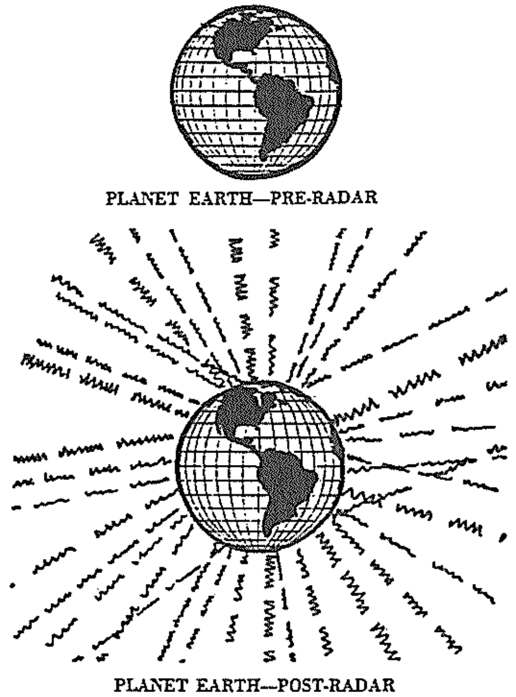

Entire planetary environment permeated with pulsed electromagnetic radiation—an unprecedented condition.

ELECTRICITY: THE MAIN CULPRIT

The main culprit in this sickness of earthly matter is none other than the electrical environment that has manifested on this planet. Certainly there were radio communications going during Dr. Kolisko's early experiments, but as this chart from Trevor Constable's *COSMIC PULSE OF LIFE* shows, the post WWII environment was heavy with these signals, culminating in today's total saturation point. Imagine what the *Steigbild* would have shown in previous centuries before the electrical force began its manifestation.

Nuclear proliferation must also be considered as adding to this phenomenon, but electricity is the degraded form of the light ether, whose source the sun is related to gold. Walter Russell, in ATOMIC SUICIDE, indicates that the spread of nuclear energy will destroy the atmosphere surrounding this planet and now scientists are scratching their heads as the hole in the planetary ozone layer gets larger. (Of course chlorofluorocarbons and New Zealand sheep farts have all been blamed for the ozone depletion, but nuclear energy is a primary cause.) This atmospheric destruction could well aid in cutting off gold from its cosmic wellspring of life.

Rudolf Steiner had many insights into the true nature of electricity, its relationship to humanity and its role in our future. Pertinent to the matter at hand let us consider the following remarks, a portion of his answer to a question about electrical treatment of animal fodder:

"You must consider the whole part played by electricity in Nature ... to the effect that human beings cannot go on developing in the same way in an atmosphere permeated on all sides by electric currents and radiations. It has an influence on the whole development of man. This is quite true; man's inner life will become different if these things are carried as far as is now intended. It makes a difference whether you simply supply a certain district with steam-engines or electrify the railway lines. Steam works more consciously, whereas electricity has an appallingly unconscious influence; people simply do not know where certain things are coming from. Without a doubt there is a trend of evolution in the following direction. Consider how electricity is now being used above the earth as radiant and as conducted electricity, to carry the news as quickly as possible from one place to another. This life of men in the midst of electricity, notably radiant electricity, will presently affect them in such a way that they will no longer be able to understand the news which they receive so rapidly. The effect is to damp down their intelligence. Such effects are already to be seen today. Even today you can notice how people understand the things that come to them with far greater difficulty than they did a few decades ago.

"It is a remarkable fact that whenever something new appears, as a rule in the early stages it is heralded as a remedy--a means of healing. Then the prophets get hold of it. It is strange, where a new thing appears, clairvoyant perception is often reduced to a very human level! Here is a man who makes all sorts of prophecies about the healing powers of electricity, where no such thing would previously have occurred to him. Things become fashionable! No one was able to imagine healing people by electricity so long as electricity was not there. Now--not because it is there, but because it has become the fashion--now it is suddenly proclaimed as a means of healing. Electricity--applied as radiant electricity--is often no more a means of healing than it would be to take tiny little needles and prick the patient all over with them. It is not the electricity--it is the shock that has the healing effect.

"Now you must not forget that electricity always works on the higher organization, the head-organization both of man and animal; and correspondingly, on the root-organization in the plant. It works very strongly there. If... you pour electricity through the foodstuffs--you create foodstuffs which will gradually cause the animal that feeds on them to grow sclerotic. It is a slow process; it will not be observed at once. Electricity will not at first be recognized as the cause; it will be ascribed to all manner of other things.

"Electricity, once and for all, is not intended to work into the realm of the living--it is not meant to help living things especially; it cannot do so. You must know that electricity is at a lower level than that of living things." (from **Agriculture** by Rudolf Steiner)

How pertinent and timely these words are today on many levels of borderland research. Electrical/electronic medicine has the glamorous position on stage at most of the conferences, and in many people's minds, in today's 'borderland' field. People are racing to buy instruments which are designed by electronic engineers with little or no training in the comprehensive constitution of humans. *It is the fashion!* There is no consideration that true healing comes from the spiritual realms, that the spirit is always healthy, and that dis-ease comes from a condition where the various subtle levels of the body are not aligned properly with the physical.

These electrical instruments have the potential to produce apparent results in treating disease symptoms. But what of the long term effects of the subsensible electric current? Considering the alarming state of affairs viewed through the information just presented I pose the question to the reader, "Do you prefer the natural or the unnatural forms of treatment?" Many people flock to electronic medicine because they feel that the petro-chemical drugs that are prescribed by the medical establishment are detrimental to their long term health. Very possibly so, but will electricity be any better? What about your own vital energy and its relation to cosmic life forces?

Electrical stimulation of the body treats a symptom and fails to allow for the body's natural healing process. As electricity is an externalized part of humanity's inner volitional force we can see that there will be definite effects from the subsensible current. But the effects will be relative and the inner healing process put on secondary mode. Electro-therapy, while emerging as a viable medicine of the future, still ignores the natural healing process.

This does not mean that we should not research electro-therapy. In what may seem a contradiction to many readers, I am 100% for further research into this area. It will no doubt be of benefit to many. Our relationship with electricity is one of the greatest questions of our time. People unconsciously use this force, and the results are such as those noted by Lilly Kolisko. Electricity is intimately connected to the Life Force, but great care must be taken not to confuse the two. Electro-therapeutic research helps define the boundaries, and thus is of interest to Borderland Researchers, but it certainly does not mean I recommend it!

Mar-Apr 1990 JBR, BSRF Bulletin Board

MWO QUESTIONS

Your comments re the MWO in the Nov-Dec 89 JBR are of much interest and raise the following questions in my mind:
1. Will you be publishing any information by Dr. Norman McVea?
2. How can the Kent units now in circulation be converted to produce a 1/2 wave? Perhaps Klark would be willing to publish an article on this or be willing to do the converting for a price.
3. Has any member ever tried to obtain any information on the circumstances of Lakhovsky's death? About three years ago a reference librarian made a serious effort for me to obtain any information re Lakhovsky, with absolutely no success. Perhaps there was an obit in the NY Times at the time he died, assuming that is where he was at the time.
4. Where in the world would we obtain copper spheres like Tesla used? Perhaps a member can suggest a source.
P.V.P.
California

In answer to your questions:
1. Norman McVea has sent us in a bit of material on his latest oxygen researches. He has opted for the oxygenation approach instead of the electrical.
2. I don't know about conversion of Kent's units to 1/2 wave, you should check with him on that. It will probably be easier to use it as it is -- they are good power supplies for experimentation -- and in the future engineer units with more refined outputs. After reading my column Klark added a rare gas tube assembly to his "Rife" units. He has always listened to other people and has constantly worked to improve his units.
3. If any readers have info on Lakhovsky's death, or his life for that matter, these pages are open to pursue the matter.
4. Anybody know where to get copper spheres? Tesla also used copper plates coated with a heavy dielectric.

September-October 1990 JBR BSRF Bulletin Board

MWO EXPERIMENTS

I have been very fascinated with your Journal. Thank you for the new world you have opened up to me. For the last few months I have begun to work on an MWO.

I have completed the antennas but did not do as well with the Tesla-coil. I decided to buy the ready made coil from

"Edmonds". I connected the secondary to the center of my antennas and did get some action in the form of a purple glow around the third and fourth smaller rings. I can only assume that I need more power to this unit although I was "Quite" pleased at the limited reaction on my first attempt. I would like to reach out to a fellow member who might be able to help me with some advice at this point.

Any advice you would send my way would be greatly appreciated.
P.D.
New York

Naturally, the quality of output of any electrical device is dependent on not only the quality of materials, but also the geometry of structure and arrangement. The better the coil, the better the antenna, the more refined the fields will be that are put out around the device. As I have pointed out in the article mentioned above in my remarks about the "Rife" devices, the types of "Lakhovsky" MWO's being built today vary quite a bit from what Lakhovsky was actually building and using himself.

Lakhovsky used very refined electrical designs and probably log-periodic antennas with both poles of a Tesla coil feeding the outside ring in such a manner that there was a logarithmic, electro-magnetic structure oscillating in space which apparently had quite beneficial effects on living organisms.

The modern variations, based on Bob Beck's schematic of 1963, feed one pole into an antenna (and the other into a second antenna), creating a dielectric field, which when it collapses, due to the rings on the antennas, creates unnecessary sparking and random radio-frequency noise with questionable effects. It is an organic system alterant, but is quite different electrically from Lakhovsksy's original equipment.

Nov-Dec 1990 JBR.

BSRF Bulletin Board

THAT MWO AGAIN

I recently purchased your MWO HANDBOOK. In the book are many references to the "Golden Ratio". I was unable to determine what the mathematical basis for the Golden Ratio is. I have other material which describe the Golden Section, and it is a ratio of .618 to 1. The closest I came to a value for the Golden Ratio is from the MWO antenna data on page 188 of your book the ratio appears to be 1.27:1. Could you please clarify the mathematics of the Golden Ratio antenna?
B.C.S., California

*** I recently purchased THE LAKHOVSKY MWO HANDBOOK. I am in the process of building an oscillator. I have all the parts except 6Kv .005mf capacitors. No electronic store in Alaska has them, or knows how to get them. I would appreciate it if you would send the address of either the manufacturer or an electronic store that carries 6Kv .005μf capacitors. I have one more question. Do you know of a meter or electronic apparatus that measures the amount of energy level the human body produces? It seems that like fingerprints, all human bodies are different in the amount of energy they use or produce. If one can coordinate the energy level of the oscillator to that of the body far less damage would be done and far greater healing potential be made.**
S.N.
Alaska

For the first question, the Golden Ratio antenna uses the golden ratio as a base, but it was engineered to conform to other considerations of antenna propagation. It is what is called a log periodic antenna. These antennas are based on the square root of phi, that is the inner ring is multiplied by the square root of phi to get the next ring, or inversely the outer ring is multiplied by one over the square root of phi to get the next inner, etc. This provides the Golden Ratio mathematics (every other ring is phi related).

MWO NEWS FROM FRANCE

I'd like to mention with regard to "Electricity the Main Culprit" that just after the war, Gurdjieff wrote a chapter in his book ALL AND EVERYTHING entitled "Beelzebub says that man's capture and subsequent destruction of electricity from nature is one of the main causes of the reduction in the human lifespan." And you are lucky in the US. In Europe we have a 230V system that is far more harmful than yours.

A final comment on the MWO which seems as popular as ever. When I lived in Paris I made the acquaintance of Lakhovsky's son and saw two of the MWO's built by his father in operation. The demonstration was very impressive but it didn't last too long because the old units interfered seriously with radio and television reception in the area. They were very big units, about as large as an early TV set. I don't know where they are now since he later loaned them to friends living in the country.
R.C.
France

Very curious quote from Gurdjieff as it concurs with what Steiner indicated, that electricity makes living things sclerotic. If perhaps you wouldn't mind searching your memory a bit more I have a question about Lakhovsky's device: You say it interfered with radio and television reception. Was this from internal spark gaps, or did any unit you see arc between antenna coils. The pictures in THE SECRET OF LIFE indicate that the then necessary spark gaps were interior to the unit, and the pictures of the effluvia coming off the antenna are obviously pictures of corona discharge and not arcing as in the units built in the US today. Bob Beck, in his 1963 article on MWO's claims he built a unit from an original, but his antenna hook-up is quite different from Lakhovsky's pictures. Many years of searching have failed to produce an original unit. The modern units are derived from the most obvious derivation of Lakhovsky's first MWO patent, which is written to both protect his design and cover infinite variations thereof. A detailed account of an original MWO in action would be of great value in this research.

SOUTH OF THE BORDERLAND

I have read Mark Clement's THE WAVES THAT HEAL and discovered a little piece of information not included in the current edition to THE LAKHOVSKY M.W.O. HANDBOOK. In WAVES I noted a short sentence indicating that one set of rings transmits energy through space which falls upon the patient's body and (it must be ascertained through experiment whether this is as I tell here or otherwise) later absorbed by the receiving resonator which resonates it for rebroadcasting into the patient's body. Thus the true Lakhovsky MWO, and UNLIKE the Beck-reconstructed model, would be a more-complex device involving features so-far unthought of. It must be ascertained what KIND of "receiving resonator" Lakhovsky employed, if it had or not special constituents not described at any of his patents included in the MWO HANDBOOK and what other features could be complementary to both.

Another feature to be considered, and which I discovered in the photo on page 30 of Clement's book is that there are two protruding spheres connected to the outermost ring of the Lakhovsky setup. In the copy of the BSRF publication I have, these two photographs are reduced and with not the same kind of resolution of the work by Clement. Each protruding sphere is connected to a cable and following the path of least resistance

the electrostatic current flows both along the metallic path of the outermost ring and the excesses accumulated from attracting ether from the environment are projected towards the patient (in the case of the transmitter) and collected for resonating (in the instance of the receiving resonator). Also notice that there are two "bracing wires" to both keep the rings in place and at the same time to fulfill the 90° division determined for Earth by the metric theorems which are the foundation of the Energy Grid! This adds another feature of complexity to the duplication of the MWO, which carries the Lakhovsky invention, again, farther from the Beck-built model.

Please mention my paper SOME FREE-ENERGY DEVICES for the MWO analysis I made is at best very incomplete, not including the feature of non-uniformity the true MWO does possess.
Jorge Resines
Buenos Aires, Argentina

Thanks for your further insights into the MWO. I feel that a continuing dialogue will eventually produce more evidence as to what type of equipment Lakhovsky used in his successful researches into cell regeneration and how to best reproduce it. The resonator antenna setup on Lakhovsky's original is more complex than the Beck MWO, as, like the primary antenna, it also has two feeder wires. The two spheres you mention, from which brush discharges can be seen, have always struck me as the nuts that held the heavy antenna onto its stand. Some mechanical contrivance was necessary and the mini-spheres fit the bill with the possibility that they may have been further configured to produce a specific capacitance, as our evidence indicates that Lakhovsky's antennas were very precisely built, probably log periodic. We have heard that Lakhovsky was a radiethesiest and would dowse the sites where he set his equipment up. He was certainly into the subtle energies, and the French school was very precise in its radiesthetic work, being the source material for much of Christopher Hills Supersensonics work. We have an unconfirmed report that the original tube antennas were filled with a rare gas such as argon. This would seem to be the case as in Lakhovsky's second MWO patent he reduced the antennas and puts them into vacuum or low pressure gas bulbs (as the first, this patent is written to be very specifically vague). This was meant to be an improvement over the original units, and probably was. What it tells us is that the "Beck" units appear to be far from what Lakhovsky was doing and where he was going, though because of more than 25 years of people building and using Beck models it is hard find information on the original. The "bracing wires" are supposedly silk. Their angle was probably determined solely by mechanical function.

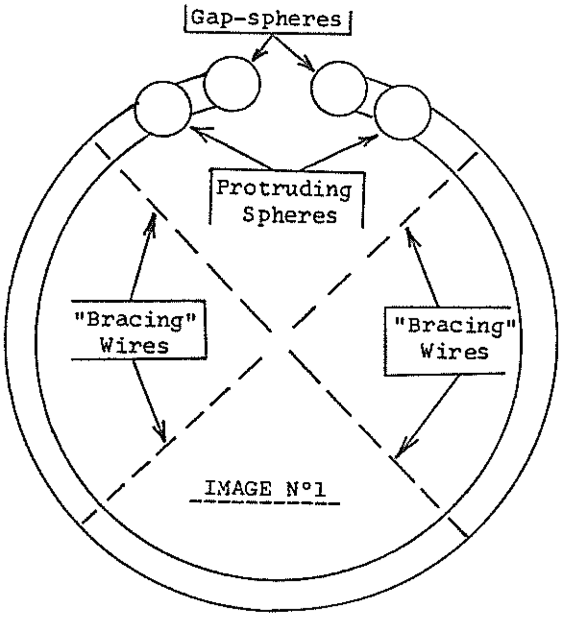

Detail of outer ring on Lakhovsky's original MWO antenna. Front view above, side view below

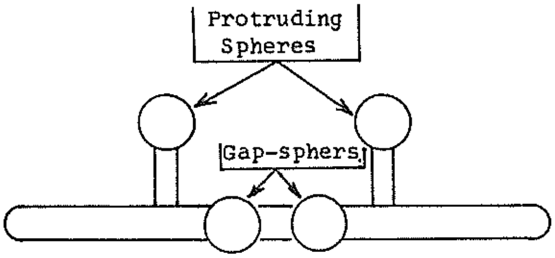

We have actually come under attack for our attempts to clarify the MWO story! In a brochure put out by MWO salesman Mike Brown, Box 25, Pleasant Hope, MO 65725, we come in for the following treatment: "Then there are the characters who, with no formal training whatsoever, denigrate tings [sic] they know less than nothing about. One outfit calling itself Borderland Research sent out a bulletin claiming that no-one today is manufacturing a Lakhovsky Multiple Wave Oscillator. These characters arrived at their conclusion by comparing photos of a 1935 unit to ones being made today. The units don't have the same external appearance. Quite often people with no formal training in electronics or electricity don't understand that an electronic device and how it operates are determined by the circuitry contained within the unit, not by the box it comes in." My comments are that there are perhaps *tings* we don't know but we do know that what is inside counts as much as the outside on the MWO. I have put five years of serious investigation into the MWO and would have to say that what we published is the result of serious consideration by a number of intelligent people--some highly trained electrical and mechanical engineers--interacting and exchanging ideas. We have published the evidence to support our contention, and have in fact stated that if we can be shown to be wrong then we would like to know, that we have no final conclusions, just some of the facts. A crude attempt to belittle BSRF's research by ignoring the main facts does not change the facts and the facts are that the units pictured in Lakhovsky's book **THE SECRET OF LIFE** are quite a bit different than the units being built today, indicated specifically by the antenna hookups, from which basic logic would indicate the need for a different interior setup. A thorough analysis of the basic differences between the two types of units can be found in The Lakhovsky Multiple Wave Oscillator Handbook, edited by myself, Tom Brown (no relation to Mike). Mike's brochure then goes on to claim that "the equipment necessary to regrow arms and legs can be purchased at Radio Shack for about $10", that is a "metal plate, copper wire, battery, and 100 K-ohm to 1.5 Megohm resistor." Heck, if you can regenerate an amputated limb with a only battery and a resistor then why do we need all the electronics in a MWO just to drive out mere cancerous tissue?!

As a final note on this, I would like to point out that electricity is a system alterant. Whether it alters the system for better or worse depends on many other factors, including mental status and diet. As Trevor Constable pointed out back in the 1960s in this Journal, the healing powers attributed to the MWO could be caused by the etheric force rushing to suppress the life-negative electro-magnetic activity. It is also possible that any benefits coming from modern day MWOs are psycho-tronic in nature -- that is the mind is doing the cure and the instrument provides a physical anchor for the psychic energies. The actual operation of the equipment, the sparking and gross electrical output, is superfluous to the matter, other than in convincing the mind that *someting* is happening.

Jan-Feb 1991 JBR,
BSRF Bulletin Board

LAKHOVSKY'S ORIGINAL

You ask whether the original Lakhovsky MWO had any arc between antenna coils. I don't remember any but to make sure I checked up with a friend who has worked on the MWO and knows it inside out. He was also present when I saw the device working. He tells me that there is definitely no arc between coils. The arc occurs inside the machine after the "Ruhmkorff coil". He also says that there are two tungsten plates which are drawn apart and brought together manually.

The other thing I remember is a sort of metal wand with a knob at one end which is used between the two units to treat people. At the time I saw it, one of the persons present had a bad back. The two MWO antennas were turned on and were facing each other about 6 ft apart. The "wand" was applied to the sore back. I seem to remember that it kind of gave off sparks, but I'm not sure of this. I touched the wand and I could clearly feel something coming through. My knowledge of electricity is nil, however, if you have further questions, I could always try my friend who should be able to answer them for you.
R.C., France

Many thanks, as this confirms the evidence of Lakhovsky's own pictures used in his books. Would you be so kind as to ask your friend if it is possible to get a schematic (even a rough drawing from memory) of the basic layout of the MWO, which we could reproduce on these pages, also any antenna construction details. Many questions remain such as what components are in the metallic tubes holding the antennas up, and such (Eric Dollard finger-dowsed the pictures of the tubes and felt it was an arrangement of capacitors and spark gaps -- his basic MWO diagram, an intelligent guess of what Lakhovsky was doing, is in this Handbook). A basic schematic would probably be the most important piece of information to date. Please read on, pot has been stirred.....

LAKHOVSKY ANTENNAS

I got my Nov-Dec 89 Journal where it shows an update on Lakhovsky's antennas. On page 181 of SECRET OF LIFE it also shows the same thing. Take a magnifying glass and you can see it. I disagree with all the experts, it seems that they are chasing a phantom ghost. Lakhovsky gave us all his secrets in his book, all we have to do is read carefully.

He could cure cancer and other ailments with a simple piece of copper wire. Lakhovsky wasn't dumb, can you see him telling doctors to cure with a piece of wire they would tell him to get lost, so he built a very complex machine to impress them, also he would have something to sell. I think Beck's machine is close except the antenna. Who came up with this aluminum foil glued to plastic? That's where they fell on their face. If you read page 75 of the same book he repeats VIBRATES 3 times. That's why he hung his rings with silk thread, so it wouldn't damp the VIBRATIONS.

Can you imagine a pipe organ with copper strips glued to a board? I have designed an antenna using copper tube of varying sizes held by fish leader that I think would work fine. I'd bet a penny to 10 bucks that George used thin wall bronze tube as it will resonate better -- but it's expensive to fabricate.
Louis A. Schad
California

Yup, I agree that Georges was no dummy. His book shows plant cancers that have been cured via use of a single copper wire single-coiled around the plant. He felt that the pre-matter cosmic rays (or what I would call the material pole of the etheric instreams) were "ringing" the copper coil, in effect creating a tone that would heal the plant. I got the impression, from studying Lakhovsky's progress towards the vacuum (or low pressure rare gas) patent, that in his equipment Georges tried to invert the process -- exciting the ring with a post-material high frequency current that would create a healing tone. By using log periodic antennas (as would be allowed under the MWO patents) a vibrant chorus would ring in the ether, the tones, via the structure of the antenna, would be such that they could harmonize the functioning of the human body.

The aluminum foil antennas were developed by Riley Crabb, as a low cost way for experimenters to make their own antennas without having to have printed circuit boards made, and many folks were quite happy with the results. The physical evidence now clearly indicates that the Beck unit is quite different from the original Lakhovsky unit for which the cell regeneration results were recorded. What I have termed the "BECK MWO" must stand on its own merits and it can lay no claim to Lakhovsky's recorded results. As stated several times in the past Beck's device is more like one of Tesla's electro-therapeutic oscillators, except Tesla used copper sphere antennas, coated with a heavy dielectric to suppress sparking. Beck's unit would be covered under Lakhovsky's original (and very generalized) patent, but research of Lakhovsky's own material shows it is not what he used for electro-therapeutic researches. To excite the antennas in the Beck manner will not create the same vibrating field that Lakhovsky's antennas excitation circuit will. But it does create an electrical field that some claim is beneficial to a live body placed in its proximity.

It makes much sense that the antenna should also vibrate physically as well as electrically. Ah... if we simply had an original unit to weigh and measure...

Lou has sent us in a MWO of his own design with a single spiral antenna made of two twisted bare copper wires. We will be running a several part article on construction and use starting next issue. This unit is reminiscent of the Sun-Ether disc of Oscar Korschelt, circa 1891. Having sat in front of a number of different operating MWO's over the last six years I have to say that Lou's unit is the most exciting thing I've ever seen like this. I could strongly feel the effect 10-12 inches away, and there was no electrical excitation of the antenna, rather it is physically vibrated. There is a low voltage shocker that can drive the antenna through two smaller copper coils. It's easy to build and use, simpler than the Beck units, and should provide borderlanders with years of workable ideas. Lou will not be selling these devices but will provide complete building plans in the JBR.

We received a call from Patrick Flanagan. He reports that he has a MWO unit built by Lakhovsky's son in Paris. It is a little box with a one tube white noise generator and it has a multi-stranded cable which wraps around the body. It shows that the evolution of Lakhovsky's device has taken a different turn among those who actually knew what he was doing. This unit is based on the Radio-Cellular Oscillator.

Also, Klark Kent (of Super Science, Box 392, Dayton OH 45409) just called to say that he has completed a run of Golden Ratio antennas from the original B.S.R.F. artwork. These are double sided (with inverse gap strips, to simulate the ball capacitance on the original tube antennas, on back) gold plated and are available from Klark for $100 each. (These antennas have produced beneficial effects on plants without any electrical excitation. Many experiments can be done and the results quantified with various protocols such as those used by Rudolf Hauschka and Lilly Kolisko.)

There is no doubt that there is a new phase of MWO research being undertaken. For 25 years people built Beck's model, with many people claiming beneficial results, and others experiencing nothing but no one keeping track of results for evaluation. Thanks to Bob Beck's schematic (published by B.S.R.F. in 1963) the idea was kept alive and now people will have many ideas to choose from in their search for measurable effects from energies ranging from the gross electrical to the fine etheric. Thanks to everyone who has interacted on this ongoing aspect of our researches!

EVEN MORE LAKHOVSKY STUFF

Some time ago I bought your excellent handbook of the Lakhovsky Multi Wave Oscillator and due to pressure of other things have only just found time to read it. Interesting because it parallels some research done by myself and friends, or should I say we duplicated some of the original. This was only to be expected as the entire principle is so outrageously simple, too simple for establishment minds to comprehend.

I should like to offer a few ideas which might be of interest and in return I would be grateful for some additional information.

I helped develop the Betatron now sold by All-Source Systems, PO Box 596, Morton, WA 98356. The Betatron uses a TV Flyback transformer and is very similar to the Master Violet Ray devices. The nicest feature is that the applicator is an ordinary clear light bulb. Even a burnt out one will do. There are differences; some makes of bulb will not work at all, perhaps due to different gas fillings. The Betatron can also be used with the Lakhovsky concentric antenna with spectacular results. This of course has been done in the book's articles although as far as I know it was not actually copied, but was an obvious choice of experiment.

I have tried out an improvement which I did not see in the Handbook or anywhere else, perhaps you will be interested. The Standard circuit whilst apparently covering a very wide range of frequencies, must, because of its inherent fixed inductance and capacitance have its own natural fundamental frequency. Our Betatron is around 21 Kilocycles (I refuse to use the meaningless term Kilo Hertz), and other units seem to be basically tuned to one fundamental even though they do stray and produce an almost infinite range of harmonics.

My first improvement is to substitute a "Jacob's Ladder" for the spark gap. This consists of two 1/4" copper tubes spaced 5/32" apart at the bottom and tapering outwards to 1" at the top then abruptly turning out at 1" radius. The spark travels up the ladder from bottom to top and is extinguished by the sudden increase of gap, then begins again at the bottom. This sequence takes place at about 2 second intervals. The whole point of this is, as the spark progresses up the ladder, the length of the tuned circuit is progressively increased, then suddenly decreased again ditto repeato. As the ladder is about 9" long the length of the tuned primary of the Tesla coil is progressively made 18" longer, then back to its original length. Thereby resonating at infinite numbers of frequencies within its range and delivering virtually the same power at all those frequencies.

A further improvement on this feature is to make the ladder a Spiral Staircase by using a solid rigid central electrode, say a ½" copper tube and the spiral be wound in an ever increasing diameter on the outside, the start being about 5/32" gap and vertical to begin with.

Alternatively the spiral can be the inside element wound on a porcelain former such as is used in radiant "bowl" heaters with a plain copper tube surrounding it or perhaps a perforated tube as used in decorative lighting fixtures. The spark winds its way around the spiral staircase and alters the frequency accordingly. The Spiral Staircase provides a much longer path and more variation in frequency range. I have not yet perfected this and ruined a few porcelain formers in the process. It requires some exact progressive increases in the travelling spark gap but there are no serious obstacles.

Both the Jacobs Ladder and the Spiral Staircase have the advantage of easily cooling the spark gap. We had experienced problems due to overheating of the electrodes in a fixed gap, even spark plugs transmitted the heat to other parts and carbonized insulation. The addition of a cooling fan (the low power kind used in computers) provides a current of air which accelerated the progress of the spark up the Jacobs Ladder and more importantly up the Spiral Staircase. The spark is sometimes reluctant to leave its easy path at the base of the ladder and even more reluctant to wend its winding way up the coil, but the air flow does the trick and keeps everything cool.

The fan also helps remove Ozone from the enclosure. Ozone has the delightful habit of attacking insulation. It has also been blamed for respiratory illnesses. I am skeptical of the latter as I have breathed plenty of Ozone and found it exhilarating.

Master Violet Ray even included an Ozone Inhaler with its de-luxe model!!!, recommended for asthma, bronchitis etc. What's more M.D.s used it on their patients. Certainly Ozone is about the only thing which will quickly digest Carbon Monoxide in a hurry and it deodorizes a room when the cat has misbehaved.

The Betatron and Lakhovsky machine both can be used to destroy troublesome warts and moles, by simply wrapping a wire around the electrode and applying the point of the wire to the offending growth. I had some troublesome warts removed in this way by a Dermatologist. It was not until I got involved with this project that I realized the Dermatologist was using a Lakhovsky device.

He anesthetized the area first but I have since got rid of warts and moles without this. It caused some "Discomfort" which as any mother knows is worse than mere pain, but it was tolerable.

I hope the foregoing will be useful and will not be in the least surprised it has not already been done, but I would like to know what you think of it.

Regarding Spark Coils. I recently found at the local Flea Market, a small spark coil about 3½" overall length 2" dia. with a sturdy vibrator assembly and a turret mounted sideways. It looked like a small steam boiler in fact. It is labelled "Coil, Booster, Battery Operated. Type C1, (24 Volts DC) Spec. #94-32182, Serial #96082, order #A,C 27073, Eclipse Aviation Bendix, N.J. Division of Bendix Corp. I don't know what it is for but on 12 volts it delivers a lively 3/8" spark. The absence of a Zip Code suggests it is fairly old, but this is the ideal driver for a compact tesla coil. I do not know if it contains condensers but I added some anyway. The adjusting screw has a click stop adjustment. Have you come across

one of these? It works well on 6 volts which is just as well as it draws a lot of juice and 8 D cells won't fire it.

Now a request for information about Lakhovsky Treatment and parts. Please Please. On my knees!! Where can I obtain High Voltage Capacitors as used in the Lakhovsky Machine?????? I have tried all the surplus stores I know of in this area but the highest voltage they have is 450. I recently paid over $50 for a 660 volt one for a 12 v trailer power supply and that was ordered specially from New York. HELP!!! I even tried making one from instructions in the Boy Electrician, a Sunday School prize from the 1930s. I used Acrylic 1/8" dielectrics instead of glass. They worked well for a time but soon tracked and failed. I don't fancy using glass plates any way but I desperately need some H V Capacitors at less than Pentagon prices.

Re Treatments. One of the contributors to the Handbook reports that treatment with the Lakhovsky Machine resulted in the dis-appearance of underarm hair. (The report of decreasing underarm hair is from the Nov. 1968 Journal of Borderland Research by Mrs. E.H., Piqua, Ohio.) This is very interesting as I am a heavy construction worker, and after a hard day at the office I like to drop in for a cool beer and fraternize with nice hot ladies. However, underarm odor which results from my daily exertion and defies Right Guard AFTER for 2 hours, is a Repellant. I did find that shaving the offending fuzz helped a lot. Tearing it out by the roots was even better; it seemed to de-activate the sweat glands.

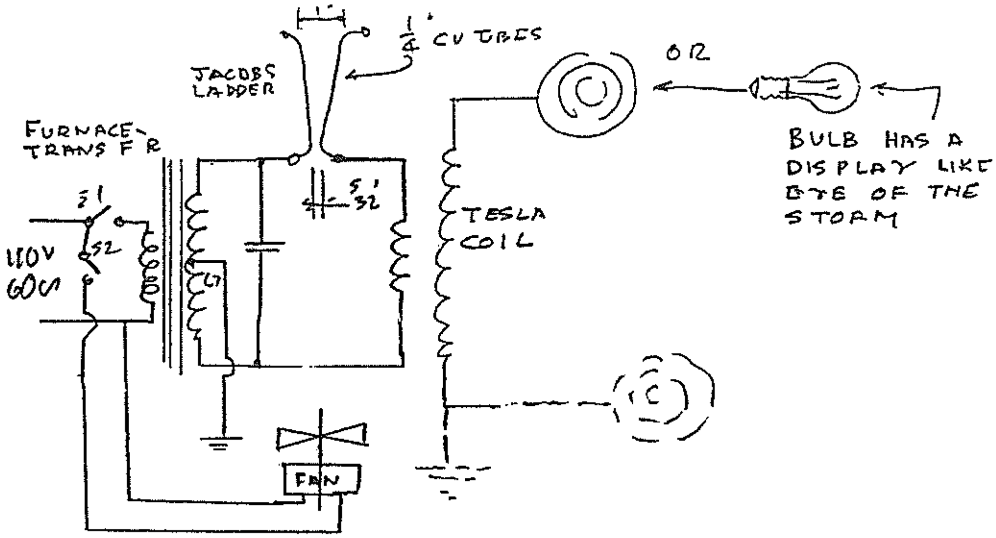

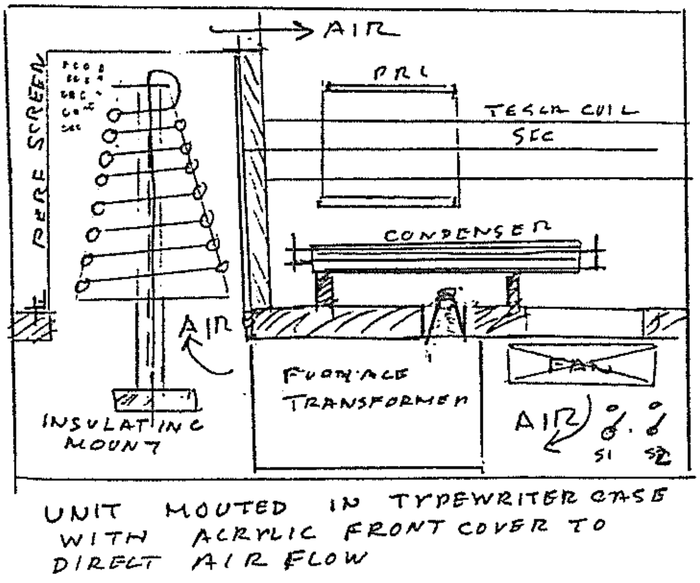

UNIT MOUTED IN TYPEWRITER CASE WITH ACRYLIC FRONT COVER TO DIRECT AIR FLOW

After I sprained my shoulder I used the Betatron all around the joint and have noticed that underarm hair now grows much more slowly, and I have much less trouble with B.O. I used the Betatron frequently and the Ozone removes any residual stink...

NOW the Big question. Has anyone succeeded in using the Lakhovsky machine to stimulate hair growth??? I have hair like Yul Brynner. Maybe some folks think its sexy but I find it repulsive and worst of all it makes me look and feel 20 years older at least without a hat. It seems to me that the machine could well be used to rejuvenate hair cells since it can reverse most other degenerative processes. Its a strange thing that men who are short on top have an overabundance of superfluous body hair in places where it is utterly wasted.

Has anyone done any research on this subject with the Lakhovsky Machine??? If so please let me know. I will be most grateful for any information and will of course inform you of the results of my own experiments.

It occurs to me that such a phenomenon could well dispense with the daily chore of shaving and make Gillette obsolete. I am forever disgusted that such a lot of life force is wasted growing hair on my face instead of where it belongs, On Top! Would be most grateful for your ideas.

I remember the Master Violet ray being demonstrated in a department store in the 1930s, I must have been about 7 years old. My mother who was with me watched the demonstration and I, of course, was fascinated. I said, "Oh Mom do buy one, it's just what you need for your rheumatism". Mother replied, "Oh no, You just want it to play with". How right she was. Well, in spite of all the quack devices which come and go, Lakhovsky in all his forms is still with us. That's recommendation enough. The AMA persecution of course confirms its effectiveness.

A Naturopathic Doctor saw me wearing a copper bracelet. He was treating me for arthritis, effectively with electro therapy etc. and asked if I really believed in the copper. I said "No", then he said "Why wear it?" "Because it works." He laughed and said, "Look at that green stain on your wrist. That's copper salts and they are being ingested; maybe that's how it works, you might have a copper deficiency, well it's cheaper than pills."

Very Sincerely,
K.V.A.R., Washington

First off, without further ado, let me congratulate you on taking a stand on refusing to use the word Hertz to designate cycles per second. Why use mumbo-jumbo designators when simple English will do. Tesla said that the acceptance of the Hertzian theory would come to be seen by future generations as the greatest aberration of the scientific mind. Of course Tesla is probably rolling in his grave (at 3600 rpm?) at the modern etherless explanations of electricity. He, and others of his day, regarded the (transverse) Hertzian waves as the "waste" put off by coils. The Tesla (longitudinal) wireless system would have been worldwide (including radio and power), with the transverse waves engineered to a minimum, whereas the radio we have today, with the transverse waves engineered to a peak, fades out as one goes around a hill or a decent distance from the emitting source. This is dealt with much more specifically in THE THEORY OF WIRELESS POWER by Eric Dollard. Regarding use of the Jacobs Ladder as a spark gap, that is an interesting idea and thanks for sending it in. As far as your hair growth problems go, perhaps our other readers have further comments. You have provided a lot of food for thought, thanks!

HIGH VOLTAGE SOURCES:

Re: Nov/Dec Journal and inquiry of S.N. Capacitor 6KV (6000V) .0047 uF (which is the standard value, not .005!) available ($4.90) from: Antique Electronic Supply, 6221 S. Maple Ave., Tempe, Arizona 85283. (602) 820-5411.

A.A.
Massachusetts

Source for high voltage equipment: Plastic Capacitors, Inc., 2623 N. Pulaski Rd., Chicago IL 60639-2190. Write for catalog (Dealer for above is Newark Electronics). Also recommended is Homemade Lightning: Creative Experiments in Electricity by R.A. Ford, ISBN 0-8306-3576-9, TAB Books, Blue Ridge Summit, PA 17294-0850. Can be ordered through your bookstore.

Simplified Tech Service
PO Box 2140A, Champaign IL 61825

In the Nov-Dec 90 issue S.N. of Alaska asks about a 6Kv .005 uf capacitor. Radio Shack sells a 2Kv .01 capacitor #272-160. Three connected in series will be 6Kv .003 uf, another 3 connected in parallel across the first three will be 6Kv .006 uf. Six of these costs about $3.

Regarding the MWO, when Hieronymus lived near me in Ft. Lauderdale I asked him about the MWO. He didn't recommend it because the broad spectrum of frequencies had both beneficial and harmful frequencies which could not be separated.

R.B.
Florida

Thanks for all the input on where to get HV capacitors. Also, thanks Ron, for Hieronymus' viewpoint. With the Beck style MWOs I would agree. A mass of radio frequency noise is generated. With Lakhovsky's original units driving structured antennas one could produce multiple wavelengths of a more refined nature. We don't recommend the MWO to anyone, that is not our purpose. This Journal acts as a forum to ideas such as these to be explored, and the MWO refuses to go away, rather interest is growing and it is time we closely examine the field. Read on...

THE BECK MWO

My MWO is a Beck model. It has successfully killed a skin cancer on the back of my hand. Yet I am wondering about its resonator. There seems to be no action, it's of no use and is as if dead. I did not realize that we are having trouble establishing a real Georges Lakhovsky MWO. This [enclosed] information from Europe may help clear up our need.

D.M.
Bryte,

A good puzzle truly has many pieces, so thanks for your experiences. I never said that the Beck MWOs didn't work, but that they are not the same as Lakhovsky built. The enclosed ad was for radionic and related equipment including MWOs from Radionic Instrumentation & Research, Dorset, England. It may be that this is the radionic type of MWO originally sold by Copen Labs. I honestly don't know a thing about these MWOs, so any info would be greatly appreciated especially a schematic. Years ago I had heard them spoken of disparagingly by people who claimed that they wouldn't work because there wasn't enough power, an argument I never bought. It would seem that the more subtle and refined the energy the more beneficial it is to the human organism. The field is still open for new ideas and fresh impulses...

KLARK KENT CHECKS IN!

Just a note to say thanks for the mention in the Journal about the availability of the Golden Ratio Antennas from Super Science. And also to say that the articles on the continuing developments of the MWO are most welcome. Especially interesting is the comparison of the Beck MWO to the Lakhovsky original!

I have never seen an original Lakhovsky MWO. Only from the patent information can we really get schematic information, circuit diagrams, antenna structures, etc. As you know from all of the free energy patents circulating around, a patent wrapper and drawing is not really enough to produce the claimed device. If it were, then we should have many free energy devices in production today!

Several months ago a contact in New York made a generous offer to procure an original Lakhovsky MWO. I passed this information along to you and to others. So far no one has come forth to my knowledge with information leading to the whereabouts of an original Lakhovsky MWO. What this means is that we must grope our own way in replicating the technology first developed by Lakhovsky.

So I guess I owe Bob Beck a word of thanks because he was the one who made it possible for researchers to actually build a MWO. The information left to the world by Lakhovsky is (in my opinion) wholly inadequate to build and construct the device. The patent information is, like most patents, just not adequate to enable someone to replicate the device. And this brings up the question: Just what constitutes the original Lakhovsky MWO? We don't have an original device (and a bounty of $100,000 has so far failed to produce one). The information given in the patents doesn't really tell us enough.

Borderland Sciences can help out by continuing to publish articles about the MWO which emphasize circuit diagrams and construction details. In this way, Borderland researchers can redevelop this important technology.

I enclose a letter we received from a customer. Since the letter is of the testimonial variety, it is best not to reference it in the Journal. But the point is that even a Beck adaptation MWO can seem to produce positive experimental results!

Klark Kent, Super Science,
P.O. Box 392, Dayton, OH 45409

Glad to hear you are open to new research. Lakhovsky left much information -- it just never got translated into English. BSRF will continually publish the results of ongoing MWO research, and will always be on the lookout for the best information available. Here is the letter submitted to Klark, from a person known to us here also:

Dear Mr. Kent,
At long last I get around writing to you and share some of my experiences with the MWO I bought from you.

The unit arrived on May 17, 1990 in excellent condition. Your packaging is superb. I inspected the unit superficially but, because of some professional problems, was not able to make use of it until October.

Let me comment first on general appearance and workmanship. While obviously not industrially produced, this is a well-built high quality machine, using fine components. Having been an electronics engineer part of my life (before I became a philosopher), I feel well qualified to make that judgement.

The instructions are generally adequate but I was glad that I did order your audio cassette, done by you with a fine dramatic flair, worthy of Christopher Reeve, who does resemble you. My suggestion: include the cassette with the MWO as a matter of course.

I used the machine about thirty times, sometimes daily, sometimes skipping a day or two for 35 min. per session, mostly in the High Power Radiator mode.

I suffered over a period of about fifteen years from a mild chronic bronchitis, some coughing in the morning with some phlegm, rest of the day just fine. The medicine men told me that I "just have to live with it". Though I am 70, I judge myself being in great health.

Now, your machine has definitely improved my condition. I know, of course, that my testimonial is worthless to you, since you cannot make, and do not make, any health claims. But still I thought you might want to know.

I do have several radionics machines (De la Warr and Copen) but we all realize that this is sympathetic magic; great stuff, though!
With best wishes, faithfully yours,
NAME WITHHELD

ELECTRONIC ENTITY REMOVAL

Judging by the letters you receive and print in the BULLETIN BOARD people are still experimenting and building MWOs. It should be accepted that no two MWOs will be identical due to different components, circuits, etc. Also, Lakhovsky did not have some mystical secret that is exclusive to his particular instruments; he built one that generated his required multi-frequencies and there are no doubt many ways to effect this.

The large MWO I built in 1966 worked as well as I could expect and, although I never use it now, its circuit can be regarded as a basic Tesla coil, starting out with a 12KV neon transformer. Other circuits I have experimented with unfortunately appeared not to give the same results. Your MWO Handbook should be the starting point for would-be experimenters keeping in mind that most of the 'transistorized' models do not have enough energy to effect the desired results.

A major point that never seems to be mentioned is that the subject should be seated ON AN INSULATED PLATFORM with the antenna about six inches from the back. In this way all the cells of the body are raised to their full potential for, say, ten minutes during which each cell can absorb energy at its one resonant frequency. The subject should ground himself for an instant before stepping off the platform. Then switch off the instrument. I found this procedure mandatory to obtaining positive results and have found it odd that this point has never been mentioned in the literature.

I might also bring up the very real possibility (discussed with Riley Crabb in the past) that at least part of the positive results obtained with the MWO could be attributed to the removal of spirit entities by the high voltage involved. Regrettably, Borderland appears to skirt this issue.

As I have stated before, the Journal is an important platform for the meeting of the minds and I wish you well in your efforts.
R.M.
Nevada

Appreciate your input and valuable ideas. I'm certainly open to the possibility that the results obtained from the MWO could be attributed to the removal of spirit entities. This should also bring up the question of whether or not what is perceived as spirit entities (negative in this sense) could be also related to their potential physical anchors, i.e. virus and bacteria. Maybe we see the same thing but interpret differently.

As far as using different configurations to achieve multiple wavelengths, different arrangements will produce qualitatively different results. A French horn and a clarinet will show the same frequency for the same note played, but the difference in quality of sound should be obvious to all. The major difference between the hookup of Lakhovsky's and Beck's units makes them fall into quite separate categories. A silver bell rings finer than tin.

MWO REPORT FROM FRANCE

I finally managed to get hold of my friend (see BSRF Bulletin Board, Jan-Feb 1991) but he is very, very busy and only gave me some of the information I wanted.

Firstly, he declined to provide any basic layout of the MWO because he say no two units were the same. Apparently Lakhovsky experimented all the time and each machine was a "one-off" model, different from the others.

The origins of the MWO are in Tesla's work. A French professor, Mr. D'Arsonval took some of Tesla's equipment and modified it. Lakhovsky used D'Arsonval's work to build his machine. Tesla heard about D'Arsonval's experiments and, convinced that D'Arsonval was stealing his patents, came to France with the idea of starting legal proceedings against the professor. However, he found D'Arsonval so courteous and charming that he returned to America having dropped the idea of court action.

My friend tells me that the whole secret of the MWO is in the "diffuseur". He further indicated two sources of information, both in French. A series of books have been written and published by the Study Group "Arkall" called "L'Emergence de l'Enel" dealing with all aspects of the MWO. The second source is George Lakhovsky's book "L'oscillation cellulaire" which has been re-edited in Belgium. Unfortunately, my friend did not know the editor, so I suggest you will have to contact the American Embassy in Belgium and they should be able to give you addresses of bookshops where it can be obtained. Apparently, this book gives all the references concerning the people I have mentioned.

I'm sorry I can't be more helpful but the whole question seems to be much more complicated that I at first thought. In short, I'm afraid you will have to wade through "L'oscillation cellulaire" and the "L'Emergence de Enel" to see how Lakhovsky himself worked out the idea of the first MWO's.
R.C.
France

Thank you, your input has been most valuable and stimulating. The "diffuseur" (literally: loudspeaker or nozzle) would probably be the antenna, so the "secret" could very well be (as we have intimated over the years) in his mathematical arrangement -- a much more important consideration under Lakhovsky's antenna feed arrangement than with Beck's.

Jul-Aug 1991 JBR,
BSRF Bulletin Board

MWO CORNER

Received the Journal yesterday and was very excited about the Spiral Coil MWO. I congratulate Mr. Schad on doing a very good job and would submit the following input for future models.

1) No aluminium tubing should be used. Copper or steel tubing should be used instead. Aluminium is wholly negative. It attracts negative radiation like honey attracts files. In this connection, I heard this morning on the British radio that a "mobile home" site had just been devastated in the USA. If I'm right, "mobile home" site means trailer park, in which there is a huge quantity of Al. These are prime targets for tornadoes and all kinds of negative influence. My chief reference for this is Dr. Starbard Webb, Box 200, Orangevale, CA 95662. She has been studying the subject since 1956 and has accumulated a mass of data. Be patient for she travels a lot in the States doing seminars etc.

2) The rest is from a Tibetan lama to whom I showed the drawing. If you're wondering what a lama would know about the MWO I would say that they seem to know about everything that matters. For instance, in my own field (medicinal plants), their knowledge is not great, it is simply colossal, far beyond anything known in the West. The Lama tells me that there should be 12 coils, no more no less (Fig. 6). The "up" wire (see illustration) should be five times longer than the diameter of the coil. No doubt further information will be forthcoming when the whole of the MWO is described.

Yours sincerely,
R.C.
France.

Glad to hear of your suggestions for further research on Lou's spiral coil MWO. His articles have stimulated a lot of work that should provide a lot of new ideas over time. Peter Lindemann has taken Lou's concept of physical vibration and is experimenting with phi ratioed antennas that are stimulated with sub-audible sound. Also, Lou called to make sure that it was clear that the antenna side facing the subject is a counterclockwise spiral. It is clockwise from the top.

Interesting thoughts on the mobile homes, which are reverse orgone accumulators. A few items congruent with your comments have showed up over the years and perhaps we can get more on that in an upcoming issue.

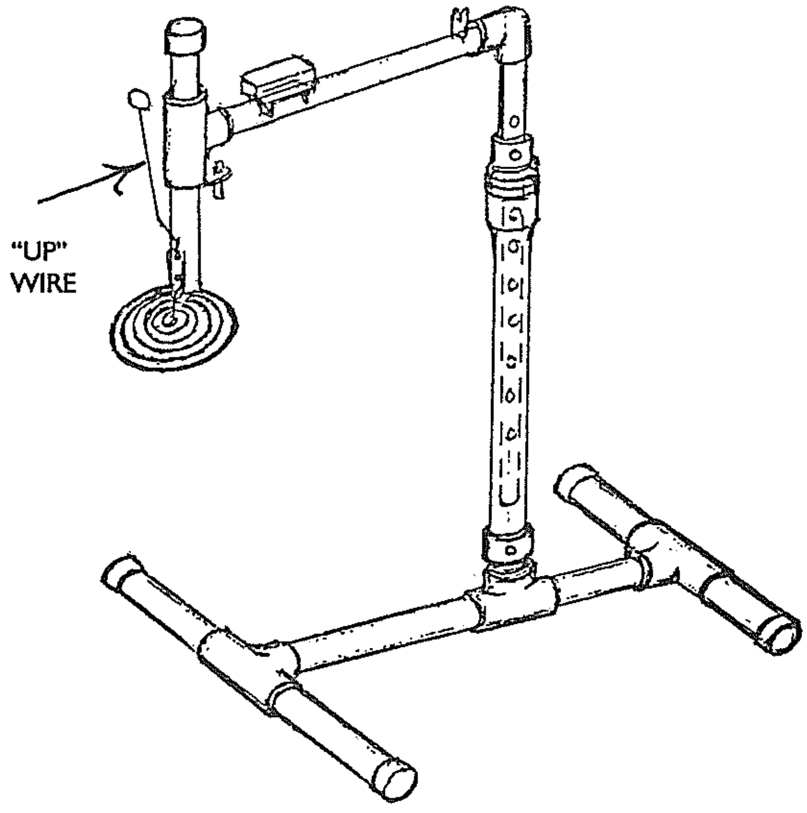

Spiral Coil Multi-Wave Oscillator

LAKHOVSKY COIL INFORMATION

The following was forwarded to us from Health Research, Box 70, Mokelumne Hill CA 95245:

On Health Research's flyer for THE HANDBOOK OF UNUSUAL ENERGIES, the artist shows the copper wire positioned incorrectly through the wood support.

I have made a number of discoveries which I even find hard to believe by using this COILED COPPER WIRE. I expect to write a booklet on my findings and you may want to represent it in your sales catalog. I have proven over and over again what I have done but, to date, have not shared anything with anyone else. I had to be ABSOLUTELY CERTAIN before I 'went public'.

When I first received the books THE SECRET OF LIFE and THE WAVES THAT HEAL from you about 4 years ago, I noted carefully just how the copper wire was COILED around the plant. Please note that the copper wire STARTS from INSIDE THE EBONITE and MOVES COUNTER CLOCKWISE and then ends ABOVE the ebonite... This is shown on page 14 of THE WAVES THAT HEAL and page 110 of THE SECRET OF LIFE.

In the past four years I have made dozens of experiments and I've gone far beyond what even Lakhovsky did or thought of. NOW, to prove my point about the proper way the copper MUST be coiled, I deliberately experimented by coiling the wire the way you have shown it and NOTHING HAPPENED in all my experiments. That's why I KNOW THAT

I AM RIGHT and it is imperative that you change your picture to show the correct way. I am not a scientist of a professional person, but I KNOW what I have done and you will have to take my word for it unless you want to YOURSELF, conduct the same experiments I have and PROVE it to yourself... I have not the slightest idea WHY the copper wire MUST be coiled this way but I DO KNOW THAT IT MUST.

You know as well as I that EVERYTHING in the COSMOS is perfect, exact, precise and absolute that the LAWS, FORCES, ENERGIES, POWERS, FREQUENCIES, RAYS... and all the rest function flawlessly and cannot allow any variations even for an instant. Life as we know it will cease forever if even an INSTANT change is made by whatever cause.

I would dearly love to see one of Lakhovsky's BELTS, COLLARS, BRACELETS he calls them OSCILLATING CIRCUITS, of course, and although the picture on page 63 of WAVES THAT HEAL is interesting, I am wondering if there is more to it that what this artist concept shows. I have yet to experiment with a HUMAN BODY as I will need to be very careful to use the RIGHT PERSON who will not think that I have gone off the deep end and headed for the 'funny farm'. But Lakhovsky certainly has plenty of evidence to show. If you know of anyone who has one of the ORIGINAL Lakhovsky oscillating circuits, let me know. It is imperative that the 'human' circuits MUST be done, and applied, correctly.

A.P.
Nebraska

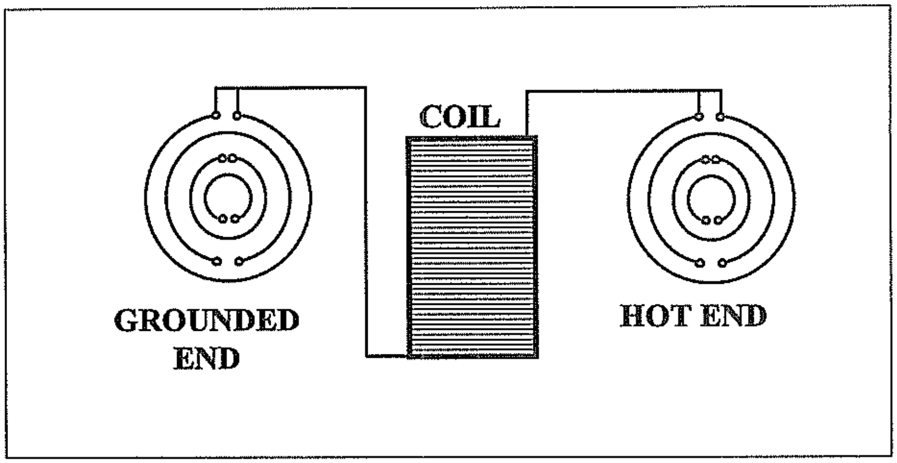

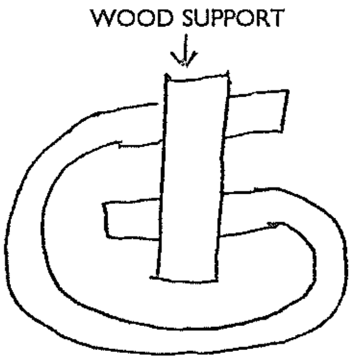

WRONG

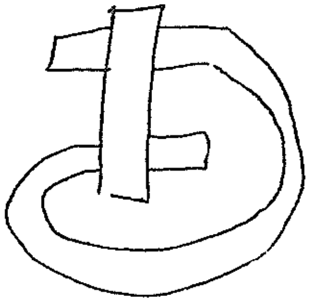

RIGHT

Nov-Dec 1991 JBR,
BSRF Bulletin Board

MWO RESEARCH

Five years ago I duplicated the MWO as best I could with my limited knowledge. I scaled up the antennas from the book WAVES THAT HEAL using copper tubing of all different sizes and soldering brass hex caps of various sizes on the ends after I rounded them off with a grinder. The outside ring was about 2 ft. across.

I was able to duplicate the exact brush discharge pattern in the picture with Tesla coil 4" diameter with about 1,000 turns of regular hook up wire. 500 turns gave a discharge of 2"-3" evenly all around the outer ring. 1,000 turns gave a discharge of 5"-6" at the bottom with less and less towards the top except for some off the two balls at the top, like in the picture.

I found it necessary to run two connections from the feed wire otherwise with one connection the spark would travel around the ring and out the other end burning anything used to suspend the ring. Two connections eliminated this problem forcing all energy towards the bottom of the ring like in the picture.

The resonator was grounded just like the Beck model to the ground end although I used two wires one going to each end of the outer ring. Even so, at this power you could draw a small spark off the outer ring and if you turned the unit off and on you could hear the resonator ring just like it was struck by a small hammer.

I used a 30mA 15,000 volt sign transformer or stronger. These tend to burn out unless you put the primaries of two in series to share the load and restrict current. Actually, I started using half burned out transformers using mid point ground and the good half for 7,500 volts which is much safer for the capacitors. I used two .005 uf 20,000 volt capacitors in parallel for .010. The stationary spark gap was air cooled with a small blower fan.

There was no spark between rings. The side effects from this unit were too strong to continue treatments so I am now working on smaller units. After 15 minutes your head would be full of mucus. The closer the antennas were to each other the stronger the effect to the point of making one very nauseous.

Also, did anyone else purchase and test Dr. Charles T. Turley's square wave sound MWO two printed antennas powered by a sound generator?

J.A.
Ohio

Thanks for your most helpful input. Would be most interested in Dr. Turley's MWO as it may very well relate to Lou Schad's spiral coil MWO, presented several issues back, which worked on physically vibrating the coil. Peter Lindemann has picked up on this and is experimenting with vibrating the antennas with sound waves and through other mechanical means. A whole new field of MWO research is opening up and it sounds like Dr. Turley is working in the same direction. Please send us any information you may have on this. Thanks.

From the March-April 1992 JBR
BSRF Bulletin Board

MWO CORNER!

Here is a nice present for you and your readers. Some tidbits about George Lakhovsky. I phoned my knowledgeable friend on Sat, 14th Dec. To my surprise not only was he at home but he invited me to spend the afternoon with him the following day.

Firstly, it is comforting to know that there are still about four original MWO's within a radius of 200 miles from Paris, and the people who have them are likely to take good care of them.

Secondly, from what I can gather, L. took very great pains with each MWO. I understand now what my friend meant when he said that each one was different. L. would calculate all sorts of parameters relating to the place where the MWO was to be used. He would then do the "fine tuning" with the help, sometimes, of an unusual cleric Abbe Breuil.

L. arrived in France around 1896. He was a friend of Tesla and when the latter came to France the first person he wanted to see was L. I don't know anything about this early period but two episodes occurred later which gave L. quite a bit of money and enabled him to finance his future research himself. 1) During the 1914 war, there was a dearth of methylated spirits in France and L. negotiated supplies, on behalf of the French government, with the Russians. 2) He invented a gadget which was used everywhere on the French railways and which brought in tidy sums in royalties. He was thus able to invite titled people to his table, the Count of this and the Duke of that and, of course, Abbe Breuil. This curious cleric was very well known in France, not for saintly endeavors on behalf of the Church but for his exploration of caves in search of traces of early man. He became the High Priest of Pre-history. However, what interests us here is that he was an extraordinary medium. For instance, L. at table, would say something like this: "Abbe, I've just thought up a new design for the MWO (passing him the drawings) what do you think of it?" And the Abbe might reply: "The antenna's too thick here, you should take 1 millimetre off and here, you are using the wrong kind of metal." In this way, L. gradually refined each model, using the Abbe's intuition to verify his own findings, until he got it as near perfect as possible.

What follows is a translation of the notes scribbled down by my friend yesterday. I offer no comments because I wouldn't know a Hertz dipole if I fell over one in the dark.

These are some of the parameters taken into consideration by L.

1) Diameter: D.
2) Thickness of the wire of the tube (not to be confused with the section).
3) Distance "e" between the two ends.
4) Orientation of the circuit with respect to the horizontal plane (29^{0} 70 N).
5) Kind of circuit. Should it be made of metal or hollow glass containing chemical salts.
6) Orientation of the distance "e" with respect to the magnetic North.
7) Study of the terrain: the mineralogical, geological, geodesic and tectonic factors.
8) Study of the nature of the surface of the circuit and whether and how it should be covered.

I see now that what I had taken for solid rods making up the antenna were in reality hollow tubes.

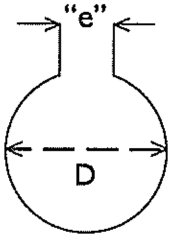

Hertz dipole

MWO antenna

I hope that you and your readers will find this information interesting. It sheds a new light on how L. operated. He used the pendulum a lot and, as we have seen, the mediumistic talents of his friend Abbe Breuil. My best wishes to yourself and to BSRF in 1992.

R.C.
France

Your input on the MWO project has been most helpful to everyone trying to find the truth of the story, and I offer my thanks and I'm sure the thanks of BSRF members the world around. Georges Lakhovsky was a complex individual whose ideas on cell regeneration have ignited many a seeking mind. It seems as though he had some rather interesting friends also, Nikola Tesla, Professor D'Arsonval, and the curious Abbe Breuil! Your friend's comments indicate some important factors in Lakhovsky's considerations: He was very concerned about the orientation towards the natural energetic lines of the earth, even considering latitude (what you have termed the horizontal plane) as a determining factor in the energetic output of his devices. He must have been aware of the interaction of the electric output with the planetary magnetic, electrostatic and etheric fields. He was also concerned about the underlying earth itself. In his **Secret of Life** he proposes his theory that cancers are more predominant in areas where the soils (such as clays) reflect the cosmic radiation back to the surface, causing interference patterns that disrupt the body's energetic flows. It is interesting to note that his antennas were constructed with all these factors involved. It is also very interesting that he relied on the pendulum and the mediumship of Abbe Breuil for his refinements. Lakhovsky was a true borderland scientist! Also of note is your point #5 mentioning the hollow glass circuit containing chemical salts. Something new to figure in our researches. We also have the open question of whether or not some of his tube antennas contained rare gases, which his MWO Tube Patent implies by its refinements in miniaturization. This has been an exciting period of discovering what Lakhovsky was really up to in his vast researches. Please read on....

A REAL LAKHOVSKY SCHEMATIC!

This schematic is from a Lakhovsky Radio Cellular Oscillator manufactured by Lakhovsky's son in Paris long after the untimely death his father in the early 1940s. It shows that the Lakhovsky's own family followed the path of the radio-cellular oscillator and not that suggested by Lakhovsky's turn to damped electrostatic power supplies in his MWO. Our many thanks go to Patrick Flanagan for providing this information to BSRF. Patrick obtained the unit, removed the seal at risk of voiding the warranty, and provided us with the circuit description you see here. Also, a special thanks to Roger Modjeski of RAM Tube Works of Santa Barbara, California for providing the pinouts and tube conversion info. As far as our extensive research notes show, this is the first real, complete schematic of an working Lakhovsky device that has been published in the US!

This unit, which fits in a small box that can be held in the hand, is a very powerful white noise generator with a treatment coil comprising of coax cable which is looped around the subject when in use. The treatment coil loop is 1 meter in length, the center wire and shield are connected. This unit will interfere with electronic communication devices up to the satellite TV range, and may be illegal to operate under FCC rules. The tube is an EL83, a French tube (Telefunken manual gives the German number as PL83) whose US equivalent is given as a 6CK6 video output pentode, but GE never made them. It is very similar in function to a 6CL6 which has a different

Lakhovsky MWO/Radio Cellular Oscillator submitted by Patrick Flanagan

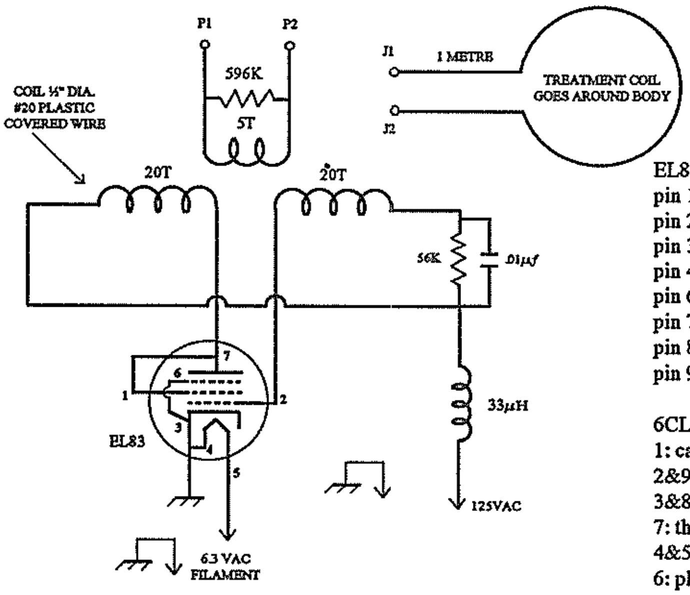

EL83/6CK6 PINOUT
pin 1: grid 2
pin 2: grid 1
pin 3: cathode
pin 4&5: filament 6.3V
pin 6: grid 3
pin 7: plate/anode
pin 8: internal shield
pin 9: not used

6CL6 PINOUT
1: cathode
2&9: first grid
3&8: second grid
7: third grid
4&5: filament
6: plate/anode

pinout. We provide the pinouts for both.

The 5 and 20 turn coils are ½ inch diameter tightly wound with #20 plastic coated wire. They are air core and all three are wound in the same right hand direction. They are spaced tightly and arranged exactly as shown. Varying the physical distance between the primary and secondary coils would vary the output strength. The resistors are ½ watt. The .01 capacitor is ceramic at 500 volts. The choke is 33 µH. The power supply is not shown, but is implied by the circuit.

We have no exact details of the interaction of this generation of Lakhovsky device with the human system, only that suggested by Lakhovsky's original RCO research and its extension through time. This data is provided for research purposes and to record yet another development in radio-biological research.

We would still be most interested in finding the schematic of one of the original MWOs pictured in Lakhovsky's various publications such as that pictured below from his book **La Terre Et Nous (The Earth and Ourselves)**, Fasquelle Editeurs, Paris, 1933, as it would provide the exact antenna feed for the concentric ring antennas and other pertinent details. Note that the antennas on this unit stand above the arm, unlike the well publicized pictures from **The Waves that Heal** and **Secret of Life.** This indicates that the protruding spheres seen in the detail on those pictures have a mechanical function, perhaps more important than their electrical function, though with Lakhovsky's methodology, he would certainly incorporate the optimum characteristics for the main antenna feeds.

We have discovered the basic construction of the Radio-Cellular Oscillator and know that Lakhovsky moved from them to the MWOs with damped electrostatic power supplies and concentric ring antennas because of the potential for thermal shortwave damage from continuous-wave radio frequencies. But then we see that Lakhovsky's son was still manufacturing RCOs after the death of his father! So we can see that the final conclusion has yet to be drawn as to exactly how the original MWOs were built, and of course we would also like to see some actual MWO Tube units that Lakhovsky patented. We also see the RCO as a potential low-cost method of researching Lakhovsky's basic theory of radio-frequency cell regeneration. This MWO Handbook contains the essential keys to all our questions and more. Much research has to be done and Borderland Sciences is willing to help any serious researcher who wishes to pursue these streams of activity.

The Journal of Borderland Research is the publication of record in the English language for these and many other research subjects. Please join in and exchange information with us. Without the help of the many borderland researchers sending us their ideas for over a quarter of a century this book would not exist. Thanks to all and best wishes on your researches -- may they be serious, straightforward, and well documented for others to learn from.

Tom Brown

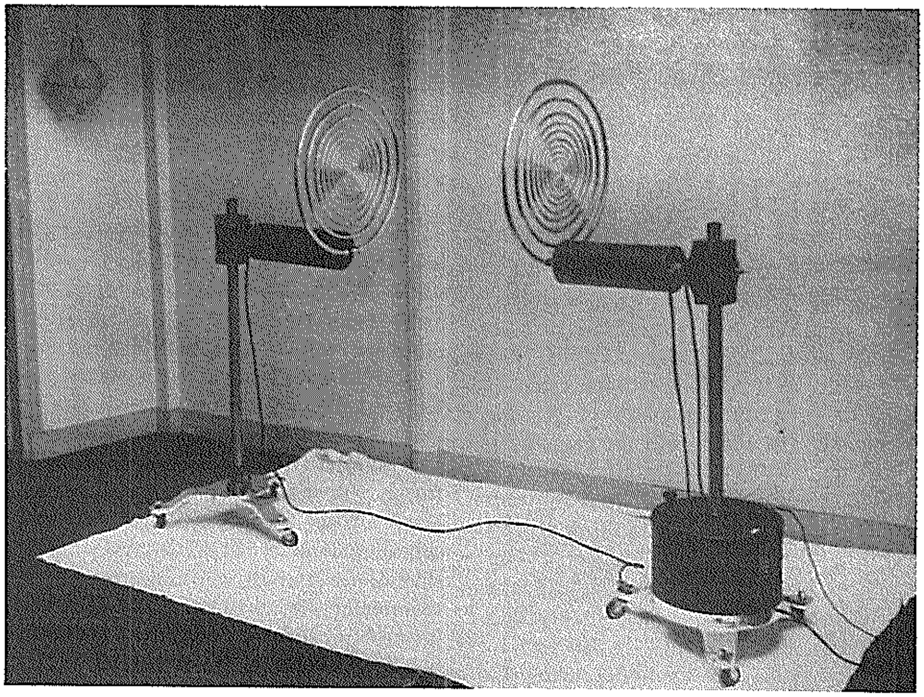

The complete Lakhovsky Multiple Wave Oscillator apparatus which creates an ambient electromagnetic field between its two *resonateurs.*

5

HIGH FREQUENCY OSCILLATORS FOR ELECTRO-THERAPEUTIC AND OTHER PURPOSES*

by Nikola Tesla

Some theoretical possibilities offered by currents of very high frequency and observations which I casually made while pursuing experiments with alternating currents, as well as the stimulating influence of the work of Hertz and of views boldly put forth by Oliver Lodge, determined me some time during 1889 to enter a systematic investigation of high frequency phenomena, and the results soon reached were such as to justify further efforts towards providing the laboratory with efficient means for carrying on the research in this particular field, which has proved itself so fruitful since. As a consequence alternators of special design were constructed and various arrangements for converting ordinary into high frequency currents perfected, both of which were duly described and are now — I assume — familiar.

One of the early observed and remarkable features of the high frequency currents, and one which was chiefly of interest to the physician, was their apparent harmlessness which made it possible to pass relatively great amounts of electrical energy through the body of a person without causing pain or serious discomfort. This peculiarity which, together with other mostly unlooked-for properties of these currents I had the honor to bring to the attention of scientific men first in an article in a technical journal in February, 1891, and in subsequent contributions to scientific societies, made it at once evident, that these currents would lend themselves particularly to electro-therapeutic uses.

With regard to the electrical actions in general, and by analogy it was reasonable to infer that the physiological effects, however complex, might be resolved in three classes. First the statical, that is, such as are chiefly dependent on the magnitude of electrical potential; second, the dynamical, that is, those principally dependent on the quality of electrical movement or current's strength through the body, and third, effects of a distinct nature due to electrical waves or oscillations, that is, impulses in which the electrical energy is alternately passing in more or less rapid succession through the static and dynamic forms.

Most generally in practice these different actions are coexistent, but by a suitable selection of apparatus and observance of conditions the experimenter may make one or other of these effects predominate. Thus he may pass through the body, or any part of the same, currents of comparatively large volume under a small electrical pressure, or he may subject the body to a high electrical pressure while the current is negligibly small, or he may put the patient under the influence of electrical waves transmitted, if desired, at considerable distance through space.

While it remained for the physician to investigate the specific actions on the organism and indicate proper methods of treatment, the various ways of applying these currents to the body of a patient suggested themselves readily to the electrician.

---

* Read at the eighth annual meeting of The American Electro-Therapeutic Association, Buffalo, N. Y., Sept. 13 to 15, 1898

As one cannot be too clear in describing a subject, a diagrammatic illustration of the several modes of connecting the circuits which I will enumerate, though obvious for the majority, is deemed of advantage.

The first and simplest method of applying the currents was to connect the body of the patient to two points of the generator, be it a dynamo or induction coil. Fig. 1 is intended to illustrate this case. The alternator G may be one giving from five to ten thousand complete vibrations per second, this number being still within the limit of practicability. The electromotive force — as measured by a hot wire instrument — may be from fifty to one hundred volts. To enable strong currents to be passed through the tissues, the terminals T T, which serve to establish contact with the patient's body should, of course, be of large area, and covered with cloth saturated with a solution of electrolyte harmless to the skin, or else the contacts are made by immersion. The regulation of the currents is best effected by means of an insulating trough A provided with two metal terminals T' T' of considerable surface, one of which, at least, should be movable. The trough is filled with water, and an electrolytic solution is added to the same, until a degree of conductivity is obtained suitable for the experiments.

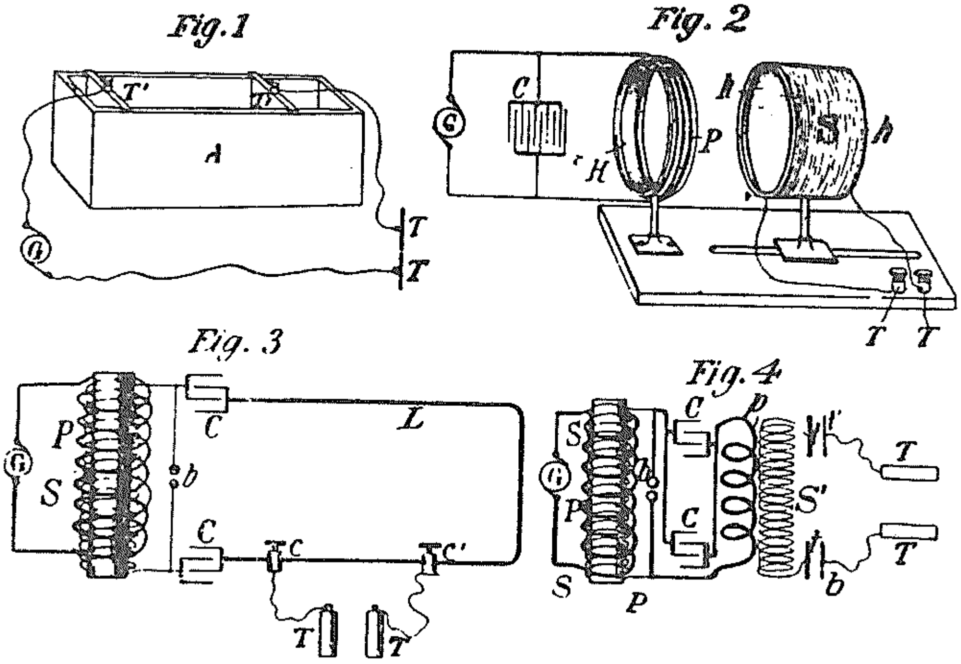

Figs. 1, 2, 3, 4.

When it is desired to use small currents of high tension, a secondary coil is resorted to, as illustrated in Fig. 2. I have found it from the outset convenient to make a departure from the ordinary ways of winding the coils with a considerable number of small turns. For many reasons the physician will find it better to provide a large hoop H of not less than, say, three feet in diameter and preferably more, and to wind upon it a few turns of stout cable P. The secondary coil S is easily prepared by taking two wooden hoops h h and joining them with stiff cardboard. One single layer of ordinary magnet wire, and not too thin at that, will be generally sufficient, the number of turns necessary for the particular use for which the coil is intended being easily ascertained by a few trials. Two plates of large surface, forming an adjustable condenser, may be used for the purpose of synchronizing the secondary with the primary circuit, but this is generally not necessary. In this manner a cheap coil is obtained, and one which cannot be easily injured. Additional advantages, however, will be found in the perfect regulation which is effected merely by altering the distance between the primary and secondary, for which adjustment provision should be made, and, furthermore, in the occurrence of harmonics which are more pronounced in such large coils of thick wire, situated at some distance from the primary.

The preceding arrangements may also be used with alternating or interrupted currents of low frequency, but certain peculiar properties of high frequency currents make it possible to apply the latter in ways entirely impracticable with the former.

One of the prominent characteristics of high frequency or, to be more general, of rapidly varying currents, is that they pass with difficulty through stout conductors of high self-induction. So great is the obstruction which self-induction offers to their passage that it was found practicable, as shown in the early experiments to which reference has been made, to maintain differences of potential of many thousands of volts between two points — not more than a few inches apart — of a thick copper bar of inappreciable resistance. This observation naturally suggested the disposition illustrated in Fig. 3. The source of high frequency impulses is in this instance a familiar type of transformer which may be supplied from a generator G of ordinary direct or alternating currents. The transformer comprises a primary P, a secondary S, two condensers C C which are joined in series, a loop or coil of very thick wire L and a circuit interrupting device or break b. The currents are derived from the loop L by two contacts c c', one or both of which are capable of displacement along the wire L. By varying the distance between these contacts, any difference of potential, from a few volts to many thousands, is readily obtained on the terminals or handles T T. This mode of using the currents is entirely safe and particularly convenient, but it requires a very uniform working of the break b employed for charging and discharging the condenser.

Another equally remarkable feature of high frequency impulses was found in the facility with which they are transmitted through condensers, moderate electromotive forces and very small capacities being required to enable currents of considerable volume to pass. This observation made it practicable to resort to a plan such as indicated in Fig. 4. Here the connections are similar to those shown in the preceding case, except that the condensers C C are joined in parallel. This lowers the frequency of the currents, but has the advantage of allowing the working with a much smaller difference of potential on the terminals of the secondary S. Since the latter is the chief item of expense of such apparatus and since its price rapidly increases with the number of turns required, the experimenter will find it generally cheaper to make a sacrifice in the frequency, which, however, will be high enough for most purposes. However, he only needs to reduce proportionately the number of turns or the length of primary p to obtain the same frequency as before, but the economy of transformation will be somewhat reduced in so doing and the break b will require more attention. The secondary S' of the high frequency coil has two metal plates t t of considerable surface connected to its terminals, and the current for use is derived from two similar plates t' t' in proximity to the former. Both the tension and volume of the currents taken from terminals T T may be easily regulated and in a continuous manner by simply varying the distance between the two pairs of plates t t and t' t' respectively.

A facility is also afforded in this disposition for raising or lowering the potential of one of the terminals T, irrespective of the changes produced on the other terminal, this making it possible to cause a stronger action on one or other part of the patient's body.

The physician may find it for some or other reason convenient to modify the arrangements in Figs. 2, 3 and 4 by connecting one terminal of the high frequency source to the ground. The effects will be in most respects the same, but certain peculiarities will be noted in each case. When a ground connection is made it may be of some consequence which of the terminals of the secondary is connected to the ground, as in high frequency discharges the impulses of one direction are generally preponderating.

Among the various noteworthy features of these currents there is one which lends itself especially to many valuable uses. It is the facility which they afford for conveying large amounts of electrical energy to a body entirely insulated in space. The practicability of this method od energy transmission, which is already receiving useful applications and promises to become of great importance in the near future, has helped to dispel the old notion assuming the necessity of a return circuit for the conveyance of electrical energy in any considerable amount. With novel appliances we are enabled to pass through a wire, entirely insulated on one end, currents strong enough to fuse it, or to convey through the wire any amount of energy to an insulated body. This mode of applying high frequency currents in medical treatment appears to me to offer the greatest possibilities at the hands of the physician. The effects produced in this manner possess features entirely distinct from those observed when the currents are applied in any of the before mentioned or similar ways.

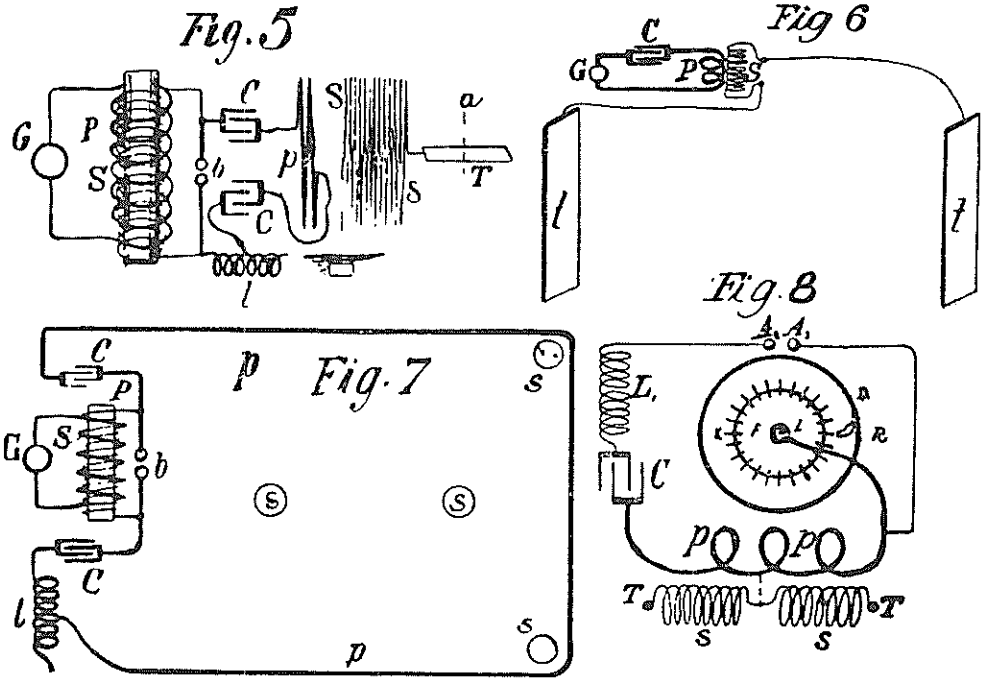

Figs. 5, 6, 7, 8.

The circuit connections as usually made are illustrated schematically in Fig. 5, which, with reference to the diagrams before shown, is self-explanatory. The condensers C C, connected in series, are preferably charged by a step-up transformer, but a high frequency alternator, static machine, or a direct current generator, if it be of sufficiently high tension to enable the use of small condensers, may be used with more or less success. The primary p, through which the high frequency discharges of the condensers are passed, consists of very few turns of cable of as low resistance as possible, and the secondary s, preferably at some distance from the primary to facilitate free oscillation, has one of its ends — that is the one which is nearer to the primary — connected to the ground, while the other end leads to on insulated terminal T, with which the body of the patient is connected. It is of importance in this case to establish synchronism between the oscillations in the primary and secondary circuits p and s respectively. This will be as a rule best effected by varying the self-induction of the circuit including the primary loop or coil p, for which purpose an adjustable self-induction e is provided; but in cases when the electromotive force of the generator is exceptionally high, as when a static machine is used and a condenser consisting of merely two plates offers sufficient capacity, it will be simpler to attain the same object by varying the distance of the plates.

The primary and secondary oscillations being in close synchronism, the points of highest potential will be on a part of terminal T, and the consumption of energy will occur chiefly there. The attachment of the patient's body to the terminal will in most cases very materially affect the period of oscillation in the secondary, making it longer, and a readjustment of the primary circuit will have to be made in each case to suit the capacity of the body connected with terminal T. Synchronism should always be preserved, and the intensity of the action varied by moving the secondary coil to or from the primary, as may be desired. I know of no method which would make it possible to subject the human body to such excessive electrical pressures as are practicable with this, or of one which would enable the conveying to and giving off from the body without serious injury amounts of electrical energy approximating even in a remote degree those which are entirely practicable when this manner of applying the energy is resorted to. This is evidently due to the fact that action is chiefly superficial, the largest possible section being offered to the transfer of the current, or, to say more correctly, of the energy. With a very rapidly and smoothly working break I would not think it impossible to convey to the body of a person and to give off into the space energy at the rate of several horse power with impunity, while a small part of this amount applied in other ways could not fail to produce injury.

When a person is subjected to the action of such a coil, the proper adjustments being carefully observed, luminous streams are seen in the dark issuing from all parts of the body. These streams are short and of delicate texture when the number of breaks is very great and the action of the device b (Fig. 5) free of any irregularities, but when the number of breaks is small or the action of the device imperfect, long and noisy streams appear which cause some discomfort. The physiological effects produced with apparatus of this kind may be graduated from a hardly perceptible action when the secondary is at a great distance from the primary, to a most violent one when both coils are placed at a small distance. In the latter case only a few seconds are sufficient to cause a feeling of warmth all over the body, and soon after the person perspires freely. I have repeatedly, in demonstrations to friends, exposed myself longer to the action of the oscillations, and each time, after the lapse of an hour or so, an immense fatigue, of which it is difficult to give an idea, would take hold of me. It was greater than I experienced on some occasions after the most straining and prolonged bodily exertion. I could scarcely make a step and could keep the eyes open only with the greatest difficulty. I slept soundly afterward, and the after-effect was certainly beneficial, but the medicine was manifestly too strong to be used frequently.

One should be cautious in performing such experiments for more than one reason. At or near the surface of the skin, where the most intense action takes place, various chemical products are formed, the chief being ozone and nitrogen compounds. The former is itself very destructive, this feature being illustrated by the fact that the rubber insulation of a wire is destroyed so quickly as to make the use of such insulation entirely impracticable. The compounds of nitrogen, when moisture is present, consist largely of nitric acid which might, by excessive application, prove hurtul to the skin. So far, I have not noted injuries which could be traced directly to this cause, though on several occasions burns were produced in all respects similar to those which were later observed and attributed to the Röntgen rays. This view is seemingly being abandoned, having not been substantiated by experimental facts, and so also is the notion that these rays are transverse vibrations. But while investigation is being turned in what appears to be the right direction, scientific men are still at sea. This state of things impedes the progress of the physicist in these new regions and makes the already hard task of the physician still more difficult and uncertain.

One or two observations made while pursuing experiments with the apparatus described might be found as deserving mention here. As before stated, when the oscillations in the primary and secondary circuits are in synchronism, the points of highest potential are on some portion of the terminal T. The synchronism being perfect and the length of the secondary coil just equal to one-quarter of the wave length, these points will be exactly on the free end of terminal T, that is, the one situated farthest from the end of the wire attached to the terminal. If this be so and if now the period of the oscillations in the primary be shortened, the points of highest potential will recede towards the secondary coil, since the wave-length is reduced and since the attachment of one end of the secondary coil to the ground determines the position of the nodal points, that is, the points of least potential. Thus, by varying the period of vibration of the primary circuit in any manner, the points of highest potential may be shifted accordingly along the terminal T, which has been shown, designedly, long to illustrate this feature. The same phenomenon is, of course, produced if the body of a patient constitutes the terminal, and an assistant may by the motion of a handle cause the points of highest potential to shift along the body with any speed he may desire. When the action of the coil is vigorous, the region of highest potential is easily and unpleasantly located by the discomfort or pain experienced, and it is most curious to feel how the pain wanders up and down, or eventually across the body, from hand to hand, if the connection to the coil is accordingly made — in obedience to the movement of the handle controlling the oscillations. Though I have not observed any specific action in experiments of this kind, I have always felt that this effect might be capable of valuable use in electro-therapy.

Another observation which promises to lead to much more useful results is the following: As before remarked, by adopting the method described, the body of a person may be subjected without danger to electrical pressures vastly in excess of any producible by ordinary apparatus, for they may amount to several million volts, as has been shown in actual practice. Now, when a conducting body is electrified to so high a degree, small particles, which may be adhering firmly to its surface, are torn off with violence and thrown to distances which can be only conjectured. I find that not only firmly adhering matter, as paint, for instance, is thrown off, but even the particles of the toughest metals are torn off. Such actions have been thought to be restricted to a vacuous inclosure, but with a powerful coil they occur also in the ordinary atmosphere. The facts mentioned would make it reasonable to expect that this extraordinary effect which, in other ways, I have already usefully applied, will likewise prove to be of value in electro-therapy. The continuous improvement of the instruments and the study of the phenomenon may shortly lead to the establishment of a novel mode of hygienic treatment which would permit an instantaneous cleaning of the skin of a person, simply by connecting the same to, or possibly, by merely placing the person in the vicinity of a source of intense electrical oscillations, this having the effect of throwing off, in a twinkle of the eye, dust or particles of any extraneous matter adhering to the body. Such a result brought about in a practicable manner would, without doubt, be of incalculable value in hygiene and would be an efficient and time-saving substitute for a water bath, and particularly appreciated by those whose contentment consists in undertaking more than they can accomplish.

High frequency impulses produce powerful inductive actions and in virtue of this feature they lend themselves in other ways to the uses of the electro-therapeutist. These inductive effects are either electrostatic or electrodynamic. The former diminish much more rapidly with the distance — with the square of the same — the latter are reduced simply in proportion to the distance. On the other hand, the former grow with the square of intensity of the source, while the latter increase in a simple proportion with the intensity. Both of these effects may be utilized for establishing a field of strong action extending through considerable space, as through a large hall, and such an arrangement might be suitable for use in hospitals or institutions of this kind, where it is desirable to treat a number of patients at the same time.

Fig. 6 illustrates the manner, as I have shown it originally, in which such a field of electrostatic action is established. In this diagram G is a generator of currents of very high frequency, C a condenser for counteracting the self-induction of the circuit which includes the primary P of an induction coil, the secondary S of which has two plates t t of large surface connected to its terminals. Well known adjustments being observed, a very strong action occurs chiefly in the space between the plates, and the body of a person is subjected to rapid variations of potential and surgings of current, which produce, even at a great distance, marked physiological effects. In my first experiments I used two metal plates as shown, but later I found it preferable to replace them by two large hollow spheres of brass covered with wax of a thickness of about two inches. The cables leading to the terminals of the secondary coil were similarly covered, so that any of them could be approached without danger of the insulation breaking down. In this manner the unpleasant shocks, to which the experimenter was exposed when using the plates, were prevented.

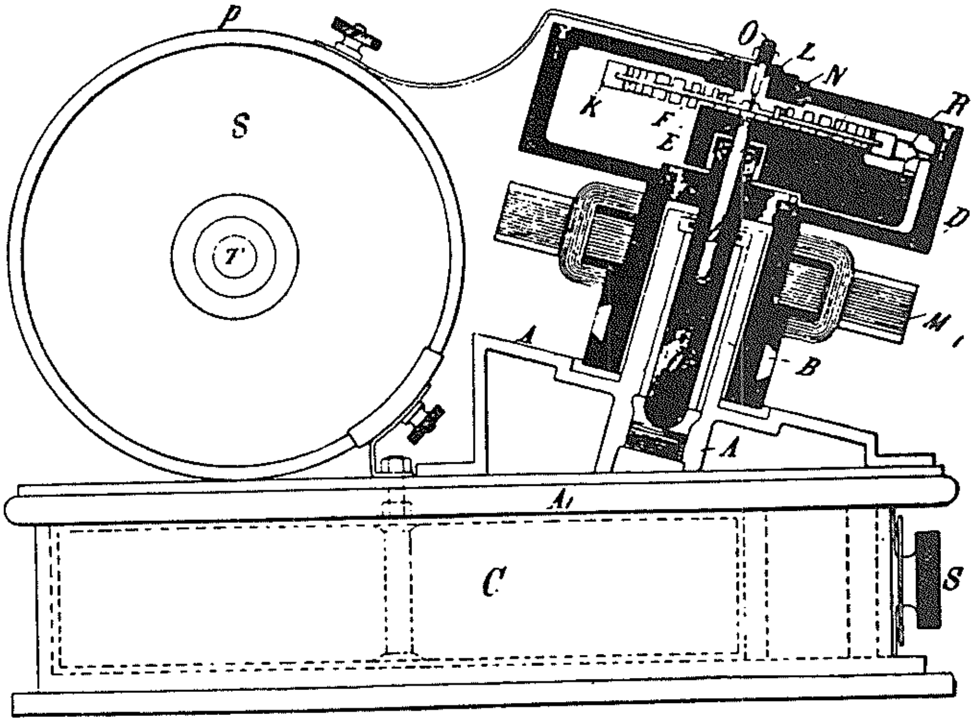

Fig. 9.

In Fig. 7 a plan for similarly utilizing the dynamic inductive effects of high frequency currents is illustrated. As the frequencies obtainable from an alternator are not as high as is desired, conversion by means of condensers is resorted to. The diagram will be understood at a glance from the foregoing description. It only need be stated that the primary p, through which the condensers are made to discharge, is formed by a thick stranded cable of low self-induction and resistance, and passes all around the hall. Any number of secondary coils s s s, each consisting generally of a single layer of rather thick wire, may be provided. I have found it practicable to use as many as one hundred, each being adjusted for a definite period and responding to a particular vibration passed through the primary. Such a plant I have had in use in my laboratory since 1892, and many times it has contributed to the pleasure of my visitors and also proved itself of practical utility. On a latter occasion I had the pleasure of entertaining some of the members with experiments of this kind, and this opportunity I cannot let pass without expressing my thanks for the interest which was awakened in me by their visit, as well as for the generous acknowledgment of the courtesy by the Association. Since that time my apparatus has been very materially improved, and now I am able to create a field of such intense induction in the laboratory that a coil three feet in diameter, by careful adjustment, will deliver energy at the rate of one-quarter of a horse power, no matter where it is placed within the area inclosed by the primary loops. Long sparks, streamers and all other phenomena obtainable with induction coils are easily producible anywhere within the space, and such coils, though not connected to anything, may be utilized exactly as ordinary coils, and what is still more remarkable, they are more effective. For the past few years I have often been urged to show experiments in public, but, though I was desirous to comply with such requests, pressing work has so far made it impossible. These advances have been the result of slow but steady improvement in the details of the apparatus which I hope to be able to describe connectedly in the near future.

However remarkable the electrodynamic inductive effects, which I have mentioned, may appear, they may be still considerably intensified by concentrating the action upon a very small space. It is evident that since, as before stated, electromotive forces of many thousand volts are maintained between two points of a conducting bar or loop only a few inches long, electromotive forces of approximately the same magnitude will be set up in conductors situated near by. Indeed, I found that it was practicable in this manner to pass a discharge through a highly exhausted bulb, although the electromotive force required amounted to as much as ten or twenty thousand volts, and for a long time I followed up experiments in this direction with the object of producing light in a novel and more economical way. But the tests left no doubt that there was great energy consumption attendant to this mode of illumination, at least with the apparatus I had then at command, and, finding another method which promised a higher economy of transformation, my efforts turned in this new direction. Shortly afterward (some time in June, 1891), Prof. J. J. Thomson described experiments which were evidently the outcome of long investigation, and in which he supplied much novel and interesting information, and this made me return with renewed zeal to my own experiments. Soon my efforts were centered upon producing in a small space the most intense inductive action, and by gradual improvement in the apparatus I obtained results of a surprising character. For instance, when the end of a heavy bar of iron was thrust within a loop powerfully energized, a few moments were sufficient to raise the bar to a high temperature. Even heavy lumps of other metals were heated as rapidly as though they were placed in a furnace. When a continuous band formed of a sheet of tin was thrust into the loop, the metal was fused instantly, the action being comparable to an explosion, and no wonder, for the frictional losses accumulated in it at the rate of possibly ten horse power. Masses of poorly conducting material behaved similarly, and when a highly exhausted bulb was pushed into the loop, the glass was heated in a few seconds nearly to the point of melting.

When I first observed these astonishing actions, I was interested to study their effects upon living tissues. As may be assumed, I proceeded with all the necessary caution, and well I might, for I had the evidence that in a turn of only a few inches in diameter an electromotive force of more than ten thousand volts was produced, and such high pressure would be more than sufficient to generate destructive currents in the tissue. This appeared all the more certain as bodies of comparatively poor conductivity were rapidly heated and even partially destroyed. One may imagine my astonishment when I found that I could thrust my hand or any other part of the body within the loop and hold it there with impunity. More than on one occasion, impelled by a desire to make some novel and useful observation, I have willingly or unconsciously performed an experiment connected with some risk, this being scarcely avoidable in laboratory experience, but have always believed, and do so now, that I have never undertaken anything in which, according to my own estimation, the chances of being injured were so great as when I placed my head within the space in which such terribly destructive forces were at work. Yet I have done so, and repeatedly, and have felt nothing. But I am firmly convinced that there is great danger attending such experiment, and some one going just a step farther than I have gone may be instantly destroyed. For, conditions may exist similar to those observable with a vacuum bulb. It may be placed in the field of the loop, however intensely energized, and so long as no path for the current is formed, it will remain cool and consume practically no energy. But the moment the first feeble current passes, most of the energy of the oscillations rushes to the place of consumption. If by any action whatever, a conducting path were formed within the living tissue or bones of the head, it would result in the instant destruction of these and death of the foolhardy experimenter. Such a method of killing, if it were rendered practicable, would be absolutely painless. Now, why is it that in a space in which such violent turmoil is going on living tissue remains uninjured? One might say the currents cannot pass because of the great self-induction offered by the large conducting mass. But this it cannot be, because a mass of metal offers a still higher self-induction and is heated just the same. One might argue the tissues offer too great a resistance. But this again cannot be the reason, for all evidence shows that the tissues conduct well enough, and besides, bodies of approximately the same resistance are raised to a high temperature. One might attribute the apparent harmlessness of the oscillations to the high specific heat of the tissue, but even a rough quantitative estimate from experiments with other bodies shows that this view is untenable. The only plausible explanation I have so far found is that the tissues are condensers. This only can account for the absence of injurious action. But it is remarkable that, as soon as a heterogeneous circuit is constituted, as by taking in the hands a bar of metal and forming a closed loop in this manner, the passage of the currents through the arms is felt, and other physiological effects are distinctly noted. The strongest action is, of course, secured when the exciting loop makes only one turn, unless the connections take up a considerable portion of the total length of the circuit, in which case the experimenter should settle upon the least number of turns by carefully estimating what he loses by increasing the number of turns, and what he gains by utilizing thus a greater proportion of the total length of the circuit. It should be borne in mind that, when the exciting coil has a considerable number of turns and is of some length, the effects of electrostatic induction may preponderate, as there may exist a very great difference of potential — a hundred thousand volts or more — between the first and last turn. However, these latter effects are always present even when a single turn is employed.

When a person is placed within such a loop, any pieces of metal, though of small bulk, are perceptibly warmed. Without doubt they would be also heated — particularly if they were of iron — when embedded in living tissue, and this suggests the possibility of surgical treatment by this method. It might be possible to sterilize wounds, or to locate, or even to extract metallic objects, or to perform other operations of this kind within the sphere of the surgeon's duties in this novel manner.

Most of the results enumerated, and many others still more remarkable, are made possible only by utilizing the discharges of a condenser. It is probable that but a very few — even among those who are working in these identical fields — fully appreciate what a wonderful instrument such a condenser is in reality. Let me convey an idea to this effect. One may take a condenser, small enough to go in one's vest pocket, and by skilfully using it he may create an electrical pressure vastly in excess — a hundred times greater if necessary — than any producible by the largest static machine ever constructed. Or, he may take the same condenser and, using it in a different way, he may obtain from it currents against which those of the most powerful welding machine are utterly insignificant. Those who are imbued with popular notions as to the pressure of static machines and currents obtainable with a commercial transformer, will be astonished at this statement — yet the truth of it is easy to see. Such results are obtainable, and easily, because the condenser can discharge the stored energy in an inconceivably short time. Nothing like this property is known in physical science. A compressed spring, or a storage battery, or any other form of device capable of storing energy, cannot do this; if they could, things udreamt of at present might be accomplished by their means. The nearest approach to a charged condenser is a high explosive, as dynamite. But even the most violent explosion of such a compound bears no comparison with the discharge or explosion of a condenser. For, while the pressures which are produced in the detonation of a chemical compound are measured in tens of tons per square inch, those which may be caused by condenser discharges may amount to thousands of tons per square inch, and if a chemical could be made which would explode as quickly as a condenser can be discharged under conditions which are realizable — an ounce of it would quite certainly be sufficient to render useless the largest battleship.

That important realizations would follow from the use of an instrument possessing such ideal properties I have been convinced since long ago, but I also recognized early that great difficulties would have to be overcome before it could replace less perfect implements now used in the arts for the manifold transformations of electrical energy. These difficulties were many. The condensers themselves, as usually manufactured, were inefficient, the conductors wasteful, the best insulation inadequate, and the conditions for the most efficient conversion were hard to adjust and to maintain. One difficulty, however, which was more serious than the others, and to which I called attention when I first described this system of energy transformation, was found in the devices necessarily used for controlling the charges and discharges of the condenser. They were wanting in efficiency and reliability and threatened to prove a decided drawback, greatly restricting the use of the system and depriving it of many valuable features. For a number of years I have tried to master this difficulty. During this time a great number of such devices were experimented upon. Many of them promised well at first, only to prove inadequate in the end. Reluctantly, I came back upon an idea on which I had worked long before. It was to replace the ordinary brushes and commutator segments by fluid contacts. I had encountered difficulties then, but the intervening years in the laboratory were not spent in vain, and I made headway. First it was necessary to provide for a circulation of the fluid, but forcing it through by a pump proved itself impractical. Then the happy idea presented itself to make the pumping device an integral part of the circuit interrupter, inclosing both in a receptacle to prevent oxydation. Next some simple ways of maintaining the circulation, as by rotating a body of mercury, presented themselves. Then I learned how to reduce the wear and losses which still existed. I fear that these statements, indicating how much effort was spent in these seemingly insignificant details will not convey a high idea of my ability, but I confess that my patience was taxed to the utmost. Finally, though, I had the satisfaction of producing devices which are simple and reliable in their operation, which require practically no attention and which are capable of effecting a transformation of considerable amounts of energy with fair economy. It is not the best that can be done, by any means, but it is satisfactory, and I feel that the hardest task is done.

The physician will now be able to obtain an instrument suitable to fulfil many requirements. He will be able to use it in electro-therapeutic treatment in most of the ways enumerated. He will have the facility of providing himself with coils such as he may desire to have for any particular purpose, which will give him any current or any pressure he may wish to obtain. Such coils will consist of but a few turns of wire, and the expense of preparing them will be quite insignificant. The instrument will also enable him to generate Röntgen rays of much greater power than obtainable with ordinary apparatus. A tube must still be furnished by the manufacturers which will not deteriorate and which will allow to concentrate larger amounts of energy upon the electrodes. When this is done, nothing will stand in the way of an extensive and efficient application of this beautiful discovery which must ultimately prove itself of the highest value, not only at the hands of the surgeon, but also of the electro-therapist and, what is most important, of the bacteriologist.

To give a general idea of an instrument in which many of the latter improvements are embodied, I would refer to Fig. 9, which illustrates the chief parts of the same in side elevation and partially in vertical cross-section. The arrangement of the parts is the same as in the form of instrument exhibited on former occasions, only the exciting coil with the vibrating interrupter is replaced by one of the improved circuit breakers to which reference has been made.

This device comprises a casting A with a protruding sleeve B, which in a bushing supports a freely rotatable shaft a. The latter carries an armature within a stationary field magnet M and on the top, a hollow iron pulley D, which contains the break proper. Within the shaft a, and concentrically with the same, is placed a smaller shaft b, likewise freely movable on ball-bearings and supporting a weight E. This weight being on one side and the shafts a and b inclined to the vertical, the weight remains stationary as the pulley is rotated. Fastened to the weight E is a device R in the form of a scoop with very thin walls, narrow on the end nearer to the pulley and wider on the other end. A small quantity of mercury being placed in the pulley and the latter rotated against the narrow end of the scoop, a portion of the fluid is taken up and thrown in a thin and wide stream towards the centre of the pulley. The top of the latter is hermetically closed by an iron washer, as shown, this washer supporting on a steel rod L a disk F of the same metal provided with a number of thin contact blades K. The rod L is insulated by washers N from the pulley, and for the convenience of filling in the mercury a small screw o is provided. The bolt L forming one terminal of the circuit breaker is connected by a copper strip to the primary p. The other end of the primary coil leads to one of the terminals of the condenser C, contained in a compartment of a box A, another compartment of the same being reserved for switch S and terminals of the instrument. The other terminal of the condenser is connected to the casting A and through it to pulley D. When the pulley is rotated, the contact blades K are brought rapidly in and out of contact with the stream of mercury, thus closing and opening the circuit in quick succession. With such a device it is easy to obtain ten thousand makes and breaks per second and even more. The secondary a is made of two separate coils and so arranged that it can be slipped out, and a metal strip in its middle connects it to the primary coil. This is done to prevent the secondary from breaking down when one of the terminals is overloaded, as it often happens in working Röntgen bulbs. This form of coil will withstand a very much greater difference of potential than coils as ordinarily constructed.

The motor has both field and armature built of plates, so that it can be used on alternating as well as direct current supply circuits, and the shafts are as nearly as possible vertical, so as to require the least care in oiling. Thus, the only thing which really requires some attention is the commutator of the motor, but where alternating currents are always available, this source of possible trouble is easily done away with.

The circuit connections of the instrument have been already shown and the mode of operation explained in periodicals. The usual manner of connecting is illustrated in Fig. 8, in which A_2 A_2 are the terminals of the supply circuit, L, a self-induction coil for raising the pressure, which is connected in series with condenser C and primary P P. The remaining letters designate the parts correspondingly marked in Fig. 9 and will be understood with reference to the latter.

THE MAGNETIC-ARC INSTRUMENT OF S.S. KNIGHT

by Mark L. Gallert

Reprinted from

NEW LIGHT ON THERAPEUTIC ENERGIES, 1976

Mr. Knight, a chemical and electrical engineer of Walnut Creek, California, was impressed with the severity and persistence of pain suffered in certain types of cancer. Realizing that pain is a sensation representative of an electrical phenomenon of excessive nerve-current flow, he decided to experiment to see if there was some way of interfering with that nerve flow through artificial impulses, in the effort to stop or alleviate the pain.

Towards this end, he developed an instrument combining several features of electro-therapy:

1. A wave pulsed at 90 times per minute, taking advantage of data obtained from experience with the older types Faradic machines.
2. A rest interval exceeding 75% of the duration of each pulsing cycle - enhances the beneficial effect.
3. Breaking the treatment by mechanical opening of contacts in such a way as to create an arc at each of the pulsing intervals. An arc produces a very wide band of frequencies, from those of very low frequency, to the highest that can be measured. The multiplicity of frequencies seems to play a part in masking or interrupting the sensation of pain.
4. A unique feature of the Knight instrument is that the arc is broken in a strong magnetic field produced by a large, powerful, permanent magnet. This introduces what is called the "Zeeman effect", that is, a magnification of the arc impulses, and a spreading of frequencies.
5. Use of very low power output - below 1/100 of a watt, so that there will be no perceptible heating of tissue when the complex of impulses from the instrument are fed to the patient through electrodes applied to the skin.

Experiments on cases which had been declared terminal cancer by medical doctors, showed that this device not only brought outstanding relief of pain in the majority of instances, but also produced some surprising "remissions" characterized by shrinking of tumour masses, and restoration of general health to the patient.

Dr. C.P. Bryant of Seattle, Washington, who used these instruments extensively in his practice over a period of years, declared that with their use he was able to restore to health 75% of all cancer cases which had not undergone surgery, X-ray or radium, and 25% of those cancer cases which had been subjected to surgery, X-ray or radium.

The Knight family has never represented these instruments as a "cancer cure". The original aim was to find some way to relieve the pain of cancer. Reports from doctors of apparent cures through use of the instruments were received with interest, as well as reports of benefits from the application of the instrument to conditions other than cancer.

Mr. S.S. Knight, a man with pronounced humanitarian ideals, made the circuit of the instrument available to all inquirers and encouraged the building of instruments by anyone who wished to copy his circuit and specifications. He was keenly disappointed when reports reached him of failure, or lack of therapeutic effect from instruments constructed by others. It developed that adjustments were quite critical, particularly concerning the making and breaking of the arc. Mr. Knight could tell by looking at the arc and listening to it, when it was "right". It was difficult to develop objective criteria for adjustment of the equipment.

A puzzling circumstance, was that the maximum effect from Mr. Knight's instruments was obtained when they were new; after a few weeks of use the effect began to diminish; after some months it was perceptibly weaker, and after a year or more, of very little use. Several years ago, the reason for the lessening of effect was determined. The isolation transformer used to block the direct flow of the arc frequencies from the electric line supply (an isolation required to eliminate radio interference) was being saturated by the magnetism from the large magnet surrounding the arc. The more saturated the transformer became, the less electricity it passed. The equipment was re-designed to place the transformer out of range of the field of the magnet.

The instrument can be classed with experimental devices, as there have not been sufficient tests made under controlled conditions to establish with certainty the percentage of cases of various types of cancer that can be alleviated, arrested, and perhaps cured, but the tests to date indicate at least that this is a promising approach, which should be explored further.

An imitation instrument has been marketed by others, using an electro-magnet instead of a permanent magnet. Mr. Knight was opposed to this substitution, having a number of reasons for believing that the permanent magnet was preferable. Other changes have also been made in constructing these imitation instruments, including the type of electrodes. So far as your compiler has been able to determine, the instruments using electro-magnets have some value in relieving pain in a variety of conditions, but the effect on cancer seems to be slower and less pronounced than that obtained from the genuine Knight instruments when the latter are functioning properly and are in adjustment.

electronics
April 20, 1962

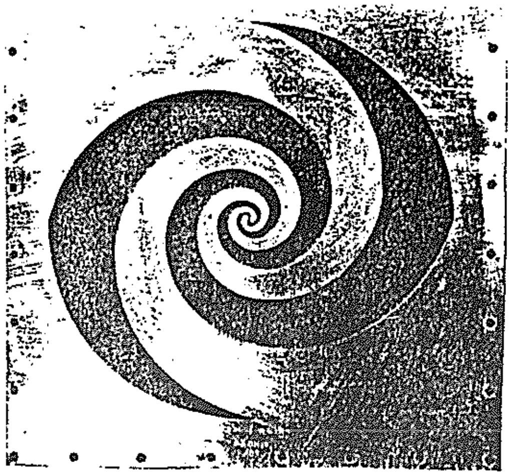

FIG. 1—Balanced planar equiangular or logarithmic spiral-slot antenna (left) radiates a broad circularly polarized beam on each side of, and perpendicular to, the plane of the antenna

Frequency-Independent Antennas

SURVEY OF DEVELOPMENT

Antennas now provide continuous coverage of large portions of the frequency spectrum. These very-wide-band antennas offer a challenge to the systems engineer to design equipment that will match bandwidths now available

By JOHN D. DYSON
University of Illinois,
Urbana, Illinois

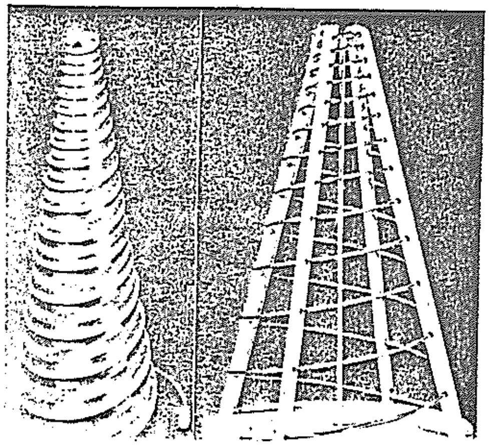

FIG. 2—Balanced conical logarithmic spiral antenna (left) with feed cable carried along one arm; wire version of balanced conical logarithmic spiral antenna (right), fed with coaxial line and balun along axis of cone

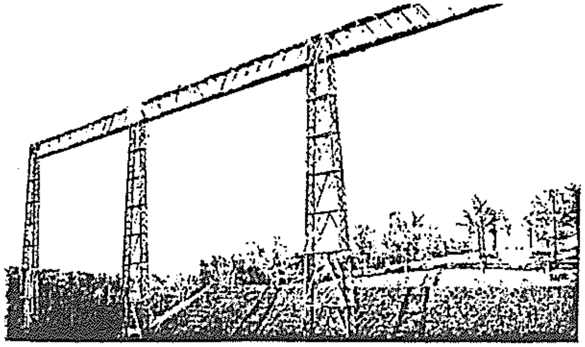

FIG. 3—The University of Illinois radio telescope. The focal line array, 153 feet above the 400-by-600-foot parabolic reflector, is visible below the wooden truss

THE NECESSITY of providing continuous coverage of large portions of the electromagnetic frequency spectrum for defense purposes has exerted great pressure on the antenna engineer to produce antennas that could be used over wider and wider bands of frequencies. Such requirements have led to an important breakthrough in the antenna field. The conception of the frequency-independent antennas at the University of Illinois and their development at this and at other laboratories has produced advances in the state of the art that now puts pressure on the electronics and system engineer to design equipment that will match the available antenna bandwidths. The purpose of this survey is to outline the development of these antennas.

As of 1945, antenna research and development had led to a few structures that operated over a 2 to 1 or slightly greater range of frequencies, with acceptable radiation and impedance characteristics. In a majority of cases, it was the radical changes in the radiation pattern with changes in operating frequency which limited the usefulness of antennas to somewhat less than this 2 to 1 range, since antennas had been developed with an input impedance that remained relatively constant with a change in frequency.

The discone antenna[1] introduced in 1946 was one of the first antennas that could be used over approximately a 4-to-1 range of frequencies. Advances during the next 7 or 8 years were limited, mainly, to increasing the bandwidth of existing antennas by a few percent. The one exception during this period was a modified helical antenna[2, 3, 4]. These unbalanced conical helical structures, fed against a ground plane, had acceptable pattern bandwidths of 4 or 5 to 1.

The spiral antenna introduced in 1953 was a major step forward in wide band antenna design. This construction consists of two flat, narrow, constant-width metallic arms, wound in an Archimedean spiral.[5] Fed at the center in a balanced manner, it radiates a broad circularly polarized lobe on each side of the plane of the antenna. Subsequent development of this configuration produced forms that had stable pattern and impedance characteristics over bandwidths up to 10 to 1.[6-10]

Arrays of the Archimedes spiral antenna provide variable polarization of the radiated field[11] and beam scanning by simple rotation of the antennas.[12-15]

In 1954 a major breakthrough occurred when V. H. Rumsey suggested that an infinite length structure based upon the equiangular or logarithmic spiral could be specified completely in terms of angles, and hence it should have characteristics independent of frequency. A study of balanced planar antennas based on these curves disclosed the equiangular spiral antenna to be the first antenna of practical size that had characteristics associated with an infinite structure.[16, 17] This structure became the first of a class now called frequency-independent antennas.[18]

This descriptive term is applied to antennas set apart from the usual concepts of operating bandwidth. The upper and lower frequencies at which these antennas may be operated are determined by independent paramters, and hence the bandwidth over which the radiation pattern and input characteristics are essentially constant is theoretically unlimited. Actually, the bandwidth is limited only by practical considerations of construction, and such antennas have been built to operate over a bandwidth of more than 42 to 1. There is no fundamental reason why this bandwidth cannot be extended.

The general philosophy behind the frequency-independent antennas is to design a radiating structure such that successive applications of a scaling factor result in structures identical with the original, or which at most, differ by some rotation about an axis through the origin of the original structure. Although the structure must be infinite in size to fulfill the scaling condition of exact equivalence, forms based upon the equiangular, or logarithmic, spiral curve may be truncated to a finite size and still retain, over an extremely wide range of frequencies, characteristics associated with the infinite structure.

A second breakthrough occurred in 1955 when R. H. Du Hamel proposed to force radiation from otherwise angular structures by periodic discontinuities. This led to a linearly polarized antenna with pattern and impedance characteristics that repeat periodically with the logarithm of frequency.[19-21] Although these logarithmic-periodic antennas scale exactly at only certain discrete frequencies, they can be designed so that the radiation characteristics do not change appreciably over the period between these frequencies. Hence they also exhibit nearly frequency-independent characteristics.

Many versions of the log-spiral and log-periodic antennas have been developed. The planar log-spiral antenna shown in Fig. 1 is bidirectional, radiating a broad circularly polarized beam on each side of, and perpendicular to, the antenna plane.[22] This spiral antenna may be made unidirectional, and hence more useful, by an orthogonal projection of the balanced spiral arms upon the surface of a cone (see Fig. 2 left).[23-25] For included cone angles less than about 45 degrees, radiation is confined to one broad circularly polarized lobe with maximum radiation off the apex of the cone. The antenna is balanced, with the feed voltage applied between the two arms at the apex of the cone. This feed voltage may be supplied by a conventional balanced line carried along the axis of the cone, or by a coaxial line and balun.

The characteristics of equiangu-

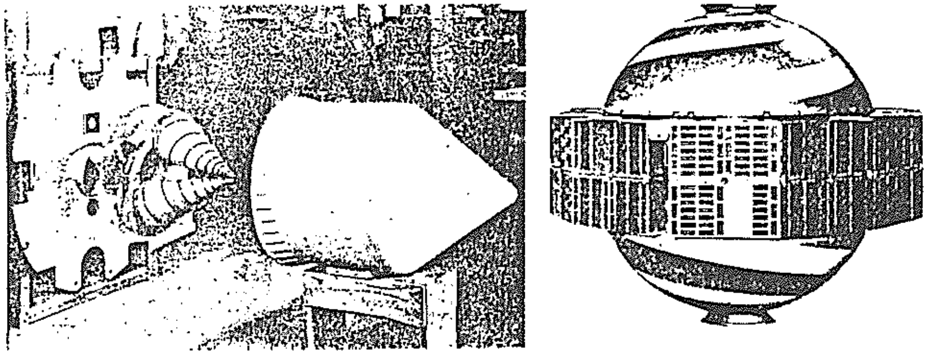

FIG. 4—First production of Dynatronics 500-to-4,000-Mc dual conical spiral Autorack feed for 40-foot diameter parabolic reflector (left); Transit 3B satellite (right) by Applied Physics Lab. of the Johns Hopkins University. The logarithmic spiral antenna projected onto the spherical surface provides for transmission at 54, 108, 162 and 216 Mc

lar spiral antennas permit a third method. The feed line, a coaxial cable, may be carried along and soldered in contact with one of the arms. The amplitude of the antenna currents on the arms, and also on the outside of the cable, fall off rapidly with distance from the apex, so the ends of the arms where the cable enters carry negligible antenna current. This optional feed, conveniently and automatically, provides a frequency-independent balun, permitting feeding the balanced antenna by means of an unbalanced line.

Conical log-spiral antennas have been constructed to operate over more than 40-to-1 bandwidths. The upper usable frequency is determined by the truncated region at the apex which must remain less than 1/4 wavelength at the highest frequency of operation. The lowest usable frequency is determined by the base diameter of the cone which, for tightly wrapped antennas, must be approximately 3/8 wavelength at lowest operating frequency.

A highly practical conical log-spiral antenna can be constructed by allowing the exponentially expanding arms to degenerate into narrow constant width structures following the logarithmic spiral curve (see Fig. 2 right). For tightly wrapped spirals this version has characteristics similar to the original antenna. This wire version, simple to construct, may be excited by bringing the feed cable up the axis of the cone. If the infinite balun feed is desired, the feed cable itself can become one of the arms, with a dummy cable becoming the second arm of a balanced structure.

The balanced conical log-spiral antenna has characteristics other than bandwith that are attractive. The electromagnetic fields radiated by these antennas are essentially circularly polarized over very wide angles off the axis of the antenna. In addition, the phase of the far field of the two-arm antenna has essentially a one-to-one relationship with the azimuthal angle around the antenna. These two properties were utilized in the line feed for the University of Illinois radio telescope shown in Fig. 3. The focal-line array for this 400-by-600-foot parabolic reflector consists of a linear array of 286 conical logspiral antennas.". The beam is steerable in declination ±30 degrees and the phase adjustment to provide this scan is secured by a physical rotation of each element of the array to a prescribed orientation. Thus it is of importance that the beam of each element be circularly polarized over at least a sector ±30 degrees from the axis and that the phase of the radiated field be linearly related to the angle of orientation around the antenna.

Multiarm versions of the balanced conical log-spiral antenna operate in several distinct modes. Thus a four-arm symmetrical structure may be excited to radiate an axial beam with an $e^{j\phi}$ variation, that is, the phase of the far field varies linearly with the azimuthal angle ϕ, around the antenna. It can be excited to produce a conical beam with an $e^{j2\phi}$ and a conical beam with an $e^{j3\phi}$ variation. Since orientation of the maximum of the conical beam may be controlled by varying the rate of spiral of the antenna arms, it can be positioned perpendicular to the axis of the antenna to provide a simple very-wide-band circularly polarized omnidirectional source. The well-behaved phase characteristics of these antennas suggest the basis for many variations of phase-comparison direction-finding systems.

The conical log-spiral antennas have been developed as feeds for dishes used in tracking applications. A modernization program is now under way to equip many of the large parabolic reflector antennas on the Atlantic Missile Range with dual conical spiral feeds designed to cover a 10-to-1 range of frequencies. The first production model of one of these feeds is shown in Fig. 4 (left).

In an interesting modification of the log-spiral geometry, the planar antenna, was projected upon a spherical surface and used on the Transit satellites shown in Fig. 4 (right). Although this geometry is no longer frequency independent, the antenna operates well over four frequencies, spaced from 54 to 216 Mc. Similarly, the planar antenna can be projected upon a myriad of surfaces. However, it can be shown that the conical surface, which includes the plane, is the only surface satisfying the required scaling principles for unlimited bandwidth.

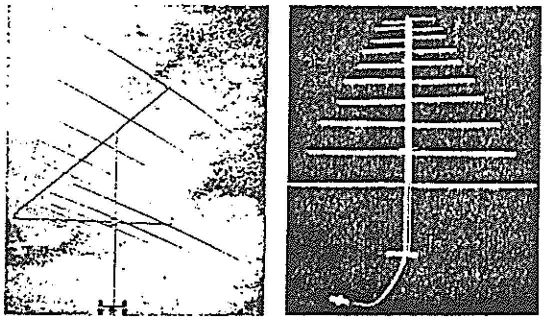

FIG. 5—Trapezoidal tooth nonplanar log-periodic antenna (left) designed by Collins Radio Company to cover 6.5 to 60 Mc; and logarithmic periodic dipole array (right)

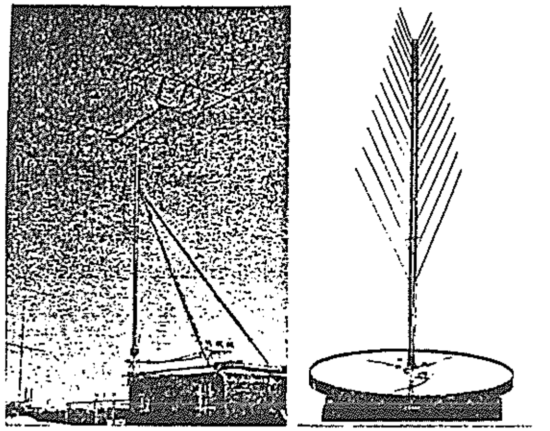

FIG. 6—Militarized version of the log-periodic dipole array (left) by Collins Radio Company; and log-periodic resonant-V array (right), an adaption of the log-periodic dipole array, designed to operate in several different modes

The log-periodic antennas have been extensively investigated. The planar antenna can be modified to provide unidirectional patterns by forming the two elements of the antenna into a V, that is, a nonplanar arrangement. A wire version of the log-periodic elements (see Fig. 5 left), and omnidirectional horizontally polarized, and unidirectional circularly polarized log-periodic antennas have been described. Multielement arrays have been described, and design information for log-periodic antennas as feeds for dishes, and as point-to-point h-f communications antennas has been given.

The dipole array of Fig. 5 (right) is a simple log-periodic structure. The length of each element in this array bears a fixed ratio to the length of the preceding element, and the adjacent element spacings bear the same ratio one to another. The pattern and impedance characteristics of the array at frequency f_n, such that the nth element is resonant, will be repeated at a higher frequency f_{n+1}, which makes the $(n + 1)$th element resonant. The characteristics repeat periodically at all frequencies given by $\tau^n f$, where n is an integer and τ the ratio of the lengths of adjacent elements. The geometry shown in this figure provides an excitation with a 180-deg phase shift between elements. Thus it radiates a single-lobe, linearly polarized beam, directed toward the apex of the array. Analysis and design information for this array has been presented. A militarized version of the antenna is shown in Fig. 6 (left).

In an extension of this work, a log-periodic monopole array has been described in which simple reactive networks produce the proper monopole current amplitude and phase distribution for unidirectional frequency-independent operation. A slot complement of the log-periodic dipole array has also been investigated.

The log-periodic resonant-V array (Fig. 6 right) is an adaptation of the log-periodic dipole array, designed to operate in several modes. In the lowest order mode, the performance approximates the log-periodic array. The active portion of the antenna centers around elements whose lengths are near a half-wavelength at the frequency of operation. As the operating frequency is increased, this active region moves toward the apex. The antenna may be designed so that as the $\lambda/2$ region moves off the apex of the antenna the large elements on the other end approach a $3\lambda/2$ resonance. Thus, as the operating frequency is increased, the active region moves through the array in the $3\lambda/2$ wave mode, then in the $5\lambda/2$ mode, etc. Forward tilt of the individual elements insures a unidirectional beam and provides increased directivity. This antenna has transition regions between modes where the characteristics deteriorate, but it can provide a gain of approximately 18 db over an isotropic radiator in the higher modes.

Triangular or zig-zag log-periodic structures have received considerable attention. A flush-mounted sinuous structure, fed by a waveguide, has been studied, and a balanced zig-zag antenna has been proposed. A vertically polarized log-periodic zig-zag antenna is shown in Fig. 7. The teeth bent parallel to the ground plane serve as a delay line to phase the vertical radiating portions.

Tapering structures, not constructed in a logarithmic periodic manner, have led to antennas of appreciable bandwidth. One tapered ladder antenna consists of symmetrical elements excited by a balanced line. In addition long slot, tapered feed, structures that can be used over a 10-to-1 bandwidth have been proposed. A radiating horn with side walls of periodically spaced wires, when fed from a ridged waveguide, can be operated over a 4-to-1 range of frequencies.

The frequency-independent antennas have opened up a new era in wide-band antennas. However, in view of their complex geometry, the experimental and empirical development of these antennas have outstripped their theoretical analysis.

The theoretical problem of the log-spiral antenna has received considerable attention. This problem has been attacked by calculating the fields due to a logarithmic spiral filament, and the fields produced by an infinite number of planar logarithmic spiral filaments. An approximate analysis for the radiation patterns of a filamentary log-spiral antenna with an assumed current distribution has been published.

A recent comprehensive analysis of the log-periodic dipole array has been reported.

A new and promising approach considers the log-periodic antenna as a locally periodic structure whose period varies slowly, increasing linearly with distance to the origin or apex. This approach indicates that the basic ingredients for a frequency-independent antenna are a slow-wave transmission medium and a series of radiating elements, satisfying the similarity condition, coupled to the transmission medium at points spaced in geometric progression. Such a structure produces a wave substantially radiated toward the feed point with characteristics that vary little as the operating frequency is varied over extremely wide bandwidths.

A general consideration of the log periodic zig-zag antennas and the conical log-spiral antennas as locally periodic, slow wave structures has led to a better understanding of their operation and to some useful design criteria.

This survey has covered the work of many people. It has been impossible to acknowledge the work of many others, however an extensive list of references has been included.

The original work on logarithmic spiral and logarithmic periodic antennas was conducted at the University of Illinois, sponsored by the Aeronautical Systems Division, Wright Air Development Division under contracts AF33(616)-3220 and AF33(616)-6079.

This article is based upon a paper by the author published in the Journal of Research of the National Bureau of Standards, Section D, January 1962.

REFERENCES

(1) A. G. Kandoian, Three New Antenna Types and Their Applications, *Proc. IRE*, 34, p 70w, Feb. 1946.
(2) P. S. Springer, End-Loaded and Expanding Helices as Broadband Circularly Polarized Radiators, *Proceedings National Electronic Conference*, 5, 1949.
(3) J. S. Chatterjee, Radiation Field of a Conical Helix, *Journal of Applied Physics*, 24, p 550, May 1953.
(4) J. S. Chatterjee, Radiation Characteristics of a Conical Helix of Low Pitch Angles, *J. Appl. Phys.*, 26, p 331, Mar. 1955.
(5) E. M. Turner, Spiral Slot Antenna, Wright Air Development Center, Dayton, Ohio, Technical Note WCLR-55-8, June 1955.
(6) B. H. Burdine, and R. M. McElvery, Spiral Antennas, MIT Research Laboratory of Electronics, Report No. 1, Mar. 1955.
(7) B. H. Burdine, and H. J. Zimmerman, The Spiral Antenna, MIT Research Laboratory of Electronics, Report No. 2, Apr. 1955.
(8) E. M. Turner, Broadband Radiators, Proc. *IRE Nat. Aero. Electronics Conference*, 1958.
(9) W. L. Curtis, Spiral Antennas *IRE Transactions*, AP-8, p 298, May 1960.
(10) J. A. Kaiser, The Archimedean Two-Wire Spiral Antenna, *IRE Transactions*, AP-8, p 312, May 1960.
(11) R. Bawer, and J. J. Wolfe, The Spiral Antenna, *IRE National Convention Record*, p 84, 1960.
(12) J. P. Jones, D. E. Taylor and C. W. Morrow, Design Techniques for a Light Weight High Power Spiral Antenna, *IRE WESCON Convention Record*, 4, Part 1, 1960.
(13) J. R. Donnellan, Second-Mode Operation of the Spiral Antenna, *IRE Trans.*, AP-8, p 637, Nov. 1960.
(14) S. R. Jones, and E. M. Turner, Polarization Control with Oppositely Sensed Circularly Polarized Antennas, Proc. *IRE National Aero. Electronics Conf.*, p 558, May 1959.
(15) R.M. Brown, Jr., and R. C. Dodson, Parasitic Spiral Arrays, *IRE International Convention Record*, 8, Part 1, p 51, 1960.
(16) J. R. Donnellan, and R. T. Close, A Spiral-Grating Array, *IRE Trans.*, AP-9, p 291, May 1961.
(17) J. R. Donnellan, A Spiral-Doublet Scanning Array, *IRE Trans.*, AP-9, p 276, May 1961.
(18) J. D. Dyson, The Equiangular Spiral Anténna, *Proc. Fifth Symp. on the USAF Ant. Research and Development Program*, Allerton Park, Monticello, Oct. 22, 1955. (Classified).
(19) J. D. Dyson, The Equiangular Spiral Antenna, *IRE Trans.*, AP-7, p 181, Apr. 1959.
(20) V. H. Rumsey, Frequency Independent Antennas, *IRE Nat. Conv. Record*, Part 1, p 114, 1957. Also Antenna Laboratory Technical Report No. 20, Contract 33 (616)-3220, Dept. of Electrical Engineering, University of Illinois, 25 Oct. 1957.
(21) R. H. DuHamel, U. S. Patent No. 2,985,879, 1958.
(22) R. H. DuHamel, and D. E. Isbell, Broadband Logarithmically Periodic Antenna Structures, *IRE Nat. Conv. Record*, Part 1, p 119, 1957.
(23) J. D. Dyson, The Equiangular Spiral Antenna, AD-145019, Technical Report No. 21, Contract AF33 (616-3220, Dept. of Electrical Engineering, Univ. of Ill., Sept. 15, 1957.
(24) N. Barbano, Phase Center Distributions of Spiral Antennas, *IRE WESCON Conv. Record*, 4, Part 1, p 123, 1960.
(25) J. D. Dyson, Recent Developments in Spiral Antennas, *Proc. IRE Nat. Aero. Elect. Conf.*, p 617, May 1959.
(26) M. S. Wheeler, Design Procedure for Equiangular Spiral Antennas, Seventh East Conf. on Aero. and Nav. Electronics, p 1.2.4, Oct. 1960.
(27) J. D. Dyson, U. S. Patent No. 2,958,-081, 1959.
(28) J. D. Dyson, The Non-Planar Equiangular Spiral Antenna, *Proc. Eighth Symposium, USAF Antenna Research and Development Program*, Robert Allerton Park, Monticello, Ill., Oct. 1958.
(29) J. D. Dyson, The Unidirectional Equiangular Spiral Antenna, *IRE Trans.*, Vol. AP-7, pp. 330-334, October, 1959. Also Antenna Laboratory Technical Report No. 33, Contract 33(616)-6079, Dept. of Electrical Engineering, University of Illinois, July 10, 1958.
(30) M. Nussbaum, Designing Broadband Conical Helix Antennas, *Electronic Design*, Sept. 3, 1958.
(31) H. S. Barsky, Broadband Conical Helix Antennas, *IRE Nat. Conv. Record*, 7, Part 1, p 130, 1959.
(32) J. D. Dyson, and P. E. Mayes, The Log-Spiral Omnidirectional Circularly Polarized Antenna, URSI-IRE Spring Meeting, Wash., D. C., May 3, 1960.
(33) J. D. Dyson, and P. E. Mayes, New Circularly Polarized Frequency Independent Antennas with Conical Beam or Omnidirectional Patterns, *IRE Trans.*, AP-9, No. 4, p 334, July 1961. Also Antenna Laboratory Technical Report No. 46, Contract AF33(616)-3220, Dept. of Electrical Engineering, Univ. of Ill., July 1960.
(34) J. D. Dyson, Multi-Mode Logarithmic Spiral Antennas—Possible Applications,, *Proc. National Electronic Conf.*, 17, p 206, Oct. 1961.
(35) J. D. Dyson, The Log-Spiral Antenna in Simple Arrays, *Proc. of the Eleventh Symposium, USAF Antenna Research and Development Program*, Allerton Park, Monticello, Ill., Oct. 1961.
(36) E. M. Turner, The Scimitar Antenna, *Proc. IRE Nat. Aero. Electronics Conf.* 1956.
(37) H. B. Riblet, A Broadband Spherical Satellite Antenna, *Proc. IRE*, 48, p 631, Apr. 1960.
(38) D. E. Isbell, Non-Planar Logarithmically Periodic Antenna Structures, *Proc. Seventh Symp. USAF Ant. Res. and Dev.*, Allerton Park, Monticello, Illinois, October 1957. Also Antenna Laboratory Technical Report No. 30, Contract AF33(616)-3220, Elect. Eng. Dept., Univ. of Ill., Feb. 1958.
(39) R. H. DuHamel, and F. R. Ore, Logarithmically Periodic Antenna Designs, *IRE Nat. Conv. Rec.*, Part 2, p 139, 1958.
(40) R. H. DuHamel, and D. G. Berry, Logarithmically Periodic Antenna Arrays, *IRE WESCON Conv. Rec.*, 2, Part 1, p 161, 1958.
(41) R. H. DuHamel, and F. R. Ore, Log Periodic Feeds for Lens and Reflectors, *IRE Nat. Conv. Rec.*, 7, Part 1, p 128, 1959.
(42) R. H. DuHamel, and D. G. Berry,

Medical Aspects of Ultra Short Waves

By C. H. West

U. S. Public Health Service

Mr. West here describes some of the effects and also some of the dangers of subjecting the human body to powerful ultra short wave fields

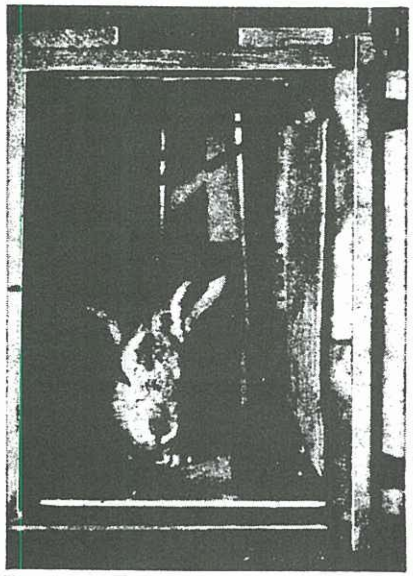

Photos courtesy J. G. Francis, U. S. Public Health Service. Member R. P. S.

Above—A rabbit about to be exposed to a powerful ultra short wave field; the sides as well as the top and bottom linings of the cabinet are of metal.

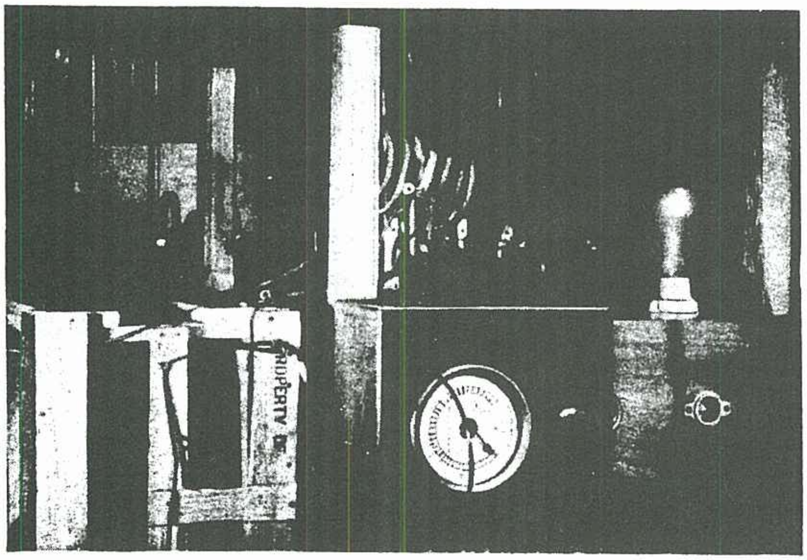

At left—The appearance of the ultra short-wave apparatus built and tested by Mr. West in his experiments.

● IT SEEMS that the general public is once again taking to radio and the art of building as a pastime. The amateurs are busy developing the five meter band, while the engineers are devoting a goodly portion of their time

to dabbling in the mysteries that the ultra short waves have created.

On the other hand, medical-scientific personages are making use of their medical knowledge, plus a large portion of electronics, in efforts to make the world a better place to live in with less aches and pains to bother us.

Many engineers employed in large electrical laboratories have made valuable use of their knowledge pertaining to the ultra short waves, and have constructed oscillators of high power for the purpose of producing fever in the human body at will. It is a known fact that fever is nature's doctor to a certain degree. If it can be produced in cases where there is no cause for a high fever, the patient's chance for recovery and cure is far better than combatting disease by injection of malaria germs, etc., which may produce reactions worse than the disease itself

Unknown Qualities

However, the production of fever by an oscillator is a simple procedure, but the application of these high wave speeds to the human body is an entirely different matter. Secondly, the output or radiation of an ultra high frequency oscillator has never been completely identified.

It seems that for every known element discovered there are many others within—undiscovered. An illustration of this fact was brought to light very recently, in which the writer witnessed a demonstration of photo-electric cell work.

A double-cell apparatus manufactured by the Weston Electrical Instrument Corporation produced readings in foot-candles from the output of a 10 k. w. X-ray tube in which the cells were completely covered by ⅛ inch lead sheet.

The roentgen rays do not penetrate lead, and as the tube was quite a distance from the cells, the reaction in the sensitive meter was not due to heat waves, but simply a current output which registered through the lead upon the faces of the cells. The X-ray is noted for producing very bad burns if not handled properly, and the deflection from the *target* is called and identified as the *roentgen ray* after the name of the discoverer. How does science know whether the burns are caused from this additional element as registered in *foot candles*, for want of a better name?

U. S. W. "Fever" Apparatus

Recently, the writer constructed and put into practice an *ultra short wave oscillator* for the production of fever or temperature. This apparatus consisted of two UX-852 tubes arranged in push-pull, but with the "back-to-back" arrangement of *Mesny-Valuri*. It is a persistent oscillator and is capable of running long hours on wavelengths from 2 to 10 meters.

An auxiliary cabinet was constructed with two leads to couple across the plate inductance. This cabinet has one adjustable side; the top and bottom being one lead, and the two sides the other and forming a large condenser.

Rabbits were used in experiments and their temperature reached 41 degrees (Centigrade) from the usual 38 degrees (normal). However, after a few weeks severe burns were noticed. Since that time research has been carried on by other factions in an effort to ascertain the direct cause of the burns, which do not seem to be due to collections of moisture, or from coming in contact with the condenser plates.

This brings to mind that experiments by the writer with the very short waves disclose the fact that some sort of a photo-electric ray could be produced and would highly expose small dental films. It is believed this elementary action was first discovered by Mr. John Reinartz, the noted radio amateur, who perceived a bluish glow within an auxiliary indicator at a certain frequency.

Before any actual application to humans may be applied, it is our duty to ascertain accurately whether other elements are prevailing. To all indications there are many, and one of them is severe burns, which may be caused by an unidentified element.

Constructional Data and Summary

As will be noted in the photographs, the apparatus is entirely enclosed and surrounded by glass sides to better view the "works" in case things tend to go the "hay-wire" route. The lower portion of the cabinet contains the necessary plate and filament transformers, which operate direct from the 110 volt, 60 cycle line. The plate power is raw A. C. at a potential of 2600 volts, the filament transformer being the customary 10 volt affair. Inductances are removable; the ones shown are at present used for 30,000 kc. work. The circuit is shown in Fig. 1, which is the conventional push-pull method.

However, the auxiliary cabinet is for purpose of holding an animal within the field and without strapping him and causing discomfort. The condenser plates are of sheet metal and insulated, the schematic diagram appearing in Fig. 2.

When a rabbit is placed in this pen and the power applied, he fails to react immediately; but as he "warms" up to the situation the veins in the ears show dilation, and in 20 minutes his temperature starts to rise and keeps going. It is only at a much later date that burns are noticed.

From the foregoing it would hardly seem advisable to subject a human to any lengthy treatment, at least, for the present; but the value of high frequency in various forms have proved essential in many cases. Secondly, the natural heavy perspiration that is produced is essential in one degree in opening the pores (some of which have probably never been opened before) and allowing the natural poisons of the body to escape more quickly than could have been accomplished with potent medicines.

The majority of high frequency apparatus produces this result. In the case of pulsating currents, which produce a series of reactions or "jumps" within the muscular system, this often is very beneficial in activating those muscles which could not have been manipulated manually without a great degree of pain.

Probably the first reaction noticed with reference to the ultra high frequencies were from early experimenters while conducting transmission tests with high power. It was noticed that the body temperature tended to rise at various times, due entirely to their proximity to a powerful oscillator. Science has been looking for a long time for some method to raise the temperature in the human body *quickly*. Ultra high frequency oscillators have solved that problem but have brought many other matters to light. Secondly, to produce the desired results the patient must be enclosed between two or more condenser plates of large proportions. The oscillator must be of high power, capable of delivering 500 watts output and the patient must be wrapped in a suitable covering to prevent his contact with any portion of the charged plates.

Various research workers who have volunteered their services as "trial patients" have received a goodly portion of burns as compensation, and are quite satisfied that ultra high frequency currents are "hot stuff!" Where heat is concerned, one could get equivalent results by setting on a hot steam radiator. The writer knows this to be true, and has experienced many a painful burn which he would have been willing to trade for a shock from a quarter kw. closed-core transformer.*

It could hardly seem probable that burns would prevail where the body is not in contact with any metallic object; but an examination under darkness of a large condenser tapped from the oscillator, will disclose a bluish aurora between the two elements, which is similar to that noted in Tesla coil experiments. Within this bluish discharge there can be many components not yet identified, which have a tendency to attack tender portions of the body and produce an internal burn, which does not come to the surface for identification until some days have passed.

It is open to discussion whether one-half of the electrical apparatus in present use actually cures the patient of pains, or whether it is the psychology of the matter in which the patient is surrounded by many cabinets of mysteries, of which he knows absolutely nothing. What cures one man does not seem to cure another with the identical complaint.

---

*Yes—but it is generally understood we believe, that the high frequency currents or field causes heat to be developed inside the body; even in organs such as the liver, etc., without heating up the whole body, which is not the case with direct application of heat as when seated on a hot radiator, etc.—Editor.

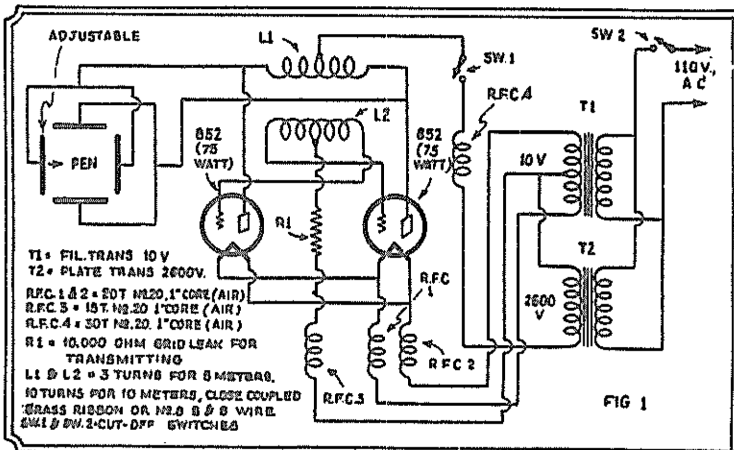
ADJUSTABLE
L1
SW 2
SW.1
110V. A C
R.F.C.4
T1
L2
852 (75 WATT)
852 (75 WATT)
PEN
10 V
R1
T2
R.F.C 1
2600 V
R F C 2
R.F.C.3
T1 = FIL.TRANS 10 V
T2 = PLATE TRANS 2600V.
R.F.C. 1 & 2 = 20T NO.20, 1" CORE (AIR)
R.F.C. 3 = 15T. NO.20 1" CORE (AIR)
R.F.C. 4 = 30T NO.20. 1" CORE (AIR)
R1 = 10,000 OHM GRID LEAK FOR TRANSMITTING
L1 & L2 = 3 TURNS FOR 5 METERS.
10 TURNS FOR 10 METERS, CLOSE COUPLED
BRASS RIBBON OR NO.8 B & S WIRE
SW.1 & SW.2-CUT-OFF SWITCHES

FIG 1

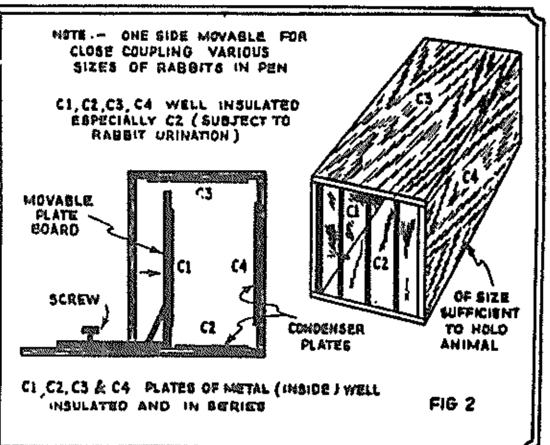
NOTE.— ONE SIDE MOVABLE FOR CLOSE COUPLING VARIOUS SIZES OF RABBITS IN PEN
C1, C2, C3, C4 WELL INSULATED ESPECIALLY C2 (SUBJECT TO RABBIT URINATION)
MOVABLE PLATE BOARD
C3
C1
C4
SCREW
C2
CONDENSER PLATES
C3
C4
C1
C2
OF SIZE SUFFICIENT TO HOLD ANIMAL
C1, C2, C3 & C4 PLATES OF METAL (INSIDE) WELL INSULATED AND IN SERIES

FIG 2

The drawings above provide details of the circuit used by Mr. West in building his ultra short-wave physiological apparatus, together with details of the apparatus itself.

reprinted by Lindsay Publications, Box 12, Bradley IL 60915-0012

Short Waves in Medicine

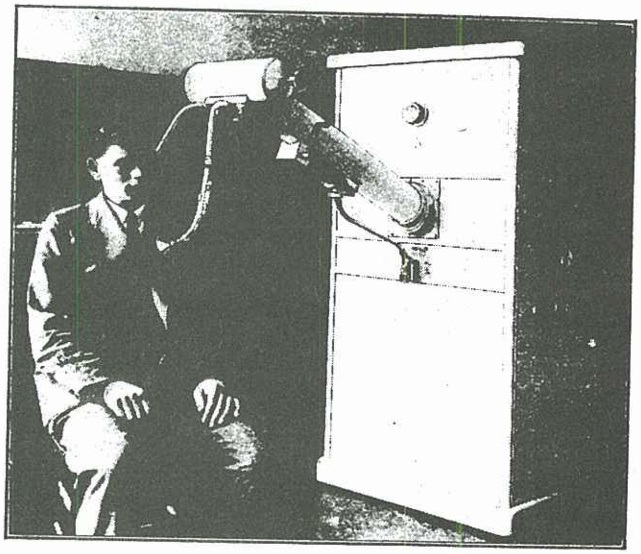

ULTRA SHORT WAVES IN *MEDICINE*

By Dr. FRITZ NOACK
(Berlin)

PROFESSOR ESAU, of the University of Jena, the well-known pioneer in the field of ultra short waves, was the one who nearly two years ago pointed out that very evidently the ultra short wave exercised beneficial medicinal effects.

For further investigation of this effect, small mice were put between the condenser plates of the sender; and Lo! after some time the mice fell victims to a sort of tetanus, which eventually led to death. The result of this fact was that Prof. Esau resolved to commission a physician to make a further study of the physiological effect of these ultra short waves. Naturally the serious investigation of medicinal effects requires experience in physical measurements, likewise an apparatus operating in a constant and easily controlled manner.

Siemens and Halske Interested

It is gratifying that a manufacturer which already possesses very great experience in constructing electro-medical apparatus and which also has the requisite experience in physical measurements, the prominent firms of Siemens and Halske took up the matter. Now this company has manufactured a first ultra-short wave *radiation apparatus* and recently exhibited it to a picked group of scientists and to the press.

The pictures show the new set which, of course, has to be considered as the first of an evolutionary series. There will still be necessary all sorts of experiments to produce the set in such form that it will correspond to all the needs of the medical profession. The present set will, for the time being, be used for the further study of the physiological effect of ultra short waves. It will be the problem of the doctors to investigate the fields of use and the conditions under which the set may be used.

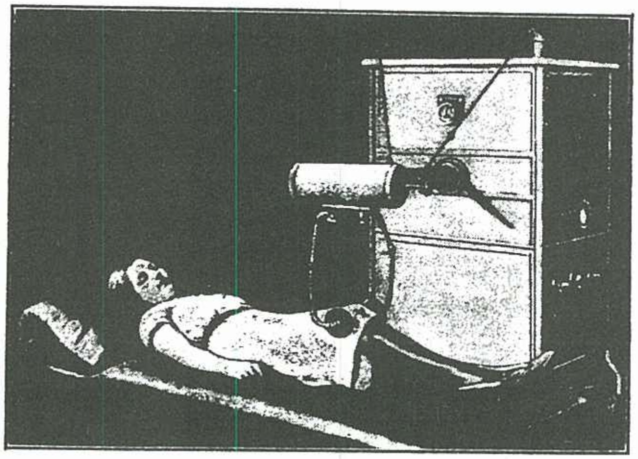

One of the "fair sex" receiving a treatment through the knee by ultra-short waves. Several treatments are given for the average ailment.

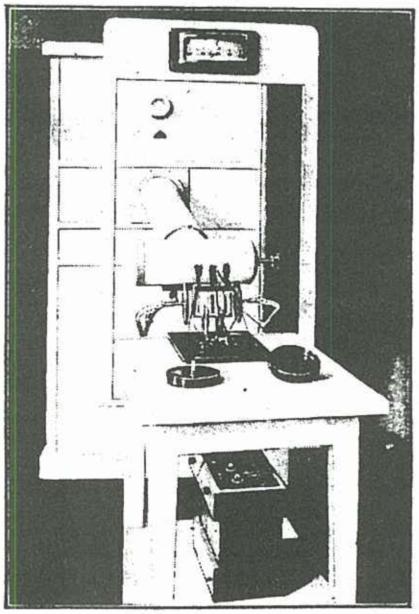

The ultra-short wave medical apparatus with its electrode arm adjusted to permit the waves passing through a solution for experiment.

Arrangement of U. S. W. Generator

In the large shielded box are the real ultra short wave generator—(actually resembling a radio transmitter) and the parts needed for its operation. The generator can work on two waves, four and eight meters. The wavelength is adjusted by inserting the proper tuning coils. The arm which projects out of the apparatus has inside it two lead wires which convey the oscillations outward; they are led to the two electrodes, which are placed at the ends of the two visible supports.

In the cross-beam at the end of the arm is a "tuning" device, which tunes the electrode oscillatory circuit exactly to the wavelength of the generator. The exact tuning can be read on a meter, which is above the arm on the box; this shows the direct plate current of the transmitter, which, with correct tuning of the electrode circuit, adjusts to a minimum; the value then indicated by the meter gives, after calibration of the a value for the electrode energy.

To be able to adjust the electrodes, as is necessary from one case to another, the electrode arms are set in ball-joints. To ascertain whether oscillations are actually present in the electrode oscillatory circuit, there is on one electrode arm a little *glow lamp* (Neon tube), which lights up in the presence of oscillations.

Electrodes Swing on Giant Arm

The arm projecting out of the apparatus can be swung in all directions, for the better use of the whole unit. To accomplish this, it is fixed to the box in a universal joint. The end cross-piece of the arm can furthermore be turned, so that the electrodes can assume any desired direction. The right side wall of the box contains the main current switch, as well as a resistor with which the filaments of the tubes are regulated. The apparatus is in fact constructed just like a vacuum tube radio transmitter. The electrodes may take various forms; in accordance with the most recent experiences, they are so shaped that they do not have to be put firmly against the body, but exercise their effect even in the presence of a layer of air between body and electrode. At the same time, the remarkable fact may be determined, that the thickness of the layer of air and the "depth effect" are dependent on each other.

Advantages of Ultra Short Waves

According to all experiences thus far, the medicinal ultra-short wave radiator is a sort of *diathermal* apparatus; therefore it serves to conduct to the part of the body, or the organ to be treated, electric wave energy, which is converted into heat and thereby exercises a healing effect. As compared with the usual diathermal sets (which operate on a wavelength of about 600 meters, using leaden electrodes, which have to be put right on the body), the *ultra-short wave set* offers a considerably better control of the direction in which the heat is to be applied, so that one can better reach any desired organ than was possible before. Furthermore, by changing the wavelength, one can also adjust for a definite deep-lying stratum. Unpleasant accompanying phenomena, which result with the usual diathermal sets, where the electrodes do not rest firmly on the body, do not occur in the case of the ultra short wave radiator. It is especially important to note that one can now, for the first time, perform a "therapy" or treatment directly on the head. The operation of the set is very simple.

Whether, besides the warming of the different strata in the body, still other biological effects occur or can be accomplished, with the new ultra short wave radiator, is, at the present time, still a matter of question.

HUMAN BEINGS *as* ANTENNAS

By DR. ERWIN SCHLIEPHAKE, M.D.
Of the Jena University Medical Clinic

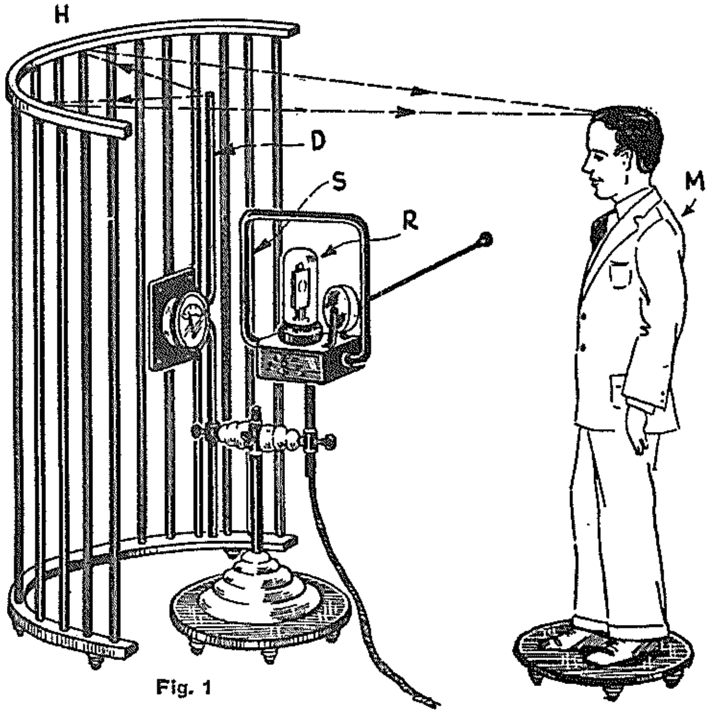

Fig. 1

Arrangement of the apparatus used by Dr. Schliephake in measuring and observing the effects of short waves on the human body, absorption, et cetera. H, parabolic reflector formed of wires; D, antenna, charged inductively by the coil S; vacuum tube at R, and beside it wave changing condenser; subject at M.

If human beings remain in the vicinity of powerful radio transmitters, they are exposed to a very strong electromagnetic field. One must therefore ask what significance these radiations have for the internal organs, and how the body reacts to them. This question is justified by the fact that all other kinds of electromagnetic rays influence the human body more or less strongly.

The character, however, of the reaction is different, according to the wavelengths. Radium rays and X-rays, for example, because of the extraordinary shortness of the waves and the high effectiveness of energy under these conditions, exercise a very strong effect on the structure of atoms; so that serious injuries to the organism can arise (for instance, under certain conditions, cancer). Less serious are the effects of light rays which, however, in the ultraviolet part of the spectrum, can still occasion serious injuries; while the visible light which we perceive is of too low a frequency to injure healthy persons. There are, however, subjects, made sensitive by illness, who suffer serious affections of the skin from strong illumination. Above the optical spectrum, there follow the infra-red and heat waves, whose effect on atoms is relatively slight; in the case of still longer waves only molecular effects are to be expected. Here we are entering the range of the Hertzian waves, electric waves in the stricter sense. Of these, the shortest are about half a millimeter (.02-inch) long, and therefore just above the heat radiations.

From direct analogy, it is to be assumed that the shortest of these waves must be absorbed by human bodies, as in the case of heat rays, and this absorption can be actually demonstrated.

DO YOU KNOW —

¶ What effect short waves have on the hair and scalp in general?

¶ Whether the current in the exciting antenna increases or decreases when a person stands in its field? Why?

¶ If the person stoops will this affect the reading of the antenna current meter?

¶ Whether 'cold" or "heat" is produced inside the body by an overdose of the high frequency waves?

¶ The effect of short waves on the nervous system?

Absorption of Radio by the Body

With specially-built transmitters, especially such as those which have been described by Kohl, electric waves only a few centimeters in length can be produced; these waves can be reflected by mirrors, exactly like light rays. If one puts the antenna in the focus of a concave mirror, then there is formed a parallel beam of radiation which can be concentrated again in another concave mirror. By means of a lens placed in the course of the radiation, the ray is made very sharp. If a person steps into this path, reception ceases. Since no reflection by the human body or only a very slight one is demonstrable, absorption of the radiation must have occurred; this effect is, moreover, demonstrable in the case of other organic substances or water. Unfortunately, not much can yet be said about the physiological effects of these radiations; the power of the transmitters of extremely short waves is still too low to produce perceptible changes.

It is otherwise with the wavelengths from three meters up, to the production of which powers of several kilowatts can be applied. Here the above described optical phenomena cannot be so well demonstrated, because the diffraction is much greater; but absorption by the human body can be very well shown.

When an antenna is inductively coupled to a 3-meter transmitter, the oscillations can be indicated by a detector even at a considerable distance away. If however, a person puts himself in the place of the antenna, the detector responds much more weakly, although the power consumed by the transmitter remains the same. Accordingly, a part of the power must have been used up in the body. The same phenomenon can be demonstrated as follows: a closed or open oscillation circuit is inductively coupled with the transmitter (see Fig. 1). The ammeter in the circuit shows a definite current strength. If a person places himself on an insulating stool beside it, the current in the oscillation circuit is reduced. This withdrawal of energy, however, depends on the length of the body; for if the subject stoops, or changes to a sitting position, the current in the other circuit increases (Fig. 2). Therefore, it appears as though, by the tuning of the subject, *to about half the wavelength, the power transmitted to him becomes much greater.*

Concentrating Power by Means of a Reflector

By means of a large concave mirror one can also collect the transmitted power to a focus. Such a mirror need consist only of parallel wires stretched between two wooden frames. Its height must be equal to the wavelength, the opening one and one-half times the wavelength. It is best to use elliptical reflectors, with the transmitting antenna at one focus; then the reflected radiation is at the other focus of the ellipse. With a "dipole" ("Hertzian" antenna) containing an ammeter, especially strong concentration of energy at this point can be demonstrated. Here a lessening of the current in the dipole is instantly shown if a person steps into the vicinity. Since the human body is to be regarded essentially as an electrolytic system, with regard to the electric wave, and I have tried to demonstrate the effect in the following manner in a model experiment.

A glass tube, of half the wave's length, was filled with an 0.5% sodium chloride (salt) solution, to which gelatine was added to prevent convection. In this jelly the temperature was measured, at different places, by thermo-elements. It was shown that the heating was greatest in the middle and least at the ends; being half as strong at the quarter points of entire length as it was in the middle. Since the maximum strength of the current is at the middle, the greatest heating is therefore connected with this.

Physiological Effect of Short Waves

Especially noteworthy, also, was a feeling of vibration, which was particularly evident if the hand was raised in front. We could establish this sensation only at our transmitter, which is operated with 50-cycle alternating current. There must be, therefore, a direct influence on the nervous system. Here too we have, therefore, another proof that the ultra-short radio waves exercised an effect on the nerves.

These vibrations are also felt if the hand is placed in a condenser's field. Many persons who remain close to the transmitter also experience remarkable sensations on the head, near the roots of the hair; these are like a peculiar prickling, the hair likewise standing up a little. In many subjects we could also observe slight increases in bodily temperature, which however did not exceed 0.5 of a degree, Centigrade (0.9 degree Fahrenheit). Since the body contains extremely fine regulators, by which the temperature is always kept constant, and since also the amounts of energy which can be conveyed into the body, even by powerful radio transmitters, give (when translated into heat units) only a relatively low number of calories, this increase in heat is probably not attributable to the received energy alone. As is to be shown later, nervous effects may play an important part in this.

The effect on the nervous system is

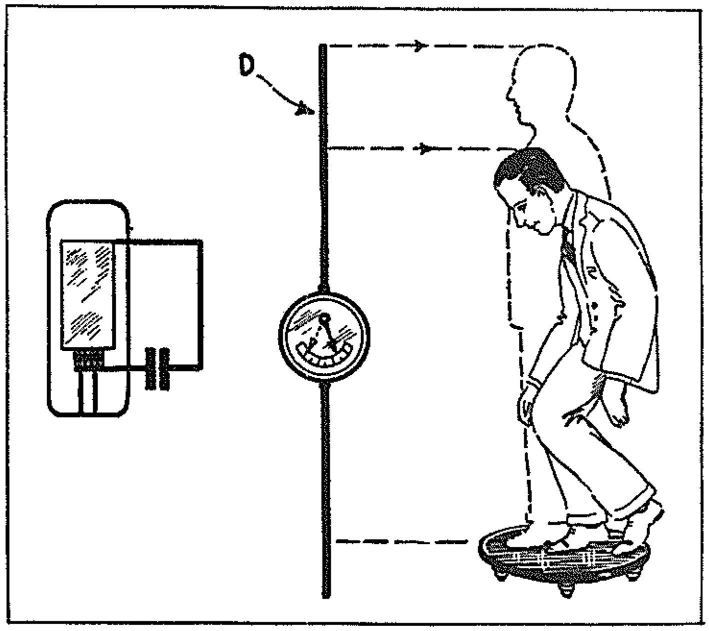

Fig. 2. One of the most interesting effects of placing a human being in the field of a short wave (high frequency) oscillator is that of absorption. The antenna "D" is excited by the power tube and loop circuit at the left. Do you think the antenna current is increased or decreased by bending the body?

Human Beings As Antennas

By Dr. Erwin Schliephake, M.D.

plainly felt by all persons who work a fairly long time with short waves. The sensations are different with the individuals; there are persons who are relatively insensitive, and others who very soon complain about the disturbances of their health. Usually there is first an increasing sleepiness; they are very tired by day, but at night they sleep badly. Several times a night they start out of their sleep, and they are tired and sleepy in the morning.

These phenomena increase more and more. Often there are also intensive headaches, particularly covering the back of the head. Many persons complain of digestive disturbances and pressure on the stomach. Most of them are furthermore easily excited and irritated, being inclined to complaining and to violence. This increased excitability of the nervous system can also be shown by electrical tests of the nerves.

Effects of the Electrostatic Field

It is much easier to study these changes by using the condenser field. Here it is not the electromagnetic wave which is used, but the electrostatic field, which always accompanies it. Here, however, the effect may be much more strongly concentrated. I have generally used plates four inches in diameter, between which the parts of the body in question were treated. The field between these plates suffices to heat 100 cc. of a 0.5% sodium chloride (salt) solution 5 degrees C. (9° F.) in one minute. If parts of the human body are introduced into this field, and the blood is then taken from some part of the body (for instance, from the earlobe) on investigation obvious changes are found.

The number of red corpuscles per cubic millimeter is very greatly increased; i. e., if they formerly amounted to 5 million, the number has risen to 6 million. The same is true of the haemoglobin (a constituent of the blood) and the white corpuscles, the number of which likewise increases. But this is not always true to the same extent, depending upon the strength of the radiation and the place treated. One frequently sees instead a lessening of these amounts, which would therefore correspond to a thinning of the blood. Likewise, the albumens of the blood undergo certain changes which I cannot discuss here.

In the case of these strong effects it is to be assumed that the tissues of the body also undergo changes; which, however, cannot be directly proved in a subject.

Results of Experiments on Animals

In experiments with animals, on the contrary, such changes are plainly recognizable. They occur particularly strong in projecting parts of the animals, such as in the ears and tips of the tails; since there the electric field is distorted. Very often, one sees, some time after the treatment, that the parts in question are dying and falling off. In the case of a rabbit whose leg had been too strongly exposed to the electrostatic field, I noticed a complete disintegration of the flesh in a ring-shaped region, so that only the bone remained; but then, after a while, that also fell off.

That the nervous sytem of the animals is also influenced is shown by the fact that many of them shudder on the switching on or off of the field, at a time when there can be no question of a strong heating effect.

Also very interesting are the disturbances of the internal heat regulation, which one can produce in animals. If the region of the neck and the back of the head of a rabbit are exposed to a limited capacitative field and the temperature is afterward measured, one can soon note a permanently increased temperature, which often lasts for some weeks. If a very powerful dose of "irradiation" is administered the opposite can occur: the bodily heat falls more and more, often below 35 degrees C. (95° F.), until these low temperatures are no longer compatible with life. At the same time, it is interesting to observe that almost all animals which have had such disturbances of their heat regulation after a few weeks developed inflammation of the lungs and pleurisy, afterward dying. It seems as though, by the disturbance of the heat regulation, the resistance of the animals to disease had suffered greatly; so that in this path throughout the central nervous system there was created a special susceptibility to colds. On investigating the spinal marrow of such animals microscopically, serious injuries to the nerve cells have been observed.

Dangers and Beneficial Possibilities

After these experiences, I have not dared to expose entire human beings to a condenser field and in this way produce artificial rises in temperature. The responsibility seemed too great. At the same time, effective heating of the body can be accomplished equally well in other ways; such as with the well-known Apostoli "condenser bed," which can be connected to any diathermal apparatus. With the method previously described, only serious dangers for the patients treated would have been conjured up, without the possibility of producing a fundamentally new effect.

On the other hand, the disturbances of the physical health, which we could observe in the field of free radiation of powerful transmitters, have never been serious. After a period of recovery of a few weeks, with no irradiation, all effects have been observed to vanish. For four years now, I have almost daily worked for several hours at a transmitter with 1½ kw. plate dissipation; and the effects, often very unpleasant, have always gone back to normal on stopping the work.

In these things the wavelength is also certainly of importance, and in fact we have the impression that the disturbances to health became stronger as the wave was shortened. Anyway, the unpleasant sensations appear much quicker with a three-meter wave than with longer ones.

From all these experiences, it is at any rate clear that treatment with electric waves can in no way be regarded as always harmless for the human body. It is plain that their incorrect use can cause serious injuries to health. Certainly such injuries are to be expected only when the frequencies are very high; that is, with ultra-short waves; even then there is nothing to fear except with fairly high transmitter power.

On the other hand, with proper use, the short electric waves seem to be a valuable means of treatment. According to our experiments to date, with bacteria cultures and infected animals, the germs of disease can be killed. There is the added point that certain defensive processes are stimulated in the body. I have also already repeatedly treated human beings; and, in about a hundred cases, I have been able to attain an extraordinarily quick cure of suppuration (pus formation).

Ills Treated by Short Waves

● IN a dispatch to the N. Y. Times, a report from London states that cures for various ailments by the use of short radio waves were claimed by Dr. Erwin Schliephake, a German physician and scientist.

Writing in the *British Medical Journal* he described how he succeeded in treating deep-seated abscess in the human body by passing ultra-short wireless waves through the patient, who was not in immediate contact with any instrument. He found, he said, that various tissues exhibit different degrees of conductivity in the presence of these waves.

Dr. Schliephake declared he has used waves to treat pulmonary abscesses after pneumonia, in pleural empyema, pneumonic tuberculosis, in certain forms of peritonitis, in migraine and acute tonsilitis.

Dr. Willis R. Whitney, research director of the General Electric Company, revealed in April, 1930, that he had developed a radio type of apparatus for killing bacteria in the body.

From THEORY AND CALCULATION OF TRANSIENT PHENOMENON &
ELECTRIC OSCILLATIONS by Charles Proteus Steinmetz, 1920

CHAPTER V.

DISTRIBUTED SERIES CAPACITY.

43. The capacity of a transmission line, cable, or high-potential transformer coil is shunted capacity, that is, capacity from conductor to ground, or from conductor to return conductor, or shunting across a section of the conductor, as from turn to turn or layer to layer of a transformer coil.

In some circuits, in addition to this shunted capacity, distributed series capacity also exists, that is, the circuit is broken at frequent and regular intervals by gaps filled with a dielectric or insulator, as air, and the two faces of the conductor ends thus constitute a condenser in series with the circuit. Where the elements of the circuit are short enough so as to be represented, approximately, as conductor differentials, the circuit constitutes a circuit with distributed series capacity.

An illustration of such a circuit is afforded by the so-called "multi-gap lightning arrester," as shown diagrammatically in Fig. 90, which consists of a large number of metal cylinders p, q . . . , with small spark gaps between the cylinders, connected between line L and ground G. This arrangement, Fig. 90, can be represented diagrammatically by Fig. 91. Each cylinder has a capacity C_0 against ground, a capacity C against the adjacent cylinder, a resistance r,— usually very small,— and an inductance L.

The series of insulator discs of a high voltage suspension — or strain — insulator also forms such a circuit.

If such a series of n equal capacities or spark gaps is connected across a constant supply voltage e_0, each gap has a voltage $e = \frac{e_0}{n}$. If, however, the supply voltage is alternating, the voltage does not divide uniformly between the gaps, but the potential difference is the greater, that is, the potential gradient steeper the nearer the gap is to the line L, and this distribution of potential becomes the more non-uniform the higher the frequency; that is, the greater the charging current of the capacity of the cylinder against the ground. The charging currents against ground, of all

the cylinders from q to the ground G, Figs. 90 and 91, must pass the gap between the adjacent cylinders p and q; that is, the charging current of the condenser represented by two adjacent cylinders p and q is the sum of all the charging currents from q to G; and as the potential difference between the two cylinders p and q is proportional to the charging current of the condenser formed by these two cylinders, C, this potential difference increases towards L, being, at each point proportional to the vector sum of all the charging currents, against ground, of all the cylinders between this point and ground.

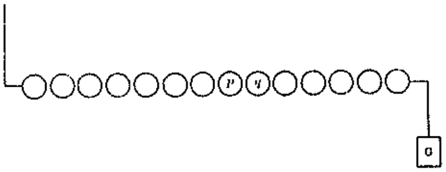

Fig. 90. Multi-gap lightning arrester.

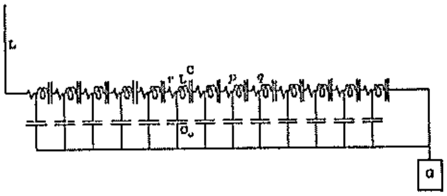

Fig. 91. Equivalent circuit of a multi-gap lightning arrester.

The higher the frequency, the more non-uniform is the potential gradient along the circuit and the lower is the total supply voltage required to bring the maximum potential gradient, near the line L, above the disruptive voltage, that is, to initiate the discharge. Thus such a multigap structure is discriminating regarding frequency; that is, the discharge voltage with increasing frequency, does not remain constant, but decreases with increase of frequency, when the frequency becomes sufficiently high to give appreciable charging currents. Hence high frequency oscillations discharge over such a structure at lower voltage than machine frequencies.

For a further discussion of the feature which makes such a multigap structure useful for lightning protection, see A. I. E. E. Transactions, 1906, pp. 431, 448, 1907, p. 425, etc.

44. Such circuits with distributed series capacity are of great interest in that it is probable that lightning flashes in the clouds are discharges in such circuits. From the distance traversed by lightning flashes in the clouds, their character, and the disruptive strength of air, it appears certain that no potential difference can exist in the clouds of such magnitude as to cause a disruptive discharge across a mile or more of space. It is probable that as the result of condensation of moisture, and the lack of uniformity of such condensation, due to the gusty nature of air currents, a non-uniform distribution of potential is produced between the rain drops in the cloud; and when the potential gradient somewhere in space exceeds the disruptive value, an oscillatory discharge starts between the rain drops, and gradually, in a number of successive discharges, traverses the cloud and equalizes the potential gradient. A study of circuits containing distributed series capacity thus leads to an understanding of the phenomena occurring in the thunder cloud during the lightning discharge.*

Only a general outline can be given in the following.

45. In a circuit containing distributed resistance, conductance, inductance, shunt, and series capacity, as the multigap lightning arrester, Fig. 90, represented electrically as a circuit in Fig. 91, let r = the effective resistance per unit length of circuit, or per circuit element, that is, per arrester cylinder; g = the shunt conductance per unit length, representing leakage, brush discharge, electrical radiation, etc.; L = the inductance per unit length of circuit; C = the series capacity per unit length of circuit, or circuit element, that is, capacity between adjacent arrester cylinders, and C_0 = the shunt capacity per unit length of circuit, or circuit element, that is, capacity between arrester cylinder and

ground. If then f = the frequency of impressed e.m.f., the series impedance per unit length of circuit is

$$Z = r + j(x - x_c); \tag{1}$$

the shunt admittance per unit length of circuit is

$$Y = g + jb, \tag{2}$$

where

$$\left.\begin{aligned} x &= 2\pi fL, \\ x_c &= \frac{1}{2\pi fC}, \\ b &= 2\pi fC_0; \end{aligned}\right\} \tag{3}$$

or the absolute values are

$$\left.\begin{aligned} z &= \sqrt{r^2 + (x - x_c)^2} \\ \text{and} \quad y &= \sqrt{g^2 + b^2}. \end{aligned}\right\} \tag{4}$$

If the distance along the circuit from line L towards ground G is denoted by l, the potential difference between point l and ground by $\dot{E}$, and the current at point l by $\dot{I}$, the differential equations of the circuit are *

$$\frac{d\dot{E}}{dl} = Z\dot{I} \tag{5}$$

and

$$\frac{d\dot{I}}{dl} = Y\dot{E}. \tag{6}$$

Differentiating (5) and substituting (6) therein gives

$$\frac{d^2\dot{E}}{dl^2} = YZ\dot{E}. \tag{7}$$

Equation (7) is integrated by

$$\dot{E} = A_1\epsilon^{-al} + A_2\epsilon^{+al}, \tag{8}$$

where

$$a = \sqrt{YZ} = \alpha + j\beta, \tag{9}$$

$$\left.\begin{aligned} \alpha &= \sqrt{\tfrac{1}{2}\{yz + gr - b(x - x_c)\}} \\ \text{and} \quad \beta &= \sqrt{\tfrac{1}{2}\{yz - gr + b(x - x_c)\}}. \end{aligned}\right\} \tag{10}$$

---

* See paper, "Lightning and Lightning Protection," N.E.L.A., 1907. Reprinted and enlarged in "General Lectures on Electrical Engineering," by Author.

* Section III, Chapter II, paragraph 7.

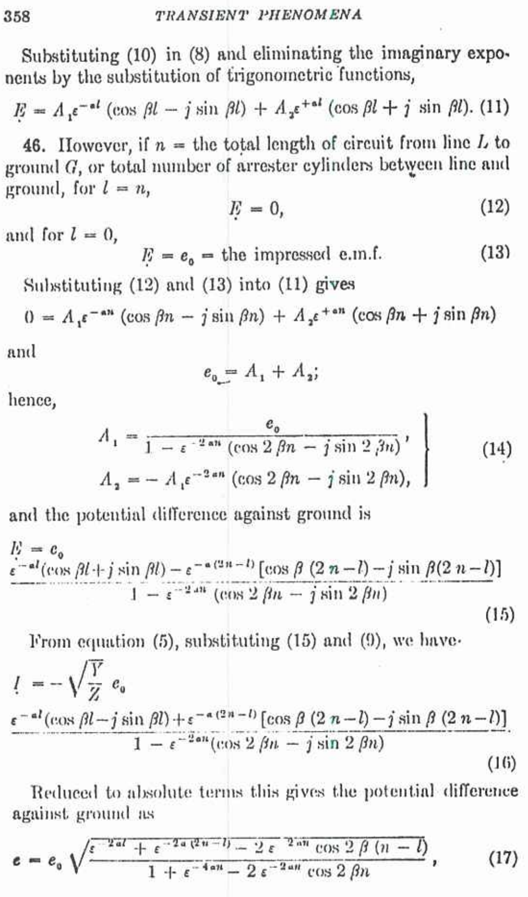

Substituting (10) in (8) and eliminating the imaginary exponents by the substitution of trigonometric functions,

$$E = A_1\epsilon^{-\alpha l}(\cos \beta l - j \sin \beta l) + A_2\epsilon^{+\alpha l}(\cos \beta l + j \sin \beta l). \quad (11)$$

46. However, if n = the total length of circuit from line L to ground G, or total number of arrester cylinders between line and ground, for $l = n$,

$$E = 0, \quad (12)$$

and for $l = 0$,

$$E = e_0 = \text{the impressed e.m.f.} \quad (13)$$

Substituting (12) and (13) into (11) gives

$$0 = A_1\epsilon^{-\alpha n}(\cos \beta n - j \sin \beta n) + A_2\epsilon^{+\alpha n}(\cos \beta n + j \sin \beta n)$$

and

$$e_0 = A_1 + A_2;$$

hence,

$$\left.\begin{aligned} A_1 &= \frac{e_0}{1 - \epsilon^{-2\alpha n}(\cos 2\beta n - j \sin 2\beta n)}, \\ A_2 &= -A_1\epsilon^{-2\alpha n}(\cos 2\beta n - j \sin 2\beta n), \end{aligned}\right\} \quad (14)$$

and the potential difference against ground is

$$E = e_0 \frac{\epsilon^{-\alpha l}(\cos \beta l + j \sin \beta l) - \epsilon^{-\alpha(2n-l)}[\cos \beta(2n-l) - j \sin \beta(2n-l)]}{1 - \epsilon^{-2\alpha n}(\cos 2\beta n - j \sin 2\beta n)} \quad (15)$$

From equation (5), substituting (15) and (9), we have

$$I = -\sqrt{\frac{Y}{Z}}\, e_0 \frac{\epsilon^{-\alpha l}(\cos \beta l - j \sin \beta l) + \epsilon^{-\alpha(2n-l)}[\cos \beta(2n-l) - j \sin \beta(2n-l)]}{1 - \epsilon^{-2\alpha n}(\cos 2\beta n - j \sin 2\beta n)}. \quad (16)$$

Reduced to absolute terms this gives the potential difference against ground as

$$e = e_0 \sqrt{\frac{\epsilon^{-2\alpha l} + \epsilon^{-2\alpha(2n-l)} - 2\epsilon^{-2\alpha n} \cos 2\beta(n-l)}{1 + \epsilon^{-4\alpha n} - 2\epsilon^{-2\alpha n} \cos 2\beta n}}, \quad (17)$$

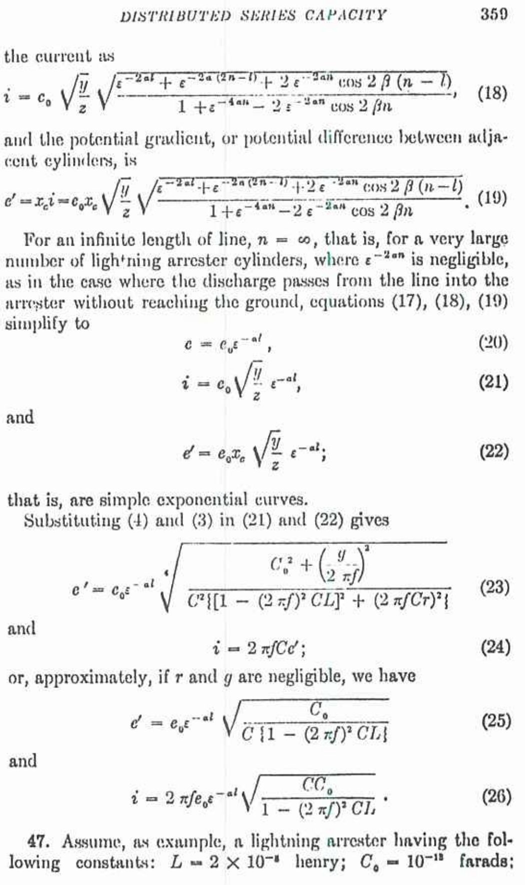

the current as

$$i = e_0 \sqrt{\frac{y}{z}} \sqrt{\frac{\epsilon^{-2\alpha l} + \epsilon^{-2\alpha(2n-l)} + 2\epsilon^{-2\alpha n} \cos 2\beta(n-l)}{1 + \epsilon^{-4\alpha n} - 2\epsilon^{-2\alpha n} \cos 2\beta n}}, \quad (18)$$

and the potential gradient, or potential difference between adjacent cylinders, is

$$e' = x_c i = e_0 x_c \sqrt{\frac{y}{z}} \sqrt{\frac{\epsilon^{-2\alpha l} + \epsilon^{-2\alpha(2n-l)} + 2\epsilon^{-2\alpha n} \cos 2\beta(n-l)}{1 + \epsilon^{-4\alpha n} - 2\epsilon^{-2\alpha n} \cos 2\beta n}}. \quad (19)$$

For an infinite length of line, $n = \infty$, that is, for a very large number of lightning arrester cylinders, where $\epsilon^{-2\alpha n}$ is negligible, as in the case where the discharge passes from the line into the arrester without reaching the ground, equations (17), (18), (19) simplify to

$$e = e_0\epsilon^{-\alpha l}, \quad (20)$$

$$i = e_0\sqrt{\frac{y}{z}}\, \epsilon^{-\alpha l}, \quad (21)$$

and

$$e' = e_0 x_c \sqrt{\frac{y}{z}}\, \epsilon^{-\alpha l}; \quad (22)$$

that is, are simple exponential curves.

Substituting (4) and (3) in (21) and (22) gives

$$e' = e_0\epsilon^{-\alpha l} \sqrt[4]{\frac{C_0^2 + \left(\frac{g}{2\pi f}\right)^2}{C^2\{[1 - (2\pi f)^2 CL]^2 + (2\pi f Cr)^2\}}} \quad (23)$$

and

$$i = 2\pi f C e'; \quad (24)$$

or, approximately, if r and g are negligible, we have

$$e' = e_0\epsilon^{-\alpha l} \sqrt{\frac{C_0}{C\{1 - (2\pi f)^2 CL\}}} \quad (25)$$

and

$$i = 2\pi f e_0\epsilon^{-\alpha l} \sqrt{\frac{CC_0}{1 - (2\pi f)^2 CL}}. \quad (26)$$

47. Assume, as example, a lightning arrester having the following constants: $L = 2 \times 10^{-8}$ henry; $C_0 = 10^{-13}$ farads;

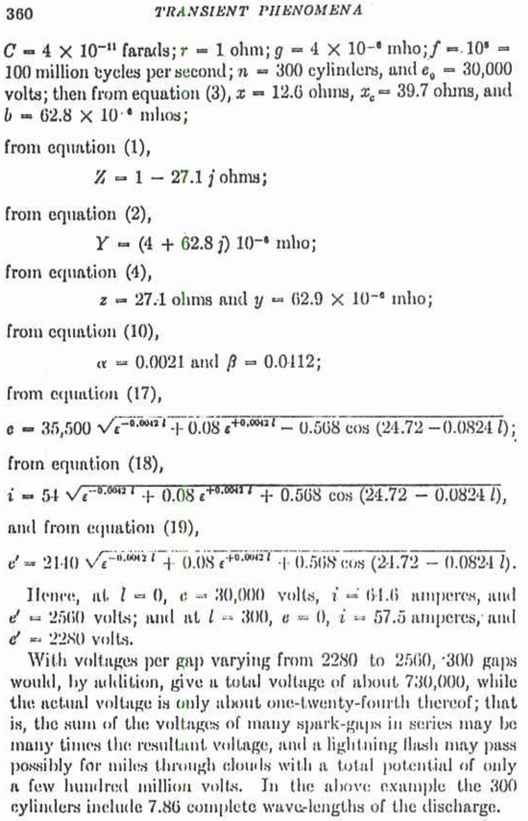

$C = 4 \times 10^{-11}$ farads; $r = 1$ ohm; $g = 4 \times 10^{-6}$ mho; $f = 10^8$ = 100 million cycles per second; $n = 300$ cylinders, and $e_0 = 30{,}000$ volts; then from equation (3), $x = 12.6$ ohms, $x_c = 39.7$ ohms, and $b = 62.8 \times 10^{-6}$ mhos;

from equation (1),

$$Z = 1 - 27.1\,j \text{ ohms};$$

from equation (2),

$$Y = (4 + 62.8\,j)\, 10^{-6} \text{ mho};$$

from equation (4),

$$z = 27.1 \text{ ohms and } y = 62.9 \times 10^{-6} \text{ mho};$$

from equation (10),

$$\alpha = 0.0021 \text{ and } \beta = 0.0412;$$

from equation (17),

$$e = 35{,}500 \sqrt{\epsilon^{-0.0042\,l} + 0.08\,\epsilon^{+0.0042\,l} - 0.568 \cos(24.72 - 0.0824\,l)};$$

from equation (18),

$$i = 54 \sqrt{\epsilon^{-0.0042\,l} + 0.08\,\epsilon^{+0.0042\,l} + 0.568 \cos(24.72 - 0.0824\,l)},$$

and from equation (19),

$$e' = 2140 \sqrt{\epsilon^{-0.0042\,l} + 0.08\,\epsilon^{+0.0042\,l} + 0.568 \cos(24.72 - 0.0824\,l)}.$$

Hence, at $l = 0$, $e = 30{,}000$ volts, $i = 64.6$ amperes, and $e' = 2560$ volts; and at $l = 300$, $e = 0$, $i = 57.5$ amperes, and $e' = 2280$ volts.

With voltages per gap varying from 2280 to 2560, 300 gaps would, by addition, give a total voltage of about 730,000, while the actual voltage is only about one-twenty-fourth thereof; that is, the sum of the voltages of many spark-gaps in series may be many times the resultant voltage, and a lightning flash may pass possibly for miles through clouds with a total potential of only a few hundred million volts. In the above example the 300 cylinders include 7.86 complete wave-lengths of the discharge.

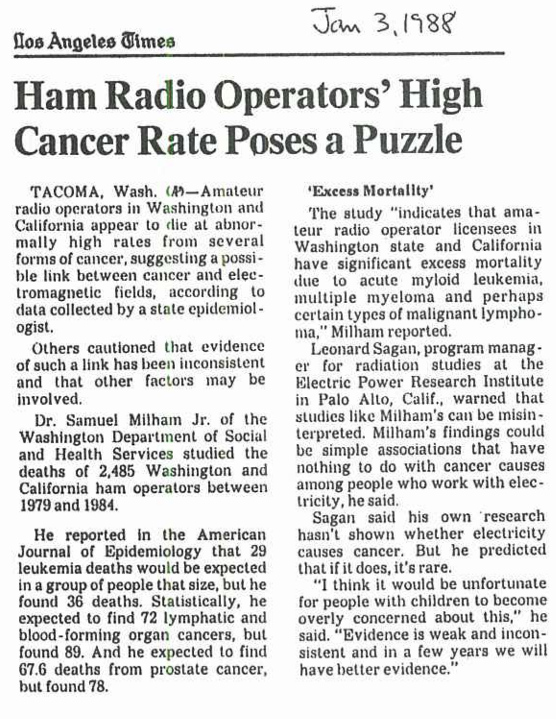

Jan 3, 1988

Los Angeles Times

Ham Radio Operators' High Cancer Rate Poses a Puzzle

TACOMA, Wash. (AP)—Amateur radio operators in Washington and California appear to die at abnormally high rates from several forms of cancer, suggesting a possible link between cancer and electromagnetic fields, according to data collected by a state epidemiologist.

Others cautioned that evidence of such a link has been inconsistent and that other factors may be involved.

Dr. Samuel Milham Jr. of the Washington Department of Social and Health Services studied the deaths of 2,485 Washington and California ham operators between 1979 and 1984.

He reported in the American Journal of Epidemiology that 29 leukemia deaths would be expected in a group of people that size, but he found 36 deaths. Statistically, he expected to find 72 lymphatic and blood-forming organ cancers, but found 89. And he expected to find 67.6 deaths from prostate cancer, but found 78.

'Excess Mortality'

The study "indicates that amateur radio operator licensees in Washington state and California have significant excess mortality due to acute myloid leukemia, multiple myeloma and perhaps certain types of malignant lymphoma," Milham reported.

Leonard Sagan, program manager for radiation studies at the Electric Power Research Institute in Palo Alto, Calif., warned that studies like Milham's can be misinterpreted. Milham's findings could be simple associations that have nothing to do with cancer causes among people who work with electricity, he said.

Sagan said his own research hasn't shown whether electricity causes cancer. But he predicted that if it does, it's rare.

"I think it would be unfortunate for people with children to become overly concerned about this," he said. "Evidence is weak and inconsistent and in a few years we will have better evidence."

Here's something to consider while using a Multi Wave Oscillator. Just "any old" radio frequency broadcast energy may not be desireable. Use this handbook in your detective work to discover Lakhovsky's secret.

THE LAKHOVSKY MUPTIPLE WAVE OSCILLATOR HANDBOOK

Borderland Scien

Is This Device
AN ELECTRONIC
FOUNTAIN OF YOUTH?!?

Is it possible to restore aging or diseased tissue? Earlier this century scientist Georges Lakhovsky conducted researches indicating that living cells can be regenerated by applying special fields of electrical energy.

It has been theorized that every cell in your body has a specific vibratory rate. When the cell ages or becomes diseased its rate changes. The Multiple Wave Oscillator produces a full spectrum of vibratory rates that resonate with cells to restore them to oscillatory equilibrium.

Earlier this century this Multiple Wave Oscillator was shown to neutralize certain types of cancer--but no exact plans are available. This book contains the records of many people who have attempted to reconstruct the original equipment and continue with this fascinating research.

This book describes researches such as the enhancement of plant grown through the amplification of cosmic waves with simple wire coils, and the neutralization of cancerous tissue with radio waves. This was done with cosmic radiations ringing wires coiled around plants, and with refined electrical fields ringing specially designed antennas. Lakhovsky's original work into the dramatic impact of the subtle forces of nature included the effects of sunspot cycles on plant and animal health, the nature of cancer and how it appears in clusters over areas which have soils that reflect cosmic radiations, thereby causing interference patters, and how diseased cells would turn to a state of health when specific natural or electrical fields were applied in a proper manner.

Nikola Tesla provided some of the early electro-therapeutic research and Lakhovsky was introduced to this through Professor D'arsonval. Lakhovsky, a brilliant scientist, combined this knowledge with his original work and produced a device known as the Multiple Wave Oscillator.

The Multiple Wave Oscillator is an ongoing research project of Borderland Sciences Research Foundation, begun in 1963. This book is the published record of our search for Lakhovsky's secrets and the potential for reversing the aging process.

ISBN 0-945685-03-3
51795>
9 780945 685036

BORDERLAND SCIENCES
RESEARCH FOUNDATION
P.O. BOX 429 ★ GARBERVILLE, CA 95440-0429 ★ U.S.A.

Electro-Biology
$17.95

The Rife Machine Report

A History of Rife's Instruments and Frequencies

www.rife.org

www.rife.org

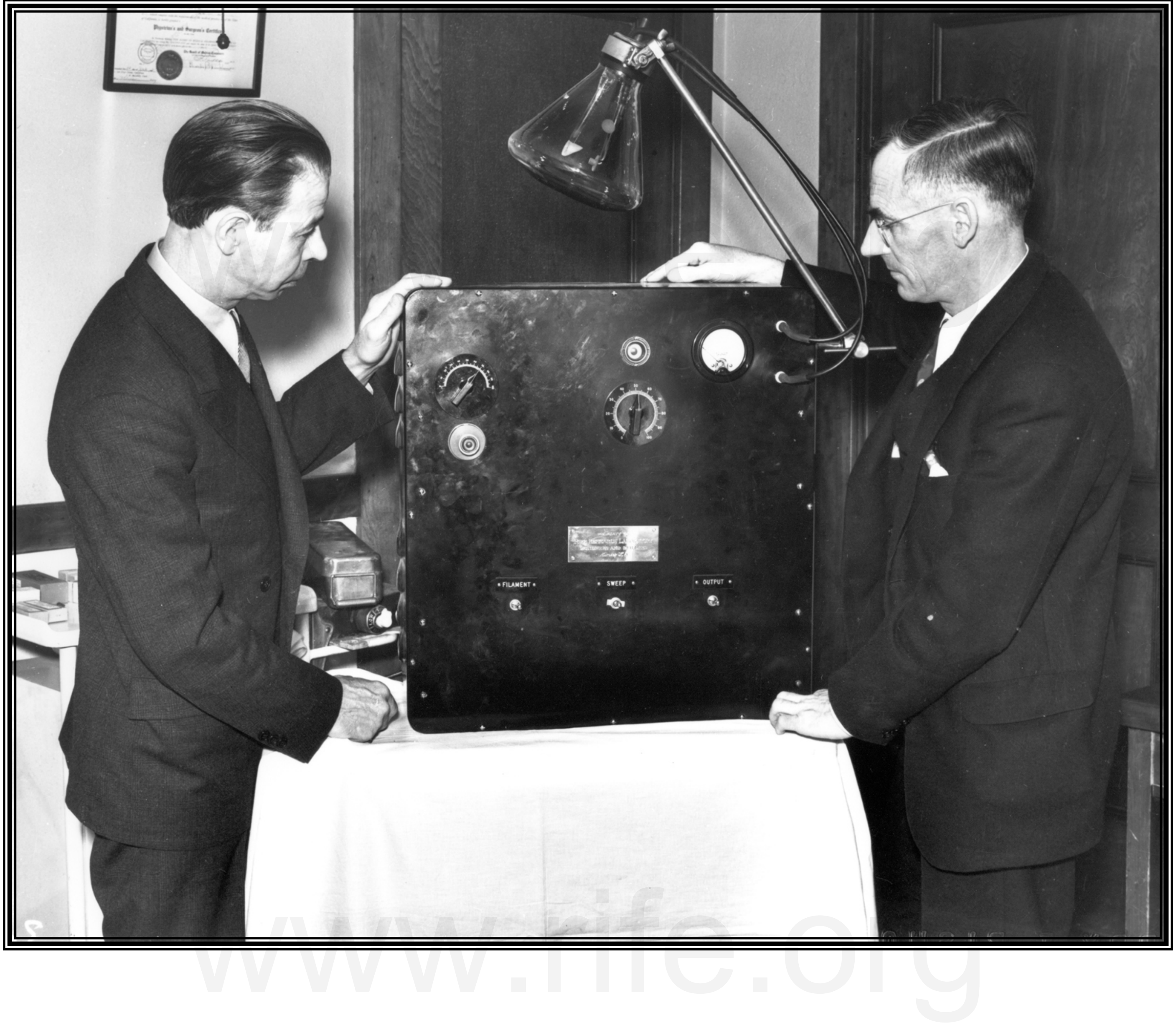
FILAMENT
SWEEP
OUTPUT

www.rife.org

www.rife.org

The Rife Machine Report

A History of Rife's Instruments and Frequencies

Updated 03/01/2013 **Copyright www.rifevideos.com 2003-2013**

www.rife.org

This report from its first writing back in 2003 has been an evolving document. The current report supersedes all previous reports that have been released.

Whenever original Rife instruments, schematics or documents have been found this report has been updated according to the information that was obtained. This report was updated in 2010 because at that time an original 1938-1939 Rife Ray #5 or Beam Ray Clinical Instrument had been found and analyzed. That instrument was built by the original 1938-1939 Beam Ray Corporation. At that time the analyzing of that instrument finally showed where the audio frequency instruments came from which Dr. Rife used in the 1950's. It was updated in 2011 and 2012 when some new information came to light about Dr. Gruner's Rife machine and John Marsh's documents. The last revision of this report was dated 12/01/2012.

This report is being updated now because more of John Marsh's documented information has come to light. Two of his ray tube instruments have been obtained that were built in the 1970's and 1980's. More extensive photos have been taken of those instruments and have been added to this report at this time. Also this report has had a great deal of updating and clarifying in almost all of its chapters. With these updates the report has increased from 153 pages to 224 pages.

In this report we will examine the way Dr. Rife's instruments were built. We will look at the evidence by quoting the sources such as Dr. Rife, John Crane, John Marsh, Dr. Couche, Dr. Lara, Dr. Stafford and Bertrand L. Comparet (Dr. Rife's attorney in the 1939 Beam Ray Corporation Trial, and later John Crane's attorney for Life Labs' trial in 1961). Hopefully anyone who reads this report will have a better understanding about Dr. Rife and the methods he used. Our goal is to try to give people information so that they will know how Dr. Rife's equipment worked so that they will not be misguided by all the misinformation and disinformation that has been published on this subject. The writer of this report has tried to explain the information in layman's terms.

This report, called by the same name, is also online at www.rifevideos.com. All of the quoted documentation is linked in the online report for those who want to study the documents used to back up the quotes used in this report. Also online are more photos of the equipment shown in this report.

We will continue to update this paper when new information is obtained.

The Rife Machine Report table of contents

Chapter 1 - What is a ray tube and how does it work? 4
Chapter 2 - What power levels did Dr. Rife use in his Rife Machines? 9
Chapter 3 - Is it necessary to use a Ray tube to output the frequencies? 12
Chapter 4 - Are Dr. Rife's RF frequencies safe to use? 17
Chapter 5 - Did Dr. Rife use audio frequencies? 20
Chapter 6 - Dr. Rife's 1920 to 1922 Rife Ray #1 Rife Machine 23
Chapter 7 - 1934 Rife Ray #3 Rife Machine used in the 1934 clinic 26
  Dr. Rife's misread lab note frequencies from before 1935 (Chart) 48
Chapter 8 - 1935 Rife Ray #4 Rife Machine 49
  Rife Ray #4 sine wave high RF frequencies (Chart) 57
Chapter 9 - 1938 to 1939 Beam Ray Corporation Clinical Rife Machine 58
  The Beam Ray Clinical instrument sideband sine wave audio frequencies based on a 3.80 MHz carrier (Chart) 90
  Beam Ray instrument sideband sine wave audio frequencies (Chart) 91
  Oscilloscope waveform readings of the original Beam Ray Clinical instrument 93
  Original Beam Ray spectrum analyzer graphs for microorganisms 94-100
  Upper harmonic frequency charts for Microorganisms 101-102
  Photos of the original 1938-1939 Beam Ray Clinical instrument when found 103-105
Chapter 10 - The Gruner schematic of Philip Hoyland's Beam Ray laboratory instrument 106
  The initial Gruner schematic work done 3 years ago 108
  Deciphering Dr. Gruner's Beam Ray instrument schematic 109
  Rebuilding Philip Hoyland's Beam Ray Laboratory instrument 110
  The modulated Audio Frequency pulsing circuit 113
  Mr. Peters' photos of the rebuilt Beam Ray Laboratory instrument 118
  Second machine photos of the rebuilt Beam Ray Laboratory instrument 119-121
  Beam Ray Clinical instrument Schematic 122
Chapter 11 - Aubrey Scoon's Beam Ray replica Rife Machine Re-evaluation 123
  Aubrey Scoon's sideband audio frequencies reconciled to Rife's original high frequency M.O.R.s. (Chart) 134
  Aubrey Scoon's Beam Ray replica spectrum analyzer graphs for microorganisms 137-144
  Photos of the rebuilt Aubrey Scoon Beam Ray Clinical instrument 145
  Aubrey Scoon's Beam Ray Clinical replica schematic 146
  Aubrey Scoon 1939 Rife Beam Ray website report 147-159
  Aubrey Scoon's Beam Ray audio frequencies (Chart) 156
Chapter 12 - Dr. Rife and Verne Thompson's 1950's AZ-58 Beam Ray replica Rife Machine 160
  Original 1950's AZ-58 frequencies used by Dr. Robert P. Stafford M.D. (Chart) 175
  1950's Beam Ray Clinical instrument optimum sideband square wave audio frequencies based on a 4.68 MHz carrier (Chart) 176
  Photos of the rebuilt AZ-58 Beam Ray Clinical instrument 179
  AZ-58 Beam Ray Clinical instrument schematic 180
Chapter 13 - Rife Machine Harmonic Audio Frequency Misunderstanding 181
  AZ-58 M.O.R. audio frequencies square wave harmonics (Chart) 184
Chapter 14 - Life Labs 1950's pad instrument (without ray tube) 186
Chapter 15 - John Marsh's 1970's Beam Ray replica Rife Machine 197
  John Marsh's Beam Ray Clinical instrument sideband square wave audio frequencies based on a 4.122 MHz carrier (Chart) 209
  John Marsh's Beam Ray Clinical instrument higher sideband square wave audio frequencies based on a 4.122 MHz carrier (Chart) 210
Chapter 16 - John Marsh's 1980's Ray Tube Rife Machine 211
Chapter 17 - Summary 219
  Rife's M.O.R. frequencies and audio sideband frequencies (Charts) 224

www.rife.org

Chapter #1

What is a Rife ray tube and how does it work?

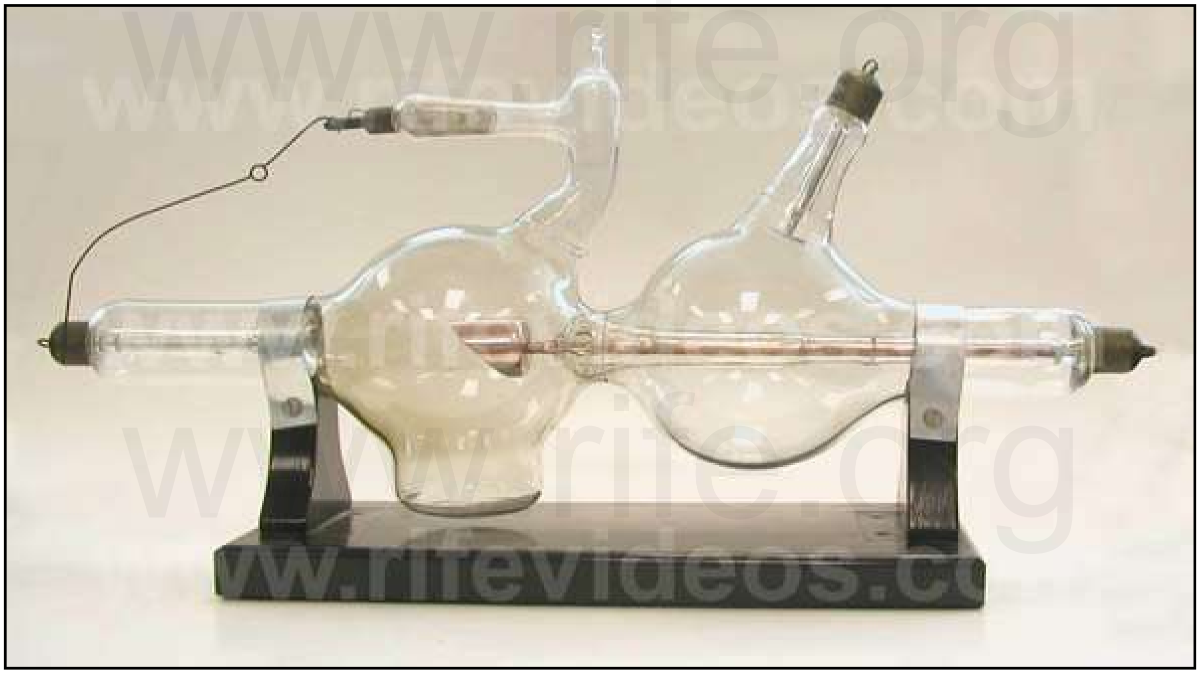

Dr. Rife used a ray tube with his Rife Machines. The photo, shown above, is a picture of the style of double-bubble ray tube which Dr. Rife used for many years in his laboratory. A ray tube was made out of glass, quartz or Pyrex and was filled with a noble gas or a mixture of noble gases.

The next photo, shown below, shows Dr. Rife's double-bubble ray tube in his laboratory right next to his microscope.

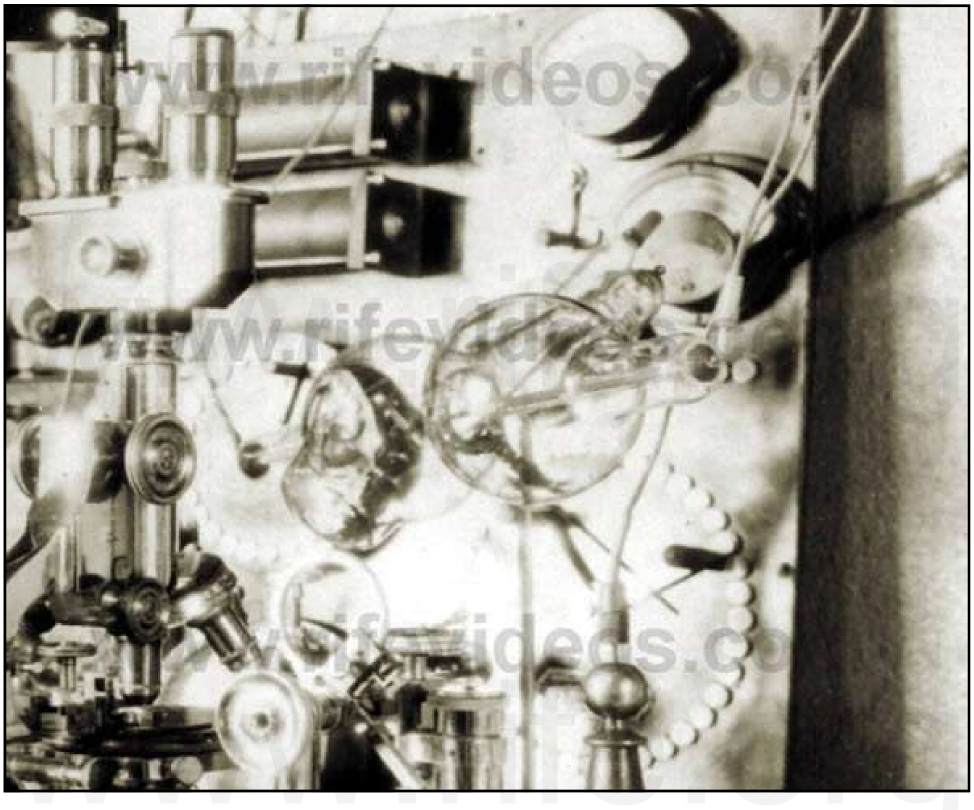

www.rife.org
www.rifevideos.com

Some people believe that there was something magical to the gas mixture that Dr. Rife used in his ray tube. There are also people who claim that they use some proprietary blend of gases that makes their ray tubes work exactly like Dr. Rife's. Others even claim that their gas formula works better than Dr. Rife's did. The truth is that none of this rings true because of what Dr. Rife has said on this subject. Dr. Rife knew more about these gases than most anyone else since he used different gas mixtures over a thirty year time span. The fact is Dr. Rife used many different mixtures of gases but eventually ended up using only helium. To verify this we will read his statement:

www.rife.org

RIFE: *"We have experimented with various inert gases and we found that helium stood up by the bombardment better than any of the other gases. That's why we use it. We don't care about the color or anything of that sort. It stood up better over many more hours of bombardment than the argon and the crypton* [krypton] *and those different gases that we tried."* (John Marsh Collection, Gonin and Siner Papers, Page 25)

From what Dr. Rife said in the above quote it is apparent that he tested many different gases over the years in his laboratory and they all worked. But we find that he decided to use only helium because it lasted longer in the ray tube.

The ray tube was connected to Dr. Rife's Machine by two wires. These wires were connected to two round metal bars that went into the glass tube and they had round disks connected to their ends. One disk was straight and the other one was on a 45 degree angle. This gave it a directional effect towards the patient.

Below is a photo showing the internal electrodes that were built into the ray tube. You can see that one is straight and the other is on a 45 degree angle.

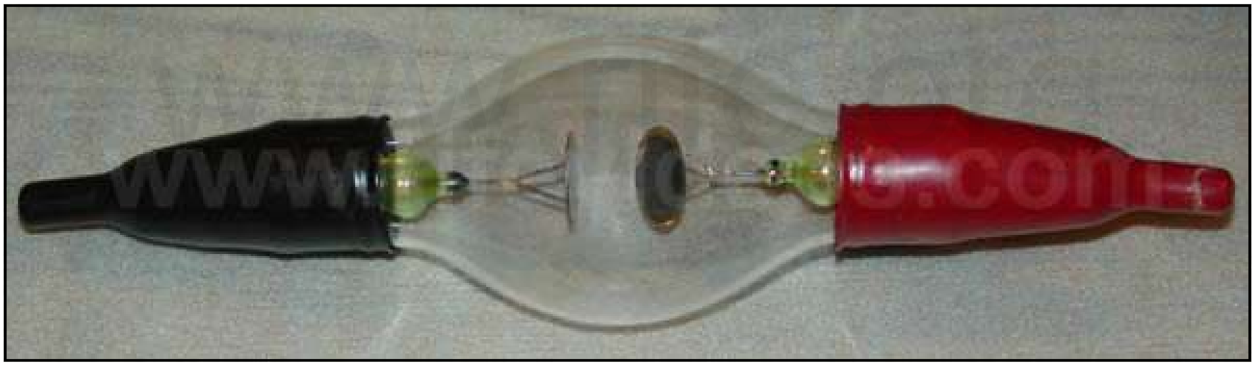

Dr. Rife stated that the ray tube was "a partial directional antenna". Because the scientific technology behind ray tubes had already been perfected, Dr. Rife worked with that technology and only had to make some adjustments for it to work the way he wanted it to in his applications. Bertrand L. Comparet, Dr. Rife's attorney, stated in an interview:

COMPARET: *"Now, the original instrument had a tube, like an X-ray tube. That was the way in which Rife developed it. You see, all the X-ray work necessarily was done with a beam projected from a tube. So, Rife worked on the same basis."* (1970's Bertrand Comparet Interview #32)

Although Dr. Rife used X-ray tubes in his work they did not put out any X-rays. Because the ray tubes used inert gases they did not produce X-rays. They work on the same principle as a neon sign. The gas that is used in a neon sign does not produce any harmful rays of any kind. Inert gases such as helium, argon, krypton, neon and others can safely be use in these ray tubes without producing X-rays. These gases are considered noble gases and there are 18 different types. The ray tube is just an antenna and the noble gas will emit the frequency when the plasma is lit. This makes a ray tube a safe method of delivering the frequencies. If you use a metal antenna and someone just happened to touch it they could receive severe third degree RF (Radio Frequency) burns.

Many people believe that ray tubes are just as efficient as metal antennas, and this may be true. They also believe that the energy emitted from a ray tube will actually travel farther with less loss than a metal antenna. Since there are no actual scientific tests comparing the output of ray tubes to metal antennas it is hard to know for sure if these assumptions are correct. For this report we will accept what is known, not what is un-known. Therefore we will compare ray tubes to metal antennas since they both are designed to emit frequencies.

There are limitations to metal antennas that need to be understood and this has to do with the laws of physics. It is referred to as the "Inverse-Square Law". This law deals with power loss and distance. We will give a simple explanation which should suffice since we are trying to stay in layman's terms and make it easy for the average person to understand this information. When a circuit is properly tuned metal antennas are very efficient. About 100% of the energy that you put into a metal antenna comes out, but only if the impedance is matched correctly.

Dr. Rife’s Rife Ray #3 instrument information which has been obtained from the Rife documents list that about 50 RF watts was input into the ray tube. If we compare this to a metal antenna this means about 50 watts would have passed through and come out of the ray tube. When it comes to metal antennas and the inverse-square law on signal loss this would mean that you would have to divide the 50 watts which come out of the metal antenna by four for every foot that you move away from the antenna. The exact power loss of a ray tube, as stated before, is not known but if a ray tube is equally as efficient as a metal antenna, and we believe it is, then the same laws of physics would also apply to it.

Because no actual scientific tests have been done with ray tubes demonstrating that they are exempt from the inverse-square law then we are left with only one conclusion, this law does apply to ray tubes. Because of this we will use this inverse-square law of power loss for a ray tube. Therefore, with a 50 watt power output at one foot away from the ray tube you would only have 12.5 watts. At two feet you only have 3.125 watts and at 3 feet you only have about .78 of a watt. The laws of physics are important to understand because Dr. Rife and the doctors that used his equipment put the ray tube within a few inches to a few feet of the patient’s body. One of Dr. Rife's 1950's business partners, John Marsh, wrote a paper in which he stated that the Ray tube should be used from 12 to 24 inches from the body:

MARSH: *"A frequency instrument with therapeutic applications which has been developed and successfully tested over a period of year’s works on the principle of stimulating tissue with low energy, low frequency pulsating current. It applies electron transmission at variable frequencies from an applicator source, which consists of either (1) a bare anode and cathode (in direct contact with the body) and constructed from metal for easy transmission flow of electrons, or (2) from an antenna broadcast source at a distance of 12 to 24 inches."* (An Explanation of the JLMSQ-1A frequency instrument and its use. Page 1-2).

Dr. Couche said that he would sometimes touch the body of the patient in the area that needed to be treated. When we discussed this with Dr. Robert P. Stafford M.D., he said that when he treated cancer patients he would put the ray tube within a few inches of the body and treat a 6 inch square area. He would move the ray tube up and down and back and forth so that the whole 6 inch area was treated. He said that he did this because of the way the phanotron (ray tube nickname) ray tube worked. The design of a phanotron ray tube makes it partially directional and concentrates its energy or power into a smaller area. Due to the "Inverse-square law" power loss it is easy to understand why Dr. Stafford, Dr. Couche, Dr. Rife and the other doctors used the ray tube right next to the body. Many

www.rife.org

people have used these RF (Radio Frequency) ray tube instruments and have notice that within several feet they will get a strong reaction but beyond this it drops of very quickly. It is apparent that there was a good reason why the ray tube was used close to the patients body.

We built both the Aubrey Scoon Beam Ray replica ray tube instrument and the 1953 AZ-58 Beam Ray replica ray tube Rife Machine. The 1953 AZ-58 (Rife instrument made by Life Labs) was built from schematics that are on Stan Truman's site, http://www.rife.org, under AZ-58 research information. This AZ-58 instrument is nearly the same as the original Rife Ray #5 or Beam Ray Clinical instrument and schematics can be found on this page.

Both Aubrey Scoon's instrument and the original Beam Ray instrument use sine wave audio frequencies and the 1953 AZ-58 uses square wave audio frequencies. We tested the AZ-58 and Aubrey Scoon's replica for penetration and found that at about 32 inches from the body full penetration of the carrier frequency emitted from the ray tube was lost with our frequency detection device. This test was done in this manner. A B&K frequency counter was used and the antenna was put in a fold of body tissue on the front of the body. The ray tube was put behind the person and move backwards until the carrier frequency could no longer be read through the person's body. John Crane listed the AZ-58 as outputting 14 watts out of the ray tube but we tested it and found it outputs at least 30 to 40 watts. The 1950's Aubrey Scoon instrument puts out about 30 to 40 watts also.

Another test was made using a crystal for testing resonance. The audio frequencies broadcast out of the ray tube from both of these machines could only resonate a crystal designed to test resonance through about two inches of tissue. From the tests made, it takes a carrier frequency of at least 0.125 watts to penetrate all the way through the body. It could take an output of 50 watts from a ray tube to resonate a crystal through 14 inches of tissue. These tests showed that it takes more power to penetrate all the way through the body when modulating an audio frequency on a carrier frequency than when a single un-modulated frequency is used. The tests were done using the AZ-58 replica and the Aubrey Scoon replica Rife machine using a phanotron ray tube outputting about 30 to 40 watts. Another interesting thing worth noting is when we turned the ray tube more than 45 degrees either to the right or the left of center we could not resonate the crystal. Another test showed we could not resonate the crystal at all on the backside of the phanotron ray tube proving what Dr. Rife said:

RIFE: *"The ray tube is a partially directional antenna."*

One interesting fact worth noting is the ray tubes that do not use the internal electrodes, like the Phanotron ray tube, have a higher field strength reading which indicates a greater output. These ray tubes use copper collars or wire wrapping around the ray tube. These ray tubes also last longer because the gas inside the ray tube does not get contaminated. The contamination comes from the internal metal electrodes. Over time the metal from the internal electrodes comes off and slowly contaminates the gas. It will also deposit on the inside of the glass making the ray tube go darker. In the photo below you can see how the ray tube has darkened.

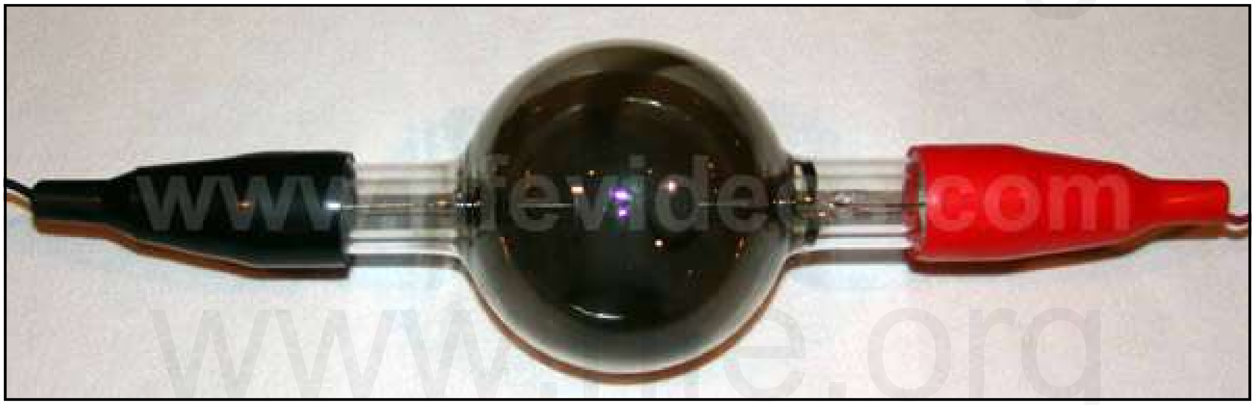

www.rife.org

The next photo, show below, is a picture of Dr. James B. Couche's Rife Ray #5 or Beam Ray Clinical instrument which he purchase from the original 1938-1939 Beam Ray Corporation. The ray tube is Dr. Rife's original double-bubble ray tube which he used for over 30 years in his laboratory. You can see how the ray tube has darkened. This was a common problem which required the ray tubes to be cleaned and re-gassed on a regular basis. A better view of this double-bubble ray tube is shown in the two photos on page 4.

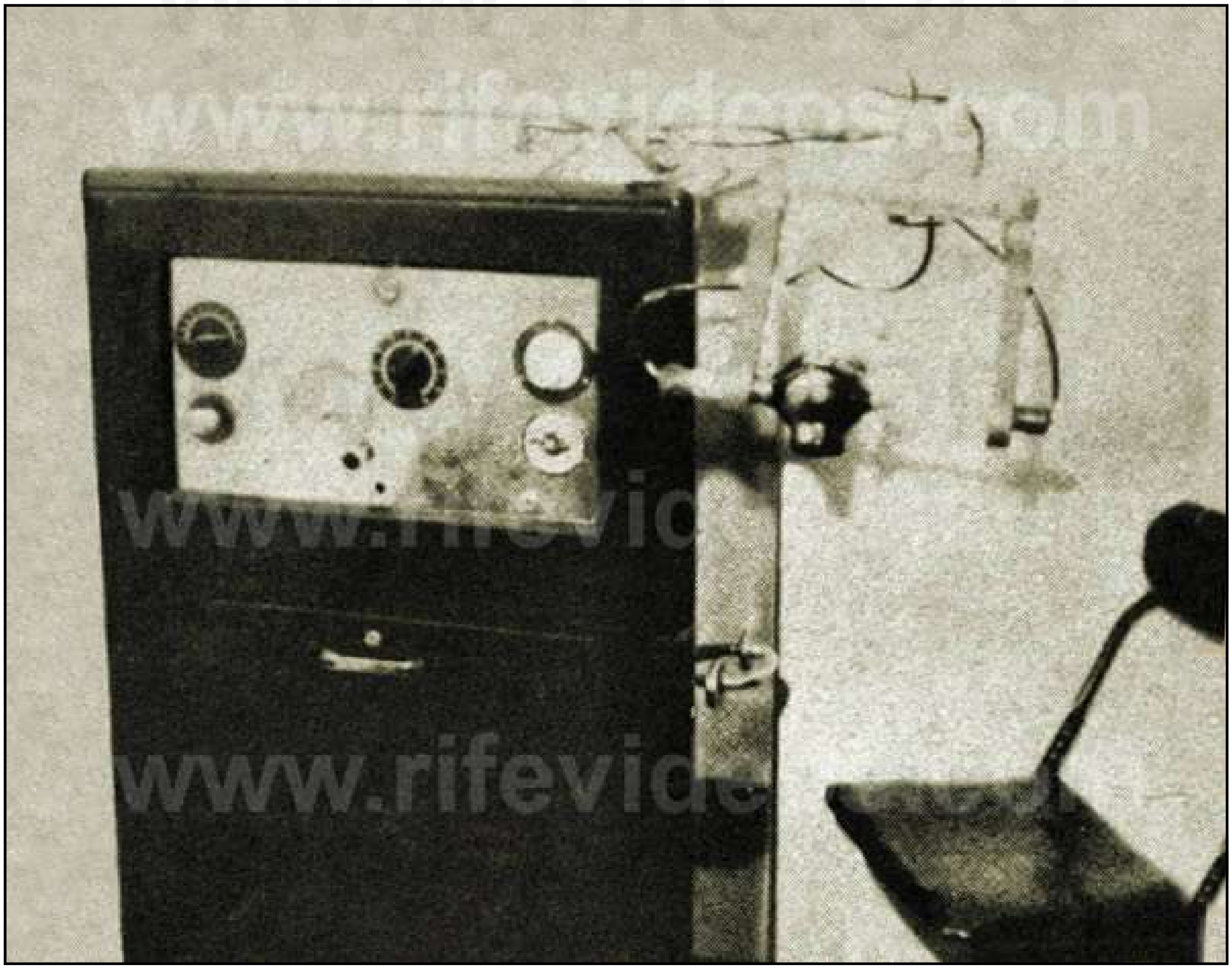

Chapter Summary: A ray tube is a plasma antenna built for the purpose of outputting frequencies. Dr. Rife built many different styles of ray tubes. The gas used inside the ray tube does not really matter as long as it is a noble gas (Helium, Argon, Krypton, Neon, etc.). Dr. Rife settled on using helium because it lasted longer in a ray tube that uses internal electrodes. The ray tube made it possible for Dr. Rife to safely use a powerful RF (Radio Frequency) frequency instrument next to a person or microscope. Dr. Rife preferred using a ray tube over pad type instruments because a great deal more power meant a better outcome for the patient.

Many people have wondered what power levels did Dr. Rife use in his instruments. In the next chapter of this report we will discuss the power levels of Dr. Rife's instruments.

www.rife.org

Chapter #2

What power levels did Dr. Rife use in his Rife Machines?

www.rifevideos.com

Many people have wondered what power levels are safe to use when using an instrument like Dr. Rife's machine. According to the documents we have Dr. Rife’s ray tube instruments, such as the Rife Ray #4 Rife Machine, was listed to be able to output as much as 400 watts. When the test for B. Typhosis was made it was listed that they used 400 watts. The milliamp meter was listed at 450 milliamps for those 400 watts. Since the equipment of that era would have only been about fifty percent efficient then the true power output from the ray tube would not have been more that about 200 watts, not 400 watts. The Rife Ray #4 was the most powerful of the five machines that Dr. Rife built. Dr. Rife did not notice any problems using instruments with power levels between 40 to 250 watts of power output. Next are two statements made by Dr. Rife about the safety of his instruments:

RIFE: *“I have operated the frequency instrument since 1921. I have watched it advance in style and performance with the advancement of electronics. In the many years I have used this equipment in my research, I have never suffered an injury or any ill effects whatsoever. I found it reliable in performance and efficient in results.”* (Letter from Dr. Rife to Dr. Justin Stein, July 2, 1956)

www.rife.org

Also on the John Marsh, Rife audio CD's Dr. Rife made this statement about his RF frequency instrument:

RIFE: *"I stood in front of that thing for thirty years finding these different frequencies that devitalize these different bacteria. And that thing* [RF ray tube] *was shooting on me right here [his chest], but it is absolutely harmless to normal tissue and each individual bacteria requiring a different frequency to devitalize."* (John Marsh Rife CD's)

The Rife Ray #5 or Beam Ray Clinical instrument built by the Beam Ray Corporation in the 1930's output, at the most, about 40 watts out of the ray tube. The 1950's AZ-58 Life Labs instrument and the Aubrey Scoon 1950's instrument also only output about 40 watts. Because some of Dr. Rife's information about instrument power levels is confusing, most have believed Dr. Rife's Machines put out 400 to 600 watts to the ray tube: however, the new information from the building of some of these instruments shows these assumptions are not correct. The problem has been that the people who wrote down this information were incorrectly giving the power usage of Dr. Rife's Machines as the output power. Dr. Rife's Machines used generally about 400 to 600 watts but they only output about 40 to 250 watts out of the ray tube. The Rife Ray #4 did output more power but this was the exception not the rule. In the paper entitled "Development of the Rife Ray" it states:

"The frequencies were generated by a tube oscillator with many stages [5 stages] *of amplification, the final stage being a 50 watt output tube."* (Development of the Rife Ray and use in devitalizing of pathogenic micro-organisms)

This part of the description is of his pre-1935 instrument. The output tube was not the ray tube. It appears from the documents that Dr. Rife's pre-1935 instruments did not output any more power than about 50 watts out of the ray tube. He said he lit the ray tube from another power source then input the frequencies into the ray tube. When Dr. Rife, John Crane and John Marsh were working on sea water conversion - a process that used frequencies - they boosted the output power in the instrument. Concerning that instrument and some 1930's Beam Ray instruments that Dr. Yale had increased the power level on, Dr. Rife said the following:

RIFE: *"Now this outfit here - the way we have it boosted up here now with an extreme lot of power behind the actual output that is coming out of the thing...I wouldn't want to use this - or I wouldn't want to use this instrument here the way it is souped up there for this salt water proposition to treat a patient with."*

GONIN: *"No."*

RIFE: *"You can get beyond the limit* [power levels in excess of 250 watts]*."*

GONIN: *"Yes, quite."*

CRANE: *"That's what Dr. Yale did. You see, he stepped it up and up and up…"*

RIFE: *"When Verne Thompson used to go down there and take care of Yale's machines - when he began stepping them up and so...where you get up into that extreme power…oh yes, that is not good. With the power that is in these* [50 to 100 watts of power coming out of the ray tube], *there is absolutely no harm because I had my microscope here - I had my tube [ray tube] right here in front of it - oh, about 11 or 12 inches away from the slide in the microscope and here I was with this thing all around like that and that tube going here and my specimens and the microscope year after year tuning that thing and it never harmed me any."* (See the photo at the beginning of this chapter for a better understanding of this statement about where the ray tube was located.) (1950's Gonin, Rife, Crane and Marsh Paper #27-32)

www.rife.org

Because Dr. Yale’s 1938-1939 Beam Ray Corporation Rife Machines were modified they were putting out a lot more power than Dr. Rife felt was safe. We do not know how Dr. Yale had his machines modified so it is not possible to know the exact power output they had. But he must have been exceeding the power output of the Rife Ray #4 which could output about 250 watts (true power). It may be that Dr. Rife was just overly cautious but his statement should be considered when anyone starts using power levels in excess of 250 watts true power output.

Chapter Summary: Dr. Rife's ray tube instruments output anywhere between 40 to 250 watts of RF (Radio Frequency) power levels. Dr. Rife found that his instruments were safe to use with power levels between 20 to 250 watts. Power levels in excess of about 250 watts (true power output) were not recommended by Dr. Rife.

In the next chapter we will discuss whether it is absolutely necessary that a ray tube should be used or if it is possible to use the contact method of delivering Dr. Rife’s frequencies which he used on microorganisms.

www.rife.org

Chapter #3

Is it necessary to use a Ray tube to output the frequencies?

www.rife.org

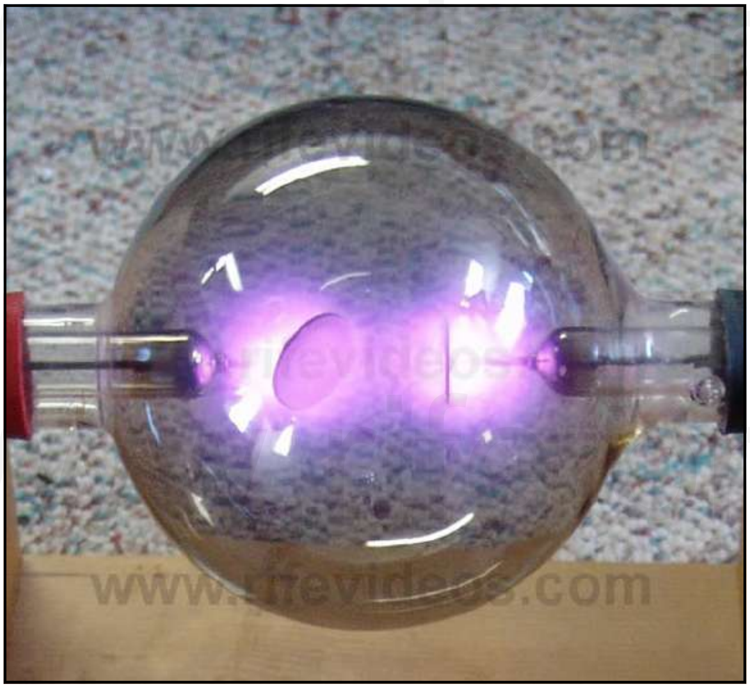

We really shouldn’t care if an instrument uses a ray tube or hand cylinders and footplates (commonly called pad instruments) as long as it will devitalize the microorganism we desire. In the next photo, shown below, we see one of the first pad instruments built by John Crane and John Marsh back in the late 1950's early 1960's. You can see the aluminum disks which were used to come in contact with the body and deliver the frequencies. In the strictest sense of the word just because a ray tube is used doesn’t mean its “Rife.” By the time you read this whole report you will find out that very few are doing exactly what Dr. Rife did. But does this mean that these instruments don’t work? Most of those who are building pad instruments are not using ray tubes, and most are not using Dr. Rife’s original frequencies. Those who are building ray tube instruments are also not using Dr. Rife’s original frequencies or methods. So where does this really leave us since very few are really doing what Dr. Rife did?

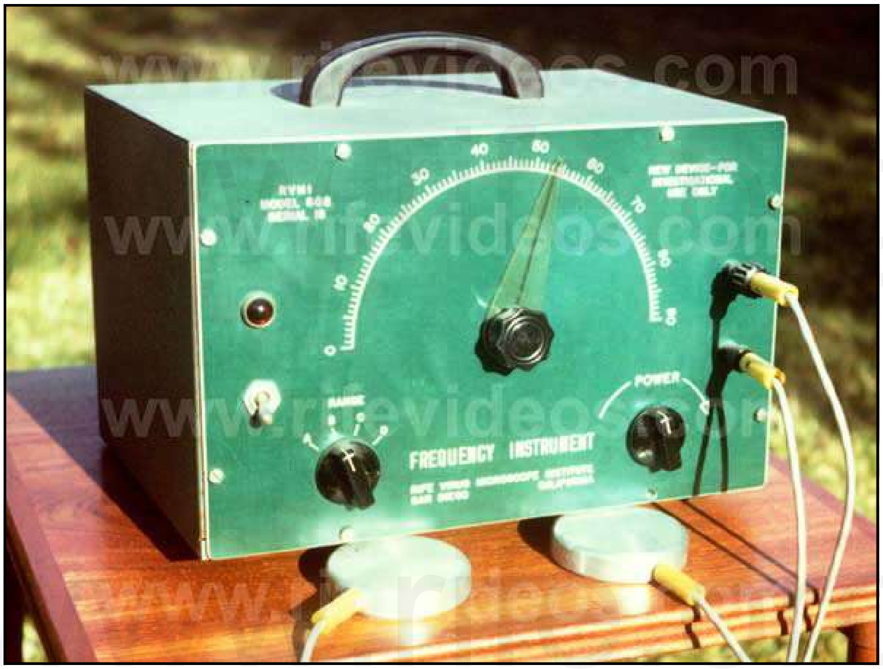

We have quite a paradox. This is the problem we face. If we were to build a ray tube Rife Machine that worked exactly the way Dr. Rife's did and use frequencies from 139,200 to 1,607,450 Hertz then we would have an instrument that could cause interference with AM radio stations if we were too close to them. These ray tube instruments may have to be used with a Faraday cage which is a conducting cage used to stop electromagnetic fields. We may be able to use them as long as we do not operate them within one half of a mile of a radio station on that stations particular wavelength or call number. We can build a pad instrument that will use all of the frequencies Dr. Rife's Machines used but then we are not using a ray tube. When we consider the problems we face today with building instruments, the least expensive instrument we can build is a pad instrument. This type of instrument can produce all of Dr. Rife's frequencies output by his Rife Machine. Therefore we should look at this method carefully and not reject it out of personal bias.

As we already said, it really shouldn't matter if an instrument uses pads or a ray tube as long as it works. With this in mind let's look at the reasons why pad instruments were built in the first place. John Crane and John Marsh had really good reasons why they built pad instruments. After nearly 50 years of research and use, there is enough evidence that a pad instrument may work just as well as a ray tube instrument, as long as there is sufficient power used. In some cases, because of the electrical stimulation like a T.E.N.S. instrument, they may work even better than a ray tube on some problems. We will now take a look at some of the reasons that prompted John Crane and John Marsh to use pads:

RIFE: *"But the principle of this thing is basically built on a coordinative vibration. Just like one tuning fork pitched to the C. Another one here—you strike this one and this one vibrates."*

DR. LARA: *"What kind of vibration is it? Electromagnetic vibration?"*

RIFE: *"We won't say magnetic, we will say electronic frequency vibration. The same as put out on a broadcasting station for the radio. The same thing you know, only it's transmitted into a tube. And the tube acts as a partial directional antenna you see."* (John Marsh Rife CDs - CD 6 track 2)

In the John Marsh papers describing his trip to Ohio we read a statement made by Dr. Rife:

RIFE: *"You know we had an idea when we had our Clinic in La Jolla, of course that was battery and motor generator operated that set, you know, and boy it would sure raise the devil with all the radios so we had a couple of cars that was equipped with car radios and we sent them out and we would take the switch of that thing, and had a code you know like an S.O.S., and one of them went up north, and one of them went south from La Jolla. Before we started in we wanted to see how far we were going to disturb things with it you know, and incidentally we had it in a steel room, a steel lined vault about this size at the old Ellen Scripp's home. It was the vault in the library of the Scripp's home where they kept their valuable manuscripts and books in all steel lined and a door on it like a safe. We had the thing inside of that too, but it didn't make much difference, but we started in, and one car lost the pick up on top of Torry Pines, and the other one half ways through Mission Beach picked it up, and then they could go a hundred feet and lose and then they would have to pick it up again. Old Henry [Henry Siner] the boy that was with us out there, one of the lab boys, boy he went up in the air. He says, "By God" he says "look, we're going to fix them up right. At two o'clock we'll hook this up to a big radio station, a big transmitting station, and at two o'clock next week we'll broadcast for tuberculosis, and at half past three the week after we will broadcast for cancer, and everybody at the radio will pick it up". See, boy I said Henry that really is an idea."* (1957 John Marsh Trip to Ohio Paper #24)

www.rife.org

In 1960 one of Dr. Rife's close friends, Ben Cullen, stated in a talk given at a Rife Virus Microscope Conference the following:

CULLEN: *"The fact is, had it not been for certain very, very unpleasant circumstances, Dr. Rife would've had an arrangement with KFI or KFSD* [Radio stations] *where he would have been broadcasting out over quite a large radius from each broadcasting station the rays which are responsible for eradicating these various viruses, which we now know as being "killing" viruses. Just imagine as you walked around shopping downtown, or out in the street, you would receive the rays of this wonderful current."* (Ben Cullen's 1960 talk given at the First International Convention of the Rife Virus Microscope #G)

These two statements made by Dr. Rife and Ben Cullen were made over 20 years after the 1934 clinic. Dr. Rife knew that the frequencies would broadcast from a metal antenna just as well as from a ray tube. The fact that he felt that Henry Siner's idea was a good idea and the fact that he wanted to use a radio station indicates that Dr. Rife knew a metal antenna would give the same results as a ray tube. It is apparent from what we have read that Dr. Rife believed it was the frequency that was devitalizing the organism and the method of application really didn't matter. Clearly, Dr. Rife understood that the frequencies could be broadcast by a radio station, using a metal antenna, if they had enough power.

When John Crane and John Marsh, Dr. Rife's two business partners in the 1950's, came to understand this, they eliminated the ray tube and used pads or hand cylinders to apply the frequencies. The pads and hand cylinders work just like an antenna except you do not want too much power so that they are safe to use. The body also becomes an antenna when you hold the hand cylinders or use the pads and this is why pad instruments work. Bertrand Comparet stated this in his interview:

COMPARET: *"Now, Crane said "Well now look, Rife himself admits that no matter how much tube and ray, and so on, you have, you can't get any results unless you've got the right frequency. Therefore the real clue to the thing is the frequency and not the means by which you deliver it." Comparet also said: "Well, Crane originally was, with more modern techniques, duplicating the Rife machine, tube and all for early experiments. And, as I say, he came to the conclusion that you just weren't getting anything additional by the use of the tube. If you didn't get the frequency, you could run the rest of it indefinitely and nothing happened. So, what Crane did, he got an audio frequency generator. Now, you could make them up yourself by an awful lot of work, or you could buy a Heathkit audio frequency generator and get all the same results with a lot less time and effort. So he was using these Heathkit generators. Now, instead of a beam projected from a tube, a ray, he simply had two wires. I think they were aluminium knobs on the end of them, which would be used. They would be put on the body in such a position that the natural flow of the current from one to the other would go through the diseased area, and he got astonishing results."* (1970's Bertrand Comparet Interview #33 & 47)

These pads or hand cylinders act just like an antenna when in contact with the body, but only if you have an RF carrier frequency. This is where John Crane and John Marsh made a critical error, we believe, and the reason Dr. Rife probably did not like their pad instrument. Without an RF carrier frequency the audio frequencies will only go through the connective tissue and not the cell. There are exceptions to this and they have to do with the wave form of the frequency. If a square wave audio frequency is used then the higher harmonics produced from this waveform may penetrate the cell to some degree. How much power from these harmonics penetrates the cell is not known. But this may explain why instruments that do not use an RF carrier frequency also seem to work well. Dr. Rife expressed his dislike for John Crane and John Marsh's instrument that did not use a carrier frequency when Bertrand Comparet asked him about it:

www.rife.org

COMPARET: *"And I asked Rife, because I thought Rife would certainly say that the way Crane was working on it then was still using the Rife principle, but he indignantly denied it.*" (1970's Comparet interview papers #32)

We know that Dr. Rife knew that a metal antenna would work and we also know that the pad instrument worked on this principle. Logically it must have been the fact that it did not use an RF carrier frequency is what upset Dr. Rife. Pad instruments that do not use a carrier frequency are limited in power. The highest power output that can be safely used from a non RF carrier pad instrument is about 1/5 of one watt (0.20 to 0.40). Any more power than this and the muscles of the body will begin to lock up. If you use an RF carrier frequency then you can output a hundred times more power safely. It is apparent that it was the lack of power that concerned Dr. Rife. All of Dr. Rife's original frequencies were in the RF (Radio frequency) broadcast band of frequencies and used with at least 40 to 50 watts of power. We will cover these frequencies and the audio frequencies along with the importance of a carrier frequency later in this article.

In the next photo, shown below, we see another pad type instrument built by John Crane back in the 1960's. You can see the same kind of disks that he used to come in contact with the body.

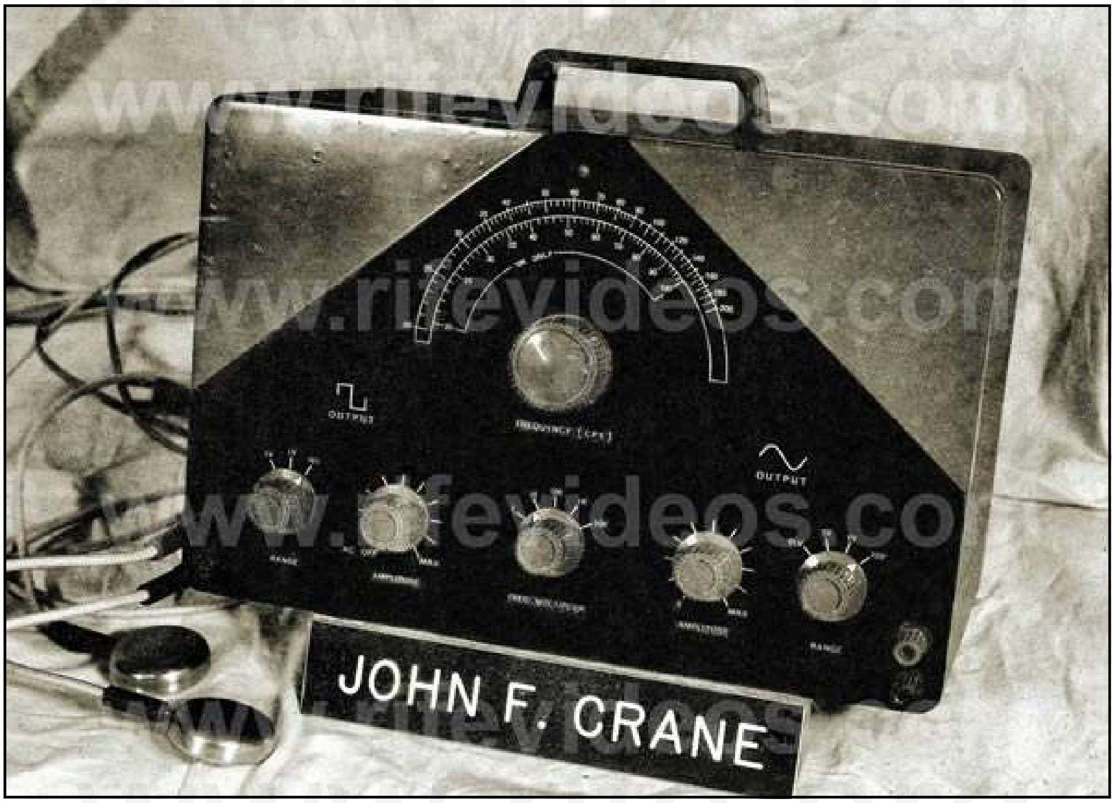

Some people have thought that it was the color of the light from the ray tube that made the Rife Machine work. But the evidence doesn't seem to support that concept either because in the Gonin Papers of John Marsh, Dr. Rife said this with regard to the light that came from the ray tube:

RIFE: *"We don't care about the color or anything of that sort."* (John Marsh Collection, Gonin Papers, Page 25)

Dr. Couche, while visiting Dr. Rife's lab with some other men, said:

DR. COUCHE: *"There was fifteen inches of concrete on the floor so as to stop any earthquake shocks from interfering with his work. And in his laboratory upon the ground floor he had a microscope with a slide on it that this group of people and myself looked at. And this was not stained, there was no killing*

www.rife.org

of the bacteria on it. It was just a fresh culture of the colon bacillus…..Well we all went down under the stairs into the cellar right immediately under the microscope upon the floor above us and the Rife machine was down in underneath there under the culture in the cellar probably I suppose about ten feet away, eight or ten feet away. And he turned the machine on and gave it less than a half minute's frequency for the colon bacillus...Then he turned the machine off and we all came upstairs and waited for ten or fifteen minutes. And presently he came back to his microscope and he said, "Well gentlemen come and look at the slide now." Well to my astonishment the bacilli all had been killed and they were all stacked up on the slide." (John Marsh Rife CDs - CD 3 track 1)

There is no possible way the light from the ray tube of the Rife Machine could have penetrated that fifteen inch concrete floor. It is obvious that the light didn't make any difference but that it was the frequencies that were broadcast through the ray tube. It is easy to see that there is more than one way to deliver the frequencies. The ray tube could be easily replaced with metal hand cylinders and foot pads. It is interesting to note here that Dr. Rife said Abrams' Oscilloclast would devitalize the BX cancer virus and it was a contact type device. The waveform the Oscilloclast produced is shown in Dr. Rife's 1936 film. John Crane and John Marsh probably used this contact method because of the success of Abrams' instrument. The Abrams' instrument proved that a contact type device would work and it was used before Dr. Rife even started using a ray tube. In fact Abrams' contact instrument predates all of Dr. Rife's work. Pad instruments like Abrams' instrument come in contact with the body. Abrams instrument worked on the same RF principles as Dr. Rife's instruments. Pad instruments with an RF carrier turn the body into an antenna and work on the same principle as a metal antenna or ray tube. People have been using pad instruments without an RF carrier for almost 50 years now and have had very good results. But, in order to work the way the ray tube instruments do, an RF carrier frequency is necessary.

Chapter Summary: Dr. Rife knew that a metal antenna could be used in place of a ray tube. Frequency generators that use hand cylinders or footpads or footplates may work as well as a ray tube instrument as long as sufficient power is used. The only way to increase the power level in a pad instrument is with the use of an RF (Radio Frequency) carrier. Dr. Rife did not approve of John Crane and John Marsh's pad instrument due to its lack of power and its lack of an RF carrier frequency. Unless you can have a frequency generator with sufficient power you are better off using a ray tube instrument with 40 to 250 watts of power output. If a person is dealing with a very serious condition they may want to use a ray tube instrument instead of a pad type instrument if they have access to one.

Many people have wondered if the frequencies that Dr. Rife used are safe. In the next chapter we will discuss the safety of these frequencies.

www.rife.org

Chapter #4

Are Dr. Rife’s RF frequencies safe to use?

Rife Ray #3 And Rife Ray #4 High Sine Wave RF Frequencies

Actinomycosis (Streptothrix)	192,000 Hz
Anthrax	139,200 Hz
B. Coli (Rod form)	417,000 Hz
B. Coli (Filterable virus)	770,000 Hz
Bacillus X or BX (Cancer Carcinoma)	1,604,000 Hz
Bacillus Y or BY (Cancer Sarcoma)	1,530,000 Hz
Gonorrhea	233,000 Hz
Spinal Meningitis	427,000 Hz
Staphylococcus Pyogenes Aureus	478,000 Hz
Staphylococcus Pyogenes Albus	549,070 Hz
Streptococcus Pyogenes	720,000 Hz
Syphilis	789,000 Hz
Tetanus	234,000 Hz
Tuberculosis (Rod)	369,000 Hz
Tuberculosis (Virus)	769,000 Hz
Typhoid Fever (Rod)	760,000 Hz
Typhoid Fever (Virus)	1,445,000 Hz

Today there are many who profess to believe in Dr. Rife and his method of coordinative resonance using frequencies. But they also claim that RF or Radio frequencies are bad or harmful. Some of these people build and sell instruments that do not use any RF carrier frequency and put in their sales information that RF frequencies are harmful. This is a great disservice to Dr. Rife. The instruments they build use low audio frequencies like those used in the Rife Ray #5 or Beam Ray clinical Rife Machine and the 1953 AZ-58 instrument. But they ignoring the importance of the RF carrier frequency these machines used. (These low audio frequencies will be discussed later).

When it comes to Dr. Rife and the method he used in his Rife Machine you cannot “have your cake and eat it too.” Dr. Rife’s principles were all based on coordinative resonance from RF frequencies. Any method used that does not use RF frequencies cannot be called Dr. Rife’s method nor can it truly be called a "Rife Machine." Today, regardless of the method used, people call their instruments “Rife Machines” no matter how they are built. Dr. Rife specifically asked that his name not be put on any instrument, yet this is exactly what people do. There is nothing wrong with people building non RF instruments, but claiming that RF is bad or unsafe just to sell their instruments is where the problem lies. Dr. Rife’s Colin B. Kennedy equipment (which will be discussed later in detail) had a frequency range of about 12,000 Hertz to about 2,000,000 Hertz. With this equipment Dr. Rife found the many frequencies (shown above in the chart) that would eliminate or devitalize the various organisms he tested.

www.rife.org

Anyone who understands anything about frequencies knows that the frequencies in that chart are all RF or radio frequencies. It is impossible to build an instrument that could output these frequencies of Dr. Rife's without it using RF frequencies. Every one of Dr. Rife's instruments from the Rife Ray #1 to the Rife Ray #5 used RF frequencies. This is a fact of history and it cannot be changed. Yet there are some who are so foolish as to state that RF instruments such as Dr. Rife's instruments are "Dangerous Rife Machines." It is hard to believe that this is being said by some who claim to be selling "Rife Machines." It is obvious that they know very little about Dr. Rife and his methods. All of the frequencies that Dr. Rife used were in the AM radio band of frequencies. If they were dangerous then we would not have any AM radio stations on the air today. AM radio stations have been around for almost 100 years. If these frequencies were dangerous then people would have been falling dead all over the place for the last 100 years. This is how foolish the idea is that Dr. Rife's RF frequencies and instruments were dangerous.

Another foolish statement that some have made is that these instruments include frequencies in the "Microwave band." The AM radio band is not in the Microwave band. Dr. Rife's Rife Ray #4 had a frequency range that went as high as 22.5 Megahertz (22,500,000 Hertz or 22.5 million Hertz). The microwave band begins at about 300 Megahertz (300,000,000 Hertz) and goes to about 3 Gigahertz. None of Dr. Rife's machines were ever in the microwave band of frequencies. It is clear that people who have written this false information clearly know nothing about Dr. Rife or his equipment. In this report we have no ax to grind other than to point out how foolish some of the false information that is being written by those who claim to know a great deal about Dr. Rife. There is a lot of nonsense being written by some people in order to try and sell some kind of so called "Rife Machine”.

In Dr. Rife’s tests he would have naturally started in the low frequency range and slowly worked his way higher until he found a frequency that would eliminate the organism. All of the frequencies that he found were in the RF range. They went from 139,200 Hertz for Anthrax to 1,604,000 Hertz for the BX cancer virus (Frequency range taken from the Rife Ray #4 documents). It was only these frequencies that Dr. Rife found that would resonate the organisms and devitalize them. Since Dr. Rife found that only these RF frequencies would resonate and kill the organism then it is impossible to separate RF from coordinative resonance. It would also be impossible to build an instrument that truly worked on Dr. Rife’s principles without the use of RF or radio frequencies. In order to prove the safety of Dr. Rife’s work we must quote him since he is the person who used these frequencies for over 30 years in his laboratory. Below are two of his quotes on the safety of using RF frequencies in the range and power level that he used in his Rife Machines. The first quote comes from a letter sent to Dr. Stein in 1956:

RIFE: *“I have operated the frequency instrument since 1921. I have watched it advance in style and performance with the advancement of electronics. In the many years I have used this equipment in my research, I have never suffered an injury or any ill effects whatsoever. I found it reliable in performance and efficient in results.”* (Letter from Dr. Rife to Dr. Justin Stein, July 2, 1956)

On the John Marsh, Rife audio CDs Dr. Rife also made this statement about his RF frequency instrument:

RIFE: *“I stood in front of that thing for thirty years finding these different frequencies that devitalize these different bacteria. And that thing [RF ray tube] was shooting on me right here [his chest], but it is absolutely harmless to normal tissue and each individual bacteria requiring a different frequency to devitalize.”* (John Marsh Rife CDs)

Dr. Milbank Johnson, M.D. also used the RF Rife Machine for many years and conducted clinics and found the instrument safe to use. Dr. James B. Couche, M.D. used the instrument in his private practice for over 22 years and also said he found the instrument safe to use. Dr. Tully, D.D.S. pur-

www.rife.org

chased one of Dr. Couche’s instruments and used it for several years and found it completely safe to use. Dr. Robert P. Stafford, M.D. used the frequency instrument for over 5 years and also expressed that he found the instrument completely safe to use. These statements along with Dr. Rife’s that we quoted above show that Dr. Rife found that his RF frequencies, in the ranges he used, were as safe to use as the frequencies output by any radio station. These frequencies are broadcast through the air day and night passing through our homes without any harm to the human body. There may be some people with RF sensitivity but this does not mean that RF is unsafe to use. This only means that these people are sensitive to RF and should avoid it if they find a problem using it.

Chapter Summary: An instrument that does not output Dr. Rife's RF frequencies is not a "Rife Machine" in the strictest sense of the word. Neither can they resonate the microorganisms that Dr. Rife proved could be devitalized with his RF frequencies. None of Dr. Rife's RF instruments ever put out frequencies in the "Microwave band" of frequencies. Dr. Rife found that his frequencies and the power levels he used in his instruments were safe to use. Non RF frequency generators appear to be very good instruments and help may people but they do not work on the same methods and principles that Dr. Rife used.

Many people have wondered if Dr. Rife ever used audio frequencies. In the next chapter we will discuss whether Dr. Rife ever used audio frequencies with his equipment.

www.rife.org
www.rife.org
www.rife.org
www.rife.org
www.rife.org

Chapter #5

Did Dr. Rife use audio frequencies?

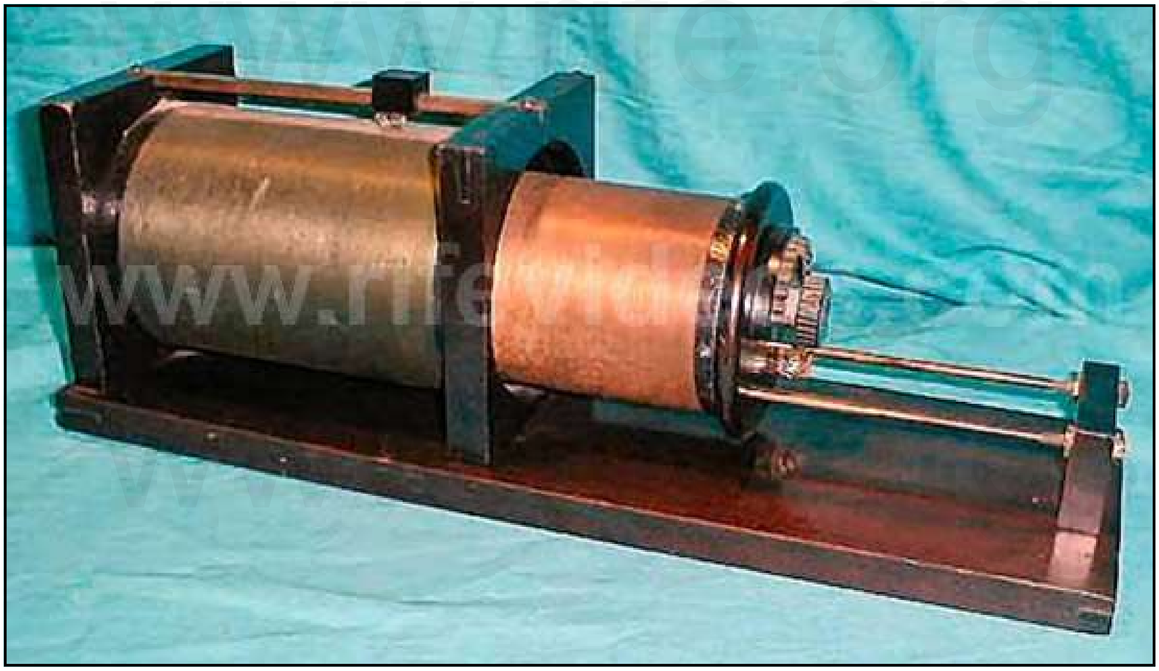

In Dr. Rife's 1961 deposition he revealed the fact that he was using some form of audio modulation from the beginning of his work with frequency instruments:

RIFE: *"Initially I worked with loose couplers to get an audio oscillation and then with the use of transmitters, I tried to balance the audio and modulate the audio on a carrier wave to transmit the audio energy."* (Dr. Rife's 1961 Deposition, answers page 5)

Dr. Rife stated that at the beginning of his work, back in the 1920's when loose couplers were used in generating frequencies, he was using audio modulation with an RF or radio frequency carrier. Above is a photo of a loose coupler courtesy of "Henry Rogers, Western Historic Radio Museum" (www.radioblvd.com). These loose couplers worked by moving one coil inside the other in order to change the frequencies.

In Dr. Rife's Rife Ray #3 and Rife Ray #4 Rife Machines he used a specific audio circuit which gated or pulsed his frequencies. This circuit modulated a fixed audio frequency with all of his high frequencies which he used on the various organisms. This circuit will be discussed later when we discuss these Rife Machines.

It is apparent that Dr. Rife first tested audio frequencies on organisms in his search for the frequency which would devitalize them. The audio range would be the logical place to start. If he couldn't find a frequency in the audio range, he then moved up into higher frequency ranges until he found a frequency that would devitalize an organism. In Dr. Rife's early lab notes, he listed only two organisms that had an audio frequency M.O.R. Later when Dr. Rife found out that he was reading his frequencies incorrectly, one of those audio frequencies was changed to a much higher RF frequency. There are other statements made by Dr. Rife which show that he tested the audio range. In fact Dr. Rife gave the full range of his frequencies:

www.rife.org

RIFE: *"Some of them are in the visible band, or I mean not only the visible band but, uh, band of frequencies audible to the human ear [audio frequencies are frequencies that you can hear as tones]. Some of them are way beyond either way. They run through a very, very large gamut. Some of them are very, very broad, long. Some of them are...not extremely short. There are none of them what we call our ultra short wave that I have found yet. Well there's many of them...we would, uh, classify in the ultrasonic band because they're not visible* [sic] *with the human ear. They're way beyond you know. And some of them are even in the broadcast band. Your cancer is very high* [1,604,000 Hertz]. *You can't hear it, the oscillation. But now you take your T.B.* [Tuberculosis Rod 369,000 Hertz & Virus 769,000 Hertz]. *Now that's down. A little more you see...if you don't have an absolute coordinative resonance, you have nothing. One tenth of one meter off and you have nothing. Its got to be absolutely correct for that individual organism. It's got to be precise...the virus of cancer has a certain frequency. And it has to be there, otherwise if it's a little one way or the other, no good, no good for nothing. Infrared will penetrate, yes, but the heat is not the thing because the heat is not the frequency, it's [Infrared] way down in the very low band of frequencies and the laboratory rate of the BX is up into the high band."* (John Marsh Rife CDs - CD 5 track 2, CD 6 track 2, CD 7 track 1 and CD 9 track 1)

In these statements Dr. Rife clearly explains the broad range of his frequencies. Some were audio and could be heard by the human ear; others were in the ultrasonic range, and some were even in the broadcast band. Cancer he said was very high. He states the frequencies have to be very accurate to work. One tenth of one meter off and they would not work at all. We will talk about this later. Here are two additional statements that also verify that Dr. Rife's Machine's could output a modulated audio frequency:

RIFE: *"You know we had an idea when we had our* [1934] *Clinic in La Jolla, of course that was battery and motor generator operated that set, you know, and boy it would sure raise the devil with all the radios so we had a couple of cars that was equipped with car radios and we sent them out and we would take the switch of that thing, and had a code you know like an S.O.S., and one of them went up north, and one of them went south from La Jolla. Before we started in we wanted to see how far we were going to disturb things"* (John Marsh Collection, Trip to Ohio Papers, #24)

In the next photo, shown below, is the Rife Ray #3 which was used at the 1934 Clinic in La Jolla.

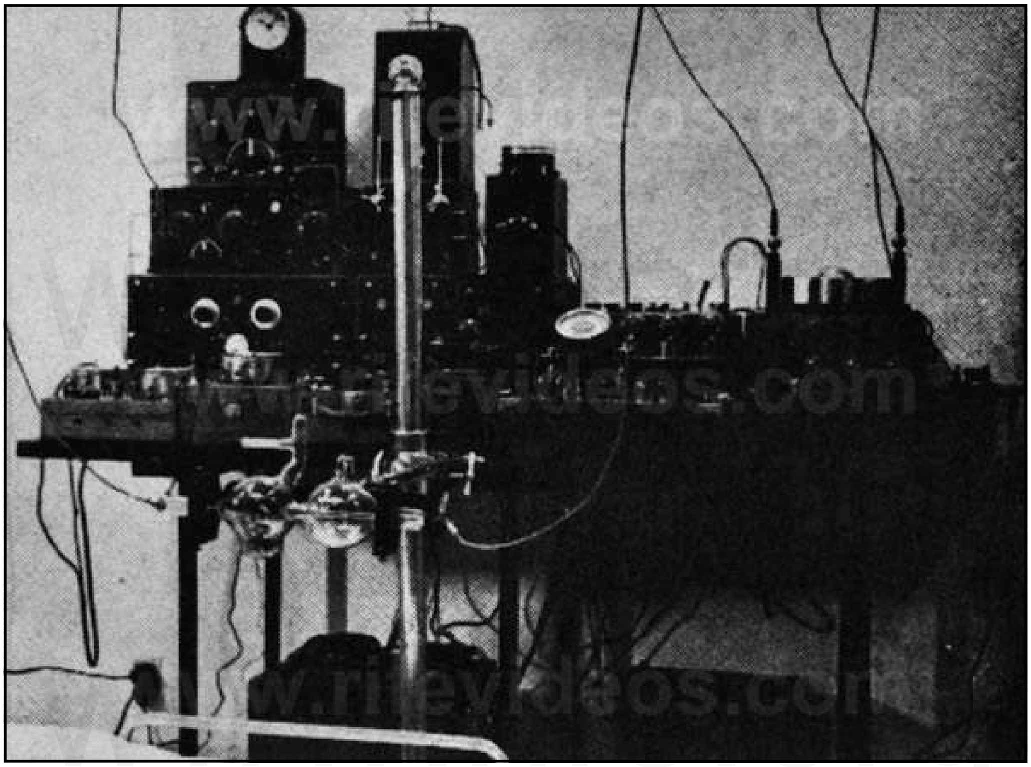

www.rife.org

In order to be able to put out an S.O.S. type signal he would have had to modulate the audio frequency onto a carrier in order for the car radios to pick up the signal. On the John Marsh Collection of Dr. Rife's audio CD's, Dr. Couche makes an interesting comment about the Rife Ray #3 instrument. He was present at the 1934 clinic sponsored by Dr. Johnson and the University of Southern California. He stated:

DR. COUCHE: *"They gave him a treatment of the Rife frequencies which are in the auditory band."* (John Marsh Rife CD's - CD 3 track 1)

The cancer frequency (Rife Ray #4 1,604,000 Hertz and 1,607,450 for the Rife Ray #5) and tuberculosis (Rife Ray #4 Rod form 369,000 Hertz and 369,433 for the Rife Ray #5)(Rife Ray #4 and Rife Ray #5 Virus form 769,000 Hertz) frequencies used in the 1934 clinic were not audio frequencies. Why would Dr. Couche make this statement? The evidence shows that Dr. Couche was getting the Rife Ray #4 and the Rife Ray #5 methods of generating the frequencies mixed up. The Rife Ray #5 or Beam Ray Clinical Rife Machine which Dr. Couche used for over 22 years used audio frequencies to create the proper sideband spacing to hit the high harmonic RF M.O.R. frequency (Mortal Oscillatory Rate or the frequency that will kill or devitalize an organism). Dr. Couche purchased two of these clinical Rife Machines and used them until 1952 when he retired. We will cover this instrument later in this paper. All the evidence that we have acquired from the quotes show that the Rife Machines from 1934 and earlier could output audio frequencies above 12,000 Hertz and that Dr. Rife tested audio frequencies right from the beginning of his work in 1920.

Although we have been able to prove that Dr. Rife tested the audio range of frequencies, as any good scientist would have done, it should be pointed out that by 1935 when the Rife Ray #4 Rife Machine was built, he no longer felt that he needed to test audio frequencies any longer. This is indicated by the fact that no variable audio oscillator was included in this new Rife Ray #4 frequency instrument. The Rife Ray #4 Rife Machine will be discussed later in this article.

In 1936 Dr. Rife's engineer, Philip Hoyland, developed a new machine which used audio frequencies to create Dr. Rife's high RF frequencies through harmonic sidebands. The audio frequencies were not the frequencies that would devitalize or kill the various microorganisms. These audio frequencies were used to produce the frequencies that would devitalize or eliminate the disease organisms. This method of using harmonic sidebands will be discussed later in this report when we examine the Rife Ray #5 or Beam Ray Clinical instrument.

Chapter Summary: Dr. Rife tested and use audio frequencies in his instruments. Primarily the audio frequency was modulated onto the high RF frequency to create a pulse in the higher RF or radio frequencies he was using to devitalize the various microorganisms. When his engineer, Philip Hoyland, developed the new Rife Ray #5 or Beam Ray Clinical instrument the audio frequencies were used to produce Dr. Rife's High RF frequencies through harmonic sidebands.

In the next chapter we will look at Dr. Rife's first frequency instrument called the Rife Ray #1 and how he was able to devitalize his first microorganism using it.

www.rife.org

Chapter #6

Dr. Rife's 1920 to 1922 Rife Ray #1 Rife Machine

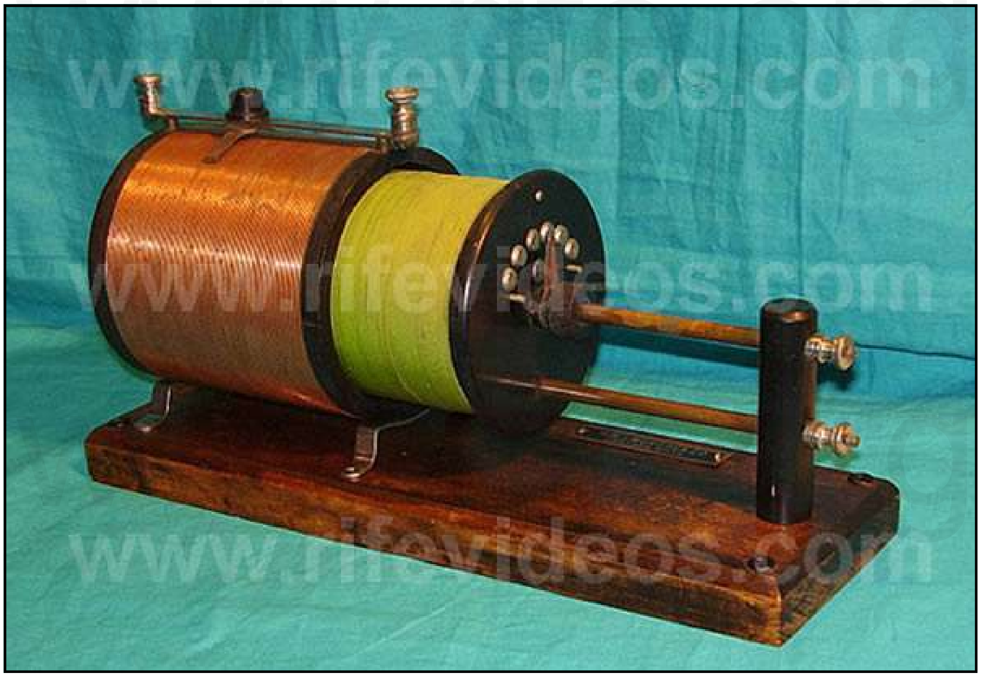

1. Used a Ray tube.
2. Frequency range was probably about 1 MHz or less using loose couplers: Could modulate audio frequencies.
3. Output to the Ray tube was 8 to 10 watts.

From information taken from many different documented sources there is finally enough information to know what Dr. Rife did in the early years, from 1920 to about 1934. Dr. Rife designated his machines in numerical order beginning with the Rife Ray #1. Even though we do not have any photos of his Rife Ray #1, we know that Dr. Rife was using loose couplers (photo shown above is a loose coupler) because of this statement:

RIFE: *"Initially I worked with loose couplers to get an audio oscillation and then with the use of transmitters, I tried to balance the audio and modulate the audio on a carrier wave to transmit the audio energy."* (Dr. Rife's 1961 Deposition, question and answer #35)

In a paper that was written in the 1930's, probably by Jack Free, one of Dr. Rife's lab assistants, we get a history of Dr. Rife's machines. This paper starts with the Rife Ray #1.

"In January 1920 experiments were started at the Rife Research Laboratory by Commander Royal R. Rife U.S.N. Ret. to determine the effect of electrical influences upon pathogenic microorganisms. Tests were made for anode and cathode polarity influences and the effect of infrared, ultraviolet and X-ray. During these experiments the idea was conceived of the possibility of devitalizing the pathogenic micro-organisms by electrical frequencies of varying wavelengths. The initial apparatus [Rife Ray #1] *for the tests along this line of experiments was constructed and used in prolonged experiments during 1921 and 1922, with results that warranted the belief that the principles involved contained possibilities."* (Development of The Rife Ray and use in devitalizing of pathogenic microorganisms).

www.rife.org

The tests that were done with the Rife Ray #1 which showed that it was possible to devitalize micro-organisms with frequencies was talked about in a 1938 Evening Tribune newspaper article. Also in this newspaper article we learn the power output of the Rife Ray #1 and the first organism that he was able to devitalize.

“Rife built a simple frequency generating apparatus of about 8 or 10 watts output. He grew some cultures of bacteria. Then he began the studies whose reported results now promise to revolutionize the entire theory and the whole treatment of the human diseases, other than those of a functional or accidental nature. Machine and cultures ready, the San Diegan anxiously, feverishly began testing his idea. Would those minute killers of men die under the frequency bombardment? It would be a patience-wracking task, for there was no way to measure what wave length or frequencies the organism might have. In the quiet loneliness of the laboratory, Rife simply had to turn and turn and turn the tuning dials of his machine and check after each bombardment the conditions of the disease organisms in his cultures to see if anything had happened to them. He just had to hunt by trial and error a frequency, which might do something to a certain organism. Then, if he found one for that disease, he would have to start all over again on the next kind.

The scientist took first a culture of b. coli, the organisms, which always seem to accompany the agency of typhoid fever yet apparently, are harmless themselves. He prepared microscope slides from the culture and saw that his little subjects were alive. Then he turned the ray on them, tuned it to a certain frequency, then took the slide back to the microscope to see if anything had happened. He did this time after time and the b. coli still remained discouragingly healthy. Then one day, Rife recounted, a culture of the organisms which had been bombarded with a certain frequency [417,000 Hertz] *appeared different under the microscope. They seemed lifeless! He tried to get them to grow, to reproduce in their laboratory media. He tried that same frequency on culture after culture of b. coli and always the results were the same. The organisms were dead. "It did kill them!" Rife told himself. And probably, cool, conservative scientists though he is, he allowed himself to hope that he, Royal Raymond Rife, had found that 'bullet" which scientist have sought for years, that "magic bullet' which would surely, certainly slay mankind's diseases.”* (Dread Disease Germs Destroyed By Rays, Claim Of S.D. Scientist-Cancer Blow Seen After 18-year Toil by Rife, The Evening Tribune, May 6, 1938)

These few statements that we have just read give us all that is really known about the Rife Ray #1. The most important development that was accomplished with the Rife Ray #1 was Dr. Rife was able to test and prove that microorganisms could be killed or devitalized using RF frequencies pulsed with an audio frequency.

The photo, shown below, is a picture of a more complex loose coupler setup.

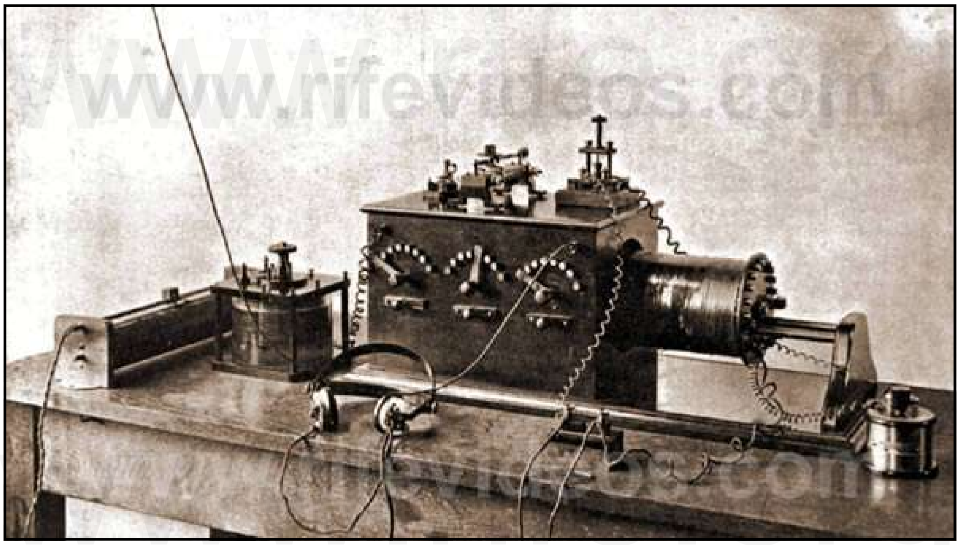

www.rife.org

By 1923 Dr. Rife assembled his next Rife Machine called the Rife Ray #2 using off-the-shelf frequency generators. The same equipment that was used with the Rife Ray #2 was also used with the Rife Ray #3. The changes made to the Rife Ray #2 to create the Rife Ray #3 appear to be mostly an increase in the power output of the instrument through the ray tube. Since the full details of the changes are not known and the same equipment that was used with the Rife Ray #2 was used with the Rife Ray #3 we will look at the Rife Ray #3 in the next chapter.

Chapter Summary: The Rife Ray #1 was a simple frequency generating device built to determine if microorganisms could be devitalize by frequencies. The instrument was built using loose couplers which when moved in and out of the coil would produce different electrical RF frequencies. Dr. Rife modulated an audio frequency on the RF frequency he was using in order to produce a pulsed waveform. With this simple frequency generator he was able to devitalize his first organism which was B. coli. Dr. Rife after proving his theory went on to build more powerful equipment which we will look at in succeeding chapters in this report.

In the next chapter we will look discuss Dr. Rife's Rife Ray #3 instrument which was used in the 1934 cancer and tuberculosis clinic which Dr. Milbank Johnson conducted.

www.rife.org

Chapter #7

1934 Rife Ray #3 Rife Machine used in the 1934 clinic

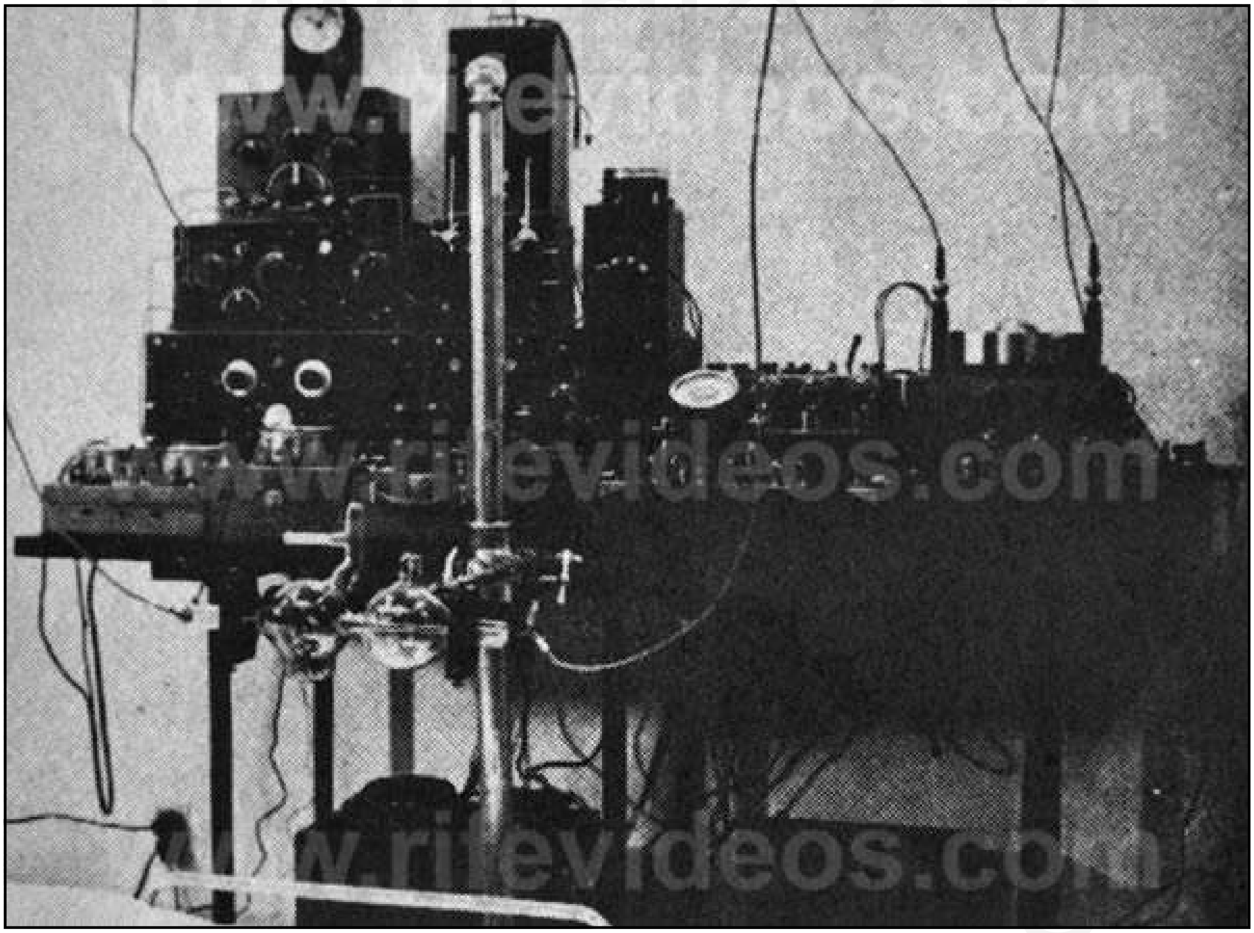

www.rifevideos.com

1. This was a regenerative instrument that used a ray tube.
2. It consisted of two Kennedy Regenerative Receivers (the model numbers were the 110 and 281). These two receivers made it possible to have a combination of one low frequency oscillator and one high frequency oscillator or two high frequency oscillators.
3. The output was sine wave.
4. Power usage was from batteries. Output to the ray tube was about 50 RF watts?

This instrument was described in a document believed to have been written by Jack Free one of Dr. Rife's lab assistance. We will quote the portion of that document that pertains to this instrument:

"In 1923 more appropriate apparatus [Rife Ray #2 when improved became the Rife Ray #3] *was assembled and used. The different frequencies were generated by a tube oscillator with many stages of amplification, the final stage being a 50 watt Telefunken tube.*

This amplified frequency was in turn fed into an output tube, and as the voltage at this point was quite small, it was found necessary to apply external voltage across the anode and the cathode of the output tube [ray tube] *to act as a carrier wave for the frequencies that were generated in the apparatus.*

The output tube was constructed with a double expansion bulb, blown from quartz, using platinum anode and cathode it having a 45° target for directional effect. No heat is generated in output tube-

www.rife.org

temperature constant. The frequency control of the instrument was exact to a fraction of a wavelength making it possible to coordinate the frequency in each pathogenic micro-organism with its own wavelength of frequency delivered from the instrument. The current supply for the whole apparatus was supplied by batteries and generators.

During the next eight years these experiments continued and with the aid of the Rife super microscope and the frequency instruments the coordinating frequencies (termed mortal oscillatory rate MOR) of most of the pathogenic micro-organisms were found and recorded including the frequencies of many of the virus or filter passing forms of these organisms." (Development of the Rife Ray and use in devitalizing of pathogenic micro-organisms).

Dr. Rife was asked by Dr. Milbank Johnson M.D. to write a description of his Rife Ray #3 instrument in 1935. Dr. Rife had Jack Free, his lab assistant; include this description in a letter that Jack Free wrote to Dr. Milbank Johnson. Below is Dr. Rife's description.

RIFE: *"The basic principle of this devise is the control of a desired frequency. These frequencies varying upon the organisms being treated.*

The frequency is set which controls the initial oscillator, which in turn is run thru six stages of amplification, the last stage driving a 50 watt output tube.

The frequency with its carrier wave is transmitted into an output tube [double-bubble ray tube] *similar to the standard X-ray tube, but filled with a different inert gas. This tube acts as a directional antenna.*

The importance in the variable control of these frequencies is that each pathogenic organism being treated is of a different chemical consistency, the consequence being they carry a different molecular vibratory rate. Each one in turn under these conditions requires a different frequency or vibratory rate to destroy." (Letter from Jack Free to Dr. Milbank Johnson M.D., December 17, 1935).

This Rife Machine was used in the 1934 clinic by Dr. Milbank Johnson. See photo of Rife Ray #3 on page 26. If you look at the bottom of that photo of the Rife ray #3 you can see part of the bed railing and mattress where they treated the patients. If you look at the table you can see that the instrument was not a one piece instrument but had many components. This Rife Machine has always been considered the best instrument used by Dr. Rife because it produced the results of the 1934 cancer and tuberculosis clinic. Those interested in the work of Dr. Rife have always wanted to know how this instrument worked. They have also wondered what equipment he used. This has been one of the biggest Rife mysteries. There have been all kinds of speculation on how his first instrument worked. What was its waveform? What was the frequency range? Could it generate audio frequencies? Was it super-regenerative (as he wrote on his lab notes), or was it just regenerative? All of these things have remained mysteries for over fifty years. It was generally believed that the 1934 instrument was custom made for Dr. Rife. However, if the equipment had not been custom made, the mystery could be solved. And today, thanks to some great detective work done by Mr. Peters, the mystery, in fact, is now solved. The instruments were not custom made. They were standard off-the-shelf frequency generating equipment that Dr. Rife purchased. The equipment and frequency ranges are now known.

A better photo of the equipment Dr. Rife used is the first photo on the next page. It was when Mr. Peters was looking at this photo that he recognized the Kennedy frequency generating equipment. Dr. Rife most likely stacked it all up on a table and took a picture of it after he started to use the newer equipment built for him in 1935. This photo, amongst others, made it possible to figure out the equipment Dr. Rife used. This photo has been provided courtesy of Jason Ringas of Rife Research Group of Canada. Here in this paper you will be able to see the actual equipment along with the selling advertisements of the 1920's that give the specifications of the equipment.

www.rife.org

We will now look at each piece of equipment and take an in-depth look at the specifications of each. All pieces of equipment except the ray tubes and possibly the five stage amplifier were considered off-the-shelf equipment. This means that this was standard frequency generation equipment which could be purchased from companies in the 1920's. Although they are regenerative receivers, they could output whatever frequency Dr. Rife wanted to use when the regenerative circuit was turned up. Dr. Rife used top-of-the-line Kennedy equipment from the Colin B. Kennedy Company, which built some of the most accurate, high quality equipment that could be purchased in 1923. It was also some of the most expensive equipment to purchase.

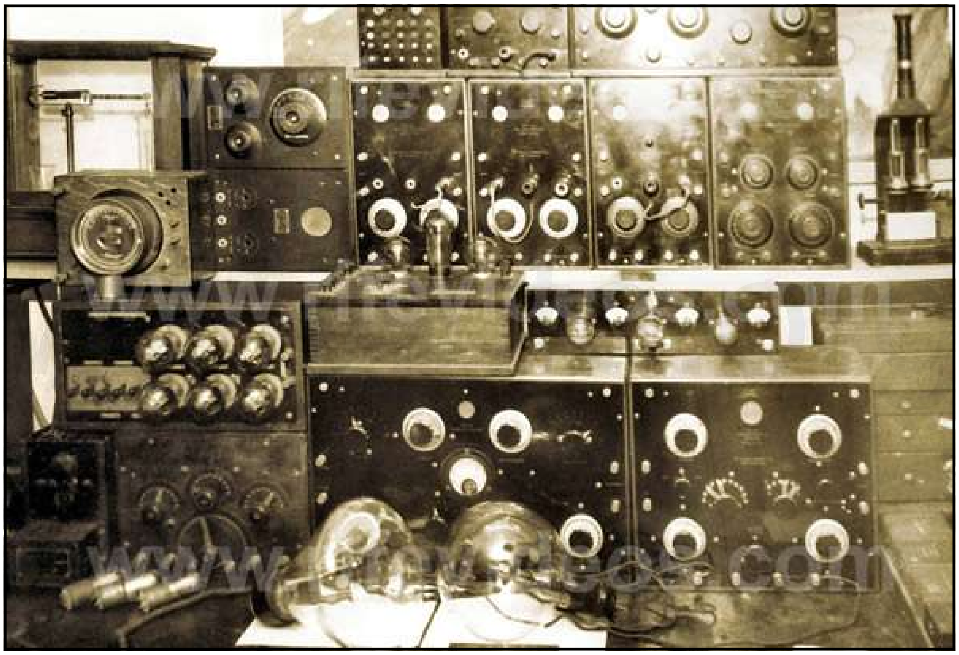

We will now take a look at the next photo, shown below. This photo is one of several pictures of Dr. Rife's laboratory instruments. The bottom two pieces of radio equipment were the Kennedy Receiver Model 110 connected to the Kennedy Two-Stage Audio Amplifier Model 525. The other piece of equipment sitting on top of the Kennedy Receiver Model 110 we will look at later. On the next page are two more photos. The first photo is a better photo of this old antique equipment. The second photo is the 1923 advertisement from the Colin B. Kennedy Company which provides the frequency range and features of this regenerative receiver. It also gives the effective frequency range from 175 to 25,000 meters or from 12,000 Hertz to 1,700,000 Hertz.

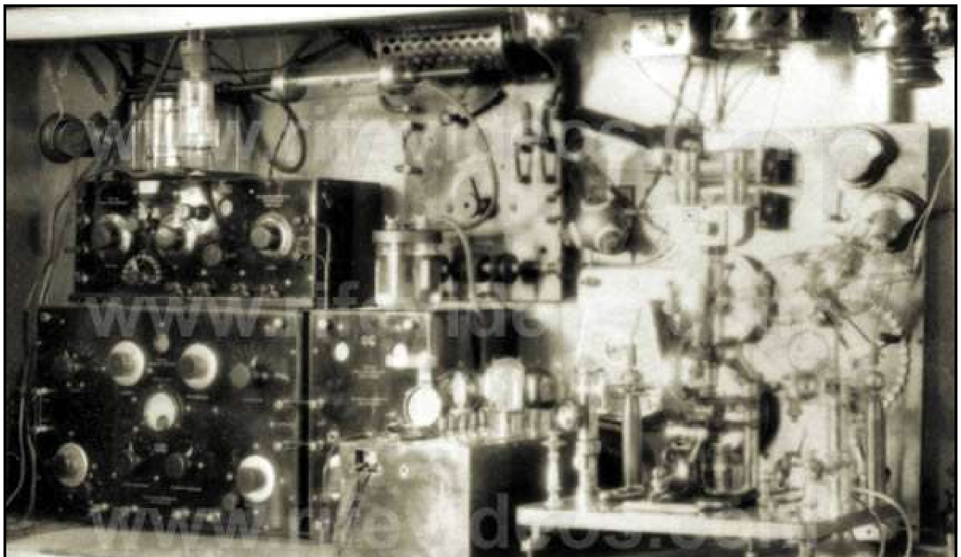

www.rife.org

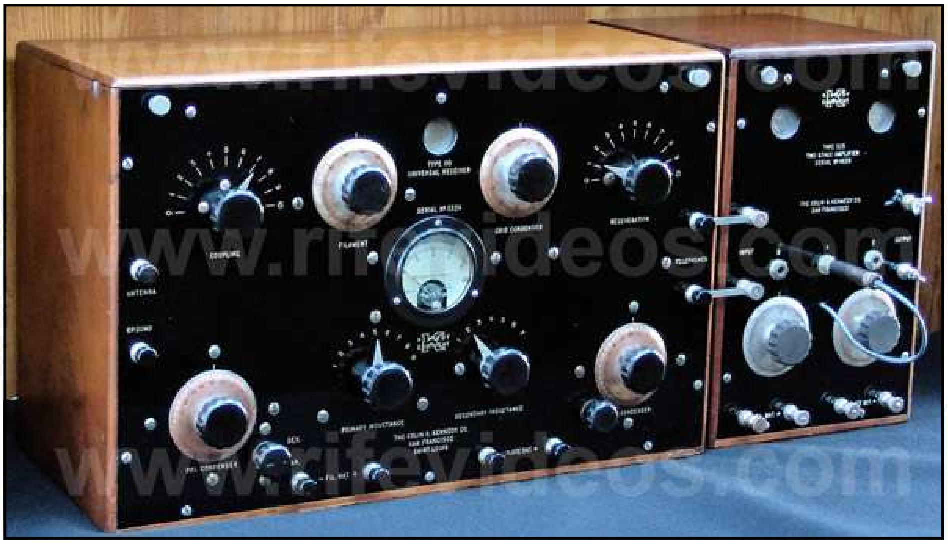

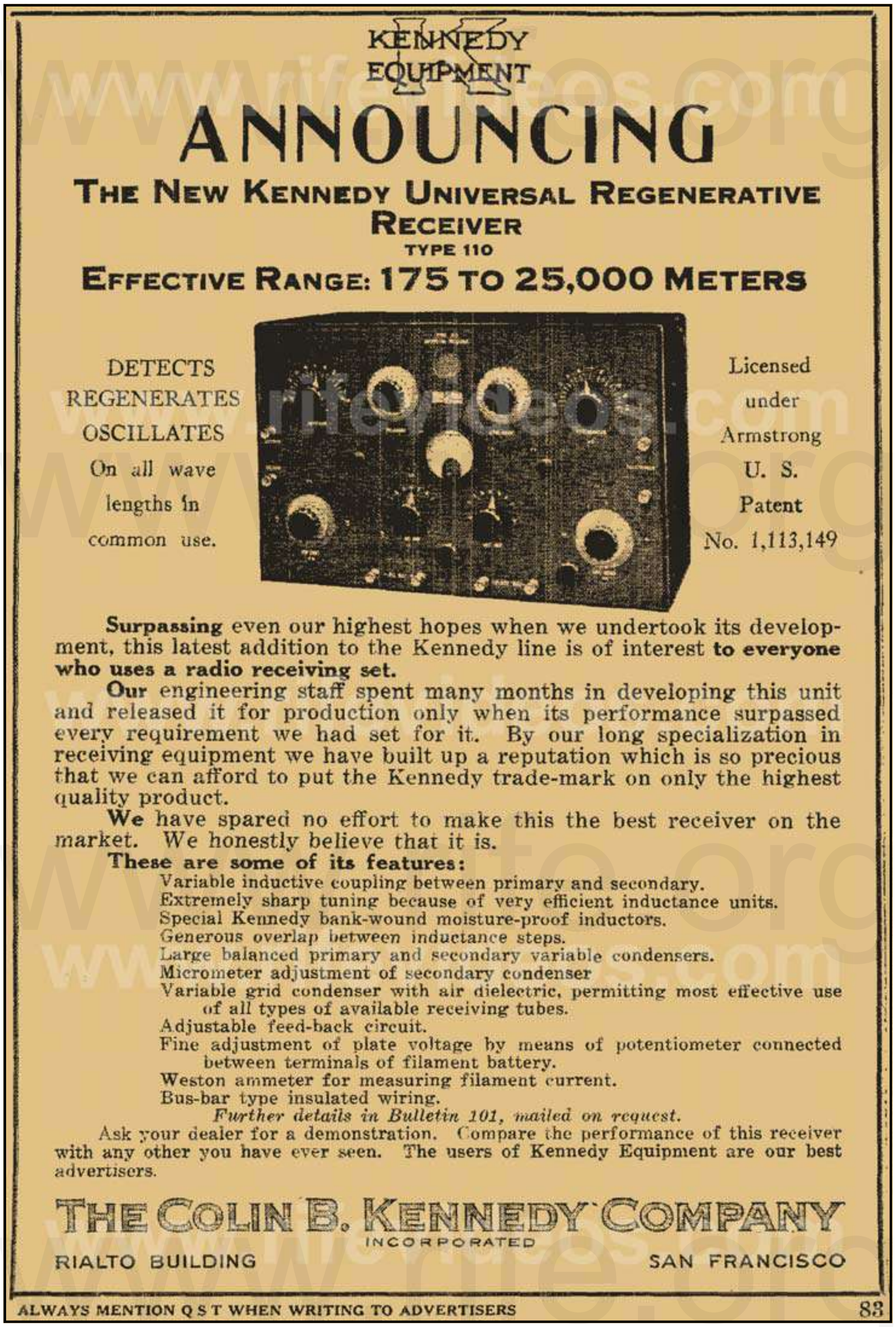

KENNEDY
EQUIPMENT

ANNOUNCING

THE NEW KENNEDY UNIVERSAL REGENERATIVE RECEIVER

TYPE 110

EFFECTIVE RANGE: 175 TO 25,000 METERS

DETECTS
REGENERATES
OSCILLATES
On all wave
lengths in
common use.

Licensed
under
Armstrong
U. S.
Patent
No. 1,113,149

Surpassing even our highest hopes when we undertook its development, this latest addition to the Kennedy line is of interest **to everyone who uses a radio receiving set.**

Our engineering staff spent many months in developing this unit and released it for production only when its performance surpassed every requirement we had set for it. By our long specialization in receiving equipment we have built up a reputation which is so precious that we can afford to put the Kennedy trade-mark on only the highest quality product.

We have spared no effort to make this the best receiver on the market. We honestly believe that it is.

These are some of its features:

Variable inductive coupling between primary and secondary.
Extremely sharp tuning because of very efficient inductance units.
Special Kennedy bank-wound moisture-proof inductors.
Generous overlap between inductance steps.
Large balanced primary and secondary variable condensers.
Micrometer adjustment of secondary condenser
Variable grid condenser with air dielectric, permitting most effective use of all types of available receiving tubes.
Adjustable feed-back circuit.
Fine adjustment of plate voltage by means of potentiometer connected between terminals of filament battery.
Weston ammeter for measuring filament current.
Bus-bar type insulated wiring.

Further details in Bulletin 101, mailed on request.

Ask your dealer for a demonstration. Compare the performance of this receiver with any other you have ever seen. The users of Kennedy Equipment are our best advertisers.

THE COLIN B. KENNEDY COMPANY
INCORPORATED
RIALTO BUILDING SAN FRANCISCO

ALWAYS MENTION Q S T WHEN WRITING TO ADVERTISERS 83

The Kennedy Model 110 could actually go from 12,000 Hertz to about 1,800,000 Hertz. The Kennedy Company was just being conservative in its advertisement. The next instrument that was on top of the Receiver Model 110 in Dr. Rife's laboratory photo is the Kennedy Short-Wave Regenerative Receiver Model 281. Below is a photo of the Kennedy Receiver Model 281. Below the 281 is a picture of the 1923 Kennedy 281 advertisement.

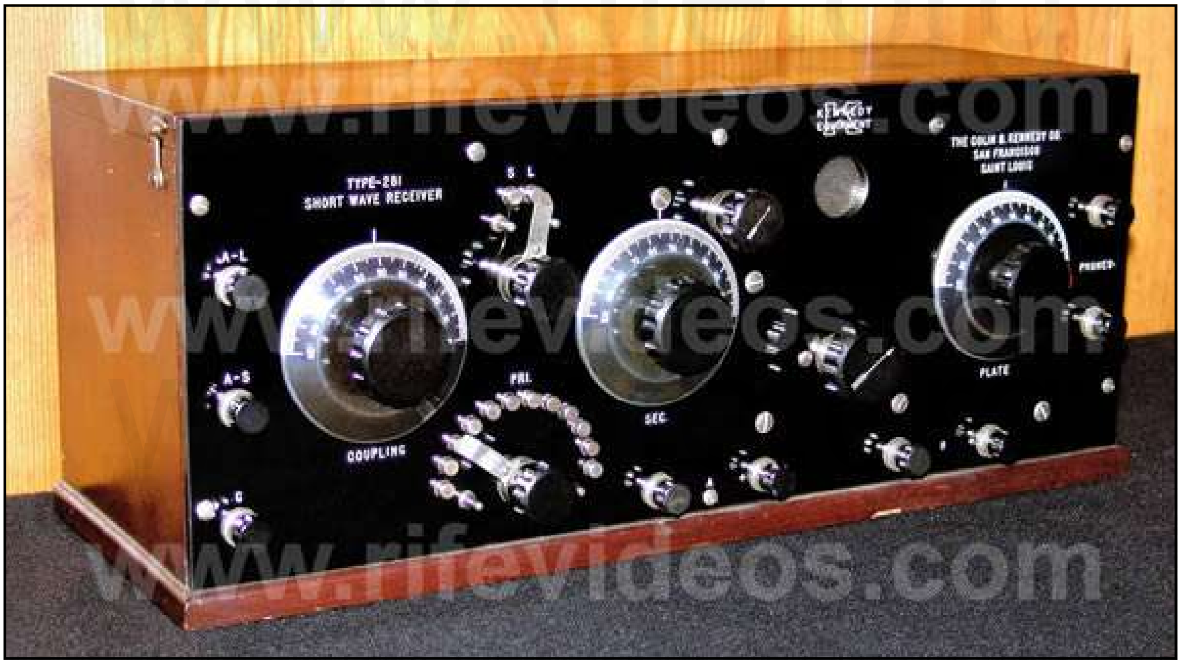

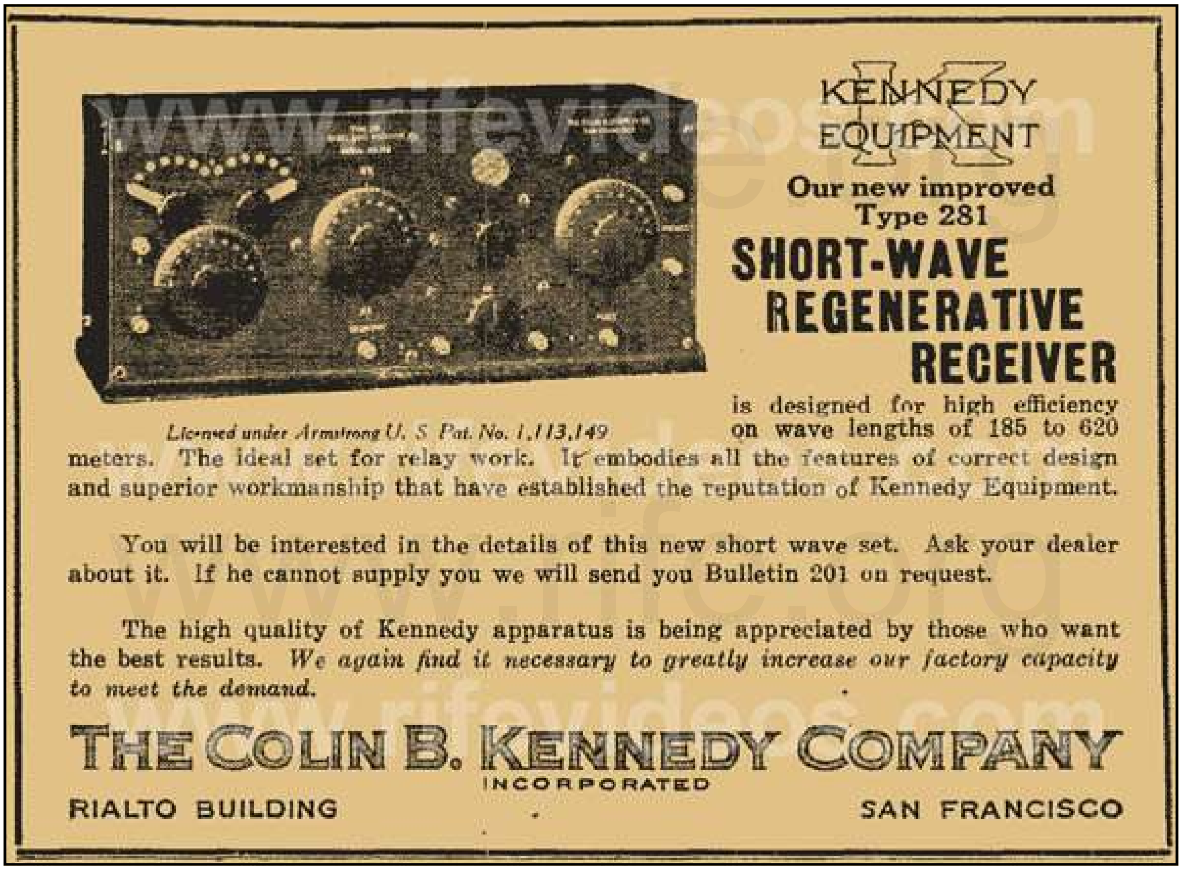

KENNEDY EQUIPMENT

Our new improved
Type 281

**SHORT-WAVE
REGENERATIVE
RECEIVER**

Licensed under Armstrong U. S. Pat. No. 1,113,149

is designed for high efficiency on wave lengths of 185 to 620 meters. The ideal set for relay work. It embodies all the features of correct design and superior workmanship that have established the reputation of Kennedy Equipment.

You will be interested in the details of this new short wave set. Ask your dealer about it. If he cannot supply you we will send you Bulletin 201 on request.

The high quality of Kennedy apparatus is being appreciated by those who want the best results. *We again find it necessary to greatly increase our factory capacity to meet the demand.*

THE COLIN B. KENNEDY COMPANY
INCORPORATED
RIALTO BUILDING SAN FRANCISCO

This Kennedy 281 instrument had an effective range from 185 meters to 620 meters or from 483,000 Hertz to 1,620,000 Hertz. This instrument could actually go to about 1,800,000 Hertz. The Kennedy Company again being conservative on its frequency range.

www.rife.org

Above in the original lab photo which showed all of Dr. Rife's equipment stacked up on a table we see another Kennedy Regenerative Receiver, this being the Kennedy Model 220. The first photo, shown below, is a photo of this instrument. Its effective frequency range was from 175 meters to 3250 meters or from 92,000 Hertz to 1,700,000 Hertz. It could also go to about 1,800,000 Hertz. All three models of this Kennedy equipment could go higher in frequency than the advertisements state. Below the Kennedy Model 220 photo we see the 1923 advertisement for the Kennedy 220 instrument.

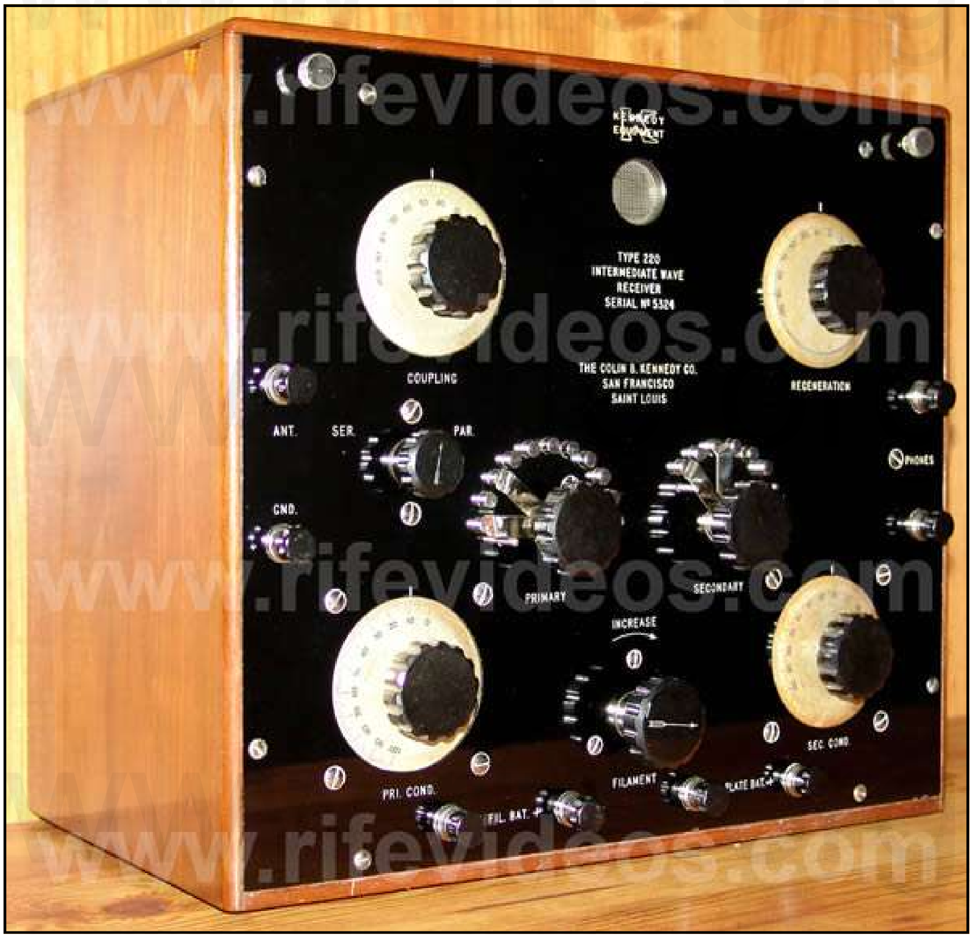
KENNEDY EQUIPMENT
TYPE 220
INTERMEDIATE WAVE
RECEIVER
SERIAL No 5324
THE COLIN B. KENNEDY CO.
SAN FRANCISCO
SAINT LOUIS
COUPLING
REGENERATION
ANT.
SER.
PAR.
PHONES
GND.
PRIMARY
SECONDARY
INCREASE
PRI. COND.
FIL. BAT. +
FILAMENT
PLATE BAT. +
SEC. COND.

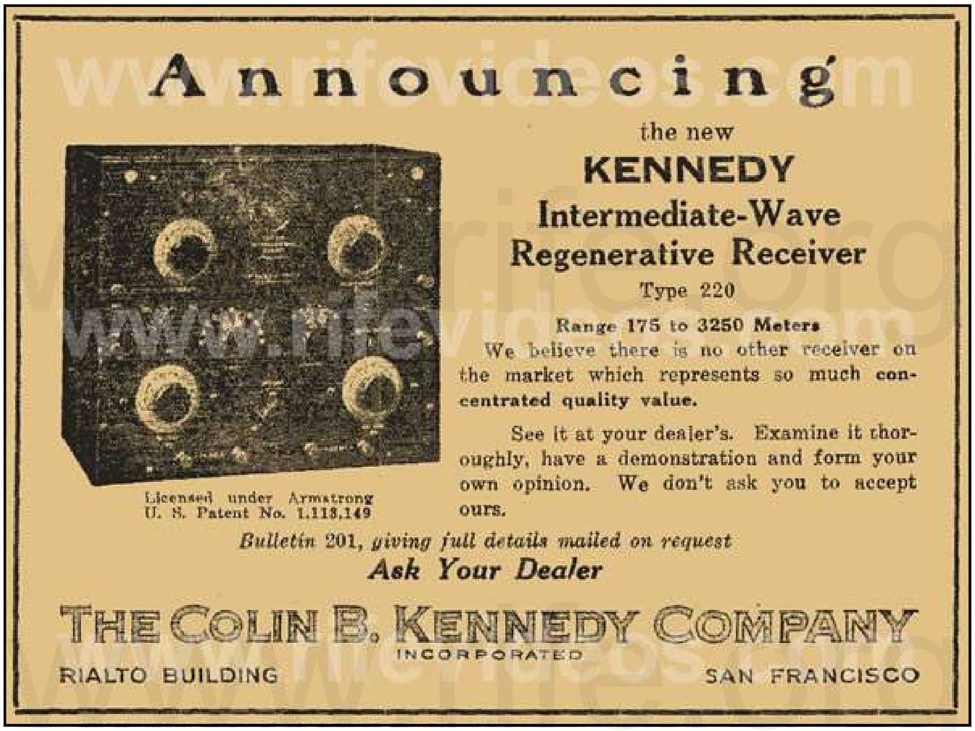

Announcing

the new

KENNEDY

Intermediate-Wave
Regenerative Receiver

Type 220

Range 175 to 3250 Meters

We believe there is no other receiver on the market which represents so much **concentrated quality value.**

See it at your dealer's. Examine it thoroughly, have a demonstration and form your own opinion. We don't ask you to accept ours.

Licensed under Armstrong
U. S. Patent No. 1,113,149

Bulletin 201, giving full details mailed on request

Ask Your Dealer

THE COLIN B. KENNEDY COMPANY
INCORPORATED
RIALTO BUILDING SAN FRANCISCO

www.rife.org
www.rifevideos.com

Now that we have all the frequency generating equipment identified we can now come to some conclusions. All of this Kennedy equipment was sine wave. Square wave was not used or even generated in this old equipment. The Kennedy Receiver Model 110 had a frequency range from 12,000 to about 1,800,000 Hertz. This shows that Dr. Rife's instruments had the ability to output audio frequencies, a fact that he mentioned in his 1961 deposition. The only audio frequencies he could have used would have been modulated from this equipment. Dr. Rife also mentioned in his 1961 deposition that he balanced the audio on a carrier which would have been a modulated waveform.

What is really surprising is the fact that none of the Kennedy equipment that Dr. Rife used could output a frequency higher than about 2 Megahertz (MHz). This fact changes a lot of things with regard to his lab notes dated before 1934. It was impossible for him to produce 11,780,000 Hertz or 17,033,000 Hertz using this equipment. These are the two frequencies that Dr. Rife listed on his lab notes for the BX cancer virus. The frequency range of the Kennedy equipment now explains why Dr. Rife's Engineer, Philip Hoyland, told Mr. Edwards, a business partner in Beam Ray Corporation, that Dr. Rife and not taken some factors into consideration when he read his frequencies prior to 1935. What this implies is that Dr. Rife may have misread his frequencies prior to 1935.

With Dr. Rife's approval, Philip Hoyland was hired by Dr. Milbank Johnson, M.D. and the University of Southern California Special Medical Research Committee in 1935 to build a more up-to-date portable frequency instrument to be used for their research. Dr. Rife's 1934 instrument was cumbersome because it was not just one, but several, pieces of equipment which were difficult to move and use. In order to build the new instrument, Philip Hoyland needed to know what frequencies Dr. Rife was using.

This is where the story gets interesting. Dr. Rife had many lab notes which had the frequencies written down on them for the various microorganisms. Dr. Rife could have just given Philip Hoyland a copy of the frequency ranges that the lab notes covered and he could have built the instrument from that information. But this is not what happened. Philip Hoyland brought his standard master oscillator into the laboratory and then Dr. Rife and Philip Hoyland went through the long process of putting many organisms under the microscope and checking to see what the frequency was when it was devitalized. If Dr. Rife had been confident that his original frequencies were correct on his lab notes this testing would not have been necessary. It is apparent that there were probably some questions in Dr. Rife's mind about the accuracy of his initial readings and frequencies which made this retesting of the organisms M.O.R. frequencies necessary.

Keeping this in mind, it was difficult to read the correct frequencies prior to this time unless you were very proficient at doing it. Philip Hoyland apparently wanted to know exactly what frequencies Dr. Rife was using in order to build the new instrument. While testifying on the stand in the 1939 Beam Ray trial, Philip Hoyland stated this about how he obtained the frequencies: (1939 Beam Ray Trial Transcript #778)

HOYLAND: *"They were taken off the last machine* [the Kennedy equipment] *that was built by Dr. Rife. I transferred them from one machine to another."*

At another point during the trial the transcript reads as follows. (#905-916):

COMPARET: *"In June of 1935 was when you made an agreement with the* [transcript missing words] *medical research to build a Rife Ray machine,* [the Rife Ray #4] *you did build it soon after that?"*

HOYLAND: *"Yes."*

www.rife.org

COMPARET: *“You had an agreement with them that all work was to be done under Dr. Rife’s direction?”*

HOYLAND: *“That’s what the contract called for.”*

COMPARET: *“Did you do this work without getting the frequencies from Dr. Rife?”*

HOYLAND: *“I recalibrated the machine according to the bacteria.”*

COMPARET: *“What specifically did you do that constituted this recalibration?”*

HOYLAND: *“I used a standard oscillator against his machine to see what frequencies he was using.”*

COMPARET: *“He set his machine and you measured his frequencies?”*

HOYLAND: *“Yes.”*

COMPARET: *“Did you make any memorandum of these particular frequencies?”*

HOYLAND: *“Yes, I gave Dr. Johnson and Dr. Rife a list of them.”*

Later during the trial Dr. Rife was asked where the frequencies came from. (#1290-1293):

JUDGE KELLY: *“When you constructed this Beam Ray machine* [from Kennedy equipment] *you had a dial representing the frequencies or harmonics?”*

RIFE: *“We had many dials on the original machine* [Kennedy Model 110].”

JUDGE KELLY: *“Is that the machine Mr. Hoyland got the frequencies from?”*

RIFE: *“Yes, he took them off that old machine* [Kennedy Model 110].*”*

From the court testimony given by Dr. Rife and Philip Hoyland we see the frequencies were read by Philip Hoyland off of the Rife Ray #3 or Kennedy Model 110 and 281 and used in the next instrument which was the Rife Ray #4 (We will be discussing this instrument next). Now let’s continue on reading the court testimony: (Beam Ray Trial Transcript #1553-1555)

COMPARET: *“Now going back to your assumption that Dr. Rife knew the frequencies, had Mr. Hoyland ever told you that Dr. Rife knew them?”*

EDWARDS: *“No, he told me that Dr. Rife only thought he had them.”*

COMPARET: *“What did you think that meant?”*

EDWARDS: *“Well, Mr. Hoyland told me about that time* [1934 and before], *that Dr. Rife measured the frequencies only by the length of the wire and that he did not take other factors into consideration.”*

Here in the court testimony we just read that Dr. Rife had not read his frequencies correctly when he measured them. This would have been a mistake easy to make in the 1920’s and 1930’s. The frequencies which Philip Hoyland read off of Dr. Rife’s #3 instrument, which consisted of the Kennedy equipment, were different from the earlier lab note frequencies recorded by Dr. Rife. This has caused a lot of confusion because the frequencies that Philip Hoyland read were all lower than 2,000,000 Hertz.

www.rife.org

Dr. Rife had written down on his lab notes frequencies as high as 11,780,000 and 17,033,000 Hertz for the BX cancer virus. However, the Kennedy Models 110, 220 and 281 could not output these high frequencies. It is apparent that Philip Hoyland was absolutely correct when he told Mr. Edwards that Dr. Rife had misread his frequencies. Also, Philip Hoyland testified in court that he gave both Dr. Rife and Dr. Johnson a list of the correct frequencies he read off of the Kennedy Model 110. This verifies the truth of what Philip Hoyland said in court.

www.rife.org

There is another verification that Dr. Rife had misread his frequencies. On the Rife audio CD's, Henry Siner, Dr. Rife's lab assistant, read from a lab note of the BX cancer virus. All the information was the same as Dr. Rife's earlier pre-1935 lab notes except the frequencies. On that corrected lab note Henry Siner read 187 meters for the wave length and 1,604,000 Hertz for the cycles per second frequency for the BX cancer virus. Both the meter frequency and the cycles per second measurement being the same frequency. However, on the pre-1935 lab note, both were different. One frequency was 11,780,000 and the other was 17.6 meters or 17,033,000 Hertz. Henry Siners reading of this corrected lab note also verifies that Dr. Rife had not read his frequencies correctly. At the end of that discussion about the BX cancer virus and the Lab note Henry Siner made this statement, quote:

SINER: *"That was a long time ago, but, and remember, I was just copying what he* [Dr. Rife] *dictated."* (John Marsh Rife CD's - MP3 track 11).

www.rife.org

This quote from Henry Siner shows that it was Dr. Rife who made the corrections to the lab note. The frequency of 1,604,000 Hertz was the frequency Philip Hoyland read and gave to Dr. Rife and Dr. Johnson and it was used in the new instrument built in 1935 called the Rife Ray #4.

There is one thing we need to consider. Dr. Rife could have read a harmonic of the frequency instead of the correct frequency. It appears this is in fact what Dr. Rife did. Dr. Rife understood how easy it was to read a harmonic frequency instead of the correct frequency and recognized that he may not have had true fundamental frequencies. He stated:

www.rife.org

RIFE: *"I've talked to you* [John Crane] *and Verne* [Verne Thompson] *and other people too that there may be some of the frequencies that we are using that may be harmonics, you know...It's not an impossibility that some of those frequencies may be a harmonic. We may not know the true frequencies of some of them. But it does the business. Maybe if we had the true frequency it would do it better because it has more power than a harmonic."* (John Marsh Rife CD's - CD 7 track 2).

The frequency that Philip Hoyland read off of Dr. Rife's 1934 Rife Ray #3 Machine was 1,604,000 Hertz. Dr. Rife had written two frequencies down on his pre-1934 lab notes. One was 11,780,000 Hertz and the other was 17,033,000 Hertz. The seventh harmonic of 1,604,000 is 11,228,000 which is close to the 11,780,000 especially if you consider that Dr. Rife was not reading his frequencies correctly. We now know Dr. Rife was not even reading the harmonic correctly. Now the eleventh harmonic of 1,604,000 is 17,644,000 which is close also to 17,033,000 Hertz. Had Dr. Rife read the frequencies correctly then both the meter frequency and the cycles per second frequency should have been the same frequency. This was the case with the new lab note when it was corrected by Dr. Rife and read by Henry Siner in the 1950's. The evidence is absolutely overwhelming that Dr. Rife was not reading his frequencies correctly because the frequencies Philip Hoyland read were used in the next Rife Machine which was called the Rife Ray #4. In the space of about 60 days all Dr. Rife's frequencies changed from from the pre-1935 lab note frequencies to the new frequencies that were used with the Rife Ray #4.

www.rife.org

We wondered where these harmonics that Dr. Rife read might of come from. Did the Kennedy Model 110 have harmonics in its waveform? Did it output a sine wave waveform? Was the waveform distorted? The only way to answer these questions was to find a working Kennedy 110 and put it on a

www.rife.org

spectrum analyzer. Jason Ringas of the Rife Research Group of Canada and I contacted Henry Rogers the owner of the Western Historic Radio Museum (www.radioblvd.com) who owns two Kennedy 110s that are still operational. Henry Rogers knew nothing about Dr. Rife but agreed to let me come visit his location to check the readings of the Kennedy Model 110. He also owns a Kennedy 220 and a Kennedy Model 281, both of which are also in working condition. The Kennedy Company built top-of-the-line equipment and we were surprised to find out even after over 80 years, they still worked as well as they did when they were new. Very little attention is ever needed to get these instruments back in working condition because of the quality of their construction. So with spectrum analyzer in hand, I went to see Henry Rogers and we put the Kennedy 110 on the spectrum analyzer to get the answers to our questions. Below is the reading of the waveform of the Kennedy Model 110 at 417,000 Hertz using a PicoScope 3205 spectrum analyzer. On the left is the waveform which proves that Dr. Rife was using sine wave.

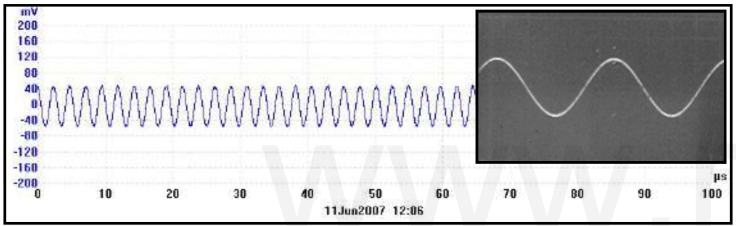

Kennedy Model 110 at 417,000 Hertz

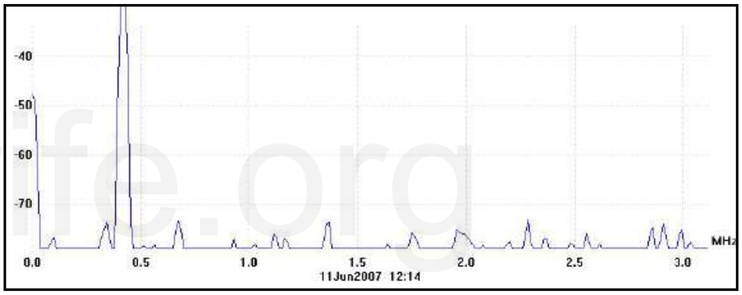

That question was finally answered. The spectrum analyzing of the frequency revealed that there were no harmonics in the waveform. The noise which shows up as little spikes are from the power supply. These old receivers ran on batteries and when they are hooked up to batteries the noise in the circuit is greatly reduced. The amazing thing about the Kennedy Model 110 sine wave waveform was that it was picture perfect. This amazed us because everyone believed that the equipment that Dr. Rife used would have had a distorted waveform. No one that I have ever talked with believed that this old equipment was capable of producing a nearly-perfect waveform. It was as good as we can do today with our sophisticated modern frequency generating equipment. The fact that it produced no harmonics also amazed us. Below are the readings of the Kennedy Model 110 at 770,000 and 1,604,000 Hertz. At 1,604,000 Hertz the sine wave was still nearly perfect and it did not produce any harmonics. We checked all frequencies out to 50 Megahertz for harmonics and found none.

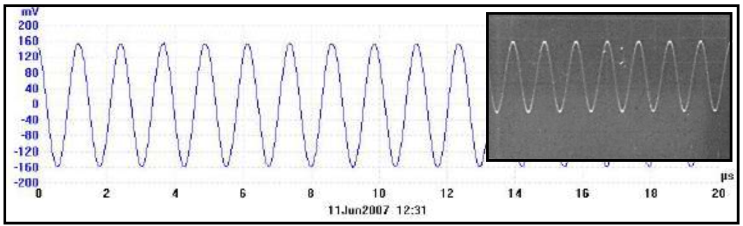

Kennedy Model 110 at 770,000 Hertz

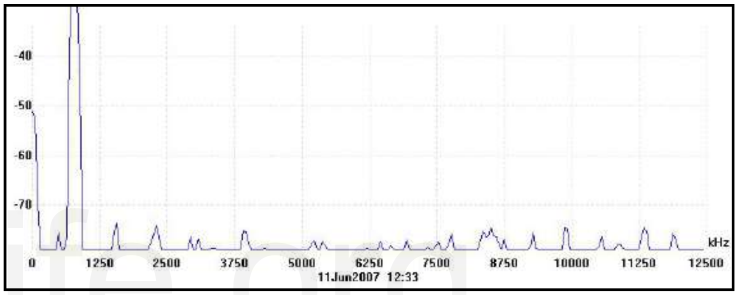

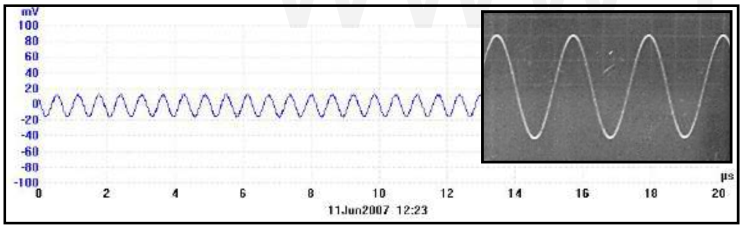

Kennedy Model 110 at 1,604,000 Hertz

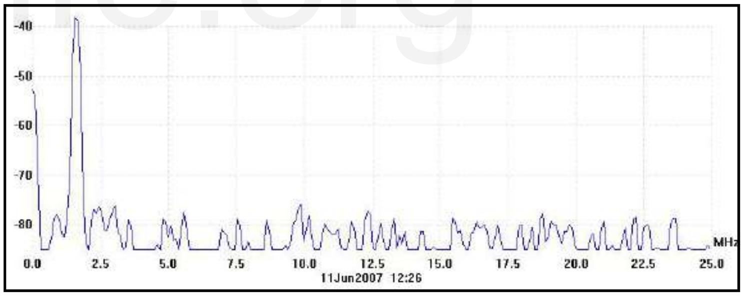

This testing showed that Dr. Rife's equipment, call a Rife Machine, output a sine wave waveform with no harmonics. So where did the frequencies come from that Dr. Rife read and recorded on his old lab notes? Why did he record two frequencies in his lab notes? We now knew what equipment he

www.rife.org

used. His pre-1935 lab notes just didn't make any sense. We know from Henry Siner's reading of the corrected BX lab note that the meter frequency and the cycles per second frequency should be the same. It is apparent that Dr. Rife used two different pieces of equipment to read his frequencies. One piece of equipment gave a reading in meters and the other piece of equipment gave a reading in cycles per second. However even knowing this did not explain where the harmonics came from.

We knew that the noble gas he used in his ray tube could double the frequency that went through it. These types of tests have been done with plasma in laboratories in the past. So we decided to make some tests. We tested the Icom 718 which we hooked up to a phanotron ray tube. This is the type of ray tube Dr. Rife used and is the only one we tested. We first tested to see what the sine wave looked like coming out of the Icom 718. We wanted to make sure that it did not produce any harmonics, and in fact, our testing showed it did not produce any harmonics. Then we hooked it up to the antenna tuner to see if the tuner distorted the waveform and produced any harmonics. We found it did not distort the waveform or produce harmonics through the antenna tuner except at 1,604,000 Hertz. This is only because the Icom is not supposed to output a frequency below 2,000,000 Hertz. Below this frequency it will produce two harmonics (see graph on page 37). The other two frequencies we tested were 11,780,000 and 17,033,000 Hertz. These were the frequencies Dr. Rife recorded on his pre-1935 lab notes and neither of these produced harmonics through the antenna tuner. Then we put it through the ray tube. The ray tube didn't just double the frequency - it also produced all the harmonics that Dr. Rife would have read. We now had the answers as to where the harmonics came from. The ray tube produces the harmonics. You can put a harmonic-free sine wave through a ray tube and get all the harmonics that Dr. Rife recorded on his lab notes. The photo below is the Icom 718 and on the next three pages are three graphs that show the readings taken in this testing.

www.rife.org

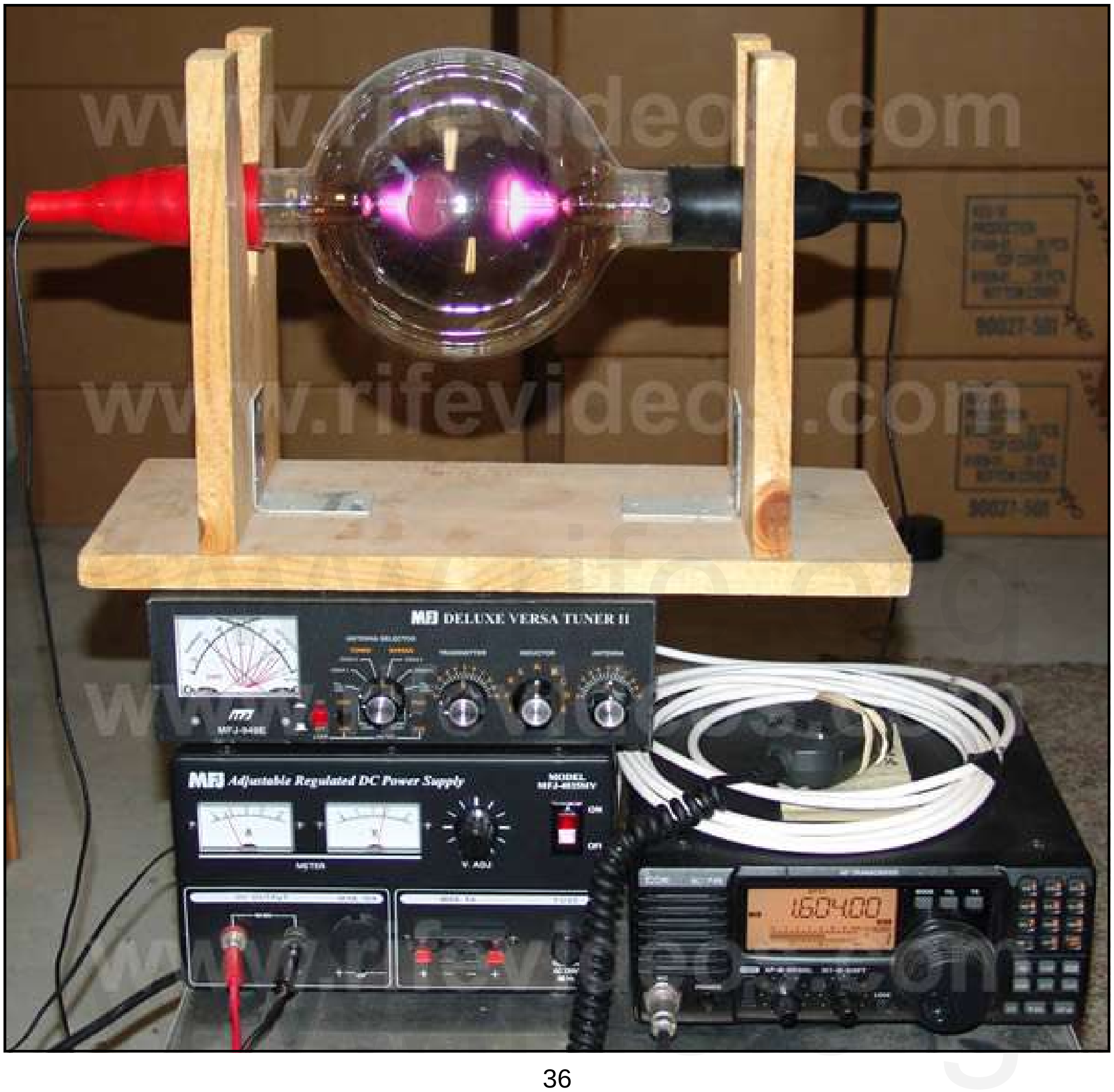
www.rifevideos.com
MFJ DELUXE VERSA TUNER II
MFJ-949E
MFJ Adjustable Regulated DC Power Supply
1.604.00

Testing done with PicoScope 3205 spectrum analyzer at 1,604,000 Hertz using Icom 718

www.rife.org

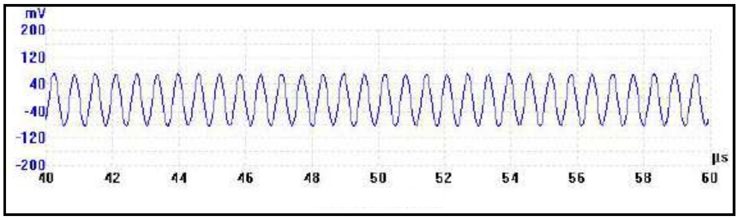

Sine wave out of Icom 718 at 1,604,000 Hertz.

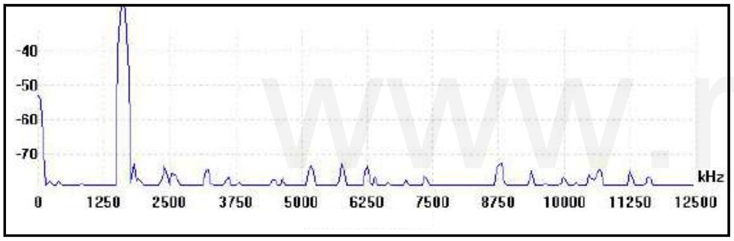

Icom 718 at 1,604,000 Hertz measured with spectrum analyzer showing no harmonics.

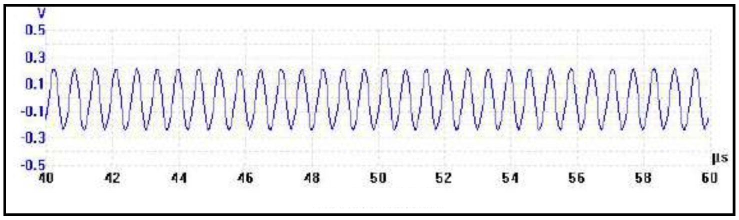

Sine wave out of Icom 718 at 1,604,000 Hertz using the antenna tuner.

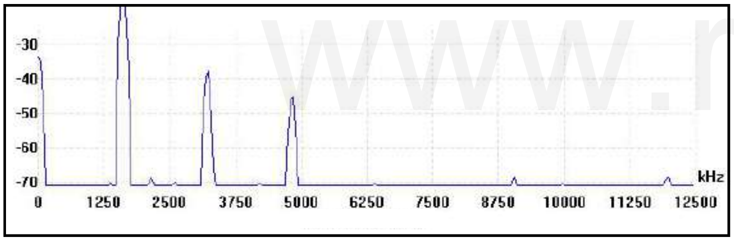

Icom 718 and antenna tuner at 1,604,000 Hertz measured with spectrum analyzer showing two harmonics. These two harmonics are only produced because the Icom is not designed to go below 2,000,000 Hertz. If you output 2,000,000 Hertz it produces no harmonics.

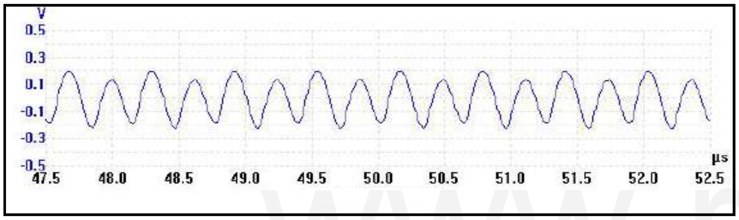

Sine wave out of Icom 718 at 1,604,000 Hertz using antenna tuner and ray tube. Sine wave is distorted. In all tests done the sine wave was always distorted when put through a ray tube.

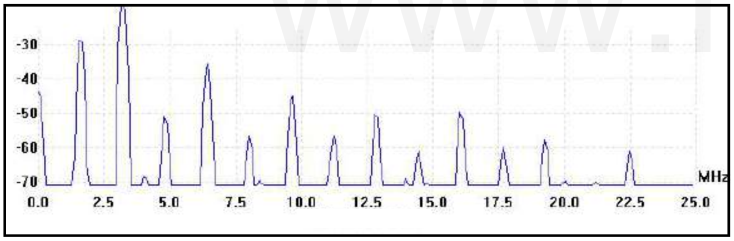

Icom 718 at 1,604,000 Hertz using antenna tuner and ray tube. Measured with spectrum analyzer showing harmonics all the way up to 22,000,000 Hertz. This shows that Dr. Rife's Kennedy Model 110 which only had a top range of 2,000,000 Hertz did produce harmonic frequencies in the 11,000,000 and 17,000,000 Hertz range.

www.rife.org

Below are the measurements taken with the PicoScope 3205 spectrum analyzer from the Icom 718 using the antenna tuner and ray tube at 11,780,000 Hertz. This was the first frequency Dr. Rife listed on his pre-1934 lab notes which was later changed to 1,604,000 Hertz.

www.rife.org

Testing done with PicoScope 3205 spectrum analyzer at 11,780,000 Hertz using Icom 718

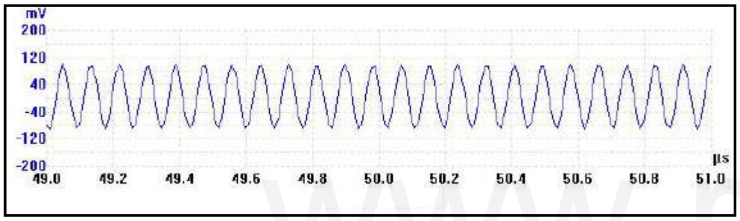

Sine wave out of Icom 718 at 11,780,000 Hertz.

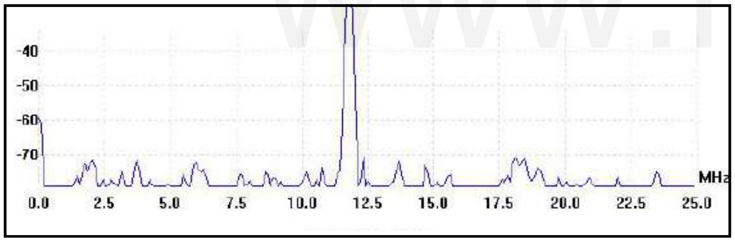

Icom 718 at 11,780,000 Hertz measured with spectrum analyzer showing no harmonics.

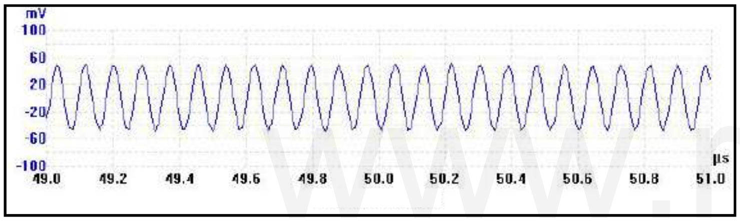

Sine wave out of Icom 718 at 11,780,000 Hertz using the antenna tuner.

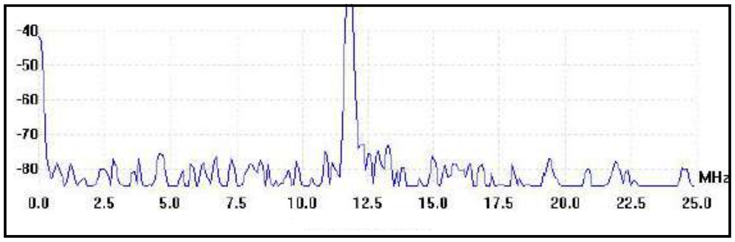

Icom 718 and antenna tuner at 11,780,000 Hertz measured with spectrum analyzer showing no harmonics.

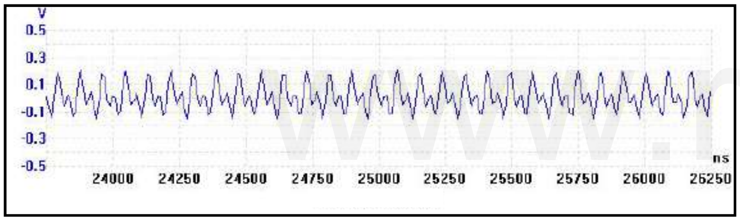

Sine wave out of Icom 718 at 11,780,000 Hertz using antenna tuner and ray tube. The sine wave does not look like a sine wave. In all cases the sine wave is distorted to some degree when put through a ray tube.

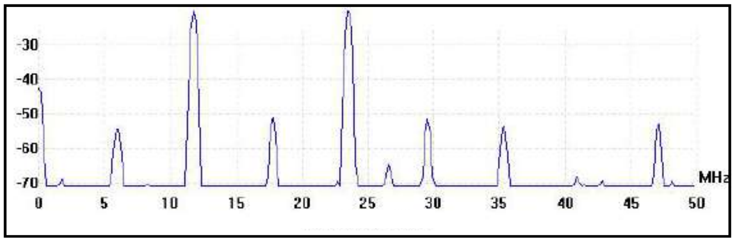

Icom 718 at 11,780,000 Hertz using antenna tuner and ray tube. Measured with spectrum analyzer showing harmonics all the way up to 50 MHz.

Below are the measurements taken with the PicoScope 3205 spectrum analyzer from the Icom 718 using the antenna tuner and ray tube at 17,033,000 Hertz. This was the second frequency on his pre-1934 lab notes which was recorded in meters. This was later changed to 187 meters which would give us a frequency of about 1,604,000 Hertz. This confirms that Dr. Rife was just reading a harmonic at 17,033,000.

www.rife.org

Testing done with PicoScope 3205 spectrum analyzer at 17,033,000 Hertz using Icom 718

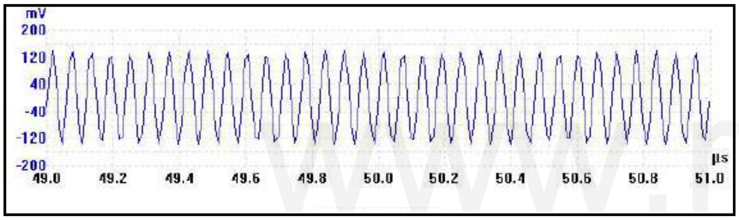

Sine wave out of Icom 718 at 17,033,000 Hertz.
Some distortion was in the sine wave.

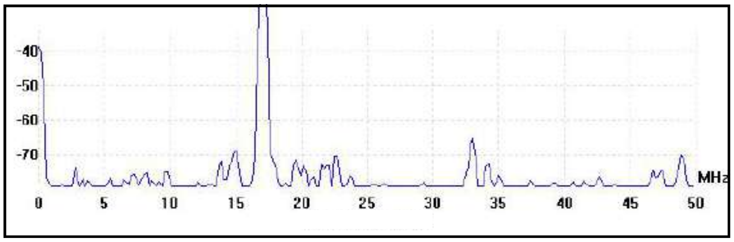

Icom 718 at 17,033,000 Hertz measured to 50 MHz with spectrum analyzer showing no harmonics.

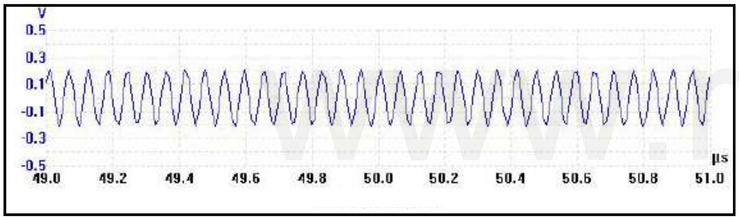

Sine wave out of Icom 718 at 17,033,000 Hertz
using the antenna tuner. Same slight distortion noticed.

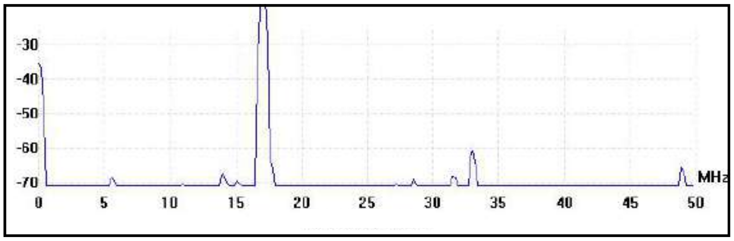

Icom 718 and antenna tuner at 17,033,000 Hertz measured to 50 MHz with spectrum analyzer showing no harmonics.

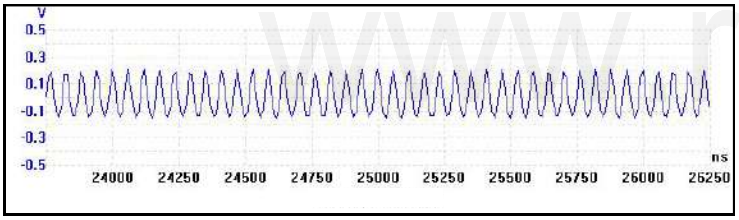

Sine wave out of Icom 718 at 17,033,000 Hertz using antenna tuner and ray tube. Sine wave was distorted even more when put through a ray tube.

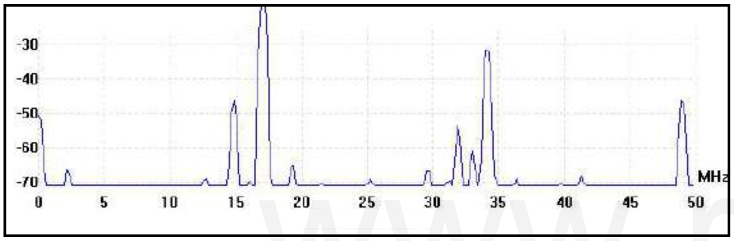

Icom 718 at 17,033,000 Hertz using antenna tuner and ray tube. Measured with spectrum analyzer showing harmonics all the way up to 50 MHz.

Westinghouse RC Receiver & Amplifier

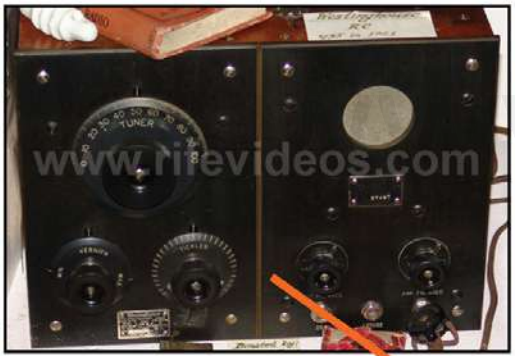

Western Electric Audio Amplifier 7A

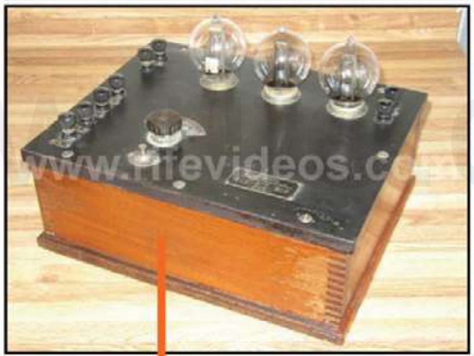

Kennedy Model 281 Receiver

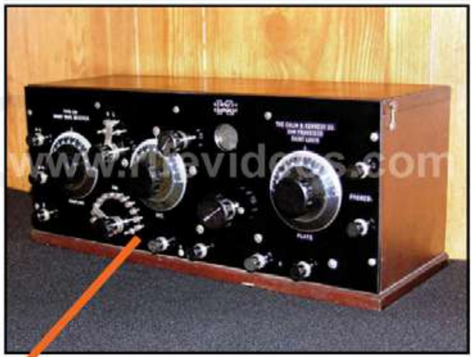

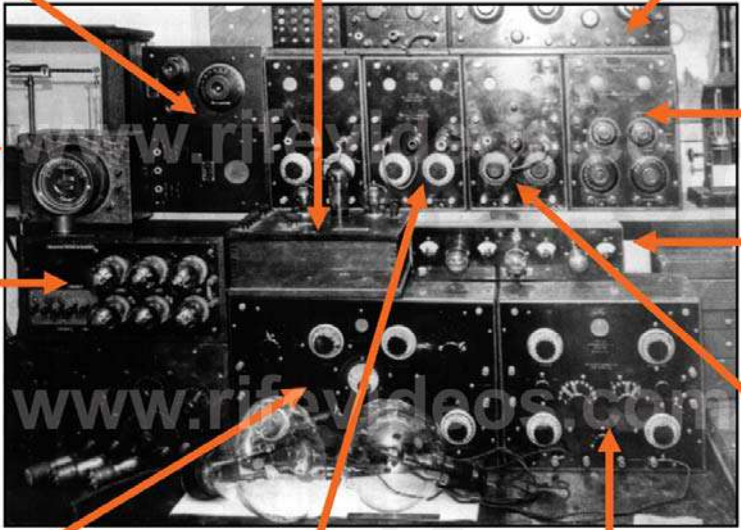

Photo of Rife's equipment that he used in his lab for doing M.O.R. work.

Five stage 50 watt class A cascade RF amplifier

The three photos of Rife's lab on this page are courtesy of the Rife Research Group of Canada.

Kennedy RF Amplifier

Remler 700 I.F. Amplifier 3.5 MHz

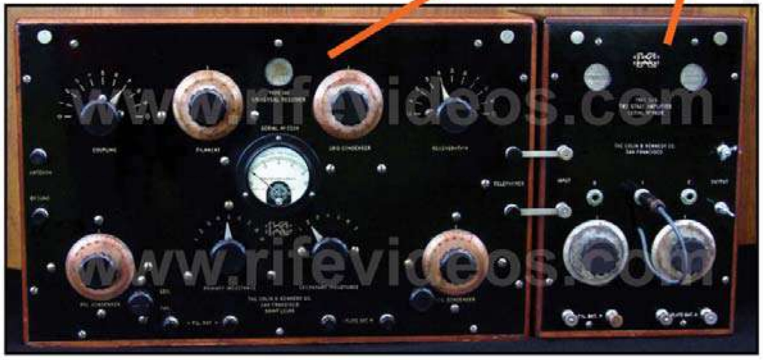

Kennedy Model 110 and 525 Audio Amplifier

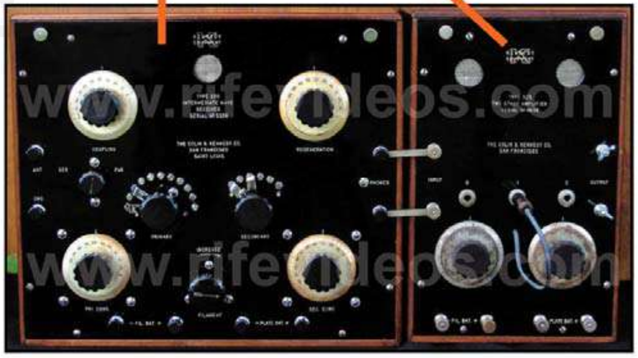

Kennedy Model 220 and 525 Audio Amplifier

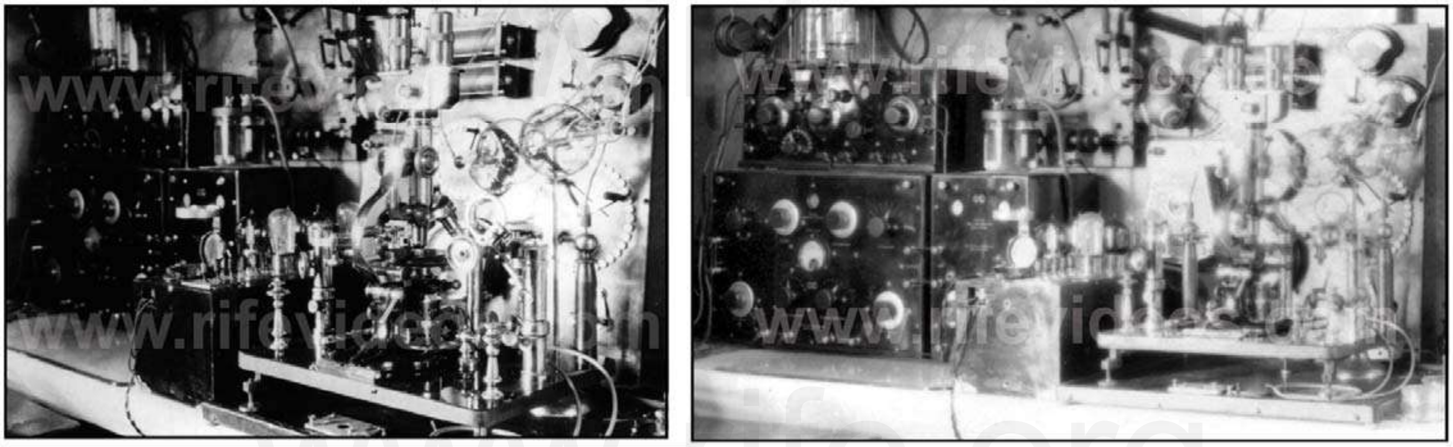

Kennedy Model 110, 281 and 525 Audio Amplifier Set Up In Rife's Lab For Doing M.O.R Work

www.rifevideos.com
www.rife.org

After having done all this spectrum analysis testing we now know how Dr. Rife misread his frequencies. The ray tube gave him the harmonics that he read. Also, he evidently did not read the harmonics correctly. Philip Hoyland read the frequencies correctly because he was an electronics engineer and had the ability to read the frequencies properly. We wish to mention that we do not feel this in any way diminishes or questions the brilliance of Dr. Rife. Even Dr. Rife himself said he was not an electronics man and never claimed to be one. He made a mistake that any untrained person could have easily made.

Having said this, let's move on to the facts. Philip Hoyland read 1,604,000 Hertz for the frequency of the BX cancer virus. Dr. Rife corrected his lab notes to this frequency. This frequency was used in the later Rife Ray #4 instrument. With these documented facts, we now know what must have happened. Dr. Rife read the seventh harmonic of 1,604,000 Hertz and recorded it on his pre-1934 lab notes. The only problem was he was unable to read the seventh harmonic correctly and misread it as 11,780,000 Hertz. It should have been 11,228,000 Hertz because this is the actual harmonic frequency that came out of the ray tube. Dr. Rife had two different pieces of equipment for reading frequencies - one which read in cycles per second and the other which read in meters. These types of meters used to measure wavelengths were common electronic test equipment, just as digital frequency counters are in common use today. Wavelength meters were much harder to use and measure frequencies with if you didn't really understand how to use them. We know that this was the case. Dr. Rife then misread the eleventh harmonic of 1,604,000 Hertz. This harmonic should have been 17,644,000 Hertz instead of the 17,033,000 Hertz. Again, we know from the corrected lab note read by Henry Siner that the cycles per second and meters frequencies should match or be the same. In these early pre-1934 lab notes none of the cycles per second and meter frequencies matched. This shows Dr. Rife used two different pieces of equipment to read the frequencies. The final fact is the Kennedy Company equipment could only output frequencies up to about 1,800,000 Hertz (far below the 11 and 17 MHz range).

When we read the Kennedy Model 110 the instrument was surprisingly accurate. Dr. Rife could have very easily hit the frequency he wanted within the tolerances he gave. He gave "one tenth of one meter" as a gage to show how close you had to be to an organism's M.O.R. At 1,604,000 Hertz this would be 858 Hertz. He said if you were off by this amount the frequency wouldn't work. With that in mind it would be necessary to be within a few hundred Hertz of the BX M.O.R. in order to make sure the frequency was effective. The Kennedy instrument could hit within 200 to 300 Hertz very easily at 1,604,000 Hertz. After changing the dials and then coming back to the same dial settings you could get within 2000 to 6000 Hertz at 417,000 Hertz. This is less than 1% inaccuracy which is quite amazing. Even Philip Hoyland, when he measured the frequencies rounded off all but one frequency to the nearest thousandth. The testing of the Kennedy Model 110 shows that the frequency for the BX is most likely somewhere between 1,600,000 and 1,608,000 Hertz, however it could be as much as 10,000 Hertz plus or minus of 1,604,000 Hertz. All of the frequencies are only close and this should be considered when using them. One fact that helps to point this out is Philip Hoyland read 1,604,000 Hertz for the frequency of the BX. He also gave 187 meters as the frequency. One hundred and eighty seven meters is 1,603,168 Hertz. This is a difference of 832 Hertz and shows why the frequencies are only close. Today's frequency generating equipment is very accurate at hitting a specific frequency but in Dr. Rife's era this was not the case. Dr. Rife's microscope gave him an advantage that we do not have. He could see the organism die.

So now that we know that Dr. Rife's Kennedy Model 110, 220 and 281 only went to 1,800,000 Hertz with harmonics going to about 20,000,000 Hertz (see graph on the top of the next page). We have to ask this question: What frequency is really the true M.O.R? Is it the 1,604,000 Hertz or a harmonic of it? The actual M.O.R. frequency could have been very easily a harmonic, and Dr. Rife would have never known it. The spectrum analyzer graph, second down on the next page, of 1,604,000 Hertz shows it could be any one of these harmonics. Since the ray tube is what produces these harmonics it may be very important to have all these harmonics. Myth Busters, a cable television program did a test

www.rife.org

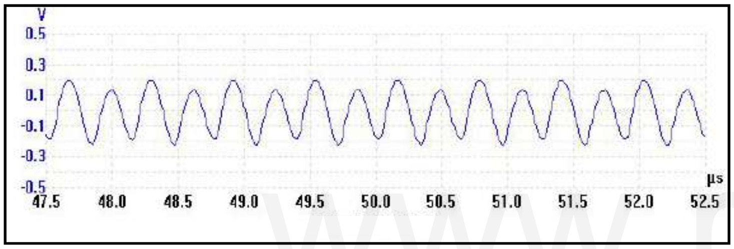

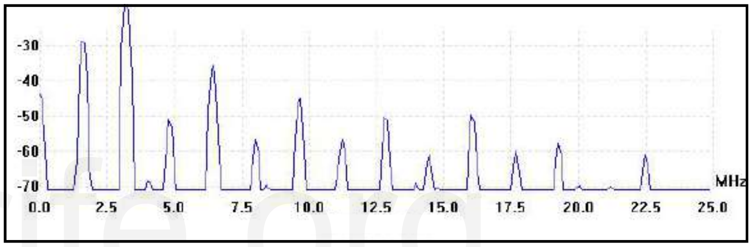

to see if they could break a crystal glass with sound waves. They found when they used only the fundamental frequency without the harmonics they could not break the glass. But when they used the harmonics along with the fundamental frequency then they were able to break the glass. This may or may not be pertinent but it is something that should be considered.

With this in mind we decided to see if there was a way that we could duplicate the harmonics without having to use a ray tube. The below reading with the spectrum analyzer shows that if we distorted the sine wave no more than what the ray tube did we could produce the same harmonics as a ray tube.

GB-4000 at 1,604,000 Hertz using PicoScope 3205 Spectrum Analyzer

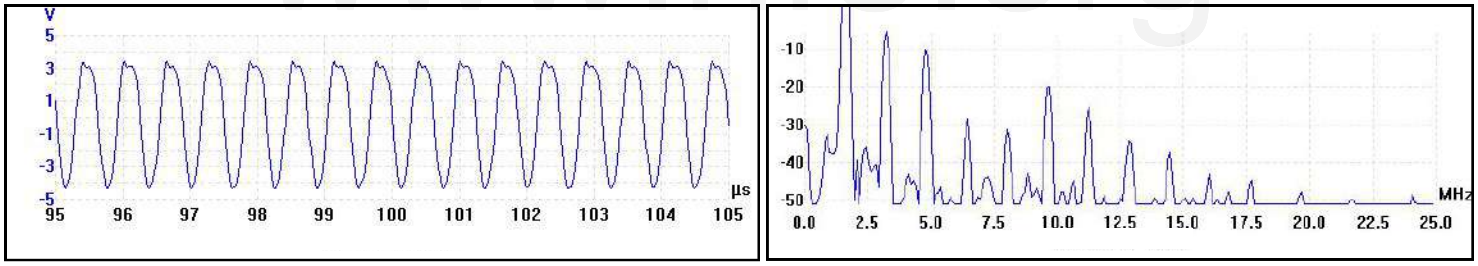

The reading was done at 1,604,000 Hertz taken from an off-the-shelf GB-4000 Function Generator. This test showed it was very easy to duplicate the harmonics produced by a ray tube. We decided to test a triangle wave since the distorted sine wave out of the ray tube resembled it. It also produced the same harmonics as a ray tube. Then we gated an undistorted sine wave and it produced the harmonics. It is apparent that any sine wave frequency from any frequency generator, when gated will produce harmonics.

Inside of Kennedy 281

www.rifevideos.com

www.rife.org

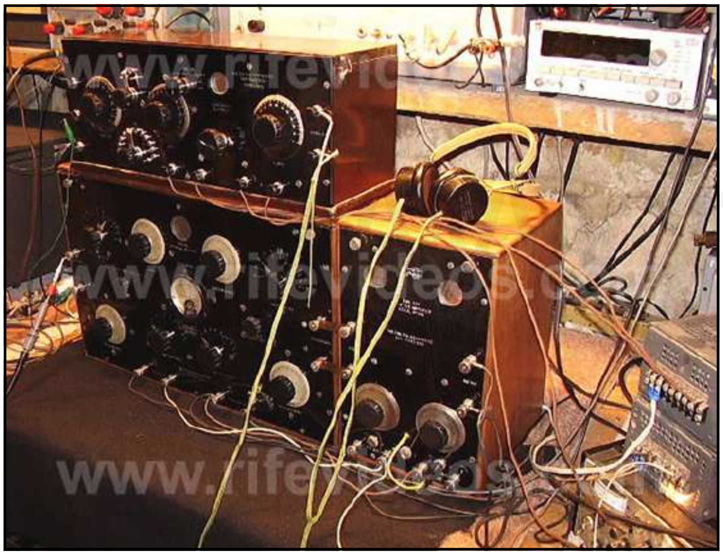

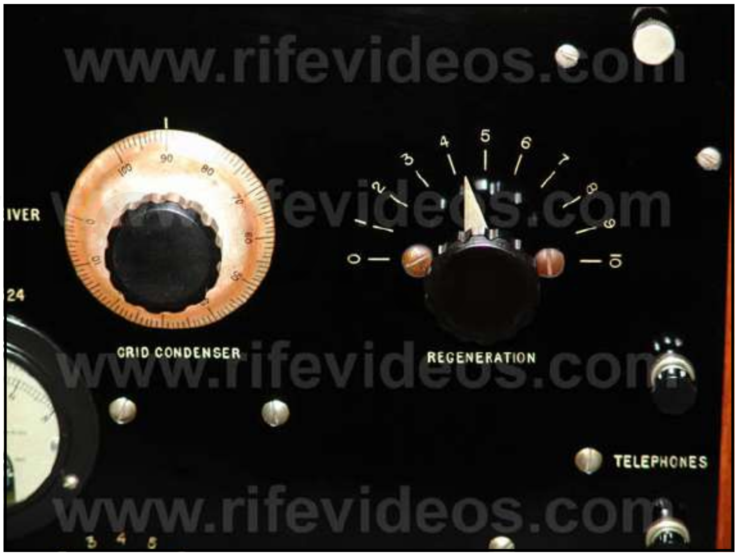

We will now discuss Dr. Rife's tuning of the Kennedy Receiver Model 110 using headphones. In the photo above, on the left, you can see a set of headphones on the Model 525 audio amplifier. Headphones were used to tune the Kennedy Receiver Model 110 and Dr. Rife's earlier instrument that he used before purchasing the Kennedy equipment. When Dr. Rife first tested the audio range of frequencies he would tune his instrument using headphones. Bertrand Comparet, Rife's attorney for the Beam Ray trial of 1939 made this statement when he was interviewed by Dr. John Hubbard:

COMPARET: *"Way back in the old days, way, way back, Rife told me that the way he used to tune his instrument, which in those primitive days was, I guess, garbled. He would hook up headphones and turn the thing. He had a very keen musical sense of pitch and so on, and he would tune it in his headphones until he got the right pitch, and that was the frequency."* (1970's Bertrand Comparet interview #89).

The headphones were used for tuning the audio frequencies in Dr. Rife's early tests when he used loose couplers. The headphones also played an important role in the tuning of the Kennedy Receivers. In the second photo, above on the right, you can see the regeneration dial of the Kennedy Receiver Model 110. When you turned up the regeneration you would listen for clicks or some static in the headphones, this would tell you that the instrument was oscillating. If you turned the regeneration up too high you would hear feedback in the headphones. This feedback meant you did not have a pure sine wave waveform. The photo below, on the left, is the waveform with the feedback from the Kennedy Receiver Model 110. It is a form of audio modulation. The other photo, below on the right, is what is produced when there is no feedback. Dr. Rife always wanted to use a pure waveform. The fact that we could actually listen to the original type of equipment that he used made it so we could understand what Dr. Rife was doing.

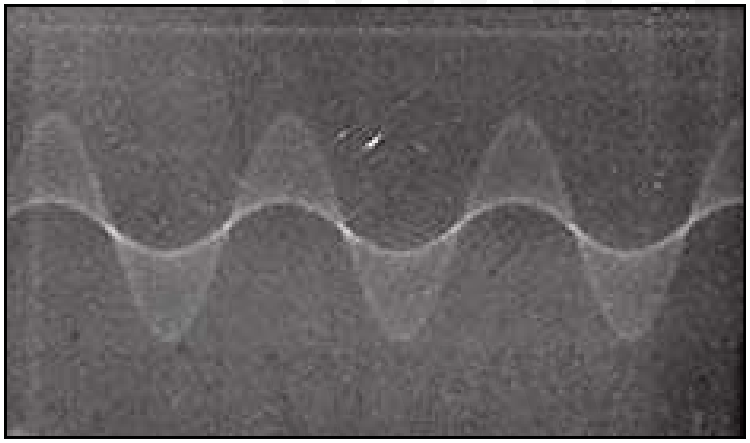

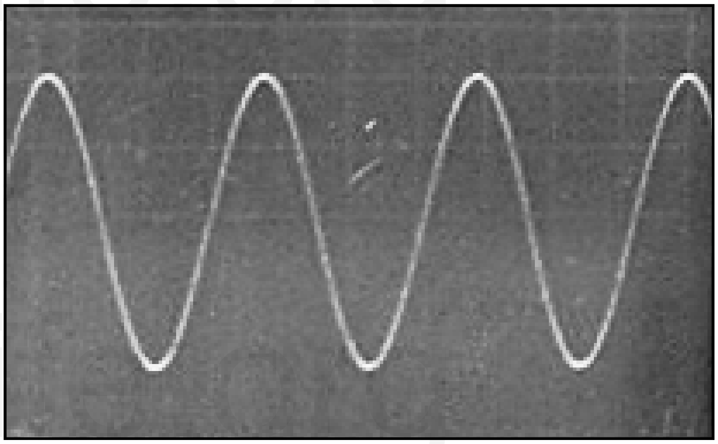

www.rife.org

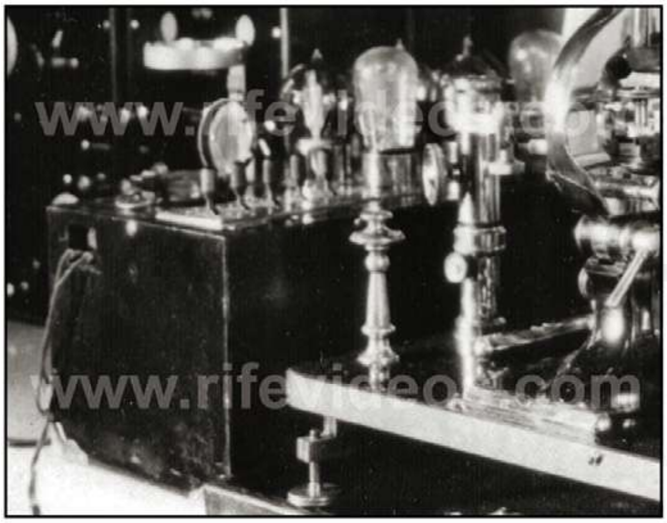
www.rifevideos.com

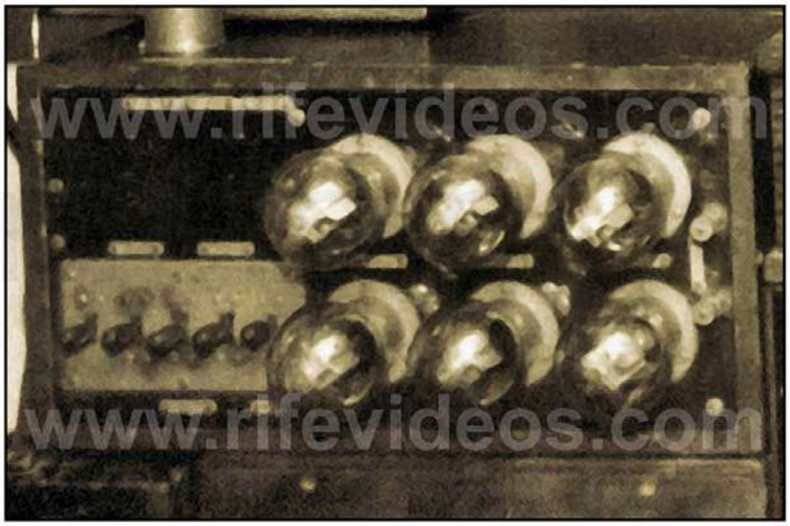
www.rifevideos.com

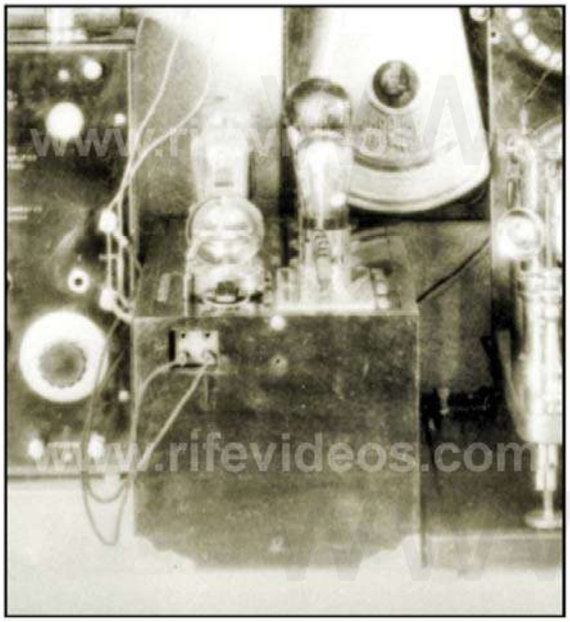
www.rifevideos.com

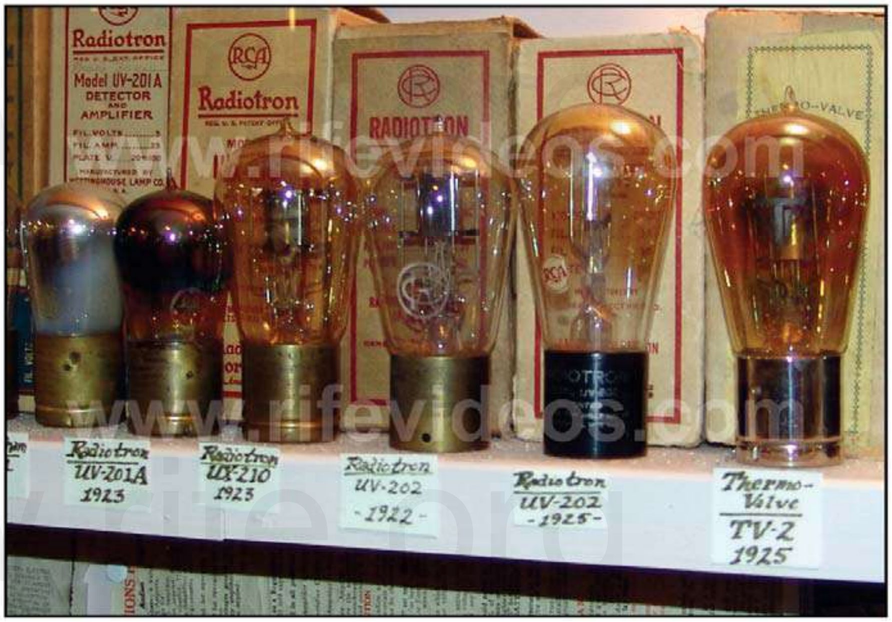
Radiotron
Model UV-201A
DETECTOR AND AMPLIFIER
Radiotron
RADIOTRON
Radiotron UV-201A 1923
Radiotron UX-210 1923
Radiotron UV-202 -1922-
Radiotron UV-202 -1925-
Thermo-Valve TV-2 1925
www.rifevideos.com

We will now discuss Dr. Rife’s multi-stage-amplifier that he used with the Kennedy equipment. This was most likely a class A RC coupling cascade style amplifier. Daven Company started building this type of amplifier back in about 1925. Dr. Rife may have had Daven custom build his multi-stage-amplifier but we cannot be sure. The two photos, at the bottom of this page, are Daven amplifiers. Both of them are three stage amplifiers. The Kennedy Receiver Model 110 only output about 1.5 to 3 volts. Dr. Rife needed to be able to amplify the signal to a high enough power level to make it effective. In the three old lab photos above we see Dr. Rife’s multi-stage-amplifier. In the above photo, bottom right, you can see the type of tubes he would have used in the early to mid 1920’s. These tubes would have made it so Dr. Rife could amplify the signal from the Kennedy Receiver Model 110 to about 50 watts in multi-stages. If you look at the above three photos of Dr. Rife’s multi-stage-amplifier you will see five switches. These five switches (representing five-stages) made it so he could choose different power levels determined by how many stages of amplification he wanted to use.

www.rifevideos.com

www.rifevideos.com

www.rife.org

With this five stage configuration he could have easily produced the 50 watts he said he used. This 50 watts, was the power level that was mentioned in the Rife CD's for this instrument.

Ben Cullen, a close friend of Dr. Rife's, mentions on the Rife CD's that Dr. Rife would light the ray tube with a separate power source. His lab photos shows a spark gap transmitter which he probably used, in the 1920's, to light the ray tube. If you look at Dr. Rife's lab photo, shown below, you can see the spark gaps. The spark gaps are right below the "Spark gap" writing.

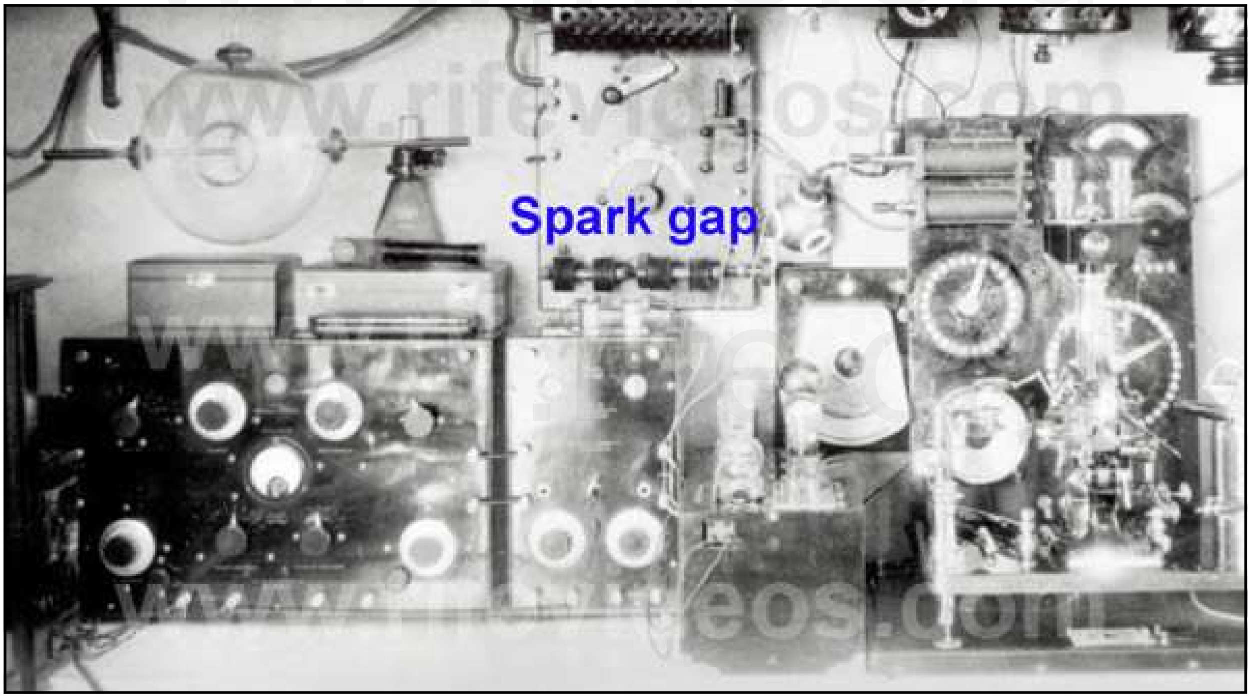

The next photo, shown below, is an up close photo of the spark gap transmitter diathermy from the 1920's. We purchased it so we could test the lighting of a ray tube with it.

www.rife.org
www.rifevideos.com

The next photo, shown below, shows the lighting of the ray tube using this spark gap transmitter. It lit the ray tube with ease and could output more power than the ray tube could handle.

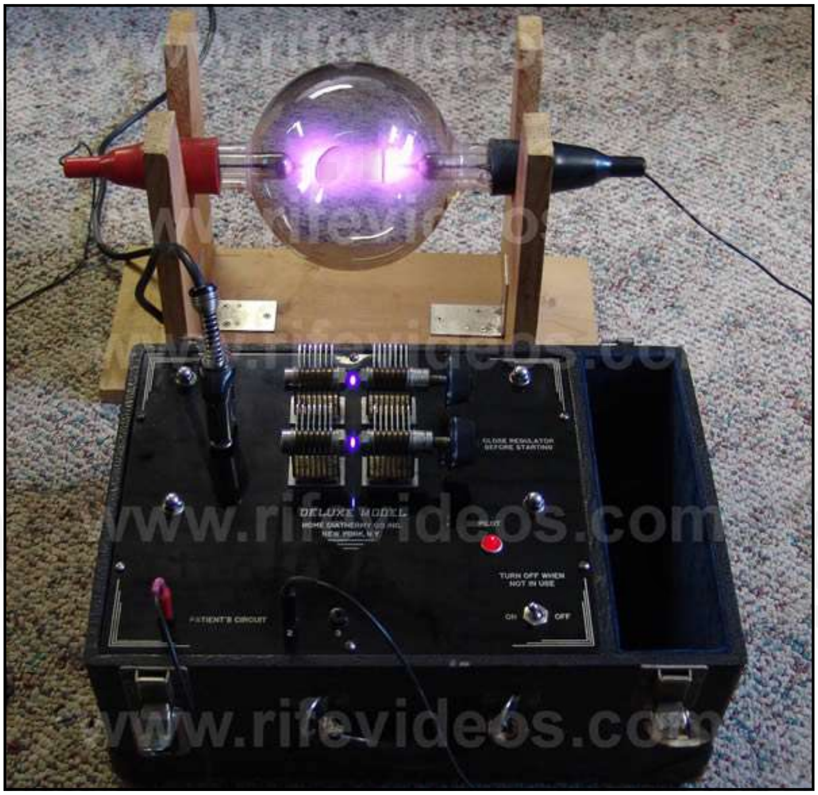

This spark gap transmitter would make it so Dr. Rife didn't have any difficulties tuning the ray tube when he changed frequencies from a low frequency of 139,000 Hertz to a higher frequency of 1,604,000 Hertz. The spark gap transmitter had a damped waveform and would have given him a damped wave carrier frequency (see the photo, below left) most likely somewhere around one Megahertz. This transmitter we purchased has a frequency of 920 KHz. Dr. Rife would not have modulated frequencies onto this carrier frequency but he would have just mixed the frequencies in the ray tube (see the photo, below right, is of sine & spark mix).

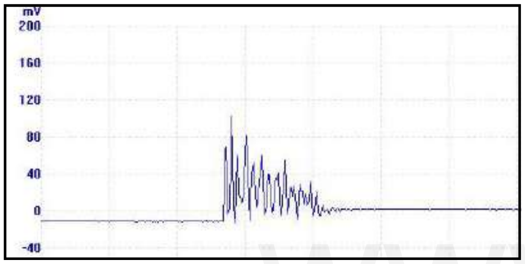

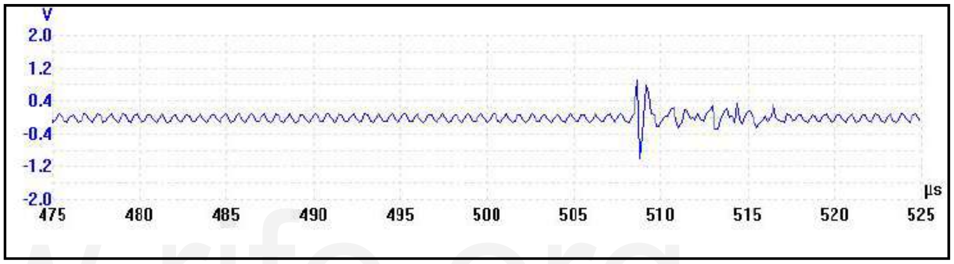

Mixing would have given him the combination of a damped wave and one or two sine wave frequencies, depending if he used two sine wave frequencies simultaneously. We do not believe that Dr. Rife continued to use a spark gap transmitter because it would have made it impossible for him to read the ray tube harmonic frequencies that his ray tube output. This is because a spark gap outputs broadband noise that makes it impossible to read any harmonic frequencies. Dr. Rife must have only used the spark gap transmitter in his early 1920's work. From the document "Development of the Rife Ray" we learn what he replaced the spark gap transmitter with.

"And as the voltage at this point was quite small, it was found necessary to apply external voltage across the anode and the cathode of the output tube [ray tube] *to act as a carrier wave for the frequencies that were generated in the apparatus."* (Development of the Rife Ray and use in devitalizing of pathogenic micro-organisms).

www.rife.org

It appears that the spark gap transmitter accomplished to important things. One: It lit the ray tub with an external voltage. Two: It produced a high potential voltage spike in the frequencies. Later he used a DC voltage transformer much like a neon light transformer to light his ray tube. Then he added an audio pulsing circuit to create the high potential voltage spike.

To better understand the reason why a high potential voltage spike is important we need to jump forward in the history of Dr. Rife's instruments to 1936-1937. The Beam Ray Laboratory instrument built by Philip Hoyland was built at this time and it had a fixed audio frequency pulsing circuit. The audio frequency that it produced was modulated with the RF frequencies it output. This audio pulsing circuit would have given Dr. Rife's frequencies a very high potential voltage spike almost identical to the damped wave of the spark gap. John Crane made this statement when he was narrating Dr. Rife's lab film.

CRANE: *"Now the spikes that you see on the frequencies are the lethal part that kill and devitalize the virus. They are the resonant peaks of the frequencies which increase the voltage to a very high potential which the cells of the virus wall can not tolerate and they break up into many pieces and are destroyed."* (Dr. Rife's Lab Film Narrated by John Crane in the 1970's).

The modulated audio frequency in the 1936-1937 Beam Ray Laboratory instrument was in the shape of a damped wave. With both the Rife Ray #4 and the Beam Ray Laboratory instruments having waveforms in the shape of a damped wave doesn't seem like a coincidence. When Dr. Rife discontinued using the spark gap and replaced it with an external high voltage current to act as a carrier frequency he would have had to develop a new method of creating this high potential voltage rise in his frequencies. It appears, with the help of Lee Deforest, that Dr. Rife must have developed this audio frequency pulsing circuit for his instruments. It is apparent that this pulsing of the M.O.R frequencies is the reason why Dr. Rife was able to devitalize the many microorganisms he tested.

Dr. Rife described the method he used to find these frequencies on the Rife audio CD's.

RIFE: "*Because when I check on that thing and look through that microscope hour after hour day after day, tuning that damn thing* [Kennedy 110] *to find something that will kill that bug. And every hour or half an hour, whatever is required, I put a new fresh culture under the microscope and keep that on and I find something that folds it up, alright!*" (John Marsh Rife CD's - CD 7 track 2).

It was a very tedious task to find a frequency that would devitalize and organism. Dr. Rife recorded all his frequencies on lab notes. Even though he misread his frequencies many people want to know those frequencies anyway. On the next page in a chart are the misread lab note frequencies which he recorded prior to 1935. Each lab note had two frequencies. One was listed in cycles per second and the second was listed in meters. For the purpose of making this report easier to understand the meter wave lengths on Dr. Rife's lab notes have been converted to cycles per second or Hertz. You will notice that there are two audio frequencies listed for organisms that are above 12,000 Hertz. They are the only audio frequencies ever listed by Dr. Rife for any organisms. One of them was changed to a higher RF frequency when Philip Hoyland read the correct frequencies in 1935 when he built the Rife Ray #4. Most likely the other audio frequency was really a higher RF frequency.

Chapter Summary: The Rife Ray #3 frequency generating equipment which Dr. Rife purchased back in 1923 was made by the Collin B. Kennedy Company. It mainly consisted of the Kennedy model 110 and model 281 to produce its frequencies. This equipment was regenerative not super-regenerative. Its frequency range, when the model 110 and model 281 were connected together, was from about 12,000 Hertz to about 2,000,000 Hertz (2 Million Hertz). Its power output through the ray tube was about 50 watts. The frequencies it output were mostly in the AM radio band of frequencies. This equipments frequency range now explains why all of Dr. Rife's frequencies were less than 2 million Hertz as

www.rife.org

listed on the Rife Ray #4 documents. The Rife Ray #3 was the instrument that was used by Dr. Rife and Dr. Milbank Johnson M.D back in the 1934 clinic on cancer and tuberculosis patients.

In the next chapter we will look at Dr. Rife's Rife Ray #4 Rife Machine and the correct frequencies which Philip Hoyland read with his master oscillator in order to build this instrument.

www.rife.org

Dr. Rife's Misread Lab Note Frequencies From Before 1935

Microorganism	First Frequency In Hertz	Second Frequency Meters To Hertz
Actinomycosis (Streptothrix)	678,000 Hz	186,554 Hz
Anthrax	900,000 Hz	272,539 Hz
Anthrax Symptomatic	400,000 Hz	16,655 Hz
B. Coli (Rod form)	683,000 Hz	317, 914 Hz
B. Coli (Filterable virus)	8,581,000 Hz	11,103,424 Hz
Bacillus X Cancer	11,780,000 Hz	17,033,662 Hz
Bubonic Plague	160,000 Hz	512,466 Hz
Catarrh	1,800,000 Hz	1,713,100 Hz
Cholera Spirillum	851,000 Hz	960,873 Hz
Contagious Conjunctivitis	1,206,000 Hz	2,025,625 Hz
Diphtheria	800,000 Hz	1,090,154 Hz
Glanders	986,000 Hz	736,591 Hz
Gonorrhea	600,000 Hz	150,649 Hz
Influenza	1,674,000 Hz	1,946,704 Hz
Leprosy	743,000 Hz	251,926 Hz
Pneumonia	1,200,000 Hz	381,901 Hz
Spinal Meningitis	927,800 Hz	1,795,164 Hz
Staphylococcus Pyogenes Aureus	998,740 Hz	555,171 Hz
Staphylococcus Pyogenes Albus	Found in Dr. Rife's papers	549,070 Hz
Streptococcus Pyogenes	1,214,000 Hz	2,111,214 Hz
Syphilis (Treponema Pallidum)	900,000 Hz	2,775,856 Hz
Tetanus	700,000 Hz	15,779 Hz
Tuberculosis (Rod form)	583,000 Hz	541,142 Hz
Typhoid Fever (Rod form)	900,000 Hz	868,964 Hz
Typhoid Fever (Filter passing)	9,680,000 Hz	13,943,835 Hz

www.rife.org

Chapter #8

1935 Rife Ray #4 Rife Machine

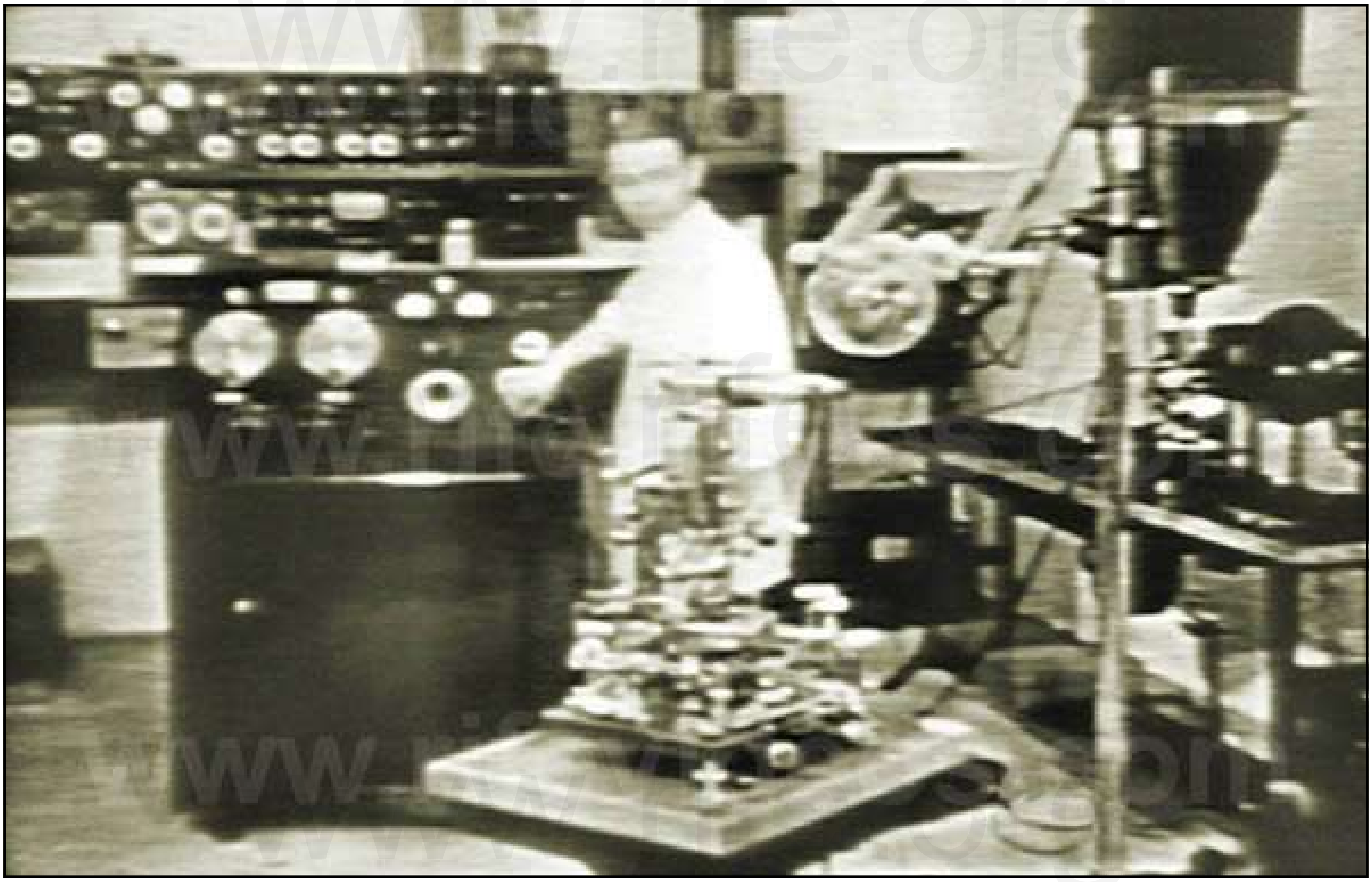

1. Used a ray tube.
2. Had two separate oscillators so it could output two frequencies at a time. Frequency range was from 87,000 Hertz to 22.5 MHz.
3. Power usage was about 450 to 600 watts. Output to the ray tube was variable (50 to about 250 RF watts).

Some have asked how we can be sure these photos we have are of the Rife Ray #4 Rife Machine. It is a simple process of deductive reasoning. John Crane, one of Dr. Rife's 1950's business partners, incorrectly dated the Rife Ray #4 as a 1942 Rife Machine. This misdating has led to some confusion which we will now attempt to clear up. In the photo, shown above, we see Dr. Rife using the instrument which John Crane dated as built in 1942. However, the lab film this picture was taken from was made in the summer of 1936 for use at a conference which Dr. Rife planned to attend in the autumn of that same year. He was presenting this film at this conference to demonstrate the isolation of the BX cancer virus. This properly dates the instrument as having been built before the summer of 1936 and shows that John Crane was incorrect. In the background of this photo, behind the instrument Dr. Rife is using, we see his Kennedy Company equipment back against the wall. Therefore this instrument could have only been built in late 1935 or early 1936. The Rife Ray #4 documents show it was completed in the fall of 1935. This logically dates the Rife Machine Dr. Rife is using in the 1936 film as the Rife Ray #4. The next picture, shown on the next page, is a better photo of the Rife Ray #4 instrument.

www.rife.org

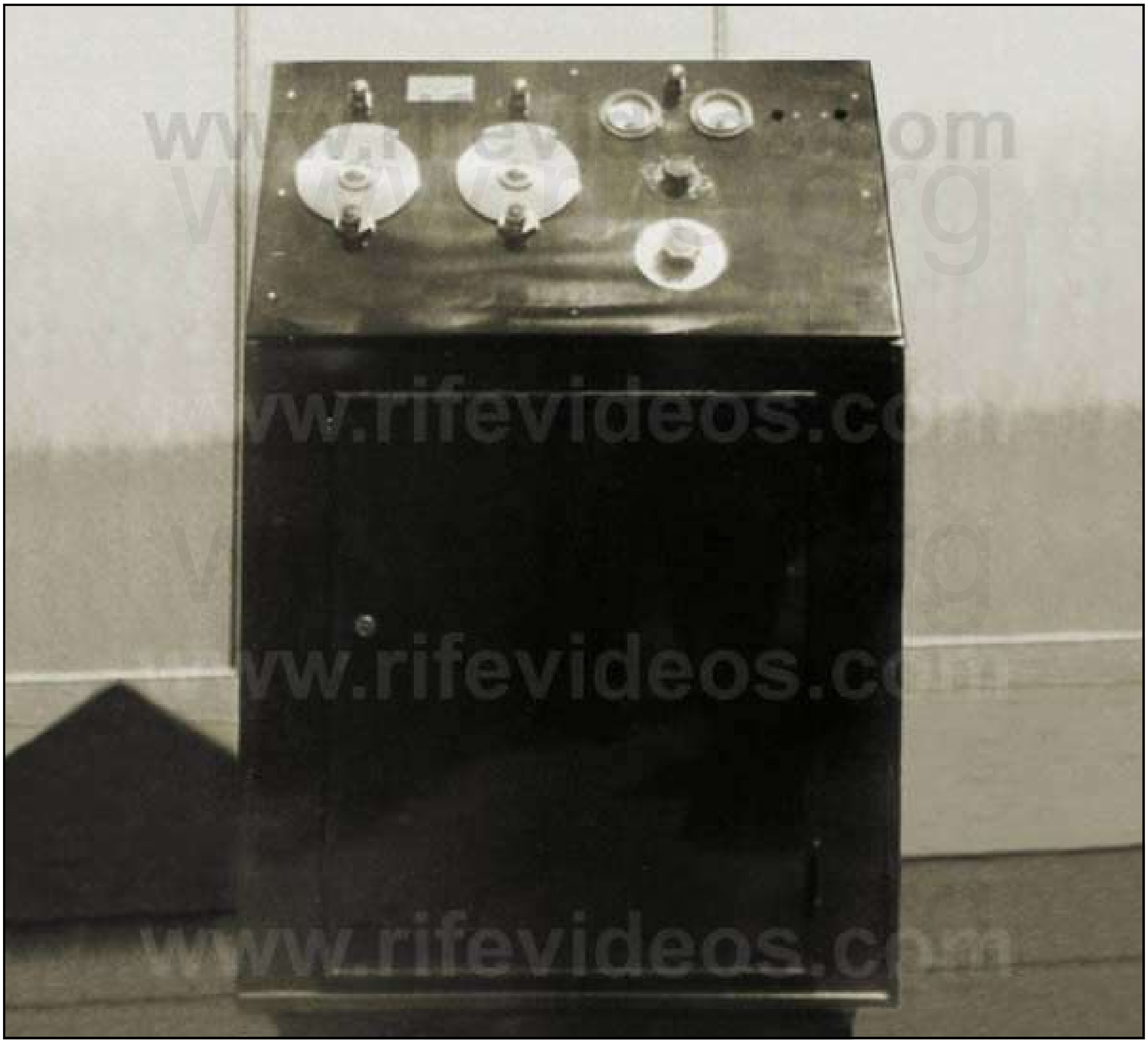

Bertrand Comparet, Dr. Rife’s attorney, said three experimental machines were built. With the proper dating of this Rife Machine, which shows it is the Rife Ray #4, we will now discuss it in detail.

Of the three experimental machines it appears that two were Rife Ray #4's and the third was the Rife Ray #5 or Beam Ray Clinical experimental instrument which was used in a clinic in 1937. We will cover the Beam Ray Clinical instrument later.

Dr. Johnson's machine was the first Rife Ray #4 that was built in 1935. After Dr. Johnson's instrument was built he states in a letter that the second #4 instrument was soon to be built for Dr. Rife:

DR. JOHNSON: *"While you have the machine* [Dr. Johnson's Rife Ray #4] *down there, I hope you and Jack are working on those bananas so we can get an M.O.R. for them, if possible. As soon as you send Phil and the machine back, I am going to start Phil on making your Machine."* (Letter from Dr. Johnson to Dr. Rife, November 15, 1936).

Now that we have established the fact that both Dr. Milbank Johnson and Dr. Rife had their own Rife Ray #4 instrument we can continue with the history of this instrument. As pointed out earlier in this report Philip Hoyland eventually became Dr. Rife's engineer and business partner. Philip Hoyland was the one who measured Dr. Rife's frequencies when he came to Dr. Rife's lab in 1935.

www.rife.org

We will also point out that the frequencies used in this instrument were transferred from the Rife Ray #3 to the Rife Ray #4 after Philip Hoyland read the correct frequencies. We have in the document "Development of the Rife Ray" a description of this instrument:

“In 1935 in entirely new application of the old principle was incorporated in an instrument built under the direction of Commander Rife by Philip Hoyland of Pasadena California. The new instrument was light socket powered and had an output of 500 Watts. Designated #4 [Rife Ray #4] [Many of the power output measurements of Dr. Rife’s equipment were misread]. *Furthermore it was equipped to deliver two distinct frequencies simultaneously and both variable. This apparatus proved to be more efficient with decidedly fewer factors of error in the laboratory tests using 75 pounds of horse meat.”* (Development of the Rife Ray and use in devitalizing pathogenic micro-organisms).

In the laboratory tests that were performed with the Rife Ray #4 they made many tests to determine its capability. Dr. Johnson talked about the many different aspects of the machine they had to test in a letter:

DR. JOHNSON: *"Also, about this time, the new Rife Ray Machine* [#4] *had arrived at its point of construction when elaborate tests had to be made in order to synchronize the M.O.R. produced by it with the M.O.R. produced by the old machine* [#3]. *Now, we are in the throes of accurately charting the 14,000 possible settings on the new machine. Our next process, beginning next week, is to test its penetration, the time required in the different exposures, the different depths of lesions. So, take it altogether we are just about as busy as a bear in berrytime."* (Letter from Dr. Johnson to Dr. Schram, dated September 25, 1935).

Dr. Rife’s Single M.O.R. Frequency

The Rife Ray #4 instrument documents show it could put out two RF or radio frequencies simultaneously. Dr. Rife’s previous Kennedy Model 110 when connected to the Model 281 could output two frequencies simultaneously like the Rife Ray #4. It is apparent that Dr. Rife still wanted this capability in this new instrument. Dr. Rife from time to time would run two frequencies or M.O.R.s. simultaneously on two different organisms. This is pointed out by Dr. Rife on the John Marsh Rife audio CD’s:

RIFE: *"We found the frequency of the virus, we found the frequency of the rod, which we had for years of course. But if we use the two of them simultaneously over the same carrier wave, the patient gets well and the Guinea pig gets well, but if you use one or either individually you either kill the patient or you don't do nothing.*"

This statement from Dr. Rife also shows that each organism only requires a single frequency to devitalize it. The rod form of tuberculosis requires a different frequency than the virus form. In the above quote from Dr. Rife he clearly states "frequency" for each organism, not frequencies. Since his instrument could only output two frequencies at the same time this clearly shows that each organism has a single frequency as its M.O.R. In another statement when Dr. Rife was talking to John Crane about tuberculosis he said:

CRANE: *"We have the frequency that will devitalize E. coli, don't we?*"

RIFE: *"We have yes. And we have the frequency also that will devitalize the filterable form of B. coli which is not necessary. There is only one organism that we have to use two frequencies simultaneously, and that as we've spoken of before John, and that is the Bacillus of tuberculosis, where it has that so-called poison molecule of Von that is released by any known methods of devitalizing the organism and that's it."* (Marsh collection, Rife audio CD’s).

www.rife.org

Dr. Rife states in this quote that there is only one organism that needs two different frequencies to devitalize it. All the other organisms only need a single frequency to devitalize them. The reason that Tuberculosis needs two frequencies is due to the fact that Dr. Rife said both the rod form and the virus form must be devitalized at the same time or the virus form will feed on the dead body of the rod form causing the patient to die from toxemia.

Even with what Dr. Rife has said many have still believed that it took two frequencies to devitalize a single organism. This misunderstanding came about because Dr. Rife had two frequencies listed for each organism on his pre-1935 lab notes. From the reading of these documents we can see this is not the case. It was Dr. Rife's misreading of his frequencies, prior to 1935, which lead to this misunderstanding. To further make the point that it only takes a single frequency to devitalize an organism we will read more statements made by Dr. Rife on the Rife Audio CD's:

RIFE: *"The frequency of each bacteria is absolutely individual. They run through a very, very large gamut. Some of them are very, very broad, long. Some of them are not extremely short. There’s none of them what we call our ultra short wave that I have found yet."* (Marsh collection, Rife audio CD’s).

DR. DICKLAND: *"Is there a different frequency between cancer and tuberculosis?"*

RIFE: *"Oh much. Every organism requires a different frequency owing to its own chemical constituents or its premodel cell or predominate chemical factor. "* (Marsh collection, Rife audio CD’s).

RIFE: *"We found the frequency of the virus, we found the frequency of the rod, which we had for years of course."* (Marsh collection, Rife audio CDs).

RIFE: "*So we’re throwing an electronic frequency through the tissues of the body that simply devitalizes the bacteria with no harm to normal tissue."* (Marsh collection, Rife audio CD’s).

These statements clearly show again that each organism has its own frequency and that it only takes a single frequency to kill, devitalize or render it harmless. As we read other statements or documents you will easily be able to see that this understanding was plainly understood by Dr. Rife and all the other doctors that used his equipment.

Because this is a very important point that needs to be understood we will read a few more documents which contain added information on this subject. When the Rife Ray #4 was finally completed Dr. Johnson wrote a letter to Dr. Rife in which he made this observation:

DR. JOHNSON: *"Maybe these Much glands are another form of the T.B.* [tuberculosis] *corresponding to our filter passing form and we will have to get an M.O.R. for them so as to destroy them at the same time that we do the rod form of tuberculosis....Now that we have a machine in which we can give two M.O.Rs at one time, it would be easy to treat all forms of tuberculosis, both for the tubercle bacilli and the Much granules."* (Letter from Dr. Johnson to Dr. Rife dated, September, 12 1935. Page 3).

Since the Rife Ray #4 only had two variable oscillators and each one was set on a different M.O.R. frequency then it would only take one frequency to devitalize each organism. In another letter this same understanding is given:

DR. JOHNSON: *"I would suggest that you set up a slide with a hanging drop planted with the cryptomyces pleomorphia. Set your machine and the microscope together. After it gets growing, start out and set on of the groups to the MOR of BX. Then with the other group, start with a certain switch on Group 2 beginning at zero. Give one minute exposure, wait about fifty minutes, and then see if there is any change. Then move the dial of Group 2 up to ten and give another exposure, waiting about fifty min-*

www.rife.org

utes and the examining it carefully. Go on that way throughout the dial and you will get through about one switch a day. Pretty soon you will find some setting that will kill the organism. Mind you, you are running at every exposure Group 1 on the BX MOR. If you got this setting started, Jack could do the microscope work and if he got any change, he could call your attention to it. That would save your eyes and your time. It is very important that we find the MOR of this fungus before I start in with the clinical work in the hospital with this new machine." (Letter from Dr. Johnson to Dr. Rife dated, October 10, 1935).

www.rife.org

When this letter was written they already had the frequency for the BX cancer virus and they were trying to find the frequency for the cryptomyces pleomorphia fungi. It is clearly pointed out that one oscillator was set on the BX organism and the second oscillator was being used to try and find the frequency for the cryptomyces pleomorphia fungi. Dr. Rife when speaking about the BX virus clearly states that it had a certain frequency not frequencies:

RIFE: *"If you don't have an absolute coordinative resonance, you have nothing. One tenth of one meter off and you have nothing. It's got to be absolutely correct for that individual organism. It's got to be precise...the virus of cancer has a certain frequency. And it has to be there, otherwise if it's a little one way or the other, no good, no good for nothing."* (John Marsh Rife CD's - CD 5 track 2, CD 6 track 2, CD 7 track 1 and CD 9 track 1).

There are even more documented examples which show that only one frequency is needed for each organism. From the horse meat tests done with the Rife Ray #4 we read:

"The test being ready, the No. 4 Rife Ray Machine was set for the MOR of B. Typhosis (rod) and also for the filter-passing form. It was turned on for an exposure of three minutes at 400 watts..." (Test of the No. 4 Rife Ray Machine, Page 2).

The frequency listed on the Rife Ray #4 documents for B. Typhosis rod form is 760,000 Hertz and the frequency for the B. Typhosis filter-passing virus form is 1,445,000 Hertz. This document also confirms that only one frequency is needed for each organism to devitalize it.

There is one more document that we will look at. This document has the misread frequency of 11,780,000 Hertz for the BX frequency. The correct frequency that was read by Philip Hoyland was 1,604,000 Hertz. It is apparent that Dr. Rife was reading the 7th harmonic of 1,604,000 Hertz incorrectly. The correct 7th harmonic frequency of 1,604,000 Hertz is 11,228,000 Hertz. Had Dr. Rife read the 7th harmonic correctly this frequency would have been on his lab note. Regardless of this mistake Dr. Rife listed only one frequency for devitalizing the BX virus. Below is his statement:

RIFE: *"Since experiments show that the Bacillus X* [BX cancer virus] *in form A exists in malignant tissue, it is theoretically possible to change the cycle to form B by application of the Argon Ray and vacuum conditions. After the cycle change has been accomplished (in theory), the application of the oscillative ray at a cycles per second vibration of 11,780,000* [correct frequency measured in 1935 was 1,604,000] *should completely destroy the B.X. in the malignant tissue."* (Rife Research Laboratory lab note on the BX).

So that no one misunderstands this quote. It was the change of the Bacillus form from A to B using the Argon Ray and vacuum conditions that was the "theory" not the destroying of the B.X. using the 11,780,000 (corrected frequency of 1,604,000) Hertz. All these document show that Dr. Rife found that only one frequency is required for each organism.

There is one other thing that apparently needs to be combined with the M.O.R. frequency and that is the pulsing of the waveform. When we discussed the Rife Ray #3 we talked about a fixed audio frequency pulsing circuit. The Rife Ray #4 also had a fixed audio frequency pulse circuit. The waveform that it output was shown in Dr. Rife's lab film and is the next photo, shown below.

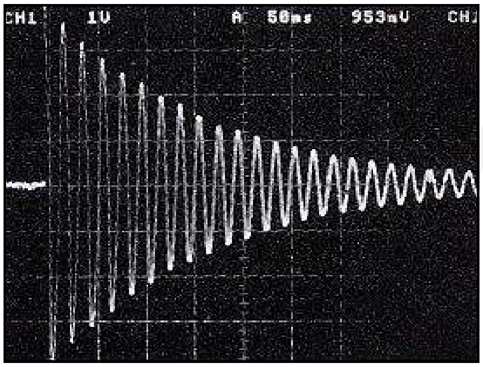

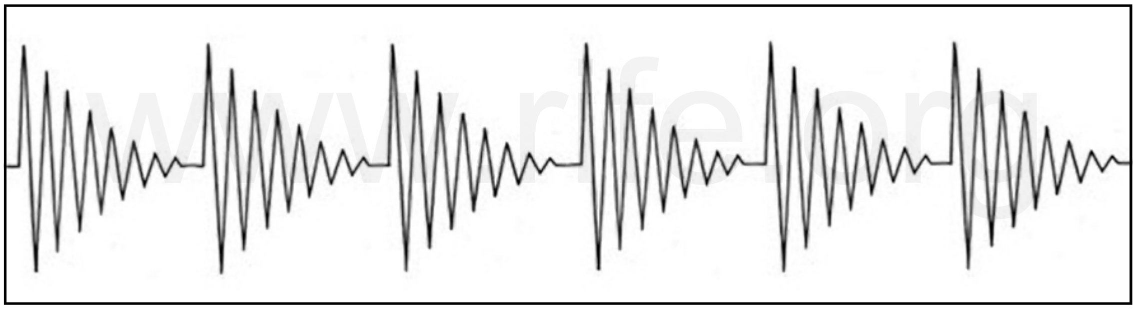

The other photo shown above, top left, is a clearer photo of what a damped wave looks like. Below those two photos is drawing of what a damped waveform looks like. How Dr. Rife created this waveform was not known until a circuit from an original Beam Ray Laboratory Rife Machine was rebuilt. This circuit will be discussed in detail later in this paper when we cover that instrument.

Dr. Rife's Audio Pulsing Circuit

Though we will discuss the building of this circuit later the importance of that fixed audio frequency pulsing circuit will be discussed now. The reason for this is without this pulsing circuit no organism could have been devitalized. This pulsing circuit was the secret to making Dr. Rife's high RF frequency instruments work. Bertrand Comparet, Dr. Rife's attorney said this in his 1970's interview:

COMPARET: *"The whole secret of the Rife thing was the audio frequency because Rife, way back in the old days, way, way back, Rife told me that the way he used to tune his instrument…he would hook up headphones and turn the thing. He had a very keen musical sense of pitch and so on, and he would tune it in his headphones until he got the right pitch, and that was the frequency."* (1970's Bertrand Comparet interview paper #89).

This audio frequency pulsing circuit was put into the Rife Ray #4 Rife Machine as a fixed audio frequency. It was not a variable audio frequency circuit. The Rife Ray #4 instrument documents show the #4 instrument had no variable audio oscillator which indicates that Dr. Rife believed that it was no longer necessary for M.O.R. work. But the fix audio pulsing circuit was absolutely necessary to be put into the instrument so that the high RF frequency M.O.R. frequencies were pulsed in order to devitalize the various organisms. All of the frequencies that Philip Hoyland read from the Kennedy Model 110 and transferred to the Rife Ray #4 were RF frequencies and needed to be pulsed. The lowest frequency was for Anthrax at 139,200 Hertz; the highest was 1,604,000 Hertz for the BX organism that caused cancer.

www.rife.org

The Rife Ray #4 would have also been a sine wave instrument just as the Kennedy Company equipment was. Since the Rife Ray #4 had two high RF frequency oscillators it would allow for the oscillators to be set at two different M.O.R.s. This instrument was probably built much like the Beam Ray Laboratory instrument which connected the ray tube between the two RF oscillator tank coils. The reason for this is the fact that the bandwidth of the RF tank coils could only pass a frequency of about 250,000 Hertz and Dr. Rife was using frequencies much higher than this.

www.rife.org

Dr. Rife’s Carrier Wave

We will now discuss Dr. Rife's use of a carrier wave in his instruments. This next statement by Dr. Rife’s verifies the fact that with the Kennedy equipment or Rife ray #3 he used a carrier wave:

RIFE: *"We found the frequency of the virus, we found the frequency of the rod, which we had for years of course. But if we use the two of them simultaneously over the same carrier wave, the patient gets well and the Guinea pig gets well, but if you use one or either individually you either kill the patient or you don't do nothing".* (Marsh collection, Rife audio CD’s).

Dr. Rife's mention of a carrier wave has lead to some confusion in the past because when we think of a carrier wave we think of a higher RF frequency that the lower RF or audio frequencies would ride on. But the documents show this was not the case. The "Development of the Rife Ray" document tells us what that carrier wave was:

"And as the voltage at this point was quite small, it was found necessary to apply external voltage across the annode and cathod of the output tube to act as a carrier wave for the frequencies that were generated in the apparatus." (Development of the Rife Ray and use in devitalizing of pathogenic micro-organisms).

Dr. Rife's carrier wave was just an external high voltage current that was used as a means to light the ray tube. The use of this high voltage current as a carrier wave was also used with the Rife Ray #4 and was not remove from the instruments until the summer of 1936 when the new Beam Ray Clinical instrument was built. This will be covered later.

Dr. Rife used the Rife Ray #4 Rife Machine in his laboratory until he closed it down in about 1947. The Rife ray #5 or Beam Ray Clinical instrument was completed in the summer of 1936 and was the new design that was used by all the doctors. Dr. Johnson used his Rife Ray #4 instrument in at least two medical trials. Dr. Johnson always had his patients sign a release card, shown on the bottom of page 56, in order to be treated with the frequencies output by the Rife instrument.

On page 57 is a chart showing the sine wave frequencies read off of the Kennedy Company equipment and used in the Rife Ray #4 instrument. We now know these frequencies were the frequencies used in the 1934 clinic since they were taken off of the Rife Ray #3 which was used in that clinic. The chart was made from The Rife Ray #4 document. This same document along with its additional page was mentioned in a letter written by Dr. Milbank Johnson to Dr. Rife dated Oct. 15, 1935:

DR. JOHNSON: *I am also inclosing a copy of a chart prepared for me by Phil* [Philip Hoyland] *which show the frequencies in kilocycles of the different M.O.R.’s we have worked with. Opposite these frequencies you will find the proper switch and settings on both groups. Now, in looking over the frequencies representing the M.O.R. of certain organisms, I find that the typhoid rod has 760 kc. And just above that you will find that the typhoid filter passing has 1445 kc. - that is almost an exact harmonic of the rod. Assuming that the banana might represent the rod form of the cryptomyces pleomorphia - on the same plan let us assume that the B.X. is the filter passing form of this organism. If that is so, we should kill the rod, it being the harmonic of the rod form, at 800 kc.* [this statement verifies that the BX

M.O.R. frequency was 1.604 kc. - or 1,604,000 Hertz]. *This 800 kc. Will come in group No. 2 on switch 6. A reference to the proper chart under Group No. 2 would give the exact setting to get 800 kc. This is just a chance and I have no assurance that it will work out, but I think it is worth trying so I think I would try on switch 6 in Group N. 2 for your first effort to kill the fungus, or rather to kill the bananas.* (Letter from Dr. Milbank Johnson to Dr. Rife, October 15, 1935).

This letter and the Rife Ray #4 document confirm that Dr. Rife had misread his frequencies prior to 1935 and that each organism only has a single frequency as its M.O.R. The letter also shows that they were using the corrected frequencies read by Philip Hoyland when he came to Dr. Rife's lab to read the Rife Ray #3 in 1935. The frequencies used in Rife Ray #4 were used to create the next version of Dr. Rife's machine.

Chapter Summary: The Rife Ray #4 was built in 1935 at the request of Dr. Milbank Johnson because he wanted an instrument that was in the comforts of a single cabinet. It had the highest frequency range and power output of any of Dr. Rife's machines. There were only two of these machines built. The first one was built for Dr. Milbank Johnson and the second one was built for Dr. Rife. The frequencies used in this machine were the same frequencies that were used with the Rife Ray #3. Each organism only required one frequency to devitalize it. Like the Rife Ray #3 the Rife Ray #4 used a fixed audio frequency pulsing circuit to create a high potential voltage rise in the RF M.O.R frequency to devitalize the organism treated with the instrument. The Rife Ray #4 was used by Dr. Milbank Johnson in two clinical trials.

In the next chapter we will take an in-depth look the next instrument that was built by Dr. Rife's engineer, Philip Hoyland. Philip Hoyland developed a new method of delivering Dr. Rife's M.O.R. frequencies.

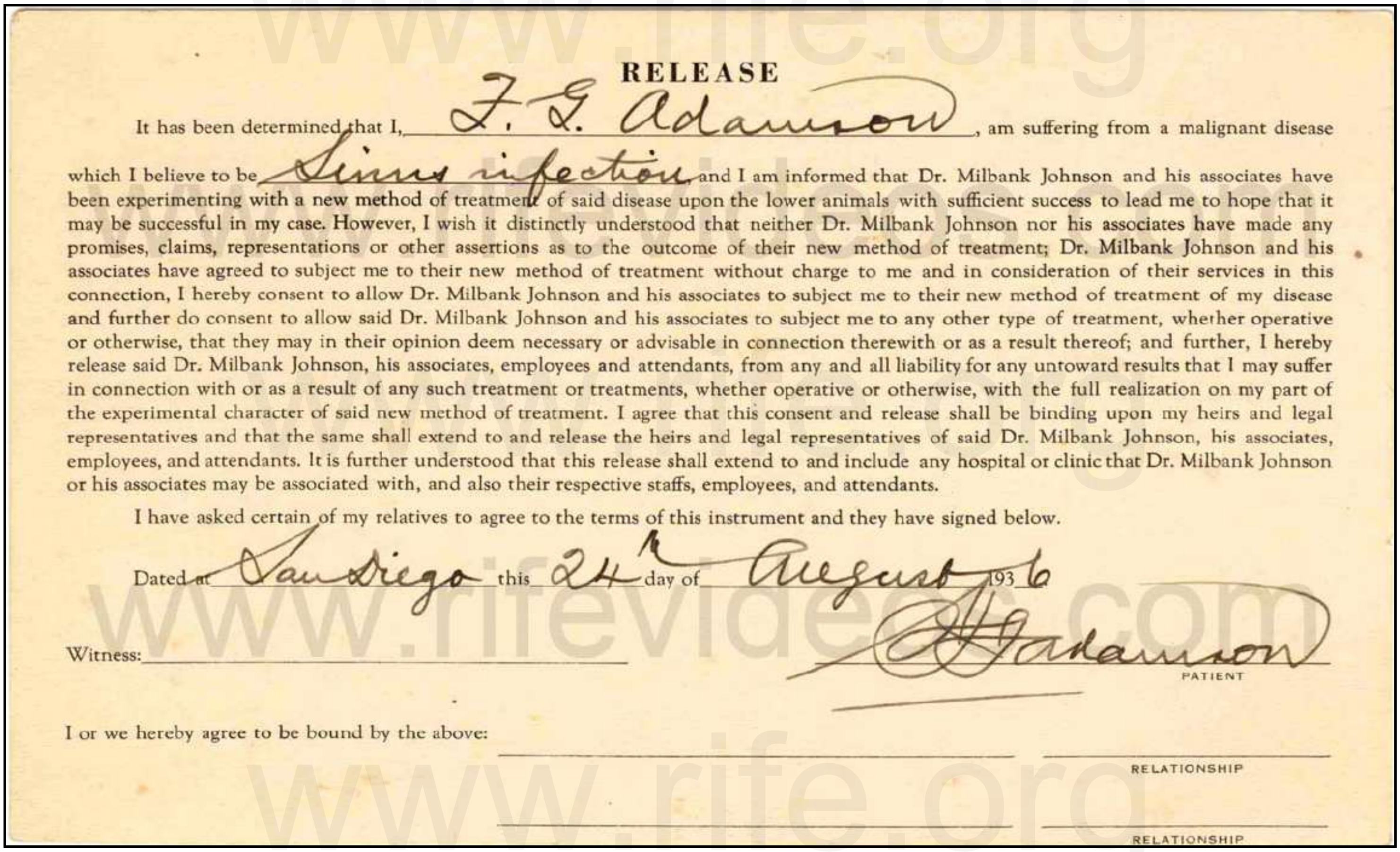

RELEASE

It has been determined that I, F. G. Adamson, am suffering from a malignant disease which I believe to be Sinus infection and I am informed that Dr. Milbank Johnson and his associates have been experimenting with a new method of treatment of said disease upon the lower animals with sufficient success to lead me to hope that it may be successful in my case. However, I wish it distinctly understood that neither Dr. Milbank Johnson nor his associates have made any promises, claims, representations or other assertions as to the outcome of their new method of treatment; Dr. Milbank Johnson and his associates have agreed to subject me to their new method of treatment without charge to me and in consideration of their services in this connection, I hereby consent to allow Dr. Milbank Johnson and his associates to subject me to their new method of treatment of my disease and further do consent to allow said Dr. Milbank Johnson and his associates to subject me to any other type of treatment, whether operative or otherwise, that they may in their opinion deem necessary or advisable in connection therewith or as a result thereof; and further, I hereby release said Dr. Milbank Johnson, his associates, employees and attendants, from any and all liability for any untoward results that I may suffer in connection with or as a result of any such treatment or treatments, whether operative or otherwise, with the full realization on my part of the experimental character of said new method of treatment. I agree that this consent and release shall be binding upon my heirs and legal representatives and that the same shall extend to and release the heirs and legal representatives of said Dr. Milbank Johnson, his associates, employees, and attendants. It is further understood that this release shall extend to and include any hospital or clinic that Dr. Milbank Johnson or his associates may be associated with, and also their respective staffs, employees, and attendants.

I have asked certain of my relatives to agree to the terms of this instrument and they have signed below.

Dated at San Diego this 24th day of August 1936

Witness: ______________________ F. G. Adamson
PATIENT

I or we hereby agree to be bound by the above: ______________________ ______________________
RELATIONSHIP

______________________ ______________________
RELATIONSHIP

www.rife.org
www.rifevideos.com

Rife Ray #4 Sine Wave High RF Frequencies

Actinomycosis (Streptothrix)	192,000 Hz
Anthrax	139,200 Hz
B. Coli (Rod form)	417,000 Hz
B. Coli (Filterable virus)	770,000 Hz
Bacillus X or BX Virus (Cancer Carcinoma)	1,604,000 Hz
Bacillus Y or BY Virus (Cancer Sarcoma) This frequency was not on the Rife Ray #4 document because it was not found at that time. It was in 1936 when this viruses frequency was obtained and it was used in the Beam Ray Clinical machine. We included it here with this explanation.	1,530,000 Hz
Gonorrhea	233,000 Hz
Spinal Meningitis	427,000 Hz
Staphylococcus Pyogenes Aureus	478,000 Hz
Streptococcus Pyogenes	720,000 Hz
Syphilis	789,000 Hz
Tetanus	234,000 Hz
Tuberculosis (Rod)	369,000 Hz
Tuberculosis (Virus)	?769,000 Hz
Typhoid Fever (Rod)	760,000 Hz
Typhoid Fever (Virus)	1,445,000 Hz

www.rife.org
www.rife.org
www.rife.org
www.rife.org
www.rife.org

Chapter #9

1938 to 1939 Beam Ray Corporation Clinical Rife Machine

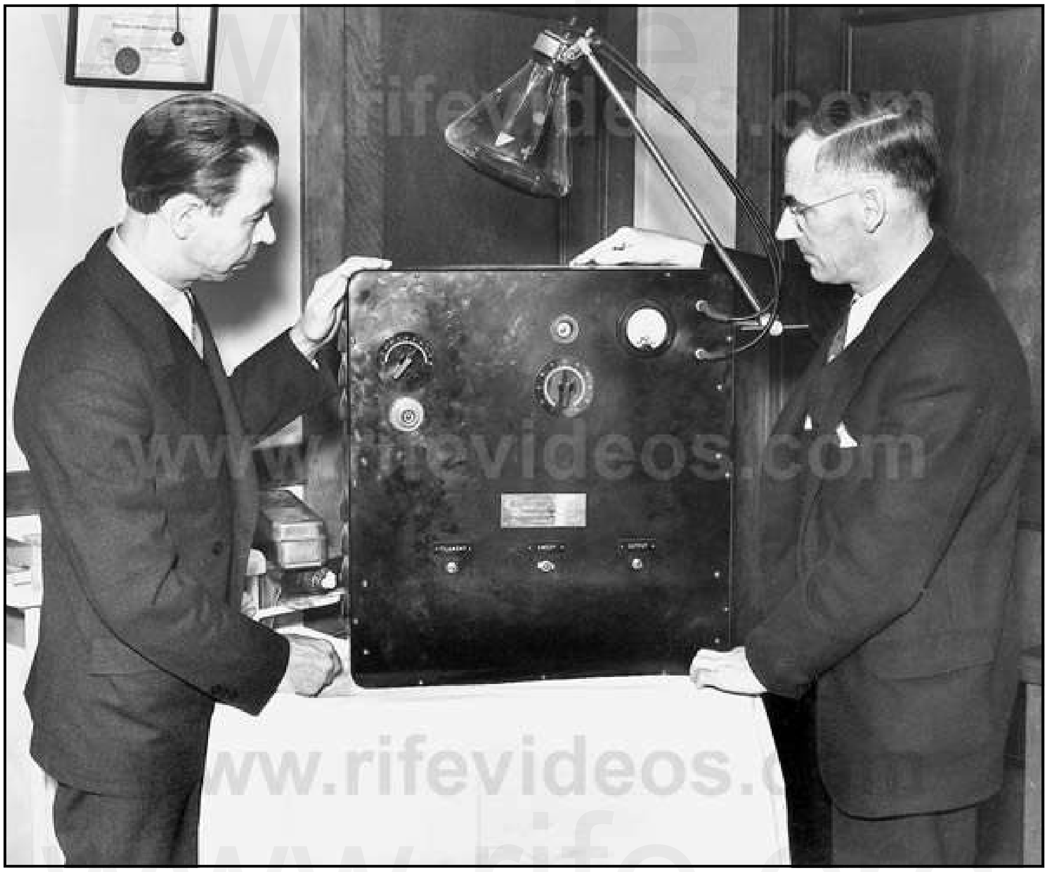

1. The instrument used a ray tube.
2. Had one variable Audio oscillator and one Fixed RF oscillator set at 3.30 or 3.80 MHz.
3. Power usage was about 450 to 600 watts. Output to the ray tube about 40 to 50 RF watts.

Just as with the Rife Ray #4 Rife Machine we must determine what the Rife Ray #5 or Beam Ray Corporation Rife Machine looked like. The reason we need to determine this is because unless we know what those instruments really looked like we may think we have a true Rife Ray #5 or Beam Ray Clinical instrument and find out later that it is not one. Beam Ray Corporation built two different instruments, one was called the Clinical instrument and the other was called the Laboratory instrument. The fact that Beam Ray built two different instruments was pointed out in the Trial: (Beam Ray Trial Transcript #209-210)

COMPARET: *"The four machines bought by the British were two so called laboratory types and two so called clinical types, what was the difference between the two?"*

HOYLAND: *"The clinical type was similar in all respects to the Rife machine except that it did not have* [word missing] *of the* [word missing] *used on Mrs. Henderson."*

We now know that Beam Ray Corporation built two different Rife Machines of which one was built using the original Rife principles and it was considered the Laboratory instrument. The other was built using a different method of generating the frequencies and we will show that it was called the Clinical instrument. First we will prove that the photo, shown above, and the next photo, shown on the top of the next page, are photos of the Rife Ray #5 or Beam Ray Clinical instrument.

www.rife.org

The instrument, shown below, is a photo of one of two Rife Machines owned by Dr. James B. Couche which he purchased from Beam Rays Corporation. Dr. Hamer also purchased one of these Clinical Rife Machines. This information was pointed out during the Trial and gives us the proof we are looking for: (Beam Ray Trial Transcript #98-99, 217-218, 1128-1131 and 2700)

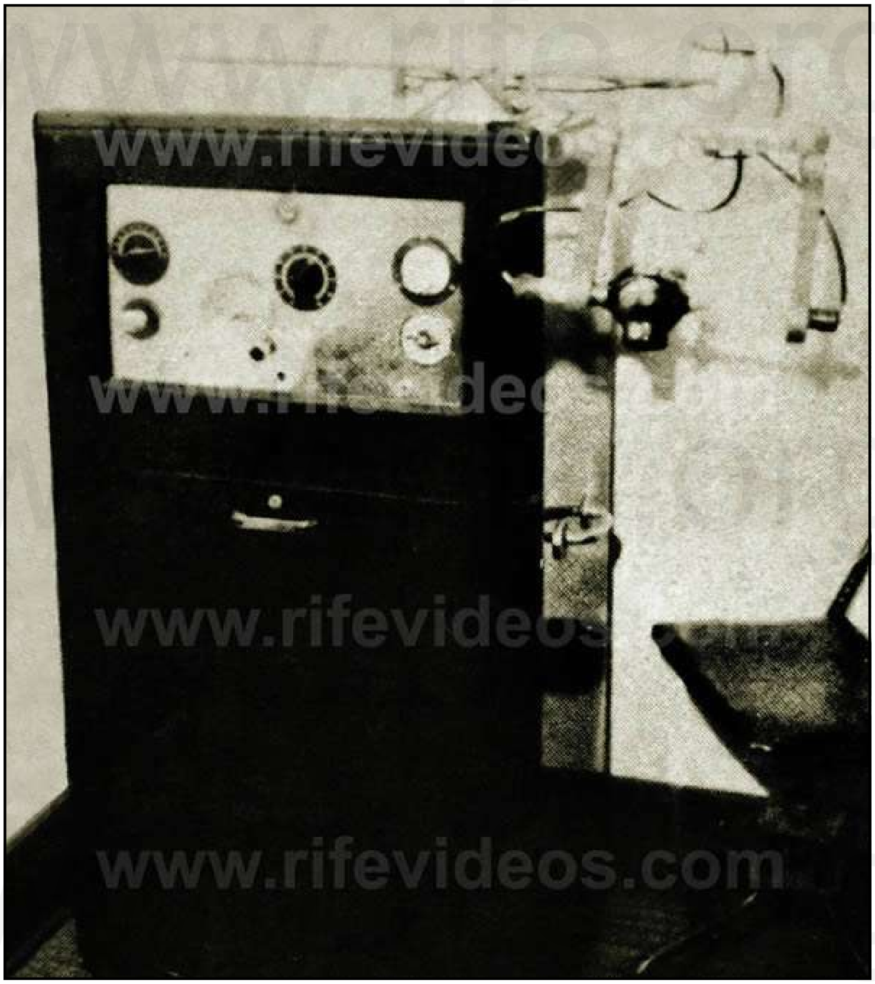
www.rifevideos.com

COMPARET: *"Before this agreement was signed did the company manufacture any Rife ray machines?"*

HOYLAND: "*They started to about the first of May* [May 1, 1938]. *Dr. Hamer was sold one."*

COMPARET: *"How was the price of these machines fixed?"*

HOYLAND: *"The price was decided from the costs of what it cost to manufacture the first machine that was sold to Dr. Hamer."*

COMPARET: *"Were the clinical machines the same as were made for Dr. Hamer?"*

HOYLAND: *"Yes."*

COMPARET: *"Was that the same as the machine used on Mrs. Henderson?"*

HOYLAND: *"No, but the same type."*

SAPIRO: *"These machines are perfectly good, they are just the same as the* [Dr.] *Couche machine and the one that gave Mrs. Henderson such relief."*

www.rife.org

These quotes show that Beam Ray Corporation sold the Clinical style Rife Machines to Dr. Hamer and Dr. Couche. In 1951 Dr. Couche sold one of his Beam Ray Clinical instruments to Dr. Tully.

The photo on the previous page of Dr. Couche's machine and the documents we have read show us that his machine is an original Rife Ray #5 or Beam Ray Rife Machine. With this machine we can make comparisons against it when looking at other instruments.

The next photo, shown below, is a picture of Dr. Rife and Philip Hoyland. Philip Hoyland was Dr. Rife's engineer and business partner in the Beam Ray Corporation. In this photo is an instrument. We will prove that this instrument is also a Beam Ray instrument by making some comparisons with other Beam Ray machines.

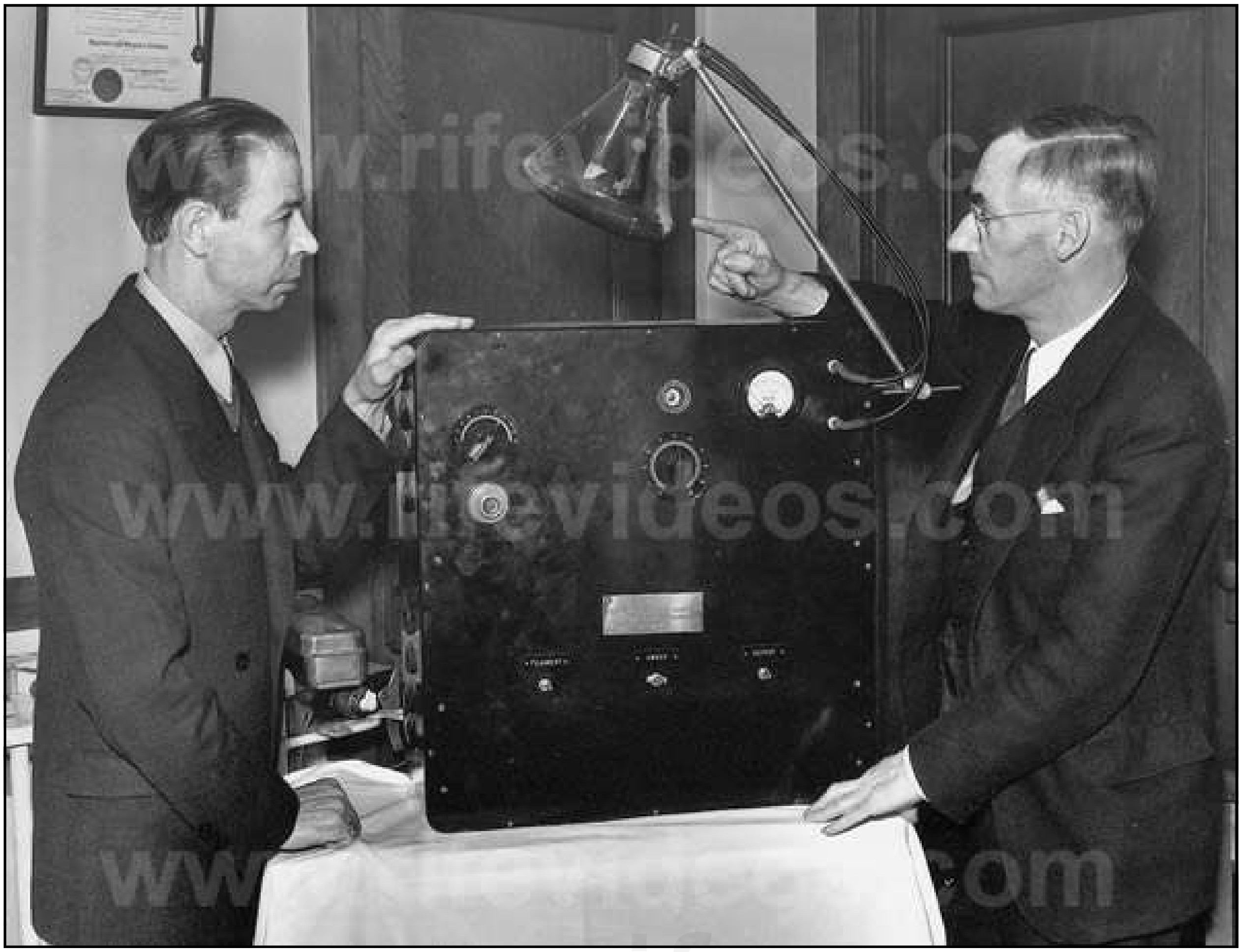

This photo of Dr. Rife and Philip Hoyland was taken for a May 6, 1938 newspaper article published by the San Diego Tribune. In the newspaper the caption below the photo said: "Royal Raymond Rife, left and Philip Hoyland with Rife ray apparatus". On May 1, 1938 the Beam Ray Corporation started selling its Rife Ray #5 or Beam Ray Clinical Machine to doctors. This front page newspaper article had the capability of selling many instruments. It is only logical they would have photographed the instrument they were selling.

The next two photos, shown at the top of the next page, are close-up photos of these instruments. You will notice the similarities of these two instruments. They are almost exactly alike except for the case. Beam Ray used both types of cases with the Rife Ray #5 or Beam Ray Clinical instruments they sold.

www.rife.org

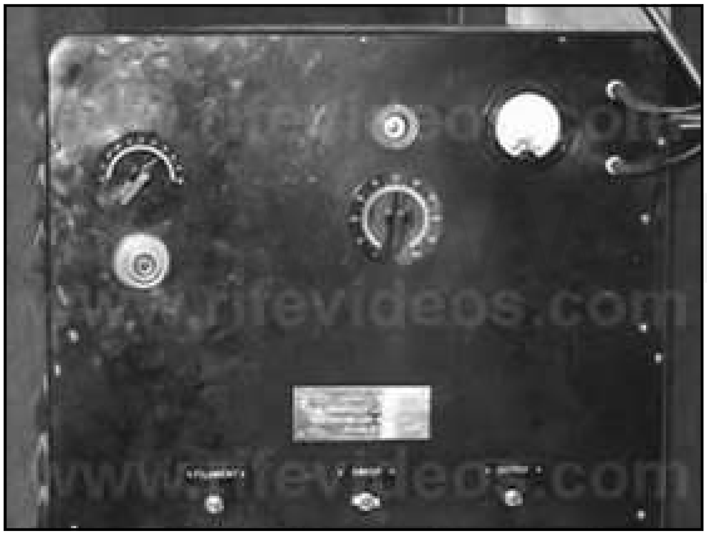

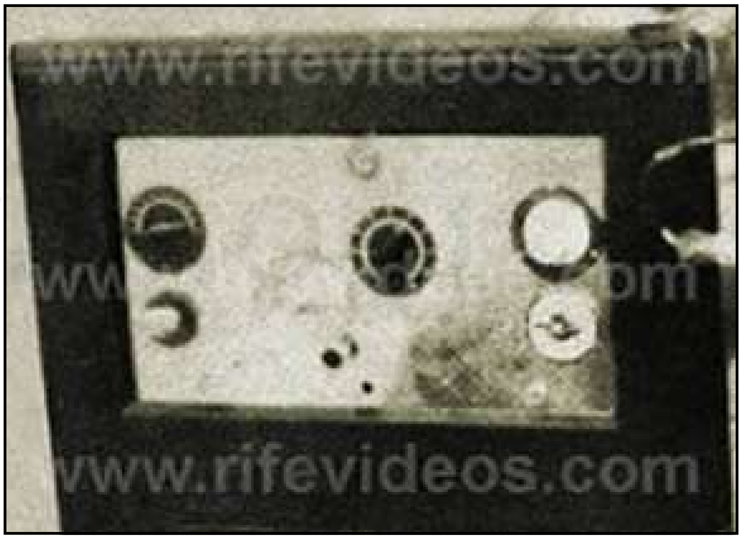

Dr. Couche’s Rife Machine, above on the right, was in a case that extended all the way down to the floor. It had handles on the side and wheels on the bottom which would make it very easy to move around. Both instruments have one oscillator dial which is located on the left side of the front panel. Below that dial on Dr. Couche’s instrument was a four position band switch and on the other instrument, above on the left, is a fine tuning dial. The fine tuning dial was replaced by a four position band switch in order to give the instrument better accuracy. The second dial, in the center of both instruments, goes to 100 and was the amplitude dial. Above that dial on both instruments is a tuning eye for calibrating the instrument’s RF carrier frequency. Below is a photo of one of these tuning eyes.

Both instruments have a milliamp power meter located all the way over to the right next to where the ray tube is connected. Dr. Couche’s instrument had a timer below the power meter to help him make sure he treated the patient for the correct amount of time. Along the bottom are the filament, sweep and output switches which are not clearly marked on Couche’s instrument but we can see what appears to be three different switches, two below the center amplitude dial and one below the timer. The comparison we have just made with Dr. Couche’s Beam Rays instrument shows they are both Beam Ray instruments. Beam Ray Corporation just put this Clinical machine into two different cases.

In the document "Development of the Rife Ray" we have a description of this Beam Ray Clinical instrument:

"In the early part of 1936 Commander Rife and Mr. Hoyland spent much time collaborating on revising some of the applications of the fundamentals of the instruments due to the advancement that had

www.rife.org

taken place in the application of electronics and it was found that the carrier wave used in the previous instruments could be eliminated. During the summer of 1936 further experiments were carried on, which resulted in an entirely new method of generating the desired frequencies and produced a constant input and output in the instruments.

During this work several new test appliances were built for further studying the different frequencies and waveforms noteworthy among these was a 9 inch Cathode-Ray oscillograph of high sensitivity, built for the purpose of photographing the different frequencies on motion picture film and thereby allowing the numerous waves to be studied at will. During the fall of 1936 Dr. Couche of San Diego and Jack Free assistant to Commander Rife conducted a clinic with one of the frequency machines treating experimentally cases of carcinoma and senile cataract..." (Development of the Rife Ray and use in devitalizing pathogenic micro-organisms).

The first thing that we learn from this document is the high voltage current carrier wave that was used to light the ray tube in the Rife Ray #3 and Rife Ray #4 was no longer used or needed with the new Rife Ray #5 or Beam Ray Clinical instrument. The document is correct because this style of instrument only uses an RF carrier frequency to light the ray tube. The second thing we learn is that this instrument was working on an entirely new method of generating the M.O.R. frequencies. This method has been referred to over the years as harmonics. But it is more complex that what was originally believed.

This new method that was used to generate the frequencies has been a mystery for the past 75 years. Finally with the location and purchase of an original Beam Ray Clinical instrument and the use of spectrum analysis the method that Philip Hoyland used has been discovered.

In the first photo, below on the left, is the Rife Ray #5 or Beam Ray Clinical Machine that Dr. Rife and Philip Hoyland were photographed with for the may 1938 newspaper article. In the second photo, below on the right, is a picture of an original Beam Ray Rife Ray #5 instrument. This instrument was obtained from Dr. Larry Low. He has owned it for over 25 years. This is the only known original Beam Ray Corporation Clinical instrument to have survived. It was used by a Medical Doctor who died in the mid 1960's.

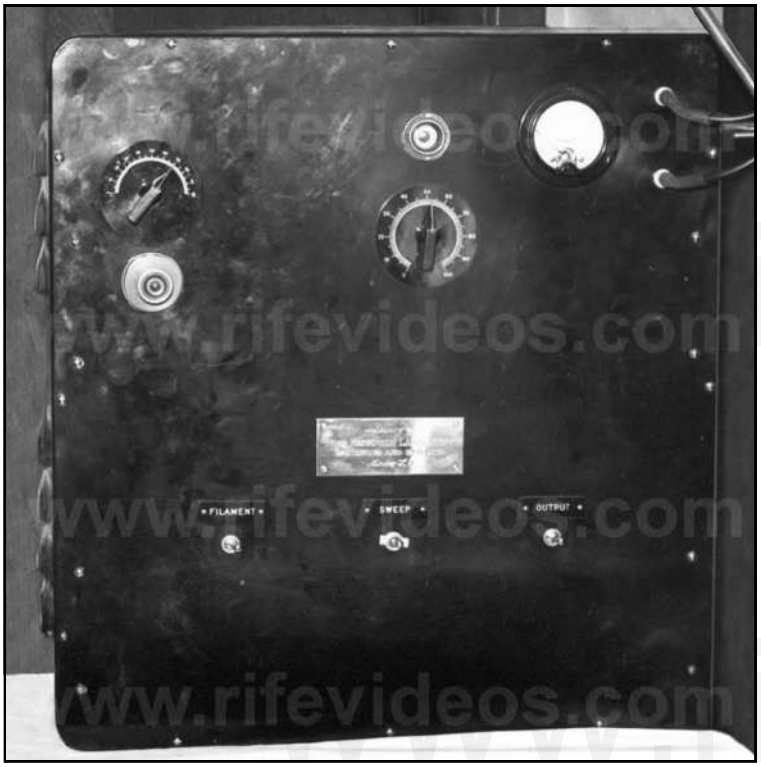

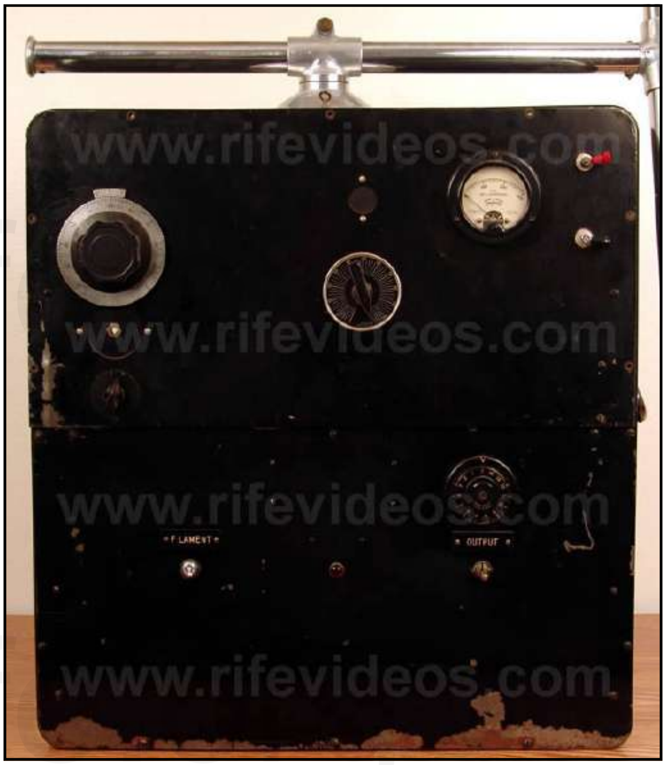

www.rife.org

We would like to thank Dr. Low for allowing us to get this instrument so we could analyze it. It is a low audio frequency instrument which uses an RF carrier frequency. This Rife Ray #5 or Beam Ray Clinical instrument is very important. The significance of this instrument is due to the fact that it is the only known original Rife Ray #5 or Beam Ray Clinical instrument to exist. There were about 14 instruments built by the original Beam Ray Company and until now no one has ever been able to find one. The fact that even one has survived is a miracle. This instrument proves beyond any doubt that Philip Hoyland was the one who first built the low audio frequency instruments such as the Rife Ray #5 Beam Ray Clinical instrument.

The next photo, below on the left, is also a Beam Ray Clinical instrument. This photo, was found inside the case of the original Rife Ray #5 or Beam Ray Clinical instrument obtained from Dr. Larry Low. The second photo, below on the right, is Dr. Couche's Beam Ray machine. Both of these Rife Machines are in the same case. If you look closely you will notice that both instruments have the same metal arm attached to the top of the instrument which holds the ray tube. This same arm is also on the top of the original machine obtained from Dr. Low.

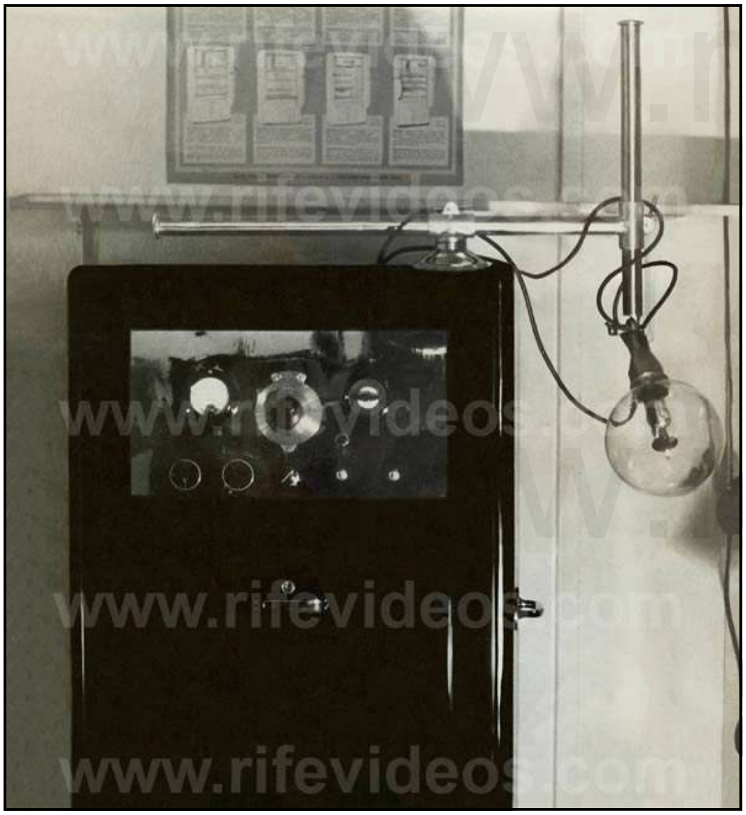

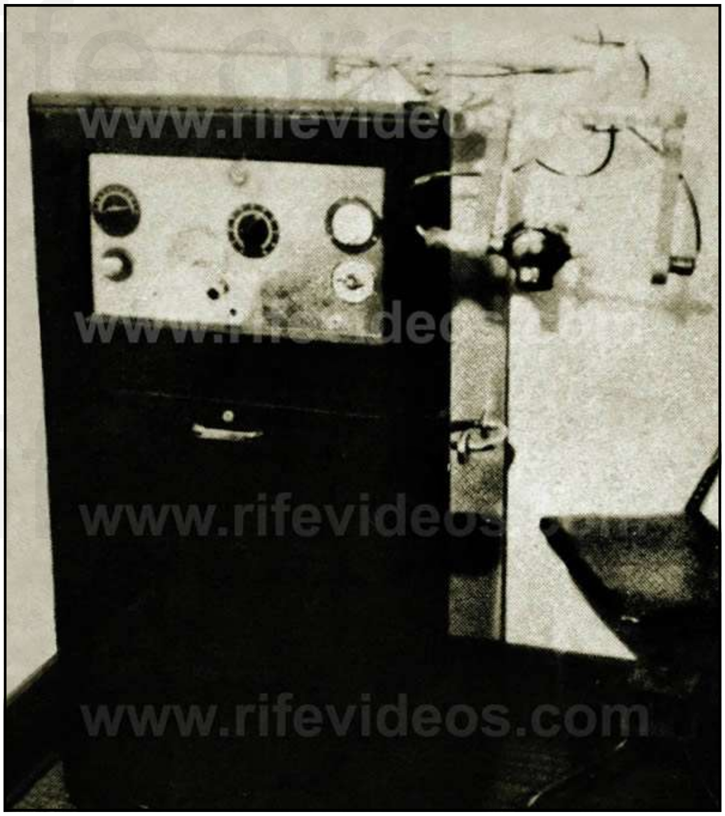

In every detail the cases are the same. The only difference between these two machines is the layout of the front panel. Though the panels are laid out differently both machines have the frequency dial, amplitude dial, band switch, milliamp meter, timer, power light and two switches. It appears when Beam Ray Corporation built their first instrument, which was shown in the May 1938 newspaper photo, they had no band switches dividing out the audio frequency range. The doctors that used these instruments complained about the accuracy problems. It is apparent from the Beam Ray Trial testimony that a four position band switch was added to help stabilize the audio oscillator and hopefully solve this problem. The original Beam Ray instrument that we obtained has a band switch with four settings. Aubrey Scoon’s Beam Ray Clinical replica instrument also has a four position band switch. We will fully evaluate the Aubrey Scoon instrument later in this report. But we will refer to it from time to time as needed as we look at this original Beam Ray Clinical instrument. The 1953 AZ-58 Beam Ray Clinical replica, which we will also evaluate later in this report, had a three position band switch. They only used a three position switch because they used audio frequencies which were ten times lower than the original Beam Ray Clinical instrument. Because of this they did not need or use a fourth band.

www.rife.org

To further analyze the two original Rife machines we will now do a comparison of the faceplates. We will look at the instrument obtained from Dr. Low and the instrument Dr. Rife and Philip Hoyland were standing next to in the 1938 news paper photo.

The first close up photos are of the variable audio oscillator control. The photo, below on the left, is the May 1938 photo. The photo, below on the right, is the original instrument we obtained. Normally the dial should only go 180 degrees from 0 to 100 as shown on the May 1938 dial. Our instrument does not have the original dial because it goes from 0 to 100 in 270 degrees.

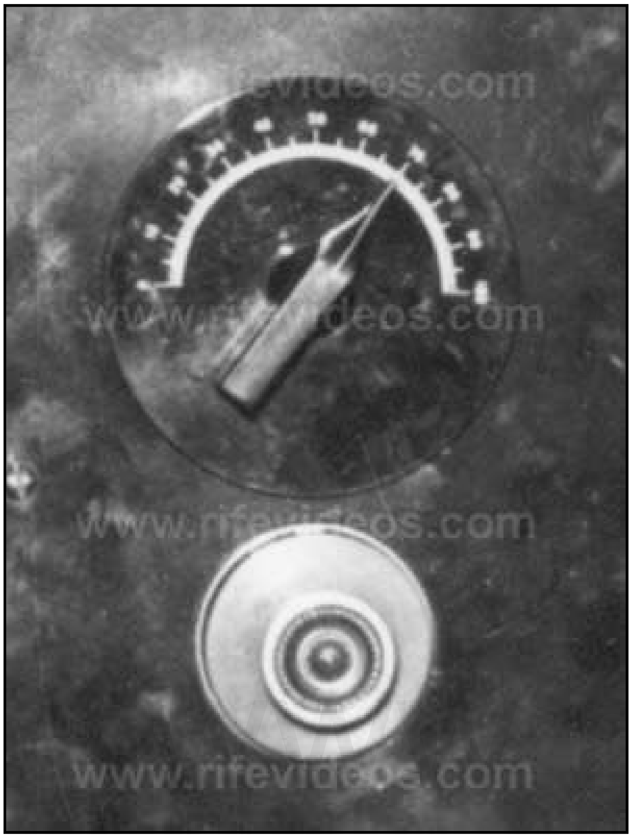

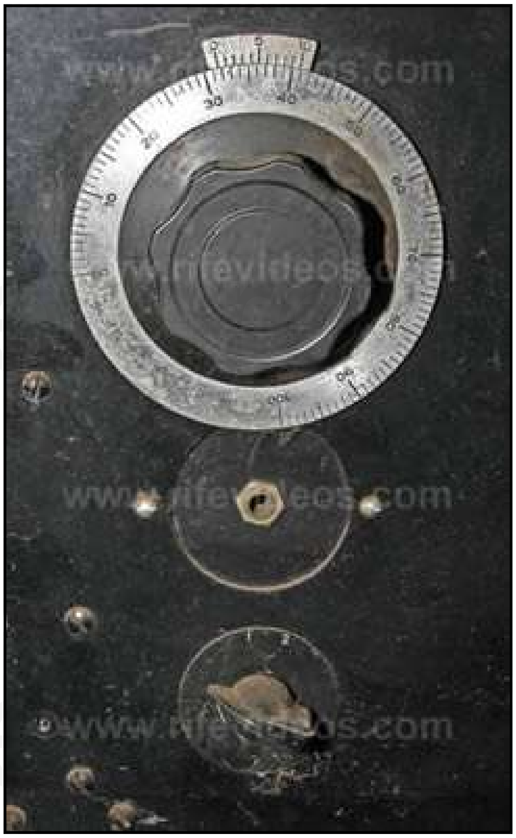

Though the dial shows 270 degrees it will only go 180 degrees. You will also notice that our dial is bigger than the original dial and partially covers where the old fine adjustment dial was located. The fact that the hole is still there, but was covered, indicates that this instrument at one time worked the same as the instrument in the 1938 photo. When they added the four position band switch the fine adjustment knob was no longer needed.

Our Rife Ray #5 or Beam Ray Clinical Rife Machine, like Aubrey Scoon's instrument, had four bands that cover these frequency ranges.

Band 1: 160 Hertz to 820 Hertz.

Band 2: 594 Hertz to 3,190 Hertz.

Band 3: 2,440 Hertz to 12,930 Hertz.

Band 4: 9,430 Hertz to 42,600 Hertz.

www.rife.org

Aubrey Scoon's Beam Ray Clinical Replica instrument band ranges, listed below, were different.

Band 1: 20 Hertz to 200 Hertz.

Band 2: 200 Hertz to 2,000 Hertz.

Band 3: 2,000 Hertz to 20,000 Hertz.

Band 4: 20,000 Hertz to 200,000 Hertz.

These four bands were discussed by Philip Hoyland and Bertrand Comparet during the trial. (#257-260):

COMPARET: *"If you wanted to treat one with typhoid for instance wouldn't you have to set the machine so that it would be on a particular frequency."*

HOYLAND: *"No, the machines were made so that they varied over a band of frequencies."*

COMPARET: *"That band used for the treatment of each disease was different from other bands for other diseases, wasn't it?".*

HOYLAND: *"The whole list of bacteria that the machine was treating was divided into four bands."*

The next two photos, shown below, are of the tuning eye and the amplitude control. The photo, on the left, is from the May 1938 instrument. The photo, on the right, is our original Beam Ray Clinical instrument. Above the amplitude dial on the photo of the May 1938 instrument is the tuning eye.

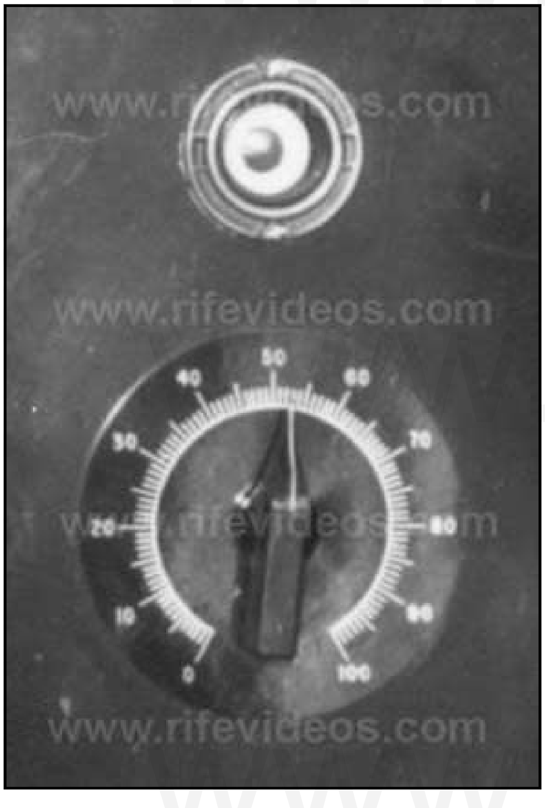

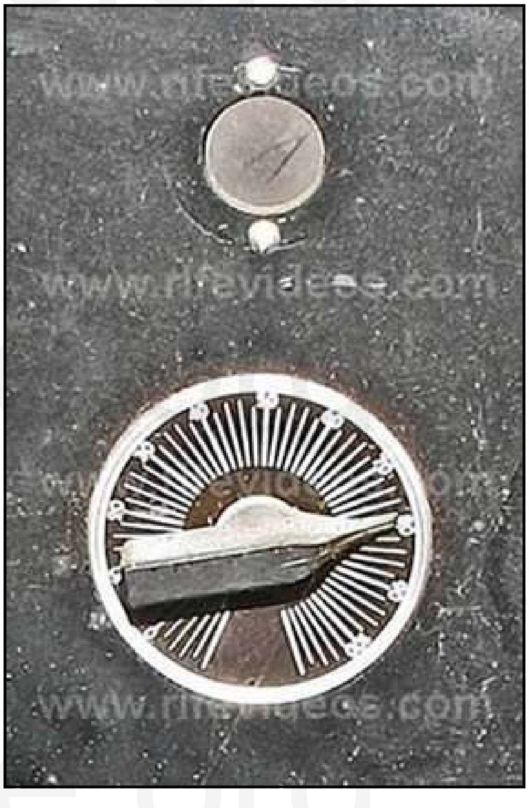

www.rife.org

On our instrument you can see that the tuning eye was removed. Notice that you can still see the outline of the tuning eye mounting ring. On the May 1938 instrument the amplitude dial went from 0 to 100 and adjusted the audio frequency modulation. On our instrument the amplitude dial also goes from 0 to 100 and adjusts the modulation of all the audio frequencies. This modulation is not a true (AM) amplitude modulation. It is more of a pulse width modulation.

The next two photos, shown below, are of the milliamp meter that goes to 300 and the ray tube hookup. The photo below, on the left, is the May 1938 instrument. The photo, on the right, is our instrument.

100
200
300
MILLIAMPERES

Until the discovery of our Rife Ray #5 or Beam Ray Clinical instrument we did not know exactly where the fixed RF tank coil was located. We thought it was probably behind the milliamp meter but now we know this is exactly where it was located.

The next two photos, shown below, are of the plaque. The first photo is of the May 1938 instrument. The second photo is our instrument. These photos show where the original plaque was on the instrument.

The plaque read; “Property of the Rife Research Laboratory, Designers and Builders.” On our instrument the plaque is missing but you can still see the four mounting screws that held the original plaque. Since this plaque had Dr. Rife's name on it this could have caused the doctor problems. So he may have removed it in order to avoid the machine being called a “Rife Machine.”

www.rife.org

In the next two photos, shown below, you can see the “Filament”, “Sweep” and “Output” switches. The first photo is of the May 1938 instrument. The second photo is our instrument. On our instrument the sweep plaque was removed and the switch was replaced with a red light that indicated that the instrument had power.

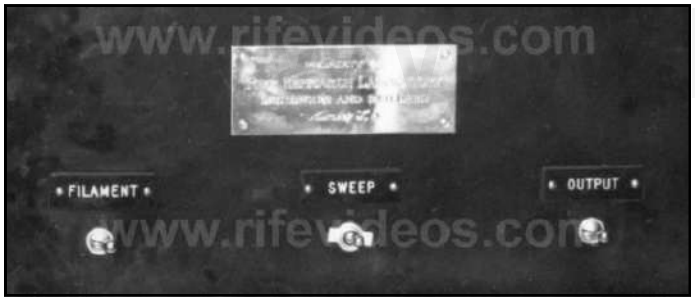

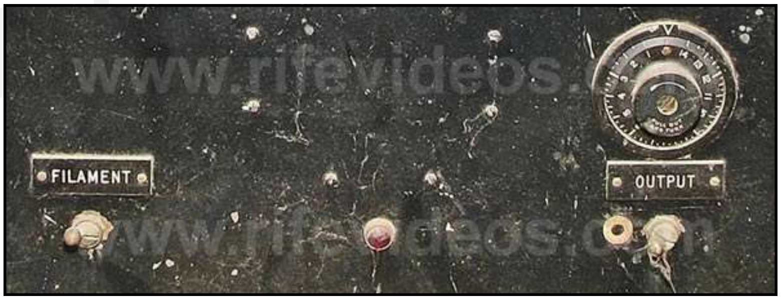

This is where the power indicator light was put on almost all of the instruments built during the 1940's and 1950's. The filament switch turned on the power to the 866 rectifier vacuum tubes and the output switch turned on the power to the 809 main amplifier vacuum tube section. Turning on the power to the 809 tube would light the ray tube. Because of the accuracy problems (due to the old (RC) resistor capacitor audio oscillator) of the Beam Ray audio instruments the sweep switch was probably used to try and help solve this problem. The sweep switch appears to have been removed when the band switch was added. Above the output label on our instrument we see a 15 minute timer for setting the desired runtime for each frequency that was used. The Original 1938 Beam Ray Clinical instrument photo does not have a timer but Dr. Couche’s Beam Ray instrument shows his instrument had a timer on the front panel. Even the photo of the other Beam Ray Clinical instrument which was found inside our instrument has a timer built into it. These photos show that a timer was put in most all of the original Beam Ray instruments. The next photo, shown below, is a larger photo of the Rife Ray #5 or Beam Ray Clinical Rife Machine with the ray tube lit.

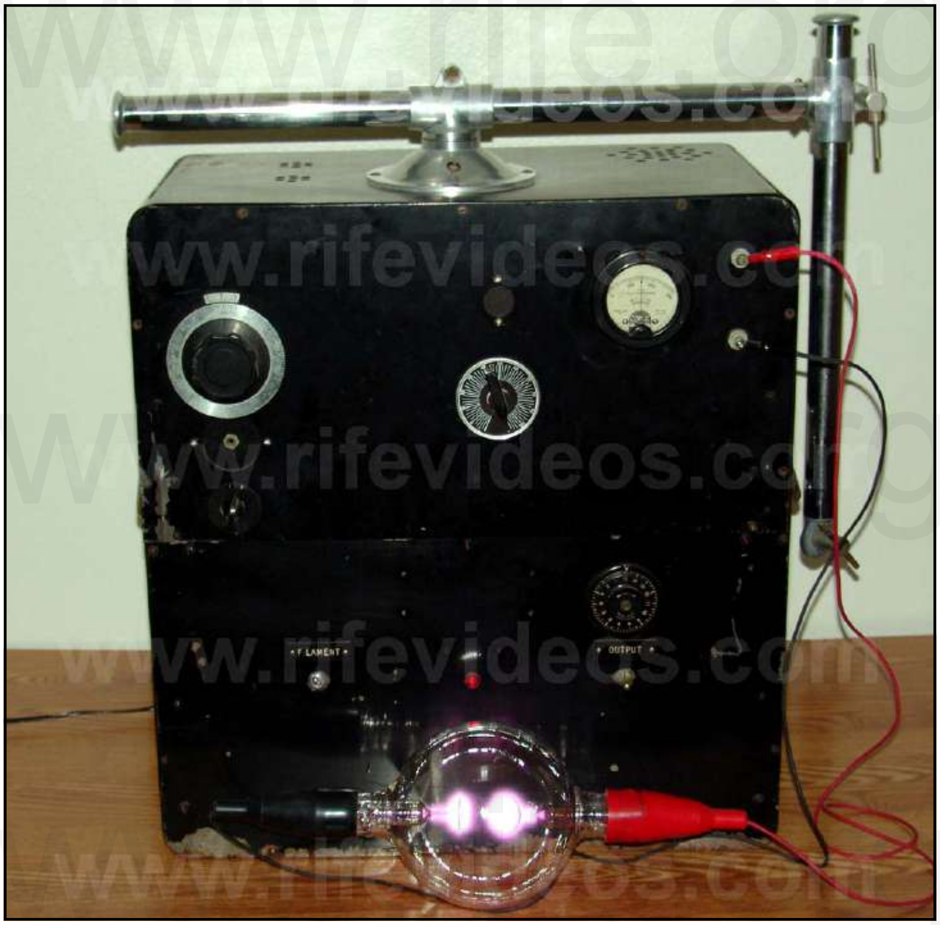

www.rife.org

The next photo, shown below, is of the inside of the Rife Ray #5 or Beam Ray Clinical instrument. The RF tank coil of our instrument was set at 3.80 MHz. The 809 was the main output power tube. There were two 866 rectifier tubes.

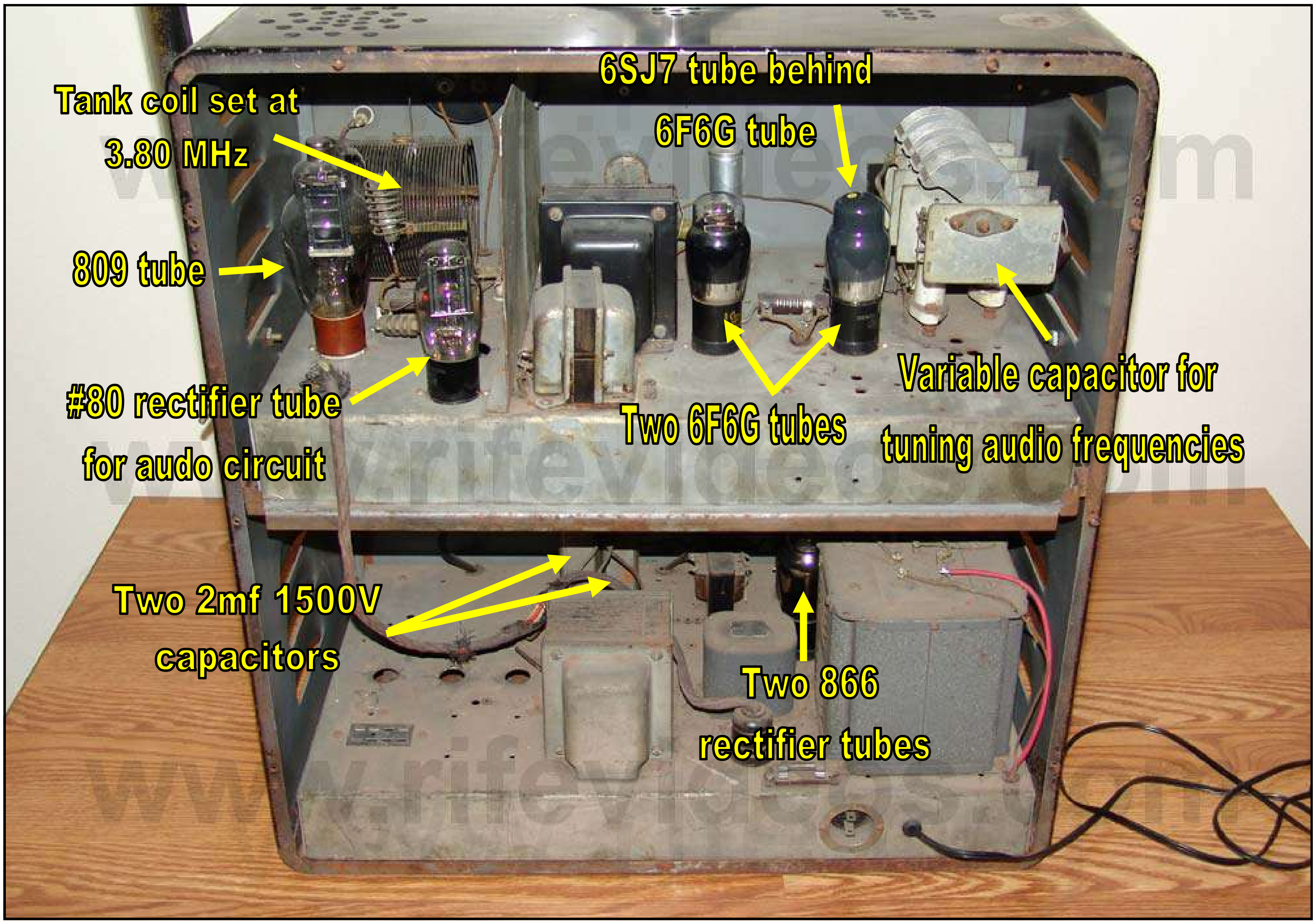
Tank coil set at 3.80 MHz
6SJ7 tube behind 6F6G tube
809 tube
#80 rectifier tube for audo circuit
Two 6F6G tubes
Variable capacitor for tuning audio frequencies
Two 2mf 1500V capacitors
Two 866 rectifier tubes

The audio section consists of 2 6F6G tubes and one 6SJ7 tube. All of the Philip Hoyland audio instrument designs built from 1936 through the 1953's AZ-58 resemble each other. Anyone looking at the different instruments can see that they are all built almost in the same way. Tubes may vary, such as the 812a eventually replaced the 809, but the workings of all the instruments are similar. Both this original Beam Ray Clinical instrument and Aubrey Scoon's Beam Ray Replica instrument have Hewlett Packard sine wave audio oscillators. Mr. Peter's and I was able to repair the audio oscillator and read the different frequency band settings. It was only a 25 watt 10,000 ohm resistor which had burned up that made it so this instrument would not work.

The readings for this report was done with this original Rife Ray #5 or Beam Ray Clinical instrument . Also with Aubrey Scoon's Rife Ray #5 instrument and the AZ-58 replica instrument that we built. They all work identically the same as the original instrument. This original Beam Ray Clinical instrument did not come with any paperwork which gave the dial settings for the various audio frequencies it used. This actually turned out to be for the best because we had no audio frequency data that could have stopped us from discovering how this instrument really worked on Dr. Rife's original high RF frequencies.

www.rife.org

www.rife.org

The Next photo, shown below, is the underside of our original Rife Ray #5 or Beam Ray Clinical instrument. The underside of the chassis shows both the audio section and part of the RF section.

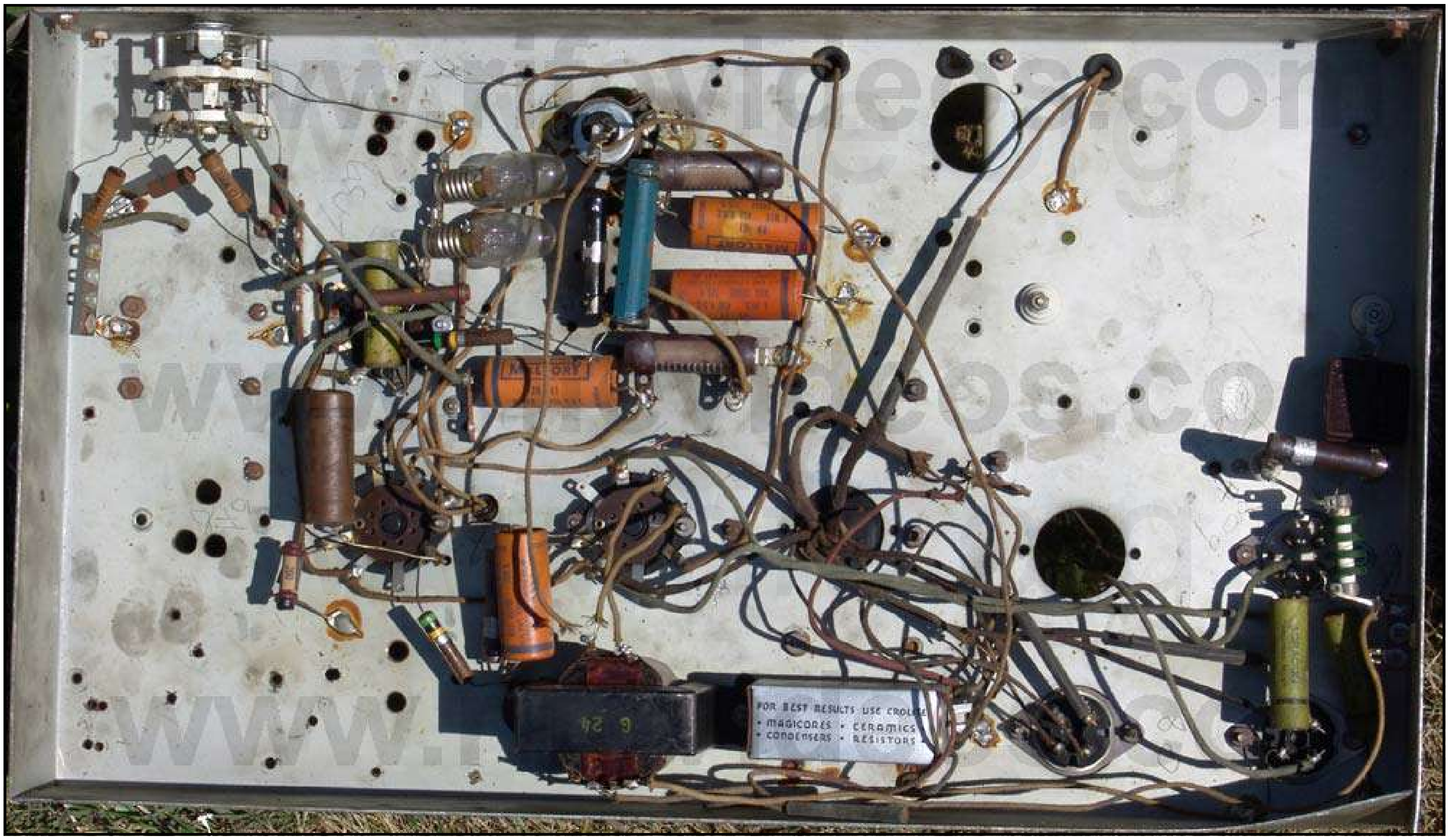

Now that we have an original Beam Ray Clinical instrument we know without any doubt that Philip Hoyland's Clinical instrument design used audio frequencies and it modulated those audio frequencies onto a fixed RF carrier frequency. Dr. Rife and Philip Hoyland had an agreement that they would share evenly on the financial profits of the instruments. Philip Hoyland stated this when he was on the stand during the Beam Ray Trial: (Beam Ray Trial Transcript #505-507)

HOYLAND: *"Dr. Rife and I had always had the understanding that we shared evenly, as I had done all the development work."*

COMPARET: *"What do you mean by that."*

HOYLAND: *"I had done all of the building and designing of the machines other than the one original machine* [Rife Ray #3 which consisted of the Kennedy equipment] *that he had in his laboratory. I had brought that to a state where it could be carried around"* [Rife Ray #4].

From these trial statements we know that all the designs were Philip Hoyland's designs. When Hoyland built the audio instrument he built it on a completely different principle or method. Philip Hoyland had changed the instrument to work on a different method which used harmonics. Dr. Rife believed they were using the Rife Ray #4 RF frequencies along with harmonics. This was pointed out in the 1939 Beam Ray Trial (#1247-1250, 1278-1281):

COMPARET: *"Has the Plaintiff* [Philip Hoyland] *ever informed you that the machines that he designed and built for the Beam Ray were not operating on the same frequencies as your own?"*

RIFE: *"They were supposed to be operating on the same with harmonics."*

COMPARET: *"Did he ever tell you that there was a fundamental difference?"*

RIFE: *"He said on one or two occasions that there was a difference in harmonics."*

SAPIRO: *"You say that the devices that were being built in the early part of 1938, the one that went to Dr. Couche and two that were in the lab were built on new harmonics?"*

RIFE: *"They were built on a different principal, we have a given wave length and it can be produced in different ways, but it should be the same no matter how it is produced."*

SAPIRO: *"You knew that these machines were being built with that machine."*

RIFE: *"Yes"*

Philip Hoyland when he was on the stand was asked (#935-938, 953-958):

COMPARET: *"I understand you say that the frequencies used in the machines put out by the corporation were not set to the same frequencies as Dr. Rife's machines* [Rife Ray #4]*."*

HOYLAND: *"That is correct."*

COMPARET: *"Did you inform the board of directors of Beam Ray that the machine you built was not the same as Dr. Rife's?"*

HOYLAND: *"I had spoken to them about it."*

COMPARET: *"Then it was during the period between September and November that you told Edwards at his home that the machines you were building were not putting out the same frequencies as Dr. Rife's machines?"*

HOYLAND: *"Yes."*

COMPARET: *"How did you explain that?"*

HOYLAND: *"In the summer of 1936 I designed a new machine, or rather I checked it there at the lab* [The Beam Ray Clinical instrument]. *I had designed it in Pasadena, and we tested it out then and the frequencies were not the same as on Dr. Rife's machine."*

COMPARET: *"Did you tell him how great the difference it was?"*

HOYLAND: *"I explained that there was quite a fundamental difference."* [Harmonic frequencies]

Comparet when asked a question by Judge Kelly made this statement. (#2673 & 2685):

COMPARET: *"Hoyland has said that the design and the frequencies of the machine itself is not that of a Rife Ray machine, and that the machine is in fact different. The company will have to have these machines junked, must draw up new designs according to Dr. Rife's ideas, must have Dr. Rife ok these designs, etc…Dr. Rife is not going to be a party to a fraud, and if the machines we sell are not the true Rife machines they are a fraud."* (Beam Ray Trial Papers).

When Edwards was on the stand he stated. (#1384-1385):

www.rife.org

COMPARET: *"Did Mr. Hoyland tell you at any time in the fall of last year that the machines he was manufacturing for Beam Ray corporation operated on a principle fundamentally different from Dr. Rife's machine?"*

EDWARDS: *"No, Mr. Hoyland told me at one time that Dr. Rife thought that he had the frequencies but he didn't have them* [Here Edwards is talking about the Beam Ray Corporation instruments not the Rife Ray #4 instrument because Philip Hoyland said, on the stand, that he gave the Rife Ray #4 frequencies to Dr. Johnson and Dr. Rife in 1935]." (Beam Ray Trial Papers).

Philip Hoyland also stated this when he was on the stand. (#800):

HOYLAND: *"Regarding the frequencies of the machine* [Beam Ray Corporation Clinical instrument], *you will remember me telling you that the frequencies used are not the same ones on the Rife machine* [The Rife Ray #4]. *They* [Meaning the Rife Ray #4 frequencies] *were in the upper bands* [139,000 to 1,604,000 Hertz]." (Beam Ray Trial Papers)

In a letter which he sent to Dr. Gonin in 1939, there are indications that Dr. Rife wanted the so-called harmonics removed:

RIFE: *"I spoke only Friday evening to a Mr. John Chamblin, a radio man now connected with Beam Ray Inc., about the redesign and building of a device according to the old Rife Ray principles; as the present instrument has been so deviated away from that old principle that it is nowhere near the same...those devices which you have are merely working on a harmonic and not a true frequency; and in our research on electronics, we definitely know that there is no possible way of controlling electrical harmonics of a frequency."* (Letter from Dr. Rife to Dr. Gonin, May 14, 1939. Page 1 of 3).

We have just read a lot of trial testimony about how this Rife Ray #5 or Beam Ray Rife Machine worked on harmonic frequencies. Also, in the trial testimony it is mentioned that these Beam Ray audio machines were tested in Dr. Rife's lab to see if they would devitalize microorganisms. It was Philip Hoyland who made the tests using Dr. Rife's organisms. These tests were also repeated again in Dr. Johnson's laboratory in the summer of 1936. From the documents we know that Philip Hoyland put a lot of work into this instrument and didn't finish it until late 1936. Benjamin Cullen said Philip Hoyland spent a lot of time at the lab and stated the following in a taped interview in the 1950's:

CULLEN: *"Philip Hoyland was in there quite a lot...Hoyland developed some few items in the lab...Hoyland seemed to help quite a lot and he got into the bacteriology side with Rife a good deal because Rife had so much to work out...he finally got to the point where he* [Dr. Rife] *had to delegate some of the work."* (John Marsh Rife CDs, CD 6 track 1).

In the Beam Ray Trial manuscript we also read the following. (#91 & 93):

COMPARET: *"Were any experimental activities carried on in the lab?"*

HOYLAND: *"Yes."*

From the trial we learn that Philip Hoyland developed and tested his harmonic instrument in the lab. How could Philip Hoyland have tested it unless he put micro-organisms under the microscope? From the "Development of the Rife Ray" document we learned that:

"During the fall of 1936 Dr. Couche of San Diego and Jack Free assistant to Commander Rife conducted a clinic with one of the frequency machines treating experimentally cases of carcinoma and senile cataract..." (Development of the Rife Ray and use in devitalizing pathogenic micro-organisms).

www.rife.org

From the trial papers we learn that Philip Hoyland didn't tell Dr. Rife what frequencies he was using in the instruments. Dr. Rife thought the instruments were using his frequencies (the upper band frequencies) but with harmonics because this is what Philip Hoyland told him. The information that we now have obtained from this original Rife Ray #5 or Beam Ray Clinical instrument shows that Philip Hoyland's instrument was working on Dr. Rife's principles and on his frequencies but in a different manner than Dr. Rife was lead to believe. This is the reason that the instrument worked so well. Philip Hoyland was still using Dr. Rife's principle of coordinative resonance but hid the truth from Dr. Rife to protect his ownership interest. Philip Hoyland was a business man and Dr. Rife was not.

The fact that these tests were done along with the fact that these instruments were used by many doctors with incredible results show that this instrument which used audio frequencies modulated onto an RF carrier frequency did devitalize microorganisms. Though Dr. Rife did not like the method of harmonic frequencies that Philip Hoyland used it was pointed out in the trial that Dr. Rife knew there were changes. Dr. Rife also didn't think that these changes would make much of a difference as long as the instrument worked: (Beam Ray Trial Transcript #2700)

SAPIRO: *"Dr. Rife said that he knew there were changes made in his machine and that they were not changes that would make any difference. Dr. Rife is a genius but he didn't know how to put the machines in a form that could be used in offices of doctors. These machine are perfectly good, they are just the same as the* [Dr.] *Couche machine and the one that gave Mrs. Henderson such relief."*

It was only with the release of the complete Beam Ray Trial manuscript and other Rife documents; found in California a few years ago that we now know why Dr. Rife continued to have this style of instrument built even in the 1950's. Though Dr. Rife originally did not like Philip Hoyland's Rife Ray #5 or Beam Ray Clinical design instrument, because of the use of harmonics, he later tested it in his laboratory and found it would work as well as his original instruments that did not use harmonics. We will quote the complete letter since all of this information is important:

RIFE: *"My first association with Dr. Yale came through an organization known as the Beam Ray Corporation. In order to acquaint you with the details of the formation of this corporation I shall bring in a little background. My assistant at that time* [1934] *in the laboratory was Philip Hoyland whom I met through Dr, Milbank Johnson M.D., Medical Director for the Pacific Mutual Life Insurance Company. I considered Hoyland as a capable electronic engineer and brought him to my laboratory in Point Loma on Alcott Street, San Diego, Calif. Hoyland became associated with a promoter named Hutcheson and Dr. James B. Couche M.D. They came to me with the idea or forming the Beam Ray Corporation to manufacture and distribute the Frequency Instrument to the medical profession. I gave this considerable thought and came to the conclusion that if these instruments were manufactured and placed into the hands of legitimate and bonafide medical practitioners, my efforts, over a period of years would derive exceedingly more benefits, so I gave this corporation permission to manufacture these devices on two stipulated conditions (1) that they would adhere decidedly to the original basic principles of the Frequency Instrument and (2) that each Frequency Instrument would be thoroughly tested before delivery to determine its true devitalizing power and effect on pathogenic bacteria. And so they went ahead. Three instruments were built. The first two were shipped to England (unwired as Hoyland wanted a trip to England) and the third went to Dr. Hamer M.D. at the Paradise Valley Sanitarium and Hospital. Dr. Hamer was the superintendent there I believe. Hoyland was like many men with whom I have associated over a period of years. In a short time he began changing the basic principles of these instruments according to his own ideas. About this time he became associated with Dr. Yale and Yale ordered and received another or the Fourth Frequency Instrument. Since I was out of the city or San Diego at the time, all of these devices were delivered without being tested by myself.*

At a much later period, I called on Dr. Yale at 333 Plaza in San Diego which was the address of his clinic at that time and told him that I did not feel that the Frequency Instrument had been calibrated

www.rife.org

properly so that it would not work. In the interim, I became associated with another electronist by the name of Verne Thompson, of San Diego. Under my supervision, Thompson rebuilt Dr Yale's Frequency Instrument which I tested in my research laboratory on pathogenic bacteria and the Frequency Instrument proved effective. Then later at different times I had this instrument checked and found it lacking in its ability to devitalize anything. I later learned that Dr. Yale had ideas of his own and would have somebody change the Frequency Instrument to suit his individual whims. I will state here definitely that I have never been associated in any way with Dr. Yale outside of the interest that I have taken in some of the patients of Dr. Yale." (Letter written by Dr. Rife, March 22, 1958).

This document shows that Dr. Rife, after the shutdown of Beam Ray Corporation, eventually tested Philip Hoyland's Beam Ray Clinical design and found that it would devitalize all the microorganisms it was tested on. This also confirms that Philip Hoyland's statement that he tested it out in the laboratory was a truthful statement.

With this information taken from the Rife documents we now know why Dr. Rife considered the Beam Ray Clinical machine his instrument and knew it worked. Add to this the fact that all of the doctors who used it had a great deal of success with this Rife Ray #5 or Beam Ray Clinical instrument design. This is why Dr. Rife, John Crane and John Marsh built this Rife Machine in the 1950's and called it the AZ-58. The only difference in the 1953 AZ-58 design was it only had a three band switch which covered a lower audio frequency range. Other than this change the 1953 AZ-58 was a replica of Dr. Couche's and Dr. Tully's original Beam Ray machine. In a letter dated July 6, 1956 this fact was pointed out:

"The Frequency Instruments used by Dr. Tully and Dr. Couche were built in 1938 and do not apply to this code. However the Frequency Instrument currently built are a copy of these earlier Frequency Instruments and are labeled "For Investigational Use Only" at the present time." (Application letter for approval of device in compliance with California Pure Drugs Act. Page 2).

This information that we have just read which shows that the 1953 AZ-58 was a copy of Dr. Couche's machine has more importance than it appears. We will explain why. The 1953 AZ-58's audio frequencies, which it used, were 10 times lower than the audio frequencies used in Aubrey Scoon's instrument. This correlation is important because it shows a direct link to the original Beam Ray Clinical instrument built by Philip Hoyland. This correlation shows that Dr. Couche's instrument used audio frequencies 10 times higher than the 1953 AZ-58. It also shows that Aubrey Scoon's Beam Ray Clinical replica instrument is also a replica or copy of Dr. Couche's Machine. This information now ties all these instruments together and shows that the original audio frequencies which came from Philip Hoyland and used in the Beam Ray Clinical instrument were the same audio frequencies used in Dr. Couche's machine and Aubrey Scoon's machine. This information also shows that these original audio frequencies were later lowered and used in the 1953 AZ-58 built by Dr. Rife, John Crane and John Marsh. Now we can understand why this letter showing that the 1953 AZ-58 is a copy of Dr. Couche's machine is so important. It ties all these machines together.

Since we now know the history of this instrument and that this instrument worked on Dr. Rife's frequencies we will go back to the harmonic method used by Philip Hoyland. From the documents that we have read we know that Hoyland's machine worked on harmonics. The problem with the Beam Ray instrument is it has to be built a certain way in order for it to work on the correct harmonics. It was from the analyzing of this original Beam Ray Clinical instrument that the mystery of how it worked was discovered. We will show how it worked after we read another important quote. Bertrand Comparet, Rife's attorney who eventually defended Dr. Rife against Philip Hoyland in the 1939 Beam Ray trial said this about Philip Hoyland's Beam Ray Clinical instrument:

www.rife.org

COMPARET: *"Well, none of us know enough about it. Now, I remember at that time Rife saying that Hoyland had not used a simple straight forward circuit, as Rife had used, but he thought he had a short cut, through use of harmonics and so on, and Rife had no faith in Hoyland's circuit".* (1970's Bertrand Comparet interview #28).

This statement by Bertrand Comparet and all the quotes we have read sums up Dr. Rife and his Beam Ray business partners understanding of how this instrument worked on harmonic frequencies. Anyone can see that only Philip Hoyland knew how this machine really worked and he would not reveal the secret to anyone.

This confusion of how the instrument worked still remained throughout the building of the 1940's and 1950's Beam Ray replica instruments and even up until the present day (2010). It appears from the Beam Ray Trial testimony that Dr. Rife and the other owners of Beam Ray Corporation really didn't know how this instrument worked. The trial testimony shows that Philip Hoyland would not tell anyone in the Beam Ray Corporation the frequencies used with the Clinical instrument or explain how it worked on harmonics. John Crane and John Marsh who worked with Dr. Rife didn't understand how it really worked either. If Dr. Rife had really known how it worked then he would not have allowed John Crane and John Marsh to have a variable RF carrier frequency in the instrument. Having a variable RF carrier, in and of its self, is not the problem. Having a variable RF carrier frequency would be a good thing as long as you also understand that the audio frequencies have to be correctly matched to the RF carrier frequency in order to produce Dr. Rife's higher harmonic M.O.R. frequencies. The problems came about because John Crane and John Marsh did not have this understanding. The reason why John Crane and John Marsh put a variable capacitor in the instrument was so that they could tune the RF carrier frequency in order to make the ray tube brighter. By doing this it also made it so they could change the RF carrier frequency from about 4.9 MHz to about 2.2 MHz. This change showed that they did not understand how the instrument worked. They mistakenly believed that the audio frequencies were the M.O.R.s. or the frequencies that would devitalize the microorganisms. They also mistakenly believed that the RF carrier frequency was not important. This belief shows without any doubt that they did not understand how the Beam Ray Clinical instrument worked. If they had really understood how it worked they would have never called any of the audio frequencies M.O.R.s.

At this point we need to explain a few things. Philip Hoyland had very good reasons why he wanted to hide how this new Rife Ray #5 machine worked. He became Dr. Rife's Engineer in 1935 when he built the Rife Ray #4. In 1936 Philip Hoyland also began building the Beam Ray Clinical instrument that would be sold in 1938 by Beam Ray Corporation. Philip Hoyland was worried about keeping the original frequencies a secret because he felt people would try to steal their technology. This concern of Philip Hoyland's was not unfounded because Mr. Parsons of the British Group did try to steal their instrument. From the trial transcript we learn they had no way to patent the instrument because everything they were doing was in public domain in regards to the frequency generating equipment. Even the frequencies themselves cannot be patented. Philip Hoyland felt that he had to come up with a way to keep anyone from finding out what the true frequencies were. So he built the instrument a different way using harmonics to hit the harmonic frequencies of the Rife Ray #4 and Rife Ray #3 Kennedy equipment. Until a genuine Beam Ray Clinical Rife Machine could be found and tested, we would never know for sure how Philip Hoyland generated and used the harmonics in his instrument. Even though we now know that Aubrey Scoon's instrument is a Beam Ray Clinical replica no one knew for sure that it really was a genuine replica when he obtained it. This confusion and the lack of having the chronological history of when and what type of instrument was built made it very hard to find the truth.

Not only would Philip Hoyland not tell anyone how his instrument worked he also would not let anyone know the new frequencies he was using. Everything he did hid the frequencies. In the Beam Ray Trial manuscript we read that he always used a code to give the frequencies. This code would work with the dials. No digital readout was available in those days like we have today.

Since no one was ever given the frequencies from Philip Hoyland this has led to many problems. Some of the later Beam Ray Clinical instrument replicas have different audio frequencies and this also has led to a lot of confusion. Even John Crane was sending people different audio frequencies other than the standard set he used with the 1953 AZ-58. Because of the different audio frequencies used in these Beam Ray replicas we really didn't know which set of audio frequencies Philip Hoyland really used. So it was not possible to use any of these frequency sets and come to any reasonable conclusions.

When we obtained this original Rife Ray #5 or Beam Ray Clinical Rife Machine, from Dr. Low, we had to put aside all the frequency lists and thinking of the past in order to figure out how the instrument worked. From everything we had read about Philip Hoyland's instrument we had come to the conclusion that the answer would be found in the math. However Hoyland came up with his idea it had to be a mathematical method. Like many others we tried to reconcile the audio frequencies as lower harmonics of Dr. Rife's original high RF frequencies. But this proved not to be the case. We tested the frequencies that were used in Aubrey Scoon's Beam Ray Clinical instrument and they didn't match either as harmonics of Dr. Rife's original frequencies. We did the same with the AZ-58 audio frequencies and they didn't match. We took the other audio frequencies that Crane had a list of and they also didn't match. There is an instrument known as the 1947 instrument which we tried to reconcile with no success either. With all these audio frequencies only a few were close harmonic matches. If all these audio frequencies were true harmonics of Dr. Rife's original high RF frequencies then they should have harmonically match up, but they did not.

Once we put all this aside and began to analyze this instrument we found out how it worked. This Beam Ray Clinical instrument is truly a harmonic instrument and works on harmonic frequencies of Dr. Rife's original frequencies which he found. Philip Hoyland was telling the truth when he told Dr. Rife that the instrument was working on harmonics. The only thing was Philip Hoyland would not tell exactly how this was being done. The fact that both the audio frequency and the RF carrier frequency were sine wave did not make it easy to understand how it worked. How could it be a harmonic instrument when it used a sine wave waveform for both the audio and the RF frequencies? The 1953 AZ-58 Beam Ray Clinical replica worked on square wave audio frequencies and the harmonics came from the square wave waveform. But this is not how the Beam Ray Clinical instrument was suppose to work.

The 1950's square wave method has been used ever since the 1950's. But how can an instrument that uses sine wave audio frequencies be a harmonic instrument? With this understanding we knew that the audio frequencies could never be the true M.O.R.s. But how did they work in the instrument to produce Dr. Rife's M.O.R.s? The harmonic square wave method has been an accepted method for many years. For many years almost everyone has understood and has generally accepted that you can take a square wave audio frequency and the harmonics it creates and hit a higher frequency M.O.R. through those harmonics. These audio frequencies are usually many hundreds of harmonic steps lower that the higher M.O.R. frequency. This harmonic method must be kept in mind as well as the fact that not once, but several times, in the Beam Ray Trial manuscript and other documents everything said about this machine showed that it was using the harmonic method. We must remember that Dr. Rife said that he believed that many of his frequencies were sub-harmonics of a higher frequency. He said if we knew the true higher frequency it may even work better. Because of how this instrument works it is logical to conclude that Philip Hoyland understood this concept and apparently found through testing that all of Dr. Rife's frequencies were sub-harmonics of higher frequencies. With this understanding he built the Rife Ray #5 or Beam Ray Clinical instrument. Philip Hoyland was being truthful when he said that this machine was using harmonics. With all this knowledge from the Trial testimony and the Rife documents we how have a good understanding of the facts behind this new instrument design. Here is what we found when we analyzed the Beam Ray Clinical instrument.

www.rife.org

The graph, below on the left, is of the harmonics from the 3.80 MHz RF carrier frequency coming out from the ray tube of our original Beam Ray Clinical instrument. It was taken using a PicoScope 3205 spectrum analyzer. The first photo, below on the right, with the black waveform comes from Dr. Rife's 1936 lab film. The photo to the right of that photo is an oscilloscope reading that we took of the waveform of the Beam Ray Clinical Rife Machine. You will notice the similarity of the waveforms. All the M.O.P.A. (Master Oscillator Power Amplifier designs used by Philip Hoyland) RF carrier waveforms we have tested look like this. When you put a non harmonic sine wave into a ray tube you will always see a distortion of the sine wave which will produce both odd and even harmonics through the ray tube. This is the reason the sine wave looks distorted.

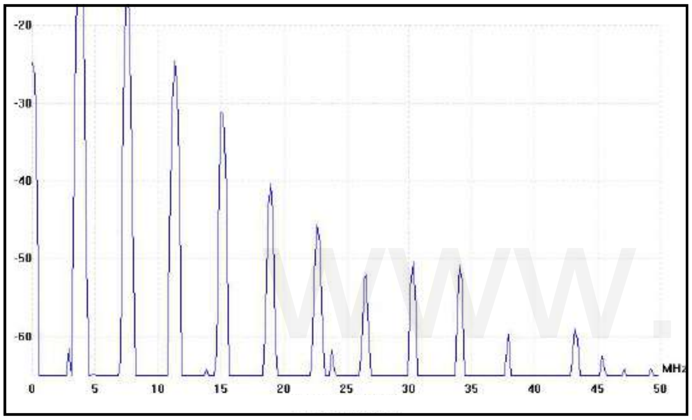

Photo one: From Rife's lab film.

Photo two: Beam Ray carrier before ray tube.

Photo one: From Rife's Lab film.

Photo two: Beam Ray carrier after ray tube.

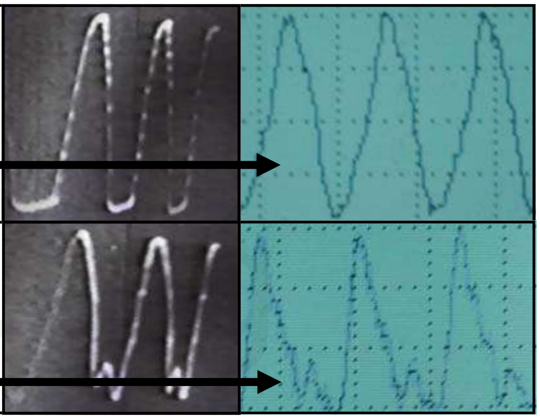

The graph, below on the left, is of a pure 2000 Hertz sine wave frequency. The oscilloscope photo, below on the right, shows a 2377 Hertz sine wave waveform coming from the audio oscillator of the original Beam Ray Clinical Rife Machine. The small photo shows it after it goes through the ray tube. You will notice that it almost looks like a square wave, but it is not. The circuit creates this waveform.

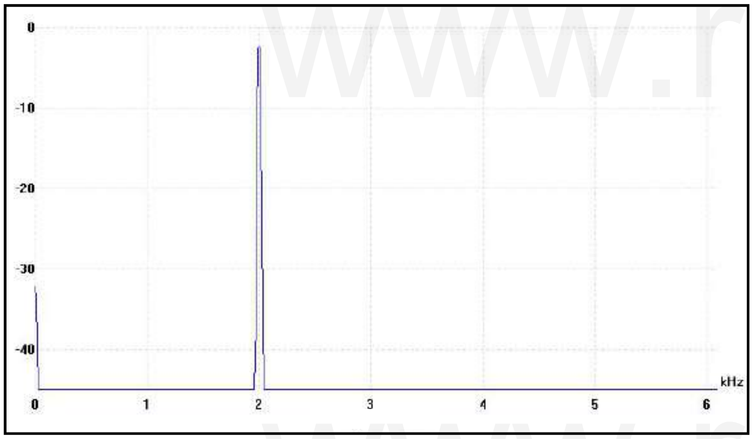

Beam Ray sine wave audio frequency before circuit and ray tube.

Beam Ray modulated sine wave audio frequency after going through circuit & ray tube

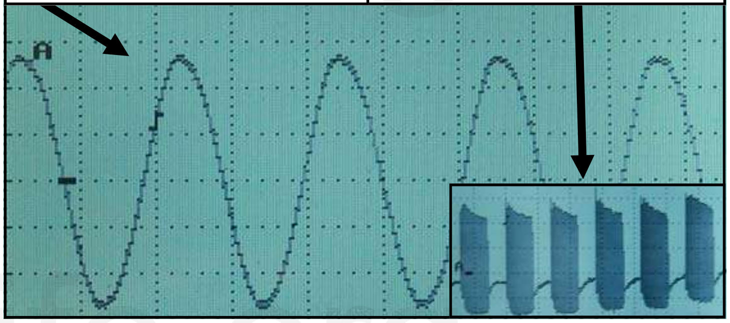

The next two graphs we will look at are on the top of the next page. The graph, on the left, was before modulation. The graph, on the right, during modulation. When we modulated 40,000 Hertz on a harmonic sine wave 3.80 MHz RF carrier frequency this is what the PicoScope spectrum analysis showed coming out of the ray tube. This was interesting to see. Not only did it create sideband frequencies 40 thousand Hertz above and below the harmonic RF 3.80 MHz carrier frequency but it created many harmonic sidebands every 40,000 Hertz. These harmonic sidebands covered a large frequency range of hundreds of thousands of Hertz.

www.rife.org

www.rife.org

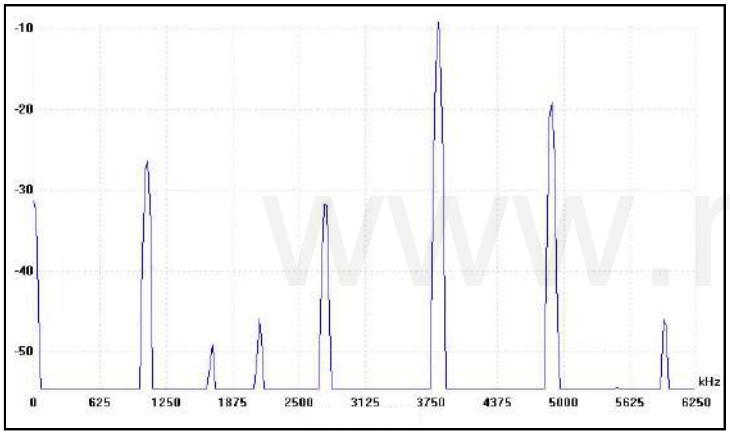

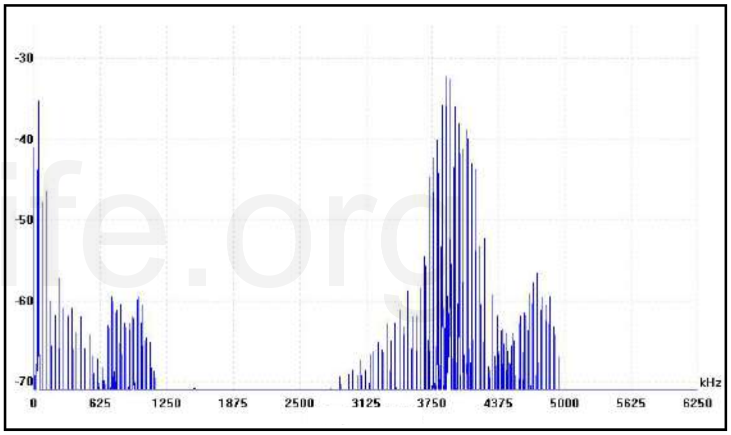

Below are two more graphs showing a closer view of these sidebands that were created from one audio frequency. The graph, below on the left, is with a 50,000 Hertz sine wave frequency. The graph, below on the right, is with a 100,000 Hertz sine wave frequency. This was done with a harmonic sine wave 2.4 MHz RF carrier frequency so you can see a closer view of these sidebands. When we saw this we knew there was only one way this Beam Ray Clinical Rife Machine could work on Dr. Rife's frequencies and principles. We will now explain how Philip Hoyland's Rife Ray #5 or Beam Ray Clinical instrument worked using a fixed RF carrier frequency of 3.80 MHz and a variable audio frequency. Philip Hoyland knew from working with Dr. Rife that his frequencies or M.O.R.s. were subharmonics of true higher frequencies. Understanding this must have given him the idea of how to build the new Rife Ray #5 or Beam Ray Clinical instrument. If we also understand this concept, that all frequencies have lower and higher harmonic frequencies, is half the key to understanding how Philip Hoyland made this machine work. What Philip Hoyland did was multiply Dr. Rife's original M.O.R. frequencies up in harmonic steps until he had the highest harmonic frequencies closest to 3.80 MHz. From the Beam Ray Trial quotes we know that Dr. Rife was told by Philip Hoyland that this Beam Ray Clinical machine was working on his frequencies with harmonics. What Philip Hoyland would not do is clarify exactly how it was working on harmonics of Dr. Rife's frequencies. Philip Hoyland also stated that his frequencies were not the same frequencies as Dr. Rife's frequencies. He was being truthful when he said this but what he would not reveal is the fact that his frequencies were exact harmonics of Dr. Rife's frequencies.

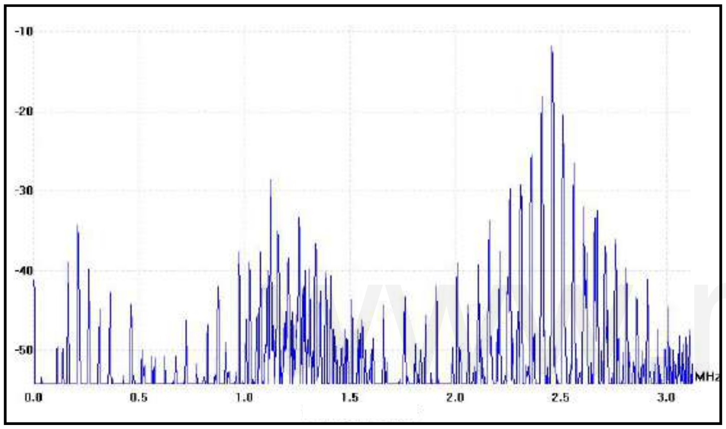

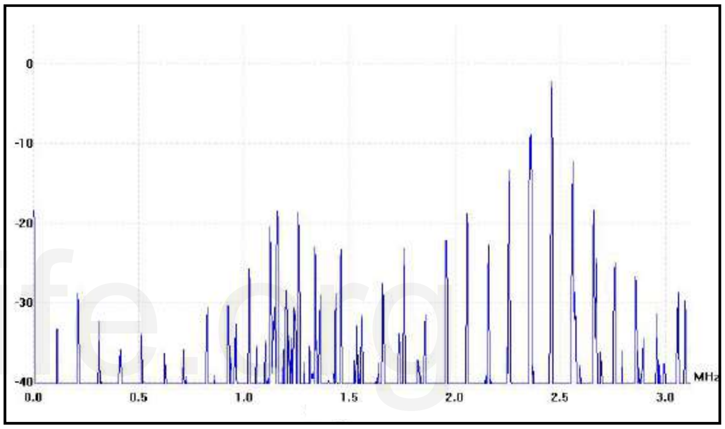

To clearly explain how Philip Hoyland used Dr. Rife's frequencies to come up with his new frequencies which were to be used in the Rife Ray #5 or Beam Ray Clinical instrument we will use the Rife Ray #4 frequency for Streptothrix as an example. That frequency was 192,000 Hertz. Philip Hoyland multiplied 192,000 Hertz by 20 times which will give you the 20th harmonic frequency at 3,840,000 Hertz. Since the RF carrier frequency of the Beam Ray Clinical instrument was fixed at 3,800,000 Hertz the difference would only be 40,000 Hertz between the two frequencies. This method of multiplying Dr. Rife's frequencies up in harmonic steps and using those frequencies was an ingenious method used by Philip Hoyland. Even though this was an igneous method it is only half the mathematical equation. The second half of the mathematical equation is probably even more ingenious than the first part.

Philip Hoyland then put into this Beam Ray Clinical instrument an audio oscillator. We must keep in mind that none of Dr. Rife's frequencies that were transferred from the Rife Ray #3 to the Rife Ray #4 and then to the Rife Ray #5 or Beam Ray Clinical instrument were audio frequencies. The audio oscillator was put into this instrument to accomplish the second half the mathematical equation. Using the audio oscillator all Philip Hoyland had to do was use an audio frequency of 40,000 Hertz. The 40,000 Hertz frequency would then create what are called sideband frequencies. These sideband frequencies would be created by the 40,000 Hertz both above and below the RF carrier frequency being used. The first upper sideband frequency would then be 40,000 Hertz above the RF carrier frequency and hit the harmonic M.O.R. frequency of Streptothrix at 3,840,000 Hertz. This method of using sidebands was the second half of the mathematical method used by Philip Hoyland in this Beam Ray Clinical instrument. Now for further clarification of these sidebands, as shown in the photo on the previous page, which extend out many times above and below the RF carrier frequency depending on the audio frequency which is used. They are like pickets on a fence and every picket represents a harmonic sideband created from the audio frequency. These harmonic sidebands are part of the harmonics that the Rife Ray #5 or Beam Ray Clinical instrument used. So Philip Hoyland took Dr. Rife's frequencies and multiplied them up in harmonic steps as close as he could get them to the fix 3.80 Megahertz RF carrier frequency. Then these new higher harmonic frequencies of Dr. Rife's original frequencies then became Philip Hoyland's new M.O.R. frequencies which he would not reveal to Dr. Rife or any of his Beam Ray Corporation partners. Next he then used harmonic sidebands to hit those new higher frequencies.

Though Philip Hoyland's new frequencies were harmonic frequencies obtained from Dr. Rife's original frequencies they did not create harmonics in the Beam Ray Clinical instrument. The harmonics came from the RF carrier frequency and the harmonic sidebands created from the audio frequencies, which Philip Hoyland used. By using both of these methods combined into and instrument Philip Hoyland could then say that the Rife Ray #5 or Beam Ray Clinical instrument was a harmonic instrument capable of outputting Dr. Rife's frequencies. His method was truly a harmonic method but he kept it as a secret form Dr. Rife and the other owners of the Beam Ray Corporation. The sideband method that Philip Hoyland used was the new cutting edge of electronic technology back in 1936. There would be very few people who could understand what Philip Hoyland was doing unless it was fully explain to them as we have explain here.

We will now explain Philip Hoyland's method further so that there can be no misunderstanding. But first what needs to be pointed out here is this IMPORTANT fact which came from the analyzing of this Beam Ray instrument. Neither the harmonic 3.80 MHz RF carrier frequency nor the audio frequencies will do anything by themselves. But when the harmonic 3.80 MHz RF carrier frequency and the audio frequencies are combined together they will produce many sideband frequencies. And one of these sideband frequencies will line up with the true higher harmonic Rife M.O.R frequency and devitalize or render harmless the harmful microorganism. To re-emphasize this so that no one misunderstands. If you just use the audio frequencies by themselves you will get nothing. If you just use the 3.80 MHz RF carrier frequency without the audio frequencies you will get nothing. The audio frequencies used in this instrument must have the RF carrier frequency of 3.80 MHz or they will not produce Dr. Rife's frequencies. This is the reason the 1953 Beam Ray Clinical instrument called the AZ-58 did not work properly. In that instrument they were only using the audio frequencies without correctly matching them to the RF carrier frequency. In other words they were not using the sideband method that Philip Hoyland originally developed to be used in this style of instrument.

We know that Philip Hoyland was trying to hide the true M.O.R. frequencies of the organisms from anyone who used the equipment. Twice in the Beam Ray Trial it was mentioned that Dr. Rife had no ability to patent the Rife Ray tube instrument. The only secret was the frequencies and Philip Hoyland was trying to protect his and Beam Rays interests. The third and final secret that Hoyland used to hide the M.O.R. frequencies will now be explained. Keeping the harmonic sideband method in mind

www.rife.org

Philip Hoyland could have just divided the 40,000 Hertz by two and used a 20,000 Hertz audio frequency. Then you would have one upper sideband at 3,820,000 Hertz and the second upper harmonic sideband at 3,840,000 Hertz. The 3,840,000 Hertz sideband would hit the 20th harmonic of 192,000 Hertz. Philip Hoyland could have divided it by three and used a 13,333 Hertz frequency. He also could have divided it by 4 and used a 10,000 Hertz frequency. If he would have divided it by five he could have used an 8,000 Hertz frequency. Divided by six he could have used a 6666 Hertz frequency.

We could go on but anyone can see the many variable frequencies that could have been used to create the correct sideband frequency. Also, because of the “one tenth of one meter” factor that Dr. Rife mentioned you could add a few Hertz to each frequency without changing the frequency enough to make it so it would not work. This means you could change the 6,666 Hertz to 6,669 Hertz just to make things a little more confusing. By using this method Philip Hoyland could use many different audio frequencies to produce which ever number of sidebands he wanted to use. This would make it impossible to figure out which sideband was hitting the M.O.R. frequency that would devitalize the organism. By using this new method of using higher harmonic frequencies of Dr. Rife’s original frequencies and then using harmonic sideband frequencies Philip Hoyland was able to build a machine that would hide the frequencies from anyone. Philip Hoyland's method as anyone can see was ingenious.

We know that Philip Hoyland's method worked very well except that the technology of the mid 1930's did have a few drawbacks. Dr. Couche said that he had to sweep the frequency dial in order to get his instrument to work consistently. This is understandable since the 3.80 MHz RF carrier frequency will wander in a sweeping motion five hundred to one thousand Hertz up or down from the fixed RF carrier frequency. The audio frequency oscillator also varied. Bertrand Comparet in his 1970's interview said this about the Beam Ray Clinical instruments inherent frequency drift:

COMPARET: *“Well, as they warmed up they’d shift frequency...Now, whether this was Hoyland’s inability to do better, or whether it was just inevitable in those days, I don’t know, but Hoyland’s devices did have that frequency shift as they warmed up. So, they had their problems. Now what Couche did, see, he would have cases where he would get an instantaneous cure, like that, and other times when the treatment just didn’t produce any results, because of the frequency shift. So, he would start in, he had from Rife (Hoyland dial settings) a set of the frequencies for several different diseases and he would tune it deliberately to one side of that frequency and then gradually tune it across to the other side making sure that somewhere in the process he crossed the correct frequency, even if the instrument wasn’t exactly in tune anyway. Well, when they hit the exact frequency they got amazing results.”* (1970’s Bertrand Comparet interview #10).

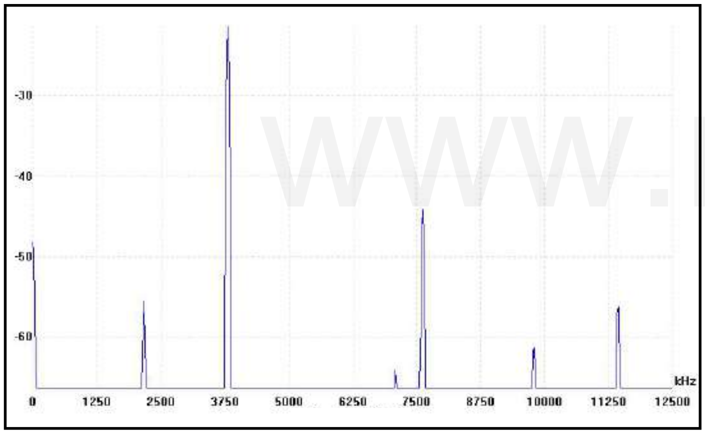

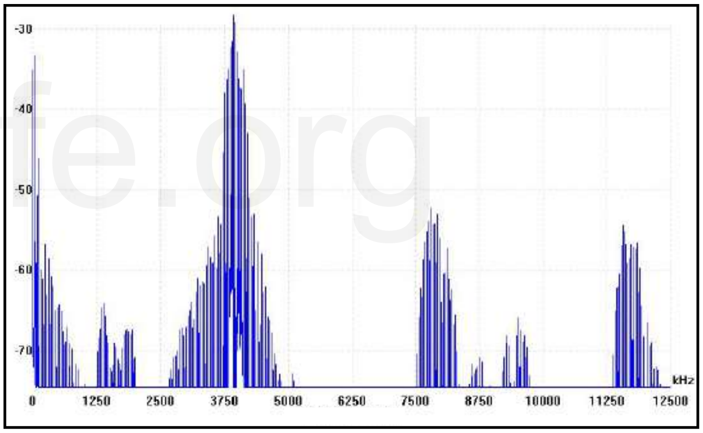

Again if you look at the first graph, above on the left, you will also notice that there are harmonic frequencies created from the 3.80 MHz RF carrier frequencies at 7.60 MHz and 11.40 Megahertz. These sideband frequencies as shown in the above graph, on the right, are created not only for the 3.80 MHz RF carrier frequency but its entire harmonics. These harmonics continue all the way up to

www.rife.org

about 12 MHz with reasonable power. This creates an interesting effect and shows that all the harmonics of 192,000 Hertz are being hit over the whole spectrum of about 12 Megahertz. This is probably why this Beam Ray harmonic instrument worked as well as it did, when they hit the correct frequency. This Rife Ray #5 or Beam Ray Clinical machine was definitely a harmonic instrument as Philip Hoyland had stated. It may have even exceeded Philip Hoyland's original concept.

Now going back to the audio frequency method of creating the sidebands that Philip Hoyland used. No one could ever determine the 20th harmonic frequency of 3,840,000 Hertz or the original Rife Ray #4 frequency of 192,000 Hertz using this method. The secret of the Beam Ray Clinical instrument and the frequencies that would devitalize the microorganisms could never be figured out unless you had the original Rife Ray #4 frequency of 192,000 Hertz. Only a spectrum analysis of this instrument made it possible to figure out what Philip Hoyland was doing. Without this modern technology a person would have to be very knowledgeable about how frequencies and their harmonics worked.

The fact that it has taken this long, almost 75 years, to figure out how this Rife Ray #5 or Beam Ray Clinical Rife Machine worked is proof enough that the secret was well hidden. As pointed out earlier, Philip Hoyland would never tell anyone how this instrument really worked. If anyone changed the 3.80 MHz RF carrier frequency then the audio frequencies would not work. New audio frequencies would have to be calculated to match the new RF carrier frequency. Also, Philip Hoyland could have changed the audio frequencies any time he wanted for any machine in order to confuse anyone who had the equipment. Since the instrument that we obtained from Dr. Low is an original Beam Ray Clinical instrument then one would assume that its 3.80 MHz RF carrier frequency is the carrier frequency that was used with all the Beam Ray Clinical instruments. But we know that this is not the case because they used a different carrier for other machines. None of the audio frequencies used in all the replica instruments from the 1940's to the 1953 AZ-58 match up to the 3.80 MHz RF carrier frequency. With this understanding we also have to assume that the audio frequencies used in this original Beam Ray Clinical instrument were different than the 1940's and 1950's instruments. Aubrey Scoon's instrument had a 3.30 MHz carrier frequency so if his machine is a replica of one of the original machines, which it is, then the audio frequencies would logically have to be matched to its 3.30 MHz carrier frequency, which they are.

The 1953 AZ-58 technical data shows that it used a 4.68 MHz RF carrier frequency. But when the variable capacitor was put into the AZ-58, in about 1956, the carrier frequency then became variable. This made is so Dr. Robert P. Stafford could change the RF carrier frequency to different carrier frequencies. In Dr. Robert P. Stafford's "Electromagnetic Field Therapy" report he said the following about the carrier frequency they used on his patients with the 1953 AZ-58:

DR. STAFFORD: *"Radio Wave transmission is used as a carrier wave. We use between 3100 KC [3.10 MHz] and 3300 KC* [3.30 MHz] *(This does not appear to be a critical value). The carrier wave is modified with specific cycles per second modulations. We believe that the CPS [audio frequencies] is a critical value and it actually may prove to be the most important factor which this research may offer."*

You will notice that Dr. Stafford did not think that the RF carrier frequency was critical. He clearly states that they believed it was the "CPS" or cycles per second of the audio frequencies which were the most important part of making the AZ-58 work. If the AZ-58 was going to work on the sideband method that Philip Hoyland developed then the RF carrier frequency and the audio frequencies had to be correctly matched. John Crane and John Marsh at this time also did not think that the setting of the carrier frequency was important to make the instrument work properly. Dr. Stafford would have gotten his understanding from them. We now know that the proper setting of the carrier frequency was important. The audio frequencies used with the AZ-58 should have been recalculated to match its 4.68 RF carrier frequency because its carrier frequency was different than Aubrey Scoon Beam Ray replica instrument carrier or our original Beam Ray Clinical instrument carrier frequency. When the math is done to check

www.rife.org

the sidebands with the audio frequencies that were used with the AZ-58 it shows that those audio frequencies will not create the correct sideband spacing which will match up to the harmonic Rife Ray #4 frequencies. This is probably why Dr. Robert P. Stafford could not get the AZ-58 to devitalize any of the microorganisms he was testing in the laboratory. Here are two of his statements:

DR. STAFFORD: *“Please excuse my format in the following letter for I intend to ramble a bit and forget strict grammatical dictum. I am writing you at this time partially because John Marsh informs me in a recent letter that you may be somewhat disheartened or at least worried about your role in the experimentations with the Rife Machine. Believe me, Dr. Edward I know how you feel for I too have been through this same feeling with this matter. I have observed clinical results after treatments with this gadget which I can scarcely believe myself. Yet, despite these good results, I have been confused by some rather simple failures such as a recent experiment which I conducted at Good Samaritan Hospital where we used the machine to treat some cultures of Staph Aureus and Strept. Fecalis. In this work we failed to inhibit growth at all or influence the cultures with the Rife Rx. I sent the results to John Marsh and asked for clarification and to be very frank I am not satisfied with John’s excuse of the failure as described by Dr. Rife. I am afraid I’m not a very good apostle for I’m getting some ideas myself on how this thing may work. I really wonder if this ultrasonic kills bacteria and virus at all or does it work like other forms of ultrasonic and merely stimulate the tissue in some unusual manner thereby improving the circulation and secondarily enhancing the body’s defenses against infection…To summarize some of this rambling: I feel that the Rife Ultrasonic Therapy has a very definitely beneficial effect on the human (and canine) body...I furthermore feel that we, as doctors of medicine, using this machine must remain constantly alert to the condition of our patient and vary the Rx as indicated.” (*Dr. Stafford letter to Dr. Jeppson April, 1 1958).

DR. STAFFORD: *“As yet, we have failed to “cure” any case of advanced, terminal malignancy. It appears in several instances that we may have impressed the disease favorably, temporarily. It is difficult to rule out the psychological, morale booster effect to the terminal patient when some definitive effort is made again in his behalf. However, several improvements have appeared to be more physical than emotional...All the patients in the series were treated with the same frequencies (e.g., 728 - 784 - 880 - 2008 - 2128). Perhaps these frequencies may be wrong, or only nearly correct.”* (John Marsh Collection, Dr. Stafford’s Report on using the AZ-58, page 4, www.rife.org)

If the incorrect audio frequencies were used with the 3.10 MHz to 3.30 MHz RF carrier frequency what Dr. Stafford M.D. reported is exactly what would have happened. From Dr. Stafford’s statement we know that he did not think the RF carrier frequency was critical. If Dr. Rife had fully understood how Philip Hoyland had designed the instrument he would have never allowed John Crane or John Marsh to put a variable capacitor in its circuit without recalculating the audio frequencies for the new RF carrier frequency they wanted to use. Also they would not have lowered the original audio frequencies by a factor of 10 times and used them. The AZ-58 would have had a new set of audio frequencies correctly match to the new 4.68 MHz RF carrier frequency it was originally given in 1953. The new set of audio frequencies would then have been the correct audio frequencies that would create the proper sideband spacing. Dr. Rife would have also made sure that Dr. Stafford knew that the carrier frequency was critical. If John Crane and John Marsh really understood how this instrument worked they never would have wanted to change the 4.68 MHz carrier frequency. The 4.68 MHz RF carrier would have been a good carrier frequency if they would have used the correctly matched audio frequencies.

It is obvious they never really understood how the instrument worked or how the audio frequencies interacted with the carrier frequency to produce the M.O.R.s. which would devitalize the organisms. They, like us, were under the false belief that the audio frequencies they were using were the M.O.R.s. They unknowingly promoted this idea because Philip Hoyland would never tell anyone how the instrument worked. We know that Dr. Rife knew that this Beam Ray Clinical machine somehow

www.rife.org

worked on his frequencies with harmonics but it is apparent that no one but Philip Hoyland really understood how the instrument worked on the harmonic sideband method.

With the new variable capacitor that John Crane and John Marsh put into the AZ-58 they could change the RF carrier frequency from about 2.40 MHz to about 4.90 MHz. Dr. Stafford told me personally, when I talked to him, that it did not make any difference which carrier frequency he used, they all worked the same. We will probably never know the full story of how these mistakes were made but it is obvious that no one but Philip Hoyland had the proper understanding of how the instrument worked. Since the AZ-58 used the square wave harmonics method rather than the harmonic sideband method it would have been better if Dr. Rife, John Crane and John Marsh had used a true lower harmonic frequency of the Rife Ray #4 frequencies, but they didn't. Using the correct sub-harmonic audio frequencies derived from the higher Rife Ray #4 RF frequencies would have been a better approach. Those Rife Ray #4 sub-harmonic audio frequencies used with the square wave harmonics method, probably would have worked better than the incorrectly match 1950's audio frequencies they were using in the AZ-58. We should do the same today with our frequencies.

All that we have discovered with the analyzing of this original Rife Ray #5 or Beam Ray Clinical instrument reveals that the audio frequencies used with this equipment are not Dr. Rife’s true M.O.R.s. They are just the frequencies needed to produce the proper sideband spacing in order to hit the correct higher RF harmonics of Dr. Rife’s original M.O.R. frequencies. If the audio frequencies were the frequencies that would devitalize the microorganisms then logically all Dr. Rife’s frequencies would have been audio frequencies, but they are not. If these audio frequencies were Dr. Rife’s true M.O.R. frequencies then an RF carrier frequency would not have be needed and he would have never even needed to build a Ray Tube instrument. Dr. Rife would have been able to do all of his work with a simple frequency generator that would have had a frequency range of only about 25,000 Hertz. But this is not the case. Dr. Rife always said that almost all his frequencies were in the upper bands which match the Rife Ray #4 frequencies which covered a frequency range from 139,200 Hertz to 1,604,000 Hertz. None of those Rife Ray #4 frequencies where audio frequencies.

It is interesting to note that this Rife Ray #5 or Beam Ray Clinical instrument did not have a dedicated fixed audio pulsing circuit. This is also the case with the 1953 AZ-58. Aubrey Scoon mentions a sixty Hertz feedback pulse into the circuit of his instrument but he could not determine if the instrument was intended to work this way or if it was just a malfunction. When we built his instrument we could not get our instrument to work in the same manner as his did in producing this feedback, therefore we believe that it most likely was malfunctioning and was not intended to work in that manner. Since this original Rife Ray #5 or Beam Ray Clinical Rife Machine that we obtained didn’t work in this manner either and none of the other replica instruments work this way we believe that our conclusion is correct.

We know that Dr. Rife’s high RF frequency instruments such as the Rife Ray #3 and Rife Ray #4 used a fixed audio frequency pulsing circuit. This fixed audio frequency modulated the high RF frequency and was used to devitalize the organism. We have shown already in this report that this pulsed waveform was needed to devitalize the various microorganisms that Dr. Rife tested. The Beam Ray Clinical instrument did not use this circuit. The logical reason why it did not us this circuit is the modulated waveform created from the variable audio oscillator must have been sufficient to create the effect. The variable audio frequencies, which created the sidebands, would also pulse the waveform and this must have been sufficient to accomplish the same result. From this we can conclude that all that is needed is a modulated or pulsed waveform with the proper M.O.R. high RF frequency to devitalize microorganisms.

www.rife.org

In the photo, shown below, is the Rife Ray #5 or Beam Ray Clinical instrument waveform. It looks almost like a modulated square wave waveform but it is not produced with a square wave. This waveform is produced using a sine wave audio frequency.

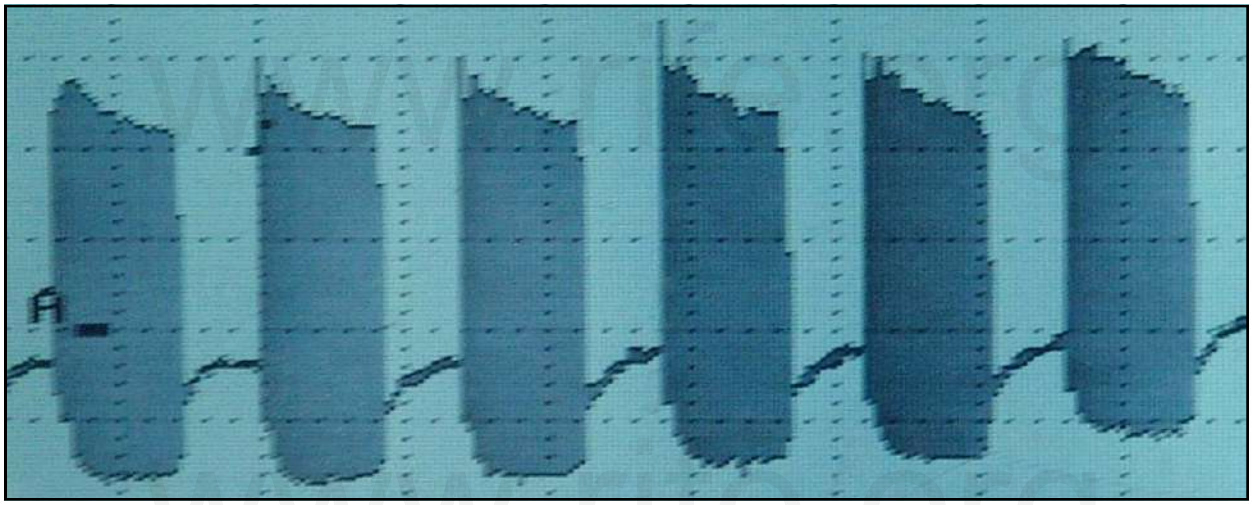

The Beam Ray instrument uses sine wave audio frequencies modulated onto a sine wave RF carrier frequency. It is the unique design of the Beam Ray Clinical circuit that produces this waveform. The M.O.P.A (Master Oscillator Power Amplifier) circuit that Philip Hoyland used was built in a different manner than would normally be used. Instead of using one vacuum tube for the RF oscillator section and a second vacuum tube for the amplifier section he only used one vacuum tube for both sections.

Because he used only one vacuum tube the circuit over oscillates and shuts off for half of the cycle. This shutting off for half of the cycle makes the waveform look almost like a square wave. This new waveform replaced the damped wave waveform that was used in the Rife Ray #4 instrument. In the 1953 AZ-58 Beam Ray Clinical instrument the variable sine wave audio oscillator was replaced with a variable square wave audio oscillator. Had they really understood how the AZ-58 instrument was really supposed to work on the harmonic sideband method they probably would not have made this change. Changing the waveform from sine wave to square wave did not change how the harmonic sidebands worked in the AZ-58. It only shaped the waveform in a true square wave waveform. The only real change that could affect the output of the instrument is not matching the audio frequencies with the new 4.68 MHz RF carrier frequency. It was this change that proved to be the biggest mistake they made.

What must be kept in mind is the treatment frequencies were not the audio frequencies in the original Beam Ray Clinical instrument. It was the sideband frequencies created by the audio frequencies when match and combined with the 3.80 MHz RF carrier frequency that would create the M.O.R. treatment frequencies. We tested both sine wave and square wave with the spectrum analyzer and there was no change in the sidebands. But in the 1950's they were under the false assumption that the audio frequencies were the treatment frequencies.

Variable Audio Frequencies

Now we will discuss the variable audio frequencies used in the Rife Ray #5 or Beam Ray Clinical instrument. When we figured out how this instrument worked on harmonic sidebands we did not choose the optimum audio frequencies. We just used audio frequencies that would show how the instrument worked. There are better audio frequencies that could be used for optimum performance to hit the Rife Ray #4 higher harmonic frequencies. In order to determine the audio frequencies that were used to create the sideband spacing frequencies we first had to determine the difference between the 3.80 MHz RF carrier frequency and Hoyland's new higher harmonic M.O.R.s. which were derived from

www.rife.org

Dr. Rife's original frequencies. In order to make it easier for the reader to understand the difference between the RF carrier frequency and Dr. Rife's higher harmonic frequency will be call that the "difference number" in this report. In most cases we divided, the difference number, by 5 but the BX was divided by 15. The BX M.O.R. higher harmonic frequency, based on the Rife Ray #4 frequency, is 3,208,000 Hertz. This frequency is only the second higher harmonic of the Rife Ray #4 frequency of 1,604,000 Hertz, so it will still work just as well as all the other frequencies.

One must keep in mind that the modern square wave audio frequency harmonics are projected to hit frequencies hundreds of harmonics up. This Beam Ray instrument has far more power in its harmonic sidebands than any square wave harmonic has in it. And the Beam Ray instrument only has to go to the 15th lower sideband to hit the BX M.O.R. frequency, not hundreds of harmonics like a square wave harmonic would. This Philip Hoyland method would be far superior to using low square wave audio frequencies since the harmonics would all be 40 harmonic steps or less.

So that you can better understand how this instrument worked we need to determine what RF carrier frequencies would be best to use with this Beam Ray Clinical instrument. Philip Hoyland would have also done these calculations back in the 1930's. The best RF carrier frequencies to use would always be calculated based on Dr. Rife highest original M.O.R. frequency. As an example the highest frequency that Dr. Rife used in the Rife Ray #4 was the BX cancer virus frequency of 1,604,000 Hertz. The best RF carrier frequency would be this 1,604,000 Hertz frequency multiplied by two which would be 3,208,000. So an RF carrier frequency in the 3.10 to 3.30 Megahertz range would be a good RF carrier frequency range. Dr. Couches' Rife Ray #5 or Beam Ray Clinical instrument, built by Philip Hoyland, used an RF carrier frequency of 3.30 MHz and Aubrey Scoon's instrument which is a copy of it also used a 3.30 MHz carrier frequency. The importance of this 3.30 MHz RF carrier frequency will be discussed later. The next best carrier frequency would then be three times this frequency which would be 4,812,000 Hertz. So a carrier frequency in the AZ-58 4.68 MHz range would also work very well. By using Dr. Rife's highest M.O.R. frequency it is easy to see that the best RF carrier frequencies can be determined. We are sure that Philip Hoyland had this same understanding.

Some of the following information was originally part of Chapter 8 but after restudying the Rife documents we found that this information, including the additional documents we found, actually belonged in this chapter. This information confirms Philip Hoyland's use of higher harmonic frequencies obtained from Dr. Rife's original M.O.R. frequencies.

When Dr. Johnson and Philip Hoyland were testing the first proto-type of the Rife Ray #5 or Beam Ray Clinical instrument in the summer of 1936 they had an interesting effect take place. At that time they were at Dr. Johnson's laboratory testing different bands of frequencies on the various micro-organisms. Dr. Johnson wrote about what happened at his laboratory in a letter he sent to Dr. Gruner and Dr. Rife on November 4, 1936:

DR. JOHNSON: *"Last summer, in hunting for the M.O.R. for the other two reproductive forms of the cryptomyces pleomorphia, we ran into a new band of oscillations which introduced itself to us by killing all three forms - those that we called BX, our filter-passing form; then a transitional form such as you found in the monocytes in the blood; and then the third or highly developed form coming from the sporangius forming from the hyphas of the mycelium. At the same time that this new wave band arrived, we broke all the glass in the laboratory of a certain shape, not only in the room where we were working but in all the other rooms...we had been troubled a great deal with a mold because in the microscope room there were no windows, but this band not only destroyed that mold, which was growing on the leather objects in the room, but every bacteriological culture that we had in the laboratory! It cleaned us out completely so we had to start from scratch and replace our losses. In fact, we were all so surprised that we began to feel each other's pulses to see if we were still alive. As no harm had been done to us, we proceeded to test the new band out on mice, rats, rabbits, guinea pigs and dogs. So far as we were*

www.rife.org

able to discover, it is not at all destructive or injurious to normal cell tissue. While we have been forced to modify our machine so as to produce this new band, still it is so much more effective clinically that we look upon it as a very advantageous discovery. However, our experience has forced us to do all of our experimenting with the new ray [Rife Ray #5 or Beam Ray Clinical instrument] *completely outside of our laboratory building or abandon all form of bacteriological experiments, because it instantly kills them all."* (Letter from Dr. Johnson to Dr. Gruner (copy sent to Dr. Rife) dated, November 4, 1936).

Dr. Johnson clearly points out in this letter that this was done with a new Rife Ray. Dr. Johnson had been using the Rife Ray #4 since the fall of 1935 on his patients but here he points out that they were using a new Rife Ray machine and that it was clinically more effective than the machine he had been using up until this time. In the Beam Ray Trial documents Philip Hoyland stated that he had built a new Rife Ray instrument in the summer of 1936: (Beam Ray Trial Transcript #956)

HOYLAND: *"In the summer of 1936 I designed a new machine, or rather I checked it there at the lab. I had designed it in Pasadena, and we tested it out then."* (Beam Ray Trial Papers #956).

Also in another document called "Development of the Rife Ray" we have supporting evidence that a new Rife Ray was built in 1936:

"In the early part of 1936 Commander Rife and Mr. Hoyland spent much time collaborating on revising some of the applications of the fundamentals of the instruments due to the advancement that had taken place in the application of electronics and it was found that the carrier wave used in the previous instruments could be eliminated. During the summer of 1936 further experiments were carried on, which resulted in an entirely new method of generating the desired frequencies and produced a constant input and output in the instruments."

With the Rife documents showing that Philip Hoyland built the new Rife Ray or Beam Ray Clinical instrument in the summer of 1936 we know that this was the machine that Dr. Johnson was referring to in his letter. In that letter Dr. Johnson clearly states that during the summer of 1936 they had tested this new machine in his laboratory. Dr. Johnson's statement also confirms Philip Hoyland's statement given in the Beam Ray Trial documents that the machine was tested in the laboratory. The documents all support each other and show that it was the Rife Ray #5 or Beam Ray Clinical instrument that they were using which produced this amazing effect. In this same letter Dr. Johnson stated, not once, but twice that this work was done with the new Rife Ray machine. Here is his other statement:

DR. JOHNSON: *"I can assure you that no one, not even myself, could help but be astounded at the results we are now obtaining with the assistance of our new machines and our new band of MOR's."*

This letter was written in November of 1936 a few months after the laboratory testing was completed on this new Rife Ray machine. They now had built more than one of these new machines as pointed out by Dr. Johnson when he said "our new machines." Again Dr. Johnson points out that this is a new style of machine and it has a "new band of MOR's." In the Beam Ray Trial documents Philip Hoyland's complete statement reveals that he was using different frequencies than Dr. Rife:

HOYLAND: *"In the summer of 1936 I designed a new machine, or rather I checked it there at the lab. I had designed it in Pasadena, and we tested it out then and the frequencies were not the same as on Dr. Rife's machine."* (Beam Ray Trial Papers #956).

This statement made by Philip Hoyland about the frequencies not being the same as Dr. Rife's original frequencies is confirmed by Dr. Johnson. So Dr. Johnson's statement in his letter clearly shows that he understands that this new machine is using different M.O.R.s. than what Dr. Rife was originally

www.rife.org

using in the Rife Ray #4. Dr. Johnson also points out that these new M.O.R.s are more effective than the original M.O.R.s found by Dr. Rife. Today we now know that these new M.O.R.s. that Philip Hoyland was using were based on Dr. Rife's original M.O.R.s. They were just higher, more powerful, harmonics of Dr. Rife's original frequencies. From the Beam Ray Trial we know that Philip Hoyland stated that the frequencies he was using in this machine were not the same as the frequencies that Dr. Rife originally found. This however does not change the fact that all Philip Hoyland did was multiply Dr. Rife's original frequencies up in harmonic steps to use them with this new Rife Ray #5 or Beam Ray Clinical instrument. Without Dr. Rife's original frequencies Philip Hoyland would have had nothing. Probably the most important fact is the higher harmonics seemed to work better at devitalizing the organisms. As mentioned before, Dr. Johnson indicates in his letter that this new machine works better than any of the previous machines they had used up until that time (Rife Ray #3 and Rife Ray #4). His letter also shows that he is very impressed with the results they were obtaining with it. Another important fact we learn from this letter is Dr. Johnson now had the M.O.R. for the cryptomyces pleomorphia fungi. He had been trying to get Dr. Rife to find this M.O.R. for many months, without success, because Dr. Rife had developed eye problems and could not use his microscope. But now he had the ability to produce its frequency with the new machine. Another thing that Dr. Johnson pointed out was the fact that they had killed all the bacteriological specimens in his laboratory with the new Rife Ray machine and that they had to use it outside of the laboratory or it would kill all the new replacement specimens they had to obtain.

Dr. Johnson in this same letter to Dr. Gruner said the following about this new Rife Ray machine which Philip Hoyland built:

DR. JOHNSON: *"We are having a great time out here. I have opened a clinic in the Pasadena Home for the Aged and am having excellent luck. Of course, what I am trying to do is to experiment with the effect of the Rife Ray on bacteria in vivo. Our laboratory work has demonstrated pretty conclusively what it will do in vitro. The clinic is held three mornings a week, Tuesday, Thursday and Saturday. Yesterday I had eighteen patients. Among them were two cases of pulmonary tuberculosis, three cases of carcinoma, two cases of old chronic varicose ulcers of the leg, and sundry other cases of more or less definite infectious origins…I certainly wish that you were here to work with me because I am afraid that even you, who know what we are trying to do, will not believe some of the yarns that I would have to tell you as to what is occurring in the clinic without actually seeing them for yourself."* (Letter from Dr. Johnson to Dr. Gruner (copy sent to Dr. Rife) dated, November 4, 1936).

Many people believe that Philip Hoyland's Rife Ray #5 or Beam Ray Clinical machine was somehow inferior to Dr. Rife's previous instruments in effectiveness in devitalizing microorganisms. But Dr. Johnson points out that this instrument was working better than any previous instrument they had used in the past. In another letter to Dr. Meyer, M.D. Dr. Johnson again talks about the effectiveness of this new Beam Ray Clinical machine:

DR. JOHNSON: *"I am conducting a very interesting clinic three times a week with the Rife Ray, treating about twenty two patients a session, and we are obtaining much valuable information from our work. We are encountering many physical problems in the production of the Ray, the greatest difficulty being to obtain uniform results from the apparatus which we have. At times the results of the Ray are absolutely astounding, causing an instantaneous sterilization of the wounds whether interior or exterior. But owing to the slight changes in the temperature of the room and apparatus, the same setting of the dials does not give us the same output, and hence our results are unsatisfactory. However, we believe now that we have developed it to a point where we can very greatly diminish the errors in output and substantially increase the beneficial results or the effect of the apparatus."* (Letter from Dr. Johnson to Dr. Karl F. Meyer dated, December 7, 1936).

www.rife.org

Even with the calibration problems with the instrument Dr. Johnson was still impressed with how well the instrument was working. If Dr. Johnson only would have had the ability that we have today to measure the frequency correctly, rather that relying on dial settings, he would have always had consistent results. With modern technology we can accurately measure a frequency regardless of any changes that the climate may make. Dr. Johnson only had the dial settings to work with back in the 1930's. The limits of the 1930's technology would have given him the problems he talked about in his letter. Even with the limits of the 1930's technology Dr. Johnson stated they had solved many of their problems with the instrument and this helped them to obtain even better results.

During this same clinic which Dr. Johnson was conducting he was also using this new Rife Ray on 30 cataract patients. In a letter that Dr. Johnson wrote to Dr. Joseph Heitger he stated the following:

DR. JOHNSON: *"I closed my clinic on May 28, having been running it for eight months. Our special effort this past winter has been working on cataracts, and while we have treated a number of other infectious conditions (if cataract is an infection), still our principal work has been on the eye. We have had about thirty cases of cataract...Every case that we have treated, with the exception of one which was a traumatic cataract where the lens was absolutely opaque and of recent origin, has been benefited...The application of the Rife Ray as we have used it , does, in the great majority of cases restore the full visual function of the eye; that is, that portion of the visual disturbance due to opacities in the lens. How it does it and why it does it, I do not know, but the above statement is an actual fact, supported by many cases…Oh Lord; How I wish we could get together and go over this work! I believe it will result in epochal changes in the profession's handling of cataract cases."* (Letter from Dr. Johnson to Dr. Joseph D. Heitger dated, June 1, 1937).

In this letter to Dr. Heitger, Dr. Johnson again points out how well this instrument worked. In this letter he also talked about the calibration problems they were having with this new Rife Ray machine. The Beam Ray Clinical machine was known for its calibration problems. Even the Beam Ray Trial testimony and other documents show that this instrument had calibration problems. Eventually Philip Hoyland divided the audio frequency range of the Beam Ray Clinical instrument into four bands. This helped but did not solve the problems completely. By the early 1940's the audio oscillator was replaced by the new stable Hewlett Packard audio oscillator. This Hewlett Packard audio oscillator was far superior to the original RC (Resistor Capacitor) style audio oscillator used by Philip Hoyland. Had they had a modern frequency counter they could have easily solved any of the calibration problems they had. Today if anyone builds this instrument they would want to have the ability to check the RF carrier frequency so that they will have consistent results.

Next we will explain how this instrument had the ability to devitalize so many organisms as described by Dr. Johnson in his letter. He stated in his letter of November 4, 1936 that it was done using a "new band of oscillations." This happened at the time they were testing this new Rife Ray #5 or Beam Ray Clinical instrument at Dr. Johnson's laboratory. Since it was a proto-type machine we do not know what the original frequency range of the audio oscillator was when this happened. What we do know is the original Beam Ray Clinical instrument which we obtained from Dr. Low had four frequency bands which covered a frequency range from 160 Hertz to about 42,000 Hertz. Whether it was one of the lower bands or one of the higher bands we do not know. But this we do know, Dr. Johnson would have had this band put into the new instrument because he stated this:

DR. JOHNSON: "*While we have been forced to modify our machine so as to produce this new band, still it is so much more effective clinically that we look upon it as a very advantageous discovery."* (Letter from Dr. Johnson to Dr. Gruner (copy sent to Dr. Rife) dated, November 4, 1936).

www.rife.org

This statement from Dr. Johnson leaves no question that this band of oscillations was included in the new Rife Ray machine which eventually became the finished product that was sold by Beam Ray Corporation. Why this new machine was able to kill all the organisms that Dr. Johnson talked about is easy to explain since we now know how this instrument worked. If you look at the spectrum analysis graph, shown below, you can see the many sideband frequencies created by a single audio frequency.

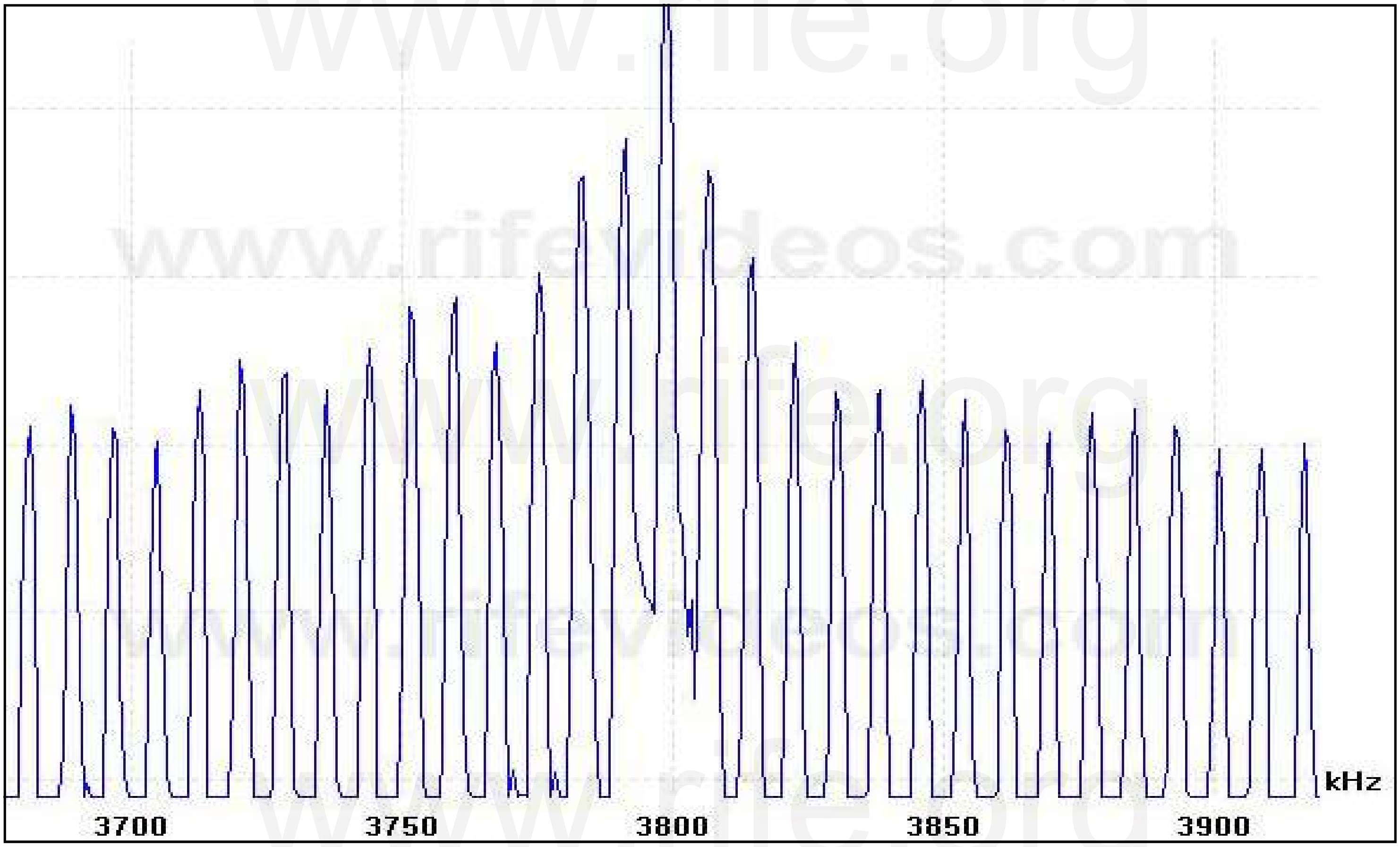

It has been said that a "picture is worth a thousand words." From this spectrum analysis graph of the Beam Ray Clinical machine we know it produced many sideband frequencies simultaneously, sometimes over 100. This explains how this effect happened. Dr. Johnson did not say that it was a single frequency that killed the organisms. He stated that it was a band of frequencies. By the word “band” we know that they must have been turning the frequency dial of the audio oscillator looking for a frequency that would kill the “two reproductive forms of the cryptomyces pleomorphia” organism. These two forms of this disease were found to be a factor in cancers and Dr. Johnson wanted the M.O.R. for these organisms. If the dial of this instrument was turned slowly, as they would have done, then they would have be sweeping the frequency. Even though you are only sweeping a single audio frequency you are also sweeping as many as 100 harmonic sideband frequencies simultaneously which were created by a single audio frequency. For an example we will explain how this works. It must be kept in mind that if you sweep an audio frequency in the Beam Ray Clinical instrument from 6000 Hertz down to 500 Hertz all of the RF sidebands produced from the RF carrier frequency will start with a distance of 6000 Hertz apart and as you sweep down to 500 Hertz these sidebands will slowly close until they are only 500 Hertz apart. During this sweep every frequency between these sidebands will be hit because these sidebands are also all going in towards the RF carrier frequency.

Not only are the sidebands closing and getting closer together but they are also moving in towards the RF carrier frequency completely overlapping each other. Both the upper sidebands and the lower sidebands produced by the audio frequency close in towards the RF carrier frequency. This produces a large broadband sweep, with these many sideband frequencies, that can cover up to a one Megahertz or one million Hertz frequency range. This instrument would be very advantageous to use

www.rife.org
www.rifevideos.com

on patients but in a laboratory setting it would be your worst nightmare because it could very easily kill all your specimens as described by Dr. Johnson:

DR. JOHNSON: "*experience has forced us to do all of our experimenting with the new ray completely outside of our laboratory building or abandon all form of bacteriological experiments, because it instantly kills them all.*"

We know that Philip Hoyland used at least two different RF carrier frequencies, 3.80 MHz and 3.30 MHz, with his Beam Ray Clinical instruments. We also know from the Rife documents that Dr. Couche's Beam Ray Clinical instrument RF carrier frequency was set to 3.30 MHz. The audio frequencies went from 1200 Hertz for Tetanus to 21,275 Hertz for the BX Cancer virus. These audio frequencies produced the correct sidebands which hit the correct frequency for each organism. It is easy to understand that if all of Dr. Rife's frequencies for the various organisms are covered over this range from 1200 Hertz to 21,275 Hertz then a slow sweep over this range would cover all the organisms' M.O.R.s. or frequencies. It is also logical to conclude that if any organism's frequency is located between about 40,000 Hertz up to about 1.80 MHz it would also be hit at some point during this sweep. This is because any organism's frequency could be multiplied up in harmonics so that it would be as close to 3.30 MHz as you can get. With this understanding it would also be logical to assume that a sweep through that complete frequency range down to about 500 Hertz would cover all Dr. Rife's M.O.R. frequencies both known and unknown. During the Beam Ray Trial it was pointed out that Dr. Rife had found the M.O.R.s. for about 40 more organisms that were not included in the Beam Ray Clinical Machine. Here is that statement:

HUTCHINSON: *"There are about 40 frequencies discovered by Dr. Rife that have not yet been released to the public, and have not been included in the machine."* (Beam Ray Trial Papers #1689)

Those 40 frequencies could have been put into the Beam Ray Clinical machine had Dr. Rife understood how Philip Hoyland built the machine using the sideband method. Though they were never put into the instrument using a dial setting this does not change the fact that the instrument could output those frequencies. Since they probably all would have been within this same sweep range then it is reasonable to assume that this same sweep range would also hit all of these organisms' frequencies. This is why we said *"All Rife's M.O.R.s. both known and unknown."*

If you wanted to cover an even larger frequency range you could sweep from 40,000 Hertz down to about 100 Hertz. The mathematical calculations show that to hit each organism's M.O.R. with sufficient time a sweep should not cover more than 25,000 Hertz in 4 hours. One important fact that should be realized is when this sweep is done from 25,000 Hertz down to about 500 Hertz, or lower, each organism's frequency is hit more than 20 times. To understand how this is possible requires the understanding that when these sidebands close and move in toward the carrier frequency each organism's frequency is hit with multiple sidebands. To explain how this happens we will take the BX Cancer virus frequency of 21,275 Hertz as an example. It is the fourth lower sideband that first hits the BX Cancer virus frequency of 3,214,900 Hertz (1,607,450 X 2 = 3,214,900) when it hits 21,275 Hertz. The fifth lower sideband again hits 3,214,900 Hertz at 17,020 Hertz. The sixth hits it at 14,183 Hertz and the seventh at 12,157 Hertz. This continues on until each organism is hit by more than 20 sidebands.

What has just been explained is why Dr. Johnson had the experience of this instrument killing all the organisms in his laboratory. Just a sweep through one of these four frequency bands with the Beam Ray Clinical instrument could wipe out a great many organisms because the sidebands would hit many organisms' M.O.R.s one right after another. This is because there can be nearly 100 sidebands generated at the same time.

www.rife.org

The reason we did not make a schematic for this original Beam Ray Clinical instrument is due to the fact that Aubrey Scoon's Rife Ray #5 or Beam Ray Clinical replica schematic is available and will build the same type of instrument. Aubrey Scoon’s instrument carrier frequency has been discovered even though his instrument was malfunctioning or had parasidic oscillations. He originally gave a 3.33 MHz carrier frequency using an 812a tube. Then he made the wrong assumption and used an 809 tube and replaced the 812a with the 809, but this did not eliminate the parasitic oscillations. The problem with the instrument malfunctioning is what made it hard to determine the true carrier frequency. Without having the correct RF carrier frequency it makes it very hard to verify the audio frequencies it used. Having done a lot of work with this style of instrument and knowing how harmonics can shift the carrier frequency we decided to test a 3.30 MHz carrier to see if it was the correct frequency it used. This 3.30 MHz carrier frequency would have been a more logical choice to use especially when it comes to making the correct sideband frequencies. The results of that decision will be shown later when we take a look at Aubrey Scoon’s 1940's (actually 1950's) Beam Ray Clinical replica instrument built by Verne Thompson.

On the next page is a chart showing the “Rife Ray #4 Frequencies” with their “Higher Harmonic Frequencies” along with the “Carrier Difference Frequencies” and the “1/10 tenth Of One Meter” mentioned by Dr. Rife, converted to Hertz. Also included in this chart are “Beam Ray Clinical Frequencies” which are audio frequencies that would work with the Beam Ray Clinical instrument. It also includes “Aubrey Scoon’s Frequencies” for a comparison even though we have not yet examined Aubrey’s Beam Ray replica machine. After you have read about Aubrey Scoon’s instrument this information will be even more understandable.

You will notice in the chart on the next page that many of the Beam Ray Clinical instrument audio frequencies, when divided with some numbers are very close to the audio frequencies used by Aubrey Scoon’s Beam Ray instrument. It is easy to see the correlation between these two machines and the sideband method used by Philip Hoyland to hit the proper high frequency M.O.R. harmonics of the Rife Ray #4 frequencies. There could be many more audio frequency sets depending how you divide the “*Carrier Difference Frequencies.” The best audio frequency set would be a list with the highest frequencies so that you have the lowest number of sideband harmonics. The reason for this is power loss: the more sidebands the less power in each sideband. The power loss after about 40 sidebands is so great, when looking on a spectrum analyzer, it is doubtful the sideband frequencies would work.

As mentioned before the higher the audio frequency used the more power there is in the sidebands. For this reason we have put in the chart, shown below, a list of the optimum Beam Ray Clinical instrument audio frequencies to be used with a 3.80 MHz RF carrier frequency and a 40,000 Hertz audio oscillator. We did this since this original Beam Ray Clinical instruments oscillator went to just over 40,000 Hertz. This would give the lowest number of sideband harmonics and the greatest power in each sideband.

Beam Ray Clinical Instrument Sideband Sine Wave Audio Frequencies			
Anthrax	20, 800 Hz	Streptococcus	33,333 Hz
B or E Coli Rod	23,500 Hz	Steptothrix	40,000 Hz
B or E Coli Virus	25,000 Hz	Syphilis or Treponema	36,250 Hz
BX Virus Carcinoma	39,467 Hz	Tetanus	28,000 Hz
BY Sarcoma	?37,000 Hz	Tuberculosis Rod	36,667 Hz
Gonorrhea	36,000 Hz	Tuberculosis Virus	?22,500 Hz
Spinal Meningitis	21,500 Hz	Typhoid Rod	100 Hz
Staphylococcus	24,000 Hz	Typhoid Virus	38,214 Hz

www.rife.org

The Beam Ray Clinical Instrument Sideband Sine Wave Audio Frequencies Based On A 3.80 MHz Carrier

Microorganism	Rife Ray #4 Frequencies	Higher Harmonic Frequencies	*Carrier Difference Frequencies	1/10 Of One Meter Freq.	Number of Sideband Harmonics	Aubrey Scoon's Frequencies	Beam Ray Clinical Frequencies
Actinomycosis or Streptothrix	192,000 Hz	3,840,000 or 20th	40,000 Hz	12 Hz	4	7,870 Hz	10,000 Hz
Anthrax	139,200 Hz	3,758,400 or 27th	41,600 Hz	6 Hz	5		8,320 Hz
B or E Coli Rod	417,000 Hz	3,753,000 or 9th	47,000 Hz	58 Hz	6	8,020 Hz	7,833 Hz
B or E Coli Virus	770,000 Hz	3,850,000 or 5th	50,000 Hz	198 Hz	3	17,220 Hz	16,667 Hz
BX Virus Carcinoma	1,604,000 Hz	3,208,000 or 2nd	592,000 Hz	858 Hz	15	21,275 Hz	39,467 Hz
BY Sarcoma	?1,530,000 Hz	3,060,000 or 2nd	740,000 Hz	780 Hz	20	20,080 Hz	37,000 Hz
Gonorrhea	233,000 Hz	3,728,000 or 16th	72,000 Hz	18 Hz	5		14,400 Hz
Pneumonia or Spinal Meningitis	427,000 Hz	3,843,000 or 9th	43,000 Hz	61 Hz	5	7,660 Hz	8,600 Hz
Staphylococcus Pyogenes Aureus	478,000 Hz	3,824,000 or 8th	24,000 Hz	76 Hz	3	7,270 Hz	8,000 Hz
Streptococcus Pyogenes	720,000 Hz	3,600,000 or 5th	200,000 Hz	173 Hz	24	8,450 Hz	8,333 Hz
Syphilis	789,000 Hz	3,945,000 or 5th	145,000 Hz	207 Hz	22	6,600 Hz	6,591 Hz
Tetanus	234,000 Hz	3,744,000 or 16th	56,000 Hz	18 Hz	5	1,200 Hz	11,200 Hz
Tuberculosis Rod	369,000 Hz	3,690,000 or 10th	110,000 Hz	45 Hz	13	8,300 Hz	8,462 Hz
Tuberculosis Virus	?769,000 Hz	3,845,000 or 5th	45,000 Hz	197 Hz	3	16,000 Hz	15,000 Hz
Typhoid Rod	760,000 Hz	3,800,000 or 5th	0 Hz	192 Hz	1	6,900 Hz	100 Hz
Typhoid Virus	1,445,000 Hz	4,335,000 or 3rd	535,000 Hz	694 Hz	14	18,620 Hz	38,214 Hz
Worms						2,400 Hz	

On page 93 you will find oscilloscope images of the original Beam Ray Clinical instrument. On pages 94 through 100 you will find the spectrum analyzer graphs showing the sideband frequencies for each organism using the Beam Ray Clinical instrument 3.80 MHz RF carrier frequency. These graphs show which sideband frequency is hitting the Harmonic Rife Ray #4 frequency that is closest to the 3.80 MHz carrier frequency. These graphs show clearly how this instrument worked.

On pages 101 and 102 are two charts that show the Rife Ray #4 frequencies and their M.O.R. harmonics so you can see the progression of harmonic frequencies as they go up in harmonics to the frequencies that were used by Philip Hoyland in this Beam Ray Clinical instrument. With these two charts you can do the same math that we did to determine the audio frequencies that will produce the proper sideband spacing. You will notice in the chart above called "The Beam Rays Clinical Instrument Sideband Sine Wave Audio Frequencies Based On A 3.80 MHz Carrier" that our list of audio frequencies for the Beam Ray Clinical instrument are in all four bands of frequencies. Philip Hoyland said:

HOYLAND: *"The whole list of bacteria that the machine was treating was divided into four bands."* (Beam Ray Trial Transcript #260)

You will notice that Aubrey Scoon's Beam Ray replica frequencies are only in three bands (bands 2, 3 and 4). This is because Dr. Rife's engineer, Verne Thompson, used a different audio oscillator with this instrument than the audio oscillator used in our original instrument. The first band which covered from 20 Hertz to 200 Hertz is not used since the lowest audio frequency was 1200 Hertz when

www.rife.org

using a 3.30 Megahertz carrier frequency. In the original Beam Ray Clinical instrument we obtained the audio oscillator was also changed but the RF carrier frequency was 3.80 Megahertz. With this carrier frequency you will get one audio frequency that has to be in band one because it is so close to the carrier frequency that only a frequency of less than two hundred Hertz will work.

IMPORTANT INFORMATION: Many people have asked why the frequencies that were used to determine how this Beam Ray Clinical instrument worked are different than those used with Aubrey Scoon’s Beam Ray replica instrument. The reason they are slightly different is this instrument was evaluated and analyzed first. It also used an RF carrier frequency which is different than Aubrey Scoon's instrument. It was the evaluating and analyzing of this instrument using the Rife Ray #4 frequencies that made it possible figure out Aubrey Scoon's Beam Ray Clinical replica instrument. It was the information that was obtained from this Original Beam Ray Clinical instrument that made it possible to obtain the most accurate frequencies from Aubrey Scoon's Beam Ray Clinical instrument. In the Beam Ray Clinical instrument the frequencies had to be calculated precisely or it would not work. The reason for this precision is, if you happen to be off 20 Hertz on the first sideband then the second sideband will be off 40 Hertz. The error will continue to compounded with every sideband. It is easy to see that this simple small error would compound and make it so the instrument would not work. It is easy to see that though the instruments frequencies would wander a little the math to determine where the sidebands needed to be had to be exact. It was Philip Hoyland’s precise math that made it so we would have the most accurate frequencies for the various microorganisms. Philip Hoyland said, on the stand, in the 1939 Beam Ray Trial that he tested the instrument in Rife's lab. His testimony revealed that he had tested the Beam Ray Clinical design using microorganisms to determine its effectiveness. Dr. Rife also tested this same instrument in his lab on micro-organisms in the early 1940's. So we know that this design worked.

Please read Chapter 11 of this report to understand how we obtained the most accurate frequencies from Aubrey Scoon's Beam Ray Clinical replica instrument.

Chapter Summary: Dr. Rife's engineer, Philip Hoyland, not John Crane, was the first person to built the instruments that used audio frequencies and an RF carrier frequency. This instrument which was called the Rife Ray #5 or Beam Ray Clinical instrument was fundamentally different than the Rife Ray #3 and Rife Ray #4 instruments because it was built with a variable audio oscillator and a fixed RF Carrier frequency. It was the modulation of the audio frequency with the RF carrier frequency which produced Dr. Rife's higher harmonic M.O.R. frequencies which devitalized the microorganisms. This new method which Philip Hoyland developed used harmonic sideband frequencies produced by the audio frequencies to produce Dr. Rife's higher harmonic frequencies. Dr. Rife’s higher harmonic frequencies were the frequencies used in both the Rife Ray #3 and Rife Ray #4 instruments.

Philip Hoyland built this Beam Ray Clinical instrument using this new method because Dr. Rife was not able to patent his equipment or his frequencies. The audio frequencies were not the frequencies that devitalized any microorganisms as many believe today. The audio frequencies had to be combined with the correct RF carrier frequency to produce the correct sideband frequencies or they would not produce Dr. Rife's frequencies which would devitalize the various microorganisms. Because Philip Hoyland hid how the Beam Ray Clinical instrument worked it resulted in John Crane and John Marsh believing that the audio frequencies were the treatment frequencies for the microorganisms. They in turn passed on this incorrect understanding of the audio frequencies which has lead to the current belief that the audio frequencies are the treatment frequencies. It was the analyzing of the original Beam Ray Clinical instrument obtained from Dr. Low which revealed how the audio frequencies produce Dr. Rife's original higher harmonic frequencies through harmonic sidebands.

In Chapter 10 we will look at the next Rife Machine built by Philip Hoyland which was call the Beam Ray Laboratory instrument.

www.rife.org

Oscilloscope Readings Of The Original Beam Ray Clinical Instrument

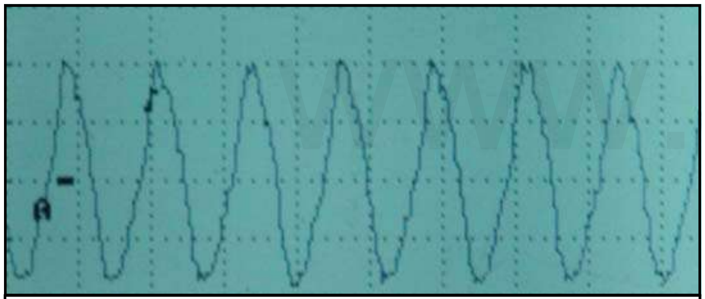

3.80 MHz carrier frequency before ray tube.

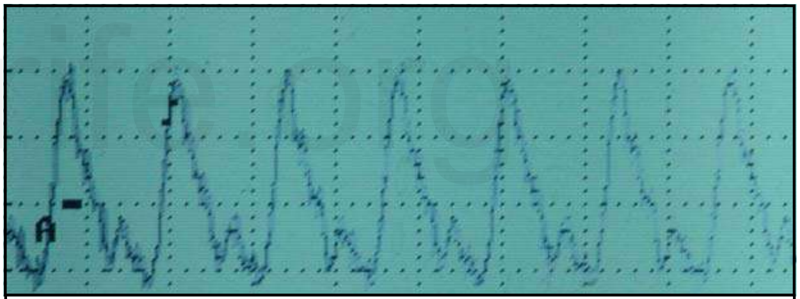

3.80 MHz carrier frequency after ray tube.

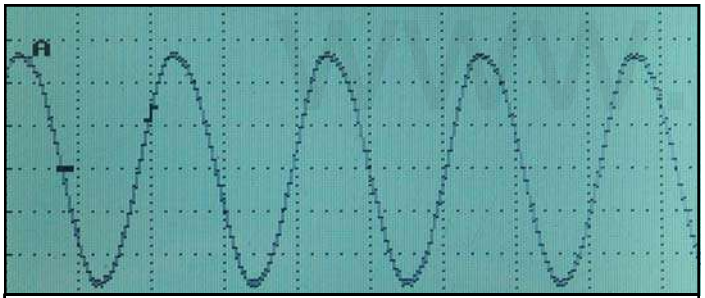

2,377 Hertz frequency from audio oscillator.

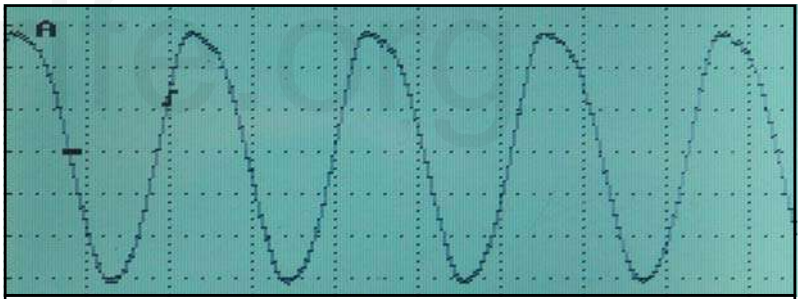

9,360 Hertz frequency from audio oscillator.

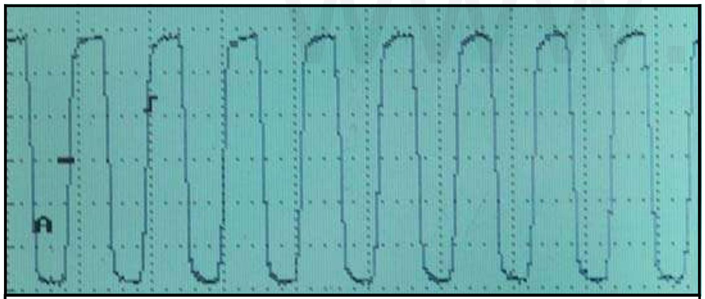

18,630 Hertz frequency from audio oscillator.

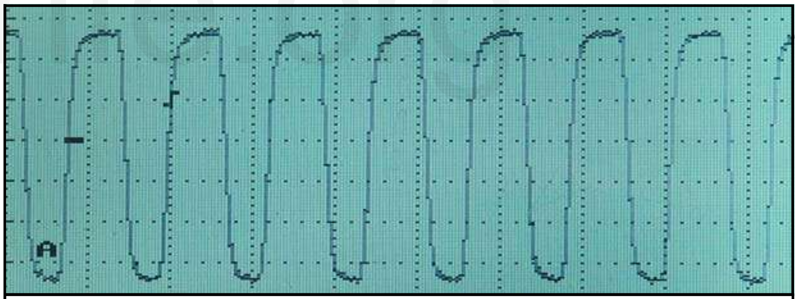

40,980 Hertz frequency from audio oscillator.

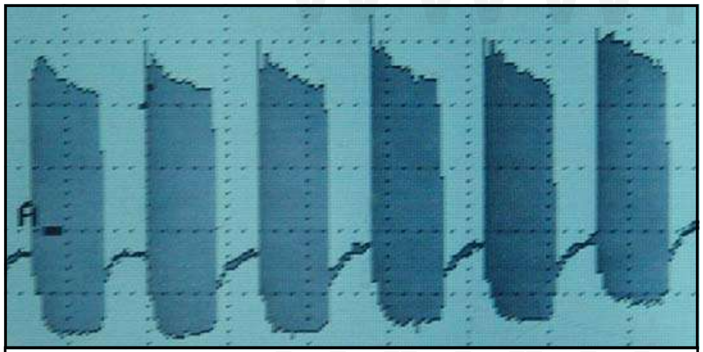

Original Beam Ray modulated waveform.

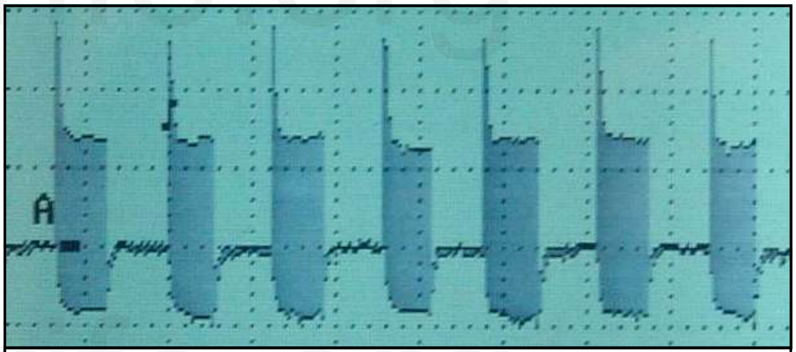

Aubrey Scoon’s Beam Ray replica modulated waveform.

www.rife.org

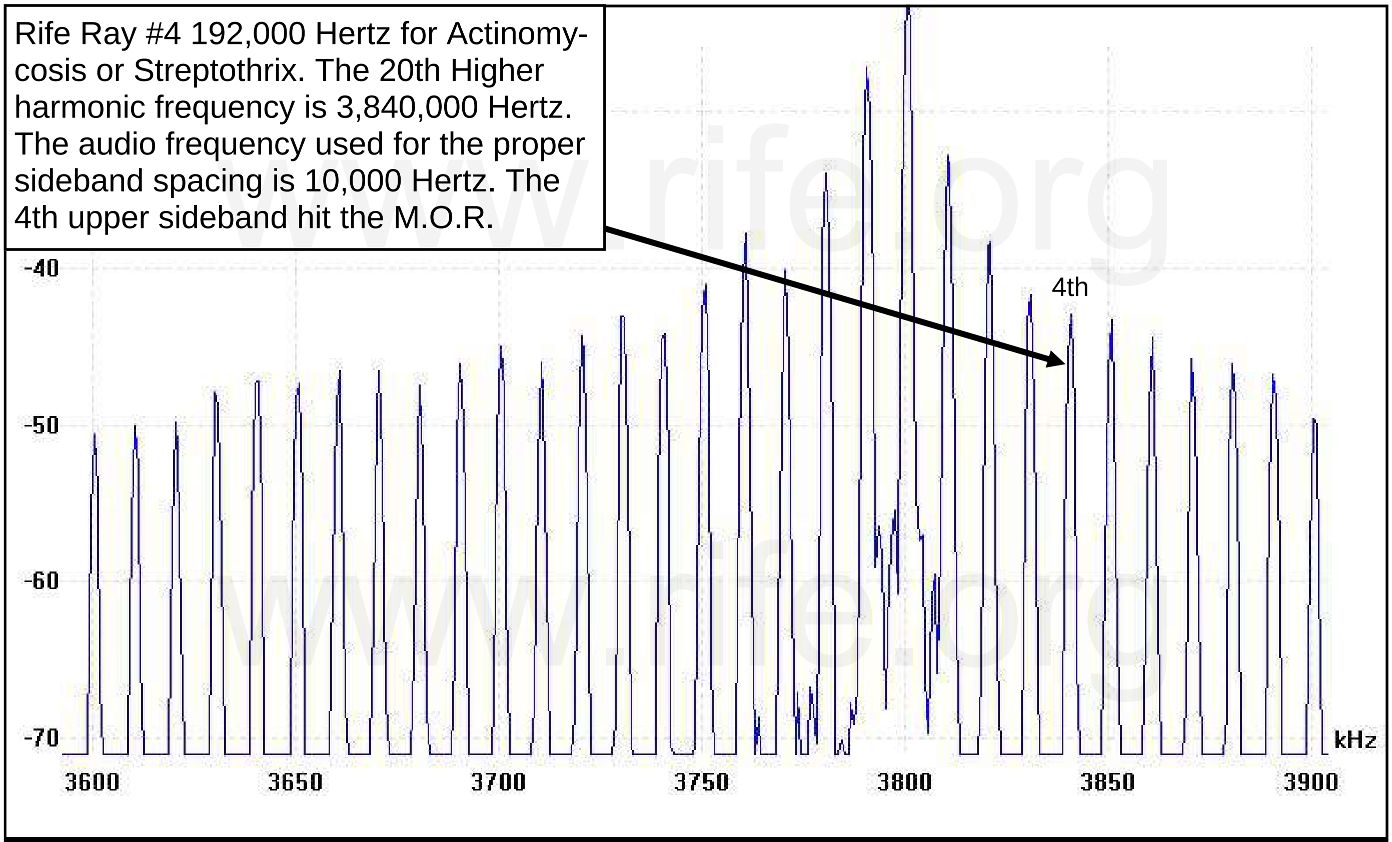

Rife Ray #4 192,000 Hertz for Actinomycosis or Streptothrix. The 20th Higher harmonic frequency is 3,840,000 Hertz. The audio frequency used for the proper sideband spacing is 10,000 Hertz. The 4th upper sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

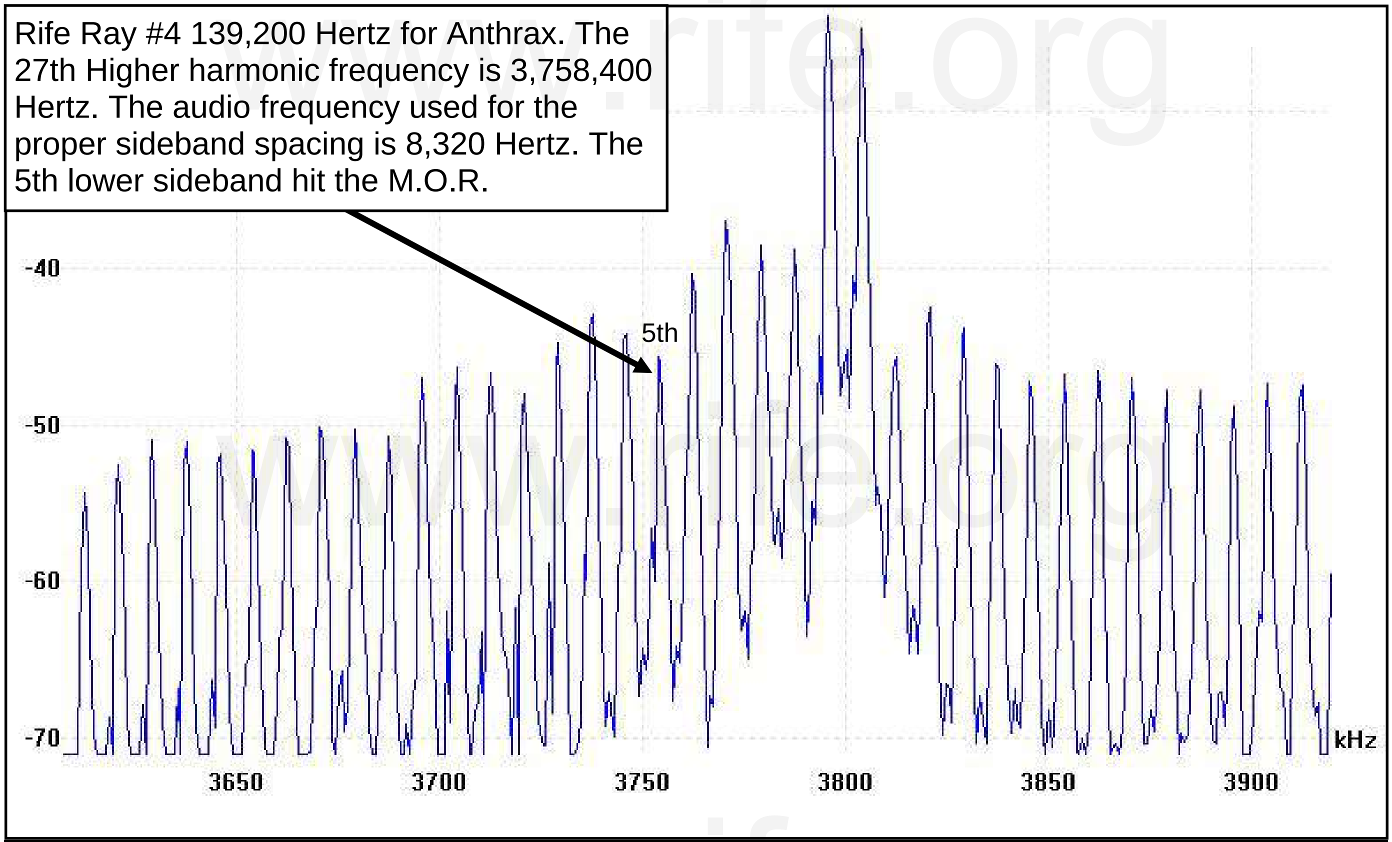

Rife Ray #4 139,200 Hertz for Anthrax. The 27th Higher harmonic frequency is 3,758,400 Hertz. The audio frequency used for the proper sideband spacing is 8,320 Hertz. The 5th lower sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

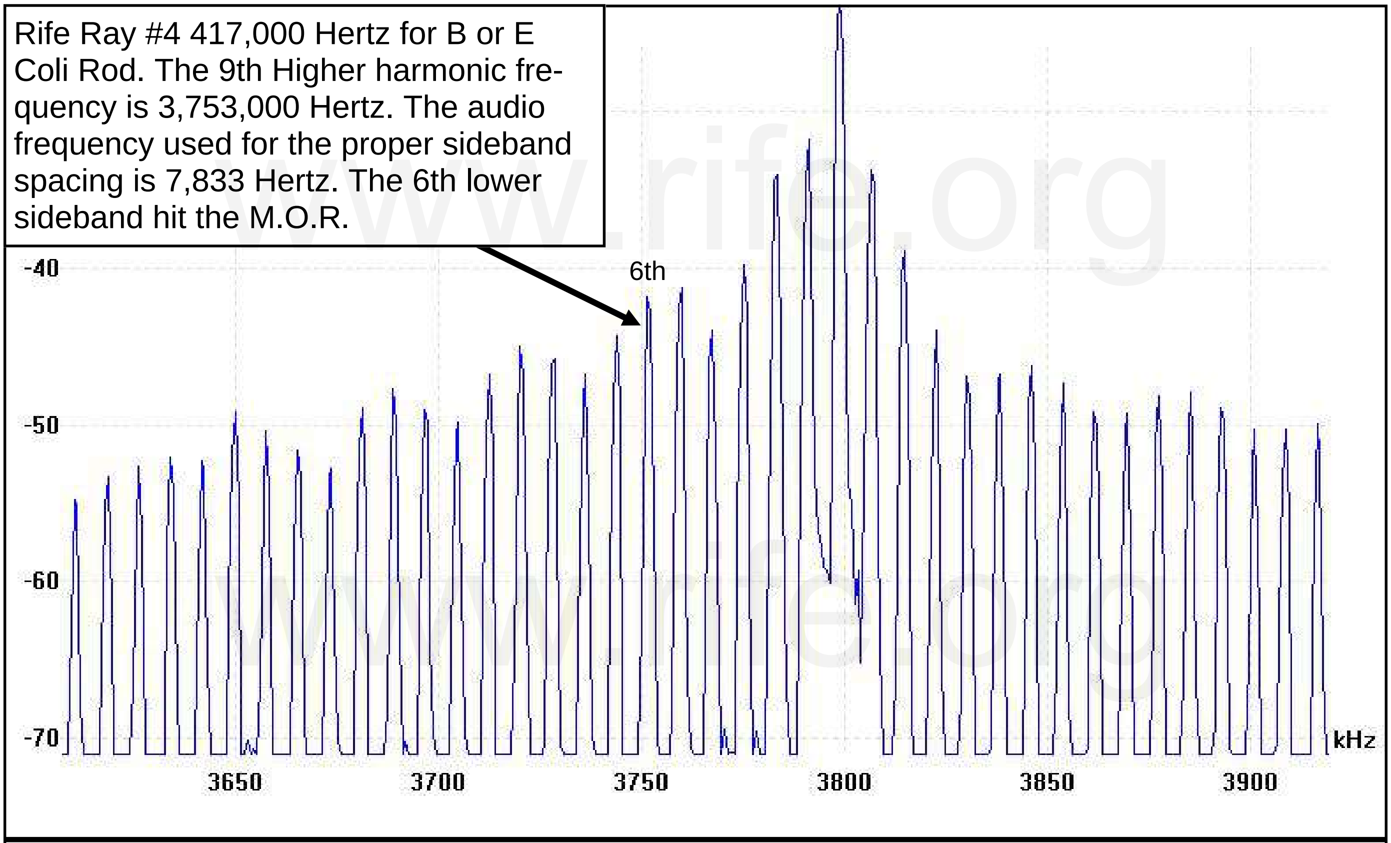
Rife Ray #4 417,000 Hertz for B or E Coli Rod. The 9th Higher harmonic frequency is 3,753,000 Hertz. The audio frequency used for the proper sideband spacing is 7,833 Hertz. The 6th lower sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

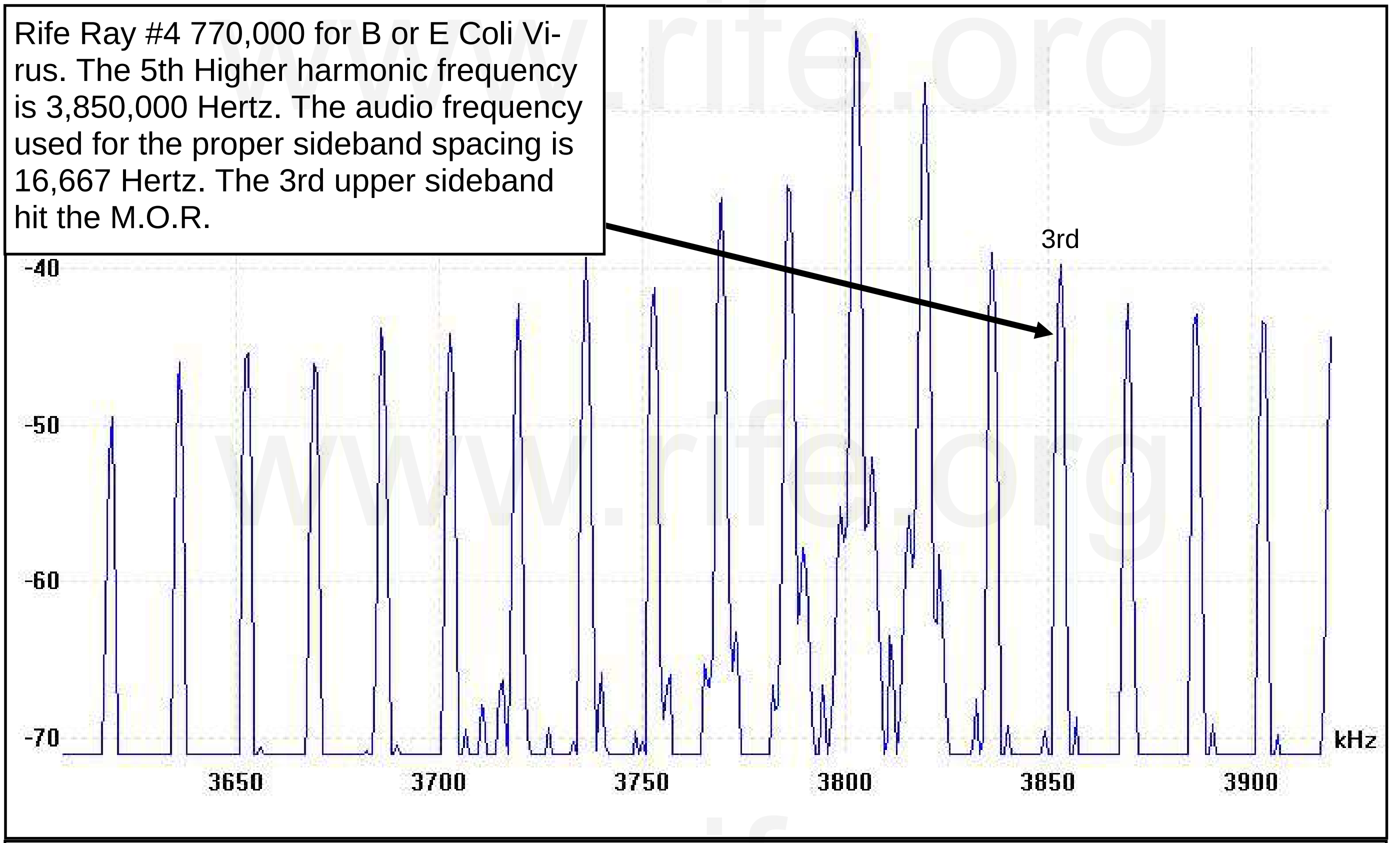
Rife Ray #4 770,000 for B or E Coli Virus. The 5th Higher harmonic frequency is 3,850,000 Hertz. The audio frequency used for the proper sideband spacing is 16,667 Hertz. The 3rd upper sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

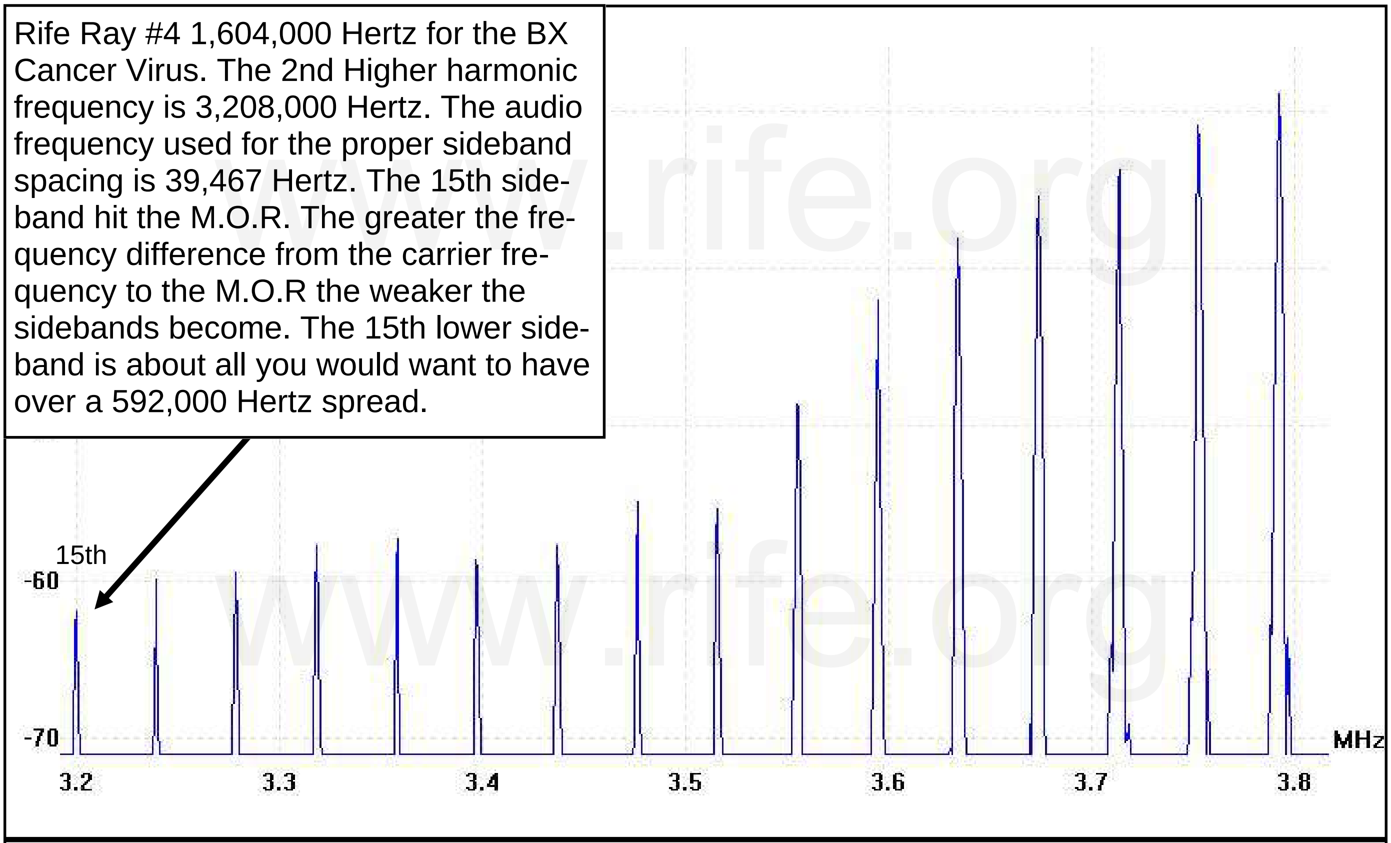
Rife Ray #4 1,604,000 Hertz for the BX Cancer Virus. The 2nd Higher harmonic frequency is 3,208,000 Hertz. The audio frequency used for the proper sideband spacing is 39,467 Hertz. The 15th sideband hit the M.O.R. The greater the frequency difference from the carrier frequency to the M.O.R the weaker the sidebands become. The 15th lower sideband is about all you would want to have over a 592,000 Hertz spread.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

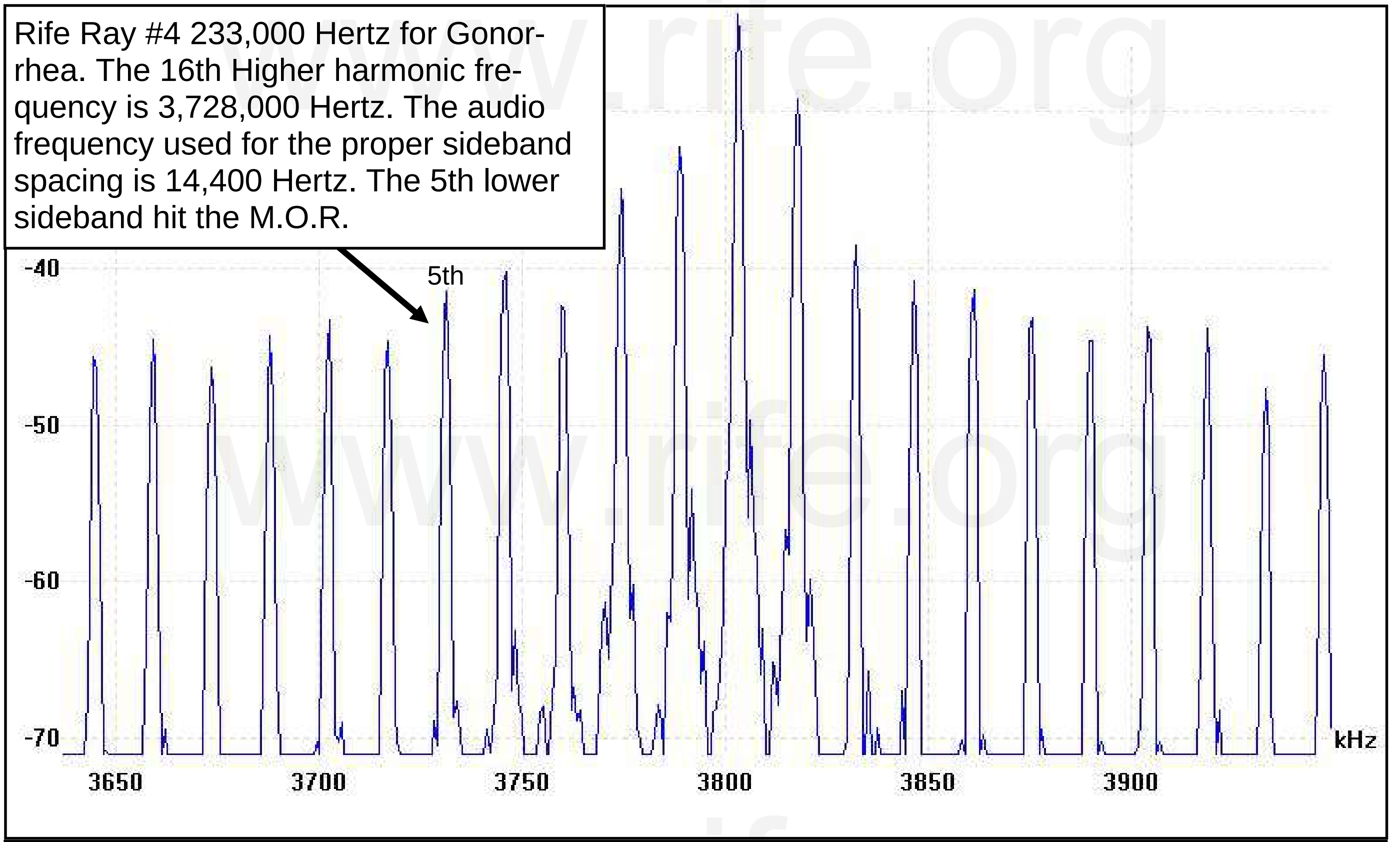
Rife Ray #4 233,000 Hertz for Gonorrhea. The 16th Higher harmonic frequency is 3,728,000 Hertz. The audio frequency used for the proper sideband spacing is 14,400 Hertz. The 5th lower sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

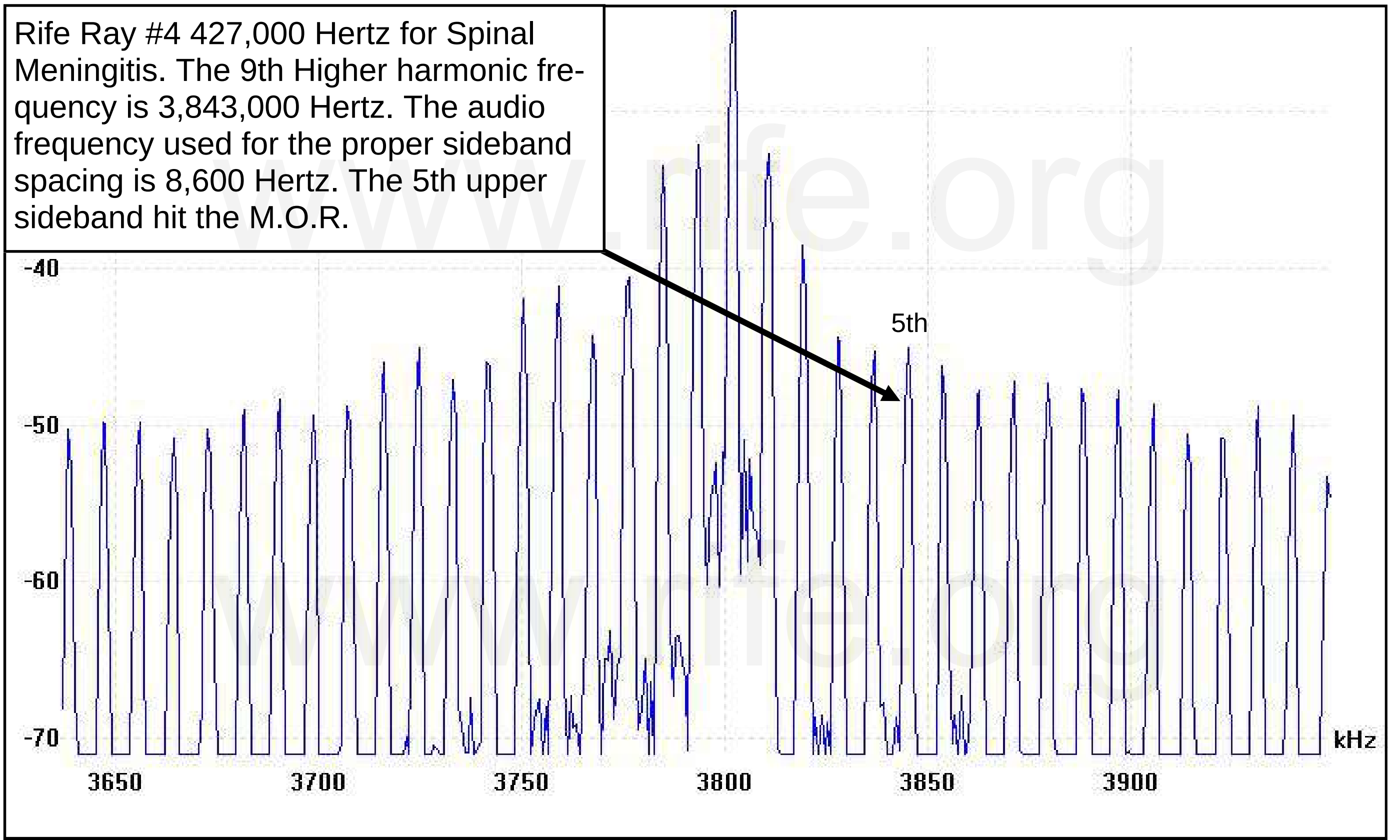

Rife Ray #4 427,000 Hertz for Spinal Meningitis. The 9th Higher harmonic frequency is 3,843,000 Hertz. The audio frequency used for the proper sideband spacing is 8,600 Hertz. The 5th upper sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

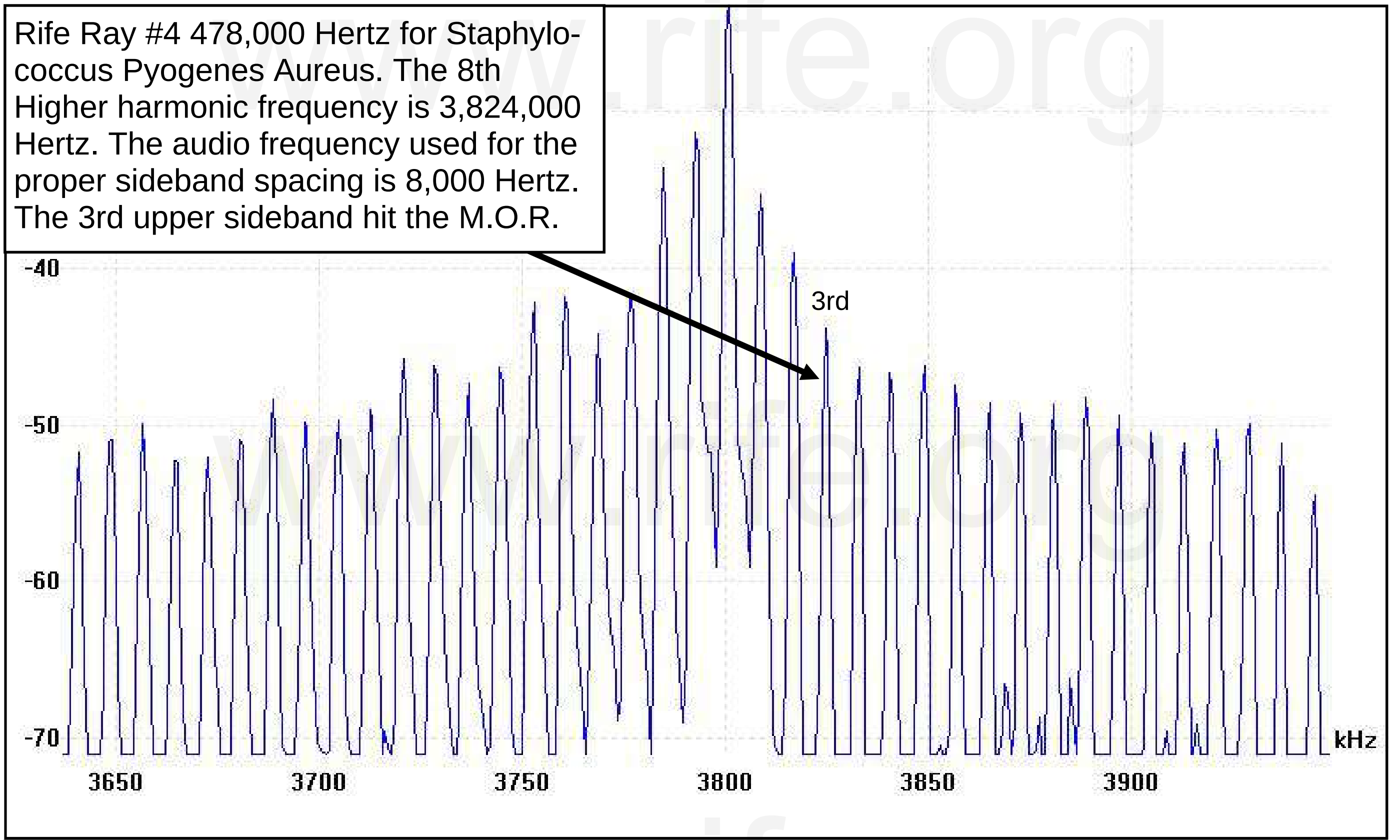

Rife Ray #4 478,000 Hertz for Staphylococcus Pyogenes Aureus. The 8th Higher harmonic frequency is 3,824,000 Hertz. The audio frequency used for the proper sideband spacing is 8,000 Hertz. The 3rd upper sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

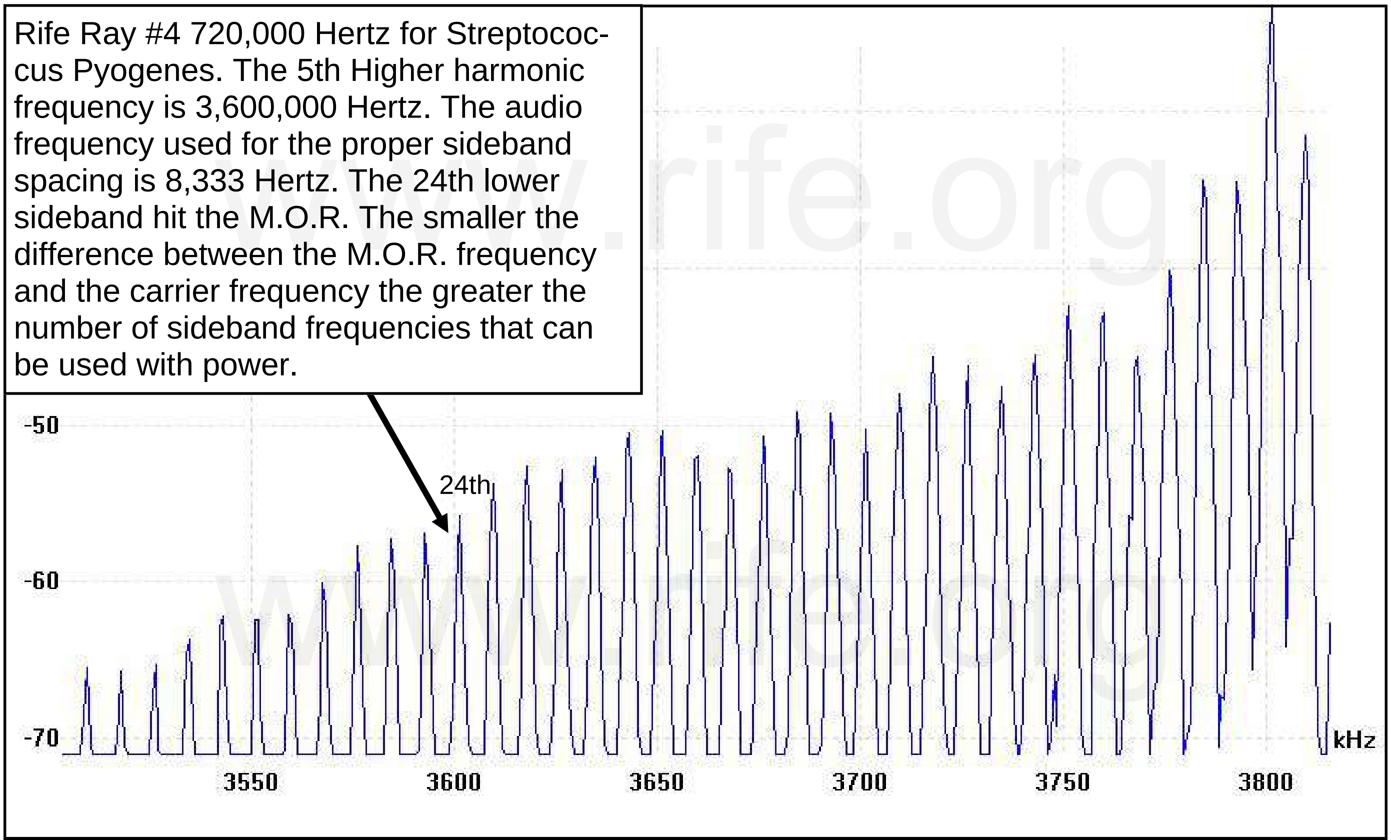

Rife Ray #4 720,000 Hertz for Streptococcus Pyogenes. The 5th Higher harmonic frequency is 3,600,000 Hertz. The audio frequency used for the proper sideband spacing is 8,333 Hertz. The 24th lower sideband hit the M.O.R. The smaller the difference between the M.O.R. frequency and the carrier frequency the greater the number of sideband frequencies that can be used with power.

Philip Hoyland's Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

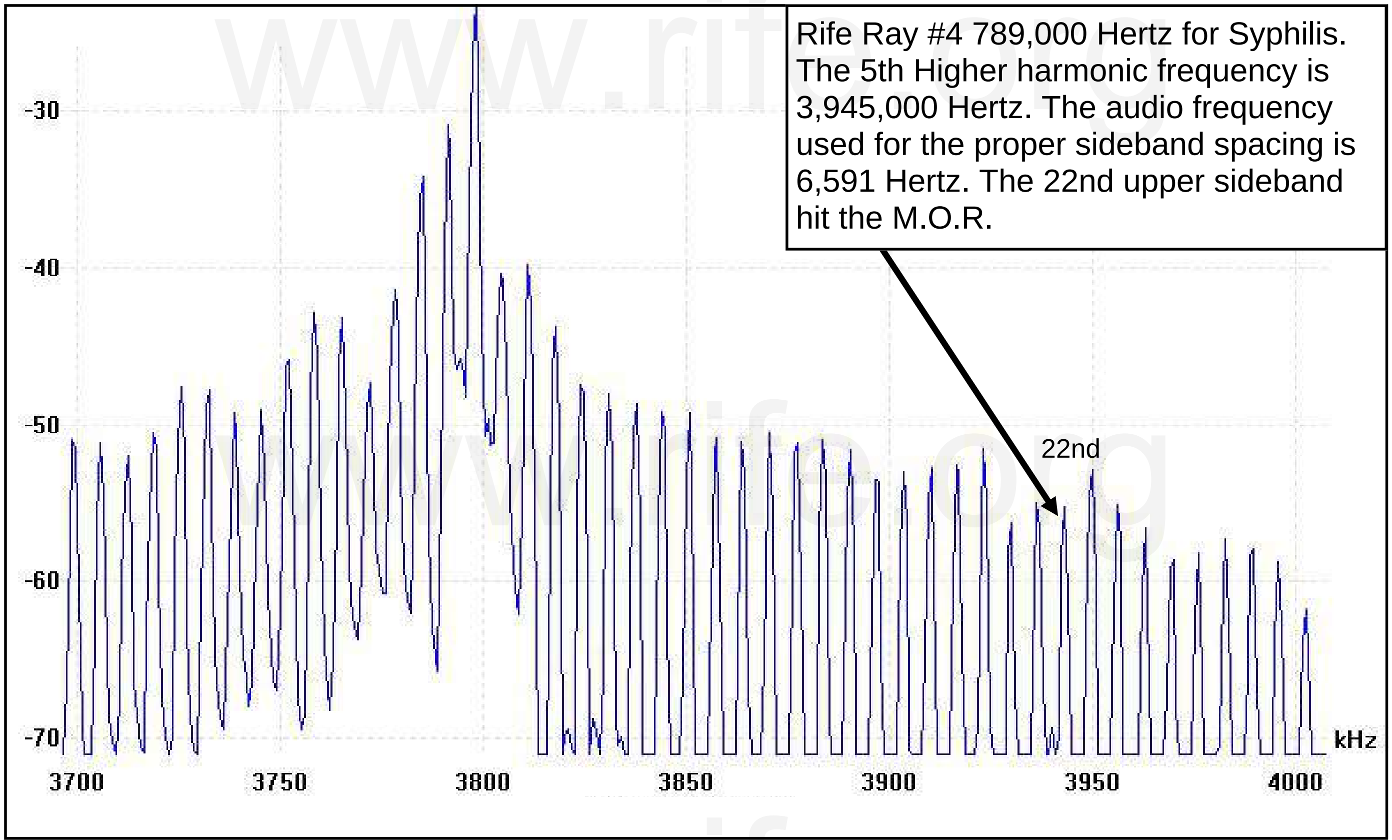

Rife Ray #4 789,000 Hertz for Syphilis. The 5th Higher harmonic frequency is 3,945,000 Hertz. The audio frequency used for the proper sideband spacing is 6,591 Hertz. The 22nd upper sideband hit the M.O.R.

Philip Hoyland's Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

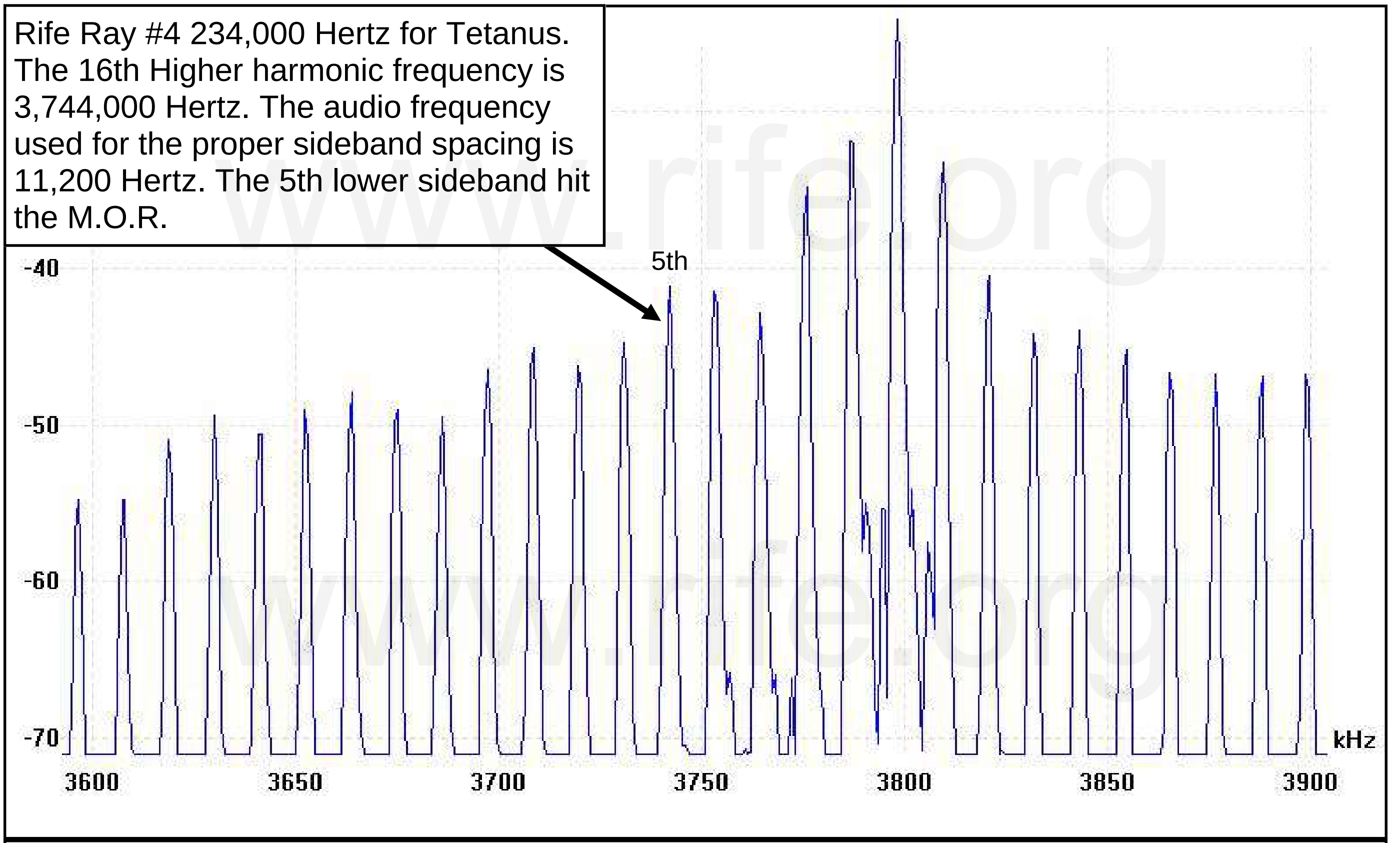

Rife Ray #4 234,000 Hertz for Tetanus. The 16th Higher harmonic frequency is 3,744,000 Hertz. The audio frequency used for the proper sideband spacing is 11,200 Hertz. The 5th lower sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

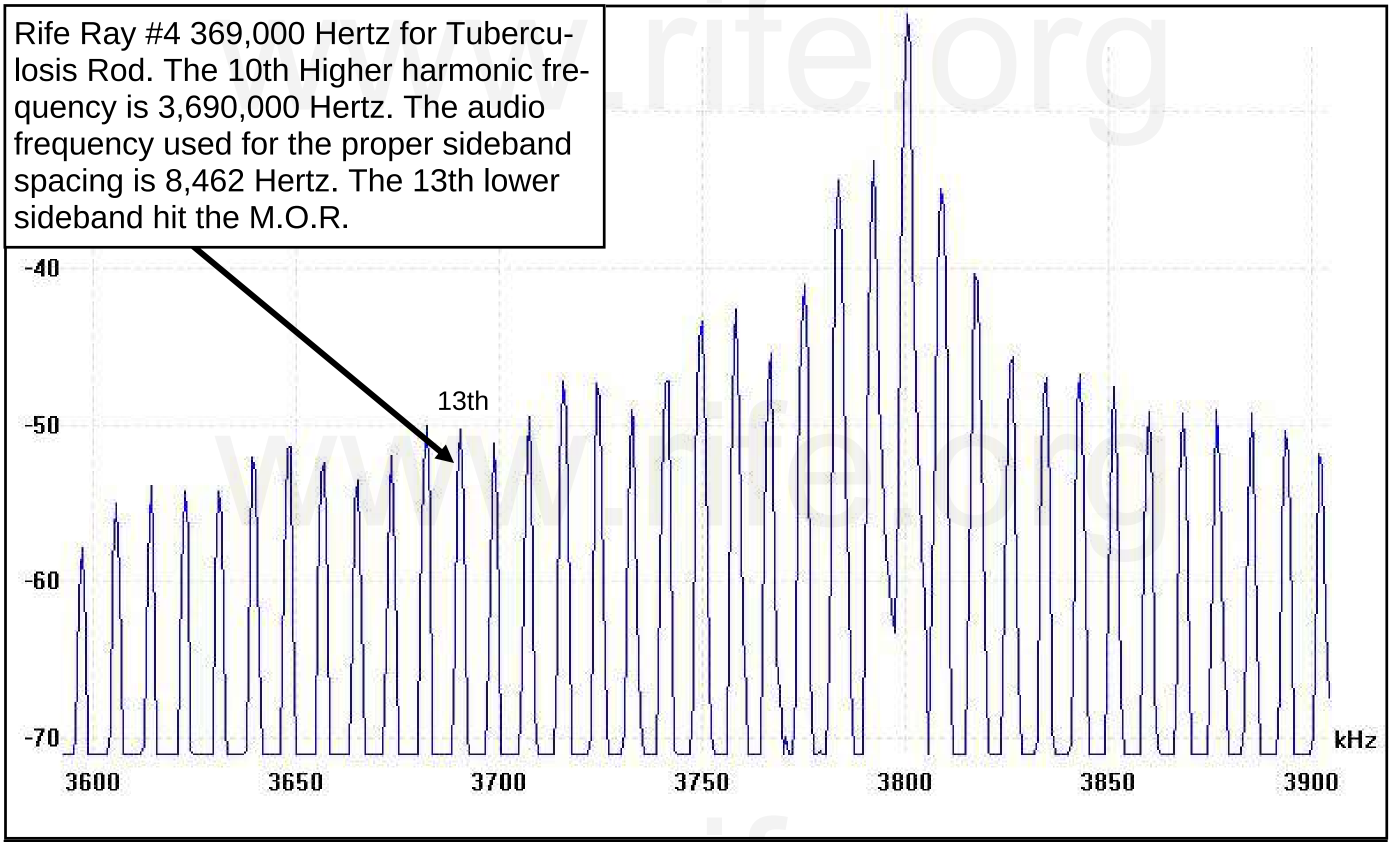

Rife Ray #4 369,000 Hertz for Tuberculosis Rod. The 10th Higher harmonic frequency is 3,690,000 Hertz. The audio frequency used for the proper sideband spacing is 8,462 Hertz. The 13th lower sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

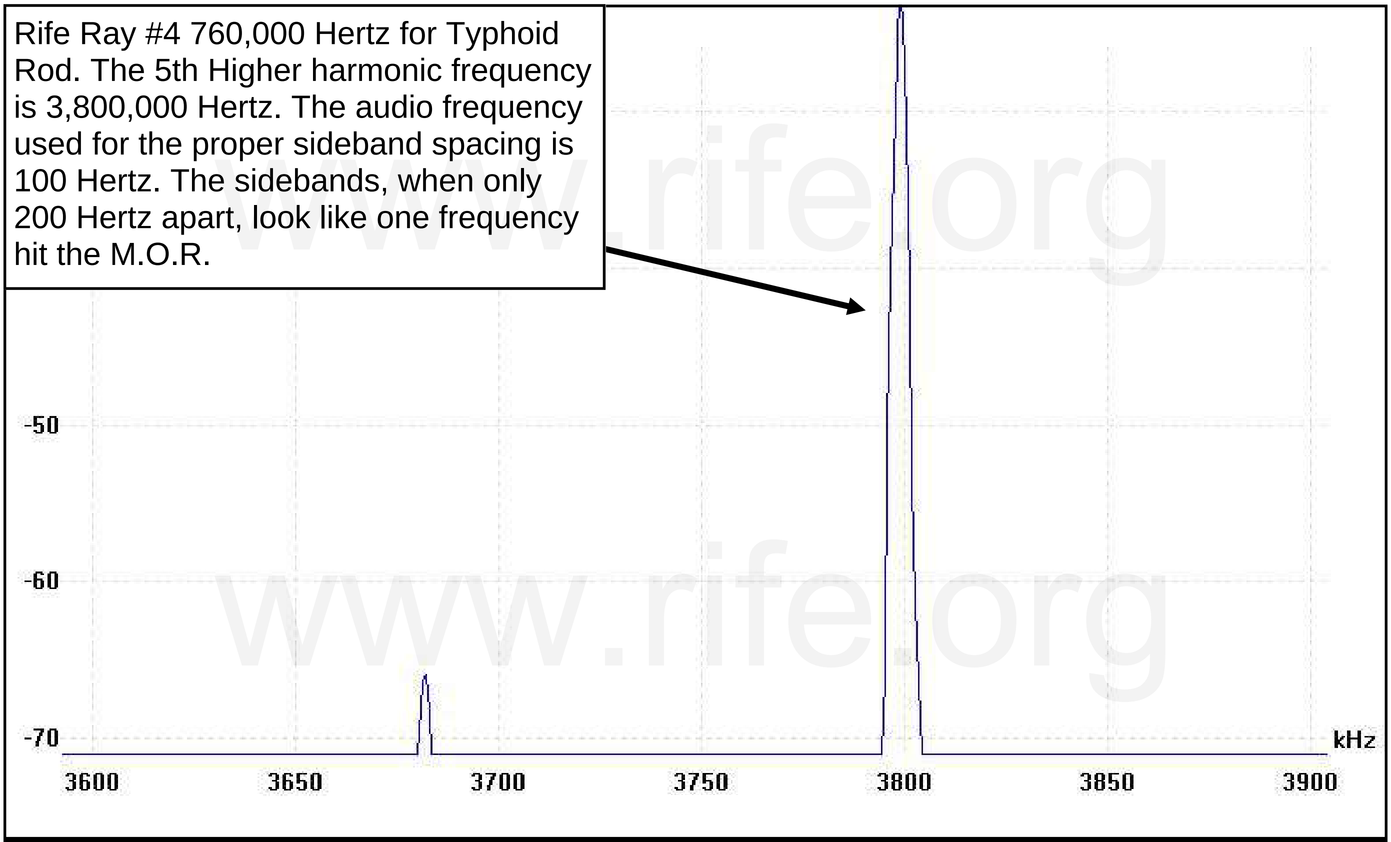
Rife Ray #4 760,000 Hertz for Typhoid Rod. The 5th Higher harmonic frequency is 3,800,000 Hertz. The audio frequency used for the proper sideband spacing is 100 Hertz. The sidebands, when only 200 Hertz apart, look like one frequency hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

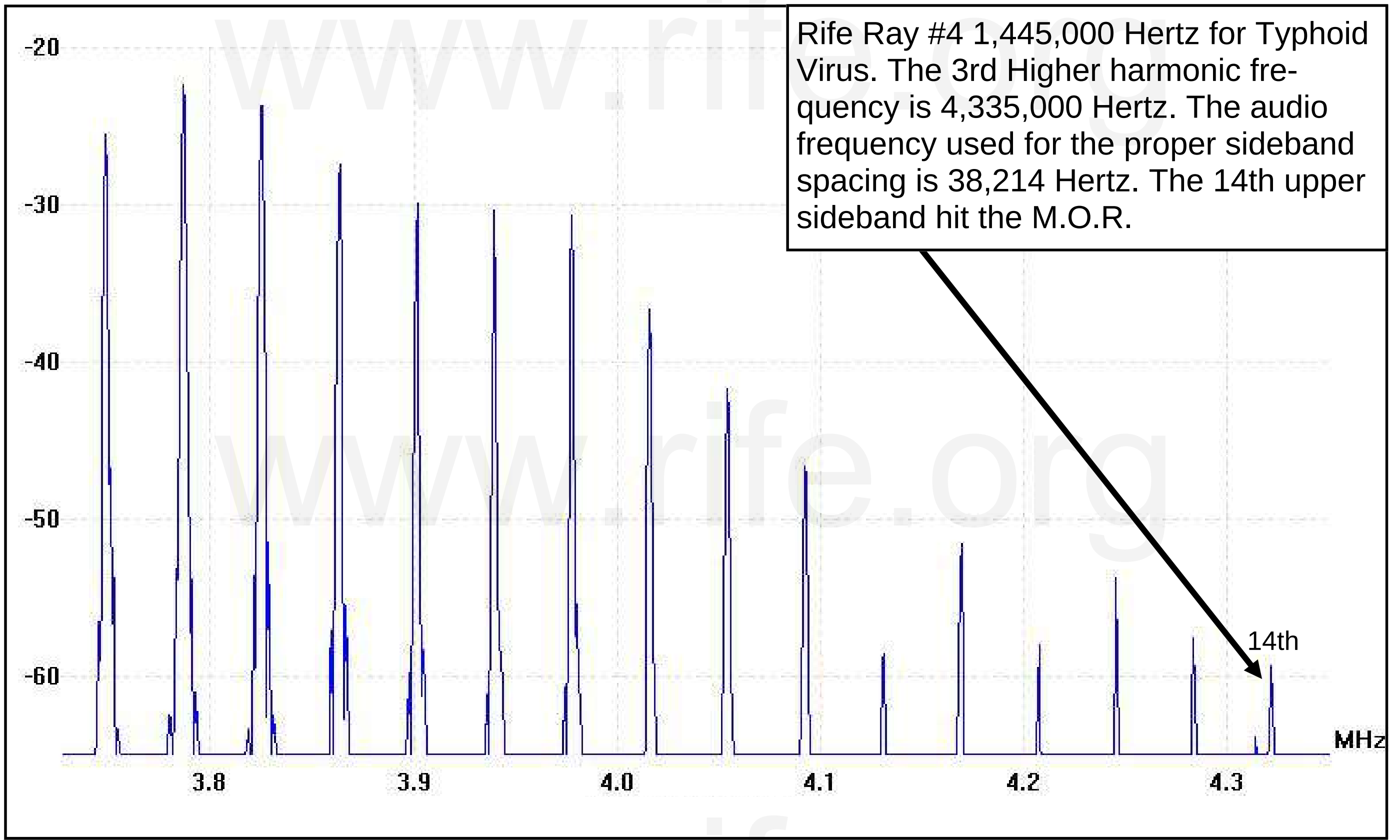
Rife Ray #4 1,445,000 Hertz for Typhoid Virus. The 3rd Higher harmonic frequency is 4,335,000 Hertz. The audio frequency used for the proper sideband spacing is 38,214 Hertz. The 14th upper sideband hit the M.O.R.

Philip Hoyland’s Beam Ray Clinical instrument 3.80 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

Streptothrix

1	192000
2	384000
3	576000
4	768000
5	960000
6	1152000
7	1344000
8	1536000
9	1728000
10	1920000
11	2112000
12	2304000
13	2496000
14	2688000
15	2880000
16	3072000
17	3264000
18	3456000
19	3648000
20	3840000
21	4032000
22	4224000
23	4416000
24	4608000
25	4800000
26	4992000
27	5184000
28	5376000
29	5568000
30	5760000
31	5952000
32	6144000
33	6336000
34	6528000

Anthrax

1	139200
2	278400
3	417600
4	556800
5	696000
6	835200
7	974400
8	1113600
9	1252800
10	1392000
11	1531200
12	1670400
13	1809600
14	1948800
15	2088000
16	2227200
17	2366400
18	2505600
19	2644800
20	2784000
21	2923200
22	3062400
23	3201600
24	3340800
25	3480000
26	3619200
27	3758400
28	3897600
29	4036800
30	4176000
31	4315200
32	4454400
33	4593600
34	4732800

B or E Coli Rod

1	417000
2	834000
3	1251000
4	1668000
5	2085000
6	2502000
7	2919000
8	3336000
9	3753000
10	4170000
11	4587000
12	5004000
13	5421000
14	5838000
15	6255000
16	6672000
17	7089000
18	7506000
19	7923000
20	8340000
21	8757000
22	9174000
23	9591000
24	10008000
25	10425000
26	10842000
27	11259000
28	11676000
29	12093000
30	12510000
31	12927000
32	13344000
33	13761000
34	14178000

B or E Coli Viurs

1	770000
2	1540000
3	2310000
4	3080000
5	3850000
6	4620000
7	5390000
8	6160000
9	6930000
10	7700000
11	8470000
12	9240000
13	10010000
14	10780000
15	11550000
16	12320000
17	13090000
18	13860000
19	14630000
20	15400000
21	16170000
22	16940000
23	17710000
24	18480000
25	19250000
26	20020000
27	20790000
28	21560000
29	22330000
30	23100000
31	23870000
32	24640000
33	25410000
34	26180000

BX Cancer Virus

1	1604000
2	3208000
3	4812000
4	6416000
5	8020000
6	9624000
7	11228000
8	12832000
9	14436000
10	16040000
11	17644000
12	19248000
13	20852000
14	22456000
15	24060000
16	25664000
17	27268000
18	28872000
19	30476000
20	32080000
21	33684000
22	35288000
23	36892000
24	38496000
25	40100000
26	41704000
27	43308000
28	44912000
29	46516000
30	48120000
31	49724000
32	51328000
33	52932000
34	54536000

Gonorrhea

1	233000
2	466000
3	699000
4	932000
5	1165000
6	1398000
7	1631000
8	1864000
9	2097000
10	2330000
11	2563000
12	2796000
13	3029000
14	3262000
15	3495000
16	3728000
17	3961000
18	4194000
19	4427000
20	4660000
21	4893000
22	5126000
23	5359000
24	5592000
25	5825000
26	6058000
27	6291000
28	6524000
29	6757000
30	6990000
31	7223000
32	7456000
33	7689000
34	7922000

Spinal Meningitis

1	427000
2	854000
3	1281000
4	1708000
5	2135000
6	2562000
7	2989000
8	3416000
9	3843000
10	4270000
11	4697000
12	5124000
13	5551000
14	5978000
15	6405000
16	6832000
17	7259000
18	7686000
19	8113000
20	8540000
21	8967000
22	9394000
23	9821000
24	10248000
25	10675000
26	11102000
27	11529000
28	11956000
29	12383000
30	12810000
31	13237000
32	13664000
33	14091000
34	14518000

www.rife.org

	Staph Pyo Aureus	Strept Pyogenes	Syphilis	Tetanus	Tuberculosis Rod	Typhoid Rod	Typhoid Virus
1	**478000**	**720000**	**789000**	**234000**	**369000**	**760000**	**1445000**
2	956000	1440000	1578000	468000	738000	1520000	2890000
3	1434000	2160000	2367000	702000	1107000	2280000	4335000
4	1912000	2880000	3156000	936000	1476000	3040000	5780000
5	2390000	3600000	3945000	1170000	1845000	3800000	7225000
6	2868000	4320000	4734000	1404000	2214000	4560000	8670000
7	3346000	5040000	5523000	1638000	2583000	5320000	10115000
8	3824000	5760000	6312000	1872000	2952000	6080000	11560000
9	4302000	6480000	7101000	2106000	3321000	6840000	13005000
10	4780000	7200000	7890000	2340000	3690000	7600000	14450000
11	5258000	7920000	8679000	2574000	4059000	8360000	15895000
12	5736000	8640000	9468000	2808000	4428000	9120000	17340000
13	6214000	9360000	10257000	3042000	4797000	9880000	18785000
14	6692000	10080000	11046000	3276000	5166000	10640000	20230000
15	7170000	10800000	11835000	3510000	5535000	11400000	21675000
16	7648000	11520000	12624000	3744000	5904000	12160000	23120000
17	8126000	12240000	13413000	3978000	6273000	12920000	24565000
18	8604000	12960000	14202000	4212000	6642000	13680000	26010000
19	9082000	13680000	14991000	4446000	7011000	14440000	27455000
20	9560000	14400000	15780000	4680000	7380000	15200000	28900000
21	10038000	15120000	16569000	4914000	7749000	15960000	30345000
22	10516000	15840000	17358000	5148000	8118000	16720000	31790000
23	10994000	16560000	18147000	5382000	8487000	17480000	33235000
24	11472000	17280000	18936000	5616000	8856000	18240000	34680000
25	11950000	18000000	19725000	5850000	9225000	19000000	36125000
26	12428000	18720000	20514000	6084000	9594000	19760000	37570000
27	12906000	19440000	21303000	6318000	9963000	20520000	39015000
28	13384000	20160000	22092000	6552000	10332000	21280000	40460000
29	13862000	20880000	22881000	6786000	10701000	22040000	41905000
30	14340000	21600000	23670000	7020000	11070000	22800000	43350000
31	14818000	22320000	24459000	7254000	11439000	23560000	44795000
32	15296000	23040000	25248000	7488000	11808000	24320000	46240000
33	15774000	23760000	26037000	7722000	12177000	25080000	47685000
34	16252000	24480000	26826000	7956000	12546000	25840000	49130000

www.rife.org

Photos of the original 1938-1939 Beam Ray Clinical instrument when it was found

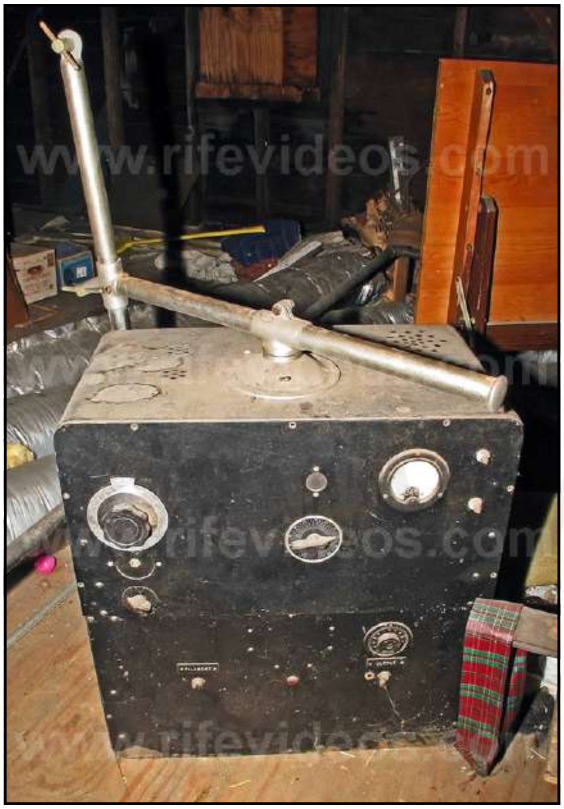

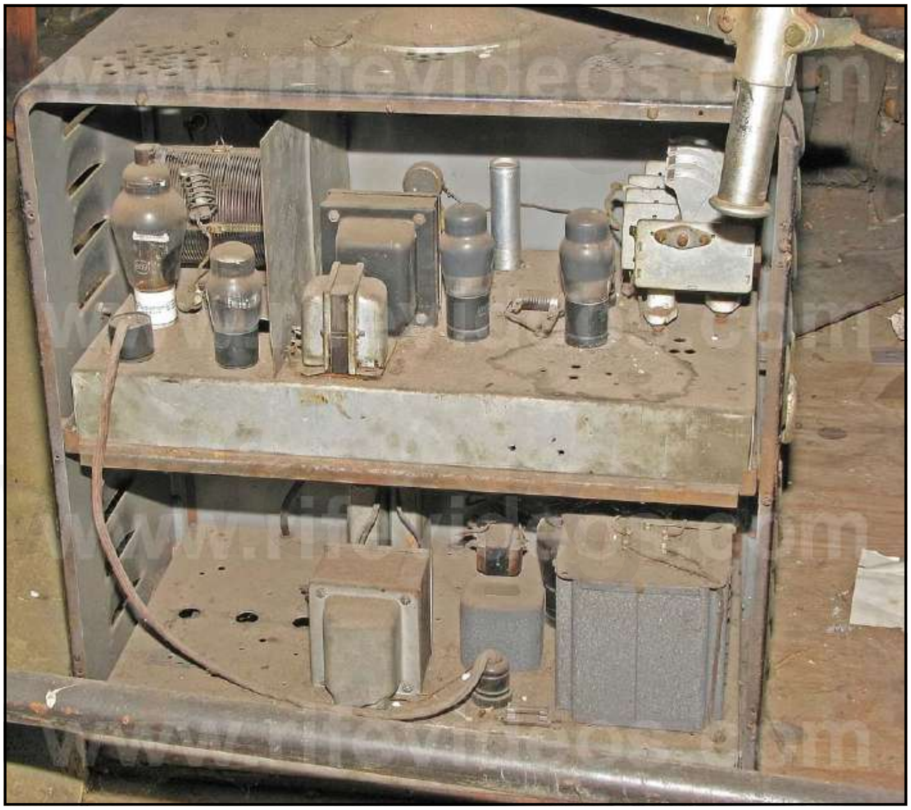

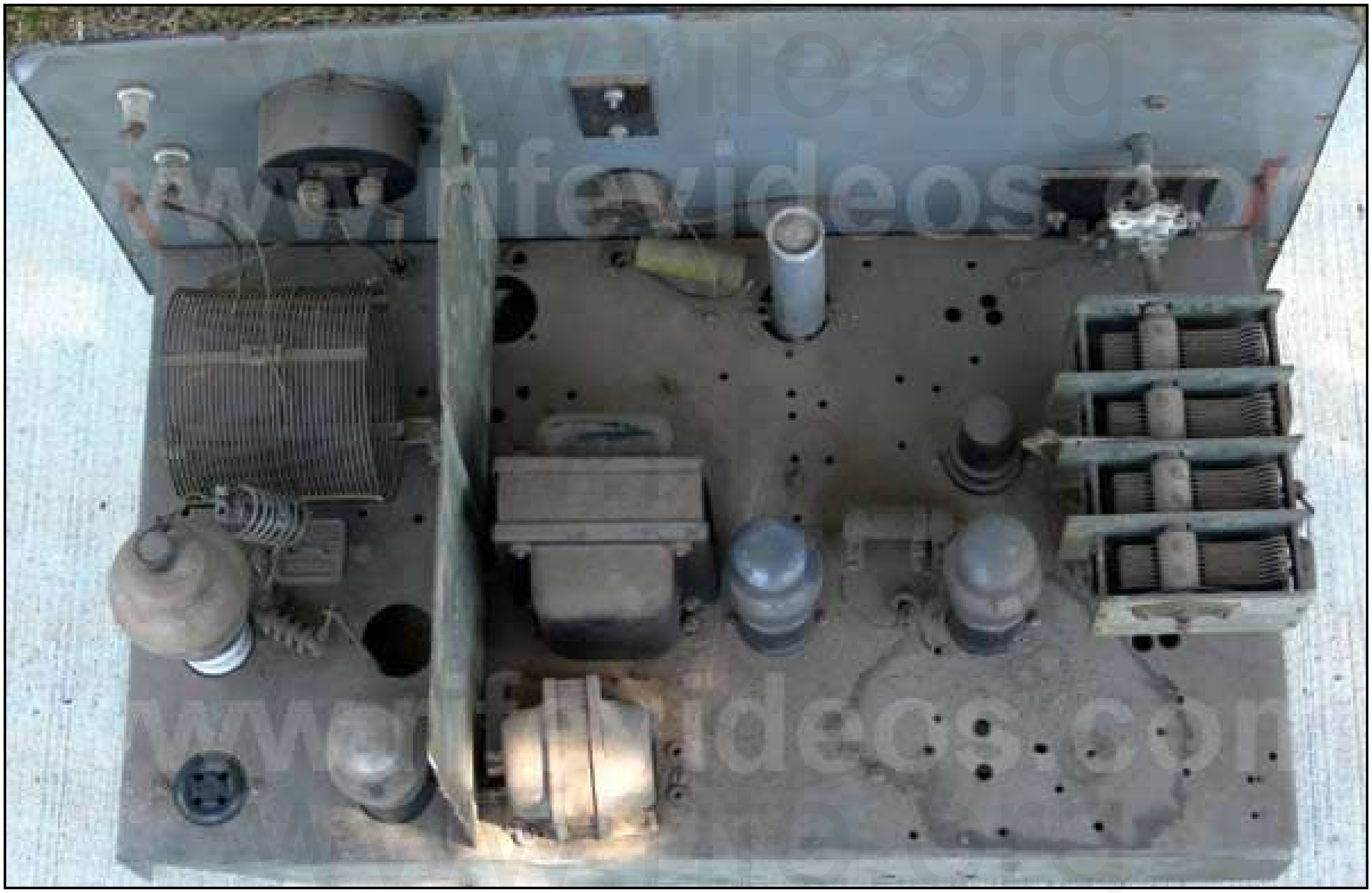

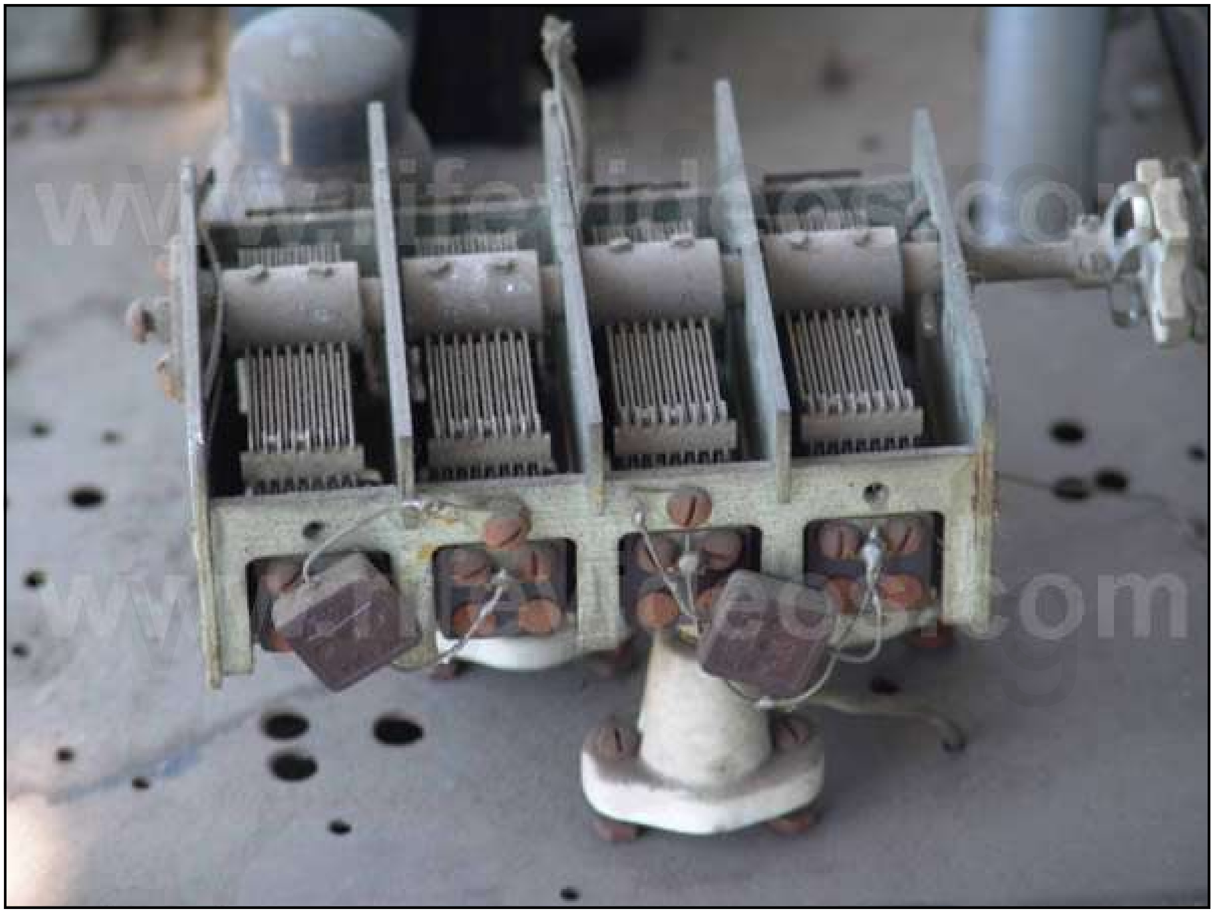

www.rifevideos.com

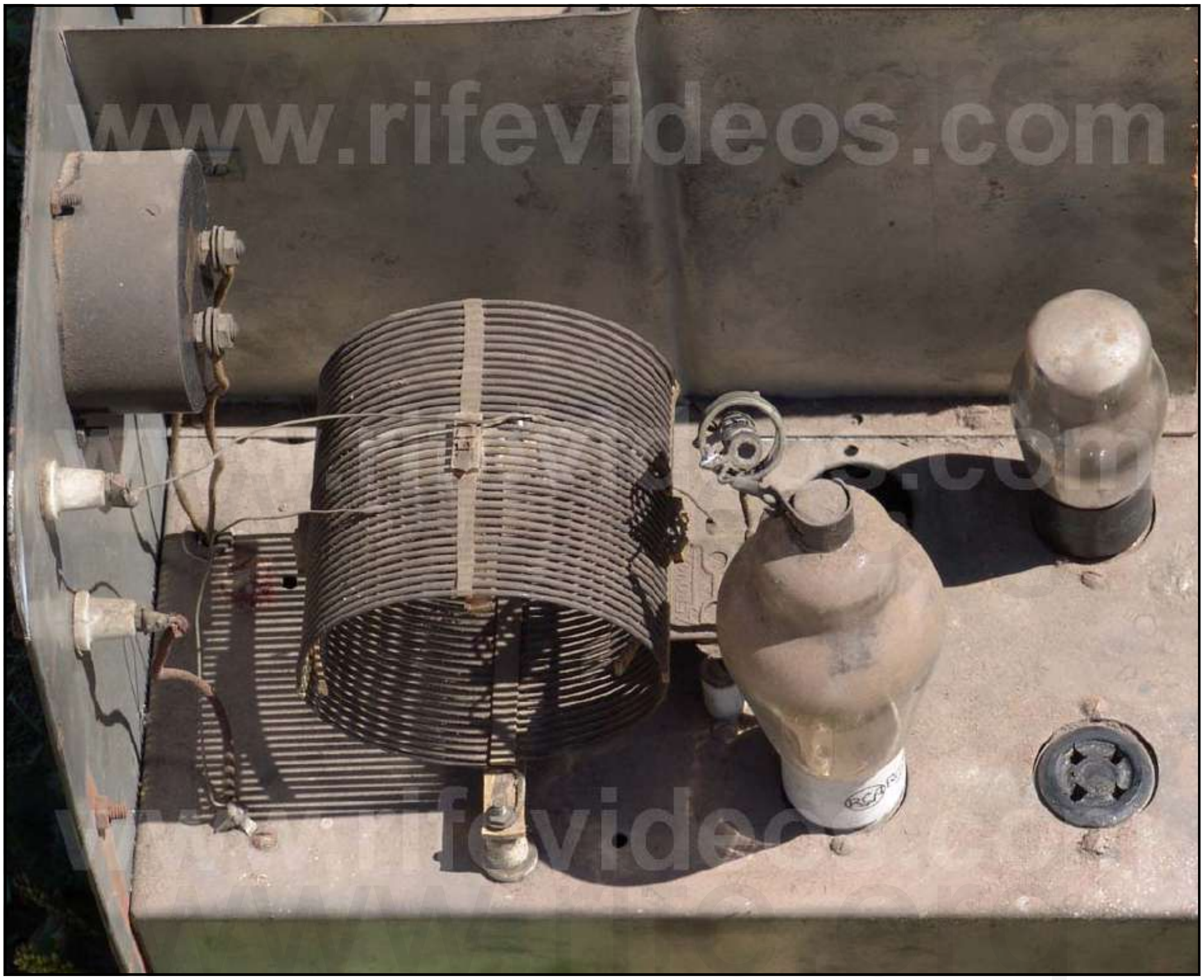

www.rifevideos.com

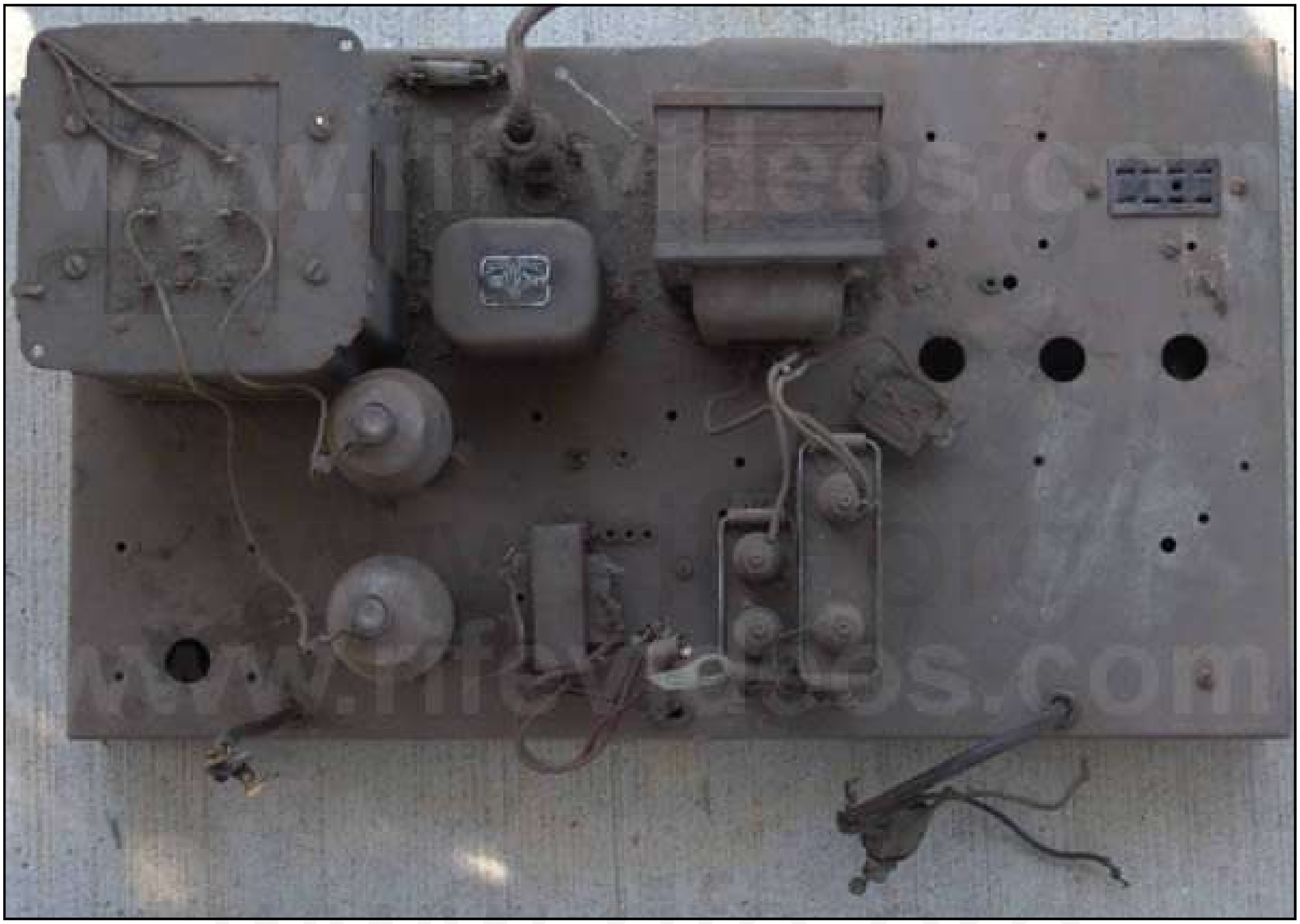

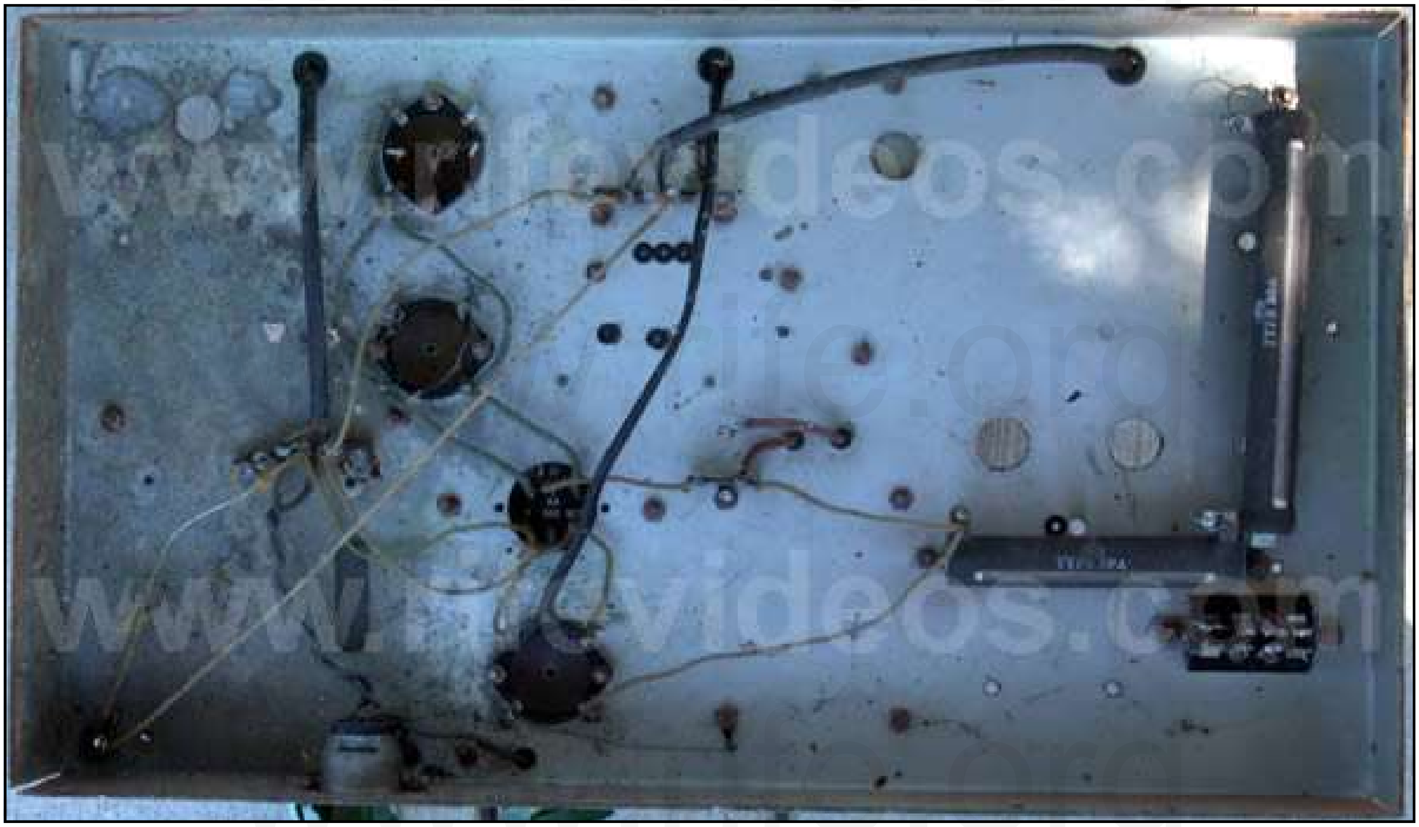

www.nfe.org

Chapter #10

The Gruner Schematic and Philip Hoyland's Beam Ray Laboratory Rife Machine

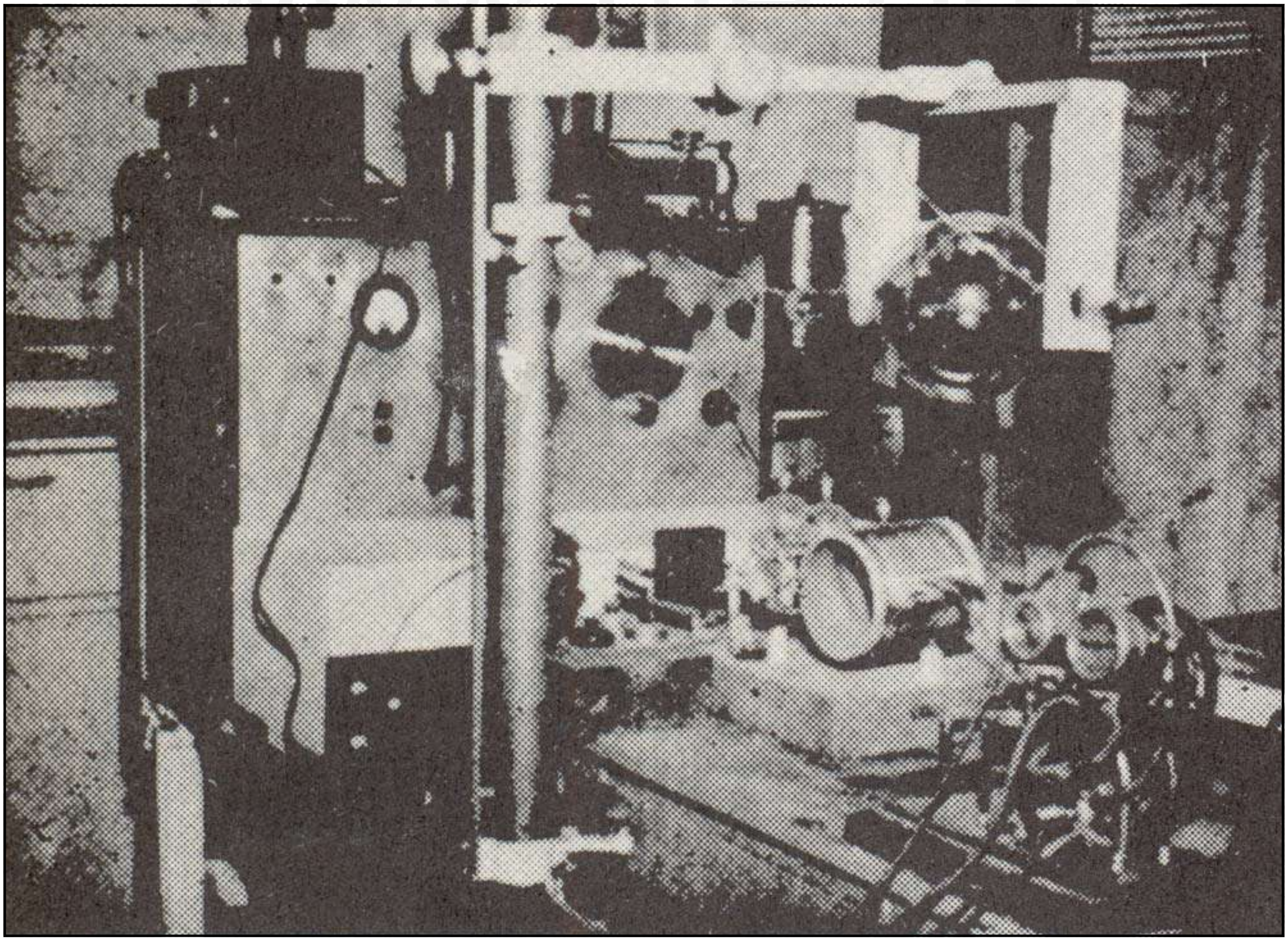

1. Used a ray tube.
2. Had one RF oscillators which was set on a fix frequency.
3. Had a fix audio pulsing circuit.
4. Power usage was about 450 to 600 watts. Output to the ray tube was probably about 40 watts.

In this chapter we will cover two instruments. First, Dr. Gruner's Rife machine and second, the Beam Ray Laboratory Rife machine. The reason for this is the work done on the Gruner instrument lead too many discoveries of how Dr. Rife's machines worked. The first instrument we will discuss is the Gruner Rife Machine. From some recent documents that were found by Mr. Ringas we have found it necessary to change some of the information in this chapter which deals with Dr. Gruner's Rife instrument.

The new documents that were found by Mr. Ringas reveal the fact that the Gruner Rife machine had one fixed RF oscillator. It was previously believed that this instrument was a Beam Ray laboratory instrument which had two high RF frequency oscillators. We now know from these documents that this was an incorrect assumption. The new information comes from two letters written to John Crane back in 1953 from Dr. O. C. Gruner. John Crane at that time was inquiring about the status of Dr. Gruner's Rife machine that was sent to him in 1942 by Dr. Milbank Johnson M.D.:

www.rife.org

DR. GRUNER: *"You refer, I suppose, to the ray machine which Milbank Johnson sent "in bits" to me in Montreal* [Canada]. *Vergil Neher assembled it, but actually I had to purchase a large amount of the radio material to be able to get it completed. I used this machine quite a bit, and it was very successful at first; but after about six months it "faded", first the tubes seemed all right. The only reason would have been changes in the glass of the globe* [ray tube] *as a result of the discharges… I could send the plan* [schematic] *of the apparatus."* (Letter from Dr. Gruner to John Crane, March 31, 1953).

DR. GRUNER: *"I enclosed the diagram* [schematic] *of the Ray Machine as made here in 1942. So I hope it will be of interest to you. It will not be necessary to return it. It is to be noted that it was made only for one frequency; obviously it would be necessary to alter the design to enable many frequencies to be used."* (Letter from Dr. Gruner to John Crane, April 29, 1953).

These two letters changed all our understanding of this instrument. John Crane had altered the schematic and put a second Hartley oscillator on the schematic which confused us when we originally built the instrument. This second Hartley oscillator, added by John Crane, which will be shown in a photo below, made it so the instrument would have had one fixed RF oscillator and a second variable RF oscillator. This second variable RF oscillator, that John Crane added to the schematic, would make it so that instrument could output many different frequencies. Logically this is what Dr. Rife, John Crane and John Marsh wanted in an instrument. But this is not how Dr. Gruner's original instrument worked.

From the analyzing of the original Beam Ray Clinical instrument, which was built by Philip Hoyland, we know that it had the capability of variable audio frequencies. We need to point out here that Dr. Gruner's ray tube instrument did not work on the Beam Ray Clinical instrument harmonic sideband method. From these two letters we now know that Dr. Gruner's instrument work on the same principles as the Rife Ray #4 instrument. That instrument output the specific frequencies directly from its two variable RF oscillators. Because the Rife Ray #4 put out specific frequencies it also used a modulated fixed audio pulsing circuit to pulse the high RF frequencies that were used to devitalize the various microorganisms. This modulated fixed audio frequency pulsing method was used with all of Dr. Rife's instruments. The Gruner schematic which has this modulated fixed audio frequency pulsing circuit reveals how Dr. Rife used this pulsing method in his instruments. With this understanding we now have a clearer understanding of how these four instruments, Rife Ray #3, Rife Ray #4, Beam Ray Laboratory and Gruner Rife machine really worked.

As we pointed out when we initially looked at Dr. Gruner's schematic, several years ago, we were under the wrong assumption, because John Crane altered the schematic by adding the second Hartley oscillator to Dr. Gruner's schematic. But that wrong assumption actually caused us to build an instrument that worked almost exactly as the Rife Ray #4 did. Except the instrument that we built from Dr. Gruner's schematic did not have as high a frequency range as the Rife Ray #4. Because of what we have just explained we feel that the information that we obtained from the initial building of the Gruner schematic, several years ago, should be kept in this report. What was learned, even under the wrong assumption, revealed how the Rife Ray #4 was built and will still be of interest to those who may want to build this instrument.

The Beam Ray Laboratory instrument would have worked like the Gruner instrument and the Rife Ray #4 instrument since both of these were to be used in laboratory work. An instrument that worked like the Beam Ray Clinical instrument, which worked on the harmonic sideband method, would have been useless in any laboratory work since you could never know the exact frequency that killed, devitalize or rendered harmless an organism. For this reason the Beam Ray Laboratory instrument is still included in this information about Dr. Gruner's Rife instrument.

www.rife.org

The initial Gruner schematic work done in 2007-2008

In one of the previous version of this report we dealt with the concept that the Gruner instrument was a heterodyning instrument. We now know that this method was not used in the Gruner Rife Machine therefore the heterodyning concept was removed from a previous rewriting of this document. Though the Rife Ray #4 Rife Machine could output two frequencies at the same time and those two frequencies did heterodyn in the ray tube, this was a byproduct of the instrument not the method used to produce the M.O.R. frequencies needed to devitalize or render harmless the microorganisms. The knowledge that we gained through our testing of this concept is the reason much of the history and work that we did at that time is still included in this report. It was that testing which gave us the understanding of how the Rife Ray #4 worked and how the Beam Ray Laboratory Rife Machine would have worked.

We do not have a picture of Dr. Gruner's Rife instrument so we have no idea what it looked like. This however did not affect building the instrument from the schematic. The Rife Machine in the photo shown at the beginning of this chapter is probably a photo of the Beam Ray Laboratory instrument built by Philip Hoyland. John Crane dated that instrument as being built in 1935 but we know that the Rife Ray #4 was built in 1935. The knowledge of the Laboratory instrument came from the Beam Ray Trial manuscript. A complete copy of this transcript was provided when Steven Ross allowed us to scan it. I would like to acknowledge his generous contribution of this information which has given us a great deal of understanding so that we could figure out the history of how and when the instruments were built.

After reading for the first time the complete Beam Ray Trial manuscript I found there was mention of a Laboratory instrument. It appeared that this Laboratory instrument was probably to be used by those who would be working in laboratories with microorganisms for testing. Dr. O. C. Gruner worked in a laboratory with organisms and worked with Dr. Rife on the organism called Cryptomyces Pleomorpha fungi. The Laboratory instrument was mentioned two or three times but no real information was given about it. Below are statements made by Bertrand Comparet and Philip Hoyland in the Beam Ray Trial that gives us some important information about this instrument. (#209-210, 217-222):

COMPARET: *"The four machines bought by the British were two so-called laboratory types and two so-called clinical types, what was the difference between the two."*

HOYLAND: *"The clinical type was similar in all respects to the Rife machine except that it did not have* [word missing] *of the* [word missing] *used on Mrs. Henderson."*

COMPARET: *"How was the price of these machines fixed."*

HOYLAND: *"The price was decided from the costs of what it cost to manufacture the first machine that was sold to Dr. Hamer."*

COMPARET: *"How much was that?"*

HOYLAND: *"I think it was four hundred dollars plus the royalty."*

COMPARET: *"Wasn't it five hundred dollars plus royalty on the clinical type and six plus royalty on the lab type."*

HOYLAND: *"I don't remember."*

Since this Beam Ray Laboratory Rife Machine was for Laboratory work it would have worked like the Rife Ray #4 but because the case was smaller it probably had a smaller frequency range. The

www.rife.org

Rife Ray #4 had nine frequency bands that covered from 87,000 Hertz to 22.5MHz (22,500,000 Hertz). The first four bands of the #4 covered from 87,000 Hertz to 2,140,000 Hertz. These four frequency bands would cover the whole list of Dr. Rife's disease organisms listed on the Rife Ray #4 documents. With this understanding we know that the Laboratory instrument at least covered this frequency range. We will not speculate if it had a higher frequency range because we really do not know if it did.

Deciphering Dr. Gruner's Beam Ray instrument schematic

Several years ago a group of us had been looking on the Gruner schematic in hopes of trying to figure out how it worked. I had built and tested both the 1953 AZ-58 Beam Ray Clinical instrument and 1950's Aubrey Scoon, Beam Ray Clinical replica instruments. To our knowledge none of these instruments ever obtained the same results as the original Beam Ray Clinical Rife Machine built by the original 1938 Beam Ray Corporation. All the documentation we had showed that there were changes made to the original Beam Ray design which compromised the1953 AZ-58, and possibly 1950's Aubrey Scoon, Verne Thompson instruments. John Crane, over the years told many people that the AZ-58 and the audio frequencies it used were Dr. Rife's original frequencies. The Rife documents we have show that what John Crane claimed was not correct. Dr. Rife was not using audio frequencies in 1934 as John Crane and John Marsh claimed. Rebuild of the 1953 AZ-58 and Aubrey Scoon's instrument partially made the rediscovery of the Beam Ray Laboratory instrument possible.

At the 2003 Rife Conference, a gentleman put up the Gruner schematic of an original Beam Ray instrument. From reading the Rife documents I knew that this schematic existed because John Crane had mentioned it in his papers. John Crane said that the AZ-58 was built from that schematic. When I saw it I knew it was important, so I took still photos of it with my video camera. Because the video camera only had a one mega-pixel capability I took many up close photos knowing I could put it back together at a later date. Back in 2004 I gave this schematic to Aubrey Scoon and his British Rife group in hopes that they could look it over and correct any mistakes that may have been made. They redrew the schematics, unfortunately without fully correcting them, and put them up on their web site. This information was released because we wanted everyone to have access to it.

Back on July 27, 2007 a Mr. Andrews, who was one of the British Rife group and I got into another conversation about the Gruner schematic. He asked me to send him another copy of the original Gruner schematic so he could look it over again. An email conversation began at that time which included Mr. Peters, Mr. Ringas, Mr. Andrews, Mr. Berger and myself. Mr. Peters immediately noticed that the schematic that had been redrawn by the British group had some errors in it. Discussions continued on and off for a few months until one day Mr. Peters noticed an over sight when looking over the schematic again. At the time we thought it had to do with the heterodyning method. But now we understand that this oversight was one of the keys to understanding how the Rife Ray #4 and Beam Ray Laboratory Rife Machines would have worked.

Mr. Peters sent Mr. Ringas and I an email. He mentioned that a possible test could be made that would determine if this observation of his was the key to understanding how the Beam Ray Laboratory instrument really worked. I called Mr. Peters and had a discussion with him and he told me how we could make these tests. I told him that we did not need to do the test with solid state frequency generators because I had conducted a similar test back when John Bedini and I were working on the AZ-58 tests. John Bedini and I knew that the original Rife Ray #4 instrument was a lot more powerful than the AZ-58 because of the Rife documents. After more accurate testing we found the AZ-58 only output about 40 watts from the ray tube. I told him that I still had my 1950's Aubrey Scoon replica and several AZ-58's on the shelf. I told John Bedini that the original Beam Ray Clinical instrument, from the documentation that we had, may have output about 50 to 60 watts from the ray tube. He told me how I could make a test, by putting two AZ-58s together, which would give me at least 60 watts out of the ray tube. I didn't know it at that time but that test was probably the way the Rife Ray #4 and Beam Ray Laboratory instrument worked. I told Mr. Peters I would connect the AZ-58 and Aubrey Scoon instruments together again and make the tests that he suggested, but this time I would use my spectrum analyzer and we would fully make tests and find out if this was the method Philip Hoyland used. This test made the two instruments work exactly the same as the Gruner Beam Ray schematic would have worked with John Crane's addition of the second variable RF oscillator. This test was the key to understanding how Philip Hoyland's Laboratory instrument and Rife Ray #4 instrument worked. I had always stated that Philip Hoyland had to have come up with his method using math because they didn't have any spectrum analyzers back in 1936. In looking back that test showed that the Rife Ray #4 was heterodyning in the ray tube. Though we now know the heterodyning method was not used by Dr. Rife the method of connecting the ray tube up to two RF oscillators showed that this was the method used with the Rife Ray #4 which had two RF oscillators.

We will now show how Philip Hoyland's Rife Ray #4 and Beam Ray Laboratory instrument worked. We can do this because we have been able to rebuild these instruments from the Gruner schematic. This information should be of great interest to all who have been interested in Dr. Rife's work.

Rebuilding the Philip Hoyland Beam Ray Laboratory instrument

When this schematic was analyzed Mr. Peters noticed that the British group had overlooked a second Hartley RF oscillator that was in the lower left corner of the Gruner schematic. This was the variable oscillator that John Crane added to the schematic after he received it from Dr. Gruner. At this time we did not know that this second oscillator was not a part of the original schematic. The British group believed this oscillator was the same RF fixed Hartley Oscillator that used the 809 tube. Below is a photo of the complete Gruner Beam Ray schematic.

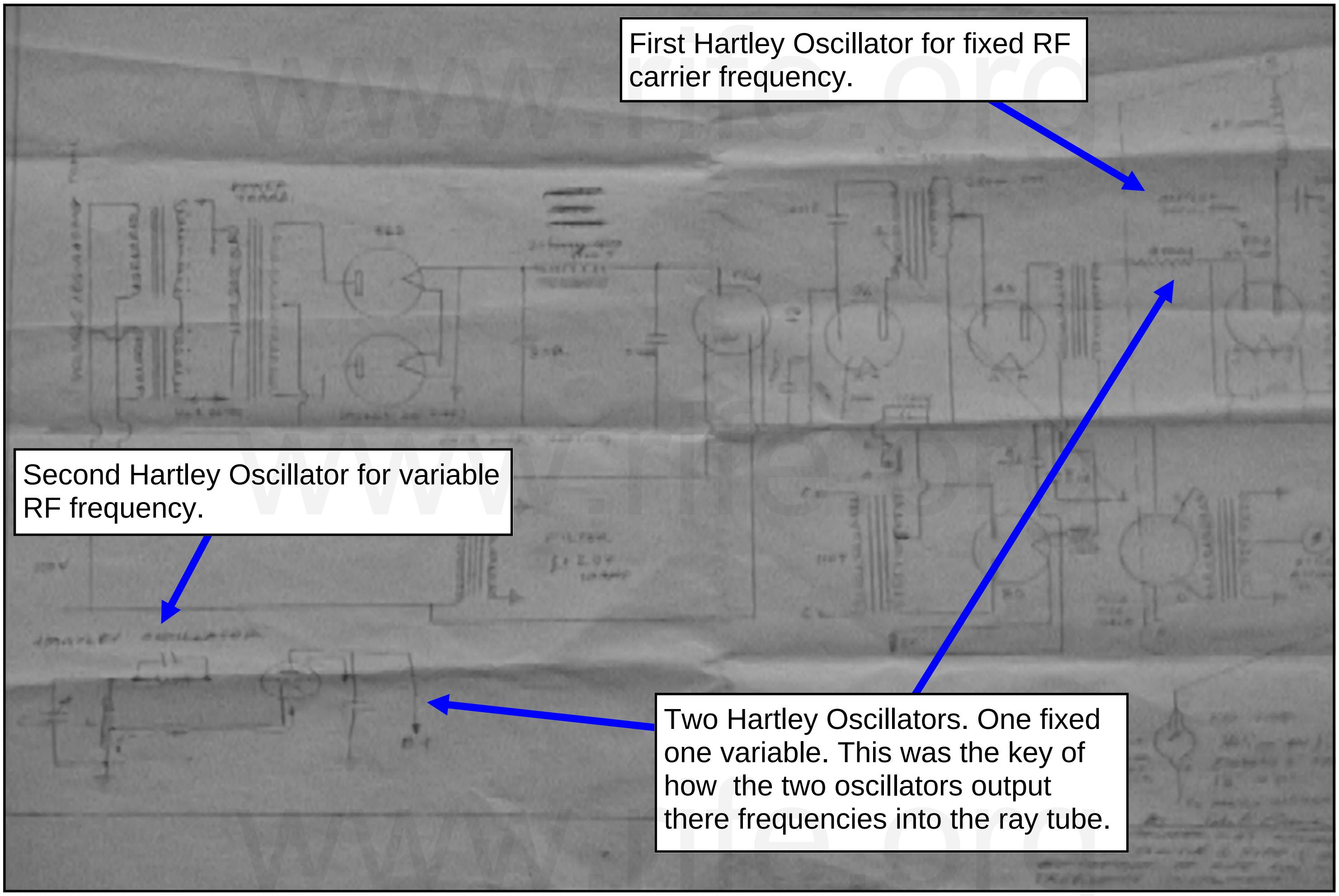

First Hartley Oscillator for fixed RF carrier frequency.

Second Hartley Oscillator for variable RF frequency.

Two Hartley Oscillators. One fixed one variable. This was the key of how the two oscillators output there frequencies into the ray tube.

www.rife.org

In next schematic photo, shown below, is the fixed RF carrier frequency section that used the 809 tube.

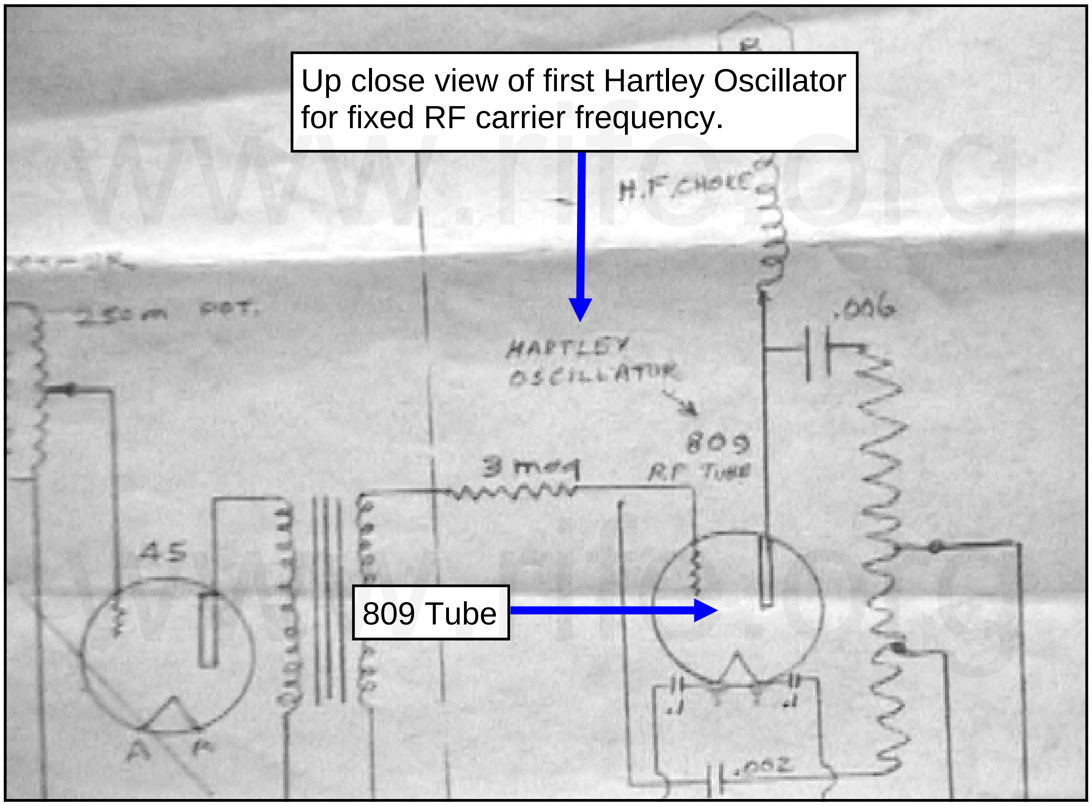

Up close view of first Hartley Oscillator for fixed RF carrier frequency.

809 Tube

In the next schematic photo, shown below, is the second oscillator that was overlooked by the British group. This was the oscillator that John Crane added to the schematic which caused us to misinterpret how the instrument worked. You will notice that the second oscillator that John Crane added says “Hartley Oscillator.” Because the first fixed oscillator that used the 809 tube is a Hartley Oscillator the British group assumed that both of these oscillators were the same. This overlooking of the second oscillator that John Crane added would have made the instrument work exactly as Dr. Gruner had stated his instrument worked. Dr. Gruner stated that the instrument was fixed on one frequency. It was the overlooking of this second Hartley oscillator, added by John Crane, that Mr. Peters noticed. If you look at the second Harley Oscillator, in the photo below, that does not use the 809 tube, you will notice that it has a variable capacitor. This variable capacitor shows that there were two Hartley Oscillators shown on the schematic. One fixed and one variable.

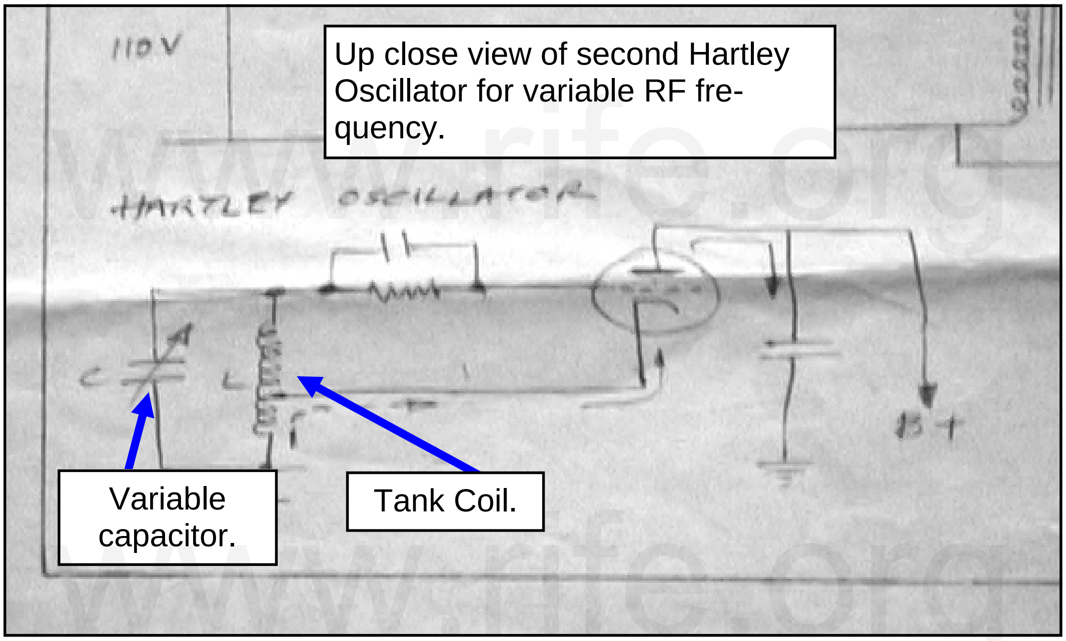

Up close view of second Hartley Oscillator for variable RF frequency.

Variable capacitor.

Tank Coil.

www.rife.org

The next schematic photo, shown below, shows that the first fixed Hartley Oscillator was connected from the tank coil to the negative side of the ray tube. If you look closely at the photo you will notice that the positive side of the ray tube was also to be connected to a Hartley Oscillator.

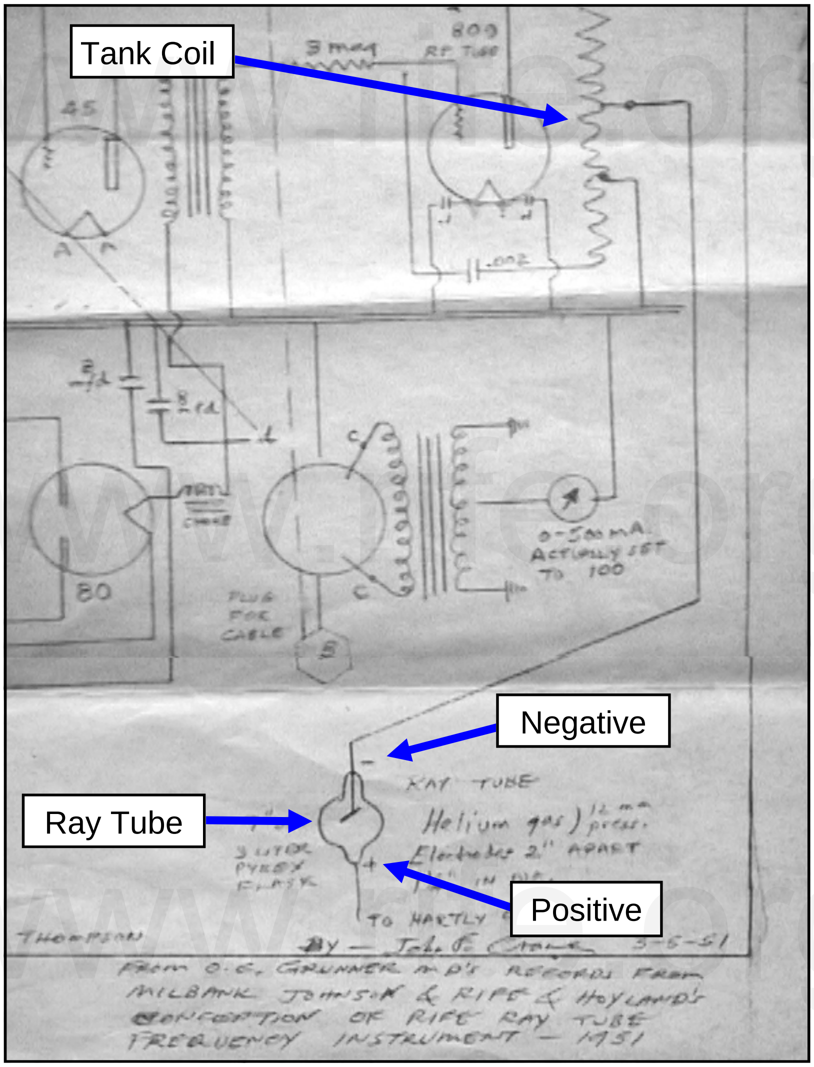
Tank Coil
809
RF TUBE
3 meg
45
A
.002
80
CHOKE
PLUG FOR CABLE
C
C
0-500 mA. ACTUALLY SET TO 100
Negative
Ray Tube
Positive
RAY TUBE
Helium gas) 12 mm press.
Electrodes 2" APART
IN DIA.
3 LITER PYREX FLASK
TO HARTLY
THOMPSON
BY — John F. Crane 3-5-51
FROM O.C. GRUNNER M.D'S RECORDS FROM MILBANK JOHNSON & RIFE & HOYLAND'S CONCEPTION OF RIFE RAY TUBE FREQUENCY INSTRUMENT — 1951

If you hooked the positive side of the ray tube back up to the same fixed Hartley Oscillator it would have only output one frequency as described by Dr. Gruner. This adding of the second oscillator, by John Crane, is why there was confusion on how this instrument worked. The positive side of the ray tube was supposed to be hooked back up to the same fixed oscillator. But with the adding of the second variable Hartley Oscillator, by John Crane, would mean that the positive side of the ray tube would have been hooked up to the second Hartley oscillator. This would have made it so the ray tube would be connected between the two Hartley Oscillators. The negative side of the ray tube connected to one oscillator and the positive side of the ray tube connected to the other oscillator.

The second Hartley Oscillator was also an RF Oscillator. It had a tank coil and a variable capacitor for changing the RF frequencies. Anyone looking at this schematic will notice that it does not have any variable audio oscillator. Philip Hoyland's Beam Ray Gruner instrument was using an RF frequency not an audio frequency. This would logically mean that Dr. Gruner was using one of the Rife Ray #4 frequencies. Because Dr. Gruner was working on the cancer organism of Cryptomyces Pleomorpha fungi his instrument was most likely set on that organisms frequency.

Because we did not know that John Crane had added this second oscillator we came to the conclusion that the logical way to build the instrument would have been to have had two Hartley Oscillators using the 809 tubes. So this is the way that we rebuilt the instrument. By using the method of connect-

www.rife.org

ing the ray tube between the two Hartley Oscillators, both variable, the instrument could output two high RF frequencies at the same time without the bandwidth problems that would have come with trying to modulate frequencies through a tank coil. The tank coil in the Beam Ray Clinical instrument can only pass modulated frequencies up to about 250,000 Hertz. Because we put the ray tube between the two Hartley Oscillators and found that the instrument could work this way we determined that Philip Hoyland most likley built the Rife Ray #4 and Beam Ray Laboratory instrument in this same way. Our instrument could output two frequencies at the same time.

The Beam Ray Laboratory instrument probably had some band switches like the Rife Ray #4. What the frequency range of this Beam Ray Laboratory instrument was is not known but it would have gone up to at least 1.80 MHz like the Rife Ray #3 Kennedy equipment did. Dr. Rife's Rife Ray #3 and Rife Ray #4 machines put out specific frequencies and this instrument, it appears, was built to do the same thing. If Dr. Rife wanted 1,604,000 Hertz he would set the oscillator to 1,604,000 Hertz. This was the type of instrument Dr. Rife used. When you look at the case of the Laboratory instrument it is a large case that could have easily held the necessary components for this instrument. The instrument would have had two large dials on the front of it. In the photo, shown at the beginning of this chapter, the pole that holds the ray tube is blocking part of the panel where we would expect to see the second dial for the second oscillator.

The modulated Audio Frequency pulsing circuit

In this report, in the previous chapters, we have mentioned a fixed audio frequency pulsing circuit. We will now cover it in more detail because the Beam Ray Gruner instrument had this circuit in its schematic. This circuit pulsed the high RF frequencies. We know it was important because it was considered the secret that made the instrument work. If you look at the next schematic photo, shown below, you will see two audio transformers. This is the only indication of any audio frequency used in Dr. Gruner's instrument.

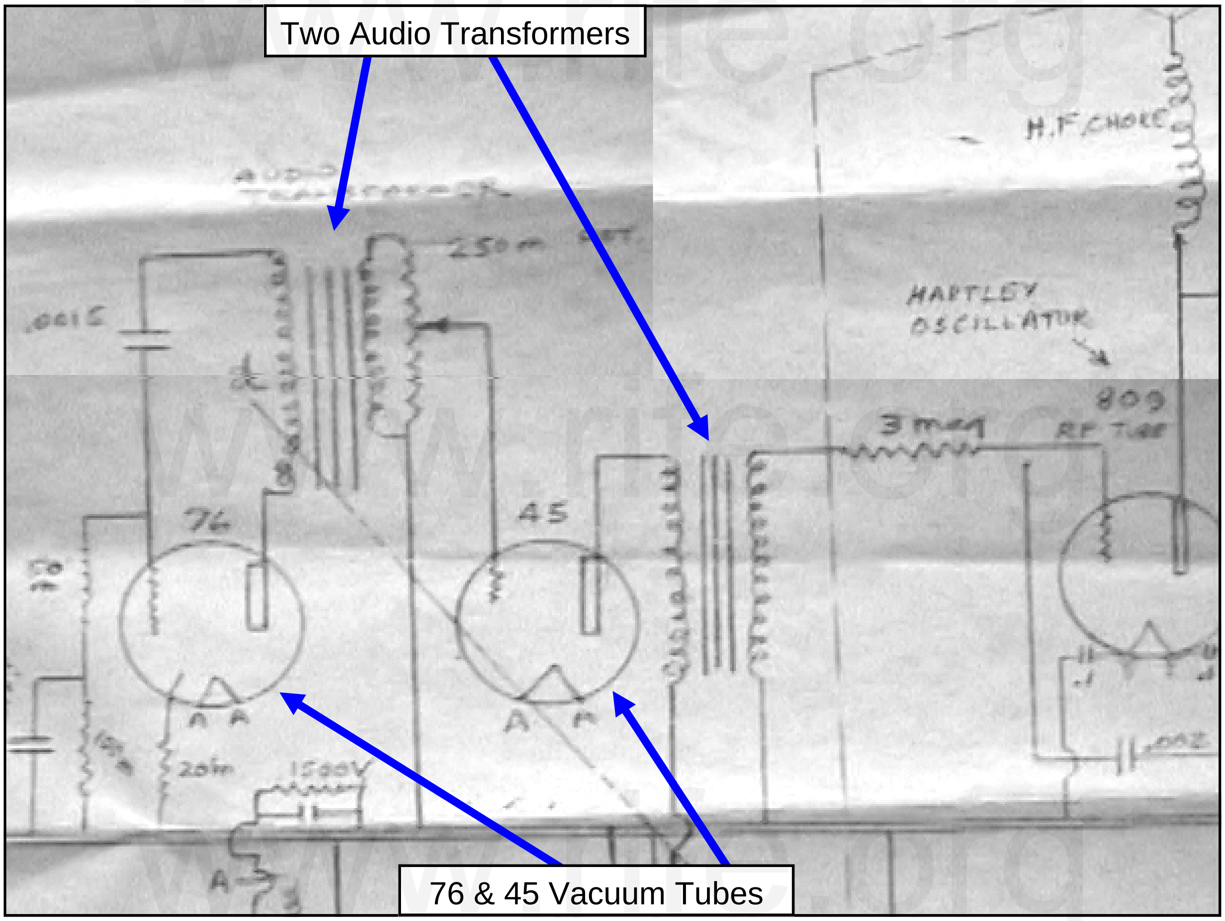

www.rife.org

It was not a variable audio oscillator but it was for a single fixed audio frequency. The 76 and 45 vacuum tubes along with the two audio transformers make up this circuit. Mr. Peters built this section and found that the frequency was at about 1330 Hertz and it pulses, through modulation, the fixed Hartley Oscillator RF carrier frequency that used the 809 tube. The next two photos, shown below, is this rebuilt circuit.

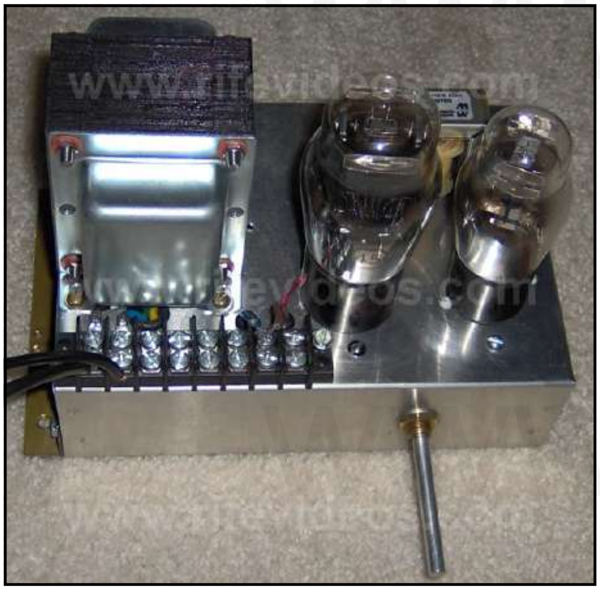

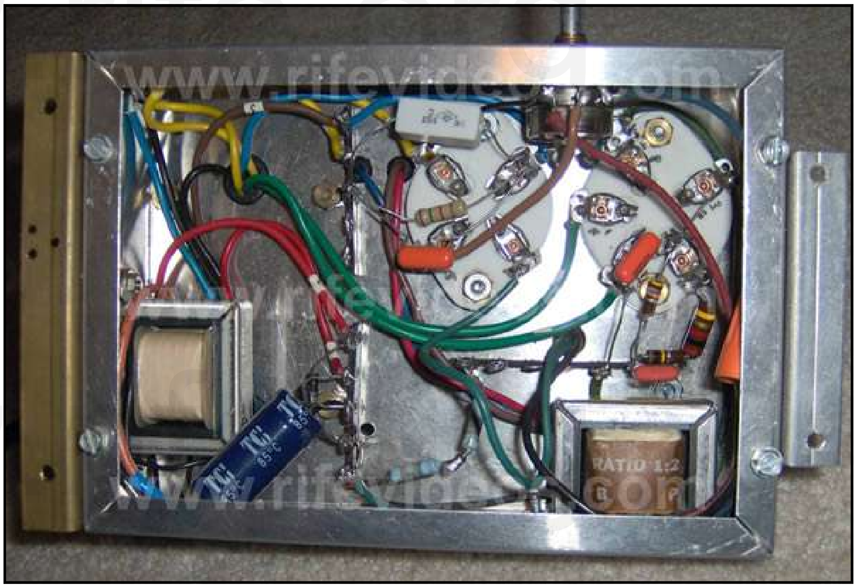

This 1330 Hertz frequency could have been a little higher or a little lower frequency because he used modern transformers instead of the original 1940's transformers. This pulsing frequency is a lot faster than the eye can see so no one would know it was in the instrument. If you look at the next photo, shown below, you will see the waveform of the pulsing frequency. It resembles a damped wave minus the ring oscillations of a true damped wave.

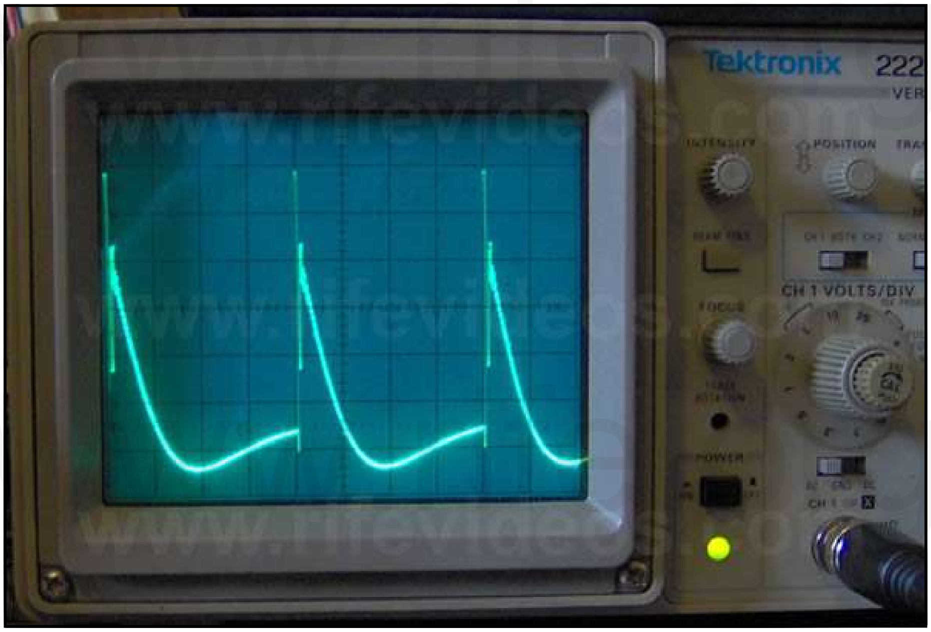

This waveform also looks like the waveform of the Rife Ray #4. This waveform would produce the effect that John Crane mentioned as he narrated Dr. Rife’s 1936 Lab video:

CRANE: *“Now the spikes that you see on the frequencies are the lethal part that kill and devitalize the virus. They are the resonant peaks of the frequencies which increase the voltage to a very high potential which the cells of the virus wall can not tolerate and they break up into many pieces and are destroyed.”* (Dr. Rife’s Lab Film Narrated by John Crane in the 1970’s).

www.rife.org

It's doubtful that this understanding of the spikes would have been something that John Crane would have known anything about had Dr. Rife not told him about it. From the statements we have read the resonant frequency of an organism is not enough to devitalize it. It's apparent that an organism's resonant frequency will not harm it unless the resonant frequency is pulsed through modulation with a waveform that produces a high potential voltage rise. It also appears that this modulated audio pulsing circuit would have been necessary for all of Dr. Rife's high RF frequencies. Logically this same type of circuit would have been in Dr. Rife's Rife Ray #3 and Rife Ray #4 and this is why the Gruner instrument also had this kind of circuit built into it. Also, logically his Beam Ray Laboratory instrument had this circuit.

In the previous chapter of this report when we covered the original Beam Ray clinical instrument. We showed that a sine wave audio frequency modulated onto a carrier frequency, in this M.O.P.A. style instrument, was sufficient enough to create the necessary pulse to devitalize the various microorganisms. The analyzing of that Beam Ray Clinical circuit showed that it almost creates a square wave frequency. The Beam Ray Clinical instrument output variable audio frequencies well above 10,000 Hertz. This indicates that the pulse rate of the modulated audio frequency is not important, only that the high RF frequency is pulsed.

There is another important effect that happens to the plasma of a ray tube when you pulse it with a low audio frequency of a damped shaped waveform or square wave waveform. Because the duty cycle is very low it allows deionization of the plasma which makes it possible for the very high potential voltage rise to be emitted from the ray tube. A square wave audio frequency of a 50% duty cycle should be just as effective as a damped waveform. A square wave has the same high potential voltage rise on the leading edge as this damped wave. Philip Hoyland found that even a sine wave frequency was sufficient to achieve the same result when used in the M.O.P.A circuit. His Beam Ray Clinical instrument circuit was modulated with a sine wave audio frequency that produced a waveform that almost looks like a square wave waveform and that waveform would devitalize the organisms. The next photo, shown below, is that waveform. We are showing this waveform so that the reader will understand what we are talking about.

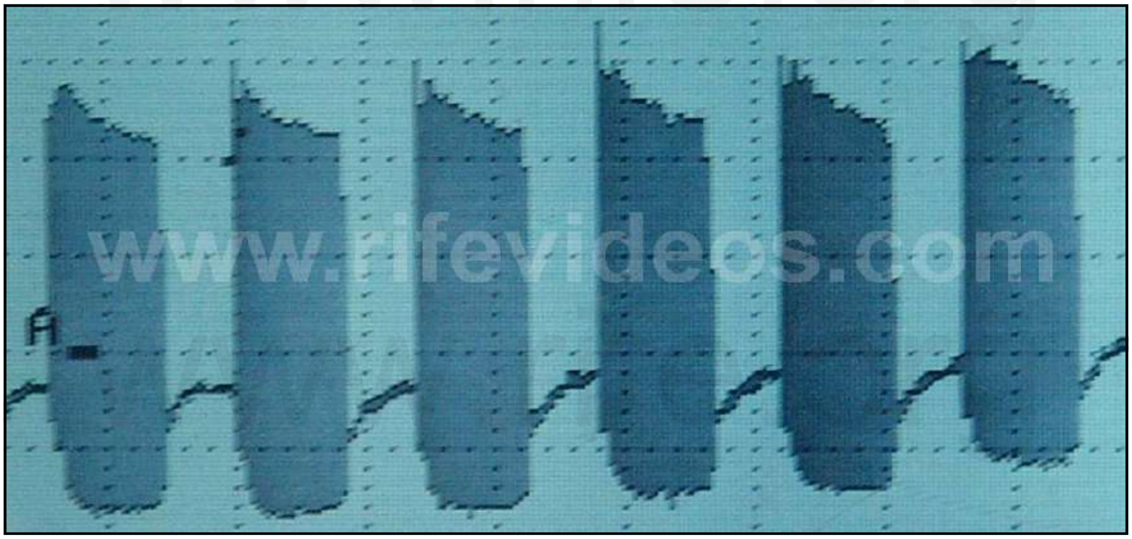

www.rifevideos.com

This all indicates that an audio frequency modulation with a high RF frequency is what makes the frequencies work on microorganisms. The waveform of the audio frequency whether it is a damped wave or square wave makes no difference in its effectiveness because both of these waveforms have been used in Dr. Rife machines.

www.rife.org

The next photo, shown below, is a picture of Dr. Rife's waveform from his Rife Ray #4 Rife Machine. We showed this waveform in a previous chapter but we need to show it again. This waveform came from Dr. Rife's 1936 lab film. The lab film shows that he put a metal plate about 3" X 8" under the ray tube and ran his oscilloscope lead to it so he could read the frequencies. This photo shows the type of waveform he was using to devitalize organisms. It also matches the waveform produced by the 76 and 45 vacuum tubes in the Gruner instrument we are looking at in this chapter. We now know from the rebuilding of Dr. Gruners instrument and from the analyzing of the Beam Ray Clinical instrument, the two waveforms that were used, how they were created, and the method that should be used for doing M.O.R. research.

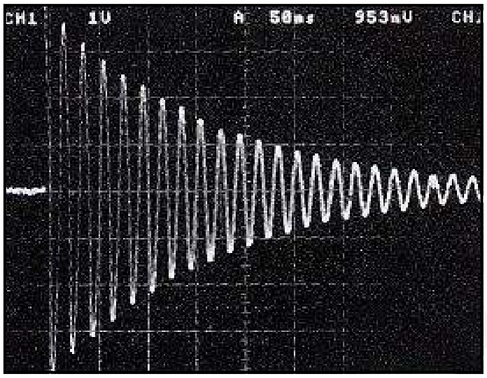

Both Mr. Peters and I rebuilt the Gruner instrument using two variable RF oscillators. We did not put the various bands in it like the Rife Ray #4 or the Beam Ray Laboratory instrument because we first believed it was a heterodyning instrument. Now that we know that Philip Hoyland's Beam Ray Clinical instrument used upper harmonics of the Rife Ray #4 frequencies it doesn't seem necessary that we put the different bands into the instrument. Our instruments will probably be adjusted so we can cover a range from about 1 MHz to about 4 or 5 MHz. Then we will multiply up the Rife Ray #4 frequencies into these ranges, like Philip Hoyland did with the Beam Ray Clinical instrument, and use them with this instrument. Other than John Crane adding the second Hartley oscillator the Gruner schematic was a complete schematic. The pulsing circuit was drawn correctly and worked. It also showed how Dr. Rife produced the damped waveform which pulsed his high RF frequencies in his instruments.

Mr. Peters built his instrument using 805 tubes and I built my instrument using 812A tubes. The AZ-58 Beam Ray Clinical replica RF section was almost exactly the same (809 vacuum tube replaced with the more powerful 812A tube) as the Gruner circuit. The reason I built the Gruner instrument using the 812A tubes is because the 809 is no longer being manufactured. We were also able to replace the modulated pulsing circuit consisting of the 45 & 76 tubes with a single high voltage switching transistor driven with a function generator which outputs the 1330 Hertz pulse frequency. With the use of this transistor we can use any audio frequency we want to use. Using this transistor we believe makes the design better because we can output all the lower audio frequencies including the original audio frequencies used in the AZ-58. My instrument has two Vernier dials which allows me to output two high RF frequencies at the same time like the Rife Ray #4 did. The ray tube is connected to both tank coils instead of having one side of the ray tube go to ground. It is the connecting of the ray tube between the two RF oscillators that makes this design work like the Rife Ray #4 would have done.

Please note: There is a modern instrument built today that is called the "Beam Ray." It works nothing like this original 1930's Beam Ray Laboratory instrument built by Philip Hoyland. We are not saying anything negative about the modern Beam Ray instrument but some people have asked us if these instruments work on the same principles and frequencies and they do not. We have only given this information so people are not confused about these two instruments.

www.rife.org

On page 118 are photos of Mr. Peters instrument which he built. On pages 119 through 121 are the photos of the Gruner instrument that we built. Because John Crane altered the schematic by adding the second oscillator we built this instrument with two oscillators. I built mine with two variable RF oscillators which made it work like the Rife Ray #4 and the Beam Ray Laboratory instrument. For this reason we refer to this instrument, in the photos below, as the Laboratory instrument.

On page 122 is a redrawn schematic of this design. We replaced the smaller vacuum tube of the second Hartley Oscillator with the 809 tube. The layout of the electronic parts of this instrument is very important because of the inherent interference problems that come with RF oscillators. Anyone wanting to build this instrument should have a good understanding of old tube technology. Some parts of this circuit use up to 2000 volts DC with substantial current and can easily kill anyone not experienced in working with this kind of current or voltage. We take no responsibility for anyone who builds this instrument. We recommend that you have professional help.

Chapter Summary: Dr. Gruner's machine was built to be set on one high RF frequency. It had no variable audio oscillator and did not work on the sideband method developed by Philip Hoyland. It had a fixed audio pulsing circuit that modulated the high RF frequency causing a high potential voltage rise in the waveform of the RF frequency being outputted to the ray tube. Because it used only one high RF oscillator it would have been set on Dr. Rife's high RF frequency for the organism Dr. Gruner was working on.

The Beam Ray Laboratory machine was built to work like the Rife Ray #4. It was built to be used in a laboratory for doing M.O.R. work on microorganisms. It would not have been built like the Rife Ray #5 or Beam Ray Clinical instrument which worked on harmonic sidebands. The Beam Ray Clincal instrument would not have worked in a laboratory setting since it devitalized all of the organisms in Dr. Johnson's laboratory due to the many sideband frequencies it would output.

In the next chapter we will now look at the next original Rife Machine replica that was purchased by Aubrey Scoon and his British Rife group.

www.rife.org

Mr. Peters' photos of the rebuilt Beam Ray Laboratory instrument

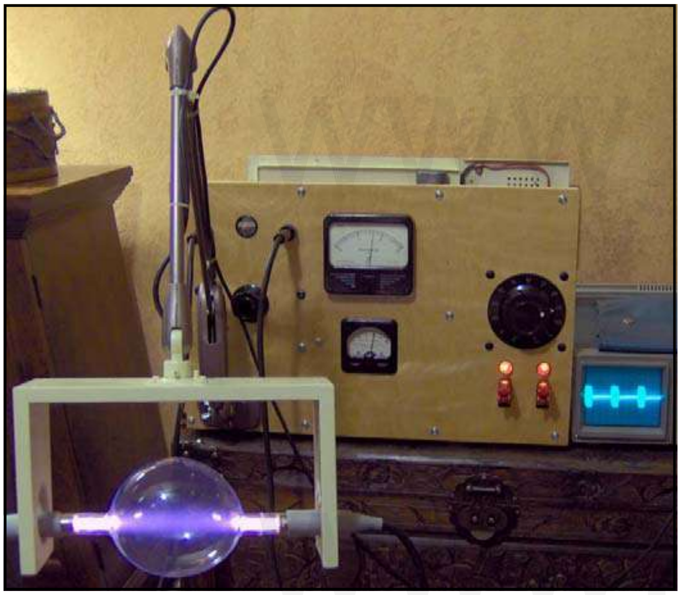

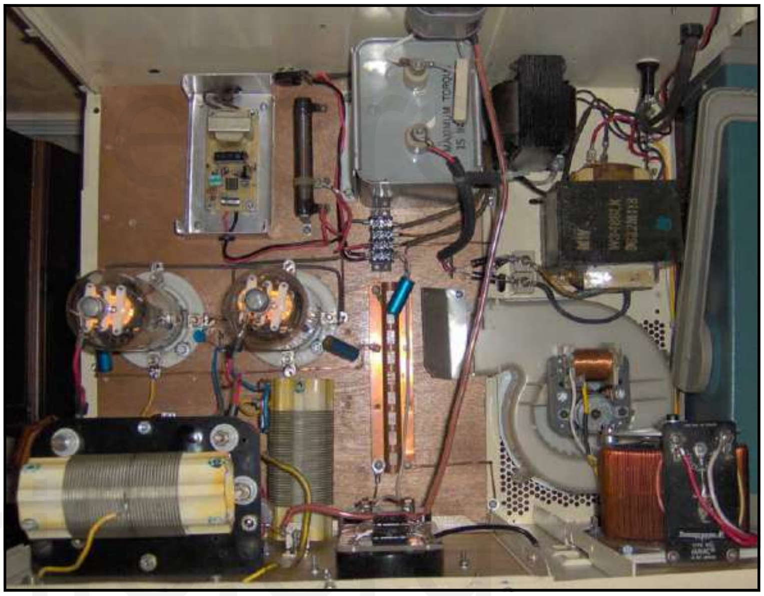

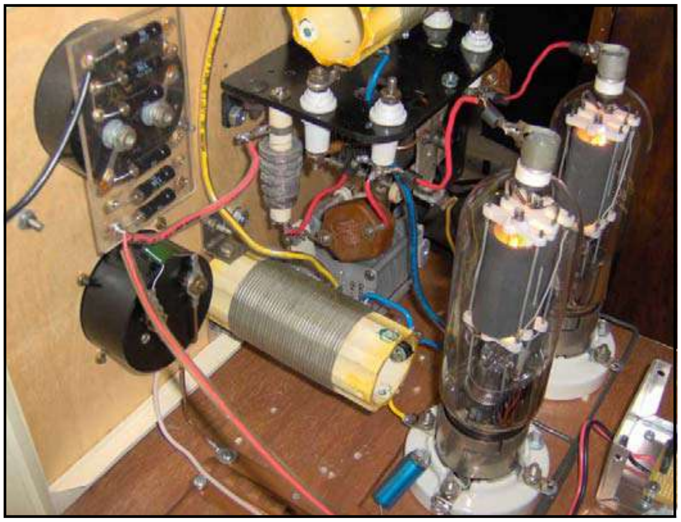

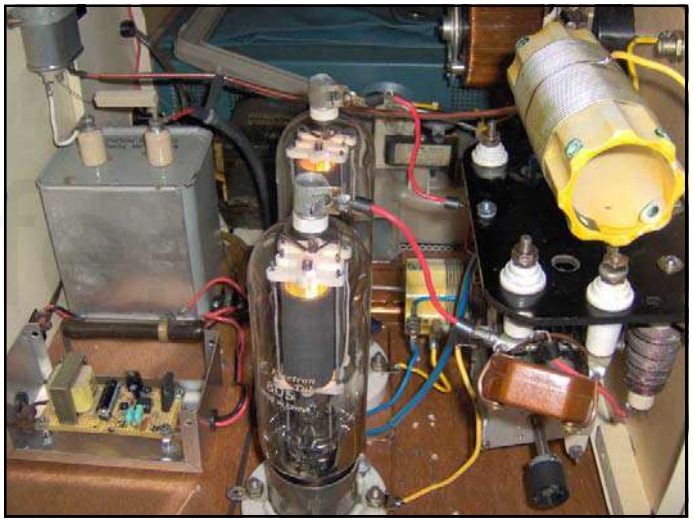

www.rife.org

Photos of the rebuilt Beam Ray Laboratory instrument

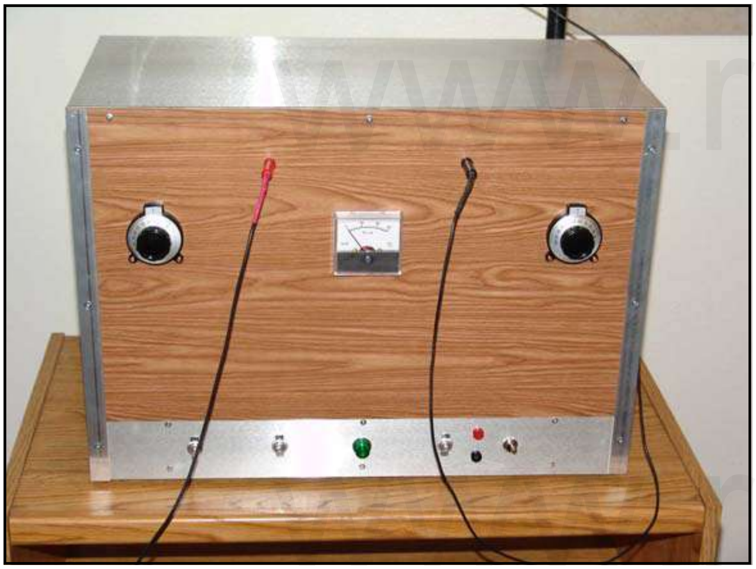

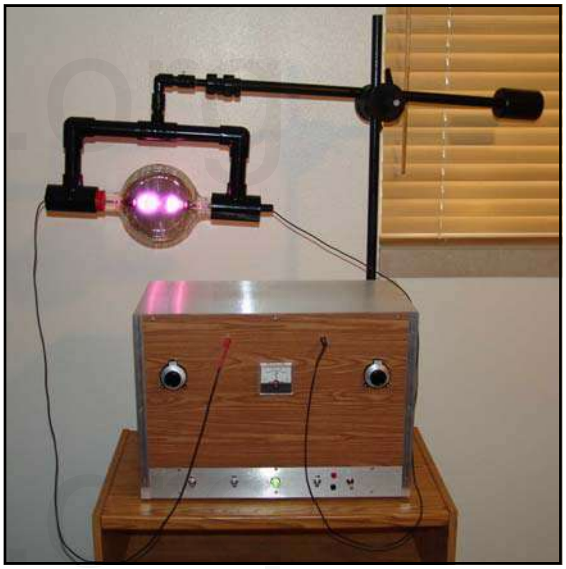

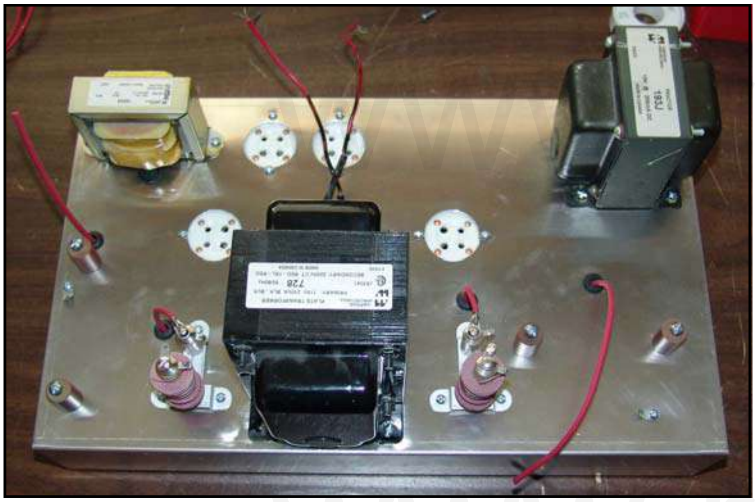

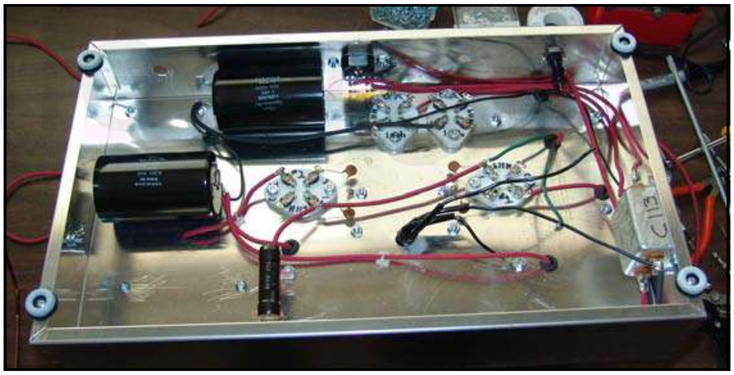

www.rife.org

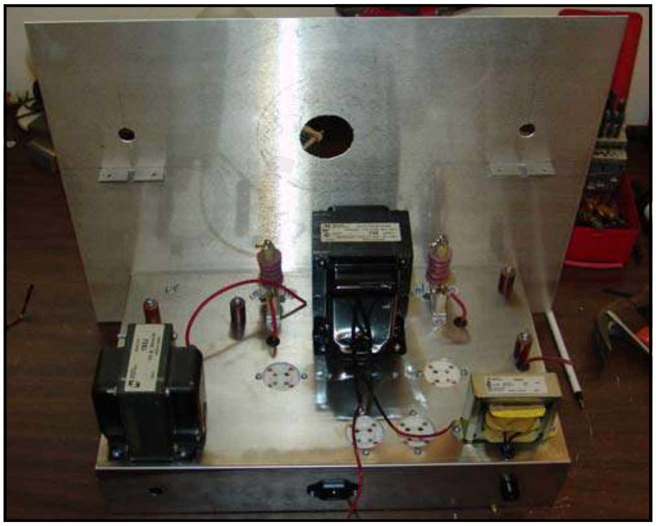

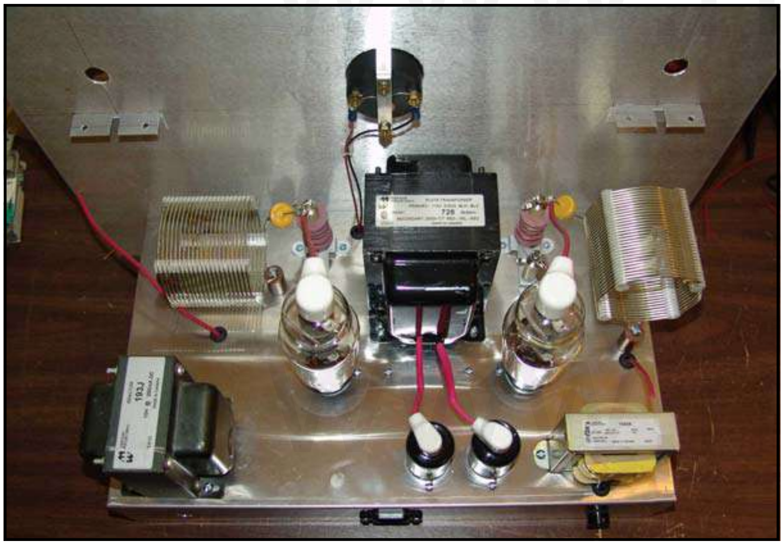

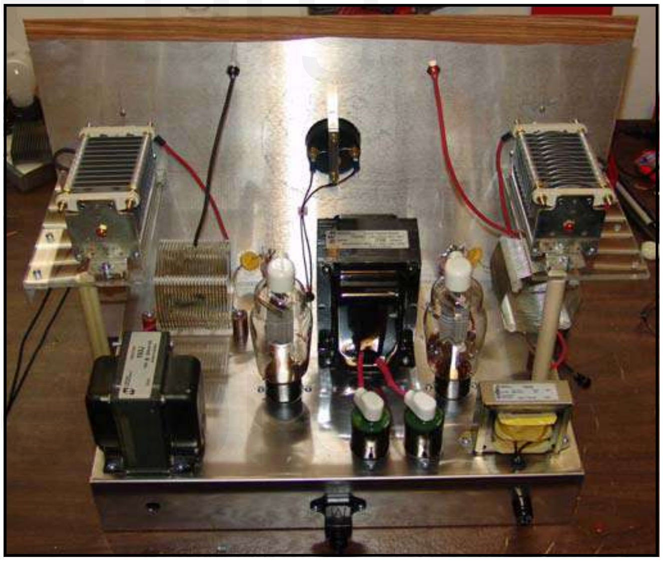

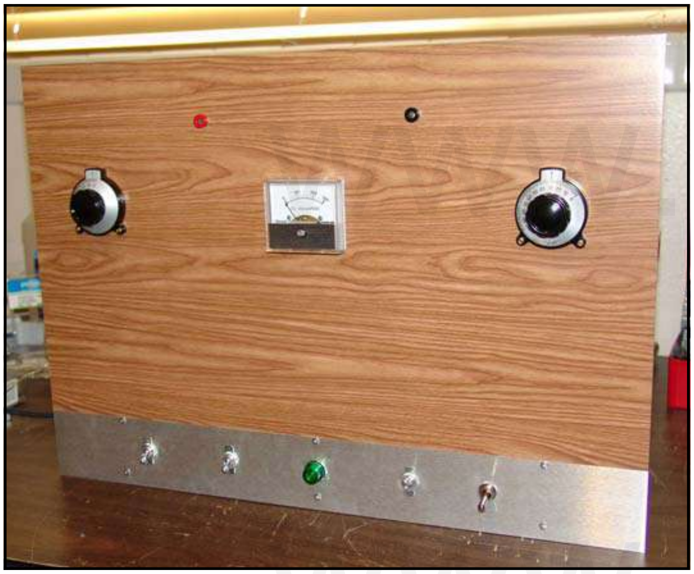

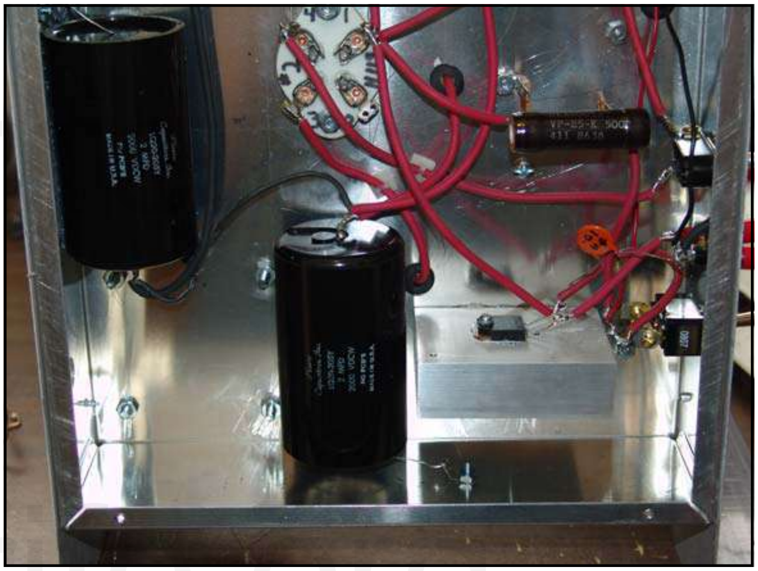

www.rife.org

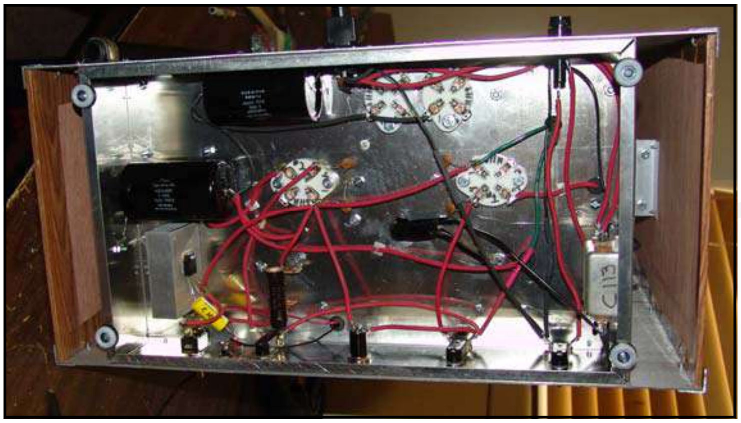

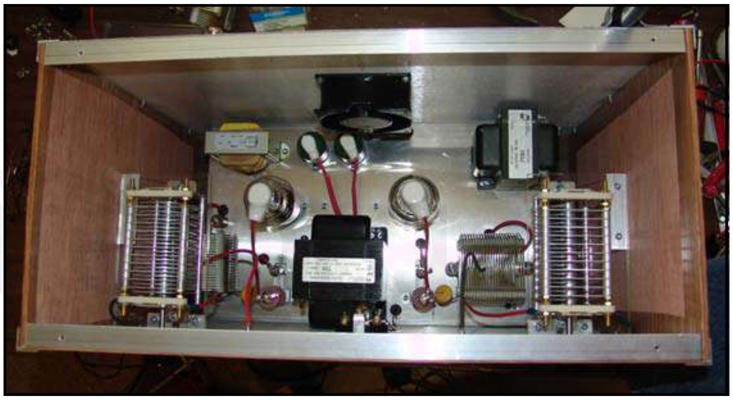

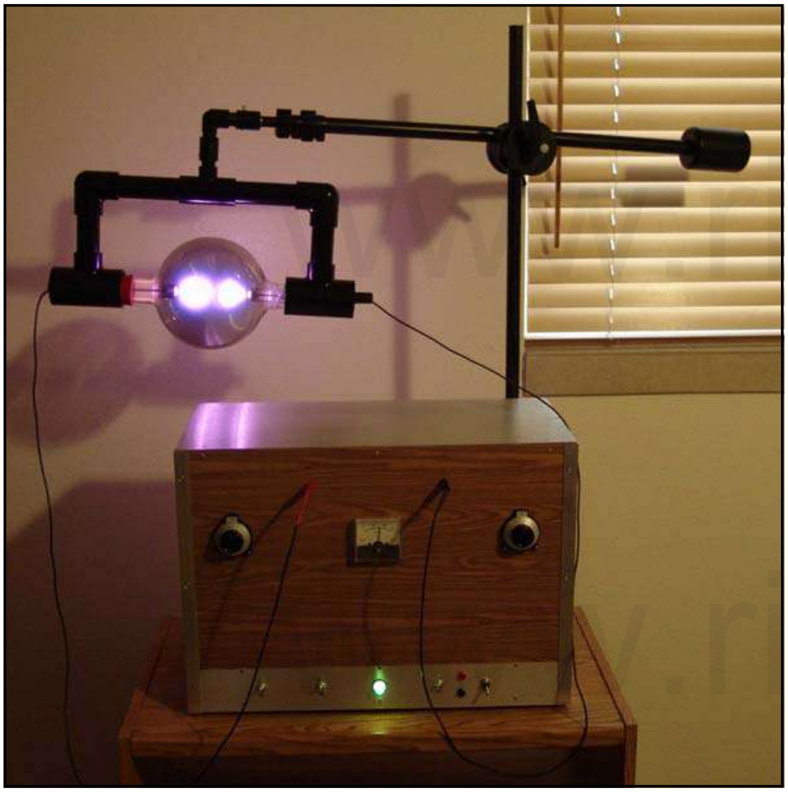

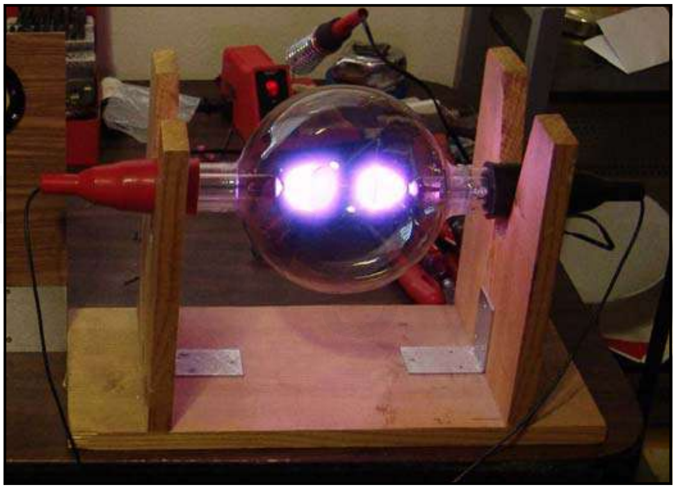

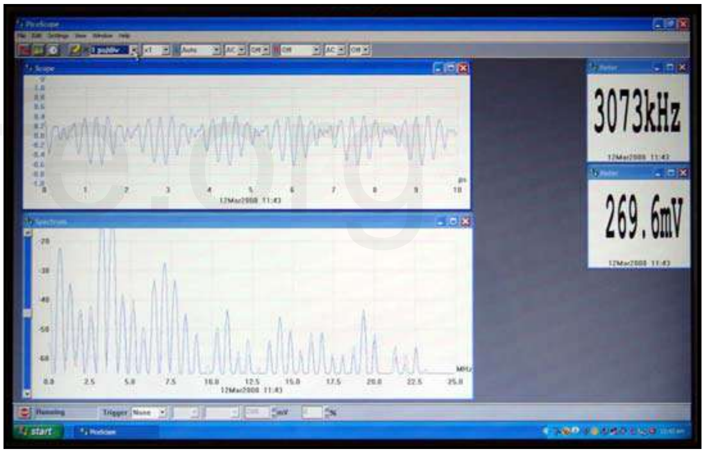
PicoScope
Scope
Spectrum
Meter
Running
Trigger None
start
PicoScope

3073kHz

12Mar2008 11:43

269.6mV

12Mar2008 11:43

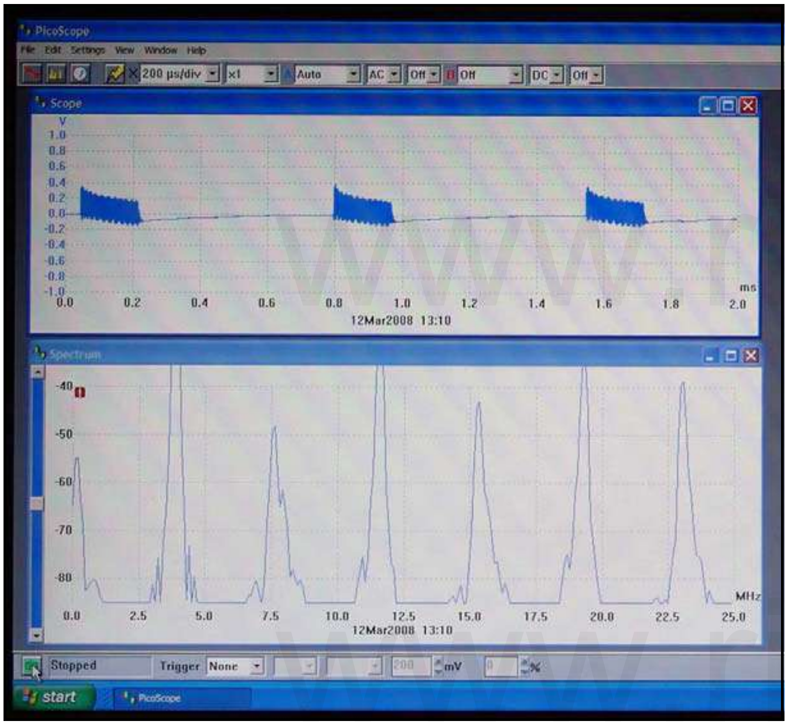
PicoScope
File Edit Settings View Window Help
Scope
V
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
ms
Spectrum
-40
-50
-60
-70
-80
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
MHz
start
PicoScope

200 µs/div x1 Auto AC Off Off DC Off

12Mar2008 13:10

12Mar2008 13:10

Stopped Trigger None 200 mV

www.rife.org

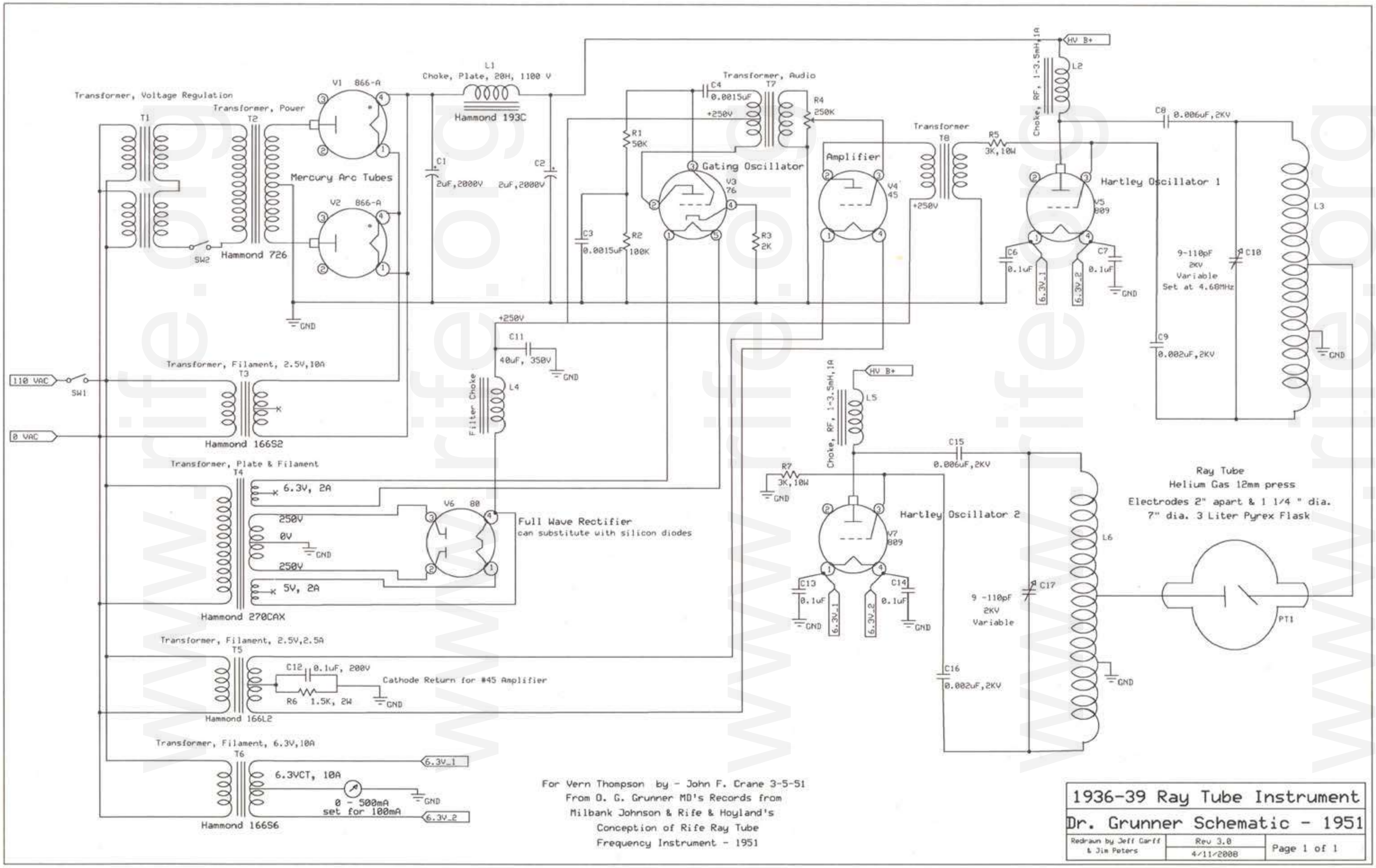
Transformer, Voltage Regulation
T1
Transformer, Power
T2
SW2
Hammond 726
V1 866-A
V2 866-A
Mercury Arc Tubes
GND
L1
Choke, Plate, 20H, 1100 V
Hammond 193C
C1
2uF,2000V
C2
2uF,2000V
HV B+
C4
0.0015uF
Transformer, Audio
T7
R4
250K
+250V
R1
50K
Gating Oscillator
V3
76
C3
0.0015uF
R2
100K
R3
2K
Amplifier
V4
45
Transformer
T8
R5
3K,10W
+250V
Choke, RF, 1-3.5mH,1A
L2
C8 0.006uF,2KV
Hartley Oscillator 1
V5
809
C6
0.1uF
6.3V_1
6.3V_2
C7
0.1uF
GND
9-110pF
2KV
Variable
Set at 4.68MHz
C10
L3
C9
0.002uF,2KV
GND
110 VAC
SW1
0 VAC
Transformer, Filament, 2.5V,10A
T3
Hammond 166S2
+250V
C11
40uF, 350V
GND
Filter Choke
L4
HV B+
Choke, RF, 1-3.5mH,1A
L5
C15
0.006uF,2KV
R7
3K,10W
GND
Transformer, Plate & Filament
T4
6.3V, 2A
250V
0V
GND
250V
5V, 2A
Hammond 270CAX
V6 80
Full Wave Rectifier
can substitute with silicon diodes
Hartley Oscillator 2
V7
809
C13
0.1uF
GND
6.3V_1
6.3V_2
C14
0.1uF
GND
9 -110pF
2KV
Variable
C17
L6
Ray Tube
Helium Gas 12mm press
Electrodes 2" apart & 1 1/4 " dia.
7" dia. 3 Liter Pyrex Flask
PT1
GND
C16
0.002uF,2KV
Transformer, Filament, 2.5V,2.5A
T5
C12
0.1uF, 200V
R6 1.5K, 2W
Cathode Return for #45 Amplifier
GND
Hammond 166L2
Transformer, Filament, 6.3V,10A
T6
6.3VCT, 10A
0 - 500mA
set for 100mA
GND
6.3V_1
6.3V_2
Hammond 166S6

For Vern Thompson by - John F. Crane 3-5-51
From O. G. Grunner MD's Records from
Milbank Johnson & Rife & Hoyland's
Conception of Rife Ray Tube
Frequency Instrument - 1951

1936-39 Ray Tube Instrument

Dr. Grunner Schematic - 1951

Redrawn by Jeff Garff & Jim Peters	Rev 3.0 4/11/2008	Page 1 of 1

www.rife.org

Chapter #11

Aubrey Scoon's Beam Ray replica Rife Machine Re-evaluation

1. Used a ray tube.
2. Used a fixed 3.30 MHz RF carrier frequency.
3. Modulated sine wave audio frequencies onto a sine wave RF carrier frequency.
4. Power usage was about 460 watts. Output to the ray tube about 40 watts.

In 2009 Aubrey Scoon passed away. His web site remained up on the web until the domain name expired last year. Since his web site is no long available and is an important part of this chapter, including this report, we have included all of his information pertaining to his machine on the www.rifevideos.com web site. To fully understand this chapter we suggest that you read Aubrey's information about his Beam Ray replica instrument at this web address:

http://www.rifevideos.com/aubrey_scoons_1939_rife_beam_ray.html

Aubrey Scoon's 1950's Beam Ray Clinical replica Rife Machine was originally mistaken for an g original 1938-1939 Rife Ray #5 or Beam Ray Clinical instrument built by Philip Hoyland and the Beam Ray Corporation. We now know that it was not an original Beam Ray instrument built by Beam Ray Corporation but it is an exact replica of that instrument. It was originally for sale on www.rife.org. John Bedini and a group of men who had worked with John Crane for a year and a half considered purchasing it. After careful examination of the information available, at that time, they concluded that this instrument was not an original Beam Ray instrument built by the original 1938-1939 Beam Ray Corporation. They concluded that it was built later, perhaps, in the 1940's by Dr. Rife's engineer, Verne Thompson. Verne Thompson eventually replaced Philip Hoyland as Dr. Rife's new engineer and he was building instruments during the 1940's and 1950's for doctors that wanted Dr. Rife's machines.

The reason we are re-evaluating Aubrey Scoon's Rife Ray #5 or Beam Ray Clinical replica instrument is this instrument is the second most important instrument that we have analyzed. When it was first analyzed by Aubrey Scoon's team they did not take their original evaluation of this instrument far enough. Had they used a spectrum analyzer there is a possibility that they may have figured out

www.rife.org

how this instrument really worked, but they did not. Once we determined how the "Original Rife Ray #5 or Beam Ray Clinical" instrument worked we were able to re-evaluate Aubrey Scoon's Beam Ray replica and show that it worked on the same principles and frequencies as the original Beam Ray Rife Machine obtained from Dr. Low. All of Dr. Rife's instruments, from the first to the last, worked on the same principles and the same high RF frequencies, or higher harmonics of Dr. Rife's original frequencies. The method of generating the frequencies may have changed but all the frequencies used in Dr. Rife's machines were based on the original frequencies he found that would eliminate, deactivate or devitalize those organisms he worked on in his laboratory.

For a better understanding of Aubrey Scoon's Beam Ray Clinical Rife Machine we need to give the history of it. Aubrey Scoon and a group of men from England purchased the above instrument believing it was an original Beam Ray Clinical instrument built by the 1938-1939 Beam Ray Corporation. The original Beam Ray Clinical instrument was known to be working on harmonics. Without really knowing it, they purchased a replica of the Beam Ray Clinical instrument. At that time no original instrument had been located. Because of this and the fact that we had no absolute concrete evidence that Dr. Rife's engineer, Philip Hoyland, built this style of audio frequency instrument it was not fully accepted as a genuine Rife instrument. Not even the later 1953 AZ-58 Beam Ray Clinical replica was accepted as a genuine instrument either.

A few years ago, before Aubrey passed away, I was communicating with him about this instrument. In the course of our communications he told me that he believed he had used the wrong main output vacuum tube in the instrument when they worked on it. Because they believed they had used the wrong vacuum tube (812a vacuum tube) the carrier frequency had parasitic oscillations which created harmonics. He said that when they discovered this mistake they put what they believed was the proper tube (809 vacuum tube) in and most of harmonics from the parasitic oscillations were gone. Aubrey Scoon mentions the change of this tube (809) on his web site. But they did not change the photos of the waveforms so we do not know how much the waveform really changed. This much we do know all of the photos of the waveforms on his web site are of an instrument that has parasitic oscillations. Using the wrong tube was a simple mistake that anyone could make but it led to a great deal of confusion causing many to believe, including myself, that this instrument was, an original Beam Ray Clinical instrument because of those parasitic oscillations. We believed these parasitic oscillations created the harmonics which the original Beam Ray Clinical instrument was supposed to be working on.

Both Jim Berger and I separately built Aubrey Scoon’s instrument with the correct vacuum tubes (812a) and found using an oscilloscope that the RF output was clean of any parasitic oscillations. In our tests the 812a vacuum tube produced no harmonics like those seen in Aubrey Scoon’s photos. At this time it confirmed to us that if the circuit was working correctly there would be no parasitic oscillations or harmonics from parasitic oscillations. The building of this instrument also showed that it didn't have the reported harmonics which the genuine Beam Ray Clinical instrument was supposed to have. Little did we know, at that time, that the harmonic concept we were looking for was there but we didn’t really understand how the instrument was supposed to work. This wrongly convinced both Jim Berger and I that Aubrey Scoon's instrument was not a genuine Beam Ray instrument. Nevertheless, this mistake does not change the fact that we now know that this instrument is a genuine Beam Ray machine replica.

At a later date when both the 809 and 812a vacuum tubes were tested in the circuit neither tube created the same kind of waveforms as Aubrey Scoon's machine was producing. This proved to us that the parasitic oscillations are due to some other problem in the RF circuit of his machine. It was also determined that the 812a vacuum tube was the correct tube that should be used, not the 809 vacuum tube. The 812a vacuum produces the correct carrier frequency which should be used with the audio frequencies Aubrey Scoon's instrument used.

www.rife.org

The original Beam Ray Clinical instrument built by Beam Ray Corporation that we obtained from Dr. Low did not have any frequency list that came with it showing us what frequency band or dial settings should be used for the various microorganisms. However, Aubrey Scoon's Beam Ray replica instrument built by Verne Thompson came with a list of frequencies which the doctor used on the various organisms. It is this list of frequencies that makes this instrument so important.

We know from the previous documents that we have read in this report that Dr. Rife had Verne Thompson rebuilding and repairing these instruments. We also read that Dr. Rife had Verne Thompson rebuilt Dr. Yale's machine in 1940. Because Verne Thompson became Dr. Rife's engineer he would periodically repaired Dr. Couche's and Dr. Tulley's machines. We also know that Verne Thompson was making copies of Dr. Couche's Beam Ray Clinical instrument for other doctors. With this understanding we know that Verne Thompson was the one who would have written down these audio frequencies that were used in Aubrey Scoon's Beam Ray Clinical instrument. Also with this understanding we know that the same audio frequencies used in Aubrey Scoon's Beam Ray Clinical instrument were used in Dr. Couche's Beam Ray Clinical instrument. This information is very important. The importance of this information will become very clear as we continue to look at this instrument and the 1953 AZ-58 Beam Ray Clinical Replica instrument.

Since this paper was updated on 9/20/2010 we decided to test a 3.30 MHz RF carrier frequency believing that it could have been the correct RF carrier frequency. When Aubrey Scoon first tested the instrument with the 812a vacuum tube (not the 809 vacuum tube) he listed 3.33 MHz as the RF carrier frequency on his website. Knowing how parasitic oscillations in an RF carrier frequency could easily have shifted the carrier frequency 30,000 Hertz, it was thought that a 3.30 MHz RF carrier frequency would have been a more logical frequency to use. This assumption proved to be correct. In the summer of 2011 we obtained more of John Marsh's documents from his nurse. In one of these documents dated November 20, 1967 John Marsh stated the following:

MARSH: *John Crane's, a simple oscillator, which can be obtained easily for about $33.00. It produces a fuzzy band. He experimented with hooking up wires with an ordinary radio speaker and produced a different musical note for each frequency. The large instrument* [Beam Ray Clinical replica instrument] *is a RF frequency generator which carriers wave oscillations at 3300 kilocycles (3.3 Megahertz) on the marine band. Pre-auditory sound waves.* (John Marsh 1967 document about Dr. Robert P. Stafford).

This letter of John Marsh's confirms that the RF carrier frequency in the original Rife Ray #5 or Beam Ray Clinical instrument, which Dr. Rife, John Crane and John Marsh made a replica of in 1953, was set at 3.30 MHz. This also confirms that the 812a vacuum tube is the correct tube. It also confirms that that the original reading done by Aubrey Scoon of the RF carrier frequency being about 3.30 (3.33) was a correct reading. It was only slightly off due to the parasitic oscillations.

As mentioned before, Aubrey Scoon's instrument had four bands. These four bands were mentioned by Philip Hoyland in the Beam Ray Trial. He stated that they covered all the frequencies for the various organisms this instrument treated. The frequency range of each of the four bands is as follows:

Band 1: 20 Hertz to 200 Hertz.

Band 2: 200 Hertz to 2000 Hertz.

Band 3: 2000 Hertz to 20,000 Hertz.

Band 4: 20,000 Hertz to 200,000 Hertz.

www.rife.org

Our original Beam Ray Clinical instrument, along with Aubrey Scoon’s Beam Ray replica instrument design, predates John Crane and John Marsh and this reveals that the audio frequencies came from Philip Hoyland not John Crane or John Marsh. This Aubrey Scoon Beam Ray replica instrument also shows a connection to the original 1938-1939 Beam Ray Clinical Rife Machine and the audio frequencies that were used in the 1953 AZ-58. This AZ-58 instrument will be discussed in greater detail later in this report. The 1953 AZ-58 Beam Ray replica instrument used almost the same frequencies as Aubrey Scoon’s instrument except Crane and Marsh divided them down by a factor of 10 times and used these lower audio frequencies in the AZ-58.

Below is a chart that has the audio sideband frequencies that were used in Aubrey Scoon's Beam Ray Clinical replica instrument. It was these frequencies that John Crane and John Marsh lowered to get the audio frequencies they used in the 1953 AZ-58. We will reconcile these frequencies in this chart to Dr. Rife's original high RF frequencies that were used in the Rife Ray #3 and Rife Ray #4 Rife Machines in this section of this report. Before we do this we need to compare Aubrey Scoon's Beam Ray Clinical replica to the original Beam Ray Clinical Rife Machine which we obtained from Dr. Low.

The list of audio frequencies, shown below, of Aubrey Scoon’s Beam Ray replica sine wave audio frequencies must be used with a 3.30 MHz RF carrier frequency. Testing of this 3.30 MHz RF carrier frequency showed that this was the correct carrier frequency for this instrument. It was this testing with the 3.30 MHz RF carrier frequency in combination with the audio frequencies which produced the correct sideband frequencies that hit the Rife Ray #3 and Rife Ray #4 higher harmonic frequencies.

What we need to point out here is this **IMPORTANT** fact which came from the analyzing of the original Beam Ray Clinical instrument. This fact also applies to Aubrey Scoon's Beam Ray replica instrument. Neither the 3.30 MHz RF carrier frequency nor the audio frequencies will do anything by themselves. But when the 3.30 MHz RF harmonic carrier frequency and the audio frequencies are combined together they will produce many sideband frequencies. And one of these sideband frequencies will line up with the true Rife M.O.R. frequency and devitalize or render harmless the harmful microorganism. To re-emphasize this so that no one misunderstands. If you just use the audio frequencies by themselves you will get nothing. If you use the 3.30 MHz RF carrier frequency without the audio frequencies you will get nothing. The audio frequencies used in Aubrey Scoon's instrument must have the RF carrier frequency of 3.30 MHz or they are useless. This is the reason the 1953 Beam Ray Clinical instrument called the AZ-58 did not work properly. In the chart, shown below, are the audio frequencies used in Aubrey Scoon's instrument. It is these audio frequencies when combined with the 3.30 MHz RF carrier frequency that will produce Dr. Rife's higher harmonic M.O.R. frequencies. It must be understood that these audio frequencies were not meant to treat these organisms.

Aubrey Scoon’s Beam Ray Clinical Replica Audio Sideband Frequencies

B or E Coli Rod	8,020 Hz	Syphilis or Treponema	6,600 Hz
B or E Coli Virus	17,220 Hz	Tetanus	1,200 Hz
BX Virus Carcinoma	21,275 Hz	Tuberculosis Rod	8,300 Hz
BY Sarcoma	20,080 Hz	Tuberculosis Virus	16,000 Hz
Pneumonia or Spinal Meningitis	7,660 Hz	Typhoid Rod	6,900 Hz
Staphylococcus	7,270 Hz	Typhoid Virus	18,620 Hz
Streptococcus	8,450 Hz	Worms	2,400 Hz
Streptothrix	7,870 Hz		

www.rife.org

The two photos below show Aubrey Scoon's RF section. The electronic components are almost identical to our original Beam Ray Clinical instrument. It used the same 866 rectifier tubes (two used) along with the 812a vacuum tube (one used).

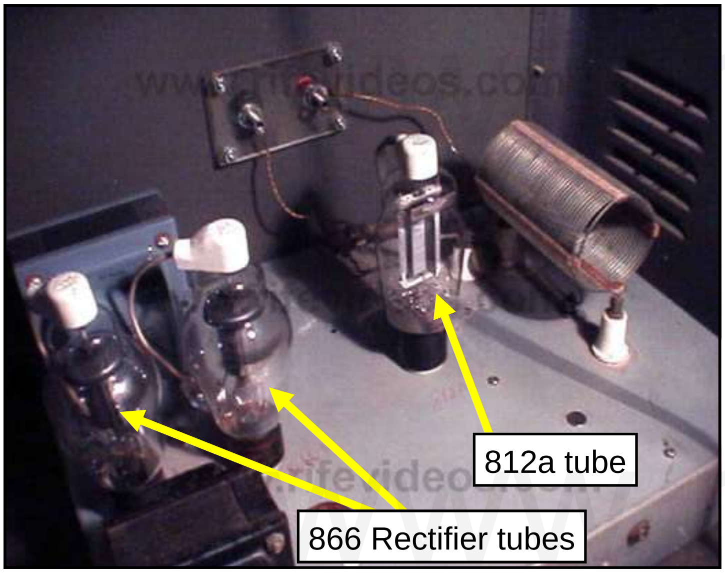

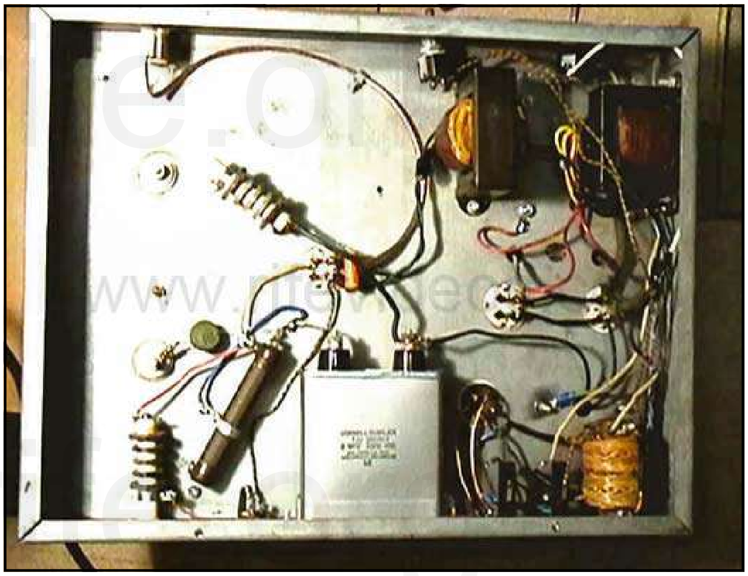

Like the original Beam Ray Clinical instrument which used a Hewlett-Packard audio oscillator, Aubrey Scoon's Beam Ray replica instrument also used a Hewlett-Packard audio oscillator design. By design we mean that this oscillator used an incandescent light bulb in the circuit (patented by Hewlett-Packard) which stabilized the audio oscillator frequencies and waveform which it output. In the next two photos, shown below, is Aubrey Scoon's Beam Ray Rife Machine audio oscillator. In the photo, below on the right, you can see the small incandescent light bulb.

In 1938 when Beam Ray Corporation built this style of instrument the Hewlett-Packard Wein Bridge audio oscillator was not invented yet. Since this newer Hewlett-Packard audio oscillator patent was not filed until July of 1939 and the Beam Ray Corporation was at this time in a court battle it is only logical that the original Beam Ray instrument did not use Hewlett-Packard's new design. It was invented in early 1939 and a patent was filed July 11, 1939. The patent was granted on Jan 6, 1942.

In 1938 and 1939 the original Beam Ray Clinical instrument would have had an RC (resistor capacitor) type of audio oscillator. These RC audio oscillators were known to be very unstable and it was replaced in the original Beam Ray instrument that we have with the newer Hewlett-Packard design. The audio oscillator section of the original Beam Ray Clinical instrument that we have has many extra

www.rife.org

holes in the chassis and shows that modifications were made to update this instrument to the newer Hewlett-Packard design.

Aubrey Scoon in his evaluation of his Beam Ray replica instrument believed that his instrument was an original machine built by the 1938-1939 Beam Ray Corporation. He was correct that it was an original design but he was wrong about the year it was built. John Bedini was correct in his belief that it was not an original machine built by the 1938-1939 Beam Ray Corporation. The fact that the Hewlett-Packard design patent was not issued until 1942 should have indicated to Aubrey Scoon that his instrument was built later than 1939. Rather than accept that the patent had not been filed until July of 1939 and issue until 1942 he believe that Hewlett-Packard had somehow allowed Beam Ray Corporation the right to use their design before they even filed a patent on it. Of course this does not make any sense since they could have lost any patent rights if they did this. Just this information alone should have indicated that this machine was not an original machine built in 1938-1939.

Aubrey Scoon’s audio oscillator update

Roger Blain, who has had a great deal of interest in this technology, uncovered some very important information about the manufacture date of the audio oscillator design that was used in Aubrey Scoon's instrument. If you look at the above two photos which show both the underside and top view of the audio section of Aubrey's instrument you will notice that there are no signs of any retro-fitting of the audio oscillator to a newer design. If it had been built in 1938 or 1939 then it would have had the original (RC) resistor capacitor design which would have been replace. It also would have showed obvious signs of the necessary changes needed to update the old style (RC) audio oscillator to the newer Hewlett-Packard design which it now has. The next photo, shown below, shows the underside of the audio oscillator of the original 1938-1939 Beam Ray machine obtained from Dr. Low.

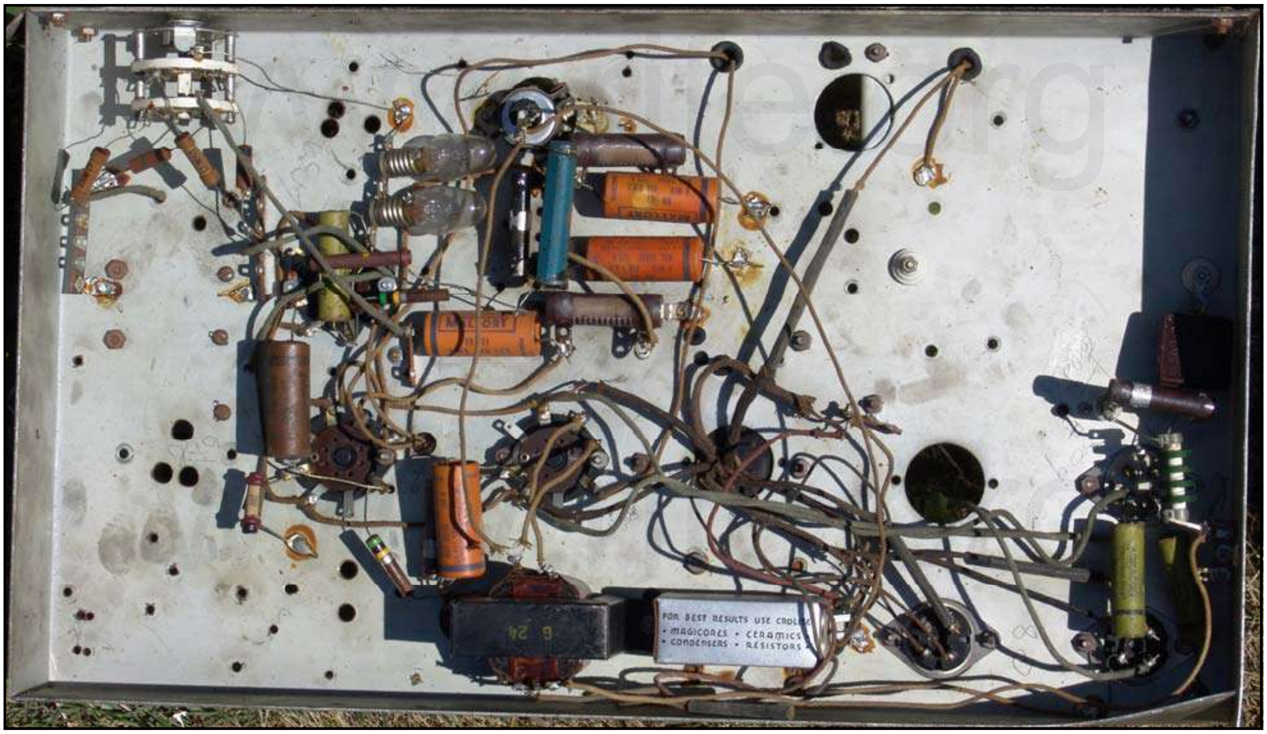

You will notice many small holes including two larger old vacuum tube holes (empty and without any vacuum tube sockets) in its chassis. This is what we would expect to see if an instrument's audio section had been replaced with the newer more accurate Hewlett-Packard design. The lack of any modifications to the audio section of Aubrey Scoon's instrument indicates that the instrument still has the original audio oscillator it was built with. Just this fact alone puts the building of his instrument after the original Beam Ray Corporation shut down in 1939.

www.rife.org

Roger Blain decided that if he could determine which design of audio oscillator was used in Aubrey Scoon's instrument then the correct dating of that instrument could be determined. In his discovery work he found that the audio oscillator that was used in the construction of this instrument was the EICO 377. Aubrey Scoon's instrument audio oscillator frequency range and the circuit design are identical to the EICO 377 including the component values, wiring and layout. The only thing that was left out was the double pole double throw switch that allowed for switching between square and sine wave. This new information leaves no doubts as to the approximate date Aubrey Scoon's instrument was built. This information also firmly put its construction date at 1952 or later instead of the 1939 date Aubrey Scoon gives in his report. The next three photos are of the EICO 377.

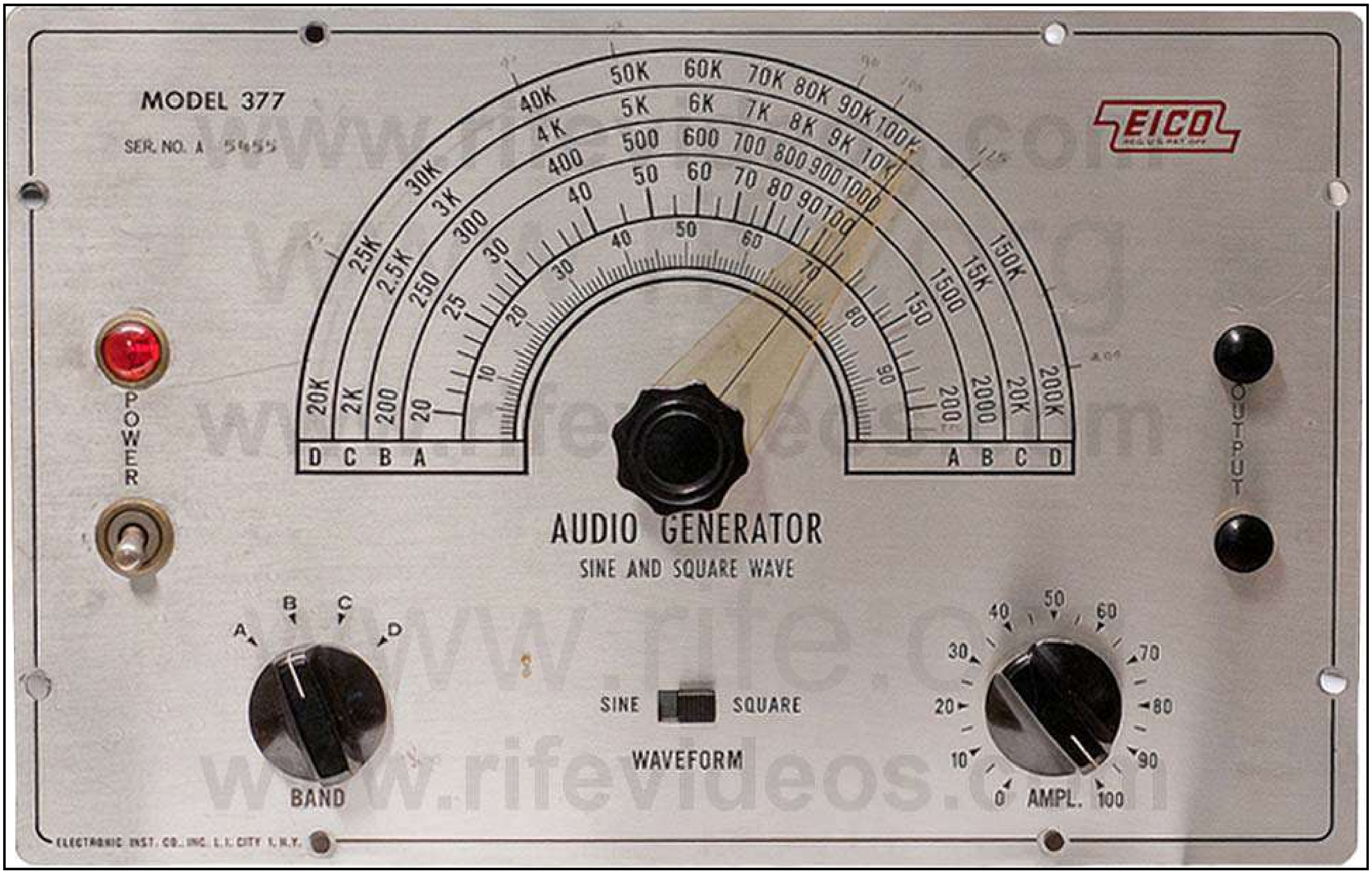
MODEL 377
SER. NO. A
EICO
AUDIO GENERATOR
SINE AND SQUARE WAVE
POWER
OUTPUT
BAND
A B C D
SINE
SQUARE
WAVEFORM
AMPL.
ELECTRONIC INST. CO., INC. L.I. CITY 1, N.Y.

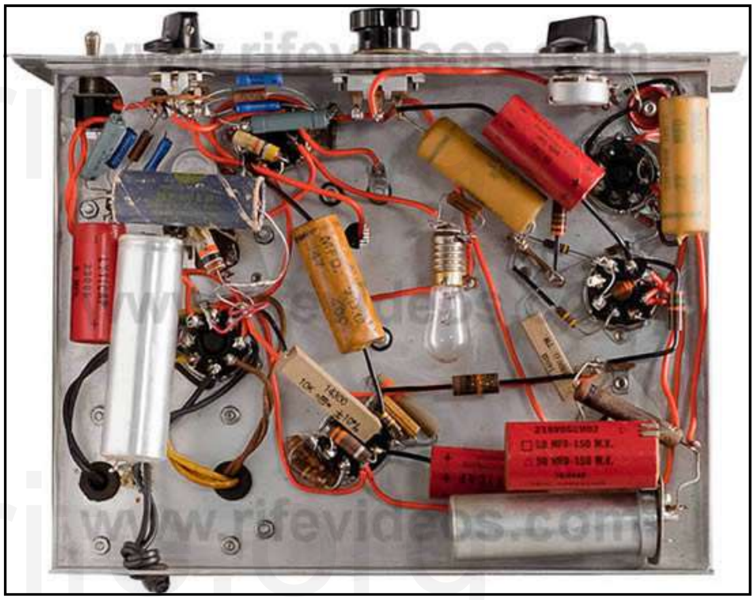

Below are three more photos. The first photo, below on the left, again shows the top view of the EICO 377 component layout. The second photo, below on the right, shows the layout of Aubrey Scoon's audio oscillator. Looking at both of these two photos you can see that the layout is identical. The third photo is a slightly different view of the second photo.

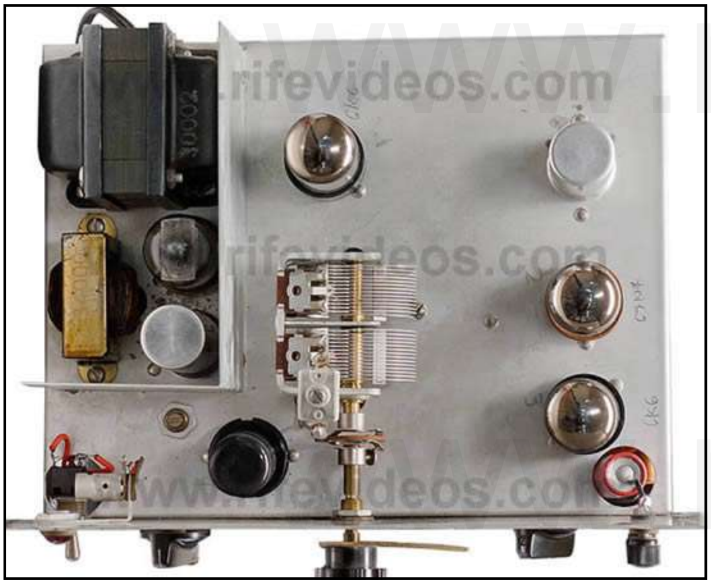

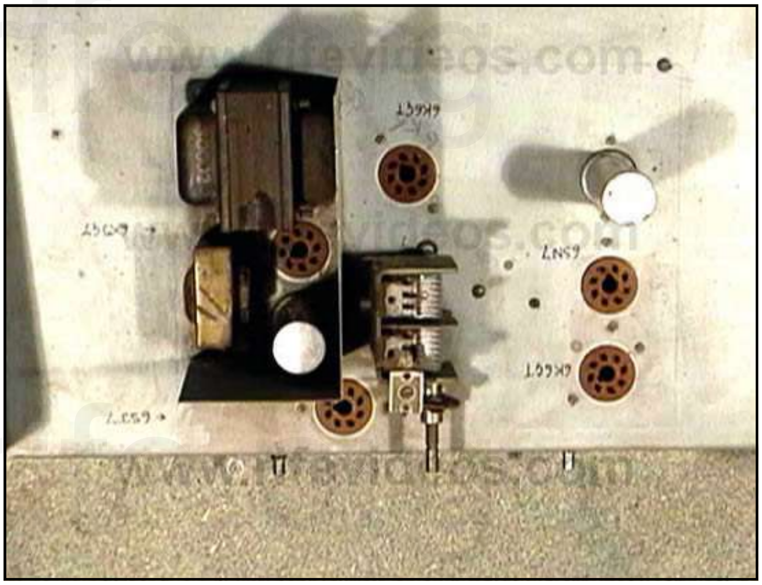

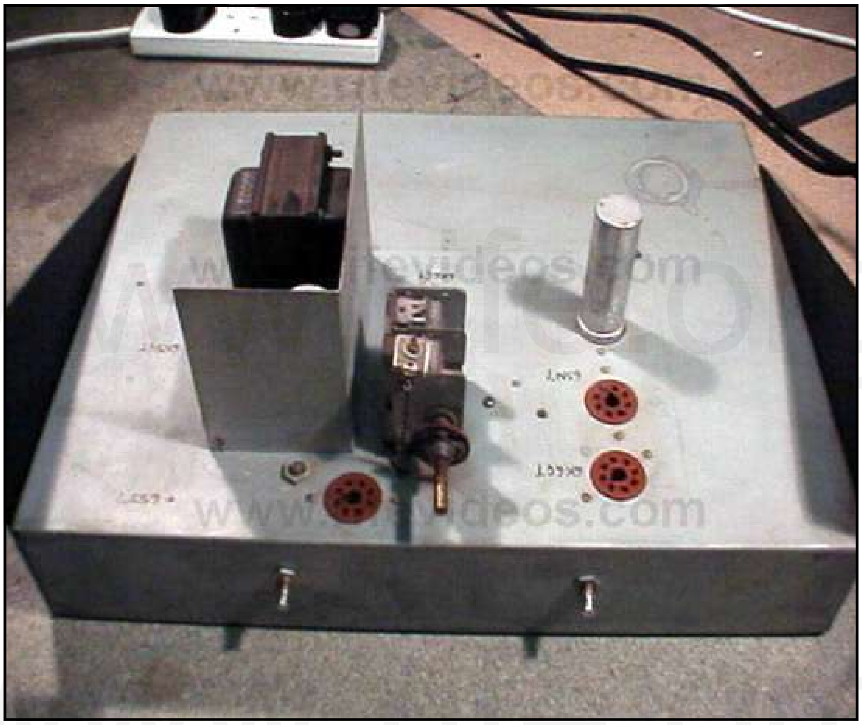

Once we knew that the EICO 377 was the audio oscillator that was used by Verne Thompson in the building of Aubrey Scoon's instrument we went through John Marsh's papers to see if they happened to mention the EICO 377 instrument. We found in his papers a document which shows that John Crane and John Marsh were using the EICO 377 as a pad instrument. This new style of pad instrument used aluminum discs to come in contact with the body of the user instead of using a plasma tube. Below is a quote found in that document:

"The device consists of: 1. An audio oscillator. These are produced by various firms such as Heath Company of Benton Harbor, Mich.; Electronic Instrument Co. of Long Island City, New York; R.C.A. ; General Radio; Knight Co.; and others as well as our own which was a Hartley oscillator initially.

www.rife.org

The model submitted herein is Model 377 EICO which is manufactured and sold by them as a commercial item." (Rife Virus Microscope Institute - EICO 377 document).

It really shouldn't be a surprise that Verne Thompson used the EICO 377 in his building of Aubrey Scoon's original Beam Ray Clinical replica design since he was using it for their pad instrument. Since the EICO 377 was first built in 1952 this dates the building of Aubrey Scoon's instrument to 1952 or later. This is also when the 812a vacuum tube was used by Verne Thompson in the newer design called the AZ-58. The AZ-58, built in 1953, was also a Beam Ray Clinical replica design. This explains why the correct vacuum tube for Aubrey Scoon's machine was not the 1930's 809 vacuum tube but the 1950's 812a tube.

In the 1940's and 1950's Verne Thompson was still building Beam Ray replica designs for those doctors who wanted to own an instrument. It is apparent that he was building instruments for anyone that wanted one and was not exclusive to John Crane or John Marsh since Dr. Rife and Verne Thompson did not meet John Crane until 1950. The Rife documents show that Rife, Crane and Marsh had Verne Thompson build their new machine which they called the AZ-58 in 1953. It used lower audio frequencies (original 1930's audio frequencies divided by a factor of 10 times) than what the original Beam Ray Clinical design used. Even though Verne Thompson was building the AZ-58 he was still building the original 1930's design which used the higher audio frequencies. The fact that Aubrey Scoon's instrument was built in 1952, or later, proves that both designs were being built at the same time in the 1950’s. The 1953 AZ-58 was using the lower audio frequencies (from 120 Hertz to 2128 Hertz) and Aubrey Scoon's original Beam Ray replica was using the higher audio frequencies (1200 Hertz to 21275 Hertz).

Another discovery made by Roger Blain is also important. It is in regards to the Logo emblem that was on the case of Aubrey Scoon's instrument. Aubrey believed that the "B" Logo found on his instrument must have been the original 1939-1939 Beam Ray Corporation emblem. Below is a photo of this emblem. This was actually the Logo of the company that built the instrument case. The company was Bud Industries. This company is still in business today and are building encloses for electronic devices. Their web site is http://www.budind.com. This discovery, by Roger Blain, also indicates that Aubrey Scoon just didn't do enough investigative work or he would have also discovered that this logo was not Beam Ray Corporations logo.

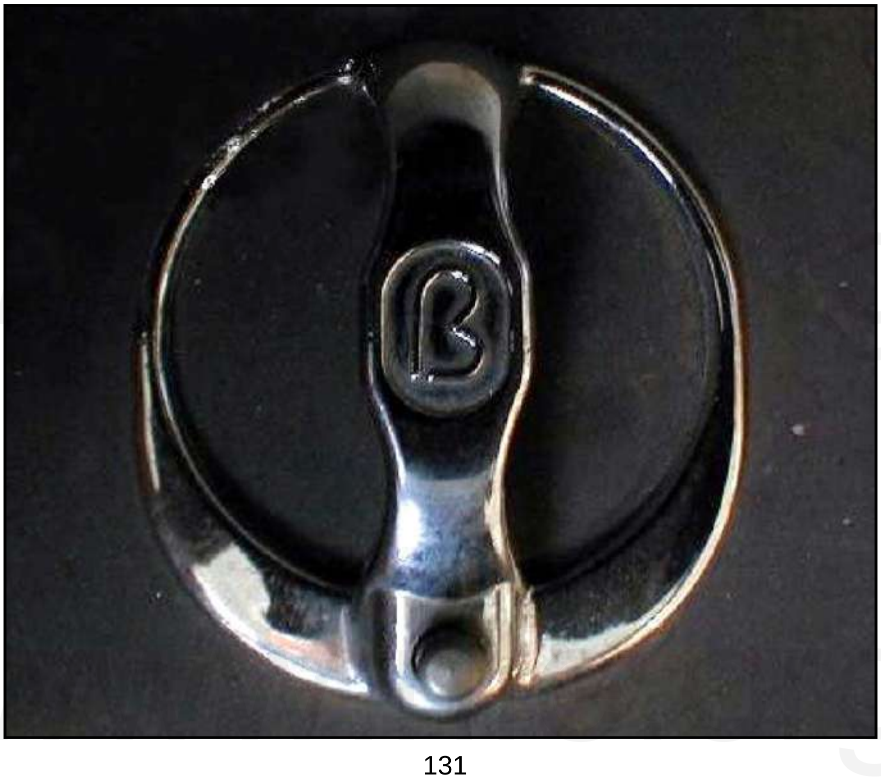

The information in the next two paragraphs should actually be in Chapter 12. But we are also including it here for those who already have a good understanding of the audio frequencies used in this type of instrument. The reason we are including this information here is some people have asked whether the low audio frequencies (120 Hertz to 2128 Hertz) used in the 1953 AZ-58 are actually the original audio frequencies used in the 1938-1939 Beam Ray machine rather than the high audio frequencies (1200 Hertz to 21275 Hertz) used in Aubrey Scoon's Beam Ray replica instrument. This is a good question and it can easily be answered with certainty. The answer is in the math which produces the correct sidebands for each organism. Only the high audio frequencies (1200, 2400, 6600, 6900, 7660, 7270, 7870, 8300, 8450, 8020, 16000, 17220, 18620, 20080, 21275) will produce the correct sideband frequencies which will produce the higher harmonic frequencies from Dr. Rife's original frequencies. Only six of the low audio frequencies (120, 660, 727, 1862, 2008, 2127-2128 or 2127.5 Hertz as given by John Crane) used in the 1953 AZ-58 when multiplied by a factor 10 times give the exact same high frequency used in Aubrey Scoon's instrument. But the other seven (712, 784, 776, 800, 803, 880 and 1552 Hertz) when multiplied by a factor of ten times will not give the correct high audio frequency. These facts reveal which frequency list came first. Since we know that Philip Hoyland designed this Beam Ray Clinical machine and hid the method of using sideband frequencies to produce higher harmonics of Dr. Rife's M.O.R. frequencies then only the list (Aubrey Scoon's list) that will produce Dr. Rife's higher harmonic frequencies could be the original list.

We will point out a few more facts. From the Beam Ray Trial we learned that no one but Philip Hoyland understood how the instrument worked. Not even Dr. Rife, Verne Thompson, John Crane or John Marsh or anyone else understood that the RF carrier frequency had to be matched to the audio frequencies in order to produce the sideband frequencies that would hit the higher harmonic frequencies of Dr. Rife's original frequencies. John Crane and John Marsh said many times the RF carrier frequency did not matter. In fact, they eventually quit using the RF carrier frequency when they built their 1950's contact pad style instrument. This clearly shows that they did not understand that the original Rife Ray #5 or Beam Ray Clinical instrument worked on harmonic sidebands. Had they understood this simple fact they never would have changed the RF carrier frequency or built their contact pad style instrument without using an RF carrier frequency. They also would not have lowered or changed any of the audio frequencies if they understood the sideband method used by Philip Hoyland. Both frequency lists would be identical except that one list would be 10 times higher than the other list. Only someone who did not understand how the audio frequencies really worked would have lowered them and then changed them. Only the list which came first would have all the correct frequencies. Again this information proves that the high audio frequency list came first. The low audio frequency list used in the 1953 AZ-58 would have come later in the 1950's because it is the list, when multiplied by 10 times, will only produce some of the correct sideband frequencies. Only someone such as Philip Hoyland could have made the high audio frequency list since the high audio frequencies are the only frequencies that will produce Dr. Rife's higher M.O.R. harmonic frequencies. As we said, the answer to this question is found in the math.

Now we will go back to the EICO 377 and Aubrey Scoon's audio oscillator. Other than the audio oscillator going to 200,000 Hertz (original Beam Ray instrument going to about 42,000 Hertz) Aubrey Scoon's instrument is a replica of the original Beam Ray Clinical instrument. With this knowledge we know that Aubrey Scoon's instrument was a faithful reproduction of Philip Hoyland's Beam Ray Clinical instrument. Just like the original Beam Ray Clinical instrument Aubrey Scoon's Beam Ray Clinical replica instrument used the sine wave waveform for both the RF carrier frequency and the low audio frequencies. The original machines waveform is shown in the first photo, on the left, at the top of the next page. The second photo, on the right, at the top of the next page shows Aubrey Scoon's instrument's modulated waveform. These two waveforms are similar. The replica that we built of Aubrey Scoon's instrument does not have any parasidic oscillations and its waveform looks like the original Beam Ray Clinical instrument waveform shown in the first photo, on the left, on the next page.

www.rife.org

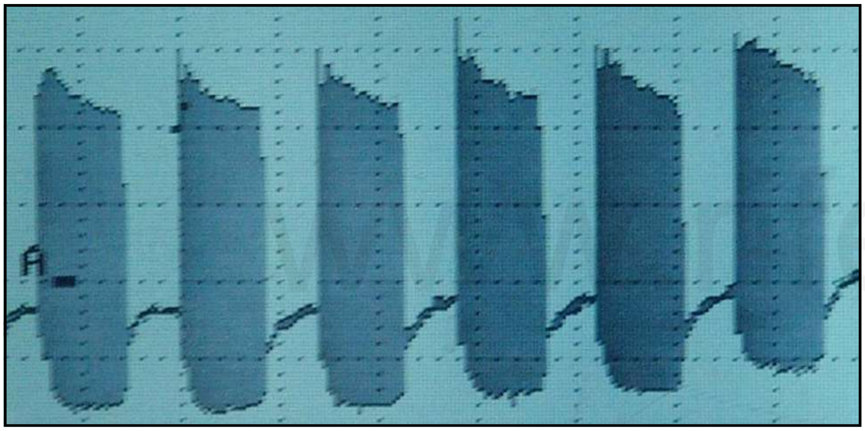

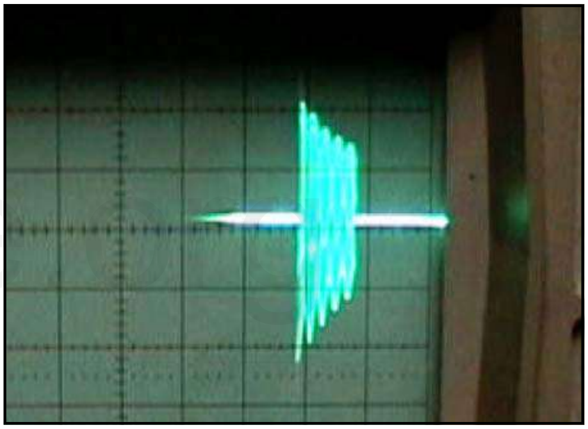

The most important information that came with Aubrey Scoon's instrument was the higher audio frequency list which the doctor who owned it used on his patients. Dr. Low's original Beam Ray Clinical instrument frequency list was lost to time. However, Aubrey Scoon's frequency list was not lost so we can use it and reconcile its audio frequencies to Dr. Rife's original high RF frequencies using the side-band method that was used in this style of instrument.

Since we know that Philip Hoyland tested this Clinical instrument in the laboratory he would have calculated the exact audio frequencies to hit the M.O.R.s. The Rife ray #4 frequencies could be one quarter of one percent off, because of the limits of the 1930's technology, if Philip Hoyland only read the frequency one time. This appears to be what happened because the frequencies read in 1935 are rounded to the nearest one thousandth. In 1935 Philip Hoyland needed the information about the range of Dr. Rife's frequencies to build the Rife Ray #4. But in 1936 when he was building the Rife Ray #5 or Beam Ray Clinical instrument he needed more accurate frequencies for this new instrument since it was using the new sideband method. When the math was done using the high audio frequencies it showed that the frequencies were not rounded to the nearest one thousandth, but they were more precise. It is apparent that the testing Philip Hoyland did in the laboratory in the summer of 1936 on microorganisms allowed him to get the most accurate frequency for each organism. In 1935 the Rife Ray #4 frequency given for Streptothrix was 192,000 Hertz but the frequency for the 1936 Rife Ray #5 or Beam Ray Clinical instrument was 191,803. Not only was this frequency just a little different, but all the frequencies were a little different. This indicates that Philip Hoyland took a more accurate reading of the frequencies of each organism in 1936 for use in his new instrument.

When we did the mathematical equations in order to determine if the audio frequencies from Aubrey Scoon's instrument could produce the correct M.O.R. frequencies when used with the 3.30 MHz RF carrier frequency the math had to be done in reverse order. Using the audio frequencies to determine the most accurate M.O.R.s through the sideband frequencies was the only way to figure out what the frequencies were for each organism listed. If these harmonic frequencies, when divided down, were within one quarter of one percent of the Rife Rays #4 original M.O.R.s. then we knew that the 3.30 MHz RF carrier frequency was the correct carrier frequency. This would also prove that the sideband method was the method of producing the M.O.R.s. in the Beam Ray Clinical instrument. Doing this would also show that the two instruments worked identically the same way. Aubrey Scoon's Beam Ray Clinical instrument with its audio frequencies would firmly prove the sideband method was the method that Philip Hoyland developed. Aubrey Scoon's instrument would also prove that Philip Hoyland used at least two different fixed RF carrier frequencies in the instruments in order to help keep anyone from figuring out the secrets of the instruments. If the carrier frequency is different then the audio frequencies will also be different because they have to be properly matched in order to create the sidebands on the correct high RF harmonic M.O.R. frequencies.

www.rife.org

Below is a frequency comparison chart of Aubrey Scoon’s Beam Ray Clinical replica instrument. In the “Rife Ray #4 Frequencies In Hertz” column are Dr. Rife's M.O.R. frequencies read by Philp Hoyland in 1935. In the “Aubrey Scoon’s Sideband Audio Frequencies In Hertz” column are the audio frequencies used to create the correct sideband frequencies to hit the harmonic Rife Ray #4 frequencies. In the “Rife Ray #4 Frequencies Based on Scoon’s Audio Frequencies” column we see the more accurate M.O.R. frequencies that these audio frequencies produce. You will notice in the “Rife Ray #4 Frequencies In Hertz” column that the frequency for Actinomycosis or Streptothrix is 192,000 Hertz and in the “Rife Ray #4 Frequencies Based on Scoon’s Audio Frequencies” column is the frequency of 191,803 Hertz. There is only a 197 Hertz difference between these frequencies. If you compare both of these columns you will notice how closely these frequencies match up. All the frequencies which are in the column “Rife Ray #4 Frequencies Based On Scoon’s Audio Frequencies” are less than one quarter of one percent off of the “Rife Ray #4 Frequencies In Hertz.” Anyone looking at this chart can see that this could not be just a coincidence.

Aubrey Scoon’s Sideband Audio Frequencies Reconciled To Rife’s Original High Frequency M.O.R.s.

Microorganism *These frequencies were not included in Aubrey Scoon’s instrument.	**Rife Ray #4 Frequencies In Hertz**	**Higher Rife Ray #4 Harmonic Frequencies In Hertz**	**Aubrey Scoon’s Sideband Audio Frequencies In Hertz**	**Higher Rife Ray #4 Harmonic Frequencies in Hertz Based On Scoon’s Audio Frequency and 3.30 MHz Carrier**	***Carrier Difference Frequency**	**Number of Sideband Harmonics**	**Rife Ray #4 Frequencies Based On Scoon’s Audio Frequencies**
Actinomycosis or Streptothrix	192,000 Hz	3,264,000 or 17th	7,870 Hz	3,260,650 or 17th	39,350 Hz	5	191,803 Hz
Anthrax	139,200 Hz	3,340,800 or 24th	*10,200 Hz	3,340,800 or 24th	40,800 Hz	4	139,200 Hz
B or E Coli Rod	417,000 Hz	3,336,000 or 8th	8,020 Hz	3,332,080 or 8th	32,080 Hz	4	416,510 Hz
B or E Coli Virus	770,000 Hz	3,080,000 or 4th	17,220 Hz	3,076,140 or 4th	223,860 Hz	13	769,035 Hz
BX Virus Carcinoma	1,604,000 Hz	3,208,000 or 2nd	21,275 Hz	3,214,900 or 2nd	85,100 Hz	4	1,607,450 Hz
BY Sarcoma	1,530,000 Hz	3,059,040 or 2nd	20,080 Hz	3,059,040 or 2nd	240,960 Hz	12	1,529,520 Hz
Gonorrhea	233,000 Hz	3,262,000 or 14th	*9,500 Hz	3,262,000 or 14th	38,000 Hz	4	233,000 Hz
Pneumonia or Spinal Meningitis	427,000 Hz	3,416,000 or 8th	7,660 Hz	3,414,900 or 8th	114,900 Hz	15	426,862 Hz
Staphylococcus Pyogenes Aureus	478,000 Hz	3,346,000 or 7th	7,270 Hz	3,343,620 or 7th	43,620 Hz	6	477,660 Hz
Streptococcus Pyogenes	720,000 Hz	3,600,000 or 5th	8,450 Hz	3,595,750 or 5th	295,750 Hz	35	719,150 Hz
Syphilis	789,000 Hz	3,156,000 or 4th	6,600 Hz	3,154,800 or 4th	145,200 Hz	22	788,700 Hz
Tetanus	234,000 Hz	3,276,000 or 14th	1,200 Hz	3,276,000 or 14th	24,000 Hz	20	234,000 Hz
Tuberculosis Rod	369,000 Hz	3,321,000 or 9th	8,300 Hz	3,324,897 or 9th	24,897 Hz	3	369,433 Hz
Tuberculosis Virus	769,000 Hz	3,076,000 or 4th	16,000 Hz	3,076,000 or 4th	224,000 Hz	14	769,000 Hz
Typhoid Rod	760,000 Hz	3,040,000 or 4th	6,900 Hz	3,037,800 or 4th	262,200 Hz	38	759,450 Hz
Typhoid Virus	1,445,000 Hz	2,890,000 or 2nd	18,620 Hz	2,890,360 or 2nd	409,640 Hz	22	1,445,180 Hz
Worms	?		2,400 Hz				

www.rife.org

We will now give a simple description of how we reconciled these audio frequencies to Dr. Rife's original high RF frequencies which were used in the Rife Ray #3 and Rife Ray #4 Rife Machines. The Rife Ray #4 paperwork gives us the Streptothrix frequency of 192,000 Hertz. Since we know that Philip Hoyland used the higher harmonic frequency closest to the RF carrier frequency in these instruments what we have to do is multiply 192,000 Hertz by 17 to get the closest frequency to the 3,300,000 Hertz. The 192,000 Hertz multiplied by 17 give us a frequency of 3,264,000 Hertz. The difference between these two frequencies is only 36,000 Hertz. This math gives us the method that Philip Hoyland used.

Now Hoyland used an audio frequency of 7,870 Hertz as the frequency to produce the proper sideband spacing in Aubrey Scoon's instrument. If we multiply 7,870 Hertz times 5 we get the frequency of 39,350 Hertz which is the closest frequency to 36,000 Hertz. If we take 3,300,000 Hertz and minus 39,350 Hertz we get 3,260,650 Hertz which would be the higher harmonic frequency which was used by Philip Hoyland on Streptothrix. Now if we divide 3,360,650 Hertz by 17 we get the true frequency of 191,803 Hertz used by Dr. Rife on Streptothrix. Aubrey Scoon's sideband audio frequencies now give us the most accurate frequencies for the organisms since they are not rounded to the nearest thousandth. Those frequencies are found in the chart, on the previous page, with the column labeled "Rife Ray #4 Frequencies Based on Scoon's Audio Frequencies". The math we did for this chart shows that Aubrey Scoon's Beam Ray Clinical Replica instrument works on the harmonic sideband method to produce the M.O.R.s.

It wasn't until we were able to get the original Beam Ray Clinical instrument and figure out how it worked did we have the ability to determine how Aubrey Scoon's replica instrument really worked. Once we understood how these instruments really worked we could finally figure out the M.O.R. frequencies for Sarcoma, Pneumonia and Tuberculosis. This was made possible because some other documents that we have gave a second reference point to work with to help us determine the correct frequency. Worms (hookworms) however, did not have a second reference point so the only way to produce the M.O.R. frequency for worms is through the sideband method using the audio frequency of 2,400 Hertz in combined with the 3.30 MHz RF carrier frequency.

Also included on pages 137 through 144 are the spectrum analyzer graphs showing the sideband frequencies for each organism using Aubrey Scoon's Beam Ray replica instrument audio frequencies with his 3.30 MHz RF carrier frequency. These spectrum analyzer graphs also include Anthrax and Gonorrhea which was not included with Aubrey Scoon's audio frequency list.

On page 145 are photos of Aubrey Scoon's instrument that we built. On page 146 is the schematic of this 1950's instrument. The 866 vacuum tubes have been replaced with solid state rectifiers. Also the old vacuum tube audio oscillator was not included in the case of the instrument. It is easier and more accurate to use Aubrey Scoon's booster amplifier and a modern function generator to produce the audio frequencies that were used in this instrument. The layout of the electronic parts of this instrument is also very important because of the inherent interference problems that come with RF oscillators. Again anyone who would like to build this instrument should have a good understanding of old tube technology. Some parts of this circuit use up to 2000 volts DC with substantial current and can easily kill anyone who is not familiar with this kind of current or voltage. We take no responsibility for anyone who builds this instrument. We recommend that you have professional help.

On pages 147-159 we have included in this report Aubrey Scoon's original report that was on his website. The reason we have included his report in this report is due to the fact that Aubrey Scoon passed away a few years ago and his website is no longer available on the internet. The information that his report gives is very important to those who are interested in how Dr. Rife's machines worked.

www.rife.org

Chapter Summary: We now know that Aubrey Scoon's instrument is a copy Dr. Couches' original Rife Ray #5 or Beam Ray Clinical instrument which Dr. Couche purchased from the original 1938-1939 Beam Ray Corporation. Aubrey Scoon's instrument was built by Dr. Rife's engineer, Verne Thompson, sometime in 1952 or later. Dr. Low's original Beam Ray Clinical instrument, which was built by the original Beam Ray Corporation, was the most important machine found because it made it so we now know that it was Philip Hoyland who built the audio frequency instrument that used a fixed RF carrier frequency. With this information we were able to prove that Aubrey Scoon's instrument was a faithful replica of the original Beam Ray Clinical instrument. It came with a frequency list that the doctor who owned it used on his patients.

When it comes to the audio frequency list Aubrey Scoon's instrument is the most important Rife machine found because it gives us the original audio frequencies used in the original Beam Ray Clinical instrument. With this frequency list we have been able to prove that Philip Hoyland used higher harmonic frequencies of Dr. Rife's original Rife Ray #3 and Rife Ray #4 frequencies in his Beam Ray Clinical instrument. This list also gives us some of Dr. Rife's frequencies which had been lost, such as the BY Cancer virus frequency. This same audio frequency list was lowered by a factor of 10 times and used in the 1953 AZ-58 which was built by Verne Thompson. With this information we have been able to prove that these audio frequencies are not Dr. Rife's original M.O.R. frequencies and that they do not devitalize any microorganisms. These audio frequencies, when used with a 3.30 Megahertz carrier frequency, will produce harmonic sideband frequencies which will hit the higher harmonic RF M.O.R. frequencies found by Dr. Rife. Though Aubrey Scoon did not live to realize how important his instrument was we now know how important this Rife instrument really is. Dr. Low's instrument and Aubrey Scoon's instrument have revealed how Philip Hoyland's Beam Ray Clinical instrument was a harmonic instrument.

In the next chapter we will discuss the 1953 AZ-58 built by Verne Thompson and the history of that instrument in detail.

www.rife.org

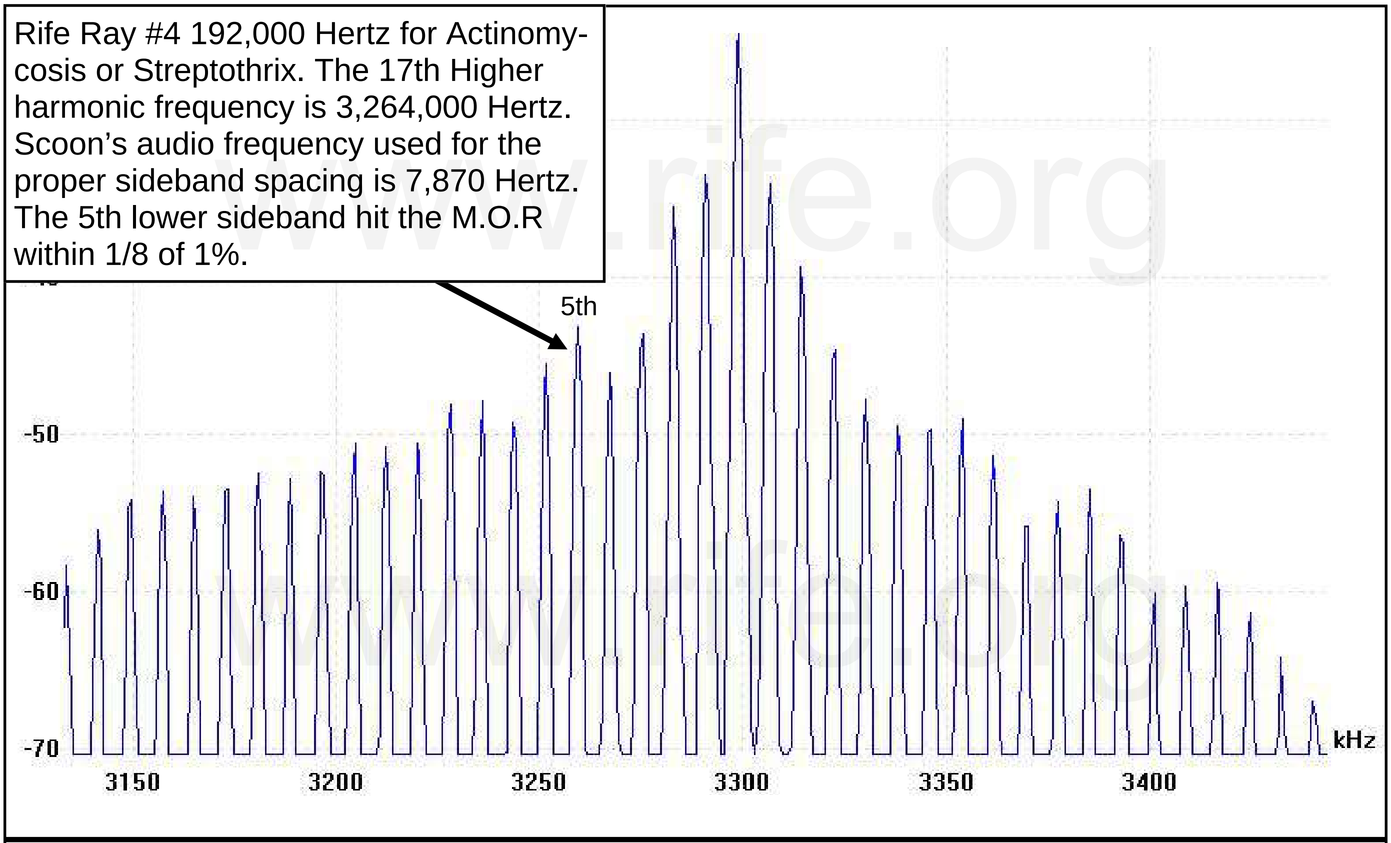

Rife Ray #4 192,000 Hertz for Actinomycosis or Streptothrix. The 17th Higher harmonic frequency is 3,264,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 7,870 Hertz. The 5th lower sideband hit the M.O.R within 1/8 of 1%.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

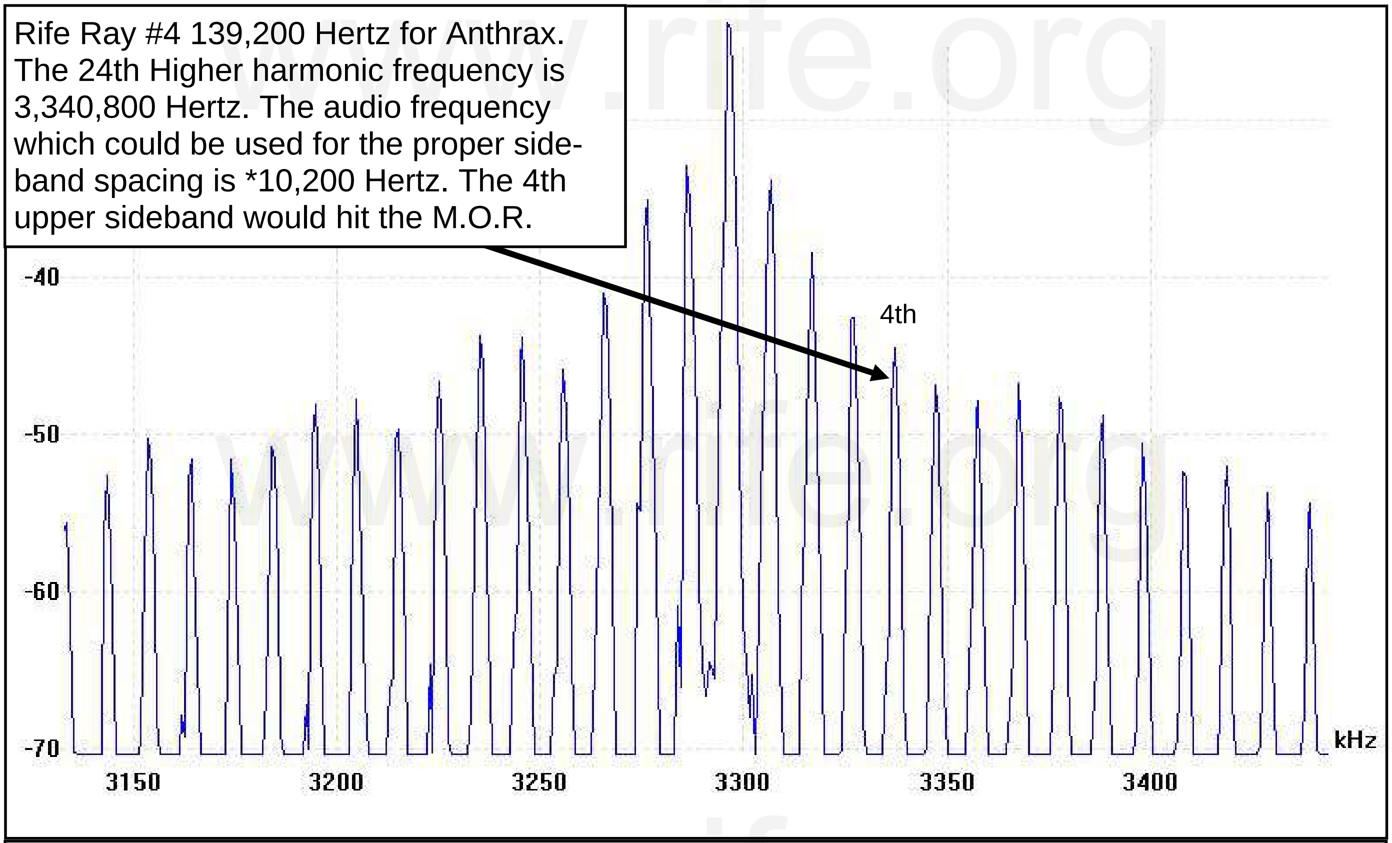

Rife Ray #4 139,200 Hertz for Anthrax. The 24th Higher harmonic frequency is 3,340,800 Hertz. The audio frequency which could be used for the proper sideband spacing is *10,200 Hertz. The 4th upper sideband would hit the M.O.R.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

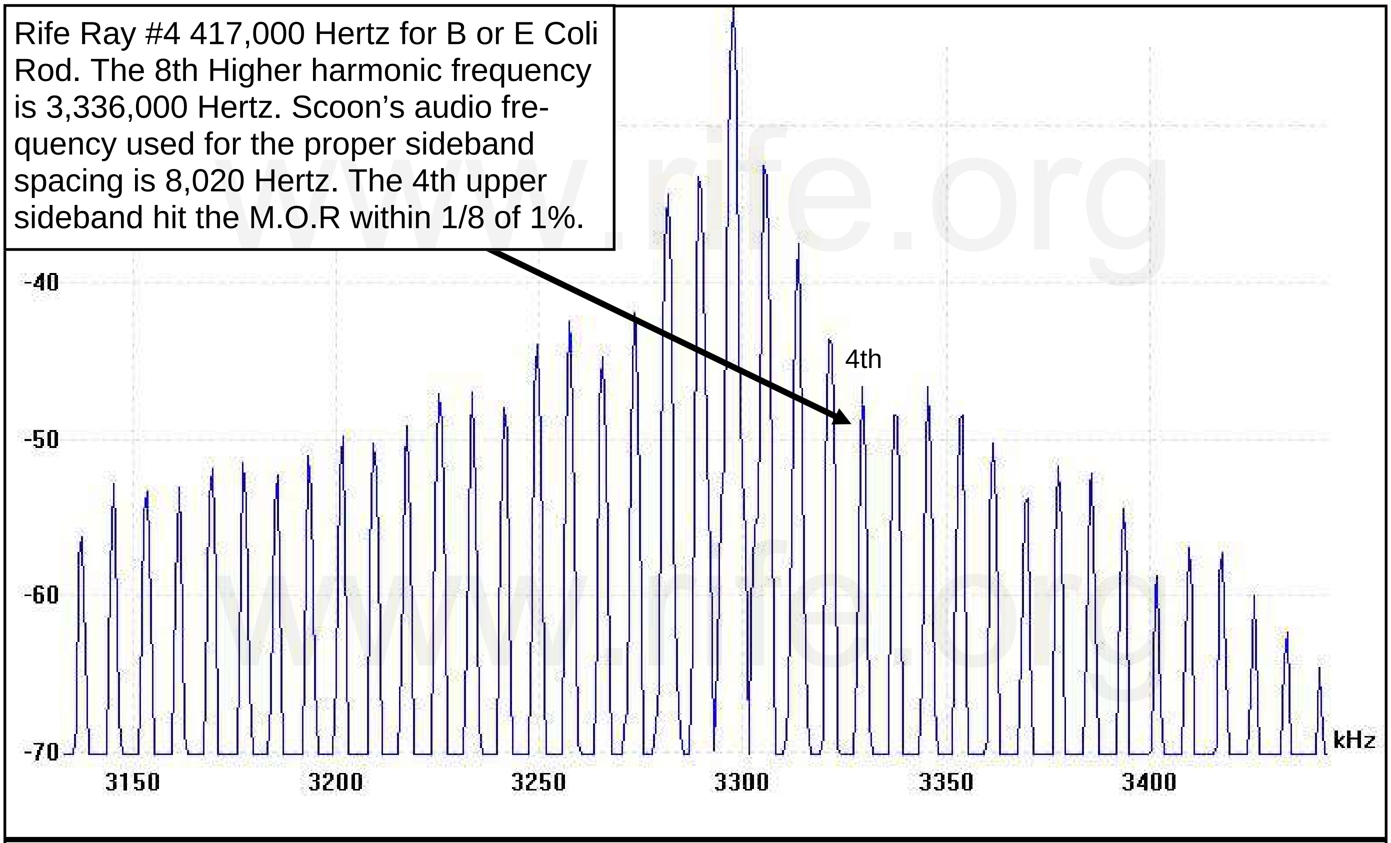

Rife Ray #4 417,000 Hertz for B or E Coli Rod. The 8th Higher harmonic frequency is 3,336,000 Hertz. Scoon’s audio frequency used for the proper sideband spacing is 8,020 Hertz. The 4th upper sideband hit the M.O.R within 1/8 of 1%.

Scoon’s Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

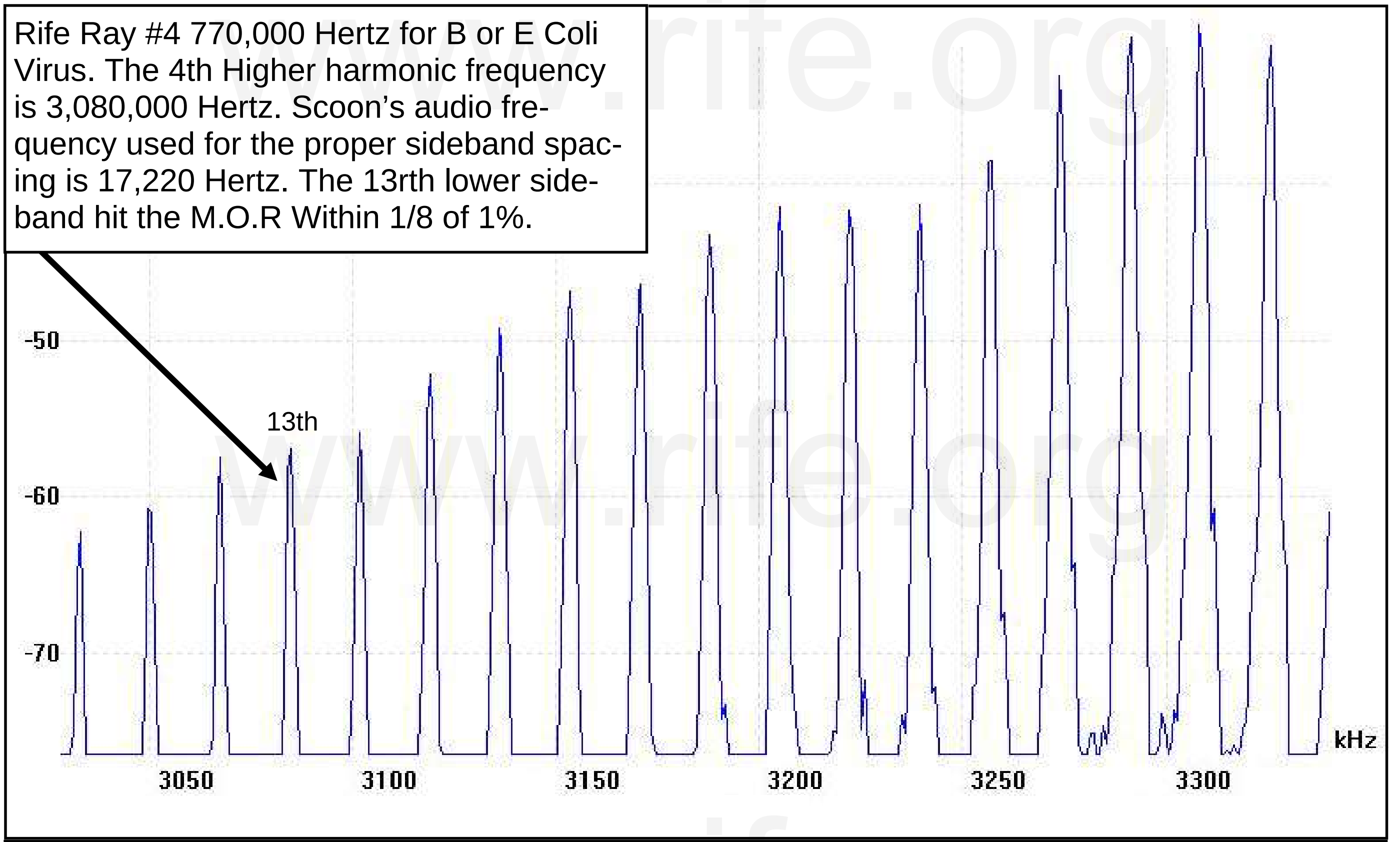

Rife Ray #4 770,000 Hertz for B or E Coli Virus. The 4th Higher harmonic frequency is 3,080,000 Hertz. Scoon’s audio frequency used for the proper sideband spacing is 17,220 Hertz. The 13rth lower sideband hit the M.O.R Within 1/8 of 1%.

Scoon’s Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

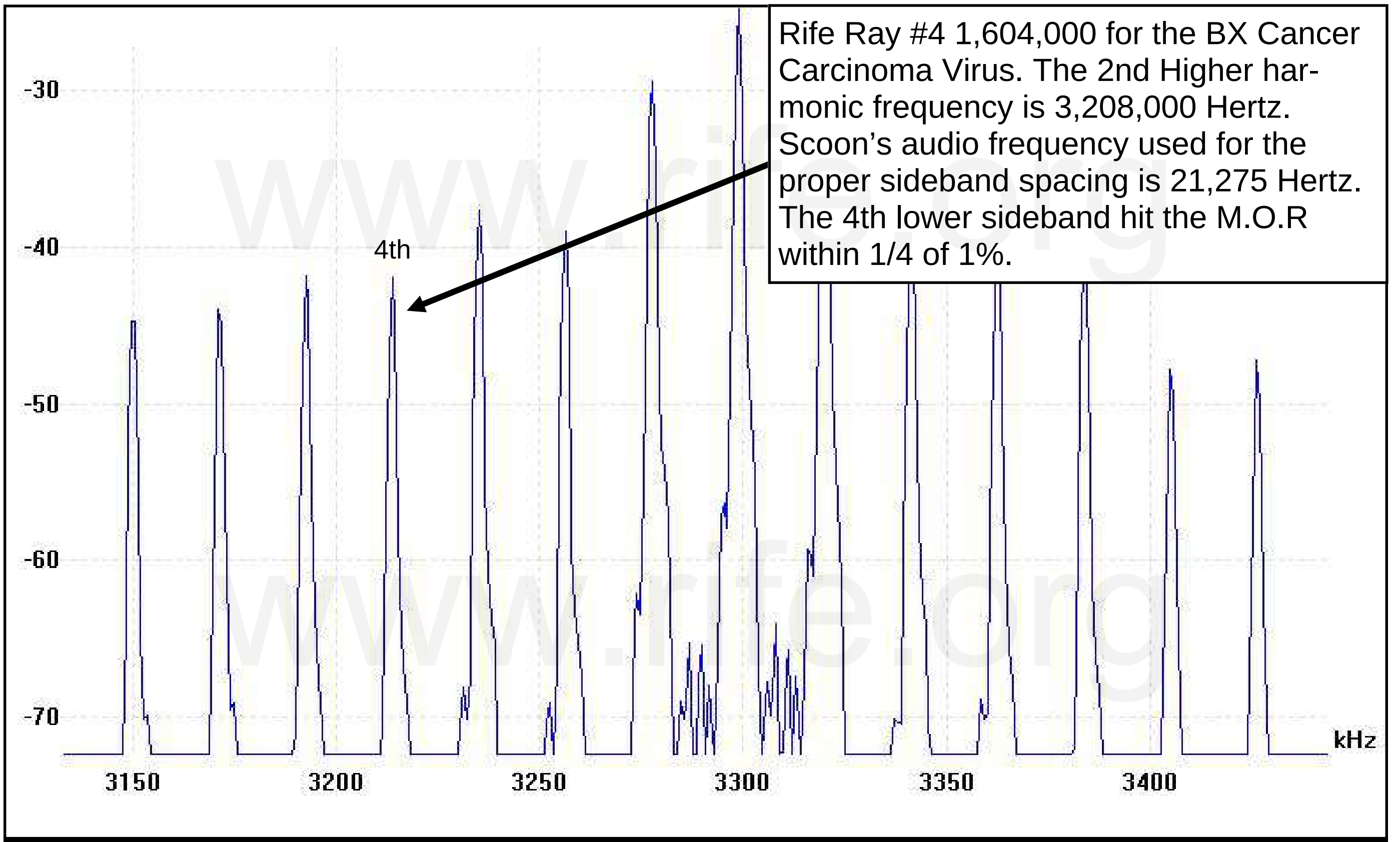

Rife Ray #4 1,604,000 for the BX Cancer Carcinoma Virus. The 2nd Higher harmonic frequency is 3,208,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 21,275 Hertz. The 4th lower sideband hit the M.O.R within 1/4 of 1%.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

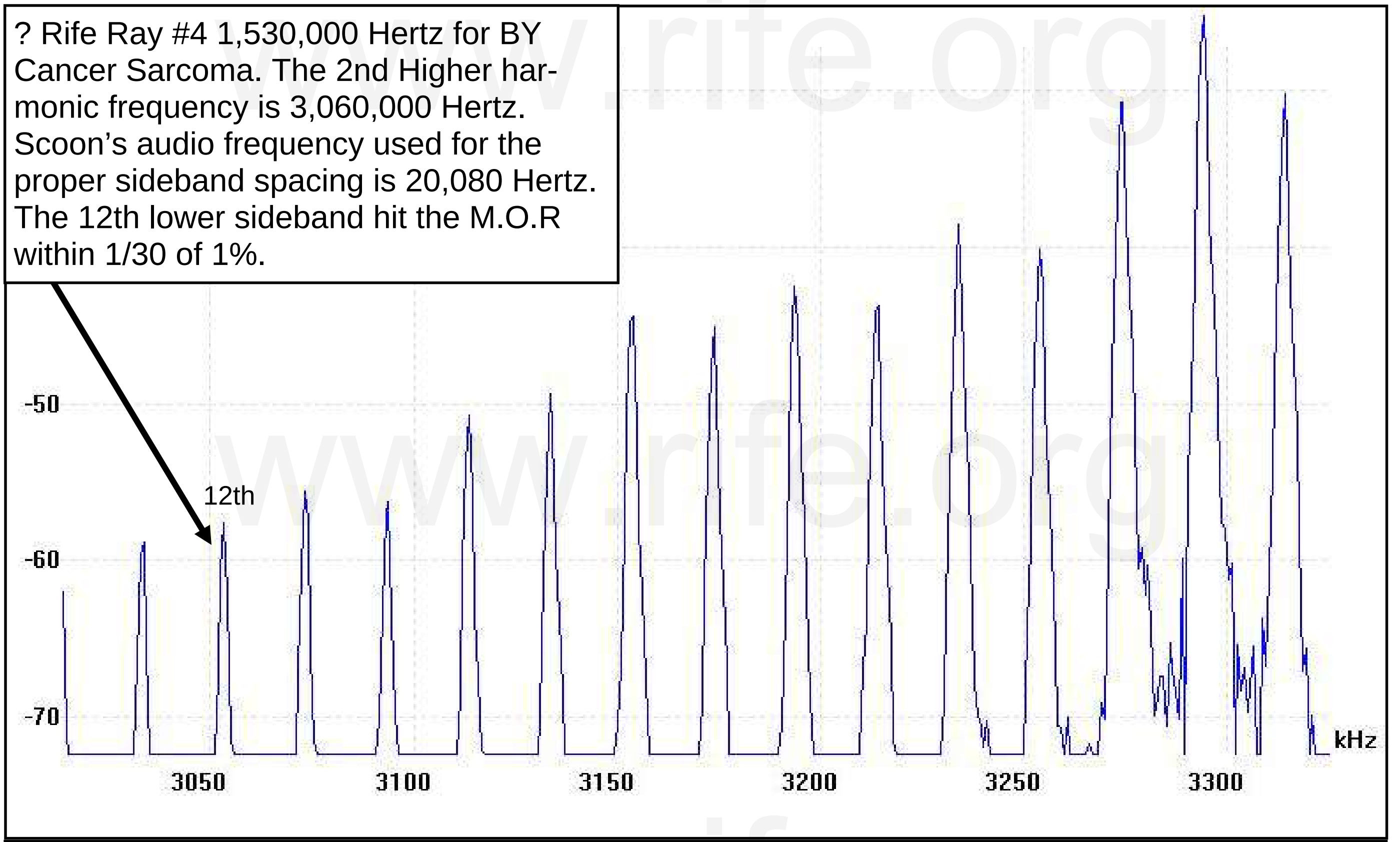

? Rife Ray #4 1,530,000 Hertz for BY Cancer Sarcoma. The 2nd Higher harmonic frequency is 3,060,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 20,080 Hertz. The 12th lower sideband hit the M.O.R within 1/30 of 1%.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

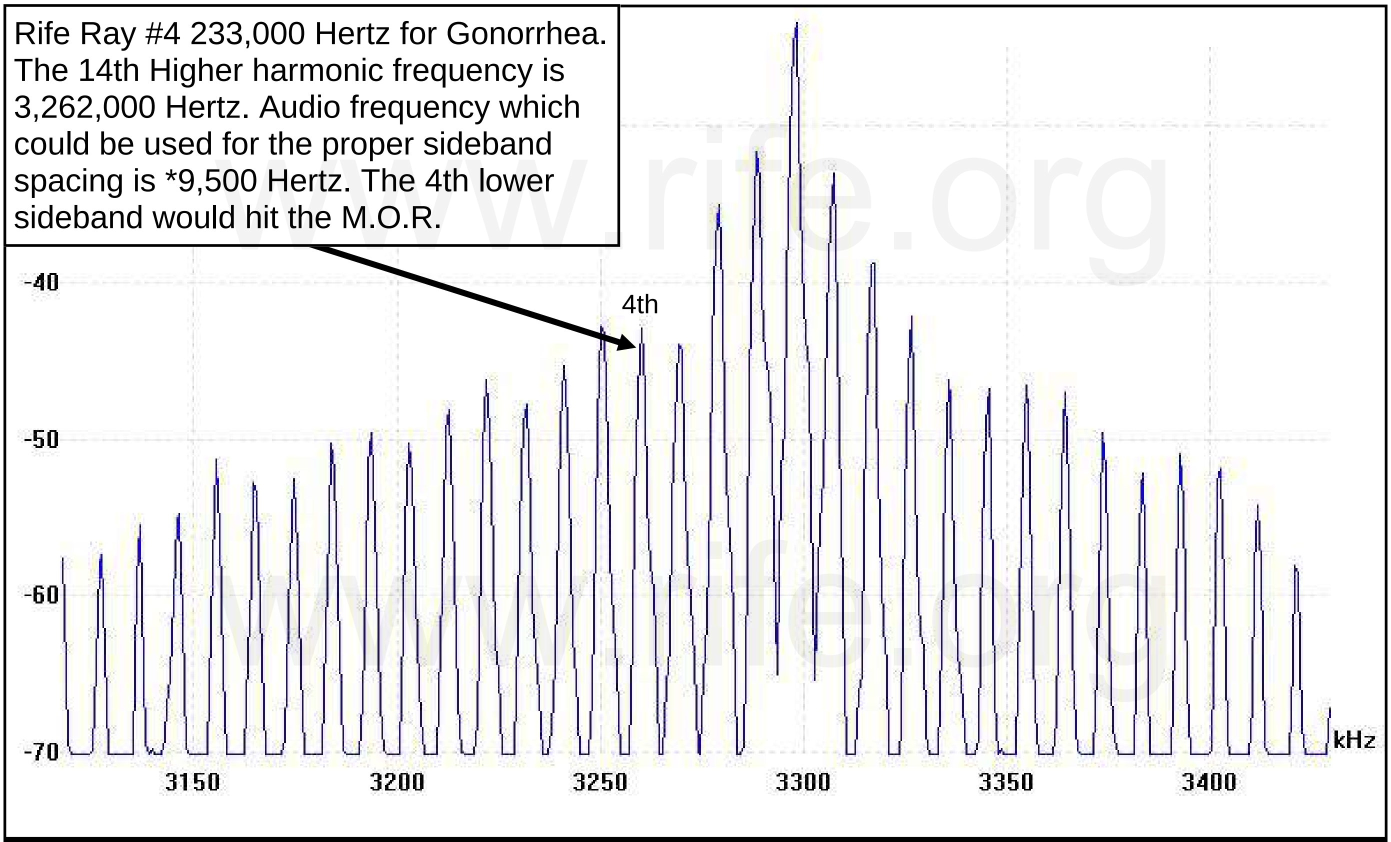

Rife Ray #4 233,000 Hertz for Gonorrhea. The 14th Higher harmonic frequency is 3,262,000 Hertz. Audio frequency which could be used for the proper sideband spacing is *9,500 Hertz. The 4th lower sideband would hit the M.O.R.

Scoon’s Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

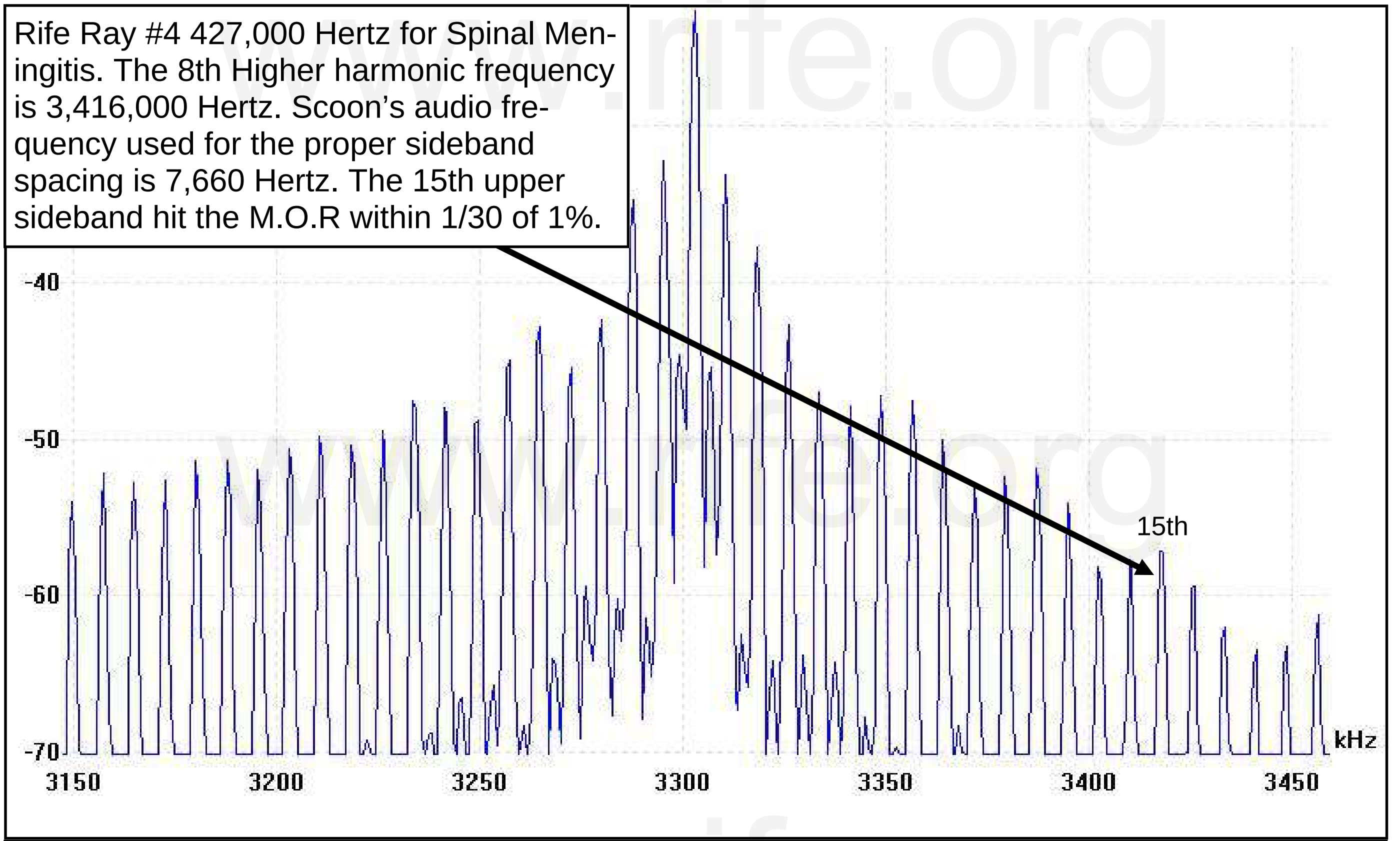

Rife Ray #4 427,000 Hertz for Spinal Meningitis. The 8th Higher harmonic frequency is 3,416,000 Hertz. Scoon’s audio frequency used for the proper sideband spacing is 7,660 Hertz. The 15th upper sideband hit the M.O.R within 1/30 of 1%.

Scoon’s Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

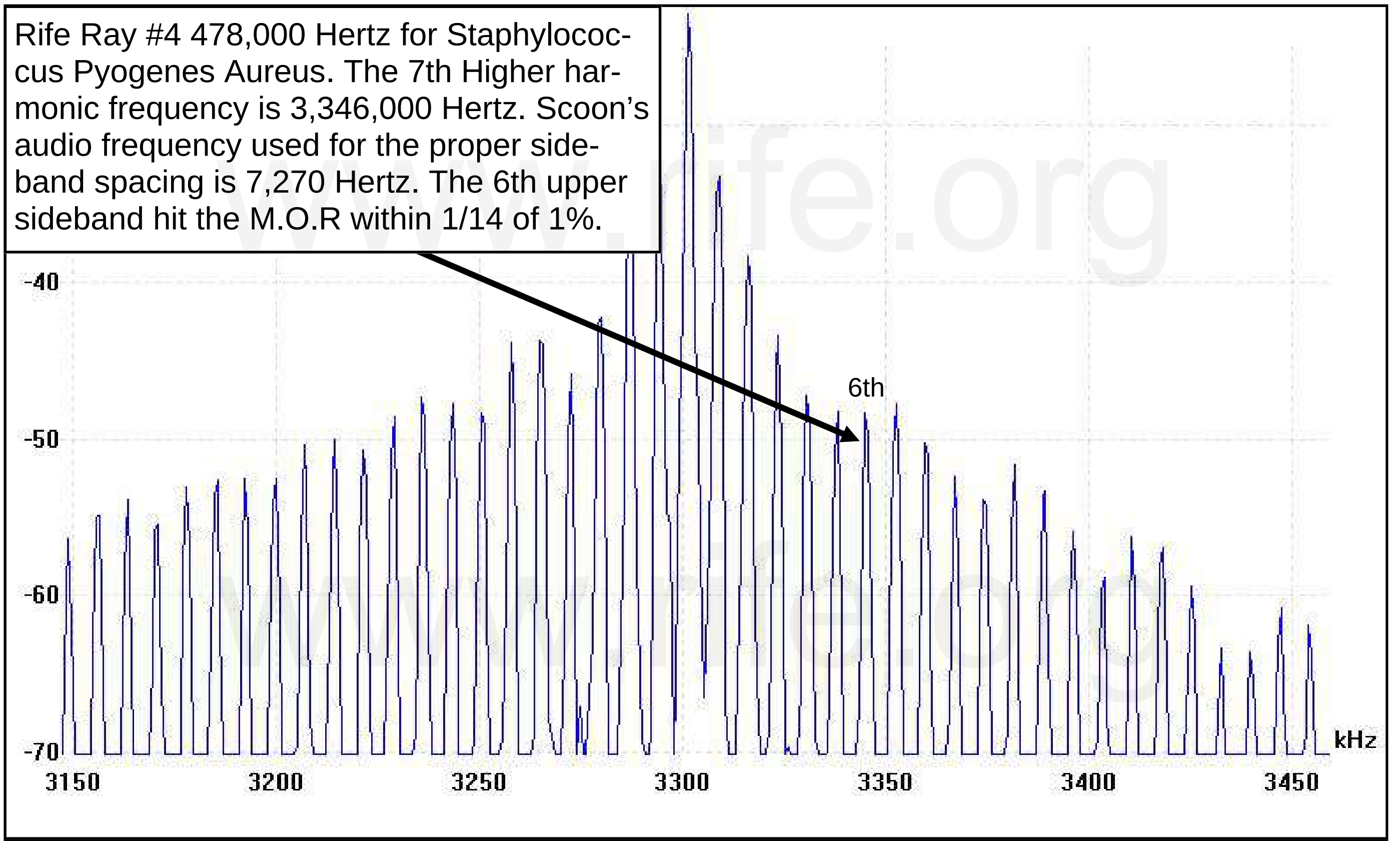

Rife Ray #4 478,000 Hertz for Staphylococcus Pyogenes Aureus. The 7th Higher harmonic frequency is 3,346,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 7,270 Hertz. The 6th upper sideband hit the M.O.R within 1/14 of 1%.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

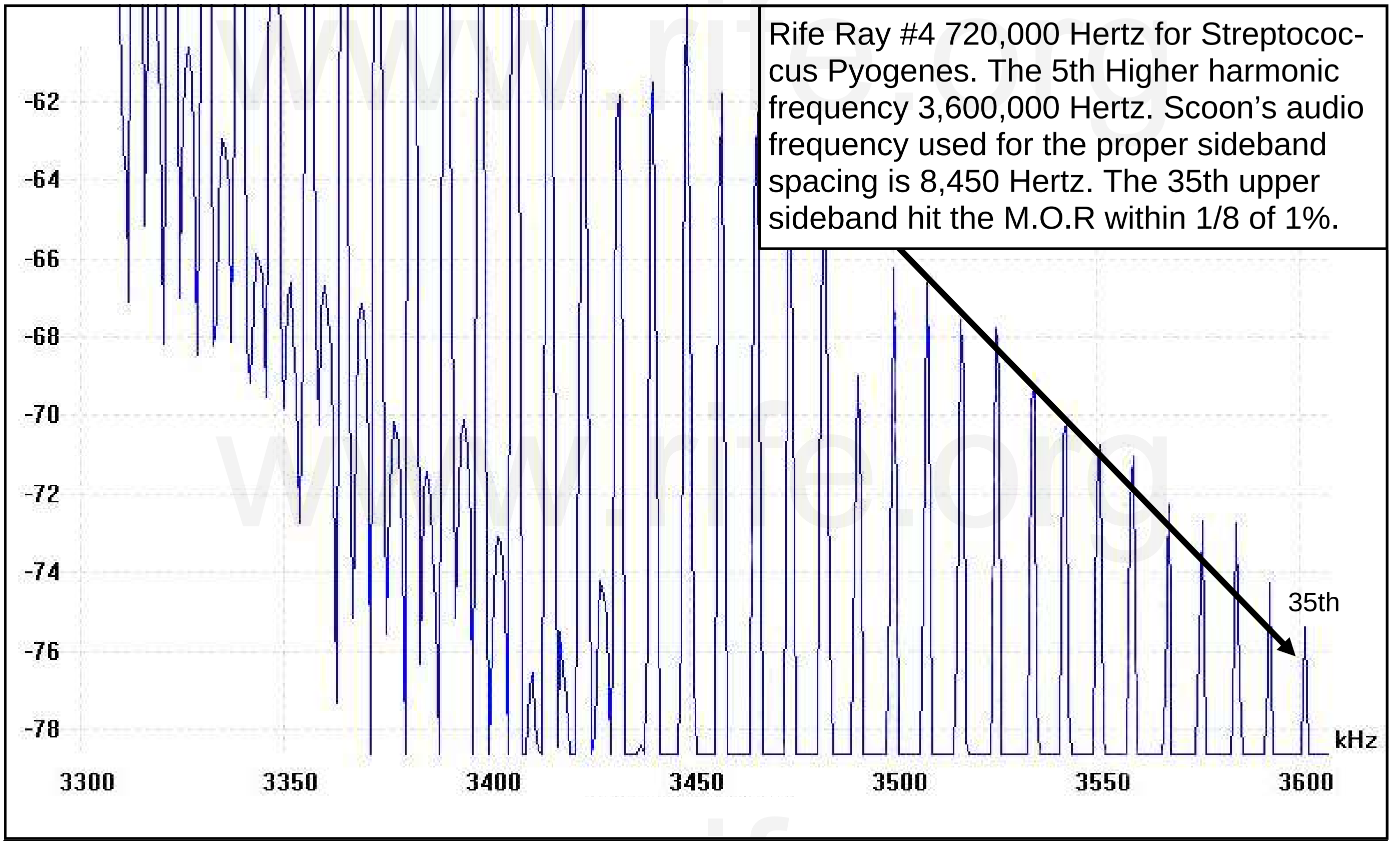

Rife Ray #4 720,000 Hertz for Streptococcus Pyogenes. The 5th Higher harmonic frequency 3,600,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 8,450 Hertz. The 35th upper sideband hit the M.O.R within 1/8 of 1%.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

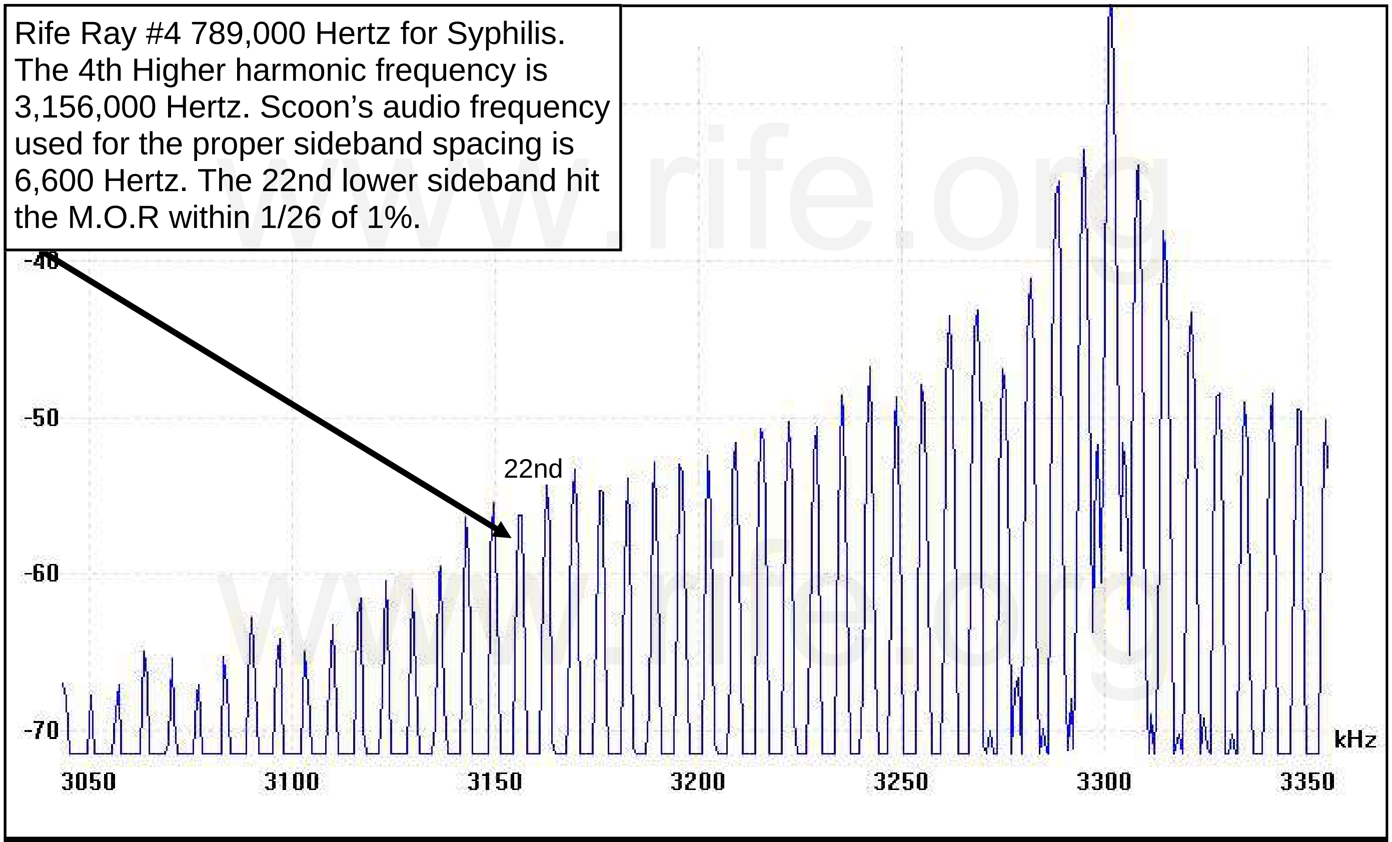

Rife Ray #4 789,000 Hertz for Syphilis. The 4th Higher harmonic frequency is 3,156,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 6,600 Hertz. The 22nd lower sideband hit the M.O.R within 1/26 of 1%.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

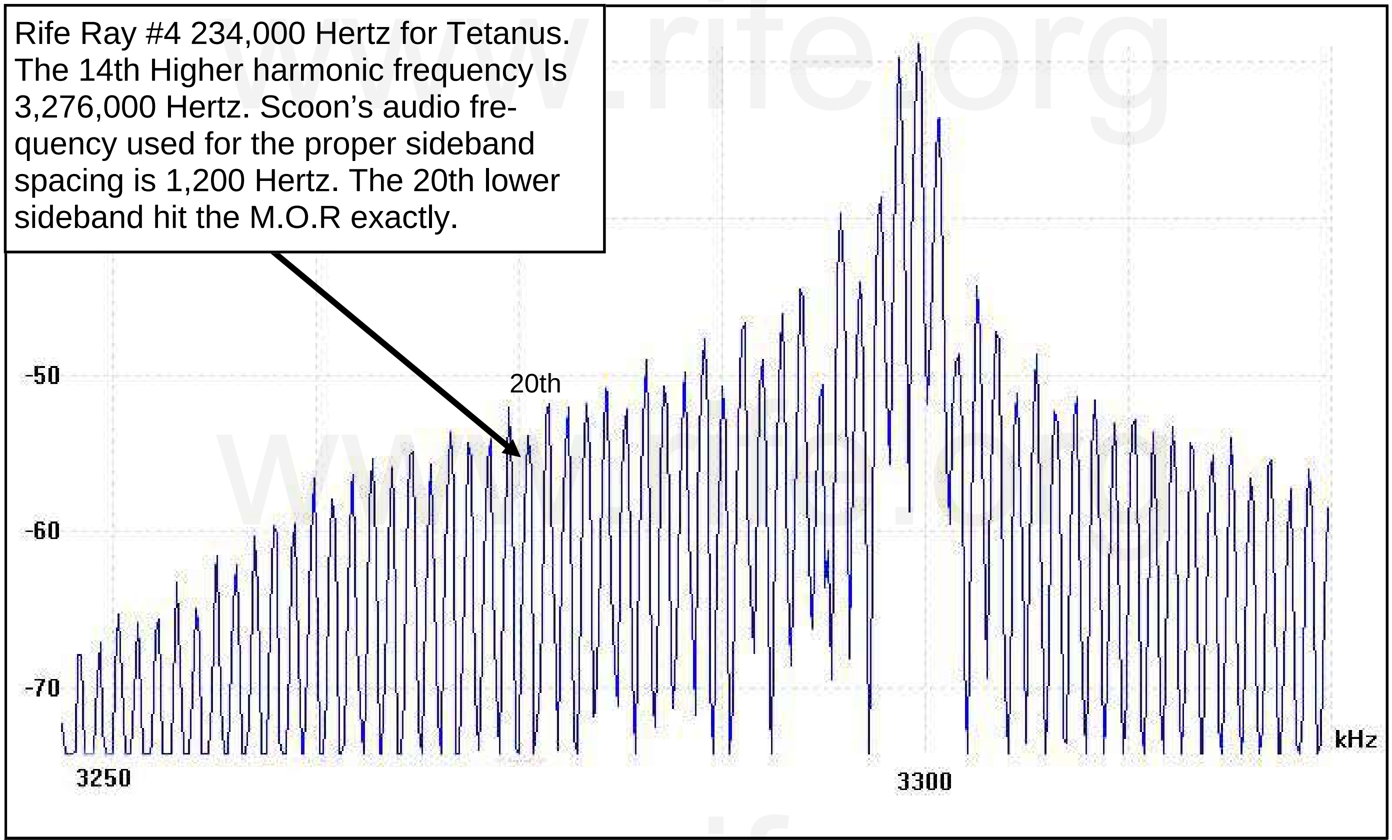

Rife Ray #4 234,000 Hertz for Tetanus. The 14th Higher harmonic frequency Is 3,276,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 1,200 Hertz. The 20th lower sideband hit the M.O.R exactly.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

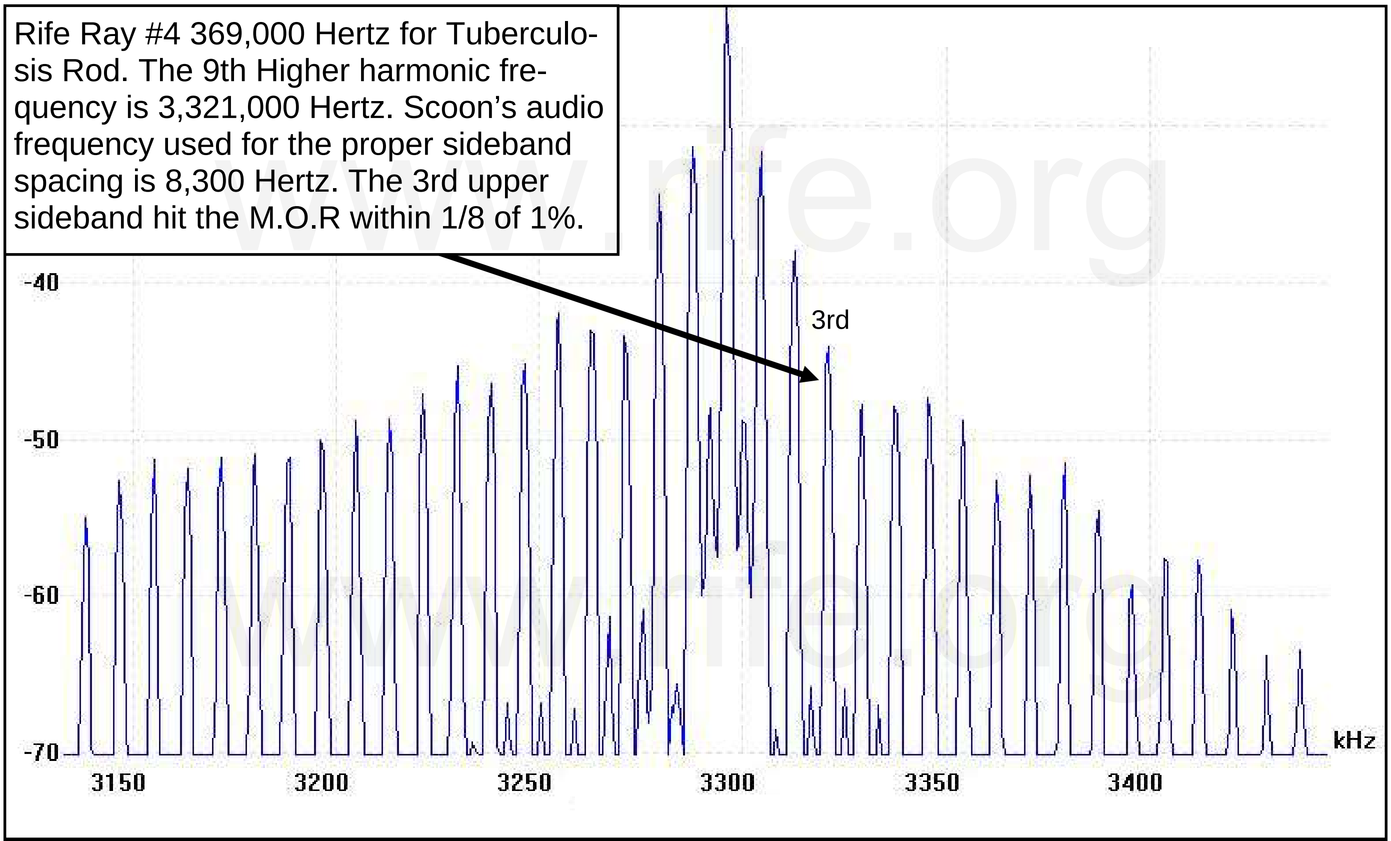
Rife Ray #4 369,000 Hertz for Tuberculosis Rod. The 9th Higher harmonic frequency is 3,321,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 8,300 Hertz. The 3rd upper sideband hit the M.O.R within 1/8 of 1%.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

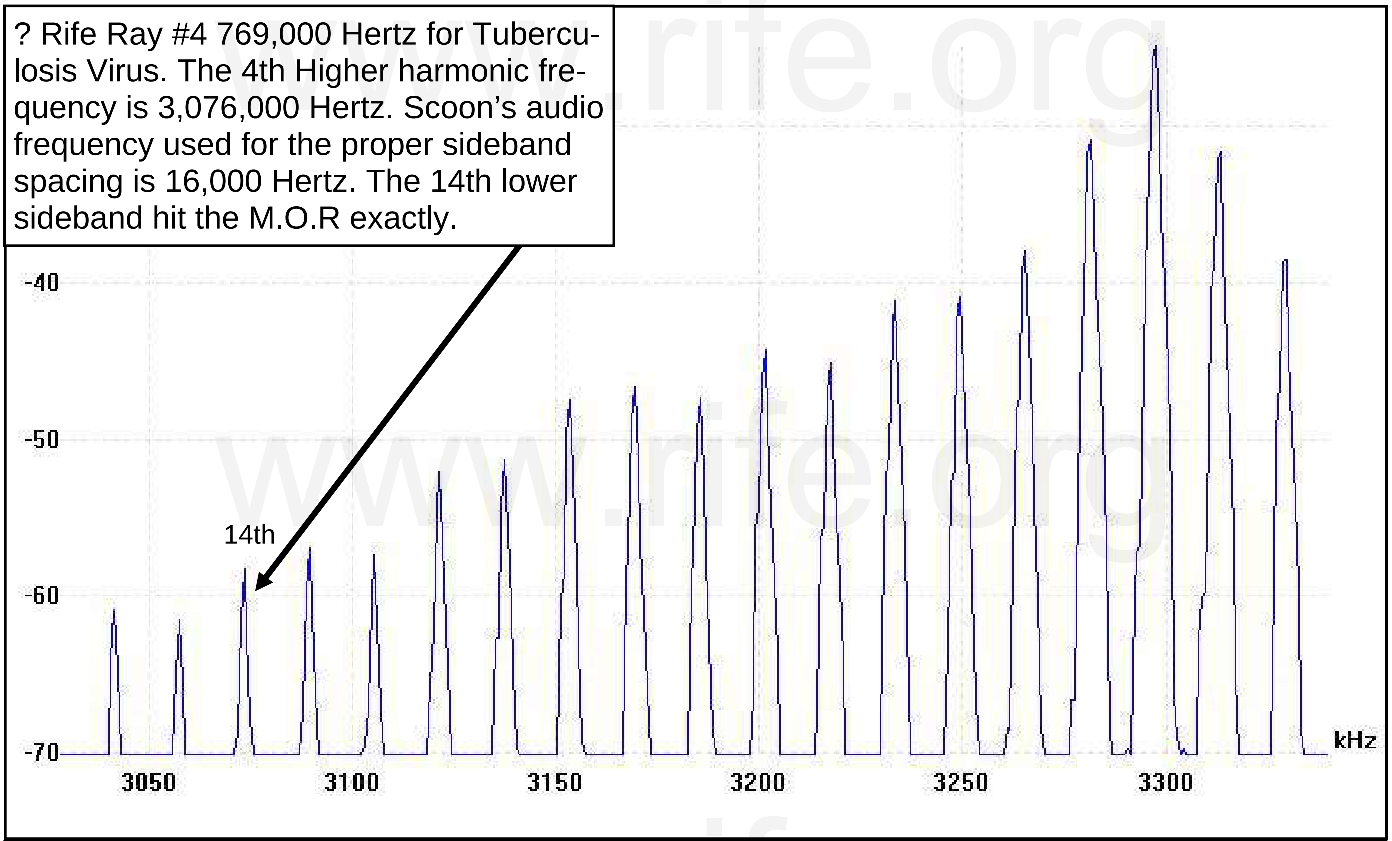
? Rife Ray #4 769,000 Hertz for Tuberculosis Virus. The 4th Higher harmonic frequency is 3,076,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 16,000 Hertz. The 14th lower sideband hit the M.O.R exactly.

Scoon's Beam Ray Clinical instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

www.rife.org

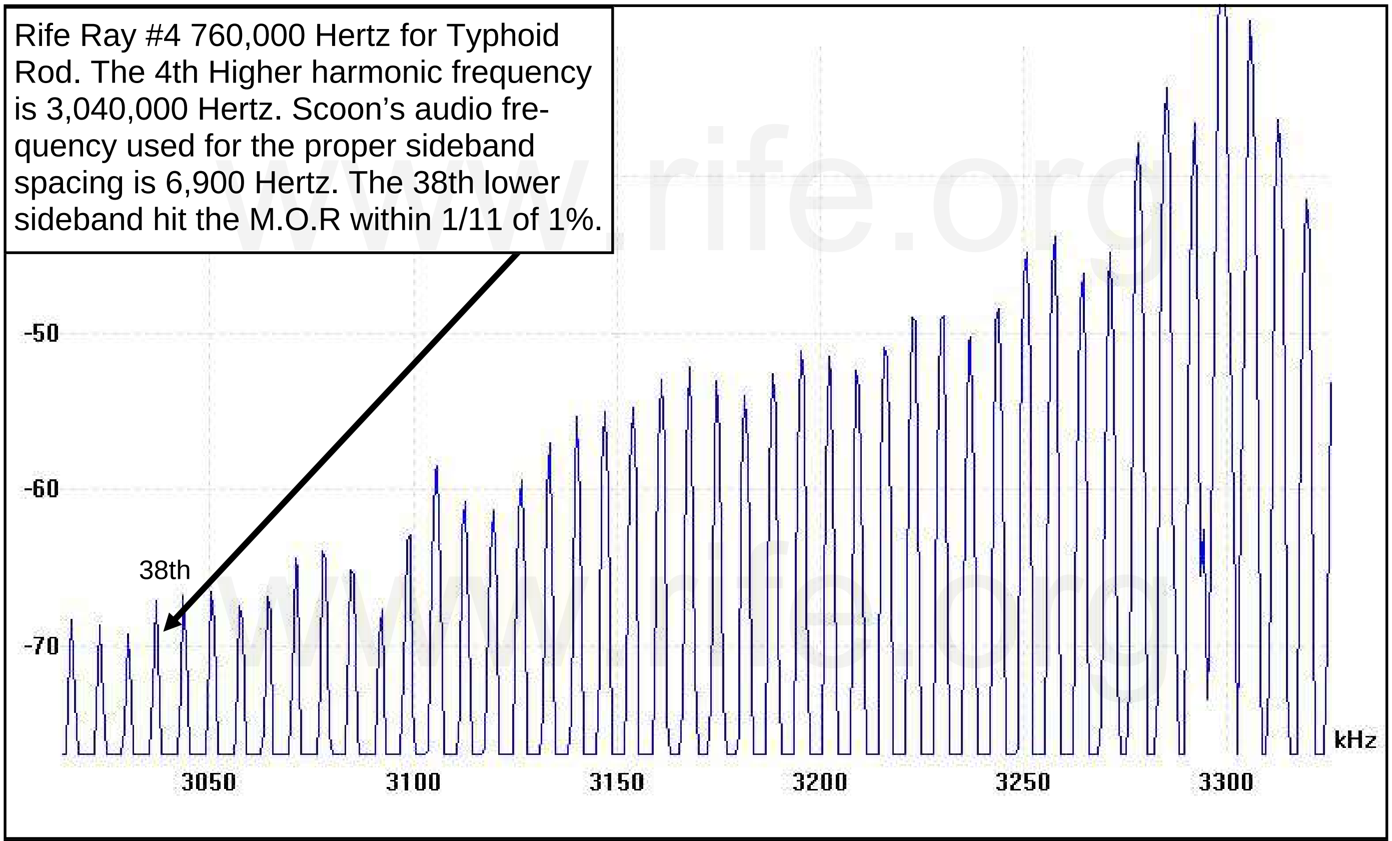

Rife Ray #4 760,000 Hertz for Typhoid Rod. The 4th Higher harmonic frequency is 3,040,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 6,900 Hertz. The 38th lower sideband hit the M.O.R within 1/11 of 1%.

Scoon's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic frequencies.

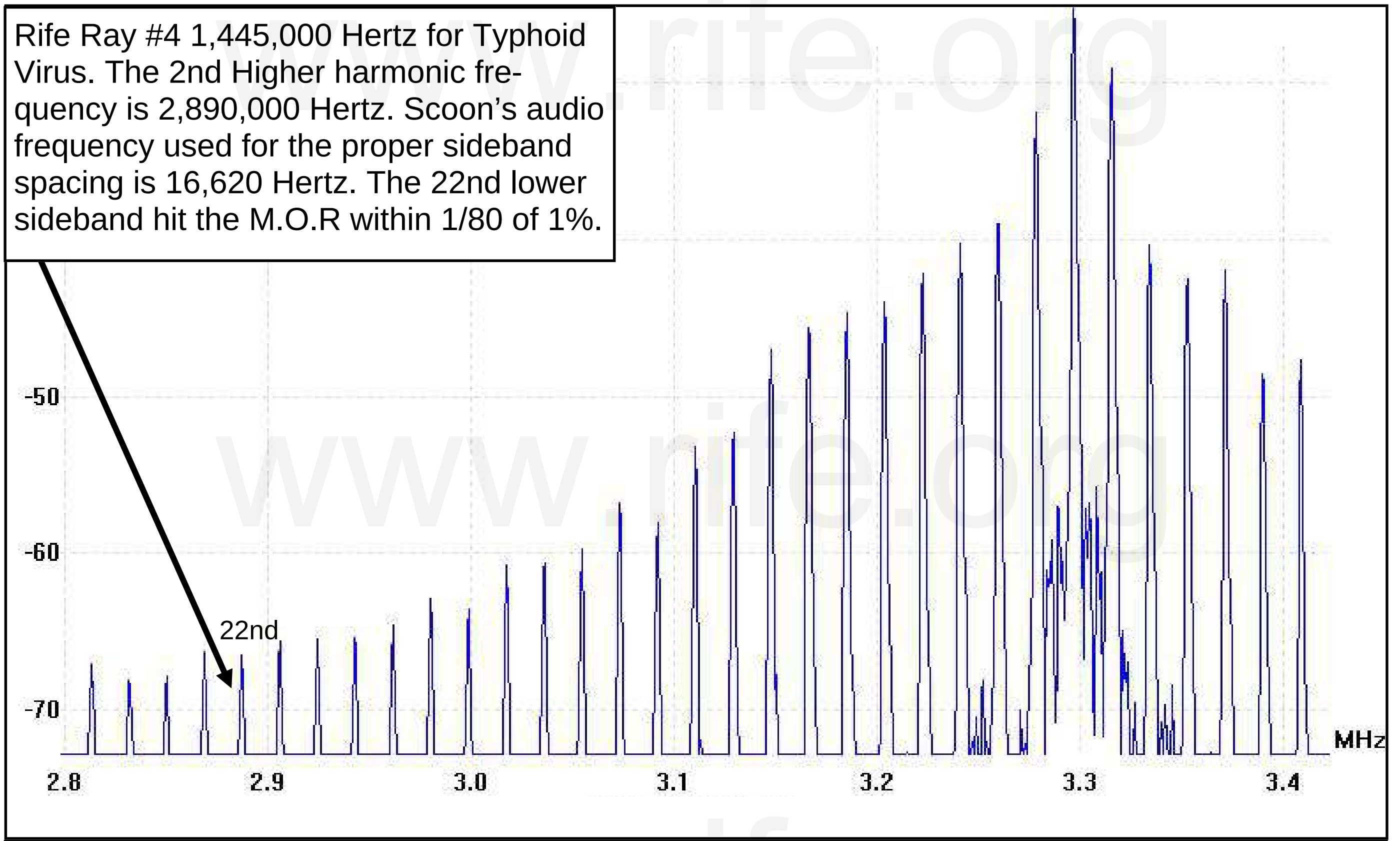

Rife Ray #4 1,445,000 Hertz for Typhoid Virus. The 2nd Higher harmonic frequency is 2,890,000 Hertz. Scoon's audio frequency used for the proper sideband spacing is 16,620 Hertz. The 22nd lower sideband hit the M.O.R within 1/80 of 1%.

Scoon's Hoyland's Beam Ray Clinical replica instrument 3.30 Megahertz spectrum analysis of sideband harmonic

www.rife.org

Photos of the rebuilt Aubrey Scoon Beam Ray Clinical instrument

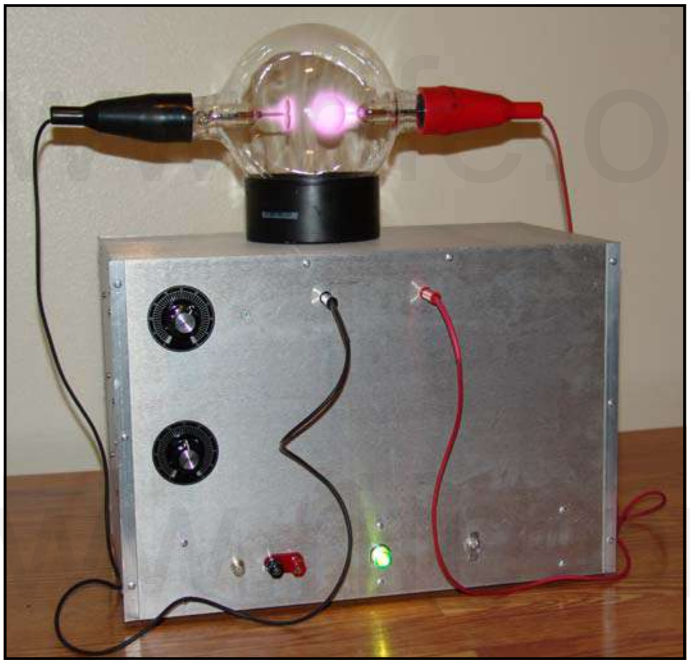

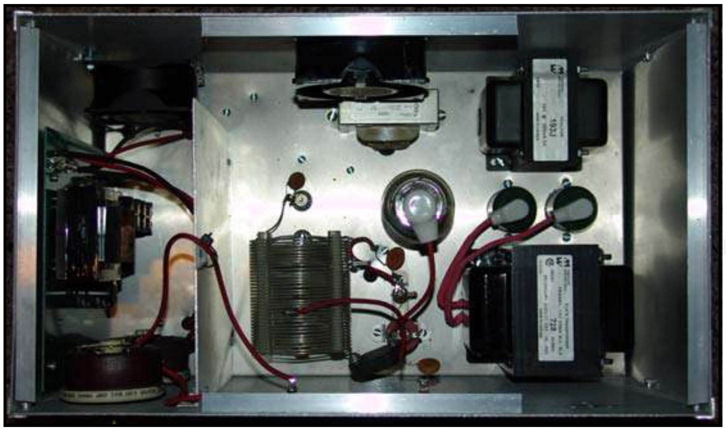

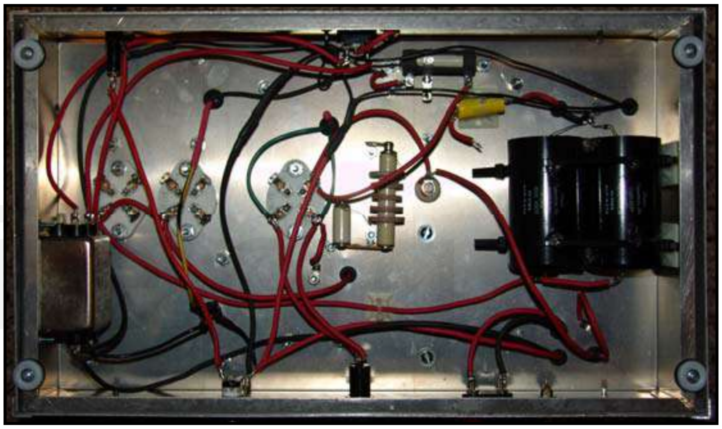

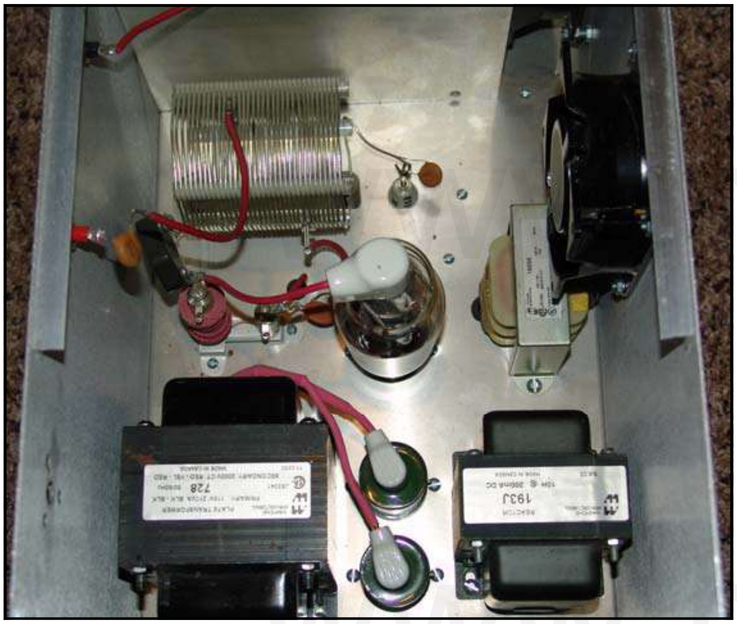

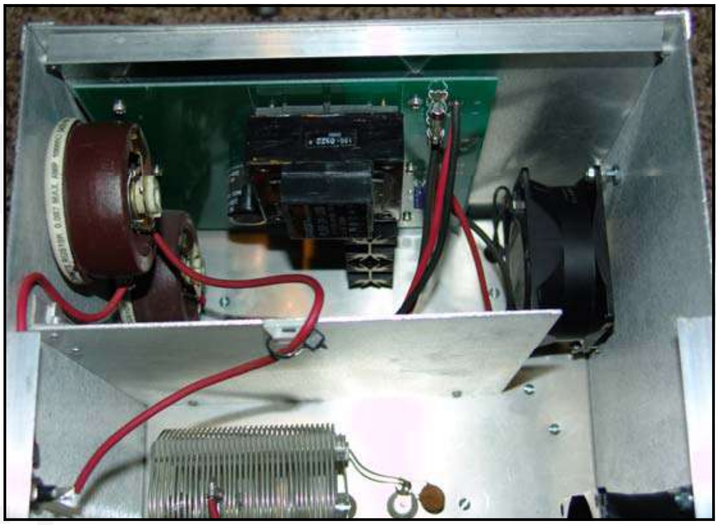

www.rife.org

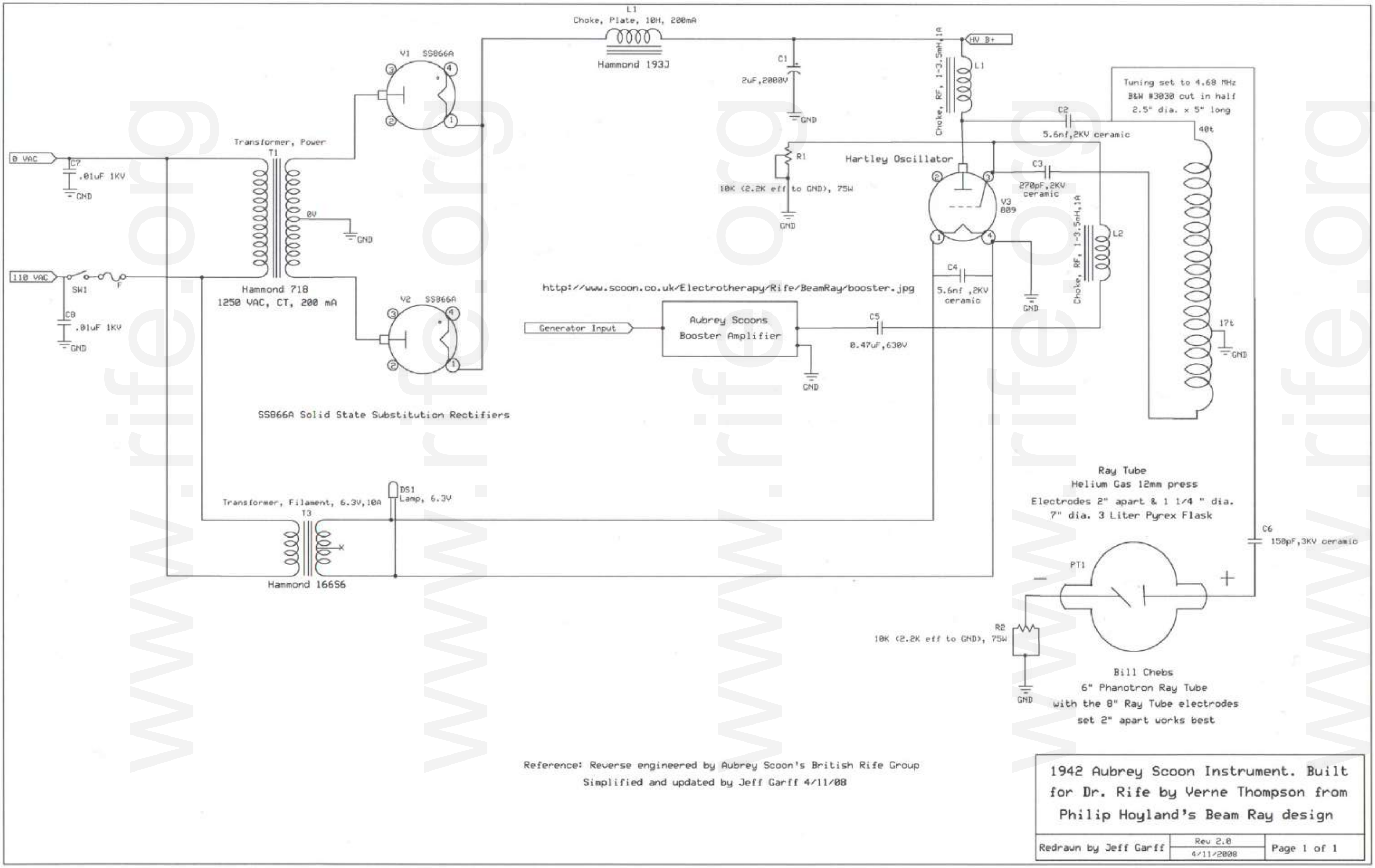
L1
Choke, Plate, 10H, 200mA
Hammond 193J
HV B+
C1
2uF,2000V
GND
Choke, RF, 1-3.5mH,1A
L1
Tuning set to 4.68 MHz
B&W #3030 cut in half
2.5" dia. x 5" long
C2
5.6nf,2KV ceramic
40t
V1 SS866A
Transformer, Power
T1
0 VAC
C7
.01uF 1KV
GND
0V
GND
110 VAC
SW1
F
C8
.01uF 1KV
GND
Hammond 718
1250 VAC, CT, 200 mA
V2 SS866A
R1
10K (2.2K eff to GND), 75W
GND
Hartley Oscillator
C3
270pF,2KV
ceramic
V3
809
C4
5.6nf ,2KV
ceramic
GND
Choke, RF, 1-3.5mH,1A
L2
http://www.scoon.co.uk/Electrotherapy/Rife/BeamRay/booster.jpg
Generator Input
Aubrey Scoons
Booster Amplifier
C5
0.47uF,630V
GND
17t
GND
SS866A Solid State Substitution Rectifiers
Ray Tube
Helium Gas 12mm press
Electrodes 2" apart & 1 1/4 " dia.
7" dia. 3 Liter Pyrex Flask
DS1
Lamp, 6.3V
Transformer, Filament, 6.3V,10A
T3
X
Hammond 166S6
C6
150pF,3KV ceramic
PT1
–
+
R2
10K (2.2K eff to GND), 75W
GND
Bill Chebs
6" Phanotron Ray Tube
with the 8" Ray Tube electrodes
set 2" apart works best

Reference: Reverse engineered by Aubrey Scoon's British Rife Group
Simplified and updated by Jeff Garff 4/11/08

1942 Aubrey Scoon Instrument. Built for Dr. Rife by Verne Thompson from Philip Hoyland's Beam Ray design

Redrawn by Jeff Garff	Rev 2.0 4/11/2008	Page 1 of 1

www.rife.org

Aubrey Scoon's 1939 Rife Beam Ray Website Report

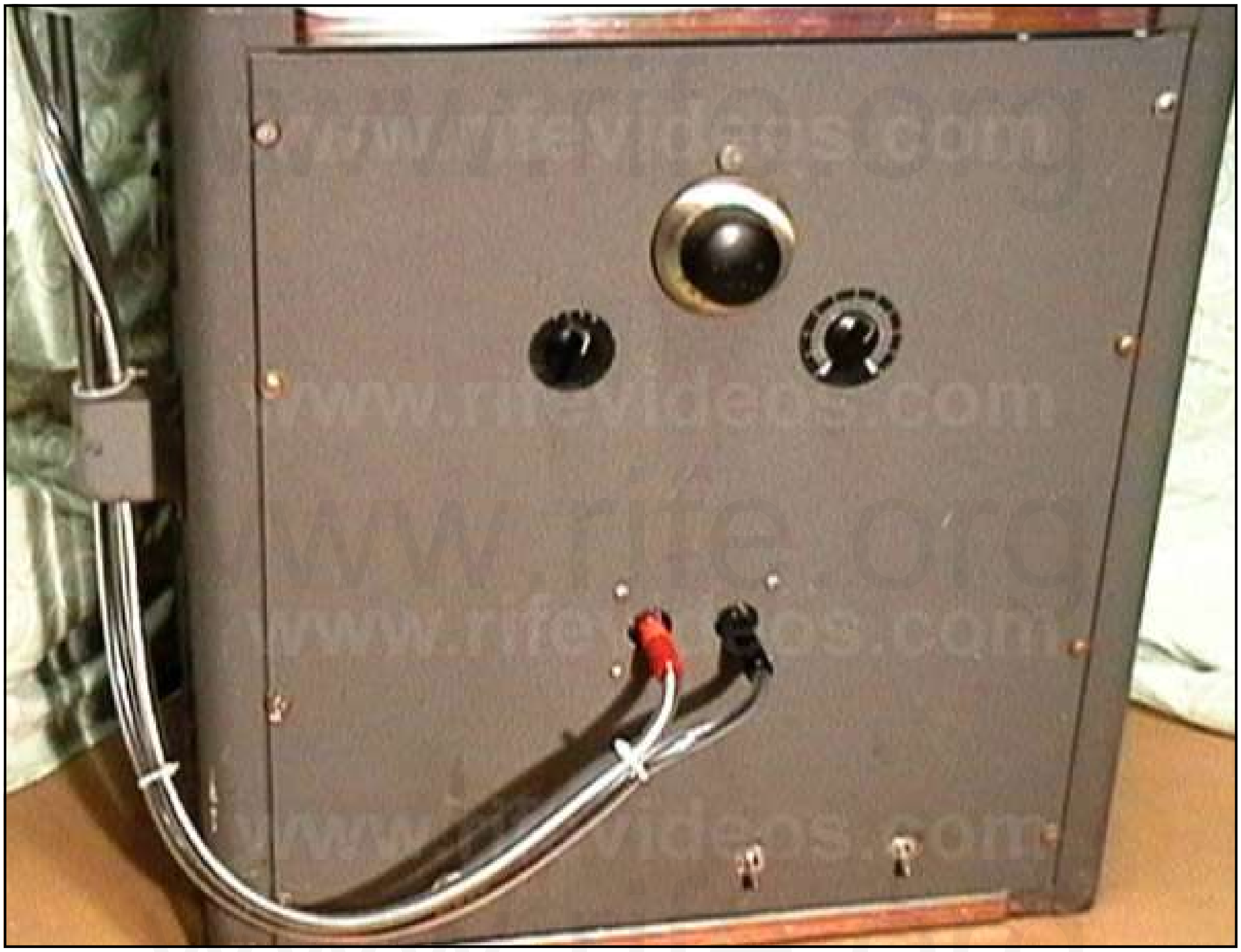

1) Used a Ray Tube.
2) Used a fixed 3.30 MHz carrier frequency.
3) Modulated sine wave audio frequencies onto a sine wave carrier frequency.
4) Power usage was about 460 watts. Output to the ray tube about 40 watts.

The 1939 Beam Ray Machine?

Aubrey Scoon’s instrument is an important part of this report. We felt that his information about this machine should be added and included in Chapter 11. All of this information was originally on Aubrey Scoon’s web site but since he has passed away his website is no longer available. This information is to important to be lost. Even though we now know that his machine was not built by the original Beam Ray Corporation it is still a copy of the original Rife Ray #5 or Beam Ray Clinical instrument. This machine was built by Verne Thompson in 1952 or later and used the original 3.30 MHz RF carrier frequency. It also used Philip Hoyland’s original audio frequencies that will produce the original sideband frequencies that will would hit the higher harmonic frequencies of Dr. Rife’s original M.O.R.s.

Below is Aubrey Scoon's written Report:

In 1938, a group of British researchers headed by Dr Bertram Winter Gonin, sought to buy some experimental machines from Rife to confirm his work. At that time, Rife had no commercial operation capable of handling the orders and at the instigation of his old friend Ben Cullen, Rife consented to the formation of a commercial company called Beam Rays Inc. At that time, the majority owner of the

www.rife.org

rights to the machine was an electronics engineer called Philip Hoyland. Hoyland had designed and developed all of the original Rife machines since late 1934. Hoyland and Rife became partners in Beam Rays Inc. The company produced a small number of machines but suffered internal conflicts because of the actions of some of the other partners and because of frictions between Hoyland and Gonin's group. Hoyland believed that Gonin and his partners were trying to steal the technology for themselves. Subsequently, the company was destroyed in 1939/1940 when Hoyland brought a lawsuit against Beam Rays in an attempt to stop one of the partners from illegal stock trading and also to dissolve the contract with the British. Fuller details of this are being written up and will be posted here in the near future.

The Beam Ray Corporation produced a number of machines. I'm not sure of the exact number as I have seen varying accounts, but I believe that about 17 machines were produced and shipped to MD's in California. Another four machines were shipped to England to Dr. B. Winter Gonin of the London School of Tropical Medicine. There was some argument between Rife and Hoyland over these machines and their operating principle, however various MD's including James Couche used one of these machines for many allegedly successful treatments of patients for many years.

The U.K. Rife Research Group managed to get access to one of the machines believed to be one of the originals that were shipped to California in the late 1930's. I reverse engineered that machine and present the results below. I have a lot of data on this machine of which this is only part, I will add things as I get time as much of the data needs to be organized properly. The exact date of the machine is unknown - we know that it was built sometime between the autumn of 1938 and the early part of 1939 - so it may actually be a 1938 machine.

I would like to express my thanks to Robert Harrison, an expert valve/tube engineer (I'm not a valve/tube expert) who gave me massive help and support during the reverse engineering process and who corrected numerous silly mistakes that I made! My thanks also to Stuart Andrews and Bob Haining for their help and support in this effort.

Machine Schematic

Below are the machine schematics and links to some photographs. The schematics have been rendered into PDF format which allows them to be zoomed to resolve fine detail.

These machine schematics are only included in the online report due to the fact that some of them would not be readable. Please go to the following web address to obtain these schematics:

www.rifevideos.com/aubrey_scoons_1939_rife_beam_ray_.htm.

WARNING: These schematics are presented for information only. Do not try to construct a copy of this machine unless you are knowledgeable and experienced in high voltage valve/tube work. The tube filaments need to be heated for a couple of minutes before switching on the HT - failure to do this will blow them. In addition, some parts of this circuit use up to 1750 volts DC at substantial current and can easily kill in inexperienced hands.

Booster Amplifier Schematic

At the same www.rifevideos.com link, given above, is the booster amplifier schematic. This is for a transistor booster amplifier I designed. It is possible to separate the Beam Ray machine output stage from the oscillator stage and to drive the output stage directly from a modern digital frequency generator. This allows much more precise and accurate frequency control than the original. But the output

www.rife.org

stage requires over 40 V p-p to drive it properly at 100% modulation and most modern generators are not capable of outputting this kind of voltage. So the booster stage allows a low level signal between 20Hz and 200Khz (the original range of the machine) to drive the output at up to approx 66 V p-p without significant signal distortion (it can be driven at higher inputs of up to 2 V p-p with clipping). An input of 951 mV will cause 100% modulation of the Beam Rays machine and an input of 1.5 V p-p will result in significant over modulation without distortion of the modulating signal itself. The booster amplifier has a bandwidth of approximately 800 KHz and so can also be used as a general purpose wideband, high voltage buffer amplifier as well.

Plasma Tube Photos

The first two photographs below show the Nazarov phanotron running on the Beam Ray Machine. And below in the third photo is a close-up of the electrodes of the original plasma tube.

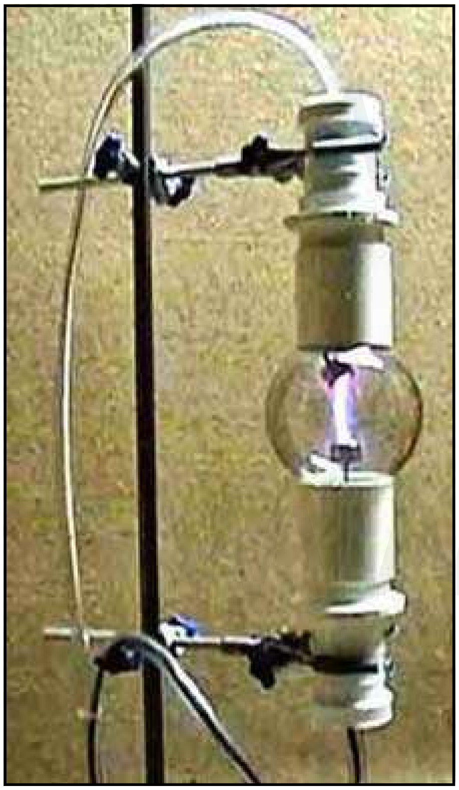

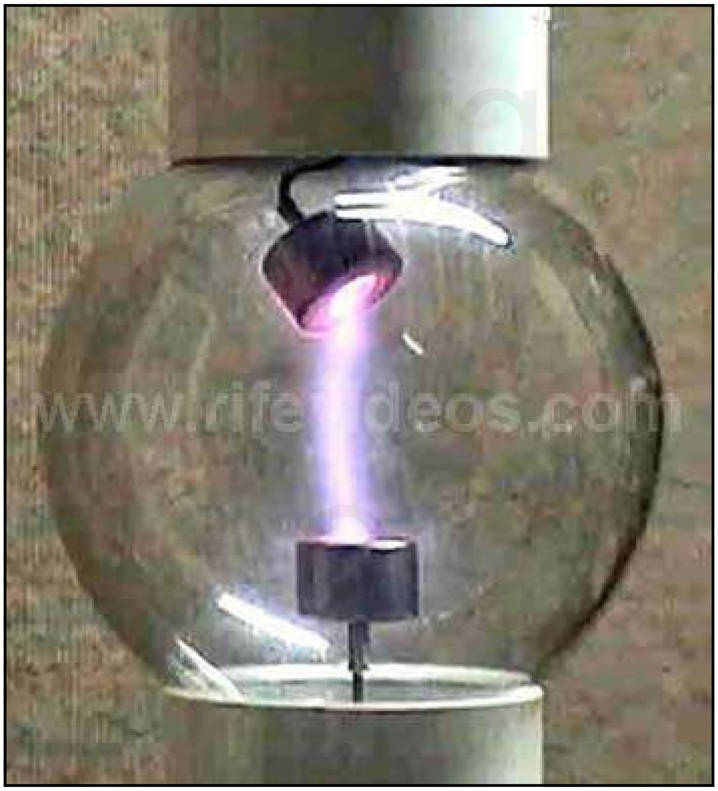

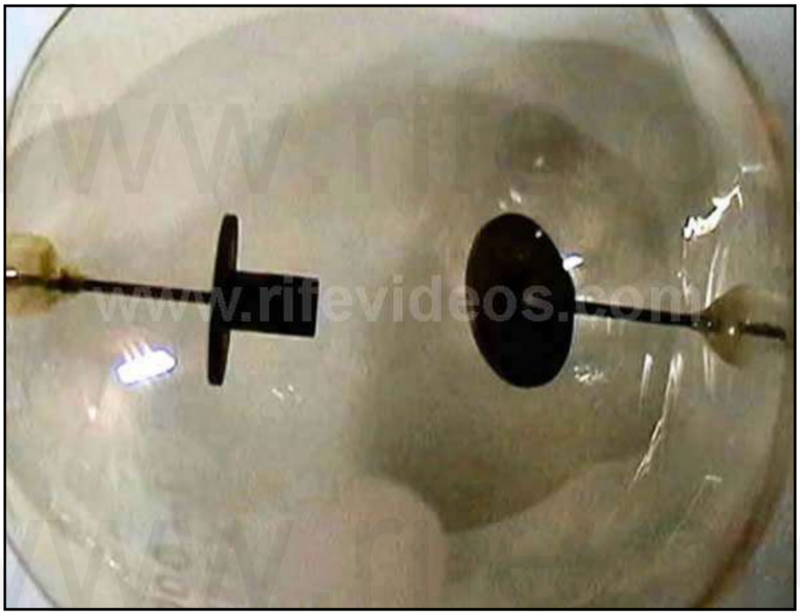

www.rife.org

Inside the Beam Rays Machine

The picture below is of the output stage of the machine. The tank coil can be clearly seen on the right with the 812 (should be 809) output triode next to it. The output jacks are in the center top. The HT transformer is at the back left with the two 866 mercury vapour rectifiers in front of it. Picture courtesy of Stuart Andrews.

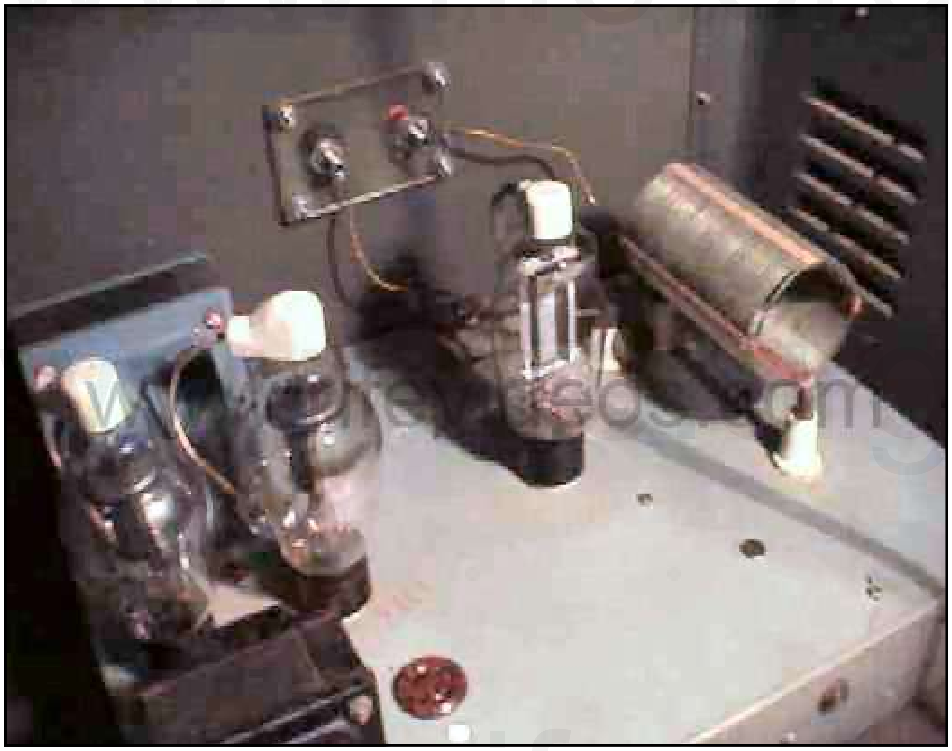

The next photo below is the underside of the output stage (note that some components on both stages had to be replaced with modern ones to get the machine working because the originals had deteriorated too much).

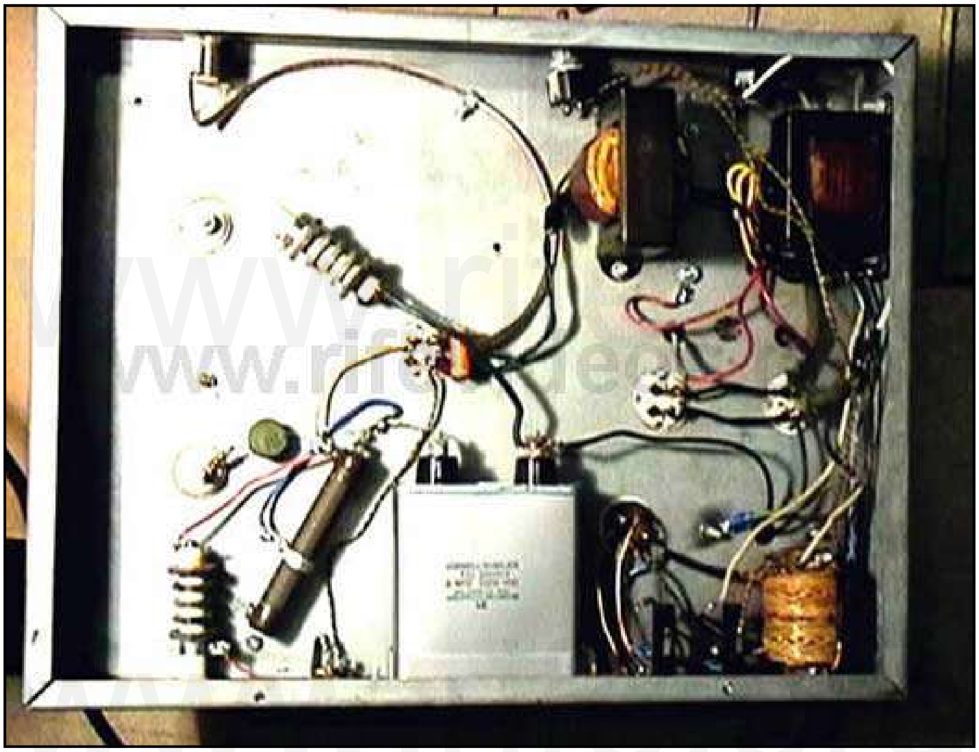

www.rife.org

And this photo below is of the underside of the oscillator stage.

Beam Ray Carrier Waveform Photographs

Here are various photographs I took of the machine's waveforms on an oscilloscope. Note: these were taken with the 812A tube in place. The waveforms produced when an 809 is used are smoother and more sinusoidal.

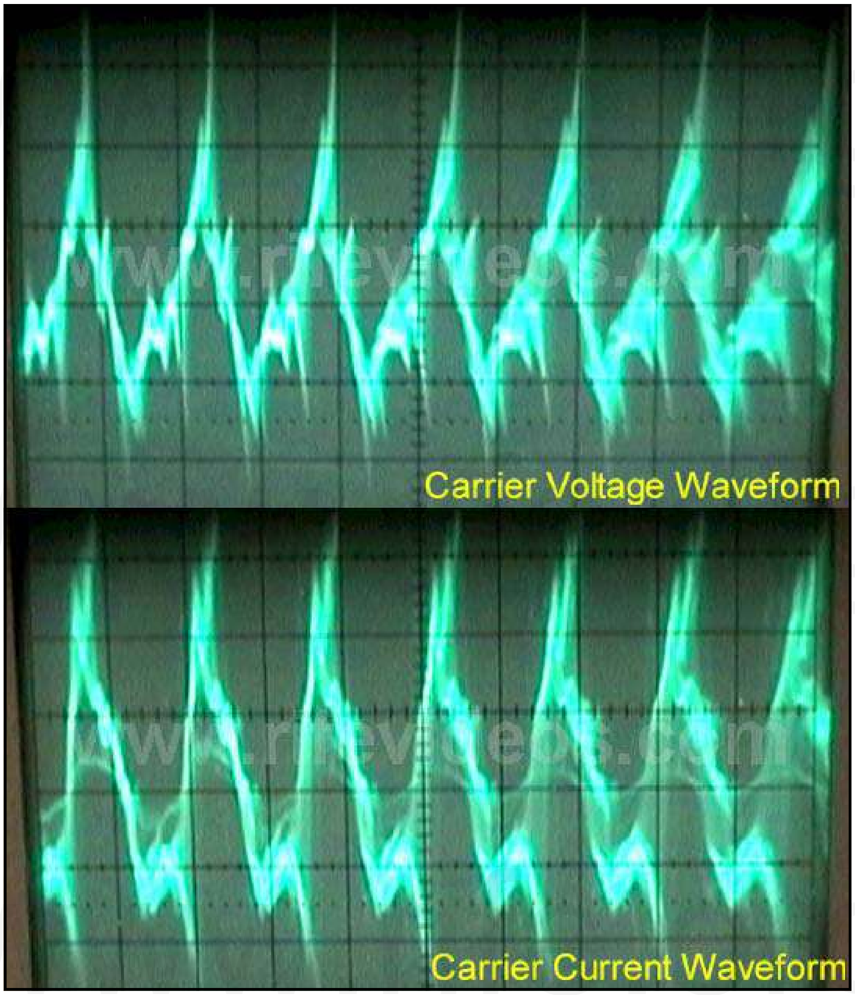

These photos on the previous page are the carrier waveforms for both voltage and current. The actual carrier frequency of the machine depends very much on the plasma tube and also on coupling effects between the tube and any person in the vicinity.

This has made it very difficult to get accurate measurements of the machine because even being in the same room as the machine causes the frequency to alter!

However, after many experiments I am reasonably certain that the resting carrier wave frequency, undisturbed by any local effects is approximately 3.33 MHz.

Because the wave is not a pure sine there are strong harmonics at many other frequencies as well. Some of the dominant harmonics have been observed at approximately 2.3 MHz, 4.6 MHz and 9.09 MHz.

Photo Comparison To Earlier Rife Machine

Comparison of wave from original Rife machine (top) with the waveform from the Beam Rays machine (bottom). Note the similarities in the wave envelopes.

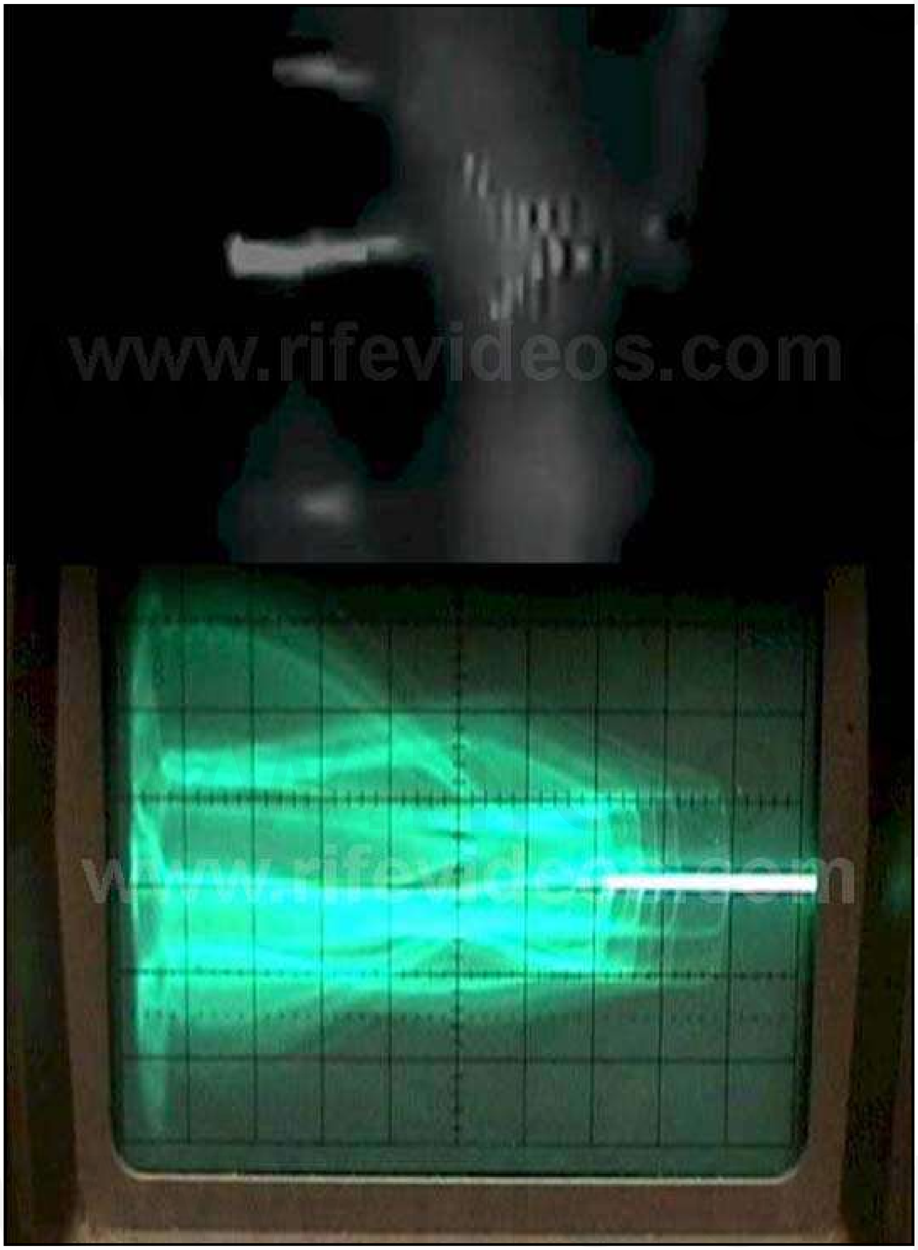

www.rife.org

Effect of Modulation Depth

Wave 1 - this is one burst of the tank waveform at 1430 Hz (scope settings on uncalibrated and adjusted for better resolution) and in excess of 90% modulation depth (sine wave).

Wave 2 through Wave 4, this is what happens as the modulation depth is successively increased - Wave 4 is pretty much 100%. The number of cycles per burst (a burst is approx 20ms apart and is set by the mains cycle) decreases with increasing modulation.

Wave 5 - as you enter over modulation, the burst breaks up into smaller chunks.

Wave 6 - at approx 50V p-p modulation the wave is over modulated and consists of single cycle pulses grouped into bursts.

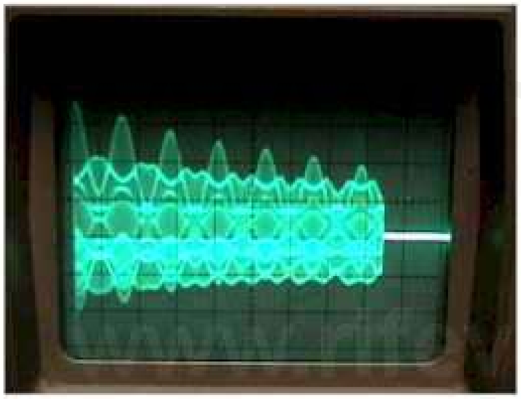

Wave 1

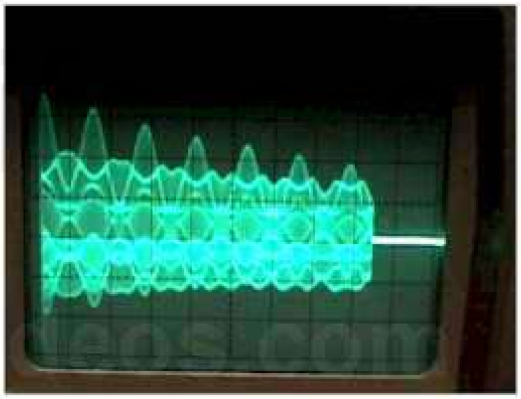

Wave 2

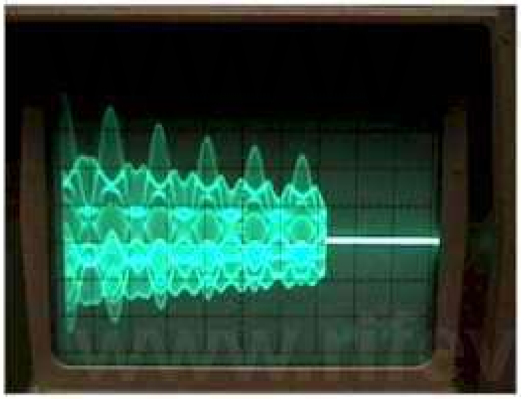

Wave 3

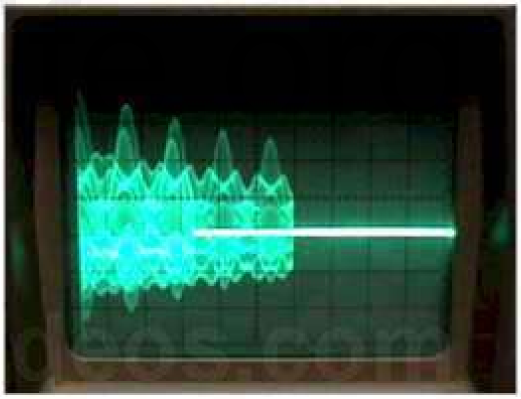

Wave 4

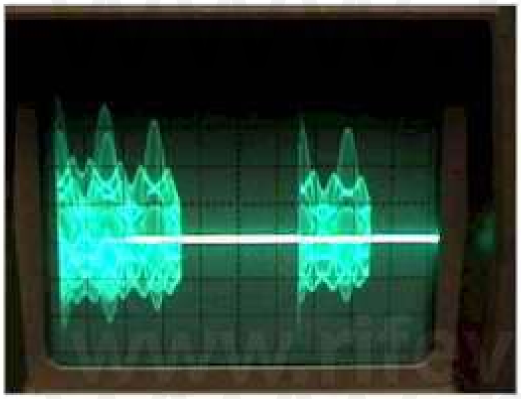

Wave 5

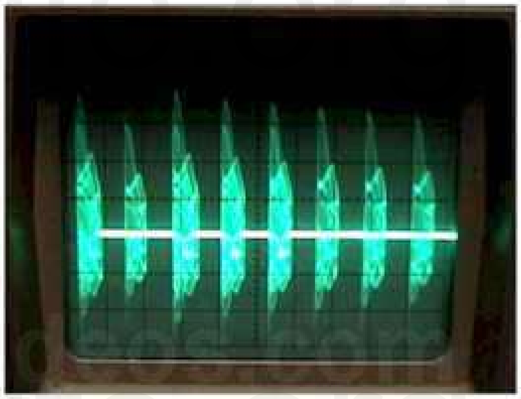

Wave 6

www.rife.org

Analysis Of Beam Ray Machine Operation

Below is a more detailed description and brief analysis of the operation of the Beam Ray machine.

The machine has 3 external controls. The leftmost one is a 4 way rotary switch with positions labeled 1 through 4. These are the (modulation) frequency bands of the machine. The total range of the modulation settings is from approximately 20 Hz to 200 KHz in 4 decades as follows:

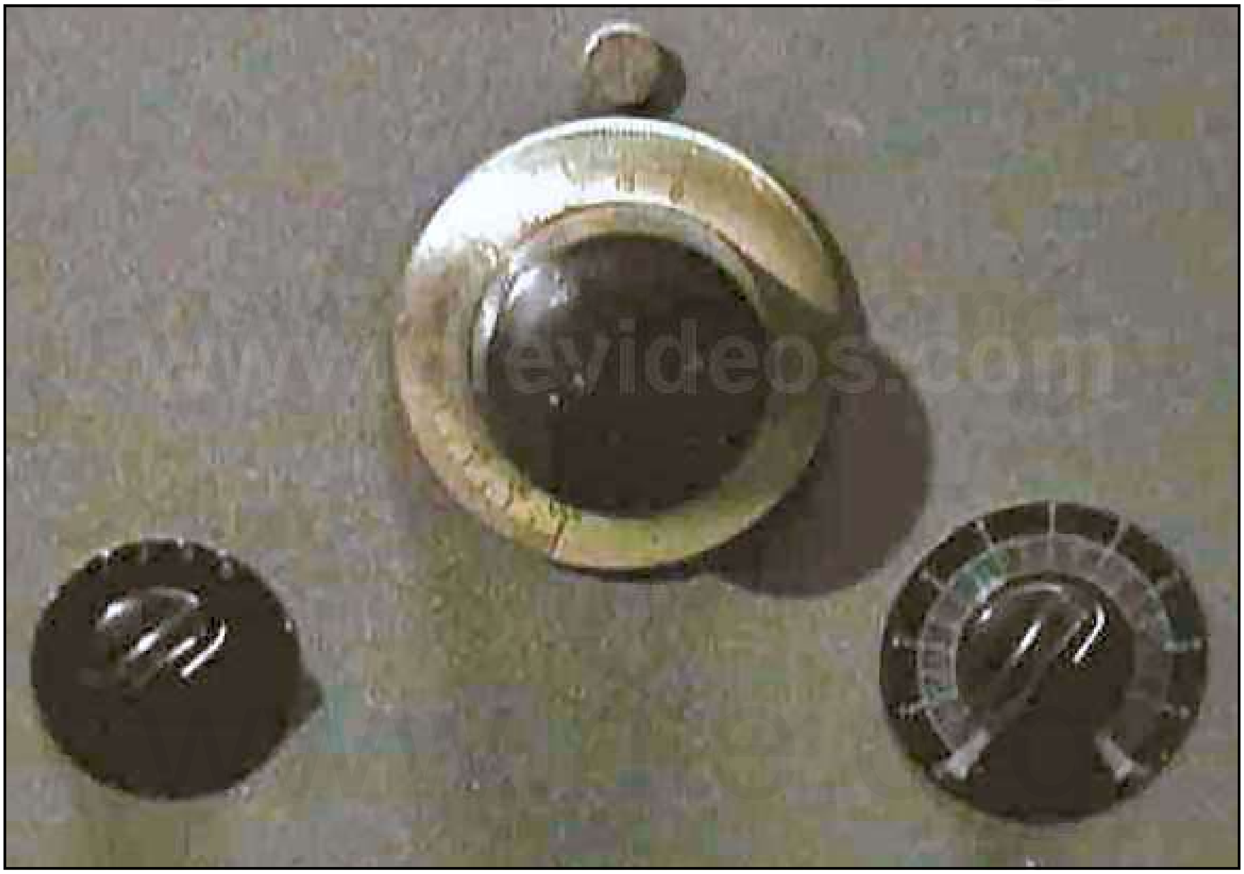

1
2
3
4
www.rifevideos.com

Band Frequency Ranges

Band 1: 20 Hertz to 200 Hertz.

Band 2: 200 Hertz to 2000 Hertz.

Band 3: 2000 Hertz to 20,000 Hertz.

Band 4: 20,000 Hertz to 200,000 Hertz.

www.rife.org

Main frequency Dial Adjustment

The center control is the main modulation frequency dial. This is calibrated in one step units from 0-100. The dial superficially appears to be a vernier dial but isn't, although it is finely marked, of good quality and is geared down, so a single turn does not move it completely from end to end. A resolution of 1/2 a division is easily possible.

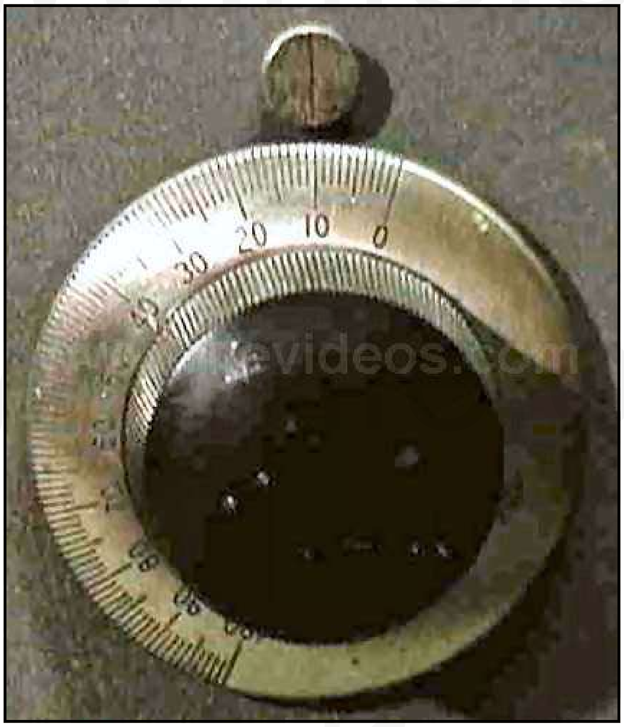

Modulation Amplitude Control Dial

The right hand control is the modulation depth control. It varies the modulation applied to the grid of the output triode from 0 to 50 V p-p. The highest setting of 50V p-p is not 100% modulation, but rather over modulation of about 115%.

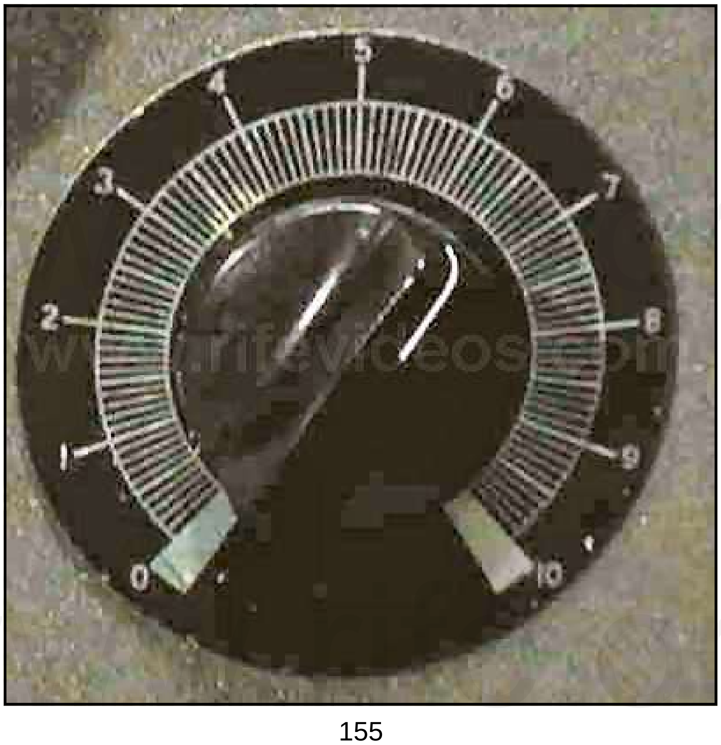

www.rife.org

Treatment Settings

The machine came with some old diary pages with a series of treatment settings scribbled on them. The year of the diary was not shown, just the months but February was shown. It was not a leap year and by matching the days of the week to the dates it had to be 1939. There were the "usual" Rife pathogens plus a few more ailments that I had never seen quoted on any other early Rife machine list. Unfortunately they were not all complete so it wasn't possible to derive the true frequencies for all of them. The anomalies included:

"V" with a setting of band 3, dial 39 - but I have no idea what this is.

"Radiation" with a strange setting of "2-17-3" which could mean band 2 or 3 dial setting 17.

Pain - "20", Iodine Poisoning - "92", Mercury Poisoning - "55", Tissue repair - "58", Scar Tissue - "13", Mucous Colitis - "38", Hemorrhoid - "88", Catarrh - "17", Inflammation - "40" and Irritation - "30".

The ones that could be clearly resolved are shown in the table Below. The treatment frequencies are all approximately 10 times the ones listed in the modern Crane derived frequency lists for the various conditions.

Pathogen	**Band**	**Dial**	**Frequency**
BX Carcinoma	4	10	21275
BY Sarcoma	4	6.5	20080
Typhoid Virus	3	94	18620
Typhoid Rod	3	58.5	6900
Tetanus	2	78.5	1200
Treponema or Syphilis	3	56	6600
Staphylococcus	3	59	7270
Streptococcus	3	63.5	8450
Streptothrix	3	61.5	7870
E Coli Rod	3	62	8020
E Coli Virus	3	89.5	17220
Tuberculosis Rod	3	63	8300
Tuberculosis Virus	3	88	16000
Pneumonia	3	61	7660
Worms	3	24	2400

www.rife.org

Audio Frequency Circuit Analysis

The most striking thing about the Beam Rays audio frequency circuit at first glance is the oscillator section. The machine oscillator is clearly a first generation Hewlett Wein Bridge circuit. What makes this particularly notable is that Hewlett (Hewlett-Packard) only invented the circuit around the time the Beam Rays machine was built. Because it was so new and had not found its way into commercial designs it tends to imply that there might have been some connection between Hewlett and Beam Rays.

I have written a detailed account of the history entitled: "The Hewlett Connection." To read this report go online to: www.rifevideos.com/aubrey_scoons_1939_rife_beam_ray_.htm.

The earlier Rife machines had used the Hartley oscillator circuit which was no where near as stable as the Hewlett Wein Bridge. So the use of this circuit was a big step forward for Beam Rays. The earlier machines had been plagued with apparent frequency instability which made consistent use very difficult. The Wein bridge circuit was an apparent solution, but in practice there was another factor that they apparently did not take into account.

The Beam Rays oscillator is remarkably stable - it drifts by only a few Hertz during normal operation and is superior to many modern analog generators. But it has one major drawback. Tuning is achieved by way of a variable capacitor. This capacitor is connected directly to the tuning dial. The tuning dial is a geared down dial that allows very precise turning of the capacitor shaft. However in practice it suffers from a slight degree of "backlash" - in other words, no matter how carefully you turn the dial, there is always some residual pressure on the rotary shaft - and left to itself for a while this residual pressure or tension causes the capacitor shaft to turn back by a small amount. This is enough to throw the tuning out by a couple of Hertz in the lowest range - and the problem multiplies by a factor of 10 for each higher range. So in the top range the backlash can throw the frequency setting off by approximately 2 KHz. In addition, because the relationship between the dial setting and the actual frequency is non-linear, the problem is always worse toward the top of the scale (i.e. the dial is calibrated

www.rife.org

from 0 to 100 - the problem is much more pronounced near 100 than it is near 0). One of the ways in which this manifests most noticeably is that turning the dial down from a higher number results in a lower overall frequency than turning the dial up from a lower number to the same final setting.

The frequency setting is obviously critical and this may explain why some frequency stability related problems were encountered even with this extremely good oscillator.

The presence of the Hewlett oscillator explains also the general build of the machine. When I examined the machine there were various minor anomalies. One of which was the size of the two chassis. They are both much bigger than they need to be. Also there are two separate chassis and the earlier Beam Rays machines had only one. Another thing is that there is a circuit on the oscillator stage which is not connected and not used. This was probably meant to be able to create square waves by sine overdrive and clipping. Finally the mounting holes on the chassis do not correspond with the mounting holes on the case - someone has drilled new holes in both chassis to remount them in the case. This could mean that the case the machine is in is not the original case it was shipped in. Alternatively it could mean that the chassis were previously mounted in a different way (maybe in a different case) prior to shipping.

Taking all these things together leads me to the conclusion that the machine I examined is probably an original prototype. This makes sense considering that the Hewlett oscillator was so new. Obviously, Phillip Hoyland or whoever built the machine decided to try making the oscillator stage separately from the output stage. The output stage is presumably the same as in earlier Beam Rays machines (it looks the same as a 1937 Beam Rays machine), but clearly they decided to build the oscillator on a separate chassis. The chassis were overlarge to allow for circuit expansion and modification. The unused circuit was probably tested at some point and found to be unnecessary. And it is likely that during development the different mountings were used on some sort of open frame to allow testing and measurement.

But this prototype machine was shipped and sold as a finished unit, why? The answer is probably quite obvious. The machine was produced during the great depression. Everything was expensive; Beam Rays was a small company and needed to keep down costs. The prototype was probably only needed during development. Once everything had been worked out satisfactorily, the prototype was no longer needed - and could be sold for a substantial profit, as it was effectively a working machine of a new design.

The 6SJ7 and 6K6(B) tubes on the combined schematic are the basic Hewlett Wein Bridge oscillator circuit. See "The Hewlett Connection" for a schematic of Hewlett's patented design. The third tube marked 6K6(A) is a simple cathode follower buffer stage, analogous to a modern transistor emitter follower circuit. This circuit has high input impedance, low output impedance and unity gain. It is designed to insulate the sensitive oscillator section from the following output stages. The 6SN7 tube is the unused circuit and as mentioned above was probably meant to be part of a fast clipper amplifier to produce square wave modulation.

The oscillator stage creates a pure sine wave from approx 20 Hz to 200KHz depending on range and dial setting. It also produces a variable amplitude output which can be adjusted from 0v right up to approx 50V peak to peak.

The output stage consists mainly of a single power triode. Although the machine had an 812A triode in it when I got it, I believe the correct original tube was an 809. The machine runs a lot more cleanly and stably with an 809 than an 812A. The stage is self-oscillating; it has a simple regenerative feedback arrangement from plate to grid via two capacitors and the tank coil. The degree of feedback can be adjusted by means of a large power resistor from the grid to ground.

www.rife.org

Note: the machine was not actually grounded, the negative end of the supplies connected to the chassis and all "grounds" were actually referenced to the chassis. I found in practice that the chassis did tend to accumulate quite a nasty residual charge after the machine had been operating and so I grounded it which did not seem to affect the operation of the machine. The output stage is actually a Hartley oscillator, although not obviously so, because the output capacitor in series with the plasma tube capacitance represents the "tuning" capacitance of the circuit. The large power resistor in series with the tube affects the oscillator loading, the output field impedance and also the Q of the resonant circuit. Because the plasma tube is an active circuit element, capacitive coupling from anybody in the vicinity of the plasma tube actually causes changes in the oscillator frequency. The "resting" frequency of the output oscillator is around 3.30 MHz using the 812A and an Argon (Nazarov) phanotron tube. When the tube was was changed to an 809 and a 15mm Helium Cheb phanotron was used, the "resting" frequency changed to 4.68 Mhz and the wave became much more sinusoidal .

The output from the modulation oscillator stage is capacitively coupled to the output triode grid via an inductor. The latter is designed to prevent the carrier oscillations from feeding back into the modulation oscillator stage.

The DC HT power for the output triode is derived from a 1235 VAC plate transformer by two 866 mercury vapour rectifiers. The DC output is smoothed via a large choke and a filter capacitor to ground. There is also an RFC choke in the line to the plate. The net voltage at the plate of the triode is only around 550V DC which is consistent with a choke smoothed circuit. However much more interesting is the other end of the tube - the filaments (which double as cathodes) are connected to a direct AC filament heater transformer which means that there is an additional modulation at the 60Hz mains frequency (cathode modulation).

In practice during operation, the machine creates the carrier waveform (which is not very clean and contains a lot of harmonics - it has a superficial similarity to a sawtooth wave).

UPDATE: the carrier waveform is smoother with the 809, but still contains some strong harmonics. The carrier is then amplitude modulated by the sine wave produced by the modulation oscillator. However, in addition, the AC cathode connection causes a further modulation at 60Hz. In effect the modulated wave is chopped into chunks or bursts that are one period of a 60 Hz cycle apart. And the envelope of the wave is effectively the first quarter cycle of a 60Hz cosine wave. In some respects this is like a very crude approximation to a damped wave. I believe that the latter is not a design flaw but rather a feature.

In the next chapter we will discuss the 1953 AZ-58 built by Verne Thompson and the history of that instrument in detail.

www.rife.org

Chapter #12

Dr. Rife and Verne Thompson's 1950's AZ-58 Beam Ray Replica Rife Machine

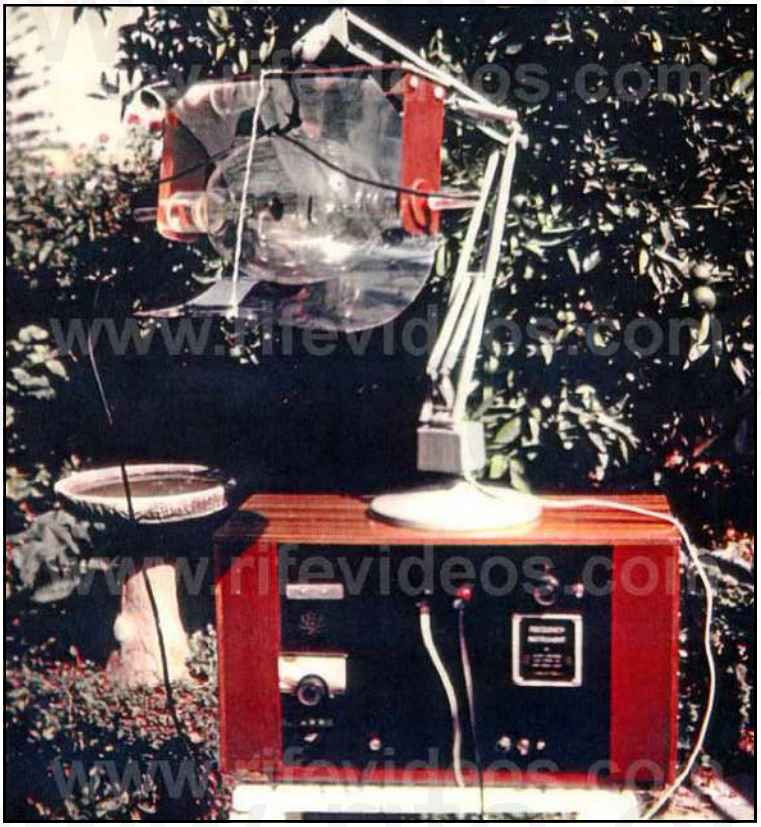

1. Used a ray tube.
2. Variable RF carrier frequency between 2.20 and 5 MHz. (First used 4.68 carrier).
3. Modulated square wave audio frequencies onto the sine wave carrier frequency.
4. Power usage was about 460 watts. Output to the ray tube about 40 watts.

With the knowledge gained from obtaining the original Rife Ray #5 or Beam Ray Clinical instrument we know that the 1953 AZ-58 design is an updated replica of Philip Hoyland's Beam Ray Clinical Rife Machine. The 1953 version was updated with newer component parts. Dr. Rife's engineer, Verne Thompson had been building this type of Rife Machine during the 1940's and 1950's. We will now cover this information in more depth. This style of instrument worked well as long as it was using the sideband method. From the stress of the Beam Ray Trial, Dr. Rife became an alcoholic and all that he had worked so hard to accomplish was almost destroyed. Many of the doctors had returned their instruments because of AMA threats. Some of these returned Beam Ray Clinical instruments Dr. Rife parted out and sold to anyone who wanted radio parts. Under these circumstances Beam Ray Corporation eventually closed its doors. Verne Thompson became Dr. Rife's engineer around 1940 and built the Beam Ray style Rife Machines during the 1940's and 1950's for anyone who wanted an instrument. The documents show that back in 1939 Dr. Rife wanted to go back to the original principles of his instrument rather than use Philip Hoyland's harmonic Beam Ray Clinical design. This is what Dr. Rife said in a 1939 letter sent to Dr. Gonin:

RIFE: *"I spoke only Friday evening to a Mr. John Chamblin, a radio man now connected with Beam Ray Inc., about the redesign and building of a device according to the old Rife Ray principles; as the present instrument has been so deviated away from that old principle that it is nowhere near the* *same...those devices which you have are merely working on a harmonic and not a true frequency; and in our research on electronics, we definitely know that there is no possible way of controlling electrical harmonics of a frequency."* (Letter from Dr. Rife to Dr. Gonin, May 14, 1939. Page 1 of 3).

Because Dr. Rife didn't understand Philip Hoyland's harmonic sideband method he believed that the harmonics could not be controlled. Sidebands can be controlled by the audio frequency which is used to create them. Had Dr. Rife been given the understanding of how the instrument worked he would have known that they were controllable. Because of this misunderstanding Dr. Rife wanted to go back to the original designs of his earlier Rife Machines. It is also a historical fact that he never did go back to using his original designs like the Rife Ray #4. There is a reasonable explanation for this. Back in the late 1930's and early 1940's the FCC began controlling the airwaves and hundreds of new radio stations were being granted licenses. The RF M.O.R. frequencies which were output by the Rife Ray #3 and Rife Ray #4 were almost all in the A.M. radio band of frequencies and would interfere with these new broadcasting stations. These events probably brought about the continued use of Philip Hoyland's Beam Ray Clinical Rife Machine since it operates out of the A.M. band and in the low ham radio range. In Chapter 9 of this report we also read that Dr. Rife tested Philip Hoyland's instrument, in about 1940, when he had Verne Thompson rebuild Dr. Yale's Beam Ray Clinical instrument. In those tests Dr. Rife found that Philip Hoyland's design would devitalize all the microorganisms. The tests that were done by Dr. Rife showed that this style of instrument worked. This meant that he really didn't need to go back to his earlier designs even though he may have wanted too.

There are a few important facts that should be pointed out. Dr. Rife at any time could have had an original Ray tube instrument built. He had a Rife Ray #4 in his lab that could have been copied. I am sure that he also had access to the Beam Ray Laboratory instrument. These facts, I believe show that it was probably because he knew that Philip Hoyland's design was a good design that he continued to use it. It could also be because of the FCC regulations Dr. Rife either had to use the Beam Ray Clinical instrument design or no longer build one. We know, from the documents, that Verne Thompson was copying Dr. Couches' instrument with its fixed RF carrier frequency of 3.30 MHz. This carrier frequency was well above the A.M. band of radio stations. Since the Beam Ray Clinical instrument used the higher harmonics of the original Rife Ray #4 frequencies this would not cause any problems with the A.M. radio stations. The FCC was, at the least, a contributing factor in why Dr. Rife continued to use Rife Ray #5 or the Beam Ray Clinical design.

With the understanding of this information let us continue on with the history of Dr. Rife's instruments. John Crane met Dr. Rife in 1950 when he inquired about purchasing a drafting set that Dr. Rife was selling. John Marsh met John Crane in 1952 when they were both working at Convair. See photos below. John Crane is the photo on the left and John Marsh is the photo on the right.

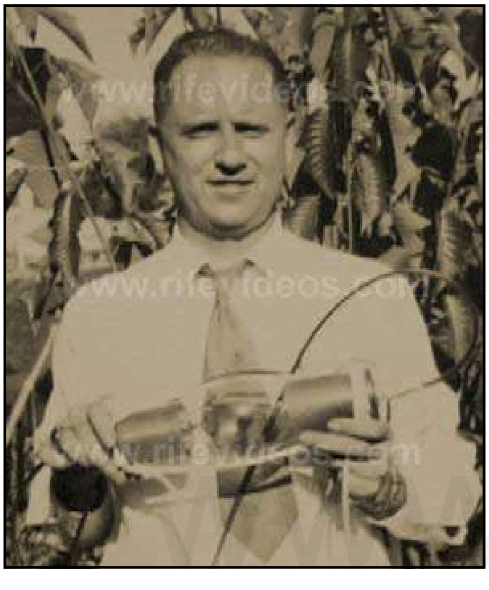

www.rifevideos.com
www.rifevideos.com
www.rifevideos.com
www.rifevideos.com

www.rife.org

They both went to talk to Dr. Rife in 1953 to see if he would be willing to help with John Marsh's wife who had cancer. In Dr. Rife's 1961 deposition #123-125 we have the verification that Dr. Rife gave John Crane his frequencies in 1950:

COMPARET: **"***Did you ever explain to John F. Crane, one of the defendants in this case, the principles upon which your electronic frequency-generator is used in the treatment of diseases?"*

RIFE: "*Yes in 1950."*

COMPARET: "*Did you also inform him of the particular frequencies which you had found to be effective in the treatment of various diseases?"*

RIFE: "*Yes. Verne Thompson and I gave the frequencies to John Crane."*

COMPARET: "*When did you furnish him with this information?"*

RIFE: "*In 1950."*

Dr. Rife always said that almost all his frequencies were in the upper bands (139,200 Hertz to 1,607,450 Hertz) so we know that he gave these high RF frequencies to John Crane in 1950. These frequencies would have been the frequencies used in the Rife Ray #4. The audio frequencies that were used in Dr. Couche's Beam Ray Clinical instrument also may have been given to John Crane at this time. There is an interesting correlation of these audio frequencies that needs to again be mentioned at this time which ties these machines together. The higher audio frequencies, which produce the side-band spacing, used in the 1950's Aubrey Scoon, Verne Thompson Beam Ray Clinical instrument were 10 times higher than those used in the Verne Thompson 1953 AZ-58 Beam Ray Clinical replica. It is apparent that these higher audio frequencies were lowered in the 1950's by Dr. Rife, John Crane and John Marsh and were used in the AZ-58 replica instrument. This gives us the knowledge that these audio frequencies came from Verne Thompson's reading of one of these Beam Ray Clinical instruments. The chart, shown below, is a list of the lowered audio frequencies which were used in the 1953 AZ-58.

AZ-58 Lower Audio Frequencies Derived From Aubrey Scoon's Beam Ray Replica Instrument.

B or E Coli Rod	800Hz	Streptothrix	784 Hz
B or E Coli Virus	1552 Hz	Syphilis or Treponema	660 Hz
BX Virus Carcinoma	2128 Hz	Tetanus	120 Hz
BY Sarcoma	2008 Hz	Tuberculosis Rod	803 Hz
Gonorrhea	712 Hz	Tuberculosis Virus	1552 Hz
Pneumonia or Spinal Meningitis	776 Hz	Typhoid Rod	712 Hz
Staphylococcus	727 Hz	Typhoid Virus	1862 Hz
Streptococcus	880 Hz		

The original Beam Ray Clinical instrument we have has an RF carrier frequency of 3.80 MHz. Aubrey Scoon's Beam Ray Clinical replica instrument's RF carrier frequency is 3.30 MHz. The audio frequencies Aubrey Scoon's instrument used, we showed earlier in this report, will not work with the original Beam Ray Clinical instrument that we have because the audio frequencies have to be matched to the RF carrier frequency. The AZ-58's RF carrier frequency was set at 4.68 MHz as written down on the schematic and as per the FCC. Aubrey Scoon's Beam Ray Clinical instrument audio frequencies would not produce the correct sideband spacing with the AZ-58 RF carrier frequency because the audio frequencies are calculate to work with a 3.30 MHz RF carrier frequency not a 4.68 MHz RF carrier frequency. This also verifies to us that Philip Hoyland was using different RF carrier frequencies in the Beam Ray Clinical instrument. It also indicates that both Rife and Verne Thompson did not understand the importance of using the correct RF carrier frequency with the correct audio frequencies.

John Crane said this in his "Crane Report":

CRANE: *"The instruments were completed by Crane and Thompson in 1953, but the test results were negative."* (The Crane Report" page 6).

With the carrier frequency set at 4.68 MHz as per their FCC license, instead of 3.30 MHz, the audio frequencies they used would not make the correct sideband spacing and the instrument wouldn't have worked like they expected. Logically, if they understood the relationship of the RF carrier frequency and the audio frequencies they would have change the RF carrier frequency back it to 3.30 MHz or recalculated new audio frequencies to work with the 4.86 MHz RF carrier frequency, but they didn't do either.

Even with the wrong RF carrier frequency Dr. Rife, John Crane and John Marsh would have first tried the higher audio frequencies in the sine wave waveform like the original Beam Ray Clinical instrument used. Apparently when this didn't work, because the carrier frequency was wrong, they lowered the audio frequencies by a factor of 10 times and then changed the sine wave waveform to a square wave waveform. This appears to be the first time square wave was used. This change from sine wave to square wave wasn't really necessary since the circuit design will create a square wave looking waveform out of a sine wave anyway. The photo, shown below, is of AZ-58 waveform.

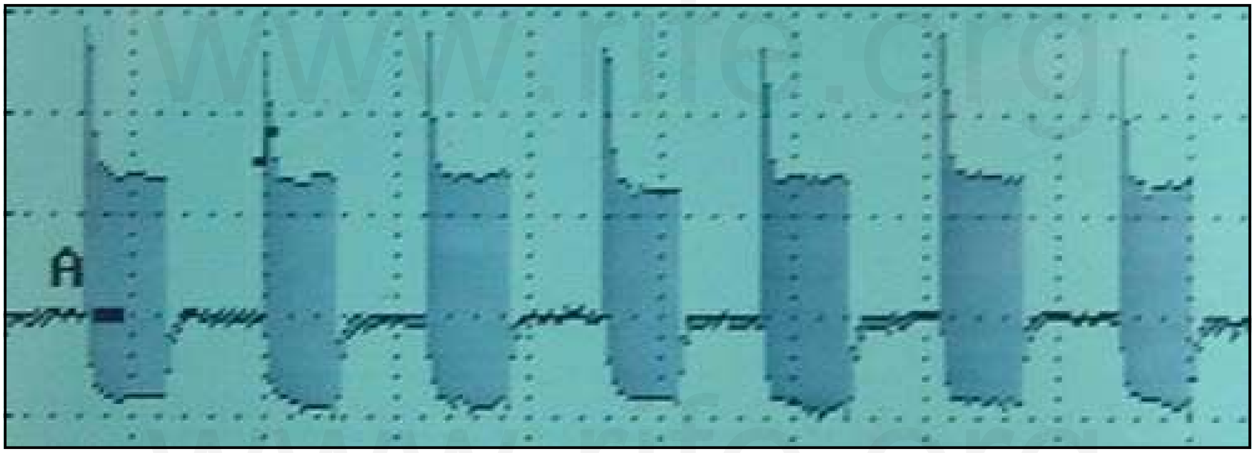

Using the square wave waveform would not have changed how the sidebands would have worked. But it would have created true square wave harmonics which the instrument did not have when it was using the sine wave waveform. John Crane mentions that this is when the instrument started to work better. So we are left to assume that he was referring to the use of the square wave waveform. It is ironic that they used a square wave waveform that produces harmonics in order to get the instrument to even begin to give them any results. Had they really understood how the audio frequencies interacted with the RF carrier frequency they would have recalculated the audio frequencies to work with the new 4.68 MHz carrier frequency instead of lowering the audio frequencies by a factor of 10 times. Then the square wave waveform change would not have been necessary. These important facts that have been pointed out again show they did not understand the relationship between the RF carrier frequency and the audio frequencies. All they managed to prove is neither the carrier frequency nor the audio frequencies will do anything by themselves.

www.rife.org

This change from sine wave to square wave also created a whole new methodology of using frequencies. Instead of relying on the interaction of the sine wave audio frequency with the sine wave RF carrier frequency to produce the correct sideband harmonic M.O.R. to devitalize the organisms they were now depending only on the harmonics from the square wave waveform to hit the correct M.O.R. frequency. This change was also another indication that they didn't understand how the Beam Ray Clinical instrument worked.

Another thing that shows that they did not understand how the Beam Ray instrument worked is the fact that they didn't recalculate the audio frequencies to properly use the square wave harmonics to hit the high frequency M.O.R.s. This shows that they believed that the lower audio frequencies, which they were using, were the M.O.R.s. of the organisms even though they had no harmonic relationship with the original M.O.R.s. used in the Rife Ray #4. We must point out that the square wave mathematical method we are talking about here is completely different than the harmonic sideband method which Philip Hoyland used.

In many of the documents we find that both John Crane and John Marsh firmly believed that these lowered audio frequencies were the M.O.R. frequencies that would devitalize the organisms. Even Dr. Rife appears to have believed that the audio frequencies were the M.O.R.s. as long as they were used with Philip Hoyland's Beam Ray Clinical design. In Dr. Rife's original equipment like the Rife Ray #3 and the Rife Ray #4 he knew that the M.O.R. frequencies were the high RF frequencies. But since Dr. Rife didn't really understand exactly how Philip Hoyland's instrument worked it appears that he accepted the concept that the audio frequencies were the M.O.R.s, or created the M.O.R.s, as long as they were used with that style of instrument. Without understanding these facts it is hard to understand the rest of the history of Dr. Rife's instruments. The fact that they believed that the audio frequencies were the M.O.R.s. compromised the effectiveness of the 1953 AZ-58 Beam Ray Clinical replica.

Below and on the next page are four photos of ray tube instruments which Verne Thompson built. The first instrument was Aubrey Scoon's 1950's Beam Ray replica. The second instrument is believed to have been built in 1947 but some of the transformers are 1960's vintage which indicate it was built in the 1950's or 1960's. The audio frequencies used in it were even lower than the 1953 AZ-58. Since all the evidence we have shows that the lower audio frequencies began with the 1953 AZ-58 this indicates that the instrument was not built in 1947 but was built sometime after 1953. The third and fourth photos, on the next page, are of two of the AZ-58 instruments built by Verne Thompson in 1953 for Dr. Rife, John Crane and John Marsh of Life Labs Co.

www.rifevideos.com

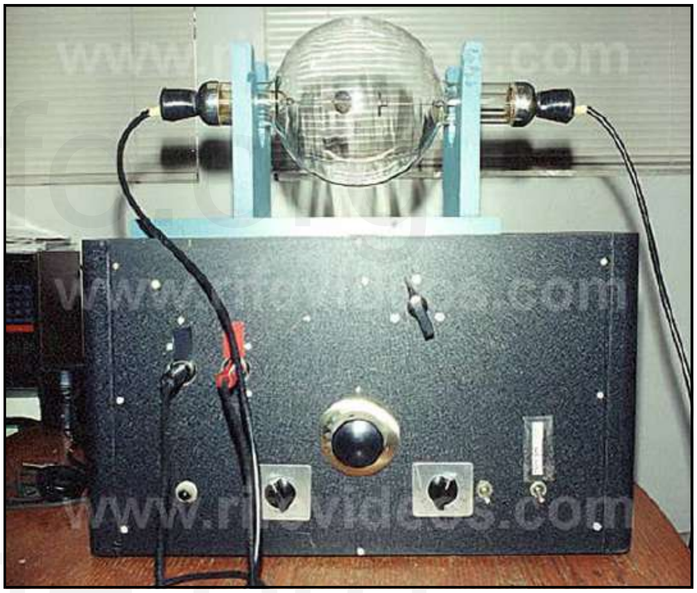

www.rife.org

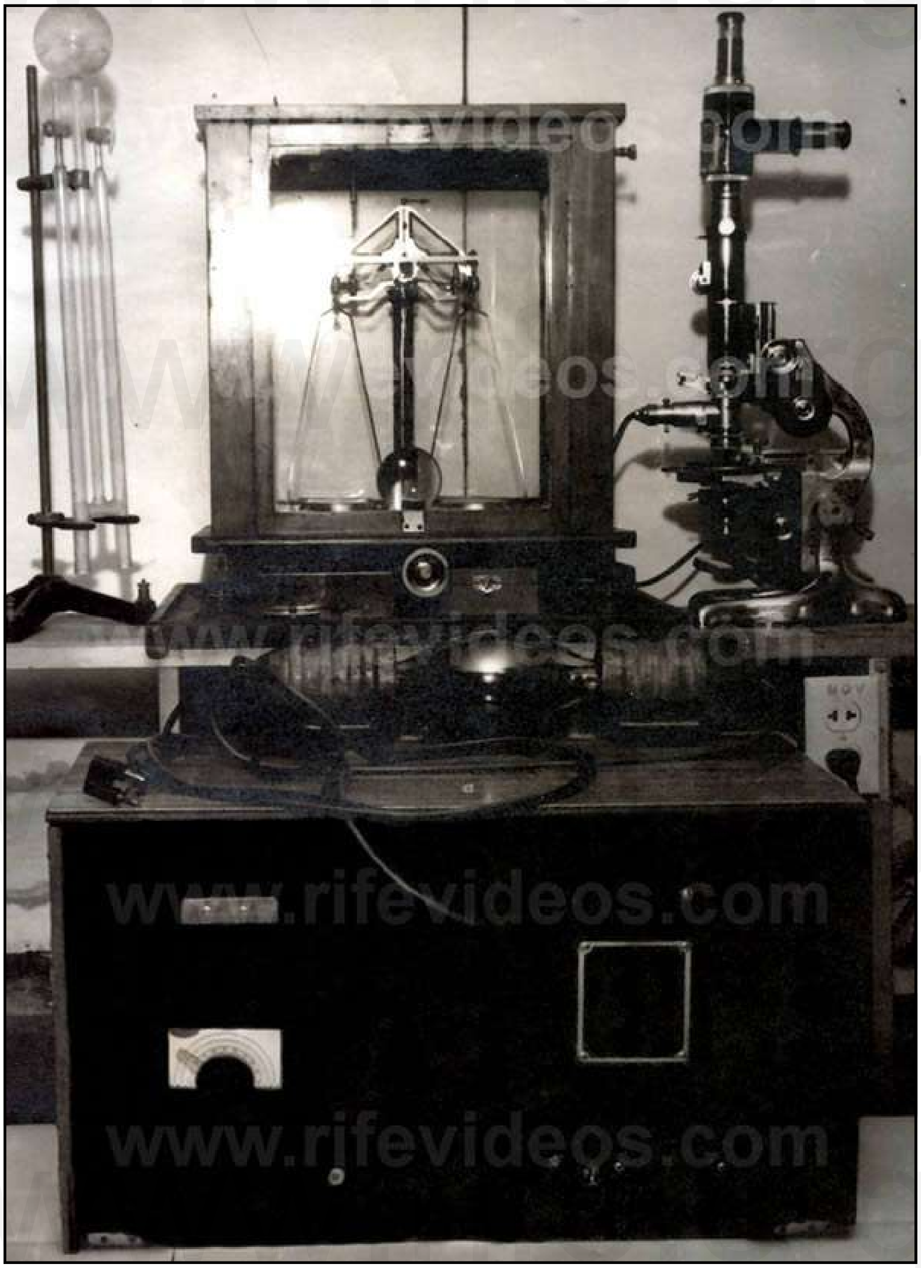

www.rifevideos.com
www.rife.org

All four of these instruments, like the Beam Ray Clinical instrument, had the fixed Hartley Oscillator section which created the RF carrier frequency. Verne Thompson changed the 809 tube to the 812 tube as he updated the instruments from the 1940's to the 1950's. All three instruments also had one main frequency dial for adjusting the audio frequencies. The other two smaller dials were for adjusting the modulation amplitude of the audio frequencies and changing the audio frequency bands. These bands would take you through the various audio frequency ranges. The AZ-58 was limited to about 6000 Hertz, using 3 bands, because they lowered the audio frequencies. The original Beam Ray Clinical instrument we have went to just over 40,000 Hertz using 4 bands. The 1950's Aubrey Scoon Beam Ray Clinical replica instrument could go to about 200,000 Hertz using 4 bands. Other than the bands there is very little difference in the way these instruments worked. We built two of the three, tested them, and found no significant difference other than the use of sine or square wave audio frequencies. At any time the AZ-58 could have been changed to work on the original Beam Ray Clinical sideband method just by adding a fourth band and then recalculating the audio frequencies to work with the 4.68 MHz RF carrier frequency. As pointed out before we made some tests with the spectrum analyzer to see if square wave could be used instead of sine wave in creating the proper sideband spacing and found that the square wave waveform produces the same sidebands as the sine wave waveform.

The basic design of Philip Hoyland's Beam Ray Clinical instrument is very apparent when you compare the schematics. Six of the 1953 AZ-58's were originally built by Dr. Rife, John Crane and John Marsh. We decided to test the AZ-58 with the spectrum analyzer. Below in the graph you can see the results of these tests.

Testing done with PicoScope 3205 spectrum analyzer at 2,400,000 Hertz using AZ-58

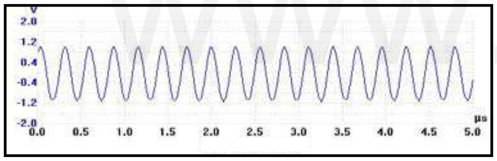

Sine wave out of AZ-58 at 2,400,000 Hertz.

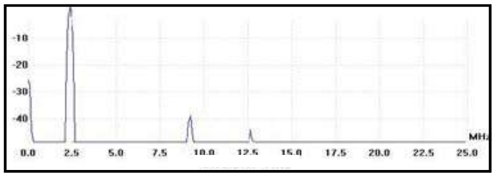

AZ-58 without ray tube at 2,400,000 Hertz measured with spectrum analyzer showing no real harmonics.

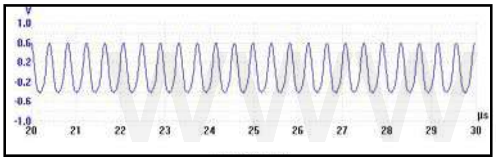

Sine wave carrier out of AZ-58 at 2,400,000 Hertz using ray tube. Sine wave carrier was always distorted when put through a ray tube.

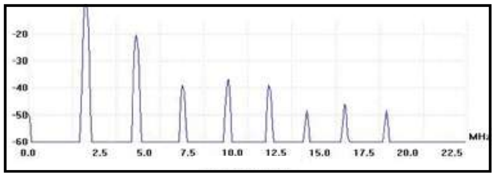

AZ-58 at 2,400,000 Hertz using ray tube. Measured with spectrum analyzer showing harmonics all the way up to 20,000,000 Hertz.

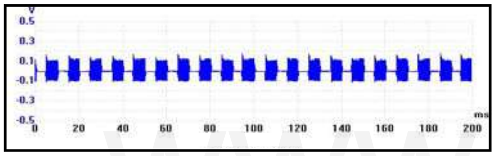

AZ-58 at 2,400,000 Hertz using ray tube. Showing 50% square wave audio frequency modulation. The square wave shows some distortion.

www.rife.org

The instrument called the AZ-58 is significant because we have more information about what took place in the 1950's than any period before this. At this time in the 1950's John Crane and John Marsh recorded interviews with Dr. Rife, Dr. Couche, Ben Cullen, Henry Siner and many other individuals who were key players in the early years of Dr. Rife's work.

We will now cover in more detail the history of the 1953 AZ-58 instrument so we have more information about this instrument. Some of the information comes from the John Marsh Collection of Rife audio CD's. As we said, in 1950 John Crane met Dr. Rife and in 1952-53 he met John Marsh. John Marsh became John Crane’s supervisor at Convair Aeronautics when John Marsh moved from Tucson, Arizona to California. John Marsh’s wife had cancer and they were not able to help her in Tucson. The doctors recommended that he take her to San Diego for specialized care. John Marsh and John Crane became friends. John Crane told John Marsh about Dr. Rife and they went to see him. Dr. Rife gave them an old Beam Ray Clinical instrument which they had Verne Thompson repair. John Crane and John Marsh then used this instrument on John Marsh’s wife and after several treatments John Marsh said she fully recovered. Here are John Marsh's statements which he made in 1976 and 1986:

MARSH: (1976) *“I met this Rife. I said Dr. Rife, I said, my name is John Marsh, I’ve got a wife that’s dying. She’s got cancer of the uterus.” Dr. Rife said: “I won’t touch that thing with a 20 yard pole.”*

After some discussion Dr. Rife said:

RIFE: *“I have an old instrument down here in the basement.”*

MARSH: *“I dug up that old instrument and of course it had tubes in it, antique stuff, and so I rebuilt the darn thing.”* (John Marsh Rife CDs, CD 10 track 1)

MARSH: (1986) *“I went to see him* [Dr. Rife], *and I talked with him and he said he didn’t want to have any part of it...I said look, I got a wife that’s dying and I need your help! And so I got him out of his cocoon, so to speak, and we took an old instrument and rebuilt it. And I treated my wife and by darn all the pain left her and she got well.” In another part of the tape he said: “I discovered that this Dr. Rife was a very great individual...I told John* [Crane], *I said look if we have any of those old instruments laying around loose, let’s rejuvenate one of them and see if we can get my wife well. Well Verne Thompson who was with the San Diego police department radios, uh, radio expert, uh, had built some instruments and they were antiques when I saw them.”* (John Marsh Rife CDs, CD 2 track 3).

John Marsh and John Crane then decided they would like to work with Dr. Rife and try to get the frequency instruments rebuilt and back into doctor’s hands. They wanted to help people who were suffering from many incurable diseases. From earlier quotes we learned that Verne Thompson had worked on Dr. Yale’s Beam Ray Clinical machine. He knew how to build these instruments and this is why John Marsh and John Crane had him repair the instrument that Dr. Rife gave them. It is apparent that Verne Thompson knew this instrument’s circuit designs very well but he didn’t understand how the Beam Ray Clinical instrument was using harmonic sidebands to hit Dr. Rife's higher harmonic M.O.R.s. From a patent application that Dr. Rife and Verne Thompson filed we know that Verne Thompson also believed the audio frequencies were the M.O.R.s. This document leads us to only one conclusion; no one in the 1950’s understood how Philip Hoyland’s sideband spacing method worked.

John Crane in his later years (1970's & 1980’s) was not being very truthful in some of his statements. He said this in his paper titled "The Crane Report":

CRANE: *"Unfortunately, Rife had enlisted the help of electronic experts in the '30's who never wrote down the details of the instruments. Rife was unable to duplicate the marvels of his earlier Frequency instruments."* (The Crane Report, page 6).

www.rife.org

The documents we have read in this report show that Dr. Rife was having Verne Thompson build and repair the Beam Ray instruments originally built in the 1930's. For this reason we know that John Crane's statement, in his report, was not truthful. The facts have shown that this instrument design came from Philip Hoyland. It was Verne Thompson who updated the Beam Ray Clinical instrument for Dr. Rife, John Crane and John Marsh in 1953. If Verne Thompson was building these Beam Ray Clinical instrument then he surely would have had a schematic for them.

Others have felt that John Crane took advantage of Dr. Rife. This may be true in some things that took place later on in the late 1960's, 1970's and 1980's but the evidence does not support this in the 1950's and early 1960's. Here is a quote from the John Marsh Collection "Trip to Dayton Ohio Papers" and Gonin Papers. Dr. Rife, John Marsh and John Crane were talking at great length about John Marsh's trip to Ohio to see Dr. Robert P. Stafford M.D. In the Gonin papers they talked about the frequencies. I would recommend that everyone read these papers because they show that Dr. Rife, John Crane and John Marsh worked as a team. They also show that John Marsh and John Crane considered the frequencies to be Dr. Rife's and the AZ-58 Beam Ray replica instrument to be Dr. Rife's instrument. The following statements confirm this:

RIFE: *"Well I have lived my life for the benefit of humanity, and it is the end result of the accomplishment."*

MARSH: *"Yes, now here is what I did tell them. They wondered where I fit into the picture. I told them I had lay outs at the base, I designed part of it. You would say that I was possibly not an exactly an inventor, but I think we are all co-inventors of a sort by adding what we think would make the instrument better and if they try to validified* [verify] *any of the statements that I have said to them please don't let me down, and say no this isn't so, which might upset what might be the truth to them. I mean just by accident. Now what I mean by that is this. I don't think that I have in my own right lied to them. I did [didn't] try to impress them with the idea that I was the one that did it. I did impress that you* [Dr. Rife], *John Crane and myself had worked together on this thing, but that you* [Dr. Rife] *were inventor and John* [John Crane] *was the designer and inventor, co-inventor and myself for putting this thing together and making it. They asked if I helped putting this thing together and making it from time to time. I couldn't tell them that I didn't, because if I had built up a feeling in them that I knew nothing about what I was doing; psychologically that could have torn down, or have caused delay the foundation that now is laid. Now I think we have a solid footing there. I under no circumstances would want that torn down, and I will not under any circumstances accept the credit for this instrument as being invented, because it is Rife's instrument as printed on the plate in front and that is one of the reasons in building you up to them, which I don't think is unwarranted; not by a darn sight, and that is why they want you there. They want to hear you talk, and they also want to know your past experiences with the people of La Jolla and also I was very happy to have received the paper concerning the Dr. etc., because I'm sure Stafford will contact every blooming Dr. that you had given him to me and I turned over all the letters to him, because I didn't want anything to stand in the way if he could contact him, now whether he would do that, before he would talk to the group, and I do not know. I suppose he will, but he wants the truth as badly as you do. Now I don't know an easier way it can be done. I don't think there is going to be an easy way to get it on, but I think I've outlined this thing. I studied the moves I was going to make before I ever went there. I studied what I was going to do if I had the opportunity to do so, which I did."*

RIFE: *"Well I think that you did a very excellent job."* (1957 John Marsh trip to Dayton, Ohio #8, 9 and 16)

And in the John Marsh Collection, Gonin Papers we read:

CRANE: *"So the frequencies* [audio] *we have written down. I will give you those or Rife will give them to you. I think you ought to have them. Each one [organism] has a different frequency, you see. I don't* *remember any of them off-hand. I should memorize them all, but I haven't yet. I've just written them down and they are in the lab."*

GONIN: *"Those frequencies that you have written down, would only apply to your own machine* [AZ-58], *wouldn't it?"*

CRANE: *"That is because they have been calibrated for each machine. Each machine has its own calibration."*

GONIN: *"And that's constant?"*

CRANE: *"Yes."* (John Marsh Collection, Gonin and Siner Papers, Page 15)

After reading these documents the facts stand out that all three of these men knew the audio frequencies and they were not John Crane's invention. Both John Marsh and John Crane considered the instrument to be Dr. Rife's. Dr. Rife also had a plaque on the front of the instrument with his name on it. It is also clear from John Marsh's papers that Dr. Rife was not on the sidelines but he a working partner in Life Labs. Dr. Rife by this time had become what some people call a "working alcoholic." This type of drinker would have a little to drink during the day to take the edge off but would not be a total drunk. The documents show that Dr. Rife knew the audio frequencies that were used in the AZ-58. But it is clear that he didn't understand how those audio frequencies worked with the RF carrier frequency with Philip Hoyland's design. On the Rife CD's all of Dr. Rife's recorded conversations were very positive about the 1953 AZ-58 instrument. This does not sound like a man who was ignorant of what was going on, as some have claimed.

Anyone who reads the documents from this period of time can tell that there were a lot of high hopes for this instrument and the lower square wave audio frequencies it was using. Now the real question is how well did this instrument work when it was only working on square wave harmonics? There were a lot of good reports on how well the instrument worked but what really counts is how well it worked in the hands of the doctors who used it on their patients. Dr. Robert P. Stafford M.D., used the AZ-58 for 5 years on his patients and wrote a report and sent it to Dr. Rife, John Crane and John Marsh. The photo below is a picture of Dr. Robert P. Stafford M.D.

www.rifevideos.com
www.rifevideos.com
www.rifevideos.com

www.rife.org

Dr. Stafford's report is very favorable on many conditions that he used the AZ-58 on but when it came to cancer this instrument did not work as well as the Beam Ray high frequency harmonic sideband method. Dr. Stafford’s report showed he treated 16 cancer patients and had a varied response rate using the AZ-58 on cancer. We must point out that the instrument did temporarily help some of his cancer patients while others did not see any benefits. Two of the women that he treated received a great deal of benefit but died from other complications. This made it so that he could not say they were cured by the instrument. When the first woman was autopsied she only had a small amount of cancer left in her breast. When Dr. Stafford started treating her she had cancer in several other places including her neck. Had she not died from a fall she probably would have fully recovered.

The second woman had cervical cancer for which other doctors used a great deal of radiation therapy. Dr. Stafford treated her at the hospital and after about four weeks she went home. She died of radiation damage that was done to the tubes that go from the kidneys to the bladder. When she was autopsied they found no cancer cells in her abdomen. In this case the cancer was completely gone. Since neither of these women lived for at least five years they could not be considered cured. So Dr. Stafford could never say that he had cured anyone of a terminal malignancy. He said this about the AZ-58 square wave audio frequency instrument:

DR. STAFFORD: *“As yet, we have failed to “cure” any case of advanced, terminal malignancy. It appears in several instances that we may have impressed the disease favorably, temporarily. It is difficult to rule out the psychological, morale booster effect to the terminal patient when some definitive effort is made again in his behalf. However, several improvements have appeared to be more physical than emotional...All the patients in the series were treated with the same frequencies (e.g., 728 - 784 - 880 - 2008 - 2128). Perhaps these frequencies may be wrong, or only nearly correct.”* (John Marsh Collection, Dr. Stafford’s Report on using the AZ-58, page 4).

It is clear that Dr. Stafford was questioning the accuracy of the square wave audio frequencies and felt that something was wrong. Everyone who has been around Rife technology for a long time has seen the very same results as Dr. Stafford. From time to time we see that someone has an amazing experience of recovery from cancer but for the majority this does not happen. Today these frequencies **728, 784, 880, 2008 and 2128** Hertz which Dr. Stafford used are the same frequencies used by just about everyone for cancer.

Many have wondered why the AZ-58 worked well for a few patients and not for others. We believe we know the reason why. The RF carrier frequency is the reason it worked. The AZ-58 RF carrier frequency was set at the new 4.68 MHz to comply with the new 1950's FCC license. Dr. Stafford changed the RF carrier frequency using the variable capacitor to operate between 3.10 MHz to 3.30 MHz with the two women who were treated for cancer. He also used this same RF carrier frequency with some of his other patients. If per chance the RF carrier frequency was set at about 3.20 MHz it would have been only about 15,000 Hertz off of the second higher harmonic of Dr. Rife's original BX cancer virus frequency, which was 3,214,900 Hertz. This second harmonic of the BX (1,607,450 X 2 = 3,214,900 Hertz) was the primary frequency used by Philip Hoyland in the Beam Ray Clinical instrument for the BX cancer virus. If an audio frequency of 2128 Hertz (AZ-58 BX audio frequency) was used, and it was, it would have created many harmonic sideband frequencies and the 7th harmonic sideband frequency would have hit Dr. Rife's higher harmonic BX frequency of 3,214,900 Hertz. We need to keep in mind that Dr. Stafford also used 728, 784, 880 and 2008 Hertz. It would have been almost impossible for him not to hit the BX frequency of 3,214,900 Hertz with one of the sidebands from all these frequencies. If we also take into account the “one tenth of one meter” tolerance frequency of 858 Hertz it gives room for the frequency to be off a few hundred Hertz and still work. What we have just explained is the reason why we think the AZ-58 worked so well on these two women.

www.rife.org

It is also known that Dr. Stafford tried other RF carrier frequencies. Once he changed the RF carrier frequency the instrument would have been relying completely on the harmonics from the square wave waveform of the audio frequency. This would have greatly affected the outcome of his testing. The fact is the square wave harmonics method did not work as well as the Beam Ray Clinical sideband spacing method. **The real problem is almost everyone believes the square wave audio frequencies are Dr. Rife's true M.O.R. frequencies.** The evidence shows that John Crane and John Marsh believed the audio frequencies were the true M.O.R. frequencies and they had given this understanding to Dr. Stafford. Without the correct understanding of how the instrument worked Dr. Stafford would not have understood how important the fixed RF carrier frequency was. If he changed it off of the 3.20 MHz RF carrier frequency many of the 16 cancer patients could have been affected negatively. Dr. Stafford followed these 16 people over many years and in some cases things looked good at first, but the people eventually died from their cancer anyway. One thing that needs to be pointed out is we do not know if Dr. Stafford treated many of his patients over a long enough time period as Dr. Couche and the other doctors did their patients.

In the 1934 clinic 16 patients who had cancer and tuberculosis were treated and considered clinically cured. This is quite a contrast; 100% success in 1934 using Dr. Rife's high RF frequency method. And a very limited success rate on cancer for the low square wave audio frequency harmonics method. Only when Dr. Stafford accidently used an RF carrier frequency close to the BX cancer virus frequency did the AZ-58 seem to get good results. Dr. Stafford used the AZ-58 instrument for a little over five years and sent his report to John Marsh. I am sure John Crane received it also. The big question that needs to be asked is why did John Crane and John Marsh continue to tell people these low audio frequencies were the frequencies which Dr. Rife used in the 1934 clinic when the medical proof showed they didn't work as well as the high RF frequencies? This again clearly shows they did not understand the sideband method. We have to ask the question, why do people today continue to say these frequencies cure cancer even after they have seen the same results? Could it be because having had a few good results they ignored the evidence and fool themselves? By the time all the evidence was available, John Marsh and John Crane were in jail on three or four different legal counts, one of which was for treating a woman without a medical license.

Dr. Rife did not want to have anything to do with all the legal trouble they were in. He was able to avoid it because he never made any claims and he would never treat anyone. The legal problems shut down Life Labs. Had this not happened I wonder if Dr. Rife would have ignored this evidence? I do not believe he would have. He would have realized that the changes they made to the instrument which depended solely on square wave audio frequency harmonics compromised its effectiveness. I think Dr. Rife would have eventually realized that they didn't fully understand how the Beam Ray Rife Machine worked and he would have gone back to the higher audio frequencies and put the RF carrier frequency back on 3.30 MHz. We will never know what he would have done because I do not think Dr. Rife ever read Dr. Stafford's report. This is because John Marsh received the report after he and John Crane were released from jail.

The troubling thing is this, because so few really understand Dr. Rife's early instruments and how Philip Hoyland's Beam Ray Clinical Rife Machine worked almost all frequency generators have been built using this limited square wave audio frequency harmonic method. The people who purchase these low square wave audio frequency instruments have been led to believe it is the same type of instrument used in the 1934 clinic. All because we didn't know the truth. Are people today just fooling themselves also? Are we trying to get these same square wave audio frequency harmonic type instruments and the frequencies they use to do what Dr. Stafford could not get them to do? Cure cancer? We know there have been incredibly good results on many other conditions using audio frequencies which show this type of instrument and method is of great worth but the truth is sometimes hard to accept.

www.rife.org

As we have already read, Dr. Stafford came to suspect that the audio frequencies were not true M.O.R.s. Another letter written by Dr. Stafford to Dr. Edward Jeppson in Salt Lake City also confirms his concerns. He wrote this letter to Dr. Jeppson because he was having the same type of results that he was having. Here is his statement from his letter:

DR. STAFFORD: *"Please excuse my format in the following letter for I intend to ramble a bit and forget strict grammatical dictum. I am writing you at this time partially because John Marsh informs me in a recent letter that you may be somewhat disheartened or at least worried about your role in the experimentations with the Rife Machine. Believe me, Dr. Edward I know how you feel for I too have been through this same feeling with this matter. I have observed clinical results after treatments with this gadget which I can scarcely believe myself. Yet, despite these good results, I have been confused by some rather simple failures such as a recent experiment which I conducted at Good Samaritan Hospital where we used the machine to treat some cultures of Staph Aureus and Strept. Fecalis. In this work we failed to inhibit growth at all or influence the cultures with the Rife Rx. I sent the results to John Marsh and asked for clarification and to be very frank I am not satisfied with John's excuse of the failure as described by Dr. Rife. I am afraid I'm not a very good apostle for I'm getting some ideas myself on how this thing may work. I really wonder if this ultrasonic kills bacteria and virus at all or does it work like other forms of ultrasonic and merely stimulate the tissue in some unusual manner thereby improving the circulation and secondarily enhancing the body's defenses against infection…To summarize some of this rambling: I feel that the Rife Ultrasonic Therapy has a very definitely beneficial effect on the human (and canine) body...I furthermore feel that we, as doctors of medicine, using this machine must remain constantly alert to the condition of our patient and vary the Rx as indicated."* (Letter from Dr. Stafford to Dr. Edward Jeppson dated, April 1, 1958).

Clearly Dr. Stafford was questioning whether the audio frequencies were correct. Little did he know they were not the same frequencies used with the Rife Ray #3 or the Ray #4 instrument? The AZ-58 could have output the higher harmonic sideband frequencies that Philip Hoyland used in his Beam Ray Clinical instrument if they had only understood how it really worked. Whatever was told to Dr. Stafford by Dr. Rife through John Marsh it did not satisfy Dr. Stafford's concerns. Dr. Rife, John Crane and John Marsh probably felt that Dr. Stafford had just made some errors in his work. The one thing that Dr. Stafford did inadvertently find out was the AZ-58 using the square wave audio frequency harmonic method did not kill organisms in the laboratory at the hospital. In the 1950's Dr. Rife no longer had a laboratory for testing any microorganisms. There is no evidence they ever tested just the square wave audio frequencies with Dr. Rife's microscopes. So they did the only thing they could. They let the doctors use the AZ-58 and tell them how well it worked. We must keep in mind that the instrument Dr. Rife gave to John Marsh and John Crane was an original Beam Ray Clinical instrument which used the harmonic sideband method developed by Philip Hoyland. It used the correct higher audio frequencies and the correct 3.30 MHz RF carrier frequency. This instrument apparently worked because John Marsh said it cured his wife of cancer. Again the fact is the AZ-58 Beam Ray Clinical instrument using square wave audio frequency harmonics never worked as well as Philip Hoyland's sideband method. It also didn't work as well as the Rife Ray #3 or Rife Ray #4.

The square wave harmonic method used in the AZ-58 produced very good results on many conditions but not the results hoped for on cancer. But still even with the changes the AZ-58 worked very well on many different conditions. These square wave audio frequencies are what people have been using for the past 50 years believing they were Dr. Rife's true M.O.R.s. All this time not knowing they were not Dr. Rife's original frequencies which he used in his earlier instruments built in the 1920's and 1930's. It wasn't until the papers from the 1939 Beam Rays Trial, John Marsh Papers, Kennedy Company equipment spectrum analysis and Philip Hoyland Beam Ray Clinical Rife Machine spectrum analysis came to light did we have the ability to finally figure out where all these frequencies came from. This information finally reveals which frequencies were the correct M.O.R.s. Notwithstanding the various setbacks Dr. Stafford was still amazed at the results he achieved with the AZ-58.

www.rife.org

Since we were able to obtain access to John Marsh's papers we were also able to discover that in 1976 John Marsh had some of these low square wave audio frequencies tested on microorganisms by a laboratory using his 1970's model of the AZ-58 Beam Ray Clinical ray tube instrument. Keep in mind that this instrument was miscalibrated like all the 1950' AZ-58 Beam Ray Clinical instruments. It was also only working on square wave harmonics. He paid for these tests and they were carried out between May 1976 and March 1977. The University that did the tests specifically stated that, whatever the outcome of the tests, John Marsh could not use their name in any way to endorse his equipment. These tests were similar to the tests Dr. Stafford conducted at Good Samaritan Hospital. Dr. Stafford's tests showed that these low square wave audio frequencies, which did not produce the correct sideband M.O.R. frequencies, would not devitalize any organisms. John Marsh's tests which he had done at this laboratory also showed that the low audio frequencies working on square wave harmonics will not devitalize any organisms. Both Dr. Stafford's and John Marsh's tests convincingly show the low square wave audio frequencies are not true M.O.R.s. Unless these instruments use the higher audio frequencies with a 3.30 MHz RF carrier frequency they will not devitalize anything.

No one that reads this information should in any way believe that this means that Dr. Rife's method of coordinative resonance is without merit. What these laboratory tests showed is that it is absolutely necessary that the instrument work on the same principles and frequencies as the original instruments that produce the true M.O.R. frequencies use in the Rife Ray #3, Rife Ray #4 and the Beam Ray Clinical instrument. These 1950's AZ-58 Beam Ray Clinical instruments were not calibrated correctly and they used the wrong audio frequencies. This miscalibration made it so the instruments did not work on the original sideband principles which produced Dr. Rife's frequencies. Because they were not calibrated correctly it rendered the instruments incapable of devitalizing any microorganism using just low square wave harmonics audio frequencies.

Even after these tests proved that the low square wave audio frequencies would not devitalize any microorganisms John Marsh continued to hold on to the false notion that these low square wave audio frequencies would devitalize microorganisms. He never would consider that something was wrong with the instruments. Today we understand that a simple correction would have solved all the problems. To read the document about this laboratory test that conclusively proves that the audio frequencies are not M.O.R.s. go to www.rifevideos.com and read this chapter and you will find the following link. (John Marsh's tests performed by: Brigham Young University Microbiology Department)

We covered this information in a previous chapter but it should be reviewed again here. Some people have wondered if the low audio frequencies (120 Hertz to 2128 Hertz) used in the 1953 AZ-58 are actually the original audio frequencies used in the 1938-1939 Beam Ray machine rather than the high audio frequencies (1200 Hertz to 21275 Hertz) used in Aubrey Scoon's Beam Ray replica instrument. This is a good question and it can easily be answered with certainty. The answer is in the math which produces the correct sidebands for each organism. Only the high audio frequencies (1200, 2400, 6600, 6900, 7660, 7270, 7870, 8300, 8450, 8020, 16000, 17220, 18620, 20080, 21275) will produce the correct sideband frequencies that will produce the higher harmonic frequencies from Dr. Rife's original frequencies. Only six of the low audio frequencies (120, 660, 727, 1862, 2008, 2127-2128 or 2127.5 Hertz as given by John Crane) used in the 1953 AZ-58 when multiplied by a factor of 10 times give the exact same high frequency used in Aubrey Scoon's instrument. But the other seven (712, 784, 776, 800, 803, 880 and 1552 Hertz) when multiplied by a factor of 10 times will not give the correct high audio frequency. These facts reveal which frequency list came first. Since we know that Philip Hoyland designed this Beam Ray Clinical machine and hid the method of using sideband frequencies to produce higher harmonics of Dr. Rife's M.O.R. frequencies then only the list (Aubrey Scoon's higher audio frequency list) that will produce Dr. Rife's higher harmonic frequencies could be the original list.

www.rife.org

From the Beam Ray Trial we learned that no one but Philip Hoyland understood how the instrument worked. Not even Dr. Rife, Verne Thompson, John Crane or John Marsh understood that the RF carrier frequency had to be matched to the audio frequencies in order to produce the sideband frequencies that would hit the higher harmonic frequencies of Dr. Rife's original M.O.R. frequencies. John Crane and John Marsh said many times the RF carrier frequency did not matter. In fact, they eventually quit using the RF carrier frequency when they built their 1950's contact pad style instrument. This clearly proves that they did not understand that the original Rife Ray #5 or Beam Ray Clinical instrument worked on harmonic sidebands. Had they understood this simple fact they never would have changed the RF carrier frequency or built their contact pad style instrument without using an RF carrier frequency. They also would not have lowered or changed any of the audio frequencies if they understood the sideband method used by Philip Hoyland. Both audio frequency lists would be identical except that one list would be 10 times higher than the other list. Only someone who did not understand how the audio frequencies really worked would have lowered them and then changed them. Only the list which came first would have all the correct frequencies. This information proves that the high audio frequency list came first. The low audio frequency list used in the 1953 AZ-58 would have come later in the 1950's because it is the list, when multiplied by 10 times, will only produce some of the correct sideband frequencies. Only someone such as Philip Hoyland could have made the high audio frequency list since the high audio frequencies are the only frequencies that will produce Dr. Rife's higher harmonic M.O.R. frequencies. As we said, the answer to this question is in the math.

www.rife.org

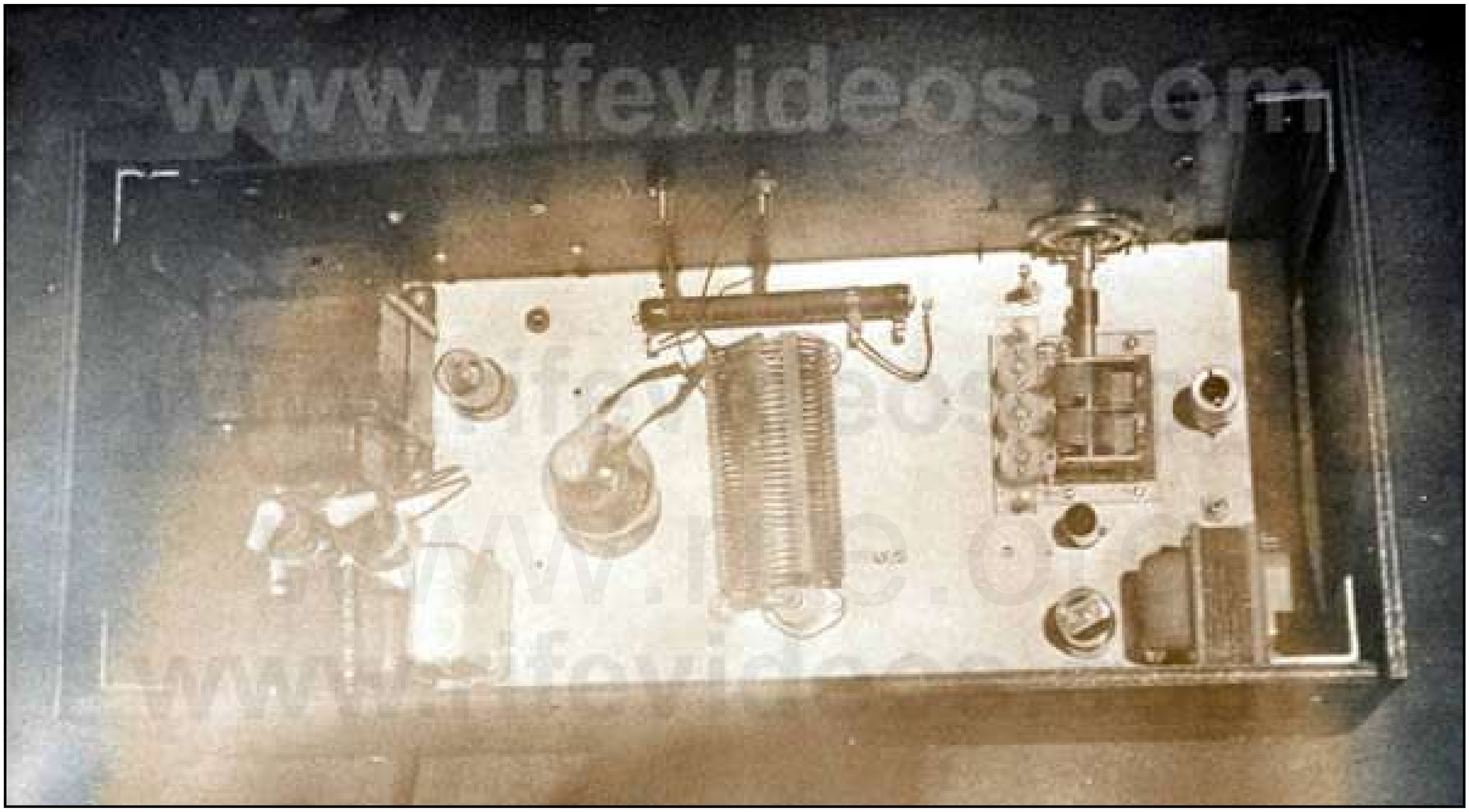

www.rife.org

Below in the chart is a list of the “Original 1950s AZ-58 Frequencies” used by Dr. Stafford. The column to the right of the "Original 1950's AZ-58 Frequencies" is based on a 4.68 MHz RF carrier and is labeled “4.68 Correct Sideband Frequencies”. They would be the correct sideband frequencies to make the AZ-58 instrument work like the original Beam Ray Clinical instrument, but, the problem is the large “Number of Sideband Harmonics”. You will notice that almost all of the low audio frequencies exceed more than 40 sideband spacing steps to hit the high frequency M.O.R.

Original 1950’s AZ-58 Frequencies Used By Dr. Robert P. Stafford M.D.

Microorganism	**Rife Ray #4 Frequencies**	**Higher Harmonic Frequencies**	***Carrier Difference Frequency**	**1/10 Of A Meter Freq.**	**Number of Sideband Harmonics**	**Original 1950s AZ-58 Frequencies**	**4.68 Correct Sideband Frequencies**
Actinomycosis or Streptothrix	192,000 Hz	4,608,000 or 24th	72,000 Hz	12 Hz	92	784 Hz	782 Hz
Anthrax	139,200 Hz	4,732,800 or 34th	52,800 Hz	6 Hz	85		621 Hz
B or E Coli Rod	417,000 Hz	4,587,000 or 11th	93,000 Hz	58 Hz	116	800 Hz	802 Hz
B or E Coli Virus	770,000 Hz	4,620,000 or 6th	60,000 Hz	198 Hz	38	1552 Hz	1538 Hz
BX Virus Carcinoma	1,604,000 Hz	4,812,000 or 3rd	132,000 Hz	858 Hz	32	2128 Hz	2129 Hz
BY Sarcoma	1,530,000 Hz	4,590,000 or 3rd	90,000 Hz	780 Hz	45	2008 Hz	2000 Hz
Gonorrhea	233,000 Hz	4,660,000 or 20th	20,000 Hz	18 Hz	28	712 Hz	714 Hz
Pneumonia or Spinal Meningitis	427,000 Hz	4,697,000 or 11th	17,000 Hz	61 Hz	17	776 Hz	1000 Hz
Staphylococcus Pyogenes Aureus	478,000 Hz	4,780,000 or 10th	100,000 Hz	76 Hz	137	727 Hz	730 Hz
Streptococcus Pyogenes	720,000 Hz	4,320,000 or 6th	360,000 Hz	173 Hz	409	880 Hz	880 Hz
Syphilis	789,000 Hz	4,734,000 or 6th	54,000 Hz	207 Hz	82	660Hz	658 Hz
Tetanus	234,000 Hz	4,680,000 or 20th	0 Hz	18 Hz		120 Hz	20 Hz
Tuberculosis Rod	369,000 Hz	4,797,000 or 13th	117,000 Hz	45 Hz	146	803 Hz	801 Hz
Tuberculosis Virus	769,000 Hz	4,614,000 or 6th	66,000 Hz	197 Hz	44	1552 Hz	1500 Hz
Typhoid Rod	760,000 Hz	4,560,000 or 6th	120,000 Hz	192 Hz	168	712 Hz	714 Hz
Typhoid Virus	1,445,000 Hz	4,335,000 or 3rd	345,000 Hz	694 Hz	185	1862 Hz	1865 Hz

The large “Number of Sideband Harmonics” means that the power level of the correct sideband that hit the higher harmonic M.O.R. would be so weak that it would never work. If the “Number of Sideband Harmonics” go over about 40 sideband steps the power level on a spectrum analyzer shows that they become too weak to work.

You will notice how closely these two sets of low audio frequencies correspond to each other. The problem is when Dr. Rife, John Crane and John Marsh lowered the audio frequencies they compromised the AZ-58. The BX audio frequency that Dr. Stafford used was 2128 Hertz and the RF carrier frequency was about 3.20 MHz. The sideband spacing to the “Higher Harmonic Frequency” M.O.R. took only a few sidebands which had enough power to work. This appears to be the reason why the two women and a few others that he treated had such amazing results. The problem is most of the organisms far exceed the 40 sideband harmonics when using these low audio frequencies. This will happen even if you use a 3.20 MHz or 4.68 RF carrier frequency. As mentioned before the higher the audio frequency used the more power there is in the sidebands. For this reason in the next chart, at the top of the next page, we have put a list of the optimum AZ-58 audio frequencies to be used with a 4.68 MHz RF carrier frequency and a 40,000 Hertz audio frequency oscillator. This would make the instrument work like the original Beam Ray Clinical instrument.

www.rife.org

1950’s Beam Ray Clinical Instrument Optimum Sideband Square Wave Audio Frequencies Based On A 4.68 MHz Carrier

Anthrax	26,400 Hz	Streptococcus	32,727 Hz
B or E Coli Rod	31,000 Hz	Streptothrix	36,000 Hz
B or E Coli Virus	30,000 Hz	Syphilis or Treponema	18,000 Hz
BX Virus Carcinoma	33,000 Hz	Tetanus	20 Hz
BY Sarcoma	30,000 Hz	Tuberculosis Rod	39,000 Hz
Gonorrhea	20,000 Hz	Tuberculosis Virus	33,000 Hz
Pneumonia or Spinal Meningitis	17,000 Hz	Typhoid Rod	40,000 Hz
Staphylococcus	25,000 Hz	Typhoid Virus	38,333 Hz

Below and on the next page are three photos of the original AZ-58. The first photo is of the inside of the case showing the variable capacitor (bottom left corner with the black knob and the black wire leading to it) which could change the RF carrier frequency. The second photo is a view of the underside of the chassis. From the 1930's to the 1950's the instrument had only a few changes made in the electronic parts. The third photo is a side view of the first AZ-58 built in 1953. If you take a close look at this photo you will see that it did not have the variable capacitor when it was first built. The variable capacitor was added in 1956 and was used for tuning the brightness or resonance of the ray tube. This was done because they didn't believe the carrier frequency made any difference. It was not added for the purpose of changing the RF carrier frequency, this was just a side benefit. Doctor Stafford used this variable capacitor to set the carrier frequency to about 3,200,000 Hertz for some of his tests. This is the reason we believe he had such good success with the two women patients who had cancer. The BX cancer frequency was 3,214,900 Hertz (1,607,450 X 2 = 3,214,900 Hertz) and even using the low audio frequencies this would have worked very well because he would have been hitting the BX frequency with powerful harmonic sideband frequencies.

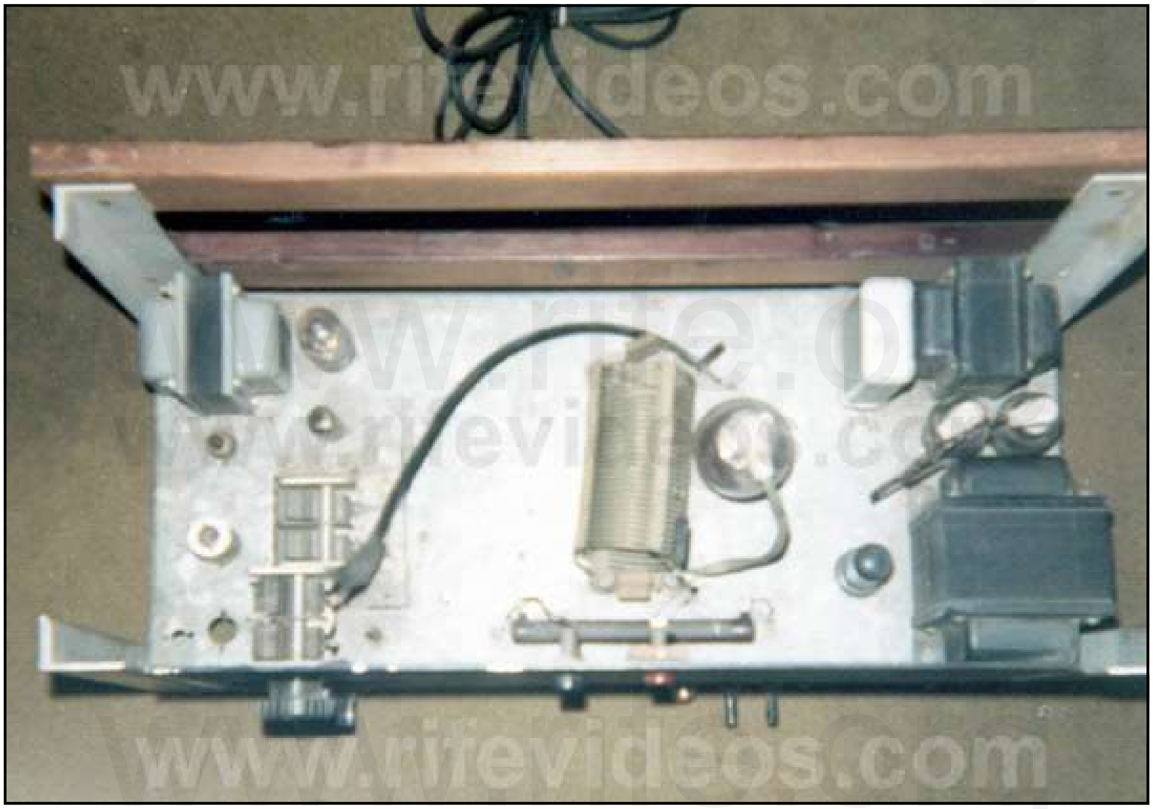

www.rife.org

www.rifevideos.com

The photos, shown on page 179, are of the AZ-58 we built back in 2000. It used to have the vacuum tube audio oscillator built into it but it never worked properly. So it was removed and replaced with Aubrey Scoon’s audio amplifier.

On page 180 is a schematic of the 1950's AZ-58 instrument. The 866 vacuum tubes have been replaced with solid state rectifiers. Also the old vacuum tube audio oscillator has been removed. It is easier and more accurate to use Aubrey Scoon’s booster amplifier and a modern function generator to produce the audio frequencies that were used in this instrument. The layout of the electronic parts of this instrument is very important because of the inherent interference problems that come with RF oscillators. Again anyone who would like to build this instrument should have a good understanding of old tube technology. Some parts of this circuit use up to 2000 volts DC with substantial current and can easily kill anyone who is not familiar with this kind of current or voltage. We take no responsibility for anyone who builds this instrument. We recommend that you have professional help.

Chapter Summary: To sum things up only a few changes were made to the original Beam Ray Clinical instrument design which produced the 1953 AZ-58 Beam Ray Clinical replica. Dr. Rife and Verne Thompson kept the original Hartley oscillator but changed the fixed RF carrier frequency from 3.80 MHz to 4.68 MHz as per the new FCC license. The RF carrier frequency section did not significantly changed with the use of the 812a vacuum tube instead of the 809 vacuum tube. They kept the variable audio oscillator which produced the low audio frequencies but lowered its range of frequencies using only three bands with a top audio frequency range of about 6000 Hertz. They then lowered Philip Hoyland's original sideband audio frequencies by a factor of about 10 times and used these lower audio frequencies in the AZ-58. Then they changed the audio frequency waveform from sine wave to square wave and depended on the harmonics produced by the square wave waveform instead of the sidebands. It is interesting that Philip Hoyland found that a modulated sine wave waveform when put thought his M.O.P.A. circuit was sufficient to devitalize organisms because it creates a wave form that looks almost like a square wave waveform. It appears that a damped wave is not really necessary. When you compare the original Beam Rays Clinical instrument and Aubrey Scoon’s 1950’s Beam Ray Clinical instrument to the 1950’s AZ-58 Beam Ray Clinical replica they are almost identical except for the audio frequency bands. When we built both of these instruments, Aubrey Scoon’s and the AZ-58, and compared them we found the AZ-58 replica appears to be the better design and easier to build.

Even with the changes Dr. Stafford had very good results which greatly impressed him. Today because of the changes that were made during the 1950's most people believe that the lowered audio frequencies that were used in the AZ-58 are Dr. Rife's original M.O.R. frequencies, but they not. Dr. Rife's original M.O.R. frequencies were the frequencies used in the Rife Ray #3 and Rife Ray #4. The Rife Ray #5 or Beam Ray Clinical instrument also worked on higher harmonics of Dr. Rife's frequencies. An instrument that could output both square wave harmonics and Philip Hoyland’s harmonic sidebands would take advantage of both of these methods.

In the next chapter we will discuss the difference between using square wave harmonics and sideband harmonics.

www.rife.org

Photos of the rebuilt AZ-58 Beam Ray Clinical instrument

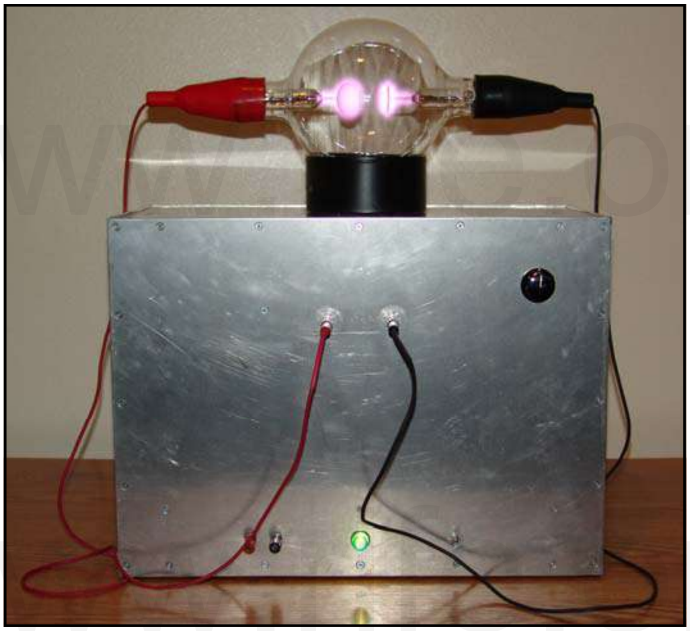

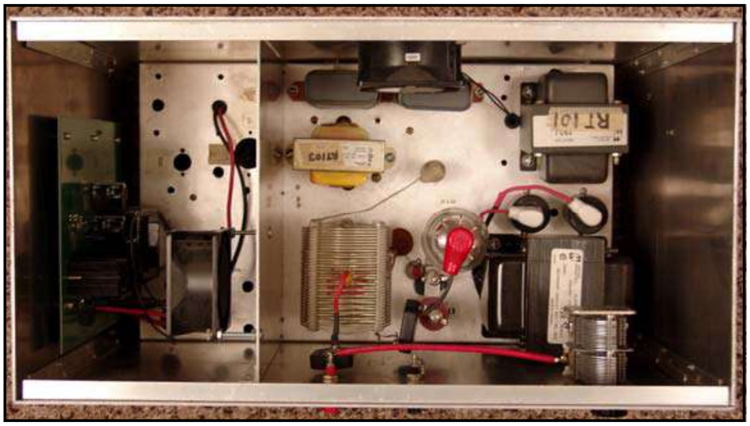
RT101
RT103

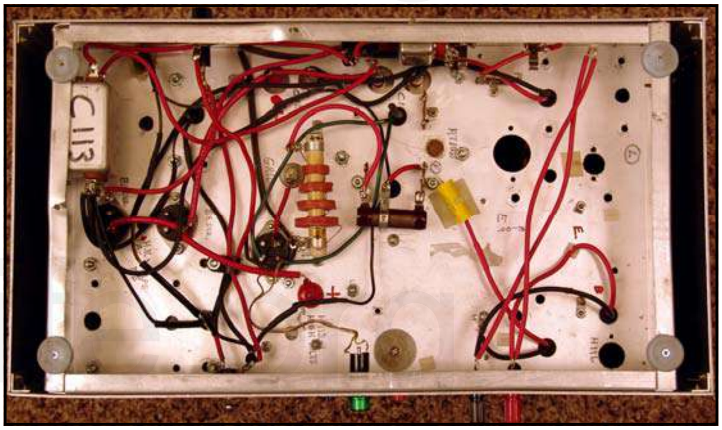

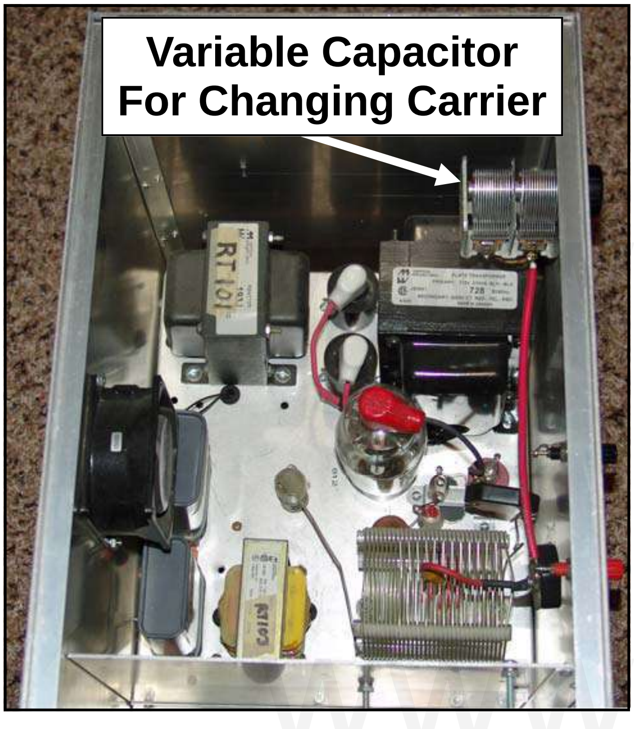
Variable Capacitor
For Changing Carrier
RT101
RT103

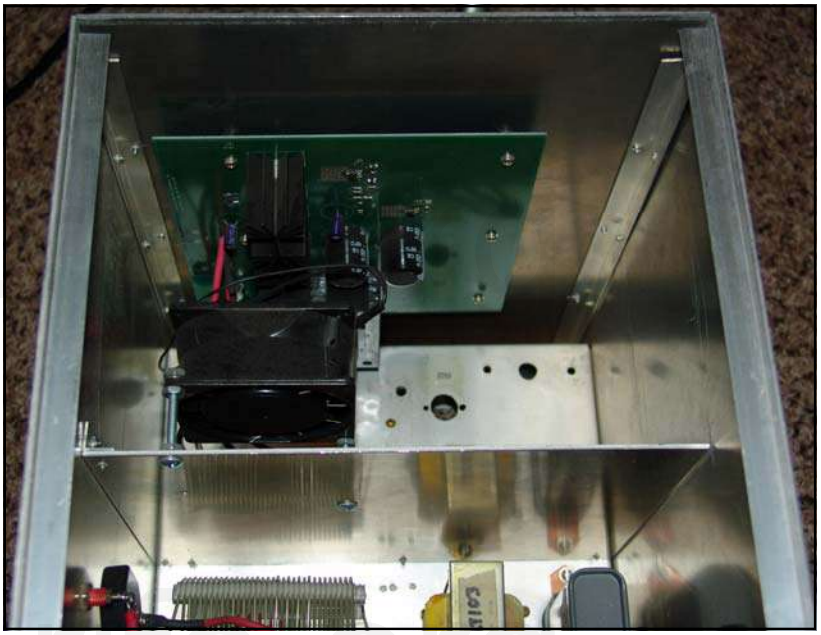

www.rife.org

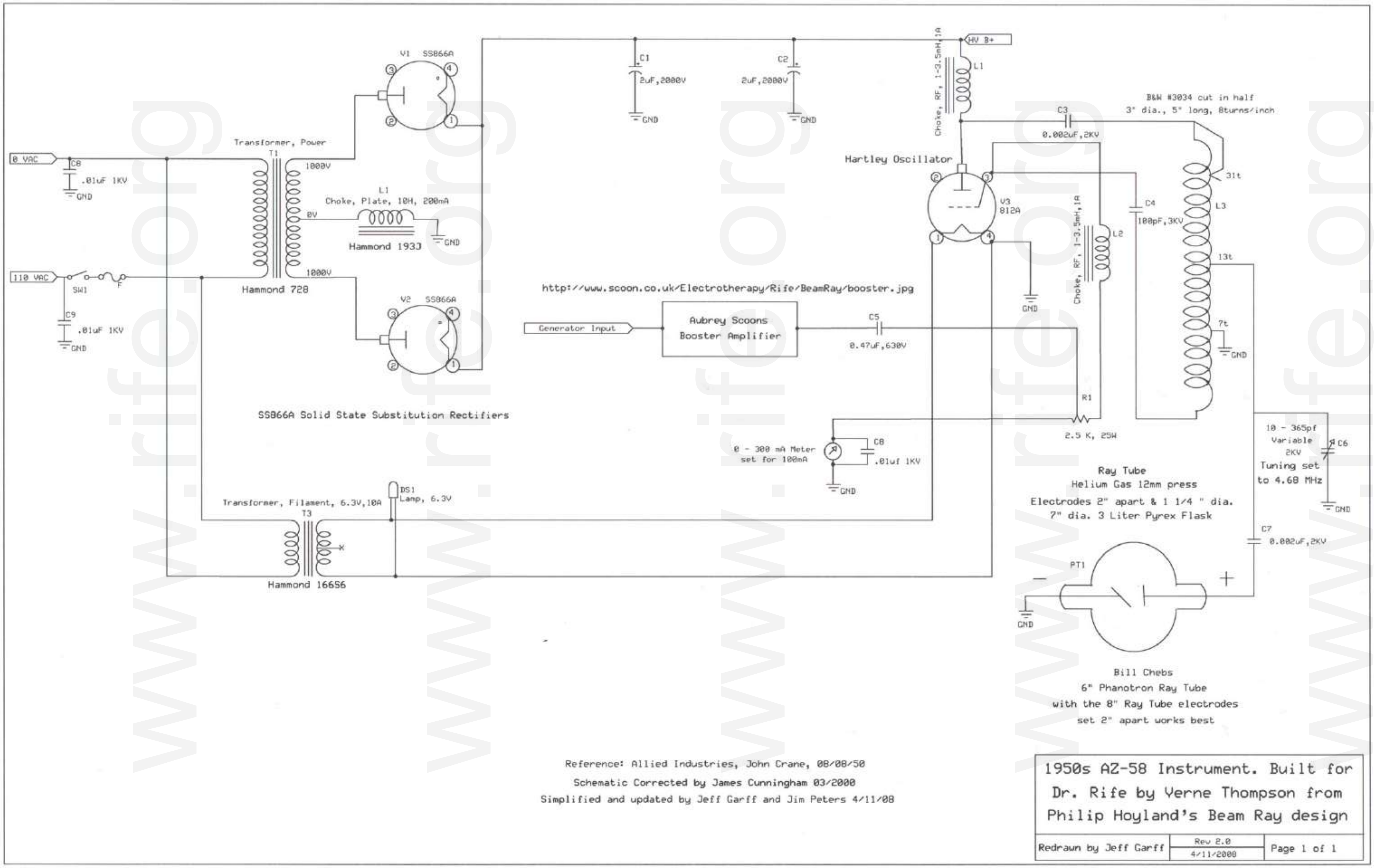
0 VAC
C8
.01uF 1KV
GND
110 VAC
SW1
F
C9
.01uF 1KV
GND
Transformer, Power
T1
Hammond 728
1000V
0V
1000V
L1
Choke, Plate, 10H, 200mA
Hammond 193J
GND
V1 SS866A
V2 SS866A
SS866A Solid State Substitution Rectifiers
Transformer, Filament, 6.3V,10A
T3
Hammond 166S6
X
DS1
Lamp, 6.3V
C1
2uF,2000V
GND
C2
2uF,2000V
GND
HV B+
Choke, RF, 1-3.5mH,1A
L1
Hartley Oscillator
V3
812A
GND
C3
0.002uF,2KV
B&W #3034 cut in half
3" dia., 5" long, 8turns/inch
31t
L3
13t
7t
GND
C4
100pF,3KV
Choke, RF, 1-3.5mH,1A
L2
http://www.scoon.co.uk/Electrotherapy/Rife/BeamRay/booster.jpg
Generator Input
Aubrey Scoons
Booster Amplifier
C5
0.47uF,630V
R1
2.5 K, 25W
0 - 300 mA Meter
set for 100mA
C8
.01uf 1KV
GND
10 - 365pf
Variable
2KV
C6
Tuning set
to 4.68 MHz
GND
C7
0.002uF,2KV
Ray Tube
Helium Gas 12mm press
Electrodes 2" apart & 1 1/4 " dia.
7" dia. 3 Liter Pyrex Flask
PT1
−
+
GND
Bill Chebs
6" Phanotron Ray Tube
with the 8" Ray Tube electrodes
set 2" apart works best

Reference: Allied Industries, John Crane, 08/08/50
Schematic Corrected by James Cunningham 03/2000
Simplified and updated by Jeff Garff and Jim Peters 4/11/08

1950s AZ-58 Instrument. Built for
Dr. Rife by Verne Thompson from
Philip Hoyland's Beam Ray design

Redrawn by Jeff Garff	Rev 2.0 4/11/2008	Page 1 of 1

www.rife.org

Chapter #13

Rife Machine Harmonic Audio Frequency Misunderstanding

www.rife.org

There is a belief that the audio frequencies that were used in the original Rife Ray #5 or Beam Ray Clinical Rife Machine and its 1940’s and 1953 AZ-58 replica Rife Machines are harmonic M.O.R. frequencies derived by dividing Dr. Rife’s original high RF frequencies down in octave steps until you reach the audio range of frequencies. For this to be true then all of the audio frequencies would have to be true harmonics of the higher RF frequencies. But this is not the case.

All the documents we have quoted and the understanding we now have, prove as a myth, the long believed concept that the low audio frequencies used in the 1950’s were created by John Crane and John Marsh by dividing Dr. Rife’s high frequency M.O.R.s down by harmonic steps until they reached the audio range of frequencies. If we take the correct frequency for the BX of 1,607,450 Hertz read by Philip Hoyland and divide it down by harmonics we do not get 2008 Hertz or 2128 Hertz. In fact, we do not get Philip Hoyland's higher audio frequencies of 20080 Hertz or 21275 Hertz either. This clearly proves John Crane and John Marsh did not create the audio frequencies by just dividing down Dr. Rife’s higher RF M.O.R. frequencies. The fact is we know that the origin of these audio frequencies came from the sideband harmonic method used by Philip Hoyland in the original Beam Ray Clinical instrument.

The documented information we have shows that Dr. Rife’s true M.O.R.s that would resonate organisms were the higher RF frequencies (139,200 Hertz to 1,607,450 Hertz) used in the Rife Ray #3 which consisted of the Kennedy Models 110, 281. These same frequencies were used in the Rife Ray #4 Rife Machine and the higher harmonic frequencies of these frequencies were used in the Rife Ray #5 or Beam Ray Clinical Rife Machine. All of Dr. Rife's Machines worked on RF frequencies not audio frequencies. From the spectrum analysis of Philip Hoyland's Beam Ray Rife Machine we know that it used audio frequencies for one purpose only, to create the method of sideband harmonic spacing to hit the higher harmonic frequencies which were harmonics Dr. Rife's true RF M.O.R.s. Henry Siner reported Philip Hoyland's design killed the organisms under microscope observation. This same instrument was tested by Dr. Rife in 1940 when Verne Thompson repaired it. Add to this the list of doctors who used the instruments and said they had many diseases cured. This shows that Philip Hoyland's harmonic method works beyond question.

www.rife.org

Dr. Robert P. Stafford was not able to kill any organisms in laboratory tests using just the square wave audio frequency harmonics. He did his tests under microscope observation. Today, like Dr. Stafford many have tried to kill the organisms which these audio frequencies correspond to, with no success. John Marsh had laboratory tests done in 1976-1977 which also confirmed Dr. Stafford's tests.

If these audio frequencies, as some have claimed, were really harmonic frequencies derived from the higher RF M.O.R. frequencies then one would think that they should be able to kill the microorganism they correspond to. If a true audio frequency harmonic of Dr. Rife's RF M.O.R. will not devitalize an organism under microscope observation, can the harmonic association be valid? No rational person would believe this. Yet today this is exactly what most people believe because they do not have all of the facts.

www.rife.org

When Dr. Rife started looking for frequencies to devitalize microorganisms he worked his way up from the lower frequency ranges to the higher frequency ranges until he found a frequency which would devitalize an organism he was working on. His Rife Ray #3 had the ability to start at about 12,000 Hertz which is in the upper audio frequency range. So Dr. Rife would have naturally started at 12,000 Hertz and moved up in the frequency range in his testing until he found the frequency that would devitalize the organism. Since the Rife Ray #4 documents show that the lowest frequency for any of the organisms was 139,200 Hertz we can conclude that he did not find any frequencies in the audio range or less than 139,000 Hertz that would devitalize any of these organisms listed on the Rife Ray #4 documents. Now this method of starting at the lowest frequency and moving up into the higher frequency ranges would be a logical method of finding the M.O.R. of an organism. What this also logically tells us is that no frequency lower than 139,000 Hertz could ever be a frequency that would devitalize any of these organisms. What this also reveals to us is that Dr. Rife had to get to a high enough frequency range before a resonant effect would devitalize an organism. What is the point that we are trying to make with this simple logic. Philip Hoyland found that higher harmonics of Dr. Rife's original M.O.R. frequencies would devitalize those organisms. But what Dr. Rife found was that there were no frequencies lower than the frequencies he found that would devitalize the organisms found in the Rife Ray #4 documents. We must keep in mind that the Rife Ray #4 did not have any variable audio oscillator. Its frequency range started at 87,000 Hertz. From this we can conclude that Dr. Rife found that the audio range up to about 87,000 Hertz was not needed because you cannot truly resonate any organism with frequencies below 87,000 Hertz using his method.

www.rife.org

With this logical understanding you would not divided down in octave or harmonic steps any of the Rife Ray #4 resonant frequencies and expect these lower frequencies to devitalize those organisms using the same method Dr. Rife used. But this is what people claim can be done. If Dr. Rife could not find any frequency lower than 139,000 Hertz to devitalize the Rife Ray #4 organisms then this concept that people are claiming is a questionable concept depending on the number of harmonic or octave steps used. It is a known fact that square waves will create higher harmonic frequencies. Some claim infinite harmonics but this could only be true if there was infinite power behind those harmonics. So power determines the number of harmonics. In most situations the harmonics from a square wave are only readable, with sufficient power, to about 9 harmonics up from the original frequency used. Understanding this you would not take any of Dr. Rife's frequencies and divide them down in either harmonic steps or octave steps more than about 9 times and expect an audio frequency with a square wave

www.rife.org

waveform to resonate those organisms. Yet this is what people are claiming can be done using audio frequencies and square wave harmonics. They actually promote the idea that you can go down hundreds of harmonic steps with Dr. Rife's frequencies and get them to resonate organisms with low audio frequencies. It is this false concept that John Crane and John Marsh promoted. This false concept has caused all of the confusion we have today. This is why so many people claim that Dr. Rife's original M.O.R. frequencies where the low audio frequencies that were used in the AZ-58.

We are not saying that low audio frequencies do not have positive effects, because they do. But they are working on a different concept than coordinative resonance. No one knows exactly how these low audio frequencies work, they just know they are beneficial. Dr. Stafford believed that they somehow stimulated the adrenals which in turn stimulated the immune system. He also believed when the immune system is stimulated the body will be able to overcome many ailments that it otherwise could not overcome. Again we will point out that it is not known how these low square wave audio frequencies work. But what is known is that they do not create a true resonate frequency effect like Dr. Rife's high RF frequencies or the higher harmonic RF frequencies Philip Hoyland used.

The evidence we have given in this report shows that if the RF M.O.R. frequency is lowered by too much it will lose its ability to devitalize an organism even though it is a lower harmonic frequency. Dr. Stafford's and John Marsh's laboratory tests showed this when they treated the organism with the low audio frequency that was supposed to devitalize it. The tests showed that the organism continued to grow even when they transferred it from one culture to another. Can there be any greater scientific proof than this? Even Dr. Rife would not have argued with this method of determining true M.O.R.s, because this is the method he used. This is the greatest proof, along with the fact that almost all the audio frequencies are not true harmonics of the original high RF frequencies. We know many other people who have made the same tests on microorganisms as Dr. Stafford and John Marsh and they told us they obtained the same results he did. John Marsh said on the Rife CD's that they came up with the frequencies using math. The documented information we now have show that the math they used was to lower Philip Hoyland's original audio frequencies by a factor of 10 times. If every audio frequency was a perfect harmonic match to its higher RF M.O.R. then we could say they were all derived from Dr. Rife's original M.O.R.s, but they are not. So this leaves us with only one conclusion. They lowered Philip Hoyland's higher audio frequencies and wrongly considered those lowered frequencies to be the true M.O.R. frequencies. Another thing we must understand is that even if the audio frequencies were derived from Dr. Rife's original RF M.O.R.s in lower harmonic steps or octave steps and they do not devitalize the organism they are not real M.O.R.s. either.

What the evidence in this report certainly proves is, without really knowing it, Dr. Rife, John Crane and John Marsh discovered that these lower square wave audio frequencies are beneficial. Even though they will not devitalize the organism they correspond to under microscope observation they still seem to help people? The fact is we really don't know why the audio frequencies are beneficial, but for some unknown reason they are. Though the method of modulating a square wave audio frequency onto a fixed RF carrier did not work as well as Philip Hoyland's sideband method, nevertheless it works very well on many conditions. Even though the audio frequencies are beneficial in many ways, this still does not prove the claim that the 1950's audio frequencies are harmonic M.O.R.s.

What needs to be kept in mind is how the original Beam Ray Clinical instruments worked. We will again point out here that neither the 3.30 or the 3.80 MHz RF carrier frequency nor the audio frequencies will do anything by themselves. But when the 3.30 or the 3.80 MHz RF carrier frequency and the audio frequencies are combined together they will produce many sideband frequencies. And one of these sideband frequencies will line up with the true Rife M.O.R. frequency and devitalize or render harmless the harmful microorganism. If you just use the audio frequencies by themselves you will get nothing. If you use the 3.30 or the 3.80 MHz RF carrier without the audio frequencies you will get nothing. The audio frequencies used in this style of instrument must be match to the RF carrier frequency of

www.rife.org

3.30 or 3.80 MHz or they are useless. This is the reason the 1953 AZ-58 Beam Ray Clinical instrument did not work properly.

Below is a chart showing the “High Frequency M.O.R.s” which were taken from the Rife Ray #3 and used with the Rife Ray #4. These frequencies were recorded in 1935 when Philip Hoyland went to Dr. Rife's laboratory to read them with his master oscillator. They were fine tuned with more precision in 1936 when Philip Hoyland built the Beam Ray Clinical instrument. This is the reason for the slight discrepancy in Dr. Rife's high frequencies. In 1935 the frequencies were rounded off to the nearest thousandth. In 1936 a more accurate reading was done of these frequencies so that the higher harmonics could be used in the Beam Ray Clinical instrument. These frequencies are the true M.O.R.s. The “Harmonic Higher Audio Frequency” column is the true harmonic of the “High Frequency M.O.R.s” column. How well these higher frequencies under 60,000 Hertz would work is anyone's guess using the square wave harmonic waveform that was used in the AZ-58 replica instrument. The “Harmonic Higher Audio Steps” column shows how many harmonic steps it takes to hit the fundamental “High Frequency M.O.R.s” using the “Harmonic Higher Audio Frequency” using a square wave waveform. As can be seen the higher the frequency the less harmonic steps it takes to hit the fundamental M.O.R. frequency. We did this chart just to show how the concept of square wave harmonics would work. Just keep in mind that any square wave audio frequency only goes up for about 9 harmonics with sufficient power. Many of the “Harmonic Higher Audio Steps” exceed 9 harmonics. With this understanding a higher frequency should be chosen that would be within 9 harmonics of Dr. Rife’s original high RF M.O.R. frequency.

AZ-58 M.O.R. Audio Frequencies Square Wave Harmonics

Microorganisms	**1950’s Low Audio Frequency For AZ-58**	**True Harmonic Low Audio Frequency For AZ-58**	**Harmonic Low Audio Steps for AZ-58**	**Harmonic Higher Audio Frequency**	**Harmonic Higher Audio Steps**	**High Frequency M.O.R.s**
Actinomycosis (Streptothrix)	784 Hz	750 Hz	256	48,000 Hz	4	192,000 Hz
Anthrax	None	1087.5 Hz	128	34,800 Hz	4	139,200 Hz
B. Coli (Rod form)	800 Hz	814.4531 Hz	512	52,125 Hz	8	417,000 Hz
B. Coli (Filterable virus)	1552 Hz	1503.9063 Hz	512	48,125 Hz	16	770,000 Hz
Bacillus X or BX (Cancer Carcinoma)	2128 Hz	3132.8125 Hz	512	50,125 Hz	32	1,604,000 Hz
Bacillus Y or BY (Cancer Sarcoma)	2008 Hz	2988.28125 Hz	512	47,812.5 Hz	32	1,530,000 Hz
Gonorrhea	712 Hz	910.15625 Hz	256	29,125 Hz	8	233,000 Hz
Pneumonia or Spinal Meningitis	776 Hz	1667.96875 Hz	256	53,375 Hz	8	427,000 Hz
Staphylococcus Pyogenes Aureus	727 Hz	933.59375 Hz	512	59,750 Hz	8	478,000 Hz
Streptococcus Pyogenes	880 Hz	703.125 Hz	1024	45,000 Hz	16	720,000 Hz
Syphilis	660 Hz	770.5078 Hz	1024	49,312.5 Hz	16	789,000 Hz
Tetanus	120 Hz	914.0625 Hz	256	29,250 Hz	8	234,000 Hz
Tuberculosis (Rod)	803 Hz	720.7031 Hz	512	46,125 Hz	8	369,000 Hz
Tuberculosis (Virus)	1552 Hz	1501.953125 Hz	512	48,062.5 Hz	16	769,000 Hz
Typhoid Fever (Rod)	712 Hz	742.1875 Hz	1024	47,500 Hz	16	760,000 Hz
Typhoid Fever (Virus)	1862 Hz	1411.1326 Hz	1024	45,156.25 Hz	32	1,445,000 Hz

Now if we take a look at the "1950's Low Audio Frequency For AZ-58” column we see the frequencies that were used by the AZ-58 back in the 1950's. These frequencies are still used today. In the “True Harmonic Low Audio Frequency For AZ-58” column we find the true low audio frequency harmonics of the “High Frequency M.O.R.s.” As you compare these columns you can see that the 1950's frequency for Actinomycosis or Streptothrix was 784 Hertz but the true harmonic frequency is 750

www.rife.org

Hertz. It is easy to see that the AZ-58 was not working on square wave harmonics of the true M.O.R.s. If you look at the “Harmonic Low Audio Steps For AZ-58” column you can see how many harmonic steps it takes to reach the true “High Frequency M.O.R.” These range from 128 harmonic steps to 1024 harmonic steps. Anyone looking at these numbers would wonder if these frequencies could ever work. Logically the higher the frequency the better the chance they will work. Also logically the less you go up in square wave harmonics the greater the chance they will work. When using the square wave method the highest frequency possible should be used.

Chapter Summary: None of the low audio frequencies used in the AZ-58 were Dr. Rife's original M.O.R. frequencies. Not even the higher audio frequencies used in the Beam Ray Clinical instrument were his original M.O.R. frequencies either. Both the lower and the higher audio frequencies are not true harmonics of Dr. Rife's original M.O.R. frequencies, as some have thought. The frequencies that Dr. Rife discovered were the lowest frequencies which he found that would resonate an organism and devitalize it. Dr. Rife knew that his frequencies were lower harmonics of a true higher frequency. Philip Hoyland used this understanding in the Rife Ray #5 or Beam Ray Clinical instrument. One thing we do know is dividing Dr. Rife's original frequencies down in octave or harmonic steps until you get into the low audio frequency range has not been proven through laboratory testing to resonate those organisms or any other organisms. Dr. Stafford and John Marsh had laboratory tests done and found that the 1950's AZ-58 square wave low audio frequencies would not devitalize any organisms they were tested on. The audio frequency range does have many beneficial frequencies but it is not understood how they really work. Dr. Stafford believed they stimulate the adrenal glands thus stimulating the immune system.

The real problem with not understanding which frequencies are Dr. Rife's original M.O.R. frequencies means that people will purchase frequency generating equipment that does not output Dr. Rife's original frequencies. It really is "buyer bewares". If a frequency generating piece of equipment cannot output both the low audio frequency range and the high RF range then you may want to reconsider purchasing it. Look for a frequency generator that can output all of Dr. Rife's frequencies. If a company will not tell you what the frequency range of their so called "Rife Machine" is then keep looking until you find a good frequency generator with the correct frequency range. Any frequency generator worth purchasing should have a range from at least 1 Hertz to 5,000,000 Hertz so you can use the full frequency range that Philip Hoyland used with the higher harmonics of Dr. Rife's original M.O.R. frequencies. If a frequency generator can go to 20,000,000 million Hertz, even better, because you can also work with many of the higher harmonics of Dr. Rife’s frequencies.

In the next chapter we will look at the pad type instrument developed by John Crane and John Marsh in the late 1950's.

www.rife.org

Chapter #14

Life Labs 1950's Pad Instrument without Ray Tube

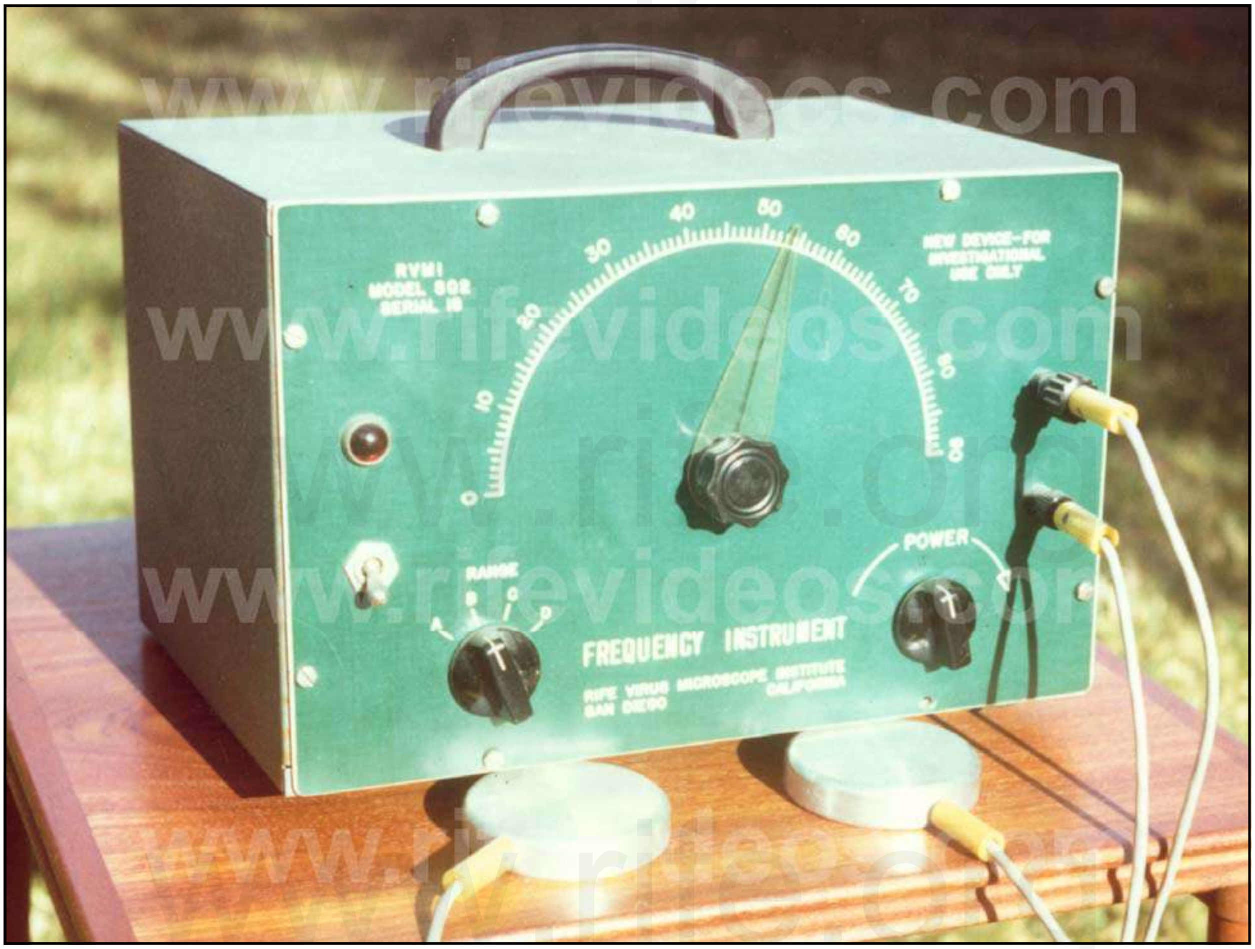
RVMI
MODEL 802
SERIAL 18
0
10
20
30
40
50
60
70
80
90
NEW DEVICE-FOR
INVESTIGATIONAL
USE ONLY
RANGE
A
B
C
D
POWER
FREQUENCY INSTRUMENT
RIFE VIRUS MICROSCOPE INSTITUTE
SAN DIEGO
CALIFORNIA

1. Used round disks that came in contact with the body. Later changed in the 1960's to hand cylinders or foot pads.
2. Had no RF carrier frequency.
3. Used the square wave audio frequencies used in the AZ-58.

It was about 1957 when John Crane and John Marsh began building instruments without a ray tube. Earlier in this article Bertrand Comparet was quoted as saying:

COMPARET: *"Now, Crane said "Well now look, Rife himself admits that no matter how much tube and ray, and so on, you have, you can't get any results unless you've got the right frequency. Therefore the real clue to the thing is the frequency and not the means by which you deliver it."* (1970's Bertrand Comparet Interview #33).

John Crane and John Marsh replaced the ray tube with two aluminum disks (shown in the above photo) which they developed that came in contact with the body. As we pointed out earlier in this report it is interesting to note that Dr. Rife said Abrams' Oscilloclast would devitalize the BX cancer virus and it was a contact type device. John Crane and John Marsh probably used this contact method because of the success of Abrams' instrument. From the documented information we have it was also the high cost of building ray tube instruments that caused them to look at doing things in a different way. In addition to being expensive to build, the ray tube could break very easily. They had many problems with them. It does not appear that Dr. Rife, at least in his early years, ever had a reason to look at doing

www.rife.org

things differently. John Crane and John Marsh did. They didn't have the kind of money to spend that Dr. Rife did. They say that "Necessity is the mother of invention!"

In the photo, shown below, we see that they eventually added handles onto the disks so that they were easier to use.

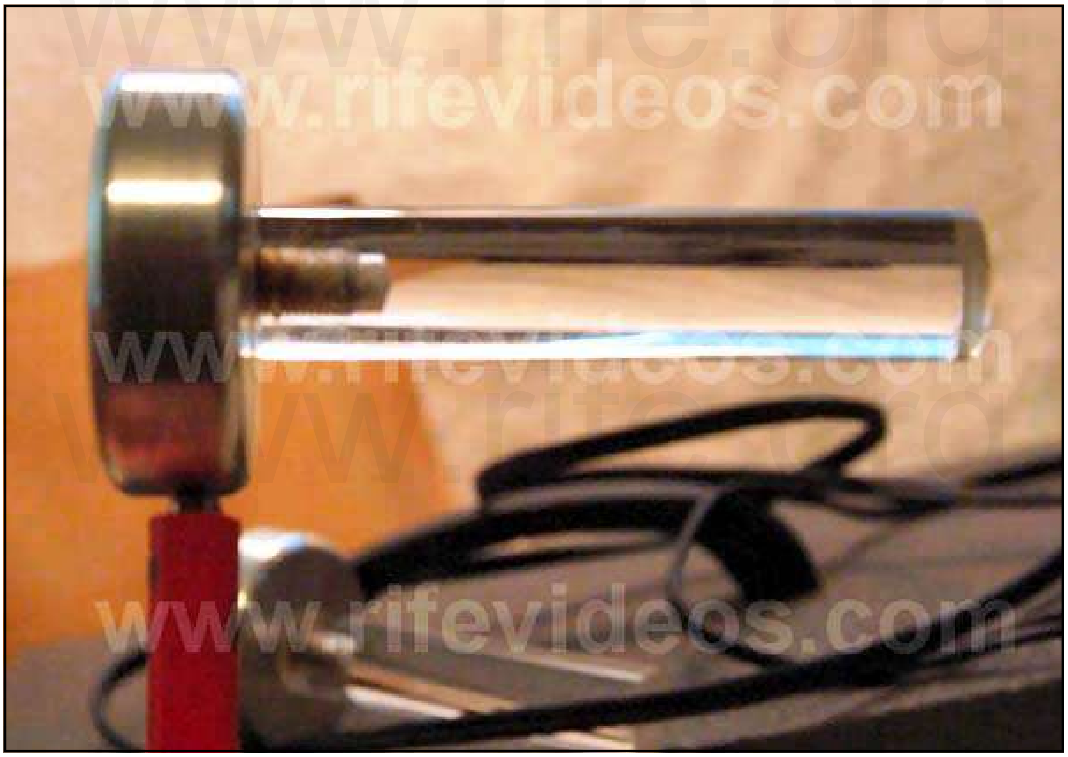

John Crane and John Marsh used a Heathkit function generator to produce the frequencies. These Heathkit function generators had no built-in carrier frequency on which to modulate the audio frequencies. Therefore, the RF carrier frequency was no longer used. The fact that they didn't feel the RF carrier frequency was necessary shows that they totally believed that the audio frequencies were the M.O.R. frequencies. This also shows beyond doubt that they never knew the importance of the RF carrier frequency or had any understanding of the harmonic sideband method used in the original Rife Ray #5 or Beam Ray Clinical Rife Machine. They could have made a pad instrument work like the Beam Ray Clinical Rife Machine if they would have used a harmonic sine wave RF carrier frequency at 3,300,000 Hertz. After the many years that these pad instrument have been used it appears that the removal of the ray tube was not as important as the removal of the RF carrier frequency.

Dr. Rife would have never approved of using an instrument without an RF carrier frequency. He knew that Philip Hoyland's Beam Ray Clinical instrument which used the sideband method somehow needed an RF carrier frequency to make it work properly. Though Dr. Rife didn't fully understand Philip Hoyland's instrument he clearly understood the importance of the RF carrier frequency. The positive thing about using a Heathkit function generator in this way is they were inexpensive (about $200) and a lot more people could afford one. Many people can thank John Crane and John Marsh for this innovative method. John Crane and John Marsh proved that the square wave audio frequencies worked the same whether applied through a ray tube or pads if sufficient power is used. Many people think that John Crane and John Marsh built the pad instrument without Dr. Rife being fully informed about it. But this was not the case. John Crane and John Marsh had talked for some time about building a smaller ray tube instrument but instead of building it they built the pad instrument. In John Marsh's Trip to Ohio Papers we read this:

RIFE: *"That is the only way that it can be handled properly."*

www.rife.org

MARSH: *“Maybe we can sell small instruments for the purpose of small diseases like colds, flu and stuff like that, which are minor, which the Dr.s prefer not treating those kind anyway, because they are chronic, and there isn’t anything they can do with them. People keep coming in and coming in and they take up his time where he could spend it taking care of a bad case, or something or other. Dr. Stafford said that he would prefer that a small instrument would be made...What do you think John? I've been doing a lot of talking not even giving you a chance to get a word in edgewise.”*

CRANE: *“There is no doubt there is going to be an awful lot of development on this design…”* (1957 John Marsh trip to Dayton, Ohio #36-38).

From these statements we learn that Dr. Rife knew that they wanted to build small instruments. Also we learn that it was John Marsh and John Crane’s idea to build the pad instruments, not Dr. Rife’s. We know that Dr. Rife was upset with John Crane and John Marsh over the building of these pad instruments because he expressed it to Bertrand Comparet during his 1961 deposition. Comparet said:

COMPARET: *“And I asked Rife, because I thought Rife would certainly say that the way Crane was working on it then was still using the Rife principle, but he indignantly denied it.”*

HUBBARD: "*All right, I see. But, getting back, you say that Rife was very indignant, that the machine that Crane was building was really his* [Cranes] *idea. I suppose he did not compromise on that, did he?*

COMPARET: *"Oh no, he just blew up."* (1970’s Bertrand Comparet interview #32 & 40).

At this time John Crane and John Marsh were working on both the ray tube instrument and the pad instrument. We know that Dr. Rife considered the ray tube instrument to be his instrument. The ray tube instrument used an RF carrier frequency on which the square wave audio frequencies were modulated. So it is clear that the pad instrument without an RF carrier frequency is what Dr. Rife was upset about. It is also clear that Dr. Rife fully knew what they were doing but he did not approve of what they were doing. Because they didn’t use an RF carrier frequency the pad instrument would not be working on Dr. Rife's principle of coordinative resonance through high RF sideband frequencies. We know that this was the method Philip Hoyland used in his design. If there is no RF carrier then there would be no sideband frequencies and this would mean that none of Dr. Rife's higher harmonic RF frequencies would have been produced. This also means the pad instrument only worked on low square wave audio frequency harmonics not high RF frequencies. However, John Marsh and John Crane’s innovation with a pad instrument proved that the ray tube could be removed and frequencies could be applied electrically through the contact method which they used.

This new method made it possible for more people to have access to a less effective form of Dr. Rife’s technology. Though it was a less effective method many people over the years have been helped by this method of only using low square wave audio frequencies. Back when John Crane and John Marsh were building these pad instruments they could have built a pad instrument capable of producing all of Dr. Rife's frequencies. The reason they did not do this is due to the fact that they believed that the low square wave audio frequencies would work as well as Dr. Rife's original high RF frequencies. Time has proven that they were not correct in this belief. Today all of Dr. Rife's original high RF frequencies and the higher harmonics of those frequencies which Philip Hoyland used can be produced by any function generator with the proper frequency range of at least 1 Hertz to about 4,000,000 Hertz.

It appears that one of the reasons why John Crane and John Marsh didn’t use an RF carrier frequency is the Heathkit function generator that they used didn’t have the capability of using an RF carrier frequency. Audio frequencies will not broadcast; therefore, they are modulated upon the RF carrier frequency so that they will penetrate the body. In laymen’s terms, modulation is piggy-backing one or more low frequencies onto another higher frequency. The frequencies travel together but still keep the components of both waveforms. Almost all of the so called "Rife Machines" built today do not use an RF carrier frequency even though Dr. Rife’s Beam Ray Clinical instrument and the 1953 AZ-58 Machine did. If Dr. Rife could have removed the RF carrier frequency from his instrument and gotten the same results, I believe he would have removed it because it would have made building an instrument a lot easier. He became upset with John Crane and John Marsh for doing this. As it is, Dr. Rife never removed the RF carrier frequency from any of his instruments. It was John Crane and John Marsh who remove it. If a person wants to try and obtain the same type of results which Dr. Rife did, then an RF carrier frequency should be considered and used with any audio frequency instrument.

We realize that there are ray tube instruments today that do not use an RF carrier frequency. These use a high electromagnetic field (EM) which will transfer the energy into the body. We do not doubt that these instruments work very well with low square wave audio frequencies. We have talked with people that are using these instruments and they say that they are getting very good results. But these instruments are still using just the low square wave audio frequencies without an RF carrier frequency. What we are talking about in this report is the way that Dr. Rife used RF frequencies and an RF carrier frequency modulated with an audio frequency to produce the proper sideband method. Some of these EM ray tube instrument builders like to compare EM devices to RF devices. This is like comparing apples and oranges. There is no comparison because if you do not have an RF carrier frequency then you cannot create any sideband frequencies. The EM devices work on a high electromagnetic field, the stronger the better. RF devices work on radio frequencies and the power output is measured in watts. Dr. Rife’s instruments were all RF instruments and could resonate an organism when the proper RF frequency was used. EM devices use low audio frequencies because they cannot output RF frequencies. RF versus EM or apples and oranges cannot be compared because they are two totally different principles. EM devices are also limited in their frequency range. Usually they can only output frequencies to about 20,000 Hertz where Dr. Rife's RF devices were putting out frequencies in the millions of Hertz or cycles per second.

There is some misinformation being put out about pad instruments. Mostly it is done by people who believe that only a ray tube will work in delivering frequencies to the body, whether it is through the RF or the EM method of delivering frequencies. These people promote the false belief that audio frequencies when used in a pad instrument without an RF carrier or without an EM field will only travel along the surface of the skin of the body. They say that the frequencies cannot penetrate the body because of the "Skin Effect." We need to point out that the "Skin Effect" has nothing to do with human tissue or human skin. The "Skin Effect" has to do with the skin or surface of a metal conductor such as a copper wire. If the human body was made of metal then the "Skin Effect" would apply. These people are either totally ignorant or willfully trying to mislead people so that they will only purchase an instrument that uses a ray tube. What we should really believe is what the scientific tests have proven in regards to the "Skin Effect" as it pertains to human or animal tissue.

In scientific studies called "Bioelectric Impedance Analysis" it has been shown that sine wave audio frequencies, without an RF carrier frequency, will enter the body but will only travel in the connective tissues around the cells. These tests have prove beyond doubt that the frequency does go right through the skin contrary to what people have claimed. Also in these scientific studies it has been shown that the closer you get to 1 Megahertz the greater the penetration of the current through the cell. At 1 Megahertz or one million Hertz the current of the frequency will go through the cell and fully penetrate the body. These type of tests show why it is very important that an RF carrier frequency be used. A virus can enter a cell and live there. An RF frequency can enter the cell where it can do the most good. These kinds of scientific studies and their importance were not understood by John Crane and John Marsh in the 1950’s and 1960’s. These "Bioelectric Impedance scientific tests also prove that what many people have been saying about the "Skin Effect" is absolutely false. In this Rife machine report we do not ask people to believe what has been said without proof. For this reason we have in-

www.rife.org

cluded one of these scientific reports for you to read. There are other reports on the web that you can read but we found that this one is the easiest to understand. But even with that said, unless you have some understand of how electricity works then it may be difficult for some to understand. Because of this we added, in blue print, a layman's understanding at the end of each section. To read this file go to the following website address.

www.rifevideos.com/pdf/skin_effect_and_bio_electrical_impedance_analysis.pdf

Below is a schematic of John Crane & John Marsh's pad instrument. It was nothing more than an off-the-shelf audio frequency generator with the faceplate changed. There was nothing special about this frequency generator because any common function generator can do the same thing that this one could do. There have been people who have copied this instrument and who charge enormous sums of money for a replica. As much as four or five thousand dollars. They claim that this is a real genuine Rife Machine. Do not be fooled. The same audio frequencies that were used in the 1953 AZ-58 were also used with this instrument. Those frequencies were 120, 660, 712, 727, 776, 784, 800, 803, 880, 1552, 1862, 2008, 2128.

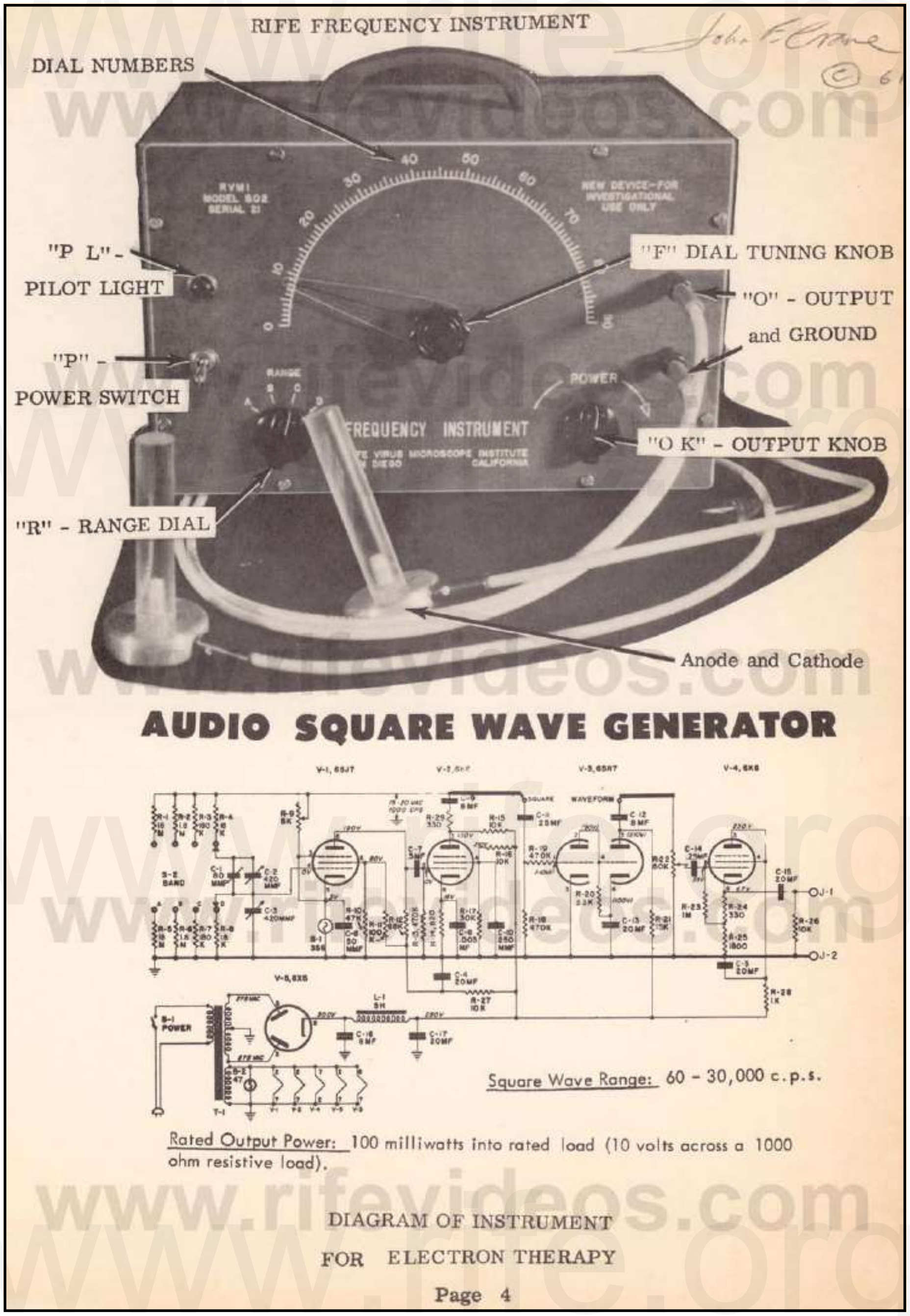

RIFE FREQUENCY INSTRUMENT

John F. Crane

© 6

DIAL NUMBERS
"P L" - PILOT LIGHT
"P" - POWER SWITCH
"R" - RANGE DIAL
"F" DIAL TUNING KNOB
"O" - OUTPUT and GROUND
"O K" - OUTPUT KNOB
Anode and Cathode
RVMI MODEL 602 SERIAL 21
NEW DEVICE-FOR INVESTIGATIONAL USE ONLY
RANGE
POWER
FREQUENCY INSTRUMENT
RIFE VIRUS MICROSCOPE INSTITUTE SAN DIEGO CALIFORNIA

AUDIO SQUARE WAVE GENERATOR

V-1, 6SJ7
V-2, 6K6
V-3, 6SN7
V-4, 6K6
V-5, 6X5
SQUARE WAVEFORM
S-1 POWER
T-1
L-1 5H
J-1
J-2

Square Wave Range: 60 - 30,000 c.p.s.

Rated Output Power: 100 milliwatts into rated load (10 volts across a 1000 ohm resistive load).

DIAGRAM OF INSTRUMENT

FOR ELECTRON THERAPY

Page 4

www.rife.org
www.rifevideos.com

Below are several photographs of the many pad type frequency generators that John Cane and John Marsh built. The first one was built by John Crane.

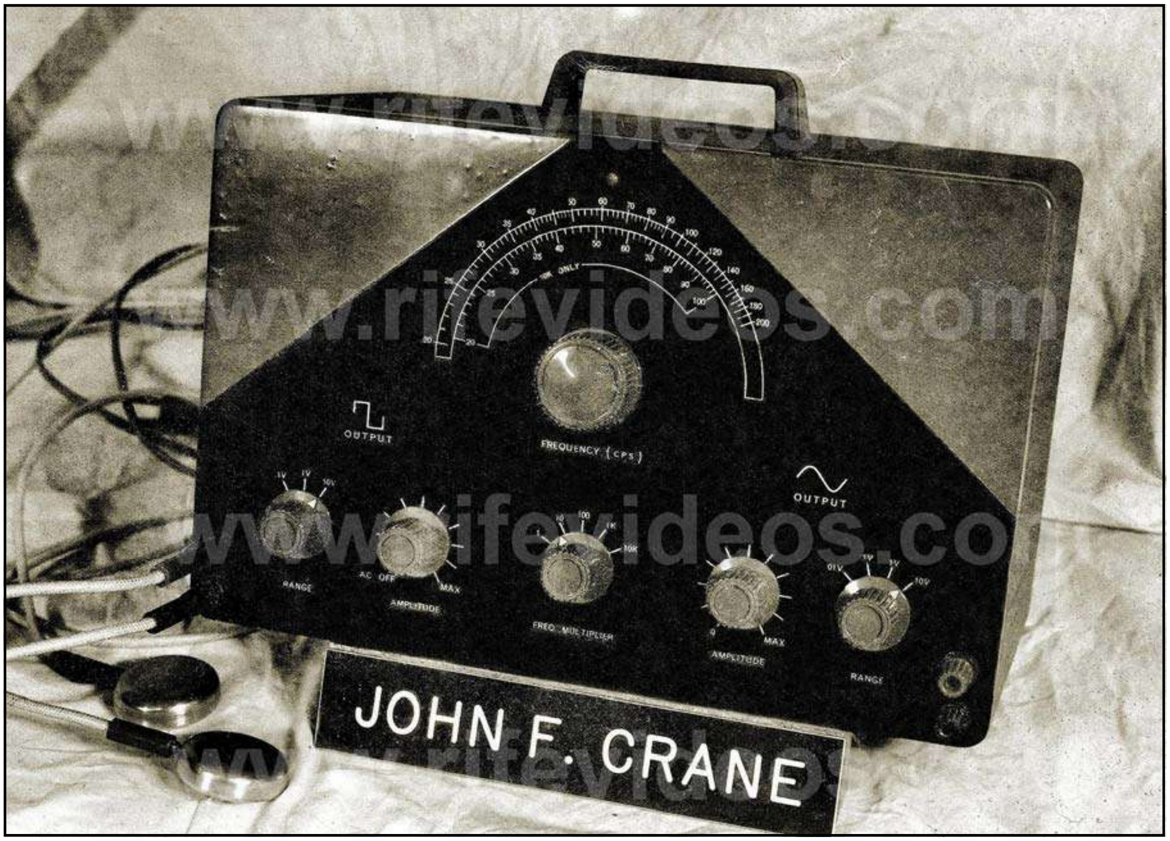
OUTPUT
FREQUENCY (CPS)
OUTPUT
RANGE
AMPLITUDE
FREQ. MULTIPLIER
AMPLITUDE
RANGE
JOHN F. CRANE

This next three photos, shown below and on the next page, are of another one of John Crane's pad instruments built in the 1960’s.

www.rife.org

JOHN F. CRANE

JOHN F. CRANE

www.rifevideos.com
www.rife.org

The next two photos, shown below, are of the pad instrument built back in 1961 by John Crane and John Marsh. They at that time were calling their organization the "Rife Virus Microscope Institute". This name is on the front of their instrument. The first photo is from back in 1961 and the second photo is a new photo of one of these instruments which has survived.

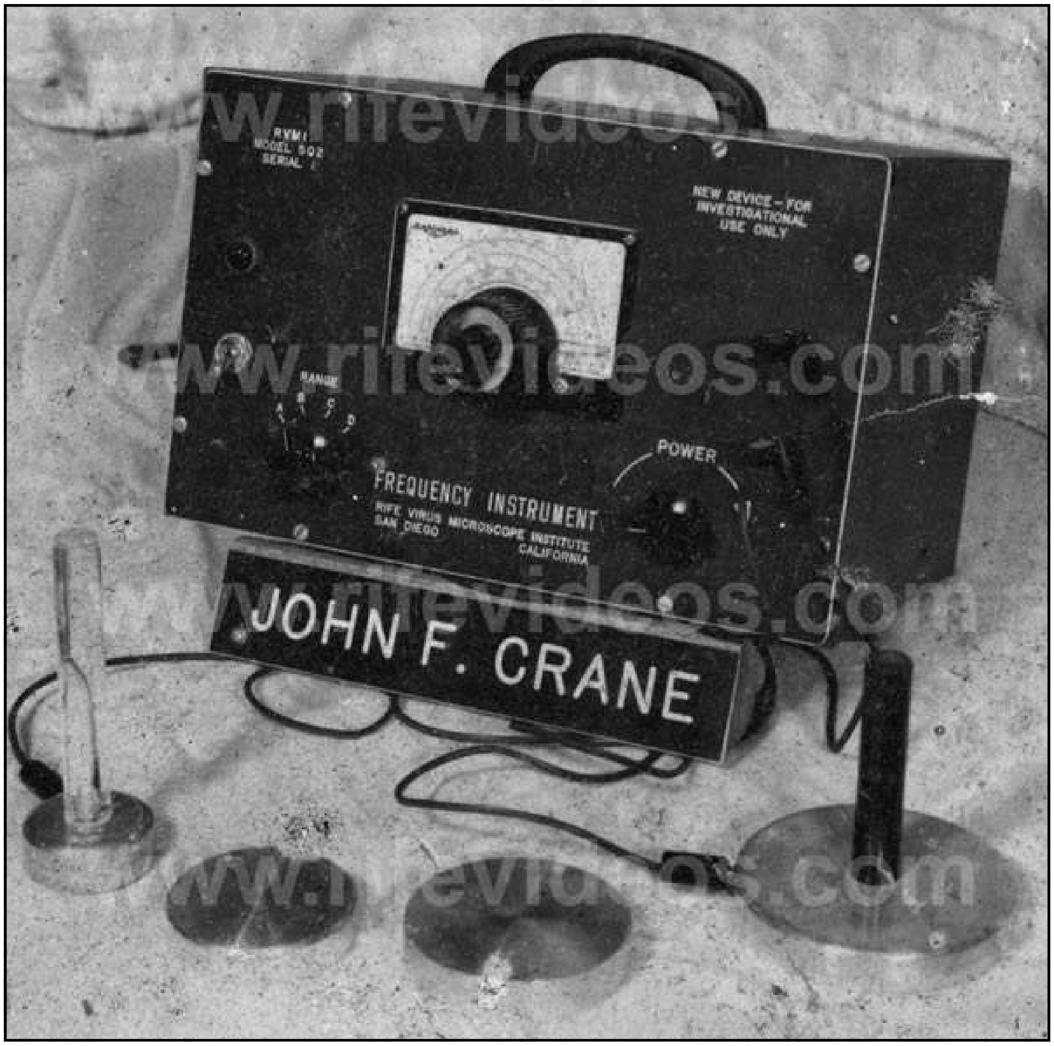
RVMI
MODEL 502
SERIAL 1
NEW DEVICE – FOR INVESTIGATIONAL USE ONLY
RANGE
A B C D
POWER
FREQUENCY INSTRUMENT
RIFE VIRUS MICROSCOPE INSTITUTE
SAN DIEGO CALIFORNIA
JOHN F. CRANE
www.rifevideos.com

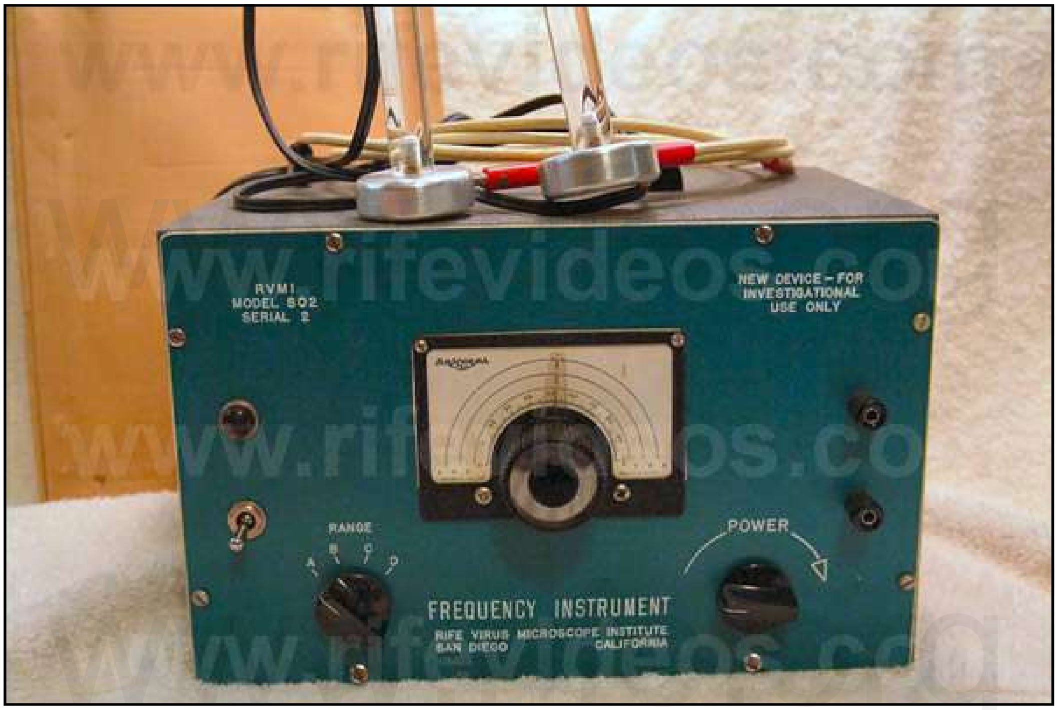
RVMI
MODEL 502
SERIAL 2
NEW DEVICE – FOR INVESTIGATIONAL USE ONLY
RANGE
A B C D
POWER
FREQUENCY INSTRUMENT
RIFE VIRUS MICROSCOPE INSTITUTE
SAN DIEGO CALIFORNIA
www.rifevideos.com

The next two photos, shown below, are pictures of one of John Marsh's pad instruments which he built in the 1980's. The first photo is from the 1980's and shows the aluminum disks which he was using at that time. The second photo shows this same instrument with another pad instrument which he built.

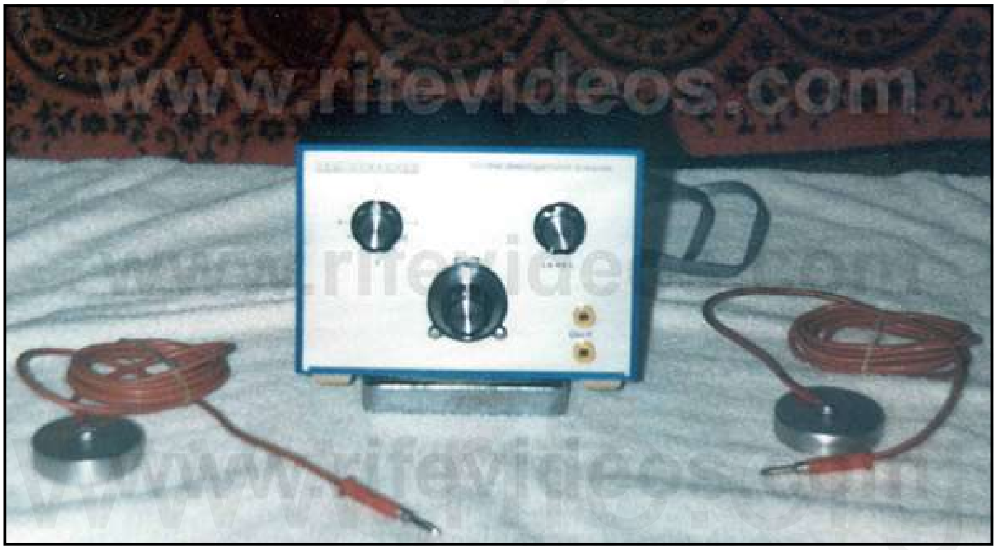

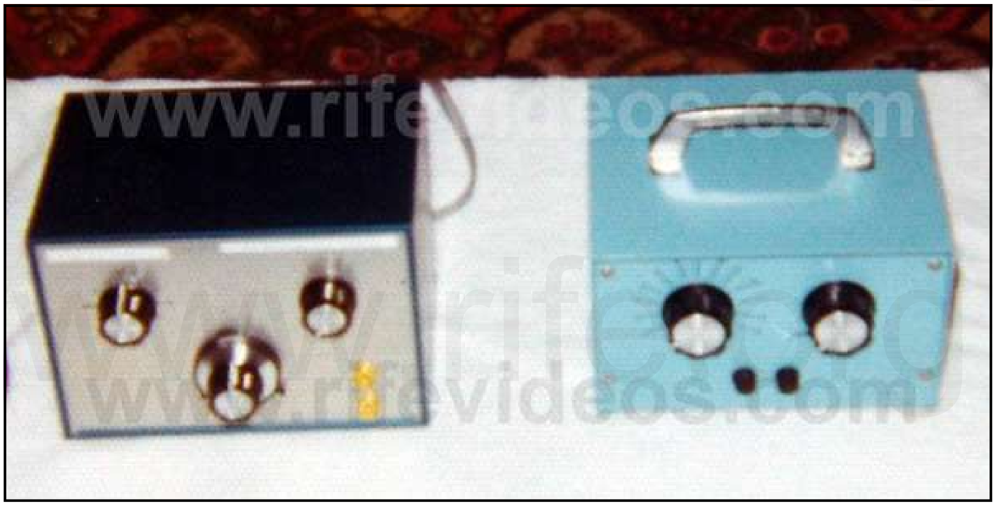

The next photo, shown below, is a new photo take of an instrument which John Marsh owned and gave to his nurse. It is similar to the instrument show in the first photo above.

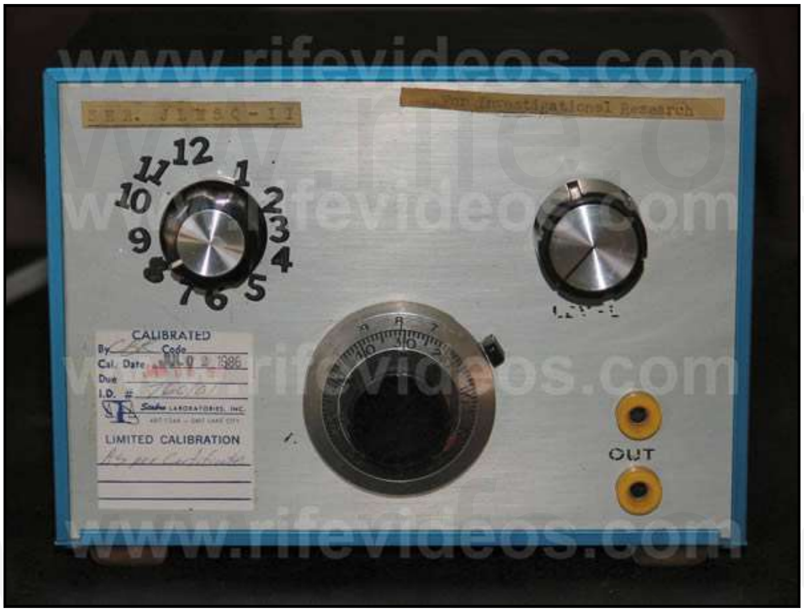
For Investigational Research
12
1
2
3
4
5
6
7
8
9
10
11
CALIBRATED
By
Code
Cal. Date JUL 0 2 1986
Due
I.D. #
LABORATORIES, INC.
LIMITED CALIBRATION
OUT

The next two photos, shown below, are of a pad machine that John Marsh made out of a 1950's Heathkit frequency generator. Almost all of the pad type machines that both John Crane and John Marsh built were really made from off-the-shelf frequency generators without any changes being made to them. Many times they would either replace the faceplate with their own or they would cover it so that no one would know that they were just using an off-the-shelf frequency generator. Above in one of the photos of John Crane's machine you can see that he covered the original faceplate so no one would know what he was doing. In the next two photos you can see that John Marsh did the same thing with this one. The first photo was taken by John in the 1980's and the second photo is a new photo of this instrument.

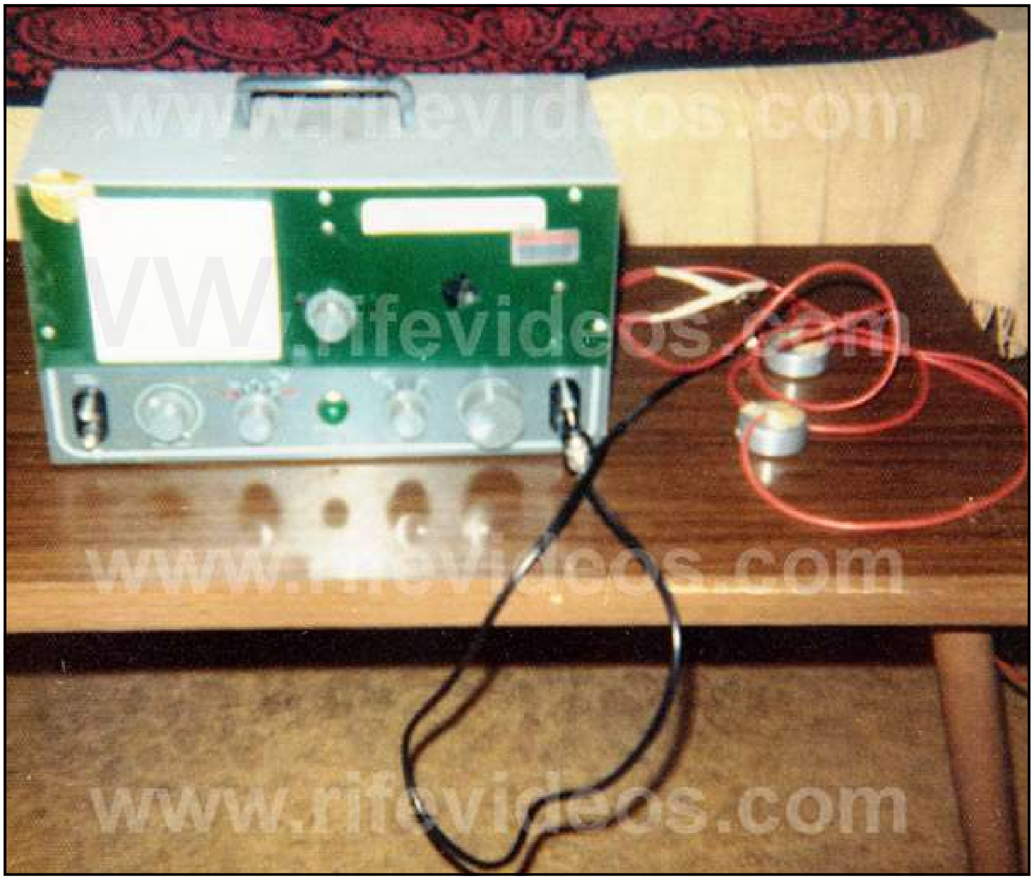
www.rifevideos.com

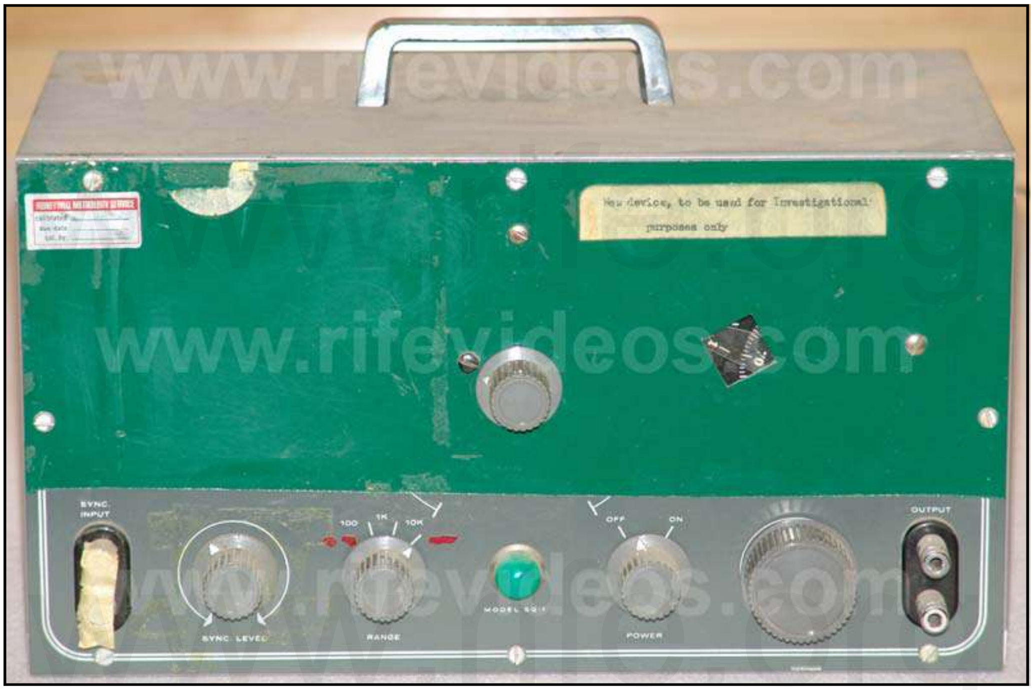
New device, to be used for Investigational purposes only
www.rifevideos.com
SYNC. INPUT
100
1K
10K
OFF
ON
OUTPUT
SYNC. LEVEL
RANGE
MODEL SQ-1
POWER

www.rife.org

Chapter Summary: Pad type audio frequency instruments have done a lot of good over the years. But without an RF carrier frequency properly match with the correct audio frequencies Philip Hoyland's sideband method cannot be used with them. These instruments need to have an RF carrier frequency or the ability to output the exact frequency that Dr. Rife used. We know from the documents that Dr. Rife preferred using the exact frequency rather than Philip Hoyland’s sideband method, but either method will work and Dr. Rife used both. Dr. Rife’s frequency range included both the audio range and the RF range of frequencies from 1 Hertz to about 1,800,000 Hertz. Philip Hoyland's design produced higher harmonic frequencies of Dr. Rife's original frequencies in the 2,000,000 and 3,000,000 Hertz range.

The so called pad "Rife Machines” that are sold today which cannot produce these frequency ranges cannot produce Dr. Rife's results. Make sure that the frequency generator you purchase can produce both the audio and RF frequencies that Dr. Rife used. Also it should be understood that the pad type instruments are not able to output the power that the ray tube instruments can. Pad style instruments that do not use an RF carrier frequency or the higher RF frequency range generally do not exceed one half of one watt of power. This is because the body cannot take more that about one fifth of one watt before the electrical current begins to lock up the muscles of the user. When an RF carrier frequency is used then the body does not react to the electricity and this makes it so that higher power levels can be used up to the 10 to 12 watt range. Dr. Rife's ray tube instruments had power ranges from 50 to 200 watts output. The Rife Ray #5 or Beam Ray Clinical instrument power output was between 40 to 50 watts.

In the next chapter we will look at an AZ-58 type of ray tube instrument built by John Marsh in the 1970's.

www.rife.org

Chapter #15

John Marsh's 1970's Beam Ray replica Rife Machine

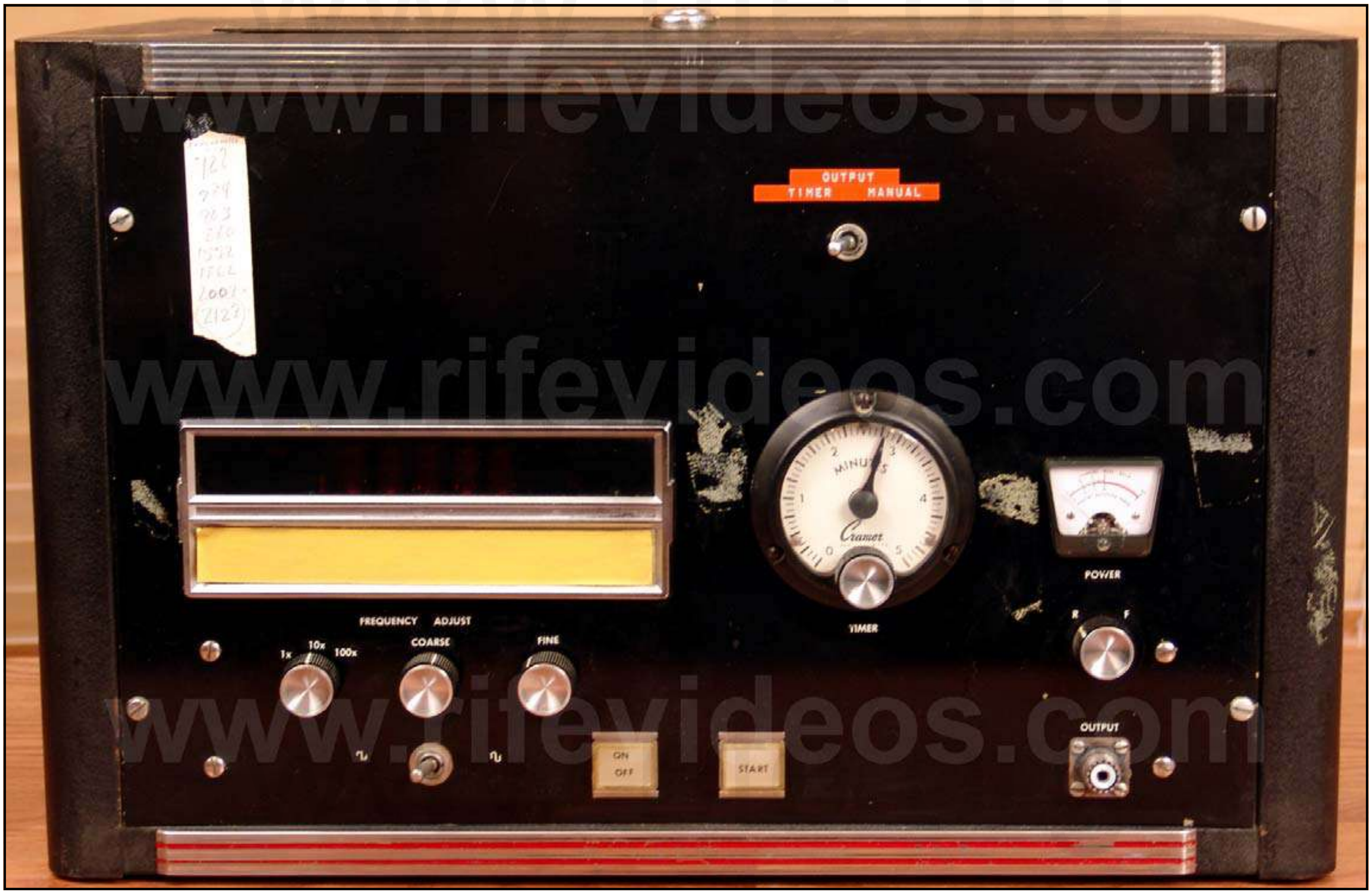

1. Used a ray tube.
2. Carrier frequency was 4.150 MHz.
3. Modulated sine and square wave audio frequencies onto the sine wave carrier frequency.
4. Power usage was about 460 watts. Output to the ray tube about 40 watts.

John Marsh had this replica of the Beam Ray Clinical Rife Machine Model #JLMSQ-1A built back in November 1971 for $3,800. It was completed by January 1, 1972. He and John Crane were under court order not to associate with each other. Because of this court order they went their separate ways but communicated often through phone calls and letters. John Crane stayed in California and John Marsh went to Colorado but eventually settled back in SLC, Utah until his death in 1987. All of his Rife Machines and Rife information were given to his nurse before his death. His equipment and documents were obtained from her in 2012. Since we were able to obtain this instrument we have been able to properly date it and take better photos of the complete instrument. Some of this information and new photos are shown below and are now apart of this report.

The photo at the top of the next page shows the two chassis that were connected by wires with their covers taken off. This instrument was a mix of both tube technology and modern solid state components.

www.rife.org
www.rifevideos.com

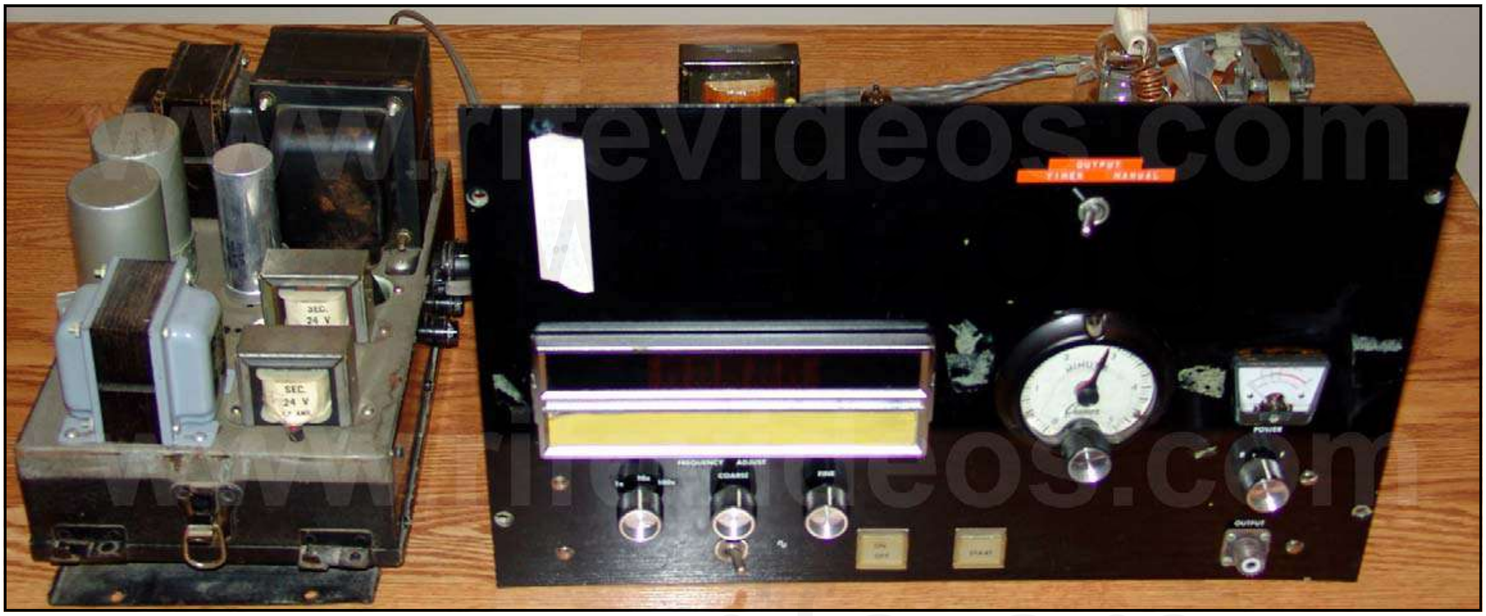
OUTPUT
TIMER MANUAL
SEC. 24 V
SEC. 24 V
MINUTES
POWER
FREQUENCY ADJUST
COARSE
FINE
ON OFF
START
OUTPUT
www.rifevideos.com

In the next photo, shown below, you can see that the instrument could output both sine and square wave audio frequencies. The frequency range of the audio oscillator went from 20 Hertz to 20,000 Hertz over three bands. John Marsh initially want it to have a frequency range from 0 to 100,000 Hertz. It had coarse and fine adjustments for the audio frequencies. Above those knobs we see the digital readout window of the frequencies. John Marsh put in a modern solid state audio oscillator with a digital readout. To the right of the digital readout we see a timer with a range of up to 5 minutes.

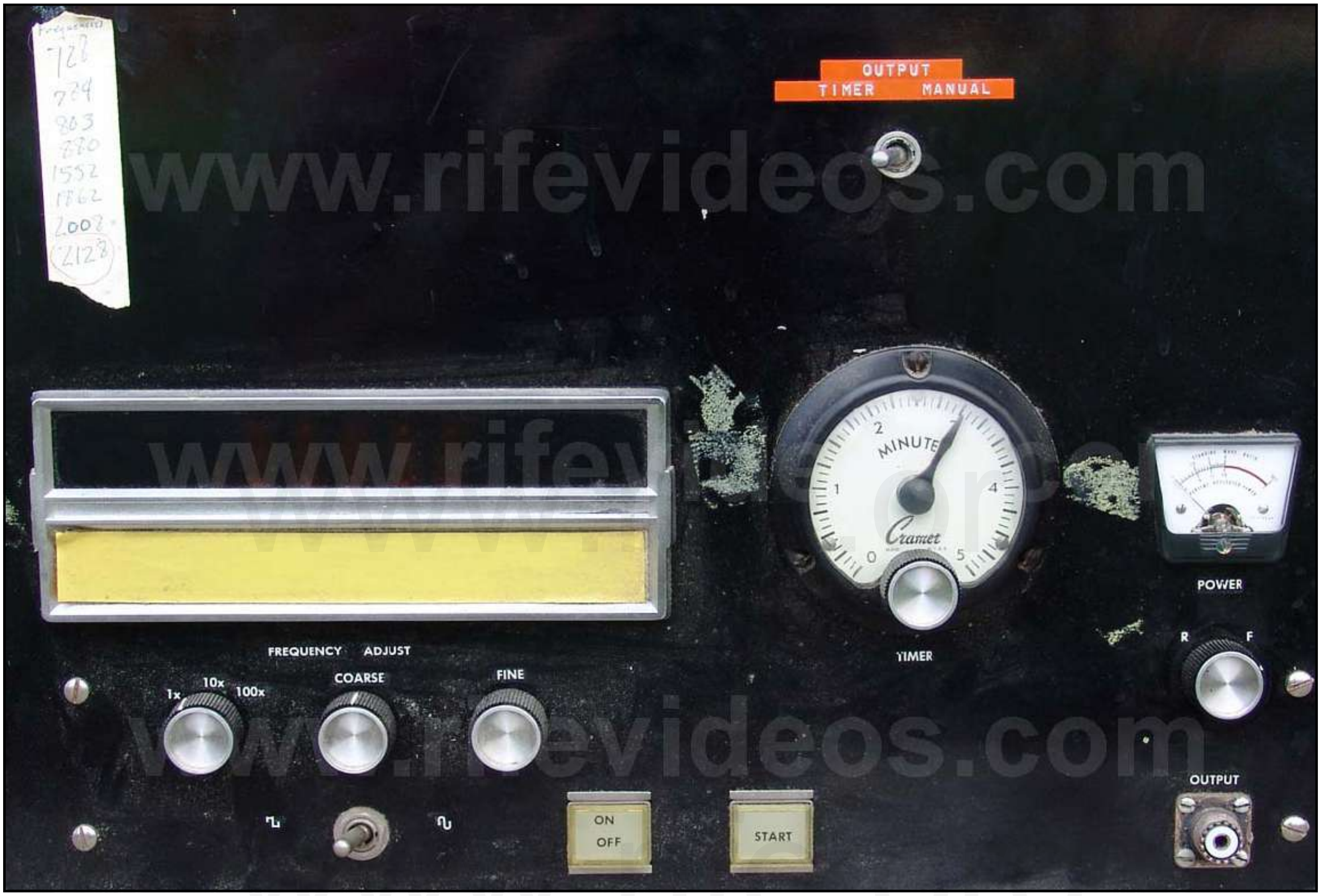
Frequencies
728
784
803
880
1552
1862
2008
2128
OUTPUT
TIMER MANUAL
2
MINUTES
1
4
0
5
Cramet
TIMER
POWER
R
F
FREQUENCY ADJUST
1x
10x
100x
COARSE
FINE
ON
OFF
START
OUTPUT
www.rifevideos.com

www.rife.org

To the right of that timer, shown below, in the first photo on the left, we see the power meter which is written on it “Standing wave ratio” and “Percent reflected power”. The knob below the meter was for adjusting the standing wave ratio. Below the knob is where the ray tube was connected. John Marsh used the CB antenna style connection instead of the banana jack method use in the original 1953 AZ-58. In the second photo, shown below on the right, you can still see an up close view of the John Marsh's masking tape with the 1950’s frequencies written on it.

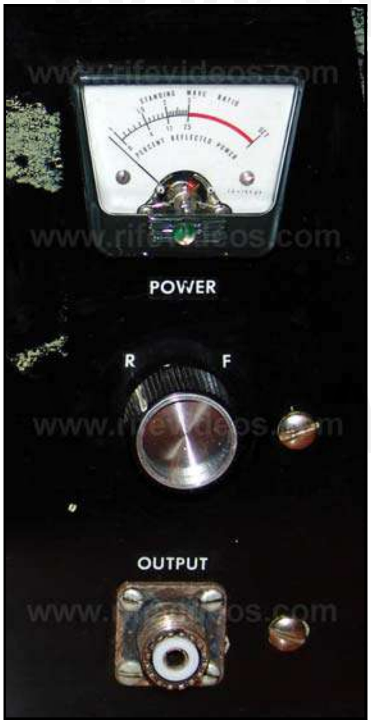
STANDING WAVE RATIO
PERCENT REFLECTED POWER
POWER
R
F
OUTPUT

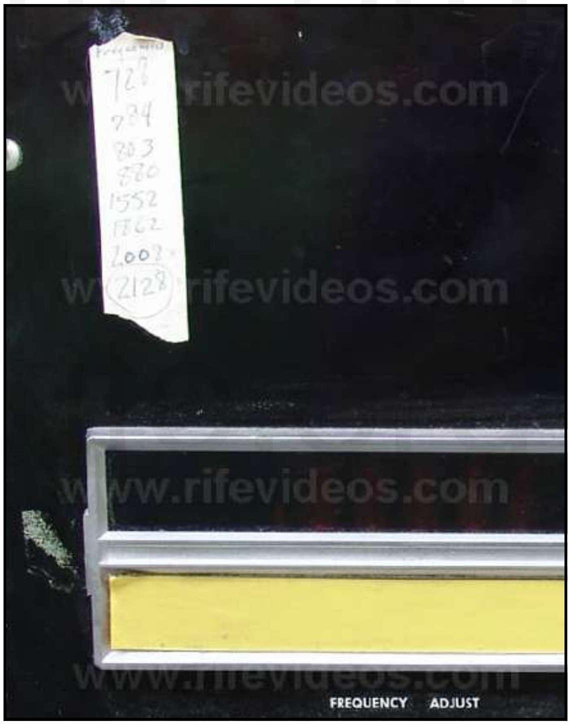
Frequencies
728
784
803
880
1552
1862
2008
2128
FREQUENCY ADJUST

The next photo, shown below, is a top view of the chassis with the case removed. You can clearly see that John Marsh used a mix of old tube technology and solid state electronics.

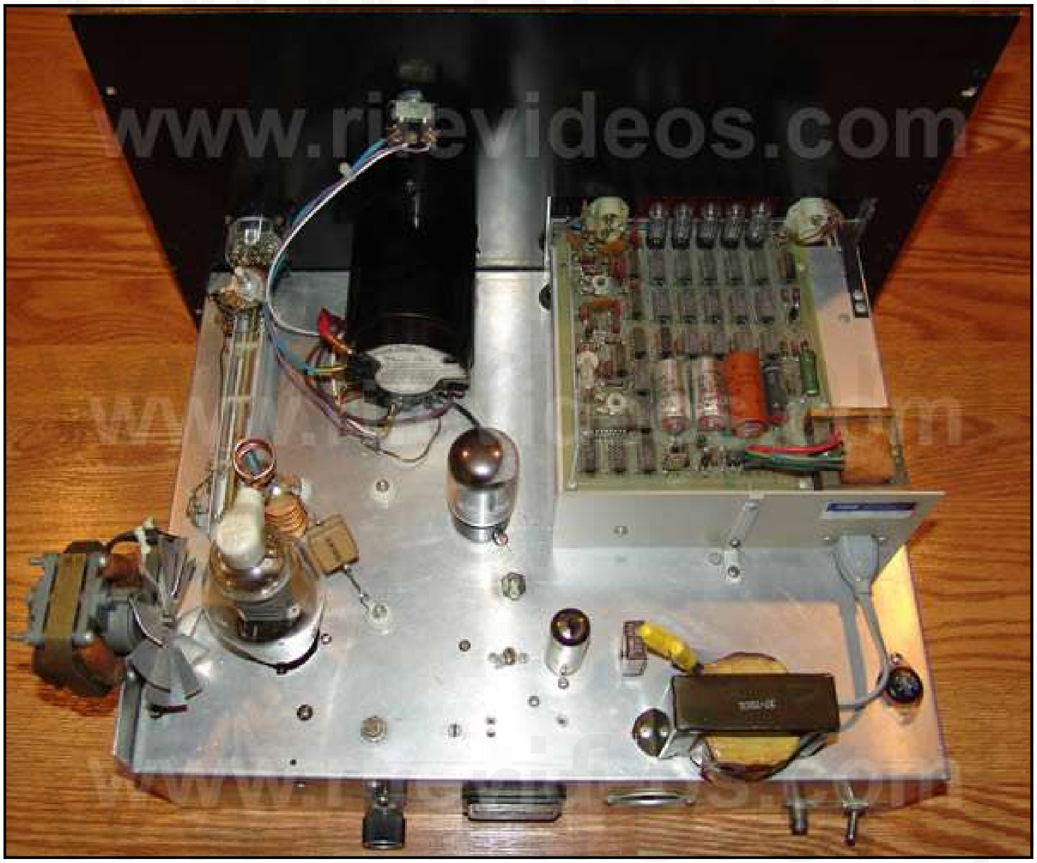

www.rife.org

The next four photos, shown below and on the next page, are close up photos of the top of the chassis. The first photo is the built in timer. The second photo is the solid state audio oscillator.

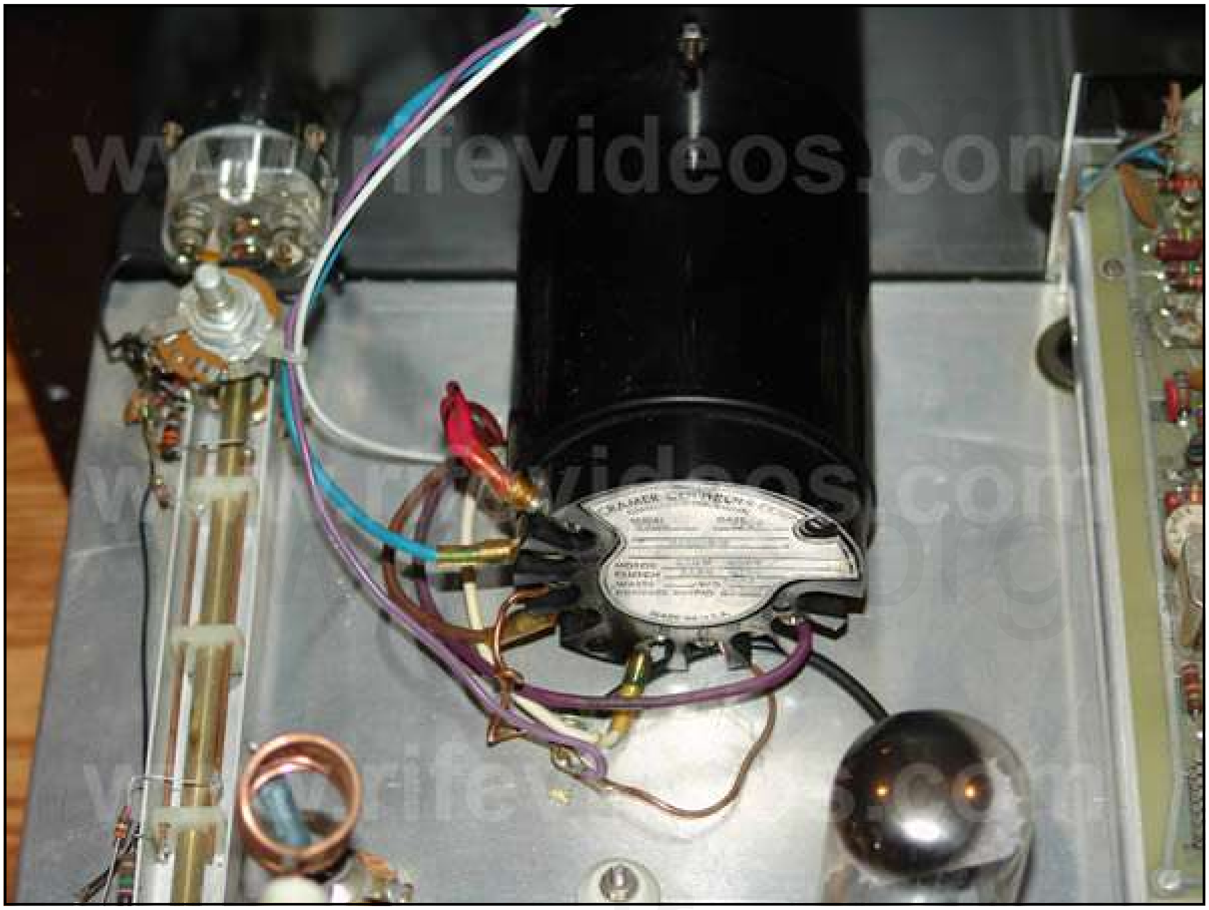

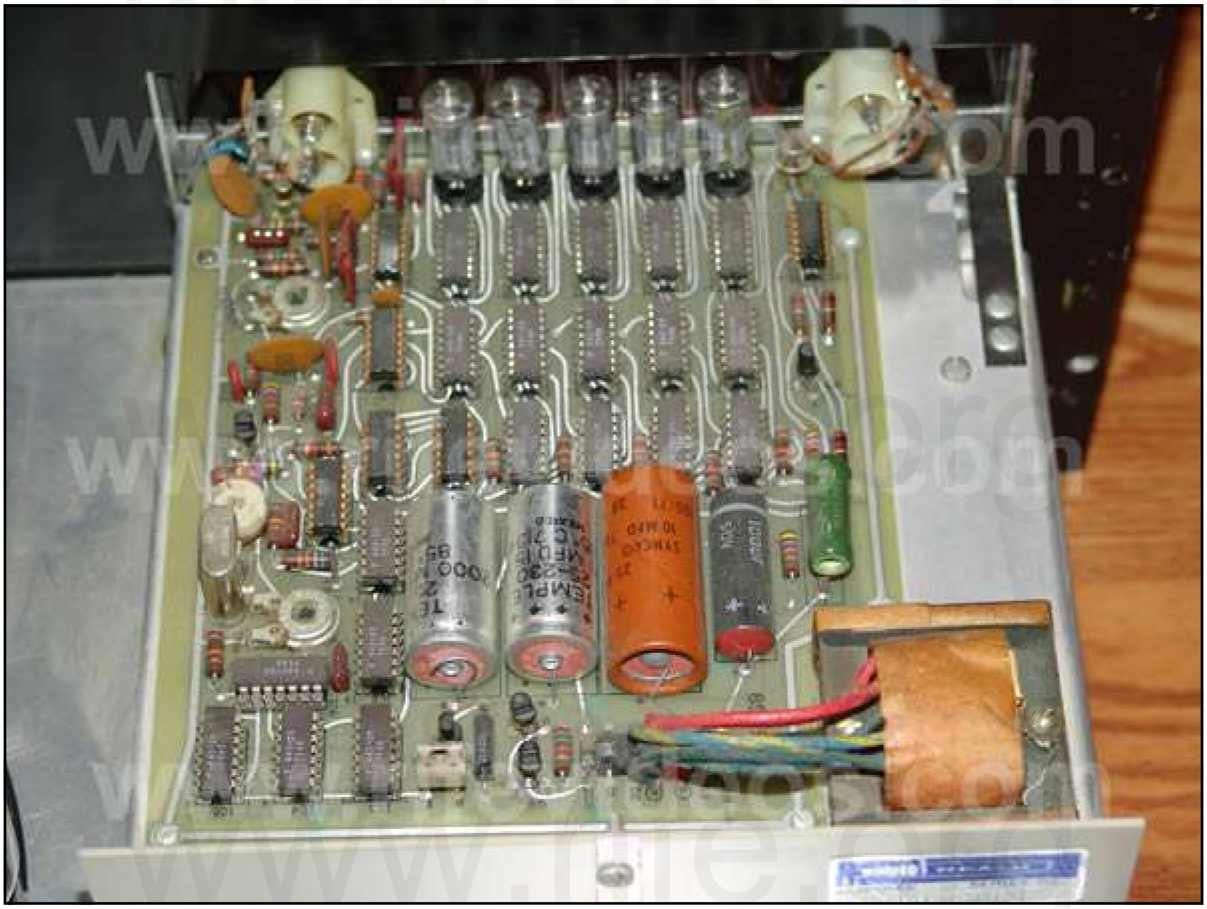

www.rife.org

The third photo shows the 811a main power output vacuum tube. The fourth shows the transformer that powers the audio oscillator board.

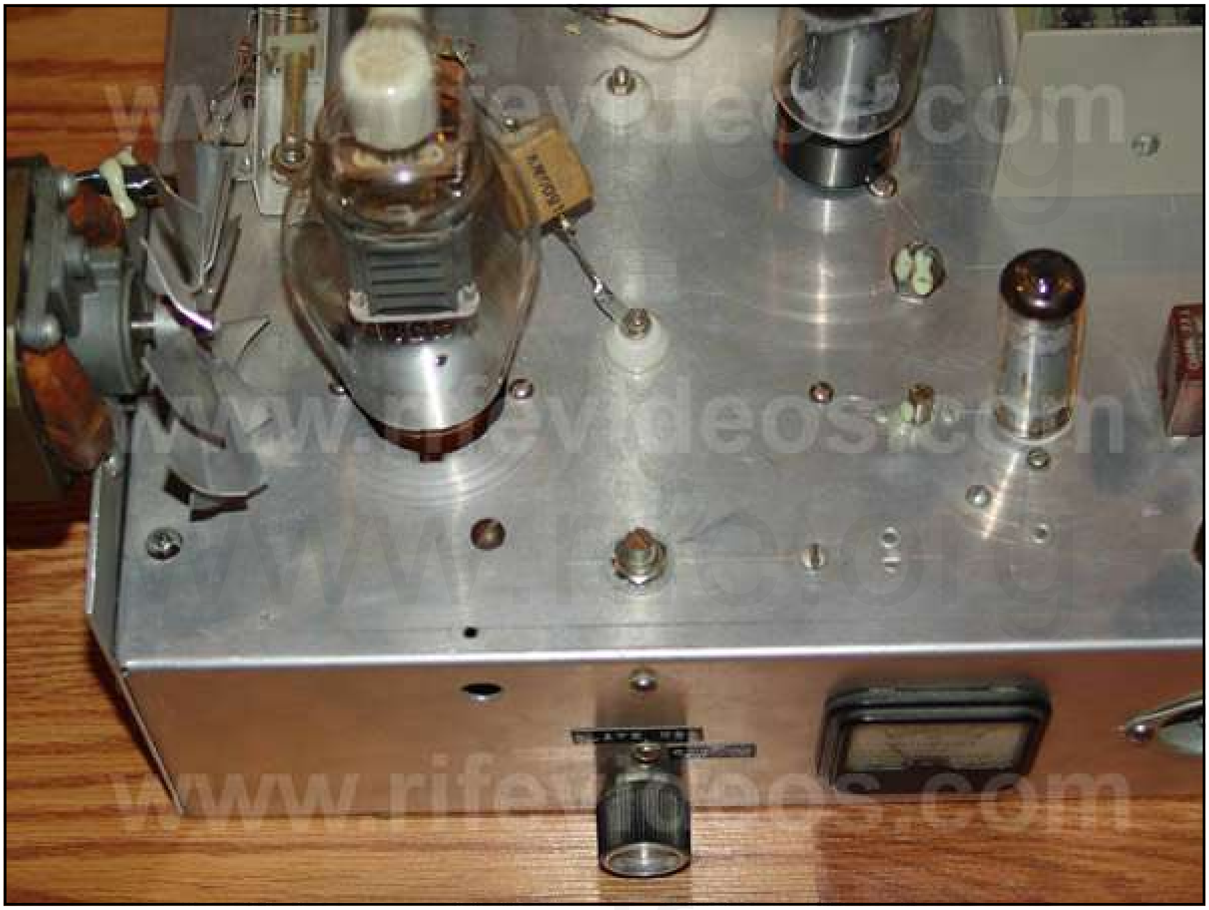

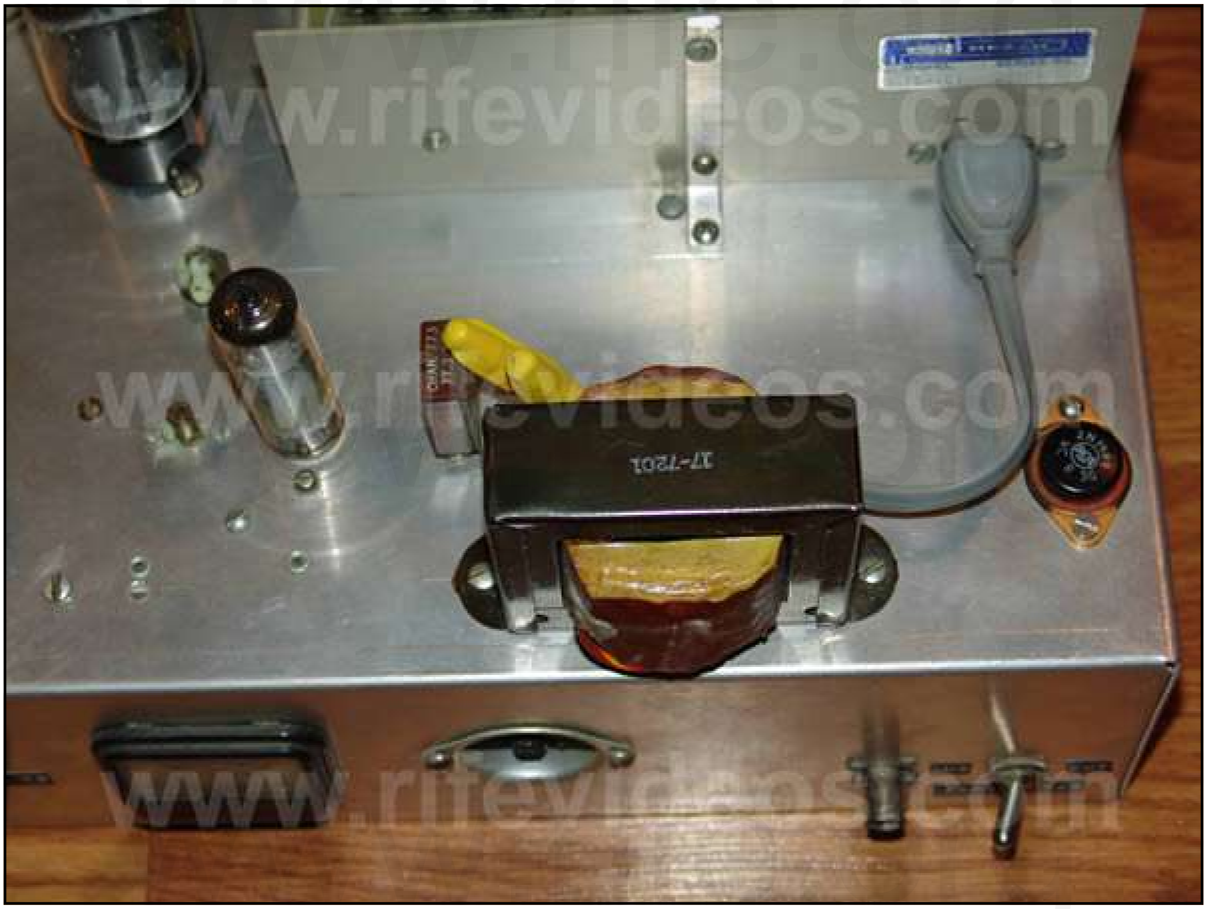

www.rife.org

The next photo, shown below, is a back view of the instrument. The meter is a D.C. milliamp meter. The socket to the right of the meter is for connecting the smaller box that has the power transformers.

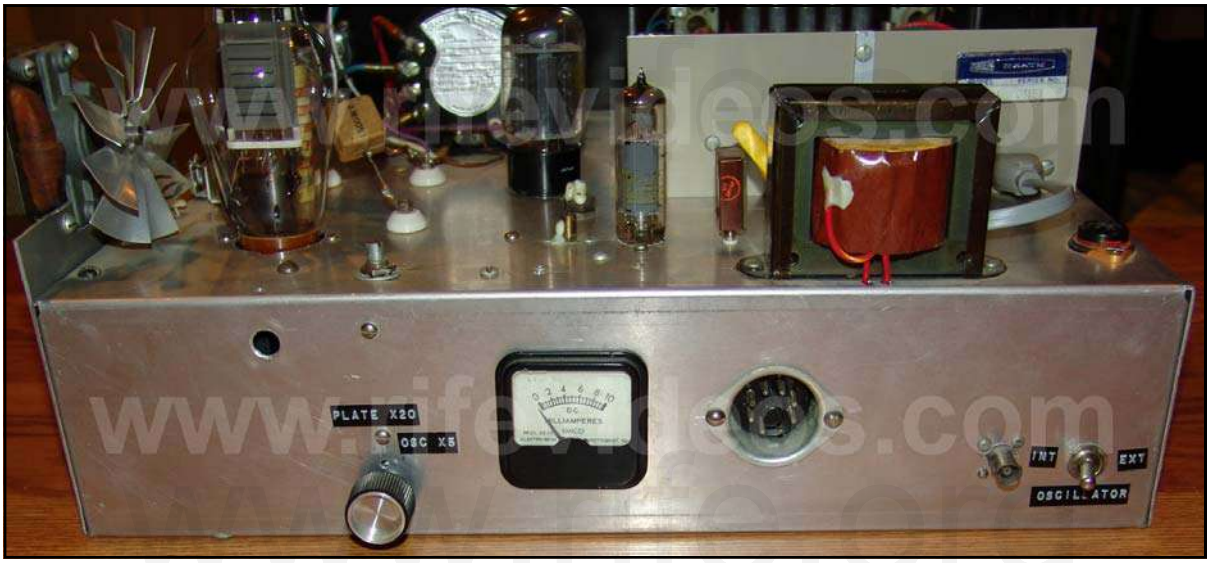
PLATE X20
OSC X5
D.C.
MILLIAMPERES
INT
EXT
OSCILLATOR

In the next four photos, shown below, give you an understand of the various components. The first photo labeled #1 is looking at the inside front of the instrument and shows the audio oscillator. To the left of the audio oscillator is the five minute timer. Photo #2, is looking at the back of the instrument and shows the three vacuum tubes. The three photos of vacuum tubes labeled #4, #5 and #6 show a clear view of the 811a, 6L6GC and 6GK6 vacuum tubes and their placement into the chassis. Photo #3 is one of John Marsh's original 1970 photo which shows the underside of the chassis where you can see the RF tank coil that was fixed at 4.150 MHz (4,150,000 Hertz).

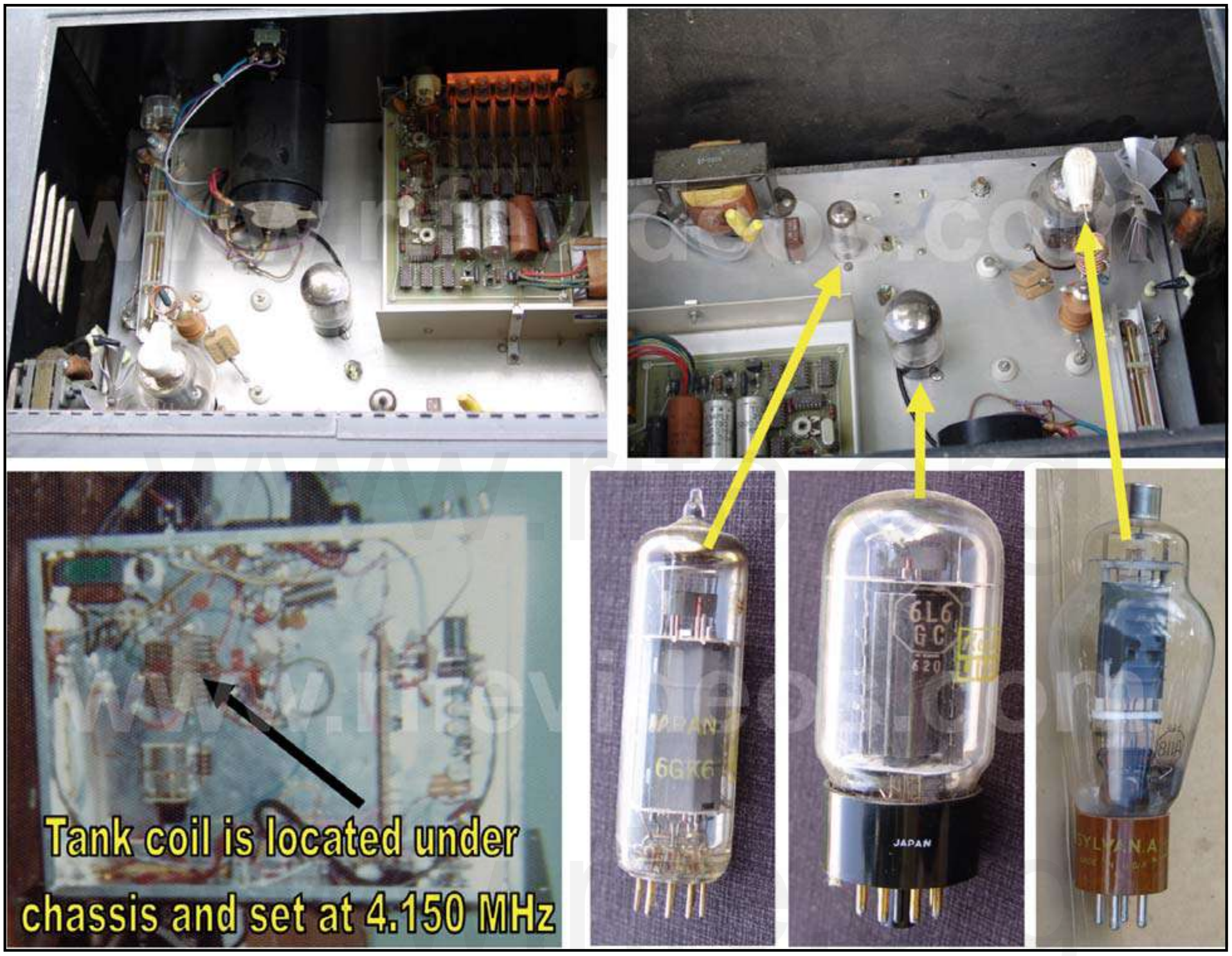
Tank coil is located under chassis and set at 4.150 MHz
6GK6
JAPAN
6L6 GC
811A
SYLVANIA

The next photo, show below, is a side view of the 811a tube. You can also see the RF choke right in front of the 811a tube with a small coil on top of it. This coil helped eliminate any parasitic oscillations.

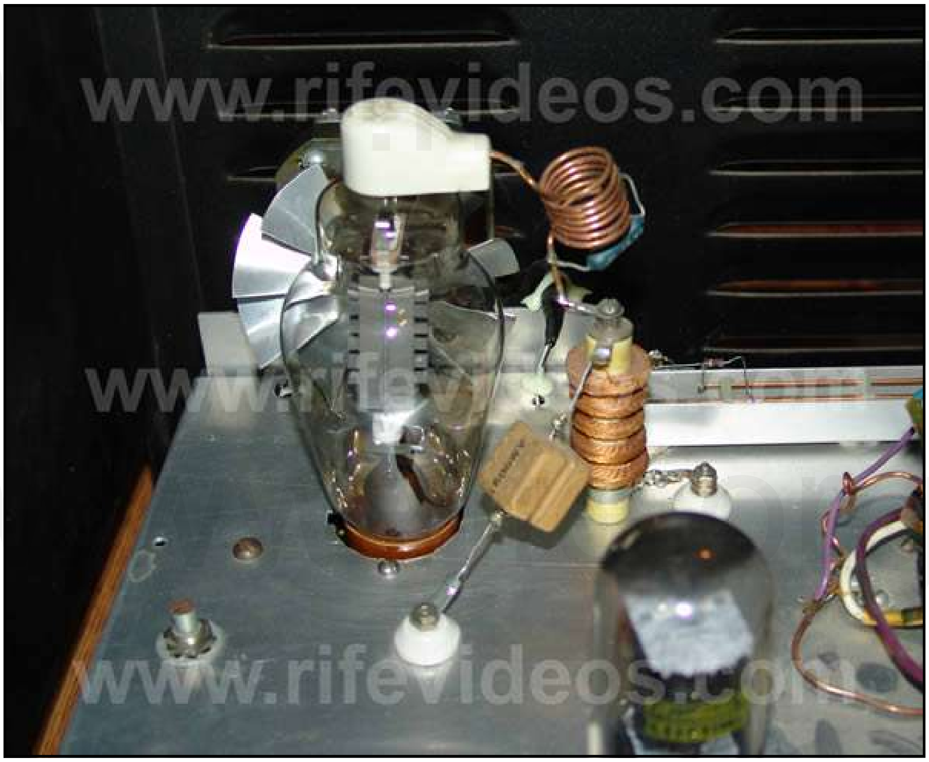

www.rifevideos.com
www.rife.org

The first of the next two photos, shown below and on the next page, is a photo which was taken of the underside of the chassis. The first photo, was one of the three pictures we have of the underside of the chassis. This photo was not very detailed and was taken back in 1971 when the instrument was built.

In the second photo, shown below, (new clearer photo) the larger coil is the RF tank coil which was set to 4.150 MHz. The variable capacitor which has the black knob was used to tune the carrier frequency to 4.150 MHz.

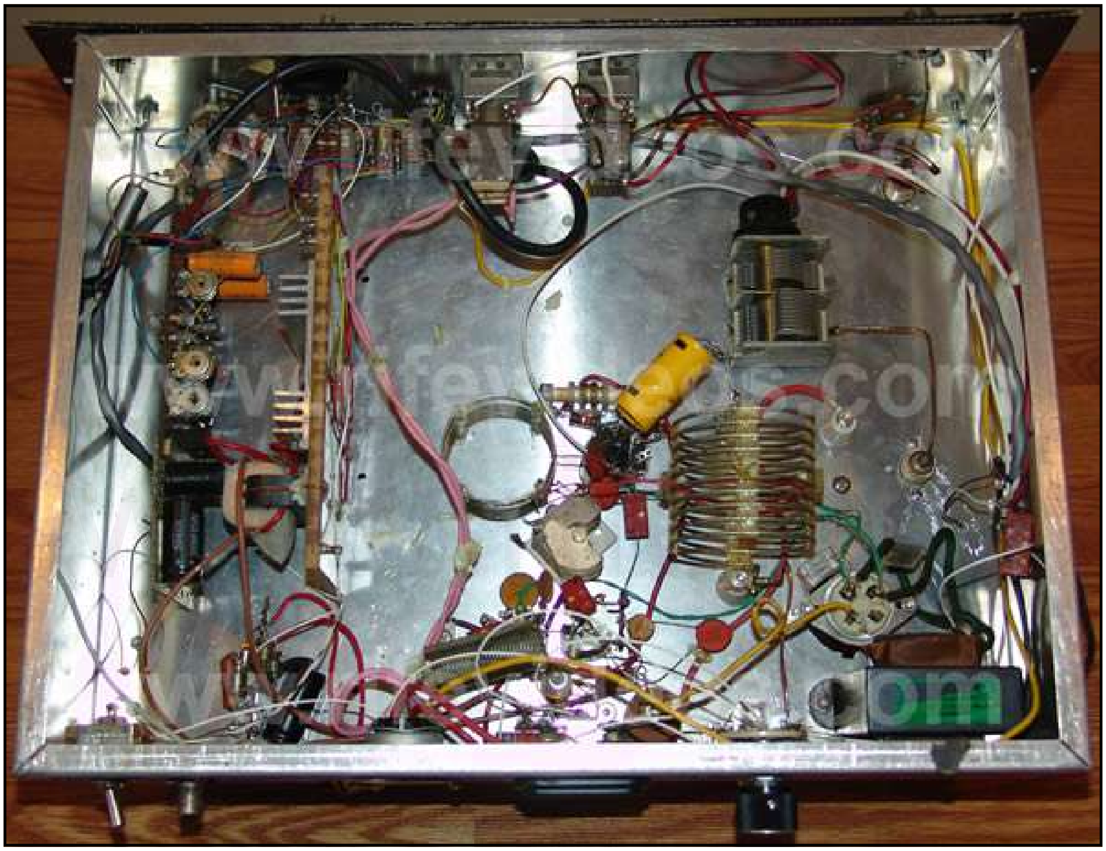

The next four photos, shown below and on the next page, are up close photos of the underside of the chassis showing the various components used to build this instrument.

www.rife.org

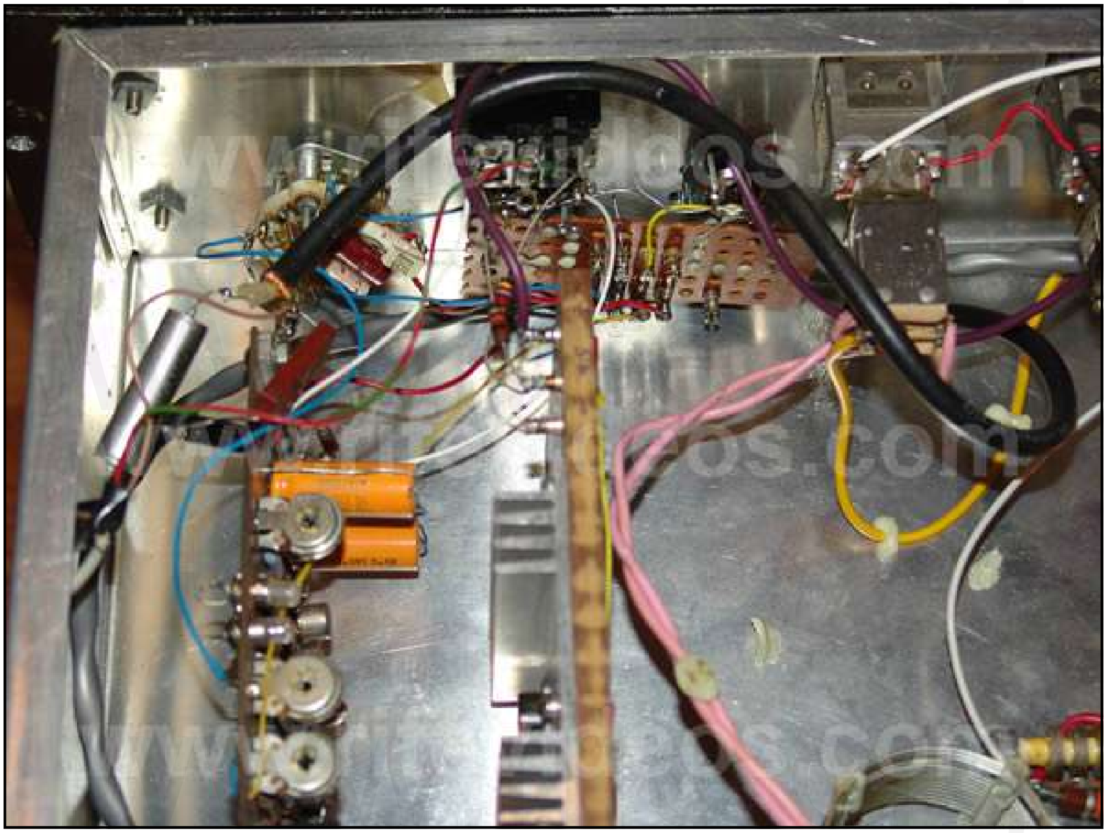

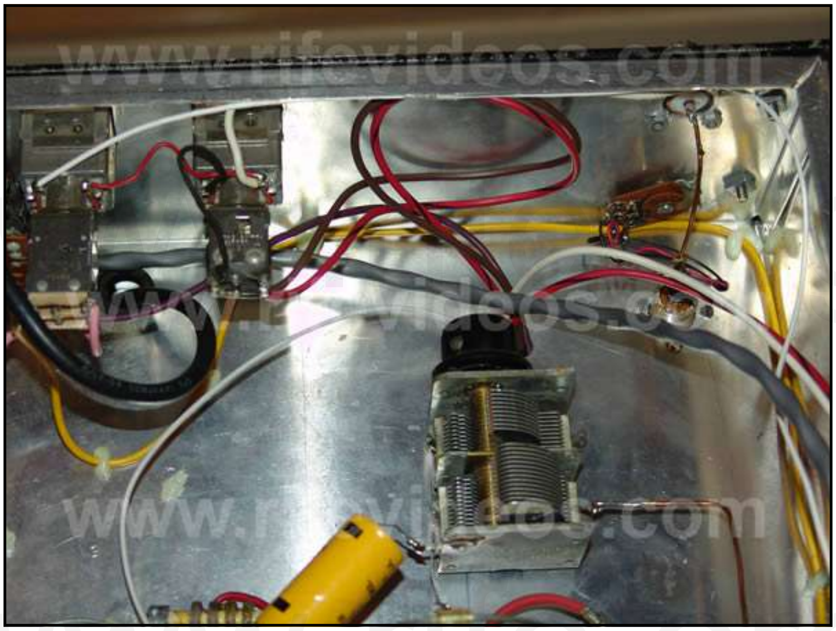

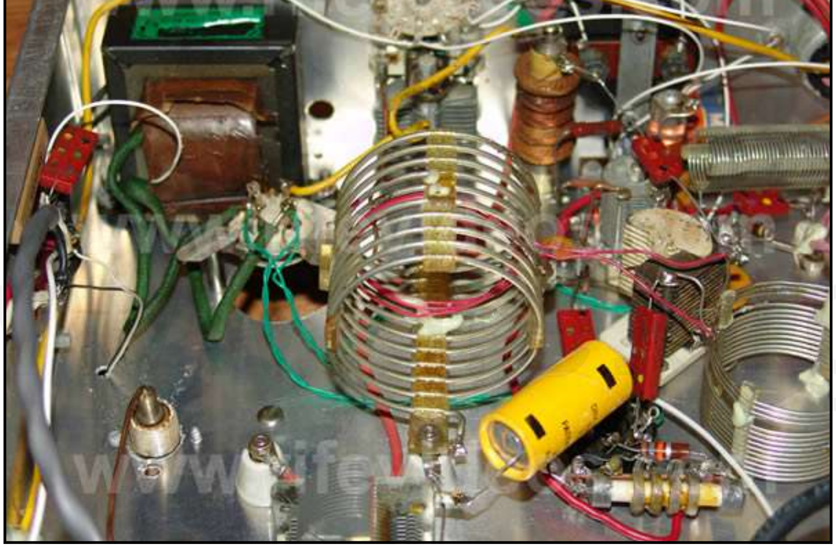

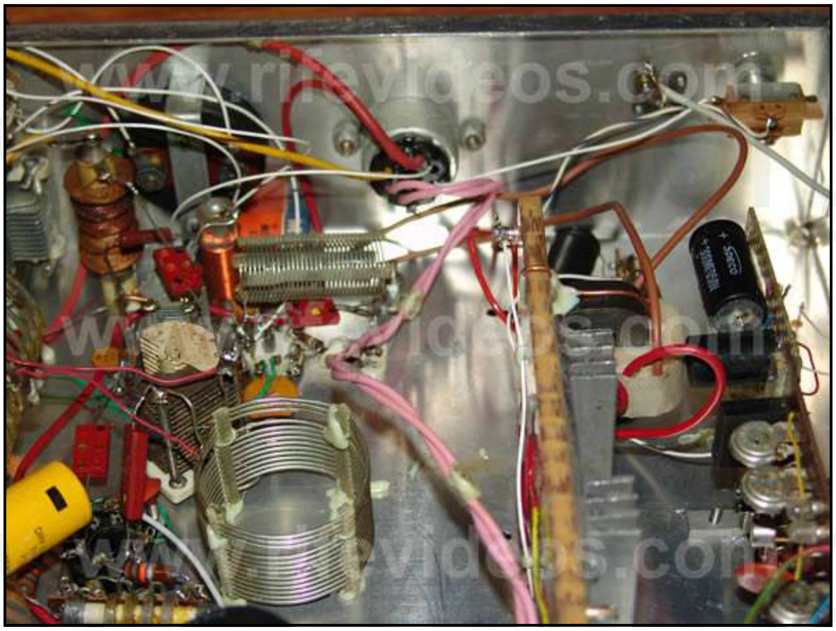

www.rifevideos.com
www.rife.org

The next two photos, shown below are of the inside of the small case. It contained almost all the transformers. The standard Beam Ray Rife Machine had two shelves in one case for components. The AZ-58 combined everything into one case but for some reason John Marsh used two cases to hold the components.

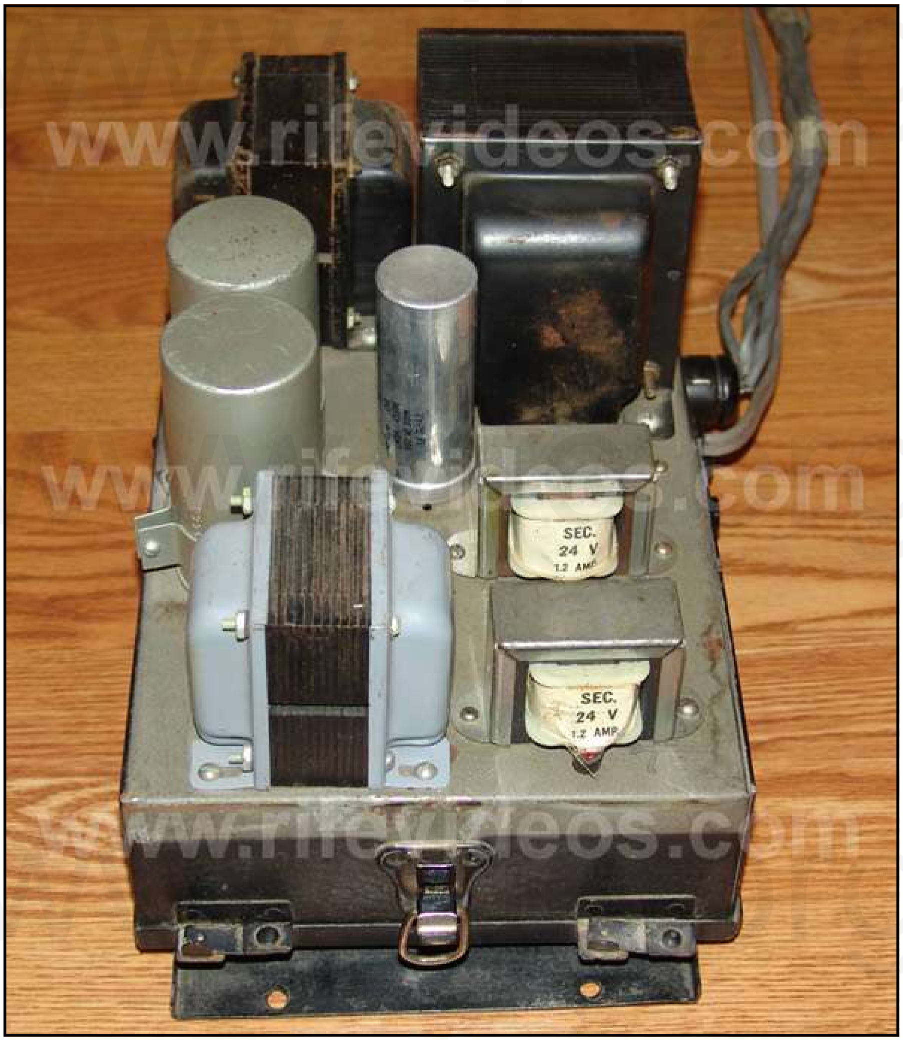

www.rifevideos.com

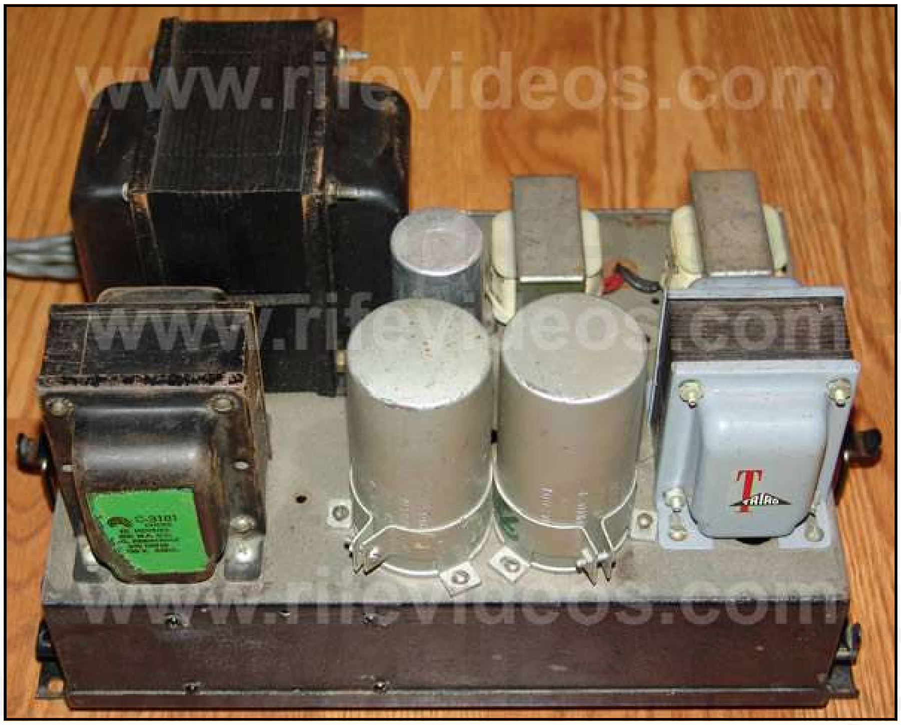

www.rifevideos.com

The next two photos, shown below, are of this instrument being used in a doctors office back in 1972 when John Marsh live in Colorado.

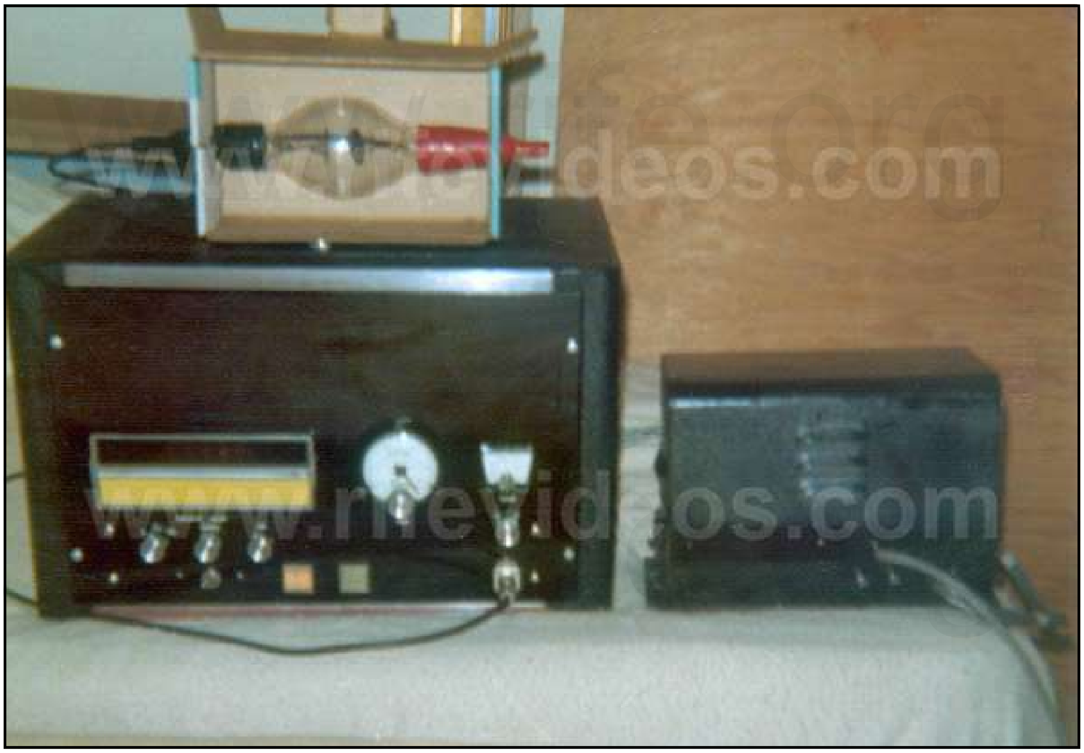

www.rifevideos.com
www.rife.org

The photo below is the ray tube which was used with this instrument. The ray tube still worked when we tested it.

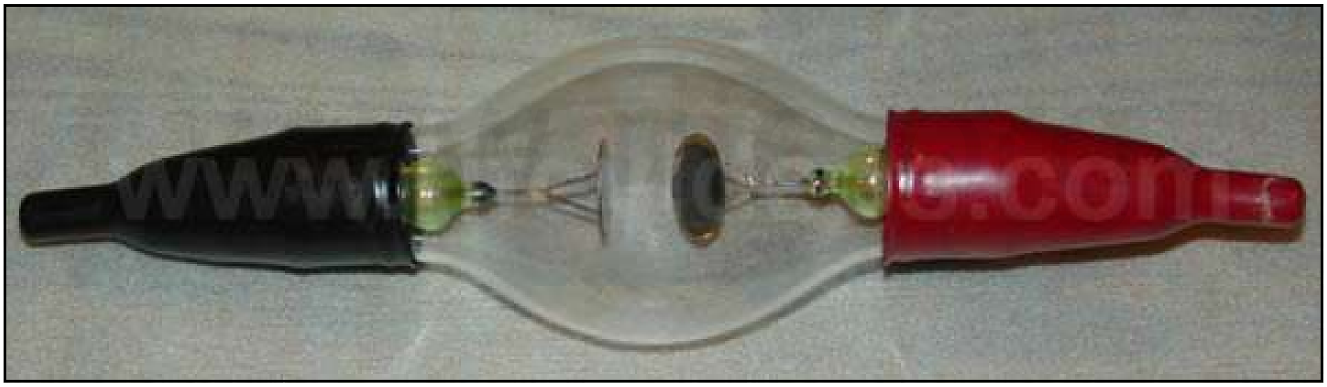

John Marsh's instrument, like the AZ-58, was a more modern replica version of the original Beam Ray Clinical Rife Machine. The carrier frequency that John Marsh chose to use with this instrument again clearly shows he did not understand the importance of the RF carrier frequency. He changed it from the 1953 AZ-58's 4.68 MHz to 4.150 MHz. The 4.150 MHz RF carrier frequency is probably one of the worst carrier frequencies he could have chosen using the AZ-58 low audio frequencies for the sideband method. In fact it would not be a good carrier for the higher audio frequencies either. We will explain again how to determine the best RF carrier frequencies to use in an instrument.

The method Philip Hoyland used to determine the best RF carrier frequency to use was by doing multiples of the BY (Sarcoma 1,529,520 Hertz) and the BX (Carcinoma 1,607,450 Hertz) frequencies. Logically, multiples of these frequencies are the best RF carrier frequencies to use because they were Dr. Rife's highest M.O.R. frequencies that he found. If you multiply the BY frequency by two you get 3,059,040 Hertz and if you multiply the BX frequency by two you get 3,214,900. So an RF carrier frequency in the 3,100,000 to 3,300,000 Hertz range would work well. Philip Hoyland used 3,300,000 Hertz. The next best range would be to multiply these two frequencies by a factor of three. The BY multiplied by three gives you 4,588,560 Hertz and the BX multiplied by three gives you 4,822,350 Hertz. So a carrier frequency in the 4,600,000 to 4,700,000 Hertz range would be the next best RF carrier frequency to use in an instrument. So you can see by the math that 4,150,000 Hertz would not be a good carrier frequency to use if you were going to used the sideband method that Philip Hoyland used when building the Rife Ray #5 or Beam Ray Clinical instrument. The RF carrier frequency should always be determined by multiples of the highest frequencies that Dr. Rife found for the various organisms. The 1953 AZ-58 had an RF carrier frequency of 4,680,000 Hertz. This carrier frequency would have worked very well had they understood the sideband method Philip Hoyland used. Since they lowered the audio frequencies instead of recalculating them to work on the sideband method then this also again reveals that they did not understand how Philip Hoyland's instrument really worked.

With the above understanding it is easy to see that the only reason you would use a 4.150 MHz RF carrier frequency is if you did not care what RF carrier frequency you used. The fact that they didn't really care what RF carrier frequency they used is without question since both John Marsh and John Crane have said in several documents and on audio tapes that the audio frequencies were the M.O.R. frequencies. The whole concept of using the sideband spacing method is to choose a carrier frequency that would work the best with all of the Rife Ray #4 higher frequency harmonics. Had John Marsh really understood the significance of the RF carrier frequency he would have chosen a different one. But just like the 1953 AZ-58 they changed it and relied on the square wave audio frequency harmonics rather than the sideband spacing method used in the original Beam Ray Clinical instrument. The Aubrey Scoon Beam Ray Clinical instrument replica was working on the sideband spacing method because the audio frequencies used with it were high enough to make the number of sideband harmonics reasonably low. So far Aubrey Scoon's instrument is the only instrument that we have seen, except for the original Beam Ray Clinical Rife Machine, which worked properly on the sideband spacing method.

www.rife.org

It is clear that the 1953 AZ-58 was not working fully on the sideband principle even though it could have. It appears that just by chance or accident some of the frequencies, like the BX frequency, worked because the RF carrier frequency was set at about 3.2 MHz by Dr. Stafford. Just the fact that they lowered the audio frequencies by a factor of 10 and then depended solely on square wave audio frequencies showed they didn't understand Philip Hoyland's sideband method. Had Philip Hoyland revealed how his Beam Ray Clinical Rife Machine worked a lot of confusion could have been avoided. This machine of John Marsh's could have easily been changed to work properly on the sideband method. The audio frequency range was designed to go to 20,000 Hertz. If the RF carrier frequency was changed to 3,300,000 Hertz, which would have been easy to do, then most of the original audio frequencies could have been used. The only two that would need to have been re-calculated would have been the BX and the BY frequencies. This also would have been easy to do.

In the chart, shown below, the frequencies have been calculated for John Marsh's instrument. You will notice that the "Original 1950's AZ-58 Frequencies" (low audio frequencies) are almost a perfect match to the "Correct Sideband Frequencies." But before we place too much significance in this coincidence we need to keep in mind the "Number of Sideband Harmonics." These numbers are so high that almost any low frequency can be divided into the "Carrier Difference Frequency" and come out within a few Hertz of the "Correct Sideband Frequency." The audio frequency needs to be a great deal higher in order to make it so the sideband frequencies will work. This is due to the fact that power is lost in sidebands. We must keep in mind that the higher the audio frequency is, the lower the number of sidebands that will be created and the better they will work. So these low audio frequencies in the chart below are two low to work with an RF carrier frequency of 4.150 MHz. In fact they would be too low to work even if they were used with a 3.30 MHz RF carrier frequency as was used in the original Beam Ray Clinical instrument. Using the higher audio frequencies like Philip Hoyland used is the

John Marsh's Beam Ray Clinical Instrument Sideband Square Wave Audio Frequencies Based On A 4.150 MHz Carrier

Organism	Rife Ray #4 Frequencies	Higher Harmonic Frequencies	*Carrier Difference Frequency	1/10 Of A Meter Freq.	Number of Sideband Harmonics	Original 1950's AZ-58 Frequencies	Correct Sideband Frequencies
Actinomycosis or Streptothrix	192,000 Hz	4,032,000 or 21st	118,000 Hz	12 Hz	151	784 Hz	781 Hz
Anthrax	139,200 Hz	4,176,000 or 30th	26,000 Hz	6 Hz	41		634 Hz
B or E Coli Rod	417,000 Hz	4,170,000 or 10th	20,000 Hz	58 Hz	25	800 Hz	800 Hz
B or E Coli Virus	770,000 Hz	3,850,000 or 5th	300,000 Hz	198 Hz	193	1552 Hz	1554 Hz
BX Virus Carcinoma	1,604,000 Hz	4,812,000 or 3rd	662,000 Hz	858 Hz	311	2128 Hz	2129 Hz
BY Sarcoma	?1,530,000 Hz	4,590,000 or 3rd	440,000 Hz	780 Hz	219	2008 Hz	2009 Hz
Gonorrhea	233,000 Hz	4,194,000 or 18th	44,000 Hz	18 Hz	62	712 Hz	710 Hz
Pneumonia or Spinal Meningitis	427,000 Hz	4,270,000 or 10th	120,000 Hz	61 Hz	155	776 Hz	774 Hz
Staphylococcus Pyogenes Aureus	478,000 Hz	4,302,000 or 10th	152,000 Hz	76 Hz	209	727 Hz	727 Hz
Streptococcus Pyogenes	720,000 Hz	4,320,000 or 6th	170,000 Hz	173 Hz	193	880 Hz	880 Hz
Syphilis	789,000 Hz	3,945,000 or 5th	205,000 Hz	207 Hz	311	660Hz	659 Hz
Tetanus	234,000 Hz	4,212,000 18th	62,000 Hz	18 Hz	517	120 Hz	120 Hz
Tuberculosis Rod	369,000 Hz	4,059,000 or 11th	91,000 Hz	45 Hz	113	803 Hz	805 Hz
Tuberculosis Virus	769,000 Hz	3,845,000 or 5th	305,000 Hz	197 Hz	197	1552 Hz	1548 Hz
Typhoid Rod	760,000 Hz	3,800,000 or 5th	350,000 Hz	192 Hz	492	712 Hz	711 Hz
Typhoid Virus	1,445,000 Hz	4,335,000 or 3rd	185,000 Hz	694 Hz	99	1862 Hz	1869 Hz

www.rife.org

method that worked in the original equipment. If the frequency you want to hit is close to the RF carrier frequency then the lower the audio frequency you can use. But if the frequency is farther away from the carrier frequency then the higher the audio frequency you will need to use in order to make it work properly. Philip Hoyland could have used even higher audio frequencies since his audio oscillator in the original Rife Ray #5 or Beam Ray Clinical instrument would go to a little over 40,000 Hertz. He could have used frequencies up in the 30,000 to 40,000 Hertz range which would have worked with even less sidebands. But Philip Hoyland was also trying to hide the method he was using. So Philip Hoyland balanced his frequencies in order to make sure they would work and also not reveal the method he was using. He did accomplish his goal.

If you look at the “Number of Sideband Harmonics” it takes to hit the correct Rife Ray #4 “Higher Harmonic Frequencies” you will understand that this instrument could never work on the sideband spacing method using these low audio frequencies. None of the “Number of Sideband Harmonics” are less than 59 sideband steps and the highest is 750. The chance of this working would be almost zero. The best method to use with John Marsh’s instrument is the audio frequency square wave harmonic method. This is the method he used with his instrument.

Below in the chart is a list of the higher audio frequencies, 20,000 Hertz or lower, that could be used with John Marsh’s instrument and make it work using the harmonic sideband method. Many different audio frequencies could be calculated to work. We did the highest audio frequency for each organism. The BX and the BY frequencies probably would not work since the sidebands would have to go nearly 600,000 Hertz to hit the correct frequency. For this reason the RF carrier frequency should be changed. The best frequencies would always be the highest audio frequency you could use within in the 20,000 Hertz frequency range of the instrument.

John Marsh’s Beam Ray Clinical Instrument Higher Sideband
Square Wave Audio Frequencies Based On A 4.150 MHz Carrier

Anthrax	13,000 Hz	Streptococcus	18,889 Hz
B or E Coli Rod	10,000 Hz	Steptothrix	19,667 Hz
B or E Coli Virus	20,000 Hz	Syphilis or Treponema	18,636 Hz
BX Virus Carcinoma	19,471 Hz	Tetanus	15,500 Hz
BY Sarcoma	19,130 Hz	Tuberculosis Rod	18,200 Hz
Gonorrhea	14,667 Hz	Tuberculosis Virus	19,062 Hz
Pneumonia or Spinal Meningitis	17,143 Hz	Typhoid Rod	19,444 Hz
Staphylococcus	19,000 Hz	Typhoid Virus	18,500 Hz

Chapter Summary: The fact that John Marsh built these Beam Ray replica Rife Machines and used different RF carrier frequencies with the same audio frequencies conclusively proves that he never understood how the instrument was really intended to work. This also shows that John Crane didn't really know how the instrument was intended to work either. John Crane was doing the same thing that John Marsh was doing. The fact that Philip Hoyland did not reveal how the Beam Ray Clinical instrument really worked has affected Rife's work in a negative way to this very day.

In the next chapter we will look at the ray tube instrument that John Marsh built back in the 1980's when he lived in Salt Lake City, Utah.

www.rife.org

Chapter #16

John Marsh's 1980's Ray Tube Rife Machine

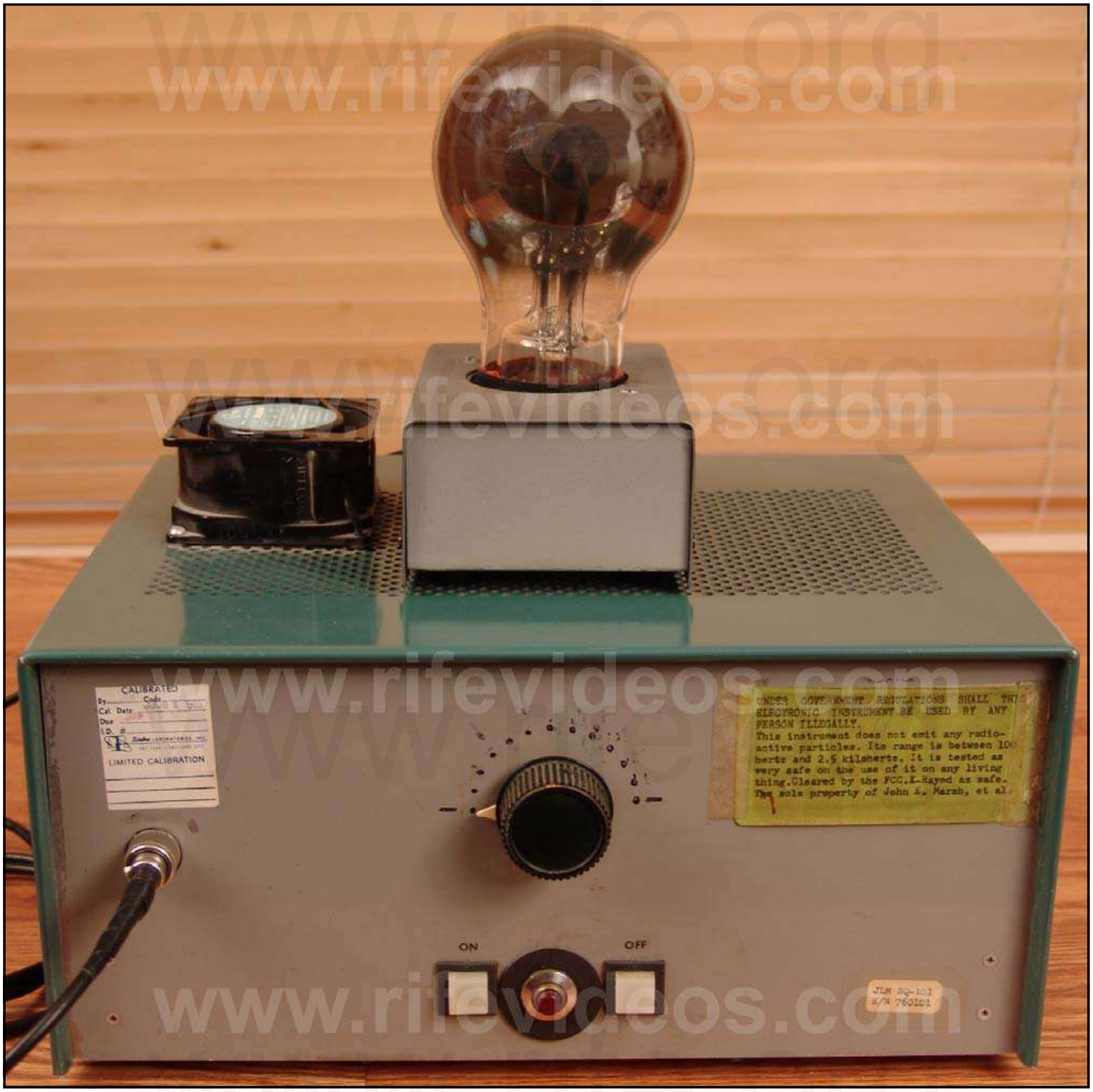
CALIBRATED
LIMITED CALIBRATION
UNDER GOVERNMENT REGULATIONS SHALL THIS ELECTRONIC INSTRUMENT BE USED BY ANY PERSON ILLEGALLY.
This instrument does not emit any radio-active particles. Its range is between 100 hertz and 2.5 kilohertz. It is tested as very safe on the use of it on any living thing. Cleared by the FCC. X-Rayed as safe. The sole property of John L. Marsh, et al.
ON
OFF
JLM SQ-101
S/N 760101

www.rife.org
www.rifevideos.com

1. Used a ray tube.
2. Carrier frequency was ?
3. Square wave audio frequencies modulated onto a sine wave carrier frequency.
4. Power usage was about 125 watts. Output to the ray tube about 25 to 30 watts.

This style of ray tube instrument Model JLMSQ-101, which was built in the 1980's, was the last design that John Marsh built before his death. All of his Rife Machines and Rife information were given to his nurse before his death. His equipment and documents were obtained from her in 2012. Since we were able to obtain this instrument we have been able date it and take better photos of the complete instrument. Some of this information and new photos are shown below and are now included in this report.

John Marsh built two different models of this style of Rife Machine. In the next two photos, shown below, we see the first model of this style of instrument. The instrument was a mix of both old vacuum tube technology and modern solid state technology. The audio oscillator was a solid state variable audio oscillator with a digital readout for displaying the frequencies. The frequency range went from about 20 Hertz to 20,000 Hertz. The instrument used the same low square wave audio frequencies that were used in the 1950’s AZ-58 Rife Machine built in 1953. Those thirteen frequencies were 120, 660, 712, 727, 776, 784, 800, 803, 880, 1552, 1862, 2008, 2128. The RF carrier frequency section was built using old vacuum tube technology. The RF carrier frequency was set at 2,200,000 Hertz. This RF carrier frequency again shows that John Marsh did not understand the harmonic sideband method that Philip Hoyland used in the original Rife Ray #5 or Beam Ray Clinical Rife Machine sold by the 1938-1939 Beam Ray Corporation. His earlier 1971 instrument which was discussed in Chapter 15 of this report used a 4.150 MHz (4,150,000 Hertz) RF carrier frequency. Both of these RF carrier frequencies (2.2 MHz and 4.150 MHz) clearly show that Philip Hoyland's sideband method was not used. What became of this instrument is not known. John Marsh most likely sold it to someone because he was building these for a few people who wanted them.

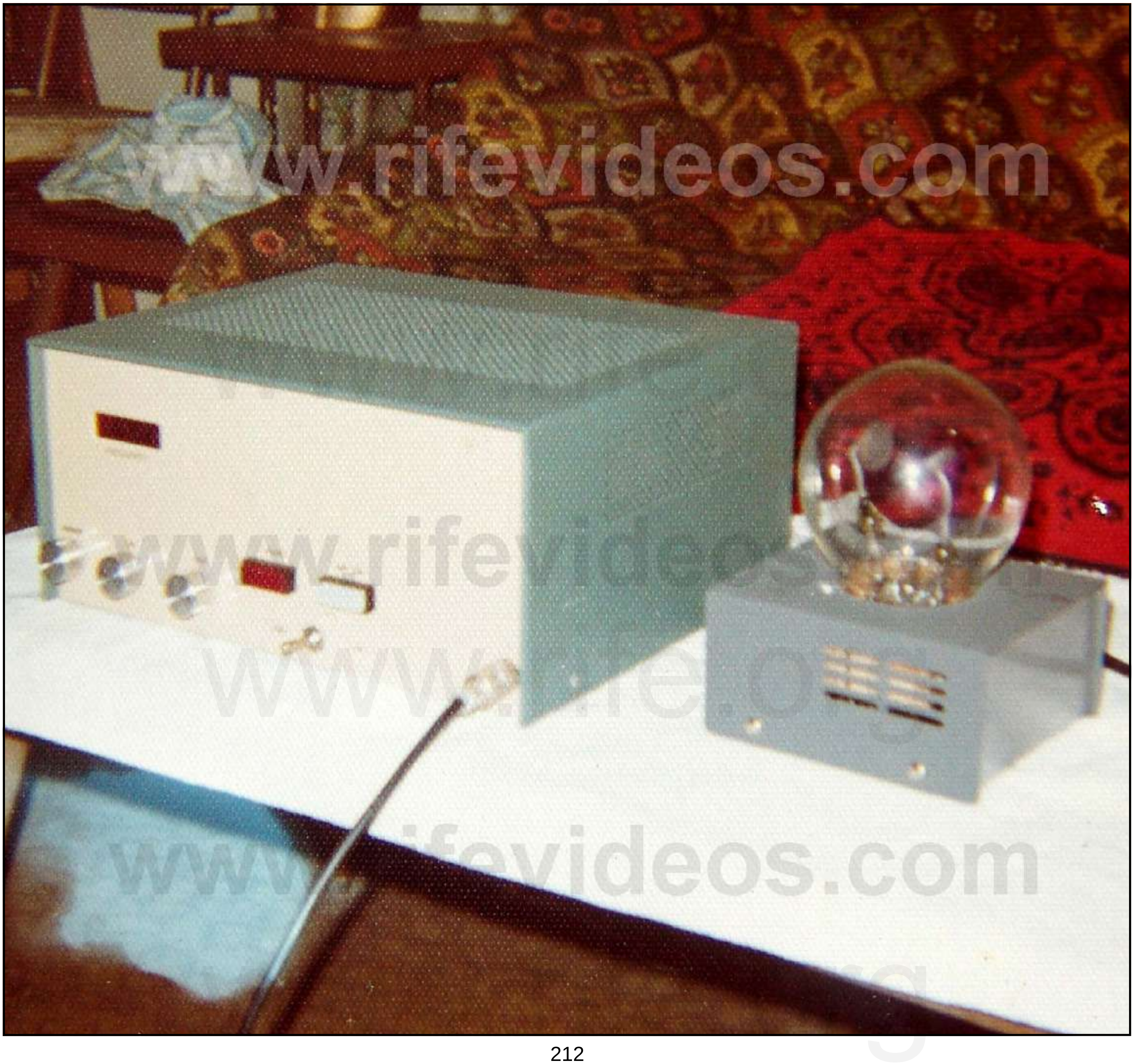

In the first photo, shown below, we see John Marsh's second style of Rife Machine. This is the same machine shown in the first photo of this chapter. This machine John Marsh used until his death in 1987. This machine was given to his nurse. We now have this machine and know how it worked. The instrument uses the same low square wave audio frequencies that were used in the 1953 AZ-58 Rife Machine. This instrument was unique because he did not have it built with a variable audio oscillator. Instead it had a dial which had 13 different positions for the 13 different audio frequencies which he used. John also had this instruments RF section built using old vacuum tube technology. The RF carrier frequency was also set to 2,200,000 Hertz (2.20 MHz). The fact that this instrument is still being used even after about thirty years speaks of its quality of construction. In the second photo, shown below, we see John Marsh with this instrument in the 1980's shortly before his death.

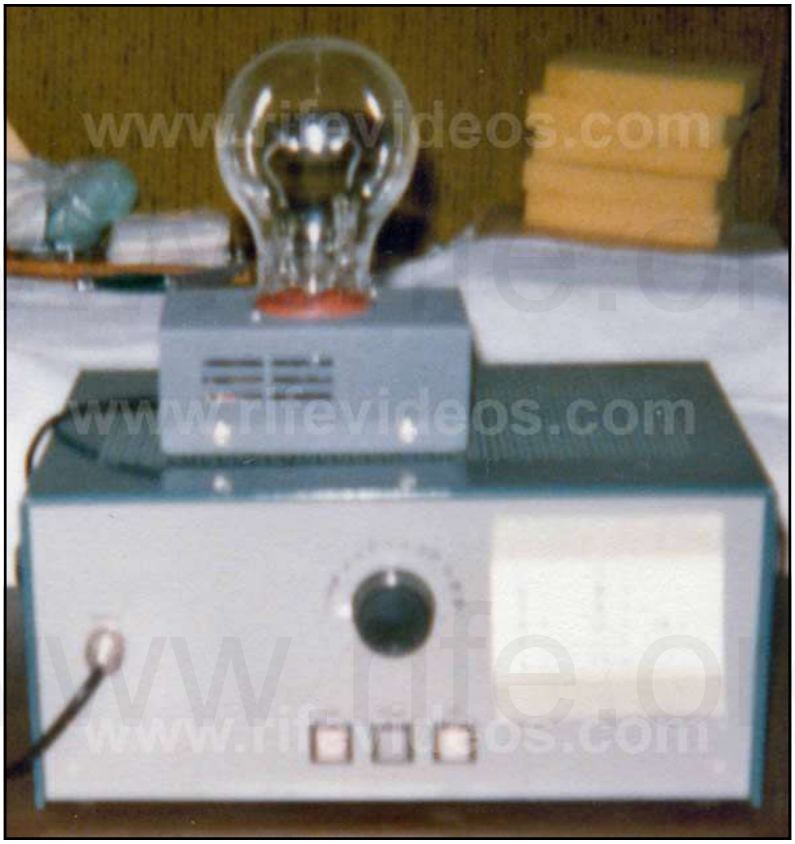

www.rife.org

The ray tube used with this instrument has an interesting design because it was designed to stand up. In the new photo, show below, you can see an up close view of this ray tube.

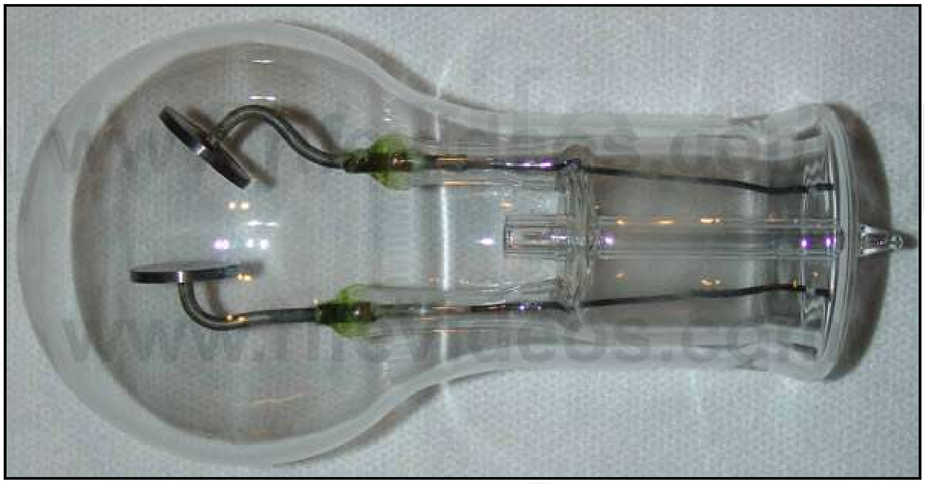

Below is a photo of the inside of the case that held the ray tube. You will notice that the ray tube is darkened. This happens to this type of ray tube because the metal electrodes are on the inside of the ray tube. When the ray tube is lit the metal comes of the electrodes and over time slowly contaminates the gas and the interior of the tube. The metal deposits or coats the inside of the ray tube and then the ray tube becomes less effective. In many cases the ray tube begins to sputter requiring it to be replaced or cleaned out and re-gassed. The photo of the new ray tube, shown above, is one of the extra ray tubes that John had built for this instrument. Many people no longer use ray tubes with internal electrodes because of this problem. The tubes without internal electrodes last for many years without any need of replacement. Some people have used them for more than 15 years and the ray tubes are still working without any problems.

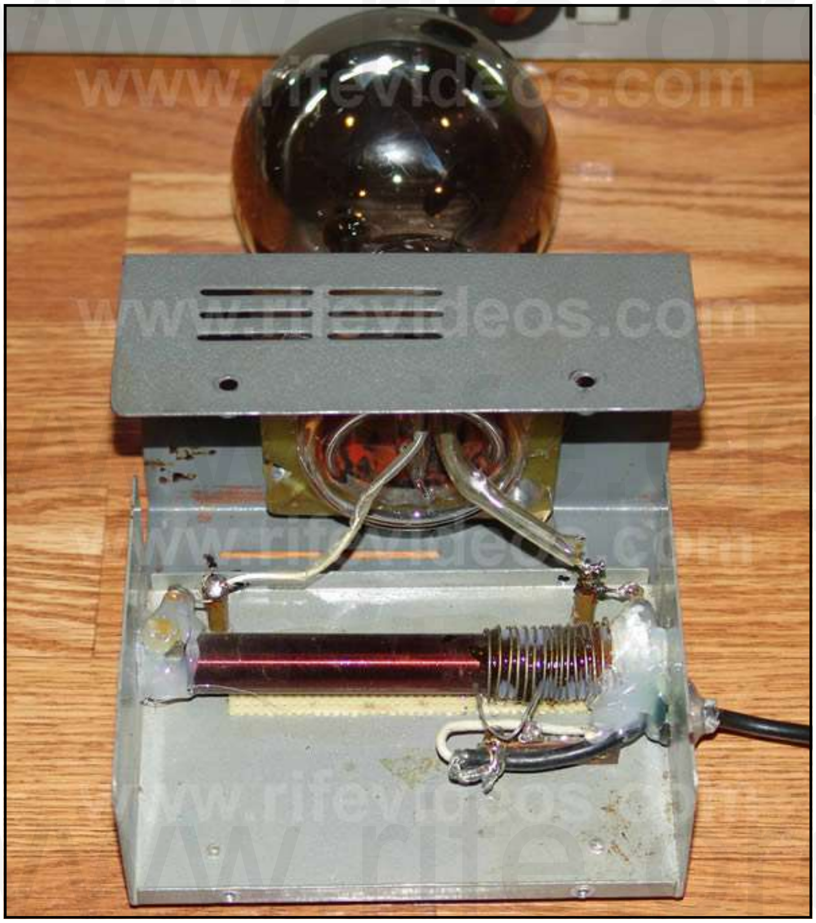

www.rife.org
www.rifevideos.com

The photo, shown below, is the underside of the instrument with the bottom removed.

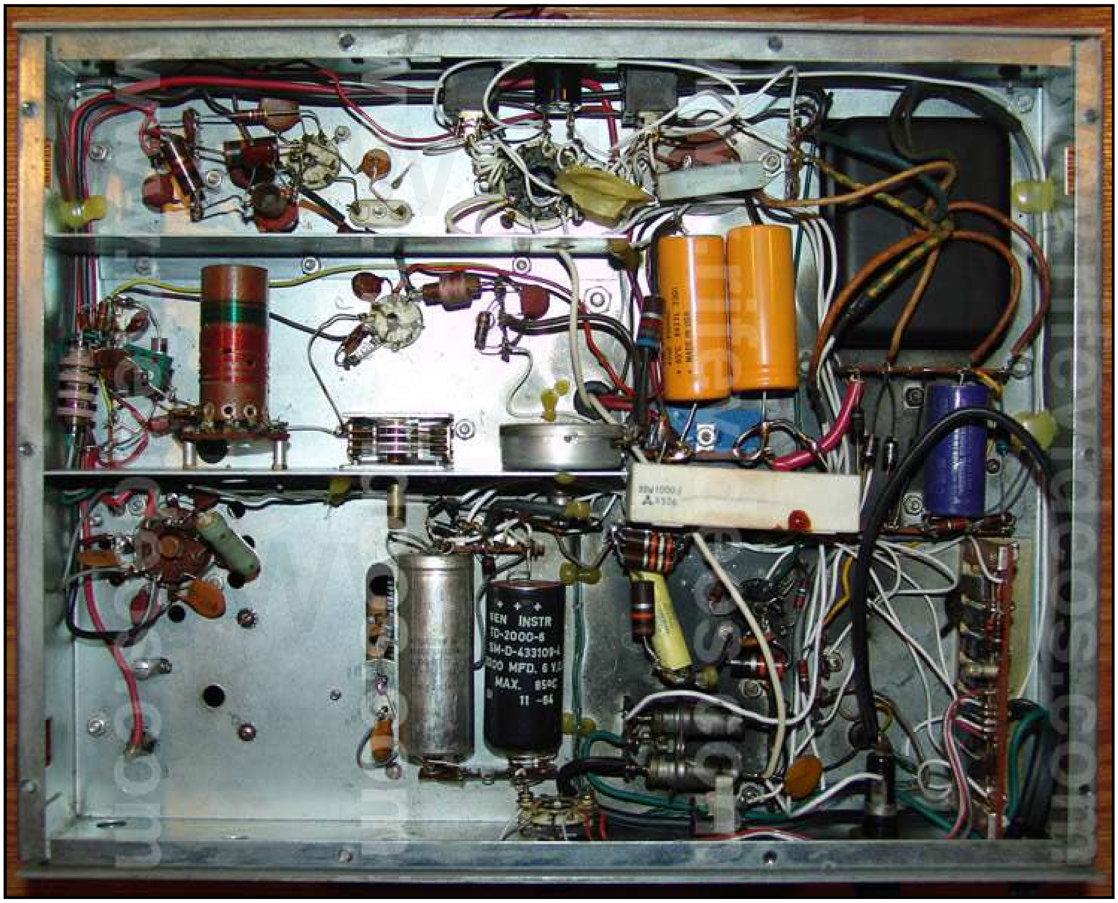

The next picture, shown below, is a photo of the instrument with the top cover removed.

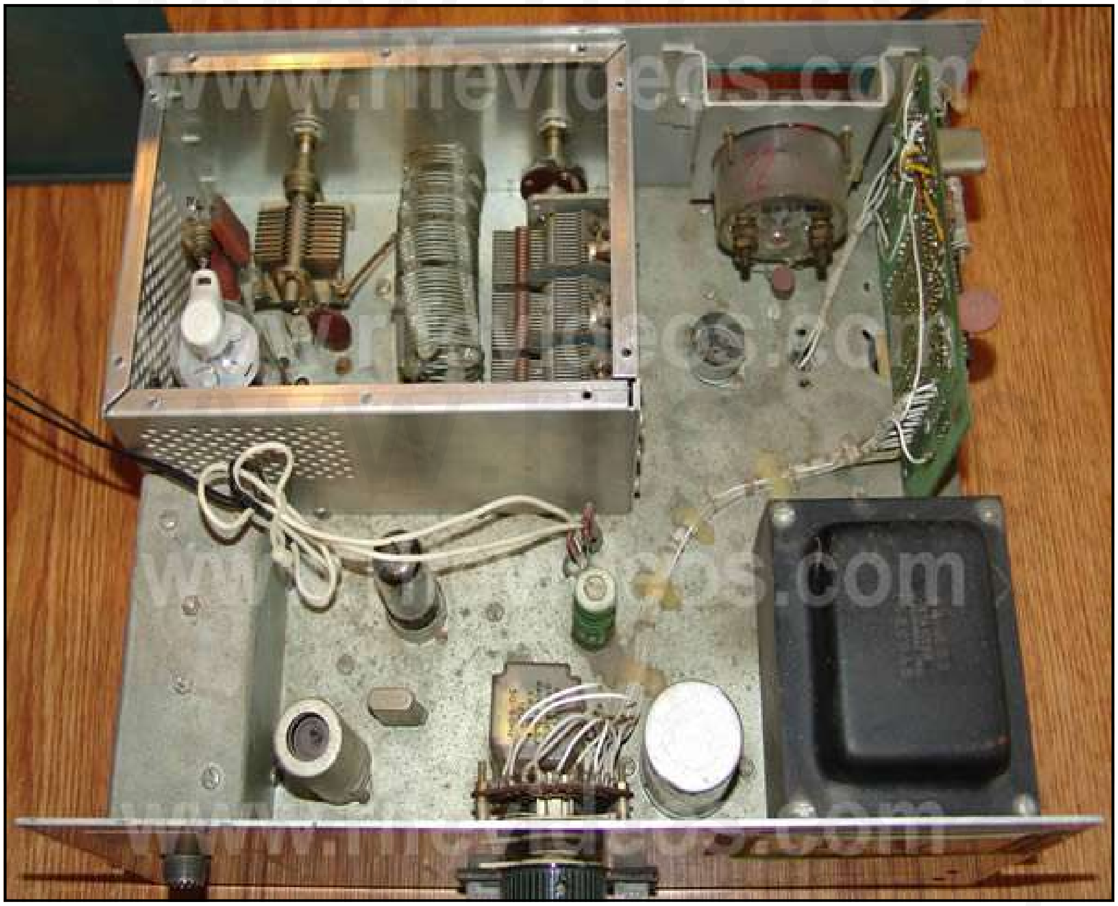

www.rife.org

The next photo, shown below, is a close up of the RF carrier frequency section showing the RF tank coil and the two variable tuning capacitors for tuning the RF carrier frequency to 2.20 MHz and for tuning the ray tube resonance.

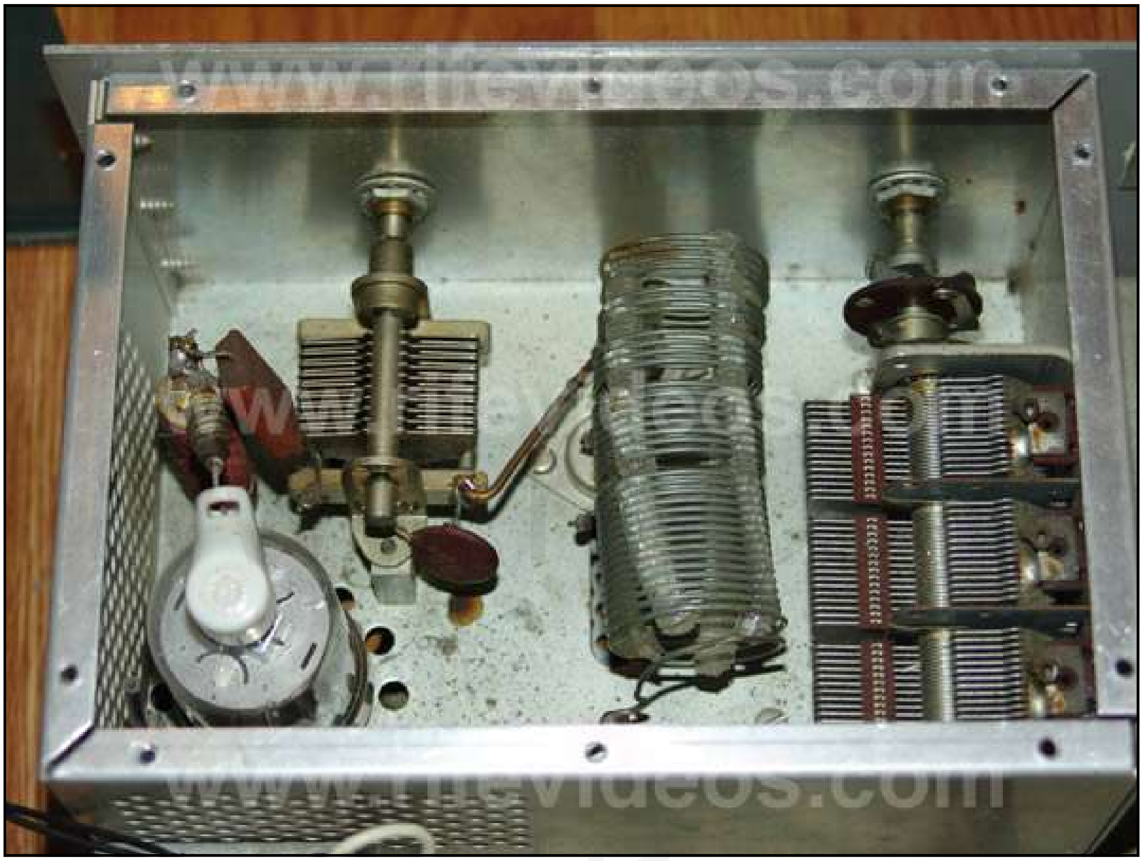

In the next photo, shown below, you can see the two brass colored nuts with regular screw driver slots. The one on the left labeled “Load” was for tuning the ray tube and the other on the right which is labeled “Tune” was for tuning the carrier frequency.

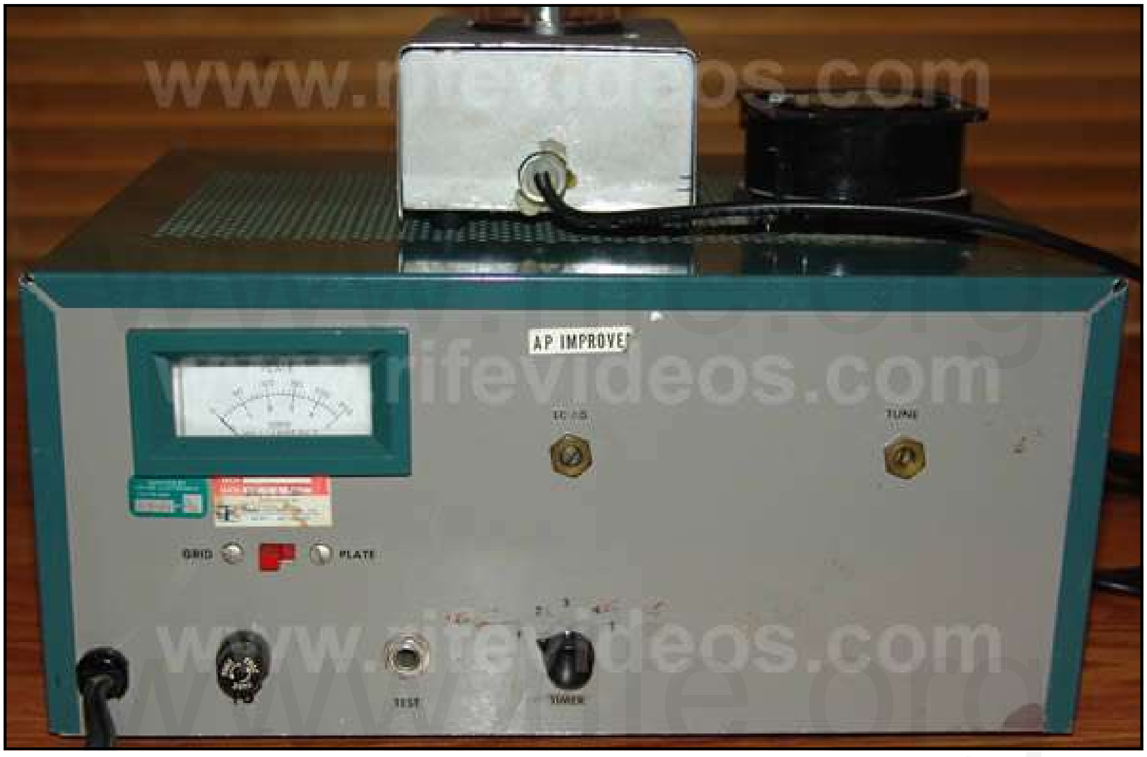

www.rife.org

The next photo, shown below, is a side view of the instrument showing the electronics for the thirteen square wave audio frequencies.

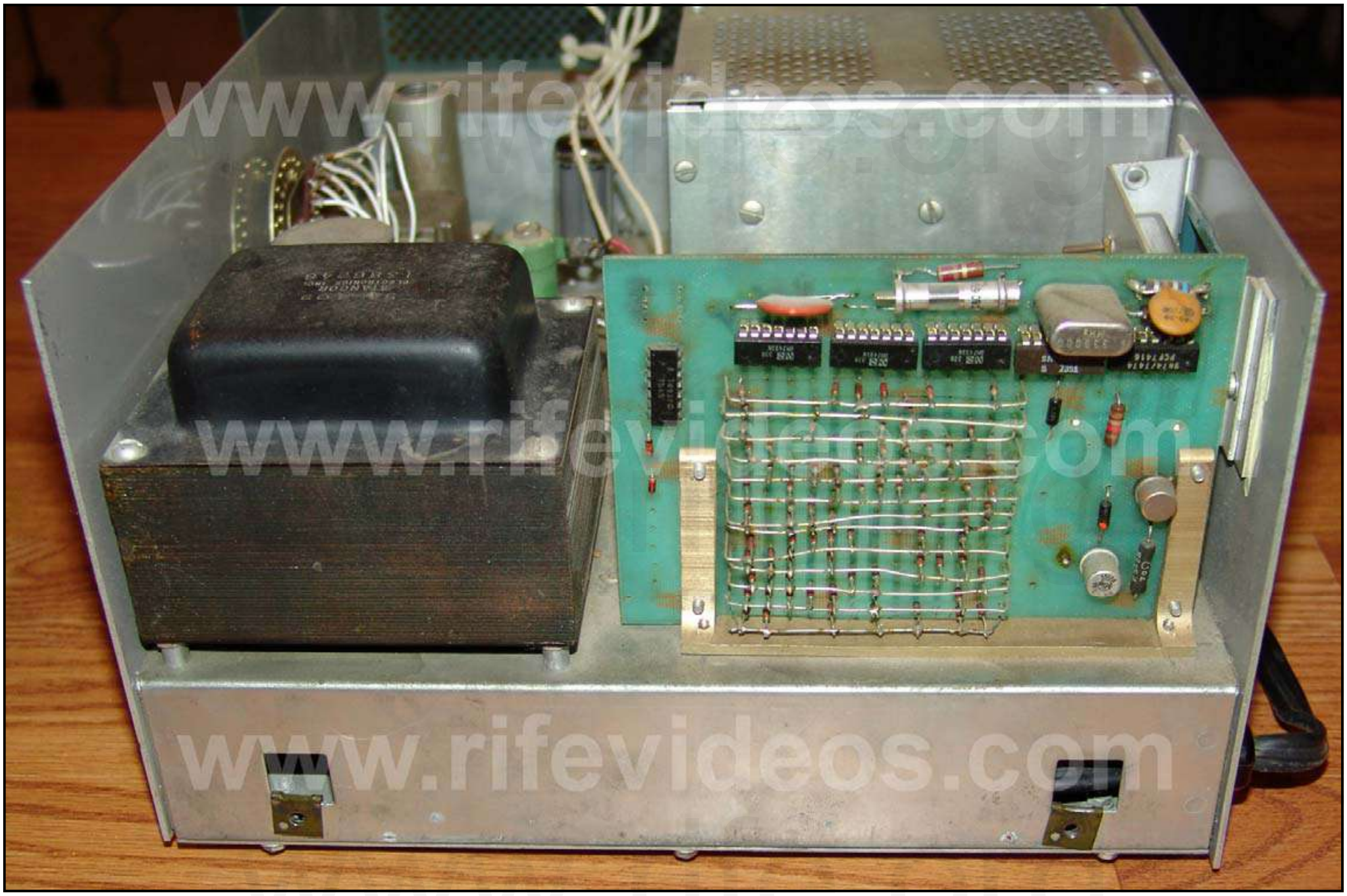

www.rifevideos.com www.rife.org

In the next photo, show below you can see the tuning dial for each of the thirteen square wave audio frequencies. The dial started with the lowest audio frequency and went to the highest. The audio frequencies went in clockwise order from the first setting to the last setting 120, 660, 712, 727, 776, 784, 800, 803, 880, 1552, 1862, 2008, 2128.

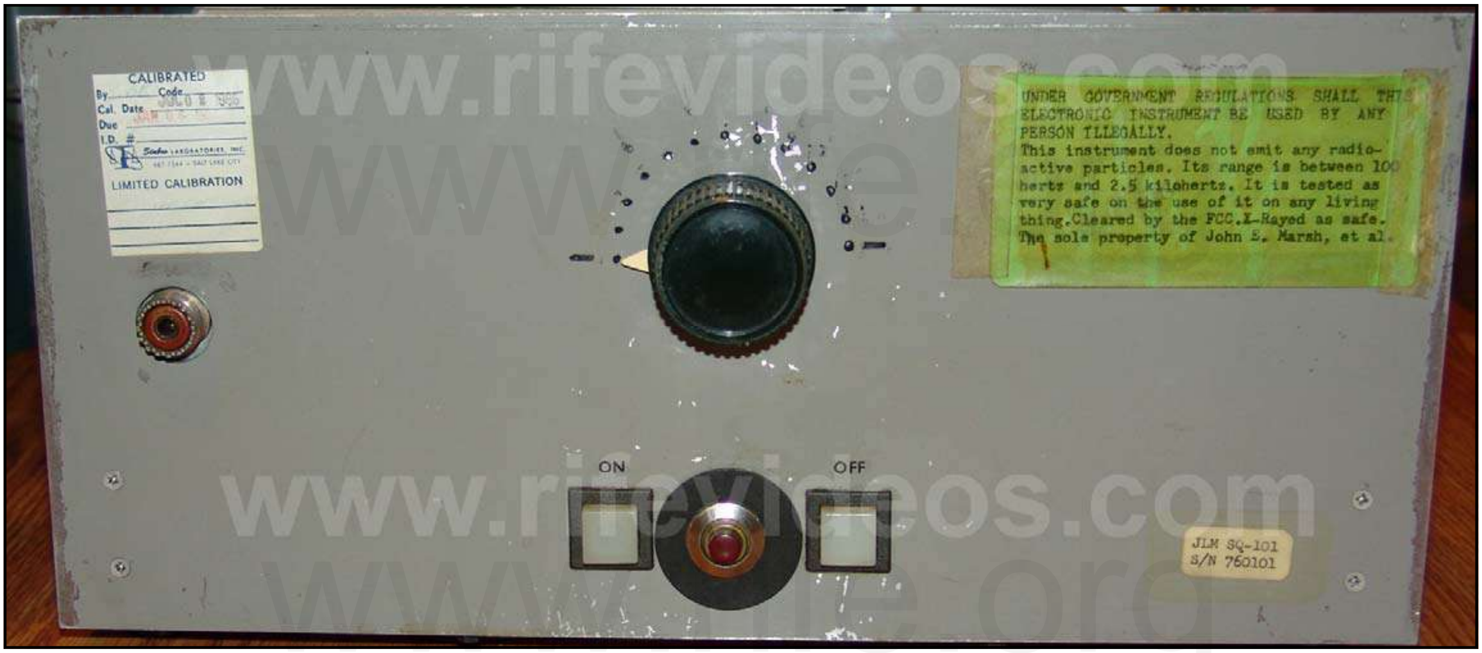

UNDER GOVERNMENT REGULATIONS SHALL THIS ELECTRONIC INSTRUMENT BE USED BY ANY PERSON ILLEGALLY.
This instrument does not emit any radio-active particles. Its range is between 100 hertz and 2.5 kilohertz. It is tested as very safe on the use of it on any living thing.Cleared by the FCC.X-Rayed as safe. The sole property of John E. Marsh, et al.

The final photo, shown below, is of the machine with the ray tube lit. The ray tube would be a lot brighter if the tube was in new condition.

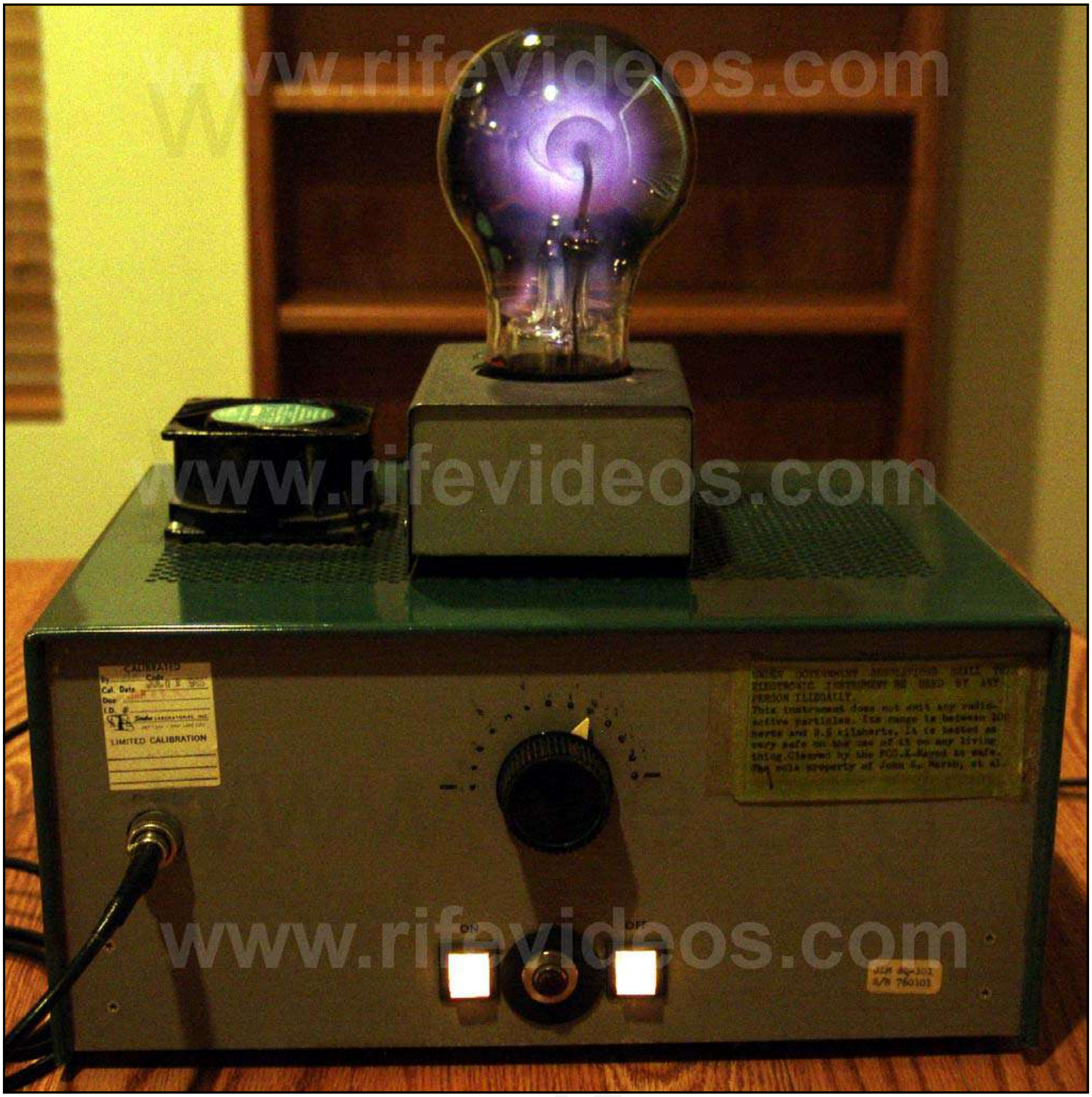

Chapter Summary: This instrument built by John Marsh used the same frequencies as the 1953 AZ-58. It also worked on the same low audio frequency square wave harmonics method as the AZ-58. The power level was lower than the AZ-58 and his 1970's AZ 58 replica discussed in chapter 15 of this report. This instrument was the last ray tube instrument John Marsh built before his death in 1987.

The next chapter will be a summary of the complete "Rife Machine Report."

www.rife.org

Chapter #17

Summery of the Rife Machine Report

In summary, with all the historical information that has come to light in the past few years we finally know the truth about which frequencies were Dr. Rife's M.O.R.s. We also understand the audio frequency sideband spacing method used in the Rife Ray #5 or Beam Ray Clinical instrument which hit the high harmonics of Dr. Rife original RF M.O.R. frequencies. Because Philip Hoyland hid how his instrument worked these audio frequencies, due to lack of knowledge, were lowered and used with a square wave waveform in the 1953 AZ-58. These lower square wave audio frequencies may not work as well as Dr. Rife's original frequencies but they have accomplished a lot of good helping many people. With the use of even more square wave audio frequencies a whole new field of frequencies are now available for our use. Having said this, we still need to remember Dr. Rife still maintained his true M.O.R. frequencies were in the RF band of frequencies. Even though Dr. Rife, John Crane and John Marsh tested these square wave audio instruments in the 1950's and early 1960's to see how well they would work. It wasn't until after John Crane and John Marsh were released from prison that they received Dr. Stafford's report showing the limited capability of the low square wave audio frequencies on cancer.

From about 1964 on, John Crane and John Marsh continued to build the audio frequency instruments even though they had the evidence from Dr. Stafford which showed the audio frequencies alone didn't work on cancer like the original high RF frequencies. Even though John Crane and John Marsh said the square wave audio frequencies were Dr. Rife's true M.O.R.s this does not change the fact that

www.rife.org

we now know the true purpose of the audio frequencies. The evidence in this report proves that Dr. Rife, John Crane and John Marsh really didn't understand how Philip Hoyland's Beam Ray Clinical instrument worked. This lack of understanding caused them to miss the truth when it was right before their eyes. We would still be in the dark had it not been for the original instruments found and the written documents that revealed Dr. Rife's high frequencies. Add to this the audio tapes which have Dr. Rife's own voice on them telling us his frequencies ranged from the audio to the broadcast bands. Dr. Rife was a pure scientist and only believed what he could prove. Had he seen Dr. Stafford's final report we feel certain he would have considered the cancer tests a failure. Dr. Rife said "he never fooled himself". It is entirely possible that Dr. Rife would have insisted they go back to his original high frequency design used in the Rife Ray #4 or put the 1953 AZ-58 RF carrier frequency back on 3.30 MHz and use Philip Hoyland's audio frequencies.

The most important information that has been obtained from the original Rife Ray #5 or Beam Ray Clinical instrument and Aubrey Scoon's Beam Ray replica instrument is the fact that the audio frequencies used in these instruments had nothing to do with the treatment of disease. To put it bluntly, all of us have been led down the primrose path because we did not understand how the Beam Rays Clinical instrument really worked. The mistakes made have major implications since the audio frequencies used with the AZ-58 (120, 660, 712, 727, 776, 784, 800, 803, 880, 1552, 1862, 2008 and 2128 Hertz) have no ability to eliminate the diseases we thought they would eliminate. In all reality these audio frequencies should be replaced with higher frequencies that are true harmonics of Dr. Rife's original high frequency M.O.R.s instead of clinging to the old dogma. Those who are really trying to do what Dr. Rife did should no longer promote the concept that these AZ-58 audio frequencies are M.O.R.s and by so doing put many people at risk. The best frequencies to use would always be the original high frequency M.O.R.s followed by lower audio frequencies that are exact lower harmonics of the high RF M.O.R. frequencies. For the most accurate list of Dr. Rife's original high RF frequencies go to the first chart on page 223. These frequencies were set by Philip Hoyland in Dr. Rife's laboratory using his microscope.

Hopefully this information will help make a change and in the future we will begin to see what Dr. Rife's original high frequency M.O.R.s will do. Many helpful people have provided the records and resources so this new information could be brought to light: the release of the John Marsh information from John Marsh's nurse; the Beam Ray Trial Papers from Steven Ross; the many photos from Jason Ringas of the Rife Research Group of Canada; the great benefit from Dr. Larry Low who allowed us to purchase the original Beam Ray Clinical instrument; the British Rife group and their work on the Aubrey Scoon replica instrument; the help of James Cunningham along with the great detective work done by James Peters in figuring out that Dr. Rife was using the Kennedy company Model 110, 220 and 281 receivers. We also want to recognize the great work Jim Peters did on the schematic of Dr. Gruners original Beam Rays Laboratory instrument. His recognition of the second variable Hartley Oscillator made it possible to rebuild an instrument that works like the original Beam Ray Laboratory instrument.

The spectrum analyzing of these machines has finally given us the answers to how all these different instruments really worked. I believe the recognition of the Kennedy equipment and the locating of the original 1938-1939 Beam Ray Clinical instrument along with the Beam Ray Laboratory instrument schematic correction and rebuilding are three of the greatest pieces of information we have yet discovered about Dr. Rife. No longer are we guessing in the dark. We have purchased the Kennedy Company equipment Models 110, 220 and 281 along with the original Beam Ray Clinical instrument for all of this testing. We plan on doing more extensive spectrum analysis work on this equipment. We have built, into one case, the Beam Ray Clinical and Laboratory instrument designs. We wish also to give special thanks to Henry Rogers the owner of the Western Historic Radio Museum (www.radioblvd.com). He allowed us the opportunity to come and test the Kennedy receivers that he owns. As more information comes out we will update this article as necessary.

www.rife.org

For those who would like a complete list of Dr. Rife’s frequencies output by the Rife Ray #3, Rife Ray #4, Philip Hoyland's Rife Ray #5 or Beam Ray Clinical instrument, Aubrey Scoon’s 1950's Rife Ray #5 and the AZ-58 Beam Ray replica instrument we have listed them in a chart on page 222. Other charts that may be of interest are included on pages 223 and 224.

None of the sets of the low audio frequencies are true M.O.R.s and were originally used to create the proper sideband spacing frequencies. The square wave audio frequencies used by the AZ-58 were used in a different manner or method relying only upon the harmonics from the square wave waveform. This method has been used with very good results over the past 50 years by many people, but, these audio frequencies have never produced the true M.O.R. effect of devitalizing organisms. The correct high RF M.O.R.s are the frequencies that should be used since we know what these frequencies are. If people are determined to use the lower audio and ultrasonic range of frequencies below 50,000 Hertz then they should at least use the highest harmonic frequency of the true M.O.R.s. At the very least we should make sure that all frequencies used are true harmonics of the fundamental M.O.R.s that Dr. Rife found.

www.rifevideos.com
Copyright © 2003 & 2013
All rights reserved.

www.rife.org

Rife's M.O.R Frequencies And Audio Sideband Frequencies

Microorganisms From Rife's Lab Notes And Other Documents.	Square Wave Frequencies for Rife, Crane, Marsh 1950s AZ-58 Beam Ray Replica. Built by Verne Thompson.	Sideband Frequencies For Aubrey Scoon's 1940s Beam Ray Replica. Built by Verne Thompson. Used with 3.30 MHz	Sideband Frequencies For Original Beam Ray Clinical Instrument. Built By Philip Hoyland 1938-1939. Recovered in 2008. Used with 3.80 MHz	High RF Frequencies For Dr. Rife's Rife Ray #3 and #4 Instruments. Correctly read in 1935 By Philip Hoyland*. Rife Ray # 4 built by*.	1/10th Of One Meter Frequency Tolerance.	Rife's Misread Lab Note Frequencies Before 1935. Frequency #1	Rife's Misread Lab Note Meter Frequencies Before 1935. Meters	Rife's Misread Lab Note Meters Converted To Hertz. Frequency #2
	Square Wave	Sine Wave	Sine Wave	Sine Wave		Sine Wave	Sine Wave	Sine Wave
Actinomycosis or Streptothrix	784 Hz	7,870 Hz	10,000 Hz	192,000 Hz	12 Hz	678,000 Hz	1,607 M	186,554 Hz
Anthrax			8,320 Hz	139,200 Hz	6 Hz	900,000 Hz	1,100 M	272,539 Hz
Anthrax Symptomatic						400,000 Hz	18,000 M	16,655 Hz
B or E Coli Rod	800 Hz	8,020 Hz	7,833 Hz	417,000 Hz	58 Hz	683,000 Hz	943 M	317,914 Hz
B or E Coli Virus	1,552 Hz	17,220 Hz	16,667 Hz	770,000 Hz	198 Hz	8,581,000 Hz	27 M	11,103,424 Hz
BX Virus Carcinoma	2,128 Hz	21,275 Hz	39,467 Hz	1,604,000 Hz	858 Hz	11,780,000 Hz	17.6 M	17,033,662 Hz
BY Sarcoma	2,008 Hz	20,080 Hz	37,000 Hz	? 1,530,000 Hz	780 Hz			
Bubonic Plague						160,000 Hz	585 M	512,466 Hz
Catarrh						1,800,000 Hz	175 M	1,713,100 Hz
Cholera Spirillum						851,000 Hz	312 M	960,873 Hz
Contagious Conjunctivitis						1,206,000 Hz	148 M	2,025,625 Hz
Diptheria						800,000 Hz	275 M	1,090,154 Hz
Glanders						986,000 Hz	407 M	736,591 Hz
Gonorrhea	712 Hz		14,400 Hz	233,000 Hz	18 Hz	600,000 Hz	1,990 M	150,649 Hz
Influenza						1,674,000 Hz	154 M	1,946,704 Hz
Leprosy						743,000 Hz	1,190 M	251,926 Hz
Pneumonia						1,200,000 Hz	785 M	381,901 Hz
Pneumonia or Spinal Meningitis	776 Hz	7,660 Hz	8,600 Hz	427,000 Hz	61 Hz	927,740 Hz	167 M	1,795,164 Hz
Staphylococcus Pyogenes Aureus	727 Hz	7,270 Hz	8,000 Hz	478,000 Hz	76 Hz	988,740 Hz	540 M	555,171 Hz
Staphylococcus Pyogenes Albus							546 M	549,070 Hz
Streptococcus Pyogenes	880 Hz	8,450 Hz	8,333 Hz	720,000 Hz	173 Hz	1,214,000 Hz	142 M	2.111.214 Hz
Syphilis (Treponema Pallidum)	660 Hz	6,600 Hz	6,591 Hz	789,000 Hz	207 Hz	900,000 Hz	108 M	2,775,856 Hz
Tetanus	120 Hz	1,200 Hz	11,200 Hz	234,000 Hz	18 Hz	700,000 Hz	19,000 M	15,779 Hz
Tuberculosis Rod	803 Hz	8,300 Hz	8,462 Hz	369,000 Hz	45 Hz	583,000 Hz	554 M	541,142 Hz
Tuberculosis Virus	1,552 Hz	16,000 Hz	16,910 Hz	? 769,000 Hz	197 Hz			
Typhoid Rod	712 Hz	6,900 Hz	100 Hz	760,000 Hz	192 Hz	900,000 Hz	345 M	868,964 Hz
Typhoid Virus	1,862 Hz	18,620 Hz	38,214 Hz	1,445,000 Hz	694 Hz	9,680,000 Hz	21.5 M	13,943,835 Hz
Worms		2,400 Hz						

www.rife.org

Dr. Rife’s Original High RF Frequencies Fine Tuned To The Precise Frequencies By Philip Hoyland.

Actinomycosis (Streptothrix)	191,803 Hz
Anthrax	139,200 Hz
B. Coli (Rod form)	416,510 Hz
B. Coli (Filterable virus)	769,035 Hz
Bacillus X or BX (Cancer Carcinoma)	1,607,450 Hz
Bacillus Y or BY (Cancer Sarcoma)	1,529,520 Hz
Gonorrhea	233,000 Hz
Spinal Meningitis	426,862 Hz
Staphylococcus Pyogenes Aureus	477,660 Hz
Staphylococcus Pyogenes Albus	549,070 Hz
Streptococcus Pyogenes	719,150 Hz
Syphilis	788,700 Hz
Tetanus	234,000 Hz
Tuberculosis (Rod)	369,433 Hz
Tuberculosis (Virus)	769,000 Hz
Typhoid Fever (Rod)	759,450 Hz
Typhoid Fever (Virus)	1,445,180 Hz

Philip Hoyland’s Audio Frequencies Used With 3.30 MHz To Produce Through Sidebands Dr. Rife’s Frequencies

B or E Coli Rod	8,020 Hz	Syphilis or Treponema	6,600 Hz
B or E Coli Virus	17,220 Hz	Tetanus	1,200 Hz
BX Virus Carcinoma	21,275 Hz	Tuberculosis Rod	8,300 Hz
BY Sarcoma	20,080 Hz	Tuberculosis Virus	16,000 Hz
Pneumonia or Spinal Meningitis	7,660 Hz	Typhoid Rod	6,900 Hz
Staphylococcus	7,270 Hz	Typhoid Virus	18,620 Hz
Streptococcus	8,450 Hz	Worms	2,400 Hz
Streptothrix	7,870 Hz		

www.rife.org

Philip Hoyland's New M.O.R.s.
Used In The Beam Ray Clinical Instrument.

B or E Coli Rod	3,332,080 Hz	Streptothrix	3,260,650 Hz
B or E Coli Virus	3,076,140 Hz	Syphilis or Treponema	3,154,800 Hz
BX Virus Carcinoma	3,214,900 Hz	Tetanus	3,276,000 Hz
BY Sarcoma	3,059,040 Hz	Tuberculosis Rod	3,324,897 Hz
Gonorrhea	3,262,000 Hz	Tuberculosis Virus	3,076,000 Hz
Pneumonia or Spinal Meningitis	3,414,900 Hz	Typhoid Rod	3,037,800 Hz
Staphylococcus	3,343,620 Hz	Typhoid Virus	2,890,360 Hz
Streptococcus	3,595,750 Hz		

www.rife.org

NEW RESEARCH FINDINGS show that all diseases have simple explanations and cures once their true cause is known. This book describes the causes of both common and extraordinary diseases and gives specific instructions for their cure.

The Cure for all Diseases

With Many Case Histories

of diabetes, high blood pressure, seizures, chronic fatigue syndrome, migraines, Alzheimer's, Parkinson's, multiple sclerosis, and others showing that all of these can be simply investigated and cured.

Hulda Regehr Clark, Ph.D.,N.D.

Electricity can now be used to kill bacteria, viruses and parasites in minutes, not days or weeks as antibiotics require.

If you have been suffering from a chronic infection or have cancer, or AIDS, learn to build the electronic device that will stop it immediately. It is safe and without side effects and does not interfere with any treatment you are now on.

The Cure For All Diseases

© Copyright 1995 by Hulda Regehr Clark, Ph.D., N.D.
All rights reserved.

Permission is hereby granted to make copies of any part of this document for non-commercial purposes provided this page with the original copyright notice is included.

Published in the United States by New Century Press
1055 Bay Blvd., Suite C, Chula Vista, CA 91911
(619) 476-7400, (800) 519-2465
www.newcenturypress.com
ISBN 1-890035-01-7

Other books by Dr. Clark available from New Century Press:
The Cure For All Cancers
Heilverfahren Aller Krebsarten (Cure For Cancers, German)
The Cure for All Cancers (Japanese Translation)
The Cure For All Cancers (Korean Translation)
Heilung ist möglich (Cure For All Diseases, German)
The Cure For HIV And AIDS
The Cure For All Advanced Cancers

30 29 28 27 26 25 24 23 22 21

Notice to the Reader:

The opinions and conclusions expressed in this book are mine, and unless expressed otherwise, mine alone. The opinions expressed herein are based on my scientific research and on specific case studies involving my patients. Be advised that every person is unique and may respond differently to the treatments described in this book. On occasion we have provided dosage recommendations where appropriate. Again, remember that we are all different and any new treatment should be applied in a cautious, common sense fashion.

The treatments outlined herein are not intended to be a replacement or substitute for other forms of conventional medical treatment. Please feel free to consult with your physician or other health care provider.

I have indicated throughout this book the existence of pollutants in food and other products. These pollutants were identified using a testing device of my invention known as the Syncrometer.™ Complete instructions for building and using this device are contained in this book. Therefore anyone can repeat the tests described and verify the data.

The Syncrometer is more accurate and versatile than the best existing testing methods. A method for determining the degree of precision is also presented. However at this point it only yields positive or negative results, it does not quantify. The chance of a false positive or a false negative is about 5%, which can be lessened by test repetition.

It is in the public interest to know when a single bottle of a single product tests positive to a serious pollutant. If one does, the safest course is to avoid all bottles of that product entirely, which is what I repeatedly advise. These recommendations should be interpreted as an intent to warn and protect the public, not to provide a statistically significant analysis. It is my fervent hope that manufacturers use the new electronic techniques in this book to make purer products than they ever have before.

Dedication

I would like to dedicate this book to all the persons who visited me professionally, from the very first person in 1963, **Mrs. R. Biehl**, to the present. I learned so much from each of you and I appreciate your confidence, your intelligence and your reluctance to be defeated.

Acknowledgments

I would like to express my gratitude to my son, **Geoffrey**, who always listened to my “crazy ideas” on Sundays, right at supper time. He was patient, kind, helpful and willing to share his expertise in electronics and computers. Without him, this book could not have been written.

A big thank you goes to **Frank Jerome, DDS**. Without the loan of his parasite slide collection many of my discoveries could not have been make, and without his development of metal-free dentistry, many of these patient histories could not have ended happily. Thank you to **Linda Jerome** for nurturing us both with personal interest and patience. And thank you to **Edna Bernstein** for her resourceful assistance.

Contents

FIGURES .... **XIII**

PREFACE .... **XVII**

THE PROMISE **1**
  Only Two Health Problems 2
  Be A Health Detective 3

THE DISCOVERY **5**
  Everything Has A Unique Frequency 8
  Selective Electrocution 9
  Bioradiation 12
  Zapping Bugs 13
  The Bioradiation Spectrum 16

BUILDING A ZAPPER **19**
  Killing The Surviving Pathogens 20
    Zapper Parts List 21
    Assembling The Zapper 23
    Using The Zapper 28
  Simple Pulser 28
    Using The Simple Pulser 30

PARASITES & POLLUTION **31**
  The Worst Parasite 33
  Pollution 36
  Solvent Pollution 36
  Metal Pollution 37
  Mycotoxins 38
  Physical Toxins 38
  Chemical Toxins 40

HOW WE REALLY GET SICK...................................... 43
Self Health 46
How To Heal 47
The Road To Wellness.. 49

PAIN FROM TOE TO HEAD.. 51
Searching For Bacteria 53
What You Will Find.. 54
Toe Pain.. 55
Balance Your pH 57
Foot Pain... 59
Heel Pain 61
Leg Pain In Children 69
Leg Pain In Adults.. 70
Leg Pain Protocol.. 73
Fibromyositis and Fibromyalgia 75
Joint Pain or Arthritis.. 78
Osteo or Common Arthritis.. 79
Rheumatoid Arthritis... 82
Allergic Arthritis... 84
Thigh Pain... 86
Hip Pain.. 87
Bone Strengthening.. 87
Groin Pain... 89
Low Back Pain.. 92
Lower Abdominal Pain... 97
Stomach Pain... 98
Appendicitis.. 100
Urinary Tract Pain.. 101
Irritable Bowel Syndrome(IBS), Colitis, and Spastic Colon 102
Crohn's Disease 103
Central Abdominal Pain.. 108
Interstitial Cystitis... 108
Uterine Pain 109
Endometriosis 109

- Contraception 112
- The Silent Cervix 112
- Menopausal Symptoms, Hot Flushes, PMS 113

Infertility 117

- Nausea 121

Prostate Problems 124

Side Pain 128

Midabdomen Pain, Stomach Pain 131

Stomach Ulcer 132

Hiatal Hernia 133

Respiratory Illness 134

- Asthma 134
- Bronchitis, Croup, Chronic Cough 140

Breast Pain 142

Breast Sensitivity 143

- Breast Lumps 144

Heart Pain 146

- Slow Pulse/Syncope (Passing Out) 149

Chest Pain 149

Upper Back Pain 150

Shoulder Pain 151

Upper Arm Pain 153

Elbow Pain 154

Wrist Pain 154

Thumb Pain and Hand Pain 155

Finger Pain 155

Back Of Neck Pain 156

Front Neck Pain 156

Temporal Mandibular Joint (TMJ) Problems 157

Tooth Ache 158

Throat Pain 159

Eye Pain 160

- Macular Degeneration 162

Headache and Migraines 162

Earache 168

Ear Noises And Ringing 169

Scalp Pain ... 171

NON-PAINFUL DISEASES ... 173

Diabetes ... 173
  Prediabetic ... 178
  Diabetes Of Childhood ... 178
Herpes ... 179
Fatigue ... 182
Skin Problems ... 188
  Sebaceous Cysts ... 188
  Skin Rash ... 189
  Hives ... 191
  Acne ... 192
  Psoriasis, Eczema ... 192
Warts ... 195
Tapeworm Stages ... 197
The Flu ... 201
Fever ... 203
Multiple Sclerosis & Amyotropic Lateral Sclerosis 204
High Blood Pressure ... 210
Glaucoma ... 213
Tooth Decay ... 214
Muscle Diseases ... 218
  Muscular Dystrophy ... 219
  Myasthenia Gravis ... 221
Universal Allergies ... 222
  Environmental Illness ... 223
Alcoholism ... 227
Seizures ... 230
Weight Problems ... 237
Sleep Problems ... 241
  Sleep Apnea ... 241
  Insomnia ... 243
Yeast Infections ... 246
  Diaper Rash ... 246
  Treat Yeast or Fungus the Same ... 247
Fluke Disease ... 249

Flukes and Solvents.. 251
Flukes Not Alone 252
Burning And Numbness.. 253
Depression.. 256
Manic Depression... 257
Schizophrenia.. 261
Autism... 263
Digestion Problems.. 266
Alzheimer's... 269
Dementias, Memory Loss... 272

PUSHING BACK AGE **275**
Diet.. 275
Increase Oxygen... 278
Niacin.. 279
TIA’s, Strokes and Purpura.. 281
Blood Pressure.. 281
Air Pollution... 282
Anemia.. 285
Acid Levels... 287
Increase Minerals.. 287
Chelation.. 289
Pulse... 289
Heart Health.. 290
Incontinence.. 293
Bad Digestion.. 295
Chewing... 295
Stomach Acid... 296
Salmonella and Shigella... 297
Other Clues.. 299
Liver Bile... 299
Constipation.. 301
Poop Your Troubles Away 301
Sugar Regulation.. 302
Elderly Person’s 7 Day Diabetic Diet.............................. 302
Breakfast.. 302
Lunch.. 303
Dinner 304

Supper 307
Diabetic Supplements 309
Diabetic Eating Out 309
Extra Diet Tips For The Elderly 309
Salt 310
Tremor 312
Tremor Remedies 313
Weakness 317
Heart Disease 318
Heart/Kidney Relationship 320
Diuretics 321
Feeling Comfortable 322
Hearing Loss 324
Insomnia 326
Healthful Habits 326
Not Old Yet 328
Super Longevity 329
Dying 330

CURING CANCER 331
Another Fluke Disease 331
Teamed With A Solvent 332
Propyl Alcohol Polluted Products 335
Getting Well After Cancer 336
Herbal Parasite Killing Program 338
Parasite Program Handy Chart 341
Maintenance Parasite Program 341
Pet Parasite Program 343
Parasites Gone, Toxins Next 345
Watch For Bacteria 346
Help Your Family, Too 347
Aflatoxin 348

CURING HIV/AIDS 351
Method Of Treatment 353
Benzene Polluted Products 354

Plan For The Future 356

CURING THE COMMON COLD 357
Tapeworm Stage or Mites.. 357
Molds and Colds... 360
The Cure 363
Mold Free Diet 365
Prevention.. 368
Curing Symptoms 369
True Origins Of Viruses... 370

BODY WISDOM AND WHY BAD FOOD TASTES GOOD373
Salt... 373
Fats 374
Starch... 374
Sugar.. 375
Dislikes... 375
Outwit The Cravings 377
Back To Normal.. 379
Toxic Food.. 379
Moldy Food 381
  Aflatoxin...382
  Eradicating Aflatoxin...383
  Zearalenone 384
  Sterigmatocystin..385
  Ergot...385
  Cytochalasin B 387
  Kojic Acid 388
  T-2 Toxin 388
  Sorghum Molds...388
  Patulin...388
  Mold Avoidance 389
  Getting Away From Grains...392
Pets Teach Health Lessons....................................... 394

EASY LIFESTYLE IMPROVEMENTS 397
Living Hand To Mouth.. 397

Better Laundry Habits 398
Better Kitchen Habits... 399
Better Housekeeping 402
Windows Open Or Closed? 405
What Kind Of Heat.. 407
Getting Rid of Mites.. 407

FOUR CLEAN-UPS 409
Dental Clean-up 409
Guidelines For A Healthy Mouth 411
Horrors Of Metal Dentistry...416
Dental Rewards 419
Diet Clean-up 420
Breakfast...420
Lunch..422
Supper 423
Beverages...424
Horrors In Commercial Beverages 428
Food Preparation...430
Food Guidelines..431
Dining Out 433
Body Clean-up 435
Stop Using Supplements...439
Safe Supplements...441
Home Clean-up 444
Your Basement...444
Your Garage...445
Special Clean-up for Freon (CFCs).......................................446
Special Clean-up for Fiberglass...447
Special Clean-up for Asbestos...447
Your House...448
Unnatural Chemicals 451

BIOELECTRONICS... 457
Making A Syncrometer.. 457
Making a Hobby Kit Syncrometer..458
Making Test Plates...460
Test Plates Parts List...460

Test Plates Assembly 461
Using The Syncrometer 462
Resonance 465
Lesson One 466
White Blood Cells 468
Making a White Blood Cell Specimen 469
Lesson Two 470
Lesson Three 470
Lesson Four 471
Preparing Test Substances 472
Impure Test Substances 472
Making Organ Specimens 477
Making a Complete Set of Tissue Samples 478
Purchasing a Complete Set of Tissue Samples 478
Body Fluid Specimens 479
Lesson Five 481
Lesson Six 482
Lesson Seven 484
Testing the Air 485
Testing Someone Else 486
Lesson Eight 486
Saliva Testing 487
Lesson Nine 487
Surrogate Testing 488
Lesson Ten 489
Lesson Eleven 491
Lesson Twelve 492
Lesson Thirteen 493
Lesson Fourteen 493
Lesson Fifteen 494
Lesson Sixteen 494
Lesson Seventeen 495
Sensitivity of Pollutant-In-Product Testing 497
Microscopy Lesson 498
Taking Pictures Of What You See 501
Using A Frequency Generator 502
The Theory 502
Lesson Eighteen 504
Lesson Nineteen 505
Lesson Twenty 506

Lesson Twenty One....507
Lesson Twenty Two....507
Lesson Twenty Three....508
Lesson Twenty Four....508
Lesson Twenty Five....509
Lesson Twenty Six....509
Lesson Twenty Seven....509
Lesson Twenty Eight....510
Lesson Twenty Nine....510

RECIPES....513
Beverage Recipes....513
Lemonade....513
Fresh Tomato Juice....513
Fresh Pineapple Juice....514
Maple Milk Shake....514
Yankee Drink....514
Hot Vanilla Milk....514
Red Milk....515
C-Milk....515
My Own Soda Pop....515
My Own Super C-Pop....516
Half And Half....516
Buttermilk-C....516
Food Recipes....516
Daily Foods....517
Two Granolas....518
Peanut Butter....519
Sweetening and Flavoring....519
Preserves....520
C Dressing....521
Cheese Sauce....521
Sour Cream-C....521
Yogurt....521
Soups....521
Fish and Seafood recipe....522
Baked Apples....522
5 Minute Ice Cream....522
Seven Day Sample Menu....523
Recipes for Natural Body Products....525

Borax Liquid Soap....526
For Laundry....526
For Dishes....527
In The Dishwasher....527
In The Sink....527
Shampoo....528
Hair Spray....529
Homemade Soap....529
Liquid Soap....530
Skin Sanitizer....530
Deodorant....531
Brushing Teeth....532
For Dentures....533
Mouthwash....533
Contact Lens Solution....533
Lip Soother....534
Foot Powder....534
Skin Healer Moisturizer Lotion....534
Other Skin Healers....535
Massage Oil....536
Sunscreen Lotion....536
Nose Salve....536
Quick Corn Starch Skin Softener....537
Cornstarch Skin Softener....537
After Shaves....537
Personal Lubricants....537
Baby Wipes....538
People Wipes....538
Recipes For Natural Cosmetics....539
Eye liner and Eyebrow Pencil....539
Lipstick....539
Face Powder....540
Blush (face powder in a cake form)....540
Recipes For Household Products....540
Floor Cleaner....540
Furniture Duster and Window Cleaner....540
Furniture Polish....541
Insect Killer....541
Ant Repellent....541
Flower and Foliage Spray....542

Moth Balls 542
Carpet Cleaner 542
Health Improvement Recipes 543
Black Walnut Hull Tincture 543
Quassia recipe 546
Emmenagogue (Menstrual Period Inducer) 546
Bowel Program 546
Constipation Tea 548
Weight Reduction 548
Kidney Cleanse 549
Liver Herbs 552
Liver Cleanse 552
Lugol’s Iodine Solution 559
Vitamin D Drops 560

PATHOGEN FREQUENCIES 561
Bandwidth Of Organism Families 562
Mold, Mold Toxin Frequencies 562
Bacteria and Viruses 563
Roundworms, Flatworms, One-celled Animals 566
Wart Frequencies 569
Tapeworms 569
Mite Frequencies 570
Miscellaneous Frequencies 571
Toxic Elements 571
Solvents 574
Pathogen Frequency Chart 576

SOURCES 581

FINALE 589

INDEX 591

Figures

Fig. 1 Selected animal bandwidths. 17
Fig. 2 Selected pathogen bandwidths. 18
Fig. 3 Zapper schematic. 22
Fig. 4 Finished zapper, outside and inside. 27
Fig. 5 Simple pulser 29
Fig. 6 Ascaris. 32
Fig. 7 (L to R) Human intestinal fluke, pancreatic fluke, sheep liver fluke, and human liver fluke. 33
Fig. 8 Fasciolopsis' normal life cycle 35
Fig. 9 Feed animals out of your yard and away from where children play. Their wastes contain pathogens. 50
Fig. 10 Your kidneys with the adrenal glands sitting like hats on top. The ureters lead to the bladder 60
Fig. 11 All disposable diapers I tested had mercury and thallium! These toxins can be absorbed through babies' skin. 69
Fig. 12 Do not use any commercial personal products, the risk of pollution is too great. 71
Fig. 13 Safe substitutes for personal products 72
Fig. 14 Tooth numbering system. 74
Fig. 15 Trichinella larvae settled in muscles 76
Fig. 16 Hookworms, strongyles, and whipworms 79
Fig. 17 Stone ground corn tortillas, high in calcium. 81
Fig. 18 Your lymph nodes are your best friends 90
Fig. 19 Colon. 97
Fig. 20 Keep personal water bottles sterile. 98
Fig. 21 Estrogen and progesterone levels (pg./ml) 114
Fig. 22 The liver has a large lobe on your right side with the gallbladder tucked inside. The left lobe is smaller. 128
Fig. 23 Gallstones 130
Fig. 24 Stomach, esophagus, diaphragm, gallbladder. 132
Fig. 25 Hiatal hernia. 133
Fig. 26 Lymph node neck glands 157
Fig. 27 Eye. 160
Fig. 28 Bottling equipment should be rinsed with ethyl (grain) alcohol, not propyl alcohol or wood alcohol. 183

Fig. 29 Tight diapers are a modern atrocity, forcing mercury and thallium into the baby's sponge-like skin. 190
Fig. 30 Some cysticercus varieties (types) have multiple heads....... 197
Fig. 31 Emerged cysticercus. 198
Fig. 32 Sheep liver flukes. Black threads in toilet are indicative of fluke remains. 222
Fig. 33 I found all calorie boosters to be polluted with wood alcohol. Make your own. 240
Fig. 34 Alcoholic beverages contain ergot and aflatoxin. Add vitamin C (1/8 tsp.) upon opening and wait 10 minutes for it to act..... 280
Fig. 35 Chlorinated water can cause mental problems...................... 283
Fig. 36 Red blood cells, top and side view. 285
Fig. 37 Water pitcher with filter. 288
Fig. 38 Sea salt flats are often roosting places for sea gulls. 311
Fig. 39 Too hot and too cold.. 323
Fig. 40 Products with propyl alcohol... 334
Fig. 41 U.S. regulations on sterilizing solutions. 348
Fig. 42 Keep both salt and vitamin C powders in closable plastic shakers. 349
Fig. 43 Grilled food has benzopyrenes as do hot dogs and "smoke flavored" foods. 352
Fig. 44 Mexican made candy with no benzene. 354
Fig. 45 Pollutants are in unlikely places.. 355
Fig. 46 Every brand of popcorn and corn chips I tested had zearalenone contamination. 355
Fig. 47 This Mexican bread has no molds. It is made from white, unbleached flour and is not wrapped in plastic..................... 362
Fig. 48 Goods baked in a panaderia (Mexican style bakery) never tested positive to molds. 367
Fig. 49 Starter set of homeopathic remedies...................................... 369
Fig. 50 All U.S. brands of pasta I tested had mold, including health food brands like the one shown here (left). No Mexican brands of pasta, like the one pictured (right), had any molds.......... 385
Fig. 51 All cold cereals I tested were full of mold toxins (besides solvents), health food varieties were worst. 387
Fig. 52 Packaged herb tea is moldy and polluted with solvents. Get yours in bulk from an herb company. 390
Fig. 53 Three safe flavorings. 391
Fig. 54 Don't eat the green on the potato... 392
Fig. 55 Potato harvest of the future. 393

Fig. 56 Smooth floors allow every bit of dust to be removed. 404
Fig. 57 Moving into the summer kitchen got you away from the accumulation of filth from winter! 406
Fig. 58 Mite 408
Fig. 59 Dental replacements 411
Fig. 60 Tops and bottoms of some metal crowns. 412
Fig. 61 Thallium excerpt. 418
Fig. 62 More dental metal 419
Fig. 63 Beautiful plastic mouth. 419
Fig. 64 Ugly metal in mouth. 420
Fig. 65 Unpolluted breakfast cereals. 421
Fig. 66 Lawful uses of solvents in food 428
Fig. 67 Some unsafe beverages. 430
Fig. 68 Some good foods 432
Fig. 69 All breads I tested had mold if they were in plastic 433
Fig. 70 Good restaurant foods 434
Fig. 71 Examples of commercial “food lube.” 437
Fig. 72 Detergents with PCBs. 438
Fig. 73 Safe soaps 438
Fig. 74 Some polluted supplements 439
Fig. 75 Pure salt 440
Fig. 76 Laboratory salt label 441
Fig. 77 Garage fumes. 445
Fig. 78 Our future, unless we act 456
Fig. 79 Syncrometer schematic 458
Fig. 80 Syncrometer parts list. 459
Fig. 81 Bottle with white blood cells taped on top. 470
Fig. 82 Some purchased parasite and tissue slides. 479
Fig. 83 Table of the future. 483
Fig. 84 Microscope, slides and coverslips 498
Fig. 85 Setup for finding resonant frequencies 503
Fig. 86 Function generators 512
Fig. 87 Light colored, roasted peanuts in the shell had no aflatoxin. 519
Fig. 88 Make a bottle of borax liquid to fill your soap dispensers and shampoo bottle. Use citric acid to rinse and condition. 528
Fig. 89 These are gallstones. 554
Fig. 90 Chart of bandwidths for organism families. 562

Preface

The sick have been held hostage for their money or intangible assets since time immemorial. Doctors, even primitive and natural healers, surround themselves with mystery as they use herbs or chemicals and incantations or “prognoses” to help the sick recover. Today, the medical industry (doctors and their suppliers and insurers) take a significant amount of the worker's earnings. Wouldn't it be nice if they could all go back to gardening or some other primitive and useful endeavor? Wouldn't it be wonderful if the sick could join them?

The most promising discovery in this book is the effectiveness of electricity to kill viruses, bacteria and parasites. Does this mean you can cancel your appointment with your clinical doctor? No it does not. Killing your invaders does not make you well instantly. But happily, at your next doctor visits she or he will be removing drugs, not adding them.

You might think that such an invention should be quickly patented. That was my universal advice. But I chose not to. It helps me, my children, and my grandchildren, if you are well. The whole world needs to come out of the dark ages of medicine and illness. And to learn the true causes of infection and disease. We must and can usher in the **new age of disease-free living.**

No diabetes, no high blood pressure, no cancer, no HIV/AIDS, no migraines, no lupus and so on!

Not a single disease is left unconquerable with this new understanding!

The Promise

Step into a new world.

A world without chronic diseases.

Step out of your old world.

It has kept you a prisoner.

Try something new.

The prison has no walls. It has only lines. Lines that mark the ground around you. Inside the lines are your old ideas. Outside are new ideas that invite you to step over and escape your prison. Dare to try these new ideas and your illness promises to recede. In a few weeks it can be gone.

If you are very ill or chronically ill you must have asked yourself many times: why have these problems chosen me? Will there never be a way to conquer them?

You may be quite familiar with your doctor's explanation of your illness or your child's illness. A *Coxsackie* virus has entered your child's brain causing inflammation (encephalitis) there. You pray that your child's immunity will overcome it. You may be familiar, but so very helpless against this microscopic invader.

If you had the proverbial 3 wishes they would be: 1) please spare my child's life; 2) please make it so my child doesn't have permanent damage; 3) please bless and guide the wonderful doctors and nurses who are keeping the oxygen tent going, and are watching my child's temperature and vital signs.

What if you could turn a dial and in 3 minutes kill every Coxsackie virus in your child's body?

What if this had no side effects?

What if the virus never came back?

In this book you will learn how to do that. You will also learn why your child got encephalitis or other disease and how to prevent it forever.

If this is too mind boggling, just take it a step at a time: First, learn about the radio-type broadcasting that all living animals do. Second, find the "station frequencies" that your particular invader(s) broadcast at. Third, learn how to "jam" their frequency until they **expire**: it takes only minutes!

Finally, learn how to make your own diagnostic and treatment devices. The instructions are simple enough for anyone. Only by putting this power in your hands will it be safe from government regulation, however well intended.

Only Two Health Problems

No matter how long and confusing is the list of symptoms a person has, from chronic fatigue to infertility to mental problems, I am sure to find only two things wrong: they have in them **pollutants** and/or **parasites**. I never find lack of exercise, vitamin deficiencies, hormone levels or anything else to be a primary causative factor. So the solution to good health is obvious:

Problem	**Simplest Cure**
Parasites	Electronic and herbal treatment
Pollution	Avoidance

It's a valiant quest: The quest for health. With optimism in one hand and determination in the other, you too can work the miracles for yourself that my clients accomplished in the case histories.

More good news is that it is not expensive. The cost will range from a few hundred dollars to only a few thousand in order to eliminate both problems and cure your chronic diseases.

Be A Health Detective

After curing your own diseases, teach your friends and family how it's done. Families are related and their problems are related. This should make the task easier. Keep a small notebook to become part of the treasured family legacy as much as photographs do. If your aunt, father and brother had diabetes as well as yourself and all were cured after introducing them to this concept and technology, isn't this worthy of notes in your family's history?

Notice what a strong line of inheritance there can be, **not due to sharing genes** but due to sharing a roof, a table, a supermarket, and a dentist!

Many problems can be disinherited. Cure yourself of retinitis pigmentosa, Muscular dystrophy (the “inherited” kind), and break down your family's faith in the gene-concept for these diseases. Bring hope to your family by proving diseases' true etiology. Bring respect back for your loyal genes that bring you hair color, and texture, not hair loss. That bring you eye color, not eye disease. Your genes brought you the good things about your ancestors, not the bad things. Parasites and pollution brought you the bad things.

Killing all your invaders is just the first step, though. It is indeed the life-saving step. But getting well is more than saving your life. Next comes the more tedious task of finding their sources. Where did they come from? Why did they invade you so massively. Why you?

The story of your personal pollution unfolds as in a book. Look closely and you see the whole panorama of your numerous tiny invaders being held at bay by your valiant immune system, your white blood cells. You can see what they are fighting besides the invaders. Your ill-chosen diet and lifestyle products!

Your heart may go out to those tiny white blood cells. Never again, you may say, will you give them arsenic and mercury and lead. Never again, cobalt and asbestos and freon.

That great body of wisdom, your body, the same as listened to your three wishes, will reward you over and over as you co-operate with it, until you have had not 3 but 30 wishes granted, each one seemingly as impossible as climbing Mt. Everest.

- Your chronic yeast infection can go away.
- Your hair can stop falling out–might even grow back.
- Your body can become pregnant–when you had already given up.
- Your fatigue can vanish.
- Your insomnia can be gone.
- Your warts can fall off.
- Your sight and hearing can sharpen.
- Your constant hunger can disappear.

Health isn't just being free of sickness. Health is feeling great, feeling like laughing at funny things. Health is feeling grateful to be alive. It is feeling happy to see the sky and to see growing things and to feel confident in human society's progress. Health is remembering the good parts of childhood and believing you still have a lot of them.

The Discovery

What makes me think I can find things in the human body that a blood test can not? What new technology makes this possible? Why is electronic testing superior in many ways to chemical methods? What are my claims of electrically killing parasites based on?

In 1988 I discovered a new way to scan a body organ. It was electronic. We already can "see" an organ with a sonogram, X-rays, computerized tomography (CAT) scan, or with magnetic resonance imagery (MRI). These techniques can identify abnormal shapes in an organ without having to explore or guess. But my new electronic technique can check for viruses, bacteria, fungi, parasites, solvents and toxins, and in addition is simple, cheap, fast and infallible. Electricity can do many magical things; now we can add detecting substances in our body to that list.

The method rests on radio electronic principles.

If you match, very precisely, the capacitance and inductance properties of an external circuit so that its resonant frequency is the same as the emitted frequency coming from somewhere else, the circuit will oscillate. This means there will be positive feedback in an amplifier circuit. You can hear it. Like when a public address system squeals.

The external circuit I use is called an *audio oscillator*, quite easy to build or buy. Your body provides the emitted frequencies. When you combine the audio oscillator circuit with your body, and you hear *resonance*, then you have detected a match! Something in your body matches something in the circuit on the test plate. By putting a laboratory sample of, say, a virus on the test plate, you can determine if your body has that virus by listening for resonance. Hearing resonance is easy if you're a radio technician or musician. Others must patiently practice. The details are given in the Bioelectronics chapter (page 457).

You do not have to be an expert in anything to learn the electronic detection method. But a keen sense of hearing helps.

In 1988 I learned a way to put anything on my skin, blindfolded, and identify it electronically in a few minutes. I could taste something without flavor and identify it electronically. The system worked fine for detecting things in the skin and tongue. Would it be reliable for internal organs, too?

A whole world of discovery lay ahead of me. I wanted to know what was in my inner ear causing tinnitus, in my eyes causing pain, in my stomach causing indigestion and a thousand other things.

But behind the daily excitement of new discoveries, a gnawing question lingered in my mind. How is this possible without some pretty high frequency energy source, radio frequency in fact, running through my circuit? My audio oscillator was only 1000 Hz (*hertz*, or cycles per second); radio frequency is hundreds of thousands of Hz. And the phenomenon could be produced with an old-fashioned *dermatron*[1], too, that only puts out DC (direct current)–no frequencies at all!

A high frequency energy had to be coming from somewhere. Was it me? *Ridiculous*!

But there was a way to test. If my own body was putting forth the high frequency energy, it could be bled off and diverted into the ground with a correct size capacitor. This should stop the feedback oscillations. This turned out to be true; it was stopped. But *ridiculous* kept ringing in my ears and I tried another test. If there was indeed radio frequency (RF) running through my circuit I should be able to block it with the right snap-on choke. It did block. I thought of a third test. If this was truly a resonance phenomenon I should be able to add a capacitance to this circuit and see the resonance destroyed. Then add an inductance and see the resonance return. It did just that. I made graphs of the relationship between capacitance and inductance. They were entirely reproducible.

---

[1] The dermatron was invented decades ago and made famous by Dr. Voll. Establishment science disdained it!

Then why couldn't I see the RF on my RF oscilloscope? Probably because it was high frequency energy, not high energy frequency, and I didn't know how to amplify it above the background noise level. It was nevertheless not convincing. Yet much too tantalizing to ignore.

I thought of yet a fourth test. If I was really producing RF radiation that could be channeled through a circuit, I should be able to interfere with it by adding another RF radiation from an outside source. I added a frequency from my frequency generator, first at 1,000 Hz. Now there was no resonance. It interfered. Did this mean that my body was not producing radiation at 1,000 Hz? Or was my 1,000 Hz radiation being matched and canceled? I raised the frequency gradually, from 1,000 to 10,000 to 100,000 to 1,000,000 Hz. There was no resonance anywhere, and I couldn't draw any conclusions. It was 5 o'clock on Sunday afternoon. Quitting time. But one last look at my generator reminded me that it could reach 2,000,000 Hz and I was just at 1,000,000. One more quick experiment wouldn't take much time. I cranked it to 1,800,000 Hz. And now a resonance screamed out! Was I "hearing things?" No more interference. I did it over and over. Why was it resonating now and not before? Had I arrived at my body's own *bandwidth* (transmission range), and this was the reason it no longer interfered?

I found the lowest frequency that resonated to be 1,562,000 Hz. All frequencies that I checked (about 2,000) from there up to 2,000,000 (my frequency generator would go no higher) also resonated.

A year later I purchased a better frequency generator to search for the upper end of my bandwidth. Any frequency between 1,562,000 and 9,457,000 Hz could be added to the circuit and produce resonance.

It seemed obvious, then, that the human body broadcasts electrically, just like a radio station, but over a wide band of frequencies and very low voltages, which is why it has not been detected and measured until now.

Everything Has A Unique Frequency

It was a busy year, now 1989. I was determined to find a bandwidth for other living things: I found them for flies, beetles, spiders, fleas, ants. They were between 1,000,000 Hz and 1,500,000 Hz; cockroaches were highest amongst insects I tested.

Then came a dismaying finding. A dead insect had a bandwidth too! Much narrower, and near the top end of the same range it had when living, but distinctly present. So it wasn't altogether a living phenomenon.

But if dead things had a resonant bandwidth, then maybe a prepared microscope slide of a dead creature could be used, and my trips to the garden and telephone calls to abattoirs (for meat parasites) could cease. That was a lucky thought. My first slide was of the *human intestinal fluke*, a huge parasite, scourge of humanity. I had just found it to be present in the liver (not intestine) of every cancer sufferer I saw. The (dead) adult parasite had a resonant frequency around 434,000 Hz. Slides of that parasite's redia resonated nearby (432,000 Hz), as did its other stages.

Dead things still resonated! The entire catalog of biological supply companies, hundreds of specimens of viruses, bacteria, parasites, molds, and even toxins, were now available to research with this new technique!

Suddenly an idea bolted out of the blue. If a person were to hold on to the frequency generator while it was generating 434,000 Hz, what would happen to the adult fluke, if you were infected with it?

I tested this plan that same week on myself, not with the fluke but with *Salmonella* bacteria and *Giardia* and *Herpes* that I carried chronically. After a 3 minute treatment, I retested myself. I could no longer find them in my organs! There were no emissions at their characteristic frequencies. I repeated and repeated. Were they really dead? Maybe they were just numbed or were suddenly hiding. But symptoms were gone quickly too. My *Herpes* lesion stopped tingling. It was all too simple and unbelievable.

But was it safe? Within three weeks I had reliable data regarding the necessary level of electrical treatment. It only took 5 volts for three minutes at the specific frequency. It is not as if you had to use house current which would kill you, along with the parasite.

Selective Electrocution

In twenty minutes (three minutes at six different frequencies) a whole family could get rid of this parasite. Cancer cases showed that in a few hours the universal cancer marker, *ortho-phospho-tyrosine* could be banished from their bodies by killing this same parasite.

"Incurable" HIV cases lost their virus in a few hours, too. Laboratory retesting for HIV came back negative! Most cases of pain got immediate relief if I could identify the correct "bug" and have its frequency found by the next office visit. This seemed to be absolute proof that living things had an essential high frequency output of some kind of energy.

What was actually happening to the bacteria or parasites? If I could kill something as large as an *Ascaris* worm or intestinal fluke, then perhaps I could kill something even larger, like an earthworm or flea, something I could see with my own eyes instead of having to imagine its demise inside my body.

Ten minutes at a frequency chosen near the top of their broadcast range seemed to anesthetize them. But they didn't die. Later I checked the body bandwidth (the range of frequencies they emit) of each. The earthworms had lost a lot of their bandwidth, both at the top and bottom. The fleas seemed hardier; they had only lost a little. However they did not recover, even weeks later, from this loss.

Could it harm humans to douse them with RF frequencies **in their own bandwidth?** Quite probably, if the voltage were high enough. There was no need to experiment, though, because the parasites we want to kill have characteristic frequencies that **do not overlap** the characteristic frequencies of a human. In fact, they are far away (see the chart on page 17).

So my electronic method attacking illness was born. Find the resonant frequency of a bacterium, virus or parasite using a slide or dead bit. Treat the living invaders inside the human body with this frequency and in a matter of minutes they are no longer transmitting their own bandwidths—they are dead or sick and will be removed by our white blood cells.

It was a worrisome truth. Perhaps the department of defense would use this knowledge and develop super high voltage devices to kill people ("enemies") somewhere in the world. But I couldn't let sick people suffer. Besides, it would probably require a voltage much like lightning to kill people from a distance. Possibly a way could be found to shield yourself from frequencies harmful to humans by wearing a choke (inductor) coil which suppresses these frequencies. Remember, there was no recovery, just a slow death for my experimental animals. It must not happen to humans!

Meanwhile, people must be alerted that they can safely kill their invaders and heal their chronic illnesses. Invaders that have been increasing exponentially due to lowered immunity in recent decades. Possibly this is true for all species on our planet. The pollution of the entire biosphere has been increasing and with it the prospect of acquired immune deficiency syndrome (AIDS) for all of us.

Remember, though, that the true challenge is not to kill our invaders but to regain our health and immunity.

More than just parasites are making us sick! Pollution is too. Selective electrocution rarely makes people completely well. Sick people always have an environmental factor that must be corrected also.

How do we do that? The ship of “progress”, of increasingly complex, processed foods and products, must be turned around and simplicity become our goal. Survival is in simplicity of food intake, simplicity of life habits. Did Ralph Waldo Emerson foresee this when he said “To be simple is to be great”? Or will daily parasite and pathogen electrocution become another crutch that makes us just enough better that we can continue a detrimental lifestyle? Yet another “Band-Aid” treatment for our poisoned planet?

Bioradiation

Strange as it appears, it now seems obvious that every living creature broadcasts its presence like a radio station, the sun, or the stars. I have named it *bioradiation*.

Perhaps it is the same energy as the Asian *chi*; perhaps it is merely related to it. Perhaps it is the energy that runs along the meridians discovered eons ago by Asian practitioners.

Perhaps it is the energy that faith healers and religious teachers know how to harness, perhaps not.

Perhaps it is the energy that psychics perceive and that drives occult phenomena, perhaps not.

What is truly amazing is that ordinary persons have discovered such energy well ahead of scientists. Persons using the “art” of kinesiology, pendulums, radionics, dousing rods and many other forms of “strange energy” have no doubt harnessed a part of this bioradiation. It is a tribute to the generally high intelligence of common people and to their open-mindedness that they discovered this energy, **in spite of** opposition from scientists of today.

Over a century ago the scientists of Europe proposed the existence of a “life force” called “élan vitale.” They were scorned out of existence (and out of jobs). Young scientists, (including myself) were systematically taught to scorn this idea. Of course we were also taught that a good scientist was unemotional, does not scorn ideas, has a completely open mind, and does not rule something out until it is disproved to their satisfaction. The youthfulness of college years is so susceptible to prejudices of all kinds, and the desire for acceptance is so great, that special effort needs to be made to teach neutrality. Or at least to distinguish between emotion and fact. Where have these basic pedagogic principles gone? I was indeed inspired with the phrase “search for truth” but then promptly led down the path of “search for acceptance.”

I do not know what bioradiation, this electrical broadcast from our cells, is made of. Only its frequency was noticed and caught (modulated) in such a way as to be measurable. And this frequency, 1,520,000 to 9,460,000 Hz (for a human infant) is in the radio frequency (RF) range[2].

Anyone who is experienced with RF knows its strange behavior. Not strange in the "unknown" sense but in the amazing sense. Circuits don't need to be complete or closed for it to travel. Bodies and objects can "pick it up" without being in the circuit. These amazing properties are due to the capacitive and inductive properties of objects all around us, including ourselves.

Zapping Bugs

By *zapping* I mean selectively electrocuting pathogens. For years I used a commercial frequency generator to "zap" one pathogen after another.

First I made a chart of the frequencies for most of the bacteria and viruses in my collection (over 80, see page 561). Then I would test the sick client for each one of these, and hope they did not have one for which I didn't have a sample. Even persons with a simple cold typically had a dozen they tested positive to (not just *Adenovirus*).

Next it was time to tune in the frequency generator to a dozen frequencies for three minutes each. The total process, testing and treatment, would take about two hours. They frequently got immediate relief. But often the relief would be temporary. What I didn't know at that time was that viruses could infect a larger parasite such as a roundworm. Until you killed your roundworm and your virus, you would keep getting the virus back promptly.

---

[2] AM radio broadcasts are from 540,000 Hz to 1,600,000 Hz (slight overlap with lower end of human band), FM is 88,000,000 to 108,000,000 Hz (out of the human range).

In 1993 my son, Geoffrey, joined me and we tried a new approach. He programmed a computer controlled frequency generator to automatically cover all the frequencies populated by all the parasites, viruses, and bacteria, from 290,000 Hz to 470,000 Hz. It spent about three minutes for every 1000 Hz it covered. This was more efficient, but it meant spending ten hours being zapped.

Again, the results were disappointing. Arthritis pain, eye pain, colds were improved, but not completely cured overnight. Months later I would find that organisms were transmitting as low as 170,000, and as high as 690,000 Hz. My specimen collection was obviously incomplete. To cover this larger range, spending three minutes for every 1000 Hz, would take 26 hours. Still worth doing if it would indeed help all our illnesses. But even this method of zapping was not 100% effective for reasons yet to become clear.

In 1994 my son built a hand held, battery operated, accurate frequency generator. The purpose was to enable everyone to kill the intestinal fluke at 434,000 Hz with a low cost device. Enough benefit would be derived from zapping at various frequencies that I thought everyone should know how to make one. When I tested it on one of my own bacteria, however, **three others at much different frequencies died also!** This had never happened before. When I tested it on others, even though they had dozens of pathogens, all were killed!

Subsequent testing showed it was not due to some unique design, or special wave form produced by the device. It was due to battery operation!

Any positively offset frequency kills all bacteria, viruses and parasites simultaneously

given sufficient voltage (5 to 10 volts), duration (seven minutes), and frequency (anything from 10 Hz to 500,000 Hz).

Before this I had always set my commercial frequency generator to alternate between positive and negative voltage. Now I tried setting it to alternate between positive and zero voltage (*positive offset*). It was just as effective as the battery operated frequency generator my son designed.

Generating positive offset frequencies is the best way to kill all pathogens quickly.

But it takes more than one treatment.

It takes three treatments to kill everything. Why? The first zapping kills viruses, bacteria and parasites. But a few minutes later, bacteria and viruses (different ones) often recur. I conclude they had been infecting the parasites, and killing the parasites released them. The second zapping kills the released viruses and bacteria, but soon a few viruses appear again. They must have been infecting some of the last bacteria. After a third zapping I never find any viruses, bacteria or parasites, even hours later.

Why didn't the virus inside the parasite die with the first zapping? It may be because electricity travels on the exterior of things. The body of the parasite shielded the interior. This is why my earlier, promising work spending hours on a frequency generator gave only partial or temporary improvement—it was only done once, not three times. And it explains why a single treatment with a frequency generator or zapper frequently gives you a cold!

Zapping does not kill shielded organisms

such as those that may be in the middle of your stomach or intestines. The electricity travels along the stomach or intestine wall, not through their contents.

So zapping is still not perfect, but can bring such manifest relief that everyone should buy or make one. Parts cost less than $25.00; the plans are in the next chapter.

The Bioradiation Spectrum

Everything emits a characteristic range of frequencies (bandwidth). In general, the more primitive the organism, the lower its bandwidth. Advanced animals have higher frequencies and the range is wider.

BANDWIDTH OF BIORADIATION OF ANIMALS

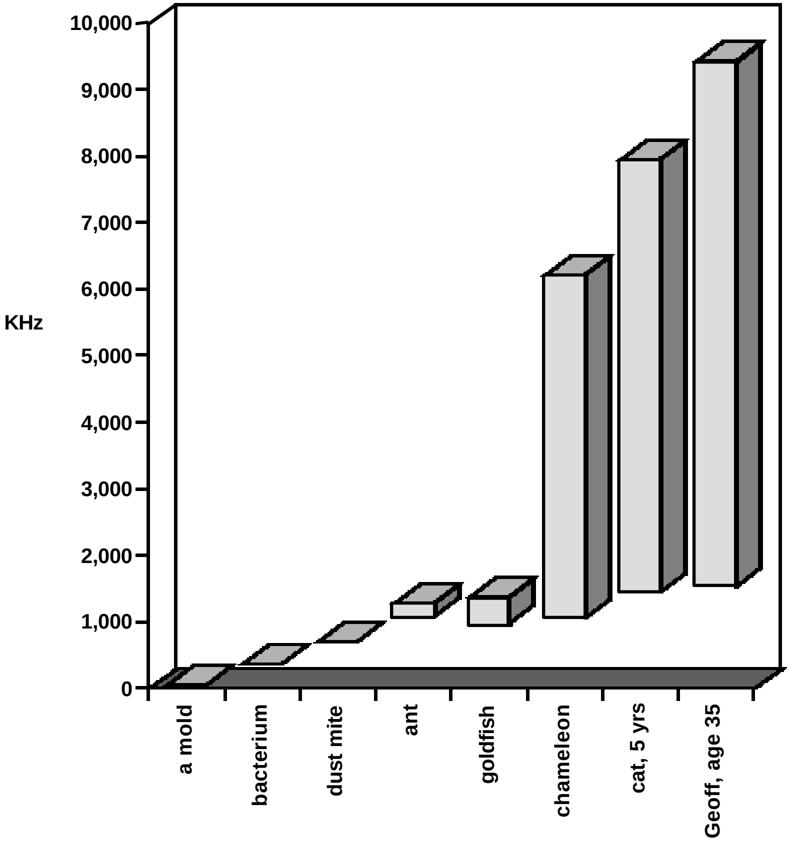

Fig. 1 Selected animal bandwidths.

The human range is from 1520 KHz to 9460 KHz. Pathogens (molds, viruses, bacteria, worms, mites) range from 77 KHz to 900 KHz. Fortunately for us we can work on zapping pathogens in the lower ranges without affecting humans in the upper range.

BIORADIATION OF TYPICAL PATHOGENS

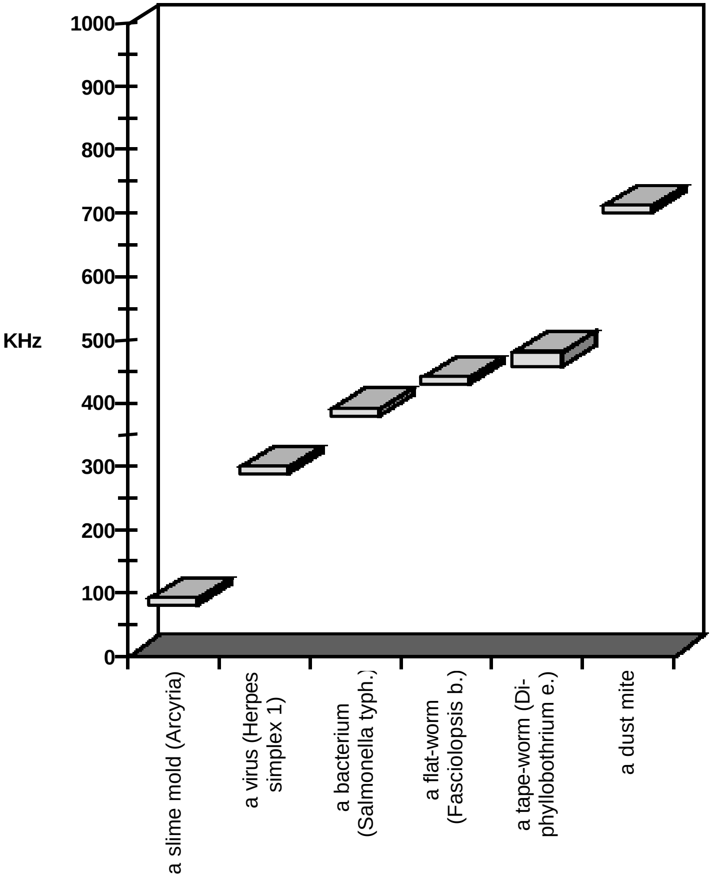

Fig. 2 Selected pathogen bandwidths.

Applying an alternating electrical voltage within an organism's bandwidth injures it. Small organisms with narrow bandwidths are extinguished quite readily (three minutes at five volts).

Positively offset frequencies can kill the entire range of small organisms (viruses, bacteria, parasites) in just seven minutes.

Building A Zapper

Being able to kill your bacteria and other invaders with electricity becomes much more of a panacea when you can do it all in three 7 minute sessions. No need to single out specific frequencies or to sweep through a range of frequencies one KHz at a time. No matter what frequency it is set at (within reason), it kills large and small invaders: flukes, roundworms, mites, bacteria, viruses and fungi. It kills them all at once, in 7 minutes, even at 5 volts.

How does it work? I suppose that a positive voltage applied anywhere on the body attracts negatively charged things such as bacteria. Perhaps the battery voltage tugs at them, pulling them out of their locations in the cell doorways (called *conductance channels*). But doorways can be negatively charged too. Does the voltage tug at them so they disgorge any bacteria stuck in them? How would the positive voltage act to kill a large parasite like a fluke? None of these questions can be answered yet.

Other fascinating possibilities are that the intermittent positive voltage interferes with electron flow in some key metabolic route, or straightens out the ATP molecule disallowing its breakdown. Such biological questions could be answered by studying the effects of positive frequencies on bacteria in a lab.

The most important question, of course, is whether there is a harmful effect on you. I have seen no effects on blood pressure, mental alertness, or body temperatures. It has never produced pain, although it has often stopped pain instantly. This does not prove its safety. Even knowing that the voltage comes from a small 9 volt battery does not prove safety, although it is reassuring. The clotting of red blood cells, platelet aggregation and functions that depend on surface charges on cells need to be investigated. But not before you can use it. Your safety lies in the short period of exposure that is necessary. Viruses and bacteria disappear in 3 minutes; tapeworm stages, flukes, roundworms in 5; and mites in 7. One need not go beyond this time, although no bad effects have been seen at any length of treatment.

The first seven minute zapping is followed by an intermission, lasting 20 to 30 minutes. During this time, bacteria and viruses are released from the dying parasites and start to invade you instead.

The second seven minute session is intended to kill these newly released viruses and bacteria. If you omit it, you could catch a cold, sore throat or something else immediately. Again, viruses are released from the dying bacteria. The third session kills the last viruses released.

Do Not Zap If You Are Pregnant
Or Wearing A Pacemaker.

These situations have not been explored yet. Don't do these experiments yourself. Children as young as 8 months have been zapped with no noticeable ill effects. For them, you should weigh the possible benefits against the unknown risks.

That is all there is to it. Almost all. The zapping current does not reach deep into the eyeball or testicle or bowel contents. It does not reach into your gallstones, or into your living cells where *Herpes* virus lies latent or *Candida* fungus extends its fingers. But by zapping 3 times a day for a week or more you can deplete these populations, too, often to zero.

Killing The Surviving Pathogens

The interior of gallstones may house parasites inaccessible to the zapping. Eliminate this source of reinfection by flushing them out with liver cleanses (page 552).

Although the center of the bowel contents is often unaffected by electric current, which lets bowel bacteria like *Shigella, Escherichia coli (E. coli)* and parasite stages survive, sometimes it is nearly all sterilized by zapping. This results in considerable shrinkage of the bowel movement. Eliminate remaining parasites and bacteria with a single dose (2 tsp.) of Black Walnut Hull Tincture, Extra Strength (see page 543).

There is no way of distinguishing between "good" and "bad" bacteria with either of these methods. However even good bacteria are bad if they come through the intestinal wall, so zapping targets mostly "bad" bacteria. The good news is that perfect bowel habits often result in a few days. Evidently, the good bacteria are benefited by killing the invasive ones. Home-made yogurt and buttermilk (see *Recipes*) are especially good at recolonizing the bowel. But it does not seem wise to culture yourself with special commercial preparations and risk getting parasite stages again when you can become normal so soon anyway. If you do decide to take some *acidophilus* bacteria to replenish your intestinal flora make sure you test for parasites like *Eurytrema* first.

When a large number of parasites, bacteria and viruses are killed, it can leave you fatigued. Try to give yourself a low-stress day after your initial zapping. But there are no significant side effects. I believe this is due to the second and third zapping which mops up bacteria and viruses that would otherwise be able to go on a feeding frenzy with so much dead prey available.

To build your zapper you may take this list of components to any electronics store (Radio Shack part numbers are given for convenience).

Zapper Parts List

Item	Radio Shack Catalog Number
large shoe box	
9 volt battery	

9 volt battery clips	270-325 (set of 5, you need 1)
On-Off toggle switch	275-624A micro mini toggle switch
1 KΩ resistor	271-1321 (set of 5, you need 2)
3.9 KΩ resistor	271-1123 (set of 2, you need 2)
low-current red LED	276-044 or 276-041 or 276-045
.0047 uF capacitor	272-130 (set of 2, you need 1)
.01 uF capacitor	272-1065 (set of 2, you need 1)
555 CMOS timer chip	276-1723 (set of 2, you need 1)
8 pin wire-wrapping socket for the chip	276-1988 (set of 2, you need 1) Note: Radio Shack is discontinuing all wire wrap sockets. Find another parts store or use 276-1995 (but the legs are much shorter and harder to attach clips to).
short (12”) alligator clip leads	any electronics shop, get 6
Microclip test jumpers	278-017 (you need 2 packages of 2)
2 bolts, about 1/8” diameter, 2” long, with 4 nuts and 4 washers	hardware store
2 copper pipes, ¾diameter, 4” long	hardware store
sharp knife, pin, long-nose pliers	

Hints for absolute novices: Don't let unusual vocabulary deter you. A “lead” is just a piece of wire used to make connections. When you remove a component from its package, label it with a piece of tape. A serrated kitchen knife works best as does a large safety pin. Practice using the microclips. If the metal ends are L-shaped bend them into a U with the long-nose pliers so they grab better. Chips and chip holders are very fragile. It is wise to purchase an extra of each in case you break the connections.

R1	1K
R2	3.9K
R3	1K
R4	3.9K
C1	.01ì f
C2	.0047ì f
U3	MC1455
LED1	2 ma LED Red
Pin 1	ground
pin 8	power

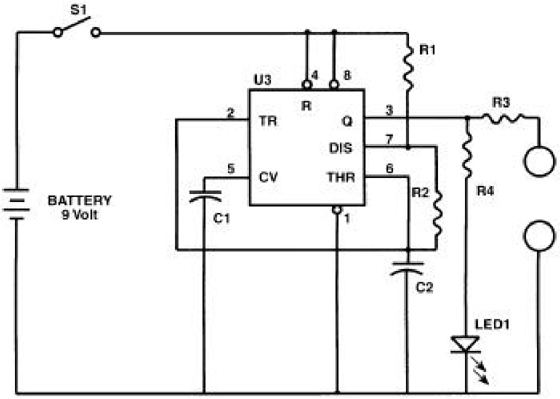

Give this to an electronics person or make it yourself in a shoebox by using the following instructions.

Fig. 3 Zapper schematic.

Assembling The Zapper

1. You will be using the lid of the shoe box to mount the components. Save the box to enclose the finished project.
2. Pierce two holes near the ends of the lid. Enlarge the holes with a pen or pencil until the bolts would fit through. Mount the bolts on the outside about half way through the holes so there is a washer and nut holding it in place on both sides. Tighten. Label one hole “grounding bolt” on the inside and outside.

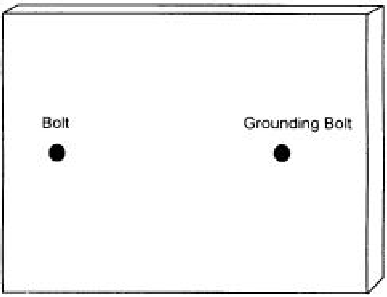

3. Mount the 555 chip in the wire wrap socket. Find the “top end” of the chip by searching the outside surface carefully for a cookie-shaped bite or hole taken out of it. Align the chip with the socket and very gently squeeze the pins of the chip into the socket until they click in place.
4. Make 8 pinholes to fit the wire wrap socket. Enlarge them slightly with a sharp pencil. Mount it on the outside. Write in the numbers of the pins (connections) on both the outside and inside, starting with number one to the left of the “cookie bite” as seen from outside. After number 4, cross over to number 5 and continue. Number 8 will be across from number 1.

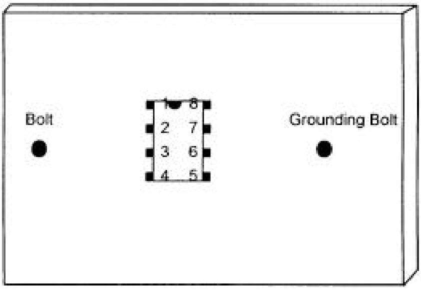

5. Pierce two holes ½ inch apart very near to pins 5, 6, 7, and 8. They should be less than 1/8 inch away. (Or, one end of each component can share a hole with the 555 chip.) Mount the .01 uF ca-

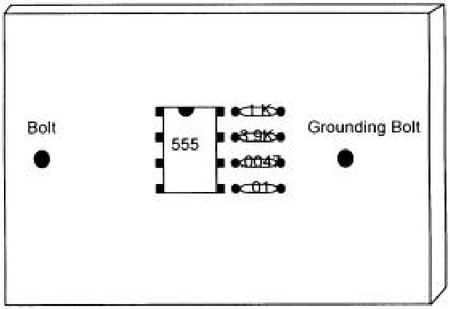

pacitor near pin 5 on the outside. On the inside connect pin 5 to one end of this capacitor by simply twisting them together. Loop the capacitor wire around the pin first; then twist with the long-nose pliers until you have made a tight connection. Bend the other wire from the capacitor flat against the inside of the shoe box lid. Label it .01 on the outside and inside. Mount the .0047 uF capacitor near pin 6. On the inside twist the capacitor wire around the pin. Flatten the wire from the other end and label it .0047. Mount the 3.9 KΩ resistor near pin 7, connecting it on the inside to the pin. Flatten the wire on the other end and label it 3.9. Mount the 1 KΩ resistor and connect it similarly to pin 8 and label it 1K.

6. Pierce two holes ½inch apart next to pin 3 (again, you can share the hole for pin 3 if you wish), in the direction of the bolt. Mount the other 1 KΩ resistor and label inside and outside. Twist the connections together and flatten the remaining wire. This resistor protects the circuit if you should accidentally short the terminals. Mount the 3.9KΩ resistor downward. One end can go in the same hole as the 1K resistor near pin 3. Twist that end around pin 3 which already has the 1K resistor attached to it. Flatten the far end. Label.

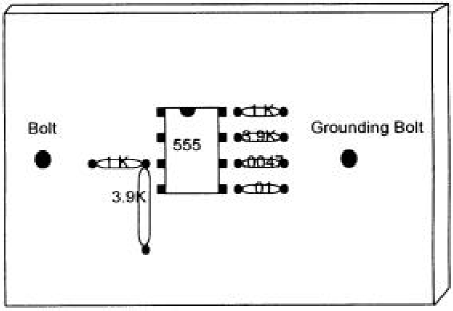

7. Next to the 3.9KΩ resistor pierce two holes ¼ inch apart for the LED. No-

tice that the LED has a positive and negative connection. The longer wire is the anode (positive). Mount the LED on the outside and bend back the wires, labeling them + and - on the inside.

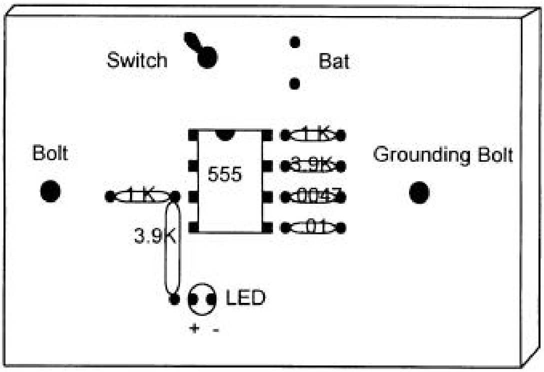

8. Near the top pierce a hole for the toggle switch. Enlarge it until the shaft fits through from the inside. Remove nut and washer from switch before mounting. You may need to trim away some paper with a serrated knife before replacing washer and nut on the outside. Tighten.
9. Next to the switch pierce two holes for the wires from the battery holder and poke them through. Attach the battery and tape it to the outside.

NOW TO CONNECT EVERYTHING

First, make holes at the corners of the lid with a pencil. Slit each corner to the hole. They will accommodate extra loops of wire that you get from using the clip leads to make connections. After each connection gently tuck away the excess wire.

1. Twist the free ends of the two capacitors (.01 and .0047) together. Connect this to the grounding bolt using an alligator clip.
2. Bend the top ends of pin 2 and pin 6 (which already has a connection) inward towards each other in an L shape. Catch them both with an alligator clip and attach the other end of the alligator clip to the free end of the 3.9KΩ resistor by pin 7.

3. Using an alligator clip connect pin 7 to the free end of the 1KΩ resistor attached to pin 8.
4. Using two microclips connect pin 8 to one end of the switch, and pin 4 to the same end of the switch. (Put one hook inside the hole and the other hook around the whole connection. Check to make sure they are securely connected.)
5. Use an alligator clip to connect the free end of the other 1KΩ resistor (by pin 3) to the bolt.
6. Twist the free end of the 3.9KΩ resistor around the plus end of the LED. Connect the minus end of the LED to the grounding bolt using an alligator clip.
7. Connect pin number 1 on the chip to the grounding bolt with an alligator clip.

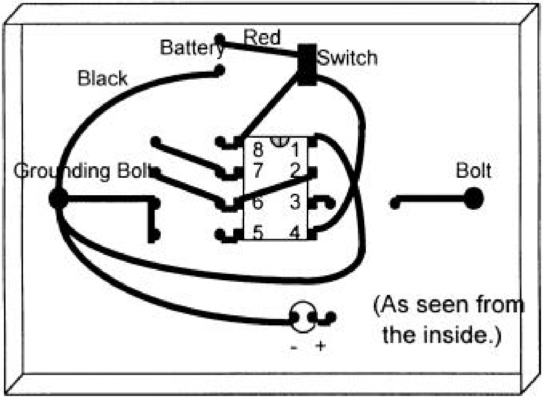

8. Attach an alligator clip to the outside of one of the bolts. Attach the other end to a handhold (copper pipe). Do the same for the other bolt and handhold.
9. Connect the minus end of the battery (black wire) to the grounding bolt with an alligator clip.
10. Connect the plus end of the battery (red wire) to the free end of the switch using a microclip lead. If the LED lights up you know the switch is ON. If it does not, flip the switch and see if the LED lights. Label the switch clearly. If you cannot get the LED to light in either switch position, you must double-check all of your connections, and make sure you have a fresh battery.

11. Finally replace the lid on the box, loosely, and slip a couple of rubber bands around the box to keep it securely shut.

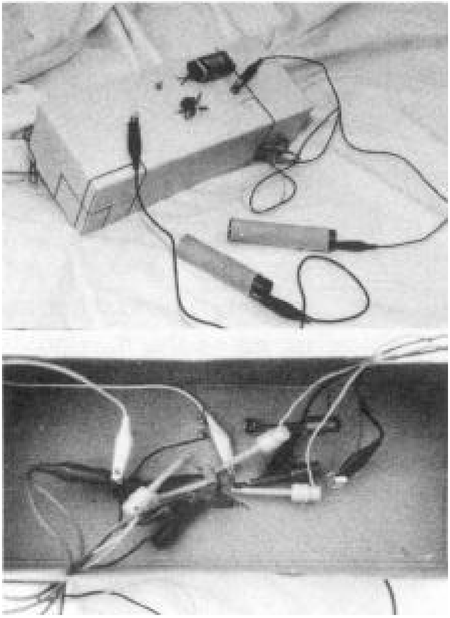

Fig. 4 Finished zapper, outside and inside.

- Optional: measure the frequency of your zapper by connecting an oscilloscope or frequency counter to the handholds. Any electronics shop can do this. It should read between 20 and 40 kHz.
- Optional: measure the voltage output by connecting it to an oscilloscope. It should be about 8 to 9 volts. **Note: a voltage meter will only read 4 to 5 volts.**
- Optional: measure the current that flows through you when you are getting zapped. You will need a 1 KΩ resistor and oscilloscope. Connect the grounding bolt on the zapper to one end of the resistor. Connect the other end of the resistor to a handhold. (Adding this resistor to the circuit decreases the current slightly, but not significantly.)

The other handhold is attached to the other bolt. Connect the scope ground wire to one end of the resistor. Connect the scope probe to the other end of the resistor. Turn the zapper ON and grasp the handholds. Read the voltage on the scope. It will read about 3.5 volts. Calculate current by dividing voltage by resistance. 3.5 volts divided by 1 KΩ is 3.5 ma (milliamperes).

Using The Zapper

1. Wrap handholds in one layer of wet paper towel before using. Grasp securely and turn the switch on to zap.
2. Zap for 7 minutes, let go of the handholds, turn off the zapper, and rest for 20 minutes. Then 7 minutes on, 20 minutes rest, and a final 7 minutes on.

Trying the zapper on an illness to see “if it works” is not useful. Your symptoms may be due to a non-parasite. Or you may reinfect within hours of zapping. The best way to test your device is to find a few invaders that you currently have (see Lesson Twelve, page 492, or Lesson Twenty Seven, page 509). This gives you a starting point. Then zap yourself. After the triple zapping, none of these invaders should be present.

Simple Pulser

If you are ill or want a reliable zapping, make the first model. However, there is another way to make a zapper if you can not afford to build the first model.

An ordinary battery is a source of positive voltage. It is the positive voltage that eliminates so many parasites at once, not a specific frequency. So although the zapper's frequency is about 30 kHz (thirty thousand “zaps” per second), even 5 Hz (five "zaps" per second), **about as fast as you can tap the battery with your hand, is moderately effective!**

You must be connected to both terminals. One will be marked + (positive) and the other – (negative). If you simply touch these terminals with your wet fingers, nothing much happens. That is because your resistance to the current starts going up right away, so less and less current passes through you.

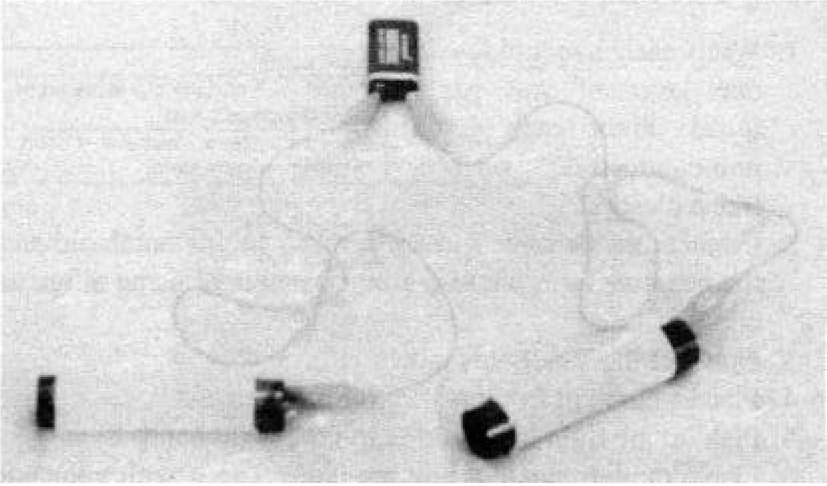

Fig. 5 Simple pulser.

However if you tap the positive terminal with your wet hand, and tap it at a fairly high rate, your body's capacitors come into play. Capacitors only take part in the flow of electricity when they are charged and discharged. Tapping the terminal starts and stops the voltage so capacitors charge and discharge. This kind of resistance to current flow is much smaller.

The faster you tap, the greater the frequency of current pulses and the lower this kind of resistance becomes. Now you can have a considerable sustained current flow through your body.

If you can tap even twice per second (2 Hz) for ten minutes without interruption you can give yourself a zapping that is moderately effective. Remember to take an intermission of twenty minutes and then repeat to avoid catching new viruses. After a second twenty minute intermission repeat zapping a third time.

Using The Simple Pulser

9 volt battery
2 short (12") alligator clip leads (from any electronics shop)
2 copper pipes, ¾" diameter, 4" long (from a hardware store)

1. Wrap each handhold with one layer of wet paper towel. Place each on a non-conductive surface, like a plastic bag.
2. Connect the positive battery terminal to one handhold and the negative terminal to the other handhold using alligator clip leads.
3. Don't let the handholds touch.
4. Place a clock in front of you to time yourself.
5. Pick up the right handhold with your right hand.
6. Leave the left handhold on the table. Tap it with your left hand, preferably the fleshy part of the palm. You may brace yourself with your fingers on the plastic. Keep up a steady pace as fast as you are able.
7. When you get tired pick up the left handhold with your left hand and tap with your right hand. Keep changing off with the least interruption.
8. Repeat a second time 20 minutes later, and a third time 20 minutes after that.

A single 9 volt battery will wear out rather quickly used this way. Put two together, in parallel, for longer lasting power. This requires two more short alligator clips. Connect positive terminals of the batteries to each other, and the negatives also.

Parasites & Pollution

The word "parasites" is used in two senses. Everything living on you or in you, not just to perch, but to take its food from you is a *parasite*. No matter what its size, it can be called a parasite.

But in some way the big worms need to be distinguished from the medium-sized amoebae, the even smaller bacteria and the smallest of all—viruses. So often the term parasite is reserved for the bigger things, from amoebae on up. In this book, the word parasite will be used in both ways as usual. You can easily guess what is meant.

Parasitic worms are divided into *roundworms* and *flatworms*. Roundworms are round like earthworms even though they may be as thin as hairs (threadworms, *filaria*) or microscopically small (like *Trichinella*). Flatworms are more like leeches. They have a way to attach themselves sometimes with the head (*scolex*) like *tapeworms*, sometimes with a special sucker like *flukes*.

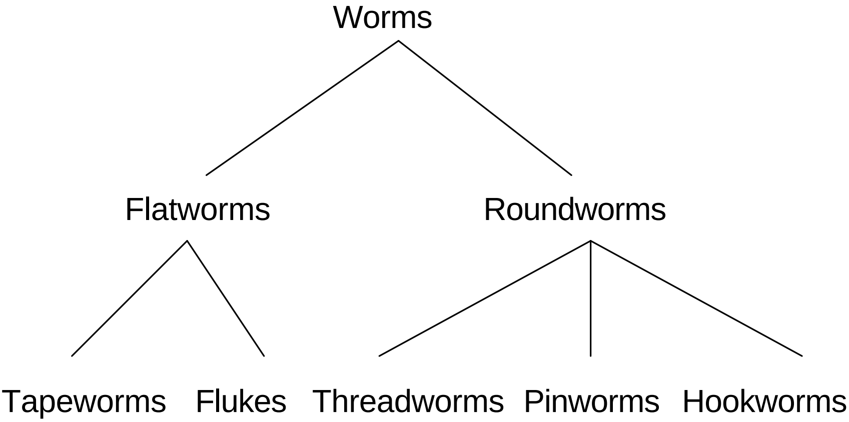

Worm parasites go through stages of development that can look very, very different from the adult.

Roundworms like *Ascaris* (common cat and dog roundworm), are simplest. The **eggs** are swallowed by licking or eating a bit of filth. They hatch into a tiny *larva*. The larva treks to the lungs. You cough it up and swallow it. Meanwhile it has molted a few times. It then crawls to the intestine where it becomes an adult, shedding eggs in your stool.

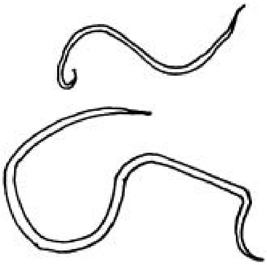

Fig. 6 Ascaris.

Worms usually have preferred locations. The favorite organ for *Dirofilaria* (dog heartworm) is the heart (even human heart). Sometimes the rules can be broken. My tests show *Dirofilaria* can live in other organs, too, if they are sufficiently polluted with solvents, metals and other toxins.

Flatworms like **tapeworms** are much more complicated in their life history. You could eat the eggs accidentally with dirt. After hatching, the tiny larva burrows into its favorite organ. Your body encases it with a cyst. The white blood cells have been taught never to attack your body...and the cyst case is your body! So the tapeworm stage has safe residence for some time. If you are a meat eater, you could eat such a cyst if it happens to be lodged in the meat you are eating! Your teeth break it apart as you crunch. The little larva is swallowed and tries to attach itself to your intestine with its head. Then it grows longer by making segment after segment. The segments with their eggs leave with the bowel contents. I often see dog tapeworm of the small variety in their human family.

Flatworms like **flukes** are also very complicated. The eggs, passed out with bowel contents were not meant to be eaten as such. They were meant to hatch in a pond where snails and minnows eat them. The larva grows up in these new "secondary" hosts. Later, the snail sheds them and they attach themselves to foliage near the pond. They over-winter in a tough *metacercarial* cyst. An unsuspecting browsing animal now eats them. They come out of their metacercarial cyst as a small adult and quickly attach themselves to the intestine with a sucker. They now have "safe haven" and can go about maturing and laying eggs.

Four common flukes are: **human intestinal fluke, human liver fluke, sheep liver fluke, pancreatic fluke of cattle.** Don't let the terms sheep and cattle mislead you. They are all found in humans.

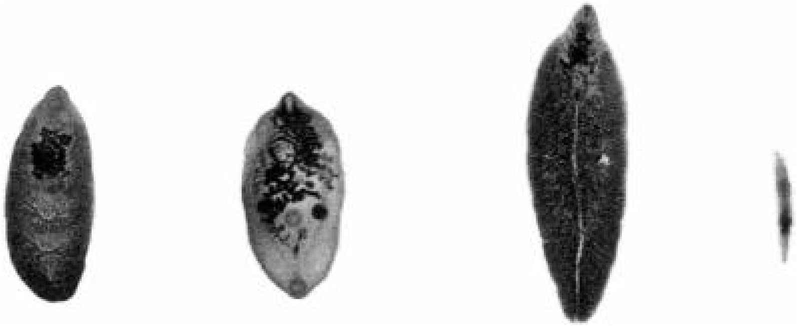

Fig. 7 (L to R) Human intestinal fluke, pancreatic fluke, sheep liver fluke, and human liver fluke.

The Worst Parasite

Fasciolopsis buskii is the fluke (flatworm) that I find in every case of cancer, HIV infection, Alzheimer's, Crohn's disease, Kaposi's, endometriosis, and in many people without these diseases. Its life cycle involves six different stages:

Stage	**Normal Life Cycle**	
1 Egg	Expelled with bowel movement onto soil. Washed by rain into ponds.	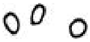
2 Miracidia	Hatches from egg in water. Has cilia, can swim vigorously and must find intermediate snail host in one to two hours or may be too exhausted to invade.	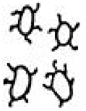
3 Redia	Develop inside miracidia as little balls until expelled. Those are "mother" redia, and each one bears "daughter" redia for up to 8 months, all still inside the snail, and living on the fluids in the lymphatic spaces. Similarly, daughter redia are continually developing cercaria.	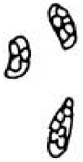
4 Cercaria	Have a tail, use it to exit from snail and swim to a plant. If the snail is feeding on a plant, cercaria can latch onto plant with sucker mouth and start to encyst (form a "cocoon") within minutes. Tail breaks off and swims away to dissolve.	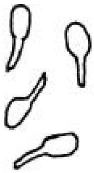
5 Metacercaria	Two-walled cyst. The outer wall is very sticky. But as you eat the plant it is stuck to, the least pressure will break it, leaving the cyst in the mouth. The "almost unbreakable" inner cyst wall protects it from chewing, and the keratin-like coat prevents digestion by stomach juices. However when it reaches the duodenum, contact with intestinal juices dissolves away the cyst-wall and frees it. It then fastens itself to the intestinal lining and begins to develop into an adult.	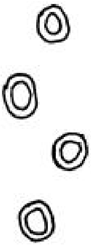

6 Adult	Lives in your intestine and can produce 1000 eggs per bowel movement and live many years.	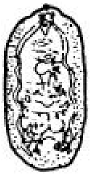

Fig. 8 Fasciolopsis' normal life cycle.

Note that the adult is the only stage that "normally" lives in the human (and then only in the intestine). *Fasciolopsis* depends on a snail, called a *secondary host*, for part of its life cycle. **But when your body has solvents in it, the other five stages can develop in you!**

If **propyl alcohol** is the solvent, the intestinal fluke is invited to use another organ as a secondary host—this organ will become cancerous. If **benzene** is the solvent, the intestinal fluke uses the thymus for its secondary host, setting the stage for AIDS. **Wood alcohol** invites pancreatic flukes to use the pancreas as a secondary host. This leads to pancreatic dysfunction which we call *diabetes*. If **xylene** (or **toluene**) are the solvents, I typically see any of four flukes using the brain as a secondary host. If **methyl ethyl ketone** (MEK) or **methyl butyl ketone** (MBK) are the solvents, the uterus becomes a secondary host and endometriosis a likely result.

This is a new kind of parasitism, based on pollution. I call the diseases caused by fluke stages in inappropriate locations **Fluke Disease**; it is discussed in more detail later (page 249).

Are tapeworms and roundworms affected by solvents this way, too? This is a fascinating and very important question. Search for the answer and help others search for the answer. I do not know yet.

Pollution

Pollutants are all the dead things around us that should not get into your body because they interfere with its work. As long as they don't penetrate your tissues, they won't interfere, like plastic eyeglasses and clothing. But if they are invasive, your body must fight to remove them.

Pollutants can invade your body via the air you breath, the foods and beverages you eat, and the products you put on your skin.

The biggest tragedy is not recognizing when a pollutant is harming you.

Two people can use the same face cream. One develops a rash, the other does not. The one who did not assumes the cream is not harmful to them…that they are like a bank vault, impregnable to that product. A better assumption is that the face cream is somewhat toxic, as evidenced by the rash that can develop, and they escaped the rash only because they had a stronger immune system. The immune system is like money, paid out of the bank vault, for every toxic invasion. When the money is gone, the bank (your health) fails.

Solvent Pollution

Solvents are compounds that dissolve things. Water is a useful, life giving solvent. Most other solvents dissolve fats and are life threatening, because fats form the membrane wall around each of our cells, especially our nerve cells.

The solvent that does the most harm is **benzene**. It goes to the thymus, ruins our immune system, and causes AIDS. The next worst solvent is **propyl alcohol**. It goes to the liver and causes cancer in some distant organ. Other major culprits of disease are **xylene, toluene, wood alcohol, methylene chloride,** and **trichloroethane** (TCE). I'll discuss each one later, with the ailment it's associated with.

Metal Pollution

Biochemists know that a mineral in raw element form always inhibits the enzyme using that mineral. Copper from the meat and vegetables you eat is essential. Inorganic copper, like you would get from a copper bottomed kettle or copper plumbing, is carcinogenic[3]. Unfortunately, the inorganic form of metals is what pervades our environment. We put metal jewelry on our skin, eat bread baked in metal pans, and drink water from metal plumbing.

Another obvious metallic threat is tooth fillings. **Mercury amalgam fillings**, despite the assurances of the American Dental Association, are not safe. And sometimes the mercury is polluted with **thallium**, even more toxic than mercury! **Gold** and **silver** seem to have fewer harmful effects, but no one should have any pure metal in or on their body.

Other prevalent toxic metals include **lead** and **cadmium** from soldered and galvanized plumbing, **nickel** and **chromium** from dentalware and cosmetics, and **aluminum** from food and drink cans, and cooking pots.

---

[3] Haleem J. Issaq, The Role of Metals in Tumor Development and inhibition. From *Carcinogenicity and Metal Ions*, volume 10, page 61, of a series called *Metal Ions in Biological Systems*, edited by Helmut Sigel, 1980.

Mycotoxins

Molds produce some of the most toxic substances known, called *mycotoxins*. One small moldy fruit or vegetable can pollute a huge batch of juice, jam or other product. Although molds are alive, and can be killed by zapping, mycotoxins are not, and must be detoxified by your liver.

But because mycotoxins are so extremely poisonous, a tiny amount can incapacitate a part of the liver for days!

Aflatoxin is the most common mycotoxin I detect. It is produced by molds that grow on quite a variety of plants. For that reason I am always cautioning people to eat only perfect citrus fruit, and never drink commercial fruit juice. Of the thousands of oranges that go into the batch of orange juice you drink, one is sure to be moldy, and that is all it takes to give your liver a setback.

A heavy dose of vitamin C helps the liver recover quickly. It also helps get rid of aflatoxin before it is consumed, right in the food container. So keep a plastic shaker of vitamin C powder handy and use it like salt on all your food.

There are thirteen other mycotoxins I have searched for in our foods. They are described in the section on moldy food (page 381).

Physical Toxins

Breathing in dust is quite bad for you so your body rejects it by sneezing, coughing, spitting up and out. Imagine breathing in broken glass particles. They cut into the lungs in a thousand places and couldn't be coughed up. They would travel. Imagine swallowing a needle or open pin. If the tip was blunt it could move through the intestine. But because it is sharp it gets caught in your tissue, then works its way deeper and deeper.

Would we ever knowingly breathe in broken glass? We are justifiably afraid of it in our food or under our bare feet. We are unaware that it fills our homes when **fiberglass** insulation is left imperfectly sealed off. Any hole made through the ceiling or wall, even if covered with cloth, lets swarms of broken glass bits into the house air. Air currents flow inward, into your living space. So all holes leading to the attic or insulated spaces must be sealed airtight. Of course, fiberglass should never be used in home construction, draperies, or around water heaters. The best advice is to have it all removed while you are away and then vacuum and dust.

Occasional exposures by house builders working outdoors does much less harm. Chronic exposure from a single small hole in the ceiling does a lot of harm, leading to cyst formation. And that cyst is a perfect place for parasites and bacteria to settle and multiply. When the intestinal fluke settles there it becomes malignant!

Cancer patients with *solid tumors* have either fiberglass or asbestos in them.

Asbestos is another tiny bit, sharp as glass, that moves through your body like a swordfish, impaling your cells until it, too, gets routed into a cyst.

We have been led to believe that we no longer have asbestos in our homes because we have outlawed the fireproofing materials it was used in. While that may be true, the source I find most often is all too prevalent: the clothes dryer belt. As it gets hot the belt releases a blast of asbestos particles that are forced through the seams of your dryer, and also openings in your exhaust hose, by the high pressure formed inside. It is now in your air.

Chemical Toxins

Chlorofluorocarbons (**CFCs**) or **freon** is the refrigerant in your air conditioner and refrigerator coils. CFCs are suspected of causing the ozone hole above the South Pole. All cancer sufferers test positive for CFCs in their cancerous organ! I have preliminary evidence that it is CFCs that attract other pollutants–fiberglass, metals, PCBs–to form a growing tumor instead of allowing their excretion. This would make it a "super carcinogen.". How could you detect CFCs leaking in your home? By the time your air conditioner or refrigerator needs recharging, you have been exposed for a long time. We desperately need an inexpensive, in-home test for this unsuspected killer.

Arsenic is used in pesticide. Why would we poison ourselves along with the cockroaches? Is it because we can't see it happening? Just as we couldn't see the fiberglass floating in the air? Our diligent scientists have studied the mechanism of arsenic poisoning in great detail. Then why are we allowed to put it on our lawns to be carried into our carpets via shoes?

Polychlorinated biphenyls (PCBs), oily compounds with wonderfully useful electrical properties, were originally used in transformers until their inability to break down into less toxic substances in our environment was spotlighted. Banned from use, I find them in most commercial soap and detergents! Is transformer oil being disposed of by selling it to soap makers?

Formaldehyde is used to cure foam. As a result, foam furniture, pillows and mattresses give off formaldehyde for about two years after manufacturing. If you sleep with your nose buried in a new foam pillow all night, you are risking major lung problems.

Every cleanser in your house probably has a toxic warning on its label. Every fluid your automobile uses is toxic. Every pesticide, herbicide and fertilizer you put on your lawn is probably toxic. Every paint, varnish, wax, lubricant, bleach and detergent will send you to the hospital if even a small amount is ingested. Why do we keep them around? See *Recipes* (page 513) for safe, old-fashioned, alternatives.

If you are ill even after zapping, it is toxins still at work. Getting rid of them is a major step toward being well.

I DON'T THINK WE SHOULD LET HUMANS INTO THE HOUSE. WE'LL GET ALL THEIR PARASITES.
YEAH, I JUST CAN'T GET RID OF THIS MANGE.

How We Really Get Sick

What if you invented a device that could search people for the presence of mycotoxins (extremely toxic substances made by food molds)? And what if you found that although many people had them, those who were sick with a cold always had at least one of them. Would you ask whether a sudden buildup of mycotoxins is what really lets colds develop? Why do some people in the same family get the cold while others do not?

- What if you found everyone with **cancer** had the human intestinal fluke in their liver, and no one else did?
- What if you found everyone with **diabetes** had the pancreatic fluke of cattle in their pancreas, and few others did?
- What if you found everyone with **environmental illness** tested positive for Fasciola (sheep liver fluke) in their liver?
- What if you found everyone with **asthma** tested positive for Ascaris in their lungs?

What if you always found every mysteriously ill person had some unsuspected parasite or pollutant?

The device is the Syncrometer^{TM}, and these “what ifs” are all true. They forced me to alter my entire outlook on what really causes some of our “incurable”, mysterious diseases.

We used to believe that diabetes was caused by over consumption of sugars, a cold by a virus you caught from somebody, cancer from carcinogen exposure, depression from poor parenting. This multicausal concept is what made the study of medicine so difficult that only a few could undertake it. And every year new syndromes are added to the list of human illnesses.

But these diagnoses are based on a description of what is happening at a particular place in your body. This is like calling a mosquito bite behind the ears by one name and a mosquito bite behind the knee by another name. If you never see the true cause, a mosquito at work, this system could be excused as somewhat sensible.

And, until now, the profession of medicine has made some sense. The new truths, however, make the old descriptive system obsolete. You can now find the true causes of all your illnesses. And you can find them yourself by building the electronic diagnostic circuit (page 457)!

Once you have seen a mosquito at work on your body you no longer need to go to the doctor for a red, itchy bump. You don't need to search for the correct diagnosis and an appropriate drug. **You put up screen doors and windows!**

Once you have seen how common house dust is implicated in the common cold you get rid of the house dust. Once you have seen the mold in your food facilitate the cold virus you throw out that moldy food. But only seeing is believing. Nothing is left to faith. The electronic resonance method described in this book will let you see all these things for yourself.

You are not a hapless pawn attacked by bacteria and viruses that dart at you from nowhere to make you ill. You are not at the mercy of diseases all around you, hoping, by chance, to escape, like a soldier hoping to come home from the war. Nature and your body make good sense.

There is no disease that can outwit you if you know enough about it. Not even Lou Gehrig's disease! Nor asthma or diabetes. Read how the people in the case histories made themselves well. Read why some people failed. You have an advantage they did not have. Their instructions were hard to carry out because they had to have faith in them. You don't. You can replace faith with your own hard headed observations by building the diagnostic circuit (Syncrometer). The great convincer is seeing it yourself. When you personally find the mold in your peanut butter, or *Shigella* in your cheese, you have the knowledge, not faith, that convinces and guides you.

All illness comes from two causes, PARASITES and POLLUTANTS.

Only two causes! This is what simplifies the picture to make it possible for you to cure yourself.

We have been taught that illness is largely our own fault. That it is due to “catching something”, not eating what we should, like roughage or vitamins, or not doing what we should, like dressing properly, exercising or going to bed on time. Somehow, it's our own fault. Either by doing something we shouldn't or not doing something we should. When absolutely no fault can be found we are told it's in our genes! True inherited diseases are extremely rare. Our genes have evolved over millions of years to produce healthy humans. Nor are genes that mutate during your lifetime at fault. Pollutants, which are known to be mutagens, are the real culprits.

Neither the parasites nor the pollutants in you are “your fault”. Notice that other people all around us are doing the same things or not doing the same things, and even share our genes and don't have our illness. The current concepts on disease causation blaming our actions and our genes are simply not logical.

But until now they appeared logical. Suppose 1000 people were bitten by a mosquito or flea, it would always be in a different place, and if you were using the location and effect system to diagnose the problem you could have a thousand diagnoses listed for them, including a defective gene. Could you be persuaded to accept a gene replacement for your red itchy bumps? The new gene might be for antihistamine production so the welts don't become large, red and hot, or lead to impetigo. If you were the research doctor, you might be tempted to alleviate a thousand persons' distress with a new gene. Fortunately, you are not and only need to solve your own problems. You can be more logical.

After you have found the parasite interlopers hiding in your body you can kill them electronically. And after you have identified the pollutants stuck in your organs you can stop eating them, breathing them or putting them on yourself. In response, your body will begin to heal, just as surely as a mosquito bite heals.

Heal from multiple sclerosis, emphysema, myasthenia? Yes!! Some healing will be swift. Some healing will be slow. Healing is not understood. It is much faster in young persons. But fast or slow you know it has begun. It will be an exciting adventure to watch yourself lose your symptoms and get stronger.

Self Health

The entire purpose of this book is to enable you to diagnose and treat yourself for any disease. You have three new approaches that make this wish a reality: the understanding that only pollution and parasites make you sick, the quick and inexpensive diagnostic circuit that lets you find which pollutants and parasites they are, and the zapper or herbal recipe that kills the parasites.

Wouldn't it be nice not to have to go to the doctor for your aches and pains? And not to be dependent on the doctor to diagnose and treat you?

Self health means keeping yourself healthy. Doing it **yourself**.

Suppose your doctor has already diagnosed you as having “Atypical Lateral Sclerosis” or “Shoulder-Hip Girdle Metastable Aplasia.” Could you re-diagnose this for yourself so as to treat and cure it? And be successful? Why not? You have already succeeded in many enterprises. You learned to talk, walk, read, get along with people. These skills took a few years to master. Learning to keep well is a new skill. It may take a few years too. After you have learned these skills you may pass them on to your children. And so a new gift is given to humanity, like the gift of music or the art of cooking.

How To Heal

Your body has been trying to rid itself of its parasites and pollutants all your life! It had its own ways. It made stones, it made mucus secretions, it made itself toxic dumpsites. These were good tactics but now of course, they are no longer necessary. Can you help your body get rid of these accumulations and sweep itself clean again?

Sweeping your liver clean is the most powerful way of helping your body to heal itself after the parasites are gone. There are thousands of bits of “trash” accumulated in the liver bile ducts. They will turn into stones (gallstones) if left in place.

The kidney, too, has made numerous small stones in its effort to keep your body clear of lead, cadmium, mercury and other impassable pollutants. You can assist the kidney to expel all these.

In days, not weeks or months, you can feel the healing effects of clearing gallstones and kidney stones from your body. But there are miles of bile ducts (50,000 ducts) in the liver; the herbal recipes that do this are used over and over, patiently, until all, the “trash” is removed. This can take several years.

So, although you can stop your disease very quickly from progressing, the healing process may not be complete for years.

Nevertheless, you are healthy again. This means your pains are either gone or greatly reduced. Your organs are functioning better. You have a new sense of well being. Your energy is up. Your desire to live and accomplish something is back.

Organs that have been damaged beyond the ability of our simple methods to reverse can be treated with the magic of modern surgery. Cataracts, bunions, old injuries are examples. Possibly, these too, are “stone” formations. But no recipe has been found to clear them up simply and in a useful time frame.

Killing parasites, removing pollutants and clearing gallstones and kidney stones from your body is a powerful combination of treatments. It is so powerful you can change yourself into a new person in half a year. And then go on improving for years more.

Should you stop taking your prescription medicine while you are treating yourself? NO. Wait until you have cured yourself of the condition that required the medicine. Reduce your medicine and eventually go off it. Will your doctor approve? Find one that will. Remember that the medicine is buying you the time to cure yourself, something to be grateful for.

The Road To Wellness

To review our new understanding of health vs. disease:

- We have only two problems: parasites and pollutants. Parasites are things that live on us, using up our food and giving us their wastes. Pollutants are toxic things in us making it difficult for our organs to do their work. These two things are responsible for all our other problems.
- Our bodies have been trying to rid us of these by making stones, making secretions, giving us swellings, inflammations and benign tumors. We develop deficiencies and disabilities.
- Finally, some permanent damage is done. Our hair turns gray, we develop cataracts, the spine bends, nerves and muscles die. We weaken.

Our strategy to undo all this will be a logical one.

First, we will kill all parasites, bacteria, viruses and fungi.

Second, we will remove the toxic molds, metals and chemicals in our foods and body products.

Third, we will clear away and wash away the stones, secretions and debris already formed, that hinder healing.

Fourth, we will use herbs and special food factors to hasten healing, being very careful to use pure products. (These act more quickly when given intravenously but the emphasis in this book will be on oral consumption.)

Finally, for repairs that are beyond our abilities, we will seek help from health and medical professionals.

What could be more exciting than finding the tremor is out of your arm or the pain is out of your shoulder? Won't it be admirable to correct your pulse and your high blood pressure, by yourself? What an adventure it could be to get rid of all your warts.

But getting rid of pain seems like a first priority to me, since pain often undermines our morale, our initiative to do things, even our interest in getting well!

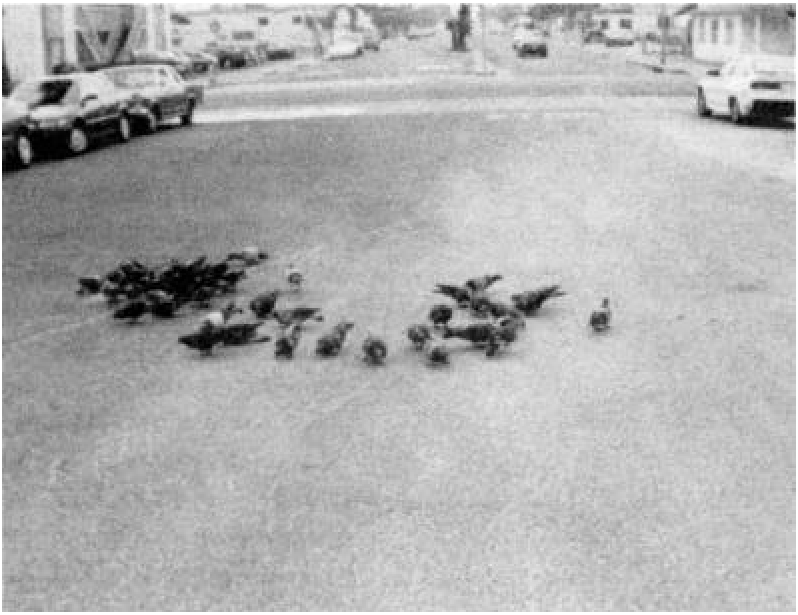

Many of our illnesses are caused by unsuspected sources. I have found pigeon tapeworm in humans, for instance.

Fig. 9 Feed animals out of your yard and away from where children play. Their wastes contain pathogens.

Pain From Toe To Head

I would rather die than endure excruciating, unrelievable pain. That puts **pain** at the very top of my priority list.

Fortunately for us, pain killers are at hand to get us through it and buy us the time it takes to solve the real problem behind it.

The pain killing industry also brought us the addiction industry. As we turn to electrical pain killing the need for addicting drugs should decline. There are other very useful pain killers: acupuncture, massage, listening to music, feedback devices, contemplation, hypnotism, and prayer.

But we will focus on getting rid of the cause of pain and healing the organs that are in pain so none of these methods are needed.

I am not talking about the pain of a broken bone, twisted ankle, bee sting or sunburn. I am not talking about the pain of a misaligned vertebra or stretch trauma in your leg muscles or arm muscles. I am referring to pain that is in one of your organs and refuses to go away. For example, arthritis. Other pains are headache, foot pain, elbow pain, hip pain, chest pain. All of these may have special names like rheumatoid arthritis, cluster headache, fibromyalgia, bursitis, tennis elbow and so on, but they are all the same phenomenon.

Knowing that parasites and pollutants are the real culprits, let us get right down to the job of finding out which they are, where they come from, and how to get rid of them.

The parasites that cause pain are not the large ones, like worms or amoebas. Nor are they the very tiny viruses. They are *bacteria*. Bacteria are the right size to get into the doorways of our cells. Our cells try to keep their doorways tight-shut but, of course, they have to open to let food in, or hormones, or other life-signals. If bacteria are swarming around the outside of cells, some will manage to slip into a doorway while it is open. A battle begins. The cells refuse to let the bacteria in. The bacteria refuse to let go of the door latch. There is probably a specific electrical attraction between them and an exact physical fit. The body's extra forces, the immune system, are called in to help quell the invasion. The bacteria multiply as fast as they can. There is swelling. There is heat produced. There is pressure against other organs. All together it is called inflammation and infection.

The answer is not to deinflame with drugs (like the cortisone variety). The answer is not to reduce swelling or body temperature. **The answer is certainly not pain killer.** The answer is to kill the bacteria. (Even this is not the ultimate answer. We must stop the source of the bacteria and your body's invitation to be invaded. We will get to this later.)

So **Step One** is to search for the bacteria nibbling at our painful regions and identify them. This gives us the clue to finding their source. **Step Two** is to exterminate them electrically. Within minutes they fall out of the doorways to our cells. Your white blood cells are waiting for them, and will gobble them up in a grand feast.

But, remember, there is another cause of pain, pollutants. **Step Three** is to find the pollutants and identify them because this gives us a clue as to their source. **Step Four** is to eliminate our pollution sources. **And the job is done. Pain is gone.**

An intriguing question will pop into your head as you search your organs for parasites and pollutants. Which came first? Pollutants can jam their way into your cell's doorways too. Does this allow the bacteria to swim in? Or do the bacteria come first, jamming open the doorways so the pollutants can enter? Both seem possible. Maybe both events go on simultaneously. Perhaps that is why bacteria and pollutants are always seen together. Viruses can land on your cell's doorways, too, and cause viral diseases, but they are not as often pain producing.

You will also notice something, as you keep testing and watching over your health. **Your body is very good at killing bacteria and viruses**. Your body kills them faster than a wave of a magic wand! The only ones that get away are those that are stuck in doorways and ‘channels with pollutants in them! This seems like evidence that pollutants do the gate-opening. But it isn't proof and we must keep all possibilities in mind.

Fortunately we do not have to know exactly how parasites and pollution make us sick in order to get well.

Searching For Bacteria

In order to find which organs have the bacteria and which bacteria are present you will need to learn the new technology that makes all of this possible. This technology is a simple electronic circuit that is capable of trapping frequencies in such a way that you can hear them. Your body's frequencies, the frequencies of bacteria, viruses and parasites are all different and can be heard as distinctly as a **mooo**, **baah**, **tweet**, or **oink** coming from a farm yard.

But do you have to do all that?

No! You don't. You could simply electrocute all these tiny invaders. But how would you know what to avoid in the future? If your pain returned how would you know if it was the same old bacteria or a new one?

Learning to test takes the guesswork out of diagnosis.

My hope is that you will find it all so intriguing, so absolutely fascinating, that in spite of some chronic pain, you will feel compelled to make the searches yourself. Find someone willing to help. Trade your information. It is less difficult than learning to use a computer.

What You Will Find

First we will study and cure pains of all kinds, starting with the toes and working our way up the body. We don't need to be very specific about the location of the pain since bacteria and toxins flow over to nearby locations anyway.

For each kind of pain, we will look into the causes so you can eliminate them. Pain could come back in an hour if you didn't know the bacteria were coming from cheese and you had a cheese sandwich after *zapping* (killing parasites electrically). In fact, the pain may not have had time to go away before the next onslaught begins and you might conclude, wrongly, that this method doesn't work.

Parasites might also come back to your specially painful place from a few far away places that are hard to reach by your zapper's electrical current. The inside of your eyeball, the testicle, the interior of gallstones, the middle of a tooth abscess or the bowel contents are such places. Your zapper current, because it is high frequency, prefers to "go around" these items, rather than through them. But with repeated zapping, and herbal parasite treatment, you can decimate them, too, and stop reinfecting the rest of your body.

In fact, it is such knowledge about reinfection and sources of our pollutants that is the most important contribution of the case histories.

After dealing with pains we will turn our attention to the diseases that aren't pain producing, like diabetes, myasthenia gravis and so on.

Toe Pain

The ends of your feet get the poorest "service" from your blood supply. They are the furthest away. The blood here has the most accumulated acid and the least oxygen supply. The body produces quite a bit of *uric acid* and this should, of course, be excreted into the bladder by the kidneys. But if the kidneys are doing a poor job of this, levels in the body and blood stream rise. The blood can only hold so much. It holds even less in acid conditions such as the ends of your feet experience. Uric acid begins to settle out or precipitate at our feet.

Hippuric acid, too, is found where pain is found. Hippuric acid is made in large amounts (about 1 gram/day) by the liver because it is a detoxification product. It makes no sense to consume *benzoic acid*, the common preservative, since this is what the body detoxifies into hippuric acid. Read all labels on food you buy. Don't buy any beverages or baked goods preserved with benzoic acid. *Citric acid* is fine. The joints of the big toes are favorite places for pain to develop. This is made even worse when circulation is poor. Take a *cayenne* capsule with each meal to improve circulation. If you cannot find your pulse just below your inner ankle your circulation is poor.

The accumulation of uric and hippuric acid invites bacteria to feed on them. As bacteria multiply the tissue fights back with

inflammation. Now the stage has been set for **pain**. Some people do not have pain although these acids and other deposits are present making their joints knobby and unbending. Bacteria have not found them yet or haven't multiplied sufficiently in them to invade your tissues. Perhaps there are other reasons as well.

If you have **toe pain** your course of action is this:

1. Kill all bacteria possible with your zapper. Repeat daily until no further benefit is seen.
2. Dissolve the deposits away. An herbal recipe, quite elaborate, is capable of doing this. It is called the Kidney Cleanse (page 549). Toe deposits are made of the same crystals as kidney stones, which is why the Kidney Cleanse works for toe pain. But because these deposits are far away from the kidney, it takes longer than merely cleaning up kidneys. It may take six months to make a significant dent on these deposits. This will at the same time remove kidney crystals so that these are no longer a source of bacteria.
3. Get teeth cavitations cleaned (cavitations are bone infections in the jaw where a tooth was pulled; it never healed; see Dental Cleanup page 409). This can “magically” stop the toe pain the same day as they are cleaned. The effect lasts for days afterward showing it is not the dental anesthetic that is responsible. It also teaches you that the bacteria in the toes can come from the teeth. But pain may return as other bacteria find the deposits.
4. Clean the liver of stones using the Liver Cleanse (page 552). This, too, can give immediate pain relief in the toes showing you they are a source for bacteria. Liver cleaning may take you two years to complete! Meanwhile your toe pain is receding. And, of course, this pays extra dividends in health for your body.

5. Reduce the acidity in your toes. First, check your acidity with pH paper meant for testing urinary pH, called Nitrazine™ paper. Ordinary pH paper, as for fish tanks, is almost as accurate and will serve as well. Tear a ½inch piece and hold it in the urine stream. Early morning is the time your urine is most acid. If this reads below 5.5 your feet must have been even lower in the night (lower is more acid).

Once deposits start, it is hard to stop them. If they start forming at 2 a.m., they are likely to continue for several hours even if the pH goes back up to normal. Your tactic will be to go to bed with an alkalizing action. Taking a calcium and magnesium supplement at bedtime, drinking milk at bedtime, using baking soda at bedtime are all remedies to be tried. They should raise your urinary pH to 6 in the morning.

Balance Your pH

Most persons with painful deposits anywhere in their feet have a morning urine pH of 4.5! At 4.5 it is safe to guess that a lot has precipitated again in the night. During the day, your body's pH swings back and forth. The urine gets quite alkaline right after a meal; this is called the *alkaline tide*. Three meals a day would bring you three alkaline tides. During these periods, lasting about an hour, you have an opportunity to dissolve some of your foot deposits. But if you allow your pH to drop too low in the night you put the deposits back again. The net effect decides whether your deposits grow or shrink.

To alkalinize yourself at bedtime, choose one of these options:

1. Two oyster shell tablets, equaling 750 mg. of calcium plus a magnesium oxide tablet, 300 mg (see *Sources*). The magnesium helps the calcium dissolve and stay in solu-

tion. Taking more calcium at one time is not advised because it cannot be dissolved and absorbed anyway and might constipate you. For the elderly only one calcium tablet is advised. Take calcium tablets with vitamin C or lemon water to help dissolve (¼ tsp. vitamin C powder; adding honey is fine).

2. One cup of sterilized milk or buttermilk, drunk hot or cold, plus 1 magnesium oxide tablet, 300 mg. (adding cinnamon is fine).

   If these two remedies work for you, your morning urinary pH will come up to 6.0 but if for some reason they don't, you need to take more drastic measures. Take the supplements and milk earlier in the day and reserve bedtime for:

3. ½tsp. baking soda in water. This is *sodium bicarbonate*. But don't use baking soda from a store because most brands I have tested are polluted with benzene! (See *Sources* for safe baking soda.) Using a combination of sodium and potassium bicarbonate in a ratio of 2:1 is actually a much more healthful potion. You can make your own or ask a pharmacist to make it for you. Mix two parts baking soda and one part potassium bicarbonate (see *Sources*) in a jar. Keep tightly closed. Label it **sodium potassium bicarbonate alkalizer** (this potion is also very useful in allergic reactions of all kinds). Take 1 level tsp. in water at bedtime. If your pH reaches 6 in the morning continue each night at this dose. If it does not, take 1½tsp. Keep watching your pH, since it will gradually normalize and you will require less and less. If you are using plain baking soda, instead of the mixture, watch your pH each morning, also, so you can cut back when the pH goes higher than 6.

Persons with a limit on their daily sodium intake must carefully count the grams of baking soda consumed in this way.

Each tsp. weighs about 2 grams, of which half (1 gram or 1000 milligrams is sodium. The sodium/potassium mixture would only give you half as much sodium (½gram per tsp.). By comparison, the usual daily intake of sodium is about 5 gm., although salt eaters consume twice that amount.

You have done five things to pull the rug out from under the bacteria living in and around the deposits in your toes. Now when you kill bacteria with your zapper, you can expect the pain to go away and stay away.

Locations at the base of toes may be painful due to a *neuroma*. Deposits and bacteria here are even more painful because this is the location of nerve centers. If the build-up is large, you may prefer some surgical help or a cortisone shot rather than wait several years for solid relief.

Foot Pain

This kind of pain does not involve as much deposits as toe pain and is therefore easier to clear up. But trauma to the foot is more important. First, even though your shoes are comfortable, change them. Get wider shoes, longer shoes, lower heels. Alternate two pairs of shoes in a single day. In your home take shoes off.

Keep feet very warm. Wear natural fiber socks, not synthetics. If your circulation is poor, take a *cayenne* capsule with each meal. When circulation is moderately poor, your feet get cold easily. When circulation is very poor, the heart pulse cannot be felt in your feet (take your pulse just below your inner ankle). Again:

- Get teeth cavitations cleaned (Dental Cleanup, page 409).
- Do the herbal Kidney Cleanse (page 549).
- Kill parasites and bacteria with a zapper.

- Then clean the liver (page 552).
- Check body pH in the morning upon rising using Nitrazine™ or other pH paper. Correct it as discussed in Toe Pain. If the urinary pH is 7 or higher, it means you have a bladder infection. Treat it immediately (page 101).

If there is any swelling around the foot or ankle, you are "holding water." It may be called *edema*. This is due to poor adrenal and kidney function. The adrenals are located on top of the kidneys and together they regulate how much salt and water stays in your body. Because they are situated so close together, they share their parasites and pollution. When the kidneys form *kidney crystals* the flow through the kidney tubes is hindered, and less water and salt can leave the body. It stays in your tissues as edema.

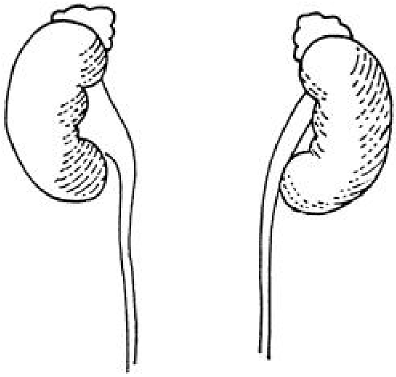

Fig. 10 Your kidneys with the adrenal glands sitting like hats on top. The ureters lead to the bladder.

The kidney herb recipe will dissolve the crystals. But you must remove toxins such as metal from tooth fillings and kill parasites. Continue doing the Kidney Cleanse after the original six weeks is up until all foot pain and edema are gone. Continue killing your tiny invaders with a zapper twice a week. You may need to cleanse the liver several times, too, before all the pain and edema are gone.

The supplement, *pantothenic acid* (see *Sources*), is particularly good for feet. Take 500 mg three times a day for several weeks to see if it helps. Massage and reflexology are also good for them.

Heel Pain

Sometimes heel pain is due to heel spurs, sometimes it is not. Heel spurs are due to deposits. The usual heel deposits are *uric acid* and various *phosphates*. Uric acid deposits become a breeding ground for bacteria. The phosphates give the deposits a rigid structure that is hard to dissolve. You may have to choose a pain killer, get specially built "orthopedic" shoes, or stop your daily walks to get relief from the piercing pains. These will not cure the problem but may "buy you some time" while you make basic changes in your lifestyle.

Stop drinking tea and cocoa because they contain *oxalic acid*. Stop drinking coffee, decafs, fruit juice and soda pop because they are contaminated with solvents. All of these must be detoxified by your body and eliminated by your kidneys. We should spare the kidneys these extra tasks when we wish them to clean up heel spur deposits. You must first dissolve the deposits, then help the kidneys eliminate them.

To dissolve them, you alkalinize your body's pH, watching over it carefully with pH paper or Nitrazine paper. To help kidneys eliminate them, use the Kidney Cleanse and drink lots of water. Only water, not a beverage, helps the kidneys to eliminate.

Drink a pint of water upon rising in the morning, and a pint of water between meals.

Do not buy "special" water. Use your cold tap water only. If you don't like the taste of your own tap water, get it from somebody else's tap. Use only glass or reused polyethylene jugs. Polyethylene is opaque, not clear. It has already lost its plasticizer into the water that stood in it earlier. (Somebody drank it!) Purchased water has traces of solvents from machinery and sterilizing equipment used in its bottling. Your own tap water is not pure (indeed it may have 500 toxic elements), but it never contains solvents in amounts I can detect.

Traces of solvents are worse than traces of other toxins.

Water filters are not the solution. They trap the pollutants and then allow a tiny amount to enter the water on a daily basis. Chronic toxin consumption is much worse for your health than periodic surges of toxins. You can use a small filter, that gets changed every month without much risk. This would dechlorinate the water, at least, and improve the flavor. The pitcher variety (it should be made of hard, inflexible plastic) and the faucet variety are listed in *Sources*.

Bottled water is popular,

and tasty, and has appealing advertising, but it is just not safe.

Why is it easier for everyone to spend dollars per day, for the rest of their life, buying water instead of insisting that their water pipes are metal-free?

Another reason not to drink water from bottles, however convenient, is that it is stagnant and is soon contaminated with our own bacteria from contact with mouth or hands. *Staphylococcus (Staph)* and *E. coli* are commonly seen. The solution is not to add still more chemical disinfectants, the solution is to drink from a flowing source, such as our faucets. If you must carry water, use glass containers; plastic is porous and much more difficult to wash clean (sterilizing in a dishwasher is effective, though).

By drinking a total of four pints of water in a day, the kidneys will notice the assistance. (Any single herb tea without added salt or sweetener counts as water.) The urine will stop having any odor and will become very light colored. This is especially important while you are dissolving the heel deposits since your body is now carrying these in the circulation.

Killing bacteria with a zapper may give you instant pain relief and is, of course, beneficial to your body. But you can't expect the pain to stay away until the deposits are gone and the source of bacteria is also removed. Go after the usual sources: teeth and stones. Another source is dairy food contaminated with *Salmonella* and *Shigella* bacteria. Sterilize all of it by boiling. Even the amount put on cereal in the morning or used in scrambled eggs is enough to reinfect you! Butter and heavy whipping cream also need boiling. Stop eating yogurt and cheese which can't be boiled.

Phosphate deposits are a mixture of three phosphates: *monocalcium phosphate* ($CaH_2\ PO_4$), *dicalcium phosphate* ($Ca_2H\ PO_4$) and *tricalcium phosphate* ($Ca_3\ PO_4$). They are formed by eating too much (way too much) phosphate relative to calcium.

Our high phosphate foods are meats, carbonated beverages and grain products like rice, cereals, breads, pastas and nuts.

The body's normal elimination tactic for phosphate is to combine it with calcium and magnesium in order to neutralize it first. Unneutralized phosphate is very acidic and would burn the delicate kidneys. But where shall the calcium and magnesium be taken from? Magnesium is often in very short supply since it comes from green vegetables in the diet and is not stored up in any special organ. So it falls on calcium to be used for this purpose since it is stored up (in your bones and teeth).

The acid condition created by phosphates dissolves your bones and allows large amounts of bone structure to be washed out with the urine: this can be directly seen with an analysis of urine (you can ask your doctor to order this at a lab). If you catch all the urine in a 24 hour period you can measure all the calcium you have wasted. You should not lose more than 150 mg calcium in a day because this is all you can absorb in a day![4] If you do lose more than 150 mg in a day, you are dissolving your bones at a fast clip. This also means there is too much calcium in your blood and lymph, from dissolving so much bone so quickly.

Once you have dissolved your bones it is not so easy to put the calcium back into them. Your body will try to put it back as soon as possible—as soon as your acid condition is gone. But your bones can't do this without *vitamin D*. Vitamin D may "come from" sunshine and from vitamin D in milk but it isn't that simple. Vitamin D must be activated by your kidneys before it can go to work! Remember, though, it was the kidneys that had a problem in the first place, allowing deposits to form! With old kidneys, clogged with crystals, hampered by heavy metal and mold toxins, and beleaguered by bacteria and parasites, is it any wonder that sunshine and vitamin D fortified milk don't supply large amounts of activated vitamin D? It takes large amounts to put back into your bones the large amount of calcium that dissolved out during the acid state you put yourself in by over consuming phosphate food.

If you can't put the calcium back into your bones promptly where is it to go? It may attach itself to tissues that were never meant to be used this way. Your arteries fill with "scale," your kidneys form calcium phosphate crystals, heels form spurs, joints become knobby with deposits.

---

[4] You absorb 5 to 10% of what you eat. If you eat 1 gram (1000 mg) you absorb 50 to 100 mg. But you absorb 25-40% if it comes from milk!

Young persons and children, with healthy unclogged kidneys, make–that is, activate–ample vitamin D, so even if they consume too much phosphate and develop an acid condition that dissolves their teeth and bones, **they can put the dissolved calcium back in its proper place.** They don't develop hardened arteries, heel spurs, and knobby joints, at least not at first. Their bodies can take a considerable amount of abuse without showing it.

In general, people eat way too much phosphate. Meat eaters eat too much meat. Vegetarians eat too much grain. Most everyone drinks phosphated beverages. In this way we set the stage for hardened arteries, joint disease, calcified tissues that no longer have flexibility. We all get kidney crystals that become stones. This is *aging*. All these deposits invite bacteria to live in them and on them, creating pain.

Old age and pain go together as if they were true partners. Yet it is just the result of bad food choices.

Try to undo as much of this false aging as possible.

1. Reduce your meat consumption. Switch to fish which supplies calcium in the tiny bones. It is true, these bones are made of calcium phosphate and one might expect, logically, to be getting a less effective calcium source. Logic isn't necessarily biologically correct. The bones of fish work nicely as a calcium source and their phosphate content is not too great. Such a diet has worked for many primitive societies. Further, I have never seen a case of mercury toxicity from eating fish; amalgam tooth fillings are our truly significant source.

2. Reduce your grain consumption. Instead of cereal for breakfast, add fruit and reduce the cereal to half. Cut bread consumption in half. Cut pasta consumption in half. Cut rice and corn consumption in half. Eat more bananas and other fruits. Eat more vegetables; always choose potato (not potato chips) instead of rice or macaroni. Always choose a leafy salad instead of pasta salad. At any restaurant or salad bar ask yourself: is this wheat, rice or corn? If so, choose potatoes or other vegetables. Choose coleslaw. Choose mixed fruit. You don't need to go off the grains, only reduce them to improve your condition.
3. Take vitamin D as a supplement. Nothing less than 40,000 units has any real impact by the time there are problems. This strength is available by prescription only (usually 50,000 units, which is close enough). To avoid getting a polluted product, ask your pharmacist to follow the recipe on page 560. (In the past some cases of poisoning by overdosing resulted in this regulation. If you overdose you will get joint and muscle pain and nausea but it is reversible.) Take one a day (not more), for the first three weeks, then two a week forever after.
4. Finally, toss the carbonated beverages right out of your diet or make your own (see *Recipes*). It is not the carbonation that is harmful, it is the added phosphate. Drink water, herb teas, homemade fruit and vegetable juices, milk or buttermilk.

Milk, like fish, is full of calcium in the form of calcium phosphate. Again, logic might speak against the effectiveness of this form of calcium. Again, logic is wrong. Evidently, the calcium and phosphate story must be much more complex than I am depicting here.

Milk works best as a calcium source, in spite of its phosphate content. Possibly the *lactose* and other complexities of its composition contribute to this. My recommendation when deposits have formed anywhere in the body, such as heel, toe, arteries, joints, is to switch to milk as a beverage. Compare the calcium level of your urine before and after the switch (allowing several weeks first). Also compare calcium levels in your blood serum. It should move towards normal. This means **up** if it is too low (below 9.2). It means **down** if it is too high (10). If you are monitoring the effectiveness of the kidney herb recipe in dissolving away your phosphate crystals, notice that drinking milk keeps them from reforming. Taking calcium tablets does not! Taking nothing lets them reform the quickest.

The milk must be 2% or higher in butter fat to be effective.

With your body fluids at their proper acid level, with your kidneys able to flush out acids, with heavy metal toxins no longer settling in, with your bone-dissolving stopped, your heel deposits can shrink. Bacteria have no place to feed and breed. You can kill them several times more with your zapper to catch stragglers. And your heel pain becomes history. Be careful not to bruise the sensitive tissue with too much walking or running immediately after the pain is gone. Wear cushioning socks and well cushioned shoes.

Names in the case histories have been changed to ones of the same sex, picked at random from a telephone directory. Other facts may have been altered in non-essential ways.

Walter Jones, a man of 67, was diabetic for 14 years. His feet and upper legs hurt so much for the past 13 years he could barely shuffle along now. They were also cold and clammy. The herb, juniper berry was added to the Kidney Cleanse recipe to make it even more effective for him and he was advised to stop smoking, using alcohol, and caffeine. He also started the parasite killing program (at that time an herbal parasite program was being used instead of a zapper) and a liver cleanse that yielded several hundred small gallstones. Three months from his first visit his legs were better and feet OK, although he still had some trouble walking due to pain in upper thighs.

Dinah Sagun, a minister's wife, had a lot of walking and standing to do besides traveling a lot. She had knee surgery 15 years ago to remove deposits but now they were getting bad again. Also her heels were sore. My tests showed she had phosphate and oxalate deposits in her kidneys. She was started on the Kidney Cleanse followed by the parasite program. She stopped using benzene polluted items, especially her Chap Stick™. Her diet was changed to include milk and leave out store-bought beverages. In 4½ months her heel pain was gone along with the bad headaches we had not even started to work on.

Julie Fernandez came with a list of 10 serious problems including foot pain and ankle swelling. Her urinalysis showed crystals. My test showed they were uric acid crystals. Her parasites included *Trichinella*. There was a house cat. Her toxic metal test showed tellurium and platinum accumulation from metal dentalware and arsenic from pesticide. She was to start on a Kidney Cleanse and add the parasite program two weeks later. She was to remove all roach hives and other pest killers from her house and arrange for dental metal replacement. At her third visit, four weeks later, she could walk without pain although she still had *Trichinella* and arsenic problems. Two months later she was aching all over again and her kidneys hurt. She still had *Trichinella* and some dental work to do. She was to do a 5-day high dose parasite treatment and treat the cat also and repeat the kidney cleanse. This relieved her pain.

Dorothy Shelley had numerous pains including foot pain in the arch, lower back pain and cramps with her period. She had oxalate and uric acid crystals in the kidneys so was started on kidney herbs. Three weeks later her foot pain and low back pain were gone.

Paul Longtin had toe cramps, calf cramps, and heel pain. After cutting down on smoking, quitting caffeine and soda pop and taking niacin (250 mg time release; 1 capsule each meal and upon rising in morning) and doing a kidney cleanse, all in three weeks, he was much improved and didn't even need his arch supports.

Juan Onley, age 58, came for his gout in both feet and one hand. He could hardly walk, in spite of soft shoes and pain killer. It started 7 years ago. He also had prostate pain and couldn't sit comfortably. My tests showed his kidneys were full of uric acid, oxalate and cystine stones. His prostate was full of *Gardnerella* and *Campylobacter* bacteria. His wisdom teeth were harboring *plantar wart #4 virus* and *Coxsackie B4* virus. *Gardnerella* often comes with fluke parasites and indeed he had intestinal flukes in his kidney! He began with the kidney herb program, then the parasite killing program. It took five months to clear them all. Then he could walk and sit without pain.

Leg Pain In Children

It is commonly accepted as normal(!) in children to have pain in the shins or calves of the leg. They may even be called "growing pains." Children may cry with the pain and never tell anyone the reason for crying. It happens mostly after napping. This may be caused by cramping of the leg or spasms of the blood vessels. *Lead toxicity* is a common cause of both. Test for the presence of both **lead** and **cadmium** in the tap water. Only your own electronic tests are helpful. Water department tests are much too crude.

If either poison is

ULTRAS
ULTRA
Pampers
Pampers
Pampers
Pampers
Pampers
HUGGIES
HUGGIES

Fig. 11 All disposable diapers I tested had mercury and thallium! These toxins can be absorbed through babies' skin.

found, test the water supply from each faucet in the house, in the morning, before it has been run. Find the offending sources, change the water pipes to polyvinyl chloride (PVC).

Also search for thallium or mercury in the child's saliva. If it is there, remove all dental metal. Stop using all commercial disposable diapers, dental floss, cotton swabs and bandages; they are polluted with mercury and thallium probably from manufacturing them in foreign countries where it is legal to sterilize with mercuric chloride. Test again, several times, after plumbing or dental work has been completed. To relieve pain:

- immerse legs in warm water
- massage legs gently
- give 25-50 mg niacin, not time-release, to dilate blood vessels.

Leg Pain In Adults

Leg pain in adults is usually associated with *cadmium* or *thallium*. Cadmium is present in tap water that runs through corroded galvanized pipes. The cadmium is probably a contaminant of the zinc used for galvanizing. Test the water, electronically, for cadmium. If you have all copper pipes but there is cadmium in the water, there must be a short piece (a **Y** or a **T** joint) made of old galvanized pipe lurking somewhere. Track it down by testing water from all your faucets.

Cadmium causes the blood vessels to spasm and it is made worse by smoking, that's why the condition is sometimes called **Smoker's Leg**. But extremely painful legs are due to chronic thallium poisoning more than any other cause!

It is very important to know exactly how toxic thallium is. Read the clipping on page 417 right now!

Where would you ever get thallium? From your very own mouth! The mercury in fillings is often itself polluted with thallium! Replace your amalgam fillings with composite.

Bandages

Sanitary napkins and tampons

Cotton swabs

limpiecitos
HAS Hg, Tl
limpiecitos

Cotton balls

ACE
First Aid
First Aid
First Aid

Floss

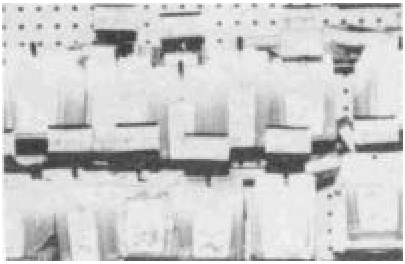

Toothpicks (the one on the right is ten years old and had no mercury or thallium)

TOOTH PICKS
Diamond
TOOTHPICKS
Forster
TOOTH PICKS

Fig. 12 Do not use any commercial personal products, the risk of pollution is too great.

Thallium has another source: it is riding along as a pollutant in cotton swabs, cotton balls, commercial bandages, toothpicks, floss, gauze, sanitary napkins, tampons, disposable diapers, and paper towels. Evidently these are being sterilized with mercuric chloride which, in turn, has thallium pollution.

- Line **disposable diapers** with a tissue
- Line **sanitary napkins** and pads with a tissue
- Use the **polyester puff** in the top of vitamin bottles in place of cotton balls. Twirl some around a plastic stirrer for a swab.
- Use pieces of **tissue and masking tape** for bandages.

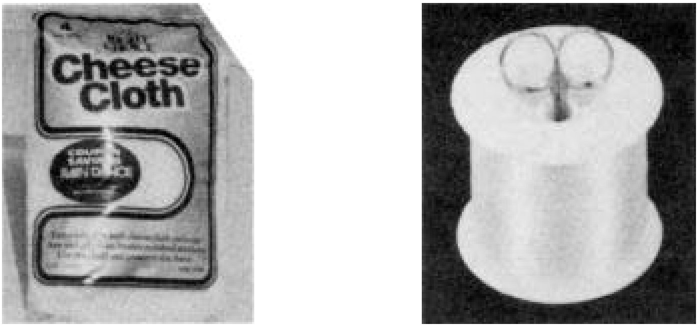

Use cheesecloth in place of gauze.

Use monofilament fish line (2 or 4 lb. test) for floss

Fig. 13 Safe substitutes for personal products.

If you do have thallium in your white blood cells and you haven't used toothpicks, etc. earlier in the day, then it is in your tooth fillings and **you have no higher priority than getting the**

amalgams out. Find a dentist immediately who will remove them, drilling deeply and widely not to miss a speck of it, thereby getting the thallium out, too. You cannot cure your leg pains without removing thallium.

Leg Pain Protocol

1. Stop smoking.
2. Repair plumbing.
3. Do dental cleanup and chelate out the mercury and thallium that has gotten into your tissues with EDTA (see *Sources*). You will need to find a chelating doctor; ask a friendly chiropractor to help you locate one. Or at least take thioctic acid 100 mg, (2 three times a day) and vitamin C (5 gm or one teaspoon) daily for a month.
4. Take magnesium oxide 300 mg (take 1 twice a day).
5. Take niacin, as much as you tolerate—time release varieties are less effective. Try 50 mg with each meal.
6. Change your diet to reduce phosphate and include milk (sterilized).
7. Do the herbal Kidney Cleanse (page 549) followed by a Liver Cleanse (page 552).
8. Zap yourself on alternate days at bedtime. If this zapping makes no difference whatever, your problem is purely spasms. But if you get relief, even if it's very short lived, you must have killed something. Bacteria must come from somewhere. Concentrate your efforts on dental health and better diet.

Jean Booth, age 30, had sore, tired legs and severe three-day headaches. She would get stabbing pains in back of her thighs. After we found thallium and mercury in her kidneys she did a Kidney Cleanse and got all her metal tooth fillings replaced. She then felt fine for one year. Suddenly she got fatigue and heavy legs again with stabbing pain at the outer thigh. She had seen a neurologist. Her chiropractor suggested it was leftover mercury so she came back to us. Indeed, she was toxic with lead, mercury, thallium, but her dentist could not find the leftover metal in her mouth. Three cavitations were cleaned; she was put on thioctic acid; eight varieties of bacteria and viruses were killed with a frequency generator and her legs became well again. Our test showed thallium at 4 teeth, but it was not a big enough deposit to show up on dental X-ray. She may eventually choose to have these redrilled.

Charlie Snelling was a picture of pain: pain in arms, elbows, shoulders, wrist, hands, chest, low back, legs, knees, and feet. He was started on kidney herbs and a few pains were reduced. He was toxic with cadmium so he changed his plumbing to plastic. However, he continued to be toxic with cadmium and thallium throwing suspicion on his numerous old tooth fillings. He got them all replaced and cavitations cleaned. He used our frequency generator to kill *beta Streptococcus*, *Pseudomonas*, *Troglodytella* and *Staphylococcus aureus* all of which were under one tooth (#15). A year later he still had bouts of leg pain. He still had numerous bacteria under his teeth because the jaw bone was not healing. He had not been taking vitamin D, nor magnesium nor drinking milk for the necessary calcium. But he had improved enough to go back to work full time.

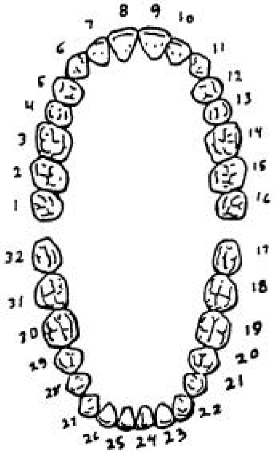

Fig. 14 Tooth numbering system.

Victor Abhay, age 16, could no longer play in high school sports because of knee pain. It began “with a virus” and high temperature two years ago. His knee started to bother him after that. He had *cysteine* kidney crystals and four parasites: *Cryptocotyl*, human liver fluke, *Echinococcus granulosus* cyst and *Echinostomum revolutum* in his white blood cells. He was to start the kidney herb recipe and follow this with the parasite program. Five months later, when we next saw him, his knee was fine.

Kim Murphy, 45, had painful legs, feet and knees. They also were swollen and itched. She was parasitized by *Trichuris*, (dog whipworm). She also had tapeworm stages (*Taenia pisiformis*) and intestinal fluke in the intestine. She started on the kidney herbs, followed by the parasite program. She stopped using zirconium-containing products (deodorant) and barium (lipstick). She stayed out of bus exhaust. (She knew she was allergic to diesel exhaust.) In one month her leg pain and itching were gone; slight swelling remained.

Nancy Tong, 80's, had edematous legs. In fact she could get no stockings on. They appeared like pillars with no taper at all. She was on diuretic medicine from her doctor. It kept her blood pressure down but she was losing ground with water excretion. This raised her general toxicity (blood BUN[5]) which made her feel bad most of the time. Yet she drank enough water, curtailed her salt, used no caffeine and had no really bad habits. She had to wear several pads for incontinence. We found she was toxic with cadmium and lead, which were probably responsible for her huge accumulation of kidney stones. The metals were in her tap water and she was unable to resolve this problem since she lived in a senior citizen center. We advised her to move, or to have her tap water carried in, but she could do none of these. Although the situation was hopeless, she did the kidney cleanse, parasite killing program and changed her metal rimmed glasses and wrist watch to plastic. She gained enough ground from these improvements to be able to wear elastic hose and thereby give some physical assistance to her body. This encouraged her to do a liver cleanse. She had a headache with the cleanse but immediately afterwards she fit into a smaller size “Keds” (elasticized stockings). A half a year from the beginning, her legs had taper to them; she didn't even mind wearing a dress.

Fibromyositis and Fibromyalgia

When pain is widespread, not just in joints or legs but in many muscles and soft tissues of your body your doctor may call it *fibromyositis* or *fibromyalgia*.

---

[5] BUN stands for blood urea nitrogen. It is a body waste and is normally kept low by the kidneys.

For bacteria to get all over your body, they must be riding along with parasites that get all over your body. Microscopically tiny roundworms can do this.

Trichinella is the most common cause of these diseases, but sometimes *Ascaris* larvae or *hookworms* or *strongyle* larvae are the main culprits. These wormlets bring hosts of **bacteria** with them, mainly *"Streps"* (*Streptococcus* varieties) and *"Staphs"* (*Staphylococcus* varieties), but also *"Clostridiums"* (*Clostridium* varieties) and "Campyls" (*Campylobacter* varieties). The bacteria are probably the pain causers. By killing all bacteria—*Staphs*, *Streps*, *Clostridiums* and *Campyls*—using a zapper, you may get relief for one hour!

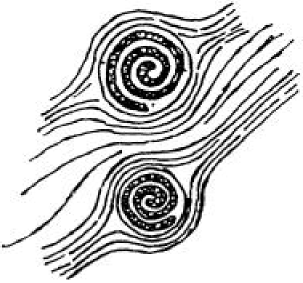

Fig. 15 Trichinella larvae settled in muscles.

By killing *Trichinella* and *Ancylostomas* (worms) first, followed by the bacteria, you may get relief for several hours. By killing the parasites and bacteria in every household member and the pets at the same time and by never putting your fingers to your mouth, you can expect permanent pain relief. It is interesting to speculate why the other family members, who are also infected with these tiny roundworms don't develop *fibromyalgia*. Perhaps the larvae stay in the intestine or go to the diaphragm (causing coughing) or the eyes (causing "lazy" eye muscles). Perhaps they merely cause **anemia**. *Trichinella*, hookworms and *strongyles* are extremely difficult to get rid of in a family.

These roundworm larvae undoubtedly cross the placenta into the unborn child during pregnancy, too. So they can be "inherited." Try to clear up the whole family before the next pregnancy.

Clearing up pets of these parasites is even harder. The best advice is to give your pets away. They will continue to harbor them even though they are on a pet parasite program. It is impossible to stay free of the parasites your pets have: they will move to your soft tissues immediately, giving you the bacteria and inflammation again.

The next most important advice is to keep fingers out of your mouth (read Hands, page 397). None of these parasites enter through your skin (this is in spite of teachings that hook-worms enter this way), you must put them into your mouth somehow! Consider your mouth off limits to anything but food and kissing.

Finally, if there is a baby in diapers in the family, be patient. When diapering days are over you will have less bowel contact, giving you an opportunity to finish your own treatment. Meanwhile, wash fingernails in Skin Sanitizer (see *Recipes*) after cleaning up children's bowel movements, diapers and yourself. Use borax liquid for soap to leave an antibacterial residue on the skin.

Try to identify your parasites before killing them so you can be on the lookout for them in the future. Get slides or dead cultures of various pathogens and search in your white blood cells. If you can't do this, at least save a saliva sample of your own; keep it frozen or preserved. Also make a saliva sample for your pet. This gives you specimens to test yourself for later. You won't know which parasite is in this saliva specimen, but if you ever test positive for it again you will know you got it back. Then zap yourself.

Brenda Byrd was diagnosed with myofibrositis two years earlier at the age of 36. Her blood test showed triglycerides slightly high (152 mg./DL), indicating the beginnings of urinary tract problems. Her urinalysis stated “hazy” (hazy with bacteria or crystals) instead of clear urine. It also listed white blood cells, red blood cells, and a few bacteria present in her urine. Our tests revealed mercury and numerous other heavy metals distributed in her thyroid, stomach, kidney, lungs, bones and bone marrow. She was also full of beryllium (usually from “coal oil”) contained in the hurricane lamps she kept in every room. She had numerous parasites, including *Strongyloides* and hookworms spread through her body tissues. She was thrilled to learn how to get her health back and started with the dental problem.

Marcia Cochran, 36, had muscles twitching all over her body. It was diagnosed as fibromyalgia. Her joints were tender and her chest felt “tight.” She had bronchitis twice a year. She was on Amytryptaline™ for muscle twitching and Bentyl™ to calm the intestine (spastic colon). She had depression with it and was on medicine for that. It all started with fever and chills that she thought was the flu but after they went away, she was left with a tremor. Sometimes she felt that little electric shocks were going through her. She had *Ascaris* and hookworm larvae widespread in her body. She had sheep liver flukes in her liver. She was started on the parasite program. She was also toxic with PVC and tellurium (dental metal). The PVC was traced to plastic storm windows applied to the inside of the window and to new shower curtains. Two months passed and she had not solved any of her problems. Then she did her first liver cleanse and got over 100 stones out. This instantly reduced her fibromyalgia to occasional attacks. She was so encouraged she decided to go ahead with dental cleanup.

Joint Pain or Arthritis

Two main kinds of arthritis are recognized clinically, *osteoarthritis* and *rheumatoid arthritis*. In osteoarthritis the joints have bacteria living on the deposits left there. In rheumatoid arthritis the bacteria come from larger parasites—wormlets actually living in these joints. The worms are the common little roundworms whose eggs hatch into microscopic wormlets that travel.

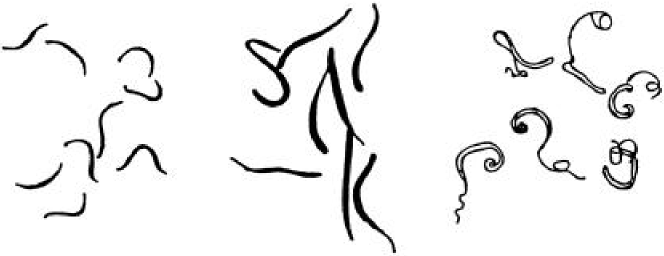

Fig. 16 Hookworms, strongyles, and whipworms.

We have four common roundworms: *Ascaris, hookworm, Strongyloides, Trichinella*. Their life cycle normally directs them to travel to the lungs but in some people they travel through the entire body, including brain, muscles and joints. More research is necessary to explain this. My suspicion is that there are toxins, like **mercury**, **thallium**, **cadmium**, **lead**, as well as **solvents**, distributed through the body, lowering immunity and allowing the tiny larvae to reside there. Once the pathway (routing) to these organs has been established, it continues to be used by other parasites as well. Soon a variety of parasites, their bacteria and viruses, and pollutants are all headed toward these organs.

Osteo or Common Arthritis

When joints are painful it is a simple matter to kill the bacteria with an electronic zapper. Treat yourself with a zapper daily until the pain is gone. Maybe it will stay away, but chances are the bacteria have a steady source.

The most common source for *Staphs* and *Streps* are small abscesses in the jaw bone, under and beside old extractions, root canals and mercury fillings. You may get immediate pain relief just from a dental cleanup, and again disappointment may follow. *Staphs* and *Streps* are such ubiquitous bacteria, they may come not only from jaw bone infections but from gallstones, kidney stones and other parasites. The correct treatment for arthritis is a complete overhaul of body health: a diet cleanup, a body cleanup, and environmental cleanup.

Start with the herbal parasite program and zapping. Follow this with a kidney cleanse, then liver cleanse. If any toxin is overlooked, especially asbestos and fiberglass, it is sure to find your joints and permit bacteria to return and cause pain. Make sure to correct your body acid levels after doing pH measurements of the urine (page 57).

Arthritic deposits contain a large amount of phosphates combined with calcium. This calcium came from some other bone, such as the base of your spine or the wrist. Here the bones are getting weaker due to this calcium loss. Calcium was taken out of your bones for the simple purpose of neutralizing the excess phosphate in your diet. Reduce phosphate consumption (meats, soda pop, grains) by half, eating fish, milk, vegetables and fruit instead. Drink three cups of milk a day. If you are allergic to milk, do several liver cleanses, switch brands of milk, use milk digestant, and use it in cooking and baking. Cheese and cottage cheese are not substitutes for milk (the calcium stayed in the whey). Dairy products must be boiled before consuming and should be no less than 2% butter fat. It takes bile to make calcium absorbable, yet milk with less than 2% butterfat does not trigger the gallbladder to empty it's bile at mealtime.

If you are not used to dairy products, start slowly and work up gradually to the 3 cups a day needed.

Carol Lachance was diagnosed with arthritis of her back and knees. (She had spurs in both places.) She was positive for *Trichinella*, *Ascaris* and *Dirofilaria* (she also had pain over the heart). They had an outdoor dog. Her blood test showed a high phosphate and alkaline phosphatase level showing she was dissolving her bones. After changing her diet to include milk, extra oyster shell calcium (one a day), magnesium oxide and vitamin B_6, and reducing her meat and grain consumption her phosphate level went down to normal (below 4). She did the kidney cleanse and liver cleanse as well as parasite program but still had pain. A toxic element test revealed lead in her bones. It was traced to a drugstore variety multivitamin tablet she had taken daily for years. When she stopped these and added prescription vitamin D (50,000 units) for three weeks to help her bones heal she got relief.

Gail Hildebrand, age 62, had painful arthritic hands and gum disease. She had 12 parasites free-loading on her and was toxic with asbestos. This was traced to her clothes dryer. Four months later, after killing parasites, her hand pain and gums were much better. She had the dryer vent taped up tighter and this got rid of her asbestos problem. She started on kidney herbs and in one month saw that her enlarged knuckles were beginning to go down.

Norma Littrell, age 53, came in for her severe arthritis of six years. Her knees, shoulders and hands were painful. I explained to her that painful shoulders did not belong to the arthritis picture but had a gallstone etiology which she could easily fix in a single night at a later time (liver cleanse). She also had mid-back, upper back and lower back pain; again the upper back pain belonged to the liver problem. She had tricalcium phosphate kidney crystals as well as uric acid crystals. She was told to go off coffee (no decaf either), decrease her meat and grain consumption (phosphate) and increase milk and use stone ground cornmeal products (genuine stone ground tortillas as well as pickled pigs' feet are high in utilizable calcium). She started on the kidney herbs. In 33 days her low back pain had improved a lot, she could wash her own hair again and she could sit down and get up from her living room floor without pain. She was elated but we recommended less stress than such exercises for her joints.

TORTILLAS
Santa Fe
Santa Fe

Fig. 17 Stone ground corn tortillas, high in calcium.

Patricia Robinson, age 76, had pain in her knees, feet, lower back, hands and wrists. Also in her shoulders and upper back which is not part of the arthritis picture. She had swollen puffy eyelids which is a telltale sign of Ascaris. She had heartburn, cold feet and insomnia as well as high blood pressure. She was on several medicines. We started her on the kidney recipe and ornithine (four to eight as needed at bedtime) for sleep. We hoped she could stop her Ativan^{TM} drug soon. She was to go off coffee, tea and decafs. Her hands were knotted and misshapen at the joints, also tender. Two months later there was little change; her doctor had put her on Prozac^{TM} and she hoped that would do a lot for her (so she stopped the kidney herbs) but it didn't. The parasite test still showed *Ascaris* and she was started on the parasite program. We also found fluoride (from toothpaste), iridium, samarium and palladium, all from her tooth implants. She was to remove as much metal as the dentist could replace, clean cavitations, and take thioctic acid, 2 a day, to help clear metal from her body. In four weeks the sharp pain in her back was gone and in three more weeks the pain in her hands was gone.

Lynne Snyder, 72, had pain in every joint and had to be on pain medicine to keep moving. Her potassium level was very low (3.6—an adrenal/kidney problem), and she was started on kidney herbs, carrot juice, vegetable juice and bananas. In ten days she could feel some new energy but her pains were terrible, especially her knees. She was taken off tomato juice, cranberry juice, citrus, pepper (she was using a lot), and given buttermilk as a beverage which she enjoyed. After three weeks of kidney herbs she was started on parasite killing herbs. In another month her arthritis was much better. She was not on any pain medicine and could get to sleep without it. She thought it was mostly going off pepper that helped.

Rheumatoid Arthritis

When inflammation and swelling affect your joints, besides pain, it is called *rheumatoid arthritis*. In addition, a blood test may reveal “rheumatoid factor” to be present. In this case, the common tiny round worms have invaded your joints. These are *Ascaris*, *Ancylostomas*, *Strongyloides*, and *Trichinella*. Their eggs are everywhere around us, in dust and dirt and the filth under fingernails and our own bowel movements.

Superior sanitation is your first defense. Rinse fingernails in alcohol after cleaning up bowel movements or changing diapers. Never, never, tolerate long fingernails in any family member. If this discipline can't be enforced, do not allow food preparation by “long nails” unless gloves are worn. He or she may not be getting ill (yet) from the family parasites, but you are. Of course, you can kill them with a zapper (internally, not the ones under the nails) but that is after you have been infected. They are easily picked up again. It would be wise to zap for roundworms every week, just in case.

Make sure your pet is treated with parasite herbs or by zapper as well. A pet that goes outdoors will quickly (the very next day) bring these roundworms into the house again. Give away your pets if possible.

Check for dental problems. Do the Dental Clean-Up (page 409). Then do a Kidney Cleanse and Liver Cleanse.

You may relieve your pain and begin to heal immediately after zapping but it is wise to do all the health programs, anyway. Change your diet. Reduce phosphate, start using sterilized milk for calcium. Switch to fish from meats. Drink much more water. Use only harmless beverages (see *Recipes*) and foods. Switch to toxin-free body products. Live in a non-toxic house. Stay on a maintenance parasite program of herbs, and zap regularly. Stay on the kidney cleanse for three to six weeks and repeat a one-week session every few months to keep removing deposits which may also choose these sick joints to settle in. Knees are a favorite location for rheumatoid arthritis. Knees are very dependent on kidneys.

To summarize, do everything as for osteoarthritis, emphasizing the roundworm parasites for elimination.

Allergic Arthritis

A prominent food toxin that is said to affect knees is *piperine*, found in the pepper family which includes black and white pepper (not cayenne). The *Solanaceae* family of plants (potatoes, tomatoes, eggplant, tobacco!) also has a common "allergen" that produces joint pain. Try going off these for two weeks to see if it helps. Two more chemicals that can trigger arthritis-like pain are *hippuric acid* and *phenylisothiocyanate* (PIT).

Your body makes large amounts of hippuric acid, up to a gram a day. It is the product of benzoic acid detoxification by the liver. Quite a few fruits contain natural benzoic acid. But we can easily triple and quadruple our benzoic acid intake by consuming commercial beverages and pastries where benzoic acid is used as a preservative. It is indeed a "natural" preservative. All of it must be detoxified, though, and this gives us way too much hippuric acid. The kidneys are unable to excrete such overloads of hippuric acid, so it distributes itself in our organs. I suspect a simple mechanism could explain its pain-triggering action: hippuric acid molecules could attach themselves to our cell's conductance channels keeping the gates jammed open. This might invite bacteria and viruses to enter there. But there are also hippuric acid-loving bacteria that feed on it. Perhaps hippuric acid can cause pain without the help of bacteria. This requires further study. It is only sensible for persons with chronic pain not to consume benzoic acid (or benzoate) preserved foods.

PIT is a food chemical found especially in chicken, eggs, the cabbage family, peas. PIT is also part of the body's own chemistry, taking place in the liver, and involving detoxification of *cyanide*-containing foods. Many vegetables, notably the cabbage family, contain such cyanides, giving them protection from insects, disease, and grazing animals. It can take the liver a week to detoxify a meal full of these cyanides: in the meantime, PIT levels are higher in the body. PIT is very reactive. In fact, it is the chemical used as a general reactant with amino acids in the well known **Edman degradation** reaction. But now, your body is the reaction flask, supplying the amino acids. Since all organs supply amino acids, it is no wonder they can all react to PIT, giving you multiple allergic reactions and pains.

If you have any kind of arthritis, stop eating the high-PIT foods and clean the liver until you are free of all allergies you are aware of. This suggests that the liver is capable, again, of detoxifying the cyanides for you in a reasonable time and you may eat them again. These foods do have many benefits, of course.

Joint pain, or arthritis, was known in antiquity long before dogs and cats were household pets and giving us their parasites. Pigs and horses harbor these roundworms too and may have been the source at that time.

Herbs and treatments that help arthritis are, therefore, plentiful. Maybe they act by killing roundworms, bacteria, and viruses, or help metabolize hippuric acid and PIT. Homeopathic treatments, as well as massage, heat and electronic devices also help. With this wide range of effective treatments dating to the distant past, why is none of them a permanent cure? The answer is simple. The common roundworms are everywhere about us, sanitation is poor, and our civilized lifestyle leads to deposit formation that invites bacteria. But knowing this, you can stop your pain and remove the causes to become one of the first humans to achieve a permanent cure.

Verna Plumb, age 46, was diagnosed with rheumatoid arthritis four years earlier. Since then she had been continuously on methotrexate™ and prednisone.™ This had caused her to gain a lot of weight from water retention. The drugs were no longer effective and she would need to do something else very soon. She had the typical causes: her body was toxic with mercury and nickel from tooth fillings. Her kidneys were full of five kinds of stones. She had numerous roundworm parasites including two kinds of *Ascaris*, two kinds of hookworm, *Strongyloides* and *Trichinella*. She would have to clean everything up to get relief. She started on the kidney herbs, killed parasites with a frequency generator and in two months noticed her swelling was receding.

Camille Franklin had hands that were swollen and hot and painful. She also had "arthritis" in shoulders and knees. We explained that shoulders were not part of the arthritis picture. She could deal with that in a single evening, soon. She also had bone spurs at sinuses which needed surgical removal. Her kidneys revealed tricalcium phosphate crystals. She was given a diet change; onto milk, fruits and vegetables, off other beverages, less meat and grains. She was started on kidney herbs. In five weeks all the swelling and redness and heat was out of her finger joints.

Thigh Pain

Inner thigh pain often stems from the *sciatic nerve* which is suffering pressure at the lower back. If this is so, chiropractic adjustments should help. The correct treatment, after killing bacteria electronically, is to clean up the entire kidney area using the kidney herb recipe.

If this is not the correct explanation, (and you're not getting pain relief) you may have a trauma—perhaps you overstretched your leg in some exercise. The minor trauma invited bacteria to settle there and give you pain.

If the pain recurs after clearing it several times, there must be a chronic source of bacteria. Since the kidneys are already cleaned, consider the teeth, as well as recurring parasites, and the liver. Clean the liver every two weeks until 2000 or more stones have appeared and no more appear. This could take 1-2 years. Be patient. Do a dental clean up. Keep killing the bacteria so they can't spread. And, of course, let the painful leg rest.

Hip Pain

Hip pain is always due to bacteria. In fact, these bacteria regularly come from two sources: the kidneys and the teeth. This simplifies the treatment since neither of these places takes a long time to clear.

Start yourself immediately on the kidney cleanse (page 549) to clean your kidneys. Continue on the recipe until the dental cleanup has been completed. Both must be clean together to stop exporting bacteria to the hip.

The dental cleanup could take several weeks if extensive work is required. During this time, kill all parasites electronically. Keep killing bacteria, especially *Staphs*, *Streps*, *Clostridia*, *Campyls*.

The dental problem is not always on the same side as the hip pain. The jaw may be very fragile and porous, full of invading bacteria. Cleaning these cavitations may give immediate pain relief in the hip (proving the bacterial source). But getting the jaw bone to heal by taking up calcium again is not guaranteed by the cleaning process.

Give your jaw bone every chance to heal:

- Start taking vitamin D (40,000 to 50,000 u.), every day for three weeks from the day of the dental work or before. After this, take it twice a week forever. **Do not take more.**
- Take vitamin C (1+ gm a day) and B_6 (250-500 mg a day) for healthy gums.
- Get one gram of usable calcium in your daily diet.

Bone Strengthening

Vegetable calcium can't be dissolved by our stomachs (*ruminants*, like cows, can dissolve it—they have an extra stomach loaded with special bacteria who do the actual retrieving of calcium). Tablet-form calcium can't be dissolved well either, especially as we age. We need predissolved calcium if a little is to go a long way.

Primitive people who lived on fish or stone-ground meal ate 4-6 grams of calcium a day.[6] Even if only 20% of this got dissolved, they would still have about 1 gram of utilizable calcium for themselves. Chances are good they dissolved even more, since they were young (life expectancy was less than 50 years). Their skull remains show beautiful, cavity-free teeth. They lived outdoors, mainly, so getting enough sunshine-derived vitamin D was not a problem. Their natural diet supplied enough vitamin B_2 to protect them from UV (ultraviolet) damage from sunshine exposure.

But these are civilized times. Our lives are stretched into old age, when our stomachs no longer produce enough acid to kill bacteria, nor to dissolve the minerals in our food. So they need to be dissolved for us. Milk is a beverage where the calcium has already been dissolved by the other ingredients. The lactic acid in milk formed during digestion gives the calcium the correct chelated structure for absorption by the intestine. Even the bile participates in calcium absorption. Milk also contains phosphate, but not too much to be useful. I recommend milk as a calcium-source to heal the jaw bone after and before dental work. You need 1+ grams a day. One quart of milk has 1 gm (1000 mg) of utilizable calcium. You absorb only 250-400 mg. The rest is excreted and eliminated. Only the absorbed calcium can heal your jaw bone. Use milk in cooking as well as a beverage. The calcium in it is indestructible.

Bones are not made of calcium alone. **Magnesium** is essential. Since magnesium is more soluble and easy to assimilate than calcium, the tablet form (magnesium oxide, 300 mg, see

---

[6] Read Nutrition and Physical Degeneration by Weston A. Price, DDS first published in 1939. At least, gaze at the pictures. Ask your library to buy a copy, available from Price-Pottenger Nutrition Foundation, (800) 366-3748.

Sources) will do. If you are not absorbing the magnesium it will stay in your intestine and act as a laxative. If this happens acidify your stomach during meals: always add fresh lemon, vinegar, or vitamin C to your food or drink to help digest milk and dissolve minerals for you. **Boron** (3 mg. once a day) and **manganese** (15 mg. once a day) are additional bone hardeners.

When your diet is improved, your dental problem is cleared, and your kidneys are clean, your hip pain will stay away and you can stop thinking about hip replacement surgery.

Mary Hammond, 48, had two painful hips. She was diabetic and drinking five cups of coffee a day. She had *Staphylococcus aureus* infections under three teeth that she no longer had. After going off caffeine (caffeine may spread bacteria by making tissues more permeable) and getting some dental work done, her right hip stopped hurting. When cavitations were cleaned, her left hip stopped hurting too, but after a week the pain returned. She still had parasites and their bacteria to kill. She had solvents to eliminate and a kidney cleanse to do but she was quite enthusiastic and enjoyed showing off how well she walked.

Groin Pain

Lymph nodes are situated here, as well as in armpits and around the neck. Lymph nodes are your best friends. They are hives of activity. Your white blood cells “nest” here. Lymph nodes sample your body fluids (lymph) at these locations, much like the water department and health department sample our milk and water, making sure they are pure and sterile. Of course, they never really are. But your white blood cells keep working at it. They are busy removing impurities like zirconium and titanium and pathogens like bacteria and viruses.

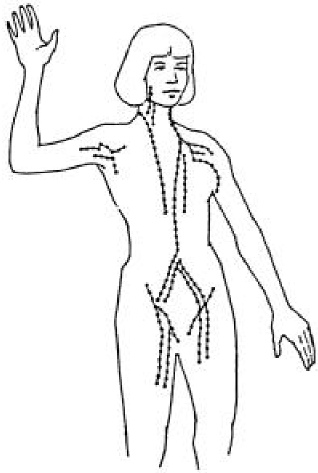

Fig. 18 Your lymph nodes are your best friends.

When an especially challenging dose of parasite or pollution arrives at the node, it enlarges, in order to handle the bigger sized task. This enlargement can be felt. It may press against other organs and you feel the pain of pressure in the groin. This should alert you, of course, to the danger. Kill the parasites and pathogens immediately with your zapper. Start the kidney cleanse. Stop using toxic products and eating toxic food. If you are being effective, the lymph nodes will begin to shrink in a few days, relieving the groin pain. If it does not come back, you probably eliminated the main cause. But if it recurs, try to diagnose it accurately. Test yourself for HIV and AIDS and then to pet saliva, dairy products and other disease specimens. Stop reinfecting from humans or pets or dairy food. Assist your body by cleaning up your dentalware, and environment. Check all supplements for toxins before using them on a daily basis. (If they show up in your white blood cells a few minutes after eating them, there is a toxin.) If your groin pain doesn't go away, you probably missed something!

There are other causes of groin pain, though, such as **hernia**. A physician will help you identify it. Have it surgically repaired. The Shouldice Hospital[7] has the best surgical record and will disclose their statistics! You won't even need full anesthesia there.

Pain from the **ovaries** is often felt near the groin. Sometimes a large cyst in the ovary puts pressure on the sensitive nearby organs. The formation of cysts, *cystic ovaries*, is a common condition of cats, too. Obviously, something is not right in the ovaries.

---

[7] Shouldice Hospital, 7750 Bayview Ave. Box 370, Thornhill, Ontario, Canada L3T-4A3, telephone (416) 889-1125.

Ovaries make hormones: **progesterone** and **estrogen**. If bacteria (like *Gardnerella, syphilis, gonorrhea*) are living there, the ovaries are handicapped and may under-produce or over-produce some hormone. The first treatment should be to kill parasites, especially the flukes. Then kill all other parasites and bacteria, especially *Gardnerella* and enteric bacteria which can migrate easily from intestine to ovary. In spite of all this, the cyst may not shrink. In order to shrink the cyst you must determine what is filling the cyst. Then stop filling it. The cyst will eliminate itself and shrink.

Many kinds of pollution can fill an ovarian cyst. Test for CFCs and PCBs. Gold is another favorite. Particulate pollutants like asbestos and fiberglass are often put into cysts by your body. Your body is wise. If these can't be eliminated through the kidneys or bile, it will at least keep them all together (in the form of a cyst) the same way a toxic dump spares the rest of the landscape. As long as you are adding to it, it cannot get smaller. Fortunately, your loyal white blood cells are trucking toxins away from your cysts every minute of every day and night. All you need to do is stop adding. Remove dental mental, clean up your body, diet, and environment.

You will notice shrinking of cysts in three weeks. What you actually notice is absence of pain, implying shrinkage. Meanwhile pain relief by killing bacteria, herbal ovarian assistance (try wild yam), chiropractic, pain killers are all useful. Decongest the area by means of a kidney cleanse. If your cyst does not shrink you have missed the pollutants. After it does shrink, stay on a regular program of parasite killing, herbal or electronic, and improved lifestyle.

Ovaries may be giving you pain even without a cyst present. By cleaning up parasites and pollutants you can eliminate it quickly; much more quickly than when a cyst has formed.

Synthetic hormones (Provera,™ Premarin,™ Ogen,™ etc.) are often used to clear up ovary problems of various kinds. Don't continue to use them when the need no longer exists. Although they have low toxicity, there are disadvantages such as the need for liver detoxification, and risk of heavy metal pollutants. Cancer acceleration has been seen when taking estrogens.

Low Back Pain

We have been told that lower back pain originates in an inherited weakness of the spine at its base because we humans walk upright instead of on four legs. And we have been told that the bony hooks that keep the spine aligned are flatter in some families, making it harder for them to hold the spine together. We are also told that “proper exercises” could have kept this part of our bodies strong so lower back pain could be avoided.

These theories become obsolete when, without surgery or exercise or change in posture, lower back pain can be made to disappear quickly and permanently.

Acupressure massage and chiropractic can bring “miraculous” relief. The most severely crippled lower back pain sufferer can shuffle lamely into a chiropractor's office and walk out normally, without pain or painkiller after treatments. So although there has been slippage of disc or spine, apparently it goes back into place rather easily.

If muscle relaxation is the clue, we must ask why these muscles spasm so easily. Any muscle spasms if you irritate or injure it suddenly. In fact, your whole body spasms and flinches if a sliver or bit of broken glass is in your shoe. If you remove these objects, the leg can walk normally.

Oxalic acid crystals are as sharp as broken glass. Use the kidney cleanse (page 549) to dissolve them and other stones. All lower back pain can be cured by removing the sharp crystals in the kidneys. It takes about three weeks to dissolve them. In some very severe cases, it may take six weeks.

Whether you have suffered a year or 20 years, the permanent cure is only weeks away.

Our bodies make eight or more different kinds of kidney "stones." The oxalic acid variety is associated with sharp stabbing pains. In its effort to eliminate this extremely vicious acid your body neutralizes it with calcium first to make calcium oxalate. Your kidneys can keep a bit of calcium oxalate in solution but not a lot. The excess hardens into crystals. A glass of regular or iced tea (not herb tea or green tea) has about 20 mg[8] of oxalic acid—way too much for kidneys to excrete. Tea is a toxic drink, not to be considered a beverage. Chocolate is very high in oxalate, too, and should not be used as a beverage (as cocoa).

Children should never drink tea or cocoa. Their delicate kidneys should not be faced with the daily burden of excreting large amounts of oxalic acid. And calcium used to neutralize oxalic acid is wasted. Calcium is a precious nutrient. It should be conserved for children's bone development.

---

[8] Taken from Food Values 14th ed. by Pennington and Church, 1985.

It isn't necessary to find which variety of kidney crystals are causing your muscle spasms. Different herbs dissolve different kinds. And by combining them into a grand herbal mixture you can be dissolving all varieties at the same time.

Wherever oxalate crystals have formed, a particular bacterium, *Proteus vulgaris*, can be found. Does that bacterium somehow thrive on oxalate crystals? Or even help them form? Does *Proteus* itself contribute to lower back pain? Is lower back pain in reality two pains in one—the sharp jabbing of glass-like particles plus the inflammatory effect of bacteria? Fortunately, you can kill *Proteus vulgaris* electronically.

By using your new diagnostic skills, you can test your kidneys for crystals. The kidney stone varieties I have tested for are: **calcium oxalate, uric acid, cysteine, cystine, monocalcium phosphate, dicalcium phosphate, tricalcium phosphate.** All these varieties can be dissolved by the herbal mixture. But all can be formed again in a week!

To prevent **oxalate** formation stop drinking oxalic acid (eating oxalate rich vegetables is not significant—spinach, chard, rhubarb and sorrel all have their place in the diet). Also take magnesium and B_6 supplements (as directed in the kidney cleanse).

To prevent **phosphate** crystals from forming, reduce phosphate consumption and drink milk as a calcium source. Keep your kidneys squeaky clean with herbs and copious water drinking. After drinking one quart of sterilized milk, two pints of water, one-half glass of homemade fruit juice and one-half glass of vegetable juice, there is little desire for additional beverages.

I have no understanding of what may cause **cysteine** or **cystine** stones (the genetic theory does not explain them either, considering that people without *cystinuria* make these stones).

Since they are sulfur-containing, and I have seen them appear after taking sulfa-drugs, it may be wise to avoid use of sulfa-drugs if you have a choice, or go on the kidney cleanse afterward.

If you have severe lower back pain you probably have several kinds of stones. Some persons have all seven kinds!

Rosie Zakar, age 30, came to see us because her mother was cured of lower back pain so severe she could do no housework for 30 years. Rosie had the usual crystals in her kidneys: oxalate, urate and phosphates. She was started on the kidney herb recipe. In three weeks she was so much better she would have missed her appointment if she had not wanted to cure her digestive problem and fatigue too.

Vera Vigneault, age 32, came mainly for help in getting pregnant but she already had lower back pain and mid-back pain. If she had gotten pregnant before clearing this up, she might have developed eclampsia and high blood pressure which are kidney-related disorders. She was started on kidney herbs for these. She chewed gum a lot and had bleeding gums. She stated her bad teeth were hereditary (meaning other family members had bad teeth also). For this she was instructed to stop chewing gum, start drinking three glasses of 2% milk a day and take a vitamin A&D perle. She was to floss her teeth once a day. (All floss varieties are polluted with mercury and thallium. Use monofilament fish line, 2 to 4 lb. weight.) Immediately after flossing she was to brush them with a new very soft toothbrush with five drops of 17½% food grade hydrogen peroxide. She was to avoid brushing teeth with metal fillings to reduce erosion. She was to brush them a second time without flossing first, this time with five drops of white iodine (potassium iodide) made up by the pharmacist, again avoiding the metal. She had only oxalate kidney stones and was to stop drinking regular tea, replacing it with single-herb teas. In five weeks her gums were better although she was still chewing a little gum and the "peroxy" had been too painful for her to use. Her low back and mid-back pains were gone too.

Gerhard Rogers, age 39, came for his lower back pain and leg cramps. He had mono, di and tricalcium phosphate crystals in his kidneys. His diet was changed to reduce phosphate (meat, cereal, breads, pasta, carbonated beverage) and increase calcium and minerals. He was to drink three glasses of 2% milk a day and to start the kidney herb recipe. In 25 days he had only minor improvement. He still had phosphate crystals. He was afraid to drink milk because he had heard so many bad things about it. He was told to boil it first to eliminate these “bad things”. A toxic element test showed a buildup of copper, arsenic, cobalt, cadmium, lead, thallium, vanadium and radon. This could easily explain his leg cramps, headaches and sleep problem, too. The arsenic came from pesticide, cobalt from detergent, thallium and copper from tooth fillings. The vanadium was fixed by having the gas pipes tightened, and radon could be reduced by improving ventilation under the house. He was thankful for the information and set about cleaning up his body and environment.

Alberta Mellos age 52, came in for lower back pain and upper back pain. It was explained to her that lower back pain was simply due to tiny stones cutting into her tissues but upper back pain was due to gallstones. She could clear her low back pain first. The kidney test showed she had oxalate and cysteine crystals. She was started on the kidney herb recipe. Nineteen days later she arrived with a cold but stated that her low back pain was gone.

Glenn Dirk, age 62, called on the telephone to say his urination had stopped, probably due to kidney stones. This had happened once before and now he was in a panic. He started our kidney herb recipe the same day and passed 117 stones the same night without bleeding or enough pain to need painkiller. After this, he could focus on his prostate enlargement and pain with sitting. He had intestinal flukes and other stages in his prostate gland as well as in his intestine. He also had *Clonorchis* (human liver flukes and their eggs) in his prostate. He had carbon tetrachloride methyl butyl ketone and TC Ethylene from food pollution accumulated there too. After stopping grocery store beverages and killing parasites with a frequency generator, he could urinate normally, freely and without pain.

Lower Abdominal Pain

The lower abdomen on the left side has the sigmoid colon as it comes down and bends. This is a favorite location for larger parasites to settle permanently. Flukes, roundworms, parasites of all kinds and their attendant bacteria and viruses can be felt if they produce gas and pain. Sometimes they live perfectly quietly, seemingly in harmony with us.

Moving the bowel more frequently expels them repeatedly and prevents their numbers from getting very high. Nature may help you with this by setting up diarrhea. Diarrhea is your clue that intestinal freeloaders are present.

The small intestine leads into the colon at your lower abdomen on the right side. At the junction is the ileocaecal valve that prevents backwash, and the appendix. The ascending colon goes up your right side then becomes the transverse colon that crosses your abdomen at the belly button level. The colon descends on your left side, leading into the sigmoid.

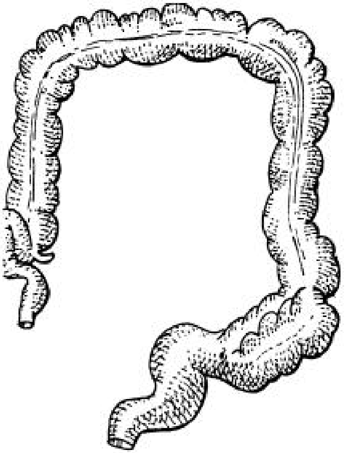

Fig. 19 Colon.

E. coli and *Salmonella* and *Shigella* are “enteric” (they live in your bowel) bacteria that can give you severe abdominal distress and pain. In fact, you can become a chronic sufferer. They can live on hands and under your fingernails, so reinfection from yourself is the most important source. Never, never touch your fingers to your lips. Most importantly, don't try to stop your frequent bowel movements. They will stop on a dime when your parasites and bacteria are dead and gone.

Other sources of *E. coli* are personal water bottles, other people's hands, hands that have changed baby diapers or cleaned bathrooms.

Hands do everything. To eliminate their threat of reinfection, cut out the section on hands (page 397) and paste it on your refrigerator.

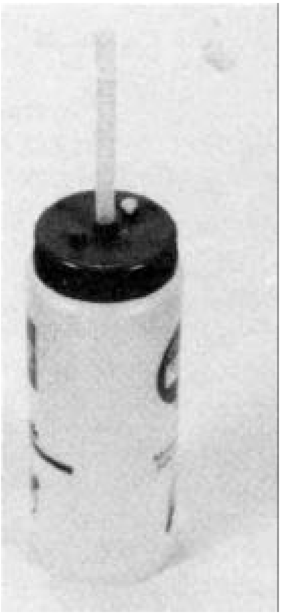

Fig. 20 Keep personal water bottles sterile.

Stomach Pain

Our dairy foods are polluted with *Salmonella* and *Shigella* bacteria. It is impossible to operate a dairy without getting some cow manure into the milk. Although udder wash contains antiseptic it does not kill all manure bacteria. Later, when milk is pasteurized, many heat sensitive bacteria are killed like the "friendly" *streps* and *staphs*, but not all the harmful *Salmonellas* and *Shigellas*. Some survive to colonize the milk, then later infect the consumer. Only milk that is sterilized is safe. A commercial source of sterilized (safe) milk can sometimes be found on the shelf (unrefrigerated). If it had any bacteria, it would not survive shelf life for more than one day!

You may not notice any discomfort from drinking milk, buttermilk, or eating yogurt without sterilizing it. Your stomach acids may be strong enough to kill them, or your liver able to strain them out of your body fluids and dump them, dead, into your bile ducts. Or they may live quietly in some part of the bowel where you feel no effects. Sterilize all your dairy foods by heating at the boiling point for 10 seconds, even if you have no symptoms.

If you are intolerant of milk it doesn't mean that you are lactase deficient, but that you are unable to kill any *Salmonellas* and *Shigellas*. These, in turn, can make you lactase deficient via frequent diarrheas. You can correct this situation by not eating any more bacteria. Give your body clean food, as intended by nature. Don't eat any deli foods, don't eat salad bar food or restaurant food unless it has been baked or cooked. Never eat chicken or red meat at restaurants. They are not thoroughly cooked.

As soon as a new abdominal pain or discomfort, or a gassy condition appears, zap bacteria and try to eliminate your bowel contents. Use the herb, *Cascara sagrada* (follow directions on label) as a laxative, or Epsom salts if necessary. Also start the Bowel Program (page 546).

If you have chronic abdominal problems, make sure you eliminate the bowel contents two or three times a day. There are herbs that can kill enteric bacteria, known to our ancestors of various cultures. **Echinacea** was a treasured herb of American Indians. **Goldenseal** is another favorite. **Turmeric** can kill *E. coli* and some *Shigellas*. **Fennel** can also kill some. But a single dose of Black Walnut Tincture Extra Strength seems to do the most. Make it yourself if possible (see *Recipes*, page 543).

If your body has lost its ability to kill *Salmonellas* and *Shigellas*, all the antibiotics and herbs and good bowel habits cannot protect you from these ubiquitous bacteria. You could ask how you lost your natural protection from them. There is evidence that common **antibiotics** that kill *Streptococcus* and *Staphylococcus* varieties are responsible.

A fraction of your bowel bacteria should be the friendly *Strep. lactis* and *Staph. epidermidis*. After repeated doses of penicillin-like antibiotics (that you may be taking for your strep throat) they are eradicated along with the “bad” *Strep. pneumoniae* bug. No amount of acidophilus culture (which contains active *Lactobacillus*) can replace these *Streps* or *Staphs*.

In mice, it takes a million Salmonella bacteria to start an infection. After giving them streptomycin, it only took 10 bacteria to infect them![9] Your intestines are similarly handicapped after antibiotics, and allow even very small amounts of *Salmonella* and *Shigella* to escape and multiply!

Try to get some natural killing power back. This means improving the stomach's ability to produce acid and the liver's ability to make bile. In turn, this means getting the toxins out of the stomach and cleaning the liver. Certain toxins accumulate in the stomach when the liver and kidneys can't keep up with elimination. Arsenic is a prominent stomach toxin. Get all sources of arsenic removed. Freon is another stomach toxin. Switch to a non-Freon containing refrigerator.

The metals from dentalware: mercury, silver, copper, thallium, first are swallowed and then land in the stomach. Clean up your dentalware. Toxins you inhale such as asbestos, formaldehyde, fiberglass, also are coughed up and swallowed to accumulate in the stomach. Test for them and clean up your environment. Any stomach can recover a significant part of its function by cleaning it up. Even though you regain your tolerance toward minute bits of filth in dairy products, do not go back to unsterilized milk products.

Appendicitis

The lower abdomen on the right side has the valve that separates the small intestine (*ileum*) from the large intestine

---

[9] Sherwood L. Gorbach, M.D., Perturbation of Intestinal Microflora, Vet Human Toxicol 35 (Supplement 1) 1993.

(*caecum*) called the *ileocaecal* valve. It is a common trouble spot because large parasites can attach themselves behind it and keep themselves safe from elimination. (A parasite's biggest worry is your elimination.) It is a favorite location for pinworms in children. It is near this point where the appendix attaches and this, too, is a favorite location of pinworms. With an appendix full of pinworms and their bacteria, is it any wonder when it gets inflamed and causes pain?

Appendicitis refers to inflammation of the appendix. Often, the pain isn't felt over the appendix but over the navel (this is an example of pain being caused at one location but felt at another; it is called *referred pain*). If there are any suspicions of appendicitis, zap pinworms and all enteric parasites and bacteria immediately. Because the current does not penetrate the bowel contents very well, zap every day for two weeks and take 2 tsp. (½tsp. for children) Black Walnut Hull Tincture Extra Strength one time. Make sure bowel movements are regular after this (see the Bowel Program, page 546, for hints) and hands are washed after bathroom use and before eating. Keep fingernails short for the whole family.

If appendicitis does not clear up it can lead to a burst appendix, spewing the dreadful contents into the abdomen. Kill pinworms and roundworms and enteric parasites regularly (once a week) in children. Keep pets on a regular diet of parasite-killing herbs.

Urinary Tract Pain

Urinary tract infections, including bladder, kidney, and urethral infections, are easier to clear up than to test for. Start by drinking a half gallon of water a day. Put yourself on the kidney herb recipe (page 549). This will dissolve the tiny crystals where bacteria hide and multiply.

Common urinary tract pathogens are *Gardnerella*, *Proteus*, *Trichomonas*, *Campylobacters*, *E. coli*, and *Salmonellas*. Sex will give you an instant dose of your partner's urinary tract bacteria. Make sure both of you clean up the urinary tract by zapping and doing the Kidney Cleanse.

Irritable Bowel Syndrome(IBS), Colitis, and Spastic Colon

are all conditions that combine parasite and bacterial problems plus an allergic trigger. If dairy foods trigger yours, you can guess it is not allergic at all but simply *Salmonella* or *Shigella* infection. Boil all dairy foods, stop eating ice cream, cheese and yogurt which you cannot boil. If eating lettuce triggers your intestinal attack, but other roughage does not, it may be a true allergen and cleaning the liver will eventually cure it. Apples, cinnamon and other “allergic” foods can be salvaged the same way. Wheat “allergy” is due to the pancreas being full of pancreatic flukes, wood alcohol, Kojic acid (a mycotoxin), and gold.

All these bowel diseases are quite easily cured by killing all parasites, bacteria, and viruses. Since reinfection is such a big problem, give your pet away until you are completely cured. Have your pet on the herbal parasite program before bringing it back. Clean up your diet, dentalware and environment. Your abdomen will be happy once more and grateful to you for your kind attention. Remember that zapping does not penetrate into the bowel contents. It kills only the outside layer of pathogens. For this reason you should zap daily for several weeks. For this reason, too, I recommend the Bowel Program (page 546) and Black Walnut Hull Tincture Extra Strength even though you may have gotten immediate relief from zapping alone.

Crohn's Disease

is somewhat more serious because the sores are higher up in the digestive tract. This is because sheep liver fluke and pancreatic fluke are commonly the main parasites and these live in the pancreas and liver. They often spill over into the upper intestine. *Salmonella* and *Shigella* are always part of the picture, too, as are various amoebae and fungi. The treatment is the same, kill all parasites and remove all pollutants, especially wood alcohol in commercial beverages. Healing of the digestive tract is very quick, often in a week. Reinfection is very quick too, if the rule about cooking dairy foods is not observed. Keep up meticulous hand sanitation.

Michelle Whorton had stomach pain at the middle of her abdomen, not related to eating. She had occasional very bad diarrhea and also daily headaches. She was started on the kidney herbs for other reasons. Her diarrhea disappeared! We found she had *Ascaris* (probably in her stomach where they cause indigestion and inflammation). She zapped them with a frequency generator set to 408 KHz for three minutes at 10 volts. She was to be very careful with sanitation since they owned a number of farm dogs. Next seen after six weeks, she stated that all her previous problems were gone but she had a different pain in the mid-lower abdomen that got worse during her period and sent pain shooting down both legs.

This turned out to be uterine in origin, but not endometriosis. Her uterus was full of asbestos, arsenic, gold, silver, titanium, propyl alcohol, benzene, styrene, toluene and carbon tetrachloride. This would invite any bacteria toward it!

Mark Lippman, age 51, came in for his irritable bowel syndrome, hoping we would find *Giardia* and put an end to it quickly. Actually, he had intestinal flukes, beginning to invade the liver. He also had propyl alcohol built up in his body giving him a precancerous condition that needed immediate attention. The flukes were killed in twenty minutes, along with *Ascaris* (he had swollen eyelids). He was allergic to milk as could be expected with so much interference with digestion. He was immediately better and did not need to come back.

Billy Henry, a 9-year-old boy, had diarrhea daily and stomach aches. He also wet the bed. There was an indoor pet dog and a bird. Electronic testing showed he had two kinds of *Ascaris* and pinworms. His young body also had a buildup of benzene, moth balls and carbon tetrachloride that he was eating, drinking, and breathing. His bed wetting stopped after killing parasites with the herbal program. His other problems recurred until he was older and could stop licking his fingers when eating.

Tom Ochs, age 36, had chronic stomach problems, alternating constipation and diarrhea, was labeled “lactose intolerant” after an elaborate test, and finally had been diagnosed with irritable bowel syndrome. Actually, he had *Ascaris*, besides other smaller parasites. He was also toxic with cesium from drinking beverages out of clear plastic bottles. This frequently causes depression and he was happy to understand his mood changes. After changing to purer food and products and killing his parasites, he did not need to come back. Five months later he was able to drink all the milk he wanted, no longer had sinus problems and lost his IBS.

Rex Callahan, age 5, had dark circles under his eyes, numerous ear infections until tonsils were removed and tubes put in, and many strep throat infections. Clinically, he was found allergic to dust mites, pollens, and animal dander. His skin got “rashy” if he drank too much fruit juice. He had frequent diarrhea. We found he had *Ascaris* parasites. They were not difficult to clear and he was soon a new person.

David Falls, age 52, had stomach pain and numerous health problems stemming, no doubt, from his diagnosed Crohn's disease. He became allergic to the sulfa drugs commonly used in this disease. We found he had sheep liver flukes and all their developmental stages in his blood and intestine. He was put on the herbal parasite program which he found difficult to follow. Nevertheless, in three months his bowel was nearly normal and the pain in his intestine much less.

Edward Marsili age 7, had bouts of stomach pain. He had intestinal flukes and a build-up of benzene in his body. This would seriously lower his immunity and ability to fight off tiny parasites. He was using a product containing an herbal oil that was polluted with benzene. His parasites were quickly killed with a frequency generator and he was put on the herbal parasite program. One month later his stomach felt much better, but he still had an occasional stomach ache. Testing showed hookworm and rabbit fluke. His benzene was now gone so the tendency to "pick up everything" was gone too. Staying on a child's maintenance parasite program would protect him.

Kim Johnson, almost two years old, had lots of ear infections. It started at eight months so the mother took her off cow's milk and wheat. This stopped her ear infections until mid-winter. She had to be back on antibiotics and a few months ago the doctor began discussing tube implants with her since she was still on antibiotics (six months). Another ear, nose and throat doctor agreed with this opinion, but was willing to wait until Autumn. The baby had been passing a lot of undigested food and was unhappy. They were vegetarians. The baby nursed. Our test showed pancreatic fluke infestation; this would easily lead to bad digestion, especially of milk and gluten in wheat. Fortunately, she was nursed throughout, in spite of going to daycare. Simply killing the parasites (in both mother and baby) solved both problems and she did not need to come back. The ear infections were probably caused by bacteria and viruses brought in by the parasites.

Cynthia Prout, age 36, brought her three children because of their poor health. They all, including herself, had stomach problems, a lot of allergies, asthma, ear infections, and milk intolerance. One boy, age 8, was intolerant of both milk and wheat and hadn't had them for years. He was infested with two kinds of *Ascaris* and pancreatic flukes. His sister, Nola, had itching legs and headaches besides; she was toxic with bismuth and antimony (from shampoo fragrance and laundry fragrance). She also showed a build up of vanadium, implying a gas leak in the home. The youngest, age 5, had frequent stomach aches and vomiting. It was a simple matter to kill *Ascaris* electronically at 408 KHz and the pancreatic fluke with all its stages (421 through 434) after which the children did fine.

Sofia Sobel had extreme ulcerative colitis although she came for her headache. She had been on Prednisone for a month with no relief. We found she had the three large flukes plus *Chilomastix*, dog whip worm, and amoebas in her intestine (but not in body organs).

There were several house dogs. Her stomach and intestines were much too sensitive to accept parasite herbs, or in fact, anything—anything except slippery elm powder. This herb (1 tbs. made into a paste first with water and then drunk as a beverage three times a day) paved the way for acceptance of two oyster shell calcium, one magnesium and one zinc tablet. Her blood test showed high phosphate levels since she was dissolving her bones to get calcium. She added sodium alginate (¼ tsp. to a cup of vegetable broth soup) twice a day to help her tolerate the parasite killing herbs. Her children were given VMF (vermifuge or parasite killing) syrup. By the 12th day of the parasite program she no longer needed colitis medicine; her bowel movements were down to twice a day, soft and formed, but still with a little blood streaking. She loved the alginate mixed with slippery elm. She was able to eat fruits and vegetables but agreed to stay off wheat and corn until her liver was cleansed. In another week she was free from all abdominal complaints except a heaviness over the uterus, possibly due to two missed periods. The thought of pregnancy put her in a panic. She was instructed to induce her period (Emmenagogue, see *Recipes*). She was sure she wanted her period, not a pregnancy and this seemed to be her God-given right. Three weeks later she had a flare up of colitis due to *Salmonella* in food; it also gave her a urinary tract infection. This time she took Quassia herb to kill invaders in addition to the maintenance parasite program which she had begun to neglect. She had been very busy, had lots of energy and wasn't on medications. She treated her urinary tract infection with betaine-hydrochloride (to acidify the stomach), began using plastic utensils to reduce her nickel intake (see Prostate Pain, page 124)) and drank a lot of water. This experience taught her valuable lessons that she was eager to learn, benefiting her family and herself immensely.

Rebecca Goetz, age 53, had ulcerative colitis and her husband had Crohn's disease. She had been on Azulfidine™ and Flagyl™ frequently. Her parasites were only intestinal flukes and their stages, and *Endolimax*, an amoeba. It was a simple task for her to clear her problems by killing them and by sterilizing her dairy foods. Wes, her husband, had three surgeries to remove sections of bowel due to Crohn's disease. He now had a colostomy but was on Advil™ for pain in the rectal area. He could hardly sit. He had been tried on anti parasitic medicine (Cypril™ and Flagyl™) but they did no good. He had intestinal flukes and all their reproductive stages in his body, also pancreatic flukes, *Capillaria* roundworm, and *Diphyllobothrium erinacea* scolex. We interrupted his testing at that point. His kidneys were full of phosphate crystals— he ate no dairy products. He was started on half-doses of kidney herbs and only part of the parasite program in view of his colostomy and possible diarrhea. Two weeks later we continued testing, finding pinworms, *Haemonchus*, *Leishmania tropica*, *Paragonimus*, *Sarcocystis*, *Stephanuris* and *Trichuris* (whip worm.) Quassia was added and doses increased. His blood test showed a high thyroid hormone level (T4), contributory to over activity of his bowel He was started on goat milk, vitamin C (3 gm. daily) and B_{12} shots. He was given magnets to sit on for pain. He was toxic with cadmium, from his old tooth fillings. But in five weeks he could sit comfortably without pain pills. There was less blood in the stool. Dental work would bring him his next big improvement.

Benito Villamar, a middle age man, had severe side pain for several weeks. He was also gassy. He had sheep liver fluke and stages in his thymus and intestine. The thymus is under the top of the breastbone and is a very important organ of immune function. It is easily damaged by benzene. He did, indeed, have benzene accumulated there. He was given a list of benzene-polluted products to avoid and was started on the parasite killing herbs after killing the flukes instantly with the frequency generator. Two weeks later his side was very much better, his benzene was gone and he was eager to rid himself of lower back pain, which he also had.

Al Vickers, age 9, had stomach aches, headaches, a constantly runny nose and asthma. He was on Slo-Bid™ medicine and allergy treatment. He had a sleep problem. He also had two dogs, one rat and two hamsters. The dogs and he had high levels of *Ascaris*. He was zapped for *Ascaris* and the four common flukes (without testing). He was put on vermifuge syrup and Rascal capsules. This ended his problems and began a new chapter of better care for his health by his parents.

Tim Melton, age 16, had several colitis attacks yearly, requiring hospitalization, from third grade to the present. He also drooled constantly, needing to spit a lot. (This is due to mercury toxicity from amalgam fillings. It is better to spit out the mercury than to swallow it.) He had intestinal flukes at a high level for which he was started on the parasite program. One month later he was very much better. He had only one diarrhea session since the last visit. But he still had sharp pains under both buttocks (probably due to kidney stones). He had been an iced tea drinker and had numerous oxalate and cysteine crystals deposited. He was appalled that a common beverage could be so harmful.

Central Abdominal Pain

can be coming from the uterus, bladder, or bowel. It is difficult to tell which is the source. The first step is to simply kill enteric (bowel) free-loaders and get into good bowel habits. Gas and bloating should be gone. If this isn't the solution to the pain there may be special bladder parasites with their bacteria. *Schistosomes* prefer to invade the bladder wall. In fact, very many parasites temporarily invade the bladder because the body is trying to excrete many of them. The whole family should be cleared of these same parasites. Kill them by zapping. Pets should not be kept indoors since they have many of these parasites, too, and they are easily transmitted to us.

Interstitial Cystitis

is one of the most painful conditions described by clients. *Schistosomes* are the real perpetrators but after the bladder wall is weakened, other parasites and their bacteria and viruses accumulate here too. To regain your bladder's health all toxins must be cleared as well. Dental metal, environmental toxins, including radon, asbestos, formaldehyde, must be cleaned up. The diet, body products and home should be carefully searched for toxins. *Schistosomes* are easily zapped but easily picked up off toilet seats and doorknobs. Always wash hands after toileting: a single droplet reinfects you!

Uterine Pain

Endometriosis

Many a woman's dreams have been shattered by her inability to have a child. Endometriosis is often the cause. It starts with painful cramps at period times. They get worse and worse until pain killers are necessary just to get out of bed and move about the house. There are flukes in the uterus! Large intestinal flukes in a rather small organ! Did they migrate to the uterus from the intestine or did they develop there from eggs?

Sometimes sheep liver fluke is seen there. Once an avenue to the uterus is established, numerous other parasites move in the same direction: *Clonorchis*, the human liver fluke and even *Eurytrema*, the pancreatic fluke, can invade the uterus wall. Why have they taken up so abnormal a living place? Because the uterus has solvents in it! This is the green light for flukes. This disarms your organs so they are left helpless against fluke stages left there by the blood and lymph. Stop eating solvent-polluted foods. There are solvents in all cold cereals. Make cooked cereal. There are solvents in purchased drinking water. Drink from your cold kitchen faucet. There are solvents in grocery store bread, grocery baked goods and cholesterol-reduced foods. Eat none of these. Buy baked goods and bread at your local bakery. Stay away from "low cholesterol" foods. There are solvents in decaffeinated and other powdered mixes for beverages. Drink nothing except milk from the grocery store (sterilize it). Milk does not have solvent pollution. The hormones, antibiotics and udder wash can be tolerated—solvents cannot.

Beverages and powdered mixtures sold at health food stores are no exception. Use no powdered mixture intended for weight loss or weight gain, nor vitality supports, nor dietary supplements. They are all polluted. Some solvents (I often see methyl ethyl ketone and methyl butyl ketone) choose the uterus to accumulate in. This sets the stage for endometriosis and fertility problems. Where there are large parasites, smaller ones soon crowd in. All bring their own bacteria and viruses. *Gardnerella,* especially, is found in cases of endometriosis, ovarian cysts and menstrual problems. The flukes evidently travel from the uterus to other parts of your body cavity, distributing bits of the uterine lining as they go. Once this distribution has occurred, can the bleeding (regular menstrual bleeding) at these extra sites ever be stopped?

It stops immediately when the flukes are dead!

Your body knows how to clean up after dead flukes and does the job perfectly. You can be free of pain in time for your next period. Zap to kill the four common flukes, *Gardnerella,* all other common parasites, and urinary tract bacteria (common ones include *Proteus, Salmonella, Campylobacter, Chlamydia, Trichomonas*). Avoid reinfection by avoiding solvents! It is impossible not to pick up parasites. If your uterus has solvent in it, they will find their way to it in a day. Without solvent, they will not.

The solvents will leave by themselves. Help your body get rid of them with vitamin C and B_2 (3 grams and 300 mg. respectively, daily, see *Sources*).

To heal the uterus so it no longer attracts parasites, clear up its internal pollution besides solvents. This means mainly the dental metal that has piled up and environmental toxins such as asbestos, arsenic, fiberglass, and formaldehyde. Gold and silver are especially attracted to the uterus. Don't wear gold rings or any metal jewelry touching your skin anymore, and, of course, get all metal out of your teeth. Never try to get pregnant before you have cleared up endometriosis.

The advice given by obstetricians to get pregnant to solve your pain problem is most unwise. Indeed, pregnancy changes your body's metabolism and without periods you get pain relief. But it seems much too risky to grow a baby in an infested, polluted uterus. Fear of birth defects is an intelligent fear. Be careful not to get pregnant while you are killing parasites and getting mercury removed from your teeth.

Healing starts as soon as all the parasites and pollutants are gone.

Joanne Biro, age 22, had severe cramping pain with her periods, diagnosed as endometriosis. She had adult intestinal flukes and the *cercaria* stage in her uterus. She had a xylene (solvent) buildup in both her brain (cerebrum and cerebellum) and uterus. She was started on the herbal parasite program following the kidney cleanse. Her next period was pain free. A check up showed she was free of flukes but had thallium in her immune system. Dental cleanup was next on her agenda.

Denise Leyva, 22, was on birth control pills to control the growth of endometrial tissue. She had laser surgery previously. She had hexanedione and methyl butyl ketone buildup in her uterus supporting the intestinal fluke and its eggs in the uterus. There were also some sheep liver flukes and human liver fluke stages there! She was advised to stop eating cold cereals and commercial beverages and kill the parasites immediately. She had no recurrence.

Anita Pierce, age 32, had numerous surgeries for her endometriosis. She also had chronic fatigue syndrome, and several allergies. She had two beautiful poodles in her home. She could not part with them so she gave them the pet parasite herbs faithfully. She had intestinal flukes, tapeworm stages, *Ascaris* and various other flukes in the uterus. She had all her dental metal removed and home detoxified. Her body was teeming with *Nocardia* and Epstein Barre Virus (EBV). In spite of repeatedly killing the flukes and bacteria with a frequency generator and making herculean efforts she was no better off eight months later. She was unable to solve the problem of reinfection from her dogs.

Christine Solton, age 27, had extremely heavy and painful periods and didn't stop bleeding between periods. A large cyst had been seen by X-ray in the uterine wall. She also had constant bladder pain. Both problems kept her in the bathroom most of the time (90 visits/day). She had the intestinal fluke in her uterus (probable cause of cyst) and *Schistosoma haematobium* (bladder parasite) throughout her body. She was started on the parasite program and in one week her bladder pain was under control but bleeding (from the cyst in uterine wall) continued. *Schistosomes* are very contagious, probably even from toilet seats and the house dust of an infected person. She had them again three weeks later. This time she zapped them and got instant relief. Her bladder and uterus were both full of propyl alcohol, tooth metal, fluoride, cobalt, zirconium, aluminum, antimony, cadmium, and formaldehyde. She was delighted, though, to understand her problem and made the dental appointment.

Contraception

There is an excellent pamphlet available at health food stores, called Wild Yam for Birth Control Without Fear[10] that informs that 3 capsules taken two times a day provides reliable (perfect) contraception provided you give it a two month head start. Also, an emmenagogue recipe is on page 546.

The Silent Cervix

The cervix is a big "trouble spot" for women just as the prostate is for men. It seldom lets you know with pain, however, that bacteria or parasites or toxins are present. Sometimes a brief needle-like pain does alert you to something going on there, but it is easy to miss. The cervix is constantly secreting a little bit of mucous and this helps it stay clean but why give it mercury and copper and gold to secrete? Many a fertility problem has been solved by stopping the toxic pollution of uterus, ovaries, and cervix. Clean up your dentalware and clean up your diet and environment.

---

[10] Willa Shaffer, published by Woodland Health Books, PO Box 1422, Provo, Utah 84603.

Kill parasites and bacteria regularly, every week, with the herbal recipe or by zapping. You should have no pain with menstruation, no bloating, fatigue or headaches. A PAP smear test should always be “good”. If it isn't, hurry to the rescue of your cervix.

Menopausal Symptoms, Hot Flushes, PMS

Insomnia, irritability, PMS (pre-menstrual syndrome), depression, anxiety, nervousness, are all not to be expected at and after menopause. They may certainly be caused by hormone imbalances. It is these imbalances that are not normal.

NO menopausal symptoms are normal

After the ovaries are done with their cycles of estrogen and progesterone production, the **adrenal glands'** hormone production was meant to “kick in” and make up any deficit.

During your fertile years, you were meant to have a peak of 100 picograms/milliliter (pg./ml) of estrogen on day 9 and day 22 of your cycle. Progesterone, on the other hand, only peaks once, on day 22, and it should reach a level 20 to 100 times as high as estrogen! After this ends, your adrenals can still keep your hormone levels regulated. Typical values are 20 pg./ml estrogen and, again, 20 to 100 times as much progesterone. Keeping these two hormones in balance is just as important as the actual amounts. **20 pg./ml estrogen is enough to prevent menopausal symptoms** including hot flushes, and to give you heart protection and bone density protection. Taking synthetic hormones usually gives you exactly such levels. But if your own adrenal glands can supply them, surely it is a better approach. (A blood test can tell you your levels; do it on day 21, 22 or 23 before menopause, after menopause the day is not critical.)

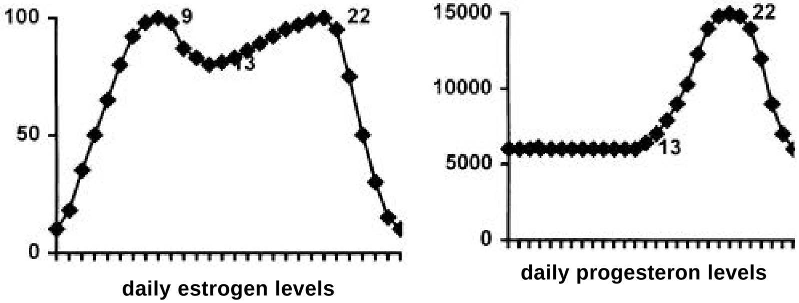

Fig. 21 Estrogen and progesterone levels (pg./ml).

Why aren't your adrenals producing them? Because they are hampered by parasites and pollutants! Kill all the parasites, bacteria and viruses, especially *Gardnerella*, *Proteus*, *Chlamydia*, *Campylobacter*, *Neisseria*, *Treponema*, *Salmonella*. Use the zapper.

Notice that these are also the favorite urinary tract bacteria! This makes good sense, because the adrenal glands sit right on the kidneys and would be geographically close to the kidney bacteria. To avoid getting them back, do a kidney cleanse (page 549) to remove all crystals where they might hide.

After this, hot flushes can be suddenly gone. If not, continue the cleanse. Meanwhile, do some permanent lifestyle improvement. Start drinking two pints of water between meals plus water and milk (sterilized) at mealtime.

Don't drink cranberry juice with its copious *hippuric acid* and its solvent pollutants. You won't need such insufficient help after cleaning up your body, dentalware, and environment. Besides giving you a better hormone supply, your newly revitalized adrenals will get you through stress in better shape and keep your blood pressure normal.

If bacteria are not kept down they will begin to invade other organs. *Gardnerella* goes to ovaries to feast on the corpus luteum after ovulation. This keeps the corpus luteum from making enough progesterone and PMS results.

Menstrual problems and PMS, in general, although they may not be pain-causers, are a sign that all is not well with the uterus and ovaries. Go on a cleanup program. Kill your invaders with herbs or the zapper. Clear up your toxic accumulations from dental metal and environmental sources.

Don't be surprised by a pregnancy! This is not the time to get pregnant, though. If you wish to get pregnant, clean up your body first, being very careful to prevent pregnancy during this time. If this is not under your control do your dental work first. Since every cleanup job increases your fertility, it is best to get the mercury, thallium, copper and nickel out of your body before your risk of conception is raised further by making other improvements. Many an amalgam replacement job had to be halted in the middle due to pregnancy! Couples just couldn't believe they could have a child as a result of cleaning up their bodies so they were careless in spite of my warnings! **A dentist will not take out mercury during pregnancy.**

Nicole Truett, 40ish, had PMS before her periods. She was a returning student and couldn't afford to feel emotionally upset. She also had recurrent yeast infections, *Herpes* and panic attacks. She complained that her thinking was weird, although she was an excellent student (she got her birth date wrong on our office forms!) She had high levels of mercury and we suggested dental work immediately. Two months later she was feeling much better and had all metal removed but was experiencing a slight return of symptoms which panicked her. She still had *E. coli* and *Bacillus anthracus* at four tooth locations, giving her chronic sinus symptoms. She needed her cavitations cleaned. Four weeks later she described how she had gotten immediate emotional relief after two cavitations were done. Our solvent test showed methyl butyl ketone, benzene and carbon tetrachloride (which we found in her Mountain Valley Arkansas Spring Water). They were lowering her immunity giving recurrent Herpes and yeast problems. These disappeared in a week. She was so happy to find the cause of her problems and yet so angry at the nature of the cause that she planned to write to the water company. We need more such environmental activists.

Monica Koziol was on Ibuprofen™ for menstrual cramps. She also got a severe sugar craving and minor depression and headache with her period. She was toxic with silver, copper, platinum, mercury and lead. She also had hookworms, pinworms, human liver fluke and cat liver flukes infesting her. A hormone test showed very low estrogen (57.6 pg/ml) probably due to all these interfering factors. She had all the metal from her fillings replaced and killed parasites with the herbal recipe. This gave her regular normal periods for several months. She decided to get pregnant but couldn't. A follow-up showed she was full of parasites again. She had stopped the maintenance parasite program. She resumed it and began kidney and liver cleanses to get longer lasting benefits.

Barbara Ashby, age 43, had suffered for 1½ years from menstrual pain. She had oxalate crystals in her kidneys and was started on the kidney cleanse. She followed this with the parasite program and dental metal replacement. Then she cleaned her liver and after three cleanses (she got over 1,000 stones the first time!) she said she felt great again.

Terri Entzminger, age 16, had a long list of health problems including painful ovaries and painful periods for which she was put on birth control pills by her doctor. She had several bacteria in her genital tract: *Neisseria gonorrhea*, *Plasmodium cyano*, *Staphylococcus* *aureus* (also at tooth #28 which had a plastic filling), *Streptococcus pyogenes* (also at tooth #28). A parasite test showed intestinal fluke adults in the uterus, not in the intestine or liver. There were also fluke *cercaria* here. She was started on the parasite herbs. Three weeks later there was no improvement. She had a rabbit, a dog and a hamster. She was to get them all onto the pet parasite program. Her diet was changed to exclude solvents. Six weeks later her periods were "great", she did not need the pill and she was keeping the whole household on a maintenance program killing parasites. Three weeks later she had cramps again. This time it was sheep liver fluke in the uterus. She probably got them from the snails in her fish tanks. She was off the maintenance program and drinking caffeine free cola again. This would fill her with solvents that make fluke disease possible. She killed them all with a frequency generator and decided to be more vigilant over parasites as long as she was such an ardent animal lover.

Azar Moya, 57, was on Premarin™ and Provera™ for hot flushes and emotional extremes, Synthroid™ for the thyroid, Xanax™ for nerves and sleep problems, something for diarrhea and something for depression. In five months she needed none of it. She had done a liver cleanse by then and got a commode-full of stones (about 1,000), she had changed her plumbing, got rid of the water softener, killed parasites and cleansed her kidneys. She still had sinus problems and some arthritis and was planning dental metal replacement and cavitation cleaning to clear them up too.

Infertility

An ominous sign in any species, infertility is not just another disease or "problem."

When birds' eggs don't hatch their species is doomed. So we learned from the DDT experiment humans did with birds in the 1960's. The DDT changed the thickness of their egg shells so they cracked when the mother bird sat on them. All changes are experiments whether intentional or unintentional. Nature by itself produces sweeping changes, too, such as droughts, wind storms, fires, ice ages, but usually living things had time or mobility to adapt to them. When there wasn't time, and they couldn't run away, the species went extinct.

Are we humans an exception to this rule of adapt or perish? Nobody feels more helpless and hopeless than the infertile couple. They can not run away, time is limited, and obviously adaptation is not occurring. More likely their lineage will perish. The couples' only wish is “Give us one child, now.” Surely, it is their birthright, as it is any living creatures', to reproduce.

Can we relax with the assurance that our intelligence, through the arm of science, will always rescue us? Are test tube fertilizations, fertility drugs, Cesarean sections, incubators for premature babies all triumphs for science? No, they are signs of reproductive failure for the human species.

When the concern is overpopulation of this planet, reproductive failure might seem less ominous. Maybe it's no worse than the natural way any species curbs its growth rate. Maybe only those who can survive parasitism, pollution and immune deficiency should survive in order to strengthen the species. But when reproductive intervention becomes a necessity, not an option, surely the danger signal is present as it was for the DDT'd birds who saw cracks develop in their eggs. The solution to our reproductive failure is not to find ever more artificial ways to conceive, to give birth, and to care for damaged babies. The solution is to fix the old fashioned way; to safeguard the natural way.

If you are unable to conceive or to provide viable sperms use an intelligent approach. Remove the obstacles. **The obstacles are parasites and pollutants, the same enemies of health we have seen before.**

Kill all large and small parasites with a zapper and the herbal parasite killing program. Don't try to keep a pet parasite free, give it away. Living close to another species is a luxury you can't afford at present. The pet can live with its parasites, you can't. Remember to kill bacteria and viruses too, especially *Gardnerella, Neisseria, Treponema*, the ancient enemies of human reproduction.

Is it safe to kill parasites if you might be pregnant?

The electronic way of killing parasites is safe if you use a frequency generator. The frequencies of parasites and bacteria are far away from human frequencies. The treatment with each frequency is short. There are no side effects.

The zapper has not been tested and should not be used during pregnancy.

The herbal way of killing parasites has been used by pregnant women without bad effects but this is not enough safeguard. I recommend waiting until the baby is born if at all possible. The treatment is long and intense. The growing baby is exposed continuously to herbs. Perhaps this is preferable to the toxins produced by parasites. You must use your own judgment. Obviously it is wiser to take a chance on herbs than to take a chance on inheriting AIDS or “genetic” diseases.

Part two of regaining your reproductive freedom to have a child is removing pollutants. Gold, silver, copper and mercury can accumulate in the reproductive organs, wrecking the delicate hormone balance between estrogen and progesterone, or wrecking the motility of sperm. Research has not been done to search for dental metal in the uterus, ovaries and testicle of infertile couples. You can do this research yourself. Slides of ovary tissue cost less than $10.00 as do other parts of the reproductive system. Search for dental metal yourself. Remove all dental metal from your mouth, and replace it with metal-free composite. Extract teeth with root canals.

A Word Of Warning!

Be extra careful with contraception during the dental cleanup. You could get pregnant the very next day! **This is no joke.** It is a serious hazard to conceive a child while mercury is loose and rampant in your body from the removal process. It may be a higher risk than leaving it untouched. If you are pregnant no dentist will want to finish the job of mercury removal! **Don't try to get pregnant yet.**

You may have tried fertility pills, in vitro fertilization, and other methods for getting pregnant over a ten year time period, all to no avail. Then you start cleaning up your body and taking your mercury out and suddenly you are pregnant before the job is complete! It may seem unreasonable and illogical to have to be careful after ten years of no worries, but play it safe.

If you fail to observe this warning and do get pregnant too soon, you may pray for miscarriage. Otherwise, take vitamin C and thioctic acid and hope for the best. Men should add daily zinc and arginine (60 mg. and 450 mg, respectively) to their diets. Both men and women should add vitamin E (200 mg.), a prenatal multivitamin and multimineral tablet, eat freshly grown vegetables for folic acid, and add vitamin C (at least a gram daily, see *Sources*). **No other supplements**! Supplements polluted with heavy metals or solvents do more harm than good. If you are not sure of their purity, test one by eating it and searching for it in your immune system five minutes later. If it is there, it is harmful; eat no more.

Nausea

of pregnancy is the scourge of expectant parents. After waiting hard and long for the desired pregnancy, the mother-to-be feels rotten, salivates and gags at the thought of food, and wants no more sex. Maybe sex is ill-advised during pregnancy, no matter how reassuring the male or male-oriented obstetrician is! Maybe salivation is actually mercury excretion being attempted by the body. Maybe nausea is all about keeping toxins out of the body and away from the developing child. These are intriguing possibilities, worthy of your research expertise.

A few decades ago the treatment for nausea was a weekly B_6 and B_{12} shot. Ask your obstetrician for this to see if it helps.

An older, herbal remedy was cinnamon tea: 2 tbs. cinnamon (bark or powder) in 2½cups boiling water, steeped for 10 minutes. Strain and add honey to taste. Dose: ¼ cup three times a day before meals.

Nausea invites starch eating—pasta, potatoes, rice and bread. Starches can absorb. Perhaps they absorb the noxious substances causing nausea. Make sure you add vitamin C to grains. In any case you must still eat additional nutritious food to grow your baby. In spite of craving a pickles/chocolate pudding/carbonated beverage lifestyle, you must eat mainly good food. Craving can take strange turns. Search for the taste you crave in good food and in long forgotten childhood foods.

These are all the fertility cases I saw in a year's time. None were left out in order to hide failure. Assess the success rate yourself:

Domilita Renshaw and her husband had been trying for six years to get pregnant. Both had been tested and treated in assorted ways. Domilita's period was irregular, a sure sign that all is not well in the area of reproduction. I gave them the usual warning about not risking pregnancy during their deparasitizing and depolluting procedures they both would be going through. Her hormone test showed slightly high (125 pg/ml) estrogen levels for day 22 (if it really was day 22!) and higher still if it was not yet day 22! Obviously, something was irritating the ovaries into overproduction of estrogen. She had oxalate and urate crystals and was put on the kidney herb recipe. She was switched to milk (3 glasses 2% a day) as her primary beverage besides water. She was toxic with nickel (dental metal) which would invite hordes of urinary tract bacteria, dangerously close to the ovaries. She made her dental appointment. She had sheep liver flukes and was started on the parasite program. She broke out in hives from a new hair spray polluted with praseodymium which got into her ovaries. She prepared to clean her liver for her frequent hives. Then she called to cancel her next appointment because she was pregnant (four months from first visit). Fortunately she had one visit with dentist completed. Nine months later she had a beautiful perfect baby.

Lindy Maloy and her husband had been trying for eight years to have their second child. They all had *Ascaris*, including, of course, their house dog. They wormed the dog monthly and did not want to part with it since they did not believe it mattered. They used the pet parasite program, but five months later she had higher *Ascaris* loads than ever. She also could not rid her uterus of intestinal fluke stages in spite of killing them with a frequency generator and using the parasite herbs. She remained full of solvents, bacteria and platinum from dentalware. Her endometriosis continued. They gave up.

Rosemary Peterson, age 33, had been trying to get pregnant for fourteen years. She had seven laparoscopies for endometriosis and very hard cramps with her period. She had intestinal flukes and sheep liver flukes in her uterus. There were sheep liver flukes and human liver flukes in her liver. There were intestinal fluke redia and cercaria in her saliva. The solvents in her uterus were methyl butyl ketone, acetone, carbon tetrachloride (from drinking store bought water), styrene (from drinking out of styrofoam cups), xylene (from carbonated beverages) and decane (from cholesterol-reduced foods). She also had a chronic yeast infection, treating it constantly with Nystatin™. She killed the flukes and yeast electronically before leaving the office and started herself on the parasite program and diet restriction. She got pregnant immediately, and did not return.

Elisabeth Tran, age 37, had tried to get pregnant for five years. She did get pregnant recently, on a special “gift” surgery program, but lost it. Her ovaries and uterus were toxic with mercury and thallium from polluted dental alloy. She also had barium and titanium in them, probably from lipstick. We did not see her again, we hope she solved these problems.

Christopher Gravely, a young man of 26 and Frederica, 22, promised faithfully not to get pregnant until their cleanup was complete. He was found by his doctor to have slow moving (low motility) sperm. He was robust and healthy looking but suffered a lot from low back pain—a clue to swarms of bacteria in the lower abdomen. An electronic search of his testicles and prostate (which had been infected once) revealed iridium, platinum and yttrium. This implicated tooth metal. He was also started on kidney herbs. Eight months later he had completed all his tasks, his low back and pain with urination had stopped, and this encouraged him to continue with his fertility program. We started him on thioctic acid two a day and zinc (60 mg.) two a day, switched him to an electric razor so he wouldn't have to use any chemicals on his face, and recommended that he ventilate his garage which was attached to the house, to reduce fumes in his home.

Meanwhile, Frederica, his wife, was also checked for toxic elements. She had antimony (from mascara) in her ovaries and breasts. She had sensitive painful breasts during her periods which were quite irregular. She also had indium and gallium, dental alloys in the ovaries and breasts. She, too, was started on the kidney herbs and instructed to get metal tooth fillings replaced. After two months they canceled their appointment. Frederica was pregnant! Not for long, though, and a wiser couple returned a few months later. Frederica finished her dental work. Both started the parasite program. Frederica's periods became regular. She was started on thioctic acid (one a day) plus zinc, (one a day), until her first missed period. Twelve months later they sent me their baby's picture: he was two months old.

Ginger Hart had been trying for three years to get pregnant. After an endometrial biopsy, a D&C,[11] and laparoscopy she was diagnosed with "inadequate corpus luteum." We found her ovaries toxic with nickel and europium from tooth fillings and strontium from toothpaste. She was delighted to understand her problem and set about correcting it.

Marjory Davis, age 28, had been on the "pill" (synthetic hormones) a long time but was off now and couldn't get pregnant. She actually got pregnant about one year ago but lost it at one month. A toxic element test showed her ovaries and uterus were full of beryllium (gasoline and coal oil), gadolinium and gallium. The metals are alloys of gold used in dentistry. She wore a lot of jewelry, just loved her chains, necklaces, rings, etc. But she agreed to go off all except two rings which did not have these alloys. To reduce fossil fuel fumes in the house she removed all gas cans and the lawn mower from the attached garage. They parked the cars outside. She was started on kidney herbs and promised to use contraception until she was done. She was to drink three glasses of 2% milk a day and take a magnesium tablet and stop drinking other beverages. Three weeks later her husband canceled her appointment because she was too embarrassed and delighted to call herself.

Prostate Problems

If urination is not complete, so you must soon go again, especially in the night, it is suggestive of pressure on the urethra from an enlarged prostate gland. Keeping a little urine from being voided is conducive to bladder and kidney infection, too, because bacteria soon find this "free food."

The prostate collects toxins as if it were a designated dump site, especially for nickel. Urinary tract bacteria quickly find "their" metal, nickel. Any supply of nickel will attract bacteria as quickly as crumbs and cheese attract mice. Urinary tract bacteria are making use of your urea as their food. To digest it, they first break the urea molecule apart into two ammonia molecules. It is the ammonia smell of the urine that gives away their presence. The urine should have no ammonia smell. Our bodies do not make ammonia. Only bacteria can do this! Their digestive enzyme is called *urease*. In order to do its job, this enzyme requires the element nickel.

---

[11] A surgical procedure, called *dilation and curettage*, meaning dilate the bladder with air and scrape away the inner lining.

Nickel is plentiful in the soil which is undoubtedly where these bacteria belong, breaking up and utilizing the urine and droppings left there by animals. They perform an essential task in our environment, destroying animal excrement and thereby cleaning-up the soil around us. What folly it is to load ourselves up with nickel so that in one short hop from the earth they are residing in us! Bacteria are all around us doing their valuable jobs. We cannot stop associating with bacteria. We were not meant to feed them, however. If we did not supply them with nickel, as if we were just another patch of earth, they could not gain a foothold in our urinary tract and then to the prostate.

How can we rid our bodies of nickel? This is the challenge. Are we eating dirt? A small amount of dirt can't be avoided on our food. But we can stop sucking on nickel as if it were a lollipop. Our spoons, forks and knives are made of nickel! Stainless steel is 8% nickel! Does it really come off as you eat? When you stick a knife into the mayonnaise jar, it is stained in a few minutes. Try it. The mayonnaise has reacted with the metal loosening the nickel. When you smear the knife over your bread, this film comes off and you eat it. If you were to put the knife in your mouth, now, you could taste the metal. You will later wash the knife but not before you have eaten enough nickel to supply all the bacteria in your body with the daily allowance of their essential element, nickel. Nickel is not our essential mineral. Even plants keep their nickel levels very low. But due to pollution of animal feed with it, even a hen's egg now has 3 mcg/100 gm of nickel in it.[12] To produce a nickel-polluted egg, the chicken must be polluted.

Especially infants and children should never be given metal cups or cutlery. They need all their immune power to combat the barrage of new bacteria and viruses that is emerging in this age of AIDS. Nickel is part of dirt and belongs there, not in our pots and cutlery.

Another large source of nickel is metal dentalware. It is used to harden gold! If you suffer from prostate problems, remove metal from your dentalware. Nickel is used to make bridges, gold crowns, retainers, amalgams. Exchange it all for plastic ware and composite buildups (see *Sources* for more dental information). Stop eating and cooking with metal utensils; use old fashioned wooden or sturdy plastic cutlery instead. Always use a plastic knife for butter or mayonnaise. Never buy foil-wrapped butter.

Nickel is fat soluble and is stored in your skin fat temporarily when a surge of it enters the body. Your skin oils may be loaded with nickel causing “allergies” in the skin. Male pattern baldness is such an allergy. The sweat tries to excrete it for you. Always wash off your scalp sweat to help with this excretion. The skin oils dissolve nickel from metal jewelry (sometimes leaving your skin with a greenish black color) and transport it into your body. Don't wear metal jewelry. Earring posts should be plastic. Metal watches and metal rimmed glasses should be replaced with all plastic types. Metal rings should be replaced. After lowering your total body nickel levels and your prostate disease is only a memory, you might notice scalp hair returning to sparse areas. Search around the edge of the hair line for the first returnees.

Bacteria cannot live in the prostate without nickel being present. You can cure your prostate problems with the simple tactic of stopping nickel pollution of your body. Notice that you get a fresh attack after accidentally using metal cutlery in a restaurant or eating mayonnaise-style salad with a metal spoon stuck in it. It might be wise to take a **histidine** capsule (about 500 mg., one a day for three weeks, see *Sources*) when this happens. Histidine is a nickel chelator. Taking **zinc** is helpful too (60 mg. two times a day for a month). Possibly, the nickel was poisoning zinc enzymes. Read about the benefits of **flaxseed**, too, but remember to test every product for pollutants before accepting it as a supplement.

---

[12] Food Values by Pennington and Church, 1985

Prostate problems of all kinds clear up when bacteria are zapped, the Kidney Cleanse is done, dental cleanup is done, and the Bowel Program is followed.

Richard Traylor, age 71, had suffered from prostate and urinary tract disease for three years. Scar tissue had to be removed occasionally from prior treatments of them. He was started on the kidney herbs and in two weeks (13 days) he had a considerable improvement in urine flow. At his follow-up visit we searched for toxins. He had radon, chromate, yttrium and strontium in his genital and urinary tract. He got rid of his water softener (such salts are polluted with chromate), toothpaste (strontium source) and opened the crawl space vents (source of radon). This cured these problems in less than two months. It also cured his stomach ulcers for which he had to take medicine. He was so pleased he decided to install a crawl space fan and pursue a parasite program and dental health just to see what extra health improvements he might get.

Omer Whitney, age 45, had always been a strong, healthy, hard worker. He could now barely walk, due to weakness and pains of several kinds; his prostate problems began several years ago. Our tests showed 4 kinds of kidney stones. He was started on only half a dose of the herbal recipe to give them a chance to dissolve more slowly. One month later he still had some stones although his leg cramps were already gone. At this time we found *Ascaris* (both kinds) which he killed with a frequency generator. We also found carbon tetrachloride and pentane (in decaf beverages) built up in him; also gasoline and TCE. He was considerably improved five weeks later and was quite eager to improve further. A toxic element test showed he was full of copper, antimony (from mineral ice massages), cobalt (aftershave), zirconium (deodorant), thulium (vitamin C fortified orange drink) and mercury (very high, from tooth fillings). He planned on cleaning all of it out of his body and regaining his lost strength.

Harvey Van Til, age 35, came in for his prostate and testicle swelling which began shortly after a vasectomy. He ached over the front right side of his abdomen. He was started on the kidney herbs and in four weeks he had eliminated his oxalate crystals and felt considerably better. We next found the adult intestinal flukes and human liver flukes in his prostate gland! After killing them immediately with a frequency generator and getting instant relief of pain, he got his own device and did not need to return.

Clayton Gamino, 26, had pain during urination which he interpreted as a left-over from a prostate infection he once had. He got all the metal out of his mouth, and did a kidney cleanse. A half year later he had no remaining pains and was able to father his first child.

Side Pain

Pain on the right side can come from problems at the ileocaecal valve or the appendix or the large intestine itself. It can also come from the liver which is higher up but is sending its pain message to your side. Pursue it as an intestinal problem first, killing parasites and bacteria and normalizing bowel movements with the Bowel Program. If the pain persists, especially if it reaches up the side to the

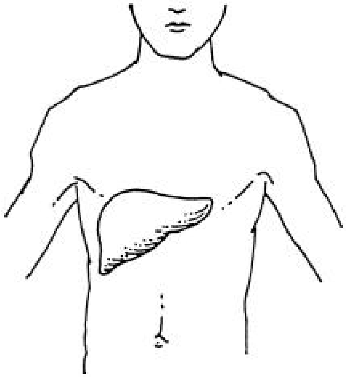

Fig. 22 The liver has a large lobe on your right side with the gallbladder tucked inside. The left lobe is smaller.

middle of the waist, it is probably due to the liver.

The liver is a large organ, mostly on the right side of the body, but with a smaller lobe on the left side. It is the chemical manufacturing plant for the body. It pulls in the food from the intestine (that you ate hours ago) and makes your body's chemicals from them. Toxic items are changed chemically into non-toxic items that the kidney is able to excrete into the bladder. Fatty things must be made water soluble for them to leave with the urine. The liver also makes bile and sends toxic items along with it to the intestine through the bile ducts. The bile enables calcium and fat to be absorbed. If the liver isn't getting much bile to the intestine, fat is left in the bowel contents and the feces will float in the toilet bowl. That is your clue to bile duct blockage.

Bile is bright green. Mixed with intestinal contents it turns the bowel movement dark brown. If the bile is not arriving in the intestine the bowel movement will stay light colored, even yellow or orange. This is another clue to bile duct blockage. Over a quart of bile should exit the body each day. Since bile is loaded with cholesterol this daily excretion of bile is a major method of keeping cholesterol levels low. If the bile ducts are choked with debris so only half as much (often only a cup instead of a quart!) bile is produced and excreted you can expect cholesterol levels to rise, and digestion to be bad. When food isn't promptly digested and absorbed the ever present, ever-ready bacteria will grab it for themselves.

Taking cholesterol-lowering drugs should be reserved for cases where natural excretion cannot be regained.

E. coli and other intestinal bacteria, which do no harm in moderate numbers, can overgrow in a few hours to give you bloating, gas and pain. Your body produces no gas. Only bacteria can produce gas. If your side pain is accompanied by bloating and gas, you know you have a digestive problem. And that this digestive problem stems from a congested liver if the pain is directly under it or over it, or if the feces are light colored or your cholesterol levels are high. Not everybody has all the symptoms.

To clear the clogged passages of the bile ducts, you simply do the liver cleanse (page 552) over and over until the problem is gone. There is one catch. If there are living parasites in the bile ducts, they will not let the bile ducts clear themselves. They are stuck fast to your ducts as a tick can be to your skin. They must be killed before they will let go. Zap them all, or you may use the herbal parasite program, staying on a twice a week maintenance program. Only after parasites are dead (after day 20 if using the herbal program) will you get a lot of “green stuff” and be able to clear “stones” out of your bile ducts. Only one large duct at a time will clean itself. We have hundreds of larger ducts and thousands of tiny ducts feeding into the larger ones! Stay on a schedule of cleansing the liver every two weeks (unless you are ill) until your side pain is gone, your digestion is normal, and you are bouncing with energy. You may also lose some weight, but only if you are overweight.

Remember that a clogged liver does not necessarily give you pain by itself. It is more likely the bacteria in the gall bladder and bile ducts, causing inflammation there and in your intestine, that cause pain. Don't wait till pain occurs over the liver. Use whatever clues you can to diagnose your clogged condition. Or just assume it is clogged. Do the cleanse, and see if you get any stones out. It can never hurt and can help a lot.

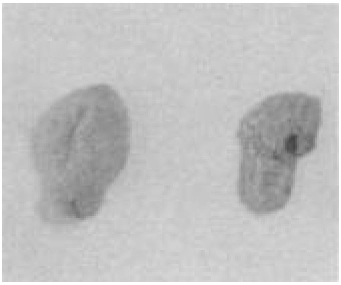

Fig. 23 Gallstones.

Bruce Hearn, 40ish, had severe side pain for several weeks and was rather gassy. He had seen a clinical doctor who found him in good health. But we found sheep liver flukes in the intestine and in the thymus! The thymus is an immunity-giving gland, so anything in the thymus is a very serious matter. He also had benzene in his thymus (inviting AIDS). He quickly switched from drinking soda to drinking milk. He went off everything in the benzene list. He killed the flukes electronically and started on the parasite program. Three weeks later the benzene was gone, his side was very much better and he could begin a kidney cleanse for his low back pain. He hadn't cleansed his liver yet! His improvement was probably due to improving his immunity which then controlled the bacteria.

Midabdomen Pain, Stomach Pain

The colon crosses over from your right side to the left side at the midabdomen. This is also the location for the bile duct to join your small intestine. Most midabdomen pain comes from either the colon or the bile duct connections. Kill parasites and bacteria by zapping or with the herbal recipe. Improve your digestion with diet clean up (off moldy food, boil dairy products). Sometimes the midabdomen pain stems from the stomach itself. The valves at the top and bottom of the stomach are meant to keep the food in. The valve at the top where the stomach joins the esophagus is a favorite location for bacteria.

You seldom feel them here though. This is just under your breastbone. They do their work quietly. Eventually, the tissue there is so weakened, the valve can't shut tightly and food is allowed to go back up the esophagus. This is called *reflux* and you may be told to sleep with your head elevated and to eat small meals, especially at bedtime. You may be given Reglan,™ a drug to help empty the stomach faster. Digestive enzymes will help empty the stomach, too, but may harbor molds. The real solution is to kill the bacteria in this area and keep them from reinvading it. Start by killing the *Salmonellas* and *Campylobacters* (zap them).

Within a day or so, however, the area can be re-invaded. Some bacteria are coming from your teeth! Some come from the liver! Often both sources are supplying bacteria to the stomach. The stomach allows a little bile to enter at the end of each meal, this is normal. But if the bile is full of live parasite stages and bacteria they may try to colonize the stomach, too. If there is insufficient stomach acid to kill them or if there is an accumulation of toxin in the stomach, they will get a foothold. Then they can burrow down deep to escape acid. Eventually, an ulcer can develop.

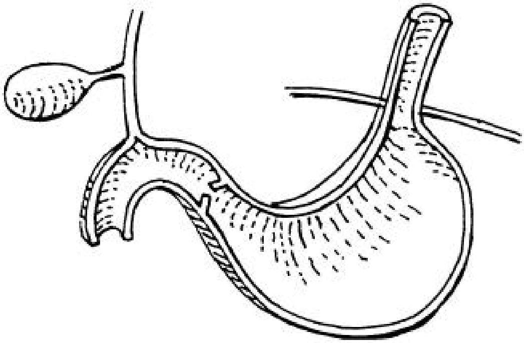

Fig. 24 Stomach, esophagus, diaphragm, gallbladder.

Stomach Ulcer

Often, such an ulcer is painless.

You are, therefore, not alerted to the parasites' presence until they have established themselves in good hiding places. Taking antacids, of course, works in their favor. The solution for both stomach pain and stomach ulcers is to kill parasites and bacteria, followed by dental and liver clean ups. One very common toxin that accumulates in the stomach is the pesticide **arsenic**. You inhale it right along with the flies and roaches you may be trying to kill with arsenic-laced pesticides.

Other inhaled toxins are fiberglass, asbestos, formaldehyde and freon. Your nose and mouth mucous traps a lot of these whereupon you swallow them and they glide into the stomach. Tooth bacteria and tooth metal get into the stomach the same way. You simply swallow them.

Clean up your air, don't use any pesticides (see other methods in *Recipes*). Your dentalware may be cleaned up in a few dental visits but the liver cleanses must go on for a year or two before it is reasonably clean. You may get pain relief in a few weeks but this should not derail your intention to revitalize yourself completely with a cleaned liver and stomach.

Hiatal Hernia

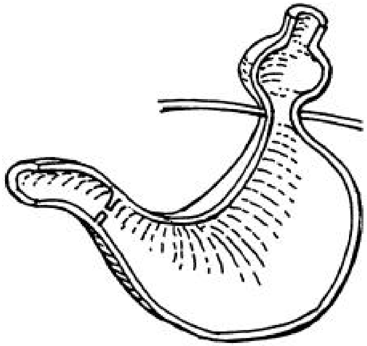

Fig. 25 Hiatal hernia.

When bacteria have spread to the diaphragm and weakened it, along with the upper-stomach valve, food is allowed to get pushed up right through the diaphragm. Pressure in the abdomen from overeating or sitting in an easy chair pushes it up. Walk after eating. Empty the stomach before going to bed. Don't sit in recliners. Very many of us, about 25%, have a small hiatal hernia. It only hurts if it is inflamed. Work at clearing the inflammation first, to get pain relief. In time, the area will strengthen and the hernia itself may get smaller.

Jeffy, 3½ months, had been screaming a lot ever since birth. He was very gassy. He could not be put down without starting to cry. He was congested and had colds. He was on mother's milk exclusively so his chronic *Salmonella* infection must have come from someone who was a carrier, possibly his mother, although she had no symptoms. The mother had platinum and tellurium in her milk (*Salmonella* can be transmitted in milk but this was not checked). The solution was to clean up the mother's milk by getting the dental metal out of her mouth, and washing hands before nursing. Jeffy's father had an *Ascaris* and intestinal fluke infestation. It is quite possible the baby had these also, giving him a nasty tummy ache in addition to the gas pains. The baby's older sister had screamed and cried the first nine months, too, so the parents were told it was a “familial thing.” She also had ear infections, colds and a flaky patch of skin occasionally. She, too, had *Ascaris*. They were very pleased to be able to clear up the whole family's problems by killing parasites and removing toxins.

Ruby Adair, age 14, filled a page with her symptoms. She was also chronically fatigued and had consumed enough antibiotic “to fill a room.” She had shooting pains from under her feet up her legs. A quick check revealed mercury and thallium in her immune system. Instead of being dismayed, this news filled her with hope for recovery. She arranged for dental metal replacement. Then she related her stomach “story”. While at a wedding, a year ago, she began vomiting with diarrhea. She thought it was the stomach “flu” but she didn't recover for six months and had to miss school. The psychologist thought it was emotional. She was still only attending school one hour a day. We found *Fasciolopsis*, the intestinal fluke, in her stomach wall as well as in her intestine. She started the parasite program and in three weeks her appetite was back, insomnia was gone, fatigue was better and a significant improvement was evident.

Respiratory Illness

Asthma

is a very old disease described in the ancient literature. The only progress we have made to date with this disease is to give drugs to soothe the symptoms.

Asthma is associated in **all** (100%) cases with tiny *Ascaris* larvae. As soon as eggs hatch (in the stomach, immediately after swallowing filth) the microscopic larvae travel to the lungs, not the intestines. Imagine the distress of lungs full of tiny worms! One tries to cough them up, of course, but in our misguided effort to be polite we teach children to swallow anything they cough up! Some swallowing is inevitable and the young worms are back in the stomach, this time to set up their housekeeping in the intestine. Some never leave the stomach, causing children stomach aches and, of course, a large entourage of bacteria which, in turn, have their viruses.

Most cases of *Ascaris* infestation also show *Bacteroides fragilis* bacteria which, in turn, carry the *Coxsackie* viruses (brain viruses). Whether or not these bacteria or viruses will thrive in you depends on whether you make a good home for them, namely have low immunity in some organ. The preferred organs for *Bacteroides* are liver and brain (brain tumors always show *Bacteroides*). The preferred organs for *Coxsackie* viruses appear to be tooth abscesses and brain.

Not everybody with *Ascaris* develops asthma, even though they always go through a lung stage. Does it depend on the age of the person when the infection develops? Or how many *Ascaris* are present? Or the time of year when lung infection is present?

That innocent cough of early childhood should not be neglected, as simply "croup." At the first sign of a cough, use a frequency generator set to 408 KHz, the frequency for *Ascaris*, or use a zapper. Pay extra attention to washing hands before meals. Pay extra attention to the animals nearby. Kill their *Ascaris* with a zapper and keep it up daily or put parasite killing herbs in their food.

Asthma sufferers become **allergic** to many air pollutants such as pollen, animal dander, smoke. The production of histamine in the lungs and the vast interconnectedness of histamine to allergies has been well studied scientifically. Although invasion by worms is known to result in both histamine production and high eosinophil counts (over 3), and **asthma clients typically have both**, they are not routinely checked for worms clinically! They are simply given drugs to enable better breathing. More and better (though toxic) drugs have been developed.

But you can put an end to your asthma by terminating your *Ascaris* infection. Then wash your hands and fingernails with grain alcohol, and let no more filth past your lips. Wear plastic gloves for a week to break a nail biting or finger sucking habit. For children wash hands before eating anything, even between meals; keep fingernails short.

Dogs, cats, pigs and horses all get *Ascaris*. Never, never let a child clean up any vomit or mess left by an animal. This could lead to massive infection, the kind that could result not only in asthma but **seizures**. Kill the worms in animal messes before you clean them up too. Never use your dustpan or broom. Use cardboard, newspaper or anything that you can afford to throw away with the mess. Squirt povidone iodine (available at drug stores) on the mess–even if it's outside–and cover with salt before cleaning up. Discard it in outside trash. Wash your hands with grain alcohol, paying special attention to fingernails. If you are a food preparer, you could infect the whole family if you don't wash carefully.

If there is an asthmatic in your family, the whole family should be treated for *Ascaris* with a zapper or with the herbal parasiticides. Even after everybody including the pets have been treated, pets should not be allowed in the bedroom of the asthmatic person. Asthma is more than parasitism. It is also an allergic reaction, to the pet and to other inhaled bits of matter.

Clean the air. Smoke of any kind, fragrance and chemicals of any kind, all household cleaners, polishes, and so forth should be removed. Store essentials in the garage, not the basement, since basement air rises. Clean up the whole house. Persons who must use hair spray or nail polish should do so outside in the summer and in their own rooms with the doors closed in winter. Use a chlorine filter for the water, especially at the shower. Install central air conditioning if possible, with maximum filtering (but never with chemicals added to the filter and never with a fiberglass filter) at the furnace. A room air filter (not fiberglass) is next best. Use it for an hour in the bedroom ahead of bedtime so the air currents can cease. Just moving dust around is worse than leaving it there. Never do dusting when an asthmatic is in the house. Lungs heal quickly when the air is clean and there is no reinfection. The best place to recover is outdoors away from trees and bushes or indoors with total pollution-free air conditioning (free of asbestos, formaldehyde, arsenic, fiberglass, pet dander).

As your asthma lessens, reduce your inhalers, but always keep them on hand. When you suddenly need them, try to identify your source of reinfection or allergens. Use this experience wisely. Try to understand the recurrence of your asthma. Keep notes. It may take half a year with a dozen recurrences to finally learn and conquer! It will feel great to breathe without spraying yourself and taking medicine. If it comes back a year later, figure out what is happening that's seasonal. A pine tree near the house, a flowering bush, the first mown grass? Stay away from these until you are completely healed.

There are traditional herbs for helping lungs. Grow your own **comfrey** and **garlic**. Make **mullein** tea from the dried herb (see *Sources*). Read herb books for more help. Dry some for winter use, being careful to do it right and not let it mold.

Suzanne Carlyle, 45, had asthma from infancy. She was currently on Albutesol^{TM} spray and tablets daily. Now she was beginning to have arthritis too. She was given arginine to replace caffeine. She had two species of *Ascaris* and was allergic to cats and other animals. She was started on the herbal parasite program after killing *Ascaris*, *Bacteroides* and *Coxsackie* viruses with a frequency generator. Her lungs showed kerosene, carbon tetrachloride, mineral oil, benzalkonium (from udder wash, she was also milk sensitive), aluminum from her cookware, and aluminum silicate from her salt. She had two extra lung parasites: *Paragonimus*, a lung fluke, and *Pneumocystis*. She was immediately improved after cleaning up these sources and canceled her future appointment.

Cay Wenkert, 63, had asthma for many years for which she took Proventil™, but this gave her such bad side effects she had to stay home now and not venture out. Her lungs were full of benzalkonium (toothpaste), arsenic (ant poison under kitchen sink), zirconium (deodorant), and nickel from tooth metal. She had dog heartworm in her lungs(!), in addition to her heart where she sometimes felt pain. She had *Ascaris* and *Naegleria*, *mycoplasma*, *Endolimax* and the intestinal fluke in her lungs! She coughed up blood, after her doctor had diagnosed *bronchiestasis* recently, meaning her lungs were not capable of sweeping out the daily refuse we all breathe in. In spite of killing these and cleaning up her environment she got no improvement. She repeatedly got parasitized. She had four or five root canal teeth which she was unwilling to have pulled. Hopefully, the tradeoff between teeth and health will soon begin to look poor.

Amy Newberry, 41, had recently begun to have asthma attacks. She had been on cortisone but now was on inhalers through the day (two puffs Ventillin™ 3x/day, plus three puffs Intel™ 3x/day plus four puffs Asthmacort™ 3x/day). She was often hospitalized for attacks. She also had stomach trouble and sinus problems (had pneumonia recently). She had *Ascaris* larvae, *Endolimax*, *Naegleria* and *Acanthocephala* in her lungs! She also had arsenic and palladium (tooth metal). Going onto homeopathic medicine for stuffiness helped her avoid some hospital visits. It took several months (5 visits) to track her arsenic source to the bedroom carpets (stain resistance!). After steam cleaning it herself and doing a liver cleanse (after first killing parasites) she was amazed at her improvement. She had not been to the hospital in a month and was only using inhalers preventively. She planned to get dental work done. She got all her former health back.

Laura Brewster, 25, lived in a very old house. The slightest exertion would give her an asthma attack. She frequently got pneumonia, too. Her lungs had beryllium (coal oil) and asbestos, and two parasites, *Paragonimus* (lung fluke) and *Ascaris*. She got rid of all the old fashioned lamps and candles in the house, but could not find the source of asbestos. She got rid of the attacks but her cough and pneumonia bouts will continue until she moves from that house.

Brett Wilsey, 70, was congested most of the time, had chronic sinus problems, was getting allergy shots for dust and mold, and was on several inhalers for his asthma plus emphysema. He had asthma for eight years. He had oxalate, urate, and all three phosphate variety crystals in his kidneys. He was started on kidney herbs. His blood test showed high "total carbon dioxide" or "carbonate" showing that his air exchange was not good. His potassium was low, showing that his adrenals couldn't keep his electrolytes (sodium, potassium and chloride) regulated, in turn, giving him muscle weakness. His LDH was very high, showing that his heart muscle) was in distress, too. Fortunately, his RBC was slightly elevated, to help with oxygen delivery. (See Anemia page 285). His eosinophil count was high, as is the rule for asthmatics since they all have *Ascaris* worms. Besides *Ascaris* he had pancreatic and intestinal flukes in his lungs(!) plus *Heterakis* (a chicken roundworm), human hookworm, and *Prosthogonimus*, another fluke. He was started on the parasite herbs. Two weeks later he felt a lot better although he still had some *Ascaris*. He was toxic with barium and hafnium (which were traced to his dentures) nickel, tin, rhenium. He stopped using commercial "denture-soak." His wife stopped using hair spray and they started leaving the car out of the garage and walked around to the front door, instead of using the attached garage. He also started drinking homemade comfrey tea. This made him feel still better.

When he removed his watch he got rid of his nickel source. Nickel has an affinity for lungs (as well as prostate and skin). He was now down to one puff of inhaler instead of two, only four times a day instead of hourly. He could breathe deeper than before. Then the lead in his water was found and traced to a single "sweated" joint in the pipes. He was started on food grade hydrogen peroxide, working up a drop at a time; now his cough became "productive", he was coughing up a lot.

He got the plumbing fixed and did a liver cleanse after which his fatigue lifted. He couldn't stop laughing and joking about the "emphysema" he was so worried over earlier.

Bea Karnes, 49, had asthma from childhood when she also had eczema. She was presently on several medicines plus allergy shots and antibiotics for it. She was toxic with antimony although she used no eye makeup and europium, tantalum, and gadolinium from dental metal. She also had hand swelling in the morning and was started on kidney herbs. She killed her intestinal flukes (in the intestine) and *Ascaris* in her lungs and was not seen for half a year. When we eventually saw her, she said her asthma had been “wonderful.” She was faithfully controlling parasites with the maintenance program and occasionally drinking kidney herb tea.

Cynthia Prout's whole family (mentioned previously) had asthma. She was told it was inherited. The three young children and herself were on inhalers, nose sprays, cough syrups and antibiotics. Nola, age 10, also had ear infections and numerous allergies. Lewis, age 8, was a slight, nervous boy; he had been off wheat and milk for many years due to intolerance. Irwin, age 5, seldom went with the family due to his frequent stomach aches and the fact he could vomit without notice. The mother and two children who were with her (Irwin stayed home) had *Ascaris* infection and Lewis also had pancreatic flukes. Their home was toxic with vanadium, namely, a gas leak. The parasites were quickly killed and gas leak repaired. The whole family got well and canceled their next appointment. Some inherited disease!

Asthma is just one of our common respiratory problems. The causes are always a combination of *Ascaris* and other parasites with pollutants (allergies).

Bronchitis, Croup, Chronic Cough

- In **bronchitis** the bronchioles are the site of the problem.
- In **croup** it's further down.
- In cases of **chronic little hacky coughs** it may be heartworm!

Bud Portillo, 62, worked on earth moving machines. He frequently had a “bronchial virus” or “bronchial infection.” He was always coughing. He had *Ascaris* infestation and arsenic pollution in his home. He also had palladium toxicity from tooth fillings. As soon as they removed the mouse bait from their home, tore down the hallway wallpaper (arsenic source) and changed wells (the well water had arsenic in it from seepage!) and killed parasites, his cough was gone. He also finished his dental work. All this took six months.

Gene Mizell, age 4, was constantly clearing his throat. His parents wondered if it was a "nervous habit." He had two species of *Ascaris* living inside his small body. He was started on a small dose of parasite herbs immediately, stirred into a daily dose of VMF syrup (see *Sources*). He also had PCBs, aluminum silicate, lutetium from recent painting in the house and xylene and toluene solvents. He was switched off soda pop and onto milk. We saw him five weeks later. He had gone through three treatments with vermifuge syrup. His throat clearing was gone, as well as his hacky cough and the blue circles around his eyes. He still had PCBs, which they later cleared by switching off detergent.

Carmen Castro, 30ish, had a chronic cough without being ill. She also had some heart irregularity. She had *Dirofilaria* (dog heartworm) and *Loa loa*. There were no indoor pets. After killing the parasites with a frequency generator and starting on the parasite herbs she was still coughing a bit but her pulse was down to 80 (from 120). She still had toluene, TC Ethylene solvents and mercury, platinum and lead in her body. Then a gas leak developed in their home. She used moth balls and painted a room. After this she had to be on antibiotics for her extreme cough. This time she had *mycoplasma*, *Haemophilus* and *influenza*. She killed these electronically and her coughing stopped. Until she gets the dental work done she will continue to pick up new infections.

Hope Feldman, 57, had been coughing for half a year. She had seen many doctors including a reflexologist. She had *mycoplasma* and *alpha strep* at tooth #17 (lower left wisdom tooth). As soon as these were killed with a frequency generator her cough stopped. She was advised to wear turtleneck sweaters for extra warmth over her bronchii and get a cavitation cleaned at tooth #17.

Teresa White, 37, had bronchitis several times each winter and was put on antibiotic for the whole season to keep it from breaking out. Her lungs were loaded with tantalum from dental metal, cobalt from detergent and thulium from her vitamin C. She had taken herself off milk, thinking it might be a factor, and used an air cleaner but without results. As soon as she had the tooth metal replaced with plastic (in less than a month) she could go off antibiotics and also was rid of a chronic sinus condition, but still had a bronchitis bout. She had *Ascaris* larvae in her lungs and phosphate and oxalate crystals in her kidneys. Her diet was changed to include milk and fish, magnesium, lysine (500 mg one a day), vitamin B_6 and a vitamin A+D capsule. She started on the kidney cleansing herbs and then the parasite program. In another month the whole problem was gone. Wisely, she got ready to cleanse her liver.

Craig Stewart, 2, had a history of respiratory problems. He had been on Ceclor™ and Ventilin™ for a long time; he had pneumonia the previous year. He had whip worm (*Trichuris*) infesting his intestine which was promptly killed with parasite herbs (as much as his parents could get down him was effective). He was toxic with asbestos. When the clothes dryer belt was changed to a U.S. variety (imported belts contain asbestos which flies into the air when the dryer is used), Craig's health turned around.

Doris Gumb, 26, was on Isoniazid™, Tussionex™ and Rifodin™ for tuberculosis. It started with coughing. She was down to 98 lb. in weight. Her lungs were toxic with beryllium (coal oil fuel), mercury, uranium, and tellurium. She began by clearing all toxic items from her house and basement and then bringing an air sample for testing. She also had *Ascaris* larvae in her lungs and pancreatic fluke in her pancreas. In three weeks she was coughing less and producing less with each cough. The “clean” air sample still had uranium, tellurium, mercury and beryllium. She had no metal tooth fillings, though. In the next three weeks she found a hole in the floor to the crawl space. It was letting up uranium dust and radon. The mercury was traced to the carpet in a child's bedroom. After throwing it out, her sputum cleared up and she was on the way to recovery, although we never found the source of tellurium.

Breast Pain

Although lumps and cancer in the breast produce no pain, they sometimes do give you little warning twinges. These twinges cross into the breast from the armpit or from any direction. It is gone in a second, leading one to believe it couldn't be serious. If the breast has any unusual sensations, painful or not, investigate immediately.

Test yourself for cancer. (You may use a specimen of mother's milk as a cancer test since this has mitotic stimulants in it—see Curing Cancer page 331). If you have purchased a slide of breast tissue (mammary gland) you can search your breast for cancer. If not, but you find "mother's milk" in your white blood cells, assume it is cancer and clear it up immediately.

If you don't have cancer, search for the pollutants giving you these twinges of pain. Search for your deodorant, cosmetics, and soap in your white blood cells. Search for dental metal. The breasts are often full of nickel. Nickel is quite soluble in fat and the breast is mainly fat. Nickel is one of the top carcinogens listed by researchers. It could even explain the high incidence of breast cancer. But titanium and barium from cosmetics, as well as asbestos and fiberglass are also quickly accumulated in the breast. Clean up your dentalware and body products. Check for exposed fiberglass. Change your dryer belt. Buy a new non-CFC refrigerator. Never try to get rid of these pains with pain killers; let the pains show you whether the clean up has been complete.

Breast Sensitivity

Breast sensitivity can be quite uncomfortable to the point of not being able to wear a bra, especially near period time. It may be due to high estrogen levels; this is also conducive to breast lumps and breast cancer. Most of your estrogen is produced by the ovaries before menopause and later by the adrenal glands. Too high estrogen levels plague the modern woman. Certain food molds, particularly **zearalenone**, causes over-estrogenization. It affects men too. I have found it in popcorn and corn chips! And brown rice. Stop eating these. Eat white rice. If you make cooked cereals be sure to add vitamin C to them before cooking (1/8 tsp. per cup), to detoxify food molds. I do not know whether taking vitamin C with your popcorn would detoxify zearalenone. Don't risk it. The excess estrogen compounds must be detoxified by the liver. Yet, the liver may be incapable of this because you ate yet another food mold! See the section on moldy food (page 381).

Over-estrogenized women are over-emotional, seemingly on a roller coaster of enthusiasm and despondence. They can develop a high pitched voice, that almost sounds squeaky. High enough estrogen levels are important for fertility but too high levels can cause infertility. Your body is eager to set the level just right, if only you will clean up the ovaries of parasites and pollution. Don't stop your clean up until the breast feels normal again throughout your cycle and you don't feel over-emotional, even just before your period.

Breast Lumps

Breast lumps may or may not be painful. If you feel one, don't wait to be more certain, don't wait to analyze it with tests, don't wait for a doctor's diagnosis or a mammogram. Obtain a frequency generator or zapper and zap yourself immediately. Also do the herbal parasite program immediately (page 338).

Your body often turns the breast into a collecting station for toxic wastes that have been drawn downward from the top of your body. From your head where shampoo and hair spray and cosmetics leave their daily deposits, from your dentalware with its constant supply of heavy metals, from neck and armpits where cologne, deodorant and soap leave their toxic residues. The lymph nodes under the armpit and the region above the breast, collect it all and let it slip into the breast where it is bundled up in a cyst. Perhaps the kidneys are clogged so toxins are forced to go to a designated dump site instead of out through the bladder. Do a kidney cleanse. Don't rest until all your breast lumps are gone. They will begin to shrink in three weeks if you are removing the correct toxins. Even radon and asbestos go to the breast, so be meticulous with your cleanup.

When the platelet count (in a blood test) is very high (over 400) there is quite a tendency to form cysts or lumps since platelets make your blood clot. The platelet count goes up when parasites are present. Maybe your blood is attempting to clot them! These clots make “nests” for fluke stages which may be why breast lumps often become cancerous. If yours is over 300, (it should be 250,000/cu mm) start patrolling parasites regularly. Stopping the use of caffeine and taking vitamin E (400 u. a day) are helpful in recovery but don't rely on these minor measures. Breast lumps definitely invite breast cancer.

Leslie Yeager, age 37, had breast soreness and "fibrocystic lumps”. She had cerium and nickel accumulated in her breasts. They cleared up in weeks after her dental metal was gone (she simply took out her retainer). Later she replaced it with a partial made of plastic.

Kari Pfeifer, age 36, had numerous cysts in both breasts and uterus. Her estrogen level was too high (187 pg/ml on day 22 of her cycle; the day of testing is important since it varies through the cycle). Her breasts were full of beryllium (coal oil from hurricane lamps) and radon. After these toxins were removed, all her breast lumps got smaller. After she did the kidney and Liver Cleanse, the lumps got softer and breasts were no longer painful. She had several root canals which filled her breasts with numerous bacteria, mainly *Histoplasma cap* (root canals develop infection around themselves). After starting her dental cleanup and killing bacteria with a frequency generator, all her breast lumps disappeared.

Claudia Davis, age 41, had breast soreness ever since a mammogram two years earlier. She had numerous other pains and indigestion. She had intestinal flukes in her intestine and fluke eggs in her blood: a dangerous situation. They might land in the breasts and start developing there. This is how cancer begins. But she did not have cancer yet. She had a buildup of niobium from polluted pain killer drugs and thulium from her vitamin C. She had *Salmonella* and several other bacteria in her white blood cells, which accounted for digestive problems. In eight weeks she had cleaned kidneys, killed parasites and gotten rid of her heavy metals. Her breast pain was better and a lump on her eyelid had also disappeared.

Stephanie Nakamura, 68, had six surgeries to remove breast lumps, going back to youth. Her recent mammogram was O.K. Her breasts were toxic with cadmium, lead, gold, radon, uranium, gallium, silver. Our tests showed she had kidney crystals and she was started on the kidney cleanse. She was given vitamin E, (400 units daily), sodium selenite (150 mcg daily) and vitamin C (1 or more grams daily). Her triglycerides were also very high showing again that she had kidney problems. She was given magnesium (300 mg daily), vitamin B_6 (250 mg daily) and lysine (500 mg daily). She killed parasites and cleaned up everything except gallium, silver, mercury, gold, cadmium. These must have come from her gold crowns. Her dentist advised against removing these and proclaimed they had nothing to do with her developing glaucoma, arthritis and stomach ulcers. It was a tough decision for her and she made the wrong choice. Perhaps if she had been up for the next breast surgery she would have gotten those “gold” crowns replaced with composite too.

Heart Pain

Pain over the heart region is usually quite real, even though an EKG does not find any abnormality. The most common cause is *Dirofilaria*, heartworm of dogs. It often begins as a pain just above the heart but spreads itself over the whole heart region. Kill it with your zapper. Parasite herbs can also be effective. If you did kill them, the pain often intensifies for a day before it leaves. Then the pain should be completely gone.

Heartworm is very easy to pick up again. If you have had heartworm, you should no longer keep a dog for a pet. Give it away.

Another heart parasite, *Loa loa*, is also a filarial worm and may be the causative factor. Both *Dirofilaria* and *Loa loa* can be obtained as slide specimens to use for testing yourself. Heart muscle can also be obtained as a slide specimen, but a chicken heart from the grocery store or snippets of beef heart (make sure to sample all 4 chambers) will do.

Follow up on your heart, even when no pain remains. These tiny heart parasites have stages that you may not be able to purchase in slide form and therefore can't test for. These stages, if not killed, will become adults so a maintenance parasite killing program, herbal or electronic, is essential. Virtually all dogs have *Dirofilaria* in spite of monthly medicine to kill it. They pick it up immediately after their last treatment for it and can give it to you again. The only way to live safely with pets is to give them parasite killing herbs daily in the feed.

Other heart problems such as irregular beat and mitral valve prolapse can clear up along with the pain. Or they may be due to bacteria (see Heart Disease, page 318).

Meredith Zackman, age 53, came for her diagnosed cardiomyopathy (heart disease). She owned a beautiful, old, very big dog, and of course she would never part from him. We knew she would lose her battle against heart disease. She had both *Dirofilaria* and *Loa loa* which we killed instantly with a frequency generator. She was on Lanoxin™, Furosemide™, Captopril™ and Metoprolol.™ We found she also had *Cytomegalovirus*, *Staphylococcus aureus*, *Streptococcus pneumonia* in her heart. The *Staph* bugs were also in tooth #17. She had copper in her heart (from tooth fillings) and cobalt and PCB from her detergent. There was zirconium from her deodorant and fiberglass from somewhere. Her pulse was typically in the 90's. Eight days later, after her dental work to replace metal was done her pulse was normal (low 70's), both worms were gone and she felt much better. But she still carried four bacteria, five viruses and two tapeworm stages: *Taenia solium* scolex in the spleen (she had chronic pain there) and *Taenia pisiformis* in the liver. She was started on Rascal (an herbal combination) for these. Six days later she had all of her problems back including *Loa loa*. She repeated everything, then she had to go off her heart medications because they lowered her blood pressure and pulse too much. She started the dog on the parasite program but continued to be heavily laden with parasites and bacteria that always found their way to her heart. She purchased her own frequency generator and was quite faithful with dog treatments. She may outlive her dog and then regain her health, finally.

Bruce Walby, age 42, had chest pain for three years. We found he had *Dirofilaria* and *Loa loa* in all four chambers of his heart. When he zapped them, the pain left a day or so later.

D'Ann Fonties, age 22, had a lot of chest pain but was told by her doctor it was simply “gas bubbles”. She also had a serious digestion problem. She had *Dirofilaria*, high levels of styrene (from styrofoam drinking cups) and benzene. This information delighted her and she planned to change her habits.

Sheila Osborn, age 27, had chest pain when she lifted objects. Her pulse was slightly elevated (81) and slightly irregular. She had *Dirofilaria*. Five weeks after starting the parasite program she was feeling much better but still had the chest pain. This time she had *Loa loa* (but not *Dirofilaria*).

Wendy Lewellen, age 28, had a chronic cough and chest pain at midsternum (the sternum is the bone attached to the ribs and runs up the middle of the chest). She had *Dirofilaria* in one chamber of her heart (right auricle). She had xylene and toluene solvents which came from her daily beverage, Mountain Dew.^{TM} She was also full of asbestos from her trips to do laundry nearby (this could not be proved, but when she switched laundromats, the asbestos went away). Two months later, after killing parasites, she was free of heartworm and her cough and chest pain were almost gone. She was probably healing very slowly due to the asbestos which was still present.

Lupita Cline, a young mother, had a chronic hacky cough and irregular heart beat. She had *Dirofilaria* in all chambers of her heart and *Loa loa* in her blood but not in her heart. She had carbon tetrachloride, propyl alcohol, hexanedione, toluene, and TC Ethylene buildup from drinking Pepsi™ and Mountain Dew.™ Her pulse was slightly elevated at 80. There were no pets in the house. She was started on the parasite program and two months later was rid of her heartworm but now had *Loa loa* in her heart and was still coughing a bit. She was full of platinum, mercury and palladium from tooth metal as well as vanadium from a gas leak in her home and paradichlorobenzene from using moth balls. She was on antibiotics for a "bronchial infection" and was happy to learn about a better solution.

Slow Pulse/Syncope (Passing Out)

Mason Heckler, 30s, was a mechanic by trade and could not afford to pass out on the job. Yet, it had been happening off and on for 10 years. He had acquired high blood pressure in his teens! This was soon followed by an extremely slow pulse (50 beats/mm). No medicine worked (he had been tried on many) so he was on none. Then he got high blood pressure, it was 160/80 currently. He also had constant chest pain around the left nipple. He had heartworm and was started on the herbal parasite program. He never had a dog. In five weeks his pulse was 72; the parasite was gone. I presume his syncopes were due to sudden blood pressure changes or missing a few heart beats in a row. He began the kidney cleanse next to lower his blood pressure.

Chest Pain

When there is a tightness or just a little pain at the middle of the chest, especially under the breastbone, you may be merely having an allergic reaction. You might be feeling little spasms coming from the esophagus, and reaching up toward the throat from gallstones. You might also have HIV/AIDS disease which has a similar symptom over the sternum. So it is very important to pay attention to even a minor symptom in the chest.

Ruling out HIV disease ranks first in importance. Search for its emission at 365 KHz, or purchase a microscope slide with the dead virus on it as a test specimen (see *Sources*). Or purchase a set of slides representing all the stages of *Fasciolopsis buskii*. Without this parasite you can't get the HIV virus. Purchase a slide of the thymus gland or make your own specimen of throat sweetbreads. Check yourself for benzene buildup in the thymus. (See Using The Syncrometer page 462.)

If you have neither the benzene nor the parasite stages, you have no risk. Your chest distress is due to something else. Improve your air quality so that your lungs are not in distress. This includes radon, chlorine (from the bleach bottle under the sink), colognes, room fresheners as well as the usual pollutants (asbestos, arsenic, formaldehyde, fiberglass, freon).

If you feel waves of pain reaching up to your throat, you probably have a gallstone stuck in a bile duct. Epsom salts can relax that bile duct in 20 minutes. Take a tablespoon in ¾ cup water but only on an empty stomach or you may feel quite ill. Taking a large dose of valerian herb (6 to 8 capsules) may also buy you a little time by relaxing the duct. If you do get relief, you can be sure it was a spasm of some kind. The magnesium in Epsom salts relaxes spasms. I would recommend cleaning the liver (page 552) a number of times to try to dislodge the sticking gallstone. The instructions for liver cleansing advise you to kill parasites and cleanse the kidneys first. But if your throat pain is severe enough, you might just zap and go ahead with it at once.

Upper Back Pain

The main pain may be a dull ache over a shoulder blade, or between the shoulder blades or running right through you from the front to the back of the chest. These are all gallstone pains coming from the liver! Get yourself ready to clean them out. Clean the kidneys first, kill all the parasites that might be residing in the bile ducts and blocking them (flukes, pinworms and roundworms) and get onto a maintenance parasite program. Then mark your calendar for your first liver cleanse. Even if your first cleanse gives you only a dozen bits of green "stuff" you have done well; you have accessed the bile ducts. The pains will probably be "magically" gone the next day, but they might start to return in two days. The bile ducts are having spasms again due to the remaining stones.

After you have cleaned out 1,000 or more, you will get permanent relief. Repeat every two weeks, unless ill, until the upper back pains are gone, permanently. If chest pain or upper back pain is severe, try going off your favorite high fat food (ice cream, butter, cheese). Also try taking 6 valerian capsules, 4 times a day including bedtime to relieve the spasms.

Shoulder Pain

Some shoulder pain is called *bursitis* and some is called *arthritis*. But it always derives from stuck gallstones in the bile ducts of the liver! You can prove this by taking a tablespoon of Epsom salts in ¾cup water at 6 p.m. instead of supper and on an empty stomach (or you may feel quite ill). If some of the pain subsides then you have evidence as to its true cause, because Epsom salts relax the bile duct valves.

Get started cleaning the liver (page 552). By the time you have chronic or acute shoulder pain you have about 3,000 Stones! Count them roughly, as they float in the toilet after the liver cleanse so you know how much progress you have made toward the final goal. Don't start cleaning the liver until you have killed parasites and spent three weeks cleaning the kidneys though. This improves elimination of liquid toxins so a liver cleanse is promptly cleaned up for you.

What emerges from the liver is the most contaminated mess imaginable, full of bacteria and viruses and parasite eggs and stages (all dead we hope) of every kind. It needs prompt clearing from your body. The diarrhea sees to the bowel elimination. But some toxins can only pass through the kidneys. Kidneys should be clean and capable of doing the job. Remember it is unwise to clean the liver before all parasites are dead, especially flukes, because they produce a substance that inhibits any action of the bile ducts!

You are only one day away from freedom of shoulder motion and sleeping on your side again. Permanent improvement, though, depends on progress with your total stone count. Your bursitis can return in a few days or a few weeks. Be patient. You may only cleanse once in two weeks, and not if you are ill. After six cleanses you can be quite sure of being relatively pain free.

Between cleanses use valerian capsules to stop the spasms. It takes 6 capsules 4 times a day to be effective.

Typically only some bile ducts are spasming, and typically those ducts have a single fatty food trigger. Stop eating the high fat food you consume the most (it's probably your favorite). If that hasn't helped in two days choose a different high fat food to omit. The most common culprits are ice cream, potato chips, salad dressings, cheese, butter, cream, and milk.

Perhaps the pain is actually caused by bacteria living in the blocked bile ducts and invading the shoulder. This point has not been clarified. Using your zapper or frequency generator does no good for this pain. Only liver cleaning brings your shoulders back to their youthful mobility.

Upper Arm Pain

Excruciating pain in the upper arm soft tissues can keep your arm hanging straight down for fear of worsening it with motion.

Magnets of high strength (2x5000 gauss) taped to your arm, under your sleeve, can get you through the day. Also try valerian capsules (6 capsules four times a day). Go off all fat in the diet to let the spasm subside. Then start your liver cleanse at 6 p.m. If you used pain killer drugs during the day, the cleanse may not yield anything but it's worth a try anyway. You might be lucky and pop out the chief culprit stone. If not, you should wait several days before trying again; this time avoid pain killers the day of the cleanse. Be sure to zap parasites the day before or earlier.

Peggy Patton, age 60, had shoulder pain and painful feet in addition to aching all over. The aching was due to *Trichinella* which both she and her husband had. It took six months on the parasite program before it stayed away. She had clay colored stools, evidence of bile duct blockage. Then two liver cleanses cured her shoulder pain, nausea and remaining pains. She started gardening again and immediately picked up hookworms and *Trichinella* again. But she learned to sanitize her hands with grain alcohol after washing away dirt and this kept her parasites in check.

Jessica Atkinson, a middle age school teacher, developed a pain in the right cheek quite suddenly. She also had pain over the right mid abdomen and right side at the waist but X-rays and scans showed nothing (she had been X-rayed three times). She struggled for seven years to stay employed. She was having severe pain attacks over the liver and described her stool as almost white after these attacks. She cleaned her liver at least 30 times before she related, one day, that her joy in living had returned. Her gallstones were exceptionally large (½ x ¾ inch). Eventually the abscesses in her upper teeth were found, clearing up her cheek pain and protecting the liver from recurrent infections from these bacteria. Only then did she get permanent and complete pain relief.

Lisa Mattie, 72, had her right arm hanging limply by her side. It was so painful she bent forward to let it hang straight down. But in seven months she had done 6 liver cleanses, getting over 3,000 stones out. All pain was gone although some numbness in that arm persisted. She could also stop using Tums^{TM}, stop coughing, and no longer was bothered by her hiatal hernia.

Elbow Pain

One variety of elbow pain is due to an inflamed tendon there; it is sometimes called “tennis elbow.” It is not due to playing tennis or any other arm use. The inflammation is caused by a liver full of stones and parasites, especially flukes which manufacture a chemical that affects tendons. Kill all flukes and cleanse the liver for quick relief. Using your elbows while they are inflamed is traumatic to them, like working with a sore thumb. Don't play tennis or do other arm exercises until they are pain free.

Parasites consume large amounts of your vitamins and minerals. Give yourself vitamin A (25,000 u daily), zinc (60 mg. daily), and B_6 (250 mg twice a day) until the pain is gone.

Wrist Pain

Tendons passing through the wrist can become inflamed from the unnatural chemicals produced by fluke parasites in the liver. Using the wrists to work further traumatizes them (injures them) making it harder for them to heal. A small hole between the tendons lets the nerve and blood vessels through into the hand. Fluke parasites also make chemicals that thicken tendons. When tendons at the wrist thicken, they can squeeze down on the nerves and blood vessels until the hand or fingers feel numb. If you have pain at the wrist or numbness in your hands, killing parasites and cleaning the liver may give you the permanent cure. Wearing a wrist bandage or support can help reduce trauma damage to the wrist while it is healing.

Numbness of hands, without wrist pain, is more often due to a brain problem with parasites and pollutants. Lead, mercury, fluke parasites are the usual culprits.

Thumb Pain and Hand Pain

can be due to liver parasites. Get yourself ready for a liver cleanse. If the pain goes away beforehand, while you are on the kidney cleanse, it shows you had deposits in your joints. You were headed for arthritis in your hands. Read the information on arthritis (page 78) to protect yourself.

Finger Pain

This is pain in a joint, often accompanied by some enlargement or knobbyness of the joint. It is not hard to recognize these as deposits of the same kind as we saw in the toes. You can test yourself to identify the variety. **Uric acid** and **phosphates** are the commonest types. Read the section on toe pain (page 55) for detailed instructions. You can greatly reduce your finger joint deposits and the size of the knobs. In six weeks after starting the kidney cleanse and changing your diet, the knobs may already be shrinking. A large magnet (5000 gauss—used only as directed) may bring pain relief but only dental cleanup and environmental cleanup will give you lasting improvement.

Back Of Neck Pain

The back sides of the neck seem to be highways that run between the teeth and liver. Both contribute to pain at the back of the neck. Pulling an infected tooth or cleaning a cavitation can bring complete relief, only to return the next time a tooth is extracted. Extractions should be followed by cleaning out the cavity created so an infection can't start here. Cleaning the liver can also bring immediate relief, only to find pain and stiffness to return months later. You must cleanse many times for permanent relief.

An allergic reaction to potatoes and tomatoes can express itself in neck pain too. When the liver can no longer detoxify the chemicals (solanine, etc.) in this food family they are free to roam the body with the circulation. Perhaps they prefer to attach themselves at a particular neck site and cause inflammation here. Perhaps an injury was already there, beforehand. Whiplash is often blamed for back-neck pain and indeed chiropractic adjustments can bring total relief. Perhaps the trauma of whiplash first invited all of these contributors. Merely killing bacteria with a zapper is not long lasting. But dental cleanup plus liver cleansing is.

Front Neck Pain

Lymph nodes under the jaw strain your body fluids of the head, removing bacteria and toxins. They are sometimes called “neck glands.” If the stream of bacteria is endless such as when they are coming from a hidden tooth infection, the lymph nodes will enlarge to do a better job. They will also try to remove toxic metals, mouthwash, and toothpaste for you.

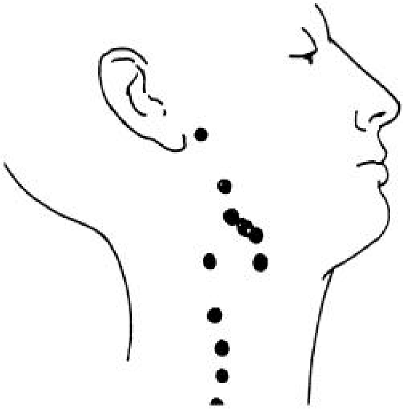

Fig. 26 Lymph node neck glands

To see what is affecting your lymph nodes, purchase a slide of lymph nodes. Since we have lymph nodes in many locations in the body, you can't single out the neck nodes for study. So you will get to see all the toxins affecting lymph nodes everywhere: PCBs in your underwear being removed by groin lymph nodes, lead in your intestinal lymph nodes from the water you are drinking, mercury in your neck lymph nodes from amalgam fillings. Eliminate all these. Don't rest as long as any of your glands are enlarged.

Roland Sanford, 23, had minor pain and a lot of stiffness along the sides of his neck. His arms had some numbness. He only had one metal tooth filling but his whole body was toxic with samarium, beryllium, indium, copper, cesium, and mercury. When it was replaced, his neck pain and numbness got better.

Audrey Doyle had severe neck pains she attributed to sitting all day and sleeping in her wheelchair. She had to take pain killers to get to sleep at night because they were so bad. She knew eating cream and butter made it worse but she had no will power, she said. After a liver cleanse, getting “thousands” of tiny stones, she was free of it.. for one day! But it was enough to convince her and she was determined to be off pain killers.

Temporal Mandibular Joint (TMJ) Problems

Pain at the angle of your jaw is not due to a poorly fitting joint. It fits perfectly but the muscles nearby are pulling it out of joint with their frequent spasms. There is always a hidden tooth infection present. Ask the dentist to search for hidden tooth infections and to clean your cavitations (you will need to find an alternative dentist, and read Dental Cleanup, page 409). Begin immediately to heal these bone lesions with vitamin D (40,000 to 50,000 units once a day for 3 weeks, followed by 2 such doses per week forever), milk-consumption for calcium, and a magnesium oxide tablet. (See bone healing, page 87).

Kill pinworms with a zapper or frequency generator in yourself and family members twice a week and keep everyone's fingernails short for extra hygiene in the family for a month to prevent reinfection. Pinworms can cause tooth grinding and contribute to TMJ dysfunction. For extra muscle relaxation, take two magnesium tablets at bedtime and valerian capsules.

Tooth Ache

Before the pain becomes acute and excruciating, kill bacteria of the mouth, including “tooth decay” and “tooth plaque” frequencies (see frequency list, page 561). Make your dental appointment immediately. If zapping bacteria several times in a few hours relieves pain enough to get you through the night or past the weekend, do not delay a single day. Zapping does not reach into the middle of an abscess—it circles around, so some bacteria are left to repopulate.

Try to understand the problem. If there are teeth with root canals near the location of pain, extract them. The toxic materials of the root canal jeopardize your total body health. Removing them always helps and may let the jaw heal normally where they were extracted. Since the pain is caused by a bulging infection pressing on a jaw nerve, and because each tooth has a related organ(s) it is especially important to clear up all infections to protect these organs.

These relationships are not understood yet. Finding that teeth have the same tissue frequency as some distant organ sheds a little light on the situation. Until the meaning and function of these frequencies is understood we can only guess that they interact somehow. Bacteria have taken advantage of this common resonance and have invaded both organs.

Heart infection by *Staphylococcus aureus* is an example. It is a common heart bacterium, causing much of our heart ailments, particularly mitral valve disease and irregularity problems. They always derive from teeth, whether present or missing! After wisdom teeth are extracted, the hole left in the jaw frequently does not heal, it picks up *Staphylococcus aureus* from the mouth and a chronic infection is started. For this reason it seems ill advised to pull wisdom teeth if it isn't necessary.

Another heart bacterium, *Nocardia*, can originate in the teeth. *Nocardia* can also invade the brain and nerves. Persons with Parkinson's disease always harbor quite a population of *Nocardia*. Zap all these bacteria before and after your dental visits. Remember that killing bacteria in itself is not a panacea. Pain and body damage will return unless you do the proper repair and cleanup work.

Throat Pain

Recurring sore throats are always improved by removing dental metal and root canals, and by cleaning up hidden tooth infections. This is essential for a permanent solution. Bacteria and viruses that cause sore throats are thriving in hiding places under and around dental metal. It is quite difficult to reach the center of such places (abscesses) with electric currents. Even if you could, you would reinfect the very next time you ate non-sterile food! Zapping is not a substitute for cleaning up your dentalware.

Keep your neck warmer than would be perfectly comfortable, gargle with hot water to bring in extra circulation. Use zinc tablets (60 mg. see *Sources*) twice a day; suck on them before swallowing. After cleaning up dentalware you may never have another sore throat!

Children with recurrent sore throats usually do not have a dental problem. They invariably have an air pollution problem. Asbestos, formaldehyde, fiberglass, Freon and arsenic are the chief culprits. Clean these out of the house. Eating moldy food is especially conducive to respiratory illness. The immunity drop can last for weeks! Go off the moldy suspects to raise your family's resistance.

Eye Pain

is caused by parasites.

The eye is a favorite location for many parasites, including *Giardia, amoebas, hookworm, schistosomes, Toxoplasma,* and innumerable others. The eye has two large bodies of watery fluid: the *aqueous humor* and *vitreous humor*, where no blood traverses to bring in extra white blood cells when the need arises. It has its own protective devices, like tears, but some parasites, like *Toxoplasma*, have overcome all the body's

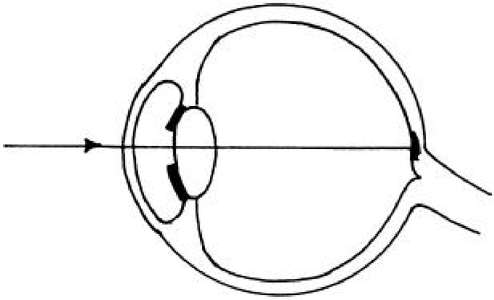

Light travels through the cornea, through the aqueous liquid, the lens, and then the vast vitreous humor, finally striking the super sensitive spot on the retina, called the macula.

Fig. 27 Eye.

defenses and regularly invade the eye if they get into the body somewhere.

We get *Toxoplasma* from cats; the infectious stage is in cat feces.

Toxoplasma infection could be the beginning of a lifetime of eye disease due to weakening of the eyes at an early age. *Toxoplasma* also invades the brain, frequently causing a dull ache or pressure at the back of the head. *Trichinella* is another common eye parasite, invading the eye muscles, so that a muscle is weakened and doesn't allow perfect alignment of the two eyes. Dogs and cats are the source of these.

There should never be a cat box in the house. Never let a child near a cat box.

By killing all the large parasites plus a few bacteria (*Staphylococci*, *Chlamydias*, *Neisserias*) the eye can become pain free in a few days. If pain returns, you missed something or reinfected yourself. Everyone in the family including pets needs to be treated for all the parasites. No indoor pets should be kept by a person of low immunity, since infecting yourself daily and then killing parasites daily is not a solution.

Herpes of the eye is not a single actor. The true host of this virus is a larger parasite, possibly a tapeworm stage. Tapeworm stages should be killed with an herbal preparation, Rascal, or with a zapper. (Not with a frequency generator. Only a zapper can kill all the segments and eggs at once, leaving nothing alive to wander about and find a new tissue to invade.)

Mary Rauch, age 60, filled a page with different pains and problems. Even her eyes had a dull ache around and behind them, sometimes reaching to the back of the head. Her teeth hurt when she ate and she had been getting B_{12} shots once a week for 36 years for pernicious anemia. She was full of *Ascaris*, amoebas and pinworms which kept her legs twitching and jerking in bed at night, even waking her up. She was so much better after the kidney cleanse and parasite program she was eager to cleanse her liver. Before two months had passed her eye pain was gone.

Jessie Healy, middle aged and in good health otherwise, had carried the anxiety of having inherited *retinitis pigmentosa* for forty years. Now her drivers' license was in jeopardy due to fast progression of her disease. Numerous heavy metals had accumulated in her retina, including cerium from dental floss, arsenic from pesticide, tin from health food brand deodorant, PCB from skin salve, cobalt from dish detergent and indium from tooth metal. She had eight parasites in the retina including *Toxoplasma* from association with cats years ago. Removal of dental metal alone arrested the disease process. Two years later she was slightly improved and still driving her car.

Macular Degeneration

Dolores Bollapragada, 50ish, had suffered from clinical ergot poisoning in the past which put her in a coma for several weeks. Ergot is a grain (especially rye) fungus, very toxic to the liver. Its byproducts are used in migraine medicine. She had overdosed, somehow. Perhaps her liver never recovered. It no longer detoxified solvents for her, allowing them to accumulate in her retina. There she had propyl alcohol, benzene, carbon tetrachloride acetone, butyl nitrite, styrene, gasoline, wood alcohol, paradichlorobenzene (moth balls), pentane, methylene chloride and decane. She stopped snacking on cold cereal through the day. She had many habits and products to change. But she was determined to salvage her sight. She did.

Headache and Migraines

Headaches can take the joy out of living. They can lower your motivation level so low that you don't even want to do anything about them! People with daily headache deserve our deepest sympathy since they must carry on with life's daily duties in spite of it. Often, no medicine helps. Although there are common headaches, sinus headaches, migraines, and others, the causes of all overlap a great deal.

Tooth infection, urinary tract infection, bowel problems, and a wormlet, *Strongyloides* are the common causes.

There are also allergic connections which include milk, eggs, citrus and salty foods. But the allergy-approach is difficult to put into practice. It is almost impossible to stay off these foods for long time periods. Eliminating bacteria and *Strongyloides* leads to a permanent solution.

Possibly the parasite wormlet comes first, since even young children can suffer with migraines. *Strongyloides* is a microscopically small threadworm that horses are plagued with, but humans and our pets pick it up easily. Every migraine sufferer (100%) has high levels of this tiny worm. Perhaps it is really the special bacteria it brings with it that cause the blood vessels to seep or to spasm in the brain, causing pain. Zap it immediately.

Bacteria hidden under a tooth filling or root canal or in a space where once a tooth was pulled can be the cause. *Staphylococcus aureus* is a favorite, but various *Clostridia*, *Streps* and others are often seen, too. You don't feel pain from this small abscess because the pressure isn't building up inside it. It has found a way into your body–namely, your brain!–so no pressure builds up to alert you. Find an alternative dentist with experience cleaning cavitations and finding small hidden abscesses. If you get immediate relief from dental work, only to lose ground again later, the abscess may have formed again (they are notorious for forming again). Go back to the dentist. Irrigate the dental wound site with white iodine (potassium iodide, 12 drops) or Lugol’s (6 drops) to ¼ cup water using a curved-tip syringe.

Do not use iodine if you know you are allergic to it.

The colon is always a thriving mass of bacteria. It should be emptied at least twice a day to keep their levels down. Use the simple herb, *Cascara sagrada*, senna tea, or magnesium oxide tablets (2 to 3 a day) to help you eliminate frequently if your own regularity is lacking.

Gallstones in the liver and a congested liver are sources of bacteria, too. Do several liver cleanses and note the effect.

Is it the toxins made by bacteria or the inflammation from the bacteria or wormlets themselves that produces the headache? Certainly, one can eat the toxins by themselves in foods like yogurt, cheese, wine, sour cream and develop "royal" headaches. Stay away from these **tyramine** containing foods. Boil all your dairy foods to prevent *Salmonellas* and *Shigellas* from swimming into your beleaguered brain. Don't eat dairy foods that can't be boiled for ten seconds.

Begin by killing all *Strongyloides* and other parasites, bacteria and viruses with a zapper. Hopefully, this will only leave a few stragglers behind in abscesses, gallstones and the colon contents. If the colon bacterium, *E. coli*, is your headache cause, start the Bowel Program. Search for the source of your *E. coli* in food or polluted water. Stop reusing water bottles. Most people get their *Strongyloides* back in a few days from pets, other family members, and themselves! Zap every day for three weeks. Try to "clean up" family members and pets. Never let a horse or pet salivate on you. Never put your fingers in your mouth. Always sanitize your hands with grain alcohol after using the bathroom.

To summarize the steps to cure migraine headaches:

1. Kill *Strongyloides* and bacteria by zapping.
2. Do the herbal kidney cleanse.
3. Clean up dentalware.
4. Do the Bowel Program.

5. Boil all dairy foods.
6. Do liver cleanses.

This has never failed to substantially reduce migraine intensity and frequency.

Headaches are also caused by toxins in your environment; especially things you breathe in. Household gas is the most common offender. You don't smell it after you get used to it! Gas pipes are notoriously leaky. Conducting gases through pipes with joints in them, where gases could escape, must be the most ludicrous of all modern “conveniences”. Would you try to conduct water through pipes with holes in them? You would soon see the water on your kitchen or bathroom floor. But gas doesn't land on the floor, it doesn't make a puddle, so you, the consumer, are left helplessly believing you don't have leaks. Every gas pipe that has a seam should have a clear plastic boot around it containing indicator compound to let it be known when gas is escaping. Other methods could be invented to make the gas utility safe. As you will see from the case histories, very many persons are living in a cloud of poisonous gas.

The telltale element is **vanadium**. When your vanadium test is positive, you have a gas leak that your body found, even though the gas company may not. Many gas companies use outmoded equipment to search for it. Four out of five gas companies miss the gas leak. Health Departments and building contractors use modern equipment that detects even the tiniest leak; call them.

If you are a headache sufferer, fixing the pipes is not good enough. Switch from fossil fuels to electric. There will be many dividends. (Remember not to use candles as part of your lifestyle, nor to be a smoker.) Your air needs to be clean: clean of perfumes and colognes, clean of potpourri and air fresheners, clean of air sprays, pesticides, formaldehyde, freon, carpet and car odor, fireplace lighter, and of wood smoke.

Janette Donovan, age 52, had migraines every day but could not tolerate any medication for it. She walked about "like a zombie" most of the time. She frequently had vomiting with them. She was gassy, had pain in her right groin for many years and chronic bladder leakage. Our kidney stone test showed urate crystals. Her urinalysis also showed urate crystals and a slight amount of blood, obviously chronic urinary tract infection. The parasite test showed *Strongyloides*, *Trichuris* and *Fasciolopsis buskii* in the intestines. Her heavy metal test showed beryllium. This was tracked down to hurricane lamps that once held coal-oil in them— it filled her house air unbeknownst to her. Thirty five days later she had done an herbal parasite killing program, done the kidney cleanse, cleared out the hurricane lamps and all fuel containers. Her headaches were "more like pressure" now than pain. She still had *Strongyloides*. Four months later she had some headaches still but not migraines. They were down to once a week. Three months after that she had done a liver cleanse and gotten abut 3,000 stones out! She was still getting some headaches. She tested positive to *Histoplasma* and *Coxsackie virus #4* (a common brain virus) probably stemming from dentalware. She had many root canals and couldn't decide what to do. Keep her teeth and risk return of migraines; or get partial dentures. After eight months of indecisiveness she was back up to ten headaches per month, although not migraines.

Patsy Olsen, age 30, had migraines daily, frequently with vomiting. She had *Strongyloides* as well as *Ascaris*, and other bowel parasites as did her husband and two children. She couldn't tolerate any medication and had to give herself an enema daily for constipation. The whole family was put on parasite killing herbs. A gas leak was found and fixed and the baby's diapers were changed to non fragrant. Ten months later, the whole family still had *Strongyloides*. By one year she was experiencing a couple of good days a month although she still had *Ascaris*, *Coxsackie* viruses, and various tooth-related bacteria. Her two and a half year old had swollen neck glands, was toxic with bismuth from disposable diapers but did not have *Strongyloides*. The eight year old was also toxic with the lotions and fragrance of baby-stuff in the home; she was constantly congested and coughing but became free of *Strongyloides* in six months. After another ten months (the baby had been potty-trained meanwhile) the mother began to have headache-free days.

Donald Schaible, age 14, had migraines, learning disabilities and severe acne. The parasite test showed *Strongyloides*, *hookworm* and *Ascaris*. In spite of being on the parasite program two weeks and zapping, he still had *Strongyloides*.

Kenneth Jones had migraines for thirty five years and had tried all the new medications. They worked for a while, then stopped helping, but he continued taking them anyway. He usually went to the emergency room for the really bad ones, once a week but lived with the constant daily variety. There were two house dogs. They and the whole family had *Strongyloides*. After cleaning up an asbestos problem, killing parasites for five months and clearing kidneys of urate stones, he was down to two to three mere headaches a week. Two months later, he was getting migraines again; they all had *Strongyloides* again. With renewed efforts, one month later his bad headaches were down to one a month, although his low level chronic headache persisted: they had the dogs on a strict schedule of parasite killing herbs as well as themselves. He had not been to the emergency room for a month.

Angelina Gander, age 46, had daily headaches, not migraines. She also had persistent urinary tract infection and sinus infection. She was put on the herbal parasite program and four weeks later was much better. She also lost her chest pain due to heartworm and regained her milk tolerance.

Gracie Arlington had a boy age 6 who wet the bed, a girl age 8 with a behavior problem at school. She was stressed by an unfaithful spouse and thought she should go back to school for a Nursing degree so she could support the family. But she was getting two or three migraines a week in addition to colitis attacks which she feared would make her unable to study. The two cats, a dog, the children and herself all had *Strongyloides*, *Ascaris*, and a variety of other intestinal parasites. The humans were promptly zapped for parasites and the boy was dry that night for the first time in his life. A few nights later he was wet again. This time the animals were zapped and put on the pet parasite program and the children's toileting was carefully supervised. When she dropped her extreme vigilance over all, they all relapsed. After a year of trying, they gave away their beloved dog, put the cat box in the porch and the mother did the cleaning herself but nothing prevented relapses. A half year later, her six year old son asked if he could launder his own sheets and blankets. This request broke her heart and she planned to give away one cat. Even with only one cat, the girl's behavior and grades fluctuated extremely, the mother had a migraine a week and the boy wet. During a vacation the mother loaned the last cat, (her cat) to which she was very endeared, to a friend, to see if it made a difference. The boy never wet again, the girl made straight A's, and the mother's headaches became sinus headaches. Some intensive dental work cleared these up. She brought her Nursing diploma to our office two years after beginning. She should have had another diploma for **Intelligent Parenting.**

Earache

Earache is particularly common in children. Bacteria, mainly *Streptococcus pneumoniae*, have built up to a high level. Zap them. They were probably introduced by some larger parasite. But why did they multiply and thrive in your child's ear? There must have been food for them and protection from the immune system. Probably the Eustachian tubes are full of mucous, providing habitat. The mucous is present because some air toxin is irritating the sensitive linings. Clean up the air: not just asbestos, fiberglass, formaldehyde, freon, and arsenic, but perfumes, fragrant school supplies, potpourri.

Being housebound, as in winter, makes the air-toxin problem worse. Use summer time to strengthen your child for the winter—spend summer outdoors. Always wash hands a lot. Stay away from moldy food. Give small doses of niacin (25 mg., niacin thins mucous so it can run away) along with vitamin C (250 mg.) at bedtime.

Putting in “tubes” is a short term solution that is better than perpetually staying on antibiotics, but neither should be necessary.

To hasten healing, as soon as earache is suspected put room temperature olive oil in both ears, tug the earlobe to let bubbles out, and stopper them with cotton wool salvaged from vitamin bottles (other types are polluted with mercury). Put on a light bat that covers the ears.

When adults get earache, the *Streps* are hiding under tooth fillings and in gallstones. Clean up these two sources and zap.

Drinking non sterile milk adds *Salmonellas* and *Shigellas* to the *Strep* ear infection! Small wonder milk is thought to make people mucousy. The ears must now battle them, too. Even a few bacteria consumed in milk can start a whole colony if bowel bacteria have been disturbed by antibiotics. If your child has become “allergic” to milk or gets mucousy, remove cheese and ice cream–not milk–from the diet. Even cheese can be eaten if baked as in pizza or casseroles. Milk must be boiled. Try to reduce the child's *Salmonella* sensitivity by avoiding unnecessary antibiotics. Keep immunity high by avoiding moldy food!

Ear Noises And Ringing

This is also called *tinnitus*. A hissing or buzzing sound is heard in one or both ears. It can be continuous or pulsing off and on. It can be as loud as roaring.

Tinnitus is caused by three things acting in partnership: toxic elements, an allergy to *salicylates* (the aspirin family) and a bacterium *Streptococcus pneumoniae* (the pneumonia bug). This “bug” can be carried in the chronic state after a bout of pneumonia or what seems to be a head-cold. It is always present in earache. It is easily activated by exposure to cold wind or drafts and certain toxic substances. For this reason I recommend keeping the ears warm in winter by keeping them covered or wearing a cotton plug or ear-muffs. The *Strep* bug can also cause *Meniere's syndrome*, congestion, loss of balance and chronic sinus problems. Often an antibiotic of the penicillin family helps immediately, then loses its effectiveness. This shows you that bacteria are involved but cannot be truly vanquished with antibiotics.

Streptococcus pneumoniae often hides in pockets under infected teeth and in holes left where teeth have been pulled—especially wisdom teeth. These can be found by alternative dentists who clean these cavitations. *Strep* also resides in the liver; clean them out with liver flushes.

Certain foods and many medicines, especially cough medicine and lozenges contain salicylate.

To cure the tinnitus we must stop using aspirin or any high level source of salicylates. We must also stop the exposure to certain toxic elements—lead, beryllium, zirconium, benzalkonium. They are present in the air at gas stations and in many of our body lotions, soaps and salves. Switch to safe varieties. Niacin tablets, such as 100 mg, taken three times daily is another treatment, intended to increase the blood supply to the inner ear. Taking a thyroid tablet, by prescription, often helps too.

These methods never fail to improve tinnitus but a complete cure is seldom possible.

William Thall, 47, had a headache daily and was on pain medicine for it daily. He had tinnitus, a loud humming noise in the right ear. He had *Shigella* (producing nerve toxins) and *Borellia* (Lyme disease virus). He had *Gaffkya*, a respiratory bacterium under two wisdom teeth (right side). He also had *Campylobacter* and *Anaplasma* growing on him somewhere. He had *Strongyloides* too. He was toxic with vanadium (gas leak) and asbestos. After ten weeks he had two cavitations on the right side of his jaw cleaned, he had been on the parasite program and his tinnitus was gone, (he still had occasional headaches indicating he still had some *Strongyloides*).

Billie Scott, 26, had a history of ear problems, and lots of headaches. She was started on the kidney cleanse. Then she added the parasite program including their two dogs. She had a cyst in her left buttock, due to PCBs traced to the well water. She stopped using this well and switched off detergents for laundry and dishes. She switched off soda pop, onto milk. In 3½ months her tinnitus was gone and the cyst was shrinking.

Larry Pelegrini, 59, had tinnitus in both ears but was otherwise a strong, tall, intelligent person, who cared for sick people, whether family or not. As soon as he saw how simple it was to kill invaders with a frequency generator, he bought one. Preventing their recurrence was his big challenge since he had neither the means nor insurance to do dental work. He was toxic with aluminum, copper and PCBs. After switching to borax for all washing purposes, he got rid of aluminum and could feel his memory improve. The kidney herbs cleared him of uric acid and oxalate and he felt more supple. He had to go off his favorite beverage to get rid of pentane and methyl ethyl ketone. Twice a week he killed two dozen parasites and bacteria, that just seemed to pop up from nowhere, in order to feel better and reduce his tinnitus. But he lived alone, had to cook, garden, take care of animals and his sick friends which gave him a lot of parasite exposure. Sometimes he would be toxic with arsenic (a new pesticide he tried out) or vanadium (gas leak) but mainly it was tooth filling metal. If only this wonderful man could afford his dental work: what a blessing to society he could be for a long time to come.

Scalp Pain

Infection anywhere in the head can cause sensitive scalp and scalp pain. Even a common cold can cause bouts of scalp pain. Clean up your dentalware and environment.

Lumps on the scalp are often called *sebaceous cysts*, actually this is where your body has sequestered PCBs (polychlorinated biphenyls, extremely hazardous, now banned in the U.S., but often a pollutant in detergents.) Get rid of them by switching off detergents for all purposes. See *Recipes* for dishwasher liquid, dishwasher detergent, and laundry detergent replacements.

Non-Painful Diseases

Explanations for many of these "mysterious" diseases become rather obvious when you see a **common parasite, or pollutant, or both** consistently show up in case after case.

Diabetes

All diabetics have a common fluke parasite, *Eurytrema pancreaticum*, the pancreatic fluke of cattle, in their own pancreas. It seems likely that we get it from cattle, repeatedly, by eating their meat or dairy products in a raw state. It is not hard to kill with a zapper but because of its infective stages in our food supply we can immediately be reinfected.

Eurytrema will not settle and multiply in our pancreas without the presence of **wood alcohol** (methanol). Methanol pollution pervades our food supply—it is found in processed food including bottled water, artificial sweetener, soda pop, baby formula and powdered drinks of all kinds including health food varieties. I presume wood alcohol is used to wash equipment used in manufacturing. If your child has diabetes, **use nothing out of a can, package or bottle** except regular milk, and no processed foods.

By killing this parasite and removing wood alcohol from the diet, the need for insulin can be cut in half in three weeks (or sooner!)

Be vigilant with your blood sugar checks. The pancreas with its tiny islets that produce insulin recovers very quickly. Even if 90% of them were destroyed, requiring daily insulin shots, half of them can recover or regenerate so insulin is no longer necessary. The insulin shot itself may be polluted with wood alcohol (this is an especially cruel irony—the treatment itself is worsening the condition). Test it yourself, using the wood alcohol in automotive fluids (windshield washer) or from a paint store, as a test substance. Try different brands of insulin until you find one that is free of methanol.

Artificial sweeteners are polluted with wood alcohol! Instead of helping you cope with diabetes, they are actually promoting it! Do not use them.

Drugs that stimulate your pancreas to make more insulin may also carry solvent pollution; test them for wood alcohol and switch brands and bottles until you find a pure one. You may not need them much longer, so the extra expense now may soon reward you.

Many persons can detoxify the amount of wood alcohol that pollutes our foods. They do not have a food mold, **Kojic acid**, built up in their bodies as diabetics do. I have found Kojic acid in coffee, and potatoes with gray areas inside. Do not eat discolored potatoes or peels, even if cooked or baked. Being able to detoxify a poisonous substance like wood alcohol should not give us the justification for consuming it. All poisons are bad for us. Do not consume them.

All diabetic persons also carry a virus, *HA virus* in the pancreas. This virus grows in the skin as a wart but is spread quite widely in the body such as in the spleen or liver besides pancreas. It is not necessary to kill this virus since it disappears when the pancreatic fluke is gone. The HA virus undoubtedly belongs to the pancreatic fluke. The question can be asked: Does the fluke or its virus cause diabetes? There might even be a bacterium, so far missed in our observations, that is the real perpetrator.

While recovering from diabetes, it is very important to check your blood sugar every day. Improvement is so rapid, you may suddenly be over-insulinized by your next shot. Cut down your dose to suit your actual need.

There are additional aspects to diabetes that have been studied by alternative physicians. For instance, allergy to wheat and other grains containing gluten is common. Perhaps the pancreas and its islets would heal much faster if grains were out of the diet for a while. Perhaps the 50% improvement that is consistently possible just by killing parasites and stopping wood alcohol consumption could be improved further by a month of grain-free diet. Eating **fenugreek seeds** has been reported to greatly benefit (actually cure) diabetes cases. Are they a specific fluke killer, virus killer, or neither? It seems like a good idea to add this to your diet if you are a diabetic. Wood alcohol also accumulates in the eyes, and there is a connection between diabetes and eye disease. **Bilberry leaves** are an herbal treatment for both diabetes and weak eyes. Do they help by counteracting wood alcohol or detoxifying Kojic acid? Make a tea for yourself, using ¼ cup leaves to three cups water. Drink ½ cup a day. Chromium is another must for diabetics (200 mcg three times a day). It helps insulin enter your cells.

Gold is attracted to the pancreas. Heavy metals should be removed from dentalware including all gold crowns and no metal should be worn next to the skin as jewelry, including all gold items.

Blythe Jenkins was on Micronase™ (5 mg) daily for her diabetes but she still had a morning blood sugar of 183. She had pancreatic flukes and sheep liver flukes in her pancreas, vanadium (a gas leak) in her home and cadmium in her water (old pipes). After killing parasites and cleaning kidneys her morning blood sugar was down to 148. She also got rid of her hot flushes, groin and leg pain. This encouraged her so much she did the rest of her body cleanup and could go off her medicine completely.

Robert Greene, age 65, had been on insulin five years already, getting two shots a day (25 u each), and even this was not controlling his blood sugar which was 288 in the morning. He was raised to 30 units without much improvement. His legs and feet were too painful to walk without a walker to lean on. He had *Loa loa* in his pancreas. (This is another example of a parasite going to the "wrong" location, this time the pancreas instead of the heart. I believe the solvent makes this possible.) Also *mumps virus*, *HA wart virus*, *mycoplasma*, *flu virus*, *chicken pox virus*, *Adenovirus* and *Shigella*, all making their home in his pancreas! This was possible because he had wood alcohol accumulated there, from drinking various beverages and using artificial sweetener. As soon as he stopped this practice and killed everything with a frequency generator his blood sugar fell below 100 in the morning and he had to reduce his insulin to 20 units. He was also on chromium 200 mcg (2 three times a day). Meanwhile, he cleared 2,000 stones out of his liver. He could now walk well again and decided to take a part time job. The whole process took one year.

Ralph Dixon, age 72, had been switched to 30 units of insulin, once a day, after six years on pills for his diabetes. He still had a fasting blood sugar of 242. He had the pancreatic fluke and a host of bacteria and viruses in his pancreas. He had a poodle. After killing the pathogens and cleaning his kidneys, his blood sugar dropped so he cut his insulin to 25 units (blood sugar was at 111) Soon he had to cut it to 20 units. He was also on chromium 200 mcg (2 three times a day), manganese 50 mg. (one a day), fenugreek capsules (2 a day), and bilberry leaves, each supplement by itself. He also lost his angina and fatigue with all this body-cleaning. But if he went off the maintenance parasite program he would promptly get a spike in his blood sugar, showing how easy it was for him to reinfect and how new parasites would immediately find his pancreas.

Melissa Bird, 54, had major illnesses including heart disease (2 angioplasties), numerous other surgeries and diabetes. She took 15-18 units of insulin once a day. Besides pancreatic flukes in her pancreas, she also had intestinal flukes there(!) plus widespread *Ascaris*. She also had warts and candidiasis. Fortunately there were no pets. Her parasites were instantly eliminated with a frequency generator and she was started on kidney herbs for her other problems. Seven weeks later she stated she had to cut down her insulin because her morning blood sugar had dropped to 90. Then she eliminated the decafs and artificial sweetener that were giving her wood alcohol, started the parasite herbs and did a liver cleanse. The day after the liver cleanse her blood sugar went up to 164 but was completely normal after that (under 100) and she did not dare take any more insulin or pills. We advised her to keep monitoring her blood sugar and be very, very vigilant and to please stop smoking.

John Angert, 65, was not on any medicine in spite of his diabetes. He was too embarrassed to tell us his blood sugar. He had a long list of other health problems, too. He was toxic with thulium (vitamin C) and palladium (from tooth fillings). He had a variety of parasites. After doing some dental work and parasite killing his fasting blood sugar dropped to a normal 98. But his leg cramps kept returning in spite of doing a kidney cleanse. Only after changing his diet to include milk did the phosphate crystals stay away and eliminate his cramps.

Cornelius Edens, age 33, came for his diabetes, although he also had fatigue, digestion problems, and headaches. He had numerous other minor symptoms like chest pain over the heart, soreness in testicles, etc. His diabetes was only diagnosed one year ago. he was on insulin (Humulin^{TM} 20 long acting plus 6 regular units) in the morning. He often omitted the evening dose which should be 14 long plus 6 regular. This brought his fasting blood sugar down to 166 in the morning.

He had pancreatic flukes in his testicles besides in his pancreas. His aflatoxin level was very high; he was told to stop eating grocery store bread, eat bakery bread only. He had silver, nickel and very high levels of gold–probably all three coming from his gold crowns–he was to have them all replaced with composite. He had fiberglass in his pancreas. Wood alcohol was present in large amounts in his pancreas as expected. He was to stop drinking all store bought beverages, whether frozen, powdered, or ready to drink. He did not test positive to benzene, propyl alcohol, *Salmonellas*, *Shigellas*, or *E. coli*. But he had *Staph aureus* growing at high levels in his pancreas. The usual source for this is in teeth, so he was told to ask his dentist to search for cavitations and clean up hidden tooth infections while replacing the gold crowns. He was to find the exposed fiberglass and seal it off. He was to zap parasites every other day for 1 month. His supplements were vitamin C (1 tsp. with each meal for constipation; then 1 tsp. daily), vitamin B-50 complex (2 per day), thioctic acid (2 per day). He was to start the Kidney Cleanse recipe for his testicle problem, and after 6 weeks do a Liver Cleanse. He was to follow up with us in 3 weeks, after dental work was completed. He never returned. Four months later we received a phone call he was too embarrassed to make himself. He needed no insulin and was doing fine. Another 3 months later he was still doing well, and off insulin.

Prediabetic

Alyce Dold, 64, came because she was worried about her blood sugar and chest pain. Indeed, a blood test showed her fasting blood sugar to be 136, just beginning to show insufficient insulin production by her pancreas. She had pancreatic flukes and wood alcohol there. Also *mumps virus* and *HA virus*. She had six more solvents accumulated due to eating raisin bran and other cold cereals each day. She was glad to be forced off this routine: she switched to 2 eggs every other day with biscuits or bread (not toast) and cooked cereal in between. Her chest pain was due to dog heartworm and *Staphylococcus aureus* bacteria that originated at teeth #16, 17, 1, 32. The worms and *Staph* were killed with a frequency generator. She was referred to a dentist for cavitations and started on kidney herbs. Two weeks later, there was still a little residual heart pain due to *Staph*; dental work was not yet done. She was given chromium (600 mcg per day) to help her insulin regulate sugar. Her LDH (See tests) was still high from the recent heart stress, no doubt. But she had accomplished a lot and planned to get it all done.

Diabetes Of Childhood

The problem is the same for diabetes of childhood as for diabetes of later onset, but much easier to clear up, provided the whole family cooperates.

Wesley Evanco, age 6, had his onset at age 11 months. Prior to that he had chicken pox and his baby shots. He had pancreatic flukes and their reproductive stages in his pancreas as well as wood alcohol. He had no other solvents accumulated. The problem was clearly due to these two enemies of his small body. Wesley's father did not have pancreatic flukes but his mother had very high levels, along with wood alcohol. She had to clear hers up before Wesley could recover. She chose not to. She couldn't believe the connection. Our most persuasive efforts were not successful. We hope Wesley can forgive us all in due time.

Herpes

The *Herpes* family of viruses includes *Herpes Simplex Virus* (HSV) 1 and 2, *Epstein Barre Virus* (EBV), *shingles or Chicken Pox* (*Varicella zoster*), *Cytomegalovirus* (CMV) and some newly discovered ones.

Herpes simplex virus 1 is the virus that breaks out around or inside the mouth. We call it a cold sore because it often follows a cold in childhood. As children, we may get HSV 1 once or twice each winter. When they come more often the child's immunity is low. Adults who get repeated attacks also have low immunity (this is obvious from a blood test where the white blood cell count is less than 5,000 per cu mm).

Herpes simplex virus 2 breaks out in the genital area. It is often blamed on promiscuous sex but I believe it has quite different origins.

I believe the virus is introduced to our bodies by another large parasite. Perhaps it is pinworms, or *Ascaris*. Perhaps it is a tapeworm stage, picked up from dust and dirt.

I have some evidence that it is released from dog tapeworm stages when these are being killed by your immune system. It is probably the same tapeworm stage as releases the *Adenovirus* (cold virus) which would explain why children frequently get "cold sores" during or after a cold. (See Curing The Common Cold, page 357.)

Nevertheless, you would not get sick from these viral releases unless your immunity was lowered.

Herpes lives in your nerve centers (*ganglia*) and it is from here that you can be attacked after the initial infection. Evidently your immune system can destroy them as quickly as they emerge. But a meal of **aflatoxin** or other **moldy food** suddenly "gags" your white blood cells and lets a viral attack happen.

The viruses can also be "triggered" which lets them out of hibernation (latency) to multiply and travel along the nerve fiber to the skin. Triggers are things that put these nerve centers to work: sudden cold and heat, trauma from chafing and friction. Never drink water with ice cubes in it. Never eat hot soup with a metal spoon. Don't tweeze hairs.

Begin your prevention program by raising the immunity of your skin; this means removing all toxins from the skin. Use only natural lotions, softeners, cleansers on your skin made from recipes in this book. Health food brands are not superior. This will get rid of nickel, chromate, titanium, zirconium, aluminum, and benzalkonium from your skin and probably your whole body! Do laundry with borax and washing soda, only, to eliminate commercial detergent as a source, too.

As soon as you feel that warning tingle or sensation of *Herpes*, zap or use a frequency generator at 293 and 345 KHz (HSV 1) or 360 and 355 KHz (for HSV 2). Immediately take a **cayenne** capsule and 8 **lysine** tablets (500 mg each). The cayenne slows down travel of the virus along the nerve.

Attacks probably occur when the triggers act at the same time as an immune drop occurs. Immune drops happen frequently due to eating moldy food. Of course, mercury in amalgam tooth fillings keeps immunity chronically lowered. Many persons report they stopped breaking out with *Herpes 1* after replacing their amalgams. Stopping wearing tight synthetic underpants helps reduce *Herpes 2* out breaks.

When you get an outbreak, mop up a droplet of the blister fluid and prepare it as a specimen for yourself. If you search for it in your white blood cells when your attack is over, it will not be found because it is in hiding inside your nerve cells. Zapping does not reach them inside your own cells. Nevertheless, you can totally eliminate them by repeated zapping provided you kill them at their earliest warning. Evidently, they haven't multiplied yet, so gradually their numbers go down.

Even after you have been *Herpes* free for a long time, stick to your preventive principles. Avoid the trigger foods, peanuts, and chocolate, or be ready to zap. Avoid cold wind or direct sunlight on your face. Don't eat abrasive or acid foods like popcorn, nuts, toast, crackers, candy, citrus. Although you may stop the virus in its tracks by zapping, healing the lesion takes time. Keep the skin softened with a cornstarch or sodium alginate recipe (see *Recipes*).

A lysine mush helps too: crush a lysine table with a large wooden spoon, add a pinch of vitamin C powder and a pinch of zinc oxide. Save part of this mixture for later use. Wet a small bit of it with a few drops of water to make a paste. Apply to lesions.

Bazezew Hailey, 38, started breaking out in the genital area after a period of antibiotic use. She was started on lysine (500 mg, 8 a day), cayenne caps (one with each meal). She was to fill her prescription for Zovirax™ from her doctor, though. By the time she had it filled, the next day, her lesion had stopped enlarging, and she could reduce her supplements. Her skin was full of mercury, lead, strontium, aluminum. Her white blood cell count was low, showing low immunity. Her ratio of segmental to lymphocyte white blood cells was low, evidence for a chronic viral condition. Her bone marrow contained lead and radon. She stopped using toothpaste (strontium), salt, deodorant, detergents (aluminum). She got the metal out of her mouth and eliminated her radon problem by opening crawl space vents. The lead source was never found, but after she moved to a new house it went away. *Herpes 2* stayed away. A year later she still hadn't opened the Zovirax™ bottle.

Fatigue

Fatigue, whether minor or extreme, is always associated with blood sugar disturbances. The more disturbance, the more fatigue. We have three organs that do most of the sugar regulating: our **adrenals**, the **liver**, and the **islets in the pancreas**. In severe fatigue, that keeps you partly bedridden, all three organs are heavily parasitized. *Epstein Barre Virus* (EBV) is running amok in your body, as a rule, even when clinical culture techniques cannot find it. With your frequency generator and Syncrometer you can find it emitting at 380 KHz. Killing the viruses is not as important as killing the larger parasites and getting your organs functioning for you again. The viruses will go away by themselves.

The liver's role in blood sugar regulation is to get it out of storage when needed. When sheep liver flukes have taken up residence they're spewing their chemicals as well as their own bacteria and viruses into your circulation; it is surprising that the liver has any sugar in reserve and can function at all. Sheep liver flukes are commonly seen in fatigue syndrome cases.

The adrenals (the outer layer called the cortex) help to regulate the blood sugar in a complex way. Some adrenal factors influence the thyroid which is another energy-related organ. Both adrenals and thyroid have toxic buildups in fatigue cases! Their work is hampered.

The heart of sugar regulation is in your pancreas in the tiny islands of cells that secrete insulin, called the *islets of Langerhans*. Here we always find the pancreatic fluke in residence. It actually breeds there when wood alcohol accumulates in it. Wood alcohol is a common pollutant of food, even in artificial sugar which is often recommended to replace real sugar and spare the pancreas. There is wood alcohol in store-bought drinking water, fruit juice, powders meant to be stirred into beverages, even if they are health food varieties. It's probably being used to clean tubes and hoses in the manufacturing plant. The only beverage you can safely buy (not safe unless you sterilize it, though) at a grocery store is milk. Make your own beverages using recipes in this book.

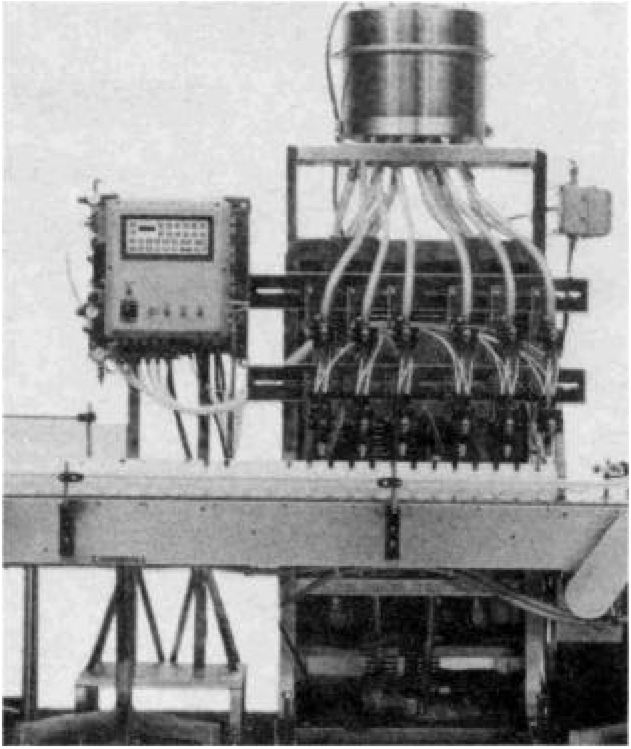

Fig. 28 Bottling equipment should be rinsed with ethyl (grain) alcohol, not propyl alcohol or wood alcohol.

Your first step toward curing your fatigue syndrome is to kill the pancreatic fluke and all other living invaders of the pancreas, liver, adrenals and thyroid. Use a zapper. Drink milk or buttermilk for several days afterward to provide lactic acid for the "good" bacteria to feed and recover on.

Your second step is removing metal from your mouth, particularly gold. Gold from teeth and jewelry readily goes to the pancreas! Do everything in the *Easy Lifestyle Improvements* chapter.

Your energy can bounce back in a few weeks by attending your liver, adrenals and pancreas. Help these organs heal and grow strong again. Avoid food molds (read Moldy Food, page 381). The most powerful assistance to the liver is a cleanse. This will eliminate liver viruses such as EBV and *Cytomegalovirus* (CMV). Be patient. Do it on a schedule until you have over 2,000 stones out. Take **vitamin C** at least 3 gm/day, to help both liver and adrenals. Take **B_2** and **B_6** to help adrenals and kidneys. Take glucose tolerance factor, **chromium** (two 200 mcg tablets 3x/day). See *Sources* for all of these. Take these supplements for three weeks, then cut the dose in half, and take on alternate days only, as a hedge against possible pollution in these.

Although your energy may be normal in three weeks, you are at higher risk for fatigue than the average person. Reinfection with anything will put the new parasites right back where the old ones were. Other bacteria, solvents and toxins will head for the pancreas, liver and adrenals again because these are weakened organs. It could take two years to build your health to its previous level, but is well worth it to have youth, initiative, and a beautiful appearance again. Going back to school is a good use of your time when your initiative has returned but your physical strength is still not up to housework or a job. When your energy comes back to you, it is tempting to overwork: to clean the whole house or to get into some gardening.

It is better to be moderate today than in bed tomorrow.

Anyone who has suffered from Chronic Fatigue Syndrome (CFS) or recurrent EBV has learned that lesson well.

June Timony, age 38, was diagnosed with Chronic Fatigue Syndrome), EBV and *Candida* around 1½ years ago by her family doctor. She also had a thyroid problem and a high estrogen level (165 pg/ml). She had severe depression at times. Our test showed her body was full of bismuth (fragrance) and silver (tooth fillings) especially in the ovaries. She cleansed her kidneys and killed parasites but could not make up her mind to do the expensive dental work.

Janice Brown, age 21, had EBV with chronic fatigue and depression along with a dozen more symptoms. Her skin, kidneys, breasts, brain, ovaries and pancreas were all loaded with mercury, platinum and other metals. She was full of radon and bus exhaust and the plumbing was shedding cadmium. Rather than do all this she and her husband decided to move. Before the moving date arrived she had cleansed kidneys, killed parasites and done dental work and was feeling noticeably better. Then they moved. She immediately was very fatigued again and worried that the move had been in vain. This time she had a liver full of Salmonella and a return of phosphate crystals in her kidneys. But it was easy to clear up and it was a very useful lesson to her to avoid unsterilized dairy products. She was much more careful after this.

Dee Safian, age 36, came especially for her low energy. She had EBV once. It left her extremely nervous. We advised going off caffeine but not substituting for it by drinking decafs. Her tissues were full of arsenic from pesticide; her urinalysis showed kidney crystals and her eosinophil count was high 5.5% (parasites). She had sheep liver flukes and stages in her pancreas due to a buildup of wood alcohol there. In four months after killing parasites and doing a kidney cleanse she was much improved.

Josefina Linzer, age 32, came for her fatigue and depression. Her tissues were full of arsenic, PCBs, chromate (eye makeup), mercury, yttrium, radon and terbium. She needed tooth metal replacement but could not schedule it immediately. She had *Ascaris* and pancreatic flukes in her pancreas and reacted to sugar in her diet quite strongly, so avoided it. She also had *Strongyloides* and *Trichinella*. She killed all these with a frequency generator and started on the kidney herbs. Her wood alcohol buildup was coming from carbonated beverages. She went off. In 6 weeks she had done everything except the mercury removal and was feeling much better.

Brigette Dawn, age 21, had Chronic Fatigue Syndrome with EBV, along with other problems. She cleaned her home and cleansed kidneys, killed parasites, and did two liver cleanses. Still the fatigue would return two days later. Meanwhile, though, her infertility problem got solved (she got pregnant) and this encouraged her to continue the battle against fatigue after the baby was born.

Hector Garcia, age 14, was getting gamma globulin injections every three weeks for his chronic fatigue syndrome. He had pancreatic flukes in his pancreas, sheep and human liver flukes in his liver and intestinal fluke in his intestine. He had a buildup of benzene, propanol, and carbon tetrachloride as well as aflatoxin from his granola breakfasts. He also had *Candida* and measles in his white blood cells. He killed parasites with a frequency generator and went off the solvent polluted items in the propyl alcohol and benzene lists. He immediately (in 20 days) felt much better.

Dana Levi, age 16, had chronic fatigue syndrome and dizziness; he was not in school. He had pancreatic fluke in his pancreas, sheep, human and intestinal flukes in his liver! Both benzene and propyl alcohol were present in his immune system. As soon as the parasites were killed (with a frequency generator) and he changed a lot of his products, he felt better but soon lost his improvement. At the next visit, our tests showed a buildup of vanadium (from burning candles in his bedroom). When he stopped this, he was better again for a while only to relapse again. His propyl alcohol level was up as was aflatoxin. But getting a taste of normal energy gave him the determination to get himself well! He embarked on a liver cleanse program and still more careful food selection.

Dennis Dillard, age 16, was beginning to do poorly at school due to fatigue. He had to stop athletics although his regular doctor pronounced him well. He was also getting chronic sinus infections. His lungs and trachea had accumulated seven heavy metals: vanadium, palladium, cerium, barium, tin, europium, beryllium. His fasting blood sugar was low (73 mg/DL) and LDH very low (90 u/L; it should have been 160). The body makes LDH in response to lactic acid levels. When the muscles aren't making much lactic acid from their normal metabolism, LDH levels will fall too. Was he developing a muscle disease? The gas leak was fixed (vanadium), the garage was sealed off from the house to eliminate barium and beryllium but the other toxic elements came from his dental retainer. As soon as his retainer came out, and they stopped using flea powder on their dog, his energy became normal and sinuses cleared up. A year later he still had not been ill.

Evelina Rojas, age 12, was having extreme fatigue with mood problems and sudden fevers. She killed *Ascaris* and sheep liver flukes with the parasite program but promptly got them back due to a benzene buildup I believe due to using products containing an herbal oil. Her high levels of *Streptococcus pneumoniae* (cause of fevers), *Staphylococcus aureus* and *Nocardia* could not be eliminated until her three baby teeth (with root canals) were pulled. After that, she was well.

Elaine Perkins, 48, came specifically for her low energy and nervousness. She was toxic with arsenic, a substance that replaces energy with nervous excitement and exhaustion. She also had a backlog of antimony (using baby oil), aluminum, rhenium (hair spray), benzalkonium (toothpaste) and radon. In four months, she had the arsenic and three other toxins eliminated and already had more energy.

Neil Youngblood, 53, was so fatigued, he had to brace himself even while sitting. A blood thyroxin level (T_4) of 1.0 instead of the normal 7.5 mcg/DL explained his fatigue. He had a mouthful of assorted dental metals which were accumulating in the thyroid, inviting viruses, particularly CMV, into it. He had to be on four grains of thyroid to feel near normal. After he had all his tooth metal removed, he only needed one grain to feel O.K. This encouraged him to clean up more of his body.

Scott Pennington, 50ish, was on thyroid medicine for his hyperthyroidism. He had been on iodine-radium earlier. He had the *miracidia* of the intestinal fluke, sheep liver fluke, and pancreatic fluke in his thyroid! Adult human liver flukes were also there! His thyroid was toxic with iridium, nickel, tellurium and mercury (metal tooth fillings) and decane, TC Ethylene and pentane solvents. He had been drinking a great deal of regular tea, which let oxalate crystals deposit in his kidney and slow down the excretion of toxins. The parasites were killed with a frequency generator, he changed his diet to get rid of solvents. In two weeks he was feeling better, had more energy and better sleep. This encouraged him to do the dental work.

Skin Problems

Sebaceous Cysts

Your body, in its wisdom, keeps toxins together and out of harm's way by making a cyst out of it. All the *sebaceous cysts* I have seen are filled with *polychlorinated biphenyls* (PCBs).

Get rid of PCB sources in order to clear PCBs from your body. This can take 6 to 12 months. Change all detergents (for dishes, laundry, and body use) to borax and/or washing soda.

Commercial detergents are wonderful cleaners...

but are unquestionably toxic. Whether you have cysts or not, it is always a good idea to use borax and washing soda instead.

If you have a particularly visible cyst, try to poultice it to the surface. Washing away some of it would be a great help to your body. Read herb books on poultices or just take whatever wild leaves you can gather; make a mush out of them with a blender, mix with honey or homemade skin softener to make it spread like peanut butter and cover the cyst with it night and day, covered with a piece of plastic. The wetness has drawing power. The plant juices have other benefits. If you can't poultice and must rely on kidney excretion, be sure to take kidney herbs at the same time.

Skin Rash

Rashes can be caused by many things. To find the cause, follow a logical pattern. It may be due to:

- HIV
- Yeast and Fungus
- Allergy
- Childhood Diseases (rubella, etc.)

HIV is the most alarming possibility. Eliminate it first by testing for it. It's often too early for a clinical test to be positive for HIV, so use the Syncrometer (page 457). I often see the rash disappear within days of eliminating the virus.

Probably the most common cause of rash is **yeast**. *Candida* has a resonant frequency of 384-388 KHz. If you test positive for it, stop all commercial soap and detergent for all possible uses. Zapping *Candida* may drive it away for a few days. The fungus is hosted by another parasite but finds your skin quite satisfactory for a home, at least while your skin immunity is low. It may be low from wearing metal jewelry, having metal tooth fillings, aluminum (from lotions and soaps), cobalt (from shaving supplies), and zirconium (from deodorant.). When all these are removed, the skin will dry up quickly in open air or under a heat lamp.

In babies' skin, immunity is low from diaper chemicals (mercury and thallium), wipes, and soap and detergent chemicals. Tight fitting, mercury treated diapers are a modern atrocity. Always line them with a tissue and keep them loose fitting.

The skin that has rash or fungus should be dried with paper towels, unfragranced and uncolored, in order not to contaminate the cloth towels, and thereby transport the tiny infectious spores to other skin locations. You may use skin healers.

Only the skin healers given in *Recipes* are safe to use. Keep them refrigerated when not in daily use. Commercial lotions, including health varieties, all contain toxic ingredients.

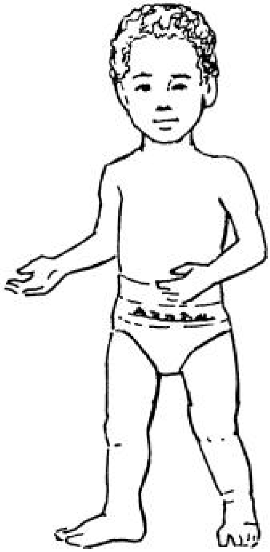

Fig. 29 Tight diapers are a modern atrocity, forcing mercury and thallium into the baby's sponge-like skin.

Many adult rashes are due to **allergies**. Allergy to nickel is common, the reaction is easiest to see under rings or watches. The metal is pulled into the body for elimination. Since nickel is used by so many bacteria, especially urinary tract bacteria, it doesn't get eliminated, it gets taken up by bacteria. It also piles up in kidneys, adrenals, bladder and prostate where bacteria thrive on it. Strangely, it also accumulates in the male scalp (and in women's scalps who have male pattern baldness).

Allergy to strawberries, perfume, deodorant or chlorinated water, however different they are, can all be expressed the same way, in a rash. The liver has refused (been unable) to detoxify the chemicals in these items and allows them to circulate in the body. Not for long, though, since great damage could be done to brain and other tissues. Mercifully, the skin grabs these chemicals into itself. Ultimately, it must still be removed. The allergic reaction and your immune system come to your skin's rescue, although bringing you discomfort.

Why didn't the liver detoxify the chemicals given to it? The answer is based on a simple experiment. Try cleaning your liver (page 552) several times or until 1,000 bits of refuse have been washed out of the bile ducts. This relieves the back pressure on that part of the liver, and allows it to do its work again. The physiological details are not understood. Fortunately, the results are instantaneous. The day before the liver cleanse you would never eat a strawberry or peanut for fear of a reaction. A few days after the cleanse, your body "knows" which food it might tolerate and as you try a bit, you notice no reaction, no rash This simple experiment suggests that the liver couldn't detoxify these foods because of the refuse in its ducts.

Each liver cleanse "cures" a different set of allergies suggesting that the liver is compartmentalized—different parts having different duties. It follows that by getting all the stones out all your allergies will disappear. Experience shows this to be true, although it can take two years to carry out such a program.

Meanwhile, avoiding the offensive food or product is very important. It is quite destructive to bathe the brain in strawberry chemicals or your toes in maple syrup chemicals. Stay off allergy-producing foods and products even if you can tolerate a little or can be "desensitized" to them with shots or homeopathic methods. Use these methods for relief, not license to continue using items that tax your body.

Certain **childhood diseases** produce a rash and this can be diagnosed by testing for the suspected disease with a slide or culture of it. Then use a zapper to kill both the bug and any larger parasites that may have brought it in.

Hives

Sometimes your body will break out in hives instead of a rash. Again, gallstones in the liver and gallbladder are responsible. Do the Liver Cleanse.

Acne

The more severe cases of acne cover parts of the body as well as face and can pit the face so badly there is hardly any clear skin left.

In ten days you can reverse this, so most of the skin is beautifully clear. There are bacteria involved and skin oils of youth feed these bacteria. Acne has been extensively studied by scientists. Perhaps the true culprit was too big to be seen with a microscope or too small (antigen) to be recognized or just too unimaginable. I inevitably find *Trichinella*, one of the four common roundworms that infect humans.

Test yourself to *Trichinella* with the slide called "larvae in tissue." Search for it in your skin. It is generally believed to reside in muscles, especially the diaphragm, but in acne cases it is in the skin.

If you only have a frequency generator, set it at 404.5 KHz and at neighboring numbers, extending 5 KHz on each side, to be sure to include all eggs and other molt stages. Or zap.

Psoriasis, Eczema

Psoriasis and eczema are both caused by *Ascaris*. Their molting chemicals are quite allergenic; perhaps it is these that are affecting the skin. Since pets pick these worms up daily, there is chronic reinfection in families with pets. Keep zapping.

Bernadette McNutt, 34, had acne on her back, chest and face. She had been treated since teen age with ultraviolet light, Retin A, and antibiotics. She already had a history of shingles and fungus, two more skin conditions. Her skin was toxic with strontium and her kidneys had cadmium, silver and beryllium deposits inhibiting excretion. She had only one parasite, extremely high levels of *Trichinella*. Her children were also infected as was the cat. In spite of using parasite herbs for months she got no improvement until the baby was out of diapers. Then she cleared up.

Royce Hamilton, 17, had acne so dense on his face there was no good spot the size of a dime anywhere. He had it about one year. His urinalysis showed "amorphous" crystals (stones of all kinds) and a trace of protein. He had *Trichinella* worms throughout his organs. He was very fatigued. He was started on kidney herbs so there would be good excretion after killing the *Trichinella*. His thyroid and kidneys were full of zirconium and titanium from all the lotions he used for his skin. He didn't need deodorant. Evidently even his armpit bacteria had been affected. It took four months to clear his *Trichinella* although there were no young children or pets in the house. His face was beginning to heal, but three months later he had a recurrence, although his parent was not a carrier. After this, he cleared it up again and his face looked as beautiful as a child's.

Evan Knight, 36, had psoriasis at elbows and knees from age 9 but now it was spreading to his fingers and scalp. He occasionally had bronchitis and puffy eyelids, indicative of *Ascaris* but at the time of his visit he had *Trichinella* fluke stages and *Echinostomum* in his skin. He was started on the parasite program and in three weeks it was clearing instead of advancing. He switched to milk for his beverage to raise his immunity and removed the arsenic, formaldehyde and thulium (from his vitamin C) by doing the necessary cleanups.

Gerry Chastain, 41, had a red nose, erupted along the sides. He had been on sulfa drugs and Emycin.^{TM} He had *Leishmania tropica*. He killed it in the office with a frequency generator and got immediate improvement but four weeks later it was back. He had four solvents built up in his body: benzene, TCE, TC Ethylene and hexanedione. This situation would make recovery impossible since he was no doubt reinfecting himself. He also had titanium, platinum and silver accumulated in his tissues and needed to replace his dentalware before expecting a permanent cure.

Floyd Oldham, 50s, was getting pimples on his nose. His whole face was red and flushed looking. His bowels were loose and he had some urinary urgency. He harbored two kinds of *Leishmania* (*braziliensis* and *tropica*). He killed the *Leishmanias* with a frequency generator and started himself on the kidney herb program. Five weeks later the pimples were gone but general redness had reappeared. The *Leishmanias* were still gone but this time he was toxic with cobalt.

Cobalt is known to promote skin cancers and also heart disease. This was discovered decades ago when an outbreak of heart disease occurred in England. It was traced to a pub (where they all partook) where cobalt was added to the beer to make the foam rise higher! In fact, the foam will just stay in place like a hairdo, if you add cobalt. It was made an illegal additive. Gradually, it has crept back into consumer's products: first toilet cake (blue), then window washer (blue), then dishwasher detergent, and now even mouthwash. If you see a blue colored product, **stay away from it**. It accumulates in the heart and skin. People with skin cancers often have cobalt build up.

Floyd stopped using all these products and cleared up again.

Grethe Driscoll, middle aged, wore tons of make up, so skillfully applied that scars from a face lift could never be detected. When she had minor breakouts, which usually occurred while away on a trip, it seemed like a catastrophe. She tried everything available but could not get to her parasite herbs until she was back home several weeks later. After one week on them (5 day high dose plus maintenance) her complexion was perfect again.

Crofton Thornton, 15, had an embarrassing case of acne. He had *Ascaris*, hookworm and *Strongyloides* (he also had migraines) all reacting in the skin. He stopped drinking commercial beverages that gave him solvents. He killed parasites electronically and with herbs and got a considerable improvement. But he still had *Strongyloides* one month later. Nevertheless, he had seen the connection and he knew it was just a matter of persistence to a clear complexion. Note, he should have had *Trichinella*—did I miss testing for it?

Warts

Could we get warts from playing with toads in childhood? We don't play with toads anymore. Yet we get warts. We don't know how we get them. But after learning how to get rid of them, you will probably know how you got them. Not all warts are the same. In fact, they might all be different: each one is made up of 5 or 6 different viruses, not just one as we had believed.

Peel a tiny fragment off one of your warts. Prepare it for testing by placing it in a small bottle. Add a few tsp. filtered water and a ¼ tsp. grain alcohol, Label it with the location you got it from: like "left middle finger knuckle. "First, search your body for other locations of this wart (organs that test positive to your sample). You can easily find them in your skin, of course. But also search electronically in your liver, spleen, muscles, stomach, heart, pancreas. **Notice how often they are present in the pancreas**. The pancreas seems to be a wart-virus heaven. Are they in the islets or the rest of the pancreas?

Without a zapper, you will need to find the frequency of each virus to completely destroy it. Attach your frequency generator and search between 400 and 290 KHz. When you find its resonant frequency, kill it by treating yourself for three minutes at 10 volts from a frequency generator. Will your warts fall off?

In a few days one or two of your warts will begin to shred. After a week you may lose one or two completely, and find that several more have become smaller. The remainder are unchanged. Continue to identify and kill them. Notice that they are not necessarily gone from the pancreas or other organs at the same time as they are gone from the skin. Perhaps warts are not the benign entities we have believed them to be. They may, in fact, ride into the body on some common bacteria, like *Salmonella*, or common parasite like pinworms or tapeworm stages.

Zapping doesn't reach all the viruses in a wart either. It takes repeated zappings to start the shredding and gradual killing of warts.

Guy Laird, age 11, had warts on his lips besides fingers. His job was feeding the three outdoor dogs. He was full of *Ascaris*. He had *Taenia pisiformis* and *Taenia solium* bladder cysts in his liver. These were shedding viruses into Guy. He was started on Rascal for six weeks (this was before the zapper was invented). Maybe his benzene buildup was responsible for letting so many parasites (and their viruses) survive and multiply in his body. He stopped using toothpaste, killed *Ascaris* (408 KHz) and some flukes (434 to 421 KHz) and improved his diet. All except one wart came off (without bleeding). He was given different chores, too, to reduce his contact with animals and their parasites.

Georgianna Mills, a middle age music teacher, broke out with warts all over her hands, at least 30 in total. A few months later she was diagnosed with bone cancer; she always wondered if there was a connection. She cleared up her cancer and killed her viruses and bacteria with a frequency generator. Nearly all her warts disappeared. But her indoor pet brought new parasites daily, especially *Moniezia* tapeworm stages. With each *Moniezia* infection (about once a month) she got new warts. She was never able to clear them completely.

I concluded that each wart is actually composed of 3 to 6 viruses and these viruses are distributed throughout our bodies! How satisfying to be able to rid our bodies of them, once and for all even in internal organs. There is a catch. Small remnants of some warts do not disappear in spite of killing most of them. More accurately, they disappear and then reappear in our internal organs. Could this suggest to us their true origin? Could it be a tapeworm stage?

Tapeworm Stages

Our bodies harbor numerous stages of tapeworms. But not the tapeworm itself, which may belong to a dog, cow, or pigeon. Tapeworms lead complicated lives, much like insects with their caterpillars, larvae, larval molts, pupae and eventual adults. Tapeworms shed eggs with the bowel movement of the animal host. The eggs blow in the dust and reside in the earth. A vegetarian animal nibbling vegetation near this filth, or licking dirt and dust off its coat, swallows the eggs. Humans, too, eat plenty of filth by licking their fingers. As children we all eat dirt simply by eating with unwashed hands.

The Jewish society discovered the great importance of washing hands before eating, thousands of years ago. But many of us choose to ignore truths that seem old fashioned. In our own relatively short life times we cannot see the whole picture as well as the prophets and seers of ancient cultures could. We eat plenty of dirt and along with it, the eggs of tape-worms. Dog and cat tapeworms are most prevalent, but sheep, cow, pig, and seagull tapeworms are also common.

There is hardly a predator species in existence that doesn't have its own characteristic tapeworm. Whatever animal species you live near, or once lived near, you probably swallowed some of its filth and some tape eggs. The **eggs** hatch in your stomach and the tiny **larvae** burrow into a neighboring organ without any consideration that this is your stomach wall or spleen or muscle. The larva's plan is not to grow into a long worm—that can wait. The larva must simply

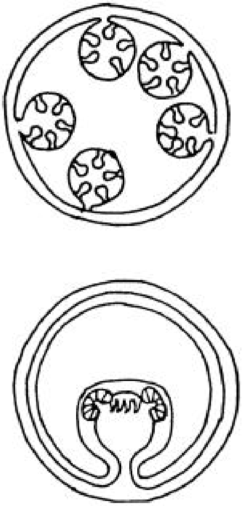

Fig. 30 Some cysticercus varieties (types) have multiple heads.

survive until you can be conveniently eaten! A wolf or a tiger will surely come along! In bygone days it did.

The larva is about ¼ inch long, surrounded by a “sac of waters,” like a tiny water balloon. Looking very closely at this sac, called a **cysticercus**, we see a head (scolex), complete with hooks and suckers, turned inside out, inside a bladder.

As the tiger's teeth bite down on the cysticercus, the pressure pops it out. The head is now right side out with hooks and suckers ready for action. Now it grows in the tiger!

It quickly hooks into a loop of intestinal wall so it can't be swept away and begins its growth into a regular long **adult** tapeworm. The tiger is the true or **primary host**. We were merely the secondary or **intermediate host**. Why does the adult tapeworm prefer the tiger instead of us? Only Mother Nature knows. But the best way to get to a carnivore is through its prey.

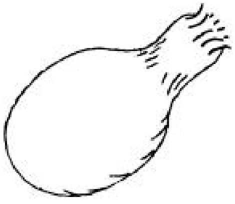

Fig. 31 Emerged cysticercus.

You can find these larval cysts in your organs using slides of the *cysticercus* stage of various common tapeworms. Search in your muscles, liver, stomach, pancreas, spleen, intestine and even brain. **You will not find even little bits of them in your white blood cells**. My explanation for this curious finding is that the tapeworm leaves no debris to be cleaned up by your white blood cells. Evidently your body builds a cyst wall around the larva to tightly encase it and prevent toxins and debris from entering your body. Thus your white blood cells are not alerted in any way. Of course, the larva is much too big to be devoured by tiny white blood cells anyway. Yet, it seems that if a pack of white blood cells had attacked the larva just as soon as it hatched from the egg they would have been able to devour it. Perhaps it enlarges too rapidly. Perhaps our white blood cells are preoccupied. In any case, we begin to load up on tapeworm stages from infancy and by the time we are middle aged we have dozens tucked away in our organs.

Some do die in the course of time. Perhaps their true secondary host is a rabbit or a mouse instead of a human. The short life span of these other hosts might mean that the life span of the *cysticercus* is also quite short, not 40 years! When they die, the white blood cells do clean them up and we can see them in our white blood cells at this time. It can take several weeks for the *cysticercus* to be completely gone by this natural method. During this time, we become ill! Numerous bacteria and viruses spring up, as if from nowhere, in our organs.

Don't be surprised if you are testing yourself during illness to find a tapeworm or two in your white blood cells! It is well worth searching for at such a time. Help your body dispatch the tapeworm stages all together with your zapper. A frequency generator is bound to miss some. Some cysticercus varieties consist of many heads, and each head has even more heads inside it! These might have different resonant frequencies. Only killing them together has the desired effect. Remember bacteria and viruses are released by killing tapeworms, so always follow with a second zapping in 20 minutes, and a third zapping 20 minutes after that. Only then can your tapeworm-related illness disappear.

If you do nothing, your body will be kept busy killing bacteria and viruses as the tape cysticercus wears down and eventually dies. You may not wish to identify all of them (but at least search for *Adenovirus*, the common cold) and just note where you are being attacked: your nose, throat, ears, lungs, bronchi. Internal organs are attacked too. It seldom takes more than three weeks, though, for your body to clean up a tape stage even without any help from a zapper. The attendant illness will be gone by then, too.

Watching these events in your body gives you insight into the very powerful forces at work, called immunity or body defense. The body "knows" a great deal more than we have surmised. There is yet so much to discover.

What initiated the death or dying process of the tapeworm stage in the first place? Has your body been trying all along and finally succeeded? Has the *cysticercus* reached the end of its life span naturally? Have its (the tapeworm's) own viruses and bacteria gotten the upper hand and killed it? Did it accidentally absorb something that killed it?

By taking a herbal combination, Rascal, you can soon find a tapeworm stage in your white blood cells where you could not find it earlier. It is now dead or dying. This proves the effectiveness of Rascal, even though it is slow.

Since we all eat dirt and inhale dust that is laden with dog feces or other animal excrement, we all harbor tapeworm stages, although none may be present in our white blood cells. **Are they harming us?** Perhaps they are living out their lives as quietly as they can in our organs, the way mice or ants try to live in our dwellings. Yet, when tapeworm stages are being killed, either spontaneously by your body or with a zapping device, we see an assortment of bacteria and viruses spread through the body, including the common cold.

Getting rid of the tapeworm stages in your organs seems a very worthwhile goal. Since each of us has been associated with dozens of animal species in our past, we probably have dozens of varieties of tapeworm stages in us. I cannot identify more than a handful due to lack of prepared slides. You can find them without identifying first, though, by listening to their emission frequencies. Their emissions are often extremely weak, possibly due to being encased in a cyst. Search between 510 KHz and 410 KHz. You may wish to "track" them for a while before killing them. You may wish to search for identical frequencies in your pet's saliva. Or you may wish to dispatch them as rapidly as possible. Use the zapper, not a frequency generator. Remember to "mop up" after your tapeworm killing by zapping again to kill bacteria and viruses that have been released from the tapeworm.

You may be disappointed not to feel any different after ridding yourself of numerous tapeworms and their pathogens. Evidently, the tapeworm stage itself doesn't make you sick; it is simply there like a wart is there, without making you sick. Its viruses can make you sick. Depending on which virus it is, it can make you very sick or not sick at all. Different viruses invade different organs. And some of these turn into warts!

The Flu

Influenza is a virus that can cause “the flu.” Does it belong to us as humans or to a larger parasite we are hosting? It is easily transmitted from person to person and in less than a year can spread across the planet. Some flu examples are Influenza A, B, C and Swine flu.

However, much that is called “flu” is actually caused by a bacterium, either *Salmonella* or *Shigella*. If someone in your family is “catching” a flu, test their saliva for the presence of dairy products, implicating the *Salmonellas* and *Shigellas*. Also test for influenza A, B, and C. Children's “flus”, especially when there is a fever, are usually due to *Salmonellas*. Even after zapping it can take an hour for the symptoms and fever to go away.

Go straight to the refrigerator and throw away all dairy products. Throw away all milk, cheesecakes, buttermilk, cream, butter, yogurt and cottage cheese, deli food and leftovers. You may wish to identify the food source of your family's bacteria first, and save the uncontaminated food. Use the sick person as a subject, searching for foods that appear in her white blood cells (or search their saliva sample for the food offender). If the flu is “going around” your neighborhood, you might wish to tell some of your neighbors which foods you found were contaminated. They may have purchased the same food! Obviously, when a contaminated shipment of dairy products arrives in your grocery stores, quite a few people will be consuming it, setting the stage for a “bad flu” that “goes around”.

Why can some people eat contaminated food without getting sick? Maybe their *Salmonellas* don't multiply rapidly. Maybe their stomach acid is strong enough to kill most of them. Maybe their bowel movements are frequent enough to expel them quickly. Maybe they haven't been on frequent treatments with anti-strep antibiotics. The answer is not known yet. After a serious bout with *Salmonellas* or *Shigellas* the body does not completely clear itself of them. They stay in hiding somewhere. When a new batch of bacteria arrives, even though very sparse, as in 1 tbs. of milk, the two subtypes can hybridize and produce a much more vigorous offspring. This is called *virulence*. You are made much sicker by more virulent subtypes of bacteria. You may have diarrhea, vomiting, illness. Especially if you believe you have “lactose intolerance,” pay attention to *Salmonella* and *Shigella*.

If your flu is due to an influenza virus, kill it with your zapper. Some family members may prefer to take a homeopathic flu “remedy” such as Oscillococcinum™ or Flusolution.™ Others may take herbs. These probably act by prying the viruses out of your cells' gateways and channels so the white blood cells can easily devour them.

But flu due to *Salmonella* is not easily zapped away. Remember, the zapper current does not penetrate the bowel contents, which is exactly where *Salmonella* lives! Besides zapping to clear them from your tissues, you must eliminate them from the bowel by using the Bowel Program (page 546).

Lugol’s iodine solution (see *Recipes*) can quite quickly get rid of Salmonella throughout the body. Use 6 drops (small drops from an eyedropper) in ½glass of water four times a day. If no more *Salmonella* is consumed, it will be vanquished in a day or two.

If your flu brings you a fever, use Lugol’s.

Fever

Fevers are there to help your body fight the invaders...up to a point. Don't use a fever medication unless the body temperature goes over 102°F, and then only enough to bring the fever down a bit.

Most fevers, especially **“fevers of unknown origin”** are due to *Salmonellas* and *Shigellas*. Your body may be young and strong enough to kill them but not strong enough to kill an everlasting supply of them coming from dairy foods you eat on a daily basis. Stop eating salads at restaurants immediately. Stop eating food made by others' hands unless it is sterilized. Stop eating dairy foods until you have cooked them. Stop eating those that can't be cooked.

Sam Ellis, age 7, had two episodes of severe abdominal pain with fever lasting two weeks. He got hyperactive with milk products, had a frequent cough and stuffy nose. (Here the picture is quite clear. The milk products were bringing him *Salmonellas*, *Shigellas* and other bacteria which grew in his intestine to produce pain. But why only Sam and not his brothers?) Sam had a buildup of benzene from using bathroom soap containing a special herbal oil. Sam also had hookworms, intestinal fluke, and rabbit fluke, probably due to his lowered immunity from the benzene. Then his mother boiled Sam's milk, removed the polluted soap (she planned to use it herself!) and killed his parasites with the parasite herbs. His fever went away and stayed away. He said he enjoyed all this because now he “could play after school” without a stomach ache and he wasn't being sent to the nurse's office because of a fever. Notice the bacteria causing the temperature went away by themselves, probably due to the return of his normally strong immune system.

Jalene McCormick, 46, had been passing lots of kidney stones (hundreds) for years and had a temperature most of the time for which she was on antibiotics. It took her six months on our kidney herb recipe to dissolve and pass so many they no longer showed up on X-ray, and to stop making them. Then her fever left, not to return.

Kristen Jane Johnson, a young mother, had recurrent fevers and fatigue but, besides EBV, her doctors found no cause. We found HIV virus and a lot of bacteria and parasites. The fever-causing bacterium was *Salmonella*. To stop her *Salmonella* attacks she had to raise her immunity besides boiling all dairy products. Moldy foods (pasta) and lunch meats (benzopyrenes) were the source of liver toxicity. Each new *Salmonella* attack immediately invaded the liver so a vicious cycle was set up. When she stayed meticulously on the parasite program, meticulously off unsterilized dairy products, and meticulously off benzene-polluted items, she cured herself of fevers and night sweats and the HIV infection. Perhaps in two years the liver will have recovered enough to kill *Salmonella* that enter it, but she is not taking any chances till then.

Although Kristen was eating food polluted with both *Salmonellas* and *Shigellas* she only “picked up” *Salmonella*, never *Shigella*! Why is that? In contrast, people with multiple sclerosis “pick up” *Shigellas*, not *Salmonellas*.

Multiple Sclerosis & Amyotropic Lateral Sclerosis

Multiple sclerosis (MS) is a disease of the brain and spinal cord. It is called *lateral sclerosis* if the disease is mainly in the spinal cord.

It is caused by fluke parasites reaching the brain or spinal cord and attempting to multiply there. Any of the four common flukes may be responsible. Kill them immediately with your zapper or a frequency generator (434 KHz to 421 KHz). They cannot return unless you reinfect yourself. Stop eating meats, except fish and seafood. All meats are a source of fluke parasite stages unless canned or very well cooked. Pets and family members are undoubtedly carriers of the same flukes, although they do not show the same symptoms. Give away your house pets. Don't kiss your loved ones on the mouth. Make sure your sex partner has also been freed of fluke parasites.

The most important question you must be able to answer is why did these parasites enter your brain and spinal cord? When the brain contains solvents, it allows flukes to multiply there. The solvents, **xylene** and **toluene** are common brain solvents always seen in MS cases. Evidently these solvents accumulate first in the motor and sensory regions of the brain, inviting the parasites to these locations.

Xylene and toluene are industrial solvents used in paint and thinners. It is also a pollutant of certain carbonated beverages (I found it in 7-Up,™ ginger ale and others that I tested). Stop drinking them.

All MS cases I have seen also harbor *Shigella* bacteria in the brain and spinal cord. These come from dairy products. They are manure bacteria. Be absolutely meticulous about sterilizing dairy products. Even one tsp. unsterilized milk added to scrambled eggs could reinfect you. Not even heavy whipping cream or butter is safe without boiling. Kill bacteria every day with a zapper. *Shigellas* produce chemicals that are toxic to the brain and spinal cord. Eliminating *Shigellas* brings immediate improvement.

All large parasites like flukes have their own entourage of bacteria and viruses. Perhaps it is these that initiate the brain's reaction, which is inflammation and scar tissue formation in the outer covering of brain cells and nerve fibers. Perhaps it is the fluke stages themselves. Your brain is trying desperately to heal these lesions, only to be assailed by a fresh batch of solvent and *Shigellas* and another generation of parasites and pathogens.

The other pollutant associated with MS is **mercury** from dental metal. The mercury that is constantly released in the mouth does not all get excreted by the kidneys or eliminated by the bowels. Some of it travels up to the brain and gets into the spinal cord as well. You will be able to eliminate and excrete more mercury by doing a kidney and liver cleanse. The mercury may itself be polluted with **thallium** which is even more toxic. For this reason mercury removal should be done extra thoroughly to be sure no thallium has been left behind.

If you are concerned about MS-like symptoms, purchase slides of the brain regions, cerebrum, and cerebellum. Or purchase pork brains at the grocery store and snip out a portion of the sensory lobe and cerebellum. Prepare these as test substances (sterilize your hands afterward). Test your daily foods and body products for their presence in these brain areas. Also test for parasites, bacteria (especially *Nocardia* and *Shigella*) and other pollutants such as arsenic and pesticides. If the disease (tremor and lack of sensation) has not progressed too far, you can cure it. In all cases you can stop it from progressing further by cleaning up dentalware, the environment and diet.

Brandi Rainey, age 34, of Amish religious culture, was diagnosed with MS four months earlier after an MRI confirmed it although she had symptoms for many years. She was told she had inherited a gene for it and that Amish folk are particularly susceptible to MS for reasons of inbreeding. She had a constant pain running down the side of her neck, and headache. Her legs were getting too heavy to get up stairs. Our tests showed her brain was full of scandium (tooth metal alloy) and fluoride (toothpaste). Her vision was getting worse; her eyes were full of wood alcohol. She lost no time in getting dentures: there were no teeth that could be saved. She had several bacteria growing in her jaw bone: *Strep G* (sore throat bacteria), *Staphylococcus aureus* (this was raising her pulse to over 100), *Clostridium tetani* (causes great stiffness), and *Shigella* (produces nerve toxins). She killed these with a frequency generator. Five weeks later the pain and stiffness in her neck were gone, her pulse was down to 100, her periods were free of pain and her hands seemed to shake less. She was put on the parasite program plus thioctic acid (2 a day) and histidine (500 mg, one a day to keep nickel levels down)and advised to cook and eat with non metal. Four weeks later her pulse was down to 80, her legs were much better. She zapped four remaining bacteria. Two months later the numbness and tremor had left; her legs still felt tired and her hands sometimes shook but she was quite reassured that MS would not claim her life. Nor had a gene betrayed her. A half year later she was walking and working normally, doing liver cleanses and keeping up her vigilance against parasites and pollutants.

Kendra Welch, 56, was diagnosed with MS a year ago, by MRI. She went to a chelating doctor and this cleared up her temporary ischemic attacks (T.I.A's) which were occurring daily. But she had lost her balance, eyesight was getting worse, her feet and hands stung. Her sister also had MS but nobody else in the family did which baffled her doctor. Her brain tissue was full of barium, europium, gadolinium, and platinum. These are dental alloys, although barium could come from bus exhaust (she wore no lipstick). She was advised to have all metal removed from her mouth immediately. Two days afterward she came into the office without any neurological symptoms. She stated she was afraid to stop her new health program, though, and this was good policy.

Lynne Ceretto, age 15, was diagnosed with an “MS-like syndrome” by MRI. She had intestinal flukes and stages, human liver flukes and *Trichinella* in the brain. She had no tooth fillings. But there was benzene in her thymus keeping her immunity low. She also had propane and asbestos in her brain from leaky pipes and a worn washing machine belt. She, too, was told she had a “bad” gene. She was barely able to walk with help. It took several months to track down the source of benzene—the drinking water. Sometimes it was polluted; sometimes it was not! When the pump was oiled, some of it dripped onto the cement platform. Rain washed it to the center and down the pipe in the well. They eagerly removed the platform, found the oil on the water surface, cleaned everything up carefully, until no benzene could be found which put her on the road to recovery. A year later she had recovered further.

Norma Luellen, a young mother, had tingling, numbness and weakness on the entire left side of her body. She was in process of clinical tests for MS. Her body was full of pentane, possibly from her workplace since 12 other persons working there also had MS (such a situation should make our government eager to jump at investigations)! She had intestinal flukes and their stages, not in the intestine or liver or thymus, but in her brain! We found her home air toxic with bismuth too, probably from cosmetics. In spite of staying on the parasite program she got reinfected with sheep liver fluke, probably from eating hamburgers. She was not able to stop her carbonated beverage habit and frequently showed xylene, acetone, methylene chloride in addition to pentane in her white blood cells.

Shannon Synder, age 44, had been getting more numb over her whole body for several years and was presently considered by her doctor to have MS. Her muscles twitched all night, making sleep impossible, and her hands shook. She had intestinal flukes in the brain (cerebrum and cerebellum) but none in the intestine! The brain also had wood alcohol from drinking Diet Coke.™ She also had bismuth (cosmetics), palladium, copper, samarium, and tellurium (tooth alloys) in her brain. She began to improve enough to be off Prednisone by her 10th day of the parasite program.

Erica Blake, age 41, was diagnosed with MS two years earlier although her symptoms went back 13 years. She was on Prednisone™ but her balance was getting so bad she had to be in a wheelchair. Chelation treatments kept her from deteriorating further. Her brain was full of gasoline; she used to work at a gas station and now was getting it from the attached garage. Her hands and feet were completely numb. She had 5 root canals extracted and a few days later was able to stand. She could now walk with a cane. She had human liver flukes, sheep liver fluke and *Trichinellas* and dog tapeworm stages in her cerebellum (motor control center). After killing parasites and starting to take thioctic acid (4 a day) and cleaning up her environment she improved enough to drive a car again, walk without a cane in her home. She regained enough feeling in her hands and legs to do her housework, too.

Kurt Nielsen, age 43, was told he had peripheral neuropathy. His feet were so numb he had to look at the clutch to drive. Also both hands were numb. He was full of kerosene and benzene possibly from fuel oil that he pumped for a living. There were fluke stages in his brain and a dozen bacteria and viruses. He also had mercury and thallium in his immune system which came from tooth fillings. However, he had all his metal fillings replaced two years earlier! He thought there was no mercury left in his mouth! Actually, he had little bits (called tattoos) left somewhere. And they were giving him the classical symptoms: numbness of hands and feet and gradual destruction of his nervous system. His peripheral neuropathy was due to thallium poisoning. The dentist couldn't find tattoos, and he was left in his predicament. (This was before I found mercury and thallium sources in many personal products.)

Duncan Wood, a middle age father of 5 young children, could still slowly shuffle along when he arrived. He could not raise his arms to eat. He had uncontrolled inappropriate laughter every minute. He was diagnosed with MS two years earlier and told he had a “bad” gene. The fact that one child was beginning to show similar symptoms strengthened their belief in the gene theory. Ten days later his inappropriate laughter stopped; he could get his right hand to his face, he walked twice as fast and had very little tremor remaining. Strong chelating treatments obtained at a Mexican clinic had drawn much of the mercury and thallium out of his brain. He killed the flukes and *Shigella* bacteria electronically and stopped consuming unboiled milk. The brain solvents, xylene and toluene were removed quickly, too, as well as asbestos. His fast improvement showed them how important it was to remove the source of these pollutants in his home.

Two days later he regressed considerably which made him feel quite depressed, since his chelating treatments had not stopped. Was there still an unknown factor? It was a return of *Shigella* bacteria! He had inadvertently eaten a non-sterile dairy food: milk added to soup when it was already done cooking! This was a valuable lesson. Nothing else had returned. He was away from the asbestos and xylene from the workshop at home. He zapped the bacteria again and applied greater vigilance to eating only sterilized dairy foods. He recouped his losses in one day.

Then they scheduled their dental work, which had already been done once two years ago! He had leftover mercury and thallium. Now, selecting a dentist with experience in finding tattoos and cleaning cavitations made much more sense to him than it had before. He also planned to do a kidney and liver cleanse after returning home. And to stay out of the workshop until the asbestos-containing belt had been replaced and the furniture painting had been moved to a different building.

Shigella also causes irritability and depression, a frequent problem for MS cases.

High Blood Pressure

High blood pressure is one of the easiest problems to correct without resorting to drugs.

The most important change to make is to stop using caffeine as in coffee, tea, or carbonated beverages. Don't use decaffeinated coffee or tea either because of the solvent pollution in them. Switch to hot milk or hot water if a hot beverage is desired, or any of the beverages given in the recipe section. If being without caffeine leaves you fatigued, take an arginine tablet in the morning (500 mg).

Blood pressure is mainly controlled by the adrenal glands which sit like little caps on top of the kidneys. Whatever is affecting the kidneys is probably affecting the adrenals, too, since they're so close to each other. You must find out what it is.

You could do your search in the kidneys since kidney tissue is available in grocery stores. Adrenal tissue is available on microscope slides. What will you find? Probably **cadmium**.

Search for the cadmium source in your drinking water! Cadmium comes from the metal pipes. In fact, you could scrape a galvanized pipe to get a cadmium test substance. Conducting or storing drinking water in containers of metal is as foolish a practice as eating food off the floor. Water picks up everything it touches simply because it is wet! You may not see what it picked up any more than you can see if it has picked up sugar or salt. The cadmium and other metal is dissolved in the water. The older the pipes the softer, more corroded they are, and the more metal is picked up as the water rushes by. If you find cadmium in your hot or cold water, you will never be able to filter it out. Nor should you switch to bottled water. The amount of cadmium in your clothing from doing laundry with this water is already too much for your adrenals and kidneys.

Change your galvanized pipes to PVC plastic. If you believe you already have plastic pipes or all copper (which leads to leukemia, schizophrenia and fertility problems) you will need to search every inch of plumbing for a very short piece of galvanized pipe left in the system! A piece as short as a 2 inch T or Y can be causing all the trouble.

The toxicity of cadmium, in fact, the high blood pressure connection, has been known a long time. After finding the cadmium start on the kidney cleanse. You might miss the cadmium problem if you don't attend to it first. Also remove all metal from your mouth.

All (100%) cases of high blood pressure I have seen could be easily cured by eliminating cadmium and other pollutants, followed by cleansing the kidneys.

To test whether you still need your blood pressure medicine, wait until your pressure is down to 140/90 or better. Then cut the dose in half. Check it again next day. If it has climbed back up you are not ready; go back to ¾or a full dose of medicine. Try again a few days later. If your blood pressure stays down, cut your medicine in half again (you are now down to ¼ the regular dose) and see if your blood pressure stays improved.

When you are down to 130/80 go off completely. But stay on the kidney herb recipe. At 120/80 try yourself on a few shakes of sea salt. The amount of salt eaten, once the pressure is down, has little influence. In fact increasing salt intake improves energy without raising blood pressure. Take no more than one teaspoon a day (2,000 mg sodium), total, including cooking. Better yet, make a salt that is a mixture of sodium and potassium chlorides (see *Sources*). Mix it for yourself in a 1 to 1 ratio or whatever your taste can accept. The sodium portion could be sterilized sea salt (test and make sure it has no aluminum silicate in it first).

Mold toxins have specific kidney effects! Especially **T-2 toxin**, found mostly in dried peas, beans and lentils. Rinse these thoroughly first, throw away shriveled ones, and add vitamin C to the cooking water. All cases of serious kidney disease show a build up of T-2 toxin. Be extra careful to avoid moldy food (read Moldy Food, page 381).

Bala Cuzmin, age 72, had high blood pressure for ten years but the upper (systolic) pressure remained high in spite of various medicines that were tried. She had three kinds of kidney stones and only one functional kidney. She stopped using caffeine, switching to arginine tablets to get over the let-down. Her diet was changed to reduce phosphate and add calcium, and she took magnesium and Vitamin B_6 to assist the kidneys. She was very anemic and her mean cell volume (MCV) was high due to *Ascaris* infestation. She killed parasites, cleansed kidneys but saw no drop in blood pressure which stayed at 150 to 170 systolic. Her adrenal glands were choked with copper and platinum. She had all the metal in her mouth replaced and promptly saw a blood pressure drop to 145-1 50. Three months later it was at 128 to 133 on half her medicine. She had not been tested for T-2 toxin yet, nor changed her copper water pipes.

Sabrina Patton, 66, had a long list of health problems, including high blood pressure for six years. She was on Corgard™ and diazide drugs which kept it down to 140-160/74-80. She had phosphate crystals in her kidneys and was started on kidney herbs and a diet change to include milk and exclude soda pop. She had high levels of mercury and copper in her immune system. She was feeling so much better after the kidney cleanse that she decided to remove her last fillings and replace her bridge, too, since it was shedding ruthenium. On her way home from the dentist, her ears stopped ringing and soon her blood pressure was down to 126/68. She was still on half a dose of drugs because she was too afraid to go off entirely. But when her pressure stayed down she found the courage to go off completely. This gave her the energy she wanted to play basketball with the grandchildren again.

Rolf Ehrhart, 61, had 80% blockage of heart arteries and high blood pressure for which he was on a Hydropres™ patch, Tenormin™, and Logol™ (diuretic). He had phosphate and uric acid crystals in his kidneys. He was started on kidney herbs followed by the parasite herbs. His *Ascaris* and flukes were zapped. He stopped using store-bought beverages. Then he could cut back on his medicines, measuring his blood pressure daily to guide him. After seven weeks it was down to 140/85, so he decided to do without medicine, a bit early. He was also getting chelation therapy and was now able to walk 2-4 miles a day. His next chore, which he approached gladly, was removal of all metal from his mouth.

Len Gerald, 45, was on Vasotec™ for high blood pressure. He was constantly sleepy; his blood test showed a low thyroid level in spite of being on Euthyroid.™ He was started on kidney herbs followed by parasite herbs. In two weeks, barely into his program, his blood pressure dropped. He had to go off his blood pressure medicine. It stayed at 126/80. He still had some *Ascaris* and other health problems but was highly motivated to clean them up, too.

Glaucoma

In glaucoma the pressure in the eyeball gets too high, putting pressure on fragile retina cells that do your seeing. The first question to ask is: "Is my blood pressure too high?," because there is a link between high blood pressure and elevated eyeball pressure.

Your blood pressure should be 120/80. Your doctor may say 140/85 is "not high." He or she is kindly refraining from giving you drugs until this level of pressure is reached. It is your tip-off, though, that something is not right and you should correct it now, when it is easy, and before other damage is done. Read the section on high blood pressure (page 210) to learn how to reduce it by going off caffeine, checking for cadmium poisoning from your water pipes, and cleansing the kidneys (page 549). Even though your doctor has explained how the tiny tube draining your eyeball is too narrow, you should ask: was it not too narrow before high blood pressure struck? Simply getting your blood pressure to normal is sufficient help for beginning glaucoma.

Antonia Guerrero, age 51, had glaucoma for five years and was deteriorating rapidly. She cleansed her kidneys, killed parasites and changed her diet to the anti-arthritic one since she also suffered from arthritis in her hands for ten years with painful enlarged knuckles. She didn't get relief from taking aspirin. She got rid of her asbestos toxins by bringing her own hair blower with her to the hairdresser. After seven months she had pain relief for her arthritis (without aspirin) and her glaucoma was pronounced stable by her ophthalmologist.

Tooth Decay

The strongest part of our body structure is our bones. The strongest bones are our teeth. How can they decay? We must look at the **enamel**, **dentine** and **root** of the tooth as well as the bone they rest in for some answers.

Scientists have already searched very hard and long for answers. But their work is hampered by commercial interests that try to shape the results. Since commerce determines which research can be done (that is, paid for) sacred territory can be ignored. For example, the effects of sugar-eating, gum-chewing, tooth brushing, fluoridation, tooth filling materials and diet can be ignored if it interferes with product sales. Trivial studies such as comparing shapes of toothbrushes, studying the chemical composition of plaque, and studies of bacterial structure and genes are done instead. Studies “at the molecular level” do not threaten existing industries.

Important research has lapsed since the 40's and 50's. Perspective on tooth health was sound and clear in the mind of Dr. Weston Price in the 1930's. His scientific studies stand as a beacon even today because truths, once found, do not change. He traveled the world over in search of good teeth. Anywhere and anytime he found them, he described the people who had them. This is excellent science. It lets you draw the conclusions. He described what he saw in a book, titled Nutrition and Physical Degeneration.[13] They came to these conclusions from the following observable facts:

1. Skulls of primitive peoples who lived along coastlines, such as Peruvians, Scandinavians and various islanders, and whose staple foods included fish daily, showed perfect teeth; not a single cavity in a lifetime. They had strong bones that didn't break even once in a lifetime of 45 years. Skeletal structure was fully developed, meaning the jaw bone was not undershot or cheek bones squeezed together, forcing the teeth to grow into a smaller than ideal space. Consequently, there was room for the wisdom teeth, and no need to crowd the remainder. They saw no crooked teeth or unerupted wisdom teeth. The authors estimated a daily consumption of **4 to 5 grams of calcium** in their fish containing diet.

Our daily consumption of less than 1 gram calcium daily is small by comparison. Our wisdom teeth erupt poorly, our other teeth are often crooked. But today bad teeth go shamefully unheeded because we don't need to chew our food, we can lap it (ice cream) or suck it, or gum it (applesauce).

2. These primitive peoples got all the calcium, magnesium, phosphate, boron and other bone builders they needed simply from eating (fish) bones. Mexican peoples got 4 to 6 grams of calcium a day from **stone-grinding** of corn for their staple, tortillas, instead of from fish.

---

[13] It is still available from the Price-Pottenger Nutrition Foundation, a non-profit organization that seeks to keep his observations alive. Their address is PO Box 2614, La Mesa, California 91943, (800) 366-3748.

Where do we get our calcium? Milk is our only supply. One quart supplies one gram. There is little excuse for a carnivorous society like ours to regularly throw away the bones of its food animals in view of our dire shortage. It leaves us dependent on milk alone. Milk has so many disadvantages. It is impossible to milk a cow by machine and not get a few manure bacteria, *Salmonellas* and *Shigellas*, into the milk. These bacteria are not completely killed by pasteurization the way more susceptible bacteria are. It takes boiling temperature to kill all of them. Why isn't milk sterilized? Water was sterilized for human consumption in distant decades. Chlorination of water is not ideal but it did sterilize the water. Milk could be sterilized by boiling or flash-heating.

Milk has other disadvantages: dozens of antibiotics, both by feed and by shot, bovine growth hormone, chemicals added in milk processing, the bad effects of homogenization, and allergy to milk. Yet, in a choice between milk drinking and bone loss, one must choose the milk. This would not be necessary if bones were properly salvaged–ground to powder and added back to the meat where it belongs–to offset the acidifying effect of the phosphate in meat. One gram of calcium is not much bone (½tsp.) but it requires a whole quart of milk. Bone powder added back to ground meat, soups, stews could greatly improve our tooth decay problem, bone density problem, and skeletal growth problems.

Softened teeth set the stage for decay; bacteria do the dirty work.

Zapping bacteria does not kill them all. The zapper current does not reach into abscesses under metal filled teeth or around root canals. *Staphylococcus aureus*, which we are constantly stuffing in our mouths as we lick our fingers, finds an immediate hiding place in a crevice where it can't be zapped. Many other bacteria hide here, too: those that cause ear ache, sore throats, bronchitis, stiff knees, joint disease. You can try zapping all the *Clostridia*, *Streps* and tooth decay or plaque bacteria. But the only way to successfully eliminate them is to pry them out of hiding and wash them away. This is a job for the dentist (see Dental Cleanup page 409).

Strep. mutans is considered to be the bacterium that causes tooth cavities. I have found it in milk, evidently another pasteurization escapee. All the more reason to sterilize dairy products.

Frannie LaSalle, 52, was getting compression fractures in her spine, but the weak bone condition was evident in her mouth (many teeth were loose—they could be jiggled!). Her gums were red and inflamed. A low thyroid condition (she needed 2½ grains a day of thyroid—in one day the normal body goes through 5 grains of thyroid products) contributed to this. Her blood phosphate level was high (4.7 mg/DL—should be below 4.0) and her alkaline phosphatase was 205, also high, showing she was dissolving her bones (including tooth sockets) at a rapid pace. Her whole system was too acid, as could be seen in elevated CO_2 levels (28, when 23-30 is normal).

Only the major minerals, sodium, potassium, calcium and magnesium can have an impact on this major disturbance. The dentist said she had to have all her teeth pulled and replaced with dentures. Her kidneys showed all three types of calcium phosphate crystals. She drank no milk. She had only three weeks before her oral surgery appointment. She was started on ½ cup 2% milk, 6 times a day plus 50,000 units of vitamin D (a prescription dose) to make sure she absorbed all the calcium. She also took magnesium oxide (300 mg. once a day) and vitamin B_6 500 mg (one a day).

She was started on the kidney cleanse to help activate the vitamin D and to help the adrenal glands make *estrogen*. Her estrogen level (5.2 pg/ml) was too low to get the calcium deposited back into her bones. She was also given licorice herb for their estrogen-like action to help with this and vitamin C, 1 gram (1,000 mg) 2 to 3 a day.

Her mouth care was to be as follows: potassium iodide (white iodine, made up by dissolving 88 gm potassium iodide in one liter/quart water). Purchase a new very soft toothbrush. Use no toothpaste or store bought floss. Use 2 lb. or 4 lb. (the 4 lb. is coarser) fish line (rinse first). Brush twice a day; floss only once at bedtime before brushing. Use 6 drops of food grade hydrogen peroxide for daytime brushing. Use 6 drops of potassium iodide for nighttime brushing. Use no mouthwash, chewing gum, candy. In three weeks her teeth could not be jiggled. Her dentist was astonished (but was not interested in how she achieved this). In six weeks her mouth looked normal and she could chew some foods. Her vitamin D was tapered as follows: Take 6 a week for the first week (miss one day). Take 5 a week for the second week (miss two days). Take 4 a week for the third week. Then 2 a week indefinitely. She never lost a tooth.

Muscle Diseases

There are a variety of muscle wasting diseases, thought to be genetic in their cause. Yet, what could be more easily inherited than a parasite? Persons living together share food, living habits. refrigerators, and parasites. Their shared genes indeed give them similar susceptibilities but if we take muscle parasites away, muscle diseases “magically” disappear.

Of course, there is no magic involved. It is actually hard work. Hard work to rid the whole family of parasites that are shared and possibly were present even at birth. Parasites that normally don't go to the muscles. For example liver flukes and intestinal flukes. They belong in the liver and intestine! Yet, in muscle disease they show up and reproduce themselves in the muscles. The reason for this becomes clearer when you see that certain solvents have accumulated there. Heavy metals, bacteria, and viruses have accumulated there, too. The host's muscles, instead of kidney and bowel, have taken on the duties of toxic dumping grounds. Could it be that the regular routes of elimination were overwhelmed? Or did the muscles, traumatized by these unusual parasites, invite the toxins? A tantalizing question.

But in seven minutes you can methodically kill everything and anything that is alive in your muscles and shouldn't be there. Zap until you are free of all parasite invaders. Your muscles will feel lighter afterward.

Muscular Dystrophy

In muscular dystrophy the solvents, **xylene** and **toluene** are seen to accumulate in muscles. These also accumulate in brain and nervous tissue! (See Alzheimer's page 269 and multiple sclerosis, page 204). Could it be that these solvents are actually present in the nerves of the muscles?

Fortunately these solvents will leave your body, by themselves, in five days after you stop consuming them! Stop drinking all store bought beverages, including water and powders that you mix, and including health food varieties. Water claims and health food powder claims sound as convincing and strong as a twelve inch plank to walk on. But if the plank leads out over the side of a ship, would you walk it?

Throw all your possible sources of solvents out. Flavored foods are the chief offenders (cold cereals, sweets and candy too). But of course, you should check in your basement or attached garage for cleaning solvents. Places where painting is done or automobiles are worked on should be off limits to you.

If you've been wondering whether you have muscular dystrophy, which I consider to be a fluke disease, search your muscles electronically. Use prepared slides of flukes along with a sample of hamburger meat to represent your muscles. If flukes have already taken up residence in them, you should diagnose yourself as “Positive”. If not, but other parasites and toxins are present, you have pre-muscular dystrophy.

Also, the likelihood of finding thallium is quite high, judging by the case histories.

Mel Rickling, age 18, had been seeing a specialist for bouts of muscular weakness for several years, but no diagnosis was given. His condition was not yet severe enough although it was difficult for him to raise an empty glass or get upstairs. It began at puberty, not too long after his first mercury tooth filling. He had asthma in childhood. The flukes attacking his muscles were liver fluke, intestinal fluke, and pancreatic fluke. Other parasites in his muscles were *Leishmanias*, several dog tapeworm cysts, and pinworm.

He also had assorted bacteria in his muscles. The solvents propyl alcohol, benzene, toluene, and xylene were accumulated there too. Ortho-phospho-tyrosine (cancer test) was already positive. His doctors had not searched for cancer in their biopsies. His drinking water contained lead and since he had lived in one house since birth he was probably drinking lead every day of his life. He also had high levels of mercury and some thallium accumulated in his muscles; these came from the tooth fillings in his mouth and could explain why his problems began after his first filling was put in.

His flukes and other large parasites were killed immediately with a frequency generator. He was started on the herbal parasite program to prevent reinfection. His diet and body products were changed to exclude solvent pollution. He could have no commercially prepared beverages except milk which needed to be boiled to kill bacteria. In twelve days his daily stomach pains were gone, so he was able to eat more and gain some much needed weight. The rash on his face was gone, the pain at his right side was gone, his muscle twitches were gone, his joints no longer ached and his mood was much better. The whole family was put on the parasite program and Mel was scheduled for dental cleanup. The plumbing repairs removed lead from the water and he was soon able to walk upstairs, in fact run upstairs.

A young man, seeing himself regain normalcy, wants nothing more than to lead a normal life” which includes reckless behavior. But after several warnings from his muscles he stuck to his restrictions and gained the weight he wanted in order to participate in athletics.

Myasthenia Gravis

is probably a fluke disease. Some chemical, possibly coming from the fluke, may affect the acetylcholine receptors, thereby causing an allergic reaction so they become inefficient. This is most noticeable in the eyelids. They droop from lack of strength to lift them.

The thymus is often involved, too. The thymus is extremely sensitive to benzene and with so much benzene pollution in our products and foods (pollution from gasoline is negligible by comparison), you will probably find benzene accumulated there. Search the thymus and the muscles for parasites, bacteria and tooth metal as well as toxins in the foods eaten daily. Kill invaders twice a week with a zapper or stay on an herbal parasite program until all danger of recurrence is past (one to two years).

Clean up dentalware, diet and environment. Keep no indoor pets since any new parasite, however tiny, will surely find the niche left behind by the flukes and give you a new myasthenia gravis-like disease. The whole family must be parasite-free to protect the member with myasthenia gravis. But it is a task easily accomplished and desirable in its own right, so discuss your plan immediately with family members. Don't delay. The flukes don't waste a single minute. They go right on feeding and breeding.

Carmen Opsal, age 37, was told by her specialist she had her myasthenia from birth since she didn't have the strength to nurse. She had pancreatic fluke stages throughout her body. Her plan was to start on the parasite killing program, clean her kidneys, remove toxic elements, kill bacteria and clean her liver. Long before she accomplished this, in one month, she was feeling better and had return of her strength on some days. She still had the solvent, methylene chloride, from drinking “pure” orange juice and praseodymium from eating foil-packaged foods, also thulium from her brand of vitamin C. She was full of auto exhaust and nickel from dental metal. There was hafnium from nail polish and hair spray and zirconium from deodorant. She planned to get rid of it all, and never need to return.

Universal Allergies

If minor allergies are due to a disabled liver, then extreme allergies must be due to an extremely disabled liver. This is the case for persons suffering from "universal" allergies, namely "everything", like the lacquer on floors, plastic chairs, the neighbor's flowers, and the grocery store.

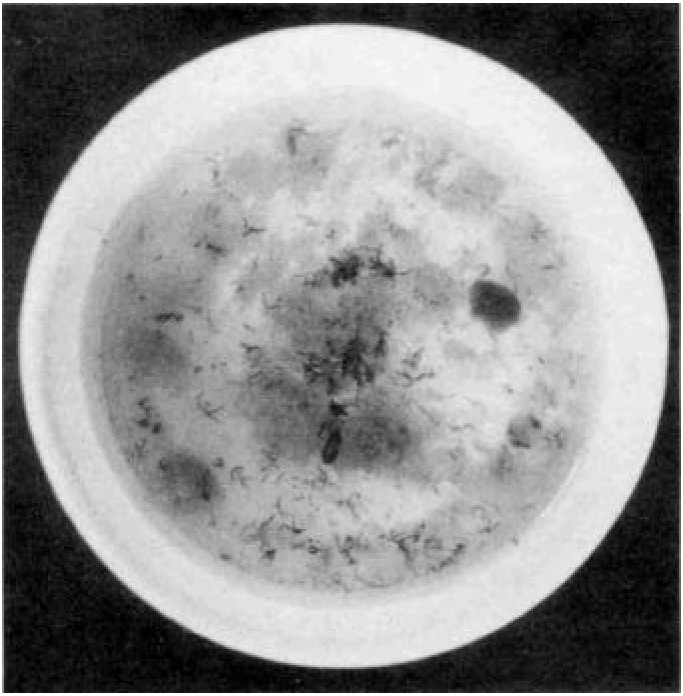

Fig. 32 Sheep liver flukes. Black threads in toilet are indicative of fluke remains.

They have more than merely clogged bile ducts. They have the sheep liver fluke living in the bile ducts! A tip-off to this situation is allergy to wool and wool fat (lanolin). A few flukes might not be noticed but a liver full of flukes that spill over into the intestine can give the worst case of allergy imaginable.

Sometimes the body manages to kill them with its own resources (maybe you ate something even too toxic for them!) They come through the bowel in a torrent. In the water of the toilet bowel they explode, spewing their infectious eggs all over in little black threads. Because these look like hairs, you may believe you passed "things with black hairy legs." These are actually burst flukes with black strings of highly infectious eggs. Why some people are literally taken over by these flukes is unclear. Amongst sheep, only certain sheep will be severely affected, being called "liver-rot." The disease in animals has been extensively studied.

Kill flukes with a frequency generator (434-421 KHz) or zapper. Come to the aid of the liver by avoiding food molds, removing dental metals, stopping chronic *Salmonella* infection and finally cleaning the kidneys and liver.

Environmental Illness

is another name for "universal" allergies. When more than a few flukes are present in the liver, they keep the liver from doing its major job: detoxifying all the food and chemicals that are taken into your body.

Different parts of the liver have different detoxifying jobs. One part detoxifies plastics and solvents, another part detoxifies perfumes and another newsprint ink, and so on. Foods have natural chemicals that need detoxification. By changing our food constantly we avoid overburdening any one of our detoxifying mechanisms. This is probably the basis for wanting different food at each meal and different meals each day. We somehow "know" when we're ready for the same food again.

Less extreme forms of allergy can be due to other flukes in the liver, such as human liver fluke (*Clonorchis*), or just plain clogging with numerous cholesterol crystals.

Anything that is lodged in the bile ducts obstructs the flow of bile. This causes back pressure in that part of the liver so it produces less bile.

The bile duct system is a gigantic tree with lots of interconnecting branches. Remember how "stringy" liver can be when you buy it in the grocery store. These strings are bile ducts. When one is obstructed, others take over its job. But when a whole section of the liver gets obstructed and it can't detoxify a whole set of chemicals you dare not get those chemicals into you again.

What if you do? These chemicals go coursing all over your body! They are taken up by various organs. The brain has special protection, called the *blood brain barrier*. But this can get broken by parasites that burrow. Now chemicals can roam through the brain. Some attach themselves and cause an "allergic reaction". Beryllium, from "coal oil," kerosene and gasoline attaches itself to the brain easily. Then other things attach themselves to the beryllium!

The extreme form of allergies simply requires killing the sheep liver fluke and other flukes inhabiting the liver. They tend to overflow the liver and inhabit the intestine, too. In this case, you might actually see some in the bowel movement after killing them. They won't let go of you as long as they are alive.

They have two attachments to hold onto you, yet, they are not difficult to kill, even with herbs. Use the herbal parasite recipe (page 338), zapper, or a frequency generator.

Sheep liver flukes might actually be breeding, that is, multiplying in the liver of the hyperallergic person. This is not normal. Sheep liver flukes "should" only spend their adulthood in our bodies. When the baby stages are also found in our bodies, (instead of in minnows or snails) there is undoubtedly a specific solvent involved. Environmentally ill persons have quite a few solvents accumulated in their organ tissues. Which one enables the sheep liver fluke to go through all of its development in the human is not known yet.

Obviously, the extremely allergic person, should remove all solvents from their diet and environment. Begin with eliminating propyl alcohol and benzene. The same products have other solvents too. It is processing of foods that puts solvents into them. Go completely natural. Dairy products are free of solvents, except for some cheeses. Remember to boil them to get rid of bacteria. Salt, olive oil, butter, and honey are free of solvents. With electronic technology, you can find solvent free products. Otherwise, if it didn't grow or you didn't make it from scratch, you must assume it has solvents! Cook from scratch, make your own pasta, bread, fruit juices, beverages.

Often, but not always, persons with sheep liver fluke, have a specific allergy to lanolin, a sheep product. Since lanolin is widely used in other products, this becomes a very broad range allergy. Such persons “can get no fat” at mealtimes or wear no wool without a considerable reaction. The allergy to lanolin does not disappear the day the flukes are all dead. But cleaning the liver with several liver cleanses (page 552) after killing parasites will start the recovery process.

Which comes first, the flukes or the solvents? That can't be answered. But what happens next is easily seen. The more flukes, the less able the liver is to detoxify solvents. The more solvents the better able the fluke is to multiply. A vicious cycle is set up that accelerates the illness.

Perhaps neither of these came first. Perhaps something else poisoned the liver so both solvents and flukes are given a home in your liver! Such a powerful liver poison would be a food mold: *aflatoxin*, *cytochalasin B*, *sterigmatocystin*, *zearalenone*, *ochratoxin, sorghum mold, griseofulvin, citrinin, T-2 toxin, Kojic acid, ergot* and others. Avoid food molds—see Mold Free Diet, page 365. The diet must be quite limited at first, to allow the liver time to "regain" its detoxifying capability.

Of course, it is assumed that environmentally ill persons have had their dental metal replaced by metal free composite. This includes gold. Gold accumulates in the pancreas, another organ of digestion. This may mean choosing partial dentures. Read the section on healing the jaw and Bone Strengthening (page 87) to ensure this move brings you success.

The liver is a versatile organ. It can regenerate itself but it won't if food molds block regeneration. Given half a chance it will become like new. After killing parasites do the liver cleanse (page 552). If it has been a month or more since you killed parasites, then go on a high dose parasite herb treatment the week before, or zap. Don't use the herbs the day of the cleanse. With one major allergy gone after each cleanse and by timing liver cleanses two weeks apart, it takes only six months to have a reasonably normal life again. You can endure indoor air again, sit on plastic chairs, read newspapers, wear cotton clothing and leather shoes without reacting. You must still be patient and careful as you take back the world for you to live in.

Delores Flores, 53, was brought by her husband to the driveway in front of the office. There she put on her mask and advanced to the outdoor bench. She did not dare to come in. Without doing any testing her condition was obvious. She must start killing liver parasites. But it seemed too simple to be believable. And she knew she'd be allergic to the parasite killing herbs (this was before the zapper). She decided to do nothing.

Patricia Humphry arrived wearing an industrial painters' mask. It smelled rubbery. Her winter boots smelled moldy. And a faint scent of mothballs came from somewhere. Even her car had an engine problem, spewing exhaust fumes into the driveway. I suggested she begin with some basic reading material on allergies. She did not return either.

Arlene Kelly, 50ish, could eat no fat—not the tiniest snippet. But at Thanksgiving she allowed herself a little gravy. The consequences were swollen eyelids, swollen face, swollen throat: quite a dangerous situation. After killing *Ascaris* and the flukes, and cleansing the liver (all in time for Christmas) she dared a little pie—and got along quite well!

Alcoholism

When the portion of liver that detoxifies ethyl alcohol (the drinking kind) is hampered you are at risk for alcoholism. The other contributors to alcoholism are **beryllium** and **ergot**. Perhaps there are even more contributors.

Beryllium is plentiful in coal products such as “coal oil”, and in gasoline to which kerosene or coal oil has been added. Fuel for “hurricane lamps” is a common offender, filling the air with beryllium. Liver blockage can force beryllium to circulate through the body; if it happens to attach itself to the addiction center of the brain, you're in a heap of trouble. Beryllium is very reactive. Any other allergen coursing by can react with it. If this happens to be alcohol–the drinking kind–you will be alcoholic unbeknownst to you.

When the liver is quite disabled, there may still be alcohol coursing through your body the day after you drank even a little bit! It is bound to find the beryllium stuck to the addiction center. Together, they turn the brain into a uncontrolled machine. Neurotransmitters are released that shouldn't be; others not released that should be. Mood is affected in a typically “alcoholic” way. Depression may be lifted—or caused!

Alcoholic persons should remove all fossil fuels from their environment, and never choose a career that exposes them to paint, cleaners, or automotive products. They should do liver cleanses until 2,000 stones or more are out.

Of course, they should never touch a drop of alcohol: not even the Black Walnut Hull Tincture in the herbal parasite program. The aqueous recipe should be made for them. Since alcohol is produced anyway, in the body, the liver should never be poisoned by molds, especially ergot: the very mold that is so abundant in alcoholic beverages! Alcohol and ergot interact to make each more toxic.

To help the brain recover its neurotransmitter status, take glutamine (500 mg.), a B-50 complex, and niacinamide (500 mg—to help detoxify ergot), with each meal.

To prevent alcoholism, protect your liver from food molds, especially ergot. Add vitamin C to nuts, pasta cereals, grains and even alcoholic beverages! Avoid fossil fuel pollution of your home by switching to all electric utilities.

Alcohol Addiction

There are many definitions of addiction. My definition is based on the special brain toxins, *beryllium* and *ergot*.

The brain has a region called the addiction center. If this center is stimulated it produces pleasure-chemicals. It is carefully controlled so that not too much pleasure or happiness can be experienced.

When a toxic substance, beryllium, is inhaled it circulates with the blood to the brain and may land at the addiction center. The more beryllium is inhaled the bigger the chance that it will occupy the addiction center. The brain cells in the addiction center have receptor sites for *glutamate* (the same glutamic acid that comes from the protein in our food). Normally, glutamate activates the addiction center. But when beryllium has “stolen” these seats, the glutamate is powerless to activate the joy and happiness center. The result is a low level chronic depression. The more beryllium there is to clog the receptors the worse the depression. Giving glutamic acid does not help but giving **glu-**

tamine does! Addicted or depressed persons should take glutamine, no less than 3 grams (3000 mg) a day. It comes in 500 mg. tablets. It is completely harmless.

When we drink alcohol or put it on the skin (as in mouthwash, tinctures, medicine) or produce it by fermentation in the intestines (*Candida* produces alcohol) a substance, *salsol,* is formed. Salsol reacts with beryllium. If the beryllium is in the pleasure center it reacts with it there. This reaction has the effect of activating the cells! Now a large amount of pleasure-chemical can be released. The amount is larger than normal because so many clogged cells are activated together. This explains the alcohol "high". In all the alcohol-addicted persons I studied, salsol was present, along with beryllium, on the receptor sites normally activated by glutamate (or NMDA or kainate). As we removed the beryllium we saw that the salsol also disappeared.

The solution to alcoholism is to avoid ergot contaminated food and avoid beryllium inhalation. We also remove the brain beryllium using **thioctic acid**. Stopping the use of alcohol may save a life or career but does not correct the problem. Even after 30 years of abstinence, I still see the beryllium present in the addiction center and the salsol, derived no doubt from endogenous sources, still attached to the beryllium. This is why the addiction is never gone even after years of abstinence.

If any member of the family is, or was, addicted to alcohol the house should be searched for beryllium sources. Hurricane lamps or antique lamps are the most common sources. Remove them permanently. Washing does not clean them. Remove all solvents, cleaners, lighters. Switch to a butane lighter. The air should be tested for beryllium. The garage door to the house should be permanently closed, and the car and lawnmower kept out of it. Addicted persons should not be painters! Nor walk into a dry-cleaning business. Soon you will see a more cheerful disposition in the addicted person and this will be rewarding for the whole family.

Miguel Alcorn's wife cleaned up the environment for Miguel, whose alcoholism was 30 years long. Even lighter fluid was removed. The garage was sealed off from the house. She added vitamin C to his meals so he didn't have to "take anything". She was very careful about moldy grains. He lost his preoccupation with drinking after killing parasites. She believes he is not sneaking any. Our hats are off to her.

Seizures

are always caused by tiny *Ascaris* larvae in the brain.

I suppose they find their way to the seizure center by accident. It is not normal for them to be in the brain, they typically travel between the stomach and lungs. *Ascaris* eggs are present everywhere in animal filth. Dogs, cats, horses and pigs all get *Ascaris*. Their excrement dries and flies about in the dust, but mostly it resides in the soil. Children playing in the dirt cannot help but pick up *Ascaris* eggs. The eggs hatch in the stomach and the tiny larvae, microscopic in size, travel first to the lungs. Here they go through a *molt*. This causes some coughing.

Whenever a child coughs for part of the day do not assume it is harmless. Use a frequency generator set to 408 KHz, or use a zapper. Children should be treated for *Ascaris* anyway, whether coughing or not, once a week. The tiny larvae are in the cough up. Children should not be taught to politely swallow this. They could be swallowing their own future seizures, asthma, or eczema. Teach children to use tissues for all spit up. The lungs are doing what they can to rid the body of these invaders. Your intelligence must cooperate with your lungs.

Ascaris do not come sweet and clean in themselves. They bring their own bacteria and viruses. One of these bacteria is *Bacteroides fragilis*. *Bacteroides* needs a host like *Ascaris* so it can be oxygen-free. Such a requirement is termed *obligatory anaerobic* meaning “must have absence of air.”

Being transported to the brain inside *Ascaris* larvae is probably the means by which *Bacteroides* gets into the brain. Brain abscesses and brain tumors usually have *Bacteroides fragilis* growing there. Brain tumors will not shrink unless all the parasites, bacteria and viruses are dead. *Bacteroides,* in turn, are big enough to house thousands of viruses. Two common viruses seen with *Ascaris* are *Coxsackie* B_1 and *Coxsackie* B_4. Perhaps it is the toxins of the *Ascaris* larvae or *Bacteroides* or *Coxsackies* that induces the seizures. Maybe it is something else about the infestation that induces them. But by killing *Ascaris, Bacteroides* and *Coxsackies* (zapper or frequency generator at 408, 325, 364, 362.5 KHz) you have eliminated the first essential link in the chain of developments that causes seizures.

The brain does not quietly tolerate these invaders. It fights back with its immune system. But the wormlets are already too big to be eaten by white blood cells. The brain fights back by producing **inflammation**.

Inflammations are intended to attract calcium so a wall can be built around the intruders. Inflammations are negatively charged regions so the positively charged calcium can find its way to the inflamed site. But lead and mercury are also positively charged! Perhaps this is how these **toxic metals** are attracted to the brain. All metals are positively charged. Perhaps this is their fateful poisonous attraction to living things. Perhaps they do not poison when no inflammations are present! In seizure cases we see many tooth metals in the brain. These must be removed so the inflamed site can heal.

Other toxic substances have also found their way to the inflammation at the seizure center: **vanadium** from leaking household gas; **PVC** from new carpets; **titanium** from face powder; **zirconium** from deodorant; **asbestos** from the clothes dryer belt or hair blower; **lead** from tainted drinking water. Do a good clean up of your environment.

Solvents accumulate here, too. Especially **toluene** and **xylene**. These are found in paint (persons with seizures should never be around fresh paint) but are also found in trace amounts in carbonated beverages. A person with seizures should drink no commercial beverages: see the Recipe section for homemade carbonated and other beverages. There are several other specific brain irritants that accumulate at the seizure center.

The food fungus, **ergot**, is always seen in seizure cases. Perhaps it can act alone to produce seizures. After all, seizures are an ancient malady, existing long before chemicals and solvents were manufactured.

MSG, which is *monosodium glutamate* should never be used in food. It was already banned in infant food in Europe a decade ago. Why are we still allowing it? Are our children not as precious? Healthy adults may tolerate it better than young children. But what about sick adults? Specifically, adults with a brain illness? Glutamic acid is a natural constituent of protein in our foods. And the sodium glutamate it must form in the body does no harm. But MSG is not biologically produced. It is lab-made. Lab-made amino acids are not the same as biologically made. Perhaps it is these “isomers”, perhaps it is the simple overdose of a natural thing that is brain-toxic. Whatever the mechanism, MSG should not be consumed by anyone, let alone a seizure sufferer. Fortunately, you can ask at restaurants, where (not whether) MSG is used so you can avoid these foods.

BHT and **BHA**, standing for *bishydroxytoluene* and *bishydroxyanisole* are food preservatives and also seizure triggers. They are often put on the boxes of cereals, rather than the cereals themselves, so the cereals can be pronounced preservative-free. Imagine how much the box must be drenched with to prevent oxygen leakage into the interior?

A more insidious seizure trigger is a perfectly natural substance, **malvin**. Malvin is the natural dye found in grapes, strawberries, plums and blueberries. Stop eating strawberry and grape jam or juice. Chickens and the eggs they lay, have lots of malvin too, stop eating chicken and eggs.

Here are foods relatively free of malvin: artichokes, asparagus, almonds, barley, beans of all kinds, green beans, broccoli, Brussels sprouts, cantaloupe, celery, nectarines, citrus, dates, mango, pears, kiwi, pineapple, Granny Smith apples.[14]

To Summarize

To stop your seizures on a dime, and not have another one:

1. Avoid malvin in food. Eat nothing colored red or blue, chicken, eggs, MSG, BHA & BHT. Boil all dairy products or don't eat them.
2. Avoid ergot in food. Eat no whole grain products; take niacinamide 500 mg three times a day to help the liver detoxify tiny bits in other foods.
3. Kill *Ascaris*, *Bacteroides* and *Coxsackie* virus and stay on a maintenance program of killing them. Avoid reinfection. It would be wise to have only outdoor pets.
4. Replace dental metal with metal-free plastic (See Dental Clean-up).
5. Clean up the home environment and body products of toxic substances.
6. Keep your fingers sanitary: spray them with 10% grain alcohol or vodka after bathroom use.

You can often tell by how you feel whether you are near to having a seizure. But some people get no warning. Don't take chances. Keep your drugs with you, even though you no longer need to take them. Drugs like Tegretol™ and Phenobarbital™ are not harmless. Once you can tell that you're better, try reducing your drug dosage slightly. See how you manage. If you have a breakthrough seizure you could fall and break a bone. Don't take chances. Even a year after your last seizure you should carry your medicine with you and have some in your house. This is because it is so easy to pick up *Ascaris* again. The highway to the brain–its seizure center–is still open. They will travel it again. It might only take two days from the time of accidental swallowing of animal filth, to having little larvae in the brain. Use a frequency generator every day at first, or zap daily to guard against this. Notice that killing parasites the first time may actually start a seizure coming. Simply stop zapping if you feel one coming. Do shortened zapping until you can tolerate a regular treatment. Have someone with you while zapping.

---

[14] Taken from *A Guide to the Identification and Treatment of Biocatalyst and Biochemical Intolerances*, 1988 by J. A. Krohn, Los Alamos Medical Center, 3917 West Road Suite 136, Los Alamos, NM 87544.

Say good-bye to seizures forever. But don't forget your susceptibility. I believe you should be extra careful for two years. If you have a recurrence, repeat the whole procedure. This time it should be curable in a single day.

Doug's mother was distraught when she brought him, age 8 months. Her doctor only seemed sympathetic with her purpose to keep the baby's temperature down the next time he catches a cold. But her doctor had referred her to the county social worker. She had been completely honest with her doctor, because she was that kind of trusting person. Since her baby had only experienced one seizure (which was during a fever), she didn't see why her beautiful first born child should be on medicine “the rest of his life”. She wanted her baby to be perfect. But the social worker had called her, talked about “the law” and being an unfit mother. She was all apart. The baby was supposed to be on Phenobarbital™ twice a day. Our tests showed Doug had *Ascaris* plus lead toxicity. He was also getting home made strawberry and grape juice. She promised to put the three cats outdoors, keep the baby off the floor, keep Doug's fingernails short and always wash his hands before eating. The parasites were easily killed. The lead was sporadically present in the water. She planned to move, and until then would filter all the drinking water so her breast milk would be free of it too. His diet was changed to eliminate eggs, chicken and the red and blue fruits. We recommended leaving the state in order to be able to peacefully raise her child. We promised to treat her child free of charge if another seizure should occur. Two years passed and we heard she was doing fine with her child.

Clara Scruggs, 50ish was losing control over her seizures and had to be hospitalized while a new medicine was tried. It changed her personality (again!) which upset her husband, too. She was started on the herbal parasite program but could only increase by one drop of Black Walnut Hull Tincture a week, instead of daily, since each new increase would give her a seizure. In half a year she was fairly free of seizures in spite of being off drugs. After each seizure, a checkup showed she had picked up *Ascaris* again sometimes with additional parasites. She could not bear to put her cat outside; Boots had been a friend in need many times. When she finally got Boots onto a regular parasite program she improved enough to go to church and church events again. When Boots finally "wanted out" she didn't mind. She decided to do a liver cleanse—this, too, gave her two seizures the next day but paid big dividends in other ways. She eventually improved to an incidence of one small seizure ("spacey" time or incoherent speech) in two weeks.

Chun Yee, age 28, has been on Dilantin™ from age 15. Now he was up to five pills a day and it didn't control his seizures. Any activity would trigger it. He was put on a low malvin diet and started on the kidney cleanse. His blood test showed he was anemic with a high MCV (104 cu microns), suggesting the *Ascaris* worms were using up all his vitamin B_{12} too. In six weeks he was down to one or two seizures per week, although he had not yet started the parasite program. He had beryllium built up in him, probably from a kerosene heater. When the pets and family were all treated for parasites he had no more breakthrough seizures and could cut his medicine in half which gave him much more energy. He was no longer in danger of losing his job.

Shiresse Nobel, age 7, was having minimal seizures but the mother did not want to start her on medicine. Shiresse had high levels of mercury in her body, although she had no tooth fillings. The whole family cleared their Ascaris infestation and in three weeks the seizures were gone but aggressive behavior remained. There was mercury in the air of some rooms but it was not in the paint. The bathroom was the worst. When everything in the bathroom was removed, the air cleared and so did Shiresse's behavior problem.

Drew Seaton, age 8, had his first seizure lasting nearly a minute. The parents were very fastidious and extremely conscientious about diet and habits. They were distraught. They all had *Ascaris*. A baby was still being diapered. Drew had arsenic (pesticide under kitchen sink), formaldehyde (some remodeling) and PVC (a new carpet) accumulated in his body. The carpets had to be steam cleaned to get rid of the arsenic. He was started on parasite herbs at once, since he was on medication that would shield him from having another seizure while killing *Ascaris*. Two weeks later everybody, except the mother, was free of *Ascaris*. In another two weeks they were all free and had cleaned up house toxins. They dared to stop his medicine. This let Drew's former happy personality come back to everyone's delight. Two years later there was still no recurrence.

Cosmo Maser, 30ish, was in a hospital across the country. He was having continuous seizures, although he had been there a week. It seemed impossible to transport him but, against doctor's orders, they got him into a station wagon. He had nothing to eat or drink that had any malvin in it (he ate four very well done hamburgers, plain, with lettuce on his trip) and his seizures stopped immediately. They felt a bit sheepish upon arrival 20 hours later since he could sit up, could tell his own story and no longer looked ill. He was without medication, too. They had left in a hurry. They immediately removed all the metal from his mouth; this cleared his mercury problem. He was started on parasite medicine and weathered the small seizures each increase gave him. He could return home in five days with his new diet and thioctic acid daily as a supplement. He occasionally had a seizure (2 a month) until they moved away from the busy street below their apartment. About a year later he could hold a job and go off Social Security support.

Weight Problems

Overweight is not the same problem as *obesity*. In fact, I do not have the answer to either problem, only a part of the answer.

In **obese** women, the ovaries, pancreas and thyroid are all involved. Perhaps the adrenals, the brain satiety center, and the liver are also involved. Maybe it's as simple as gold accumulation in all these places. Perhaps it is bacteria in all these places. When you weigh close to 300 pounds obviously some organ isn't working right. Try several things, but not a starvation diet.

The cause is not eating too much. Try removing all gold: gold teeth, gold jewelry and gold rings. Replace them with non metal varieties. After removing the gold, pull the remaining gold out of your tissues with thioctic acid (2 or 3 a day for several months). Make sure kidneys are able to excrete the gold instead of making crystals by doing a kidney cleanse. Gold accumulates in the pancreas, the brain (possibly in a control center here) and the ovaries (causes some infertility here). Also try clearing the body of all bacteria and parasites by regularly using a zapper. Use the Bowel Program (page 546) to evict the last of the *Shigellas*. Be very careful to avoid nonsterile dairy products. Try cleansing your liver by doing liver cleanses. Get 3,000 stones out.

Make sure you are getting enough nutritious food; make carrot and vegetable juice; use no commercial beverages. Avoid moldy food—don't take risks. If all these measures bring your weight down to the level of mere overweight give yourself good grades.

Overweight is a low energy condition. Your food is being turned into fat instead of energy. The decision not to make energy is being made in the liver mainly, but perhaps other organs as well.

Try cleaning the liver (page 552) until no more stones come out: get at least 2,000 stones. Notice that as the liver gets cleaner you get more and more energy right after each cleanse. Some cleanses have a dramatic effect. Others do not. This suggests that a certain part of the liver is the responsible part. Soon after the cleanse–within a week–the same old lassitude sets in. But during these few days notice how your body feels. It feels light. Your abdomen feels tight, like it's a part of you again. You're mind isn't on food throughout the day. It's very easy to lose weight, in fact you may lose five pounds in these few days without dieting or exercising. Definitely, your long-lost weight regulation is back in force. But then it vanishes. Fortunately, a bit of the weight loss stays with you, and by repeating cleanses (only once in 2 weeks, though) you can shed the pounds you want and gain energy in a permanent way. It is probably the way nature intended.

Try increasing your bowel movements. Notice how cats and dogs seem to derive energy from emptying their bowels. A cat walks to its litter box; after emptying its bowels and carefully covering it up, it jumps from the box and runs away. It now has its playful mood. A body chemical, *acetylcholine*, plays a role in emptying the bowels. Acetylcholine is a necessary operator for many of our muscles. Is there a disturbance in our acetylcholine metabolism in overweight conditions?

Coax your body to release more acetylcholine, at least in the intestines, by using a herbal laxative like Cascara sagrada. Other varieties are useful, too, but don't use a drug laxative: it burdens the liver more. Try to have three bowel movements a day.

Shigella bacteria can cause dreadful constipation. Immediately, the body feels sluggish, abdomen feels disconnected or hanging out. A nameless hunger sets in. Keep a close watch on dairy foods to make sure they are sterile. Zap when in doubt and do the Bowel Program.

Exercise helps as long as you keep at it. Strict dieting works as long as you keep doing it. But there's the problem. When you stop, you lose ground. It doesn't solve the problem of an inefficient body metabolism.

Raising thyroid levels helps but this can be dangerous. Raising thyroid levels naturally, by removing toxins is a very effective method—provided it was low to begin with. Overweight people often have a low body temperature, showing that the thyroid is involved: it is under producing. **But giving the body extra thyroid doesn't solve the thyroid's problem. It only temporarily solves the rest of the body's problem.** The thyroid will have viruses and heavy metals in it. The metal in your mouth drains downward to the stomach passing very close to the thyroid. Perhaps its iodine uptake is inhibited. The chlorine in water and bromine in bread may inhibit iodine uptake by the thyroid, too. After all, they are all halogens. Stop eating bleached bread and filter out the chlorine in your water. If this raises your body temperature you could expect better weight control.

The traditional herb, *Fucus*, was used to treat thyroid problems (and overweight) in days when herbs ruled medicine. Herbalists made a point of discouraging use of plain iodine. Fucus, they said, was much more effective. See *Recipes*.

If all these measures don't work for you, at least you have improved your health trying.

There are some advantages of being overweight. Overweight people seem to weather illness better. They laugh more. But one look in the mirror or at the scales ruins it. Put away the long mirror and scales. Don't ruin your whole life over it. Make a reasonable effort and then let go. Enjoy your stay on this planet.

Underweight can be just as difficult as overweight to correct. Once the stomach has been trained to say “full” or “full enough,” even after a few mouthfuls, it is difficult to heal. *Salmonellas* in the stomach wall are often seen. *Giardia* and other common parasites, that don't belong in the stomach, are seen here in cases of underweight.

The stomach is invaded by bacteria and parasites when its immunity is low. The common culprit is tooth metal draining down continuously into the stomach. Remove every bit of mouth metal. Don't cook or eat with metal. Stimulate your appetite with B-vitamins. And, again, the liver plays a role. If it is toxic with mold, it may say “eat no more” and the body obeys.

A stage of cancer illness is weight loss. This is called *cachexia*. The victim believes he or she is eating enough. They may not even have a poor appetite. Yet weight drops steadily. This metabolic problem has been studied scientifically. A chemical, **hydrazine sulfate** (prescription only), can reverse it to some degree. Use 65 mg three times a day for 30 days. Perhaps a lesson taken from cachexia metabolism could be applied to obesity. For the cachexia sufferer, life would end sooner if high calorie supplements were not used. But don't use canned “calorie boosters”. They are polluted—often with the very solvent that makes the condition worse. Instead, make an eggnog: ½cup boiled milk, ¼ cup boiled whipping cream, a raw egg (exterior carefully washed), 1 tsp. olive oil, a banana, honey, cinnamon, cloves, nutmeg to taste. Mix all in a blender. Drink 1 cup a day. Vary it daily to keep it interesting.

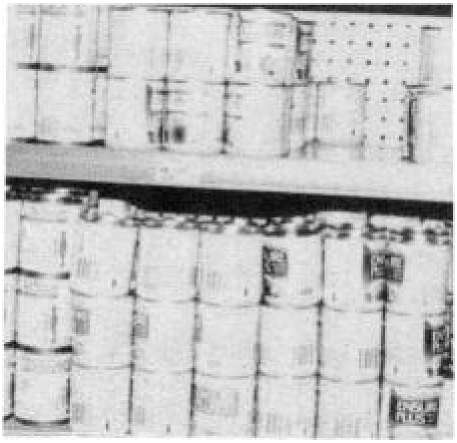

Fig. 33 I found all calorie boosters to be polluted with wood alcohol. Make your own.

If all this fails, give yourself credit for achieving health improvement and shift your focus to a different project. Underweight is not all bad either.

Sleep Problems

All big animals sleep, but some sleep by day instead of night. When humans do this, that is, work the night shift, they don't feel as well. Humans need about seven hours of sleep out of the twenty-four. Younger ones sleep more; newborns sleep much more. When we are deprived of sleep we are grouchy, think less clearly next day and have less energy. In spite of lots of research at "sleep labs" sleep problems are not understood, except for sleep *apnea*. Breathing should be even. When breaths are missed it is called apnea. It is especially disturbing when a baby shows apnea.

Sleep Apnea

Since breathing is regulated by acid levels in the blood and this is influenced by air quality, air toxins should be searched for first. **Cigarette smoke** is an air toxin. **Vanadium** from a gas leak is a very serious air toxin and can go unnoticed. Do your own checking since gas companies give wrong answers four out of five times. Ask a home construction company to check for gas leaks or the Health Department. **PVC** from new carpeting may be polluting the air. **Arsenic** from "treated" carpets and drapes and furniture also pollutes. **Asbestos** from clothes and hair dryers may be the toxin responsible. Perhaps even **fiberglass**, **formaldehyde**, or **freon**.

Adults with sleep apnea show swollen throat tissues: not necessarily pain. This makes the air passage smaller; long gasps of air are taken to try to make up for the missed oxygen and the carbon dioxide build up.

Swelling of the throat is a common allergic reaction. The possibility of allergy should get second consideration after air quality. Drug reactions, even in a nursing baby, where only the mother is using a medicine could be the problem. Allergy to food, chemicals has been suggested, as well as a simple lack of vitamin C (implicating mold and medicine which consume vitamin C in the detoxification process).

A third possibility is infection. Many bacteria and viruses can cause throat swelling. Redness of the throat is a telltale sign. It doesn't necessarily hurt. Kill all invaders with a zapper and try to understand the basis of low immunity in the throat.

Keeping metal in the mouth constantly, is a cause of low throat immunity since it must drain past the throat.

If you snore, you can deduce that your throat is swollen, even if you don't have sleep apnea. Pursue all three possible causes (air toxin, allergy, infection).

Overweight and obesity have been emphasized as causes. This may apply to some cases but certainly not to babies.

Whatever you do, don't do nothing. Keep removing bad things until you find the cause of irregular breathing.

Chester Fannon, 5Oish, was quite overweight and wore a mask at night with an air blower to assist his breathing. He had been referred to a sleep center for sleep apnea. He had extreme dryness of his throat at night and some hearing loss in one ear. He was toxic with arsenic (roach killer), bismuth (cologne), tin (toothpaste), and thallium (polluted dentalware). He was infested with both species of *Ascaris* and had a hacky cough. He had four solvents accumulated in his tissues. He was growing nine pathogens: *Mycoplasma*, *Haemophilus inf.*, *Streptococcus pneu*, *A-strep*, *Nocardia*, *Staphylococcus aureus*, *Bacillus cereus* and Flu virus, over half of them in his throat. These were killed with a frequency generator and a general cleanup was done. After two teeth were pulled he no longer needed his mask, he no longer had apnea.

Peter Day, middle age, was set for throat surgery in a few days. He had a couple of scary nights when he thought he was dying and couldn't risk many more nights with his obstructed airways. He was overweight. His throat was red and swollen, although he felt nothing. He never even had colds. His diet was completely changed, to things he rarely ate (bananas, milk, soup, oatmeal) and off things he ate daily (hamburgers, fries, tea, pancakes with genuine maple syrup). In two days, his throat was quite clear, the tissues having become unswollen. Maybe it was the molds in the maple syrup, maybe it was the oxalic acid in the tea, or something else he could not detoxify in these foods. He was certainly happy not to live the rest of his life with an artificial voice box.

Insomnia

Another sleep disturbance is waking in the night and not being able to go back to sleep for hours. Or not being able to get to sleep.

I believe these problems are caused by a high ammonia level in the brain. This belief is based on two observations. **Ornithine**, an ammonia reducer, induces a wonderful sleep in sleep-deprived persons. It is also observed that after killing parasites, which produce ammonia, sleep is much improved. Our metabolism does not produce ammonia. We produce urea which is excreted by the kidneys along with water and then called urine. When we are parasitized, our metabolism is burdened with ammonia, though, made by the parasites. We have to turn it into urea in the liver and kidneys so we can excrete for them. But this can't be done in the brain! The brain lacks an essential enzyme, *ornithine carbamyl-transferase*, for this bit of biochemistry. The brain was never meant to be parasitized or infected and has no defense. Most of our parasites come from animals we associate with. We weren't meant to live with horses, cows, sheep, pigs, monkeys, guinea pigs, cats, dogs and chickens nor to come in contact with dozens more at a zoo. We do so at our own peril.

It is known that ammonia is a strong brain irritant. In fact, a person can be awakened from a coma by being made to smell ammonia "smelling salts." Ornithine reacts with ammonia, mopping it up like a sponge. **Arginine**, another amino acid, also reacts with ammonia, but does not put you to sleep. So there is more to insomnia than mere inability to reduce ammonia levels. Arginine results in alertness and therefore should be used in the morning, when needed. Ornithine, given at bedtime, may take ½ hour to do its magic. Both are perfectly safe, since they are natural to your body, and a food constituent.

Start by taking two ornithine capsules (each 500 mg.) on the first night. Take four the next night. Take six the night after and choose the dose you like best. Sometimes it takes five days to "catch up" on everything that needs to be done for the brain and get you sleeping. Meanwhile, of course, you are planning to kill your parasites and be done with insomnia in the most effective way of all.

Another sleep aid is herbal. A couple of herbs, **valerian** and **skullcap**, are known for such action. The mechanisms are not understood and this makes for nonuniform action. Some persons sleep well with them, others do not. Simply try them to find out. We are all so different in our metabolism details, we respond differently to herbs. But it is a blessing that the mechanism is not understood. Herbs, a tradition that precedes civilization, need to be forever off limits for intervention by government agencies.

Tryptophane, another amino acid, is about twice as powerful as ornithine, but was taken off the market a few years ago. Some persons taking it daily were seen to become quite ill and some deaths ensued. Since tryptophane had been used in prior years without noticing toxicity, something unusual should have been suspected. My tests showed extreme pollution of tryptophane capsules. They contained PCBs, mercury, ruthenium, strontium, praseodymium, aluminum, and benzalkonium. I can only speculate that a mixing vat broke, dumping its precious load onto the floor—but it was salvaged. Or that the mixing vat wasn't cleaned thoroughly from it's last use.

Persons with illness due to taking tryptophane developed an extremely high eosinophil count in their blood test—an index of parasitism, too. Parasitism, that would have led to insomnia in the first place! Were these unfortunate victims seeing the cause or the result of their tryptophane use? This tragic event should have led to a discovery of the heavy pollution, a revelation of the industrial manufacturing process, and a safeguarding against any repetition. It has not been done (certainly not publicly).

Foreign countries' manufacturing processes do not come under U.S. scrutiny or jurisdiction, although some imported products must pass tests. There are no safeguards against repetition of the tryptophane experience. It behooves us to demand safe supplements and medicines. It is not the list of ingredients that informs. Lot analysis, after bottling, would give us the necessary safeguard. The presence of filth contamination and toxins cannot be completely avoided but the consumer can make informed choices if he or she knows it is there. Disclosure, of course, is the bane of the manufacturing business. Interest rate disclosure was the bane of the money lending business. Such important matters can't be left to “self-regulation” policies. The consumers must simply demand to know what they are consuming.

Ruby Adair, 14, ached all over, had ringing in her ears, sinus problems and chronic fatigue. She couldn't get to sleep, ever; and had been half a year out of school already. She had intestinal flukes in her stomach. In three weeks she had eliminated them with parasite herbs and she could go to sleep naturally.

Yeast Infections

The most common yeast in humans is *Candida albicans*. *Candida* has always been around. It flies in the air, searching for a place to land and reproduce. It can invade a variety of human tissues like the mouth (called thrush), skin (including some kinds of diaper rash), vagina, and the digestive tract. We all have some yeast in our digestive tract, but when it gets out of hand, it's called *candidiasis*.

Yeast is a **fungus**. It needs dampness to survive and sugar to grow.

Our immune system, white blood cells, are capable of eradicating yeast provided it isn't growing too fast. And provided the white blood cells aren't immobilized or preoccupied with something else.

Diaper Rash

A baby's rash is an example of the white blood cells being preoccupied. When chemicals are used in the diaper, the white blood cells go after the chemicals and let the yeast grow. Drying the baby's skin helps since the yeast must have dampness. This should be done with air, sunlight and a heat lamp, not with more chemicals! Certainly not with cortisone containing salves that further reduce the immune competence of white blood cells.

Use a heat lamp for five minutes at a time, several times a day. Switch to cloth diapers; do not bleach them with chlorine bleach, the residual chlorine trapped in the cloth is a chronic irritant, setting the stage for another rash and future chlorine-allergy. Cloth diapers should be sterilized, not bleached. Use the hottest water your laundry system is capable of producing. Add ½ cup borax for the washing process. If you have homemade Lugol’s iodine (made by your pharmacist or by yourself, see *Recipes*), add a tsp. to the wash or rinse. Vinegar is a yeast inhibitor, add it to the rinse. Dry diapers at the hottest setting. Dry to kill. Kill all the yeast spores in the diapers.

To strengthen the baby's skin against future infection, do not put chemicals on the skin. Do not use any soap, fragrance, bath oil, ointment or lotion. Do not use cotton balls or baby wipes. Do not give a daily bath. Wash bottoms gently, with borax followed by a vitamin C rinse. Vitamin C is acid and is our natural healing agent but it will sting on a broken skin surface. Use it as dilute as necessary to be tolerated. Zinc oxide is another natural healer because it competes away the iron that fungus and bacteria need for their reproduction. Never use commercially available zinc compounds though, simply purchase your own zinc oxide powder, mix it with cornstarch and keep in a large old salt shaker, dust it wherever there is moisture or fungus growth.

Treat Yeast or Fungus the Same

Other fungus growths, like *Tinea* (crotch itch), ringworm (not a worm at all), athlete's foot, along with *Candida*, can be similarly eradicated:

1. Deprive the fungus of moisture.
2. Deprive the invaders of iron.
3. Deprive the fungus of sugar.
4. Strengthen the skin's immune power.
5. Strengthen the skin's healing ability.

It may be impossible to deprive the fungus of moisture, for example if your feet sweat and you must wear socks. Take your socks off as soon as you are at home, treat your feet with a heat lamp. Use zinc oxide or cornstarch to powder and dry the skin. Boil your socks when laundering. Dry them to tinder-heat (too hot to touch). Launder with borax only (soaps and detergents contain aluminum which pollutes the skin). Rinse skin with vitamin C water. It takes all these measures used simultaneously to clear up athlete's foot fungus. And great persistence. They may have developed a foothold underneath the toe nail where a steady supply of moisture, iron and sugar is available to them. Nevertheless, your white blood cells will eventually gobble them up if you let them.

In thrush (yeast infection of the mouth) you must again outwit its growth by doing everything possible at one time. Eat no sugar, drink no fruit juice, stay off antibiotic. Avoid trauma like eating abrasive foods (crusts, popcorn, nuts, lozenges) or sucking on things. Floss teeth only once a day (using monofilament fish line), followed immediately by brushing with white iodine (or Lugol's, but this may temporarily stain). Hydrogen peroxide is not strong enough. Remember to sterilize your toothbrush with grain alcohol or iodine. You may also rinse your mouth with Lugol's (6 drops to ¼ cup of water). Or apply 6 drops directly to the tongue and rub it in lightly with your lips.

Do not use Lugol's iodine if you have been told you are allergic to iodine.

Kill *Candida* daily with a frequency generator or zapper. Since reinfection is constant, you must continue to do all the treatments given to permanently cure yourself of fungus disease.

Since *Candida* grows right into your living cells (which you are not attacking!) you cannot kill it all at once. Only surface fungus can be accessed by either Lugol's or electrical "zapping". But as the top layer is killed, exposing the next layer, you will make progress. It will take a month of daily treatment to clear it.

Clearing up fungus at one location but not another will not bring you a permanent cure, either. Damp locations like under the breasts, under the belly fold, groin and crotch need to be kept dry with cornstarch daily. Keep it up long after it seems to be cured.

Fluke Disease

Flukes, or flatworms, have a complex life cycle with many stages. Although sheep, cattle, pigs and humans can be "natural" hosts to the **adult** stage, the other stages are meant to develop outdoors and in secondary hosts. When fluke stages other than the adult are able to develop in us, I call it *fluke disease*.

Or, when an adult that "normally belongs" to another species is able to develop in us, I also call that fluke disease. Or even with adult flukes in their "normal" host, when they move from the organ that they "normally" colonize to other organs in the body I call this fluke disease, too.

Four fluke varieties engaged in this extra territorial pursuit are the **intestinal fluke**, **sheep liver fluke**, **pancreatic fluke**, and **human liver fluke**.

As you can see from their names, scientists have studied them well, and know exactly which animals are the "normal" hosts, and which organ in that animal is the adult fluke's "normal" home. Fluke disease is when any of these is "wrong."

Flukes don't have eyes to see with or legs to walk with, so how can they find and travel to the organ they want in the middle of your body? Scientists do not know for sure. However it's concluded from many scientific studies that the liver fluke, *Fasciola*, for example, has no trouble seeking out and colonizing the liver.

Here are some examples of what can happen when flukes go "wrong:"

- Adult flukes (any of the four mentioned) in the uterine wall causes cramping and bleeding when it is not menstrual period time. If an adult crosses the wall to the inside and then manages to get out through the fallopian tubes to the abdominal cavity it takes some endometrium with it—causing **endometriosis**.

- If adults develop in the kidneys, it can cause **lupus** or **Hodgkin's disease**.
- If adults complete their cycle in the brain, **Alzheimer's** disease and **multiple sclerosis** result.
- If the intestinal fluke (*Fasciolopsis buskii*) becomes adult in the liver it causes **cancers** of many (hundreds) kinds.
- If the pancreatic fluke completes its cycle in the pancreas it leads to **diabetes**. This is not an example of flukes straying into the wrong organs, but of having its stages reproducing where they never could before.
- If flukes develop in the thymus, immunity is lowered. If it happens to be the intestinal fluke, **HIV** (Human Immunodeficiency Virus) is released there. In turn, HIV invades other tissues, like penis and vagina.
- These four flukes can also invade the muscles, causing **dystrophies**.

As dissimilar as we always thought these diseases to be, it's obvious to me that they are but one disease—fluke disease

Considering the size of these flukes (adults are easily visible), it is not surprising that they can quickly lay waste a human's organs. Yet a human is big and makes a valiant effort to kill the stages, block access to tissues and otherwise battle them.

But only the human's intelligence can be counted on to defeat them. The intelligent approach is to discover what enables these mighty monsters to do their reproducing in our bodies instead of the pond with its snail/minnow secondary hosts.

Flukes and Solvents

The explanation I have found for two of these flukes is the presence of solvents in our bodies. The presence of *isopropyl alcohol* is associated in 100% of cancer cases (over 500 cases) with reproduction of the intestinal fluke stages in a variety of organs causing cancers in these organs.

The presence of *benzene* is associated in 100% of HIV cases (over 100 cases) with reproduction of intestinal fluke stages in the thymus.

The presence of *wood alcohol* is associated in 100% of diabetes cases (over 50 cases) with reproduction of pancreatic fluke stages in the pancreas.

The presence of *xylene* and *toluene* is associated in 100% of Alzheimer cases (over 10 cases) with the reproduction of intestinal fluke stages in the brain.

Much more work needs to be done to examine the relationship between fluke reproduction, the solvent and the chosen organ. But it seems probable that the solvent allows it all to happen. And our intelligence, to save us, must find a solution.

Stopping use of these solvents seems to me to be the most urgent advice. Finding which foods and products are polluted with them is the first step. It is imperative that you test everything you use or eat for solvent pollution. The Syncrometer makes that an easy task. Ideally, we should all pool our results, adding to the body of knowledge I have begun.

In my observations, when the big sources of solvents are stopped, the body's levels go back to zero. In other words, the minute amounts that we inhale here and there do not accumulate to the point of serious damage. We have to eat, drink or absorb them on a daily basis to injure us! So where can they come from?

The sources of benzene and propyl alcohol that I found are given in special lists (page 354 and 335). The sources of wood alcohol are not as well known, but include commercial beverages, cold cereals, artificial sweetener, vitamins, and drugs. Other solvents are even less studied. But a pattern is emerging: foods and products that require sterilization of bottles and machinery to fill these bottles are polluted with propyl alcohol or wood alcohol. Foods and products containing flavorings or oils are polluted with benzene. Let the buyer (you) be wary! Test your own products if possible. If not, do not purchase them.

There are many other flukes and many other diseases. Are there other fluke/solvent/disease trios? Has fluke disease been going on for a long time or is it a recent phenomenon? Certainly cancer is 100 years old, so is the use of propyl alcohol. Diabetes is quite old as an illness, too, and so is its associated solvent, wood alcohol. But HIV, AIDS and Alzheimer's are recent diseases. Should we conclude that benzene, xylene and toluene were used much less in the past?

Fluke diseases could be eradicated with some simple actions: monitoring of solvents in foods, feeds and products. Hopefully, this will begin. It is in the interest of the consumer to have her or his own independent way of monitoring too. Chemical ways can be devised, besides the electronic way presented in this book. Imagine a small test strip like a flat toothpick which turns color when in contact with propyl alcohol. Keep a pack in your pocket and never be unknowingly dosed again...all in tomorrow's world.

Flukes Not Alone

There are other families of parasites. The roundworms and tapeworms are gaining ground too. Are they associated with solvents? Or with yet undiscovered factors? Are they changing their life cycles to take advantage of our lowered immunity? These are important questions. But you are armed with excellent technology. The answers will be found.

And along with answers there will surely develop a new industry. An industry that not only proclaims purity for its products but provides the proof to your satisfaction.

Burning And Numbness

Burning sensations in the skin let you know that nerves are involved. Mercury is the most common offender. Mercury may have started the trek of a host of other toxins as well into your nervous system: pesticide, automotive chemicals, household chemicals, fragrance and even food chemicals. Some people can get a burning sensation after a car trip, some when exposed to perfume, some when walking down the soap aisle in a grocery store. When the affected nerves don't go to skin but instead go to an organ like toes, you might feel a toe cramp or finger cramp instead. Remove your mercury sources.

Burning skin is an ancient malady—maybe even the basis of concepts like “hell.” St. Anthony's fire was caused by *ergot* (food fungus) ingestion. Molds don't necessarily come singly. Maybe other mold toxins can go to your nerves, too (see Moldy Food, page 381). Maybe the mold toxins interfere with **pantothenic acid** used by your body, because giving pantothenate (500 mg three times a day) can sometimes relieve the condition and, of course, this is good for your body.

Monosodium glutamate (MSG) can cause burning, especially of the face and lips. Sometimes swelling occurs too. MSG is used as a flavor enhancer. It was found decades ago to be a brain toxin and was taken out of baby food. But what about adults? Especially those who already have a brain problem. Throw it all out of your kitchen. Ask at restaurants which foods have it.

Numbness has a similar cause. Numbness of fingers or feet has become quite common since thallium and mercury toxicity has spread so widely.

If you have burning and numbness, can multiple sclerosis (MS) be far away? Remove all the metal in your dentalware immediately, replacing with composite (see Dental Cleanup, page 409). Hopefully, your immune system is still strong enough to clear the bacteria growing around the metal and in pockets in the jaw. Check and clean cavitations. Start your jaw healing with the milk, magnesium, vitamin D diet. Use thioctic acid to help clear your tissues of remaining metal (3 to 6 a day).

The most common nervous system bacterium is *Shigella*. Its beginning can already be seen in cases of burning and numbness. It is deeply entrenched in cases of MS. Three kinds of *Shigella* are readily obtainable on slides: *Shigella dysenteriae*, *Shigella flexneri, Shigella sonnei*. *Shigella flex* causes depression and irritability. All can cause gas and bloating. Zapping doesn't kill all of them because they inhabit the bowel. You must empty your bowels frequently, two or three times a day. And do the Bowel Program until all symptoms are gone. Be very careful not to put contaminated dairy food in your mouth again. Boil all milk products. They must be at boiling point for 10 seconds. This includes cheese, cottage cheese, buttermilk and regular milk. Eat only home made yogurt or home made buttermilk. Butter and whipping cream need this treatment, too. Never use raw dairy products.

Nana Hughes, 48, had numbness of the whole right arm, hand and right side of her head; it was particularly bad in the last four months. She was toxic with PCBs, titanium and dysprosium (paint). She was on thyroid, Xantac^{TM} (for stomach) and chlorazipate medicines. We also found dog heartworm (she had chest pain over the heart). She started on the parasite program, stopped using nail polish, and stopped all detergents for dishes or laundry. In three weeks her numbness was greatly reduced. She still had titanium buildup from dental metal (a partial bridge). Seeing she was on the right track, she stopped use of MSG in her food, switched from concocted beverages to milk and cut down on her smoking. She brought her arm pretty near to normal.

Maria Santana, 45, had numbness in both arms; they would tingle and "go to sleep" a lot. It was spreading to one leg. Her muscles were toxic with thallium. It was in some very old pesticide, still effective and still in use in the house. It was also in the well water, probably, from pesticide seepage. She went off all commercial body products, did a kidney cleanse and killed parasites. She had difficulty getting rid of *Prosthogonimus* but in two months she had everything cleaned up. Her legs, arms, sleep problem, urinary tract problems were all gone and she could focus on her last problem, digestion.

Candy Donaldson, 44, had numbness from her shoulder to the wrist of one arm, it started a year ago. She was toxic with iridium, lithium and vanadium from a gas leak. She was advised to stop caffeine use and switch to milk (her calcium level was low: 9.0 mg/DL) and a magnesium tablet (300 mg daily). Her triglycerides were high, implying a kidney problem. She had urate and phosphate crystals in her kidney. She decreased the phosphate in her diet (meat, nuts, grains, soda pop) and started the kidney cleanse. When the gas leak was fixed, both her lithium and vanadium toxicity disappeared. In six weeks she had also killed parasites and her periods became regular for the first time. An ovarian cyst had disappeared, too, as checked by ultrasound. Soon her PMS was gone, including hot flushes. After four months she had done three liver cleanses and suddenly her numbness improved. After a few more there was no numbness at all.

If cleaning cavitations brings you immediate improvement you know that these bacteria were part of the problem. Have them checked again if problems return; dental bacteria are notorious for returning. If kidney cleansing makes it worse for a day and then better, you know kidney bacteria are partly responsible. If liver cleanses (page 552) make matters worse for a day and then better, you know bacteria are entrenched in the liver. Continue cleansing until none of these sources exist any longer. Kill all bacteria at least once a week electronically. Burning and numbness can be cured, not just arrested. In other words, your nerves can be cleaned up and healed.

Depression

All persons I have seen with clinical depression had small roundworms in the brain. Is it any wonder the brain can't make enough neurotransmitters or gets them out of balance? The usual worms are hookworms (*Ancylostoma*), *Ascaris* of cats and dogs, *Trichinellas* and *Strongyloides*.

Although it is commonly believed that hookworms penetrate the skin when walking barefoot on earth, this appears to be a negligible route. The important routes are eating animal filth and inhaling filthy dust. Our pets pick these worms up daily. We get them and give them in constant exchange with our pets and family members. Diapering babies is an especially hazardous, though necessary, business. Letting little children clean up after their own bowel movements is even more hazardous. Hands should be sanitized with grain alcohol after dealing with bowel contents, whether your own, your child's, or an animal's. If you clean up a messy diaper and then your hands well with soap, then go to make the chopped salad for dinner, you're sure to give each family member a dose of whatever the baby has. It was hiding under the fingernails. Animals clean up the easy way: they simply lick the youngster's bottom. But, we humans are not strong enough to take on a dose of bottom with each meal. **We must sanitize our hands**.

Cleaning bathrooms is also dangerous. Wear gloves. If nobody suffers from depression, you can use bleach (stored in the garage) to disinfect the stool, otherwise use alcohol (50% grain alcohol). Keep doorknobs and faucet handles wiped with alcohol, too.

If you suffer from depression use your zapper to immediately kill these four roundworm species: *Ancylostoma*, *Ascaris*, *Trichinella* and *Strongyloides*.

Other family members should be cleared of these four worms on the same day or as close to it as possible. Reinfection always occurs. In the depressed person, the microscopic parasites travel immediately to the brain. In others, they may simply reside in the intestine or lungs or liver, or other organs. Pathways (routes) to the brain have become established for the depressed person. These must heal before there is any tolerance to reinfection.

Solvents and other toxins, also in the brain, slow down or prevent healing. For this reason the depressed person should do the *Four Clean-ups*.

The bacteria of the *Shigella* family are always seen in depression cases. Kill all *Shigellas* and avoid reinfection by boiling dairy products and not eating those that can't be boiled. Go on the Bowel Program (page 546). Do not eat deli food or hand prepared salads at salad bars. Finally, do the Liver Cleanse, repeating every two weeks. Depression, even of long standing, can lift within days after the brain finally has its territory to itself. Look in the mirror and smile at yourself for your success in vanquishing your invaders. Never again, let these creepy crawlers into the happiness-center of your brain.

Manic Depression

This variety of depression is associated with *Strongyloides*, as the main parasite in the brain. Plus **chlorine** as an allergen! *Strongyloides* is the same worm that causes migraines and other severe types of recurrent headache. It probably depends on where these tiny worms have set up their “housekeeping” which symptoms you get: manic depression or migraines. Maybe chlorine is the deciding factor.

These tiny wormlets can pass the placenta into the unborn fetus. Is it any wonder that these brain disturbances seem to be inherited? Of course, there is ample opportunity to simply eat them off hands, other people's hands, and off floors during childhood. The amazing truth is that some family members do not get infected with it or at least do not get brain symptoms! It is very difficult to eradicate *Strongyloides* in a whole family and thereby let the depressed person get well. It is, in fact, impossible if there is a pet or other animal connection. Getting rid of the chlorine allergy is also very challenging. Step one is to zap all parasites.

I usually see an accumulation of **bromine** as well. Since bromine, fluorine and chlorine are all *halogens,* maybe getting too much of any one could saturate the liver's ability to detoxify chlorine. Stop eating brominated (bleached) bread. Stop taking drugs containing bromides. That is the easy part. You must also stop even washing your face in chlorinated water (use a pure carbon filter system). You inhale it as you wash. Never drink it or use it for any purpose. Of course, there should be no bleach container in the house, even when tightly closed; nor should bleached clothing be worn.

Do the *Four Clean-ups*. Sometimes the cloud lifts and the mind clears on the way home from the dentist. Recovery can be very rapid—less than a week. And recurrence just as rapid when a tiny bit of chlorine is inhaled or drunk. But your careful vigilance pays off. In half a year you can expect no recurrence.

Humans, it seems, must lick fingers with the same compulsion that cows lick their noses and cats lick their rears. The single, most significant advance in human hygiene would most assuredly be stopping the hand to mouth habit. We must always eat with utensils. And never leave a bathroom without washing. These are difficult changes but the new age of parasitism makes it necessary. Together with the new pollutants, solvents, and heavy metals, parasites will overtake us unless we change.

Don't go off lithium and other medicines until your doctor agrees you are ready, about a half year. After this, make sure you still keep it handy. Although you may be free of manic depression in a day, reinfecting yourself weeks later will attack your brain like a hurricane; it has not yet healed, the routes are open. Be patient; healing will happen as it did for these happy persons.

Lena Constantine, age 39, had a history of migraine, TMJ, heavy clotting with periods and numerous pains but it shouldn't have made her try suicide 1½ years ago. She was put on Prozac™ afterward. She was parasitized by intestinal flukes (in the intestine), dog whipworm, *Strongyloides* and human liver flukes. She was started on a parasite program and kidney cleanse. This made her feel so good she took herself off Prozac™ and landed in the hospital for reasons she couldn't remember. After 42 days spent there she got out, wiser than before. She set to work again, leaving no detail undone, because she could remember how good it felt to be free of depression (not drugged out of it). Three months later she still had *Strongyloides* (she had a cat) but she did her first liver cleanse anyway. She got over 500 stones out. Her depression was gone. She substituted 4 ornithine and 2 ginseng capsules daily (more if tension was not relieved) for Prozac and cured her problem.

Mona Zabala, 33, was extremely depressed over her job. She was full of mercury, arsenic, PCBs, chromate from eye liner, wood alcohol from drinking colas and she had *Ascaris*, pancreatic flukes, *Trichinellas* and *Strongyloides*. She thought she was a hopeless case. But in less than three months, when only half her clean-up chores were done, she was already saying positive things about her job.

Acey O'Hara, a young graduate student, was very careful to avoid junk food, caffeine, fragrance, and body products because he had learned the hard way that he didn't tolerate them. When “crying” depression hit him he was not only surprised but angry that his good health habits “hadn't paid off.” He was buying his drinking water; it came in clear plastic containers. He had cesium in his tissues and brain. When he switched back to plain tap water (filtered in small quantities) the depression lifted in a week and he was no longer crying over anything.

Leisa Underwood, 46, was clinically depressed and in therapy. She believed it was due to going through menopause. She had a heavy burden of *Strongyloides* but no other worms. Only one of her two dogs had *Strongyloides* (saliva test) and the cat was free of them also. She was full of cesium (from drinking refrigerator water) and vanadium (from a gas leak). In two months she had accomplished the impossible: all pets and herself were free of *Strongyloides*, they had repaired three gas leaks and her depression was just a memory.

Roland Greeley was diagnosed as manic depressive a few years ago and was put on Ativan™ and Prozac.™ This controlled him but did not relieve depression. He had a prominent tremor. He had *Ascaris* and sheep liver fluke stages in the brain. Also *Trichinella*, *Strongyloides* and human liver fluke stages. Styrene (from styrofoam cups), methyl ethyl ketone (beverage) and carbon tetrachloride were in his brain also, probably setting the stage for parasite reproduction. He had high levels of mercury and silver but highest of all–throughout his body–was chlorine (from bleach and tap water). In four weeks he had all the metal removed from his mouth. He stated that this "unchained" him. He could already tell on his way home from the dentist that something special had happened. His depression was simply gone "the way you lift a blanket off a bed". He resolved to clean up his whole body and recover from his illness using logical methods, like ours. Staying away from regular chlorinated water was a fine challenge to his resolve but with whole house filtering now available he may have done it.

Darren Knox, age 48, had been on Thorazine™ for 36 years but was recently changed to Desyral™ and Valium™ and lithium. He had *Ascaris* and hookworm and two dozen more assorted parasites including fluke stages. His brain was also hosting *Bacteroides* and *Nocardia*. All parasites were killed in half an hour by frequency generator at his first visit whereupon he immediately announced himself free of depression; better than the last eight years. He had many metals in his brain, also chlorine. In four weeks his tremor was way down, after dechlorinating his tap water. His depression never came back after dental work was done.

Schizophrenia

Much more mold toxin was seen in schizophrenic families than in other kinds of illness. They usually had four or more kinds of mold toxins at the same time, meaning that one toxin was not detoxified before the next was already eaten. It may reflect on the liver as much as the food selected. The initial injury, though, to the liver may have occurred in early childhood.

Schizophrenia does not require mercury or other dental metal pollution for its expression. This pattern is logical when it is seen that young children can have schizophrenia. Schizophrenia is an ancient illness, being described in some very old literature, before dentistry existed. But parasites existed. And moldy food existed. And copper existed!

Ergot is always seen in sick persons. Other mycotoxins are also present, including *sterigmatocystin*, *cytochalasin B*, and *aflatoxin*. As the mycotoxin panorama changes, brain symptoms can change from compulsive hand washing to paranoia or from hearing voices to meanness in disposition.

It is my belief that the current increase in violent crime in U.S. society can be attributed to the especially high levels of mold consumption in foods and beverages, with the resultant effect that behavior erupts into violence with almost no provocation, merely some frustration. It would not be difficult or expensive to experiment with a mold-free diet in our prisons.

A bacterium, *Mycobacterium phlei*, may also be a cause because I have seen a high correlation between it and schizophrenia cases. What its specific effects are, I don't yet know. This bacterium may be hiding under teeth. It is also seen in dogs.

Shigella bacteria from nonsterile dairy foods are part of the problem, too. They produce brain and nerve toxins resulting in irritability, depression, anger. In fact, anger can be so intense, it erupts in violence.

Lead and copper are commonly seen in schizophrenia sufferers. The usual source for these is the household water (household plumbing may have lead solder joints). Change plumbing to PVC, then take thioctic acid (100 mg, twice a day).

Parasites always found in schizophrenia are hookworms (4 *Ancylostoma* varieties) in the brain.

Instructions:

1. Stop eating all grain products, nut products and syrups immediately.
2. Sterilize all dairy products.
3. Search for lead and copper in the water.
4. Stop drinking commercial beverages, including bottled water.
5. Zap the parasites in the whole family for three days, followed by repetitions twice a week.
6. Eliminate *Shigellas* with the Bowel Program (page 546).
7. Give away pets. Schizophrenia is too serious to risk reinfection.
8. Do the dental cleanup (page 409).

The need to zap repeatedly arises from reinfection. Nothing needs to be killed twice. But reinfection from self, family members, pets and food happens every hour. One cannot rely on zapping to stay well. Do a thorough diagnostic search of all foods eaten at the last meal, the water drunk, the air breathed. Only when all these are clean, will the brain heal.

When can you eat grains again? When you are well. Find a cereal or pasta that has no mold. Test it for the mold frequencies (77, 88, 100, 126, 131, 177, 188, 232, 242, 277, 281, 288, 295

KHz). Search especially for ergot (295). Remember, ergot makes *lysergic acid diethylamide* (LSD), not a good thing for a schizophrenic brain. Honeys, too, have ergot. Ergot takes longer (up to 20 minutes) to detoxify with added vitamin C. (Read the section on Moldy Food, page 381). If you become ill again after starting grains, go off again. Take 500 mg. niacinamide 3 times a day to speed up ergot detoxification by the liver. Avoidance is much easier.

Healing of the brain is very rapid; in less than one week feelings and behavior are more normal. Perhaps there are herbs that hasten healing; considering how old the illness is, there must surely be several useful herbs. But considering that herbs, too, can be moldy, be very careful to search for molds electronically before using any herbs. Blue vervain, sage, and ginseng are herbs worth trying.

A question may have popped into your mind as well as mine. If the whole family is eating moldy food, why aren't they all schizophrenic? In fact, family members usually do suffer from some symptoms that are similar to the victim. But every person's collection of parasites and pollutants is unique. Certainly, the whole family should obey the moldy food rules, in order to function better.

Autism

Childhood brain disorders suggest an inherited genetic defect. Yet numerous parasites and pollutants are able to pass into the unborn child through the placenta. Even some bacteria and viruses can.

Lead accumulation is always seen in children with autism. Was it acquired before birth?

Mercury can be transmitted from mother to child, too. If mercury fillings were not removed before pregnancy, have them replaced as soon after delivery as possible (never during pregnancy since the removal itself causes a surge of heavy metal to enter the body).

The common tiny worms such as *Ascaris*, hookworm, *Strongyloides* and *Trichinellas* easily enter the brain. They are present in dirt. Dirt enters buildings via shoes. Shoes carry animal filth. Don't let children put their own shoes on until they have learned to avoid touching the soles. Don't let your child crawl on the floor of a public building, even though it looks glossy. Don't set shoes on furniture or table! A preferred habit is to leave shoes at the door.

Once wormlets have found a pathway to the child's brain it is difficult to reroute them. They must all be killed repeatedly since there is daily reinfection from putting hands in mouths. All family members should kill these parasites weekly to protect the child with autism. When lead and parasites are gone consistently for several weeks the pathway to the brain heals and reinfection no longer sends them to the brain and your child can resume a normal life.

I have not treated enough cases to point to a particular parasite or pollutant. For this reason you must do a total cleanup: body, environment, dental, diet (especially solvents and molds).

Leon Dickson, age 10, appeared normal at birth but he crawled and walked late. At age 3½ he began having seizures and was also diagnosed with autism. (Seizures are caused by *Ascaris* larvae, they probably began the routing to the brain). He was started on Phenobarbital™ by his clinical doctor; switched to Dilantin™ then two other drugs were tried. Now he was on Tegretol™ plus Depacote.™ The mother used no anti nausea medicine during pregnancy, no caffeine, no alcohol or nicotine, not even a single aspirin. In spite of his medicines he seemed to have headaches, continued to have seizures which disturbed his sleep and he wasn't trying to talk. He vomited a lot. He would take no pills or drops (no herbs even mixed with honey) and our frequency generator method was not discovered at that time. He couldn't kill his parasites and *Shigellas*. His brain was full of thallium (not mercury!)

besides the roundworm larvae. He must have gotten this from Q-tips™ or cotton balls or some other mercury-sterilized product. Leon's favorite food was grilled chicken which he ate nearly every day, as well as two eggs a day.

His diet was changed to exclude chicken, eggs, bacon, chips, preservatives and colors in foods, grape jelly and strawberry jam. One month later he had not improved, nor had they been able to kill his parasites with the herbal recipe. The diet change was extremely difficult; he was screaming for his favorite junk food and the whole family was upset over his restrictions. But we encouraged the mother to stick to her purpose, get a different baby-sitter who would obey her, and to try to get some parasite herbs and thioctic acid (100 mg. daily, stirred into honey, to take out lead) down him. The first week the new baby-sitter succeeded in getting him to take thioctic acid. He had only one seizure that week. This encouraged the mother to enforce the parasite program and the diet rules. He became attentive in one month and tried to voice sounds.

The problem with eggs is that they are contaminated with *Salmonellas*. I find, however that it is the outside of the eggshell and the carton that is contaminated. The safe way to handle eggs is to remove them and return the carton to the refrigerator, then wash the eggs and your hands before cracking them. Eggs also contain malvin, a known seizure trigger.

Kirk Peeples, age 5, did not have any words yet but he would point to something and voice M-M-M to mean he wanted it (usually food). He was in a special school and “doing well.” He had the brain toxin MSG, and antioxidants BHT and BHA accumulated in him. In other words, his liver was not able to detoxify these common food chemicals. At that time, I had not found the mold toxins yet; I would now surmise that they were responsible for the liver's disability. Besides going off these food additives he was “desensitized” to them with homeopathic drops by an alternative allergist. The result was immediate. In one week he tried several new sounds and managed his first word: “box.” After killing parasites with the herbal program over a four week time period he could speak 19 words. The parents were bewildered with joy. They cleaned the toxins out of their home so thoroughly, there was no place to sit and nothing familiar to eat. But their son could say things and the parents loved each new sound as if it came from a newborn baby. In two months he was saying two-syllable words and putting phrases together. He needed us no more (he was on homeopathic thyroid drops, too).

Geoff Berkely was diagnosed autistic. He went about his business of examining everything in the office without a word. He was infested with both species of *Ascaris* (there was a pet dog) and was started on the herbal parasite program: just a little less than the adult doses. (Children's' syrups are not as effective.) He was toxic with mercury, too. He had 2 or 3 “baby” root canals. These were to be pulled out. He was also taken off food color, MSG, chicken and eggs. The parents accomplished all this very quickly and called to say their child had become normal but they didn't want us to spread the good news to the “autism” club they attended for fear of criticism for doing unorthodox things. Only in the USA could such thinking occur!

Digestion Problems

Burping, bloating, and being gassy are signs that your digestion isn't perfect. What has gone wrong? Burps are gases escaping upward. Bloating is due to bubbles of gas causing pressure. Your body does not make gas! Gas can only be made by bacteria. The immediate conclusion is that bacteria are growing in your digestive tract (stomach and intestines) that should not be allowed to do so. They are likely to be the common enteric (digestive tract) bacteria: *Salmonellas*, *Shigellas*, *E. coli*, *Bacteroides fragilis*.

You may wish to identify them before killing them with the zapper. Or you can sweep through the whole bacterial and viral range killing all with a frequency generator. The good effects can be felt in an hour, although the last gases may take days to get rid of. *Salmonella* and *Bacteroides fragilis* are two bacteria that can eat bile, and can do without oxygen, so they are commonly seen in the liver. *Bacteroides* probably escaped from its roundworm host, *Ascaris*. *Salmonella* probably came with non-sterile dairy food.

If you have an intestinal problem involving digestion or pain, start immediately to boil all dairy foods. Stop eating those that can't be boiled: cheese sandwiches, yogurt, ice cream. It must be boiled for 10 seconds. The bacteria are in the liver because your liver attempted to strain them out of your blood and lymph in order to kill them with bile. Instead, they turned around and "ate" the bile, turning it brown as evidence. Now, every time the liver lets down bile into the intestine (and stomach), a population of these bacteria goes with it. The intestine becomes a seething mass of bacteria, bubbling away as they produce CO_2, SO_2, H_2S, CO (carbon dioxide, sulfur dioxide, hydrogen sulfide, carbon monoxide). Some of these gases are quite toxic.

Help your liver expel its bacterial overload with liver cleanses (page 552) until all the bile is a beautiful bright green. This is evident as dark brown bowel movements. Without the green color of bile added to your intestine, the bowel movement remains light colored, such as tan, yellow or orange! By stopping eating polluted food, killing bacteria and cleansing the liver, digestion becomes normal again.

Of course, there must be enough acid in the stomach and digestive enzymes produced to make good digestion possible. Otherwise, leftover food goes to feed the waiting bacteria.

Persons with a chronic digestion problem may also find they harbor lead, cadmium, or mercury in the intestine! This is very good in a sense. Your body has kept these toxins in the intestine, preventing it from getting into your vital organs. The bad news is that their presence in the intestine could start an intestinal disease. Toxins in the intestine would inhibit your immune system (here the white blood cells are in clumps called *Peyer's patches*), from gobbling up the "bad" enteric bacteria. Maybe the “good” bacteria found your intestine too toxic to live there. Clean up your dentalware. Search for lead and cadmium in your drinking water. After improving your lifestyle, continue to do liver cleanses. Keep your goals high: a flat tummy that feels energetic, no burps or gas.

Stomach ache (page 98) and **hiatal hernia** (page 133) are also digestive problems but are dealt with under pain. (See also lower abdominal pains, page 97, and appendicitis, page 100).

Alan Barth had his two young sons with him. One of them had a very sensitive stomach, a poor appetite, wanting nothing but sweets or chips to eat. The 16 year old had aches and pains that were keeping him out of sports. The father had a tonsil problem; they swelled if he drank milk. One problem was obvious. Their milk was tainted with *Salmonellas* and *Shigellas*, setting up throat problems for the father, stomach problems for one child and a pain syndrome for the other child. Boiling all their milk, not bringing raw chicken into the house (*Salmonella* Source) and stopping eating yogurt and cheese was the solution. There were traces of lead in their tap water and the house air had vanadium in it, announcing a gas leak. They were all heavily parasitized. When these problems were cleaned up, the whole family's health improved.

Kae Nakajima, middle age, had a history of ulcers and stomach problems. She took a lot of Zantac™ and Tagamet™ over the years. She was drinking coffee, tea and colas. She had *Ascaris* and cadmium in her stomach. Five months later she had cleaned up everything except dentalware and was feeling very good. She didn't need any medicines. She still had arthritis and sinus problems but felt so encouraged she had the dental work scheduled.

Sven Lippencott, age 4, had been tube fed for several years due to weak stomach action. He was quite underweight for his age. He had a population of intestinal fluke in his stomach along with arsenic (pesticide). The rest of the family also had intestinal flukes. After killing parasites and washing the upholstered furniture, Sven's appetite went up so he ate more by mouth and grew immediately. But he soon picked up *Ascaris* and cat liver fluke (there was an indoor cat). Then the whole family got intestinal flukes again. They all, including Sven, drank carbonated beverages *ad*

> *libitum* (whenever they wished). With both parents working it would not be possible to discipline the habit. Sven had wood alcohol, methyl butyl ketone, hexanedione, methylene chloride and toluene buildup making his recovery hopeless.

Alzheimer's

What was once a rare disease has become a household word because it now affects so many. Two new pollutants of the brain, inviting an old parasite to a location it would not normally be, is the explanation. The new pollutants are solvents that seek the brain.

Xylene and **toluene** are pollutants of popular beverages, decaffeinated powders and carbonated drinks. At first, the body can detoxify these but with a steady stream of solvent arriving, detoxification slows down and parasites begin to build up in the brain. Common fluke parasites which we eat in undercooked meat and perhaps get from our pets, can now reach the brain and multiply there.

Other toxins are also present, such as aluminum, mercury, freon, thallium, cadmium. Aluminum buildup is seen in all Alzheimer's sufferers (100%). This is undoubtedly part of the true cause. Did it come before or after the parasites?

Whatever the answer, your job is clear. Remove every bit of aluminum from the food and environment. Throw out the pots, the aluminum foil, the cookie sheets, the tea ball. Throw out the kitchen salt, the pickles, the baking powder. Buy things made with baking soda (not baking powder), use a plastic salt shaker, buy salt without added aluminum. Stop using commercial soaps and lotions. Make the soap recipes in this book. Finally, tape over all aluminum handles in the bathroom and elsewhere (e.g. the walker) with masking tape. Then find a chelating doctor to help remove aluminum from the brain. Also use thioctic acid (100 mg; take 2 three times a day).

Kill the four common flukes with the first frequency genrator or zapper you can get your hands on. Prevent reinfection from meats and pets. Stop all commercial beverages, including water. The processing has left xylene and toluene in them. They are not put in intentionally. For the same reason, health food beverages are similarly polluted. Only milk is safely bought from the store. You must still sterilize it, however. Make your own fruit juices. Select beverages from the list of recipes given. Drink water from your cold water tap, filtering it with a small pure carbon filter as in a filter pitcher (see *Sources*).

As much as xylene and toluene are brain-seeking solvents, *Shigella* is a brain-seeking bacterium. The symptoms it causes are not always the same since they depend on the location of infection. Sometimes they cause tremor, sometimes loss of balance, sometimes speech problems. But they are very serious problems. Kill *Shigellas* every day at bedtime with your zapper. Start the Bowel Program (page 546). When improvement is lasting you know you have stopped reinfecting from your own bowel or from polluted dairy products.

Remove dental metal and use thioctic acid as a help to clear tissues of metal. Use vitamin C (3 grams) and B_2 (300 mg) to assist the liver with detoxification. Use B complex (2 a day) to assist the liver generally. Avoid food molds; ergot especially has strong mental effects (see Moldy Food, page 381).

Start a kidney cleanse (page 549) as soon as you can. Follow this with a liver cleanse (page 552). Clean up environment and diet. Your beloved family member or friend with Alzheimer's can regain her or his mental function to a considerable degree. **Most important is stopping the mental deterioration** before it is not reversible.

Lisa Anne Reed, 60ish, was tentatively diagnosed with Alzheimer's 10 years ago. She needed complete care at present but was able to walk (could disappear quickly) and eat. She could occasionally say her name. Her brain had intestinal flukes and their eggs and was toxic with aluminum (cooking pots) and aluminum silicate (salt). One week later she still had the parasites because nobody could skillfully give her the parasite program. She was also toxic with benzene so reinfection was occurring. In addition she was toxic with bromine (from brominated bread?), chlorine (chlorinated water?) lithium, bismuth, vanadium, and tungsten, all known to have strong brain effects. She was also toxic with moth balls and iridium. She harbored *Naegleria*, another brain parasite. In another week there still were no changes due to inability to administer the treatment.

Isabelita Ufford, 77, was in a wheelchair, brought by her two daughters who took turns caring for her. She was on Clanopin™ medicine, did not try to speak and needed total care, including feeding. She had been ill about seven years. She had intestinal flukes and their stages in her brain (the cerebrum) as well as intestine. She also had isopropanol solvent, aluminum, chromate and high levels of arsenic in her body. She was given the parasite herbs plus instruction to get rid of solvents and metals but the plans could not be carried out. The parasites could not be killed without considerably more help than was available. She was a dear, sweet person. The daughters were highly motivated but were overwhelmed with the size of the task.

Beth Hamm, 60ish, arrived led by her ever-vigilant, ever-caring husband. She was started by medical doctors on EDTA chelation to remove aluminum from her brain. My tests showed aluminum, toluene, sheep liver flukes, asbestos and *Shigella* bacteria. The parasites and bacteria were zapped immediately and her husband began the difficult task of excluding non-sterile dairy products from the diet provided. In four days she was able to walk by herself, knowing where she was going. She could finish a short sentence and comply with directions to sit down and get up. Then she had a set back—she had acquired *Salmonellas* in the brain from a bit of dairy food that had slipped by his attention. She was given Lugol's and she improved further. In ten days she was a new person; an interview of twenty minutes length did not reveal Alzheimer symptoms. Will she be able to hold on to her gains? Only if the aluminum and asbestos are removed from her home environment, his vigilance with dairy food keeps up, and she stays on a maintenance parasite program. But her husband appeared intent on her recovery. She is a person again. And there seems to be a delightful companionship.

Ruben Camberos was brought by his wife and a friend for Alzheimer's. The first day he arrived, the intestinal flukes in his brain were found and killed. He was started on EDTA chelation. They were warned about non sterile dairy foods. His *Shigella* was zapped. Four days later he spoke his first meaningful sentence: it was three words long. He could pay attention to the appointment proceedings. Two days after that he was reading a newspaper. This was utterly shocking to his wife. In another three days, he could hold a conversation consisting of very short sentences. Larger ones became hopelessly garbled. He was started on ornithine (4) and valerian capsules (6) at bedtime: this produced a beautiful nights sleep (especially for his caretakers!) and his days were less agitated. There were still setbacks later but his wife was determined to get him well.

Dementias, Memory Loss

Memory loss is progressive with age but not due to aging. There are plenty of nonagenarians and centenarians with clear minds and good memories to prove that age is not the deciding factor in the dementias. Why do some people deteriorate much sooner? Could you prevent personal deterioration of mental abilities? You probably can. You will know it by noticing memory improvement. Telephone numbers that left you with no recall, unless you wrote them down, number by number, now form groups as you hear them, and you can jot them down the way you always did! This is a good sign of memory improvement. Your writing can improve. The jagged, crooked, misaligned words can be smoothly written again! You can remember things that happened earlier in the day and talk about it later, at mealtime. You can finish your thoughts in conversation.

Mental deterioration of the elderly is not as complicated as is generally believed. Although circulation and blood pressure play a role, **the effect of toxins is much greater**. The action of toxins is greater in age than in youth. The same polluted water and food causes disorientation in the elderly when it only gives a young person a stomach ache.

The liver's detoxification capability may be the real issue. Indeed, the liver may age, in accordance with the calendar date. Perhaps the liver is the only truly aging organ. It may even determine your life span. The answer, then, is to stop giving it toxic substances and shortening your life span.

As the liver is less able to detoxify them, common toxins are allowed to roam the body with the circulation, doing harm to all the organs. The brain feels disoriented or dizzy; there is memory loss. At first, the liver can “catch up” its work and finally clear the toxin for excretion. But, eventually, it can't catch up or keep up. The body, notably the brain, is bathed in toxic chemicals that interfere with its functioning. Now, the elderly person must use a cane for stability, must walk very carefully not to fall, must write everything down to remember it, calls people by their wrong names, can't “find” the right words to speak with, can't finish sentences, must write on a calendar to keep the days straight, starts talking to themselves to help think of things, develops tremors and unsteady gait, acquires a passive personality, loses weight, gets stooped, stops reading the newspaper.

All these signs of aging (dementias) can be reversed by simply removing the common toxins with which we are already familiar.

Of primary significance are food molds. These cause brain hemorrhages. Clean up diet, mouth, body, environment, very meticulously.

Of course, an elderly person cannot bring these changes to herself or himself. If you have a loved one with symptoms of aging, and this person is willing to cooperate with you, you can honestly promise them numerous improvements. Spend a good deal of your effort on persuasion since living longer or being healthier may not seem worth giving up a coffee and doughnut breakfast. On the other hand, they might respond to the goal of needing fewer pills, getting into their own apartment again or becoming freed from a walker.

Walter Heffern, 64, had been to various neurologists but could not find any help. He appeared to have the same kind of mental deterioration as his mother, but at a much earlier age. He couldn't understand an ordinary conversation; he constantly spoke about winning money, walked hesitantly and had to be left undisturbed to accomplish anything—even eating and dressing. He needed a lot of care. We found he had *Ascaris* larvae in the brain—in the cerebrum, where you think. He also had *Acanthocephala*, *Dipetalonema* (a chicken roundworm), amoeba (*Entamoeba histolytica*) and *Fischoedrius* in the thinking part of his brain. He had been in the poultry business all his life: his mother probably shared this exposure, as well as other lifestyle habits that gave them solvents and pollutants besides parasites. He had constant ringing in his ears, this could affect hearing an ordinary conversation. He had a water softener that would have supplied a daily dose of aluminum to the brain, too. There always were dogs in the house. Perhaps the marvel is that he was no worse off, a tribute to human strength in general.

Pushing Back Age

This chapter is a tribute to Jimmy, Nazy, Michele, Suzanne, Marlena and all the others who achieved excellence with daily care-taking of Mary Austin, deceased at 97.

It's true that we have to die sometime. But why die before our life span is up? If many people can live to 100 years, then surely this is the human life span, not three score and ten. Some scientists think the true human life span is closer to 140 years! And that we all lead shortened lives. The shortening is due to failure of some organ in us. Other organs are dependent on the failing organ and begin to fail also. When the brain fails, death occurs, sometimes in five minutes.

If we knew which organ is failing, we could come to its assistance and prevent the collapse of the whole body. Often it is easy to see which organ is failing. But whether this is the true beginning of the body's problems we cannot know. Before death there may have been appetite loss. Before the appetite loss there may have been a broken hip. Before the broken hip, dizziness. Before the dizziness a blood pressure or blood sugar problem. Before these, an episode of “flu” or a dental “repair.” Sometimes we know what started it all. But often we don't. Just make a guess and begin somewhere.

Diet

If your aging friend or relative is in a home for the elderly, you may be able to persuade him or her to choose a diet that is wiser than the average diet people eat there. This can help a lot. Just stopping drinking the coffee, decaf, iced tea and carbonated beverages that are served, and switching to the recipes in this book could get them off some of their medicines.

The beverages to encourage are sterilized milk and hot water—delicious with whipping cream, honey and cinnamon. This gets them away from solvents, oxalic acid and caffeine.

Old age is not a time when you “no longer need milk.” Calcium losses increase in old age. Milk has the organic form of calcium, chelated with lactic acid, and it has the cream to promote absorption. For this reason, milk should never be reduced in fat content (not less than 2%). The cream is necessary to improve calcium absorption.

In old age it is downright dangerous to be taking many calcium tablets. The stomach does not have the acid necessary to dissolve them. They pass into the intestine, disturbing its function and acid levels. With tablets, too, one must be careful with dosages, while food is self limiting. No elderly person would be able to drink more than one cup of milk at a time. This contains 250 mg. of calcium.

Milk, however, requires stomach acid to curdle it as the first step in digestion. If there is not sufficient acid, it will pass undigested into the intestine, causing new problems. We must listen to the elderly when they say milk gives them gas or other troubles.

Having the milk warm to hot helps in getting digestion started in the stomach. Milk served hot with cinnamon accomplishes two purposes: it will stimulate acid secretion and the cinnamon is an insulin aid. Milk served hot with honey adds the nutritive value of honey, displacing the need for other unnatural sweets. The meal should always include something sour to curdle the milk. It does not have to be added to the milk; it can simply be included with the meal somewhere.

Lemon juice or vinegar can be put in certain foods but the most reliable way to get it into the diet is to put 1 tablespoon into the water glass along with a teaspoon of honey. This gives the water a “sweet and sour” flavor, enough to make it interesting throughout the meal. The fresh lemon juice or white distilled vinegar and a honey dispenser that is easy to use should always be on the table. Bring these two items to your loved one at the "home" if it cannot be provided regularly and reliably. Pop in at mealtime to check up on it. Powdered vitamin C (¼ tsp.) is another useful acid if the first two are not effective enough.

The lemon and honey habit, alone, can add years (healthier years) to an elderly person. The extra acid taken with lunch and supper (the stomach has its own best supply of acid in the morning, for breakfast) improves overall digestion and helps dissolve the calcium, magnesium, iron, zinc, manganese, and other minerals in the food so they can be absorbed.

The habit of using vinegar and honey in water as a beverage was made famous by Dr. Jarvis in his book *Folk Medicine,* circa 1960. He recommended apple cider vinegar for its extra potassium. In those days, vinegar was made of good apples. Now, all the regular vinegars have mold in them. The toxin, *patulin,* in moldy apples has been carefully studied by scientists. It taints the vinegar as well as apple juice and concentrate made from them. I have not tested patulin to see if it can be detoxified by vitamin C. We must use only white distilled vinegar, even though it lacks potassium, aroma and popularity. Using a variety of honeys can make up for the need to vary the flavor. Get orange blossom, linden blossom, buckwheat, wildflower, and sage honey, besides clover blossom.

But honey is not perfect food. It usually has *ergot* mold, a very serious toxin. To detoxify the ergot, you simply add vitamin C to the honey as soon as it arrives from the supermarket. This gives it plenty of time to react with the ergot before you eat it. Bring your "fixed" honeys to the home.

If your elderly loved one has not tolerated milk in years, start with the vinegar and honey beverage, or lemon and honey, and be patient until that is accepted. Then add only ¼ cup milk to the day's diet, (in the morning, on homemade cereal). Go up very gradually and only when digestion allows it. Of course, the milk must be sterile.

If it is not sterile, the final warming will only increase the bacterial count. You must be sure of its sterility. Boil the milk yourself. Near-boiling is not hot enough. It must be heated until it bubbles up and almost goes over the container for ten seconds. Use a non-metal pot that holds one to two quarts. You may throw away the skin. Then cool and refrigerate. Supply it to the home, too.

Milk that is marketed in paper containers that need no refrigeration has been sterilized; it is safe.

Once the body, even an aged body, finds a nutritious food that does not cause troubles of its own, it asks for more. Your loved one will accept it and drink it without forceful coaxing, if there is no problem with it. As long as your loved one tries to avoid drinking it, your challenge is to find the problem and solve it. It is not a matter of taste or habit. It is a matter of digestibility and lack of toxicity. When your loved one is drinking three cups of milk (or buttermilk or whey) a day and three cups of water, there will be no room (nor request) for the usual coffee and tea and other bad beverages.

We all must die of something. But it needn't be a stroke, or heart failure, or cancer. Choose what seems to be the most pressing problem to work on. Common problems that plague the aged are brain problems, incontinence, bad digestion, diabetes, tremor, weakness, feeling cold, sensitivity to noise, losing the sense of taste and smell, hearing loss, insomnia, kidney and heart failure.

Increase Oxygen

Brain problems include memory loss, communication deficit, dementia (calling things by inappropriate names and saying inappropriate things). The brain is simply not getting enough oxygen and food to work right. It is like having a pocket calculator with rundown batteries: it will give you wrong answers (without telling you they are wrong). Not enough oxygen to the brain is the main cause of memory loss, inability to find the right words, getting words mixed up and not being able to speak in sentences. You can prove this by providing oxygen from a tank; modern equipment is very easy to use and inexpensive. If your loved one responds well to a few hours of oxygen, you have proof of the problem.

How can you increase oxygen in the brain?

1. Open the blood vessels wider.
2. Increase pressure of blood running through the blood vessels.
3. Raise the oxygen level in the air that is breathed. Less carbon dioxide, tobacco smoke and auto exhaust. All these, including a gas leak from the pipes in the house, compete with oxygen.
4. Increase the oxygen delivery system to the brain by raising hemoglobin levels. Cure anemia and low iron levels.
5. Raise oxygen saturation of blood by keeping body acidity down.
6. Correct a slow and irregular heart beat.

Niacin

Open blood vessels wider by giving **niacin**. Give it early in the morning, upon rising, as soon as the feet are set on the floor. Keep it at the bedside, use small capsules or tablets and combine this chore with water drinking. The water should not be cold and should have nothing added to make it a beverage. (Drinking water within minutes of sitting upright may also move the bowels soon.) A 250 mg. time-release niacin tablet (see *Sources*) is a good choice. The elderly have little side effects. Even the niacin-flush, which reddens the face and neck is welcomed since it gives a sensation of warmth.

The flush is intensified by giving hot liquids or acids (even vitamin C) to drink. The flush is reduced by giving cold liquids. The opening of blood vessels by niacin only lasts a few hours. If you see it has a good effect on your loved one, give several a day. Do not use a prescription variety, since they are polluted with heavy metals; use only the brand in *Sources*, or a brand that you have tested pure. You can freely experiment with niacin to find the best dosage and variety; it is not toxic in this amount; but the size of the tablet should not turn it into an unpleasant chore. Reduce it if it seems too large to swallow. Don't cut tablets in half, the rough edges can scratch the throat.

Hawthorn berry is an herb that opens blood vessels, particularly to the heart. Carefully watch the effect on blood pressure when using it.

Food mold, particularly ergot, has the opposite effect of niacin. Brain blood vessels are made narrower, cutting down the oxygen supply. Ergot is a common contaminant of grains: don't provide rye or pumpernickel breads or crackers. Don't provide wine or other alcoholic beverages; they are too contaminated with ergot and aflatoxin. Narrowing the blood vessels in the brain can lead to stroke. If you notice an attack of dementia coming, try a niacin tablet (100 mg, not timed release) immediately.

Fig. 34 Alcoholic beverages contain ergot and aflatoxin. Add vitamin C (1/8 tsp.) upon opening and wait 10 minutes for it to act.

TIA’s, Strokes and Purpura

Sometimes, the elderly person is aware of the onset of a brain attack; they may see stars or peculiar shapes or lose vision temporarily. These are called *temporary ischemic attacks* (TIA's). Immediately give a 100 mg tablet of niacin, 1 gram vitamin C, and a B-complex in this order of importance.

TIA’s are caused by spasming in the brain blood vessels. If this causes them to spring a tiny leak somewhere, a part of the brain will not get its usual oxygen and nourishment. A stroke results.

The spasming was probably caused by sorghum molds. Cooking during the manufacturing of sorghum syrup kills the mold but its toxic byproducts (mycotoxins) are still present. Other syrups may have sorghum added, polluting them. Brown sugar is also polluted with sorghum molds, but fortunately you can detoxify this mold with vitamin C as usual. Mix well ¼ tsp. powdered vitamin C with each new (1 lb.) box of brown sugar.

Purple patches, like bruises, on the hands or arms of an elderly person are called *purpura*, and is also caused by sorghum molds. It weakens blood vessels so they break easily.

When an elderly person has purpura or TIA’s and the mold source isn’t obvious, you must track it down. Test in a saliva sample for all the sweetenings used recently (at least an hour ago). At the very least, keep notes on all the sweeteners being used. Look for a common one at the next attack. Never let the trigger food be eaten again.

Blood Pressure

Increase blood pressure if it is habitually too low. Blood pressures below 110 may prevent strokes but ruin the quality of life. Low blood pressures are often due to toxins from foods. But severe salt deprivation in the diet can also cause this. If the adrenal glands are not conserving salt properly, too much sodium is allowed to leave the body with the urine, resulting in low sodium levels in the blood and a blood pressure drop. Adrenal performance is improved by taking the kidney herb recipe.

The kidney herbs (page 549), at half dose level (½cup a day instead of a whole cup) can be given daily for three weeks and then on alternate days indefinitely. Be very careful to keep the herbal tea sterile by reheating. Elderly people seem to sense the improvement given by kidney herbs. Try other kidney herbs from time to time: **shave grass, cedar berries, juniper berries, butcher's broom, cornsilk**. None interfere with drugs. Remember, they also have a diuretic effect. Be prepared to use extra paper padding in underwear to help catch the extra urine output. The first few nights may be disturbed by extra urination. Interpret this positively.

After the blood pressure comes up to 115 (systolic) mental performance will be greatly improved. Don't try to bring it up to 120 since this raises the stroke potential without giving you much improvement in performance. Use an electronic device to measure blood pressure, one with a finger cuff, not an arm cuff which can itself induce broken blood vessels. Purchase a device that needs no adjustments of any kind and has automatic cuff tension control (see mail order catalogs if your pharmacy does not have one).

Air Pollution

Improve the quality of the air by lowering the pollution level. Check into carbon dioxide, carbon monoxide, fumes from a gas stove pilot light, auto exhaust from an attached garage or nearby highway, arsenic from household pesticide, PVC from new plastic curtains or carpeting, formaldehyde from wearing new clothing before washing it, asbestos from hair dryers, freon from a refrigerator, fiberglass, and chlorine from running tap water. Some of these displace oxygen, some are simply toxic to the body and lungs. Clean up the air according to the general principles of environmental cleanup (see Four Clean-ups, page 409).

For the elderly pay special attention to **chlorine** in the air. Shower water puts a lot of chlorine into the bathroom air which then distributes itself through the rest of the house. Notice whether your elderly person goes into the bathroom in fair shape mentally but comes out confused, unreasonable. Not every day's chlorine exposure will have the same effect. Trust your judgment. Attach a carbon filter to the shower head (see *Sources*). Purchase a variety that has very simple-to-replace cartridges. Figure out how long it should last and write the date for replacement on the outside of it for your own convenience.

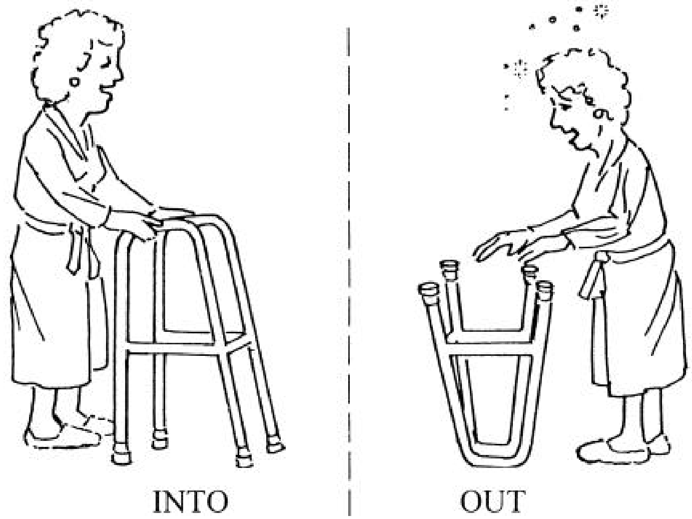

When into and out of the bathroom is a different performance suspect chlorine. Put a filter on your faucets.

Fig. 35 Chlorinated water can cause mental problems.

Washing hands and face in chlorinated water can give off enough chlorine to trigger a manic episode in a manic-depressive person. Certainly, it is enough to cause mental effects in an elderly person. Of course, the chlorine bleach bottle should not be kept under the sink. It should be kept in a closed plastic bag in the garage. It should not be used while the elderly person is in the house and never for his or her laundry. Use chlorine-free bleach.

Don't pollute the air with fresh flowers, potpourris, or room fresheners. These can induce a dizzy spell.

Room air conditioners may have a fiberglass filter! This fills the house with tiny particles of glass to be breathed by everybody. The body makes tumors out of them in order to stop them from cutting through your tissue. Replace the filter with a foam sheet. This sheet is 1/8 inch thick and washable. I have not found these types of foam filters to emit formaldehyde.

Room air filters are not the answer to polluted air. Removing the pollution source is. Air filters may remove some of the toxic elements but by blowing the air (and dust) around vigorously the remaining toxins are made much more vicious in their effect. The noise of a filter motor and fumes it may put out itself adds misery to the simple job of breathing. Using a non-fiberglass filter at the furnace is a better idea.

Make sure all fragrances are removed from the air, even though family members “like” them. They don't belong in air. The lungs treat them like toxins to be coughed up or removed by the kidneys and immune system. This includes colognes, scented tissues, soap and shampoo and shaving supplies. If you can walk into the bathroom blindfolded and know you're in the bathroom, it's not clean enough. Everything in the bathroom pollutes the air of the whole house. People who must use fragrance should apply it outdoors to keep the indoor air less polluted.

Anemia

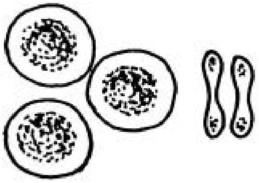

Fig. 36 Red blood cells, top and side view.

Oxygen gets to the brain on a carrier system of tiny rafts called *red blood cells*. They were meant to be an exact shape and size to fit the most oxygen molecules onto them.

Each red blood cell is shaped like a doughnut without the hole. This fits a lot more oxygen, O_2, than round balls would. Yet, if there isn't enough vitamin B_{12}, the dimple isn't put into them to make them doughnut shaped. This reduces the body's oxygen supply and the disorder is called *"pernicious" anemia*. The changed shape of the red blood cells is reflected in a bigger volume called *mean cell volume* (MCV). The correct volume for red blood cells is about 90 cubic microns. Many elderly persons have a MCV over 100!

I have seen pernicious anemia to be associated with *Ascaris* infestation. Kill *Ascaris* on a frequency generator (408 KHz) or zap. The source of *Ascaris* is usually a pet, owned in the past. Once infected, the tiny worms do not leave your body on their own. The infestation may date back to childhood. What a relief for the bone marrow whose job it is to make red blood cells to have enough vitamin B_{12} again! What does *Ascaris* do with your B_{12}? B_{12} is a beautiful rose colored vitamin. Some worms are actually pink from absorbing your B_{12}! Giving B_{12} shots is the current clinical treatment for B_{12} deficiency. Killing *Ascaris* twice a week by zapping and taking B_{12} lozenges (see *Sources*) is a better solution. (If your loved one is getting shots, make sure that no isopropyl alcohol is used on the skin beforehand. Provide vodka yourself in a small pocket flask or 70% grain alcohol for this purpose. Unfortunately, the shot itself may contain traces of this harmful solvent—take a sample home for testing.)

Other kinds of "anemia" can deprive your elderly person of oxygen. A low red blood cell count (under 4.4 million/mm^3) is the "garden" variety. Sometimes iron levels are low and can explain the low red blood cell count. Sometimes they are not.

Most regular anemias, including low iron levels, are associated with hookworm infestations.

Kill all *Ancylostomas* along with *Ascaris* twice a week. It is not wise to take iron pills, even if they do raise hemoglobin levels, except in life-threatening situations. Remove the tiny creatures that cause microscopic bleeding instead. Iron in the form of pills is too easily snatched up by bacteria who also need it, making them more virulent to the body. Use grain alcohol rinse in the bathroom to kill *Ascaris* and hookworm eggs under fingernails. Don't have pets in the house or keep them on daily parasite killing herbs.

It takes nutritious food to build the blood back up to its normal hemoglobin level. Eggs and meats (all very well cooked) are the richest sources of iron and other minerals used in blood building. B_6 and other vitamins are also involved and can be given as a B-complex (see *Sources*). Do not use black strap molasses as an iron source, or any molasses, since it contains toxic molds. (The molds could be detoxified with vitamin C the same way as honey. However, I have not tested enough molasses for solvents and you cannot risk these.) In the past, when the nutrient properties of molasses were discovered, the molasses manufacturing was a small, carefully conducted business. Now it has molds which cause platelet destruction, (purpuric spots) internal bleeding, and immune failure.

Acid Levels

Oxygen must first jump onto its raft, the hemoglobin, in the lungs. Later, in the brain, oxygen must jump off again to enter the brain cells. A difference in acid levels makes this possible. Acid levels operate the latching system that decides whether oxygen will be attached to hemoglobin or let go! Acidity unlatches oxygen. There should be no acidity in the lungs so oxygen can attach here. Sometimes, the entire body is too acid! Diabetics, asthmatics, arthritics, especially, suffer from total body acidity.

Acid was meant to be removed from the blood and loaded into the stomach at mealtime for digestion. When this isn't happening, it was meant to be shipped out of the body with the urine. But the kidneys may be doing a poor job because they are clogged with tiny crystals and because not enough water is drunk, so the body's acid levels rise. You can test total body acidity by measuring the pH of the morning urine. It should not be under (more acid than) 5.5. If the body acid level is too high, help the kidneys excrete it by adding more water to the diet and more minerals to neutralize the acid. The main minerals for this purpose are calcium and magnesium.

Increase Minerals

Adding water to the diet could be the most difficult of tasks if your elderly loved one “doesn't like it.” Calcium should be in the form of milk, magnesium as a tablet. When tablets cannot be swallowed; use magnesium oxide powder (see *Sources*). Use 1/8 tsp. added to cooked cereal, soup, stew, pudding. Magnesium, being a mineral, does not get destroyed as vitamins may. You can add it anywhere in the diet where it won't be tasted. Notice how calming it is to have extra magnesium in this gradual way. And how much better the sleep is at night.

When water "doesn't taste good," there is probably a valid reason. The body may be trying to reject chlorine or other toxins in it. In this case, filter it with a small all-carbon unit that is changed right on schedule. A plastic pitcher (not clear plastic or flexible plastic) with a carbon pack fitted into the top is best. Sterilize it once a week by putting a cup of water and one tbs. of grain alcohol in it and turning it upside down so the filter can soak for 15 minutes. Flush out the alcohol with two pitchers of water. Make sure the temperature suits the person. Temperature can mean everything to the never-thirsty person. Don't allow ice cubes, however, nor beverage making, with the essential water. Adding lemon or vinegar (white distilled) and 1 tsp. honey is probably the best way to stimulate both thirst and appetite.

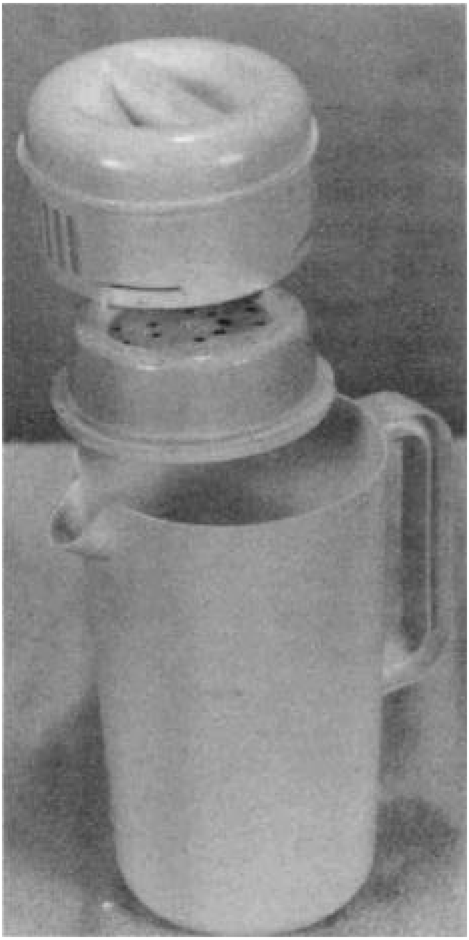

Fig. 37 Water pitcher with filter.

When blood is properly oxygenated it takes on a bright red color, unoxygenated blood is more purple. A chelation doctor can easily see the state of oxygenation.

Chelation

Chelation is a powerful way to quickly improve oxygenation of blood. The most important rule to observe, though, is to take the treatment slowly. Especially if *ethylene diamine tetra acetate* (EDTA) is being used to remove heavy metals, it is important to take the treatment over a two hour time period. Mini doses may be given in a shorter time. Generally, you are in charge of the flow rate. Discuss it with the nurse. Weekly chelations can correct many problems of the elderly that no other treatment could.

Because of hostility from insurance companies who do not wish to add another cost to their ledger and doctors indoctrinated with misinformation, bad publicity is given to this wonderful, life-prolonging mode of treatment. Clinical doctors who have no time to really investigate the statistics of chelation treatments and for whom this is purely competition may feel antagonistic to these treatments. Your loved one should not be the one who must suffer from medical politics. Go to see for yourself what chelation is all about. The receptionist should be glad to show you around. The secret is to talk to the patients themselves. They are usually sitting around a room, eating their lunch and reading as their IV's drip. The pulse, blood pressure and blood chemistry is also carefully monitored. Sit down with them to find out their stories. Get a realistic picture of benefits and costs.

Pulse

The pulse reflects the heartbeat. A **slow pulse** can give weird brain symptoms besides great fatigue. The cause is usually a drug that is being taken to correct a **fast pulse**! Check with the nurse. Read the insert included with packaging for all drugs used. The drugs responsible are usually “beta blockers”, used for the purpose of smoothing out the heart beat, that is, making it regular. Often the drug can be changed.

Less than 60 beats per minute will lead to trouble. For a young person it is a good sign to be as low as 60, provided no drug is involved. But for the elderly it does not reflect a strong athletic heart beat.

The heart is made of four separate “chambers” or compartments each pulsing in turn. They are like four horses pulling a wagon. Unless they pull evenly, the wagon feels jerky, and irregular. The wagon will wear out sooner with jerky pulling. To smooth them out you simply slow them down. Apparently they sense each other better and can pull evenly now.

A heart that is beating 100 times per minute, not unusual for a weak old heart, can be so irregular that it misses every fourth beat. That creates a terrible deficiency of oxygen. Imagine your four cylinder car or lawnmower missing one out of four engine strokes! Beta-blockers have some quite undesirable side effects but heart regularity has a higher priority. So drugs are the immediate choice. Later, when heart health is improved, the heart will beat regularly without drug use. In the meantime, watch over the pulse. When the pulse drops below 60 the new danger is slowness. Take the pulse daily when a new drug has been added, or when you are working on heart health, without getting your loved one anxious about it.

Heart Health

To improve heart health, the first steps of course would be to go off caffeine and to kill parasites and bacteria. This alone could drop the pulse from 120 to 80 in a few days. Obviously, the need for a drug is gone. Cut the drug dosage in half immediately. Don't wait for a doctor's appointment to O.K. it. If you waited another day the pulse could be below 60.

Many common bacteria, especially *Staphylococcus aureus*, choose the heart as their favorite location. Their nesting place, though, will be under a missing tooth in the jaw (cavitation). Heartworm and *Loa loa* are two very common heart parasites. You can have all these killed in a day, without side effects and your heart is once more free to beat regularly. Don't take a chance on over-medicating. As soon as the beat is regular and under 100 per minute, reduce your heart drugs. Stop them when you are regular and under 80. Fatigue will leave and the brain will work better.

Raising potassium levels slows the pulse. Try to do this with diet by eating more potassium rich food and by conserving on potassium losses. The adrenals are in control of losses. Give them a thorough cleaning. The adrenals are situated right on top of the kidneys where all toxic things are being excreted. Being this near to the urinary tract will result in shared toxins. Urinary tract bacteria, small kidney stones, moldy foods and metal from dentalware are the chief offenders. Switch to composite tooth fillings. Use non-metal jewelry. Cook and eat with non-metal ware. Don't handle metal unconsciously throughout the day. Aluminum objects that must be touched should be wrapped in masking tape: this includes walker, shower door, bathroom supports.

Door knobs, taped walker handles, and cane handles should be wiped daily with a grain alcohol solution.

Treat the adrenal glands to 5 supplements

1. **Vitamin C:** shake some into all foods that can absorb a bit of the sour taste, even cooked cereal and vinegar water.
2. **Pantothenic acid:** 500 mg., one daily.
3. **B_6:** 250-500 mg. daily.

4. **Folic acid:** 800 mcg. daily.
5. **B_2:** 150 mg. daily. These are minimum dosages.

If no capsules or tablets can be swallowed put a three day supply in a heavy plastic bag. Pound with a hammer; then roll finer with a big glass jar. Use about 1/3 of the mix each day. If you are trying to do all this in a nursing home, feed it to your loved one while visiting. Put the powder mix in a plastic (not styrofoam) cup, add honey and stir until you get a paste. Feed the paste directly by spoon. Often the elderly prefer it this way in order not to bother with pill taking at meal time. Remind them it will turn the urine yellow.

When the brain problems are corrected for an elderly person, be sure to relate the improvement to him or her. Improvements should not be dismissed lightly. Keep notes. This encourages the elderly, letting them know their existence and quality of life is important to you. Improvements are not merely steps to yet further improvements. Enjoy each bit of progress; it is often too subtle for your loved one to notice even when it is glaringly obvious to you. Before and after a chelation treatment can show a dramatic change in mood, energy, appetite and communication ability, yet get no comment from your loved one. Point it out explicitly so your loved one can look forward to it, too.

Aging is no fun. The elderly long for a brisk gait, laughter, and a picnic at a park again. They dwell on dying, though, because they know it isn't far away and wonder what to do about it. They dare not talk about it because it is too painful a subject for the loved ones. And the immediate problems are too pressing to allow much contemplation of future problems. Talk about aging in its positive aspects. Let your loved one express feelings about it.

Incontinence

plagues most elderly persons. It begins to plague women much earlier—after childbirth, for instance. Surgically shortening the bands that hold the bladder in position (called bladder "lifting") can give temporary relief, but the surgeon may be the first to tell you that it is a temporary fix. Still, it is so shocking not to be able to run a few steps or sneeze or cough without wetting the underwear, that anything seems better than doing nothing. Surgeons will tell you that the bands have been "overstretched."

The real reason why nothing, not even surgery, is permanent is that the support bands are weak. Bacterial invasion causes most of this weakness. Low potassium levels (due to excess potassium losses by the adrenals) causes more weakness. When you kill bacteria (and *Schistosomes* and *Ascaris* and other parasites that bring in bacteria) and blood potassium levels go up, the problem is solved. Overnight you may throw those pads away. Even though you needed three pads to be "safe" you will not need any. Whether you have killed bacteria permanently determines whether you have permanently cured the condition. Make sure all dairy foods are absolutely sterile. Ask that the milk be boiled for ten seconds and other foods that can't be sterilized are not on the menu, like sour cream. Sour cream has too much *tyramine* to be safe. Tyramine is a bacterial by product that is quite toxic; it is rather high in aged cheese, also. With the food bacteria, *Salmonellas* and *Shigella*, out of the way and parasites being killed regularly, you can focus attention on the adrenals which control potassium levels.

Be careful not to rave about the foods that your loved one cannot eat.

Eating more potassium in food is a good nutrition project. Bananas are the top choice. Fresh fruit salad and baked potatoes and soup also provides a lot. Mixing potassium salt with regular salt, half and half, for the shaker is another easy trick, even if you only use it in cooking where the taste cannot be detected. Potassium by prescription is often used by clinicians to conserve body potassium during diuretic use. This need not be stopped (if the pills are not polluted) although taking potassium pills is less useful than salting it in because the adrenals will let any big dose escape anyway. A sign of too much potassium is a slow pulse.

It may be necessary to wear some kind of incontinence underwear. Try to avoid them at night, though, so the skin can breathe freely. Bring a commode near the bed for the night, rather than diapering your loved one (but don't call them diapers; say “underwear”). Absorbent pants of all kinds are heavily chemicalized. This is absorbed by the skin and adds to the toxin level. Less will be absorbed if you powder the skin with cornstarch first. Use them minimally and line them with tissue or paper towel. Chair and bed pads, too, are chemicalized. Don't sit on them with bare skin. To facilitate getting to the commode quickly in the night, dress the elderly in a short night shirt, no pajamas or long gown. Bed socks on the feet help with warmth.

Wash the body parts daily, around the urinary and rectal outlet, using borax water. Follow with 5% grain alcohol. Put washcloth in laundry after a single use. Nothing, not even brain improvement, impresses and encourages an elderly person as much as seeing the incontinence lessen. This bit of progress will put him or her solidly on your side. When they believe in you, it makes your task more rewarding. Remember to enjoy and celebrate your achievements together; don't make a grim business out of it.

Bad Digestion

with its excessive gas and burping is another plague of the elderly. They would rather not go to church nor visit a friend than embarrass themselves in that way.

Chewing

It all begins with the stomach although chewing food well is essential for really good digestion. Dentures should fit perfectly so the mouth does not develop sores. Using denture cream is not a good substitute for correct fit and is toxic. Denture plastic is often toxic, even containing mercury in its composition! Toxins in plastic can seep! Such toxins lower the immunity of the mouth and throat and stomach since it all flows down into the stomach. Low immunity in the mouth permits throat infections to be chronic. If your elderly loved one has a red-looking mouth or throat, instead of pink, an infection is going on in spite of no coughs and no complaints.

It will do no good to keep zapping bacteria when reinfection is so easy. First kill the bacteria in the dentures by soaking in 70% grain alcohol. Then test the dentures for toxins. Soak the dentures in water for several hours. Rinse and soak again in fresh water. Repeat a third time to insure that any toxin found came from the dentures, not the saliva. Save this water for testing. Search for heavy metals in the denture water. If you find any, you know the dentures are toxic! Get new ones, made of uncolored methacrylate (see Dental Cleanup, page 409).

The denture-soak should kill bacteria each night. Plastic has tiny pores where bacteria can hide. Use 70% grain alcohol which you make yourself or plain vodka which is about 50% alcohol. Since alcohol evaporates and is expensive, use a wide mouth jar with close fitting non-metal lid for all this. Fish them out with your toothbrush so it gets sterilized too. It only takes minutes to kill everything. Commercial denture cleaners are much more toxic than grain alcohol; don't use them.

Use food grade hydrogen peroxide or salt water to brush teeth in your mouth, never **toothpaste**. Toothpaste has toxic metals (tin, fluoride, strontium) besides benzene pollution. See the section on brushing teeth (page 532) for details and sources. If you are responsible for this daily chore, use homemade floss (2 pound to 4 pound nylon fish line) first; then brush. If your loved one is seated they may be able to handle the brush by themselves, giving them pride in the achievement.

If an elderly person refuses or can't wear dentures, provide food that is soft and without chunks since this decides whether the stomach can digest it. The stomach is the weak point of the digestive process for the elderly because nearly all don't produce enough acid to get the job done.

Stomach Acid

The body produces hydrochloric acid (HCL) which gets pushed into the stomach from the blood! The enzyme, *carbonic anhydrase*, a zinc enzyme, is involved. Not many ways are known to stimulate this whole process. Drinking water before meals stimulates it in unknown ways but is hard to do for the elderly. Next best is to provide acid.

Because strong HCL would dissolve teeth it is not available as a solution to aid digestion. Ask a pharmacist to make a 1% HCL solution and use 10 drops of it in a beverage at mealtime once a day. HCL as a tablet ("Betaine HCL") is available but doesn't have enough HCL in it.

Using a lemon or vinegar and honey beverage helps with digestion although this provides citric or acetic acid, not hydrochloric. These acids are completely metabolized so they don't add to the body acid level. But the fact that it is not hydrochloric means that it can't kill bacteria and parasites in the stomach like regular hydrochloric acid could. The stomach becomes a haven for *Salmonellas* and other bacteria and this is **the biggest digestive plague of the elderly**. *Salmonellas* dig deep into the stomach wall, safe from antibiotics and stomach acid and aren't washed away with the food. When they take over the region near the top of the stomach, it weakens the esophageal sphincter and food keeps coming back up a bit—a most uncomfortable development, especially after supper or when lying down.

When the Salmonellas spread out further to invade the diaphragm around the sphincter, the diaphragm weakens, and lets a bit of the stomach up through the hole.

This causes *hiatal hernia* distress. Don't settle your loved one in an easy chair after supper. This presses the stomach upward and the food up, too. Leave them sitting at the table a while, then walk a bit, to get the food down lower. The food will sink lower if some of it can leave the stomach at the lower pyloric end. But if *Salmonellas* are entrenched here, too, the lower end does not have enough action to push the food through the valve. Drugs like Reglan™ are given to speed this up.

What helps most is getting digestion completed. This sets up the natural cues for emptying. Digestive enzyme tablets have been in popular use to help digestion. But they may not be safe since they have not been sterilized. Always try the vinegar and honey method first. Coughing during eating is a sign that the diaphragm is irritated (by a hiatal hernia). If drinking water starts the coughing, omit it at the beginning of meals. Work in sips during the meal.

Salmonella and Shigella

Some *Salmonella* infections can bring dizziness to your elderly person. Dizziness is another plague of the elderly, keeping them from going shopping, getting to church and even from getting around their own homes. Drugs such as Antivert™ are given but only take the edge off the problem. Feeling dizzy can make your loved one home bound and stuck to a walker for every move.

Salmonellas, along with *Shigellas*, produce very toxic substances that cause dizziness. There are three common *Salmonella* varieties: *Salmonella enteriditis*, *Salmonella paratyphi*, and *Salmonella typhimurium* (386, 380, 354 KHz). Kill *Salmonellas* daily for a month by taking Lugol’s iodine (6 drops in a half cup water, after meals and bedtime, see *Recipes*). Unfortunately, this will not kill *Shigellas*; follow the Bowel Program (page 546) to get them.

During this time set up a system of sterilizing all dairy products (see Milk, page 425) since this is the source of reinfection. Set up a system of rinsing fingers (and fingernails) in 10% grain alcohol in the bathroom. Deli food and restaurant salads carry *Salmonellas* and *Shigellas*, too. Kill them, routinely, after eating such food due to necessity. A warm stomach full of food at a neutral pH is just the right culture condition for these bacteria. It's like putting yeast into a bowl of warm water, flour and sugar. In half an hour it is overflowing with growth.

Once *Salmonella* is entrenched in an organ it is difficult to eliminate. Only an electric zapper can kill them all (in an organ, not the bowel). If your body has the right conditions (like a low acid stomach) to let them grow you dare not swallow another one! *Shigellas* arrive with dairy foods, too, but prefer the lower intestine as their headquarters. Indigestion that starts right after eating is probably due to *Salmonellas*. If your indigestion comes in the night, this suggests *Shigellas*, since they've had time to reach their favorite place further down.

Campylobacter and *E. coli*, other digestive bacteria, are sometimes the culprits. The Bowel Program is effective against these also. Besides getting digestive improvement you get mental improvement, less depression, less dizziness, less irritability after clearing these up. Remember that eating bacteria and killing them later will not solve the problem. Stopping eating them will.

Other Clues

Digestion problems that remain after eliminating bacteria can be diagnosed in a rational way. Ask these questions:

- Is the stool orangish-yellow, or very pale, instead of greenish brown? If so, bile isn't getting delivered to the small intestine from the liver.
- Is there abdominal pain? (More about this on page 97). It may be due to *Ascaris*, flukes, or other parasites.
- Is there constipation? This will let wastes accumulate, all the longer for bacteria to thrive on them.
- Is there bloating? This is due to gas made by bacteria.
- Does the stool float? If so, it must be lighter than water and contain fat or a great deal of undigested material.

Liver Bile

Bile is necessary for digestion. Absorption of fat and calcium depends on bile mixing with the food. When fat isn't absorbed, it stays in the intestine. Fat is lighter than water; it makes the stool float. Feces should not float. When the stool floats you can assume that calcium isn't being absorbed either, leaving the blood in a deficit which will be taken from the bones.

If the stool floats or is orangish in color prepare your elderly person for a liver cleanse (page 552) to clear a bile duct of obstruction. They get quite fond of these cleanses and will ask to have one. Liver cleanses are completely safe, even for persons in their 80's. One of the stones pictured on page 554 came from a woman age 97. The general rules apply to the extremely elderly: kill all parasites first by zapper if possible, otherwise by herbal parasite killing. Do a kidney cleanse (page 549) first, using half a dose instead of the regular dose, for three to six weeks. Attend your loved one in person for the liver cleanse, have a commode at bedside, protect bedding from accident: use paper underwear if necessary. Share the joy of getting gallstones out painlessly with your loved one; let them see and count them if they wish before you flush them (use a flashlight).

Be extra careful with the skin cleansing. Hot water soothes and heals. Use starch skin soother to dispense onto the wet paper towel, besides borax solution and alcohol. Don't use ordinary soap. The starch skin softener gives the smoothness of soap, and prevents the pain of friction. An elderly person may have no diarrhea at all with the Epsom salts! Evidently the body absorbs all the magnesium so eagerly, none is left in the intestine to absorb water and create diarrhea. It is especially important though to rehydrate your elderly person after a diarrhea. This time they do not balk at water consumption. The liver cleanse, it seems, gives them new thirst as well as new appetite. But it doesn't last long. As the stones from the far corners of the liver move forward, they compact into larger stones and plug the ducts again. Their previous symptoms return. Try to give a cleanse once a month until the dark color of the stool returns and it no longer floats.

The benefits of a liver cleanse will last longer if **valerian herb** is taken the day after the cleanse and from then forward. It may be preventing spasm of the bile ducts. Use 2 oz. of the herb (cut) in 3 cups water. Simmer for 5-10 minutes, let settle or strain. Add honey to sweeten. Give a few tbs. every 4 hours (or 6 capsules) for several days followed by a daily dose at bedtime.

Constipation

If constipation is a problem, use an herbal product rather than a drug until you have removed the cause. **Cascara sagrada** (half dosage) or **prunes** work for many people. Adding **roughage** to the diet is a good solution but often doesn't work. If you try bran, you should add vitamin C and boil it, first, because it is very moldy. But even eating tree branches for supper won't move a bowel that has the wrong bacteria in it.

Bacteria are part of the cause; and part of the result! Constipation increases the bacteria level which causes further constipation! You may solve the constipation problem immediately by zapping. Even though this kills some “good” with some “bad” bacteria, no harm is done. The stool is recolonized in one to two days.

Poop Your Troubles Away

Two bowel movements a day are the minimum necessary for good health. The first one should be in the morning. The morning cup of water, drunk at the bedside has the magical ability to move the bowels. Cold water may fail. But the water effect only works in the early morning. Waiting until after breakfast may not work. Notice the energy lift your loved one gets from this most primitive body cleanse. Take advantage of this to exercise them. Go immediately for the morning walk. This might be the only time of day they can enjoy their walk.

Walking and **liver cleansing** are the most health-promoting activities you can do for your loved one.

Make walking as essential as eating. Walking is not merely walking about the house or shopping. Walking should be done outdoors. Walking is a brisk exercise, done as speedily as possible and lasting at least ½hour. Only if the weather doesn't allow outdoor walking can an indoor walk be substituted. Don't let your elderly person choose whether they will walk that day. To overcome resistance, find a cheerful neighborhood person willing to do this task for pay. The need to respond to a new stranger energizes the elderly more than your persuasion can.

Sugar Regulation

Diabetes is a common development in the elderly. If your loved one is already on a pill for beginning diabetes, take this as your challenge never to let it get worse. It is not a chronic metabolic deterioration. It is a destruction of the pancreas (specifically the islets) by the pancreatic fluke which is attracted to the pancreas by wood alcohol. Zap flukes and eliminate wood alcohol as described in the section on diabetes (page 173).

Use no artificial sweetener and no beverages besides milk, water and the recipes given in this book. This is one regimen your loved one will not resent. They are well motivated to prevent the need for giving themselves daily shots of insulin. Use this motivation to acquire the taste for new foods and beverages.

Elderly Person’s 7 Day Diabetic Diet

Daily supplements: 4 chromium (200 mcg. each), 1 or 2 B-50 complex, vitamin C (1,000 mg. or ¼ tsp.). Take with food.

Breakfast

Choose any one; they need not be in order.

1. Two eggs, (replace carton and wash hands and eggs before cracking), wheat-free, corn-free bread, not toasted (special breads can be found in the freezer section of natural food

stores), with butter and 1 tbs. orange blossom honey. 1 peeled pear, raw, with whipping cream. One cup hot milk with cinnamon.

2. Old fashioned oats, with 1/4 tsp. cinnamon and 1/8 tsp. vitamin C, stirred in just before serving. Mix whipping cream with sterilized milk to make a "half n half" for the cereal. 1 tbs. honey. 1 banana. One half cup milk, one half glass water with honey and vinegar.
3. Fried potatoes with 2 eggs (use only butter, olive oil or lard), 1 cup hot or cold milk. A quartered orange (wash the orange before quartering).
4. Cream of rice, with homemade "half n half" or whipping cream, cinnamon and vitamin C stirred in. 1 cup milk, 1 banana.
5. Cottage cheese, cooked in covered skillet to sterilize. Add chives or peeled fruit (not canned). Wheat-free, corn-free bread, or rice bread with 1 tbs. honey. 1 nectarine or piece of melon. 1 cup hot milk. Water with vinegar and honey.
6. Pancakes or waffles with butter and eggs (no sweetening). Fruit (peeled), hot milk, water.
7. Fruit cup, large bowl of peeled, chopped mixed fruit with whipping cream and 1 tbs. honey and rice bread or other wheat-free, corn-free bread with unsalted butter. 1 cup hot milk with cinnamon.

Remember, all honey must be pretreated with vitamin C. All fruit is peeled and free of blemishes and soft spots. All milk, cream, butter must be sterilized for 10 seconds at full boiling point. Butter must not be "raw". Get wheat-free, corn-free bread at a natural foods store.

Lunch

It is better to have most of the day's calories in the middle of the day than at the end. Arrange for dinner at noon if possible.

Dinner

Choose any one.

1. Green beans with potatoes, meat dish, cabbage apple salad, water with lemon juice and honey, 1 cup hot milk. Water.

   Fresh green beans, especially fava beans contain a substance that is described in old herbal literature to be especially beneficial to diabetics. Don't overcook them—it might harm this substance.

   For the same reason, don't use canned green beans. If “fresh” isn't possible, choose “frozen” but rinse the chemicals off before cooking. Potatoes (not overcooked), peeled to make sure there are no blemishes (contain mold and pesticide) can be cooked with the beans. Cook with onions and oregano for flavoring. Add fresh chopped parsley to the sauce or butter for both green beans and potatoes. Fresh parsley has special herbal goodness (high magnesium, high potassium, diuretic.)

   The meat dish should be overdone. “Fast food” is plopped from the freezer into the boiling grease which browns the outside nicely but can easily leave the inside undercooked. Meat must never be “rare.” There should be no redness near bones! Canned meat is safe from parasites but may have smoke flavoring added (contains benzopyrene) or nitrates. Avoid these chemicals. Avoid MSG too. Whenever a meat dish is not accepted, substitute sardines. Let them choose from a display of six kinds. Purchase the flip-top cans to avoid eating metal grindings from the can opening process.

   Cabbage for salad should be chopped fine enough to be digestible. Add finely chopped apples (peeled) and a few apple seeds and whipping cream for the dressing.

Sweet things are reserved for dessert. Since a diabetic's tissues are not absorbing sugar, they crave it more and more. As the diabetes improves they crave it less. For dessert, serve 1 tbs. of honey to satisfy this craving without endangering their blood sugar regulation. It can be used in the hot milk or in other ways. Undercooking the vegetables also helps slow down the sugar release. Never serve mashed potatoes for this reason.

The drinking water should always have a little vitamin C, lemon juice or vinegar added, and 1 tsp. honey if desired.

2. Asparagus, potato, raw salad, fowl dish, fruit, water with vinegar and honey, 1 cup hot milk.

   The asparagus can be fresh or canned. Bake the potato: not in aluminum foil, not baked until fluffy. Don't let the skin be eaten. Use genuine butter, only, or a homemade sour cream dressing (see *Recipes*). Fresh chopped chives may be added but no regular sour cream since this is very high in *tyramine*, a brain toxin.

   The raw salad should be chopped small enough to be edible by dentures. Use homemade salad dressing with a preference for oil and vinegar styles.

   The fowl dish should be very well done, never “fast food”.

   For dessert, fresh fruit chunks dipped in a homemade honey sauce (honey, water and cinnamon). Less sweets are consumed if you dip the fruit rather than pour the sauce over. Limit the total to 1 tbs. honey. Don't serve grapes or strawberries due to the intense mold problem.

3. Soup, sandwich, fruit, hot milk, water.

   Soup should be homemade from scratch. Add bones and 1 tbs. vinegar (white distilled) or a tomato to the kettle to ensure some calcium leaches out of the bones. A fish chowder serves this purpose very well, too.

The sandwich has lettuce, real butter, and whatever else tastes good (no cheese, bacon bits or condiments). The bread is wheat-free, corn-free, stored in freezer. Home-made salad dressing can be added.

The fruit may be chopped with whipping cream, cinnamon and honey sauce (not more than 1 tbs. honey).

The water may be plain if there was vinegar in the soup.

4. Fish, green beans, potatoes, other greens, fruit, hot milk, water.

   Fried or baked fish is served with lemon or lime. Green beans are served with a cheese sauce so a lot will be eaten. (Cheese sauce: add milk, olive oil to a block of cheese. Melt and cook at least 10 seconds.) Serve au gratin potatoes or scalloped potatoes or any kind of potatoes that will be enjoyed. The extra greens can be beet greens, collards, mustard greens or spinach served with a favorite dressing to make sure it's eaten. (No croutons or bacon bits, though.)

   Never serve dessert if the plate has not been cleared. Your loved one isn't hungry enough. If appetite is very poor, sweets will only worsen the problem. Try to change the menu to stimulate the appetite. Acid foods stimulate; spices and B-vitamins (especially B_1) stimulate; hot foods stimulate. Much appetite is controlled by the liver and brain. Toxins at either location (especially food-derived toxins) tell the body to stop eating. Suspect food molds first, bacteria and chemical additives next.

5. Asparagus, meat dish, white rice (brown rice contains mold), coleslaw, milk, water, ice cream.

   A hot meat dish (no pasta, no wheat flour, no regular gravy) can be fried, cooked or baked, but not grilled. Asparagus is fresh, frozen or canned. Rinse if frozen. Fix it differently than last time. Season rice with parsley and minimal salt and pure herbs like thyme; no MSG or mixed seasoning, make butter sauce. Dessert is homemade ice cream (see *Recipes*).

6. Fish or seafood hot dish. Green peas or peas and onions. Peeled sweet potato with butter (not canned). May switch sweet potato with rice on asparagus day. Sliced tomatoes or cucumber or other raw vegetables with or without dressing. Milk, 1 tbs. honey (can be used on sweet potato).
7. Chili or stew with unlimited rice-bread and butter. If chili produces gas, stay away from it. Serve no canned varieties. Grated carrot salad with stewed raisins added and heavy cream. Milk, water as usual. Blueberry pie, sweet potato pie, custard pie.

If more bread is requested, provide a wheat-free, corn-free variety; but limit bread eating to “after main dish” eating. If not enough milk is drunk: make custard pudding or rice pudding so the daily amount (3 cups) is consumed.

Supper

1. Tuna with salad dressing or tuna salad (no pasta). Non-wheat raisin bread (from natural foods store) and butter. Milk, water.
2. Custard, cooked greens, baked potato. Rice-bread and butter. Herb tea with milk added (single herb only, not mixed herbs).
3. Vegetable soup, homemade, from scratch. Leftover meat sandwich with rice-bread (no deli or cold cuts or luncheon meat). Milk, water.
4. Baked squash with butter, rice pudding with cinnamon, raisins, and honey. Canned salmon or sardines with wheat free, corn free bread. Milk, water.
5. Chili or stew leftovers with wheat free, corn free bread. Custard pudding sweetened with honey, and nutmeg/cinnamon mix. Milk, water.

6. Sardines and rice bread or other wheat free, corn free bread. Homemade tomato juice with celery, strained. Milk, plain water. Pie a la mode (homemade pumpkin or squash pie and homemade ice cream).
7. Potato salad. Leftover meat dish and beans, stewed tomatoes, squash. Baked apple with cinnamon, cream and honey. Milk, water.

Many diabetics lose 50 points (mg/DL) of blood sugar in a few days on this diet. This is why: there is less bread than a diabetic would prefer. There is very little cheese (it must always be boiled in a sauce to sterilize). There is no fruit or vegetable juice except homemade, and not much of that because it crowds out milk and water. If your elderly loved one can't eat all this, make sure there are no snacks consumed between meals that are forbidden.

There is no pasta anywhere. Pasta is mold-ridden. Even after cooking, it may be toxic. There is no wheat or corn bread. The menu is heavy on green beans and asparagus and cinnamon. If by chance, your elderly person hates these and starves themselves to get your sympathy, add a lot more potatoes and rice (never brown) to raise calories.

There is no sweetening other than honey (5 kinds). There are no syrups or sugars. Honey is self limiting—the taste for it is all gone after 1 tbs. Not so for other sugars. The heavy use of cream and butter is offset by no deep fat fried food and little cheese.

Keep in mind that this diet may reduce the need for insulin almost immediately. You may have to cut it in half! But how would you know this? The morning blood sugar test is essential to keep track of changing circumstances. Don't neglect it. Be careful not to use rubbing alcohol when making the finger stick (use vodka or grain alcohol). Your elderly person will feel it is all worthwhile (doing without coffee or pie) if one less insulin shot is needed or if they can go back on tablets instead of insulin. Or even just the knowledge they are staying well controlled and will never have to take insulin shots. Remember to give yourself the credit for a fine accomplishment when your loved one's diabetes improves.

Diabetic Supplements

Several supplements are especially good for diabetics:

- Fenugreek seeds, 3 capsules with each meal.
- Fresh vegetable juice made of raw green beans and carrots (½cup total).
- Bilberry leaves. Maybe they have something in them that helps detoxify wood alcohol, since bilberry leaves are good for eyes, too. Get in capsules or make a tea.

Diabetic Eating Out

Since the rules are always somewhat relaxed when “eating out” a diabetic loved one will badger you to go out with them. If rules are sure to be broken, calculate it into the rest of the day so you can compensate for it.

Extra Diet Tips For The Elderly

Food should taste good.

Eating is a fundamental pleasure of living. In old age there is no other pleasure that can equal enjoyment of food. It is a time when we long for the foods of our own childhood, too. Ethnic foods often had to be given up when children were raised (switched to hot dogs and pizza) but with this diversion gone, a return to family food would be most welcome and most healthful. Ethnic foods were made from scratch. And they certainly were made at home where cleanliness and “persnickitiness” are at their finest! A speck in the batter gets noticed. Not so in a commercial mixing vat. Pots and pans are sanitized with hot water, not chemicals.

Good advice is to return to old fashioned home cooking: with its flour and butter, lard and cream, homemade pasta, olive oil and soup, coarse cereal grains and plain fruit. Gone are the fruit juices, flour mixes, crackers and sweets that fill grocery shelves. What about convenience? Old fashioned cooking took most of the day. It does take 3 or 4 hours to make a soup from scratch. But you then get 3 days off! Each day you reboil it, it is sterile again. Or freeze half of it (take the potatoes out first). It does take a whole morning to make pasta or some ethnic dish. Freeze it in plastic sealed containers so the delicate flavor isn't spoiled. Baking homemade bread is automated now. Do at least this much to get away from the mold-ridden grocery store loaves. Make your own ice cream and nut butters. They last many days and free your schedule.

Time is the great inhibitor but if you have the means or the help, the best advice, nutritionally, is a return to old-fashioned cooking and recipes. Use your new insights to improve them where you can. Don't use your mother's aluminum ware; use her enamel ware or the new glass and ceramic ware. Don't use her copper-bottomed tea kettle or gold-rimmed cups or “silver”. Use her wooden spoons, glass glasses, and plain dishes, her wooden and straw bowls and enamel pots and pans.

Salt

Should you avoid salt? No. But a good salt rule is to either cook with it or have it on the table, but not both. Use aluminum-free sea salt, and make sure the salt is sterilized by heating five minutes at 400°F in a glass pie plate to kill mold. (Sea gulls fly over the salt flats where sea salt is gathered. Their droppings provide a medium for mold.)

Fig. 38 Sea salt flats are often roosting places for sea gulls.

The best salt is a mixture of 1 part of your aluminum-free, sterilized sea salt and 1 part potassium chloride (another kind of salt, see *Sources*). Potassium ousts sodium (salt) from your body, so you can use twice as much of this kind of salt! Also, the extra potassium helps lift fatigue and has other benefits.

Always use a non-metal salt shaker with a closable lid to keep out moisture. Don't put rice in your salt because it invites mold.

Since you, the cook, know where the salt is, (mostly in soups and stews), don't serve as much of these when there is heart and kidney illness or high blood pressure. Don't put salt in cereal, cooking vegetables, or other dishes. Just leave it out! Use herbs instead. **Tip:** encapsulated herbs stay fresher and are more potent. Fenugreek and thyme are the most beneficial of the common cooking herbs. Just open a capsule and season.

Tremor

Tremor is a symptom, not a natural part of aging. The nerves controlling the hands and arms are poisoned. The nerves originate in the brain where the poison has accumulated. What is the poison? Did it happen long ago? No! It could have happened as long as two weeks ago but not longer!

Tremor is the result of ongoing poisoning! It is important to find the poison as soon as you can since the rest of the body will soon be affected, too. Search your memory for the new things that happened in the last two weeks. It is a herculean task but only gets harder each day, so keep notes as you ask: Is there new carpeting? Is there a new furnace? Is there a different water supply? Is there a new hair dresser? Did somebody bring a vase of fresh flowers? Is there a new laundry person? Was the place sprayed for insects? Is there a new medicine (drug) or supplement? Was remodeling done? Is there a new food?

The list is endless and the situation looks hopeless because so many new things can happen in two weeks.

Rather than asking individual questions like these, let's ask only **five general questions** and have the assurance that one of them will catch the culprit.

1. Is it in the **air**? This will catch insecticide, flowers, carpets all together.
2. Is it in the **water**?
3. Is it in the **medicines** or **supplements**?
4. Is it in the **clothing**?
5. Is it in the **food** or on the dishes?

To answer each question, test the item using your Syncrometer searching technique. Make a test substance. Then search a saliva sample for it.

To test the air, take a dust sample off the kitchen counter or table (this gives you fresh dust). Pick up the dust with a few wipes by a small, two inch square of damp paper towel. Place in a resealable baggy.

To test the water, make samples by putting about a tsp. of hot, cold, or filtered water into a resealable baggy with a bit of paper towel in it. Try to get the first morning water before it has run.

To test the medicines and supplements put one of each in resealable baggies.

To test clothing (laundry) use a bit of it, (such as a sock) rolled up tightly.

Testing food is the biggest job. If there are leftovers in the refrigerator or freezer, this helps. You can combine all the leftovers in a single baggy. Frozen things don't need to be thawed for testing. Still, the chance of missing a food culprit is quite high. Be sure to test everything eaten in a two week time period: unusual things like popcorn, candy, crackers, cookies, health foods and special powders. A consolation is that you will find a number of bad foods that are not necessarily the tremor causes but which cause other health problems.

Tremor Remedies

1. Let us imagine that the air (dust) sample proves toxic (resonates with the saliva sample). What is it in the air that is toxic? These are the biggest suspects:

- CFCs (from leaking refrigerant, check refrigerator and air conditioner by removing them from the house and retesting the dust after three days, or simply buy a non-CFC variety)
- vanadium (from leaking house gas from a fuel line—repair)
- arsenic from pesticide (switch to boric acid)
- solvent from a lamp or can of lighter fuel (take them out of the house)

- chlorine from water (use carbon filters; remove bleach bottles)
- asbestos (clothes dryer belt)
- paint thinner, carpet cleaner (remove from house)
- fiberglass from bare insulation somewhere (fix holes in ceiling or wall)
- formaldehyde (new foam bedding, new fabric or clothing)

2. Suppose the water proves toxic (appears in your white blood cells); search for lead, copper, and cadmium. Although municipal water tests occasionally detect small amounts of propyl alcohol, benzene, or wood alcohol, I have never detected them—you need not search for them.

3. If the medicines are toxic just change brands. Your pharmacist can help you find a replacement brand. Find which ones are toxic and stop using them. (Don't use up any of the polluted pills or supplements.) They do more harm than good.

4. If the clothing or dishes are toxic (appear in the saliva), suspect

- cobalt
- PCB
- aluminum

Stop all detergent use. Use borax, or washing soda, or use paper plates and plastic cups (not styrofoam).

5. If the food is toxic, suspect

- food molds
- aluminum
- bacteria

Don't eat that food any more.

6. Any bacterium or toxin that invades certain brain centers can cause tremor. The most common culprits are *Shigella*, mercury, thallium and arsenic. Try to identify these for your sake as well as your loved one.

The *Salmonellas* and *Shigellas* will have come from some food. Be sure all dairy foods, including butter and whipping cream, are sterilized. Parasites come from meats. Be sure no undercooked or fast food meats are eaten. Kill all bacteria, viruses and parasites with a zapper. Sterilize fingernails with grain alcohol to prevent reinfection. Use Lugol's (see *Recipes*) and the Bowel Program (page 546) to clear them up.

Bacteria, coming from teeth and jaw (bone infections, called *cavitations*) may not seem as recent as two weeks. Indeed, they may have been there for ten years. But something recent may have aggravated them, so they now can enter more easily into the blood and brain. You may never find out what caused this. It is wisest to check this possibility with a dentist before doing weeks of other testing. See a dentist who can find and clean cavitations. Do dental repair according to the principles described in the dental section.

Going after a tremor problem in this logical way always finds the cause of tremor

whether its a simple short attack or a situation of long standing tremor with head shaking and drooling. If your situation is extra difficult, you will at least improve it and stop its progression. **This includes Parkinson's cases.**

In cases of Parkinson's disease I often find the bacterium *Clostridium tetani*, well known for causing stiffness. It hides under tooth fillings, too.

Remember, there is a consolation for doing all this work. When you find the culprit, you not only will be stopping the tremor, you will be improving a lot of other conditions along the way. Conditions like hesitant speech, shuffling walk, getting up stiffly and slowly from a chair. These are extra dividends for your loved one. And you have learned which things to avoid for yourself too.

One more thing, don't take no for an answer.

If the problem is a drug, don't accept “No, it can't be changed.” Everything can be changed. Go to higher levels of authority. Doctors are very understanding and sympathetic with your intent. If the problem is leaking house gas, don't accept “We can't find a leak,” implying there is none. Go to a building contractor or the Health Department. Their equipment is better. If you get nowhere, change to electric utilities. If the problem is auto exhaust, don't accept “But we have to use the kitchen door to the garage, because it's cold outside.” Conveniences vs. tremor is no contest. Everyone benefits by excluding auto fumes from the house. Lock the door and cover it with plastic. By the time you have identified the culprits (probably 20 hours of work) surely you have won the right to make changes.

Often others are not impressed. Even when the tremor lessens and the elderly person plainly states they feel better, family members may disregard your recommendations. **Get tough!** It's your loved one. You have something at stake too. Be sympathetic with negative responses. But very firm. Make their choices clear:

- Either the inside door to the garage gets sealed off or the cars and lawn mower get parked outside and anything containing gasoline or solvents gets put in a detached shed.
- Either the girls use their hair dryers in their bedrooms with the door closed or you'll get them all new ones that don't contain asbestos.
- Either the clothes dryer gets a new belt or it is taped securely shut and the laundry is dried on the line or taken to a Laundromat.
- Either the plumbing gets changed (to plastic) or each faucet in the house gets a filter. (If your plumbing is corrod-

ing, not even filters are a good idea—you would have to change them every week!)

Weakness

should not be taken for granted in the elderly. Especially if they themselves complain about it. It isn't normal for them. Sometimes they will describe “spells” of weakness. This is an important clue. Check the pulse immediately. Count for 30 beats at least. Are there missing beats? There should not be! Missing two beats in a row certainly can produce a weak or “sinking” spell. The brain and body need every pulse of blood sent out right on time.

Check into caffeine use first. Take it all away. Caffeine speeds up the heart; then the overworked heart has to “take time out” for itself by missing a beat. Don't switch to decafs because this introduces solvents and new problems. If no other natural beverages appeal, serve hot water with cream and cinnamon. After stopping caffeine use ask: Is the pulse too slow or too fast? The pulse should be between 60 and 80 beats per minute.

If it is lower than 60, a medicine may be at fault. Ask the clinical doctor about it immediately. A slow pulse could certainly bring about weakness. A young athlete may have a slow pulse legitimately, due to having a very strong efficient heart, but your elderly person does not fit this category. The cause must be found.

If the pulse is quite high, over 100 perhaps, this will wear the heart out much sooner than necessary. Ask why it is beating so fast? A probable answer is that it is so weak that it has to beat faster to keep up with its job of circulating the blood. What is making the heart so weak?

Heart Disease

When the heart is enlarged, the valves don't quite close where they should, making the work harder for it and weakening it. It may be called “congestive heart failure.” Why is it enlarged? Possibly because it is so weak! Yes, it becomes a vicious cycle, getting worse and worse. But you can break into this cycle and get it all reversed again. The real culprit is parasite invaders and toxic pollutants.

The most common parasite heart invaders are *Dirofilaria*, heartworm “of dogs” and *Loa loa*, another small filaria worm. At one stage these worms are so tiny that they can slide through the smallest blood vessels. They are very contagious. Even persons who don't live with a house dog can pick up heartworm. *Loa loa* is thought to be a tropical parasite but it is alive and thriving in the USA! The source of *Loa loa* seems to be tapeworm stages; this is not a certainty.

Both heartworm and *Loa loa* are very easy to kill with a zapper and both are very easy to pick up again. Treat your elderly person twice a week if there is any heart problem. It makes no difference that the house dog is getting monthly preventive treatments for heartworm. They pick it up daily and have thirty days to develop it and give it to others between treatments. Killing the dog's parasites twice a week with a zapper would be very helpful to you. These heart parasites may not cause any pains, yet disturb the rhythm or the pulse of the heart and cause it to enlarge.

Staphylococcus aureus is a bacterium hiding out in far away places like pockets left under teeth when they were extracted or along root canals. Make sure extractions heal and don't leave permanent cavitations where bacteria can live. Ask a dentist familiar with cavitations to do a mouth search. Once the mouth source is cleaned up, the bacteria do not come back to the heart (after one last zapping). If they do, go back to the dentist!

Killing these three invaders (heartworm, *Loa loa, Staphylococcus aureus*) should cure an irregular heart beat immediately (within a day).

If the elderly person is on a heart-slowing drug, check the pulse twice a day after zapping to make sure it doesn't drop too low. They may need to be off their heart medicine. Nobody will notice the relief of going off this medicine as much as you. The sunshine breaks out! Your loved one can smile again at little things! Even interest in sex returns so watch out! Life is normalized when drugs, especially beta blockers are gone. (Other heart medicines, such as Digitoxin,™ don't have this depressive effect. They are used to make the heart beat stronger, not to affect the rate of beating.)

The pulse should be around 70 beats per minute and perfectly regular.

If it isn't, there is still something wrong. Get rid of toxic body products and house pollutants. Test your air for gas leaks frequently. Gas heat, gas stoves and gas water heaters are notoriously leaky. Weather changes, namely temperature changes make pipes expand or shrink—leaving cracks! The gas is toxic and a small amount can't be smelled. What a predicament! Delivering poisonous house gas to our homes in pipes that are not fail-safe is an archaic practice. Especially when the blood test shows a high “total CO_2" level, near the upper limit, search for an air pollutant like house gas or auto fumes. And read the sections in this book on pulse (page 289) and brain problems (page 278) very closely for more things to check.

With the heart regular again, it will be much stronger, too, since it doesn't have to work against itself. This strength is necessary to push the blood into the farthest “corners” of the body, especially the hands and feet, and warm them up! If your loved one constantly has cold hands or feet, try to improve circulation. Dissolve phosphate crystals with the kidney herb recipe. Give niacin (page 279). Give cayenne capsules (one with each meal). Blood thinning drugs to improve circulation are dangerous—use only if the doctor insists. Monitor blood clotting time if your loved one is on a thinner.

Heart/Kidney Relationship

A strong heart is necessary, too, to push the blood through the kidneys. Often a kidney problem is linked to the heart disorder. Kidneys are made of tubes that get finer and finer. It takes pressure, namely strength, to push the blood through them so wastes and extra water can be let down the kidney tube. Think of the kidneys as a colander full of tiny holes of various sizes that let certain things through them but not bigger things. These holes are constantly being adjusted by the adrenals which sit right on top of the kidneys and "supervise".

If the elderly person is not producing four cups of urine in a day (24 hours), it is not enough. The body cannot get enough cleansing action from less than four cups. More liquid must be consumed. If most of the urine is passed in the night this reflects on unhealthy kidneys. Use the kidney herb recipe—but only half a dose (so it will take six weeks instead of three to see good effects). As the tiny "colander" holes open up there is freer flow and many more trips to the bathroom result. The urine loses its awful odor (no ammonia, acetone and bacteria!) and gets a clear look that shows no sediment. Now that water and wastes (urea and uric acid and other acids) can leave the body quickly through more holes, it takes less pressure from the heart to get blood pushed through the kidneys. This brings relief to the heart because its work is easier. The heart and kidneys work together. Like horse and wagon the heart provides the power and the kidneys follow.

This is why heart medicine and diuretics are commonly given together. *Diuresis* (urine flow) helps the heart and a stronger heart helps the kidneys. Similarly, they fail together. In the old days this was called *dropsy*. Urine that should have left the body is backed up in the tissues. Sometimes it shows up in pockets that hang like giant oranges from the skin.

Even (especially) when the strongest diuretics (Maxzide^{TM}, Lasix^{TM}) fail to work, even when coupled with strong heart medicine (Digitoxin^{TM}), the kidney herb recipe can bail you out of the emergency.

The secret is in the varied actions of different herbs. This makes them work together. Be very careful to keep the herb tea sterile by reheating every fourth day. Freeze unused larger amounts. If too much is drunk at once, especially on the first day, a stomach ache can develop and a pressure felt in the bladder that is most uncomfortable. Go extra slow on the first few days, even though you find it quite tasty, so there is no discomfort (only lots of bathroom visits).

Diuretics

As soon as urine is flowing better, blood pressure may drop. Keep track of this twice a day with a modern electronic finger device (not an arm cuff that itself can break blood vessels). Cut down on drug diuretics gradually, using only ¾dose the first day, then ½dose, then ¼ dose. The amount of urine produced or the weight of the person can be used to assess how effective your method is. **Your goal is to not need any drug diuretics.** Again, mood will improve dramatically when diuretic drugs are removed for your loved one. The sense of humor comes back:

be prepared for some new jokes and new laughter. Give yourself good grades for this accomplishment.

Don't throw out the drug diuretics. After being off them for a while, they become more potent again. So if an emergency (sudden edema) should arise they could again be useful.

With a parasite and pollution-free heart and a low-resistance, freely flowing kidney, some reserve strength will soon be built up. Weak spells are gone and forgotten. Your loved one is walking better, needing less sleep, and a "golden age" finally arrives. It is free of pain, free of medicine, free of shots and doctor visits, free of dementia, free of the dreadful weakness that demands so much help. They are free to enjoy family and friends again. Seeing themselves gain strength and be able to do more for themselves gives the elderly a sense of pride.

Give them the credit for improving their health. When they balk at having to take herbs or vegetable juice, remind them of the days they were on a handful of pills and still had heart failure, pain and kidney disease.

Feeling Comfortable

If all this improvement doesn't warm up their feet or hands (feel them yourself) put extra socks on. Warming up their feet might have the effect of warming up hands, too. But if it doesn't, raise the room temperature.

Being forced to feel cold is an undeserved misery of the aged.

Others must accommodate to the elderly's need for body warmth. The thermometer does not tell all. Comfort is paramount for each of us. Younger persons can undress for comfort. The elderly can't make changes for themselves. Usually, by the time they are complaining of cold, they have suffered a long time, feeling too guilty to request a temperature hike. Ask your loved one. They will appreciate it. A shawl, a lap-blanket, woolen sweater, long underwear and fleecy thermal outerwear help a lot. But if your hand still senses cold feet, the thermostat must still go higher. Cold body temperature is an invitation for fungus and viruses to multiply. Viruses escape into the body daily. Don't give them the advantage.

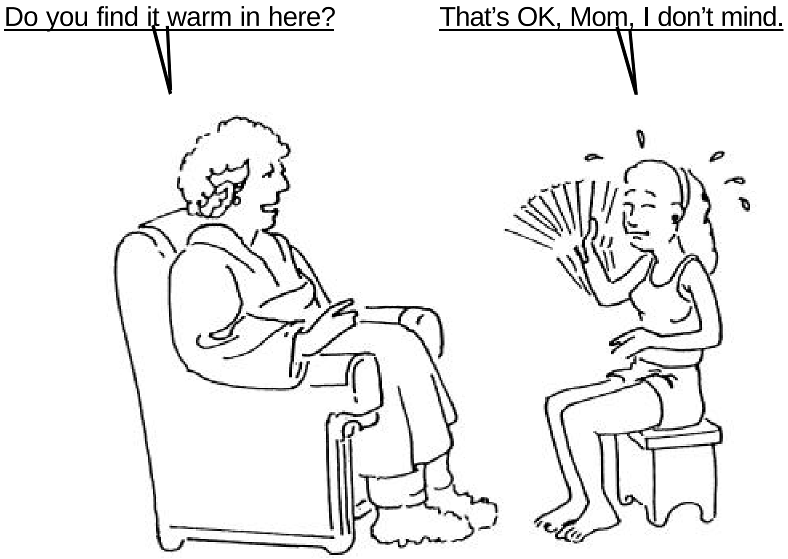
Do you find it warm in here?
That’s OK, Mom, I don’t mind.

Fig. 39 Too hot and too cold.

But unless all these things are tried, don't raise the temperature. It is much healthier to be warmly dressed and breathe cool air than to be lightly dressed in an 80°F room. The heart, especially, benefits by the cooler air temperature.

For this reason, it is important to have air conditioning during the hot season. Keep your elderly person warmly dressed, away from air conditioner or fan drafts, but keep it cool. As cool as their body temperature and comfort will allow. Don't seat an elderly person under a fan in restaurants. The blowing air not only chills them, but brings filth in dust that blows in the air currents. Bacteria and viruses are soon to follow.

Being comfortable, knowing you are there to care for them, brings out the best in your elderly person. They might get well enough to long for a genuine relationship again. And put you to the test. The test of listening. This can be very rewarding if they are still able to communicate and distill their life experience into wisdom for you. If you can listen and be interested in their distillations or their ramblings their longing for relationship will be fulfilled. There's just one catch, if the new found interest in communication can't be expressed. If they haven't communicated much for a long time, it would be easy for you to miss a simple fact. They can't hear!

Hearing Loss

The hearing deficit in an elderly person is always much greater than they or you realize. Try to understand these communications:

What you (caregiver) say	**What the elderly person hears**	**What they think**	**What they respond**
"Look at the sunshine, isn't it a nice day?"	"Look at something, isn't it a nice bay?"	What a stupid sentence. But she wants an answer. So here goes,	"MmmHmmm" (Meaning, "yes.")
"We'll have such a nice walk."	"Have a…rock."	Rock what? I'm not in a rocking char. I better not say anything.	
"What kind of soup should I make?"	"What kind of hoop should I make?"	I haven't used a hoop for embroidery for a long time. Why is she digging that up. I better say I don't know.	"I don't know."

Don't let this ruin your relationship. It takes the fun out of their life as much as yours. **Get hearing aids**.

Don't ask your loved one if they need a hearing aid. The answer may be “I'm not deaf.” You can go about it more persuasively. Every time you talk, come quite near to the person's ear and speak loudly. They can see that you must come near. If they are fighting against the whole business, they will say things like, “You don't have to shout. I can hear you.” (It's what you're saying that's wrong!)

Arrange for a hearing test. It is free. That will appeal. The results of a hearing test, as it is told by a salesperson, is much more persuasive than you can be. Let the salesperson use his or her special talents to sell your loved one on hearing aids. But you make the choice on quality. Both you and your elderly person deserve the best tone quality that is made. Plus a regular cleaning service. Most companies do offer this but don't tell the customer because of the dreadful amount of time it would take if everybody took advantage of it. You take advantage of it. Clogged hearing aids are the most troublesome feature of any of them—and never mentioned! Make it a rule to buy your batteries at the same hearing aid office where they are cleaned free of charge. This repays them and serves the elderly best. Hearing loss is too subtle to leave to chance; have the hearing aids cleaned each time you buy fresh batteries (about three months). Take your loved one to a nurse for ear cleaning every six months after hearing aids are begun. Wax and debris accumulate behind the aid because the channel is stopped up.

With hearing aids that hear, and kidneys that flush and a heart that beats strongly, your elderly person may choose to attend concerts again, go to church or gatherings—and leave you out of the picture. Give yourself good grades for this achievement. Get them incontinence pants, get regular taxi service. **Do whatever it takes to get your loved one out into the world again!**

If the excitement of a night out keeps him or her from sleeping use **ornithine** and **valerian** capsules. They are good for the health anyway. Hot milk and a piece of cake (homemade, never chocolate) may do as much.

But if insomnia is the rule, not the exception, you need to go after it as a special problem.

Insomnia

is always associated with high levels of parasites and bacteria. This leads me to believe it is their waste products, namely ammonia, that really causes insomnia. Read the section on Insomnia under Sleep Problems for more on this (page 243).

A sound sleep refreshes as nothing else can. Your elderly person will have more energy throughout the day and a better mood if sleep was good.

Of course you must guard against daytime napping if nights are sleepless. Sleep at night is more important than sleep by day. They are not equal. Try to turn your loved one's cycle back to night-time sleeping if it has become turned around by bad kidney function (frequent urination at night) or too much daytime napping. Sleep can be disturbed by taking vitamins at bedtime too. Sleep is enhanced by taking magnesium. Limit bedtime supplements to magnesium, ornithine, valerian (6 capsules) taken with hot milk.

Healthful Habits

If your loved one had his or her way, they would drive the car forever, wear the same cosmetics forever, smoke or chew tobacco forever and eat their favorite dessert forever.

But if you are the caretaker, you know that some things must stop. You also know that gentle persuasion is useless; it merely erodes your relationship. Here are a few tips.

1. Ask your loved one to ask their doctor (clinical doctor or trusted medical advisor) the following question: "Would it be better for my lungs to stop smoking?" Be present so the question does not turn into "Will smoking a few cigarettes once in a while kill me?" Hearing it from the doctor is what's needed.
2. Don't ever purchase something you believe is detrimental to your elderly person. Whether it's coffee, cigarettes, beer or lipstick, say "That is something I can't buy for you; it's against my principles." Don't be surprised if you cave in a few times to some super ruse they use on you. But the next time, have your answer ready.
3. Let your family and other caretakers know you are no longer supplying these items (the car keys, the wine bottle, the codeine-containing pain pills). Try to get cooperation. Discussing it with your loved one may do more harm than good. If they start the discussion, you end it. This is not a task for the timid! After it's done, you'll wonder what was so difficult.
4. Don't buy a wheelchair if your loved one can still walk with your help. Stay with a cane as long as possible. Then the walker. Stay with a walker as long as possible. Then your personal help. Once a wheelchair has been accepted, the last bit of exercise, walking, is lost. Fight against it. Hide it in a far away closet.

Aging is necessary but chronic illness and pain are not.

If you have managed to free your loved one from having to take pills or from certain disabilities that would soon require pills, you can give yourself great credit. Perhaps you, too, will find the needed natural help when you are aged and have lost your authority and your way mentally. Our lives are all foreshortened, much like the life of a domestic steer's. Does a captive animal learn from seeing its companion disappear? It does nothing to escape its fate. Should we accept our fate with the same docility? None of us can remember how things were in precivilized times. We are eager to believe the present is the best time that has ever been. The steer, too, has its feed provided, its water provided, its shelter for the night provided, seemingly the best time it ever had. Perhaps the price we pay for civilization, like the steer's price, is simply too high. There must be other ways. As a society, we should search for our lost longevity.

Not Old Yet

For persons merely over 40 and not ill a most important rule is: **don't overdo**.

It's easy to injure tendons and muscles by pulling or stretching them too far. Don't try, with almighty determination to open a jar. Leave it for stronger hands. Don't try to stretch a gigantic stretch to reach something on a high shelf. Wait for someone taller to come along and reach it down for you. Don't do exercises, that have left you with strains in the past. Do different exercises. Don't do new exercises, with a determined approach; start moderately; stay moderate. If you start a new job that uses a foot, leg, hand or wrist a lot, more than it was used before, even though it's in ordinary use, it can feel injured. These stresses and strains invite bacteria toward themselves. Bacteria from the liver or your own intestines find these strained tissues immediately and intensify the pain. Kill the bacteria with a zapper, cleanse the liver, and start the Bowel Program if this has already happened to you.

Don't kneel on the floor to do some cleaning job. Don't sit on the floor at a gathering. Don't let your grandchildren take “rides” on you. Don't bend over as far as you can to pick something up. Let your family know you are aiming to reach 100 years in healthy condition.

Super Longevity

We should be able to live to 140 years of age. Middle age might begin at 65, old age at 90. What is the clue to long life? Surely it is keeping all your cells healthy so they can coordinate the constant tasks of nourishing themselves, removing their wastes, plus whatever job that cell was meant to do.

Since your cells divide and therefore start again at age zero, even though you are 90, why do you age at all? Only the nervous system does not divide. Is it the nerve that ages then, and decides death for each of us?

It would never do to live forever. It would clutter the planet. Are evolutionary forces at work preventing this? But we have never cluttered the planet yet. So how could evolutionary forces have “learned” to establish death to prevent overpopulation? Living to age 140 is not living forever. Wisdom comes with age. Wisdom would serve our society well. Perhaps wisdom is needed more than ever for humans to survive. Part of the human tragedy is war. Wisdom gathered from a knowledge of history, might help end wars! But history can't be gathered and understood in less than 100 years. If centenarians can't think well or express themselves, their perspectives are lost. Longevity seems a very useful trait if only it were accompanied by health!

Dying

Although it can't be avoided it can be helped. Dying alone is the most barbaric of all society's practices.

It is still like death in the forest amongst chimpanzees. Unable to feed, to run, to call for help brings pain, fear, loneliness and finally, death. When your loved one is in the mood to talk about dying, listen. Especially if it is about their own imminent death. This doesn't obligate you to carry out any of their wishes. When death approaches the important thing is simply to be there. They may not have requested this, out of a sense of guilt or masochism, or plain dementia. But it is the most primitive of needs, the same as having a loved one nearby during childbirth. It is just a presence that counts. How can you be there if you have a job or are attending a family's needs? It is a time of great frustration for you. All your hard work and successes are culminating in one grand failure!

The good news is that it need not be you who attends your loved one every minute of the last week. Pay for someone to sit—someone who is recognized. But arrange for immediate privacy when you return. No matter how much your loved one admonishes you to go about your business, you will know when you share the final minute that your presence helped. The loneliness of the last coma, the last silence is unspeakable. Give yourself the reward of knowing you shared the pain, fear and silent cry for help.

Curing Cancer

Cancer is no longer the deadly disease it once was. In fact, you can clear it up in less time than it takes to get a doctor's appointment for a check up. If you notice a lump or think you might have cancer, don't rush to see a doctor first. Rush to clear it up, yourself, first. By the time your doctor's appointment arrives, you can be sure of a negative test.

Another Fluke Disease

Cancer is so easily cured because it is a parasite-caused disease. Kill the parasite and you have stopped the cancer. This does not mean you have also stopped being ill. If the cancer damaged your ovaries or prostate, you still must heal these organs. The Ca-125 or PSA test will not drop to normal unless you begin to heal them. These markers reflect the condition of the organ, not its malignant nature. Remember that killing a mosquito does not remove the lump it caused. That will take its course. It will heal beautifully if you let it. But if you scratch it mercilessly, it will take longer. Removing toxins from the affected organs lets them heal.

Fortunately for us, **cancer is not like a fire, unstoppable once it has started**. It takes only 7 minutes to zap all the parasite adults and their stages which cause your cells to multiply.

The responsible parasite is *Fasciolopsis buskii,* the **human intestinal fluke**, a flatworm. It is a human parasite although it can also parasitize other species. It normally lives quietly in the intestine. (The goal for all larger parasites is to live quietly. After all, your demise is their demise.) They were meant to pass their thousands of eggs with your bowel movement, outside, to some pond where snails live. But when the eggs hatch before they leave your body and are allowed to continue their development inside you, the setting is right for cancer to develop.

If the fluke eggs and other stages go through their development in your breast it can become breast cancer. If it is in the prostate it can become prostate cancer. And so on. Each different kind of cancer means the developmental stages of the intestinal fluke are present there. Only one more thing is needed to bring about an avalanche of reproduction, so that hundreds of little larvae turn into hundreds more in a short time: a growth factor. It makes them multiply and your cells are similarly affected. This growth factor, *ortho-phospho-tyrosine* (and possibly, also, *epidermal growth factor* and *insulin-dependent growth factor*) really begins your cancer.

Teamed With A Solvent

The good news is that this growth factor, which is essential for cancer to develop, cannot be made, without the presence of an abnormal solvent, **propyl alcohol** (or more exactly, isopropyl alcohol).

Without taking in propyl alcohol you could never get cancer. It takes two things, together, to give you cancer: propyl alcohol and the human intestinal fluke parasite.

Since it takes a frequency generator (3 minutes at each of 434, 432, 427, 425, 423, 421 KHz) or zapper mere minutes to kill the fluke and its stages you will be stopping the production of growth factors immediately.

Zap yourself every day for three seven minute periods, until after you are completely well.

After killing the flukes, those growth factors already formed will disappear in one hour. Your malignancy is stopped. It cannot return unless you infect yourself with the parasite again!

Getting rid of propyl alcohol is also a simple task. Once you have stopped using it, the last remnants leave your body within three days.

We must marvel at the body's wisdom and capabilities for restoring health. You are not permanently damaged by this large and hungry parasite. Given half a chance your body will throw the rascals out and restore order in your tissues.

Read the list of foods and products that are polluted with propyl alcohol. It is not, of course, usually on the label. If it is used as an ingredient, it is on the label, easy for you to avoid. It is the pollutants not appearing on the label that pose sinister hazards.

INFUSIUM23
INFUSIUM23

All the store bought shampoo I tested had propyl alcohol. Health brands were no exception

CRYSTAL GEYSER

All bottled water is polluted with antiseptics from the bottling procedure. This is a main source of propyl alcohol.

Cosmetics are laced with propyl alcohol. Use the recipes in this book to make your own.

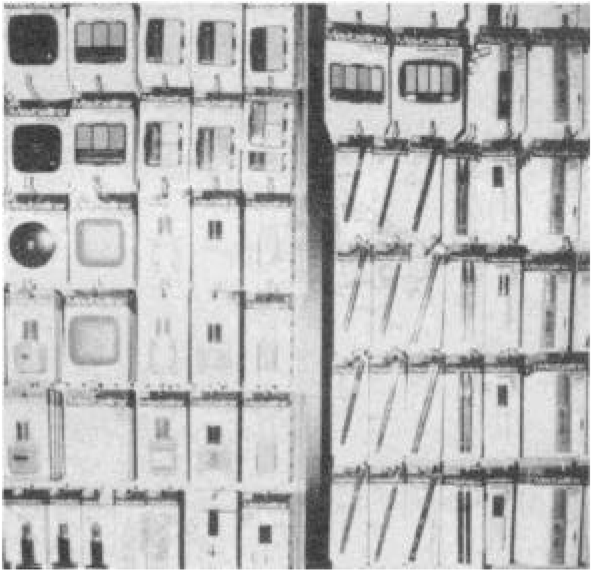

Fig. 40 Products with propyl alcohol.

Propyl Alcohol Polluted Products

THROW THESE OUT

even if propyl alcohol is not listed on the label!

- **shampoo**, even health brands
- **hair spray** and **mousse**
- **cold cereals**, even "natural" granolas
- **cosmetics** (make your own)
- **mouthwash**
- **decaffeinated coffee**, Postum^{TM}, herb tea blends (single herb teas are OK)
- **vitamins, minerals and supplements** (unless you test them)
- **bottled water**, distilled water, or spring water
- **rubbing alcohol**
- **white sugar** (brown is OK if detoxified)
- all **shaving supplies** including aftershave
- **carbonated beverages**
- store-bought **fruit juice**, including health food brands

Tear out this page, put it on your refrigerator, and make a copy to stick on your medicine cabinet. Remember propyl alcohol is also called propanol, isopropanol, isopropyl alcohol and rubbing alcohol. You won't drop dead from getting propyl alcohol, but your cancer will flare up with each small addition.

Avoid the entire list, meticulously. Even using one of these, **like your favorite shampoo or bottled water**, will result in failure. Your body will continue to make *human chorionic gonadotropin* (hCG) and the pathology will remain "indeterminate"—not cancerous but not well, either. If you have learned the new bioelectronic technology described in this book, you can test all your foods and products for isopropyl alcohol.

When you find a beverage that is free of propyl alcohol, it may have other pollutants. Xylene and toluene invite parasites to the brain. Wood alcohol invites them to eyes and pancreas. The diseases caused by these are not as frightening, perhaps, as cancer, but entirely avoidable. For this reason, I suggest in this book that you go back (actually "forward") to self made products, unprocessed food and a limited number of tested supplements.

Getting Well After Cancer

The ravages of cancer must be healed once the malignancy has been stopped. This is where *carcinogens* play a role. The lung lesions will not heal unless cigarette smoking, freon, asbestos, and fiberglass exposure is stopped. Carcinogens were thought to be the cause of cancer. Actually, they drew the cancer to the organ. Nickel draws cancer to the prostate. Barium found in lipstick draws cancer to the breast. And so on.

The following toxins can be present in any organ! I consider these to be our most serious threat. Starting with the worst, they are:

1. **Freon** (same as **CFCs** or refrigerant). I have not found a single person to be entirely free of it, including persons without cancer or disease. But in cancer sufferers it is always concentrated in the cancerous organ, and facilitates the accumulation of other toxins. This makes your refrig-

erator, if it leaks even microscopically, the top cancer hazard in the country!

2. **Copper** from water pipes. All cancer sufferers have an accumulation in the cancerous tissue. This makes plumbing the second greatest hazard.
3. **Fiberglass** or **asbestos** is present in about 25% of the cancer victims I see.
4. **Mercury** as in tooth fillings.
5. **Lead** from solder joints in copper plumbing. All colon cancer sufferers test positive.
6. **Formaldehyde** as in foam bedding and new clothing.
7. **Nickel** usually from dental metal. Most prostate cancer sufferers test positive.

At first, tumors are benign—what a relief to find your tumor was benign. Its true nature is still unknown, but it contains freon, other toxins, and later propyl alcohol! Since tumors are often large, many centimeters in diameter, and toxins do not occupy much space, there is much unidentified substance. These tumors can multiply and enlarge, as in fibrocystic breast disease, all without being malignant. But what a convenient place for baby stages of the fluke to hide out and multiply, safe from your immune system. After being colonized by fluke stages, the tumors become malignant. Eventually they also become infected when metastasis sets in.

All malignancies have the same two fundamental causes: intestinal flukes and propyl alcohol. Whether you have a rhabdomyosarcoma or a mesothelioma or melanoma, you can cure it quickly, never to return.

Herbal Parasite Killing Program

Flatworms, roundworms, protozoa, even bacteria and viruses are remarkably easy to kill using a combination of zapping and this herbal program. Thus it is not just for cancer, but a general treatment that can benefit almost every illness

1. **Black Walnut Hull Tincture Extra Strength** (see *Recipes,* page 543):
   Day 1: (this is the day you begin; start the same day you receive it)
   Take one drop. Put it in ½ cup of water. Sip it on an empty stomach such as before a meal.

   Day 2: Take 2 drops in ½ cup water same as above.
   Day 3: Take 3 drops in ½ cup water same as above.
   Day 4: Take 4 drops in ½ cup water same as above.
   Day 5: Take 5 drops in ½ cup water same as above.
   Day 6: Take 2 tsp., all together in ¼ cup water. Sip it, don't gulp it. Get it down within 15 minutes. (If you are over 150 pounds, take 2½ tsp. Do not take more than 3 tsp. because no additional value has been observed.)

   This dose kills any remaining stages throughout the body, including the bowel contents, a location unreachable by electric current. The alcohol in the tincture can make you slightly woozy for several minutes. Simply stay seated until you are comfortable again. You may put the tincture in lukewarm water to help evaporate some of the alcohol, but do not use hot water because that may damage its parasiticide power. Then take niacinamide 500 mg to counteract the toxicity of the alcohol. You could also feel a slight nausea for a few minutes. Walk in the fresh air or *simply rest until it passes. You may add more water or honey or a spice to make it more palatable.*

For a year: take 2 tsp. Black Walnut Hull Tincture Extra Strength every week or until your illness is but a hazy memory. This is to kill any parasite stages you pick up from your family, friends, or pets.

Family members and friends should take 2 tsp. every other week to avoid reinfecting you. They may be harboring a few parasite stages in their intestinal tract without having symptoms. But when these stages are transmitted to a cancer patient, they immediately seek out the cancerous organ again.

You may be wondering why you should wait for five days before taking the 2 tsp. dose. It is for your convenience only. You may have a sensitive stomach or be worried about toxicity or side effects. By the sixth day you will have convinced yourself there is no toxicity or side effects.

***Going faster**. In fact, if you are convinced after the first drop of the restorative powers of Black Walnut Hull Tincture Extra Strength, take the 2 tsp. dose on the very first day.*

***Going slower**. On the other hand, if you cringe at the thought of taking an herb or you are anxious about it's safety, continue the drops, increasing at your own pace, until you are ready to brave the decisive 2 tsp. dose.*

2. **Wormwood** capsules (should contain 200-300 mg of wormwood, see *Sources*):

Day 1: Take 1 capsule before supper (with water).
Day 2: Take 1 capsule before supper.

Day 3: Take 2 capsules before supper.
Day 4: Take 2 capsules before supper.

Continue increasing in this way to day 14, whereupon you are up to seven capsules. You take the capsules all in a single dose (you may take a few at a time until they are all gone). Then you do 2 more days of 7 capsules each. After this, you take 7 capsules once a week forever, as it states in the Maintenance Parasite Program. Try not to get interrupted before the 6th day, so you know the adult intestinal flukes are dead. After this, you may proceed more slowly if you wish. Many persons with sensitive stomachs prefer to stay longer on each dose instead of increasing according to this schedule. You may choose the pace after the sixth day.

3. **Cloves**:
   Fill size 00 capsules with fresh ground cloves; if this size is not available, use size 0 or 000. In a pinch, buy gelatin capsules and empty them or empty other vitamin capsules. You may be able to purchase fresh ground cloves that are already encapsulated; they should be about 500 mg. Grocery store ground cloves do not work! Either grind them yourself or see Sources.

   Day 1: Take one capsule 3 times a day before meals.
   Day 2: Take two capsules 3 times a day.
   Days 3, 4, 5, 6, 7, 8, 9, 10: Take three capsules 3 times a day.
   After day 10: Take 3 capsules all together once a week forever, as in the Maintenance Parasite Program.

Take **ornithine** at bedtime for insomnia. Even if you do not suffer from insomnia now, you may when you kill parasites.

Parasite Program Handy Chart

Strike out the doses as you take them.

	Black Walnut Hull Tincture Extra Strength Dose	**Wormwood Capsule Dose (200-300 mg)**	**Clove Capsule Dose (Size 0 or 00)**
Day	drops 1 time per day, like before a meal	capsules 1 time per day, on empty stomach (before meal)	capsules 3 times per day, like at mealtime
1	1	1	1, 1, 1
2	2	1	2, 2, 2
3	3	2	3, 3, 3
4	4	2	3, 3, 3
5	5	3	3, 3, 3
6	2 tsp.	3	3, 3, 3
7	Now once a week	4	3, 3, 3
8		4	3, 3, 3
9		5	3, 3, 3
10		5	3, 3, 3.
11		6	3
12		6	Now once a week
13	2 tsp.	7	
14		7	
15		7	
16		7	
17		Now once a week	
18			3

At this point you do not need to keep a strict schedule, but instead may choose any day of the week to take all the parasite program ingredients.

Continue on the Maintenance Parasite Program, indefinitely, to prevent future reinfection.

Maintenance Parasite Program

YOU ARE ALWAYS PICKING UP PARASITES! PARASITES ARE EVERYWHERE AROUND YOU! YOU GET THEM FROM OTHER PEOPLE, YOUR FAMILY, YOURSELF, YOUR HOME, YOUR PETS, UNDERCOOKED MEAT, AND UNDERCOOKED DAIRY PRODUCTS.

I believe the main source of the intestinal fluke is undercooked meat. After we are infected with it this way, we can give it to each other through blood, saliva, semen, and breast milk, which means kissing on the mouth, sex, nursing, and childbearing.

Family members nearly always have the same parasites. If one person develops cancer or HIV, the others probably have the intestinal fluke also. These diseases are caused by the same parasite. They should give themselves the same de-parasitizing program.

Do this once a week. You may take these at different times in the day or together:

1. **Black Walnut Hull Tincture Extra Strength**: 2 tsp. on an empty stomach, like before a meal.
2. **Wormwood capsules**: 7 capsules (with 200-300 mg wormwood each) at once on an empty stomach.
3. **Cloves**: 3 capsules (about 500 mg. each, or fill size 00 capsules yourself) at once on an empty stomach.
4. Take **ornithine** as needed.

	Black Walnut Hull Tincture Extra Strength Dose	**Wormwood Capsule Dose (200-300 mg)**	**Clove Capsule Dose (Size 0 or 00)**
Day	1 time per day, on empty stomach	capsules 1 time per day, on empty stomach	capsules 1 time per day, on empty stomach
1	2 tsp.	7	3
2			
3			
4			
5			
6			
7			
8	2 tsp.	7	3
9			
10			

11			
12			
13			
14			
15	2 tsp.	7	3
and so	**on...**		

The only after-effects you may feel are due to bacteria and viruses escaping from dead parasites! Be sure to zap after taking your maintenance parasite treatment. After-effects also let you know that you did indeed kill something. Try to discover how you might have picked up parasites and avoid them next time.

Pet Parasite Program

Pets have many of the same parasites that we get, including *Ascaris* (common roundworm), hookworm, *Trichinella*, *Strongyloides*, heartworm and a variety of tapeworms. Every pet living in your home should be deparasitized (cleared of parasites) and maintained on a parasite program. Monthly trips to your vet are not sufficient.

You may not need to get rid of your pet to keep yourself free of parasites. But if you are quite ill it is best to board it with a friend until you are better.

Your pet is part of your family and should be kept as sweet and clean and healthy as yourself. This is not difficult to achieve. Here is the recipe:

1. **Parsley water**: cook a big bunch of fresh parsley in a quart of water for 3 minutes. Throw away the parsley. After cooling, you may freeze most of it in several 1 cup containers. This is a month's supply. Put 1 tsp. parsley

water on the pet's food. You don't have to watch it go down. Whatever amount is eaten is satisfactory.

All dosages are based on a 10 pound (5 kilo) cat or dog. Double them for a 20 pound pet, and so forth.

Pets are so full of parasites, you must be quite careful not to deparasitize too quickly. The purpose of the parsley water is to keep the kidneys flowing well so dead parasite refuse is eliminated promptly. They get quite fond of their parsley water. Perhaps they can sense the benefit it brings them. Do this for a week before starting the Black Walnut Hull Tincture.

2. **Black Walnut Hull Tincture (regular strength):** 1 drop on the food. Don't force them to eat it. Count carefully. Treat cats only twice a week. Treat dogs daily, for instance a 30 pound dog would get 3 drops per day (but work up to it, increasing one drop per day). Do not use Extra Strength.

If your pet vomits or has diarrhea, you may expect to see worms. This is extremely infectious and hazardous. Never let a child clean up a pet mess. Begin by pouring salt and iodine[15] on the mess and letting it stand for 5 minutes before cleaning it up. Clean up outdoor messes the same way. Finally, clean your hands with diluted grain alcohol (dilute 1 part alcohol with 4 parts water) or vodka. Be careful to keep all alcohol out of sight of children; don't rely on discipline for this. Be careful not to buy isopropyl rubbing alcohol for this purpose.

Start the wormwood a week later.

---

[15] “Povidone” iodine. topical antiseptic, is available in most drug stores.

3. **Wormwood capsules:** (200-300 mg wormwood per capsule) open a capsule and put the smallest pinch possible on their dry food. Do this for a week before starting the cloves.

4. **Cloves:** put the smallest pinch possible on their dry food.

Keep all of this up as a routine so that you need not fear your pets. Also, notice how peppy and happy they become.

Go slowly so the pet can learn to eat all of it. To repeat:

- Week 1: parsley water.
- Week 2: parsley water and black walnut.
- Week 3: parsley water, black walnut, and wormwood.
- Week 4: parsley water, black walnut, wormwood, and cloves.

	Parsley Water	**Black Walnut Hull Tincture Dose**	**Wormwood Capsule Dose**	**Clove Capsule Dose (Size 0 or 00)**
Week	teaspoons on food	drops on food, cats twice per week, dogs daily	open capsule, put smallest pinch on food	open capsule, put smallest pinch on food
1	1 or more, based on size			
2	1 or more	1		
3	1 or more	1 or more, based on size	1	
4	1 or more	1 or more	1	1
5 and onward	1 or more	1 or more	1	1

Parasites Gone, Toxins Next

Healing is automatic when you clean up your body tissues. Killing parasites and bacterial and viral invaders is fundamental. Removing toxins which invite them into your organs is even more fundamental.

How do you know which toxins are responsible for your cancer? Unless you use a Syncrometer to test, you can not know. So learn to use a Syncrometer, it will save your life.

The only other alternative is to move to a safer environment. Go on vacation. Stay in hotels less than ten years old (so the plumbing hasn't started to corrode) but not in new construction (give the carpets time to outgas). Do not have a freon containing refrigerator in your room. Do not use the hair drier they supply. Do not let them spray your room with anything. Launder the sheets and towels yourself at a Laundromat with borax and/or washing soda. Don't hesitate to drink the municipal water. Get busy with your dental cleanup.

When you get better on vacation, let that be your inspiration to move from your home. All of the toxins come from a civilized lifestyle. Resolve to leave it behind. Select a warm climate where you can spend your time outdoors in the shade most of the day. Have no refrigerator, air conditioner, clothes dryer, hair dryer, new clothing, detergent. Check that the plumbing is relatively new and that no pesticide is being used. Make the Easy Lifestyle Improvements (page 397). Throw away non-essential health supplements (unless tested) and drugs. The risk is greater than the benefit.

Watch For Bacteria

In the later stages of cancer the tumors are more and more infected with the common bacteria *Salmonella*, *Shigella*, and *Staphylococcus aureus*. Killing parasites prepares a feast for these ubiquitous bacteria. Now, more than ever, must you stay off dairy products (except for boiled milk), do the Bowel Program, take Lugol’s. Remove *Staphs* by doing the dental cleanup (page 409). Don't delay.

Help Your Family, Too

If you had cancer, your whole family should be freed of intestinal fluke parasites to protect you! They may not be getting cancer (yet!) but your closeness puts you at risk. Kissing on the mouth could reinfect you. Request that family members zap themselves and take at least one 2 tsp. dose of Black Walnut Hull Tincture Extra Strength while living with you.

Pets, too, can be a source of cancer fluke stages (in their drool). They get propyl alcohol from their feed. Use unflavored, uncolored feed for them to minimize pollution from processing. Add 1 tbs. vitamin C powder to the top of a 5 lb. bag; it will stick to the pellets as they pour out. Zap your pet along with yourself by holding them and touching a bare spot such as nose or paw.

Never eat rare meats or fast food chicken. Parasite stages survive heat up to the boiling point. If you have taken a risk, zap yourself as soon as you get home to minimize the damage.

Cancer could be completely eliminated in the entire country if laws required testing for solvents in animal feeds and human food and products. Presently it is allowed in the United States Code of Federal Regulations (CFR) (see page 428)!

Another reason for propyl alcohol pollution (and other pollutants) in our food are the chemicals used by manufacturers to sterilize their food handling equipment.

> **21 CFR 178.1010 (4-1-94 Edition) Sanitizing solutions.**
>
> Sanitizing solutions may be safely used on food-processing equipment and utensils, and on other food-contact articles as specified in this section, within the following prescribed conditions:
>
> (a) Such sanitizing solutions are used, followed by adequate draining, before contact with food. **[Note rinsing or drying is not required!]**

> (b) The solutions consist of one of the following, to which may be added components general recognized as safe and components which are permitted by prior sanction or approval.
>
> [Now comes (1) through (43) permissible sterilizing solutions, including several with isopropyl alcohol, like:]
>
> (25) An aqueous solution containing elemental iodine (CAS Reg. No. 7553-56-2), potassium iodide (CAS Reg. No. 7681-11-0), and isopropanol (CAS Reg. No. 67-63-0). In addition to use on food processing equipment and utensils, this solution may be used on beverage containers, including milk containers and equipment and on food-contact surfaces in public eating places.
>
> [Then in paragraph (c)(19) the exact concentration of the iodine is specified. Nowhere is the concentration of the isopropanol specified. It can be as strong as desired.]

Fig. 41 U.S. regulations on sterilizing solutions.

Even if there were regulations governing removal of sanitizing solutions, the overwhelming truth is missed: that **nothing can ever be completely removed after it has been added**. Or perhaps the lawmakers didn't miss this fact. Perhaps they believed that small amounts–too small to measure with an ultraviolet spectrophotometer–could surely do no harm.

Meanwhile, protect yourself by avoiding propyl alcohol. And by observing 2 extra rules:

1. Never eat moldy foods. **2. Always take vitamin C.**

Aflatoxin

A common mold found on bread, nuts and fruit and in beer, apple cider vinegar and syrups, produces *aflatoxin*. This is what prevents you from detoxifying tiny bits of propyl alcohol that get into your body!

Buy bakery bread or bake your own. Roast nuts after washing in vitamin C water to destroy aflatoxin and then store in refrigerator. Keep moldy fruit out of the refrigerator where the spores can spread. Use only white distilled vinegar. Use honeys instead of syrups and even add vitamin C to them.

Vitamin C helps your body detoxify all the mold toxins I have tested, including aflatoxin.

Keep powdered vitamin C in a salt shaker. It belongs on the table with salt and pepper, and at the stove. Put it in everything possible, from cereal to soup to rice (1/8 tsp. is enough). Besides this take 1/8 tsp. powdered vitamin C with each meal (500 mg).

Developing cancer is a chain of events. This explains why it is a disease of aging. For ten years or more you poisoned your body with freon, fiberglass, asbestos, mercury, lead, copper, etc. You continually ate moldy food (chips, nuts, etc.) which was toxic to the liver. But the liver regenerated the pieces that died. Eventually a mold toxin prevents the liver from regenerating. Your aflatoxin-ed liver then lets propyl alcohol build up after using it. You use more and more propyl alcohol. Somewhere, over the years, you pick up the intestinal fluke in a hamburger or from a pet or person. The stage is set and cancer is inevitable.

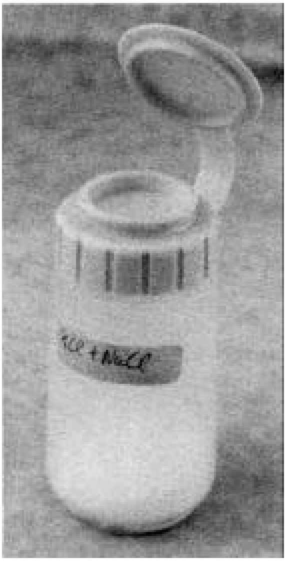

Fig. 42 Keep both salt and vitamin C powders in closable plastic shakers.

But pull out any one of the links in this chain and cancer is impossible. **Pull them all out for a return to good health**.

So cure yourself, prevent reinfection, heal the damage and go through life without this sword hanging over you.

Over 100 consecutive case histories of cured cancer victims are the subject of another book[16] along with more detailed instructions and suggestions.

---

[16] *The Cure For All Cancers*, by the same author, New Century Press 1993.

Curing HIV/AIDS

HIV and AIDS disease are very similar to cancer. That is why cancer is so often seen with HIV/AIDS. It is caused by the same parasite but the solvent is **benzene** instead of propyl alcohol.

HIV stands for
Human Immunodeficiency Virus

AIDS stands for
Acquired Immune Deficiency Syndrome

Many researchers believe HIV causes AIDS. I observe them both caused by parasites and solvents!

When the body can no longer detoxify benzene it soon may not be able to detoxify propyl alcohol. Food mold, at the base of the propyl alcohol problem, is also at the base of the benzene problem. Zearalenone, a mycotoxin I find in popcorn, corn chips, and brown rice specifically inhibits detoxification of benzene. None of us should be eating these foods untested.

Several common mold toxins inhibit the immune system, too, specifically those white blood cells that are supposed to eat and destroy viruses. Consequently, the HIV virus cannot be checked once it is introduced in the body.

Benzene goes to the bone marrow where T-cells are made, and to the thymus where T-cells are programmed, two big blows to the immune system.

Benzene, a most unthinkable pollutant, is widespread in extremely small amounts. But when **benzopyrenes** are added, the total is entirely too much for the liver and it begins to accumulate in your thymus.

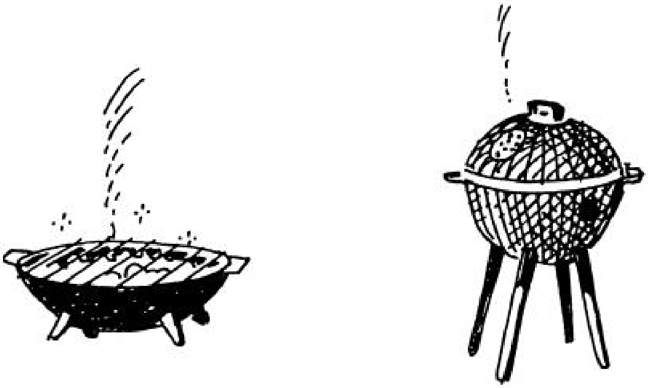

Fig. 43 Grilled food has benzopyrenes as do hot dogs and "smoke flavored" foods.

Benzopyrenes are made right in your food by direct flame-heat. Grilled food, smoked food, hot dogs and lunch meats with "smoke flavor" all have benzopyrenes—even toast has it. Food fried in a skillet does not (even if you burn it black!). As the thymus deteriorates from its benzene burden so does your immunity.

Without benzene in your body
you can't get AIDS.
Without the intestinal fluke in your body
you can't get HIV.

After killing the parasite and its stages the HIV virus will be gone in 2 hours, almost as quickly as the cancer marker, ortho-phospho-tyrosine. It is not necessary to zap the virus itself at 365 KHz.

But the AIDS is not gone. This depends on removing benzene from your body and keeping it out of your body. Keeping it out on alternate days or 5 days of the week will not suffice. It must be out consistently. Only then can your thymus and bone marrow recover.

As your thymus recovers your immunity returns. Immunity to all the diseases that lurk at very low level in our bodies. They are already in us. We have accumulated them in a lifetime. Mumps, measles, chicken pox, CMV, *Staphylococcus aureus*, *E. coli*, are all there. Until now they have been controlled–kept down–by an ever vigilant immune system. If this system fails you are an easy target for any and all of these invaders.

Method Of Treatment

As soon as you find you are HIV positive, don't panic. It was inevitable. It is also inevitable for many others with unexplainable risk factors. If you have a source of infection with the intestinal fluke and at the same time a source of benzene, HIV and AIDS are a logical result. You must stop the source of both.

Purchase or make a zapper or use somebody's frequency generator (434 KHz to 421 KHz giving 3 minutes to each KHz). Kill your flukes immediately. Take vitamin B2 immediately to help detoxify benzene (three 100 mg. tablets 3 times a day, 9 altogether). Zap daily until you feel completely well: no night sweats, no coughing, no symptoms of any kind.

Remember parasites are all around us. We pick them up daily. And your weakened immune system lets everything gain a foothold in your body. Avoiding benzene will restore your immunity. Use the benzene pollution list to guide you.

Benzene Polluted Products

THROW THESE OUT

Your health is worth more than the fortune you spent on them!

- **Flavored food**, yogurt, jello, candies, throat lozenges, store-bought cookies, cakes
- **Cooking oil** and shortening (use only olive oil, butter and lard)
- **Bottled water**, whether distilled, spring, mineral, or name brand. Bottled fruit juice.
- **Cold cereal**, including granola and health brands
- **Toothpaste**, including health brands
- **Chewing gum**
- **Ice cream** and frozen yogurt
- **Pills and capsules**. At least a third of all I test are polluted. This includes herbal extracts and prescription drugs. Test yours and switch brands until you find a safe one.
- **Vaseline products** (Noxzema™, Vick's™, Lip Therapy™), chap stick, hand cleaners
- **Vitamins** and other **health supplements**, unless tested.

- **Rice cakes**, even the plain ones
- **Personal lubricant**, including lubricated condoms
- **Baking soda and cornstarch** (see *Sources*)
- **Soaps, hand creams**, skin creams, moisturizers
- Flavored pet food, both for cats and dogs
- Bird food made into cakes

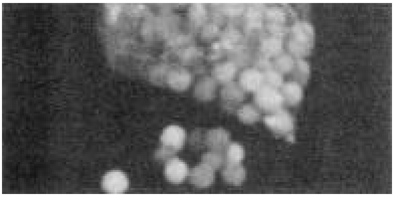

Fig. 44 Mexican made candy with no benzene.

CLOROX
simple green

Fig. 45 Pollutants are in unlikely places

It is impossible for me to have tested every batch of every food and product, but so many test positive you simply can not risk any of the foods and products on the list.

Learn to use the Syncrometer to do your own testing. While you are learning, observe the rules perfectly. There is no half way measure with benzene. Take an attitude of over-compliance, not "getting away" with as much as possible. Over-compliance can bring back radiant health, strength you never knew you had, a beautiful normal body again! You will be able to resume your plans for education, professional life, personal relationships, free of the sword hanging over you. You will not be infectious to anybody. Let over-compliance enthuse you.

Lay's
popcorn
popcorn
popcorn

Fig. 46 Every brand of popcorn and corn chips I tested had zearalenone contamination.

There are 2 extra rules that help eliminate benzene buildup:

1. Never eat moldy or grilled foods including popcorn, corn chips, grocery store breads, pasta, nuts, alcoholic beverages, toast, brown rice (white is OK).
2. Always take Vitamin B_2 (three 100 mg tablets three times a day) and vitamin C (1/8 tsp. with each meal).

Plan For The Future

After you are well again, you may wish to indulge in some philosophy. How did an obscure virus–a snail virus!–become a human virus? This is not so bizarre. After all, rabies virus comes to us from animals, and many encephalitis viruses come from mosquitoes. How did HIV spread so rapidly? How did a pollutant as dangerous as benzene get to be in our very food? Are other parasites getting set to spring on us? What must be done to protect ourselves and loved ones from future disasters? Would getting away from fossil fuel be a big answer? Would getting away from preoccupation with chemistry be a big answer? Would more disclosure of industrial practices be an answer? Should the government agencies responsible for food and product safety be depoliticized? Should public inspection of food manufacturing be a right? Should disclosure of foreign origins of food ingredients be mandatory? Would communicating with other cancer and AIDS survivors be useful? The computer age would make communication possible. Communication leads to answers. Answers might lead to new policies. Policies supported at the grass roots level can bring about change.

Over 50 case histories of cured HIV victims are the subject of another book[17] along with more detailed instructions and suggestions.

---

[17] *The Cure for HIV and AIDS*, by the same author, New Century Press, 1993.

Curing The Common Cold

Sometimes you can zap *Adenovirus*, the common cold virus, at 393 KHz for three minutes with a frequency generator and be rid of your cold, magically, in five minutes. But magic and luck are not really responsible for this. And most of the time the cold will reappear a few hours later.

Homeopathy, too, can immediately banish your cold symptoms. The reason is not mysterious either. But, this time, the cold reappears later in a different location. Choose the right homeopathic remedy for that, and you can chase it away again.

Tapeworm Stage or Mites

The fascinating story of how we really "catch" a cold kept me spellbound for a year. I was hot on the heels of *Adenovirus* wherever it might be in my body. Sometimes I saw it; sometimes I didn't. Sometimes I had cold symptoms; sometimes I didn't. Sometimes I could zap it; sometimes it did no good.

It is now apparent to me that *Adenovirus* isn't our (a human) virus at all! It belongs to other parasites. Parasites as varied as tapeworm stages and mites. Perhaps it belongs to many other parasites, as well. My evidence comes from a tapeworm stage, cysticercus of *Diphyllobothrium erinacea*, the mites *Sarcoptes* and *Dermatophagoides*, and our own colon bacteria, *E. coli*.

The tapeworm stage flies in the dust as eggs, you can trap these by setting out a pint jar with a little water in it. In three days' time you are likely to find its frequency near 487 KHz in your jar. You are also likely to find it on your kitchen sponge, since you wipe up dust each day. To test it, place it in a plastic bag, wet it thoroughly and search from 510 KHz downward, one KHz at a time. The various tapeworm stages emit between 510 and 440 KHz.

If you have a household pet, you will always be able to find a tapeworm stage in your sponge or in a dust sample you collect from the table or kitchen counter in the morning. Gather dust with a damp bit of paper towel, put it in a plastic bag. Then wash your hands or you may accidentally eat some.

This, of course, happens to every household member. Eating the dust off the tables, inhaling the dust, and eating off surfaces wiped by the kitchen sponge happens to everyone. And everyone "catches" colds. If you search for *Adenovirus*, though, in your dust sample, it isn't there!

Similarly, you can search for the mites in your house dust. Search near the frequencies given for them. There is a good chance you will have one that is not given, because the list is so incomplete. Name it after yourself. Compare notes with others; maybe it is common, maybe it's a rare one. Again, you will not find *Adenovirus* beeping its characteristic frequency out of your mite specimen. Why not? Possibly, it is too faint; it must multiply and create a loud chorus before you can hear it. But multiply it will, if given a chance, in you. You must, of course, first eat or inhale the dust.

Then the tape eggs hatch into the *cysticercus* stage, which promptly gets to the liver. Sometimes it gets to other organs, like the muscles, the spleen, the pancreas. Presumably the liver screened it out of the blood originally.

Soon you will zap them, wherever they are. If you are using a slide specimen of *cysticercus* you can locate it in your body. If you are only listening to its beeps, you can't. If you can do both, you may be able to see which organ allows the virus to replicate after it emerges. Maybe only the respiratory organs do; maybe they start to replicate in the organ where they emerge, such as the liver and then get to the respiratory tract. This is a fascinating avenue I have not yet explored.

Mites are inhaled or swallowed or both, just as tapeworm eggs are. They are on your kitchen sponge, and in any food or dishes that stand uncovered anywhere in the home. Never drink water from a glass that has "stood out" all day. After finding one, you will notice it beeping in you for several days. Then the beep disappears; presumably the mite is dead.

The tapeworm stage beeps may disappear in a few days, too, presumably dead. Except in cases of disease. Muscles that are diseased will take in the newcomer and allow it to survive adding to the parasites and pollutants already there! Evidently the immune power of such diseased locations is way down.

On the day that the mite stops beeping, the day it dies, *Adenovirus* appears!

Not many *Adenovirus* at first. You will need to search several times during the day to find it in your white blood cells. (And you have no symptoms, yet, either!) Is it a coincidence that *Adenovirus* appeared directly after a tapeworm stage or mite died?

You can find out by waiting until a time when you have a tapeworm stage or mite and no *Adenovirus*. Then kill your tapeworm stage or mite by zapping for seven minutes. Within minutes after that, the beeping of the tapeworm stage or mite is gone and *Adenovirus* can be heard, loud and clear at 393 KHz! And minutes later you may feel a stuffy nose, a slight congestion developing, a certain head feeling that is different. You are "catching" a cold!

Will you really get this cold? Will it become a full blown cold of the usual kind? After seeing this happen dozens of times after killing a mite or tapeworm stage I concluded that *Adenovirus* really belongs to them; the virus is scuttling its dying host like people jumping off a sinking ship into the ocean. Our bodies are the ocean for them. They too, immediately swim and search for a hospitable island. Our respiratory tract is such an island; perhaps other organs, too.

Yes, this "baby cold" will develop into a full blown cold if, but only if, you have a mold in you!

Molds and Colds

This is part two of the cold story. You may have *Adenoviruses* quietly slipping into your blood stream and tissues from a tapeworm stage or mite you inhaled, or *E. coli* bacteria that strayed into your tissues, and which is being slowly killed by your immune system. Your immune system can keep up with them quite easily provided you don't have a mold in you at the same time. The significance of the mold is that it lowers your immunity, specifically and generally.

Mold eluded, is health improved.

This has already been studied extensively for a number of food molds. There are a variety of ways that mold toxins lower immunity. Some simply kill white blood cells. Others seem to "bind and gag" them so they just can't go about eating viruses.

So with mold toxins present, *Adenovirus,* fleeing the dead tapeworm stage, mite, or *E. coli* is not gobbled up. It has time to get to its favorite organ and enter the cells there. It may get in your lungs if they're full of arsenic or formaldehyde, in your throat if it's full of mercury from your fillings, in your spinal cord if it's full of thallium. Sometimes you feel the viral attack, sometimes you don't.

When *E. coli* is the source of your *Adenovirus*, a question pops up. Why don't you have a perpetual cold, since these bacteria are always in your colon...and should be! As long as *E. coli* stays dutifully in your colon, no *Adenovirus* is seen. But as soon as any cross the colon wall to invade your body, your white blood cells pounce on them. After this, *Adenovirus* appears and again you are catching a cold. They may go to your internal organs where you don't feel them.

One place you do feel an attack is in your respiratory tract: lungs, bronchi, sinuses, nose, Eustachian tubes, inner ear, eyes or head. And the size of the attack depends on whether you recently ate moldy food.

Human food (in general, in the U.S.) is very, very moldy. We do not taste it because manufacturers have been using more and more flavorings in food. This covers up small amounts of mold or "off" flavor. Measures to reduce mold are not effective enough.

Bread is a good example. Calcium propionate is added to bread-stuff to inhibit molding. That's fine. But then the bread is encased in plastic to hold in moisture and keep it "fresh". The moisture acts to incubate mold spores and overwhelms the inhibitor. Vinegar is used instead of calcium propionate in some breads but, again, the plastic ruins its effectiveness.

Another good mold inhibitor is lime water. This is used in making tortillas. None of the old fashioned tortillas (made with just corn, water, lime) that I tested had any mold, even without propionate added! Other tortillas made of flour and calcium propionate frequently had molds.

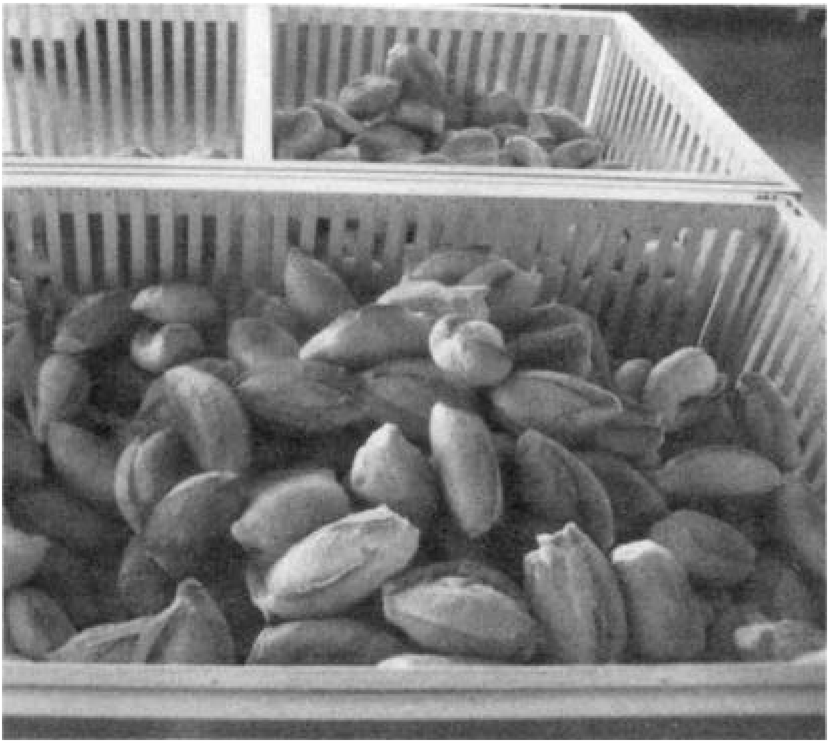

Fig. 47 This Mexican bread has no molds. It is made from white, unbleached flour and is not wrapped in plastic.

Bread is such a staple we must correct its mold problem immediately. The two likely sources for the mold spores are: in the flour to begin with, or just flying about the bakery and landing on the newly baked loaves. Bread flour in the grocery store is quite free of mold spores, so maybe it is the bakery that needs to change. Perhaps it is not possible to bake 24 hours a day in the same building, year after year, without bits of flour and moisture accumulating in the millions of tiny cracks and crevices that all buildings have and germinating mold. Yet bread from small neighborhood bakeries does not have mold!

Only buy fresh bread not wrapped in plastic.

As soon as you feel a cold coming, ask yourself: what did you eat recently that might have been moldy? Cold cereal, hot cereal, bread, crackers, cookies, rice, other grains, fresh fruit, store bought fruit juice, nuts, syrups, pasta, honey? This takes up a large part of any person's diet, even in a single meal. The answer is yes, no doubt you ate moldy food recently.

So you can't catch a cold directly from some sneezy companion, or *E. coli*, or tapeworm stages, or mites. You have to eat moldy food first. This lowers your immunity, allowing any *Adenovirus* to invade your weakest tissues. If that happens to be your respiratory system, you get classic symptoms. (Those same “cold” symptoms can be caused by bacteria, for which I have preliminary evidence of arriving in a similar fashion.)

As fascinating as this whole story is, the bottom line is: how can you stop a cold, in record time?

The Cure

Remember, zapping does not kill your cells. So anything hiding in them will not be harmed by the electric current either. Viruses live inside your cells while they reproduce. You can only kill the outsiders: those stuck in your cell gateways. The rest will be killed by your own cells in time. Not much time. Five or six hours at most. Your cells do it with mucous secretion, inflammation and other unknown ways. So zapping is an imperfect solution.

Using a frequency generator to electrocute *Adenovirus* (393 KHz for three minutes) is not effective either because you are not killing the mites and tapeworm stages at the same time, and the *Adenoviruses* are coming from them.

The best advice to stop a cold is to:

1. **Eat sterile food for 24 hours.** Follow the Mold Free Diet on page 365. Do not eat one questionable item.
2. Take **vitamin C** (10 grams or 2 tsp.), a **B50 complex** (2 tablets), and niacinamide (3 500 mg tablets) to help detoxify the mycotoxins already in you. See *Sources*. It will still take five or six hours for your white blood cells to recover their ability to capture viruses, for the "gag" to wear off.
3. **Zap** for seven minutes, killing all viruses, tapeworm stages and mites together. **Wait** twenty minutes to let viruses and bacteria in the dead larger parasites emerge. **Zap** a second time for seven minutes to kill those viruses and bacteria. **Wait** twenty minutes to let any viruses infecting the killed bacteria emerge. **Zap** a third time for seven minutes to kill the last viruses.
4. Now you need only wait for your tissues to decongest and stop making mucous, etc.
5. Immediately start the Bowel Program (page 546) in case yours is an *E. coli* cold. You can't, and wouldn't want to, kill all the bacteria in your bowel. Zapping kills the escapees, though, to give a bit of relief, and the Bowel Program stops the invasive *E. coli*.
6. Do additional zapping as time permits until the Bowel Program has stemmed the invasion.

In five or six hours your cold could be gone.

Of all these measures, stopping mold consumption is the most important. If you eat peanut butter now, your cold will return with a vengeance. If you eat cheese it will add *Salmonella* to your illness and you may develop a fever. Your own immune system is the most powerful cold killer, so stop handicapping it.

Test yourself for the presence of molds to see if you are accomplishing your goal. The various molds I have tested had these emitting frequencies: 77, 88, 100, 126, 133, 177, 181, 188, 232, 242, 277, 288, 295 KHz.

In about five hours, some relief will be felt. The time it takes depends on how much mold you ate. But if you stop immediately and eat only perfectly safe food, your illness will be over in the shortest time. Shorter than merely zapping.

Mold Free Diet

Stick to this for 24 hours, even if you feel better much sooner.

Before starting to cook sterilize your kitchen sponge (microwave it for three minutes), and wash hands.

Breakfast

- 1 or 2 eggs any style. The egg carton and egg exterior have *Salmonella* on them, so remove the eggs, replace the carton, wash the exterior of the eggs and then your hands again before cracking them. You don't want a *Salmonella* infection added to your cold.
- Hash brown potatoes
- Pancakes or waffles with artificial maple syrup made with brown sugar and vitamin C.
- 1 cup herb tea with vitamin C, cinnamon, or other spice.

- Water with a tsp. of vinegar and honey.

Lunch

- Soup without noodles, without rice or any grain product.
- Homemade biscuits or bakery bread with butter.
- Homemade pudding or custard, all ingredients well cooked.

Supper

- Well cooked fish or seafood (canned O.K., like sardines, salmon and tuna).
- Fresh green vegetables, in perfect condition.
- Cooked vegetables with olive oil and salt (pure, see *Sources*).
- Canned beans, any variety.
- Baked or boiled potato (don't eat skin) with olive oil and salt as described on page 310.
- Hot water with whipping cream and cinnamon.
- A perfectly unblemished banana.
- Water with lemon.

You can “mix and match” these safe foods. If you get a hefty dose of mold at the outset of your cold, the toxicity lasts quite a long time. Repeat the diet the next day and the next until you are well. In animal experiments reported by scientists, toxicity from mold usually lasted three weeks. Sometimes the real damage was only seen after three weeks! Keep up your vigilance. When you decide to take some risks, make sure vitamin C has been added to the new food and mixed with it thoroughly.

Fig. 48 Goods baked in a panaderia (Mexican style bakery) never tested positive to molds.

Prevention

What is the best way to prevent a cold? Stop eating moldy food. We were never meant to eat rotting, spoiling food. We were meant to be repelled by the bad odor, the bad appearance, the bad taste. Our parents were supposed to teach us in childhood to distinguish between good and bad food. This system no longer works. By mixing spoiled food with good food, by adding flavors and changing the food consistency, parents can't distinguish substandard foods, nor teach their children to.

We rely on government agency assurances, like beef grades, expiration dates, approved food colors and additives. And when they fail? We land in a debacle such as the present one, where large segments of society are ill with uncontrollable behavior (called crime), suffer from hormone imbalances and sexual disturbances, are sidelined by chronic fatigue and new illness. But the greatest social disaster is lowered immunity—AIDS.

We cannot individually control the destiny of our species. But we can get ourselves out of the path of the falling human tide. We can dissent. We can say **NO** to rotten moldy food.

What about exposure to a cold? Won't you catch one if you're in a roomful of coughing, sneezing people? Or you shake hands with a coughing, sneezing person? Or you drink from the same glass as a person with a cold? Or your child plays with the same toys as sick children?

Yes, you will catch the virus, but you won't get sick. If you are tracking *Adenovirus* using the electronic techniques in this book, you will see that it infects you immediately after eating coughed-on food. It stays in you about twenty minutes. Then it disappears, evidently eaten up by your white blood cells, **provided there is no mold toxin in you**. But if you do have a mold toxin in you, the virus spreads, multiplies and gives you a cold! I have only followed *Adenovirus* behavior. There are numerous other "cold bugs" left for you to track. Each illness in your family could provide you with a specimen to research.

This throws light on self infection, too. When we cough into our hands and then eat with them, we self-infect. But it only makes us sicker if we harbor molds.

Curing Symptoms

Cold symptoms alone can be cured much faster than the cold. Taking drugs proves that. Most of the drugs on the market can cure some symptoms in less than one hour. But not without a price, such as a harmful side effect.

Homeopathy can cure cold symptoms too, without side effects. There are three or four favorite homeopathic remedies for colds and eight or nine less common ones. To use them you read the symptoms listed and take the remedy with the closest match.

Homeopathic Remedy	**For These Symptoms**
Aconitum	early cold with fever, headache, hoarse cough
Allium	clear runny nose with burning of lips or eyes
Arsenicum	sneezing cold, frontal headache, tickling cough
Belladonna	high fever cold with flushed face, throbbing head
Kali bi	thick post nasal drip, colored discharge, sinus headache
Spongia	croupy cough

Fig. 49 Starter set of homeopathic remedies.

There are lots more remedies with fascinating symptoms to try to match with your own. Books suggest that you start with a 6X or a 12X remedy, but success is more certain with 30X. Use three remedies simultaneously. If you get total relief at some point, stop. If the cold has moved or won't budge, try another set of three after an hour. Don't take them within 15 minutes of food or anything else. Homeopathy makes very interesting reading.

Homeopaths say they stimulate the immune system specifically. My results show they do much more. They go right to the gateways of your cells and evict the tiny parasite, bacteria or virus stuck to the latch and trying to get in. Your immune system would be able to gobble them all up if they weren't gagged by the food mold you ate.

Different homeopathic remedies go to different tissues, so you can only clear one tissue at a time. If you plan on trying this for yourself, order the set of cold remedies listed above (see *Sources*). The homeopathic method would be a beautiful cold cure if it weren't for the mold intrusion.

Herbs, too, can cure symptoms rather quickly. **Elderberry tea** mixed with **peppermint** is cited in herbal textbooks and it could probably do a lot if it weren't for the mold immunosuppression. If you plan on trying these start with a set of **thyme**, **fenugreek**, **sage** (for throat). Since both herbs and homeopathic remedies work on the principle of ejection, they could eject each other. Maybe the last one to arrive takes over. This is an exciting field for you to explore.

Ultimately, the length of time your own white blood cells are bound and gagged decides how soon you are really cured of your cold. Remember taking vitamin C (10 grams or more) helps detoxify the molds. If you find a recipe that works for everybody in less than five hours, be sure to let everybody know.

True Origins Of Viruses

Your body can eliminate any virus in a short time, such as hours or days. It can even keep up with a steady stream that is coming from tapeworm stages!

Let us apply these new insights to several diseases we are familiar with. Let us speculate what could really be happening.

Epstein Barre Virus (EBV) is a mysterious disease because in spite of building up antibodies, it attacks repeatedly. I only see EBV if the person also has *Eurytrema* (pancreatic fluke) in the pancreas. When the fluke is gone, I no longer find EBV in the white blood cells. This suggests that the virus comes from the fluke. As each fluke dies and is finally removed, the body's white blood cells can catch up with the viruses and you begin to feel better again, which can be as quick as one day. But reinfection with an EBV-carrying *Eurytrema* (plus a wood alcohol containing beverage to allow the fluke to go to the pancreas) can spark the next recurrence.

Shingles is a recurrence of chicken pox. I always find *Ascaris* in persons with shingles! Unfortunately, killing the *Ascaris* does not cure shingles. *Herpes Zoster* (the shingles/chicken pox virus) is known to hide in nerve cells. Perhaps *Ascaris* facilitates it's release, or simply suppresses the immune system in a way that allows it to suddenly multiply.

Polio was once a scourge. At that time, we can theorize that a new large parasite was making its appearance. Was a new animal association taking place in the early 1900's? Life was indeed becoming more urban with horse manure on all streets. Owning a dog for a house pet was becoming an acceptable lifestyle. Could the tapeworms of these animals give us a tapeworm stage that hosts polio virus? Many polio sufferers also had migraines. These are caused by *Strongyloides*, a horse threadworm. Does *Strongyloides* host polio virus?

The **HIV** virus infects us during the time the intestinal fluke, *Fasciolopsis buskii*, is being hosted. Persons test negative to HIV shortly after the fluke and its stages are eliminated.

Coxsackie viruses give us some of our encephalitises. I never see *Coxsackie* viruses without the bacteria, *Bacteroides* *fragilis*, and I never see *Bacteroides fragilis* without *Ascaris* (a roundworm.) I conclude that one lives inside the other!

We may be deriving viruses from all the roundworms, flukes, tapeworms and bacteria that infect us! It would be a fascinating study, simply to examine each of these parasites singly, searching for their viruses with an electron microscope. They could also be searched for using immunological methods.

Fortunately, your health improvement does not need to wait on such studies. Your electronic technique can detect them in your body long before you are made ill by them. You don't need to know their hosts in order to stop hosting them yourself.

It is a time of great change for this planet as pollution spreads from pole to pole. The growth of industrial activity, mining, chemical manufacturing, the food "industry", and personal habits like smoking have spread new chemicals to every corner of the globe. The element *polonium*, which is radioactive and in tobacco smoke, is harmful to human lungs, but may not be harmful to a small lung parasite, like *Pneumocystis carnii*.

Benzene, which is a solvent and extremely harmful to humans, may not be harmful to fluke parasites living within us. Propyl alcohol facilitates the intestinal fluke but is toxic to us.

Parasites are doing abnormal things. Is this because of pollution?

The tables are gradually being turned against us in favor of our parasites and pathogens. Such large changes are called evolutionary. Is the human species doomed, or will some of us "adapt". Will some of us survive to pass on our "better" genes to a new population of cancer-resistant, AIDS-resistant, Alzheimer's-resistant, etc.-resistant humans? How many of us will die trying?

The common cold should not be so common!

Body Wisdom and Why Bad Food Tastes Good

Don't let yourself eat junk food just because your body craves something in it. Try to figure out what it is your body craves. Could it be fat? Could it be salt? Could it be starch?

Salt

If it's salt, you might also love popcorn and other salty food. This implicates the adrenal glands. Maybe they're letting too much salt (sodium chloride) leave the body through the kidneys. Maybe they're letting too much potassium chloride through, too. A diuretic pill could certainly have a similar effect.

Help the adrenal glands do their job of regulating sodium and potassium chloride by cleaning them up. Let salt-hunger be your signal to do a kidney cleanse (page 549). This will clean adrenals too. Even a slight drop in sodium and potassium chloride in the blood (body fluids) can make you too fatigued to tie your own shoelaces.

Remember, when your body craves potato chips, it craves something in the potato chips. If you let yourself eat highly salted food while you're giving the adrenals a clean up, at least add potassium chloride to your diet. Make yourself a mixture of equal parts of sodium chloride and potassium chloride. Part of salt hunger is actually potassium hunger. Let your body (your taste) decide on the amount of potassium chloride to add. Maybe one part potassium chloride to two or three parts sodium chloride is a better mixture for you. After mixing, store it in the original containers (re-label them) to prevent caking. If you put it in an ordinary salt shaker, it will cake soon. Use a shaker with a lid that closes.

Fats

Maybe you like French fries because of the fat. If you deprive yourself of the “good” normal greases that come from plant or animal sources which would ordinarily make up 25% of your calories, of course you'll crave grease. But what a bad trade it is. Now you are getting lab-made (hydrogenated) grease with a non-biological structure, and loaded with the carcinogen nickel.

So if you're body tells you that you need grease, go back to olive oil, butter, cheese (baked only), lard, avocados, nuts and nut butters (homemade only) and seeds. Humankind has been eating these natural fats long before cholesterol was vilified. The key to cholesterol control is not fat avoidance, but a liver cleanse!

Starch

If switching to natural greases doesn't satisfy your “fat-tooth”, maybe its the potato in the French fries that your body craves. Plain, pure starch. Do you also love bread and pasta (more pure starch though very inferior to potatoes)? Pure starch is very easy to digest and has a large adsorptive capability for toxins. In fact, if any family member should accidentally eat something poisonous, drinking cornstarch will quickly mop it up and keep it stuck so it can't enter your tissues. (This doesn't work for all poisons.) By craving pure starches, your body could be telling you about a need to improve your digestion (liver disorders) or to eat and breathe less toxic things.

Maybe a stomach-full of baby *Ascaris* is telling you to eat only food that doesn't need a lot of acid: "just potatoes, bread and pasta, please, and skip the sauce.” *Ascaris* inhibits acid production by the stomach. This can result in an aversion to meat.

It doesn't take much acid to digest pure starch and get it on its way out of the stomach. And out of the stomach means relief: relief of the pressure on the diaphragm and liver, heartburn, that too-full feeling, and other digestive disturbances.

Sugar

Your body runs on sugar. If you are short on sugar it will turn fat into sugar. If you are short on both, it will turn your muscles into sugar. However eating more sugar doesn't cure the craving. You have to find out why you are so short, in spite of eating it.

The first thing to try is 1 mg chromium (five 200 mcg tablets, see *Sources*) per day. If you still crave sugar after a week the problem is something else. Perhaps you have pancreatic flukes upsetting your sugar regulation. Kill them and go off commercial beverages that may contain wood alcohol. Sugar regulation is very complex, but these two approaches help most of the time.

Dislikes

Respect your body's opinion when it says, “No, I don't want to eat that.” Our education about nutritive value of food may be sound but there are other facts to consider. We should take a lesson from nursing babies: when they refuse to nurse, there is something unpalatable in the mother's milk. Usually the mother has eaten onions or members of the cabbage family. The baby tries it once, and learns to reject it immediately. The baby's liver, in its wisdom, does not want the baby to eat what it can't properly digest. The mother may feel: “Now, this breast milk is good for you and drink it you must, or you shall go hungry.” Unfortunately, this works for 2-year-olds and up. They are forced to eat carrots, peas, and other vegetables; vegetables that taste terrible, (modern agriculture has ruined the flavor). They alone taste the bitterness of PIT, a cyanide-related chemical, and very difficult for the liver to metabolize. Broccoli and onions may burn the tongue with its sulfur-containing acids. Green beans, onions, garlic, eggplant, all have unique chemicals in them. If you or your child are not ready to eat them, avoid them carefully, so you don't get a surprise dose of the toxic chemical.

The more mold a child eats, inadvertently, in peanut butter, bread, potato chips, syrups, the less capable the liver is of detoxifying foods. This will certainly increase the “pickiness” of a child's appetite. If your child has too many foods on her or his personal “off list”, let this signal you to improve liver function. Stop the barrage of chemicals that comes with cold cereals, canned soup, grocery bread, instant cheese dishes, artificially flavored gelatin, canned whipped cream, fancy yogurts and cookies or chips. Move to a simpler diet, cooked cereal with honey, cinnamon and whipping cream (only 4 ingredients), milk (boiled), bakery bread, canned tuna or salmon, plain cooked or fried potatoes with butter, and slices of raw vegetables and fruit without any sauces, except honey or homemade tomato sauce, to dip into.

It is frustrating to cook “a fine meal” for the family and find everybody likes it except Ms. Picky. The good news is that she can usually think of something she would rather eat. If it's nutritious, be thankful. If it's not say No.

Adults should hide their junk food, including everything off limits to children. Don't “hide” your junk food in the refrigerator and lower level cupboards! Treat yourself as well as your child. If a food tastes bad, don't eat it. If you crave it, try to understand the message.

Outwit The Cravings

Here are some examples. Suppose you crave these items:

- **Pickles**. They supply vinegar and are often loved by persons with little acid in their stomachs or a lot of yeast (vinegar is a yeast inhibitor). Start drinking water with lemon juice or vinegar and honey.
- **Bacon**. The fat soothes the stomach and slows down digestion. Switch to butter and cream, with meals.
- **Sugar coated cereals**. Loved by persons with disturbed sugar regulation. Kill parasites, avoid wood alcohol, use chromium tablets and a lot of cinnamon.
- **Crunchy munchies**. Your jaw and teeth want some work to do. Try salads, an apple, raw sunflower seeds (beware of moldy seeds, nuts and dried fruit).
- **Ice cream**. Ice cold food stimulates the thyroid; loved by low thyroid persons. Clean up the thyroid by doing dental work and liver cleanses.
- **Caffeine**-laced beverages. Stimulate many body tissues, raise blood pressure. Loved by low energy people. Do a general body and environment cleanup. (There are people who say coffee puts them to sleep. Insomnia has better solutions than caffeine, though.)
- **Candy**. The more you eat the more you crave because chromium is being used up as you eat it and yet it is necessary to utilize more sugar. Give yourself chromium (GTF) tablets totaling 1 mg. (1,000 mcg.) a day and watch your sugar craving shrink.
- **Pretzels**. You want salt plus crunch.

- **Potato chips**. You want salt, grease, starch and crunch. No wonder they are so popular!

If your body still has its wisdom, or most of it, why can't it detect the mold in peanuts, crackers and bread for us? There is a very sound reason. Our food manufacturers have gone to great lengths to fool our native senses. Salt and sugar, roasting and flavoring, do most of it.

Your body is accustomed, natively, to interpret sugar, salt, and flavors as “good, good, good.” Of course, the mold is “bad, bad, bad.” But when you mix them, what is your body to read? The “goods” always win; manufacturers don't stop until they do. More flavorings are added. The result is that you can be eating rotten moldy food without knowing it.

Food that is predominantly concocted can't be interpreted by your body wisdom. You must use your second-best ally, your intelligence.

How would you interpret these situations, taken from real life:

- An elderly person can't stand butter, wants and enjoys margarine.
- A child prefers canned spaghetti for the real thing.
- A child wants to eat only sweets, everything else must be coaxed down.
- A young man needs “his” beer to enjoy a cook-out.
- A young man with serious mental illness drinks half a gallon of Mellow Yellow™ a day.
- A child wants ketchup on everything.
- A pregnant woman puts herself on a pickled pigs feet and white bread diet.

Back To Normal

Body wisdom was not meant to substitute for parental teaching. They must both be there. But when there is conflict, trust your wisdom. Will you ever get your primitive body wisdom back and enjoy vegetables, fruit, simple styles of cooking and baking them? Yes, to a considerable extent. You'll hold your nose at all seeds and nuts and most flour (they smell so rancid). You'll back away from cookies and cakes, especially icing (they taste like you imagine shortening tastes). You'll retreat from deep fat fried foods: the dripping grease will just undo the appetite. You'll say “No thanks” to canned food and metal cutlery (you can taste the metal now). You'd have to force yourself to eat grocery bread (it's so doughy and sour smelling). Wheat germ smells terrible (rancidity). Even vitamin tablets may smell awful.

Has life been ruined, now that plain potatoes and butter taste good? Your body wisdom has returned. You are the true gourmet. It is as different as corn on the cob is from the canned variety. Perhaps you are “spoiled” now by eating un-rotten food. Accept the charge with humor and dignity—as long as you're not expected to eat any more spoiled, moldy food.

Toxic Food

Grilled food develops *benzopyrenes* in it that are very toxic. In an age of lowered immunity, it makes little sense to deliberately poison the food with benzopyrenes. Especially for children, who will be faced with new viruses and parasites in their lifetimes. Will they be able to overcome them or succumb at middle age? Only the strength of their immune system decides this.

Benzopyrenes must be detoxified using the liver's valuable benzene-detoxification system. With so many benzene-polluted items, there is hardly enough detoxification capability to get it all taken care of. NAD enzymes (the N stands for niacin) come into play too. These are essential for alcohol detoxification. If you have consumed alcohol, like a can of beer, NAD enzymes must be shared between the alcohol in the beer and the benzene in the beer. It takes longer to detoxify both the benzene and alcohol. The time delay is a time of lowered immunity and facilitates a growth spurt for parasites and pathogens.

Foods that are raised to very high temperatures, made possible with a microwave oven, produce benzopyrenes. Ordinary bread-toasters can do this too! Old fashioned toasters had a layer of metal separating the bread from the wires. Toasting bread in a frying pan or a stove top surface protects it too. But your stove grill, whether electric or flame, will produce benzopyrenes in your food unless there is a separating wall between them.

It does not matter what kind of fuel is used, the benzopyrenes develop due to lack of shielding between the food and heat source. A metal wall between them absorbs some of the heat.

Do not burn your food in a microwave oven. Since the temperature may go higher than your regular oven, you can produce benzopyrenes. Your regular oven is thermostated so that it cannot go higher than it states. Baking and browning is never done above 425°F. This is your safety feature. If anything in your microwave has turned dark brown or black or has melted plastic, throw it out!

Supplements that help your body to detoxify the benzopyrenes are:

1. Niacin or niacinamide. These are NAD-builders. Take 50 mg with each meal.
2. Vitamin B_2 (riboflavin). Take 50 mg with each meal.

For unpolluted vitamins see *Sources*.

It would be wise to teach children the habits that maximize their immune strength. Avoiding food toxins that are specifically immune-lowering is most important. Besides the benzopyrenes, certain mold toxins and solvents do this and are found in foods.

Moldy Food

Everything that's animal or vegetable can get moldy. While living things are alive, the mold attackers can be held at bay. As soon as they are dead, molding begins. First it molds; then bacterial action sets in. This is what makes things biodegradable. It is a precious phenomenon. It does away with filth—in an exquisite manner. Without mold and decay the streets of New York would still be full of horse manure from the days of the horse and buggy and our lakes too full of dead fish to swim in.

Every grain has its molds; every fruit has its molds; tea and coffee plants have their molds; as do all herbs, and vegetables. Nuts have their molds; nuts grown in the ground (peanuts) are especially moldy because the earth is so full of mold spores. But the wind carries these spores high up into trees, and even up to the stratosphere. Molds are not very choosy. They have their preference for certain plants and conditions. But the same molds can grow on many plants. This is why aflatoxin, for instance, is found not just in your cereal, bread and pasta but in nuts, maple syrup, orange juice, vinegar, wine, etc. Where is it not? It is not in dairy products or fresh fruit and vegetables, provided you wash the outside. It is not in meat, eggs, and fish. It is not in water.

Although I find aflatoxin in commercial bread, I do not find it in carefully screened wheat that has had its discolored, shriveled seeds removed before using it for making bread, cereals and noodles. It is not in baked goods bought at bakeries, left open to air. Evidently the system of wrapping baked goods in plastic keeps moisture trapped and starts the molding process. In spite of adding mold inhibitors, American bread-stuff is far inferior to Mexican baked goods in which I do not find aflatoxin!

Here is some good news for cooks: if you bake it yourself, adding a bit of vitamin C to the dough, your breads will be mold free for an extended period (and rise higher).

Aflatoxin

What is so important about molds? Some of them produce very, very toxic chemicals wherever they grow. They produce some of the most toxic chemicals known to exist. Aflatoxin is one of these. My tests show it is always present in cancer patients; in other words it has built up due to the body's inability to detoxify it in a reasonable time. A great deal of research has been done on aflatoxin. Any library would have more information.

Aflatoxin reaches the liver and simply kills portions of it. After a hefty dose the liver is weakened for a long time—possibly years. Hepatitis and cirrhosis cases always reveal aflatoxin. The liver fights hard to detoxify aflatoxin and manage its own survival. It manages for 2 to 3 weeks; then a portion of it succumbs. So the toxic effects of a dose of aflatoxin aren't even noticeable for several weeks! And without a taste or smell to guide you, how would you know to stop eating the moldy peanut butter or spaghetti? The answer is:

1. make and bake things for yourself
2. test the things you dearly love but can't make
3. treat things that are treatable for molds
4. throw the rest out of your diet

Treatments mentioned in the industrial research journals are hydrogen peroxide, strong alkali such as lime-water, metabisulfite (a common reducing agent) and high heat. I have tried heat and vitamin C, which is also a reducing agent.

Just heating a food to the boiling point does not kill the molds. Boiling for many minutes at a higher temperature or baking does kill them (but not ergot, another mold) and also destroys aflatoxin they produced and left in the food. For foods you can't heat that high, for example nuts that are already roasted, or vinegar, vitamin C comes to the rescue. I suppose it acts a lot like the bisulfite; chemically destroying the mold toxin molecules.

Eradicating Aflatoxin

Simply sprinkling vitamin C over **roasted nuts** is not effective because the molds have penetrated the surface. Rinse the nuts in water first (a lot of mold is removed in this simple way). Cover the nuts with water, add about ¼ tsp. vitamin C powder (for a pint of nuts) and mix. Let stand for 5 minutes. The water penetrates the nuts, taking the vitamin C with it and detoxifies them. Pour off the water and dry the nuts in the oven at low heat. (Don't burn them or you will make benzopyrenes.)

Rice and **pastas** can be demolded partly by cooking and partly by adding vitamin C before or after cooking. There is no need to add so much it affects the flavor. Brown rice is especially moldy.

Vinegars can simply have vitamin C added and placed in the refrigerator.

Honey can be warmed and treated the same way (¼ tsp. per pint).

Bread cannot be salvaged. Switch to bakery breads or homemade. Use it up in a few days, left in its paper bag. Or slice and place in a plastic bag in the freezer.

Since all foods have both their own and others' molds, there must be thousands of molds. Very many have been studied besides aflatoxin producers.

Zearalenone

> Zearalenone, an anabolic and uterotrophic metabolite, is frequently found in commercial cereal grains and in processed foods and feeds, and is often reported as causative agent of naturally occurring hyperestrogenism and infertility in swine, poultry and cattle.[18]

What this means is, in animals, “zear” looks likes extra estrogen to the body. Does it affect humans the same way? Are high estrogen levels a problem for us? I find nearly every breast cancer case shows a too-high estrogen level for years before the cancer is found! It starts females maturing too early, too. It could cause PMS, ovarian cysts and infertility. Not everybody gets all of these effects. And what is the effect on men and boys of eating an estrogen-like mycotoxin in their daily diet? This female hormone could have a drastic effect on the maturing process even in small amounts.

Zearalenone ("zear") and aflatoxin both have immune lowering effects. Zearalenone can induce thymic atrophy and macrophage activation.[19] If you have low immunity (low T-cells, low white blood cell count, and so forth), immediately go off moldy food suspects.

---

[18] Bottalico, A., Lerario, P., and Visconti, A., Production of Zearalenone, Trichothecenes and Moniliformin By Fusarium Species From Cereals, In Italy. From *Toxigenic Fungi, Vol 7*, edited by H. Kurata and Y. Ueno, copublished by Kodansha Ltd, Tokyo and Elsevier Science Publishers B.V., Amsterdam, 1984, page 199.

[19] Luster, M.I., Boorman, GA., Korach, K.S., Dieter, M.P., and Hong, L. 1984. Myelotoxicity toxicity resulting from exogenous estrogens evidence for bimodal mechanism of action. Int. J. Immunopharmacol. 6:287-297.

"Zear" is the mycotoxin that prevents you from detoxifying benzene. Every AIDS sufferer I see has a crippled ability to detoxify benzene; they also have zear!

The main zear sources I have found so far are popcorn, corn chips, and brown rice. But it was absent in fresh corn, canned corn, corn tortillas, and white rice, making me wonder how it gets in our processed corn products.

Sterigmatocystin

Sterigmatocystin ("sterig") is plentiful in **pasta**. Emphasize baked pasta dishes, not boiled. This raises the temperature much higher than boiling. Better yet, make your own pasta with a pasta maker. U.S. bread flour is quite free of mold; the mold in our pastas must come from using inferior quality flour. Always add vitamin C to pasta before or after cooking.

Fig. 50 All U.S. brands of pasta I tested had mold, including health food brands like the one shown here (left). No Mexican brands of pasta, like the one pictured (right), had any molds.

Ergot

A food mold that causes strange feelings and behavior is *ergot*. Although laws regulate the amount of ergot allowed in foods,[20] this is not enough protection. Ergotoxins, for example LSD, are active in extremely minute (less than a microgram, about one thousandth of a fly speck) quantities. They are not destroyed by heat and are especially toxic to children. I found traces in cereals, whole grain breads, wines, and honey. It can be detoxified by adding vitamin C but takes longer; about 10 minutes. Detoxify all your **honey** as soon as it arrives in your house. Warm it slightly and add vitamin C (1/8 tsp. per cup). Stir with wood or plastic.

Ergot toxicity could explain "Jekyll and Hyde" behavior in children, commonly attributed to "allergies". In fact, the mechanism, inability by the liver to keep up with detoxification, fits well into the "allergic" concept. If your child has undesirable behavior, try going off the moldy food suspects for three weeks (cold cereals, nuts and nut butters, store bought breads and baked goods, syrups). Substitute cooked cereals, bakery breads, potatoes, and honeys. Add vitamin C to honey, pasta and cooked cereals. Pancakes and waffles made from scratch would be O.K.

Combining alcohol with ergot is more toxic than either is alone. Alcohol seems to drive the toxin deeper into your tissues. I have found ergot and aflatoxin in beer and wine! Perhaps some of the bizarre behavior and speech of intoxication is really due to the mold-alcohol combination. By delaying alcohol detoxification, the mold could even be responsible for deaths "due to" alcoholism. It would be safer to brew your own alcoholic beverages. Start with pristine fruit. Or at least add vitamin C (1/8 tsp. per cup) to the store bought container you are consuming.

Older children and adults are quite susceptible to ergot too. If bizarre behavior shows up, such as saying mean and cruel things, expressing unusual, irrational thoughts, feeling emotionless or unreal, try the same diet changes, but put alcoholic beverages, soy sauces and other sauces, and other grain derived foods on the "off" list. Try this diet on yourself if you have a temper or crying spells or frequent colds! Ergot can make you super religious, hearing voices of command or threat. Ergot also causes seizures!

---

[20] Canada allows one ergot grain per 300 grains of #3 or #4 wheat.

1919
APPLE JACKS
APPLE JACKS
APPLE JACKS
APPLE JACKS

Fig. 51 All cold cereals I tested were full of mold toxins (besides solvents), health food varieties were worst.

Cytochalasin B

Cytochalasin B ("cyto B") is another immune lowering fungus. I find it mostly in pasta. It stops cells from dividing. Dead portions of the liver cannot regenerate as they otherwise would after a toxic encounter!

Kojic Acid

Kojic acid is a mycotoxin that appears to be responsible for wood alcohol build up. In other words, the toxic effect prevents you from detoxifying wood alcohol. This leads to pancreas damage, invites pancreatic fluke infestation, and typically results in diabetes. I find it in potatoes; don't eat potato skins. If you are a potato lover fix your own so you can peel them and remove any gray parts. I have also found it in regular coffee.

T-2 Toxin

T-2 toxin is a mold I have found in all cases of high blood pressure and kidney disease. It is present on dried peas and beans but it can be detoxified in 5 minutes by adding vitamin C to the water they are soaked in. Remember to throw away imperfect ones, first.

Sorghum Molds

Sorghum and millet carry these. Don't buy sorghum syrup. Rinse millet in vitamin C water before cooking, or add vitamin C to the cooking water.

These mold toxins cause hemorrhaging, appetite loss, and inability to swallow. Elderly people are more easily poisoned than others; their hemorrhages show up as strokes and purple blotches on the skin.

Patulin

is the major fruit mold toxin. It is present in most common fruits if they are bruised. It is particularly hazardous since the mold that produces it can actually grow in your intestine in patches. At these locations, bowel bacteria, *E. coli* and *Shigella*, can climb through the colon wall to invade you. These bacteria are then free to spread to regions of injury and tumors.

If you have cancer or bowel disease go off fresh fruit (bananas and lemons are OK) for a few weeks. Then choose your fruit meticulously. Peel everything so you can see and avoid every bruise. Also take a 2 tsp. dose of Black Walnut Hull Tincture Extra Strength. This kills these bowel fungi. But you can reinfect with a single soft grape.

Mold Avoidance

We should be much more critical of our food.

Crackers are notoriously moldy. Never let your child eat crackers. Make crisp things in the oven from left over bakery goods. Just sprinkle with cinnamon.

Dried fruits are very moldy. Soak them in vitamin C water. Rinse and bake to dry again. Then store in the refrigerator or freezer. When fresh fruit gets overripe, don't quickly bake it or preserve it. It's too late.

Peanut butter (store bought) and other nut butters can't be detoxified by adding vitamin C due to the mixing problem, even if you stir it in thoroughly. Make your own. Making your own peanut butter is a great adventure (see *Recipes*). Mix it with home made preserves, honey, marmalade, not very homogeneously so the bright colors and individual flavors stand out in contrast. Having three or four such spreads in the refrigerator will give your children the right perspective on food—homemade is better. Store bought jams are sweeter and brighter in color but strangely low in flavor and often indistinguishable from each other. Let your children eat the polluted foods that friends and restaurants serve (but not rare-cooked meats) so they can experience the difference. Their livers are strong enough to detoxify occasional small amounts.

Tea is quite moldy if purchased in bags. Although I used to recommend single herb teas (tea mixtures have solvents), I can now only recommend single herb teas from fresh sources in bulk (see *Sources*). This also gets you away from the benzalkonium chloride and possibly other antiseptics in the bag itself. When you get them, store them in their original double plastic bag. These herbs are so fresh, you'll only need half as much to make a cup of tea. Use a bamboo strainer (non metal). Bake the strainer occasionally or put through the dishwasher to keep it sterile.

Fig. 52 Packaged herb tea is moldy and polluted with solvents. Get yours in bulk from an herb company.

It comes as a surprise that pure, genuine **maple syrup** has the deadly aflatoxin and other molds. You can often see mold yourself, as a thin scum on the surface or an opaque spot on the inside of the glass after the syrup has stood some time, even in the refrigerator. Some mold spores were in it to begin with. Others flew in. After some time they grew enough to be visible. In my testing, aflatoxin can be cleared with vitamin C but sterig and others need to be treated with a high temperature as well. Fortunately, this is easy to do with a syrup. Heat to near boiling while in the original jar with the lid removed. Keep refrigerated afterwards.

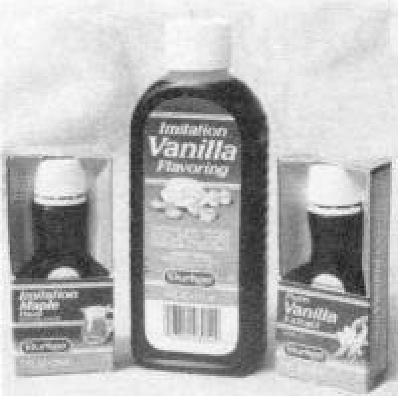

Fig. 53 Three safe flavorings.

Artificial maple flavor did not have benzene, propyl alcohol or wood alcohol, nor molds. **Turbinado sugar** had none of these contaminants either. **Brown sugar** had sorghum mold. **White sugar** had propyl alcohol pollution. You can make your own syrup, safely, with artificial flavor and turbinado sugar. Of course, you'll be missing the taste and nutritional minerals provided by the natural maple product but in a contest between nutritional value and toxicity, always choose the safe product.

The mold in our **hot cereals** can be spotted. Pick out all dark colored, shriveled bits. This represents most of it. Add honey, and salt while it's cooking—this raises the boiling temperature and detoxifies more. At the end, turn off heat and add a sprinkle of vitamin C powder. Rolled oats never showed molds in my testing, although they have their characteristic fungi, too. Don't let grains mold on your shelves simply from aging. Nothing should be more than six months old. Remember you can't see or smell molds when they begin. Molds must have a degree of moisture. As soon as you open a cereal grain, put the whole box in a plastic bag to keep moisture out. This keeps out Weevils too, so you won't have to put the box in the freezer later to kill them.

Anything that is put in the refrigerator or freezer and then taken out develops moisture inside. Store cereals in kitchen cupboards or the freezer.

No government agency can test for all of these mycotoxins in all of our foods. Production and storage methods must be better regulated so as to be fail-safe. Simply sending inspectors out to look into the bins at grain elevators is not sufficient. Crusts of mold, sometimes several feet thick, that form on top of grain bins can be simply shoveled away before the inspector arrives. The humidity and temperature of stored grain should be regulated, requiring automated controls. This would soon be cost effective, too, in terms of reduced spoilage losses and higher quality prices earned. I believe that zear, aflatoxin and ergot require special regulations. Products that are imported should be subjected to the same tests as ours. Test results should be on the label.

Getting Away From Grains

In view of the many molds that are grain-related, and because these cannot be seen or smelled in pastas, breads, cold cereals, it would be wise to steer away from grain consumption. Always choose potatoes, because it is a vegetable instead of a grain, if you have a choice. The potato appears on your plate the way it was harvested. Whereas grain was hulled, stored for quite a long time, perhaps degerminated (the bran and germ picks up

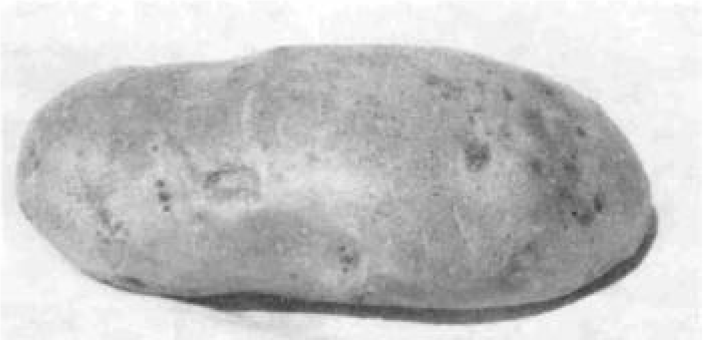

Fig. 54 Don't eat the green on the potato.

mold the fastest). Then it was mixed with assorted chemicals (fumigation, anti oxidants), each polluted in its own way, packaged again and stored again. Grains have a more tortuous history than potatoes that simply get sprayed.

The spray isn't simple, of course. Scrub it off under the tap. If potatoes weren't heavily sprayed they'd be sprouting in the stores. The spray accumulates in the eyes. Cut away all the eyes. By the time you have done this you may as well have peeled them. But no blemish, no cut, no dark spot inside may be left for you to eat. Don't buy potatoes that show a tint of green on them (the green color is due to *scopolamine*; it is toxic). Red potatoes have different chemistry that doesn't produce the green toxin, buy these often. Store potatoes out of the light, to slow down the greening process. They are still a nutritious, vitamin C-rich food—provided you don't fry them in benzene-polluted, hydrogenated grease!

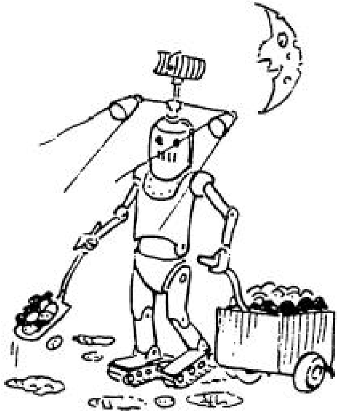
Fig. 55 Potato harvest of the future.

Potatoes have their molds but they are nicely visible. And washing and peeling does away with them. Old literature advises that potatoes should be harvested by moonlight so the green drug isn't produced in the white varieties. With modern mechanized harvesting this should pose no problem. But perhaps this must await the age of robots.

Pets Teach Health Lessons

Dogs don't eat hay and cats don't eat fruit. This is not simply due to their inability to digest them. Nor to training. Their body wisdom guides them. But we can trick them into eating corn and soybeans by adding the flavors they like and thereby defeat their wisdom the same way we defeat ours. A concoction is made for them that is called “complete nutrition” and we feed this meal after meal, day after day, a most unnatural situation. The liver is deluged with the same set of pollutants time after time and never gets a rest. Humans still obey their body wisdom about varying their meals. This gives the liver a chance to catch up with detoxifying one pollutant while the new one builds up. If the liver is absolutely unable to handle something, you are informed quite quickly with an allergic reaction to the food.

Cats and dogs with their monolithic diet get no opportunity to reject food (except by vomiting or starvation). It is not surprising they are getting cancer with increasing frequency, a situation where the liver can no longer detoxify isopropyl alcohol, a common pollutant in their food.

Should we go back to the old days and make their food for them? Yes, they deserve pure food, they deserve variety. Table scraps would be much less toxic for them than their commercial feed. But what if they like and prefer their monolithic “scientific”, “complete”, polluted diet? If our food was doused with sugar for breakfast, lunch and supper, we wouldn't care much about what was under the sugar either. And we'd continue wanting sugar, sugar, sugar the way a pet might want its favorite food and nothing else. Such is the deception of flavorings.

All change should be brought about slowly and with kindness for animals and humans alike. Learn what makes a good pet diet. Cats and dogs are both meat eaters. Cook chicken in a pressure cooker to kill all parasites. Put portions in the freezer. Add table scraps, dressed with a little butter, cheese or lard.

Don't wash the pet dishes with your own—dishwashers don't reach the boiling point. Serve fresh water daily. Standing water picks up bacteria. Don't let food get more than a day old in the dishes. It picks up molds. Don't feed pets at the table, keep them outdoors during mealtime.

After your pets have stopped eating propyl alcohol polluted food and are not getting propyl alcohol in their shampoos, there is no way they can get cancer. Whatever cancer they have will clear up by this change in diet and by giving them the pet parasite program.

Now they are back to a natural state and do not host human flukes. What a relief it is not to worry about reinfection from your pets.

Easy Lifestyle Improvements

None of us likes to change a habit. But once it is changed you are back to an automatic way of doing something. By selecting wise habits your improved lifestyle pays you back for the rest of your life.

Living Hand To Mouth

Hands do everything. They pick up things from the floor. They handle money. They touch other peoples' hands. They clean up bottoms. They touch all kinds of door knobs. And then they pick up food to eat. Some people even lick their fingers when they're sticky or just to turn a page!

What is on the hands that you don't want to eat but can't see? **Bacteria** and **viruses** from coughing and sneezing into your hands! And **cysts**! Cysts are the "eggs" made by parasites. Cysts are so tough not even bleach kills them. They hide under our fingernails when we wash our hands. Then we eat them along with our food. This is called *oral-fecal route*. They hatch in the stomach and go to the intestine to live.

To stop reinfecting yourself the little cysts under the nails need to be killed. Food grade alcohol solution kills them. Buy Everclear^{TM} or Protec (potable) alcohol and make a 5% solution (add ¼ cup of 95% alcohol to a quart of water). Keep it in a small pump bottle at the sink. After using the bathroom and washing your hands, treat your fingernails with alcohol. Pump alcohol into one palm. Put the fingernails of the other hand into it. Scratch a bit. Pour it into the other palm and do the remaining fingernails. Rinse.

- **Don't** eat with your hands! Use a fork.
- **Never** eat food off the floor!
- **Always** wash hands after petting an animal!
- **Never** touch the bottoms of shoes! Keep shoes off couch or bed or chair.
- **Always** cough or sneeze into your clothing or a tissue, not your hands.
- **Keep** your fingers out of your mouth. Don't lick your fingers to turn pages or open plastic bags.

Sick persons need a 50% alcohol solution. Add ½cup 95% alcohol to ½cup cold tap water or buy plain vodka, 80 to 100 proof. Pour the vodka into the pump bottle. Be careful that nobody tries drinking it. If there are teenagers in the house, add a hefty dose of cayenne to it.

Lugol's iodine will also sterilize your hands. However most commercial Lugol's is polluted with isopropyl alcohol. Ask your pharmacist to make it from scratch for you (there are only two ingredients and water, see *Recipes*). Then make a solution to wash in (1 tsp. to a quart of water). This can stain some things. Do not use "tincture of iodine."

Better Laundry Habits

Boil your underwear. In long-ago days, all sheets, towels, table cloths, and underwear were separated and boiled.

With the convenience of our electric washing machine, we tend to overlook the fact that underwear is always contaminated by fecal matter and urogenital secretions and excretions. Mixing these with socks and towels and dishcloths is all right if you are going to kill everything anyway. But if you don't kill them, as in cold water washes, you are mixing the yeast, parasite eggs, bacterial spores, and fungus from underwear with all the other clothing you and your family wear. An enlightened system would be to add an antiseptic to the wash or rinse cycle. Lime water (calcium hydroxide) or iodine based antiseptics seem obviously simple methods to accomplish this. In the absence of this protection, use dryer heat to do your sterilizing. Underwear should be dried until too hot to handle.

Bleach can kill a lot, but doesn't kill *Giardia* spores and a lot of types of fungus. Don't rely on bleaching. Besides, your skin absorbs it from clothing, it is quite toxic to you, and can cause mental effects.

Commercial detergents are polluted with PCBs and have cobalt added. Both of these are easily picked up through your skin. Use **washing soda** and **borax** in the wash cycle. They do not clean quite as well as modern detergents, but there is less static cling, eliminating the need to put more chemicals in your dryer. For spot removal use homemade bar soap.

Better Kitchen Habits

Once a day, sterilize the sponge or cloth you use to wipe up the table, counter tops and sink. This little piece of contaminated cloth is the most infectious thing in the house, besides the toilet. It's more dangerous than the toilet because you do not suspect it. Sometimes it has a slight odor at first, which may warn you, but most pathogens do not have an odor! As we wipe up droplets of milk, we give the milk bacteria, *Salmonellas* and *Shigellas*, a new home to multiply and thrive in. We add crumbs, picking up molds this way. We add dust, picking up parasite eggs and stages. They all feed on the milk and food residue.

As the counter and table and stove get wiped “clean” a film of contamination is left everywhere. A few varieties may die but most of them don't. The general moisture in the kitchen is enough for them to survive. The cloth or sponge recolonizes the kitchen and dining room table several times a day.

No doubt, the last thing you do before leaving the kitchen is squeeze it dry with your hands. Now all the pathogens are on your hands!

Where do your fingers go? To your mouth to remove a hull or bit of something from your teeth. Or to eat a last bite of something. Or to turn a page of the telephone directory. You have just eaten a culture sampling from your own kitchen sponge. In two hours they are already multiplying in the greatest culture system of all: your body! You have given yourself your next sore throat, or cold or headache. The worst possible habit is to wipe a child's face and hands with the kitchen cloth. Or to have a handy towel hanging from the refrigerator handle.

To sterilize the sponge: **drop it into a 50% solution of grain alcohol** at the end of each day. Keep a wide mouth glass quart jar handy just for this. Keep the jar tightly closed and out of the reach of children. Dunk your sponge and plop it onto the sink. If you stand it on end in the sink it will partly dry overnight.

Another way to sterilize the sponge or cloth is to **microwave it**, after wetting it, for 3 minutes. Any shorter time simply warms and cultures the pathogens and multiplies them. Or **boil the cloth** like our grandparents did. Drying out the dish cloth helps kill many–but not all–pathogens. It takes three days of drying to kill all! Another strategy is to **use a fresh cloth** or sponge each day, putting the used one to dry until laundry day.

During the day, set the sponge on end to start drying and slow down culturing.

Don't eat food directly off the counter top or table top. You wouldn't slice a tomato or egg directly on the counter top. It would pick up something: some little particle of dust or dirt. Treat bread the same way. Always on a new clean surface, such as a plate. The counter and table top have on them whatever is in the kitchen dust and on the wipe cloth. Dust is always falling! And the sponge is always culturing. Don't eat the dust!

Keep the cutting board sterile like dishes. Wash it the same way and keep it in the cupboard.

Keep food containers closed. Milk or water glasses are picking up dust as soon as you set them out. Dust is everywhere. Every step on the carpet sends up a puff of dust. Vacuuming sends up a hurricane of dust and distributes bathroom dust to the kitchen and kitchen dust to the bedrooms. So if one person has brought in a new infection, the whole family is exposed to it in hours via the dust.

It is very helpful not to eat the new infectious pathogen. Breathing it is not so damaging. Our noses collect such pathogens and we blow them out again. Touching the infected person is not very damaging either; the pathogens can't get through our skins and since we wash hands before eating we are not at great risk of infection this way. But eating the pathogen is 100% effective in infecting us. The new pathogen is in the dust. The newly contaminated dust drops into your ready and waiting glasses on the table and the open foods. Of course, there is no defense if somebody should cough or sneeze at the table.

Teach children to cough and sneeze into a suitable collecting place like a tissue, not their hands. Pathogens live bountifully on hands. Hands not only provide moisture but often food from the last meal. Hands are second only to the dish cloth in contamination level. If you must cough or sneeze and a tissue is not within reach fast enough, use your clothing! That's what clothing is for—to protect you. Cough and sneeze into your own clothing; this protects the cougher and sneezer, as well as everybody else. A sleeve is handy for children. The inside of your T-shirt for T-shirt wearers. The inside of coat for suited persons. The inside of the neck line for dresses. Of course, paper is best, but in emergency use cloth. Never, never your hands unless you are free to immediately dash into the washroom and clean the contamination off your hands.

Teach children this old rearranged verse:

If you cough or sneeze or sniff
Grab a tissue, quick-quick-quick!
And if you're sitting at the table
Do it in your sleeve if able.

Better Housekeeping

Throw out as much of the wall to wall carpeting as you can bear to part with. It is injurious to everyone's health, even though it's comforting to bare feet and looks pretty.

Carpets clean our shoes. Modern shoes, with their deep treads, bring in huge amounts of outdoor filth which settles deep down into the carpets. In spite of vacuuming every week, the filth accumulates.

Vacuum the carpets when the children, the sick and elderly are out of the house. The dust raised and distributed throughout the house isn't just dirt, it's infectious dirt. It lands on tables and counters. These get wiped with a cloth or sponge and then applied to dishes. The dust in the kitchen falls on open food and into open containers.

Clean carpets with a “steam cleaner”. When you see how much filth is in the water and realize how much dirt you were living with, you might be willing to trade in the “beauty” of carpets for the cleaner living of smooth floors. Don't add chemicals (commercial cleaning solutions) to the steam cleaning machine; these chemicals leave a residue in the carpet which dries and flies up into the air to add to the dust. Popular stain resistance treatments contain arsenic. Cobalt, which adds “lustre” to carpets, causes skin and heart disease after it has built up in your organs. Use borax instead of detergent. Use boric acid to leave a residue that kills roaches and fleas (but not ants). Add vinegar that leaves a residue to repel ants. Nothing controls fleas reliably, except getting rid of the carpets and cloth furniture (keep pets out of bedrooms). Removing all the borax is what brings luster to the carpet. Use citric acid in the rinse water for this purpose. Adding lemon peel to the rinse also adds luster and ant deterrence. Just drop the whole lemon in the tank so it can't block the hoses.

Fleas and other vermin in the carpet simply crawl below the wetness level when you wash the carpet. Spraying a grain alcohol solution with lemon peel in it (it needs to extract for a half hour) on the damp carpet will reach and kill a lot of these, together with the residual bacteria. The damp carpet lets it spread evenly and reach all the crevices.

We are trapped in our dwellings. Primitive peoples were mobile. This got them away from accumulations of filth and rubbish in their living space. Much living was done outdoors, the cleanest space of all.

Now, air conditioning has made indoor living more comfortable. But also has added new hazards. The strong currents of air blow the dust about continuously. Molds and bacteria that grow right on the air conditioning unit get blown about for all to inhale. Never, never use fiberglass as a filter or to insulate your air conditioner around the sides. It is a carcinogen. And the danger of freon escaping from a tiny leak is another major health hazard.

Forced air heating systems are undesirable, too. All dirt brought into the house by shoes gets circulated throughout the house by forced air systems of heating or cooling. Old fashioned radiant heat from radiators or a stove did not distribute the dust so effectively. A return to linoleum floor covering for kitchen and bathroom and hardwood for other rooms would be a good step of progress for a health conscious society. Mopping, instead of the vacuum cleaner, keeps dirt to a minimum. Throw rugs at doors and bedside, easy to clean, would "catch the dirt" as was the original intention. Carpets were intended to help keep filth out of the air. These smaller rugs should be laundered weekly.

Furniture should be wood, cane, or plastic, with cushions to soften the impact. These can be washed weekly if the covers are removable. Modern cloth furniture with its foam interior is a repository of filth and fumes and a constant source of infectious dust.

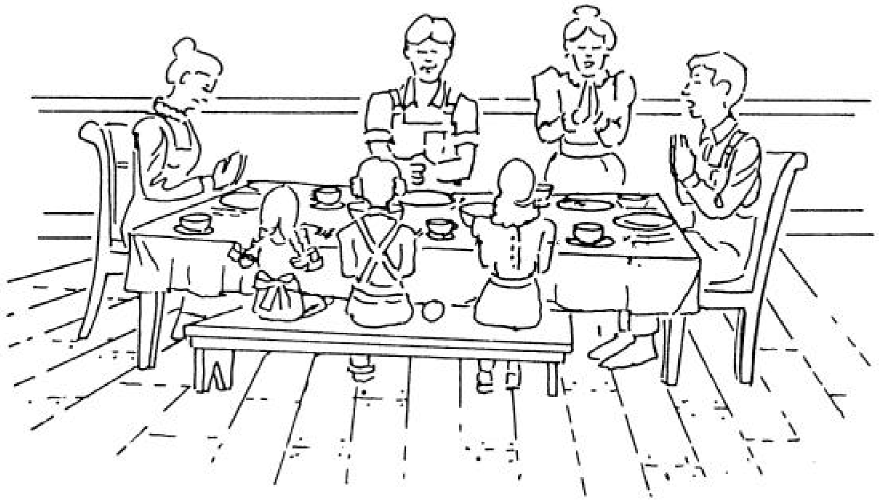

This old fashioned setting is more progressive than our carpeted, modern homes.

Fig. 56 Smooth floors allow every bit of dust to be removed.

Dust your furniture with a damp paper towel. You are picking up and removing highly infectious filth (*Ascaris* and pinworm eggs, pet parasites, "dander" and house mites). Instead of distributing these from room to room, throw the paper towel away after each room is done. Use plain water or vinegar water (50%), not a chemical combination which further pollutes the air.

Clean windows with vinegar water, too. Use a spray bottle.

Keep your dishes in cupboards. This keeps them free of dust. This principle is ancient. It is tempting to leave some of them out. If you must keep the juicer or dishes outside of cupboards, keep them covered or placed upside down so they don't catch dust. Even inside cupboards, store them upside down. When using the “good” dishes or glasses, that haven't been used in a while, wash them first.

Windows Open Or Closed?

In places like Chicago where you can smell the air as you approach the city, it is wiser to keep your windows shut. You can't breath the industrial “soup” all day and night and expect to stay healthy. Of course, it all enters the houses anyway. Central air conditioning and a plain carbon filter at the furnace location (see *Sources*) may be the best solution in spite of blowing dust around the house. Keep the vents to the bedrooms closed to reduce the air turbulence there but leave the cold air return open. Clean the vents in other rooms each week along with floors and carpets by pulling up the grating and reaching down the passage as far as possible.

If you believe the air is free of highway exhaust and industrial smoke open the windows every day. This will let some of the indoor toxins blow away. Asbestos, fiberglass, freon, radon and plain dust can be reduced to a minimum by keeping windows open. If you are ill, sit outdoors (on the porch) as much as you can. Escape to a suitable climate that makes this possible.

Just a few decades ago, many people had summer living quarters that were different from winter living quarters. Gone was all the accumulated infectious dust of half a year of habitation.

Fig. 57 Moving into the summer kitchen got you away from the accumulation of filth from winter!

Don't have a basement where you stockpile toxic items. Basements invite mold, mice and radon besides toxic things. Fumes travel upward where you live! Keep your toxic things in the attic. If there is no attic, store them in the utility room. Close off the ventilation between utility room and the rest of your house. If you have none of these, perhaps because you live in a senior citizen community or condominium, don't keep any toxic things stored anywhere. Don't save any leftover paints, solvents or cleaners. Buy such small quantities that you can afford to throw it all away when you are done with them.

Live on top of the earth as was intended by nature.

Never have a basement room "finished" for actual living space. Don't buy a house that has a "lower level" built into the earth. This will be the most polluted and dangerous room in your house. If you are ill, move out of such a room. There is no way that it can be "cleaned up". Move to the other end of the house and furthest away from an attached garage door.

What Kind Of Heat

The worst is coal. The best is none. Breathing coal fumes during the beginning of the industrial age may have brought the new lung diseases: tuberculosis (TB), and pneumonia. It may also have worsened alcohol addiction (beryllium toxicity). Choose electric heat if possible. Even though electricity is based on other fuel consumption, you don't have to breathe those fumes directly.

Wood stoves can be made safe by making sure the chimney works properly. Never use a lighter fluid. Don't fill the house with smoke when stoking.

Minimize your use of fossil fuels in every way you can.

Getting Rid of Mites

We do not tolerate external parasites like bedbugs, lice, ticks, leeches. Bedbugs were once a scourge amongst northern Europeans. I remember our parents spraying for them (kerosene) in the bedroom. This only "controlled" them. What eliminated them was a law against sale of used mattresses. Lice were originally "controlled" by frequent washing, louse combs, and ironing the seams of clothing. What eliminated them was the cutting of long hair as a societal practice. But what about **mites**? They live with us and other animals.

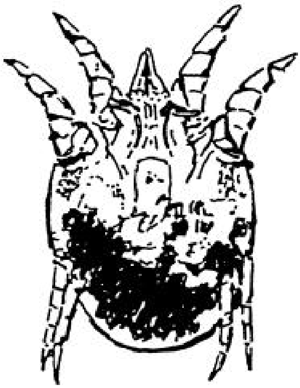

Fig. 58 Mite.

Mites are too tiny to see, tiny enough to ride on a dust particle as if it were a magic carpet. They resemble insects. Chiggers are really mites. Mange in animals is a mite infestation. Dust mites live on our dander (scales of dead skin).

Get rid of their breeding places: beds, cloth covered chairs and soft sofas. Humans leave enough dander behind in these places to support these ultra small insects. Cover mattresses with plastic covers. Use throws on easy chairs and sofas and wash them often. Never allow a pet into the bedroom or the dust will have tapeworm eggs as well as mites. Throw out rugs that have been pet-beds. Spray the air with a mist of 50% grain alcohol before vacuuming. If you have an illness wear a mask to vacuum. Deep, soft, wall to wall carpets compromise an ancient concept: everything should be washable and cleanable, without throwing the dirt into the air for humans to inhale. Vacuuming a carpet blasts mites and tape eggs into the air. Never shake bedding or rugs where the dust will blow back into the house behind you.

Mites don't bite us but we inhale them as they float in the ever present dust in our homes. The mucus in our lungs traps them and in a few days they die, only to release a drove of *Adenoviruses* (common cold virus) in us.

Four Clean-ups

Chronic health problems are not due to exposures of the past. They are ongoing. Your body is constantly fighting to remove pollutants. In order to stay sick, you must be constantly resupplied! These four clean-ups–dental, diet, body, home–are aimed at removing parasites and pollutants at their source. Only then can your body heal.

Dental Clean-up

(This section on dentistry was contributed by Frank Jerome, DDS)

Dr. Jerome: The philosophy of dental treatment taught in America is that teeth are to be saved by whatever means available, using the strongest, most long lasting materials. Long-term toxic effects are of little concern. The attitude of the majority of dentists is: whatever the American Dental Association (ADA) says is OK, they will do.

A more reasonable philosophy is that there is no tooth worth saving if it damages your immune system. Use this as your guideline.

The reason dentists do not see toxic results is that they do not look or ask. If a patient has three mercury amalgam fillings placed in the mouth and a week later has a kidney problem, will she call the dentist—or the doctor? Will they ever tell the dentist about the kidney problem or tell the doctor about the three fillings? A connection will never be made.

It is common for patients who have had their metal fillings removed to have various symptoms go away but, again, they do not tell the dentist. The patient has to be asked! Once the patient begins to feel well they take it for granted, and don't make the connection, either. If everybody's results were instantaneous, there would be no controversy.

Find an alternative dentist. They have been leading the movement to ban mercury from dental supplies. Not only mercury, but all metal needs to be banned. If your dentist will not follow the necessary procedures, then you must find one that will. The questions to ask when you phone a new dental office are:

1. Do you place mercury fillings? (The correct answer is NO. If they do, they probably don't have enough experience in the use of non-metal composites.)
2. Do you do root canals? (The correct answer is NO. If they do, they do not understand good alternative dentistry.)
3. Do you remove amalgam tattoos? (The correct answer is YES. Tattoos are pieces of mercury left in the gum tissue.)
4. Do you treat cavitations? (The correct answer is YES. By cleaning them.) The complete name of *cavitations* is *alveolar cavitational osteopathosis*. They are holes (cavities) left in the jawbone by an incompletely extracted tooth. A properly cleaned socket which is left after an extraction will heal and fill with bone. Dentists routinely do NOT clean the socket of tissue remnants or infected bone. A dry socket (really an infected socket) is a common result. These sockets never fully heal. Thirty years after an extraction, a cavitation will still be there. It is a form of *osteomyelitis*, which means bone infection.

Ninety percent or more of dental offices will not be able to answer ANY of the above questions correctly. If you allow the work to be done by a dentist who does not understand the importance of the above list, you could end up with new problems. Find the right dentist first even if you must travel hundreds of miles. There are 6,000 to 10,000 dentists who should be able to help. Some can do part of the work and refer you to a specialist for the rest. Five hundred to one thousand of these dentists can do it all.

Normal treatment cost is about $1,000 for replacement of 6 to 8 metal fillings including the examination and X-rays. For people with a metal filling in every tooth, or for the extraction of all teeth (plus dentures), it may be up to $3,000 (or more in some places).

Remember, the simpler the treatment, the better. If the dentist says that he or she can change your metal fillings to plastic but it would be better to crown them, say "NO!"

Guidelines For A Healthy Mouth

If you have	**What to do**
Metal fillings	change to plastic fillings
Inlays and onlays	change to plastic fillings
Crowns (all types)	change to plastic crowns
Bridges	change to plastic crowns, partials
Metal partials	change to plastic partials(Flexite™)
Pink dentures	change to clear plastic
Porcelain denture teeth	change to plastic denture teeth
Badly damaged teeth	become extractions
Root canals	become extractions
Braces and implants	avoid
Cavitations	need to be surgically cleaned
Temporary crowns	use plastic
Temporary fillings	use Duralon™

Fig. 59 Dental replacements.

The guidelines can be summarized as:

1. Remove all metal from the mouth.
2. Remove all infected teeth and clean cavitations.

Dr. Clark: *Removing all metal means removing all root canals, metal fillings and crowns. Take out all bridge work or partials made of metal and never put them back in. But you may feel quite attached to the gold, so ask the dentist to give you everything she or he removes. Look at the underside. You will be glad you switched.*

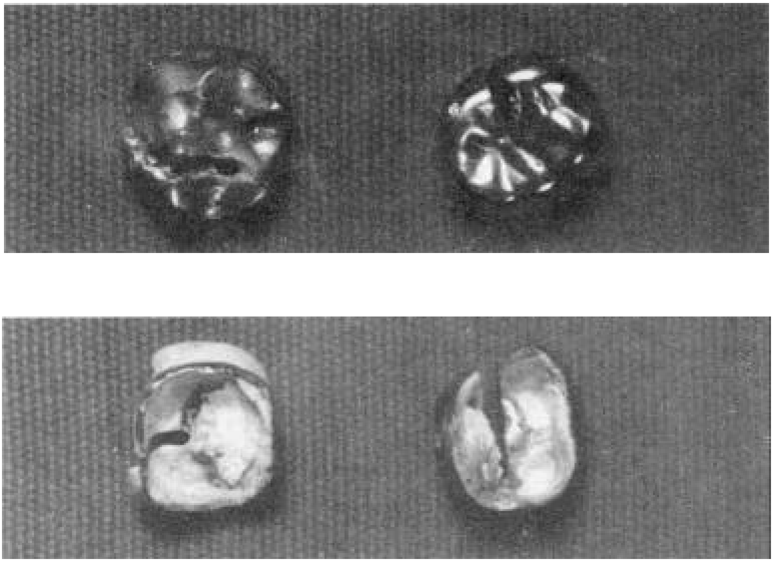

The top surfaces of tooth fillings are kept glossy by brushing (you swallow some of what is removed). Underneath is tarnish and foulness. Ask to see your crowns when they are removed.

Fig. 60 Tops and bottoms of some metal crowns.

The stench of the infection under some teeth may be overwhelming as they are pulled. Bad breath in the morning is due to such hidden tooth infections, not a deficiency of mouthwash!

All metal must come out, no matter how glossy it looks on the surface. Metal does not belong in your body. It is an unnatural chemical. Do this as soon as you have found a dentist able to do it. Find a dentist with experience and knowledge about this subject. It is more than replacing acknowledged culprits like mercury-amalgam fillings. This is metal-free dentistry. Only metal-free plastic should be put back in your mouth.

Dr. Jerome: If your dentist tells you that mercury and other metals will not cause any problems, you will not be able to change his or her mind. Seek treatment elsewhere!

Your dentist should do a complete X-ray examination of your mouth. Ask for the *panoramic* X-ray rather than the usual series of 14 to 16 small X-rays (called full mouth series). The panoramic X-ray shows the whole mouth including the jaws and the sinuses. This lets the dentist see impacted teeth, root fragments, bits of mercury buried in the bone and deep infections. Cavitations are visible in a panoramic X-ray that may not be seen in a full mouth series.

The cost of removing metals should be viewed in the proper light. It took years or decades to get into your present condition. When you do a lot of dental repair in a short time, it can seem to be costly. Unfortunately, many people are in a tight financial position because of the cost of years of ineffective treatment, trying to get well.

Your dentist may recommend crowning teeth to “protect” or strengthen them. Unfortunately, the very concept of crowning teeth is flawed. First, the enamel is removed from a tooth to prepare for the crown. This is permanent and serious damage! Many teeth, up to 20%, may die after being crowned and will need to be extracted. For this reason, you should only get REPLACEMENT crowns and NO NEW crowns. Your metal crowns can be changed to plastic. (Remember, no metal must be left under the crown.)

If you have many crowns, you should have them all removed as quickly as possible. But you should not spend more than two hours in the dentist's chair at any one time. That is too much stress for your body.

Dr. Clark: *Don 't accept intravenous (IV) treatments during amalgam removal. Both IV bags and the supplements used in them are polluted with propyl alcohol, benzene, and wood alcohol.*

Dr. Jerome: It is quite all right to have temporary crowns placed on all teeth that need them in the first visit. You may then go back and complete treatment over the next 6 to 12 months. It is common to find a crowned tooth to be very weak and not worth replacing the crown, particularly if you are already having a partial made and could include this tooth in it.

Dr. Clark: *We are accustomed to thinking that plastic is metal-free. This is wrong. The original dental plastic, methyl methacrylate was metal-free. But modern plastic contains metal. The metal is ground up very finely and added to the plastic in order to make it harder, give it sheen, color, etc.*

Dr. Jerome: Dentists are not commonly given information on these metals used in plastics. The information that comes with dental supplies does not list them either. Most dentists never look at a dental materials book after they graduate. The ADA, however, has a library full of such information.[21]

Dr. Clark: *There are many lanthanide (Rare Earth) metals used in dental plastic. Their effects on the body from dentalware*

---

[21] Call the American Dental Association at (800) 621-8099 (Illinois (800) 572-8309, Alaska or Hawaii (800) 621-3291). Members can ask for the Bureau of Library Services, non-members ask for Public Information.

have NOT been studied. Yet their cancer-promoting ability is known in many cases.[22] *Only metal-free plastic is safe.*

Dr. Jerome: These are the acceptable plastics; they can be procured at any dental lab.

- Plastic for dentures: Methyl Methacrylate. Available in clear and pink. Do not use pink.[23]
- Plastic for partial dentures: Flexite™ Available in clear and pink. Do not use pink.
- Plastic for fillings: Composite Materials. This is the material that has been used in front teeth for 30 years. It has been used in back teeth for 10 years. There are many brands and there are new ones being marketed constantly. The new ones are very much superior to those used 10 years ago and they will continue to improve. They do, however, contain enough barium or zirconium to make them visible on X-rays. There are no alternatives available without these metals.

Dr. Clark: *Composites with barium are not good, but I haven't seen enough barium toxicity from fillings at this time to merit advising extraction instead. Hopefully, a barium-free variety will become available soon to remove this health risk.*

Dr. Jerome: Many people (and dentists too) believe that porcelain is a good substitute for plastic. Porcelain is aluminum oxide with other metals added to get different colors (shades). The metal DOES come out of the porcelain! It has many technical drawbacks as well. Porcelain is not recommended. Sometimes the white composite fillings are called porcelain fillings but they are not. They also require more tooth structure to be removed.

---

[22] Thulium and ytterbium have been studied for their tumor-seeking ability. See page 321 in the book *METAL IONS In BIOLOGICAL SYSTEMS, Vol. 10, Carcinogenicity and Metal Ions*. Editor Helmut Sigel 1980.

[23] The pink color is from mercury or cadmium which is added to the plastic.

If you have a large bridge, it cannot be replaced with a plastic bridge because it isn't strong enough. A large bridge must be replaced with a removable partial (Flexite^{TM}).

The methods used to remove metals and infections are technical and complicated. See dental information in *Sources*.

Dr. Clark: *I'd like to thank Dr. Jerome for his contributions to this section, and his pioneering work in metal-free dentistry. I hope more dentists acquire his techniques.*

Horrors Of Metal Dentistry

Why are highly toxic metals put in materials for our mouths? Because not everyone agrees on what is toxic at what level. Just decades ago lead was commonly found in paint, and until recently in gasoline. Lead was not less toxic then, we were just less informed! The government sets standards of toxicity, but those "standards" change as more research is done (and more people speak out). You can do better than the government by dropping your standard for toxic metals to zero! Simply remove them.

The debate still rages over mercury amalgam fillings. No one disputes the extreme toxicity of mercury compounds and mercury vapor. The ADA feels that mercury amalgam fillings are safe because they do not vaporize or form toxic compounds to a significant degree. Opponents cite scientific studies that implicate mercury amalgams as disease causing. Many dentists advocate mercury amalgam fillings simply because they are accepted by the ADA, which they believe protects them from malpractice litigation. Why risk your health and life on their opinions? Remember everything corrodes and everything seeps, so amalgams must too.

Cadmium is used to make the pink color in dentures! Cadmium is five times as toxic as lead, and is strongly linked to high blood pressure.

Occasionally, thallium and germanium are found together in mercury amalgam tooth fillings. Thallium causes leg pain, leg weakness, and paraplegia. If you are in a wheelchair without a very reliable diagnosis, have all the metal removed from your mouth. Ask the dentist to give you the grindings. Try to have them analyzed for thallium using the most sensitive methods available, possibly at a research institute or university.

I was astonished to find thallium in mercury amalgams! It couldn't be put there intentionally, look how toxic it is:

> ***TEJ500*** ***HR: 3***
>
> **THALLIUM COMPOUNDS**
>
> Thallium and its compounds are on the Community Right To Know List.
>
> THR: Extremely toxic. The lethal dose for a man by ingestion is 0.5-1.0 gram. Effects are cumulative and with continuous exposure toxicity occurs at much lower levels. Major effects are on the nervous system, skin and cardiovascular tract. The peripheral nervous system can be severely affected with dying-back of the longest sensory and motor fibers. Reproductive organs and the fetus are highly susceptible. Acute poisoning has followed the ingestion of toxic quantities of a thallium-bearing depilatory and accidental or suicidal ingestion of rat poison. Acute poisoning results in swelling of the feet and legs, arthralgia, vomiting, insomnia, hyperesthesia and paresthesia [numbness] of the hands and feet, mental confusion, polyneuritis with severe pains in the legs and loins, partial paralysis of the legs with reaction of degeneration, angina-like pains, nephritis, wasting and weakness, and lymphocytosis and eosinophilia. About the 18th day, complete loss of the hair on the body and head may occur. Fatal poisoning has been known to occur. Recovery requires months and may be incomplete. Industrial poisoning is reported to have caused discoloration of the hair (which later falls out), joint pain, loss of appetite, fatigue, severe pain in the calves of the legs, albuminuria, eosinophilia, lymphocytosis and optic neuritis followed by

> atrophy. Cases of industrial poisoning are rare, however. Thallium is an experimental teratogen [used to induce birth defects for study]. When heated to decomposition they [sic] emit highly toxic fumes of Tl [thallium]. See also THALLIUM and specific compounds.[24]

Fig. 61 Thallium excerpt.

Thallium pollution frightens me more than lead, cadmium and mercury combined, because it is completely unsuspected. Its last major use, rat poison, was banned in the 1970s. Every wheelchair patient I tested was positive for thallium! One current use for thallium is in Arctic/Antarctic thermostats. When added to mercury the mercury will stay liquid at lower temperatures. Are mercury suppliers then providing the dental industry with tainted amalgam?

The cancer causing or carcinogenic action of metals has been studied for a long time, although it doesn't get attention by our regulatory agencies. A scientific book on this subject was published in 1980.[25] One table from this book is shown on page 431. We can see that chromium and nickel compounds are the most carcinogenic metals. Nickel is used in gold crowns, braces, and children's crowns!

Note that the form of the metal is very important. For instance chromium is an essential element of *glucose tolerance*

---

[24] *Dangerous Properties of Industrial Materials*, 7th ed. by N. Irving Sax and Richard J. Lewis Sr., Van NOSTRAND, Reinhold N.Y. 1989.

[25] The title is *Carcinogenicity and Metal Ions*. It is volume 10 of a series called *Metal Ions in Biological Systems*, edited by Helmut Sigel. A university chemistry library should have this book. It has a fascinating chapter on the leukemias by two scientists from the Academy of Sciences of the USSR, E. L. Andronikashvili and L. Mosulishvili. Their brilliant work and discussion was largely responsible for my pursuit of the whole subject of cancer.

factor, but most of its other compounds are extremely toxic. In general, xenobiotic compounds (foreign) are to be avoided! Metal doesn't belong in our foods or in our bodies.

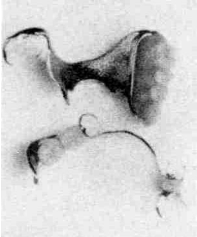

Fig. 62 More dental metal.

Dental Rewards

After your mouth is metal and infection-free, notice whether your sinus condition, ear-ringing, enlarged neck glands, headache, enlarged spleen, bloated condition, knee pain, foot pain, hip pain, dizziness, aching bones and joints improve.

Keep a small notebook to write down these improvements. It will show you which symptoms came originally from your teeth. Symptoms often come back! So go back to your dentist, to search for a hidden infection under one or more of your teeth, or where your teeth once were! That infection can be the cause of tinnitus, TMJ, arthritis, neck pain, loss of balance, and heart attacks!

Dentures can be beautiful. Of course, plastic isn't natural, but it is the best compromise that can be made to restore your mouth. At least it isn't positively charged like metals; it can't set up an electric current nor a magnetic field in your mouth, all of which may be harmful.

Do not be swayed by arguments that plastic is not as strong as metal. You see dentures everywhere and they seem strong enough to eat with. You will be told that "noble" metals like gold and platinum

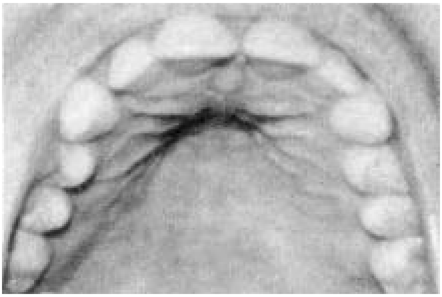

Fig. 63 Beautiful plastic mouth.

and silver are OK, that they are "inert" and do not corrode or seep. Nothing could be more untrue. You may be keeping them glossy by the constant polishing action of your toothpaste. But if you look at the underside, the view is frightful. **Everything tarnishes and everything seeps.** You wouldn't expect even a gold or silver coin that was dropped in a fountain 50 years ago to be intact. As metal corrodes your body absorbs it!

In breast cancer, especially, you find that metals from dentalware have dissolved and accumulated in the breast. They will leave the breast if you clear them out of your mouth (and diet, body, home). **The cysts shrink and are simply gone.** No need to do surgery!

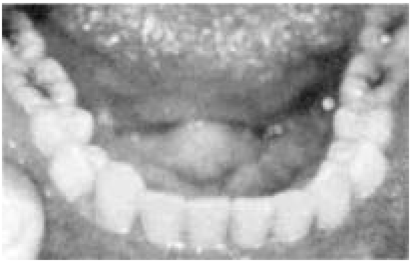

Fig. 64 Ugly metal in mouth.

Diet Clean-up

Breakfast

Cook your cereal from scratch. Don't eat cold cereal; it has numerous solvents and molds. Buy hot cereals that say "no salt added," like cream of wheat, steel cut oats or old fashioned oats,[26] millet, corn meal, cream of rice, or Wheatena. Cook it with milk to add nutritive value. Add your own (non-aluminum) salt and a pinch of vitamin C before cooking. Make granola from a recipe (see *Recipes*). Use honey, or brown sugar. Add raisins that were soaked for 5 minutes in vitamin C water. Use whipping cream or butter (both boiled) if you need to gain weight. Isn't this a delicious way to start your day! Add cinnamon to flavor, or frozen fruit and honey.

---

[26] Rolled oats have 235 mcg nickel per serving of 4 ounces, picked up from the rollers, according to *Food Values* 14th ed. by Pennington and Church, 1985. I have only found nickel in the "one-minute" or "instant" variety of oats, however.

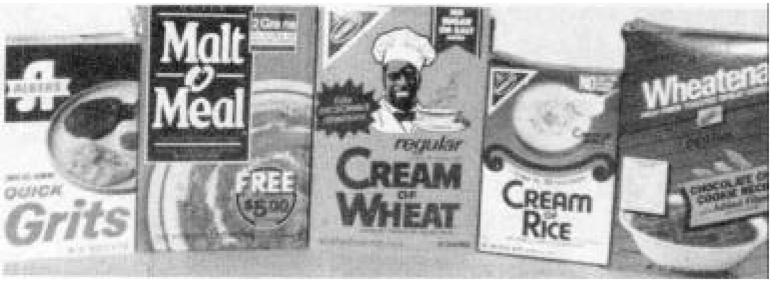

Fig. 65 Unpolluted breakfast cereals.

Or start your day with fried potatoes, an egg, and glass of milk. Don't worry about cholesterol since you will be doing liver cleanses anyway. (We have been told that eggs carry *Salmonella* bacteria. I found *Salmonellas* only on the outside shell and the egg carton—never inside! Could the researchers have accidentally transferred the bacteria from the shell to the inside while they were testing?)

The milk should be 2% or more butterfat because the calcium in milk cannot be absorbed without at least this much fat. Eat homemade yogurt and add honey or homemade preserves yourself. You need 3 cups of a milk product each day. Homemade buttermilk is fine. If you don't tolerate milk, and get diarrhea from it, try a milk digestant tablet to go with it. Start with only ¼ cup at a time. Do not choose chocolate milk. There is no substitute for milk; calcium tablets are not satisfactory. Vegetable matter, although high in calcium, does not give you *available* calcium either, unless you buy a juicer and make vegetable juice out of it. Eating fish can give you a lot of calcium, but it is in the tiny bones hidden in the fish. Don't try to remove them. Canned salmon has a lot of calcium; tuna does not. On a day that you eat fish, you would not need milk. Goat milk is probably better than cows' milk, but more difficult to get used to.

Lunch

Cook your food from scratch. Don't start with cans or packages or frozen items to make some recipe. In fact, don't bother with any fancy recipes. Just cook two or three vegetables for lunch and eat them with butter and salt or homemade sauces. Bread and milk rounds it out, plus fruit (not canned or frozen). Soup is a nice change. Cook it with all the vegetables you can find. Don't start with a can or packet or cube. Use a bit of onion and genuine herbs to give it zest. Thyme and fenugreek, together, make a flavorful combination you can purchase in capsules. Just pull apart and season.

If all this is too much work, make fresh vegetable juice once a week and freeze enough so that you can have a daily nutritious meal just by pouring a glass of it, together with bread and yogurt or milk.

Never diet during illness

This is a rule based on common sense. A weight-loss diet must wait at least two years.

Bake your own bread! I found aflatoxin in commercial bread after just four days in my bread box, but none in homemade bread even after two weeks! Aflatoxin is a most potent carcinogen and immunosuppressant. Aflatoxin is the toxin in your diet that keeps you from clearing propyl alcohol from your body (see 382)! Aflatoxin is a substance made by mold; bread starts to mold on the grocery store shelf. Don't buy plastic-wrapped bread. Don't ever salvage moldy food, whether it is fruit, breads or leftovers in the refrigerator. Throw them out. Buy a bread maker. It can do everything, including baking the bread. Use unbleached (unbrominated) flour and add ½ tsp. vitamin C powder per loaf to help retard mold further (it also makes the bread rise higher).

Supper

Cook your supper from scratch. Emphasize fish for animal food, not beef, pork, turkey or chicken. Don't buy bread crumbs, use your own. Don't buy batter, make your own. Use genuine eggs, not substitutes. Wash your hands after handling raw meat or eggs.

Make your own salad and salad dressing out of olive oil, fresh lemon juice or white distilled vinegar (apple cider vinegar has aflatoxins), honey, salt and herbs to flavor. If your digestion isn't strong enough for raw vegetables or fruit, make juice. Get a sturdy juicer and make your juice about half carrot juice and half from vegetables like celery, squash, lettuce, and broccoli. Make your own tomato sauce with pure herb seasoning, not from ajar or can (home canned foods are fine, of course, as long as they are not made in a big aluminum pot with aluminum-containing salt).

Cook real potatoes, not instant varieties. Peel them to get rid of Kojic acid (mycotoxin) and scopolamine (the green part). Make mashed potatoes from scratch, with milk, not box potatoes, nor chips nor French fries. Chips and fries were made in chemical grease called “hydrogenated.” There is a large amount of nickel in hydrogenated fats.[27] Fry your potatoes in butter, lard or olive oil. Find butter that is not wrapped in foil and is not salted. Salt your own butter, using aluminum-free salt.[28] Don't wrap potatoes in foil to bake, coat with butter or olive oil. Don't eat the peels.

Eat no meat that hasn't been cooked as thoroughly as if it were pork.

Other animals are as parasitized as we, full of flukes and worms and *Schistosomes* in every imaginable stage, and if the blood carries these, would we not be eating live parasites if we eat animals in the raw state? We have been taught to cook thoroughly any pork, fish or seafood. Now we must cook thoroughly any beef, chicken or turkey. It must be at cooking temperature (212°F or 100°C) for 20 minutes. Freezing is not adequate. Canned meats are safe from living parasites, but are not recommended due to added chemicals.

Beverages

Drink 6 kinds of beverages:

- milk
- water
- fruit juices
- vegetable juices
- herb teas
- homemade (see *Recipes*)

---

[27] 114 mcg/100 g. Taken from *Food Values* 14th ed. by Pennington and Church, 1985.

[28] Salt has aluminum in it to keep it from caking. Buy salt that has magnesium carbonate as its anti-caking agent. Sea salt must be baked for 5 minutes at 400°F to destroy molds. Or buy chemically pure salt (see *Sources*).

This means getting off caffeine. And if you are already fatigued, this means you might be even more fatigued for a short time. You might have headaches from withdrawal, too. But they will only last 10 days. Mark your calendar and count off the days. Take headache medicine, if necessary, but make sure it does not contain caffeine. For energy, to replace caffeine, take one arginine (500 mg, see *Sources*) upon rising in the morning and before lunch. Soon you won't need it.

Cutting down on coffee, decaf, soda pop and powdered drinks won't do. You must be completely off. They contain very toxic solvents due to careless, unregulated production methods. Much is imported and can't be sufficiently regulated.

Even though grain (drinking) alcohol is the recommended substitute for propyl alcohol, that doesn't mean you may safely drink it. It is inadvisable to drink any form of alcohol at least until you are fully recovered.

1. Milk: 2% or higher, drink three 8 oz. glasses a day. Alternate brands. Buttermilk will do. Homemade yogurt is fine. Goat milk is also fine. Start with ¼ cup and increase gradually, if you are not used to it. If you do not drink milk because it gives you more mucous, try to drink milk anyway. If you have other reactions, like diarrhea, try milk digestant tablets (available at health food stores). Milk is too valuable to avoid: there are many unwanted chemicals in most brands of milk, but it is solvent-free, mold-free and very nutritious. The only exception should be for serious symptoms, like swelling, colitis, flu, or chronic diarrhea.

But all milk, whether goat or cow, is contaminated with *Salmonella* and *Shigella* bacteria as well as fluke parasite stages. Cattle are immunized against *Salmonella* but it does not prevent its persistence in the bowel. All these are very harmful. Pasteurization does not kill all of them. Only heating to a rolling boil makes milk safe. To do this in the easiest way, pour 1 or 2 quarts milk into an enamel double boiler or microwavable glass jar. Stay in the kitchen while the heat is on. When the bubbles have risen to indicate boiling, turn off the heat. You may throw away the “skin”. Pour into glass jar and refrigerate. Another easy way is to use a pressure cooker that holds several pint jars of milk. All dairy products that have only been pasteurized are still contaminated. Ultrapasteurization does not improve matters. Dairy products that cannot be sterilized should not be consumed. It may be possible to find sterilized milk in paper containers on the store shelf—not in the refrigerator; if it wasn't sterile it would go foul in a day! Canned milk has solvent pollution. Powdered milk has both solvent and bacterial pollution.

2. Water: 2 pints. Drink one pint upon rising in the morning, the other pint in the afternoon sometime. The cold water faucet may be bringing you cadmium, copper or lead, but it is safer than purchased water, which inevitably has solvents in it. Let it run before using it. Filters are rather useless because water pollution comes in surges. A single surge of PCB contaminates your filter. All the water you use after this surge is now polluted, so you will be getting it chronically, whereas the unfiltered water cleans up again after the surge passes. Until you can test your own water for solvents, PCBs and metals, no expensive filter is worth the investment. An inexpensive pure carbon filter that is replaced every month may improve your tap water. Inflexible plastic pitchers fitted with a carbon filter pack are available (see *Sources*). Never buy filters with silver or other chemicals, even if they are just added to the carbon. Keep the filter sterile by soaking in diluted grain alcohol weekly.

3. Fruit juice: fresh squeezed only. Some stores make it while you wait. If they freeze some of it, you could purchase the frozen containers. Bottled fruit juices have traces of numerous solvents, as do the frozen concentrates, as do the refrigerated ones, don't buy them. You have to see it being made, but watch carefully: I recently went to a juice bar where they made everything fresh, before your very eyes. And I saw them take the fruit right from the refrigerator and spray it with a special wash "to get rid of any pesticides," then put a special detergent on it to clean off the wash! So instead of getting traces of pesticide, I got traces of propyl alcohol![29] Another grocery store had a machine that squeezed the oranges while you watched. But if you did not watch them filling the jugs, you missed seeing them add a tablespoon of concentrate, from a bottle out of sight, to give it better flavor. It still qualifies as "Fresh squeezed 100% orange juice," but thanks to that concentrate it now has toluene and xylene in it! Best of all, buy a juicer, select completely unbruised fruit, wash with plain water, and make your own juice (enough for a week—freeze it in half pint plastic bottles). For stronger flavor, leave some of the peel in the juice.

4. Vegetable juice: fresh or frozen only. If you or a friend would be willing to make fresh juice, this would be much better than purchased juice. Start with carrot juice. Peel carrots (don't scrape them, it's too easy to miss small dirt spots) and remove all blemishes carefully, then rinse. Drink ½glass a day. After you are accustomed to this, add other vegetables and greens to the juice to make up half of it. Use celery, lettuce, cabbage, cucumber, beet, squash, tomato, everything raw that you normally have in your refrigerator. Then drink one glass a day.

5. Herb tea: fresh or bulk packaged. Tea bag varieties are moldy. Buy a non-metal (bamboo is common) tea strainer. Sweeten with honey or brown sugar with vitamin C added.

6. Homemade beverage. If you will miss your coffee or decaf, try just plain hot water with boiled whipping cream. Sweeten with honey. Please see *Recipes* for many more suggestions.

---

[29] Yes, I took a sample of the wash to test.

Horrors In Commercial Beverages

Commercial beverages are especially toxic due to traces of solvents left over from the manufacturing process. There are solvents in decaffeinated beverages, herb tea blends (not single herb teas), carbonated drinks, beverages with Nutrasweet^{TM}, flavored coffee, diet and health mixes, and fruit juices, even when the label states "not from concentrate" or "fresh from the orchard," or "100% pure."

It is allowable to use solvents to clean machinery used in bottling (please look again at page 347)! It is also allowable to use solvents to make spice oleoresins, which are used as flavoring.

> **21 CFR 173.240 (4-1-94 Edition) Isopropyl Alcohol.**
>
> Isopropyl alcohol may be present in the following foods under the conditions specified:
>
> (a) In spice oleoresins as a residue from the extraction of spice, at a level not to exceed 50 parts per million.
> (b) In lemon oil as a residue in production of the oil, at a level not to exceed 6 parts per million.
> (c) [Discusses its use in hops extract.]

Here is a summary of other solvents mentioned:

Solvent	Allowable residue in spice oleoresins	Paragraph in 21 CFR
Acetone	30 PPM	173.210
Ethylene dichloride	30 PPM	173.230
Methyl alcohol	50 PPM	173.250
Methylene chloride	30 PPM	173.255
Hexane	25 PPM	173.270
Trichloroethylene	30 PPM	173.290

Fig. 66 Lawful uses of solvents in food.

I have found all these solvents and others in commercial beverages! Some of the solvents I have found are just too toxic to be believed! Yet you can build the test apparatus yourself (page 457), buy foods at your grocery store, and tabulate your own results. I hope you do, and I hope you find that the food in your area is cleaner than mine! Remember that the Syncrometer can only determine the presence or absence of something, not the concentration. There may only be a few parts per billion, but a sick person trying to get well cannot afford any solvent intake. For that matter, none of us should tolerate any of these:

- **Acetone** in carbonated drinks
- **Benzene** in store-bought drinking water (including distilled), store-bought fruit juice (including health varieties)
- **Carbon tetrachloride** in store-bought drinking water
- **Decane** in health foods and beverages
- **Hexanes** in decafs
- **Hexanedione** in flavored foods
- **Isophorone** in flavored foods
- **Methyl butyl ketone** and **Methyl ethyl ketone** in flavored foods
- **Methylene chloride** in fruit juice
- **Pentane** in decafs
- **Propyl alcohol** in bottled water, commercial fruit juices, commercial beverages.
- **Toluene** and **xylene** in carbonated drinks
- **Trichloroethane** (TCE), **TC Ethylene** in flavored foods
- **Wood alcohol** (methanol) in carbonated drinks, diet drinks, herb tea blends, store-bought water, infant formula

If you allowed a tiny drop of kerosene or carpet cleaning fluid to get into your pet's food every day, wouldn't you expect your pet to get sick? Why would you not expect to be sick with these solvents in your daily food? I imagine these solvents are just tiny amounts, introduced by sterilizing equipment, the manufacturing process, and adding flavor or color. Flavors and colors for food must be extracted somehow from the leaves or bark or beans from which they come. But until safe methods are invented, such food should be considered unsafe for human consumption (or pets or livestock!).

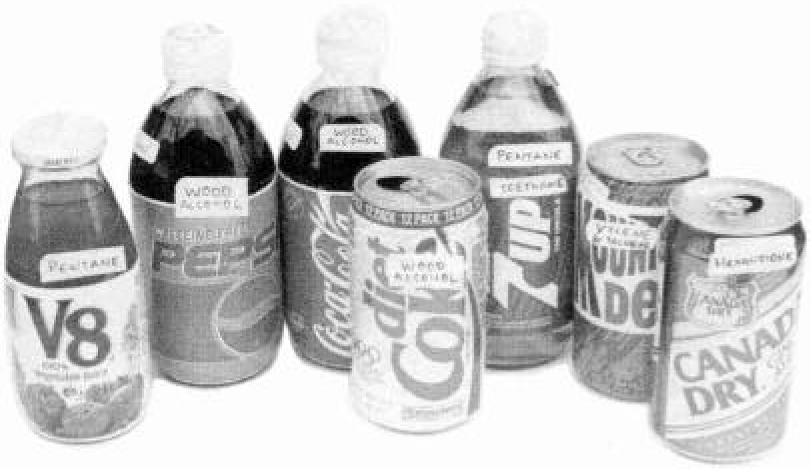

Fig. 67 Some unsafe beverages.

Food Preparation

Cook your food in glass, enamel, ceramic or microwavable pots and pans. Throw away all metal ware, foil wrap, and metal-capped salt shakers since you will never use them again. If you don't plan to fry much (only once a week), you might keep the Teflon™ or Silverstone™ coated fry-pan, otherwise get an enamel coated metal pan. Stir and serve food with wood or plastic, not metal utensils. If you have recurring urinary tract infections, you should reduce your metal contact even further; eat with plastic cutlery. Sturdy decorative plastic ware can be found in hardware and camping stores. Don't drink out of styrofoam cups (styrene is toxic). Don't eat toast (many toasters spit tungsten all over your bread and make benzopyrenes besides). Don't buy things made with baking powder (it has aluminum) or baked in aluminum pans. Choose goods made with baking soda and sold in paper or microwavable pans. Don't run your drinking water through the freezer or fountain or refrigerator. Don't heat your water in a coffee maker or tea kettle. Don't use a plastic thermos jug (the plastic liner has lanthanides) the inside must be glass. Don't drink from a personal water bottle (it begins to breed bacteria) unless you sterilize it daily.

Why are we still using stainless steel cookware when it contains 18% chromium and 8% nickel? Because it is rustproof and shiny and we can't see any deterioration. **But all metal seeps!** Throw those metal pots away. Get your essential minerals from foods, not cookware.

Never, never drink or cook with the water from your hot water faucet. If you have an electric hot water heater the heating element releases metal. Even if you have a gas hot water heater, the heated water leaches metals or glues from your pipes. If your kitchen tap is the single lever type, make sure it is fully on cold for cooking. Teach children this rule.

Food Guidelines

It is impossible to remember everything about every food, but in general do not buy foods that are highly processed. Here are a few foods; see if you can guess whether they should be in your diet or not.

breads	Yes, but only from a bakery, and never wrapped in plastic.
toast	No. It has benzopyrene and tungsten. Yes, if made on a cookie sheet or in a frying pan.
cheese	Yes, if used in baked dishes.
chicken	Only if cooked for 20 minutes at boiling point, as in soup, or canned (never prepare raw chicken yourself).
wine with dinner	No.

peanut butter	Yes if you grind it yourself and add ¼ tsp. vitamin C powder as you grind.
cottage cheese	No, it can't be sterilized easily.
desserts	Yes, but again, only if flavored with safe extracts.
rice	Yes, if vitamin C is added before cooking. Use white only, brown is too moldy.
pasta	Yes, with homemade sauce and vitamin C.
Jell-O™	No, it has artificial flavor and color.
egg dishes	Yes, but not “imitation”, cholesterol-free or cholesterol-reduced varieties.
fish, seafood	Yes!
soy foods (tofu)	No. It's the extensive processing that taints it.
soup	Yes, if seasoned only with herbs (no bouillon cube).
sugar	Yes, turbinado or brown if treated with vitamin C.
herb tea	Yes, if not in a bag and not in a mixture of herbs.
cheesecake	Yes.

Fig. 68 Some good foods.

Choose brands with the shortest list of ingredients. Alternate brands every time you shop.

Fig. 69 All breads I tested had mold if they were in plastic.

Dining Out

Restaurants (excluding fast food) are generally quite safe to eat at. Here are some do's and don'ts;

Do carry your own aluminum-free salt and vitamin C powder with you.

Do ask for plastic cutlery.

Do drink the water if from the tap.

Do ask for boiled, not just steamed, milk.

Don't eat or drink from styrofoam. If getting food "to go," get it in clear plastic containers, or ask them to line the styrofoam container with paper or plastic wrap, and line the styrofoam cup with a plastic baggy.

Don't use their ketchup and condiments (they have been standing out too long).

Here is a list of things that are generally safe to order:

pancakes, French toast, waffles	Don't use their imitation syrup (has benzoate), use honey instead.
eggs	Any style except soft boiled and scrambled. The white should be solid.
hash browns	If lightly fried, not deep fried.
soup	Only if nothing else is available. (It probably came in a can and was cooked in an aluminum pot and is full of aluminized salt.)
vegetarian sandwiches	But no soy products (too processed).
baked or boiled potatoes	Use only cheese sauce, bring your own salt, don't eat the skin.
cooked vegetables	Broccoli, Brussels sprouts, beets, corn, squash, and so forth.
vegetable salads	Don't eat the croutons, bacon bits, and anything that doesn't look fresh.
vegetarian dishes	But no soy ingredients and or sauces. Fresh ketchup OK.
bread and biscuits	White only, not toasted, not “cholesterol-free” varieties.
fish and seafood	Anything but deep fried (the oil may have benzene) is fine: baked, steamed, fish cakes, seafood cocktails, etc.
Mexican food	Any of the numerous baked dishes.
Chinese food	Except dishes with tofu or MSG.
fruit cup	With honey and cinnamon.
fruit pies, cobblers	But not with ice cream (every flavor has benzene).
lemon or lime meringue pie	Indulge yourself.

Fig. 70 Good restaurant foods.

As you see your symptoms disappear, one after another, you will feel the magic of healing. Many sick persons have 50 or more symptoms to start out! They could fill two sheets of paper, one symptom to each line. It can be quite shocking to see a list of all your symptoms.

Sometimes a new symptom appears as fast as an old one disappears. The coincidence makes it tempting to believe that one symptom turns into a different one. But it is not so. If a new symptom appears, it is because another pathogen has become activated due to a new toxin. Try to identify the new item. Stop using any new food, supplement, or body product, even if it is a health variety, and see if it goes away.

Body Clean-up

We are living in a very fortunate time. We are not expected to all look alike! The 60's brought us this wonderful freedom. Freedom to dress in a variety of styles, use make-up or no make-up, jewelry or no jewelry, any kind of hair style, any kind of shoes.

You will need to go off every cosmetic and body product that you are now using. Not a single one can be continued. They are full of titanium, zirconium, benzalkonium, bismuth, antimony[30], barium[31], strontium[32], aluminum, tin, chromium, not to mention pollution solvents such as benzene and PCBs.

Do not use any commercial salves, ointments, lotions, colognes, perfumes, massage oils, deodorant, mouthwash, toothpaste, even when touted as "herbal" and health-food-type. See *Recipes* for homemade substitutes.

---

[30] Breast cancer cases show titanium, zirconium, benzalkonium, bismuth and antimony accumulation in the breast.

[31] Barium is described in the *Merck Index* as "Caution: All water or acid soluble barium compounds are POISONOUS." 10th ed. p. 139 1983.

[32] This element goes to bones.

People are trying desperately to use less toxic products. They seek health for themselves. So they reach for products that just list herbs and other natural ingredients. Unfortunately, the buyers are being duped. The Food and Drug Administration (FDA) requires all body products to have sufficient antiseptic in them. Some of these antiseptics are substances you must avoid! But you won't see them on the label because manufacturers prefer to use quantities below the levels they must disclose. And by using a variety of antiseptics in these small amounts they can still meet sterility requirements. The only ingredient you might see is “grapefruit seed” or similar healthy-sounding natural antiseptic. This is sad for the consumer of health food varieties.

- I have seen rocks sold as “Aluminum-Free Natural Deodorant”. You rub the rock under your arms. It works because the rock is made of magnesium-aluminum-silicate.
- Men's hair color has lead in it.
- Lipstick has barium, aluminum, titanium.
- Eye pencil and shadow have chromium.
- Toothpaste has benzene, tin, and strontium.
- Hair spray has propyl alcohol and PCBs. BEWARE! Stop using it today.
- Shampoo, even health varieties, has propyl alcohol! BEWARE! Stop using it today.
- Cigarettes have lead, mercury, nickel and *Tobacco Mosaic virus*.
- Chewing tobacco has ytterbium

Some of the unnatural chemicals listed are present because of residues in the manufacturing process, but others you will actually see listed on the label!

Propyl alcohol and wood alcohol are present because the tubing used to fill the bottles is sterilized and cleaned with them. Ice cream machines are “oiled” with a gel containing petroleum products. This could explain why I always find benzene in ice cream.

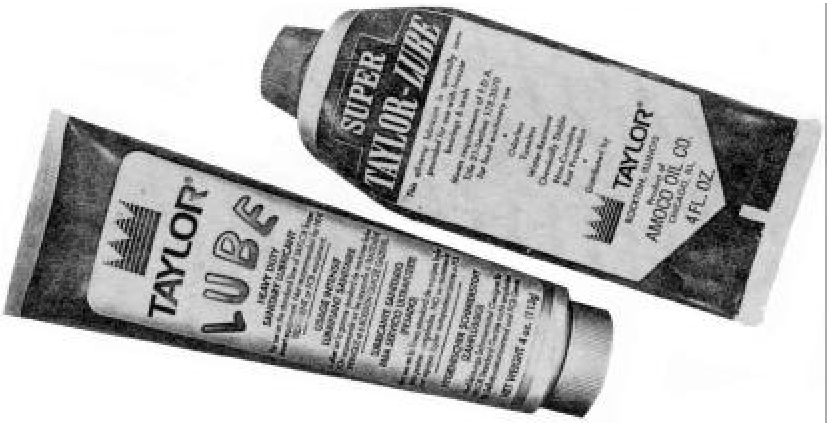

Fig. 71 Examples of commercial "food lube."

How can propyl alcohol in shampoo get into your body in significant amounts? The skin is more absorbent than we realize, and time and time again I see cancer victims who have gone off every body product except their favorite shampoo. They harbor propyl alcohol until they make that final sacrifice. It is better to switch shampoos than to not need any due to radiation and chemotherapy!

See *Recipes* for easy-to-make, natural cosmetics. But you might consider just stopping them all. Especially if you're going on vacation.

Use nothing that you wouldn't use on a new-born baby. This is a permissive age. You will be the only one feeling "naked." Others won't even notice. Don't forget advertising is aimed at you, even if other people's eyes are not!

Don't even use soap unless it is homemade soap (see *Recipes*) or borax straight from the box. Borax was the traditional pioneer soap. It is antibacterial and can be made into a solution. It is also a water softener and is the main ingredient in non-chlorine bleach. Borax can remove grease, too, and some stains.

But even borax is not natural to your body and it is therefore wise to use as little as necessary. See *Recipes* for anti-bacterial borax soap.

Dishes and clothing are primary source of PCBs.

Fig. 72 Detergents with PCBs.

MULE 20 TEAM
BORAX
100 YEARS

homemade soap pure borax

Fig. 73 Safe soaps.

Don't use toothpaste, not even health-food varieties. To clean teeth, use plain water or chemically pure baking soda (see *Sources*)—but dissolve it in water first, otherwise it is too abrasive. Or brush with hydrogen peroxide food grade, not the regular variety (see *Sources*). Don't use floss; use 2 or 4-pound monofilament fish line. Floss has mercury antiseptics (with thallium pollution!). Throw away your old toothbrush—solvents don't wash away.

Don't use mouthwash. Use saltwater (aluminum-free salt) or food grade hydrogen peroxide (a few drops in water).

Don't use hair spray.
Don't use massage oils of any kind. Use olive oil.
Don't use bath oil. Take showers, not baths, if you are strong enough to stand. Showers are cleaner.
Don't use perfumes or colognes.
Don't use commercial lotions or personal lubricants.

Stop Using Supplements

Stop using your vitamin supplements. They, too, are heavily polluted. This is the saddest, most tragic part of your instructions. I have found solvents, heavy metals and lanthanides in 90% or more of the popular vitamin and mineral capsules and tablets I test. These substances will do more harm in the long run than the supplement can make up for in benefits.

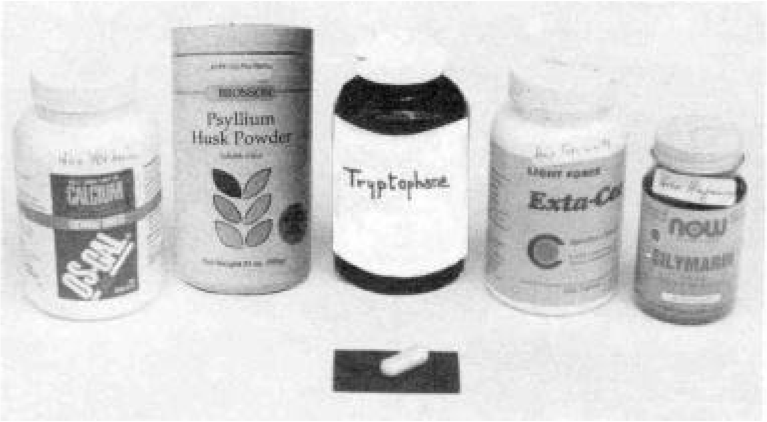

The capsule in the foreground is a notorious tryptophane capsule. It had the following pollutants: PCBs, mercury, ruthenium, thulium, strontium, praseodymium, aluminum, benzalkonium.

Fig. 74 Some polluted supplements.

Most of the varieties of vitamin C that I have tested are polluted with thulium! Until all vitamins and minerals and other food supplements have been analyzed for pollutants, after they are encapsulated or tableted, they are not safe. We need more disclosure on our products. No manufactured product is pure. We can't expect that. But at least we should be able to tell what impurities we are getting, and how much.

It is possible to do detailed analysis of foods or products at a reasonable price. Look at the bottle of common table salt, sodium chloride, that is used by beginning chemistry students to do experiments. It must be thoroughly analyzed for them because minute impurities affect their results. (Those minute impurities, like lead, affect you, too.) Look at the label on the bottle in the picture. Even after all these tests, the cost of laboratory salt is only $2.80 per pound.[33]

S-271
Certified A.C.S.
Sodium
Chloride
Fisher Scientific Company

Fig. 75 Pure salt.

It is most important not to be fooled by ingredient claims, like “made from organically grown vegetables”. Sure that's great, but the analysis *I* trust would be done on the final, cleaned, cooked and packaged product on the shelf. The package is a major unlisted ingredient.

Toxic solvents like decane, hexane, carbon tetrachloride and benzene will get more flavor or fat or cholesterol out of things than metabolizable grain alcohol. Of course, the extraction process calls for washing out the solvent later. But it can't all be washed out, and a detailed analysis on the final product would give the public the information they need to make informed choices.

---

[33] You will pay about $8.00 per pound (Spectrum Chemical Co.) for USP (United States Pharmaceutical) grade. But the same analysis is done on the cheaper grades, and my point is that the analysis is cost effective enough that it should be done on our daily foods.

All supplements must be tested for purity by yourself. If this cannot be done,
don't take them.

Polluted supplements do much more harm than good. Get your super-nutrition by juicing vegetables of all kinds and making herbal teas.

Safe Supplements

There are, no doubt, lots of safe supplements to be had. The problem is knowing which they are. The nature of pollution is such that one bottle might be safe, while another of the same brand is not. In view of this, as I found a polluted bottle, I stopped using any more of that brand. That is why I am reduced to recommending only the ones in the *Sources* at this time.

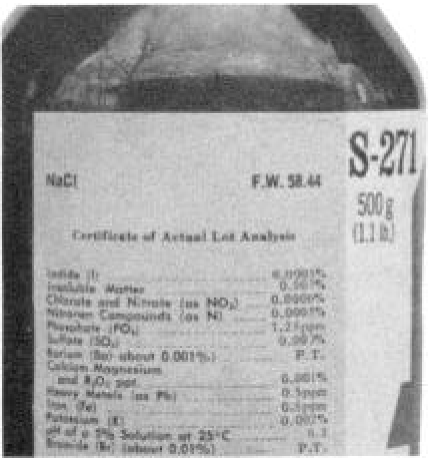

Common salt for student use is thoroughly analyzed for pollution. The label gives you the final "Actual Lot Analysis" of the product. It is not expensive.
Fig. 76 Laboratory salt label.

Vitamin C, in crystal form, is a must in your lifestyle. It helps the liver, and possibly other organs, detoxify things. It also helps retard mold, and perhaps even destroys aflatoxins! Keep some next to your refrigerator so it is handy when you put away groceries. Add 1/8 tsp. to maple syrup, vinegar, cooked cereal, fruit juice, leftovers. Have ¼ tsp. (1 gram) with each meal.

Vitamin B_2, 100 mg. This is the vitamin that helps detoxify benzene! Take one to three a day. While recovering from AIDS you need 3 tablets three times a day.

Vitamin B_6, 250 mg, and B-complex, 50 mg. Undoubtedly these help the liver and kidneys in many ways. One a day.

Magnesium oxide, 300 mg., is another must. Take one or two a day. It is a major mineral; all of our cells need lots of it. Only leafy vegetables provide it.

Hydrogen peroxide, food grade. It is advantageous to kill bacteria and viruses to some extent every day. Hydrogen peroxide lets you do this. It should never come in contact with metal, including its container or metal tooth fillings. If you get a few drops on your skin it may turn white and sting, but does no harm, so simply wash it off. Instructions for its use come with the product.

Herbs. These are excellent supplements, both in bulk and capsules, but not extracts, concentrates, or concoctions. There are many books that describe their uses.

Thioctic acid or lipoic acid. Presumably this chelates (traps and prepares for elimination) heavy metals, and helps the liver in detoxifying obscure and deadly poisons. Everyone would benefit from 100 mg. per day. I find it outstanding, and give it to many ill persons, even when not mentioned in the case histories. It comes as a 100 mg. capsule (see *Sources*). I use it at doses from one capsule, 3 times a day, to 2 capsules, 5 times a day. I have seen no side effects at these dosages, even in very sick persons.

Lugol's Iodine Solution (see *Recipes*) is old-fashioned "iodine." Iodine has a distinctive trait: it hangs up on anything and everything. In fact, it attaches itself so quickly we consider everything it touches as "stained." This is just the property we want to make it safe for use. The amount you use is immediately hung up, or attached, to your mucous and can not be quickly absorbed into the blood or other organs. It stays in the stomach. And for this reason it is so useful for killing vicious bacteria like *Salmonella*.

Do not take Lugol's iodine if you know you are allergic to iodine. It could be fatal.

Six drops of Lugol's solution can end it all for *Salmonella*. If you have gas and bloating, pour yourself ½glass of water. Add 6 drops of Lugol's (not more, not less), stir with wood or plastic, and drink all at once. The action is noticeable in an hour. Take this dose 4 times a day, after meals and at bedtime, for 3 days in a row, then as needed. This eradicates even a stubborn case of *Salmonella*.

Notice how calming 6 drops of Lugol's can be, soothing a manic stage and bringing a peaceful state where anxiety ruled before.

Lugol's is perfectly safe (if not allergic) to take day after day, when needed, because of its peculiar attaching property. It arrives in the stomach, reattaches to everything in proximity. Doomed are all *Salmonellas*; doomed also are eggs of parasites that might be in the stomach (cysts).

Naturally, one would not leave such medicine within the reach of children. Also, one would not use anything medicinal, including Lugol's unless there were a need, like cancer, AIDS, or bowel disease. When the gas and bloating problem has stopped, stop using Lugol's. If one or two doses of Lugol's cures the problem, stop. Store it in a perfectly secure place. In the past, 2/3 of a teaspoon (60 drops) of Lugol's was the standard dose of iodine given to persons with thyroid disease. Six drops is small by comparison.

Turmeric and **fennel** are herbs also used as cooking spices. They can eradicate invasive *E. coli* and *Shigella* bacteria! They are completely harmless, and are part of the Bowel Program.

Other supplements. The concept of supplementing the diet is excellent, but the pollution problem makes it prohibitive. Use only supplements and brands recommended in *Sources*, although the best approach is to test them yourself with your Syncrometer. I can't guarantee the brands in *Sources* will stay pure. If in doubt, leave it out.

Home Clean-up

This is the easiest task because it mostly involves throwing things out. Hopefully your family and friends will jump to your assistance.

- The basement gets cleaned.
- The garage gets cleaned.
- Every room in the house gets cleaned.

Your Basement

To clean your basement, remove all paint, varnish, thinners, brush cleaners, and related supplies. Remove all cleaners such as carpet cleaner, leather cleaner, rust remover. Remove all chemicals that are in cans, bottles or buckets.

You may keep your laundry supplies: borax, washing soda, white distilled vinegar, bleach and homemade soap. You may keep canned goods, tools, items that are not chemicals. You may move your chemicals into your garage. Also move any car tires and automotive supplies like waxes, oil, transmission fluid, and the spare gas can (even if it is empty) into your garage or discard them.

Seal cracks in the basement and around pipes where they come through the wall with black plastic roofing cement. In a few days it will be hard enough to caulk with a prettier color. Spread a sheet of plastic over the sewer or sump pump.

Your Garage

Do you have a garage that is a separate building from your home? This is the best arrangement. You can move all the basement chemicals into this garage. Things that will freeze, such as latex paint, you may as well discard. But if your garage is attached, you have a problem. Never, never use your door between the garage and house. Walk around the outside. Don't allow this door to be used. Tack a sheet of plastic over it to slow down the rate of fume entrance into the house. Your house acts like a chimney for the garage. Your house is taller and warmer than the garage so garage-air is pulled in and up as the warm air in the house rises. See the drawing.

In medieval days, the barn for the animals was attached to the house. We think such an arrangement with its penetrating odors is unsavory. But what of the gasoline and motor fumes we are getting now due to parked vehicles? These are toxic besides! This is even more medieval.

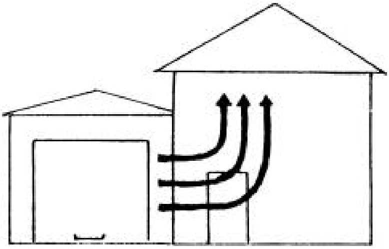

Fig. 77 Garage fumes.

If your garage is under your house, you cannot keep the pollution from entering your home. In this case, leave the cars and lawnmower outside. Remove cans of gasoline, solvents, etc. Put up a separate shed for these items.

Special Clean-up for Freon (CFCs)

Because I consider Freon to be the top health hazard in our home, I recommend turning in your refrigerator for a non-CFC variety. Dispose of all spare units. Remove window air conditioners or test the dust in your home (page 485) for Freon. Have your car's air conditioning system checked. Dispose of old pressurized cans. Even one whiff is too much. It never leaves the body because the body has no detoxification methods for it!

Only one useful reaction with Freon comes to mind. Freon is thought to be responsible for the ozone “hole” at the South Pole. Would Freon react with ozone supplied to your body and thereby become biodegradable? Indeed, it does! But only if you drink it as ozonated water. Other ozone routes, as intravenous or rectal, have not been observed to be as effective.

If you are following your progress with the Syncrometer, you will see that Freon now appears in the liver for the first time. (Before this, it was marooned in the parathyroids, thymus, and other organs.) You may also detect a feeling like indigestion. You must come to the assistance of your liver. Even ozonated Freon is extremely burdensome to the liver.

A combination of herbs (Liver Herb Drink in *Recipes,* page 552) rescues the liver from its plight, and prevents the indigestion. After drinking liver herbs you will see that the Freon now appears in the kidneys. Yet it is marooned there unless you assist them. Take the kidney cleanse to assist the kidneys so they can finally excrete the Freon into the urine.

It's an elaborate detoxifying program and usually takes six to eight weeks to get most of the Freon out. Afterward, continue the programs at one fourth dosage for half a year.

Forane is one of the new refrigerants. Although toxic, at least I observe it in the liver directly, suggesting that your body is capable of handling it. Remember your new refrigerator will still be using a toxic coolant, and it would be best to keep it outside or at least vented to the outside.

Special Clean-up for Fiberglass

Fiberglass insulation has microscopically small bits of glass that are free to blow into the air. When house drafts pull it into the air you will inhale them. They cut their way through your lungs and organs like millions of tiny knives, spreading through your body, since there is no way out for them. You smell nothing and feel nothing. This makes it a very sinister poison. Your body, though, recognizes these sharp, pointed bits and tries to stop their spread by sequestering them in cysts.

Most solid malignant tumors contain fiberglass or asbestos, another glass-like particle. In nearly all cases a hole can be found in the ceiling or walls, leading to fiberglass insulated parts of the house. When these holes are sealed in an air-tight manner the house air no longer is positive for fiberglass. Covering with paneling is not sufficient. Check your dwelling for uncovered fiberglass. Repair immediately. Search for small screw holes intended for pictures, or electric outlet plates that are missing.

Also remove fiberglass jackets from water heater and fiberglass filter from furnace. Replace with foam or carbon. Best of all, hire a crew to remove it all from your home, and replace insulation with blown-in shredded paper or other innocuous substance.

Never build a new house using fiberglass for any purpose.

Special Clean-up for Asbestos

The biggest source of asbestos is not building materials! It is the clothes dryer belt and hair dryer! To be safe, remove the belt from your dryer and check to see if it says “Made in USA” on the belt itself. If so, it is OK. If not, it is imported, and probably contains asbestos. Exchange it for a USA belt (see *Sources*).

Hair dryers, too, may be imported and shed asbestos. It is especially hazardous to be aiming a stream of hot asbestos right at your face! If you can't find a safe model (see *Sources*), or are unsure, don't use any. If you have cancer or are ill, no one in the house should use an unsafe hair dryer.

Turn off radiators and electric heaters and cover them with big plastic garbage bags, or paint them, or remove them. They give off asbestos if their paint is old.

Your House

To clean the house, start with the bedroom. Remove everything that has any smell to it whatever: candles, potpourri, soaps, mending glue, cleaners, repair chemicals, felt markers, colognes, perfumes, and especially plug-in air “fresheners”. Store them in the garage, not the basement. Since all vapor rises, they would come back up if you put them in a downstairs garage or basement.

Do not sleep in a bedroom that is paneled or has wallpaper. They give off arsenic and formaldehyde. Either remove them or move your bed to a different room. Leave the house while this is being done. If other rooms have paneling or wallpaper, close their doors and spend no time in them.

Next clean the kitchen. Take all cans and bottles of chemicals out from under the sink or in a closet. Remove them to the garage. Keep only the borax, washing soda, white distilled vinegar and homemade soap. Use these for all purposes. For exact amounts to use for dishwasher, dishes, windows, dusting, see *Recipes*. Remove all cans, bottles, roach and ant killer, moth balls, and chemicals that kill insects or mice. These should not be stored anywhere. They should be thrown out. Remember to check the crawl space, attic and closets for hidden poisons also.

To keep out mice, walk all around your house, stuffing holes and cracks with steel wool. Use old-fashioned mouse traps. For cockroaches and other insects (except ants) sprinkle handfuls of boric acid[34] (not borax) under your shelf paper, behind sink, stove, refrigerator, under carpets, etc. Use vinegar on your kitchen wipe-up cloth to leave a residue that keeps out ants. Do this regularly. To wax the floor, get the wax from the garage and put it back there. A sick person should not be in the house while house cleaning or floor waxing is being done.

Remove all cans and bottles of "stuff" from the bathroom. The chlorine bleach is stored in the garage. Someone else can bring it in to clean the toilet (only). Leave only the borax soap, homemade soap, and grain alcohol antiseptic. Toilet paper and tissues should be unfragranced, uncolored. All colognes, after shave, anything you can smell must be removed. Family members should buy unfragranced products. They should smoke outdoors, blow-dry their hair outdoors or in the garage, use nail polish and polish remover outdoors or in the garage.

Do not keep new foam furniture in the house. If it is less than one year old, move it into the garage until you are well. It gives off formaldehyde. So does new clothing; it is in the sizing. Wash all new clothes before wearing. If you have a respiratory illness, move all the clothes in the clothes closet out of your bedroom to a different closet.

Do not use the hot water from an electric hot water heater for cooking or drinking. It has tungsten. Do not drink water that sits in glazed crock ware (the glaze seeps toxic elements like cadmium) like some water dispensers have. Do not buy water from your health food store that runs through a long plastic hose from their bulk tank (I always see cesium picked up from flexible clear plastic). Also ask them how and when they clean their tank. Best is to observe that it is done with non-toxic methods.

---

[34] Boric acid is available by the pound from farm supply stores and from Now Foods. Because it looks like sugar keep it in the garage to prevent accidental poisoning.

If your house is more than 10 years old, change all the galvanized pipe to PVC plastic. Although PVC is a toxic substance, amazingly, the water is free of PVC in three weeks! If your house has copper pipes don't wait for cancer or schizophrenia to claim a family member. Change all the copper pipe to PVC plastic immediately. If the pipes are not accessible, ask a plumber to lay an extra line, outside the walls. This is less expensive, too.

If you have a water softener, by-pass it immediately and replace the metal pipe on the user side of the softener tank. Softener salts are polluted with strontium and chromate; they are also full of aluminum. The salts corrode the pipes so the pipes begin to seep cadmium into the water. After changing your pipes to plastic, there will be so little iron and hardness left, you may not need a softener. If the water comes from a well, consider changing the well-pipe to PVC to get rid of iron. While the well is open, have the pump checked for PCBs. Call the Health Department to arrange the testing. If you must have softening after all this, check into the new magnetic varieties of water softener (although they only work well when used with plastic plumbing).

The cleanest heat is electric. Go total electric if possible. If you must stay with gas, have a furnace repair person check your furnace and look for gas leaks before the heating season starts. Don't call the gas company even though it is free. The gas company misses 4 out of 5 leaks! The Health Department does not miss any; call them! House builders and contractors are also reliable in their gas leak detection.

Unnatural Chemicals

And where I found them...

ARSENIC

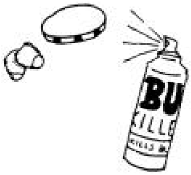

in ant & roach hives,
grains of pesticide

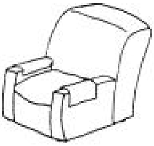

in carpet & furniture "treated"
for stain resistance

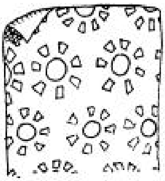

in wallpaper

BARIUM

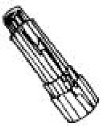

in lipstick

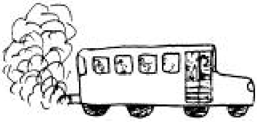

in bus exhaust

MOLYBDENU

in "molys"

COBALT

NEW!
WITH
MORE
BLUE
GRAINS

in laundry
detergent

in dishwater
detergent

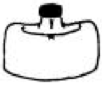

in skin bracer

in mouthwash

ANTIMONY

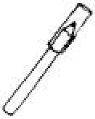

in eye liner

CADMIUM and COPPER

in water running
through old metal pipes

TITANIUM

in face powder &
other powders, and
in metal dental ware

PCB's

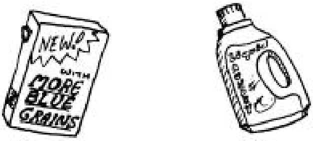
NEW!
WITH
MORE
BLUE
GRAINS

`in regular and`
`health store detergents`

LEAD

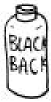

`in men's hair`
`color restorer`

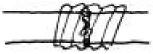

`in solder at joints`
`of copper pipes`

CHROMIUM

`in eyebrow`
`pencil`

WATER
SOFTENER
SALT

`in water`
`softener salts`

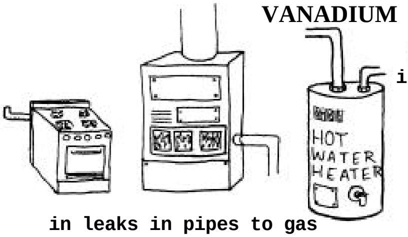
HOT
WATER
HEATER

`in leaks in pipes to gas`
`stove, furnace, water heater`

VANADIUM

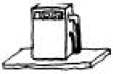

`in diesel fuel`

`in candles`
`(even when`
`they're not`
`burning)`

NICKEL

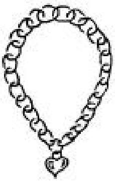

`in metal jewelry`
`worn on the skin`

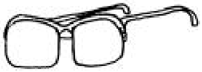

`in metal glasses frames`

`in metal`
`watch bands`

`in metal tooth fillings`
`and retainers`

FIBERGLASS

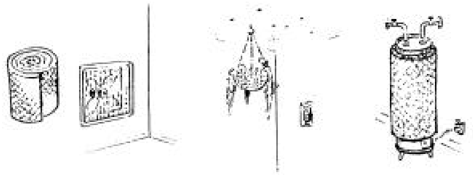

from insulation behind holes in ceiling or uncovered outlets, water heater jackets, stuffed around fans and air conditioners, insulation

CFC's (FREON)

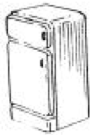

in refrigerators

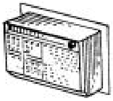

in air conditioners

in spray cans

MERCURY and THALLIUM

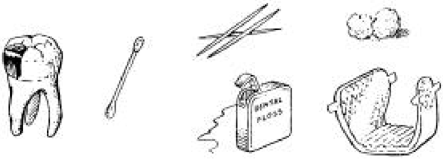
DENTAL FLOSS

in tooth fillings, sanitary napkins,
cotton balls, dental floss,
toothpicks, cotton swabs

THULIUM

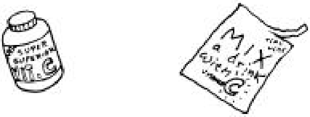
SUPER VITAMIN C
MIX a drink with VIT C

in most brands of
vitamin C I tested

DYSPROSIUM and LUTETIUM

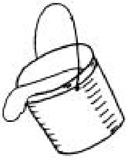

in paint, varnish,
shellac

HOLMIUM

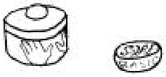

in hand cleaners

HAFNIUM

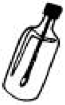

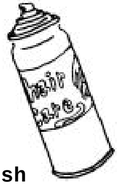

in nail polish
& hair spray

RHENIUM

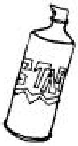

in spray starch

BISMUTH

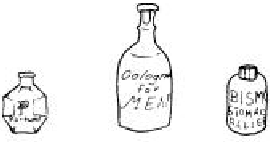
Cologne for MEN
BISM STOMACH RELIEF

in cologne and
stomach aids

CESIUM

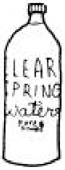
CLEAR SPRING water

in clear-as-glass
plastic

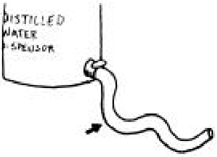
DISTILLED WATER DISPENSOR

after running through
long plastic hose

TIN and STRONTIUM

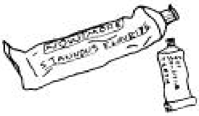

in toothpaste

YTTERBIUM, ERBIUM, TERBIUM

GREEN
ENERGY

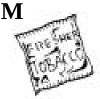
FRESHER
TOBACCO

in foil packaging

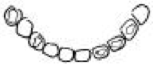

in plastic tooth fillings

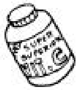

in vitamins and supplements

PRASEODYMIUM, NIOBIUM, NEODYMIUM, YTTRIUM

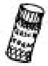

in over the counter drugs

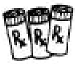

in prescription drugs

RADON, URANIUM, THORIUM

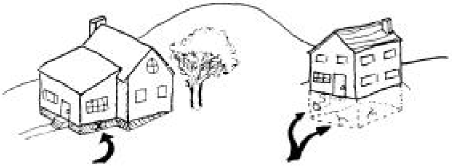

in crawlspace without open vents

in holes and cracks in basement

LANTHANUM

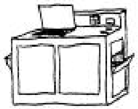

in duplicator and copier ink

BENZALKONIUM and ZIRCONIUM

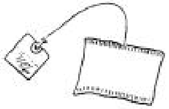

in tea bags

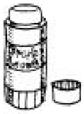

in deodorant

STANNOUS FLUORIDE

in toothpaste

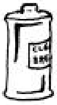

in mouthwash

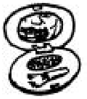

in cosmetics

ALUMINUM

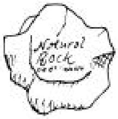
Natural
Rock

in "natural" deodorant advertised as aluminum-free

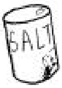

in salt

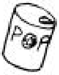

in cans

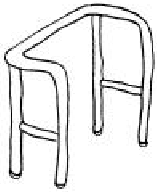

in walkers for the elderly

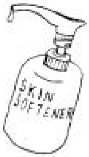
SKIN
SOFTENER

in lotions

TUNGSTEN

ELECTRIC
HOT
WATER
HEATER

corroded rod in
electric water heaters

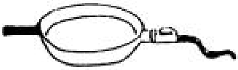

in electric
frying pans

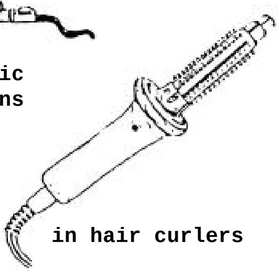

in hair curlers

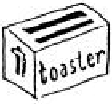

in toasters

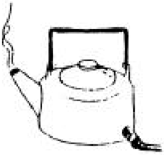

in tea kettles

BERYLLIUM

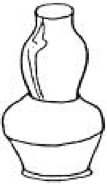

in hurricane
lamps

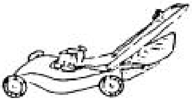

in lawn mowers

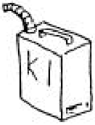

in kerosene

FORMALDEHYDE

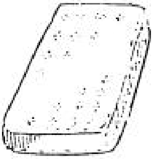

in foam
mattresses

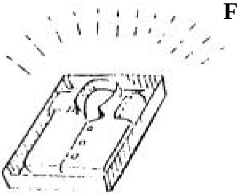

in new
clothing

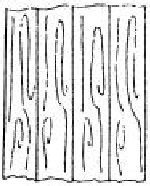

in paneling

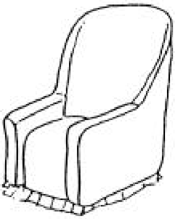

in foam chairs

ASBESTOS

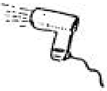

in hair blowers

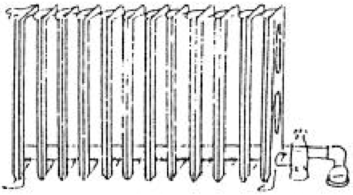

in radiator paint

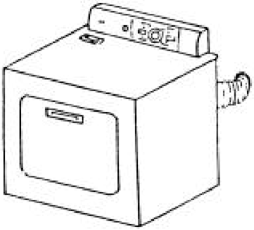

in clothes dryer belts

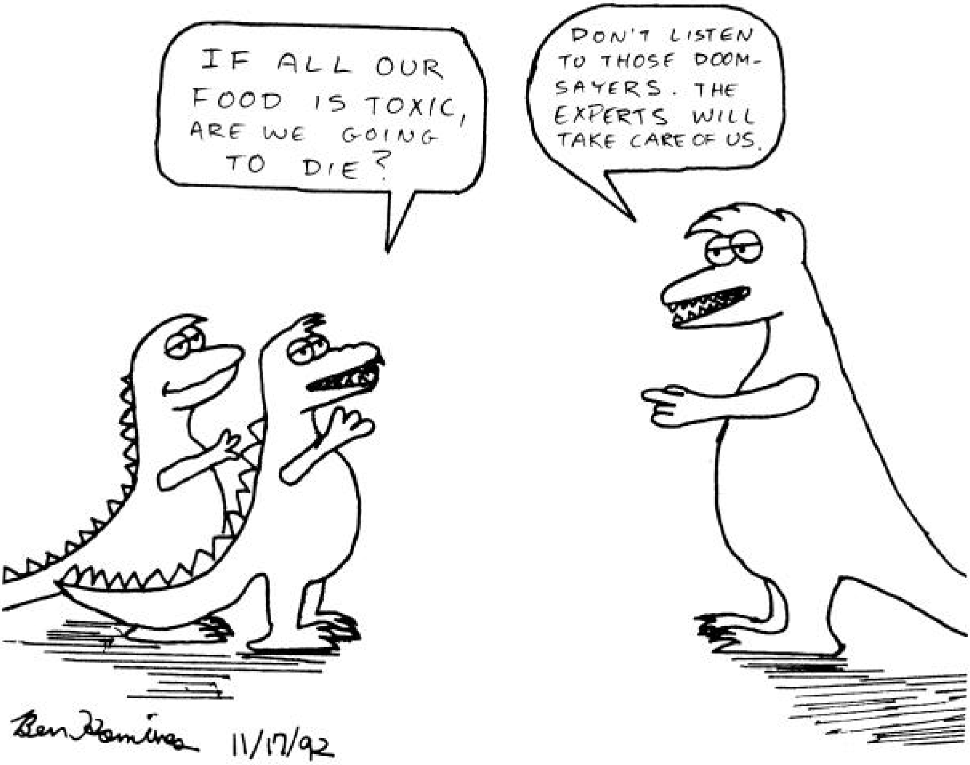
IF ALL OUR FOOD IS TOXIC, ARE WE GOING TO DIE?
DON'T LISTEN TO THOSE DOOM-SAYERS. THE EXPERTS WILL TAKE CARE OF US.
11/17/92

Fig. 78 Our future, unless we act.

Bioelectronics

The most important electronic device to make or buy is a **zapper**—a compact pulse generator operating from a common 9 volt battery whose output is about 30 KHz. It kills all parasites, bacteria, viruses, molds, and fungi even though their individual resonant frequencies are either higher or lower (50 KHz to 900 KHz). Building a zapper is described in an earlier chapter.

The next most useful device to have is a **Syncrometer.™** It lets you diagnose yourself and monitor your progress until you are cured. It consists of an audio oscillator circuit with your body as part of the circuit. Utilizing samples of parasites or pollutants, it lets you test for them in any product or body tissue. I include a design that you can make yourself.

A third very useful device is a **frequency generator**. You can use it to electrocute individual organisms, or together with the Syncrometer™ to find an organism's particular frequency. You need one that operates in the parasite, bacteria and virus ranges, from 50 KHz to 900 KHz. It must also be able to select a particular frequency, like 434 KHz, quickly and accurately. Frequency generators are available for as little as $300.00, but it is worth paying a little more to get a digital display of the frequency.

Making A Syncrometer

This is an *audio oscillator* circuit in which you include yourself by means of a handhold. You listen to the current in your circuit with a loudspeaker. Other oscillator circuits will work, too. A lot of fascinating opportunities present themselves with this concept.

I have previously published three ways to build the Syncrometer circuit.[35] Here is the circuit diagram:

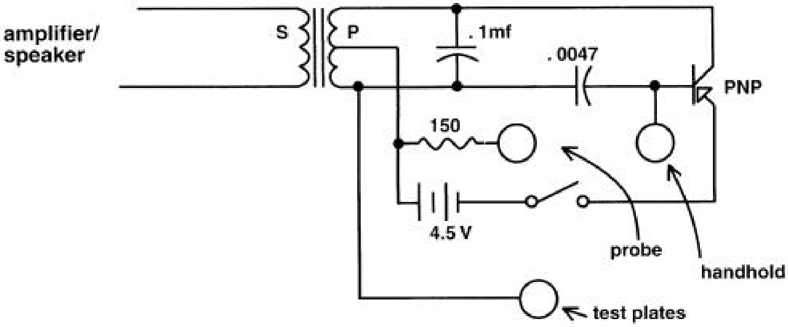
amplifier/
speaker
S
P
. 1mf
. 0047
PNP
150
4.5 V
probe
handhold
test plates

4700pf moly cap
.1 mfd 16v cap
2N2907 PNP
min output transf 8-900Ù
150 .25w res
3 AA batteries
speaker (like Radio
Shack 277-1008C)

Fig. 79 Syncrometer schematic.

If you are not an electronics enthusiast, you can still assemble a Syncrometer using a hobby kit. No soldering is required. Here is what you need:

Making a Hobby Kit Syncrometer

Item	**Radio Shack Cat. No.**
200 in One Electronic Project Lab by Science Fair	28-262
3 AA 1½ volt batteries	

---

[35] In *The Cure for all Cancers* and *The Cure for HIV/AIDS*.

Alligator clip test jumpers	You need 2.
Handhold. A four inch length of ¾ inch copper pipe, like for plumbing. These dimensions are critical to assure maximum skin contact.	
Probe. A banana plug.	Precision Mini-Hook Test Lead Set (contains two, you only need one) 278-1160A
Pencil, new.	

Fig. 80 Syncrometer parts list.

Build The Electrosonic Human in the *200 in One Electronic Project Lab*. It takes about 10 minutes.

> Later, when you use the probe to press against your knuckle you may find getting the right sound is painful. In this case try substituting the .005 microfarad capacitor for the .01 microfarad capacitor in the circuit.

Attach the Probe. The Archer Precision Mini-Hook Test Lead Set has a banana plug for the probe on one end and a mini-hook on the other end for easy attachment to the circuit. Tape a long, new pencil to the probe to make it easier to hold. Connect the Probe to middle post of the primary side of the transformer (it also connects to the negative battery post). You will not be using the two connections the instructions tell you to hold.

Attach the Handhold. Clip the Handhold to one end of an alligator clip test jumper, and clip the other end to the base (B) of the transistor used in the circuit.

Attach an alligator clip to the post of the transformer that connects to the two capacitors. This will go to the **test plates**.

Final test. Turn the control knob on and keep turning the potentiometer to nearly the maximum. (This reduces the resistance. The schematic has a 150 ohm fixed resistor in place of the potentiometer.) Make sure you have good batteries installed. Test the circuit by briefly touching the Probe to the Handhold. The speaker should produce a sound like popping corn. If it does not, check that your alligator clips are not bending the spring terminals so much that other wires attached there are loose. Finally, turn switches OFF.

Making Test Plates

This is the box you attach to the basic Syncrometer circuit. It has test plates to put your test substances and tissue samples on. The wiring in it is arranged so that you can test for a toxin in a product, as well as search in yourself. This means you can search for *Salmonella* in the milk or cheese you just ate, not just for *Salmonella* in your stomach.

Only if the resonant frequency of an item on one plate is equal to the resonant frequency of an item on the other plate will the entire circuit oscillate or resonate! This implies the two plates have something in common. By putting a known pure sample on one plate you can reliably conclude the other sample contains it if the circuit resonates.

You may build a test plate box into a cardboard box (such as a facial tissue box) or a plastic box. Here are the instructions for the cardboard box model.

Test Plates Parts List

- Stiff paper.
- Aluminum foil.
- A facial tissue box is easiest. A plastic project box, about 7” x 4” x 1½” makes a more durable product, but requires a drill, and you should discard any metal lid it comes with.

- 3 bolts (tapered heads) about 1 inch long, 1/8 inch diameter and 6 washers and nuts to fit.
- toggle switch with OFF-ON positions.
- Alligator clip test leads.

Test Plates Assembly

Cut two 3-1/2 inch squares out of stiff paper such as a milk carton. Cover them with 4½ inch squares of aluminum foil, smoothed evenly and tucked snugly under the edges. You have just made yourself a set of open capacitors. Turn the box upside down and draw squares where you will mount them at the ends of the box. Don't actually mount them, to save wear and tear on them, until the rest of the box is complete.

Mount the ON OFF switch on the front of the box, underneath the right hand plate. Line it up so ON is downward and OFF is up. (An electronics shop can determine this for you at the time of purchase.) Label the box with ON and OFF signs.

Two bolts will be reserved for the plates. The third bolt is used as a terminal where the current from the oscillator circuit will arrive. Make a hole on the side of the box, near the left hand plate and mount the bolt so it sticks half way inside and halfway outside the box. It does not matter whether the head is inside or outside. Tighten it there with a nut on each side of the box. Label it TERMINAL. It merely means connecting place.

Mark the center of each square that you drew and each capacitor you built. Pierce first with a pin; follow with a pencil until a round hole is made at the center. Mount each plate with a bolt, fastening it below with a nut. Washers are optional.

The left side connection (terminal) gets attached to the left plate (bolt) with an alligator clip. Use another clip to attach the same left plate (bolt) to the ON OFF switch (there are two connections, use either one). Finally attach the ON OFF switch connection you didn't use to the right plate (bolt). Make sure the connections at the switch are not touching each other; you might tape them to guard against this.

All these connections should be checked carefully to make sure they are not touching others accidentally. But if you leave the box open so you can see any problems and use clear tape around connections to prevent accidental touching to the wrong connection, it should work OK.

Finally, trace your current. It comes in from the Syncrometer at the main terminal on the left. It is brought to the left plate. When the switch is ON it is simultaneously brought to the right plate. Notice that the plates are not connected to anything else. They are simply capacitors, letting current in and out momentarily and at a rate that is set by the frequency of the oscillator circuit, about 1,000 hertz. This frequency goes up as the resistance (of the circuit or your body) goes down.

The probe and handhold allow you to include yourself in the Syncrometer circuit. You grasp these when testing. This makes you part of the circuit.

The speaker lets you “listen” to the current. As resistance drops, current goes higher and frequency goes up. As frequencies go higher in the circuit, pitch goes higher. You will be comparing the sound of a standard “control” current with a test current.

Using The Syncrometer

Fill a saucer with cold filtered tap water. Fold a paper towel four times and place it in this dish. It should be entirely wet.

Cut paper strips about 1 inch wide from a piece of white, unfragranced, paper towel. Dampen a paper strip on the towel and wind it around the copper pipe handhold to completely cover it. The wetness improves conductivity and the paper towel keeps the metal off your skin.

- Start with the test plate switch at OFF.
- Turn the control knob (potentiometer) on, and to near maximum.
- Touch each plate with the probe, while holding the copper pipe with one hand. Only the left plate should give you a sound from the speaker. Turn the test plate switch ON. Now both plates should give you a sound when the probe touches them.
- Turn the test plate switch OFF again.
- Pick up the handhold, squeeze it free of excess water.
- Pick up the probe in the same hand, holding it like a pen, between thumb and forefinger.

Dampen your other hand by making a fist and dunking your knuckles into the wet paper towel in the saucer. You will be using the area on top of the first knuckle of the middle finger or forefinger to learn the technique. Become proficient with both. Immediately after dunking your knuckles dry them on a paper towel folded in quarters and placed beside the saucer. The degree of dampness of your skin affects the resistance in the circuit and is a very important variable that you must learn to keep constant. Make your probe as soon as your knuckles have been dried (within two seconds) since they begin to air dry further immediately.

With the handhold and probe both in one hand press the probe against the knuckle of the other hand, keeping the knuckles bent. Press lightly at first, then harder, taking one half second. Repeat a half second later, with the second half of the probe at the same location. There is an additive effect and you get two chances to listen to the current. All of this takes less than two seconds. Don't linger because your body will change and your next probe will be affected.

Subsequent probes are made in exactly the same way. As you develop skill, your probes will become identical. Plan to practice for one or two hours each day. It takes most people at least twelve hours of practice in order to be so consistent with their probes that they can hear the slight difference when the circuit is resonant.

For reference you may wish to use a piano. The starting sound when you touch down on the skin should be F, an octave and a half above middle C. The sound rises to a C as you press to the knuckle bone, then slips back to B, then back up to C-sharp as you complete the second half of your first probe. If you have a multitester you can connect it in series with the handhold or probe: the current should rise to about 50 microamps. If you have a frequency counter the frequency should reach 1000 Hz. You should arrive at C-sharp just before the probe becomes painful.

Two things change the sound of the probes even when your technique doesn't change.

1. The patch of skin chosen for probing will change its properties. The more it is used, the redder it gets and the higher the sound goes when you probe. Move to a nearby location, such as the edge of the patch, when the sound is too high to begin with, rather than adjusting the potentiometer.
2. Your body has cycles which make the sound go noticeably higher and lower. If you are getting strangely higher sounds for identical probes, stop and only probe every five minutes until you think the sound has gone down to standard. This could take five to twenty minutes. Learn this higher sound so you can avoid testing during this period.

You may also find times when it is impossible to reach the necessary sound without pressing so hard it causes pain. You may adjust the potentiometer if that helps.

All tests are momentary.

This means less than one second. It is tempting to hold the probe to your skin and just listen to the sound go up and down, but if you prolong the test you must let your body rest ten minutes, each time, before resuming probe practice!

For our purposes, it is not necessary to locate acupuncture points.

Resonance

The information you are seeking is whether or not there is *resonance*, or *feedback oscillation*, in the circuit. If there is the test is **YES** (positive). You hear resonance by comparing the second probe to the first. You can never hear resonance on the first probe, for reasons that are technical and beyond the scope of this book. You are not merely comparing pitch in the two probes. During resonance a higher pitch is reached faster; it seems to want to go infinitely high.

Remember that more electricity flows, and the pitch gets higher, as your skin reddens or your body changes cycle. These effects are not resonance.

Resonance is a small extra hum at the high end of the probe. As soon as you hear it, stop probing. Your body needs a short recovery time (10 to 20 seconds) after every resonant probe. The longer the resonant probe, the longer the recovery time to reach the standard level again.

Using musical notes, here is a NO (negative result): F-C-B-C# (first probe) F-C-B-C# (compare, it is the same sound). Here is a YES (positive) result: F-C-B-C# (first probe) F-D (stop quickly because you heard resonance). In between the first and second probe a test substance will be switched in as described in lessons below.

It is not possible to produce a resonant sound by pressing harder on the skin, although you can make the pitch go higher. To avoid confusion it is important to practice making probes of the same pressure. (Practice getting the F-C-B-C# tune.)

Making Pure Water for Testing Purposes

Since the water you purchase is likely to have solvents in it and since your tap water may be polluted with heavy metals and since your (or a store's) filtration system may be clogged, it is important to make your own pure water.

Purchase a “filter pitcher” made of hard, opaque plastic, not the clear or flexible variety (see Sources). Fill the pitcher with cold tap water, only, not reverse osmosis, distilled, or any other water, since solvents do not filter out as easily as heavy metals. The filter should be made of carbon only. To make test substances, use fresh water in the pitcher and pour.

If your water has lead, copper or cadmium from corroded plumbing, the filter will clog in five days of normal use. So use this pitcher sparingly, just for making test substances and for operating the Syncrometer.

Lesson One

Purpose: To identify the sound of resonance in the circuit.

Materials: Potentized (homeopathic) solutions. Prepare these as follows: find three medium-sized vitamin bottles, glass or plastic, with non-metal lids. Remove any shreds of paper sticking to the rim. Rinse well with cold tap water. Then rinse again with filtered water.

Pour filtered water into the first bottle to a depth of about ½ inch. Add about 50 little grains of table salt using the tip of a plastic knife. This is a “pinch.” Replace the lid. Make sure the outside is clean. If not, rinse and dry. Now shake hard, holding it snugly in your hand. Count your shakes; shake 120 to 150 times. Use elbow motion so each shake covers about an eight inch distance. Shaken samples are different from unshaken ones, that's why this is so important. When done label the bottle on its side and lid: SALT #1. Wash your hands (without soap).

Next, pour about the same amount of filtered water into the second and third bottles. Open SALT #1 and pour a small amount, like 1/4 to 1/2 of a teaspoon (do not use a spoon) into the second and third bottles. Close all bottles. Now shake the second bottle the same as the first. Clean it and label it SALT #2. Do the same for the third bottle. Label it SALT #2 also and set aside for Lesson Four.

These two solutions have unique properties. SALT #1 always resonates. Use #1 to train your ear. SALT #2 shouldn't resonate. Use #2 to hear when you (your body's internal resistance) have returned to the standard level.

1. Turn the Syncrometer ON.
2. Place the SALT #2 bottle on the right test plate.
3. Start with the plate switch OFF.
4. Make your first probe (F-C-B-C#).
5. Flip the plate switch ON, taking only one half second. Brace your hand when switching so it is a fast, smooth, operation.
6. Make the second probe (F-C-B-C#). Total probe time is 2½ seconds. Count it out, "a thousand and one (done with first probe) a thou. (done with switching) a thousand and one (done with second probe)."
7. The result should be a NO (negative). If the second probe sounds even a little higher you are not at the standard level. Wait a few more seconds and go back to step 3.
8. If the first result was NO, remove SALT #2 and put SALT #1 on. Put the test plate switch back to OFF and repeat the test. This time the circuit was resonating. Learn to hear the

difference between the last two probes so that a resonant probe can be terminated early (reducing rest time).

9. The skin must now be rested. When SALT #1 is placed in the circuit there is always resonance whether you hear it or not. Therefore, always take the time to rest the skin.
10. How can you be sure that the skin is rested enough? Any time you want to know whether you have returned to the standard level, you may simply test yourself to SALT #2 (just do steps 3 through 6). While you are learning, let your piano also help you to learn the standard level (starts exactly at F). If you do not rest and you resonate the circuit before returning to the standard level, the results will become aberrant and useless. The briefer you keep the resonant probe, the faster you return to the standard level. Don't exceed one half second when probing SALT #1. Hopefully you will soon hear resonance within that time.

This lesson teaches you to first listen to the empty plate, then to SALT #2, to check for standard state. Then to compare the empty plate to SALT #1 to check for resonance. In later lessons we assume you checked for your standard level or are quite sure of it.

Practice hearing resonance in your circuit every day.

White Blood Cells

Checking for resonance between your white blood cells and a toxin is the single most important test you can make.

Your white blood cells are your immune system's first line of defense. In addition to making antibodies, interferon, interleukins, and other attack chemicals, they also “eat” foreign substances in your body and eliminate them. By simply checking your white blood cells for toxins or intruders you save having to check every other tissue in your body. Because no matter where the foreign substance is, chances are some white blood cells are working to remove it.

It took me two years to find this ideal indicator, but it is not perfect. **Tapeworms are a notable exception.** They can be encysted in a particular tissue which will test positive, while the white blood cells continue to test negative. Also, when bacteria and viruses are in their latent form, they do not show up in the white blood cells. Fortunately, in their active form they show up quite nicely. **Freon** is an example of a toxin that is seldom found in the white blood cells; but typically, the white blood cells are excellent indicators of toxins.

Making a White Blood Cell Specimen

Obtain an empty vitamin bottle with a flat plastic lid and a roll of clear tape. The white blood cells are not going into the bottle, they are going on the bottle. The bottle simply makes them easy to handle. Rinse and dry the bottle. Make a second specimen on a clean glass slide if available. Squeeze an oil gland on your face or body to obtain a ribbon of whitish matter (not mixed with blood). Pick this up with the back of your thumb nail. Spread it in a single, small streak across the lid of the bottle or the center of the glass slide. Stick a strip of clear tape over the streak on the bottle cap so that the ends hang over the edge and you can easily see where the specimen was put (see photo). Wipe the lid beside the tape to make sure all white blood cells are covered. For the slide, apply a drop of balsam and a cover slip (see *Sources*). Both types of preparation will give you identical results. The bottle type of white blood cell specimen is used by standing it on its lid (upside down) so that the specimen is next to the plate. The lid is used because it is flat, whereas the bottom of most bottles is not.

Lesson Two

Purpose: To add a white blood cell specimen to the circuit and compare sound.

Method:

1. Turn the Syncrometer ON.
2. Start with test plate switch OFF.
3. Place the white blood cell specimen on the left plate. Place some junk food in a plastic baggy on the right plate.
4. Eat some of the junk food.
5. After ½minute listen to the current. Flip the plate switch ON and listen again.
6. If the circuit is now resonating, the junk food is already in your white blood cells. It is toxic.

Take vitamin C and a B-50 complex to clear it rapidly; it may have had propyl alcohol or benzene in it. Test every 5 minutes afterward to see how long it takes to clear out.

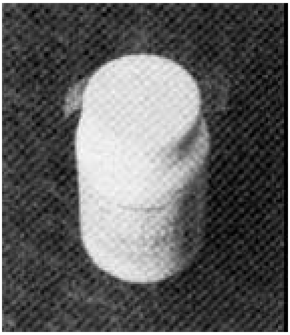

Fig. 81 Bottle with white blood cells taped on top.

Lesson Three

Purpose: To determine the purity of the filtered water you are making.

Method: Pour a few tsp. of filtered water into a bottle or plastic bag. Place your white blood cell specimen on one plate and the water sample on the other. Listen to your circuit. Taste your filtered water. After ½minute, listen to your circuit again, just as in Lesson Two. If it appears in your white blood cells at any time you can conclude the water is not pure. You must have pure water available to you before continuing.

Lesson Four

Purpose: To determine your percent accuracy in listening for resonance.

Materials: The SALT #1 and SALT #2 solutions you made for Lesson One.

Method: Move the SALT #1 and SALT #2 labels to the bottom of the bottles so you can not tell which bottle is which.

1. Turn the Syncrometer ON.
2. Start with the test switch OFF.
3. Mix the bottles up, select one at random, and place it on the right plate.
4. Listen to the current.
5. Flip the plate switch ON and make your second probe.
6. Resonance indicates a SALT #1, no resonance indicates SALT #2. Check the bottom. Remember to rest after the SALT #1, whether or not you heard resonance.
7. Repeat steps 3 through 5 a number of times. Work toward getting three out of three correct. Practice every day.

Trouble shooting:

a) If you repeat this experiment and you keep getting the same bottles "wrong", start over. You may have accidentally contaminated or mislabeled the outside of the bottle, or switched bottle caps.

b) If you get different bottles wrong each time, the plates may be contaminated. Wash the outside of the bottles and rinse with filtered water and dry. Wipe the plates very gently too, with filtered water and dry. Or replace the plates.

c) If all the bottles read the same, your filtered water is polluted. Change the filter.

Preparing Test Substances

It is possible to prepare dry substances for testing such as a piece of lead or grains of pesticide. They can simply be put in a plastic bag and placed on the test plate. However, I prefer to place a small amount (the size of a pea) of the substance into a ½ ounce bottle of filtered water. There will be many chemical reactions between the substance and the water to produce a number of test substances all contained in one bottle. This simulates the situation in the body.

Within the body, where salt and water are abundant, similar reactions may occur between elements and water. For example, a strip of pure (99.9% pure) copper placed in filtered water might yield copper hydroxide, cuprous oxide, cupric oxide, copper dioxide, and so forth. These may be similar to some of the reaction products one might expect in the body, coming from a copper IUD, copper bracelet or the copper from metal tooth fillings. Since the electronic properties of elemental copper are not the same as for copper compounds, we would miss many test results if we used only dry elemental copper as a test substance.

Impure Test Substances

It is not necessary to have pure test substances. For instance, a tire balancer made of lead can be easily obtained at an auto service station. Leaded gasoline and lead fishing weights also make good test substances for lead. There is a disadvantage, though, to using impure test substances. You are including the extra impurities in your test. If your lead object also has tin in it, you are also testing for tin. Usually, you can infer the truth by some careful maneuvering. If you have searched your kidneys for leaded gasoline, fishing weights and tire balancers and all 3 are resonant with your kidneys, you may infer that you have lead in your kidneys, since the common element in all 3 items is lead. (You will learn how to specify a tissue, such as your kidneys, later.)

Using pure chemicals gives you certainty in your results. You can purchase pure chemicals from chemical supply companies (see *Sources*). Your pharmacy, a child's chemistry set, a paint store, or biological supply company can also supply some.

The biggest repository of all toxic substances is the grocery store and your own home.

You can make test substances out of your hand soap, water softener salt, and laundry detergent by putting a small amount (1/16 tsp.) in a ½ounce glass bottle and adding about 2 tsp. filtered water. (Or for quick testing just put them dry or wet in a sealed plastic baggy.) Always use a plastic spoon.

Here are some suggestions for finding sources of toxic products to make your own toxic element test. If the product is a solid, place a small amount in a plastic bag and add a tablespoon of filtered water to get a temporary test product. For permanent use put it in a small amber glass bottle. If the product is a liquid, pour a few drops into a glass bottle and add about 2 tsp. filtered water. Keep all toxic substances in glass bottles for your own safety. Small amber glass dropper bottles can be purchased by the dozen at drug stores (also see *Sources*). Seal your test bottles with tape for safety and to prevent evaporation.

Aflatoxin: scrape the mold off an orange or piece of bread; wash hands afterward.

Acetone: paint supply store or pharmacy.

Arsenic: 1/16 tsp. of arsenate pesticide from a garden shop. A snippet of flypaper.

Aluminum: a piece of aluminum foil (not tin foil) or an aluminum measuring spoon.

Aluminum silicate: a bit of salt that has this free running agent in it.

Asbestos: a small piece of asbestos sheeting, an old furnace gasket, 1/4 inch of a clothes dryer belt that does not say "Made in USA", or a crumb of building material being removed due to its asbestos content (ask a contractor).

Barium: save a few drops from the beverage given clients scheduled for an X-ray. Lipstick that has barium listed in the ingredients.

Benzene: an old can of rubber cement (new supplies do not have it). A tsp. of asphalt crumbs from a driveway.

Beryllium: a piece of coal; a few drops of "coal oil" or lamp oil.

Bismuth: use a few drops of antacid with bismuth in it.

Bromine: bleached "brominated" flour.

Cadmium: scrape a bit off a galvanized nail, paint from a hobby store.

Cesium: scrape the surface of a clear plastic beverage bottle.

CFCs (freon): ask an electronics expert for a squirt from an old aerosol can that used freon as a cleaner. (Squirt into water, outdoors, put the water in a sample bottle.)

Chromate: scrape an old car bumper.

Cobalt: pick out the blue and green crumbs from detergent. A sample of cobalt containing paint should also suffice.

Chlorine: a few drops of pure, old fashioned Clorox.™

Copper: ask your hardware clerk to cut a small fragment off a copper pipe of the purest variety or a ¼ inch of pure copper wire.

Ergot: a teaspoon of rye grains, or rye bread. Add grain alcohol to preserve.

Ether: automotive supply store (engine starting fluid).

Ethyl alcohol (grain alcohol): the purest "drinking" alcohol available. Everclear™ in the United States, Protec™ (potable) in Mexico.

Fiberglass: snip a fragment from insulation.

Fluoride: ask a dentist for a small sample.

Formaldehyde: purchase 37% at a pharmacy. Use a few drops only for your sample.

Gasoline: gas station (leaded and unleaded).

Gold: ask a jeweler for a crumb of the purest gold available or use a wedding ring.

Kerosene: gas station.

Lead: wheel balancers from a gas station, weights used on fishing lines, lead solder from electronics shop.

Mercury: a mercury thermometer (there is no need to break it), piece of amalgam tooth filling.

Methanol: paint supply store (wood alcohol).

Nickel: a nickel plated paper clip, a washed coin.

Patulin (apple mold): cut a sliver of washed, bruised apple.

PCB: water from a quarry known to be polluted with it (a builder or electrical worker may know a source).

Platinum: ask a jeweler for a small specimen.

Propyl alcohol: rubbing alcohol from pharmacy (same as propanol or isopropanol). Use a few drops only, discard the rest. Do not save it.

PVC: glue that lists it in the ingredients (polyvinyl chloride).

Radon: leave a glass jar with an inch of filtered water in it standing open in a basement that tested positive to radon using a kit. After 3 days, close the jar. Pour about 2 tsp. of this water into your specimen bottle.

Silicon: a dab of silicon caulk.

Silver: ask a jeweler for a crumb of very pure silver. Silver solder can be found in electronics shops. Snip the edge of a very old silver coin.

Sorghum mold: 1/8 tsp. sorghum syrup.

Styrene: a chip of styrofoam.

Tantalum: purchase a tantalum drill bit from hardware store.

Tin: scrape a tin bucket at a farm supply. Tin solder. Ask a dentist for a piece of pure tin (used to make braces).

Titanium: purchase a titanium drill bit from a hardware store.

Toluene: a tube of glue that lists toluene as an ingredient.

Tungsten: the filament in a burned out light bulb.

Vanadium: hold a piece of dampened paper towel over a gas stove burner as it is turned on. Cut a bit of this paper into your specimen bottle and add 2 tsp. filtered water.

Xylene: paint store or pharmacy.

Zearalenone: combine leftover crumbs of three kinds of corn chips and three kinds of popcorn.

This list gets you off to a good start. Since few of these specimens are pure, there is a degree of logic that you must apply in most cases. If you are testing for barium in your breast, a positive result would mean that a barium-containing lipstick tests positive and a barium-free lipstick is negative.

A chemistry set for hobbyists is a wonderful addition to your collection of test specimens. Remember, however, the assumptions and errors in such a system. A test for silver using silver chloride might be negative. This does not mean there is no silver present in your body; it only means there is no silver chloride present in the tissue you tested.

You are bound to miss some toxins; don't let this discourage you. There is more than enough that you can find.

The most fruitful kind of testing is, probably, the use of household products themselves as test substances. The soaps, colognes, mouthwash, toothpaste, shampoo, cosmetics, breads, dairy products, juices and cereals can all be made into test specimens. Put about 1/8 tsp. of the product in a small glass bottle, add 2 tsp. filtered water and ¼ tsp. grain alcohol to preserve it. For temporary purposes use a plastic baggy and water only. If you test positive to your household products in your white blood cells you shouldn't use them, even if you can not identify the exact toxin.

For a list of toxins and solvents I use, see page 571. To order pure substances see *Sources* for “chemicals for testing.”

Making Organ Specimens

To test for toxic elements or parasites in a particular organ such as the liver or skin, you will need either a fresh or frozen sample of the organ or a prepared microscope slide of this organ. Meat purchased from a grocery store, fresh or frozen, provides you with a variety of organ specimens. Chicken, turkey, beef or pork organs all give the same results. You may purchase chicken gizzards for a sample of stomach, beef liver for liver, pork brains for brain, beef steak for muscle, throat sweet breads for thymus, tripe for stomach lining. Other organs may be ordered from a meat packing plant.

Trim the marrow out of a bone slice to get bone marrow. Scrub the bone slice with hot water to free it of marrow to get a bone specimen. Choose a single piece of meat sample, rinse it and place it in a plastic bag. You may freeze it. To make a durable unfrozen sample, cut a small piece, the size of a pea, and place it in an amber glass bottle (½oz.). Cover with two tsp. filtered water and ¼ tsp. of grain alcohol (pure vodka will do) to preserve it. These need not be refrigerated but if decay starts, make a fresh specimen.

Pork brains from the grocery store may be dissected to give you the different parts of the brain. Chicken livers often have an attached gallbladder or piece of bile duct, giving you that extra organ. Grocery store “lites” provides you with lung tissue. For kidney, snip a piece off pork or beef kidney. Beef liver may supply you with a blood sample, too. Always wash hands and rinse with grain alcohol after handling raw meat.

I use ½ oz amber glass bottles with bakelite caps (see Sources) to hold specimens. However, plastic bags or other containers would suffice. After closing, each bottle is sealed with a Parafilm^{TM} strip to avoid accidental loosening of the cap. You may use masking tape.

To make a specimen of skin, use hangnail bits and skin peeled from a callous, not a wart. A few shreds will do. Remember, they must be very close to the test plate when in use; add 2 tsp. filtered water and ¼ tsp. grain alcohol.

Making a Complete Set of Tissue Samples

My original complete set was made from a frozen fish. As it thawed, different organs were cut away and small pieces placed in bottles for preserving in filtered water and grain alcohol. In this way, organs not available from the grocery store could be obtained. The piece of intestine closest to the anus corresponds to our colon, the part closest to the stomach corresponds to our duodenum. The 2 layers of the stomach and different layers of the eye, the optic nerve and spinal cord were obtained this way.

Another complete set of tissue samples were obtained from a freshly killed steer at a slaughter house. In this way the 4 chambers of the heart were obtained, the lung, trachea, aorta, vein, pancreas, and so forth.

Purchasing a Complete Set of Tissue Samples

Slides of tissues, unstained or stained in a variety of ways for microscope study give identical results to the preparations made by yourself in the ways already described. This fact opens the entire catalog of tissue types for your further study. See *Sources* for places that supply them.

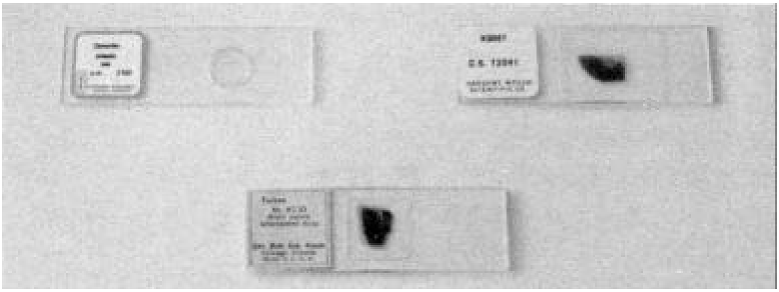

Fig. 82 Some purchased parasite and tissue slides.

You now have a set of organ samples, either fresh, frozen, preserved or on slides. You also have a set of test substances, whether chemical compounds, or elements, or products. Your goal is to search in your own organs and body tissues for the substances that may be robbing you of health.

Keeping yourself healthy will soon be an easy, daily routine.

Body Fluid Specimens

Each of these fluids should be prepared by putting about ¼ tsp. in a ½oz amber glass bottle. Add about 2 tsp. filtered water and ¼ tsp. grain alcohol for preservation. Undiluted specimens do not work for reasons that are technical and beyond the scope of this book. It is important *not* to shake the specimen, but to mix gently.

Urine. It is desired to have a pure, uninfected urine sample as a tissue specimen. Since this cannot be proved with certainty, obtain several urine samples from different persons whom you believe to be healthy and make several test specimens in order to compare results. Label your specimens Urine A (child), Urine B (woman), Urine C (mine), and so forth.

Semen. A sample from a condom is adequate. Aged specimens (sent by mail, unpreserved and unrefrigerated) work well also. Use one to ten drops or scrape a small amount with a plastic knife.

Blood. One to ten drops of blood should be used. Clotted or chemically treated blood is satisfactory. A blood smear on a slide is very convenient.

Milk. Cow's milk is too polluted with parasites to be useful. Electronically, a dead specimen is equivalent to a live specimen, so that pasteurization of the milk does not help. A human milk specimen is preferred.

Saliva. Use your own, if you have deparasitized yourself and test negative to various fluke stages. Otherwise find a well friend or child.

Specifying a tissue is the most powerful investigative technique in your arsenal. Any of your tissue samples can be tested for any of your toxic substances.

You Can Now Test Three Different Ways!

When you test with a substance on one plate and nothing on the other, you are searching your entire body for that substance. Such a test is not very sensitive.

By putting a tissue sample on the other plate you are testing for the substance specifically in that tissue, and this is much more sensitive. The tissue need not be the white blood cells. To find mercury in your kidneys you would use a mercury sample on one plate, and a kidney sample on the other. The technique is the same as when you use white blood cells.

If you put a substance on each plate, a resonating circuit means the two samples have something in common. For example, if you have mercury on one plate and some dental floss on the other, a positive result indicates mercury in the floss.

Lesson Five

Purpose: To watch substances travel through your body.

Materials: Prepare a pint of brown sugar solution (white sugar has propyl alcohol pollution) using filtered water. Use about 1 tsp. brown sugar, 1/8 tsp. vitamin C (to detoxify sorghum mold), and a pint of filtered water. Do not shake it; gently mix. Make a sample bottle by pouring about ½inch into a clean used vitamin bottle. Rinse and dry the outside of the sample bottle. Finally wash your hands with plain water.

Method:

1. Test your skin for the presence of brown sugar, using the newly made sample bottle and your skin specimen. It should not be there (resonate) yet.
2. Prepare a paper applicator by tearing the corner from a white unfragranced paper towel. Fold it to make a wick.
3. Dip the paper wick in the pint of sugar water and apply it to the skin of your inner arm where you can rub freely.

Rub it in vigorously for about 10 seconds (otherwise it takes minutes to absorb). Leave the shredded wick on the skin and tape it down with a piece of clear tape about 4 inches long (this increases the time you have to work). Quickly wash your fingers.

4. Place your skin tissue specimen on one plate and the sugar specimen bottle on the other plate.
5. Probe for resonance every 5 seconds. As soon as you hear resonance, implying that the skin has absorbed the sugar solution (which may take a full minute), replace the skin specimen with one of liver and listen for resonance again. There should be none, yet.
6. Alternate between the skin and liver. Soon the skin will be clear and the liver will resonate. Also check the pancreas and muscles to see how quickly sugar arrives there.
7. Check white blood cells and kidneys. It should not appear here (unless it is polluted with a toxin).
8. After five to ten minutes the sugar will be gone from all of these tissues and your experiment is ended. Wash your arm with plain water.

> Notice that you have only a few minutes to get all your testing done after the skin has absorbed the test substances.

Lesson Six

Purpose: To verify the propyl alcohol and benzene lists.

Method: We will use the Syncrometer to test for a toxin in a product. Assemble the products named in the propyl alcohol list (page 335) and benzene list (page 354)...as many as you can find. Also make sample bottles of benzene and propyl alcohol.

1. Place the propyl alcohol test substance on one plate and your products, in turn, on the other.

2. Listen to the current with only one of the plates in the circuit. Then listen with both plates in (the test plate switch ON). This method can detect one part per quadrillion in concentration. It is not as sensitive as the skin test (Lesson Five).
3. Repeat, with the benzene test substance.

Even tiny amounts of solvents are toxic! They must not be consumed or be left in our environment.

I have found that too many unsuspected products test positive to benzene. This is such a global tragedy that people must protect themselves by using their own tests. Rather than assurances, regulatory agencies should provide the consumer with cheap and simple tests (dip sticks and papers so we need not lug our Syncrometers around). Even if some test should fail, not all tests would fail to find an important pollutant like benzene. It would come to public attention much faster than the present debacle has.

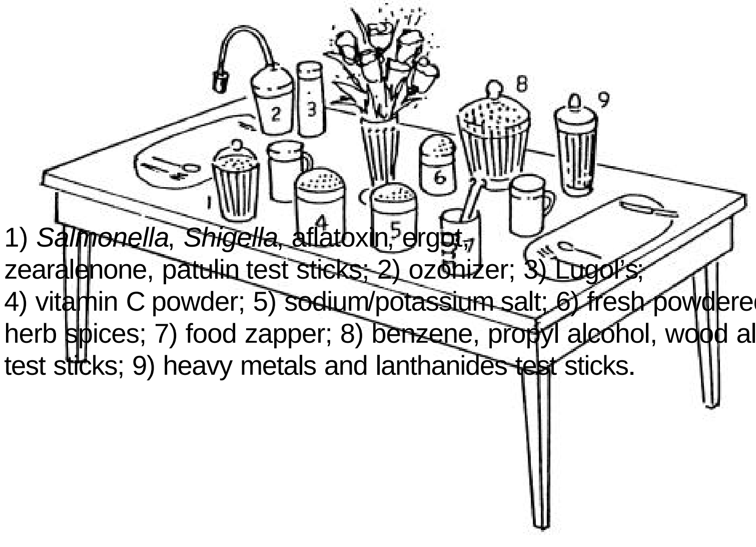

1) *Salmonella*, *Shigella*, aflatoxin, ergot, zearalenone, patulin test sticks; 2) ozonizer; 3) Lugol's; 4) vitamin C powder; 5) sodium/potassium salt; 6) fresh powdered herb spices; 7) food zapper; 8) benzene, propyl alcohol, wood alcohol test sticks; 9) heavy metals and lanthanides test sticks.

Fig. 83 Table of the future.

Lesson Seven

Purpose: To test for the presence of aluminum in your brain and your foods.

Materials: An aluminum measuring spoon, a tsp. of free flowing aluminized salt, a square inch of aluminum foil, a package of pork brain from the grocery store, kept frozen. (Other animal sources will do). Or a stained slide of cerebrum, cerebellum or other brain tissue.

Method:

1. Cut a piece of brain tissue (about 1 tsp.) and place in a plastic bag.
2. Place the aluminum samples in separate plastic bags. Add filtered water to each, about 1 tbs. Keep all surfaces and your hands meticulously clean (do not use soap).
3. Place the aluminum sample on one plate and the brain sample on the other plate.
4. Probe for resonance. If the circuit resonates you have aluminum in your brain.
5. If your aluminum specimen actually has cadmium or copper in it, you are also testing for these in your brain. Repeat the aluminum test with other aluminum objects. If they all resonate, you very, very likely have aluminum in your brain. If you can, test yourself to cadmium and copper, separately. If you don't have these in your brain, the aluminum test result is even more likely to be correct.
6. Of course, it would be desirable to have absolute certainty about this. To achieve this, purchase pure aluminum or an Atomic Absorption Standard. These are available from chemical supply companies.

If you do have aluminum in your brain, where is it coming from?

7. Leave your purest aluminum test substance on one plate, and replace the brain sample with these items, testing them one at a time. Remember to rest after each positive result.
   - a teaspoon of cottage cheese or yogurt taken from the top of a container of a foil-capped variety
   - a piece of cream cheese or butter that was wrapped in foil
   - a chip of bar soap or a bit of hand lotion
   - a piece of cake or rolls baked in an aluminum pan
   - a piece of turkey skin or hot dish that was covered with aluminum foil
   - anything baked with baking powder
   - a carbonated beverage from an aluminum can

Alternative Lesson:

To test for dental metal in your tissues. Use a piece of amalgam from an old tooth filling. This tests for the rest of the alloys in amalgam fillings as well as mercury. If you can't get a piece of mercury amalgam, use a mercury thermometer (don't break it, just put the bulb on the plate). Choose tissues like kidney, nerves, brain, liver, in addition to white blood cells.

I have never dissected human tissues and subjected them to confirmatory laboratory tests. It seems reasonable that because skin and tongue are directly provable, that other tissues work similarly.

Testing the Air

Fine particles and gas molecules that are in the air stick to the dust and eventually fall down onto the table, kitchen counter, and other places. Every night a film of dust accumulates, even though you can't see it. To test for air pollutants, gather some dust. Wipe the kitchen table and counter with a dampened piece of paper towel, two inches by two inches square. Place it in a resealable baggie. Do not get old dust, like from the top of the refrigerator or back shelves, because it does not represent the current air quality.

Testing Someone Else

Seat the person comfortably with their hand resting near you. Choose the first knuckle from the middle or first finger just like you do for yourself. Since you are touching this person, you are putting yourself in the circuit with the subject.

To exclude yourself, you need to add inductance to yourself. A coil of about 10 microhenrys, worn next to the skin, works well and is easily made. Obtain insulated wire and wrap 24 turns around a ball point pen (or something about that size), closely spaced. Cut the ends and tape them down securely. Keep it in a plastic bag, even when in your pocket. A commercial inductance of 4.7 microhenrys, worn touching your skin also works well. It can be worn on a string necklace. (Remember to remove the necklace when testing yourself.) The inductance acts as an RF (radio frequency) choke, limiting the alternating current that can flow through you while testing another person.

Test your inductor in this way. Repeat Lesson One with the coil next to your body. No resonance, even to SALT #1, should occur. If it does, make the coil bigger. Remove the inductor when you are not testing others.

Lesson Eight

Purpose: To detect aluminum in the brain of another person.
Materials: same as previous lesson, you wear the inductor.

Method:

1. Place the aluminum sample on one plate and the brain sample on the other plate.
2. Give the other person the handhold. You use the probe. Hold their finger steady in yours.
3. Probe the other person for resonance. The first probe is with only one plate in the circuit. The second is with both plates in the circuit. Resonance implies there is aluminum in the person's brain.

Saliva Testing

This may become your most useful test. The saliva has in it a bit of almost everything toxic that is in you. But it is not the first tissue to carry the HIV virus or a bit of a tapeworm stage. Nevertheless, *Salmonella* in your liver, mercury in your kidneys, aluminum in the brain all show up in the saliva, too. And saliva can be sent by mail or stored in the refrigerator. It should be frozen for long storage to prevent mold invasion. Or it may have grain alcohol added to preserve it. This test is not as sensitive as having the person present in the circuit, though.

To make a saliva specimen, place a two inch square piece of white, unfragranced paper towel (tear, don't cut) in a lightweight resealable baggy. Hold the open baggy near your mouth. Don't touch the paper towel with your fingers. Drool or spit onto the paper towel until half of it is damp. Zip it shut. Before testing, add enough filtered water to dampen the whole piece of paper.

Lesson Nine

Purpose: To search for shingles or *Herpes*.

Materials: A saliva specimen from the person being tested; they may be thousands of miles away. Also a specimen of the virus. This can be obtained from someone else's lesions—one droplet is enough, picked up on a bit of paper towel. The whole thing, towel and all, can be pushed into a glass bottle for preserving. Water and alcohol should be added. It can also be put on a slide, *Herpes, homemade*. A homeopathic preparation of the virus does not give accurate results for this kind of testing, due to the additional frequency imposed on it by potentizing. (However, homeopathic preparations can be used if the potency matches the tissue frequency where it resides. Hopefully, some way of using homeopathic sources will soon be found.)

Method: Place the saliva specimen in its unopened baggy on one plate. You may wish to open it briefly, though, to add enough filtered water to wet all the paper and add ¼ tsp. grain alcohol to sterilize or preserve it.

Place the virus specimen on the other plate and test as usual (like Lesson Six). A positive result means the person has active *Herpes*.

The main disadvantage of saliva testing is that you do not know which tissue has the pathogen or the toxin. You can only conclude that it is present. Usually this is enough information to carry out a corrective program.

Surrogate Testing

Although saliva testing is so easy, it is also possible to use an adult as a surrogate when testing a baby or pet. The pet or baby is held on the lap of the surrogate. A large pet may sit in front of the person. The handhold is held by the surrogate and pressed firmly against the body of the baby or pet. It can be laid flat against the arm, body or leg of a baby and held in place firmly by the whole hand of the adult. The paper covering should be wet. For a pet, the end is held firmly pressed against the skin, such as between the front legs or on the belly. The other hand of the adult is used for testing in the usual way. The adult must wear an inductor for surrogate testing as well as you, tester.

An ill or bedridden person may be tested without inconvenience or stress. He or she rests their whole hand on the skin of your leg, just above the knee. A wet piece of paper towel, about 4 inches by 4 inches is placed on your leg, to make better contact. You must use an inductor for yourself with this method. You may now proceed to probe on your hand instead of the ill person's.

Lesson Ten

Purpose: To test for cancer.

Materials: Ortho-phospho-tyrosine. Here are four ways to obtain some:

1. Order a pure sample from a chemical company (see *Sources*). Place a few milligrams (it need not be weighed) in a small glass bottle, add 2 tsp. filtered water and ¼ tsp. grain alcohol.
2. All persons with cancer have ortho-phospho-tyrosine in their urine as well as in the cancerous tissue. It is seldom found in other body fluids. Obtain a urine specimen from a friend who has active cancer. Freeze it if you can't prepare it immediately. Keep such specimens well marked in an additional sealed plastic bag. Persons who have recently been treated clinically for cancer are much less likely to have ortho-phospho-tyrosine in the urine.

   Urine cannot be considered a chemical in the same way as a sugar or salt solution. Urine is a tissue and has its own resonant frequency as do our other tissues. If combined with another tissue on the test plates, it will not resonate as if a solution of pure ortho-phospho-tyrosine were used. To use urine as an ortho-phospho-tyrosine specimen, you must:

   a) Pour a few drops of urine into your specimen bottle
   b) Add about 2 tsp. of filtered water
   c) Add a few drops of grain alcohol

   Gently mix, do not shake. Rinse and dry the outside of the bottle. Label it “urine/cancer”.

3. Cancer victims also have other growth factors being produced in their bodies. These are the same as can be found in mother's milk—for example, epidermal growth factor and insulin-like growth factor. Obtain a sample of mother's milk and use it to make another test substance for cancer. A few drops is enough.
4. There is still another way to prepare an ortho-phospho-tyrosine test sample. Common snails from a fish tank or outdoor snails are the natural hosts for *Fasciolopsis buskii* (human intestinal fluke) stages. The stages will produce ortho-phospho-tyrosine when the snails are fed fish food polluted with propyl alcohol. Over half the fish food cans I purchased had propyl alcohol pollution. Buy several brands of fish food. Test them for propyl alcohol and benzene. Obtain some snails, put them in a tank, feed them propyl alcohol polluted fish food. (Feed a separate group of snails benzene polluted fish food to obtain samples of HIV.) After two days put each snail in a zipped plastic bag, and test them individually against someone diagnosed with cancer or their saliva. The snails that the person tests positive to have ortho-phospho-tyrosine. Put these snails in the freezer to kill them humanely, then crush them and place in a specimen bottle with 50% grain alcohol to preserve. The bottles can be kept sealed and at room temperature.

   Similarly, your benzene snails can be tested against someone known to be HIV positive. Any snails that test positive can be used to prepare an HIV test specimen in the same way. The fish food must be tested for both benzene and propyl alcohol pollution, and separated appropriately, or you run the risk of making specimens that have both ortho-phospho-tyrosine and HIV.

Method:

1. Test for cancer by placing the test sample you just made (any of the four) on one plate and a white blood cell sample on the other plate.
2. If you resonate with both samples in the circuit you have cancer. Immediately, search for your cancer in your breast, prostate, skin, lungs, colon, and so forth.
3. To be more certain, test yourself to the other kinds of test samples. You should not resonate.

As you know by now, you can confirm the cancer by testing yourself to propyl alcohol and the human intestinal fluke in the liver. You should eliminate propyl alcohol from use, and zap all parasites. Keep testing yourself for cancer until it is gone. It should take less than one hour. Also continue to test yourself for propyl alcohol and the intestinal fluke in the white blood cells; make sure they are gone. Also test yourself for aflatoxin and freon.

Lesson Eleven

Purpose: To test for HIV.

Materials: Purchase a few milligrams of Protein 24 antigen (a piece of the HIV virus core) or the complete HIV virus on a slide (see *Sources*). You may use the vial unopened if only one test specimen is needed. To make more specimens, use about 1 milligram per ½ ounce bottle. Add 2 tsp. filtered water and ¼ tsp. grain alcohol. Or prepare an HIV specimen from snails as described in the previous Lesson.

Method: Search in the thymus (throat sweet breads), vagina and penis for the virus because that is where it will reside almost exclusively for the first year or two. If you don't have those tissue specimens, you could search in urine, blood, saliva, or white blood cells, but only a positive result can be trusted. Also search for the human intestinal fluke and benzene in the thymus. Of course, a positive test in these tissues is very significant. If you are positive, zap parasites immediately. You should test negative in less than an hour. Remove benzene polluted items from your lifestyle. Also test yourself to several varieties of popcorn, brown rice, and corn chips as an indication of zearalenone, which must be eliminated in order to get well. Follow up on yourself every few days to be sure your new found health is continuing. Test yourself for freon.

Lesson Twelve

Purpose: To test for diseases of all kinds.

Materials: Use slides and cultures of disease organisms. Homemade preparations of strep throat, acute mononucleosis, thrush (*Candida*), chicken pox, *Herpes* 1 and 2, eczema, shingles, warts, measles, yeast, fungus, rashes, colds, sore throats, sinus problems, tobacco virus, and so forth can all be made by swabbing or scraping the affected part. A plastic spoon or bit of paper towel works well. Put a small bit on a slide. Add a drop of balsam and a cover slip. Or put the towel in a bottle, add water and alcohol as described previously. Microscope slides can greatly expand your test set (see *Sources*).

Method: Test yourself for a variety of diseases, using your white blood cell specimen first. Then search in organs like the liver, pancreas, spleen. Notice how many of these common illnesses don't “go away” at all. They are alive and well in some organ. They are merely not making you sick!

Lesson Thirteen

To test for AIDS.

Materials: Benzene sample, slides of tissue samples like thymus, liver, pancreas, penis, and vagina. Also a collection of disease specimens such as the ones used in the previous lesson.

Method: Search in the thymus for benzene. If it is positive throughout the day, you are at risk for developing AIDS, although you may not be ill. Search other tissues for benzene. The more tissues with benzene in them the more serious the situation. Immediately search all your body products and foods for benzene.

Stay off benzene polluted items forever.

Tally up the diseases you tested positive for in Lesson Twelve. Test at least ten. If you had more than half positive you already have AIDS. (50% is my standard, you may set your own; an ideal standard for defining a healthy person should be 0% positive.)

Lesson Fourteen

Purpose: To test for aflatoxin.

Materials: Do not try to purchase a pure sample of aflatoxin; it is one of the most potent carcinogens known. Having it on hand would constitute unnecessary hazard, even though the bottle would never need to be opened. Simply make specimens of beer, moldy bread, apple cider vinegar, and any kind of peanuts using a very small amount and adding filtered water and grain alcohol as usual.

Method: Test yourself for these. If you have all of them in your white blood cells and the liver then you very, very probably have aflatoxin built up. Next, test your daily foods for their presence in your white blood cells. Those that test positive must be further tested for aflatoxin. Notice the effect of vitamin C on aflatoxin in your liver. Find a time when your liver is positive to aflatoxin (eat a few roasted peanuts from a health food store and wait ten minutes). Take 1 gram vitamin C in a glass of water. Check yourself for aflatoxin every five minutes. Does it clear? If not, take 5 or 10 grams vitamin C. How long does it take?

Lesson Fifteen

Purpose: To test for parasites.

Method: If you test positive to your pet's saliva, you have something in common–a parasite, no doubt. You must search your muscles and liver for these, not saliva or white blood cells, because they are seldom seen in these. Zap yourself for parasites until you no longer test positive to your pets' saliva.

Tapeworms and tapeworm stages can not (and should not) be killed with a regular frequency generator. Each segment, and probably each scolex in a cysticercus has its own frequency and might disperse if your generator misses it. Only zapping kills all and is safe for tapeworms.

Be sure to treat your pet on a daily basis with the pet parasite program.

Lesson Sixteen

Purpose: To test for fluke disease.

A small number of intestinal flukes resident in the intestine may not give you any noticeable symptoms. Similarly, sheep liver flukes resident in the liver and pancreatic flukes in the pancreas may not cause noticeable symptoms. Their eggs are shed through the organ ducts to the intestine and out with the bowel movement. They hatch and go through various stages of development outdoors and in other animals. But if you become the total host so that various stages are developing in your organs, you have what I term *fluke disease*. I have found that cancer, HIV, diabetes, endometriosis, Hodgkin's disease, Alzheimer's disease, lupus, MS and "universal allergy syndrome" are examples of fluke disease.

You can test for fluke disease in two ways: electronically and by microscope observation.

Materials: Cultures or slides of flukes and fluke stages from a biological supply company (see *Sources*) including eggs, miracidia, redia, cercaria, metacercaria. Body fluid specimens to help you locate them for observation under a microscope.

Method: Test for fluke stages in your white blood cells first. If you have any fluke stages in your white blood cells you may wish to see them with your own eyes. To do this, you must first locate them. Place your body fluid samples on one plate, your parasite stages on the other plate, and test for as many as you were able to procure, besides adults. After finding a stage electronically, you stand a better chance of finding it physically with a microscope.

Lesson Seventeen

Purpose: To see how sensitive your measurements can be. (How much of a substance must be present for you to get a positive result?)

Materials: filtered water, salt, glass cup measure, 13 new glass bottles that hold at least ¼ cup, 14 new plastic teaspoons, Your skin tissue sample, paper towel.

Method: Some of the best measurement systems available today are immunological (such as an ELISA assay) and can detect as little as 100 fg/ml (femtograms per milliliter). A milliliter is about as big as a pea, and a femtogram is 1/1,000,000,000,000,000th (10^{-15}) of a gram!

1. Rinse the glass cup measure with filtered water and put one half teaspoon of table salt in it. Fill to one cup, stirring with a plastic spoon. What concentration is this? A teaspoon is about 5 grams, a cup is about 230 ml (milliliters), therefore the starting concentration is about 2½(2.5) gm per 230 ml, or .01 gm/ml (we will discuss the amount of error later).
2. Label one clean plastic spoon "water" and use it to put nine spoonfuls of filtered water in a clean glass bottle. Use another plastic spoon to transfer one spoonful of the .01 gm/ml salt solution in the cup measure to the glass bottle, stir, then discard the spoon. The glass bottle now has a 1 in 10 dilution, and its concentration is one tenth the original, or .001 gm/ml.
3. Use the "water" spoon to put nine spoonfuls of filtered water in bottle #2. Use a new spoon to transfer a spoonful of salt solution from bottle #1 to bottle #2 and stir briefly (never shake). Label bottle #2 ".0001 gm/ml".
4. Repeat with remaining bottles. Bottle #13 would therefore be labeled ".000000000000001 gm/ml." This is 10^{-15} gm/ml, or 1 femtogram/ml.
5. Do the skin test with water from bottle #13 as in Lesson Five. If you can detect this, you are one hundred times as sensitive as an ELISA assay (and you should make a bottle #14 and continue if you are curious how good your sensitivity can get). If you can not, try to detect water from bottle #12 (ten times as sensitive as ELISA). Continue until you reach a bottle you can detect.

Calculate the error for your experiment by assuming you could be off by as much as 10% when measuring the salt and water adding up to 20% error in each of the 13 dilutions. This is a total error in bottle #13 of 280%, or at most a factor of 3. So bottle #13 could be anywhere from 0.33 to 3 femtogram/ml. If you can detect water from bottle #13, you are definitely more sensitive then an ELISA, in spite of your crude utensils and inexpensive equipment! Note that the starting error of using 2.5 gm instead of 2.3 gm only adds another 10% error.

If you want to calculate how many salt molecules you can detect, select the concentration at the limit of your detection, and put 2 drops on a square inch of paper towel and rub into your skin. Assume one drop can be absorbed. If you can detect water from bottle #13, you have detected 510,000 molecules (10^{-15} fg/ml divided by 58.5 gm/M multiplied by $6.02x10^{23}$ molecules/M divided by 20 drops/ml). Water in bottle #12 would therefore have 10 times as many molecules in one drop, and so forth. Even if your error is as much as a factor of 2 (100%), you can still get a good idea of what you can measure.

Atomic absorption standards start at exact concentrations; it is easy to make a more exact dilution series with them. When testing for iridium chloride by this skin test method, I was able to detect 3025 molecules!

Troubleshooting:

Always extend your set until you get a negative result (this should happen by at least bottle #18). If you always "detect" salt, then you shook the bottle!

Never try to reuse a bottle if you spill when pouring into it. Get another new bottle.

Sensitivity of Pollutant-In-Product Testing

Get some slides of *Salmonellas* and *Shigellas* and find some milk that tests positive to at least one. Make a dilution series of the milk up to bottle #14, being careful not to shake the bottles. Start with 2 drops of milk in bottle #1. Use an eye dropper to deliver 2 drops to subsequent bottles. Begin testing at bottle #14, using the slide that tested positive. You will learn to search by frequency later. My sensitivity was routinely around bottle #12, for a variety of pathogens. It was the same for toxic elements starting with standard solutions, about 1000 µg/ml, showing this method is less sensitive than skin testing.

Microscopy Lesson

Purpose: To *observe* fluke stages in saliva and urine with a microscope.

Materials:

a. A low power microscope. High power is not needed. A total of 100x magnification is satisfactory for the four common flukes, *Fasciolopsis*, *sheep liver fluke*, *human liver fluke* and *pancreatic fluke*.
b. Glass slides and coverslips.
c. A disposable eye dropper.

PRECLEANED
MICROSCOPE SLIDES
PLAIN GLASS
WARD'S
GLASS SLIDES

Fig. 84 Microscope, slides and coverslips.

d. For sanitation purposes (wiping table tops, slides, microscope and your hands) a 50% to 70% alcohol solution (not rubbing alcohol!) is best. Dilute 95% grain alcohol 7 parts alcohol plus 3 parts water. Vodka or 76% grain alcohol can be used undiluted.

e. Formaldehyde, 20%. Formaldehyde 37% is commonly available at pharmacies. Dilute this with equal parts of filtered water to get 18½%, which is close enough to 20%, for the purpose of "fixing" (killing) the specimens. Store in a glass bottle in the garage, away from sunlight. Label. Specimens that are fixed properly do not lose their life-like appearance.

f Iodine solution. This is only useful for the urine specimens. Lugol's iodine and tincture of iodine are both useful. Ask a pharmacist to prepare *Lugol's Iodine Solution* for you, as follows:

- 44 grams (1½oz) iodine crystals
- 88 grams (3 oz) potassium iodide crystals

Dissolve both in 1 liter (quart) filtered water. This may take a day of frequent shaking.

Method for saliva:

1. Pour the 20% formaldehyde into a small amber bottle or other receptacle to a depth of about 1/8 inch. Keep tightly closed.
2. The person to be tested is asked to salivate into the bottle so the organisms are immediately "fixed" without undergoing cooling first. The total volume should be about double the original amount of formaldehyde used. Make a mark on the container so the subject knows how much to produce. The resultant concentration of formaldehyde will be about 10%.
3. Shake the bottle a few times. Set it aside for 24 hours to settle (less if testing is urgent).

4. With a dropper, draw up some of the bottom settlings. Put one drop on a slide and apply a coverslip.
5. View under low power of microscope. Compare objects you observe with specimens obtained on slides from biological supply companies.

Note: Persons with HIV and moderate AIDS will show about one to ten parasite stages per slide. It requires several hours of searching. Persons with HIV and severe AIDS show 10 or more fluke stages per slide; this makes the task of finding them much easier. Persons with terminal untreated cancer have many more fluke stages than relatively well persons.

Method for urine:

1. Prepare bottles of formaldehyde fixative ahead of time. Put about ¼ to ½ inch of 20% formaldehyde in each. Keep tightly closed.
2. Add freshly voided[36] urine from cancer or HIV sufferers to the formaldehyde in approximately equal amounts, resulting in a 10% formaldehyde solution. Shake immediately. Let settle several hours. The sediment has a higher number of fluke stages. Cancer victims with cervical or prostate cancer will show higher numbers of stages in urine than other cancer types.
3. Staining the slide is optional. It helps to outline fluke stages slightly. Prepare Lugol's solution as described above.

   Slides may be stained in either of these two ways:

   - Put a drop of "fixed" urine on a slide. Add a drop of 50% Lugol's (dilute 1:1 with filtered water). Apply coverslip.

---

[36] Urine that has cooled even slightly below body temperature does not show *miracidia* and *redia* in their original shapes.

- Put a drop of "fixed" urine on a slide. Apply coverslip. Add 1 to 3 drops of 50% Lugol's to edge of coverslip and allow it to seep in.

Note: persons who have been treated for cancer or HIV using any of the known drugs may show only 1 to 2 fluke parasite stages per drop of saliva or urine. For this reason, you may need to search through 20 or more slides to find flukes. Very ill persons may show up to 10 parasites per drop (slide).

Taking Pictures Of What You See

You may be unsure of what you see even if you have the microscope slides of labeled flukes and their stages to study and compare. In real life, they vary so much in shape and size that absolute identification is difficult without experience. Unfortunately in a few hours, just as you are getting proficient, your magnificent specimens will be drying out and unfit for observation. To preserve them longer you can seal the edges by painting around the coverslip with fingernail clear enamel. Or dribble hot sealing wax along the edges and then place them in sealed plastic bags (one per bag). Melt sealing wax in a metal jar lid. Make an applicator from a piece of coat hanger wire bent in the shape of a small square to fit around the coverslip and a handle.

Or take photographs. To take pictures of what you see under the microscope you will need a photomicrographic camera, which costs $200.00 and up (see *Sources*). It is easy to use. Remember to label your pictures so you know which slide they came from.

Even photographs do not scientifically prove identity of parasite stages, but it is very good evidence. Proof would require that the saliva or urine sample could be cultured and seen to produce the known parasite stages.

Using A Frequency Generator

Frequency generators come in all sizes and costs and capabilities. If you can purchase one that reads out the frequency for you in numbers (digital type) and lets you produce a fraction of a kilohertz by turning a dial, it meets your most elementary needs. It should also be possible to set it on positive offset (100% positive) and still give you 5 volts. Then it can be used as a zapper. You should also be able to select a sine wave or square wave.

The advantages of having a frequency generator are that you can do your own diagnosing. You can find, in a few minutes:

1. Which invaders you have by dialing in to their frequencies.
2. At what frequencies you are broadcasting to the world.
3. What frequencies are used by other living things.

The Theory

Every living animal and every cell type produces its own frequencies and responds to these frequencies as well. We may speak of frequencies but we really mean waves, waves of energy. All waves have a frequency associated with them, so it's not really misleading to say frequencies when we really mean waves.

When the animal is alive it produces them, when it is dead it still responds to some of them. It is like the opera singer and the glass goblet. The opera singer produces frequencies in the air. The goblet responds to them because of its structure, not because it's alive. The goblet "picks up" on that particular frequency of sound because its own "resonant" frequency is exactly the same. If the singer sings loud enough the goblet shatters from the vibrations set up in it. An identical goblet, made of plastic, would not have the same resonant frequency.

There is not merely a structural and chemical difference between the living and non living. Living things are transmitting energy of some unique kind. And with your simple device, you will be able to trap this energy. And measure its exact frequency. This is not the same as understanding its makeup and source. We must leave that to others. But we can observe and use our observations to track down bacteria and other parasites. We can track down pollutants. We can measure our health quantitatively and perhaps in the future predict life expectancy.

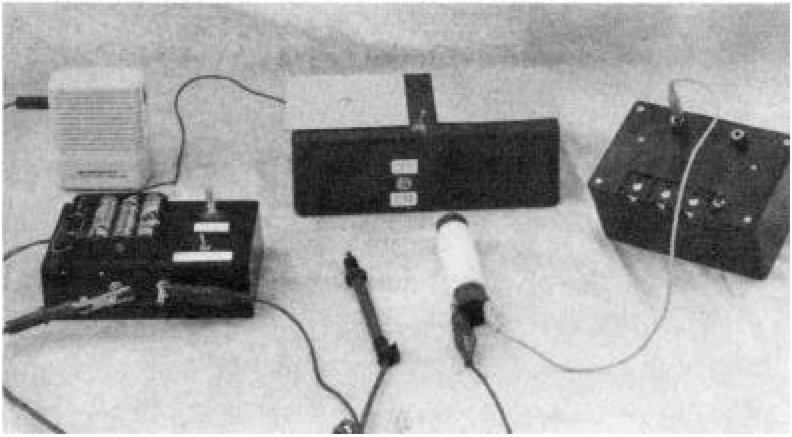

The Syncrometer at left is attached to the handhold. The frequency generator at right is also clipped to the handhold.

Fig. 85 Setup for finding resonant frequencies.

The Syncrometer traps the frequencies that match the ones in the material on the test plates and delivers them to an audio speaker in a range that you are able to hear. **Instead of test tissues or pathogens, we are now going to use pure test frequencies!** Remarkably, few things have overlapping frequencies, making this technique extremely useful for identification, even without a specimen!

Lesson Eighteen

Purpose: To search for the intestinal fluke in your body by listening to its broadcast frequency at 434 KHz.

Method: Turn on the frequency generator, select a frequency a short distance above the one you are interested in, like 438 KHz, turn the voltage (amplitude) down to less than one volt. Select sine waves. The lead coming from the frequency generator will have two connections, usually red and black (ground). We will not use the black (ground), just tape it out of the way. Pick up the handhold and probe in the usual way. Attach the red lead coming from the generator to your handhold. This makes two wires attached to your handhold. Although there is nothing on the test plates, they must be connected as usual with the switch at OFF (one plate is still ON).

Turn the Syncrometer ON. Probe yourself as usual. Your body's waves are being sent to the capacitor (plate) in the test plate box. The frequency from the Syncrometer is sent there, too. And now the 438 KHz waves from the generator are being sent there as well. Three different frequencies are mingling on the plate! If the two from your body and the generator are the same, the circuit will oscillate, and you will hear resonance. Turn the generator to 437 and probe again. Next, 436.

Sometimes, you can hear the resonance start to build. Continue on.

Next try 435. Then 434.

If your body is emitting a frequency of 434 KHz (coming from a live intestinal fluke inside you) it will be reinforced by the generator's 434 KHz. The reinforcement will put oscillations or resonance in the circuit, the same as you are accustomed to hearing with the Syncrometer. If there was none, you don't have the intestinal fluke anywhere in your body. Confirm this by starting at 430 KHz and working your way up.

If you hear resonance, you do have it. You may wish to verify this independently using a prepared slide of the fluke. Kill your flukes immediately as described in the next lesson.

Lesson Nineteen

Purpose: Killing the intestinal fluke with a frequency generator.

Materials: A frequency generator, two handholds with alligator clip leads for them.

Method: Wrap a single layer of paper towel over each of the two handholds. Wet them under the tap; squeeze out excess water. Clip them to the red and black wires of the frequency generator. (We use both wires for this purpose.) Dial up 434 KHz. Set the amplitude (voltage) at 10 volts. Grasp the handholds in each hand and hold on for three minutes. That is all. You have killed whatever tiny invader has a resonant frequency the same as the setting on the generator. Remember to zap all the stages, too; see *Pathogen Frequencies*.

If your frequency generator has a **positive offset** capability, you can use it like a zapper, and a single session will kill all pathogens, provided it is 100% offset and can give at least 5 volts at this setting. When using this technique, the generator can be set to any frequency from 2 KHz to 800 KHz, and you should go for 7 minutes. But even a small percentage of negative voltage will ruin this effect and do more harm than good! To be certain your generator is set correctly it would be best to observe the output on an oscilloscope.

Experiment with other voltage settings. Notice that less than one volt is also effective.

Lesson Twenty

Purpose: To find the bandwidth of a small living animal.

Materials: A fly, beetle or other insect, Syncrometer, frequency generator.

Discussion: Persons using a Syncrometer might have already tried putting a small insect on one of the plates. The circuit always resonates when you join the circuit at the handhold and probe. Even the tiniest ant placed in a glass bottle or plastic baggy will resonate the circuit. Unless it is too far away from the plate. If it has climbed up the side you will lose the resonance. At least one foot must be touching the bottom of the bottle. If the animal is dead this ceases. Obviously the living thing is affecting the circuit differently before and after death. Is it some kind of a wave form energy? To find its frequency you must add another frequency that will reinforce or interfere with the frequency already on the plate. Adding the generator frequency does just that.

Method: Use the same method as described in the last Lesson; however for an ant or fly, start at 1,000 KHz and proceed upward in big steps like 10 KHz. Use the right test plate which is controlled by the ON-OFF switch. Always listen to the current with the switch OFF, first, then ON. Move the frequency up and repeat. Continue until you hear resonance. Stop immediately. Rest your skin and go back down to the nonresonant frequency region. Move up in smaller steps this time. Repeat and repeat until you feel sure you know just where the resonance begins. But where does it end?

Start testing well above the suspected range taking big steps downward until you reach a resonant frequency. Rest and repeat until you find the upper limit of resonant frequencies. Record the bandwidth, for example, 1009-1112 KHz.

Lesson Twenty One

Purpose: To see if similar living things interfere with each other when put on the plate together.

Materials: Two identical living insects or very small living things.

Method: Find the broadcast range of each one separately and then together on the plate.

Note: Identical living things do not interfere with each others' frequencies.

Lesson Twenty Two

Purpose: To see if different living things interfere with each other when put on the plate together.

Method: Find the lower and upper end of the broadcast range of two different living things, such as a fly and a beetle or 2 kinds of flies or beetles. Then put them on the plate together. Notice there is no resonance in the accustomed range for either of them. They are interfering with each other on the plate.

Now add the 2 lower ends, then the two upper ends. Also subtract the 2 lower ends, then the two upper ends. For example imagine two insects, one with a spectrum of 1000 to 1090 KHz, the other with a range of 1050 to 1190 KHz. Adding the lower ends gives us 2050 KHz. Subtracting the lower ends gives us 50 KHz. Adding the upper ends gives 2280. Subtracting the upper ends gives 100. Now search for resonance at 50, 100, 2050, 2280 KHz. (These last two may be outside the range of your frequency generator. Choose more primitive life forms which have lower frequency bandwidths to stay within your limit.)

Notice that you hear resonance at exactly these frequencies and not above or below them. This is evidence for modulation of the frequencies: namely fusing them together.

Lesson Twenty Three

Purpose: To find your own bandwidth of emitted frequencies.

Materials: A frequency generator that goes up to 10 MHz. If yours only goes to 2 MHz you can still investigate the lower end of your band.

Method: You do not need to put yourself on the plate, since you are already there by being in the circuit at the handhold. However, if you are measuring someone else, they can simply touch the plate with a finger. Attach the frequency generator to the circuit at the handhold.

Since human adults begin to emit at about 1560 KHz, start searching at 1550, going upward in 1 KHz steps until you hear resonance.

Younger or healthier humans start emitting at a lower frequency and sometimes end at a higher frequency. In other words, they broadcast on a wider band.

Very young infants begin their band at about 1520 KHz. Could you ever regain this ability?

Most adults terminate at 9375 KHz

By eliminating molds from my diet, killing as many parasites and removing as many toxins as I became aware of, I have been able to expand my bandwidth from an initial 1562-9457 KHz in 1990 to 1520-9580 KHz in 1994! I hope this challenges you to accomplish a health improvement reflected in an even broader bandwidth for yourself.

Lesson Twenty Four

Purpose: To find the effect of a variety of things on the lower end of your spectrum, such as body temperature, eating, time of day, rainy weather, feeling sick. Notice that you may not change for weeks at a time, then suddenly see a shrinking of your bandwidth. You may assume you have eaten a mold.

Search for mold frequencies from 75 KHz to 295 KHz. Or test in your liver with mold samples. If this is positive go on a mold free diet—watching carefully for mold in your white blood cells. Even after removing the mold from your diet, so that no molds appear in your white blood cells, notice that your bandwidth does not recover. It regularly takes 2-3 weeks for mine to recover.

Surely, this sheds light on the poisonous effect of eating bad food.

Lesson Twenty Five

Purpose: To find an emission spectrum using a saliva sample.

Materials: A regular frequency generator.

Method: Search for the bottom of the resonant frequency band as in the previous lesson.

You may store it in the refrigerator for a few weeks without seeing a change. After that the band begins to shrink.

Lesson Twenty Six

Purpose: To observe the effect of dying on the bandwidth.

Method: Freeze the insect you tested in Lesson Twenty to kill it humanely. Repeat the search for its bandwidth. Note the bandwidth has become very narrow.

Note the bandwidth also depends on the accuracy of your particular frequency generator.

Lesson Twenty Seven

Purpose: To find unknown invaders of your body.

Method: Start at 900 KHz and proceed down to 77 KHz in 1 KHz steps, to search for all pathogens. If you find a resonating frequency, go to the Pathogen Frequency Chart (page 561) to identify likely candidates for it. Verify the identity of the invader by using a slide or culture specimen. If your pathogen remains unidentified, add it to the chart. This lets you determine whether the next illness is new or a recurrence of this one. Or just kill it.

Use the frequency generator at one pathogen's frequency. Wait ten minutes and retest all of the ones you found. Only that one will be gone. Now **zap**, wait ten minutes, and test again for all of the ones you found. Notice they are all gone. After one hour, search yet again for the pathogens you had. Any that are back must have come from an internal source not reached by the zapper current, like from the bowel or an abscess.

Lesson Twenty Eight

Purpose: To observe the action of a positive offset frequency on a very small animal. Does the animal die or is it just incapacitated?

Materials: A slug or small earthworm.

Method: Place the small animal in a plastic container like a cottage cheese carton. Add a few tsp. of water to wet the bottom. Attach a metal teaspoon to each of the generator clips. Place them on opposite sides in the carton and fasten with tape. Set the generator to positive offset at a frequency of about 30 KHz and 5 to 10 volts. Experiment with different voltages and compare effectiveness. Measure the time it takes for the animal to seem lifeless. You may try to revive it by keeping it for some time in the presence of food. Retest its emission band.

Lesson Twenty Nine

Purpose: To kill the bacteria in dairy products.

Materials: A glass of regular pasteurized milk, a carton of cottage cheese. A zapper.

Method: Search for *Salmonellas* and *Shigellas* first in the milk and cottage cheese. Search by frequency, using the chart, or with slides of these bacteria. If you don't find any, search different dairy foods until you find some bacteria. Attach metal teaspoons to the red and black leads of the generator. Place them inside the milk glass or cottage cheese carton, across from each other. Secure with masking tape. Attach the zapper. Zap them for 7 minutes. Remove the electrodes and wait 5 minutes. Test again for the same bacteria. They should be gone (but the food is not safe to eat due to the metal released from the teaspoons).

These experiments point to some exciting possibilities. Perhaps water supplies as well as foods and medicines could be sterilized this way. Perhaps sewage could be treated more efficiently, electrically. Best of all, maybe you could protect yourself from unsanitary products. If you do decide to explore this possibility, remember not to put metals in your mouth or food. Nor to use currents greater than 10 milliamps, or for longer than 10 minutes.

There are many commercially available function generators that can meet your needs. Two are shown here.

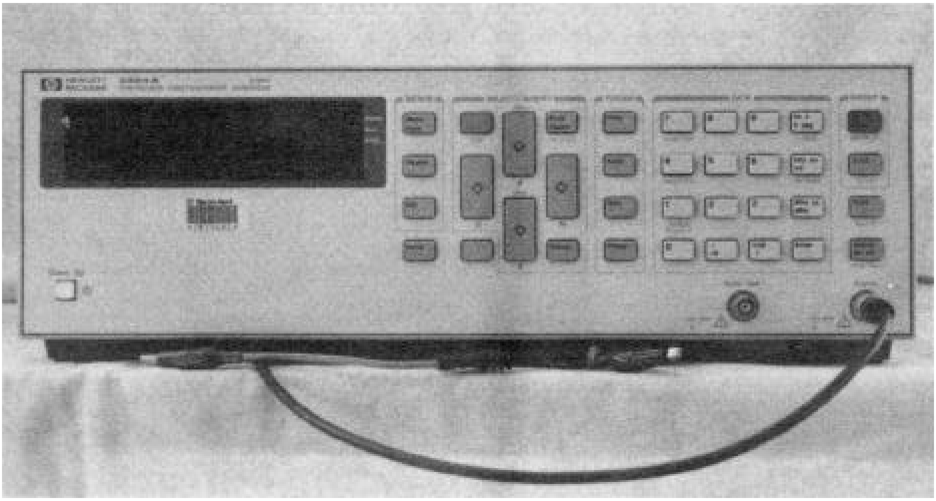

Fig. 86 Function generators.

Recipes

Read old recipe books for the fun and savings of making your own nutritious food. Change the recipes to avoid processed ingredients. Here are some I found:

Beverage Recipes

Anything made in your own juicer is fine. Experiment with new combinations to create different flavorful fruit and vegetable juices. Consider the luxury of preparing gourmet juices which satisfy your own individual palate instead of the mass-produced, polluted varieties sold at grocery stores. Remember to wash all fruit, including citrus before juicing. This removes the ever-present pesticides and common fruit mold.

Lemonade

1 cup fresh lemon juice, 1 cup honey, 1½quarts water. Bring honey and water to a boil if you plan to keep it several days. Then add lemon juice and store in the refrigerator.

All honey and maple syrup should have vitamin C added to it as soon as it arrives from the supermarket. Warm it first; then stir in ¼ tsp. per pint.

Fresh Tomato Juice

Simmer for ½hour: 12 medium-sized raw, ripe tomatoes, ½ cup water, 1 slice onion, 2 ribs celery with leaves, ½bay leaf, 3 sprigs parsley. Strain these ingredients. Season with: 1 tsp. salt (aluminum-free), ¼ tsp. paprika, ½tsp. honey. Serve thoroughly chilled. Makes about 4 servings.

Fresh Pineapple Juice

Peel a pineapple. Remove all soft spots. Cut it into cubes. Extract the juice by putting the pineapple through a food grinder or a blender. There will be very little pulp. Strain the juice and serve it on ice with sprigs of mint. Makes about 1½cups of juice. Mix the pulp with an equal amount of clover honey and use as topping (kept in freezer) for homemade ice cream (below), pancakes, or yogurt.

Maple Milk Shake

For each milk shake, blend or shake together: 1 glass of milk and 2 tablespoons maple syrup.

Remember, all milk gets boiled.

Yankee Drink

Mix together 1 gal. water, 3 cups honey, ½cup fresh lemon juice or distilled white vinegar, and 1 tsp. ginger.

Hot Vanilla Milk

Add one inch of vanilla bean and one tsp. honey to a glass of milk and bring to a near boil. You may add a pinch of cinnamon or other pure spice. You may even use vanilla extract (see *Sources*).

Red Milk

Equal parts fresh carrot juice (use a juicer) and sterilized milk. Save the carrot pulp for salads and soups.

C-Milk

Milk can absorb a surprising amount of vitamin C powder without curdling or changing its flavor. Try ½tsp. in a glass of cold milk.

My Own Soda Pop

Excellent for stomach distress. Put 1 tsp. citric acid, 2 tbs. vegetable glycerin (see *Sources*), 2 tbs. honey, and 1 lemon, juiced by hand, into a quart jar and fill with cold water. Refrigerate until ready to use. Then add 1 tsp. baking soda (chemically pure only, see *Sources*) and shake a few times, keeping the lid tight. Pour over a few ice cubes. Many variations are possible: other fruit concentrates, made in the blender, can be used along with some lemon juice; for example, 2 blended whole apples (peeled), blended pineapple, orange or grapefruit. Always add a bit of lemon to give it zip. You may add a pinch of ginger or other pure spice.

Note: The amount of sodium in ½tsp. baking soda is .476 grams. If you have heart disease, high blood pressure, or edema, use potassium bicarbonate instead. Ask your doctor what an acceptable amount of sodium or potassium bicarbonate is. I would suggest limiting yourself to one glass of soda pop a day, even if you do not have heart disease.

Another Note: the citric acid kills bacteria, while the carbonation brings relief.

My Own Super C-Pop

An excellent way to get lots of vitamin C into a child and relieve stomach distress at the same time. Squeeze 1 slice of lemon and 1 whole orange into an 8 ounce bottle that has a tight lid. Add 1 tsp. vitamin C powder (ascorbic acid), ¼ tsp. citric acid, and 2 tbs. vegetable glycerin (you may also experiment with honey for sweetness). Fill the bottle to the top with cold water. Then add ½tsp. chemically pure baking soda and close tightly. Shake briefly and serve immediately.

Half And Half

Mix equal parts whipping cream and milk or water. Boil and chill.

Buttermilk-C

Stir 1 tsp. vitamin C powder into a glass of milk. Add a pinch of potassium chloride. Additional seasoning may be pepper and herbs. Stir and enjoy.

Food Recipes

Despite the presence of aflatoxins, benzopyrenes, and solvents in many foods, it is possible to have a delicious and safe diet. Many persons need to gain weight, and with the emphasis in today's society on losing weight, consider yourself lucky in this respect. Help yourself to lots of butter, whipping cream, whole milk, avocados, and olive oil. Make your own preserves and baked goods, including breads. Remember, when you are recovering from a major illness it is essential not to diet to lose weight. You must wait two years after you are recovered to try to lose weight.

Daily Foods

Dairy products should contain at least 2% fat to enable you to absorb the calcium in them.

All milk should be sterilized

by boiling it for 10 seconds. If it makes mucous, you already have a chronic respiratory infection. Try to clear this up.

Change brands every time you shop to prevent the same pollutants from building up in your body.

If frying or cooking with fat, use only olive oil, butter or lard (the BHT and BHA preservatives in lard are OK except for seizure sufferers). Mix them for added flavor in your dishes. Never use margarine, Crisco™ or other hydrogenated fats. Do not cook over flames or grill, even when electric.

Eat lots of **fresh fruits and vegetables**. Wash them off only with cold tap water, not commercial food "wash". Scrub hard with a stiff bristled brush. Then cut away blemishes. Always peel potatoes, apples, and carrots. Modern dirt is full of chemicals and is toxic to you.

Be sure to drink plenty of **plain water** from your cold faucet throughout the day, especially if it is difficult for you to drink it with your meals. If you don't like the taste of your own tap water, try to get it from a friend with newer plumbing. Use a polyethylene (opaque) water jug from a grocery store to transport it. Never drink water that has been run through a water softener or copper plumbing or has traveled through a long plastic hose. Don't drink water that has stood in a container for a day. Dump it and sterilize the container. To further improve flavor and to dechlorinate attach a small faucet filter made of carbon only. Or buy a filter pitcher (see *Sources*). Don't drink water that has stood in the filter pitcher very long, either.

Because commercial **cold cereals** are very convenient, but have solvents, here are two replacements.

Two Granolas

7 cups rolled oats (old fashioned, not quick)
1 tsp. salt
1 cup wheat germ (fresh, not defatted)
½ cup honey
½ cup sunflower seeds, immaculate quality
½ cup milk (no need to sterilize, it is being baked)
½ cup melted butter
1 cup raisins, rinsed in vitamin C water

Mix dry ingredients together. Mix liquid ingredients and add gradually, while tossing until thoroughly mixed. Place in large ungreased pans and bake in slow (250°) oven. Stir occasionally, baking until brown and dry, usually 1-2 hours. Store in airtight container in freezer.

6 cups rolled oats
½ cup raw wheat germ
1 cup sesame seeds
1 cup sunflower seeds (raw, unsalted)
1 tsp. cinnamon
½ cup melted butter
½ cup honey

Preheat oven to 250°. Toss all ingredients in mixing bowl. Spread thinly on a baking sheet and bake 20-25 minutes. Stir often in order to brown evenly. When golden, remove and let cool. Makes 12 cups.

If you would like to add nuts to your granola recipes, rinse them in cold tap water first, to which vitamin C powder has been added (¼ tsp. per pint). This removes aflatoxins.

Peanut Butter

Use fresh unsalted roasted peanuts—they will be white on the first day they arrive at the health food store from the distributor. (Ask when they will arrive.) Or shell fresh roasted peanuts yourself, throwing away all shriveled or darkened nuts. Grind, adding salt and vitamin C (¼ tsp. per pint) as you go. For spreadability, especially for children, grind an equal volume of cold butter along with the peanuts. This improves spreadability and digestibility of the hard nut particles. This will probably be the most heavenly peanut butter your mouth has ever experienced.

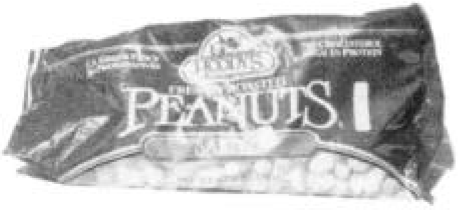

Fig. 87 Light colored, roasted peanuts in the shell had no aflatoxin.

Sweetening and Flavoring

Brown Sugar. Although I am prejudiced against all sugar from a health standpoint, my testing revealed no benzene, propyl alcohol, wood alcohol. However it does contain sorghum mold and must be treated with vitamin C to detoxify it. Add ¼ tsp. to a 1 pound package; knead until well mixed.

Maple syrup. Add vitamin C to newly opened bottle, ¼ tsp. to retard mold. Keep refrigerated and use promptly.

Flavoring. Use maple, vanilla (both natural and artificial), and any pure spice. They are free of molds and solvents.

Honeys. Get at least 4 flavors for variety: linden blossom, orange blossom, plain clover and local or wild flower honey.

Add vitamin C to newly opened jar to detoxify ergot mold (¼ tsp. per pint).

Jams and jellies. They are not safe unless homemade.

Fruit syrup. Use one package frozen fruit, such as cherries, blueberries or raspberries. Let thaw and measure the amount in cups (it might say on the package). Add an equal amount of clover honey to the fruit. Also add ¼ tsp. vitamin C powder. Mix it all in a quart canning jar and store in the refrigerator. Use this on pancakes, cereal, plain yogurt and homemade ice cream too. Use to make your own flavored beverages in a seltzer maker or to make soda pop. If you wish to use fresh fruit, bring it to a boil to sterilize. Use it up in a few days or boil to sterilize it again.

Note for diabetics

Diabetics must not use artificial sweeteners. Nor can they use all the sweeteners listed. Try stevia powder instead.

Preserves

Keep 3 or 4 kinds on hand, such as peach, pineapple, and pear. Peel and chop the fruit. It should not have any bruises. If you use a metal knife, rinse the fruit lightly afterwards. Add just enough water to keep the fruit from sticking as it is cooked (usually a few tablespoons). Then add an equal amount of honey, or to taste and heat again to boiling. Put in sterile jars in refrigerator. Make marmalade the same way, slicing the fruit and peel thinly. Always add vitamin C powder to a partly used jar to inhibit mold. Never use up partly molded fruit by making preserves out of it. Throw it out.

C Dressing

½ cup olive oil
¼ cup fresh lemon juice or white distilled vinegar
1 tsp. thyme, fenugreek or both (capsules are freshest)
1 tsp. vitamin C powder
½ tsp. brown sugar

Combine the ingredients in a clean salad dressing bottle. Shake. Refrigerate. The basic recipe is the oil and vinegar in a 2:1 ratio. After mixing these, add any pure spice desired. Or add fresh tomato chunks for creaminess.

Cheese Sauce

Add milk to cheese in equal amounts. Gradually heat to boiling while stirring. Add more of either to obtain the desired consistency. Boil 10 seconds. Use immediately.

Sour Cream-C

2 cups heavy whipping cream, previously boiled
¼ tsp. citric acid
¼ tsp. vitamin C powder
1 tsp. fresh onion juice or other seasoning (optional)

Stir until smooth, refrigerate 2 hours.

Yogurt

Buy a yogurt maker. Be sure and use boiled milk.

Soups

All home made soups are nutritious and safe, provided you use no processed ingredients (like bouillon), or make them in metal pots. Use herbs and aluminum-free salt to season. Always add a dash of vitamin C or tomato juice or vinegar to draw out calcium from soup bones for you to absorb.

Fish and Seafood recipe

Any kind of fish or seafood is acceptable, provided it is well-cooked. Don't buy food that is already in batter. The simplest way to cook fish is to poach it in milk. It can be taken straight from the freezer, rinsed, and placed in ¼ inch of milk (unboiled is fine) in the frying pan. Heat until it is cooked. Turn over and repeat. Throw away the milk. Serve with fresh lemon and herbs.

Baked Apples

Peel and core carefully. Remove all bruises (this is where the patulin is). Cut in bite-sized pieces, add a minimum of water and cook or bake minimally. Add a squirt of lemon juice when done. Serve with cinnamon, whipping cream and honey.

Ice creams

from the grocery store are loaded with benzene and other solvents. Fortunately there are ice cream makers that do everything (no cranking)! Or try our recipe which uses a blender. Be sure not to add store bought flavors, except vanilla or maple.

5 Minute Ice Cream

(Strawberry) Use 2 half pints of whipping cream, previously boiled, 1 package of frozen strawberries (about 10 oz.), and ½ cup clover honey. Pour frozen strawberries into blender. Pour whipping cream and honey over them. Blend briefly (about 10 seconds), not long enough to make butter! Pour it all into a large plastic bowl. Cover with a close fitting plastic bag and place in freezer. Prepare it a day ahead. Try using other frozen fruits, such as blueberries and cherries. Keep a few berries out of the blender and stir them in quickly with a non-metal spoon before setting the bowl in the freezer. There are many ice cream recipes to be found in old cook books. Avoid those with raw eggs or processed foods as ingredients. You may add nuts if you rinse them in vitamin C water.

Cookies, cakes and pies

Bake them from scratch, using unprocessed ingredients. Use simple recipes from old cook books.

Seven Day Sample Menu

Because processed foods have many toxins, you must cook as much from scratch as possible. So for convenience sake, keep your meals simple in preparation. You may want to prepare ahead and refrigerate your dressings and toppings. Or you could make a hot soup for dinner, refrigerate, and eat the leftovers for lunch. Don't save leftovers more than two days. Make sure they are covered. Try baking several potatoes at one time, refrigerate and put them in a salad the next night. Variety is the spice of life, so combine the allowed foods in the most creative ways you can imagine. And don't forget herbs and spices; learn to use them from old cook books.

	Breakfast	**Lunch**	**Dinner**
Day 1	Granola and honey with milk, half n' half or whipping cream 1 cup fresh squeezed fruit juice Water Milk	Fresh ground peanut butter and preserve sandwich Soup Milk Water	Orange roughy fish Fresh green beans with butter Baked potato with Sour Cream-C topping or fresh chives Pie (homemade) Milk 1 cup fresh squeezed or frozen vegetable juice Water

Day 2	Egg (limit is 2) Fried potatoes 1 glass milk Peppermint herb tea Fresh orange juice	Bagel (from bakery) Sour Cream-C Tomato 1 cup vegetable juice Water Milk	Homemade bean or lentil soup Sardines Dinner roll and butter Salad Homemade dressing Ice cream (homemade) Water
Day 3	Cream of Wheat™ cooked with raisins and milk Banana Peppermint herb tea ½ cup milk Water	Tuna sandwich with olives and butter Soup Milk Water	Baked sweet potato with butter and sweetening. Fresh broccoli with cheese sauce Bread and butter Chopped, peeled pear and whipping cream 1 cup vegetable juice Milk Water
Day 4	French toast with maple syrup Egg Homemade grapefruit juice Milk Water	Avocado and sour cream sandwich ½ cup vegetable juice Bread and butter Water	Lobster or sautéed shrimp Fresh asparagus Potatoes, any style ½ cup vegetable juice Water Milk
Day 5	Cooked cereal 1 glass milk Sliced banana with whipping cream and honey 1 glass water	Cold potato salad with C Dressing Soup ½ cup vegetable juice Custard Water Milk	1 can sardines or salmon in easy-open can (can openers shed metal) Salad of lettuce, tomato, olives, avocado with homemade dressing Bread with butter Ice-cream (optional) ½ cup vegetable juice Water Milk
Day 6	Egg and homemade biscuit with honey and butter Milk Fruit juice (homemade) Water	Homemade peanut butter sandwich ½ cup vegetable juice Milk Water	Gourmet pizza: home baked bread topped with olive oil, sliced tomato or homemade sauce, grated cheese, sardines or anchovies, chopped vegetables, garlic and onion Salad Milk Water

Day 7	Pancakes Banana or chopped fruit with cream Milk Water	Salmon sandwich (from flip top can) ½ cup vegetable juice Milk Water	Stir-fry vegetables: broccoli, carrots, cabbage, in olive oil and butter Bread and butter Pie (optional) Milk Water

Remember, take vitamin C and B-complex with each meal.

Too Sick To Cook, Too Tired To Eat

Pick three meals from the sample menu that need no cooking and eat them every day.

Recipes for Natural Body Products

You can use just borax (like 20 Mule Team Borax^{TM}) and washing soda (like Arm & Hammer Super Washing Soda^{TM}) for all types of cleaning including your body, laundry, dishes and your house! You don't need all of those products you see in commercials for each special task!

Even if you have dry skin, difficult hair or some other unique requirement, just pure borax will satisfy these needs. A part of every skin problem is due to the toxic elements found in the soaps themselves. For instance aluminum is commonly added as a “skin moisturizer”. It does this by impregnating the skin and attracting water, giving the illusion of moist skin. In fact you simply have moist aluminum stuck in your skin which your immune system must remove. While borax won't directly heal your skin or complexion, it does replace the agents that are causing damage, so that healing can occur.

Borax Liquid Soap

Empty 1 gallon jug
1/8 cup borax powder
Plastic funnel

Funnel the borax into the jug, fill with cold tap water. Shake a few times. Let settle. In a few minutes you can pour off the clear part into dispenser bottles. This is the soap!

Easier way: use any bottle, pour borax powder to a depth of a ½ inch or so. Add water. Shake. When you have used it down to the undissolved granules, add more water and shake again. Add more borax when the undissolved granules get low.

Keep a dispenser by the kitchen sink, bathroom sink, and shower. It does not contain aluminum as regular detergents and soaps do, and which probably contribute to Alzheimer's disease. It does not contain PCBs as many commercial and health food varieties do. It does not contain cobalt (the blue or green granules) which causes heart disease and draws cancer parasites to the skin. Commercial detergents and non-soaps are simply not safe. Switch to homemade bar soap and borax for all your tasks! Borax inhibits the bacterial enzyme *urease* and is therefore antibacterial. It may even clear your skin of blemishes and stop your scalp from itching.

For Laundry

Borax (½ cup per load). It is the main ingredient of non-chlorine bleach and has excellent cleaning power without fading colors. Your regular laundry soap may contain PCBs, aluminum, cobalt and other chemicals. These get rubbed into your skin constantly as you wear your clothing. For bleaching (only do this occasionally) use original chlorine bleach (not “new improved” or “with special brighteners”, and so forth). Don't use chlorine if there is an ill person in the house. For getting out stubborn dirt at collars, scrub with homemade bar soap first; for stains, try grain alcohol, vinegar, baking soda.

For Dishes

Don't believe your eyes when you see the commercials where the smiling person pulls a shining dish out of greasy suds. Any dish soap that you use should be safe enough to eat because nothing rinses off clean. Regular dish detergents, including health brands, are now polluted with PCBs. They also contain harmful chemicals. Use borax for your dishes. Or use paper plates and plastic (not styrofoam) cups.

In The Dishwasher

Use 2 tsp. borax powder pre-dissolved in water. If you use too much it will leave a film on your dishes. Use vinegar in the rinse cycle.

In The Sink

Use a dishpan in the sink. Use ¼ cup borax and add a minimum of water. Also keep a bit of dry borax in a saucer by the sink for scouring. Don't use any soap at all for dishes that aren't greasy and can be washed under the faucet with nothing but running water. Throw away your old sponge or brush or cloth because it may be PCB contaminated. Start each day by sterilizing your sponge (it harbors Salmonella) or with a new one while the used one dries for three full days. Clean greasy pots and pans with a paper towel first. Then use homemade bar soap.

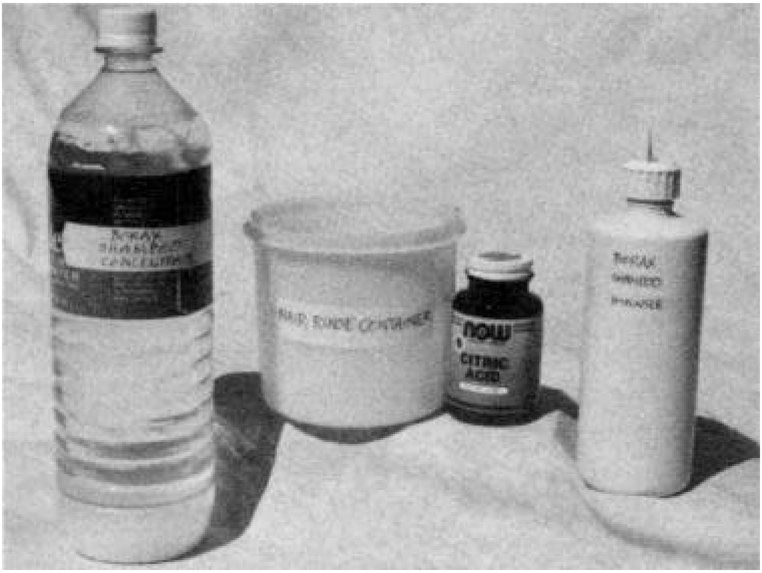

Fig. 88 Make a bottle of borax liquid to fill your soap dispensers and shampoo bottle. Use citric acid to rinse and condition.

Shampoo

Borax liquid is ready to use as shampoo, too. It does not lather but goes right to work removing sweat and soil without stripping your color or natural oils. It inhibits scalp bacteria and stops flaking and itching. Hair gets squeaky clean so quickly (just a few squirts does it) that you might think nothing has happened! You will soon be accustomed to non-lathery soap. Rinse very thoroughly because you should leave your scalp slightly acidic. Take a pint container to the shower with you. Put ¼ tsp. citric (not ascorbic) acid crystals (see *Sources*) in it. For long hair use a quart of rinse. Only **citric acid** is strong enough to get the borax out, lemon juice and vinegar are not. After shampooing, fill the container with water and rinse. Rinse your whole body, too, since citric acid is also anti-bacterial. All hair shampoo penetrates the eye lids and gets into the eyes although you do not feel it. It is important to use this natural rinse to neutralize the shampoo in your eyes. (Some people have stated that citric acid makes their hair curlier or reddens it. If this is undesirable, use only half as much citric acid.) Citric acid also conditions and gives body and sheen to hair.

Hair Spray

I don't have a recipe that holds your hair as well as the bottle of chemicals you can buy at the store. Remarkably a little lemon juice (not from a bottle) has some holding power and no odor! Buy a 1 cup spray bottle. Squeeze part of a lemon, letting only the clear juice run into the bottle. Fill with water. Keep it in the refrigerator. Make it fresh every week. Spraying with just plain water is nearly as good! For shinier hair, drop a bit of lemon peel into the bottle.

Homemade Soap

A small plastic dishpan, about 10” x 12”
A glass or enamel 2-quart sauce pan
1 can of lye (sodium hydroxide), 12 ounces
3 pounds of lard (BHT and BHA are OK here)
Plastic gloves
Water

1. Pour 3 cups of very cold water (refrigerate water overnight first) into the 2-quart saucepan.
2. Slowly and carefully add the lye, a little bit at a time, stirring it with a wooden or plastic utensil. (Use plastic gloves for this; test them for holes first.) Do not breathe the vapor or lean over the container or have children nearby. Above all use no metal. The mixture will get very hot. In olden days, a sassafras branch was used to stir, imparting a fragrance and insect deterrent for mosquitoes, lice, fleas, ticks.

3. Let cool at least one hour in a safe place. Meanwhile, the unwrapped lard should be warming up to room temperature in the plastic dishpan.
4. Slowly and carefully, pour the lye solution into the dishpan with the lard. The lard will melt. Mix thoroughly, at least 15 minutes, until it looks like thick pudding.
5. Let it set until the next morning; then cut it into bars. It will get harder after a few days. Then package.

If you wish to make soap based on olive oil, use about 48 ounces. It may need to harden for a week.

Liquid Soap

Make chips from your homemade soap cake. Add enough hot water to dissolve. Add citric acid to balance the pH (7 to 8). If you do not, this soap may be too harsh for your skin.

Skin Sanitizer

Make up a 5 to 10% solution of food grade alcohol. Food grade alcohols are grain (ethyl) alcohol or vodka. Find a suitable dispenser bottle. Mark it with a pen at about one tenth of the way up from the bottom. Pour 95% grain alcohol (190 proof) to this mark (for 50% grain alcohol or vodka make your mark one fifth of the way up). Add water to the top. Keep shut. You may add a chip of lemon peel for fragrance.

Use this for general sanitizing purposes: bathroom fixtures, knobs, handles, canes, walkers, and for personal cleanliness (but use chlorine bleach for the toilet bowl once a week). Always clean up after a bowel movement with wet toilet paper. This is not clean enough, though. Follow with a stronger damp paper towel. This is still not clean enough; use a final damp paper towel with skin sanitizer added. After washing hands, sanitize them too, pouring a bit on one palm and put finger tips of the other hand in it, scratch to get under nails, repeat on other hand. Rinse with water.

Do not use this recipe, nor keep any bottles of alcohol in the house of a recovering alcoholic.

Deodorant

Your sweat is odorless. It is the entrenched bacteria feeding on it that make smells. You can never completely rid yourself of these bacteria, although they may temporarily be gone after zapping. The strategy is to control their numbers. Here are several deodorants to try. Find one that works best for you:

Vitamin C water. Mix ¼ tsp. to a pint of water and dab it on. Then dab on cornstarch.

Citric Acid water. Mix ¼ tsp. to a pint of water and dab it on. Then dab on cornstarch.

Only a few drops of these acids under each armpit are necessary. If these acids burn the skin, dilute them more. Never apply anything to skin that has just been shaved!

Corn starch. Many people need only this. Dab it on.

Use only unpolluted **cornstarch** (see *Sources*).

Baking soda has been deleted as a deodorant because benzene was found in some boxes.

Lemon juice. This acid is not as strong, use what you need.

Pure alcohol (never rubbing alcohol). The food grade alcohols are grain alcohol and vodka. Dab a bit under each arm and/or on your shirt or blouse, then dab on cornstarch. If the alcohol burns, dilute it with water. Be very careful not to leave the bottle where a child or alcoholic person could find it. Pour it into a different bottle!

Pure zinc oxide. You may ask your pharmacist to order this for you. She or he may wish to make it up for you too, but do not let them add anything else to it. It should be about 1 part zinc oxide powder to 3 parts water. It does not dissolve. Just shake it up to use it. After you get it home, you can add cornstarch to it to give it a creamy texture. Heat 3 tsp. cornstarch in 1 cup of water, to boiling, until dissolved and clear. Cool and add some to the zinc oxide mixture (about equal parts). Store unused starch mixture in the refrigerator. Only make up enough for a month.

Alcohol and zinc oxide. This is the most powerful deodorant. Apply alcohol first, then the zinc oxide, then dab on cornstarch.

Remember that you need to sweat! Sweating excretes toxic substances, especially from the upper body. Don't use deodorant on weekends. Go to the sink and wipe clean the armpits like our grandparents did. Often, just plain cornstarch is enough! These homemade deodorants are not as powerful as the commercial varieties–this is to your advantage.

Brushing Teeth

Buy a new toothbrush. Your old one is soaked with toxins from your old toothpaste. Use only water or chemically pure baking soda if you have any metal fillings. Put a pinch in a glass, add water to dissolve it. Use food-grade hydrogen peroxide (see *Sources*) if you have only plastic fillings. Dilute it from 35% to 17½% by adding water (equal parts). Store hydrogen peroxide only in polyethylene or the original plastic bottle. Use 4 or 5 drops on your toothbrush. It should fizz nicely as oxygen is produced in your mouth. Your teeth will whiten noticeably in 6 months. Before brushing teeth, floss with 4 or 2 pound monofilament fish line. Double it and twist for extra strength. Rinse before use. Floss and brush only once a day. If this leaves you uncomfortable, brush the extra times with plain water and a second “water-only” toothbrush. Make sure that nothing solid, like powder, is on your toothbrush; it will scour the enamel and give you sensitive teeth, especially as you get older and the enamel softens. Salt is corrosive—don't use it for brushing metal teeth. Plain water is just as good.

For Dentures

Use salt water. It kills all germs and is inexpensive. Salt water plus grain alcohol or food-grade hydrogen peroxide makes a good denture-soak.

Mouthwash

A few drops of food grade hydrogen peroxide added to a little water in a glass should be enough to make your mouth foam and cleanse. Don't use hydrogen peroxide, though, if you have metal fillings, because they react. Don't use regular drug store variety hydrogen peroxide because it contains toxic additives. Health food store varieties contain solvents from the bottling process. See *Sources*. Never purchase hydrogen peroxide in a bottle with a metal cap.

For persons with metal tooth fillings, use chemically pure baking soda or just plain hot water. A healthy mouth has no odor! You shouldn't need a mouthwash! If you have breath odor, search for a hidden tooth infection or cavitation.

Contact Lens Solution

A scant cup of cold tap water brought to a boil in glass saucepan. After adding ¼ tsp. aluminum free salt and boiling again, pour into a sterile canning jar. Refrigerate. Freeze some of it.

Lip Soother

For dry, burning lips. Heat 1 level tsp. sodium alginate plus 1 cup water until dissolved. After cooling, pour it into a small bottle to carry in your purse or pocket (refrigerate the remainder). Dab it on whenever needed. If the consistency isn't right for you, add water or boil it down further. You can make a better lip soother by adding some lysine from a crushed tablet, vitamin C powder, and a vitamin E capsule to the alginate mix. If you have a persistent problem with chapped lips, try going off citrus juice.

Foot Powder

Use a mixture of cornstarch and zinc oxide poured into a salt shaker with a lid. Add long rice grains to fight humidity. You may also try arrow root or potato starch. If you don't have zinc oxide use plain cornstarch.

Skin Healer Moisturizer Lotion

1 tsp. sodium alginate
1 cup water

Make the base first by heating these together in a covered, non-metal pan until completely dissolved. Use low heat–it will take over an hour. Use a wooden spoon handle to stir. Set aside. Then make the following mixture:

¼ tsp. Vitamin C (ascorbic acid) (You may crush tablets)
¼ tsp. lysine (crush tablets)
2 tbs. pure vegetable glycerin
2 vitamin E capsules (400 units or more, each)
1 tsp. apricot kernel oil (olive oil will do)

1 tbs. lemon juice from a lemon or ¼ tsp. citric acid (this is optional)
1 cup water

Heat the water to steaming in a non-metal pan. Add vitamin C and lysine first and then everything else. Pour into a pint jar and shake to mix. Then add the sodium alginate base to the desired thickness (about equal amounts) and shake. Pour some into a small bottle to use as lip soother. Pour some into a larger bottle to dispense on skin. Store remainder in refrigerator. (See *Sources* for sodium alginate, vegetable glycerin and apricot kernel oil. Sodium alginate is also available in capsule form at some health food stores.)

Other Skin Healers

Vitamin C powder (ascorbic acid, not the same as citric acid). Put a large pinch into the palm of your hand. With your other hand pick up a few drops of water from the faucet. Rub hands together until all the powder is dissolved and dispensed. It may sting briefly. Do this at bedtime, especially for cracked, chapped hands. Include lips if they need it.

Vitamin E oil. Vitamin E oil from Now Foods was not polluted at the time of this writing, but for the future it would be safer to rely on capsules. Snip open a capsule and rub into skin.

50% Glycerin. Dilute 100% vegetable glycerin with an equal amount of water. This is useful as an after shave lotion.

Vitamin C liquid. Mix ¼ tsp. vitamin C powder in one pint water (crushed tablets will do). This is useful as an after shave lotion and general skin treatment.

Apricot Kernel Oil. This is a very light oil, useful as an after shave lotion and general skin treatment.

Cornstarch (see *Sources*). Use on rashes, fungus, moist or irritated areas and to prevent chafe.

Combining several of these makes them more effective.

Dry skin has several causes: too much water contact, too much soap contact (switch to borax), low body temperature, not enough fat in the diet, or parasites.

Massage Oil

Use olive oil. It comes in very light to heavy textures. Pick the right one for your purpose. Alginate mixtures can be used instead of, or added to, oil. Starch solutions are good, too.

Sunscreen Lotion

Purchase PABA (see *Sources*) in 500 mg tablet form. Dissolve 1 tablet in grain alcohol or vodka. Grind the tablet first by putting it in a plastic bag and rolling over it with a glass jar. It will not completely dissolve even if you use a tablespoon of the alcohol. Pour the whole mixture into a 4 ounce bottle of homemade skin softener. Be careful not to get the lotion into your eyes when applying it. A better solution is to wear a hat or stay out of the sun. Remember to *take* PABA as a supplement, too (500 mg, one a day).

Nose Salve

(When the inside of the nose is dry, cracked and bleeding.)

Pour ½tsp. pure vegetable glycerin into a bottle cap. Add ½ tsp. of water.

Applicator: use a plastic coffee stirrer or straw; cut a slit in the end to catch some cotton wool salvaged from a vitamin bottle and twist (cotton swabs, cotton balls and wooden toothpicks are sterilized with mercury which in turn is polluted with thallium). Dip it into the glycerin mixture and apply inside the nose with a rotating motion. Do each nostril with a new applicator.

Quick Corn Starch Skin Softener

4 tsp. corn starch (see *Sources*)
1 cup water

Boil starch and water until clear, about one minute.

Cornstarch Skin Softener

1 tsp. lysine powder or 8 tablets, 500 mg each
1 tsp. Vitamin C powder (ascorbic acid); or 8 tablets, 500 mg each
3 tsp. cornstarch (see *Sources*)
Vitamin E, 1 capsule 400 mg
¼ tsp. apricot kernel oil (optional)
1 cup water

Boil starch and water until clear, about one minute. Add other ingredients and stir until dissolved. Cool. Pour into dispenser bottle. Keep refrigerated when not in use. Apply after washing dishes and after showering.

After Shaves

Vitamin C. ¼ tsp. vitamin C powder, dissolved in 1 pint water.

Apricot kernel oil.

Vegetable glycerin. Equal parts glycerin and water or to suit your need.

Personal Lubricants

Heat these together: 1 level tsp. sodium alginate and 1 cup water in a covered non-metal pan until completely dissolved. Use very low heat and stir with a wooden spoon handle. It takes a fairly long time to get it perfectly smooth. After cooling, pour into a small dispenser bottle. Keep the remainder refrigerated.

Or, mix and heat 4 tsp. cornstarch and 1 cup water until completely dissolved in a covered saucepan. Use non-metal dishes and a non-metal stirring spoon. Cool. Pour some into dispenser bottle. Refrigerate remainder. This is many person's favorite recipe.

Baby Wipes

Cut paper towels in quarters and stack in a closable plastic box. Run tap water over them, drain the excess. Add 1 tsp. grain alcohol and/or borax liquid on top. Close. Put a dab of the Quick Cornstarch Softener recipe on top of each wipe as you use it.

People Wipes

¼ tsp. powdered lysine (you may crush tablets)
¼ tsp. Vitamin C powder (you may crush tablets)
¼ cup vegetable glycerin
1 cup water

Prepare wipes by cutting paper towels in quarters. Use white, unfragranced towels that are strong enough to hold up for this use. Fold each piece in quarters again and stack in a plastic zippered baggy. Pour the fluid mixture over the stack and zip. Store a bag full in the freezer to take on car trips. If you want to keep them a month or more, add 1 tbs. grain alcohol or vodka to the recipe.

For bathroom use, dampen a roll of paper towels under the cold tap first. Then pour about ¼ cup of the mixture over the towel roll around the middle. Store in plastic shopping bag or stand in plastic waste basket.

Recipes For Natural Cosmetics

Eye liner and Eyebrow Pencil

Get a pure charcoal pencil (black only) at an art supply store. Try several on yourself (bring a small mirror) in the store to see what hardness suits you. You may need to wet it with water or a vitamin E perle first. Don't put any chemicals on your *eyelids,* since this penetrates into your eye. To check this out for yourself, close your eye tightly and then dab lemon juice on your eyelid. It will soon burn! Everything that is put on skin penetrates. Otherwise the nicotine patch and estrogen patch wouldn't work. Not even soap belongs on your eyelids! Charcoal pencils are cheap. Get yourself half a dozen different kinds so you can do different things.

You could also use a capsule of activated charcoal. Empty it into a saucer. Mix glycerin and water, half and half, and add it to the charcoal powder until you get the consistency you like. Use a brush for eyelashes; use a finger for eyebrows.

Lipstick

Beet root powder (see *Sources*)
100% vegetable glycerin

Combine 1 tsp. vegetable glycerin and 1 tsp. beet root powder in a saucer. Stir until perfectly smooth. Then add ½tsp. of vitamin E oil. Snip open vitamin E capsules or buy vitamin E oil (see *Sources*). Very thick olive oil can be substituted. Apply liberally with your finger or a lipstick brush. Do not purse or rub your lips together after application. To make the lipstick stay on longer, apply 1 layer of lipstick, then dab some corn starch over the lips, then apply another layer of lipstick. Store in a small glass or plastic container in the refrigerator, tightly covered in a plastic bag.

Face Powder

Use cornstarch from the original box. You may also try arrow root starch or potato starch. Use your fingers or a tissue to apply because applicators can carry bacteria.

Blush (face powder in a cake form)

Add 50% glycerin to cornstarch in a saucer to make a paste. Slowly add beet root powder to the desired color. Use part of a charcoal capsule to darken it, if desired. A drop of food grade alcohol will also darken it. To make 50% glycerin, add equal parts of glycerin and water. Try to make the consistency the same as your brand name product, and you can even put it back in your brand name container.

Recipes For Household Products

Floor Cleaner

Use washing soda from the grocery store. You may add borax and boric acid (to deter insects except ants). Use white distilled vinegar in your rinse water for a natural shine and ant repellent. Do not add bleach to this. For the bathroom floor use plain bleach water—follow the label. Never use chlorine bleach if anybody in the home is ill or suffers from depression. Use grain alcohol (1 pint to 3 quarts water) for germ killing action instead of chlorine.

Furniture Duster and Window Cleaner

Mix equal parts white distilled vinegar and water. Put it in a spray bottle.

Furniture Polish

A few drops of olive oil on a dampened cloth. Use filtered water to dampen.

Insect Killer

Boric acid powder (not borax). Throw liberal amounts behind stove, refrigerator, under carpets and in carpets. Since boric acid is white, you must be careful not to mistake it for sugar accidentally. Keep it far away from food and out of children's reach. Buy it at a farm supply or garden store (or see *Sources*). It will not kill ants.

Ant Repellent

Spray 50% white distilled vinegar on counter tops, window sills and shelves and wipe, leaving residue. Start early in spring before they arrive, because it takes a few weeks to rid yourself of them once they are established. If you want immediate action, get some lemons, cut the yellow outer peel off and cover with grain alcohol in a tightly closed jar. Let stand at least one hour. Use 1 part of this concentrate with 9 parts water in a spray bottle. Mix only as much as you will use because the diluted form loses potency. Spray walls, floors, carpets wherever you see them. The lemon solution even leaves a shine on your counters. Use both vinegar and lemon approaches to rid yourself of ants.

To treat the **whole house**, pour vinegar all around your foundation, close to the wall, using one gallon for every five feet. Expect to damage any foliage it touches. Reapply every six months.

Flower and Foliage Spray

Food-grade hydrogen peroxide. See instructions on bottle.

Moth Balls

I found this recipe in an old recipe book. Mix the following and scatter in trunks and bags containing furs and woolens: ½lb. each rosemary and mint, ¼ lb. each tansy and thyme, 2 tbs. powdered cloves.

Carpet Cleaner

Whether you rent a machine or have a cleaning service, don't use the carpet shampoo they want to sell, even if they "guarantee" that it is all natural and safe. Instead add these to a bucket (about four gallons) of water and use it as the cleaning solution:

Wash water	Rinse water
1/3 cup borax	¼ cup grain alcohol
	2 tsp. boric acid
	¼ cup white distilled vinegar **or**
	4 tsp. citric acid

Borax does the cleaning; alcohol disinfects, boric acid leaves a pesticide residue, and the vinegar or citric acid give luster. If you are just making one pass on your carpet, use the borax, alcohol, and boric acid. Remember to test everything you use on an unnoticed piece of carpet first.

Health Improvement Recipes

Black Walnut Hull Tincture

This new recipe is four times as strong as the previous one, so it is called **Black Walnut Hull Tincture Extra Strength**.

Your largest enamel or ceramic (not stainless steel, not aluminum) cooking pot, preferably at least 10 quarts
Black walnuts, in the hull, each one still at least 50% green, enough to fill the pot to the top
Grain alcohol, about 50% strength, enough to cover the walnuts
½ tsp. vitamin C
Plastic wrap or cellophane
Glass jars or bottles

The black walnut tree produces large green balls in fall. The walnut is inside, but we will use the whole ball, uncracked, since the active ingredient is in the green outer hull.

Rinse the walnuts carefully, put them in the pot, and cover with the alcohol. Sprinkle on half the vitamin C. Seal with plastic wrap and cover. Let sit for three days. Pour into glass jars or bottles, discarding walnuts, and divide the remaining vitamin C amongst the jars. If the glass jar has a metal lid, first put plastic wrap over the top before screwing on the lid. Potency is strong for several years if unopened, even if it darkens.

You have just made Extra Strength Black Walnut Hull Tincture. It is stronger than the concentrate made with just a few black walnuts in a quart jar (my earlier recipe), because there are more walnuts per unit liquid. In addition, you will not dilute it before use (although when you take it, it will usually be in water).

When preparing the walnuts, rinse only with cold tap water. You may need to use a brush on areas with dirt. If you are not going to use all of them in this batch, you may freeze them in a resealable plastic bag. Simply refrigerating them does not keep them from turning black and useless. The pot of soaking walnuts should not be refrigerated. Nor does the final tincture need any refrigeration.

Exposure to air does cause the tincture to darken and lose potency. To reduce air exposure, fill the pot as much as possible, without touching the plastic wrap, while still keeping a snug fitting lid. Even more importantly, the glass jars or bottles you use to store your tincture should have as little air space as possible, without touching the plastic wrap on top. A large jar should be divided into smaller ones when you are ready to use it. The idea is not to have partial jars, with a lot of air space, sitting for longer than a month or so.

There are several ways to make a 50% grain alcohol solution. Some states have Everclear,^{TM} 95% alcohol. Mix this half and half with water. Other states have Everclear that is 76.5% alcohol. Mix this three parts Everclear to one part water. Yet another method is to buy vodka that is 100 proof. This is already 50% alcohol.

Remember, never use any kind of purchased water to make tincture.

Black Walnut Hull Tincture (Regular Strength)

This is the potency I used originally. It is included here in case you prefer it or wish to treat a pet. The Extra Strength recipe is four times as potent as the original recipe, so it must be diluted in quarters. (Similarly, if you have a lot of the Regular Strength left and want to use it in place of Extra Strength, simply take four times as much.)

Black Walnut Hull Tincture Extra Strength
Grain Alcohol, about 10%

Mix one part extra strength tincture with three parts of the 10% alcohol. Store in glass containers same as described above.

There are several ways to make a 10% grain alcohol solution. Some states have Everclear,™ 95% alcohol. Mix this one part Everclear to nine parts water. Other states have Everclear that is 76.5% alcohol. Mix this one part Everclear to seven parts water. Yet another method is to buy vodka that is 100 proof (50% alcohol) and mix one part vodka with four parts water.

Black Walnut Hull Extract (Water Based)

Because you do not know how commercially available extracts were made, and may not be able to test for solvent pollution, it is wisest to make it yourself!

This recipe is intended for alcoholic persons: cover the green balls in the 10 quart (non-metal) pot with cold tap water. Heat to boiling, covered. Turn off heat. When cool, add vitamin C, cover with plastic wrap, and the lid. Let stand for 1 day. It will be darker than the tincture. Do not dilute. Pour into freezable containers. Refrigerate what you will use in two days and freeze the rest. Add vitamin C after thawing or during refrigeration (¼ tsp. per quart).

For use: in programs calling for Extra Strength Black Walnut Hull Tincture use four times as much of this water based recipe (8 tsp. instead of 2 tsp. Extra Strength).

Important Note: do not use bottled or purchased water to make this tincture or you could pollute it with benzene!

Quassia recipe

Add 1/8 cup quassia chips to 3 cups water. Simmer 20-30 minutes. Pour off 1/8 cup now and drink it fresh. Refrigerate remainder. Drink 1/8 cup 4 times/day, until a total of ½cup of chips is consumed. Flavor with spices.

Emmenagogue (Menstrual Period Inducer)

Here are four herbs that can each bring on your period. They can be started anytime but the most-effective time is before your next calculated period time (count days as if you never missed a period).

1 oz sassafras bark
1 oz rue (cut)
1 oz marjoram herb
1 oz blue cohosh root
4½ cups boiling water

Add the herbs to the boiling water and turn down to simmer, covered, for 20 minutes. Do not boil. Strain and refrigerate in sterile glass jar. Pour one cup for yourself in the morning. Let warm to room temperature, and sip between meals, making it last until supper.

Bowel Program

Bacteria are always at the root of bowel problems, such as pain, bloating and gassiness. They can not be killed by zapping, because the high frequency current does not penetrate the bowel contents.

Although most bowel bacteria are beneficial, the ones that are not, like *Salmonellas* and *Shigellas*, are extremely detrimental because they have the ability to invade the rest of your body and colonize a trauma site or weakened organ. These same bacteria colonize a cancer tumor and delay healing after the malignancy is stopped.

Another reason bowel bacteria are so hard to eradicate is that we are constantly reinfecting ourselves by keeping a reservoir on our hands and under our fingernails.

- So the first thing to do is **improve sanitation**. For a serious problem, use 50% grain alcohol (100 proof vodka) in a spray bottle at the bathroom sink. Sterilize your hands after bathroom use and before meals.
- Secondly use **turmeric** (2 capsules 3 times a day, this is the common spice) which I find helps against *Shigella*, as well as *E. coli*. Expect orange colored stool.
- Third use **fennel** (1 capsule 3 times a day).
- Fourth use **digestive enzyme tablets** with meals as directed on the bottle. (But only as long as necessary, because these frequently harbor molds.)
- Fifth use a single 2 tsp. dose of **Black Walnut Hull Tincture Extra Strength**. Add it to a ½glass of water and sip over a 15 minute period. Stay seated until any side effect from the alcohol wears off.
- Sixth take **Cascara sagrada** capsules if constipated (start with one capsule a day, use up to maximum on label). Remember to drink a cup of hot water upon rising in the morning. This will begin to regulate your elimination.

It can take all six to get rid of a bad *Shigella* problem in a week. Afterward, you must continue to eat only sterile dairy products. Note that the Kidney Cleanse is often effective with bowel problems. Try it also.

You will know you succeeded when your tummy is flat, there is not a single gurgle, and your mood improves!

Constipation Tea

Constipation is often caused by *E. coli* and *Salmonella* from dairy foods, or from killing "good" bowel bacteria with antibiotics (killing a few by zapping actually restores good flora). Eat foods that restore the body's good bowel flora: vegetables, sterilized milk (the milk sugar is essential), lots of water.

There are a lot of remedies for constipation, but many people enjoy this tea:

1 tbs. senna tea leaves
½ tsp. mint leaves

Boil for one minute in a quart of water, add a dash of vitamin C and brown sugar to taste. Sip through the day to avoid "bellyache". It can take years for the body's flora to "right themselves" after an antibiotic session, be patient.

Weight Reduction

Here are two ancient herbal recipes for obesity. I have not personally determined their effectiveness.

Fucus

2 oz *Fucus vesiculosus*, cut (see Sources)
3 cups cold tap water

Boil for 15 minutes, covered. Cool. Dose: ¼ cup four times a day on an empty stomach. After one week increase dose to ½cup. You may add any flavoring desired.

Watch the pot carefully as it comes to a boil. If it boils over, you will have a month of stove-cleaning to do. The odor of *Fucus* boiling is wretched. So is the taste. Maybe garlic (fresh) would improve it.

Fennel

1 oz fennel seed (crushed or powdered is fine)

3 cups cold tap water

Boil water, pour over herb. Steep 30 minutes. Strain. Add 4 oz honey (optional). Drink one cup each day.

You could take them both together, along with the Bowel Program, to be more successful, but the best single weight reducer is the Liver Cleanse.

Kidney Cleanse

½ cup dried Hydrangea root
½ cup Gravel root
½ cup Marshmallow root
4 bunches of fresh parsley
Goldenrod tincture (leave this out of the recipe if you are allergic to it)
Ginger capsules
Uva Ursi capsules
Vegetable glycerin
Black Cherry Concentrate, 8 oz
Vitamin B6, 250 mg
Magnesium oxide tablets, 300 mg

Measure ¼ cup of each root and set them to soak, together in 10 cups of cold tap water, using a non-metal container and a non-metal lid (a dinner plate will do). After four hours (or overnight) add 8 oz. black cherry concentrate, heat to boiling and simmer for 20 minutes. Drink ¼ cup as soon as it is cool enough. Pour the rest through a bamboo strainer into a sterile pint jar (glass) and several freezable containers. Refrigerate the glass jar.

Boil the fresh parsley, after rinsing, in 1 quart of water for 3 minutes. Drink ¼ cup when cool enough. Refrigerate a pint and freeze 1 pint. Throw away the parsley.

Dose: each morning, pour together ¾cup of the root mixture and ½cup parsley water, filling a large mug. Add 20 drops of goldenrod tincture and 1 tbs. of glycerin. Drink this mixture in divided doses throughout the day. Keep cold. Do not drink it all at once or you will get a stomach ache and feel pressure in your bladder. If your stomach is very sensitive, start on half this dose.

Save the roots after the first boiling, storing them in the freezer. After 13 days when your supply runs low, boil the same roots a second time, but add only 6 cups water and simmer only 10 minutes. This will last another 8 days, for a total of three weeks. You may cook the roots a third time if you wish, but the recipe gets less potent. If your problem is severe, only cook them twice.

After three weeks, repeat with fresh herbs. You need to do the Kidney Cleanse for six weeks to get good results, longer for severe problems.

Also take:

- Ginger capsules: one with each meal (3/day).
- Uva Ursi capsules: one with breakfast and two with supper.
- Vitamin B6 (250 mg): one a day.
- Magnesium oxide (300 mg): one a day.

Take these supplements just before your meal to avoid burping.

Some notes on this recipe: this herbal tea, as well as the parsley, can easily spoil. Heat it to boiling every fourth day if it is being stored in the refrigerator; this resterilizes it. If you sterilize it in the morning you may take it to work without refrigerating it (use a glass container).

When you order your herbs, be careful! Herb companies are not the same! These roots should have a strong fragrance. If the ones you buy are barely fragrant, they have lost their active ingredients; switch to a different supplier. Fresh roots can be used. Do not use powder.

- Hydrangea (*Hydrangea arborescens*) is a common flowering bush.
- Gravel root (*Eupatorium purpureum*) is a wild flower.
- Marshmallow root (*Althea officinalis*) is mucilaginous and kills pain.
- Fresh parsley can be bought at a grocery store. Parsley flakes and dried parsley herb do not work.
- Goldenrod herb works as well as the tincture but you may get an allergic reaction from smelling the herb. If you know you are allergic to this, leave this one out of your recipe.
- Ginger from the grocery store works fine; you may put it into capsules for yourself (size 0, 1 or 00).

There are probably dozens of herbs that can dissolve kidney crystals and stones. If you can only find several of those in the recipe, make the recipe anyway; it will just take longer to get results. Remember that vitamin B_6 and magnesium, taken daily, can prevent oxalate stones from forming. But only if you stop drinking tea. Tea has 15.6 mg oxalic acid per cup[37]. A tall glass of iced tea could give you over 20 mg oxalic acid. Switch to herb teas. Cocoa and chocolate, also, have too much oxalic acid to be used as beverages.

Remember, too, that phosphate crystals are made when you eat too much phosphate. Phosphate levels are high in meats, breads, cereals, pastas, and carbonated drinks. Eat less of these, and increase your milk (2%), fruits and vegetables. Drink at least 2 pints of water a day.

---

[37] Taken from *Food Values* 14ed by Pennington and Church 1985.

Cleanse your kidneys at least twice a year.

You can dissolve all your kidney stones in 3 weeks, but make new ones in 3 days if you are drinking tea and cocoa and phosphated beverages. None of the beverage recipes in this chapter are conducive to stone formation.

Liver Herbs

Don't confuse these liver herbs with the next recipe for the Liver Cleanse. This recipe contains herbs traditionally used to help the liver function, while the Liver Cleanse gets gallstones out.

6 parts comfrey root, *Symphytum officinale* (also called nipbone root)
6 parts tanner’s oak bark, *Quercus alba* (white oak bark)
3 parts gravel root, *Eupatorium purpureum* (queen of the meadow)
3 parts Jacob’s staff, *Verbascum thapsus* (mullein herb)
2 parts licorice root, *Glycyrrhiza glabra*
2 parts wild yam root, *Dioscorea villosa*
2 parts milk thistle herb, *Silybum marianum*
3 parts walnut bark, *Juglans nigra*, (black walnut bark)
3 parts marshmallow root, *Althea officinalis* (white mallow)
1 part lobelia plant, *Lobelia inflata* (bladder pod)
1 part skullcap, *Scutellaria lateriflora* (helmet flower)

Mix all the herbs. Add ½cup of the mixture to 2 quarts of water. Bring to a boil. Put lid on. Let sit for six hours. Strain and drink 1½cups per day. Put the strained herbs in the freezer and use them one more time.

Liver Cleanse

Cleansing the liver of gallstones dramatically improves digestion, which is the basis of your whole health. You can expect your allergies to disappear, too, more with each cleanse you do! Incredibly, it also eliminates shoulder, upper arm, and upper back pain. You have more energy and increased sense of well being.

Cleaning the liver bile ducts is the most powerful procedure that you can do to improve your body's health.

But it should not be done before the parasite program, and for best results should follow the kidney cleanse and any dental work you need.

It is the job of the liver to make bile, 1 to 1½quarts in a day! The liver is full of tubes (*biliary tubing*) that deliver the bile to one large tube (the *common bile duct*). The gallbladder is attached to the common bile duct and acts as a storage reservoir. Eating fat or protein triggers the gallbladder to squeeze itself empty after about twenty minutes, and the stored bile finishes its trip down the common bile duct to the intestine.

For many persons, including children, the biliary tubing is choked with gallstones. Some develop allergies or hives but some have no symptoms. When the gallbladder is scanned or X-rayed nothing is seen. Typically, they are not in the gallbladder. Not only that, most are too small and not calcified, a prerequisite for visibility on X-ray. There are over half a dozen varieties of gallstones, most of which have cholesterol crystals in them. They can be black, red, white, green or tan colored. The green ones get their color from being coated with bile. Notice in the picture how many have imbedded unidentified objects. Are they fluke remains? Notice how many are shaped like corks with longitudinal grooves below the tops. We can visualize the blocked bile ducts from such shapes. Other stones are composites–made of many smaller ones–showing that they regrouped in the bile ducts some time after the last cleanse.

At the very center of each stone is found a clump of bacteria, according to scientists, suggesting a dead bit of parasite might have started the stone forming.

As the stones grow and become more numerous the back pressure on the liver causes it to make less bile. Imagine the situation if your garden hose had marbles in it. Much less water would flow, which in turn would decrease the ability of the hose to squirt out the marbles. With gallstones, much less cholesterol leaves the body, and cholesterol levels may rise.

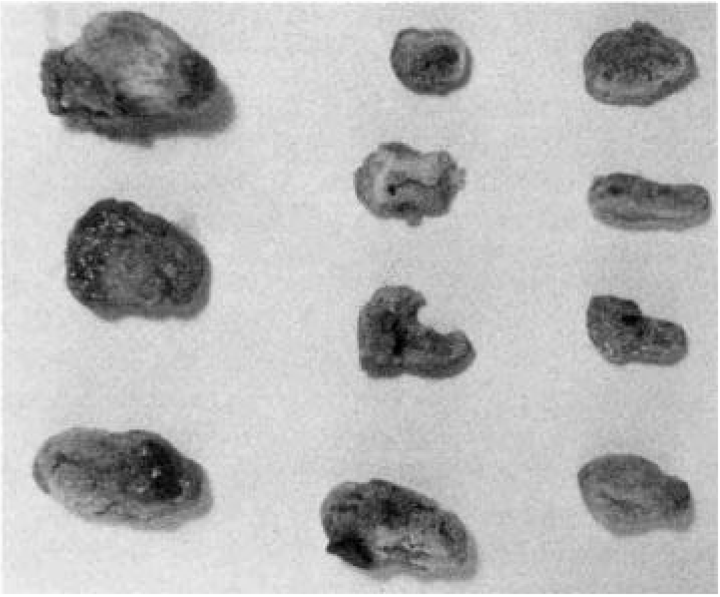

Fig. 89 These are gallstones.

Gallstones, being porous, can pick up all the bacteria, cysts, viruses and parasites that are passing through the liver. In this way "nests" of infection are formed, forever supplying the body with fresh bacteria. No stomach infection such as ulcers or intestinal bloating can be cured permanently without removing these gallstones from the liver.

Cleanse your liver twice a year.

Preparation.

- You can't clean a liver with living parasites in it. You won't get many stones, and you will feel quite sick. Zap daily the week before, or get through the first three weeks of the parasite killing program before attempting a liver cleanse. If you are on the maintenance parasite program, you are always ready to do the cleanse.
- Completing the kidney cleanse before cleansing the liver is also highly recommended. You want your kidneys, bladder and urinary tract in top working condition so they can efficiently remove any undesirable substances incidentally absorbed from the intestine as the bile is being excreted.
- Do any dental work first, if possible. Your mouth should be metal free and bacteria free (cavitations are cleaned). A toxic mouth can put a heavy load on the liver, burdening it immediately after cleansing. Eliminate that problem first for best results.

Ingredients

Epsom salts	4 tablespoons
Olive oil	half cup (light olive oil is easier to get down)
Fresh pink grapefruit	1 large or 2 small, enough to squeeze 2/3 to 3/4 cup juice

Ornithine	4 to 8, to be sure you can sleep. Don't skip this or you may have the worst night of your life!
Large plastic straw	To help drink potion.
Pint jar with lid	

Choose a day like Saturday for the cleanse, since you will be able to rest the next day.

Take no medicines, vitamins or pills that you can do without; they could prevent success. Stop the parasite program and kidney herbs, too, the day before.

Eat a no-fat breakfast and lunch such as cooked cereal with fruit, fruit juice, bread and preserves or honey (no butter or milk), baked potato or other vegetables with salt only. This allows the bile to build up and develop pressure in the liver. Higher pressure pushes out more stones.

2:00 PM. Do not eat or drink after 2 o'clock. If you break this rule you could feel quite ill later.

Get your Epsom salts ready. Mix 4 tbs. in 3 cups water and pour this into a jar. This makes four servings, ¾cup each. Set the jar in the refrigerator to get ice cold (this is for convenience and taste only).

6:00 PM. Drink one serving (¾cup) of the ice cold Epsom salts. If you did not prepare this ahead of time, mix 1 tbs. in ¾cup water now. You may add 1/8 tsp. vitamin C powder to improve the taste. You may also drink a few mouthfuls of water afterwards or rinse your mouth.

Get the olive oil and grapefruit out to warm up.

8:00 PM. Repeat by drinking another ¾cup of Epsom salts.

You haven't eaten since two o'clock, but you won't feel hungry. Get your bedtime chores done. The timing is critical for success; don't be more than 10 minutes early or late.

9:45 PM. Pour ½cup (measured) olive oil into the pint jar. Squeeze the grapefruit by hand into the measuring cup. Remove pulp with fork. You should have at least ½cup, more (up to ¾ cup) is best. You may top it up with lemonade. Add this to the olive oil. Close the jar tightly with the lid and shake hard until watery (only fresh grapefruit juice does this).

Now visit the bathroom one or more times, even if it makes you late for your ten o'clock drink. Don't be more than 15 minutes late.

10:00 PM. Drink the potion you have mixed. Take 4 ornithine capsules with the first sips to make sure you will sleep through the night. Take 8 if you already suffer from insomnia. Drinking through a large plastic straw helps it go down easier. You may use ketchup, cinnamon, or brown sugar to chase it down between sips. Take it to your bedside if you want, but drink it standing up. Get it down within 5 minutes (fifteen minutes for very elderly or weak persons).

Lie down immediately. You might fail to get stones out if you don't. The sooner you lie down the mote stones you will get out. Be ready for bed ahead of time. Don't clean up the kitchen. As soon as the drink is down walk to your bed and lie down flat on your back with your head up high on the pillow. Try to think about what is happening in the liver. Try to keep perfectly still for at least 20 minutes. You may feel a train of stones traveling along the bile ducts like marbles. There is no pain because the bile duct valves are open (thank you Epsom salts!). **Go to sleep**, you may fail to get stones out if you don't.

Next morning. Upon awakening take your third dose of Epsom salts. If you have indigestion or nausea wait until it is gone before drinking the Epsom salts. You may go back to bed. Don't take this potion before 6:00 am.

2 Hours Later. Take your fourth (the last) dose of Epsom salts. Drink ¾cup of the mixture. You may go back to bed.

After 2 More Hours you may eat. Start with fruit juice. Half an hour later eat fruit. One hour later you may eat regular food but keep it light. By supper you should feel recovered.

How well did you do? Expect diarrhea in the morning. Use a flashlight to look for gallstones in the toilet with the bowel movement. Look for the green kind since this is proof that they are genuine gallstones, not food residue. Only bile from the liver is pea green. The bowel movement sinks but gallstones float because of the cholesterol inside. Count them all roughly, whether tan or green. You will need to total 2000 stones before the liver is clean enough to rid you of allergies or bursitis or upper back pains permanently. The first cleanse may rid you of them for a few days, but as the stones from the rear travel forward, they give you the same symptoms again. You may repeat cleanses at two week intervals. Never cleanse when you are ill.

Sometimes the bile ducts are full of cholesterol crystals that did not form into round stones. They appear as a “chaff” floating on top of the toilet bowl water. It may be tan colored, harboring millions of tiny white crystals. Cleansing this chaff is just as important as purging stones.

How safe is the liver cleanse? It is very safe. My opinion is based on over 500 cases, including many persons in their seventies and eighties. None went to the hospital; none even reported pain. However it can make you feel quite ill for one or two days afterwards, although in every one of these cases the maintenance parasite program had been neglected. This is why the instructions direct you to complete the parasite and kidney rinse programs first.

CONGRATULATIONS

You have taken out your gallstones without surgery! I like to think I have perfected this recipe, but I certainly can not take credit for its origin. It was invented hundreds, if not thousands, of years ago, THANK YOU, HERBALISTS!

This procedure contradicts many modern medical viewpoints. Gallstones are thought to be formed in the gallbladder, not the liver. They are thought to be few, not thousands. They are not linked to pains other than gallbladder attacks. It is easy to understand why this is thought: by the time you have acute pain attacks, some stones are in the gallbladder, are big enough and sufficiently calcified to see on X-ray, and have caused inflammation there. When the gallbladder is removed the acute attacks are gone, but the bursitis and other pains and digestive problems remain.

The truth is self-evident. People who have had their gallbladder surgically removed still get plenty of green, bile-coated stones, and anyone who cares to dissect their stones can see that the concentric circles and crystals of cholesterol match textbook pictures of "gallstones" exactly.

Lugol's Iodine Solution

It is too dangerous to buy a commercially prepared solution. It is certain to be polluted with propyl alcohol or wood alcohol. Make it yourself or ask your pharmacist to make it up for you. The recipe to make 1 liter (quart) is:

44 gm (1½ ounces) iodine, granular
88 gm (3 ounces) potassium iodide, granular

Dissolve the potassium iodide in about a pint of the water. Then add the iodine crystals and fill to the liter mark with water. It takes about 1 day to dissolve completely. Shake it from time to time. Keep out of sight and reach of children. Do not use if allergic to iodine. Be careful to avoid bottled water for preparation.

Vitamin D Drops

1 gram cholecalciferol (see Sources)
10 cups olive oil

Mix in a non-metal container. It may take a day of standing to dissolve fully. Refrigerate. Ten drops contain 40,000 iu. Use within a year.

Pathogen Frequencies

Living creatures emit a range of frequencies, also called *bandwidth*. As they age, the bandwidth shrinks. When they die sometimes all that is left is a single frequency.

Most of the organisms listed below are dead on commercially available and prepared slides (see *Sources* for biological supply companies). However they still exhibit a 5 KHz bandwidth, probably due to testing with a frequency generator that was only accurate to 100 Hz, and also due to using more voltage than necessary (like when a powerful radio station comes in at its own frequencies and ones nearby, too). Some testing was done with a more accurate frequency generator at a lower power level so some bandwidths are reported much more narrowly.

If the same person retests the same specimens with the same equipment within a few days, the results will be absolutely identical (within 1 Hz) 90% of the time. Why a few of the results will not be identical is not known. However different people, and even the same person at different times of the year, can notice that the perceived frequencies shift by as much as 3 KHz (still less than 1% change).

Some specimens have more than one range listed; this may be characteristic of the organism or may be due to having an undocumented organism on the same microscope slide.

Blank locations represent organisms for whom there are prepared slides available, but whose bandwidth has not been determined.

Bandwidth Of Organism Families

In general, the smaller the organism the lower the frequency and narrower the bandwidth. This chart shows the major families studied and where they fall in the spectrum.

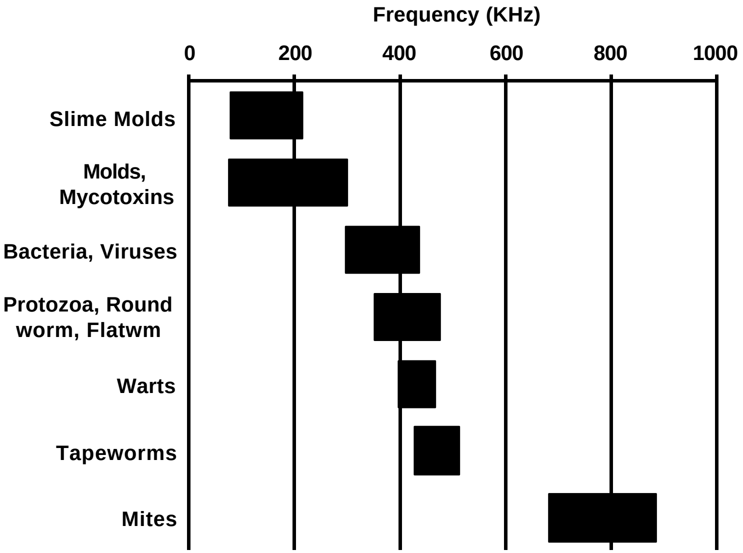

Fig. 90 Chart of bandwidths for organism families.

Mold, Mold Toxin Frequencies

Other molds and mold toxins	**KHz**
Aflatoxin	177,188
Cytochalasin B	77,91
Ergot	295
Griseofulvin	288
Sorghum syrup	277
Sterigmatocystin	88,96,133,126
Zearalenone	100

Slime Molds	**KHz**
Arcyria	81
Lycogala	126
Stemonitis	211

Bacteria and Viruses

Including locations where I commonly found them.

	Low Freq (KHz)	**High Freq (KHz)**	**Use freq gen for 3 min @**
Acetobacter aceti			
Adenovirus	393	393	393
Adenovirus (2nd range)	371.45	386.90	
Agrobacterium tumefaciens			
Alcaligenes faecalis			
Alpha streptococcus	369.75	385.4	380,375
Azobacter chroococcum			
Bacillus anthracis causes anthrax in cattle (tooth)	393.5	398.05	395,364,368
Bacillus anthracis (2nd range)	363.2	365.3	
Bacillus anthracis (3rd range)	359.4	370.5	
Bacillus anthracis spores	386.95	391.45	388
Bacillus cereus	373.65	375.85	374.5
Bacillus megaterium			
Bacillus sterothermophilus			
Bacillus subtilis spores			
Bacillus subtilis var. niger	371.85	387.1	385,380,375
Bacteria capsules (capsular strain)	416.05	418.75	417.5
Bacterial capsules	357.6	362.4	360
Bacteroides fragilis found with common roundworm Ascaris	324.3	325.0	325
Bacteroides fragilis (2nd range)	325.7	326.0	
Beta streptococcus (tooth)	380.6	387.4	385
Blepharisma	405.65	407.45	406.5
Bordetella pertussis “whooping cough” (tooth)	329.85	332.25	331
Borellia burgdorferi Lyme disease	378.95	382.0	380
Branhamella (Neisseria) catarrhalis (has hole at 398)	394.9	396.7	396
Brucella abortus			
Cabbage Black Rot			
Campylobacter fetus smear	365.3	370.6	368
Campylobacter pyloridis	352.0	357.2	355
Candida albicans (pure powder) common yeast	384.2	388.4	386
Caulobacter vibrioides			
Central spores (bacillus smear)	372.45	378.65	376
Chlamydia trachomatis	379.7	383.95	381
Clostridium acetobutylicum	382.8	391.15	389,384
Clostridium botulinum (tooth) causes food poisoning	361.0	364.55	362

Clostridium perfringens			
Clostridium perfringens spores	394.2	398.1	396
Clostridium septicum	362.05	365.6	364
Clostridium sporogenes			
Clostridium tetani (tooth) causes tetanus			
Corynebacterium diphtheriae (tooth) causes diphtheria	340	344	342
Corynebacterium pseudodiphthericum			
Corynebacterium xerosis	315.65	316.8	316.0
Coxsackie virus B-1 found with Bacteroides fragilis	360.5	366.1	364
Coxsackie virus B-4 found with Bacteroides fragilis	361.45	363.7	362.5
Coxsackie virus B-4 (2nd range)	363.9	364.9	
Crithidia fasciculata			
Cytomegalovirus (CMV) antigen	408.35	410.75	409
Cytophaga rubra	428.1	432.2	430
Diplococcus diphtheriae	357.95	264.0	361
Diplococcus pneumoniae	351.65	368.45	365,360
Eikanella corrodens	379.5	384.3	382
Enterobacter aerogenes intestinal bacterium	374	374	374
Epstein Barre virus (EBV)	372.5	382.85	380,375
Erwinia amylovora	347.2	352.1	350
Erwinia carotovora	368.1	377.0	373
Escherichia coli (E. coli) intestinal bacterium	356	356	356,393
Escherichia coli (E. coli) (2nd range)	392	393	
Gaffkya tetragena causes respiratory infections	344.85	352.5	350
Gardnerella vaginalis ovarian and genital tract infection	338.0	342.55	340
Haemophilus influenzae bacterial meningitis, infects joints	336.41	336.41	336
Hepatitis B antigen	414.55	420.8	418
Herpes simplex 1	291.25	293.05	292,345.5
Herpes simplex 1 (2nd range)	345.35	345.75	
Herpes simplex 2 (fresh smear)	353.9	362.9	360,355
Herpes Zoster “shingles”	416.6	420.2	418
Histomonas meleagridis (liver)	376.55	378.7	377
Histoplasma capsulatum	298.3	304.85	302
HIV	365	365	365
Influenza A and B (flu shot)	313.35	323.9	320,315
Iron Bacterium Sphaerotilus			
Klebsiella pneumoniae causes pneumonia	398.45	404.65	401,419
Klebsiella pneumoniae (2nd range)	416.9	421.9	
Lactobacillus acidophilus (tooth)	346.05	351.65	349

Leptospira interrogans spirochete	397.05	401.1	399
Lumpy Jaw			
Measles antigen	369.5	373.0	371
Micrococcus luteus			
Micrococcus roseus			
Mumps antigen	377.6	384.65	382
Mycobacterium para TB			
Mycobacterium phlei	409.65	410.65	410.0
Mycobacterium smegmatis			
Mycobacterium tuberculosis (infec nodule) causes tuberculosis	430.55	434.2	432
Mycoplasma	322.85	323.9	323.5,346
Mycoplasma (range 2)	342.75	349.3	
Neisseria gonorrhea causes gonorrhea	333.85	336.5	334
Neisseria sicca			
Nocardia asteroides found in Parkinson's Disease	354.95	355.35	355.1,368
Nocardia asteroides (2nd range)	363.7	370	
Propionobacterium acnes	383.75	389.0	387
Proteus mirabilis	320.55	326.0	324,349
Proteus mirabilis (2nd range)	345.95	352.1	
Proteus vulgaris urinary tract pathogen	408.75	416.45	413,336,328
Proteus vulgaris (2nd range)	333.75	339.15	
Proteus vulgaris (3rd range)	327.2	329.5	
Pseudomonas aeruginosa found in open wounds	331.25	334.6	333
Pseudomonas fluorescens			
Respiratory syncytial virus	378.95	383.15	380
Rhizobium leguminosarum			
Salmonella enteriditis intestinal infection	329	329	329
Salmonella paratyphi	365.05	370.1	368,385
Salmonella typhimurium food poisoning, nervousness, apathy	382.3	386.55	355,386,390
Serratia marcescens	349.45	352.1	351
Shigella dysenteriae intestinal problems	390.089	390.089	390.089
Shigella flexneri depression	394	394	394
Shigella sonnei invades tumors	318	318	318
Sphaerotilus natans	388.4	393.45	391
Spirillum itersonil			
Spirillum serpens	378.35	382.8	380
Spirillum sinuosum			
Spirillum volutans			
Spores in bacteria spore stain			
Staphylococcus aureus (culture)	376.27	380.85	
Staphylococcus aureus (slide) source is tooth infection, causes abscesses, heart disease, invades tumors	381	381	378,381

Staphylococcus epidermidis infects skin and mucous membranes			
Streptococcus lactis occurs in milk	382	387	385
Streptococcus mitis lung infection, tooth infection, abscesses, causes stiff knees	313.8	321.1	318
Streptococcus pneumoniae causes pneumonia and inner ear disease	366.85	370.2	368
Streptococcus pyogenes (tooth)	360.5	375.3	373
Streptococcus sp. group G (tooth)	368.15	368.85	368
Sub terminal spores bac. smear	385.15	385.95	
Terminal spores bacillus smear			
Tobacco mosaic virus (tobacco)	427.15	429.55	428
Treponema pallidum causes syphilis	346.85	347.4	347
Troglodytella abrassari	377.75	385.2	383,419
Troglodytella abrassari (2nd range)	416.9	422.2	
Veillonella dispar	401.75	405.2	403
Vibrio (photobacterium) fischeri			

Roundworms, Flatworms, One-celled Animals

	Low Freq (KHz)	**High Freq (KHz)**	**To kill, use freq. gen for 3 min. at these frequencies**
Acanthamoeba culbertsoni			
Acanthocephala			
Anaplasma marginale	386.4	388.0	387,422
Anaplasma marginale (2nd range)	415.3	424	
Ancylostoma braziliense (adult)	397.6	403.25	401
Ancylostoma caninum	383.1	402.9	400,393,386
Ancylostoma duodenale male			
Anguillula aceti			
Ascaris larvae in lung common roundworm of cats and dogs	404.9	409.15	408
Ascaris lumbricoides (m and f)			same
Ascaris megalocephala (male)	403.85	409.7	408
Babesia bigemina			
Babesia canis smear			
Balantidium coli cysts	458.8	462.9	460
Balantidium sp. trophozoites (from guinea pig) parasitic ciliate			
Besnoitia (lung sect.) protozoan	352.8	361.4	358

Capillaria hepatica (liver sect.)	424.25	430.65	428
Chilomastix cysts (rat)	388.95	390.7	389,426
Chilomastix cysts (rat) (2nd range)	425.2	427.3	
Chilomastix mesnili (trophozoites)			same
Chilomonas, whole mount	393.75	400	398
Clinostomum metacercaria			
Clonorchis metacercariae			
Clonorchis sinensis	425.7	428.75	427
Clonorchis sinensis eggs			
Cryptocotyle lingua (adult)	409.95	416.0	414
Didinium			
Dientamoeba fragilis	401.35	406.05	404
Dipetalonema perstans (microfilaria human blood)			
Dirofilaria immitis dog heartworm	408.15	411.15	409
Echinoporyphium recurvatum	418.55	423.9	421
Echinostoma revolutum	425.5	429.65	428
Eimeria stiedae			
Eimeria tenella			
Endamoeba gingivalis trophozoite	433.8	441.0	438
Endolimax nana trophozoites and cysts	394.25	397.1	396,432
Endolimax nana trophozoites and cysts (2nd range)	430.5	433.35	
Entamoeba coli cysts			
Entamoeba coli trophozoites	397.0	400.35	398
Entamoeba histolytica trophozoite	381.1	387.8	385
Enterobius vermicularis	420.95	426.3	423
Eurytrema pancreaticum	420.35	422.3	421
Eurytrema pancreaticum stages			
Fasciola hepatica	421.35	427.3	425
Fasciola hepatica cercariae	423.8	430.6	427
Fasciola hepatica eggs	422.0	427.6	425
Fasciola hepatica metacercariae			
Fasciola hepatica miracidia	421.75	424.7	423
Fasciola hepatica rediae	420.6	427.5	425
Fasciolopsis buskii adult	427.7	435.1	434
Fasciolopsis buskii eggs	427.35	435.45	434
Fasciolopsis buskii eggs unincubated			
Fasciolopsis cercariae	429.5	436.25	434
Fasciolopsis miracidia	427.35	435.2	434
Fasciolopsis rediae	427.3	433.0	432
Fischoedrius elongatus	441.75	443.2	442
Gastrothylax elongatus	451.9	457.1	455
Giardia lamblia (trophozoites)	421.4	426.3	424
Giardia lamblia cysts			
Gyrodactylus	378.75	381.8	380
Haemonchus contortus	386.8	395.5	393
Haemoproteus			

Hasstile sig. tricolor (adult)	448.05	455.1	453
Heterakis			
Hypodereum conoideum	424.45	429.55	427
Iodamoeba butschlii trophozoites and cysts	437.85	448.5	445,402
Iodamoeba butschlii trophozoites and cysts (2nd range)	398.15	404.75	
Leishmania braziliensis	400.05	405.1	403
Leishmania donovani	398.0	402.65	400
Leishmania mexicana	400.2	403.8	402
Leishmania tropica	402.1	407.4	405
Leucocytozoon	397.45	402.55	400
Loa loa	360.551	360.551	361
Macracanthorhynchus	438.85	442.8	440
Metagonimus Yokogawai	437.35	442.1	440
Monocystis agilis			
Myxosoma	409.6	416.95	414
Naegleria fowleri	356.9	364.35	362
Naegleria fowleri (brain sec.)			
Necator americanus (infect larvae)			
Notocotylus quinqeserialis			
Onchocerca volvulus (tumor)	436.3	442.1	440
Paragonimus Westermanii adult	437.8	454.2	452,447
Passalurus ambiguus	428.8	444.15	441,437
Pelomyxa carolinensis			
Plasmodium cynomolgi	417.3	424.5	422
Plasmodium falciparum smear	372.3	373.8	373.0
Plasmodium vivax smear	438.15	445.1	442
Platynosomum fastosum adult			
Pneumocystis carnii (lung)	405.75	409.15	407
Prosthogonimus macrorchis(eggs)	396.85	404.75	401
Sarcina lutea			
Sarcocystis	450.55	454.95	452
Schistosoma haematobium	473	473	473
Schistosoma japonicum eggs			
Schistosoma mansoni	353	353	353
Stephanurus dentalus (ova)	457.35	463.1	461
Stigeoclonium	404.25	415.25	412,407
Strongyloides (filariform larva)	398.4	402.0	400
Strongyloides parasitic females			
Toxocara (eggs)			
Toxoplasma (human strain)	395.0	395.0	395
Trichinella spiralis (muscle)	403.85	405.57	404.5
Trichomonas muris			
Trichomonas vaginalis	378.0	383.6	381
Trichuris sp. (male)	388.3	408.9	406
Trypanosoma brucei	423.2	431.4	429
Trypanosoma cruzi (brain tissue)	460.2	465.65	463

Trypanosoma equiperdum	434.6	451.25	448,442,438
Trypanosoma gambiense	393.75	398.7	396
Trypanosoma lewisi (blood smear)	424.5	426.0	425
Trypanosoma rhodesiense	423.5	428.55	426
Urocleidus	442.35	450.0	447

Wart Frequencies

(Most of these are from homemade slides.)

	Low Freq	**High Freq**	**Use freq gen for 3 min @**
Wart BS	402	406	404
Wart CC	426	432.35	430
Wart FR	459.3	464.75	462
Wart HA	434.8	444.1	442,437
Wart HRCm	438.9	448.55	446,441
Wart human papilloma plantar	404.7	406.75	405
Wart human papilloma virus	402.85	410.7	407
Wart JB	418.75	422.4	420
Wart L arm	343.65	345.95	344
Wart papilloma cervix smear	404.05	404.6	404.3

Tapeworms

Tapeworms are segmented. The first segment is the head, called the *scolex*. Tapeworms grow by adding a new segment to their body.

Tapeworms can have very large bandwidths (range of frequencies), and it varies by the length of the specimen! It is as if each new segment has a unique, and slightly lower, frequency.

Do not use a frequency generator to kill tapeworms. If you accidentally kill middle segments instead of working your way up from the bottom, you may conceivably promote dispersion! Use only a zapper.

	Low Freq	**High Freq**
Cysticercus fasciolaris	436.4	440.05
Diphyllobothrium erinacei (Mansoni) (scolex)	467.25	487.55
Diphyllobothrium erinacei eggs		
Diphyllobothrium latum (scolex)	452.9	472.3
Dipylidium caninum (proglottid composite)	439.55	444.3
Dipylidium caninum (scolex)	451.95	472.15
Echinococcus granulosus	451.6	461.5
Echinococcus granulosus (cysts)	441.15	446.5
Echinococcus granulosus (eggs)		
Echinococcus multilocularis	455.85	458.35
Heterophyes heterophyes		
Hymenolepis cysticercoides	478.0	481.75
Hymenolepis diminuta	445	481.15
Hymenolepis diminuta ova		
Hymenolepis nana eggs		
Moniezia (scolex)	430.35	465.2
Moniezia expansa (composite)	430.35	465.2
Moniezia expansa eggs		
Multiceps serialis	453.6	457.8
Pigeon tapeworm		
Taenia pisiformis (cysticercus)	475.2	482.1
Taenia pisiformis eggs (ova)	465.2	469.7
Taenia saginata (cysticercus)	476.5	481.05
Taenia saginata eggs		
Taenia solium (cysticercus)	475	475
Taenia solium (scolex)	444.0	448.9
Taenia solium eggs		

Mite Frequencies

These are the organisms that cold viruses ride in with!

Mite	**KHz**
Demodex folliculorum follicle mite	682
Dermatophagoides dust mite	707
Meal mite	718
Ornithonyssus bird mite	877,878
Sarcoptes scabei itch	735

Miscellaneous Frequencies

	KHz
Blue-green Algae	256
Bryozoa cristatalla	396
Mucor mucedo	288
Rhizobium meliloti	330
Rotifer	1151

It's easy to make homemade slides when you or a family member is ill. Finding out the frequencies of these illnesses helps you identify them (use the Pathogen Frequency Chart) and also lets you know if you are chronically getting them back.

Unidentified pathogens	**Low Freq**	**High Freq**
A cold virus HRC	395.8	395.8
Fungus EW	362.0	364.9
Fungus JWB	397.2	400.75
Tooth decay	384.3	387.2
Tooth decay (N)	367.9	375.05
Tooth decay (N) (2nd range)	326.95	331.5
Tooth decay (N) (3rd range)	293.2	297.4
Tooth plaque I	378.8	383.05
Tooth plaque I (2nd range)	294.7	298.25
Tooth plaque I (3rd range)	233.1	238.2
Tooth plaque II	384.95	387.05
Tooth plaque II (2nd range)	278.75	284
Tooth plaque II (3rd range)	212.15	218
Tooth plaque II (4th range)	340.15	344.8
Tooth plaque II (5th range)	305.5	310.35

Toxic Elements

Although not living, solvents and toxins must exhibit characteristic frequencies, otherwise how could the Syncrometer detect specific ones? This needs further exploration.

Most of the toxic elements I use are metals, heavy metals and lanthanides. But some are not; examples are PCBs and formaldehyde.

Some important elements are missing, like iron, zinc and manganese. This is because I never could find them present in the white blood cells, and I finally gave up searching for them.

Below is a list of the 70 or so toxic elements I use. Most of them were obtained as Atomic Absorption Standard Solutions and are, therefore, very pure. This prevents mistakes in identifying a toxin. They were stored in ½ounce amber glass bottles with bakelite caps and permanently sealed with plastic film since testing did not require them to be opened (they get close enough to the frequency field). The exact concentration and the solubility characteristics are not important in this qualitative test. The main sources of these substances in our environment are given beside each item.

Toxic Substance	**Sources**
Aflatoxin B	beer, bread, apple cider vinegar, moldy fruit, nuts
Aluminum	cookware, deodorant, lotions, soaps
Aluminum silicate	salt, water softener
Antimony	fragrance in lotions, colognes
Arsenic	pesticide, “treated” carpet, wallpaper
Asbestos	clothes dryer belt, hair blower, paint on radiators
Barium	lipstick, bus exhaust
Benzalkonium chloride	toothpaste
3,4 Benzopyrene	flame cooked foods, toast
4,5 Benzopyrene	flame cooked foods, toast
Beryllium	hurricane lamps, gasoline, dentures, kerosene
Bismuth	colognes, lotions, antacids
Boron	
Bromine	bleached “brominated” flour
Cadmium	galvanized water pipes, old tooth fillings
Cerium	tooth fillings
Cesium	clear plastic bottles used for beverages
Chlorine	from Chlorox™ bleach
Chromium	cosmetics, water softener
Cobalt	detergent, blue and green body products
Copper	tooth fillings, water pipes
Dysprosium	paint and varnish
Erbium	packaging for food, pollutant in pills
Europium	tooth fillings
Europium oxide	tooth fillings, catalytic converter

Fiberglass	dust from remodeling or building insulation
Formaldehyde	foam in mattresses and furniture, paneling
Gadolinium	tooth fillings
Gallium	tooth fillings
Germanium	with thallium in tooth fillings (pollutant)
Gold	tooth fillings
Hafnium	hair spray, nail polish, pollutant in pills
Holmium	usually found in presence of PCBs
Indium	tooth fillings
Iridium	tooth fillings
Lanthanum	computer and printing supplies
Lead	solder joints in water pipes
Lithium	printing supplies
Lutetium	paint and varnish
Mercury	tooth fillings
Molybdenum	auto supplies
Neodymium	pollutant in pills
Nickel	tooth fillings, metal glasses frames
Niobium	pollutant in pills, foil packaging for food
Palladium	tooth fillings
Platinum	tooth fillings
Polychlorinated biphenyl PCB	detergent, hair spray, salves
Polyvinyl chloride acetate (PVC)	glues, building supplies, leaking cooling system
Praseodymium	pollutant in pills
Radon	cracks in basement cement, water pipes
Rhenium	spray starch
Rhodium	tooth fillings
Rubidium	tooth fillings
Ruthenium	tooth fillings
Samarium	tooth fillings
Scandium	tooth fillings
Selenium	
Silver	tooth fillings
Sodium fluoride	toothpaste
Strontium	toothpaste, water softener
Tantalum	tooth fillings
Tellurium	tooth fillings
Terbium	pollutant in pills
Thallium acetate	pollutant in mercury tooth fillings
Thorium nitrate	earth (dust)
Thulium	pollutant in many brands of Vitamin C
Tin	toothpaste
Titanium	tooth fillings, body powder
Tungsten	electric water heater, toaster, hair curler
Uranium acetate	earth (dust)
Vanadium pentoxide	gas leak in home, candles (not necessarily lit)

Ytterbium	pollutant in pills
Yttrium	pollutant in pills
Zirconium	deodorant, toothpaste

Solvents

This is a list of all the solvents I use together with the main source of them in our environment. These are chemicals, very pure, obtained from chemical supply companies, unless otherwise stated. Those marked with an asterisk (*) were the subject of a recent book *The Neurotoxicity of Solvents* by Peter Arlien-Soburg, 1992, CRC Press.

Solvent	**Source**
1,1,1, Trichloro ethane* (TCE)	flavored foods
2,5-Hexane dione*	flavored foods
2 Butanone* (methyl ethyl ketone)	flavored foods
2 Hexanone* (methyl butyl ketone)	flavored foods
2 Methyl propanol	
2 Propanol (propyl alcohol)	see the propyl alcohol list
Acetone	store-bought drinking water, cold cereals, pet food, animal feed
Acetonylacetone (2,5 hexanedione)	flavored foods
Benzene	see the benzene list (page 354)
Butyl nitrite	
Carbon tetrachloride	store-bought drinking water, cold cereals, pet food, animal feeds
Decane	health food cookies and cereals
Denatured alcohol	obtained from pharmacy
Dichloromethane* (methylene chloride)	store-bought orange juice, herb tea blends
Gasoline regular leaded	obtained at gasoline station
Grain alcohol	95% ethyl alcohol obtained at liquor store
Hexanes*	decaffeinated beverages
Isophorone	flavored foods
Kerosene	obtained at gasoline station
Methanol (wood alcohol)	colas, artificial sweeteners, infant formula
Mineral oil	lotions

Mineral spirits	obtained from paint store
Paradichlorobenzene	mothballs
Pentane	decaffeinated beverages
Petroleum ether	in some gasolines
Styrene*	styrofoam dishes
Toluene*	store-bought drinking water, cold cereals
Trichloroethylene* (TCEthylene)	flavored foods
Xylene*	store-bought drinking water, cold cereals

Pathogen Frequency Chart

Use this chart if you know the frequency and wonder what the pathogen might be.

	KHz		5	0			1	0	0			1	5	0			2	0	0			2	5	0			3	0	0			3	5	0
Pathogen	Low	High																																
Cytochalasin B	77.00	77.00			m																													
Arcyria	81.00	81.00				s																												
Sterigmatocystin	88.00	88.00				m																												
Cytochalasin B (2nd)	91.00	91.00					m																											
Sterigmatocystin (2nd)	96.00	96.00					m																											
Zearalenone	100.00	100.00						m																										
Lycogala	126.00	126.00								s																								
Sterigmatocystin (4th)	126.00	126.00								m																								
Sterigmatocystin (3rd)	133.00	133.00									m																							
Aflatoxin	177.19	177.19													m																			
Stemonitis	211.00	211.00																s																
Sorghum syrup	277.00	277.00																						m										
Griseofulvin	288.00	288.00																							m									
Herpes simplex 1	291.25	293.05																								v								
Ergot	295.00	295.00																								m								
Histoplasma capsulatum	298.30	304.85																								b	b							
Corynebacterium xerosis	315.65	316.80																										b						
Shigella sonnei	318.00	318.00																										b						
Streptococcus mitis	313.80	321.10																										b	b					
Influenza A and B (flusho	313.35	323.90																										v	v					
Mycoplasma	322.85	323.90																											b					
Bacteroides fragilis	324.30	325.00																											b					
Proteus mirabilis	320.55	326.00																											b					
Bacteroides fragilis (2nd)	325.70	326.00																											b					
Salmonella enteriditis	329.00	329.00																											b					
Proteus vulgaris (3rd)	327.20	329.50																											b					
Bordetella pertussis	329.85	332.25																											b	b				
Pseudomonas aeruginos	331.25	334.60																												b				
Haemophilus influenzae	336.41	336.41																												b				
Neisseria gonorrhea	333.85	336.50																												b				
Proteus vulgaris (2nd)	333.75	339.15																												b				
Gardnerella vaginalis	338.00	342.55																												b	b			
Corynebacterium diphthe	340.00	344.00																													b			
Herpes simplex 1 (2nd)	345.35	345.75																													v			
Wart L arm	343.65	345.95																													w			
Treponema pallidum	346.85	347.40																													b			
Mycoplasma (2nd)	342.75	349.30																													b			
Lactobacillus acidophilus	346.05	351.65																													b	b		
Proteus mirabilis (2nd)	345.95	352.10																													b	b		
Erwinia amylovora	347.20	352.10																													b	b		
Serratia marcescens	349.45	352.10																													b	b		
Gaffkya tetragena	344.85	352.50																													b	b		
Schistosoma mansoni	353.00	353.00																														p		
Nocardia asteroides	354.95	355.35																														b		
Escherichia coli (E. coli)	356.00	356.00																														b		
Campylobacter pyloridis	352.00	357.20																														b		
Loa loa	360.55	360.55																															p	
Besnoitia (lung sect.) prot	352.80	361.40																														p	p	
Bacterial capsules	357.60	362.40																														b	b	
Herpes simplex 2 (fresh s	353.90	362.90																														v	v	
Coxsackie virus B-4	361.45	363.70																															v	

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

	KHz																			
			3	5	0			4	0	0			4	5	0			5	0	0
Pathogen	Low	High																		
Diplococcus diphtheriae	357.95	364.00	b	b																
Naegleria fowleri	356.90	364.35	p	p																
Clostridium botulinum	361.00	364.55		b																
Coxsackie virus B-4 (2nd)	363.90	364.90		v																
Bacillus anthracis (2nd)	363.20	365.30		b																
Clostridium septicum	362.05	365.60		b																
Coxsackie virus B-1	360.50	366.10		v																
Diplococcus pneumoniae	351.65	368.45	b	b																
Streptococcus sp. group	368.15	368.85		b																
Nocardia asteroides (2nd	363.70	370.00		b	b															
Salmonella paratyphi	365.05	370.10		b	b															
Streptococcus pneumoni	366.85	370.20		b	b															
Bacillus anthracis (3rd)	359.40	370.50	b	b	b															
Campylobacter fetus sme	365.30	370.60		b	b															
Measles antigen	369.50	373.00		v	v															
Plasmodium falciparum s	372.30	373.80			p															
Enterobacter aerogenes	374.00	374.00			b															
Streptococcus pyogenes	360.50	375.30		b	b															
Bacillus cereus	373.65	375.85			b															
Erwinia carotovora	368.10	377.00		b	b															
Central spores (bacillus s	372.45	378.65			b															
Histomonas meleagridis (l	376.55	378.70			b															
Staphylococcus aureus (c	376.27	380.85			b	b														
Staphylococcus aureus (s	381.00	381.00				b														
Gyrodactylus	378.75	381.80			p	p														
Borellia burgdorferi	378.95	382.00			v	v														
Spirillum serpens	378.35	382.80			b	b														
Epstein Barre virus (EBV)	372.50	382.85			v	v														
Respiratory syncytial viru	378.95	383.15			v	v														
Trichomonas vaginalis	378.00	383.60			p	p														
Chlamydia trachomatis	379.70	383.95			b	b														
Eikanella corrodens	379.50	384.30			b	b														
Mumps antigen	377.60	384.65			v	v														
Troglodytella abrassari	377.75	385.20			b	b														
Alpha streptococcus	369.75	385.40		b	b	b														
Sub terminal spores bac.	385.15	385.95				b														
Salmonella typhimurium	382.30	386.55				b														
Adenovirus (2nd range)	371.45	386.90			v	v														
Streptococcus lactis	382.00	387.00				b														
Bacillus subtilis var. niger	371.85	387.10			b	b														
Beta streptococcus	380.60	387.40				b														
Entamoeba histolytica tro	381.10	387.80				p														
Anaplasma marginale	386.40	388.00				p														
Candida albicans	384.20	388.40				y														
Propionobacterium acnes	383.75	389.00				b														
Shigella dysenteriae	390.09	390.09					b													
Chilomastix cysts (rat)	388.95	390.70				p	p													
Clostridium acetobutylicu	382.80	391.15				b	b													
Bacillus anthracis spores	386.95	391.45				b	b													
Escherichia coli (E. coli)	392.00	393.00					b													
Adenovirus	393.00	393.00					v													
Sphaerotilus natans	388.40	393.45				b	b													
Shigella flexneri	394.00	394.00					b													
Toxoplasma (human strai	395.00	395.00					p													
Haemonchus contortus	386.80	395.50				p	p													
Branhamella (Neisseria)	394.90	396.70					b													
Endolimax nana trophozo	394.25	397.10					p													
Bacillus anthracis	393.50	398.05					b													
Clostridium perfringens s	394.20	398.10					b													

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

	KHz																			
			3	5	0			4	0	0			4	5	0			5	0	0
Pathogen	Low	High																		
Trypanosoma gambiense	393.75	398.70					p													
Chilomonas, whole mount	393.75	400.00					p	p												
Entamoeba coli trophozoi	397.00	400.35					p	p												
Leptospira interrogans	397.05	401.10					b	b												
Strongyloides (filariform l	398.40	402.00					p	p												
Leucocytozoon	397.45	402.55					p	p												
Leishmania donovani	398.00	402.65					p	p												
Ancylostoma caninum	383.10	402.90				p	p	p												
Ancylostoma braziliense (	397.60	403.25					p	p												
Leishmania mexicana	400.20	403.80						p												
Wart papilloma cervix sm	404.05	404.60						w												
Klebsiella pneumoniae	398.45	404.65					b	b												
Prosthogonimus macrorc	396.85	404.75					p	p												
Iodamoeba butschlii (2nd)	398.15	404.75					p	p												
Leishmania braziliensis	400.05	405.10						p												
Veillonella dispar	401.75	405.20						b												
Trichinella spiralis (muscl	403.85	405.57						p												
Wart BS	402.00	406.00						w												
Dientamoeba fragilis	401.35	406.05						p												
Wart human papilloma pl	404.70	406.75						w												
Leishmania tropica	402.10	407.40						p												
Blepharisma	405.65	407.45						b												
Trichuris sp. (male)	388.30	408.90				p	p	p												
Ascaris larvae in lung	404.90	409.15						p												
Pneumocystis carnii (lung	405.75	409.15						p												
Ascaris megalocephala (	403.85	409.70						p												
Mycobacterium phlei	409.65	410.65						b	b											
Wart human papilloma vir	402.85	410.70						w	w											
Cytomegalovirus (CMV) a	408.35	410.75						v	v											
Dirofilaria immitis	408.15	411.15						p	p											
Stigeoclonium	404.25	415.25						p	p											
Cryptocotyle lingua (adult	409.95	416.00						p	p											
Proteus vulgaris	408.75	416.45						b	b											
Myxosoma	409.60	416.95						p	p											
Bacteria capsules (capsul	416.05	418.75							b											
Herpes Zoster	416.60	420.20							v	v										
Hepatitis B antigen	414.55	420.80							v	v										
Klebsiella pneumoniae (2	416.90	421.90							b	b										
Troglodytella abrassari (2	416.90	422.20							b	b										
Eurytrema pancreaticum	420.35	422.30								p										
Wart JB	418.75	422.40							w	w										
Echinoporyphium recurva	418.55	423.90							p	p										
Anaplasma marginale (2n	415.30	424.00							p	p										
Plasmodium cynomolgi	417.30	424.50							p	p										
Fasciola hepatica miracidi	421.75	424.70								p										
Trypanosoma lewisi (bloo	424.50	426.00								p										
Enterobius vermicularis	420.95	426.30								p										
Giardia lamblia (trophozoi	421.40	426.30								p										
Fasciola hepatica	421.35	427.30								p										
Chilomastix cysts (rat) (2	425.20	427.30								p										
Fasciola hepatica rediae	420.60	427.50								p										
Fasciola hepatica eggs	422.00	427.60								p										
Trypanosoma rhodesiens	423.50	428.55								p										
Clonorchis sinensis	425.70	428.75								p										
Hypodereum conoideum	424.45	429.55								p										
Tobacco mosaic virus	427.15	429.55								v										
Echinostoma revolutum	425.50	429.65								p										
Fasciola hepatica cercari	423.80	430.60								p	p									
Capillaria hepatica (liver s	424.25	430.65								p	p									

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

	KHz																			
			3	5	0			4	0	0			4	5	0			5	0	0
Pathogen	Low	High																		
Trypanosoma brucei	423.20	431.40									p	p								
Cytophaga rubra	428.10	432.20									b	b								
Wart CC	426.00	432.35									w	w								
Fasciolopsis rediae	427.30	433.00									p	p								
Endolimax nana (2nd)	430.50	433.35										p								
Mycobacterium tuberculo	430.55	434.20										b								
Fasciolopsis buskii	427.70	435.10									p	p								
Fasciolopsis miracidia	427.35	435.20									p	p								
Fasciolopsis buskii eggs	427.35	435.45									p	p								
Fasciolopsis cercariae	429.50	436.25									p	p								
Cysticercus fasciolaris	436.40	440.05										t	t							
Endamoeba gingivalis tro	433.80	441.00										p	p							
Onchocerca volvulus (tu	436.30	442.10										p	p							
Metagonimus Yokogawai	437.35	442.10										p	p							
Macracanthorhynchus	438.85	442.80										p	p							
Fischoedrius elongatus	441.75	443.20											p							
Wart HA	434.80	444.10										w	w							
Passalurus ambiguus	428.80	444.15									p	p	p							
Dipylidium caninum	439.55	444.30										t	t							
Plasmodium vivax smear	438.15	445.10										p	p							
Echinococcus granulosus	441.15	446.50											t							
Iodamoeba butschlii troph	437.85	448.50										p	p							
Wart HRCm	438.90	448.55										w	w							
Taenia solium (scolex)	444.00	448.90											t							
Urocleidus	442.35	450.00											p	p						
Trypanosoma equiperdu	434.60	451.25										p	p	p						
Paragonimus Westermani	437.80	454.20										p	p	p						
Sarcocystis	450.55	454.95												p						
Hasstile sig. tricolor (adult	448.05	455.10											p	p						
Gastrothylax elongatus	451.90	457.10												p						
Multiceps serialis	453.60	457.80												t						
Echinococcus multilocula	455.85	458.35												t						
Echinococcus granulosus	451.60	461.50												t	t					
Balantidium coli cysts	458.80	462.90												p	p					
Stephanurus dentalus (ov	457.35	463.10												p	p					
Wart FR	459.30	464.75												w	w					
Moniezia (scolex)	430.35	465.20										t	t	t	t					
Moniezia expansa (comp	430.35	465.20										t	t	t	t					
Trypanosoma cruzi (brain	460.20	465.65													p					
Taenia pisiformis eggs (o	465.20	469.70													p					
Dipylidium caninum (scol	451.95	472.15												t	t	t				
Diphyllobothrium latum (s	452.90	472.30												t	t	t				
Schistosoma haematobiu	473.00	473.00														p				
Taenia solium (cysticercu	475.00	475.00														t				
Taenia saginata (cysticer	476.50	481.05														t	t			
Hymenolepis diminuta	445.00	481.15											t	t	t	t	t			
Hymenolepis cysticercoid	478.00	481.75														t	t			
Taenia pisiformis (cystice	475.20	482.10														t	t			
Diphyllobothrium erinacei	467.25	487.55													t	t	t			

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

	KHz																																	
			5	5	0			6	0	0			6	5	0			7	0	0			7	5	0			8	0	0			8	5	0
Pathogen	Low	High																																	
Demodex folliculorum foll	682.00	682.00															x																		
Dermatophagoides dust	707.00	707.00																	x																
Meal mite	718.00	718.00																		x															
Sarcoptes scabei itch	735.00	735.00																				x													
Ornithonyssus bird mite	877.00	878.00																																	

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

Sources

This list was accurate as this book went to press. Only the vitamin sources listed were found to be pollution-free, and only the herb sources listed were found to be potent, although there may be other good sources that have not been tested. The author has no financial interest in, influence on, or other connection with any company listed, except for having family members in the Self Health Resource Center.

Note to readers outside the United States of America:

Sources listed are typically companies within the United States because they are the ones I am most familiar with. You may be tempted to try a more convenient manufacturer in your own country and hope for the best. I must advise against this! In my experience, an uninformed manufacturer most likely has a polluted product! Your health is worth the extra effort to obtain the products that make you well. One bad product can keep you from reaching that goal. This chapter will be updated as I become aware of acceptable sources outside the United States. Best of all is to learn to test products yourself.

Item	Source
Amber bottles, ½ ounce	Drug store, Continental Glass & Plastic, Inc. (large quantities)
Apricot Kernel Oil	Now Foods

Arginine	Now Foods, Jomar Labs
Artemesia (wormwood) seed	R. H. Shumway
Baking soda (sodium bicarbonate)	Spectrum Chemical Co.
Beet root	San Francisco Herb & Natural Food Co.
Belts for clothes dryer	Three that tested negative to asbestos are: Maytag^{TM} 3-12959 Poly-V belt, Whirlpool^{TM} FSP 341241 Belt-Drum Dr. (replaces 660996), and Bando^{TM} V-Belt A-65. Bando American makes other belts, some of which might be the right size for your dryer. Call for a dealer near you, make sure it says "Made In America", right on the belt.
Black cherry concentrate	Health food store
Black Walnut Hull Tincture	Self Health Resource Center, New Action Products
Borax, pure	Grocery store
Boric acid, pure	Now Foods, health food store, pharmacy
Cascara sagrada	Natures Way, health food store
Chemicals for testing.	Aldrich Chemical Co., Spectrum Chemical Co., ICN Biomedicals, Inc. (research chemicals only, including genistein), Boehringer Mannheim Biochemicals (research only)
Cholecalciferol	Spectrum Chemical Co.
Citric acid	Now Foods or health food store
Cloves	San Francisco Herb & Natural Food Co. (ASK for fresh)
Cornstarch	Lady Lee is the only unpolluted brand found so far
Dental information	Tooth Truth, by Frank J. Jerome, available from New Century Press.
EDTA	Spectrum Chemical Co.
Electronic parts	A Radio Shack near you.
Empty gelatin capsules size 00	Now Foods, or your health food store

Filters, pure charcoal	Pure Water Products (pitchers), Seagull Distribution Co. (faucet, shower, whole house)
Flaxseed oil	Omega Nutrition USA, Inc.
Fucus	San Francisco Herb & Natural Food Co.
Ginger capsules	Now Foods
Goldenrod tincture	Dragon River Herbals, Blessed Herbs
Grain alcohol	Liquor store, as vodka, Everclear^{TM}
Gravel root (herb)	San Francisco Herb & Natural Food Co.
Histidine	500 mg from Jomar Labs
Homeopathic remedies	Dolisos America, Inc., and others
Hydrangea (herb)	San Francisco Herb & Natural Food Co.
Hydrogen peroxide 35% (food grade)	New Horizons Trust
Iodine, pure	Spectrum Chemical Co.
Lugol's iodine	For slide staining (not internal use) from Spectrum Chemical Co., and farm animal supply store. For internal use must be made from scratch.
Lysine	Bronson Pharmaceuticals
Magnesium oxide	Bronson Pharmaceuticals
Marshmallow root (herb)	San Francisco Herb & Natural Food Co.
Microscope slides and equipment	Carolina Biological Supply Company, Ward's Natural Science, Inc., Southern Biological Supply Company, Fisher-EMD
Microscopes	Carolina Biological Supply Company, Ward's Natural Science, Inc., Edmund Scientific Co.
Milk Thistle	Herb companies.
Mullein tea	San Francisco Herb & Natural Food Co.
Niacin	100 mg or 250 mg time release, Bronson Pharmaceuticals
Ornithine	Now Foods, Jomar Labs
Ortho-phospho-tyrosine	Aldrich Chemical Co.
Ozonator	Superior Health Products, LLC
P24 antigen sample	Bachem Fine Chemicals Inc.
PABA (para amino benzoic acid)	Bronson Pharmaceuticals
Pantothenic acid	Bronson Pharmaceuticals

Peroxy	See Hydrogen peroxide.
Photo-micrographic camera and film	Ward's Natural Science, Inc.
Potassium bicarbonate	Spectrum Chemical Co., or pharmacy
Potassium chloride	Now Foods, Spectrum Chemical Co.
Potassium iodide, pure	Spectrum Chemical Co.
Rascal	Kroeger Herb Products, New Action Products (as Raz-Caps)
Salt (sodium chloride), pure	Spectrum Chemical Co., get USP grade
Sodium alginate	Spectrum Chemical Co. or health food store
Stevia powder	Now Foods
Tea	San Francisco Herbs, buy in bulk, any single herb variety.
Thallium, homeopathic	Dolisos America, Inc.
Thioctic acid (also called lipoic acid)	Maypro Industries
Uva Ursi	Natures Way or health food store
Vanilla extract	Durkee's^{TM} or Mexican brands
Vegetable glycerin	Now Foods or health food store
Vitamin B Complex	Bronson Pharmaceuticals
Vitamin B_2 (riboflavin)	Nutrition Headquarters (250 mg), or Bronson Pharmaceuticals (100 mg)
Vitamin B_6	Bronson Pharmaceuticals, EDOM Labs
Vitamin B_{12} lozenges	Bronson Pharmaceuticals
Vitamin C (ascorbic acid)	Self Health Resource Center
Vitamin D 50,000iu	From dentist, by prescription, or see Recipes.
Vitamin E capsules	Bronson Pharmaceuticals
Vitamin E Oil	Now Foods
Washing soda (sodium carbonate)	Grocery store
Water filter pitchers	Pure Water Products
Wormwood capsules	Self Health Resource Center, Kroeger Herb Products, New Action Products
Zinc	Bronson Pharmaceuticals
Zinc oxide	Spectrum Chemical Co.

Aldrich Chemical Co.
P.O. Box 355
Milwaukee, WI 53201
(414) 273-3850

Bachem Fine Chemicals Inc.
3132 Kashiwa St.
Torrance, CA 90505
(310) 539-4171

Bando American Inc.
1149 West Bryn Mawr
Itasca, IL 60143
(800) 829-6612
(630) 773-6600

Blessed Herbs
109 Barre Plaines Rd.
Oakham, MA 01068
(508) 882-3839

Boehringer Mannheim
Biochemicals
9115 Hague Rd.
P.O. Box 50414
Indianapolis, IN 46250
(800) 262-1640
(317) 849-9350

Bronson Pharmaceuticals
Div. of Jones Medical Industry
1945 Craig Road
P.O. Box 46903
St. Louis, MO 63146-6903
(800) 235-3200 retail
(800) 525-8466 wholesale

Carolina Biological Supply Co.
2700 York Rd.
Burlington, NC 27215
(800) 334-5551
(919) 584-0381

Continental Glass & Plastic, Inc.
841 West Cermak Rd.
Chicago, IL 60608-4582
(312) 666-2050

Dolisos America, Inc.
3014 Rigel Ave
Las Vegas, NV 89102
(800) 365-4767
(702) 871-7153

Dragon River Herbals
PO Box 74
Ojo Caliente, NM 87549
(800) 813-2118
(505) 583-2118

Edmund Scientific Co.
101 E. Gloucester Pike
Barrington, NJ 08007
(856) 573-6250

EDOM Labs, Inc.
860 Grand Boulevard
P.O. Box 780
Deer Park, NY 11729
(516) 586-2266
(800) 723-3366

Fisher Scientific EMD
485 S. Frontage Rd.
Burr Ridge, IL 60521
(800) 955-1177

ICN Pharmaceuticals, Inc.
Biomedical Division
3300 Hyland Ave.
Costa Mesa, CA 92626
(714) 545-0113
(800) 854-0530

Jomar Labs
251-B East Hacienda Avenue
Campbell, CA 95008
(800) 538-4545
(409) 374-5920

Kroeger Herb Products
(wholesale:)
805 Walnut St.
Boulder, CO 80302
(800) 225-8787
(303) 443-0261
(for retail call:)
Hanna's Herb Shop
5684 Valmont Rd.
Boulder, CO 80301
(800) 206-6722

Maypro Industries
2700 Westchester Ave.
Purchase, NY 10577
(914) 251-0701

Nature’s Way
10 Mountain Springs Pkwy
Springville, UT 84663
(800) 962-8873
(801) 489-1500

New Action Products (USA)
P.O. Box 540
Orchard Park, NY 14127
(800) 455-6459
(716) 662-8000

New Action Products (CANADA)
PO Box 141
Grimsby, Ont. L3M4G5 Canada
(800) 541-3799
(716) 873-3738

New Century Press
1055 Bay Blvd., Suite C
Chula Vista, CA 91911
(800) 519-2465
(619) 476-7400

New Horizons Trust
53166 St. Rt. 681
Reedsville, OH 45772
(800) 755-6360
(614) 378-6366

Now Foods
395 Glen Ellen Road
Glen Ellen, IL 60108
(630) 545-9098

Omega Nutrition USA, Inc.
6515 Aldrich Rd.
Bellingham, WA 98226
(800) 661-3529

Superior Health Products, LLC
13549 Ventura Blvd.
Sherman Oaks, CA 91423
(800) 700-1543
(818) 986-9456

Pure Water Products
10332 Parkview Ave.
Westminster, CA 92683
(800) 478-7987

R. H. Shumway
P.O. Box 1
Graniteville, SC 29829
(803) 663-9771

San Francisco Herb & Natural Food Co.
1010 46th St.
Emeryville, CA 94608
(800) 227-2830 wholesale
(510) 601-0700 retail

Seagull Distribution Co.
3372 Baltimore St.
San Diego, CA 92117
(858) 483-0264

Self Health Resource Center
1055 Bay Blvd., Suite A
Chula Vista, CA 91911
(800) 873-1663
(619) 409-9500 (USA)
(780) 475-2403 (Canada)

Southern Biological Supply Company
P.O. Box 368
McKenzie, TN 38201
(800) 748-8735
(901) 352-3337

Spectrum Chemical Co.
14422 South San Pedro Street
Gardena, CA 90248
(800) 772-8786
(310) 516-8000

Ward's Natural Science, Inc.
5100 West Henrietta Road
Rochester, NY 14586
(716) 359-2502
(800) 962-2660

Finale

I hope you have come to the same happy conclusion that I did a few years ago! That we humans don't have hundreds of different maladies and disturbances. **We only have two!** Things that crawl or climb into us. And toxins: unnatural chemicals that we unknowingly inhale or consume.

The living things are both large and small: from worms we can see, to microscopic bacteria, viruses and fungi. The non living things are pollutants in our air, food, dental metal and body products. Taking in a lot of pollutants hampers the body's ability to kill and get rid of the invaders. And so, gradually, as we get older or sicker, the body's invaders get the upper hand and take over. Don't be discouraged if you have lupus, cerebral palsy, cirrhosis, or any complex-sounding disease. Every disease is an example of the same process.

The good news is that our body can reclaim its sovereignty by throwing the rascals out. We must assist by throwing the pollutants out. Fortunately, we don't have to do the whole job ourselves, we only need to assist. Our body has miraculous powers to clean itself up.

By reducing chronic disease to two problems, it becomes manageable. Ill health, even aging, can be reversed. Health comes back and rewards you.

With the new electronic insights and technology, our parasitic invaders can be vanquished with the closing of a switch. Preventing reinfection is the bigger challenge. Similarly, pollutants can be uncovered in as short a time as days, short enough to turn any sick, even terminal, sufferer's verdict around before tragedy occurs.

The tragedies of surgery, organ replacements, radiation, chemotherapies, doses of drugs, even death can be avoided. Reversing illness and turning into a shining example of your former healthy self could be the most exciting adventure of your life.

Killing your invaders is an easy matter: you simply purchase or build the device that can do that and take the proper herbs. Cleaning up dentalware is under your control, too—a financial expense not beyond your reach, hopefully. Trading your body products for unpolluted varieties is a job but not insurmountable. Cleaning your environment may be the stumbling block. If you can't unpollute your air, water, carpets, furniture, move. Move to a healthier dwelling! Get rid of it all. Like a cat that moves her kittens to a safer place, just move.

The healthiest house is no house. If you have been quite ill, move far enough south to avoid heating and cooling. Sit outside in the shade all day. Use your new wisdom and sharp eye to choose a new dwelling as free of pollutants as you can. No refrigerator indoors. No window air conditioner. No fiberglass insulation. No fossil fuels. No attached garage. No carpets or stuffed furniture, no foam bedding, no fresh paint, no pesticide. Go primitive. Health is primitive. You were born primitive—with health. Even if you weren't, you can undo much of the "inherited" damage.

Don't listen to the new doomsayers, who persuade you nothing can be done unless a gene is replaced. Your genes have been reliable for millions of years. Genes are the most reliable of all biological chemicals in your body. They are not faulty. They are hampered in their tasks. They are commandeered by metals and other species' genes. Those of parasites, bacteria, and viruses. They wouldn't be there if pollutants weren't there. They allow invaders into the most jealously guarded recess of your being: your genes. But now you can throw the rascals out and reclaim your territory. It is not new genes that you need. You simply need your own genes back on the job, directed by your own body, working for you.

Index

A

abdominal pain, 99, 203, 268, 299
abscess, 54, 79, 135, 153, 158, 159, 163, 164, 216, 231, 510
Acanthocephala, 138, 274, 566
acetic acid, 296
acetone, 122, 162, 208, 320
acetylcholine, 221, 238
acidophilus, 21, 100, 564, 576
acne, 167, 192, 193, 194
acupuncture, 51, 465
Adenovirus, 13, 176, 179, 199, 357, 358, 359, 360, 361, 363, 368, 563, 577
adrenal gland, 60, 113, 114, 143, 210, 212, 217, 281, 291, 373
aflatoxin, 38, 177, 180, 186, 225, 261, 280, 348, 349, 381, 382, 383, 384, 386, 390, 392, 422, 491, 493, 494, 519
AIDS, 11, 35, 36, 90, 119, 126, 131, 149, 252, 351, 352, 353, 356, 368, 372, 385, 442, 443, 458, 493, 500
air conditioner, 40, 284, 313, 323, 346, 403, 446, 590
air filter, 137, 284
albuminuria, 417
alcoholism, 227, 228, 229, 230, 386
alkaline phosphatase, 80, 217
alkaline tide, 57
allergy, 85, 102, 105, 107, 111, 126, 135, 139, 140, 163, 169, 175, 190, 191, 216, 222, 223, 224, 225, 226, 242, 246, 258, 386, 495, 553, 558
aluminum, 37, 112, 137, 141, 171, 180, 181, 187, 189, 212, 244, 247, 269, 271, 274, 305, 310, 311, 314, 415, 421, 423, 424, 430, 431, 433, 434, 435, 436, 438, 439, 450, 461, 473, 484, 485, 486, 487, 514, 521, 525, 526, 533, 543
aluminum silicate, 37, 138, 141, 212, 271
alveolar cavitational osteopathosis, 410
Alzheimer's, 33, 219, 250, 252, 269, 270, 272, 372, 495, 526
amalgam, 37, 65, 71, 107, 115, 157, 180, 409, 410, 413, 414, 416, 417, 418, 475, 485
amalgam tattoo, 410
American Dental Association (ADA), 37, 409, 414, 416
amino acid, 85, 232, 244
ammonia, 125, 243, 244, 320, 326
amoeba, 106, 274
Ancylostoma, 256, 257, 262, 566, 578
anemia, 76, 162, 279, 285, 286
anesthesia, 90
angina, 176, 417
animal dander, 104, 135
animal feed, 125, 347, 574
ankle swelling, 68
ant repellent, 540
antibiotic, 134, 141, 170, 181, 248, 548
antimony, 105, 112, 123, 128, 140, 187, 435
antiseptic, 98, 344, 399, 436, 449
aorta, 478
appendicitis, 101, 268
appendix, 97, 101, 128
appetite, 134, 240, 268, 275, 288, 292, 300, 306, 376, 379, 388, 417
apple cider vinegar, 277, 348, 423, 493, 572
apricot kernel oil, 534, 535, 537
arginine, 120, 137, 210, 212, 425
arm, 49, 51, 118, 153, 154, 254, 255, 282, 321, 481, 482, 488, 531, 553, 576
armpit, 89, 143, 144, 193, 531, 532
arsenic, 4, 40, 68, 96, 100, 103, 110, 132, 137, 138, 141, 150, 160, 162, 168, 171, 185, 186, 187, 193, 206, 236, 242, 259, 268, 271, 282, 313, 314, 361, 403, 448
arthralgia, 417
arthritis, 51, 78, 80, 81, 82, 83, 84, 85, 86, 117, 137, 146, 151, 155, 214, 268, 419
artificial sweetener, 173, 176, 177, 252, 302, 520, 574

asbestos, 4, 39, 80, 81, 91, 100, 103, 108, 110, 133, 137, 138, 142, 143, 145, 148, 150, 167, 168, 170, 207, 209, 210, 214, 231, 271, 282, 314, 316, 336, 337, 349, 447, 448, 474, 582
Ascaris, 10, 32, 43, 76, 78, 79, 80, 82, 86, 103, 104, 105, 107, 111, 122, 127, 134, 135, 136, 137, 138, 139, 140, 141, 142, 162, 166, 167, 176, 179, 186, 187, 192, 193, 194, 196, 212, 213, 227, 230, 231, 233, 234, 235, 236, 242, 256, 257, 259, 260, 264, 266, 267, 268, 274, 285, 286, 293, 299, 343, 371, 372, 375, 404, 566, 578
ascorbic acid, 516, 534, 535, 537, 584
aspirin, 169, 170, 214, 264
asthma, 43, 44, 105, 107, 135, 136, 137, 138, 139, 140, 220, 230
athlete's foot, 247, 248
Atomic Absorption Standard, 484, 572
atrophy, 384, 418
Atypical Lateral Sclerosis, 47
autism, 263, 264, 266
auto exhaust, 222, 279, 282, 316

B

back pain, low, 68, 81, 96, 123, 131
back pain, upper, 81, 96, 151, 553, 558
Bacteroides fragilis, 135, 230, 231, 266, 372, 563, 576
baking powder, 269, 430, 485
baking soda, 57, 58, 269, 431, 438, 515, 516, 527, 532, 533
baldness, 126, 190
bandwidth, 7, 8, 10, 16, 18, 506, 508, 509, 561, 562
barium, 75, 123, 139, 143, 187, 207, 415, 435, 436, 474, 476
bath oil, 247, 439
beef, 147, 368, 423, 424, 477
beer, 194, 327, 348, 378, 380, 386, 493, 572
beet root, 539, 540
benzalkonium, 137, 138, 170, 180, 187, 244, 390, 435, 439
benzoic acid, 55, 84, 583
benzopyrene, 304, 431
beryllium, 78, 124, 138, 142, 145, 157, 166, 170, 187, 192, 224, 227, 228, 229, 235, 407
beta blocker, 289, 319
BHA, 232, 233, 265, 517, 529
BHT, 232, 233, 265, 517, 529
bile, 47, 48, 80, 88, 91, 98, 100, 129, 130, 131, 132, 150, 151, 152, 153, 191, 223, 224, 266, 267, 299, 300, 477, 553, 554, 555, 556, 557, 558, 559
bile duct, 47, 48, 98, 129, 130, 131, 150, 151, 152, 153, 191, 223, 224, 299, 300, 477, 553, 557, 558
bioradiation, 12, 13
birth control pill, 111, 116
birth defect, 111, 418
bismuth, 105, 166, 185, 208, 242, 271, 435, 474
black cherry concentrate, 549
Black Walnut Hull Tincture, 21, 101, 102, 228, 235, 338, 339, 341, 342, 344, 345, 347, 389, 543, 544, 545, 547, 582
bladder, 55, 60, 101, 108, 112, 124, 129, 130, 145, 166, 190, 196, 198, 293, 321, 550, 552, 555, 559
bleach, 40, 150, 246, 256, 258, 260, 284, 314, 397, 437, 444, 449, 526, 530, 540, 572
bleeding, 95, 96, 110, 112, 196, 249, 286, 536
bleeding gum, 95
bloating, 108, 113, 129, 254, 266, 299, 443, 546, 555
blood pressure, 19, 49, 75, 82, 95, 115, 148, 149, 210, 211, 212, 213, 273, 275, 280, 281, 282, 289, 311, 321, 377, 388, 417, 515
blood sugar, 173, 175, 176, 177, 178, 182, 187, 275, 305, 308
blood test, 5, 77, 80, 82, 106, 107, 114, 139, 145, 178, 179, 213, 235, 245, 319
blood vessel, 69, 70, 154, 163, 279, 280, 281, 282, 318, 321
body lotion, 170
bone cancer, 196
bone infection, 56, 79, 315, 410
bone marrow, 78, 181, 285, 351, 353, 477
bone spur, 86
Borellia, 170, 563, 577

boric acid, 313, 403, 449, 540, 541, 542
boron, 215
bowel, 20, 21, 32, 34, 35, 54, 77, 82, 83, 97, 99, 101, 102, 103, 104, 106, 108, 128, 129, 152, 163, 166, 169, 197, 202, 219, 223, 224, 238, 254, 256, 267, 270, 298, 301, 331, 338, 364, 388, 389, 425, 443, 494, 510, 530, 546, 547, 548, 558
bowel disease, 102, 389, 443
bowel movement, 21, 34, 35, 77, 82, 83, 98, 101, 106, 128, 129, 197, 202, 224, 238, 256, 267, 301, 331, 494, 530, 558
Bowel Program, 99, 101, 102, 127, 128, 164, 202, 237, 238, 254, 257, 262, 270, 298, 315, 328, 346, 364, 444, 546, 549
braces, 418, 476
breast, 142, 143, 144, 145, 146, 235, 332, 336, 337, 342, 376, 384, 420, 435, 476, 491
breast cancer, 143, 145, 332, 384, 420
breast milk, 235, 342, 376
breastbone, 107, 131, 149
bromine, 239, 258, 271
bronchiestasis, 138
bronchitis, 78, 140, 141, 193, 217
burning, 186, 253, 254, 369, 534
burping, 295, 550
bursitis, 51, 151, 152, 558, 559
butyl nitrite, 162

C

cachexia, 240
cadmium, 37, 47, 69, 70, 74, 75, 79, 96, 107, 112, 146, 175, 185, 192, 210, 211, 213, 267, 268, 269, 314, 415, 418, 426, 449, 450, 466, 484
caffeine, 67, 68, 75, 89, 117, 137, 145, 185, 210, 212, 213, 255, 259, 264, 276, 290, 317, 377, 425
calcium, 57, 63, 64, 65, 66, 67, 74, 80, 81, 83, 87, 88, 93, 94, 96, 106, 129, 158, 212, 215, 216, 217, 231, 255, 276, 277, 287, 299, 305, 361, 399, 421, 517, 521
calcium propionate, 361
calf cramp, 68
Campylobacter, 69, 76, 110, 114, 170, 298, 563, 576, 577
Candida, 20, 185, 186, 189, 229, 246, 247, 248, 492, 563, 577
candidiasis, 176, 246
candy, 181, 218, 219, 313, 354
carbon dioxide, 139, 242, 267, 279, 282
carbon filter, 258, 270, 283, 314, 405, 426
carbon monoxide, 267, 282
carbon tetrachloride, 96, 103, 104, 116, 122, 127, 137, 148, 162, 186, 260, 440
carbonic anhydrase, 296
carcinogen, 40, 44, 374, 403, 423
cardiomyopathy, 147
cardiovascular, 417
carpet cleaning fluid, 429
Cascara sagrada, 99, 164, 238, 301, 547, 582
cat liver fluke, 116, 268
cavitation, 117, 141, 156, 291, 410, 533
cayenne, 55, 59, 84, 180, 181, 320, 398
cercaria, 34, 111, 117, 122, 495
cerebellum, 111, 206, 208, 484
cerebral palsy, 589
cerebrum, 111, 206, 208, 271, 274, 484
cerium, 145, 162, 187
cervix, 112, 113
cesium, 104, 157, 259, 260, 449
cheese sauce, 306, 434, 524
chelation, 213, 271, 272, 288, 289, 292
chemotherapy, 437
chest, 51, 74, 78, 148, 149, 150, 151, 167, 177, 178, 192, 254
chest pain, 51, 148, 149, 151, 167, 177, 178, 254
chewing gum, 95, 218
chicken, 84, 99, 126, 139, 147, 176, 179, 233, 235, 265, 266, 268, 274, 347, 353, 371, 394, 423, 424, 431, 477, 492
chicken pox, 176, 179, 353, 371, 492
chiropractic, 86, 91, 92, 156
Chlamydia, 110, 114, 563, 577
chloride, 37, 70, 72, 139, 162, 208, 221, 269, 311, 373, 390, 428, 429, 440, 475, 476, 497, 516, 573, 574, 584

chlorine, 137, 150, 239, 246, 257, 258, 260, 271, 282, 283, 288, 314, 437, 449, 526, 530, 540
chlorofluorocarbons (CFCs), 143, 313, 446
choke, 7, 11, 486
cholesterol, 109, 122, 129, 130, 224, 374, 421, 432, 434, 440, 553, 554, 558, 559
chromium, 37, 176, 178, 184, 302, 375, 377, 418, 431, 435, 436
chronic fatigue syndrome, 111, 186
cigarette, 336
cinnamon tea, 121
circulation, 55, 59, 63, 156, 160, 182, 273, 320
cirrhosis, 382, 589
citric acid, 403, 515, 516, 521, 528, 530, 535, 542
Clostridium, 76, 206, 315, 563, 564, 577
clothes dryer belt, 39, 142, 232, 314, 447, 474, 572
clotting, 19, 259, 320
cobalt, 4, 96, 112, 128, 141, 147, 162, 189, 194, 314, 399, 474, 526
coffee, 61, 81, 82, 89, 174, 210, 268, 274, 275, 278, 308, 327, 335, 377, 381, 388, 425, 427, 428, 431, 536
cold cereal, 109, 111, 162, 178, 219, 252, 335, 376, 386, 387, 392, 420, 518, 574, 575
cold sore, 179, 180
colitis, 105, 106, 107, 167, 425
cologne, 144, 242
colon, 78, 97, 131, 164, 337, 357, 361, 389, 478, 491
comfrey, 137, 139, 552
composite, 71, 119, 126, 146, 177, 226, 254, 291, 416
condom, 480
congestive heart failure, 318
constipation, 104, 166, 178, 238, 299, 301, 548
cooked cereal, 109, 144, 178, 287, 291, 376, 386, 442, 556
cookie, 23, 269, 431
copper, 22, 26, 30, 37, 70, 96, 100, 113, 115, 116, 119, 128, 147, 157, 171, 208, 211, 212, 261, 262, 310, 314, 337, 349, 426, 450, 459, 462, 463, 466, 472, 474, 484, 517
corn, 66, 81, 106, 144, 215, 282, 302, 303, 306, 307, 308, 351, 355, 356, 361, 379, 385, 394, 420, 434, 460, 476, 492, 532, 537, 539
cornstarch, 181, 247, 249, 294, 354, 374, 531, 532, 534, 537, 540
corpus luteum, 115, 124
cortisone, 52, 59, 138, 246
cosmetics, 37, 143, 144, 208, 326, 335, 437, 476, 572
cottage cheese, 80, 201, 254, 432, 485, 510, 511
Coxsackie virus, 1, 2, 135, 137, 166, 233, 371, 564, 576, 577
cramp, 69, 111, 249, 253
crawl space vents, 127, 181
Crohn's disease, 33, 104, 106
crotch itch, 247
croup, 135, 140
crown, 411, 413, 414
Cryptocotyl, 74
cutlery, 126, 127, 379, 430, 433
cyst, 32, 33, 34, 39, 74, 90, 91, 92, 112, 145, 170, 171, 188, 189, 198, 200, 255
cysteine, 74, 94, 96, 108
cysticercus, 197, 198, 199, 200, 357, 358, 494
cystine, 69, 94
Cytochalasin B, 387, 562, 576
Cytomegalovirus (CMV), 147, 179, 184, 564, 578

D

decaffeinated, 109, 210, 269, 335, 428, 574, 575
decane, 122, 162, 188, 440
dementia, 278, 280, 322, 330
dental cleanup, 73, 78, 79, 87, 120, 127, 145, 155, 156, 220, 262, 346
dental floss, 70, 162, 481
dental metal, 68, 70, 78, 110, 112, 115, 116, 117, 119, 122, 134, 140, 141, 143, 145, 159, 162, 187, 206, 222, 223, 226, 233, 254, 261, 270, 337, 419, 485, 589
denture, 139, 295, 411, 533
denture cream, 295
deodorant, 75, 128, 138, 143, 144, 147, 162, 181, 189, 190, 193, 222, 231, 435, 531, 532, 572, 574

depilatory, 417
depression, 44, 78, 104, 113, 116, 117, 185, 186, 210, 228, 254, 256, 257, 259, 260, 261, 262, 298, 540, 565
dermatron, 6
dessert, 305, 306, 326
diabetes, 3, 35, 43, 44, 55, 173, 174, 175, 176, 177, 178, 250, 251, 278, 302, 305, 309, 388, 495
diabetic, 67, 89, 174, 175, 305, 308, 309
diaper, 189, 246, 256
diaphragm, 76, 132, 133, 192, 297, 375
diarrhea, 97, 103, 104, 107, 108, 117, 134, 152, 202, 300, 344, 421, 425, 558
diesel exhaust, 75
diet clean-up, 80
diet drink, 429
diet, 7 day diabetic, 302
diet, 7 day sample, 523
digestion, 34, 88, 104, 105, 129, 130, 131, 148, 177, 226, 255, 266, 267, 276, 277, 278, 287, 295, 296, 297, 299, 374, 377, 423, 552
digestive problem, 95, 130, 146, 268, 559
digestive tract, 103, 246, 266
Dirofilaria, 32, 80, 141, 146, 147, 148, 318, 567, 578
diuretic, 75, 213, 282, 294, 304, 321, 373
dizziness, 186, 275, 297, 298, 419
dropsy, 321
drug, 44, 82, 131, 136, 234, 238, 289, 290, 301, 312, 316, 319, 321, 322, 344, 393, 473, 533
dry skin, 525
dry socket, 410
duodenum, 34, 478
dysprosium, 254

E

E. coli, 21, 62, 97, 98, 99, 102, 116, 129, 164, 177, 266, 298, 353, 357, 360, 361, 363, 364, 388, 444, 547, 548, 564, 576, 577
ear, 6, 104, 105, 134, 140, 168, 169, 170, 217, 242, 325, 361, 419, 467
ear infection, 104, 105, 134, 140, 169
ear, inner, 6, 170, 361
earache, 169
earthworm, 10, 510
Echinacea, 99
eclampsia, 95
eczema, 140, 192, 230, 492
edema, 60, 322, 515
EDTA, 73, 271, 272, 289, 582
egg dish, 432
elbow, 51, 154, 467
electronic testing, 5
ELISA, 495, 496, 497
emphysema, 46, 139
enamel, 214, 310, 413, 425, 430, 501, 529, 533, 543
encephalitis, 1, 2, 356
Endolimax, 106, 138, 567, 577, 579
endometriosis, 33, 35, 103, 110, 111, 122, 250, 495
endometrium, 250
enema, 166
environmental illness, 43
eosinophil, 136, 139, 185, 245
eosinophilia, 417
Epsom salts, 99, 150, 151, 300, 555, 556, 557
Epstein Barre Virus (EBV), 112, 179, 182, 184, 185, 186, 204, 371, 564, 577
ergot, 162, 226, 227, 228, 229, 232, 233, 253, 263, 270, 277, 280, 383, 385, 386, 392, 520
esophageal sphincter, 297
esophagus, 131, 132, 149
estrogen, 91, 113, 116, 119, 122, 143, 144, 145, 185, 217, 384, 539
ether, 575
europium, 124, 140, 187, 207
Eurytrema, 21, 109, 173, 371, 567, 578
Eustachian tube, 168, 361
exercise, 2, 86, 92, 301, 327
eye, 3, 14, 76, 140, 160, 161, 162, 175, 186, 259, 478, 497, 498, 528, 539, 590
eye disease, 3, 161, 175
eye makeup, 140, 186
eyelid, 146, 539

F

face, 36, 123, 181, 192, 193, 194, 209, 220, 227, 231, 253, 258, 280, 283, 369, 400, 448, 469, 540

face powder, 231, 540
fallopian tube, 250
Fasciola, 43, 249, 567, 578
Fasciolopsis buskii, 33, 150, 166, 250, 331, 371, 490, 567, 579
fatigue, 2, 4, 73, 95, 111, 113, 134, 139, 176, 177, 182, 183, 184, 185, 186, 187, 204, 245, 289, 311, 368, 417
fennel, 444, 547, 548
fenugreek, 175, 176, 370, 422, 521
fetus, 258, 417
fever, 78, 201, 203, 204, 234, 365, 369
fiberglass, 39, 40, 80, 91, 100, 110, 133, 137, 143, 147, 150, 160, 168, 177, 241, 282, 284, 314, 336, 349, 403, 405, 447, 590
fibromyalgia, 51, 75, 76, 78
fibromyositis, 75
filaria, 31, 318
finger, 86, 136, 155, 195, 253, 282, 308, 321, 463, 486, 487, 508, 530, 539
fingernail, 501
fireplace lighter, 165
fish line, 95, 218, 248, 296, 438, 533
flatworm, 33, 331
flavoring, 304, 378, 428, 548, 574, 575
flaxseed, 127
floss, 70, 72, 95, 162, 218, 296, 438, 481, 532
flour, 298, 306, 310, 361, 362, 379, 385, 423, 474, 572
flu, 78, 134, 168, 176, 201, 202, 203, 275, 425, 564, 576
fluke disease, 117, 219, 221, 249, 250, 252, 494, 495
fluke egg, 145, 332
fluoride, 82, 112, 206, 296, 573
foam, 40, 194, 284, 314, 337, 404, 447, 449, 533, 573, 590
foam bedding, 314, 337, 590
foam filter, 284
foam furniture, 40, 449
folic acid, 120
Food and Drug Administration (FDA), 436
foot, 51, 57, 59, 60, 68, 247, 248, 328, 419, 506
formaldehyde, 40, 100, 108, 110, 112, 133, 137, 150, 160, 165, 168, 193, 236, 241, 282, 284, 314, 361, 448, 449, 499, 500, 571
fragrance, 105, 136, 166, 185, 247, 253, 259, 284, 529, 530, 550, 572
Freon, 100, 160, 336, 446, 469
Fucus, 239, 548, 583

G

gadolinium, 124, 140, 207
Gaffkya, 170, 564, 576
gallbladder, 80, 128, 132, 191, 477, 553, 559
gallium, 123, 124, 146
gallstone, 81, 150
galvanized pipe, 70, 210, 211, 450
ganglia, 180
Gardnerella, 69, 91, 102, 110, 114, 115, 119, 564, 576
garlic, 137, 376, 548
gas hot water heater, 431
gas leak, 105, 140, 141, 149, 165, 166, 170, 171, 175, 187, 241, 255, 260, 268, 279, 319, 450, 573
gas pipe, 96, 165
gas stove burner, 476
gas, household, 231
gasoline, 124, 127, 162, 208, 221, 224, 227, 316, 416, 445, 472, 572, 574
gene, 3, 46, 206, 207, 209, 590
genital tract, 116
germanium, 417
Giardia, 9, 103, 160, 239, 399, 567, 578
ginger, 205, 514, 515
ginseng, 259, 263
glaucoma, 146, 213, 214
glucose tolerance factor, 184, 419
glutamic acid, 228
glutamine, 228, 229
gluten, 105, 175
glycerin, 515, 516, 534, 535, 536, 537, 538, 539, 540, 549, 584
goat milk, 107
gold, 102, 103, 111, 113, 124, 126, 146, 175, 177, 184, 226, 237, 310, 412, 418, 419, 420, 475
goldenrod, 549
gonorrhea, 91, 116, 565, 576
gout, 69
granola, 186, 354, 421, 518
gravel root, 552
grilled food, 356
groin, 90, 157, 166, 248

growing pains, 69
gum disease, 81
gums, 81, 87, 95, 217

H

HA wart virus, 176
Haemophilus, 141, 242, 564, 576
hafnium, 139, 222, 573
hair dryer, 241, 282, 316, 346, 447, 448
hair spray, 122, 136, 139, 144, 187, 222, 335, 439, 573
hamburger, 219, 349
hand cream, 354
handhold, 26, 27, 30, 457, 462, 463, 464, 487, 488, 503, 504, 506, 508
hCG, 336
head, 31, 32, 52, 112, 131, 144, 156, 161, 169, 171, 184, 198, 199, 254, 283, 315, 359, 361, 369, 417, 461, 557, 569
headache, 51, 75, 105, 116, 162, 164, 165, 167, 170, 206, 257, 369, 400, 419, 425
hearing, 4, 6, 7, 242, 261, 274, 278, 324, 325, 387, 468, 504
heart, 4, 32, 59, 80, 114, 138, 139, 141, 146, 147, 148, 149, 159, 168, 176, 177, 178, 182, 194, 195, 213, 254, 278, 279, 280, 290, 291, 311, 317, 318, 319, 320, 321, 322, 323, 325, 403, 419, 478, 515, 526
heart disease, 147, 176, 194, 403, 515, 526
heart failure, 278, 318, 322
heartburn, 82, 375
heartworm (see also Dirofilaria), 32, 138, 140, 141, 146, 147, 148, 149, 167, 178, 254, 318, 319, 343, 567
heel, 61, 63, 65, 67, 68
hemoglobin, 279, 286, 287
herb tea, 63, 66, 93, 95, 140, 321, 335, 365, 390, 424, 428, 429, 432, 524, 551, 574
herbicide, 40
hernia, 90, 133, 154, 268, 297
Herpes, 9, 20, 115, 116, 161, 179, 180, 181, 182, 371, 487, 488, 492, 564, 576, 578
hexane, 440
hexanedione, 111, 148, 193, 269, 574
hiatal hernia, 133, 154, 268, 297
high blood pressure, 49, 82, 95, 149, 211, 212, 213, 311, 388, 417, 515
hip, 51, 87, 89, 275, 419
hippuric acid, 55, 84, 85, 115
histamine, 135
histidine, 127, 207
Histoplasma, 145, 166, 564, 576
HIV, 10, 33, 90, 149, 150, 189, 204, 250, 251, 252, 342, 351, 352, 353, 356, 371, 458, 487, 490, 491, 495, 500, 501, 564
hives, 68, 89, 122, 191, 553
hookworm, 78, 79, 86, 105, 139, 160, 167, 194, 260, 264, 286, 343
hormone, 2, 91, 107, 113, 115, 116, 119, 122, 216, 368, 384
hot flush, 114, 115, 117, 175, 255
human liver fluke, 33, 74, 96, 109, 111, 116, 122, 128, 186, 188, 207, 208, 224, 249, 259, 260, 498
hurricane lamp, 78, 145, 166, 227, 572
hydrazine sulfate, 240
hydrochloric acid, 296, 297
hydrogen peroxide, 95, 139, 218, 296, 383, 438, 532, 533, 542
hydrogen sulfide, 267
hyperactive, 203
hyperesthesia, 417
hyperthyroidism, 188
hypnotism, 51

I

ice cream recipe, 522, 523
ileocaecal valve, 97, 101, 128
immune failure, 286
impetigo, 46
incontinence, 75, 278, 294, 325
indigestion, 6, 103, 145, 298, 446, 557
indium, 123, 157, 162
inductance, 5, 7, 486
inductor, 11, 486, 489
infant formula, 429
infertility, 2, 117, 144, 186, 237, 384
inhaler, 139
insect, 8, 506, 509, 529
insecticide, 312
insomnia, 4, 82, 134, 244, 245, 278, 326, 340, 417, 557
insulin, 173, 174, 175, 176, 177, 178, 182, 276, 302, 308, 332, 490
interferon, 468

interleukins, 468
intestinal problem, 128, 267
iodine, 95, 136, 163, 164, 188, 202, 218, 239, 246, 248, 298, 344, 348, 398, 399, 442, 443, 499, 559, 583
iridium, 82, 123, 188, 255, 271, 497
iron, 247, 248, 277, 279, 286, 450, 572
irritability, 113, 210, 254, 262, 298
irritable bowel syndrome (IBS), 102, 103, 104
islets of Langerhans, 182
isopropyl alcohol, 251, 285, 332, 335, 336, 348, 394, 398

J

jam, 2, 38, 52, 233, 265
jaw, 56, 74, 79, 87, 88, 156, 157, 158, 159, 170, 206, 215, 226, 254, 291, 315, 377
jaw bone, 74, 79, 87, 88, 206, 215
joint, 65, 66, 70, 82, 84, 139, 155, 157, 217, 417
joint pain, 84, 417

K

kerosene, 137, 208, 224, 227, 235, 407, 429, 572
kidney stone, 48, 56, 75, 79, 94, 95, 96, 108, 127, 166, 204, 212, 291, 552
knee, 44, 68, 74, 419, 489
knuckle, 195, 459, 463, 464, 486
Kojic acid, 102, 174, 175, 226, 388, 423

L

lactic acid, 88, 184, 187, 276
Lactobacillus, 100, 564, 576
lactose, 66, 104, 202
lamp oil, 474
lanolin, 223, 225
lanthanide, 414
lard, 303, 310, 354, 374, 394, 424, 517, 529, 530
laser, 111
lateral sclerosis, 204
lead, 4, 22, 26, 37, 46, 47, 60, 69, 74, 75, 79, 81, 96, 101, 105, 116, 136, 139, 141, 146, 157, 170, 181, 197, 220, 231, 232, 234, 262, 264, 265, 267, 268, 275, 280, 290, 314, 349, 356, 416, 417, 418, 426, 436, 440, 466, 472, 475, 504
Leishmania, 107, 193, 568, 578
lemonade, 557
lesion, 9, 181
leukemia, 211
lip, 534, 535
lip soother, 534, 535
lipoic acid, also see thioctic acid, 442, 584
lipstick, 75, 123, 207, 327, 336, 476, 539, 572
lithium, 255, 259, 260, 271
liver disorder, 374
liver fluke, 33, 43, 74, 78, 96, 103, 104, 107, 109, 111, 116, 117, 122, 128, 131, 175, 182, 185, 186, 187, 188, 207, 208, 218, 220, 222, 223, 224, 225, 249, 259, 260, 268, 271, 494, 498
Loa loa, 141, 147, 148, 149, 176, 291, 318, 319, 568, 576
loss of balance, 169, 270, 419
lotion, 247, 485, 535, 536
Lou Gehrig's disease, 44
Lugol’s, 163, 202, 203, 246, 248, 271, 298, 315, 346, 398, 442, 443, 499, 500, 501, 559, 583
lung, 40, 135, 138, 336, 372, 407, 477, 478, 576, 578
lung fluke, 138
lupus, 250, 495, 589
lutetium, 141
lye, 529, 530
Lyme disease, 170, 563
lymph, 64, 89, 90, 109, 144, 156, 157, 267
lymphocytosis, 417
lysergic acid, 263
lysine, 142, 146, 180, 181, 534, 535, 537, 538

M

magnet, 155
magnetic resonance imagery (MRI), 5
maintenance parasite program, 83, 105, 106, 116, 151, 176, 271, 340, 341, 555, 558

malignant, 39, 331, 337, 447
malvin, 233, 235, 236, 265
mammary gland, 143
mammogram, 144, 145, 146
manganese, 89, 176, 277, 572
Maple Milk Shake, 514
maple syrup, 191, 243, 365, 381, 390, 441, 513, 514, 524
margarine, 378, 517
marshmallow root, 549, 551, 552, 583
massage, 51, 70, 85, 92, 435, 439
massage oil, 435, 439
mean cell volume (MCV), 212, 285
measles, 186, 353, 492
medication, 166, 203, 236
melanoma, 337
memory loss, 273, 278
Meniere's syndrome, 169
menopause, 113, 114, 143, 260
menstruation, 113
mental problem, 2, 283
mercuric chloride, 70, 72
mercury amalgam, 409, 416, 417, 485
mesothelioma, 337
metabolism, 111, 187, 238, 239, 240, 243, 244
metacercaria, 495
metal tooth filling, 73, 123, 142, 157, 188, 189, 442, 472, 533
metastasis, 337
methanol (wood alcohol), 173, 174, 429
methyl butyl ketone, 35, 96, 110, 111, 116, 122, 269, 574
methyl ethyl ketone, 35, 110, 171, 260, 574
methyl methacrylate, 414
methylene chloride, 37, 162, 208, 221, 269, 574
microscope slide, 8, 149, 210, 477, 501, 561
migraine, 162, 163, 164, 165, 168, 259
milk digestant, 80, 421, 425
millet, 388, 420
miracidia, 34, 188, 495, 500
mite, 358, 359, 360, 408, 570, 580
mitral valve, 147, 159
moisturizer, 525
molasses, 286
mononucleosis, 492
monosodium glutamate (MSG), 232, 233, 253, 255, 265, 266, 304, 434
moth ball, 104, 141, 149, 162, 271, 448
mother's milk, 133, 143, 375, 490
mouthwash, 156, 194, 218, 229, 335, 413, 435, 438, 476, 533
multiple sclerosis, 46, 204, 219, 250, 254
mumps, 176, 178
muscle, 66, 78, 93, 94, 139, 147, 158, 161, 187, 197, 218, 220, 477
muscle pain, 66
muscular dystrophy, 219, 220
myasthenia gravis, 55, 221
mycoplasma, 138, 141, 176
mycotoxin, 38, 102, 261, 351, 384, 385, 388, 423
myofibrositis, 77

N

Naegleria, 138, 271, 568, 577
nail polish, 136, 222, 254, 449, 573
nausea, 66, 121, 153, 264, 338, 557
neck, 89, 144, 156, 157, 160, 166, 206, 280, 402, 419
Neisseria, 114, 116, 119, 563, 565, 576, 577
nephritis, 417
nerve, 36, 59, 86, 154, 158, 170, 180, 181, 205, 206, 262, 329, 371, 478
nervousness, 113, 187
neurologist, 73
niacin, 68, 70, 73, 168, 279, 280, 281, 320, 380
niacinamide, 228, 233, 263, 338, 364, 380
nickel, 37, 85, 106, 115, 122, 124, 125, 126, 138, 139, 143, 145, 177, 180, 188, 190, 207, 222, 374, 418, 420, 424, 431, 436, 475
nicotine, 264, 539
night sweats, 204, 353
niobium, 146
nitrate, 573
Nocardia, 112, 159, 187, 206, 242, 260, 565, 576, 577
nose, 22, 24, 40, 105, 107, 133, 140, 193, 199, 203, 347, 359, 361, 369, 379, 536
numbness, 154, 157, 207, 209, 254, 255, 256, 417
nursing, 134, 242, 292, 342, 375
nut, 23, 25, 262, 310, 374, 386, 389, 461, 519
nut butter, 310, 374, 386, 389

O

obesity, 237, 240, 242, 548
optic nerve, 478
optic neuritis, 417
ornithine, 82, 243, 244, 259, 272, 325, 326, 340, 342, 557
osteoarthritis, 78, 83
osteomyelitis, 410
ovarian cyst, 91, 110, 255, 384
ovary, 90, 91, 92, 119
overweight, 130, 237, 238, 239, 242, 243
oxalate, 68, 69, 93, 94, 95, 96, 108, 116, 122, 128, 139, 142, 171, 188, 551
oxalic acid, 61, 93, 94, 243, 276, 551

P

P24 antigen, 583
PABA, 536, 583
palladium, 82, 138, 141, 149, 177, 187, 208
pancreatic fluke, 33, 35, 43, 102, 103, 105, 107, 109, 140, 142, 173, 174, 175, 176, 177, 178, 179, 182, 183, 186, 188, 220, 221, 249, 250, 251, 259, 302, 371, 375, 388, 494, 498
pantothenic acid, 60, 253
para amino benzoic acid (PABA), 536, 583
paradichlorobenzene, 149, 162
Paragonimus, 107, 138, 568, 579
paralysis, 417
paraplegia, 417
paresthesia, 417
Parkinson's disease, 159, 315
parsley, 304, 306, 343, 344, 345, 513, 549, 550, 551
pasta, 66, 96, 121, 204, 225, 228, 262, 306, 307, 308, 310, 356, 363, 374, 375, 381, 385, 386, 387, 432
patulin, 277, 522
PCB, 147, 162, 188, 314, 426, 475, 527, 573
peanut butter, 45, 189, 364, 376, 382, 389, 432, 519, 523, 524
penis, 250, 491, 493
pentane, 127, 162, 171, 188, 208
perfume, 190, 253
peripheral neuropathy, 209
personal lubricant, 439
pesticide, 40, 68, 96, 132, 162, 171, 185, 236, 253, 255, 268, 282, 304, 313, 346, 427, 472, 473, 542, 572, 590
pet food, 354, 574
pet parasite program, 77, 117, 122, 167, 395, 494
Peyer's patches, 267
phenylisothiocyanate (PIT), 84
phosphate, 63, 64, 65, 66, 67, 68, 73, 80, 81, 83, 86, 88, 94, 95, 96, 106, 107, 139, 142, 177, 185, 212, 213, 215, 216, 217, 255, 320, 551
phosphate crystal, 65, 67, 86, 94, 95, 96, 107, 177, 185, 212, 217, 255, 320, 551
pinworm, 220, 404
piperine, 84
placenta, 76, 258, 263
platelet, 19, 145, 286
platinum, 68, 116, 122, 123, 134, 141, 149, 185, 193, 207, 212, 419
PMS, 113, 115, 255, 384
Pneumocystis, 138, 372, 568, 578
pneumonia, 138, 142, 147, 169, 407, 564, 566
polio, 371
polonium, 372
polyethylene, 62, 517, 532
polyneuritis, 417
popcorn, 144, 181, 248, 313, 351, 355, 356, 373, 385, 476, 492
porcelain, 415
pork, 206, 423, 424, 477, 484
potassium, 58, 59, 82, 95, 139, 163, 211, 217, 218, 277, 291, 293, 294, 304, 311, 348, 373, 499, 515, 516, 559
potassium bicarbonate, 58, 515
potassium chloride, 211, 311, 373, 516
potpourri, 165, 168, 448
praseodymium, 122, 221, 244, 439
pregnancy, 76, 106, 111, 115, 119, 121, 122, 264
probe, 28, 459, 462, 463, 464, 465, 467, 468, 471, 487, 489, 504, 506
progesterone, 91, 113, 114, 115, 119
propane, 207
prostate, 69, 96, 112, 123, 124, 125, 126, 127, 128, 139, 190, 331, 332, 336, 337, 491, 500
prostate cancer, 332, 337, 500

Prosthogonimus, 139, 255, 568, 578
protein, 193, 228, 232, 553
Proteus, 94, 102, 110, 114, 565, 576, 578
Pseudomonas, 74, 565, 576
psoriasis, 193
purpura, 281
PVC, 70, 78, 211, 231, 236, 241, 262, 282, 450, 475, 573
pyloric, 297

Q

quassia, 546

R

rabbit fluke, 105, 203
radiation, 7, 437, 589
radio frequency, 6, 7, 13, 486
radon, 96, 108, 127, 142, 145, 146, 150, 181, 185, 186, 187, 405, 406, 475
rash, 36, 189, 190, 191, 220, 246
rat poison, 417, 418
red blood cell, 19, 77, 285, 286
Red Milk, 515
redia, 9, 34, 122, 495, 500
reflexology, 61
reproductive organs, 119
respiratory, 140, 142, 160, 170, 358, 360, 361, 363, 449, 517
retina, 160, 162, 213
retinitis pigmentosa, 3, 162
rhabdomyosarcoma, 337
rhenium, 139, 187
riboflavin, 380, 584
rice, 63, 66, 121, 144, 303, 306, 307, 308, 311, 349, 351, 356, 363, 366, 383, 385, 420, 432, 492, 534
ringworm, 247
room freshener, 150, 284
root canal, 79, 119, 138, 145, 158, 159, 163, 166, 187, 208, 216, 266, 318, 410, 412
roundworm, 14, 32, 76, 83, 86, 107, 139, 257, 265, 267, 274, 343, 372, 566
rubella, 189
ruthenium, 212, 244, 439

S

salicylate, 170
salmon, 307, 366, 376, 422, 524
salve, 162
samarium, 82, 157, 208
Sarcocystis, 107, 568, 579
sassafras, 529, 546
scalp, 126, 171, 190, 193, 526, 528
scandium, 206
schizophrenia, 211, 261, 262, 450
sciatic nerve, 86
scopolamine, 393, 423
seafood, 205, 307, 366, 424, 432, 434, 522
sebaceous cyst, 171, 188
seizure, 230, 231, 232, 233, 234, 235, 236, 265, 517
semen, 342
sex, 67, 121, 179, 205, 319, 342
shampoo, 105, 144, 284, 333, 335, 336, 437, 476, 528, 542
shaving supplies, 189, 284, 335
sheep liver fluke, 33, 43, 78, 103, 104, 107, 109, 111, 117, 122, 131, 175, 182, 185, 187, 188, 208, 223, 224, 225, 249, 260, 271, 494, 498
shingles, 179, 192, 371, 487, 492, 564
shoulder, 49, 150, 151, 152, 153, 255, 553
silicon, 475
silver, 37, 100, 103, 110, 116, 119, 146, 177, 185, 192, 193, 260, 310, 420, 426, 475, 476
sinus, 104, 116, 117, 138, 139, 142, 163, 167, 168, 170, 187, 245, 268, 369, 419, 492
skin cancer, 194
skullcap, 244, 552
sleep problem, 96, 107, 117, 241, 255
Smoker's Leg, 70
smoking, 67, 68, 70, 73, 177, 255, 327, 336, 372
soap, homemade, 437, 438, 444, 448, 449, 530
soda pop, recipe, 515
sodium, 58, 59, 106, 139, 146, 181, 212, 217, 232, 282, 311, 373, 440, 515, 529, 534, 535, 537, 582, 584
sodium alginate, 106, 181, 534, 535, 537
sodium bicarbonate, 58, 582
sodium carbonate, 584

sodium chloride, 373, 440, 584
sonogram, 5
sore throat, 20, 159, 160, 206, 217, 400, 492
sorghum, 226, 281, 388, 391, 475, 481, 519
soup, 106, 180, 209, 243, 287, 294, 306, 307, 310, 324, 349, 376, 405, 431, 432, 434, 521, 523, 524
sour cream, 164, 293, 305, 524
soy, 387, 432, 434
spasm, 70, 93, 150, 153, 163, 300
spastic colon, 78
speech problem, 270
sperm, 119, 123
spinal cord, 204, 205, 206, 361, 478
spine, 49, 80, 92, 93, 217
spleen, 148, 174, 195, 197, 198, 358, 419, 492
stainless steel, 431, 543
Staphylococcus, 62, 74, 76, 89, 99, 116, 147, 159, 163, 178, 187, 206, 216, 242, 291, 318, 319, 346, 353, 565, 566, 577
starch, 121, 300, 373, 374, 375, 378, 531, 532, 534, 537, 539, 540, 573
Sterigmatocystin, 385, 562, 576
sternum, 148, 149
stomach problem, 104, 105, 268
strep throat, 100, 104, 492
Streptococcus, 74, 76, 99, 117, 147, 168, 169, 170, 187, 242, 566, 576, 577
stroke, 278, 280, 281, 282
Strongyloides, 78, 79, 82, 86, 163, 164, 166, 167, 170, 186, 194, 256, 257, 258, 259, 260, 264, 343, 371, 568, 578
strontium, 124, 127, 181, 192, 244, 296, 435, 436, 439, 450
styrene, 103, 122, 148, 162, 430
styrofoam, 122, 148, 260, 292, 314, 430, 433, 475, 527, 575
sulfur, 95, 267, 376
sweat, 126, 247, 528, 531, 532
sweetening, 303, 308, 524
Syncrometer, 43, 45, 150, 182, 189, 251, 312, 346, 355, 429, 444, 446, 457, 458, 459, 460, 462, 466, 467, 470, 471, 482, 503, 504, 506, 571
syphilis, 91, 566

T

Taenia pisiformis, 75, 148, 196, 570, 579
Taenia solium, 147, 196, 570, 579
tantalum, 140, 141, 475
tapeworm, 20, 32, 50, 75, 111, 147, 161, 179, 195, 196, 197, 198, 199, 200, 201, 208, 220, 318, 357, 358, 359, 360, 363, 364, 370, 371, 408, 487, 494, 570
tapeworm stage, 20, 32, 75, 111, 147, 161, 179, 195, 196, 199, 200, 201, 208, 318, 357, 358, 359, 360, 363, 364, 370, 371, 487, 494
TC Ethylene, 96, 141, 148, 188, 193, 429
tea kettle, 310, 431
tellurium, 68, 78, 134, 142, 188, 208
temporary ischemic attacks (TIA), 207, 281
tendon, 154
teratogen, 418
terbium, 186
testicle, 20, 54, 119, 128, 178
thallium, 37, 69, 70, 71, 72, 73, 74, 79, 95, 96, 100, 111, 115, 123, 134, 189, 190, 206, 209, 220, 242, 254, 255, 264, 269, 314, 361, 417, 418, 438, 536, 573
thioctic acid, 73, 74, 82, 120, 123, 178, 207, 208, 229, 236, 237, 254, 262, 265, 269, 270, 442, 584
threadworm, 163, 371
throat, 20, 100, 104, 105, 141, 149, 150, 160, 199, 206, 227, 241, 242, 243, 268, 280, 295, 354, 361, 370, 400, 477, 491, 492
thrush, 246, 248, 492
thulium, 128, 141, 146, 177, 193, 221, 439, 440
thymus, 35, 36, 107, 131, 150, 207, 208, 221, 250, 251, 351, 352, 353, 446, 477, 491, 493
thyroid, 77, 107, 117, 170, 182, 183, 185, 187, 188, 193, 213, 217, 237, 239, 254, 266, 377, 443
thyroxin, 187
tin, 9, 139, 162, 187, 242, 296, 435, 436, 472, 473, 476
Tinea, 247
tinnitus, 6, 169, 170, 171, 419

titanium, 89, 103, 123, 143, 180, 193, 231, 254, 435, 436, 476
TMJ, 157, 158, 259, 419
toast, 178, 181, 352, 356, 430, 431, 434, 524, 572
toaster, 573
tobacco, 84, 279, 326, 372, 436, 492, 566
Tobacco Mosaic virus, 436
toe, 56, 59, 67, 68, 155, 248, 253
tofu, 432, 434
toluene, 35, 37, 103, 141, 148, 205, 209, 219, 220, 232, 251, 252, 269, 270, 271, 336, 427, 476
tomato juice, 82, 308, 521
tongue, 6, 248, 376, 485
tonsil, 268
tooth metal, 112, 123, 133, 138, 142, 149, 162, 186, 187, 206, 221, 231, 240
toothbrush, 95, 218, 248, 295, 438, 532
toothpaste, 82, 124, 127, 138, 156, 181, 187, 196, 206, 218, 242, 296, 420, 435, 438, 476, 532, 572, 573, 574
Toxoplasma, 160, 161, 162, 568, 577
trachea, 187, 478
tremor, 49, 78, 206, 207, 209, 260, 270, 278, 313, 314, 315, 316
Treponema, 114, 119, 566, 576
Trichinella, 31, 68, 76, 79, 80, 82, 86, 153, 161, 186, 192, 193, 194, 207, 257, 260, 343, 568, 578
trichloroethane, 37
Trichomonas, 102, 110, 568, 577
Trichuris, 74, 107, 142, 166, 568, 578
tryptophane, 244, 245, 439
tuberculosis, 142, 407, 565, 579
tumor, 40, 337, 415, 547
tungsten, 271, 430, 431, 449
turkey, 423, 424, 477, 485
turmeric, 547
tyramine, 164, 293, 305

U

udder wash, 98, 109, 137
ulcer, 132
underweight, 240, 268
universal allergy syndrome, 495
uranium, 142, 146
urate, 95, 122, 139, 166, 167, 255
urea, 75, 125, 243, 320
urethra, 124
uric acid, 55, 61, 68, 69, 81, 94, 171, 213, 320
urinalysis, 68, 77, 166, 185, 193
urinary tract, 77, 102, 106, 110, 114, 122, 125, 127, 163, 166, 167, 190, 255, 291, 430, 555
uterine wall, 112, 249
uterus, 35, 103, 106, 108, 109, 110, 111, 112, 113, 115, 117, 119, 122, 123, 124, 145

V

vagina, 246, 250, 491, 493
valerian, 150, 151, 152, 153, 158, 244, 272, 300, 325, 326
vanadium, 96, 105, 140, 149, 165, 170, 171, 175, 186, 187, 231, 255, 260, 268, 271, 313
vanilla, 514, 519, 522
Vanilla Milk, 514
vasectomy, 128
vegetable juice, 66, 82, 94, 237, 308, 309, 322, 422, 424, 513, 523, 524, 525
vegetarian, 197, 434
vinegar, 89, 276, 277, 288, 291, 296, 297, 303, 305, 306, 348, 349, 366, 377, 381, 383, 403, 405, 423, 441, 444, 448, 493, 514, 521, 527, 528, 540, 541, 542, 572
vitamin A, 95, 142, 154
vitamin B, 80, 88, 142, 146, 178, 217, 235, 285, 353, 551
vitamin B1, 235, 285
vitamin B12, 235, 285
vitamin B2, 88, 353
vitamin B6, 80, 142, 146, 217, 551
vitamin C, 38, 58, 73, 87, 89, 107, 110, 120, 121, 128, 141, 144, 146, 168, 177, 178, 181, 184, 193, 212, 218, 222, 228, 230, 242, 247, 263, 270, 277, 280, 281, 286, 291, 301, 302, 303, 305, 347, 348, 349, 356, 364, 365, 366, 370, 382, 383, 385, 386, 388, 389, 390, 391, 393, 421, 423, 427, 432, 433, 440, 441, 470, 481, 494, 513, 515, 516, 518, 519, 520, 521, 523, 525, 531, 534, 535, 537, 538, 543, 545, 548, 556, 573, 584

vitamin D, 64, 65, 66, 74, 81, 87, 88, 158, 217, 218, 254
Vitamin D Drops, 560
vitamin E, 120, 145, 146, 534, 539
vomiting, 105, 134, 166, 202, 394, 417

W

wallpaper, 141, 448, 572
wart, 69, 174, 176, 195, 196, 201, 478
washing soda, 57, 58, 180, 188, 269, 314, 346, 399, 431, 438, 444, 448, 515, 516, 525, 527, 532, 533, 540
water heater, 39, 319, 431, 447, 449, 573
water softener, 117, 127, 274, 437, 450, 473, 517, 572, 573
weakness, 92, 127, 139, 207, 220, 278, 293, 317, 322, 417
whip worm, 105, 107, 142
white distilled vinegar, 277, 349, 423, 444, 448, 521, 540, 541, 542
white iodine, 95, 163, 218, 248
wine, 164, 280, 327, 381, 386, 431
wood alcohol (see also methanol), 37, 102, 103, 162, 173, 174, 175, 176, 177, 178, 179, 183, 185, 186, 206, 208, 240, 251, 252, 259, 269, 302, 309, 314, 371, 375, 377, 388, 391, 414, 436, 475, 519, 559, 574
worm, 10, 105, 107, 142, 147, 163, 197, 247, 257, 318
wormwood, 339, 342, 344, 345, 582
wrist, 74, 75, 80, 154, 155, 255, 328

X

xenobiotic, 419
xylene, 35, 37, 111, 122, 141, 148, 205, 208, 209, 219, 220, 232, 251, 252, 270, 427, 429

Y

Yankee Drink, 514
yeast, 4, 115, 116, 122, 189, 246, 247, 248, 298, 377, 399, 492, 563
ytterbium, 415, 436
yttrium, 123, 127, 186

Z

zearalenone, 144, 225, 355, 492
zinc, 70, 106, 120, 123, 127, 154, 160, 181, 247, 277, 296, 532, 534, 572
zinc oxide, 181, 247, 532, 534
zirconium, 75, 89, 112, 128, 138, 147, 170, 180, 189, 193, 222, 231, 415, 435

A New Diagnostic Device…

With an audio oscillator circuit and a trained ear, you can detect parasites and pollutants in a way that no one has done before–electronically.

Leads To New Discoveries…

In every case of the "mysterious" disease diabetes, you find the not-so-mysterious parasite *Eurytrema*, and the fairly common pollutant *wood alcohol*. Every case! And never in healthy people! Similarly in cancer, HIV, Alzheimer's, endometriosis, to name a few, you can find specific parasites and/or pollutants at work.

And New Cures…

You don't need dangerous, expensive prescription drugs to get rid of the causes of your illness. Once you know what you are fighting you can pick herbal, electronic, or avoidance methods. Except for the cost of dental work, you can get well for under $100!

And New Hope…

Follow the advice in this book preventively, and never worry about your health again! Get your whole family well and keep them well. Start today!

It Took An Independent Research Scientist

to find the answers. Hulda Regehr Clark began her studies in biology at the University of Saskatchewan, Canada, where she was awarded the Bachelor of Arts, Magna Cum Laude, and the Master of Arts, with High Honors. After two years of study at McGill University, she attended the University of Minnesota, studying biophysics and cell physiology. She received her Doctorate degree in physiology in 1958. In 1979 she left government funded research and began private consulting on a full time basis. Six years later she discovered an electronic technique for scanning the human body. Today she puts her methods, results and conclusions before you. Read and recover.

New Century Press
Chula Vista, CA

1 (800) 51902465
www.newcenturypress.com

ISBN 1-890035-01-7
$21.95

About this electronic book

While creating this electronic version of The Cure for all Diseases, I encountered some inconsistencies and (small) errors in it. Below follows an overview.

Niels Willems, december 2002.
niels.w@hccnet.nl
www.rupa.demon.nl

Page	Line	Hard copy	Electronic version
18	Fig.2	(Di-phyllobothruim e.)	(Di-phyllobothrium e.)
25	last-2	both with a alligator clip	both with an alligator clip
255	last-3	liver cleanses(page 549) make	liver cleanses (page 552) make
360	last-7	health improv-ed	health improved
451	5	furniture"treated"	furniture "treated"
452	last	refigerators	refrigerators
467	14	SALT # 1	SALT #1
519	last	honey	honey.
562	last-7	Argyria	Arcyria
563	16	Bac. antracis spores | 391.45 | 386.95	Bac. antracis spores | 386.95 | 391.45
564	6	Corynebacterium diptheriae	Corynebacterium diphtheriae
565	18	Norcardia asteroides (2nd range)	Nocardia asteroides (2nd range)
570	last-4	folicle mite	follicle mite
575	2	Paradiclorobenzene	Paradichlorobenzene
576	(17)	missing: Aflatoxin (2nd) 188 188	(not corrected)
576	last-19	Corynebacterium diptheri	Corynebacterium diphthe
580	4	folliculorum foli	folliculorum foll

HOW TO CONSTRUCT A BOB BECK BLOOD ELECTRIFIER CIRCUT WITH A BREADBOARD

FOR BEGINNERS

PLEASE VISIT

https://archive.org/details/TakeBackYourPower-BobBecksuppressedMedicalDiscoverieS

TO VIEW THE ENTIRE BOB BECK PROTOCOL HANDOUT

AND WATCH

http://archive.org/details/DrBobBeckVideo-LectureOnCuringCancerAndHiv

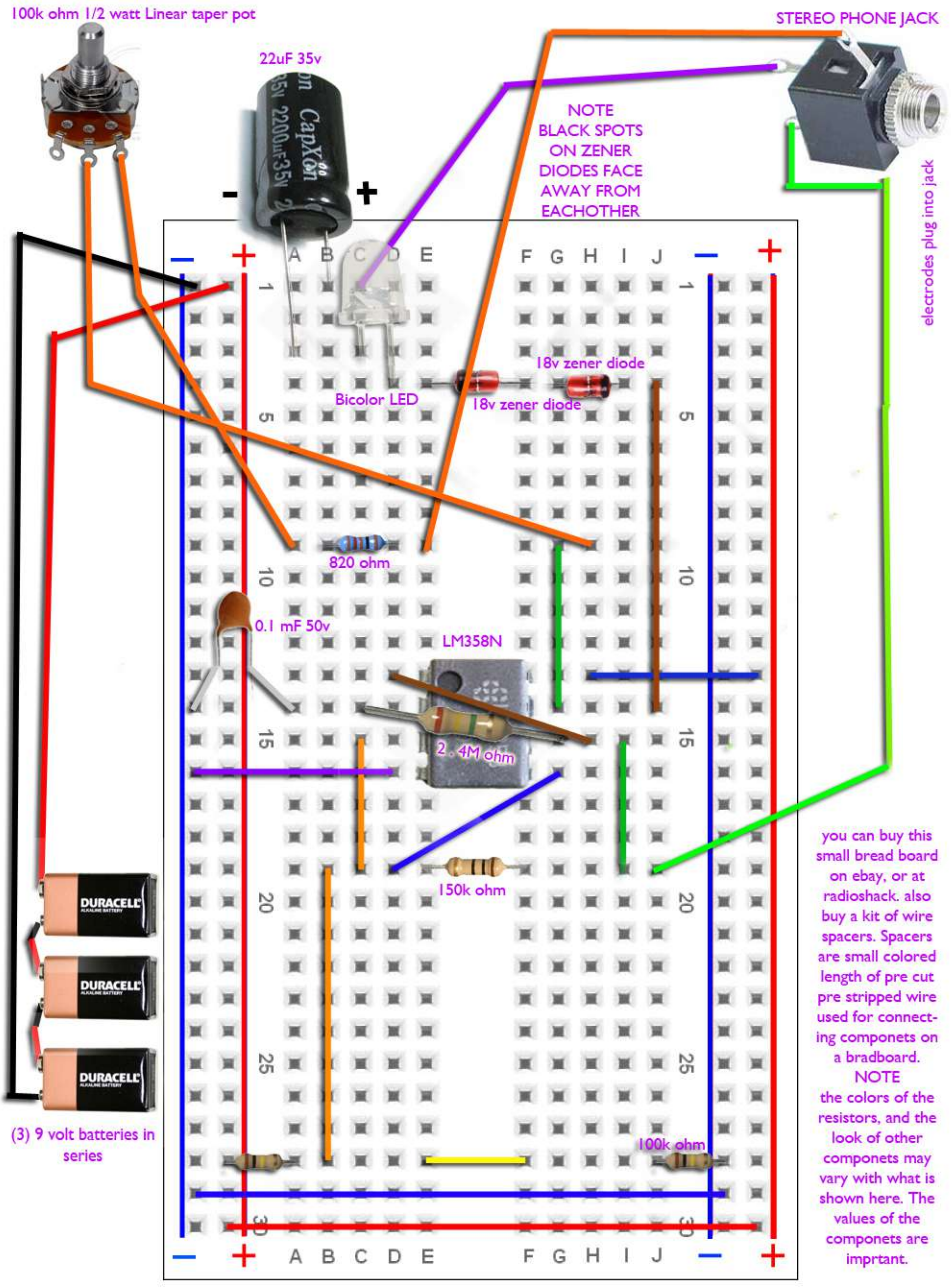
100k ohm 1/2 watt Linear taper pot
22uF 35v
-
+
STEREO PHONE JACK
electrodes plug into jack
Bicolor LED
18v zener diode
18v zener diode
820 ohm
0.1 mF 50v
LM358N
2 . 4M ohm
150k ohm
100k ohm
(3) 9 volt batteries in series
A B C D E F G H I J
1 5 10 15 20 25 30
DURACELL

NOTE
BLACK SPOTS
ON ZENER
DIODES FACE
AWAY FROM
EACHOTHER

you can buy this small bread board on ebay, or at radioshack. also buy a kit of wire spacers. Spacers are small colored length of pre cut pre stripped wire used for connecting componets on a bradboard.

NOTE
the colors of the resistors, and the look of other componets may vary with what is shown here. The values of the componets are imprtant.

COMPONENT DESCRIPTIONS, SOURCES, AND CURRENT PRICES

Revision November 24, 1996. Copyright (c) 1991/1999 Robert C. Beck, D.Sc.

Note: These data are for theoretical, informational and instructional purposes only and are not to be construed as medical advice. Consult with your licensed medical practitioner. Hundreds have been built successfully if duplicated exactly as shown.

LM358 CMOS Operational Amplifier (OP-AMP) chip (generic)		U1	$ 1.00
2.4 megohm 1/4 watt 5% resistor		R1	0.07
150 kohm 1/4 watt 5% resistor		R2	0.07
100 k ohm 1/4w 5% resistor	2 required @ 0.15	R3, R4	0.14
100 kohm linear taper pot, 1/2 watt Cattronics P-68		R5	2.56
150 ohm 1/4 watt 5% resistor		R6	0.07
820 ohm 1/4 watt 5% resistor		R7	0.07
0.1 mF 50 V (or higher) ceramic capacitor		C1	0.45
22 mF 35 V (or higher) electrolytic capacitor		C2	0.25
18 Volt Zener diodes (NTE5027A)	½ Watt, 2 @ 0.79	D1 & D2	1.58
Bulb, 6.3V 0.075 A	Type 7377	B1	1.34
Bicolor LED Red/Green	RS# 276-012	LED 1	$ 1.19
Jack for electrode leads (3.5mm or 1/8" mono-phone jack)	RS# 274-251 3/$1.59	J1	0.53
Jack for silver wire leads (2.5mm or 3/32" mono-phone jack)	RS# 274-292 3/$1.59	J2	0.53
3 Alkaline 9 V batteries, type 1604 etc.	3 for		3.75
3 x 9 V battery snaps (clip-on connectors)	RS# 270-325 pkg.	5/$1.29	1.29
Misc. wire, solder, etc.			0.50
Box, if used			2.29
Lead wire with 3.5 mm plug, 6 ft., Mouser or Calrad Electronics			0.35
Electrodes, stretch elastic, Velcro, cotton flannel, alligator clips, etc.			
Total Cost **for all** components for improved **do-it-yourself** project			**$22.47**

All components listed above check your local electronics store.

Ensure the IC chip U 1 (LM358) is wired correctly. The location of Pin #1 is shown below:

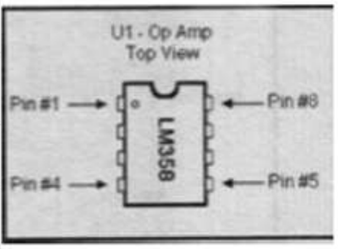

NOTE
RESISTOR
COLORS NOT
TO SPEC.
CHECK
RESISTOR
COLOR VALUES

electrodes:
3/32 x 1
stainless steel cotter pins -hardware store-
Snap pins in half across the curve. and wrap the headphone wires comming from the mono plug around the curve of each half of the pin. then hou glue into place.

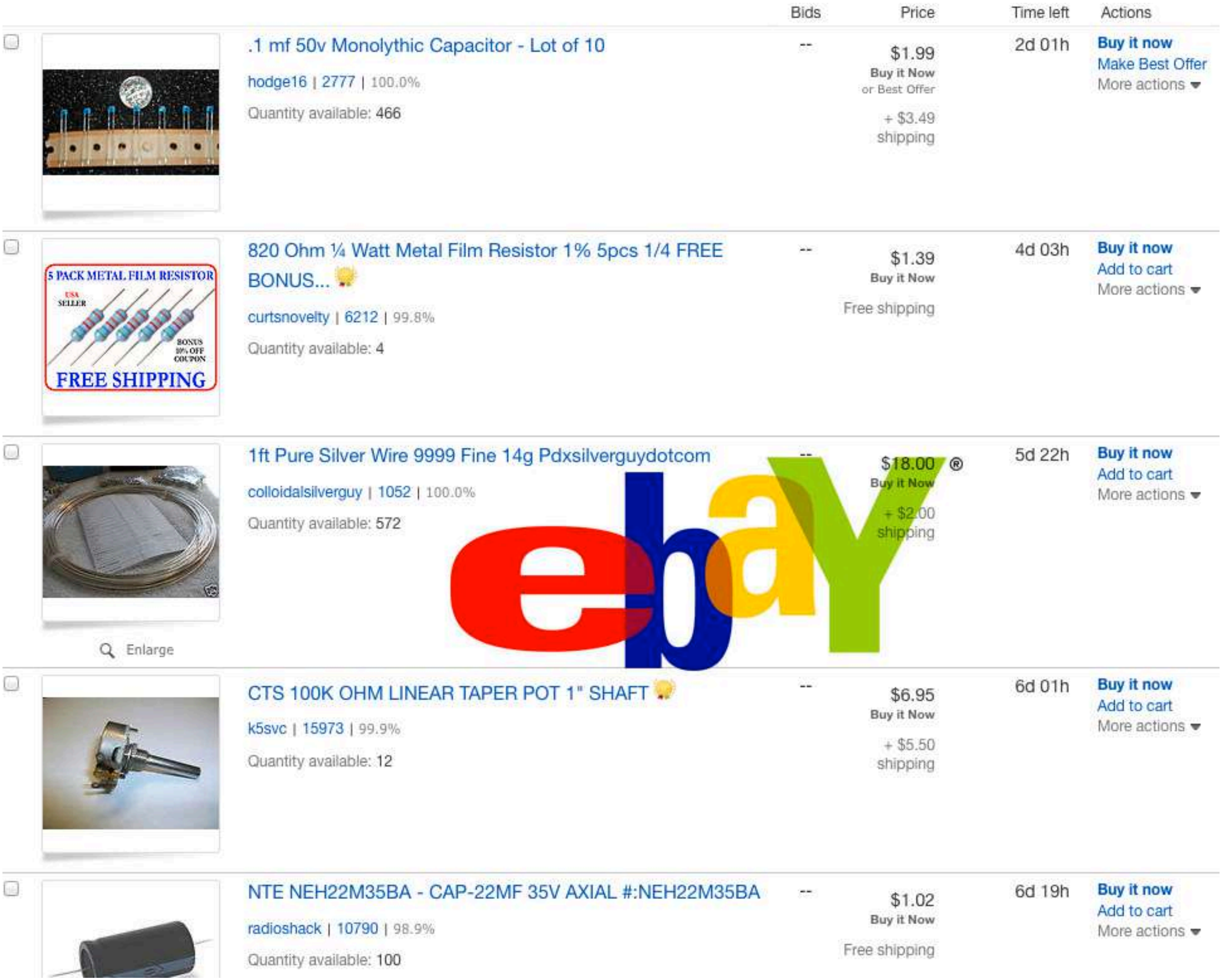

		Bids	Price	Time left	Actions
☐	.1 mf 50v Monolythic Capacitor - Lot of 10 hodge16 | 2777 | 100.0% Quantity available: 466	--	$1.99 Buy it Now or Best Offer + $3.49 shipping	2d 01h	Buy it now Make Best Offer More actions ▾
☐	820 Ohm ¼ Watt Metal Film Resistor 1% 5pcs 1/4 FREE BONUS... curtsnovelty | 6212 | 99.8% Quantity available: 4	--	$1.39 Buy it Now Free shipping	4d 03h	Buy it now Add to cart More actions ▾
☐	1ft Pure Silver Wire 9999 Fine 14g Pdxsilverguydotcom colloidalsilverguy | 1052 | 100.0% Quantity available: 572 Enlarge	--	$18.00 Buy it Now + $2.00 shipping	5d 22h	Buy it now Add to cart More actions ▾
☐	CTS 100K OHM LINEAR TAPER POT 1" SHAFT k5svc | 15973 | 99.9% Quantity available: 12	--	$6.95 Buy it Now + $5.50 shipping	6d 01h	Buy it now Add to cart More actions ▾
☐	NTE NEH22M35BA - CAP-22MF 35V AXIAL #:NEH22M35BA radioshack | 10790 | 98.9% Quantity available: 100	--	$1.02 Buy it Now Free shipping	6d 19h	Buy it now Add to cart More actions ▾

ebay®

Blood Electrification

and

Ionic/Colloidal Silver

Science & Vie Magazine *September 1998*

"White cells (leukocytes) kill bacteria and pathogenic fungi by electrocuting them. Discovered by Jacques Schrenzel and Karl Heinz Krause, Geneva University Hospital, Switzerland, this defense mechanism is triggered as soon as receptors of the white cell detect the microbe's presence. This signal activates an enzymatic system on the leukocyte's membrane, composed of a "sleeping" enzyme (NADP II oxidize) and co-enzyme (NADP II), the reaction begins by the enzyme "waking up", receiving electrons from its co-enzyme, and transporting them through the membrane. Ejected out of the white cell, this electronic flux is focused toward the bacteria with oxygen molecules and the microbe dies. This discovery might open new lines of research on immunity deficiencies in patients who are deprived of this enzymatic system."

Translated from Science & Vie, Issue #972, September, 1998 page 44.
Telephone: 01-46-48-48-48 and Fax: 01-46-48-48-67

Looking at this totally as a "physics" or "engineering" problem, a few structured electrons can act as the equivalent of trillions of white cells and restore function to one's immune system!
The electrons cost pennies.

US Patents

Amazingly similar discoveries have been made by scientists and are a permanent record in the US Patent Office since the 1880's. Why don't we know of these? Many work far more effectively than expensive antibiotics and have no side effects and practically no costs.

Since the Abraham Flexner propaganda funded by pharmaceutical company owner John D. Rockefeller in 1910 eliminated electromedicine from accredited medical universities, researchers have not pursued breakthrough discoveries. They are usually discredited. Instead the public spends billions of dollars annually with "medicines" that may create more side-effects than the original symptoms.

Here is a condensed reference list of US Patents which professional researchers may find interesting:

1 Direct application of electricity to kill or attenuate microorganisms in liquid.

1) Kaali	5,188,738	
2) Kaali	5,139,684	
3) Lathrop	5,133,352	
4) Ellis	4,019,510	
5) Myers	3,753,886	
6) Jones	592,735	(1897)
7) Dawson	3,965,008	
8) La Comme	672,231	(1901)

II Applications of Magnetic Fields to Kill or Attenuate Micro–Organisms.

1) Hofmann	4,524,079
2) Middleton	5,645,697
3) Forrest	5,248,437

III Application of Magnetic Fields in the Treatment of Cancer.

1) Costa	4,665,898
2) Liboff	5,437,600
3) Gordon	4,106,488
4) Gordon	4,303,636
5) Gordon	4,622,952

IV Application of Electricity and Magnetic Fields to produce other Biological Effects.

1) Cadossi	4,683,873	(Fuse Bone)
2) Ryaby	4,266,532	(Modify Cells)
3) Findl	4,428,366	(Reduce Serum Glucose)
4) Chang	4,822,470	(Cell Electroporation)

SPECIAL SECTIONS

LONGEVITY

DECEMBER 1992

LONGEVITY/DECEMBER 1992, Page 14

OUTER LIMITS

ELECTROCUTING THE AIDS VIRUS A SAFER-YET BLOOD SUPPLY

Despite official reassurances about the safety of the nation's blood supply, concern lingers that small amounts of HIV—infected blood may be sneaking through, especially since current screening detects only antibodies to the virus, which can take months to form. But now a new electrical process for cleaning blood of viruses may solve the problem.

At the Albert Einstein College of Medicine in New York City, Steven Kaali, M.D., has found that most of the AIDS viruses in a blood sample will lose their infectious capability after being zapped by a very low—level current. Repeated exposure appears to leave blood virtually free of HIV, as well as hepatitis—without harming blood cells.

Kaali cautions that it will take years of testing before a virus—electrocuting device is ready for use. But, ultimately, he predicts, it could be used not just to purify blood, but to treat people with AIDS, by channeling their blood out of the body, exposing it to virus—killing current and then returning it. "By lowering the viral count," he says, "we may be able to lengthen and improve the quality of AIDS patients' lives."

—Sharon McAuliffe

MARCH 30, 1991 Page 207

Biomedicine

Shocking Treatment proposed for AIDS

Zapping the AIDS virus with low–voltage electric current can nearly eliminate its ability to infect human white blood cells cultured in the laboratory, reports a research team at the Albert–Einstein College of Medicine in New York City.

William D. Lyman and his colleagues found that exposure to 50 to 100 microamperes of electricity—comparable to that produced by a cardiac pacemaker—reduced the infectivity of the AIDS virus (HIV) by 50 to 95 percent. Their experiments, described March 14 in Washington, D.C., at the First Inter–national Symposium on Combination Therapies, showed that the shocked viruses lost the ability to make an enzyme crucial to their reproduction, and could no longer cause the white cells to clump together—two key signs of virus infection.

The finding could lead to tests of implantable electrical devices or dialysis–like blood treatments in HIV–infected patients, Lyman says. In addition, he suggests that blood banks might use electricity to zap HIV, and vaccine developers might use electrically incapacitated viruses as the basis for an AIDS vaccine. For scientists working to create contraceptive devices that repel sperm with electricity, the new study also hints at a lifesaving side effect: protection against HIV.

Reprinted with permission from SCIENCE NEWS,
the weekly newsmagazine of science,
copyright 1991 Science Service

A-10 /The Houston Post/Wednesday, March 20, 1991 YOUR HEALTH/MEDICINE

Scientists say electric current may help fight AIDS

REUTER NEWS SERVICE

NEW YORK — Doctors at a prestigious New York medical center are testing a new way to fight AIDS— using electrical energy to weaken the killer virus— and say their first results are encouraging.

Researchers William Lyman and Steven Kaali of the Albert Einstein School of Medicine said Tuesday that initial laboratory tests have shown electrical current can weaken the virus believed to cause acquired immune deficiency syndrome.

The two men said they plan to move to the next phase of the experiment in April using blood samples from people with AIDS.

If their tests are successful, the researchers hope it could lead to a new way to treat AIDS patients, possibly involving a dialysis—type machine in which an AIDS patient's blood would be treated with electrical current outside the body.

"What we have done is expose the AIDS virus in laboratory circumstances to electrical current and then incubated the virus with white blood cells susceptible to the virus. We found that the virus became much more ineffective," Kaali, a specialist in the medical use of electrical current, said Lyman, an AIDS researcher and associate professor of pathology at Einstein, likened the new technique to chemotherapy.

"You are not going to get rid of the tumor, but you could get rid of enough of it to help the patient lead a normal life. This is not a cure but a new tool," Lyman said.

He added that the use of electrical energy has no toxic side effect and that a similar technique has been used as a treatment for reducing herpes.

US005188738A

United States Patent [19]

Kaali et al.

[11] Patent Number: 5,188,738

[45] Date of Patent: * Feb. 23, 1993

[54] ALTERNATING CURRENT SUPPLIED ELECTRICALLY CONDUCTIVE METHOD AND SYSTEM FOR TREATMENT OF BLOOD AND/OR OTHER BODY FLUIDS AND/OR SYNTHETIC FLUIDS WITH ELECTRIC FORCES

[76] Inventors: Steven Kaali, 88 Ashford Ave., Dobbs Ferry, N.Y. 10522; Peter M. Schwolsky, 4101 Cathedral Ave., NW., Washington, D.C. 20016

[*] Notice: The portion of the term of this patent subsequent to Aug. 18, 2009 has been disclaimed.

[21] Appl. No.: 615,437

[22] Filed: Nov. 16, 1990

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 562,721, Aug. 6, 1990, abandoned.

[51] Int. Cl.^{5} B01D 35/06; A61K 41/00

[52] U.S. Cl. .... 210/748; 128/419 R; 128/421; 128/783; 128/784; 204/131; 204/164; 204/186; 204/302; 210/243; 422/22; 422/44; 604/4

[58] Field of Search .... 210/243, 748, 764; 128/419 R, 421, 783, 784; 604/4; 422/22, 44; 204/131, 164, 186, 242, 275, 302, 305

[56] **References Cited**

U.S. PATENT DOCUMENTS

592,735	10/1897	Jones	204/242
672,231	4/1901	Lacomme	204/275
2,490,730	12/1949	Dubilier	204/305
3,692,648	9/1972	Matloff et al.	204/129
3,753,886	8/1973	Myers	204/186
3,878,564	4/1975	Yao et al.	210/648
3,965,008	6/1976	Dawson	422/22
3,994,799	11/1976	Yao et al.	210/321.64
4,473,449	9/1984	Michaels et al.	204/101
4,616,640	10/1986	Kaali et al.	128/130
4,770,167	9/1988	Kaali et al.	128/788
4,932,421	6/1990	Kaali et al.	128/831
5,049,252	9/1991	Murrell	210/243
5,058,065	10/1991	Slovak	128/783
5,133,932	7/1992	Gunn et al.	210/748

FOREIGN PATENT DOCUMENTS

995848	7/1983	U.S.S.R.	210/243

OTHER PUBLICATIONS

Proceedings of the Society for Experimental Biology & Medicine, vol. 1, (1979), pp. 204–209, "Inactivation of Herpes Simples Virus with Methylene Blue, Light and Electricity"—Mitchell R. Swartz et al.

Journal of the Clinical Investigation published by the American Society for Clinical Investigations, Inc., vol. 65, Feb. 1980, pp. 432–438—"Mechanisms of Photodynamic Inactivation of Herpes Simplex Viruses"—Lowell E. Schnipper et al.

Journal of Clinical Microbiology, vol. 17, No. 2, Feb. 1983, pp. 374–376, "Photodynamic Inactivation of Pseudorabier Virus with Methylene Blue Dye, Light and Electricity"—Janine A. Badyisk et al.

Primary Examiner—Robert A. Dawson
Assistant Examiner—Sun Uk Kim
Attorney, Agent, or Firm—Charles W. Helzer

[57] **ABSTRACT**

A new alternating current process and system for treatment of blood and/or other body fluids and/or synthetic fluids from a donor to a recipient or storage receptacle or in a recycling system using novel electrically conductive treatment vessels for treating blood and/or other body fluids and/or synthetic fluids with electric field forces of appropriate electric field strength to provide electric current flow through the blood or other body fluids at a magnitude that is biologically compatible but is sufficient to render the bacteria, virus, parasites and/or fungus ineffective to infect or affect normally healthy cells while maintaining the biological usefulness of the blood or other fluids. For this purpose low voltage alternating current electric potentials are applied to the treatment vessel which are of the order of from about 0.2 to 12 volts and produce current flow densities in the blood or other fluids of from one microampere per square millimeter of electrode area exposed to the fluid being treated to about two milliamperes per square millimeter.

31 Claims, 6 Drawing Sheets

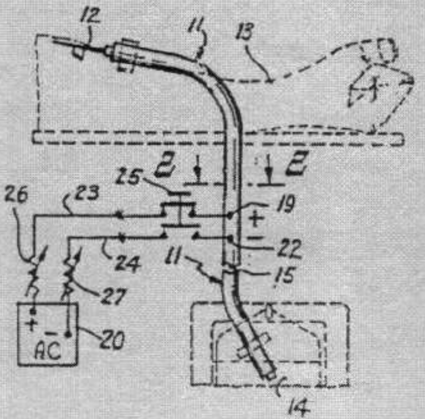

Fig. 1

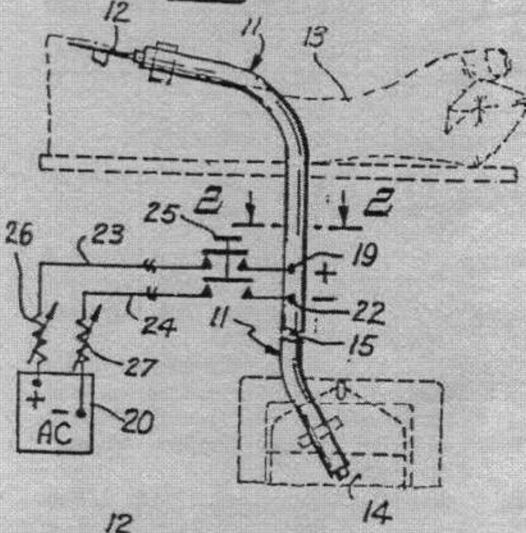

U.S. Patent 5,188,738

ALTERNATING CURRENT SUPPLIED ELECTRICALLY CONDUCTIVE METHOD AND SYSTEM FOR TREATMENT OF BLOOD AND/OR OTHER BODY FLUIDS AND/OR SYNTHETIC FLUIDS WITH ELECTRIC FORCES

FIELD OF INVENTION

This is a continuation-in-part application of prior U.S. patent application Ser. No. 07/562,721 filed Aug. 6, 1990, now abandoned

This invention relates to novel electrically conductive methods and systems employing electrically conductive vessels provided with electrically conductive surfaces for use in subjecting blood and/or other body fluids and/or synthetic fluids such as tissue culture medium to direct treatment by alternating current electric forces.

BACKGROUND PROBLEM

It is now well known in the medical profession and the general public that blood collected in a blood bank from a large number of donors may be contaminated by contaminants such as bacteria, virus, parasites and/or fungus obtained from even a single donor. While screening of donors has done much to alleviate this problem, the screening of donors can and does miss occasional donors whose blood is unfit for use. When this occurs and the unfit blood is mixed with otherwise usable blood, the entire batch must be discarded for transfusion purposes. Because of this problem, the present invention has been devised to attenuate any bacteria, virus (including the AIDS HIV virus) parasites and/or fungus contained in blood contributed by a donor to the point that any such contaminant is rendered ineffective for infecting a normally healthy human cell, but does not make the blood biologically unfit for use in humans. Similar problems exist with respect to treatment of other body fluids, such as amniotic fluids. The treatment method and system is also applicable to mammals other than humans.

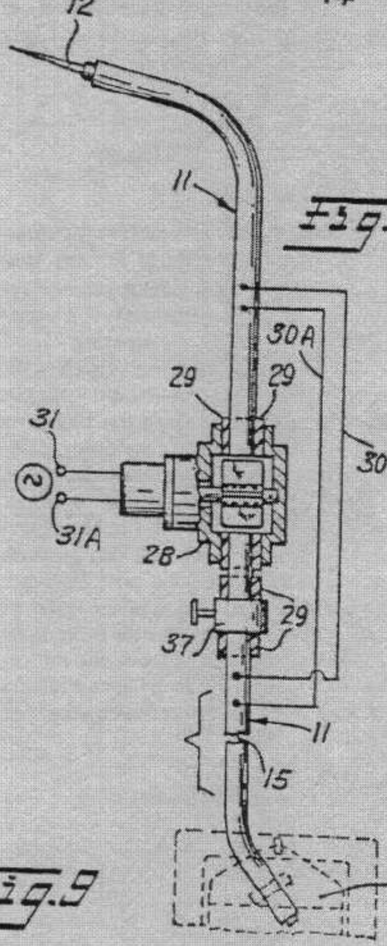

Fig. 6

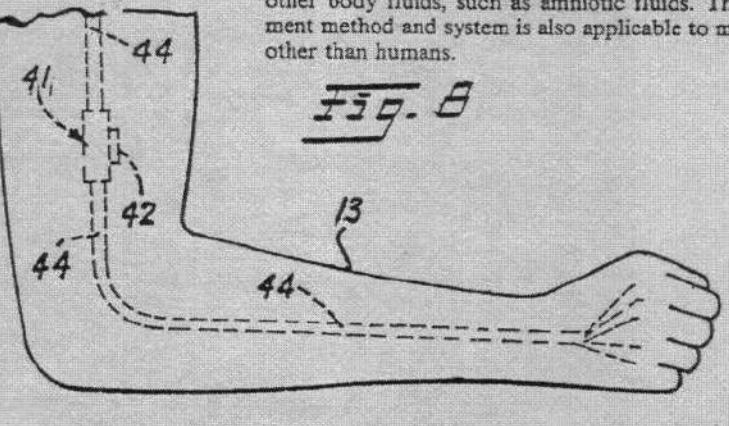

Fig. 8

Fig. 9

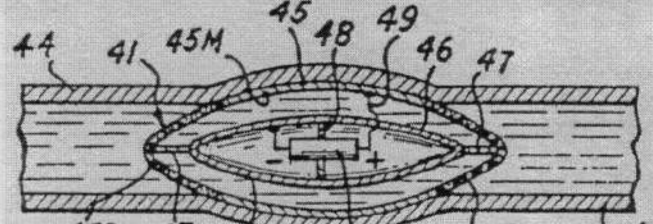

Fig. 10

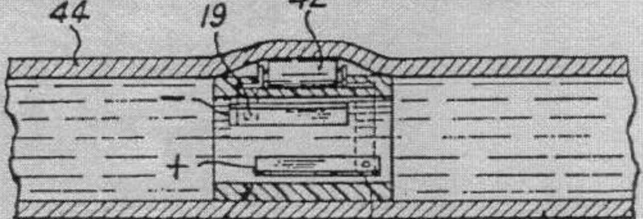

A PROPOSED EXPERIMENTAL/THEORETICAL, NON-INVASIVE, NON-PHARMACEUTICAL, IN VIVO METHOD FOR POSSIBLE NEUTRALIZATION OF HIV VIRUS IN HUMAN SUBJECTS.

Revision March 1999. Copyright 1991/98 by Robert C. Beck, D.Sc.

In a remarkable and startling discovery at Albert Einstein College of Medicine, NYC in 1990, it was shown that a minute current (50-100 microamperes) can alter outer protein layers of HIV virus in a petri dish so as to prevent its later attachment to receptor sites. *(Science News.* March 30, 1991 pg. 207.) It may also reverse Epstein—Barr (chronic fatigue syndrome), hepatitis, Lupus, cancer and many others. HIV—positive users of this enclosed information may expect a dramatic reduction of symptoms after about 45 days. ***This is reminiscent of a well proven cure for snakebite by application of electric current that instantly neutralizes the venom's toxicity.*** *(Lancet,* July 26, 1986 pg. 229.) And there may be several other diseases as yet undiscovered or untested viruses neutralizable with this discovery; perhaps more surprisingly, even the common cold.

This very simple and valid blood clearing treatment proved of great promise as a positive method for immobilizing known strains of HIV still present and contaminating some European and US blood bank reserve supplies. It was further suggested that infected human HIV carriers *could recover* by removing their blood, treating it electrically and returning it by methods similar to dialysis or by surgically implanting electrode arrays with miniature batteries sewn inside blood vessels as described in US patent #5,188,738. Dr. S. Kaali, MD, projected that "years of testing will be in order before such an *in vitro* (blood removed for treatment) device can be made ready for widespread use." *(Longevity,* Dec. 1992, page 12.) This paper reveals an alternate do-it—yourself approach for electrifying/purifying blood *with no dialysis, implants, or medical intervention.*

In the writer's opinion both blood and lymph can be cleared *in vivo* (which means blood isn't removed or skin ever penetrated) simply, rapidly and inexpensively with similar but *non—invasive do—it-yourself techniques* described herein. All are fully disclosed in this paper. Also included are proven schematics, parts lists, electrode construction and complete instructions. Electronic and controlled electroporation approaches may easily make vaccines, antibiotics and pharmaceuticals obsolete.

In a public lecture (Oct. 19, 1991) the writer proposed this theoretical do—it—yourself method for eliminating HIV, parasites, fungi, viral and pathogens *in vivo*. Subsequently, his original modalities and protocols have been refined, simplified and made universally affordable. These three simple treatments used in tandem are offered for research purposes to determine the extent to which they nullify known electro-sensitive pathogens residing in blood, lymph, and other body tissue and fluids. Following is a summary of several years of testing with this non—iatrogenic, do—it-yourself, simple and inexpensive experimental solution to the ever—escalating "incurables" **dilemma.** ***There are no known side effects since milliampere currents applied to skin are much lower than those in FDA approved TENS, CES and muscle stimulators which have been in safe daily use for many years.*** Battery replacement costs are about $10.00 per month per user or about 330 per day for a typical 30 day "spontaneous remission." No pharmaceuticals, shots, or medications appear necessary.

One compact battery—powered blood clearing instrument is basically a miniature relay driven by a timer chip set to —4 Hertz. Its 0 to 33V user adjustable biphasic output minimizes electrode site irritation. The described system delivers stimulation through normally circulating blood via electrodes placed at selected sites on wrist or arm. Optimum electrode positions are reliably located by feeling for strongest pulse (page 19.) Micro-current treatment is of such low amplitude that it creates no discomfort when used as directed and is demonstrated to have no harmful side effects on healthy blood cells or tissue.

A major obstacle to this simple, proven and obvious solution is disbelief. Treating approximately 120 minutes per day for four to twelve weeks should in the writer's opinion effectively neutralize well over 95% of HIV plus any other electro-sensitive viruses, parasites, bacteria, pathogens or fungi in blood.

In heavy infections, shorter application times will prevent stressing patients with toxins. Simply treat for a greater number of days or also ingest ozonized water. In time, the restored immune system plus silver colloid may handle residual problems. In the special case of diabetically impaired circulation longer treatment times may be indicated. (See expanded instructions pages 22-23). Immobilized viruses may be expelled naturally through kidneys and liver. More rapid neutralization is possible but ***not*** recommended because of potential excessive toxin elimination (Herxheimer's syndrome.) T—cell counts usually drop initially because of lysing and scavenging by macrophages but should recover and increase after a few months. There is too much controversy around the PCR test and it has not been proven as a reliable measurement for HIV.

Latent/germinating HIV reservoirs in body's *lymph* or other tissues and glands may theoretically be neutralized with a second and separate device by the strategy of generating a very high intensity (~20 kilogauss) short duration (-10 uS) magnetic pulse of >~32 Joules by discharging a modified strobe light's capacitor through an applicator coil held against body locations over lymph nodes, thymus, kidneys, adenoids and other possible internal sites of latent infection (see page 43). By the physics of Eddy current/back e.m.f. "transformer action" (Lenz' law) the desired criteria of minimum current induced through infected tissue on the order of >100uA should be readily attained. Several pulses repeated at each site over several weeks may insure a reliable "overkill" for successful disease neutralization. A magnetic "pulser" is very simple to build. Full instructions are on page 41-42.

Subjects must assume responsibility for their own health. In today's society, we are conditioned to look for answers only to a medical establishment. Unfortunately current medical practices have been unable to cure so many illnesses today.

These "theoretical solutions" are being disclosed under constitutional freedom of speech guarantees. ***Data* is offered only as theoretical and no medical claims are made or implied.** "See your health professional!" Anyone at his discretion and assumed responsibility should be free to build, use (on himself) and network his "research" results. With this data, someone with a little knowledge of electronics should confidently be able to assemble *both* theoretical blood and tissue clearing modalities in about three hours. Components are widely available. After assembling, the only additional cost is for batteries. If electronically unskilled, "busy" or technically illiterate, call an "Amateur Radio Supply" store (yellow pages) or find a ham radio operator, hobbyist or TV repairman. After "spontaneous remissions" some users may wish to interest their doctors. But be advised that *electronic* cures may be vigorously suppressed or ignored because there is presently no credibility nor drug cartel profit in this inexpensive solution. Also the 1910 Rockefeller—Flexner Report attempted to discredit electro—medicine with intent to give pharmaceuticals a monopoly.

I'm definitely *not* soliciting funds. This was independently developed by me, Bob Beck, at my private expense and offered freely for "theoretical, informational, and educational purposes only" and with absolutely no profit motive.

Non-FDA approved devices are illegal to use within the USA except via little known FDA loopholes. *Researchers* are allowed to use *anything* on patients if safe, they build it themselves and don't sell them. (Code of Federal Regulations 21 § 807.65 subsections {d} & {f} . See actual text in footnote on page 21). Although we will offer technical updates and always welcome feedback from users, please respect the writer's privacy and never attempt to contact him for additional help, advice or information. *Everything* users need to know is included herein.

Feedback: SOTA Instruments Inc. gathers feedback from individuals and health practitioners on the use of The Beck Protocol.

1-800-224-0242 or 250-770-2023

Fax: 250-770-1999, www.sotainstruments.com

IMPROVED SCHEMATIC

This 11/24/1996 schematic describes a **"Plant Growth Stimulator"** improved since my 1991 design. User tested for over two years, it is solid state (no relays), uses three (not seven) batteries, makes colloids, is much smaller, lighter, silent, with battery saving features.

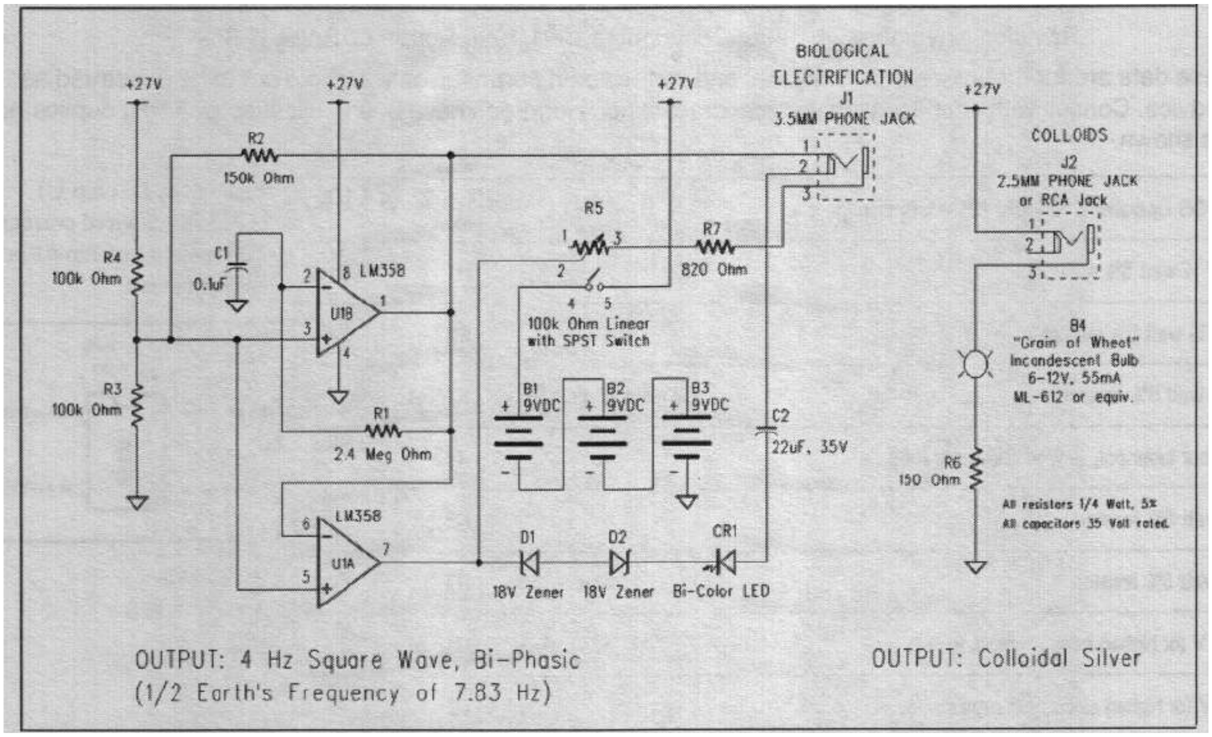
BIOLOGICAL ELECTRIFICATION
J1
3.5MM PHONE JACK
+27V
+27V
+27V
+27V
R2
150k Ohm
R4
100k Ohm
C1
0.1uF
8 LM358
U1B
R3
100k Ohm
R1
2.4 Meg Ohm
R5
100k Ohm Linear with SPST Switch
R7
820 Ohm
B1 9VDC
B2 9VDC
B3 9VDC
C2
22uF, 35V
LM358
U1A
D1
18V Zener
D2
18V Zener
CR1
Bi-Color LED
COLLOIDS
J2
2.5MM PHONE JACK or RCA Jack
B4
"Grain of Wheat" Incandescent Bulb 6-12V, 55mA ML-612 or equiv.
R6
150 Ohm
All resistors 1/4 Watt, 5%
All capacitors 35 Volt rated.
OUTPUT: 4 Hz Square Wave, Bi-Phasic
(1/2 Earth's Frequency of 7.83 Hz)
OUTPUT: Colloidal Silver

Copyright © 1996 Robert C Beck, D.Sc.

The first section (U1B) of the LM358 dual op amp is a 50 volt peak to peak square wave oscillator. The second section (U1A) reverses polarity and provides ±27 Volts DC output of low impedance. This delivers a Bi-Phasic, sharp rise time output of Hz (not critical) for the biological cotton covered stainless steel electrodes saturated with salt water before applying. Sharp rise time is considered necessary to provide higher odd harmonics to the stimulus, although "rounded" waveforms will feel different.

The third section is a current limited 27 Volts DC output from a separate RCA (or 2.5mm) jack for rapid generation of excellent colloidal silver in water. A three minute cycle in 8 Oz. of room temperature water makes a ppm concentration.

Op amp section U1B's 4 Hz oscillator frequency is set by C1 (0.1 uF) and R1 (2.4 meg W). It is configured as a comparator with hysteresis determined by R2 (150 kW). Charging and discharging of C1 is done by the 180° out-of phase signal through Rl. R3 and R4 provide a set point 1/2 the V+ to the comparator. This insures a 50% duty cycle square wave with an amplitude of slightly less than the —27 Volt supply.

U1A, the *second comparator,* is used to invert the output of oscillator U1B. A —50 Volt peak to peak signal will be generated between the op amps due to their outputs being 180° out-of phase. U1A's current is limited by potentiometer R5 (100 kW) and R7 (820 W) and is set to individual user's comfort.

The *power indicator* circuit consists of a bicolor (red green) LED (CR1) and the series combination of two 18 Volt Zener diodes, D 1 & D2, with power limited by C2 (22 uF, 35 Volt). This section of the device is automatically disabled when the 3.5 mm plug is inserted into its jack. Therefore the LEDs flash only when batteries sum is over —21 Volts. If LEDs are dim or extinguished, replace with three fresh 9 Volt Alkaline batteries. C2 used as a limiter allows the LED to flicker on at 1/8 second intervals only as the square wave output reverses polarity.

Users find this newer design highly satisfactory, trouble free and most efficient.

IMPROVED SCHEMATIC - Parts List
PROPOSED THEORETICAL IN-VIVO BLOOD, HIV, PATHOGEN, PARASITE AND FUNGI NEUTRALIZING DEVICE

COMPONENT DESCRIPTIONS, SOURCES, AND CURRENT PRICES

Revision November 24, 1996. Copyright (c) 1991/1999 Robert C. Beck, D.Sc.

Note: These data are for theoretical, informational and instructional purposes only and are not to be construed as medical advice. Consult with your licensed medical practitioner. Hundreds have been built successfully if duplicated exactly as shown.

LM358 CMOS Operational Amplifier (OP-AMP) chip (generic)		U1	$ 1.00
2.4 megohm 1/4 watt 5% resistor		R1	0.07
150 kohm 1/4 watt 5% resistor		R2	0.07
100 k ohm 1/4w 5% resistor	2 required @ 0.15	R3, R4	0.14
100 kohm linear taper pot, 1/2 watt Cattronics P-68		R5	2.56
150 ohm 1/4 watt 5% resistor		R6	0.07
820 ohm 1/4 watt 5% resistor		R7	0.07
0.1 mF 50 V (or higher) ceramic capacitor		C1	0.45
22 mF 35 V (or higher) electrolytic capacitor		C2	0.25
18 Volt Zener diodes (NTE5027A)	% Watt, 2 @ 0.79	D1 & D2	1.58
Bulb, 6.3V 0.075 A	Type 7377	B1	1.34
Bicolor LED Red/Green	RS# 276-012	LED 1	$ 1.19
Jack for electrode leads (3.5mm or 1/8" mono-phone jack)	RS# 274-251 3/$1.59	J1	0.53
Jack for silver wire leads (2.5mm or 3/32" mono-phone jack)	RS# 274-292 3/$1.59	J2	0.53
3 Alkaline 9 V batteries, type 1604 etc.	3 for		3.75
3 x 9 V battery snaps (clip-on connectors)	RS# 270-325 pkg.	5/$1.29	1.29
Misc. wire, solder, etc.			0.50
Box, if used			2.29
Lead wire with 3.5 mm plug, 6 ft., Mouser or Calrad Electronics			0.35
Electrodes, stretch elastic, Velcro, cotton flannel, alligator clips, etc.			5.00
Total Cost **for all** components for improved **do-it-yourself** project			**$22.47**

Ensure the IC chip U 1 (LM358) is wired correctly. The location of Pin #1 is shown below:

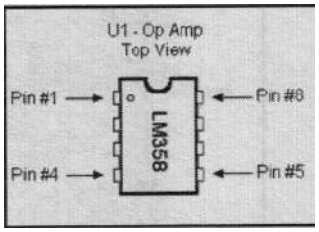

All components listed above check your local electronics store.

Map of Blood Vessels in Extremities for Optimum Electrode Access

Gray's Anatomy pgs. 540, 541, 584, 591

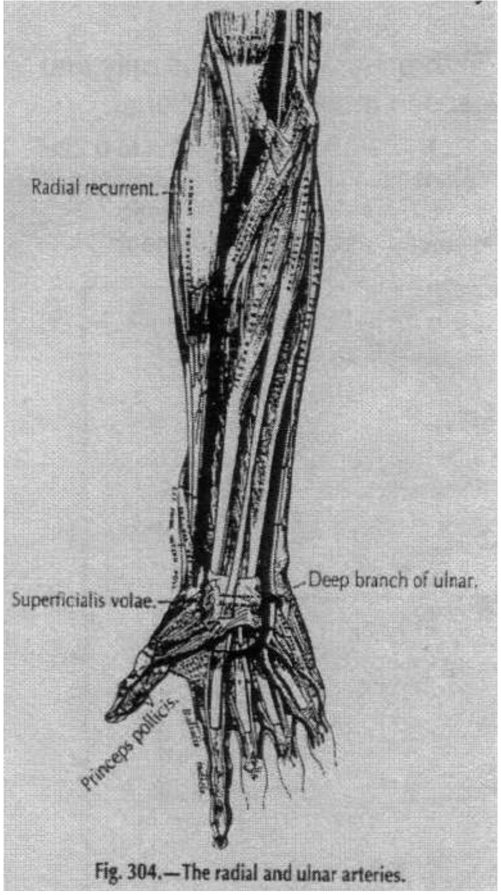
Radial recurrent.
Superficialis volae.
Deep branch of ulnar.
Princeps pollicis.

Fig. 304.—The radial and ulnar arteries.

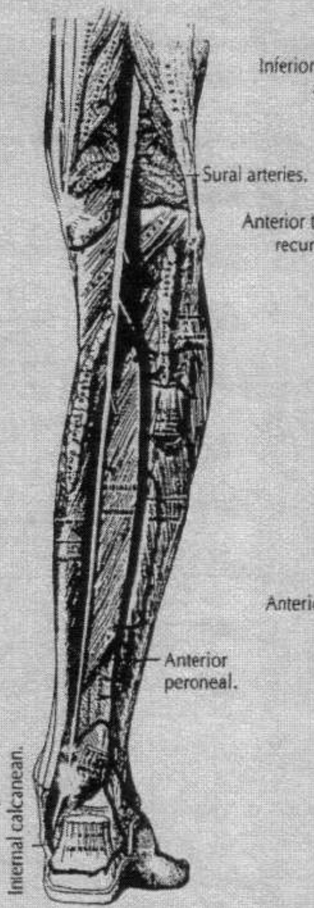
Sural arteries.
Anterior peroneal.
Internal calcanean.

Fig. 320.—The popliteal, posterior tibial, and peroneal arteries.

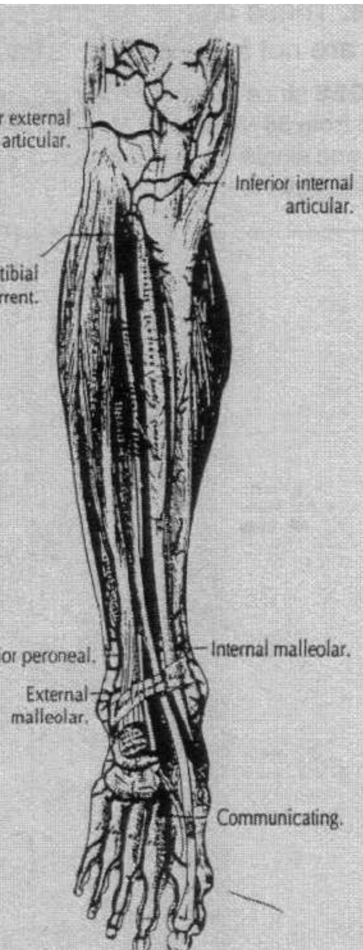
Inferior external articular.
Inferior internal articular.
Anterior tibial recurrent.
Anterior peroneal.
Internal malleolar.
External malleolar.
Communicating.

Fig. 321.—Surgical anatomy of the anterior tibial and dorsalis pedis arteries.

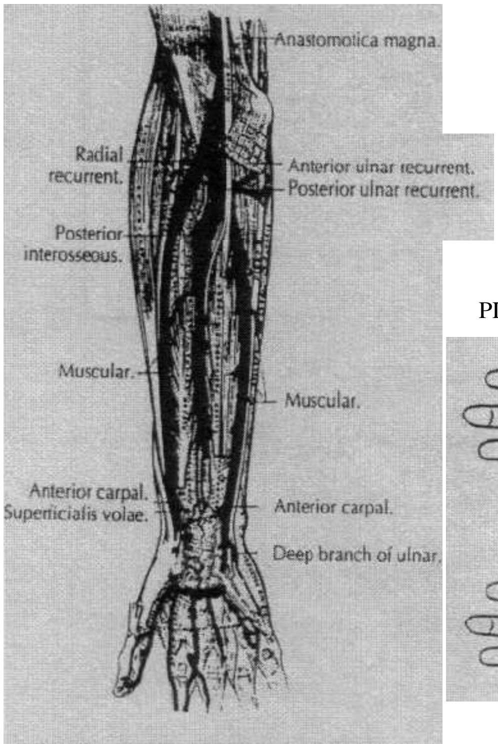
Anastomotica magna.
Radial recurrent.
Anterior ulnar recurrent.
Posterior ulnar recurrent.
Posterior interosseous.
Muscular.
Muscular.
Anterior carpal.
Superficialis volae.
Anterior carpal.
Deep branch of ulnar.

Fig. 305.—Ulnar and radial arteries. Deep view.

PLACEMENT OF PRO 5 (ELECTRODES): Over *Arteries* not blue veins.

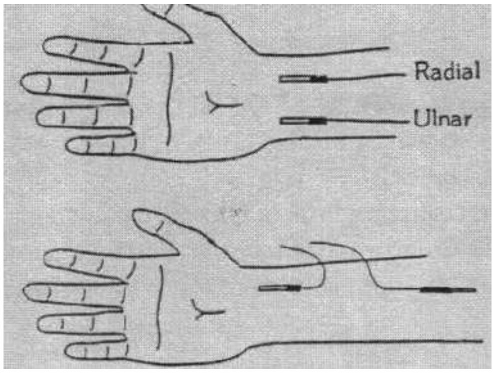
Radial
Ulnar

Align Electrodes Parallel to Pulse and Directly Over Pulse Paths

The figure illustrates the hypothetical placement of the two electrodes on the same wrist. It s easier to place the electrodes first place the strap around the wrist and then insert the electrodes under the strap.

This figure illustrates the hypothetical placement of the electrodes on the same forearm and wrist. Position probes about 6" apart on same Artery

Choose from Two Ways to Align Probes Parallel to Pulse

EXPERIMENTAL IN VIVO BLOOD CLEARING DEVICE FOR ELIMINATING VIRUSES, PATHOGENS, MICROBES, BACTERIA, FUNGI, AND PARASITES.

Revision March 16, 1996. Copyright a 1991/1999 Robert C. Beck

Note: These data are intended for informational, instructional, and research purposes only and are not to be construed as medical advice. Consult your licensed medical practitioner.

CHANGES since previous editions: Pulse Repetition Rate from 0.67 Hz to —4 Hz (Not critical). C2 from 1uf to 0.22uF. Voltage from 36 Volts to 27 Volts. Treatment time increased to 2 hours daily for at least 30 days. Improved electrode design and single wrist placement. SW2 added to extend battery life. There are NO "errors" in this schematic. Hundreds have been constructed successfully when duplicated exactly without user attempted "improvements".

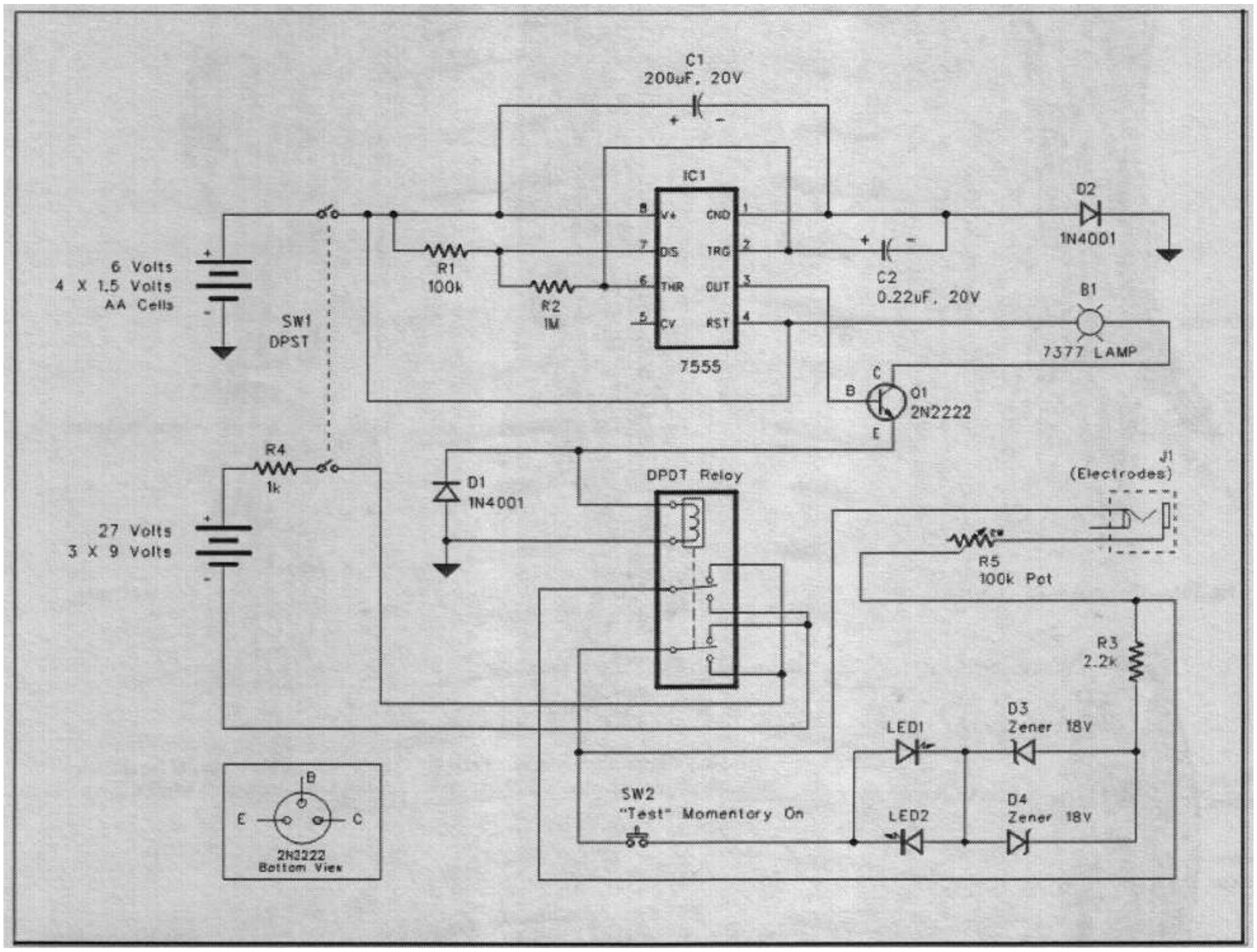
C1
200uF, 20V
+ -
IC1
8 V+
7 DIS
6 THR
5 CV
1 GND
2 TRG
3 OUT
4 RST
7555
D2
1N4001
+ -
C2
0.22uF, 20V
B1
7377 LAMP
6 Volts
4 X 1.5 Volts
AA Cells
SW1
DPST
R1
100k
R2
1M
C
B
E
Q1
2N2222
R4
1k
D1
1N4001
DPDT Relay
J1
(Electrodes)
27 Volts
3 X 9 Volts
R5
100k Pot
R3
2.2k
LED1
D3
Zener 18V
SW2
"Test" Momentory On
LED2
D4
Zener 18V
B
E
C
2N2222
Bottom View

SPECIAL PARTS

B1: Filament type incandescent bulb; 6.3V 0.075A type 7377 (Ballast and current limiter).

Relay: 5V 50 ohm coil, PCB mount DPDT; Selecta Switch SR15P207D1.

D3 & D4: Zener Diodes, 18V ½ Watt; NTE 5027A.

R5: 100 kohm 1/2 Watt linear potentiometer; Caltronics P-68 or equivalent.

LED1 & 2: Combined as Bi-Color Red & Green in same housing; Radio Shack #276-012

SW2: "Test" SPST Momentary ON subminiature Push-Button; Radio Shack #275-1571

PROPOSED THEORETICAL IN-VIVO BLOOD, HIV, PATHOGEN, PARASITE AND FUNGI NEUTRALIZING DEVICE

COMPONENT DESCRIPTIONS, SOURCES, AND CURRENT PRICES

Revision March 13, 1996. **Copyright © 1991/1999 Robert C. Beck, D.Sc.**

Note: These data are for theoretical, informational and instructional purposes only and are not to be construed as medical advice. Consult with your licensed medical practitioner. Hundreds have been built successfully if duplicated exactly as shown.

This design is basically a 7555 IC timer chip set for 50% duty cycle and -3.9 Hz. driving a sub-miniature relay.

Electrode polarity continually reverses -1/4 second positive / 1/4 second negative.

Frequency is not critical.

7555 CMOS timer chip (generic)		ICI	$ 1.80
100 k ohm 1/4 watt 5% resistor		R1	0.07
1 meg ohm 1/4 watt 5% resistor		R2	0.07
2.2 k ohm 1/4 watt 5% resistor		R3	0.07
1 kohm 1/4 watt 5% resistor		R4	0.07
100 k W linear taper pot, 1/2 watt Caltronics P-68		R5	2.56
200 mF 20 V (or higher) electrolytic capacitor		C1	0.45
0.22 mF 20 V (or higher) Tantalum capacitor		C2	0.25
1N4001 diode	2 required © 0.15	D1 & D2	0.30
18 Volt Zener diodes (NTE5027A)	1/2 Watt, 2 @ 0.79	D3 & D4	1.58
NPN Transistor	Generic 2N2222	Q1	0.30
Bulb, 6.3V 0.075 A	Type 7377	B1	1.34
Relay, 5 V 50W coil PCB Mount DPDT Selecta Switch	SR15P207D1		5.45
Misc. wire, solder, etc.			0.50
Bicolor LED Red/Green Radio Shack #276-012		LED 1	$ 1.19
Jack for electrode leads	274-251 3/$1.59	*J1*	0.53
DPDT switch, 275-626 or equiv. (used as DPST)		SW1	2.55
SPST Subminiature momentary ON push button switch ER-SW101 "Test"		SW2	0.59
Battery holder, 4-AA cells, Radio Shack	270-383		1.29
4 Alkaline M cell batteries, #23-552	4 for		2.89
3 Alkaline 9 V batteries, type 1604 etc.	3 for		3.75
3 x 9 V battery snaps (clip-on connectors)	270-325 pkg.	5/$1.29	1.29
Box, if used			2.29
Lead wire with 3.5 mm plug, 6 ft., Mouser or Calrad Electronics			0.35
Electrodes, stretch elastic, Velcro, cotton flannel, alligator clips, etc.			5.00
Custom printed circuit board #PS-PCB:			15.00
Total **Cost for all components for do-it-yourself project**			**$49.24**

EXPANDED INSTRUCTIONS FOR EXPERIMENTAL/THEORETICAL BLOOD ELECTRIFICATION *HYPOTHETICAL PROTOCOLS FOR EXPERIMENTAL SESSIONS*

Revision January 15, 1997. Copyright © 1991/1998 Robert C. Beck, D.Sc.

PRECAUTIONS: Do NOT use wrist–to–wrist with subjects using cardiac pacemakers. Any electrical signals may interfere with "demand" type heart pacers and cause malfunction. Single wrist/forearm locations should be acceptable. Do NOT use on pregnant women, while driving or using hazardous machinery.

Users MUST avoid ingesting anything containing potentially toxic medications, nicotine, alcohol, recreational drugs, laxatives, for several days before starting because blood electrification can cause electroporation which makes cell membranes pervious to small quantities of normally harmless chemicals in plasma. The effect is the same as overdosing. See Electroporation: a General Phenomenon for Manipulating Cells and Tissues J.C. Weaver, Journal of Cellular Biochemistry 51: 426-435 (1993.) The blood purifier can cause electroporation. Avoid ingesting alcohol 24 hours before using and keep in mind that high potency, synthetic vitamins as well as junk foods can be a problem when magnified.

Do NOT place electrodes over skin lesions, abrasions, new scars, cuts, eruptions, or sunburn. Do NOT advance output amplitude to uncomfortable levels. All subjects will vary. Never put bare metal on skin. The magnetic pulser should be safe to use anywhere on body or head.

Drink an 8 oz. glass of filtered or preferably ozonated water 15 minutes before and immediately following each session and drink at least four additional glasses daily for flushing during "neutralization" and for one week thereafter. This is imperative. Ignoring this can cause systemic damage from unflushed toxic wastes. ***When absolutely essential drugs must be ingested, do so a few minutes after electrification then wait 24 hours before next session.***

If subject feels sluggish, faint, dizzy, itchy, headachy, light-headed or giddy, nauseous, bloated or has flu–like symptoms or rashes after exposures, reduce pulsing per session and/or shorten applications of electrification. Drink more water—preferably ozonized—to speed waste oxidation and disposal. Use extreme caution when treating patients with impaired kidney or liver function. Start slowly at first, about 20 minutes per day to reduce detoxification problems.

To avoid shock liability, use only batteries with blood cleaner. Do NOT use any line–connected power supply, etc. with blood electrification device. However a.c. supplies are OK with well–insulated magnetic pulse generators (strobe lights.)

Health professionals: Tobacco, the most damaging (41/2 times more addictive than heroin) and deadly substance of abuse known, disrupts normal cardiovascular function. True vegetarian diets are missing several essential amino acids absolutely necessary for the successful rebuilding of AIDS–ravaged tissues. Secondary gains and hidden agendas (sympathy/martyrdom, work avoidance, free benefits, financial assistance, etc.) can play large roles with many seriously ill patients.

SUPERIOR ELECTRODES: Excellent, convenient and vastly superior electrodes, reusable indefinitely, can be user-made by butt–soldering connecting wires to ends of 1-1/8" long by 3/32" dia. blanks cut from type 316 stainless steel rods available from welding supply stores (Cameron Welding Supply, 11061 Dale Ave., Stanton, CA 90680.) Use "Stay Clean" flux before soldering (zinc chloride/ hydrochloric acid.) Shrink–insulate TWO tight telescoping layers of 3/8" long shrink tubing over soldered joints to prevent flexing/breaking and lead/copper ions from migrating.

PREPARATION FOR USE: Wrap three or four turns of 100% cotton flannel around rods. Spiral–wrap with strong thread starting from wire side to end. Tightly pinch cloth over rod's end so as to leave no metal exposed by wrapping 6 or 7 turns of thread TIGHTLY just off end of rod, then spiral wrap back to start and tie with four knots then cut off excess cloth at end close to pinch-wraps. Treat end and knots with Fray Check® (fabric & sewing supply stores) to prevent raveling. Saturate before each use in a solution of sea salt (not table salt). Add a few drops of Colloidal Silver for disinfectant. Store for reuse. Tape soaking wet electrodes tightly over arterial pulse sites with paper masking, Transpore™ or Micropore® tape or make 1" wide stretch elastic bands with tabs of Velcro® at ends to fasten. Electrodes should closely conform precisely along blood vessels, not skewing ever so slightly over to adjacent flesh. This insures better electrical conductivity paths to circulating blood and insures very low skin impedance. (-2000Q) Rinse and blot-dry electrodes and skin after each use. NEVER allow bare metal to touch skin as this will cause burns manifested as small red craters that heal slowly. The objective is to get maximum current into blood vessels, not leak it over to adjacent tissue. Therefore never use any electrode with cotton cover wider than about 1/8 inch.

ELECTRODE PLACEMENTS: Locate blood vessel paths (*NOT* to be confused with acupuncture, Chapman or pressure points) on either wrist (Ulnar to Radial) or else wrist to forearm by feeling for maximum pulses. Scrub skin over chosen sites with mild soap and water. Position each electrode lengthwise precisely along same branch of Radial artery. A 7" long, 1" wide elastic stretch–band with two 1-1/4" lengths of 3/4" wide Velcro® sewn to ends of opposite sides makes an excellent wrist band for holding electrode snugly in place. Place wet electrodes between ulnar and radial pulse points on opposite sides on inside of ***same*** wrist. Then with electrode cable unplugged, turn switch ON and advance amplitude control to *maximum*. On original blood electrifier design, push momentary SW 2 "test" switch and see that the red and green light emitting diodes flash alternately. This verifies that polarity is reversing –4 times per second (frequency in NOT critical) and that batteries are still good. When LED's don't light, replace all three 9V alkaline batteries. When the white incandescent bulb dims or appears yellowish, or relay isn't clicking, replace all four AA cells. Zener diodes will extinguish LED's when the three 9V battery's initial 27V drops below –20V after extended use. Never use any electrode larger than 1-1/8" long by 1/8" wide to avoid dissipating current through surrounding tissue. Confine exactly over blood vessels. Apply drops of salt water to each electrode's cotton cover as needed to keep cotton covers damp and insure optimum current flow. Later devices are solid-state, use only one battery and no relays, and are much smaller. Electrify blood 2 hours daily for 4 weeks or longer as indicated. (See diagram on page 19.)

SETTING VOLTAGE: Now rotate amplitude control to *minimum* (counter-clockwise) and plug in electrode cable. Advance dial slowly until feeling a "thumping" and tingling. Turn as high as tolerable but don't advance amplitude to where it is uncomfortable. Adjust voltage periodically when acclimatizing to comfort level after several minutes. It is normal to feel different sensations with time. You may notice little or no sensation at full amplitude immediately, but feeling will begin building up to maximum after several minutes at which time amplitude can be decreased. Typical adapted electrode–to–electrode tissue impedance is on the order of 2000 Ohm. Typical comfortable input (to skin) is ~3mA, and maximum tolerable input (full amplitude) is ~7mA. However, this "reserve" margin although harmless is unnecessary and can be uncomfortable. Current flowing through blood is much lower than this *external* input because of series resistances through skin, tissue and blood vessel walls, but 50 to 100uA through blood is essential.

FREQUENCY OF USE: Apply blood electrifier for about two hours daily for at least 1 to 3 months. Use judgment here. The limiting factor is detoxification. Carefully monitor subject's reactions (discomfort, catarrh, skin eruptions, weeping exudites, rashes, boils, carbuncles, coated tongue, etc.) With very heavy infections, go slower so as not to overload body's toxic disposal capability. Drinking ozone–bubbled water oxidizes wastes and speeds detoxification. With circulation–impaired diabetics, etc., you may wish to *extend* session times. Again, have subjects drink lots of water. Recent changes in theoretical protocol being currently tested suggest continuing the protocol for several weeks. One health practitioner with extensive use of both the blood electrification and magnetic pulsing devices reported to Sharing Health From the Heart Inc. to suggest prolonged use to ensure the stronger pathogens have been neutralized. He points out the danger of stopping too soon if only the weaker pathogens have been dealt with. He has suggested building up to using the blood electrification device for 6 days continuously to neutralize the longest-lived pathogens. Remember to remoisten electrodes regularly. If you absolutely ***must*** ingest prescription drugs, do so immediately ***after*** turning off instrument and allow 24 hours before next treatment to let chemical concentrations in blood plasma decay to lower levels.

Remember, if subjects ever feel sleepy, sluggish, listless, nauseous, faint, bloated or headachy, or have flu-like reactions, they may be neglecting sufficient water intake for adequately flushing toxins. We interpret this as detoxification plus endorphin release due to electrification. Let them rest and stabilize for –45 minutes before driving if indicated. If this detoxing becomes oppressive, treat every *second* day. Treatments are shown to safely neutralize viruses, fungi, bacteria, parasites and microbes in blood. See US patents #5,188,738, 5,091,152, 5,139,684, 5,328,451, 3,753,886, 4,524,079, 4,665,898 and others as well as numerous valid medical studies which are presently little known or suppressed.

SILVER COLLOID: Also, ingesting a few oz. of –5 parts per million silver colloid/silver ion solution daily can give subjects a "second intact immune system" and minimize or eliminate opportunistic infections during recovery phase. This miracle substance is pre-1938 technology, and unlike ozone is considered immune from FDA harassment. Silver colloid can *easily* be made at home electrolytically in minutes and in any desired quantities and parts per million strength for under 10 per gallon plus cost of water. It is ridiculous to purchase it for high prices. Unlike silver proteins, metallic colloid has no side effects. Silver colloids won't produce drug resistant strains as will all other known antibiotics. No reasonable amount can overdose or injure users either topically, by ingesting, or professional medical injection. Refer to page 31 for complete instructions on successfully making your own.

Preparation to Use the Electrodes for Blood Electrification

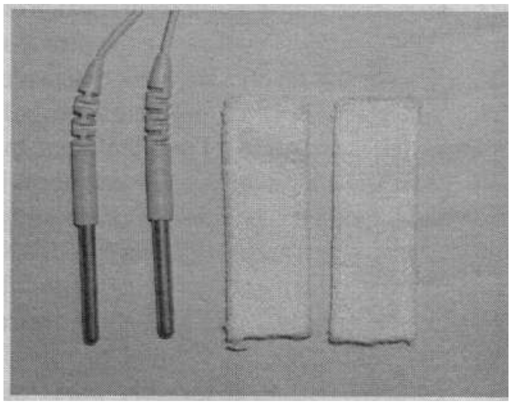

Step 1: Prepare cotton sleeves or cotton covers for the electrodes.

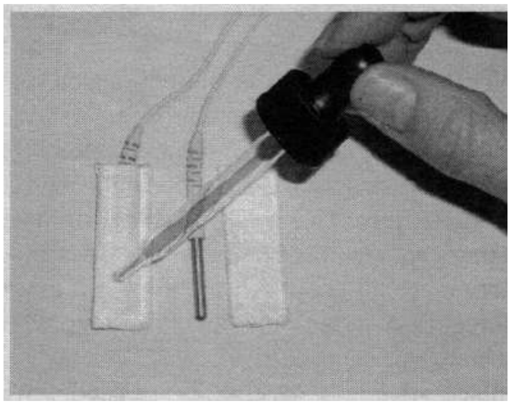

Step 2: Place the cotton sleeves over the electrodes. A dropper bottle is a handy way to wet the cotton and to keep the covers damp during use.

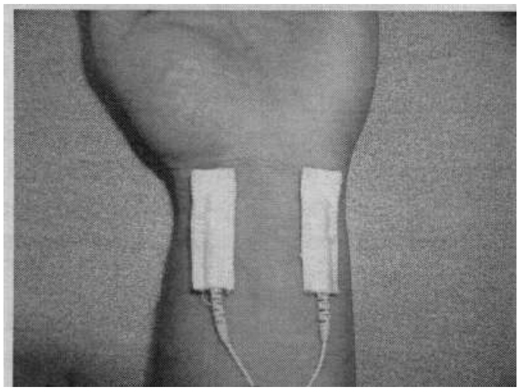

Step 3: Place the covered electrodes directly over the two arteries on the wrist: the radial and ulner arteries.

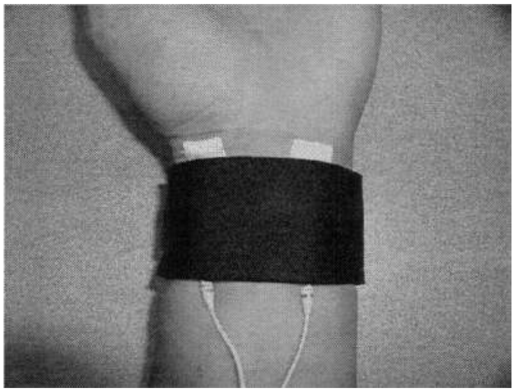

Step 4: Hold the electrodes in place with a wrist strap. This one is made of waterproof neoprene with velcro to hold it firm.

Ensure the wrist strap does not get wet between the electrodes. If it does the current will travel between the two electrodes rather than through the skin to the arteries.

Total Cancer Remissions through Blood Electrification combined with Silver Colloid

Theories Offered for Information and Educational Purposes Only and are the Author's Opinions

Copyright © 1996 Robert C. Beck, D.Sc. re Re-issued May, 1998

Opinions expressed in this article are based on the author's first hand experience and do not necessarily reflect those of this publication. [Also published in *Explore!* Magazine]

My archives contain a tantalizing report from several decades ago describing an authenticated record of an older man who was struck by lightning, survived, and subsequently grew a third set of teeth and a bushy head of youthful new dark hair. His grossly metastasized, inoperable cancers vanished. He threw away his glasses and cane, and appeared much younger and was totally healthy for the first time ever. This fascinated scientists and years later almost encouraged some highly illegal and bizarre human experiments in an abandoned aircraft hangar in Wendover, Utah where Tesla coil research with ball lightning was underway. The incident generated wide speculation, but few insights at the time. This mystery remained sleeping until 1990 when an astounding discovery was reported at Albert Einstein college of Medicine in NYC by Drs. Kaali and Lyman. Not surprisingly, the revolutionary data were apparently immediately suppressed (See *Science News:* Mar. 30, 1991 pg. 207; and *Longevity:* Dec. 1992 pg. 12.)

As a totally unexpected and unpredictable outcome of the writer's self-funded research into "blood electrification" with micro currents for AIDS (currently showing excellent results), a growing number of users previously unknown to me began independently reporting "spontaneous remissions" of numerous *other* diseases ***including cancer.*** Most involved no doctors, medication, or time off. Recoveries occurred after subjects had self-administered an altered do-it-yourself blood treatment process first described in US Patent #5,188,738 issued to Dr. Steven Kaali in '93. We were puzzled to find explanations as to why this worked. This preliminary report offers a possible theory. Magnetic pulsing success with cancer were independently proven again in 1984 and described in US Patent #4,665,898 plus many other patents dating back to 1890!

The Einstein disclosure describes removing blood from one arm, electrifying it, and returning it to the other arm in a process similar to dialysis. It also describes surgically implanted active electrode chambers containing miniature batteries sewn inside blood vessels. This author's preferred approach leaves all blood in the body, is totally non-invasive, costs practically nothing and is safely accomplished in about a month with –two hours per day exposures as one goes about his normal activities. It handles pathogens while blood flows normally through the –60cc volume of the electrified forearm's ulnar branch artery from wrist to elbow. Without medications, invasive techniques or doctors, most pathogens, viruses, microbes, parasites and fungi just tend to disappear. Progress can be readily observed in blood with dark-field and phase-contrast microscopy. The entire process and simple apparatus with parts list is fully described in my '91 paper reprinted in recent issues of *Explore!* (Vol.7 #1) and in this paper. Also simple instructions for self-made silver colloids of far better quality than you can usually buy are given in Vol. 7, #2 of *Explore!* also *Explore More,* Issue #15. Back issues: Box 11510, Prescott, AZ, 86303, 928-541-1920. You can turn any glass of tap water into a 3 ppm ionic electrolytic silver colloid in about two minutes anywhere with a shirt pocket instrument by simply running 27V (three 9V transistor radio batteries) between two short lengths of pure silver (never "Sterling") wires submerged in water. To date many "spontaneous remissions" of dozens of "incurable" illnesses including HIV have been reported by users and researchers of this "blood purification" when combined with ingestion of pennies-per-gallon instantly self-made ionic silver colloid plus magnetic pulsation and ozonized drinking water. Since none of dozens of friends using these apparent miracles has experienced infections, colds, flu, pneumonia, or lost a single day's productivity in over three years, evidence strongly suggests restored immune systems or dramatically improved blood functioning. It is also fascinating to note that several pet owners report their cats now refuse to drink water if silver colloid is not added. Trips to veterinarians with previously recurring infections were cut dramatically. It is as though the Creator had left a secret "back door" method for mankind to finally conquer the adversary - incurable diseases plaguing us since the beginning of time. To avoid legalities, these data are being offered under first amendment freedom of speech rights and should not be construed as medical advice.

It has long been known that dissections of cadavers dying of natural causes reveal many have had cancer several times during their lifetime resulting in "spontaneous remissions" generally without their knowledge and without ever visiting a doctor. An optimally functioning immune system somehow automatically "handles" diseases of which the subject seldom becomes aware. Several promising broad-spectrum natural immunological agents like interferon and interleukin are produced by healthy immune systems but would cost thousands for patients with already overloaded or "shut down" defenses although many such neuropeptides could speed cures. Other respected researchers describe "pleomorphic" forms of cancer pathogens which evolve through several stages—even mycotoxin involvement—all of which surrender to blood electrification. For persons unable to self-assemble the simple electrifier (about two hours and –$50), dozens of people are currently custom building them and several companies are providing excellent and reliable combination blood electrification and colloid generators ready to use. Most retail ready-to-use from $135 to $200. But this health breakthrough is politically incorrect and may never be FDA approved because of the millions of dollars and years needed to sort through an approval process and the billions invested in treatment facilities, pharmaceuticals, and in clinical, diagnostic equipment This discovery gives power over diseases *back to the individual*. The only dangers lie in too rapid detoxification avoidable by increasing ozonized water intake for flushing and oxidizing wastes erxheimer's syndrome), plus avoiding potentially toxic substances during blood electrification because of vastly enhanced cell absorption due to electroporation. (See J.C . Weaver: Harvard-MIT in *Journal of Cellular Biochemistry,* 51:426-435; 1993).

All drugs, alcohol, tobacco, pharmaceuticals and coffee *must* therefore be discontinued for at least two days before starting and for the duration of blood electrification or magnetic pulsing. This minimizes substances in your blood plasma which may become toxic at –20X their normal levels.

Electrification is now being successfully used underground around the world. One example—cervical cancer alone kills –1/3 of all victims in the third world, and has long been known to be caused by the papilloma virus. Electrification has the potential to eliminate these toxic "fellow travelers' coexisting in our blood and may handle 'innumerable other diseases previously considered "incurable." Ebola or other possibly genetically engineered biological warfare "designer" plagues may be unleashed someday per some theories of Gulf War Syndrome diseases which are currently immune to all other known countermeasures. Blood electrification, colloids and ozone may be effective against these viruses as well. Unfortunately in today's society, millions of dollars are needed to do the research for more people to be helped.

The most reasonable theory of why electrification is so surprisingly effective for so many conditions lies in the now proven fact that when correctly applied directly into blood *(not* into other body tissue like palms of hands, soles of feet, or organs) the process neutralizes all microbes, pathogens, fungi, parasites, viruses, bacteria, mycotoxins and coexisting foreign life forms and alien invaders and their byproducts. In my opinion, this should never be confused with Royal Rife, Hulda Clark or others' modalities. Effective results are found to require a *minimum* of 27 Volts measurable under load with low impedance output which must deliver up to several milli-amperes measurable current into skin to produce the required 50 to 100 micro-amperes internally through blood after the inevitable series resistance losses through vessel walls plus several layers of tissue. Electrical currents in blood can be measured with an AC microvolt meter by IR drop using partially insulated hypodermic needles inserted –6 inches apart into the same artery. Clark's "syncro-zap" running at her standard 30 kHz (considered many octaves too high to be effective) actually measures only –2.6V peak to peak under load (~2000 ohms) at palms. The syncro-zapper's current is unmeasurable directly in bloodstream and physically *cannot* produce the essential 50 to 100 uA required internally. Using the syncro-zapper may only mask readouts of parasite presence radionically. Unfortunately the live bugs remain undisturbed and are still there and will still be observed in stool and microscopic blood diagnosis. To function at all, electrification requires cotton-covered salt water saturated stainless steel electrodes never over 3/32" wide and 1" long. Electrodes must be carefully positioned directly over and precisely in line with specific arterial pulse points. This maximizes current into *blood* by not diffusing it into surrounding tissue. Square or round TENS, EKG, EEG, EMG; etc. electrodes work only marginally and should never be substituted. Preferred instrument pulse-repetition rate is –3.9 Hz biphasic with steep rise time and 50% duty cycle. Rate is *not* critical although much higher frequencies and certainly higher harmonics of the essential square wave output are degrades by "skin effect" where currents travel around the outside of body instead of internally. This is demonstrated by lighting a bulb in one hand while touching a Tesla coil with the other and not getting shocked.

Careful electrification causes no known harmful side effects to healthy cells or tissue. Blood electrified for 18-24 hours a day for three months were observed to live for well over a month when sealed under cover slips on microscope slides while the average life of "normal" drawn blood is under 4 days. This strongly suggests that even aging bodies may easily and rapidly be made impervious to many hostile, toxic, infectious, antibiotic-resistant and even yet undiscovered invaders.

Like all revolutionary ideas, this incredible breakthrough barely survived initial ridicule and rejection because it is too startling, effective, inexpensive, simple and foolproof to be generally believable. Almost universally, people simply refuse to take responsibility for their own health. They think the "Doctor" should know what's "best" for them. So acceptance of blood electrification, through word of mouth, is just now emerging to enthusiastic acceptance from those who've actually experienced the results. Being profit-motivated, the establishment *must* resist anything like this. But we now have our "hundredth monkey." Skeptics have only to use this technology to directly enjoy immensely better health. ***Take back your power!*** This works! The writer is a researcher, not a practitioner. **I have nothing to gain from this!**

Blood Electrification The Easy In-Vivo Way

Revision May 1998 Informational use only. Not intended as medical advice. Copyright © 1997/99 Robert C. Beck

Several years of experimentation and many reports of successes have resulted in simple, fast, proven ways to implement this most important step in self-healing.

Electrify two hours daily for a minimum of four weeks.

1. Build or purchase a ready–to–use device, or put finishing touches on a nearly–completed kit.

2. Prepare and label a sea–salt–in–water solution for electrode wetting. Use 1/8 teaspoon of sea salt in a 2 ounce dropper bottle. Greater salt concentration can cause osmotic skin burns, irritation and rash. Add water and few drops of colloidal silver. Let salt dissolve. Filter through paper towel to clarify this storable conductive interface between cotton covered electrodes and skin. NEVER let bare metal electrode or any small area of metal touch skin directly, or it will burn. Use natural salt only, not table salt containing iodine for goiter and aluminum and silicates to insure easy pouring.

3. For best electrical conductance scrub skin at electrode locations with soap and water to eliminate skin oils, grime, dead skin cells, etc. Rinse and dry. With fingertip rub a drop of salt water into skin at each electrode site.

4. Referring to illustrations on page 19, carefully feel for pulses and trace a line about 1 inch long at each wrist site precisely on top and in line (parallel) with located blood paths. Pulses on Ulnar location opposite thumb are harder to feel. Never place electrodes over new lesions, cuts, abrasions, or sunburn. Muscle twitching in palm and fingers is normal and experienced occasionally.

5. Dip into bottle to saturate initially. Position wet electrodes not over –3/32" wide 1-'/4" long to wrist precisely over traced pulse paths. Slide each from forearm side underneath a snug 1" wide stretch elastic band with Velcro® retaining overlapped ends. One electrode positions on radial (thumb) side, the second on other (ulnar) inside of same wrist. Current is confined to blood in lower forearm. Very little electrification is detectable elsewhere thus making it safer for heart-pacer users. The wrist electrode placements are more convenient and faster than any other positions.

6. Put larger units in pocket and run electrode cord down sleeve or strap the smaller electrifier—single battery units—with stretch–band to forearm. Plug in electrodes, turn on and advance slowly to comfortable level. I prefer the small, convenient, unobtrusive unit that uses one instead of the three 9V battery units. Neither interferes with normal activities.

7. Re-adjust power occasionally to maximum comfortable level. You can now even sleep with it on without fear. When the treatment (about two hours per session daily for a month or more but only after detoxifying) is done, turn it off and put it aside until tomorrow. When red and green LED's flash alternately with electrodes unplugged you know it's working properly. Blood cleansing can be speeded with heat. Example: wrapping forearm with electrodes in heating pad set to high.

8. Keep electrodes wet by re–moistening with drops of salt water occasionally using eye dropper. When finished, rinse wrists. Wash electrodes periodically with soap, water and soft toothbrush to eliminate skin oils and soil. Soaking overnight will dissolve caked salt. Discoloration at ends is normal. When frayed or worn, discard old covers and re-wrap stainless rods with 3 turns of 100% cotton flannel. Wrap tightly with a few turns of thread to end, spiral back to beginning and tie. Electrodes should last for months, but wire leads break and must be replaced eventually.
   [Editor's Note: If skin is particularly sensitive, try using conductive gel on the electrodes or reduce the amount of salt in the water to minimize skin irritation. Apply a healing salve after each session.]

What's inside the box?

Preferred electrifiers must generate a 3.9 Hz (not critical) biphasic sharp-rise-time square wave, ± 27V peak adjustable output, 50% duty cycle, capable of delivering several milliamperes into a low resistance load at skin surface (—S 2000W impedance) which after losses in tissue resistance delivers the necessary 50 to 100 microamperes through flowing blood.

Precautions:

Badly debilitated patients such as full—blown AIDS victims should begin at less than 20 minutes every second day and flush by drinking lots of pure water, preferably ozonized because their systems will go into rapid detoxification causing physical problems called Herxheimer's syndrome.

Users taking ANY medications should minimize such presence in blood for at least two days before starting and avoid irritants including coffee, tea, alcohol, tobacco, recreational drugs, etc., during the several weeks of recovery. "Electroporation" is shown to increase dosage levels up to 20X of anything drunk, shot, or ingested thus causing problems. This documented by J.C. Weaver, Harvard-MIT *Jr. Cellular Biochemistry,* 51: 426-435; 1993. Patients needing essential medications should take them immediately ***after*** turning off electrification and wait 24 hours before next blood cleansing. This lets their residues decay to minimum levels in plasma before re—electrifying.

If detoxing becomes disturbing, proceed even more slowly. Symptoms may include fever, giddiness, dizziness, headaches, light—headed, vagueness, nausea, skin rashes, eruptions, itching, boils, coughing, kidney and liver discomfort, aches, general malaise, inflammations, frequent urination, and sluggishness. Use caution when detoxing patients with impaired liver or kidney function. But remember it's far better to force wastes *out* of your system than leaving them stored where they may have been hiding for years.

Treat slowly if initial discomfort occurs. Electrification will profoundly affect your health and provoke your deepest mind—sets such as everyone's unconscious conflicting death wishes. This generally causes noticeable anxiety and depression.

RECOMMENDED SUPPLIER

For Electronic Parts: Check your local electronics supply store.

For Complete Units:

Robert (Bob) C. Beck, D.Sc. worked closely with Russ Torlage of SOTA Instruments Inc. in the years preceding his passing. He recommended SOTA as a supplier for The Beck Protocol units.

Contact SOTA at: 1-800-224-0242,250-770-2023 Fax: **250-770-1999** or

Write: PO Box 20019, Penticton, BC V2A 8K3

www.PositiveOffset.com

Electronic Kit & Parts for Educators, Researchers & Hobbyists

30 kHz / 1 kHz Zapper Kit with Color Illustrated Instructions

$50.00 Add to Cart

Our Zapper Kit contains all the parts necessary to build either a 30,000 Hz or a 1,000 Hz Positive Offset Square Wave Frequency Generator Unit. This breadboard kit is based on Dr. Hulda Clark's zapper circuit design that was recently updated and published in 2003. Copper Pipes and 9 Volt Battery are not included in kit. For copper pipes and individual electronic parts, click the links in the right column.

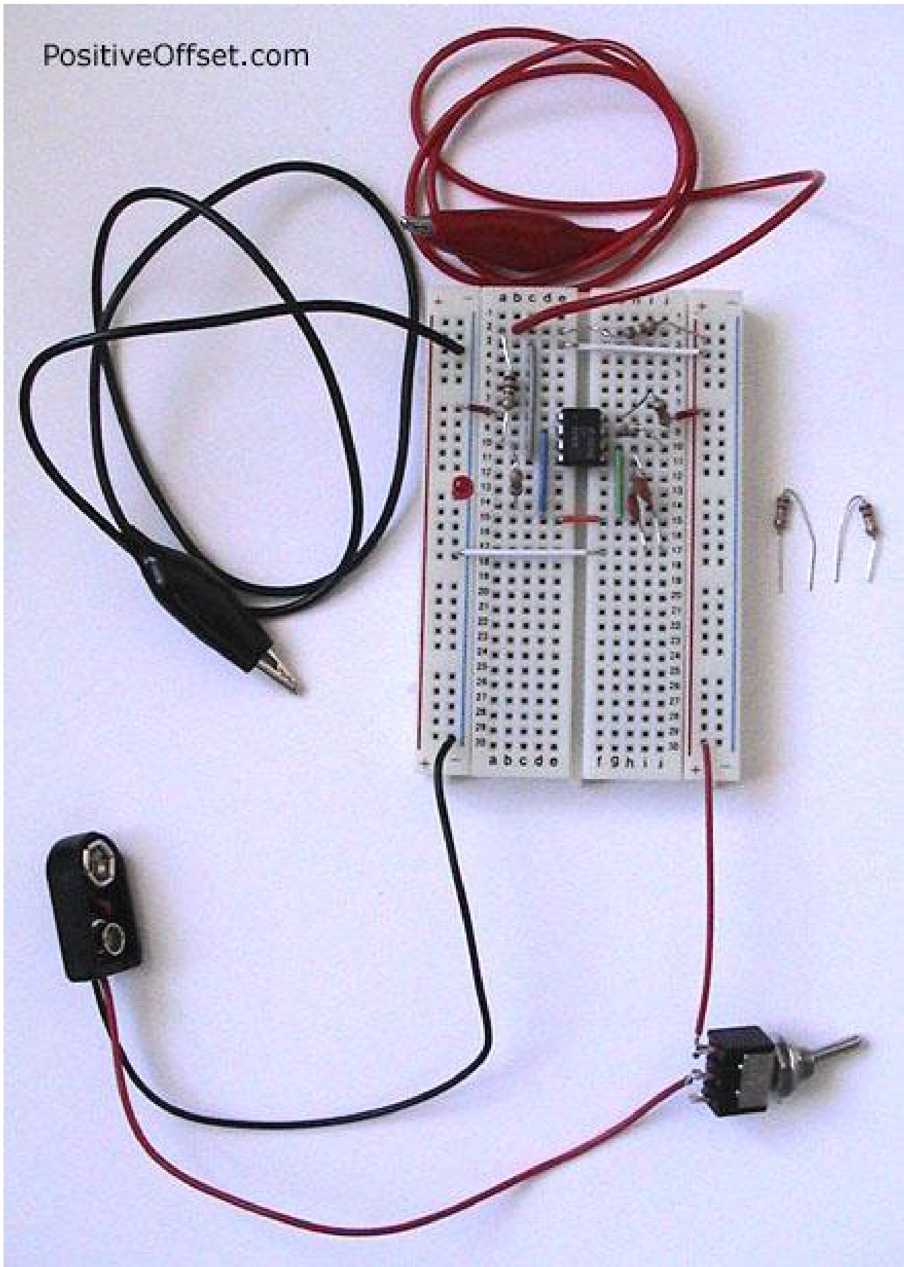

Hulda Clark Zapper Kit with Color Illustrated Instructions
$50.00 Add to Cart

Zapper Kit includes: Breadboard, 555 CMOS Timer, (2) 1K Ohm resistors, 2.2K Ohm resistor, 4.7K Ohm resistor, 39K Ohm resistor, (2) 270K Ohm

www.PositiveOffset.com

- 30 kHz/1 kHz Zapper Kit
- .0047 uF Capacitors
- 1 uH Inductors
- 1K Ohm Resistors
- 2.2K Ohm Resistors
- 3.9K Ohm Resistors
- 4.7K Ohm Resistors
- 39K Ohm Resistors
- 270K Ohm Resistors
- 555 CMOS Timer
- 8 Pin Wire Wrap Socket
- 9 Volt Battery Snap
- Alligator Clip Converter
- Alligator Leads
- Aluminum Square
- Banana to Alligator Leads
- Breadboard
- Color Coded Jump Wires
- Copper Pipes
- Nuts and Bolts
- On/Off Toggle Switch
- Red Mini LED
- Switching diode 1N914
- Video Instructions

Dual Frequency
1KHz 30KHz
www.FrequencyGeneration.com

www.PositiveOffset.com

VIEW CART

resistors, (2) .0047 uF Capacitors, Mini Red LED, (9) Color Coded Jump Wires, 9-Volt Battery Snap, (2) Alligator leads, and an On/Off Switch. All this comes in a sturdy little cardboard box you can use to house your finished circuit. The full-color instructions will help you build either a 30,000 Hz or 1,000 Hz circuit.

It's so easy to build. No soldering required! Watch the video instructions right here (36 minutes).

www.PositiveOffset.com
Breadboard
555 CMOS Timer
Jump Wires (9)
1 long red, 2 white, 1 gray, 2 red,
1 blue, 1 orange, 1 green
On/Off Switch
9-volt Battery Clip
.0047 uF Capacitors
Two 6" Wires 1 red, 1 black
1/4 Watt Resistors (7)
1K
2.2K
4.7K
39K
270K
Red LED
Alligator Leads (2)
Instructions
23:54
HD
vimeo

- YouTube video instructions: 9 easy steps to build this circuit
- Build a zapper Frequently Asked Questions and Answers
- Already have electronic parts? Download PDF Instructions

Remember the Hulda Clark Zapper has not been licensed as a medical device by the US FDA. Zapper Kit is sold as an educational tool only, no medical claims are intended by the sale of these electronic parts. Our Zapper Kit is a great project for a school science fair, or to learn the basics of building simple electronics.

Please note that when ordering electronic parts and kits we can assume no liability on the customer's ability to use them. Safety, craftsmanship, experience, and ability are the user's responsibility.

Electronic parts are small and fragile. We can not offer refunds or replacements on electronic parts. Buy electronic parts and kits only if you believe you can accomplish the work. If you experience difficulty with your build, or have little experience in electronics, please order replacement parts along with your kit.

The 555 CMOS Timer is the most likely part to need replacement, so it may be smart to order extras when building a kit.

The 555 CMOS Timer is vulnerable to static electricity. If you walk on a carpeted floor and the 555 receives a static shock it may burn out and need replacement.

555 CMOS Timers are sold separately for those who want to have several at hand during a kit build, or to troubleshoot a circuit that is not working.

If you are new to electronics, or just want to be cost effective, you may want to purchase extra 555 CMOS Timers when building your first kit. We can not offer refunds or replacements on electronic parts.

View Shopping Cart / Checkout

View Shopping Cart / Checkout

Customer Service Information

We accept all major credit cards via our secure shopping cart. We also accept checks and money orders through the mail. All packages are shipped United States Postal Service Priority Mail with Delivery Confirmation. If you have an existing order and would like to track your package's Delivery Confirmation, click: **USPS Delivery Confirmation**. If you have any other question or concern about your order, please send an email to Customer Service.

Natural Health Supply .com
CELEBRATING 17 Years 2001-2018

www.PositiveOffset.com
Electronic Kit & Parts for Educators, Researchers & Hobbyists

Hulda Clark 30 kHz / 1 kHz Zapper Kit
with Color Illustrated Instructions
Order Zapper Kit Here

Hulda Clark Zapper Kit includes: Breadboard, 555 CMOS Timer, 7 resistors: (2) 1K Ohm, 2.2K Ohm, 4.7K Ohm, 39K, (2) 270K Ohm, (2) .0047 uF Capacitors, Red LED, (8) Color Jump Wires, (2) Alligator Leads, and Soldered Battery Snap/On-Off Switch. All this comes in a sturdy little cardboard box you can use to house your finished circuit. The full-color instructions will help you build either a 30,000 Hz or 1,000 Hz circuit. Kit does not include copper pipes. Click here to order copper pipes.

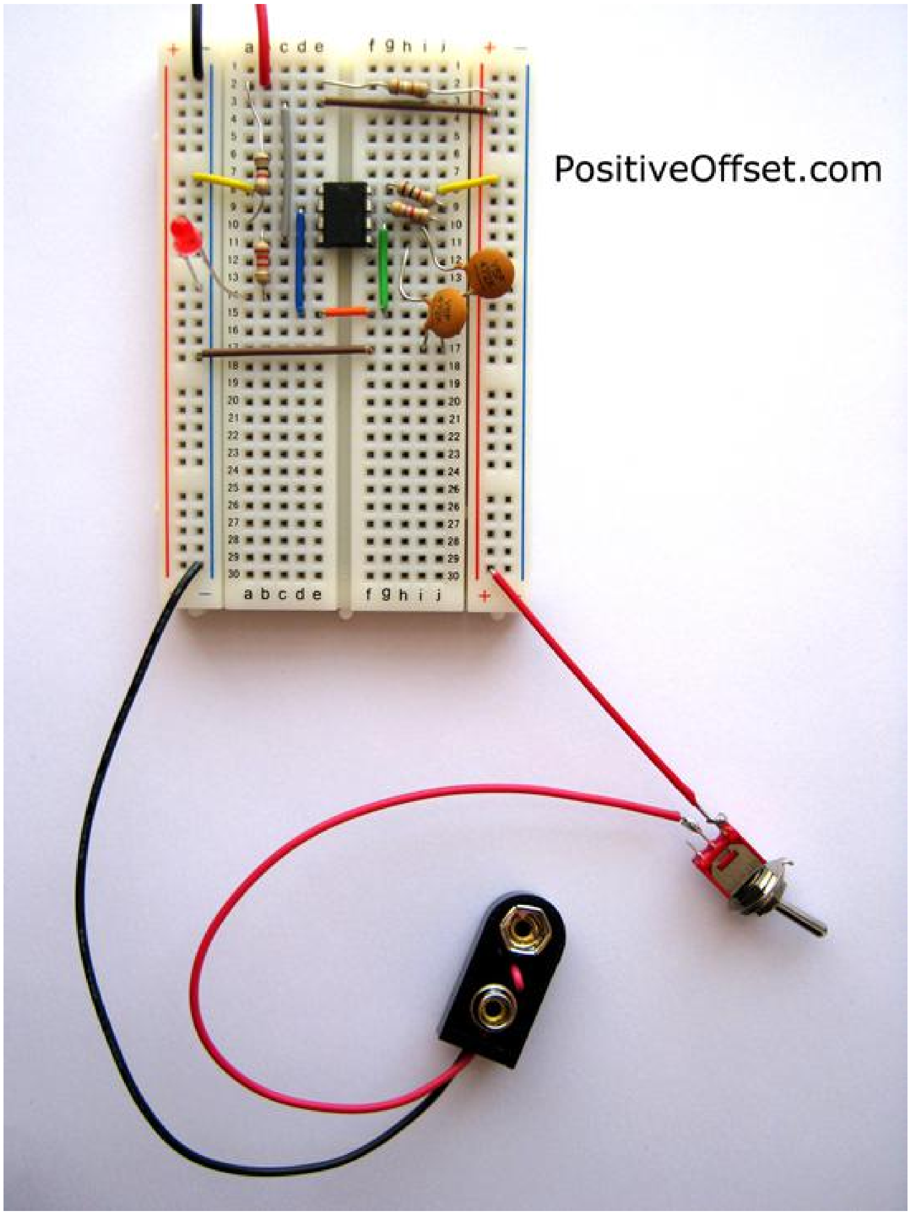

Hulda Clark Zapper Kit with Color Illustrated Instructions
Order Zapper Kit Here

If you are new to building electronics, you may also want to watch the instruction video. The 'Build A Zapper Video' shows how to build this Clark Zapper breadboard kit. Watch the video below. (Note: There are minor updates to the "Build Your Own Zapper" Kit. We have adjusted the jump wires for a

www.PositiveOffset.com

- 30 kHz/1 kHz Zapper Kit
- 555 CMOS Timer
- Switching diode 1N914
- Mini Alligator Leads
- Copper Pipes
- Dial-A-Zap
- Chair Zappicator
- Pet Zappicator
- Video Instructions

www.ClarkZapper.com

VIEW CART

better fit. We have also pre-soldered the 9-volt battery snap to the on/off switch. These updates are not shown in this video, but are clearly described in the kit instructions.)

It's so easy to build. No soldering required! Watch the video instructions right here (36 minutes).

More Video Instructions

- Build a Zapper, Plate Zapper, Food Zappicator, and Toothbrush Zappicator
- Zapper Kit Frequently Asked Questions and Answers
- Already have electronic parts? Download PDF Instructions

The Hulda Clark Zapper has not been licensed as a medical device by the US FDA. Zapper Kit is sold as an educational tool only, no medical claims are intended by the sale of these electronic parts. Our Zapper Kit is a great project for a school science fair, or to learn the basics of building simple electronics.

Please note that when ordering electronic parts and kits we can assume no liability on the customer's ability to use them. Safety, craftsmanship, experience, and ability are the user's responsibility.

Electronic parts are small and fragile. We can not offer refunds or replacements on electronic parts. Buy electronic parts and kits only if you believe you can accomplish the work.

The 555 CMOS Timer is the most likely part to need replacement, so it may be smart to order extras when building a kit.

The 555 CMOS Timer is vulnerable to static electricity. If you walk on a carpeted floor and the 555 receives a static shock it may burn out and need replacement. If you ever touch the 9 volt battery backwards to the battery snap, it may burn out the 555. You may want to protect your 555 by adding a 1N914 Switching Diode to your kit (diode sold separately).

555 CMOS Timers are sold separately for those who want to have several at hand during a kit build, or to troubleshoot a circuit that is not working. If you are new to electronics, or just want to be cost effective, you may want to purchase extra 555 CMOS Timers when building your first kit. We can not offer refunds or replacements on electronic parts.

View Shopping Cart / Checkout

Customer Service Information

We accept all major credit cards via our secure shopping cart. We also accept checks and money orders through the mail. All packages are shipped United States Postal Service Priority Mail with Delivery Confirmation. If you have an existing order and would like to track your package's Delivery Confirmation, click: **USPS Delivery Confirmation**. If you have any other question or concern about your order, please send an email to Customer Service.

www.ZapperLab.com

Order A Hulda Clark Zapper - **Order Dr. Clark's Books** - **Order Electrical Supplies**

Welcome to the Zapper Lab. Here we have the step-by-step instructions for building your very own Hulda Clark Zapper, reprinted with permission from the books by Dr. Hulda Clark. Dr. Clark has improved her original zapper circuit and published this update in 2003. Knowing her readers have varying experience with building electronics, this recent update released several different sets of instructions on how to build a zapper, depending on one's electronic expertise.

- The original published Zapper instructions require soldering or twisting the parts together. It is best when soldered, which does require some special tools and skill. These instructions are found reprinted below from Dr. Clark's 2003 update.
- A newer, easier to build design employs the use of an electronic *breadboard*. A breadboard does not require soldering, as one can push the electronic components into place onto the breadboard to easily build a circuit. Dr. Clark published designs for the 30 kHz Zapper, and the 1000 Hz Zappicator to be built on a breadboard. The instructions for building both breadboard zapper circuits are at ClarkZapper.net along with a nifty on-line video to help guide you through the process.
- And last, for advanced electricians, Dr. Hulda Clark published a schematic for building the updated Zapper (see photo below).

If you do not know anyone who could put one together for you, you can order a pre-built Zapper at the web site, www.ClarkZapper.com. Remember, the Hulda Clark Zapper is not a licensed medical device. It has not been tested by the FDA or AMA for use on humans or animals. The Clark Zapper can be sold and used only as an experimental device, and no claim can be made as to a diagnosis, cure or treatment for any medical condition or disease. Dr. Clark's book states to not experiment with a zapper if you are either pregnant or wearing an electronic pacemaker. Please be sure to read one of Dr. Clark's books before using a Hulda Clark Zapper.

Order A Hulda Clark Zapper - **Order Dr. Clark's Books** - **Order Electrical Supplies**

How To Build A Hulda Clark Zapper

click here to print the 2003 schematic for building a zapper

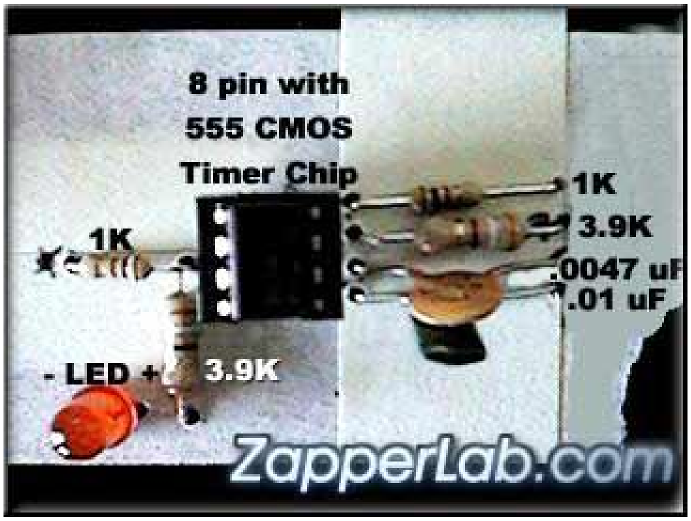

Order A Hulda Clark Zapper

Instructions for building your very own zapper can be found reprinted below from the books by Dr. Hulda Clark. You can get all the parts you need at online electrical supply stores such as: PositiveOffset.com. If you would rather order a pre-built zapper, just click: Clark Zapper.

Dr. Clark has released new devices called the *Food Zappicator* and the *Toothbrush Zappicator.* For the instructions on how to build this new type of zapper, visit: FoodZappicator.com and ToothZappicator.com

NEW! Build A Hulda Clark Zapper Video Instructions
Nine short videos that take you step-by-step through the process of building a Hulda Clark Zapper.

How To Build A Hulda Clark Zapper

***The Cure For All Cancers* © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.**
Zapper instructions have been updated for 2003 by Dr. Hulda Clark

Hints for absolute novices: Don't let unusual vocabulary deter you. A "lead" is just a piece of wire used to make connections. When you remove a component from its package, label it with a piece of tape. A serrated kitchen knife works best, as does a large safety pin. Practice using the micro clips. If the metal ends are L-shaped bend them into a U with the long-nose pliers so they grab better. Chips and chip holders (wire wrap sockets) are very fragile. It is wise to purchase an extra one of each in case you break the connections. The "555" timer is a widely used component; if you can't locate this one, try another electronics shop.

The Shoebox Way

This circuit has been improved since the one given in earlier books. A resistor has been added that gives every pulse an added Positive Offset of 1/4 volt. You no longer need to operate your zapper so daringly close to a Negative voltage.

Get from Radio Shack, (serial number included to make it easy). Order these parts from their web site at: www.RadioShack.com. If any electrical parts are unavailable at Radio Shack, visit another on-line electrical supplier of these items, such as: www.PositiveOffset.com to buy your electrical parts.

Parts List for the 30,000 Hz Zapper Circuit

Item	Radio Shack Catalog Number
Use Shoebox or Buy a Plastic Project Box	270-1809
9 volt battery	
9 volt battery snap connector	270-324 (set of 5 you need 1)
on-off toggle switch	275-624A micro mini toggle switch
If not available, choose any toggle switch with holes in the contact points or Radio Shack 275-612	
1 K Ohm resistor, brown-black-red-gold	271-312 (500 piece assortment) use 2
3.9 K Ohm resistor, orange-white-red-gold	Use 2 from 500 piece assortment
39 K Ohm resistor, orange-white-orange-gold	From 500 piece assortment
low-current red LED	276-044
.0047 uF capacitor	272-130 (set of 2, you need 1)
.01 uF capacitor	272-131 (set of 2, you need 1)
555 CMOS timer chip (TLC 555)	276-1718 (you may wish to buy a spare)
8 pin wire-wrapping socket for the chip	900-7242
If only 16 pin sockets are available, cut one in half OR leave half empty. Editor's Note: These parts have been discontinued at most Radio Shacks. 8 pin wire wrap sockets are available at many other electrical supply stores such as: www.PositiveOffset.com	
short (12") alligator clip leads	any electronics shop, get 10
If not available, use 14" length from Radio Shack, 278-1156	
Micro clip jumper wires	278-017 (you need 2 packages of 2)
If not available, use mini-clip jumper wires 278-016	
2 bolts, about 1/8" diameter, 2" long, with 4 nuts and 4 washers	hardware store
2 copper pipes, 3/4" diameter, 4" long	hardware store
Sharp knife, pin, long-nose pliers, tape, 4 twist ties or rubber bands.	

Editor's Note:

The 8 pin wire-wrapping socket has been discontinued at many Radio Shacks and readers have asked what to do. We know a great electrical supply store where 8 pin wire wrap sockets are available, as well as easy-to-build science kits for making a 30,000 Hz Frequency Generator with all the parts listed above in one simple package.

A Hulda Clark zapper is basically just a common Frequency Generator set at 30,000 Hz, with a 100% positive offset square wave. The cost for the electrical parts to build one is about $30-$40 for them all. Visit the electrical suppliers at PositiveOffset.com for all your electrical needs. They have everything you'd need to build a basic 30,000 Hz Positive Offset Square Wave Frequency Generator, and their kit is much easier to build than the instructions given below, because it uses a breadboard so there is no need to twist or solder wires, you simply push the electrical parts into place on the breadboard. **-editor.**

Assembling The Zapper

If you have tools such as a drill, needle nose pliers, and small drill bits, buy one of the plastic project boxes on the list, otherwise build your zapper in a shoe box, or a box half the size of a shoe box.

1. You will be using the lid of the shoe box or plastic lid of the project box to mount the components. Save the box to enclose the finished project.

2. Pierce two holes near the ends of the lid. Enlarge the holes with a pen or pencil until the bolts would fit through. Mount the bolts on the outside about half way through the holes so there is a washer and nut holding it in place on both sides. Tighten. Label one hole "grounding bolt" on the inside and outside.

3. Mount the 555 chip in the wire wrap socket. Find the "top end" of the chip by searching the outside surface carefully for a cookie-shaped bite or hole taken out of it. Align the chip with the socket and very gently squeeze the pins of the chip into the socket until they click in place.

4. Make 8 pinholes to fit the wire wrap socket. Enlarge them slightly with a sharp pencil. Mount it on the outside. Write in the numbers of the pins (connections) on both the outside and inside, starting with number one to the left of the "cookie bite" as seen from outside. After number 4, cross over to number 5 and continue. Number 8 will be across from number 1. The pins are numbered like this:

1	8
2	7
3	6
4	5

5. Pierce two holes ½ inch apart very near to pins 5,6,7, and 8. They should be less than 1/8 inch away. (Or, one end of each component can share a hole with the 555 chip.) Mount the .01 uF capacitor near pin 5 on the outside. On the inside connect pin 5 to one end of this capacitor by simply twisting them together. Loop the capacitor wire around the

pin first; then twist with the long-nose pliers until you have made a tight connection. Bend the other wire from the capacitor flat against the inside of the shoe box lid. Label it .01 on the outside and inside. Mount the .0047 uF capacitor near pin 6. On the inside twist the capacitor wire around the pin. Flatten the wire from the other end and label it .0047. Mount the 3.9 K Ohm resistor near pin 7, connecting it on the inside to the pin. Flatten the wire on the other end and label it 3.9. Mount the 1 K Ohm resistor and connect it similarly to pin 8 and label it 1K.

6. Pierce two holes ½ inch apart next to pin 3 (again, you can share the hole for pin 3 if you wish), in the direction of the bolt. Mount the other 1 K Ohm resistor and label inside and outside. Twist the connections together and flatten the remaining wire. This resistor protects the circuit if you should accidentally short the terminals. Mount the 3.9 K Ohm resistor downward. One end can go in the same hole as the 1K resistor near pin 3. Twist that end around pin 3 which already has the 1K resistor attached to it. Flatten the far end. Label.

7. Next to the 3.9 K Ohm resistor pierce two holes ¼ inch apart for the LED. Notice that the LED has a positive and a negative connection. The longer wire is the positive (anode). Mount the LED on the outside and bend back the wires, labeling them + and - on the inside.

8. Near the top pierce a hole for the toggle switch. Enlarge it until the shaft fits through from the inside. Remove nut and washer from switch before mounting. You may need to trim away some paper with a serrated knife before replacing washer and nut on the outside. Tighten.

9. Next to the switch pierce two holes for the wires from the battery holder and poke them through. Do not attach the battery yet.

10. An inch away from the switch pierce two holes 1/4 inch apart. Mount the 39 K Ohm resistor on the outside and label it inside and outside as "39 K, Positive offset." Flatten the wires on the inside.

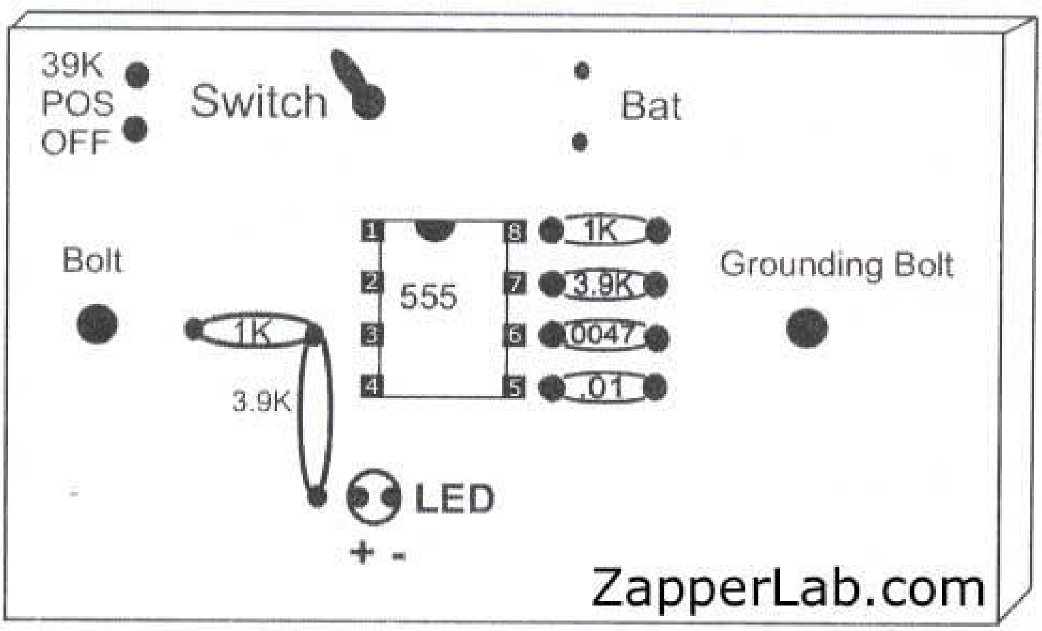

Now To Connect Everything

1. Twist free ends of the two capacitors .01 and .0047 together. Connect this to the Grounding Bolt using an alligator clip.

2. Bend the top ends of pin 2 and pin 6 (which already has a connection) inward towards each other in an L shape. Catch them both with an alligator clip and attach the other end of the alligator clip to the free end of the 3.9 K Ohm resistor by pin 7.

3. Using an alligator clip connect pin 7 to the free end of the 1 K Ohm resistor attached to pin 8.

4. Using three micro clips connect pin 8 to one end of the switch, pin 4 to the same end of the switch, and one end of the offset resistor to the same end of the switch. (Put one hook inside the hole and the other hooks around the whole connection. Check to make sure they are securely connected.) Connect the free end of the offset resistor to the bolt using an alligator clip.

5. Use an alligator clip to connect the free end of the 1 K Ohm resistor (by pin 3) to the bolt. It is the **output resistor**.

6. Twist the free end of the 3.9 K Ohm resistor by pin 3 around the plus end of the LED. Connect the minus end of the LED to the grounding bolt using an alligator clip.

7. Connect pin number 1 on the chip to the grounding bolt with an alligator clip.

8. Attach an alligator clip to the outside of one of the bolts. Attach the other end to a handhold (copper pipe). Do the same for the other bolt and handhold.

9. Connect the minus end of the battery snap connector (black wire) to the grounding bolt with an alligator clip.

10. Connect the plus end of the battery snap connector (red wire) to the free end of the switch using a micro clip lead. Attach the battery very carefully. Before attaching the battery to its snap connector, cover one terminal with tape. After snapping in one terminal, remove the tape to attach the other terminal. This is to prevent accidental touching of terminals in a backwards direction. If the LED lights up you know the switch is ON. If it does not, flip the switch and see if the LED lights. Label the switch clearly. If you cannot get the LED to light in either switch position, double-check all of your connections, and make sure you have a fresh battery. Even if it does light up, check every connection again.

11. Finally tie up the bunches of wire with twist-ties or rubber bands

and replace the lid on the box. Slip a couple of rubber bands around the box to keep it securely shut. For safer storage, place it inside a larger box.

Note: Having gained this much experience, you may prefer to build your next zapper on a piece of cardboard folded and able to fit inside a shoebox or plastic electronic project box for more protection.

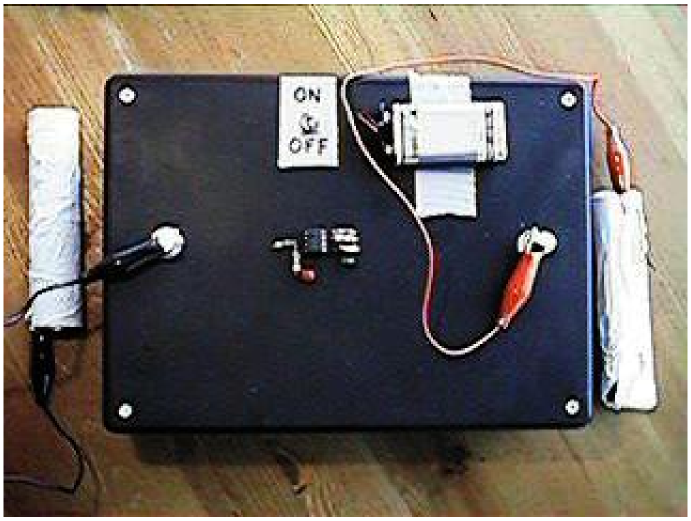

- Optional: measure the frequency of your zapper by connecting an oscilloscope or frequency counter to the handholds. Any electronics shop can do this. It should read between 20 and 40 kHz. The shop can also read the voltage (peak to peak) and the amount of Positive offset (on the .5 volt-per-division scale). The voltage output should be about 8 volts.
- Note: a voltage meter will only read 4 to 5 volts because it displays an average voltage.

Using The Zapper

1. Wrap handholds in one layer of wet paper towel before using. More will reduce the current. Grasp securely and turn the switch on to zap. Keep a bottle of water handy to keep them wet.

2. Zap for seven minutes, let go of the handholds, turn off the zapper, and rest for 20 minutes. Then seven minutes on, 20 minutes rest, and a final seven minutes on. This is the routine for **regular** zapping.

Tell us about your zapper experience at the zapper message board.

To Visit The Zapper Museum - Click Here!

Order A Hulda Clark Zapper - Click Here!

See The '*Build A Zapper*' Video - Click Here!

Learn about Dr Clark's latest zapping technique called: Plate Zapping

Learn about Dr Clark's latest inventions:
The Food Zappicator and The Toothbrush Zappicator

Order Electronic Parts at RadioShack.com and Frequency Generator Kits at PositiveOffset.com

All information is for educational purposes only. These statements have not been evaluated by the FDA. Not intended to diagnose, treat, prevent, mitigate or cure disease. Visit www.HuldaClark.com for more information on Dr. Clark's Syncrometer™ Science, and the Zapper. Legal Disclaimer Privacy Policy

www.HuldaClark.com

The Cure For All Cancers © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.

KEELYNET

Zapper Schematics
No claims are made by KeelyNet for any of these plans
They are provided for informational purposes only!
KeelyNet 01/10/02

Simple Zapper Schematic

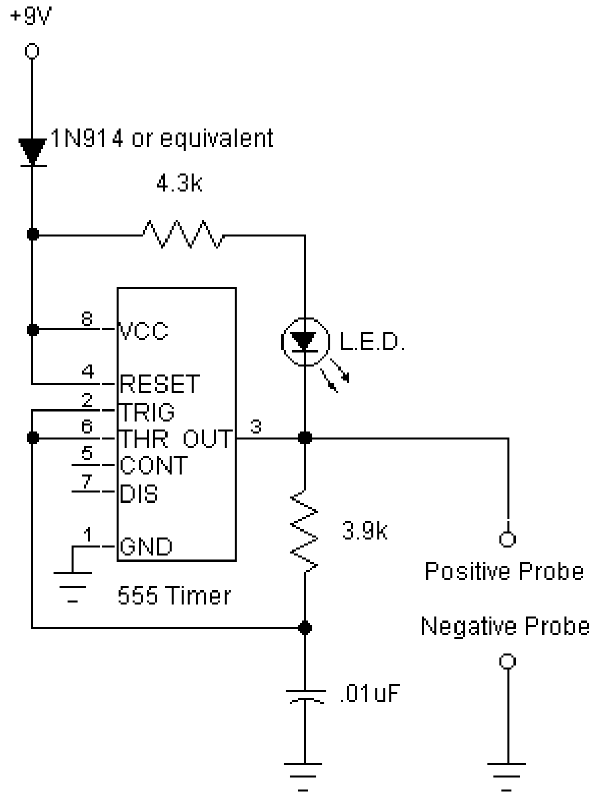
+9V
1N914 or equivalent
4.3k
8 VCC
4 RESET
2 TRIG
6 THR
OUT 3
5 CONT
7 DIS
1 GND
555 Timer
L.E.D.
3.9k
Positive Probe
Negative Probe
.01uF

The Research of Dr. Hulda Regehr Clark, Ph. D., N.D.

I have recently (9/15/2000) come across the research of Dr. Hulda Regehr Clark, Ph. D., N.D. In her books, she states that many diseases, among which are cancer and HIV/AIDS, can be cured by the use of, among other things, certain herbs and a frequency generator called a "Zapper" that was invented by her and her son.

Speaking of mothers :), my mother, 62, had fatigue and arthritic symptoms in her hands before I found out about and told her about Dr. Hulda Clark. She is a seamstress, so the arthritic symptoms hindered her work. After doing the Parasite Cleanses and using the Zapper, she doesn't have to take naps like before and her hands are flexible again. She has done the complete Parasite Cleanse only twice but she Zaps practically every day.

Simple Zapper Information

Schematic updated on 12/21/2001

Simple Zapper Information text updated on 12/21/2001.

I forgot to mention that the Simple Zapper that I'm using has a 1 Megohm potentiometer instead of the 2.2k ohm resistor in order to vary the output frequency. If you use a potentiometer with the circuit put it in series with the 2.2k ohm resistor because the chip gets hot when there is too little resistance.

The duty-cycle is pretty consistent with a range of 51% to 57%. With the oscilloscope's frequency analyzer in storage mode, I hold the Zapper probes and put the oscilloscope probes somewhere on my body and then I vary the frequency of the Zapper using the 1 Megohm potentiometer.

When testing my body on a spectrum analyzer while using the Simple Zapper with an output of around 5 Volts (tested on a DC voltmeter), I noticed that the frequencies around 4,000 hz to around 18,000 hz had larger peaks telling me that my body at the time was more conductive (less reactive?) and/or less resistive to frequencies in this range at the given voltage.

With my body being more conductive (less reactive?) at these frequencies, I believe that there is a better dispersal of the Simple Zapper's output within my body. I did this recent test with one Zapper probe in each hand (as opposed to both in one hand) and the oscilloscope probes (RCA to 1/4" phone plug adaptor) in my mouth near my wisdom teeth with my lips around the RCA ground (like a lollipop).

The oscilloscope's frequency analyzer was in storage mode. With the Simple Zapper circuit on this page, change the 2.2k ohm resistor to 4.3k ohm if you want a ~15,000 hz frequency. I have recently built a couple of Zappers with optimized output frequencies of around 11,000 hz. I haven't tested anyone using the Simple Zapper other than myself, so there is the possibility that your body's response may differ.

For more information, check out my Simple Zapper webpage at:

www.members.aol.com/mas1911/index.html

or for really interesting free circuit schematics:

www.members.tripod.com/~mas1911/zappers20011221tripod.html

You can write the Author of Simple Zapper if you have comments or questions.

The following document, written by Luke Parrish, provides instructions on building an astable multivibrator in DC pulse mode, identical in output to the Zapper, described in The Cure For All Diseases by Dr. Hulda Clark.

Plans For Zapper

This page tells how to make a Hulda Clark-style "Parasite Zapper". It's a simpler and more efficient version, but works just as well. We make no theraputic claims for this device. It is for experimental purposes only. We provide a ready made version for those who prefer not to spend so much time building it.

Materials

- **A CD4069 hex inverter (Radio Shack link)**
- **1 1M resistor, 1 10k resistor, and 1 1k resistor (Radio Shack link) (Resistor color code)**
- **1 1000 pF capacitor (Radio Shack link)**

NOTE: The size of the resistors and capacitors can vary, because they determine the frequency of the zapper, which doesn't really matter much according to Hulda Clark.

The following parts may be substituted according to preference, though something similar will be needed:
- **Solderless breadboard (Radio Shack link)**
- **9-volt battery**
- **4 alligator clip-leads sets Radio Shack link**
- **2 copper pipes (hand-holds)**

Instructions

Plug the components into the solderless breadboard according to the following schematic. Make sure none of the wires touch each other. To hook up the power, attach an alligator clip-lead to each side of the capacitor that stretches from hot to ground. Clip the one that goes to hot to the positive terminal on the battery and the one that goes to ground to the negative terminal. The two hand-holds are attached with the other two clip-leads. (One goes to ground, the other to the end of a resistor that goes to pin 12.)

Schematic

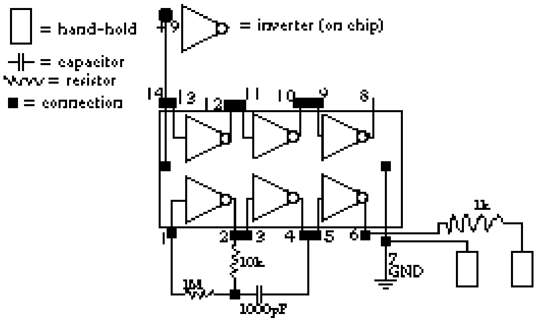

Schematic most recently updated 5-17-2001

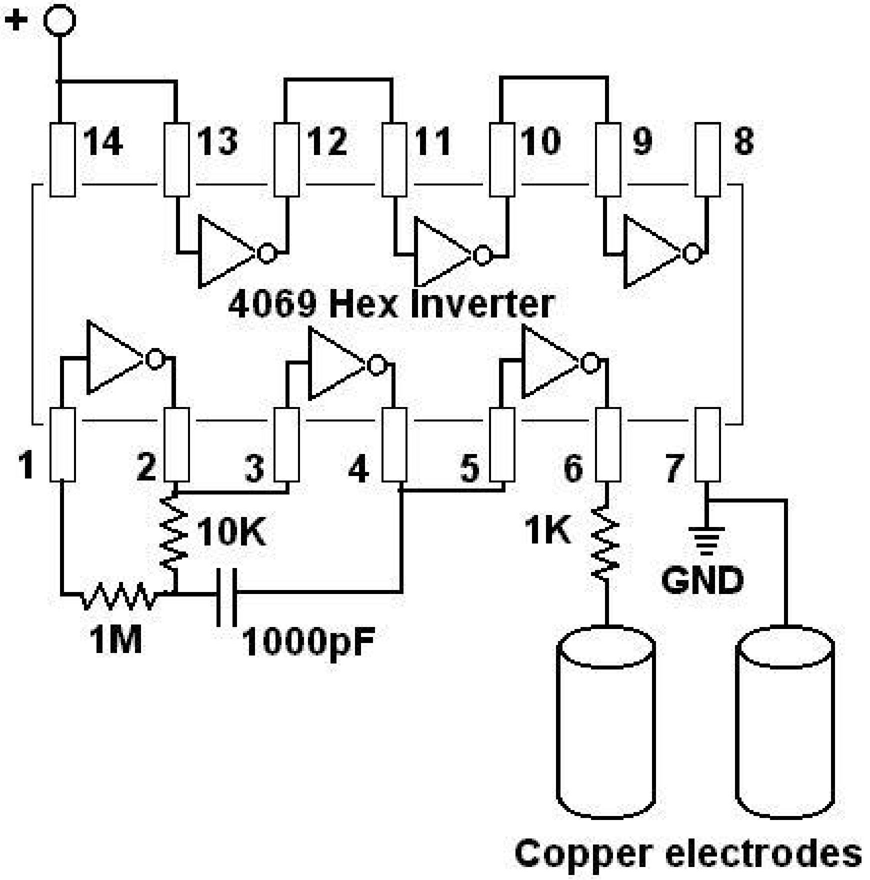

The above instructions are not the only way to do it. You can, for example, solder all the components directly to a 14-pin socket, like we do in our ready made version. This is a lot more trouble and more time-consuming for most people, which is why we recommend a breadboard. The resistors and capacitor don't all have to be the exact sizes mentioned here to make a working Zapper. However, any variation in capacitance or resistivity will change the frequency. (There are certain limits on how high of frequencies the chip can handle, so don't stray too far from the specified values.) If you have an oscilliscope, that's handy to check if it's working.

Ready made Zappers

Free plans from Zapperlab.com

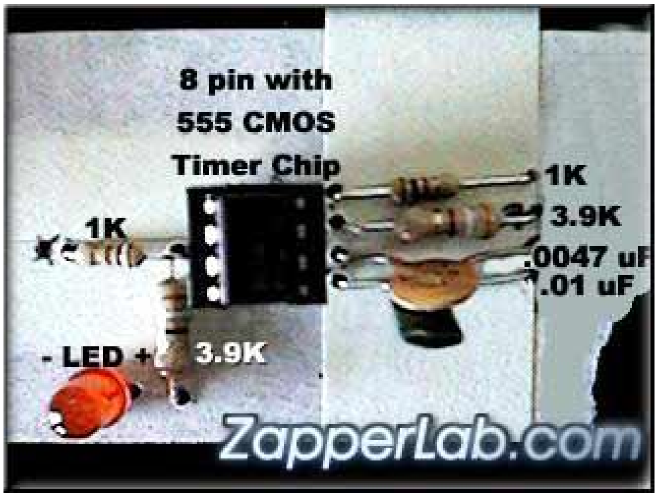

Here we have the easy step-by-step instructions for building your very own Hulda Clark Zapper reprinted with permission from her books (provided below). For advanced electricians there is a schematic for building a Hulda Clark Zapper as well. If you do not know anyone that could put one together for you, you can now order the zapper Hulda Clark uses as described in "The Cure For All Cancers" and "The Cure For All Diseases" at: www.SyncroZap.com. The Dr. Clark Zapper can be sold only as an experimental device, and no claim can be made by us legally as to a diagnosis, cure or treatment for any medical condition or disease.

Build A Dr. Clark Zapper

Order A Dr. Clark Zapper

Instructions for building your very own zapper can be found reprinted below from the books by Dr. Hulda Clark. If you would rather order a pre-built zapper, just click here.

How To Build A Zapper from *The Cure For All Cancers*:

The Cure For All Cancers © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.

Get from Radio Shack, (serial number included to make it easy).
Order these parts from their web site at:
www.RadioShack.com

___ Black plastic Project Box Radio Shack # 270-1809

___ 9 volt battery

___ 9 volt battery clips: Radio Shack # 270-325

___ On off switch: Radio Shack # 275-624A

___ 1K Ohm resistor: RS# 271-1321

___ 3.9K Ohm resistor: RS# 271-1123

___ low current red LED: RS# 276-044

___ .0047 uF capacitor: RS# 272-130

___ .01 uF capacitor: RS# 272-1065

___ 555 CMOS timer chip: RS# 276-1723

___ 8 pin wire wrapping socket for CMOS chip: RS# 276-1988

___ two packs of Microclip test jumpers: RS# 278-017

___ one pack of 14" alligator clip leads RS# 278-1156C

Get from any Hardware Store:

2 Bolts 1/8" diameter, 2" long with 4 nuts and washers.

2 pieces of ¾" copper pipe, cut to 4 inches long each.

Editor's Note:

Remove 8 Pins From A 16 Pin Wire Wrap Socket

The **8 pin wire-wrapping socket** has been discontinued at many Radio Shacks and readers have asked what to do. Some say it is still available on their website at www.RadioShack.com, if not:

There are a couple options. One can still get the **16 pin wire-**

wrapping socket. Pull out the bottom grouping of 8 pins, and you now have an 8 pin socket. Just make sure when you plug the 555 Timer into it, you line it up with the remaining 8 pins. Very simple.

There is also a **Low Profile 8 pin socket**. The pins are so small one must solder to them.

The wire-wrap socket has long pins for easy twisting onto other wires. The low profile is for those that solder electronics.

If you are soldering your zapper together get the low profile 8 pin socket. Since the socket is just a saddle for the 555 Timer, it is not a vital part of the circuit. It's just a way to put long or short pins on the 555 Timer for easy twisting. If you cannot find one at your local Radio Shack, simply get the 16 pin wire-wrap socket, and pull out 8 pins. -editor.

Assembling The Zapper

If you have tools such as a drill, needle nose pliers, and small drill bits, buy one of the plastic project boxes on the list, otherwise build your zapper in a shoe box, or a box half the size of a shoe box.

1. You will be using the lid of the shoe box or plastic lid of the project box to mount the components. Save the box to enclose the finished project.

2. Pierce two holes near the ends of the lid. Enlarge the holes with a pen or pencil until the bolts would fit through. Mount the bolts on the outside about half way through the holes so there is a washer and nut holding it in place on both sides. Tighten. Label one hole "grounding bolt" on the inside and outside.

3. Mount the 555 chip in the wire wrap socket. Find the "top end" of the chip by searching the outside surface carefully for a cookie-shaped bite or hole taken out of it. Align the chip with the socket and very gently squeeze the pins of the chip into the socket until they click in place.

4. Make 8 pinholes to fit the wire wrap socket. Enlarge them slightly with a sharp pencil. Mount it on the outside. Write in the numbers of the pins (connections) on both the outside and inside, starting with number one to the left of the "cookie bite" as seen from outside. After number 4, cross over to number 5 and continue. Number 8 will be across from number 1. The pins are numbered like this:

1 8

2 7

3	6
4	5

5. Pierce two holes ½ inch apart very near to pins 5,6,7, and 8. They should be less than 1/8 inch away. (Or, one end of each component can share a hole with the 555 chip.) Mount the .01 uF capacitor near pin 5 on the outside. On the inside connect pin 5 to one end of this capacitor by simply twisting them together. Loop the capacitor wire around the pin first; then twist with the long-nose pliers until you have made a tight connection. Bend the other wire from the capacitor flat against the inside of the shoe box lid. Label it .01 on the outside and inside. Mount the .0047 uF capacitor near pin 6. On the inside twist the capacitor wire around the pin. Flatten the wire from the other end and label it .0047. Mount the 3.9 K Ohm resistor near pin 7, connecting it on the inside to the pin. Flatten the wire on the other end and label it 3.9. Mount the 1 K Ohm resistor and connect it similarly to pin 8 and label it 1K.

6. Pierce two holes ½ inch apart next to pin 3 (again, you can share the hole for pin 3 if you wish), in the direction of the bolt. Mount the other 1 K Ohm resistor and label inside and outside. Twist the connections together and flatten the remaining wire. This resistor protects the circuit if you should accidentally short the terminals. Mount the 3.9 K Ohm resistor downward. One end can go in the same hole as the 1K resistor near pin 3. Twist that end around pin 3 which already has the 1K resistor attached to it. Flatten the far end. Label.

7. Next to the 3.9 K Ohm resistor pierce two holes ¼ inch apart for the LED. Notice that the LED has a positive and a negative connection. The longer wire is the positive (anode). Mount the LED on the outside and bend back the wires, labeling them + and - on the inside.

8. Near the top pierce a hole for the toggle switch. Enlarge it until the shaft fits through from the inside. Remove nut and washer from switch before mounting. You may need to trim away some paper with a serrated knife before replacing washer and nut on the outside. Tighten.

9. Next to the switch pierce two holes for the wires from the battery holder and poke them through. Attach the battery and tape it to the outside.

Now To Connect Everything

Twist free ends of the two capacitors .01 and .0047 together. Connect to Grounding Bolt using an alligator clip.

Bend pin 2 and pin 6 together inward, using an alligator clip, catch them and connect to free end of 3.9K Ohm (by pin 7).

Alligator clip Pin 7 to free end 1K Ohm (near pin 8)

Using two microclips Pin 8 and pin 4 to one end of switch (use hole to attach both microclips).

Free end of 1K Ohm (by pin 3) to the Bolt using an alligator clip.

Alligator clip the free end of 3.9K Ohm (by pin 3) to plus end (long) of LED.

Minus end (short) of LED to Grounding Bolt with an alligator clip.

Alligator clip pin 1 to the Grounding Bolt.

Alligator clip Black battery wire to Grounding Bolt.

Micro clip the Red battery wire to free end of switch.

Hook up the battery, and the light should turn on, click the switch if it does not, check connections if LED does not light up.

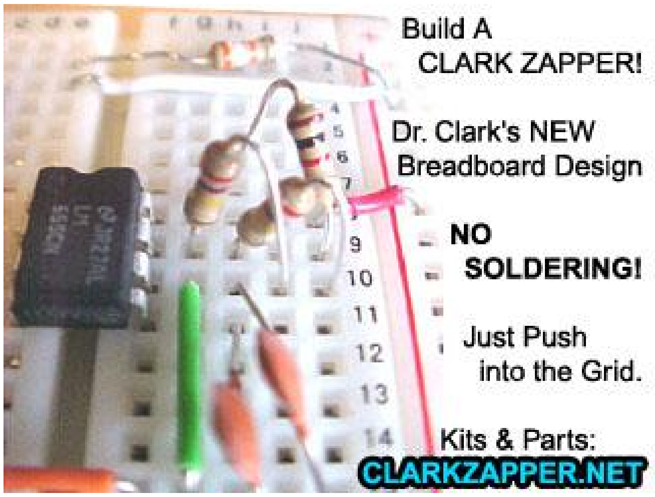
Build A
CLARK ZAPPER!
Dr. Clark's NEW
Breadboard Design
NO
SOLDERING!
Just Push
into the Grid.
Kits & Parts:
CLARKZAPPER.NET
4
5
6
7
9
10
11
12
13
14

Attach a long lead wire to the grounding bolt & bolt, and then to the copper handles. Wrap handles in paper towel, clip with lead wire, wet handle when you are ready to zap.)

How To Zap

Wrap handles in wet paper towel, or wet cotton sleeves, and hold in each hand. If you have wrist cuffs, run them under water to wet them, attach to each wrist.

Attach handles or wrist cuffs to zapper.

Turn zapper on, zapping for 7 minutes.

At end of 7 minute zapping rest 20 minutes.

Do this a minimum of three 7 minute zapping sessions, with a 20 minute break in between each session. Some people zap five times in extreme cases.

Order A Dr. Hulda Clark Zapper - Click Here!

Learn about Dr Clark's latest zapping technique at: www.PlateZapping.com

All information is for educational purposes only. These statements have not been evaluated by the FDA. Not intended to diagnose, treat, prevent, mitigate or cure disease.

This Circuit designed by Geoff Clark and approved by Dr. Hulda Clark

SOTA Instruments Inc. presents...
"The Zapper" as per Dr. Hulda Clark, Ph.D., N.D.

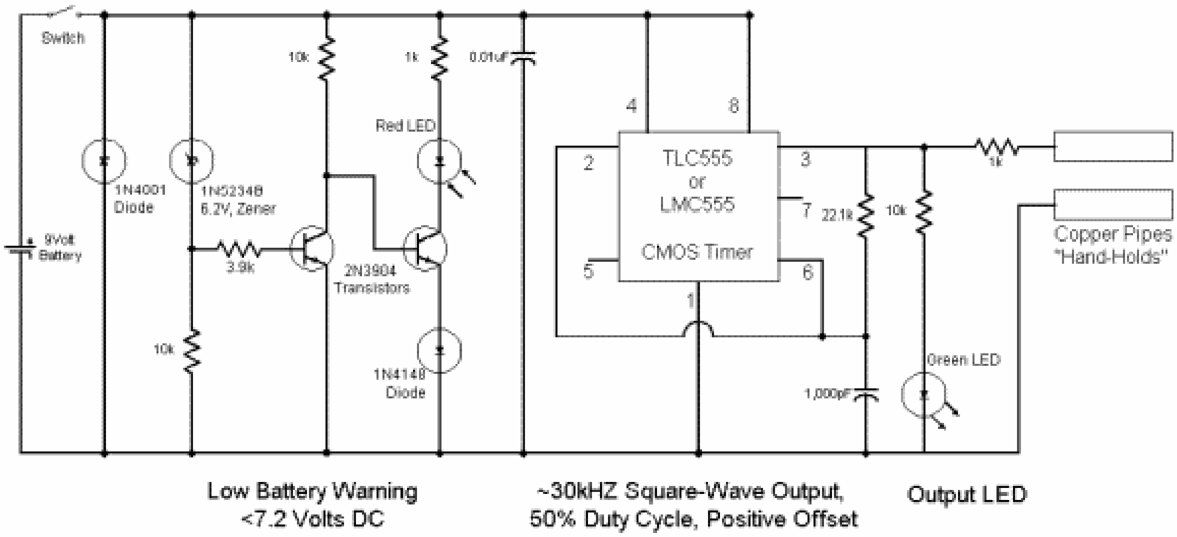
Switch
10k
1k
0.01uF
4
8
Red LED
2
TLC555
or
LMC555
CMOS Timer
3
1k
7
22.1k
10k
1N4001
Diode
1N5234B
6.2V, Zener
9Volt
Battery
3.9k
2N3904
Transistors
5
6
1
Copper Pipes
"Hand-Holds"
10k
1N4148
Diode
Green LED
1,000pF
Low Battery Warning
<7.2 Volts DC
~30kHZ Square-Wave Output,
50% Duty Cycle, Positive Offset
Output LED

Certified and Approved by Dr. Clark's son Geoff Clark.
Designed & Drawn by Russell J. Torlage, CTech, President, SOTA Instruments Inc.
Model: ZHC3 Rev 4.0
Revised: June 9, 1999

All Resitors are 1/4W, 5% Tol. 22.1k resistor is 1%.

Copyright 2000 SOTA Instruments Inc.
This schematic is not to be used for commercial applications without the express written permission of SOTA Instruments Inc.

If you found this file useful or interesting, please consider a donation or a purchase to help keep KeelyNet online and providing free information. Even a dollar will help. Others sell it, we prefer to share it, thanks!

PayPal DONATE

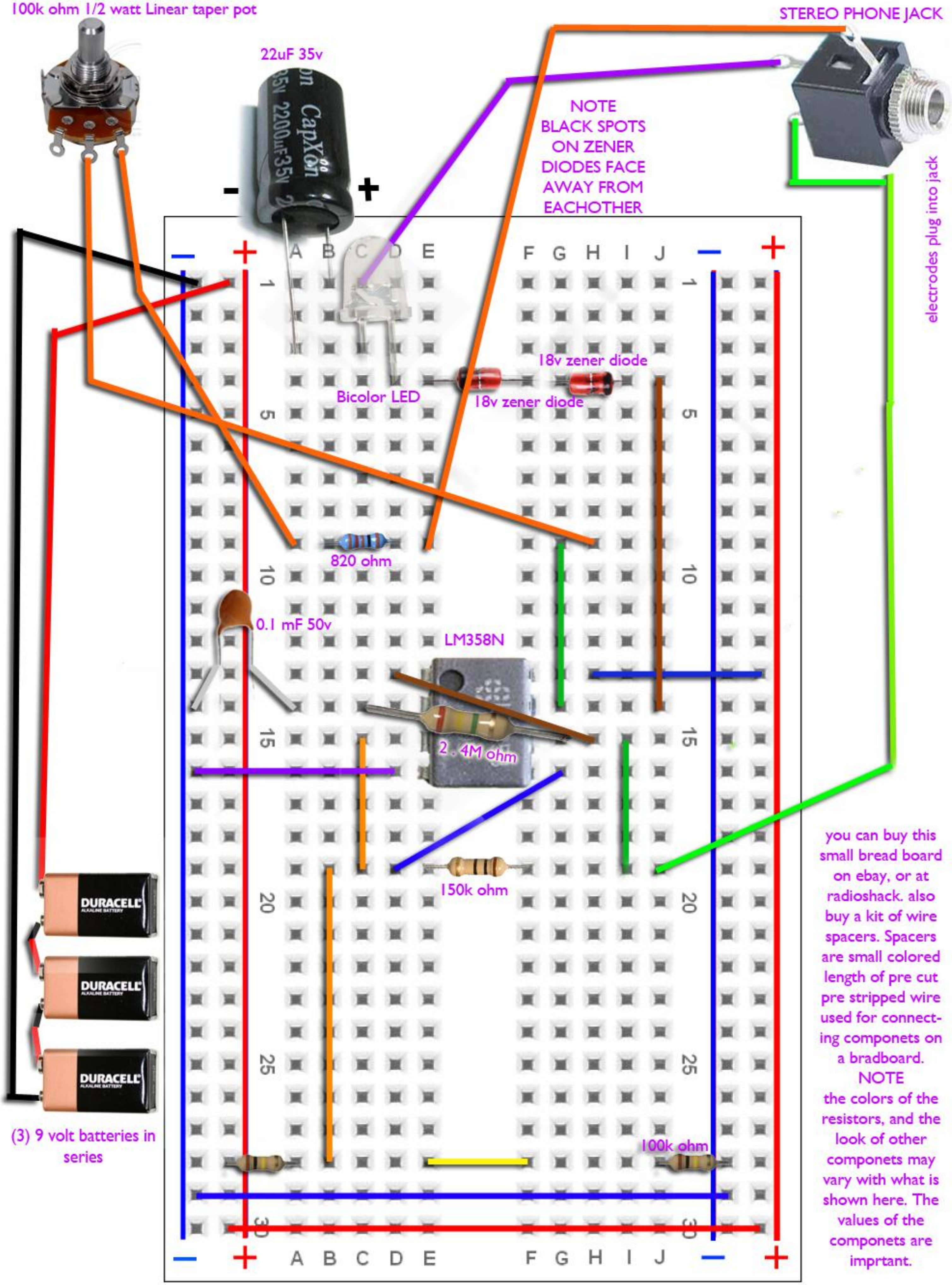
100k ohm 1/2 watt Linear taper pot
22uF 35v
-
+
STEREO PHONE JACK
NOTE
BLACK SPOTS
ON ZENER
DIODES FACE
AWAY FROM
EACHOTHER
electrodes plug into jack
A B C D E F G H I J
1
5
10
15
20
25
30
Bicolor LED
18v zener diode
18v zener diode
820 ohm
0.1 mF 50v
LM358N
2 . 4M ohm
150k ohm
100k ohm
DURACELL
DURACELL
DURACELL
(3) 9 volt batteries in series

you can buy this small bread board on ebay, or at radioshack. also buy a kit of wire spacers. Spacers are small colored length of pre cut pre stripped wire used for connecting componets on a bradboard.

NOTE

the colors of the resistors, and the look of other componets may vary with what is shown here. The values of the componets are imprtant.

COMPONENT DESCRIPTIONS, SOURCES, AND CURRENT PRICES

Revision November 24, 1996. Copyright (c) 1991/1999 Robert C. Beck, D.Sc.

Note: These data are for theoretical, informational and instructional purposes only and are not to be construed as medical advice. Consult with your licensed medical practitioner. Hundreds have been built successfully if duplicated exactly as shown.

LM358 CMOS Operational Amplifier (OP-AMP) chip (generic)		U1	$ 1.00
2.4 megohm 1/4 watt 5% resistor		R1	0.07
150 kohm 1/4 watt 5% resistor		R2	0.07
100 k ohm 1/4w 5% resistor	2 required @ 0.15	R3, R4	0.14
100 kohm linear taper pot, 1/2 watt Cattronics P-68		R5	2.56
150 ohm 1/4 watt 5% resistor		R6	0.07
820 ohm 1/4 watt 5% resistor		R7	0.07
0.1 mF 50 V (or higher) ceramic capacitor		C1	0.45
22 mF 35 V (or higher) electrolytic capacitor		C2	0.25
18 Volt Zener diodes (NTE5027A)	% Watt, 2 @ 0.79	D1 & D2	1.58
Bulb, 6.3V 0.075 A	Type 7377	B1	1.34
Bicolor LED Red/Green	RS# 276-012	LED 1	$ 1.19
Jack for electrode leads (3.5mm or 1/8" mono-phone jack)	RS# 274-251 3/$1.59	J1	0.53
Jack for silver wire leads (2.5mm or 3/32" mono-phone jack)	RS# 274-292 3/$1.59	J2	0.53
3 Alkaline 9 V batteries, type 1604 etc.	3 for		3.75
3 x 9 V battery snaps (clip-on connectors)	RS# 270-325 pkg.	5/$1.29	1.29
Misc. wire, solder, etc.			0.50
Box, if used			2.29
Lead wire with 3.5 mm plug, 6 ft., Mouser or Calrad Electronics			0.35
Electrodes, stretch elastic, Velcro, cotton flannel, alligator clips, etc.			
Total Cost **for all** components for improved **do-it-yourself** project			**$22.47**

All components listed above check your local electronics store.

Ensure the IC chip U 1 (LM358) is wired correctly. The location of Pin #1 is shown below:

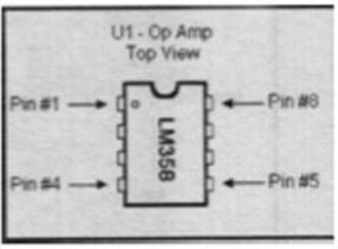

NOTE
RESISTOR
COLORS NOT
TO SPEC.
CHECK
RESISTOR
COLOR VALUES

electrodes:
3/32 x 1
stainless steel cotter pins -hardware store-
Snap pins in half across the curve. and wrap the headphone wires comming from the mono plug around the curve of each half of the pin. then hou glue into place.

Written by
MARK HYMAN, MD

Can Autism be Cured?

TODAY MOST PEOPLE BELIEVE that Autism is a genetic brain disorder. I'm here to tell you that this isn't true. The real reason we are seeing increasing rates of autism is simply this: Autism is a systemic *body disorder* that affects the brain. A toxic environment triggers certain genes in people susceptible to this condition. And research supports this position.

Think about it. Rates of autism have skyrocketed over the years, from an estimated 1 child in 3,000 to just 1 in 150 kids today. Sure, wider criteria for diagnosis and better detection might explain some of it but not an increase of this magnitude.

Dramatic scientific discoveries have taken place during the last 10 to 20 years that reveal the true causes of autism, and turn conventional thinking on its head. For example, Martha Herbert, MD, a pediatric neurologist from Harvard Medical School has painted a picture of autism that shows how core abnormalities in body systems like immunity, gut function, and detoxification play a central role in causing the behavioral and mood symptoms of autism.

Recently I treated a 2 ½ year old boy named Sam. He was born healthy but diagnosed with autism after his vaccination for measles, mumps and rubella at 22 months.

We use cookies and similar technologies to run this website and help us understand how you use it. See our privacy policy here.

When I first saw him, this little boy was deep in the inner wordless world of autism. Watching him was like watching someone on a psychedelic drug trip. So we dug into his biochemistry and genetics and found many things to account for the problems he was having.

He had very high level of antibodies to gluten. He was allergic not only to wheat, but to dairy, eggs, yeast, and soy — about 28 foods in total. He also had a leaky gut, and his gut was very inflamed. Sam was deficient in zinc, magnesium, and manganese, vitamins A, B12, and D, and omega-3 fats. Like many children with autism, he had trouble making energy in his cells, or mitochondria.

His amino acids — necessary for normal brain function and detoxification — were depleted. And his blood showed high levels of aluminum and lead, while his hair showed very high levels of antimony and arsenic, signs of a very toxic little boy. His levels of sulfur and glutathione were low, indicating that he just couldn't muster the power to detox all these metals. In fact, his genes showed a major weak spot in glutathione metabolism, which is the body's main antioxidant and major detoxification highway for getting rid of metals and pesticides.

Sam also had trouble with a key biochemical function called methylation that is required to make normal neurotransmitters and brain chemicals and is critical for helping the body get rid of toxins. This showed up as low levels of homocysteine (signs of problems with folate metabolism) and high methylmalonic acid (signs of problems with B12 metabolism). He also had two genes that set him up for more problems with this system.

Finally, he also had very high levels of oxidative stress or free radical activity, including markers that told me that his brain was inflamed and under free-radical fire.

This may all seem complicated, but it really isn't. When I see any patient, I simply work through the 7 Keys to UltraWellness (based on functional medicine) to see how everything is connected, create a plan to get to the causes of the problems, and then

We use cookies and similar technologies to run this website and help us understand how you use it. See our privacy policy here.

To create a roadmap for recovery you just take away what's bothering the patient and give his body what it is missing and needs to thrive (based on the individual's biochemical uniqueness). Then the body does the rest. Here is the roadmap I used to help Sam recover.

Step 1: Fix His Gut and Cool the Inflammation There

This step included a number of different tactics including:

- Taking away gluten and other food allergens
- Getting rid of his yeast with anti-fungals
- Killing off the toxic bacteria in his small intestine with special antibiotics
- Replenishing healthy bacteria with probiotics
- Helping him digest his food with enzymes

Step 2: Replace the Missing Nutrients to Help His Genes Work Better

In Sam's case we:

- Added back zinc, magnesium, folate, and vitamins A, B6, B12, and D
- Supported his brain with omega-3 fats

Step 3: Detoxify and Reduce Oxidative Stress

- Once his biochemistry and nutrition was tuned up, we helped him detoxify and

We use cookies and similar technologies to run this website and help us understand how you use it. See our privacy policy here.

As I said before, the keys of UltraWellness can help, no matter what the disease or condition. Biology has basic laws, which we have to follow and understand. All the details of Sam's story fit into these laws. We just have to dig deep, peel back the layers, and understand what is going on. When we do this the results are nothing short of miraculous…

After following a gluten-free diet and treating his gut for 3 weeks, Sam showed dramatic and remarkable improvement. He was getting back much of his language skills and showing much more connection and relatedness in his interactions.

After 4 months, he was more focused, unstuck and verbal.

After 10 months, his bowels were back to normal, he was verbally fluent, mainstreamed in school and he "lost" his diagnosis of autism.

After 2 years all his abnormal tests were normal including the high metals, gut inflammation and damage to his mitochondria and free radicals.

And more importantly, the child was totally normal. Not every child has such a dramatic recovery but many improve, and some improve dramatically using the approach of functional or systems medicine.

This is just the beginning. Read the full article on Why Current Thinking About Autism Is Completely Wrong.

To your good health,

Mark Hyman, MD

We use cookies and similar technologies to run this website and help us understand how you use it. See our privacy policy here.

Mark Hyman MD is the Director of Cleveland Clinic's Center for Functional Medicine, the Founder of The UltraWellness Center, and a ten-time #1 New York Times Bestselling author.

If you are looking for personalized medical support, we highly recommend contacting Dr. Hyman's UltraWellness Center in Lenox, Massachusetts today.

SHARE ON FACEBOOK

SHARE VIA EMAIL

PROGRAMS

Eat Fat, Get Thin

10-Day Detox Diet

Blood Sugar Solution

Testimonials

We use cookies and similar technologies to run this website and help us understand how you use it. See our privacy policy here.

The Broken Brain Podcast

FOLLOW DR. HYMAN

Facebook
Instagram
Twitter
Pinterest
YouTube
Community

ABOUT DR. HYMAN

Bio
Contact
The UltraWellness Center
Cleveland Clinic’s Center for Functional Medicine
Institute for Functional Medicine

DR. HYMAN

© 2016 Hyman Digital, LLC. Terms of Use | Privacy Policy | Shipping Policy | Return Policy | Disclaimer | FAQ
Statements on this site have not been evaluated by the Food and Drug Administration

We use cookies and similar technologies to run this website and help us understand how you use it. See our privacy policy here.

Healing the symptoms known as Autism

Second
Edition

PDF
E-Book

Kerri Rivera

with Kimberly McDaniel & Daniel Bender

Jim Humble • Dr. Andreas Kalcker
Dr. Marco Ruggiero • Robert L. Sands

Table of Contents

Foreword by Lorna B. Ortiz, PhD.
President of Curando el Autismo, CEA vii

Preface by Kimberly McDaniel ix

Terminology & Units of Measure xiii

Acknowledgements xv

Introduction: Autism is Avoidable, Treatable and Curable xxi

Chapter 1: Kerri's Story 1

Chapter 2: Yes We Can!!! (Testimonials) 19

Chapter 3: Step 1 – The Diet 41

Chapter 4: An Introduction to Chlorine Dioxide 77

Chapter 5: Step 2 – Chlorine Dioxide (CD) 81

Chapter 6: CDS – Another Way of Delivering Chlorine Dioxide 145

Chapter 7: CDH – The Research Continues 155

Chapter 8: Step 3 – The Kalcker Parasite Protocol 165

Chapter 9: Step 4 – Other Supplements 243

Chapter 10: Step 5 – Chelation 255

Chapter 11: Step 6 – Hyperbarics 263

Chapter 12: Step 7 – GcMAF & Autism 293

Chapter 13: Beyond Recovery: The Maintenance Plan 319

Chapter 14: Miscellaneous Information You Should Know 323

Chapter 15: Final Thoughts 347

Chapter 16: Healing Beyond Autism 349

Appendix 1: More Miracles & Testimonials 357

Appendix 2: CD Autism Worldwide 441

Appendix 3: Accidental Cure by Optimists 443

Appendix 4: Autism Treatment Evaluation Checklist (ATEC) 447

Appendix 5: Molecular Mimicry 449

Appendix 6: Measuring the Strength of Your CD, CDS & CDH 467

Appendix 7: Diluting HCl Concentration 471

Appendix 8: Other Uses of Chlorine Dioxide 473

Appendix 9: Blank Kalcker Parasite Protocol Chart 475

Appendix 10: Lunar Phase Calendar for Kalcker Parasite Protocol 477

Appendix 11: Health Benefits of Diatomaceous Earth 485

Appendix 12: Summary of Protocols 491

Appendix 13: Body Burden – The Pollution in Newborns 499

Appendix 14: Avoiding Autism 505

Appendix 15: Recipes - Cooking with Ana 513

Appendix 16: Websites We Like 517

Appendix 17: Direct Help from Kerri Rivera 521

References 523

Index 529

About the Author 563

Foreword

Autism, not the one described by Kanner in 1943, but the one that we see today diagnosed in 1 of 50 kids is a combination of immune disorders that need to be treated biomedically. There is a lot that we still need to understand about why and how these immune dysfunctions affect our kids' development, causing near complete impairment of their social interaction and communication.

There is a long, hard road ahead to fully comprehend the integrated system that comprises "autism," but our kids cannot wait. It takes more than just medical, professional, or even scientific interests working hard day and night to find a suitable and effective solution to help our kids. It takes drive, passion, and guts to do the right thing; to hear and read the overwhelming stories of parents from all over the world and not turn our backs, but help. It takes an eternal and extreme will to help, even when your own child is on the spectrum.

Kerri's protocols have been indispensible to the full recovery of many of the kids in our Curando el Autismo (CEA) Foundation. These protocols represent a readily available and effective solution to alleviate most of the pathogenic insults to the immune system. Years from now, when "autism" diagnoses do not exist, when we find ourselves fully knowledgeable about the now mysterious immune-brain-behavior connection, I will remember Kerri not only as a friend but as one of the first courageous leaders that dared to change the path for our ill kids. She goes against all odds, sharing knowledge, experience and simply making it happen.

~ Lorna B. Ortiz, PhD.
President of Curando el Autismo, CEA

“Based on parent reports, the prevalence of diagnosed ASD in 2011–2012 was estimated to be 2.00% for children aged 6–17. This prevalence estimate (1 in 50) is significantly higher than the estimate (1.16%, or 1 in 86) for children in that age group in 2007.”

US Department of Health & Human Services
National Health Statistics Reports
Number 65, Pg. 2 - March 20, 2013
www.cdc.gov/nchs/data/nhsr/nhsr065.pdf

Preface

by Kimberly McDaniel

Darkness cannot drive out darkness; only light can do that.
Hate cannot drive out hate; only love can do that.
~ Martin Luther King, Jr.

Welcome to the Second Edition of *Healing the Symptoms Known as Autism!* We are absolutely thrilled to share with you the latest protocol updates and everything else that has happened since May, 2013.

You may be thinking: Why a second edition so soon? Our first edition gave us the structure to explain the protocol, and a foundation to build on. This book has already helped many families around the world, and in fact some parents read it and recovered their children without even contacting us until afterwards! Keep in mind that when it was released in May, 2013 it was absolutely up to the minute, but as we mentioned, this protocol will continue to evolve until we have something that is consistently recovering people on the spectrum of all ages. As of January 2014, we again are sharing the latest updates, as well as a whole lot more information that we hope will be as interesting and beneficial to you as it was to us.

Here are some of the exciting new additions:

- Olive Kaiser of www.GlutenSyndrome.net has written a section on gluten and its role in molecular mimicry and autoimmunity. Since many of you are not new to the autism community, a gluten free diet for your children is nothing new. However, you may be interested to find out just how damaging gluten can be for folks who are off the spectrum as well.

- Scott McRae has contributed a chapter on CDH (Chlorine Dioxide Holding [Solution]). A new method of preparing chlorine dioxide that many families are now using with success. This gives us an even wider variety of preparations available to accommodate the needs of our families.

- The Kalcker Parasite Protocol chapter now has some beautiful charts that spell out the timing of all the components for the 18 days a month that a child will be on the parasite protocol. Thanks to Dan Bender, a lot of the confusion surrounding how to make it all fit will be cleared up. You will also find years of lunar calendars to make it easy to see when the protocol is active. You won't have to check Google again to see when the full or new moon is coming.

- The "Worm Whisperer" Robin Goffe shares with us part of her journey towards healing for her 19-year-old son. Her advice is for the Extreme Cases—Self Injurious behavior, aggression, violence, etc. If you are the parent of an older child on the spectrum, or a child who displays these behaviors, or you know someone who is living with a child like this, you owe it to yourself to read Robin's suggestions. They are full of hope and wisdom.

- The one and only Marco Ruggiero, lead researcher on GcMAF, (Gc Macrophage Activating Factor—an immune system supplement) has written an entire chapter on GcMAF and its applications for autism. A must read.

- Last but not least, a whole new crop of testimonials that will make you cry. If after reading this book you still have doubts about giving this protocol a chance, I highly recommend you revisit these testimonials. If I had to pick one part of the book that was absolute favorite, this would be it. It may be because I collect a lot of them, and usually have a little correspondence with the families to get their permission, and genuinely get to feel their excitement, their sense of accomplishment when they see their child start to heal, and not to mention their endless gratitude for being able heal their own children.

  I pray that all of these testimonials find their way to the people who need them and as a result children stop suffering because their parents see them reflected in those words and understand that if other children can heal, their children can as well.

  To me, there is nothing more real than hearing from someone that has walked a mile in your shoes. There is nothing more inspiring that to hear someone say, "I know it's possible, because I did it, I lived it, and I am here to tell you about it."

The families that courageously walked this path and then took the time to share their stories are pioneers and heroes to their own families and to our entire community and beyond. They are blazing a trail for others to follow and countless lives will benefit from their diligence, fortitude and dedication. The stories of healing come from all over the world... from children and adults of all ages. We hope that they move you as much as they moved us. We are eternally grateful for the service these families have provided to humanity and thankful that they were generous enough to take the time out of their lives to pay it forward and share.

- The book now as an extensive index making it much easier to use as a reference.

As of this printing, families are healing their children with autism using this protocol in over 58 countries! Our Facebook groups have over 3,500 members in many of these countries. CD officially knows no borders. It is absolutely thrilling to us that people are coming together with the common goal of healing their children themselves, and to help others do the same! It fills our hearts with joy to be a part of this and to be able to see and feel the love that is shared every single day.

The development of this protocol has been a growing grass roots effort, in sharp contrast to how modern medicine usually works. The reason for this is clear: Modern medicine has not really helped to heal autism and may very well be part of its cause. While we don't have any double blind studies to rest on, we do have a slew of anecdotal evidence, which may not mean much to those in the arena of modern medicine or modern science, that doesn't make it any less real. Time and time again, our *results*—children dropping ATEC points—are being duplicated by families all over the world. For a family who is using this protocol, there is absolutely nothing more real than watching their child come back to them. Ask any autism parent what they would rather have... a long-term double-blind study published in a peer reviewed journal... or a healthy child. My money's on the latter.

As exciting as it is to be a part of this, and as wonderful as it is to hear about gains and read uplifting testimonials, we know there are still families working to heal children that are very sick, and our groups share the highs as well as the lows. Kerri will tell you that some months are better than others and the gains ebb and flow. I urge you to read and reread the testimonials. These are real stories of real healing and if you haven't already I invite you to believe that your child can be one of the success stories in those testimonial pages.

All of our families are proving what many were told was impossible: Children with the symptoms known as autism can heal! This grass roots movement is creating a paradigm shift in the way the world views healing from autism. The second edition of this book will be translated into at least 13 languages including Spanish, Portuguese, French, Flemish, German, Czech, Norwegian, Arabic, Polish, Italian, Hungarian, Bulgarian and Serbian. This is exciting. This is real. And, this is a living protocol. Every family that uses it every single day is helping to shape the future and heal the children who are affected today, and quite possibly prevent children from becoming affected tomorrow. For this we are eternally grateful.

Here's to you and the continued healing of humanity,

Kerri

Terminology & Units of Measure

Throughout this book we talk about "CD," which is an abbreviation for *chlorine dioxide*, a well-established oxidizer. Chlorine dioxide is also often referred to as MMS, which is the common name given to it by Jim Humble, the man who discovered various applications of chlorine dioxide. There are many books, videos, blogs and articles using the name, "MMS," which is surrounded by a fair amount of controversy. We choose to not get tangled into that debate since our focus is on helping our children recover from autism. For us, our only concern is that (1) it is safe for our children, and (2) that it works. Based on extensive use of CD on thousands of children with autism, we can confidently say that both those statements are true. If this were not so, this book would not exist.

Units of Measure

Throughout this book we talk about various measured treatment components and containers using fluid volumes and weights. Since this book primarily addresses a US audience, we sometimes talk about measures in the US system of pounds and fluid ounces; while also using the internationally recognized metric system, which is frankly... easier to use.

The common abbreviations of measure you will see throughout this book include:

l = liter (volume)

mg = milligram (weight)

ml = milliliter (volume)

lb = pound (weight)

lbs = pounds (weight)

ppm = parts per million (concentration)

fl. oz. = fluid ounce (volume)

net. wt. oz. = net weight ounce (weight)

For easily measuring small amounts, syringes (without needles) are a great tool. Don't bother going to a chain store pharmacy... they won't sell them to you without a prescription. Instead, check your local medical supply store, animal supply store, private pharmacy or protocolsupplies.com. Cost: Surprisingly dirt cheap. The complete set of five below cost about $1.00 in Mexico—less than most candy bars—and not much more in the US. Note: Some brands of syringes have printed on scales that easily rub off, especially if your hands are a little oily. To prevent from losing the markings, cover the scale with some clear tape or clear nail polish.

SensiMedical
SensiMedical
SensiMedical

When it comes to accurately measuring larger volumes, you can buy a set of 7 Polypropylene Graduated Cylinders. Of course this is not a requirement. You can use common kitchen measuring utensils, but these are more accurate and easier to read. If you decide to get them, avoid those with printed on lettering (it rubs off). Raised lettering is best, although sometimes a bit difficult to read. Typical cost on ebay™ or Amazon® for a set is around $25USD. The sizes range from 10ml to 1000ml. You can also buy individual cylinder sizes made of plastic or glass.

Acknowledgements

Gratitude makes sense of our past,
brings peace for today,
and creates a vision for tomorrow.
~ Melody Beattie

Thank you to my Mother who taught me by example the value of volunteerism and to help others who are less fortunate. Thank you for always telling me that I am the best and that I could do anything. I believed you :)

Thank you to my husband of nineteen years for supporting me on this journey. As challenging as it is, you have always stood by me. You are the love of my life, and father of the sweetest boys on the planet. Thank you for giving me the strength to be strong enough to do a mammoth job. It would not be possible without a pillar of your magnitude.

Thank you to Alex, the best big brother in the world who supports and enjoys what we do as a family and who sees the big picture. I couldn't be prouder than I am of you. I am so grateful you chose me as your Mom, and you chose to be on this journey with me, the road less traveled. I love you forever!

Thank you to Patrick for bringing to our family a higher good. You are an angel. You have brought light to all who meet and know you. Thank you for choosing me to be your mother. You and your brother have taught me more about myself and about life than I could have ever dreamed. I am grateful to you and I Love You more than words can say. Through me, you have given people their lives back. You have forever changed the face of autism.

Thank you to Linda for giving me my sister and one of my greatest friends.

Thank you Slimmie for your support, love and dedication, you are my BFF and we speak the same languages. Thank you for the support of my projects and always having me look prepared.

Lorna, thank you for supporting me, believing in me, trusting me, being faithful to your heart, and for being one of the most honest people I have ever known. You are my sister of the heart and I love you very much.

Susan Wiseman, thank you for offering your connection on that day when I most needed it. That one act of humility changed the future forever. I will never forget what you did for me.

Norrah Whitney, thank you for pointing me toward the direction of recovery, and explaining to me what I needed to do. You are without a doubt my first angel on this journey.

Ana Meckes, thank you for being another much needed angel and teaching me how to advocate for my son. I have not stopped ever since.

Anju, thank you for being my friend, for sharing with me, and for believing in me. You are a unique treasure.

Pina, thank you for helping and supporting me, we will be friends forever. I will always remember what you did for me, the clinic, and the kids.

Kenny, thank you for you excellent books full of sage advice and making biomed user friendly. Most of all, thank you for sharing fever therapy with us.

Carolina, Yamileth y El Jefe... gracias por apoyarme, defenderme y creer en mi. Son amigos en todo el sentido de la palabra. Estoy orgullosa de ustedes, moviendo montañas con el amor que solo una familia puede tener por un hijo….y ustedes lo han hecho por un pais. Les quiero mucho.

Bob Sands, thank you for seeing something special in me, thank you for taking me to meet Bernie and Mrs. Rimland. That day changed my life forever. Thank you for giving us the best hyperbaric chamber in the world. With that chamber we have seen so many miracles.

Jim, you have brought light to many and hope to all who have had the fortune to know the miracle that is CD (Chlorine Dioxide). I am grateful to you every day of my life. I have witnessed first-hand the miracle that is autism recovery. Thank you for staying the path that has never been easy. But it is as you say, “The right thing to do...” I love you.

Andreas, my dear friend, thank you for your dedication, your research and for your willingness to help. You are the science, the reason and the truth that is chlorine dioxide, the molecule that has the power to save humanity. Thank you to you and Miriam for your unfailing support over the years. Your contributions to the world of autism are changing the way the world views autism forever. And the lives of so many children are now recovered thanks to your contribution to The Protocol. Without you both, these recoveries would not have been so abundant.

Dan Bender, thank you for seeing the big picture and for giving so generously of yourself to help us help children with autism. Your selflessness has allowed this protocol to reach more families all over the world. Thank you for making the second edition of this book a reality, we never would have gotten it off the ground without you.

Thank you Michael Harrah for standing by me and working tirelessly to share the information that has had such a positive effect on so many families all over the world. Your wisdom and know-how have been invaluable to me in this book, our website and forums. You were a friend when I most needed one. I am grateful that you are in my life.

Dr. Bernard Rimland, even though you have left us all too soon, you moved mountains while you were here. Thank you for letting me train as a DAN! Clinician, and for allowing us to translate the protocol to Spanish. Thank you for giving us Infantile Autism in 1964, changing forever the thinking that autism was caused by the refrigerator mother. And setting into motion biomedical interventions for healing our children I wish we had more like you. I try to think "What would Bernie do?" and I usually get the answer. Always helping and always available. Humility. You set the standard for humanitarian.

Thank you to all the families on our forums for blazing a trail for others to follow and fighting for the health of your children every single day. You are an inspiration.

To all the Mods; Ginette, Caryn, Joy, Alison, Heidi, Michael, Pam, Katya, Carolina, Nilesh, Mirena, Robin, Debbie, Sue, Susan A., Brandi, Don, Clint, Maggie, Claire, Amber, Dawn, Naomi, Maryann, Susan R., Stacey, Jessi, Lina, Boris and Susanne, Olive, Dana, and Pat, for being the best mods in the world. Your help changes lives for the better each and every day. This is for you:

> *It's in our interest to take care of others. Self-centredness is opposed to basic human nature. In our own interest as human beings we need to pay attention to our inner values. Sometimes people think compassion is only of help to others, while we get no benefit. This is a mistake. When you concern yourself with others, you naturally develop a sense of self-confidence. To help others takes courage and inner strength.*
>
> ~ The Dalai Lama

Thank you Joy for sharing with Alison that you had heard about CD for autism through the seminar that Jim had in the Dominican Republic. That gesture opened the gates to the north and from there to everywhere. You are a special healer.

Teri and Ed Arranga, thank you for the platform and for helping so many families find what they need for their children, and never blinking, even when it gets really scary.

Thank you Doll, for a much-needed distraction from autism when I really need it most. BFFs are good at doing that.

Thank you to the Mansours, for believing in my projects and me. Your support made this book and our website possible.

Thank you to the various proof readers who helped with that fun process, including Michael Harrah, Pam Gotcher, Joy Whitcomb, Charlotte Lackney, Don Kalland, James Beyor, Cathy Fuss, Jeremy Horne, Ph.D., Luane Beck, Candace, Andreas Schreiber, Olive Kaiser, Susan, and Clint Melanchuk.

Thank you Mads, for your website magic and our beautiful logo.

Thank you Carolyn Unck for helping get this book into shape and all your support and advice.

Thank you Marco and Stefania for thinking out of the box and preserving the truth because it works.

Thank you so much Pam Gotcher for going above and beyond everyday to make sure our readers get their books!

Thank you Scott McRae, Brenda McRae, and Charlotte Lackney for contributing the CDH chapter. This new method of preparation has already been beneficial to many of our families, and is an exciting new addition to this book.

The purpose of life is to contribute
in some way to making things better.

~ Robert F. Kennedy

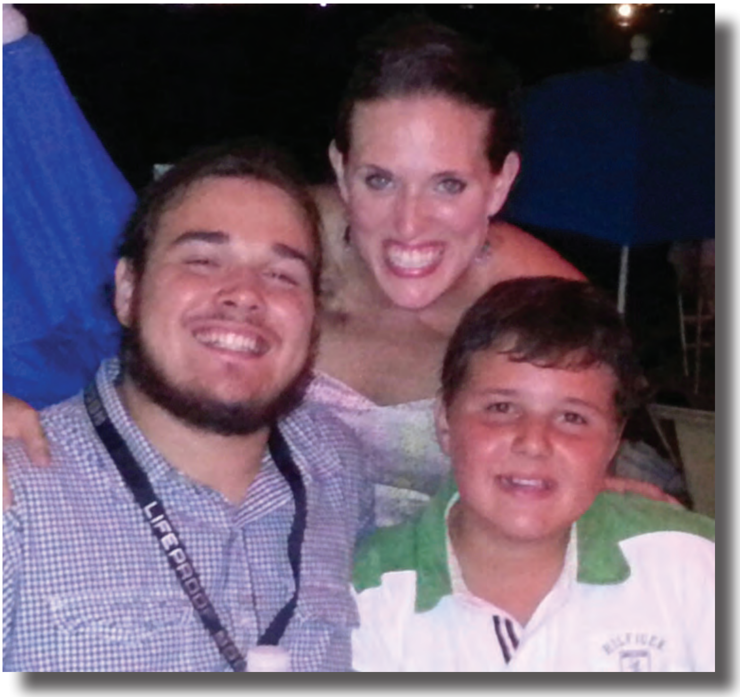

Alex Rivera, Kim McDaniel (Kerri’s sister), & Patrick Rivera

Introduction

Autism is Avoidable, Treatable and Curable

Your body's ability to heal is greater
than anyone has permitted you to believe.
~ ***Anonymous***

Congratulations on finding this book, and welcome to the world of autism recovery. This book comes to life as more and more children with an autism diagnosis respond and recover in more than 58 countries around the globe. This book gives families a do-it-yourself guide to an *Autism Spectrum Disorder* (ASD) recovery program with answers all in one place.

I have, on my ASD recovery journey with my son Patrick, personally been frustrated with the lack of information and answers leading to lost time and money. For example, when I first knew that Patrick was no longer developing "normally" (in 2003), I was unable to get a diagnosis. Seven years, dozens of interventions and hundreds of thousands of dollars later I was still searching for pieces to Patrick's autism puzzle.

I learned over the years of many people recovering their children using various protocols and interventions, and I looked into all of them. Some gave us improvements but not recovery (most specifically for Patrick, diet). Some gave us nothing.

My goal with this book is to alleviate that frustration and loss of time and money for other parents.

I became interested in chlorine dioxide (CD) in 2010 but I was unable to find any information about using it with autism on the Internet. Since I knew that almost every child with autism suffers from similar pathogens (viruses, bacteria, candida, and parasites), heavy metal toxicity, inflammation and allergies I researched those conditions in combination with chlorine dioxide—removing "autism" from my vocabulary.

I realized with further research that CD would be excellent for curing the symptoms collectively known as autism. When Patrick was first diagnosed in 2004, his *Autism Treatment Evaluation Checklist* (ATEC) score was 147 and after six years of biomed he was at a 63. (The diets made the biggest difference in those initial dropped ATEC points). 2½ years of CD later he is at a 21. CD has made all the difference in his life, in my life, and in so many lives around the world.

I brought CD to my *Defeat Autism Now!* style Clinic in Puerto Vallarta in 2010. Today more than 115 children globally have lost their diagnosis of autism (meaning an ATEC score of under 10 points). Additionally, thousands of children around the globe have dropped points on their ATEC and are moving towards recovery. 27 children in less than 1 year in Venezuela alone lost their diagnosis of autism with a combination of diet, CD, and ocean water to the surprise of the doctors who diagnosed them in the first place. Many of those same doctors are now looking at CD for other patients.

It is my dream that every family of a child with autism be given this information so that they can decide for themselves whether they want to give it a try.

This book is a protocol for all of us. Some of you may be completely new to autism recovery; some of you may be a veteran like me and/or parents of older children and adults on the spectrum. This protocol works for even the classic "non-responders" and for those who are so close to recovery yet can't seem to get through the door. This book is for you. CD helps the body heal the symptoms we call "autism" across the board—it is an equal opportunity healer.

I know from experience that an autism diagnosis is devastating on many levels. The initial regression of a neurotypically developing baby takes away eye contact, speech, and the emotional connection between parent and child. Then, when strange new behaviors such as flapping, squealing, rocking, spinning or even self-injurious behaviors appear, you know in your deepest mommy (or daddy)-gut that your child was not born like this. It seems to take forever to get the truth about what happened to your happy, healthy baby. It takes even longer and is more confusing when you have to figure out how to heal that sick child. Far too many "health" providers in the autism field are focused on making money so we can't blindly trust anyone. We must do our own homework. The journey itself is a lot of trial and error coupled with misinformation.

Many supposed "autism experts" don't know much about recovery, or the effective order of treatments and end up costing our children time and the parents money. The less time a child spends chronically ill, the easier and faster it is to recover them. Not to mention the child spends less of their lifetime suffering the physical, emotional and mental effects of autism.

I feel that all parents who begin this protocol should expect a full recovery from autism because this protocol treats what causes this diagnosis. Our research indicates that every person with a diagnosis of regressive autism has virus, bacteria, candida, parasites, heavy metals (biofilm), inflammation and allergies. This protocol handles every one of these issues, and that's why it has been so successful. Some recover faster than others. But, every day we are one step closer to the end of autism.

How do you know if the protocol is working and how long does it take to see results?

The ***Autism Treatment Evaluation Checklist*** (ATEC) is our measure. The ATEC is an online survey that evaluates the severity of a child on the spectrum. For more information see Appendix 4, page 447. Many families notice changes from day one, while others take longer. You will get results when you correctly apply the interventions in this book in the proper order and without breaks.

What are the results you can expect from the protocols in this book?

I would love to say that everyone who follows the protocol will get their child down to an ATEC of 10 or less—what we call a recovery—and we have 115 of those already. The majority of those I'm in contact with report substantially significant improvements, even if they have not reached recovery. In the case of my son Patrick, he started at 147 in 2004 and has come down to 26. I'm optimistic as I continue to search for further answers, and will continue to share what I find.

If you are not getting results and have gone through this book including the Frequently Asked Questions (FAQs) and troubleshooting, please contact me through the forum at...

www.cdautism.org

...there are always tweaks we can make to keep moving towards recovery.

I recommend you read this book straight through in the order that it is written as that is the order in which it is to be applied. Jumping around the interventions can lose time for your child and waste money for you. Doing interventions in the correct order, when your child is ready for them, is the best to way to achieve recovery. Diligence and perseverance win the race every time.

My mission is to share with whoever is interested, the blessings that I have received. If the information presented here feels right and resonates with you, then please, give it a try. It might just be what your child needs.

This book in a nutshell:

If you want the best chance at recovery, here's the overview of how to do it:

1. The Diet: Eliminating gluten, dairy, soy, sugar, and toxins, to stop inflammation and reduce the overall toxic burden.
2. The CD Protocol to kill pathogens while using a multimineral like ocean water.
3. The Kalcker Parasite Protocol.
4. Explore and implement other potentially synergistic supplements to aid in speech facilitation, neurotypical behavior and/or seizure reduction.
5. Consider gentle chelators.
6. After 3 parasite protocols and consideration of all above steps, find a hyperbaric chamber (1.75ata).
7. Consider adding in GcMAF.

It is important to consider all of the pieces of information in Chapter 14, *Miscellaneous Information You Should Know* (page 323) and apply them from the beginning when they make sense for your child. You also may find the *Summary of Protocols* in Appendix 12 a good resource when you do not have time to reread a chapter to find something specific.

On the right is what we call the *Stairway to Recovery*. Joy Whitcomb, one of our amazing moms, came up with this so you can see just how each step rests on the previous steps, and without them you would not reach the top step… RECOVERY!

Author's Note: It is never my intention to "change someone's personality" or take away their character by healing autism. I see it quite the opposite. When children start to recover, their personality starts to shine through. The behaviors that we saw before (spinning, flapping, squealing, smearing feces, auto aggression, lining things up, tantrumming, etc.) are not personality traits but symptoms of a sick body. These symptoms start to disappear once the body begins to heal and our children can express who they really are through smiles, eye contact, words, gestures, etc. They can show us what they need and want and play an active role in their own lives. It is my dream that every child has the opportunity to mature and choose the life they want for themselves, and be responsible for making their own decisions. I truly believe this is possible for all of our children and adults on the spectrum, and I want families to have the opportunity to offer healing to their children.

Throughout this book we have used the pronoun "he" when referring in general to "a child on the spectrum." This is not to alienate families with girls or women on the spectrum. It is simply a question of fluidity. To use *he/she* or *his/hers* every time we chose to use a pronoun seemed cumbersome, so therefore we are using "he" or "his" throughout the book. We chose "he" rather than "she" because autism is five times more common among boys than among girls. As of March 2013 the CDC revealed the results of a new study conducted during 2011 and 2012 which surveyed 95,000 families and estimated the prevalence of autism at 1 in 50 children.

The acronym DAN! is no longer applicable to *Defeat Autism Now!*, as it belongs to *Divers Alert Network*. The acronym has been used in several places in this book as several of the personal anecdotes are from a time when its use was still appropriate. Today, a "DAN! Doctor" could be defined as a practitioner who received training through the network formerly known as *Defeat Autism Now!*.

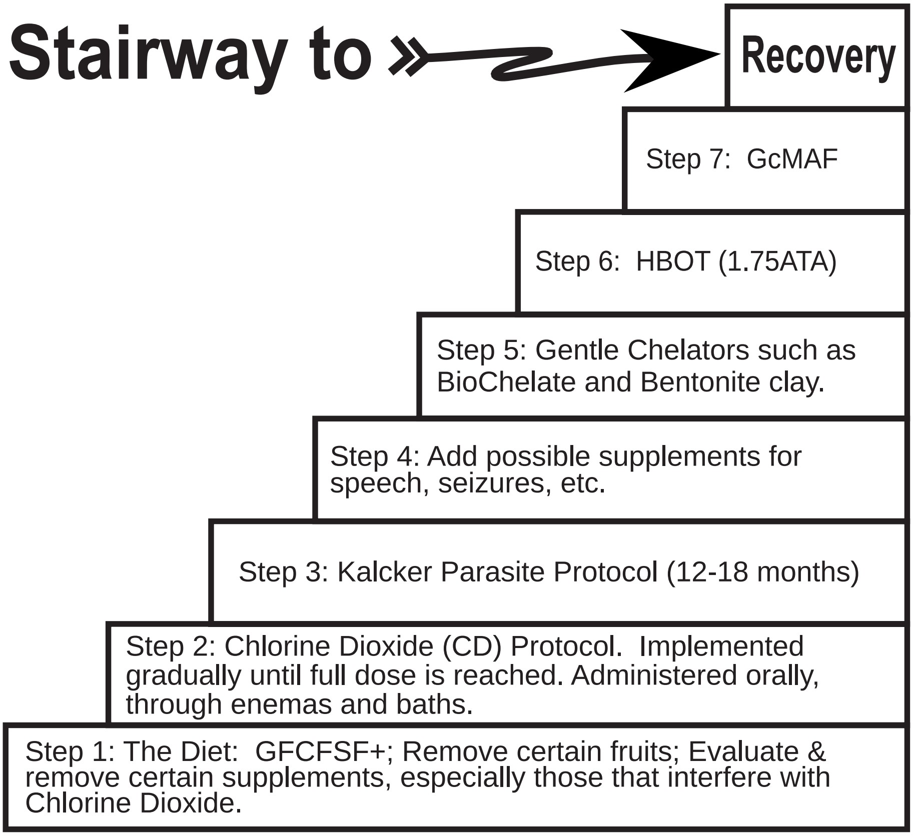

Stairway to » Recovery

Step 7: GcMAF

Step 6: HBOT (1.75ATA)

Step 5: Gentle Chelators such as BioChelate and Bentonite clay.

Step 4: Add possible supplements for speech, seizures, etc.

Step 3: Kalcker Parasite Protocol (12-18 months)

Step 2: Chlorine Dioxide (CD) Protocol. Implemented gradually until full dose is reached. Administered orally, through enemas and baths.

Step 1: The Diet: GFCFSF+; Remove certain fruits; Evaluate & remove certain supplements, especially those that interfere with Chlorine Dioxide.

Concept by Joy Whitcomb.

Everything yields to diligence.

~ Antiphanes

IMPORTANT NOTICE

Please keep in mind that the protocols in this book are still evolving, and will continue to be improved as new discoveries are made. We will release new editions to incorporate those new discoveries. The topics of CDS & CDH are particularly new and rapidly evolving. This book is current as of January 2014. Please be sure to check this book's website for important corrections and updated information beyond this and subsequent editions:

HealingTheSymptomsKnownAsAutism.com

Chapter 1

Kerri's Story

"The impossible is now declared possible as soon as you agree to it. It's just a flip of the mind, letting go really, nothing more."

~ Stuart Wilde

"What happened, what did you do to Patrick?" That was the first thing my husband asked me when he first saw our son, after returning from a weeklong trip. This was just five days after our youngest son Patrick received his last vaccine—the DPT (diphtheria-pertussis-tetanus) + Hepatitis B + Influenza B (what is known in Mexico as the *Pentavalente*) on the 13th of August 2002 at two years and one day old.

That question was the first of many to propel us down our path paved with autism. I told Memo (my husband) that we shouldn't worry. The nurse mentioned that we could expect a fever, or he could be listless. These were completely normal reactions. Contrary to what she said, what we observed during those first days and weeks were loss of eye contact, flapping, toe walking, high-pitched marine noises along with excessive drooling—drooling that would soak the front of his clothing.

Patrick had also lost all of the speech he had acquired; Mama, Pa, *agua* (*water* in Spanish), letters of the alphabet, numbers... all of it. The only thing he wanted to do was watch videos while running back and forth in his bedroom squealing the "ambulance sound," flapping and banging his gut, and drooling through all his clothes.

Not being well versed at the time in these problematic symptoms, we chalked it up to a case of the terrible twos. But these particular terrible twos led to loss of sleep for Patrick as well as the rest of the family, antibiotic use for green nose and eye mucous, and raging diarrhea that was so acidic it would burn his skin upon contact. This would be the story for the rest of the third year of his life.

The first of many to associate Patrick's behavior with autism would be my great aunt. She mentioned to me that she believed Patrick had autism after observing him at a family get together in April of 2003 while we were visiting relatives in Chicago. It was the most ridiculous thing I had ever heard.

Nevertheless, I went home that evening and immediately Googled the symptoms of autism. Laid out on the website were symptoms like spinning objects, lining up objects, self-injurious behavior, lack of socialization, and several other factors that I felt in no way described my son. I dismissed the comment, and continued to watch my son display strange behaviors, without understanding why.

I was back in Chicago a few months later in July. While going out for a run, I saw a girlfriend of mine who has a child the same age as Patrick. We stopped to chat and she asked me how he was doing. I said, "fine." She asked me if he was talking yet. I told her he had developed language, but since March he had lost all the vocabulary that he had previously acquired. My girlfriend looked at me and said "Oh…" with a look of concern on her face. I asked her, "What's wrong? What does that mean?" Now I was extremely nervous. "Well," she said reluctantly, "loss of language is a red flag for autism." There was that word again. I told her that we had already gone through all that, because by then I had taken Patrick to a neuro-pediatrician in Guadalajara, a psychologist with a huge center in Guadalajara, and a local psychologist in Puerto Vallarta. All three of these experts had told me that he was fine; they didn't see any problems with his development.

I started on my jog, and about halfway through, it hit me that Patrick did have autism, so I sprinted home, sat down at the computer and went to the *Autism Society of America*'s website and found a checklist of 16 symptoms of autism. The directions said if your child had 12 or more than most likely he has autism, Patrick had exactly 12. Looking back, he most likely had 14 or more, but I wasn't ready to see those just yet.

That same day I called his pediatrician in Puerto Vallarta and I told her I thought my son had autism, and she told me, "No, I have never seen anything like that in your son, but bring him and I will take another look." When we went in she observed him. She said that he didn't line things up, he didn't hit his head, he would still come to me, he was 'playing' with some toys in the waiting room, so therefore my son didn't have autism.

The pediatrician told me to go home that day to wait for him to straighten out. All the experts told me he was the way he was because he was the baby of a bilingual family that—according to them—traditionally produces late talkers; his parents and brother were all late talkers. He was spoiled. He had a nanny. And, he's a boy—boys are late talkers, etc., etc. The pediatrician successfully talked me out of the diagnosis once again.

As nothing was wrong, that fall we put Patrick into a kindergarten. His teacher would tell me Patrick isn't doing this activity or that, and I replied that we had taken him to see all the specialists, and he was just a late talker. She was super sweet about it and every few months she would make a comment to me because she saw him drastically different from his classmates.

Then one day it happened, I got a call from the director of the school saying that her friend, a neuropsychologist was in town from the United States, and she would like her to see my kids. At 6pm on March 12, 2004 I had an appointment with this neuropsychologist. I had already figured that this woman wanted to see Alex, my older son because he hadn't been doing well at school. He hadn't been sleeping well since Patrick hadn't been sleeping, which led to some poor performance in school. I had already gotten the news that Patrick was fine, so when she started focusing on him and his behaviors I was slightly confused. We sat down in his classroom, and she started asking me if he always… runs in circles, flaps, drools excessively, squeals like a dolphin, etc., (we actually used to joke that his mother must have been a dolphin). After that, I told her how we had already gone to all of these specialists and they said he was fine.

I was tired of getting the run around, and then everybody asking me what was wrong with my kid when he was just taking some extra time getting going. That's when she said to me, "I can't believe they didn't tell you your son has autism."

These words changed my life forever. Of course I asked her if she could be wrong, and she said, sure there exists a possibility, but she had done her post grad in autism, had seen hundreds of cases and this diagnosis was her professional opinion. That opened the floodgates for a river of tears that didn't end for years to come.

Being a positive person, I asked her what I should do. She said, "I would like to introduce you to a group of psychologists that are in town." The next day I went with her to a hopeless place, with hopeless people and I asked her if this is something that we can cure, and she said, "No, you can do therapy with these psychologists, and that's it." Children are born with autism and they will die with it, was the basic sentiment. I knew for sure that my son was NOT born with autism. He was the smartest, bright-eyed baby ever, and we had the photos and the videos to prove it. He was not born the ghost of a child that we had now. I knew that I would keep searching until I found something else for Patrick. I became proactive and never returned to that place.

The very next day I saw another girlfriend and she mentioned that she had a book on ADD and autism, so I immediately picked up the book and it was all about diet. Gluten-free, casein-free to be specific, and I decided to start it right away. Truth be told, Patrick's diet was horrible, he had self-restricted to only dairy and wheat. Bread products and cheese were his staple foods, but the good news was he still ate potatoes. Even as an autism novice, I knew I couldn't take him to a fast food chain because those fries were coated in gluten. We started off with just homemade fries with sea salt, because that was the only thing left on his diet that he would eat. After three days of being on the diet he said three words, the first three words he had said in over a year. I knew we were onto something.

The next week I ran into a friend from tennis, I really didn't want to say hi, because I was really depressed, but I heard a voice saying to me "It's not about the road but about smelling the flowers along the way." So, I forced a smile on my face, and went to say hi to her in my great depression. Well, she started to complain about her week, so I listened patiently, and then I told her about my week; Thursday my identity was stolen on the Internet, Friday my son was diagnosed with autism, and Saturday my dog of 14 years had to be put to sleep.

When she heard all this she turned off her car and told me how sorry she was. She said that she would put me in touch with her friend that had opened the *Early Autism Center* in Toronto, Canada. When I woke up the very next morning, I had a long email from Norah Whitney waiting in my inbox. She would turn out to be the first of many autism angels in my life.

I received a lot of precious information from that email, but what was maybe the most important detail for me was that what had happened to my son was an effect of all the vaccines he had received, he wasn't born with autism, as I knew. I also hadn't caused this, and I needed to start letting go of some of my guilt. Norah also told me that we can treat autism and I needed to immediately contact a DAN! Doctor, and Dr. Bobby Newman, Board Certified Behavior Analyst and Licensed Psychologist. Norah said that they were the best, and that this group of DAN! Doctors were curing autism.

I got in touch with everybody and that same month we started our own *Applied Behavior Analysis* (ABA) program. I also took Patrick to the States to see his first DAN! Doctor. When I got back home from that trip I had bought nearly $5,000 (US) of supplements and injectables. That didn't mean I knew how to use them, and by no means was I watching my son improving before my eyes. This was June of 2004, and we went through the rest of that summer with some supplements, other biomedical interventions, and 40 hours a week of ABA therapy.

That fall, some friends mentioned to me that their father was receiving chelation in San Diego, and I had just heard someone else saying that chelation was working for kids with autism, due to their extreme metal toxicity. By the end of the evening I had the phone number for the clinic that was providing his chelation. When I called for information, they told me that they referred all of the children's chelations to Dr. Woeller in Temecula California, so I made Patrick an appointment. In March of 2005, I took the whole family to Temecula to see Dr. Woeller, because basically nothing was happening with the supplements I had been giving to Patrick. I knew I needed to keep exploring other avenues.

Several thousand miles and various tantrums later, we arrived at Dr. Woeller's office. I told the receptionist that we had come to see Dr. Woeller. To my chagrin she said, "No, you have an appointment with some other doctor, Dr. Woeller isn't even in town right now." My husband was sure I had messed up the appointment and was extremely upset with me. Meanwhile, Patrick was screaming, crying and taking his clothes off in the waiting room. They finally let us through to speak with the other doctor and after going through Patrick's autism we explained that we wanted to chelate. We had the understanding at that point that his autism was coming from the mercury in the vaccines. She told us point blank that we couldn't do everything at once, and before we could chelate we would have to clean up his gut. I flew home with my family totally deflated. We started on the impossible journey of cleaning the gut.

My first ever autism conference turned out to be *AutismOne* in late May of 2005. I met a lady who was a rescue angel for *Generation Rescue*, and I told her I wanted to chelate my son with those DMPS drops, which were fashionable at the time. She told me that when it came to my son's recovery I shouldn't take no for an answer. I had to advocate for my son, and not let myself get pushed around. With that wisdom, I called Dr. Woeller's office again and told them I wanted my appointment with Dr. Woeller himself, and I would not be given the run around. The manager listened to me and said that she would put me through to him. After a long c'onversation, Dr. Woeller agreed to take Patrick's case and that he would help me chelate my son. He agreed with me that it was silly to think that we could have the gut completely under control before beginning chelation. Eventually I got the DMPS drops for Patrick that I had so coveted, and I felt we were back on track. However, after about six months of using the drops we still saw no change.

Flashback to that same *AutismOne* conference, where I had attended a lecture about the *Specific Carbohydrate Diet* and the removal of all grains from the diet—a diet that helped Patrick inch forward minimal improvement by minimal improvement.

In November of 2005, I had a phone consult with Dr Woeller, and having been disappointed with the transdermal DMPS I decided to ask if there was anything new in the world of autism. The answer that would end up changing all of our lives was hyperbarics or HBOT (hyperbaric oxygen therapy).

I learned that there was a really nice man by the name of Bob Sands in San Diego who owned a corporation that manufactured hospital grade hyperbaric chambers. I phoned Bob's office because Patrick was going to need 40 sessions right away, and I wanted to check prices and see if there was a package discount for hyperbarics. The answer was yes, in fact there was a discount. I scheduled Patrick's HBOT, and took both of my kids to San Diego, where we stayed for 20 really long days to get Patrick his first 40 sessions of hyperbarics—two a day—everyday, morning and night.

During that time, my husband and mom would call and ask me if Patrick was better, but he was still naked in front of the TV, jumping up and down, flapping and squealing. It wasn't until we got back home that he began to pronounce the first syllables of all of the words for things he wanted, like "ap" for apple. We considered hyperbarics a great success, but like I said that wasn't until a couple of weeks after we finished the 40 sessions. That's when we really began to see the changes in Patrick. Bob always says that HBOT is the gift that keeps on giving; you can see benefits for up to two months after you finish your sessions.

Meanwhile, Bob and I hit it off right away. He had a real jovial and family-oriented atmosphere in the clinic. I told him my story and how I wanted to help people know that autism is avoidable, treatable and curable. I shared with him that there was no information, much less Biomed or autism recovery in Mexico (and most of Latin America). I wanted to help people and to share with people that there is a lot we can do to help our children heal.

The very next day he walked into the office and changed the course of my life forever. He told me that he was in fact friends with Dr. Bernard Rimland, "Bernie," the Grand Godfather of Biomedical treatment for autism, and the author of *Infantile Autism*, *Dislogic Syndrome*, and the founder of the *Autism Research Institute*. I told Bob that meeting Bernie today, would be akin to having met Mick Jagger when I was 15. He then announced that we would be having lunch with none other than Bernie himself and Mrs. Rimland. For the first time in my life, I admitted to Bob that I was so excited that I didn't know what to say. Bob told me that when it was my turn to say something that I should ask Bernie "What can I do for DAN!?"

In a tizzy I ran over to Marshall's, the discount retailer; bought a new suit, hose, and even a pair of high heels. The next afternoon, I put on my new ridiculous clothes, ditched my kids with my friend's maid, and hopped in Bob's Jag to meet Bernie and Gloria (Dr. and Mrs. Rimland)—an absolute dream come true. When we arrived at their favorite restaurant, I was completely overdressed. Gloria ordered a salad, and the guys had tilapia because Bernie didn't really like vegetables. The conversation ranged from Kinotakara patches, to the difference between soft-sided hyperbaric chambers and hard hyperbaric chambers.

I waited until there was a lull in the conversation and I went for it, I asked him what could I do for DAN!, and he said I should translate the DAN! Protocol and take it to all of Latin America. His words left me totally dumbfounded. Had I heard right? Latin America? I was thinking my town of Puerto Vallarta, or maybe Jalisco, (the state we lived in) and maybe in my wildest dreams all of Mexico, but this was much bigger than I had imagined. At this point, though there was no stopping the momentum that we had started. Within months we translated the Protocol to Spanish, and donated it to the Autism Research Institute, for dissemination throughout Latin America. Much later I would find out that one of my dear friends, Yeroline, would heal her son from autism using the translation of the DAN! Protocol, as well as others.

During that same trip my husband and I spoke to Bob about buying one of his chambers for the not-for-profit autism clinic we were planning on opening in Puerto Vallarta. The plan was to run the clinic on a non-profit basis, but anchor the clinic by charging the public for their sessions in the chamber, allowing all the profits to pay for children with autism to go for free in the chamber.

It was a green light all the way. We deposited the money for the chamber in March of 2006, and it arrived on October 31, 2006. AutismO2—Hyperbaric Clinic officially opened its doors on Dec 1, 2006. We threw a party for the inauguration, invited friends, family, Patrick's therapists, and our local priest came to Bless This House. We all wore white, and in honor of Patrick, who was our motivation for opening the clinic, we all had on nametags that said *Kerri—Patrick's Mom*, *Memo—Patrick's Dad*, etc. It was a special evening for all of us, but me especially, as it solidified what would be the path that I would still be walking as I write this book seven years later.

The clinic had a Sand's Hyperbaric Chamber, an allopathic physician with a specialty in Hyperbarics, 2 psychologists, 2 hyperbaric technicians, and me, to initially meet with the parents. As for the chamber, we always knew that it was going to help a lot of kids with autism, but it has helped a lot of others as well. The chamber was what gave language back to Patrick in 2006.

Later that same year we sent our allopathic doctor with our naturopathic hippie doctor to Ixtapa, Mexico for a hyperbaric conference, where they serendipitously met Dr. Giuseppina Feingold, (Dr. Jo) a DAN! doctor using hyperbarics for children with autism. As soon as our guys got back to Puerto Vallarta they started insisting that I had to contact her because she is curing kids with a protocol that included hyperbarics. Never one to waste time, I emailed her immediately, but at the time I didn't know she wasn't an email person and, because I am just not a phone person, we didn't connect.

Fast forward to January 2007, my husband, Memo, was buying a 1951 Desoto on eBay. We had to pay for the Desoto with a check, and therefore started to talk to Bryan, the seller of the car. During their conversation, Memo, my husband, said to Bryan "If you don't want to sell, because it's worth more than your selling price I will understand." Bryan had already decided to let it go, but at the same time he was interested in where the car was going, what Memo did for a living, etc. Memo told him about our life in Puerto Vallarta, and the classified ad magazine business that he owns. That's when Bryan stopped him and said, "I've heard of that magazine!"

It turned out that this Bryan was a nurse who works with a doctor who does hyperbarics and heals kids with autism. Well, that piqued Memo's interest, so he told Bryan about Patrick's autism and then Bryan says you must call Dr. Jo, yes, the very same Dr. Jo that our clinic doctor had met months earlier at the Hyperbaric Conference in Ixtapa. So, I called her immediately and when she answered I said who I was, and I told her Bryan told me I had to call her. I asked her if she believed in God and she said yes. I went on to tell her that in September she had met with my clinic's doctor and how I had emailed, but never heard back. At the time, Dr. Jo had been receiving so many emails that sometimes she couldn't get to all of them.

Finally we had connected, and we hit it off right away. I immediately told her all about the clinic and Patrick, hoping that she could come to Vallarta. She told me that I had to go see her in New York first, and we would get Patrick some treatment. So off we went to freezing cold New York in March of 2007.

We immediately started treating Patrick with IV chelation. The theory at the time was the main factor causing autism were heavy metals coming from vaccines that were damaging the methylation pathways. During my daily treks to Dr. Jo's office I finally met Bryan, the owner of the Desoto that my husband bought on eBay. We started talking about my clinic and what we could do collaboratively if Bryan were to move there. Bryan had already been in nursing for almost 30 years and was an expert in ozone and other alternative therapies. He told me he was ready for some changes, and we were very interested to have someone of Bryan's caliber and relaxed personality to join us at the clinic.

Within a week of my return to Puerto Vallarta Bryan came for his first visit, to see if he could call Vallarta home. Two months later he came back with all of his equipment to set up his office. We were a great fit, and Dr. Jo would fly down from time to time to help out with patients.

From 2007 to 2008 we were treating Patrick with IV chelation on top of his regular GF/CF/SF, diet, supplements, and hyperbarics. In May of 2008, I met a mom who had recovered her autism spectrum child with homeopathy. Almost immediately I started working with a world class homeopath from June 2008 to May 2009, but I didn't see anything to make me feel that the cure to autism was down that path. That same year I met a doctor doing the Yasko protocol, and she had some ideas, so we gave it a try from August 2009 until the end of May 2010. At that point, Patrick honestly looked worse than before we had started giving him 80 supplements a day.

By now I had become disenchanted with *Defeat Autism Now!* based megavitamin protocols. A precious few of the families that we helped recovered their children with diet, supplements, chelation and hyperbarics. However, the overwhelming majority still had an autism diagnosis after tons of work by their parents, and usually a lot of money spent on supplements and treatments. I started feeling like it was a fraud, telling people to follow this protocol, which I knew wasn't going to be enough to recover most of the kids. That's not to say that we weren't seeing vast improvements, but counting all of the children that we worked with, only two children recovered.

By July of 2010, I was totally disillusioned and befuddled and I didn't want to continue doing what I was doing the way I was doing it. So, I asked the Universe/God/Angels—whoever was listening, for help. If my mission truly was to help families recover their children from autism then I was going to need a tool to work with. One that was available on every continent and that was affordable to everyone, because what we had just wasn't doing the job.

No magic voice came, thank God! Because that would have really freaked me out. However, I started to remember these colorful little bottles of chlorine dioxide that I never used. I decided to research their use on Google. Disappointingly, there was absolutely nothing on the Internet about autism and Miracle Mineral Solution (MMS) aka chlorine dioxide (CD). So I started thinking about what autism is made up of. So, I Googled chlorine dioxide with virus, bacteria, candida, heavy metals, blood brain barrier, allergies, and inflammation. The evidence was overwhelmingly positive, which showed me that CD could treat all of the components that make up autism. I had hope once again.

I was especially interested because at the clinic, we specialize in oxidative therapies such as hyperbarics and ozone. As chlorine dioxide is more benign than what we were already using, I decided to investigate further.

No side effects except a possible Herxheimer reaction; which is not a side effect of chlorine dioxide itself, but can happen with any detoxification protocol. I decided to speak to my husband and son Alex, who were also excited. The next day at the clinic, my husband's best friend's cousin and his wife were getting out of the chamber. I said, "Hi!" and she immediately said to me, "I'm taking CD." She didn't say "Hi!" or "Kerri!" just, "I'm taking CD." That was the definitive moment for me—my Aha! moment. I told her I had been researching it for weeks, and I was extremely interested. She was having great results, so my husband said he would try it first. If after three days on the drops he was still alive then we would start Patrick on them.

I contacted Jim Humble, discoverer of CD. I was hoping he would help me to better understand how to dose CD for children with autism. I explained to him that there was nothing on the Internet for kids. He helped me do just that. He gave the following recommendations: 1 drop 8x a day for children under 25lbs, 2 drops 8x a day for children under 50lbs, and 3 drops 8x a day for children under 100 lbs. He told me that the more doses that we can get in one day the better, 8 doses are the minimum.

That first week Patrick vomited (classic Herxheimer reaction), because I went too fast with the dosing. On the Internet, the only protocols I found gave high doses a few times a day, and as I found that week low and slow doses all day was the way to go. However, despite the Herxheimer reaction I caused Patrick (through lack of a low and slow dosing protocol) he was still noticeably better. Seven days later my son had improved eye contact and was asking for things that he had never requested in his life. At 9pm he looked me straight in the eye and said, "I want bed." With my jaw hanging open in disbelief, I followed him upstairs to his room. When we got there he turned to me and looked me straight in the eye again and said, "I want take bath." I knew I was not dreaming, and I had really just heard that. After his bath he looked me straight in the eye and said "I want brush teeth" and the whole time we brushed his teeth he was giggling... so I asked him what he wanted and he said, "I want 'kanket'" so I said, "blanket" and he repeated, "blanket, yes" and ran to the bed, and jumped on it to enjoy the blanket. He had never jumped into the bed in his life before that. This was the first seven days on CD. I was blown away.

By September of 2010, every person that was previously only using herbs or medications to kill viruses, bacteria, candida and other pathogens was about to hear about CD. That's when things really started to happen.

Back in 2007, I learned how Dr. Anju Usman was having great success with her biofilm protocol. She concluded that virus, bacteria, candida, parasites, and heavy metals are all joined together in the biofilm (more on this in Chapter 5, starting on page 117). And when I saw that CD killed pathogens and neutralized heavy metals and so many other things that make up the core of autism, I knew we would kill a lot of birds with one stone. I was also hoping to sidestep a few pharmaceutical pieces of the biofilm protocol, i.e. antifungals, antibiotics, antivirals and be able to use something with no side effects (a Herxheimer reaction is different than a side effect).

I was in the process of discovering how we could use this extremely inexpensive oxidizer that is available all over the world to help the body heal from autism. The other important part of CD is that you don't need to take your child somewhere for treatment, as in the cases of hyperbarics, ozone or IV chelation... there is no doctor needed, or trips to visit doctors in other countries. It's as simple as taking a supplement and you modify your dose depending on what you feel and see. Basically, any family with access to the internet, diet, CD and a few choice supplements can heal their child with autism.

After having limited success with my son on different biomedical treatments, even with the best doctors in the world, it was time for a change. With CD we are attacking the biofilm all day as CD destroys the electron shells of the different molecules making up pathogens, therefore releasing toxins into the bloodstream. This release of toxins is the principal reason one must go slow and build up the dose, to avoid a Herxheimer reaction as so many of these children are very toxic. If we kill too many pathogens at once, too many toxins enter the bloodstream. The body will immediately look to eliminate them, most notably through diarrhea and vomiting. This is unpleasant and totally avoidable.

CD is so benign you can use it on your skin, hair, ears, eyes, orally, rectally, vaginally, inhaled, etc. At the doses we use CD in aqueous solutions; it is not detrimental to healthy cells. It specifically targets pathogens due to their negative charges. Once I understood the basics, and Patrick was still improving, I started to share with others to understand how to use CD. Very quickly we were having success that many other clinicians weren't having. Kids on the spectrum were improving, some started recovering, and we had to sit up and take notice. I was blown away again.

In November, a child recovered and then in December another child recovered. Their families took them to psychiatrists and their doctors to have their diagnosis removed. These were the very important first steps that convinced me that this is something that we must continue to do. We started to spread the word that this was an inexpensive treatment modality available in every continent of the world.

With my background in biomed I learned that you watch for reactions while you load the dose. *Low and slow* is the rule. We arrived at the one-drop at a time dose and as kids recovered, parents shared their stories with other parents and more and more people started using it. It was very grass roots.

This is when the explosion happened. I started to think that this might be the missing piece of the puzzle that we had been looking for. In all seriousness, there is no one cure for all kids with autism, which is why each child's protocol and path to recovery is different. While we have seen great success with CD today I keep working to discover new modalities that help these children heal as non-invasively as possible. Now that we had had success in the Spanish-speaking world, I needed to share these treatments with families of children with autism all over the world.

Around this time something very interesting started happening with the CD enemas... parasites; more specifically roundworms were coming out with the CD enemas of the parents and their children using the Protocol. Today I have hundreds of photos, which were sent to me from parents all over the world (first world and third world nations) who have seen worms passed in stools.

At the moment laboratory testing is woefully inadequate, but a keen veterinarian can easily check for the presence of parasites in a stool sample using a high-powered microscope. The children who have had stool samples reviewed by microscopy have come back positive. Pinworms, roundworms, tapeworms and hookworms are the most common findings. Stool analysis done by laboratories have consistently come back negative even when worms have clearly been seen and photographed and seen under the microscope. In fact, one mother I know sent in a live, moving worm that her child had passed. The result? Negative for parasites! At this point it is simply not enough to trust a coprological analysis when looking for parasites.

Dr. Andreas Kalcker and Miriam Carrasco have been instrumental in this piece of the puzzle and have designed an amazing parasite protocol that has already helped many children including Patrick. In October of 2011, Andreas gave me the first parasite protocol and families in Spain, Mexico, Venezuela and others throughout Latin America began to use it. We'll talk about it in depth in Chapter 8, page 165 and how it has affected my life as well as the lives of so many other families with children on the spectrum.

In January of 2012, I got in touch with Teri Arranga and I was invited to speak at *AutismOne* in May of 2012. This would be the first time I would present at *AutismOne* in English, and obviously the first ever presentation on CD. After eight years of biomed and six years of helping families in Latin America I would be flying above the radar, knowing full well there would be a tradeoff. While I would reach families all over the US, for the first time I would end up taking a lot of flak in the blogosphere. We survived!

When I first came across CD in 2010 and began to watch it work it's miracles with autism, I expected parents, doctors and professionals who dealt with autism to be excited. I assumed they would begin to do research as to how and why the chlorine dioxide molecule was healing/curing autism. Much to my chagrin, many people were disinterested. Some even went so far as to say that what I had seen was impossible or that CD was toxic. Well, to that I would say that healing/curing autism with a toxic substance is impossible. Since then a handful of some of the best doctors in the world have become interested (that number is growing) and quite a few parents are taking note. Hundreds in fact went to my presentation at *AutismOne*. Several told me in hindsight that they had thought about not attending because the title of the presentation sounded too good to be true... *40 Children Recovered in 21 Months*.

The people who did attend were pleased with the information and many began the Protocol. However, what was to happen the days and weeks after my presentation blew my mind and was absolutely beyond my wildest nightmares. Some parents attacked me in print and on the Internet. I received threatening and accusatory emails filled with hate speech and foul language.

In most cases these emails and blogs were from anti-biomed parents. Others in the biomed movement told me not to worry, that these people were notorious for doing this sort of thing to others. They jump on the newest and brightest intervention in order to get their time in the limelight. Stealing attention away from the treatment, altering the truth, and in some cases lying to people about what is happening just to get other parents up in arms. Never could I have imagined anything like what transpired. However, as time passed, so did the threats, negative blog posts, etc.

Update for 2014

One of my favorite quotes states:

> *"All truth passes through three stages. First, it is ridiculed. Second, it is violently opposed. Third, it is accepted as being self-evident."*

Fortunately, the 2013 *AutismOne* conference was a totally different experience—which may indicate we are slipping into that third stage. In May of 2013, we were at 93 recoveries, and during my presentation at *AutismOne,* some courageous parents took the stage with me to share their children's stories of healing and recovery. There were no attacks. As the first edition of this book launched at the conference, many of our wonderful moderators were on hand to answer questions, and assist parents who were interested in getting started. I had a book signing along with the pleasure of meeting many parents I had up until then only known through email or Facebook. By January 2014, the first edition had already sold thousands of copies. If you searched for "autism" in the category of "books" on Amazon, it was showing up in various positions of the first two pages of over 10,000+ results, with the majority being five-star reviews. If you changed the order to "Average Customer Review," it was in the top 10, sometimes in position #1.

This has always been a grassroots parent driven movement, and today help is available in 7 languages online to answer questions and offer support. As with anything, if you attract enough attention, you will also attract some "haters," however CD has already earned itself a place in the treatment modalities that are healing the symptoms known as autism.

CDS (chlorine dioxide solution) was introduced in the first edition, as we were still hoping that it was something better than it turned out to be. It was a better tasting and more tolerated form of CD, and still remains an excellent preparation choice for those who are extremely sensitive, and have trouble tolerating even one drop of classic CD. However, we found that over the long haul, only one child so far has recovered with strict CDS use—the other 114 were with classic CD.

In this edition we are introducing *Chlorine Dioxide Holding [Solution]* (CDH). When this preparation technique was introduced it was touted as something similar to CDS… having a better taste, is better tolerated, along with less Herxheimer reactions. However, there is one big difference; CDH still contains a small amount of the raw materials required for the preparation of CD (sodium chlorite and citric/hydrochloric acid). On the other hand, CDS was chlorine dioxide gas ONLY dissolved in water. That small amount of raw materials in the CDH preparation may be what makes the difference. After 90 days of CDH use with over 70 families, it has not failed. Gains have not plateaued, and parents seem to be having an easier time increasing their children's dose, without any Herxheimer reactions. Another amazing thing about CDH is that the natural sweetener *Stevia* can be added to improve the flavor, while the potency of the dose does not change. This can be a game changer for kids that have taste aversion to classic CD. It should be noted that not all brands of Stevia are created equal and there may be some that can't be used. We are still testing various brands.

CDS and CDH have both earned their place in methods of chlorine dioxide preparation, thus allowing more people to benefit from the healing properties of CD who might not have otherwise been able to tolerate it.

People are always interested to know how my son Patrick is doing, and I am happy to share a little bit about what has been going on in his life lately. This past August 2013, Patrick turned 13. I had expected him to be recovered by now; however, we are still working towards a full recovery. He is better every month and his current ATEC is somewhere between 22 and 24. Patrick is very social, he loves a party. This Halloween, my sister threw a party and it was 11:30pm before he was finally ready to go home. He also loves spending time with his family. Every night he tells me, "I love you Mommy, gimme kissy." That is not only his way of telling me he wants a kiss, but that he wants me to come spend time with him before we go to sleep.

He has been preparing his own food in the kitchen, and while he has always liked to help chop, his being able to place his selection in the toaster oven and heat it himself is new. We didn't show him how to do this. He decided on his own that he would heat and serve himself his dinner one day. Another major advancement is that he is now able to clean himself after toileting, which is something he always asked for help with before. He will even wear headphones when listening to YouTube videos or watching DVDs if someone has to make a phone call.

We don't have any conduct issues, and if no one told you I had a son with autism, and you saw us out, you would never know. Apraxia remains the biggest factor delaying Patrick's recovery. That said, Patrick does communicate more than ever, and attempts more language than ever before.

As far as where I see this movement heading in the future, I believe that if truth does pass through 3 stages then we have finally entered the third stage of "self-evidence." The CD Protocol has now recovered 115 children (as of

December 2013); is used in 58 countries; and has already helped over 5,000 people on the spectrum, with more and more being added every day. The power of social media allows for parents to share with other parents their successes with the protocol, thus forming a stronger bond. Parents in the autism community trust other parents above doctors, and rightfully so.

At this point we are breaking many stereotypes associated with healing autism. For example, we now know that after the age of 9, recovery is still very possible (a 31 year old man is nearing recovery as I write these words). You do not have to be rich to recover your child from autism. You do not have to speak English—there are Facebook groups in 7 languages and this book will be translated into at least 13 languages. We now know that autism is not a psychological disorder. It is biomedical... viruses, bacteria, candida, parasites, and heavy metals cause the behaviors that lead to an autism diagnosis. Once you remove what is causing the symptoms, you can remove the diagnosis.

I witness on a daily basis what was not supposed to be possible: the healing of autism.

The future is bright, and it is up to us to share it!

Best in health,

Kerri

Alex, Kerri & Patrick Rivera

Fundación Venciendo

Jim, Andreas and myself in Venezuela where 2 of the most amazing women in the world have a foundation to help Venezuela heal from the autism epidemic. I am so grateful to Yamileth Paduani and Carolina Moreno for their foundation and hard work. 28 recovered children from their foundation with this protocol in their first year of service. Thank you ladies. I am so proud to be your sister of the heart.

Healing
the symptoms known as
Autism

Ty reading and healing

Chapter 2

Yes We Can!!!

Without faith, nothing is possible. With it, nothing is impossible.
~ Mary McLeod Bethune

Before we embark on the path of healing autism, it may provide you some comfort and encouragement to hear from those who have gone before you.

Some of the parents of the children who lost their diagnosis through this Protocol were generous enough to share a photo of their children along with a "thank you note." You will see that a couple of the older children wanted to write their own notes and share their success with the world.

①

My son gets to
play little League!

Thanks to ATEC of 4,
autism IS treatable!

THANK YOU
KERRI!

②

OMG! I'm still rubbing my eyes. Nathan just played with cars appropriately! He pushed them along while saying, "vroom vroom!" Until now, he would simply mouth, spin or tap toys.

-Week 5 of MMS

3

Today in my son's school they had an "I Can Run" day for good health and to collect food for families in our community. Last year he couldn't make it 1 mile - they run / walk around a track. This year he ran 2 miles... Whhaattt? 7 months on mms and 6 months treating parasites.

4

I love you Miss Kerri,
Thank you for giving me CD.
Matthew

5

Wow! Here is something to be grateful for! It was time to do my son's one year ATEC doing MMS (7 months detoxing parasites) and his score was...drum roll please...15! You read it right, a 15! I am tearing up right now writing this. My son is 14 years, 4 months old. He was at a 27 when we started, then went up to a 34 when we started treating parasites (behaviors worsened), then down to 21 three months ago, and now a 15! Thank you so much Kerri! Without you, who knows where he'd be.

6

Before MMS we were able to get my sons atec score down to a 24. Although he mainly functioned as a "neurotypical" child, he still would have behavioral problems. He was easily agitated and would struggle with some anxiety and OCD. We were constantly battling yeast, constipation, metals and come to find out, parasites! We began MMS very slowly. We noticed his moods were improving. He was smiling more and over all just really pleasant to be around. We decided to retake his atec after only 1 month. It was a shocking 4!!! It dropped 20 points in 1 month!!! We have our boy back and will forever be grateful! We will be starting the parasite protocol this month and looking forward to an atec of 0!!!

7

I just wanted to share that today my daughter said a full sentence!

She normally only makes two word sentences and often stutters and stammers, with a lot of articulation problems.

My husband came home after a week away for work and she walked up to him as he sat at the computer and she said: "I want... to sit with daddy."

And in the bathtub (I'm doing the MMS steambath as both my kids are coughing right now), she turned to her brother and said "Look brother, cup." and she held the plastic cup in front of him. She has never been this coordinated and 'present', nor articulate. I am so grateful.

Even though we still have a long way to go, I feel like a huge weight has been lifted off my shoulders.
Thank you for your hard work.

8

Thank you for helping me to rescue my son. God bless you Kerri Rivera!

9

My son is at a three-day camp with 100 NT 5th graders! We have provided all his food, and go up to give supplements twice a day... other than that, he is on his own sleeping with friends for the first time in his life! I have no words other than Praise The Lord for bringing me Kerri Rivera and also for giving me and my family the strength to never give up! Our little bird had taken flight... thanks be to God!

Thank you God for your fidelity, thank you my beautiful girl for giving me your gaze and your smile every day, thank you Kerri McDaniel de Rivera for walking with me.

⑪

Hello, I'm Silvia, Alejandro's mom, we live in Spain. I was giving MMS to my son for 10 months without seeing any changes. Kerri told me that I wasn't doing something right... and that was correct. Being the warrior she is, she made me tell her everything I was giving Alejandro and she got it, Alejandro drank pineapple juice all day long, he would take more than a liter a day and that made the MMS not work, I stopped giving him pineapple juice 2 months ago, and he is a new child... He pays attention, his comprehension is almost 100% I must say my baby did not speak, but with this he started saying: "Come on, mom!" "My mommy" All this because I started giving him the MMS as I should. I want to say NO juice near MMS because it anulls its effectiveness. I thank God for putting Kerri in my path. My son improves everyday, slowly but you can see that he is saying goodbye to his little world. Thank you Kerri for being in our lives and for making us see the light. Also thank you JIM HUMBLE for giving hope to our lives.

⑫

Thank God for putting Kerri Rivera in our path, and giving us our prince back, totally recovered.

⑬

My son turned 12 today - hooray!! This is truly the very first time since his 1st birthday that my heart didn't shatter into a million pieces because it's another year down the road, and still so far away from recovery!! He decided that he was staying at home today to play with his birthday toys, he told his teacher (several times over), and when she asked if she could take the day off too, he said "nooooo!!. When his grandpa called from Holland to sing Happy Birthday, he sat quietly listening with a smile on his face - the usual is "no singing!" Besides that - here's my brag for the day - his teacher told me that she had to move him and his best class buddy away from each other because they were copying each other's work!!! YAHOO - we were SO excited - yes I know, we're crazy, but hey - if it's typical, I don't care what it looks like, I WILL TAKE IT!!! If you think this is weird, you should've seen me when he bit a child when he was 3!! I want to encourage you all today, look for the small things, check in with the people who are in contact with your kids and keep a journal of the changes. You will be amazed at how much we miss, because we spend every waking moment watching over them! I feel SO blessed!!

14

Dear Kerri,

Thank you so much for bringing our son's smile to life. The happiness you see in these eyes comes from listening to you and your amazing protocol.

We were stunned when his ATEC dropped 18 points in 3 weeks, and has now reached a staggering 1 from the original 36 in less than one year.

I thank God every day for having me in that AutismOne lecture in 2012 where you opened my eyes to what else we could do to help him. I want every parent out there to know this is a real Protocol with real results.

Thank you again and God Bless everything you do for our Children.

You Rock! Love Maryann

* * *

15

Greetings from Monterrey.

Just to let you know, we went to visit you in Puerto Vallarta on June 21st, and my child started treatment in July with MMS. As of today my son eats almost all on his own, and wipes his mouth whenever he needs to. For about a month now he has been doing funny things; like covering up his cards when we play as if to hide them, if he sees me squatting on the floor he runs over and jumps on me. He shows expectation on his face when he knows that I am coming to tickle him. He hides behind a wall when we play hide-and-seek.

In both of the institutes where he goes they have told me he is able to pay attention for longer periods of time, his way of being has changed, and they have also said that his eye contact is much better. According to his therapist he is now a candidate for speech therapy because he is able to pay attention.

For us, Kerri, this is a miracle come true, to see our son waking up little by little and to see him eat all on his own. Thank you so much for sharing all of this with us.

⑯

Since starting MMS my daughter's play skills improved, she actually looks into my eyes and gives me eye contact which means a lot. She started answering questions, more engaging and now I started to feel as if the end of the tunnel is not too far. I am so grateful to Kerri for showing me a way to recover my child. GOD BLESS.

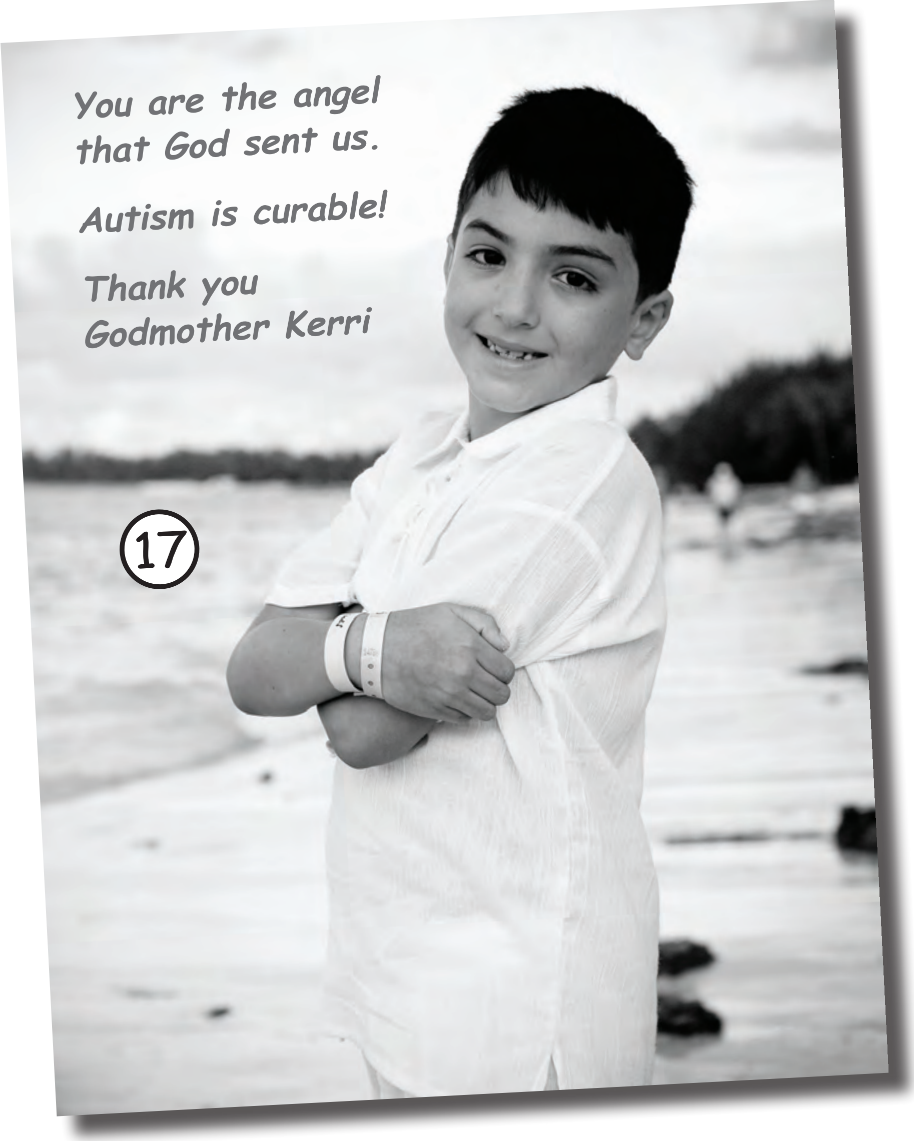

⑰

You are the angel that God sent us.

Autism is curable!

Thank you Godmother Kerri

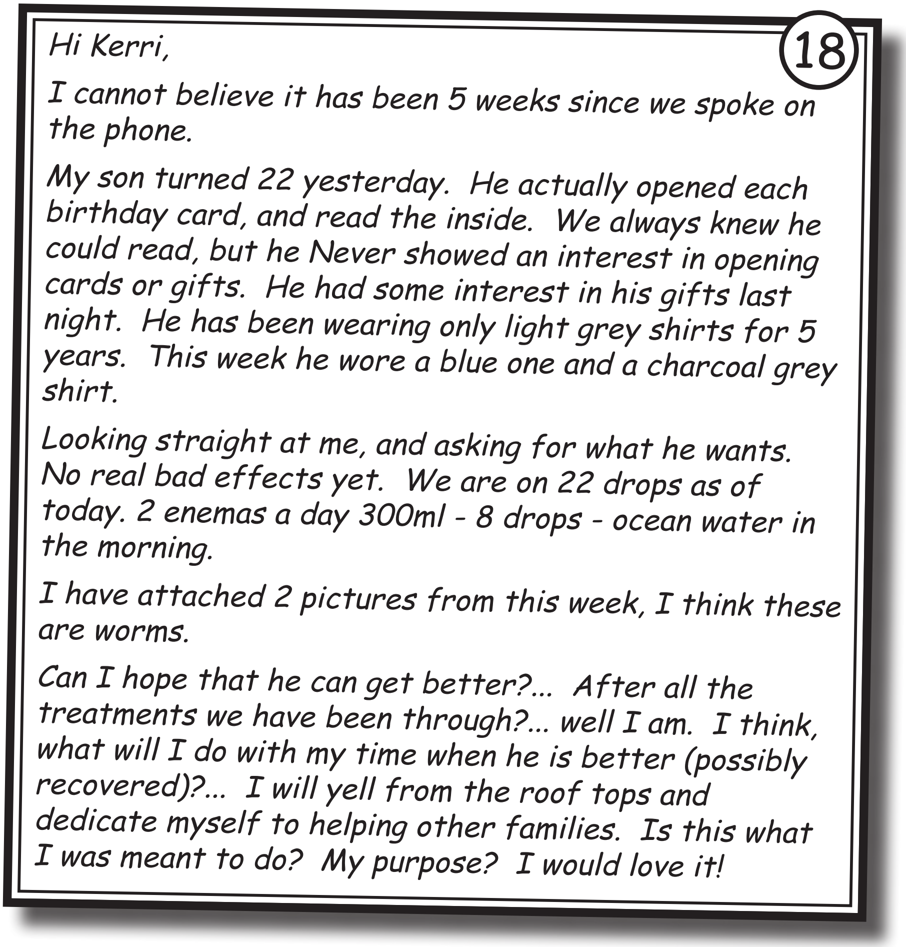

18

Hi Kerri,

I cannot believe it has been 5 weeks since we spoke on the phone.

My son turned 22 yesterday. He actually opened each birthday card, and read the inside. We always knew he could read, but he Never showed an interest in opening cards or gifts. He had some interest in his gifts last night. He has been wearing only light grey shirts for 5 years. This week he wore a blue one and a charcoal grey shirt.

Looking straight at me, and asking for what he wants. No real bad effects yet. We are on 22 drops as of today. 2 enemas a day 300ml - 8 drops - ocean water in the morning.

I have attached 2 pictures from this week, I think these are worms.

Can I hope that he can get better?... After all the treatments we have been through?... well I am. I think, what will I do with my time when he is better (possibly recovered)?... I will yell from the roof tops and dedicate myself to helping other families. Is this what I was meant to do? My purpose? I would love it!

19

Thank you for this look.
I share this look with
you, the same look
that you have given
back to my Mom, this
photo would not be, if
it hadn't been for you.
I love you very much,
your Godson Gabriel

⑳

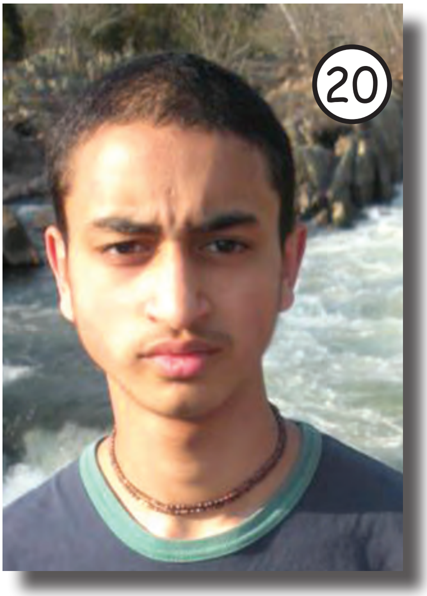

My family and I are very grateful to Kerri Rivera for making available the CD protocol to treat ASD which has recovered myself and made a huge difference to thousands of kids and families all over the world.

CD can recover even a teenager like me whom others had given up on. Of all the bio-medical interventions, CD cost the least and made the biggest difference.

Muchas Gracias por salvar mi vida Kerri.

㉑

This is Benjamin. 4 years 5 months old, current atec of 4 and dropping weekly :) Diagnosis atec at 3 years 2 months was 134. Atec when starting CD protocol was 18 and was stagnant after full on bio-med/diet intervention. Dropped all supps except melatonin when starting protocol and obviously can't be happier with that decision. Thank you to Kerri Rivera of course for giving me a way to save my baby boy... We will continue with protocol until atec = 0 :)

22

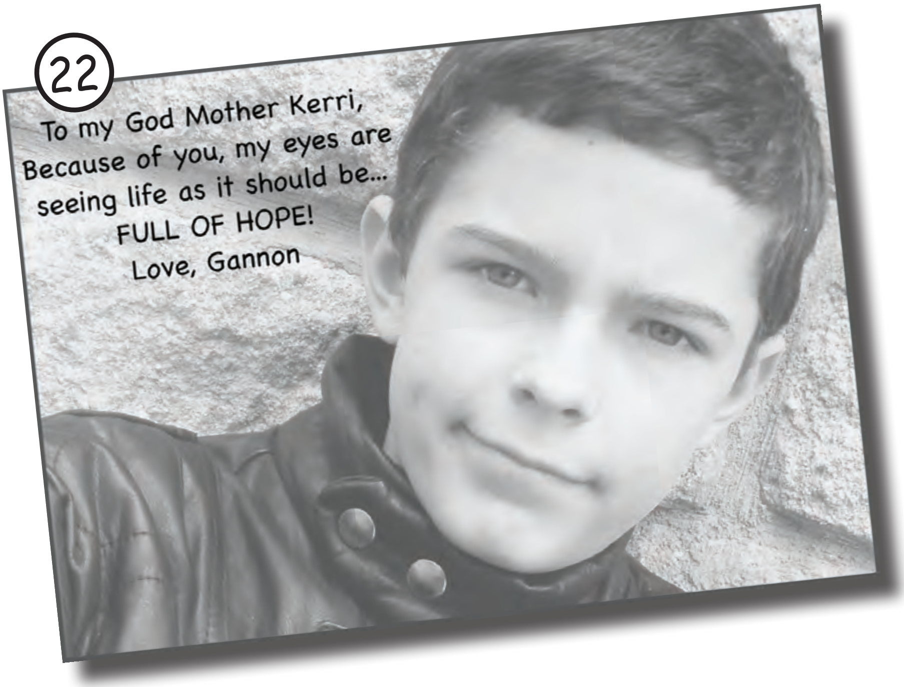

To my God Mother Kerri,
Because of you, my eyes are
seeing life as it should be...
FULL OF HOPE!
Love, Gannon

23

My sons started using MMS exactly 2 and a half months ago and we saw improvements almost immediately, my twins attention got better just days after we started the treatment and at 10 days after starting my son Juan Pablo started calling me mom! He has never done that before. (Neither of my twins were verbal, meaning they were not able to speak a single word.) A few days later Jesus Alejandro said "Mom, water" while pointing to the fridge, it was shocking for me.

My two sons can now say 4 or 5 words and their communication skills have gotten a lot better, even if it's not verbal they are very good at expressing what they want. There is a lot of improvement in comprehension, following orders and accomplishing small tasks, they react when you say their names, they are controlling their bowel movements, and in general their whole quality of life has gotten so much better since they started with MMS.

I know we have a long road to go but we are on the right track and I trust that thanks to the science that discovered the great effects of this wonderful formula for our children with autism we will recover them, as it truly does detoxify!

Ing. Artebys Cedeño
Mother of Jesus Alejandro and Juan Pablo

OAHU
Surf Academy

24

Dear Godmother Kerri,

How can we ever thank you enough for crushing autism out of TJ's life? We now have a smart, happy, helathy boy that has great confidence in himself!

We are forever grateful to you!!!

25

We started using MMS a month ago. We are not even at the full dose yet but have been seeing some amazing gains.

I took my son to the mall about a week ago because it is one of his favorite things to do. When we got home he did something he has NEVER done before. He came up to me, hugged me and said "Thanks". Wow.

Two days ago my husband had to go to the local farm supply store because something broke in the barn. So we asked my son if he wanted to go and he literally jumped off the couch and excitedly came with us. While he was there he picked out a couple of books he was interested in and a couple of small vehicle toys. When we got home I was helping him with his coat and boots he said "you're great mom". Never have I ever had spontaneous comments like this and the only thing I can attribute it to is the MMS because we stopped everything else. This gives me inspiration to continue the protocol and I look forward to what the future holds for my son and our family.

Thank you Kerri!

26

If it were not for you Kerri, my mother would still be worried about my future.

You are my Guardian Angel!!!

Autism is curable!

Thank you Kerri
Gracias Kerri!

27

His speech therapist told me today that his sentences are so much better, more organized, also his ability to converse back and forth and also his concentration have improved!! Sooooooooo..... Yay!
Oh, before we went to see his speech therapist he said: Mummy, my speech is getting so much better!
Yup, he said that! He knows!

28

I can't believe it! I just did my daughter's ATEC and in one month it dropped from 72 to 48!!! I am so shocked! Is this a dream? We revised the scores over and over with my husband and accepting that some of our answers could be just very positive feelings, still the improvement is huge!! I am so thankful Kerri Rivera! I have no words!! I am praying this is not a dream!! THANK YOU SO MUCH!!!! I believe there is more to come but right now I am so much in shock!

㉙

I have to post our Sunday gain, my daughter is and always has been on sensory overload, those of you who deal with this know that teeth brushing, hair combing and showers can be torturous for child AND parent. I literally dread washing her hair. She screams so loud she could break the windows. She smacks her wet thighs so hard that they've been bruised for years. This morning she gets in the tub, sits down and sees me roll my sleeves up (which is usually the moment she loses her mind because she knows I'm about to grab the shower head), she stood up, looked me dead in the eye, smiled at me and said "R. shower". I began to wet her hair down, still convinced that the screaming was about to begin any minute. I started shampooing and she said "well done", I smiled and said "VERY we'll done baby". My husband peeked in to see why it was so quiet and she looked at him and said "Daddy shower", smiling from ear to ear as to say 'I'm taking a shower Daddy'..... what a great way to start our Sunday. HUGE, HUGE deal for her and for us. I feel like I finally exhaled today. MMS rocks and so do all of you!!! Happy Sunday ;-))))

㉚

Dear Kerri,

Thank you SO MUCH for helping us with our daughter's complicated health issues. The CD/PP protocol allowed my daughter to sleep through the night, for the FIRST time in 7 years! One month on the parasite meds, she began sleeping perfectly and has slept perfectly ever since!

We are FOREVER grateful! Also, the more her pathogen load decreased on the protocol, the more foods she would eat and LIKE!

My daughter eats SALAD now! AND will try any food I give her. I thank you from the bottom of my heart and will be forever grateful!

With love,

The WHOLE Clark Family!

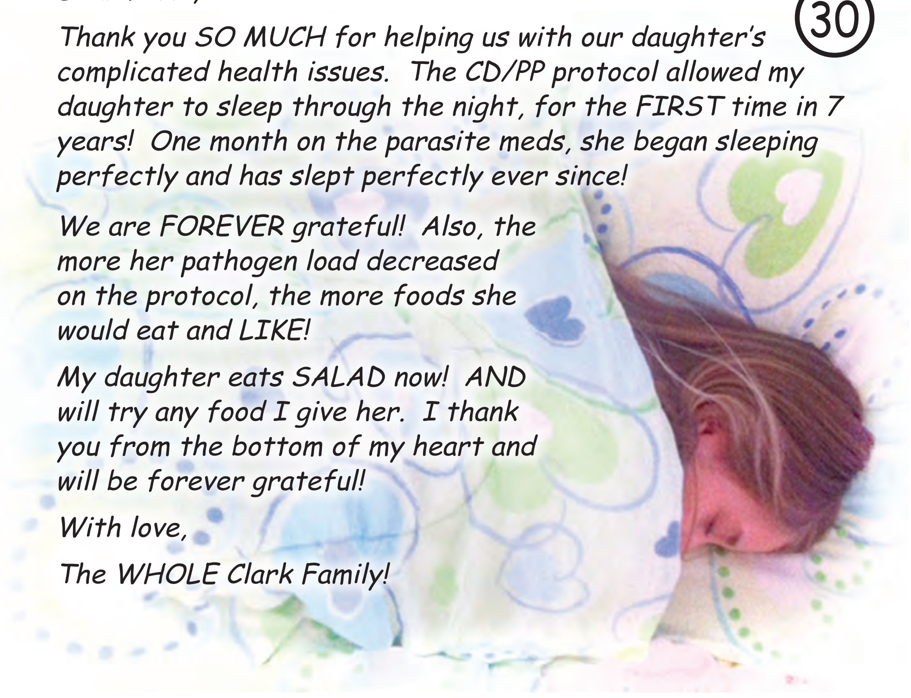

㉛

I was cured,
Yes we can.

I love you
very much
Godmother
Kerri!

㉜

My son had a great week - yesterday we went bowling as a family. It's difficult with him to notice gains sometimes because he is really close - his ATEC was a 7 last time I did it and the changes in him may be subtle and harder to notice. While we were bowling he was calm (despite the loud music, lights and commotion) he sat in between his turn watching everyone, (and the music videos playing) knew when it was his turn and cheered for me (I was beating my husband at the time - lol). He carried his ball down the lane and swung it with one arm - omg!! It was the most NT I have seem him! He now is a brown belt in Tae Kwon Do and this was the first time he was able to swing the ball - I used to have to carry the ball to the ramp, help him lift it and count to 3 for him to push the ball. He was diagnosed hypotonic at 3.5 years, MMS and PP are what helped him get some strength, energy and endurance. I increased his MMS by a drop about 2 weeks ago and we are also ramping up on GcMaf.

Great FUA moment today. My son took his first ever lesson.... swim lessons. He did amazing!!! Followed all instructions, waited in line, smiled, talked, and let the instructors help him. Amazing.... I cried almost the entire time. At one point he smiled, waved, and said "hi mommy." ㉝

㉞

Thank you Kerri! Because we followed the protocol with obviously great results.

㉟

My son has continued to improve on his handwriting and his anger has melted away. We had an IEP today and we lost 2.5 hours of services because he has done so well he does not need those services. They were surprised at how well he is doing. And he fed his little brother soup tonight to get him to eat his dinner.

Thank you Kerri for
helping us to recover our son!

㊱

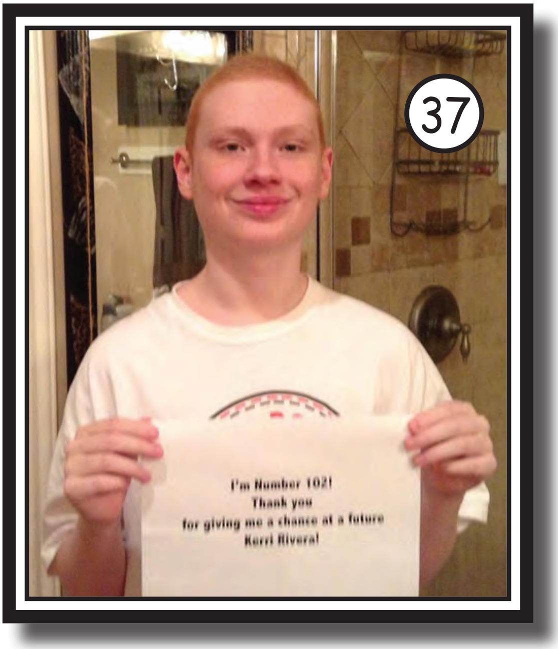

I'm Number 102!
Thank you for giving me a chance at a future
Kerri Rivera!

On any given day, I probably rattle out the words I LOVE YOU, directed at my boys at least a dozen times. My NT boy almost always will reply back. My ASD boy usually says, YES YOU DO as his answer. And I have completely accepted that that is his way of receiving my I love yous.

On very rare occasion he will say I love you to me...

Well, by now you know where this is going.

This morning I was walking past him and he said to me.... I LOVE YOU!

I wasn't at first sure what he said as he didn't shout it, So I asked him, what was that?

He repeated, I love you.

I said thank you son, I love you too!!!

38

Wait for it...

Bammity bam bam BAM!!!

Celebrating every little big thing. Appreciating my 3 favorite words. Loving every minute of every day. Even loving those big red lips this morning knowing he is detoxing...

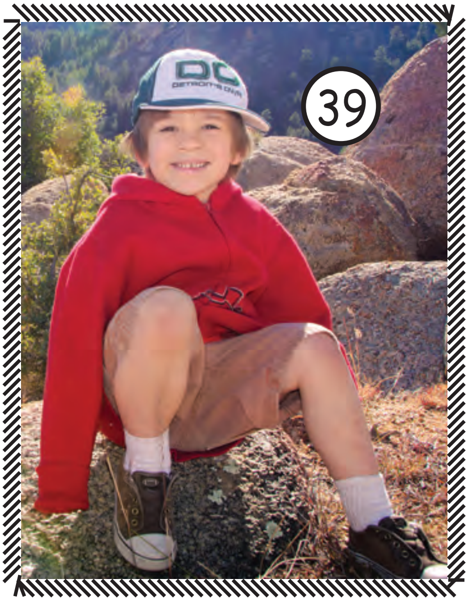

Our son was quickly slipping away from us. We were lost. This protocol gave us direction and ultimately brought our son back to us. He looks at us with smiling eyes and tells us he loves us. He has a full childhood now. Words could never convey the magnitude of our gratitude for Kerri and all who support her efforts to make a future for our kids. See Gunnar's full testimonial on page 378.

Gracias Madrina Kerri!! I now have an ATEC of 10!

㊶

We started the DAN protocol when our 10 year old daughter was 18 months old and remained on it for 9 years. Before starting CD, our daughter appeared neurotypical, but she was heavily dependent on supplements and her progress had stagnated for a few years. She had a distended belly, poor growth, severe constipation, anxiety, a touch of OCD, and focus/concentration/fogginess issues.

Her ATEC before starting CD was 24. Just 26 days later, her ATEC is 7. A 17 point drop in 26 days! We haven't even had time to do a parasite protocol or a 72/2 weekend yet! Her belly is much flatter and her anxiety, minor OCD, and focus issues are greatly improved! We're going to keep her on CD and I plan on starting the protocol myself this week. As a friend said, "DAN got her to third base and CD brought her home!"

㊷

My son just saw me reading this page, and asked what CD Autism was. I told him it was the medicine he's taking. He said, "I think my autism is gone. How do you know someone has it still?" Soon, baby, soon!

Spiro and Peter (9 year old twins)

After years of doing many interventions for our twin boys, and having moderate success at best, we learned about CD. It took about a year before we felt comfortable enough to try it. I can honestly say that, it is the BEST intervention that we've used. We started CDS in mid July 2013. In the 4 short months that we've been using CDS and now CD, we have had more gains in every area, than in the last 6 years of doing so many different treatments. What I love about this protocol, is the simplicity of it, less is more, and that was a huge change in thinking for me.

Just to give you an idea on how well our boys are doing, here are their ATEC scores so far:

Spiro: He was number 100 to recover on the "Healing Train" and the 1st child to recover using CDS.

ATEC prior to starting CDS - 22.

ATEC one month after starting CDS - 16.

ATEC two months after starting CDS - 9.

Peter: He was pre verbal prior to CDS and now uses single words and the odd 2 word combination to communicate. Well on his way to healing.

ATEC prior to starting CDS - 69.

ATEC one month after starting CDS - 55.

ATEC two months after starting CDS - 46.

We finally know, deep in our hearts, that our boys are well on their way towards "true healing". We don't have that stress and anxiety about what the future holds anymore. We finally feel like we can breathe easier, and that day by day, the gains keep coming consistently. The gains stay, and that feels better than anything.

Thank you, to my FB friends who introduced me to MMS and to Kerri Rivera who we have such admiration and respect for. I thank GOD everyday that we were fortunate enough to find out about this treatment. Our boys are coming back to us and their beautiful personalities are emerging. We now look to the future with such HOPE. Kerri, we will forever be grateful to you. xoxoxoxoxoxoxoxoxoxo

Tim Hortons
PICKERING SOCCER CLUB

㊸

㊹

Whoa... My severe kid walked in the door and wrapped his arms around my neck and gave me a normal hug for the first time in his life (6.5 years). Umm... This is crazy. He wiped his butt for the first time EVER... no prompting, no asking! I wasn't even in the room with him! When I got excited and asked, "Did you wipe???" I swear to GOD... I heard "Yeah, Mom!" Kid's never said a word ever. Could be dreaming, but WOW!!! What a day!!!

Please share your experiences!

Testimonials are one of the best ways to share your experiences with this protocol. Perhaps you learned about it by reading or watching a video testimonial?

If you don't tell us your experiences, we can't share them or take action on issues that need improving or correcting...

Send your testimonials to:

testimonials@cdautism.org

Also, let us know if we are free to publish your testimonial, with or without your name.

More miracles & testimonials starting on page 358.

Chapter 3

Step 1 – The Diet

"Let food be thy medicine and medicine be thy food."

~ Hippocrates

If I could choose the one part of *The Protocol* that would be the single most important piece of the recovery puzzle it would be—*The Diet*. By *The Diet,* I mean the dietary plan that I recommend to all of the families that want to start *The Protocol*. It is a combination of the classic gluten-free, casein-free, and soy-free diet along with the elimination of sugar, corn syrup, coloring, preservatives, and other harmful foods. *The Diet* is the basis for the rest of The Protocol; similar to laying the foundation of a house that the rest of the structure will rest on. Adhering to *The Diet* is critical to the effectiveness of the rest of *The Protocol*.

When a family of a child with autism comes to see me, the first thing they want to know is:

What can I do to help my child?

I always, always start with *The Diet*. In fact, I send them away after that first meeting with hope that they will be able to recover their child, but only if they commit to *The Diet* 100%. I explain to the families that they must think about food the same way their great grandparents thought about food. In generations past, food came directly from the Earth, with little processing. Fruits, vegetables, nuts, and meats were dietary staples for our great grandparents and they should be for our children as well. We must think whole foods and not ~~processed foods~~. There is no point in going from regular junk food to gluten-free/casein-free junk food!

After my first meeting with a family, they leave with the list of permitted foods and I have them email me when they have a week straight with zero exceptions or "errors" in *The Diet*. That is where the parents who are truly hungry for recovery are separated from those who are interested in having someone else "fix" their child for them.

One of two things generally happens during that first week. In the first scenario, I get an email from an ecstatic mother or father saying, "I can barely believe it, Johnny slept through the night for the first time in years," or "Johnny had a normal bowel movement," or "Johnny said two new words yesterday!" That's what I hope for.

The second scenario is... I never hear from the family again. There may have been tantrums, or an adjustment period related to *The Diet* that proved too much for them, and they decided to pursue another avenue. That's not to say that everyone sees a miracle, or that there's no middle ground. Some of the results are less obvious, such as more eye contact or less redness in the face, but generally speaking, we see positive changes. Any change is a good sign. *The Diet* is only the first piece of the puzzle. We must continue from here, layering in interventions until we get the desired, end result.

> Kerri... Bless you! When my son was 3 years old he was prescribed Ritalin®, Risperdol® and Clonazepam... can you believe it? Obviously I never gave them to him, it hurt me to see my child "drugged." 2 days after starting the diet he slept through the night, the sparkle is back in his eyes... I can't wait to start with MMS, I am so happy!!!

Throughout this book "miracles," which are actual emails and forum posts from parents about their children's improvements, have been set in text boxes (like the one above) and placed in the chapters that they accent. Please note that as many of these miracles were collected some time ago the contributors have used the term MMS instead of CD. Due to time constraints it was impossible to request permission from these contributors to change the acronym of MMS to CD in their testimonials. MMS and CD are the same substance. Where the acronym MMS is used please know that chlorine dioxide is the substance being discussed and is responsible for the healing.

Author's note: I'm not against all prescription drugs, especially those necessary to bring about healing. However, I don't condone medicating a child with drugs, masking the symptoms known as autism. So, when I see an email like this, I know we've got a great chance to heal that child because we have a parent who has committed herself to healing her child, and who is excited to see what changes the next tool will bring.

Your child's doctor may not have heard of *The Diet* or may be misinformed of its benefits. The references section at the end of the book lists several studies and articles discussing dietary intervention for ASD's. You can use them as a jumping off point for your own study. If you are consulting with a doctor, it is important to choose someone who is familiar with autism and the recovery of autism.

Many doctors do not have time to study up on what's new in autism recovery and if they have zero recoveries then they are in dire need of overhauling their protocols. Another major problem in the mainstream medical community is gigantic egos. If what we are doing is not working to recover our children, then we need to look at what we are doing and why we are doing it. Let's check our egos, update our protocols, and look at what has worked for families who have recovered their children.

Researchers at the New Jersey Medical School's Autism Center found that "children with autism were more likely to have abnormal immune responses to milk, soy, and wheat than typically developing children," which is published in a chapter of Cutting-Edge Therapies for Autism 2011-2012, by Siri and Lyons.[1] In addition, interest is growing in the study of the link between autism and gastrointestinal (GI) ailments. Siri and Lyons also relate a study by the University of California, Davis Health System, where they found that children with autism born in the 1990s were more likely to have gastrointestinal problems, including constipation, diarrhea, and vomiting than children with autism who were born in the early 1980s.

If your doctor is uninformed, or tells you there is no evidence to prove that *The Diet* will help your child, do the research for yourself since only you are in charge of your child's diet. Heck, why not do *The Diet*? It doesn't cost you any money and it just might help your child heal. Whether he eats cheesy puffs or fruits and vegetables will ultimately come down to you. You are the one with the money in your pocket. The only way to know for sure if your child is going to be one of the individuals who recovers with *The Diet*, is to try. It can take gluten six months or more to be removed from the microvilli or "shag carpeting" of the small intestine. As I mentioned previously, some children have obvious changes in two to three days, but even if your child's evolution is taking a little longer than most, DO NOT GIVE UP! At the writing of this book, I have helped about 5,000 or more families of children with autism; all of the children who have lost their diagnosis, as well as those coming close to recovery, have used varied protocols depending on their symptoms. The one thing that they all—unequivocally—have in common is: ***The Diet!***

In my opinion, it makes little difference what other interventions you apply to your child if you can't manage *The Diet* 24/7/365. Hyperbarics, chelation, ABA, etc., simply will not have the desired effects if you are still feeding "drugs" (aka garbage foods) to your child.

Why do I say drugs? Because that is what gluten and casein become in the bodies of our children on the spectrum; more specifically, gluteomorphin (also called gliadorphin) and casomorphin, which are similar to morphine. Gluteomorphin and casomorphin are produced in the gut due to improper digestion of peptides (as we will explain more in detail later). In a person with "leaky gut syndrome" (increased intestinal permeability), they are able to leave the intestine and cross the blood-brain barrier where they act exactly like morphine or heroin. Would you purposely give your child street drugs? NO!

Once we have this information and understand the severity of this issue, we have a responsibility to our children to do better. We must take away the foods that are keeping them ill. Invite your family and your child's school to help heal your child's ailing body. Explain that they can no longer give your child these items, and that if they do, it's like giving them a dose of morphine. If it sounds drastic... that's because it is!

Researchers have found an abnormal amount of these undigested peptides (gluteomorphin/casomorphin) in the urine of children with autism, proving their existence in the body. Among others, Dr. Knivsberg and colleagues in Norway have found that urine samples from people with autism, PDD, celiac disease, and schizophrenia contained high amounts of the casomorphin peptide.[2] Similarly, Gliadorphin (gluteomorphin) has been verified by mass spectrometry techniques to be present in unusual quantities in urine samples of children with autism.[3]

Do *The Diet*!

There are still plenty of food options that are permitted. I promise you—your child won't starve! We have included some recipes in Appendix 15 on page 513 to get you started. *The Diet* is absolutely the most important piece of the puzzle. If we can't remove what is directly linked to brain and gut inflammation, as well as immune-allergic reactions to offending foods, it is nearly impossible to heal a child on the spectrum. I have personally never seen a family recover a child without dietary intervention. That's not to say it hasn't happened, but I have never seen or heard of it.

Like I said before, when we started with my son Patrick, his only legal food was homemade French fries. So that's what he ate. Little by little he began to accept more foods. I assure you, the adjustment period will fade and they do eat. Stop the cycle of inflammation and addiction. Only then will your child begin to heal, and once your child starts to feel better he will accept more foods.

In the case of my own son, his acidic diarrhea and sleepless nights stopped the week we started *The Diet*. From that moment on I was hooked, not only on *The Diet*, but also on biomedical treatments for curing autism. I have never looked back. I strongly encourage you to experience *The Diet* for your child, or even as a family. It is an amazing thing.

> I thought the following might encourage some of you who have recently made big dietary changes for your kids and are struggling with all that involves. I put my son on a GF/CF/SF… well, basically, EVERYTHING-free diet when he was 11 months old, following a major vaccine-induced crash. My choice to put him on a diet was criticized by nearly everyone in my life, beginning with his pediatrician who told me I was "just imagining" that he'd regressed and that he "probably wasn't allergic" to what I had taken out; she called a week later with the test results and an apology, and a long list of other things he was allergic to! He's now just a little over seven and doing GREAT. I just delivered a GF/CF muffin to him, and he said, "You know what I love best about you, Mommy? I love that you are so nice to us, and that you always make sure we have food that doesn't make us sick." So, don't feel bad about not giving your kid that ice cream cone when Mr. Softee comes by or not loading their Easter basket up with food dye this weekend. It may take years, but they will see that you loved them enough to NOT give them these things.

Some other tips include the following:

Review Your Child's Supplement List

Vitamin A
Vitamin B
Folic acid
Vitamin C
Vitamin D
Vitamin E
Vitamin K
Minerals

Is your child taking 30+ supplements a day? We will talk more about this throughout the book, but this protocol is geared towards eliminating excesses and not about supplementing deficiencies. It's important to carefully review your child's supplements. In order to maximize the benefits from this protocol consider removing any antioxidants, (they kill CD—more on that in Chapter 5, page 88), calcium, magnesium (which feeds the biofilm), as well as iron and B_{12} (which are favorite foods of parasites). Supplements for increasing speech, reducing seizures, enzymes, and probiotics obviously have their place in individual protocols. B_{12}, while known to feed parasites, has also been known to greatly boost speech in some children. If your child saw benefits in speech from B_{12}, it may be worth keeping in. As always, these are decisions that each family has to make for their child on an individual basis.

If a certain supplement proved beneficial to your child, then don't remove it (the exception being antioxidants and the other aforementioned supplements). The goal of this step is to remove unnecessary supplements as they are most likely feeding parasites and creating "excess noise" in the body. Anything that is not serving a purpose needs to be eliminated.

Keep a Journal

Keep a record of the foods you have removed and those you have added in. Then, take note of the types of reactions you observe: rash, more or less hyper, more or less stimmy (self-stimulatory behavior), sleep patterns, frequency and consistency of bowel movements, tantrums, acceptance of new foods, eye contact, language, etc. Write everything down so that you can follow any relationships or patterns that will allow you to identify when something is working, or when it is not. Record these observations in a notebook because it helps guide us, especially those of us who are stressed, frazzled, and sleep deprived.

Don't Stress About Raw Foods or Perfectly Balanced Meals...Yet!

When you are first transitioning, feed your child whatever legal foods he/she will eat. We are taking baby steps. Once your child is on *The Diet* 100%, you can start adding in new foods, bite by bite, if necessary.

Do the ATEC

Keep track of your progress with the ATEC checklist (see page 447), which you can find at...

www.autism.com/index.php/ind_atec

This is an excellent way to see how you are doing. Every time that you begin a new intervention, it's a good idea to do the ATEC, and then repeat it every three months or so to see how your child is evolving. Sometimes our children are recovering right before our very eyes and we don't even know it. The ATEC can be used to measure the effects of all new interventions, not just *The Diet*.

Using the ATEC is a great way to measure improvement. Many of us are tired, burned out, or don't have the best memory. A formal questionnaire can help us to discern when an intervention is working or not, and it only takes about seven minutes to fill out. See Appendix 4, page 447 for a sample survey and more information about ATEC.

Always Read Labels

Read the label on the back of all packages—not just the one on the front that says "gluten-free." Many times they have sugar, yeast, carrageenan, or other items that are not allowed. We must know what we are putting into our children's bodies. If you can't pronounce it... you don't want to put it into your child. Be careful of hidden offenders like malt, natural flavors, artificial flavors, whey, and numbers (red 40, E-441), etc.

Have a Support System

It doesn't matter whether it is a friend, family member, rescue angel, or neighbor. Have a shoulder to lean on. Autism recovery is a marathon not a sprint, and no one should have to go it alone. There are so many amazing parents out there who have walked this path and are willing to help newbies! As of this writing we have several public forums open, including...

www.cdautism.org

and

www.facebook.com/groups/AutismCD

This is an excellent resource for finding the latest info, protocol changes, and share hope and/or frustration with parents all over the world who are walking the same path. Don't forget to sign up for our newsletter as well by visiting the home page at www.cdautism.org!

Take it One Day at a Time

One day
at a
time

Defeatist thinking will only harm your resolve to help your child "I can't do this for the rest of Johnny's life!" or "How will I get through this year?" are defeatist thoughts. Go hour-by-hour or minute-by-minute, and know there are victories every day in the world of autism. We must draw strength from those victories, even when they aren't our own. Both this book and my website have a section of miracles and testimonials (see Chapter 2, page 19 and Appendix 1, page 357), which are real correspondence I have received from parents describing the advances their children have made as a result of doing *The Protocol*. We receive these miracle emails and forum posts every day, but we had to pick and choose which ones to share in the book. You will also see them sprinkled throughout the chapters of this book. Take some time to read them, know that children are recovering every day, and believe your child can be next. If something doesn't go as planned, wake up the next day and know that, *THIS IS the first day of autism recovery.*

Why Does My Child on the Spectrum Need to Go on a Gluten-Free/Casein-Free/Soy-Free Diet?

As I said before I have never seen a child recover without *The Diet*. The following information comes from *www.gfcfdiet.com* and explains why *The Diet* is so important for people with autism.

Scientific studies have shown the presence of high levels of peptides derived from casein and gluten proteins.[4,5] The digestion process is considered "normal" in terms of protein digestion for most people, as they make smaller particles called peptides that are further broken down into amino acids. However, in people diagnosed with autism it is more difficult to properly digest some of these proteins, thus allowing them to enter the blood directly as peptides. This often happens due to a lack of enzymes that help with the optimal assimilation of food and can be a factor in causing exaggerated bowel permeability (aka, leaky gut syndrome). This results in poor digestion, which facilitates the entry of these harmful proteins directly into the bloodstream, where they can cross the blood-brain barrier.

Leaky gut syndrome has been attributed to many causes including viruses; Candida; gluten, which produces zonulin, a protein thought to cause leaky gut; amongst other things.[6]

Gluten-containing foods may destroy the digestive system if they are consumed excessively or are introduced too early into a child's diet. Wheat is hybridized through artificial processing, resulting in inadequately prepared grains. Casein-containing foods may also destroy the digestive system because they are pasteurized and/or homogenized. These processes may result in damage to the enzymes that break down gluten or casein, thus causing incomplete digestion of these proteins.

Recently, Dr. Andreas Kalcker introduced us to the theory that the inability to properly digest these proteins may stem directly from parasites in the intestinal tract. These parasites may contribute to a leaky gut and thereby play a role in the development of allergies towards gluten, casein, soy, etc.

Gluten is found primarily in wheat, rye, barley, oats, spelt, malt, most breads, cakes, muffins, cereals, flour tortillas, pizzas, pastries, and donuts, etc. Gluten is also found in food starches, semolina, couscous, malt, some vinegars, soy sauce, teriyaki sauce, flavorings, artificial colors, and hydrolyzed vegetable proteins. Casein is contained in milk from cows, sheep, goats and any or their derivatives such as yogurt, butter, ice cream, or cheese. No form of cow's milk may be consumed as it causes inflammation and mucous. Even if the product claims to be lactose free, cream free, or casein free—it is not allowed.

Improperly digested gluten and casein fragments can both enter the bloodstream and cross the blood-brain barrier. Because of their opioid properties, these peptides can react with opiate receptors in the brain to cause effects similar to those of an opiate drug such as heroin or morphine.[7] These opiates are called gluteomorphin (or gliadorphin) and casomorphin, and can react with some parts of the brain, for example, the temporal lobes, which are actively involved in the process of the integration of language and hearing. Interestingly, these are two of the areas most affected by autism.

Besides their effects on the brain, opioid-forbidden foods cause inflammation in the gut and brain. When a child has an allergy to a food—in this case gluten, casein, and/or soy, etc.—as it enters the body the immune system sees it as an invader and reacts by trying to protect the body from the substance.

The first time the body is exposed to a food allergen, the immune system makes antibody-specific immunoglobulin E (IgE) against the allergen. IgE antibodies circulate in the blood and adhere to types of immune cells called mast cells and basophils. Mast cells are found in all body tissues, especially in the nose, throat, lungs, skin, and gastrointestinal tract (GI). Basophils are found in the blood and in tissues that have been swollen due to an allergic reaction.

The next time the body is exposed to the same food allergens, the allergen binds to IgE antibodies that bind to mast cells and basophils. These connections between allergens and antibodies direct the cells to release large amounts of chemicals, one of which are histamines. After the liberation of a histamine by an activated mast cell, the permeability of the vessels near the site increases. Therefore, the blood fluids (including leukocytes, which are also involved in immune response) enter the area causing inflammation. Histamine release also causes the release of cytokines and inflammatory mediators by leukocytes. These chemicals, in turn increase the inflammatory response.

Welcome to The Diet!

Now that we've covered the science behind *The Diet*, it's time to get started. First, make sure you complete an ATEC and save the results so you have a baseline score. Next, prepare a shopping list based on the following allowed items.

Author's Note: Organic products are better but not required.

Permitted Foods List

Proteins:

- Beef
- Chicken
- Eggs
- Fish (small not large size)
- Pork
- Turkey
- No processed meats or cold cuts (hot dogs, bologna, etc.)
- No shellfish (preferably—full of toxins)

Fruits:

- Most fresh fruits are permitted (except citrus, mango, pineapple, kiwi and limit berries).
- Frozen fruit without added cream or sugar.
- NO canned fruit (nothing canned ever).
- Be careful of dried fruit as it may contain sugar.

Author's Notes:

Juice is not permitted on this protocol as it has proven time and time again to annul CD and cost families precious time and money. The amount of fruit necessary to make a glass of juice also causes it to be high in sugars, albeit natural, and they can still affect the immune system. Children on this protocol need to drink water and generally speaking most former juice drinkers transition from drinking flavored beverages to water without much issue. If your child must drink a flavored beverage it would be acceptable to blend and strain or juice one apple, for example, and add it to some water and if you choose you can sweeten it with stevia (SweetLeaf® or KAL® brands are ok).

Fruit should not be consumed after a meal as a dessert due to its rapid digestion. If it is eaten after other more slowly digested foods (i.e. meat, grains, etc.), it can ferment in the stomach, causing bloating, gas, or discomfort. Fruit is best eaten before a meal or separate from meals.

Vegetables:

- All vegetables!!!
- Including French fries, however, not frozen fries or fries from fast food chains; these are often coated in flour.

Author's Note: I don't condone the prolonged use of potatoes, as they convert to sugar in the body. However, it was the only "legal" food my son would eat when we began the GF/CF/SF diet.

Nuts:

- Almonds
- Cashews
- Coconut
- Hazelnuts
- Walnuts
- Plus, they all make great milks!

Grains:

- Amaranth
- Buckwheat
- Corn
- Millet
- Quinoa
- Rice
- Sorghum
- Tapioca
- Xanthan gum

Author's Note: I prefer a grain free diet for autism recovery, especially if your child suffers from "grain brain" (addiction to grains, undesireable behaviors and/or inability to concentrate after eating grains). If your child seems to be addicted to carbs, i.e. fruits and starches, you may want to consider the Ketogenic diet or the Rosedale diet.

Even though these grains are allowed, they can be difficult to break down, and can easily ferment in the intestinal tract due to excessive Candida overgrowth, bacteria, and parasites.

Beans:

- All beans—EXCEPT soy
- Garbanzo
- Lentils
- Navy
- Peanuts

Sweeteners:

- Stevia (This is the best of all the sweeteners, but make sure it does not contain erythritol—a sugar alcohol!)
- Agave syrup
- Honey
- Maple syrup (natural/real—not made from corn syrup)
- Xylitol
- NO piloncillo (unrefined sugar)
- NO sugar

Author's Note: Honey is allowed, however, it can create insulin spikes.

Prohibited Foods List

After working with thousands of families, and conducting independent research, it is my personal opinion that the products on the following list should be avoided if your intention is to heal autism:

- Acetic acid (E260)
- Artificial flavoring
- Artificial sweeteners
- Bouillon cubes
- Bread
- Cacao/Cocoa
- Candy
- Cane sugar
- Carrageenan
- Catsup
- Chocolate milk
- Coloring
- Corn flakes
- Corn syrup
- Cow's milk in any form (even lactose-free milk products)
- Flour tortillas
- Gelatin
- Malt
- Margarine

- Mayonnaise
- Microwave popcorn
- MSG (Monosodium Glutamate)
  (MSG goes by many names—too numerous to mention here. A detailed list is available at www.truthinlabeling.org/hiddensources.html)
- Natural flavoring
- Noodle soup
- Oatmeal (except for Bob's Red Mill GF oats)
- Pasta
- Children´s nutritional shakes
- Play-Doh™
  (Contains gluten—Gluten-Free Dough is available at www.discountschoolsupply.com)
- Preservatives
- Processed meats (hotdogs, ham, sausage, cold cuts)
- Sodas
- Soy/fruit beverages
- Soy milk
- Soy sauce
- Sports drinks
- Sugar
- Yeast

NO Cow's Milk:

- Not casein free
- Not lactose free
- Not organic
- Not raw
- Not evaporated
- No No No cow's milk!
- Sorry, no goat's milk either! It too has casein.
- All cow's milk, regardless of what is removed from it, causes mucous production in the body. This, in turn, provides an ideal environment for pathogens.

If you're thinking that once you go GF/CF/SF + that your child will starve, trust me they won't! The list of foods they eat may get shorter, but they will keep eating. As I mentioned earlier, my son Patrick ate homemade French fries for the first three weeks of *The Diet* until he began accepting different fruits, nuts, and chicken again, but we did it... we transitioned. What I have seen in my own son, and in other children, is that when they begin to get healthier in the weeks and months after starting *The Diet*, they become more open to accepting new "legal" foods.

Wow... quite amazing Kerri...

I have always believed that "you are what you eat" and have been very careful with what I give to my kids, so their diet has been quite healthy, except for the occasional ice cream, etc. My son even spits out candies when someone offers him one because he really doesn't like them. He is not into sweets at all, thank God.

But I must say this diet is just what he needed! Yay!

There are a few loopholes to the above list. If you make your own catsup, mayonnaise, etc. with items from the permitted foods list, then your child can obviously eat those foods. The list refers to those items that come from the grocery store that contain white sugar, preservatives, and other problematic ingredients.

There are many items on the forbidden foods list that don't contain gluten, casein, or soy. Here is the explanation behind a few of those items:

White refined sugar: Refined sugar has been found to reduce immune system function as well as contribute directly to obesity and Type II diabetes.

The following list is excerpted from *Suicide by Sugar.*[8]

- Sugar can suppress your immune system.
- Sugar upsets the mineral relationships in the body.
- Sugar can cause juvenile delinquency in children.
- Sugar eaten during pregnancy and lactation can influence muscle force production in offspring, which can affect an individual's ability to exercise.
- Sugar can cause hyperactivity, anxiety, inability to concentrate and crankiness in children.

MSG: An excitotoxin that can literally "excite neurons to death." We need all of our neurons!

Yeast: Feeds Candida and other fungi.

Carrageenan: Contributes to inflammation in the body and studies have linked it to colon cancer in rats.[9]

Artificial sweeteners containing sucralose: In making sucralose, the chlorine in the sweetener bonds to carbon, producing a chemical known as a chlorocarbon. "According to physician and biochemist, Dr. James Bowen, chlorocarbons are never nutritionally compatible with our metabolic processes and are wholly incompatible with normal human metabolic functioning."[10]

Processed Meats (hot dogs, ham, cold cuts, and sausages): The fats found in these meats may contain high amounts of toxins such as heavy metals, pesticides, and herbicides. In addition, processed meats contain sodium nitrite, which can harm the liver and pancreas. They may also contain corn syrup and flavoring.

Natural Flavoring: The exact definition of natural flavorings & flavors from Title 21, Section 101, part 22 of the Code of Federal Regulations is as follows:

> *The term natural flavor or natural flavoring means the essential oil, oleoresin, essence or extractive, protein hydrolysate, distillate, or any product of roasting, heating or enzymolysis, which contains the flavoring constituents derived from a spice, fruit or fruit juice, vegetable or vegetable juice, edible yeast, herb, bark, bud, root, leaf or similar plant material, meat, seafood, poultry, eggs, dairy products, or fermentation products thereof, whose significant function in food is flavoring rather than nutritional.*[11]

Basically if you start with a natural ingredient, you can process or manipulate it any way you choose. No matter how many chemicals or solvents are added it will be labeled as a "natural flavor."

Notice they can come from meat, seafood, dairy, and wheat, etc., can also contain MSG.

Coloring: Food dyes have been linked to allergic reactions, hyperactivity in children, and even cancer. Red No. 2, for example, was banned in 1976 after it was suspected to be carcinogenic. Red No. 40, in a number of tests, was shown to damage DNA in mice.

When we are dealing with children on the spectrum, their immune systems are already compromised, and their detox pathways can be blocked or impaired. The idea is to decrease their burden with life-giving, nutritious foods, rather than add more stress to a body that is already maxed out.

Corn Syrup/High Fructose Corn Syrup: Simply because corn is a permitted food does not mean corn syrup/HFCS are "legal." Both of these shut down the immune system, as refined sugar does. They are chemically similar in composition. There is much controversy surrounding the use of mercury in the refining process, and how much if any, is in the final product.[12]

> I wanted to update you on everything with my daughter.
>
> We had a mishap 2 days ago. A substitute teacher gave my daughter a regular granola bar and a flavored beverage containing sucralose. You know what Kerri, when she came home from school that day she was in her own world and was ignoring me. I knew right away she had something she shouldn't have had. I was angry and saddened by what I was witnessing. Is there anything we should do since she had those 2 items that she's not supposed to have. Or just continue to move forward??
>
> After a while on the diet it becomes obvious when an infraction happens, if you are present when it happens you can give an enzyme to help break down the food, if not, we move forward and learn how to prevent it from happening in the future.

Common Errors:

"It's no big deal if we break the diet every once in a while."

Not True! Every time you break the diet, further inflammation is caused in the brain and gut. When gluten and casein proteins are not properly broken down, the resulting peptides reach the brain as gluteomorphin and casomorphin. It takes three days for casein to be eliminated by the body, but months for gluten to be eliminated by the body.

"My child can drink milk as long as it doesn't have casein."

Not True! As long as the milk comes from a cow, your child cannot drink it. It doesn't matter how it's labeled. If it comes from a cow it's off limits. Cow's milk can provoke the body to produce mucous, thus providing an ideal environment for pathogens which can cause chronic inflammation.

"The allergy panel says my child isn't allergic to gluten or casein, therefore he can eat them."

Not True! If your child has autism, or is on the spectrum, he must avoid gluten, casein, and soy. You should also observe your child carefully after adding a new food, or a food he hasn't eaten in a while. In one particular case, even though the child didn't test positive for an orange allergy he continued to produce symptoms of an allergic reaction whenever he ate one. We must remember that the body is changing constantly and that any test is only good for a couple of months, if at all.

I have seen the same story repeated over and over again with mangoes, oranges, bananas, apples, and corn, etc. Observing the conduct of your child after consuming a food you suspect he may have an allergy/intolerance to is the best way to measure whether the food is acceptable or not. As our children move towards healing, foods that at one point produced allergy-type symptoms can be tolerated without an immune system response.

"My doctor says that autism has no cure, and that the diet doesn't work."

Not True! Run from any doctor who says he will take your money even though he feels that autism is incurable. The first question you need to ask a doctor is:

"How many children have your recovered from autism?"

If the answer is zero, keep looking!

The Diet is free—no one earns money if your child is on *The Diet*—so do it! You have nothing to lose, and everything to gain. However, if you choose not to do it, you may further lose the health or your child. Commit 100% to *The Protocol*, and your child may be one of the next recovery stories.

> My son had what can only be described as a miraculous reaction to the GF/CF/SF diet. Since starting 2 weeks ago, all the bumpy red areas on his face disappeared, tantrums went from about a 10 to a 2, and diarrhea is all but gone.

"We tried the diet, but Johnny didn't get any better."

***The Diet* is only one piece of the puzzle.** However, it is the foundation of everything we are going to do. Without maintaining *The Diet*, it is difficult to know which intervention is actually helping.

Candida will be dying and a reaction to *The Diet* may be a Herxheimer. We must go beyond the diet to heal autism, but diet is the first piece of the recovery process.

When I receive an email complaining that a child isn't getting better I ask for a detailed list of exactly what the child is eating. I always find errors in parents' application of the GF/CF/SF + diet. What I typically uncover is that *The Diet* did not fail but that the application was flawed.

"I should remove gluten and casein from my child's diet gradually."

Not True!!! Remove them from the diet immediately and watch your child improve before your very eyes. Foods that cause IgG or IgE allergic reactions cause many other problems in the body including inflammation and psychotic behaviors, in response to gluteomorphin and casomorphin. The sooner you can remove these foods the faster your child will recover.

> As soon as we started working on the diet, my son received a developmental evaluation, although he was 20 months old, the results were that of a child of 15 months. 4 months after implementing the diet, the changes have been dramatic. My boy smiles again, he speaks in his own way, he points, has eye contact, interacts with others, he is very close to being recovered, although he is still missing a little bit. He is occasionally nervous and sometimes distant, but the most important thing is that the therapist reapplied the developmental test and my son is now reacting like a child of 24 months (his current age), even in some areas surpassing it! And this is only with the diet because he hasn't taken any supplements, just a probiotic that was ineffective.

"A gluten-free/casein-free/soy-free diet is good enough."

It's Not! Like I said before, I always find errors when a parent goes step-by-step with me through their child's diet. The list on page 52 is a detailed list of what our children should not eat under any circumstances. Many of these foods contain neurotoxins/excitotoxins. These foods can negatively affect developing and mature nerve tissue. Keep them out of your child's diet.

Beyond GF/CF/SF +

Sometimes, *The Diet* as explained is not enough and we need to go beyond it. If your child suffers from constipation, diarrhea, or seizures it is recommended to do a period on the Specific Carbohydrate Diet™ (SCD™). The following text was sourced from the website...

www.breakingtheviciouscycle.info

...by Elaine Gottschall. A complete list of references is also available at that website.

The Specific Carbohydrate Diet™ has helped many thousands of people with various forms of bowel disease and other ailments vastly improve their quality of life. In many cases people consider themselves cured. It is a diet intended mainly for Crohn's disease, ulcerative colitis, celiac disease, diverticulitis, cystic fibrosis and chronic diarrhea. However it is a very healthy, balanced and safe diet that has health benefits for everyone. The foods that are allowed on the Specific Carbohydrate Diet™ are based on the chemical structure of these foods. Carbohydrates are classified by their molecular structure.

The allowed carbohydrates are monosaccharides and have a single molecule structure that allow them to be easily absorbed by the intestine wall. Complex carbohydrates which are disaccharides (double molecules) and polysaccharides (chain molecules) are not allowed. Complex carbohydrates that are not easily digested feed harmful bacteria in our intestines causing them to overgrow producing by products and inflaming the intestine wall. The diet works by starving out these bacteria and restoring the balance of bacteria in our gut.

Autism & GI Problems

Altered intestinal permeability was found in 43% of autistic patients, but not found in any of the controls (Harvard University). Intestinal permeability, commonly called "leaky gut", means that there are larger than normal spaces present between the cells of the gut wall. When these large spaces exist in the small intestine, it allows undigested food and other toxins to enter the blood stream. When incompletely broken down foods enter the body, the immune system mounts an attack against the "foreigner" resulting in food allergies and sensitivities. The release of antibodies triggers inflammatory reactions when the foods are eaten again. The chronic inflammation lowers IgA levels. Sufficient levels of IgA are needed to protect the intestinal tract from clostridia and yeast. The decreasing IgA levels allow for even further microbe proliferation in the intestinal tract. Vitamin and mineral deficiencies are also found due to the leaky gut problem.

In a healthy intestinal tract the small intestine and stomach are not inhabited by bacteria. When the flora balance in the colon is lost, the microbes can migrate into the small intestine and stomach, which hampers digestion. The microbes compete for nutrients and their waste products overrun the intestinal tract. One of the toxins produced by yeast is actually an enzyme that allows the yeast to bore into the intestinal wall. The yeast also produce other toxins such as organic acids, which can also damage the intestinal wall.

Bacterial growth in the small intestine destroys enzymes on the intestinal cell surface, which prevents carbohydrate digestion and absorption.

The last stage of carbohydrate digestion takes place at the minute projections called microvilli. Complex carbohydrates that have been broken down by the enzymes embedded in the microvilli can be absorbed properly and enter the blood stream. But when the microvilli are damaged, the last stage of digestion cannot take place. At this point only monosaccharides can be absorbed because of their single molecule structure.

In the small intestine, the body should absorb the nutrients needed from what is eaten. But in the case of malabsorption, the undigested carbohydrates left in the small intestine cause the body to draw water into the intestinal tract. This pushes the undigested carbohydrates into the colon where the microbes can feast on it. This allows for even more proliferation of the unwanted microbes and continued increase in malabsorption problems.

Low intestinal carbohydrate digestive enzyme activity was found in 43% of patients with autism. (Horvath) Recent studies point out that ongoing carbohydrate malabsorption keeps the digestive system constantly weakened, leading to systemic disorders. Suspected carbohydrate malabsorption should be treated to ward off further damage to the body's digestive system. (GSDL)

Most intestinal microbes require carbohydrates for energy. The Specific Carbohydrate Diet™ limits the availability of carbohydrates. By depriving these microbes of their food source, they gradually decrease in number. As the number of microbes decreases so do the toxic by-products they create.

The Specific Carbohydrate Diet™ (SCD™) is intended to stop the vicious cycle of malabsorption and microbe overgrowth by removing the source of energy from the microbes. The SCD™ allows simple monosaccharides that do not need to be broken down in order to be absorbed.

By following the SCD™, malabsorption is replaced with proper absorption. Inflammation is decreased and the immune system can return to normal. Once the immune system is returned to adequate levels, it can begin to keep in the intestines microbes in proper balance.

The SCD™ allows simple carbohydrates, but prohibits complex carbohydrates. The diet is started by following an introductory diet, which consists of a limited selection of foods. After the introductory diet, the next stage of the diet allows many more foods, but requires that all fruits and vegetables be peeled, seeded and cooked in order to make them more easily digested. Raw fruits, vegetables, nuts and seeds are added to the diet later. To properly follow this diet, it is imperative to read Breaking the

> *Vicious Cycle by Elaine Gottschall. The book details the progression of allowed foods as well as providing many delicious recipes.*[13, 14]

Salicylates/Phenols

Phenol is a naturally occurring chemical found in many of the foods we eat, such as fruits and vegetables, nuts, and in bioflavonoids and cartenoids (carotene, lutein, lycopene, xanthophylls, and zeaxanthin), etc. Phenols can be found in toothpaste, hair dye, and disinfectants, etc. Many foods have phenols, and they are impossible to avoid completely. Salicylates are a subgroup of phenols, related to aspirin. There are several kinds of salicylate, which plants make as a natural pesticide to protect themselves from insects, fungi, and harmful bacteria. Foods high in natural salicylates are tomatoes, apples, peanuts, oranges, cocoa (chocolate), red grapes, coffee, all berries, and peppers, to name a few.

You may also need to consider a low salicylate and/or phenol diet, as many children on the spectrum may have issues with these items. The following was sourced from *www.scdlifestyle.com*.

> *Dr. Feingold is probably the most widely known individual to study this chemical, as he developed what is now referred to as the Feingold Diet. He started out in the 1960s as a pediatrician and allergist studying children's negative reactions to aspirin. Through his work, he found that many other dietary chemicals were causing physical and even behavioral reactions in his patients. He developed the Feingold Diet to eliminate all food additives, colorings, and salicylates.*
>
> ***Why Do People React to Them?***
>
> *In a normal body that has the correct levels of sulphates and liver enzymes, phenols and salicylates are easily metabolized. The body utilizes what it needs from the chemicals and properly disposes of the rest through the bowels. In those whose levels are not normal or in the case of "leaky gut syndrome", intolerance to this chemical family can occur rather quickly.*
>
> *Many people with gut issues such as yeast/bacteria overgrowth or digestive diseases can develop salicylate intolerance as a result of "leaky gut" syndrome. Leaky gut is a result of various digestive problems and occurs when the small intestine becomes too damaged to properly filter the size and types of food particles or chemicals that enter the bloodstream. [For more on leaky gut syndrome see http://scdlifestyle.com/2010/03/the-scd-diet-and-leaky-gut-syndrome/] When these improper particles are allowed to repeatedly enter the bloodstream the*

body tries to get rid of them by triggering an immune system response. Because phenols/salicylates are so common in most foods, a person with a leaky gut will have much higher than normal levels of these chemicals in their blood and can very quickly develop intolerance´s to these specific particles.

Why Can Phenols Affect Children on the Autism Spectrum More Than Others?

Research by Dr. Rosemary Waring at the University of Birmingham found that children on the autism Spectrum have low levels of the enzyme phenol-sulfotransferase-P (the enzyme that breaks down the phenol and amine families, also known as PST) and the substrate it uses: sulfates. Sulfates are a key tool that the body uses in the process of detoxification and break down phenols such as salicylates. Without normal levels of sulfates in the body, the sulfotransferase enzyme cannot do the task it was created to do: metabolize salicylates. So there are two problems with PST deficiency: low sulfate levels and low enzyme levels. PST deficiency alone can cause problems in children (remember phenols are normal) but factor in any intestinal damage resulting in leaky gut along with it and your child's body can easily be overwhelmed. The end result is a salicylate intolerance and the subsequent physical and behavioral reactions that come with it.

Reactions Caused by Phenols

Salicylates stimulate the central nervous system in people that react to them. This can often bring with it an emotionally extreme high followed by a very low, low. Other reactions to the phenol family can occur anywhere from immediately after consumption up to 48 hours after the consumption of the chemical, depending on the immune response. Physical reactions can include: dark circles under the eyes, red face/ears, diarrhea, headache, difficulty falling asleep at night, night waking, and in some cases excessively tired and lethargic. Behavioral symptoms of a reaction can be: hyperactivity, aggression, head banging or other self-injury, and even inappropriate laughter. Hyperactivity is more common in children's reactions, while adults generally experience symptoms similar to chronic fatigue.[15]

For additional information on leaky gut syndrome you may wish to look at (search for) *Dr. Peter Osborne*:

www.glutenfreesociety.org

If your child suffers from self-injurious behaviors limit these foods that are moderate to high in phenols:

- Almonds
- Apples
- Avocados
- Bananas
- Cacao
- Cantaloupes
- Cherries
- Cider vinegar
- Coconut oil
- Colored fruits
- Dates
- Food coloring
- Honey
- Mangos
- Mint
- Oranges
- Oregano
- Peanuts
- Pineapples
- Powdered chili pepper
- Processed meats
- Raisins
- Strawberries
- Tangerines
- Tomatoes

For more information on this you can visit:

www.scdlifestyle.com

and

www.feingold.org

* * *

She's quite the book worm and has an impressive collection which includes classic favorites such as Brown Bear, Llama Llama Red Pajama, Barnyard Banter and Goodnight Moon... But lately, she has been drawn to her Mama's new favorite read... Perhaps she is hoping that someday her story of recovery will be told too...

Gluten Syndrome

For many parents of children on the spectrum, having their children on *The Diet* is something that is second nature to them. However, as more and more people are using the protocol in this book to heal their own health ailments we have decided to include the following section on how gluten can negatively affect many more than just the autism community. You will get a very interesting look into how the immune system works and why so many sufferers have misunderstood gluten syndrome. Thank you Olive for contributing to our understanding of the subject.

Molecular Mimicry
What It Is & How It Relates to the Gluten Syndrome

by Mrs. Olive Kaiser

Who am I?

I am a married stay at home mom, blessed with a wonderful husband and seven fantastic kids. In 2003, after decades of searching, we learned about gluten reactivity through our daughter's nursing school training and eventually confirmed that we are a gluten syndrome family. Our daughter and my husband had the most obvious symptoms, but we all had manifestations and antibodies. Additionally our oldest son reacted to his MMR vaccination and probably other shots, which added high functioning ASD/ADD to the mix, and he developed type 1 diabetes at age 19. Two other sons had various shades of ADD/ADHD. My own school age vaccinations in the 1950's may have led to repeated bouts of strep throat until I reacted to a strep antibiotic injection about age 10. I developed PANDAS from that reaction (Pediatric Autoimmune Neuropsychiatric Disorder Associated with Streptococcus). What a struggle! Decades later it has responded somewhat to diet changes and now the CD/parasite protocol. I give thanks to God for His guidance along the way.

How did I get into this community and this health project?

We tested for gluten syndrome (we called it celiac disease back then) using standard tests recommended by celiac experts and received confusing results. Then, our daughter had a troubling experience with a gluten challenge that did not match the celiac story we'd been taught. I delved into medical literature and networked extensively with the gluten syndrome community looking for help. In that desperate discovery and prayer process I found practitioners and researchers who stepped outside the "villi damaged celiac only" box. They were able to explain why we received, in the midst of obviously gluten-induced incidents, false negative results from our celiac blood tests and villi biopsy. Those tests significantly led us astray and eventually I put up a website...

www.TheGlutenSyndrome.net

...to warn others of discrepancies we stumbled upon in the diagnostic process.

What is Molecular Mimicry?

Molecular mimicry is a recognized medical theory which explains very nicely why gluten reactions may potentially inflame and damage so many different parts of the body, leading to very different symptoms in different people. It also clarifies why gluten antibodies may cross react with other foods and infections, and why it only takes a small exposure to trigger them.

When we understand molecular mimicry we are better equipped to deal with tempting social situations. Gluten syndrome has its own rules, which do NOT make sense unless this concept is understood.

The following is a brief introduction to Appendix 5, page 449, which goes into much more detail, along with references, about the following questions:

1. How does the gluten syndrome reaction actually damage our bodies?

> Molecular mimicry. The molecular structure of gluten resembles the molecular structure of many of our body tissues. When the immune system attacks gluten it may also attack body tissues that "look like" gluten. Even if you do not read the other detailed answers, learn the details for this question on page 450.

2. Does gluten always damage the villi of the small intestine as the celiac story teaches? Many other tissues such as thyroid, pancreas, liver, joint, brain, nerves, heart, bone, blood vessel walls, etc., are involved in this disorder. Does all that other damage only arise from poor nutrient absorption from injured gut villi?

> No, according to published research, many researchers and practitioners believe the villi are not always damaged in an autoimmune gluten reaction. Where there is no villi damage, injury to other organs CANNOT be due to nutrient deficiencies caused by villi damage. Molecular mimicry provides a mechanism for direct autoimmune gluten damage to many other tissues and organs when the villi are fine, OR other tissues/organs, *including villi*, may be directly damaged through molecular mimicry.

Dr. Vojdani's abstract in his editorial *The Immunology of Gluten Sensitivity Beyond the Intestinal Tract*, supports that the villi are not always injured. To quote his editorial abstract, "Evidence has been accumulated in literature demonstrating that gluten sensitivity or celiac disease can exist even in the absence of enteropathy [gut/villi damage], but affecting many organs."

3. I don't have damaged villi, and my tTG/gliadin tests were negative, but I feel so much better gluten-free. Why?

The tests were likely false negative. That is very common. As gluten digests, it breaks into more pieces than we have tests developed to check them, and the immune system makes a separate antibody for each piece. Standard tests only check 2-3 antibodies. You may have others (Cyrex Labs tests 28 antibodies). Your villi may be fine, but you may be injured somewhere else—for example: thyroid, nerves, heart, etc.

4. Why do many gluten syndrome patients not only react to wheat, barley, and rye but also at times to other foods, particularly oats, milk, corn, soy, egg, yeast, coffee, sesame, rice, chocolate and others?

These foods "look like" gluten closely enough in their structure that the immune system may mistake them for gluten. This situation may also cause your gluten antibodies to run high after you go gluten-free. The immune system may misrecognize other foods, such as yeast, corn or milk, and others for gluten because they resemble gluten molecularly.

5. The diet seems excessively strict? Why does it take so little gluten to start a reaction?

Our perspectives are skewed. We accept that miniscule amounts of venom injected by a bee sting, or a tiny exposure to peanuts in allergic individuals can set off immediate life threatening allergic reactions. Many medications are contained in very TINY pills, but they have powerful effects in our bodies. Immune gluten reactions are also that sensitive. "Crumbs matter."

6. Why do many people react to gluten, proven by antibody tests, but they have few or no warning symptoms for a long time and then they crash with something serious, usually autoimmune?

Gluten is famous for slowly injuring nerves by molecular mimicry, and in many cases, the nerves are silenced by that injury. The patient does not realize there is a problem until the tissue or organ that those nerves supply begins to fail.

7. Why do so many of us react to gluten today, when for centuries most people appeared to be fine with wheat, barley, rye and oats? After all, wheat and barley are mentioned positively in the Bible and other historical documents.

Today's gluten is altered, violently, by nuclear radiation and chemical mutation within the past 60 years*, plus our toxic and poorly nourished bodies do not have optimal digestive capabilities to break it down. Weak, toxic, leaky body barriers/membranes, particularly leaky gut, set the stage for gluten induced molecular mimicry.

*Nina Federoff, *Mendel in the Kitchen*

8. Why do specialists and researchers insist that the gluten-free diet must be life long? Can't we heal this problem and go back to our beloved wheat bagels, croissants, and brownies?

Our scientists still insist that gluten-free is a strict lifelong commitment. I agree. For me it is not worth playing with today's wheat. There is something strange and unpredictable about it. The memory B cells in the immune system never forget what the enemy "looks like", and fresh exposure retriggers antibodies.

9. Traditional peoples soaked and/or sprouted their wheat berries and then made sourdough bread with them. Does that process alter the gluten sufficiently for gluten syndrome patients to safely consume this bread, particularly spelt or einkorn?

No. These processes and ancient wheat grains do make the bread more digestible, but not gluten-free and still unsafe.

10. Should I substitute all the gluten foods I routinely eat with gluten-free substitutes?

No, not routinely. The gluten-free community finds that they are still mainly expensive high carb processed food (i.e., junk food).

11. What are gluten withdrawals?

Occasionally, gluten breaks into specific "pieces" in the gut that resemble opiate drugs. When a person goes gluten-free, they may experience temporary, but unpleasant, withdrawal symptoms for a few days as these pieces disappear from the blood stream.

12. What are the risks of formal gluten challenges?

Many patients avoid these challenges. Occasionally a patient tries the gluten-free diet for an extended period of time and then the patient or doctor decides to run tests to confirm gluten reactivity. The standard advice to restart the production of antibodies is to consume gluten products 4 x per day for 4-6 weeks, and then run the standard blood test, followed by a villi biopsy if the blood work is positive. This is called a gluten challenge and has created some very dramatically unhappy reactions, some of them neurological/psychological.

Please see Appendix 5 (page 449) and www.GlutenSyndrome.net for more info and references. When we understand molecular mimicry our understanding of the gluten syndrome comes into focus. It explains why gluten-free diets and beyond are important tools to reduce inflammation and promote healing. As time goes on, gluten-free diets are easier to manage in public, tests are better and social awareness has grown. The Just Eat Real Food movement and others play into healthy gluten-free dining with wonderful recipes that avoid processed foods and incorporate healthy fats and nutrient density. This is a happy, encouraging era as we watch our children heal and adults find better stability in the midst of a health crisis. Bon Appetit!!!

* * *

The Diet - FAQs

The following FAQ's were printed with permission from gfcfdiet.com:

Slight changes were made to match the formatting of this book; however, the content remains unaltered. Please visit gfcfdiet.com for complete list of the references contained within the following text.

1. My doctor has never heard of any of this and she is extremely skeptical. I'm embarrassed to tell her I'm considering this approach. What do you think?

Skepticism is a good thing in a medical doctor or scientist. However, since there is preliminary evidence to support this safe, non-invasive intervention, it is up to you to educate her, state your wishes, and ask for her support. For a doctor, it is better to wait until all of the data is published in peer-reviewed journals before advocating a treatment. However, for a parent, it is reasonable to want to help one's child without waiting for all of the results of the "double-blind placebo" studies. Because this approach does not include any unusual supplements, invasive drugs, or expensive treatments, your pediatrician should be supportive. Explain that you would like to try this for a few weeks, and agree that you will be objective about recording your child's progress while on the diet.

If you feel that you need to support your case legally with the scientific and medical documentation that is currently available, please see the medical links at www.gfcfdiet.com or at www.autismndi.com.

2. What is casomorphin?

Casomorphin (or caseomorphin) is a peptide derived from casein, a milk protein. Casein is one of the major proteins in the milk of all mammals including cows, goats, and humans. When casein is digested properly, it breaks down into large peptides like casomorphin, and should then be broken down further into smaller amino acids.

However, Dr. Reichelt in Norway, Dr. Cade at the University of Florida, and others found that urine samples from people with autism, PDD, celiac disease, and schizophrenia contained high amounts of the casomorphin peptide in the urine.[4] In its peptide form, casein has opiate properties similar to morphine, and may plug into the same opiate receptor sites in the brain. Researchers have found that these peptides may also be elevated in other disorders such as chronic fatigue, fibromyalgia, and depression based on anecdotal reports of symptom remission after exclusion of wheat and dairy.

3. What is gliadorphin?

Gliadorphin (also called alpha-gliadin or gluteomorphin) is a substance that resembles morphine. Ordinarily, this is a short-lived by-product from the digestion of gluten molecules (found in wheat, barley, rye, oats, and several other grains). Gliadorphin is very similar to casomorphin. Gliadorphin has been verified by mass spectrometry techniques to be present in unusual quantities in urine samples of children with autism, and are believed by many to be a central part of the system of causes and effects that cause autistic development.

The most probable reasons for the presence of these molecules are:

- One or more errors in the breakdown (digestion) process caused by enzyme deficiency and/or
- Abnormal permeability of the gut wall (that would allow these relatively large molecules to enter the bloodstream from the intestine in abnormal quantities).

4. I am confused about allergy vs. intolerance. I understand that our children may be sensitive to corn, soy, and other foods as well as gluten and casein. Does this mean that they will eventually start turning these foods into the morphine-like compounds too? If this were the case, would they show up as an allergy on a RAST test? Or were our children always allergic to these foods (a regular allergy

that may cause behavioral changes in our children), and we just didn't know because the gluten and casein were hiding the allergy?

To a traditional physician or allergist, "allergy" is used to describe a reaction of the IgE part of the immune system, resulting in hives, swelling, or breathing problems. However, the words "allergy" and "intolerance" are often used to describe any inappropriate reaction to foods or substances that should normally be harmless to the body.

There are at least three different ways that a child with autism may have a problem with foods that contain gluten or casein, and it's important to understand the distinction:

1. An IgE ALLERGY commonly results in skin problems, hives, swelling, and breathing problems, etc. This can be tested using a skin test or blood test.

2. An INTOLERANCE (usually mediated by the IgG or IgA part of the immune system, or by an enzyme insufficiency such as lactose intolerance) has more varied or vague symptoms like discomfort, stomach problems, sleep problems, joint pain, ear infections, or hyperactivity and behavior problems. Sensitivity to these substances can be tested with an ELISA blood test.

3. PEPTIDUREA (peptides in the urine) is caused by the inability of the body to properly break down certain proteins. It is hypothesized that certain peptides, notably from milk and wheat proteins, are plugging into the opiate receptor sites of the brain and disrupting brain and nervous system function. Urine testing for this is still experimental, and many parents believe that the best way to find out if this is what is causing a child's autism is a strict trial period on the GF/CF diet.

On the GF/CF diet, gluten and casein are avoided because they are strongly suspected of having a direct pharmacological effect. When these proteins are only partly broken down, some of the resulting fragments can be strikingly similar to morphine, and act in more or less the same manner. (This type of reaction can co-exist along with a classic type of allergy towards the same foods.)

Recent research indicates that protein from both corn and soy may also contain some molecular sequences that could, if the patient has an enzyme deficiency, be broken down into something closely resembling opioid peptides. Even spinach protein has been found to have some opioid activity.

Products made from soy or corn will also often contain metabolic end products made by microscopic organisms like bacteria, molds, or other fungi. Some of these are suspected of being harmful to a small number of people who are genetically predisposed to autism. The amount of danger will depend on individual conditions AND on the quality of the corn or soybeans used in the production process. Soy OIL (lecithin) may be worse than most other soy products since this product will look and taste okay, even when made from moldy raw material, and since it is commonly made from the "bottom grade" of the harvest.

Some people also think that one of the natural pigments in corn (lutein) may cause problems for reasons that are not properly understood (see: "Sara's Diet"). This must be regarded as highly speculative.

5. I don't think my child has allergies, or that allergies could cause autism. Why should I try removing foods from his diet?

Although parents have been reporting a connection between autism and diet for decades, there is now a growing body of research that shows that certain foods seem to be affecting the developing brains of some children and causing autistic behaviors. This is not because of allergies, but because many of these children are unable to properly break down certain proteins.

6. Milk and wheat are the only two foods my child will eat. His diet is completely comprised of milk, cheese, cereal, pasta, and bread. If I take these away, I'm afraid he'll starve.

There may be a good reason your child "self-limits" to these foods. Opiates, like opium, are highly addictive. If this "opiate excess" explanation applies to your child, then he is actually addicted to those foods that contain the offending proteins. Although it seems as if your child will starve, if you take those foods away, many parents report that after an initial "withdrawal" reaction, their children become much more willing to eat other foods. After a few weeks, most children surprise their parents by further broadening their diets.

7. Isn't milk necessary for children's health?

Americans have been raised to believe that this is true, largely due to the efforts of the American Dairy Association, and many parents seem to believe that it is their duty to feed their children as much milk as possible. However, lots of perfectly healthy children do very well without it. It is calcium children need, not milk. Cow's milk has been called "the world's most overrated nutrient" and "fit only for baby cows." There is even evidence that the cow hormone present in dairy actually blocks the absorption of calcium in humans.

Be careful. Removing dairy means all milk, butter, cheese, cream cheese, and sour cream, etc. It also includes product ingredients such as "casein" and "whey," or even words containing the word "casein." Read labels—items like bread and tuna fish often contain milk products. Even soy cheese usually contains caseinate.

For more information on dairy-free living, there's a very good book called *Raising Your Child Without Milk* by Jane Zukin. Another great book called *Don't Drink Your Milk* by Frank Oski (the late head of pediatrics at Johns Hopkins and author of *Essential Pediatrics*). This book cites the results of several research studies that conclude that milk is an inappropriate food for human children. It is available for $4.95 from Park City Press, PO Box 25, Glenwood Landing, NY 11547, ISBN #0671228048.

8. How do I know which foods he's allergic to?

Try an allergy elimination diet. For example, keep common allergens out of his diet for a few days and then re-introduce them, one-by-one. If you see symptoms, either physical or behavioral, try again in a few days. By being systematic you will be more certain of which foods are causing problems and correctly rule them out. Two excellent resources, probably available at your library, are Doris Rapp's book, *Is This Your Child*, and William Crook's, *Solving the Puzzle of Your Hard to Raise Child*.

9. Aren't eggs dairy?

Many years ago, most of us were taught that eggs and dairy were part of the same section of the food pyramid under "dairy." However, they are not. Eggs are free of any dairy. Dairy can be from cows, sheep, or goats. The eggs we eat are normally from chickens. Dairy and eggs happen to be side-by-side in the refrigerator section of the grocery store, which may add to the confusion.

10. What do I do when we go to a party or out to another person's house and are trying to maintain my child on a GF/CF diet? At times we find it impossible to catch him before he ingests something he shouldn't.

It is a good idea to bring some favorite GF/CF goodies with you. Keep them hidden until that moment of "competition." Show the surprise treat to your child before your child eats one of the gluten and/or casein laden foods. Bring more than one goodie to be on the safe side. Also, there are two small pamphlets that are great for handing out to friends and family members that give a simple explanation for your child's special dietary requirements. These are: "Alternative Treatments for Children Within the Autistic Spectrum: Effective, natural solutions for learning disorders, attention deficits, and autistic behaviors," by Deborah Golden Alecson, and "Leaky Gut Syndrome: What to do about a health threat that can cause arthritis, allergies, and a host of other illnesses," by Elizabeth Lipski, M.S., C.C.N. Both of these books cost $3.95 each and can be found at health food stores.

11. Do we need to worry about gluten-containing lotions, shampoos, and toothpaste being absorbed through the skin?

Nicotine patches, birth-control patches, and other transdermal applications of medications are proof that the skin does absorb many things (and pass them to the bloodstream). However, the molecules in gluten are too large to pass through the skin, according to John Zone, MD (a dermatologist quoted in the Spring 2003 issue of *Living Without* magazine).

Most often the problem occurs from hand to mouth (i.e. touching colored modeling dough and then touching one's face), which is a good reason to avoid giving our kids much access to anything that poses a threat. If your child's school has a sensory table, you can request that it be filled with rice or dried beans, instead of macaroni or gluten grains. Likewise, we suggest that you volunteer to be the "Play-Doh™ parent" and keep the classroom supplied with a safe version that you don't have to worry about.

We are not too concerned about shampoo, unless you have a child who is likely to try to drink the stuff. But toothpaste is an entirely different matter, since some of it is likely to be ingested, rather than spit out.

12. I'm already worried about my child's nutrition, and his "allergies" are causing me to further reduce his choices. If apple juice and bananas are the only fruits he will eat, and he's reacting to them, how is he supposed to get by?

Fruit contains water, sugar, fiber, and vitamins. If he will not eat other fruits then he needs to get these things from other sources.

13. I am wondering if anyone else is having problems with school and keeping them on the diet? I have sent a note saying: "Please don't give my son dairy, wheat, corn, or soy." Today I came in and they were giving him Popcorn!

My kid doesn't eat ANYTHING that isn't sent from home. He's almost ten now, and this policy has been good for him because succumbing to temptation outside of the house has simply never been an option. I don't think it would occur to him to take food from anybody without our permission.

We always made sure there are backup snacks and birthday party treats at school for emergencies. I think the school has always taken this seriously because they can see how seriously we take it.

We once got a note from a mom who said that her son had been strictly GF/CF for months without an improvement. I asked what he was eating, and she gave me a long list including expensive GF prepackaged products. Then she ended her letter by saying, "Of course, that's just what he gets at home. I have no idea what he gets in his school lunch, or what they're using for food reinforcers in his ABA program."

Definitely try this—you might see a real difference.

14. What percentage of children will respond to dietary intervention?

DAN! doctors used to try to be conservative about this, and say at least a third of children, and then, after seeing more patients, they said two thirds. Now that they have seen hundreds or thousands of patients, most tell us that they believe that almost every ASD child will benefit from this diet. Many will need further modifications (i.e. removing grains or sugar) before the full benefits are realized.

However, age plays a big part in how quickly results will be seen. We can probably say that the response will be dramatic in more than two thirds of the children under three, and perhaps more subtle, but still helpful, in at least two thirds of older children. We think those are pretty good odds.[3]

Some Additional FAQs...

Kerri, yesterday the therapist was playing with Play-Doh™ with my daughter. She could not resist eating it. We tried very hard to stop her with no luck. In the end we had to stop her from playing with it. People say this is due to a zinc deficiency. Do you know if this is correct? Will giving her Zinc stop her from wanting to eat Play-Doh™? If yes, how much should I give?

She is not eating Play-Doh™ because of a zinc deficiency. Play-Doh™ contains wheat. She is addicted to gluteomorphin, a by-product of her eating gluten, which is contained in Play-Doh™. So, she broke her diet. She cannot touch nor eat Play-Doh™.

What exactly is Candida?

Candida albicans is an opportunistic fungus (or form of yeast) that is the cause of many undesirable symptoms ranging from fatigue and weight gain, to joint pain and gas.

The Candida yeast is a part of the gut flora, a group of microorganisms that live in your mouth and intestine. When the Candida population starts getting out of control it weakens the intestinal wall, penetrating through into the bloodstream and releasing its toxic by-products throughout the body.

What Causes Candida Overgrowth?

Candida is an opportunistic pathogen that can rapidly take over when a person is under a course of antibiotics. Antibiotics destroy beneficial gut flora but have little effect on Candida, giving this normally harmless yeast the chance to take over dominance of the gut environment very quickly.

Babies born via C-section, or to mothers who were treated with IV antibiotics during labor are especially vulnerable to the ravages of Candida overgrowth. This is because they are not exposed to a healthy balance of gut flora on the way through the birth canal prior to the moment of birth.

…disaccharides, or double sugars, are present in many carbohydrates including ALL grains—not just gluten containing ones. An inflamed, imbalanced gut overridden with Candida is unable to digest double sugar molecules completely because the lack of beneficial gut flora has compromised the function of the enterocytes.

According to Dr. Natasha Campbell-McBride MD, author of *Gut and Psychology Syndrome*, and one of the key scientists at the forefront of gut restoration research today, enterocytes are the cells that reside on the villi of the gut wall and produce the enzyme, disaccharidase, which breaks down the disaccharide molecule into easily absorbed monosaccharide molecules. When the enterocytes are not nourished and strengthened properly by adequate beneficial flora, they become weak and diseased and may even turn cancerous. Unnourished, enterocytes cannot perform their duties of digesting and absorbing food properly.[16, 17]

The critical importance of the enterocytes to health cannot be overstated!

Weak and diseased enterocytes also have trouble digesting starch molecules, which are very large molecules that consist of hundreds of monosugars, connected in long branchlike strands. People with weak digestion due to Candida overgrowth and messed up enterocytes have a terrible time digesting these complex molecules leaving large amounts of it undigested—the perfect food for pathogenic yeasts, bacteria, and fungi like Candida to thrive upon. Even the starch that manages to get digested results in molecules of maltose, which is—you guessed it—a disaccharide! This maltose also goes undigested due to a lack of the enzyme disaccharidase and becomes additional food for Candida.

We are a few days into the diet. Can someone comment on what they send to daycare with their kids for snacks? Our daycare is nut free as well. We are ok for lunches but what should we take for snacks. The gluten-free pretzels, etc. we were taking he can't have now because of yeast and soy.

Here a few suggestions: Hard boiled eggs, cut up raw veggies and fruit, air popped popcorn, apples, bananas, dried fruits (with no sugar added). Also you can join "CD/CDS MOMS WHAT IS MY KID EATING TODAY" Facebook group. You can find more ideas there.

Chapter 4

An Introduction to Chlorine Dioxide

by Jim Humble

> *"Condemnation without investigation is the height of ignorance"*
>
> ***-Albert Einstein***

Jim Humble, the man behind chlorine dioxide for healing, was kind enough to write the following introduction to our CD section...

I just can't tell you how pleased and proud I am that Kerri has developed protocols using chlorine dioxide as the main ingredient to bring health and normalcy to many children who had the symptoms called autism. Their resulting healing is one of the greatest stories of this and the last century. Kerri's ability to look at something new and to evaluate the value of the new item without being totally affected by the dictates of present established medical science is truly rare. Kerri decided to do something since modern medicine has yet to come up with a reasonable priced globally available solution to the problem of autism that yielded good results. Medical science says that autism is incurable, that millions of children have to be left to suffer and live without reaching their potential. At this writing 115 children have recovered, hundreds more are near recovery and thousands have had documented improvements as a result of being treated with Kerri's protocol.

Here are some basic points:

Chlorine dioxide (the molecule) was discovered in 1814, with many giving credit to Sir Humphrey Davy for its creation, and today several patents exist for healing various ills with it (see list on page 85).

MMS (Master Mineral Supplement) – I rediscovered chlorine dioxide in 1996 and since then more than 1,020 people have traveled from more than 95 countries to the Dominican Republic, Mexico, and other countries, to be trained in its use. By our estimation, more than 10 million people worldwide have used chlorine dioxide.

The highest cost of a set of bottles of sodium chlorite + activator is less than $25 for 4 ounces of each. This means that almost any place in the world the average dose of chlorine dioxide is less than 4 cents, or less than $.40 a day or $25 for 2 months. In most places it is even less than that. So far there is no one in the world who can't afford it or who wouldn't be able to get it merely by asking if they can't afford it.

The chlorine dioxide molecule is the weakest of all oxidizers used in the human body and thus has very little effect on human cells. This molecule has the unique ability to recognize and oxidize (kill) harmful bacteria. Hans Christian Gram, a scientist of the 19th century, discovered that most harmful bacteria have a negative charge. He was able to dye positive and negative bacteria two different colors. His techniques are still used in laboratories and universities. All oxidizers, including chlorine dioxide, have a positive charge that will attract and kill negatively charged bacteria while repelling positively charged bacteria. Chlorine dioxide is thusly able to kill bad bacteria without destroying the good bacteria. This is simple high school science; like charges repel and unlike charges attract.[1]

Most chemicals and many foods in this society are tested to determine their poison index, which is designated by giving them a toxicity number known as LD_{50}. Sugar, table salt, butter, arsenic, cyanide, vinegar, wine, Clorox, window cleaner, and dozens of other household and industrial chemicals are rated as to their toxicity by force feeding 10 rats with the food or chemical in question until ½ of them die. The amount of chemical or food that it takes to kill 5 of the 10 rats determines the LD_{50} number. Every chemical has a toxic rating and is indicated as an LD_{50} number. All chemicals are toxic in large enough quantities and there are a lot of unlucky rats in our society.

There has never been a recorded death caused by ingestion of chlorine dioxide or sodium chlorite in humans. In the case of chlorine dioxide, it would take 1000 times more chlorine dioxide than is used in a daily treatment dose to reach a toxic dose for a human being. The following information is from the *Toxicological Profile for Chlorine Dioxide and Chlorite* (published by the U.S. Dept., of Health & Human Services) which contains LD_{50} toxic numbers for chlorine dioxide along with documentation that no deaths have been caused by chlorine dioxide (Emphasis in italics is the author's):

> *No information was located regarding death in humans following oral exposure to chlorine dioxide or chlorite. Shi and Xie (1999) indicated that an acute oral LD_{50} value (a dose expected to result in death of 50% of the dosed animals) for stable chlorine dioxide was > 10,000 mg/kg in mice.*[2]

I should say something about bleach as a number of critics insist that MMS (chlorine dioxide) is a bleach and thus is a poison. Chlorine dioxide, the chemical that is MMS, has never been used domestically as a bleach for cleaning toilets. It is an industrial bleach when used 4000 times more concentrated than MMS. It is never used for bleaching purposes in the homes of people of this world.

Legally, ethically, morally, and logically, if any item does not have the characteristics of a particular item, it cannot be said to be that kind of an item. This means that a chemical that does not have the characteristics of bleach cannot be called bleach. So a chlorine dioxide solution being 4000 times weaker than that of bleach cannot be called bleach. This is a fact accepted in courts of law of the world.

I sincerely hope you will take the time to understand the basics of chlorine dioxide as explained here. Any time one has helped someone overcome the symptoms of autism he has helped save a life, as those suffering from the symptoms of regressive autism are not living their full lives. The end result of learning this data is that the time may come when you will need chlorine dioxide to save your own life or that of a loved one.

Jim Humble

Archbishop Jim Humble

A message from Kerri to Jim Humble

"With all my love, respect and gratitude. You have swum against the current for so long, and I am so fortunate that you waited for me. I will continue to recover these generations of children that are the victims of a senseless epidemic. As you have done, I will do. We can't change the past, but the future is ours. You have brought light to where there once was none."

Chapter 5

Step 2 – Chlorine Dioxide (CD)

"All truth passes through three stages. First, it is ridiculed. Second, it is violently opposed. Third, it is accepted as being self-evident."
~ ***Arthur Schopenhauer***

CD (chlorine dioxide) is the most amazing intervention that I have ever tried with my son. It by far packed the biggest punch in the shortest amount of time—we saw the positives in under a week with no side effects! (A Herxheimer reaction is not a side effect.) After seven days on CD, my son was looking at me in the eyes and requesting non-preferred activities using four word sentences. I have never, with any other intervention, seen a change in my son come about so quickly, especially since Patrick was your typical non-responder before CD.

The chlorine dioxide molecule (CD).

Why CD for Autism?

We know that the symptoms known as autism are caused by:

- Viruses
- Bacteria
- Candida (yeast)
- Parasites
- Heavy Metals
- Inflammation
- Food allergies

CD catapults children into recovery because it kills these exact things. As of this writing, over three years since we started using CD in our protocol to heal autism, 115 children have lost their diagnosis. Thousands of children around the world are healing, and that means their families are healing as well, and for this I am eternally grateful.

CD has confirmed my faith and has shown me time and time again that autism IS curable. My life has changed since the day my son was able to look me in the eyes, smile, and request things. I could not keep information of this magnitude a secret. I absolutely had to share it with anyone who would listen.

CD is very affordable as you can purchase a two to three month supply from multiple worldwide suppliers on the Internet for around $25USD. The cost difference between the CD Autism Protocol and a megavitamin/supplement based biomedical protocol for a family living without insurance (or in a third world nation) could be the difference between recovering their child and not recovering their child. The results are also vastly different.

You can think of CD as a shortcut to recovery because it kills all pathogens with one shot. Hypothetically, you could use antibiotics, antifungals, and antivirals to get the same results; however, you could also cause damage due to harmful side effects from these products (liver and kidney stress, etc.). Also, antibiotics, antifungals, and antivirals can only target certain bacteria, viruses, and fungi. CD targets all types of pathogens based on their electrical charge and pH. The particular strain is irrelevant.

CD causes no side effects, leaves no toxic residues, and causes no liver damage. It causes no new stress to already stressed bodies. Everything else we use to kill pathogens leaves something behind. For example, antibiotics also kill your beneficial bacteria, and antiviral use comes with contraindications and side effects. CD is gone from the body one hour after your last dose, leaving no trace behind. We dose hourly throughout the day because CD is only active in the body for up to one hour. We must continue to dose if we want to stop the proliferation of pathogens and eliminate them from the body.

A few children with autism in Venezuela recovered using only CD, The Diet, and a probiotic. That doesn't happen every time, but it can.

False Information and the Fear it Creates

The most common reason that families have told me they are afraid of trying CD is because of false propaganda. The FDA, anti-biomed parents, and others have claimed that MMS (Jim Humbles term for a weak concentration of chlorine dioxide in aqueous solution) is a "poison," comparing it to chlorine or chlorine bleach, which it is not.

If you are interested in a deep dive into the science behind chlorine dioxide therapy (also known as MMS), please spend some time on Dr. Andreas Kalcker's websites...

www.medicasalud.com

...mostly in Spanish, though it has a few pertinent documents in English, and...

www.andreaskalcker.com

Also check Jim Humble's website:

www.jimhumble.org

There, you can also find a forum (G2Cforum.org) of people discussing the use of CD (MMS) for the healing of a multitude of ailments.

In a nutshell, chlorine (Cl) and sodium hypochlorite (NaOCl, aka chlorine bleach) are as different from chlorine dioxide (ClO_2) as "required-for-life" oxygen (O_2) and "will-damage-the-lungs" ozone (O_3) are different from each other. Chlorine (Cl) and sodium hypochlorite (NaOCl) destroy pathogens through "chlorination," while chlorine dioxide (ClO_2) destroys pathogens by "oxidation." The by-products of chlorination can bond with other molecules and form potentially carcinogenic trihalomethanes. The only byproducts of oxidation by chlorine dioxide are two neutral oxygen atoms and a chlorite ion, which can bond with sodium in the body to form table salt (NaCl).

CLOROX
SPLASH-LESS

Laundry Bleach

OR

Chlorine Dioxide

IS NOT

Pool Chlorine

Your doctor is most likely not a chemist, and may or may not be fluent in the language of oxidizers. If you are working with a doctor, please take the science to them when you speak with them about adding CD to your child's protocol. The best thing we can do for our sick children is to make informed decisions. Understanding the difference between chlorine and chlorine dioxide, or chlorine bleach and chlorine dioxide may mean the difference between your child recovering or not.

I have shared this with thousands of families around the globe whose children have used chlorine dioxide in very dilute aqueous solutions, and NONE have had any injuries from CD. Sure, we have had discomfort in the form of a Herxheimer reaction. This is something we have learned to minimize, and can often be prevented by going low and slow. I have learned that what may appear negative at the outset, often turns out to be a healing crisis that leads to incredible gains in the long run.

Many of the first world children who have done the CD protocol have done it in parallel with other autism protocols. They have had regular blood, urine, and stool tests and generally test healthier while using CD than before they started. CD is fortifying immune systems, improving liver function, reducing bacterial and viral loads, *Candida* markers, and inflammation.

> Just did round of blood tests on my son,
> Liver Function Test best ever after 7 months of MMS.

Critics also frequently cite the fact that sodium chlorite and chlorine dioxide can be used for industrial bleaching processes, as a way to frighten parents out of using it. Just because a substance has the power to remove color, or bleach something, doesn't mean it's "bleach." Lemon, sunlight, hydrogen peroxide, and toothpaste all remove color from textiles. We would never call toothpaste bleach, even though it's capable of removing color, nor would we call sunlight bleach, even though that's what it can do to your clothes if you leave them outside long enough. From here we can draw a very similar analogy to using chlorine dioxide for health. The dilutions we are using for internal consumption in aqueous solution (water) are very weak. Day one starts with just one drop in eight fluid ounces of water.

The 1984 study, *Controlled Clinical Evaluations of Chlorine Dioxide, Chlorite and Chlorate in Man*, found no deleterious effects when taking a concentration of up to 50ppm over several hours a day for three weeks.[1] As Jim Humble mentioned (see page 79), the concentrations used for industrial purposes are are thousands of times more concentrated than what families have been using for their children on the spectrum.

Any substance can become toxic if taken in too high a concentration—even water.[2,3] Chlorine dioxide is no exception. We always increase CD dosages, *low and slow*, and have a respect for the power that chlorine dioxide can have over pathogens, always resisting the urge to ramp up our dosing too quickly.

> We started MMS 68 Days ago and my son's ATEC has dropped from 50-23… Now that's a MIRACLE!!! He is So much more "with us" and I'm seeing improvements daily. He is just so much more fun to be with now:) He did have a plateau, but is back on track. We have only done 4 enemas to date, but they made a big difference! Baths do as well:) I can't wait to see what he is like on his 4th birthday (still 39 days away).

Chlorine Dioxide was discovered in 1814 by Sir Humphrey Davy as an effective disinfectant. Since then, numerous companies have patented its use for countless applications, including food and health.

The following is an excerpt from Jim Humble's Nov 21, 2012 newsletter:

> Dr. Andreas Kalcker spoke of patents obtained by different multinational pharmaceutical companies, to cash in on this product or to prevent it from being marketed.
>
> Some of those patents include:
>
> - Nontoxic Antiseptic (Pat 4035483/1977)
> - For combatting human amoebas (Pat.4296102/1981)
> - Against dementia caused by AIDS (Pat.5877222/1999)
> - For curing all types of illness of the skin (Pat 4737307/1988)
> - For disinfecting live blood (Pat. 5019402/1991)
> - For curing injuries more rapidly (Pat. 5855922/1999)
> - For all types of oral care (Procter & Gamble) (Pat. 6251372B1/2001)
> - Against infections caused by bacteria (Pat. 5252343/1993)
> - For treatment of severe burns (Pat.4317814/1982)
> - For the regeneration of bone marrow (Pat. 4851222/1989)
> - Treatment of Alzheimers, dementia etc. (Pat. 8029826B2/2011)
> - To stimulate the immune system in animals (Pat. 6099855/2000)
> - To stimulate the immunological system (Bioxy. Inc.) (Pat. 5830511/1998)

> ...The danger that may occur is if one mistakenly used a highly concentrated form. Highly concentrated forms of virtually every substance can cause death. In his conference, Dr. Andreas Kalcker points out that 70g of even plain table salt can cause death.
>
> There have been thousands of reports worldwide of excessive doses of MMS being ingested, however, with the hundreds of millions of tons of this substance used each year worldwide for water purification, preservation of vegetables, sterilization of slaughtered meat, and hundreds of other uses, there has never been a report of a death caused by ingestion of chlorine dioxide.
>
> This proves chlorine dioxide to be one of the safest products known.

There are many products that contain chlorine dioxide (ClO_2) that are currently approved by the Food and Drug Administration (FDA) in US. They are manufactured by *Frontier Farmaceutical*, *Alcide*, *Bioxy*, and others for skin and oral care. A number of these product names are listed below:[4]

- DioxiRinse™ Mouthwash
- DioxiBrite™ Toothpaste
- DioxiWhite™ Pro Teeth Whitener
- WhiteLasting™ Maintenance Gel
- BioClenz™ Dental Unit Waterline Cleaner
- Penetrator™ Periodontal Gel
- Simply Clear™ Acne Treatment
- DioxiWhite™ Home Teeth Whitener
- Cankers Away™ Canker Sore Cure
- DX7™ Skin Protectant Gel
- Periodontitis Treatment
- DioxiSmooth™ Facial Exfoliant
- Gingivitis Treatment Surface Disinfection
- Fire Fighter™ Burn Pain Reliever
- DioxiGuard™ Spray Disinfectant
- Nail-It™ Nail Protector

Chlorine Dioxide is an Oxidant…
But aren't Antioxidants Good for Us?

In 1954, Denham Harman published an article called the "Free Radical Theory of Aging,"[5] claiming antioxidants would slow the aging process. However, he later found that mitochondria determined life span, and that antioxidants do not enter the mitochondria, and subsequently published the

"Mitochondrial Theory of Aging,"[6] in 1972. The "Free Radical Theory of Aging" is still responsible for billions of dollars in antioxidant supplement sales a year. While writing this book, a fascinating article was published in the journal *Open Biology* by Dr. James Watson (of Watson and Crick) called "Oxidants, antioxidants and the current incurability of metastatic cancers."[7]

Watson sets forth a hypothesis that links the presence of antioxidants in the body with late-stage (metastatic) cancer. He points out that some successful cancer treatments use "free radical" molecules to treat cancer—the very same molecules that antioxidants attack and kill. Watson urges his readers of the new paper to consider the following: "Unless we can find ways of reducing antioxidant levels, late-stage cancer 10 years from now will be as incurable as it is today."[8]

The following was excerpted from the same article in *Open Biology*.[8] Please see the original document for references contained within the text.

> Free-radical-destroying antioxidative nutritional supplements may have caused more cancers than they have prevented.
>
> For as long as I have been focused on the understanding and curing of cancer (I taught a course on Cancer at Harvard in the autumn of 1959), well-intentioned individuals have been consuming antioxidative nutritional supplements as cancer preventatives if not actual therapies. The past, most prominent scientific proponent of their value was the great Caltech chemist, Linus Pauling, who near the end of his illustrious career wrote a book with Ewan Cameron in 1979, *Cancer and Vitamin C,* about vitamin C's great potential as an anti-cancer agent [52]. At the time of his death from prostate cancer in 1994, at the age of 93, Linus was taking 12 g of vitamin C every day. In light of the recent data strongly hinting that much of late-stage cancer's untreatability may arise from its possession of too many antioxidants, the time has come to seriously ask whether antioxidant use much more likely causes than prevents cancer.
>
> All in all, the by now vast number of nutritional intervention trials using the antioxidants β-carotene, vitamin A, vitamin C, vitamin E and selenium have shown no obvious effectiveness in preventing gastrointestinal cancer nor in lengthening mortality [53]. In fact, they seem to slightly shorten the lives of those who take them. Future data may, in fact, show that antioxidant use, particularly that of vitamin E, leads to a small number of cancers that would not have come into existence but for antioxidant supplementation. Blueberries best be eaten because they taste good, not because their consumption will lead to less cancer.

Don't Oxidants Cause Oxidative Stress?

Not necessarily. Dr. Andreas Kalcker presented us with another theory explaining why oxidative stress is caused by pathogens and parasites rather than oxidants. If you take an oxidant such as CD, it is likely to upset the parasites living in the intestinal tract, causing them to defecate. Waste from worms can contain many toxic substances including MDA (malondialdehyde), formaldehyde, and ammonia. These toxins are known to cause oxidative stress.[9,10,11,12] When parasites are present in the body they are releasing these toxins, even if we don't kill them. Parasitic presence in the body is responsible for many of the symptoms known as autism.

Oxidants themselves are not directly responsible for oxidative stress in the body. However, they can exacerbate oxidative stress, which can cause parasites in the body to release toxins. The solution is to get rid of the parasites. Please see Chapter 8 for an effective Parasite Protocol.

Chlorine Dioxide & Antioxidants

Although antioxidants have their place in personal health and are common in many foods and supplements, they cannot be taken at the same time with chlorine dioxide—they negate each other. This is why vitamin C, juices, and other antioxidant containing foods are prohibited in *The Diet*.

> Just a testimonial to MMS: For those of you new to the group and still unsure of the MMS protocol I wanted to tell you about us. I was afraid at first, but we decided to try MMS. Within 3 weeks my son went from an ATEC of 36 to 18! That was amazing. Then we were afraid of the enemas. We said no way, we would just do oral MMS. Of course, after seeing others progress here we decided to do them. He would calm down! Then for months we agonized over the PP. We were afraid of it. We were scared of the meds, everything. Finally we decided to try it too. We are in the middle of our first PP and our son, who used to hit, kick, bite and scream at me over the smallest things all the time, has now been a perfect joy. When he gets mad it is over quickly and does not escalate into world war 3. Also these bumps he had on his face have cleared up along with his attitude! So if you are unsure of MMS like I was, here is one more testimonial of just how good the protocol really is. Thanks Kerri, and all the other moms that came before me!

Herxheimer Reaction (Die-Off Reaction)

There are many stories on the Internet about the "side effects" of CD. The definition of a side effect is:

The unintended or secondary effects of a substance.

When people describe what they believe to be side effects of CD, they are actually talking about a Herxheimer reaction. Jim Humble's original chlorine dioxide (MMS) protocols called for higher doses per hour ramped up much faster than we titrate up for autism. In some cases that resulted in nausea, vomiting, and/or diarrhea. Critics of CD claim that a Herxheimer reaction, which is the result from ramping up too quickly, is the body's reaction to being 'poisoned' by chlorine dioxide.

This is absolutely not the case, as the aforementioned study by Judith R. Lubbers and colleagues found that when healthy individuals were given minimal doses of chlorine dioxide in increasingly higher concentrations, no deleterious effects were noted.[2]

A Herxheimer reaction indicates that the patient is suffering from an overload of pathogens/toxins. If the person taking CD has a fair amount of pathogens, the rapid killing of these invaders produces excess toxins in the bloodstream. It is these substances that the body wants to rid itself of in short order. People with autism often have impaired methylation cycles that reduce their ability to eliminate toxins, making it even more difficult to handle an influx of toxins. So LOW AND SLOW is the way to go.

Children with autism typically don't (or can't) tell you they have nausea. However, discomfort may be visible from seeing that they're eating less, their eating habits have changed, or they're more listless than usual. That's usually a sign of detox, and in which case, we won't give the child any more CD that day. Detox is great, it's what we want, but without stressing the body further. We need to stay just beneath a Herxheimer reaction.

A Herxheimer reaction can present itself in many forms including:

- Blisters
- Bloating
- Burping
- Chills
- Diarrhea
- Fatigue
- Flatulence
- Flushing
- Headache
- Heart Burn
- Hiccups
- Nausea
- Runny nose
- Sleep changes
- Vomiting

If any of these symptoms appear, suspend the remaining doses of CD for the rest of the day. The following day, return to the number of doses given when the child was stable, which is usually the dose that was given the day prior to the Herxheimer reaction.

To mitigate a Herxheimer reaction from chlorine dioxide, you can give vitamin C or orange juice, as either one of these will cut the effects of CD immediately. To help bind the excess toxins you may want to try activated charcoal, bentonite clay, (be aware that these both can cause constipation), or burbur (a liquid detox extract).

Using Chlorine Dioxide for Autism

As I developed this protocol, it became clear to me that we needed to focus on excesses causing the symptoms of autism, rather than deficiencies caused by them. In the autism biomed world, we are accustomed to supplementing deficiencies (low iron, low B_{12}, low calcium, etc.), without questioning why the deficiency exists in the first place (i.e., presence of pathogens/parasites).

As soon as a family completes one week on *The Diet*, we get right into CD. Everyone starts with the "baby bottle" method, one drop of CD in eight fluid ounces of water, which makes the first day's dose 1/8 of a drop at a time. We have greatly minimized Herxheimer reactions by using this method, but it remains very important to monitor your child while they are using chlorine dioxide. It's very important to go low and slow. Every drop we give is helping to heal our children.

What is Chlorine Dioxide (CD)?

Refer to the diagram on the right for a visual overview.

You can't actually buy CD. You have to make it. Fortunately, it's not a big deal. If you can make coffee... you can make CD, which stands for *chlorine dioxide*.

Chlorine dioxide (ClO_2), is a gas produced through a chemical reaction when two liquids, sodium chlorite ($NaClO_2$) and an acid (such as hydrochloric acid (HCl) or citric acid ($C_6H_8O_7$)), are mixed. The acid brings the combined pH level below five, causing the sodium chlorite to become unstable and release chlorine dioxide (ClO_2).

If any of this sounds complex, it isn't. Go through the following explanation and don't get bogged down by the chemical symbols.

Conceptual Overview of Chlorine Dioxide (CD)

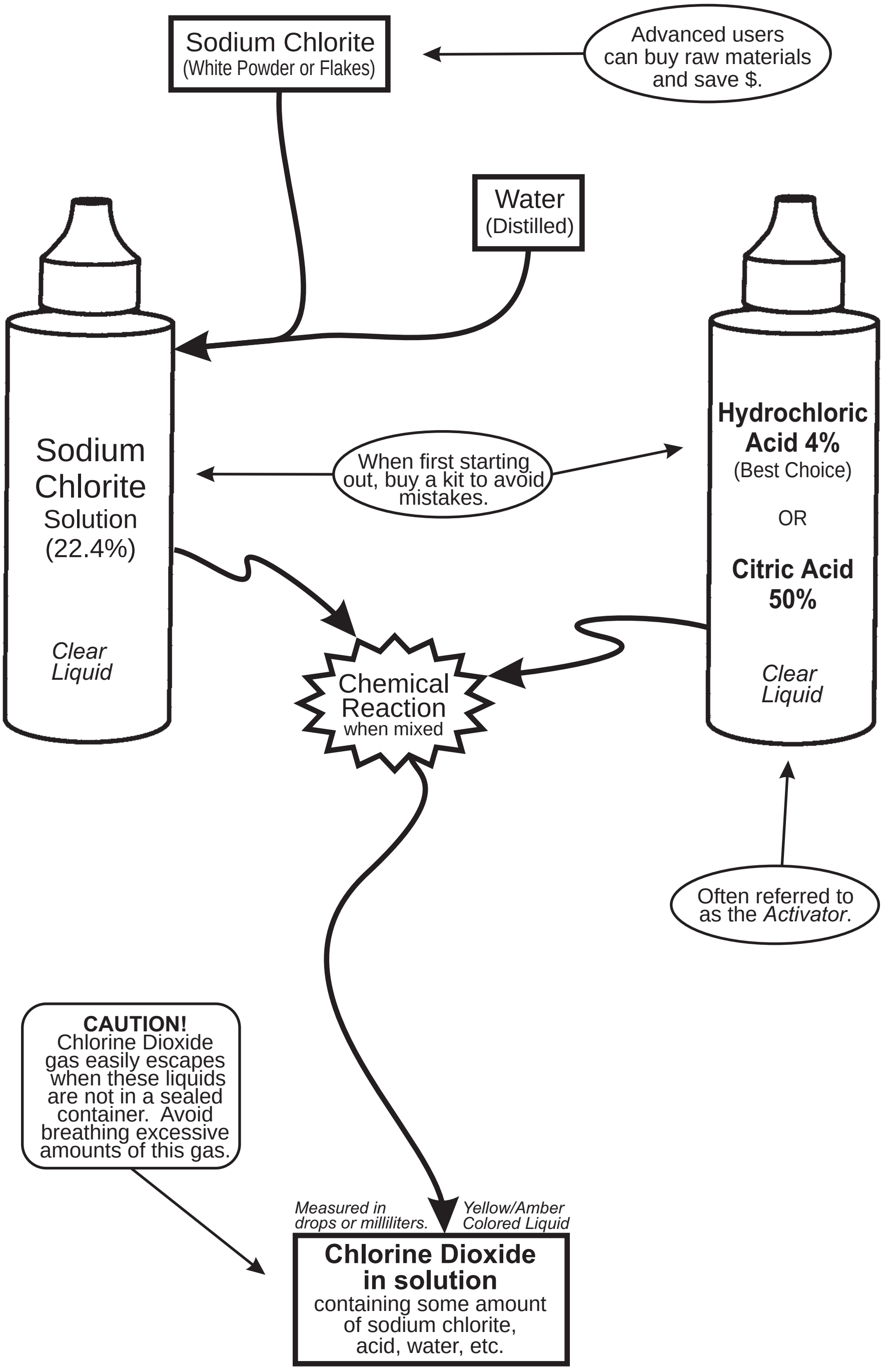
Sodium Chlorite
(White Powder or Flakes)
Advanced users can buy raw materials and save $.
Water
(Distilled)
Sodium Chlorite Solution (22.4%)
Clear Liquid
When first starting out, buy a kit to avoid mistakes.
Hydrochloric Acid 4%
(Best Choice)
OR
Citric Acid 50%
Clear Liquid
Chemical Reaction
when mixed
Often referred to as the Activator.
CAUTION!
Chlorine Dioxide gas easily escapes when these liquids are not in a sealed container. Avoid breathing excessive amounts of this gas.
Measured in drops or milliliters.
Yellow/Amber Colored Liquid
Chlorine Dioxide in solution
containing some amount of sodium chlorite, acid, water, etc.

Sodium Chlorite ($NaClO_2$)

This is the first chemical you must have to make CD.

You can get it in one of two forms:

- **Sodium chlorite liquid in water:**
  - If you are beginning, get a starter kit of this.
  - Expect to pay about $20-$30 for a 4 fl. oz. bottle set (usually includes acid activator).
  - 22.4% sodium chlorite solution (from 28% technical grade sodium chlorite salts).
  - Recommended: Buy quart or liter sized quantities of liquid once you understand the principles.
- **Sodium chlorite powder or flakes:**
  - FOR ADVANCED USERS ONLY!
  - Cheaper in the long run.
  - Recommend buying 5 lbs. of flakes/powder for a cost of about $100-$200.
  - Look online or check with your supplier for mixing instructions.
  - Avoid buying 90% sodium chlorite flakes. You want 80%. Most suppliers that sell liquid also sell flakes.
  - NEVER mix any acid directly with powder/flakes!

What's the 28%? 22.4%?

This is not a critical thing for you to worry about unless you are mixing your own. But here is the jist: sodium chlorite powder is not 100% sodium chlorite. It's usually 80% sodium chlorite and 20% salt and other inert ingredients. When diluted in water, you end up with 28% solution of powder, but only 22.4% is sodium chlorite. Labels usually show one or both of these numbers.

Sodium Chlorite Flakes vs. Powder

Sodium chlorite flakes and sodium chlorite powder are the same thing. The resulting physical consistency may vary by different manufacturers. We've even seen clumps of powder. It's still the same chemical. The important part is that it is technical grade, and 80% of the powder/flakes should be sodium chlorite ($NaClO_2$). Mixing info is at: *www.youtube.com/watch?v=t0C8XVm9QVE*.

Acid - The Activator

To produce CD, you must add the second chemical, which is a food acid, to the liquefied sodium chlorite (22.4% solution in water). Chlorine dioxide gas is produced when the pH of the sodium chlorite goes below five. We call the acid used to lower the pH of sodium chlorite an "activator" because it activates the sodium chlorite to produce the chlorine dioxide.

There are several activators to choose from. These are food acids in that they are commonly found in food preparations, or are actually food:

- **Hydrochloric Acid (HCl)**
  - The same acid found in your stomach, and therefore, the most biocompatible option.
  - Working concentration is between 4% and 10%.
  - Available on the market in liquid concentrations of 4%, 5%, 10%, 32%, 35%, and 37%.
  - Any concentration over 10% must be reduced to a working concentration of between 4% and 10%. See Appendix 7, page 471. (Note: We use 4% for CD and CDH; & 10% for CDS.)
  - Only available in liquid form.
- **Citric Acid**
  - Commonly available around the world.
  - Often found in soft drinks to give them a tangy flavor.
  - Some people have problems tolerating this acid.
  - Usually used in 50/50 concentration by weight mixed in water.
  - Available in liquid and powder form.
  - Inexpensive; can be purchased online.
- **Lemon or Lime Juice**
  - Consider this an emergency alternative to be used if the first two choices are not available.
  - Can be found in most stores carrying produce.
  - Should be from pure, organic and freshly squeezed lemons (or limes).
  - Relatively weak acidity requiring five drops of lemon juice to every one drop of sodium chlorite solution.
  - Substantially longer activation time.
  - Lower ClO_2 output.

The amount of acid used, and the length of time you wait to dilute the mixture, will depend on the acid activator and its strength. The following chart gives you an idea.

Activator	Strength	Ratio	Time
Hydrochloric Acid	10%	1:1	40 secs
Hydrochloric Acid	4%	1:1	60 secs
Citric Acid	50%	1:1	60 secs
Citric Acid	33% or 35%	1:1	60 secs
Lemon Juice	Pure	1:5	120+ secs

You are unlikely to be using lemon juice, but we've included it here for completeness.

Families should use whichever activator is available to them depending on where they live and what they have access to. All kids are different, so if yours doesn't respond well to 4% HCl, by all means, try citric acid, or vice versa.

If the HCl you purchased is anything other than 4% or 5%, then see Appendix 7 for instructions on reducing concentration.

If you are using one of the two most commonly available activators, then you are likely to be using a 1:1 mixing ratio (1 drop sodium chlorite + 1 drop activator).

In the case of 10% citric acid or pure lemon juice, you would be using five drops of activator to one drop of sodium chlorite, and waiting about three times as long as if you were using one of the newer, more concentrated activators.

Purchasing Product

When you first start out, you should buy a starter kit and get familiar with the chemistry. A starter kit usually comes in two bottles of four fluid ounces each.

1. Sodium chlorite, (4 fl. oz.)
2. Activator (hydrochloric acid or citric acid), (4 fl. oz.)

The bottles look something like those shown in the diagram on page 91. Getting the hydrochloric acid activator is slightly preferable.

You will also need an 8 fl. oz. baby bottle. This is a handy tool for storing a day's worth of CD doses with the added benefit of having marks that show how much it contains. Glass bottles are much better than any kind of plastic bottle. Lifefactory.com (shown on the right) produces a bottle we especially like. It is made of silicon-covered glass and has a reliable top. This particular bottle is marked as having a capacity of nine fl. oz., but we use it up to only eight fl. oz. Plastic should only be used for lids; avoid bottles with metal tops. Over time, CD will oxidize metals so they can't be used.

Another item that most people have at home is a shot glass or wine glass that tapers down in the middle like a "U" shape. This will be used to mix the chemicals and the tapered shape insures that even as few as two drops will find each other. You can also use a cup with a flat bottom, but must tilt it so both solutions pool together.

lifefactory

A standard shot glass with a "U" shaped bottom is the preferred mixing container for CD & activator. A glass baby bottle, such as the one made by Lifefactory®, is the ideal dosing bottle.

Terminology of Available Products

There can be a bit of confusion for some as to what to look for and buy. The following chart shows the actual chemical, and some of the common names found on the label. There may be other names we are not aware of. In most cases, the chemical and its strength are identified on the bottles.

Common Activators & Their Various Names on the Market	
Chemical	**Common Names**
Sodium Chlorite	Chlorine Dioxide – Part 1 (Purifier) Master Mineral Solution Miracle Mineral Solution Miracle Mineral Supplement MMS MMS Drops MMS Solution MMS Water Purification Drops MMS1 NaClO2 Natriumchlorit (German) Pathogen Biocide Sodium Chlorite 28% Water Purification Drops Water Purificator
Hydrochloric Acid	Activator Chlorine Dioxide – Part 2 (Activator) MMS Activator
Citric Acid	Activator Citric Acid – 20% Citric Acid – 50% MMS Activator

Checklist for Preparing a CD Dose

Assuming you have received the starter kit, it's time to make your first dose. Here's the check list of what you now should have on hand before you begin:

- ☐ Bottle of sodium chlorite solution (28% sodium chlorite salts / 22.4% sodium chlorite)
- ☐ Bottle of acidic activator (hydrochloric acid or citric acid)
- ☐ A clean and dry shot-glass with a "U" shaped bottom.
- ☐ 8 fl. oz. glass baby bottle.
- ☐ Distilled or reverse osmosis water (filtered water will do, but no alkaline water!).

A Few Points Worth Noting!

Looking at the short checklist above, it should be apparent that this is not going to be rocket science. Nevertheless, a few pointers are in order:

- In the past, people who used CD for various things would mix a fresh batch each time they took CD. This involved a lot of extra work that could be avoided by producing one batch big enough to last the entire day. For this exercise, we are assuming you are starting out and going to do things low and slow. We are going to produce 8 fl. oz. of water solution containing one drop of CD, with the intention of your child consuming one fl. oz., of water (containing 1/8 drop of CD) eight times throughout the day.

- When you mix sodium chlorite solution and the acidic activator, you are mixing equal drops of each substance (unless you are using one of the less concentrated activators mentioned earlier).
- CD doses are measured in drops of sodium chlorite. So one drop of sodium chlorite + one drop of activator = one drop of CD.
- On your first day, start with one drop of sodium chlorite and one drop of activator (in other words—ONE drop of CD).
- If you put in too many drops of one solution or the other, start over with a clean and dry shot glass.
- A batch of CD is good for 24 hours. However, you must keep the cap on tight!
- Keep the bottle with the daily CD batch out of direct sunlight. A brief exposure won't ruin it, but it will become weaker. Why? The chlorine dioxide gas will come out of the water and sometimes form a cloud in the air pocket of the bottle. No need to worry if you see the cloud—it's a common occurrence.

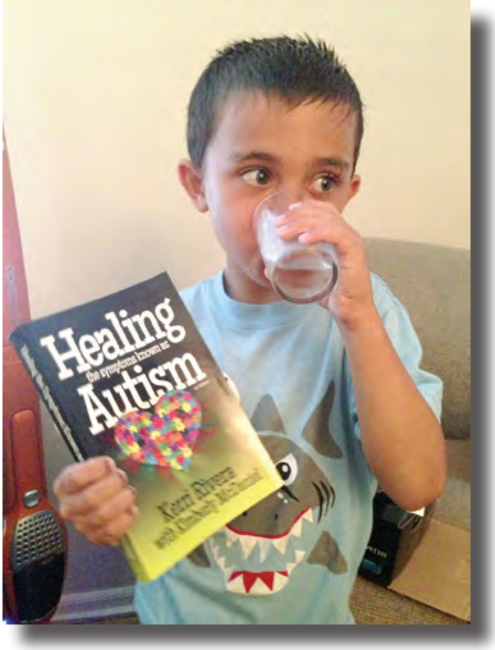

Vincent taking his hourly dose.

Preparing Your First One drop CD Batch for the Day

1.	Start by filling the baby bottle with 8 fl. oz. (237ml) of water (distilled or reverse osmosis). NO alkaline water!	
2.	Place 1 drop of sodium chlorite solution into the CLEAN and DRY shot glass.	
3.	Add 1 drop of acidic activator (hydrochloric acid or citric acid) to the shot glass containing the drop of sodium chlorite. The number of drops may be higher if you are using a weaker activator. See chart on page 94.	
4.	Now wait the appropriate time for the mixture to react (see chart on page 94). You should see the color change from clear to slightly yellow. If there were more drops in the shot glass, the color change is more noticeable. You are also likely to notice the chlorine-like smell coming from the shot glass. Remember, this is NOT chlorine, but rather chlorine dioxide.	
5.	After the activation time has passed, pour a little water from the baby bottle into the shot glass and let it mix. This mostly stops the chemical reaction and insures you get most of the mixture out in the next step.	
6.	Lastly, pour all of the watered down mixture in the shot glass back into the baby bottle, and seal it tightly with the cap—don't leave it sitting open for any length of time. Think partially used soda pop and how you would want to keep that closed.	

Voilà! You have just produced your first batch of CD!

This is now a ONE drop batch, that makes EIGHT hourly doses—each 1/8 of a drop in strength. If your batch is larger; your hourly dose is stronger:

- 2 drop batch = 8 hourly doses of 1/4 drop
- 4 drop batch = 8 hourly doses of 1/2 drop
- 8 drop batch = 8 hourly doses of 1 drop

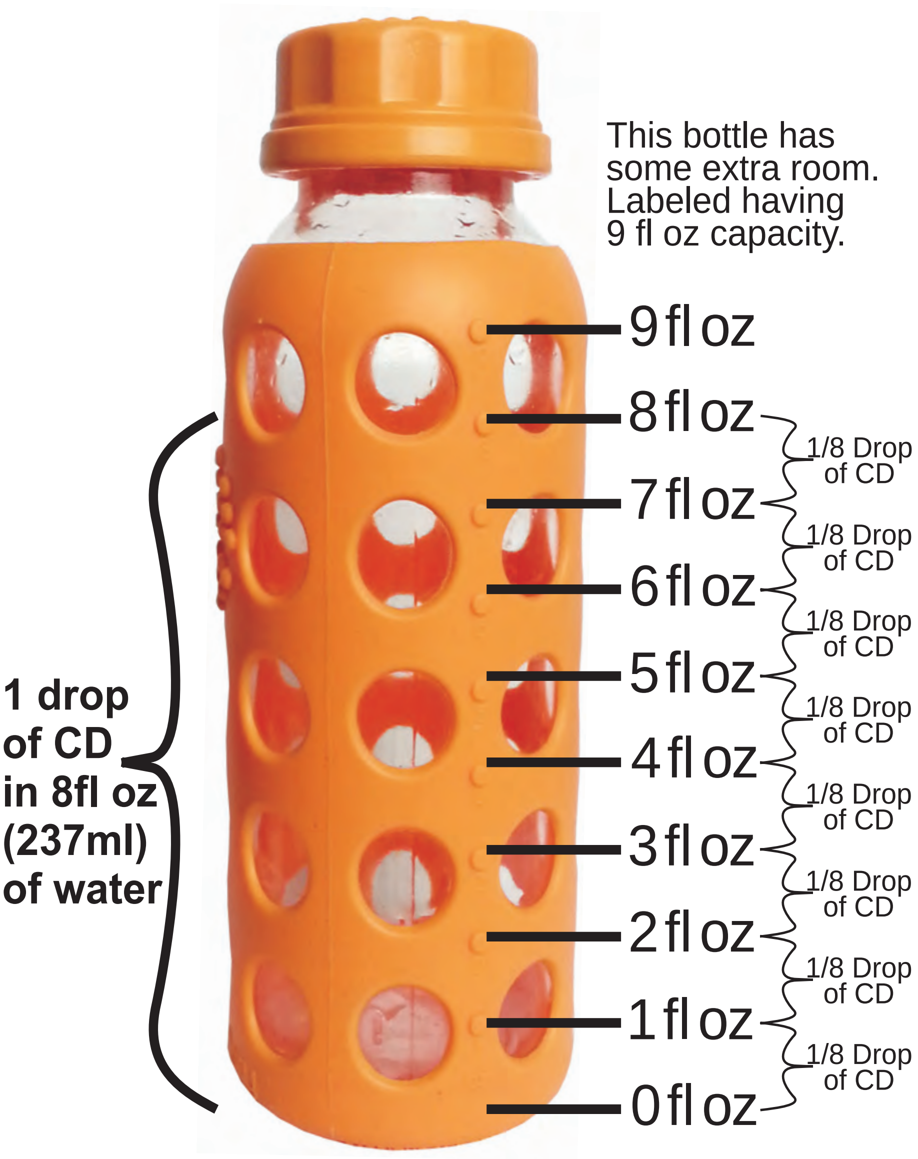
This bottle has some extra room. Labeled having 9 fl oz capacity.
9 fl oz
8 fl oz
7 fl oz
6 fl oz
5 fl oz
4 fl oz
3 fl oz
2 fl oz
1 fl oz
0 fl oz
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1 drop of CD in 8fl oz (237ml) of water

Oral Protocol for CD in Autism Spectrum Disorders (ASDs)

In the previous exercise, you learned how to prepare the dose for Day One. Children and adults all start by taking only one drop of CD divided into eight hourly doses. On Day Two, you make a batch of two drops sodium chlorite plus two drops activator and mix that into eight fluid ounces of water. Increase this dosage by one drop per day, unless you see a Herxheimer reaction (see page 89 for more information).

The chart on the right allows you to determine the estimated full oral CD dose based on the weight of the individual. This gives you an idea of the maximum goal number of drops per day for your child. Once the child has been at the full suggested dose for several months, we may need to increase again, by a few drops, to prevent a plateau.

Important Note: Some children can't go up by one drop each day and need to go slower. A child may have a relatively higher toxic load, or he may be extremely sensitive. Whatever the case, resist the urge to try to get to full oral dose as quickly as possible. This is not a race! Some children need to spend a few extra days at the same dose to avoid a Herxheimer reaction. Even if you must remain at one level for a few days or more does not mean you are making no progress. Whatever the level of dosage, you are detoxing the body.

When you get within two to three drops of your child's maximum dose by weight, go ahead and spend three days at each dose, while carefully observing your child, moving up only when your child is stable. For example, if the full dose is 16 drops in the 8 fl. oz. (237ml) baby bottle, stay at 14 drops for a few days to see how that dose level is tolerated, and then increase every third day till reaching the full dose. If the person is better at a drop or two below the calculated full dose, go back to that lower dose that they tolerated best.

> Just had a fascinating appointment with my son's functional neurologist. My son had this third Qeeg today, which clearly showed that previously under-stimulated parts of his brain are 'waking up'. I told our doctor about CD very early on and he believes that it is THIS that is having such a huge impact on his brain. He got quite emotional and said 'every time you get a worm, you are saving your son. Keep doing what you're doing, cause he clearly isn't autistic. He's poisoned'. Goosebumps!!

Estimated Full Oral CD Doses by Weight

Use these numbers as a guide only. You may need to go up by as much as 50% or more over the indicated drops. Read chart as: **POUNDS / KILOGRAMS → DROPS OF CD (per 8 fl. oz. water)**

25/11→8	62/28→17	99/45→24	136/62→29	173/78→33	210/95→37
26/12→8	63/29→17	100/45→24	137/62→29	174/79→33	211/96→37
27/12→9	64/29→17	101/46→24	138/63→29	175/79→34	212/96→37
28/13→9	65/29→18	102/46→24	139/63→30	176/80→34	213/97→37
29/13→9	66/30→18	103/47→24	140/64→30	177/80→34	214/97→37
30/14→9	67/30→18	104/47→25	141/64→30	178/81→34	215/98→37
31/14→10	68/31→18	105/48→25	142/64→30	179/81→34	216/98→37
32/15→10	69/31→18	106/48→25	143/65→30	180/82→34	217/98→37
33/15→10	70/32→19	107/49→25	144/65→30	181/82→34	218/99→38
34/15→10	71/32→19	108/49→25	145/66→30	182/83→34	219/99→38
35/16→11	72/33→19	109/49→25	146/66→30	183/83→34	220/100→38
36/16→11	73/33→19	110/50→26	147/67→31	184/83→35	221/100→38
37/17→11	74/34→19	111/50→26	148/67→31	185/84→35	222/101→38
38/17→11	75/34→20	112/51→26	149/68→31	186/84→35	223/101→38
39/18→12	76/34→20	113/51→26	150/68→31	187/85→35	224/102→38
40/18→12	77/35→20	114/52→26	151/68→31	188/85→35	225/102→38
41/19→12	78/35→20	115/52→26	152/69→31	189/86→35	226/103→38
42/19→12	79/36→20	116/53→26	153/69→31	190/86→35	227/103→38
43/20→13	80/36→21	117/53→27	154/70→31	191/87→35	228/103→38
44/20→13	81/37→21	118/54→27	155/70→31	192/87→35	229/104→38
45/20→13	82/37→21	119/54→27	156/71→32	193/88→35	230/104→39
46/21→13	83/38→21	120/54→27	157/71→32	194/88→35	231/105→39
47/21→14	84/38→21	121/55→27	158/72→32	195/88→36	232/105→39
48/22→14	85/39→21	122/55→27	159/72→32	196/89→36	233/106→39
49/22→14	86/39→22	123/56→27	160/73→32	197/89→36	234/106→39
50/23→14	87/39→22	124/56→28	161/73→32	198/90→36	235/107→39
51/23→15	88/40→22	125/57→28	162/73→32	199/90→36	236/107→39
52/24→15	89/40→22	126/57→28	163/74→32	200/91→36	237/108→39
53/24→15	90/41→22	127/58→28	164/74→32	201/91→36	238/108→39
54/24→15	91/41→23	128/58→28	165/75→33	202/92→36	239/108→39
55/25→16	92/42→23	129/59→28	166/75→33	203/92→36	240/109→39
56/25→16	93/42→23	130/59→28	167/76→33	204/93→36	241/109→39
57/26→16	94/43→23	131/59→28	168/76→33	205/93→36	242/110→39
58/26→16	95/43→23	132/60→29	169/77→33	206/93→37	243/110→39
59/27→16	96/44→23	133/60→29	170/77→33	207/94→37	244/111→40
60/27→17	97/44→24	134/61→29	171/78→33	208/94→37	245/111→40
61/28→17	98/44→24	135/61→29	172/78→33	209/95→37	246/112→40

Serving Cup Dilution

Don't have your child drink out of the baby bottle. Pour the hourly dose into a separate cup and add extra water if needed. Make sure to add water to each dose to ensure that there is a **minimum** of one fluid ounce of water per activated drop of CD. However, there is no maximum amount of water for dilution. To be absolutely clear about this, note the following chart which assumes eight fl. oz., daily batches containing eight hourly doses.

Drops of CD in Batch	Hourly Dose	Min. Water Dilution
1	1/8 drop	1 fluid ounce
2	2/8 drop	1 fluid ounce
3	3/8 drop	1 fluid ounce
4	4/8 drop	1 fluid ounce
5	5/8 drop	1 fluid ounce
6	6/8 drop	1 fluid ounce
7	7/8 drop	1 fluid ounce
8	1 drop	1 fluid ounce
9	1 1/8 drops	2 fluid ounces
10	1 2/8 drops	2 fluid ounces
11	1 3/8 drops	2 fluid ounces
12	1 4/8 drops	2 fluid ounces
13	1 5/8 drops	2 fluid ounces
14	1 6/8 drops	2 fluid ounces
15	1 7/8 drops	2 fluid ounces
16	2 drops	2 fluid ounces
17	2 1/8 drops	3 fluid ounces
18	2 2/8 drops	3 fluid ounces
19	2 3/8 drops	3 fluid ounces
20	2 4/8 drops	3 fluid ounces
21	2 5/8 drops	3 fluid ounces
22	2 6/8 drops	3 fluid ounces
23	2 7/8 drops	3 fluid ounces
24	3 drops	3 fluid ounces
25	3 1/8 drops	4 fluid ounces
26	3 2/8 drops	4 fluid ounces
27	3 3/8 drops	4 fluid ounces
28	3 4/8 drops	4 fluid ounces
29	3 5/8 drops	4 fluid ounces
30	3 6/8 drops	4 fluid ounces
31	3 7/8 drops	4 fluid ounces
32	4 drops	4 fluid ounces

CD is active in the body for about one hour and, as result, hourly doses give us the best progress. This is difficult on school days, but we recommend that you dose two to three times before school and the rest after to distribute the pathogen-fighting more evenly. Eight doses a day is the minimum requirement, but the more you can fit in the better. Try to dose every hour as much as possible.

Special Cases: "The Double Dose"

As a child goes through the detox process, they may exhibit conduct which can include hyperactivity, anxiety, OCD, aggressiveness, and night wakings. In these cases, a double dose is 2 fl. oz. (60ml) of the baby bottle given at once. For example, if your child receives one drop per fluid ounce of the baby bottle method, they will be receiving 2 drops. Let's say for example your child wakes up at 1am, give them a double dose. If your child is still awake at 2am, give another dose. The most common scenario is that after 1 or 2 double doses, the child will fall back asleep. This helps oxidize the toxins from the die-off. During the day if your child has a tantrum or other issues with conduct, we can double dose every hour as well.

CD has two functions with pathogens; one is that is oxidizes the pathogen itself, and the second is that it oxidizes the toxins released by the pathogens while they are living, dead, or dying.

Advanced Protocol – The Enema!

In early 2011, we added enemas to the protocol to kill the pathogens causing dysbiosis in the large intestine (we didn't know about parasites yet). We wanted to get the chlorine dioxide into the blood stream so it could kill the biofilm that exists in the blood. In this way, the blood can carry the CD past the blood-brain barrier to kill pathogens in the brain

When we are detoxing, it is absolutely critical to keep the colon moving and avoid the reabsorption of toxins through the intestinal walls. Enemas allow us to do just this. Some toxins can exit the intestine through the intestinal wall (more so if leaky-gut syndrome is present), and cross the blood-brain barrier, therefore affecting cognition and behavior. When we cleanse the colon, we get those out before they can cross into the brain, and we detoxify the lymphatic system, liver, and gallbladder.

The Enema Stigma

When I say the word enema to a family new to the protocol, I often hear things like "*I could never do that!,*" or "*My son would never let me do an enema.*" I assure you it's not traumatic as it has been hyped up to be. Modern society has gotten away from it, but for millennia, cleansing the colon was considered part of daily hygiene. Enemas were mentioned in the Vedas, the Bible, and by the Mayans. The reason for doing CD enemas is NOT for relief of constipation. Of course, when constipation is present, CD enemas can relieve the cause of constipation.

> Great news! First enema today...
>
> My son was so happy that we did and the first thing he said while sitting on the toilet was: "Mommy! IT WORKED!!!" hahaha He is so used to sitting on the toilet and having such a hard time. Poor thing, he was so happy that IT WORKED!
>
> And he even said he wants to do it again tomorrow. He felt so much better and ate so much more for dinner with an empty stomach. I think he got rid of stuff that was there for ages! His tummy is now soft and not hard as usual. Quite amazing and it didn't bother him at all.
>
> Today his molar hasn't bothered him so a perfect day to start the enemas! His appetite has gone up quite a bit. He has seldom asked for food before and now he asks a few times a day for something to eat. The diet is going wonderfully and he is now used to the change of foods. He asks about cheese and milk but I try to talk about something else to distract him so he eventually stops asking. hahaha He hasn't gotten upset about food he wants (that he can't have) not even once. YAY!

I would like to clear up one thing here before we get in the mechanics of the enemas. More than one person has asked in the blogosphere, "*If Kerri Rivera loves enemas so much, why doesn't she do them on herself???*" As a matter of fact I do! I have done my own enemas for years, so has my husband, my sister, and plenty of the parents using this protocol for their children. We understand the benefits of a fit colon, and have no problem applying our own enemas or going for a colonic from time to time.

Actually, I would strongly advise you to practice on yourself. That way you know from first-hand experience that this is really no big deal. You will also discover how much better you feel after cleaning out your colon, and you will understand when your child feels better as well.

When to Start Giving Enemas

Enemas should be started when the child reaches full oral dose.

However, if your child suffers from constipation you can add enemas and baths on day one (see page 113), before you get to the full oral dose. However, alternate them on opposite days or opposite ends of the day.

Frequency of Enemas

I recommend enemas no less than once every other day. You may need to do them more often, particularly if you see parasites coming out. Doing as many as four in a day is not uncalled for if each one results in parasites being released. Some of the older children seem to pass a lot of parasites in one day, typically in or around a full or new moon.

Timing of Enemas

If your child is going to school, avoid giving an enema in the morning since they may need to visit the toilet a few times after. If the child is homeschooled or too young or sick to attend, the timing is up to you. Ideally, do the enema after the child has a bowel movement.

Enemas — What You Need

The first thing you are going to need is some kind of a large plastic, liquid measuring container. Glass is not recommended since it can get knocked over and break in the confines of a bathroom.

You will also need all the same things you used for making CD batches:

- Water. Use reverse osmosis, filtered, or distilled.
  DO NOT use tap water or alkaline water!
- Sodium chlorite solution
- Acidic activator
- Shot glass
- Natural lubricant (coconut oil or any other simple food oil). NO perfumed baby oil or chemical lotions such as Petroleum Jelly or K-Y® Jelly.
- Enema kit (see your options on the following page)

Enema Options (Most to Least Preferred)

Gravity Bag or Bucket with hose

PiC indolor
IRRIGADOR
ENTEROCLISE
KLISTEERSPUIT
ΚΛΙΣΜΑ
CE

Pros:

- Holds plenty of liquid
- Can be refilled without taking it down
- Flow is easy to control
- Can be hung on a door
- Can be used for both enema and douching
- PIC Indolor: Folds for easy travel
- PIC Indolor: Has valve
- PIC Indolor: Two sizes of cathetors

Cons:

- Bucket doesn’t fold for easy travel

Enema / Douche Bag / Hot Water Bottle

Hot / Cold
Water Bottle System
Hot/Cold Therapy Anytime... Anywhere!
Safe Durable Hygienic

Pros:

- Holds plenty of liquid
- Flow is easy to control
- Can be hung on a door
- Can be used for both enema and douching
- Easy travel
- Has two sizes of cathetors

Cons:

- Can’t be refilled without taking down
- Usually made of rubber, which tends to dry out quickly, especially from CD.

Syringe & Catheter

Pros:
- Flow is easy to control
- Can be used for both enema and douching
- Good for small children

Cons:
- Holds only 60ml-140ml
- Need more than one.
- Have to swap them out to go beyond the capacity of one syringe.
- The graduations rub off easily (suggest putting clear shipping tape over them to protect the markings if they are important to you).

Enema Bulb

Pros:
- Portable
- Can be used for both enema and douching
- Good for “implants” of mebendazole

Cons:
- Volume too small for an effective treatment.
- Hard to clean inside
- Can’t tell if dirt is on the inside.
- Holds only about 200ml.

Fleet Enema (Last Resort)

Fleet
Enema
saline
Twin Pack
LAXATIVE
ready-to-use
latex free
Soft, Flexible Comfortip
EASY SQUEEZE

Pros:
- Cheap
- Can find it in most pharmacies
- Portable
- Can be used for both Enema and Douching

Cons:
- Volume too small for an effective treatment.

Enema Kits

As shown in the previous two pages, here are your enema kit options listed from best to less than good. Prices are estimates in US Dollars and are likely to be higher outside the US:

★★★★★ PIC Indolor Gravity Bag (~$15-$30)
★★★★☆ Enema Bucket (~$5-$10)
★★★⯪☆ Multipurpose Enema/Douche Bag/Hot Water Bottle (~$15-$30)
★★★☆☆ 100ml+ Syringes plus Catheter (~$5-$10)
★★☆☆☆ Enema Bulb (~$5-$15)
★☆☆☆☆ Fleet Enema (emptied out) (~$5-$15)

You must have a separate enema kit for each person doing enemas. This insures that you are not cross-contaminating parasites.

In my own home we have recently started using the PIC Indolor enema/douche bag. This has made the process faster, and we can tolerate more water before feeling full and desiring to release the enema. This is probably the most practical kit, especially because of the built-in shut-off valve. The enema bucket is a close second, but the sliding clamp that comes with it doesn't last long and breaks easily. You can always just kink the hose and control the flow that way.

The multipurpose enema/douche/hot water bottle type is a good option, but it is usually made of rubber. This material often goes bad over time and CD just speeds up that process. Rinse it out several times before and after use to insure that no breakdown or CD residue is left. Don't use if you notice the rubber becoming brittle.

Even two-year-olds still use about 400 to 500ml in an enema. The intestine is long. Parents with small children sometimes like the 100ml syringe/catheter method. The enema bulb or fleet enema are not really practical for our purposes due to the small volumes they hold. However, if you want to get started right away, they are usually available in most major pharmacies.

You can find syringes, catheters, gravity bags, multipurpose enema/douche bags on various online sites, pharmacies and medical supply stores around the world.

Preparing and Administering the Enema

Make all your supplies easily accessible in your work area before bringing your child in the bathroom. If this is your first time and you are working out the process, you may wish to ask your spouse or someone else for assistance. It's a good idea to do a dress rehearsal without the child, to insure that it goes smoothly when the child is present. Once you and your child have the process down, it will be easy. Being prepared is the key to success.

Please note: The eight fluid ounce baby bottle is NOT involved in any way with enemas!

1. Ideally, time the enema just after a bowel movement.
2. Start by washing your hands to insure you are not introducing any new bacteria into the colon. Everything you are working with should be clean. Wearing gloves and a face/surgical mask is HIGHLY recommended.
3. Assuming you are using one of the suggested gravity bags, find a good and stable location to hang it up. The best location really depends on your particular bathroom's layout and where you are going to place the child during the procedure. I hang mine from a towel hook in the bathroom, or you can use the shower curtain rod. The idea is to have the bag/bucket about three feet (~one meter) over where the enema will be administered. Most parents put a towel on the floor of the bathroom and lay their child on the towel (smaller children can lay across your lap).
4. Refer to the chart below and measure out the appropriate amount of water for the individual receiving the enema. The ratio is two drops of CD per 100ml of filtered warm water.

Age/Size	Water Volume	Drops of CD
Child	1/2 Liter (500ml)	10
Adolescent	1 Liter (1000ml)	20
Teen/Adult	up to 2 Liters (2000ml)	40

5. Don't add any CD yet. The next step is to heat the water to body temperature (98°F / 37°C). The water should feel warm to the touch—NEVER hot! Similar to the temperature of milk in a baby bottle... test it on your wrist. If you are concerned about getting the temperature right use a quality thermometer. Be aware... many cheap thermometers are notoriously inaccurate!

6. After the heating step, pour the water back into the big measuring cup.

7. Referring again to the chart on the previous page, prepare a CD mixture in the clean dry shot glass that's appropriate for the volume of water needed. For example, if you have an adolescent and are preparing one liter, then you will need to mix 20 drops of sodium chlorite and 20 drops of acidic activator to end up with 20 drops of CD! Note: The most important part of the enema is the ***water***. Resist the urge to add more than two drops per 100ml of water.

8. After the appropriate activation time, pour the concentrated CD from your shot glass into the large measuring cup. You may want to rinse out the shot glass with some water from the large mixing cup so you get all the concentrated CD solution into the water.

9. The bag/bucket should now be in its working location with the hose firmly attached, Place the end of the hose so it is over the top of the bag/bucket. A clothes pin can come in handy to hold it up over the bag/bucket. Now, fill the bag/bucket about 3/4 full with the CD water mixture you just prepared.
10. Make sure the hose is filled with water. You may want to let a bit run out into the sink or some other container and then clamp it. The bucket type enema has a plastic clamp for this purpose. The bag type has a valve where the catheter attaches. What you don't want is for there to be a bunch of air in the hose that then is transferred into the colon.
11. If all of the above is ready, it is time to bring your child into the bathroom.
12. Explain what you are going to do. Some parents say that they are going to put the catheter in his bottom, and that it has "special water" in it to help him feel better. One little boy called it "butt medicine" while another called it "butt wash." We recommend you say that you are the only person who is allowed to give this to him. Just say whatever you need to say to explain what is happening.
13. If the child is on the floor (on a rug or towel), they should lie on their stomach. If the child is small then laying them on your lap is best. You can also have them in a doggy position on the floor.
14. Place one or two drops of coconut oil (or similar) on their anus so the area is well lubricated.

15. Now take the hose and gently insert the tip, (one to two inches is plenty—more is not necessary) as the sphincter muscle will hold the tip of the catheter in place. If you are having any problems getting the catheter in, then you may need to add a bit more oil and/or adjust the angle.

16. Release the clamp or open the valve to allow the liquid to flow. The goal is *fill & release.*—NOT *fill & hold*. We are imitating the colonic process that is made up of gallons of water, yet in the home environment.

17. Your child will most likely hold the contents on their own—they will squeeze hard so as not to "go." This kind of happens naturally.

We do a fill and release method. The first part of the enema is usually stool. After that, we usually see parasites and or biofilm.

It's a learning experience and as time goes on, you will get the hang of it.

> *Quote from a teenager using the Protocol: "I like the enemas. Before them I am mad. But then when I get one, I feel better. My mad leaves in my poop.*

Syringe/Catheter Method

Before I started to use the gravity bag, the syringe/catheter method was my favorite. We have a YouTube video on our channel that shows how to fill the syringes and switch them out while using the catheter. If you don't have access to that video here is an explanation on how to do it:

1. Fill a bowl with body temperature filtered water and fill your syringes. If your child uses 500ml for an enema, fill all five syringes and dump out whatever water is still left in the bowl.
2. Squirt all the water from the syringes back into the bowl. You now have exactly 500ml of purified water.
3. Activate your drops of CD in a clean dry shot glass. If you are using ten drops, activate them and then use the water from your bowl to stop the activation process.
4. Pour the contents of the shot glass back into the bowl.
5. Refill your syringes with the activated CD solution.
6. Lubricate the tip of your catheter with coconut oil, and attach the first syringe.

Note: If you wish to preserve the graduation marks on the syringes, cover them with some clear plastic shipping tape. The printed on marks are often easily rubbed off and the tape will prevent that.

Fleet Enema Method

The third option for enemas is to use a Fleet enema, which can be purchased at any drug store. This method is acceptable if you don't have access to the other items but want to start immediately, as, generally speaking they are easy to find. An enema bulb is a fancy fleet enema without being prefilled with saline solution. Skip step one if that's what you are using:

1. Dump out the contents of the enema (basically saline water). Since a Fleet enema only holds 137ml, two drops of CD are the maximum number you can add. To use more water, empty additional Fleet enemas.
2. Activate your two drops of CD and dilute them with filtered water (not tap water).
3. Funnel the activated CD solution back into the Fleet bottle. Insert the applicator and squeeze the bottle until all or most all of the contents are empty.

Comments & Suggestions from Parents

Most people are surprised at how well this goes. Some parents have asked if they can do the Protocol without the enemas, and I say yes, but why? The results are just that much better with them. Enemas have, in fact, been around for thousands of years, and colon hydrotherapy is becoming a growing trend in autism recovery.

> I was really intimidated at first by the enemas.
> The syringe/catheter system, however, is brilliant. I lube the tip up with coconut oil, attach a filled syringe, and ask my son to put it in. He's 7 and very low verbal. He started inserting the tip by himself maybe on our 3rd or 4th enema with this system and I would depress the syringe. A few months later of every other day (ish) enemas, and he's doing the syringe depressing part himself. All I do is fill and swap syringes out. We are now to 11 100ml syringes, with 3-5 releases per enema (4 months into enemas). He requests enemas frequently on days I don't plan to give him one by getting the box of supplies and handing it to me.

Here are one mom's enema guidelines:

> *Be low-key about it, your attitude sets the tone. Have all your enema gear ready to go and in the bathroom for when the moment strikes. Put a nice, fluffy, washable blanket on the floor; bring in iPad, iPhone, whatever occupies your child for a few minutes. I've been doing my son's enemas after an afternoon poop if I get one because I figure it is easier as he is already in the bathroom, the bowel should be emptier, leaving more room for CD. Also the one time I did an enema close to bedtime, I swear he was up for an extra hour, however, another child might have the opposite reaction. I kneel on the floor, lay a beach towel over my lap, lay my son across me so he's on the ipad, pull his knees up a bit and then in it goes. You can put coconut oil on the anus or on the tip of enema. I ask my son to count to try to hold it in. (On the 4th enema, we only got to 40 seconds.) Then plop them on the toilet, and there you go.*
>
> *If possible, start the enema while lying on the left side, then rolling to the right. In this way, more water is able to reach the colon. (However, this may not be possible for all children.)*

CD Baths

Taking a bath with CD is another good way of detoxing your child. The CD gets into the pores and helps with skin rashes common with autism detoxification. This additional process can be added when you start with the enemas after reaching full oral dose, but avoid giving the child a CD bath right after an enema.

CD baths are done on non-enema days, unless you are doing daily enemas, in which case they are done on opposite ends of the day.

Enemas Every Other Day

Sun	Mon	Tue	Wed	Thur	Fri	Sat
Enema	Bath	Enema	Bath	Enema	Bath	Enema

Enemas Every Day
(Enemas can also be in the morning and baths at night.)

Sun	Mon	Tue	Wed	Thur	Fri	Sat
Morning Bath	**Morning** Bath	**Morning** Bath	**Morning** Bath	**Morning** Bath	**Morning** Bath	**Morning** Bath
Evening Enema	**Evening** Enema	**Evening** Enema	**Evening** Enema	**Evening** Enema	**Evening** Enema	**Evening** Enema

The CD baths can go from as low as 10 drops to as high as 80 to 100 drops; it just depends on the person, and the size of your tub. With younger/smaller children we start with ten drops. Since we're doing baths every other day, start with ten drops on Monday, 11 drops on Wednesday, and 12 drops on Friday. Just keep going up until you get to 20 drops. The bigger and older the person is, the more drops they can tolerate.

Fill the tub to a level that maximizes skin contact.

If they start getting nauseous/listless or have loose stool, you should back down and find out what their signature dose is on the baths. Some people who have suffered from chronic constipation have done CD baths with up to 100 drops and had a BM during the bath or immediately after, without the use of drugs, laxatives, or suppositories.

CD Steam Bath

Inhalation of CD for children is generally impossible, because you just can't tell a child, especially a child with autism, "just inhale a little bit." That's not going to work. So we do the following:

In a bathtub, with the bathroom door and windows closed, plug the drain, put 20 activated drops on the floor of the tub, and turn on a hot (to tolerance) shower water. The steam will bring the CD into the air, making a really nice steamy bathroom.

When the tub is full, turn it off and place the child to soak in the water for 20 minutes. This way, the child is absorbing it through their skin as well as inhaling it through the air. This technique creates a very dispersed, low concentration dosage. I've seen a lot of children that were going back and forth to the hospital, getting steroid injections and lots of antibiotics due to bronchitis, coughs, asthma, etc., and it was rounds and rounds. This is how this particular method of steam baths came to be a very gentle way of inhaling the CD. It works beautifully; particularly if the child has a really bad cold and is congested, do the oral, enemas, and steam baths. Steam baths are great for asthma, flu, first signs of a cold, runny nose, coughing, etc., putting an end to running to the hospital or pharmacy.

CD Foot Baths

Sometimes it is impossible to give a CD bath, and in some places not all homes have bathtubs. To remedy this you can use a CD foot bath. Fill a small plastic container with hot (to tolerance) water and add your activated drops of CD.

If you have a smaller child, start with ten activated drops, and work up from there. If you have a larger child or an adult, start with 20 drops and go up. Soak feet for 20 minutes every other day, alternating with enemas, or opposite ends of the day if using daily enemas.

72/2 Protocol (Weekend Protocol)

When we give a child eight oral doses of CD throughout the day we are killing more and more pathogens. But at night when we sleep, the pathogens don't. So in order to get ahead of them, we can dose throughout the night. Imagine the progress we can make if we don't give them the chance to gain strength overnight!

Hence, the 72/2 Protocol, which involves giving one dose of CD every two hours for 72 hours straight—including the middle of the night. Here are some additional thoughts and guidelines:

- Start protocol when you pick up your child from school on Friday.
- Give them the last 72/2 dose when you drop them off at school on Monday.
- Avoid giving CD enemas or CD baths during this protocol, UNLESS they are dumping parasites, in which case you might need to reduce the amount of CD on both the oral and enema dose to insure the child doesn't have a Herxheimer reaction.
- Why not give hourly doses during the day at the usual dose? It's simply too much if you are doing it all night as well.
- Watch for improvements on Tuesday or Wednesday each week.
- Ideally, get your spouse or significant other to help with every other nightly dose.

The key is to observe your child and their reactions to the various treatments while keeping the pressure on eliminating the pathogens. You can do this every weekend or not at all. Some families are so happy with the results that they dose throughout the night for months on end.

Ocean Water

This protocol concentrates on excesses and not deficiencies. While we are using CD to reduce pathogen populations and heavy metal toxicity in the body, we need to remineralize the body. Remineralization is extremely important as vital nutrients may be lost along with toxins while detoxing. Children with autism are often mineral deficient to begin with due to these same pathogens and parasites, as well as gut dysbiosis which can prevent adequate nutrient absorption. We need to fortify our cells and tissues so that they can overcome the deficiencies and eventually return to a state of vitality.

Ocean water has helped us achieve mineralization and has led to increased brain connectivity; not to mention the parasites hate it. The importance of ocean water to this protocol cannot be understated. On page 388 of the Testimonials appendix you will see before and after lab tests proving that this protocol, when used correctly, can correct imbalances. More than one family has reported healthy nails, hair and skin as well as appropriate growth and weight gain.

We've all been taught that drinking salt water from the ocean is dangerous. Let me make it clear: We are not advocating drinking glasses full of straight ocean water nor water mixed with table salt. That would most likely make anyone sick, since it would be too much salt for the body. Rather, we are talking about diluted ocean water. Also, there is no comparison between real ocean water and just taking plain salt and mixing it into a glass of water!

The ocean is a complete source of minerals that are bioavailable to the human body. I prefer ocean water to any synthetic minerals, since the body easily recognizes and assimilates it.

Ocean water contains trace amounts of over 74 bioavailable elements including:

- sodium
- iron
- selenium
- sulfur
- calcium
- molybdenum
- cobalt
- boron
- silver
- fluorine
- potassium
- chromium
- zinc
- silicon
- chlorine
- vanadium
- nickel
- lithium
- iodine
- magnesium
- manganese
- copper
- phosphorus
- bromine
- gold
- etc.

We can use water directly from the ocean (diluted to 25%) and drink this to remineralize ourselves. However, the water must be from clean locations where schools of small fish swim and/or near ocean vortices. These locations provide safe and clean ocean water from which we can harvest. Most of us do not have access to oceans. Fortunately, clean ocean water is available commercially from multiple vendors online.

For further information about how and why ocean water is specifically beneficial to all of us, please research Rene Quinton's work on how he healed the sick using ocean water.

Using Ocean Water with CD

You can add ocean water in before starting CD, or after reaching full CD dose. We use a one part ocean water plus three parts drinking water mixed together (1:3). If we take too much ocean water, we can get loose stool. Start low and slow, 5ml ocean water mixed with 15ml drinking water and go up to tolerance.

Use the following chart as a rough guide:

Ocean Water Dosing

Size	When	Pure Ocean Water	Purified Water	Total Volume
Child	Start at	5ml	15ml	20ml
	Go to	30ml	90ml	120ml
Adolescent	Start at	10ml	30ml	40ml
	Go to	50ml	200ml	250ml
Teen / Adult	Start at	15ml	45ml	60ml
	Go to	75ml	225ml	300ml

Ocean water should be given three times a day, with the first dose being given in the morning, after the first dose of CD. You want to separate the dose of ocean water from doses of CD by at least five minutes. Larger children can take anywhere from between 70 and 120ml, depending on the child. Some adults take up to a cup (plus dilution water) per day. It is important to watch for loose stool/diarrhea when you are increasing the dose of ocean water, as that will be an indicator that the person is at their upper limit. The ratio of ocean water to drinking water is 1:3 (one part ocean water to three parts drinking water). It is important to stay under diarrhea. So we need to add slowly, and to tolerance. If someone suffers constipation, with the right amount of ocean water that will quickly change. So low and slow as we increase the ocean water.

Some parents have reported that their children will not drink ocean water, even when diluted, due to its salty taste. You can try adding a little flavored SweetLeaf® Stevia, or if necessary adding it into a beverage or soup. As some ocean water purists say, heating it kills some qualities of the ocean water. Nevertheless, your child will still be receiving all of the minerals.

Biofilm

One of the pioneers in autism recovery, Dr. Anju Usman gave the autism community one of the critical pieces of the puzzle—biofilm. The following information is used with permission from her presentation, *Gut Biology and Treatment*.

What is Biofilm?

A biofilm is a collection of microbial communities enclosed by a matrix of extracellular polymeric substance (EPS) and separated by a network of open-water channels. Their architecture is an optimal environment for cell-cell interactions, including the intercellular exchange of genetic material, communication signals, and metabolites, which enables diffusion of necessary nutrients to the biofilm community.

The matrix is composed of a negatively charged polysaccharide substance, held together with positively charged metal ions (calcium, magnesium, and iron). The matrix, in which microbes in a biofilm are embedded, protects them from UV exposure, metal toxicity, acid exposure, dehydration, salinity, phagocytosis, antibiotics, antimicrobial agents, and the immune system.

Ok, so in layman's terms, these are microbial (bacteria, Candida, virus, etc.) communities that have a protective layer around them, which makes them 100 to 1000x harder to kill than if they did not live in these protected communities.[13]

Why are they so hard to get rid of?

- Microbes impart genetic material to one another to maintain resistance.
- Colonies communicate with one another through the use of quorum sensing molecules.
- Colonies fail to express OMP (outer membrane proteins).[14]

These colonies are fairly amazing in their development. They communicate with each other, share genetic material to prolong their own survival, and don't express outer membrane protections. This last statement is very important to understand. This is the main reason that the immune system doesn't attack these colonies, they don't express themselves as a threat, and therefore the immune system, which in many children with autism is already compromised, is not able to detect or eliminate them.

In addition to this we must be careful with what supplements we choose to give to our children. Some supplements and nutrients may inadvertently feed the biofilm. "When trying to kill bugs, if you take calcium, you may not be making headway," Usman said. "Calcium, iron, and magnesium block our efforts to dismantle the biofilm."

Why CD Works on Biofilm

The microorganism communities attach to surfaces (in many cases the gut), and are held together by polysaccharides (sugars/carbohydrates) and by iron, calcium, and magnesium. EDTA is a known chelator, and can therefore break the bonds (iron, calcium and magnesium) that hold the matrix together. Neither an antibiotic alone nor a chelator alone was effective against a staph (bacterial) biofilm. However, when combined, EDTA broke the bonds exposing the bacteria to the antibacterial agent, allowing the body to expel the biomass.

Chlorine dioxide oxidizes inorganic compounds by effectively removing their charges. Here is how it works: Fe_2+, Ca_2+, and Mg_2+ are positively charged ions that hold the negatively charged matrix together. When these substances are oxidized by CD, the bacteria are then exposed to the CD which can oxidize (kill) the bacteria, allowing the body to finally expel the biofilm mass.

Many parents using chlorine dioxide orally, and with enemas, have found biofilms in their children's stool. Biofilm can look thick, mucousy, cloudy, whitish, greyish, sometimes like pantyhose, etc.

There is some more interesting information available through the AutismPedia about Dr. Usman's work with biofilm.[14]

We had our appointment yesterday and to all our surprise when he tested her for Lyme it was negative. Just 3 months ago and before starting MMS she was testing very positive for it. He was pleased with the outcome. Thank you so much for your guidance and we owe this to the miracle of MMS and your dedication to helping me with my questions. I really, really appreciate your help in bringing us one step closer to the cure.

Also, this article "Lyme-Induced Autism Conference Focuses on Biofilm and Toxicity"[15] by Mary Budinger discusses the effects of pathogens present in the biofilm for Lyme disease and autism. The following is a notable quote from the aforementioned article by Dr. Stephen Fry of Fry Labs, which explores the possible pathogenic causes of chronic disease.

> *I could be barking up the wrong tree, but maybe not. Remember that we used to think stomach ulcers were caused by too much acid production. Then Barry Marshall and Dr. Robin Warren turned medical dogma on its head by proving that a bacterium was the cause. The pair identified*

> *the bacterium H. pylori and proved how it causes inflammation, then ulcers. Maybe in 10 years we will be smart enough to know that the 'auto' in 'autoimmune' actually means pathogen and the whole concept of autoimmunity will change. Chronic inflammation is chronic infection. In autoimmune disease, my model is that there is a chronic infection that cannot be eliminated, thus the immune system is always switched on. The self antibodies are due to apoptosis and death of host cells with host immune response.*

I couldn't agree more, and am so grateful that we have CD to combat pathogens, and restore the body to a state of health.

Mitochondrial Dysfunction

Children on the spectrum are also more likely to have mitochondrial dysfunction than neurotypical children of the same age group. The following is an excerpt from *Science Daily*.[16]

> Children with autism are far more likely to have deficits in their ability to produce cellular energy than are typically developing children, a new study by researchers at UC Davis has found. The study, published in the Journal of the American Medical Association (JAMA), found that cumulative damage and oxidative stress in mitochondria, the cell's energy producer, could influence both the onset and severity of autism, suggesting a strong link between autism and mitochondrial defects.
>
> "Children with mitochondrial diseases may present exercise intolerance, seizures and cognitive decline, among other conditions. Some will manifest disease symptoms and some will appear as sporadic cases," said Cecilia Giulivi, the study's lead author and professor in the Department of Molecular Biosciences in the School of Veterinary Medicine at UC Davis. "Many of these characteristics are shared by children with autism."
>
> Dysfunction in mitochondria already is associated with a number of other neurological conditions, including Parkinson's disease, Alzheimer's disease, schizophrenia and bipolar disorder.
>
> Mitochondria often respond to oxidative stress by making extra copies of their own DNA. The strategy helps ensure that some normal genes are present even if others have been damaged by oxidation. The researchers found higher mtDNA copy numbers in the lymphocytes of half of the children with autism. These children carried equally high numbers of mtDNA sets in their granulocytes, another type of immune

> cell, demonstrating that these effects were not limited to a specific cell type. Two of the five children also had deletions in their mtDNA genes, whereas none of the control children showed deletions.[10]

In our MMSAutism webinar, Dr. Andreas Kalcker mentions that CD changes the mitochondrial membrane electrical potential.[17] This leads to increased power in every cell in the body, including immune system cells. The mitochondria are the electric generator or "powerhouse" of each cell. When we turn off this electrical generator, cancer can develop. If the cell cannot produce electrical energy through oxidation it will use fermentation.

In practice, what I have seen is that children who suffered from lack of energy, listlessness, and other symptoms that can be related to mitochondrial dysfunction have perked up with the use of CD. They are able to do exercise again, walk without having to be carried, and can cognitively sustain activities without tiring. You will see a few testimonials mentioning "mito issues" at the end of the book. It is wonderful to think that this is another means by which CD can help people on the spectrum heal.

ATEC Statistics

Through some of our Facebook groups we have been collecting ATEC data for the past year. Just to give you an idea of what others are experiencing in terms of test results, here are some statistics:

- 246 parents have given us 2 ATEC scores. Of those, the average drop between ATEC1 and ATEC2 is 15 points. We take ATEC data quarterly, so in general that drop is in three months or less. The more impacted a child is the bigger their initial drop – kids starting with an ATEC over 100 averaged a 26 point drop between ATEC1 and ATEC2.
- Of those 246 ATEC reports, we had 32 children (13%) gain in points in the first quarter. The average gain among those 32 was 7 points. All 32 reported ATEC scores three months later, and 29 of them dropped points between their ATEC2 and ATEC3 scores. Further, 13 of those 29 dropped enough points to put them at or below their ATEC1 starting number. Which means at the 6 month mark, only 19 out of 246 (or 7%) had gained points. 93% of children who do this protocol have improved their ATEC score at the six month mark.
- We have now had 163 parents provide four consecutive ATEC scores (encompassing one year on the protocol). Of those 163, the average ATEC drop between ATEC1 and ATEC4 was 24 points.

- Of those 163, a total of 9 or 6% reported an increase in ATEC score between ATEC1 and ATEC4. The average increases was 6 points. 94% of children who do this protocol have improved their ATEC score at the one year mark.

- In a Facebook group collecting ATEC data, we have 25 recoveries as defined by the ATEC dropping to 10 or under in a year. So we have a rate of 15% recovery as defined as ATEC score under 10 in the parents who have taken this protocol through one year.

Thus showing that the longer they are on the protocol the more points they lose.

Although we all would like to see "high level" statistical research done, this is no doubt a major step in the right direction, and demonstrates that the protocol does work. But, research takes money.

DOs & DON'Ts of CD for Autism Recovery

DO review all current supplementation that your child is taking and remove all supplements which contain any antioxidants including vitamin C, vitamin E, vitamin A, vitamin K, ALA, Coenzyme Q_{10}, and colloidal silver as these will neutralize the CD, rendering it ineffective.

DO consider removing supplements that include iron and vitamin B_{12}. These are principal foods for the parasites, which explains why many children are low in them. If you are seeing a chronic iron or vitamin B_{12} deficiency, revisit the parasite chapter. These levels should normalize when parasites are no longer consuming nutrients.

DO use the Baby Bottle Method in an eight-ounce glass bottle with a completely sealing lid and start at one drop, regardless of weight. Increase by one drop/day, until you reach the optimal dosage.

DO dose frequently throughout the day. The less time we leave the pathogens to proliferate the better. Do a minimum of eight doses a day, but try to get in more doses if possible. Go for 16 in cases of PANS/PANDAS and acute situations like colds and flu.

DO activate CD for the proper amount of time for the acid you use. HCl and citric acid both activate sodium chlorite in 60 seconds.

DO supplement minerals using ocean water.

DO completely avoid the following: citrus fruits, corn syrup, fruit juices, green tea, pineapple, vitamins C, E, A, and K. Only give coconut milk and coconut water a minimum of one hour after the last CD dose at night so they will have no chance of affecting the CD. Coconut milk and coconut water are very alkaline.

DON'T give fruit juice of any form—not fresh, organic, homemade or store bought.

DON'T give highly antioxidant foods including chocolate (cacao/cocoa), coffee, green tea, kombucha, citrus fruits, pineapple, mango, or kiwi. Berries (if you have to give them at all) can only be given at night, one hour after the last CD dose.

DON'T mix anything but water with your CD drops.

DON'T give CD with food.

DON'T use a vitamin C shower water filter—yes, there is such a thing to neutralize chlorine in tap water. Unfortunately, it also neutralizes chlorine dioxide.

CD Troubleshooting Checklist

Before you decide that CD isn't working for your child please take a look at following list of common errors. If it looks like you are doing everything right, but you aren't seeing gains, please, before giving up, use one of the support options shown in Appendix 17, page 521. You are NOT alone! We have a network of parents who are ready and willing to help you.

☐ Are you preparing your CD in a clean, dry shot glass, or other type glass container that insures both chemicals are mixing?

☐ Is the CD mixture turning yellowish brown? There should be a chlorine-like smell. If there isn't, something is not right. Perhaps you have bad or incorrect chemistry? Maybe the supplier did something wrong in preparing the solutions.

☐ Are you using the correct ratio of sodium chlorite to acidic activator drops? See chart on page 94 for the correct mixture ratio depending on the type and concentration you are using.

☐ Are you allowing the correct activation time before dilution with water after mixing the sodium chlorite and acidic activator? See chart on page 94 for the minimum activation time you must wait for the chemical reaction to occur.

☐ If you question the potency of your CD, then you may wish to test its strength. See Appendix 6, page 467 for testing information.

☐ Is your CD or activator cloudy or not changing color after they are mixed? If so, there is something wrong with one or both chemicals. Check with your supplier.

☐ Did you leave your baby bottle solution in direct sunlight? A brief exposure to light is no big deal, but if you left it in a hot car for an hour, it may have lost potency.

☐ Are you keeping the same baby bottle solution for too many hours? It should be used on the day it was prepared.

☐ Are you giving antioxidant supplements or a multivitamin containing antioxidants? Vitamin C, vitamin E, etc., should not be given. Cod liver oil and fish oil supplements in general have lots of antioxidants to keep them from spoiling. Make sure those are out as well. Mixing any of these will cancel out the effect of the CD—and the vitamin. You CAN give them a few hours apart.

☐ Have you checked all your labels of supplements and foods to make sure they don't contain antioxidants?

☐ Have you removed all juices and all citrus fruits from the diet (including oranges, pineapple, mango, kiwi)? Berries (if you have to give them at all) can only be given at night, no less than one hour after the last CD dose.

☐ Are you dosing CD 30 minutes apart from food, as a minimum (one hour is optimal, but not always possible)? Grazers do 15 minutes and we dose CD 16 times a day.

☐ Are you mixing anything with your CD, such as baking soda, juice, etc.? The only thing you can add into your dose of CD is more water!

☐ Are you using alkaline water to prepare your CD? Alkaline water kills CD. Some expensive water filtration systems are designed to produce alkaline water, which we must totally avoid!

☐ Are you breaking The Diet? Could there be dietary infractions at school or with relatives while you're not there?

☐ Are you dosing so high that you are detoxing your child too fast and getting a Herxheimer reaction? Some people need to increase every three days—NOT every day.

☐ Are your doses too low to accomplish anything?

☐ Are you giving too few doses of CD? We need a minimum of eight doses per day or more.

☐ Have you started the Kalcker Parasite Protocol? If you have done three Parasite Protocols (PP) it is time to start looking into supplements, chelation, HBOT, GcMAF, etc.

Frequently Asked Questions (FAQs)

Can I use less than one ounce (30ml) of water to administer a dose of CD?

One ounce of water, or more, is best. You can always add more water, but at least an ounce is the minimum per drop.

We have been on the Protocol and are at five activated drops of CD in the baby bottle. My son's "usual" autistic symptoms got exacerbated to the maximum. He has horrible constipation, stimming, scripting, and he seems to be in a bad shape. Our target dose is 15 drops. What should I do?

The general protocol is to wait with CD enemas and CD baths until one is at their full target oral dose. His negative symptoms are probably from reabsorption of toxins from his unpassed stool. You need to resolve his constipation to stop this vicious cycle. You can administer CD enemas and CD baths on consecutive days. This should help the constipation, and flush the toxins out of his system. Then, once he is stable you can resume slowly titrating up to full dose.

What should I do if my child can handle "double doses" (two ounces of the baby bottle) all day with no herxing?

If your child is taking "double doses" all day, then it is NOT a double dose—it's his actual dose! Depending on his gains and what else you are folding into your protocol you can consider going up from there.

Is 72/2 only if you have done the parasite protocol? We are full oral dose but won't do PP until next full moon.

You can do 72/2 protocol before the Parasite Protocol, but you must have reached full dose of CD.

We are really struggling to get in enough doses of CD with school. We do well to get in 8; we really need more (way, way more). Night dosing is not an option (he won't go back to sleep; our mornings start early enough around here as it is), nor is coming to school during the day to dose him (we both work). Advice?

Some parents are dosing every 45 minutes to get in their doses when they are with their children. However, we need to be killing pathogens around the clock. It is not possible to be effective at killing pathogens when we are giving them 16 hours without a dose of CD. You need to get 1 to 2 doses in before school starts. Then get the other 6 (ideally more) in after school is over and in the evening.

My child hates the flavor of CD. Can I give Stevia with CD to improve the taste?

We recently discovered that some brands of Stevia are compatible with CD, CDS and CDH (i.e., SweetLeaf® & KAL® brand)—it doesn't reduce the potency. However, past tests were not so positive. Some added ingredients may be why this was so in our initial tests. Try to use only pure Stevia or check with the Facebook groups on what others are using successfully.

SweetLeaf
Sweet Drops
Sweetener
STEVIACLEAR
KAL
PURE
Stevia
ORGANIC
extract

Preliminary tests show that some brands of Stevia can be used with CD to improve taste without losing potency. The two brands we have tested in-house are SweetLeaf® and KAL®. Some of our moms report having successfully used flavored versions of SweetLeaf®.

When you say do two back-to-back baby bottles of 11 drops... you mean do a total of 22 drops in 16 hours? Correct? I'm going slowly up on oral CD because of behaviors. If I go from 11 to 22 drops overnight, won't he have strong detox and behaviors?

This is spread out over 16 hours or more, such that every 60 minutes it is the same amount. It is different than putting 22 drops in 8 fl. oz. Behaviors associated with bacteria (PANDAS/PANS) are often reduced with the tolerated dose spread over longer periods of time. Remember that CD is only active in the body for about one hour.

My child seems very uncomfortable. What should I do?

If you see a detox like this, stop dosing for the day. The next day, return to the last dose when the child was stable. Pathogens are dying and toxins are being released into the bloodstream on their way out of the body. This causes discomfort in some people.

How do you know if I should go past full oral dose of CD?

A Herxheimer reaction will be your biggest indicator if you have gone too far. If you have reached full oral dose and it seems like your child has plateaued, please schedule a consult with Kerri to discuss your options. Note: The full oral dose chart (page 101) is only a general guide and not a one size fits all.

Is an hour enough time to separate CD neutralizers from the CD dose?

CD is only active in the body for about one hour, however, supplements like vitamins A, C, E, K and ALA, CoQ10, and GSH (glutathione) need to be avoided. Also, orange juice, pineapple juice, and other high antioxidant fruits are a problem even after one hour. So, the best thing to do is eliminate everything that kills the CD. If not, it is pretty much like not using CD. I have seen it too many times and it's not worth it. Antioxidants destroy CD. There is no period of time that is safe to use antioxidants with CD.

I know you have to start the parasite protocol on a certain day of the month, but how about the first dose of CD?

You can start giving CD any day of the month. That said, it might be easier to start on a weekend, or a day when you will be with your child all day to make sure to get in all of their doses, and keep an eye on them to make sure they tolerate it well, i.e. no Herxheimer reaction.

What is 72/2?

72/2 is when you give a dose every two hours around the clock for 72 hours. We use a baby bottle and a half to get through the night hours. This can be added in once you are comfortable at your full oral dose and have added in enemas and baths. See page 115 for more information.

Do you recommend 72/2 for everyone?

Once a child is at 100% of their target oral CD, enema, and baths then they can do a 72/2 weekend. They are exhausting for the parents, but some kids do wonderfully. A lot of families get a nice boost. But, if you do a couple of weekends and you don't see anything afterwards, then forget it. The benefits are usually seen two to three days after the 72/2 is over.

My son suffers from constipation. Can I start the enemas before I get to his full oral dose?

Anyone who suffers constipation needs enemas right away. These can be combined with oral doses. Still start low and slow, but it is important to get the bowels moving with any detox protocol. When constipation is present, we do not wait to start enemas.

What is the maximum amount of water to use with enemas? How many times do you hold and release?

First of all, we don't encourage holding. Rather we do a fill and release, which is kind of a home colonic. If your child can hold for a few seconds without fuss, then great. The number of cycles depends on the results of each. If all you are getting out is water, you may be finished, unless the person is seriously constipated. The amount of water really depends on the individual's size and the number of cycles administered. My basic philosophy is: *Any enema is a good enema.* So, even if not much comes out, you are still getting some water and CD in.

Is it possible that my son has been getting rid of mucous after his CD enemas? What is up with the mucous? Why is it there in the first place? And, is it a good or a bad thing that it is coming out?

Mucous is always a sign of inflammation and can contain pathogens and parasites. Mucous gives a home to pathogens. The inflammation and mucous can be caused from the allergies and pathogens. It is common to see mucous come out in the beginning of the enemas since our children have so much intestinal dysbiosis.

What is a good enema method for a 17 year old boy? I am having trouble understanding the enema part of the book. The video in the file really helps, but I am having a hard time imagining doing it on my son. How do other people with older children do it? Can it be self-administered eventually? Kerri mentioned a gravity bag, so you don't have to keep changing syringes. Has anyone tried that? Where do you get the bag?

Here is some advice one mom gave to another mom on one of our boards:

My son is 14 and tends toward constipation. I had done suppositories in the past to help him have a bowel movement. So, I told him that the enema was "fanny medicine" to help him have a BM and make him better. I started with just showing it to him after a BM after I wiped him. Then, I just put it next to his anus the next day. Then closer every day. He was still constipated so just inserting the tip would help him poop and then later I was able to push the liquid with him standing next to the sink and it would run all down his leg. After several weeks of this, it was a full moon and he could not poop at all and was lying on his bed in pain. I said let's try it here because it will really help you. I got out a hundred towels and did it right there and he pooped. It was a slow process over two months but now we do them AM and PM always right after the shower. I counted over 20 feet (of parasites) coming out this last PP. And that is just what I could see. I don't dig, just pull out obvious ones that are a foot long. Main advice is keep calm with soothing voice and have tons of patience and keep telling your son how much you appreciate his patience too. I also tell him this will get rid of his autism. He is non-verbal but said with RPM that he is mad he has autism. Also, the soft tip is key. I use syringes because they go fast.

My son is not having daily bowel movements. What should I do?

Do CD enemas every day until the stools normalize to daily. If you absolutely cannot do enemas, use CD baths in high doses (50+ drops depending on the size of the tub). Ocean water in high doses will move the bowels as well. Castor oil also helps.

I am scared of the CD enemas. Will CD harm the intestinal lining?

CD stays active only for about one hour and can gently, but effectively remove biofilm and kill pathogens in the intestine. Our society has shied away from this healing method in the last few decades. However, CD enemas have been proven to be positive turning points in the treatment, time and time again. Lots of kids will actually ask for enemas, as they provide them relief and comfort. If you are nervous, apply an enema to yourself first. You will see it is no big deal, and feel more at ease applying one to your child. Some parents have shown their higher functioning children how to apply their own enemas, so no one has to be there, and to make them feel more comfortable.

What is the difference between colonics and enemas?

Colonics cleanse the entire length of the colon while enemas cleanse the lower part of the colon. Colonics involves multiple infusions of water into the colon, while enemas involve a single infusion of water into the colon (which can be repeated). With colonics, fecal material leaves the body via a tube. Colonics normally involves going to an office and receiving assistance by a trained colonics hydrotherapist, while enemas are free and done in the privacy of your own home.

I have never done an enema before. How do you administer it?

There are several positions that work well, but there will be one that works best for the individual. The enema instructions are on page 109. Some examples would be on all fours or lying on your left side on a towel, or a small child could lay across your lap.

Enema clean up question: What are we boiling? What are we spraying with *Everclear*? What are we throwing away?

After you are finished looking for worms you can throw away any plastic plates/forks you may have used. Pour boiling water on your specimen collector, enema nozzle and catheter. Then spray with *Everclear* grain alcohol. If you are using syringes, you may pour boiling water over them as well.

How far do I need to insert the catheter for an enema?

Not very far at all; one to two inches is plenty. The anal sphincter is about one inch past the anus. As soon as the catheter passes this muscle, it will hold it in place and there is no need to go further.

What is intestinal biofilm supposed to look like?

Biofilm is fluffy, cloudy, and mucousy in appearance. There are no rules on the color. It can look like pantyhose. But, it is usually mucousy and a cloudy whitish/grey color.

How far away should we be doing enema from food? If we miss the morning enema, can I give it right after lunch?

They can go together; there is no need to separate enemas from food. Any enema is a good enema.

I'm having trouble with my syringes (for enemas). One plunges really smoothly, but the other two are almost impossible to move. Do I need to grease them up somehow before each use? How do you clean them?

I use *Everclear* grain alcohol to clean them, but note, it will dry out rubber parts. You can use coconut oil to lube the plunger gasket.

Is it normal to see black specks coming out after using enemas? We saw white specks for a couple of days too.

Black specks can be heavy metals. Oxalates are known to bond to metals and would come out with metals. White dots are generally parasite eggs.

Does ocean water need to be refrigerated after opening?

No it does not, it is fine at room temperature.

If I harvest my own ocean water do I have to sterilize it?

No. Sterilizing is not necessary or desirable in our experienc—we don't want to denature it. Pouring it through a coffee filter to remove any potential particulate matter is fine. Do NOT use charcoal filters (such as a Brita® water filter), because they will alter the characteristics of the natural ocean water. Just make sure to collect the water far from any harbours or large river outlets.

I bought pure sea water in Australia. It sounds like a great mineral drink. Would this be beneficial for killing parasites or just putting goodness back into the body?

Actually, it does both.

How do I avoid feeding the pathogens?

CD kills pathogens. I would use CD to rid the body of pathogens rather than try not to feed them.

Is it possible to use CD, CDS, or CDH during pregnancy? If it is ok, how much can be used?

CD/CDS/CDH are not advisable during pregnancy or nursing.

Can I add CD to breast milk to the bottle to help the baby detox?

CD must be given in water. Mixing breast milk and CD will reduce or potentially cancel out the potency of the CD.

How many drops go in a CD bath?

It depends on the weight of the child (20-100 drops). If he is a little guy I might start at 10 or 15 activated drops and work up over the course of the following couple of weeks. The size of your tub also plays a part. The bigger the tub, the more CD you are going to need. See page 113 for more information.

I've heard that some people mix CD directly in capsules and then swallow them. Can we use that on our children?

NO! Not a good idea! Since many children with autism can't speak, they won't be able to tell you when the capsule gets stuck. CD needs to be diluted in water before taking as instructed in this chapter.

Can I use warm water instead of cold water to store CD and how long will it remain potent mixed in the baby bottle?

CD is a gas dissolved in water and therefore has to be preserved in a sealed container. If not, it will lose its potency, similar to how soda goes flat if the cap to the bottle isn't screwed on tight. Colder is better because the CD will gas out quicker at higher temperatures. If you have doubts about CD potency, get the Lamotte ClO_2 test strips to be sure. See Appendix 6, page 467 for more information.

Are purified, distilled and filtered water all the same thing?

They are different. However, we can use purified, distilled or filtered water. But, alkaline water should never be used!

Must the activated CD water mix be stored in a glass bottle, or is a no-BPA plastic bottle with an airtight cap ok?

I prefer glass to any plastic. Over time, chlorine dioxide can degrade plastic, and that means you or your child will be consuming it. Yes, the bottle needs to remain sealed. Plastic is acceptable for lids only! Metal lids should never be used because they quickly rust, even if covered with a plastic coating.

Is it a problem to transfer one ounce of CD to a steel bottle briefly before drinking it?

That is fine for a transfer, but I wouldn´t store CD in a steel bottle.

Is the GF/CF/SF "clean" diet essential to the CD protocol? My son doesn't have bad reactions and I just do not see how we could implement it.

A healthy, clean diet is very important to the healing process. Unfortunately, our "standard diet" is filled with preservatives, colorings, and other potentially harmful chemicals. When the goal is to heal, the diet is necessary. Dairy especially can cause inflammation and mucous production. Mucous provides protection to parasites.

What is a good CD-friendly multivitamin?

All multivitamins have antioxidants. What we need to supplement during detoxification are the minerals. Ocean water has 90 bio-available minerals, and it is what I prefer for this protocol.

Are there other specific foods, veggies or fruits that need to be avoided in addition to juices?

When you are following the CD Protocol, all antioxidants need to be avoided. Chocolate is a strong antioxidant. As far as vitamins are concerned; A, E, K, and ALA are off limits. High antioxidant supplements such as curcumin have also been a problem for many people. In the case of vitamin D, it is not an antioxidant, but it has caused aggression in previously calm children. Therefore, as a supplement, it must be used with care, if at all.

Vegetables and legumes which are high in antioxidants have not shown to be very detrimental to the effects of CD. However, citrus fruits, pineapple, mangoes must be avoided. Berries are best given at night, one hour after the last dose of CD, if you must give berries. Juice, any sort of fruit juice, is completely prohibited. Not only can the antioxidants involved kill CD, but the high sugar content (albeit natural) shuts down the immune system, which is already impaired in children with autism.

Can our kids drink anything other than water for the duration of the CD protocol or just not drink juices down with the CD?

Some people are not used to drinking water and find it difficult. Some people drink nut milks or rice milk. On occasion, some folks put a piece of fruit in the blender with an apple and blend it, strain it, water it down and have that sometimes. Juicing is taking 5 apples to make a glass of juice and that is a lot of sugar (albeit natural) as well as a lot of antioxidants. Antioxidants kill CD and sugar slows the immune system. You want to totally avoid alkaline water.

Can I use rice milk to give the CD instead of water?

Water is what we need to use to give the CD. If you are giving CD mixed with another beverage you are reducing its potential. By combining CD with liquids other than water we will not get the desired results. Rice milk might be okay, but you would need to test how it affects the potency of the CD. You can get the *LaMotte High Range Chlorine Dioxide Test Strips* (#3002) at www.amazon.com. This way you can see for yourself how mixing CD with any beverage affects its potential. However, I do not recommend putting CD in anything but distilled/purified water.

Can I give my child fresh pressed green juices (celery, kale, cucumber, apple)?

Anything that is very nutritious to the human body is nutritious for the pathogens/parasites. So, a healthy green drink is very healthy for the pathogens. If you are giving "green" juices make sure they are not loaded with sugar from apples, carrots, etc. In my opinion, it is preferable to avoid these for the first few months of the Parasite Protocol (PP), at least.

Can I use coconut and almond milk with CD?

You can drink coconut and almond milk 60 minutes after your last dose of CD, but I would not combine anything other than purified water with the dose of CD.

Can my child drink coconut water if he is taking CD?

If you want to give your child coconut water, give it one hour after your child's last dose of CD, at night, before bed.

Does lemon interfere with CD, even after it is cooked?

Yes, in some cases. A family of a child who suffered from SIB noticed a direct correlation with lemon use in their meals with the return of his SIB. I would avoid it at all costs.

My son barely drinks anything, especially all at once. This is why we have been unsuccessful with detox programs. Is CD different?

You should make sure that your child is well hydrated, but for the CD alone one ounce (30ml) of water is sufficient. I find that it is easier to drink CD cold rather than at room temperature. Also, if the smell is a problem, put the dose in a plastic syringe and squirt it right into your child's mouth. This prevents them from having to smell it before they swallow it. Enemas are a great tool for hydration.

My child's appetite has changed since starting CD. Is this normal?

Yes. On average larger kids tend to slim down, while thinner kids tend to gain some weight.

Can I give green tea or kombucha while taking CD?

That is out of the question in my humble opinion. Green tea is high in antioxidants and caffeine. Kombucha is similar. I don`t know exactly how long the antioxidants in either one are active in the body, so with autism, if we are looking for healing, avoid them both.

Can I use original Nutriiveda™ with CD?

Nutriiveda™ is derived from whey, a dairy product. In humans, all dairy causes inflammation and mucous, which provides a protective refuge for pathogens. That makes this product contraindicated for this protocol, plus it contains antioxidants.

Can I use Chia seeds with CD?

Yes. But do not mix them directly into the CD.

How do I dose my child when he/she is at school from 8:30 am to 3 pm? I work full time so "dropping in to dose" is not an option.

Give a dose upon waking, one at the door of the school at 8:30 and the third dose of the day at 3pm at the door of the school. The last dose of the day is at bedtime, there should be at least four hours between school pickup and bedtime to get in the other four doses. Once you get the hang of it, it is not that hard, it just takes a little getting used to and some prior planning.

What should I do if my son goes away for 5 days with school? Do you think it is possible to prepare a bottle of CDS for 5 days and he takes only one portion each evening?

It's not ideal, but better than nothing.

My son is refusing the CD. What do I do?

Have you tried adding more water with the dose, using cold water, using a syringe or a straw? Here is some advice from a Mom whose daughter refused oral CD:

We just went through a REALLY rough patch with my daughter when she absolutely refused to take oral CD, and I know it gets super stressful for both you and your child if you try to force it… I had to respect her as clearly she was telling me in her own way, "Mom, this does not make me feel good!" for whatever reason and after maybe a week we were back on track. During that time I was able to give her a few doses (dropped from 14 drops of CD to 5) here and there, during really good moments. For example, in the bath or swing that she loves and I think that helped her associate CD with feeling better again. You WILL be back on track in no time!!!!! Good luck! Keep calm and dose on!

I am on 30 activated drops/day with my teenage boy. The taste is simply horrible. He refuses to drink his doses. What should I do?

You can dilute your baby bottle as far as a liter bottle if necessary; it will not impact its effectiveness. Obviously, one dose will be more like a half a cup or a full cup in this case. However, be mindful of your conversion rate so that you get your dosage right. Cold water helps with the taste. You can add approved Stevia brands to cover the taste of CD. See page 126.

If the country I live in doesn't have glass baby bottles, can I use a plastic bottle?

No, do not use plastic for storing CD. If the cap is plastic, we can live with that. In almost any country you can find a glass water bottle (with a plastic cap). You can use a shot glass to measure out eight ounces, and even mark them on the outside of the bottle with permanent marker. Many parents in Venezuela do this, as neither glass baby bottles nor LifeFactory® bottles are available in their country. NEVER use a Stainless steel (or other metal) bottle or container!

Do you think this Protocol could work for a child who already has a low ATEC score (18)?

Yes. I started with a 12-year-old, who last February 2012 had an ATEC of 18. He did not budge from there after years of biomedical interventions. He could not stop with fears, phobias, anxiety (all parasitic in nature) and we started him on CD in February 2012 and the Parasite Protocol in June 2012. His ATEC is now a one!

My son's ATEC is not going down as I had hoped it would by now. It has hovered around 65 the past few ATEC's. I think it is probably due to the fact that if I am not home my husband often forgets to give him his doses or his PP meds. Does anyone have any suggestions on how to stress the importance of this protocol to him?

The following responses are from moms on our public Facebook group:

You will never know what following the protocol will do for your son until you follow the protocol. When I first started the protocol a little over a year ago, I remember I had been using a special juice that so many people were raving about and I was paying $140 monthly for this miracle juice and I really wanted to keep in in my son's protocol. I asked my husband to give it to my son at 4 am when he got up, and that was 4 hours before he would get any CD doses... how could that effect our results? The juice was given 4 hours before starting CD for the day, we were following the protocol exactly otherwise, and even doing the enemas. Well, nothing happened, nothing, until about 3 weeks later when I said, well let's stop the juice and see... BAM! Immediately my son's ATEC results dropped and he blossomed. I mean not only did his constant pacing stop, but his eyes lit up, he not only started talking more, more, more but even his laugh changed to a typical sounding teenager's laugh. It was incredible. His ATEC dropped instantly and kept dropping, 68, 25, 13, 7, 5, 3. It's a 6 today but this a super bad new moon for him. You will never know until you follow the protocol.

We have also seen our son stop pacing, stop suffering minute to minute, in the beginning we made our mistakes, missed a dose, didn't get this or that done, we now stick to it, the cumulative effect of staying on protocol means the difference in drops in ATEC scores and staying the same or going up, it did for us anyway. I could feel the slipping, we are still working on getting this right, it is an evolving protocol that requires the entire family's buy in. I have kept my son home from school to make sure he is dosed properly but that's just us. In November he is supposed to begin going to school full days, if they do not dose him correctly, I will hire a homeschool teacher, that's how important this is. You will not get your child well, Kerri told us 16 doses a day, he got better, she told us to start *enema's early, that made a big difference, we started our first PP before he was at full dose, it's all made a huge difference, email her, keep watching your child to see what's working, this protocol works, it may take time, but what is the alternative????*

My husband is not good about giving meds either. When I'm gone my older girls give it. I text them reminders. Sometimes setting a timer can help. I also set all of the meds that go along in a morning and night dose daily case. Often when I organize it all for them it's much easier for the meds to be given properly and in time.

My child has horrible nasal congestion. Can CD help?

Yes, you can use CD and make nose, eye, and eardrops. Put one activated drop of CD into one ounce of water. Use one drop of the mixture in the nose (eyes or ears) every 15 minutes till symptoms disappear. NOTE: DO NOT USE PURE ACTIVATED CD WITHOUT DILUTING IT!

You can also do a steam bath, which is done by closing off the bathroom (windows and door shut), and putting 20 activated drops on the floor of the tub. Run hot shower water on top of the drops until the tub fills; the air will be full of CD smell. Now, turn off the shower, put the child in the water to soak for 20 minutes, where he will breathe the light CD air.

Is there a seasonal-allergy protocol for using CD? I suffer from extreme itching in my eyes, redness, puffiness, sneezing, sinus congestion etc. Will CD provide relief quickly or does it need to be used over a long period of time?

Being on full CD, ocean water and the parasite protocol the allergies should begin to fade away.

Can I try the CD ear protocol if my child has not started oral doses?

Yes!! Please start the eardrops, if there is an infection, every hour until the symptoms disappear. If you catch an infection at the onset of its symptoms, give one drop of the mix every 15 minutes. We have seen earaches clear up in a couple hours. In any case, work towards a full oral dose, at least one drop eight times a day to combat the infection on all fronts. The eye, ear, and nose protocols are all the same: one activated drop of CD in one ounce of water in a sealed dropper bottle. Apply one drop of this mix every hour until symptoms diminish.

Can CD help arthritis?

Absolutely! CD is amazing for arthritis. Include CD baths in your routine. Use CD in conjunction with DMSO.

Our son has a runny nose and sneezes a lot, but has no fever. He hasn't been sick in well over a year! I'm wondering if it is from doing CD.

This sounds like a very typical die-off reaction. Monitor his progress, make sure he is hydrated, and continue with the Protocol. If his reactions remain manageable, you don't need to back down or change much. This could also be parasites, depending on whether the symptoms are chronic, etc.

Has anyone seen kids who had never been sick start getting sick as they heal?

Yes, it means immune system is waking up and fighting back against the crud they've been carrying around for a long time. It's called a healing crisis and is not uncommon and is actually a good thing.

Does CD, CDS or CDH work on HIV? Also, are there any contraindications of CD and HIV drugs?

We do not focus on HIV or its treatment and therefore we cannot give you a definitive answer. Anecdotally, there are many testimonials that indicate CD having a positive impact on the disease. In addition, we have spoken to doctors in various parts of the world who use CD to treat their HIV/AIDS patients and have positive results from what we understand.

Can I use CD in a Netti pot when doing an ear, eye or nose wash?

No. Do not use a Netti pot with CD. Instead, use the steam bath method. See page 114 for more information.

I´m wondering if CD can be used in a nebulizer? With our daughter's chronic mycobacterium issues, we can never get rid of the cough. I thought maybe it would be helpful to get it into her lungs or is this not safe?

Try a steam bath. Inhaling CD directly has to be done very delicately. This is why the steam bath is recommended, as the particles have time to disperse into a very large area rather than being inhaled directly. Some people are using humidifiers with 35 drops of CD per gallon. See page 114 for more information.

What supplements do I need to avoid while using CD?

Antioxidants (vitamins C, E, A, K, ALA, and CoQ10) need to be avoided because they kill CD. We avoid iron and vitamin B12 because parasites feed on these. So if you do not have something in place to kill pathogens/parasites, like CD, then the majority of your supplementation goes to strengthening the parasites instead of your child.

I really don't understand something basic... Vitamin C in food. I know not to give citrus or mangoes. Also, kiwis are rather high in vitamin C. If my child eats a kiwi or other high vitamin C containing food, how long does this impact CD? When will it wear off?

Kiwis and other fruits high in vitamin C are a tough call. We do our best to avoid the biggest offenders. Unfortunately, I have seen families lose months because they continued to give fruit juice or supplements high in antioxidants. As soon as these were pulled, their children began to improve. A few years ago, a popular omega high in antioxidants became popular for autism. Families that I was helping all over the world noted regression in their children. As soon as we figured out what the common thread was, and pulled that supplement, the improvements began again, but it was a nightmare. To answer the second part of your question, I don't know exactly how long vitamin C from a natural fruit source lasts in the body. A few different websites estimate up to 24 hours.

We have been using gluthatione every day and modified citrus pectin as a binder. Are those compatible with CD?

Gluthatione is a very powerful antioxidant and cannot be used as a supplement with CD. Modified citrus pectin is fine.

I have heard that CD helps with oxalate problems. Is that true?

Rompepiedras (RP) is very effective against oxalate-crystal formation. There are connections between parasites and oxalate problems, which this protocol seems to be helping. Ridding the body of parasites helps with high oxalates.

I know that CD neutralizes heavy metals. What does that actually mean? What is the difference between chelating and neutralizing?

The mechanism of how CD neutralizes and removes heavy metals such as mercury is not clearly known (or understood). However, a number of recovered children have had heavy metal burden tests performed before and after the use of CD, and their heavy metal burden dropped. We also know of an adult in our circles who was extremely toxic with mercury and was able to normalize his levels after all other medical interventions had failed him.

So is CD enough and we don't need to chelate?

For many it is not necessary to chelate while on CD. There are some children though who do IV chelation and CD at the same time. This method seems to have good results; there are big metal dumps, when administering chelation challenge testing. As time goes by, we add interventions until we reach recovery.

Are all chelators compatible with CD?

In many cases there is no need to do both. To answer your question, EDTA is fine as well as DMSA and DMPS with CD.

You may want to consider Bio-Chelat™, bentonite clay or zeolite clay. These are great light chelators.

We have tried every biomed option available, but nothing has helped. What makes CD different?

Biomedical treatments in general provide large amounts of antioxidants. CD is an oxidizer and kills all pathogens in the body. The premise is that once your body does not have to feed and house pathogens, it will work much better and will heal. This protocol focuses on excesses and not deficiencies, which in itself makes it different from many other biomedical interventions. With CD and the Parasite Protocol we are eliminating the pathogens and parasites that cause the symptoms known as autism.

Most biomedical interventions seem to be only applicable to small kids. Does CD work on "older" kids, teenagers or young adults?

Of course! Families with older teens and adults have had success. The body wants to heal at any age be it 7 or 70! This year, two 17-year-olds lost their diagnosis through the use of CD as part of their biomedical protocols. A 32-year-old male is doing excellent on the Protocol as well. Time will tell, and parent/caregiver dedication is the key to success.

How long should I do the protocol to see if it is working for us?

Within 30 days you will know. If you do the Protocol 100%, and in the order laid out here you should see improvements. Some see improvements with their first dose!

Has anyone had an increase in PANDAS/PANS stuff while on CD? Our ATEC didn't budge this time around and it is mainly due to the OCD and other PANDAS stuff that has increased. I get in 16 doses on weekends, but not on weekdays—maybe I get in 12 on a good day during the school week. Has anyone done anything to treat this PANDAS stuff?

Parasites typically coexist with bacteria, so it is very important to treat bacteria and parasites at the same time. It is extremely important to get in 16 doses a day, whether you get in doses earlier in the morning, or send doses to school. Also, many children with PANDAS/PANS need to be treated month long for parasites. If you are giving a probiotic, you can try doing a period without it and see if you child does not improve. We must keep in mind that if you are seeing parasites coming out with the enemas, we may not be seeing gains until we have eliminated a good number of the parasites living in the body. The toxins that they excrete—living and dead—can also cause some of the behaviors we relate with PANDAS/PANS.

Has anyone seen elevated liver enzymes while on CD?

Liver enzymes will not be high from CD. I have watched people do the liver enzyme tests for two years and the kids are in the normal range. However, when they are dumping worms or doing parasite meds is when we may see the enzyme levels go up. Within a couple of weeks they go back to normal, in the handful of times that I have seen them go up. Mebendazole and Combantrin® are non-systemic, and therefore are not absorbed; they basically travel through the intestinal system and are eliminated in the stool and urine.

My son still has a lot of mercury (DMSA provocation). Is it possible that the level can go down from using this protocol? How long should I wait before testing him again?

It is very common for the heavy metal load to go down. If you want to do labs, you can do them in three to four months.

How young is too young for CD?

There is no age that is too young. If a baby is showing signs of cold/flu, etc., you can start with the baby bottle method, one drop in 8 fl. oz. of water, so they are getting 1/8 of 1 drop per dose. The protocols are based on weight.

Ok we have had my son on this protocol for 11 months now. We started at an ATEC of 82 and had a dramatic decrease right away. Our second score was 26! Now we have had 3 consecutive ATECs of 33, most points being in the speech category. Where should I go from here? More diet changes and or supplements? We have been GFCFSF for years, considering going grain free next. We already use omegas and ocean water, and have done nine parasite protocols.

It may be time to look into hyperbarics or GcMAF, as you have *The Diet*, CD and parasite protocol firmly in place. If these 2 options are not financially possible, you need to look at the speech supplement list. Go in order, and see if your child doesn't see gains from some of those. You can also look at chelators. Also, please check that your omega does not have any antioxidants in it, as that will be killing your CD.

Is it normal to see diarrhea in the beginning of the CD protocol?

I consider there to be two types of diarrhea. One is the water faucet, which we want to avoid. If you see water faucet diarrhea, stop for the day, and give a lower dose of CD the next day, when you start up again. The other type of diarrhea is loose, unformed stool, which is normal during detox. As the body is attempting to eliminate the excess toxins quickly, the digestive process will speed up, and not all of the excess water will have a chance to be absorbed through the intestinal tract, causing loose stool.

How do you move CD through inspections at the airport?

Don't carry concentrated sodium chlorite and activator bottles into the cabin. They must go into your checked luggage. Make sure each bottle is tightly sealed. Double bag each bottle SEPARATELY and place at opposite ends of your luggage but not right up against the sides. You may wish to stick each bottle into a pair of socks as extra protection. Surround each bottle with plenty of clothes to insure they will be protected if the bag is roughly handled. According to TSA regulations, you can bring aboard up to 3.4 fl. oz., (100ml) of a liquid in your carry-on luggage, which can hold 3 doses from your baby bottle batch.

Can I use CD as a toothpaste? Should I use regular toothpaste in addition?

I use CD spray (ten drops of activated CD per ounce) on the toothbrush first. Then, brush the teeth, and follow with fluoride-free toothpaste. CD is great for healthy teeth, tongue, and gums.

Is it okay to go swimming when on CD? I worry about the chlorine.

They do have a molecule in common; however one has nothing to do with the other. Some doctors don't let ASD kids swim in chlorinated pools. But, as far as taking CD goes, the chlorine in the pool will not deactivate the CD in your body, nor react with it. A child on CD can swim in a chlorinated pool, just as a child not on CD can swim in a chlorinated pool.

Can I give colloidal silver with CD?

No. Colloidal silver is active for up to 24 hours in the body, and will lower the potency of CD. Therefore the two are not compatible.

Where is the autism clinic in Venezuela?

Puerto Ordaz, Fundacion Venciendo el Autismo. Carolina Moreno is the president and one of my best friends. Her email is venciendoelautismo@hotmail.com. We have over 36 recovered children there. It is wonderful.

My son has been sick and has really bad diarrhea. Can we skip the enemas for a few days or still do them?

We don't use enemas just for constipation, but for overall colon irrigation and health. CD kills pathogens in the colon, and thereby helps heal autism. As the sickness causing the diarrhea is most likely pathogen induced, CD enemas will continue to kill these pathogens and help your son get over this acute situation faster.

My daughter just drank undiluted CDS. What do I do?

First, get her to drink plain water right away. Next, give orange juice or vitamin C to neutralize the CDS. Give burbur or activated charcoal to mop up the toxins that were released. Considering the strong taste, it is unlikely she would have consumed very much.

Can CD be harmful?

In the past 3 years that families have been using CD for autism, we have not seen any children or adults harmed by using this protocol correctly. Liver enzyme tests, nutritional tests, metal porphyrins tests, etc., have consistently shown improvement in the health of the children. Since Jim Humble started using chlorine dioxide for health, no one has died from ingesting CD (or MMS). The story that is often shared around the internet from Vanatu, concerning a woman who unfortunately passed away, was not attributed by the coroner to be the result of ingesting chlorine dioxide. There was another case of a 25-year-old male who attempted suicide with 10g (nearly the contents of a four ounce bottle of sodium chlorite at 22.4% solution) of sodium chlorite (inactivated CD). He developed methemoglobinemia, which he received treatment for, but survived his suicide attempt. That said, as we have mentioned before... anything used incorrectly or recklessly can harm you. Water, table salt, etc. The reason we outlined the protocol in such detail is to prevent mistakes and help families use it responsibly to recover their children with autism.

Does clay or bentonite clay interfere with CD? How far apart should it be taken from CD?

One hour apart from CD is fine. I use diatomaceous earth about ten minutes apart from a CD dose.

Is ibuprofen ok to take while doing CD protocol?

We have never seen a drug that was contraindicated with CD. Always consult your physician for prescription contraindications.

Is it ok to add Epsom salts to a CD bath?

We do not use epsom salts baths in this protocol. We use CD baths. See page 113.

Is it ok to take SAMe (S-adenosylmethionine) with CD? I took out all antioxidants, but not sure about SAMe.

No, it contains magnesium and vitamin C. For a mineral supplement, we take Ocean Water. Vitamin C kills the CD. Magnesium feeds biofilm, and therefore is counterproductive to our goals of healing.

Can I use DMSO in a CD enema to help drive the CD into the parasites?

Warning: Never use DMSO in an enema! Never introduce it into the body rectally. If it is applied rectally, it will carry toxic fecal matter into the bloodstream through the intestinal wall.

How do I use DMSO?

Every chance you can, apply DMSO to a clean body for 20 to 25 minutes and let it soak in. Apply it to affected parts first then remove rings/jewelry to cover all parts of the hands. Allow it to dry while applying to another body area (i.e. right arm, left arm, right leg, left leg). Apply to clean skin with clean hands. Use natural fiber clothing if it is going to come in contact with the DMSO, which can dissolve synthetics. Dr. Stanley Jacob has proven that DMSO brings about healing of rheumatoid arthritis, and it is great for migraine headaches, etc. You can research his work at...

www.dmso.org

...or find a DMSO retailer at...

www.protocolsuppliers.com

The 99% pure stuff is best, but never use that strength on the skin directly. It must be diluted to 70% or less for topical use.

Does DMSO neutralize CD?

No, it does not. However, DMSO has not proven itself to be a tool for autism recovery. In general, we use it for self-injurious behavior (SIB). The "S" in DMSO (dimethyl sulfoxide) is "sulf" as in sulfur. Many children with autism have a spirochete (amongst other pathogens), which feeds on sulfur. We saw many setbacks when adding DMSO. It is used on a case-by-case basis.

Chapter 6

CDS

Another Way of Delivering Chlorine Dioxide

There is more than one way to skin a cat.

~ Seba Smith

In October 2011, I received an email from Jim Humble. After a visit with Dr. Andreas Kalcker, he had some fantastic new information for healing autism. Jim wanted me to fly to the Dominican Republic to see him yet again that year so he could share this new information.

The big news surrounded the creation of *Chlorine Dioxide Solution* (CDS)—a new method of producing and using chlorine dioxide; substantially different from the CD mixing process we were using (described in Chapter 5). Jim was excited about all of the great things that Andreas was developing.

Andreas was contacted by a cattle farmer who was frustrated by the health problems of newborn calves he was receiving. These calves were suffering from infections, diarrhea, ear problems, cysts, coccidiosis, bovine respiratory syndrome, etc. His yearly veterinary drug bill was around €28,000, not including the cost of feed. The farmer had heard about the wonders of MMS in humans and wondered if it might help his animals as well.

Andreas thought, "Sure! Why not?" However, he quickly learned that cattle were very different from humans when it came to their digestive system. Cattle digest through fermentation, which CD disrupts, so he had to find a way to bypass their digestive system and go directly to their blood stream. But, that idea had another big problem. The pH of standard CD was too acidic and therefore not compatible with the calves' blood stream. Injecting even dilute CD caused great pain and could result in damage to the veins. There had to be another way!

After much thought and research, Andreas came up with a distillation process to extract the chlorine dioxide—the key ingredient—from the CD mixture so it would not contain any sodium chlorite or acidic activator. He named the resulting liquid *Chlorine Dioxide Solution* (usually abbreviated CDS). Note: CDS can still contain trace amounts of sodium chlorite and whatever acidic activator used to produce it, but usually not enough to be an issue.

This new CDS solution was injected into the 800 cattle with very positive results. The animals health improved and the vet bill dropped like a rock.

In one case, the farmer called Andreas about a particularly sick cow. He advised him on a dosage, but apparently the farmer misunderstood the instructions and gave the cow 10 times the recommended amount. The result? The cow was a bit "high" with ears and tail standing up, but subsequently became healthy with no ill effects from the accidental overdose.

On the human side, CDS also solves the problem some people have with sensitivity to citric acid as well as the taste of CD. Not having the original chemicals (sodium chlorite/citric or hydrochloric acid) in CDS makes a big difference for some. Another advantage of CDS was a drastic reduction in Herxheimer reactions.

After a long weekend in the Dominican Republic, we went home with new information to try out with Patrick.

I began by swapping out drops of CD for milliliters of CDS. Initially, I didn't notice improvements or regression, but I did notice that Patrick was waking earlier and having a less profound sleep. After about 30 days of trying the new miracle I began to think that, "If it ain't broke, don't fix it." So we went back to the original CD that had given us all of the recoveries and improvements. Patrick went back to sleeping very well and I never looked back. I was a confirmed CD user and CDS, for me, had no obvious advantages.

A little over a year later, Jim was at our home, and on December 26th, 2012, his new assistant showed up from the Middle East to help him. She could be called the "CDS poster child." We differed on this topic right away. But, she was very insistent. She asked me why I didn't like CDS and I told her. So, she told me that I had been using "unstabilized" CDS and now they "stabilize" the CDS.

When the CDS is "stabilized" (with some sodium chlorite added back into the CDS liquid), it gives an extra kick, such that when the chlorine dioxide molecule oxidizes a pathogen, there is sodium chlorite on hand to react with the acid that is created by the death of the pathogen. The result is more chlorine dioxide released on the spot.

I decided to try stabilized CDS with Patrick. We arrived at his full dose based on tolerance. His sleep has been perfect. He sleeps as any pre-teen should—like a rock. And, he has great energy all day long. I can say that there was no noticeable difference over the four months that he was taking CDS, from the days when he took CD. However, he is a little bit "lighter," as if he has no cares in the world. It seems as if something that might have made him uncomfortable no longer bothers him. Little by little, a few families switched to CDS to see if their children did better, worse, or as well as they did with CD.

After a couple of trial months with about 20 families, I think that it is safe to say that CDS has its place in the buffet of different methods of chlorine dioxide administration. We still have the original CD, made with citric acid or HCl. Now, we have CDS and CDH (Chapter 7). Today, we no longer use stabilized CDS, as I feel that CDS is only for hypersensitive individuals and to be used only until they can switch over to CD or CDH. Whatever formula we choose, it will help to heal the body of unwanted pathogens, and helps with parasites as well as heavy metals. Each person is different, and it is up to us as parents to observe and decide which method of chlorine dioxide administration is best.

You are in no way encouraged to switch to CDS if what you're doing with CD is working for you. For some people it may be better, but as of now, only one of our recoveries has come from CDS—all the rest are from CD. Some children may try CDS and come to realize that CD is still better for them.

When to Use CDS

CDS is used when someone just can't get past 1 drop divided in 8 fl. oz., over a day without experiencing a Herxheimer reaction. This allows us to get past the road block and still detoxify the body. Eventually, we want to go back to the standard CD Protocol which I believe to be more effective.

This is not a common situation, but it does occur.

Making vs. Buying CDS

Most people are a bit concerned about making CDS since it involves a distillation process where you produce chlorine dioxide gas and then cause it to go into water. When Andreas first came up with making CDS, it involved two containers, a hose, heat and lots of ventilation. There are still lots of videos on YouTube explaining that original process. However, there is now a much better, easier and safer way to produce CDS at home. So, if you see a CDS making video or read instructions that talk about using a hose—forget it! The new process is much easier and safer and doesn't involve any heat nor hoses.

There are some online sellers of CDS, but the cost of it is high and making your own is simple. It just takes a little effort in learning the process and having the right equipment. In addition, by making your own, you will have a better understanding of the behavior of chlorine dioxide.

Some Important CDS Points

When it comes to CDS, it is all in the concentration of chlorine dioxide in water; measured in *parts per million*—abbreviated **ppm**.

When buying or making your own, the goal is to end up with a bottle of CDS having a concentration strength of 3,000ppm. This bottle should ideally be made out of glass and kept in the refrigerator. You never use this concentrated solution directly. It must be diluted.

When we say, for example, use 10ml of CDS, we mean mix 10ml of 3,000ppm CDS with whatever indicated amount of water, and use that diluted mixture for the indicated purpose.

Make sure you read and understand Appendix 6, page 467 which explains more about concentration and measuring ppm.

There are several names attributed to this new, better and simpler method of making CDS, such as:

- The Shot Glass Method
- The Overnight Method
- The New Method

There are some demonstrations on YouTube showing this new method along with write-ups on various forums. These methods use a plethora of different containers having different shapes and sizes. The specific tools used to produce CDS can have a significant impact on the concentration.

It should also be noted that the concentration strength is reduced every time you open the CDS source bottle as chlorine dioxide gas escapes the liquid and evaporates into the air space above the liquid. When you open the source container, some of the chlorine dioxide gas escapes. So, the strength of a half full CDS bottle can be substantially less than what you started off with when the bottle was full. This is similar in concept to a pop bottle that loses its fizz if not used up quickly.

To substantially reduce the evaporation of the chlorine dioxide gas, the CDS source bottle should be stored in the refrigerator, which stabilizes and significantly reduces the amount of CD gas escaping the CDS solution. But, no matter what the temperature, some gas will always escape and reduce the concentration each time the bottle is opened. A narrow neck bottle helps maintain the integrity of the CDS for longer periods.

It is better to store the concentrated source CDS in smaller 300ml bottles than keep it in a larger one liter bottle. That way you use up what is in one bottle faster with less concentration loss. A good storage bottle is shown and described on page 160.

The New CDS Making Method — The Basic Concept

First, let's go over the basic principles of the CDS making process and then we will go into details.

We start with two containers (both ideally made of glass). One container is larger and has some kind of a lid that can be tightly closed. The other is a smaller container having NO lid. The relative size of each container is such that the smaller container fits comfortably inside the larger container while allowing the lid to properly close on the larger container when the smaller one is inside. See page 151 for an example of a well-made glass kit distributed by wps4sale.com.

The large container is then partially filled with filtered/distilled water. The level of water must be low enough so as to not overflow into the smaller container if it is placed into the partially filled larger container. The photos on page 151 show the concept of the two bottles visually.

Now that you have the basic idea, let's make some CDS...

Using the Overnight CDS Generating Kit

The following directions assume you have the glass *Overnight CDS Generating Kit (1000 ml)* from wps4sale.com (as shown on page 151). You are of course not required to use their kit. However, based on our own experience of looking high and low for a good one, theirs has the key features:

- Made of glass
- Smaller glass fits inside of larger container AND 1000ml of water inside the big container will not flood into the smaller container once placed inside.
- Lid fits tight and no gas escapes during the chemical reaction time.
- Lid is plastic—NEVER USE METAL LIDS!

Note: There are plastic CDS kits that work. You may be able to find a combination of containers made of plastic but not glass where you live. In fact, I made a YouTube video where I show how to make CDS using a plastic container. Just understand that chlorine dioxide will deteriorate plastic containers over time, so it should be avoided. Even if you have to use a plastic jar to make the CDS, store the final product in glass. The less time the CDS is in plastic, the better! As for lids, we have no choice. Only use plastic lids. Metal lids will oxidize VERY rapidly! Even metal lids with a plastic coating on the inside will deteriorate.

Steps to Making CDS:

1. Start by pouring one liter of distilled or reverse osmosis water into the large glass container. The water should be around room temperature.

2. Place the empty dessert cup inside the large container making sure no water spills into the cup.

3. Measure 75ml of sodium chlorite solution (22.4% $NaClO_2$ in water) and pour into the dessert cup, while being careful not to spill any sodium chlorite into the water surrounding the cup.

4. Measure 75ml of 10% HCl or 50% citric acid solution and pour into dessert cup.

5. IMMEDIATELY screw on the lid to the large container while being careful to not jiggle the container which could potentially cause the CD solution

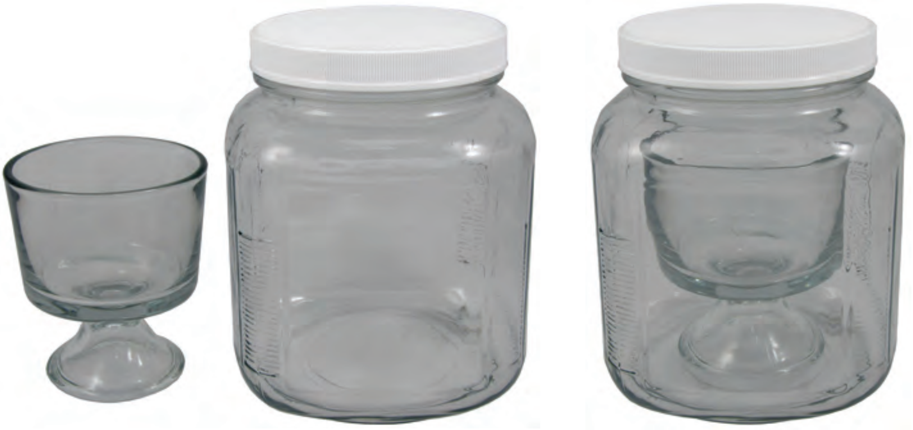

wps4sale.com sells this Overnight CDS Generating Kit that produces 1 liter of CDS at 3,000ppm from 75ml of sodium chlorite plus 75ml of activator. The only part not made of glass is the plastic lid. The glass dessert cup is sized just right to fit inside the larger glass jar with enough room left over for 1 L of water that won't overflow into the dessert cup.

to mix with the surrounding water. Make sure the cover is tightly screwed down so as to not have any leaks. If you smell chlorine dioxide (beyond the little that escapes while you are putting on the lid) then something is wrong with the lid and/or container.

6. If you did everything correctly you should see the CD solution turn a dark amber/brown within about a minute (see below) and notice that the water starts to turn a light yellow several minutes later. It is also normal to see bubbles form in the dessert cup. It is normal for the pressure to change during the chemical reaction. We find that citric acid produces a vacuum and HCl can produce a pressure at first and then a vacuum.

7. Cover the kit with a towel to reduce its exposure to light. It doesn't have to be in a pitch black location, but it definitely should not be in direct sun light. Feel free to check it from time to time and see how the colors change. Of course keep this kit out of the reach of children or anything that could disturb it. I keep mine in the cabinet overnight with a towel on it.

8. When the color in the dessert glass matches that of the surrounding water, the chemical reaction is complete. There is no harm in waiting a bit longer—nothing more will happen. This usually takes from 12-24 hours depending on room temperature and the strength of the acid used.

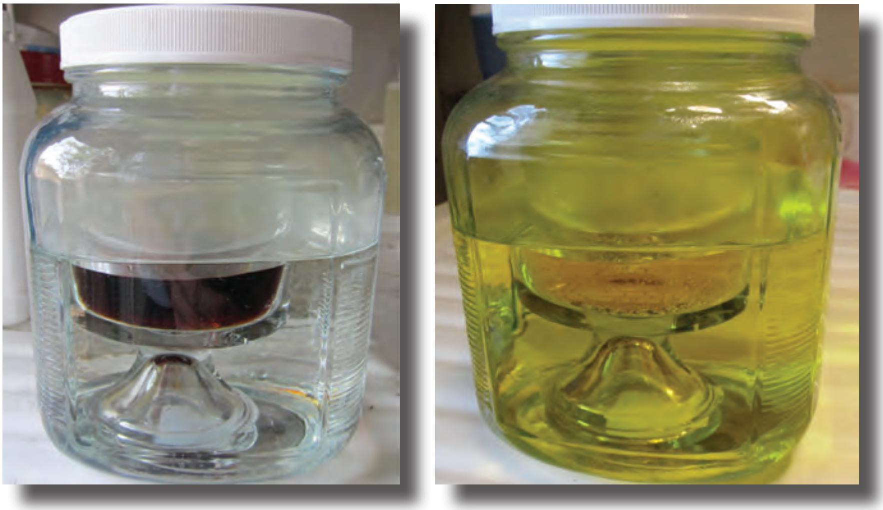

After the sodium chlorite and activator mix, the color will quickly turn a dark amber, almost black color (left). After about 12 hours, the color of the solution in the dessert cup and that of the water surrounding it should be identical (right), which indicates the chemical reaction has completed and the CDS is ready to be poured into a storage container.

---

9. Have your glass storage bottle(s) clean and ready to receive the CDS solution.

10. Before opening the kit make sure your work area is well ventilated. Perform the steps that follow outside or have the door and/or window open nearby. Having a fan lightly blowing any gas that escapes away from you is a good idea.

11. Slowly open the container being careful to not jostle it which could cause the CD solution to mix with the freshly made CDS. If the chemicals mix, you will have to start over again, so be careful! You will most likely smell the chlorine dioxide gas that was sitting in the air space when you first open it. This is normal, hence the need for good ventilation.

12. Slowly remove the dessert cup while being careful to not spill its contents into the CDS. Immediately dispose of the contents in the toilet and rinse off the dessert cup.

13. Now pour the CDS into the storage bottles while being aware that some chlorine dioxide gas will come out of the water into your working area.

14. Tightly seal the storage bottle(s) and place in the refrigerator.

Note: 10% HCl is the recommended activator when making CDS. However, you can also use 4% of HCl, but the reaction time will be slowed down. It may take as much as a full day before the color of the CD solution and the water match.

Dosing CDS

As mentioned earlier, CDS is particularly well-suited for unusually sensitive individuals who cannot tolerate even low doses of CD.

Start with a baby bottle filled with 8 fl. oz., of water and add 1ml of CDS. This will provide eight doses of 1/8ml. If all is well on the first day, increase to 2ml on the second day. Each day you increase by one milliliter. Increase the amount based on tolerance. If the person is having a problem with a certain level of CDS, then go down to where they were fine and stay there a few days before increasing again.

Tolerance is the key here. Once you can no longer increase you have reached that person's full oral dose of CDS.

Slowly Switching Back to CD

Once you have found the full oral dose where they are stable and cannot go higher, it is time to fold CD back in.

The goal of using CDS is not to replace CD, but rather do some fundamental detoxing so you can bring CD back and reach full oral dose of CD.

This is accomplished by taking out one milliliter of CDS and replacing it with 1 drop of CD. For example, if the individual can tolerate 20 milliliters of CDS, then an ideal transition would be represented in the chart on page 154.

Of course this chart represents an unlikely scenario where everything goes exactly as desired. However, it does demonstrate what we are striving for—switching out CDS for CD and continuing to increase the dose of CD as described in Chapter 5 (page 81).

To accomplish this process effectively requires careful observation of your child and making decisions as to when you can go up; when to hold; and when to back down.

Hypothetical CDS to CD Transition

Transition Day	CDS in milliliters	CD in drops	Comments
1	20	0	Last CDS only day.
2	19	1	Transition starts
3	18	2	Transition continues
4	17	3	
5	16	4	
6	15	5	
7	14	6	
8	13	7	
9	12	8	
10	11	9	
11	10	10	
12	9	11	
13	8	12	
14	7	13	
15	6	14	
16	5	15	
17	4	16	
18	3	17	
19	2	18	
20	1	19	
21	0	20	Transition completes
22	0	21	CD increases each day
23	0	22	
24	0	23	
25	0	24	

CDS Enemas

CDS enemas can be useful for sensitive individuals when starting the protocol. Follow the same enema instructions in Chapter 5, page 103, keeping in mind that 1ml of CDS has a relative strength of approximately 60% of 1 drop of CD. With that in mind you can work up to 40ml of CDS per liter of enema water.

Chapter 7

CDH - Going Beyond CD & CDS

"The progressive development of man is vitally dependent on invention. It is the most important product of his creative brain."

~ Nikola Tesla

This protocol is constantly evolving, and until every person with autism recovers, we will continue to search for things to add or tweak to the protocol so we reach that goal. A few months before the release of this second edition, I was made aware of a new method of CD preparation, known as *Chlorine Dioxide Holding (solution)* or simply CDH.

CDH can almost be described as a hybrid between classic CD and CDS. Where CDS has no raw material left in the final product, but only chlorine dioxide gas dissolved in water, CDH contains some raw material (similar to classic CD) along with the chlorine dioxide in the final product. This new process allows the sodium chlorite to react with the acid for a significantly longer period of time, thereby reducing much of the remaining amount of unactivated sodium chlorite and activator. Some people cannot tolerate citric acid so CDH is usually made with 4% HCl. Initial reports indicate that CDH is better tolerated than CD.

Another interesting benefit of CDH is that it appears to mix well with the permitted natural sweetener stevia (SweetLeaf® brand) without reducing the potency of CDH. This can help children who have an aversion to the taste of CD. Many families have also reported that they were able to increase their child's dose without producing a Herxheimer reaction, in contrast to classic CD. Older children and severely affected children have benefitted as well from the CDH preparation; you can read more about that on page 221.

Currently, this new method is being used by a relatively small group of families (around 70 as of November, 2013). Many of them are reporting that CDH continues to produce results for their children on the spectrum, and they are seeing even better things than before.

As with all new things, it is important to test with a diverse group of families over a long enough time period to make sure that gains are sustained. So we ask that you please keep in mind that CDH is literally the bleeding edge of this protocol. While we are excited to share a new option with you, the decision to use CDH must not be taken lightly. If you are doing well with classic CD you may never need to use CDH. Consider the old saying, "If it ain't broke, don't fix it." Of the 115 children who have lost their autism diagnosis, 114 have done so with CD and 1 with CDS. So far, 0 with CDH (4 months in use for ASD). I expect this number to change soon as we are seeing good things from CDH. Time will tell. Stay tuned.

Scott McRae, his wife Brenda, and Charlotte Lackney have been pioneers in developing the CDH method. The following section is written by Scott where he discusses how CDH evolved along with detailed instructions on how to prepare and use it.

* * *

My wife, Brenda, and I first learned about CD early in 2009, through Bhante Vimalaramsi, an American born Buddhist monk. He's now a very close friend of ours, and even as I'm writing this, I'm at a meditation retreat which is based upon his understanding of the open-minded, experiential teaching of the Buddha.

After receiving our first set of CD bottles from an online seller, we started by trying to get to 15 drops of CD activated with lemon juice for 3 minutes, twice a day, which was the protocol at the time. Though we gave it a good effort, we couldn't get past 6 drops before vomiting and having diarrhea, and this caused us to initially stop taking CD. Then, after moving to Jakarta, Indonesia (Brenda's home town), I experienced almost monthly illnesses, which I believe was due to being in a new and tropical part of the world (I'm from San Diego, CA), riding in public transportation twice a day and teaching in a school of over 300 students. After a year and a half of constantly being sick, I remembered our bottles of CD which we had fortunately brought with us from the US, and I started on Protocol 1000. By that time, *Protocol 1000* had been developed to be what it is now—3 drops of CD (activated for 20 seconds with 50% citric acid) per hour, 8 times a day—and so I decided to do the 3 week cleanse. By doing the cleanse, I experienced a huge improvement in my health. Not only did I immediately stop getting sick, but I also felt that my energy level had increased by about 25%! Obviously, I was excited by these great results and so was my wife because she had started taking CD again as well. However,

Scott & Brenda McRae from Jakarta, Indonesia

even though our results were great and our experience was far better than before with the old, "try to get to 15 drops twice a day protocol," we still were having some nausea and bouts with diarrhea when we took CD. This was especially true at the times when we felt the symptoms of a cold or flu and tried to take more than the 3 drops doses in order to overcome it. Still the results we were getting with CD on Protocol 1000 outweighed the nausea and diarrhea that we experienced and so we continued using it, especially when we felt some sickness coming on.

After taking CD for about a year in this way, CDS hit the CD world. Being the experimental type, I decided to give it a try. After searching all over Jakarta for the plastic tubing for a couple of days and finally finding it, I made our first batch of CDS. We both tried CDS for about six months but found it to be less effective than CD, so we stopped with CDS and went back to CD and Protocol 1000.

Nevertheless, there were 2 things I really liked about CDS: (1) it NEVER caused us any nausea or diarrhea, and (2) it was so easy to use since it was pre-made (no mixing of chemicals before each use). The main thing for me was not having any more nausea because I really disliked feeling sick. So, I started thinking about CDS and my experience with making it and I came to the conclusion that possibly, the reason why CDS wasn't nauseating was because there wasn't any unactivated sodium chlorite in the solution—it was just chlorine dioxide dissolved in water. I remembered that when I made CDS myself, even after the activation process had gone on for an hour in a heated condition, if I stirred or swirled the activation chamber bottle, more ClO_2 would still come out of it. This got me to thinking... if the chemical reaction between the sodium chlorite and the 50% citric acid solution used to make the CDS was still able to produce more chlorine dioxide after an hour in a heated condition, then surely the 24 drops that I was using to make my daily CD dosing bottle weren't being fully activated after just 20 seconds.

Therefore, I decided to increase the activation time and tried it out on myself first and then my wife (aren't I considerate?). We both found there was no longer a nausea problem, even when taking more drops per hour than we had ever taken before.

Since I was in an experimental mode, I decided to see if I could make larger quantities of concentrated CD all at once to make it more convenient. I mixed equal amounts of sodium chlorite and 50% citric acid in a bottle, let it activate for about a minute, and then added a specific amount of hot water to further encourage the activation process. In the end, I had a total of 140 ml of this concentrated CD solution. Later, I found that hot water wasn't necessary and that room temperature water worked even better because there was less CLO_2 gas lost in the process.

The finished chemistry was indeed very strong and still didn't cause any nausea, so I was compelled to post my discovery on the Genesis II Forum. I named it *7 Day Fridge MMS (CD)* because it provided a 7 day supply of Protocol 1000. The ingredients added up to 140 ml, so each 20 ml was 1/7th of the total, or a one day supply of "pre-activated" CD similar to Protocol 1000. This made taking CD every day really easy. Just pour out 20ml of the new concentrated solution into a water dosing bottle, and then pour out 1/8th of the bottle every hour into some water in a glass, and drink it.

The *7 Day Fridge MMS (CD)* method worked great for my wife and me. Others on the Forum also tried it and liked it. In addition, I gave it to some people at the school where I work to overcome their colds (usually overnight) as well as other diseases within a short time.

About 18 months later, Charlotte, my Forum friend, began testing the *7 Day Fridge MMS (CD)* process to determine the actual ClO_2 content using her *Sensafe*^{TM} Chlorine Dioxide Photometer. Over several months, we worked together to further refine the 7 Day Fridge method. The result of all of our testing and refinements is this new and exciting CDH product.

The final formula to produce CDH using the 1 Bottle Method came out to be:

22 parts water	91.6%
1 part sodium chlorite	4.2%
1 part HCl (4%) or Citric Acid (35%)	4.2%
Total Solution	**100%**

I believe this new way of making CD will be of great benefit to humanity because it will allow people to gradually double or even triple the amount of CD they can take with little or no stomach upset. By increasing their doses to higher levels, people will be able to overcome CD treatable diseases more quickly than ever before.

As with any new technology, more changes and developments are likely. For example, we now know that CDH tastes much better if activated with hydrochloric acid instead of citric acid.

Producing CDH Using the One Bottle Method

Making CDH is really quite simple. You can use any size of bottle, and produce any quantity you wish, as long as you follow the basic instructions and keep the proportions the same. However, before diverting into different quantities, it is recommended you follow these directions exactly to insure you have the process down correctly.

Note the following volume equivalents:

Fluid Ounce (U.S.) Approximate Equivalents

30ml = 1 fl. oz. US (2 Tablespoons)

660ml = 22 fl. oz. US

720ml = 24 US fl. oz. US

750ml = 25 US fl. oz. US

Equipment Needed

- One 750ml (25 fl. oz. US) glass bottle with an airtight cap. DO NOT use a metal cap (even if it is lined with plastic). The best caps are plastic or even a synthetic cork in good condition. A common 750ml wine bottle is ideal for this as you can get them colored to reduce possible UV exposure (the darker the better). But, a colored bottle is not absolutely necessary and using a clear one will allow you to actually see the chemical reaction as it turns from clear to yellow.
- Three 240ml (8 fl. oz. US) bottles ideally made of colored glass to protect the CDH from UV light (the darker the better), but colored bottles are not absolutely necessary. You can also use smaller bottles if you like—these are just used to divide up the resulting 720ml of CDH into smaller bottles to help retain the ClO_2 concentration while opening and closing the bottles during dosing. *Schweppes*™ sells 6-packs of Ginger Ale, Club Soda and Tonic Water in 10 fl. oz. glass bottles with plastic lids that are great for this purpose, and hold 300ml easily. Only the Ginger Ale is in a colored bottle (green), the others are clear.

Schweppes
GINGER ALE

- One measuring cup or graduated cylinder to accurately measure liquid in either milliliters or fluid ounces.

Ingredients Needed

Ingredients should be at room temperature—not cold. If CD and/or activator are right out of the refrigerator make your water warmer to offset, or allow ingredients to warm up before using.

- 660ml of distilled or purified water at about 70° - 90°F (21° - 32°C).
- 30ml of sodium chlorite (22.4% solution) near room temperature or slightly above.
- 30ml of 4% Hydrochloric Acid (HCl) or 35% Citric Acid ($C_6H_8O_7$) near room temperature or slightly above.

Note: The amounts indicated above add up to 720ml, while the wine bottle easily has room for an additional 30ml and more. Refer to the chart on right if you wish to produce a different batch amount.

The CDH Formulation Table and Different Acid Concentrations

What if you have 10% Hydrochloric acid or 50% Citric Acid (very common)? Or, you wish to use a different size of bottle? Not a problem. These acids can still be used. However, the formula changes accordingly.

The table below is a great tool for determining the formula for a given bottle size. To use the table, start by circling the size of the bottle you wish to fill in the left most column. Next look at the acid you have and its labeled concentration. Match that with one of the 4 options across the top. Below the matching acid/concentration you will find the 3 formulation numbers for water, sodium chlorite (labeled "SC"), and whatever acid you are using. Just go down the appropriate 3 columns to where the bottle size line intersects and you will have the numbers you need. Substitute these numbers in the following preparation instructions if your situation calls for it.

CDH Formulation Table

(Applies to the One Bottle Method of making CDH ONLY!)

SC=Sodium Chlorite / HCl=Hydrochloric Acid / CA=Citric Acid

Water Bottle (ml)	Hydrochloric Acid 4%			Hydrochloric Acid 10%			Citric Acid 35%			Citric Acid 50%		
	Water (ml)	SC (ml)	HCl (ml)	Water (ml)	SC	HCl (ml)	Water (ml)	SC (ml)	CA (ml)	Water (ml)	SC (ml)	CA (ml)
10	9.2	0.4	0.4	9.4	0.4	0.2	9.2	0.4	0.4	9.3	0.4	0.3
20	18.3	0.8	0.8	18.8	0.8	0.3	18.3	0.8	0.8	18.6	0.8	0.6
30	27.5	1.3	1.3	28.3	1.3	0.5	27.5	1.3	1.3	27.9	1.3	0.9
40	36.7	1.7	1.7	37.7	1.7	0.7	36.7	1.7	1.7	37.2	1.7	1.2
50	45.8	2.1	2.1	47.1	2.1	0.8	45.8	2.1	2.1	46.5	2.1	1.5
60	55.0	2.5	2.5	56.5	2.5	1.0	55.0	2.5	2.5	55.8	2.5	1.7
70	64.2	2.9	2.9	65.9	2.9	1.2	64.2	2.9	2.9	65.0	2.9	2.0
80	73.3	3.3	3.3	75.3	3.3	1.3	73.3	3.3	3.3	74.3	3.3	2.3
90	82.5	3.8	3.8	84.8	3.8	1.5	82.5	3.8	3.8	83.6	3.8	2.6
100	91.7	4.2	4.2	94.2	4.2	1.7	91.7	4.2	4.2	92.9	4.2	2.9
150	137.5	6.3	6.3	141.3	6.3	2.5	137.5	6.3	6.3	139.4	6.3	4.4
200	183.3	8.3	8.3	188.3	8.3	3.3	183.3	8.3	8.3	185.8	8.3	5.8
250	229.2	10.4	10.4	235.4	10.4	4.2	229.2	10.4	10.4	232.3	10.4	7.3
300	275.0	12.5	12.5	282.5	12.5	5.0	275.0	12.5	12.5	278.8	12.5	8.7
350	320.8	14.6	14.6	329.6	14.6	5.8	320.8	14.6	14.6	325.2	14.6	10.2
400	366.7	16.7	16.7	376.7	16.7	6.7	366.7	16.7	16.7	371.7	16.7	11.7
450	412.5	18.8	18.8	423.8	18.8	7.5	412.5	18.8	18.8	418.1	18.8	13.1
500	458.3	20.8	20.8	470.8	20.8	8.3	458.3	20.8	20.8	464.6	20.8	14.6
550	504.2	22.9	22.9	517.9	22.9	9.2	504.2	22.9	22.9	511.0	22.9	16.0
600	550.0	25.0	25.0	565.0	25.0	10.0	550.0	25.0	25.0	557.5	25.0	17.5
650	595.8	27.1	27.1	612.1	27.1	10.8	595.8	27.1	27.1	604.0	27.1	19.0
700	641.7	29.2	29.2	659.2	29.2	11.7	641.7	29.2	29.2	650.4	29.2	20.4
720	660.0	30.0	30.0	678.0	30.0	12.0	660.0	30.0	30.0	669.0	30.0	21.0
750	687.5	31.3	31.3	706.3	31.3	12.5	687.5	31.3	31.3	696.9	31.3	21.9
800	733.3	33.3	33.3	753.3	33.3	13.3	733.3	33.3	33.3	743.3	33.3	23.3
850	779.2	35.4	35.4	800.4	35.4	14.2	779.2	35.4	35.4	789.8	35.4	24.8
900	825.0	37.5	37.5	847.5	37.5	15.0	825.0	37.5	37.5	836.3	37.5	26.2
950	870.8	39.6	39.6	894.6	39.6	15.8	870.8	39.6	39.6	882.7	39.6	27.7
1000	916.7	41.7	41.7	941.7	41.7	16.7	916.7	41.7	41.7	929.2	41.7	29.2

Preparation Instructions

Follow the steps of mixing ingredients in this order (assuming you are using the 750ml glass bottle):

1. Pour 660ml of purified water into the 750ml glass bottle.
2. Add 30ml of sodium chlorite to the 750ml glass bottle.
3. Add 30ml of 4% HCl or 30ml of 35% citric acid to the 750ml glass bottle.
4. Immediately cap/cork the bottle tightly so that no ClO_2 gas can escape and give it a good shake to thoroughly mix the ingredients (and a few more times later on if possible).
5. Store the bottle in a dark place with a temperature of 70°F to 90°F (21°C to 32°C) for 12 to 24 hours or more (24 hours or longer if you're in a cold environment—below 70°F (21°C)).
6. After the storage time has passed, place the bottle of CDH into the refrigerator (not the freezer) and allow it to cool down for 3+ hours before opening it for the first time. The temperature of the solution should be no more than 51°F (10.5° C). Note: If you are using a significantly smaller bottle, cool down time can be reduced because the smaller volume of solution will cool faster.
7. Finally, pour the CDH from the 750ml bottle into the smaller bottles, cap tightly and keep refrigerated until ready to use. The smaller bottle are easier to dose from and they also help reduce the number of times gas can escape and reduce the potency of the solution.

Using CDH

The CDH is now ready for use. Each milliliter of CDH solution contains 1 pre-activated drop of CD. It can be used for ANYTHING that CD is used for; viruses, bacteria, yeast, parasites, heavy metals, enemas, tub baths, gums & teeth, skin care, infections, etc.

When used orally, each 1ml of CDH should be added to at least 30ml of water. You can add even more water if you notice slight throat irritation with higher doses. If taste is an issue, a little bit of stevia may be added to each dose to improve taste by sweetening it. Some of the moms are adding SweetLeaf® Natural Stevia Sweetener to their daily bottle and report no negative impact on ppm level.

Although CDH is strong, it has shown to be gentler on folks who have issues with nausea when using traditional CD. You should be able to start at whatever drop dose you were on with CD and switch to an equivalent milliliter dose of CDH and gradually increase to tolerance. Typically, people are able to take 2 to 3 times as much of CDH as traditional CD, without experiencing nausea.

The CDH bottle should be kept refrigerated and only taken out to extract doses. Since 720ml is a fairly large quantity that may take many days to finish, it's a good idea to take the 720 ml and divide it up into 3 smaller 240ml bottles (720ml ÷ 3 = 240ml) so you won't lose much of the ClO_2 each time you open the bottle. By doing this, you will conserve as much of the ClO_2 as possible. It's also easier to extract doses from a smaller bottle using a syringe or pipette.

Also, keep the CDH bottle out of direct or indirect sunlight to prevent loss of ClO_2. If you make a dosing bottle for the day, it's best to keep it cold, but it is not absolutely necessary.

A little note about taste: Most people who complain about the bad taste of CD, CDS or CDH (which does not have any taste at low doses) are actually reacting to the smell of the ClO_2 which can lead to developing a long-term aversion to any of the treatment solutions. So, if you can minimize the gas floating around your nose, you will have an easier time with drinking the CD, CDS or CDH dose. To accomplish this, Charlotte suggested avoiding the use of a cup or wide mouth drinking bottle. Instead, use a bottle with a small opening such as a common water bottle (preferably made of glass). Of course, if the smell doesn't bother you, this is a moot point, but at least you have this little trick if it does.

Well, that's all you need to know to get started on this great new way to make and use CD. May this new CDH formulation bring you and your family much health and happiness.

Scott McRae
Jakarta, Indonesia
November, 2013

* * *

We appreciate Scott, Brenda and Charlotte's contribution to the variety of ways to produce and use chlorine dioxide. Be sure to check the Facebook groups and forums for the latest developments on CDH.

Some of you may feel overwhelmed by what you just read. So, let me give you my "Easy-Peasy" single paragraph method of making CDH: I take a 600ml Lifefactory™ bottle, put in 550ml of water; add 25ml of sodium chlorite; followed by 25ml of 4% HCl (or 50% citric acid). Leave it 12 hours in a cabinet, after which it goes in the fridge for 2 or more hours. Done! ☺

Healing
the symptoms known as
Autism
Kerri Rivera
with Kimberly McDaniel
Healing the Symptoms Known as Autism
Kerri Rivera
With over 120 testimonials from around the world!

Healing & beating autism.

Chapter 8

Step 3
The Kalcker Parasite Protocol

> *All of our fathers had a treatment for parasites as a part of their cultural practice. We have gotten away from this because of our reliance on modern practice. We would do quite well to relearn the ways of our ancestors in this area and keep ourselves in relatively good health always.*
>
> ***~ Chief Two Trees***

The word "parasite" comes from the Greek word meaning, "one who eats off the table of another." Parasites, to Ancient Greeks, were those who sat at another's table, and paid for their meal with flattery.

As I mentioned earlier in the book, a very interesting thing started happening with the CD enemas. What we had previously believed to be mucous or biofilm coming out with the enemas, turned out in many cases to be worms (helminths) (aka parasites)—in rare cases they were still alive and wiggling in the toilet! We believed the most common were *Ascaris lumbricoides* (roundworm). However, what we are now seeing more and more appear to be rope parasites, a potential new species of helminth discovered by Dr. Gubarev, Dr. Alex Volinsky, and coworkers (submitted January 14, 2013). DNA testing is the only way to definitively say, but at $25,000 USD for each analysis, with a minimum of 100 test cases, it is rather cost prohibitive at the moment.[1]

In addition to Ascaris and rope parasites, parents have also seen hookworms, pinworms, tapeworms, and flukes, among others. This is an extremely important piece of the puzzle for so many of our children. We have been led to believe that in first world nations, parasites are not a problem. This is absolutely not the case.

A well washed parasite. You can almost feel the texture.

Andreas Kalcker,
co–author of the
Kalcker Parasite Protocol

Miriam Carrasco Maceda,
co–author of the
Kalcker Parasite Protocol

I was honored to have Dr. Andreas Kalcker and Miriam Carrasco Maceda share a chapter from their upcoming book, *Parasites: The Silent Enemy*. Andreas explains the importance of lifelong deworming, and shares with us a protocol that has helped many children and adults become healthier; and for some children on the spectrum was the last piece added that led them to recovery. The version included here has been tailored specifically for children and adults with ASDs. The original protocol can be found at:

www.andreaskalcker.com/
index.php/en/health/parasite

Very few details have been changed, but it is important to note that the Parasite Protocol here is what has been proven to help many of our children on the spectrum, including many of the recovered children.

Thank you Andreas and Miriam for your valuable contributions to this movement, selflessly sharing your findings, and for always taking the time to help.

HOW TO DETECT AND TREAT A PARASITIC INFECTION

Parasitic infections are more common than most people think, and may or may not result in serious health problems. We may be infected with multiple types of parasites, which vary in size and location, on or in the body.

Parasites can be classified as either microparasites, such as malaria that are only visible under the microscope, or large macroparasites such as round or flat intestinal worms (roundworms, tapeworms, etc.). These can be seen by the naked eye, and can reach great sizes. Internal parasites are found, not only in the intestines, as is generally thought, but anywhere in the body, including the lung, liver, muscle, stomach, gallbladder, brain, blood, skin, joints, and even in the eyes.

In recent history, the great migratory movements of the human population via rapid transportation and widespread trading have shortened the distances that previously had separated people and diseases. Formerly localized diseases have thus become universal ailments. Parasites previously confined to very specific geographical areas now appear in other locations, far away from their initial homelands. Unfortunately, conditions typical of the lower socioeconomic strata, (under which a large percentage of the global population lives) tend to favor the transmission of diseases and parasites.

A high percentage of the world population suffers from infections by parasites, which the WHO (World Health Organization) estimates are responsible for 15 million child deaths annually. In addition to the great cost represented by deaths, chronic and persistent infections have increased as parasites have developed multiple mechanisms of evasion and resistance to specific immunity. This allows them to circumvent and cancel the host immune response.

Persistent parasitic infection in human hosts leads to chronic immune reactions, which can result in tissue damage and altered immune regulation. Ninety percent of the world population is infected with one or more parasites, and up to five different types may coexist in the same host.

This situation becomes dangerous when the internal balance within the host is upset, the number of parasites skyrockets, and the host begins showing signs of serious illness that may even result in death. However, in some cases, parasitic worm infections do not result in disease, in fact, a number of carriers are found to be healthy.

A really good look into a parasite. You can see the actual intestines of the parasite. Also known as a helminth.

The same parasite (as above) from a different angle.

Considering that most people are not even aware of their infections, parasites have become silent killers, claiming the lives of many unsuspecting victims going about their lives oblivious to the danger. Some doctors in Western Europe and the U.S. seem unwilling to even contemplate that we may be infected.

Taking into account the recent increase in travel, immigration, and trade across continents, it is not hard to see how the problem has now become magnified to an alarming level. Parasites, especially the modern "toxified" versions, may well be causing many of the rare diseases now becoming more prevalent, as well as other recently identified or growing problems such as chronic fatigue, fibromyalgia, and arthritis.

The most common verminosis (infestation with or without obvious symptomatology of disease caused by parasitic worms) is intestinal. People who have them not only suffer from a large quantity of lost nutrients (absorbed by the parasites), but also from perforations made by worms in the digestive tract that can open the door to various infections and possible autoimmune deficiencies. Intestinal worm infections are very common and can affect everyone, not only people with poor hygiene habits. Helminths (worms) are transmitted by ingesting the eggs or larvae of parasites, which then hatch in the intestinal tract.

A parasitic infection or reinfection can be acquired through one or more of the following avenues:

- From more or less direct contact with an infected person (fecal or sexual).
- From self-infection, for example, through anal-hand-mouth contact. By scratching the anal area, eggs can become lodged under the fingernails.
- From congenital transmission (mother to fetus).
- From commonly contaminated objects.
- From soil contaminated by human or animal excrement.
- From eating contaminated raw or undercooked meat.
- From eating raw fish.

In some countries, raw fish is included in traditional foods. We can avoid the consumption of the larvae or worms by freezing the meat or fish for at least twelve hours, depending on temperature.

A great photo of a dead parasite, believed to be the Ascaris lumbricoides, or possibly a rope worm in the "seaweed" stage.

A very good look at how long some helminths are. The more worms the people pass, the healthier they get, and the improvements come faster and faster. This child passed a lot of parasites in the beginning. Then after a few months was no longer passing them and now has an ATEC score of 5. Meaning, he no longer has the diagnosis of autism.

This is the parasite that measured 32 inches.

Another long, well washed parasite for the collection. The road to recovery is paved with many dead parasites. Adios Autism…

This is a 32-inch parasite that a young boy passed. He went on to have a great day after getting this out.

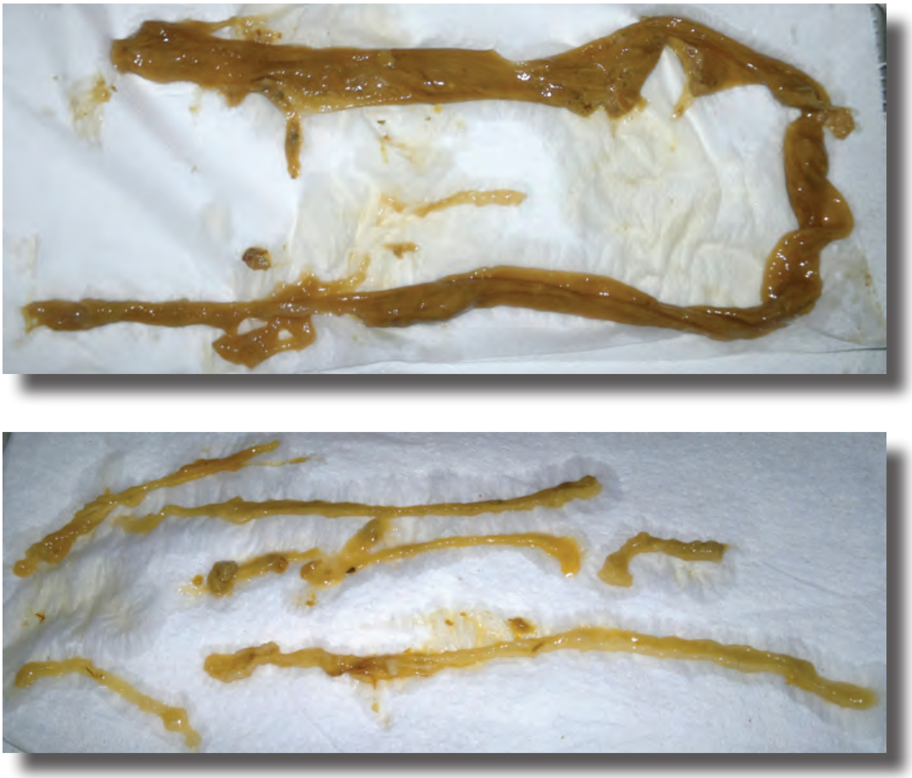

These are rope parasites as confirmed by Dr. Alex Volinsky.

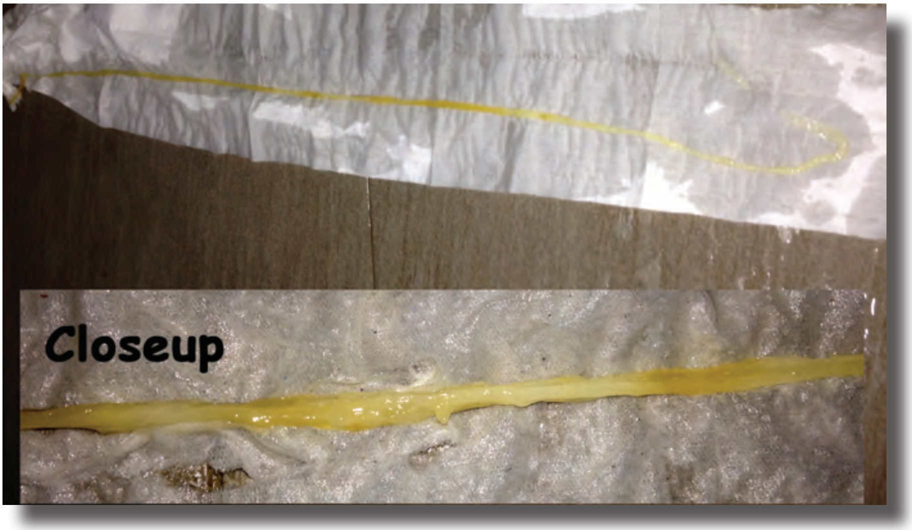

60cm long worm (Oct 1, 2013) from child; 6 months on CD; 2 drops every 1-2hr; no parasite protocol; 1tblsp DE; 2-3 vials of Quinton.

This 33 inch parasite (aka "Chester") was discovered by a woman using the protocol on herself.

The bubble visible in this photo leads us to believe this is a late stage rope worm.

After passing this parasite, the child went on to have a fabulous day at school. It is so great to feel good and healthy.

The FDA recommends freezing and storing [fish] at -4°F (-20°C) or below for seven days (total time), or freezing at -31°F (-35°C) or below until solid and storing at -31°F (-35°C) or below for 15 hours, or freezing at -31°F (-35°C) or below until solid and storing at -4°F (-20°C) or below for 24 hours is sufficient to kill parasites. FDA's Food Code recommends these freezing conditions to retailers who provide fish intended for raw consumption. Note: These conditions may not be suitable for freezing particularly large fish (e.g. thicker than six inches).[2]

▶ From drinking contaminated water.

▶ From consuming contaminated vegetables or fruits. Often we eat poorly washed (parasite infested) vegetables or fruits. There is a common misconception that vegetables from organic farming are free from any problems, pesticides, or chemicals. The danger is that the eggs or larvae of the worms reach the farm soil through animal waste, decomposed forms of natural compost, and manure (fertilizer) added to the field. There are eggs, such as Ascaris lumbricoides, which can survive in soil under extreme temperatures for five years. It is very important to perform a thorough cleaning of fruits and vegetables, and never eat anything raw and straight from the ground, however healthy it may seem.

▶ From parasite infested animals. Parasitic infections are very easy to spread by contact with pets. Veterinarians are quick to insist upon the quarterly deworming of our animals, but there are steps we must take on our own to avoid contamination.

Parasites come in all shapes and sizes. Of course, they can also come out in pieces.

Suggested: Deworm your pet at least every three months for life, as directed by your veterinarian. During the first month, it should be done every week. Prevent pets from eating raw viscera. If animals eat raw meat or raw bones the best option is to freeze the food in advance for at least 12 hours (See citation above). If the deworming treatment is working, the animal will expel the worms in the feces or vomit, which must then be burned or buried, during the eight-day treatment.

Avoid being licked in the mouth by animals as they are in direct contact with feces, soil, and their own anus. When petting an animal, wash your hands with soap and water before eating or handling food, as the eggs of the parasites remain in the animal's hair.

- Do not walk barefoot or with open toe shoes in soil or sand.
- Avoid Hippotherapy (horseback riding)

Symptoms of Parasitic Infections

The different types of worms and toxic waste produced by parasites in our body may cause the following common problems:

Parasites come in all sizes. Some are very long. These get washed with hot water for photos. This way we get the best look at them.

Blood Disorders & Blood Parasite Diseases

Parasites absorb essential nutrients from the body, such as iron, vitamin B_{12}, and sugars, which may result in certain blood disorders. In addition, some diseases are known to be caused by blood parasites:

Blood disorders:

- Anemia
- Dizziness
- Hypoglycemia
- Weakness

Blood parasite diseases:

- African Sleeping Sickness
- Babesiosis
- Chagas Disease
- Malaria

Fatigue:

The toxic waste produced by the parasites themselves (including ammonia and psychoactive substances), can stress the detox organs and cause disorders of the central nervous system such as:

- Chronic fatigue syndrome (CFS)
- Cold in the extremities
- Dizziness
- Extreme weakness
- Internal cold
- Lethargy
- Low energy
- Night waking
- Restless sleep

Gastrointestinal Symptoms

- Abdominal pain or tenderness
- Blood in stool
- Burning in the stomach
- Chronic constipation
- Chronic diarrhea or diarrhea caused by poor absorption of food
- Colitis
- Cramping
- Cravings for greasy foods and sugary foods, lots of carbs and bread, fruit, fruit juices, alcohol, or vinegar
- Digestive problems
- Distended belly
- Eating more than normal but still feeling hungry
- Excessive bowel movements
- Fever
- Frequent vomiting and nausea
- Gas and bloating (noted after eating)
- Hemorrhoids
- Irritable bowel syndrome (IBS)
- Intestinal irritation
- Intestinal obstruction
- Leaky gut
- Malabsorption syndrome
- Mucous in stool
- Pancreatitis
- Passing a worm in stool

Growth Problems, Weight, & Appetite

Parasites usually live without being detected by the host. They rob the body of many of the essential nutrients in the food consumed. Many overweight people, who are infected with parasites, go hungry for lack of essential nutrients, causing them to eat in excess due to their parasitic infection. Furthermore, depending on the type of infestation, many people are malnourished and cannot gain weight. The following is a list of some possible symptoms:

In children:

- Poor growth
- Poor physical and intellectual development consistent with their biological age

In children and adults:

- Chronic burping
- Craving white flour products; cookies, cakes, pastries, etc.
- Feeling hungry after a meal
- Inability to gain or lose weight
- Long-term obesity
- Loss of appetite
- Obsession and/or compulsion to eat sweets or very specific foods (wheat, sugar, dairy)
- Ravenous appetite
- Uncontrollable hunger to eat more than usual
- Weight gain (specifically around the time of the full moon)
- Weight loss

Mood Problems & Anxiety

Toxins that are released by parasites can irritate the central nervous system. Anxiety and nervousness are often caused by parasites that migrate throughout the body. Some of the problems caused are:

- Anger and irritability
- Anxiety
- Confused thinking (brain fog)
- Depression
- Disorientation
- Forgetfulness
- Lack of coordination
- Mood swings

- Nervousness
- Obsession
- Restlessness
- Slow reflexes

Muscle & Joint Pain

Parasites can travel almost anywhere in the body. When they migrate to the joints and the muscles they can cause cysts and inflammation. These can often be mistaken for arthritis and/or muscle pain.

Toxins from parasites can also accumulate in the joints and muscle tissue causing:

- Chest pains
- Fibromyalgia
- Joint pain
- Muscle cramps
- Muscle spasms
- Numbness of the hands or feet
- Pain in the back, thighs, or shoulders
- Pain in the navel
- Rapid heartbeat
- Restless leg syndrome
- Seizures

Parasites in Children (including children with ASDs)

Parasites can be found in the body in asymptomatic and symptomatic stages. The former are usually found in adults. Symptomatic stages occur mainly in children, in whom we can often observe the following:

- Anorexia
- Anxiety
- Bruxism (teeth grinding)
- Cramping
- Diarrhea that alternates with periods of constipation
- Excessive Flapping
- Growth retardation
- Headaches
- Inability to gain weight
- Itching/Burning/Picking of the anus
- Nasal itching and/or anal urticaria (hives/rash)
- Nervousness and irritability

- Nose Picking
- OCD (Obsessive Compulsive Disorder)
- Rage
- Smearing feces
- Unexplained laughter or weeping
- Verbal stims
- Weight loss

Tapeworms, and some other parasites, have an affinity for B_{12} and iron. Therefore, lab results that show deficiencies in B_{12} and/or iron can be indicators of parasitic infections.[3,4] Due to its size, the tapeworm consumes enormous amounts of food that it obtains by taking the child's food. This can affect the child's normal development.

Treatment is simple, but it requires that the head of the tapeworm be removed, otherwise it will continue to grow. Tapeworm treatment is separate from this protocol and usually requires niclosamide. However, the only way to be sure the head has been removed is to identify it in the stool.

Respiratory Disease

The passage of larvae through the respiratory system or larval invasion in the lungs may cause symptoms such as:

- Acute bronchitis
- Asthma
- Drowsiness
- Dyspnea (shortness of breath; air hunger)
- Chronic/irritative cough
- Pneumonia
- Shortness of breath or respiratory failure

Sexual & Reproductive Disorders

Immune dysfunction as a result of a parasitic infection can lead to:

- Candida - yeast infections
- Cysts and fibroids
- Erectile dysfunction
- Fluid retention
- Male impotence
- Menstrual problems
- Premenstrual syndrome
- Prostate problems
- Urinary Tract Infections

Skin Disorders & Allergies

External parasites (lice, bedbugs, scabies, etc.) that penetrate the skin can cause itching, redness, and/or rashes etc. However, internal parasites can be responsible for skin disorders as well. Parasites create toxic metabolic waste, and because the skin is the largest organ, the body tries to eliminate them through it, resulting in many skin problems.

Some symptoms may include:

- Allergies (to foods, dust, mold, etc.)
- Anal itching
- Brittle hair
- Crawling sensation under the skin
- Dermatitis
- Dry hair
- Dry skin
- Eczema
- Eruptions
- Hair loss
- Itchy nose
- Itchy skin
- Jaundice
- Psoriasis
- Skin ulcers
- Sores
- Swelling
- Urticaria (hives; skin rash)

Sleep Disorders

The body reacts to parasites during rest periods because at night is when parasites are most active. Nocturnal awakenings are common, especially between 2 and 3am, when the liver tries to rid the body of toxins produced by parasites. This in turn may produce:

- Insomnia
- Teeth grinding
- Bedwetting
- Drooling while sleeping
- Sleep disturbances - multiple awakenings during the night
- Restless sleep

Other Problems Associated with Parasites

- Bad breath
- Blurred vision
- Body odor
- Breathing problems
- Chronic infections: viral or bacterial
- Circulatory problems, numbness in the extremities, difficulty in moving
- Cough or coughing up blood
- Difficulty swallowing
- Excessive salivation
- Fever
- Fluid build-up or retention during the time of the full moon.
- Low immune response
- Peritonitis
- Sensation of a foreign body or discomfort in the throat
- Swollen eyes
- Weight gain during the full moon.

Blood Analysis

The following markers may be present when a person is suffering from a parasitic infection or the resulting allergies:

- Anemia/low iron
- Elevated immunoglobulin (IgE)
- Elevated eosinophils (The eosinophil is a specialized cell of the immune system, more specifically it is a proinflammatory white blood cell. According to the Registry for Eosinophilic Gastrointestinal Disorders (REGID), their known functions include movement to inflamed areas, trapping substances, killing cells, antiparasitic, and bactericidal activity, participating in allergic reactions, and modulating inflammatory responses.)
- High ammonia
- High oxalates
- Low vitamin B_{12}

Measures in the Home Environment to Prevent Reinfection

It is important to treat all people and pets that live in the same environment to prevent someone from infecting others. Reinfection occurs via underwear, bedding, towels and household items such as children's toys or animals that have been in contact with eggs. It is important to wash all clothing that has contact with intimate body areas at a temperature not below 60°C (140°F).

All bed linen and underwear must be washed daily (or to the extent possible) while performing anti-parasitic therapy. Affected individuals should not share their swimwear with other members of the family, and should use a separate cloth to wash his anal area. It is best to sleep wearing both underwear and pants to avoid involuntary scratching during the night. This will prevent infection through anal-hand-mouth contact because by scratching the anal area, eggs can become lodged under the fingernails. Keep pets away from the place of rest of their owners, such as beds, sofas, blankets, and cushions.

Thoroughly wash fruits and vegetables in water and soak them in CD or CDS solution for a few minutes. Clean the sink with alcohol, as the eggs of many parasites are immune to the pH of normal cleaning products such as soap or bleach. It is important to note that parasites do not leave any kind of immunity behind in the host, therefore, once cured, the person who has suffered can suffer from them again. The only surefire method of killing the eggs of *Ascaris lumbricoides* is in water above 60°C (140°F) or with 96% grain alcohol (Everclear).

Evolutionary Cycles of Intestinal Parasites

Although there are many more, here is a description of the three most common types of intestinal parasites that can be found in developed countries:

Ascaris Lumbricoides **(Roundworm)**

Ascaris reproduce quickly, as a single female can lay up to 200,000 eggs each day. This parasite is very common, especially in damp conditions, and when hygiene measures are inadequate. It can affect the entire population, but mostly affects children, seriously disrupting their development and growth. It's so infectious that the WHO estimates that there are about 700,000,000 people infected worldwide, of which around 60,000 cases end in death per year, mainly children.[5]

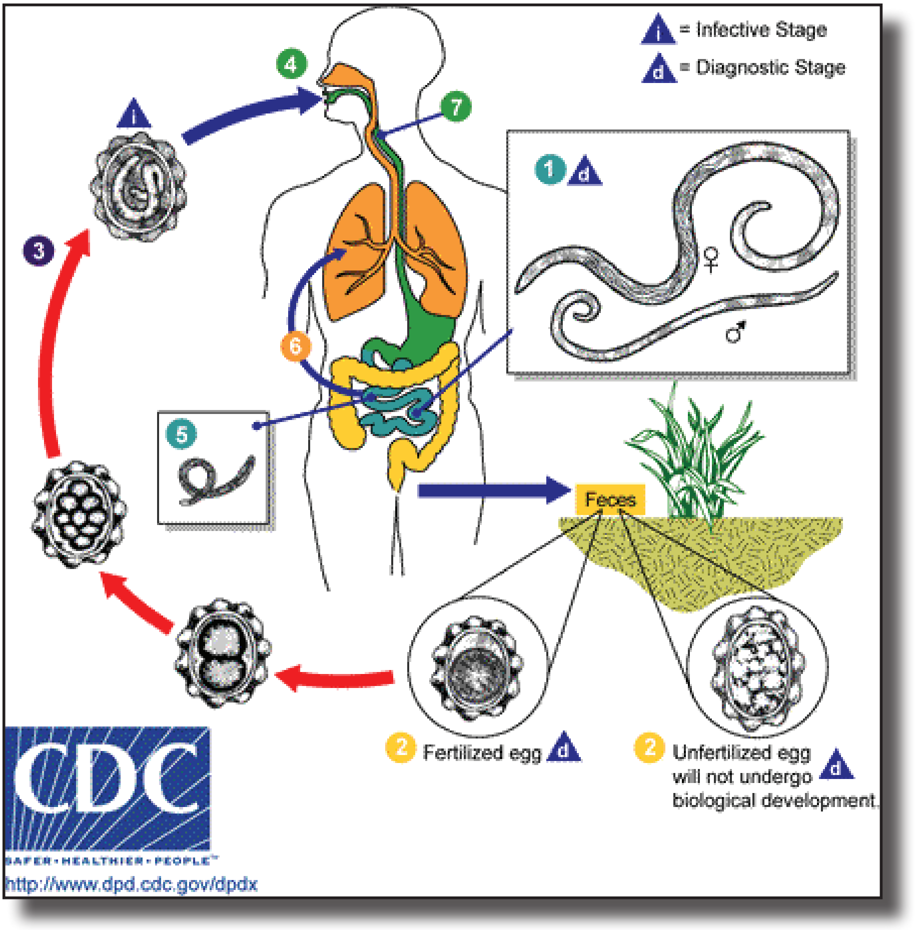
i = Infective Stage
d = Diagnostic Stage
4
7
i
1 d
♀
♂
3
6
5
Feces
2 Fertilized egg d
2 Unfertilized egg will not undergo biological development. d
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx

Parasite eggs reach the duodenum through the mouth of the host. Gastric juices rid the eggs of their shells and release the larvae. These larvae, which are highly mobile, penetrate the duodenal mucosa and migrate to the liver. From there, they continue their migration to the heart, reaching the lungs through pulmonary circulation, and finally become trapped in the pulmonary capillaries. Here, the larvae break the thin walls of the capillaries and penetrate the alveoli, bronchioles, and bronchi at which point they are able to travel up through the pharynx. Once the eggs pass the epiglottis (see diagram above), the larvae are swallowed such that they return to the duodenum, where they complete their maturation process. This process takes approximately two to three months to complete; therefore we calculate that to find ourselves completely parasite free, the initial treatment should be done for at least 12-18 months, possibly longer. From there on out, you may only need to follow a routine deworming two to four times a year.

Keep in mind that eggs are expelled through the feces (human or animal) into the environment where they can survive even in harsh conditions, favoring the persistence of the parasite. They are resistant to low temperatures, desiccation, strong acids, soaps, chlorine, formaldehyde (pH between 2 and 11.5), and can live in planted soils for five or more years, creating a "parasite hotbed" that makes them almost indestructible. Once dry, they are transported through the air, fly in air currents like dust that can be inhaled and/or swallowed. We have recovered eggs from nasal mucus, paper currency, potting soil, dust, and in indoor airborne particles, etc.

***Taenia Saginata* and *Taenia Solium* (Tapeworm)**

Taenias reach humans when humans eat their eggs, by the consumption of tissue from infected cattle or pigs. In infected animals, the larvae are encysted in the muscle tissue. If the infected animal is consumed, development may proceed in the human digestive system. Humans are hosts for *T. saginata* (beef tapeworm) and *T. solium* (pork tapeworm). The tapeworm is considered to be solitary, because usually no more than four species are found in any one individual host. The danger of this parasite is that the larvae can migrate to the brain, or other vital organs (cysticercosis). Tapeworms may be detected by identifying segments in the host's stool that the worms discard as they grow. However, tapeworms may go undetected for many years, living asymptomatically within their host.

According to classification, they can vary in size ranging from 2 to 12 meters in length. They consist of a head called the scolex, which attaches to the intestine by means of suction cups, and a body consisting of repeating units called proglottids. A single Taenia can grow from 1,000 to 2,000 proglottids, depending on the type. A tapeworm produces an average of 720,000 eggs per day.

Pinworms (*Enterobius Vermicularis*)

Humans are considered the only host of what are commonly called pinworms (*Enterobius vermicularis*). This type of worm is the most typical in the family because it propagates easily. It is common for children to become reinfected in schools, through contact with others, or through anal-hand-mouth contact. Pinworms have an elongated shape, are whitish in color, and are about 1cm long. They inhabit the large intestine of humans. Female pinworms leave eggs around the anus. Once deposited, the eggs are infectious for a period of up to 20 days. Once in the intestine, it takes between five and eight weeks to develop into adult worms. The most important symptom is intense itching that occurs in the anal area, especially at night. In women, inflammation of the vulvar area is very common.

A pinworm infection is, generally speaking, not very serious. Unlike other parasites, they infect only humans. Transmission from person-to-person happens by handling clothing, bed sheets, towels, and environmental surfaces (such as curtains, carpeting) contaminated with pinworm eggs, which are so light that they are able to become airborne. A small number of eggs can be integrated into air particles that when inhaled follow the same developmental process as ingested eggs. Enemas are extremely useful in removing this parasite from the large intestine.

Graham's method is a simple method of detection. Just after waking and before a bowel movement, press a piece of tape against the anal folds. The tape will catch the remains of eggs and/or parasites that are situated there. With the naked eye we can see small worms no more than an inch long, but with a microscope, many transparent eggs from females and even other species may be seen.

The Importance of Lifelong Deworming

Once we begin the process of deworming, we should recognize that we must maintain this habit of cleaning for the rest of our lives to enjoy good health. It is common among people who have pets, to follow the recommendations of their veterinarians, and deworm their pets every three months. It is interesting to ask why family physicians do not give the same advice to humans. Perhaps some physicians ignore this information, or simply do not consider it important to eliminate these pests, which are just as harmful to people as they are for animals.

It is true that many parasites are not endemic or common outside certain climates, but human migration and global marketing of food products have facilitated the spread of many parasitic pests silently. It is important that we understand the lifecycle of each parasite, from birth to death including reproductive and death stages. This information is crucial for the complete elimination of the parasite. For example, in the case of intestinal parasites treated here with this protocol, some can live in the host for up to ten years, as in the case of a single Taenia, while others may remain in the host for a lifetime, reproducing again and again, as in the case of the pinworms or the well-known Ascaris.

The Kalcker Parasite Protocol & Lunar Cycle Timing

In the modern civilization in which we live, we have lost touch with much of the ancient wisdom of the past. One of the things we have forgotten is how the natural cycle of the moon influences many of nature's routines. This is especially true for the behavior of parasites. They are known to sync their life-cycle with that of the lunar cycle. Your child may demonstrate extreme behaviors on certain lunar cycle days… especially on the full moon and sometimes even the new moon.

Therefore, to maximize effectiveness, this Parasite Protocol is specifically timed to the lunar cycle. Appendix 10, page 477 provides an easy reference for you to look up the days to perform the Parasite Protocol, which is administered over 19 days—numbered 0 to 18—each month; starting three days prior to the full moon and continue during the waning moon. This period of the moon's cycle is very effective for deworming because many nematodes (parasitic worms) travel back into the intestine to mate at this time.

Length of Treatment

This protocol is not a one-time treatment. You should plan for at least 12 to 18 months to insure you have purged multiple parasite life-cycles and continue beyond 12 months if your child is still expelling parasites.

Building on CD

This protocol builds on what you have already learned using CD. During treatment, it is absolutely necessary to continue CD dosing, CD baths and CD enemas.

Tape Worms

This protocol is specifically designed for the deworming of large intestinal parasites, especially round nematodes such as Ascaris. It is effective for most nematodes, but may not be effective against tapeworms. In the case of infestation by Taenias, the recommended treatment is Niclosamide, the preferred medication due to its low toxicity.

Components of the Kalcker Parasite Protocol

This protocol uses some of what you have already learned and should already be doing, along with a set of ingredients you will need to have on hand before you start (shown in bold below). Here is an overview list of the items you will need to have on hand, and the items are described in detail on the pages that follow:

- Meal time (1, 2, 3)
- CD / CDS (4-19)
- CD Baths (20)
- CD Enemas (21, 22)
- Ocean Water (23, 24, 25) (see page 115)
- **Diatomaceous Earth** (26, 27)
- **Lepidium Latifolium Extract (Rompepiedras)** or **Chanca Piedra (Stone Breaker)** (28, 29)
- **Pyrantel Pamoate (Combantrin®)** (30, 31)
- **Mebendazole** (32-36)
- **Castor Oil** (37)
- **Neem** (39, 40)
- **Probiotic (usually THERALAC®)** (41)

Check the following website for the latest information on where to find these products:

www.ProtocolSuppliers.com

You may have noticed one or more numbers in parenthesis following each of the previous items, such as "(40)" for the probiotic. The timing of what to give when, is covered in great detail in the daily charts starting on page 198. These numbers match those found on the daily sample charts to make it easier for you to connect the dots and also be able to identify related notes under each chart. They have nothing to do with the quantity/dosing of any substance. The use of these numbers allows us to comment on specific items and when and where they come into play as shown on the daily sample charts.

We now detail each one of the items on the list above and discuss what you need to know about them and how to acquire them (including their associated numbers on the charts).

Meals (1, 2, 3)

Obviously meals are part of everyone's day. The purpose in mentioning them here is that many of the following steps are related to meal timing. Some actions or ingredients are taken before breakfast, others during or after.

In our example charts, we make some assumptions which include:

- Breakfast (1) is at 7:30am
- Lunch (2) is at 12 noon
- Dinner (3) is at 5pm

CD, CDS or CDH (4-19)

As you learned in Chapter 5, you continue to dose CD (or CDS/CDH) as before during the Parasite Protocol. Nothing changes in that regard with the Parasite Protocol building on top of those steps already in place.

The sample charts assume your child is going to school, and so you may not be able to administer doses of CD during his school day (7-12) unless you are home-schooling, in which case you are encouraged to give hourly CD doses, even if the total exceeds 8 for the day.

CD Baths (Optional) (20)

Our sample charts assume you administer a CD bath just before bedtime. See pages 113 for more information about CD baths.

CD Enemas (21, 22)

Ideally, give your child a CD enema in the morning (21) and another one in the evening (22). However, if your child is going to school, a morning (21) enema may not be a good idea due to the possibility of an “accident.” Therefore, consider the morning enema optional, but the night time enema a must do! See page 103 for more information.

Ocean Water (23, 24, 25)

Supplementing ocean water minerals is important to support the body through the detoxification process. See page 115 for more information on ocean water.

Dosing:

(23) A dose of ocean water should be administered upon waking, but five minutes apart from CD dosing.

(24) One dose of ocean water immediately after school (or at lunch time if at home).

(25) One dose of ocean water, 15 minutes before or after dinner.

Diatomaceous Earth (DE) - Food Grade (26, 27)

Diatoms are unicellular plants that existed by the trillions in our oceans over 300 million years ago. They are encased by a cell wall that is made of silica. When diatoms die, this microscopic coating deposits at the bottom of oceans. Over time, they pile up in banks forming deposits thousands of meters in size. With the receding of the oceans, these deposits have been uncovered. Through compression, and ultimately fossilization, these silica deposits have given rise to a chalk rock called diatomaceous earth.

DE is an inert, nontoxic compound, which contains a number of minerals such as manganese, magnesium, iron, titanium, calcium silicates, and others. Properly ground, the skeletons of microscopic diatoms become sharp silica needles, harmful to parasites, fungi, yeast, worms, and amoebas. However, these needles are harmless to humans and other warm-blooded animals. Although it is safe to consume diatomaceous earth continuously, the best method (as with everything else) is to allow for periods of rest. During the 18-day treatment, take two teaspoons (5ml) twice a day.

Dosing: ½ to 1 teaspoon twice a day for smaller kids, 1 teaspoon three times a day for adults and bigger kids. Mix with a little water and drink. Given on days 1 through 18. DE mixes well with water but never dissolves. Stir the DE/water slurry vigorously and drink immediately before the DE settles to the bottom. Some people take heaping tablespoons in water, but larger amounts are not necessary. DO NOT take dry!

Note: In rare cases, DE can cause constipation, which can usually be managed by reducing the dosages to 1/2, 1/4 or even 1/8th of a teaspoon. If that doesn't resolve the issue, remove DE from the protocol and continue with all other directions.

Source & Cost: Search online for "Food Grade Diatomaceous Earth." Buy at least 1 pound and expect to pay about $20 more or less. Better yet, buy a five pound bag, which will reduce your cost per pound. DE does not expire or degrade, but should be kept in a dry container. Note: Diatomaceous Earth is often used as non-toxic element in pool filtration systems. You DO NOT want to use this kind of DE since it has been processed. Only get "Food Grade" Diatomaceous Earth!

Lepidium Latifolium Extract (aka Rompepiedras or Pepperwort) or Chanca Piedra (aka Stone Breaker) (28, 29)

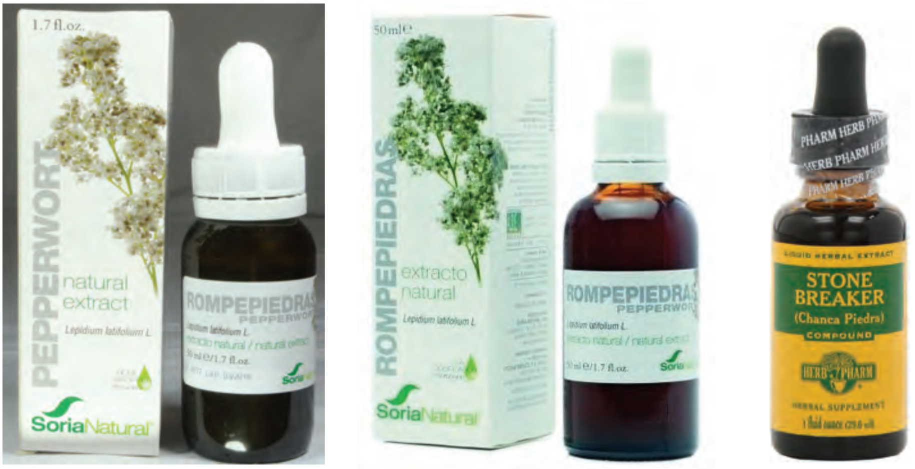
1.7 fl.oz.
PEPPERWORT
natural extract
Lepidium latifolium L.
SoriaNatural
ROMPEPIEDRAS
PEPPERWORT
50ml℮
ROMPEPIEDRAS
extracto natural
Lepidium latifolium L.
SoriaNatural
ROMPEPIEDRAS
PEPPERWORT
50 ml ℮/1.7 fl.oz.
HERB PHARM
LIQUID HERBAL EXTRACT
STONE BREAKER
(Chanca Piedra)
COMPOUND
HERB PHARM
HERBAL SUPPLEMENT

Lepidium Latifolium (Rompepiedras) and Chanca Piedra (Stone Breaker) both break up hard substances in the body. The reason we use it in this protocol is two-fold: It breaks up the protective outer coating of parasites, and annihilates oxalates, which many of our kids have an abundance of because oxalates are produced by parasites.

Note #1: It is often simply referred to as "RP" in our discussion forums.

Note #2: If you are having trouble finding this particular ingredient, don't let its absence stop you from starting the protocol with all other ingredients.

Dosing: 15 drops for a 100 pound child mixed in with the diatomaceous earth/water. Use seven drops for smaller children.

Source & Cost: You have a choice of two herb extracts; Lepidium Latifolium Extract (aka Rompepiedras or Pepperwort) and Chanca Piedra (aka Stone Breaker). One US source is www.mightyguts.com, which sells a 50ml dropper bottle of Pepperwort for about $30. A primary manufacturer in Europe is Soria Natural from Spain that labels their product Rompepiedras, while they also have an English labeled box that says Pepperwort. They both show Lepidium Latifolium on the box.

Pyrantel Pamoate (30, 31)
(Trilombrin/Combantrin®)

Pyrantel pamoate is a broad spectrum anthelmintic, which works by causing a neuromuscular block that produces spastic paralysis of the parasite, and its subsequent expulsion by intestinal peristaltic action, without excitation of the parasites or encouragement of their migration. Pyrantel pamoate acts over a short duration, and tends to be completely eliminated from the body in the feces and urine within three to four days. Pyrantel pamoate is poorly absorbed from the gastrointestinal tract, and approximately 6 to 8% total is found in the urine, with the remainder in the feces. The recommended dose is one daily dose of 10 mg per kilo.

Pyrantel pamoate is incompatible with the use of piperazine, because the two substances neutralize each other. Thus Pyrantel pamoate should not be combined with pumpkin seeds, which contain piperazine, or with antiparasitic drugs that contain piperazine in their formulation.

Dosing: Pyrantel pamoate is given only twice during one cycle of the parasite protocol; once during breakfast on day one (30), and again during breakfast on day five (31). Dose is based on weight and calculated by multiplying your child's weight in kilograms times 10mg of pyrantel pamoate. To make it easy, refer to the following chart:

Pyrantel Pamoate (Trilobrin/Combantrin®) Dosing by Weight		
Pounds	**Kilograms**	**Dose in mg.**
20	9	91
40	18	181
60	27	272
80	36	363
100	45	454
120	54	544
140	64	635
160	73	726
180	82	816
200	91	907
220	100	998
240	109	1089

Pyrantel pamoate (often just referred to as Combantrin®) is available in three forms:

- **Liquid:** where each milliliter contains a certain number of milligrams. For example, one available formulation contains 144mg/ml. So a 100lb child would take 3ml.
- **Tablets:** where each usually contains 250mg.
- **Capsules:** where each usually contains 250mg.

You will have to read the label or the particular product you acquire and determine the milligrams to use.

Source: Combantrin® is available by prescription in the US. Most other countries have it available over the counter.

The preferred source of pyrantel pamoate is a compounding pharmacy, as to avoid coloring and flavoring. If you are unable to find it without coloring/flavoring then I would personally use mebendazole for the entire 18 days rather than risk giving your child an ingredient which may cause regression.

Note #1: Some brands of Combantrin® include mebendazole. You want the stand-alone Combantrin®!

Note: #2 Pumpkin seeds should not be consumed with pyrantel pamoate because they neutralize its effects.

Mebendazole (Vermox®/Lomper®) (32-36)

Mebendazole is a drug used in treating diseases caused by helminths (parasites of the gastrointestinal tract). This drug prevents the parasite from using glucose, which results in a decrease in energy and therefore death of the parasite.

Mebendazole is a non-systemic drug which means it is only absorbed, to a limited extent, in the gastrointestinal tract (approximately 5 to 10%). However, if it is consumed with fatty foods then more absorption occurs.

Approximately 2% of the administered mebendazole is excreted in the urine, while the remainder is excreted in the feces. The appropriate dose of mebendazole may be different for each patient as it depends on the type of parasite causing the infection.[6] The most frequently recommended dose is 100mg for children, 200 mg for adults, two times a day for seven of the first nine days of the Protocol.

Adverse effects from mebendazole are generally rare due to its poor absorption. However, it may cause nausea, vomiting, abdominal pain, and diarrhea. Normally these effects are in fact a result of the release of toxins from the very death of the parasite itself. Anti-parasitic drugs can be administered very effectively by diluting them in water, putting the mixture in a small bulb enema, and inserting the anally. This is especially suitable in the case of oxyuriasis (pinworms). Read more about this "implant method" on page 220.

Dosing:

Note: Indicated weights should only be considered a rough guide.

Small Children (20-40lbs.): Days 2, 3, 4, 6, 7, 8 & 9 — Take as little as 25mg of mebendazole with breakfast and dinner, and do NOT do a blitz on Day 9.

Children (41-70lbs.): Days 2, 3, 4, 6, 7, & 8 — 50mg with breakfast and dinner. On Day 9 they can take 50mg, 50mg and 25mg.

Adolescents (71-100lbs.): Days 2, 3, 4, 6, 7, & 8 — 100mg during breakfast and 100mg during dinner. Day 9 is "Mebendazole Blitz Day" where you administer ONE 200mg dose during breakfast; ONE 200mg dose at lunch; and a final 100mg dose at dinner.

Teens & Adults (101 lbs. and up): Days 2, 3, 4, 6, 7, & 8 — 200mg during breakfast and 200mg during dinner. Day 9 is "Mebendazole Blitz Day" where you administer ONE 500mg dose during breakfast and no dose at lunch or dinner for the remainder of the current cycle.

Source: Mebendazole is available by prescription in the US and over the counter in other countries.

Note #1: Some brands combine mebendazole with Combantrin®. You want the stand-alone mebendazole!

Note #2: Mebendazole is mostly sold in tablet form, but it is also available in liquid. DO NOT buy the liquid form—stick with the tablets! I have seen horrible reactions from the "inert" vehicles used in the liquid products.

Castor Oil (37)

Castor oil is extracted from the seed of a plant called Ricinus communis ("Higuera del diablo"). Its seeds contain between 50-80% oil, which has a high content of ricinoleic acid, which has excellent laxative and purgative properties. Once you begin anti-parasitic treatments, spastic paralysis may occur in some parasites and many together may form a "knot" of worms that can cause intestinal obstruction. It is important to help your body

Apollo
CASTOR
OIL

purge them by using castor oil. Castor oil should be taken in the morning, two hours after breakfast and other medications. If your child goes to school they take it as soon as they walk in the door from school. A typical dose for a child is 1/2 tsp to 1 tsp, or up to tolerance. The adult dosage is 15 to 30ml (two tablespoons), two hours after breakfast and other medications. If you experience any intestinal distress, mineral purgatives such as Epsom salts, or vegetable purgatives such as senna leaves, can be used.

Castor oil is also available in gelcaps for those who dislike the taste.

Dosing: The amount to administer varies and really depends on the individual's tolerance. A good starting point is ½ teaspoon for smaller children and up to two tablespoons for larger children and adults. Only experimentation will determine the right amount, if castor oil causes diarrhea.

Source & Cost: Readily available in liquid form at most pharmacies in the laxative section. Usually under $10 for 16oz. Also available online.

Neem (Azadirachta indica), Caps or Tea (38, 39)

PREMIUM
Nature's Way
HERBAL
Neem
Leaves
100 Capsules / Certified 475 mg
DIETARY SUPPLEMENT
Recommendation: Take 2 capsules three times daily, preferably with food.
Supplement Facts
Serving Size 2 Capsules
Servings Per Container 50
Amount Per Serving
% Daily Value
Total Carbohydrate
1 g
<1%†
Neem (leaf)
950 mg
Other ingredients: Gelatin (capsule)
Actual Capsule Size

The neem tree is a great natural inheritance of mankind. References in Sanskrit scriptures and ayurvic medicine practices indicate the use of neem since ancient times in Hindu medicine. Even today, Hindus living in rural areas call the neem tree the "village pharmacy" for its ability to alleviate many diseases and is currently endorsed by authorities in India for its use in medicinal preparations. Neem is one of the purifying and detoxifying plants with the greatest potential. Neem has been used to combat all forms of body parasites, external and internal parasites alike. To prepare neem, boil four leaves (normally the contents of an envelope) in one liter of water for five minutes. Drink the tea throughout the day over the course of each parasite protocol.

Dosing: Neem is given each parasite protocol from day 10 through Day 18.

You have a choice of caps or tea. I prefer caps over tea because the taste is strong and unpleasant, so some kids will buck drinking tea.

Caps (assuming 475mg each): An adult takes six in a day, three times two caps at meal times.

Follow the directions on the bottle. Give a full dose for teens and adults 100 lbs. and over. Small children receive ¼ to ½ dose.

If using caps, give one dose during breakfast and one dose at dinner.

Tea: Give four doses throughout the day. Prepare a tea from the leaves, one tea bag in one liter of water (add stevia if needed to cover some of the bitter taste). One tea bag usually contains approximately four leaves. If using loose leaves, then make one liter of tea with four neem leaves. If using crushed leaves, then use approximately one slightly heaping teaspoon.

Neem Tea Dosing

Weight of Person	Total Daily Amount (To be split into 4 doses)
20-34 lbs	100 ml
35-49 lbs	200 ml
50-64	300 ml
65-84	400 ml
85-109	500 ml
110 and over	600 ml

Source & Cost: Neem capsules cost about $8 for 100x 475mg caps. Look in your local health food store. One popular brand in the US is Nature's Way®. BioPure™ sells a product called *Neem Synergy* which contains a few additional herbs that do not affect the neem.

Neem tea can be purchased in bags or as dried leaves.

THERALAC® (Probiotic) (40)

THERALAC
PROBIOTIC MASTER SUPPLEMENT
5+2
BIOTHERAPY
30 CAPSULES

TruFlora
Probiotics + Enzymes
ACID-PROOF

THERALAC® is a probiotic that should be given during the parasite protocol to help reestablish good gut flora. Ideally, it is rotated every other month with THERALAC® TruFlora®.

The key reason for recommending THERALAC® over other probiotics is explained in this paragraph excerpted from their website:

> *THERALAC® probiotics survive transit through the harsh acidity of the stomach and arrive alive in the intestinal tract. THERALAC's ACID PROOF™ technology utilizes sodium alginate from seaweed in a unique formulation that assures survival at pH 1.6 for 90 minutes, the most severe stomach acid conditions normally experienced. Other probiotics that claim acid resistance are tested at pH 2.5 – 3.0, or >10 times less acidic than pH 1.6 – not a fair test. THERALAC's ACID PROOF® technology is protected by US Patents 7,122,370 and 7,229,818. This technology goes beyond resisting stomach acid and involves keeping the probiotic cells together in a viscous alginate-gel moving in group-force, not as separate individual cells like other probiotics, deep into the intestinal tract while retaining key formula ingredients, LactoStim® and Sodium Alginate in close proximity.*

Dosing: One capsule of THERALAC® is given each and every night at bedtime, irrelevent of age/weight. This probiotic is also to be given outside of the 19 Parasite Protocol days.

THERALAC® can be given at the same time as CD if your child swallows capsules. However, if you are using the THERALAC® powder form then give it at least five to ten minutes after the last CD dose of the night. See their website for more information about the powder form.

Special exception: Some people do not tolerate probiotics, in which case we have no choice but to leave them out. In some of these cases, sauerkraut and fermented veggies may help to cultivate beneficial bacteria.

Source & Cost: Amazon carries "THERALAC® 30 caps by Master Supplements Inc." for $37.

Sample Parasite Protocol Daily Calendar

To assist you in understanding the Parasite Protocol and how it changes from day to day, we have prepared the following visual set of daily charts showing how all the pieces fit together. To prepare this set of charts, we made up the following times for example purposes only:

- Your child's awake time is between 6am and 10pm.
- Breakfast is at 7:30am
- Your child goes to school, which starts at about 9am
- School is out at about 2:30pm
- Dinner is at 5pm

As stated earlier, each dose or activity is labeled with a unique number. Those numbers relate to notes below each chart AND they match longer descriptive notes in the previous pages.

You are encouraged to copy/enlarge the blank chart in Appendix 9 (page 475) and workout your real-life schedule based on these example charts.

Refer to the lunar calendar (Appendix 10, page 477) for the exact days the protocol should be followed, as well as to get an idea of behaviors related to parasites around the full and new moon.

> After four years of biomed and no major gains, we decided to try mms/cd protocol. We started in April and are currently on our second parasite protocol. Initially, we were really intimidated by the enema part but realized if we were apprehensive our 7 yr old daughter would pick up on that. We told her that what we were about to do would help her feel better we gave her the iPad right before the procedure and a new fav toy (beanie boos) and she was game. After it was over she said "no more booboo in my tummy". We all cried. To make a three-month adventure short and sweet, her team at school cannot figure out how her fragmented speech has gone from 2-3 word utterances to long, drawn out sentences. Her expressive language and receptive have increased. Her auditory processing speed has quadrupled. The most amazing part thus far has been the increase in social skills. Or just the fact that she is interested in engaging. Socially she was nonexistent and is now at a four year old level. We just spent a week with family and she was saying hi and engaging without being prompted!!!! She now has a million questions, wants to know when and where we are going and the sequence in which events will take place. All this from a girl who never asked questions before mms! STOP thinking about trying! Just Do It!! For the first time in seven years, I can hear the bells of freedom ringing!!!! Thank you Kerri Rivera !!!!!

Day 0 (3 days before Full-Moon)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1											
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3											
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 0 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
40	Probiotic at the end of the day

Day 1 (2 days before Full-Moon)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28	30				
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29					
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 1 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
30	Dose of Combantrin with breakfast
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 2 (1 day before Full-Moon)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 2 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
40	Probiotic at the end of the day

Day 3 (Full-Moon)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 3 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 4

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 4 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
40	Probiotic at the end of the day

Day 5

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28	31				
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29					
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 5 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
31	Dose of Combantrin with breakfast (NO Mebendazole today!)
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 6

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 6 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
40	Probiotic at the end of the day

Day 7

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 7 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 8

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 8 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
40	Probiotic at the end of the day

Day 9 (Mebendazole Blitz Day)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		34			
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM									35			
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		36			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 9 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
34	Blitz Day - Small Child Dose 1: 200mg Mebendazole with breakfast. Teen/Adult Dose: 500mg Mebendazole with breakfast ONLY!
35	Blitz Day - Dose 2: 200mg Mebendazole with lunch (or right after school) Teen/Adult Dose: N/A
36	Blitz Day - Dose 3: 100mg Mebendazole with dinner. Teen/Adult Dose: N/A
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 10

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 10 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Start dosing Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 11

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 11 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 12

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 12 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 13

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 13 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 14

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 14 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 15

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 15 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 16

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 16 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 17

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 17 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 18

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 18 (Last Day!) - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Parasite Protocol *Off Days*

Days "19" through to the next "Day 0" are "Off Days" where you discontinue parasite meds and herbs. Here's a simple chart showing what to continue doing and what to stop during this *off* time:

Continue with these:
CD / CDS / CDH Dosing
CD / CDS / CDH Baths
CD / CDS / CDH Enemas
Ocean Water
Probiotic

Stop these:
Diatomaceous Earth
Lepidium Latifolium
Pyrantel Pamoate (Combantrin®)
Mebendazol
Neem
Castor Oil

Of course this represents a hypothetical case and your situation may call for taking other meds or supplements.

Detection of Parasites in Stool

It is necessary to detect the parasites by observing the stool carefully. For that we use a small plastic basin, and a plastic stick or fork for examination.

Author's Note: One of the moms who is a part of our forum came up with some guidelines for processing your child's stool for parasite identification. She calls it, "Everything You Wanted to Know About Sorting Through Poop." Here are her suggestions:

Supplies:

- rubber gloves
- paper plates
- plastic forks (plastic sticks, chop sticks, or plastic back scratcher)
- a pen
- a coin
- toilet hat

I like to use a plastic "toilet hat" also known as a specimen collector or a specimen collection unit, which goes under the toilet seat and collects the stool before it sinks into the bottom of the toilet (available on Amazon.com).

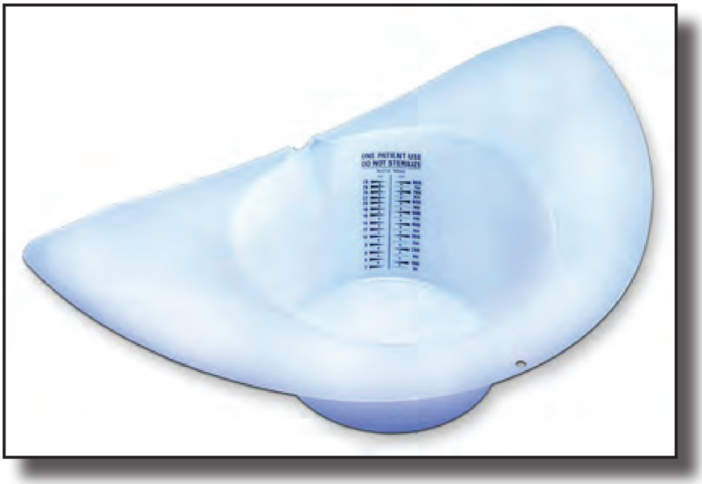

The Collection:

When your child poops, it is collected in the toilet hat (specimen collector). After I get my child cleaned up and taken care of, I remove the toilet hat from the toilet and put the specimen on a paper plate with my plastic fork. I have a look for anything interesting and then transfer that part to a clean plate using my trusty plastic fork. I discard the remainder of the specimen into the toilet, flush and put that dirty paper plate in the bathroom garbage. (We now line with kitchen garbage bags, and I change it after each of these poops.) On the clean plate with the suspected worm, I may add a bit of water and swish it around to get the worm cleaner. I then may transfer the worm to a third plastic plate to get a clear picture. On the clean plate with the washed worm, I write the date, and the initials of the person the worm came from. If you need help identifying the worm, place a penny next to the worm (for size context), snap a picture, and mail the image to kerri@cdautism.org, (Kerri collects the photos for documentation purposes, so send those worm pictures). Then flush the worm and put all paper plates, gloves, and fork in garbage and take it outside. Now, you can go find out what your child has gotten into while you were doing all this.

Clean Up:

Use HOT water (60°C/140°F), and sterilize with 96% (180 proof) grain alcohol (Everclear).

Microscope

It is very useful to have a microscope for diagnosis because it enables you to see both the small parasites that may appear in the blood, as well as eggs or larvae in the feces. This way we are more accurately able to determine if the number of parasites decreases. A simple microscope that costs about $100 is suitable for this kind of identification. The easiest method to determine what you see is to compare your sample to images you find on Google. This way you can enlarge the images, see various samples from different angles, and get a much wider variety of samples than if you were comparing a sample with many textbooks.

Bulb Enemas (aka "Implant")

To prevent anal itching from pinworms, night wakings, etc., you can use a bulb enema or small catheter/syringe with a dilution of 50mg of mebendazole in 10 to 15ml of water for small children, or 100mg of mebendazole in 15 to 20ml of water for larger children/teens/adults. The best way to do this is to introduce the medication together with the water in the rectum immediately before bedtime and hold overnight. If you are using the "implant," a morning enema is mandatory.

So we did the 100mg of Mebendazole with 20ml of water. He kept it in and we did it right before bed... Next day a ton of worms came out!!! Better yet he started singing songs with a tune - 3 different songs over the last 2 days and counting on his fingers very purposefully! His aide at school said he did amazing today - no behaviors, very focused, better articulation and better social interest.

The Worm Whisperer

The following words of wisdom are from a mom who has earned the nickname *The Worm Whisperer* (although her ability to destroy parasites may make the *Worm Ninja* more apropos). Her dedication and diligence have not only made the difference in the life of her son, but countless other children who are what she calls "extreme cases." The information presented here may be the difference between an older/aggressive child living with their family or going to a group home for the care that their family is no longer able to provide them. Thank you Robin, for being a pioneer and for never giving up. Thank you for having the guts, the know-how, and the generosity to share what you have learned with those who need it most.

Extreme Cases by Robin Goffe

This section is for older children who may be violent, self-injurious, destructive, physically aggressive, high risk and/or bedridden.

Some families have children who are older when they first start treatment. My son was 18 years old. We call ourselves the "last chance" group. We say this because our children have lived their whole lives infested, unbeknownst to us. Our children lived as happy, relatively easily manageable, learning-disabled children: perhaps slightly annoying with their routines of movie dialogue scripting or attachment to Disney movies, maps, or little known facts. They were mostly friendless but pretty easy to take care of. However, during the hormonal teen years, things took a grave turn. We may have chalked up their newfound solitude to just giving up socially for lack of friends. They did not fit in. We had no idea that something else was brewing. The hormones within them began to have a bitter war with the parasites living inside them, and a war would soon break out. Their mental state would diminish considerably. What we are faced with are grown, strong children who are mentally ill and violent, self-injurious, and destructive. Some have even become so ill that they had to stay in bed.

This "last chance" is an opportunity to try ONE MORE THING before putting them in a group home; to give them away—for their own safety and the safety of their families.

These things that I am going to share with you are unique to the standard treatment outlined in the book. It is a more aggressive treatment because it is needed. Our children are so highly infested that it is shocking to tell you what we have found. There is a lot to do, but there is a method. It is time

The Goffe family.

consuming and it is a lot of work. But here I will outline what we did to get our son from an aggressive, self-injurious, destructive young man to one who has regained patience, language, understanding, and reasoning; but best of all, the opportunity to remain in our home. With continued treatments, he has the hope for a future: a job perhaps, and maybe even a family of his own.

Beginning treatment for the older child and especially one who is aggressive, self-injurious, violent, and destructive is one done starting low and slow.

Day one = one drop.

I kept a very detailed journal, and I will simply tell you what we did. I will tell you what worked for us and what did not. I will also tell you what I shared with others that worked for them as well. I am not a physician. I am not a chemist. I am a mother in love with my son and his utmost comfort was my main concern. My only goal was to clear his parasites. I had no idea that the things I learned by digging through his stool for 9 months would eventually give me the nickname, "The Worm Whisperer." Although intestinal parasites have been around longer than man, and they are smart enough to live within a human being lifelong, they can go completely undetected. I was ready to plot out a war against them. By the beginning of our 9th month of treatment, I estimate that I had cleaned and examined 35 pounds (16 KG) of solid parasites. Hard to believe? I have 80% of them photographed. The last two months in jars total seven pounds and a running total of the combined length will be at over 200 feet (61 meters), at the publishing of this book. There is not a classroom in the world that can teach what I learned in my bathroom day after day in latex gloves and a hospital mask. There was a fan blowing on me to manage the stench and by the third week my gag reflex was under control.

I examined the parasites, the pieces, how did it die? What was my son feeling at the time? Was he violent? Sweating? Slapping his legs from the pain of the parasites moving around and torturing him from the inside? I needed to eliminate these worms without allowing them to cause my son pain as they were killed. I had no idea the monsters I was to face.

In March of 2013, we started our son on one drop of chlorine dioxide. One of the most feared things that the parents of adult sized children face is the dreaded enema. It is just not something that most of us grew up with as standard care. But even so, my husband and I knew it was something that made sense. There are toxins inside these kids. They must be flushed out. The bowel is the way out. So on the very first day and the very first drop, we also explained to our severely autistic child that we were going to help his belly feel better. Our son, by this time, was so severe in regression that I can only best explain him as non-human. He was no longer speaking, unable to react to us, had stopped answering to his name, and could no longer contain the saliva in his mouth. He walked with a 12-inch drool hanging from his mouth. He ate like an animal, glaring at us. He was frequently violent: jumping on cars and denting them in, kicking down fences and destroying property.

This was the child that we were going to start giving enemas to. If we could do it, anyone can.

Each day we went up a drop; two drops on day two, three drops on day three, and so on. He was edgy and tired. We went for walks, as per usual, to deal with his aggression, but by day seven he began to have a runny nose and had a great deal of fatigue. We were so happy since we knew this was a sign to look for; that the immune system was kicking in. He slept for about 15 hours as we woke him and dosed him every hour and by the next morning there were about 25 white, hairy, thin looking objects in his stool. We knew these were worms and that we had our answer. After three weeks we began to see 4-6 inch parasites in all shapes and sizes, yet the aggression continued. It was here that I learned that what we were experiencing was POWS (Pissed Off Worm Syndrome). By this point we were at 13 drops, and while we were killing the smaller parasites, we were only pissing off the larger ones. The parasites do not like their environment disturbed, and subsequently they cause distress to the host. Parasites excrete ammonia, (possibly, leading to hyperammonemia and possibly seizures), morphine, and a meth-like substance. As these toxins enter the body, they can also cause aggression and anger. We had many possessions broken during these tantrums. We stopped valuing material objects.

It was here that we were told about—double dosing. So even though our son was getting about the equivalent of two drops an hour, if we saw increased aggression, OCD or self-injurious behavior we would give him four (see page 103 for an explanation of double dosing). If he did not calm down, we gave him another four drop dose. This was the perfect return attack for the parasites. Within a few moments the aggression settled. His red face and wild eyes diminished, and we had made it past another hurdle.

We used the double dose method dozens and dozens of times over the following months, and this helped tremendously to curtail the aggression caused by the parasites.

By the third week of treatment, we were concerned about the time during the night that our son would not be getting doses. So for the next six months our family had a dosing schedule, dosing additionally at midnight, 2am, 4am and picking back up to hourly doses from 6am to 10pm. This schedule was shared by not only my husband and I, but also our son's siblings,who would also set their alarms and take their turns, while we all rotated lack of sleep.

By the second month we found that our son's behavior not only changed around the full moon but the new moon as well. The standard protocol uses mebendazole over the full moon. However, with the older, more aggressive children we have had success with including mebendazole over the new moon cycle as well. These mini mebendazole courses feature shorter bursts of 5-6 days each, over the new and full moons beginning four days before each. We also deemed it necessary to start earlier in the moon phase than with the younger kids, since the infestation was greater, and the parasite movement occurred earlier, causing behaviors sooner.

For the child that has aggression, every parent must remember that the behavior is parasite related. During treatment you will see a variety of behaviors stemming from the parasites attempting to control their environment aka—the host. We found that increasing the frequency of the dose dampened the violence and aggression. It can be difficult to trust your child who is lost in a rage. You feel deceived and hurt. Those were our frequent emotions as well. It does pass. It took about 4-5 months for the rages to stop. They happened everywhere and were happening even on the toilet while trying to expel the parasites. Have a plan. Music worked very well for us. So did reading books. Find something to distract, and always give soothing, comforting tones. We frequently told our son how much we loved him and we rubbed his head and back. These herxing behaviors may be avoided through the use of a new method of preparing CD called CDH (see Chapter 7, page 155 for more info on CDH).

CDH: This new preparation method has really been a huge asset for the older kids that need to go higher in drops. At around 20 drops, the volume of the regular CD became unpleasant, making some children nauseous. My son was one of them. With the CDH preparation, I firmly believe that the older, tough-nut children will have more success with getting to the appropriate levels needed to start expelling the larger parasites. It is my personal belief that the amount needed in many of the non-seizure older kids with aggression, violence, self-injury, and destruction to be between 75-100ml of CDH daily. This amount may have a connection to the voltage needed in order for the mitochondria in the white blood cells to be given the power to not only kill the parasites, but to also destroy the bacteria. The job is two-fold, and must be done quickly as once the parasite is dead, the bacteria want to devour them immediately. I say these things because I have seen the condition of the parasites every day. I found that if I can kill the parasite with the least amount of disturbance or preparedness by the parasite, then there is no time for them to fight back and cause herxing. The bacteria also come into play and must be dealt with. Higher amounts of CDH kill both, without so much as a frown or a concern on the face of my son.

Stevia: I found that by adding the sweetener stevia to the CDH, the process can be more pleasant and does not affect the properties or effectiveness of the dose. As proven by Lamotte ClO_2 test strips (see page 467 for information on using the test strips).

Humidifier: In the beginning months, the infestation level is so high that nightly doses may have to be sacrificed in order to start breaking down the will of the parasites. It is not about seeing them every day, but slowly and consistently breaking their will. The constant aggressive irritation of the CDH will kill parasites, stopping them from making gains. There is another layer of nighttime treatments, and that is the humidifier. We fill a cool air humidifier that holds one gallon of water, and put 35 drops of activated CD (NOT CDH) and let it run all night near our son's head. This too was a method we used for maximum parasite elimination.

Spices and Herbs: It is urgent that we work constantly with our older kids to rid the body of parasites and pathogens. We have found that for kids that can easily swallow pills, filling empty gel capsules with the recommended herbs and spices in this chapter under Other Medicinal Plants (page 231) helpful. We have found great results by filling the gelcaps with the variety mentioned there: black walnut, ginger, rue, wormwood, and yarrow and giving two to four different ones with each meal as a toxic feast to the parasites. Don't narrow your selection to those mentioned here—get them all. There is no plan or pattern. Use any combination.

Colonics: Professional colonics have been an important addition to healing and cleansing parasites in our older kids. We started these a few months after our son was accustomed to the routine enemas. This has been a wonderful addition, and we try to do at least one weekly. Some places will allow you to do inserts, and we have taken our CDH there working up to 50-100ml CDH with great success.

Limiting Undesirable Behaviors

Probiotics: If you are using probiotics and your child is aggressive or becomes aggressive after restarting the probiotics, consider removing them as a first step. There could be such a high infestation rate that the child's system does not distinguish between good bacteria and bad bacteria.

Enemas: CD enemas should start on day one of treatment for the aggressive/older child. Since they have such a high rate of infestation, it is necessary to clear the pathogens on a continuing cycle from the very beginning. Please follow instructions for enemas starting on page 103.

Diatomaceous Earth (DE): In some children, the DE may shred the larger parasites to death, causing them to spew additional toxins into the child's body, which results in herx symptoms. Discontinuing DE for a few months may help to ease the distress, and give a cleaner kill.

Salt Baths: There are times that the body is unable to pull out the toxins fast enough. Be careful about killing too fast (dosing without or not enough enemas). If you have done enemas, and there still seems to be distress, you can buy a 40-pound bag of plain pool salt, take 9 pounds of salt and put it into a hot bath. Allow the salt to dissolve and have the child soak for one hour. If the child sweats during the bath, this is a good sign, as it shows toxins are being pulled out. Do not use Epsom salts. They contain magnesium. Magnesium feeds parasites as well as biofilm. Plain pool salt is cheap, and helps to remove toxins.

Dandelion Tea: This is an incredible detoxing source as well. There have been times that our son was in and out of a salt bath all day drinking dandelion tea to get the built-up toxins out. This is a tea ready made by Traditional Medicinals® (traditionalmedicinals.com).

Tips on What to Expect During the Parasite Elimination

Intestinal parasites have been on Earth longer than humans. Their plan is to find a host, and to continue to create life. They have the home advantage by living inside of your child. There are many things that I have learned while at war with them. They have a plan and so should you.

Situations We Have Come Across

Peetox: As you kill the parasites, they are alerted that you have taken away their joyful life. As a result, they do things to the host to show their displeasure. One of the things the parasites do is to spew toxins, including a morphine-like substance that can cause your child to not feel that he needs to urinate or that he is urinating. Naturally, this can lead to accidents. This is not the child's fault, but only a part of the parasite elimination process. One of the most difficult things we did was to put our 18-year-old son in adult diapers. It was heart breaking because it felt that we were moving backwards instead for forwards. However, this was only temporary, and for his own dignity. Especially if they go to school; perhaps being in diapers for a while is the answer. Our son was in and out of them for about two months. This does end.

Sleepytox: As you are going higher in drops, the CD is assisting the immune system to get the parasites out. It will take a lot of energy to do this. Remember, your child is the host and there is a war going on! Sometimes, your child may be tired for days. Our son slept for nearly the entire summer at the beginning of his treatment. Allow the rest they seek. Now our son gets higher doses of CD on the weekends, and he is very tired and sleeps then, too. By Monday morning, we can usually get a big fat rope worm after our efforts. Sleepytox is great because you can get a lot done on your home "to do" list.

Behaviors/OCD/Tics: Some may become alarmed because their child starts having behaviors that they did not have before treatments, or they feel that the behaviors have gotten worse. Parasites have a plan. It is to stay alive and to procreate. They want their eggs to return back into a host to repeat the cycle. So, you may see behaviors of playing with feces, touching their anus to their mouth, trying to put their fingers into your mouth or another person. There are finger and nail chewing, and spitting. These are all parasite controlled behaviors. We have found that eliminating parasites helps to minimize this behavior and it will eventually subside. There may also be a mineral deficiency attached as well. We have found that giving PLENTY of ocean water can help to tame these behaviors. Our 19-year-old son gets over 200ml OW mixed with 400ml spring water per day.

POWS (Pissed-Off Worm Syndrome): POWS presents itself when you have been dosing, and your child seems weepy, is pacing, and/or unhappy. Sometimes it is really that you have pissed off the worms, but are not killing them. Always check with Kerri regarding your concerns. But many times the answer is to go up in drops not down. You need to kill them—not piss them off, or they will fight back.

Clear Zone: This time period can be about five to seven days before the full moon; once infestation is brought down after six months or more of treatment. You may be able to start seeing glimpses of your child being lucid. This is a good time for an ATEC test.

High Volume Eating: Often times, the kids reach a point where they realize that they have been starving. The parasites have been taking a bulk of their nutrients, causing their mental deficiencies. They may even be eating non-stop all day. It is tough to juggle the timing and effectiveness of the CD with the continued eating. Allowing the child to eat high volumes of food while dosing with CD will still kill the parasites, but the CD levels may not be high enough to avoid a herx. Though the parasites may be killed the CD level may not be high enough to also kill off the bacteria eating the dead parasite. One solution we found was to give our son homemade bone broth with ocean water many times a day. This will help to replace the lack of nutrients, meanwhile reducing the overeating enough to continue the CD without risking a herx reaction.

Weight Loss: Our son lost 50 pounds during the first seven months of treatment. He was very over weight at 5'9" and 200 pounds. The weight loss brought concern to people who had not seen him for long periods. During this time I estimate that 35 pounds was solid parasites. His stomach was very bloated, as was his face. Our children are full of parasites and pathogens (i.e., yeast, bacteria, etc.), and all of these things can make them gain weight or be bloated. Our son is now in the normal weight range for his height, and has a flat stomach.

Night Terrors: During the night you may see that your child has trouble settling down, or they may wake in the night screaming your name. They may even do night walking, possibly leaving out the front door in a sleepwalk state. We saw these behaviors years ago, as our son was heading into a higher infestation level. Now that we are on the healing end, many of the behaviors return as sort of a rewind. The behaviors that came into play going down the ladder are sometimes repeated heading back up the ladder to wellness. We feel that our son was experiencing separation anxiety, and the recognition that he had total control again. This new gained independence may be frightening. At a point, our son entered the peetox stage (see above). He had no feeling or sensation to urinate, so he was put into adult diapers. As our son's infestation rate decreased, the normal behaviors like waking in the night and feeling the urge to urinate returned. Then, as he would get up to go to the bathroom, he would suddenly stop, and feel perhaps lost and afraid to continue this task alone. He would scream out or wander around whimpering. Although night wakings were disturbing, we found them to be a necessary passage for bringing our son to reality. In order to give comfort, at night we continued the CD humidifier, as well as CD ear and nose drops. Many times a CD bath before bed was calming as well.

Emotional Blindness: Because the toxins that the parasites excrete are substances that have the effects of morphine, ammonia, and histamine, these toxins can cause emotional confusion in your child. They may be confused by your laughter, smiles, and affection. Your positive and loving encounter may be perceived as threatening or challenging. Your smile may bring a rage. Any spoken words by you no matter how soothing they may be, could be perceived as a confrontation. This is difficult on both parent and child, as the child may see you as a threat and perhaps even as mocking them; while you as the parent feel rejected in your affections. Many times the child's behaviors evolve into a stoned-out loss of reality and presence. These are all related to the toxic build up and defense of the parasites. Their motivation is to remain in the host; they show their displeasure to the discomforts we have given them in our elimination attempts. Many times during these episodes, we have found it best to avoid words or eye contact, but rather to offer double doses as outlined by Kerri, and keep a calm environment. We have found that these toxic levels can be lowered through salt baths and dandelion tea.

Despite the behaviors that may be brought on by a parasite protocol, the results are amazing. In our family we are finally getting our son back. I would recommend staying the course. As we know, these kinds of parasitic infestations in immune-comprised children will not go away on their own. Children do not "grow out of" worms. In fact, the worms will continue to grow inside of them. The following parasites were expelled by my older son, by making the aggressive modifications aimed at older children. By removing these parasites we are seeing the first stages of recovery. The jars on the following page represent seven pounds of worms in less than two months.

Yours in healing,
Robin

Jar #2 Sept 30th
Oct 13 Rope

7 lbs. of 35 lbs. (total) of the parasites extracted in 9 months (160 feet in total above).

28
29
30
31
32
33
34
35
36
37
38
39

39+ inch long worm.

Other Medicinal Plants

Author's Note: The following information is extremely powerful for our continued health and well-being. However, these plants and foods ALONE have NOT been proven to heal autism. Rather, follow the 18-day deworming protocol presented starting on page 187. When this is done for 12-18 months it has proven to be an important part of the protocols that have led to the recovery of many children. With that in mind, some parents of older children, and aggressive/self-injurious children have started to implement a few of the plants and foods listed below, in addition to the aforementioned 18-day protocol. In many cases, these additions have led their children to increased gains through increased parasite elimination.

A number of other plants are also effective for deworming. If, after three months of treatment, the problem persists we can change the type of plant, or repeat any plant that was effective in previous months. We can use them in combinations, mixing several plants at once, or take them individually. Plant formulations that should be considered are alcoholic extracts, in oil or by infusion, and include the following plants:

- Clove (*Syzygium aromaticum*)
- Common Rue (*Ruta graveolens*)
- Dandelion (*Taraxacum officinale*)
- Gentian Root (*Gentiana lutea*)
- Mint (*Mentha sativa*)
- Mugwart/Common Wormwood (*Artemisia vulgaris*)
- Pomegranate Root Bark (*Punica granatum L.*)
- Southernwood (*Artemisia abrotanum*)
- Sweet Wormwood (*Artemisia annua*)
- Sweetflag/Calamus Root (*Acorus calamus*)
- Tansy (*Tanacetum vulgare*)
- Walnut shell (*Juglans*)
- White Fraxinella (*Dictamnus albus*)
- Yarrow (*Achillea millefolium*)

Preventative Food & Diet

There are groups of foods that should be avoided if you have a parasitic infection. For example, dairy products in general, refined sugars (sucrose, fructose, corn syrup), flour (especially refined), and overly sweet foods in general. The list of foods and plants below promote good internal balance of the body, hence becoming our allies. With good production of stomach acid, a normal level of healthy bacteria, and proper bile production, it is impossible for parasites to survive for long. Worms need an acidic environment that comes from the breakdown of sugars and putrefaction from the ingestion of unhealthy or processed foods. It is very important to eat raw vegetables and fruit juices, which provide us with enzymes and other elements necessary to protect us.

Choucroute/sauerkraut (fermented cabbage in salt).

Many people have low levels of stomach acid, which is the cause of many intestinal problems, because the body is unable to defend itself against intruders. Sauerkraut juice or cabbage/sauerkraut is one of the most powerful stimulants for your body to produce stomach acid. The use of unpasteurized fermented foods (water kefir, soy sauce, miso, etc.) is highly recommended for its stimulation of the beneficial bacterial flora that is responsible for generating control over parasites. Take a few spoonfuls of cabbage juice before meals, or better yet sauerkraut juice, because it will do wonders to improve your digestion.

Author's note: I do not recommend soy, in any form, for anyone with an ASD.

Garlic

Garlic, eaten regularly, turns the stomach and intestine into a lethal environment for parasites, providing constant protection. Garlic is the quintessential home remedy to eliminate intestinal parasites naturally. It has been used by many different cultures such as Chinese, Greek, Roman, Indian, and Babylonian.

Garlic is still in use today by practitioners of modern medicine. It is used both fresh and as an oil. The simplest treatment is to eat three cloves of garlic every morning, or take a teaspoon of garlic oil. Alternatively, mix crushed garlic in a little cold water and drink the mixture immediately. Another recipe is to cut and crush four cloves of garlic, place them in milk, and allow the mixture to sit overnight. Take the liquid while fasting the next day.

Pumpkin Seed

Pumpkin seeds contain a substance called piperazine. It acts by paralyzing the parasites, which allows them to be removed easily.

We can find piperazine commercially in pharmacy drug formulations or naturally, as we said, in the seeds of the pumpkin. This traditional method of deworming has been used around the world since man can remember. There are several effective traditional formulas, below we describe one of them:

Use one cup of peeled and mashed pumpkin seeds (about 80 seeds). Mix them with coconut water and two tablespoons of honey. Take the mixture over three hours on an empty stomach. Do not eat during this three-hour period. At the end of the three hours, take castor oil in order to quickly eliminate the parasites.

Papaya and Papaya Seeds

Papain is a digestive enzyme contained in papaya that is capable of breaking down the outer layer of adult parasites. The milky juice of unripe papaya is a powerful agent for destroying roundworms. The adult dose is one tablespoon of fresh green papaya juice, an equal amount of honey, and three or four tablespoons of hot water. Two hours later, administer a dose of castor oil mixed with warm milk. This treatment should be repeated for two days if necessary. For children seven to ten years of age, half of this dose should be administered. For children under three years of age, one teaspoon (5ml) of the mixture is sufficient.

Papaya seeds are also useful for this purpose as they are rich in papain and caricin. For every tablespoon of crushed, fresh seeds, add an equal amount of honey. Take the dose of one teaspoon (5ml) daily in the morning or at night on an empty stomach for ten days, rest five days and repeat the cycle three times. We recommend the use of a purgative.

Ginger

Ginger not only helps to combat intestinal parasites but also reduces nausea and can help calm nerves. For hundreds of years, fresh ginger has proven to be highly successful in destroying intestinal worms. The most common way to consume ginger is raw or by infusion. Ginger extract may also be sprinkled on a variety of foods.

Propolis

Propolis is a resin like substance gathered by bees from the bark and leaf buds of trees, to help disinfect, build and maintain their hives. Propolis has been used for at least 3,000 years. Its use dates back to the Egyptians and the Romans, and remains in use today. To the Greeks we owe the name pro, meaning "before" and polis, meaning "city." This translates as "defenses before the city," or "defender of the city." Thanks to the antibiotic action of propolis, which protects against the activity of viruses and bacteria, the hive is one of the most sterile places known to nature.

Many scientific studies have proven the antiparasitic activity of propolis, therefore it is recommended for treatment of: *Giardia*, amoebas and roundworms, and also for intestinal infections caused by gram-positive bacteria.

Take propolis, diluted in water or fruit juice, for treatment of parasites, for seven days, on an empty stomach. Use Propolis standardized at 30% in either propolis tincture or capsules. Take three drops per kilo of weight, or three capsules one half-hour before each meal. A seven-day treatment cycle should include seven days on, followed by seven days off; repeat three to five times to ensure complete elimination of parasites or bacteria. Repetition of the treatment is essential to halt bacterial reproductive cycles. By repeating the treatment at least three times, the effective elimination of parasites is ensured. The benefits of propolis are that it has no side effects, is well tolerated, and is highly effective.

Pomegranate Bark

Pomegranate bark contains an alkaloid known as punicine, which is highly toxic to earthworms. It is used by decoction of the root bark, stem, or fruit. The root bark is preferable because it contains a greater quantity of the alkaloid than the bark of the trunk. This alkaloid is also highly toxic to tapeworms. A cold decoction of root bark, preferably fresh, should be given in quantities of 90ml to 180ml three times per day (for adults), with one-hour intervals between cups. A purgative should be taken after the last glass. For children, a dose of 20ml to 60ml is appropriate. A decoction is preferably used to eliminate solitaires (tapeworm, *Taenia Solium*).

Carrots

Carrots are another effective home remedy for eliminating intestinal parasites in children. The chemical constituents of carrots attack pests by preventing their development. It is one of the most effective natural treatments for children, when given a small cup of grated carrots each morning until the problem desists.

Condiments

Seasoning plants are also powerful weapons to keep in mind in our everyday cooking. Since time immemorial, mankind has used them to control parasitic diseases. The following are most interesting because of their effects:

- Cayenne
- Cinnamon
- Cloves
- Paprika
- Pepper
- Tarragon
- Thyme
- Turmeric

Thank you Andreas and Miriam for sharing what I know will be very enlightening, not only to the families of children with autism, but to people all over the world suffering from mysterious symptoms consistent with parasitic infections. I want to share one interesting tidbit about parasite infections before we get into the FAQs on parasites and the parasite protocol.

"We have a tremendous parasite problem right here in the U.S. It is just not being addressed."[5] - Dr. Peter Wina, Chief of Patho-Biology in the Walter Reed Army Institute of Research in 1991. (The problem existed in 1991, and with modern globalization parasites are more prevalent than ever, yet they still not being addressed.

Parasite Protocol FAQs

I showed my family practitioner pictures of the worms we have found. He thinks they are not a parasite, but just mucous. How can I be certain?

It can be difficult to identify parasites, and most general practitioners are not trained to do so. Some parents have success with a local vet, who analyzes samples in their office. As a crude test, you can pour boiling hot water onto your specimen. If it falls apart, it is probably biofilm or mucous. If it withstands the wash, it is probably a parasite. Unfortunately, most stool analyses are notorious for false negatives. If you put hot water on parasites, they do not dissolve. Mucous dissolves in any water—hot or cold.

Is it ok to do the Kalcker Parasite Protocol during pregnancy?

No. Don't do any kind of detox or parasite protocol while you are pregnant or nursing. Any detox will release toxins into the blood stream that could potentially negatively impact the developing fetus or nursing baby. If you are planning on getting pregnant, it would be advisable to do whatever detox or deworming procedure before pregnancy.

When is it appropriate to stop the parasite protocol? In other words, does every child on this protocol need both the CD and the other parts of the Parasite Protocol or are there some that only need CD? This is a critical question since these are two protocols joined together and not every child may need both. We have done the Parasite Protocol and have never passed parasites so I am uncertain about my child.

Usually the Parasite Protocol is repeated for 12-18 months, sometimes less. The most important thing is to make sure that there are no fertilized eggs left that could hatch at a later date. As it is very difficult to know if there are remaining eggs, it is important to complete the 12-18 month treatment. The Parasite Protocol is a complete protocol heavily researched to achieve the best results. The Parasite Protocol in conjunction with CD has been shown to be one of the most effective methods to heal regressive autism. Parasites are cyclical and need to be treated as such. You continue with the parasite protocol until no more parasites are seen—be that in behavior or actually in the stool.

What does the process of worms toxifying the brain look like?

Ascaris lumbricoides (roundworm), for example, produces at least five different toxins: Malondialdehyde, ammonia, histamine, formaldehyde, and morphine. Malondialdehyde is responsible for oxidative stress, and is mutagenic.[7] Ammonia, which can lead to hyperammonemia, can be responsible for seizures, tremors, flapping, poor coordination, growth retardation, combativeness, lethargy, and other symptoms. Comparing hyperammonemia with the symptoms known as regressive autism yields overwhelming similarities. Formaldehyde has been shown in some laboratory studies to affect the lymphatic and hematopoietic systems. Morphine inhibits nerve reactions and slows intestinal peristalsis. It also keeps the immune system from finding parasites, and from doing anything about them. This is a reason we often can't identify parasites. We only find parasites when the infection is acute, not chronic. This is because IgE or IgM reactions are altered by morphine. Histamine can lead to chronic inflammation in the body.

As a family, we expect to be doing the CD and parasite protocols for some time. How do we know when to stop? Should we wait until we have a few months of no symptoms?

We do it until full recovery is reached.

We have been doing full oral dose CD and the Parasite Protocol. We are seeing some amazing changes, but we don't see worms. Are we doing something wrong?

CD affects parasites of different sizes. You may be clearing parasites that would be undetectable to the naked eye, or nematodes that are very small and therefore hard to detect in the stool.

Why is my child so deficient in vitamins?

As a general rule, most children with autism are deficient in vitamins. In the first place many pathogens feed off of vitamins intended for the host, and helminthes especially love B12 and iron. Since many of us, especially those of us living in cold climates, do not receive enough sun, we are deficient in vitamin D. Calcium is used by the body as an antagonist for acid inflammation; all acidity in the body is compensated by calcium, and therefore it is usually low with the kids on the spectrum. As our children heal, these deficiencies fade and homeostasis returns.

I just gave mebendazole, Not pyrantel this AM. What should I do?

Don't stress. Tomorrow is a new day, and you can begin again. Give mebendazole tomorrow as that is tomorrow's scheduled dose.

Aren't there some parasites that are good for us and help our immunity and gut healing?

A parasite is defined as a living being that is dependent on a host for survival, to the detriment of that host.

For those of you that continue the parasite protocol thru the new moon, how many days do you stay on it? We are now 3 days past the new moon.

Some people treat month long. The Kalcker protocol itself is 18 days long. Each family has to find a protocol that suits them. Some folks use mebendazole 3 days before the new moon, the day of, and three days after. Other families use herbal remedies on the off days. If you consult with a healthcare provider, this would be a question to ask them. Each child is unique and has different needs when it comes to treating parasites.

Why are live worms not digested? Poisons?

Live helminths are protected by a glucosoid mucous with positively charged ions, making them resistant to stomach acid, or digestive fluids. When they die, the mucous separates from the helminth, leaving them open to digestive enzymes. We often see the mucous, as well as semi-digested helminths, in the stool of people using the protocol.

Does everyone in the family need to do the parasite protocol at the same time? With or without CD, enemas, diatomaceous earth, Rompepiedras/stone breakers, and castor oil? My (neurotypical) daughter will take the anti-parasitic drugs, mebendazole, and pyrantel pamoate, but not these other parasite protocol ingredients, which is why I ask.

Your entire family needs to do the protocol at the same time, or you risk reinfection. Neurotypical family members should do as much of the protocol as possible. If someone will only take the meds, so be it. However, I feel the more the better.

Do I need to separate the neem from CD?

Yes, I give neem with food, and would separate from CD by at least an hour.

If I don't see any worms, does that mean my child doesn't have parasites, or are we just missing the eggs?

Just because you don't see parasites doesn't necessarily mean you don't have any. For example, *Toxocara canis* or *Toxocara cati*, which are very common in our pets (and can infect humans as *Toxocariasis*), are not expelled in fecal matter. Some families on the protocol did not see actual worms until month seven.

To what extent do the rest of us need to do the CD protocol? Is it the same as our ASD children? And, is it necessary to do the whole nine yards with the full Parasite Protocol for the whole 12 months? Are there any shortcuts for healthier individuals?

All of the family members, including pets, need to be on anti-parasite treatment for a year and thereafter lifetime maintenance is best. As mentioned before, with neurotypical family members, you do as much as you can, but the anti-parasitic medications are crucial to preventing reinfection.

Is diatomaceous earth a binder and therefore must it be given apart from food, medication, and supplements?

No. It is not a binder. It is fine to give with or without food, and does not affect medication or supplements.

Do you have to do the Protocol forever if you are fully recovered or is there a maintenance plan?

When the child gets to where you want him, you can start to pull things. Then, we do maintenance dosing of CD; one dose on Monday and one on Thursday. We do the Parasite Protocol for a week every three months since we live in a world of parasites. It is good preventative medicine for all. Ocean water is also good for all of us. See Chapter 13 (page 319) for the complete maintenance plan.

Does CD kill parasites?

Yes. CD kills amoebas, *Giardia lamblia*, and other smaller parasites. CD does not kill the larger macroparasites due to their higher oxidative stress resistance.

When is it ok to start the parasite protocol prior to reaching full dose of CD?

Healing autism is a marathon not a sprint. The point is to heal with the least amount of aggravation. Therefore, you should be at a full dose of CD before starting, and have added in enemas and baths. If you are seeing Herxheimer reactions then don't start yet. Obviously, everyone must make his or her own decisions.

Where in the gut do helminths live? Do CD enemas (300 ml) reach them?

It depends on the type of helminth. There are more than 300 different helminths so they could be in many different places. However, most live in the small intestine. Some lay their eggs in the rectum. The size of the person will dictate how high an enema will reach.

Do you tell your kids that they have worms? Do you show them? Are they scared? My son is very interested about everything in the world and asks 4,000 questions a day. What do I tell him?

How to handle this situation is unique to each and every family. There is no right or wrong answer here. The following are some suggestions we have received from families on the protocol:

Nope. My husband was just telling me a story from when he was a kid about another kid who got teased all the time about having worms. He still remembers it. So, no I will not tell my kids.

My kids (ages 10 & 7) know about worms because we have talked about their Lyme disease and other symptoms for a long time. I think it helps them understand why they feel grumpy or sick and why the enemas help. But, I think every child is different. The risk is that they might tell their friends. I tell them that it is private information and that pretty much everyone has worms but some people have more and get sick from them. And, I told them that even most doctors don't know this but slowly people are learning and getting smarter about this and a good diet, etc... My boys like to see the big ones, and try for new records. They like to show their brother the records. I think that's a boy thing.

I have a 10 year-old, and I tell him that we have parasites. I have not showed him anything, I ask him to leave the room before I start to sort the poo. ;)

My son had another bowel movement FULL of worms, pinworms—and I mean hundreds. This is only our second day on CD (1/2 drop in 8 fl. oz.) so a bit thrown to the deep end. Has anyone else had a start like this? How did everything go from there on?

I have seen it a handful of times. I get VERY excited about this. Everyone has autism a bit differently. But, if we see worms early, in many cases the recovery begins early too. That may be the case for your child.

I understand that even organic produce needs to be washed properly. What is the best way to wash my fruits and vegetables?

The Ascaris eggs are resistant to UV, and can withstand a pH from 2 to 11.5. Eggs are killed by heating to approximately 60°C (140°F). Spray your fruits and veggies with CD spray. 10 drops per 1 fl. oz., of water, no need to rinse off.

Anyone notice a huge growth spurt after starting the parasite protocol?

Many parents have reported their children resumed growth as well as weight gain. No real surprise considering they are getting back what the parasites were taking from them.

Do we need to treat our pets if they are taking heartworm medication?

Yes, heartworm medication does not address intestinal parasites, which can infect us and our children. We need to deparasitize our pets, as well as ourselves.

We put "cascara sagrada" in place of castor oil because for my son and me, it is difficult to take the oil. Is it ok?

Cascara sagrada is not the same thing as castor oil and is not part of the protocol. An alternative solution to liquid castor oil is to buy castor oil in softgels. You will need many but it does solve the taste problem.

My son is very high in Trichinella spiralis, which is a roundworm from undercooked meat or pork. It's supposed to be rare in the US, so I'm wondering how my child would have picked this up. Any ideas? No undercooked meat here for the kids, and we don't eat pork.

At this moment we have not seen a relationship between *Trichinella spiralis* and autism. It may be a multi-parasitic infection, or a misdiagnosis. Apart from the standard protocol, your child might need something else. I would consult the practitioner for the appropriate dosage, and only if your child has a definite Trichinella spiralis diagnosis.

How do we know if there are parasites in the brain, and what can we do about it?

Parasites in the brain are very rare. (According to the CDC, *Cysticercosis* is a parasitic tissue infection caused by larval cysts of the pork *tapeworm.* These larval cysts infect brain, muscle, or other tissue, and are a major cause of adult onset seizures in most low-income countries. An individual acquires cysticercosis from ingesting eggs excreted by a person who has an intestinal tapeworm.)[6] Larvae can be seen on a scan. Many people assume that the problem causing behavioral issues, or mental issues, must be located in the brain. However, the problem likely exists elsewhere. If the blood contains toxins, this will affect the brain. For example, if you ingest too much alcohol, your brain, nervous system, etc. will be affected, but the problem is not in the brain itself. We can think the same way about parasites. The chemicals that they produce will have effects on the brain, however, rarely will the parasites be located there.

Do parasites cause autism? If so, why doesn't every child with parasites have autism?

Regressive autism has been called parasitological vaccinosis by Dr. Andreas Kalcker. This is a cross reaction between a child with parasites that receives certain vaccines. Further research is needed for a definitive answer.

Why don't normal lab tests find parasites?

Generally speaking, lab tests need to find living creatures or eggs. The eggs are only present certain days of the month, and even on those days they won't necessarily be present in a particular stool sample. Living worms are extremely rare in stool because they are generally passed only once they die. They are very good at avoiding being excreted in stool. Furthermore, if they die internally, they can be partially or fully digested within our bodies before being expelled.

Are some labs better than others? Is it worth taking a stool sample to a vet if they send it out, or only if they look under a microscope?

We haven't found a lab that consistently finds parasites. Metametrix™ has identified parasites, but in our experience and the experience of the families I have helped, we have not found a lab that is consistent. It is only worth sending a sample to a vet if they are identifying parasites with a microscope.

Combantrin® is available in tablet and liquid form. Which is better?

I hate the liquid and never use it. It may have added colors and flavors. If you absolutely can't find it without additives, I would use only mebendazole instead of using pyrantel with additives.

Why are there no systemic drugs in the protocol?

They are not needed with this Protocol. Treating parasites without systemic drugs is much safer and easier on the body. The suggested treatments using mebendazole and pyrantel pamoato are nearly unabsorbed by the body, meaning that we do not add more toxins to an already overloaded body.

When does a child need systemic drugs?

This is necessary only for certain parasites like cysticercosis, which is caused by tapeworms, hookworms, Trichinella spiralis (from pork), or other hard to kill helminths. These may be identified by a blood test. The practitioner may prescribe systemics depending on the situation.

Should I start on the new moon if my child's behavior declines then?

The full 18-day Kalcker Parasite Protocol always starts before the full moon. However, some parents have found that by treating parasites for 3 days over new moon that they are able to get through the new moon as well as the full moon with limited issues if any.

Is there a time in the moon calendar, where it is normal that there are no worms in the stool? (We don't have any at new moon.)

Further research is necessary to determine this definitively.

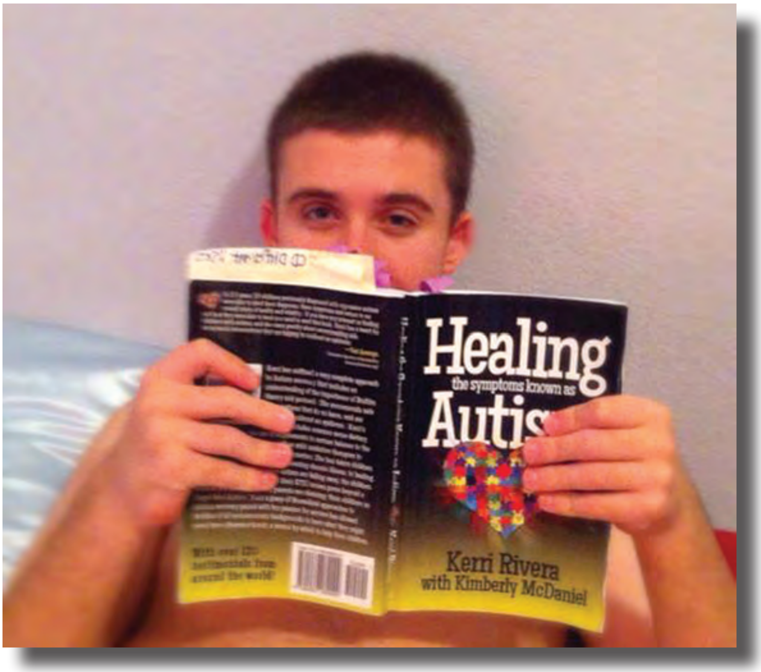
Healing
the symptoms known as
Autis
Kerri Rivera
with Kimberly McDaniel

Hero Guy taking a peek at *The Protocol.*

Chapter 9

Step 4 - Other Supplements

"One of the first duties of the physician is to educate
the masses not to take medicine."
~ William Osler

Most of the families that followed this Protocol, and recovered their children, used a combination of these additional supplements along with CD. Each supplement combination is completely unique, depending on the symptoms of the child, and how the child reacts to each intervention. If after *The Diet*, full *CD Protocol*, and three *Kalcker Parasite Protocols* we are still dealing with an autism diagnosis, then families start adding in supplements. This section is about "unfinished business." I like to do three Kalcker Parasite Protocols before adding back in supplements because parasites love supplements, especially vitamin B_{12} and iron.

I feel when we supplement deficiencies in children with underlying parasitic infections, it can result in resistant parasites. We aren't making the child healthier; we're making the parasites healthier and stronger. If significant supplementation does little or nothing to change lab values in subsequent blood, urine, and stool tests, then it's fair to say that we have a flaw in our approach. Pathogens and parasites can cause vitamin/mineral deficiencies. By killing them, instead of feeding them with additional supplementation, we are focusing on the root cause of the symptoms that we call autism.

When I came back from our first *Defeat Autism Now!* doctor visit, I had thousands of dollars in supplements that I did not know how to use. We kept going to different doctors hoping to find the one that had the answer. At one point, Patrick was on a protocol, given to me and supervised by a doctor, which called for 70+ different supplements a day. He looked worse during the entire nine-month period that we followed that protocol. In hindsight, all those years of supplementation made it harder to eliminate Patrick's pathogens.

That is why this protocol focuses on the excesses instead of the deficiencies for children with an autism diagnosis. If by eliminating the excesses of viruses, bacteria, candida, parasites and heavy metals, we can see a reduction in the symptoms known as autism, then eventually we can reach full healing.

Commonly, when a new family contacts me, they send me a list of all the supplements their child is taking. I frequently see lists of 30+ supplements that may be combined in ways that negate their effects. For example, probiotics are given in the morning with breakfast, or GABA with food. Both are incorrect, probiotics go alone at night right before bed, and GABA is always taken without food. If there is a legitimate reason for a certain supplement, then it absolutely must be given correctly or it's just excess noise in the body, and more stress on the detox organs.

The following list of supplements can be considered after *The Diet*, ocean water, CD, and three months of the *Kalcker Parasite Protocol* (except probiotics and enzymes which may start day one). This chapter is to be used as a starting point for your own research on these supplements. This list contains references and the websites used in this research with some of the more pertinent information about products.

Probiotics

The vast majority of the families who have recovered their children with this Protocol have used THERALAC® brand probiotic. It is a prebiotic plus probiotic in a capsule that can withstand stomach acid, which allows it to pass to the small intestine, where the beneficial bacteria are needed. THERALAC® contains three *lactobacillus* and two *bifidobacterium* strains.

The paragraph below was slightly modified from www.theralac.com where you can find more information. Please visit their site for details on references contained within the text.[1] The following two paragraphs were taken from different sources, and help explain benefits of THERALAC®, specifically for children on the spectrum.

> Probiotics are defined as: "Live microorganisms that when administered in adequate amounts confer a health benefit to the host" (World Health Organization 2001). According to the International Probiotics Association (IPA), the benefits of probiotics can include reduction in diarrhea caused by antibiotics and rotavirus, alleviation of symptoms of lactose intolerance, alleviation of symptoms of food and skin allergies in children, reduction of recurrent ear and bladder infections, and other positive indications.

How Does THERALAC® Benefit My Immune System?

> THERALAC®'s five probiotic strains work in unison to support the mucosal immune system on the intestinal surface; activation signals are then sent to the body's systemic immune system. This is a gentle and controlled process because THERALAC®'s strains are beneficial. Essentially, the immune system is put on alert so that it is ready to act quickly if there is an appearance of pathogenic microorganisms.[2]
>
> Many children with autism have chronic digestive problems. In fact, gastrointestinal symptoms in autistic children often first appear in conjunction with initial changes in emotion and behavior during the onset of autism, leading researchers to suspect a gut-brain connection.[3]

Normally, proteins are digested in stages by enzymes; first to peptides, and then to smaller amino acid components, which are absorbed into blood capillaries in the gut mucosa. The larger peptides are generally unable to cross this mucous membrane barrier, if they do, however, they can act as opioids affecting neurotransmitters in the brain causing abnormal behaviors and/or activity. These incompletely digested peptides—known as exorphins, casomorphins, and gluteomorphins usually come from milk proteins such as casein, or from wheat (gluten), and are structurally similar to morphine. The formation of excess peptides in the gut is possibly associated with sub-optimal enzyme activity, or an insufficient supply of enzymes required to breakdown these peptides. So if we repair the imbalance of beneficial bacterial organisms in the gut and the gut lining, while killing the pathogens causing the dysbiosis, we have the opportunity to heal children with autism. Probiotics can be used to improve the quality of the gut mucosa.[2]

In addition to aiding in the repair of the gut lining and improving digestion, there is also evidence that probiotics can help with detoxification of heavy metals such as toxic mercury.

The paragraph below was taken from the presentation "Gut biology and Treatment" by Dr. Anju Usman:

> Emerging literature is showing the beneficial effect of oral probiotics on mood and anxiety symptoms. In a double-blind, placebo controlled randomized parallel group study, daily use of probiotics reduced psychological distress.[5] A number of studies have shown the anti-anxiety effects of probiotic use in patients with medical conditions.[6]

> Dr. Vincent Young, University of Michigan (2009): "the gut ecosystem needs to be preserved and that changing the ecosystem through stresses such as antibiotics could irreversibly change the ecosystem, with deleterious effects." Dr. Young has studied the effects of antibiotics on the microbes in our gut. He found that mice when given particularly strong antibiotics completely wiped out all their normal gut microbes. Even more striking, clostridium species and fungal species are then able to overgrow without the bacteria there to fend them off.

There is no doubt in my mind that probiotics helped get my son's gut back in order, helped improve my own health, and that of thousands of families of children on the spectrum. Probiotics should be given without food, directly before bedtime. This way, they have the entire night to proliferate in the small intestine where they are needed.

Special note for those with PANDAS/PANS: We have received some reports where some PANDAS/PANS and/or older children were becoming violent and having more SIBs while on probiotics. Be observant and watch for behavior changes when adding probiotics. In addition, know that we may need to remove them for a period of time!

Omega-3 & Omega-6 Fatty Acids

Omega-3 and omega-6 fatty acids are polyunsaturated fatty acids and are considered essential fatty acids. Essential fatty acids must be consumed because your body cannot produce them. They are important for brain development, immune system and cardiovascular function, and normal metabolism. Omega-3 and omega-6 fatty acids are commonly found in marine and plant oils. While a healthy diet often provides adequate supplies of both, supplementation is sometimes necessary.

Omega-3 fatty acids also help to regulate energy levels, as well as normalizes blood sugar levels. They help improve concentration as well as mental vividness. Using omega-3 fatty acid supplements helps with anxiety, mood swings, and depression. They are especially significant in the treatment of autism. The entire body benefits from ingesting omega-3 fatty acids; from physical development, to relaxation from anxiety and jitteriness, to developing brain cells; all of these aspects can aide in the treatment of autism.[6]

Omega-6 fatty acids promote healthy brain function and assists with skin and hair growth, bone development, and metabolism. A healthy balance of omega-3 and omega-6 fatty acids promotes heart health and minimizes inflammation. Omega-6 may also be used to treat allergies, eczema, osteoporosis and premenstrual syndrome.[7]

Unlike omega-3 fatty acids, excessive consumption of omega-6 fatty acids can have negative effects. Therefore, their consumption must be monitored. Omega-6 fatty acids, in large quantities, can promote inflammation in the body that may lead to flare-ups of eczema, acne, and the aches and pains associated with arthritis. It is recommended that you maintain about a 4:1 ratio of omega-6 to omega-3 fatty acids.

Some research has been done that suggests improvements in overall health, cognition, sleep patterns, social interactions, and eye contact when children on the spectrum were given an EFA (Essential Fatty Acid) supplement.[8]

Furthermore, another study showed the positive effects of EFA's on dyspraxia (a motor disorder frequently associated with cognitive, behavioral, and social challenges). The participants showed improvements in reading, spelling and behavior during the treatment period.[9]

I have personally found that I really like omegas from YES™ Supplements. Dr. Brian Peskin has some very informative videos on YouTube discussing the merits of vegan versus animal based omega sources. I highly recommend you have a look for yourself. My own son, and many other children have built up to taking one tablespoon, two times a day.

L-carnitine

L-carnitine is a vital amino acid that promotes healthy neural levels of acetylcholine, an important neurotransmitter that aids memory and proper brain function. Research suggests that an L-carnitine deficiency may be implicated in a number of conditions, including ME/CFS (chronic fatigue syndrome), diabetes, Alzheimer's, dementia, and autism.

L-carnitine is composed of two essential amino acids, lysine and methionine. It is produced in the liver and kidneys and is contained in most cells of the body. It is required for the proper metabolism of fat and helps with mental concentration, and energy production. It transports long-chain fatty acids across the mitochondrial membrane so that they can be burned to produce energy.

A clinical trial studying the use of L-carnitine as a potential therapy for autism demonstrated the children taking L-carnitine showed improvements in their ability to relate to people, body use, adaptation to change, listening response, verbal communication, sociability, sensory/cognitive awareness, and health/physical behavior. There are a few articles that may be helpful for a better understanding of the importance of this amino acid. Check out an article by Henke Schultz, "L-carnitine helps kids with autism, study finds" and another by Emily Singer, "Defects in carnitine metabolism may underlie autism."

The average dose for L-carnitine is anywhere from 250-1000 mg per day with food, depending on weight and reactions.

GABA

GABA (gamma-aminobutyric acid) is an amino acid and a neurotransmitter (a type of chemical responsible for carrying information from one cell to another). It is produced naturally in the body, but is also widely available in supplement form. Manufacturers claim that GABA supplements can help boost the brain's GABA levels and, in turn, treat anxiety, stress, depression, and sleep problems.

The GABA system acts as something of an information filter to prevent the nerves from becoming over stimulated. It has long been suspected that this filtering process is compromised in many autistic children. Impairment of the GABA system could overwhelm the brain with sensory information, leading to many of the behavior traits associated with autism. GABA is also believed to play a key role in the early development of the brain.[10]

GABA is also involved with the production of endorphins in our brain, which make us feel positive and upbeat. GABA can reduce stress, relieve anxiety, and increase alertness. GABA can be helpful for behavior, language, and possibly even seizures.

GABA is used for control of seizure activity and works on the same receptors as the drug Keppra®. GABA is also useful for speech, helps facilitate language in nonverbal children, and improves language in children that are beginning to speak.

If you are going to give GABA to your child, the max is 5,000 mg per day, divided into two doses, one in morning and one at night, always on an empty stomach. In the evening, GABA can be given 15 minutes after your last dose of Chlorine Dioxide, with your probiotic. I start at 250 mg two times a day, one in the morning and one at night. I then titrate up the dose daily or every other day if everything looks good. If you observe sleepiness or irritability it is a good idea to back down.

5-HTP (5-Hydroxytryptophan)

5-HTP, also known as oxitriptan, is a naturally occurring amino acid. It is a chemical precursor to the neurotransmitters, serotonin and melatonin from tryptophan. 5-HTP works in the brain and central nervous system by increasing the production of the chemical serotonin. Serotonin can affect sleep, appetite, temperature, sexual behavior, and pain sensation.

From *Saving Eli: One Family's Struggle* - Vitamin Research Products website:

> While not clearly understood, researchers know that serotonin pathways are disturbed in autism, contributing to sleep disorders and mood. Tryptophan has been shown to help but has been banned by the FDA since 1989. We found that 5-HTP helped immensely, calming tantrums and increasing communication with our son.[11]

5-HTP is used between 50 mg and 200 mg, divided into two doses, morning and evening with food. I have seen it help children with autism for attention, focus, sleep, and cravings.

L-theanine

L-theanine or gamma-glutamylethylamide or 5-N-ethyl-glutamine, is an amino acid commonly found in tea. Theanine is able to cross the blood–brain barrier, and is reported to have positive effects on mood, stress, and cognition.

The following is excerpted from the L-Theanine website:[12]

> Many people experience stress on a daily basis and are looking for natural and safe ways to manage it. For thousands of years, it has been suggested that drinking green tea will make one relaxed. Recently this relaxation effect was found to be true and works because of an amino acid called L-Theanine which is found in the tea. Clinical research will show that ingesting up to 200mg of L-Theanine will promote the creation of the very important "anti-stress" neurotransmitter in the brain called GABA or gamma-aminobutyric acid. This results in a relaxed, clear and alert mental state.
>
> When used in the correct amounts in supplemental form, L-Theanine may:

- Reduce Stress
- Reduce Occasional, Simple Nervous Tension
- Promote Relaxation without Drowsiness
- Promote Mental Clarity and Focus
- Promote Positive Mood
- Promote Alertness
- Promote Learning and Memory
- Help Prevent Jitters Caused By Caffeine

I have found that when we can reduce these symptoms in individuals with ASDs we can get an increase in speech, learning, focus, concentration, and attention. L-Theanine is given with GABA upon waking, and can also be given at bedtime if needed, always without food. You can work up to 200 or 250mg/day.

Pycnogenol®

Pycnogenol® (pic-noj-en-all) French maritime pine bark extract acts as a potent blend of antioxidants, it is a natural anti-inflammatory, stimulates generation of collagen and hyaluronic acid and help with natural dilation of blood vessels by supporting production of nitric oxide.[13]

Pycnogenol® is not a supplement I tend to use very often, because it is an antioxidant, and therefore cannot be combined with chlorine dioxide. However, in some people it has proved necessary and helpful for speech and reduction of seizures. It should be given in the morning with food at a dose between 25 and 400 mg/day depending on need. Pycnogenol® may not be suitable for children who are sensitive to phenols.

L-Carnosine

L-Carnosine is a dipeptide of the amino acids beta-alanine and histidine. It is highly concentrated in muscle and brain tissues.

L-Carnosine is believed to stimulate the frontal areas of the brain, resulting in overall improved levels of functioning. More and more research shows that the frontal lobes and temporal lobes in the brain control emotion, epileptic activity, cognitive, expressive speech, and abstract thinking. Studies have shown Carnosine to improve language, socialization, and overall level of functioning in individuals within the autism spectrum. It has also been shown in studies to have anti-seizure properties without the side-effects of prescription anti-seizure medications.

Research indicates it prevents the formation of the beta amyloid plaque that is found not only in neurological conditions such as Alzheimer's, Parkinson's and autism, but also in the eyes in degenerative eye conditions and in the pancreas in diabetes. It is a neuroprotectant, with a study indicating it reduced severity of damage in stroke patients.[14]

Chez and coworkers found that after eight weeks on L-carnosine, children showed statistically significant improvements on the Gilliam Autism Rating Scale (GARS). They relate this to the likely ability of L-Carnosine to enhance neurological function, perhaps, in the enterorhinal or temporal cortex.[15] This enhancement in neurological function has led to speech in many of the families that I have worked with.

I have also seen L-Carnosine help reduce seizures in some children and increase language in others. It is given twice a day with breakfast and dinner at a dose of 200 to 400mg, 2 times a day. If you note hyperactivity, remove it.

Taurine

Taurine is an amino acid, a chemical that is a required building block of proteins. Taurine is found in large amounts in the brain, retina, heart, and in blood cells called platelets. It can be consumed through eating meat and fish.

Autism and Low Taurine

"Are You Dangerously Deficient in Taurine"[16] explores possible problems associated with a taurine deficiency. Among the problems noted in the article is autism and low taurine levels. In the article, Leonard Smith MD writes about the benefits of taurine, some of which can be of interest to people dealing with autism spectrum disorders:

- Brain and nervous system function
- Helps eliminate toxins
- Stabilizing the brain (can be effective in treating seizure disorders)

Taurine, like all amino acids, must be given without food. The maximum dose is between 500 and 1,500 mg.

DMG

DMG, or dimethylglycine, is an amino acid that can be found in many common food items like meats (especially liver), various grains, and beans. It has been classified as a type of amino acid that is closely linked to vitamin B.

There have also been interesting developments in the use of a DMG supplement for children with autism. Because many children who have autism are unable to tolerate eye contact, and some have problems forming complete sentences and thoughts studies investigated whether an increase in DMG consumption could help to alleviate many of these problems. In fact, the study found that when children with autism were given supplemental DMG they appeared less frustrated and showed a marked increase in their speaking and cognitive abilities.[17] At present, this research remains novel and therefore further investigation is necessary to determine what the long-term effects of DMG supplementation might have on children on the spectrum.

The following paragraphs were excerpted from *Defeat Autism Now!*[18]

> ### Dimethylglycine (DMG) for Autism
>
> For over 20 years ARI has been hearing from parents who have tried DMG on their autistic children. In many cases remarkably good results have been seen, especially in enhancing speech. In some cases, drug-resistant seizures have been stopped by DMG. (See New England Journal of Medicine, 10-21-82, pgs 1081-82).
>
> There is an extensive research literature on the safety and health benefits of DMG. Many studies have shown that DMG enhances the effectiveness of the immune system, improves the physical and athletic performance of humans and other animals (e.g. race horses) and has, all in all, a very wide range of beneficial effects. It is very safe. I have seen no evidence of any toxic or significant adverse effects.
>
> Many parents have reported that, within a few days of starting DMG, the child's behavior improved noticeably, better eye contact was seen, frustration tolerance increased, the child's speech improved, or more interest and ability in speaking was observed.

A full dose of DMG is 900 mg per day, taken without food upon waking. However, it is best to start at a lower dose and slowly work up to 900 mg over a week or so. If you see an increase of hyperactivity (which is rare), reduce the dose. If no improvement is seen within a month I would switch to TMG.

TMG - Trimethylglycine (Betaine Anhydrous)

TMG (Betaine anhydrous} is a chemical that occurs naturally in the body, and can also be found in foods such as beets, spinach, cereals, seafood, and wine.

How does it work?

A form of betaine called betaine anhydrous helps in the metabolism of homocysteine, a chemical involved in the normal function of many different parts of the body, including blood, bones, eyes, heart, nerves, and the brain. Betaine anhydrous prevents the buildup of homocysteine seen in people who have problems with its metabolism from birth.[19]

The following is excerpted from the *Autism Canada Foundation*.[20] Please visit their site...

www.autism.org

...for full references contained within the text.

The benefits of taking DMG or TMG range from behavioral changes, reduction of seizures, and decreased obsessive-compulsive behaviours to improved language. DMG and TMG have been reported from thousands of families to be quite beneficial to many individuals with autism.

Research on humans and laboratory animals has shown that DMG and TMG enhance the effectiveness of the immune system. Some children and adults with autism have seizures, and there are published reports of decreases in seizure activity as a result of DMG. A double-blind placebo-controlled study by Drs. Shin-siung Jung, Bernard Rimland, and Stephen M. Edelson involving 84 participants documented a significant decrease in behavioral problems.

It should be noted that some kids tolerate DMG but not TMG. TMG is given upon waking without food, in a dose of 500mg. If DMG didn't yield improvements in language, then we switch to TMG.

FAQ's

There are times when I do see undigested food in my son's stools. I am particularly concerned. He never complains of stomach pains. But I do see some at times, especially cashews. Should I bother with an enzyme? Or is it better without one. If so, which brand?

> I love enzymes. Kirkman has one with Isogest®, 851/180 is the number, and it's broad spectrum. Also, Biofilm Defense® is great for dissolving the biofilm. A number of the parents online have used Ness® enzymes Gastric comfort formula #601 very successfully.

How should I administer THERALAC®?

The best time to administer probiotics is at bedtime. Children's THERALAC® is a granular formula so you can sprinkle it on yogurt or mix it into a smoothie and still obtain the same benefits! So how do you take it? We recommend that you take a level 1/4 teaspoon and fold it into yogurt, applesauce, or food of a similar consistency and let it sit for a minute. This is to keep the granules as close together as possible, thus allowing our acid proof gel matrix to form around the product. Visit the THERALAC® website for more info:

www.theralac.com/childrens-theralac.aspx

Supplement Dosing Overview

Supplement	Dose	Time of Day	Empty Stomach (ES) or With Food (WF)
Probiotics	1 cap	Before Bed	ES
Omega-3/ Omega-6	1 Tbsp	With any meal 1-3x / day	WF
L-Carnitine	250-1000mg/day	With any meal 1-3x / day	WF
GABA	Up to 2,500mg 2x/day	Upon waking & at bedtime	ES
5-HTP	50-200mg	Morning & Evening	WF
L-Theanine	Work up to 200-250mg/day	Mornings or Mornings & Nights	ES
Pycnogenol	25-40mg a day/as needed	Morning	WF
L-Carnosine	200-400mg 2x / day	Morning & Evening	WF
Taurine	500-1500mg/day	Morning, Noon & Night	ES
DMG	900mg/day	Morning	ES
TMG	500mg/day	Morning	ES
Enzymes	1 cap w/meals	Morning, Noon & Night	WF

Chapter 10

Step 5 - Chelation

So, let's say we're clipping along; full Diet, CD, Kalcker Parasite Protocol, and we've added in whatever supplements the child specifically needs and the child still has autism. Recovery is still not a reality... yet! At this point, it's time to look at chelation.

Why use Chelation for Autism?

Chelation became very popular in the world of autism about a decade ago, when things were heating up around the Thimerosal/Autism connection. Today, many of our children are still metal toxic. Metal challenges (testing that shows metals in urine) from any of the thousands of children whose families I have helped show the same thing—extremely high levels of mercury, lead, aluminum, as well as sometimes tin, cadmium, and other metals.

These heavy metals can come from various sources such as:

- Coal burning power plants. According to the EPA, coal-fired power plants in the United States emit about 48 tons of mercury into the air every year, where more than half of this mercury falls within five miles of the plant itself. When it reaches the water, microorganisms consume it and convert it into a substance called methyl mercury.
- Drinking water
- Our food supply
- Cookware
- Deodorant
- Beauty products
- Dental amalgams
- Vaccines, etc.

Heavy metals are known to accumulate in different parts of the body including organs, bones, joints, and the brain, etc. Metal toxicity can provoke inflammation, kill neurons, cause behavioral changes, affect the thyroid and other master glands, lower T-cell counts, and cause a myriad of other symptoms. Furthermore, new research shows that fluoride in drinking water makes the aluminum that we ingest more bio-available. As was reported in the journal *Brain Research*, the combination of aluminum and fluoride causes the same pathological changes in brain tissue found in Alzheimer's patients.[1]

Many children with autism also have impaired methylation cycles. The result of this impairment is an inability of the body to rid itself of excess metals, thereby prolonging chronic illness.

What exactly is methylation?

Methylation reactions are those that involve the transfer of a methyl group from one compound to another. The methylation cycle is the name given to a biochemical pathway that contributes to a range of crucial bodily functions, including:

- Detoxification
- Immune function
- Maintaining DNA
- Energy production
- Mood balancing
- Controlling inflammation

Impairments or mutations on the methylation cycle can lead to problems with:

- ASD's
- Alzheimer's
- Diabetes
- Allergies and Asthma

An overload of toxins (including heavy metals), can contribute to the impairment of the methylation cycle, and if the methylation cycle is impaired the body is unable to detoxify as needed, therefore creating a vicious catch 22.

We need to help our children's bodies rid themselves of their excess metals as part of the healing process. Dr. Usman has shown us that heavy metals are present in the biofilm, and Dr. Klinghardt has shown that removing mercury can be directly related to a reduction in chronic infections.[2]

Let's get those metals out!

What exactly is Chelation?

Chelation is a process used to rid the body of heavy metals. It is described as a chemical process in which a substance (chelator) is used to bind molecules, such as metals or minerals, and hold them so that they can be removed from the body.

Some of the most common chelators used in the world of autism are:

EDTA

This is an amino acid that attracts lead, other heavy metals, and some minerals from the bloodstream and expels these toxic elements in the urine. EDTA works to remove excess lead from the body, but it is not specific to mercury or methyl mercury as are DMSA or DMPS. It can be taken orally, by rectal suppository, or IV.

DMSA

This is an FDA approved drug that can be used in children when lead toxicity is suspected, however it can also be effective at removing other heavy metals including mercury and arsenic. It can be taken orally, transdermally, or given as a suppository.

DMPS

DMPS is given to remove mercury from the body. It can be given IV, intramuscularly, subcutaneously, transdermally, or by suppository.

Patrick has been prescribed all of these at different points of his life, and I can't say that I saw miraculous results during the time that I used them. However, some families have seen results, and this is something you may want to discuss with your doctor. After years of using chelators, and seeing other families use chelators, I have opted for a gentle approach. I like to use two products: bentonite clay baths and Bio-Chelat™. These products are strong enough to help the body rid itself of heavy metals, but have not been shown to stress the liver or provoke undesirable detox symptoms.

Knowing now that metals are in the biofilm, I feel it is short sighted to focus solely on metals or heavy metals rather than all the pathogens in the biofilm.

Bentonite Clay Baths

Bentonite clay is sedimentary clay composed of weathered and aged volcanic ash. Bentonites are more widely known as healing clays used for detoxing, cleansing, and drawing out impurities. They are used in many everyday products such as toothpaste, antacids, and cosmetics.

Indigenous people have used bentonite clay for centuries; Dr. Weston A. Price in his book, "Nutrition and Physical Degeneration,"[3] stated that when studying the diets of native tribes he examined their knapsacks. Among the tribes examined in the High Andes, in Central Africa and the Aborigines of Australia, Dr. Price reported that some knapsacks contained balls of volcanic ash clay, a little of which was dissolved in water.

Bentonite is known as "swelling clay." When bentonite clay absorbs water and swells up, it is stretched like a sponge. Toxins are drawn into these spaces through electrical attraction and bound. In fact, according to the Canadian Journal of Microbiology, bentonite clay can reportedly absorb pathogenic viruses, as well as herbicides and pesticides.[5]

One of my personal favorites has been *Even Better Now*®'s product available at:

www.evenbetternow.com

> "EBN® Cleansing Clay is 100% pure sodium bentonite clay which has the highest cation exchange capacity (CEC of 98-107 meq/100g) of any bathing clay that we tested on the market. This clay is high purity air-classified sodium bentonite, selectively mined, consisting of micronized particles, which is a free-flowing powder. EBN® Cleansing Clay is 100% pure, hypoallergenic, and free of viruses, bacteria, yeast, and mold, as well as having a high cation exchange capacity."

There are other quality clays on the market, but EBN® tests every batch for heavy metals when it comes in. No, I do not receive financial gain from the company.

Bio-Chelat™

Another product I really like for gentle (low and slow) chelation is Bio-Chelat™. This is a German product containing a minimal amount of EDTA, which is FDA approved as a food substance. According to a clinical trial carried out in Germany:

> The therapeutic value of the Bio-Chelat™ in the context of other chelators that are currently on the market is seen as follows: Chelators work relatively fast, but they are also very strong with a relative high washout of important trace elements and a high degree of specific side effects. Bio-Chelat™ works much gentler than most common chelators.

> Although during the treatment a significant decrease of the body's heavy metal ion load was seen, this is accomplished without greatly disturbing the mineral and trace element relationships. The reduction of zinc should be looked at with caution and may easily be corrected throughout the treatment.[5]

I have seen very positive changes in many of the children who added in either bentonite clay baths, Bio-Chelate™, or both. As with any other intervention, add them separately while closely observing your child. Keep a log to record any changes, progress, or reactions you observe. Even if you don't see anything right away, I would keep using the products for a minimum of three months. Again, since they work LOW and SLOW changes may not be apparent at the outset, but that doesn't mean they are not helping to heal your child's body.

I use the CD baths every other day and then two times a week I do bentonite clay baths. I don't do the CD bath and the bentonite clay bath on the same day. I do them at opposite ends of the week.

Three days or so later, we'll add in the Bio-Chelat™. Again, follow the instructions on the package, and you can always start low and work your way up. Since Bio-Chelat™ doesn't alter the CD, the drops can be added to a single dose of CD. In addition, since the drops have no flavor, it is suitable to add them to water or any other drink your child might consume throughout the day.

Common Errors in Chelation:

Urine analysis without provocation:

Many families will do urine analysis to see if there are heavy metals present. If you have not done any chelation (oral or intravenous) you really don't get a good picture of what kinds of metals are actually in the body. Simply because the metals don't show up in urine, doesn't mean they are not present in the bones, the intestinal tract, the brain, and organs, etc. If we don't do any sort of provocation, we may end up with a false negative, leading a family to believe that their child is not metal toxic, when in fact they are. Bentonite clay baths and Bio-Chelat™ are gentle enough that they will not cause any harm to a child (metal toxic or no) but can greatly help reduce the heavy metal load, as well as other toxins in the body.

Healing the gut before starting chelation:

A doctor once told me that we had to heal the gut before we could start the chelation process. However, in the children that I have seen recover, probably up to the last month before they recover, their gut might still not yet be perfect. For example, one of the little girls that recovered continued to suffer from constipation at 30 days before she lost her autism diagnosis. However, at this stage something seemed to switch, she started having regular bowel movements and within 30 days she lost her diagnosis.

Basically we're healing the intestines as we're chelating, as we're healing everything else. They are all parts of the puzzle. I would never put chelation before the CD or parasite protocols. However, it's definitely something that's part of our road when, after these other approaches, children have not yet recovered. At some point after we have the CD, ocean water, parasite protocol and some supplements is when we can add in chelators. It's not like I'm saying, "Oh we'll talk about that in a couple of years." We want our children recovered in less time. It's like a marathon, we keep the pace, never lose our heads, or break into a sprint.

Do not chelate with constipation:

If the child suffers from constipation, it's really important that we start moving the intestines when we're using chelators. There are a few options open to us for moving the intestines. I have found that the most effective is through the use of CD enemas. If your child isn't having a daily bowel movement, I would do an enema once a day. If you absolutely can't do enemas, get that child in a CD bath. We're finding that 50 to 100 drops in a bathtub is allowing children who previously had constipation to have bowel movements.

There are a few children/adults on the protocol that will not do enemas, or the parent will not. Whatever their situation may be, we have to get the bowels moving again. In this case, we do 50 to 100 drops in their bath water along with their max oral dose every day. The result has been that they're having regular bowel movements. That's one way we can move it along and then you can keep chelating. It's really important to have a daily bowel movement with any kind of chelation because otherwise the toxins are simply reabsorbed through the intestinal wall, whether they're heavy metals or other pathogens.

IV chelation is not the only means by which to cure autism:

I think at one point between 2006-2008, we were all focused on the problem in the world of autism being the heavy metals, because there was mercury/Thimerosal in the vaccines. We had to get the mercury out of our children. Well, part of the problem is they're not detoxing. As mentioned before, methylation cycles are impaired in children on the spectrum. I met a young girl whose parents were both allopathic doctors and they really believed at that point, as most of us did, that heavy metals were the problem with autism. If we could get the heavy metals out then they would recover. The parents did 122 IV chelations and their child never improved. At this point I have lost touch with them, but these are things that we have to learn along the way. I know now that it is NOT just about chelation, and why metals are NOT the only important piece. All of the pieces play a role and likely interact to cause autism in our children. There is a symbiotic relationship between the viruses, bacteria, candida, and parasites that exist in the biofilm. Since CD can break down biofilm and kill the pathogens living in it, as well as neutralize heavy metals, it has been a very powerful tool for us.

Doing a blood test to look for heavy metals:

Generally speaking, within three days of heavy metals entering the bloodstream, the body will deposit the heavy metals into tissues, organs, and eventually bones. For this reason, a blood test is not the best place to look for heavy metals in children with autism. Unless the child has very recently been exposed to heavy metals, we are unlikely to see elevated levels of heavy metals through a blood analysis.

The articles at the end of this book (Appendices 13 & 14) will provide further insight as to why chelation can be so important for our children on the spectrum.

Healing
the symptoms known as
Autism
Kerri Rivera

Luca

Chapter 11

Step 6 - Hyperbarics

by Bob Sands (Introduction by Dr. Beatrice Golomb)

> *"What Oxygen is to the lungs, such is hope to the meaning of life."*
> ~ ***Emil Brunner***

We were fortunate enough to have the amazing Bob Sands write this chapter on hyperbarics explaining its importance in healing ASD's.

Note: Throughout this section of the book, the author, Robert Lyne-Sands uses the actual names of the people with whom he came in contact with. In some instances it is with their permission, in others it is not. The accuracy of quotations, opinions expressed within this chapter, and the names used are solely the responsibility of Robert Lyne-Sands. Great care has been taken to remain within the bounds of current journalistic integrity and practice.

Robert Lyne-Sands new book, *OXYGEN: The First Medicine, Volume 1 CANCER*, is expected to be released January 2014.

Robert Lyne-Sands author of the Hyperbaric section of the book, with one of his young patients, getting ready for a session.

Robert Sands is one of only about five licensed manufacturers of hyperbaric oxygen therapy (HBOT) machines in the US, with extensive experience in clinical hyperbaric treatment, whose expertise will be an exceptional asset to this project. He has designed and manufactured HBOT machines for the US, Australia, Britain, Thailand, and many other nations, for naval/military use, civilian diving use, and medical use. He has the theoretical understanding of HBOT therapy that derives from intimate involvement in the process of designing and manufacturing these machines. He has the practical experience that derives from running a set of HBOT centers, and personally overseeing greater than 20,000 hours of HBOT sessions over 35 years.

~ Dr. Beatrice Golomb, MD, PhD.
Professor of Medicine. UCSD 2011.

Beatrice Alexandra Golomb, MD, PhD (PI) is singularly qualified to lead this study. She has expertise in GWI including authorship of four RAND reports related to GWI29-31, and work in the scientific literature. She has been a member, since its inception, of the Department of Veterans Affairs Research Advisory Committee on Gulf War Veterans' Illnesses, and served as Scientific Director of this Committee. As a licensed internist, she is a primary care physician for a panel of ~280 veterans, including GWV, at the VA San Diego Healthcare Center.

Recently, I sought the answer to something that was puzzling me. My friend, Dr. Bernard Gunter, is a revered and renowned psychologist in his eighties and, if anyone would know the answer, Bernard would.

"Does God, Great Spirit in the Sky, The Universe or whatever unseen deity you worship give you freedom for personal choices or is everything predestined?"

Bernard's eyes twinkled. He was never in a hurry to answer a question. In my mind, Dr. Bernard Gunter is sort of the American version of the Dali Lama.

He thought for a moment, selecting his words carefully, and then replied, "Funny you should ask that—just the other day I was talking with Eckhart Tolle and asked him that very same question. Tolle replied, 'God allows you to think that you have free will and free choice—and you do, by and large. However the roadmap is laid out for you beforehand.'"

Just as I suspected.

Dr. Bernard Rimland, Kerri Rivera, and the Hand of God.

On January 20th, 2006, a tall, glamorous woman walked into my San Diego clinic; her name was Kerri Rivera. She had brought her small five-and-a-half-year-old son, Patrick all the way from Mexico for hyperbaric oxygen therapy. The little fellow was absolutely gorgeous, but had a vacant stare and made no eye contact with any adult, rarely even with his mother. Day after day Patrick spent time in the chamber.

Kerri explained to me that Patrick developed normally for almost the first two years of his life. Then, to comply with vaccine regulations, Patrick received the DPT, HepB, and Influenza B (Pentavalente) vaccines in one day. Patrick fell into a complete zombie-like state after the injections. By the time Kerri arrived in San Diego, she had already done her homework. She blamed the vaccinations and had already undertaken the task of healing her son.

After her own experience with autism, Kerri wanted to help people and she was already doing the DAN! Protocol, etc., as Dr. Rimland has done so much for the children and she had a lot of respect and admiration for him and his work.

One day, I walked into the clinic and told her, "We're having lunch with Bernie tomorrow. Get a babysitter." At that point, totally shocked, she said, "Oh my God, yes, but I don't have anything to wear!"

"Trust me, what you have on would be just fine." A flustered Kerri rushed off to purchase some dressy clothes—she was, after all, a woman originally from Chicago. The very next day, we were sitting with Bernie and his wife Gloria in a quaint little restaurant, *The Green Tomato*. If the reader wishes to see that day, just go to your computer and look up Bernie on Wikipedia and there you will see Kerri in her black business suit (http://en.wikipedia.org/wiki/Bernard_Rimland).

Over lunch, Bernie and Kerri bonded immediately. None of us at the table realized that Bernie was suffering from cancer and had less than a year to live. Driving the short distance back to my clinic I made the observation to Kerri that I suspected she was about to become a scientist, and no longer just a trophy wife. "Be careful," I warned her, "Bernie Rimland's mantle will be heavy on your shoulders."

Kerri turned her face toward me and said, "I will carry that mantle willingly and with honor."

After a few days, Bernie told me that he thought Kerri had what it takes to carry the torch and help families heal their children with autism. I saw a lot of Bernie in the next nine months. He would come down to the clinic for regular hyperbaric treatments because it eased the pain of his cancer. Just before his death, he told me, "I will not die before I have found the reason for and defeated autism." But die he did on November 21, 2006, at the age of 78.

As I drove to Bernie's grave, I anticipated many, many people lining the way with some hundreds even thousands thronging the memorial park, similar to Giuseppe Verdi's funeral, in 1901, where 200,000 people gathered. At his graveside, the thousands of mourners spontaneously burst into the "Slaves Chorus" from his Opera, Nabucco.

I was stunned upon arriving. Probably less than thirty or so weeping caregivers of children with autism were present at the burial, along with Gloria and Bernie's three children. WHERE WERE ALL THE PEOPLE, THE MOTHERS AND FATHERS, THE RESEARCHERS AND SCIENTISTS? I asked myself. It was my first Jewish funeral and I think that the booming sound of the shovel full of earth that I threw down onto his plain casket made me look heavenward and ask, "God, where are you now?"

I drove home, eyes unfocused with tears, and stone eggs of the birds of unhappiness in my heart, stopping only to purchase two fine bottles of wine to toast Bernie. I knew one would not be enough. Arriving at my house, I found the fine crystal goblet my Mom had left me, filled it, and played Neil Diamond's, Morningside over and over, with each playing, another glass, held high to Bernie...

Morningside
The old man died
And no one cried
They simply turned away

And when he died
He left a table made of nails and pride
And with his hands,
He carved these words inside
'For my children'

Morning light
Morning bright
I spent the night
With dreams that make you weep
Morning time
Wash away the sadness
From these eyes of mine
For I recall the words an old man signed
'For my children'

And the legs were shaped with his hands
And the top made of oaken wood
And the children
That sat around this great table
Touched it with their laughter
Ah, and that was good

Morningside
An old man died
And no one cried
He surely died alone
And truth is sad
For not a child would claim the gift he had
The words he carved became his epitaph
'For my children'

In fact, as I sit here writing this, I play Diamond's melody and listening to the words makes my heart swell and my eyes become damp.

The morning after Bernie's funeral, hung-over and miserable, I pondered: Did my beloved friend Bernie die feeling that he failed? I now doubt it for he had passed his mantle to carry on his search for the cure without fear, without the need for accolades and rewards and to do so with integrity. I did not know it, but before his passing, Bernie and Kerri had been in communication. Bernie had given her permission to translate his *Defeat Autism Now!* Protocol into Spanish, Kerri's other language, as well as other tasks.

In review, I now see God's Hand at work—and it started one afternoon at the Green Tomato. Thus, my question to my friend on free will versus predestination right at the start of my notes on this section.

Brain overload and frozen fingers.

At the time, when Kerri asked me to write this section of her book about hyperbaric oxygen therapy and the treatment of challenged children, my mind went into overload. Why? Kerri was (in my view and that of others) the most efficient and effective health scientist with a high normalization rate of challenged children (autism, etc.). With her own hospital-quality hyperbaric oxygen center, Kerri had first-hand knowledge that to normalize such an afflicted person required a lot more than some trips in a oxygen chamber or some Silver Bullet response from a pharmaceutical company.

My brain whistled like an express train rushing through a tunnel. I thought of the dozens of children that hyperbaric oxygen therapy has helped at my different centers. Some of the little guys were so badly damaged they would repeatedly punch at their own head, or chewed on their own wrists. I'm thinking of Marco as I write this. Ten years later, he still punches. A clearly defined casualty of a hotshot thimerosal vaccination, normal until three, his parents won a multimillion-dollar lawsuit against Big PHARMA. There has been some progress over the years; but it will be a matter of unrelenting attention to Marco.

I remember another little guy; a handsome, articulate smiling boy of about 12 years of age named Chase. Meeting him and his parents, you would never know he had autism or ADHD. Chase was a "runner." When his treatment in the chamber was over, he would promptly disappear out of the clinic doors. Eventually, Chase would be carted back down by one of the upstairs nurses. "Chase" by name and chased by nature.

After one such session in the chamber, he rocketed out and into the restroom of the clinic. I was sitting at the controls when he came back. Chase climbed up onto my knee and said, "I love you Bob," and he rubbed his hands across my face. From the terrible odor I knew in an instant what was all over my face and my lips. Chase had used his hands instead of toilet paper and bypassed the washbasin on the way out.

"Oh goodness," came from his horrified mother. "He got you. I am so sorry. It is one of his favorite jokes."

"Not to worry," I said, "I have eaten and tasted this in a variety of forms most my life." So, whenever I hear somebody say, "Sands is not a Doctor so he doesn't know s—t from Shinola," I think of Chase.

Normalizing kids is so rewarding, but you have to take a chance on s**t.

My fingers were not ready for the keyboard as I tried to wade through dozens of experiences with these children, my own personal experiences, and finally adding input from experts in the field of pediatric abnormalities. (Many of these experts became my close, personal friends as the years went by.)

For any parent or caregiver to delve into all of the opinions and conclusions of the so-called experts, it is a bewildering journey through a lot of contradictory and self-serving educated guesses and erroneous assumptions. To arrive at something that the reader will identify with, and believe in, I have to go way back. After 70 years on the face of this planet, I have reached certain conclusions about the treatment of children that enables me to write this section. Mind you, I'm not stuck in my ways and I am open to change my thoughts and conclusions as the science improves on treating children with autism.

Little Patrick Rivera, Little Bobby Sands Something in Common?

Here is something that I doubt my adult children even know about. In 1947, polio was raging throughout Australia. I was seven years of age and had developed a sore throat and fever. My father, a wealthy man, called the very best doctor to my bedside—yes physicians did make house calls in those days. I can even remember the doctor's name, Dr. Bradfield. His father had been the engineer that designed the Sydney Harbor Bridge. My dad, with all of his money drew a lot of water in those days and could get almost anything he wanted. Bradfield came in and examined me briefly; after all he had other house calls that night.

Bradfield told my Mom and Dad that he did not think I had polio, but perhaps diphtheria and that I should spend time in isolation in the infectious disease hospital overlooking Botany Bay, "just in case." The next morning an ambulance came and, in spite of my fear, pleading, misery, and tears they carted me away. The three-hour journey became a six-hour journey. The ambulance broke down outside the Sydney airport. I remember looking out of the rain spattered window and seeing the aircraft clearly visible over the fence. We sat on the side of the road for almost two hours before the replacement ambulance came to transport me, and my journey continued to a huge colonial building complex, built by convicts and perched on rolling, green lawns that looked down onto the Pacific. The complex was known as, The Coast Infectious Disease Hospital, which later became Prince Henry Hospital. I was terrified and clutched my teddy bear, *Threadbare*, under the blankets. Threadbare became my only friend for the next four months, and a glimpse of airplanes over the fence became an enduring memory of the trip.

My parents were already there and had obviously filled out all of paperwork. I never got to say goodbye to them since it was an infectious disease hospital. As they walked past a big hopper window, they paused momentarily and Mum waved through the window. I saw them maybe three times in the next four months when the nurses wheeled my bed up against the window that looked out into the hall. They would tap on the glass and smile at me but because my father was quickly bored, after about five minutes they would wave and be gone for another couple of weeks. It was a long way for them to travel.

The night of my arrival, a team of doctors and nurses arrived and they rolled me onto my side. Full syringes lay on a tray; in those days they were glass with little thumb and finger holes on each side and each needle was reused and supposedly resharpened before their next use. This meant that the needles were thick and long. One-by-one these needles were inserted into my upper thigh on the left-hand side. The pain was agonizing and I'm sure I screamed and screamed. I remember at least three nurses holding me down. At seven, I was just a little fellow. There had to be at least 20 of these very large syringes. Checking the history books, I believe that they were mostly penicillin and sulfa drugs. I might add, it was already known to Doctor Bradfield that I was highly allergic to sulfa and remained so for my entire life.

The next morning, when I awoke, I found that I was paralyzed from the feet up to just under my arms. Threadbare was my only friend for the next four months. The days all blurred together. One thing I do remember vividly is that I used to deliberately wet the bed. I did so only when a lovely red-haired nurse with a band of freckles across her nose was on duty. This nurse was kind and gentle, probably in her twenties. After she had changed the sheets, she would stroke my cheek and tell me that everything was going to be okay and not to worry about wetting the bed. The other older nurses scolded any of the other pediatric patients if they had an accident. At some point later, my gentle nurse told me that I really did not need to wet the bed to get her attention. I soon stopped and I still got a cuddle from her. Such a display of care and affection was unknown to me from my parents back at home.

After that stay in the hospital, I was never the same. I know that there was constant discussion amongst my family members that I should be placed in some sort of an institution because my behavior was "contrary and stubborn." I was either talking my head off, trying to make my family love me, or distant, and as such, I was said to be "uncooperative." Soon after, I was placed in a Hogwarts-style boarding school.

Could that childhood experience have made me what I am today, forever alone, but rarely lonely? As the years ticked by, to my surprise (and I think everybody else's), the school counselors found that I had a genius IQ. Amongst the 1,200 boys, I was second from the top.

After seventeen years of treating so-called challenged children, I have yet to find one that was not as smart as a whip in something.

I managed to burn every social bridge at the boarding school, thrown out of every special social club at Newington. On Saturday afternoon, after mandatory sports such as Cricket or Rugby, the other boarders would get to do things they wanted like billiards, photography, or gardening (these were clubs that you had to apply for and join). However, I had been expelled from each of them so I wandered alone each Saturday afternoon and finally picked the lock on a large room-sized cage full of Sulfur-Crested Cockatoos and other Australian parrots. I would sit alone with them and be jeered at by the boys in the garden club that were planting and weeding. Within weeks, these birds would happily climb all over me, enjoying my company. The birds loved me, but the boys and masters did not. To me, it was a two-way street. Animals made better friends than people. Eventually some of the other boys would come and beg to be allowed to accompany me into the cage so they too could have that same privilege.

Do not get the wrong impression of my beloved Newington College in Sydney Australia; even today I dream of the sanctuary of that place. It was better than what I had with my family back home.

The really curious thing is that my mind had blotted my hospital stay out, or had hidden it in some black cabinet in my brain where the really bad things are kept to allow me to get on with the reality of NOW. I remembered the illness and isolation only when Kerri asked me to write this. The details have returned with clarity: the old iron beds, the smiling face and freckles of my favorite nurse, being cold at night, the colors on the wall, the distant sight of the ocean through a window on the other side of the big ward, the frame of the window so far away that it looked like a painting hanging there. If I close my eyes as I write I can see it now.

I now ask myself, "Is this the way children with autism feel?"

I don't know. What I do know is that my old friend and confidant, teddy bear, is sitting on the desk as I write this. I reach out to rub Threadbare's head and a great calmness washes over me. The little fellow is 64-years-old now and there is a great competition amongst my adult children to see who gets him after I die.

Botched diagnosis
The incorrect pharmaceutical response.

With almost 40 years of working in medicine behind me, and now in my retirement, I do know that neither diphtheria nor polio were the problem. The suffocating "Bull Neck" or the "Strangling Angel of Children," diphtheria, had not happened to me, nor did the withered limbs of poliomyelitis set in after paralysis.

Simply put, Dr. Bradfield cocked it all up in a big way by guessing, not taking enough time and arriving at the wrong medical response on his visit. The relentless attack on me with sulfa drugs caused paralysis, which was probably what is commonly known today as Guillain–Barré Syndrome, a disorder affecting the peripheral nervous system. Ascending paralysis, weakness beginning in the feet and hands and migrating towards the trunk, is the most typical symptom; the disease is usually triggered by an infection.

It sure sounds similar to what happened to little Patrick, and probably to the reader's challenged child as well.

Migrating to America in 1990 with my tribe of offspring.

I arrived in the United States over twenty years ago. If the reader is interested enough, a quick look at a funny but classy "roast" presentation is on YouTube. You can find it by searching "Robert Lyne-Sands" on YouTube.

After completing the United States Navy transportable chamber development project, I arrived in swank, Pacific Palisades, on the outskirts of Los Angeles; living amongst all of the Hollywood greats. I was going nuts with boredom and became the "Bob Clampett" of the area; jogging down the leafy streets I always said Gudday to any celebrity. I rarely remembered their names but knew their faces. They never replied on the first go past, but then I would jog backwards, keeping pace with them (big stars like Madonna always ran with a group of minders) and I would yell out "I said Gudday. My goodness, your manners are appalling." I might add, it was not just me; they never ever nodded to each other when their paths crossed either.

Around this time, I opened the very first freestanding hyperbaric oxygen therapy center. It was more or less hidden away in Santa Monica on 6th Street. Tucked away in what used to be an old motel, you could almost refer to it as a "speakeasy" center, akin to something from the days of prohibition, where you had to knock on the door three times before being let in. I would only treat people who had been referred by another patient. Most of these people were from the entertainment industry in Hollywood, many with the HIV virus, and lots of folks with chronic fatigue syndrome.

The kiss of death in Hollywood is to be known to be ill. Hyperbaric oxygen therapy works so well that the word soon got around, and I was inundated with patients, some famous and well-known, and others just hardworking folk in the movie industry. Thus, I ended up being known for some years as "Mr. Oxygen," and because of the remarkable effects of HBOT, I befriended many of these celebrities, who subsequently invited me as their guest to many of the big Hollywood functions (the only one that I missed was the Academy Awards).

One of the first of these talented people that became my close friend was Harold Michelson. Harold was an award-winning art director working out of DreamWorks. He was suffering the aftermath of a mild stroke. His wife Lillian had told me that her Harold loved crossword puzzles, but since his stroke he would put the across answer in the down column. This caused him immense frustration since his answer was usually correct, just in the wrong place. Five treatments in the little chamber were all it took for Harold to get his verticals and horizontals in the right place. The big thing that he noticed was that his blood pressure stabilized, his lung problems normalized, and he needed less medication for his diabetes.

Harold still stood well over 6-feet–tall, though stooped with old age. We hit it off immediately because Harold had a quick wit and loved a good laugh. The stories and banter between us were constant. Harold was extremely talented yet self-effacing. He really was a hero in the truest sense of the word. As young man, he made over 40 flights across the English Channel to bomb Germany. As navigator, Harold sat in the nosecone of the B-17 and either froze with the cold on the trip or froze with fear when making the lower-level bombing run. "Bob, I was so frightened of the flak from the antiaircraft fire from the Germans on the ground. As the puffs of smoke would erupt around the front of the aircraft, shrapnel would pierce plexi-glass windows in front of me. I would hold the navigational map in front of me and try to believe that the thin paper would stop the lethal fragments."

The first challenged child to be treated in a hyperbaric chamber—Emily.

I need to jump forward in time for a moment before getting back to Harold Michelson and his wife, Lillian. I went back into production of hyperbaric chambers after I realized the great need for hospital-grade dispensing of oxygen to the public. There were no rules preventing me from doing this even though I'm not a physician. All my patients needed was a prescription from a registered physician; that was the easy part. My first center was in San Diego.

In the first few months of patient treatments we received a call from a father in Canada. His daughter, Emily, had cerebral palsy. The father had read that hyperbaric oxygen therapy would perhaps help his little girl normalize, and so when he asked me what I thought I gave him my usual answer, "I do not know, I am not a physician." And then I added, "I would hate for you to waste your money on a trip from Canada to San Diego if it doesn't work."

The father replied, "I have two choices. I can go to England where there is snow on the ground or I can come to San Diego. It is an easy choice to make."

So father and daughter duly arrived. Emily was the most delightful child. When she walked through the door she had a crash helmet on, knee guards, and wrist-guards to prevent serious injury when she would crash, face first, into the floor. We learned a lot in the first ten days of treating Emily. One of the first things was the fact that fathers can never properly do a good job on their daughter's hair. I felt for the father since Emily had lots and lots of hair and was never still for a moment (I have a few daughters myself). Emily would arrive looking like a rag doll; her little blonde head would have pins, bits of string and ribbon through it. So the very first task of our nurses would be to grab Emily (usually tearing around the clinic), and comb her hair out to make her more presentable.

Now, the really big thing that we noticed was that on day ten she arrived at the clinic without any of her body armor. Yes, she still "toe-walked" because her heels did not touch the floor. To compensate, she would park her little fists against her chest, push her elbows out to keep a center of gravity, and rush around the clinic. On day 11, father and daughter spent five hours at the San Diego Zoo without the usual body protection. Emily never lost balance or tumbled even once. Follow-up visits over the last 17 years find that Emily is living a normal life. Yes, she still has some walking problems, but never needed to undergo the mutilating surgery that was recommended to her by her Canadian doctors. During 20 days of hyperbaric oxygen treatments in San Diego, something had changed in Emily's brain.

Likely the most important introduction in my life... The Great-Spirit-in-the Sky's (God's) hand at work.

Listening to the improvements that little Emily had made, Harold Michelson was both enchanted and excited.

"Did you know that Lillian and I have an autistic son? We need to introduce you to a couple of people in San Diego. Dr. Eric Courchesne, who runs the autism research center for the University of San Diego, known as the Autism Center of Excellence or ACE."

Lillian chimed in and said, "I really must call them about you. Eric is a gem and his wife Rachel is also a research doctor. They work side-by-side and will be so helpful."

With Emily's positive results, and more children making gains at the center, I called the Courchesne couple at the ACE and offered to provide them with the use of a $150,000 hospital-grade hyperbaric chamber for one full year. There was no response to my offer.

While you may imagine that the UCSD-ACE Research Center is located in some lofty center of learning, it is in fact to be found in a delightful millionaire's village known as La Jolla in Southern California, a block or so from the Pacific Ocean. On the top floor of a three-story building, an elegant balcony ran across the front of the building, while red and white umbrellas and tables make it look more like a posh restaurant than a research center.

I waited about a month and called again and repeated my offer of a hyperbaric chamber, operated by trained technicians for a year, or whatever time it took to make a decision on efficacy, at zero cost to the ACE. About an hour after making the second call, my secretary told me that a Dr. Rachel Courchesne was on the line and needed to speak to me on an urgent matter. I picked up the telephone, anticipating a good conversation but this is the way that it went, all one-way at me. Dr. Rachel was in fine fury and, looking back, the only way I could have put a word into that conversation would have been to fold the word flat and slip it into the tirade, sideways. She told me that it was amateurs such as myself that muddied the scientific waters of autism research and that, essentially, I should crawl back into whatever hole I came out of.

She made it clear to me that the mission of the ACE was to perform research to find a cure with pharmaceuticals. In fact, the ACE receives funding from drug companies, and that I should not bother them again. I was taken aback, to say the least.

A decade later, Dr. Eric is apparently doing fine work. For example, in his recent research, quoted widely in the media, he has found that the frontal lobes of an autistic child's brain is abnormally large. This is an observation, perhaps useful in the future, but only time will tell. But does it matter to you the reader whether this is so? Maybe yes, maybe no, so let's hold off on that sort of thing until we get into the scientific part of the book.

Let's get back to Dr. Rachel's harangue of me. In my mind, there is no doubt about the fact that for her, at least, it was all about the money and the potential for a patented silver bullet for Big PHARMA. If successful, a drug company could charge you a huge amount of money to heal your child with autism. At worst, there would be the ability to raise more money for research despite scant evidence of clinical results on children.

An example of this ability to raise money on a potential medical marketplace is the fact that in the year 2012, the National Institute of Health (NIH) allocated $100 million over a five-year period to a bunch of ACE research across the United States. So much for Lillian's glowing endorsement of Dr. Rachel.

However, the Michelson's other introduction proved to be a link in the chain.

Autism research with integrity—Dr. Bernie Rimland.

"A dear friend of ours, Dr. Bernard Rimland, also has a research center in San Diego which he calls *Defeat Autism Now!*, or DAN! for short. Bernie's son, Marc, is a true autistic savant. The movie with Dustin Hoffman and Tom Cruise, Rain Man, was modeled around Bernie and Gloria's son. Bernie was the technical advisor for the film. You must meet Bernie."

Lillian was especially grateful and fond of Bernie and his wife, Gloria. Back in the early 1950s, autistic children were attributed to cold, hard, heartless mothers. Many psychologists and psychiatrists agreed with the "Emotional frigidity" hypothesis, which suggested that the mothers were the cause of what we now know as autism. A lot of children in the United States were taken by force from their parents, and placed in isolated institutions to "protect" them from their parents. Bernie Rimland alone stood against this hypothesis; after all Bernie did have degrees in psychology, including a PhD. In fact, it was Bernie that took away the shame, the blame, and the guilt of many loving parents who were bullied into believing labels put upon them by the so-called "learned."

Driving forces behind autism research centers.

Unlike the La Jolla location for the ACE, Bernie's Autism Research Institute (ARI) was about five minutes drive from my San Diego hyperbaric oxygen dispensary; two minutes of freeway, a quick turn into Adams Avenue, and I had arrived. Again with an erroneous assumption, I had thought that Bernie's ARI would be similar to the ACE La Jolla location. In fact, I drove past Bernie's place three times before I found it. An unassuming, clay-colored, converted storefront, with a little sign on the front door that said, "Autism Research Institute." I knocked on the door and walked in. Instantly, the smell of antiquity and old books overwhelmed me and beguiled me, like something out of a Harry Potter movie. Stacks of papers were high on the floor almost in a willy-nilly fashion. A gentleman dressed in a checkered shirt came towards me. "Bob Sands? The Michelson's said to expect you." He peered over his glasses and gray whiskers at me. In that one glance, we created a bond of friendship that lasted for more than a decade. At least once or twice a month we would lunch together along with his wife, Gloria; always two blocks from his Center, at the *Bleu Boheme*. Bernie and Gloria would carefully examine the menu and then they would both look out and recommend, "Let's all order the fried green tomatoes."

Bernie was interested in clinical results. He would turn over any stone that would perhaps unlock the mystery of autism. When I first met Bernie, 1 in 850 children born in the United States was afflicted with this mysterious and terrifying disorder. As I write, autism spectrum disorders have spiked to 1 child in every 88. So when the phone rang at the ARI, Bernie would often answer it and ask many questions, if the person believed that they had found a cure or a clue. He never discarded anything, in stark contrast to my encounter with Dr. Rachel at the ACE in La Jolla. I once listened in when he spoke with a lady who believed that you could cure autism by filling crystal glasses to different levels and rubbing your fingers around the edges to create pure tones.

Bernie put the telephone down and said to me, "It could be worth looking into. Music does create new neuronal wiring in small children, for example, or a mother singing to a child." I had already mentioned to him that when a child with autism was in the chamber, if you played unfamiliar sounds such as whale songs, or bagpipes, the unresponsive child would often brighten up and look around.

"Maybe new neuronal connections are being established," Bernie pondered as he rubbed his whiskers.

In those days, I worked on weekends. If I wanted to chat with Bernie, Sunday was the best day. No staff to answer the telephone, so he always did. I once asked him whether he took a day off and his response was succinct:

"When autism takes a day off so will I."

Bernie, a $2.6 million donation and his ARI.

Now, here is the most interesting of Bernie Rimland's driving forces—to me at least. It was not about dollars or ego, not in the slightest.

At that time, I was also friends with another Bernie, the renowned anthropologist, Dr. Bernard Aginsky. To avoid confusion between the two Bernie's I will refer to him as Dr. Aginsky. He was in his late 90s at the time, as rich as Croesus. For many years, he had leased the penthouse of the La Jolla Shores Hotel, right on the beach of one of the most desired parts of Southern California. Again, like most of my friends, Dr. Aginsky was a humble man and thought nothing of his wealth other than as a means of accomplishment. Out of the blue, one day in my office, Dr. Aginsky looked at me thoughtfully and then asked, "I have $2.6 million and I want to give it to somebody that would make a difference. Can I give it to you?"

I immediately thought of Bernie and the children with autism and the nonprofit status of the Autistic Research Institute. "Hold that thought Dr. Aginsky. I will give you an answer in a couple of days."

A couple of days later, Bernie and I were traveling to take part in a medical internal review board. I was wearing a shirt with French cuffs and a silk tie. Bernie had on his usual checkered shirt. He confided in me once that he had to make a choice between wearing a tie or trimming his Santa Claus-gray beard. The beard always won. Wondering about his shirts, I respectfully asked him whether he ever received a salary from his ARI.

"Good gracious. I could never pay myself a salary from ARI funds and donations. You must remember that I am on a full Navy pension so it would be wrong to double dip."

After the meeting, as we traveled back along the I-405 to San Diego, I broached the subject of Dr. Aginski's offer of a $2.6 million donation for the Autism Research Institute. Bernie looked out the window for a few moments and fiddled with his beard. I could see he was calculating in his head. He turned to me and said, "We have so much money in the ARI at this time, donations and such, that I think that amount of money could be used elsewhere. Find another recipient."

Dr. Bernard Aginsky, a widower, passed away in January 2000. I have no idea where his donation ended up. I do know whether Bernie Rimland ever had a second thought about not taking that money. He never mentioned our conversation again.

DAN! changing to parallel other autism research.

Research costs money. Bernie used his resources wisely and was not interested in empire-building. He just wanted to find the cause and the answers. Bernie cured kids.

On the anniversary of his death, I went to his Kensington ARI center. I had two bottles of good wine, a bunch of wine glasses, and a whole lot of chocolates. I walked in and was astonished. The whole tenor and tone of the place had changed. The clutter was gone. Bookshelves were neat and organized. The smell of scholarly interest in documents and old books had evaporated. It was as if the ghost of Dr. Bernard Rimland seemed to have moved to another place completely. The staff were delighted to see me, and happily munched on the yummy chocolates that I had brought for them. I commented on the neatness and the vast changes that I observed. One of the staff shrugged her shoulders, grimaced, and said,

"You have no idea just what changes have taken place, and what is about to happen."

An example of the transformation in attitude is exemplified by an offer for inflatable, aka soft-sided, chambers for DAN! while Bernie was still alive. One day, he called me up and explained to me that a group was coming to do a presentation on these soft chambers, and suggested that I plant one of my staff at the ARI to listen to the blurb from the manufacturer, Oxyheal. My secretary, Crystal was waiting amongst Bernie's staff to greet the technicians, doctors, and such that accompanied the owner of the company. They arrived at the ARI, set up the inflatable chamber, and gave their sales pitch, extolling the virtues of hyperbarics at low pressure and at a low cost.

Crystal reported back to me that Bernie looked dour during the presentation and asked, "What would happen to this with a razor blade?" Bernie and I discussed the concept of inflatable chambers the next time he was at my center, after he had finished his treatment in a chamber. He was not convinced about the soft-sided chambers.

"I worried that the scientific data that they quoted was all from hospital quality chambers. I think you are correct Bob, the physics do not hold up. I could not recommend this device to parents for their children."

The "discovery" (observation) of God's gunpowder + Bernie Rimland and Kerri Rivera.

Throughout history, air was taken for granted. You just breathed in and out and there it was; bugs and birds flew in it, mostly useful to keep your floors and ceilings apart. Even in this day and age, many folks think of it in the same way.

Let´s revisit what is now called the Age of Enlightenment, the 17th and 18th Centuries. About 1774, the kindly and soft spoken clergyman/scientist Joseph Priestly and his close friend Benjamin Franklin. They spent time drinking coffee together once a week in a London coffee house along with other scientists. Together they realized that air was actually a mixture of gases, but mostly nitrogen and oxygen. Priestly and Franklin could kill a mouse in a glass jar in less than five minutes by depriving it of one of those gasses. Put the poor little mouse in a glass jar with a lit candle (which consumed the oxygen) and the flame of life in the mouse departed by about the time the candle went out.

Using precise scales, Joseph Priestly and the French chemist Antoine Lavoisier, gave oxygen its name. It could be weighed on a scale just like you would weigh out a cup of rice. Each unseen gas has a weight. Baros is the Greek word for weight. Words that we are familiar with come from this: barometer, barometric pressure, and hyperbaric (as in hyperbaric oxygen therapy, hyper meaning more than and baric meaning weight of the gas).

Weigh oxygen?

The weight of air (mixed oxygen and nitrogen) is 14.7 pounds per square inch at sea level. Most folks like me would prefer to use the word pressure, however, the math is easier if we use the doctors measure (the same as that used in measuring blood pressure – millimeters of mercury or mmHg). An average blood pressure measure is typically 120/80 mmHg. Let's look at it this way; if air at sea level is 760 mmHg, then 20% of that air (oxygen) is 20% of that amount. So multiply 760 by 0.20 and you will realize that vital, life-giving oxygen going into your lungs is about 150 mmHg.

Oxygen - You cannot see it, or smell it but you have to "eat" a kilogram of it each day to stay alive.

Make no mistake. If you take a substance into your body, ALL of your 60 to 90 trillion cells will be affected. Snort it, chew it and swallow it, breathe it, or inject that substance and it will have a positive or negative effect on your whole body. Cells "consume" the substance. Another word for consume is "eat." We will discuss this process, known as oxidative metabolism shortly and, especially how to produce stem-progenitor cells.

Eat oxygen?

Oxygen is a cellular food. The adult human consumes approximately 0.6 kg of oxygen each day (about 1.2 pounds). We can live about three weeks without the food that we chew and swallow, survive three days without water, but we can only live approximately three minutes without oxygen. Body cells start to die—the brain is the most vulnerable. Even a minute or so without oxygen can cause permanent damage.

Don't worry about the earth running out of oxygen. Back in the 1700s Ben Franklin and Joe Priestly also discovered the fact that sunlight on green vegetation produces oxygen—a process called photosynthesis. They showed this with the mouse and candle in the glass jar. As the candle flame seemed to flicker, about to go out and the mouse lay, apparently dead, if there was a sprig of a green plant, both the candle and mouse came back to life. All of the oxygen that you are going to eat in the next 24-hours is replaced by just one square meter of grass producing oxygen through photosynthesis during that same length of time. In populated areas, where there is not much greenery, life-sustaining winds carry oxygen over the oceans and hills to all humans and animals.

Captain Cook, Joseph Priestly, Bernie Rimland, and Kerri Rivera—Explorers

Before we stop talking about Joseph Priestly, he made another observation about gas and this applies to our subject of autism, Bernie, and Kerri.

Living next door to a brewery, Priestly was intrigued about the gas given off by the fermentation of beer in the barrels. He put a dish of water under the vent-pipe coming out of the barrel. When he tasted the water it was full of refreshing bubbles we now call fizz. In 1767, Priestly discovered how to carbonate a liquid. He taught the Navigator, Captain Cook, how to carbonate water and, thinking that drinking that bubbly water would prevent scurvy, Cook set sail with it in 1772. It did not prevent scurvy, but, just like Bernie and Kerri, everything is worth investigating.

We now know that any gas becomes soluble if it is concentrated or made thick. In this instance, the gas was an extremely light gas known as carbon dioxide (yes, the stuff we breath out and, unlike the 21% oxygen, there is only about .033% of carbon dioxide in the air we breathe). However, if you concentrate it to 100% some of it melts into the water and becomes carbonic acid, as is the case with soft drinks. Pop the lid and the acid turns back into carbon dioxides gas, forming little bubbles in the liquid. Joseph Priestly should have patented his observation. Ten years or so later, Joseph Schweppes patented the process in Germany and most folks think of him, not Priestly, when they have a gin and tonic.

Thereby hangs another Newtonian rule about gas that cannot be changed. "Gas under pressure becomes soluble in a liquid according to its density (or thickness)." – Henry's Law.

And this is how oxygen works in a hyperbaric chamber – Newtonian Gas Laws

Stay with the soda can thought and the thicker gas melting into the sugary drink. When a human climbs into a hyperbaric oxygen chamber, we substitute oxygen for carbon dioxide. Believe it or not, you do not feel squashed by the pressure other than you have to "pop" or equalize your ears. Humans are mostly water and other fluids that cannot be compressed like a gas. What is happening in the chamber is that the oxygen (it should be 100% oxygen) is being made thicker or concentrated. And, just like the soda, Henry's Law kicks in. The oxygen melts and the body supersaturates with oxygen.

Another curious thing about a well-oxygenated body is that it becomes slightly alkaline. Your powers of hydrogen (pH) read about the same as seawater, which is alkaline because of the dissolved oxygen in it. Yes, even fish need their daily meal of oxygen!

The word "saturate" comes from the Latin word meaning "to completely fill." Since only our red blood cells (RBC) carry oxygen, they are almost always full of oxygen in a healthy human. The little pulse-ox device that clips on your finger when you are in hospital almost always reads 99% unless you are anemic.

It takes between one and three minutes for each RBC to leave your heart, make its complete circuit, and be back in your lungs for its next load of oxygen. Remember that Henry's Law also applies here, therefore, oxygen is a gas in your lungs but turns into molecular oxygen when it enters your body's fluids. This is so complex that there is no chemical symbol for it, but rest assured, your body's cells feed on it with a great appetite. (More on oxidative cellular metabolism shortly.)

Supersaturate?

When you fill a glass of water to the brim, it is full or saturated. Right? Yes. Holding it under the faucet and running more water into the glass will only mean that the water slops over the edges. The glass is still full. However, this is not the case with any gas, particularly oxygen breathed in by a human. Sure, just like a glass of water, oxygen can be measured by volume, but there is plenty of room to over-fill our bodies with oxygen—supersaturated.

To illustrate a complex process, let's paint a simple picture—An underground subway system.

Think of your lungs as a subway station. Passengers get off the carriages and others get on. In this instance, remember that air is made up of (for easy math calculations) 80% nitrogen and 20% oxygen. Visualize each gas molecule of either type as a gas-person. The body is already full of nitrogen at sea level and since it does nothing at all (it is inert), the nitrogen-people just come in and out your lungs. However the oxygen-people are dynamic and bustle to climb on board the little red-blood-cell carriage (which has a special oxygen-people magnet—an iron molecule), and away the red blood cell rushes with the oxygen-people ready to jump off to the hungriest of the body's cells.

Since Henry's Law allows a gas to melt into clear fluids (and our body has lots of clear fluids), there is a whole lot of room to add more oxygen-people. For example, if we barred entry to our lungs of the nitrogen-people (simple to do—just put on a mask and breath oxygen out of a 100% oxygen tank), then our lung-subway station would have five times more oxygen-people and that would mean that we increased the oxygen-people with each breath from 160 mmHg to 760 mmHg. Remember that they are dynamic, and that each red-blood-cell carriage has an oxygen magnet in it. The thick crowd oxygen-people all bustle to get onto that carriage, through the wet soapy surfactant in the wall of your lungs, even though it was already full.

Essentially, the oxygen-people fill the red-blood-cell carriage, and into the plasma where they are quickly swept along in that clear plasma fluid. Then, when even more oxygen-people continue to crowd on back at the subway station (i.e. the lungs), the plasma, which normally does not carry oxygen-people (molecular oxygen), then needs to put the overload of oxygen somewhere. From here the oxygen-people go out through the rail tunnels (i.e. the circulatory system) and into the clear fluid (i.e. interstitial fluid) that bathes our body cells. As this fluid fills with oxygen, it dumps it off into the body's largest water compartment—the body's cells (the intra-cellular compartment). These cells now have a veritable feast of oxygen available.

About 90 minutes after you start breathing 100% oxygen, you are supersaturated with five times the normal amount of molecular oxygen. No, you do not swell up since we are talking about molecules in existing body fluid, not added fluid. Curiously, if you put the little pulse-ox back on your finger, it would still only read 99% since it is just measuring the oxygen in your red blood cells. However, using high-tech oxygen measuring equipment (transcutaneous oxiometry), you can even measure molecular oxygen in your urine, saliva, and tears—something once thought impossible.

Now comes the hyperbaric chamber part. You climb in, the door is shut, and 100% oxygen floods in. The technician can look at the physician's prescription and increase the thickness of the gas in that chamber. Double the thickness (i.e. weight or pressure) to 1520 mmHg; ninety minutes later you will climb out of that chamber with ten times the number of oxygen-people (molecular oxygen) in your body. Or, put another way, since your body's cells are going to consume (eat) that additional oxygen, this is the only way you can increase the weight of oxygen-food available in a 24-hour period. One hundred percent pure oxygen at sea level and you will "eat" between just under half a kilogram to just over one kilogram of oxygen.

The deeper/thicker the oxygen is in the chamber the more you get. In fact, following Dalton's Universal Gas Laws, by doubling the pressure, you can actually get your daily oxygen meal at almost two kilograms.

Now the nifty thing about oxygen supersaturation.

Unlike many pharmaceuticals that are weight dependent, (too much of a certain drug will sicken or kill a smaller patient, not enough of the same drug and there will be no therapeutic benefit to the patient) with hyperbaric oxygen no individual can suffer from an "overdose" of oxygen, regardless of whether the patient is a mouse, rabbit, horse, or human. All climb out of the chamber, (depending on the pressure/thickness of oxygen they were treated at) with the same amount of molecular oxygen in their body. Not one person has suffered from an "overdose" of oxygen.

Our oxygen dispensary centers have treated, according to their physician's prescription, tiny babies, the elderly, and everyone in between; we have never seen a negative side effect. Instead, we have observed clinical positive effects for all.

What happens when we actually give people an extra "oxygen feast" by putting them in the chamber? Again, an extremely complicated process has to be simplified for understanding. However, before making things simpler we must first look at the science. Chatelier's principle predicts the effect of a change in conditions on a chemical equilibrium. The principle is named after Henry Louis Le Chatelier, who observed that if a chemical system at equilibrium has a change in concentration, temperature, volume, or partial pressure then the equilibrium shifts to counteract the imposed change and a new equilibrium is established. Any change in the status quo prompts an opposing reaction in the responding system.

Time to change just a couple of words for better understanding. Let us start with the word equilibrium and now call it "the way it was." We can also use that term for status quo and counteract can be changed to "pushing back." So what is really being said here is, if we change something in any chemical environment it will push back hard to be the way it was. Now it is time to make things even easier.

We are all chemical systems.

We have already mentioned cola in a can and advised against giving it to anybody, your child or yourself, because of the acidity of the liquid in the soda. But there is an even more compelling reason—the sugar. There are 39 grams of sugar, about ten teaspoons of sugar, in one can of soda. Would you feed your child that much dry sugar?

"Come on," you might think to yourself, "No way. It does not taste that sweet." It is true. The manufacturers of sodas use a lot of phosphoric acid to disguise the sickly sweetness of the drink. If you still doubt this, take an ordinary penny, or an egg and put it in a glass full of cola overnight. Within a few days, you will find all tarnish removed and instead a bright, shiny penny; the eggshell will be soft and pliable, all its protein dissolved.

Visualize what happens when you allow your small child a glass of cola (or any soda drink). Your little kid becomes supercharged, bouncing off the wall, and for some time you cannot get the child to sit still. In other words, your child has become hyperglycemic. Too much sugar, of any sort, whether it be corn fructose, beet or cane sugar, the results are the same according to Chatelier's principle. Remember, we are all chemical systems and by adding all of the sugar to a small child's body, all of the cells in that body will push back to return to the way they were before the additional sugar. The most efficient way to return to this state is to increase physical activity, causing cells to consume all of the additional added sugar.

Run, baby, run!

Another way, of course, is to rely on the body to use its own chemicals to burn off the sugar. This is predominantly the job of the hormone insulin. If an adult drinks a lot of soda each day (or even has one gin and tonic, which has about half as much sugar as a soda) and does not increase physical activity, diabetes eventually sets in as the body become insulin exhausted, unable to produce more. Then the risk of blindness in old age, amputations, and other terrible consequences may become a reality.

The process of cells using the nutrients we consume is called cellular metabolism. Overloading cells with nutrients that we do not need such as excessive raw sugars (or even food that is easily converted by the body into glucose) and fats, results in obesity; your brain will become sluggish. Now, add extra oxygen and your body will certainly consume it. This is called oxidative metabolism. The added oxygen will push all cells into overdrive; repair cells use the added oxygen to work harder in an effort to restore their state prior to entering the hyperbaric chamber. Repair cells help get rid of the toxins, acids and, excess/unneeded sugars.

After people have been in a hyperbaric chamber, healthy adults comment on the fact that their brain is working more efficiently. Memory becomes vivid and problems are easier to solve. This state is commonly known as the “cappuccino” effect or “brain brightening.”

As for small children with autism, almost all of them have a neuronal oxygen transport problem. So the additional feast of oxygen actually makes the good cells, the repair cells, work harder for just a little while after the treatment. This increase in oxidative metabolism dissipates after a couple of hours, but that little bit of improvement lasts: brain cells wire up into complex circuits, blood vessels within the brain become robust, and stem cells are produced. (More on that shortly.)

This is why all oxygen treatments need to take place over consecutive days. There is no point in attending one hyperbaric treatment and then not following it up for another week. Sort of like a seed sprouting from the earth. With each treatment, the positive changes begin to accumulate, and positive results are visible with the passing days.

Can oxygen therapy alone cure autism?

The short answer is no; it cannot. I make this statement based on the treatment of scores of challenged children that have come to my centers across America, and in other parts of the world. Something more is needed—an interdisciplinary approach.

How many treatments in the chamber does a challenged child need to be part of that recovery program?

As noted earlier, we treated our first challenged child, Emily, in one of my centers in the United States. Just one treatment per day at 2 ATA, the pediatrician's decision, made a vast difference to little Emily. Then along came a physician, Dr. Paul Harch, a very bright health scientist. He came to the conclusion 15 years ago that two 60-minute treatments each day at 1.75 ATA for 20 days could cure children with autism. He was so sure of himself that he even took out a patent on that protocol.

I was nonplussed and puzzled. I did not think that you could patent a natural phenomenon such as the effects of hyperbaric oxygen therapy or gravity. Of interest is the fact that the Supreme Court came to the same conclusion in 2012 with what is now known as the "Prometheus" ruling. In essence, the courts had taken all of the enforceable teeth out of Dr. Harch's protocol. This is not to say that there isn't some merit in the treatment pressure, nor is it to imply that Dr. Harch was ignorant of the science of hyperbaric rights in the therapy. So my centers took notice of the fact that he is a physician, and for a little while at least, we did use the Harch protocol on children.

While our little patients saw useful and positive results, there was never a normal recovery by any of the children in that 20-day double ride. This left a lot of the parents disappointed, some angry at what they felt were false claims about hyperbaric oxygen therapy, and rightfully so.

It is significant to note that Dr. Harch, and other clinicians, no longer use the two treatments a day protocol for challenged children, and he no longer claims that recovery is possible in just 20 days.

Now, onto our wonderful Jennifer; the ballet dancer and champion ice-skater. Her mom, Vickie, a bright spirit and highly intelligent woman (actually that describes the whole family), independently arrived at the conclusion that diet was a most important part of the normalization program. All together, over a 12-year period, Jennifer had 135 rides in the hyperbaric chamber. She ice skated at 2am (when the rink was clear), and focused on diet. In a nutshell, while the extra oxygen was feeding one side of cellular metabolism in spurts of 20 treatments at a time, Jennifer's cells were getting the correct diet of nutrients on the other. All of Jen's cells were giving off the right amount of heat and energy, particularly her repair cells.

Bernie and Kerri proved scientifically that the addition of proper nutrients, combined with the removal of deleterious/toxic foods is probably more important than hyperbaric oxygen therapy for children and young adults on the spectrum. Yes, the oxygen boost helps propel the patient to "normal," but the other nutritional changes are really the rocket fuel. It is a long journey, not all that expensive, and certainly worthwhile.

When a hyperbaric chamber is not a hyperbaric chamber.

One final comment about treating neuronal disorders at 1.75 ATA pressure is of utmost importance; this is that 100% oxygen is essential to the protocol. Since Dr. Harch arrived at this treatment depth, there have been thousands of inflatable chambers, so-called mild hyperbaric chambers, sold in the United States and around the world. These are not the same as hospital grade, hard-sided chambers. These bag-chambers do not deliver the needed 100% oxygen. In my opinion, at the very best, if you modify them (against FDA rules) you will only get 34% oxygen.

What have we learned thus far? Time for a review and then THE BIG SECRET REVEALED!

- Oxygen cannot be seen, but we can weigh it.
- Our bodies "eat" oxygen every day because it is a food for our cells.
- The only way we can get additional oxygen is in a hyperbaric chamber.
- This is because any gas becomes soluble and melts into a fluid (i.e. Henry's law); the same reason we can carbonate soda drinks.
- In a single treatment in a hyperbaric-oxygen-therapy chamber, we cannot overdose (and I have never witnessed any negative side effects).
- With the extra oxygen feeding the cells, our bodies get a boost of energy, particularly the repair cells.
- For oxygen in the chamber to be effective, it must be 100% oxygen. Inflatable chambers do not deliver 100% oxygen.
- Hyperbaric oxygen therapy alone will not cure autism, but it does make a big difference.
- There is no such thing as a "one-size fits all" treatment protocol in a hyperbaric chamber that works for all patients. For example, twice a day in the chamber for 20 days will make some difference but not necessarily "normalize" a challenged child.

Did you miss the big secret?

It was already mentioned about one page back—the discoveries of Bernie and Kerri. Let us repeat the secret for you and then push the edges of their really stunning discovery that is backed up by the latest science.

Bernie and Kerri proved scientifically that the addition of the proper nutrients, in the absence of offending foods, in addition to eliminating pathogens and parasites, are most likely more important than hyperbaric oxygen therapy for children and young adults. Yes, the oxygen boost helps propel the patient to "normal" but the other dietary interventions are really the rocket fuel. Not to mention freeing the body of offending pathogens/parasites which wreak havoc on the body. It is a long journey, not all that expensive and, certainly worthwhile.

Nutritional "padding." Hyperbaric oxygen therapy and the production of progenitor and stem cells.

When a baby is growing inside its mother, it has progenitor cells and stem cells. The function of these cells is considered to be, as yet, undecided. Although studies continue to elucidate exactly which mechanisms are involved in determining the fate of these cells, it is generally accepted that their surrounding cells will dictate their development. For example, the big arteries that crisscross the chest area make a decision sooner or later to turn into a heart with all of its compartments. Then, as development moves on different parts of the baby appear. These are called progenitor or stem cells.

There are various estimates on how many different types of cells the developed human has in total; however, it is predicted there are somewhere between 210 and 300. Mind you, when the baby takes its first breath of air, all but about four of these different cell groups stop reproducing. For example, the brain, with all of its neuronal pruning and circuitry needs about 20 year's worth of stimulation to properly wire all of the connections. Surprisingly, a lot of the neurons will be shed or "pruned" since they will not be put to use in the adult. However, neuroplasticity allows new skills to be added or repairs to be made to and by the brain in a lifetime.

However, medical scientists have now realized the great need for progenitor or stem cells (cells that can make decisions to replace worn or damaged cells within the body), and the search is on for these decision-making cells. Many scientists are looking for something they can patent and as a result are missing the point that it is already here. Or they are ignoring it, because but they cannot patent it.

It is called hyperbaric oxygen therapy.

> Penn Study Finds Hyperbaric Oxygen Treatments Mobilize Stem Cells
>
> **Science Daily** — According to a study to be published in the American Journal of Physiology-Heart and Circulation Physiology, a typical course of hyperbaric oxygen treatments increases by eight-fold the number of stem cells circulating in a patient's body. Stem cells, also called progenitor cells are crucial to injury repair. The study currently appears online and is scheduled for publication in the April 2006, edition of the American Journal.
>
> Stem cells exist in the bone marrow of human beings and animals and are capable of changing their nature to become part of many different organs and tissues. In response to injury, these cells move from the bone marrow to the injured sites, where they differentiate into cells that assist in the healing process. The movement, or mobilization, of stem cells can be triggered by a variety of stimuli including pharmaceutical agents and hyperbaric oxygen treatments. Where as drugs are associated with a host of side effects, hyperbaric oxygen treatments carry a significantly lower risk of such effects.
>
> "This is the safest way clinically to increase stem cell circulation, far safer than any of the pharmaceutical options," said Stephen Thom, MD, PhD, professor of emergency medicine at the University of Pennsylvania School of Medicine and lead author of the study." This study provides information on the fundamental mechanisms for hyperbaric oxygen and offers a new theoretical therapeutic option for mobilizing stem cells."
>
> "We reproduced the observations from humans in animals in order to identify the mechanism for the hyperbaric oxygen effect," added Thom. "We found that hyperbaric oxygen mobilizes stem/progenitor cells because it increases synthesis of a molecule called nitric oxide in the bone marrow. This synthesis is thought to trigger enzymes that mediate stem/progenitor cell release."[1]

Nutritional padding: It has long been known that when hyperbaric oxygen therapy is applied, existing cells, the good ones, the repair cells, turn the body into its own pharmaceutical company, providing the right sort of chemicals at the right time. Above all, there is a dramatic increase in the production of the molecule called nitric oxide from the bone marrow. Probably the best way of describing the addition of hyperbaric oxygen therapy, as an adjunct to the correct nutritionals, is simply by putting fuel onto an already existing smoldering fire. Out of the glow of embers and smoke you will see a vigorous combustion as your fireplace heats up and the flames appear.

Now imagine other ingredients that have yet to be discovered as beneficial. For example, stem cells that repair the damaged brain of a person with autism. It could be that adding external stem cells to the recipe of recovery and normalization, with simple nutritional low-cost padding, and putting the little patient in the chamber would be a faster way of achieving what all parents want for their children—to eventually become fully functional and independent adults.

All of this by just a little bit of nutritional padding and helping the body to rid itself of pathogens/parasites.

Now, it is time to turn back to Kerri so she can tell you what she has seen when hyperbaric oxygen therapy is added to the correct nutritional padding, along with chlorine dioxide to keep pathogens under control. That is the big secret to helping your child through the challenges that living in this century have visited upon you.

* * *

As Bob mentioned, I have seen miracles happen when hyperbaric oxygen therapy was applied to a child who is ready. Two girls, in particular, lost their diagnosis completely within weeks of completing treatment in the chamber. As I mentioned before, hyperbaric oxygen therapy gave my own son his speech back. Before his first 40 dives he had several single-syllable sounds, but after, he began stringing words together to request his favorite foods.

Today we are seeing that one dive a day at 1.75 ATA (atmospheres absolute) for 90 minutes at depth is better than two 60-minute sessions were first prescribed for Patrick seven years ago. Every three months after the first treatment of 20 sessions, apply ten sessions of 90 minutes, each at 1.75 ATA, until your child is recovered.

A lot of parents ask when is a good time to do hyperbarics. To this I reply that we must first have diet and chlorine dioxide totally under control, and be at least three months into the Parasite Protocol. At that point it's time to start looking for a chamber.

Make sure to do an ATEC before your child's hyperbaric oxygen therapy sessions, and repeat two months after the sessions are over; the results will be obvious.

Healing
the symptoms known as
Autism

This is my Asperger son Tobias, 10 years old, and me... when we received the book this summer.

Chapter 12

Step 7 - GcMAF & Autism: State of the Art & Future Perspectives for a Natural Cure

by Marco Ruggiero, MD, PhD.
Specialist in Diagnostic Radiology
Full professor of Molecular Biology
Dept. of Experimental & Clinical Biomedical Sciences
University of Firenze, Italy

"The voyage of discovery is not in seeking
new landscapes but in having new eyes."
~ Marcel Proust

My family and I have been in biomed for so long, that there aren't too many interventions we haven't already tried. A few years back, Dr. Usman mentioned to me that GcMAF might be something good for Patrick. She explained that GcMAF is a human protein that occurs naturally in a healthy body. It is a special supplement that replaces the missing part of the immune system, and also acts as the body's own internal medicine.

IMMUNE GCMA
/ML 2.2ML
ORE AT 4°C
CH M1210-25
ETAINED VIAL

Based on her recommendation, we ordered a vial, and applied it for about two and a half months. I saw less obsessive compulsive eating, but when the vial ended and I hadn't seen amazing results, I wasn't extremely motivated to continue. Not only that, funds were scarce, so continuing really wasn't an option.

Two years went by and I spoke to Dr. Usman again. She said we really should give it four to five months before making a final decision. So, we scraped together the money for a second order. We started again in November of 2012, and reached full dose by December, 26th 2012. Seemingly, as time went on, Patrick started to do new things.

By springtime of 2013, Patrick's BFF's mom was saying how well he was doing, and reminded me that during the prior month Patrick had not been doing the things that he was now doing. As I write this in October of 2013, all of my money now goes to GcMAF. ☺

One of the things that I feel makes GcMAF so exciting is the fact that it helps regrow dendrites in a matter of hours after it is injected. The dendrites are the very things that got fried when mercury entered our children's bodies. With the help of GcMAF, we are now able to restore what was lost. Not to mention, GcMAF boosts the immune system, which we have come to find is often sorely lacking in people with ASD's.

Many people have asked me if GcMAF is now an official part of the protocol. When working towards recovery, I feel that we need to try everything at our disposal that has been proven to heal autism—in the correct order, or course. GcMAF fits that description. According to www.gcmaf.eu, "As of February 2013, on the American National Library of Medicine alone, 142 eminent scientists from 8 nations have published 59 major GcMAF research papers, which can be viewed on the US Government's Pubmed system." There is now enough research to prove that GcMAF has secured its place as a powerful biomedical intervention for healing autism. Therefore, after three months on the Kalcker Parasite Protocol, it is time to start thinking about adding in supplements, HBOT and GcMAF.

We are blessed to have the one and only Dr. Marco Ruggiero, the lead GcMAF researcher, write the remainder of this chapter about this substance, it's benefits and applications in autism and beyond.

The information he presents is extremely important because I feel that GcMAF is the future of medicine—today. What other intervention boasts a cancer recovery rate of 80%? None! Dr. Bradstreet is reporting 85% responders and 15% recovery rate from autism with GcMAF.[1]

Imagine adding GcMAF to the protocol already outlined in this book at the right moment (after working down some of the pathogen load), we should start seeing even higher rates of recovery than 15%. I look forward to seeing what transpires during the upcoming year as more families add GcMAF to their children's protocols. I am excited for more and more people to lose their autism diagnoses.

For the past 20 years, Dr. Marco Ruggiero has worked on the vitamin D axis, a metabolic pathway that includes the vitamin D binding protein-derived macrophage activating factor or GcMAF. In the past 3 years he has published studies on the immunotherapeutic effects of GcMAF in cancer, HIV/AIDS, chronic fatigue syndrome and neurological conditions.

Marco Ruggiero holds a PhD in Molecular Biology, is a certified Medical Doctor specializing in Clinical Radiology and is a professor of Molecular Biology at the Department of Experimental and Clinical Biomedical Sciences of the University of Firenze, Italy.

He served in the Army as Medical Lieutenant with specific training in chemical, biological and nuclear warfare. He worked at Burroughs Welcome Co. North Carolina, USA publishing a seminal paper with Nobel Laureate Sir John Vane and, subsequently, at the National Cancer Institute of the NIH in Bethesda, working with Dr. Stuart A. Aaronson and Dr. Peter Duesberg.

Since 1992, he holds the chair of Molecular Biology at the University of Firenze where he leads a research group of about 10 researchers. He has published more than 150 peer-reviewed scientific papers on signal transduction in a variety of experimental and spontaneous pathologic system related with cancer, chronic kidney disease, chronic fatigue syndrome and neurological conditions.

The role of nutrition in the treatment and eradication of the symptoms of autism is well established and other chapters of this book treat this topic in detail. What is far less known is the fact that certain natural components of the diet, certain nutrients, have profound effects on all the systems that are affected in autism and in particular on the brain and the immune system. Their effects are so dramatic that for about 20 years, these natural components have been labelled as "drugs" even though they are essential parts of our bodies, just like hormones or neurotransmitters. I am referring here to the family of molecules that is known as GcMAF, an acronym that stands for Gc-protein-derived Macrophage Activating Factor.

GcMAF has been the object of intense research in the past 20 years for its therapeutic role in cancer, autoimmune diseases, and HIV infection (for review, see: Anticancer Res. 2012 Jan;32(1):45-52). But, only one year ago, in 2012, its dramatic effects on the eradication of the symptoms of autism were published in a prestigious peer-reviewed scientific journal (Autism Insights 2012:4 31–38).

In this chapter I shall provide an overview on GcMAF with particular reference to its therapeutic role in autism and to the future perspectives deriving from the basic and applied research that is revolving around this fascinating molecule.

A Brief History of GcMAF

The first publication on GcMAF in a peer-reviewed journal indexed in the US National Library of Medicine is dated 1994 (J Immunol. 1994 May 15;152(10):5100-7). In this article, a group of researchers from the Department of Biochemistry at the Temple University School of Medicine in Philadelphia, USA, described the effects of a protein, defined as GcMAF, on rat macrophages. Macrophages are key elements of the immune response and the research group led by Dr. Nobuto Yamamoto, the first author of this paper, showed that GcMAF activated these cells. Given the central role of macrophages in the control of all immune responses, GcMAF was considered a powerful regulator of the immune system and the research on GcMAF initially was focused on its immune-stimulant properties. In fact, one year after the initial observation on rats, Dr. Yamamoto, who in the meantime had moved to the Laboratory of Cancer Immunology and Molecular Biology of the Albert Einstein Cancer Center in Philadelphia, authored a paper suggesting that a defect in endogenous GcMAF production contributed to immunodeficiency in AIDS patients (AIDS Res Hum Retroviruses. 1995 Nov;11(11):1373-8.). Consequently, he provided GcMAF to Medical Doctors treating HIV patients and, 14 years later, after 7 years of follow-up, he published a seminal paper describing eradication of HIV infection through GcMAF administration (J Med Virol. 2009 Jan;81(1):16-26).

Immunodeficiency, however, is not unique to HIV/AIDS and, back in 1996, Yamamoto and colleagues, hypothesized that deficient endogenous GcMAF production was responsible for the relative immunodeficiency that is typically observed in cancer patients (Cancer Res. 1996 Jun 15;56(12):2827-31). Quite logically, this observation of 18 years ago, led to the proposal of administering GcMAF to cancer patients with the goal of empowering the immune system that in turn would fight the cancer growth. This approach is known as "immunotherapy" and it is much older than GcMAF since, in modern times, it stems from the observation of Dr. William Coley in 1891. However, it could be argued that the idea of stimulating the immune system to fight diseases is even older since, as it is stated in this recent article, "…cancer regresses when associated with acute infections such as bacterial, viral, fungal, protozoal, etc. Acute infections are known to cure chronic diseases since the time of Hippocrates…" (Indian J Cancer. 2011 Apr-Jun;48(2):246-51).

Whatever the case, administration of GcMAF to patients with metastatic advanced cancer proved efficacious and there are now scores of papers published in the peer-reviewed literature describing the dramatic therapeutic effects of GcMAF in all types of experimental or spontaneous tumors.

Nowadays, the interest for GcMAF in the immunotherapy of cancer is so high that the prestigious scientific journal "OncoImmunology" dedicated the cover of its August 2013 issue to GcMAF. The molecular rendering of GcMAF is on the left of the cover with a caption linking it to an article by Thyer and colleagues published in the same issue and describing successful treatment of advanced cancer patients (Figure 1).

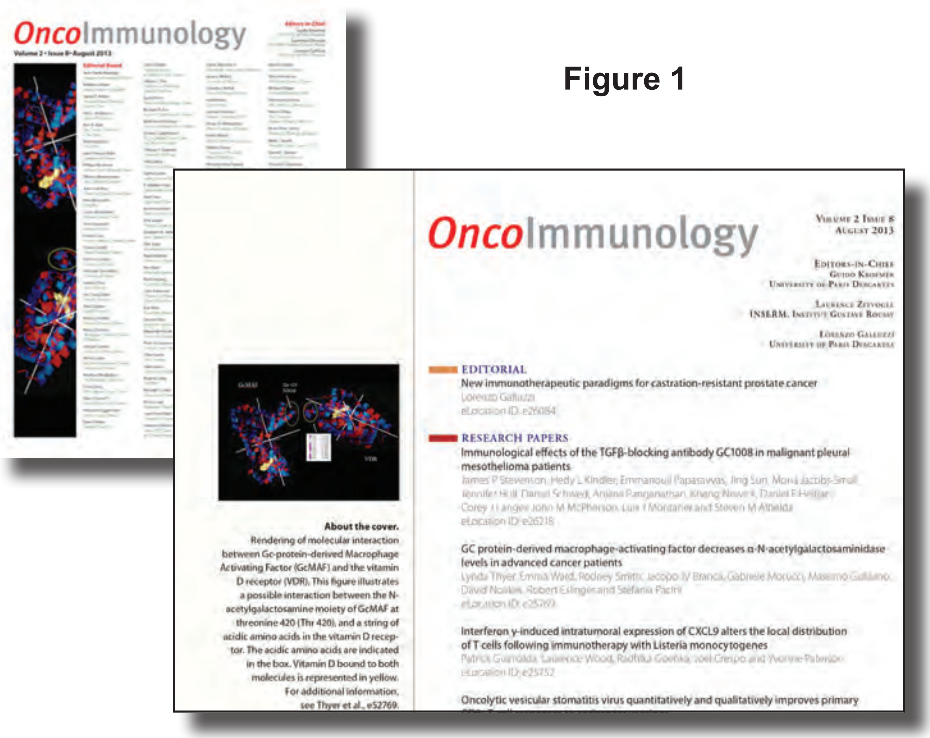

Figure 1

OncoImmunology
Volume 2 • Issue 8 • August 2013
Editorial Board

OncoImmunology
Volume 2 Issue 8
August 2013

Editors-in-Chief
Guido Kroemer
University of Paris Descartes

Laurence Zitvogel
INSERM, Institut Gustave Roussy

Lorenzo Galluzzi
University of Paris Descartes

EDITORIAL

New immunotherapeutic paradigms for castration-resistant prostate cancer
Lorenzo Galluzzi
eLocation ID: e26084

RESEARCH PAPERS

Immunological effects of the TGFβ-blocking antibody GC1008 in malignant pleural mesothelioma patients
James P Stevenson, Hedy L Kindler, Emmanouil Papasavvas, Jing Sun, Mona Jacobs-Small, Jennifer Hull, Daniel Schwed, Anjana Ranganathan, Khang Newick, Daniel F Heitjan, Corey J Langer, John M McPherson, Luis J Montaner and Steven M Albelda
eLocation ID: e26218

GC protein-derived macrophage-activating factor decreases α-N-acetylgalactosaminidase levels in advanced cancer patients
Lynda Thyer, Emma Ward, Rodney Smith, Jacopo JV Branca, Gabriele Morucci, Massimo Gulisano, David Noakes, Robert Eslinger and Stefania Pacini
eLocation ID: e25769

Interferon γ-induced intratumoral expression of CXCL9 alters the local distribution of T cells following immunotherapy with Listeria monocytogenes
Patrick Guirnalda, Laurence Wood, Radhika Goenka, Joel Crespo and Yvonne Paterson
eLocation ID: e25752

Oncolytic vesicular stomatitis virus quantitatively and qualitatively improves primary

About the cover.
Rendering of molecular interaction between Gc-protein-derived Macrophage Activating Factor (GcMAF) and the vitamin D receptor (VDR). This figure illustrates a possible interaction between the N-acetylgalactosamine moiety of GcMAF at threonine 420 (Thr 420), and a string of acidic amino acids in the vitamin D receptor. The acidic amino acids are indicated in the box. Vitamin D bound to both molecules is represented in yellow. For additional information, see Thyer et al., e52769.

Source: https://www.landesbioscience.com/journals/oncoimmunology/oncoimmunology_2-8.pdf

Consecration of GcMAF in the Olympus of science, however, occurred a few days later, on September 14, 2013, when another article by Thyer and colleagues was published in "Nutrients", an international, peer-reviewed open access advanced forum for studies related to human nutrition (Nutrients. 2013 Jul 8;5(7):2577-89). In fact, this article describing the effects of GcMAF on human breast cancer cells, was ranked in the top 5% of all articles ever tracked by Altmetric. The Altmetric score is a general measure of the attention that an article, book or dataset has received online and reflects the quantity of attention received as well as the quality of that attention. By the time that the GcMAF was ranked in the top 5% of all scientific articles ever tracked, Altmetric had analyzed 1,510,524 articles across all scientific journals. Therefore, we may safely state that in this chapter we are describing one of the hottest topics in today's world science (Figure 2).

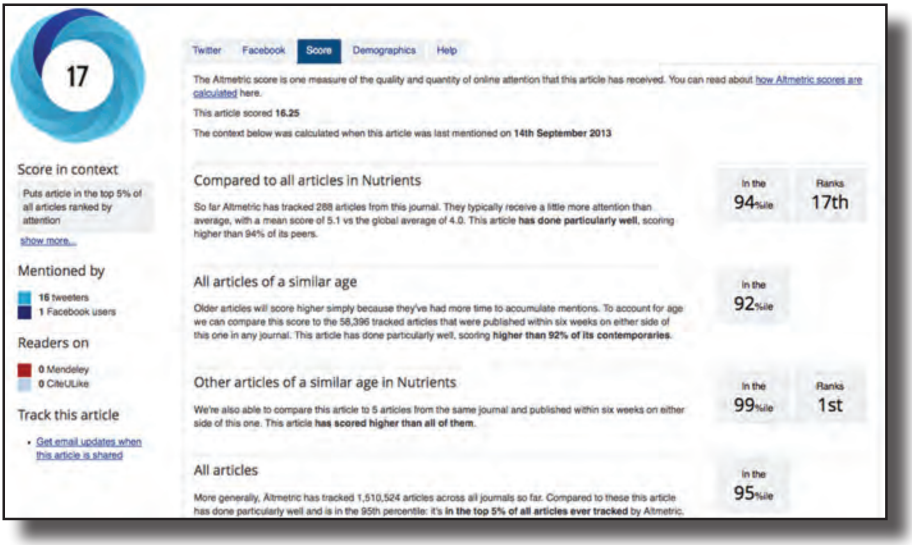

The Altmetric score is one measure of the quality and quantity of online attention that this article has received. You can read about how Altmetric scores are calculated here.

This article scored **16.25**

The context below was calculated when this article was last mentioned on **14th September 2013**

Score in context

Puts article in the top 5% of all articles ranked by attention

show more...

Mentioned by

16 tweeters
1 Facebook users

Readers on

0 Mendeley
0 CiteULike

Track this article

- Get email updates when this article is shared

Compared to all articles in Nutrients

So far Altmetric has tracked 288 articles from this journal. They typically receive a little more attention than average, with a mean score of 5.1 vs the global average of 4.0. This article **has done particularly well**, scoring higher than 94% of its peers.

In the 94%ile
Ranks 17th

All articles of a similar age

Older articles will score higher simply because they've had more time to accumulate mentions. To account for age we can compare this score to the 58,396 tracked articles that were published within six weeks on either side of this one in any journal. This article has done particularly well, scoring **higher than 92% of its contemporaries**.

In the 92%ile

Other articles of a similar age in Nutrients

We're also able to compare this article to 5 articles from the same journal and published within six weeks on either side of this one. This article **has scored higher than all of them**.

In the 99%ile
Ranks 1st

All articles

More generally, Altmetric has tracked 1,510,524 articles across all journals so far. Compared to these this article has done particularly well and is in the 95th percentile: it's **in the top 5% of all articles ever tracked** by Altmetric.

In the 95%ile

Figure 2

The interest for the therapeutic effects of GcMAF in a variety of conditions derives from the early observation by Yamamoto and colleagues in 1995, who showed elevated level of serum alpha-N-acetylgalactosaminidase (Nagalase) in the serum of AIDS patients. Nagalase is an enzyme that deglycosylates (degrades) the precursor of GcMAF that is the circulating Gc-protein, also known as vitamin D-binding protein (AIDS Res Hum Retroviruses. 1995 Nov;11(11):1373-8). This results in the loss of GcMAF precursor activity and consequent dysfunction of the immune system. Increased Nagalase activity, however, is not observed uniquely in AIDS patients. For example it was demonstrated that the increase of Nagalase activity in the serum of cancer patients is due to the fact that cancer cells release Nagalase and, therefore, Nagalase activity reflects tumor burden, aggressiveness and progression of the disease. Consequently, determination of Nagalase activity is currently proposed as a reliable way of evaluation of cancer severity (Cancer Lett. 2009 Oct 8;283(2):222-9).

In short, in cancer, HIV/AIDS, and in many other chronic conditions, elevated serum Nagalase, degrades Gc-protein that is the precursor of GcMAF. Therefore, GcMAF is not produced inside the body to a sufficient extent and this leads to immunodeficiency. Serum Nagalase, however, does not degrade GcMAF, and this justifies the administration of exogenous GcMAF in all those conditions where its endogenous production is insufficient due to elevated Nagalase activity.

Until 2002, it was thought that GcMAF exerted its effects only on macrophages and that its therapeutic efficacy was due to stimulation of macrophages that in turn would attack and destroy cancer cells as well as virus-infected cells. In 2002, however, a novel effect of GcMAF was described. Kanda and colleagues, working at the Department of Molecular Microbiology and Immunology, Division of Endothelial Cell Biology, Nagasaki University Graduate School of Medicine, Nagasaki, Japan, published a paper, co-authored by Yamamoto, describing the inhibitory effects of GcMAF on angiogenesis (J Natl Cancer Inst. 2002 Sep 4;94(17):1311-9.). Angiogenesis is the process of building new blood vessels that provide the tumor mass with the required supply of blood and nutrients that it needs to grow and metastatize. Therefore, since 1971 it was postulated that inhibition of angiogenesis could be a successful strategy to deprive the growing tumor mass of its supply of blood and nutrients (N Engl J Med. 1971 Nov 18;285(21):1182-6). While there are scores of inhibitors of angiogenesis, some of which in use in the therapy of cancer, GcMAF offers the advantage of performing more than one anticancer effects; in fact, not only it inhibits tumor-induced angiogenesis (Cancer Immunol Immunother. 2011 Apr;60(4):479-85), but it also stimulates macrophages that attack and destroy the cancer cells (Nutrients. 2013 Jul 8;5(7):2577-89).

If these combined anticancer effects were not enough, in 2010 Gregory and colleagues, working at the Department of Ophthalmology and Visual Sciences of University of Kentucky in Lexington, Kentucky, USA, demonstrated that GcMAF directly inhibited proliferation and metastatic potential of human prostate cancer cells (PLoS One. 2010 Oct 18;5(10):e13428). Slightly more than one year later, we were able to demonstrate that GcMAF not only inhibited proliferation and metastatic potential of human breast cancer cells, but it also reverted their neoplastic phenotype; in other words, cancer cells treated with GcMAF became normal (Anticancer Res. 2012 Jan;32(1):45-52).

On the basis of the peer-reviewed scientific literature quoted above, we can now state that GcMAF exerts multiple effects that are responsible for its efficacy in anticancer therapy:

1. It stimulates macrophages that attack and destroy cancer cells, as originally postulated by Yamamoto and colleagues.

2. It inhibits tumor-induced angiogenesis thus depriving the growing tumor mass of blood and nutrients.

3. It directly inhibits cancer cell proliferation and metastatic potential and it retro-transforms cancer cells into normal healthy cells.

Given this plethora of effects on cancer, it is no wonder that the majority of GcMAF research has been focused on its anticancer properties. There are, however, a few research articles suggesting that GcMAF might exert also other biological actions with therapeutic implications. In addition to the studies on HIV and AIDS quoted above, GcMAF has received attention for its potential role in the therapy of bone disorders (Blood. 1996 Oct 15;88(8):2898-905), and autoimmune diseases such as Lupus Erythematosus (Clin Immunol Immunopathol. 1997 Mar;82(3):290-8).

Despite these scanty evidences on the possible role of GcMAF in conditions other than cancer and AIDS, it was not until 2012, however, that GcMAF proved dramatically effective in the therapy of autism.

GcMAF & Autism

GcMAF, being derived from Gc-protein, also known as vitamin D binding protein, belongs to the so called "vitamin D axis", a nutritional metabolic pathway described in detail in a recent review (European Nephrology, 2011;5(1):15–9). My research group had been working on the vitamin D axis since 1996 (Epidemiol Prev. 1996 Apr-Sep;20(2-3):140-1), that is since the beginning of GcMAF research. Just like our colleagues in the USA and Japan, we were mainly interested in the role of the vitamin D axis in bone metabolism and cancer (Radiol Med. 1996 Nov;92(5):520-4. Oncol Res. 1998;10(1):43-6). Therefore, it came as a surprise when Dr. Nobuto Yamamoto first mentioned to me the effects of GcMAF on neurodegenerative diseases such as Parkinson's and Alzheimer's diseases.

In the hot summer of 2010 we visited Dr. Yamamoto in his hometown of Philadelphia, and we had the honor and the pleasure to spend two very intense days with this old-style exquisite gentleman. Dr. Yamamoto gave us a detailed history not only of his decade-long research activity, but also of the entire field of immunotherapy, starting with the pioneering work of Dr. William Coley at the end of the nineteen century. During these intense and fruitful talks, Dr. Yamamoto mentioned the dramatic effects that he had personally observed by administering GcMAF to a patient suffering from Parkinson's disease. According to Dr. Yamamoto, the effects had been almost immediate and dramatic up to the point that he himself was amazed. Regrettably, he had been unable to clinically document the case and, therefore, this remained an anecdote. From this anecdote, however, an entire new field of GcMAF research was about to rise; that is the study on the role and implications of GcMAF in treating neurologic and neurodegenerative disorders.

The first solid evidences for a role of GcMAF in treating a neurodegenerative disorder were simultaneously presented by Dr. Paul Cheney and Prof. Kenny de Meirleir at the meeting of the International Association for Chronic Fatigue Syndrome and Myalgic Encephalomyelitis (CFS/ME) held in Ottawa in September 2011. A few days later, in collaboration with Dr. Paul Cheney, we presented in Padova, Italy, a study demonstrating that food-based GcMAF dramatically improved the symptoms of CFS/ME, and this work was published in the official journal of the Italian Society of Anatomy and Histology, one of the oldest and most respected European scientific societies, founded at the beginning of the twentieth century (It. J. Anat. Embryol. Vol. 116, No 1, 2011).

CFS/ME is a complex disorder that shares many features of autism. In fact, it is characterized by immune system dysfunction, widespread inflammation, and multi-systemic neuropathology (J Intern Med. 2011 Oct;270(4):327-38. In Vivo. 2013 Mar-Apr;27(2):177-87). Dysfunction of the immune system involves abnormal functions and distributions of T lymphocytes, B lymphocytes, natural killer cells, and monocyte/macrophages (Nihon Rinsho. 1992 Nov;50(11):2625-9. Brain Behav Immun. 2012 Jan;26(1):24-31). The etiology of CFS/ME, just like the etiology of autism, is still unknown and multiple factors are thought to be responsible for its onset and progression, thus lending credit to the hypothesis that both etiology and pathogenesis are multifactorial (J Intern Med. 2011 Oct;270(4):327-38). Heavy metal exposure and viral infections are among the factors that contribute to CFS/ME etiology and pathogenesis, and a role for human endogenous retroviruses has been recently hypothesized (In Vivo. 2013 Mar-Apr;27(2):177-87). Both chronic heavy metal exposure and viral infections are considered responsible for the immune system dysfunction and neuropathology that are typical of CFS/ME (Med Hypotheses. 2012 Sep;79(3):403-7. In Vivo. 2013 Mar-Apr;27(2):177-87).

In addition to these areas of overlapping, it had been noticed by Doctors treating autism and/or CFS/ME, that in several families the two conditions are often associated with parents of autistic children suffering from CFS/ME. Interestingly, this happens also in families with adopted children with autism, thus lending credit to the hypothesis that environmental or infectious factors may contribute to both conditions.

Whatever the case, it had been observed that CFS/ME patients had elevated levels of serum Nagalase together with a number of immunological alterations, and it was hypothesized that they could benefit from GcMAF treatment. The results presented in Ottawa and in Padova confirmed this assumption and ever since immunotherapy with GcMAF is a stronghold of CFS/ME treatment.

At about the same time, Dr. Bradstreet, a leading authority in the field of autism, began to study the levels of Nagalase in autistic children and their responses to GcMAF treatment. The results obtained in a first cohort of 40 subjects were reported in a seminal paper published in 2012 (Autism Insights 2012:4 31–38). This paper represents a turning point in the understanding of autism pathogenesis and treatment.

The first significant finding is the observation that the average level of Nagalase in autistic subjects was more than two folds higher than the normal level, thus lending credit to the hypothesis that immunodeficiency or immune system dysfunction is a hallmark of autism. The second finding, probably the most important in recent years, is the observation that GcMAF treatment significantly improved the symptoms of autism and this improvement was associated with a decrease of Nagalase levels.

The therapeutic response to GcMAF was statistically significant and showed the typical bell-shaped curve that is expected when a biologically effective treatment is at work. About 15% of subject did not respond; this means that their symptoms were not ameliorated by GcMAF treatment. However, more than 15% of children responded so dramatically that all the symptoms of autism were eradicated in a matter of weeks up to the point that the children could no longer be defined as "autistic." According to the very word of the authors, in the children in this group, "This response was demonstrated at school, during therapies, home and outside the home as substantial improvement to the point that many or most of the criteria of autism were no longer present."

The remaining 70% of children were distributed according to the so-called Gaussian distribution in the groups showing "slight," "moderate," or "considerable" improvement.

This paper demonstrated for the first time that GcMAF treatment was able to eradicate the symptoms of autism in a matter of weeks, thus reinforcing the idea brought forward in many chapters of this book that autism is a curable condition indeed.

Quite obviously, the dramatic results obtained by Dr. Bradstreet did not go unnoticed and, in April 2013, he presented further results at the First GcMAF Immunology Conference held in Frankfurt, Germany. According to Dr. Bradstreet, as the time of follow-up and the number of subjects increased, a shift toward the right (that is the "best") part of the curve could be observed. This means that the percentage of those showing "considerable" to "very considerable" improvement increased.

From the point of view of the general practitioner, the issue could be considered closed: GcMAF cures a significant percentage of autistic children now and that figure is rising.

However, since I am a researcher in addition of being a Medical Doctor, the successes of Dr. Bradstreet in treating autism and the research papers published by Immuno Biotech laboratories are in the middle of an exciting series of observations that are leading to a completely new understanding of how GcMAF works and what autism truly is.

In fact, the undisputable efficacy of GcMAF in treating autism raises a number of questions on the mechanisms involved at the cellular and molecular levels. The most important questions are:

1. The efficacy of GcMAF in curing autism is due only to its immune-stimulating effect?
2. Or, is GcMAF acting directly also on neurons as it does on cancer cells?
3. Does GcMAF revert the neuro-anatomical alterations that are typical of the autistic brain?
4. Why a certain number of autistic subjects do not respond to GcMAF?
5. Can we devise strategies to have these non-responders respond?
6. Can we devise strategies to further improve this already dramatic therapeutic effect?

The Effects of GcMAF on Human Neurons

GcMAF is a protein and, as such, it was thought that it would stimulate cells from the outside. In other words, until 2013, it was conjectured that, on their external surface, cells had a receptor that bound GcMAF and conveyed the signal to the inside of the cell and, ultimately, to the nucleus and the DNA, thereby modifying the cell behavior. However, despite 20 years of research, such a receptor for GcMAF had not been identified. And this might appear truly odd considering that the entire human genome had been sequenced, that is studied in every single detail, more than 10 years ago.

In our first publication with Dr.Yamamoto, presented at the XVIII International AIDS Conference in Vienna in 2010, we had demonstrated that the response to GcMAF in human mononuclear cells (that is macrophages) was dependent on polymorphisms (that are the individual variations among humans) of the vitamin D receptor (VDR) gene. This observation was not surprising considering that vitamin D, its receptor, and GcMAF all belong to the vitamin D axis.

In order to further investigate this relationship between GcMAF and the VDR we began studying the molecular assembly of GcMAF that is the "shape" that the molecule assumes in its physiological conformation. Having done this, we looked for complementary areas in the GcMAF and VDR molecules and… voilà! The two molecules complemented each other as two elements of a puzzle.

We published this observation in the "Nutrients" paper that is now in the top 5% of all articles ever tracked, and the picture of the two molecules interacting with each other is in the front cover of the OncoImmunology August 2013 issue. The molecular mechanisms through which GcMAF and VDR interact and the consequent complex web of intracellular signaling are topics for specialists and rather difficult to divulge; the basic point, however, is that we had demonstrated how is it possible that GcMAF exerted so many different effects that can be exploited to treat a number of diseases so different from each other.

From this observation we deducted that all cells having the VDR would have responded to GcMAF and, given our interest for the effects of GcMAF in neurological conditions such as autism and CFS/ME, we began our study on human neurons. In fact, human neurons express (i.e. "have") the VDR and therefore, they are candidate to be stimulated by GcMAF.

We cultured human neurons in Petri dishes and we challenged them with ultra-pure GcMAF; in this type of experiments there were no other variables. That is, there were only neurons and GcMAF; the cells of the immune system or other molecules were not involved. By doing so, we would have been certain that whatever effect we observed, this was due uniquely to GcMAF.

As I have shown at the congress "Curando el Autismo" in San Juan de Puerto Rico in August 2013, GcMAF at extremely low concentration significantly increased neuronal cell viability and metabolic activity. These effects on cell viability were associated with dose-dependent intracellular cAMP production (cAMP is an intracellular second messenger). This means that GcMAF increased the production of energy inside human neurons and made them more active, more viable. Increased viability and metabolic activity following GcMAF stimulation as well as cAMP formation were accompanied by morphological changes, meaning that stimulated neurons changed their shape.

In the absence of GcMAF, human neuronal cells under the microscope appeared as small, relatively undifferentiated cells with large nuclei. After 24h stimulation with GcMAF, neurons showed a significant change in morphology that was consistent with the induction of neuronal differentiation and increased connectivity. The cytoplasm was enlarged and cytoplasmic elongations could be observed. After 72h incubation with GcMAF, these morphological changes were more evident and well differentiated cells could be observed.

After incubation with GcMAF, the cells appeared to establish contacts with each other. Taken together these results indicated that GcMAF increased neuronal cell viability, metabolic activity and differentiation, with the first effects being observed at 24h. It is worth noticing that the assay that we used to determine cell viability, measured mitochondrial activity, and mitochondrial dysfunction is known to be hallmark of ME/CFS (Int J Clin Exp Med. 2012;5(3):208-20), and possibly of autism as well. Therefore, the effects of GcMAF that we observed explain the reason why GcMAF is so effective in CFS/ME and autism; it counteracts the molecular alterations that are the pathogenetic basis of both conditions.

I wish to spend a few words here on the effects of GcMAF on neuronal connectivity. It is well assessed that this is one of the basic alterations in the brain of autistic subjects; neurons do not establish contacts with each other and therefore the neurological signal cannot be transmitted from one area of the brain to another thus causing a "disconnection" (Rev Neurol. 2012

Feb 29;54 Suppl 1:S31-9). GcMAF re-establishes neuronal connectivity by stimulating the formation of what are called "cytoplasmic elongations" that are like small "arms" that protrude from the neurons and establish contact with each other as if neurons were tending arms to "touch" each other and exchange information.

In Figure 3A, human un-stimulated human neurons in culture can be observed; they are small and roundish. After stimulation with GcMAF (Figure 3B), they are much larger because increased energy production leads to increased protein synthesis in their cytoplasm. But, perhaps more important, they have a highly irregular, elongated, triangular, shape with spikes that touch each other. The areas of contact between the two neurons can be clearly observed in the central part where they establish connecting bridges. Please notice that both pictures were taken with the same magnification.

Figure 3

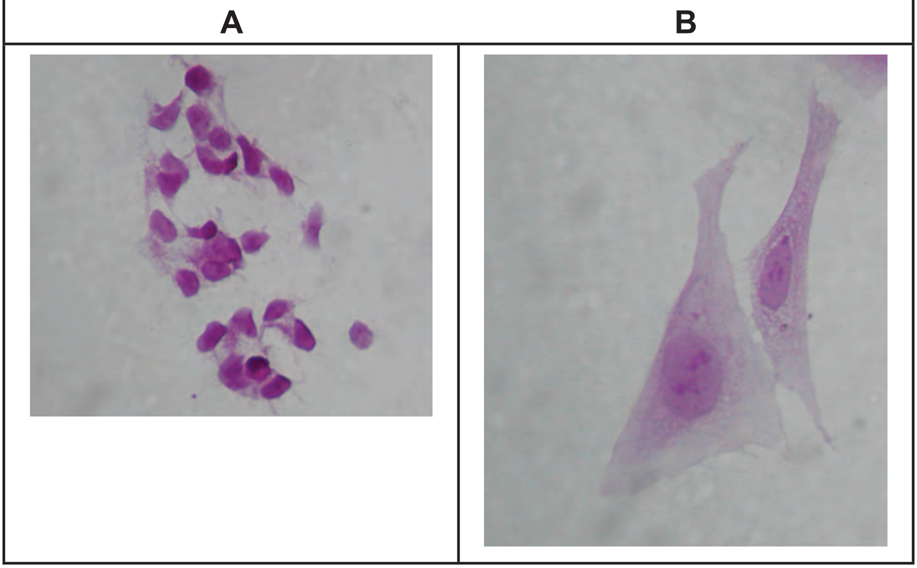

These observations provide an answer to the first two questions raised by the article published by Dr. Bradstreet: the therapeutic effects of GcMAF in autism can be ascribed to at least two different actions of GcMAF. On one side, it rebalances the immune system that is typically dysfunctional in autistic subjects; on the other side, it directly stimulates human neurons increasing their metabolic activity and their connectivity. But, as it always happens in research, these "answers" led to other "questions": does GcMAF cross the blood brain barrier? That is, does it arrive to the neurons of the central nervous system? Before challenging this question further, however, let's proceed with order and let's try to provide an answer to the other questions that were raised by the Bradstreet's paper.

GcMAF and the Neuro-Anatomical Alterations of the Autistic Brain

If the etiology of autism is still undefined, the pathogenesis and the brain alterations associated with autism are better understood. Several studies using sophisticated imaging techniques demonstrated that there is atypical structural brain connectivity within cortical white and grey matter and it is postulated that such an alteration of brain connectivity may represent one of the principal neural substrates underlying autistic symptoms (PLoS One. 2013 Jun 18;8(6):e67329. Proc Natl Acad Sci U S A. 2013 Jul 22.).

The observed atypical brain connectivity is localized to particular areas of the brain and the alterations in the fronto-temporal regions are associated with the severity of social and repetitive symptoms of autism (Proc Natl Acad Sci U S A. 2013 Jul 22.). Atypical brain connectivity in autism is also associated with reduction of grey matter volume in different areas of the brain (Cerebellum. 2013 Apr 10.). In the bilateral superior temporal gyrus significantly decreased functional connectivity was accompanied by the strongest trend of grey matter volume decrease (PLoS One. 2013 Jun 18;8(6):e67329), thus reinforcing the hypothesis that study of grey matter volume may provide indication of functional connectivity.

However, the studies on brain connectivity and grey matter reduction quoted above were conducted with sophisticated Magnetic Resonance Imaging (MRI) equipments using complex softwares. Because of their cost and complexity, these techniques are not suitable for routine examination and this appears to be in contrast with the need to provide the general practitioners with the instruments to study autism. In fact, given the current epidemiological trend, it is highly probable that even primary care providers will see with increased frequency a patient with autism spectrum disorder in their clinic (Curr Probl Pediatr Adolesc Health Care. 2013 Jan;43(1):2-11). This consideration prompted us for the search of accessible methods to study brain alterations in autism and to study the response to different treatment including GcMAF.

To this end, we developed an easily accessible method to study the brain cortex using transcranial ultrasonography. This technique was first applied to the diagnosis and treatment CFS/ME a syndrome that, as we know, share neuropathological similarities with autism (Toxicology. 2008 May 2;247(1):61-72). In particular, we proposed to use ultrasounds to study the grey matter in the temporal lobe, and to evaluate signs of neuronal layer disorganization and brain inflammation (Med Hypotheses. 2012 Sep;79(3):403-7). It is worth noticing that these same neurological alterations, i.e. minicolumn structural changes and inflammation are frequently observed in autism (Front Immunol. 2013 Jun 10;4:140. doi: 10.3389/fimmu.2013.00140. Print 2013). Therefore, we hypothesize that ultrasounds could be used in autism in order to study the brain alterations that have been described using more sophisticated and less accessible techniques.

First of all, we assessed the safety of the technique; to this end we used sub-thermal ultrasounds. These are considered inherently safe and have been used for fetal imaging in utero, and virtually every part of the body, including brains of newborn babies through fontanelles. Their inherent safety and any lack of correlation with the onset of autism have been confirmed by meta-analysis of thousands of cases. A recent randomized controlled trial on over 1.400 cases demonstrated beyond any reasonable doubt that there is no link between the prenatal ultrasound scans and the autism phenotype (J Autism Dev Disord. 2012 Dec;42(12):2693-701).

Then, in a collaborative work with Dr. Bradstreet, we decided to focus our attention on the temporal cortex in order to compare our results with those already published by Hameroff and colleagues and by ourselves using this technique (Med Hypotheses. 2012 Sep;79(3):403-7. Brain Stimul. 2013 May;6(3):409-15). In addition, having to choose a particular area of the brain for this initial study, the temporal lobe was particularly interesting because of its accessibility to ultrasounds and its involvement in the pathogenesis of autism (PLoS One. 2013 Jun 18;8(6):e67329. Print 2013).

In the typical adults subject, meninges appeared as a well-organized array of layers of about 2.6 mm thickness. The meninges were separated from the cortex by a structure that shows echogenic and morphological feature consistent with the sub-arachnoid space. The cortex of the temporal lobe appeared as a well-organized array of alternate, hyper-echogenic/hypo-echogenic, layers. The thickness of the cortex was 5.0 mm. This value of thickness and the anatomical positioning of the probe led us to hypothesize that we were observing the temporal areas designated by von Economo as TG and TE, i.e. those areas involved in the control of eye movements and balance in standing position (area TE), social behavior, mood and decision making (area TG) (von Economo C. Temporal lobe. In: Triarhou LC, editor. Cellular structure of the human cerebral cortex. Basel; 2009. p. 114-132). It is worth noticing that the layers within the column-like structure of the healthy subject appeared well-organized and parallel to each other.

However, the images obtained in a representative autistic subject showed that the thickness of the cortex of the temporal lobe was significantly reduced, as it had been observed with MRI (PLoS One. 2013 Jun 18;8(6):e67329. Print 2013). Other images showed a concomitant increase in the spinal fluid in the sub-arachnoid space that was twice as much as in the healthy subject. These findings are consistent with the very recent observation that infants who developed autism spectrum disorder had significantly greater extra-axial fluid characterized by excessive cerebrospinal fluid in the sub-arachnoid space (Brain. 2013 Jul 9.).

Taken together, the results obtained in autistic subjects demonstrate that our technique was able to visualize the excess extra-axial fluid in a manner comparable to MRI; however, at variance with MRI, the observation with ultrasounds was immediate and intuitive and could be easily performed with no discomfort during a routine examination.

Another finding that we observed was a relative disorganization of the neuronal layer arrangement in the temporal cortex of autistic subjects. It was possible to observe that the neuronal layers had a different orientation and that there were black areas of about 0.4 mm that could represent areas of poor connectivity.

Since this technique has been applied only recently to the study of autism, there is no statistical analysis indicating whether GcMAF or any other effective treatment can reverse the anatomical alterations that we have observed in the brain of autistic subjects. Therefore, there is no sound scientific answer to this question; there are only some anecdotic evidences suggesting that this might indeed be the case. Autistic subjects treated with GcMAF and showing considerable clinical improvement showed a normal brain anatomy. Future observation will tell whether these anecdotic evidences have a statistical meaning.

Ultrasounds, however, in addition to their role in early diagnosis and follow-up, might also have a role in the therapy of autism and in particular in improving the response of those subjects who do not appear to benefit from GcMAF treatment.

GcMAF Non-Responders

As odd as it may appear to non-researchers, the presence of non-responders is often the best assurance that we are dealing with a real biological phenomenon and not with placebo effects or worse. In fact the bell-shaped curve presented in the article by Bradstreet and colleagues it typical of any effective treatment and it confirms the validity of the results. This having said, however, there is the need to provide an answer to those who do not respond. And, in the effort of providing an answer to non-responders, improvements in the overall efficacy of the treatment for everyone will eventually follow.

At the moment, there are two hypotheses that could help explaining why certain subjects do not respond to GcMAF in terms of improvement of symptoms of autism.

The first hypothesis is related with the VDR gene polymorphisms. As we proposed at the AIDS conference in Vienna in 2010, and later published in 2012 (J Nephrol. 2012 Jul-Aug;25(4):577-81), the individual responsiveness to GcMAF depends on the individual variations (polymorphisms) in the VDR gene. In other words, there are subjects harbouring variants that are associated with little or no response. However, this lack of response is relative and not absolute. In fact, it has been observed with vitamin D that these individuals respond to higher doses of the vitamin, thus indicating that the VDR is "less" responsive, but still functioning. This would mean that the so-called non-responders might need a higher dose or a longer time of GcMAF treatment to obtain the desired therapeutic effects. Prolonged observation of a higher number of cases will tell whether this hypothesis is correct.

The second hypothesis is closely associated with the first one as well as with the question briefly outlined in a preceding section: "does GcMAF cross the blood brain barrier?" In a recently submitted paper, we proposed that the interconnection between GcMAF and VDR signaling is responsible for the transport of GcMAF through the blood brain barrier. In fact, transport of macromolecules across the blood brain barrier requires both specific and nonspecific interactions between macromolecules and proteins/receptors expressed on the luminal and/or the abluminal surfaces of the brain capillary endothelial cells (Int J Cell Biol. 2013;2013:703545. doi: 10.1155/2013/703545). Since VDR signaling appears to enhance brain to blood transport at the blood brain barrier through both genomic and non-genomic actions, it can be hypothesized that the interaction between GcMAF and VDR (Nutrients. 2013 Jul 8;5(7):2577-89. doi: 10.3390/nu5072577) may favor GcMAF transport into the brain. This means that subjects with a less responsive VDR because of unfavorable genetic variants, would also have a less efficient transport of GcMAF to the brain.

This latter consideration, however, opens the way to a possible solution that might also help increasing the effectiveness of GcMAF in those who already respond to it.

In fact, now that we know the physiological molecular assembly of GcMAF, we can design specific GcMAF molecules that are more efficient in crossing the blood brain barrier and, therefore, are more effective in treating neurological conditions such as CFS/ME and autism. These molecules exploit the fatty acid binding site in the GcMAF molecule, and use one of the most beneficial fatty acids, oleic acid, the basic component of olive oil, to target brain cells.

Most interestingly, GcMAF/oleic acid molecules do not only stimulate human neurons with a much higher efficiency than GcMAF, but they also regulate glial cell metabolism. Glial cells are the connective tissue of the brain and glial cell dysfunction is involved in the pathogenesis of autism. In particular, glial cell dysfunction seems to be responsible for the neuronal underconnectivity typical of the autistic brain (Neuron Glia Biol. 2011 May;7(2-4):205-13). In short, we can postulate that the development of these novel GcMAF/oleic acid molecules will prove beneficial not only for the GcMAF non-responders, but also for improving the therapeutic efficacy of GcMAF in all neurological conditions where GcMAF treatment is indicated.

In addition to these novel, more physiological, brain-targeted, molecules, there are other strategies that can improve GcMAF transport into the brain with consequent increased therapeutic efficacy. Just a few days ago, in a journal dedicated to advanced radiological research, it was published a report demonstrating that ultrasounds, just like those described in the preceding section, transiently open the blood brain barrier. This phenomenon is exploited to favor transport of chemotherapeutic drugs into the brain of individual with brain tumors and it proved safe and efficacious after extensive testing.

On the basis of this observation, one could devise a simple treatment protocol using the same ultrasound system that is currently used for transcranial ultrasonography. In analogy with what is performed to deliver chemotherapeutic drugs to brain tumors, ultrasounds could be used to facilitate the transport of GcMAF through the blood brain barrier and to focus its delivery into those areas of the brain that are known to be altered in autism.

In this manner, during a regular examination, the health professional could exploit our technique to assess the neuro-anatomical alterations in the autistic brain and to deliver GcMAF to those areas. The use of more physiological, brain-targeted GcMAF molecules that can be administered sublingually, would greatly improve the efficacy of the treatment and it will most likely reduce the percentage of the so called non-responders to GcMAF.

As noted above, however, the use of ultrasounds to treat neurological conditions is not completely new. In fact, we and others have demonstrated that non-thermal ultrasounds have significant effects on mental states (Brain Stimul. 2013 May;6(3):409-15), modulate neurotransmission (Neuropsychobiology. 2012;65(3):153-60), cardiac frequency and blood pressure and improve skeletal muscle strength (The Journal of IiME, vol. 6 (1), p 23-28, 2012). These effects can be observed with ultrasounds alone, that is in the absence of other treatments.

In a protocol for the treatment of autism and CFS/ME, we propose to use our ultrasonographic technique together with timely administration of sublingual GcMAF/oleic acid molecules. In this manner, we are able to exploit the direct effects of ultrasounds on the neuronal microtubules (Brain Stimul. 2013 May;6(3):409-15), together with the ability of ultrasounds to favor the transport of GcMAF/oleic acid into the brain and, more precisely into those areas where the neuroanatomical alterations are observed.

What is GcMAF?
A drug, a hormone, a natural component, a nutrient?

How to define a molecule that is formed every time a wound is licked? This is the most evolutionary conserved protective behavior in animals, humans included. The enzymes in the saliva convert the Gc-protein in the blood that is on the wound into active GcMAF that prevents infection of the wound and promotes repair (PLoS One. 2013 Jul 18;8(7):e69059).

For many years researchers considered GcMAF only a key regulator of the immune system because of its activity on macrophages and its denomination reflects this early conjecture. However, as it often happens in science, the original denomination of a molecule proves incorrect with the passing of time and the accumulation of scientific evidences. Vitamin D is a notable example; it is not an vitamin and it does not contain any amine group (even though the suffix "vita" applies, in that it is indispensable for life). Now it is rightly considered a hormone produced by the body and, to be more precise, a secosteroid hormone that has a mechanism of action similar to that of classical steroidal hormones. GcMAF, being a part of the vitamin D axis might soon become another example of scientific mislabeling. In fact, we now know that it exerts its actions in many other cell types in addition to macrophages and that its role go well beyond activation of macrophages.

And even though there are no doubts that it is a powerful stimulant of the immune system, it would not be surprising to realize that this might not be its most relevant physiological role. In this respect, GcMAF shares the destiny of its precursor, the Gc-protein. Gc-protein is a serum alpha 2 glycoprotein composed of a single polypeptide chain, structurally related to serum albumin, that performs a variety of actions in several cell types. It is synthesized in the liver and has been detected on the surface of several cell types, yolk sac endodermal cells, and some T lymphocytes. In B cells, it participates in the linkage of surface immunoglobulins. In addition to vitamin D, it binds actin and acts as an actin scavenger. These characteristics are of paramount importance in human pathology since actin is the most abundant protein in eukaryotic cells and it is a major cellular protein released during cell necrosis that may cause fatal formation of actin-containing thrombi in the circulation if the actin scavenging capacity of Gc-protein is exceeded. Not surprisingly, therefore, several studies suggest an association between Gc-protein and resistance or susceptibility to chronic obstructive pulmonary disease, thyroid diseases, diabetes, multiple sclerosis, and sarcoidosis (for a review on Gc-protein and the vitamin D axis, see: European Nephrology, 2011;5(1):15–9).

From these considerations it derives that the potential indications for GcMAF treatment might be far too numerous than those currently envisaged. And also this parallels the fate of vitamin D, once thought only to prevent rickets and now involved in almost all aspects of human physiology and pathology up to the point that a (serious) scientific article on this subject was entitled: "Does vitamin D make the world go 'round'?" (Breastfeed Med 2008; 3: 2392–50.).

Apart from issues of biomedical terminology, it is now emerging the concept that GcMAF is an essential component of that nutritional pathway that is the vitamin D axis. It is not a case or a coincidence that the article on GcMAF that has received the greatest attention was published in a scientific journal named "Nutrients". With the proceeding of research it is becoming apparent that GcMAF participates in such a variety of physiological events that each proposed definition is limited.

But, if we accept the concept that GcMAF is a natural nutrient, indispensable for brain and immune system development and function (and possibly for many more essential physiological events), two questions spontaneously arise: "Where do I find it in nature?" and "How can I consume it?"

The second question already has an answer: now that the physiological molecular assembly of GcMAF is known, there are on the market several GcMAF products that can be used sublingually or simply eaten. Modern biotechnology has reproduced the physiological conformation of GcMAF so that it can be absorbed in a natural way without recurring to intravenous or intramuscular injections. GcMAF injections are reserved for those cases where a strong and rapid stimulation is required.

The other question is more intriguing and leads to the final section of this chapter.

My Own Personal Experience with GcMAF

As it happens in many people of my age exposed to the radiations from the Chernobyl incident, I have several thyroid nodules. Being a radiologist, it was easy for me to perform ultrasonography on my thyroid and what I saw on the screen was not nice. In addition to small lesions in my right lobe, in my left lobe I observed a massive mixed (solid and cystic) protruding nodule substituting for most of the lobe. The anterior-posterior diameter was about 3 cm, whereas the longitudinal diameter was greater than the length of the probe (4 cm). The structure was grossly inhomogeneous with cystic areas and hypo-reflecting areas alternated with iso- or hyper-reflecting areas. The margins bordering the residual normal thyroid tissue appeared well defined with an area of edema (inflammation) surrounding the nodule. The blood vessels were located around and inside the nodule and few or no blood vessels could be observed in the large hypo-reflecting areas suggesting the presence of metabolically inactive tumor tissue. Consistent with these findings, laboratory tests showed signs of immunodeficiency and anemia. And, it has been reported that anemia is among the predictive elements of malignancy in thyroid nodules (Tunis Med. 2002 Sep;80(9):536-41).

It was immediately clear to me that I needed a powerful stimulant of the immune system as wells as an anticancer agent, and GcMAF was my first thought. However, at that time, GcMAF was synthesized only in academic research laboratories and it was not available on the market as the pure, highly active, quality controlled product that we know today. Therefore, we decided to look for natural sources of GcMAF.

We already knew that GcMAF is naturally formed when a wound is licked because saliva contains the enzymes that convert blood Gc-protein into active GcMAF. Licking my thyroid, however, quite obviously, was not an option.

Because of these hindrances, we hypothesized that GcMAF could be obtained through a natural food fermentation process that exploits the characteristic of certain microbial strains to produce the enzymes that convert food Gc-protein into GcMAF (Hollis and Draper, 1979). According to our hypothesis, fermented food-derived GcMAF would activate the macrophages widely diffused in the walls of the entire gastrointestinal tract and, in particular, in the Waldeyer's-Pirogov tonsillar ring (or pharyngeal lymphoid ring). Our first goal was to assess the feasibility of obtaining natural GcMAF through home-made food processing procedures. To this end, we devised a strategy to prepare a fermented food product using those microbes, known to mankind for their healthy properties for millennia, that express the enzymes involved in GcMAF synthesis (Hollis and Draper, 1979).

After having prepared this fermented food at home, I began consuming it on a daily basis.

Since at that time we too considered GcMAF mainly for its immune-stimulating properties, the first objective was to assess the efficacy of this procedure on certain markers of immunodeficiency, the so called CD4 lymphocytes. The general public became familiar with CD4 lymphocytes because of their involvement in HIV infection and AIDS, and CD4 cell number is considered a reliable marker of immune system function. The role of CD4 cells in the immune system has been widely popularized up to the point that the webpage of the AIDS Authority of the Government of the United States of America states; "CD4 cells or T-cells are the "generals" of the human immune system. These are the cells that send signals to activate your body's immune response when they detect "intruders," like viruses or bacteria. Because of the important role these cells play in how your body fights off infections, it's important to keep their numbers up in the normal ranges..."

In our observation, the reference laboratory considered normal value for CD4 cell count 500 - 1500 cells/μL. Before consumption, my CD4 count was low, close to the threshold of AIDS definition (372 cells/μL). CD4 count rose from the abnormal value of 372 to the normal value of 609 cells/μL after three weeks of consumption of the fermented food that I had prepared. Such an increase continued up to eight weeks with CD4 cell count reaching 853 cells/μL. After many years, they are still in the normal range.

The trend toward normalization of blood parameters was observed also in other abnormal haematological values. My values (low or abnormal) changed as follows:

	Before Consumption	After 3 Weeks	Normal Values
Red Blood Cells			
(x106/uL)	4.55	4.85	4.6 - 6.2
Haematocrit, %	42.9	45.9	40 - 52
Haemoglobin, g/dL	14.8	15.6	13 - 18
Platelets/uL	165,000	206,000	140,000 - 440,000

The neutrophil cell count decreased from 77% to 55% in about 12 weeks, indicating a normalization of parameters suggestive of a pro-inflammatory state. Overall, the observed changes in haematological parameters were consistent with those positive modification of immune responses in humans that might represent the potential mechanism by which food fermented by probiotic bacteria confer health benefits.

Being a researcher, I am very well aware that this report is a mere anecdote that, in the era of evidence-based medicine, has low value or is not even considered to be medical evidence. However, a very recent study on the evaluation of clinical practice challenges the deprecation of stories like that reported above. Thus, it is well known that some studies present large and impressive statistics obtained from many observations while others report a small number of noteworthy events, as we did in this study. However, according to this authoritative epistemological approach, "all of these stories become evidence of what works in medicine" (J Eval Clin Pract. 2011 Oct;17(5):920-6).

Conclusions

In August 2013, a paper published in the very prestigious American Journal of Immunology reported a series of clinical cases successfully treated with GcMAF (Am. J. Immunol., 9: 78-84. 2013). Among these, there were pathologies already known to be responsive to GcMAF treatment, such as cancer, autism, and CFS/ME. Other neurodegenerative diseases such as multiple sclerosis and amyotrophic lateral sclerosis, however, had not been treated with GcMAF before, and this is the first report on such a successful treatment. This paper, together with all the other recognitions reported above, places GcMAF as one of the most effective tool to fight diseases that were considered incurable until very recently.

In other chapters of this book it is clearly stated that autism IS curable, and the successes of GcMAF in curing autism further strengthen this concept. What is emerging and what I hope to have been able to convey, is that GcMAF is not an exogenous drug that "fixes" some problem, but rather a very natural component of our body that is essential for, at least, brain and immune system development and function. In many conditions, including autism, GcMAF is not produced to a sufficient extent and therefore it has to be administered from the outside. With the advent of a courageous little biotech company that has perfected the synthesis of natural, fully physiological, GcMAF molecules that can be eaten, GcMAF treatment will become all the more and flexible to fit the individual needs. Eventually, GcMAF will become part of an integrated nutritional plan as it was intended by nature during evolution.

Consistent with this concept, I wish to conclude with an example from the realm of nature.

According to the Genetic Science Learning Center of the University of Utah, "Some mother rats spend a lot of time licking, grooming and nursing their pups. Others seem to ignore their pups. Highly nurtured rat pups tend to grow up to be calm adults, while rat pups who receive little nurturing tend to grow up to be anxious. It turns out that the difference between a calm and an anxious rat is not genetic—it's epigenetic. The nurturing behavior of a mother rat during the first week of life shapes her pups' epigenomes. And, the epigenetic pattern that mom establishes tends to stay put, even after the pups become adults."

It seems that licking their pups while breastfeeding in the first week of life is the decisive factor; in the saliva there are the enzymes that convert milk Gc-protein into GcMAF. What is more natural than this?

* * *

Subcutaneous GcMAF has been available through FIRSTiMMUNE for several years. Their website is:

www.gcmaf.eu

They now also produce a sublingual version of GcMAF called GOleic. They also have Bravo probiotics, and more. Please visit their website for more information.

"If you're not stubborn, you'll give up on experiments too soon. And if you're not flexible, you'll pound your head against the wall and you won't see a different solution to a problem you're trying to solve."

~ Jeff Bezos
Founder of Amazon

Chapter 13

Beyond Recovery: The Maintenance Plan

"Success is not final, failure is not fatal:
it is the courage to continue that counts."
~ Winston Churchill

We see these questions on our boards from time to time:

Will my son have to take CD for the rest of his life?

or

Will we be sorting through poop forever?

The answer is No!

Once your child has recovered, you can move to the maintenance phase. Since we live in a toxic world, it is important to continually detox the system, and keep the body free of parasites. While this can look different for every child and every family, the following are some general guidelines designed to keep our kids healthy once they are recovered.

Diet

While some families go back to giving their children foods that aren't permitted here, many do not. Wheat, dairy, and soy aren't "healthy" foods for any of us to be consuming. After reading Olive Kaiser's section on gluten, you may have a tough time giving your child any bread products ever again. That said, once the gut is healed, it should be able to tolerate many of these foods again, to the extent that a "healthy" person would.

CD

Even into the recovered stage of autism, many of the families keep using CD (oral, baths, enemas) to reach an ATEC of zero... or until there are no behaviors left from the autism. Once that is accomplished, then a maintenance plan of one oral dose on Monday and one oral dose on Thursday is enough. Typically, such a dose is 3 drops for children and 6 drops for adults—all in one dose on each of those 2 days. However, each family does this a little differently depending on weight and what the optimum dose was while on the full protocol.

The more I learn about how toxic our planet is (i.e. GMOs, toxins in the air, drinking water, parasites in the environment, etc.), the more I realize how critical it is to be on a at least a limited long-term detoxing program. This is especially true for someone who was, or may still be immunocompromised. I do believe that once a child recovers from autism, their immune system can do the job that it was designed to do. However, healing takes time, and every case is different. Some people ask me if recovered children still take CD once recovered—and to be honest—many do not.

Parasite Protocol

It is impossible to avoid parasites in our world. As Andreas and Miriam explained to us, their larvae are nearly indestructible in nature and can live for years. A strong immune system, as well as sufficient stomach acid, may prevent a parasitic infection from taking hold in the body. To make sure, it is best to do a "mini protocol" every 3 months. You can do the full 18 days every 3rd month or a 1-week mini version—whichever you prefer.

Other Supplements

While ocean water is necessary when detoxing with CD, it is extremely beneficial to all of us as our soils are almost completely devoid of nutrients and are contaminated with heavy metals and toxins. Also, it has been theorized that many GMO crops are lacking in nutrition as well. The 90 bioavailable nutrients in OW can certainly go a long way in keeping our bodies healthy. I especially love that it helps with conductivity in the brain. And, it is conducive to speech and cognitive function as well.

Depending on each individual's needs it may behoove a family to keep giving MCT oil or Vegan 3,6,9 omegas. Certain supplements may still be required if there is a history of seizures.

Chelation

Gentle chelators like Bio-Chelate™ can continue to be used into the recovery phase as heavy metals are present in our air and water.

Hyperbarics

Nothing brings down inflammation quite like a hard hyperbaric chamber. However, 10 to 20 sessions can be very cost prohibitive for some families, making it a very personal decision as to whether or not to continue with HBOT after recovery.

GcMAF

I never felt better than when I did a course of GcMAF injections on myself. Energy went up and so did my ability to concentrate. Again, GcMAF is a fairly expensive intervention and that will factor into each family's decision as to whether to keep using it. or not

In a nutshell, I would at least continue with CD two times a week, and a mini parasite protocol every 3 months, just as a preventative. Staying healthy is as important as getting healthy.

> Um, YEAH so just took an ATEC. We have officially been on the protocol for 12 months now and finishing up our 11th PP. ATEC of THREE!!!!!!!!!!!!!! 57 to 3 in 12 months. Henry's speech is what is keeping us from 0. He has great conversation with people that are patient and don't interrupt him so we are so close to a BIG ZERO!!! 57 to 3 in 12 months!!!! Far cry from the severe autism diagnosis he got at 2.5 years old. Here is to the doctor that told me my son should be medicated, would never speak, and will probably need to be institutionalized. I hope YOU never say those things to another person!!!!! All you need is hope, determination, and LOVE.

Heali
the symptoms kno
Auti
l Rivera
mberly McD

Chapter 14

Miscellaneous Information You Should Know

"Always bear in mind that your own resolution to succeed is more important than any one thing."

~ Abraham Lincoln

Seizures

An estimated 40% of children with autism suffer from some form of seizures. Seizures can be absolutely terrifying for both child and family and deleterious to the health of a child who is already suffering from various issues. I have talked to many different professionals about seizures and in turn have heard just as many different opinions on why this happens. Many doctors are unsure, while some simply don't express their opinions. Some say it is a virus in the brain or heavy metals, inflammation, or pathogens/parasites. In a presentation by Dr. Andreas Kalcker, he stated that high levels of ammonia in the body can cause seizures and the high levels of ammonia in the body come from parasites. Since we know that CD kills pathogens/parasites, neutralizes heavy metals, and helps to reduce inflammation, it is not surprising that we have noticed reduced, or even eliminated seizure activity in children with seizure disorders that use CD in combination with the Parasite Protocol.

MCT (medium chain triglyceride) oils, which come from the Ketogenic Diet out of UCLA, are also a very important piece to healing the body and eliminating seizure disorders. There is plenty of information on the Internet that discusses the results of people (on the spectrum and otherwise) using MCT oils for seizure disorders.

The following is a list of things that parents on The Protocol have used successfully to reduce/eliminate seizures in their children.

- CD
- Parasite Protocol
- MCT oil

- GABA
- Hyperbarics at 1.75 ATA
- Ketogenic diet
- SCD diet (grain free)
- L-Carnitine
- P5P with magnesium
- L-Carnosine
- Taurine
- DMG
- Pycnogenol®

The first four in this list have proven to be the most important and have shown consistent results. If you have a child suffering from seizures, you would be wise to review and consider these interventions immediately. The others have also proven very helpful and can be added in one-by-one, (always leaving three days between each new supplement) while you watch for a decrease in seizures.

When we introduce GABA, we increase the dose daily in search of the appropriate dose for the individual. While establishing the correct dose for one intervention it is not appropriate to add in other supplements. Watch out for sleepiness or undesired behavior with GABA. In these cases, I would personally discontinue its use, as with everything there is not one formula that is right for every child. Similarly, if you see hyperactivity or undesired behavior with DMG discontinue its use.

All of the above supplements/interventions have proven very important for eliminating seizures, in different combinations, for different families. Whether the seizures are grand mal (tonic clonic) or petit mal (absence), they start to fade away as a result of the Protocol.

Sometimes, when a family starts corresponding with me, their child may already be taking some of these supplements, and they continue to do that. They may have already taken some of these and seen no improvement. Some other families live in parts of the world where these supplements are unavailable, and in that case we focus on CD and killing parasites.

As I mentioned, a critical piece for some of the children with seizures has been the implementation of the Parasite Protocol. This has also led to a reduction or cessation of seizure activity in many children with seizures.

Interestingly, once some moms were made aware of the parasite/seizure connection it was easy to time seizure activity/anxiety/aggression to full and/or new moons. As Dr. Andreas Kalcker talked about in the parasite section, the full moon period is when parasites are mating, many in the gastrointestinal tract, and therefore "wormy behaviors" (anxiety, SIBs, depression, tantrums, aggression, and teeth grinding to name a few) will generally be higher around that period. New moons can also cause an increase in these behaviors.

Children who have a seizure disorder or are self-injurious need to avoid food coloring, eggs, tomatoes, pork, and chocolate. When adding in the supplements for the seizures that we listed above, follow the same rules as before, never adding more than one supplement every three days, and always carefully observing and keeping track of any and all changes in your notebook so we know what causes what.

After starting anti-parasitics with the help of her daughter's doctor, one veteran mom introduced the CD protocol; a combination that has been extremely beneficial for her daughter, who suffered from frequent grand mal (tonic clonic) seizures.

Her Advice:

Most of my understanding of how parasitic infections can lead to seizures involves my own tracking of seizure activity patterns in my daughter, which coincided with new moon and full moon periods. Then, since MOST doctors want references to what they feel are 'reliable sources,' I started by Googling "NIH parasites seizures," which brought me to:

ncbi.nlm.nih.gov/pubmed/18754958
Helminthic parasites and seizures. [Epilepsy. 2008][1] *- PubMed - NCBI*

I would strongly suggest that you document patterns of seizure activity against the Farmer's Almanac of Full Moons (you can even look up past Full Moons) and marry that information with info from sites like the one above. Also, I would highly recommend you hunt down a pediatric infectious disease specialist in your local area (start by researching pediatric staff at local hospitals) and bring with you all information/insight to your first appointment.

Her Testimonial:

From her very first drop of activated CD in early July 2012 (actually her very first one-eighth drop to be exact), I witnessed an improvement in my daughter. Within about two hours of drinking her first ounce of the activated CD water solution (one activated drop of CD mixed into eight ounces of water), I witnessed her sitting still in our kitchen and observing me in such a way unlike she had ever done before. She was calmly observing me instead of running around the kitchen. She was taking my actions in. She was studying me. Her eyes were focused on me, rather than constantly scanning everything else in the room. Both her facial expression and her body language were calmer than I had ever witnessed before. She was more focused, more centered. She appeared content.

This change that I witnessed in her was so noticeable and so sudden, that I couldn't help but think to myself, "So this must be why the first word in MMS is Miracle." It was that amazing of a change in those moments.

Shortly after those incredible moments, my daughter went on to eating contently and then jumping on the neighbor's netted trampoline. She has enjoyed the neighbor's trampoline for years; yet on that particular day, only a few hours after her first drink of a tiny amount of CD, I noticed something quite different in her behavior. As I observed her from the windows of our house, I could see that she was jumping less than usual and instead looking around a lot more. She was calmly looking around at the neighbors' houses, down at some birds in the grass, up at the trees, the clouds, etc. It was all as though she was perhaps truly noticing them for the first time. She had glanced at these things before that day, sure, but she never before appeared to be studying them in the manner she was in those moments on the trampoline that July afternoon. She did engage a bit in some of her tic-like spitting movements with her mouth while on the trampoline, but to nowhere near the intensity, nor duration that would more typically have consumed her prior to that day.

When she came back into the house after using the trampoline and observing her world, I asked her, "why were you spitting out there?" To my surprise, she then very calmly replied, "because it's sensory." Wow! She truly was more connected somehow to her world than ever before. Something good was happening to my girl—only hours after drinking one-eighth of a drop of activated CD with one ounce of water. From those moments on I felt encouraged in my heart and soul that CD could possibly be a very important catalyst toward improving my daughter's health and well-being.

With the help of a caring local doctor, we had already begun some trials of anti-parasitic medications on our daughter in early May 2012, prior to starting CD, in response to behavioral struggles that had been consistently intensifying around full moons every month. (Later, we noticed new moon behavioral intensities as well). We had seen some improvement with anti-parasitic medication, yet CD seemed to somehow boost our daughter toward further improvements that much quicker. In fact, our daughter's ATEC score before starting CD measured 79 on July 8th, 2012 (the day before starting CD), but as of September 30th had dropped down to 55 (only two and a half months later).

Now, in early December 2012, our daughter still appears to be on an overall upward trend of improving health and stability, attributable to the synergy of CD and anti-parasitic medications. Natalie's interpersonal skills have improved, her growth rate has improved, her focus has improved, and her frequency of seizures has lessened.

In fact, at this writing she is 53 days seizure free, which is quite an improvement considering before she was on a terrifying pace of grand mal (tonic clonic) seizures occurring every 2-3 weeks. These are very dangerous, life-threatening events. Her seizures were rapid onset, heading toward grand mal within seconds. They could happen anywhere, i.e. while in her car seat coming back home from a Dr.'s appointment, while driving away from Costco, while running in the backyard playing with the dog, or swinging high on the swing set in the backyard with a neighbor girl. Yet another occurred while wading in the lake (with life vest on), another hit while in our neighbor's four-foot swimming pool with me and her brother. She was right next to me, walking around and then suddenly, she was in the throes of a rapid-onset grand mal seizure that required me to lift her out of pool into her NT (neurotypical) brother's arms. May these be the last my daughter ever knows of seizures.

It is our hope that, in time, CD will be able to take on the lead role as the primary health-support for our daughter, if not, perhaps even someday the sole healthcare support to her. We're looking forward to the days when we may be able to remove or reduce many, or perhaps even almost all, of her current supportive medications (i.e. antibiotic, antiviral, anti-seizure, etc.). With the caring support of parents of The Thinking Moms' Revolution, as well as some excellent local doctors, we suspect that our daughter's life will continue to improve. In the almost 14 years of caring for our daughter in her various states of medical fragility, we have never ever had this much Hope. We feel fortunate to have been swept up into this Revolution and embraced warmly by so many Thinking Moms who are also in the process of working toward more healing for their children. Working to heal illness is so much nobler than working to mask illness, wouldn't you agree? God Bless Jim Humble, Andreas Kalcker, Kerri McDaniel de Rivera and all the TMR families!

UPDATE: January 30th, 2013

I too am so thankful for Kerri and her whole 'crew'. Thanks to Kerri for bringing CD to all of our kids on the spectrum worldwide. My Natalie is now OFF of her former roller-coaster ride of prescription anti-fungals, OFF of her former roller-coaster ride of prescription antibiotics, OFF of prescription antiviral, OFF of prescription low-dose naltrexone, etc. Her belly distention has disappeared, she's gained weight, she's gained height and she's now almost 103 days seizure free (as of 6:16pm tonight). She still has a ways to go toward more improvements, but she's happily climbing up from bottom of the trenches; she's climbing up from what seemed like the very bottom of the abyss that vaccine injury/injuries had placed her in. Thank you Kerri, Andreas Kalcker, Jim Humble and all of the moderators on the supportive websites!!!

UPDATE: February 1st, 2013

My Natalie's long run of 105 seizure-free days has ended. After being off of Albenza® less than 24 hours, BOOM! a grand mal (tonic clonic) seizure occurred yesterday evening. This same thing happened the last time we stopped Albenza®, sooooo we're not sure if it's withdrawal related or parasitic activity related. We're going to closely watch her behaviors this week for signs of parasitic burden. We put her on Alinia® now, just to give her some sort of parasitic coverage while we wait to likely restart Albenza®, soon. She did fine on Alinia® 200mg BID in the past. I'll keep you all posted, but for now we unfortunately need to restart the clock on my Natalie's 'seizure-free days' tracking. I'm bummed, but thought I should keep you in the loop on this; as I've been soooooo excited over the past three months about her freedom from seizures, which I believe was due to the synergy of CD and Albenza®. I'll touch back if/when we learn more.

UPDATE February 2nd, 2013

O.K. so, I just wanted to kind of round out this long thread a bit with my thoughts, for whatever they're worth: I really truly suspect that perhaps too sharp of a drop off (lack of weaning) from the systemic med, Albenza® lead to the seizure. Natalie is stable right now on a low-dose interim of Alinia® 200mg BID for now until all insurance approvals are in for the delivery of more Albenza®. This leads me to believe that the several courses of Albenza® over the past months likely HAS lowered her overall parasitic load/burden fairly well thus far (it's a pretty broad-spectrum anti-parasitic, from what I understand). However, I don't think for a *minute that Natalie's broad spectrum anti-parasitic medicine needs are finished. For this reason, we'll resume Albenza® as soon as we're able.*

Although further discussion/negotiation between myself and Natalie's pediatric infectious disease doctor is required, my thought is that she may need an entire year on Albenza. However, certainly in the future, whenever we stop Albenza®, I'm going to plan for a slow weaning off period so as not to disrupt Nat's delicate nervous system. Also, I'm kind of thinking that perhaps if our kids 'wig out' when CD is eventually dropped that maybe we've not lowered the overall parasitic burden enough, as yet to allow their immune system to 'kick in gear' and properly function??? I don't know, obviously, just theorizing with the brain matter God has given me.

SUMMARY & UPDATE FOR SECOND EDITION: December 1st, 2013

From birth to age 7 years, my daughter Natalie unfortunately had her share of health and developmental struggles. Her diagnoses began with 'Failure to Thrive' shortly after birth, followed thereafter by a Mitochondrial Encephalopathy Partial Complex III Deficiency diagnosis by age 7 months, then Hypothyroidism as well as global developmental and cognitive delays obvious from there. Despite significant medical and developmental challenges, Natalie entered a Kindergarten designed for the 'mild cognitively impaired' (formerly referred to as educatably mentally impaired 'EMI') with daily opportunities for mainstreaming with the typical Kindergarten classroom next door. After a fall 2005 Flu shot early in that same Kindergarten year, however, all bets were off. She regressed very rapidly in every way: medically, developmentally, and behaviorally. Within months of that shot, a laundry list of additional diagnoses began to form: Sensory Integration Disorder, Autism, Epilepsy, P.A.N.D.A.S., P.A.N.S., and Combined Immune Deficiency, to name a few. After much study by a large team of educational specialists, Natalie was reclassified by school professionals by the following April of 2006 to the 'SXI" category of disabilities (Severely Multiply Impaired), having deteriorated terribly from her former classifications of POHI (Physically or Otherwise Health Impaired) and 'CI' (Cognitively Impaired) within only months of that Flu shot. Her significant healthcare needs and behavioral challenges since have necessitated assignment of a one-on-one Para-educator to her case, even while she's been schooled within high adult-to-student ratio basic education special- needs classrooms over the past several years.

In 2011, between the ages of 12 and 13, I began to notice patterns of apparent P.A.N.S. flares in Natalie, which seemed to coincide with the New Moon and Full Moon weeks. When I began to study this phenomenon a bit more closely, I also recognized a pattern of seizure activity that tracked these timeframes as well. In reading about potential parasite activity that is thought to link to the New Moon and Full Moon timeframes, I requested anti-parasitic medications for Natalie. Thankfully, a local Pediatric Infectious Disease doctor kindly accommodated my request. Each time that the doctor prescribed various anti-parasitic medications for Natalie, she would improve. The improvements were noticeable both at school and at home. It was incredible to witness. However, each and every time that she was off of such medications, she would regress back to her former state of illness (displayed primarily by significant behavioral regressions). So, we found that we needed to try to keep her on various anti-parasitic medications more often than off of them, in order for her to enjoy overall improved health and neuropsychiatric stability. Even so, there was still some sort of physiological strain that she was very *definitely experiencing, but we were unable to decipher what it could be. This was frustrating, of course; to witness her continued rollercoaster of ups and downs in health and her various psychiatric disturbances each month.*

It wasn't until a friend suggested we get Natalie started on CD, that the health and behavioral 'rollercoaster ride' began to show signs of smoothing out. In July 2012, after her first drink of CD (only one-eighth of a drop), Natalie began to show initial signs of relief from whatever it was that was still ailing her. Her connectedness and overall psychological state of being changed within hours of that first one-eighth drop of CD. After that first day, we of course then continued with a very slow and conservative ramp-up of CD (orally, then baths, then enemas) alongside various anti-parasitic medications and all of the components of the Kalcker PP (with frequent swap out of Mebendazole with Albendazole, predominantly). With each passing month over the past sixteen months, we've been able to witness improvements in Natalie; perhaps a bit more slowly than some of the younger and/or potentially less health-compromised kids that Kerri's protocol has healed, but improvements nonetheless. Before CD, Natalie's ATEC score was a 79. As of this writing, her ATEC score is a 39. Even more astounding to us than that improved ATEC score have been the changes witnessed thus far in her overall health status and health regimen. Here is a list of some of the improvements that I can very easily recall as I write this (there are likely many more that I could comment upon, if I had more time):

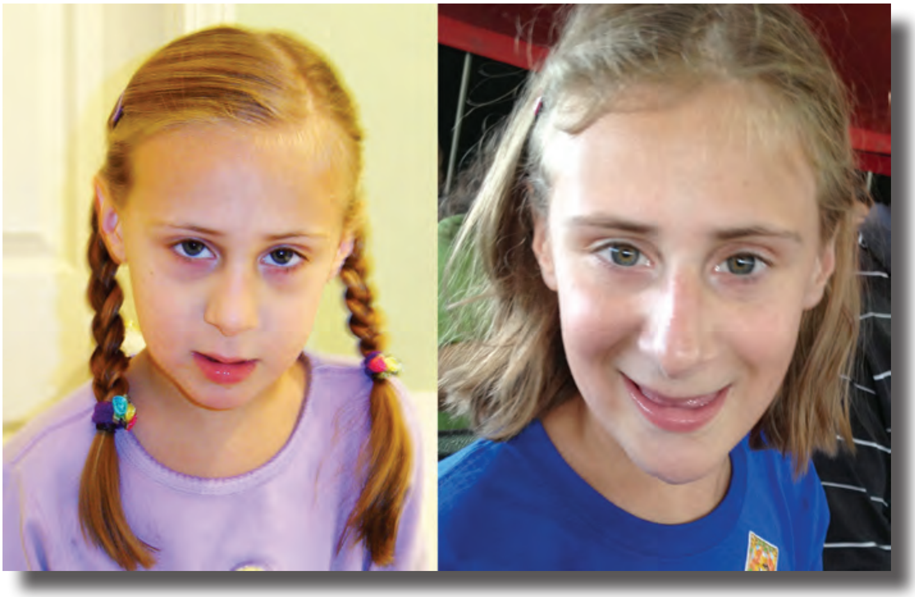

Natalie in 2006 (left) and in 2013 (right)

1. *Daily prescription anti-fungal medications no longer needed.*
2. *Daily prescription antibiotic medications no longer needed.*
3. *Daily prescription antiviral medication no longer needed.*
4. *Daily prescription LDN (low dose naltrexone) immune modulator removed.*

5. *Daily prescription Leucovorin removed without ill effects noticed, as well as several other daily supplements that might have been ultimately not helpful.*
6. *Prescribed Immune Globulin (weekly subcutaneous, formerly monthly intravenous for years) removed.*
7. *Distended belly gone & improved posture (no longer hunched over).*
8. *Formerly very frequent tongue sores/lesions lessening more & more.*
9. *Seizures formerly as high as every 17-21 days in her worst times; now she's enjoyed seizure-free stretches as high as 105 days (3.5 months) followed shortly thereafter by a seizure-free stretch of 168 days (5.5 months). Prescription anti-seizure medication reductions have also been enjoyed as well as fewer headaches resultant of those reductions.*
10. *P.A.N.S. flares controlled with the combination of CD and anti-parasitic meds. Neuropsychiatric symptoms of flares very dramatically declined.*
11. *Overall appearance of improved health, improved language and interpersonal skills, improved attentiveness (per Neurology's opinion at recent appointment).*
12. *Improved stamina and strength (core, legs, etc.) noticed at Special Needs Performance Cheerleading practices, physical therapies, etc. Her ability to peddle adaptive bike has also improved.*
13. *No more nose bleeds, no more ear infections, no more apparent seasonal allergies, fewer headaches (as noted above).*
14. *Incredibly improved sleep and need for naps only once in a while now (as opposed to formerly daily).*
15. *Improved growth rate and normalized appetite.*
16. *Improved language (more descriptive words, more sentences).*
17. *Healthier-looking skin overall; softer, less dry, improved color & ability to tan.*
18. *Improving oral muscular abilities, able to swallow pills/capsules.*
19. *Capable of successful dentist appointments and now even Orthodontist appointments, with calm and ease.*
20. *Increased success rate during therapies (Speech Therapy, Occupational Therapy, Physical Therapy, etc.).*
21. *Less rigidity overall, fewer transition issues, cooperates more often than she ever used to.*

With all of the well-documented improvements noticed in Natalie at home, at school and in doctors' offices, I don't have the same level of fear as other parents might surrounding our open use of CD, as Natalie's doctors are all well informed and quite supportive. We've also formally notified her social worker (both verbally at a home meeting and in writing) of our use of Kerri's CD Protocol (including Kalcker's Parasite Protocol and of course a GFCFSF/low sugar diet) while under Natalie's local medical team's support and medical monitor. We're 'out of the closet', so to speak, when it comes to our use of CD for improving Natalie's health. It's been quite obvious to Natalie's doctors; actually, that Kerri's protocol has been extremely complementary to their medical efforts for improving Natalie's health and overall quality of life.

There's of course no way to predict with certainty how far my sweet Natalie will be able to ultimately improve via this protocol, but based upon what we've witnessed thus far, I do feel that her current best hope for continued improvement lies within the pages of this book; as these healing strategies of Kerri Rivera's have brought Natalie toward more health (physical, neuropsychiatric, emotional and even some initial cognitive improvements) than any other medical effort attempted since her 2005 Flu shot injury. All other medical efforts for Natalie over the years, I feel, perhaps only temporarily patched her severe illness. Kerri's methods actually work to attack disease, kill it and remove it. Yes, that's right. Eight years after Natalie's tragic rapid spiral downward, it's now finally a time of healing for Natalie. Thank you, Kerri Rivera, for all of your help given to our girl and to so many children around the world.

Self-Injurious Behaviors (SIBs)

Self-injurious behaviors (SIBs) can be absolutely devastating for families with children on the autism spectrum. Common self-injurious behavior includes: headbanging, hand biting, pinching, slapping head/ears, eye poking, hair pulling, scratching/rubbing, throwing one's body on the floor, etc. The horror of watching your own sick child further inflict pain onto himself, and the helplessness of not knowing what to do to help him is heartbreaking. I have met with families through the clinic who have had terrifying episodes of SIBs in front of me as well as at home, and many of these families as well as others are now using the Protocol to help their children with very good results.

There are many theories as to why some children have SIBs. The article "Self-Injurious Behavior" by Stephen Edelson on ARI's website discusses several different theories and supporting research.[2] The theory that most fits what I see on the ground is that of pain alleviation. Many of our kids are in chronic pain and cannot tell us. Sadly, we haven't always had answers to help them.

If we follow this thinking, and assume that many of these children suffer from severe gastrointestinal pain, or headaches (supported by head banging, gut clutching, etc.) we can hypothesize that much of this pain can come from pathogenic overload, parasites wreaking havoc on the body and the resulting inflammation.

Families applying CD in combination with the Kalcker Parasite Protocol are reporting a dramatic reduction in SIBs in their children. I believe there is an important connection to be made here, as we eliminate pathogens and parasites SIBs decreases. More research is needed, but for families facing the horrors of self-injurious behavior, this combination could be of great value. (More testimonies on decreased SIB's can be found in Chapter 2, and the Miracles and Testimonials section at the end of the book starting on page 357.)

> Prior to starting CD, we had tried almost every autism biomedical protocol. We saw several DAN and GI specialists and numerous other practitioners. We spent tens of thousands of dollars on consults, supplements, diets, medicine and medical procedures not covered by insurance. My daughter either had no response or negative ones and she continued to get worse. She developed extreme self-injurious behaviors that progressed to the point of her needing to be restrained several times a day in an attempt to prevent her from pinching or biting herself until she drew blood. Her body was covered with bruises and cuts. She would also pull out big clumps of her hair. She would be up for hours screaming in the night with these behaviors. She was severely malnourished despite the fact that two registered dieticians had deemed her daily caloric intake adequate. She was 14 years old and weighed a little over 70 pounds. I honestly did not think that CD would be any different than anything else we had ever tried but figured things could not get much worse.
>
> Since starting CD her self-injurious behaviors are almost entirely gone. She went from having to be restrained almost daily to once in the last 5 months. Her former multiple daily 45 minute self-injurious tantrums are now a few minutes of crying or protest and happen once a week if at all. We have noticed gains in language and cognition (20 point drop in her ATEC score). After years of fluctuating between loose stools and constipation, she is now regular and has gained almost 10 pounds. She sleeps through the night and we are able to take her places that we would have never dared to before. If my daughter can make gains this drastic in 5 months at age 14, I can't wait to see what the future holds.
>
> Julie, Colorado, USA

PANS, PANDAS and PITAND

PANS stands for *Pediatric Acute-onset Neuropsychiatric Syndrome*. It encompasses PANDAS (Pediatric Autoimmune Neuropsychiatric Disorder Associated with Streptococcal Infections) and PITAND (Pediatric Infection Triggered Autoimmune Neuropsychiatric Disorders), where PANDAS is a subset of PITAND. Dr. Susan Swedo, Chief of Pediatrics & Developmental Neuroscience Branch at the National Institute of Mental Health (NIMH), and her research team identified a subtype of OCD in children during the 1990s that is triggered by infection. Streptococcal infection has been most studied; however, any infection can trigger this condition.[3]

I have run across these diagnoses a lot since I have started helping families in first world nations. There seems to be an overlap in some autism and PANS/PANDAS/PITAND cases. I have heard that around 25% of children with ASD's also suffer from PANS/PANDAS, and/or PITAND. Some practitioners would put that figure even higher. However, since further research is needed, and this field is still developing it is impossible to say for certain at the moment.

According to the International OCD (Obsessive Compulsive Disorder) Foundation some of the symptoms of PANDAS/PANS may include:

- Acute sudden onset of OCD.
- Challenges with eating, and at the extreme end, anorexia.
- Sensory issues such as sensitivity to clothes, sound, and light.
- Handwriting noticeably deteriorates.
- Urinary frequency or bedwetting.
- Small motor skills deteriorate - a craft project from yesterday is now impossible to complete.
- Tics
- Inattentive, distractible, unable to focus and has difficulties with memory.
- Overnight onset of anxiety or panic attacks over things that were no big deal a few days ago, such as thunderstorms or bugs.
- Suddenly unable to separate from their caregiver or to sleep alone.
- Screaming for hours on end.
- Fear of germs and other more traditional-looking OCD symptoms.

I know several moms who have been able to reduce/control/eliminate their children's symptoms with CD. In fact, families began to eliminate the antibiotics that they were using as they reached full dose of CD. I am by no means suggesting anyone reduce or remove a prescription medication without a doctor's supervision. I am simply reporting on what I have witnessed.

We know that CD kills bacteria, so we could hypothesize that CD could have a similar effect on the PANS/PANDAS/PITAND symptoms that an antibiotic would. What has worked in these cases is upping the number of doses of CD given a day so as to kill pathogens around the clock. For PANDAS/PANS/PITAND use two baby bottles a day. If your max dose is eight drops in an eight ounce bottle, and a PANDAS flare presents, prepare two baby bottles with eight drops in each and give one dose every hour over the course of 16 hours. Some families are doing 16 doses a day even if there is no flair and the child is doing excellent.

As noted previously, in the absence of research, we cannot recommend doing nothing for these children. Their suffering is intense and the risk of long-term damage cannot be ignored. We also do not know if some of these children remit spontaneously later in adolescence. We do not know if some of them grow into adults with the most treatment resistant OCD due to permanent basal ganglia damage. While we research those questions, offering treatment options at the local level, especially in the first few months after onset, could be critical to the long term success for these kids.[4]

We parents are pioneers in the research surrounding PANS/PANDAS/PITAND, so it is critically important that we document what we are doing with our children and what is working for them. In fact, A MOM was the first to discover the connection between strep and OCD/tics. Her child was the first documented PANDAS child. Dr. Susan Swedo reinforces that physicians should listen to what the mothers have to say![5] Speak up to your doctor, share your findings, join our forum to learn, and share what is working for you. Parents are driving the biomed movement (and always have) for ASDs as well as PANS/PANDAS/PITAND. In fact Dr. Bernard Rimland was a parent and the first to suggest biomedical interventions for people on the spectrum. We as parents can help countless other families by sharing our experiences.

You will find more info on PANS/PANDAS/PITAND at:

www.pandasnetwork.org

and

ocfoundation.org/PANDAS

Fever Therapy

Dr. Kenneth Bock and Nellie Sabin, in their groundbreaking book *The Road to Immunity: How to Survive and Thrive in a Toxic World*, explain how to use fever therapy and how it is an incredible tool for healing.[6] When a mom emails me to tell me her child has a fever, it tells me that the immune system is reacting to an invader, and we have an excellent opportunity to kill pathogens in the body. Having a fever is part of an effective immune response. This is also true in infants as well as small children. Using over-the-counter (OTC) fever reducers interferes with the body's own ability to fight infection.

This section was excerpted from the aforementioned book.

> Fever has long been recognized as an important tool in the process of healing. The process of heating the body is called "hyperthermia." Hyperthermia, or ordinary fever, can be very effective against cancer cells, various pathogens, or invading organisms that cannot "tolerate the heat." Hyperthermia may not kill all pathogens in the body, but may reduce the number of invaders in the body to a manageable level for the immune system. Hyperthermia may also be responsible for stimulating the immune system to produce cytokines i.e. antibody production, and release of toxins from fat cells.

What Temperature Constitutes a Fever?

According to Dr. Robert Sears, the following are normal temperatures in children over 3 months:[7]

Normal temperature: 97° to 99° F (36° to 37.2° C)
Low-grade fever: 99° to 100.9° F (37.3° to 38.3° C)
Common fever: 101° to 103.5° F (38.4° to 39.7° C)
High fever: 103.6° F (39.8° C)

Fever Therapy in Practice

If your child develops a fever, it is not a failure of the body that needs to be corrected, but a mechanism that brings the body closer to healing. Rather than administer an OTC remedy to reduce fever, apply a cool, damp washcloth to the forehead and neck, replace them as they warm, and let the body do its job. Some parents put damp socks on their child's feet.

During a fever you can dose orally, as well as do one CD bath and one CD enema in one day. Keep dosing CD every hour (16 to 24 doses a day), unless the person stops eating. Even if you must reduce the number of drops, even if you are only giving one drop, keep dosing every hour. CD baths, as well as baths in general, are also your friend if your child has a fever. Some parents have given several baths throughout the day while a fever is present with great results. Bath the child if the fever spikes. Always be vigilant that the temperature does not rise too high or too fast so as to cause febrile seizures, we have to use common sense and trust our instincts. Bath and keep a cool damp wash cloth going 24/7 until the fever pases.

According to the National Institute of Neurological Disorders and Stroke febrile seizures are convulsions brought on by a fever in infants and small children. The majority of children with febrile seizures have rectal temperatures greater than 102° F. Approximately 1 in 25 children (4%) will have one febrile seizure. Although febrile seizures can be frightening to parents, the vast majority of febrile seizures are fast and harmless. There is no evidence that short febrile seizures cause brain damage. Large studies have found that children with febrile seizures have normal school achievement and perform as well on intellectual tests as their siblings who don't have seizures.[8]

If your child has a febrile seizure, experts recommend that you try to stay calm, keep the environment around your child safe (place them on the floor if they are not, and clear objects out of the way), and allow the seizure to resolve spontaneously without restraint. Try to look at your watch so you will know how long the seizure lasted. Do not put anything in your child's mouth as it may break, obstructing their airway.

Skin Issues

I have been hearing about skin issues right and left: rashes, hives, redness, acne, eczema, psoriasis, pustules, etc. It is very easy to assume that since skin is on the outside of our bodies that these reactions are caused by exterior factors (i.e. irritation from a detergent, mold, etc.). However, the skin is our largest detox organ, and many pathogens seek exit through the skin causing many of the above conditions.

The scenario that I keep hearing is the following: We just started the Kalcker Parasite Protocol. Little Johnny has a rash with white heads on his legs. What do I do? When the families keep up with the parasite protocol, and continue dosing daily with CD these eruptions disappear, and this is generally accompanied by improvements in behavior, or hyperactivity/anxiety.

We need to think about skin issues as outward manifestations of inner toxicity, and use our protocols accordingly. Skin issues can be a good indicator that something is being flushed from the system, or something needs to be flushed from the system. The Kalcker Parasite Protocol combined with CD tends to eliminate skin issues at the root.

The following paragraphs were excerpted from leakygut.co.uk, written by Dr. Gloria Gilbere ND DA hom PhD[9]:

> Experts agree that most chronic skin disorders, specifically psoriasis, is the external manifestation of the body's attempt to eliminate internal toxins that have accumulated within the lymphatics and blood stream by "seeping" through intestinal walls in a condition known as leaky gut syndrome.
>
> In working with clients around the world, the protocols that have been most effective are those that include repairing the intestinal walls by first beginning a protocol to eliminate candida yeast and parasites.
>
> The most effective cleansing protocol is obtained by clearing out these "uninvited" health-depleting microorganisms in order to release the accumulated pressure on the liver and the intestines—allowing free passage and thus reduction of overall toxic load.

The following paragraphs, authored by Dr. Edward F. Group III, DC, ND, DACBN, DABFM, were excerpted from parasite-cleanse.com/intestinal-parasite-symptoms.html[10]:

> **Symptoms of Intestinal Parasites Related to Skin**
>
> The presence of intestinal parasites, at times, manifest as allergies, tumors, and various other skin conditions.
>
> Intestinal parasites cause the human body to release hormones and immune system defenders, which in turn irritate the skin and cause various skin conditions such as acne, hives, rashes, itching, weeping eczema, ulcers, swelling, sores, lesions, blisters on lips, and dermatitis. It also deprives the skin of its usual glow.
>
> Intestinal parasites cause irritation and inflammation in the stomach making it difficult to digest certain types of foods. The presence of undigested food particles in the stomach causes the body to produce increased levels of eosinophil, an immune system defender. These eosinophils inflame the body's tissue and cause skin rashes.
>
> Intestinal parasites trigger the body's immune system to develop a tumor-like mass known as granuloma to encase parasitic larvae or eggs. Granulomas usually develop in the colon, rectal wall, lungs, liver, peritoneum, and the uterus.

Btw, you were bang on re his allergies and PP (parasite protocol). He had this horrible red lip thing because of food allergies. He would make it worse by rubbing it. It would get really bad. After first PP (Parasite Protocol) - NO MORE RED LIP ISSUES!

Families often tend to panic when dealing with sudden "unexplained" skin issues. If you are seeing chronic or acute skin issues I think there is enough evidence to look towards parasites as the main or at least a contributing culprit. Therefore, if we work towards removing the parasites causing the symptoms, rather than treating the symptoms topically we can expect an alleviation of the symptoms.

If you are seeing acute or chronic skin issues and you have not started a Parasite Protocol, I would suggest you revisit the parasite chapter and go over the recommendations with your healthcare provider. Keep dosing CD, in some cases upping the dose has helped resolve the issue faster, and in the case of severe irritation using a CD spray topically can provide relief (10 activated drops per ounce of water in a spray bottle) and a CD bath can also help. I have seen most, if not all skin issues resolved with CD and/or the Kalcker Parasite Protocol.

Healing Adults with Autism

Common thinking in the biomed community and the autism community at large is that autism is only "curable" or "recoverable" up to a certain age. I have heard that two to nine years old is the "golden window" for autism, and that after nine years of age improvement would be possible, but not recovery. The neural pathways would be set, the damage done, and the child would retain the autism diagnosis for life.

Not anymore. This year a 15-year-old recovered from autism, as well as two 17-year-olds. In addition, I am helping an amazing mother of a 31-year-old with autism that began the Protocol and for the first time in 28 years was seeing real progress with her son. In fact that young man's ATEC score dropped from a 65 to somewhere in the 20s in his first three months on The Protocol. I want to prove that autism recovery is possible at any age. It may take longer, and the road may prove more challenging, but it's already happening. Families that have been using biomedical interventions for more than ten years are finding the Protocol and sharing their gains.

Older children and adults on the spectrum start the same way as the little kids—clean up the diet first, and then on Day one of CD, give one drop in eight ounces of water over eight doses throughout the day. It is so important to start low and slow even with adults because they have often been chronically ill much longer than young children and have deep rooted pathogenic and parasitic infections. This means that when we start giving CD they might experience a detox reaction considering the amount of pathogens that will be dying.

Working with Neurotypical Adults

Can CD benefit parents and other adults not on the spectrum? Absolutely. You can literally find thousands of testimonials on the Internet (Jim Humble's website, www.jimhumble.org or www.curezone.com, etc.) on people who successfully detoxed, or healed a myriad of illnesses using CD.

Many parents who have seen improvements in their children have decided to use CD, or the Kalcker Parasite Protocol for themselves. I have done CD oral, baths, and enemas on myself, especially when I feel "something coming on" and with great results.

It's important to keep in mind that just because we weigh more than our children doesn't necessarily mean we should jump in and go to full dose from one day to the next. The best way to start an adult who may have 40 or 50+ years of accumulated toxins/pathogens/parasites is low and slow. Use the baby bottle method and start day one with one drop spread out over eight doses and go up from there. Chart your progress, see how you feel, go low and slow. When we as adults use the protocol it can be great for us, but also a window into how our children are feeling.

I love to see a mom or dad begin CD with their child. Both begin at one drop on day one in eight ounces of water, 1/8 or a drop per dose, for eight doses throughout the day. Starting with our children gives us a much better understanding as to how our child is feeling, and when we up the dose how that feels as well. This includes enemas. Once parents see how good they can feel after an enema, it helps garner confidence, if we had doubts about using them for our children. Applying our own enemas can help to remind us just what a powerful tool they are for healing.

Many of the parents doing the protocol for themselves write me to tell me about their remission from fibromyalgia, chronic fatigue syndrome, Lyme disease, life-long constipation, etc.

We all do the parasite protocol in my house so as not to re-infect my son, as parasites can be highly contagious. Last year I did a round of parasite medications prescribed to me by a doctor. During the course of treatment I felt a tremendous amount of anxiety and some depression (feelings I had never felt before) right before I passed worms, I can assure you it felt awful. That experience motivated me to encourage parents to get these invaders out of their children's bodies.

As soon as I started passing worms and detoxing more with CD, I started to feel good again. The cough that I had had on-and-off for years miraculously disappeared. Seriously, go low and slow, a toxic body can only process so much at a time. Too much too fast can cause headaches and diarrhea. The way I see it, there is no need for that, listen to your body and go at whatever pace you need. Don't just "push through," low and slow wins the race. Always remember to drink water when doing any sort of detox!

Tax on a Marriage

During the first five years after an autism diagnosis married couples face an estimated 80% divorce rate, and a 90% chance of divorce beyond the tenth year of marriage. My family was almost a statistic. Even before we had a diagnosis, our family was suffering. Due to Patrick´s constant screaming and tantrums, he wasn't sleeping at night, and that meant that no one else was sleeping either. He had also stopped being the affectionate baby that he was for the first two years of his life, which was emotionally challenging for all of us. We were all over stressed and sleep deprived, not to mention having symptoms of post traumatic stress disorder (PTSD).

If I could give any advice to families dealing with a new diagnosis it would be to start The Protocol and CD immediately. The sooner the affected family member starts to heal, so will the rest of the family, similar to the sun coming out after a storm—this too shall pass.

Financial Stress

Not only is an autism diagnosis emotionally stressful on a marriage, but it can be extremely financially stressful as well. The Protocols outlined here can be extremely cost effective, even if including ocean water. Recently, one of the moms using The Protocol did some calculations comparing what she spent previously (see table on page 342) using another biomedical intervention with many supplements, etc., to what she spends now (see table below).

The Protocol - Estimated Cost

Product	Size	Price	Reorder Schedule	Current Cost per Month
CD	1 bottle	$22	1/mo	$22.00
Diatomaceous Earth	2.5 lb	$15	1/3mo	$5.00
Rompepiedras	50 ml	$24	1/2mo	$12.00
Vortex Ocean Water	1 L	$40	1/mo	$40.00
THERALAC®	30 caps	$36	1/mo	$36.00
Natures Way® Neem	100 caps	$4.09	1/5mo	$0.75
Pyrantal Pamoate (Combatrin®)	32 oz	$35	1/6mo	$5.00
NOW® Castor Oil	16 oz	$8.23	1/6mo	$1.38
Mebendazole	90 tabs	$28	1/5mo	$5.60
TOTAL				**$127.73**

Defeat Autism Now! based Protocol - Estimated Cost				
Product	**Size**	**Price**	**Reorder Schedule**	**Current Cost per Month**
Actifolate	60	$15.00	1/2mo	$7.00
Calm PRT	120	$89.00	1/5weeks	$78.00
Caprilex™	90	$16.00	1/6weeks	$10.60
Cytoflora™	4 oz	$79.00	1/mo	$79.00
Designs for Health Lithium Synergy	120 caps	$22.00	1/mo	$6.00
DMG	90	$16.00	1/5mo	$10.60
Douglas B-Complex	60	$21.00	1/6mo	$21.00
Douglas CoQ10	60	$49.81	1/6mo	$12.50
Enhansa™	150 caps	$58.00	1 / 75 days	$29.00
Enzymedica Virastop™	120 caps	$32.11	1 / 4 months	$8.00
Garden of Life Probiotics	81g	$32.17	1/mon	$32.17
GI Revive	225g	$52.00	1/mon	$52.00
Homocycsteine Supreme	60 caps	$20.00	1/4months	$5.00
Kirkman Digestive Enzyme	200 caps	$60.00	1/3months	$20.00
Kirland Vit E	500	$13.65	1/16months	$0.84
Klaire Multivitamin	180	$28.00	1/3months	$9.00
Liver Life	4 oz	$50.00	1/2months	$25.00
Magnesium	240	$13.25	1/6months	$7.00
Metagenics Adreset®	60	$33.00	1/4months	$8.00
Milk Thistle	100	$25.00	1/3months	$8.00
NDF+	1 oz	$80.00	1/6months	$15.00
NeuroProtek®	4 bottles	$140.00	4/4months	$35.00
NeuroScience Avipaxin	60	$58.00	1/2months	$28.00
NeuroScience EndoTrax	1 bottle	$27.00	2/month	$54.00
NeuroScience Kavinace	120	$65.00	4/4months	$32.50
New Beginnings Copper	50 ml	$20.00	1/year	$1.75
NOW MCT	32 oz	$17.00	1/6months	$2.80
Nutribiotic GSE	4 oz	$16.99	1/6months	$3.00
OSR	30	$60.00	1/2months	$30.00
Oxycell Glutathione Cream	1 jar	$33.00	1/2months	$16.50
Pharmax Cod Liver Oil	8 oz	$13.33	1/mon	$13.33
Source Naturals B6 (p5p)	240	$15.00	1/6months	$2.60
Source Naturals Evening Primrose Oil	90	$7.45	1/6weeks	$4.97
Source Naturals Gaba	100g	$18.00	1/4months	$4.50
Source Naturals Lysine	200	$10.00	1/mon	$10.00
Source Naturals Quercitin	200	$25.80	1/3months	$8.25
Speak EFA	60	$42.00	1.5/month	$63.00
TheaNAQ	90	$54.00	1/month	$54.00
Therabiotic	120	$69.00	1/2months	$35.00
TravaCor Jr.	120 caps	$24.00	1/4months	$7.00
Vit B12 shots	1 refill	$45.00	1/3months	$15.00
Wobenzyime®	120	$60.00	1/2months	$30.00
TOTAL				**$894.91**

Tips for a Non-Toxic Home

Many of our children on the spectrum have chemical sensitivities, and even if they don't, it's wise to limit their exposure to toxins during the healing process as much as possible. Toxins are not good for anyone at any point in their life. The benchmark investigation "Bodyburden: The Pollution in Newborns" by the Environmental Working Group[11] (see Appendix 13, page 499) explains why this is so important for children with developing brains and immune systems. The following is a list of products to use instead of toxin filled, name brand cleaning products.

Floor cleaner: Add one cup of white vinegar and baking soda to a bucket of water (add a squirt of liquid soap if you like).

Air Freshener: To remove odors from carpets spread on baking soda and then vacuum the carpet. Simmer cloves and cinnamon as air freshener.

Removing Toilet odors: Borax

Trashcans: Spread baking soda in the bottom, then spray to disinfect with a solution of borax and water.

Dish soap: I use liquid dishwashing soap and white vinegar to cut grease. If you buy dish soap make sure it's biodegradable.

Disinfectants: Borax with water is an excellent disinfectant, or spray CD.

Laundry detergent: Soak the clothes for an hour in borax soap before putting into the washer.

Oven Cleaner: Dissolve two teaspoons of liquid soap (not detergent) and two teaspoons of borax with enough warm water to fill a spray bottle.

Windows and Mirrors: Club soda in a spray bottle

Toilets and Tubs: Baking soda with apple cider vinegar

96% Alcohol: Kills parasites on enema catheters, etc., and is great for cleaning bathrooms.

Videos

I have had many parents ask me if I think it is a good idea that their child watch "Genius Baby" or other videos that are supposed to help with their development. I do not believe these videos to be appropriate for a developing brain. Studies were done in Cuba that found that excessive TV watching before three years of age can cause the symptoms of autism. When the TV was removed from the children that presented symptoms of autism, the symptoms disappeared. They also reported that TV watching could cause neurological discharges in the brain.

I can also tell you that the only time my son stims or flaps is when he is allowed to watch TV. When he was younger, and deep in the throes of autism, the only thing he wanted to do was watch TV, while running back and forth and squealing. Any other activity was met with screaming and tantrums. He is much more interested in other activities now, and easily transitions from one to the next, including school. However, he still requests to watch YouTube in the evenings when all of his activities are complete, and that is truly the only time he flaps his hands anymore. To me it is very obvious that TV is a source of over stimulation for him. However, it is hard to deny him a truly preferred activity, especially when he enjoys videos on cooking and foreign language! The least amount of screen time (TV, computer, iPad, etc.) the better, however, we do what we can.

Therapy Options:

ABA is not the only therapy for children with autism.

Many parents worry because ABA/IBI is either not available, or not affordable to them. I firmly believe that something is better than nothing. At the same time, I have seen many children recover without any ABA or IBI. Here are some of the options that have worked for families in conjunction with the protocol:

- ABA - Applied Behavior Analysis
- Occupational Therapy
- DIR® Developmental - Individual Difference, Relationship-based Model
- Floortime Therapy
- TEACCH - Treatment and Education of Autistic and Related Communication Handicapped Children
- Son-Rise®
- PROMPT - Prompts for Restructuring Oral Muscular Phonetic Targets
- Sensory Integration Therapy
- RPM - Rapid Prompting Method
- Supportive Typing

There are many other options not mentioned here. Human interaction is key for our children on the spectrum to find joy with us in our world.

I have known many families to have varying degrees of success with all of the above therapies. Some are better for younger children and some for older children. Depending on the areas most lacking in your child's development, you can investigate any one of these types of therapies. Families in many countries where insurance doesn't cover any type of therapy have had to be trained themselves in how to apply therapy. Sometimes an aunt, sister, cousin, or neighbor steps up to help out. No ONE therapy has been fundamental to all of the recovered children and for that reason I say something is better than nothing when it comes to therapy.

Even household chores can be therapeutic and task oriented. Simply because a person has autism doesn't mean they can't have responsibilities and learn new skills while helping out in the home.

A word on laboratory testing

I can't stand to watch people waste money that they could be using for healing their children through various interventions on laboratory testing. If you have all the money in the world, and want to chart your child's progress through laboratory testing, so be it. However, the rest of us sometimes have to choose between expensive testing and treatment. So many families have come to me after spending thousands of dollars they didn't have on testing to tell me their child has viruses, bacteria, and candida. Am I surprised? Never. I have yet to see a child with regressive autism that does not have pathogens/parasites. No matter which pathogens are present, we use CD.

Testing for parasites is woefully lacking at this time, and in fact, the only practitioners who have consistently diagnosed parasites have been veterinarians and practitioners of alternative medicine (also Metametrix labs have identified parasites in some cases).

Generally speaking, if a child doesn't have viruses, bacteria, yeast, parasites, heavy metals and allergies they don't have autism.

Stay Up-To-Date

Subscribe to our Newsletter!

Just go to www.cdautism.org and enter your name and email address...

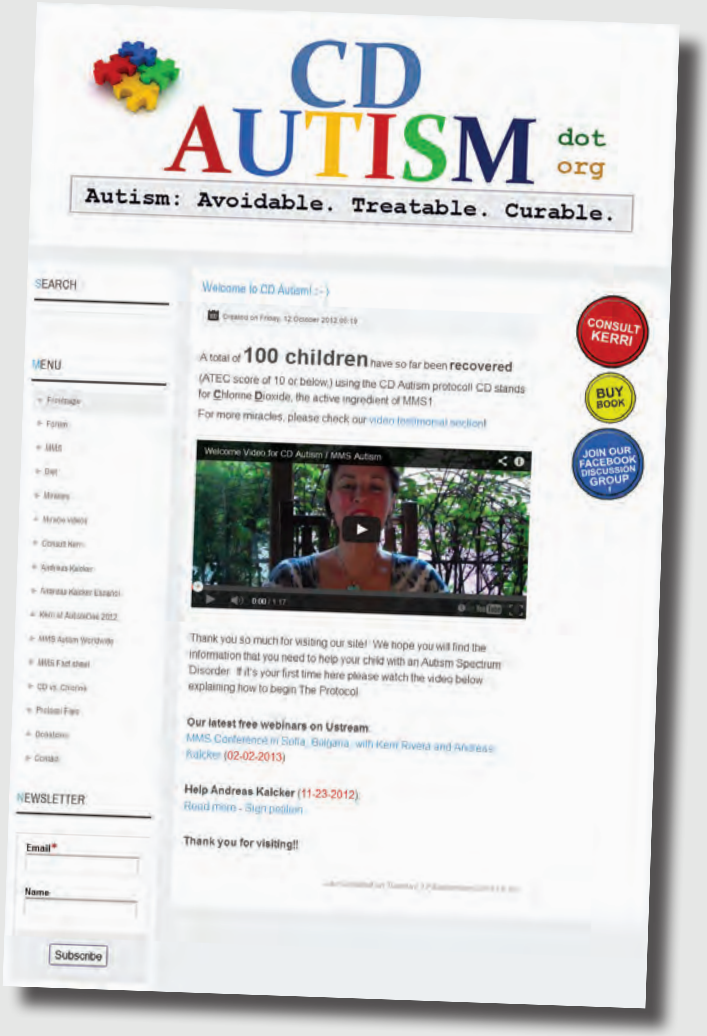

CD AUTISM dot org

Autism: Avoidable. Treatable. Curable.

SEARCH

MENU

- Frontpage
- Forum
- MMS
- Diet
- Miracles
- Miracle videos
- Consult Kerri
- Andreas Kalcker
- Andreas Kalcker Español
- Kerri at AutismOne 2012
- MMS Autism Worldwide
- MMS Fact sheet
- CD vs. Chlorine
- Protocol Files
- Donations
- Contact

NEWSLETTER

Email*

Name

Subscribe

Welcome to CD Autism! :-)

Created on Friday, 12 October 2012 06:19

A total of **100 children** have so far been **recovered** (ATEC score of 10 or below,) using the CD Autism protocol! CD stands for Chlorine Dioxide, the active ingredient of MMS1

For more miracles, please check our video testimonial section!

Welcome Video for CD Autism / MMS Autism

0:00 / 1:17

Thank you so much for visiting our site! We hope you will find the information that you need to help your child with an Autism Spectrum Disorder. If it's your first time here please watch the video below explaining how to begin The Protocol.

Our latest free webinars on Ustream:
MMS Conference in Sofia, Bulgaria, with Kerri Rivera and Andreas Kalcker (02-02-2013)

Help Andreas Kalcker (11-23-2012)
Read more - Sign petition

Thank you for visiting!!

CONSULT KERRI

BUY BOOK

JOIN OUR FACEBOOK DISCUSSION GROUP!

Chapter 15

Final Thoughts

If you knew how powerful your thoughts were
you would never think a negative thought again.
~ Peace Pilgrim

What would be your message to families of children with autism who have just read the book, but not started the protocol?

Do not wait! Get started now! Autism is Avoidable, Treatable and Curable! You don't have to spend fortunes on lab testing, imported foods nor supplements. Even if the medical community can't decide if your child has autism, PDD-NOS, ADHD, Asperger's, PANDAS, etc.... don't wait! Start the diet. Even if you cannot do more than that, it will help. Some children have lost their diagnosis with just diet. There are parents all over the world doing biomedical protocols for autism, get together with them, join the forum at...

www.cdautism.org

...or the CDAutism Facebook group. The Internet helps us all to be close and share information. Don't let anyone tell you the diet doesn't work, just keep going and don't stop. There is never an excuse to not do what is right for your child. "Their future depends on our courage."

The final chapter of this book is yet to be written. Until the day we have recovered every victim of regressive autism, there is no success. Just progress. Here I have expressed what I have seen and what the road of autism recovery has brought to me. I have learned a lot and continue to learn. Our children need us to keep learning and searching until we find what unlocks them from their autism. Not everything works for everyone. However, I have seen that this protocol works for most. Especially when it is done exactly as it should be. If you have found this book you have an advantage that I did not have with my son. When I first learned of Patrick's diagnosis I lost precious time and money that I didn't have to lose, trying interventions that didn't yield results. In the nine years that my son has been in recovery I have yet to come across another protocol that has helped as many children to recover from autism in such a short amount of time—less than three years, to be exact! As of the date of this writing in May of 2013, 93 children have lost the diagnosis of autism and every month more and more are recovering.

Please remember that if your journey is autism you will never walk alone. I am here too. Just one click away...

www.cdautism.org

I wish everyone who is ill to have the fastest recovery possible from whatever it is that ails them.

To be continued…

Dr. Andreas Kalcker, Kerri Rivera, Jim Humble, in Puerto Vallarta, Mexico 2012.

Chapter 16

Healing Beyond Autism...

When TRUTH is found, understood and accepted is when we realize it
was not even close to what we thought it would be...
~ Miriam Delicado

As Jim Humble has been proving for many years, chlorine dioxide is a molecule capable of healing many common ailments. When combined with the other elements of this protocol, in the correct order, it has shown to be even more effective. We have decided to include this bonus chapter because it is time to start sharing this information with everyone, as pathogens and parasites are the root cause of many of the diseases that are causing suffering in our world today.

As parents saw improvements in their children's health from the protocol, some of them decided to use it on themselves as well. Similarly to when the flight attendant tells you to put on your own oxygen mask first before helping others, it is important to take care of yourself if you want to be able to help your children heal, grow, learn and mature.

Over time, parents who followed in their children's footsteps started to have some undeniable gains in their own health. This was an added bonus many didn't expect, and we couldn't resist sharing their stories with you.

After reading the parasite chapter, you may have been surprised to see how many common symptoms can be caused by parasitic infections. Imagine what would happen if we could get this information into the hands of people who have seen many doctors, used various drugs, and applied countless protocols for years while only masking the symptoms of the underlying cause. So, there should be no surprise if going after the real cause was life changing in real ways.

The following testimonials are a small sample of how some of these people's lives have been changed. Serious afflictions such as Lyme disease, fibromyalgia, chronic fatigue syndrome (CFS), depression, cancer, and chronic infections, are just some of the conditions that improved or completely cleared up. By cleaning up toxic pathogens and parasites, as well as removing foods that feed these invaders, it should be no surprise that disease triggering inflammation is reduced so true healing can take place.

I am so excited for you to read these, and I know that some of you will recognize yourselves here. Please know that healing is possible! We have a public Facebook group called *CD Health* where adults using the protocol for themselves can share questions, doubts and success stories. I highly recommend that you check it out if you are looking for support or some guidance getting started.

Testimonials Beyond Autism

After we saw our son recover from ASD using CD, I tried CD myself and it helped me recover from Lyme and its co-infections. Some of our extended family heard about our success and they have had very good results using CD for treating Hypertension, Diabetes and Liver problems. My younger NT son used to bang his head in sleep since he was born, CD has now resolved the problem. I think anyone who tries CD sincerely, can only benefit in treating so many hard problems. One very beneficial side effect I have seen with CD is the resolution to seasonal allergies and improvement in the immune system to the point where kids do not fall sick or catch colds. In this regard CD is much better than yearly flu shots which have so many toxic elements in them and still do not protect adequately.

—Nilesh

To say I was anxious about the impending birth of Baby #2 is an understatement. I felt like I had failed our first born on so many levels and was a mess believing that I'd do the same for our new daughter.

But alas, she was born a perfect and healthy baby girl. So was Callum, our son, but unlike Lila who had zero interventions, Callum had antibiotics from my labor, the Vitamin K shot, HiB, and DTaP. And it was the DTaP vaccine at 15 months that took the shine from our sweet boy's eyes.

We discovered Kerri's protocol when Lila was 3 months old and more worries cropped up about how to best protect her from the die off that was going on in our house.

At 5-months, she developed an eczema patch underneath her chin, became chronically constipated, and her normally cheerful disposition became irritable.

As soon as she hit the 20-lb mark, I started giving her 20 drop CD baths, low dose oral CD, and 3 drop CD/150 mL water enemas. To my horror and surprise, Lila started passing 1" baby roundworms almost daily.

In a family with both NT and ASD children, the NT child always has some type of quirky behavior. Now we know why! Parasites do not only affect the ASD

population, they affect everyone. Having CD in my tool chest gives me the power to prevent so many things that may have been in my daughter's future.
Don't be afraid to treat your babies! Her eczema is gone and she's pooping regularly once more.

I write and share my story in that it might help or save someone else's life. That it may just be the key to help someone see what may be what is happening for them as it was and is happening to me. If it helps just one person onto their path to recovery and health then it will have been worth sitting down at this computer and sharing it.

For the past 10 years or so my world has been collapsing in on me, and getting smaller and smaller, and accelerating the past 5 years. I have felt as if I were under a thick coffee mug on the edge of a table about to fall off. On a scale of 0 to 20 with 0 being death and 20 being fully alive and cognizant, I felt as thought I was at 3 staring at 0. For the past 5 years at least I have done so much research, listened to so many radio interviews, done so many different modalities, bought so many products, paid a lot of money to different therapists, ate weeds, made my own concoctions, etc. My mother would always say to me wow you should be so healthy to which I would reply no mum, actually I am not. I began feeling like a phony. I ate real good took all the right stuff, got plenty of rest foraged for my own food and spring water, got exercise, walked barefoot, etc, but always felt exhausted and looked worse and worse. I was fighting with everyone it seemed and chasing people away including family.

Four months ago my wife decided we should go away on a weekend getaway and relax. We fought the whole drive up there and when we got there she threw me my room key and said you're on your own. I spent the whole weekend in bed and on the trip home we both decided to end our 20 year relationship and marriage. I was good with that and looking forward to being on my own. I was finally chasing the last person who mattered to me away.

A week later an angel came into my life. I have been listening to oneradionetwork.com for the last 5 years and on this day Patrick Timpone had a lady by the name of Kerri Rivera on and she spoke about how she was curing kids with autism and that the kids had parasites. I looked up the symptoms of autism and realized I had quite a few of them. I immediately began the mms enema protocol which she recommends. I could not believe what then came out of me. I kept this up for a week and still there was no let up with the amount of parasites, biofilm, and who knows what coming out of me. After a week of this and seeing no let up on what was coming out of me, and having just done an mms enema I laid on my bedroom floor and the grief just poured out of me. My wife came in and lay down next to me and just held me. I knew then that something had changed, that my life and my marriage were now on a healing path. That the pain, the toxins and parasites were now leaving my body.

It has now been just over 3 months and there has been no let up on what has been coming out of me. If I were to have scooped it all up and put it in a jar I am sure it would fill a one gallon jar with hardly any water in it. I now feel I am at about a 9 with 20 in my sights. I thank God that this info came into my life and I took action on it. It has saved my life and my marriage. I love my wife, and cannot imagine not having her with me. And I love my life and am looking forward to not only getting *to the 20 mark but what is beyond the 20. What is possible when one's brain is at full capacity and not being inaundadated with toxins? Much love and appreciation goes out to Patrick Timpone at oneradionetwork.com, Kerri Rivera and her work with autistic kids (adults too Kerri), and to Andreas Kalcker. Thank you all for showing me and pointing to the keys I needed to free myself from what has been a nightmare.*

A big part of my healing has been reviewing my life and realizing that this has been going on for most of it and answers so many questions I have had. As a child growing up I had pain in my stomach and what the doctors eventually did was take out my appendix, which most of us know now is needed to keep the healthy flora in our intestines. And they told my parents that it was all in my head. So whenever I complained about stomach aches and pains was told that it was all in my head. School was a nightmare for me, as socially I felt I didn't fit in or belong, and a lot of easy concepts that most others got, I just couldn't get. I always had the feeling like I was missing something. I felt as though I was stupid or something just wasn't right with me. I now know there are many just like me. The words Autism, add, adhd, ocd, etc were not a reality back then. And I know those are just words or diagnosis of toxins and parasites. Get rid of the toxins and parasites and the diagnosis goes away. Easy peasy. Thank you all for listening to this chapter of my story. I look forward to many more chapters in my life story. Please feel free to pass this along to anyone who it may help. Much love.

—Bruce

When starting the protocol for my son's Autism, I decided to do it along with him.
I was suffering from Hypothyroidism, Adrenal Fatigue, a Hormonal Imbalance, and Severe Anemia. My cortisol was at a 3. One week into the protocol, my cortisol was at a 14, and my iron levels were normal without supplements. My thyroid medication dosage was decreased twice. My vitamin levels were all normal. My uterine fibroid cysts were completely gone after a few months on the protocol. I continue to see improvement in my health. This protocol has been a Godsend for my family.

—Maggie Kaye

CD has been life changing for my 18 year old son who has autism and Lyme, although his ATEC today is only a 3, down from a 68 at the start of cd, but I never thought it would be so instrumental in my dad's life also.

My dad, William, has a malignant melanoma in his eye and has received radiation for treatment for the last few years. I talk about cd all the time but they never took this treatment seriously, at least for anything besides autism until now. My dad gets terrible ear infections that he sometimes has to go to the hospital for. They tell him each time that he has "swimmers ear" and they give him abx, sometimes it seems to work, sometimes not, and he goes back to the drs or hospital for more abx....over and over. Recently he was leaving here for his 7-hour trip home, and he was very worried that he would not make it home with the painful infection that had begun, *and he would have to stop on the way and visit a strange hospital. So I mixed up some cd, and told him to drop a few drops of the cd mixture in his ear periodically on his way home. To his surprise, the infection cleared up by the time he got home and there was no need to even call a doctor.*

Of course, my secret plan is to get him on the cd protocol to finally deal with the tumor in his eye and that's the plan we have started. Whether he uses it and shrinks the tumor and saves his life with cd or not, he's a true believer now that cd cures infection.

I have been taking CD and the parasite protocol for 5 mo., and here are my observations. I have had long-term chronic sinus issues which has caused sever sleep problems. Also, a few months ago, I had a severe toothache in the upper left area. The dentist examined and found no fault with the tooth, even though there was severe swelling in the area. The infection originated from my sinus, and by taking my CD regularly hourly, was quickly solved. (My hourly CD doses were compromised by travel) I have also noticed improvement in my sinus condition, resulting in better sleep, although I have not fully recovered yet. Right now my sleep is somewhat interrupted by peetox. I will continue CD and PP for the 18 mo. Previously; I had used several naturopaths, with either negative results or limited results that disappeared when the medications were discontinued. So far I am happy with CD and would recommend it to people to try.

—Ed Canada.

A friend started eardrops with her 13-year-old pug 3 weeks ago. The pug Lucy was almost deaf from ear infection after ear infection, on constant drops, had seizures. After 3 weeks on CD eardrops, went to the vet, no ear infection, hearing is improving!! Winning!!

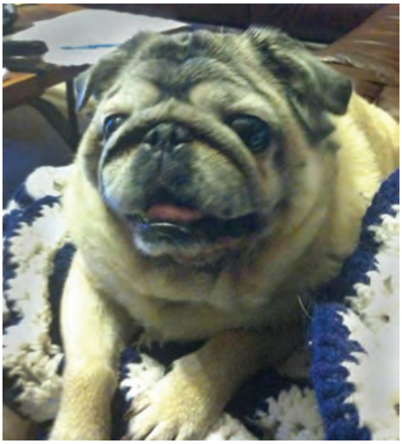

Since I was a child, I had been anemic and reading the book "Healing the symptoms known as autism" on the Parasite chapter, I saw that anemia is a sign of parasite infection. Since my son is taking CD I also took CD to know how my son feels and maybe cure my anemia. I had no problem going up to 24 drops in less than 2 weeks. I stopped taking my iron pill after 1 month of taking CD. I also did parasite protocol as I do not want to reinfection my son. Since, I stopped taking my iron pill I never had any dizzy spells and I feel great.

I used to wake up at night too to use the bathroom, now with CD no more night waking. In 2009, I was informed by my Doctor that I need to go and see a specialist as there was some suspicious stuff in my Pap Test and if not removed it *can lead to cervical cancer. I was booked in the hospital for the procedure. After taking CD reaching full dose, I had been discharging this horrible smell, like rotten egg, whenever I urinate. It took about a month before it stopped. Whatever that was, CD cleansed me and I think that discharge is related to my diagnosis on the pap test result. I am very grateful for CD for giving me my health back. I am still continuing taking CD as I want to be healthy and I will also try to complete the Parasite protocol for 12 months, I just finished my 5th PP and have been seeing lots of parasites mainly pinworm and ascaris. Now, I know the root cause of my anemia especially. I would encourage people to try the Protocol and see where the healing leads you. Thanks to Kerri Rivera, Jim Humble, Andreas Kalcker and people who used the CD before me as you gave me courage to use the CD.*

—G.M, Canada.

Two days before our first set of cd bottles arrived (we bought it to start Kerri's protocol with our autistic daughter), our son (6) got sick with scarlet fever. We didn´t give him antibiotics because we just finished our research on cd and were excited to use it instead. My sons face was swollen and red, he had a moderate fever and rash all over his body. After the first dose of 1 activated drop of cd within one hour his swollen red face subsided completely and in the evening the fever was gone. His rash got dry and after 3-4 days it was not visible anymore. Our physician told us that scarlet fever without antibiotics lasts up to three weeks. We were done within 4 days, without damaging his gut flora, as it would have using antibiotics. CD has taken care of many infections in our family since then and most important - brought our daughters ATEC down from 100 to 46 in nine months. I´m beyond grateful for its discovery.

For the past six years, my life has been plagued with anxiety/fear, depression, pressure in my chest, making it difficult to lay flat and sleep, bloating which caused me to look 7 months pregnant and which made my abdomen very tight and uncomfortable. Headaches, weight gain, sore joints, no energy, fibromyalgia pain, and dizziness causing me to feel drugged and off balance.

The anxiety and fear were debilitating literally making it impossible for me to sleep. I would go night after night not sleeping, just dead on my feet suffering with so much anxiety and very frightened wondering what in the world was wrong with me.

When Kerri's book, Healing The Symptoms Known As Autism came into our lives, I read it wanting to help or autistic son. I soon realized I had so many of the symptoms that I needed to do the protocol.

I started on the cd taking a very small amount, 2 drops every hour, about 12 doses per day, and worked up to between 60 and 90 drops a day, I also took cd baths and did enemas. I increased the cd every day by just a drop or two. Very soon I started to calm down significantly. I started sleeping! I started losing weight! For about 5 weeks I thought I had died and gone to heaven. In two months I lost 20 lbs.

But when my anxiety started coming back I set up a consultation with Kerri. I learned I needed to begin the parasite protocol to kill the parasites. The cd just killed the little ones.

With Kerri's help in several consultations I was able to tweak the protocol in her book and apply it to me for my personal needs. I started on low doses and increased it for what I needed to sleep.

As of today I have only been taking Mebendazole, the medication that kills parasites, for 3 and a half weeks. I sleep at night. Some nights are a little choppy but I get 4-5 hours of sleep. I am not depressed. My emotions are not all over the place. I am coping. I'm functioning without sever headaches, only mild ones. I still have bloating but not as severe. I feel like I will live and that I am not dying anymore. I hardly ever feel depressed.

To date I take 200mg Mebendazole at breakfast, lunch, dinner and bedtime, plus a 200mg implant right before I get into bed. The implant has been the turning point in my healing. (Consult with Kerri for directions regarding the implant.) I also take 2 neem tea capsules with each meal. Also, one dose of Combantrin everyday at lunch. I take castor oil as needed each day 1-2 Tablespoons to keep things moving, plus a cd bath at bedtime, 150 drops soaking for 20 minutes.

Trying to differentiate what is die-off and what are actual symptoms from the parasites is challenging. Pay close attention to how you feel after you do or take something. After an hour note how you feel. Worse, better? You have to listen to your body. Consult with Kerri often and keep her informed by email. You may have to tweak the protocol every few days so write everything down, day, time, how much, how you feel etc.

All of my symptoms are not gone 100% but they have greatly improved. The anxiety, which was the worst and most debilitating symptom, is the greatest improvement I have had. I know I will be able to heal completely and get my life back. I have just begun!! I have had so much progress in just a few short weeks that I still cannot believe it!! And...I have lost 2 more pounds!!

Thank you Kerri for all of your love, support and guidance!! I could not have attempted this without your knowledge and help!! You are an angel!!!! I AM HEALING!!!!!!

Your devoted friend, Marcie

My mother 84 yrs of age- who has dementia was very, very serious last week. Her lungs were filled with mucous and she could not spit it out. We thought this was the end and my sister invited the priest to bless her. The doctors suggested that I place her in a hospice, but I refused! They also offered antibiotics, but I bought some CDS and gave it to her every hour. I placed a humidifier with 40 drops of activated CD by her bedside day and night. And massaged different parts of her body with CD and DMSO. soaked her feet just once in CD as it is difficult. She cant sit straight. To everyone's surprise she made it through. Just a little bit more to clear up! Thank you everyone for sharing wonderful tips on loving and caring for our loved ones! This is a wonderful extended family! love the CD!!!!!!!!!!

* * *

I want to thank all of those who were kind enough to share their personal stories with us. There are many more, but you get the idea. I am so excited for you to read them, and I know that some of you will recognize yourselves here. It just goes to show that healing challenging illnesses is possible. We have a public Facebook group called *CD Health* where adults using the protocol for themselves can share questions, doubts and success stories. I highly recommend that you check it out if you are looking for support or some guidance getting started.

Appendix 1

More Miracles & Testimonials...

The only limits you have are the limits you believe.
~ Wayne Dyer

To date, 115 children have lost their diagnosis of autism by using CD and some or all of this Protocol. Autism IS curable!

This section is comprised of testimonials, miracles, and short blurbs on gains made by children using The Protocol. In many cases, these are candid emails from parents to Kerri sharing their joy at their child healing. These were gathered over the past 3 years, and have come from families all over the world. Some were translated by the parents of the child to English, and some were translated by my sister or I. We have done our best to honor these exactly how they came to us, unless a correction was needed for clarification, these were not edited. Many are anonymous for the protection of the families and their children. Names were used when permission was given. These families were generous enough to share their experiences with all of us to spread hope and joy, as well as to prove that children all over the world of all ages are healing from the symptoms know as autism. Please enjoy and share.

Please note that many of these testimonials were collected some time ago and therefore some contributors use the term MMS instead of CD. Due to time constraints it was impossible to request permission to change the acronym of MMS to CD in their testimonials. As we mentioned previously, MMS and CD are the same substance. Where the acronym MMS is used in the following testimonials please know that chlorine dioxide is the substance responsible for the healing.

We have added many new testimonials since the first edition came out. Those new testimonials are located at the end of chapter 2 and continue on the next page, followed by those appearing in the first edition.

(continued from Chapter 2, page 40)

45. *Things I wish I'd known before starting this protocol:*
    - * *My son would drop juice and pick up a water regimen immediately.*
    - * *My son would sleep through the night every night.*
    - * *My son would not be constipated again.*
    - * *My sons ATEC score would be halved in 7 weeks.*
    - * *My son's speech would be at age level.*
    - * *My son wouldn't need his daily rx med again.*
    - * *My son would be excited about and participate in his birthday celebration.*
    - * *My son would come into the bathroom and say, "I'm ready to wash my butt" unprompted.*
    - * *My son would drink ocean water.*
    - * *My son would get acclimated to school and would initiate conversation with his peers.*
    - * *My son's dark circles under his eyes would disappear. He would start to look and act like a healthy boy, and others would take notice.*
    - * *My son would be affectionate, tell me he loves me, crack jokes, and be connected to me like never before.*
    - * *The people who opposed our diet and protocol choices would be unable to deny that this works and we're on the right track.*
    - * *My marriage would be saved by rolling our sleeves up and tackling this together.*
    - * *I would find a way to heal my own ailments.*

    The future is bright

---

46. *My son with autism is 31 years old. Although he is fully verbal, he had a lot of issues: SIB, Aggression, he was not able to sleep. We have been trying different kinds of protocols since he was 3 years old. Unsuccessfully. I heard about Kerri Rivera and the CD Protocol last September. We started right away. We were continuously in touch with Kerri who kept giving us the faith, that even at that old age it is possible to heal and even recover. In September, when we stared CD he had an ATEC of 87, 4 months later we measured 63, and now, after 8 mounts, he has an ATEC of 28. There are some SIB's and aggression left, but it happens only rarely. If we could get rid of it, my son would be recovered. He became an interested young man, his memory improved. He keeps giving proper answers to my questions, something he never did before. I can see his real personality now: he became a nice, lovable, open-minded young man from a very sick raging, roaring, screaming, suffering boy. There are no words to say thank you to Kerri to show us the end of the tunnel. I think I'm on the best way to recover my 31-year-old son!*

    Mark, 31, Hungary

---

47. *Dear Kerri, I am very grateful to you for your benevolent service of helping to recover kids with Autism like myself. You have really improved the quality of lives of many hopeless kids. I really feel the difference now that my ATEC score is a*

I and I am still making inroads in executive functioning and social skills. Without someone like you, we would have not known about CD and it would have taken several years to cure me solely with chelation. By that time, I would have been out of high school and probably in a community college taking rudimentary general education courses and would have had a bleak future ahead of me. But now, I am doing well in school and am taking college level courses. And I am not the only one in my family that has had benefitted from using CD. My brother used to have mild seizures in the form of head banging and most of that has now stopped. My father has used it because he contracted Lyme's disease twice. And my grandparents have used it to treat Alzheimer's, Diabetes, Blood pressure and Hepatitis C. So again, I cannot thank you enough for everything you have done that has not only changed my life forever but has made a profound difference in the lives of others afflicted with Autism.

---

48. My seventeen-year-old son was a healthy happy baby until he received vaccinations at age three. The day after the vaccinations he began having horrible anxiety problems. Day by day he lost skills and regressed. He developed problems with fine and gross motor skills, and social skills, memory and sensory integration. He also had multiple learning disorders and struggled in school. He spent all of his time in his room by himself and he did not make eye contact.

We tried many interventions to try and help him. Biomedical interventions, antibiotics, therapies homeopathy and Biofeedback all had limited benefits for my son. The only thing that helped him some was a gluten free, casein free, sugar free diet. While reading Healing the Symptoms Known as Autism we learned that most of the interventions we tried did not work because we were feeding the parasites that were causing my son's autism with vitamins and supplements.

While I was watching videos about Chlorine Dioxide I came across a video of Kerri Rivera talking about the protocol she uses with autistic children. As I watched the video everything began to make sense and I knew the protocol would help my son.

We bought Kerri's book Healing the Symptoms Known as Autism, gathered the supplies and my whole family started the protocol. To our amazement we saw improvements in my son overnight. We had taken the ATEC test for my son right before we started the protocol and it was a 43. In just one month of Kerri's protocol my son's ATEC dropped to a 5! We will continue with the protocol for the recommended time and we expect to see my son's ATEC drop to zero before we finish.

His social skills, motor skills, thinking, behavior, attention span and stimming have all improved immensely. His sleep patterns, digestion, physical appearance and weight have also improved tremendously.

Kerri Rivera's book gave us the tools we needed to reach in and pull my son away from the grip that autism had on him. We will forever be grateful to Kerri. She devotes her life to helping children with autism and always has an open line of communication for anyone who needs help. She is the voice for so many children that don't have a voice. Thank you from the bottom of my heart!

---

49. FUA!!!! BIG moment... crazy morning, ughhh we miss the Bus I was so mad because my other daughter had preschool at 9:30 so my morning was a mess, I had to leave in Pj's and make them eat breakfast in the car....but then in our way to school my son saw a School Bus passing by our side and said "school Bus is gone, L is sad" OMG OMG OMG......did he just said this 30 words sentence?????

(actually 7 words but seems like 30 to me!!!) Ok...WAIT then I get to school not mad anymore, I walked with him still in shock, I couldn't believe what I just heard....and when he saw the line of kids by the wall he point at them and said "my friends" OMG OMG OMG....ok WAIT....then we get there and he takes the teacher hands, gives me a kiss and said "bye mommy" OMG OMG OMG .... WAIT then the teacher ask him L, remember back pack by the wall, so he goes to the wall with his back pack in his back and does exactly what he was ask to do!!!! O M G

FUA...THAT'S RIGHT.... 223 days in this Protocol, searching poop, dosing, enemas, writing in his food diary every little thing he eats, therapies ABA, speech, occupational, hours hours of work, hard work...all just to witness this miracle of healing...to see him think, react, analyze a situation, be happy, spontaneous and smart by himself (no prompt or reinforcer necessary)....THANK YOU Kerri Rivera since the first day I found you in YouTube I have never stopped believing....I LOVE YOU WITH ALL MY HEART!!!! YES COMADREEEEE!!!!! I can feel it...we are close...it's happening.... so happy I missed the Bus this morning... this is the best day in my son's way to recovery so far!!!!!

HOPE FOR ALL!!!!

---

50. My son is 23 and we have tried many, many approaches to heal him over the years. Typically we would see some initial progress, but then the rate would slow down and he would plateau again, slipping further and further behind. Lyme disease and co-infections, as well as various viral, fungal and metabolic issues complicated everything. We had to resort to psych meds, even, for his depression, anxiety and mood disregulation.

We started Kerri's protocol in the summer of 2012, and had to cut the number of drops again and again to something he could tolerate, since he has MTHFR genetic issues with detoxing, until he was at 1/64th of a drop per dose. By October 2012, he was at full dose and after two weeks, began to talk about learning to drive, to offer to do a chore at home, to draw again. So it has been about a year for us, and the improvements are continuing, his ATEC score went from 77 pre-meds to 11 now, and our lives are all much better thanks to this book and Kerri's dedication to the cause of healing the children. Please, if you are reading this, don't be afraid to try one more thing!

---

*51. My son started Kerri's protocol 10 months ago. At the time we started the protocol he was doing a home based therapy program for 60 hours a week. We did not send him to school because he was not able to handle all of the outside stimulus, unable to follow directions, and unable to communicate effectively. He stimmed most of his day and only spoke in 2-3 word sentences to get needs met only. When he started the protocol his ATEC was 57. It quickly dropped and the improvements were mind blowing. In 5 months time his ATEC was a 9. Today at 10 months after starting the protocol his ATEC is a FIVE!!!! He now attends school with out any kind of extra support and exceptions made for him. He is truly just one of the other children. I have been volunteering in his school and I have had 5 different office people tell me that he stops and greets them on his way to different specials (p.e, library, art, gardening). They told me he initiates the greetings and says: hello, good morning, buenas noches, and all with a big smile on his face. I have even been told by 2 of them that he greets them by name. Did I mention that he leaves the classroom on his own to go to specials??? Yep, my one time *severely autistic* child finds his name and places it in that specials category* *basket, leaves the classroom and goes to specials on his own???? When he is done with specials he goes back to his classroom and does his work. Holy CRaP!!!! A dream come true.*

Thank You Kerri!!

---

52. *My 6 year old son's ATEC went from 45 to 22!!! Wow I knew he was doing better but more then half!!*

    Thanks Kerri Rivera!!

---

53. *I don't know what to chalk it up to--we just finished our 3rd PP and did our 2nd shot of GcMAF on Saturday--but my 9 y/o son has started taking pictures with my iPhone. They are not just random shots; he is finding objects or angles and taking pictures. They are FASCINATING! He has also figured out how to use the reverse feature, and has taken a series of self-portraits, which he flips through and says to himself "Who's that?"*

---

54. *My 16-year-old son is doing great on your CD protocol. ATEC down from 89 to 37. Best thing that ever happened after doing 8 years biomed with him.*

---

55. *We started CD with my precious 5 year old 'O-ster' in June 2013. Some of the completely new developments we have seen since we started....... He is aware of how to burp and toot, and thinks it is funny. He is apologizing for things he does wrong. Even a day later, he will remember and apologize. He goes to the bathroom on his own. He is completely pottytrained, even at night. He does household chores happily. He sent people on my contact list emails that said 'I love my mommy.' He has better problem-solving, more motivation to draw pictures. If he gets messy, he just goes to the bathroom and washes his hands and wipes off his shirt by himself without saying anything. He dresses himself. He does somersaults (just a few months ago, it was impossible for him to even get in somersault position), can open baby gates, can get in crab walk position, can jump off of furniture with both feet. He greets us when we come home and is excited to see us. He MISSES us when we are gone and talks about it during the day. He's making up 4 line rhymes. He cares what he wears and picks out his clothes! He is not running back and forth as much. We have normal trips to restaurants. I can leave our front door open now, and not worry that he will run away. He tries to help put our dog in his crate and fetches with him. He wants to come with me if I leave the house to go somewhere. He will now eat almost anything I put on his plate. He used to vomit at the mere thought of things like GF pasta or rice, and he wouldn't allow any food on his plate to touch another type of food on his plate. He will now eat bowls full of tomato sauce, onions, burger, rice, garlic and peppers all mixed together. Our family can eat dinner together now. I no longer have to make separate meals. He even eats raw peppers! He yells 'watch me!' and does things like diving onto the bed or jumping off of something. He looks at cloud formations to see what pictures he sees in them. Last week he heard a toddler crying in Target. He wondered why the child was crying and wanted to go find her. He misses his little sister when she is not at home. He wants her to play with us when we are playing. He is completely aware of what is going on now. He is much more aware of danger. He notices the weather. There is a sense of calmness about him that he did not have before. His tantrums are much less severe and much less frequent. I was supposed to be his aide this year in preschool, but I have been able to just drop him off at school. There are many more new things that he is saying and thinking and doing.*

Our quick biomed history- He regressed after vaccines at 12 months. Starting at 18 months of age, he was gfcfsf for 1 year, then scd for almost 3 years. It feels like we have tried everything with supplements over the last 4 years. Some things helped a little, but nothing really tackled his pathogen load, even though he was on an awesome diet. His ATEC was 66 in Dec 2010. Low dose chelation brought it down to 33 in 3 months. The ATEC stayed at 33 for 2.5 years. We did 112 rounds/weekends of low dose chelation during that time. ATEC June 2013 was 34 (before starting CD), July 2013 was 30, October 2013 was 20! We are on our way down. This is the first time that I have physical proof that we are getting rid of the pathogens that have controlled and poisoned his body and brain for most of his life. It feels so good to finally help him. I am hoping to write our complete recovery story in Kerri's next book. I thank God that we saw Kerri at AutismOne.

56. *Nick texted me yesterday for the first time, on his own. "Hi from Nick" - that's a 2013 FUA fo sho. – Alison*

    (this text was accompanied by an ATEC drop from 69 at 6 years old to 14 at 8.5 years)

57. *Yesterday my son told me I worried too much, I said you worry too much Too Noah! He responded, my hair is red, yours is gray, so who is the worrier, hahaaa! The windshield wipers keep wiping away the autism, what's staying around it seems, is completely lucid speech and unbelievable comebacks! He never could say things like that before, not at all, he really was in there somewhere.*

58. *Something really beautiful is going on in our household right now...Caleb is totally blossoming socially on every level. There is a calmness in his body. He is alert, responding to his name, sharing his toys with other kids, wanting to give affection without prompting, soothing Lilia when she is crying rather than batting at her. We even have gotten a few spontaneous silly phrases!!!! Oh baby, oh baby, oh baby!!! We are on our way!!!!!*

59. *We started the diet on the 1th of March, CD on the following week. Adam started to ask questions in April. Since then he keeps asking continuously…His ATEC dropped down from 49 to 38.*

    Tunde, Hungary

60. *OMG OMG OMG!!!!! My son is singing!!! He was ZERO verbal... I mean ZERO!! His shadow teacher told me a week ago that he started to mimic, I said to myself "yeah he was babbling and she thought he is mimicking!..." Yesterday she told me he started to sing and I said to myself, "Yeah right, she is absolutely illusionist or exaggerating!!!" She knew I will not believe her because he is just humming at home! So she sent me a video which was a shock for us. He is singing the song "A is for apple, a a apple... B is for ball, b b ball... C is for cat, c c cat... D is for dog, d d dog..." I don't have enough words to thank you and thank Kerri Rivera and this incredible protocol!!! Thank you for saving my child. xoxo*

61. *...My son is doing his first CD PP. We picked him up from preschool he was extremely happy. Hugs & kisses. Then...runs up to a classmate and touches his shoulder and says, "bye Johnny", runs to another classmate and says, "bye Kevin." Then he skips to the car. These improvements in his socialization is new. We are so pleased and I wanted to pass along these social gains with you all. His atec at start is 58.*

62. *My little sweetpea is almost 4...and completely NON VERBAL...upon picking her up from therapy today, her teacher said she was having a great last few days... and both therapists said her eye-contact has been different...like in a really good way!!! I've also noticed her looking at me...straight into my eyes for what seems like forever...almost as if she has never seen me before?! They also said that she has been looking at them and has been opening her mouth trying to make sounds...as if she wants to say something very badly...but can't...but with a lot of prompting...she did say "Hi!"...all I know is that somewhere in there is this amazing person waiting to be let out! I often worry if we are ever gonna make it...finances and healing your family can be very overwhelming...especially on one income. But for the love of my daughter we just cannot give up...feeling very hopeful and blessed today...and very grateful that something is finally working.*

---

63. *Hi Kerri,*

I wanted to write and thank you for discovering and sharing your protocol. This protocol has saved my almost 5-year-old son.

My son suddenly started having absence seizures that would freeze his breath. From the day he had his first noticeable seizure, he was hospitalized every seven weeks for uncontrollable status epilepticas, a life threatening condition. They would pump him with seizure meds, Ativan, and Valium for hours before the seizures would finally stop.

Since starting CD five months ago my son has not been hospitalized once. He went ten weeks seizure free. He had a few mild seizures that he brought himself out of without any extra medications. He currently broke this record and has made it another thirteen and a half weeks seizure free. I true miracle.

Everything we searched for to treat his Lyme disease in the past would either cause seizures or not help to prevent them and I now know why. As you predicted after starting your protocols we discovered that he not only has Lyme, he is also incredibly loaded with parasites. So much so, that it scares me to think of what may have happened to him if we didn't find you.

I'm so amazed to watch him subtly emerge. Since we started, he went from saying about 20 words to saying 69 words, along with two and three word combinations, and over the past month he has begun to answer simple questions. His vestibular system has gotten much better; he'll actually swing. He just started jumping off curbs. His eye contact is perfect. He's gained two pounds in the past three months, something he hasn't' done since being diagnosed with failure to thrive at a year old. He is playing with his three year old sister; it's so adorable to watch them chase each other around the house, hug, chase each other, hug, get in the dog crate together, hug, take turns pushing each other in their wagon, and kick the ball back and forth. Last night the three of us played catch together for over 10 minutes. I'm so happy for her to have her brother to play with.

This protocol changed his life, my life, and our family's life.

There are no words to thank you enough for all your help along the way – he's on his way…

Love, Love, Love

Update: as of December 3rd, 2013 this little guy is 18 weeks seizure free!

---

64. *My name is Dawn. Below is a shortened version of what we've been through prior to and since Kerri Rivera coming into our lives and changing our family forever.*

I am Noah and Eli's mom, caregiver, advocate etc., that's been my job for 17 years. Prior to having children with special needs I had a career that I loved and my plan was to continue my career, how deluded I was, my life has been forever changed by my children and their symptoms as my entire existence over these years has been to save my son.

My son Noah born August 19, 1996 has struggled and suffered with what is now known as the symptoms of autism. His symptoms began at 8 months of age with a rare blood disorder known as neutropenia, he had a low white blood count. As he continues to improve, I too, along with our entire family seem to be healing as we bore witness to his daily struggles of self injurious behaviors, aggressions and rages, learning disabilities, speech and auditory delays, 2 steps forward and 4 steps back.

Healing the symptoms know as autism with the Chlorine Dioxide and Kalcker Protocol have been a cathartic exercise for me as well, no amount of therapy would have soothed my soul like witnessing my son's return from the daily brink of death.

My hopes and dreams for my son dashed with vaccines before he was old enough to speak. I'm 50 this year, however I'm older then those years, and my views and outlook on life have changed forever. Imagine fighting for my son's life against those trusted factions of our society supposedly put in place to protect us, i.e., government agencies, churches, schools, teachers, principles, school boards, doctors, hospitals. I'm changed forever by the daily battles, but somehow they are small now in comparison to the beast of the symptoms that ravaged my boy with red hair, often referred to by me as my favorite Martian. As he continues on his healing path, so do those of us who have loved him so dearly, despite his outwardly awkwardness, he is the innocent victim left to suffer for what is often referred to as the "greater good", he's responsible for so much more in my life then just the symptoms known as autism.

It was a Saturday morning in May of 2003. Noah, at the time 6 yrs. old, had slept with us because of seizures on and off through the night. Like so many times before, I was very concerned he would go 'status epileptus', a state of continuous seizures. He ate cantaloupe for breakfast that day. I remember it vividly even 11 years later. His seizures would come about every 20-25 days. First he would complain of a stomach ache, and I could hear the gurgle of his stomach. They usually lasted a few minutes and then went away. This day, the Saturday before Mother's Day, he didn't come out of the seizure. We administered his anti-seizure medication but it didn't work. We called 911, the fourth such call in Noah's brief life. When EMS arrived, Eli, then 3 years old, ran to his bedroom upstairs. I remember the technicians working on stabilizing Noah. They ripped the shirt from his body as they worked on him and carried him out to the ambulance. As I followed them to the ambulance I happened to look back to see Eli watching from his upstairs window as they placed his brother in the ambulance. He stood motionless. Darryn scooped him up and followed us as we drove swiftly through traffic to Texas Children's emergency room, about 25 miles away. I remembered thinking, why don't people heed the right of way? I looked back as they worked on Noah. His seizures continued until we arrived at the hospital. Once we arrived he was taken straight to the ER, ICU room, where *four doctors began to work on him. Darryn and Eli arrived as I stood next to the bed while the doctors worked to save his life. I took Eli out of the room and we sat in a chair looking in at what was happening. His eyes were as big as I have ever seen them and he sat perfectly silent in my lap, we colored. I overheard the doctors explain to my husband that they were having problems stopping the seizure. They wanted permission to use a drug that could stop his heart. Darryn asked if there were any other options. The doctor shook his head. He asked what they would do if the heart stopped. The paddles. So he was given the drug. No heart stoppage. His breathing became shallow and the seizures finally stopped after 6-7 straight hours. A couple of hours later Noah woke up in his hospital bed, connected to monitors and the like, in a drugged stupor. By this time my parents had arrived (like so many times before) and had taken Eli home. Darryn stood next to Noah's bed rubbing his chin nervously. Noah asked if he could use the restroom. Darryn said yes, and in the next second Noah leapt from the bed, almost a perfect swan dive, toward the floor. Luckily, Darryn was there to catch him. Noah's life was saved twice in the same day. After the drugs wore off, he was released from the hospital. Noah missed a week of school while he learned to walk again.*

I never knew the impact this had on Eli until Father's Day that same year. About a month after the seizure we planned to attend with some friends one of those workshops at Home Depot where the kids make a gift for Dad. He was excited to do the workshop. As we approached the store we noticed the local fire station had an area of the parking lot roped off for tours of ambulances and fire trucks. Eli became frightened and started crying, and began backing up, finally screaming and contorting. This from a child that hadn't given us a peep of trouble before. I had to leave our friends standing in the parking lot as my second son became inconsolable. At the time it hadn't occurred to me that he remembered the incident with his brother just a month prior. He too suffered, and sadly he was 3 years old at the time, unbeknownst to me, Eli has probably remembered more then I realize even today.

Noah had his first seizure in March 1998 at 18 months of age, shortly after a round of vaccines. (His doctors continually assured us that vaccines had nothing to do with Noah's problems, but I know otherwise.) Seizures were not Noah's only problem, however. He was diagnosed with neutropenia (the lack of neutrophils or a type of white blood cells) in 1997. He has suffered from acid reflux for as long as we can remember. He was hospitalized for 4 days at 4 years old for a roto-virus infection, something most children get through at home in a couple days. He had a cyst removed from his thyroid. His tonsils were removed due to frequent infections. In 2005, he was diagnosed with Lyme disease. In 2008 his pediatrician discovered his blood was missing carnitine, a key component for the body's metabolism. We took him to a specialist who diagnosed him with mitochondrial disease. His seizures stopped after we started giving him medication for the carnitine deficiency. But his behavioral symptoms continued and even worsened which always indicated a deeper underlying illness plaguing Noah. But what?

Early in 2013 Noah began to show some psychiatric symptoms that frightened us to the point we had to consider putting him in a home. He had frequent panic attacks (all day, minute to minute), tics, and sudden bursts of outrage over insignificant things. He often didn't make sense while speaking, and secluded himself in his bedroom. He fantasized uncontrollably about his videos, and spoke over and over about what we were going to do about the characters (not what *we would do if we were in the stories). He had repeated events of unrealistic fears. We took him to see an energy medicine specialist. The energy medicine protocol involved various supplements and dietary changes, as well as electrical pulses to adjust his body voltage (an approach similar to acupuncture). During one of our visits we learned of MMS, which this doctor indicated could be used in small doses to treat colds. Symptoms improved, but very little. But then they worsened again dramatically. We took him to his pediatrician who diagnosed PANDAS, and prescribed antibiotics. This helped but only temporarily. His psychiatrist prescribed anti-anxiety medications, with only marginal results. He was unable to attend school through all of this, and needed someone by his side constantly. His pediatrician prescribed Albenza thinking parasites may be involved (despite a negative stool test). This improved his behaviors some but only marginally. There was also the side effect of hair loss. Noah became completely bald, like a cancer patient.*

To see your son regress into psychotic behavior is something no human can ever imagine. Facing people everyday as they watched Noah regress and comment, it was a slow slip and then suddenly he was in a dive toward insanity. I felt he was slipping away. I scoured the Internet for answers. I had heard of Kerri Rivera but thought antibiotics were the only treatment for my son, like the vaccines, I had been told by doctors that the only treatment for Lyme was antibiotics. I emailed her and she answered immediately to "start the protocol." We began the CD protocol in June 2013. The protocol takes complete dedication, it requires steadfast commitment, and it's the best gift I have ever given Noah. I have learned so much about illness and it's roots through treating my son with the protocol.

Noah's behavioral symptoms started to improve immediately. He had fewer and fewer tics and outbursts. He was noticeably calmer. He now answers the phone, takes messages, and helps me scan food items at the grocery store. He even got behind the wheel of our car for a practice drive in the school parking lot. He asks a lot of questions. So much has improved in such a brief time. Prior to the CD protocol I scored Noah as a 38 on the ATEC test. Since then, he has improved to an 8. Darryn (Noah's Dad) scored Noah a 55 before we started, and two months after beginning it was a 22. It's amazing that a few brief months ago we were considering having him put him into an institution. Although we try not to think too far in advance, now we can hope our son can someday lead the normal life he was meant to.

I can't help one final story about Noah. Yesterday he asked me what autism was. We've never explained to him that he has it, because he would not have been able to understand it even if we tried. While I searched for an answer, he said "you know, Mom, Sponge Bob couldn't get his boating license because of his autism." I belted out a loud laugh and said "Ok. Noah, you know you have a little bit of autism too." And he said "Yes Mom, I know that," and smiled.

While he still has much to make up, years of life to regain, experiences to live, we're so pleased and grateful to Kerri and the others responsible for these protocols for these improvements, his life is being restored. I would encourage anyone to follow the protocols as closely as possible and do not delay. It is so fitting that the only opportunity I've ever been given to tell a little of what my family has been through would be afforded to me by the very person that actually saved my son…she gains nothing from what she has done for us, except the knowledge that she saved a boy's life that she has yet to meet, by saving my son, she has saved my family.

65. *Shay is doing so well it's surreal. In the past week the balance has shifted so that he is now recovered more of the time than not. He is having a 2-4 hour block every day or two where he is back in his old state (weeping, raging, zero-to-flipout, shrieking, obstinate, defiant, OCD, mean, potty-mouthed, etc.). The rest of the time he is lucid, has perfect mental clarity, is as funny as can be, does his math quickly, has perfect handwriting, gets along with his sister, is generous, expresses himself clearly, plays well, understands how others feel, etc. It is a mind blower Kerri.*

Although Shay doesn't have the awareness to know when he is in the autistic state, he is now clearly aware when he is in the recovered state. He comes up to me and says quietly, "Mom, it's gone again. Do you think it will come back?" I tell him yes, that it will be like this for a while, coming and going, until one day it will be gone and it won't come back. Shay and I have been discussing which "number" he is going to be in the line of children that recover. :) He is also planning a big "Goodbye Autism" party for himself.

Bella is getting over her meltdowns much more quickly. Sometimes I say "no" to something and then flinch, waiting for the screaming, but it's small, or sometimes she cries and gets over it within minutes. This is a big improvement for her. In her own words, she has noticed that she feels less anxious.

---

66. *We started the protocol 9 months ago. My son had 2, 3 horrible days every month. He wouldn't stop crying all day long, and he could not get any sleep at night. After talking to Kerri, we realized that these terrible days were always the days of the full moon. We saw improvement straight after the first PP. After 6 months, all of the bad behaviors were gone. I realized that the PP helped my child to improve month by month.*

And to prove that it is not only me who can see the changes, my son's therapist from school told me the other day: "Well, when is full moon again? So that you would start the Mebendazol? I love when he gets it, he is doing so well with it".

Staring atec in February was 50, now his atec is 34. Thank you so much, Kerri!

Betti, Hungary

---

67. *We are celebrating 3 months on CD and 2 PP's done! Our little angel has (or I should say had) extreme ADHD...we are not seeing much of that anymore. Our life altering issue is her lungs. This child has undergone over 35 surgeries in 8 years. 17 of those to surgically remove mucous from her lungs. She is hospitalized every 6 weeks with pneumonia. We have O2 in our home. She even had a port placed for IV meds at home. Our life has revolved around illness. In 8 years the docs have done every test in the book to figure out the cause. We've turned up nothing. In July, we flew to NYC to see a medical mystery doctor. He told us it's all nutritional and had us do some parasite testing. The testing turned up zilch, but thanks to Jean Marie, we ended up here. Long story short, we are making huge strides. We are not there yet, but to give you an idea, normally her pneumonias take 4 months to clear, this time it was only one month! She's only needed one antibiotic in 3 months, normally it's 8 antibiotics in 3 months. She's doing better in school too, which is just a bonus. Today she came home and told me she was able to run in gym, something she hasn't done for years! Thank you Kerri Rivera and Jean Marie and all the mods for welcoming us to this group and for helping us to take back this child's life. Bless it! Bless all of you!*

68. *30 days of gratitude, Day 4...I am thankful for an evening alone with my son. It's rare that we have time together alone. He jumped on the bed and danced around while I packed for my trip in the morning, talking and singing the whole time. The dinner table was lonely by ourselves so we decided to eat in front of the fire. I said, "This is fun, isn't it?" and he said, "I love this and I love you, Mom" Moments like this were non-existent before our journey brought us to Kerri Rivera who is an angel on earth. She gave us the direction we needed to heal our son.*

---

69. *I observed my kiddos at school today. Get this........my once severely autistic 6 year old choose and received a Spanish lesson that lasted 27.5 minutes (I timed it). The whole lesson he was sitting on his bottom, following instructions, answer questions correctly (in Spanish), smiling, great eye contact, and really enjoying his time with the Spanish teacher. FUA to the max! This time last year Henry was speaking in 1-2 word sentences to get needs met. Now he is talking in complete sentences in English and Spanish!!!!!!!*

---

70. *In the past, if I gave Nate a hug and told him I loved him, he would sort of hug me back, awkwardly jutting his backside out (to reduce physical contact) and patting me on the back and saying, "I love you, too." He would then escape as soon as possible.*

    After a few days of being on 7 drops (last week) Nate approached me with full eye contact, threw his arms around me, gave me a prolonged bear hug and spontaneously said, "I love you, Mom". I hugged him back and we stood there, just enjoying the moment, while I quietly balled my eyes out.

    Thirteen years. I had waited 13 years for that moment.

    And now he has been doing it almost every day since that first hug.

    This alone is the world.

---

71. *There's a lot of people jumping on the "Thankfulness Train" this time of year. Here's my contribution. There's only one major thing I can think of this year that I am thankful for. Of course, there's so much to be thankful for, but one thing that really stands out. That would be this protocol. God led us to see Kerri Rivera at Autism One this year. I only went because it was basically the one thing I didn't know much about. The day before Kerri presented, I saw my friend Laura at Karaoke Night (so fun by the way!). She told me she was doing CD and I'm pretty sure I looked at her like she was crazy! Saw Kerri the next day and I was hooked. We started the protocol June 3rd. Laura got me in this group. Thank you Laura, for your patience with me and for letting me know about this group! Thank you, Debbie who has talked to me many times about the protocol, about my son's aggression, reassuring me that things will get better and they have! Much, much better. We still have further to go, but there is light at the end of the tunnel. My four kids and myself are doing the protocol. We all have Lyme. We are all feeling so much better. We are no longer spending thousands of dollars a month on empty treatments. We are getting our lives back; our finances under control, and my son with autism went from an ATEC of 47 to 20 in four months. Thank you and love you all. And to Kerri Rivera, there are no words. Thanks and love isn't enough. You can be assured of my prayers for you, Patrick and your family for life. That's the least I can do for all you have done for all of us.*

---

72. *Small progress this week! My son had his ABA therapy yesterday and at the end the therapist gives him a "high five" before he leaves. The therapist asked my 3-year-old normally non-verbal son for a "high five". My son looked at him, gave him a high five and said "HIGH FIVE!" plain as day. My husband almost fell over sideways in shock! He also put his toys away without being asked to, and that too is a big change. CD is the best!"*

---

73. *We have been on the protocol since Aug 30th of this year. We've done one PP. My DS is 8 with severe apraxia and autistic.*

    ATEC: 4/2013: 62 starting some biomed

    ATEC: 8/15/2013: 51 before CD

    ATEC: 10/6/13: 24

    Some WOW moments for us so far:

    * *Playing soccer with other kids at school three weeks after the start of the protocol. *Really "on" in conversations...adding information when we didn't think he was listening because he was playing on his iPad.*

    * *Talking about other things in his day besides Minecraft or Plants Vs Zombies.*

    * *Going from a battle with homework that would last over three hours to picking him up at daycare with his homework already finished. And correct. That he did on his own.*

    * *His ST saying he is able to follow three step directions better and was able to talk/ask Who and What questions for nearly the entire hour...I actually had to tell the ST what we are doing because it was obvious to her it was more than just diet.*

    He's just happier. We're all just happier.

---

74. *We came back from our Intensive in December of last year on a high. After a week with the Son-Rise® team, Jordan was saying 20 different words; he was more confident, and more mature. He had such an amazing time learning and growing. Every morning he would take his facilitator by the hand and walk him/her into the playroom. He really grew up that week. I loved watching him grow through love and laughter. What an amazing, wonderful thing to witness.*

    We, also, learned at our IAHP appointment that Jordan was reading at about a third grade level. What a trip! The evaluator held up a paragraph for Jordan to read. It was so fast I didn't even have time to read it. Then, she held up a question for him to read, then, a sheet with the answers. She, then, asked him to point to the answer and he did. He was three for three. My son taught himself to read. What?! We were, also, very surprised to find out that between running our Son-Rise® program and incorporating exercises from IAHP, Jordan improved 433% from September to December. WOW!

    Between the two experiences on the east coast, I felt like I didn't know my own son. I always believed he can accomplish anything and yet I had tremendously underestimated his ability. I was so excited to get home and start his program. I knew that if we focused and put everything into his therapy for two years, he would be ready to go to mainstream kindergarten.

    I was so excited; I was beyond hopeful; I believed he was on his road to healing.

    On Christmas Eve, Jordan was crouching down and just froze there. I called his *name, no response. I ran over to him and noticed his lips were turning blue. Immediately, I tipped him over, hit him on the back and did a swipe in his mouth to get out what he was choking on. There was nothing there, but he gasped for air and began breathing again. Did he just have a seizure?*

January 15 has become a day I will never forget. We woke up to Jordan throwing up and he didn't want to get out of bed. He just wanted to sleep; he must have caught the flu and needed to sleep it off? Finally, around 10am he got up and ran around the house like nothing was wrong so we began therapy. While I was in the playroom with him we were jumping on the trampoline and he suddenly stopped and looked at me with a panicked look on his face and reached out for me. Clearly, jumping was too active for his upset stomach so I brought him to the family room with Kaiyan and my helper.

Suddenly, he looked off and froze again. He then threw up again and passed out. Oh shit, he just had a seizure, and another one, and another one, and another one. All the while staying unconscious. I brought him to the ER where he slept and had another mild seizure. While he was passed out they gave him a CT head scan to make sure there wasn't a bleed or something in his brain. Everything was fine.

Shortly after, he woke up. He was playing, running around the ER as if nothing had happened. He went through what little water I had brought and was asking for food and water. A good sign, I thought. The ER wouldn't allow him to eat or drink anything "just in case". Anyone who knows anything about seizures knows that no food and drink could very well cause a seizure…and it did.

This time the ER Doctor went into full-blown panic mode, probably because she saw what happened. They began trying to get blood out of him. On a good day this is almost impossible. Now, they are trying on a kid that didn't have any fluids in his system to help expand the veins.

And so it went, needle poke, dig around in the arm searching for a vein, Jordan would hold his breath from the pain and have a seizure. Seizures are typically the brains way of trying to get oxygen; the pinpricking was inadvertently inducing seizures because of the breath holding. This went on for two hours: needle poke, dig, hold his breath, seizure. They poked him 27 times: in his neck, head, ankles, thigh, anywhere they could think of.

This of course wasn't enough. They insisted on doing a spinal tap to check for meningitis, which we had been declining all day. Now they are telling us that the way he had been seizing for the past two hours is classic for meningitis and not doing it could be fatal. They crunched him in the tightest ball they possibly could and held him so tight he couldn't move…or breath. After the fifth poke, I began to get faint and had to sit down. They didn't know what they were doing. When I sat down I could really see how they were holding him and asked if he was breathing. The nurse looked down and said, "he's blue, get the oxygen" and they put oxygen near his face.

I immediately thought, "they're killing him. They don't know what they're doing." That's when I reached out to every FB group I'm on and asked for prayers. That was the only thing that was going to save him.

When they, finally, let him go he had the worst seizure he's had to date. He was actually convulsing. As they went for blood the 28th time, I told the nurse that's enough. This has gone on far too long. She agreed, ran and got the Dr., they wheeled him into to the CPR room, and called for a crash cart. I think I said, "what the fuck?" She said it's "just in case".

They proceeded to get a drill and drill a hole through his shin and into his shinbone so they could put in an IO (IV into the bone). They pumped him full of all kinds of drugs and, at long last, drew blood.

He looked up at me with a longing in his eyes and reached out for me. I went over and put my hand on his chest and he passed out. He was transferred by ambulance to our local children's hospital; we were allowed to follow the ambulance, but couldn't ride with him. He remained unconscious until he arrived into his private room at the hospital.

The first thing he did when he woke up was reach out for my hand and give me a kiss. My little boy!

The blood they had finally taken revealed low electrolytes and bicarbonate at a life-threatening low level. Screaming and crying can drop bicarbonate, which he did for two hours in the ER; lack of food and drink can drop bicarbonate levels, which he wasn't allowed to have in the ER. Low bicarbonate levels can cause seizures. What a recipe for disaster!

The hospital ran every test they could think of and no reason was found for the low bicarbonate or the seizures. We came home after five days, followed by several follow up appointments over the next two weeks. Everything seemed to check out fine until I brought his labs to his DAN (Defeat Autism Now) Dr., she looked at the blood test, pointed to something and said, "Lyme, every time I see this in one of my kids, it's Lyme."

We had the (one of many) $800 test at the best lab in hopes we wouldn't get a false negative, which is very common with Lyme. Jordan came back positive! Now we had something to go on. I was excited to finally have an answer to what was going on with my little boy. Lyme is known for seizures and abnormal electrolyte levels.

We scheduled an appointment with a Lyme Literate Medical Doctor (LLMD) and were ready to start healing our guy. Two days before the appointment, we woke up to Jordan vomiting. He didn't want to get out of bed. He finally did and ran around like nothing was wrong and then he had another seizure. We were told to take him for blood tests immediately the next time he had a seizure. We took him to his DAN Dr. an hour away, but knew they would be able to get his blood without difficulty. We went to lunch, came home, he had a seizure, turned blue, threw up, passed out, had a seizure, turned blue, threw up, passed out, had a seizure, turned blue, threw up, passed out. His DAN Dr. just happened to call to check on him in the middle of this and told us to get him to the hospital before he had a heart attack. We went to the ER...

And so the story goes every seven weeks TO THE DAY, seizure, turn blue, vomit, pass out, seizure, turn blue, vomit, pass out, admitted to the hospital where he was pumped full of drugs for hours to try and stop the seizures, and they ran the same tests over and over again. The hospital didn't think he had Lyme even with a positive test and thought that seizures every seven weeks to the day was a coincidence, although, they could still not offer any explanation.

We tried treating him with ABX that the LLMD prescribed. One teaspoon and he had a seizure, turned blue, threw up, passed out, had a seizure, turned blue, threw up, and passed out. This time the Valium I administered did stop the seizure cycle, most likely because it was induced from the ABX and not part of his regular seven week seizure cycle. He had multiple mild seizures every day for two weeks from that one-teaspoon.

Clearly, ABX were not going to work. I emailed with a woman who coauthored the book, "The Lyme Autism Connection" and tried the homeopath that she used down in San Diego. This was definitely working with detoxing, but it wasn't helping with the seizures. He went from seven weeks to the day to three weeks to the day. This treatment was working on opening Jordan's detox pathways so his body could rid itself of the Lyme. I liked the idea of treating this way, but the process could take four years. I didn't want Jordan having what the neurologist told us were life-threatening seizures for four years. I was desperate to find something to help him.

We knew what was wrong, but needed something safe to treat him. Nothing was covered by insurance and we were running out of money. I had lost hope. Every treatment seemed to make him worse and I couldn't bear to see my boy suffer the way he had been for the past six months. I didn't know what to do.

I, literally, got on my knees and prayed for help every night. In the same week, a woman who attended the same Son-Rise® start-up emailed me about a newly discovered treatment she was doing. I had signed up for a Lyme conference that mentioned the same treatment, and I had signed up for an autism telesummit that unbeknownst to me had Kerri Rivera as one of the speakers. (Kerri is the woman who discovered this treatment for autistic kids.) I emailed back and forth with Kerri, I posted on the FB group, trying to get all my questions answered. I was terrified to treat Jordan with anything given his seizure history.

Then, he was hospitalized again for uncontrollable seizures and I realized I couldn't let fear get in the way of helping him. I started the day we were home from the hospital at a really low dose that Kerri recommended. She has held my hand through this process to help prevent him from having any seizures.

Since we started in May, he went 10 weeks seizure free and had several mild seizures that he brought himself out of without any extra medication and NO hospital. He is currently at a record-breaking 15 weeks seizure free.

Through this treatment we have learned that in addition to Lyme he has three different visible parasites. He is loaded—he passes more parasites than most of the teenagers in the FB group. It scares me to think what would have happened to him if we weren't led to this treatment. Thank God for Kerri Rivera sharing what she has learned and being so gracious and generous with her time.

I'm incredibly grateful we attended the Son-Rise® Intensive before all of this began. His team of facilitators created a base for him that continued to blossom even through this tumultuous year. Coupled with the parasite treatment, Jordan has gone from 20 to 90 words—he's picking up new words almost daily. His balance is amazing, his muscle tone is getting better, and he understands everything. Even more exciting, he's playing interactively with his sister. He is on the road to recovery.

We are finally able to start picking up the pieces of where we left off back in December. I felt stuck; I had PTSD; I needed help believing we can move past (what I had deemed) the horror of this year. The first thing I did was turn to Son-Rise®; I started having monthly dialogue sessions with Bears (Barry Neil Kaufman)—the co-founder of Son-Rise®, which has been incredible. This is one of those times that happen in life that can make you a better person or break you. Bears is guiding me to be better. I just arrived back from another Son-Rise® training where I realized that staying in fear and being nervous about getting excited or anxious about feeling hopeful again will prevent me from being the *bridge to healing that I want to be for Jordan. I felt a relaxation that I haven't felt all year. I've been wound real tight, waiting, being on the look out for that next seizure.*

This year has had a huge impact on Kaiyan as well. She'd see her brother sick and then her mommies would disappear. I'm noticing she's wound pretty tight as well. Her favorite moments are when the entire family is together. I'm hoping talking with her about what happened and spending more one on one time with her will help. She asks daily is Jordan throwing up, is he having a seizure? She doesn't really know what it all means, but is really aware it's not good. She, recently, said she wanted to have autism. After I explained to her that autism is why Jordan isn't talking or riding a bike, she decided she didn't want autism. She wants to be a boy and insists on wearing Jordan's clothes almost daily. I think it's her way of healing and feeling included in the process.

We've been doing Son-Rise® again and it feels great. I do feel like we have some time to make up for since this year has been spent dealing with health issues rather than cognitive issues so we are working up to having a full time program, having quarterly consultations, and doing another Intensive.

Son-Rise® and Kerri's protocol's are not only bringing our son back, they're bringing me back.

---

75. *Healing From Autism*

Hospital birth, bottle fed, fully vax until age 4. Digestive disorders beginning at age 1. Low tone, speech and language delay. Mild early education interventions. GF since age 2, GF, Soy free, sugar free since age 4. Majority of autism symptoms presented in 2013; in the fall of 2012, our son did not qualify as being in need of special education, and thus, could not attend school in our district at all.

Long winter days in a tiny home had me sending him out to play, in our tiny backyard, which had a patch of grass and a bush, under which small animals and birds lived. In his boredom, he sprinkled the dirt all over himself; his head, face, clothes, shoes, on several different days.

Later in the early spring, the symptoms of autism appeared more and more obvious: disobedience, disinterest in sitting for a story or coloring, lost interest in puzzles of more than 24 pcs, jumping on and off furniture for hours, despite scoldings and time-outs. Sudden relapse in bedwetting. Then by summer, the symptoms were very obvious: longer tantrums, rages, anger issues, hitting, fighting, a mean spirit; running out of the house without permission, with no respect for danger or regard for safety. Weepiness, hypersensitivity to touch, suddenly repeating the same phrase, over and over all day. Lost balance and coordination, bumping into walls and furniture, stubbing toes. Shoulder shrugging, all day, but not able to stop or say why he was doing it. Lost dexterity—could not put cards in a box. Lost organizational skills—could no longer put toys or clothes away. Extreme nighttime drooling and bedwetting. The need for sameness, and to have all food served cold. He frequently did not seem to be in his right mind.

In desperation, I asked a group of special needs moms for advice, and ordered a handful of books on Autism, including Healing the Symptoms of Autism by Kerri Rivera.

It was the best money I ever spent.

I read the book thoroughly, ordered the supplies and began the CD Protocol on July 8, 2013. Starting ATEC: 46. We ramped up slowly on drops, serving doses 8 times per day. My son has had two low dose/partial PP's, to minimize the emotional meltdowns, the gas and discomfort. I did not realize how very deficient in minerals he was, and did not give him enough ocean water. Now I order the big 10-liter jug of ocean water, so I am not tempted to be stingy. He needs more than 100 ml ocean water per day for his digestive system to handle the parasite dumping during the PP. His previous digestive trouble was unexplained chronic constipation, since age one, and that is really linked to his high need for the OW. Interestingly, his doctors had all said to keep giving Miralax, and that he might possibly need that laxative every day for the rest of his life. We stopped the laxative when we started the CD program, and began the enemas right away. When I sifted the stool and saw the number of parasites, I understood finally why he couldn't poop well on his own for the last 4 years, despite diet interventions. We are looking forward

to PP #3, in which I hope to ramp up to full dose on all components.

Poop sifting revealed what 6 pediatricians and 2 GI specialists could not tell us. Our son is infected with at least 6 different kinds of parasitic flukes and 3 kinds of parasitic worms. His stool has had a very large amount of sediment, with odd, tiny, but defined shapes. A search for "rodent and avian parasites" reveals photos of very tiny creatures.

I just completed his ATEC at Kerri's request: November 12, 2013: he now scores a 10, with a 1 in language, 2 in social, and 7 in health.

What does his recovery look like? We enrolled our son in a small private school this fall, with one teacher for 6 students, no aide. He now memorizes poems, Bible verses, and songs. He sings, knows his alphabet and all the sounds of the letters, and can count to 47 with precision. He shows kindness to others, is mindful of safety, and has glowing reports from his NT school. He sits with me for stories, and brings books to 'read' to his baby sister. He is not literate quite yet, but enjoys word search games, where he can complete levels and advance to more challenging puzzles. Twice now he has played the word search game on Leap Frog for 45 minutes. He has regained his balance and coordination. The rages, anger, and shoulder shrugging disappeared within the first 10 days on CD. He identifies printed numbers and letters, and does well in math, phonics, penmanship and reading class. He is learning to write numbers and letters in cursive, and follow directions on drawing worksheets. He operates the TV and DVR unaided, recording and watching programs he chooses.

He is happy; he has a patience and joyfulness about life that was previously lacking. He tells me jokes, and now has the coordination to operate a hand puppet while speaking for the puppet!! He is thoughtful and obedient. He has regained fine and gross motor skills, and motor planning skills.

There are still improvements we hope for: further speech, language, and vocabulary improvements. Social graces. More improved digestive health; I can see that he still has parasites in his stool. We will continue the CD Protocol, and we look forward to further healing. Thank you all for your support and encouragement.

The CD Autism Program has done more than cure our son of his brief affliction with Autism; it addresses the pervasive, underlying gut disorder and parasite issues that have plagued him all his life. Where complex testing in expensive medical centers has failed, a simple home health program succeeds. Thank you Kerri Rivera!!

76. *Well...we too did an ATEC! Our son was 11.5 when we started the CD Protocol July 2012 First fully loaded (All PP meds on board) was November 2012*

VERY First ATEC: Jan. 2006= 124

ATEC in July 2012: 64

ATEC August 2012: 52

ATEC Oct. 2012: 48

ATEC Dec. 2012: 45

ATEC May 2013: 40

ATEC August 2013: 39

ATEC NOV. 2013: 33

I actually did one myself as I have done throughout all the previous, but I had hubby do one and combined and divided by the 2! Resulting in the 33! Oddly, it's the "SPEECH" that remains consistently around 14 & 15 throughout ALL of these! All the others continue to go down! We've had some ups and downs, but with continued evacuation of some MIGHTY big worms...we continue to see major gains!

This is the ONLY Protocol that has EVER given us the progress we've seen in one year's time let alone in 6 months! Thank you Kerri Rivera and especially Patrick Rivera for being such pioneers in this journey to recovery! You are SOOO loved and appreciated beyond words! Thank you for never giving up! Thank you for giving so much, and bringing healing to our son! He's come a long way, but we have a ways to go! But with the progress he's making...I honestly feel it will seem like nothing from where we've come!

Again THANK YOU! XOXOXO!

---

77. *We had a great day today. For the first time, he joined me in throwing something at a target (blocks at carpet circles). He laughed and smiled and pulled me to play with him better than he has in weeks. I heard a lot more sounds and approximations today. When I couldn't figure out what he was saying, he kept trying until I did.*

My favorite part of the day, and I know this is odd, was when his head collided with my mouth. I exclaimed in pain and he was clearly worried/empathetic. He kept trying to get close to see my lip and giving me kisses. I've seen him concerned before but never that clearly empathetic.

---

78. *Just wanted to share that on 21/8/13 my sons ATEC was 46, just redid because we're at the start of our first pp & his score has come down to 31! The big improvement came after introducing CD enemas 3 weeks ago. Although he's been on oral CD for about 2 months, it was always the case that if I missed a dose within 1/2 hour he'd be sensory, inflexible, super-anxious, shouting, aggressive etc & I'd have to syringe it down him but within 10mins of dosing the lightswitch would flick on & he'd be in control & a calm, loving boy again. Since the 3 enemas I gave him 3 weeks ago (only 3 because then his brother had one, panicked, screamed & totally freaked out my 4yr old who up til then had said he liked them bc they made him feel good afterwards) his personality has been shining through again, there's less panick/anxiety, less sensory issues & there's no 'reaction' if I miss a dose or more of CD. I've seen more improvements this week, on Monday he decided to sit down & colour & proceeded to colour neatly,*

inside the lines - he was so pleased with himself! (He used to colour so neatly but about 6 months ago his colouring became scribbly & he frustrated that he couldn't do it neat anymore). Then tonight, when I asked him to hop out of the bath, HE DID! He stood up, stepped out & as I handed him a towel he took it & said 'thank you mummy' and wrapped himself in it & walked out - all completely relaxed - like it was usual!!! WWWOOAHAHH! Yes! He has never done this, it's always tantrums, always arguing, needs to bathe longer, has to be made to get out, refuses to step out, has to be lifted, needs me or hubby to wrap & dress him, & usually all while panicking, crying, demanding, arguing or screaming because it's cold, different, he doesn't want to, it's not right etc. I just can't believe what happened tonight! I'm still astounded. It's like the Finlay we 'know' isn't Finlay & now we're beginning to see who he really is!

So I'm REALLY excited about our 1st pp. We're already seeing full moon behaviours & although tough it's been great to see them because it's a reminder of what Finlay was like every day until 3 weeks ago!

He's already healing. It's soooo exciting! Thank you Kerri - may God bless you & all the children on this protocol. Healing, here we come!

---

79. True Hope by Maggie Kaye

Most people don't know that babies are born with about 100 million brain cells. That is more cells with more connections than there are stars in the galaxy! This is difficult to picture when you consider an infant's brain only weighs about 12 ounces at birth, about 25% of it's full size. But, there is a logical explanation for this: Brain cells; like the brain itself, have yet to grow. As neurons (brain cells) mature, they become larger, stronger, and form tentacle-like branches that form synapsis (connection and communication) with other brain cells. Brain growth is directed by neural activity. The neural activity of the cerebral cortex (the largest, most complex part of the brain) molds the way a child will think, behave, and acheive.

Conventional wisdom long held the belief that the human brain cannot change… that it is hard-wired at birth. Scientists today have proven that this is not the case. They have found that the brain is malleable. It has the ability to change shape, size, number of branches, number of connections, and even the strength of it's connections over time. The potential growth, or neuroplastisity (neuro means neurons or brain cells, plastic meaning changable) is far-reaching. At birth, the brain is simply a blueprint of what it will become. No one can predict the amount of growth that a brain has yet to acheive.

For the parents of a child with Autism, the diagnosis can be devastating. But, the biggest and most harmful blow can come from the prognosis: There is no cure. Kenn and I were given a diagnosis of an Autism Spectrum Disorder for our son, but we were not given this prognosis. In fact, the word "Autism" was avoided at our son's diagnostic evaluation until the therapists heard us say it first, and realized we were comfortable with the word. For many parents, a diagnosis is the end all, be all. For far too many parents, the words "No cure" are scary and harmful. We did not hear these words nor would we have accepted them as truth if we had. We do not want to change our son. We want him to be his best self. We are not looking for a cure. We are in search of treatment and *healing for his symptoms. We know and understand that a Therapist is not God. A Therapist cannot predict the future for our son. He or she can simply assess where Gunnar is developmentally and give us options for a treatment plan for a healthy, successful future. Our son's diagnosis was not an ending. It was simply the begininning of a journey through a world of strength, challenges, support, opposition, and the love of a parent that cannot ever be fully defined.*

Our son, Gunnar was born with an Autoimmune disorder, Hypothyroidism. He was tested at birth, was diagnosed, and has been treated by an Endocrinologist since he was three days old. We were not upset by the Hypo diagnosis, but rather ready for the course of action to regulate all of Gunnar's bodily systems. What did surprise us was the constant shrugging off of our concerns over our son's developmental delays and gastrointestinal symptoms over the last few years. We were continually disappointed by the compartmentalization of our son's physical and mental challenges by every Pediatrician and Specialist we saw. To us, Gunnar's symptoms all had to be as connected as are all of his body parts. But, we were treated as if our concerns were dumbfounded…as if digestive, behavioral, and physical challenges were not intertwined.

Let's be clear...The brain does not function in isolation. It is a team player. It needs vital nutrients and informational input to perform at optimum levels. To fill these needs, the brain relies heavily on complex interactions between the immune, endocrine, and gastrointestinal systems. Gunnar, like so many other children on the Autism spectrum, suffers from gastrointestinal issues uncluding chronic constipation, diarrhea, and abdominal discomfort. Scientific studies and countless parent reports are pointing to gut inflammation caused by casein, gluten, and soy in our foods. Gunnar has shown clear physical and behavioral reactions to these foods, and improvement in these areas with the removal of Gluten, casein, and soy from his diet. This is not an allergy from the perspective of a traditional allergist. This is what is called a T Cell Inflammatory response.

The use of a gluten, casein, and soy free diet as a means of treatment for Autism symptoms has gotten mixed responses. Many parents and doctors who deal with Autism, Immune, and Endocrine disorders rave about it's benefits, including me and my husband. Still, others call it "junk science". Clearly, if someone is looking purely at one body part at a time, they will not be open to the truth that the body and brain are codependent. I, personally, have been challenged by the choice to follow this diet for my own thyroid disease and for my son's Autism. I was even told on an online support forum that the GFCFSF diet is "fake science" and is causing "false hope" for parents of children on the Autism Spectrum.

False Hope?

This may be one of the biggest oxymorons I've ever heard. This ranks right up there with "God Hates". I cannot say what causes people to be so oppositional to our choice to follow this eating plan. Perhaps the resistance to this way of eating is due to a simple misunderstanding of the leaky gut syndrome that plagues so many children on the Autism Spectrum. This leaky gut syndrome does not cause food allergies. It causes food sensitivities.

In food allergies that many parents are famililar with, the symptoms arise immediately in the form of hives, watery eyes, sneezing, or difficulty breathing. They can even produce life-threatening anaphylactic shock. Only about 10% of *autistic kids have this type of allergy. Some 85% of children with Autism have food sensitivities. Rather than immediate physical symptoms, these sensitivities produce and inflammatory response that results in more subtle mental and behavioral symptoms that can take anywhere from 6-72 hours to appear. The brain's primary fuels are oxygen and glucose which are manufactured from nutrients in our food supply. It only makes sense that we get out of our brains what we put into our bodies.*

When it comes to the treatment of any neurological disorder, many parents and doctors focus solely on the obvious cognitive impairments and fail to realize that correcting their child or patient's underlying intestinal imbalances can; and often does, lead to significant overall improvement. Leaky gut syndrome which is so prevalant in children with Autism Spectrum Disorders causes a long list of vitamin and mineral deficiencies because the inflammation process damages the various carrier proteins normally present in healthy GI systems. Our plan is to treat our son from the inside out; Reduce the inflammation, test for vitamin and mineral deficiencies, provide necessary supplements, test for toxin poisoning and remove if necessary, provide therapy as needed, and set Gunnar up for the best possible outcome. When you feel better, you perform better. Who can argue with this rationale?

Many people have differing opinions on what does and doesn't work for families facing Autism. There is a saying in the Autism community, "If you have met one child with Autism, you have met one child with Autism" There is a broad spectrum with severe to mild physical and behavioral symptoms. One person's treatment protocol might not work for another person. I am not judging what others choose to do to treat their child, I am simply sharing what has worked for us thus far. I am not searching for or promising a cure, I am finding a treatment for and healing the symptoms that affect our son. Whether or not you agree with how my family and I choose to approach this journey does not change the outcome. If we have helped one person understand this complex neurological disorder, we have helped spread understanding, and that makes us happy. And, if you are a family facing Autism and don't know which avenue of treatment to pursue first, our advice is: Go with your gut.

---

80. Recovery #107

Having a sick child is a full time job. Even when they are well, your mind is filled with ways to keep them from getting sick again, ways to improve their current state, ways to help others who were where you were just a short time ago. It is my hope to do all three.

Autism entered my life a few years ago, but I didn't recognize it. My son, Gunnar was three when my husband and I realized he wasn't progressing as he should. He wasn't asking why, what, where, or when questions. He wasn't growing. He was lining up his trains and cars, screaming if we touched one of them. Gunnar was oppositional and refused to go into a crowded place. He didn't look anyone in the eye, and it seemed to make him uncomfortable to try. He had several sensory issues such as covering his ears when music played, even when the volume was low. He had an aversion to touch and would often yell, "ouch!" when someone touched his arm softly. He was constipated. He smeared his feces on the wall. He often woke in the middle of the night screaming. My husband or I *would find him in the fetal position and he couldn't or didn't respond when we asked what hurt. Gunnar went ballistic when we tried to wash or cut his hair or brush his teeth. He was indifferent to me leaving for work, and what was most painful for me, he ran and hid when I returned. He didn't much care for his own mom and that was heart breaking for me.*

We also noticed some regression. Gunnar stopped showing interest in potty training. He went from saying, "Gotta go pee pee!" to not even recognizing he'd soiled his pants. He started refusing foods he previously enjoyed. He went from using several words in succession to not speaking more than two words strung together. He started to be extremely picky when it came to the foods he would eat and spent most of his time at the dinner table staring out of the window. We were losing our son.

Our son was born with an autoimmune condition, Hypothyroidism. He started taking medication just three days after he was born. Initially my husband and I interpreted some of our son's Autism symptoms as side effects from his medication dosage fluctuations. After all, we voiced our concerns to his Pediatric Endocrinologist, and she assured us he was fine. She said the medication was likely making him lethargic. She said boys took longer to potty train. As time went on, Gunnar's symptoms multiplied and worsened, and other caregivers took notice and shared their concerns. At our son's next Endocrinologist appointment, I insisted on a referral for a developmental evaluation.

Our son was diagnosed with an Autism Spectrum Disorder in January of 2013. He was 4 years old. I immediately went into research mode reading everything I could get my hands on that was related to healing Autism. Two weeks after his diagnosis, we started a gluten, dairy, soy free diet and saw digestive issues disappearing. The night-waking was going away. He turned toward us when we called his name. He spoke in short sentences, requesting to watch a movie or eat a snack.

A month after Gunnar's diagnosis we had further testing done to see where he was on the spectrum. When we showed up for the appointment, the diagnostic team was impressed with our son's progress. He responded when they asked him to point to specific items, and he had no trouble matching items together. I remember thinking he was much smarter than I'd realized. Clearly the problem with our son was physical.

The test results showed Gunnar's speech and language to be at a 23 month old's level with his gross and fine motor skills not much further along. We were determined to heal our son and scheduled an appointment to have all kinds of testing done including food allergies, comprehensive stool analysis, and genetic mutation tests. These results showed that our son was sick with food allergies, leaky gut caused by an overgrowth of bacteria, and pathogens. We knew what we were up against. We were on the right track with diet, but had to find out how to get rid of the bacteria and pathogens. In my research, I stumbled upon the CD forum. It was the first treatment program that made complete sense to me.

We received the book in June. It was the first book our son ever showed interest in. He kept saying, "This is it!" and, "This book is the best!" He didn't want to put the book down. My husband and I looked at each other with tear-filled eyes. We knew Gunnar was right; this was it.

We started the CD Protocol immediately...just five months after our son's diagnosis. His starting ATEC was an 80. Immediately we saw regular bowel movements, diminished dark circles under his eyes, and healthy skin coloring. Gunnar began to sleep through the night. Four days into the protocol, our son started taking turns and sharing with other kids. That night, he danced at a graduation party. Every day, his vocabulary improved.

Four weeks into the protocol, our son no longer needed his thyroid medication he'd been on his entire life. He started using the bathroom unprompted. He dressed himself unprompted. He brushed his teeth on his own. Gunnar didn't mind getting his hair cut. He began to engage in reciprocal conversation.

Seven weeks into the protocol, we saw so many behavioral improvements; we decided to take the ATEC test again. Gunnar's score was a 40. It was cut in half in seven weeks!!! Eight weeks into the protocol our son celebrated his 5th birthday. It was the first time he showed anticipation for a party. He was excited to open his gifts and thanked everyone individually. Normally, Gunnar would run and play by himself when guests came over. He had always covered his ears and yelled, "Stop singing!" at previous parties. Gunnar was happy when everyone sang Happy Birthday to him. He was healthy and present. It was one of the happiest days of our lives.

Twelve weeks into the protocol, our son started mainstream preschool. All previous attempts to take him to school resulted in anxiety and complete meltdowns. He started speaking in complete sentences and his speech was relevant to the situation. He started showing interest in books. Fifteen weeks in, our son read his own bedtime story with no help. He picked out his own Halloween costume. He brought home his artwork and proudly showed it to us. His teachers continually voiced their pleasure at his progress. His speech therapist said he was at age level for speech and language. (that's 3 years worth of progress!)

Twenty-four weeks into the protocol, our son's ATEC score is a 10. He continues to go to mainstream preschool and is interested in learning. He is reading a variety of books. He is counting in English and Spanish. He knows his US geography, and does puzzles that are advanced for his age. He asks questions. His speech is appropriate and relevant with continued improvement in vocabulary and articulation. He loves to sing, dance, and paint. Our son is smart, funny, and very affectionate. He spontaneously hugs and kisses me and tells me he loves me. And the best part of all? When I come home from work, Gunnar yells, "Moooommmmyyyy!", runs up to me, jumps into my arms, and wraps his little arms around my neck. I have my son back.

I could never sufficiently thank Kerri Rivera for all of her selfless help. She and her team of parents give freely of their time and energy to bring kids and their families out of the depths of Autism. It is my hope to pay it forward. All children deserve a healthy childhood. If you are holding this book, you are likely in the position my family was in just 10 months ago. This is the way to turn your feelings of desperation into determination. This is the way to change the future for your family. This is the way out of Autism.

—Maggie Kaye

81. *Ok, my brother is 17 and started CD at the end of october. We are on 9 drops as of today. Non verbal. My mom called me crying and my brother said "I want snack please." My mom asked if he wanted something and he "no, no, no." He got his yogurt (df) out and said "thank you." This is just so amazing. Thank you God for putting us on this path and around so many amazing warrior parents to help us with our journey. Happy healing to all!!!*

---

82. *At the age of 3 our son was absent. He talked a lot... to himself. Scripting. It was almost impossible to get him to look into our eyes when we spoke to him. He almost never answered back, and when he did, it was scripting again (from a show on TV or a movie he was into.) He played by himself. He refused to potty train. He would wake up in the middle of the night crying and holding his stomach but was unable to tell us what was wrong. He stopped eating whatever we put in front of him and seemed to only want fruit snacks, American cheese, hot dogs, chicken nuggets and fries. And thats it. He didn't seem to be growing anymore, both physically and mentally. He seemed to be slipping away from us. You could see it in his eyes. Those eyes that were pointed off in a different direction. It was gradual...hard to notice right away. But definitely a change, and not one we were happy with.*

Up until this time he was progressing like any other child. He ate well. He slept well. He played well. Was even starting to potty train. A typical child. But not anymore. Then there was the diagnosis. Autism. No parent wants to hear that, but we knew. Every parent knows what I mean. You just know. And at this point we (as parents) have to make a choice.

We can choose to do nothing. Just live with it. We can listen to our doctor, or a specialist in the field. You know, someone who claims to know everything about you and your child in the 5 minutes they spend with you. Or you can talk to other parents. Parents that spend every moment with their children. Not just a few minutes every six months. You can network. You can research. Ask whats working for them, or more importantly what's not.

We chose the latter, and because of this our son is recovering. He looks into our eyes. He asks questions. He sleeps at night. He's potty trained. He sets the dinner table with his older sister and asks to be excused when he's done eating. He's in pre-school now and plays with all of the other kids. He calls them by name. He draws. He paints. He reads books. He does all of the things a typical five year old does.

Our son is recovering, and I need to give two heartfelt thanks for this. The first to my wife for the countless hours of reading and research she did immediately following our sons Autism diagnosis. She is an incredible woman who refuses to give up on anything. Ever. She refused to believe that we had to accept what was handed to us. She knew in her heart and mind that we could make a difference. We just had to do it. Because of this determination, she found Kerri Rivera, whom is the next person I want to thank. She is a true pioneer, someone who refused to keep thinking inside the box, someone who is helping families like ours all over the world.

" We have the world at our fingertips. Every book, every encyclopedia for us to read at our local library, in our own homes, and now on (most) of our cell

phones. It's a beautiful tool for us to use. Use it. Do your own research. Talk with parents. Find out what works and do it.

The only thing I really have to say about Kerri's book and the protocol it explains is that it works. We have our son back.

—Kenn Kaye, Father of Gunnar Kaye

83. *Last night, my 19 year old son came out into the kitchen, picked up a purple straw from the dish drainer, and pretended to play it like a clarinet with great enthusiasm. Pretend play. For the first time in 19 years.*

84. *We've been playing the "healing game" with my son for some time, (7 years 11 months old) but we've never really healed anything. To be honest we've never known what we were up against...until now. Like everyone here, my child is very sick. He doesn't grow, he has maldigestion and malabsorption issues and he's chronically constipated. His history of constipation is long and extensive. ANYTHING that disrupted his gut flora would result in this extreme constipation; whether it was a cold, a new food, treating the GI. To say he has leaky gut is a tremendous understatement. So, when we started Kerri's protocol 6/15 my biggest health markers for initial healing were growth and for him to poop daily on his own. With his history being what it is I wasn't surprised to see he was instantly constipated. How could he not be? We are killing pathogens and parasites like gang busters. We started enemas on day two and have done them daily. But, an amazing thing is happening! He is healing. He has grown two inches in these last three months and he starting to poop on his own. BTW, these aren't just lame little poops they are total TROPHY poops. They are soft, formed, have no undigested food and there is lots (I'd show you a picture but we barely know each other)*

Yep... It's taken 3.5 months to get here, but his little; fragile and damaged body is doing what it is supposed to do.

Trust the process... It's not easy; it's not always fast. There will be ups and downs. But, my gosh, watching the body heal is an extraordinary thing. It's nothing short of a miracle.

85. *Hi Kerri,*

I started my son on the CD last night. I'm taking it with him.

Well, I for one, feel fabulous. My lungs are clearing up, the low-level depression I felt is evaporating. Wow!!!

I've noticed some subtle changes in my son. He is more "with us" in the world and not looking out the window at cars quite as much. Also, a little less flapping his hands.

He and I have been doing a ketogenic version of the GAPS diet (no sweet vegetables, no dairy, no sweeteners) since January of 2013, and I think that's why we've been so successful with the CD! :)

I was thinking of upping it to two drops per 8 oz bottle--one for him, one 8 oz bottle for me.

I am hoping that Sulfur is Ok to still give him. We used to cut it with Kombucha, but that is now a no-no.

Thanks and Huzzah!

86. *Ryan is the first of our 3 children to recover. Ryan was not considered autistic by most standards but did score a 19 on the ATEC test prior to starting the CD protocol. He is 9 years old and has been in mainstream classes without an IEP. We did hold him and his twin brother back one year prior to starting kindergarten so they could have more time to heal from Lyme Disease which we had been treating since they were 4.*

Ryan was a healthy baby but I did notice after saying his first words at the appropriate age, he stopped talking a few months later. (I now know that is a symptom of autism!) And both he and his brother were very late with language despite knowing their alphabet at a very young age. Ryan did eventually talk but was often hyperactive and had a hard time sitting still in class or at the dinner table. It was often hard to get his attention and to get him to listen to requests. He was constipated since he was a baby and was a very picky eater. I dreaded mealtimes because it was so hard to get him to eat. We often had to spoon-feed food into his mouth even when he was 7 years old. Treatment for Lyme Disease helped many of his symptoms but he still had some remaining symptoms such as weakness, stomachaches, seeing orange and blue spots at night (more noticeable in the dark), nightmares, bedwetting, difficulty falling asleep and early awakening, and overly emotional and irrational behavior. These symptoms worsened dramatically when our house well water became contaminated with parasites and bacteria. It took us two years to figure out that the water was the cause of our deteriorating health. During this time, Ryan became unable to play sports or have play dates without melting into irrational tantrums when things didn't go his way. When playing mini golf, he refused to finish the round after getting a score of "10" on one hole and had a crying tantrum. In tennis class, he would sit down on the court and start crying if he didn't get to hit as many balls as the other children on the court with him. We didn't even attempt to put him in team sports for fear that he would melt down and storm off, refusing to play the rest of the game.

Then we discovered CD and Kerri's protocol, and his and our lives changed forever. Within three weeks of taking CD, Ryan's ATEC dropped to an 8. This improvement was the difference in Ryan being able to play sports. He was able to play on a baseball team and while he did get angry if he struck out or made an error, I was there with a dose of CD and that would keep him from a complete meltdown. As the season continued, Ryan's emotions became more and more stable and normal. His energy increased greatly and now after 6 months on the protocol, Ryan is playing three different sports, taking piano lessons and excelling in above grade level reading and math. His ATEC is a 3 but we will continue until it is 0. His first report card since starting the protocol was the best of his life. We were most excited to see that his scores for behavior and attention were above expectations. Last year, these same scores were all below expectations because Ryan was so hyperactive and unable to focus due to the parasites. Ryan still sees a few orange spots at night but they are diminishing and we know they will be gone so long as we stick with this protocol until he is completely healed. We are so grateful that Kerri blazed the way for us all by developing this protocol. I have been searching for the past 7 years for a way to fully heal my children and the CD was the last but most significant piece of the puzzle for us. We will be forever thankful to you Kerri!

87. *I first heard about CD, then called MMS, to treat autism in May 2012 and wanted to try it out with my then 10-year-old son. I bought supplies, and did the then-current protocol for a few months, but then school ended and summer vacations started and I put CD on the back burner or did it very sporadically. My son was verbal and high functioning, but still had lots of autistic behaviors. His ATEC score in 2012 was 27.*

    I started reading about CD again in the summer of 2013. This time there was more information out there. Kerri had written her breakthrough book and lots of facebook friends in the autism community were mentioning it. I bought the book and dove in, first using my old supplies (CD with citric acid) then buying HCl and trying that. My son did very well with the new protocol, starting low and getting to full dose around the time of his 12th birthday at the end of August. I started seeing worms before we even started our first Parasite Protocol! When I started the Parasite Protocol and added in ocean water, things kept getting better! No set-backs or regressions. School was seemingly much easier for my son. He was in a mainstream class with resource room help for reading and writing for the second year, but this year, math homework was no longer taking several hours, but 20 minutes! His reading comprehension was improving as well, and more importantly, I didn't worry about hearing about behavioral issues anymore!

    I wanted to complete three months of CD before I took another ATEC. When I filled in out in November, at the start of our third Parasite Protocol, I was shocked to see a score of 7! My son had dropped 20 points from his 2012 score! Seeing the single digit score was so amazing! My son had been diagnosed with autism at age 2 ½ and there were no options given to me at the time. An early IQ test given by a speech therapist had shown he was mentally retarded and that was that. However, I looked into everything I could, and was not going to give up on my son. So glad I didn't! Nine years of autism, but the beginning of his 12th year marks the end of it! I know his ATEC score will be 0 soon!

---

88. *"thank you for your help in recovering me"*

    This is what my son just typed to me this morning on his iPad with the app we use for supportive typing!

    Thank you Kerri & Patrick for all you have done and continue to do!

    Without you BOTH, we would NEVER be in the place we are right now... HEALING & RECOVERY!

    Just know you are LOVED!

    —Melissa in Indiana

---

89. *Sophia was born on March 9, 2006. She was a typical happy baby who slept well and seldom got sick. She met all of her developmental milestones and laughed and made great eye contact. When she turned one just after her vaccines she became violently ill and vomited with fever for 2 days straight. After that she was never the same, her smile was rare and words were gone. When she was nineteen months after more shots, we took her to an ear doctor out of concern that she was not responding to her name or speaking clear words. Her hearing was tested and we were told she was fine and to give her time. After a few months of no change her pediatrician determined that she was developmentally-*

delayed and had all the symptoms of Autism. We were stunned and unprepared for this news. I went into warrior mode and ended up with a DAN dr, who started her on lots of supplements after initial lab results showed heavy metal toxicity and high amounts of intestinal yeast. We followed the program and advice of this doctor for the last 4 years with the only change being dietary fluctuations. Sophia did improve but hit a plateau where no new gains were made in the last year or so, so in May 2013 her DAN doctor suggested that we do another round of stool testing after a urine acid test showed bad bacteria overgrowth and terrible mito issues. Her comment to me was "this kid's DNA is so messed up, I haven't seen a test this bad since the 1990's." This was after thousands of dollars and too many supplements to count.I was done at that point. There were no answers there obviously. I came across this protocol on a Facebook page for autism parents and purchased Kerri's book in September 2013. I immediately knew that this was the answer for my daughter. I had a gut instinct that she would be healed . We started the diet immediately and CD just after. My daughter had been a juice addict before this and I was worried about her taking the CD and being able to drink the water. This was only an issue for a day or so. I have not missed a day of giving her 8 or more doses. We get up very early and dose her ocean water and 2 CD doses before school. On weekends we are doing 16 doses/day. In the second week we started seeing some herxheimer/die off, very sleepy followed by fevers. This is when the transformation started! I followed the fever protocol and did enemas and she got better. She also had an episode around week 7 of very red throat and fever which we double dosed for and resolved. We have seen a withdrawn, self -focused child with little interest in the world outside of herself disappear and turn into a child interested in friendship with peers, reading books and understanding the story (huge!), and having relevant conversation with great eye contact. She gets all A's at school and has made a complete turnaround with behavior. We have so much hope for her future and she is blossoming into a beautiful, happy girl!

---

90. *Here are a few of the emails I mentioned from my son's therapist in regard to what's happening at therapy! :)*

This is an email from the exact day I restarted the CD, which was August 21 of 2013.

We discovered a new game today. He was being a turkey and running up the hall to try to get around me to get places he wasn't supposed to go. I started singing (to the tune of rolling down the river) about whatever he was doing "walking, walking, walking down the hallway" "jumping, jumping jumping down the hallway" climbing, climbing, climbing up the ladder. He thought it was hilarious, especially when I'd change in the middle of a verse "running running, jumping down the hallway.

—Your son's therapist

Great day today! Vail seems to be understanding first, then sequencing and I think I finally have a bribe for him--he loves our fish.

Also, he made a friend today. One of our other pt's was playing in the gym and he was watching him, imitating him a little and said his name 2x!

—Your son's therapist

We tried to play the game and he wasn't interested. Instead we ended up *pushing the fisher price horse, sheep, goat, and cow off the treehouse. He would "randomly" produce sounds which we took for requests for the animal associated. Then we would give him the animal "requested." Reviewed his sound cards and he produced the /n/ and /h/. Coulda sworn he said go a couple of times, and mom said she heard a whispered "home." Listen and be watchful for whispers! Oftentimes children with apraxia can get one system working but not the other, ie. articulate with no sound, or vocalize with intonation and no articulation, like when he says "ah uh oo" for "I love you." Yep, he does that!*

Cheers!

---

91. *Hi had to share this moment today, just wish daddy wouldn't have missed it!*

Well, we were seeing daddy off for work this evening, when Vail decided that he wanted to go run around in the grass. I told him we needed to put on socks and shoes and then we could play outside. He walked straight to the door, we walked in, I pointed at his socks and shoes, and FOR THE FIRST TIME EVER (always knew he could, just to stubborn to do it unless he really wants too!) he promptly walked over, picked up his socks and shoes and turned and handed them to me!!! (all the while doing it with an attitude!!!"lol)

This is Just from MOM to therapists. :)

CD and Kerri have been life changing for our entire family. We now can give Vail choices, usually of two items. For example I showed him two pairs of pants and asked which he wanted to wear, and he grabbed one pair and started to put them on even though he was still in his pj's. I showed him two shirts, and he picked a shirt. Same thing with his shoes and socks, he is able to choose now on his own. I look forward everyday to him doing something differently or completely new, and he almost always does now!!! This was not possible for CD!!! THANK YOU KERRI !!! WE LOVE YOU FOR SHARING AND FIGHTING THE FIGHT!!!

LOVE,

—The parents of Vail

---

92. *Best day of the entire school year! My son initiated verbal interactions 5 times today and responded to his teacher in a group over 5 times, WITHOUT A SINGLE PROMPT! Most days his TA is poking him to pay attention and ask other kids for directions, rather than an adult, all day long! She is tracking these interactions as they are his IEP goals.*

---

93. *Casey is a young man who is 17 years old today, and the estimated ATEC score from 2006 is from when he was 10 years old. Proving once again that recovery is possible at any age.*

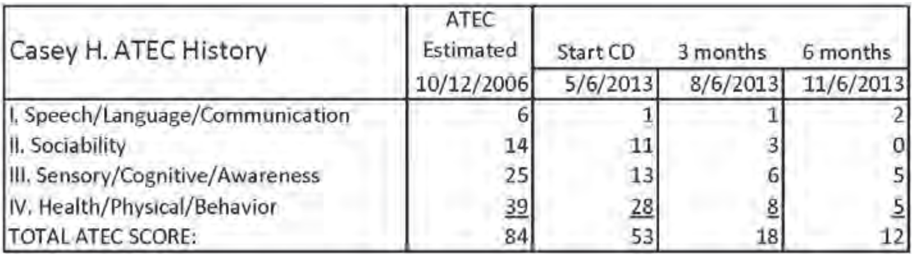

Casey H. ATEC History	ATEC Estimated	Start CD	3 months	6 months
	10/12/2006	5/6/2013	8/6/2013	11/6/2013
I. Speech/Language/Communication	6	1	1	2
II. Sociability	14	11	3	0
III. Sensory/Cognitive/Awareness	25	13	6	5
IV. Health/Physical/Behavior	39	28	8	5
TOTAL ATEC SCORE:	84	53	18	12

94. *Daniel - beginning ATEC June 2013 - 101. Today ATEC Nov 12, 2013 - 44! We are thru our 3rd PP and starting speech supplements. Speech, by far, has been the hardest thing. In all other areas he is under "10" on the ATEC! Kerri Rivera words are coming. This week alone he is saying "Hi" and "Bye" and trying to say others. I can't wait to hear "I love you Mom"! You are the Bomb Diggitty Kerri. We all love you!*

   UPDATE: December 3rd, 2013

   I officially want to tell the world that my son, Daniel, who has autism is nonverbal no more! He is starting to talk. Praise God! This battle has been the struggle of my life, but i have learned to trust God, be patient, and enjoy the journey! Where some take a lot for granted regarding their children, I have been waiting almost 7 yrs to hear his beautiful voice! And thank God he has one! Autism is treatable and there is hope!

---

95. *We had a super exciting trip, we went to a therapist on the 28th who teaches Rapid Prompting Method. We have tried since by daughter was diagnosed to get her to speak, but haven't had any luck yet, she is now 11. We thought we would give RPM a try to see if the therapist could get her to communicate.*

   We not only found out that she is able to make correct choices, but can actually SPELL! Her very first word that she spelled was CLOUD!

   Later on in the session the therapist asked her what her favorite color was... she spelled out PINK!

   I was sitting on the floor in awe... For the first time in 11 years, I finally know what my girl's favorite color is... it was surreal... it was amazing... she is so smart, but had no way of letting us know... I am soooooooo proud of her!

---

96. *Diagnosed at 20 months old in August 2011*

   Started supplements under Dan in September 2011

   Started CD / PP on July 1, 2013, after seeing rope in stool

   Starting ATEC on July 1, 2013, was 46

   Present ATEC 22

   Gained 5 pounds from July 1st to present!

   Too many improvements to list!!!!!

   Also doing GcMAF which has also helped……but, I believe GcMAF may have helped more with the viral load. I suspect this because we actually started IV GcMAF in January 2013, before CD . Huge gains with GcMAF but after 35 shots, her gut was still a mess and GcMAF definitely didn't help with the gut. Found Kerri in July 2013, and began our healing journey.

The following lab results are proof of just how well the child described above is doing since starting the protocol. Here is what her mom had to say:

> *"The original test was taken last August of 2012, which showed a deficiency in almost everything despite the fact that my daughter had been on supplements for over a year at that point. The re-test was just taken October 2013, only 4 months after CD / PP and removal of ALL supplements. Wow!"*
>
> *You can see in the second test that her high viral titers were reduced as well.*

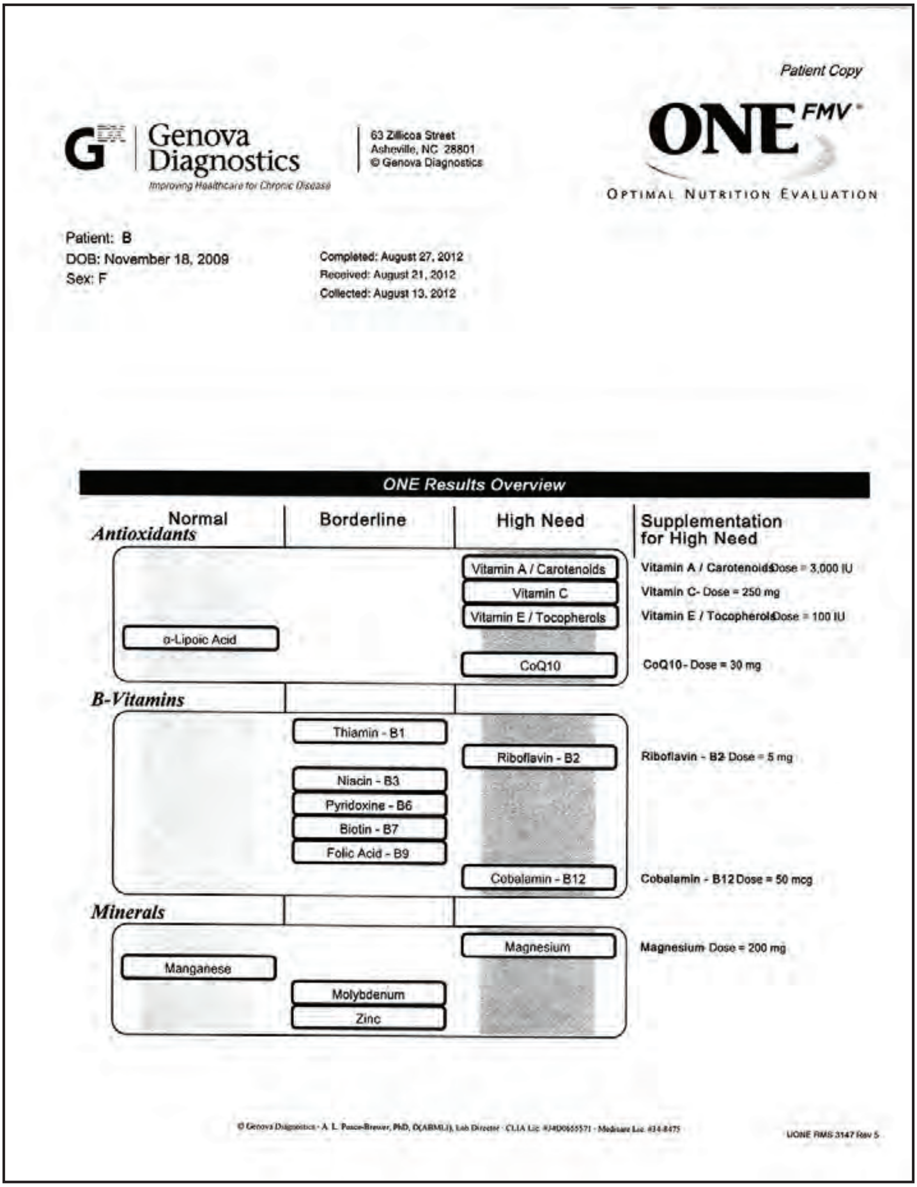

Patient Copy

ONE FMV®

OPTIMAL NUTRITION EVALUATION

Genova Diagnostics
Improving Healthcare for Chronic Disease

63 Zillicoa Street
Asheville, NC 28801
© Genova Diagnostics

Patient: B
DOB: November 18, 2009
Sex: F

Completed: August 27, 2012
Received: August 21, 2012
Collected: August 13, 2012

ONE Results Overview

	Normal	Borderline	High Need	Supplementation for High Need
Antioxidants				
			Vitamin A / Carotenoids	Vitamin A / Carotenoids Dose = 3,000 IU
			Vitamin C	Vitamin C- Dose = 250 mg
			Vitamin E / Tocopherols	Vitamin E / Tocopherols Dose = 100 IU
	α-Lipoic Acid			
			CoQ10	CoQ10 - Dose = 30 mg
B-Vitamins				
		Thiamin - B1		
			Riboflavin - B2	Riboflavin - B2 Dose = 5 mg
		Niacin - B3		
		Pyridoxine - B6		
		Biotin - B7		
		Folic Acid - B9		
			Cobalamin - B12	Cobalamin - B12 Dose = 50 mcg
Minerals				
			Magnesium	Magnesium Dose = 200 mg
	Manganese			
		Molybdenum		
		Zinc		

© Genova Diagnostics · A. L. Peace-Brewer, PhD, D(ABMLI), Lab Director · CLIA Lic #34D0655571 · Medicare Lic #34-8475

UONE RMS 3147 Rev 5

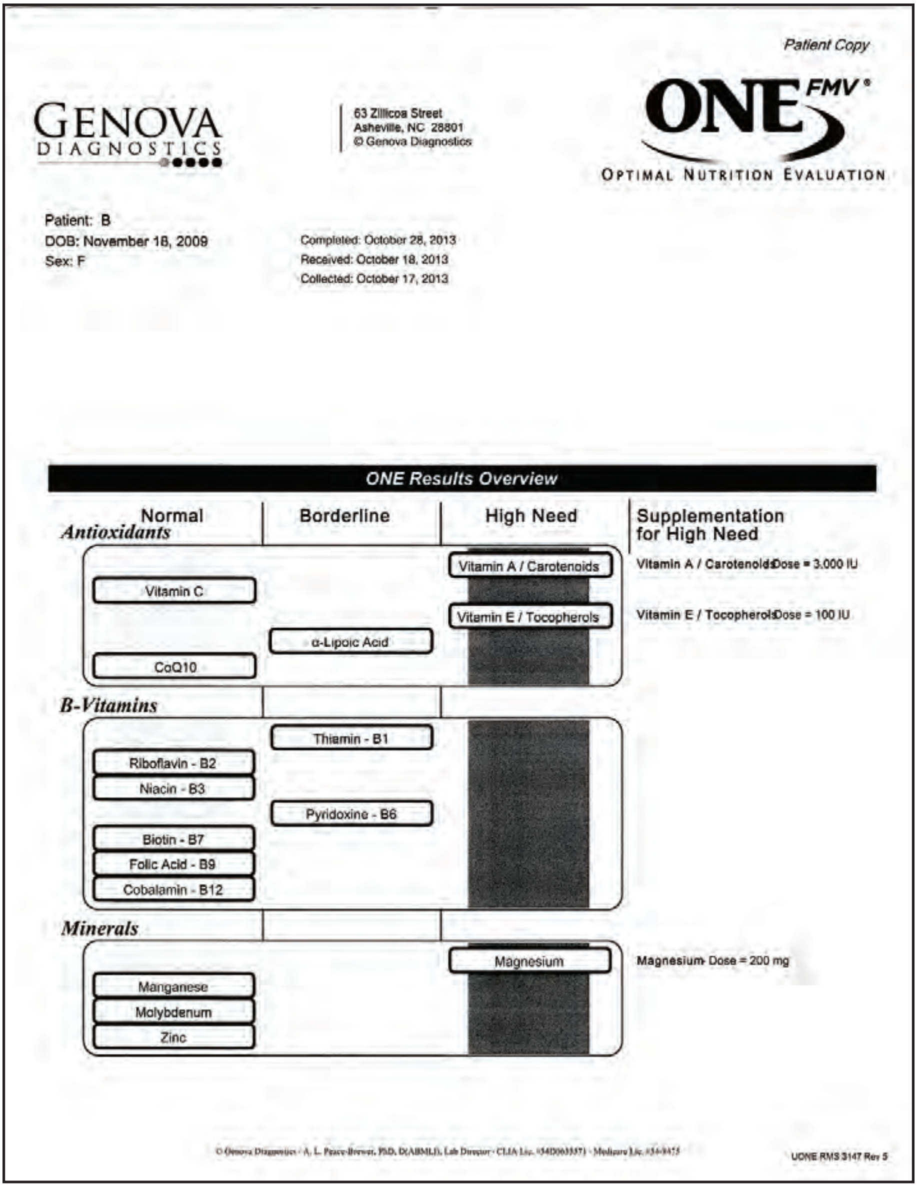

Patient Copy

GENOVA
DIAGNOSTICS

63 Zillicoa Street
Asheville, NC 28801
© Genova Diagnostics

ONE FMV®
OPTIMAL NUTRITION EVALUATION

Patient: B
DOB: November 18, 2009
Sex: F

Completed: October 28, 2013
Received: October 18, 2013
Collected: October 17, 2013

ONE Results Overview

	Normal	Borderline	High Need	Supplementation for High Need
Antioxidants				
			Vitamin A / Carotenoids	Vitamin A / Carotenoids Dose = 3,000 IU
	Vitamin C			
			Vitamin E / Tocopherols	Vitamin E / Tocopherols Dose = 100 IU
		α-Lipoic Acid		
	CoQ10			
B-Vitamins				
		Thiamin - B1		
	Riboflavin - B2			
	Niacin - B3			
		Pyridoxine - B6		
	Biotin - B7			
	Folic Acid - B9			
	Cobalamin - B12			
Minerals				
			Magnesium	Magnesium Dose = 200 mg
	Manganese			
	Molybdenum			
	Zinc			

© Genova Diagnostics · A. L. Peace-Brewer, PhD, D(ABMLI), Lab Director · CLIA Lic. #34D0655571 · Medicare Lic. #34-8475

UONE RMS 3147 Rev 5

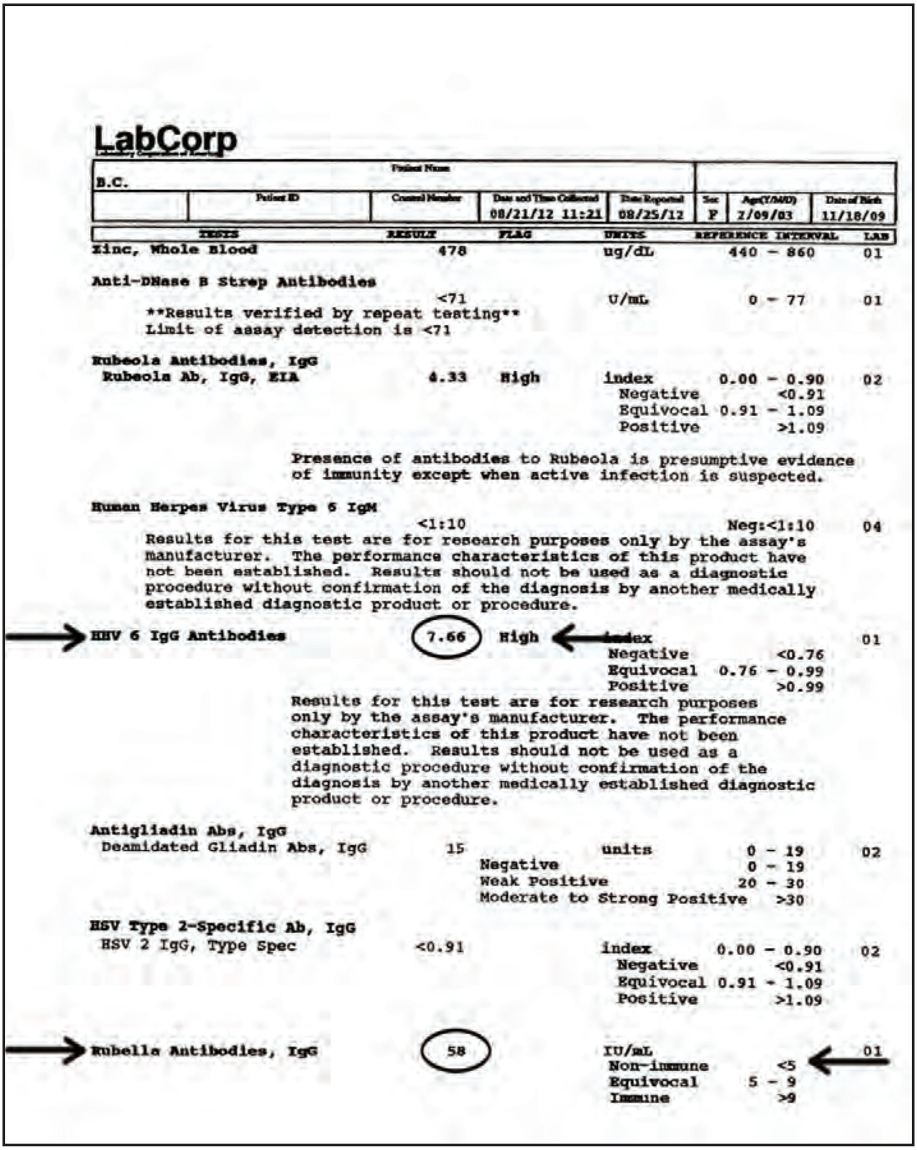

LabCorp
Laboratory Corporation of America

Patient Name							
B.C.							
	Patient ID	Control Number	Date and Time Collected	Date Reported	Sex	Age(Y/M/D)	Date of Birth
			08/21/12 11:21	08/25/12	F	2/09/03	11/18/09

TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL	LAB
Zinc, Whole Blood	478		ug/dL	440 - 860	01
Anti-DNase B Strep Antibodies					
	<71		U/mL	0 - 77	01
Results verified by repeat testing Limit of assay detection is <71					
Rubeola Antibodies, IgG					
Rubeola Ab, IgG, EIA	4.33	High	index	0.00 - 0.90	02
			Negative	<0.91	
			Equivocal	0.91 - 1.09	
			Positive	>1.09	
Presence of antibodies to Rubeola is presumptive evidence of immunity except when active infection is suspected.					
Human Herpes Virus Type 6 IgM					
	<1:10			Neg:<1:10	04
Results for this test are for research purposes only by the assay's manufacturer. The performance characteristics of this product have not been established. Results should not be used as a diagnostic procedure without confirmation of the diagnosis by another medically established diagnostic product or procedure.					
HHV 6 IgG Antibodies	7.66	High	index		01
			Negative	<0.76	
			Equivocal	0.76 - 0.99	
			Positive	>0.99	
Results for this test are for research purposes only by the assay's manufacturer. The performance characteristics of this product have not been established. Results should not be used as a diagnostic procedure without confirmation of the diagnosis by another medically established diagnostic product or procedure.					
Antigliadin Abs, IgG					
Deamidated Gliadin Abs, IgG	15		units	0 - 19	02
		Negative		0 - 19	
		Weak Positive		20 - 30	
		Moderate to Strong Positive		>30	
HSV Type 2-Specific Ab, IgG					
HSV 2 IgG, Type Spec	<0.91		index	0.00 - 0.90	02
			Negative	<0.91	
			Equivocal	0.91 - 1.09	
			Positive	>1.09	
Rubella Antibodies, IgG	58		IU/mL		01
			Non-immune	<5	
			Equivocal	5 - 9	
			Immune	>9	

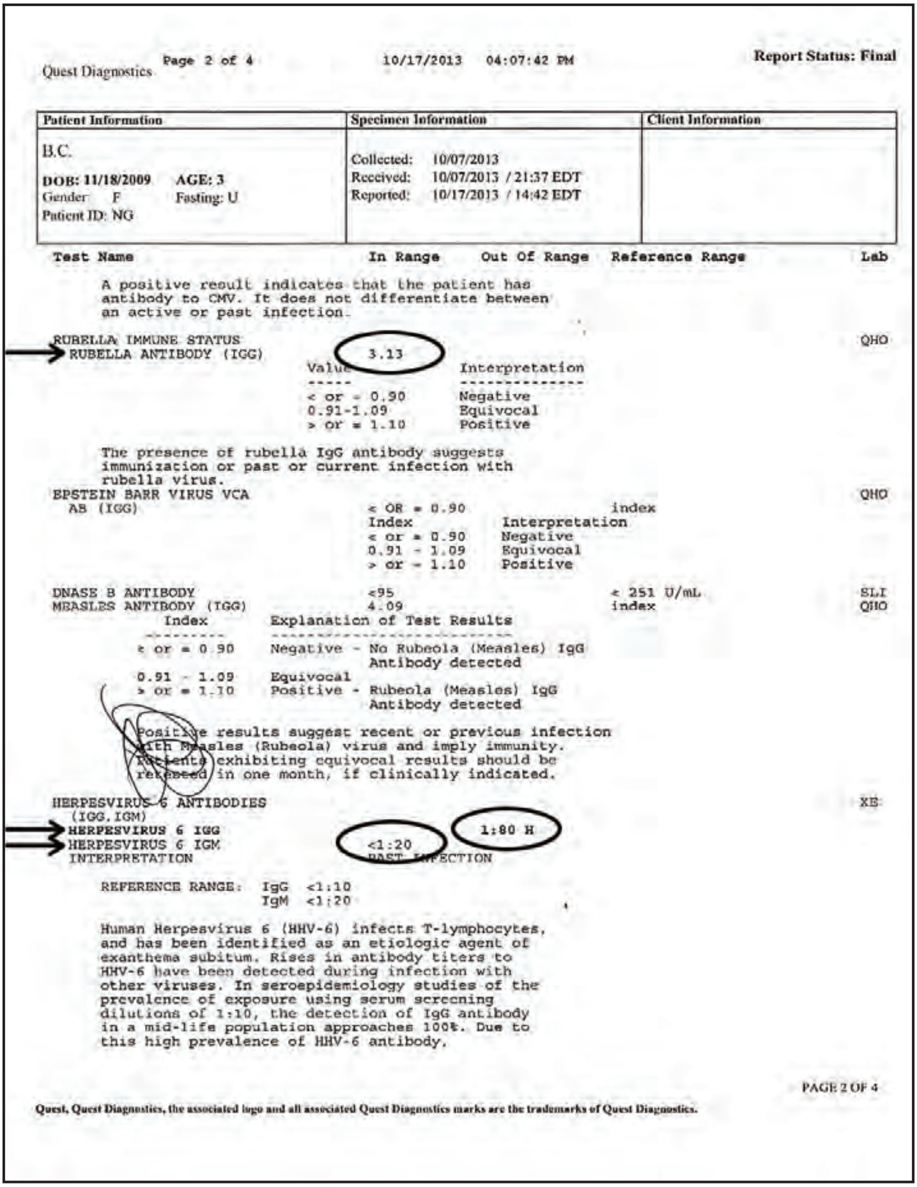

Quest Diagnostics Page 2 of 4 10/17/2013 04:07:42 PM **Report Status: Final**

Patient Information	Specimen Information	Client Information
B.C. **DOB: 11/18/2009** **AGE: 3** Gender: F Fasting: U Patient ID: NG	Collected: 10/07/2013 Received: 10/07/2013 / 21:37 EDT Reported: 10/17/2013 / 14:42 EDT	

```
Test Name                           In Range    Out Of Range   Reference Range         Lab

       A positive result indicates that the patient has
       antibody to CMV. It does not differentiate between
       an active or past infection.

RUBELLA IMMUNE STATUS                                                                  QHO
  RUBELLA ANTIBODY (IGG)              3.13
                            Value              Interpretation
                            -----              --------------
                            < or = 0.90        Negative
                            0.91-1.09          Equivocal
                            > or = 1.10        Positive

       The presence of rubella IgG antibody suggests
       immunization or past or current infection with
       rubella virus.
EPSTEIN BARR VIRUS VCA                                                                 QHO
  AB (IGG)                          < OR = 0.90               index
                                    Index        Interpretation
                                    < or = 0.90  Negative
                                    0.91 - 1.09  Equivocal
                                    > or = 1.10  Positive

DNASE B ANTIBODY                    <95                       < 251 U/mL               SLI
MEASLES ANTIBODY (IGG)              4.09                      index                    QHO
            Index        Explanation of Test Results
          ---------      ----------------------------
          < or = 0.90    Negative - No Rubeola (Measles) IgG
                                    Antibody detected
          0.91 - 1.09    Equivocal
          > or = 1.10    Positive - Rubeola (Measles) IgG
                                    Antibody detected

          Positive results suggest recent or previous infection
          with Measles (Rubeola) virus and imply immunity.
          Patients exhibiting equivocal results should be
          retested in one month, if clinically indicated.

HERPESVIRUS 6 ANTIBODIES                                                               XE:
  (IGG,IGM)
  HERPESVIRUS 6 IGG                             1:80 H
  HERPESVIRUS 6 IGM                 <1:20
  INTERPRETATION                    PAST INFECTION

       REFERENCE RANGE:  IgG  <1:10
                         IgM  <1:20

       Human Herpesvirus 6 (HHV-6) infects T-lymphocytes,
       and has been identified as an etiologic agent of
       exanthema subitum. Rises in antibody titers to
       HHV-6 have been detected during infection with
       other viruses. In seroepidemiology studies of the
       prevalence of exposure using serum screening
       dilutions of 1:10, the detection of IgG antibody
       in a mid-life population approaches 100%. Due to
       this high prevalence of HHV-6 antibody,
```

PAGE 2 OF 4

Quest, Quest Diagnostics, the associated logo and all associated Quest Diagnostics marks are the trademarks of Quest Diagnostics.

Testimonials from the First Edition Continue Here...

97. *Meeting a Rescue Angel*

The way OUT of AUTISM

My 5 year old little daughter, Anna's story has been long and paved with lots of suffering and doubt, compared to her age, but prayers have always helped us in our journey. We are aware that it may take years to get out of the dark depths of Autism but now we know we are going in the right direction. Finally! Thanks to Kerri Rivera, the Rescue Angel of so many suffering children. She showed us the way out!

We have wasted so much time by trying to face the fact, that unfortunately there is a problem with our perfectly healthy-born daughter. After the consultation with many doctors, they first suspected a hearing problem. As the result of this, she went under many surgeries and antibiotic-treatments, which made her state worse mentally and physically as well. The medical professionals could not find any evidence of physical abnormalities, so they choose the easy way out, the one word diagnosis of Autism! Dear Mom, please accept it, you cannot do anything about it. Try some developmental therapies, but do not put your hopes high!

It is very hard to accept the unexplainable. What does autism mean? If you google it, you will find the verdict immediately: "Autism is a lifelong developmental disability characterized by impaired social interaction and communication, and by restricted and repetitive behavior" How is it possible that a perfectly developing child changes overnight? She stops saying the words she said before; she stops paying attention when called by her name and simply forgets some of her skills she knew so well before. How can one accept that? For us, it was impossible. We took up the gauntlet!

We spent so many sleepless nights on the Internet searching for solutions, to be shortly shocked to learn that in the US, there is a totally different concept and approach to autism. A group of medical professionals and scientists are proving the crucial role of environment and vaccines in the rapid rise of autism, accepting genetics as not the single cause but a factor of susceptibility to the disease. More and more so called DAN (Defeat Autism Now) clinics and practitioners are treating children on the spectrum all over the world with different biomedical interventions with various success rates. Whichever method one chooses in their realm, the base line to all is a strict diet, especially free from gluten, casein and sugar. There is no way out without eliminating these from the diet of an autistic child.

We immediately changed her diet from a bread and butter with chocolate milk style to a much stricter one. The first items to remove were sugar, gluten and casein, but these were shortly followed by potato, rice, corn and all artificial additives. We slowly changed to SCD (Specific Carbohydrate Diet). A few days into the diet and Anna slept through the night, and what's more, the painful eczema spots that had been on her face since she was a baby have shortly disappeared. Parallel with the diet modifications we also started the DAN interventions. We spent a fortune on lab testing, supplements and to see a DAN practitioner in Germany. Unfortunately none of these brought us significant improvements.

But I knew and believed that there has to be a solution.

And yes! I cannot keep it to myself and I would like to spread the word happily, powerfully and self-confidently, that there is a solution! This year in August, I have found the presentation of Kerri Rivera on the internet. Autism is preventable, avoidable, treatable and curable."

At the beginning "Kerri Rivera rose against the wind rather than with it when, after a frustrating slew of hits and misses, her youngest son, Patrick, was diagnosed with autism nine years ago. He was two at the time. Surrounded by literally no local expertise or any type of peer support circle, she did what most mothers would do: She became informed and involved in the biomedical community. It was through the Autism Research Institute (www.autism.com), a non-profit organization dedicated to autism research, diagnosis and treatment for the past 40 years, that she learned about a specialized biomedical protocol involving a special diet and supplements, physical therapy, applied behavioral analysis, chelation and hyperbaric treatment. There was light for Patrick, Kerri's son at the end of the tunnel."

But Kerri did not stop here! As she told me, she felt that there was a missing piece to the autism puzzle. A huge piece! It shortly turned out to be the MMS piece. She developed MMS protocol which can be done at home. Following Kerri's protocol, 68 children lost their autism diagnosis in the past two years! This result is outstanding, considering the fact that this is said to be an incurable disease.

The day I saw her presentation I e-mailed her and I got an answer within a half an hour. Kerri Rivera's MMS protocol is the missing piece of the puzzle to cure autism, ADHD, asthma, allergies, seizure disorders. Even mothers of children with Down Syndrome have reported improvements with this protocol.

My whole family was in Mexico in August. We spent two months with the person, who is also a mother of a child with autism, who sacrificed part of her life to help others. For free, out of faith, humanity and love. Approximately 8-900 children are following Kerri's protocol currently all over the world. The protocol is available for everybody and can be done at home (www.mmsautism.com).

Anna is getting better every day. Although we have heard very promising sounds from her, speech is not developed yet. However she has improved a lot in the last 3 months in many other areas like mood, communication, learning ability, focus and attention span are all better and she started pointing. There were no signs of these at the beginning of the summer.

We are fully aware that there is still a long journey ahead of us, but now we know we are on the right track. Thanks to God and Kerri Rivera to guide us to recovery!

---

98. Hi Kerri,

PLEASE use our testimony. MMS kept my daughter's PANDAS under control after switching from Silver Hydrosol BUT the biggest is the PP's (parasite protocols) in combo with MMS. She began sleeping through the night, THEN, sleeping in her own room after sleeping in our room for 9 months. THEN she began putting herself to sleep for the first time in 7 years. We always laid with her--her WHOLE life--I KNEW it was NOT "behavorial", knew it was due to *something going on interally, AND NOW, we read, kiss, and WALK OUT of her room! It's SURREAL really. Seriously SURREAL!*

Also, she was a chronic chewer, nose picker. Almost GONE! Peaks at new and full moon, BUT goes away. She was chewing about 20 straws a day before PP! Mangling them! UNREAL! I have pics of the straws... cant even recognize them as a "straw" once she was done with them. CLEARLY parasites. CLEARLY on the right track. NO other interventions. Except for how the MMS PP protocol is outlined MINUS Rhompe etc.. and Neem. NO sleep intervention supps etc.. None.

---

99. *Favy is my God sent angel. He is 4 years old and since he was born he came to change lives, to make families stronger and to help others forgive. As many of your stories, everything was "perfect" (besides of milk intoleration, delay speech and sleeping problems) the rest was great, milestones on time.. Then gradually after 13to 18 month something was clearly wrong, we loose him and then the bomb!: diagnosed in Puerto Rico by a developmental pediatrician and a neurologist as PDD-NOS on December 6, 2010.*

We moved fast and start the diet GFCFSFYF no sugars.. Etc Thanks to CEA non- profit organization who help parents with kids with autism. Soon we saw positive changes, then probiotics and a natural anti fungal. Difficult days... But always seeing little good things. On march 2011 I met an angel: Kerri Rivera and we start MMS protocol!!! Then we start to see the light! We moved to FL and they diagnosed him (now I know he was having a die off when reevaluate) he was diagnosed as Autism. For me , not a big differences, still in the spectrum and he still sick!!! We need to cure him!!!

After MMS the big changes came ,and still coming!!

first from 2 to 4 hours straight per night to 6to 9 straight during night plus a nap during day!!!!! Eye contact was much better and he had less tantrums. He learned fast 3 signs to communicate. He was more happy. I didn't see more great results because I was using Mms incorrectly for almost a year (with juice, antioxidants and vit c) When a couple of months ago I started doing it right (just water and no vit c nor antioxidants etc) he started to improve again and better.. After a lot of efforts, dedication, patience and hours of hard work, with Kerri Rivera MMS protocol Favy to improved in many areas.

This is all he have improved:

Educational:

* *Respond to his name when called and say his name when ask.*
* *Recognize people: Mommy, Daddy, Sister, Teachers, Therapists, friends*
* *Name and identify some body parts: eyes, nose, mouth, ears, belly, feet*
* *Recognize some animals (Ipad application or tv) : cow, horse, duck, chicken, zebra turtle.*
* *Recognize some colors: blue, yellow, purple, red.*
* *Recognize and count 1-10 (can repeat 11, 13)*
* *Recognize and/ or use some common objects: plate, spoon, fork, bottle, chair, pants, shirt, shoes, backpack, lunch box, blanket, pillow, tv, books, some toys, ball, cars.*
* *Stack objects like blocks, books*

- *Match pictures, objects or things.*
- *Color independently but refuses, he can trace but hand over hand.*

Social / Emotional:

- *He enjoyed for the first time his birthday Party, he was happy! He opened his presents and show emotions!!!*
- *He is beginning to interact more with others.*
- *He can follow direction of 1 step and some cases of 2. (clean up your plate to the sink and push your chair.)*
- *He is affectionate : kisses, hugs, eye contact is better.*
- *Wave and say bye.*
- *Greet with a good morning or good night.*
- *Play around and with others: chase, tickles, peek-a-boo,*
- *Dancing and imitating some song rituals like: The wheels on the bus, If you are happy, head shoulders knees and toes, and some in Spanish too.*
- *Throw, push, blow, pull, jump, run, walk, go up and down stairs, climb, open, close, zip, unzip,*
- *He can watch an ABC DVD seated or lay down , dancing or repeating.*
- *He can be seated for almost 5 to 8 minutes with his favorite books.*
- *He enjoys to go to playgrounds, go up and down slide or swing.*
- *He enjoyed the trick or treating this year, going house to house picking candy and wore all night his costume.*

Comunication:

- *I think this is the area of more improvement. From 0 to almost 35 words and different ways to label, request, interact and communicate!*
- *He can answer questions with yes or no (yeah- nah)*
- *He signs (modified) more, all done, juice*
- *He can use 3 words at a time (mommy water please, mommy come here, I love you, mommy ipad please, mommy hungry, eat)*
- *Say independently—not all of the sound clear but very similar sounds.*
- *Mommy*
- *Dadda*
- *On*
- *Yes(yeah)*
- *No (nah)*
- *Please (peas)*
- *All done (ah dah)*
- *More (mor)*
- *Water (at-ter)*
- *go*
- *Four*
- *Three*
- *Tickles*

- *Zebra (eebah)*
- *Turtle (totlo)*
- *Uh oh!*
- *Oh No!*
- *I Love You!*
- *Under*

When Prompting or Repeat:

- *Tv (tetete)*
- *Dvd (dededah)*
- *Off*
- *Come (she)*
- *Itch (each)*
- *Scratch (ach)*
- *Bubbles (babah)*
- *Bye (bah)*
- *Swing (win)*
- *Push (pah-sh)*
- *Five (hi)*
- *Nine (nun)*
- *Ten (ttn)*
- *Baby*
- *Ble (buh)*
- *Purple (por-po)*
- *Yellow (ye-yo)*
- *Chase (ch sound)*
- *Play (ay or pah)*
- *He can imitate some animal sounds like tiger, horse and wolf.*
- *He make choices between to objects . (playing, dressing, eating)*
- *Independent Functioning (he needs some prompting)*
- *He eats independently with spoon and clean up after finished.*
- *He can drink of a regular cup.*
- *With pictures or social stories he can transition to different activities.*
- *He is showing understanding of routines and rules. (dvd or toy stay home, 2 minutes and we are all-done)*
- *He can pull up and down his pants (multiple prompt)*
- *He can put his shirt on but need assistance to take it off.*
- *He can pick up and put on hamper his clothes.*
- *He can put his shoes (crocs) independently and cooperates with other type.*
- *He move furniture to reach what he wants.*

Health:

Favy still in the Gluten, Casein, Soy, Yeast, Antibiotics and Hormones free diet. He is trying new textures, flavors and consistency in a variety of food allowed.

He is drinking more water. MMS help us every time he have a cold, a mosquito bite aor nasal allergies. Is the best product we can ever find!

We are excited about all he learned, do and know during this last year because of the use of the protocol of MMS. We are proud of his efforts, but we still have areas to improve. We will continue with MMS, diet and wil start the parasite protocol with faith because MMS Rocks!!!!

We love you Godmother Kerri! God Bless you and your Family.

---

100. Conversation at pick up:

Aide: "What did you spike his breakfast with this morning? I've never seen him like this!"

*Me: *smile* "Oh glad he had a good day"...in my head...'three doses of MMS before drop off'...*

This child is going to figure out he has wings and start to SSSSOOOOAAAARRR!!!!

---

101. I have a 4-year-old son on the autism spectrum and MMS has been the most effective treatment we have tried for him. He had bad eczema on his legs for over a year that I was never able to get rid of. We tried diet, medication, cream and nothing seemed to clear it up. Within 2 weeks of starting MMS it was completely gone and has not returned. My son has made great strides in the past 5 months of being on MMS and the parasite protocol. He is more interactive with all the members of the family, talking more, and becoming more successful with his OT goals.

---

102. Kerri; Thank you all for your great job. Because through the Webinar—Autism and MMS—the information has come to me and gave me the strength to start a treatment.

---

103. Hi Kerri,

My wife and I did an ATEC on 8 year old son and it came in at 7. Wow.

We're learning so much from you about how this whole recovery thing works. The child's health takes three steps forward then two steps back then three steps forward then two steps back...but over time that one extra step forward seems to slowly add up. His pee is smelly again but, as you taught us, it's pee-tox and this is a good thing. His focus is still a bit of a challenge at school and he still won't shut up :-) but we are feeling that he truly is slowing getting better.

We've got a ways to go yet but we feel that we are truly moving forward. Thanks so much for all of the help you've given us and any help you might be able to give in the future.

---

104. My son is 19 years old. We started MMS in July. For 2 months my son did not sleep. The end of August at full oral dose we started the baths. At night you could see he was uncomfortable. I don't know how we did it but we talked him into doing enemas. My husband does it. They induce him to have a bowel movement.

Before enemas my son would have nights where you heard him walking back and forth from his bed to the toilet. In the morning I would say you had a bad night *last night, what was your problem? He would spell constipation. He doesn't seem to be doing this since the enemas.*

On Saturday 11-10 our son went to the first day of the season Special needs Ice Hockey. He is just learning. There were no problems with him about anything. In fact for the first time he put on his hockey gloves without objecting, which we have been trying to do the last 2 years. He followed directions from a volunteer on the ice, using his hockey stick to hit the puck back and forth to him and towards the goal. When hockey was over we went outside and hung out at the car and did nothing for about 20 minutes and had a dose of MMS. He did not care one bit. He entertained himself by watching the sea gulls. He was not impatient.

Next we walked over to the bowling alley, which is next door to the ice rink. He participated in special needs bowling. They have the bumpers up so the kids don't get gutter balls. My son actually threw some balls that didn't hit the bumpers. There were times in the past where every ball he threw hit the bumpers. Again there was not a single problem, incident, or signs of agitation from him while we were there. This is the first time in his life where we were able to go from one activity to another and not have an upset. He enjoyed himself.

What was really interesting was that he usually stims with a container of pumpkin seeds when he is in the car or at the bowling alley. This Saturday he left the container of seeds at home and not once did he even ask for them while we were out.

My son asks for a heating pad, which he has been doing for years. We wonder is it worms that have been causing all this pain for him? We would love to eliminate the pain, the repetitive behaviors, and the prompting through self-help and activities. He does lay around a lot. Everything he does is at a slow pace. He lacks stamina and speed for appropriate activities.

---

105. Hi Kerri, my daughter's first atec was 85. We just did it again and it is now 55! :) :) (approximately 2 months on MMS)

---

106. Hi Kerri,

I just wanted to let you know that we are up to 30 drops per 8 oz bottle and now I'm seeing some new things. He used future tense this morning (he sticks larry boys super suction ear to his window and waits for him to fall) This morning he said "Larry boy will fall." Usually when he falls he says in his high pitched voice "are you ok". Today when Larry Boy fell he said in a deep 11 year old voice "ouch". I just burst out laughing.

---

107. Dear Kerri,

In the spirit of giving thanks, I wanted to say how deeply grateful I am to you and your amazing work with mms and recovering children. I have never felt such hope for my son and such joy sifting through poop!!!

What a truly tremendous gift you have given us all...

Sending love and Thanks

---

108. My son is getting tiny bit more normal every day!!! I could kiss you!!!!!

109. *There was once a day not too long ago when all my son would do is scream and cry. No words at all. No communication. Today he asked me "mommy who holds up the clouds so they don't fall?" :) xoxo love that question! It means so much!! Autism is big. But God is bigger!! Don't give up ladies!! God led us to mms & to Kerri for a reason!!*

---

110. *Kerri is fabulous. She is very loving and supportive and will do anything to help our daughter recover. Using her MMS and parasite protocols have resulted in our eleven year old girl's ATEC scores dropping by half in less than sixth months. She is fully connected and much more responsive and affectionate. She herself is now motivated to succeed! We look forward to full recovery through Kerri's guidance.*

---

111. *I think we are over the initial die off hump!!! We are getting good reports from school about better focus and independence. Also, my son shampooed and combed his hair independently this morning. He even used a Q tip in his ears! I only asked him to try and shampoo by himself. He rewarded me with the combing and Q tip! Awesome mms!*

     Thank you again!

---

112. *I am so grateful to have found you!! MMS here in our house is a MUST and my boy gets better by the day because of it.*

---

113. *6 months ago, my son was on antibiotics for PANDAS, he was suffering with OCD and anxiety, as well as completely distracted - like a zombie.....his teachers came out one day with tears in their eyes because my son said he feels like monsters are inside his head and they are scaring him! They were so worried about him, as we all were!*

     We started MMS in April. My son got up to full dose relatively quickly and we added in MMS baths and MMS enemas and then I pulled the antibiotics—cold turkey! I was emailing Kerri at the time and was afraid to stop the antibiotics. She guided me and gave me the confidence that was needed! As suggested by Kerri, we also did night dosing of MMS. This really helped my son. Keeping a constant flow of MMS was the key.

     My son is flourishing!! He is HAPPY and confident. His focus has increased greatly! Since the start of the new year, he has had great day reports everyday in school! SO many gains in all areas! His coordination and strength have improved - he can do 11 sit ups (In June he couldn't do 1) He was chosen as Student of the Month in his Tae Kwon Do school (he is now a purple belt). The best news to report is his ATEC is now a 10!!!!

     Our family is so grateful to Kerri!!! This protocol has made the BIGGEST impact on my son's recovery!

     Thank you so much Kerri - My son's Fairy Godmother!!!

     Xoxoxo

---

114. *I was praying for a new direction to go in to heal my boy. I literally woke up the next morning with a link to Kerri's presentation in my inbox from a friend whose daughter is making beautiful gains with mms. We are on day 3 here and things are going great, I am so excited! Thank you!!*

---

115. *Yes, my son is making real progress - slow and steady but real progress. The MMS & Parasite Protocol has been very beneficial for him...that's why I am trying to stick to it very closely. Also, I have wanted to tell you that one of the best things that we have ever done for my son is the enemas!!! He is a gut kid with chronic constipation and your enema program has really improved his daily life.*

Thanks again for all of your help and guidance—always truly appreciated!

---

116. *Hi, I'm the mother of Lizbeth Hernandez Saray Miramontes and I want to share my testimony in case it is useful to you.*

My 13-year-old daughter started The Diet in January 2012, in June the full moon protocol to deparasite and withinin two months, she expelled a worm. This helped her regain 8 kilos of weight. (She was 10 kilos underweight when beginning the protocol) Her attitude has changed completely, from being very passive and low energy, she began to smile and be more active.

Thank you and God bless!

---

117. *From the ATEC at www.autism.com*

Total Score: 5

I. Speech/Language/Communication: 1

II. Sociability: 1

III. Sensory/Cognitive Awareness: 2

IV. Health/Physical/Behavior: 1

Started at 56 score in March.

I´m still having problems in environmental awareness, because in his mind sometimes he is Super-Hero (he wants to help to all the people in disasters, his own words). Sometimes he is eating portions like me, not such big portions but he is burning calories in the swimming classes. His meals are so healthy.

So, my homeopathic doctor told me (and recommends me) to wait until January next year to start the PP. He wants to see the effects of MMS, Probiotics and some homeopathy that he will provide at the end of this month.

We continue using as I said MMS, Theralac, GFCFSF diet, NAET and A.B.A therapy. I am so happy to share it with you.

Hugs my dear God Mother

---

118. *My daughter is doing well on the MMS. Her ATEC went from 53 to 26*

in three months. The special education teacher at her school was very impressed with how well she is doing. She told me to keep doing whatever it was that I was doing, because it is working!

Thank you for your help.

---

119. *Kerri, good day*

I want to tell you about the progress of my child, the treatment began in July and the progress that we have seen is:

-He has gone from not wanting to eat ANYTHING, TO EATING BY HIMSELF, SITTING DOWN he grabs his spoon and he cleans his mouth every time he gets dirty

- *He has stopped wearing a diaper, and goes to the bathroom by himself, almost like any other typical child.*

- *He lets people other than my wife and I hug him.*
- *He now gives hugs and has started imitating a few other things.*
- *He can now obey simple instructions that he couldn't before.....no, pick it up, get down, get up, give me your hand, etc.*
- *At the institute where he receives therapy they have told us that they are seeing great improvement in his eye contact, as he now focuses his vision more clearly on an object.*
- *He still will not play with other children, but at least he is where they are.*

Thank you for everything and we are together in this fight day after day, winning small battles everyday to win the war.

God bless

---

120. My son was a non responder. We worked with a DAN! Dr. for 14 years. We did the DAN! protocol tirelessly with very, very little gains. Did 3 years of ABA and 2 of SonRise. Did the diet. Still do the diet. We started the MMS protocol in April of this year. My son has had more gains since April of this year than with 14 years of DAN protocol combined. My local children's hospital specialist told me to put my son in a home and forget I ever had him. What a jerk. We pressed on and I could write a book about the gains since April. My son's Dr said he had two near impossible "tough nut kiddos to crack open...one being my son." Well, MMS works. If my son can improve I know yours can too. Whatever you have to do to do enemas, and get him to drink it...do it. My son is drinking 35 drops a day. Trust me, it is yuck. But he understands that if he drinks it, it makes the autism go away and he slams it down and says, "yuck." He totally knows I am helping him and cooperates. Be encouraged.

---

121. Hi Kerri,

We're now at 1/4th of a drop per oz and today my daughter for the first time EVER answered a 'Where is...' question. I was putting out the garbage and when I came back in I remembered I had forgotten to put away a black marker

and when I found it was missing, I could only hope the property damage would not be too bad this time. I checked the living room and the kitchen but could not find any traces and meanwhile my daughter had taken the photo camera

and was taking pictures of herself in the mirror. Amazingly she seemed to have a way better understanding of the thing. I asked her twice where the black marker was, and after the second time she pointed towards the kitchen and said "There."

SHE HAS NEVER DONE THIS BEFORE! I checked the kitchen and found the black market next to the fruit. The pears, melon and mango all had black circles and stripes. I was a bit overwhelmed I should say.

Thank you, thank you, thank you for never having given up your search!

---

122. My son Fernando Latin is 8 years old and was diagnosed at age 3 and half. When I started this struggle with autism all the relevant studies were performed. We visited a neurologist, psychologist and a gastroenterologist, he began taking the supplements according to DAN! protocol, but the list of pills he was taking was increasing and my boy was restless. Then I went to a homeopath, *the treatment worked a little but it was so expensive I couldn't continue, it is needless to say how much money and time I spent on appointments and tests, plus the hyperactivity problem was growing each day. Until God put in my path an extraordinary friend MARIA CAMPERO and she talked to me about a foundation in Port Ordaz "VENCIENDO EL AUTISMO" FOUNDATION (Carolina, Yamileth and all that beautiful team) they brought Kerri Rivera to lecture here in Venezuela, on 04/02/2011. I will never forget these words AUTISM IS PREVENTABLE, TREATABLE, AND CURABLE. A new hope that a mother's heart never looses… In that conference she also talked about MMS and its benefits I was lucky Kerri evaluated my child, I started giving him MMS plus some supplements and have won this since:*

- * *Better pronunciation, for example he used to say Dowg and now he says Dog.*
- * *He obeys orders.*
- * *Spends more time sitting.*
- * *Evacuates normally with no bad smell.*
- * *He stopped biting his clothes.*
- * *He stopped making noises like: wiiiii yiiiii.*
- * *He stopped playing with his hands.*
- * *We go somewhere and he is relaxed.*
- * *If we take a picture, he looks at the camera.*
- * *He is more aware of his environment.*
- * *He is improving a lot and really fast with his ABA therapy*

These are all the benefits I've seen with my son after starting with the MMS, I know there is a lot to do but know a have real hope. Some advice for parents that are starting this road: please accept what is happening, inform yourselves. I always think that to get your children through this you need the four-legged table:

- * *Diet*
- * *MMS, Supplement treatment*
- * *Therapy*
- * *Help at home*

I've always asked God to give me enough wisdom to help my child and to put me in a direction were I could find people that can help me, and believe me he has listened: Venciendo al Autismo Foundation, Kerri Rivera, my sons therapist Milagros Alcanta, Danny her teacher-shadow, Maria Campo, a woman not only committed to her son but to all of those who she can help and guide me, including Marilyn Aparcedo and her team. All of them are my bases to keep going and as Kerri and Carolina say: This is an endurance test and there is no giving up!

I have nothing but thanks these wonderful and professional people, Thank you!

Norkis Pinto, El Tigre—Venezuela.

---

123. Alexander was a normal baby until he was 6 months old, after that he suffered a change that, when we got home from work it was like no one came in, he was very uneasy, we talked to him but he didn't seem to care, first we thought that he didn't listen, the pediatrician and neurologist recommended hearing tests, *we did them and his results were normal. The pediatrician also recommended a psychiatrist but I thought it was very early for that so I took him to a psychologist, but he didn't diagnosed him, instead he referred me to a psychiatrist specialized in autism, and with a simple test he confirmed the suspicions: your child has autism and this has no cure, take him to special classes and therapy. I quickly signed him to a special camp and school, but they they just took my money and my sacrifice to travel 60 kilometers daily, they didn't do what they promised, they charged me triple just to give me a diagnosis and I believe he didn't even receive therapy. In my despair I went to the internet, I made an email account, a facebook page and started to look for parents that were going through the same things I was, then I contacted a Dominican mom living in Puerto Rico and she told me about Kerri Rivera and the diet. She told me that Kerri was the person to go to about this and Kerri answered immediately and asked me if my boy was on the diet and I said he was only 50% and she told me that if I wanted my boy to get better I had to do 100% so I wrote her back when I was doing the diet entirely. She gave me the first three supplements MMS1, THERALAC AND QUINTON WATER. Everything has been so difficult to get but if you persevere you get want you're looking for. I got the MMS1 but I was scared because of the strong smell of chlorine but I gave it to him with faith, I took it and my youngest baby girl started taking it. She had had a cold for about a month and it disappeared, Alex started to get better in every aspect but the biggest changes began 15 days ago, since I give him his MMS1 treatment. I start at 3 o'clock and finish at 12 at night. On weekends I start at 1 and I complete 12 doses of one drop mixed with citric acid and give him that daily. My 3 year and 4 month old baby boy is saying his first words. He is more connected, pays attention, he is aware of what happens to him and he is more loving than ever. I am so happy; I thank God, Kerri Rivera, Zeneida, Cynthia. But I mainly thank MMS1, because with a supplement so cheap, I'm getting my son back, the one that Autism wanted to steal from me.*

---

124. *Juan Eduardo is 9 years old and he was diagnosed with Autism at age 2. It has been a long road in which we have tried everything. 7 years after trying and failing, in April of this wonderful 2011 we started in the right way. MMS came to our lives. Juan Eduardo was a self-centered boy and with a lot of fixations to objects. His improvement in school has been slow, but in 5 months he has improved what he improved in 5 years. MMS in our lives has been wonderful, despite the hyperactivity and lack of sleep at the beginning of the protocol, the life of Juan Eduardo and our lives have changed drastically in a very positive way. He will start regular school at the beginning of next year. He is a happy boy, he sleeps all night, he talks a lot more even if there is a long way to go, his quality of life is so much better. He shares with whoever is around him, a lot of times he surprises us with his qualities. He has adapted to the MMS in a great way. We took the panel viral test and the results were amazing, there was no virus. When I saw the great positive effects the MMS has made, I started to do protocol 1000 for maintenance with 100% success. Last year I suffered from several episodes of flu in which I had to take antibiotics, but this year I feel excellent, Thank God and Kerri and MMS I haven't been ill. I thank God, Kerri for putting you in our lives; I'm sure he has given back everything you selflessly have done for everybody. The smiles and looks in our children's eyes are priceless. Thank you Kerri and Jim Humble. A big hug.*

---

125 Yes, I have increased the drops. Honestly, since we last spoke, he has had some of his best days...super happy. He actually read a book on his own on the way to school this morning. Seems very calm. Thank you.

---

126. *Hello Kerri! I have not written you in some time. I must tell you I have already been using MMS for two months. We have seen a lot of improvement. He suddenly speaks, says single words. If I sing a song or the vowels, he tries to imitate me, although it is still not very clear. But he tries to imitate. He smiled back at me when I smiled at him. And so those are several things. Tantrums are also less.*

---

127. *She is doing excellent! Every day we see small but important progress. I am not sure if she has noticed the changes in herself, the best will be when she starts school and classmates and teachers can give me their opinions. Family members have also commented on the changes we have seen in her.*

She is also answering questions about what she wants... With yeah, nah or the sign for all done with and then says dddon!

Now she wanted to see Mickey and told me tttttiii and I showed her the movies and she said "ickey!" ;) We are communicating!!

---

128. *We are a family that lives in a country where living is not easy at all. The maternal death rates and child deaths are very high, and there is hardly any chance of getting access to special medical treatments and therapies. Our son started to show signs of autism since he was 2 years 8 months old. He had begun speaking but at that moment he stopped talking, he started to show no expression with his eyes, to roll around trees, to scream and to tantrum, and to not allow contact with anyone. Little by little he started to be very hyperactive and to have serious sleeping problems (he didn't want to wake up and he could not go to sleep at night).*

The start of my son's autism went along with colds, severe coughing and infectious diarrhea. Our son's teacher had always had excellent reports about our son's behavior, but since this severe cold with synovitis in his hip, the transformation begun. Since then, the awful diagnosis and the absence of guidance led us to an abyss. Our son presented total absence of self, almost no language, hyperactivity and many sicknesses. 6 months went by in which he met with a real parade of doctors and therapists that gave us bad treatments, bad diagnoses and almost no information, our life became a living hell. We went into a state of denial in which we didn't want to see anything on the Internet regarding Autism. In that time our son had been diagnosed with severe sinusitis, complications with his adenoids, and limitations that he developed because of his hearing capacity.

By the end of the year 2007, I alone broke the pact I had made with my husband, and started to search the Internet and found Yeroline Ruiz's page where she commented about the biomedical protocol through which she had gained her son back in months. Little by little I tried to convince my husband, who at the beginning was completely reluctant to the possibility, but by December 2007 we made our first order of supplements and quelators (magnetic clay from the United States). The supplements took forever to get to us and when they arrived we gave them to our son without any clue or guidance and hardly any diet. The results were horrible, he got worse and worse, and this lasted about 8 months. During this time the therapists gave us a terrible prediction about our son's future. Our family problems were also getting worse and the only guide we had was Yeroline and the Curando el Autismo foundation. Thank God we didn't give up and we passed the hard times.

We eliminated some of the food that is forbidden in the protocol and we started to use Enzymes. Our son started to improve, his health got a lot better and we stopped visiting traditional doctors and I decided to take care of his treatment, the process was not easy because there aren't even the basics for the protocol in my country, from the flour to the milk, everything is really hard to get. Obviously a DAN! doctor or supplements are unthinkable. However even if his health had improved, the Autism symptoms were still present, the speech therapy did not help and sometimes even made it worse. With a lot of effort we brought Rosa Dominguez (Tomatis Therapist) and we begun applying this technique to our son. The economic effort was very big but the results were worth it, he was more in control of his body, the clearest sign of this was the control of his sphincter that happened almost magically. His language also improved even if he still didn't use it socially, but little by little we verified that our son had no sign of retardation compared to other kids his age. Even with the improvement, the tantrums, frustration and limitations of his language remained, we decided then to make a new effort and traveled to Panama to the Stem Cells Institute to make a stem cell transplant.

Our debt was huge, and despite the effort our baby showed two contrary signs: a positive one that was that our boy was completely connected to his environment (he is never outside of the world anymore) and a negative one: he became aggressive. He had never shown aggression, auto-aggression or violent behavior, but from the moment the anesthesia was applied these began.

Our child started hitting himself when things didn't go his way and he bit other people out of frustration. We remained on our path with despair because we could not see the light. We tried hyperbarics for 70 days in a row, eqinotherapy, and others. In 2011 we came across Puerto Rico's CEA conference and we found Kerri's topic and the SCIA topics very interesting. We decided to do the SCIA tests and we confirmed that our boy also had virus and an inflammation even though his immune system didn't seem very affected. Because of our country's limitations we couldn't complete the SCIA tests but then we had a lead on how to start a new journey. All we needed was a doctor that could apply the SCIA protocol.

But then there was a miraculous day in which we saw Kerri's talk and hope grew inside us, we contacted her immediately and she answered immediately too, with good will and all the indications as to how give him MMS and how to apply the protocol. After two weeks we started seeing a change in our son. Eye contact was way better; he speaks fluid sentences and gives structure to his words.

We are now three weeks into the protocol, we can see almost no aggressions and tantrums, we teach him letters and he recognizes them, reads and writes his name (he does it with some difficulty) his receptive language is %100 active, he is very loving, specially with me. (His mom), he hugs me and kisses me and tells me he loves me. We know we still have a lot to go but I'm certain that we will make it. My son has now Borderline Autism; his connection with the world is total. We are still thinking about doing all the SCIA tests but just to find out our child is cured. The light and hope that Kerri and the MMS have given us does not compare to anything. Thank God! Kerri is a blessing in our lives!

129. *This email was after an ATEC drop of 64 to 44:*

This is truly AWESOME!!! He lacks in the "Speech/Language/Communication" area the most!!

I honestly thought when I was filling this ATEC out that he would have regressed because I just thought he might have...it seems everything else we have EVER done it is always too good to be true! It actually is not this time!! Praise the Lord Jesus Christ! \o/ I even had my oldest daughter give her thoughts on the questions too so that I doubled checked myself regarding his behavior etc...you know to make sure I was not out of touch! :)

We definitely have a ways to go, but I have never had my son at this level of 'awareness' before! We are doing RDI (Relationship Development Intervention) and have been since Feb. 2011...he's made great gains with it, but he is FLYING now only because of the MMS!!

~Melissa, IN, USA

---

130. *Hello, my name is Silvia and I have a 5-year-old son named Ricardo who was diagnosed with Autism 2 ½ years ago. From that moment he has been on a gluten and casein free diet and a therapy routine that help him day by day. Three months ago I met Kerri Rivera and the Venciendo el Autismo Foundation, and started giving Ricardo MMS. From that moment on he has improved significantly, his language is better and his behavior is a lot more controlled, plus with the biomedical treatment he goes to the bathroom regularly, this makes me very happy because he was always constipated.*

Silvia Morales — Caracas.

---

131. *Since I was a baby I was a very sick child ... I remember I had many gastrointestinal problems, I made my mom's life miserable with all the diarrhea and vomiting ... I couldn't gain weight and I had no appetite. All this accompanied me for many years; the diarrhea became chronic constipation, fluid retention and dry skin.*

All of this without knowing I had a serious problem, until I gave birth to my first child. The constipation was a nightmare, my baby got ill, when I fed him breast milk he would get a rash, and then he had convulsions for about 20 seconds, when I fed him again the same would happen but now the convulsions lasted for nearly a day. The doctors didn't know what was going on either. In one of my researches I ran into a page that talked about Autism and Candida being one of the main causes, this is when I find out that my son had chronic candidiasis and that he got it from me when I breast fed him. I tested my child and myself and we both had Candida, and I have had it since I was a baby.

One day I heard about the existence of a woman named Kerri Rivera who is devoted to helping children with Autism and I contacted her and she gave me all the help to treat my child. I will be forever grateful. She talked to me about the MMS and how to give it to my son. I also started taking it and in a matter of days I had a white layer in my tongue, it was the Candida getting out. I have gained my energy back, the headache I had in the mornings is gone, I feel vivid and almost all the Candida is gone. I've been with MMS for a month and I've never felt so satisfied with a treatment. This simple and yet powerful remedy has given my health and family back and with this I can assure, by my own experience that MMS works.

132. *Hi Kerri, I want to tell you that Erik has started with the speech therapist, and psychologist, when she finished the session with Erik, I asked if she thought that Erik could still be diagnosed with autism. She looked me in the eye and said, autism? Are you sure that he was correctly diagnosed? I found her the forms and she told me today Erik has eye contact, he socializes, and is cooperative. She said he does seem small for his age, and that we were going to have to work on strengthening the muscles of his mouth and the tongue as well because they are very week. She said that not chewing would affct the development of one's language, and Erik has only been chewing for 8 months, which started after we did the chamber with you in Mexico. That was the answer she gave me, and that's something, the fact that she told me that this is not autism is really something, even though she doesn't want to contradict the diagnosis they gave two years ago. The truth is, I am so glad he has come to improve, and that is since you KERRI have started to help my child. It's been a year since we started with the MMS protocol and the improvements have been incredible, he still lacks language as he mispronounces many words, but the worst is over. You can publish this if you want, thank you very much for all your help.*

Kisses

---

133. *We had first formed stool at 4 drops. It was so inspiring. Since then somedays we have formed stool, some days we have loose stool but I relate this to detox - I can clearly see his gut healing. His expressive speech hasn't changed yet but he obviously understands much, much better and is able to follow complex comands like "come and give me one teaspoon from the drawer" or "go and put this in the dishwasher" or "wait for daddy to get out the baby from the car and then you can lock the door and put on the alarm". Things like that, that were unthinkable before. I hope he will go beyond his few words soon and will begin to talk.*

Although speech is the thing that interests me most of all, perhaps the most impressive thing that MMS brought to us all of a sudden is the sense of smell - our boy never seemed to feel any smell, good or bad, in his entire life. He got this sense a week ago. We are thrilled - something is waking up in his brain, perhaps some pathogens in some part of his brain or in his nose or whatever died and now he is able to sense a smell. Things are getting moved.

Thank you Kerri.

---

134. *My son is 8 1/2 years old. I have spent six years and well into six figures to help my son. While I would see gains, nothing was really moving him forward. Three months with MMS, DE and ocean water; the junk that is coming out of him is both horrifying and thrilling. I think he is finally starting to feel better, and because he feels better he is more with us. He is trying so hard to speak and everyday he is saying at least four new words. He also watched a movie and laughed at a funny part!*

---

135. *We started MMS 68 Days ago and my son's ATEC has dropped from 50-23... Now that's a MIRACLE!!! He is So much more "with us" and I'm seeing improvements daily. He is just so much more fun to be with now:) He did have a plateau, but is back on track. We have only done 4 enemas to date, but they made a big difference! Baths do as well:) I can't wait to see what he is like on his 4th birthday (still 39 days away).*

Kelly, CA, USA

136. *Is it just an amazing day or is the MMS already working? Its just day 1! I know people say that once your child starts talking, they never stop, but holy cow... Today in the grocery store, it took me 3x's longer to shop because he was talking about everything he saw around him and needed a response on all of it from me! So unlike anything we have ever experienced with him before. Could it really be the MMS already? Also, potty training has been going well (slow & and steady), but today, he didn't need much prompting! Many times, he just stopped what he was doing and ran to the bathroom to go all by himself! This is amazing, because he never stops a preferred activity to go potty! When i thought it had been too long since he went last and said "its time to go potty." he just ran in there and peed, washed hands etc. No tantrums or weight dropping, he just ran in and went! God is answering my prayers! :) I'll keep you posted on what happens tomorrow!*

---

137. *You know Kerri, I had to email you directly I waited until today to retake my son's atec...just did his first Parasite Protocol and did the 72/2 hr this weekend, so waited a few days to pass by before we did...so this atec was based SOLELY on this past monday, tuesday & today. I got an 11, hubby got a 13. so average: 12. (he was 32 before MMS, then 27 after 3 weeks of MMS...now 12) I had to email you to thank you b/c well...because. What is there really to say :) cant think of the right words really :) thank you?? not really enough if u know what i mean :) ...But to think i was going to skip your presentation at A1 b/c I thought it's too good to be true. But you caught my attention when you started tearing up as you got up on stage :) you were the only person/speaker to show emotion like that. AND have NOTHING TO SELL ON YOUR TABLE :) You were there only for our children...getting nothing back in return! :) I know I've been emotional, but I'm prego and my sons getting better so cut me some slack :)*

I'm SO grateful for all you do Kerri. If there is something I can do in return, Please let me know. I know you have a lot on your plate. You have a heart of gold; such a giving, compassionate person. a "rare-ity" (this word is in my dictionary!!) You & Patrick are always in our prayers. I seriously want to go visit you in Mexico w/my family so my son can meet you when he's recovered. I told my hubby & he's up for it hehe :)

Anyways, long email but really just wanted to let you know how well he's doing and to 'thank you'. It blew my mind after I got the results. He still has issues in socialization and getting his vocab/speech up to his 4 year old level. I think much of that will change as he learns English better: he'll interact better with his peers.

Thank you Kerri from the bottom of me & husbands heart :) You are loved & covered in prayer. :)

---

138. *We started to see a setback in January with our 1 year 8 month old girl. Until this date she had had a normal development, like her sisters (she is our third child) but then, she stopped talking, she just screamed and twirled. She would twirl like 20 times a day for no reason until she got dizzy and she fell to the ground. She didn't look at anybody in the eyes and could be behind a curtain hours comparing her toothbrushes; she wasn't hungry nor asked for her milk, we had to push the food into her mouth so she would eat it. I was desperate so I went in to a forum and a friend from Venezuela told me to look for Kerri Rivera in my country so I did. She immediately answered and gave me hope, since the*

three doctors I had seen before told me that Autism had no cure, that all we could do is to teach her how to survive and be a little bit more independent.

I found my hope in Kerri, I read the testimonies on her page and sent her a message that she answered right away. She told me to do the diet, remove milk and everything else, I went to her clinic and within a week of the diet Abby could look at me and smile (she had stopped doing that long ago) When I went to her clinic Autismo2 in Puerto Vallarta she looked at Abby with all the patience in the world, she gave me MMS, the supplements and all the hope in the world. Everything was new to me but I had faith that I was in the right place and from that moment on, while always counting on Kerri's advice, my baby girl started to get much better. Her first ATEC was 95 and a year and a half after it's a 15. My girl now smiles, talks, dances, she obeys and stopped screaming, her only tantrums are of a normal girl her age. I did hyperbarics and am still doing MMS, supplements and deworming her. She is now diagnosed with Pervasive Developmental because the Autism symptoms are gone. She doesn't run from one place to another, she plays, looks for other kids to play with and has friends. The only thing left is for her to establish a conversation, and to put together the words to make sentences, but I know she is going to make it.

Once she sat down on ant hill and a bunch of ants bit her, we had to go to the hospital and the doctors gave her cortisone and she was seriously ill, now when a bug stings or a mosquito bites she tells me where it bit her, and she asks me to apply a calendula ointment that I use for the mosquito bites. She sets the table and can eat by herself, she doesn't use a diaper anymore, we can go shopping or to parties and she is well behaved, there is no more tantrums or fixations, this is wonderful! I know there is a long way to go, I've been in this fight for a year and a half, but with KERRI RIVERA'S HELP I will rescue my baby from autism, even though some say that Autism has no cure I know it does. Kerri has been a blessing for our lives and I thank God for putting her in my path.

Thank you Kerri, I have no way to repay what you have done for Abby.

---

139. Our 10 year old son, Ben, showed symptoms of autism since he was an infant. Until he was 6, he never connected normally with family or peers. As his mother, I ached for a hug, to be called "Mommy," or for a lingering look of love. I never got them.

Until he was diagnosed with autism at age five, I was utterly befuddled. What had I done? What was wrong with Ben? Why didn't he want to play with other children? Why did he drool endlessly, laugh constantly for no reason, and bolt from play dates and parties with no warning? Why was he constantly lethargic, to the point where we called him "the playdough boy?"

Ben was always happy, dwelling far from us, in his own perfect little heaven. My husband and I were depressed, even despondent, despite attempts at gluten-free, casein-free diets, ABA therapy, and many supplements, because we saw no substantive and lasting changes in our child.

Finally we found the Son-Rise Program®, which brought Ben back to us. After three-plus years of a mostly full-time Son-Rise Program®, Ben became highly interactive, engaged, and interested in his peers. That alone was a palpable miracle. As we say in Hebrew, Dayeinu! That would have been sufficient. But we constantly experienced regressions back into autistic symptoms. For no apparent reason, and despite endless biomedical interventions, including the DAN!

protocol—to the tune, I might add, of tens of thousands of dollars, multiplied over four to five years—Ben would suddenly revert to autistic behaviors, and we would lose contact. He became once again non-responsive, uninterested in family or peers, and seemingly unable to connect. Only the day before, he had been totally connected.

After being on the SCD and GAPS diets, and finding even them ineffective to alleviate these regular and devastating returns to autism, we were advised by Dr. Campbell-McBride that it sounded to her as if Ben had parasites. A friend of mine who was also a Son-Rise mom had mentioned to me that she and another Son-Rise mom friend of ours were trying something called "MMS." When I heard the words for which the acronym stood, I gave an internal scoff. "Master Mineral Solution? Is this a joke?"

I can tell you, in all seriousness, that now, five and a half months into our MMS autism protocol, MMS is definitely no joke. Since starting MMS, we have seen nothing but progress in Ben. We are still doing GAPS, but avoiding foods we know Ben cannot yet digest. Ben's ATEC score has dropped from 43 when we started MMS to a startling 20. We are seeing sustained, natural, unsolicited and consistent eye contact, like never before. Ben is consistently responsive, certainly on par with most 10 year old boys. He is playing hockey in our driveway with friends. He is solicitous of others, and compassionate when others are hurt in any way. He is funny, expressively loving, and eager to please and be loved. It makes me cry to write this. I wanted this so badly, and now, with MMS leading the other interventions, I have what I prayed for.

We are not perfect on the MMS. Ben is in his first year of school post-Son-Rise, and dosing during the school day is not easy. But we are tremendously hopeful and excited.

I truly believe, as Kerri Rivera has often said, that parasites are the missing piece of the autism puzzle. Like so many other warrior parents, we have done everything. But everything minus one was not enough for Ben. Now, I think we truly have the tools, and it is only a matter of time, faith, and love, to fully and permanently recover our beloved Ben.

---

140. Caitlin has been on the MMS protocol for 3 months now. Caitlin is now off her 100 plus supplements a day that were the glue holding her together so to speak, with no regression. She is happier, more affectionate and focused. She no longer has a distended belly. She used to have terrible smelling gas on a daily basis and very foul smelling bowel movements. This is gone! She has increased muscle tone (has a diagnosis of low muscle tone and bilateral hip anteversion). She has been seizure free as well! She continues to make slow but steady gains in her reading and math skills at school. She isn't conversational in her speech yet but can comment on things and express her wants in longer and more elaborate sentences. She is falling asleep faster and sleeping longer.

---

141. Eczema!!

At last my son (4 yrs) is living now without eczema, he was born with severe eczema all over his body, we used every allergic cream and med and even the magical Dead Sea Clay but it didn't work, but now after using mms and pp (parasite protocol) for 2 months... ITS GONE!! Thank you MMS!!"

Thank you!

---

142. *This is the story of our son who is 4 years and 9 months old today. Our story is a desperate struggle to defeat Autism, the same struggle that a lot of families are going through to free their sons and daughters. Guillem started to have ear inflammation and laryngitis when he was a few months old and we medicated him with antibiotics, he had his vaccines according the vaccine calendar, and by the time he was 9 months old he was showing signs of Autism. He would put his fingers in his throat until he choked, he would hit his head on the wall and he was always mad, his tantrums were frequent, longer and more intense, he started to look at the lights and to twirl with no sense, he showed less interest in people day by day and he would get really mad if someone said anything to him. He was hyperactive, couldn't stop running and going up and down the stairs. The nights were terrible too, he didn't sleep well and woke up screaming as loud as he could plus he could not say a word.*

My father and my brothers blamed me for this; they said that my boy wanted to dominate me and to get my attention. I had a hard time recognizing that it wasn't this, that my boy had something else. Once I accepted it I spent hours and hours reading on the Internet, and I read about the gluten-free diet and that was our first step. When Guillem was 2 and half years old we changed the diet and it worked. Guillem started to respond to orders, there was less screaming and his first words appeared. Three months into the diet we contacted Kerri, this has been a key moment in our history, because she has been our guide to get our son out of this long tunnel. The first thing she did was to check what our son was eating, I thought that I was doing a perfect diet but I wasn't, there were some foods with sugar, soy and additives, the next thing we did was to talk about the supplements and MMS, but I was very scared to use it and it wasn't until three months later, the day Guillem turned 3 that I started giving it to him. Now we have been using MMS for a year and 9 months. We give him MMS daily and every weekend we can but minimum one weekend a month we do the 72/2. His improvement was amazing, his language got better, his dressing skills, he learned how to use the bathroom, his sleep was normal, and the screaming was almost gone. The next step was to start with the MMS enemas, every enema made Guillem more alert, more present, we started with the enemas 4 months after starting with the MMS. We saw that the enemas expelled a great amount of biofilm and the more he expelled it, the more abilities Guillem gained back, I saw the importance of healing the intestine and that has been my obsession since.

The next important step was the sessions in the hyperbaric chamber, his first sessions were at age 3 and 9 months, and the changes we saw were mainly in the obsessions he had, he paid much more attention to everything, his social skills were very affected but after this he made his first friend. Now we do hyperbarics with seawater every three months. Another piece in the puzzle were the parasites but with Kerri's help we've been on the protocol against parasites for 7 months and every deworming cycle our son improves. I know that we have to keep fighting but I also know that we are closer to the exit. Now Guillem sings, he plays with his brothers, we go to the movies, we do puzzles, we paint, and he is very happy, he has learned to spell his name and he walks joyfully down the street telling you about his day. He goes to school, plays football and more importantly, we are a family again.

For all of this we are deeply thankful to Kerri, because she has always been there for us, without Kerri and MMS our story would be completely different, so we owe them EVERYTHING. I know that we will someday say Guillem is cured, and all the other children with Autism can say that too.

UPDATE: As of December, 2012 Guillem no longer has a diagnosis of autism, here is an email his Mother sent to Kerri when she knew he had recovered:

> *KERRI!!!*
>
> *My son has an ATEC of 10!!!!*
>
> *I have it repeated several times because I did not believe it...*
>
> *I have copied it below...I'm going to have it framed!*
>
> *The truth is that this Christmas we have enjoyed our whole family, we've been to the Zoo, to the fair, and during family meals he was very well behaved and enjoyed the Magi and the magic of Christmas. Christmas last year was totally different. For all this we give infinite thanks! There are not enough words to tell you how we feel! Without you our story would be different. We will continue working to get to zero!*
>
> *A big hug!*

---

143. I started with MMS on April 2012, I was afraid but had a lot of faith. I did the ATEC for my son and it was 92. He could not use the bathroom alone nor say a single word. The day I started I thought there was nothing different, but when we were at the beach on vacation I noticed that when I called his name he would turn, and come to where I was. All this with one drop! Before that day I always had to stand up and find him so he would pay attention to me. He wouldn't look at me but he would listen to instruction. Days went by and when we came back from the beach he went straight to the bathroom… I couldn't believe it; I sat down and asked my self WHAT IS THIS? What happened? And I silently thanked God and Kerri. This was decisive for me to continue. After this he improved day by day, after 5 weeks his ATEC had gone down to 76 points.

I made a few mistakes at the beginning because I didn't know the protocol very well, and wanted to rush things. Again, Kerri with her infinite patience and love for others led me to the right path, and I started to see more, way more.

My son started writing, he knows his letters, he reads some words (he has memorized them) he is great in the bathroom, he listens, his eye contact has improved significantly, his last ATEC was last month and it was 60 points. I know we have a long way to go, but I every month there is improvement and even though there are some bumps and setbacks the good things overcome the bad.

My son has also started saying words, names of movies, he can say over 100 words (names of movies or things he likes). Some will think it doesn't mean anything to know the name of a movie but for me, it is to know how my son's voice sounds and that, is priceless. Now my next goal is to hear him say "MOM"

~Myrna Sterling, Mother of Gianmarco, age 5.

Myrna is the president of the GM foundation for Autism

---

144. My wife and I have always been fond of traveling, getting to know different cultures and people, this is why we had our little one we decided t o keep traveling but to adapt to his rhythm.

When he was 9 months old he went on his first international journey. Everything was perfect, he was the happiest baby on the plane, he played with the people in the seats behind us, the stewardess wanted to stay with him to play, all the plane ride went along perfectly.

Something happened between his 15th and 18th month, his smile was gone, we no longer had happiness in our home, only nerves and less language. Maybe the excessive vaccines were destroying his immune system and leaving the door open for parasites to take hold.

When he was 18 months we went on a second trip and it was an absolute nightmare. He wouldn't stop crying and screaming on the plane, he didn't want to play with anybody. We spent a few days getting to know this new place while our baby fell further into this abyss in front of our astonished eyes.

Something was wrong, and after an episode of auto aggression we knew that this was not normal. After watching a couple of videos on the Internet we started to understand, and we went to a psychologist who didn't give us a diagnosis because our son was so young.

Later on we went to a neuro-pediatrician who also used the excuse of our son's young age to dismiss us, and finally a last doctor who didn't explain anything either.

Not even our family listened because they thought we were exaggerating. We were left alone and our biggest ally was and still is the Internet.

After the medical failure and their "anti-diagnosis" we took the best and most important decision of our lives, that was not to listen to any of the diagnoses, or to wait another 6 months to be called by the neuro-pediatrician again. We started down this road without an official diagnosis but with all the symptoms and behaviors that define autism.

We started with the casein, gluten and soy free diet with no experience on how to do it and our baby suffered awful withdrawal symptoms. We couldn't believe what we were seeing, his nervousness and aggression took us to our limits. After a few days on the diet his tantrums went down little by little until he was stable. But we needed something else because everything was out of control; this is when we went back to the Internet.

After three months of false doctors and pseudo medical experts we found Kerri Rivera.

She gave us hope, and filled us with confidence, but most importantly she was clear with us from the beginning. Everything she told us was logical and everything made sense as to why it was done so we trusted her without a moment's doubt.

He got better and better everyday. Kerri told us how to do the ATEC and what it meant. Our first result was 57 points when he was 2 years old, which was very high for us because of his young age. We have been doing everything Kerri suggests, improving the diet, and using everything she recommends, and the results have been spectacular, right now he is 2 years and 10 months old, his ATEC is 8 and he is doing great. Our goal is to have him at 0, and I'm certain that once we free his intestines of all the parasites, he will be a strong, healthy and happy boy.

He doesn't have any behaviors like the ones he had one year ago, now he smiles again, he talks and asks for things, he plays a lot and he is happy.

This was possible thanks to his Godmother Kerri.

Kerri, you have done and given us everything in exchange for nothing, and that is your defining quality as a human being and as a person.

Every day you give hope to all that YES WE CAN defeat Autism and now we have a lot of puzzle pieces to complete our puzzle. We may find a new one soon because there are a lot of people investigating and that makes us feel alive, to fight against the entire circus that has sprung up around this problem.

I would recommend that if anyone is not seeing progress in their children, to come to Kerri, and try everything she says word for word. She has done it all and has the best investigators and doctor on her side. The ones that aren't intimidated by the absurd laws of the upper echelon, those who only want our children to be sick all their lives so they can sell medicine to us.

We will keep up the fight.

---

145. Hi Kerri, I just got the third MMS Autism newsletter, and I tell you it couldn't have arrived at a better moment. For the last few days my nephew has had a strange infection, he was hospitalized and the doctors couldn't figure how to fight it because it was resistant to many antibiotics. I told my brother-in-law about MMS and he started using it. After two days my nephew was healed, the doctor asked me for a little so he could examine it, and his answer was that the way MMS functions is amazing, he couldn't believe that such a simple and powerful product hasn't come out to the public eye.

---

146. My husband and I were on the couch having our morning coffee and my son came bolting out of the room and ran up to my husband's face and said..."Oh Hi Daddy, I'm back!" And then ran back to his room...this is our once in a while "word" little man. After 3 non-progressive PPs (parasite protocols) the 4th one was a charm. A nightmare, but quite an eliminator... Lots of Worms Gone! And he is slowly coming back...we are seeing it in fact he is even reassuring us that he "is back".

~The previous email was accompanied by an ATEC drop from 95 to 35 in 6 months.

---

147. Dear Kerri,

It's a pity I don't speak English and that's why I can't write you on my own. Currently, a lot of good things are happening to my son (31 years old). I have followed all your advice and as a result he has completely calmed down. I have learned to prepare CDS at home with the help of Jim Humble's video. The CDS is a 600 ppm solution; my son drinks 30 ml per day with the feeding bottle method. I suppose this is the daily dose for a person of his weight. We are past the third PP and preparing for the fourth one.

Unfortunately I won't be able to attend the congress in Prague, since I can't find anyone to look after my son. If it were in the summer I would risk going by car and taking him with me.

I am very grateful for your attention to our life and progress and for your help in curing my son. God bless you for helping children with autism by sharing your knowledge and experience with the world. I will try to give a regular account on the progress of my son. I have no doubt he will be cured, too.

148. *Three more tiny little gains on our step by step path to recovery....yesterday Nick got a haircut and did not cry at all AND was relaxed enough that it was the first time it wasn't like a sheep shearing, he actually got to get a little style (so cute!). Also, he was doing some pretty terrible screaming in the car in the parking lot at Whole Foods while I tried to find a parking space (I wanted to scream too) and I let him know he scared me. When we got into the store he spontaneously said "I'm sorry" -total first! Lastly, He has requested a hug from me three times recently - HUGE new thing and I love it. Thank you MMS and Kerri.*

---

149. *Thank you Kerri. For all of your help, answering my emails, and for all that you're doing to help all of us, help our kids!!! I was at a breaking point. Nothing was helping her anymore, her organs continuing to get worse, hadn't grown in 4 years, and had become Jeckyll/Hyde this past 4 years...A violent, raging Hyde being increasingly present more and more each year, each month. I knew something had a hold of her all these years. And it amazes me how 2 endocrinologists, 3 pediatric GI specialists, and many other doctors had no clue.*

Thank you for ALL you are doing to get the truth out there. You are an angel!!!

Merry Christmas - xoxo

---

150. *Puerto Ordaz, November 4th, 2011*

My name is Carolina Moreno and my husband is Rafael Colmenares, we are the lucky parents of a little 6 year-old whose name is Ana Victoria, she came to this world to give us the world's greatest lesson, through her we have learned the real value of life, to be steadfast, and brave. We learned that it's our job to fight for our dreams if we want to accomplish them, and over all we've learned that God is who holds us up and gives us the strength and tools to move forward.

My journey with Autism began the 27th of march of 2007, my daughter Ana Victoria was 23 moths old, her initial development was normal, but little by little she stopped talking, she didn't understand anything, she didn't miss her parents, she didn't play with other children or even her toys, she didn't like to be kissed or hugged, she acted as she was deaf but we didn't know that all this were symptoms of autism. It wasn't until the day that we went to a neurologist in our area, that someone told us that this was Autism, an incurable condition, and that our daughter would have it all her life. It was a very hard moment for us; all of our plans to be a happy and healthy family went down in a matter of hours.

As parents we didn't want to give up, after recovering from the horrible news, we began looking for information on how to help Ana Victoria. In our country of Venezuela there were 4 specialists in Autism and they were all using the gluten and casein free diet, vitamin therapy, as well as expensive tests performed in the United States to check candida and bacteria. Thanks to them I learned how important the diet was to get Ana back from Autism, without it, recovery would be impossible.

We did some testing through Great Plains labs, and the results indicated candida, bacteria, inflammation and much more . We started with Ketazol and Flegyl. My daughter was taking up to 400 mg of Ketazol even though she was barely three years old. Now with everything I've learned I can not believe that a physician could prescribe such a high dose to a child who is only 3 years old and weighs only 16 kg. That alone should be punished and criticized

However, I felt that I wasn't doing the right thing, years passed, Ana Victoria was 5 and her comprehension and attention levels were very low, my husband always thought there had to be something else we could do!

One day my dear friend Carolina Garcia talked to me about Kerri Rivera's blog, she told me that it explained the reason for the supplements and the hours they had to be taken, for example that the antifungals had to be given away from food and the pro-biotics worked better if they are given at night. After I read her blog and visited her website in July 2010 I called the Clinic Autism02 and Kerri Rivera herself answered me. I explained my daughter's case, and she asked me to give her an opportunity to help, and if in 2 months we didn't see anything different with our daughter we could continue doing what we always did. I told my husband and we agreed to try. I felt a great sense of peace when when Kerri told me, "I'm here, and I promise I will walk with you until we recover your daughter"

In August Ana Victoria was on Kerri's full protocol and in November God gave me the opportunity to go to Puerto Vallarta, México. An entire month in this city were my daughter received the hyperbaric chamber treatments, we started with MMS, plus we got together with Kerri at her home every night with pen and paper, she explained to us step by step how we could heal our children. From that moment a beautiful friendship was born between Kerri and I.

This is when I understood that Autism is caused by a weakness of the immune system and this brings as consequences; viruses, bacteria, candida, intestinal and brain inflammation, food allergies and contamination with heavy metals, and that it is necessary to attack every aspect to be able to heal autism.

The 23rd of November Ana Victoria Started to take MMS and in 30 days of treatment we noticed the firsts changes. My daughter's stools were consistent; there were no food residues or foul smells. When I came back from Vallarta I was full of new knowledge and a lot of MMS bottles, Ana Victoria had already a very well defined protocol and the best thing was that it wasn't a long list of supplements, only 7, MMS and Theralac. Everyone in my family noticed a great change in our daughter, better eye contact, controlled hyperactivity, she started saying her first words with sense and purpose; I want park please, I want food please, I want to sleep. She was doing much better in therapy and at school. Anita had begun her way back to the world she left when she was 1 and half years old.

When I saw this improvement in my daughter I decided to share my experience with other members of my community and it was there with my dear friend Yamileth Paduani and her husband Alejandro Teran that Fundación Venciendo el Autismo was born. Mr. Alejandro asked if Kerri could come to Venezuela and talk to parents in a simple way so everyone could understand that Autism is curable. Kerri was delighted to accept the invitation and in only 7 moths of work we had 2 conferences and many evaluations and nowadays there are 700 kids on Kerri's protocol that is: casein, soy, gluten and sugar free diet, some supplements depending on every child, MMS and Theralac. The results are very positive and there hundreds of families that have started with their children's recovery.

Then came the great discovery of the parasite protocol, for me, it was the missing piece in my daughter's puzzle. The first few months of the PP were severe (fever, laughter, hyperactivity) and many worms in her stool. However, after all these worms came out out I saw how her mind cleared, and month by month we waited for the full moon to deworm ...

Ana Victoria is now improving very quickly, she is a girl that enjoys being with her family, the park, a day out, she goes to school and she says to me that at school she now does her work in silence without crying, when I arrive home she greets me with a smile, she runs to me and calls me Mom, she tells me what she wants and what she doesn't, she can use the computer like an expert, she takes her own shower and dresses herself, she even picks out her clothes, she knows when Christmas is coming and that there will be gifts. Her therapists and her shadow teacher are really happy to see how Anita is progressing day by day.

Yesterday we were all at the table eating lunch ... and she points with her finger …"Papa's food, Mama's food and Ana Kikota's (Victoria's) food…yummy… heeheehee. My husband and I were laughing with excitement and happiness. Some time ago we felt that speech would be nearly impossible for Ana Victoria.

My advice: do not be afraid of MMS ... be afraid of autism; that which is unforgiving and has the power to destroy lives and families if we don't attack it on time.

We are grateful to God because He has never abandoned us and has always given us the strength and tools to fight and to believe that it is possible to recover our children. We thank God for putting Kerri in our path and we ask God to bless her everyday for having such a big and noble heart, she who is always helping others in a selfless way. We also thank Jim Humble and Andres Kalcker for helping us find the missing piece in this great puzzle that is Autism. We can do it!!!

To all of you, many sincere thanks for helping the children in my country.

Blessings,

Carolina Moreno, President of Fundación Venciendo el Autismo. (The Defeat Autism Foundation)

Venezuela.

---

151. Good Day, first of all I want to thank God for giving me all the necessary tools to be able to help my son, Jesus Soto who is 8 years old today. Jesus was born normally with no complications, loved by his parents and family, everything was going fine, but when he turned 18 months old he became ill; laryngitis, and fever, amongst other things.

He was diagnosed with Autism at age 3. As a worried mother, I took him to the psychologist because I observed several behaviors that were not normal. After the terrible diagnosis we started to take him to a specialist after specialist, they were all very good doctors, his DAN! doctor in Venezuela, and my sons gastroenterologist were very helpful. I met a very special friend Mayerling Aparcedo who brought the ABA method to Venezuela. My son had many improvements but we still felt that we needed something else. I spent my time on the internet finding out as much as I could until I found Kerri Rivera that without even knowing me helped me via Internet, talked to me about MMS and about the supplements and how to take them. My friend Carolina Moreno had the joy to travel to Puerto Vallarta and wrote to me from there, telling me that MMS was helping a lot of children. The 14th of December I ordered my first MMS and the 27th of January it arrived to my country, from that date on my son is taking MMS.

The 2nd of April Kerri arrived to Venezuela and I went to see her, I took all the tests I had and she checked them, after only two months of taking MMS his viral *load went down to near normal levels, and his liver profile showed no problems. I made some adjustments that day and I continued with the treatment seeing a lot of improvement in his language, socialization, understanding of things, his hyperactivity has gone down and the laughter for no reason stopped. I'm very happy with the results and now all the family is taking MMS.*

I'm very grateful to God for putting my husband in my life, as he has been a very special part of this battle that we are winning, I also want to thank our family, Kerri, Carolina, Yamileth, Mr. Alejandro and everyone from the Fundación Venciendo al Autismo.

Yours truly,

Maria Campero

---

152. *My baby girl Yuliangel Nazareth Quijada Montero was diagnosed with autism when she was one and a-half years old. From that moment she was given a treatment to allow her to sleep and to control her tantrums: "Risperdal" which had absolutely no effect, I gave her up to three pills before bed and even them nothing happened, I was just damaging her brain. She didn't sleep night or day, I was worried because she screamed all the time and she hit herself.*

The 2nd of April 2011 I attended the conference of Fundación Venciendo el Autismo, where I heard for the first time that Autism had a cure. Kerri explained in a very simple way how I could cure my girl, even though I didn't have all the economic resources necessary for her treatment, the Foundation helped me with a part of the treatment, MMS and multivitamins. From there I started noticing changes, with a lot of sacrifice and effort my girl was on a strict casein, soy and gluten free diet. When I started giving her the MMS I would give her 72/2 on weekends and then I noticed that there was a lot of mucus in her stool, from there she started talking, she slept the entire night, and took naps in the afternoon, she pays attention, she knows how to count and all the colors with only three years of age. For me MMS is a miracle, since I started giving it to her she is a very healthy little girl, with no colds or any of the diseases she used to have, she is now considered to be very high functioning (14 pts. on the ATEC scale) with MMS, Vitamins, and the diet my child is getting out of the Autism spectrum. She does not go to therapy because of my finances but my sister Yenitze Montero interacts a lot with my girl, and her grandparents Raiza and Julian help me a lot with the diet, my nephews Jesus and Luis are also helping me with the home therapy and of course her dad and me fill her with love and care. My daughter is healthy, she has meaningful speech, she recognizes herself, she sleeps well with no pills and we are all very happy with the results. God Bless the entire team at Fundacion Venciendo el Autismo, especially Kerri Rivera for bringing MMS to my country, it is absolutely miraculous. Blessings.

Yurliana Montero and Jonatha Quijada

---

153. *Martin is 6 years old, weighs 19 kilograms and is diagnosed with severe non-verbal Autism. He was diagnosed when he was 2.*

He almost made no progress with therapy, when he turned 4 everything seemed to be lost, but when he was 5 we tried a new therapy and he started to say a few words. He also got very sick almost everyday, he coughed a lot, had an ear inflammation and fever, he couldn't go to school.

He started taking MMS the 29th of June 2011, when he was only taking two doses of one drop and two drops every third day has intestines were not moving and he had a psychotic attack (something that I've never seen before) and this lasted 2 hours.

Children with Autism are severely intoxicated so his reaction to the MMS was very clear. After 15 days of taking 8 takes of a drop a day Martin stopped coughing, he is more alert, connects all the time, his expression changes, he started to make friends, he starts to play with the computer, I hear from school that he'll be a preschooler because all the improvement he has had and that he didn't have all year. (I was almost going to get him out of school) he was attending school with a shadow teacher but had no apparent social communication or increase in cognitive capabilities.

This past October, in a meeting with his team of therapists, I asked them all what they thought about Martin's improvements and they were very happy, I asked 2 of his speech therapists who have worked with him for more than three years, "Did Martin have severe autism?" "Yes." They both responded. "And now?" I asked them. "Noooo" they both agreed. The rest of the team, who have only known Martin for a few months looked at each other, and said that they could not imagine Martin like that.

The path out of autism is scary, but MMS is giving us great joy and hope.

Martin has not been sick anymore, the last time he was sick, before MMS, we had to admit him to the hospital, his cough was so bad that he was vomiting non-stop.

First off we give thanks to God, we thank Jim, Andreas, my dear Kerri who guides me with all her love, and to every single mother that is in this struggle. I also thank Mr. Luis from Buenos Aires, Argentina who has made it easy for me to get MMS; he is a wonderful human being who helps those in need.

We are on the way to full recovery for Martin, I'm sure of that.

---

154. So I just redid my daughter's ATEC. She was a 71 a month and a half ago and is now a 60!!! I thought she was making nice improvements but it really set in when I saw her score. Some things that have changed for her are now she is pooping on the toilet every day and telling me. She used to just go every few nights in her diaper. There's no more diarrhea, she's not constipated, she's sleeping great. She is just starting to show signs of imaginative play and a big one for us is that she just learned to blow. She couldn't blow no matter what we did to try to teach her. Now she's got it. Hopefully more good things to come.

---

155. My now 8-year-old daughter was dx with autism at 18 months old. Started GF/CF diet and started with a DAN! She slowly made progress with speech, had a few words, a few 2 and 3 word sentences, and despite having autism, she was always a happy, calm, sweet little girl. Shortly after her 3rd birthday, she "changed.' She became more of a Jekyll / Hyde personality, and all progress in speech and improvements in eye contact and sociability started to fade. Seemed like everything we did was just like yo-yoing back and forth between improving, and then crashing yet once again.

Last year at this time, we had just left the doctor we had been taking her to. For years she was on antivirals, antifungals, antibiotics on and off, SSRI's, Tenex, and when her violent SIB had become so bad the second year into treatment with him, he put her on Abilify. SSRI's and the Abilify did nothing for her except make *her behave worse. Neither did any of the other things he had her on. The first year she was doing pretty good, but possibly because the antivirals were acting as an anti-inflammatory. She always had chronic constipation with loose stools/diarrhea. Nothing helped that either. We finally left that doctor who yelled at us in his office for 20 min straight, in front of my daughter, because he did not like that I had been challenging him on all of this for the past 6 months. I was no longer going to let him put the blame on us and everyone else... that her ABA and speech therapists didn't know what they were doing, and telling us she acted out this way because we didn't know how to give proper time outs or how to correctly give "pep talks" at bedtime. I was convinced that she had PANDAS/PANS and/or parasites that were causing her ever increasing SIB, loss of all progress she had made in speech over the past 4 years, and the fact that she had not grown in 4 years...did not gain 1 pound, feet did not grow at all, and she only grew 1/2 inch in height in that 4 year span. We had taken her to a very well respected endocrinologist the previous summer who ran all the tests. Could find nothing wrong as to why she wasn't growing. But he decided to put her on HGH anyway. Thankfully, we left and decided not to give her the HGH. We also took her to one of the best Ped Psychiatric doctors at UCLA who specialized in autism that summer. He observed her, her SIB happened right there in his office for him to see. He said her SIB is coming from pain, discomfort, something medical going on inside, most likely stemming from her gut. He suggested we get her scoped upper and lower. So we did. The only thing that came up was on the upper endoscopy, and that was that she had esophageal gastritis. So the GI dr gave her Nexium.*

All the parasite stool tests we had done over the years had always come back neg except for one at age 3 came back with Giardia. But that was it. So last January, when we had finally left that her doctor of 3 years, her liver AST and ALT were at 130 and 135, her Creatinine (kidneys) were severely elevated at a dangerous level, she now had hypothyroidism, and after testing her cortisol thru blood and saliva we discovered she was barely making any at all. She had dark circles under her eyes, grinding her teeth for 3 years now, and agitated beyond belief ALL the time. Mind you when we started with him 3 years prior, her thyroid, liver, and kidneys were all in normal range. She did have dark circles, many many food allergies, some SIB, but nothing compared what we were dealing with now, or what was to come.

So now, what to do. We took her to a very highly recommended Pediatric GI in Los Angeles who was very nice. Listened to what we had to say. Ran the most sensitive and comprehensive stool analysis. Came back with nothing at all. Then took her to the head of pediatric Infectious Disease at Cedars Sinai. She obviously had her mind made up before she even walked in the door. She saw the word AUTISM on my daughter's chart, and that was it. I showed her every blood and stool test, explained the not growing at all for 4 years despite all the endocrinology labs said everything related to growth was in normal range, her severe SIB. She said to us..."I know you want to help your child. You want to find something "medical" that has caused her autism so you can fix it. But the truth is, your daughter has autism. You need to accept that, and take her to a pediatric psych dr who can help her with these SIB's. There are lots of meds that can help her be calmer." Had she not listened to one damn word I told her!!! All the meds we had tried in the past did nothing, already took her to a psych dr who specialized in autism telling us something internal was wrong. She would not run any blood tests, or any other kind of Infec Disease tests. So with that, my *husband, my sister, and I completely dumbfounded started to walk out. The dr then had the nerve to say..."She hasn't been vaccinated since 18 months (she was 7 here) Let's get her caught up right now before you go. I can give her 6 vaccines today." I looked at her and said you have got to be crazy! I am not giving my very sick child vaccines full of the poisons that did this to her."*

So a month later we started with doctor who "gets it." Who listens, and treats parents with respect. Values their opinions. And the networking I had been doing for the past 6 months on a few different bio-med autism groups with other moms was paying off. The moms are the ones who KNOW!!! That's how I found my daughter's current doctor and how I found Dr Maile Pouls last June to help her with nutritional and metabolic healing. Dr Pouls ran a $100, 24 hour collection urine analysis, and we found out she had severe malabsorption, Ph was too Alkaline, she was catabolic, extremely electrolyte and mineral deficient, severely Vit C and D deficient, and had severe bowel toxicity. Working with Dr Pouls and her new doctor lead them to suspect parasites/worms. I was encouraged to look into mms, but of course I had heard nothing but bad things like it's bleach. I was very hesitant to try it, but I researched it, and sought out other moms on Facebook who were doing it. The toxins from the parasites had completely taken over her body and brain. She was SIB almost all day, every day by now. It was pure hell. We thought we were going to lose our minds it was so bad. We had stopped taking her anywhere but school and dr's appointments for the past 2 years because she would just suddenly out of nowhere, for no reason, violently freak out and there was no way to help her calm down. Her doctor said she's so toxic, and her body is so sick that she can't detox. That the mms will not only help kill, but will neutralize the toxins and help her calm down. So we started mms, and I could not believe how much calmer, happier, and more present she was in just a week. MMS is literally saving her life, and bringing her back to us. Before starting mms, we did just parasite meds like Alinia and Mebendazole. The parasite meds alone did not help her, even though she was dumping worms, but she was still completely psychotic, and getting more crazy and manic every day. Since starting mms she has dumped hundreds of worms, some 10-12" long, ascaris eggs, TONS and TONS of the shedded skins of the worms, tons of liver flukes, and hundreds of tapeworm segments. We had not done any PP yet. This had all been with MMS only! We actually started Albendazole for treating tapeworm about a month ago and saw immediate improvements with it. It really seems like this is her biggest beast right now to deal with.

These are her labs from before MMS, and after starting MMS:

2/7/2012 - one month after leaving scumbag dr of 3 years

AST = 79 ALT = 108 Creatinine = 1.24 EOS = 11.2 Sed Rate = 9

~ Started MMS 11/23/2012 ~

12/4/2012

AST = 73 ALT = 64 Creatinine = 1.32

1/15/2013

AST = 61 ALT = 60 Creatinine = 1.30

2/27/2012

AST = 51 ALT = 42 Creatinine = 0.92 EOS = 5.1 Sed rate = 12

Ref Range for AST is [15 - 46] ALT is [3 - 35]

Ref Range for Creatinine (kidney) is [0.60 - 1.20] Shows how well kidneys are working. Anything close to or over the 1.20 is considered to be of serious concern

*Ref Range for EOS (eosinophils) is [0.00 - 3.0] * High EOS are always seen with parasites*

Ref Range for Sed Rate (marker for inflammation) is [0 - 10]

Her doctor said her Sed Rate is likely high because when killing off pathogens and detoxing, inflammation will go up temporarily.

So I say never give up. Don't listen to doctors who don't listen to you. Kerri, thank God you saw those bottles of MMS at the clinic that day and asked what they were. And that you bought some out of curiosity and tried it. And that you selflessly have taken what you've learned and experienced, and shared it. Given so much of your time to help others, to help heal and recover their kids. You are truly an amazing person. We finally have the answers to our daughters autism after almost 7 years of numerous doctors and specialists, so many tests, so much money and valuable time wasted... and now we finally have real hope...real results. We know without a doubt she is going to get healthy, which will in turn give her the happy joyful life back that she once had before she got "autism." ~ xoxo

---

156. *Today was just wonderful! We have been on MMS since 12/11 and this is our second round of Ivermectin from our Dr, just started today. My son has been very rigid lately, but not today. For the first time in over 6 months we went to the park and had a great time (no stimming), then we were able to run an errand and pick up books from the library. (he usually gets upset unless we do one specific route in the car). Felt like a regular mom running kids to the park and one errand. So lovely!!!*

---

157. *Just a testimonial to MMS. For those of you new to the group and still unsure of MMS protocol I wanted to tell you about us. I was afraid at first, but we decided to try MMS. Within 3 weeks my son went from an ATEC of 36 to 18! That was amazing. Then we were afraid of the enemas. We said no way, we would just do oral MMS. of course after seeing others progress here we decided to do them. He would calm down! Then for months we agonized over the PP. We were afraid of it. We were scared of the meds, everything. Finally we decided to try it too. We are in the middle of our first PP and our son, who used to hit, kick, bite and scream at me over the smallest things all the time, has now been a perfect joy. When he gets mad it is over quickly and does not escalate into world war 3. Also these bumps he had on his face have cleared up along with his attitude! So if you are unsure of MMS like I was, here is one more testimonial of just how good the protocol really is.*

    Thanks Kerri, all the other moms that came before me!

---

158. *Dear Kerri and the readers,*

    I would like to say that we started giving MMS to my son Filip, autistic/PDD NOS, when he turned 5 years old, after having tried lots of things (diet, many different supplements, HBOT, some sort of behavioral therapy Teach and ABA), we have seen enormous changes already in the first month.

    I had heard about MMS from a mom of an autistic child. I started reading about it, but was not impressed by the comments I read on the internet. I contacted *a few mums that I knew from the Son-Rise programme in the US and they have reassured me that it works miracles.*

I was however very sceptical of MMS, as I have read so many "negative" comments on the internet. This is why I decided to start it myself for a few months to see what would happen to me. Also, I had to figure out how I was gonna dose MMS 8 - times a day, when I am working 10 hours a day and Filip is going to the kindergarden.

In fact, I was just feeling great when I started using MMS.

I wanted to start with my son when on holiday, but my husband displaced MMS bottles by mistake. I was so mad. I have only found them the last day when we were returning home. I have immediately decided to give one drop to Filip before we started to drive. I need to underline that my husband was very sceptic. Filip would at that time speak, but it would mostly be scripting and rarely would he say things in context. Otherwise he also had other autism symptoms- tantrums, no friends, low muscular tone, no drawing, obsessive occupations (in Filip's case watching cartoons and reading books), no sense of danger, wandering out, nose picking, scratching but, toe walk etc. But suddenly Filip started commenting everything in the car, asking questions, observing the things he would see from the car etc. Me and my husband looked at each other and my husband commented " Maybe it is working after all".

Filip's ATEC was 48 at the time we started. At start, my mother-in-law helped out in dosing and we were able to dose 8 times a day. Filip's ATEC fell to 32 the first month. Unfortunately, my mother-in-law read some horrible stories on the internet and the FDA's advice was detrimental for her decision to stop dosing Filip, even though she was seeing wonderful results, she was scared and said that she is only believing the official medicine, admitting that the official medicine has nothing to offer our son.

I was very disappointed, but continued to dose as much as I could, however not reaching the recommended 8 times a day. I was therefore only able to dose 4-5 times a day. I started telework on Friday's in order to at least get 8-10 doses in on Fridays, Saturdays and Sundays and 4-5 doses the rest of the week. My son was great in November (the volunteers working on Son-Rise with Filip were commenting on unprecedented success), but his ATEC only fell to 28, I would say due to low dosing the progress was slower. Filip was having some hard times in January, when he was more hyperactive. At that time we have also introduced the GcMaf injections, which could have added to Filip's hyperactiveness.

But since two weeks now Filip is back to his shape. He is so communicative, uses unseen vocabulary, amazes us every day, he started drawing without being prompted, communicates with everyone, was invited to a birthday from a girl at the kindergarden, started to dress without prompting, started to tell me:" Common mummy, hurry up, let's go." or "Where are you mummy? What are you doing?" or when he occasionally wakes up during the night " Is it day already? or " I should not go on the street, because a car can hit me" or " Daddy, you are not sad, you look angry" or just yesterday when he draw a picture in the kindergarden he said to his teacher " I want to show it to my mum". The list goes on and on and every day I have to thank the all mighty for having led me to Kerri and to the MMS protocols. I know we are on the right path to recovery and I am not scared of the future any more. And, finally, i just did his ATEC to check out his ATEC- it is 16!!!!!! LOVE IT! LOVE IT!!!!!!

Kerri, thank you!!!!!!!!!

159. *My daughter has had really bad SIB for over 4 years. We've been in the ER for massive goose-egg bumps on head, and she always had severe bruising on her tailbone and spine from dropping herself down, butt first and legs in a "W" position, onto the floor, and then would fling and flail herself into walls, tables, furniture, bang her head on all of the above and the floor, a rage like she was the girl from the Exorcist. Nothing could stop it. It just had to run its course...like a tornado. Comes out of nowhere, and it will disappear when it's ready to. Nothing can stop it.*

But after starting MMS, it was 80% better right away. Then after 3 weeks it started coming back. For a month it was really bad again, and we could not move up past 2 drops without her going crazy. Then after you had us switch to the HCl, she was remarkably better in about 5-6 days. She has had a few "episodes" in the past 2 weeks, but now they can last 30 min max, where before mms they would last anywhere from 15 min - 2 hours literally non-stop. The year before starting mms, there were times I had to lay her on the bed or couch and just sit on top of her, like on her butt so she couldn't hurt herself. I would just lose strength after about 30-45 min of trying to hold her.

160. *Hey everyone! Just wanted to give a positive report about the HCl activator. My son is 12, non-verbal, very hyper and lots of food intolerances. I got mine in the mail today, gave a quick 3 drop dose to my son before he went out the door to school, noticed he was a bit more compliant and calm than usual after school. have given 2 more doses since coming home and he has been doing his usual, but he is much more calm today than he has been in months!!! I gave him a snack and he took it and started to eat and I told him: can you say thank you mom? and he actually responded with a sorta grumbled "graaaa mooooom" which was to me "thanks mom."*

161. *Thanking God for Kerri Rivera and happy here too. My 15 y/o is in the chorus of his HS play. He has to do 5 costume changes, his own hair and makeup and several song/dance numbers, which he did perfectly on opening night! He is not your typical MMS/autism kid. He was pretty much recovered, starting with biomed at age 2.5, until age 11 when he hit puberty and he started getting rigid and anxious. We had some supplements going and were starting to look at stem cells when we found Kerri. He is not on the strict MMS protocol but takes it when he can and does PP and enemas each month. He is so much better since starting this last August! So thankful for all the info and support from this group.*

162. *Hi Kerri,*

Last week I contacted you regarding our 8 year old son and his focus issues and the disappearing act his mathematical skills have taken, and we've done our best follow through on your advice to up the number of doses to 16 per day. I visited my son's Integrative Medicine doctor this morning, who has been his biomed doctor for the last 3 or so years to do a live cell microscopy, and to see if he had any ideas about the focus issues our son has been having. I told my son's doctor in detail what we've been up to over the last 10 months (MMS Protocol, Parasite Protocol) and he took it all in.

When he looked under the microscope he was shocked to see that our son's nasty dysbiosis issues were gone! He told me that he's had almost no luck completely getting rid of dysbiosis in a case like my son's. The yeast beast is gone! He still has evidence of leaky gut but the drainage looks great. He was so excited, our 20 minute scheduled visit turned into 40 minutes as he wanted *to know all about what we did so he can find a way to use it in his practice. He also remarked that you must be one bright lady to come up with the protocol, but we already knew that. :-) He had us fill out an ATEC when we first visited him back in 2010 and our son was at 55. He dropped into the 30's when we started him on Carnitine but the MMS protocol has brought him down to the 6. The doctor also noticed the distinct lack of autistic behaviours, so it's nice to get some outside feedback to help confirm what we are showing on the ATEC. I asked him directly if he thought we should change anything at this point and was direct...don't change anything keep going. And so we continue on our quest for the zero ATEC. :-) About two years ago we did a chelation challenge test (DDI) and found our son extremely high in lead and so we proceeded with IV chelation at that time. He ants us to do a chelation challenge test to check the metals status just to rule out lead and the possibility that metals might be still at the heart of our son's lack of focus and more ADHD like symptoms.*

He wants me to send him some information on your MMS protocol. He attempted to use chlorine dioxide about 10 years ago and found that he couldn't get the results that you are getting. I'll send him to MMS Autism website but can I include your contact info (email)?

Thank you for being so patient with us and everyone. As you know full well when it involves your own kid the emotions come out and it's hard to remain logical.

---

163. Drum roll....

I have been dying to share...

Been off MMS for a few weeks...AND HOLY SMOKES we are ALLLLL GOOD!!!!!!!!!!!!!!!!

Summary...

My daughter—PANDAS—MAJOR sleep issues—fluctuate between diarrhea and constipation

Me—Acne—horrible—and nail fungus

ALL GONE. ALL.

NOBODY could EVER, ever, ever tell me why I broke out so much esp. with our squeaky clean diet.

First PP, OMG it was REALLY bad—worst ever—and became cystic—painful—so for SURE worse before better in my case BUT believed in protocol and KEPT GOING.

By month 4, GONE and yet to return

Fungus—gone in month 1, never to return

My daughter:

Well, you know, AMAZING!!!!!

YOU, Kerri, YOU helped me figure out FINALLLLLY after a zillion docs, WHY MY KID could NOT SLEEP! EVER! 7 LONG years... long long—exhasuting years—and it was PARASITES NO DOUBT in my mind! Month 1 of PP, literally like a light switch, SLEEPING —falling asleep ON HER OWN—(not with a circus routine to "get her to sleep") AND staying asleep ALL NIGHT LONG! UNREAL! This is HUGE!

AND bye bye PANDAS (Siver/Aloe nipped it, MMS maintained her gains, *BEAUTIFULLY!) SO I do believe in our Silver Aloe protocol for PANDAS, 100% worked for us! AND MMS. Both.*

AND since OFF MMS, my daughter is 100% pooping NORMAL—OMG first time in her WHOLE LIFE!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! (on MMS it was still fluctuating...)

AND she is eating things she would have NEVER ever touched! EVER! Raw veggies, nuts, trying EVERYTHING I make... unreal... this is HUGE too... the "healthier" you get, your tastes DO change... and one becomes less "picky"

I have learned that SOME probiotics can literally throw a child into a PANDAS rage--and also brought back MY symptoms—literally within 12 hours of consuming. (Like "Regarding Caroline" posted) This clearly happened to us. Mental note: NOT all prob's are appropriate for everyone. We react to fermented food—so I am going to experiment with no prob's and with THeralac __ONLY THeralac BTW

Now,

We are back to silver/aloe, daily, as prevention

We will still take DE (prevention)

Vit D3

Trace Mineral Drops

Enemas as needed

AND I am going to go BACK to the herbs for 3 months around full moon "just in case"—

I CANNOT WAIT to meet you in person, THANK YOU, and share our story with EVERYONE that will "hear me".

Again, my daughter did not have 'autism' per se, but autoimmune issues, meaning that, perhaps, her treatment protocol is 'shorter' then someone who is "more sick' etc. if that makes sense.

We also did a boat load of work prior to this. Meaning diet in place, for instance, for 4 years already—so healing began. This was the final layer for us. The frosting on the cake!!!!!

I have 2 brand new bottles of MMS in my cabinet :) AND will ALWAYS have it on stock!

With MUCH LOVE and respect to you Kerri.

-Jeni, IL, USA

---

164. *My daughter was born Sept 9 2010 and she had pretty good health until she had 15 months, she was able to say five to ten words by then, she had all her vaccines on time as recommended by our pediatrician, and she was really lovely and happy but suddenly every start to change slowly, we noticed she start losing the words she had learned, she use to wave people saying hi, and suddenly she was not interested on waving anymore and she start to isolate herself little by little to the point she stop being a happy girl, she barely looked at us and we started to worry about it, anyway, I believe the first thought every parent may have think in this very same situation is that her personality is different than other kids and is something temporal.*

When she had 18 months we went to the regular doctor check and they asked us to get the opinion of a development pediatrician since her social and language skills were behind the expected, and this started the worst nightmare ever.

It seems like every single development pediatrician has a waiting list huge enough to have more than a couple months totally scheduled, we called on May and the first available appointment was 4 months later, we called like 4 or 5 of the recommended pediatricians and institutions but it was all the same thing, but we finally got a pediatrician, not form the recommended list, but he made room for us in a couple of weeks and we were able to see him in a couple of weeks.

The diagnostic was mild autism, the recommended treatment was ABA therapy and the follow up was going to be 1 year later since he did not have may expectations of positive changes in the next 6 months, we were devastated by this, my wife was in denial, and I started our journey to recovery the very same day, I started researching and suddenly I found tons of information about autism and people who claimed to have recovered their children from autism.

I sent emails everywhere and I never imagined I would have such a great response from every person I wrote extending me a helping hand on all my doubts with the most sincere and honest recommendations, 2 or 3 days later I found Autismo2.com website and I saw all the videos available.

I bought 3 or 4 books on amazon, I saw many hours of recordings and seminars in Spanish and English, and I was amazed about some kids losing their diagnostic, but also about how they were affected by heavy metals, bacteria, viruses and all sort of things.

At the beginning I thought my daughter could not be sick, she has never had a single cold, how come? but a home heavy metal screen test gave us the first clue, and some lab tests confirmed the very same problems all our kiddos share, mercury levels on the roof, candida, yeast, you name it.

I started the GFCFSF diet the first weekend after her confirmed diagnosis and just one week later we started noticing changes, then I started with baths with bentonite clay, and noticed more improvements, we also started ABA on July, and little by little our daughter started to improve, later on we introduced MMS baths, Quinton water, DE, probiotics, and she improved every day.

I honestly believe that the early diagnosis and prompt reaction from our side played a big role on her fast recovery, It happened that we never canceled the first appointment that was made with the most recommended development pediatrician, by then we were looking for a second opinion and a new evaluation due to her huge improvements, we never imagined this pediatrician will say what.

Although my history with this disease or condition has been very short nobody knows how much pain, stress and suffering had caused to our family, we now enjoy our daughter more than ever when we see her playing with her old sister and taking away every single concern we had, we continue keeping an eye on her development but she is doing really well, I wish nothing more than every single mother I have hear on the discussion groups I follow have the very same results we have had with all the guidance from such smart and courageous parents like Kerri Rivera and many others that shed us the light on our darkest night.

I am more than grateful with all of those parents, I would do anything on my reach to help them, we need to help every single kid and family that goes through this disease to recover our kids.

UPDATE:

I'm very happy to forward the last ATEC from my daughter, today we came from her 2 year check up with the new pediatrician and she found her pretty well, she just told us" It's pretty hard to believe she was diagnosed with mild autism keep *doing whatever you are doing, since I would say she is acting like a regular kid for her age."*

I would say that keeping in mind that most of her ATEC score is on the Language section, for a 2 year old and from what we came from she is on her way to full recovery (Previous ATEC was 55, at this writing was a 23)

---

165. *I found Jim on the Internet 2 years ago, and I immediately started Jim's protocol for MMS. Unfortunately we were so loaded with mercury, so when I reaches 8 drops of mms, and my son 2 drops (after more than 2 months on slow upgrading) we were really sick, again. All the heavy metals that where released with mms could not get out of us. That´s when I found zeolites, and that was our way back.*

Now, two years later, my arthritis is gone, and now I know it was mms that killed the clostridium bacteria that I had been carrying since I was 13 (I am 52 now). 6 months ago my son fell off his bike and got an open wound. It was infected and some hours later the blood was black in his hand. Now I had a choice, to go to the doctor and give him antibiotics, which I knew would kill the bacteria but make him worse, or do it my own way. I gave him 2 x 3 drops of mms, and the blood was ok in some hours, I thought. Some days later he started to wave his hand. I thought he had eaten something else that he could not digest, but it got worse. After 4 weeks he was not only waving/flapping both his hands, but also his head. Then I started get suspicious. I tested him for bacteria (I use a Rayonex PS 10 for testing, bioresonance) and I found Clostridium botulism.

Now I had a choice again. I had a fully booked week, but his father came and took him home to his island, where they could use MMS every day and be in a calm environment. He got 2 x 3 drops of mms, for 6 days. After 3 days his father called and said "Eva this is scary, the flapping is getting worse". And I smiled and calmed him down, the bacteria are dying and the toxins are being released, I said. The zeolites took care of the toxins, and after 4 week he was ok again.

My lesson in this was, apart from the clostridium toxins, that I also found an imbalance in my son on Condrogenesis with the Rayonex. When I was 13 I stepped on a rusty nail. The blood was poisoned in the foot, but I did not dare tell my parents. After 3 days the whole leg up to my knee was black. I got antibiotics, of course. At 18 I was diagnosed arthritis in my knee. I think I had small amounts of bacteria left, as my son had, and the toxins from clostridium bacteria degenerates the cartilage near the infection. After some years one develop arthritis. I did, but it's gone now!

Eva, Sweden

---

166. *Drum roll please.......my son is 17 years old. He's been on mms for 5 months now, I'm so excited I don't even know, could it be 6? I think we started late July. He started with a 63.......he is now a 7. He went from 63 to 25 in just a few weeks, then a 13 a couple months later, now a 7. I feel like I'm cheating every time. I go back and look, argue with myself, nitpick....but whether he's a 5, 7 or 9, it's nothing short of a miracle because just one year ago, he was WAY over 100.*

The ATEC is not perfect, not even close. There are things that the ATEC does not show. All that is true, but this is incredible improvement that no doctor I know would have ever been able to accomplish. Been there, did that, spent the 425K+.

And no, it's not just MMS, but that has been the main treatment and everything else we do supports MMS and general health because I believe that in the end, *it's a battle between your immune system...and all that's attacking it. We are winning and I sure do wish they would STOP CALLING IT AUTISM. That alone hurts so many.*

Here's a little background on my son: Matthew was very high functioning from age 6-12. That was after diet since 3, floortime play therapy, some aba, AIT, IVIG, chelation for a year, OIG, years of infusions of glutathione, vitamin C, B-com, etc...immune stuff. Then he did very well GFCF ages 6-12, ran track (very well), 2nd degree black belt, boxing lessons, etc....Then puberty hit, and what I did not know about, Lyme and co-infections RAGED, and for around 3 years, I lost him cognitively, worse, worse, worse. At 15, he was like an advanced Alzheimer's patient, ADVANCED, he had lost ALL short-term memory, it was horrible, he struggled to think and could not.

He became more and more violent; choking me, kicking me, did I say he was a blackbelt? Even though he lost everything cognitively, like the ability to answer a question, those karate skills popped back into his head and he nearly killed me.... really close a couple times.

That was a year ago when his ATEC was over 100. So, here's what worked: 1. Low amylose diet helped with the constant urinating. (he acted like a diabetic and had 14 of 16 PANDAS symptoms so auto-immune) 2. Biofeedback helped calm him some. 3. Cholestyramine helped him psychologically - I could see him clear up mentally in minutes. (chemical/mold/dust sensitivities big time) By this time, his doc had him on Ketamine for pain. 4. PEMF (pulsed electromagnetic field device) stopped the pain in 6 weeks, got him off of Vicodin, which was not working anyway at double doses, and all the Ketamine. I hated drugging him. Got him off the psych drugs over 6 months time. Then he got a lot better but communication did not really come around until 5. MMS. BAM. Big improvements in communication and PERSONALITY. With MMS, even his laugh changed to a typical sounding teenager's laugh, and for the first time in years, he can sit still and stop pacing, pacing pacing....the parasites were eating him alive....and now more recently, I know 6. Hydro-colonics is taking us to another level of clear communication. Now there's lots of spontaneous language.

I write things on Facebook so that I can remember and maybe write a book one day. I don't care how bad things get, believe me, your kids can get better and it's amazing, they really do store up all this info that you didn't think was there. It is there, and they will share it with you one day.

I hope one day Matthew can share his perspective with others. It's just that it has taken a very long time for us to help him understand that the things he did back then were not in his control so I'm careful not to talk about it. For months he would out of the blue apologize and feel terrible, over and over. Then there's memories of his dad losing it also....protecting me.

Thankfully, Matthew has blacked out or forgotten a lot of it. He did say he remembers being locked in his room. We had to turn the lock around and sometimes lock him in there. Bad memories. We were all traumatized. But not anymore. He is a different person today but he's always been the sweetest boy I know. I knew that, even back then. One day, he'll probably talk more about it.

---

167. Hi Kerri,

First of all a HUGE big thank you for what you are doing with our children! My daughter has been on MMS two weeks now and she is like a different child!!!!!!!! Every day she gets better...By leaps and bounds. She WILL be one of your recovered ones ;)

UPDATE:

My daughter has now been on MMS for three weeks. ATEC down from 78 to 44. This IS a miracle. We have been in absolute desperation for almost three years trying to figure out what in the world is going on with our child. After hundreds of thousands of dollars, experts all over the world, LA, New York, San Fransisco... and nothing would work. Minor improvements but nothing that I could say really worked. Everyone always agreed that she probably does not have "autism" but what is making her behave that way?!?!?! So a long story short... If I could I would wrap my arms around all of you helping my daughter and so many others through MMS until I was sure you truly understand what your dedication means and that you are saving lives...literally <3 This is by far the biggest thing that has ever happened to my family. We are forever thankful.

UPDATE #2

I sent you my daughter's ATEC a week ago but redid it today. We started MMS in December and I decided I will do the test every month. Here are the results, 4 more points gone!!!! In 1 month from 78 to 40... Almost half off!!!!!! Her team is blown away ;) all we need now is that language score to come down... Kerri my goal is to have her recovered by A1!!!

Update #3

We are seeing such amazing improvements every single day. I think the true testament is the daily feedback from people working with her who do not know we are doing MMS. Here is part of a note from her SLP " her communication on the whole is truly remarkable- words, iPad, eye contact...everything is improving!!!!!!! So exciting. Has anyone told you lately that your daughter is a superstar? You should be so proud. :)" And I am...

Our main challenge has always been anxiety and today we had a breakthrough on that front. Instead of clinging to me for her dear life when I separate from her at school she took her friends hand, waived AND said goodbye and walked away... Here is the best part, she then turned her head, waived and gave me the biggest smile . Her face at that moment is forever engraved to my heart.

Also for the first time EVER my son was able to have a friend come over from school in our car without her having a major meltdown. Instead she shared her iPad with the friend to show him her program and use it TOGETHER!

Honestly I think our major challenge at the moment is for the rest of the family to heal with her... We had lost all hope and settled for life with Autism. We can now look at the future very differently and actually look forward to it! Love our new life

UPDATE #4

I have never really posted gains but this is too huge not to share. We started MMS 3,5 months ago and at that point I could hardly take my daughter to the grocery store. And even if I did it was a guaranteed meltdown. Well we just got home from a two week long vacation that included travel by air, time difference, different language, going from hot weather to freezing temps and piles *of snow, totally different foods, nothing but strangers to her, visit after visit to new places...The list goes on and on. The BEST part - the vacation was full of excitement, fun, laughter and happiness every single day   I have to pinch myself to believe this is my life after such a short time on MMS. Thank you all for your support and guidance   I have my life back and can't wait for the future now! We still have a looooong way to go but we'll get there.*

Today is a wonderful day already

---

168. MMS Works!

It's my number one pathogen fighter. Our family continually suffers from Strep, my son specifically from Pandas. MMS remains and continues to keep our strep symptoms at bay. I have used everything for our bacterial flares (Abx, herbals, supplements even homeopathics) and nothing has worked so effectively like MMS does for us.

Recently I changed over to the HCL activator and within in a few days we heard my severely apraxic son's first WORD "Mama" Its been just over a month on the HCL activator and he is now stringing two words together!! MMS is a huge part of our recovery protocol. I feel relief beyond measure that I found a post on MMS and decided to try it, and mostly I'm immensely grateful to Kerri for putting the MMS protocol together and for her love for our kids.

---

169. We just had a good meeting with all of Amor's teachers and therapists. In the last few months there has been big jumps in Amor's cognitive skills.

I used to worry that Amor might have dyslexia. I never pushed her to read even if was behind her peers. I know frustration will only make things worse. Her 'x' and 'k', 'b' and 'd' were always mixed up. I have been observing this mix up for the past 2 years. Early this year, her speech therapist agreed that dyslexia might be an issue and we were ready to start studying dyslexia management.

However, Amor now shows interest in reading both phonics and sight words. She can answer letter names and sounds precisely. The same speech therapist now says that we shouldn't look at dyslexia anymore. It might have simply been a mix-up as part of early learning. Amor's dyslexia issues in 2 years all of a sudden went away in 2 months??

Amor is also doing well with number recognition and improving in one-to-one correspondence. This is something her former speech therapist (back in Japan) was working on for a year...with little progress.

All these seem so simple but these are great strides from someone who has vision issues. Amor also wears glasses.

If I were to plot Amor's improvements with time, the biggest jumped occurred in the past few months, which of course can easily be explained by MMS.

MMS clears the gut of bad bacteria, yeast and parasites. All these affect sensory processing thus inhibiting the child's ability to understand her environment and people around her. Clean the gut, remove autism...in Amor's case, remove GDD.

We are also running a son-rise lifestyle program. I.e. we don't put fixed hours into her program but use the son-rise attitude and son-rise her in the weekends.

---

170. *Luca came into our lives July 26, 2010, he was premature, but all my kids were born in week 35, so I was not surprised. He was perfect, I have thousands of videos of him looking into my eyes, smiling, playing with his sisters, following daddy to the door when he was ready to leave to work and saying bye bye. He walked at one, and was a perfect little baby. It is hard to know exactly when he changed but I can tell you when I started to get worried, and saw him behaving differently. I had a neighbor the same age as my son, and when they got together to play I was surprised at how my neighbors son followed simple directions and looked into my eyes every time I called his name. When I had play dates with my friends all the moms were so relaxed, and while they were talking and kids were playing with toys, I was always chasing Luca and trying to keep him in the same room because he was not interested in toys or playing with other kids.*

My first trip to the Zoo with my three kids was a disaster, Luca didn't want to move from the Monkey's area and had his first melt down, all over the floor screaming and crying. I remember seeing all the families with kids the same age as my son, and seeing how they understood, and paid attention to what their mothers or fathers were saying. For me that was enough to feel for the first time something was not right, and it was not that I was a bad mother, I knew deep in my heart it had nothing to do with discipline.

I enrolled him in a school thinking that would help. The first day when I picked him up the teacher said they needed to move him with the babies because he was not ready for the 2-year-old class. That day I came home just knowing - and despite all my family, friends, husband and Pediatrician thinking I was crazy, we got him evaluated and the rest is history. On November 15th, 2012 it was official and we got the papers. Classic Autism, by that time we had read so much, watched hundreds of videos and searched every day on the Internet for interventions, schools, therapy etc.

I remember I wanted to start ABA the same day that I got the diagnosis, but the waiting list was really long and we would have to wait. So we waited and started an early intervention program in the Kent District. Surprisingly, I think we took the diagnosis really well; we just wanted to help him as soon as possible.

After the first month I stopped asking for a miracle or cure. I even asked all family and friends to do the same. I needed to accept it!!! My miracle was already here. My husband wanted to do the Gluten free diet and I was not even 100% in agreement...why??? He is such a blessing; let our miracle eat whatever he wants!!! I agreed to do the diet anyway, but wasn't really interested. I just had the feeling my husband was not "getting IT" or accepting autism = our new precious miracle. Now I can recognize and see clearly I was in total denial too. I was so afraid to let anybody know, even myself, that I didn't want this for Luca. I tried and tried and tried, but a voice inside of me was not in peace. I spent a lot of time crying and looking for recovery videos on YouTube, imagining Luca being one of them, and was fascinated with those stories.

Then one day I found Kerri Rivera's conference in Bulgaria, it was a "Godcidence," (like my father always says) I wasn't looking for MMS, I didn't know anything about that. I remember I watched the conference 3 times, and I felt butterflies for the first time since I got the diagnosis. It was like that voice was finally quiet and listening to every single word! "I want this."

I did my homework and found a lot of negative information about MMS but I did not care…that voice was so strong that I knew, the same way I knew about

Luca being autistic, that this was something I needed to explore. I looked for her in FB and to my surprise she answered me in minutes, when I asked her about the testimonies she connected me with some moms with recovered kids, REAL moms that decided to believe too. After I saw pictures from before and after and reading about their journeys I felt Hope!!! I knew deep in my heart this was the way and the path we should take. So I got my case together and when I was ready I talked to my husband about it.

It was easier than I expected, we wanted so badly to recover our son!!! It was not an easy process but finally after almost 3 months I GOT IT…WE GOT IT...THERE IS NOTHING WRONG TO WANT THIS, that is ALL we want for Luca!!!! For the first time I recognized my kid's autism was not a blessing, he was sick; GI problems, extreme constipation, food issues, sensory problems... and that voice (God voice) sounds really clear and loud; I NEED TO HELP MY KID TO RECOVER!!!!!!

On February 18th we started a Dietary Intervention with all Kerri's recommendations, the one diet that we were doing was full of sugar and carbohydrates, so we started to be really serious about all the food choices. We kept a very detailed food diary and ordered the MMS. By March 12th when our bottles arrived at our doorstep we were ready to start Kerri Rivera's Baby Bottle Protocol. March 13th we did our first bottle and that same day I asked Luca to look for a ball in the other room, and give it to me, and he did.

It was like he was sleeping soundly, and then he woke up. He was extremely alert and he was looking in a different way. At first I thought I was crazy, but then my husband noticed it too and the changes keep coming. He improved dramatically in all his therapy sessions, his waiting time, sitting at the table, responding to commands, pointing, eye contact, everything improved!

Kerri recommended doing an ATEC when we started back on February 18th, so I did, his score was 64 points. By April 1st he dropped 20 points! He now follows simple directions, plays with his sisters, recognizes them by name, hugs me and says mommy and daddy, has close to 200 words, and every day is a surprise for us.

I'm proud to say my son is on his way to RECOVERY AND HE WILL RECOVER, HE WILL BE RECOVERED!!!!!!!! I'm loving Kerri Rivera's Protocol, I'm loving my kid dropPing 20 points in his ATEC in a month, I'm loving searching his diaper, and inspecting his poop finding parasites, I know it seems impossible but I truly get excited! (Probably some mothers would understand) I'm loving this side of the picture FULL of hope. HOPE. I don't think I can leave this road now, I don't think I want my son to live in his "perfect" world and not in mine... I'm a fighter and thanks to Kerri I'm a believer ...THAT... is the blessing right there HOPE AND FAITH IN GOD IS ALL I HAVE AT THE MOMENT… Luca is really special, but it's not the autism that makeS him special, autism sucks!!! Overcoming it and all the obstacles he encounters every day, and doing what he is doing now is what makes him THE MOST SPECIAL KID ON THE PLANET, HE ALWAYS WILL BE SPECIAL TO ME NO MATTER WHAT, AUTISM OR NOT….I see it in his eyes now, he is a fighter like me, he is telling me what to do, recovery is soooooo in the picture!!!!!! I can actually see it!!!! I'm a HAPPY Mom!!!!!!!!!!!!!!!!!!!!!!!!

-Alma

171. *Today is a wonderful day already. My daughter got up to her alarm, took a shower without help, got dressed, packed her lunch, woke up her sisters, and is now getting breakfast for everyone ready. It amazes me how independent she is becoming, and how much our life is changing because of her new independence.*

She is 10 and we started MMS in Sept. She had done 5 PP and following the PANDAS protocol the last week after emailing Kerri Rivera last weekend with concerns about eye fluttering. She is also on Fish oil, garlic, Magnesium, and GABA. Before she was very hyper, spacey, a lot of language, but not using it correctly. She was extremely impulsive, and would fixate on object to play with and ignore others. Now she does reciprocal play, she talks back and forth with her friends, she cracks jokes that are now actually funny. She is focused with her school work and not hyper - except in typical situations.

This week my husband forgot to pick her up to take her to therapy and she came to my classroom to tell me how irresponsible he was, and that she was disappointed in his inconsistent behaviors. I had to laugh, but at the same time I was so proud as I explained to her the whole situation because we where having a fabulous discussion about life. She is doing so well. Thank you to all of you who have restored some of my hope that there is a cure.

---

172. *We are now in our 7th month of using MMS (CD) and in our 4th year of treating B.D., our son, who was diagnosed Autism. Our first week of MMS (CD) brought on a skill that we had spent years trying to accomplish, over night potty training. Week 1 of MMS (CD) and B.D. was out of pull-ups and in underwear. We weren't even close to full dose and already seeing positives. We knew we were on to something big at this point. One month later, we all witnessed my son blow out his birthday candles and open birthday gifts. Probably the first time he knew this event was a birthday and it was meant for him. As a gift to him, we took a trip to Disneyland. Not only did B.D. walk the entire park at our side, but he never once cried, whined, or objected to anything on the trip. Prior to this, he had to be in a stroller or wagon at all times.*

B.D.'s skills continued to flourish over the next few months. We enrolled B.D. in swimming lessons hoping he would feel comfortable without a life jacket on and clinging to an adult. This hope turned into a major accomplishment. B.D. has passed all Beginner levels in swimming and is now on to Intermediate. He is now swimming unassisted in the deep end of the pool! I used to hold my breath during his sessions hoping he wouldn't do anything that would scare the other kids in the pool. I still hold my breath, but now I'm at the end of my seat recording him as he's giving every single bit of effort to learn from his instructor and pass another level!

Just the other day, we went to a birthday party that was held in a warehouse full of screaming kids, loud music, inflatable jumpy houses and slides. This place can over stimulate just about anyone! In the past, my son would vocally stim, cover his ears, run behind the jumpers where the generators are (trying to turn them off) and require me, or his Dad to keep him engaged with the rest of the kids. Well, not anymore! He blended right in with the rest of the kids! Not a single stim, didn't cover his ears, stood in lines, listened and followed the rules and instructions, stood in lines and PLAYED with the other kids. His Dad and I were able to hang out with all the other parents, watching him run and whiz by with all the other kids as if that is how it has always been. Only it hasn't. He has come so very far in the past 7 months. We have a 6 year old son who no longer has sound sensitivity, loves to eat whatever is placed in front of him, DRINKS *tons of WATER, has manners, shows empathy, is careful and aware of other kids (especially babies and toddlers), making progress academically, and so very affectionate. B.D. isn't recovered… YET, but we expect it to happen now more than ever. We have to put our shades on because the end of the tunnel just got closer and brighter!*

---

173. *My husband and I were told to come to the 2nd grade assembly this morning. They gave awards to 2 kids in each of the 4 second grade classes. One was for outstanding in academics, and the other was for most improved in Self Control, which is the "character trait of the month" for the whole school. Kayla was given the award for most improved in Self Control in her class! She walked up and got her award and stood up there with the other kids holding theirs. Up until a few months ago, she would have violent, self-injury tantrums at school every day... throughout the day, lasting anywhere from 10 - 90 min. She could never get through a class singing performance without losing it and tantruming, screaming through the whole thing. Her teacher came up to me on Wed. and said she just can't get over how happy my daughter is lately, and that all the teachers and staff in the school have really noticed too. FUA!!!! I'm so proud of my girl!*

---

174. *Our third child (K) was diagnosed with moderate autism at 21 months of age. K was very ill. He suffered from vomiting and diarrhea mixed with severe constipation for years. At first, we began every intervention we could afford. The first step was removing all gluten, caseine and soy. We saw a decrease in stimming behaviour and an increased ability to learn. We implemented many vitamins and B12 injections, cranio sacral therapy, anti-fungals, HBOT and an even stricter diet removing all artificial colours, flavours, preservatives and most sugar. Our son was making slow but steady gains. ABA therapy continued from when he was two until just this past March (4.5 years). K was seeming much better but despite these gains, was never close to "normal".*

In July of 2012 a friend suggested MMS. I did my research and so it began. In three days K started to sleep through the night. After less than one week at a very low dose, K stopped the constant verbal scripting and obsessive speech-his most evident sign of autism. In that same week he began dressing himself independently. In the coming months, K's ATEC began dropping steadily. He began to make jokes, initiate social interactions, show empathy, cope with noise and busy places and best of all- understood safety. The risk of him running away was a memory. We could go places and he would stay near us. Teachers and Educational assistants were raving about his progress, unaware of the protocol we had begun. His most significant changes happened after a three months of the Parasite Protocol, using mms, Vermox, rompepiedras, neem and diatomaceous earth. Through enemas, he expelled worms as long as 24 inches to my husband and family doctor's disbelief. What K expels in his stool now looks normal and ordinary.

His ATEC in July 2012 was 46. Ten months later he has a score of 6.

He is undecipherable within his classroom and amongst his peers. He laughs, plays and acts like any other child his age. His academics are at par with his peers.

Our son does not appear to have autism anymore and only a trained eye would suspect that he ever had it to begin with. We thank Kerri Rivera, her Facebook group for all their support and MMS for being the final, and most significant leg in our sons' journey to recovery.

Maria A., Ontario Canada

---

175. *My severe autistic boy has been now doing wonderful!!! This protocol has completely changed my life!! He's been dumping tons of parasites since we started MMS about three months ago we're now in the second pp...he's now much calmer more happier!! His ATEC dropped down from 80 to 48 in three months only...that was truly amazing he is now gaining weight, I feel like I'm getting my child back to normal again!! I wish I could tell the whole world how Kerri Rivera has changed thousands of people's lives!!! God bless u Kerri for all your hard work to help others!!!*

---

176. *My son Nick started MMS just after his 7th birthday. Within 6 months he had cut his ATEC score in half from 66 to 37. Nick showed speech gains from the very first drop of MMS after having made no changes in speech for two years. Really, Nick hadn't made any improvements at all for two years. He was in a total stall until we began MMS. We raised Nick's oral dose of MMS very slowly taking two months to get to his target dose and never experienced die off that was unreasonable. Three significant changes are obvious in Nick's health at this point; he has been off of his Mitochondrial cocktail (for several years he has had blood labs pointing towards Mito dysfunction) for over 8 months and we haven't seen any of the lethargy that used to concern us so much, he was able to wean off of his seizure medication for his absence seizures and we haven't seen any return of seizure activity, and his thyroid appears to be beginning to shift from his long standing diagnosis of hypothyroid. The nicest changes we see are in the vibrancy of his health and how happy he is! His eyes sparkle and he has lots of energy for the day. He is no longer sitting on the couch staring out the window waiting for someone to figure out how to help him feel better. Since beginning MMS, every day Nick shows us just a little bit more of the boy that he is underneath the veil of autism and sickness. He tells us what his favorite things are, what he wants to do, and his sense of humor is coming shining through. MMS has given me a ton of hope that we just might get our boy back!"*

---

177. *After 7 years of biomedical treatments, numerous therapies, and more doctors than I care to think about, our son, who was diagnosed with autism, could speak a full sentence, was potty trained, and could finally go to the mall and movie theater without screaming and putting his hands over his ears.*

Victory??...yes...however, we knew we had so much farther to go. You see, as your child gets older, you realize that he is missing the key things in life that are crucial for independence. Our son had become "stuck" if you will, he was not socially maturing, and still lacked empathy, mature connections with peers, and proper communication with teachers and the other adults in his life. He would still yell when he didn't win a game, and he did not always react in an age appropriate manner, if something did not go his way.

My research went on as always...searching, searching, searching...we were not getting anywhere with his current doctor, and I knew that we had to change our current plan for treatment.

A good friend of mine called me to tell me about a new product that some of our warrior mom friends were trying. The letters MMS were spoken to me for the first time in March of 2012. My son was 11 years old and I was desperate for help. The first thing I asked was, "How much does it cost?" $25.00...what?? It can't be! We have spent our life savings, and now I am looking at something that cost less than a night out for dinner?? I decided to check it out!

I emailed Kerri Rivera and told her that I wanted to start the MMS autism protocol, and I'll never forget her quick response..."YIPEE", she said, let me know if you have any questions, I am here for you.

I read, studied, and called my friend about twenty times over the next couple of days, and then we began the next chapter in our journey.

About three weeks into the protocol, our son seemed calmer, as well as aware of others on a more mature level...we were excited! Once he was at full dose, we began, baths and enemas, and he continued to improve. Our next move was to try the 72/2 protocol, which is giving MMS every two hours around the clock for 72 hours.

Well, it was as if someone had kicked our sons maturity clock into high gear. We saw HUGE gains!

"Hey mom, I like your dress...you look really nice today"....WHAT??

"I'll try that new kind of chicken for lunch"...WHAT??

"I feel like I'm beginning a new life"...WHAT??

"It's OK that I didn't win the race, I had fun"...WHAT??

I wanted to call the Good Morning America Show and say, LOOK AT WHAT MY SON IS DOING!

He also began to grow and put on weight for the first time in three years! He put on 10 pounds and 2 shoe sizes in about 2 months!

The gains continue to come, and our boy is coming alive...recovering...finally it's here! My eyes well up with tears as I write these words...the words I have dreamt about for eight long years.

He celebrated his 12th birthday on January 3rd, and has told us that 2013 is going to be the best year of his life...YES...it is!

He is our angel, our fighter, our miracle. Thanks be to God for MMS!

We are blessed and forever grateful~

---

178. Once upon a time there was a little boy who had a strange mind, he hooted and he danced and he did not learn like the other children around him. So, he was sent to a special school. For his whole educational life he was separated, because he could not express correctly the answers put before him.

As he grew older he seemed sad and very much in pain. He appeared to try to stop the pain by clawing at his face and arms. He also stopped talking. We, his mom and dad, were very sad.

Then an angel came into our lives and told us of a magic potion.

"But" said the angel. "You must give him the potion every hour of the day and night for him to be better."

So the little boys' mom and dad have been doing that now for 9 weeks and 4 days. Every hour.

Now in those hours the mom and dad still cry. Some, because the little boy has been hurting for so long, and..... now they know why. But they also cry because they see that their little boy is not hurting like before. They are happy for his healing, and that he does not claw at his body any more.

And they are also happy because their little boy is talking again. And he is laughing. And the things that he knows surprise them. We thought he was not learning…turns out he was.

179. *Feeling really good still about my Natalie's progress from the combination of anti-parasitic medication, the supportive components of the PP and CD (HCL version). Natalie is about 63-64 lbs and she takes 2 drops of CD an hour both at home and at school. We don't wake her for CD in the middle of the night; but if she wakes for the restroom or something, then we give her another 2 drops (we just keep it pre-mixed in the bathroom near her bedroom). CD has been the single most important catalyst to Natalie's improvements/healing over the past 10.5 months; utilizing all three modes of delivery (oral, baths and enemas). Kerri having escorted CD into the ASD community is like a Miracle, in my opinion. Sorry... I don't mean to be corny nor impose my Beliefs on anybody else; but I sometimes find it hard to contain my Joy now and I get chills just thinking about all of the children now feeling better and/or beginning to feel better and better with CD. Before CD, it was hard to feel Hope for my special little Natalie; who was formerly suffering very extreme PANS flares and re-flares monthly as well as rapid-onset grand mal (tonic clonic) seizures every 2.5-3 weeks on top of her ASD and global developmental disabilities. Now Natalie's PANS 'flares' if/when they occur are waaaayyy less intense and better-controlled/'snufffed out' with CD than the former roller-coaster of oral antibiotics that her Drs had tried for years, her seizure activity is decreasing significantly and we've even begun a slow reduction in her anti-seizure med Lamictal ODT. We're optimistic that Natalie will be able to remain stable on less and less of that med, as we feel it might be causing her some nasty side effects. Our Hope for her is that she might even be able to come fully off of Lamictal ODT, as we continue CD and PP for her. Time will tell. While she's still a 'work in progress' [Note: She's 14 years old and been ill since infancy and it's a bit harder to heal longterm infection(s) and/or infestation(s)], we do see lowering ATECS in Natalie's future and brighter and brighter days. Patience is a virtue. So Grateful. Thank you Kerri and all you supportive Moms & Dads. More and more healing to come to ALL of our children!*

---

180. *Hi Kerri,*

Thank you for helping my daughter, age 15, recover from Autism. Thank you for the MMS/CD program and the support to carry it out. Here is a summary of her recovery:

June 19, 2012 Atec=52

Started with MMS slowly, adding one drop a day

GF/CF/SF diet

It took her 1 1/2 months to get to full dose.

Noticed improvement in physical health & behavior

Began parasite program

Sept. 18, 2012 Atec=24

Continued with 6-8 doses/day & baths

More improvements

School discontinued IEP, because she was doing so well academically and socially

January 15, 2013 Atec=19

My daughter, who is not vaccinated, comes down with Chicken Pox

No one at school or in community had it before her. It appears to be an isolated case.

It may have been in her system and released as parasites died?

Fever of 102 degrees for two weeks was the lead up to the Chicken Pox. Seemed like her immune system woke up! After chicken pox, huge gains! That is when we noticed she was recovered!

April 25, 2013 Atec=6

Yeah!

* * *

Please share your experiences!

Testimonials are one of the best ways to share your experiences with this protocol. Perhaps you learned about it by reading or watching a video testimonial?

If you don't tell us your experiences, we can't share them or take action on issues that need improving or correcting...

Send your testimonials to:

testimonials@cdautism.org

Also, let us know if we are free to publish your testimonial, with or without your name.

It's in our interest to take care of others. Self-centredness is opposed to basic human nature. In our own interest as human beings we need to pay attention to our inner values. Sometimes people think compassion is only of help to others, while we get no benefit. This is a mistake. When you concern yourself with others, you naturally develop a sense of self-confidence. To help others takes courage and inner strength.

~ Dalai Lama

Appendix 2

CD Autism Worldwide

Since September of 2011 over 5,000 families in 58 countries have used CD to help with an autism spectrum disorder. Here is a list of all the countries who have at least one family using CD to help heal the symptoms known as autism. This list is constantly updated at www.CDAutism.org.

- Argentina
- Australia
- Austria
- Belgium
- Bolivia
- Brazil
- Bulgaria
- Canada
- Chile
- China
- Colombia
- Costa Rica
- Cuba
- Czech Republic
- Dominican Republic
- Ecuador
- Egypt
- England
- France
- Germany
- Ghana
- Greece
- Hungary
- India
- Iraq
- Ireland
- Italy
- Japan
- Jordan
- Kenya
- Kuwait
- Malaysia
- Mexico
- Montenegro
- Morocco
- New Zealand
- Norway
- Panama
- Peru
- Philippines
- Poland
- Portugal
- Puerto Rico
- Qatar
- Russia
- Scotland
- Singapore
- Slovakia
- Slovenia
- Somalia
- South Africa
- Spain
- Swaziland
- Sweden
- Tanzania
- Turkey
- United Arab Emirates
- United States
- Venezuela

If you are using CD for autism in a country not listed above please email us so we can add you to the list... kim@cdautism.org.

Conformity is the jailer of freedom and the enemy of growth.

~ Robert F. Kennedy

Appendix 3

AutismOne on Healing Autism: Accidental Cure by Optimists

by Simon Yu, MD

Accidents happen. We accept the reality of them as is and we usually move on. On the other hand, accidental discovery is another story. For some people, they ask questions like, “why me?” Or keep asking for a deeper meaning: Is this really an accident or a message to understand? An accident is a fertile ground to find out if you are a pessimist or optimist.

Tom Jacobs, story teller from Kansas City, recently told a story about a pessimist and optimist: Two sales men were assigned to Africa in the early years of the shoe industry to sell their shoes. One week later, one salesman telegrammed to his boss, “Business situation in Africa is hopeless. Nobody wears shoes”. On the other side of the continent, the other salesman excitedly telegrammed to his boss, “Unlimited potential in Africa. Nobody wears shoes!”

I don’t know about you but I would rather be an optimist. I always encourage my patients to look at the bright side and be cautiously optimistic no matter how grim their conditions might be. Hope is a powerful driving force to promote healing. Pessimists do not call themselves pessimists. Just realists.

Today, I saw a 77 year old patient with a history of chronic pulmonary fibrosis, bronchiectasis and rheumatoid arthritis with a recent diagnosis of stage one bladder cancer. She has been depressed and joined a cancer support group because of feeling a sense of hopelessness.

After a long discussion, I told her how lucky she is that her lung and rheumatoid conditions have been stable and she has only a relatively benign early stage of bladder cancer. I told her if you are going to have cancer, this is the kind of cancer to have and she should be thankful. All of a sudden, her doom and gloom mood was lifted with a broad smile. The rest of our session was uplifting.

The AutismOne 2013 conference was held in Chicago during Memorial Day weekend. I was invited to give a talk on parasites, allergies, and autism. Autism is not my field. I usually see adults as an Internist. However, beginning in 2012, I have been seeing autistic children with rather interesting responses. In May 2013, I wrote an article about a medical hypothesis of a relationship between parasite infection and autism in preparation for my lecture for the AutismOne conference.

I attended a full day of lectures and was awestruck by the thousands of autism parents from all over the country attending the lectures. Multiple lectures were conducted by many autism specialists at the same time covering a large variety of topics. This conference was driven by a grass roots movement of parents with a hope to help their children. Here is a short synopsis of the lectures.

Dr. Anju Usman, MD from the Chicago area covered gut-brain connection and biofilms by pathogens in the gut. Symptoms such as depression, anxiety, poor attention and focus, and obsessive compulsive behaviors may be related to the delicate balance of bugs which produce a mucous slime known as biofilm.

Andreas Ludwig Kalcker, Ph.D. Bio-Physicist from Spain, who studied under German bio-physicist Fritz Albert Popp, Ph.D., discussed parasites in depth. He demonstrated a successful treatment for more than 65 children around the world with his parasite protocol over a one year period.

Kerri Rivera, from Mexico, one of the main leaders of the group running the AutismOne conference, showed many cases of successfully treated autistic children including her child. She has been using chlorine dioxide in conjunction with diet, nutritional supplements, detox, and hyperbaric oxygen.

By the time I presented my medical hypothesis on parasite infection and autism, they had already had numerous discussions about parasites and had been saying all along how parasite infection might be one of the major underlying problems for autism that has been overlooked. I felt like I was repeating what had already been presented.

Dr. Andreas Kalcker and Kerri Rivera collaborated using chlorine dioxide for two years with prescribed parasite medications, albendazole and pyrantel pamoate, to turn around autistic children. Professionally, I have no experience using chlorine dioxide on my patients. The difference with my therapy was that I was using acupuncture meridian assessment as a guide to detect and treat parasites.

If what they are reporting is even partially true for these autistic children, it would be a major breakthrough in the autism community. Some of the audience was crying in excitement but some were skeptical and saying it sounded too good to be true. I could feel the excitement of the audience but also a sense of fear that Kerri Rivera and Andreas Ludwig Kalcker will be attacked by special interest groups. Chlorine dioxide is too inexpensive to a fault.

Is this an accidental finding leading into an accidental cure for autism? It is too early to tell but any chronic medical conditions like cancer, heart disease, diabetes, arthritis, Alzheimer's dementia, or autism seems driven by epigenetic influences from environmental toxins, parasites, hidden dental problems, and faulty diet and nutrition.

I spent several hours with Andreas Kalcker after our lectures were over. We found a common ground for treating parasites: his experience as a bio-physicist and my experience as a military medical officer. Few people truly understand and are aware of the magnitude of parasite problems. This man was very passionate to rescue these children from the scourge of autism.

Kerri Rivera just published a book called [Healing the Symptoms Known as Autism]. I just finished the book. This book is for everyone but especially for both autism parents and all medical professionals involved in the care of autistic children. The book contains much important information that is not available in the main stream autism community.

Most pediatricians might be sympathetic but they are rather pessimistic regarding the care of autistic children. Andreas Kalcker and Kerri Rivera are true optimists in the midst of skeptics and pessimists. They see unlimited potential for the cure for autism based on diet, nutrition, and parasite eradications using chlorine dioxide and parasite medications. If you want to know more about autism and what is possible, I highly recommend reading [Healing the Symptoms Known as Autism] by Kerri Rivera. The book explains in detail the Kalcker parasite protocol.

Dr. Simon Yu, M.D. is a Board Certified Internist. He practices Internal Medicine with an emphasis on Alternative Medicine to use the best each has to offer. For more articles and information about alternative medicine as well as patient success stories, and Dr. Yu's revolutionary health book Accidental Cure: Extraordinary Medicine for Extraordinary Patients, visit his web site at...

www.PreventionAndHealing.com

...or call Prevention and Healing, Inc., 314-432-7802. You can also attend a free monthly presentation and discussion by Dr. Yu on Alternative Medicine at his office on the second Tuesday each month at 6:30 pm. Call to verify the date. Seating is limited, arrive early.

Appendix 4

Autism Evaluation Treatment Checklist (ATEC)

The next page contains a copy of the original *Autism Evaluation Treatment Checklist* (ATEC) developed by the late Dr. Bernard Rimland, the founder of the *Autism Research Institute*. We are not suggesting you use this particular copy. It is only included here for those not familiar with the evaluation so you have an idea what the ATEC involves since we mention it throughout this book.

The survey consists of a series of questions to be answered by a parent about what they are observing in their child's CURRENT behavior. It is not uncommon for both parents to take the survey separately and compare/average the scores.

The best way to take the survey is online at the *Autism Research Institute's* website. Here is the direct link to the current survey:

www.autism.com/index.php/ind_atec_survey

After completing the questions, a score is generated that can range from 0 to 180. 0 means no autism while 180 would mean that the answer to every question was in the most negative extreme.

We consider a child recovered from autism if their score is between 0 and 10.

You are encouraged to complete an ATEC before starting the Protocol, and every 3 months thereafter so you can clearly see the progress you are making.

Also, please share your results with us; we collect data on the children whose parents follow the Protocol as explained in this book. Send your name, date of the survey, general location (city, state, country) and score to:

atec@cdautism.org

If you have a history of ATEC results while on the protocol, please include them all.

NOTE: Do not send questions to this email address! You are unlikely to get any response from this address! It is just a collection point.

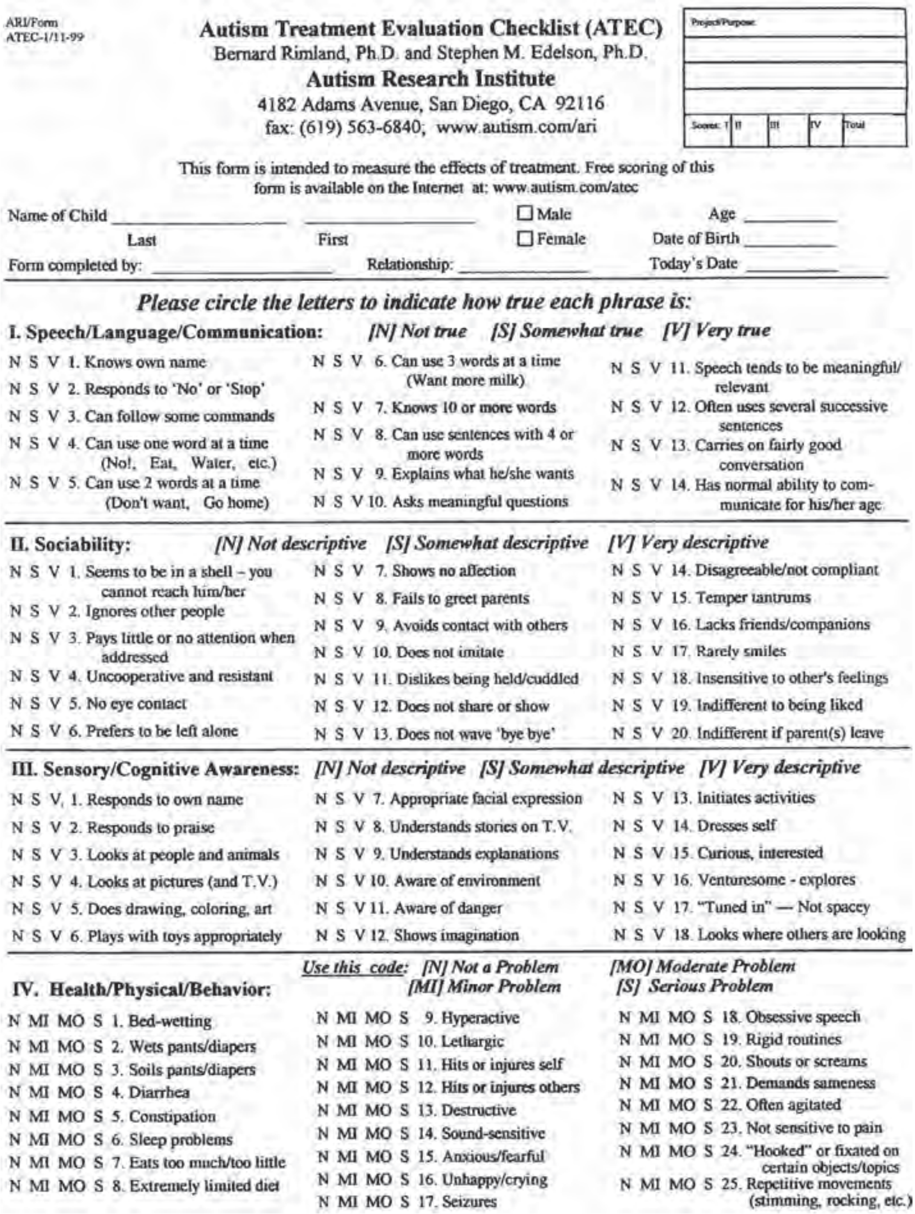

ARI/Form
ATEC-1/11-99

Autism Treatment Evaluation Checklist (ATEC)

Bernard Rimland, Ph.D. and Stephen M. Edelson, Ph.D.
Autism Research Institute
4182 Adams Avenue, San Diego, CA 92116
fax: (619) 563-6840; www.autism.com/ari

Project/Purpose:

Score: I | II | III | IV | Total

This form is intended to measure the effects of treatment. Free scoring of this form is available on the Internet at: www.autism.com/atec

Name of Child ____________ ____________ ☐ Male Age ________
Last First ☐ Female Date of Birth ________
Form completed by: ____________ Relationship: ____________ Today's Date ________

Please circle the letters to indicate how true each phrase is:

I. Speech/Language/Communication: *[N] Not true [S] Somewhat true [V] Very true*

N S V 1. Knows own name
N S V 2. Responds to 'No' or 'Stop'
N S V 3. Can follow some commands
N S V 4. Can use one word at a time (No!, Eat, Water, etc.)
N S V 5. Can use 2 words at a time (Don't want, Go home)
N S V 6. Can use 3 words at a time (Want more milk)
N S V 7. Knows 10 or more words
N S V 8. Can use sentences with 4 or more words
N S V 9. Explains what he/she wants
N S V 10. Asks meaningful questions
N S V 11. Speech tends to be meaningful/relevant
N S V 12. Often uses several successive sentences
N S V 13. Carries on fairly good conversation
N S V 14. Has normal ability to communicate for his/her age

II. Sociability: *[N] Not descriptive [S] Somewhat descriptive [V] Very descriptive*

N S V 1. Seems to be in a shell – you cannot reach him/her
N S V 2. Ignores other people
N S V 3. Pays little or no attention when addressed
N S V 4. Uncooperative and resistant
N S V 5. No eye contact
N S V 6. Prefers to be left alone
N S V 7. Shows no affection
N S V 8. Fails to greet parents
N S V 9. Avoids contact with others
N S V 10. Does not imitate
N S V 11. Dislikes being held/cuddled
N S V 12. Does not share or show
N S V 13. Does not wave 'bye bye'
N S V 14. Disagreeable/not compliant
N S V 15. Temper tantrums
N S V 16. Lacks friends/companions
N S V 17. Rarely smiles
N S V 18. Insensitive to other's feelings
N S V 19. Indifferent to being liked
N S V 20. Indifferent if parent(s) leave

III. Sensory/Cognitive Awareness: *[N] Not descriptive [S] Somewhat descriptive [V] Very descriptive*

N S V 1. Responds to own name
N S V 2. Responds to praise
N S V 3. Looks at people and animals
N S V 4. Looks at pictures (and T.V.)
N S V 5. Does drawing, coloring, art
N S V 6. Plays with toys appropriately
N S V 7. Appropriate facial expression
N S V 8. Understands stories on T.V.
N S V 9. Understands explanations
N S V 10. Aware of environment
N S V 11. Aware of danger
N S V 12. Shows imagination
N S V 13. Initiates activities
N S V 14. Dresses self
N S V 15. Curious, interested
N S V 16. Venturesome - explores
N S V 17. "Tuned in" — Not spacey
N S V 18. Looks where others are looking

IV. Health/Physical/Behavior:

Use this code: [N] Not a Problem [MI] Minor Problem [MO] Moderate Problem [S] Serious Problem

N MI MO S 1. Bed-wetting
N MI MO S 2. Wets pants/diapers
N MI MO S 3. Soils pants/diapers
N MI MO S 4. Diarrhea
N MI MO S 5. Constipation
N MI MO S 6. Sleep problems
N MI MO S 7. Eats too much/too little
N MI MO S 8. Extremely limited diet
N MI MO S 9. Hyperactive
N MI MO S 10. Lethargic
N MI MO S 11. Hits or injures self
N MI MO S 12. Hits or injures others
N MI MO S 13. Destructive
N MI MO S 14. Sound-sensitive
N MI MO S 15. Anxious/fearful
N MI MO S 16. Unhappy/crying
N MI MO S 17. Seizures
N MI MO S 18. Obsessive speech
N MI MO S 19. Rigid routines
N MI MO S 20. Shouts or screams
N MI MO S 21. Demands sameness
N MI MO S 22. Often agitated
N MI MO S 23. Not sensitive to pain
N MI MO S 24. "Hooked" or fixated on certain objects/topics
N MI MO S 25. Repetitive movements (stimming, rocking, etc.)

Appendix 5

MOLECULAR MIMICRY What It Is & How it Relates to The Gluten Syndrome

by Mrs. Olive Kaiser

After reading the short answers in *The Diet* Chapter (page 64), we understand (simply) the basics of how some of these departments of the immune system function. It is easier to grasp how, through molecular mimicry, gluten can damage so many different tissues in different people, and sometimes cause other foods that *look like* gluten to also be reactive. Before we go any further, I would like to thank the following professionals for their contributions to our family's well-being and a wider understanding of the gluten syndrome and other related topics. If you would like to conduct further research into this topic, the body of work that the following professionals have contributed would be a great place to start. There is so much more to learn.

Dr. Alessio Fasano, MD, shook America awake on gluten awareness in the 1990's and in 2003 published a landmark paper[1] on one small subset of the gluten syndrome, villi damaged celiac disease. That study provided the impetus that brought gluten awareness to the table (or sadly, OFF many tables) across our nation.

Dr. Thomas O'Bryan, DC, Functional Medicine, and international gluten syndrome educator - www.thedr.com. Dr. O'Bryan is my functional medicine doctor. He taught me much of what I learned about gluten and other topics and led me to his mentor, Dr. Aristo Vojdani. Later Dr. O'Bryan's amazing research review seminars on gluten contributed hugely to my knowledge base and bolstered my confidence to manage our new lifestyle in a then difficult social era.

Dr. Aristo Vojdani, PhD., MsC, Immunologist, CEO owner of *Immunosciences Lab*, Los Angeles, CA. Dr. Vojdani is also chief scientific advisor of *Cyrex Laboratories*, Phoenix, AZ. Dr. Vojdani has published nearly 150 excellent research papers.[2,3,4] They fit the community like a glove.

Dr. Rodney Ford, MD, pediatric gastroenterologist, NZ, grasped the extent of neurological injury,[5] and "silenced nerves phenomenon."[6] This compellingly suggests why many of us do not recognize symptoms of gluten damage until we are in deep trouble. This builds on the previous work of Dr. Marios Hadjivassiliou, Professor of Neurology, Sheffield, UK.[7]

Dr. Kenneth Fine, MD, gastroenterologist, Dallas, Texas, owner of an investigative research lab, *Enterolab* (www.Enterolab.com). His lab ran the only accurate gluten antibody tests our family received back in 2004.

Without the courage of these astute researchers, we would still be wandering in the dark. Thank you all, and other individuals who thought outside the box. Kerri you are one of those thinkers. Thank you!

Lastly, special thanks to **LuEllen Giera**, my support group leader, for leading me to Dr. O'Bryan and Dr. Vojdani.

1. What is Molecular Mimicry? How does gluten damage us?

The structure of gluten resembles the structure of many of our body tissues. When the immune system attacks gluten or partly digested "pieces" of gluten it may also attack body tissues that "look like" those pieces of gluten. There may also be other processes that we do not yet understand.

What is gluten? - Gluten is a stretchy protein found in some bread grains. The problematic types are found in wheat, barley, and rye, and now early research suspects possibly rice, corn and oats (Dr. Peter Osborne, DC, CCN, www.glutenfreesociety.org).

What are proteins? - Proteins are a class of materials found in living tissues, such as hair, nerves, enzymes, etc. Molecularly, all proteins look like necklaces of beads strung into various color sequences. The different sequences make the proteins different, and the "colored beads" represent 22 separate amino acids. Our digestive system uses enzymes to cut up these necklaces into single beads so they are small enough to cross the gut wall properly and be restrung into new proteins. See image on page 452.

Unfortunately, many folks today are toxic and poorly nourished and do not have strong digestion, so the gluten "necklaces" may never completely break down. As weaker bonds between the beads break, "pieces", called "peptides," of gluten are formed. Gliadin is a well known gluten piece and there are many others.

Our immune system - has "departments" to protect us in various ways, including IgA, IgG, IgE, IgM, IgD and others. IgE can cause immediate allergic reactions to bee stings or peanuts, etc. IgA, IgG and IgM may react more slowly with less drama. They all manufacture "workers" or "soldiers", called antibodies, each custom designed to patrol our bodies, looking for the bead sequence of one particular enemy. When they find that bead sequence, they "tag or stick onto" it. Our killer white blood cells interpret the antibody tag as a condemned sign and know to surround and destroy that protein.

If an antibody test lists "Gliadin – IgA", "Gliadin – IgG", and"Gliadin – IgM" it means the test checked for gliadin antibodies in the IgA, IgG and IgM departments. If the antibodies are high it means the immune system is does not like gliadin and is working to destroy it.

Weak, leaky barriers - Our gut wall and other barrier membranes such as the skin, lung, placental, and blood brain barriers are held together with "tight junction proteins" that act like velcro. Inflammation, parasites, gluten, medications, infections, electrosmog, etc, may damage the "velcro" or open them up too much. Substances may slip through them into places they should not be and cause trouble.

Unfortunately if the gut wall does not hold together well, i.e., is "leaky", pieces of incompletely digested gluten strings (and others) may slip through and run into the immune system—our dutiful "guard dog" on the other side. Due to their too-large size "he" may raise the alarm. The invader strings are "frisked out" i.e., examined. If they are rejected, one or more of the immune departments make matching antibodies to "tag" them so the killer cells know to go after them.

It is at this point more problems may arise. A gluten antibody may "run by" a natural body tissue, for example a nerve in the heart. It may see in that innocent nerve tissue a sequence of "beads" that matches or partly matches the gluten sequence it was designed to "tag" and stick to that section of the nerve protein instead. This, unhappily, attracts killer cells to the misidentified nerve, resulting in autoimmune injury to the nerve. This is molecular mimicry.

Other foods and proteins beside gluten may also initiate molecular mimicry when they prematurely cross a leaky gut, and may prompt allergies (IgE) or intolerances (IgA, IgM, or IgG). However specialists agree that gluten's particular "bead" (amino acid) sequences are uniquely guilty in their ability to upset the immune system and instigate molecular mimicry between "their" antibodies and our tissues.

Molecular mimicry can take place between non-food proteins and body tissues also. Infectious microbe sequences such as strep, for example, are believed to partially match, (in the case of strep), heart and joint proteins and so injure the heart muscle or valves, and (rheumatic) joints. Ditto is suspected for root canal and cavitation infections that may circulate and injure specific tissues, including the heart. Flu microbe sequences can resemble gluten and may trigger or surface gluten syndrome. There are numerous other examples of mimicry between infections, foods and body tissues.

Consider the image below. An antibody may be made to seek the shorter "bead" sequence of a gluten piece that crossed the gut wall, but it might also recognize a similar sequence, a partial match, in the longer nerve protein sequence. The antibody "tags" or "sticks to", the part of the nerve tissue that matches the gluten piece that it seeks.

<table>
<tr><th colspan="16">Gluten Peptide (and antibody) Sequence
Note: the different colors represent different amino acids.</th></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td>pink</td><td>blue</td><td>white</td><td>brown</td><td>brown</td><td>brown</td><td>red</td><td>red</td><td>green</td><td></td><td></td></tr>
<tr><th colspan="16">Nerve Sequence</th></tr>
<tr><td>orange</td><td>orange</td><td>yellow</td><td>green</td><td>red</td><td>red</td><td>blue</td><td>white</td><td>brown</td><td>brown</td><td>brown</td><td>red</td><td>orange</td><td>green</td><td>blue</td><td>yellow</td></tr>
</table>

It gets more complicated, but eventually the victim tissue is attacked by killer cells on an ongoing basis. This process may be slow but stealthy. In time, the nerve may malfunction and the organ it serves may also begin to fail; for example, heart failure. This is termed autoimmunity. The body attacks its own tissue.

Molecular mimicry can occur in hundreds of places in the body since gluten can break up into numerous pieces having amino acid sequences that partly match many of our natural tissues. This may explain why gluten syndrome presents so differently in different people.[2]

Once this process starts, like elephants who never forget, the memory B cells that manufacture the antibodies never forget the sequence of the "invader" gluten pieces. If gluten is removed, antibody manufacture eventually stops, but any time the immune system sees even a tiny amount of that protein sequence, antibody production is retriggered. This is why specialists insist the gluten-free diet is strict and lifelong.

Gluten removal stops production of the initial antibody, which decreases the inflammatory process significantly. Literature shows the gluten-free diet may reverse or improve many serious conditions, but it may not always arrest an advanced autoimmune disease. Prevention is key.

Example: Once a ball is pushed down a hill, further pushing will only send it downhill faster, but it will roll on its own. To correlate, molecular mimicry between the gluten antibody and victim tissue starts the ball rolling. Removal of gluten stops pushing the process, and the sooner the better. The longer the exposure[8], and to other factors as well in a triad of autoimmunity (a. environmental stress, toxins, infections, etc.; b. a faulty blood brain barrier; and c. susceptible genetics), the higher the chance of autoimmune disease.

Dr. Vojdani comments, "If it is detected early enough and steps are taken early enough the condition may be reversed. If any condition is advanced enough you can reach a point when simple removal may not be enough. Autistic children generally have autoimmune reactivity rather than full-blown autoimmune disease, which is why they show great improvement upon the removal of gluten."

2. Does gluten always damage the villi of the small intestine as the celiac story teaches? Many other tissues such as thyroid, pancreas, blood vessels, joints, brain, nerves, liver, bone, etc. may be involved in this disorder. Does all that other damage arise due to poor nutrient absorption from injured gut villi?

No, according to published research many researchers and practitioners now believe the villi are not always damaged in an autoimmune gluten reaction.[2] Where there is no villi damage, injury to other organs cannot be due to nutrient deficiencies caused by villi damage. Molecular mimicry provides a mechanism for direct autoimmune gluten injury to many other tissues and organs when the villi are fine. It is also possible that several tissues and organs, *including villi,* may be injured and then nutrient deficiencies from poor villi absorption may affect other areas.

Dr. Vojdani's abstract in his editorial, *The Immunology of Gluten Sensitivity Beyond the Intestinal Tract,*[2] supports that the villi are not always injured. To quote his editorial abstract, "Evidence has been accumulated in literature demonstrating that gluten sensitivity or celiac disease can exist even in the absence of enteropathy" [gut/villi damage], "but affecting many organs."

3. I do not have damaged villi, and my tTG/gliadin blood tests were negative, but I feel so much better gluten-free. Why?

The tests were likely false negative. That is very, very common. As gluten is processed in the digestive system, it can break into more pieces than we have tests developed to check them, and the immune system makes an antibody for each separate piece. The test probably missed your antibodies because they were different from the one or two the test checked. Many patients have proven this because they received negative results to the standard test, and then found a number of positive gluten related antibodies when they ran a more comprehensive panel. Cyrex Laboratories (www.cyrexlabs.com), runs 28 gluten related antibodies in 3 areas of the immune system.

Your properly performed* biopsy was probably negative because likely your gut villi are fine.** Your gluten injury may have targeted other organs or tissues, not the villi.*** For instance, if gluten has damaged the heart, snipping intestinal villi will not help find the heart injury. Damage in either place is not ok.

More detail on False Negative Tests

The blood and saliva tests - Standard celiac blood and saliva tests only check tissue transglutaminase IgA, (tTG – IgA) as an initial screener. Most doctors give up if that test is negative. However, the literature shows that standard tTG-IgA tests are only elevated when the villi are completely destroyed. Dr. Vojdani finds that form of tTG is often not found in other gluten injured tissues. Therefore the standard tTG-IgA test is a poor screener for most patients, and returns many false negatives. More forms of tTG have now been discovered.

Doctors might also order a standard test for deamidated gliadin IgA and possibly IgG. Unfortunately these tests also have a miserable failure rate for picking up gluten syndrome because many patients have other antibodies beside those two. The gluten "string" can break up into many other pieces, plus other "departments" of the immune system may react.

---

* If there IS villi damage it is possible for the biopsy to miss it if the damage is patchy or further down the duodenum than the villi samples are taken. Most good "gluten aware" gastroenterologists take a number of samples to try to avoid this possibility.

** It is possible that very early villi damage may slowly accumulate long before it shows up on a pathology report. An elevated IEL count, (intraepithelial lymphocyte, an early sign of inflammation) of the villi tissue indicates this process is underway. Ask the gastroenterologist to specify that the pathologist do an IEL count.

*** I am careful to specify the "villi of the gut". There may actually be gut damage, but not to the villi. There are many types of tissue in the gut, such as nerves, for instance. The villi may be fine, but it is still possible that some other gut tissue may be damaged by molecular mimicry or other processes we do not yet understand. For instance, if the nerves that control peristalsis (wave like gut motion) are injured, it could result in chronic constipation.

Villi biopsy vs. positive antibodies for diagnosis? - To add insult to injury, even if antibodies are discovered, many celiac, villi focused specialists may still insist that the villi biopsy is the final word for "celiac" diagnosis and they trust a negative villi biopsy over positive antibodies. Many folks have positive antibodies and suspicious symptoms, including improvement upon gluten removal, but no apparent villi damage. Their doctors assure them they can eat gluten! This is a serious mistake for many patients.

More on the villi biopsy – By now it is easy to understand why even properly performed* villi biopsies are only useful for the relatively small subset of patients in which the villi of the duodenum are injured. Most patients do not happen to have damaged villi. Their damage is somewhere else in their body, other organs, nerves, etc., or some enzyme or functional item has been damaged by gluten antibodies. Scientists such as Dr. Aristo Vojdani[2,3,4,] do NOT recommend villi biopsies for gluten syndrome diagnosis (there may be other reasons to scope the gut, such as tumors, etc). Dr. Vojdani explains that if there are elevated antibodies to gluten, the immune system is screaming "I do NOT want this substance! I am manufacturing antibodies to tag and destroy it." That is sufficient reason to remove gluten from the diet. It is less expensive to check an extensive array of gluten antibodies than undergo an endoscopy anyway, and is much less invasive.

4. Why do many gluten syndrome patients not only react to wheat, barley, and rye but also sometimes to other foods, particularly oats, milk, corn, soy, egg, coffee, sesame, yeast, chocolate and others.

These foods have similar amino acid sequences to gluten. Now that we grasp molecular mimicry, this is logical. The immune system may misrecognize them for gluten, causing cross reactions which may keep the gluten antibodies running high even on a gluten-free diet. Happily, this does not always occur. Cyrex labs, Array # 4, checks a list of cross reactive and gluten substitute foods, IgA and IgG.

5. The diet seems excessively strict? Why does it take so little gluten to start a reaction?

We understand that miniscule amounts of bee venom, or peanut, can trigger emergency allergic reactions, and very tiny medication pills can cause major effects on our bodies. This is also true of gluten reactions, both immediate allergic IgE, and delayed IgA, IgG, and IgM.[3] "Crumbs matter." The difference is that delayed reactions are not as obvious as the allergic reaction. They may go unnoticed for hours to decades, but may be a long, slow, serious process. Their lack of drama robs these reactions of the compliance respect they deserve.

6. The silent syndrome - Why do many people react to gluten, proven by antibody tests, but they have few or no warning signs, or seemingly unrelated symptoms they do not recognize or connect with their diet for a long time. Then they crash with something serious, often or usually autoimmune?

Gluten is famous for slowly injuring nerves[5] by molecular mimicry, and in many cases the nerves are silenced by that injury.[6] The patient does not realize there is a problem until the tissue or organ that those nerves supply begins to fail. Furthermore some areas of the body do not contain many pain nerves, so we may not feel the damage. Slow silent damage is understood in other diseases. Heart damage, cancer, and aortal aneurysms are examples of conditions that develop silently and then suddenly flare. Prevention is best.

For years, celiac literature recognized 2 neurological conditions, peripheral neuropathy and gluten ataxia. In the wider perspective of gluten syndrome, nerve damage may be one of the most important areas of injury.[5]

Dr. Rodney Ford, MD, suggests in his book, *The Gluten Syndrome: Is Wheat Causing You Harm*,[6] that this is primarily a neurological disease[5], injuring and in some cases silencing nerves, compromising the health and function of the tissues they serve. This idea was reinforced by an astute observation made by Dr. Ford of one of his patients, an elementary school child who had not achieved bowel control. After she went gluten-free, the problem resolved. Dr. Ford realized the child now recognized the signal to visit the toilet, and accidents were avoided. The nerves in the lower bowel apparently "woke up" once the antibodies that injured them disappeared. This is an interesting and logical theory to explain "silent gluten injury", and it fits the community. Here is an abstract of Dr. Ford's published paper which discusses a possible widespread neurological focus:[5]

The Gluten Syndrome: A Neurological Disease, by R.P. Ford

Hypothesis: gluten causes symptoms, in both celiac disease and non-celiac gluten-sensitivity, by its adverse actions on the nervous system. Many celiac patients experience neurological symptoms, frequently associated with malfunction of the autonomic nervous system. These neurological symptoms can present in celiac patients who are well nourished. The crucial point, however, is that gluten-sensitivity can also be associated with neurological symptoms in patients who do not have any mucosal gut damage (that is, without celiac disease). Gluten can cause neurological harm through a combination of cross reacting antibodies, immune complex disease and direct toxicity. These nervous system affects include: dysregulation of the autonomic nervous system, cerebella ataxia, hypotonia, developmental delay, learning disorders, depression, migraine, and headache. If gluten is the putative harmful agent, then there is no requirement *to invoke gut damage and nutritional deficiency to explain the myriad of the symptoms experienced by sufferers of celiac disease and gluten-sensitivity. This is called "The Gluten Syndrome."*[5]

7. Why do so many of us react now, when for centuries most people appeared to be fine with wheat? After all, wheat is spoken of positively in the Bible and many other historical documents.

There may or may not be a conclusive answer to this question, but a few factors may play a role.

a. Many folks have higher toxin levels now, their nutritional status is worse, and digestive strength is weaker.

b. Today's wheat is different. Gluten grains have been subjected to a lot of changes, some genetically violent, according to Nina Federoff, a pro GMO scientist. She asserts in her book, *Mendel in the Kitchen*, that gluten grains were altered with nuclear radiation and chemical mutation by the 1950's – 1960's.[9] Recently a stash of old blood samples from that era, stored in a freezer by the military, were checked with standard celiac antibody tests. The incidence of positive antibodies was much less in those samples than is typically found in the general public today.

c. Wheat seed is sometimes treated with mercury. Might this play into unintended results?

8. Why do specialists and researchers insist that the gluten-free diet must be life long? Can't we heal the gut and go back to our beloved wheat bagels, croissants, and brownies?

Our scientists insist that gluten-free is a strict lifelong commitment. The memory B cells in the immune system never forget what the enemy "looks like." Today's gluten is a violently altered substance according to the very scientists who defend genetic alterations.[9] It is not worth playing with today's wheat. There is something strange and unpredictable about it. Even if a leaky gut has healed, future circumstances, toxins and stress might injure it and retrigger the syndrome, perhaps silently. Researchers say gluten creates leaky gut for a few hours in everyone.

Particularly since gluten appears to be the "bad boy" that predisposes to other intolerances, I prefer to walk away for life and concentrate on other nutrient dense foods. However, it is wise to consume whole foods rich in B vitamins, (ex., liver or bee pollen), and silica (the herb horsetail, "equisetum hymale" species, is a source,) as gluten grains contain those nutrients.

9. Traditional peoples soaked and/or sprouted their wheat berries and then made sourdough bread with it. Does that process alter the gluten sufficiently for gluten syndrome patients to safely consume it today, particularly organic spelt and einkorn?

No. These methods make bread more digestible, but the fermented products and ancient gluten grains still contain gluten and can trigger reactions in research trials. Even if there are no visible reactions, silent injury cannot be ruled out without long term research. Additionally, the preparation process before fermentation is completed is definitely gluten based, so significant cross contamination issues come into play in the kitchen. Healing the gut is a challenge. Other areas crop up that need cleanup too, such as the parasite issues. Gluten specialists advise us to go gluten-free, stay that way, and move on.

10. What are gluten withdrawals?

Rarely, a few days to a few weeks after going gluten-free, or after being glutened, a patient may experience a few hours to a couple of weeks of a parade of varied and unusual symptoms including dizziness, black pit depression, crying, physical or emotional exhaustion, even as in difficulty getting up to use the rest room, and other odd symptoms. In severe cases they may experience an inability to socialize, make eye contact, make decisions or hold a normal conversation. Children or teens may act out in extreme ways during this situation. Often patients cannot bring themselves to discuss this experience afterward. However, a patient reluctantly described her experience 2 years later as "encountering an empty white board with nothing on it." The rest of life around her seemed to be "across the Grand Canyon." This appears to be a temporary crisis that resolves anywhere from a few hours to a couple of weeks or so according to folks who contact me about them. It is assumed that this phenomenon is due to particular pieces of gluten strings called gluteomorphins with amino acid sequences which resemble opiate drugs. When these gluten pieces disappear from the blood stream the patient may experience "withdrawal," very much like a drug withdrawal. Another theory for why this may happen involves changes in blood flow to the brain that may create a temporary neurological crisis. Autistic children may suffer these withdrawals and may take longer to stabilize, but they usually make nice gains after the crisis passes. Happily, once withdrawal is over the patients are usually much better, and they are VERY vigilant with their diet.

In the rare event that this type of reaction might occur, family and friends of the person do well to understand that the person may (or may not) be able to prepare their own food, for instance, but not be able to verbally communicate much, make eye contact, hold a conversation, answer questions, and may be uncharacteristically snappy particularly if others attempt to communicate with them. Family members may wish to be quietly and unobtrusively nearby until the person passes through this stage. If the patient has children to care for, help may be needed, and also solitude, rest, and simply prepared "real" nourishing food that is gluten-free and easy to digest such as bone broth. Probiotics or old fashioned fermented raw veggies or sauerkraut from the refrigerated section of the health food store may be helpful. Family and friends should not take the person's temporarily withdrawn personality personally.

For details on these unstudied rare reactions, or if you need support during a crisis, see...

**www.theglutensyndrome.net
/Adverse_reactions.htm**

...or contact me at *jka8168@sbcglobal.net*. I collect testimonials, so feel free to contribute if you have experienced this type of reaction. We hope these reactions will eventually be studied.

11. Should I replace all the gluten foods I routinely eat with gluten-free substitutes?

No, most gluten-free substitutes are still mainly expensive processed food (aka junk food). For the first few weeks it is normal for newbies to look for substitutes to replace their "old gluten friends" and it helps them make the very real emotional transition, but there are better, more nutrient dense food choices. The substitutes are still high carb, (potato, corn, tapioca, rice), and most contain sugar, processed gums, GMOs and ingredients that do not help us get well. We may actually eat more of them BECAUSE they are gluten-free. Additionally they are expensive and not always easy to prepare at home.

Many experienced gluten-free folks gradually wean themselves off a constant diet of the rather junky substitutes in favor of other real unprocessed foods, such as meat, eggs, veggies, fruits, nuts, fermented vegetables, bone broths, and so on according to their digestive abilities. The gluten-free substitutes become the occasional treat, such as pizza crust. Gluten free families are wise to find the healthiest GF versions of just the items they miss the most, and skip the rest. For example, at our house, I had served gluten spaghetti for years, but when I dropped that habit, no one noticed. However, my husband wanted his breakfast toast, so we found a gluten free brand he liked and continued that tradition. Each family works out these adjustments for themselves.

Lettuce wraps are crunchy and yummy instead of sandwiches. Hamburger patties without the bun but with onion for the onion lovers, are still tasty and they feel better afterward. In our family we now use watermelon or a watermelon basket for birthday "cake", complete with candles and a bow in the summer. Fruit pizza with a nut flour or nut/date crust, avocado chocolate or other dairy-free or fruit based pudding for the sauce, and an artistic fruit/coconut topping works for parties. There are lots of ways to celebrate happy healthy gluten-free, junk free birthdays.

12. Warning! Formal gluten challenges for testing purposes – risky!

A word of caution. Antibody tests must be performed while the patient is still consuming gluten or shortly after going gluten-free. Sometimes doctors advise patients who are already gluten-free to go back on gluten, 4 slices a day for 4-6 weeks to restart the antibodies for testing purposes. Patient experience has shown that once the system is fairly clean of gluten, going back to perform a formal gluten challenge for testing purposes may be risky. The secondary reactions for some individuals can be significant, particularly neurologically. Challenges have been known anecdotally to trigger psychological black pit crashes, fibromyalgia, and other organ injury. Autistic children may have a hard time when they just go gluten-free initially or consume gluten accidentally. A planned challenge for them may be very unwise.

13. Intermittent infractions (aka cheats) are seriously unwise and may increase injury.[10]

This is not a fad, or cheater's diet. The gluten-free diet is a medical diet to treat or control serious autoimmune, inflammatory, and often neurological diseases. There is no room for casual infractions. Research suggests that repeated, intermittent cheats, even every few weeks, over time may actually influence mortality rates.[10] This is not a reason to avoid the diet, but to take it seriously. Once the strictness of the diet is embraced the patient or family adjusts and discovers that it is doable. Gluten-free food bars, nuts, a packet of gluten-free soy sauce, and GlutenEase™ or other brands of DPP IV enzymes are good to keep in the glove compartment or backpack to handle emergencies. NOTE: DPP IV (pronounced DPP 4) digestive enzymes help break down gluten but they do not stop a reaction and are NOT a reason to cheat. However if there is a possibility of exposure it makes sense to take them for whatever help they might afford. In the case of a confirmed gluten exposure, take the enzymes, keep the bowels moving, stay calm and deal with it. Worry and drama makes everything worse.

14. Discrepancy – The celiac focus uses villi biopsy for diagnosis. The wider perspective relies on positive antibodies or improvement on the diet. Question? How did villi damage become the gold standard?

The first place the medical profession found conclusive damage by gluten in the 1950's – 60's was to the villi of the gut. After the endoscopy tool was developed, villi damage and subsequent healing upon gluten removal could be observed. They concluded that gluten damaged the villi. This was true. The particular subset of cases they scoped had villi damage, but their conclusion that the villi were the only target of damage for everyone with gluten syndrome was too narrow according to research today.[2] In most patients the target damage was NOT the villi, but the bones, joint lining, heart, thyroid, pancreas, liver, brain, almost any organ, blood vessel walls, nerves almost anywhere in the body, and so on. The patients might even have injury to multiple tissues, BUT NOT NECESSARILY ALWAYS TO THE VILLI. When the villi biopsy was declared the gold standard for gluten syndrome diagnosis, it cut out most of the patients who were reacting to gluten. Snipping villi does no good if the damage is in the thyroid or brain. For the next 60 years very few patients were prescribed a gluten-free diet because most of them did not have damaged villi, (or the doctor never thought to look at all). Their gluten-induced injury was somewhere else in their body, so they were never diagnosed.

15. Discrepancy - The villi damaged celiac disease story teaches that celiac disease is autoimmune and much worse than "non autoimmune" non-celiac gluten syndrome (NCGS). The wider gluten syndrome perspective teaches that both are autoimmune and serious.[2,3,4]

A significant disagreement exists over the autoimmunity of non-celiac gluten syndrome (NCGS). In the beginning, antibodies to gluten could not be found in NCGS patients, therefore it was assumed that NCGS was not autoimmune. However, Dr. Vojdani insists NCGS IS autoimmune. The NCGS patients have plenty of antibodies, just different ones than the standard tests check. His tests, which check and find more antibodies,**** the illnesses these patients develop, and recoveries or improvements on the diet all prove his point. He also asserts that NCGS can indicate a gut wall in worse shape than the celiac villi damaged subset.[2,3,4] The damage is simply somewhere else, not to the villi. See the link below for diagrams of these reactions, and compare the condition of the gut wall in the celiac diagram with the gut wall in the gluten intolerance diagram.

**www.TheGlutenSyndrome.net/
VojdaniDiagrams.htm**

**** Dr. Vojdani's research found several more forms of both tTG, and gliadin antibodies, (alpha, gamma, omega), gluteomorphins, glutinins, and others. His saliva and blood panels check 12 separate antibodies in 3 different immune departments and 2 mediums, totaling 28 gluten related antibodies. The two panels, run together, nearly always find antibodies if they are present in the patient, translating to far fewer false negatives.

> This disagreement fosters confusion among NCGS patients that villi damaged celiac disease is the "big bad boy" to avoid (true, it is bad), but that NCGS is less severe, not autoimmune. This translates in real life to gluten birthday cake for the Friday night party and then back on the gluten-free bandwagon Saturday morning, "so I won't eventually develop villi damaged celiac disease." Yes, that could happen also, particularly if the person seesaws on the diet, but the notion that NCGS is not as bad as villi damaged celiac disease is a misunderstanding according to Dr. Vojdani.[2,3,4] To repeatedly seesaw off and on gluten indiscriminately is unwise according to medical literature,[10] and community experience. Gluten exposure needs to be an accident, which happens occasionally even to the most vigilant. Casual cheats are a more risky mindset, usually meaning "more frequent."[10] Take it seriously, nerves, blood vessels, and organs are precious.

Gluten Tests with a Good Track Record

Some families cannot afford to test or good tests are not available. In the case of autism this protocol and most others require gluten-free and that is the end of the matter. Kerri takes this position partly due to cost and the miserable record of false negatives that standard tests return. The protocol does not work without the diet. Period.

Some adults are also willing to go gluten-free without a test. Their bodies tell them what they need, they listen, and are happy to find answers. Social pressure does not sway their decision or ability to comply.

In many other situations, a positive test is very meaningful. It gives patients confirmation that gluten-free is right for them, silences critics, and helps them comply with the diet. In the case of children it provides proof to the other parent, grandparents, therapists, doctor, etc. It may also come in handy if the child turns into an invincible teen who doubts he ever needed the diet in the face of pizza and beer, ditto for his future spouse and in-laws. The catch is that the test must be adequate. A false negative misvalidates the skeptics. These are decisions every individual or family must evaluate for themselves. Thankfully there are test panels with good track records now.

Absolutely, regardless of any test result, if the body is able to communicate a reaction to gluten or its removal, then that is the final answer. Be grateful and go gluten-free.

This testing section is for those who wish to test for their own confirmation or social support.

NOTE: If a patient has been ill for a very long time it is possible for the immune system to be so worn out that few antibodies are manufactured. Lower antibody counts might show a false negative but not prevent injury to innocent tissues. Also, due to the wide variety of antibodies a patient may happen to make, it is possible to run any antibody panel and miss those particular antibodies. The more antibodies that are checked the less likely this may occur, but should be considered in the event of a negative test that the patient or practitioner questions. In this case a gene test may be helpful.

Enterolab - Stool and 1 part gene tests

For 10 years Enterolab's mail in home collection research stool test stood in the gap for thousands of patients who received false negative standard tTG and gliadin blood tests. It has saved many lives, and gave our family the social confirmation we needed. This unpublished research test checks stool for tTG-IgA and gliadin-IgA only. The use of stool as the testing medium appears to pick up those antibodies most of the time, much much more often than standard tTG or gliadin IgA blood tests. However, since only two antibodies are checked, it may miss in some cases.

Dr. Fine's lab also offers a one part gene test which he believes is adequate for a reasonable price. Dr. Fine finds if a gene is present, nearly always so are the antibodies, and two genes are worse. Villi damaged celiac specialists recognize only HLA DQ 2 and 8 as gluten related, but Dr. Fine includes 1 and 3 and their subsets 5, 6, 7 and 9. In fact according to Enterolab, HLA DQ 4 is the only DQ gene that does NOT correlate with gluten syndrome. According to him, a patient needs 2 copies of the HLA DQ 4 gene to miss the predisposition. This translates to 81% of the Caucasian population with a predisposition to trigger gluten syndrome at some point in their lifetime, including before they are born.

Cyrex Labs - Better blood and saliva tests

Cyrex Labs (www.CyrexLabs.com) opened their doors in 2010 in Phoenix, Arizona. They run much more complete antibody panels designed by Dr. Aristo Vojdani, PhD., their scientific advisor, an immunologist, and researcher. Dr. Vojdani found a wider variety of gliadin antibodies, (alpha, gamma, omega) and variations of tTG in other tissues, plus gluteomorphins, and several others. He also checks an IgM antibody due to possible malfunctions in that system. Cyrex blood and saliva panels, Array #1 and #3, combined, test for 28 gluten related antibodies between 3 immune departments, 2 mediums, plus IgA insufficiency. They rarely miss a diagnosis because they look for so many antibodies. (Note: Vojdani believes stress, toxins and infections, i.e. environment, can trigger a gluten reaction without the genes.)

Gluten Free Society - Two Part Gene Test

A third approach espoused by Dr. Peter Osborne, *The Gluten Free Society*, Sugarland, Texas, (www.glutenfreesociety.org), is to run only the gene test, since any antibody panel may theoretically miss the particular ones the patient may have. He uses a 2 part test, the most complete method, and looks for both celiac and gluten sensitivity genes. A positive gene test does not prove a current immune response as do antibodies, but predisposition to it. Presence of 2 genes indicates a more severe case. Gene tests have an advantage in that they can be run anytime, gluten consumption is unnecessary, and depending on the results, useful information can be gleaned for immediate and extended family members.

Elimination Diet

The elimination diet is inexpensive. Often/usually it demonstrates improvement upon removal of gluten, or worsening upon reintroduction. Many "heads up" practitioners accept this as reason enough to go gluten-free.

There are occasional complications or interpretation issues for the elimination diet as follows:

a. Reintroduction of gluten (gluten challenge) can trigger stronger, sometimes risky reactions.[10] Stop a challenge upon negative symptoms, including depression and emotional instability, or best, don't challenge. An accidental infraction may come up that provides insight.

b. Occasionally it takes several months to see the difference, or a silent reaction may mask symptoms.[6]

c. If the patient later decides to test and is already "clean" on the gluten-free diet, blood tests will not work unless gluten is reintroduced for many weeks.[10] No! Stool/gene tests are safer.

d. There is no lab confirmation to silence naysayers.

Other Related Tests

Cyrex Labs Array # 2 - Intestinal Permeability Panel focuses on specific causes of leaky gut. This helps strategize treatment, and is an improvement over the old lactulose/mannitol test.

Cyrex Labs Cross Reactive Foods and Gluten Substitutes, Array # 4. - This specific list checks foods that commonly cross react with gluten and also foods commonly used to replace gluten. It helps customize an anti-inflammatory diet.

Cyrex Labs Predictive Antibody panels determine if or which tissues are currently under antibody attack. This predicts autoimmunity years ahead of time and gives the patient advance notice in order to address trouble spots.

Enterolab (www.Enterolab.com) offers several stool based food antibody panels and gut related stool tests.

What are the lab instructions?
Do I need to consume gluten for testing?

Antibody tests, (blood, saliva, stool), prove a reaction and require recent gluten consumption. Ideally, test first, then go gluten-free. If the patient is off gluten, call the lab for advice on the time window before the test will not work.

Cyrex Labs (www.CyrexLabs.com) tests require prescriptions and a saliva specimen and/or blood draw. If a doctor is needed to write the script, check www.thedr.com for a partial list of practitioners who are familiar with Cyrex Labs. Results are sent to the prescribing doctor.

Enterolab stool specimens are ordered online, kits are sent, home collected and mailed. No prescription is required. Results arrive on email. Enterolab's test works for several months after going gluten-free.

Gene tests do not require gluten consumption or a script and can be run at any time. Genes prove a predisposition to gluten reaction.

Enterolab's gene test is a relatively inexpensive "one part" test. It is a mail in cheek swab and reports the patient's actual genes.

Gluten Free Society (www.Glutenfreesociety.org) gene test is a "two part" (more complete) mail in cheek swab and reports yes or no for both celiac and gluten sensitivity genes.

For information on testing (I have no financial interests) see:

www.TheGlutenSyndrome.net

The latest version of this article and a complete list of references is available at:

www.TheGlutenSyndrome.net /Molecular_Mimicry.pdf

Author's note: While the information Olive has provided can be extremely helpful for adults or children who are not on the spectrum or who are suffering from other ailments, a gluten-free diet is non-negotiable for healing autism. We need to remove foods that cause inflammation and mucous, and which can produce gluteomorphin in the body. I have seen many families waste time and money doing testing only to get confused by false negatives—time and money they could have spent on recovering their child.

Online Gluten Summit – A Chance to Support CDAutism

Recently, in November 2013, Dr. Thomas O'Bryan, an internationally recognized gluten educator, put together an amazing online Gluten Summit. Twenty nine iconic experts on Celiac Disease and its wider perspective, Gluten Sensitivity, also called Gluten Syndrome, covered every possible aspect before 115,000 appreciative viewers. This set of interviews plus extra materials is available for purchase at only $3.30 per speaker and is a priceless addition to all our libraries. It also is a great gift and a good way to inform friends and family. Often our closest loved ones accept new information better from a respected professional rather than from us.

Arrangements are in place to contribute half of your purchase price of The Gluten Summit interviews to the CDAutism project. Go to the link below to order and also credit CDAutism with your contribution.

gg110.infusionsoft.com/go/tgso/Kerri

Appendix 6

Measuring the Strength of Your CD, CDS & CDH

by Charlotte Lackney

CD, CDS, and CDH work because all of them contain chlorine dioxide (ClO_2). ClO_2 is made when a 22.4% solution of sodium chlorite ($NaClO_2$) is activated with an acid, usually 4% hydrochloric acid (HCl) for CD and CDH; and 10% HCl for CDS.

Sometimes it might be useful to measure the amount of ClO_2 in those solutions, although that has not been done in the past when using CD. And, it may only be necessary to know the amount of chlorine dioxide if you are not getting the expected results. There could be a problem with the ingredients or process and knowing the amount of ClO_2 in the solution could be helpful in determining what might be wrong.

The amount of ClO_2 in a solution is measured in *parts per million* (ppm). It is always necessary to specify the volume of solution when talking about ClO_2 ppm because the ppm will vary depending on the dilution of the solution.

If using CDH or CDS with the protocol in this book, it is assumed that you are preparing them at 3000ppm. If using CDS, please keep in mind that the relative strength of one milliliter of CDS is approximately 60% of one drop of CD.

However, if the ingredients are of good quality and the CD or CDH protocol is closely followed, there is usually no need to know or measure the ClO_2 concentration.

Both CD and CDH may continue to activate inside the body from whatever sodium chlorite has not been activated outside the body. That is not the case with CDS, as it has no sodium chlorite to continue activating inside the body. Some believe that this is why many people can tolerate a higher dose of CDS than an equivalent dose of CD.

Measuring the concentration of ClO_2 can easily be done at home. You will need to purchase ClO2 test strips made by *LaMotte* or others. I have only used LaMotte Insta-Test® High-Range Test Strips, 0 to 500ppm (code #3002). The container's label has a chart showing seven colors on it wih each color indicating a ClO_2 concentration in ppm. Note: LaMotte recently changed the appearance of their labels, as shown on the right. Both are the same product. For more information, see their website:

www.lamotte.com/
en/water-wastewater/test-strips/3002.html

Directions for how to use the LaMotte test strips are printed on the side of the container and need to be followed in order to get an accurate ClO_2 concentration reading.

The numbers below each of the seven colors represent 0, 10, 25, 50, 100, 250 and 500ppm. Many people who use the test strips think the 50ppm color is the best one to use for color matching, because they can see the difference in color below and above that color more easily than the others.

If the ClO_2 solution you want to measure is suspected to be 3000ppm, how do you measure that if the test strips highest ClO_2 concentration color is 500ppm? Recall that I had said that it is always necessary to specify the volume of solution when talking about ClO_2 ppm, because the ppm will vary depending on the dilution of the solution.

So if we dilute the suspected 3000ppm ClO_2 stock solution with more water, the same amount of ClO_2 will now be evenly spread out in the larger amount of solution if stirred. There will be less ClO_2 in any one spot than before dilution, because the ClO_2 is now dispersed into a larger volume of water. Visualize what happens when a drop of red food coloring falls into a glass of water. It disperses throughout the water if stirred, and that dark red color is now much lighter because it is diluted in the water.

In order to read 3000ppm with a test strip, and have it match the 50ppm color on the container, we will need to dilute a small sample of the stock solution.

Stock solution (1ml)	÷	Desired dilution in ppm	=	Volume of water for dilution.
3000ppm		50ppm		60ml

Insta-TEST
50 Test Strips
High Range
Chlorine Dioxide
For testing sanitized water
Just dip & read!
LaMotte
InstaTest ANALYTIC
High Range
Chlorine Dioxide
50 Test Strips
Code 3002
LaMotte

The chlorine dioxide test strips (code #3002) made by LaMotte are available on the market with two different labels, although both are the same product. Both labels show the same measuring color chart on the outside of the container.

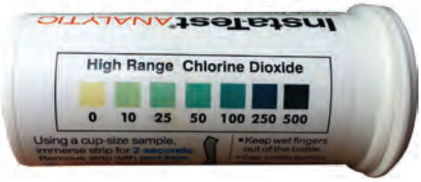
InstaTest ANALYTIC
High Range Chlorine Dioxide
0
10
25
50
100
250
500
Using a cup-size sample, immerse strip for 2 seconds.
Keep wet fingers out of the bottle.

So, if we take a 1ml sample of the stock solution and add it to 60ml of water, then the stock solution is diluted 60 times its original volume. Technically, one should add the 1ml of stock solution to 59ml of water, for a total of 60ml, but using 60ml is easier to measure and you would not see the difference in ClO_2 readings if 59ml of water was used.

To measure, pour 60ml of distilled water into a small glass and then add 1ml of stock solution. Stir well to evenly distribute the ClO_2 sample in the water. Then take one of the test strips—taking care not to touch the pad of chemicals on the end of the strip—and dip it in the solution for two seconds. Keep the strip in one place in the solution and do not move it around during the two-second period.

Without flicking any of the solution off the strip, remove the strip from the test solution with the pad facing up and wait ten seconds. Now compare the color of the test strip pad to the color chart on the side of the test strips container (as shown above).

If the stock solution is 3000ppm, then the color on the test strip will match the 50ppm color on the container. Multiply 50 times 60 (the amount of dilution) and you get 3000ppm.

If the color on the test strip does not match the 50ppm color on the container, try to estimate the ppm number between container colors and multiply that number by 60 to get the ClO_2 concentration.

Tip: You can cut the strips lengthwise to get 100 strips instead of 50; just be sure not to touch the strip's pad with your fingers or your readings may not be accurate. Also it is important to keep the container tightly capped to keep moisture out which can affect the readings.

Healing
the symptoms known as
Kerri Rivera

Dana one year after starting Kerri's protocol.

Appendix 7

Diluting HCl Concentration

The protocols in this book use two different concentrations of hydrochloric acid (HCl): 10% & 4%. Some suppliers sell both concentrations, while others only one. You may also find a source with a different concentration and wish to dilute it to 4 or 10%. The math to determine how much to dilute your stock solution is rather simple. You should have no problem if you know how to multiply, divide, add and subtract. Note: The same calculation works for citric acid if you have 50% and need to go down to 35%.

What you will need is:

- Bottle of concentrated HCl (10% or higher)
- Receiving bottle to store the resulting diluted HCl
- Graduated cylinder or other measuring cup

On the next page is a graphical representation of the dilution equation. It may look intimidating at first glance, but it is actually quite simple and is made up of three simple mini-calculations. Each box is labeled with a letter which refers to the graphic. Just fill in the numbers with a pencil and follow the directions.

We'll go over one example here: Let's say you have 10% HCl and want to make 1 liter of 4% HCl. Here's what you would have to calculate:

Desired Volume		Desired HCl %		Temporary Number
1000ml	×	4%	=	4000

Temporary Number		Stock HCl %		Amount of 10% HCl
4000	÷	10%	=	400ml

Desired Volume		Amount of 10% HCl		Amount of water
1000ml	–	400ml	=	600ml

So, in this example, you would need 400ml of 10% HCl and 600ml of water. to make 1 liter of 4% HCl. If you add them together, you should have 1000ml (1 liter).

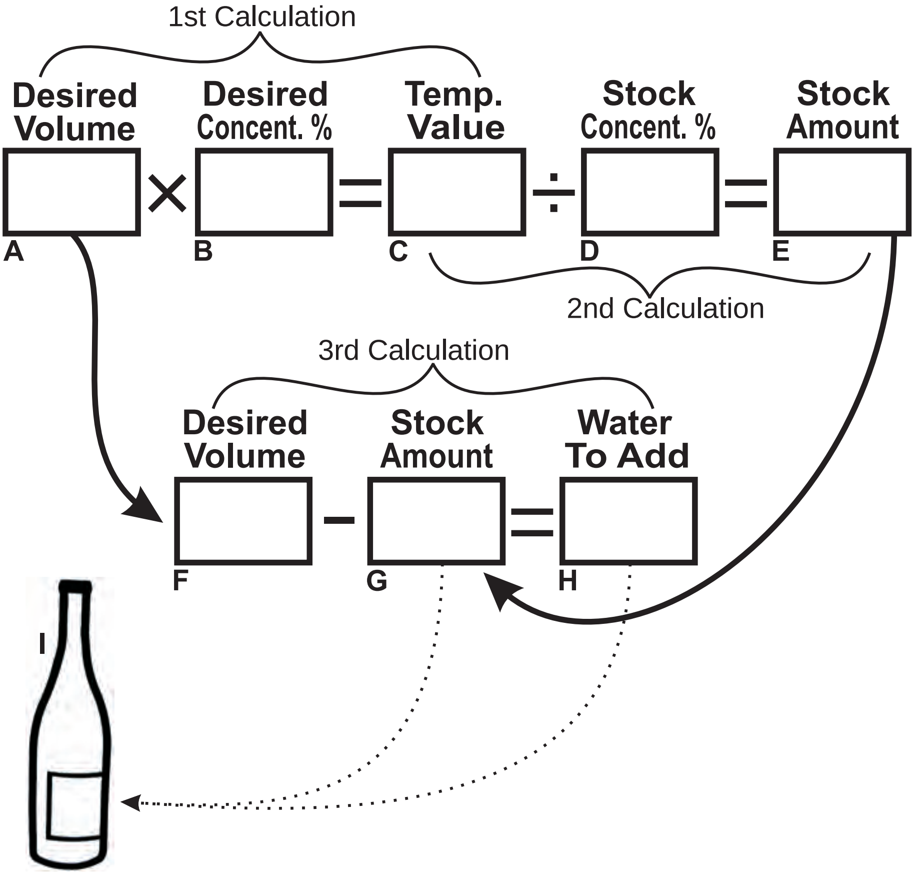
1st Calculation
Desired Volume
Desired Concent. %
Temp. Value
Stock Concent. %
Stock Amount
×
=
÷
=
A
B
C
D
E
2nd Calculation
3rd Calculation
Desired Volume
Stock Amount
Water To Add
–
=
F
G
H
I

A	Determine the volume of the receiving bottle in milliliters. For example, if you have a 1 liter bottle, you would enter the number “1,000” for 1,000ml.
B	This is where you enter the % of the final solution. If you are making 4%, then enter “4” in this box.
C	Multiple **A** and **B** and enter the result into box **C**. This is a temporary value you will need for the next calculation.
D	This is where you enter the percentage of your concentrated stock solution–the HCl you want to dilute. If you have 10%, then enter a “10” in this box.
E	Now divide the temporary value in **C** by **D** and enter the result in **E**. This is the amount of concentrated stock HCl (in milliliters) you will need.
F	Copy what you entered in **A** to **F** for the next calculation.
G	Copy the result from **E** to **G**.
H	Do the subtraction of **F** – **G** and enter he value in **H**. This is the number of milliliters of distilled or filtered water you will need.
I	Measure the resulting volume of water calculated in **H** and pour it into the receiving bottle **I**. Finally, measure out the value in **G** and pour it into the receiving bottle **I**.

Appendix 8

Other Uses of Chlorine Dioxide

Bladder Infection Protocol (for Adults)

The protocol for bladder infection is three oral drops of CD eight times a day with a douche of 10 drops CD three times the first day, and then once every day thereafter, for 30 days (for females).

Ear, Eye, Nose Protocol

Make a mixture that contains one ounce of water for every one drop of CD. Use one drop of that mixture in the affected orifice every hour until symptoms disappear. If the issue is acute you can use one drop every 15 minutes. This method can be used in the ears for infections, in the nose for congestion/infection, and the eye for infection or conjunctivitis.

General Fever Protocol

If the person has a fever, or if you want clarification of any kind, email Kerri, but in the meantime, this will hopefully get you started:

1. Give baths interspersed with enemas, approximately eight hours apart (e.g. bath in the evening and enema in the morning or vice versa).
2. Give oral CD doses (using the person's usual number of drops) 12 to 16 times or more a day (essentially every hour); unless they stop eating, in which case, stop oral dosing. Enemas and baths can continue.
3. Have them drink lots of water.
4. Cool clothes on forehead and neck.

Please see Chapter 14, page 335 for more information on fever therapy.

Teeth

Ten activated drops of CD per one ounce of water. Put this in a spray bottle and spray on the toothbrush.

Throat Spray

Ten activated drops of CD (chlorine dioxide) in one ounce of water. Spray on throat once/hour. Recommended in the case of a sore throat.

Vaginal Douching for Yeast Infections:

Put six activated drops of CD in 60 ml of water. Hold this in for two minutes every day for a few weeks or between periods. Six drops MAX in the vagina.

CD Skin Spray

Use ten drops of activated CD per one ounce of water in a spray bottle.

CD spray is great for a variety of ailments.

- Skin
- Dandruff
- Scrapes
- Cuts
- Rashes
- Toe Fungus

Burns

Any burn should be sprayed with full strength sodium chlorite solution (no acid added) directly from a small spray bottle. Do not add citric acid or HCl at all! If you don't have a spray bottle available drip sodium chlorite solution directly onto the burn, making sure the area is soaked with it.

Wait up to five minutes, but no longer before rinsing off with room temperature water. If you fail to rinse off the sodium chlorite solution, the burn will continue to hurt and get worse.

This technique will cause the burn to heal in 1/4 the time normally required. The pain should stop almost immediately or reduce to almost zero within several minutes.

Sunburns should be treated the same way. Spray the red area, wait 1 to 5 minutes, and rinse off. If the area is still sore, in about an hour spray the area again and wait 5 minutes before rinsing off. Remember, DO NOT allow the sodium chlorite to remain in place. It must be rinsed off! The pain should be gone in a couple of minutes. Generally two doses will overcome most sunburns, but on rare occasions if the discomfort is not all gone you can use a third dose. Be sure to rinse it off.

Note that sodium chlorite is alkaline and burns are acidic. Therefore the sodium chlorite neutralizes the acidity that resides in the burned areas. This is part of the reason why burns heal rapidly after sodium chlorite applications.

Appendix 9

Blank Kalcker Parasite Protocol Chart

The next page contains a blank Kalcker Parasite Protocol Daily Chart identical to those found on pages 199 through 217. You are encouraged to copy this chart for use in your parasite protocol planning. Setting the copy machine to enlarge at 125% should fill a standard letter-sized page.

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin®)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM												
6:30 AM												
7:00 AM												
7:30 AM												
8:00 AM												
8:30 AM												
9:00 AM												
9:30 AM												
10:00 AM												
10:30 AM												
11:00 AM												
11:30 AM												
12:00 PM												
12:30 PM												
1:00 PM												
1:30 PM												
2:00 PM												
2:30 PM												
3:00 PM												
3:30 PM												
4:00 PM												
4:30 PM												
5:00 PM												
5:30 PM												
6:00 PM												
6:30 PM												
7:00 PM												
7:30 PM												
8:00 PM												
8:30 PM												
9:00 PM												
9:30 PM												

Appendix 10

Lunar Phase Calendar for Kalcker Parasite Protocol

As covered in Chapter 8, the Kalcker Parasite Protocol is administered over a period of 19 days and scheduled according to the lunar cycle. We have included an easy to follow calendar starting on the next page for the years 2014 through 2016.

Each chart page shows 6 months with each month broken up into 3 columns: Column 1 is the day of the month; column 2 indicates the lunar phase; and column 3 shows the corresponding Parasite Protocol day.

The full moon is identified with an "F" in the second column and the new moon is identified by an "N" in each month section, which usually occurs towards the end of each protocol cycle.

We have used New York as our center point since it is roughly the center of the majority of those using this book in the US and Europe. The moon phase may occur 1 day before or after the dates shown if you are in another part of the world, but the difference is not great, so feel free to use the dates shown. However, you are welcome to be precise for your particular location. The Farmer's Almanac is a great source and provides an online calendar that can be adjusted to your location:

www.almanac.com/moon

You may also wish to install a moon phase app into your smart phone. Many free options exist.

2014 - Lunar Phases for New York, NY

Jan			Feb			Mar			Apr			May			Jun		
1	N	18	1			1	N	18	1			1	N		1		
2			2			2			2			2			2		
3			3			3			3			3			3		
4			4			4			4			4			4		
5			5			5			5			5			5		
6			6			6			6			6			6		
7			7			7			7			7			7		
8			8			8			8			8			8		
9			9			9			9			9			9		
10			10			10			10			10			10		0
11			11		0	11			11			11			11		1
12		0	12		1	12			12		0	12			12		2
13		1	13		2	13		0	13		1	13		0	13	F	3
14		2	14	F	3	14		1	14		2	14		1	14		4
15	F	3	15		4	15		2	15	F	3	15		2	15		5
16		4	16		5	16	F	3	16		4	16	F	3	16		6
17		5	17		6	17		4	17		5	17		4	17		7
18		6	18		7	18		5	18		6	18		5	18		8
19		7	19		8	19		6	19		7	19		6	19		9
20		8	20		9	20		7	20		8	20		7	20		10
21		9	21		10	21		8	21		9	21		8	21		11
22		10	22		11	22		9	22		10	22		9	22		12
23		11	23		12	23		10	23		11	23		10	23		13
24		12	24		13	24		11	24		12	24		11	24		14
25		13	25		14	25		12	25		13	25		12	25		15
26		14	26		15	26		13	26		14	26		13	26		16
27		15	27		16	27		14	27		15	27		14	27	N	17
28		16	28		17	28		15	28		16	28		15	28		18
29		17				29		16	29	N	17	29		16	29		
30	N	18				30	N	17	30		18	30	N	17	30		
31						31		18				31		18			

2014 - Lunar Phases for New York, NY

Jul			Aug			Sep			Oct			Nov			Dec		
1			1			1			1			1			1		
2			2			2			2			2			2		
3			3			3			3			3		0	3		0
4			4			4			4			4		1	4		1
5			5			5		0	5		0	5		2	5		2
6			6			6		1	6		1	6	F	3	6	F	3
7			7		0	7		2	7		2	7		4	7		4
8			8		1	8	F	3	8	F	3	8		5	8		5
9		0	9		2	9		4	9		4	9		6	9		6
10		1	10	F	3	10		5	10		5	10		7	10		7
11		2	11		4	11		6	11		6	11		8	11		8
12	F	3	12		5	12		7	12		7	12		9	12		9
13		4	13		6	13		8	13		8	13		10	13		10
14		5	14		7	14		9	14		9	14		11	14		11
15		6	15		8	15		10	15		10	15		12	15		12
16		7	16		9	16		11	16		11	16		13	16		13
17		8	17		10	17		12	17		12	17		14	17		14
18		9	18		11	18		13	18		13	18		15	18		15
19		10	19		12	19		14	19		14	19		16	19		16
20		11	20		13	20		15	20		15	20		17	20		17
21		12	21		14	21		16	21		16	21		18	21	N	18
22		13	22		15	22		17	22		17	22	N		22		
23		14	23		16	23		18	23	N	18	23			23		
24		15	24		17	24	N		24			24			24		
25		16	25	N	18	25			25			25			25		
26	N	17	26			26			26			26			26		
27		18	27			27			27			27			27		
28			28			28			28			28			28		
29			29			29			29			29			29		
30			30			30			30			30			30		
31			31						31						31		

2015 - Lunar Phases for New York, NY

Jan			Feb			Mar			Apr			May			Jun		
1		0	1		1	1			1		0	1		1	1		2
2		1	2		2	2		0	2		1	2		2	2	F	3
3		2	3	F	3	3		1	3		2	3	F	3	3		4
4	F	3	4		4	4		2	4	F	3	4		4	4		5
5		4	5		5	5	F	3	5		4	5		5	5		6
6		5	6		6	6		4	6		5	6		6	6		7
7		6	7		7	7		5	7		6	7		7	7		8
8		7	8		8	8		6	8		7	8		8	8		9
9		8	9		9	9		7	9		8	9		9	9		10
10		9	10		10	10		8	10		9	10		10	10		11
11		10	11		11	11		9	11		10	11		11	11		12
12		11	12		12	12		10	12		11	12		12	12		13
13		12	13		13	13		11	13		12	13		13	13		14
14		13	14		14	14		12	14		13	14		14	14		15
15		14	15		15	15		13	15		14	15		15	15		16
16		15	16		16	16		14	16		15	16		16	16	N	17
17		16	17		17	17		15	17		16	17		17	17		18
18		17	18	N	18	18		16	18	N	17	18	N	18	18		
19		18	19			19		17	19		18	19			19		
20	N		20			20	N	18	20			20			20		
21			21			21			21			21			21		
22			22			22			22			22			22		
23			23			23			23			23			23		
24			24			24			24			24			24		
25			25			25			25			25			25		
26			26			26			26			26			26		
27			27			27			27			27			27		
28			28			28			28			28			28		0
29						29			29			29			29		1
30						30			30		0	30		0	30		2
31		0				31						31		1			

2015 - Lunar Phases for New York, NY

Jul			Aug			Sep			Oct			Nov			Dec		
1	F	3	1		4	1		6	1		7	1		8	1		9
2		4	2		5	2		7	2		8	2		9	2		10
3		5	3		6	3		8	3		9	3		10	3		11
4		6	4		7	4		9	4		10	4		11	4		12
5		7	5		8	5		10	5		11	5		12	5		13
6		8	6		9	6		11	6		12	6		13	6		14
7		9	7		10	7		12	7		13	7		14	7		15
8		10	8		11	8		13	8		14	8		15	8		16
9		11	9		12	9		14	9		15	9		16	9		17
10		12	10		13	10		15	10		16	10		17	10		18
11		13	11		14	11		16	11		17	11	N	18	11	N	
12		14	12		15	12		17	12	N	18	12			12		
13		15	13		16	13	N	18	13			13			13		
14		16	14	N	17	14			14			14			14		
15	N	17	15		18	15			15			15			15		
16		18	16			16			16			16			16		
17			17			17			17			17			17		
18			18			18			18			18			18		
19			19			19			19			19			19		
20			20			20			20			20			20		
21			21			21			21			21			21		
22			22			22			22			22		0	22		0
23			23			23			23			23		1	23		1
24			24			24		0	24		0	24		2	24		2
25			25			25		1	25		1	25	F	3	25	F	3
26			26		0	26		2	26		2	26		4	26		4
27			27		1	27	F	3	27	F	3	27		5	27		5
28		0	28		2	28		4	28		4	28		6	28		6
29		1	29	F	3	29		5	29		5	29		7	29		7
30		2	30		4	30		6	30		6	30		8	30		8
31	F	3	31		5				31		7				31		9

2016 - Lunar Phases for New York, NY

Jan			Feb			Mar			Apr			May			Jun		
1		10	1		12	1		11	1		12	1		12	1		14
2		11	2		13	2		12	2		13	2		13	2		15
3		12	3		14	3		13	3		14	3		14	3		16
4		13	4		15	4		14	4		15	4		15	4	N	17
5		14	5		16	5		15	5		16	5		16	5		18
6		15	6		17	6		16	6		17	6	N	17	6		
7		16	7		18	7		17	7	N	18	7		18	7		
8		17	8	N		8	N	18	8			8			8		
9	N	18	9			9			9			9			9		
10			10			10			10			10			10		
11			11			11			11			11			11		
12			12			12			12			12			12		
13			13			13			13			13			13		
14			14			14			14			14			14		
15			15			15			15			15			15		
16			16			16			16			16			16		
17			17			17			17			17			17		0
18			18			18			18			18		0	18		1
19			19		0	19			19		0	19		1	19		2
20		0	20		1	20		0	20		1	20		2	20	F	3
21		1	21		2	21		1	21		2	21	F	3	21		4
22		2	22	F	3	22		2	22	F	3	22		4	22		5
23	F	3	23		4	23	F	3	23		4	23		5	23		6
24		4	24		5	24		4	24		5	24		6	24		7
25		5	25		6	25		5	25		6	25		7	25		8
26		6	26		7	26		6	26		7	26		8	26		9
27		7	27		8	27		7	27		8	27		9	27		10
28		8	28		9	28		8	28		9	28		10	28		11
29		9	29		10	29		9	29		10	29		11	29		12
30		10				30		10	30		11	30		12	30		13
31		11				31		11				31		13			

2016 - Lunar Phases for New York, NY

Jul			Aug			Sep			Oct			Nov			Dec		
1		14	1		16	1	N	17	1		18	1		18	1		
2		15	2	N	17	2		18	2			2			2		
3		16	3		18	3			3			3			3		
4	N	17	4			4			4			4			4		
5		18	5			5			5			5			5		
6			6			6			6			6			6		
7			7			7			7			7			7		
8			8			8			8			8			8		
9			9			9			9			9			9		
10			10			10			10			10			10		0
11			11			11			11			11		0	11		1
12			12			12			12			12		1	12		2
13			13			13		0	13		0	13		2	13	F	3
14			14			14		1	14		1	14	F	3	14		4
15			15		0	15		2	15		2	15		4	15		5
16		0	16		1	16	F	3	16	F	3	16		5	16		6
17		1	17		2	17		4	17		4	17		6	17		7
18		2	18	F	3	18		5	18		5	18		7	18		8
19	F	3	19		4	19		6	19		6	19		8	19		9
20		4	20		5	20		7	20		7	20		9	20		10
21		5	21		6	21		8	21		8	21		10	21		11
22		6	22		7	22		9	22		9	22		11	22		12
23		7	23		8	23		10	23		10	23		12	23		13
24		8	24		9	24		11	24		11	24		13	24		14
25		9	25		10	25		12	25		12	25		14	25		15
26		10	26		11	26		13	26		13	26		15	26		16
27		11	27		12	27		14	27		14	27		16	27		17
28		12	28		13	28		15	28		15	28		17	28		18
29		13	29		14	29		16	29		16	29	N	18	29	N	
30		14	30		15	30	N	17	30	N	17	30			30		
31		15	31		16				31						31		

Three things cannot be long hidden:
the sun, the moon, and the truth.

~ Buddha

Appendix 11

Health Benefits of Diatomaceous Earth

by Dee McCaffrey, CDC

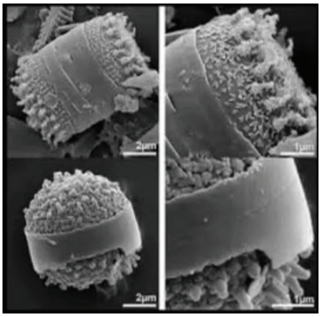

You've probably never heard of one of the best and most economical supplements to improve your health and cleanse your body. In fact, this supplement is perhaps one of the best kept secrets ever. This relatively unknown "food" supplement is called Diatomaceous Earth, also known as DE. It is a completely natural substance that is rich in naturally occurring silica, a mineral whose list of documented health benefits continues to grow as more research is being conducted.

Some of the most recent studies show that it can strengthen bones and joints, prevent osteoporosis and restore bone health if you already have osteoporosis, boost the immune system, ward off Alzheimer's, prevent premature aging and wrinkling of the skin, and strengthens the arterial walls to maintain good heart health. Another of the benefits of silica is that it helps to destroy bad fats in the body. Used as a daily treatment, diatomaceous earth can alleviate the potentially deadly risks of high cholesterol, high blood pressure, and obesity.

Additionally, because of its physical structure, diatomaceous earth is a highly effective anti-inflammatory and internal cleansing agent for the body. It can remove intestinal bacteria, parasites, e-coli, viruses, pesticides, heavy metals, and other toxins. It has also been known to assist with vertigo, headaches, tinnitus, and insomnia.

Sound too good to be true? Read on to learn more about this overlooked multi-purpose supplement and how it can improve many aspects of your health at a fraction of the cost of other supplements.

What is Diatomaceous Earth?

Diatomaceous earth is the fossilized shells of microscopic water-dwelling organisms known as diatoms. Diatoms are one-celled plants (algae or phytoplankton) that are the primary food source for marine life. These plants have been part of the earth's ecology since prehistoric times, yet their species still exists today in both salty and fresh waters all over the world.

Diatoms use soluble silica (sand) from their environment to make their shells, or exoskeletons. As diatoms die, their shells fall to the bottom of the bodies of water in which they live. It is believed that 30 million years ago the diatoms piled up to form thick beds of chalky fossilized sediment. These beds, known as diatomite or diatomaceous earth, were discovered after the waters receded.

Today, hundreds of large deposits of diatomaceous earth occur all over the world. Some are still underwater and some are found in ancient dried lake bottoms. It has been estimated that one cubic inch of diatomaceous earth may contain as many as 400 million shells, that's how small diatoms are!

Diatomite is mined and ground into a powder that looks and feels like talcum powder. This all natural powder is called Diatomaceous Earth (DE) or fossil shell flour. It is made up of approximately 33% silicon, 19% calcium, 5% sodium, 3% magnesium 2% iron and many other trace minerals such as titanium, boron, manganese, copper and zirconium.

What Do Fossilized Shells Have to Do With Improving Your Health?

It's all about the silica.

As I mentioned earlier, the shells of diatoms are made up mostly of silica (known by chemists as silicon dioxide). Silica is one of the most abundant minerals on the planet, but most of it is in a form that is unabsorbable by humans, and there are limited numbers of foods that contain an adequate amount of the absorbable form to supply the quantity our body needs. Silica is the most important trace mineral for human health. It plays an important role in many body functions and has a direct relationship to mineral absorption. For optimal health, the average human body needs to hold approximately seven grams of silica, a quantity far exceeding the figures for other important minerals such as iron and calcium.

Silica can be found in such foods as alfalfa, beets, brown rice and oats. Bell peppers and leafy green vegetables also provide silica, while asparagus, Jerusalem artichokes, parsley, sunflower seeds and grain husks, such as barley, millet and wheat contain smaller amounts.

Years ago, the silica found in our foods was adequate, but with today's depleted soils, only about one-third of the silica needed by our body is supplied in our food.

In our youth, our tissues absorb and maintain high levels of silica— enabling our bodies to remain flexible, resilient, and energetic—but as we age, and as our dietary sources of silica are not meeting our needs, our silica levels steadily decline until they become almost non-existent. In fact, 80% of all of our body's silica is used up by the time we become adults.

Our bodies need silica regardless of our age and even when diet is not the primary factor in cases of deficiencies, we often become deficient in this essential trace element simply thorough the aging process. The effect of this steady decrease in silica levels is a progressive decline in health, increasing fatigue and acceleration of the aging process.

Diatomaceous earth, which is very high in the absorbable form of silica, can replenish silica levels in the body, which can vastly improve your health and reverse many chronic problems. Studies show that high levels of silica can:

Improve Bone Health and Stop Osteoporosis: Calcium and vitamin D alone are not sufficient for bone growth, density, strength, and flexibility. In fact, the body cannot absorb and use calcium without the presence of silica. Recent data suggests that instead of promoting healing, calcium supplements actually speed up the leeching away of bone calcium and accelerate the degenerative process of osteoporosis and similar diseases that affect the connective tissues in the human body!

To re-mineralize and repair damaged bones, it is now advised that a sufficient silica supplement be taken daily because bones are composed of mainly of the minerals phosphorus, magnesium and calcium; however, these minerals need the presence of silica to be deposited into the bones, especially calcium. Silica hastens the healing of fractures and also diminishes scarring at the location of a fracture. A great deal of research evidence indicates that silica has the ability to "morph" itself into calcium through a transmutation process. Yes, silica actually has the ability to be turned into calcium when there is a calcium deficiency and the body needs it!

That fact alone should be reason enough to supplement with DE!!

Ward Off Alzheimer's: Scientists and researchers have long hypothesized that Alzheimer's disease is linked to a build up of aluminum in the brain, and links between aluminum in drinking water supplies and Alzheimer's have now been ascertained. A factor that had been overlooked is that silica reduces the accumulation of aluminum. When researchers added silica to aluminum-laced water supplies, it inhibited the aluminum from being absorbed. It also caused a proliferation in the excretion of aluminum in urine and lowered aluminum concentrations in the brain, liver, bone, spleen and kidneys. Silica, therefore, may be important in supporting neurological health.

Improve Heart and Lung Health: Silica can hinder the effects of coronary disease by fortifying blood vessels. Studies confirm that with age, silica disappears from the aorta, the heart's key blood vessel—thus weakening its critical connective tissue, and resulting in a greater cardiac risk. Studies have shown diatomaceous earth to significantly lower cholesterol by removing plaque and keeping arteries and veins supple. It also helps to regulate blood pressure. Silica also aids in the repair and maintenance of vital lung tissues and defending them from pollution.

Stop Premature Wrinkling and Sagging Skin: Silica is excellent for supporting bones and connective tissue. Your body needs healthy connective tissue for internal organs and the largest organ which is your skin. And you probably know that degrading connective tissue, with age is why wrinkles appear and skin begins to sag. Perhaps it is not age, but rather not enough silica that makes this happen.

Collagen, which is mostly made up of silica, is the glue that holds us together. Collagen accounts for up to 75% of the weight of the dermis and is responsible for the resilience and elasticity of the skin. Our connective tissues consist of collagen, elastin, mucopolysaccharides and mucous carbohydrates which aid in moisture retention. Their capacity to hold on to moisture keeps the connective tissue resilient and has apparent importance in the prevention of premature aging. All these valuable molecules house large quantities of silica. Also, many people with advanced arthritis suffer from bone deformation when tendons and ligaments in the joints lose flexibility due to loss of collagen. Tendons and ligaments need silica for health and flexibility.

Boost the Immune System: Silica may play an important role in the immune system and its biological response to harmful stimuli. Silica is necessary for the body to produce antibodies that fight off viruses, bacteria, allergens and other invaders that the body views as foreign.

Other Ways Diatomaceous Earth Can Improve Your Health

One of the most sought after benefits of diatomaceous earth is to cleanse the digestive tract. Many people consume a variety of processed foods, which contain various harmful chemicals and toxins. Plus, as air quality has decreased, we are constantly putting more toxins in our body that need to be removed.

Diatomaceous earth can purge any parasite, virus, bacteria, or toxin that is clinging to our digestive tract. While diatomaceous earth feels like a soft powder, the truth is that diatomaceous earth is actually a small cylinder with extremely sharp edges. As we consume diatomaceous earth, these sharp edges scrape away parasites, toxins, and viruses clinging to the lining of our digestive tract. Plus, any toxins or bacteria floating in our digestive tract are absorbed and trapped by diatomaceous earth. It has also been shown to cling to bad fats in the body, effectively reducing bad cholesterol. Diatomaceous earth is then expelled through our bowel movements and these harmful materials are removed.

Our skin and organs are strong and thick enough not to be affected by these sharp edges, which is why we feel no pain taking diatomaceous earth. After a few days, the body's digestive system can be thoroughly cleansed and can operate much more efficiently.

Are All Forms of Diatomaceous Earth the Same?

No! There are two types of DE—Food Grade and Pool Grade.

Most deposits of diatomaceous earth are from salt water sources, while only a few are from fresh water sources. Of the hundreds of DE deposits that exist worldwide, only 4 of them can be called food grade. The DE sourced from fresh water beds are of extremely high purity—so pure that it is called Food Grade Diatomaceous Earth, or also known as fossil shell flour. Food grade diatomaceous earth, the only form safe for human consumption, is a white powder, while other forms of diatomaceous earth may have a brown/reddish tint.

Also, there are two types of naturally occurring silica—crystalline silica and non-crystalline (amorphous) silica. Crystalline silica can be very dangerous, especially to our lungs if inhaled, while non-crystalline silica is completely safe, even for human and animal consumption. The fresh water diatoms contain mainly non-crystalline silica, while salt water diatoms can contain both types, with higher amounts of the crystalline form.

PERMA-GUARD™ is the trade name known worldwide for using a grade and quality of Diatomaceous Earth (DE) that is extremely pure. Brands that use Perma-Guard™ food grade Diatomaceous Earth contains less than 0 .5% crystalline silica. It is important to have a consistent shape diatom and no unwanted sediment. The shape of the diatom must be tubular with holes on the walls. It must be from fresh water because the fresh water diatoms form a harder shell and are less fragile that those from salt water. Salt water deposits contain a mix of diatom species. These deposits shapes and sediments are inconsistent making them unusable for safe human consumption.

Other types of diatomaceous earth have other industrial uses. It is heated to a very high temperature (about 1000°C or 1800°F). This type of DE is called "calcined." It is used for pool filters and other types of filtering but it is also used as a filler and can end up in paints, cosmetics, drugs, chemical insecticides, and other things. Food grade DE is never heated.

Bottom Line: ONLY use Food Grade Diatomaceous Earth for health purposes!!! Food grade diatomaceous earth is the purest form of diatomaceous earth and can be consumed and used by humans. Food grade diatomaceous earth is heavily regulated and must contain less than a certain amount of specific minerals.

Pool grade DE is calcinated and is a stronger, more potent form of DE. It is not to be used by humans and pool grade DE must be handled with gloves and a mask to ensure no throat irritations occur.

All DE should have a label that tells you if it has been calcined and how much crystalline silica it contains.

How Much to Take—How Safe Is It?

Food grade diatomaceous earth is a very fine powder and is very light due to its high porosity. It mixes easily into liquids and foods. Most people take a teaspoon or a tablespoon two or three times a day (up to a total of one quarter cup per day) for best results. Mix it into water, juice, smoothies or other foods.

It sort of resembles putting a spoonful of baking soda in water and drinking it!

Silica is water soluble; hence, once you get it into your body it easily absorbed via the intestinal wall and is also rapidly excreted. It does not accumulate in the body, so consistent daily supplementation is important. Studies have not found any negative side effects from too much silica. Its safety and extensive range of uses makes silica one of the most important minerals used in complementary therapy and alternative medicine. Regular supplementation could make a significant difference in your health.

Being approved by the FDA, Diatomaceous Earth has absolutely no dangerous side effects. It can be used by anyone, as long as the person doesn't suffer from a serious illness. As with any other health supplement, talk to a doctor before using DE. You might have some intolerance to it, so it's advisable to not risk your health. Pregnant and breastfeeding mothers can use it as long as their doctor has agreed.

What Other Uses Does DE Have?

One of the most commonly known uses for diatomaceous earth is as a non-toxic bug killer. DE is almost pure silica (with some beneficial trace minerals); under a microscope, it looks like shards of glass (glass is made from silica). On any beetle-type insect that has a carapace, like ants, fleas and cockroaches, the DE works under the shell and punctures the body, which then dehydrates their innards and the insect dies. DE is totally nontoxic. There is no buildup of tolerance like there is to poisons because the method of killing is PHYSICAL, not chemical. But rest assured, food grade DE doesn't hurt people or animals. In fact, it makes a great flea killer for your pets, and can also be added to their food as de-wormer. Many farmers use it to keep farm animals healthy, both inside and out. Food grade DE is also added to grains in storage because it keeps bugs from eating the grain. So if you eat any grains, you're probably already eating a little bit of DE!

Where to Get It!

You can buy food grade diatomaceous earth online. Processed-Free America sells it through its online store.

Written by Dee McCaffrey, CDC, who lost 100 pounds and has kept it off for over 20 years. For more info, go to: www.processedfreeamerica.org

Appendix 12

Summary of Protocols

The purpose of this list is to have all the protocols in one place, so you don't have to be hunting and flipping pages to find something. This information is also available on our website www.cdautism.org if you would like to print anything full size.

Diet

Permitted Foods List:

Note: Organic products are better but not required.

Proteins

- ☐ Beef
- ☐ Chicken
- ☐ Eggs
- ☐ Fish (small not large size)
- ☐ Pork
- ☐ Turkey

Fruit

- ☐ Most fresh fruits are permitted (except citrus, mango, pineapple, kiwi, and limit berries).
- ☐ Frozen fruit without added cream or sugar.
- ! NO canned fruit (nothing canned ever).
- ! Be careful of dried fruit as it may contain sugar.

Vegetables

- ☐ All vegetables are fine!!!
- ☐ Including French fries, however, not frozen fries or fries from fast food chains; these are often coated in flour.

Nuts

- ☐ Almonds
- ☐ Cashews
- ☐ Coconut
- ☐ Hazelnuts
- ☐ Walnuts

Grains

☐ Amaranth
☐ Buckwheat
☐ Corn
☐ Millet
☐ Quinoa
☐ Rice
☐ Sorghum
☐ Tapioca
☐ Xanthan gum

Beans

☐ All beans—EXCEPT soy
☐ Split Pea
☐ Garbanzo
☐ Lentils
☐ Navy
☐ Peanuts

Sweeteners

☐ Stevia (Best choice of all sweeteners)
☐ Agave syrup
☐ Honey
☐ Maple syrup (without added sugar)
☐ Xylitol

Prohibited Foods List

- Acetic acid (E260)
- Artificial flavoring
- Artificial sweeteners
- Bouillon cubes
- Bread
- Candy
- Cane sugar
- Carrageenan
- Catsup
- Chocolate milk
- Coloring
- Corn flakes
- Corn syrup
- Cow’s milk in any form (even lactose-free milk products)
- Flour tortillas
- Gelatin
- Malt
- Margarine
- Mayonnaise
- Microwave popcorn
- MSG
- Natural flavoring
- Noodle soup
- Oatmeal (except for Bob’s Red Mill GF oats)
- Pasta
- Piloncillo (unrefined sugar)
- Children´s nutritional shakes
- Play-Doh™
- Preservatives
- Processed meats (hotdogs, ham, sausage, bologna, cold cuts)
- Shellfish (full of toxins)
- Sodas
- Soy/fruit beverages
- Soy milk
- Soy sauce
- Sports drinks
- Sugar
- Yeast

Preparing Your First One drop CD Batch for the Day

1.	Start by filling the baby bottle with 8 fl. oz. (237ml) of water (distilled or reverse osmosis). NO alkaline water!	
2.	Place 1 drop of sodium chlorite solution into the CLEAN and DRY shot glass.	
3.	Add 1 drop of acidic activator (hydrochloric acid or citric acid) to the shot glass containing the drop of sodium chlorite. The number of drops may be higher if you are using a weaker activator. See chart on page 94.	
4.	Now wait the appropriate time for the mixture to react (see chart on page 94). You should see the color change from clear to slightly yellow. If there were more drops in the shot glass, the color change is more noticeable. You are also likely to notice the chlorine-like smell coming from the shot glass. Remember, this is NOT chlorine, but rather chlorine dioxide.	
5.	After the activation time has passed, pour a little water from the baby bottle into the shot glass and let it mix. This mostly stops the chemical reaction and insures you get most of the mixture out in the next step.	
6.	Lastly, pour all of the watered down mixture in the shot glass back into the baby bottle, and seal it tightly with the cap—don't leave it sitting open for any length of time. Think partially used soda pop and how you would want to keep that closed.	

Estimated Full Oral CD Doses by Weight

Use these numbers as a guide only. You may need to go up by as much as 50% or more over the indicated drops. Read chart as: **POUNDS / KILOGRAMS → DROPS OF CD (per 8 fl. oz. water)**

25/11→8	62/28→17	99/45→24	136/62→29	173/78→33	210/95→37
26/12→8	63/29→17	100/45→24	137/62→29	174/79→33	211/96→37
27/12→9	64/29→17	101/46→24	138/63→29	175/79→34	212/96→37
28/13→9	65/29→18	102/46→24	139/63→30	176/80→34	213/97→37
29/13→9	66/30→18	103/47→24	140/64→30	177/80→34	214/97→37
30/14→9	67/30→18	104/47→25	141/64→30	178/81→34	215/98→37
31/14→10	68/31→18	105/48→25	142/64→30	179/81→34	216/98→37
32/15→10	69/31→18	106/48→25	143/65→30	180/82→34	217/98→37
33/15→10	70/32→19	107/49→25	144/65→30	181/82→34	218/99→38
34/15→10	71/32→19	108/49→25	145/66→30	182/83→34	219/99→38
35/16→11	72/33→19	109/49→25	146/66→30	183/83→34	220/100→38
36/16→11	73/33→19	110/50→26	147/67→31	184/83→35	221/100→38
37/17→11	74/34→19	111/50→26	148/67→31	185/84→35	222/101→38
38/17→11	75/34→20	112/51→26	149/68→31	186/84→35	223/101→38
39/18→12	76/34→20	113/51→26	150/68→31	187/85→35	224/102→38
40/18→12	77/35→20	114/52→26	151/68→31	188/85→35	225/102→38
41/19→12	78/35→20	115/52→26	152/69→31	189/86→35	226/103→38
42/19→12	79/36→20	116/53→26	153/69→31	190/86→35	227/103→38
43/20→13	80/36→21	117/53→27	154/70→31	191/87→35	228/103→38
44/20→13	81/37→21	118/54→27	155/70→31	192/87→35	229/104→38
45/20→13	82/37→21	119/54→27	156/71→32	193/88→35	230/104→39
46/21→13	83/38→21	120/54→27	157/71→32	194/88→35	231/105→39
47/21→14	84/38→21	121/55→27	158/72→32	195/88→36	232/105→39
48/22→14	85/39→21	122/55→27	159/72→32	196/89→36	233/106→39
49/22→14	86/39→22	123/56→27	160/73→32	197/89→36	234/106→39
50/23→14	87/39→22	124/56→28	161/73→32	198/90→36	235/107→39
51/23→15	88/40→22	125/57→28	162/73→32	199/90→36	236/107→39
52/24→15	89/40→22	126/57→28	163/74→32	200/91→36	237/108→39
53/24→15	90/41→22	127/58→28	164/74→32	201/91→36	238/108→39
54/24→15	91/41→23	128/58→28	165/75→33	202/92→36	239/108→39
55/25→16	92/42→23	129/59→28	166/75→33	203/92→36	240/109→39
56/25→16	93/42→23	130/59→28	167/76→33	204/93→36	241/109→39
57/26→16	94/43→23	131/59→28	168/76→33	205/93→36	242/110→39
58/26→16	95/43→23	132/60→29	169/77→33	206/93→37	243/110→39
59/27→16	96/44→23	133/60→29	170/77→33	207/94→37	244/111→40
60/27→17	97/44→24	134/61→29	171/78→33	208/94→37	245/111→40
61/28→17	98/44→24	135/61→29	172/78→33	209/95→37	246/112→40

Advanced Protocols (After Full Oral Dose is Reached)

CD Enemas:

I recommend enemas no less frequently than every other day. You can do them more often particularly if you see parasites coming out. The ratio is 1-2 drops chlorine dioxide per 100 ml of filtered warm water. Work your way up to 500 mL – 1.5 L of water per enema, depending on the size of your child. You have the following options for methods of administering the enemas: a catheter and syringes, a multipurpose enema/douche bag, empty out a fleet enema, or a gravity bag. (note, if your child suffers from constipation you can add enemas and baths in before you get to full dose, alternating them on opposite days or opposite ends of the day).

Enemas should be started when the child reaches full oral dose. However, if your child suffers from constipation you can add enemas and baths on day one (see page 113), before you get to the full oral dose. However, alternate them on opposite days or opposite ends of the day. The following chart is an approximate amount of water based on the size of the person, and calculated the maximum of 2 drops of CD per 100mL of water.

Age/Size	Water Volume	Drops of CD
Child	1/2 Liter (500ml)	10
Adolescent	1 Liter (1,000ml)	20
Teen/Adult	up to 2 Liters (2,000ml)	40

CD Baths:

10-100 Activated drops of CD in a hot (to tolerance) bath with enough water to cover the child's body. Soak for 20 minutes. Apply on alternate enema days, unless you are doing daily enemas, in which case apply on opposite ends of the day.

The CD baths can go from as low as 10 drops to as high as 80 to 100 drops; it just depends on the person, and the size of your tub. With younger/smaller children we start with ten drops. Since we're doing baths every other day, start with ten drops on Monday, 11 drops on Wednesday, and 12 drops on Friday. Just keep going up until you get to 20 drops. The bigger and older the person is, the more drops they can tolerate. Fill the tub to a level that maximizes skin contact.

72/2 Protocol

Give one dose of CD every 2 hours for 72 hours straight, including the middle of the night. Apply every possible weekend.

This involves giving one dose of CD every two hours for 72 hours straight—including the middle of the night. Here are some additional thoughts and guidelines:

- Start protocol when you pick up your child from school on Friday.
- Give them the last 72/2 dose when you drop them off at school on Monday.
- Avoid giving CD enemas or CD baths during this protocol, UNLESS they are dumping parasites, in which case you might need to reduce the amount of CD on both the oral and enema dose to insure the child doesn't have a Herxheimer reaction.
- Why not give hourly doses during the day at the usual dose? It's simply too much if you are doing it all night as well.
- Watch for improvements on Tuesday or Wednesday each week.
- Ideally, get your spouse or significant other to help with every other nightly dose.

Supplement Dosing Overview

Supplement	Dose	Time of Day	Empty Stomach (ES) or With Food (WF)
Probiotics	1 cap	Before Bed	ES
Omega-3/ Omega-6	1 Tbsp	With any meal 1-3x / day	WF
L-Carnitine	250-1000mg/day	With any meal 1-3x / day	WF
GABA	Up to 2,500mg 2x/day	Upon waking & at bedtime	ES
5-HTP	50-200mg	Morning & Evening	WF
L-Theanine	Work up to 200-250mg/day	Mornings or Mornings & Nights	ES
Pycnogenol	25-40mg a day/as needed	Morning	WF
L-Carnosine	200-400mg 2x / day	Morning & Evening	WF
Taurine	500-1500mg/day	Morning, Noon & Night	ES
DMG	900mg/day	Morning	ES
TMG	500mg/day	Morning	ES
Enzymes	1 cap w/meals	Morning, Noon & Night	WF

Parasite Protocol

Please see the complete explanation of the parasite protocol starting on page 187. Day-to-day protocol charts can be found on page 198.

Chelation:

I suggest that when you decide to add clay baths into your child's protocol that you do so on the off days of the CD baths.

Follow the instructions on the package, and follow septic tank precautions if you have one.

Three days or so later, you can add in the Bio-Chelat. Again, follow the instructions on the package, and you can always start slowly and work your way up. Since Bio-Chelat™ doesn't alter the CD, the drops can be added to a single dose of CD. In addition, since the drops have no flavor, it is suitable to add them to water or any other drink your child might consume throughout the day.

Hyperbarics:

Option #1:

Two sessions of 60 minutes each for 20 days for a total of 40 sessions at 1.75 ATA.

Option #2:

One 90-minute session a day for twenty days at 1.75 ATA.

Instead of thinking outside of the box, get rid of the box.

~ Deepak Chopra

Appendix 13

Body Burden
The Pollution in Newborns

When I came across the following investigation I knew it was the perfect precursor to the incredible "Avoiding Autism" by Dr. Anju Usman and Beth Hynes. Babies are exposed to toxins in the womb and this study by the Environmental Working Group demonstrates so perfectly why the advice in the "Avoiding Autism" article is so very crucial to health of our children and for our species as a whole.

Body Burden — The Pollution in Newborns

A benchmark investigation of industrial chemicals, pollutants and pesticides in umbilical cord blood.

Environmental Working Group, July 14, 2005

Summary. In the month leading up to a baby's birth, the umbilical cord pulses with the equivalent of at least 300 quarts of blood each day, pumped back and forth from the nutrient- and oxygen-rich placenta to the rapidly growing child cradled in a sac of amniotic fluid. This cord is a lifeline between mother and baby, bearing nutrients that sustain life and propel growth.

Not long ago scientists thought that the placenta shielded cord blood — and the developing baby — from most chemicals and pollutants in the environment. But now we know that at this critical time when organs, vessels, membranes and systems are knit together from single cells to finished form in a span of weeks, the umbilical cord carries not only the building blocks of life, but also a steady stream of industrial chemicals, pollutants and pesticides that cross the placenta as readily as residues from cigarettes and alcohol. This is the human "body burden" — the pollution in people that permeates everyone in the world, including babies in the womb.

In a study spearheaded by the Environmental Working Group (EWG) in collaboration with Commonweal, researchers at two major laboratories found an average of 200 industrial chemicals and pollutants in umbilical cord blood from 10 babies born in August and September of 2004 in U.S. hospitals. Tests revealed a total of 287 chemicals in the group. The umbilical cord blood of these 10 children, collected by Red Cross after the cord was cut, harbored pesticides, consumer product ingredients, and wastes from burning coal, gasoline, and garbage.

This study represents the first reported cord blood tests for 261 of the targeted chemicals and the first reported detections in cord blood for 209 compounds. Among them are eight perfluorochemicals used as stain and oil repellants in fast food packaging, clothes and textiles — including the Teflon chemical PFOA, recently characterized as a likely human carcinogen by the EPA's Science Advisory Board — dozens of

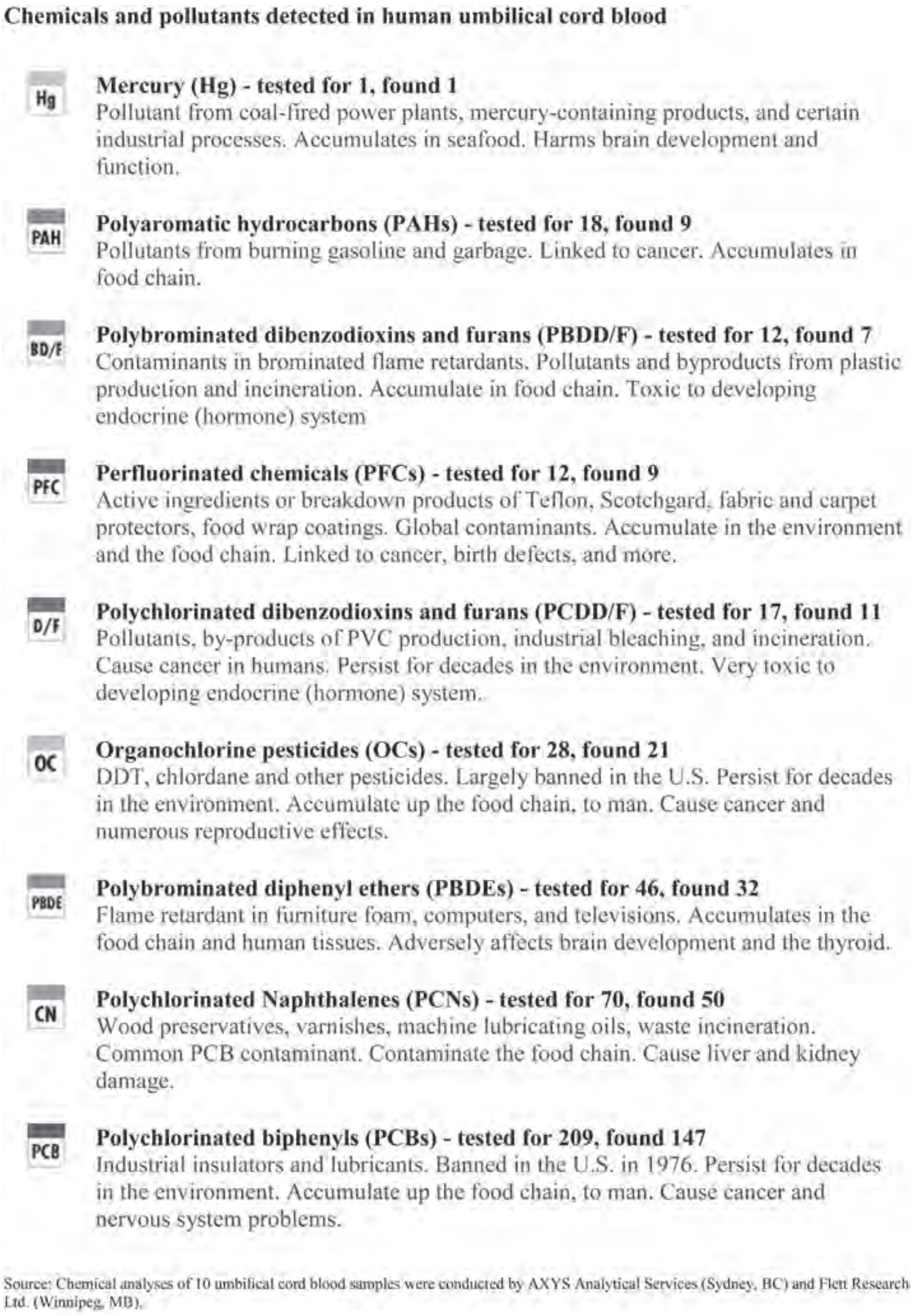

Chemicals and pollutants detected in human umbilical cord blood

Hg **Mercury (Hg) - tested for 1, found 1**
Pollutant from coal-fired power plants, mercury-containing products, and certain industrial processes. Accumulates in seafood. Harms brain development and function.

PAH **Polyaromatic hydrocarbons (PAHs) - tested for 18, found 9**
Pollutants from burning gasoline and garbage. Linked to cancer. Accumulates in food chain.

BD/F **Polybrominated dibenzodioxins and furans (PBDD/F) - tested for 12, found 7**
Contaminants in brominated flame retardants. Pollutants and byproducts from plastic production and incineration. Accumulate in food chain. Toxic to developing endocrine (hormone) system

PFC **Perfluorinated chemicals (PFCs) - tested for 12, found 9**
Active ingredients or breakdown products of Teflon, Scotchgard, fabric and carpet protectors, food wrap coatings. Global contaminants. Accumulate in the environment and the food chain. Linked to cancer, birth defects, and more.

D/F **Polychlorinated dibenzodioxins and furans (PCDD/F) - tested for 17, found 11**
Pollutants, by-products of PVC production, industrial bleaching, and incineration. Cause cancer in humans. Persist for decades in the environment. Very toxic to developing endocrine (hormone) system.

OC **Organochlorine pesticides (OCs) - tested for 28, found 21**
DDT, chlordane and other pesticides. Largely banned in the U.S. Persist for decades in the environment. Accumulate up the food chain, to man. Cause cancer and numerous reproductive effects.

PBDE **Polybrominated diphenyl ethers (PBDEs) - tested for 46, found 32**
Flame retardant in furniture foam, computers, and televisions. Accumulates in the food chain and human tissues. Adversely affects brain development and the thyroid.

CN **Polychlorinated Naphthalenes (PCNs) - tested for 70, found 50**
Wood preservatives, varnishes, machine lubricating oils, waste incineration. Common PCB contaminant. Contaminate the food chain. Cause liver and kidney damage.

PCB **Polychlorinated biphenyls (PCBs) - tested for 209, found 147**
Industrial insulators and lubricants. Banned in the U.S. in 1976. Persist for decades in the environment. Accumulate up the food chain, to man. Cause cancer and nervous system problems.

Source: Chemical analyses of 10 umbilical cord blood samples were conducted by AXYS Analytical Services (Sydney, BC) and Flett Research Ltd. (Winnipeg, MB).

widely used brominated flame retardants and their toxic by-products; and numerous pesticides.

Of the 287 chemicals we detected in umbilical cord blood, we know that 180 cause cancer in humans or animals, 217 are toxic to the brain and nervous system, and 208 cause birth defects or abnormal development in animal tests. The dangers of pre- or post-natal exposure to this complex mixture of carcinogens, developmental toxins and neurotoxins have never been studied.

Chemical exposures in the womb or during infancy can be dramatically more harmful than exposures later in life. Substantial scientific evidence demonstrates that children face amplified risks from their body burden of pollution; the findings are particularly strong for many of the chemicals found in this study, including mercury, PCBs and dioxins. Children's vulnerability derives from both rapid development and incomplete defense systems:

- A developing child's chemical exposures are greater pound-for-pound than those of adults.
- An immature, porous blood-brain barrier allows greater chemical exposures to the developing brain.
- Children have lower levels of some chemical-binding proteins, allowing more of a chemical to reach "target organs."
- A baby's organs and systems are rapidly developing, and thus are often more vulnerable to damage from chemical exposure.
- Systems that detoxify and excrete industrial chemicals are not fully developed.
- The longer future life span of a child compared to an adult allows more time for adverse effects to arise.

The 10 children in this study were chosen randomly, from among 2004's summer season of live births from mothers in Red Cross' volunteer, national cord blood collection program. They were not chosen because their parents work in the chemical industry or because they were known to bear problems from chemical exposures in the womb. Nevertheless, each baby was born polluted with a broad array of contaminants.

U.S. industries manufacture and import approximately 75,000 chemicals, 3,000 of them at over a million pounds per year. Health officials do not know how many of these chemicals pollute fetal blood and what the health consequences of in utero exposures may be.

Had we tested for a broader array of chemicals, we would almost certainly have detected far more than 287. But testing umbilical cord blood for industrial chemicals is technically challenging. Chemical manufacturers are not required to divulge to the public or government health officials methods to detect their chemicals in humans. Few labs are equipped with the machines and expertise to run the tests or the funding to develop the methods. Laboratories have yet to develop methods to test human tissues for the vast majority of chemicals on the market, and the few tests that labs are able to conduct are expensive. Laboratory costs for the cord blood analyses reported here were $10,000 per sample.

A developing baby depends on adults for protection, nutrition, and, ultimately, survival. As a society we have a responsibility to ensure that babies do not enter this world pre-polluted, with 200 industrial chemicals in their blood. Decades-old bans on a handful of chemicals like PCBs, lead gas additives, DDT and other pesticides have led to significant declines in people's blood levels of these pollutants. But good news like this is hard to find for other chemicals.

The Toxic Substances Control Act, the 1976 federal law meant to ensure the safety of commercial chemicals, essentially deemed 63,000 existing chemicals "safe as used" the day the law was passed, through mandated, en masse approval for use with no safety scrutiny. It forces the government to approve new chemicals within 90 days of a company's application at an average pace of seven per day. It has not been improved for nearly 30 years—longer than any other major environmental or public health statute—and does nothing to reduce or ensure the safety of exposure to pollution in the womb.

Because the Toxic Substances Control Act fails to mandate safety studies, the government has initiated a number of voluntary programs to gather more information about chemicals, most notably the high production volume (HPV) chemical screening program. But these efforts have been largely ineffective at reducing human exposures to chemicals. They are no substitute for a clear statutory requirement to protect children from the toxic effects of chemical exposure.

In light of the findings in this study and a substantial body of supporting science on the toxicity of early life exposures to industrial chemicals, we strongly urge that federal laws and policies be reformed to ensure that children are protected from chemicals, and that to the maximum extent possible, exposures to industrial chemicals before birth be eliminated. The sooner society takes action, the sooner we can reduce or end pollution in the womb.

Copyright © Environmental Working Group, www.ewg.org. Reprinted with permission.

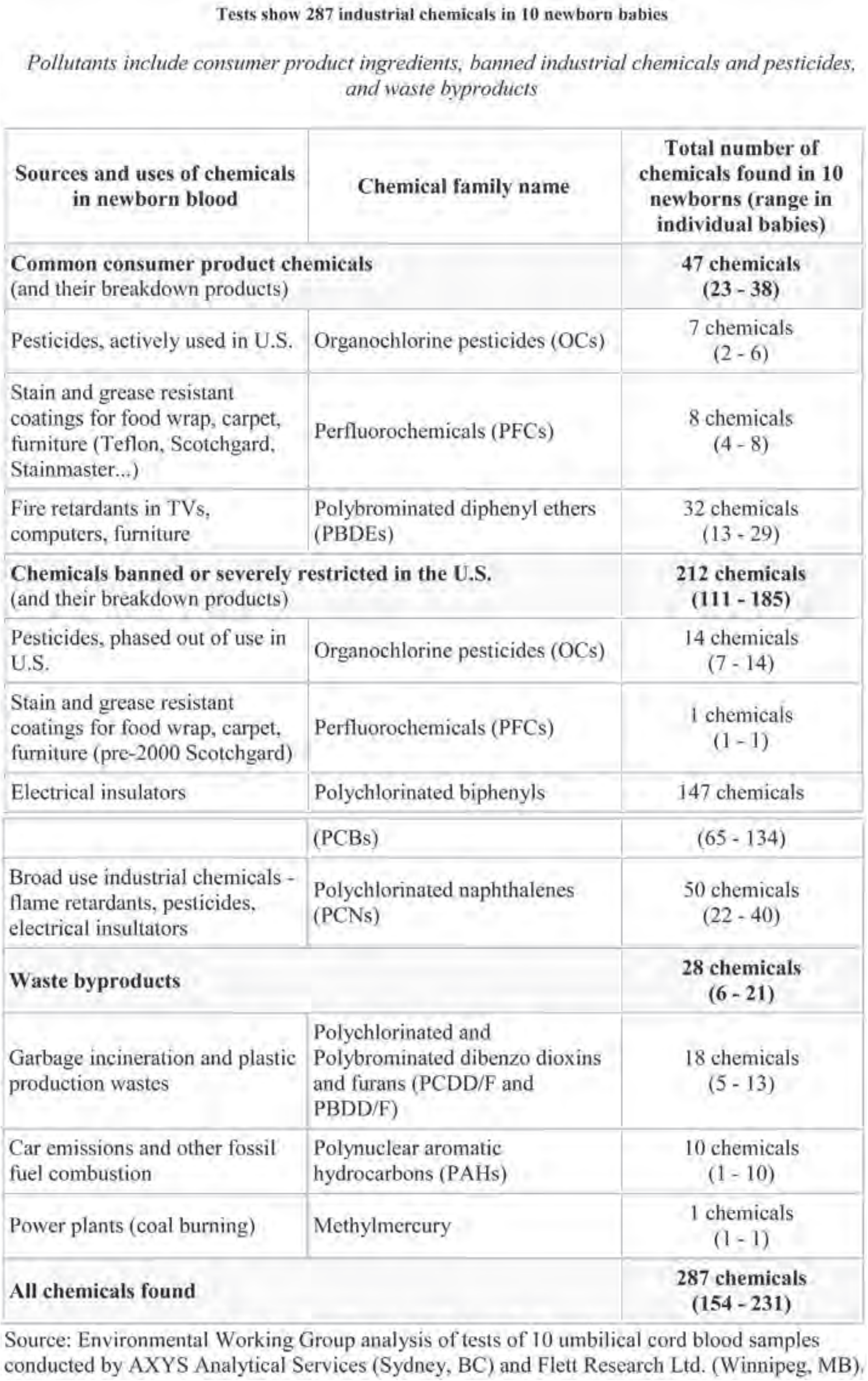

Tests show 287 industrial chemicals in 10 newborn babies

Pollutants include consumer product ingredients, banned industrial chemicals and pesticides, and waste byproducts

Sources and uses of chemicals in newborn blood	Chemical family name	Total number of chemicals found in 10 newborns (range in individual babies)
Common consumer product chemicals (and their breakdown products)		**47 chemicals** **(23 - 38)**
Pesticides, actively used in U.S.	Organochlorine pesticides (OCs)	7 chemicals (2 - 6)
Stain and grease resistant coatings for food wrap, carpet, furniture (Teflon, Scotchgard, Stainmaster...)	Perfluorochemicals (PFCs)	8 chemicals (4 - 8)
Fire retardants in TVs, computers, furniture	Polybrominated diphenyl ethers (PBDEs)	32 chemicals (13 - 29)
Chemicals banned or severely restricted in the U.S. (and their breakdown products)		**212 chemicals** **(111 - 185)**
Pesticides, phased out of use in U.S.	Organochlorine pesticides (OCs)	14 chemicals (7 - 14)
Stain and grease resistant coatings for food wrap, carpet, furniture (pre-2000 Scotchgard)	Perfluorochemicals (PFCs)	1 chemicals (1 - 1)
Electrical insulators	Polychlorinated biphenyls (PCBs)	147 chemicals (65 - 134)
Broad use industrial chemicals - flame retardants, pesticides, electrical insultators	Polychlorinated naphthalenes (PCNs)	50 chemicals (22 - 40)
Waste byproducts		**28 chemicals** **(6 - 21)**
Garbage incineration and plastic production wastes	Polychlorinated and Polybrominated dibenzo dioxins and furans (PCDD/F and PBDD/F)	18 chemicals (5 - 13)
Car emissions and other fossil fuel combustion	Polynuclear aromatic hydrocarbons (PAHs)	10 chemicals (1 - 10)
Power plants (coal burning)	Methylmercury	1 chemicals (1 - 1)
All chemicals found		**287 chemicals** **(154 - 231)**

Source: Environmental Working Group analysis of tests of 10 umbilical cord blood samples conducted by AXYS Analytical Services (Sydney, BC) and Flett Research Ltd. (Winnipeg, MB).

The ultimate ignorance is the rejection of something you know nothing about and refuse to investigate.

~ Dr. Wayne Dyer

Appendix 14

Avoiding Autism

By Anju Usman, MD and Beth C. Hynes JD, MBA

Whether or not you are the parent of a special needs child, please pay special attention to this article, we do believe autism is totally preventable and that starts with pregnancy. Here's to your health!

Every parent's dream is to have a healthy baby. Over the past ten years of working with families devastated by the diagnosis of autism, we, along with our patients' parents, have realized one key fact: if we knew then what we know now, the diagnosis of the disease called autism could very well have been prevented. Considering that the cause(s) of the current epidemic of autism spectrum disorders (ASD) are unknown (and scant research dollars are being dedicated to finding the most likely culprits), there is no surefire method to avoid having your infant regress into this devastating state. That sad fact, however, does not mean that there are no proactive steps parents can take to try to mitigate the chances of this disease overtaking their baby and stealing their hopes and dreams for that child's life.

Having reviewed the medical and laboratory test results of thousands of children afflicted with ASD, clinicians using a biomedical approach to treating autism see repeating patterns in these children—impaired methylation and detoxification, mitochondrial dysfunction, gastrointestinal distress, and immune dysregulation compounded by chronic viral, fungal and bacterial infections and a burden of heavy metals and other toxins. From this perspective, we have come to view autism as a multifactorial medical disease that can be treated and overcome and, therefore, possibly avoided. Coming straight from the heart, this advice is for thinking, active parents and prospective parents who want to know what they can do to reduce the risk that their infant will regress into autism. In our practice, we have helped many parents go on to have healthy children even though they already have a child with autism. Considering that the information provided here is not harmful but, to the contrary, is helpful in promoting improved maternal and fetal health, there is little downside to pursuing these strategies while simultaneously minimizing the risk of having your child slip away into the autism epidemic. According to the latest statistics out of California, that epidemic continues to grow from the already shocking 1 out of 150 children. The aim here is to provide an analytical construct that will guide parents in making choices for both their own health and that of their baby before, during, and after pregnancy.

Think of the analogy of the straw that broke the camel's back. Human beings start life as infants with some "straws" in their saddle – these are predispositions to be harmed by certain chemicals, to be susceptible to certain diseases, or to have a less robust detoxification capacity. The goal of gestating and parenting an infant in today's toxic world is to not accumulate additional straws during your baby's gestation and infancy, thereby avoiding adding the straw that breaks the camel's back. Autism seems to occur when the infant's body is so overwhelmed by toxins, viruses and/or pathogenic elements that typical development is arrested and/or derailed.

Children diagnosed with autism have a variety of medical problems, the underpinnings of which can be ameliorated and/or healed. The symptoms associated with autism occur through malfunctioning of the infant immune, gastrointestinal, and central nervous systems ("Infant Body Systems") – all of which, when functioning properly, contribute to the healthy development of the brain. Viruses, heavy metals, toxins, fungi, and bacteria together form a weapon that disrupts the development of, and harms the functioning of, these critical Infant Body Systems. As researchers have recently found, "a single toxicant may promote different immune-associated diseases that are dependent upon the specific window of early life exposure, the gender of the exposed offspring, and the genetic background of the offspring." Our construct will consist of actions you can take to promote the healthy development of these Infant Body Systems as well as steps you can take to avoid upsetting the healthy development of those systems.

Due to the symbiotic relationship between a mother and her developing fetus, care must be taken by the mom before, during, and after pregnancy (if nursing) to both avoid exposure to harmful elements and to promote optimal maternal detoxification processes so that the host body remains as clean an environment as possible within which vigorous infant development can unfold.

Only recently have researchers begun closely linking environmental factors with developmental delays in our children. A host of studies show how environmental factors including diet, nutrition, air quality, clean water, heavy metals (e.g., mercury, lead, arsenic, and aluminum), xenobiotics (e.g., chemicals, pesticides, plasticizers, and perchlorates), pharmaceuticals, and vaccinations can change your child's inborn risk.

This area of study is called epigenetics – the molecular changes in our DNA caused by environmental factors. In 2003, researchers at Duke University found that rats carrying a gene for obesity, cancer, and diabetes, when fed a diet high in B-12, folic acid, choline, and methyl-rich foods, had babies that were born healthy and with a reduced risk of developing disease; this research revealed that these genetic changes persisted across four generations. Thus, when you eat a healthy diet, you are affecting not only the health of your own kids but also your grandchildren, great-grandchildren, and great-great-grandchildren. Researchers are finding hundreds of unwanted and potentially toxic chemicals in the placenta and cord blood of newborns as well as in mothers' breast milk. That means that today's children are born with a toxic burden their infant bodies must contend with from day one. Studies out of China, Texas, and San Francisco show that if pregnant mothers can decrease their exposure to pollutants during gestation, they can improve health outcomes for their children – especially brain development. Compounds such as mercury, cadmium, nickel, trichloroethylene, and vinyl chloride in the air around the birth residence caused a 50% increase in the risk of autism.

Numerous studies document the neurotoxic effects of heavy metals – especially dangerous to the developing brain is mercury. Yet, many of our children are actually injected with the mercury-containing preservative thimerosal in their infant vaccines. Although thimerosal has been removed from, or perhaps phased out of, many childhood vaccines, it still remains in the flu vaccine, which is recommended by health authorities to pregnant woman as well as to children under age 2. Mercury exposure is such a concern that pregnant women are advised by health authorities not to eat fish because of potential mercury contamination – yet injecting thimerosal is much more damaging than the exposure suffered through eating it in food such as fish. Vaccines also contain aluminum, formaldehyde, and other toxic elements as well as live viruses that, in a subset of infant bodies, can trigger harmful impacts. Heavy metals have also been identified as factors affecting human fertility. Diagnosing and reducing the heavy metal burden of women improved the spontaneous conception chances of infertile women. Women with many dental amalgams had a higher incidence of miscarriages and a higher excretion of mercury when given the chelating agent DMPS (2,3-Dimercapto-1-propanesulfonic acid). It has been found that DMPS was a useful and complementary method to increase fertility compared to hormone therapy in infertile women.

With all this evidence mounting against various environmental toxins affecting human development, an international assembly of scientists, doctors, toxicologists, and researchers sponsored by the

World Health Organization (WHO), the European Environmental Agency, the Centers for Disease Control (CDC), and the National Institutes of Health (NIH) gathered in the Faroe Islands in May 2006. Their consensus statement was clear: "it is time to take action, now." As expected, the declaration highlighted the results of hundreds of studies that showed early fetal exposures to toxic substances, even at low concentrations, can cause health problems later in life. There are a growing number of chemicals that affect the developing embryo, fetus, and infant, including the pesticides DDT, atrazine, methoxychlor, and vinclozolin. Also suspect are plastics and the epoxy resin bisphenol A, plasticizing agents called phthalates, mercury, lead, arsenic, organotoxins, polychlorinated biphenols (PCBs), carbon monoxide, smog, tobacco smoke, and alcohol. The health problems caused by exposure to these substances comprise a dizzying array of maladies including cancer, diabetes, obesity, asthma, and allergies as well as reproductive, cardiovascular, neurological, cognitive, endocrine, psychological, immune, and respiratory troubles.

Other troubling chemicals include parabens (found in various creams and lotions) and triclosan, an antibacterial agent (found in toothpaste, soaps, and cleaning products). Although parabens and triclosan are not ingested, absorbing chemicals through the skin can be far more dangerous than swallowing because transdermal applications are transported directly to the bloodstream. Most women absorb 51 pounds of chemicals every year from their cosmetics alone! Additional exposures emanate from daily lifestyle, profession, and locale. For those elements for which we have the flexibility to make choices, we must choose wisely. After all, it is not just our health that is being affected by our choices but the health of our children and progeny as well.

Basic principles to guide you in decision making surrounding pregnancy are as follows:

1. You are what you eat (and drink); therefore, make healthy choices in what you consume.
2. Skin is the largest organ in the body, so be very careful of what you rub into yours and your baby's.
3. Think beyond "green": What is "green" for the environment is not always what is healthiest for the body, but what is healthiest for the body is always "green."

Suggestions to promote these principles are below and are cumulative, so keep following each set of suggestions as this journey into and through parenthood progresses.

Pre-Pregnancy

While planning a pregnancy, we recommend that you clean up any toxicity in your body and begin to follow a more organic, healthy lifestyle. Remember, the less toxic you are, the better for you and your future baby. Undertake a sequential detoxification program that targets the liver and colon; this type of program can take six months or more and should not be done while pregnant. In a careful manner, with an experienced dentist, remove amalgams from your teeth, which also can take six months or more. Take the time now to find organic, nontoxic makeup, hair, and body products that you like and start integrating them into your daily life. Ask your doctor to run some tests to determine any additional specific supplementation you may need to optimize levels within your body; a good place to start is your copper-zinc ratio (1:1 is the ideal ratio), thyroid function (TSH, free t3, free t4), Vitamin D 25 OH, vitamin A levels, and total cholesterol (low cholesterol is associated with preterm births).

Next, clean up your living and working environment. Remove all harmful chemical cleaning agents from your cleaning routine at home and at the office, and instead use cleaning products labeled level 1 by the EPA. Do not forget to include products for dishwashing and clothing detergent in your cleanup, and avoid toxic dry cleaning as much as possible. Finally, improve your nutrition with a targeted vitamin supplementation program to include omega-3 essential fatty acids, sublingual methyl B-12, folinic acid, vitamin D3, zinc, and antioxidants.

DO	DON'T
• Eat organic, hormone-free food • Drink organic green tea, filtered water, and antioxidant rich organic juices • Use stevia, raw organic honey, and xylitol as sweeteners • Go for walks and get some sunshine daily • Use aluminum-free natural deodorant in them • Use natural hennas to color your hair • Use cast iron, glass, or stainless steel cookware • Use chemical-free cleaning products in your home	• Consume fish or foods w/MSG or food dyes • Drink soda, carbonated beverages, or alcohol • Consume artificial sweeteners • Be exposed to lawn chemicals or second-hand smoke • Use moisturizers or makeup with chemicals or parabens • Use chemical dyes, perms, or other such hair treatments • Cook with pans that are non-stick, Teflon coated, or made from aluminum

During Pregnancy

We would suggest that you find a holistic health practice to guide you during pregnancy and delivery as there are many decisions to make during this time. In general, we suggest avoiding medications to the extent possible as well as acetaminophen because it hinders normal detoxification. Add zinc, calcium, essential fatty acids, and prenatal vitamins to your daily supplement intake. Discontinue use of nail polish and any makeup (including lipstick) products that contain parabens and other toxins. Use fluoride-free toothpaste as fluoride interferes with iodine metabolism, which is an issue implicated in mental retardation worldwide.

DO	DON'T
• Yoga and engage in stress management techniques, such as massages and listening to soothing music • Eat fermented foods, cook with organic coconut oil, use organic raw apple cider as salad dressing, and consume healthy fats, and cold pressed oil • Use natural remedies for pain, like homeopathic arnica • Drink kombucha and take probiotics • Run an air filter in your bedroom while you sleep	• Start a rigorous exercise program, sit in a sauna, or get dental work (not even cleanings) • Wait until labor arrives to discuss NOT subjecting your baby to vaccines • Talk on a cell phone without a headset or work with a laptop computer on your lap • Undertake a major home renovation (concern is for lead and other toxins in the process)

Infancy

Healthy babies grow and develop perfectly when nature's biochemical processes are allowed to unfold uninterrupted. The sequence of events that leads an infant to begin to focus, develop gross and fine motor skills, begin to speak and to walk is based on a delicate but deliberate series of biochemical processes and chemical interactions. A primary goal of raising a healthy infant is to avoid disruption or interference with this sequence. Babies look fragile and, indeed, they are. An infant's immune system, which guards the rest of the body from harm, is immature and requires time and peace to mature maximally. Thus, help your infant progress uninterrupted by avoiding the introduction of toxins, viruses, allergens, and heavy metals into their bodies. In light of this, parents should consider an alternate, gradual vaccine schedule of carefully thought through vaccines for infants being raised in the domestic United States.

If you have amalgams and plan to nurse, you should send a sample of your breast milk to a specialty lab for heavy metal testing. If you plan to use formula, use those containing DHA (docosahexanenoic acid), an essential fatty acid critical to the healthy development of the central nervous system. In terms of food introduction, organic baby food is recommended. Start feeding with organic rice cereal. Avoid the introduction of soy, gluten, or dairy until after the baby turns two years of age. After one year, supplement your baby with a quarter teaspoon of mercury-free cod liver oil – again this is to increase the supply of omega-3 essential fatty acids.

In terms of coping with baby colds and other minor illnesses, unless symptoms are severe, less is more when it comes to treatment. For fevers over 101˚F, treat with a tepid bath or dye-free ibuprofen. Fevers, while nerve-wracking for new parents, are the response of a healthy immune system reacting to kill off an invading virus through heat. Antibiotics should be used sparingly and only for confirmed bacterial infections (they do not alleviate viral infections). Remember, antibiotic use disrupts the normal gut flora and promotes the overgrowth of yeast and resistant organisms that, in turn, harms the optimal functioning of the immune system. Bear in mind that most ear infections are viral and are thus not treatable with antibiotics. Use homeopathic ear drops to help ease the symptoms associated with ear infections and colds.

DO	DON'T
• Invest in an organic baby matress, bedding and pillows and hypo allergenic encasements • Bathe your baby daily in warm filtered water — enjoy the experience with your baby! • Feed with all organic and hormone-free products • Feed baby using glass bottles • Wearing a hat, walk outside with baby to get 10 to 15 minutes of sunshine daily • Run an air filter in baby's bedroom	• Clothe the baby in pajamas soaked in flame retardant chemicals • Use soaps, moisturizers, or other "baby products" on the skin as such items are unnecessary and contain harmful chemicals • Introduce dairy, gluten, or soy until after age two • Feed baby from plastic bottles or cups or microwave formula or breast milk • Take your infant into polluted and heavily populated locations

Source: http://www.autismfile.com/what-is-autism-facts/autism-symptoms/avoiding-autism

Reprinted with permission.

When a truth is not given complete freedom,
freedom is not complete.

~ Vaclav Havel

Appendix 15

RECIPES

Cooking with Ana

Since 2001 Ana Bobadilla has been a chef to Kerri Rivera and family, including her son Patrick, who has a diagnosis of regressive autism. Over the years she has mastered gluten-free/casein-free to vegan to Atkins and back again. Her dishes always impress and nourish guests and family members alike.

One of the most important pieces to the autism recovery puzzle is diet. As Kerri always says, if your goal is to heal autism, you must do *The Diet.* However, that is easier said than done for many families whose children have self-limited to grilled cheese sandwiches and pizza.

A book is in the works containing dozens of delicious recipes using natural ingredients that can be found in almost any country. Ana has found unique ways to make foods that children on the spectrum can enjoy, without breaking the bank. The ingredients that she uses can be found all over the world, not just the US. They are presented here in a very approachable way so even if you are not a whiz in the kitchen your child will be able to enjoy what you serve.

The following is a sneak peek from *Cooking with Ana*:

Almond Milk

Amount: 1 L of milk

- 1 cup of peeled almonds
- 1 L of filtered water
- Stevia and vanilla to taste

Preparation:

To a saucepot of boiling water, add the almonds. Remove immediately from the fire. Allow almonds to soak for 20 minutes, remove them from water, rinse, dry and blend in a liter of filtered water. Strain the liquid through a piece of cheesecloth until all the water is drained. Add agave syrup and vanilla to taste.

Vegetable Fries

3 Servings

Ingredients:

- 3 Carrots
- 1 cup oil
- Salt to taste

Preparation:

Peel the carrots and then form very thin slices using the same peeler. Pat dry with a piece of cloth to remove any excess water. Heat oil in a frying pan and fry carrots. When frying be very careful because they can quickly burn, be sure to have a slotted spoon ready to remove from the oil as soon as they are crisp. After frying carrots remove excess oil with a paper towel and sprinkle some salt on top of them.

You can also do this recipe with banana, potato, sweet potato or jackfruit. This is a great alternative to French fries or if you want to eat something crunchy.

Beef or Chicken Tostadas

4 Servings

Ingredients:

- 5 ground chicken breasts or beef filets
- 1 cup of rice or coconut flour
- 1 cup of oil
- Salt and pepper to taste

Preparation:

In a mixing bowl, combine the ground chicken, flour, salt, pepper and a ½ cup of oil. Mix thoroughly until a uniform dough results. Take approximately 2 tablespoons of the dough and form small tortillas/flat circles with it. They should be approximately ¼ inch think and 4 inches wide. Fry in a layer of oil until cooked through, and a thin outer crust forms, or bake in a 350 degree oven, after pouring a dollop of oil on top of each tostada.

Coconut Bars

15-18 bars

Ingredients:

- 3 cups of vegan butter (or coconut oil)
- 1 cup of honey
- 1 cup of raisins
- 1 cup of grated of coconut
- 1 cup of walnuts
- 1 tablespoon of salt

Preparation:

In a saucepan over medium heat combine honey and butter, stirring constantly for a few minutes until thoroughly combined. Remove from heat and mix with all other ingredients. Spread in a 9x13 casserole pan without oil, and bake in an 350°F (180°C) over for 25 minutes until it sets. Allow to cool, and cut into small rectangles.

Get people back into the kitchen
and combat the trend toward
processed food and fast food.

~ Andrew Weil

Appendix 16

Websites We Like

General:

- **Our website containing the latest information about our research, videos, forums, etc.**
  www.cdautism.org

- **The Spanish version of our website.**
  www.autismo2.com

- ***Curando el Autismo* is a Puerto Rican organization created by three mothers of children recovered from autism who have an interest in helping the community learn about their options.**
  www.curandoelautismo.com

- **Defeating Autism - Non-Profit Organization in Venezuela that utilizes Kerri's Protocols.**
  www.fundacionvenciendoelautismo.blogspot.mx

- **Autism Research Institute originally founded by Dr. Bernard Rimland. Home of the ATEC survey.**
  www.autism.com

- **Our public Facebook page.**
  www.facebook.com/groups/AutismCD

- ***CD Health*: Facebook group dedicated to using the protocol for non-autism related issues.**
  www.facebook.com/groups/mojoother

Diet:

- **Website forcused on the *Specific Carbohydrate Diet*™ (SCD) in English.**
  www.pecanbread.com

- **Top 10 FAQs about the *Specific Carbohydrate Diet* (SCD) in Spanish.**
  www.pecanbread.com/pandenuez/10preguntas.html

- **Autism Network for Dietary Intervention**
  www.autismndi.com

- **Gluten Free/Casein Free Diet Intervention website.**
  www.gfcfdiet.com

- **Intestinal health through diet with the Specific Carbohydrate Diet™**
  www.breakingtheviciouscycle.info

- **Resource for naturally healing digestive disease, reducing stress and living a long, healthy life.**
  www.scdlifestyle.com

- ***The Feingold® Association* is a non-profit organization helping provide dietary management and generate public awareness of the role food and synthetic additives play in behavior, learning and health problems.**
  www.feingold.org

Chlorine Dioxide (CD):

- **Jim Humble's MMS (CD) website.**
  www.jimhumble.org

- **Another popular site about MMS (CD).**
  www.mmswiki.org

- **Spanish site selling various books about MMS (CD)**
  www.voedia.com

- **Andreas Kalcker's website (in English)**
  www.andreaskalcker.com

- **CD research site by Andreas Kalcker (in Spanish)**
  www.medicasalud.com

- **A huge public health forum with thousands of discussions about every imaginable health topic. Also includes surveys about various topics.**
  www.curezone.com

- **Good sources of CD and other related supplies.**
  www.wps4sale.com
  www.mightyguts.com

- **The latest information on CDH.**
  http://www.mmsinfo.org/infosheets/infosheet_cdh.pdf

Hyperbarics:

- **Good source of information about Hyperbaric Oxygen Therapy.**
  www.hyperbaric-oxygen-info.com

Other Valuable Source of Information:

- **PubMed comprises of more than 23 million citations for biomedical literature from MEDLINE, life science journals, and online books.**
  www.ncbi.nlm.nih.gov/pubmed

- **Parents Helping Parents - P.A.N.D.A.S. Network: A Resource Library of Medical Studies and Case Histories**
  www.pandasnetwork.org

- **International OCD Foundation**
  www.ocfoundation.org/PANDAS/

Videos / Audio of Interest:

- **Video Testimonials. This link contains various health testimonials related to MMS, many are about autism.**
  www.youtube.com/mmstestimonials

- **AutismOne 2012: 38 Children Recovered in 20 months with MMS by Kerri Rivera**
  www.autismone.org/content/38-children-recovered-20-months-mms-0

- **Video about Kerri, produced by Daniel Bender in 2012.**
  www.youtube.com/watch?v=HnCiX5bIl-Y

- **Kerri speaks in Bulgaria, 2013 (Voice in English, slides in Bulgarian)**
  www.youtube.com/watch?v=4OxNGIEWzds

- **Kerri on the Robert Scott Bell Show**
  http://www.youtube.com/watch?v=-1zJhr5VhxE

- **MMS Autism Webinars**
  http://www.ustream.tv/channel/mms-autism-webinar%20

- **Patrick Timpone interviews Kerri Rivera (8/15/2013)**
  http://oneradionetwork.com/health/kerri-rivera-healing-the-symptoms-known-as-autism-children-are-not-born-with-autism-so-they-should-not-die-with-it-august-15-2013/

- **Patrick Timpone interviews Kerri Rivera (11/18/2013)**
  http://oneradionetwork.com/health/kerri-rivera-encore-interview-healing-the-symptoms-known-as-autism-using-parasite-protocols-and-chlorine-dioxide-november-18-2013/

- **Bulletproof Executive Radio with Dave Asprey (11/27/2013)**
  http://www.youtube.com/watch?v=R7s7lYj2SCs

Appendix 17

Direct Help from Kerri Rivera

Consults

Since the release of the first edition of the book, we have found that when families decide to implement The Protocol some do not need further assistance beyond reading the information in this book, using our online support groups, YouTube videos, and attending conferences. However, there are some of us who need a more personalized approach and support to navigate the information, or would like ideas on dosing or timing.

For those of us who need or want that additional support, consults are perfect. We can work together to map out a plan for the parent/family and child, brainstorm through challenges, or just know that you will never walk alone. Whatever the case may be I am available if you need me. I am fluent in English and Spanish, and we can use a translator for other languages. We can speak over the phone or on Skype.

As time goes on, more and more practitioners will be able to support families while using this protocol. In the meantime, If you would like to set up a consult with me, please click on the "Consult Kerri" button on...

www.CDAutism.org

...and follow the instructions there.

> Kerri Rivera's undivided attention + discussing your child + $100 = priceless. Tonight, instead of living it up in Mexico, Kerri spent her Friday evening on the phone doing a consult with me. Not only was the consult informative but also highly enjoyable. Kerri's energy is contagious, her vision boundless and her sense of humor edgy (just the way I like it). Kerri thank you for rocking our world! You are an amazing lady and my family is eternally grateful for your love of our children. Peeps, if you are new, stuck, confused or you want a better understanding of the protocol PLEASE, PLEASE, PLEASE take advantage of this opportunity. I've done a million consults over the years but this was by far my favorite.

About the Author

Kerri Rivera is a native of Chicago, but has lived the last 19 years of her life in Puerto Vallarta, Mexico with her husband. Both of her sons were born there, and 13-year-old Patrick is currently in autism recovery. Kerri is the founder of AutismO2, a non-profit autism clinic based on what is formerly known as the *Defeat Autism Now!* approach. Located in Puerto Vallarta, the clinic opened in 2006. Kerri is the clinic's biomedical consultant as well as a consultant for Curando El Autismo (Latin America), and Venciendo el Autismo (Venezuela).

At the request of Dr. Bernard Rimland, she was responsible for translating the *Defeat Autism Now!* protocol to Spanish and donating it to the Autism Research Institute, so it could be applied throughout Latin America. Since then she has helped over 3,500 families in over 58 countries to improve the lives of their children on the autism spectrum. She has become the foremost expert on the use of CD (chlorine dioxide) for spectrum disorders. Since the addition of CD to the biomedical protocol she has seen 115 children recover from autism in the past 2.5 years.

Since 2007 Kerri has lectured internationally on the Biomedical Protocol for autism and CD for autism. She has been featured on *One Radio Network* with Patrick Timpone, *The Bulletproof Executive* with Dave Asprey, *The Mother Cub Show*, *Thought for Food*, *Enlightened Health Radio* with Justin Elledge, *Voice America*, and *The Robert Scott Bell Show*. She is the Mexican Liaison for AutismOne. Kerri graduated as a Certified Homeopath in June of 2013.

Our greatest weakness lies in giving up.

The most certain way to succeed is always to try just one more time.

~ *Thomas Edison*

The world of autism is rocking again. Kerri Rivera has done it. In this comprehensive book she has condensed and simplified the core elements of the biomedical approach, those that work 90% of the time on 90% of the children. Her focus on using a unified and affordable approach to reduce the body burden of chronic infections and infestations is addressing the very core issue of most autistic children and adults.
It has made her method the most powerful tool to this day to help many children recover safely, inexpensively and solidly. Her research is validated by the enormously positive response from parents from all over the world.

—Dietrich Klinghardt MD, Ph.D.

EXCITING UPDATES TO THIS EDITION INCLUDE:

A new method of CD preparation • New parasite protocol charts
Lunar calendar • Many new testimonials
A section on gluten's role in molecular mimicry and autoimmunity
Information on helping older or violent / aggressive children

In the seven months since the release of the first edition of *Healing the Symptoms Known as Autism*, 22 more children have lost their autism diagnosis and returned to a state of health, for a total of 115 recovered children and counting. Hundreds more have lost ATEC points, as well as shown many cognitive, behavioral, emotional and physical gains thanks to the protocol outlined in these pages. This edition includes all protocol updates and a whole lot more information that can benefit families of children and adults on the spectrum.

In 3 years, 115 children previously diagnosed with regressive autism were able to shed their diagnosis, their symptoms and return to an overall state of health and vitality. If you have any interest in finding out how they were able to heal you need to read this book. Kerri has a heart for children with autism, and she cares greatly about recommending safe, economical interventions that are helping to confront an epidemic.

—Teri Arranga
Executive Director of AutismOne (www.autismone.org)

With over 180 TESTIMONIALS from around the world!

ISBN 978-0-9892890-2-3
52499>
9 780989 289023

Educate-Yourself

The Freedom of Knowledge, The Power of Thought ©

Current News | Introduction | Colloidal Silver | Chemtrails | Sylphs | Emerging Diseases | Forbidden Cures | Ozone | Immunity Boosting | Nutrition | Tone Gen
Mind-Body Connection | Ozone | Bioelectrification | Story on Drugs | Vaccine Dangers | Cancer | Newsletter | New World Order | NWO News | Pam Schuffert
James Casbolt | Phil Schneider | Al Bielek | Trevor Constable | Mind Control | Brice Taylor | Ted Gunderson | Metatech.org | Free Energy | Tim Hicks & Phil Ledoux
Dr. Robert Bitzer | T. Lobsang Rampa | Ruth Drown | ZS Livingstone | David Brandt | Global Warming | The CIA | Veterans Awaken | Eustace Mullins
Bill Cooper | Depleted Uranium| Dowsing | Police & Tasers | British Israel | Born Again Brainwashing | Amy Goodman Gatekeeper | Rethinking Noam Chomsky
Red Elk | Project Blue Beam | Bush Family & Nazis | Otto Skorzeny | Insights on Aliens | Cell Towers | WiFi/Cell Phone Dangers | EMF Radiation News
CPS/DCF Tyranny | Adrenal Burnout | The Women Warriors | Orgone Adventures | Dr. John Coleman | Railroading Dr. Jeffrey MacDonald | Henry Makow |
Radio Interviews | Ten Best Sites | Metal Free Dentistry | Dr. Hulda Clark Books | Planet X Sequel | 'Undocumented Immigrants' | Donald Marshall & Cloning Statio
War on Terror | Tavistock | U.S. Concentration Camps | FEMA | Aliens Are Coming! | Guiding Principles | Global Warming | Gang Stalking | Monoatomic Gold
Seth/Jane Roberts | Power of Thought | Spiritualism | Hope | Healing Thought Forms | Vanquish Fear | Prevent Alien/Demon Attacks | Rockefeller File | War a Racke
Letters | Sodom-LGBT Exp'd | 21 Illuminati Goals | Strawman Explained | Daily Blog | | Hydrogen Peroxide | Protocols of Zion | Eric Jon Phelps Exposed |
Israeli Soldier's Story | Home | Japan Radiation Hysteria | Agenda 21 | Treat Parasites | Swine Flu Hoax/Vaccine | Links | *Contact Us/ E-mail* | Jane Tripp Time Travel | Dis

Want to Contact the Editor? First read this

How to Detect & Treat a Parasitic Infection

> *"Ninety percent (**90%**) of the world population is infected with one or more parasites in their body, being able to coexist in the same host with up to five different types. The danger comes when the balance is upset, within the host, skyrocketing the number of parasites and the host starting to show signs of serious illness and even death in some cases. Parasitic infections caused by parasitic worms do not always cause disease in man, a number of carriers have been found who are fully healthy. Parasites are silent murderers claiming unsuspecting victims living in a people world that does not even imagine they exist.* "

By Dr. Andreas L. Kalcker, Ph.D & Miriam C. Maceda
English Translation: Mercy Acevedo, Proof Reading: Karl Wagner
http://educate-yourself.org/cn/detectandtreatparasites12aug12.shtml#top
August 12, 2012, E-Y posted Feb. 14, 2015

How to Detect & Treat a Parasitic Infection (Feb. 14, 2015)

Facebook Censorship
To post this article on **Facebook**, link to the TinyUrl seen below. Facebook will remove any article identified as coming from educate-yourself.org
http://tinyurl.com/hl54enf

[Ken Adachi Note: Dr. Hulda Clark will eventually be acknowledged by history as one of the greatest medical sleuths of the modern era for her recognition of the pivo role which **parasites** play in all disease conditions. Her Cure for All Diseases book goes into many details about eliminating parasites, but she her 18 page published synopsis called "New Concepts" explained her latest stretegies for detecting and eliminating parasite infestation that you will find very informative and useful.

I did not realize that Andreas Kalcker had a web site until I recently watched a few videos that he and Kerri Rivera had made in 2014 at an autism conference where he made reference to his web site. Once I found it, I was delighted to find this gem on getting rid of parasites with simple home remedies. Andreas made an important speculation in one of his videos that the reason that some vaccinated children get autism and others don't, may be the degree and extent of the vaccine's complex *inter-reaction* with parasites in the child's body (as **all kids with autism are *loaded* with parasites**, that once removed, results in a kid without the symptoms of autism). **90 of the world's population has parasites** (including you) and most of us don't know it because we don't see any outward signs of the parasite's presence - except the asthma, or fatigue, or bloating, or diabetes, (or the mysterious anal itch that occurs periodically about every two weeks at full moon or new moon) etc. As you will lear below, **it's not difficult to remove parasites using the inexpensive, natural remedies recommended here**, although you may have to keep at it for 6 months to a yea until all of your parasites (and their eggs) are completely removed. Re-infection can happen easily, so we have to keep up a maintenance schedule for the rest of our liv if we want to enjoy a parasite-free (and healthier) existence. You won't discover how many parasites are in *your* body until you make the effort to take the herbal remed described and see the evidence yourself in the toilet bowl. Once that happens, you'll be on your way towards a healthier and happier you because you now know what y have to do. I took the liberty to slightly tweak the English and added some links to make it easier to find the parasite remedies online. I hope that those of you who mak the effort and see the results will send me an email to report on your discoveries. See the Dr. Hulda Clark parasite examination videos embedded at the bottom of this article.]

http://andreaskalcker.com/en/health/parasite/116-parasite.html

Version 1.1 – Revision 1.02 - Copyright © 2012

12/04/13, 5:15 am EST
14 yr old 66 lb girl

11/08/13 @ 5:15 am
N.A. 14 yr old girl

10
6
2
200
100
APPROX MEAS. O

Parasitic infections are very common, more than you thi and they may or may not ca us serious health complications. We could su from various types, dependi on the size of the parasite or whether they are inside or outside the body. There are *micro*parasites such as mala that are only visible under t microscope or large *macro*parasites, such as roundworms or flat intestinal worms (tapeworms) that can reach great sizes. Parasites are not only found in the intestine, as is generally thoug but anywhere in the body: the lungs, liver, muscle, stomach, brain, blood, skin and even in the eyes.

The great migratory movements of population, rapid transportation, and trading, have shortened the distances and diseases that used to be common localized diseases, now turned into universal common ailments and vice versa. Parasites previously confined to very specific areas, now appear in other locations far away from their initi focus. This situation is favored by the socioeconomic conditions of the masses of people on the planet. A high percentage of the world's population suffers from infections by parasites, which are responsible for the death of **15 million children** annually.

To this situation we can add the fact that the parasites have developed multiple mechanisms of evasion and resistance to specific immunity, which allows them to circumvent and cancel the host immune response, resulting in chronic and persitent infections. The persistence of parasites in human hosts leads to chronic immune reactions that can damage tissue and cause alterations in immune regulation.

Ninety percent (90%) of the world population is infected with one or more parasites in their body, being able to coexist in the same host with up to five different typ The danger comes when the balance is upset, within the host, skyrocketing the number of parasites and the host starting to show signs of serious illness and even death some cases. Parasitic infections caused by parasitic worms do not always cause disease in man, a number of carriers have been found who are fully healthy. Parasites a silent murderers claiming unsuspecting victims living in a people world that does not even imagine they exist.

Doctors in Western Europe and the U.S. do not contemplate the possibility that we may be infected with parasites, but with modern circumstances like the immigration of people across continents, the problem has become much more common than anyone can imagine and often parasites are causing many of the rare disea or other conditions such as chronic fatigue, fibromyalgia or arthritis that are causing so much damage lately in our communities.

The most common Vermiosis are **intestinal**. People who have them lose large quantity of nutrients, which are absorbed by the parasite, while erosions that originate in digestive tract can be the gateway to various infections. Intestinal worm infections are very common and affect many, not only people with dirty and poor hygiene hab Worms are parasites that are transmitted by **ingesting eggs** produced by the parasites themselves. These eggs hatch in the intestinal tract.

A parasitic infection can be acquired through one or more of these scenarios:

- From someone else, more or less direct contact (by faeces or sexual).
- From self-infection, for example, in the anal-hand-mouth way because the scratching of the anal area can drag the eggs.
- From congenital transmission (mother to fetus).
- From common contaminated objects .
- From from soil contaminated by human or *animal excrement* [E.g. cat litter box].
- From eating contaminated raw or undercooked meat.
- From eating raw fish.
- The traditional foods of some countries include it in their recipes. We can avoid consumption of the larvae of worms by **freezing** the meat or the fish for **twelve hours**. [Seen note 1 at end of document].
- From drinking contaminated water
- From contaminated vegetables or fruits:

    Many times we eat **poorly washed** vegetables or fruits. There is a misconception when we think that vegetables from organic farming are free from any problems, pesticides or chemicals. The danger is that the eggs or larvae of the worms reach the ground through animal waste and decomposed forms of natural compost and manure in the field. There are eggs, such as Ascaris, which can survive in soil under extreme temperatures for as long as five years. It is very important to perform a thorough cleaning of fruits and vegetables; and never eat anything raw, straight from the ground, however healthy it may seem.

- From parasitized animals:

    Infections are very easy to spread **by contact with pets**. Veterinarians can't insist enough about the quarterly worming of our animals, but there are steps we must take on our own to avoid contamination.

    De-worming your pet **at least every three months for life**, as directed by your veterinarian during the first month, should be done every week.

    Prevent pets from eating raw viscera. If animals eat raw meat or raw bones the best option is to freeze the food in advance for twelve hours. After deworming the animal, **their feces must be destroyed during the 8 days of treatment**. It can be burned and buried. If the treatment is working, the animal would be, expelling the worms throughout the faeces or vomit. **Avoid being licked in the mouth by the animal** as these are in direct contact with feces, soil and its own anus. When petting an animal, wash your hands with soap and water before eating or handling food, as the eggs of the parasites remain in the animal's hair.

- Do not walk barefoot or with open toe shoes in soil, wet soil or sand.

Symptoms of parasitic infections

The different types of worms and toxic waste produced by parasites in our organism may cause the following widespread problems:

- Chronic diarrhea or diarrhea caused by poor absorption of food
- Chronic Constipation
- Gas and bloating (Important after meals)
- Digestive problems
- Excessive bowel movements (Intestinal movement after eating)
- Abdominal pain.
- Mucus in stool
- Leaky gut
- Frequent vomiting and nausea
- Hemorrhoids
- Burning in the stomach
- Blood in stool
- Intestinal irritation
- Intestinal obstruction
- Swollen eyes
- Fever
- Pancreatitis
- Malabsorption syndrome
- FATIGUE - The toxic waste metabolic overload overworks the organs of elimination and causes central nervous system disorders like:
    1. Chronic Fatigue Syndrome
    2. Low energy
    3. Lethargy
    4. Extreme weakness
    5. Internal cold
    6. Cold in the extremities

SKIN DISORDERS & ALLERGIES:

Parasites that penetrate the skin cause **itching**. When the tissue is inflammed by these parasites, white blood cells increase the body's defense. This reaction causes skin rashes and food allergies are not far away. Parasites also create toxins and the skin, being the largest organ, tries to eliminate them therefore many skin problems are fel Symptoms include:

- Dry Skin
- Dry Hair
- Brittle hair
- Hair loss
- Allergies
- Itchy nose
- Anal itching
- Itchy skin
- Urticaria
- Jaundice
- Allergic reactions to foods
- Crawling sensation under the skin
- Eruptions
- Eczema
- Skin ulcers
- Swelling
- Sores

MOOD PROBLEMS AND ANXIETY:

Toxins that are released by the parasites can irritate the central nervous system. Anxiety and nervousness often are caused by parasites that run throughout the body. So of the problems caused are:

- Mood swings
- Anger and irritability
- Nervousness
- Depression
- Forgetfulness
- Confused thinking
- Restlessness
- Anxiety
- Slow reflexes
- Obsession
- Lack of coordination
- Disorientation

SLEEP DISORDERS:

The physical presence of the parasites are a nuisance to the body. The body reacts to them during rest periods, since during the night is when the parasites are most acti Nocturnal awakenings are common, especially between 2 and 3 am, when the liver tries to rid the body from toxins produced by parasites. This in turn produces:

~ Insomnia
~ Grinding of teeth during sleep
~ Bedwetting
~ Drooling while sleeping
~ Sleep disorders - multiple awakenings during the night

GROWTH PROBLEMS, WEIGHT AND APPETITE:

The parasites usually live without being detected by the hosts, rob the body of all essential nutrients of the food eaten. Many overweight people who are infected with parasites, go hungry for lack of essential nutrients and this forces them to eat in excess due to parasites. Furthermore, depending on the type of infestation, many peopl are malnourished and can not gain weight.

Some symptoms are:

- In children, poor growth, physical, age and intelectual development Obsession and compulsion to eat sweets or very specific foods
- Weight gain
- Weight loss
- Long-term obesity
- Loss of appetite or uncontrollable hunger to eat more than usual
- Inability to gain or lose weight
- Muscle and joint pain

Parasites can travel to almost all soft tissues, particularly at the joints and the muscles that cause cysts and inflammation and often get mistaken for arthritis and muscle pain. Toxins from parasites can also accumulate in the joints and muscle tissue causing:

- Muscle spasms
- Muscle pain
- Joint pain
- Muscle cramps
- Numbness of the hands or feet
- Chest Pains
- Pain in the navel
- Pain in the back, thighs or shoulders
- Arthritic pains
- Rapid heartbeat
- Fibromyalgia
- Seizures

BLOOD DISORDERS:

The parasites absorb the good vitamins from the body like iron and sugars, resulting in:

- Hypoglycemia
- Anemia

SEXUAL AND REPRODUCTIVE PROBLEMS:

A weakened immune system because of parasites and their waste can also experience:

- Male impotence
- Erectile dysfunction
- Premenstrual syndrome
- Candida - Yeast Infections
- Urinary Tract Infections
- Cysts and fibroid
- Menstrual problems
- Prostate problems
- Fluid retention

RESPIRATORY DISEASE:

The passage of larvae through the respiratory system or larval invasion on the lungs may cause symptoms such as:

- Drowsiness [sleep apnea]
- Shortness of breath or respiratory failure
- Dyspnea [shortness of breath with slight exertion]
- Acute bronchitis
- Asthma
- Pneumonia
- Irritative cough

OTHER PROBLEMS:

- Foreign body sensation in the throat or discomfort
- Difficulty swallowing
- Cough or coughing up blood
- Circulatory problems. numbness in the extremities, difficulty in moving.
- Excess of saliva
- Blurred vision
- Bad breath
- Low immune response
- Fever
- Breathing problems
- Peritonitis
- Chronic symptoms viral or bacterial
- Body odor

PARASITES IN CHILDREN:

We can speak of asymptomatic and symptomatic stages. The former are usually found in adults. Symptomatic stages occur mainly in children, in which we can usually observe: Weight loss, anorexia, growth retardation, headaches, cramping, diarrhea that alternates with periods of constipation, nervousness and irritability, nasal itching and / or anal urticaria. The **Tapeworm**, because of its size, consumes enormous amounts of food that it obtains from the child's consumed nutrients. This can affect thei normal development. Treatment is simple, but it requires that the head of the tapeworm is removed, otherwise it will continue its growth.

MEASURES IN THE HOME ENVIRONMENT TO PREVENT REINFECTION:

It is important to treat all people and pets who live in the same environment to prevent that person from infecting others. Reinfection occurs through underwear, beddin towels and household items such as children's toys or animals that have been in contact with eggs. It is important to wash parts at a temperature not below 60 ° C [140 ° F], including all clothing that has maintained contact with intimate body areas. You have to wash all bed linen and underwear daily (or to the extent possible) while performing antiparasitic therapy. Affected individuals should not share their swimwear with other members of the family and would have to use one cloth to wash his anal area and another one for the rest of the body.

- It is necessary to sleep with underwear and pants if possible to avoid involuntary scratching of the anal area at night, so the eggs will not get under the fingernails Keep pets away from the place of rest of their owners, such as beds, sofas, blankets and cushions.
- Thoroughly wash fruits and vegetables in water and soaking them in MMS or CDS solution for a few minutes.
- Clean the sink with alcohol as the eggs of many parasites are immune to the pH of the cleaning products such as soap or bleach.
- It is important to note that the parasites do not leave any kind of immunity behind, therefore, once rid of them, the person who has suffered can become reinfecte again by exposure to the eggs.

EVOLUTIONARY CICLES OF INTESTINAL PARASITES:

There are many more, but these are the three most common types of intestinal parasites that can be found in developed countries.

ASCARIS LUMBRICOIDES:

In the case of the Ascaris, these worms reproduce easily, and a single female can produce up to **200,000 eggs each day**. This parasite is very common, especially in dar conditions and when hygiene measures are not adequate. It can affect the entire opulation, but mostly children, seriously disrupting their development and growth.

[Note: Dr Hulda Clark says on page 43 of her book, The Cure for All Diseases (on page 67 of pdf), that "...***everyone with asthma tested positive for Ascaris in their lungs***" Another Hulda Clark quote from page 43: "***What if you always found every mysteriously ill person had some unsuspected parasite or pollutant?***"]

Ascaris life cycle

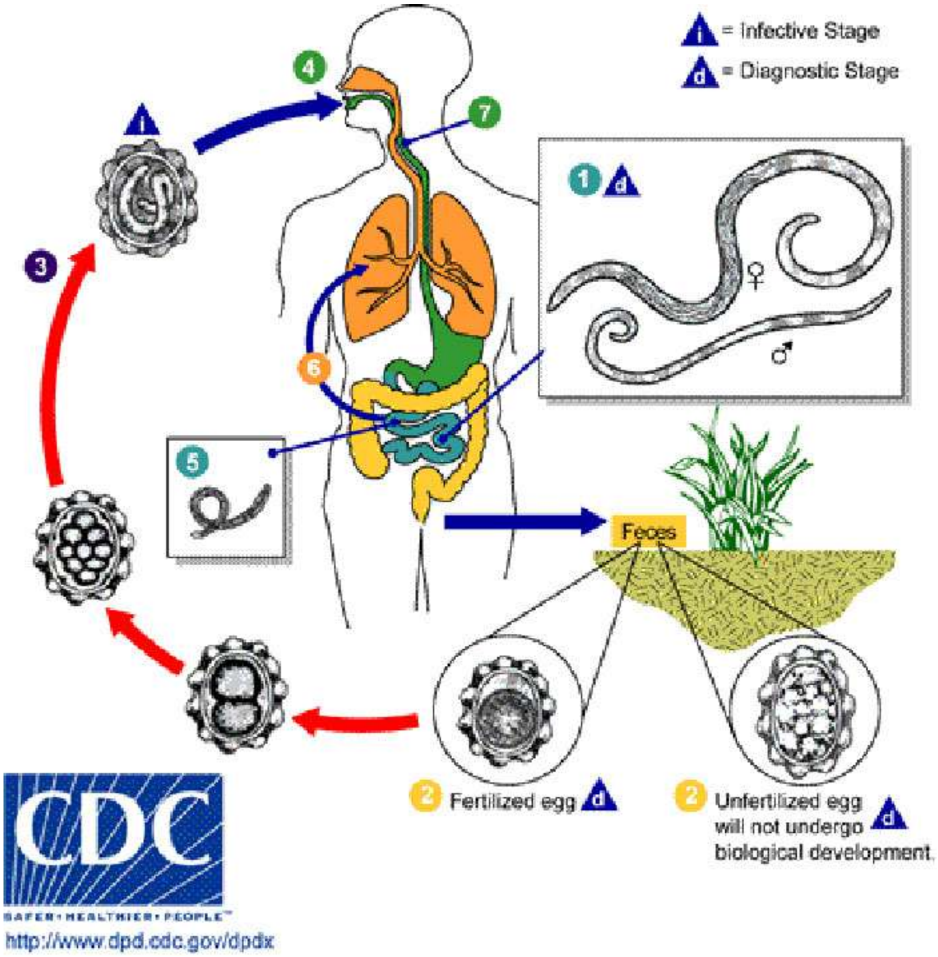
i = Infective Stage
d = Diagnostic Stage
4
7
1 d
i
3
6
5
Feces
2 Fertilized egg d
2 Unfertilized egg will not undergo biological development. d
CDC
SAFER • HEALTHIER • PEOPLE™
http://www.dpd.cdc.gov/dpdx

Ascaris is so infectious, the World Health Organization estimates there are about 2,000 million [**two Billion**] people infected worldwide, of which about 20% of ascaari cases end in death. The eggs of the parasites reach the duodenum through the mouth of the host. Gastric juices digest the egg envelope releasing the larvae. These larva (which have high mobility) penetrate the duodenal mucosa, going for the liver. They then continue their migration to the heart, reaching the **lungs** through the pulmona circulation, reaching the **pulmonary capillaries**, where they are trapped. There, the larvae break the capillary endothelium penetrating the **alveoli**, **bronchioles** and **bronchi** up through the pharynx. Once the epiglottis is flanked [See note 2], larvae are swallowed, returning again to the duodenum, where they complete their maturation process. This process takes approximately **two to three months to complete**, therefore we should calculate that to find ourselves completely parasite free **t** **initial treatment should be done for one year**. Then you only need to follow a routine de-worming of two to four times a year.

Ascaris intestines

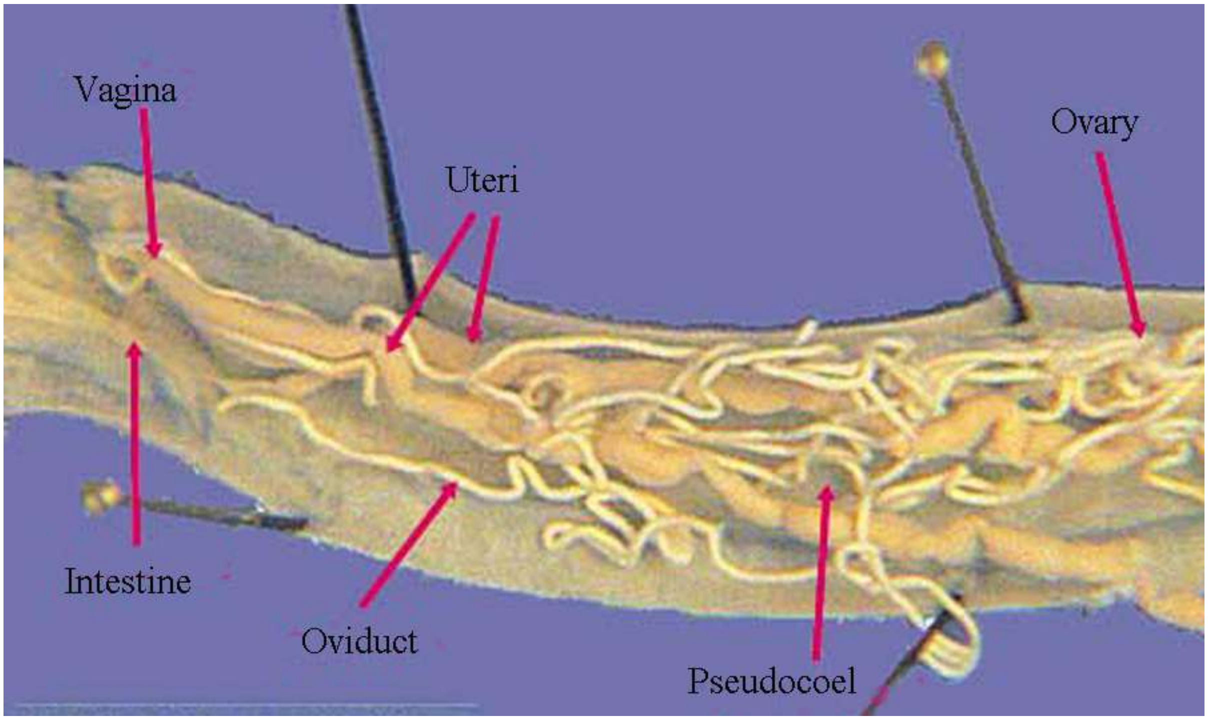

Keep in mind that eggs are expelled with the feces (human or animal) to the environment where they can survive even in harsh conditions, favoring the persistence of t parasite. They are resistant to low and high temperatures, desiccation, strong acids, soaps, chlorine and formaldehyde (pH between 2 and 11) can live in planted soils between 7 and 12 years, creating a "hotbed" and making them almost indestructible.They get **transported through dust flying** within the air currents and are **inhaled and / or swallowed**. From these environments we have recovered eggs from nasal mucus, paper money, potting soil, dust in rooms, etc.

Taenia Saginata and Taenia Solium [tapeworms]:

The taenias reach humans when humans eat their eggs through the tissue of cattle infected and pigs. The larvae encysted in the muscle tissue of the animal is swallowe and their development takes place in the human digestive system. Humans are hosts for T. saginata and T. Solium. The tapeworm is also known as *Solitary* usually no more than four specimen are found in each individual. The danger of this parasite is that the larvae can migrate to the brain or other vital organs (cysticercosis). It can b detected because as the parasite grows it is **discarding segments that can be seen in the stool**. They can usually live in their host for many years in a totally asymptomatic way detected only by the white flat strips in the stool. According to the classification they have a variable size between **2 and 12 meters in length**. It consists of a head called **scolex**, which attaches to the intestine by means of suction cups, and a body consisting of repeating units called **proglottids** it can reach up to 1,000 or 2,000 proglotids, depending on type. A tapeworm can produce an average of **720,000 eggs per day**.

Pinworms (Enterobius vermicularis):

Humans are considered the only host of what are commonly called **Pinworms** (Enterobius vermicularis). This type of worm is the most typical in the family because its easy propagation. It is common for children to re-infect over and over in schools or in their games by scratching the anus and introducing the dirty hands in their mouths. Pinworms have an elongated shape, are whitish and about 1 cm long. They inhabit the large intestine of humans. Females place eggs on the outside, around the anal opening. Once deposited, the eggs are infective for a period of up to **20 days**. Once in the intestine, it takes between **5 and 8 weeks** to develop into adult worms. The most important symptom is intense itching that occurs in the anal area, especially at night. In women inflammation of the vulvar area is very common . Unlike oth parasites, they infect only humans. Eggs are laid between the folds of the anus. Reinfection occurs when infected eggs reach the mouth via our own hands that have scratched the anus. Transmission from person-to-person happens when handling contaminated clothing, bed sheets, towels and environmental surfaces contaminated w pinworm eggs (such as curtains, carpeting) as they are extremely volatile. A small number of eggs can be integrated into the air and then inhaled and follow the same developing ingested eggs. Enemas are extremely useful in removing this parasite of the large intestine.

Graham's method is a simple method of detection. Just after waking and before stool, using **tape** to hold the remains of eggs or parasites which are situated between t anal folds. At first glance we can see small worms no more than an inch, but if we have a microscope, we can see the many transparent eggs from females and even oth species.

The importance of lifelong deworming:

Once we begin the process of deworming we should be aware that we must maintain this habit of cleaning for the rest of our lives to enjoy a good and balanced health. is common among people who have pets, to keep the processes recommended by veterinarians, deworming their pets **every three months**. It is interesting to ask why family physicians do not advise or alert individuals about the existence of this problem among humans. Maybe we have to consider the possibility that they ignore it or are unaware of the importance to eliminate these harmful pests, as harmful for animals as for people. It is true that many parasites are not endemic or common outside certain climates, but human migration and marketing of food products traveling around the world have propagated the spread of many parasitic pests silently. Each parasite has its own life cycle and involves the development of a parasite during different stages of their life since the beginning of his life until he reaches maturity, reproduces and dies.

In the case of intestinal parasites, which are treated here, they can live in the host for **up to ten years**, as in the case of a single Taenia, or the **entire lifetime** of the host reproducing again and again, as in the case of the Pinworms or the well known Ascaris.

THREE MONTHS DEWORMING PROTOCOL:

In the kind of fast society we live today, we have lost touch with the ancient wisdoms. One of the things we have forgotten is to follow the natural cycle of the moon fo many of our routines, so it is important to **begin treatment within three days of the full moon** day and continue until the waning moon. The effectiveness of this time period is greater in this cycle because much of the nematodes travel back into the intestine to mate during the full moon period. During treatment, especially at first, it i completely necessary to use **MMS enemas daily** or fairly continuous; also purging with **castor oil**, a mineral purgative such as **Epson salts** or a preparation of the plar and the leaves of **Senna** available online [also called Senn]. This treatment is specially designed for the deworming of large parasites, especially round nematodes such Ascaris. It is effective for most nematodes but may not be effective against tapeworms. In the case of infestation by Tenias the treatment recommended is **Niclosamide** [available online] , being the medication preferred for its low toxicity.

Day 1:

Pyrantel-Pamoate (single dose in the morning) 10 mg / kg, administered as a single dose together with any liquid.

- If purchased in liquid form, a 5ml teaspoon contains 250 mg (for 60 kg three 5ml teaspoons). In tablets take 3 pills for 60 Kg
- **Diatomaceous earth** [available online 10 lbs food grade Best Buy, 2 lbs food grade, 1 lb. food grade] (two doses) A dessert teaspoon (5ml) twice daily with mea preferably with liquid. Morning and afternoon.

Day 2:

- **Mebendazole** [sold online as Vermox] (two doses) **100 mg every twelve hours**. A pill in the morning and one at night. [See Note 3]
- Diatomaceous earth (two doses) A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon.
- Enema.

Day 3:

- **Castor oil** [cold pressed organic castor oil is sold online], **two tablespoons** (unflavored from the store) while fasting (See note 4).
- Mebendazole (two doses) 100 mg every twelve hours. A pill in the morning and one at night.
- Diatomaceous earth (two doses) A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon.
- Enema.

Day 4:

- Mebendazole (two doses) 100 mg every twelve hours. A pill in the morning and one at night.
- Diatomaceous earth (two doses) A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon.
- Enema.

Day 5:

- Pyrantel pamoate (single dose) 10 mg / kg, administered as a single dose together with any liquid. In theevent that is purchased in liquid form, a 5ml teaspoon contains 250 mg (for 60 kg three 5ml teaspoons).
- In tablets take 3 pills for 60 Kg
- Diatomaceous earth (two doses) A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon.
- Enema.

Day 6:

- Castor oil, two tablespoons (unflavored from the store) on an empty stomach.
- Mebendazole 100 mg every twelve hours. If in tablet, one in the morning and one at night.
- Diatomaceous earth (two doses) A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon.
- Castor oil, two tablespoons (unflavored from the store) on an empty stomach.
- Enema.

Day 7:

- Mebendazole 100 mg every twelve hours. A pill in the morning and one at night.
- Diatomaceous earth (two doses) A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon and evening.
- Enema.

Day 8:

- Mebendazole 100 mg every twelve hours. A pill in the morning and one at night.
- Diatomaceous earth (two doses) A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon.
- Enema.

Day 9 to 18 (first month):

- Castor oil, two tablespoons (unflavored pharmacy) while fasting. Should be repeated depending on the needs of each person.
- Diatomaceous earth (two doses) A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon.
- **Neem** Infusion (Azadirachta Indica) [available online as organic bulk leaves and as organic neem leaf powder] (9 days).

- Three teaspoons of Neem tea in one liter of water. Boil for 5 minutes and taken throughout the day.
- Enemas as often as possible.

Day 9 to 18 (second month):

- Castor oil, two tablespoons (unflavored pharmacy) fasting. Must be repeated again depending on the needs of each person.
- Diatomaceous earth- A dessert teaspoon twice daily with meals, preferably with liquid. morning and evening.
- Infusion of **Epazote** (Chenopodium ambrosioides) (3 days) [available online as 1 lb. bulk, 0.1 oz package, three 0.5 oz packages, 1.0 oz bottle, 1.4 oz jar, 5 lb. bulk. and seeds]
- Boil 1 or 2 tablespoons Epazote leaves per 1 liter of water for 10 minutes, let stand and strain.
- Drink 1 cup fasting for 3 days straight
- The remaining days take the **Aloe Vera gel** with juice or water while fasting
- Enemas as continuously as possible.

Day 9 to 18 (third month):

- Castor oil, two tablespoons (unflavored pharmacy) fasting. Should be repeated depending on the needs of each person.
- Diatomaceous earth-
- A dessert teaspoon twice daily with meals, preferably with liquid. Morning and afternoon.
- Infusion of Neem. for 9 days or take alternative antiparasitic infusion.
- Enemas as often as practical.

Materials required

Enema of 2 liters:

According to the British Royal Academy of Medicine, 90% of disease and discomfort is directly or indirectly related with the dirty colon. We consider the colon as the sewer of the body and toxins found there infiltrate the blood and seriously impair health. Usually in the colon of adults over 40 years old between 2 and 12 kilos of sediment can be found . In this thick unpurged garbage, large parasites move, poisoning the body of its host slowly but surely. Of the food that reaches the digestive tra parasites consume most of its good nutrients, often leaving only "bites" to the host. So while we try to eat healthy by taking vitamins, supplements and other products, do not get any better. Indeed, sometimes quite the opposite: we feed the parasites.

One of the best methods for elimination of this bulk is called "colon hydrotherapy," usually performed by a technician. Even with just using water, a 2-liter enema treatment can help cure many diseases. Used by many cultures since ancient times, it's an easy and natural way of cleaning the colon gradually. The introduced liquid helps eliminate toxins accumulated in the intestine by the death of the parasites. In the case of a mass death of worms, fever and fatigue may increase, and one may fee quite ill. To prevent toxins from reaching the blood stream, we must empty the lower digestive tract as soon as possible.

The enema is one of the essential parts of this treatment. The parasites produce a **biofilm** or **mucus** in the human intestine where they can **hide** to avoid being attacked any medicine, hence the importance of enemas with MMS. MMS is able to undo this layer, also destroying the eggs and larvae they host. It should be **performed daily from the second day** and can be **spaced every 3 or 4 days** progressively. The most important thing is to listen to your body and it will set the guidelines for your need The proper amount to use is between **12 and 20 drops of activated MMS** or its corresponding **10-15ml of** CDS water added to **2 liters of water**. You can gradually increase the dose according to body's acceptance. The most important thing is to **retain the liquid** in as long as possible (**10-12 minutes if possible**).

Detection of parasites in stool:

It is necessary to detect the worms by observing directly the stool. For that we have a plastic basin not very large in diameter and a stick for examination whenever we defecate.

Microscope:

A microscope can be very useful for diagnosis because it can see both the small parasites that may appear in the blood, as well as eggs or larvae in the feces. This way know if the number of parasites decreases.

Bulb - enema

You can use a bulb - enema with a dilution of 15 ml of mebendazole or pyrantel pamoate if you have anal itching. The best way is to introduce the medication together with about 30 ml of water and hold overnight. So the itching caused by pinworms goes away.

Medications for Deworming:

Mebendazole (sold online as Vermox source 1, Vermox source 2) is a drug used in treating diseases caused by helminths (parasites of the gastrointestinal tract). This drug prevents the parasite from using the glucose, causing a decrease in energy and the death of the parasite.

http://webstore.blue/buy-vermox-usa.html

http://allofantibiotics.org/mebendazole/

Mebendazole is absorbed to a limited extent in the gastrointestinal tract (approximately 5 to 10%). Absorption is enhanced when administered in conjunction **with fatt foods** [milk, cheese, butter, fats]. It is metabolized in the liver. Approximately 2% of the administered mebendazole is excreted in urine. The remainder is excreted in feces. The appropriate dose of mebendazole may be different for each patient. It depends on the type of parasite that causes the infection. The most frequently recommended dose is **100mg, 2 times a day for 3 days**.

The adverse effects of mebendazole are generally rare due to its poor absorption. May cause nausea, vomiting, abdominal pain and diarrhea. Normally these effects are generated by the release of toxins from the very death of the parasite itself. The drugs can be administered very effectively diluted with water in a **small bulb-enema** anally. **10 ml in warm water** is introduced into the intestine and is allowed to **act overnight**. It is especially suitable in the case of **oxiurasis**.

Pyrantel pamoate (Trilombrin):

Pyrantel pamoate can be purchased online. It is a broad spectrum **anthelmintic** which works by causing a neuromuscular block producing spastic paralysis of the paras and its subsequent expulsion by intestinal peristalsis action without producing excitation of the parasites, or encourage their migration.

Pyrantel pamoate has a short duration of action and tends to a complete elimination from the body in the feces and urine within **3 to 4 days**. Pyrantel pamoate is poorly absorbed from the gastrointestinal tract, and approximately 6 to 8% total is found in the urine and the remainder in the feces.

The recommended adult dose is **one daily dose**. Children of 12 years and older, 40 to 75 kg: 3 tablets. Adults over 75 kg: 4 tablets. **Its incompatible with the use of piperazine**, a substance naturally found in **pumpkin seeds** or the antiparasitic containing this substance in their formulation since they cancel or neutralize each other .

Plants and minerals for deworming:

The parasite problem has been and still remains a major concern in many cultures around the world. From immemorial time there have been native plants with which people that have done this type of cleaning treatments. In the West and other developed countries parasites are not being treated, making us more vulnerable. There are minerals and substances very useful in combating parasitic infections and we can make various treatments. The bentonite, vegetable charcoal, common clay or diatomaceous earth are the most important examples. In this case we will use Diatomaceous earth as an important intestinal parasitic destructive agent.

Diatomaceous Earth:

The treatment with **Diatomaceous Earth** [available online 10 lbs food grade Best Buy, 2 lbs food grade, 1 lb. food grade] should be continued for **18 days. Diatoms** a one-celled plants living in the oceans for about 300 million trillion years. They formed a little shell composed of the same water extracted silicates. When diatoms die, this microscopic coating was deposited at the bottom of the oceans. Through the ages, they piled up in banks forming deposits of thousands of meters. Over time the ocean receded and these deposits were eventually uncovered. Compressed and fossilized, the shells give rise to a chalky rock called diatomaceous earth.

Diatomaceous earth is an inert, nontoxic compound, which contains a number of minerals such as manganese, magnesium, iron, titanium, and calcium silicates among others. Properly ground, the skeletons of microscopic diatoms become sharp silica needles, harmful to parasites, fungi, yeast, worms, and amoebas. These needles are harmless to humans and other warm-blooded animals. Although it's safe to take it continuously, it's best (as with everything else) to also rest for periods of time. Take one teaspoon twice a day during the 18 days of treatment.

Castor Oil:

Castor oil [cold pressed organic castor oil is sold online] extracted from the seed of a plant like the fig tree, called *Ricinus communis* ("Higuera del diablo"). Its seeds contain between 50-80% oil, which itself has a high content of ricinoleic acid, which has excellent **laxative and purgative** properties. Once treatment begins with plan and medicines, spastic paralysis may occur in some parasites and if there are many together, they can form a "knot" of worms and cause intestinal obstruction. Thus the importance of using castor oil. You need to take it on a **empty stomach in the morning** with juice, tea or milk. The adult dosage is 15 to 30 ml (two tablespoons) fasti one hour after breakfast and medication. In the event of not having good tolerance, mineral **purgatives** such as Epson salts or **Senna leaves** available online [also calle Senn] can also be used.

Neem (Azadirachta indica):

The **Neem** tree [Neem available online as organic bulk leaves and as organic neem leaf powder] is one of the great natural patrymony of mankind and has been used as medicinal remedy for healing since ancient times. There are references in Sanskrit scriptures and Ayurveda has used it since antiquity. The healing and medicinal properties of Neem have been exploited since ancient times in Hindu medicine. Even today Hindus living in rural areas call the Neem tree "The People's Pharmacy" fo its ability to alleviate many diseases and is currently endorsed by the authorities in India for its use in medicinal preparations.
Neem is one of the purifying and detoxifying plants with the most potential. Neem has been used to combat all forms of body parasites, external parasites and internal parasites alike. Boil **four neem leaves** in **one liter** of water for 5 minutes. Take an infusion throughout the day. We will use it for **9 or 10 days** during the first month.

Epazote (Chenopoidum ambrosioides):

Epazote (Chenopodiun ambrosioides) is a plant that grows wild in South American countries [available online as 1 lb. bulk, 0.1 oz package, three 0.5 oz packages, 1.0 bottle, 1.4 oz jar, 5 lb. bulk. and seeds]. This plant has healing properties used to relieve upset stomach, gas colic, and for the expulsion of parasites and intestinal wor **Ascaridol** is the substance in epazote acting and producing a paralyzing, narcotic effect on intestinal parasites, causing it to detach from the intestine to which they we strongly attached. Do not exceed the dosage because it may cause toxicity. Make a decoction drink by boiling 1 or 2 tablespoons of leaves [available online in bulk, in **one liter** of water for 10 minutes. Let it stand and drink 1 cup while fasting for 3 days.

OTHER MEDICINAL PLANTS:

There are other plants that we can use effective for deworming. If after three months of treatment the problem persists, we can change the type of plant or repeat any pl that was effective in previous months. We can use them in preparations, mixing several plants at once or individually. Plants are considered as alcoholic extract, in oil o infusion, including the following plants:

- Artemisia annua,
- Walnut shell,
- Calamus Root Rue,
- Artemisa Absinthium (Ajenjo),
- Southernwood Gentian (Gentiana lutea),
- Mint (Mentha sativa),
- White Fraxinela (Dictamnus albus),
- Tansy (Tanacetum vulgaris),
- Yarrow (Achilea millefolium L.),
- Dandelion Clove (Syzygium aromaticum),
- Pomegranate root bark.

PREVENTIVE FOOD AND DIET:

There are groups of foods we **should avoid eating** if we have a parasite problem such as **dairy products** in general, **refined sugars** (sucrose, fructose corn syrup), **flo** (especially refined) and **overly sweet foods** in general. On the other hand we also have a good list of foods and plants that promote good internal balance of the body hence becoming our allies. With a good production of stomach acid, a good level of healthy bacteria and proper bile production, it is impossible for parasites to survive for long. **Worms need an acidic environment** by the breakdown of sugars and putrefaction caused by the ingestion of unhealthy or processed foods. It is very importa to **eat raw vegetables and fruit juices** that provide us with enzymes and other elements necessary to protect us.

The choucroute / sauerkraut (fermented cabbage in salt):

Many people have **low levels of stomach acid** [Type A blood especially], which is the cause of many intestinal problems, because the body is unable to defend itself against intruders. **Sauerkraut juice or cabbage / sauerkraut** is one of the most powerful stimulants for your body to produce acid. The use of unpasteurized fermente water kefir, soy sauce, or miso is highly recommended for its stimulation of beneficial bacterial flora responsible for generating the control over parasites. Take a few spoonfuls of cabbage juice before meals, or better yet, sauerkraut juice of fermented sauerkraut will do wonders to improve digestion.

Garlic:

Garlic eaten regularly turns the stomach and intestine into a lethal environment for parasites, providing constant protection. Garlic is the quintessential home remedy to eliminate intestinal parasites naturally. It has been used by different cultures such as Chinese, Greek, Roman, Indian and Babylonian. It has not become obsolete as it is still used by practitioners of modern medicine. It is used both fresh and its oil, the simplest treatment is to eat three cloves of garlic every morning or take a teaspoon o garlic oil. Can also be used by mixing crushed garlic in a little cold water and drink the mixture immediately. Another recipe is to cut in pieces and crush four cloves o garlic in milk overnight and take the liquid fasting the next day.

Pumpkin seed:

Pumpkin seeds contain a substance called **piperazine**. It acts by paralyzing the parasites, which allows you to remove the parasite easily . We can find piperazine commercially in pharmacy drug formulations or naturally, as we said, in the seeds of the pumpkin. This traditional method of deworming has been used around the wo since man can remember. There are several effective traditional formulas and this is one of them:

> Use one cup of pumpkin seeds (about 80 seeds), peeled and mashed. Mixed with water from a coconut and two tablespoons of honey. Take within a period of **three hours** on an empty stomach. Don't eat during this three hour period. At the end, take **castor oil** in order to quickly eliminate the parasites.

Papaya and papaya seeds:

Papain is a digestive enzyme contained in papaya and is capable of breaking the outer layer of an adult parasites. The milky juice of **unripe green papaya** is a powerf agent for destroying **roundworms**.

A) The adult dose is **one tablespoon** of **fresh** juice of **green papaya**, the same amount of honey and three or four tablespoons of hot water.

B) **Two hours later**, administer a dose of **castor oil** mixed with warm milk. This treatment should be repeated for **two days** if necessary. For children aged seven to te years of age should be administered half the dose. For children under three years, a spoonful of the mixture is sufficient.

Papaya seeds are also useful for this purpose. They are rich in papain and **caricin**. Use a mixture of **crushed fresh seeds**. For every tablespoon of seeds, one tablespo of honey is added. The dose of a dessert spoonful daily in the morning on an **empty stomach** or before bed for **ten days**, then **rest five days**, and repeat the cycle three times [10 days on, 5 days off - repeated three times]. We recommend the use of a purgative.

[**Additional notes on eating Papaya seeds from Ken Adachi**: I've looked at a number of Youtubes videos on using Papaya seeds and they seem to be very effective, easy to obtain and relatively inexpensive. Many talk about taking 1 tablespoon of dried crushed seeds (use coffee grinder or blender) with 1 tablespoon of honey added warm water and drink, which seems to work OK. Other say to eat the seeds **fresh**, blended with the fruit pulp to make a pudding and eat. Blending the seeds with the p will both crush the fresh seeds and make it easy to consume the seeds, which do not have a pleasant taste (pepper-like and pungent) when eaten alone. An inspiring and likable Florida woman named Tasha Lee (great videos and blogs!) with major parasite problems has a video showing her eating all the seeds of a large papaya straight, without adding honey or mixing with the pulp, but it is hot and not easy to do unless you get used to it. You will obviously kill more parasites by eating all the seeds (about 1/4 -1/3 cup worth) of a large papaya at one time, but I think it's wiser to start off with one tablespoon and work your way up to a larger amount of seeds over time.]

Ginger:

Ginger not only helps to combat intestinal parasites but also reduces nausea and can help calm nerves. Fresh ginger has shown great success in destroying intestinal worms for hundreds of years. The most common way to consume is **raw ginger** or by infusion [brewed with hot water, drunk as a tea]. Ginger extract may also be sprinkled on a variety of foods.

Propolis:

Bee Propolis has been used, for at least 3,000 years. Already known to the Egyptians and the Romans, its use continues into modern times. To the Greeks we owe the name: pro, meaning "before" and polis, meaning "city". This translates as "defenses before the city" or "defender of the city." Thanks to the antibiotic action of propoli which protects against the activity of viruses and bacteria, the bee hive is one of the most sterile places known in nature. Many scientific studies have proven the antiparasitic activity of propolis, so it is recommended for treatment of: **Giardia**, **amoebas** and **roundworms**. It's also used for intestinal infections caused by **gram-positive bacteria**.

Take propolis for the treatment of parasites for **7 days**, fasting, diluted in water or fruit juice. Propolis at 30% in Propolis tincture or capsules. Take 3 drops per kilo of weight, or 3 capsules half an hour before each meal. Follow 7-day tratment cycle with 7 days off, then repeat **3 to 5 times** to ensure complete elimination of parasites o bacteria. It is very important to repeat the treatment to cut the reproductive cycles.

By repeating the treatment at least 3 times, the effective elimination of parasites is ensured.The benefits of propolis are, that it has **no side effects**, is well tolerated, and **highly effective**.

Pomegranate Fruit:

Pomegranate bark contains an alkaloid known as **punicine**, which is highly toxic to earthworms. Used by decoction [extract with boiled water] of the bark of the root, stem or fruit. The root bark is preferable, because it contains a greater quantity of the alkaloid than the bark of the trunk. This alkaloid is highly toxic to **tapeworms**. A cold decoction of bark, preferably fresh, should be given in quantities of 90 to 180 ml three times (for adults), with intervals of one hour between cups. A purgative should be taken after the last glass. The dose for children is 20 to 60 ml. The decoction is preferably used to get rid of Solitaires [another name for tapeworms].

Carrots:

Home treatment with raw carrots to eliminate intestinal parasites in children is another effective home remedy. The chemical constituents of carrot attacks pests by preventing their development. Carrots are one of the most effective natural treatments for children. The child should be provided with a small cup of grated carrots in t morning until the problem is fixed.

Condiments:

Seasoning plants are also a powerful weapon to keep in mind in our everyday cooking. Since time immemorial mankind has used them to control parasitic diseases. M interesting because of their effects are:

- turmeric,
- pepper,
- tarragon,
- thyme,
- cinnamon,
- paprika,
- cayenne and
- cloves.

Annex to the Protocol:

It's difficult to detect a parasite or worm in the toilet bowl because once they are dead, they look like intestinal mucus because they are half digested by our body. But t differ from intestinal mucus which dissolves in water. Try with a stick and bowl, and if it doesn't dissolve (See photos), it's a parasite (Ascaris).

Version and Revisions:
First release – Version 1.0, Revision 1.0 - August 2, 2012.
#
Minor corrections – Version 1.0, Revision 1.01 – August 2, 2012.
#
Added Annex to the protocol – Version 1.1, Revision 1.0 – August 3, 2012.
#
Minor text/notes corrections sent in by Michael Harrah - Version 1.1, Revision 1.01 – August 12, 2012.
#
Added note 4 - Version 1.1, Revision 1.02 – August 12, 2012.

Notes:

[1] From FDA site: Freezing and storing at -4°F (-20°C) or below for 7 days (total time), or freezing at -31°F (-35°C) or below until solid and storing at -31°F (-35°C) below for 15 hours, or freezing at -31°F (-35°C) or below until solid and storing at -4°F (-20°C) or below for 24 hours is sufficient to kill parasites. FDA's Food Code recommends these freezing conditions to retailers who provide fish intended for raw consumption.

Note: these conditions may not be suitable for freezing particularly large fish (e.g. thicker than six inches)

http://www.fda.gov/Food/GuidanceComplianceRegulatoryInformation/GuidanceDocuments/Seafood/FishandFisheriesProductsHazardsandControlsGuide/ucm091704.

[2] http://www.embryo.chronolab.com/tongue.htm

[3] Mebendazole dosage varies from 100 mg 2x/day to 500 mg 3x/day depending on which parasite(s) are being targeted.

See: http://allofantibiotics.org/mebendazole/

[4] Definition of Fasting
Fasting: : going without food or drink. Patients may be advised to fast for a certain period of time prior to surgery, medical procedures, or certain blood tests.
http://www.medterms.com/script/main/art.asp?articlekey=97585

If you want you can make a translation of this article into another language.

Dr. Andreas Kalcker at AutismOne 2014 - Successful Treatment of Autism
https://www.youtube.com/watch?v=rj5XjATNHuw#t=1281

Published on Jun 11, 2014
Successful Treatment of Autism
Presentation by Dr. Andreas Kalcker at AutismOne Conference on May 24, 2014. Dr. Kalcker speaks about how the combination of parasites and vaccines is contributi to the autism epidemic of today. Kalcker explains how he discovered this link and the positive results of treating parasites in autistic children around the world.

HEAL YOURSELF
THE NATURAL WAY

Overcome Diseases
& Create Superior Health

WALTER LAST

Also by Walter Last

Heal Yourself
Healing Foods
The Self Help Cancer Cure Book
The Heal Yourself Series of Self Help books

Copyright © 2010
by Walter Last

Any part of this book may be freely reproduced for
non-commercial use and healing activity.

Editing, Artwork & Cover design by Leonie Ganivet ~ Austpac Productions

Published by

Huckleberry Enterprises

…a division of

Austpac Productions
P. O. Box 842, Moss Vale 2577 NSW, Australia
Phone: +61 2 4869 4285
Email: info@austpac-productions.com
Web: www.austpac-productions.com

If you are unable to order this book from your local bookseller,
you may order online from the publisher on CD Rom in PDF format.

See: www.the-heal-yourself-series.com

Heal Yourself - The Natural Way
Overcome Diseases & Create Superior Health.
By Walter Last.

Includes bibliographical references and index.
1. Naturopathy. 2. Healing. I. Title.

ISBN 978-0-9803281-7-2

HEAL YOURSELF
THE NATURAL WAY

Overcome Diseases & Create Superior Health

WALTER LAST

Published by

Huckleberry Enterprises

Disclaimer

The aim of this book is to provide information on using natural healing methods to improve health and overcome illness. The author or publisher cannot accept any legal responsibility for any problem arising from experimenting with these methods. For any serious disease, or if you are unsure about a particular course of action, seek the help of a competent health professional.

* * * * * * * * * * *

Contents

Introduction

Part 1: Healing the Body

Step 1:
The Road Towards Disease
Become aware of the factors that cause your illnesses………………………………………

Step 2:
The Road Towards Health
Learn the principles of natural living to restore your health…………………………………..

Step 3:
Intestinal Sanitation
Clean out and sanitize your gastrointestinal tract…………………………………………....

Step 4:
Fasting and Cleansing
Become clean on the inside by removing accumulated wastes and toxins………………....

Step 5:
Allergy Testing
Discover and overcome your food allergies and chemical sensitivities……………………..

Step 6:
The Acid-Alkaline Balance
Bring your body back into balance by testing for and correcting any
overacid or over-alkaline condition……………………………………………………………

Step 7:
Living Water
Living water has a great healing potential—learn how to use it…………………………....

Step 8:
The Practice of Healthy Living
Improve your environment—your house, clothes, even your teeth………………………......

Step 9:
Exercise
Get into the habit of exercising regularly……………………………………………………......

Step 10: The Healing Crisis
Sooner or later you are bound to have a healing crisis, so use it to your advantage……....

Part 2: Natural Healing Methods……………………………………………………………

Step 11:
General Examination
Examine yourself to assess your state of health……………………………………………..

Step 12: Iris Diagnosis
Study your eyes to find out more about internal conditions of your body…………………...

Step 13: Muscle Testing
Test your muscle strength to see whether a food or remedy is good or bad for you………

Step 14: The Endocrine Pattern
Simple arm and leg measurements can show the condition of your endocrine glands…...

Step 15: Reflexology
*Press reflex points on your hands and feet to diagnose and
treat conditions within your body…………………………………………………………………*

Step 16: Ear Acupuncture
Try this easy method to improve the function of any diseased or unbalanced organ……...

Step 17: Color Therapy
Use colours to balance body conditions and your emotions…………………………………..

Step 18: Magnet Therapy
Take advantage of the fact that both poles of a magnet have useful biological effects…...

Step 19: The Electronic Zapper and Magnetic Pulser
Use electrotherapy to keep your body free of undesirable microbes and parasites……….

Step 20: Spinal Therapy and Massage
Improve the activity of all your organs and glands by massaging your spine and body…..

Step 21: Hydrotherapy, Packs, and Colonics
Use water to relax or alleviate painful conditions……………………………………………....

Step 22: Urine and Urea Therapy
It may not be your "cup of tea," but many swear by this ancient therapy……………………

Part 3: Nutrients and Remedies………………………………………………………………….

Step 23: Vitamins
*Find out what these essential nutrients do and if you have special
requirements for any of them……………………………………………………………………..*

Step 24: Minerals
Find out what minerals do and if you have increased requirements for them………………

Step 25: Amino Acids
*They're the building blocks of all body proteins.
Adequate levels are essential for good health…………………………………………………*

Step 26: Digestive Enzymes
*These vital substances are needed for digestion, but poor diet and
cooking practices can deplete your supply……………………………………………………...*

Step 27: Nutritional Supplements
*Concentrated nutrients in pill form can shore up nutritional
deficiencies and build optimal health……………………………………………………………*

Step 28: Herbs
You may benefit from using herbal remedies instead of pharmaceutical drugs……………

Step 29: Homeopathy
Homeopathic remedies are effective, and you can even make them yourself……………...

Step 30: Colloidal Silver, Copper, Zinc, and Gold
Use these powerful remedies whenever you have an infection or inflammation…………....

Step 31: Oxygen Therapy
Oxygen, therapeutically applied, helps you quickly overcome any infection…………….....

Part 4: Healing Foods……………………………………………………………………

Step 32: General Dietary Rules
Understand the principles of healthy food selection, preparation, and combination……….

Step 33: Food Groups
Learn the health characteristics of the various food groups and apply them for healing….

Step 34: Health Diets
Adopt a health diet suited to the specific requirements of your body………………………..

Step 35: Special Foods
Take advantage of the health-giving properties of bee pollen, sprouted seeds, fermented foods, and purple foods……………………………………………………………

Step 36: Water and Juices
Heal yourself with fresh vegetable and grass juices…………………………………………..

Step 37: Healing Recipes
Transform healthy foods into enjoyable meals…………………………………………………

Part 5: Problem Foods and Food Problems………………………………………………

Step 38: Cow's Milk Products and Lactose
Have a closer look at cow's milk products and lactose. In some forms they can be beneficial, but as commonly used they tend to create a lot of health problems…...

Step 39: Wheat and Gluten
Wheat may be the staff of life, but in many people it is actually unfavourable to health......

Step 40: Sweet Foods
Your sweet tooth may be your undoing, so consider reducing your intake of sugars……..

Step 41: Meat and Fat
Overconsumption of meat and fats can create serious health problems. Try cutting back on your intake………………………………………………………………...

Step 42: Chemicals in Your Foods
Minimize your exposure to toxic chemicals, water additives, and yeasts and moulds—your body will thank you for it…………………………………………………

Step 43: Metabolic Types, Pathways, and Diets
Know your individual metabolic type and your specific dietary requirements………………

Part 6: Specific Health Problems………………………………………………………...

Step 44: Self-Care Help for Health Problems—Practical Tips for Relief
If you have a specific disease, this summary of typical treatment protocols can show you what to do………………………………………………………………………..

Step 45: Deficiency Symptoms
Check out this list of deficiency symptoms to see if you are lacking in key nutrients—and start replacing them……………………………………………………....

Step 46: Cancer
If you have cancer, here are some useful suggestions……………………………………….

Step 47: Emotional-Mental Shocks
The secret to preventing illness is to undo shocks that unbalance the body……………….

Step 48: Candida Infections
If you have an overgrowth of Candida, here is what to do about it…………………………...

Step 49: Cardiovascular Disease
Managing cholesterol and improving your heart health - some practical tips………………

Part 7: Energies - Dealing with Harmful and Helpful Ones………………………………

Step 50: Electromagnetic Energies
Protect yourself from the harmful effects of electromagnetic radiation
and make sunlight your friend…………………………………………………………………

Step 51: Bioenergy
Bioenergy is the true energy of healing, so for optimal health learn as
much about it as you can………………………………………………………………………

Step 52: Cultivating the Energies
Learn to sense subtle bio-energies and use them to energize your body…………………...

Step 53: Energy Accumulators
Use orgone accumulators, pyramids, cones, or water to concentrate bioenergy…………...

Step 54: Energy Healing
Give energy or remove excess energy to facilitate healing yourself and others…………....

Step 55: Meridian and Acupressure Therapy
Balance the flow of energies along your acupuncture meridians by tracing their circuits....

Step 56: Healthy Sexuality
Channel your sexual energies into pleasure rather than frustration…………………………

Part 8: Healing Your Emotions…………………………………………………………...

Step 57: Understanding Emotions and Disease
Learn how you came to be so out of touch with your feelings, what
emotional problems and diseases this causes, and how this is expressed in your body….

Step 58: Healing Relationships
Try out these tools to heal your personal, family, and social relationships……………….....

Step 59: Heal in Groups
Join a group of like-minded individuals with the goal of healing your
feelings and helping others do the same……………………………………………………..

Step 60: Learning to Feel
Feel what your body and your soul tell you, and use your feelings to
improve your health rather than destroy it……………………………………………………

Step 61: Take the Love Cure
Lead a radiant life filled with love and joy, and practice the Love Cure……………………...

Part 9: Healing Your Mind………………………………………………………………

Step 62: Harness the Power of Mind
Appreciate the power of your subconscious mind. It is the real power
behind the throne of the ego…………………………………………………………………...

Step 63: Helpful Mind Tools
Use relaxation, regression, reprogramming, prayer, meditation,
and inner communication for a more fulfilled life………………………………………………

Part 10: Spirituality and Health………………………………………………………..

Step 64: The Spiritual Dimension
Enter the spiritual dimension of life and make it your home…………………………………..

Step 65: Life on the Path
The spiritual path is the crowning glory of human endeavour, and by following the 65 steps you are already on it…………………………………………………..

A Review of the 65 Steps……………………………………………………………………….

Endnotes…………………………………………………………………………………..

Bibliography……………………………………………………………………………………

About the Author……………………………………………………………………………...

* * * * * * * * * * *

INTRODUCTION

This book is for those who are searching for a way to live a healthy and fulfilled life. I have no doubt that we can do this, live a healthy and fulfilled life, regardless of age, and that this is actually our natural birthright. We just need to avoid or remove the artificial living conditions that prevent us from being as we really are. A modern problem for genuine health and spiritual seekers is the great flood of advice available in these areas. Much of it is contradictory or, alternatively, there are so many different methods to choose from that it seems almost hopeless to find the right one.

I have sifted through much of this information and experimented with many different healing methods in my 30 years as a natural therapist, with the aim of finding out what works for whom and in which conditions. Even more important for me was my spiritual quest. Here I share the fruits of my experience as a lifelong health and spiritual seeker.

While this book does not go into great detail about the treatment of specific diseases, by using the outlined methods and following the provided guidelines to rebuild your health, you are likely to improve anyway, whatever the name of the disease. Rather than fighting your disease, I recommend you focus all your resources on improving your health. Your disease may then disappear on its own.

To show how simple, natural methods can be very effective in overcoming medically incurable diseases, I mention an example from The Food and Health of Western Man, by Dr. J. L. Mount. In five cases of bowel cancer, surgery revealed that the cancer had already metastasized all over the body. Therefore, these patients were just closed up again and sent home to die. But instead of doing that, these five people, independently of each other, became keen gardeners and from then on ate only home grown, organically raised food. When they finally did die, 21 to 30 years later, no traces of cancer could be found in post mortem examinations.

Such cures without medical intervention are regarded as “spontaneous remissions.” It should be comforting for those with advanced cancer and other “incurable” diseases to know that spontaneous remissions can be induced with such simple natural methods. However, while not mentioned in Dr. Mount’s book, I suspect that there was another important factor at work: These five had made peace with themselves and the world.

Similarly, you may use this book to improve your health or to lead a more happy and fulfilled life, more at peace with yourself and the world. Ideally, it is meant to provide guidance to heal not only your body, but also your emotions and mind as part of your spiritual growth. By following this path and the 65 steps I present, you will need less and less to rely on this book or any other “expert” advice. Instead you will realize that all the answers for your healing and spiritual growth are within you. You just need to ask and learn to listen. In order for you, the reader, to gain a better understanding of the purpose of this book and the principles on which it is based, I will share with you my basic philosophy on health and spirituality.

My most basic assumption of existence is that it has purpose. I came to this conclusion gradually through observing, experiencing, and thinking. Countless people share this view, but for most it is just a belief, commonly adopted from parents and society. In order to make it a knowing, our beliefs have to be consciously evaluated and tested.

On the other side of the fence is the orthodoxy of the scientific community and its followers who assume that existence evolved according to inherent physical laws and does not have a purpose. This, too, is just an untested belief. In line with this belief about the nature of our existence, we regard our health problems and other problems in our life as accidental, coincidental, or just due to biological considerations and without a deeper meaning or purpose in our life.

I believe that our health and social problems are part of the purpose of our existence. Therefore, in order to overcome any health problems and be able to lead a healthy, happy, and fulfilled life, we need to have a basic understanding of the purpose of our life.

It has been my healing experience, especially in serious conditions such as cancer, that with apparently the same amount of effort and dedication some people get well while others do not. I believe that success or failure in healing depends only partly on doing the right thing on the biological level and to a larger degree on our emotional, mental, and spiritual constitution. This applies even more to the degree of happiness, fulfilment, and joy we have in our lives.

Therefore, after starting out as a biochemist and nutritionist, I gradually moved towards a holistic perspective, which increasingly included the spiritual dimension. I came to the conclusion that important events in our lives, and especially life or death decisions, are never “accidental.” Not only do they follow the laws of cause and effect, but also the dictates of our higher guidance as to whether this body is still useful for the purpose for which it was created and inhabited.

Here are two examples to illustrate this point. One of my early lung cancer patients recovered well with only biological therapies. Soon afterwards he had surgery for an unrelated problem. The tumour had disappeared but he died from the surgery. Another patient was in a wheelchair with permanent unbearable pain after a spinal accident. We found a way to stop the pain. So he went out drinking with a mate to celebrate and smashed his chin in a fall. This landed him back in the hospital with constant pain.

These and similar cases are not scientific proof of anything, but for someone looking for an answer they are signposts to show the way. Also consider an experience related by the American psychic Edgar Cayce. He routinely saw auras around people. Once he entered a department store elevator but immediately backed out again because he sensed that something was very wrong. Soon after, the elevator crashed and all in it were killed.

Cayce realized what was wrong. None of the occupants had an aura. (An aura is an energy field around a living body that can be seen by some individuals as a nebulous outline.) For those who know that auras exist, this can mean only one thing: It was no accident that these people died together. The guiding consciousness of every single occupant had decided beforehand to dispose of the body in this seemingly accidental way. In an odd sense, they were already dead. I am convinced that if Edgar Cayce had remained in the lift, he would have been the only miraculous survivor and without any serious injury.

If you can follow me to this conclusion, you may ask: How can we find the purpose of our lives and live accordingly? That is what spirituality is all about. We need spirituality, not only to overcome serious diseases, but also to lead a healthy and genuinely happy and fulfilled life. But what is spirituality? Different people have different ideas about it. I can only tell you my version. This may change somewhat over time, as my understanding is still growing.

In addition to an individual, more narrowly defined purpose, we all have the same general and overall purpose to become more complete or perfect as human beings. Normally, this is a very slow evolution for the whole human race. However, if we do this consciously, then our snail-pace evolution becomes a personal revolution and we call it “the spiritual path.”

As I see it, our higher guidance tries to lead us towards or along this spiritual path. There are two possibilities. We may be led gently if we ask and listen for this guidance as with prayer and meditation or just with our general attitude and conduct. The alternative is more common and painful. If we stray too far from the desired direction, we may run into a brick wall that forces us to change course. We may encounter a disease or calamity in our social or professional life that causes us to re-evaluate our life. Many former cancer sufferers have said that cancer was the best thing that ever happened to them because it helped them to end a spiritually meaningless life and start a new, meaningful, and more fulfilling spiritual life.

There are different levels of our existence, as for instance the biological, emotional, and mental levels. Following the spiritual path means that we consciously work on improving ourselves on all of these levels. This is the focus of the 65 steps in this book. On the biological level, this means that we cleanse the body of lifelong deposits of metabolic wastes and -toxins. We also improve our living conditions by minimizing harmful factors and maximizing beneficial factors. On the emotional level we do the same: We cleanse our emotional body of blocked and negative emotions and cultivate uplifting feelings and emotions. It is similar on the mental level: We cleanse ourselves of negative thoughts and belief systems and learn to generate positive and beneficial thoughts and beliefs.

After achieving a certain level of control over our body, emotions, and mind, we are able to move along on the spiritual path. As a reward, our body will most likely be much healthier; we will generally be content, happy, and increasingly joyful; and we may even achieve a state of mental illumination with a flood of intuitive insights. By this time also our individual purpose for being in this body will have become clearer and we may act accordingly.

The overall theme of this book is health and harmony - learning to live in harmony with our biological, social, and spiritual nature. This book will give you the tools to achieve this worthwhile goal. Do not be concerned that it seems far away. The most interesting part of life is the journey, so enjoy it, make it fulfilling and meaningful. It does not matter how fast you master the steps; it is more important that you enjoy doing them and in the process get to know your body and your mind and become friends with both. Learn to love yourself, and you will find it easier to love everyone else as well. If in doubt about any particular course of action, choose what on reflection appears to be more natural or will lead to greater harmony or balance.

In this book you will find a wide range of natural health improvement and healing methods with easy-to-follow, do-it-yourself instructions. I suggest that you experiment with as many of these as you can in this venture. On the biological level, these are intestinal sanitation, allergy testing, and cleansing. Some idealistic individuals may embrace these measures enthusiastically, while others will try a little here and there but generally wait until poor health launches them more determinedly into a program of health improvement. That is quite all right. Just keep this book handy until you are ready. For those who want to start right now, however, here is the "prescription."

The main purpose of this book is to serve as a practical instruction manual for creating and maintaining good health. To get a basic overview and understanding of the underlying principles of natural living and holistic healing, I recommend that you start by reading the book in a selective way. Read more carefully those steps that provide a basic understanding of healthy living, nutrition, emotions, and beliefs, but skim over the details of specific exercises and procedures, or diseases in which you are not currently interested.

Then start working through the steps in a practical way. I regard all of the practical steps in Part 1 as essential for establishing a solid base for improving and maintaining health. Try to do them as thoroughly as possible. The steps in Part 2 are designed to help you identify and overcome specific health problems. Therefore, you may use these steps in a selective way; concentrate on those that you expect to be most suitable with your current problems. It is similar with Part 3. Use this information to select suitable supplements and remedies.

Part 4 is another fundamental chapter. Adopting a natural diet - fresh, raw, and organic - may be the most difficult, but also the most rewarding part of the whole program. Do it gradually. Part 5 provides a deeper understanding of nutrition and helps you to fine-tune your diet for your individual requirements. In Part 6 you need only study the information that is relevant to a current health problem.

The practical steps in Part 7 help you feel and direct your inner or life-force energies. This is a key requirement for improving health beyond what is possible with a healthy lifestyle. Continue practicing in this area for the rest of your life. In Parts 8 and 9, select and experiment with the methods that most appeal to you for healing your emotions and adopting an appropriate belief system. Part 10 helps you understand and reapply your health improvement efforts as part of your spiritual path. Do this review when you feel ready for it.

Finally, use this book as a reference manual whenever you encounter a specific health problem. In Part 6, look for specific diets, remedies, and treatment modalities, and in addition use suitable therapies described in Part 2 and other parts.

Here is a vision I have of our near future. I see health clubs and healing groups springing up all over the world, like mushrooms after a warm rain. Many people come forward, willing to accept self-responsibility for their health and eager to learn. Friends meet regularly, testing each other, experimenting with new healing methods, discussing results, exchanging recipes, sharing books, exercising and meditating together, and studying spiritual philosophy for healing.

Many become healers in their community, guided initially in workshops by experienced healers. City groups rent suitable buildings to establish healing and meditation centres. Communal farms produce organic food and are used as healing and spiritual retreat centres, where visitors can share the healthy life, learn to heal themselves, and find guidance on the spiritual path.

I see the physical and spiritual rejuvenation of nations starting at the grassroots level in a worldwide movement for social and environmental responsibility. All you rejected and disillusioned individuals unite, to make the world a better place for everyone!

* * * * * * * * * * *

Part 1

HEALING THE BODY

Initially, most health seekers are mainly concerned with overcoming a specific health problem, using the specific method or remedy promising the quickest result. However, in the long run and as we get older, more health problems or diseases pop up and become increasingly difficult or impossible to overcome just by using a “magic bullet.” This then makes the idea of serious lifestyle changes towards natural living and holistic healing more attractive, giving us a much better chance to become and remain healthy and fit into old age.

I have no doubt that it is our natural birthright to be healthy and that it is unnatural to have a disease. A disease means that we (or possibly our parents) did not live according to our biological, social, or spiritual nature. At present most of humanity lives in unnatural conditions surrounded by a minefield of technological and chemical health hazards. Even worse is our ignorance about the nature of these hazards that are mainly hidden from our awareness.

Therefore, in order to live more naturally we need to become aware of the main health hazards in our environment and learn to avoid them or minimize their effect on us. We can learn from our own experiences and from those of others and we can also develop our intuition or ask for guidance through prayer and meditation. If in doubt about a choice, try to imagine which course of action is more natural or more in harmony with your biological, social, and spiritual nature.

The main biological influences on our health are nutrition, exercise or how we use our body, and environmental factors. In addition to improving these conditions, there are numerous natural healing methods to help us in our quest for good health. These include herbs and other natural remedies, working on muscle and bone structure, and using electrical, magnetic, and vibrational medicines and approaches.

* * * * * * * * * * * *

Step 1

THE ROAD TOWARDS DISEASE

Become aware of the factors causing your illnesses.

People in Western society are rarely healthy. Real health - the perfect functioning of all parts of body and mind - is so rare that people mistake the temporary absence of disease symptoms for health. Formerly, people suffered mainly from violent but short-lived infectious diseases; now many are afflicted with chronic degenerative diseases for a lifetime.

The main factors causing widespread health degeneration are as follows:

1. Unnatural nutrition: food refining; lack of “living food” and of vitamins, minerals, and enzymes; GM food, junk food, incorrect food combining, unsuitable diet for metabolic type.
2. Chemicalisation: pesticides, herbicides, fertilizers, medical drugs, food additives, water and air pollution, plastics, and even contact with synthetic garments.
3. No exercise: lack of sufficient physical activity or exercise in natural, unpolluted surroundings.
4. Harmful negative feelings and emotions: resentment, fear, discontent, greed, envy, and lack of positive feelings and emotions, such as gratitude, devotion, contentment, joy, unselfish love.
5. Harmful mental attitudes: lack of spiritual or humanistic goals, lack of faith.
6. Emotional damage: sustained during birth, childhood, or adolescence; lack of positive role models.
7. Inherited genetic problems: genetic makeup of poor and deteriorating quality.

The general sequence leading to the development of poor health and chronic diseases commonly is as follows: Emotional injuries received during birth and childhood, as well as negative attitudes and emotions experienced during later life, cause muscle tension which block the free flow of bioenergy within the body and especially within the acupuncture meridians. This weakens the blood circulation and the glandular system, as well as the functions of affected organs. Inherited metabolic weakness and poor nutrition as well as chemical pollution combine to cause an inefficient use of biochemical building blocks and poor cellular energy production. Disease is the result.

Metabolic Residues: As a result of this metabolic weakness and the overuse of unsuitable food, toxics and metabolic residues accumulate in the body. Toxics (toxic waste products from our chemicalised food) accumulate mainly in the fat tissue. Metabolic residues consist of protein debris, mucus, fatty sludge, and organic acids. Here are some further complications:

- **Protein Debris:** This accumulates when we eat more protein than we metabolize. Small protein fragments clog up capillaries and lymph channels through which nutrients move into cells and waste is removed. Current research shows that virtually all chronic diseases are strongly associated with protein debris accumulating within cells.

- **Mucus:** This originates from lactose (milk sugar) in dairy products, gluten (in wheat products), refined carbohydrates, fat, moulds, and food allergens. Mucus has a special affinity to mucous membranes, causing catarrh, ear and sinus problems, respiratory diseases, and malabsorption of food.
- **Fatty Sludge:** Mainly formed from oxidized (rancid) fats and oils, fatty sludge leads to poor circulation and heart problems by clogging the blood vessels and congesting the lymph system.
- **Organic Acids:** General overacidity contributes to the accumulation of organic acids in the tissues and induces inflammations and mineral deficiencies. Overacidity results from insufficient cellular oxidation and from the formation of lactic acid after eating sweet foods or foods to which we are sensitive or allergic.

Infections: Wastes accumulate wherever the blood circulation slows down, whether because of emotional blocks leading to permanently contracted muscles (called muscle armouring), injuries, or insufficient use of muscles. Up to 30 percent of the volume of cells in elderly people may be filled with waste material, usually in the form of old-age pigments (oxidized fats and proteins). This is comparable to 30 percent of a city or a household being filled with refuse dumps. Just as dumps become infested with rodents and insects, the waste accumulations in the body provide favourable breeding grounds for microbes and parasites.

Infections are healing efforts of the body designed to destroy any foreign invaders and reduce the heavy load of toxic and obstructive residues (mucus) accumulated in the body. The use of prescription or over-the-counter drugs in fighting these infections prevents this cleansing action and reduces our vitality still further; the body's immune system - its ability to defend itself against foreign invaders - is also damaged.

If the body is still reasonably strong, it fights these invaders and simultaneously reduces the load of stored wastes. This is experienced as an acute (sudden and severe) infection, often in the form of a cold or fever. However, as the body becomes weaker and the amount of stored wastes increases, the body does not fight back anymore. It can then only limit the invasion to a certain part of the body, and chronic (low-key but long-lasting) infection results.

Finally, the immune system becomes too weak to limit the chronic infiltration of toxic substances and its distress becomes more generalized as autoimmune diseases, cancer, and leukaemia. However, it should be understood that this invasion or infiltration does not necessarily occur from the outside. Especially in the case of cancer, it is more an "uprising of germs from within" created by the disintegration of diseased body cells. It is similar to a revolution spreading through a country with dissatisfied inhabitants.

The German cancer researcher P. G. Seeger, M.D., D.S.c., demonstrated in thousands of experiments that the virulence of cancer cells depends on the nutritional status of the host. Another interesting study found that viruses breeding in hosts that are deficient in protective nutrients, such as selenium or vitamin E, tend to mutate and change from benign into very virulent and lethal forms. With sufficient levels of these nutrients, someone may be immune to the benign form of the virus but not to the mutated form that is produced within a deficient host. For example, most of the influenza epidemics with new virulent strains that periodically sweep the world originate in selenium-deficient parts of China[1]. Furthermore; the African HIV epidemic appears to have emerged in northern Zaire where many people are selenium deficient. The lesson from this is that poor nutrition breeds deadly viruses.

Food Addictions and Allergies: Health deterioration, closely linked to the weakening of the immune system, also takes place with the development of addictions and allergies. These are extremely widespread in our society. Conventionally, an allergy is seen as an immunological reaction against a foreign protein. However, in most instances, and especially in the early stages, food allergies may be better described as food sensitivities. This means that the body is sensitive to a particular food and when confronted with it, enters a state of stress. The glandular system is induced to pour adrenal hormones into the bloodstream and dopamine into the brain, and this may be felt as an uplifting stimulation. Eventually, the body begins to crave this stimulation and this leads to an addiction to that particular food. In our society, these food cravings develop mainly for sweet and fried foods, bread, dairy products, and red meats, but also for other foods that we frequently eat.

After many years of exposure to an addictive food, the glandular system is markedly weakened and stronger stimulants are required. You then become addicted and sensitive or allergic to alcohol, coffee, tea, tobacco, or various drugs. Eventually, even these will not achieve the desired effect or only for ever-shorter periods.

The body then becomes chronically fatigued and disease symptoms begin to surface more forcefully. Some symptoms can be linked directly to incompletely digested foods that are absorbed through a weak intestinal wall and now cause strong immunological reactions; these symptoms are the result of a food allergy. Disease symptoms develop selectively in parts of the body with hereditary defects or in those weakened by misuse or injured by accidents.

Problem Foods: Almost any food, if used to excess by susceptible people, can cause allergies. However, certain foods, such as rice and vegetables, rarely cause allergies, while others carry a high risk of causing them. These high-risk items are called “problem foods” because, in addition to allergies, they frequently cause health problems in people who are not susceptible to allergies.

People who are prone to food allergies generally have hypoglycaemia (low blood sugar) or diabetes, low blood pressure, a weak immune system, or a lack of digestive enzymes. People who have alkaline blood chemistry, as opposed to acidic, rarely have allergies. (For an explanation of the acid-alkaline balance, see Step 6.) Instead, they are insensitive to pain and skin irritants and may suffer from chronic degenerative diseases or become victims of an unexpected heart attack or stroke. However, the same problem foods are major contributing factors in the development of disease of both groups.

The main problem foods are as follows:

1. Cow’s milk and its products; also lactose or milk sugar added to processed foods.
2. Wheat and to a lesser extent the other gluten grains oats, rye and barley (e.g. in beer).

3. Sweet foods, especially sugar, fructose used as sweetener, and synthetic sweeteners.
4. Hydrogenated fats (for example, margarine) or heated oils and fats.
5. Overuse of alcohol, coffee, tea, tobacco, drugs, and other stimulants.

These five groups should generally be used sparingly. Foods that are incompatible with our blood-type group often cause problems. In addition, any food that contains pesticide residues or added synthetic food chemicals is a potential health hazard. I do not believe in a "safe" limit for these chemicals, especially as limits are set for each item individually while ignoring the cumulative effect of multiple chemical residues and additions.

Social and Economic Factors: With the present poor eating habits deeply ingrained in our society and the all-out chemicalisation of our lives, it is almost impossible to remain reasonably healthy - and very few people do. Almost everyone has obvious or hidden health problems. From attending to children, parents, and grandparents of the same family, I have seen how the level of health deteriorates in successive generations.

This trend is also clear from statistics and general observations for anyone who wants to see it. Statistics showing increased life expectancy due to greatly reduced infant mortality are irrelevant for judging the health of a population. These figures simply mean that more babies survive the early childhood infections, that we have better plumbing, and that more elderly people are kept artificially alive by organ transplants or life-support machines.

But this is not health. More relevant statistics show that most people in Western countries suffer from chronic diseases and that the percentage of chronically disabled people is increasing rapidly. Arthritis, cancer, and Parkinson's disease used to be diseases of old people. Now they are affecting children and juveniles. Health authorities so far have ignored the claims of natural medicine that it is the superior form of treatment for chronic and medically incurable diseases. The very fact of a high rate of chronic disease in our society attests to the inability of the conventional medical profession to successfully treat these diseases.

This is highlighted by the embarrassing phenomenon of the falling death rate during doctors' strikes. Statistics show that with a strike by doctors, the death rate in the affected population fell dramatically. In 1976 during a doctors' strike in Bogota, Colombia, the death rate fell by 35 percent. In Los Angeles County, California, it fell by 18 percent during a strike in the same year, while in Israel it fell by 50 percent during a strike in 1973. Only once before had there been a similar drop in the death rate in Israel and that was during another doctors' strike 20 years earlier. After each strike, the death rate jumped again to its normal level.

During these strikes, emergency care was provided. This means the patients who lived longer during the strikes were those with chronic diseases. They obviously benefited from a reduced exposure to medical drugs, operations, and other technological interventions. There is an abundance of statistics, estimates, and case reports of the great number of patients who are in the hospital or worse because of disease caused by medical drugs and medical technology.

Selfish Motivation: Our medical and economic leaders do not want to face reality. They brainwash the public into believing that the present health situation is completely normal. Most importantly, the economic structure of Western civilization is largely based on the production and distribution of goods and services that are contributing to poor health. These include chemicalised agriculture and food processing, the pharmaceutical industry, technological medicine, and the petrochemical and plastics industries.

The guiding motto for industry is profit, while for the consumer it is convenience. The price we all pay is the loss of our health. This situation is the natural outcome of a society based on selfish motivation. A change for the better can come only when more and more people realize that ultimately they harm themselves with selfish attitudes. Recognizing this is a first step towards improving your health.

* * * * * * * * * * * *

Step 2

THE ROAD TOWARDS HEALTH

Learn the principles of natural living to restore your health.

The first and most important step in reversing the deterioration of your health is to realize that you may have caused it yourself. Due to ignorance and habit, we select the wrong foods; out of laziness, we do not exercise; from lack of self-control, we create destructive emotions. Therefore, only the realization that poor health is the result of a lifetime of wrong living, which includes the living habits of our parents, can initiate a real change in us for the better.

This understood, the next step is to find guidance on your road towards health. There is so much contradictory advice available today that it is confusing for the aspiring health seeker. Worldwide, medicine is in a state of chaos. The sterile monopoly of the conventional medical establishment is beginning to break down and multitudes of different healers are rapidly filling the gap. Find out by reading, listening, meditating, and experimenting which healing system is most appropriate for your needs. Then start a determined healing program using careful self-observation to monitor your progress.

Measuring Health: It can be seen from the above discussion that disease is not something that happens suddenly, that one day you are healthy and the next day you are sick because cancer has been detected. Instead, your state of health has greatly deteriorated over many decades to only a fraction of what it might be in true health. Even suddenly appearing infections are possible only because of a gradual deterioration of your immune system.

Health should be measured as vitality, as the body’s ability to regenerate itself. An indication of the inherent vitality of the body is given by the speed of wound healing or bone mending and possibly hair and nail growth; in a more scientific way, we could measure it by the electric potential of body cells.

According to this yardstick, hardly anyone alive today is likely to be more than 90 percent healthy, while most people of “normal” health live at below 50 percent of their personally attainable health potential. When a chronic degenerative disease becomes manifest, the health potential drops to below 10 percent of optimal, and at death, of course, it is zero.

It is meaningless to be elated about the success of conventional medicine in greatly reducing the number of people contracting certain infectious diseases. It is the overall number of unhealthy people that counts. If one disease is eliminated, unhealthy people will contract a different one. What is needed is a general health improvement that will effectively reduce the total number of people becoming sick and ultimately hinder the development of any disease.

Eventually, with better genetic stock and improved living conditions, we should be able to live an active and enjoyable life until the age of about 120 years and then quietly withdraw from the body. This is the presently unknown "natural death." My reason for taking this position is the fact that some individuals have managed to reach an age of between 110 and 120 years without even trying, just due to their genetics. I interpret this as meaning that we all can reach this age if our gene pool is sufficiently improved. I believe that we will be able to surpass this limit if, in addition, we also consciously adopt a healthy lifestyle. However, for me it is irrelevant how long I "live," because I know that I live equally well without a biological body. The important part for me is how well I live and not how long.

In regard to the postulated natural death, I know personally of some individuals who were active until a ripe old age and then just died in their sleep, sometimes even saying goodbye to their relatives beforehand. For most of us who have serious inherited weaknesses and have misused our body for many decades, this may not be possible to achieve. Nevertheless, better health and an enjoyable life are possible for almost everyone. The key to success is in your mind!

Table 1-1: Opposite Body Forces

If the body is out of balance, then a condition may be improved by using a remedy from the same column.

Condition	**Condition**
sympathetic nervous system	parasympathetic nervous system
tension, contraction	relaxation, expansion
aggression, extroversion	placidity, introversion
anger, fear, activity	resentment, rest
rapid pulse and breathing	slow pulse and breathing
pupil dilated, mouth dry	pupils narrow, mouth waters
good muscle tone, body rigid	poor muscle tone, body soft
poor digestion	good digestion
overexcited, hyperactive	low energy, depression
high blood pressure	low blood pressure
farsighted	nearsighted
overactive thyroid/adrenals	under active thyroid/adrenals
inflammation, pain	debility, weakness
diabetes, schizophrenia	hypoglycaemia, allergies, arthritis
Remedy	**Remedy**
blue, indigo, purple	red, orange, yellow
vegetarian diet	animal protein
more potassium, magnesium	calcium (sodium), iodine
lecithin, choline, insulin	adrenaline, cholesterol
sedation, releasing energy	stimulation, giving energy
magnetic south-pointing pole	magnetic north-pointing pole
deep, slow acupressure	rapid, light acupressure

Basic Rules for Healthy Living

Avoid or minimize your contact with:

- Microwave cooking, frying; use of aluminium, non-stick and stainless steel cookware
- Wheat, gluten, cow's milk, sweetened foods
- Meat from feedlots, or processed and seasoned meats such as sausages
- Acid foods in contact with metal
- Soy protein except if sprouted or traditionally fermented such as miso
- Genetically modified foods
- Margarine, cooking oils, trans-fatty acids
- Chemical additives; artificial sweeteners such as aspartame (Nutra-sweet)
- Soft drinks and processed fruit juices
- Foods listed as inappropriate for your metabolic type or blood type (see part 5)
- Starches mixed with food high in protein or acids
- Smoking; overuse of coffee, tea, alcohol, drugs
- Precooked or reheated meals
- Chlorinated and fluoridated water
- Metal beds; innerspring mattresses; sleeping near live electric wiring or appliances
- Fluorescent lighting
- Synthetic clothing, especially worn close to the skin (such as underwear)
- Mercury amalgam tooth fillings; root-canalled or otherwise dead teeth; dental metals inside the mouth
- Mobile/ cell phones, microwave stations for telephone and computers.
- Vaccinations

Maximize your contact with:

- Fresh, raw, and organic food; purple foods (e.g., black grapes or red beets)
- Sprouted seeds; fermented foods with acidophilus and bifidobacteria; sourdough baked goods
- Seeds and nuts sprouted or soaked before eating or cooking
- Cooking in glass or enamel utensils; eating food soon after it is cooked
- Freshly ground linseed (flaxseed); extra virgin olive oil; coconut oil
- Fresh vegetable and grass juice and blended green leaves.
- Barley or wheat grass juice powder, spirulina, chlorella, pollen, lecithin, kelp
- Neutralize fruit acids or vinegar with dolomite; have a high intake of magnesium and trace minerals
- Bitter herb teas to improve digestion; hydrochloric acid as supplement if fingernails are soft
- Energy devices: crystals, magnets, pyramids
- Energized food and water
- Sleep with the head towards north, northeast, or east; pull out electric plugs close to bed before sleeping
- Sleep on latex, natural fibre, or air mattress; energize water in water beds
- Increase outdoor activity and exposure to mild sunlight without sunglasses or sunscreen
- Press tender points on feet and body; squat for bowel movements; exercise; meditate

The Health Improvement Program: The basic requirements for real healing involve the reversal of the factors previously listed (Step 1) as the cause of health degeneration. Then add natural living, good nutrition, proper breathing, sufficient physical activity, improved emotional and mental attitudes, as well as specific procedures, uppermost of which is cleansing - the purification of body and mind.

A fundamental aim of all healing is to provide the body with more vitality or a higher charge of freely circulating bioenergy. True healing is always a holistic process, involving all levels of the personality - physical, bio-energetic, emotional, mental, and spiritual.

Using the types of self-testing described in this book and possibly with the help of a suitable health professional, you can work out an individualized health improvement program for yourself as well as for interested members of your family. This program may have to be changed and modified as you progress and gain experience. Preferably, it should include the following steps:

1. Improve your nutrition: discover and break addictions and allergies; avoid problem foods, chemicalised, and refined food; pay attention to food combinations; select foods suitable for your body type and condition; use suitable supplements.
2. Remove metabolic wastes and toxins: drink plenty of unpolluted water and use colonics and laxative foods; stimulate the skin; have repeated cleansing periods.
3. Eliminate harmful microbes and parasites from your body and repopulate your intestines with healthy lactobacteria.
4. Strengthen the body: use regular exercises such as yoga, tensing and stretching, or suitable outdoor activity.
5. Improve energy flows: work out specific problems with reflexology, acupressure, ear acupuncture, magnet and meridian therapy, massage, and homeopathy.
6. Create natural living conditions: minimize the use of synthetics and plastics; reduce your exposure to pollution and harmful radiation in your home; frequently be in fresh air with mild sun exposure; bathe in the sea or unpolluted streams; walk barefoot on the grass.
7. Learn meditation and mind control: learn to express your negative emotions appropriately; practice generating positive feelings and emotions; love yourself as well as your neighbour; adopt a positive or spiritual philosophy of life; find a worthy goal or ideal and strive to manifest it.
8. Form or join a health or healing group: support yourself, heal yourself, and support others in their healing.

In addition to these self-help measures, it is advisable to seek expert professional help for specific problems. It is self-defeating, however, to view your health improvement efforts as a deadly serious matter, a “life-or-death” struggle. Like everything else in life, they are best conducted in a light and playful way, by exploring and enjoying yourself, even if it sometimes hurts. Remember: True healing means growing - learning and applying the laws of nature.

Balancing: An important general principle in healing is to aim for an overall balanced condition. Generally, every condition in our body is the result of the interaction of two opposing forces; in Chinese medical philosophy these are called yin and yang. Yin is cool and receptive; yang is hot and active. In our Western thinking, we see body processes as an interplay between overactive (yang) and underactive (yin) conditions, between tension and relaxation, and the sympathetic and parasympathetic nervous systems. However, yin can be overactive and yang underactive, depending on the individual.

As part of our healing, we aim for balance not only on the biochemical level, but also in our emotions and thinking. If you are an extrovert, for instance, aim to go into yourself by meditating, learning to listen, feel, and understand. If you are an introvert, then try to become more outgoing and open in your relationships and social interactions. Table 1-1 shows some of the opposing forces that rule our body.

Many natural therapists believe that most disease starts in the bowels or, more generally, in the gastrointestinal tract. This certainly is a common factor in many diseases and requires our initial attention. Allergy (food and substances) testing is most important for sensitive individuals and children with health problems. Fasting and cleansing are required to remove the heavy load of metabolic wastes and toxins that most of us have stored in the body. This cleansing process often leads to unpleasant symptoms - healing reactions, also called the healing crisis - which we frequently must endure before we can advance to a higher level of health. This is especially the case for those with a chronic degenerative disease.

* * * * * * * * * * * *

Step 3

INTESTINAL SANITATION AND ANTIMICROBIAL THERAPY

Sanitize your intestines and rid your body of harmful microbes.

Most diseases are caused or aided by microbes and parasites. Due to antibiotics, steroidal drugs and chemotherapy our natural intestinal bacteria commonly have been replaced with pathogenic microbes. Also microbes from childhood vaccinations may still be present, and there tend to be infection foci in root-canal treated teeth and possibly in scars due to surgery. All of these greatly weaken our immune system and allow fungi, viruses, bacteria and parasites to invade the blood and internal organs. Now we become susceptible to frequent acute infections, or to chronic infections and immune disturbances such as autoimmune diseases, allergies, diabetes, heart disease and cancer.

This antimicrobial therapy is designed to rid our body of pathogenic microbes and parasites, and at the same time regenerate our immune system. The necessary steps are intestinal sanitation or re-establishing a healthy intestinal flora, followed by a longer period of systemic antimicrobial therapy. For long-term healing also root-canal filled teeth may need to be removed.

INTESTINAL SANITATION

One of the first and most important causes of our health deterioration at the biological level is an inability to digest all the food we eat. This may be because of habitual overeating; insufficient chewing; consumption of too much milk, bread, meat, and other problem foods; wrong food combinations; emotional problems; vitamin and mineral deficiencies; or inherited weakness of our digestive organs. However, presently the main cause of intestinal distress is the use of antibiotics and other drugs that damage our protective intestinal flora.

This has two immediate effects: a proliferation of undesirable bacteria, fungi and other parasites as well as putrefaction in the large intestine. A major part of our immune system is located in the small intestine with an important centre in the appendix. For years before the onset of illness, there will be fierce fighting between our lymphatic cells and the proliferating intestinal microbes, resulting in chronic inflammation of parts of the intestinal wall.

This may cause intestinal discomfort and various digestive disorders, and in children it can often result in appendicitis. This chronic inflammation weakens the intestinal wall and allows proteins that are only partly broken down, as well as bacterial toxins, to be absorbed and cause allergies. Apart from many unpleasant symptoms, this will greatly weaken the immune system and predispose us to rheumatoid arthritis and other autoimmune diseases, and it makes us susceptible to infections and cancer.

While the intestines are often overactive during childhood due to high bacterial activity, with increasing age we tend to become more and more constipated. This is largely due to a lack of dietary fibre and a decrease in the bile flow, indicating a deficiency of lecithin and choline. The walls of the intestines become weak and deformed as with diverticulitis; hard crusts begin to cover the intestinal walls and restrict their flexibility.

Eventually, putrefaction products may cross the weakened walls of the large intestine and enter the bloodstream. This can poison the whole body, though the liver, heart, and brain usually bear the brunt of the attack. Irregular heartbeat, headaches, and fuzzy thinking are frequent initial symptoms. Every part of the nervous system may be affected. In addition, the valve between the large and small intestines (called the ileocaecal valve) may become inefficient and allow bacteria from the large intestine to invade the small intestines and the bloodstream.

When the sewerage in a city breaks down or is not working efficiently, refuse accumulates in streets and on properties and greatly contributes to the spread of diseases. It is similar in a body with inefficient intestinal elimination. Waste products accumulate in the bloodstream and tissues and cause or contribute to the development of most diseases as well as to the general deterioration of health. Colon cancer is the most obvious and direct result.

A diet of predominantly raw foods will, in time, correct these intestinal problems by supplying plenty of high-quality fibre and strengthening the intestinal walls. I do not recommend the habitual addition of bran to meals, as this makes minerals unavailable; instead, I recommend freshly ground raw linseed (flaxseed). Add one tablespoon or more to meals to ensure at least one and preferably two or three bowel movements daily.

Another consequence of microbial overgrowth, often in combination with food allergies, is inflammation of the pancreas. This leads not only to a deficiency of digestive enzymes and problems of malabsorption, but it is also a main cause of insulin-dependent diabetes. A first step in this process is the bottle-feeding of infants from birth. This deprives them of important protective factors that are normally transmitted in breast milk.

To overcome these problems we must improve our eating habits, learn to chew well, eat less, combine foods correctly, eat more raw foods, especially vegetables, eliminate harmful microbes, and reintroduce beneficial bacteria.

Garlic Flush: A key step in improving digestive function, the immune system, and overall health is the restoration of healthy intestinal flora. This is important for all health improvement but especially if chronic diseases are present, and in particular AIDS, cancer, and autoimmune diseases.

For this we reduce the pathogenic microbes in the intestines with a suitable microbicide followed by probiotics or healthy lactobacteria. I prefer fresh, raw garlic which has powerful antimicrobial and especially fungicidal properties. To carry the garlic through the length of the intestinal tract and also to minimize unpleasant reactions, such as headaches, nausea and other discomfort due to microbial die-off, it is preferable (but not essential) initially to take the garlic together with a strong laxative. This is called a flush. An effective flush is with a tablespoon of Epsom salts (magnesium sulphate) in a large glass of water. Drink some more water afterwards.

If you are constipated or overweight you may take 2 tablespoons when using it for the first time to make sure that it quickly comes out at the other end. It tastes less bitter if you refrigerate the dissolved Epsom salts overnight. In the morning you crush a large clove of garlic, mix and drink it with a small amount of water, and follow this with the refrigerated drink and some more water. You may minimize the garlic odour by crushing cloves while they are submerged in lemon juice.

During the following mornings you may use less Epsom salts or take a milder laxative instead, such as psyllium. If you dislike garlic you may put the crushed or chopped garlic into gelatine capsules. If you seem to be allergic to garlic, then use another antimicrobial, such as Lugol’s iodine solution, a teaspoon of sodium bicarbonate, 1 or 2 teaspoons of 3% hydrogen peroxide, olive leaf extract, pau d’arco extract, or wormwood. Similar if you cannot use garlic because of social restraints. Nevertheless, I believe that initially garlic is best suited to disinfect the gastrointestinal tract.

Other laxatives that may be used for a flush are Glauber Salz or sodium sulphate, castor oil, aloe vera, senna, or an isotonic flush: within about 5 minutes drink a litre of water in which is dissolved a heaped teaspoon of salt. You may need to experiment with your chosen laxative to find the right amount to take with the garlic so that it is effective within a few hours.

Psyllium is excellent and strongly recommended for long-term use. It is especially effective for removing endotoxins from the intestines and reducing allergic and autoimmune reactions. It also helps to carry any added microbicide through the whole intestinal tract. To a large glass of water add a teaspoon of sodium bicarbonate and a heaped teaspoon of psyllium hulls, stir and drink immediately followed by another glass of water. Until you start with the systemic antimicrobial therapy in the next step add chopped garlic or another antimicrobial to the psyllium drink but continue adding a teaspoon of sodium bicarbonate as long as there are signs of pain, inflammation, Candida or overacidity.

Probiotics and Ferments: This leads us to the next step: replacing the decimated pathogenic microbes with desirable or "friendly" lactobacilli. This must be done in a massive dose 30-60 minutes after each flush; otherwise the unwanted microbes may grow back to full strength after the next meal. Commonly probiotics contain acidophilus and bifido bacteria in addition to other lactobacteria.

These bacterial cultures are available from health food shops as powders or capsules that should be refrigerated. Preferably buy milk-free cultures, especially with Type 1 diabetes and allergies. High-potency cultures should contain 10 Billion or preferably 25 Billion live bacteria per gram or per capsule (keep refrigerated).

However, these dried cultures are dormant and relatively inactive as compared to bacteria in fresh or live ferments, and often have only a limited number of strains. Therefore experiment with making your own ferment as explained in Part 4. Alternatively obtain genuinely fermented sauerkraut or lactic acid fermented foods from health shops or the Internet. If you do not regularly use lactose-containing products then it is better to repopulate your intestines with bacteria grown on grains or vegetables.

Kefir grains can be used to ferment not only milk but also vegetables. For an excellent information site see users.chariot.net.au/~dna/kefirpage.html. It has the advantage of containing a wide range of protective microbes and also to work at room temperature. I do not recommend using commercial yogurt to sanitise the intestinal tract because of the high content of mucus-forming lactose and limited strains of lactobacteria.

Commonly these ferments are rather acidic and may cause digestive problems in sensitive individuals. Therefore it is preferable for sensitive individuals to partly neutralise this acidity by adding some alkalizer before ingestion, such as bicarbonate until it is only slightly sour, or you may let the ferment stand for a few hours with the addition of some dolomite powder and decant it shortly before use.

Take a cupful of ferment about 30 to 60 minutes after each flush, and possibly use more before other meals. Continue taking either probiotic cultures or ferment several times daily during the intestinal cleanse and the following systemic antimicrobial therapy. Continue this Intestinal Sanitation for a minimum of two weeks but preferably use ferments on and off for the rest of your life.

Raw Food; The type of microbes that will readily grow and possibly dominate in our intestines depends very much on the nutrients that they get, meaning it depends on the food that we eat. A negative example is the strong influence that sugar and grain products have on the growth of yeasts, and especially Candida albicans.

The opposite effect can be found in fresh raw foods, they tend to suppress harmful microbes and encourage beneficial ones. Besides raw lactic acid fermented foods, the strongest beneficial effect have green leaves and sprouted seeds. In addition to their own range of good bacteria they also provide the fiber and other nutrients for them to thrive in our intestines.

As an added bonus fresh raw foods and in particular green leaves are strongly anti-inflammatory. Therefore they are the best healing agent for inflammatory intestinal conditions and (I believe) in sufficient amounts will prevent and probably cure bowel cancer.

In addition to raw vegetable salads, a relatively easy way to ingest more fresh green leaves is as a blended pulp: blend a handful of leaves at high speed with sufficient water or juice and drink without straining. For best results the leaves should be organically grown and used as fresh as possible. Even a handful of fresh grass (cut in short length) or garden weeds may be used.

You may also add sprouted seeds to the blend but not much fruit or other sweet foods as the combination of raw vegetable fibre and sugar can produce strong bowel fermentation. To a degree that can be good for those who have problems with constipation but it needs to be properly regulated. Also fibrous fresh fruits on their own tend to stimulate bowel activity by providing food for bowel bacteria.

Hydrochloric Acid: In the elderly and especially those with chronic degenerative diseases, or often even young asthmatics, the body frequently loses its ability to produce an adequate amount of hydrochloric acid in the stomach. This leads to incomplete digestion of proteins and poor absorption of vitamins B1 and B12 as well as minerals. The resulting mineral deficiency usually causes fingernails to be soft; also the hair may be unhealthy.

Hydrochloric acid is also needed to activate pepsin (a protein-digesting enzyme in the stomach) and kill microbes in food. An inadequate amount of hydrochloric acid (HCl) in the stomach often results in overgrowth of the digestive tract with undesirable and pathogenic bacteria and fungi. This can cause or contribute to malabsorption, intestinal distress, and allergies to products resulting from bacterial and fungal breakdown, and in turn it will weaken the immune system.

Symptoms associated with insufficient gastric acid are belching, discomfort after eating or food lying heavily in the stomach, and anaemia. Often the lymphatic fluid is either over-alkaline or overacid (too far above or below neutral) and the energy level is low. In mild cases it may be helpful to add some lemon juice or cider vinegar as well as a small amount of magnesium chloride to a protein meal or drink half a cup of a bitter herb tea before meals. However, with obvious mineral deficiencies and lack of hydrochloric acid, typical of most cases of chronic degenerative disease, it is beneficial to use hydrochloric acid supplements with cooked meals that include meat, fish, eggs, cheese, legumes, or nuts.

You can easily buy hydrochloric acid tablets, commonly with added pepsin. Use one or two tablets depending on the amount of cooked protein. Alternatively you may also buy and dilute hydrochloric acid yourself. Commercial diluted hydrochloric acid is usually 20 percent HCl, while the concentrated acid (from a hardware store) is 35 percent. To make a two percent solution that is safe to use, mix one part of the 20 percent acid with nine parts of water, or one part of the 35 percent acid with 16 parts of water. You may then mix one to two teaspoons (use plastic) of two percent HCl with your meal. It may be mixed with the protein portion or used as part of the salad dressing. Alternatively, you may dilute the acid with half a cup of water or herb tea and drink it after the meal. You may protect your teeth by drinking this through a straw.

However, habitually taking acids will soon make the body overacid. Further, those who do not produce enough stomach acid also do not make enough sodium bicarbonate to neutralize an increased amount of acid in the duodenum. While the stomach content needs to be acidic, the contents of the small intestines need to be alkaline, otherwise nutrients are not properly absorbed. Therefore, it is advisable to take half a teaspoon of sodium bicarbonate in some water about two hours after a small meal and three to four hours after a large meal (with flesh food or other proteins).

SYSTEMIC ANTIMICROBIAL THERAPY

Systemic antimicrobial therapy involves cleaning the bloodstream, lymph fluid, and internal organs of pathogenic microbes. I believe the best and most powerful combination is a 3-week course of Lugol’s iodine solution followed by a gradually increasing dose of MMS. This may be followed by a longer period on

Frequently, especially with Candidiasis, there is a die-off or Herxheimer reaction during intestinal sanitation or systemic therapy. You may feel weak and nauseous. In this case temporarily interrupt or reduce the intake of antimicrobials, but make special efforts to clean out your bowels. There is some anecdotal evidence that sipping a litre of water with the addition of half a teaspoon of salt may help eliminating headaches associated with a reaction.

Lugol's Solution

Compared to MMS, iodine is much better tolerated. Therefore you can use iodine to remove much of your microbial load in a relatively easy way, and with fewer side effects than with MMS. Furthermore, iodine has a special affinity for mucous membranes which are also an attractive hiding place for fungi. Iodine is an all-round antimicrobial and an excellent fungicide.

It is commonly used as Lugol's solution with 10% of potassium iodide and 5% iodine, and may be called 5% iodine solution. Each drop contains about 6.5 mg of iodine/iodide. Other names for Lugol's solution are Aqueous Iodine Oral Solution BP or Strong Iodine Solution USP (both of which contain 130 mg iodine/iodide per ml). Iodine. Topical Solution (USP) contains 2% iodine and 2.4% potassium iodide. Therefore if you use this 2% iodine solution take 2.5 times more than the amount recommended for Lugol's solution. In the US you may also buy Iodoral tablets, each with the equivalent of 2 drops of Lugo's solution. I do not recommend pharmaceutical types of iodine dissolved in alcohol or other solvents as they often cause side-effects.

To test for allergy (which is very rare) take a drop of iodine in liquid other than just water. If there is no reaction, gradually increase to 6 to 8 drops 4 times daily or 10 drops 3 times daily in liquid or mixed with food. Do not take the iodine directly with antioxidants but it is fine to have antioxidants five or more minutes later during a meal. You may use more or less iodine according to body weight. I believe that a 3-week course adjusted for body weight is generally safe for children. However, it may not be suitable if the thyroid is overactive or with goitre; with these conditions the amount of iodine may need to be increased much more gradually and adjusted according to symptoms.

Continue for 3 weeks with the full dose, but interrupt or temporarily reduce the dose if you develop a serious reaction; if necessary repeat the course after several months. In the meantime continue with several drops daily for a much longer time. For more information on iodine see www.health-science-spirit.com/iodine.html.

MMS – Sodium Chlorite

Sodium chlorite is being used in many countries as an antimicrobial treatment in the food industry, as disinfectant, mouthwash, in toothpastes, and for water purification. As a 3.5% solution it was called 'Stabilized Electrolytes of Oxygen', and as a nominally 28% solution it is known as Miracle Mineral Supplement or MMS, or as water purification drops.

In solution sodium chlorite (NaClO2) is very alkaline and stable but when acidified it forms the gas chlorine dioxide (ClO2) which is one of the strongest all-round antimicrobial and parasite remedies. The only residue left in the body after treatment with MMS is a small amount of table salt or sodium chloride.

After intestinal sanitation and a course of iodine you may start with MMS therapy. The original procedure involves starting with 1 or 2 drops of MMS and gradually increasing this up to 15 drops once or twice daily. However, this often causes nausea to develop at higher doses. In this case you may absorb the gas from the mouth or the rectum as this usually does not cause nausea.

An alternative method for intensifying the antimicrobial program or for overcoming an acute or persistent infection is by taking 3 to 4 drops of acidified and diluted MMS every two waking hour for several days. Temporarily reduce the dose if any nausea should occur.

In all cases mix the MMS with five times the number of drops of lemon or lime juice but preferably with a more effective 10% solution of citric acid in water. You can make this yourself by dissolving 1 part of citric acid crystals in 9 parts of water. Less recommended is white vinegar (but not cider vinegar which can sometimes aggravate fungal problems).

Three to four minutes after adding the acids add half a glass of water or herb tea, or juice without added vitamin C, e.g. apple or grape juice but not orange juice. Do not take any antioxidant supplements close to MMS; generally keep them 3 or more hours apart. If it tastes too acid then you may wait 5 minutes longer and add some sodium bicarbonate shortly before drinking. Watch that it remains slightly acid because when the solution becomes alkaline chlorine dioxide converts back into sodium chlorite.

Acidified MMS acts strongest on an empty stomach but that also easily causes nausea. Generally it is better to have some food in the stomach and be able to take a higher dose. If nausea occurs temporarily reduce the dose. Commonly nausea, vomiting and diarrhea may occur sooner or later and are beneficial for cleaning out but may be avoided or minimized with the preceding intestinal sanitation and iodine therapy.

With an acute and serious infection, including malaria, influenza and urinary tract infection, you may immediately use a high dose, such as 10 or 15 drops, possibly followed an hour later by the same amount. You will vomit and be very uncomfortable for about a day but it may immediately overcome the infection. However, for chronic conditions and general health improvement it is best to increase by 1 drop each day until you feel some nausea.

Then cut back by 2 drops and stay at this level for several days until increasing again by 1 drop a day. In this way you gradually work your way higher, reducing and then increasing again to keep nausea under control. You may reduce problems by dividing the daily dose into a morning and a bedtime portion, but after some time always try to edge higher until you start feeling the nausea.

To quickly stop nausea you may take 1000 mg or more of vitamin C, but this also stops the antimicrobial activity. To avoid or minimize oxidative damage I recommend taking MMS only in the morning and evening or only once a day and use a combination of antioxidants, including vitamin C, at lunchtime or several hours away from the MMS.

The oral absorption method is designed to avoid nausea and seems also effective with smaller doses then swallowing it. To 1 part of MMS add 5 parts of 10% citric acid and after 3 or 4 minutes dilute with about 30 ml or a big mouthful of water. Wait another 5 minutes and then add up to 8 parts of a 10% sodium bicarbonate solution to protect the teeth from acid attack. This will give a pH of about 5 to 6, and one can keep it in the mouth for 5 to 20 minutes before swallowing or spitting it out. It should not become neutral or alkaline or it will become much less effective. You make the 10% bicarb solution by dissolving one level spoonful of bicarb in 9 spoonfuls of water. Instead of water you may also use fruit juice as mentioned before, or herb tea, or sweeten with Xylitol or Stevia.

I regard 6 drops of MMS as a good strong dose for oral absorption. For an acute infection you may follow up with a second dose As a general cleanse you may use this or a lower or higher dose for 2 to 3 weeks, best at bedtime. With a persistent infection continue on and off for a longer time, you may experiment with one week on a high dose of MMS alternated with several weeks on a high antioxidant intake.

The rectal absorption method is good with problems in this area, such as rectal or prostate cancer, irritable bowel, and infections, cysts and cancers of the female organs. This is similar to using coffee enemas for cancer patients.

Preferably have a bowel movement and a cleansing enema, and then try 2 acidified drops in 250 ml of water. Hold for several minutes, then expel, clean again with an enema and finally insert a larger number of drops, perhaps 4 to 8 activated drops in 250 ml water and hold for a longer time, up to 30 minutes. As with oral absorption you may also add some bicarbonate to make it less acid. Protect the anus with some fat or Vaseline. There tends to be no direct discomfort, but for several hours afterwards energy may be low, and several cleanout bowel movements may follow. You may try this once a week with gradually increasing doses.

In case of vaginal thrush you may also try one activated drop suitably diluted and later followed by some probiotics. Also try the mouthwash solution.

Mouthwash: You may make a mouthwash by diluting a teaspoon of MMS in 500 ml of water. This is only slightly alkaline and tends to release small amounts of chlorine dioxide in contact with acid-forming bacteria. It is also commercially promoted as the most effective method for removing bad breath or halitosis. It does this by oxidising smelly sulphur compounds in the mouth to non-odorous sulphates. Swish a mouthful around for a short time, gargle, and spit it out. You may also flavour the solution or make it weaker. Some people claim that regular use has protected them from ‘catching’ infections.

Furthermore, occasional oral absorption of chlorine dioxide and regular use of MMS mouthwash tend to eliminate pathogenic microbes and inflammation in the mouth. These may originate from root canals, deep tooth pockets or gum inflammations and are strongly implicated in the causation of arteriosclerosis, heart attacks and other heart diseases, rheumatoid arthritis, diabetes, prostate cancer and other cancers.

To combine the stronger antimicrobial effect of oral absorption with the convenience of a mouthwash add a drop or two of lemon juice or citric acid to a teaspoonful of mouthwash solution and immediately start swishing this in the mouth for a minute or two before spitting it out. This has a relatively mild effect on the taste buds. One teaspoonful of mouthwash contains about one drop of MMS.

A recent addition to the MMS arsenal is MMS2 or calcium hypochlorite. It is especially good for killing microbes and parasites in the intestinal tract. This is beneficial with Crohn’s disease and other inflammatory bowel conditions as well as cancer of the colon and prostate, and possibly of the pancreas and liver. For details see http://www.health-science-spirit.com/MMS2.htm.

Preferably inform yourself further by reading my MMS Internet articles at www.health-science-spirit.com, and visit http://www.jimhumble.com and http://jimhumble.biz/; Jim Humble discovered and developed the MMS treatment.

Other Antimicrobials

Fungi can be controlled with alkalizers. Due to its alkalinity sodium bicarbonate is a strong fungicide, and potassium bicarbonate is even stronger. Half to one teaspoon of bicarbonate in a glass of water taken on an empty stomach or 2-3 hours after meals can be used several times daily to sanitize the intestinal tract and alkalize the lymphatic system. Bicarbonate solution is very effective in direct contact such as for rinsing out the mouth, sinuses or vagina or as packs over sites of fungal infection and tumours.

Some additional remedies for systemic antimicrobial treatment are concentrated olive leaf extract, concentrated Pau d'arco extract, and wormwood. These may not be as strong as the combination of Lugo's solution and MMS but are milder acting and easier to use for a longer period.

The following herbal dosages are for average body weight, overweight individuals may take more and underweight individuals or children less. These remedies may not be suitable in high doses during pregnancy, in this case consult a health professional.

Further supportive methods are oxygen therapy, coconut oil or caprylic acid (Capricin), an electronic blood purifier and magnetic pulser. Some other good anti-microbial remedies are propolis, extract of Neem leaves, and the oils of tea tree, Neem tree, cinnamon, eucalyptus, mustard, oregano, shark liver, and thyme. You may try using 1 to 2 drops of essential oil four times daily diluted and ingested or kept in the mouth.

Colloidal silver is best for direct treatment of affected areas, such as douching or gargling, as after ingestion it may not be effective past the stomach. For asthma, lung infections and lung cancer I would also consider spraying colloidal silver into the mouth during deep inhalations.

Even if systemic infestations with Candida and other microbes are successfully eliminated there may still remain problems in isolated areas, such as fungi or parasites in skin patches, vagina or under toenails. These should be intensively treated with local application using a wide range of the mentioned remedies.

If you continue to have persistent long-term microbial or parasite problems use a combination of different antimicrobials and frequently vary the composition of the treatment as well as the dosage, also add alkalizing and possibly electronic zapper and pulser. Also read about the importance of the Basic Cleanse in Step 4.

Olive leaf: Olive leaf is an effective anti-microbial. It can be used as tea or powder but commonly it is now used as an extract. The active ingredient is oleuropein. Preferably use up to 1000 mg of oleuropein daily as part of the anti-microbial rotation program. Dried olive leaf has about 30 mg of oleuropein/gram.

As the tea is rather bitter it is now rarely used: simmer 30g of dried leaves in 1 litre of water for 1-2 hours, top up, strain and refrigerate, and drink 3-5 cups daily.

In powder form olive leaf extract has 200mg to 220mg/g. Use up to 5 g of powdered extract daily. You may stir half a teaspoon in water or other non-protein liquid two or three times daily and take before meals (oleuropein may be inactivated by high amino acid levels). The normal retail form is as 500mg capsules of extract, and you may use up to 10 capsules daily. Commercial liquids commonly contain only 4 mg of oleuropein per ml.

Pau d'arco; Pau d'arco (also called taheebo or lapacho) is a strong fungicide from the inner bark of a South American tree. It has the advantage of tasting reasonably pleasant. Either use extract or make tea: one tablespoon of bark or a heaped teaspoon of powder in a large cup of water. Bring to boil and let simmer or steep for 5 to 15 minutes; drink 3 cups a day. Powdered pau d'arco in capsules or cold water does not work but you may use the extract in capsules. Use 3 capsules of extract 3 times daily. For severe Candidiasis or other systemic fungus problems try to obtain 1 kg of extract powder and take several teaspoons daily.

Wormwood: Wormwood is effective against most microbes and parasites. Commonly used for Candida is Artemisia absinthium, but also Artemisia annua or Sweet Annie can be used. Both herbs have different active ingredients, and it may even be good to use a mixture of both. They should not be used during pregnancy.

To avoid the bitter taste of A. absinthium stir a rounded teaspoon of powder in some cold water or juice and drink immediately. Use a teaspoonful three times daily. However, the bitter qualities of wormwood tea stimulate the digestive system - stomach, pancreas, liver and gall bladder. It has been used as a folk remedy to cure jaundice.

Therefore, you may sometimes drink it as tea during or after a meal. Make the tea by adding a teaspoon of wormwood to a cup of boiling water and steep for ten to fifteen minutes. It is easiest to take by swallowing it quickly and, before breathing again, drinking something pleasant afterwards. For children you may also use wormwood powder in capsules (3 to 5 capsules 3 times daily), or use only Sweet Annie.

Conclusion

While orthodox medicine uses specific drugs for specific microbes, in natural medicine we use broad-based remedies that commonly cover viruses, bacteria, fungi and parasites at the same time. Medical antibiotics tend to kill our friendly gut bacteria, which then allow Candida and other fungi to invade the body and cause numerous health problems. This can be easily prevented by also taking fungicides in the form of natural remedies while using antibiotics, and adding probiotics soon afterwards.

Broad-based remedies are cheap and effective and ideally suited for self-healing. But this does not mean that they are harmless. While damaging side-effects are much less common than with drug medicine, problems can occur, especially with high doses of MMS, and for those with overactive thyroids also when using Lugol's solution. This means, you need to use common sense, observe how your body reacts, and cut back or use another remedy if indicated. This is an experimental approach in which you need to be flexible and take responsibility for your health and your actions instead of just following the orders of a doctor.

* * * * * * * * * * * *

Step 4

FASTING & CLEANSING

Become clean on the inside by removing accumulated wastes and toxins.

Fasting is a period of restricted food intake and cleansing is the removal of waste products from the body as a result of fasting. This combination of fasting and cleansing is the most effective general healing method. It is essential for overcoming chronic diseases and for rejuvenating an aging body. It is also excellent for ensuring continued health and minimizes the risk of a disabling disease later in life.

In recent years a steady stream of research articles has pointed to the accumulation of metabolic waste products inside of cells as a main factor in the causation of practically all chronic diseases. In addition to weakening the cell and its energy metabolism, these waste products are also the breeding ground for internally evolving or endogenous microbes that are a basic cause of cancer and autoimmune diseases.

Cellular Waste Products

The body recycles millions of cells and their components every second, usually without a problem. However, if cells are unhealthy, then some of their proteins, especially nucleic acids from DNA and RNA, are abnormal and cannot be properly broken down and recycled. Instead they accumulate as cellular sludge. Various researchers have shown that out of this sludge can now arise primitive microbes.

It appears that new viruses arise from RNA fragments surrounded by a cell-wall, while mycoplasma are from DNA fragments without a proper cell-wall, and nanobacteria develop from damaged mitochondria, the cellular energy factories. These primitive life forms are pleomorphic which means that they can change into higher forms of bacteria and fungi which then are at the basis of cancer and autoimmune diseases.

A large part of these cellular waste products are composed of oxidized and peroxidized or polymerized polyunsaturated fatty acids, especially trans-fatty acids, and combinations with degraded or wrongly folded proteins. Included are also metals, such as mercury, aluminium and iron. Residues with oxidized fats are commonly called age pigments, and more scientifically ‘lipofuscin’. In the aging skin it can be seen as yellow-brown or brown ‘age spots’ or ‘liver spots’.

Another waste product associated with vitamin E deficiency is called ‘ceroid’, while ‘amyloids’ are protein residues that accumulate mainly in the brain and in nerve cells, causing. Alzheimer’s disease, Parkinson’s disease and amyotrophic lateral sclerosis (ALS). Up to 70% of the volume of nerve cells may consist of such waste until cells are finally killed, although they would have lost most of their normal functions much sooner.

These waste products impair cellular functions and our overall health in many ways: they interfere with the movement of nutrients into the cell and metabolic wastes out of the cell, they poison the oxidative energy production or respiratory chain of the mitochondria and cause us to be low in energy, they interfere with the signalling capacity of nerve cells and the normal function of muscle cells, and they provide the raw material out of which the pleomorphic microbes arise that weaken our immune system and cause autoimmune attacks and cancer.

If you now understand the devastating impact of cellular wastes on your long-term health you will also realize the importance of repeated fasting and cleansing periods in your health program. Even medical research can now show that removing protein debris from cells keeps aging organs young and healthy.

The main causes leading to the accumulation of these wastes are:

1. devitalized food due to processing, cooking, and long-term storage
2. lack of enzymes due to cooking. processing and overeating
3. eating highly heated proteins and oils
4. antioxidant deficiencies
5. exposure to unbiological chemicals in food, water, air and drugs.

What to do

Just by fasting alone many chronic diseases can be healed. The only disadvantage is that it can be unpleasant, especially for beginners. Fasting is not indicated for anyone who is weak, fatigued, or underweight with difficulty gaining weight. We need a reasonably good energy level to fast and cleanse successfully.

When we stop having normal meals our digestive enzymes become available to digest and remove the sludge in our cells. While it is essential to minimize the intake of food, it should be as high as possible in vitality and enzymes, meaning that it should be fresh and raw. Ideal would be freshly cut wheat or barley grass, or leaves freshly picked in a garden. Eat any food in small amounts or drink any juices in small sips spaced out over a longer period. This is especially important for fruits or any sweet juices as from carrots etc. Keep each bite or sip in the mouth for several minutes.

Start with several one-day fasts, possibly once a week, before trying a three-day fast. Then have a three-day fast once a month; if you live according to normal working hours, have your first fast day on a Friday, as the second day often feels the worst, and this way you'll be at home.

Once or twice a year you may fast for seven days or longer. Instead of a series of strict fasts you may have longer periods on a raw food diet and use a strict fast only occasionally. With chronic diseases the need for cleansing is much greater than for general health improvement. After a year or two of good health, the number of cleansing periods may safely be reduced to twice a year, unless you use much processed food. With the onset of a cold, flu, or other infectious disease, start a fast immediately.

The more often you have a fast, the less you normally are affected by it in a negative way; a sense of peace and well-being may develop instead. As long as you are still congested during a fast, your tongue will be strongly coated and your breath may have a strong odour when you get up in the morning. Conversely, if your breath and tongue are clear in the morning, your body is reasonably clean and no more fasting is needed. To freshen your breath, you can scrape or brush your tongue clean whenever it is heavily coated; use a sharp-edged spoon for scraping.

For several days before a longer fast, start reducing your amount of food and preferably do not eat flesh or processed foods on the last day before a fast. Best have only raw fruit and vegetables on that last day. In the same way, increase the amount of food only gradually after a fast and have only raw food on the first day after a fast. The longer the fast, the less you should eat at the first meal and the longer it should take to return to the usual diet; an apple is a good food for breaking a juice or water fast. Drink plenty of fluids during the fast.

It is important to keep the bowels moving during a fast, and you may use laxatives, enemas or colonics as required. With a juice fast it is advisable to use several glasses of psyllium in water. Keeping the bowels clean is not a problem if fasting on fresh raw salads, fruits, and blended green leaves.

The effectiveness of the fast can be further increased by using a hot castor-oil pack for 3 or more days during the fast. Coat the liver area (the bottom of your right rib cage) and upper abdomen with castor oil, cover with a woollen cloth folded in four layers, and keep the area warm for an hour or two with a hot-water bottle. At bedtime drink two tablespoons of olive oil, possibly shaken in a jar with lecithin and fresh orange or lemon juice. The next morning take sufficient laxative to produce a good clean-out within a few hours.

The topical castor oil stimulates the liver’s capacity to detoxify, while the olive oil induces the gallbladder to empty and expel fat-soluble toxins from the liver, and the laxative removes the poisons quickly from the bowels. This method of cleansing causes a minimum of discomfort as the fast might otherwise result in headaches, nausea, and weakness.

Take it easy during fasts; rest whenever you feel weak, or just stay in bed if necessary. Skin brushing, deep-breathing exercises, mild sun exposure, and frequent showers are helpful.

Learn to listen to your intuition and the language of your body, rather than just following expert advice or sticking to a pre-determined program. Sometimes it may be better to ease the program and have some cooked food, You also need to keep your 'inner child' reasonably happy. Ask it for cooperation, and in return promise and deliver some treats for any hardships that have to be endured.

With chronic diseases cleansing may trigger healing reactions, lasting for days and sometimes weeks. During a reaction, pain or areas of discomfort may shift around to different body locations. During this time you may ease any strict program and do whatever is necessary to make yourself more comfortable. For a better understanding of this important subject see Step 10.

The Basic Cleanse

I recommend that you use this most effective cleanse from time to time, perhaps once or twice a year. It is designed to clean out and sanitize your gastrointestinal tract, liver, and gallbladder. Eat organic apples only for the first three days, as many as you like. Granny Smith or other tart apples are preferable, but if you feel that they are too acidic for you, then use a sweeter variety. If organic apples are not available, then remove the skin and core of conventionally grown apples.

In the evening place a hot castor-oil pack on your liver and upper abdomen and take a tablespoon of olive oil orally at bedtime. However, on the evening of the third day, drink half a cup of olive oil in this way. Then lie down immediately on your back or right side. After the first bowel movement the next morning, look in the toilet bowl for gallstones. There may be hundreds of little greenish wax balls of different sizes floating on top of the water.

First thing in the morning on the second to fourth day, drink a glass of water mixed with a crushed clove of garlic and one tablespoon of Epsom salt (to cause a good bowel clean-out within a few hours). If you cannot take Epsom salt, then try another strong herbal laxative (e.g., senna) or use the isotonic flush. If you have trouble taking raw garlic, put it into empty gelatine capsules, or use hydrogen peroxide instead. About one hour after consuming the Epsom salt and garlic, take a large dose of acidophilus-bifido culture in the form of high-potency capsules, powder, or homemade yogurt, or some self-made ferment.

If you do not have the time or energy for the full cleanse, then you can at this point return to your normal diet and do the rest at another time. Otherwise, from the fourth morning on stop using the apples, Epsom salt, and garlic, and for the next five days drink only freshly pressed vegetable juices, although these may be flavoured with apple or some other fruit. Every three hours (5 times daily) take a heaped teaspoon of psyllium hulls together with two to three teaspoons of ground linseed (flaxseed). Mix this in a glass of water and drink immediately, followed by more water. The linseed should be freshly ground in a blender or coffee grinder and stored frozen or refrigerated. The purpose is to fill the bowels, with a moist mass that will soften and dislodge any hard crusts and putrefying matter. This is essential and very effective with all inflammatory bowel conditions, diverticulitis, intestinal parasites, and cancer.

Take a spoonful of lecithin granules and plenty of protein-digesting enzymes together with the juices. With every glassful of juice, you can take two or three tablets such as pancreatin, bromelain, or papain. Alternatively and even better, if available, use mature green papaya, when it turns yellow inside and the seeds are already black. You can make a Smoothie by blending it with some vegetable juice or yogurt; use up to half a medium-size papaya a day. You could also use the juice of papaya leaves or immature fruit, but this may need to be flavoured and well diluted, as it is bitter.

On the morning after the last juice day (or possibly at bedtime the night before) take a tablespoon of Epsom salt in water or generally enough laxative to produce a strong bowel clean-out during the morning. Often long strings and parts of compacted faecal matter, which look like sections of bowel, are eliminated. It is advisable to increase the variety of food very gradually while doing allergy testing at the same time (see Step 5).

There are many possible variations of this Basic Cleanse. For instance, if you do not have a juice extractor, you can either continue longer with apples or have a few small raw salads instead. If you cannot get organic apples but have plenty of wheat or barley grass and vegetables for juicing, then you can extend the juice schedule from five to eight days; have the castor-oil packs, olive oil, garlic, Epsom salt, and acidophilus during the first three or four days. Also highly recommended are fresh green leaves blended in water for part or all of the cleanse.

Another possibility, especially good for those with high blood pressure, is to use freshly pressed lemon juice instead of the vegetable juice and possibly instead of the apples as well. Use psyllium hulls and ground linseed as for the last five days of the Basic Cleanse, but instead of vegetable juices drink the juice of five to eight lemons diluted in three or four quarts of pure water. You may do this for the full eight days or instead of the vegetable juice for the last five days. If doing this for the full eight days, then use castor-oil packs, olive oil, Epsom salt, garlic, and acidophilus (as described for the apple days).

Another often recommended additive is Bentonite, a form of powdered or liquid absorbent clay. This helps to remove plaque or old coating from colon walls and also absorbs toxins. Each time you use the ground linseed and psyllium, add either one teaspoon of powdered or one tablespoon of liquid Bentonite, which you can consume orally.

Lymphacising

A good way to greatly increase the cleansing action during a fast is lymphacising - gently bouncing on a rebounder or mini-trampoline. This speeds up the lymph flow to many times its normal rate. Lymph is the blood fluid that filters through the walls of arteries and capillaries into the tissue, delivering oxygen and nutrients to individual cells and carrying away carbon dioxide and other waste products. However, if the cellular waste production is high and lymph flow is slow, cells begin to clog up and become inefficient, preparing the ground for the development of a disease. Increasing the lymph flow greatly helps to reduce this danger; bouncing on a rebounder can do this.

Cleansing the body involves dislodging wastes and toxins from within and between cells and transporting them with the lymph fluid to the large veins near the neck; there they enter the bloodstream to be discharged through the liver and kidneys.

A greatly reduced food intake is the main way to dislodge stored wastes and toxins. Muscle activity propels the lymph flow; castor-oil packs, fruit acids, olive oil, bitter herbs, and sulphur compounds stimulate liver detoxification of fat-soluble waste while a high fluid intake and herbs for the kidneys, such as dandelion, help remove water-soluble wastes.

During fasting we usually do not have much energy for muscle activity, but lymphacising can come to the rescue as it does not consume much energy and one feels energized afterwards. With each bounce on the mini-trampoline the lymph is pushed towards the neck and prevented from falling back again because the valves through which it pushes are one-way (also see Step 9).

Another way of speeding up your lymph circulation is by inverting your body. Special inversion equipment, on which you can hang upside down, is available, or you can use an inclined board, simply put your legs up against a wall while lying on the floor, do the leg motions of bicycling while in the inverted position, or do a shoulder stand or headstand as described in hatha yoga.

THE LONG CLEANSE

A Long Cleanse of about 6 weeks duration on either a small amount of vegetable juice or grapes has been used to shrink or dissolve tumours and normalize lymph and blood cancers in thousands of cancer victims. If the body is starving it digests its own non-essential tissues, and it starts with diseased parts and tumours. Famous examples of this kind of cleanse are the Breuss Cure on half a litre of fresh vegetable juice; the grape cure of Johanna Brandt; the wheat grass cure of Ann Wigmore; and even the urine fast of John Armstrong.

It works even better with large than with small tumours as these are a source of protein, but it also eliminates the large amount of protein debris that is a main cause of cancer. This long cleanse is usually the best way to eliminate or greatly reduce tumours without strong and unpleasant side effects. It is especially suited for overweight individuals.

On the ***Breuss Cure*** you drink half a litre of fresh organic juices daily, extracted in a non-centrifugal way. The usual composition is 3 parts of beetroot, 1 of carrot and 1 of grass juice or celery, and preferably a small amount of horseradish, radish or turnip or their leaves.

As an alternative you may use 1-2 litres of juice consisting mainly of about equal parts of green juice as from grass and leaves, and beetroot-carrot juice flavoured with ginger and pollen, in addition to unstrained blended green leaves. Hot onion broth may be used in the evening, purple onions are best; cook the skins as well. This larger amount of juice is more suitable for maintaining energy without compromising the cleansing effect. The weight loss during a 6-week fast is usually between 5 and 15 kg, mainly due to losing water and from half to one kilogram of fat per week.

With the ***grape cure*** eat black or purple grapes. Eat them with skins and pips. Preferably chew the pips to release their antioxidants. Soak the grapes best with the addition of some hydrogen peroxide to kill and remove any fungi. Eat about 1kg daily according to appetite. If a temporary aversion to grapes develops, have a few days of juices in between. Eat only a few grapes at a time to avoid a strong rise in the blood sugar level. Store grapes cool and dry, you may hang them over a washing line in a garage.

If you feel that you definitely need something savoury to keep going, prepare and drink a warm strained vegetable broth in the evening, including plenty of purple onions. Also simmer the outer purple onion skins. Furthermore, you may make a tea from savoury herbs and spices or turmeric and ginger.

Different Cleansing Methods

Apples are good for the first three-day fast. Preferably obtain unsprayed apples, otherwise either peel or scrub them in warm soapy water and rinse well. Do not ingest waxed skins. Eat as many apples as you like, generally the more the better to keep the bowels moving. Apple fasts are usually well tolerated, and they are also good for children.

Fasting on freshly pressed vegetable juices is somewhat more severe but generally well tolerated. Preferably use plenty of green juice made from wheat or barley grass, vegetable leaves, and celery, flavoured with sweet vegetables such as carrot, beet root, or pumpkin, and possibly an apple. This juice may be diluted with water for a total fluid intake of two to three litres daily. If you are sensitive to pain, cold, or skin irritants, minimize the amount of sweet juices you consume and possibly add some alkalizers.

For sensitive, weak, or underweight individuals, a rice diet lasting ten or more days may initially be better. Eat only cooked brown rice, lightly flavoured with salt and olive oil and supplemented with preferably one to three tablespoons per meal of unheated ground linseed (flaxseed). Use sufficient linseed to have three easy, bulky bowel movements daily. You do not need to eat “proper” meals; just have a few spoonfuls whenever you feel hungry, and chew it very slowly. Preferably follow up the rice diet with a period of eating raw food only; or alternate periods on cooked rice and raw foods several times. However, if brown rice is not tolerated, try white rice. A rice diet helps to overcome allergy problems.

For non-sensitive individuals, a three-day lemon-juice fast is excellent and especially good for improving liver functions; it is also recommended for those with high blood pressure and who are overweight. About ten times during the day drink a large glass of water to which is added the fresh juice of half a lemon. You may flavour this with a small amount of fresh grapefruit or orange juice. Later you can extend this fast to seven days or longer. This is especially helpful with diabetes.

The “grape cure” is not only known as a cancer cure, but is also good for coronary artery disease and other chronic diseases. You eat about one kilogram of grapes spaced out in intervals during the day to avoid any strong rise in the blood sugar level. Continue for several weeks. Black grapes are best, otherwise use purple or red.

Unfortunately, it is difficult to obtain unsprayed, organic grapes. If you have to use sprayed grapes, wash them thoroughly in warm soapy water and rinse well. If you are sensitive to moulds, soak the grapes in diluted hydrogen peroxide or expose them to sunlight before a last rinse and eating.

For a longer cleansing period, use a mixed fast. You may, for instance, start with several days on a rice diet and follow this with a period of raw food only, with a time on apples or grapes and a further period on juices. Then repeat this sequence in reverse order with days of eating only fruit, raw food, and finally rice before returning to your normal diet. The cleansing period may last for several weeks, or even months, with repeated sequences.

Unless you have an acute infectious disease, do not start a strict fast when you feel weak. In this case, you can alternate between a rice/vegetable diet and a raw food diet. Those with acute infections should have only fresh fruit, fresh salads or vegetable juice until health improves and good appetite returns.

A **urine fast** has the advantage of providing vital nutrients and activating the immune system while ingesting a fluid that is already charged with your own life force, It is not as difficulty to do as it may sound because after a few days it is so diluted that it just tastes like water. For details see Step 22.

If you suspect your body is overloaded with chemicals or that your liver’s ability to detoxify is reduced, do only mild fasts initially, such as the rice diet and raw food diet. Also supply yourself with plenty of sulphur in the form of MSM (see Part 3) or, alternatively, use onions, raw egg yolk, and possibly the amino acids cysteine and methionine, or even powdered sulphur. Regular cleansing is the more important the older and physically less active you are.

Mucus problems are very common. This is obvious when we have a cold with a congested or runny nose, but at other times it may be noticed as congested or inflamed sinuses, ear or hearing problems, including tinnitus and “glue ear” in children, glaucoma, and throat and lung problems, especially asthma. To overcome these conditions, see Steps 38 and 44.

An **oil rinse** can be helpful. Fat-soluble toxins can be removed from the system before they reach the liver. First thing in the morning take a large sip of a light oil, such as sunflower or grape seed (stored in glass bottles), and vigorously swish it around the mouth for 10 to 20 minutes; do not swallow it; spit it out (be careful: it may clog the sink). Rinse your mouth several times with warm water. Continue this daily for several months. This rinse is especially important if the lymph system is congested, as it commonly is with cancer and other chronic diseases, and it is very effective with health problems in the head and neck area, including dental problems.

At other times, you can use distilled water to filter out water-soluble toxins. After rising in the morning and again at bedtime, use a cup of distilled water. Take a mouthful and hold it under the tongue for about two minutes. Then spit it out and repeat the process until the cup is empty.

* * * * * * * * * * * *

Step 5

ALLERGY TESTING

Discover and overcome your food allergies and chemical sensitivities.

Usually we are addicted to any daily-used food or substance to which we are also allergic. The reverse is also true, that we have allergies to addictive foods. Initially when we eat a sensitizing food (one that may produce an allergic reaction), we experience stimulation due to the release of adrenal stress hormones. However, eventually the adrenal glands become exhausted, and we develop a chronic degenerative disease and lack energy. In this book, I use the term “food allergy” in a general way to include not only genuine immunological reactions, but also food intolerances and chemical sensitivities.

In 1936, Hans Selye, a Canadian professor of experimental medicine and surgery, discovered the allergy-stress mechanism. He was the first to describe this mechanism, which he did in his pioneering book The Stress of Life. So far the medical profession has not yet grasped the importance of his fundamental discovery for understanding the diseases of our society, I assume because it is not profitable to do so. However, in light of this stress model, we now know what we need to do to heal ourselves.

Initially an environmental challenge such as an incompatible food or emotional stress causes an alarm reaction. The adrenal glands release inflammatory hormones, and the sympathetic nervous system is overstimulated. This may result in an acute allergic reaction or a general inflammatory condition, hyper-excitability, increased blood pressure, palpitation, aggressiveness, anxiety or anger, and poor digestion. This incidentally, is the basis for the connection between junk food consumption and criminality.[2]

If we continue eating the same problem food nearly every day or continue to be plagued by stressful memories or conditions, then the stress becomes permanent and the body adapts by releasing anti-inflammatory hormones. The symptoms of the alarm reaction with its inflammatory tendency subside. This is the resistance phase, a state of adaptation for a hidden or masked allergy. Commonly, the alarm reaction occurs in early childhood when we are first introduced to wheat, cow’s milk, maize, or soy milk and then settles down to several decades of hidden allergies with minimal outward symptoms.

However, eventually the capacity of the adrenal glands to produce sufficient anti-inflammatory hormones becomes exhausted and we enter the exhaustion phase of the allergy. Now we have a maladaptation to allergens and other stressful conditions with chronic and generalized inflammations. These can manifest as arthritis, cancer, heart disease, and any of the other diseases typical in our society. Finally, even this chronic inflammatory condition subsides and we have the stage of advanced old age with senility, debility, and the final non-sensitive stages of degenerative diseases.

This sequence of ill health shows us the requirements for genuine health improvement. By permanently removing the offending problem food or other chronic stressor and at the same time supporting the body with high-quality nutrition we slowly begin climbing back to health. We leave the exhaustion phase and re-enter the resistance phase with a period of quiet recovery, and finally we move back into the alarm phase with a series of acute but beneficial healing crises. At the end of this long healing process, we may have reacquired the adrenal function and disease resistance that we had as a small child before we started getting colds, allergic reactions, and digestive upsets due to consuming problem foods.

Sometimes energetic individuals, often males with blue-white irises who cannot relax, experience a prolonged period of lethargy when avoiding wheat or gluten products. During this time, the adrenal glands have time to recover. Others who are already closer to adrenal exhaustion feel much stronger as soon as they avoid allergenic foods. The main allergenic foods tend to be the same as the most commonly eaten foods and include wheat, cow's milk, eggs, soy, corn, peanuts, and synthetic food additives (for a compilation of possible food allergy symptoms, see Table 5-2).

Addictions and allergies are part of the stress syndrome. Prolonged severe stress of any kind leads to adrenal exhaustion in the same way that an addictive allergy does. The only real solution lies in avoiding excessive stress factors such as allergenic foods, work pressures, and so forth.

Discovering your allergies and addictions is an essential step in your healing process. You may do this by adopting a five-day water fast or rice diet as described in the following pages.

After complete avoidance of an allergenic food (one you suspect or know produces allergy symptoms) and while reducing the overall allergenic load from other influences at the same time, one usually becomes hypersensitive to that food for a period lasting from a few weeks to several months. After this, the sensitivity to the offending food gradually declines. If during the hypersensitive period you are occasionally exposed to the allergenic food, then an immediate strong reaction can be triggered, resulting in pain, weakness, a rash, or whatever its characteristic effect is on your body.

If this is done repeatedly but separated by at least four days between exposures, this may delay the complete recovery, but is not likely to result in serious internal damage. However, permanent damage will result if you use this food so frequently that you do not get an immediate reaction any more or only a much weaker reaction than initially.

Generally, however, it is essential to avoid the offending food as much as possible for an extended period in order to heal yourself completely. Depending on the severity of your initial symptoms or degenerative disease this may be from a few months up to several years. How soon you can overcome your allergic tendencies depends mainly on the effort you put into adopting a comprehensive health improvement program.

Types of Allergies: Depending on its severity and mode of manifestation, we may distinguish five types of allergy: addictive, cyclic, fixed, multiple, and trigger.

An *addictive allergy* induces us to eat our favourite food every day or we are threatened with withdrawal symptoms. Therefore, we have a craving for a certain food and eating or drinking it makes us feel better. The person usually is not aware of the addiction and for this reason this type of allergy is also called a hidden or masked allergy, corresponding to the period of adaptation.

In a *cyclic allergy*, we have a reaction to a specific food only if it is eaten in excess, either by eating large quantities of it at any one time or by eating it on several successive days. In addition, a reaction to a cyclic allergen can be triggered by other adverse factors, especially by stress as from emotional and mental problems or from environmental agents such as chemical fumes, tobacco smoke, and so on. Combining two different cyclic allergens in the same meal can also cause a reaction due to the effects of both. Poor food combining or a generally acidic body condition may be further contributing factors.

If you realize that you have a cyclic allergy - the food agrees with you sometimes but not always - then eat it only on a rotation basis, not more often than once every four or five days. Fruits and the lactose content of dairy products often belong in this category. They may not cause immediate reactions, but eating fruits on successive days can cause an increasing degree of overacidity and skin problems in sensitive individuals, while an oversupply of lactose can lead to a build-up of mucus. Both, in turn, can trigger secondary symptoms such as arthritic pains.

With a *fixed allergy*, one reacts every time the offending food is eaten, no matter how small the quantity. Fixed allergies, therefore, are non-adapted allergies and thus easy to recognize. One soon becomes efficient in avoiding offending items. Fortunately, only a small percentage of all allergies are of the fixed variety; most are addictive or cyclic in nature.

With *multiple allergies*, we react to a wide range of foods and chemicals and quickly become allergic to any replacement foods. This has sometimes been called being allergic to the twentieth century. The main causes are deficiencies of zinc, digestive enzymes, and gastric acid, weak liver function, and bacterial overgrowth of the digestive tract, often with systemic Candida infestation. The solution is intestinal sanitation, a raw food diet, and a high intake of zinc and sulphur compounds. Before testing for allergies by elimination, sanitize the intestines and minimize any deficiencies by using vitamin and mineral supplements.

We can also speak of a fifth type, trigger allergy. Attacks of asthma or hay fever are often triggered by inhaled substances such as pollen or house dust with residues from house mites. However, such reactions would not occur if one’s mucous membranes were not already hypersensitive because of underlying food allergies, mucus congestion, and various deficiencies. (A deficiency is a state of higher nutrient requirement than is available.) In a similar way, food chemicals can act as triggers for hyperactivity. A hypersensitive skin reacts very strongly to insect bites because of its high histamine levels due to underlying food allergy and overacidity. Removing the triggers will help to some degree, but the underlying health deterioration continues and a new trigger can readily be adopted.

Elimination Testing: There are various forms of allergy testing, some technical, others a simple skin scratch method, and there is also muscle testing or pulse testing. None of these, however, is as reliable as proper elimination testing, called the elimination diet. Here’s how to do it:

For five days or longer, abstain from all your commonly eaten foods, as well as alcohol, nicotine, or other drugs. During this time, try to minimize any other possible allergenic influences such as exposure to solvent or car exhaust fumes, house dust, moulds, gas, or cigarette smoke - basically anything with a strong or unusual smell.

Starting on the second day, there may be more or less troublesome withdrawal symptoms, with cravings for the food or stimulant to which you are addicted. You may have headaches, dizziness, weakness, and other problems. On the fourth day, these symptoms usually subside and you begin to feel much better. Often symptoms of long-standing diseases such as arthritis, heart pain, or schizophrenia disappear as well.

If you still do not feel right on the fifth day, it may be better to extend the program until you do feel well. Chemicals and drugs, be they medical or recreational, take much longer to clear from the body. It will greatly speed up the elimination and lessen your discomfort during the fast if you take a strong laxative on the first morning of the fast, using an isotonic flush, herbal laxative, or a tablespoon of Epsom salts in water; in addition, drink a lot of pure water.

The most reliable fast is on pure water only, which means unchlorinated and unfluoridated, and not stored in plastic containers. In a less strict form, it should be all right to use a weak, unsweetened herb tea of a variety you do not habitually use. Non-sensitive (non-reactive) individuals can become more balanced by having a fast on diluted lemon juice, while sensitive individuals can use cooked white or brown rice only, preferably unflavoured; otherwise use a small amount of olive oil, herbs, and possibly salt with the rice. If rice has been habitually eaten before, then use cooked millet, sago, or arrow root instead. Eat as much as you like and drink about two litres of fluid daily. If you use a cleansing period as recommended in Step 4 then you may do elimination testing when gradually re-introducing different foods.

If you fast on fluids, use another laxative on the last day of the fast, normally the fifth day. On the morning of the sixth day, eat cooked rice as the first test meal, possibly lightly salted but with no other additions. As explained in Part 2, check your pulse rate for a full minute before a meal, and 30 and 60 minutes after the meal. Always test sitting in the same position and when the pulse rate is not increased from previous physical activity.

Before and after each test meal, make the pulse tests and any other relevant tests. If there is a reaction involving a combination of newly tested foods, repeat testing with the individual food components at a later date. If the pulse after a test meal is significantly higher than after the first test meal of rice, say more than 10 percent, then this indicates a possible allergy or incompatibility; the higher the pulse rises after the meal the more likely is it that a food is allergenic.

Make other tests relevant for your condition. If you have weak eyesight, you can check your reading ability with a wall chart; schoolchildren can be tested for clarity of writing and speech. Check the range and pain level of impaired joints or muscles, and watch for any sign of bodily or emotional discomfort. Keep a detailed diary of the pulse rates, measurements, and any unusual happenings, feelings, discomforts, and so on. Those with diabetes should also test their blood sugar level, and those with hypertension preferably should check their blood pressure (you can easily buy a sphygmomanometer to do this).

With each subsequent meal, expand the basic rice meal with one or more additions, such as oil and various vegetables. Test a sprout salad and then other basic foods. You can have four or five small meals a day. In the beginning, test mainly foods that you expect to be safe. Test suspected allergenic foods at the last meal of the day, so you can sleep off any prolonged reactions. Use suspect foods only individually or with previously tested foods.

You can stop a strong food allergy reaction by taking one level teaspoon of sodium bicarbonate (baking soda) or, better, two parts baking soda with one part potassium bicarbonate. Also take one tablespoon of Epsom salt or milk of magnesia as a laxative if the reaction occurs soon after the meal. On a normal diet, however, reactions can be delayed and manifest up to two days later.

Reactions can also be stopped using homeopathy. Liquefy a sample of the allergenic food, dilute one part of it with nine parts water, and shake in a bottle or jar about 40 times with a strong downward movement. Dilute one part of this with nine parts water and so forth until the dilution process has been done four times in all. Then place one teaspoon of the fourth dilution in your mouth and hold it there for a few minutes before swallowing. Then make two more dilutions and keep a teaspoon of each of these in your mouth in the same way. Take additional teaspoons of all or just the last dilution some time later if still required.

Test offending foods again three to six months later. Commonly, small, infrequently consumed amounts can then be tolerated. Eat suspect or problem foods no more than once a week. If conditions deteriorate again after some time, return to testing or adopt a safe, low-allergen diet. The body can be made less sensitive to allergic reactions by making the body more alkaline, as in Step 6.

* * * * * * * * * * *

Step 6

THE ACID-ALKALINE BALANCE

Bring your body back into balance by testing for and correcting any over-acid or over-alkaline condition.

Our blood is slightly alkaline, and the body makes every effort to maintain this alkalinity at a constant level. For this purpose, we normally have an ample reserve of alkalizing minerals. Most of our foods supply minerals. The total balance of minerals in a particular food can be either acid or alkaline. Mineral salts are composed of an acid group (anion), such as chloride, citrate, or phosphate, and of an alkaline group (cation), mainly metal ions such as sodium, calcium, or potassium. If one of these groups is stronger, then the salt is either acid or alkaline.

In plants, metal ions such as potassium are combined with organic acids such as citric acid. In the body, the organic acids are converted to energy, while the metal ions remain to produce an alkaline reaction. Therefore, we say vegetables and fruits are alkalizing. Animal tissue, on the other hand, contains a high percentage of phosphoric acid bound to proteins. Some of the proteins are converted to energy and an acid remains. Accordingly, we can classify our food as either alkalizing or acid-forming. Here's a rule of thumb:

Alkalizing foods: vegetables, fruits, sprouted seeds, almonds, most legumes

Acid-forming foods: meat, fish, eggs, cheese, most grains and nuts

Neutral foods: refined starches and sugar, fats, oils

In order to maintain an amply alkaline body reserve, eat approximately four times the weight of alkalizing food compared to acid-forming food, or 80 percent alkalizing to 20 acidifying food. There are food tables available to show the amount of acid or alkaline equivalents in different foods. However, I regard these as useless because some of the listed foods may change their values in the body due to specific, individual metabolic problems.

Sugar and chemically neutral and normally alkalizing fruits become acid-forming in sensitive individuals, while a high intake of meat and fat can lead to a lack of acids required for energy production. No matter how alkalizing a food is supposed to be, if you are sensitive or allergic to it, then it is acid-forming for you. Even the same food may change from alkalizing to acid-forming if it is incorrectly combined with other foods or if consumed when you are emotionally upset.

All of this is contrary to what you would expect by using acid-alkaline food tables. The values in these tables reflect the typical mineral content of the food, but this varies greatly depending on soil and storage conditions, production methods, and cooking. Therefore, the only reliable and meaningful method is to observe your body: skin sensitivity, tendency towards inflammation, allergic reactions, and possibly to test your urine and saliva for acidity.

The Regulation of Body Acidity: The body tries to keep the blood acidity constant within a very narrow band, between pH 7.35 and 7.45. This is actually a slightly alkaline condition. The pH is a measure of acidity or alkalinity and indicates the hydrogen ion concentration in a fluid. These hydrogen ions are the originators of acidity. A pH of 1 indicates the highest hydrogen ion concentration and with this the highest acidity. At the other end of the scale is a pH of 14, with the lowest hydrogen ion concentration and the highest alkalinity. A pH of 7 is neutral.

In our younger years, our metabolism often is too fast; too many acids are produced and the body becomes chronically overacid. Later in life, however, with decreasing muscle activity and vitality, not enough metabolic acids are produced, potassium is lost from the cells, and the metabolism becomes sluggish. With this, the body often becomes chronically over-alkaline and loses its ability to regenerate and heal infections or wounds.

By excreting more acids or more alkalis through the urine, the body tries to maintain a normal pH level. The body also does so by changing the breathing rate: An increased volume of breathing makes the blood more alkaline by releasing additional carbon dioxide into the air; decreased breathing, on the other hand, increases the carbon dioxide level in the blood and makes it more acid. The balance in the blood between carbon dioxide and sodium bicarbonate (baking soda) is the main buffer system to keep the blood pH constant.

With an efficient metabolism, the pH of the urine depends mainly on the acid-alkaline balance of the diet. A diet high in acid-forming foods will naturally produce acid urine. This is not dangerous as such but it is an indicator that the diet is unbalanced and may in time lead to exhaustion of the body’s alkaline reserve.

Similarly, on a diet high in alkalizing foods, the urine is always alkaline. This can lead to stone formation in the urinary tract, and it may show that the cells don’t have enough metabolic acids for efficient energy production. A normal mixed diet containing about 80 percent alkalizing foods and 20 percent acidifying foods should give a slightly acid morning urine, while urine formed after meals is naturally slightly alkaline as a result of the production of gastric acid.

Test Your Body Acidity: To obtain an indication of the acid-alkaline balance of your body, test your first morning urine from time to time. In evaluating the results, take into account whether your present diet in general and the last meal in particular were balanced or acid or alkaline forming. If you test the urine during the day, the result may be much more variable because of the temporary alkaline tide after eating. Urine paper should have a range from pH 5.5 to 7.5. You can obtain it over the internet. For an approximate test, you can use litmus paper which turns red with an acid reaction and blue when it is alkaline.

Instead of using pH strips you may use the yellow powder turmeric or possibly curry powder. Dissolve a teaspoon of powder in half a litre of methylated spirits, shake and let settle to produce a clear yellow solution. To make a test pour a teaspoon of the yellow solution into a small drinking glass. Add a teaspoon of urine or saliva, if it turns ruby red then what was added was alkaline with a pH greater than 6.8, if it stays yellow then it is still acid and less than pH 6.8.

Also red cabbage juice can be used as pH indicator. The colour at pH 2 is red, at pH 4 purple, pH 6 violet, pH 8 blue, pH 10 blue-green, and pH 12 green-yellow. Finely chop some red cabbage, pour hot water over it and soak for 10 to 20 minutes, then filter through some gauze or tissue. In a small clear glass add a spoonful of filtered juice to a similar amount of a test liquid, stir gently and compare with the colour chart. Refrigerated the rest of the juice, use within a week or two, or keep frozen. Also other purple juices work in a similar way, just experiment.

It is also valuable to check saliva acidity, especially if there are any digestive difficulties and if you do not seem to get sufficient value from your food. An allergic reaction makes the saliva too acid, so you can use this test as part of your allergy testing program.

Before a meal, the normal acidity of the saliva is 6.4 to 6.7, while 30 to 60 minutes after the meal, it should be above 6.8. If test results remain too low you may be to acid or produce not enough stomach acid. For testing first clean your mouth with plain water, fill it with saliva, swallow, and then test the second lot of accumulating saliva.

Acidity Problems: The main reasons people become overacid are:

1. Inefficient metabolism
2. Stress reactions due to food allergy, chemical sensitivity, emotional upset
3. Eating sweet foods when the blood sugar regulation is poor
4. Acids accumulating after eating fruit
5. Strenuous muscle activity
6. Candida albicans infestation

Chronic overacidity has a serious effect on the body. It causes a continual loss of minerals needed to neutralize excess acid and the body becomes hypersensitive. The more acidic the body becomes, the more histamine (the allergy chemical) is released. This means acidity causes and greatly increases inflammations and allergies, and it makes the skin very sensitive to insect bites and other irritants. Remember that degenerative diseases are associated with chronic inflammation.

Overacidity of the body is often felt as pain; it is most prominent in arthritis and rheumatism. If the alkaline reserve is insufficient to neutralize these acids, more calcium will be mobilized from the bones for this purpose, thereby causing the bones to become brittle and the tissues and joints to calcify. Further, tumours grow only if the surrounding tissue is too acid. Acidity around a tumour may be due to the lactic acid production of the tumour itself; however, sensitive individuals are generally too acid and this promotes rapid tumour growth and pain.

Table 1-2: Symptoms of Calcium–Phosphorus Imbalance

Calcium level too high	**Phosphorus level too high**
Colds and flu are rare	Frequent colds and flu
Insensitive to pain and noise	Sensitive to pain and noise
Insensitive skin	Sensitive skin
Signs of magnesium deficiency	Signs of calcium deficiency
Blood pressure raised	Tendency to low blood pressure
Blood is too alkaline	Blood is too acid
Caries of the crowns	Caries near the gum
Tartar on teeth	Gums very red and tender: pyorrhoea,
gingivitis	
Calcium kidney stones	Tendency to inflammations & swellings
Stiffness, rigidity	Red-rimmed eyes
Fusion of vertebrae	Headaches, nausea
Arthritic deposits causing bone deformations	Acute arthritic attacks
Arteriosclerosis	Allergies, palpitations

Alkalizing the Body: Fresh green vegetable juices and vegetable broth are strongly alkalizing foods, especially the broth of boiled potato peelings. While fresh fruits increase the acidity of sensitive individuals, they are usually beneficial for those with a normal carbohydrate metabolism, raised blood pressure, and insensitive skin.

Habitually taking several grams of vitamin C in the form of ascorbic acid can deplete mineral reserves or increase body acidity. This may lead to inflammation, increased pain, and other problems in sensitive people. Therefore, for such people it is advisable to use the neutral sodium or calcium ascorbate.

The best alkalizing food cannot bring lasting improvement if you continue to eat food to which you are allergic or if you eat sweet foods and fruits when you are overly sensitive.

Alkalizing with Sodium and Potassium: Sodium is the highest mineral in the blood, and sodium bicarbonate is well tolerated but should not be taken with meals as it then reduces our stomach acid. Potassium citrate can be mixed with meals and it as well as potassium carbonate and bicarbonate have the strongest alkalizing effect of any mineral supplements.

Whether sodium or potassium is preferable for alkalizing depends on some additional factors. The following conditions favour the use of potassium:

1. A high sodium and low potassium intake as in a conventional Western diet
2. Raised and especially high blood pressure
3. Kidney problems or water retention/edema.

Sodium bicarbonate, on the other hand, is called for:

1. With low blood pressure
2. With a diet high in fruits and vegetables and low in salted foods.

Those who do not fit into one of these categories may take both alkalizers. Commonly a ratio of 2 parts of sodium bicarbonate to 1 part of potassium citrate or bicarbonate is being used, but also 1 : 1 or an excess of potassium is alright. These may be mixed and taken together in the same drink, or they may be taken separately during the day. If you use more than 1 teaspoonful then space the intake out during the day. Do not take more than 3 level teaspoons of potassium salts in a day mixed with meals as this may cause irregular heart beat.

Alkalizing with Calcium and Magnesium: A recommended way of obtaining more calcium and magnesium and at the same time alkalizing the body is to dissolve dolomite powder in vinegar or lemon juice. In the body, the organic acid tends to be broken down and converted to energy while the mineral remains to alkalize the body.

Keep in a jar several tablespoons of finely powdered dolomite mixed in a glassful of vinegar. You can stir or shake the mixture occasionally to speed up the reaction. Drink about 50 ml of the decanted liquid daily, either before one meal or divided before two meals, best diluted with water or other liquid. This supplies approximately 500 mg of calcium and 300 mg of magnesium.

The day before all the liquid is used up, start a fresh mixture of dolomite with vinegar. You can use apple cider vinegar, preferably organic, but white vinegar is adequate for this purpose as well. It is also possible to use lemon juice instead of vinegar.

Table 1-3: Calcium/Phosphorus Ratio of Foods
You can minimize inflammations by selecting foods with a high ratio of calcium to phosphorus

	Ca/P ratio mg/100 g	Ca/P ratio mg/100 g	
Bone	36,700/16,400	Dried apricots	67/108
Kelp, dry	1200/300	Raisins	62/101
Molasses, blackstrap	684/84	Sweet potatoes	31/52
Sesame seeds	1160/616	Filberts	209/337
Carob flour	352/81	Cottage cheese	90/165
Turnip greens	246/58	Salmon	188/328
Kale	249/93	Beet root	16/33
Parsley	203/63	Cauliflower	25/56
Mustard greens	183/50	Almonds	234/504
Dandelion greens	187/66	Soybeans, dry	226/554
Endive	181/54	Pumpkin	21/44
Watercress	151/54	Capsicum	12/28
Beet greens	119/40	Spirulina	396/1023
Leaf lettuce	68/25	Buckwheat	114/282
Green barley powder	1100/590	Bananas	8/26
Grass juice	150/75	Beans, dry	135/460
Spinach	93/51	Jerusalem artichokes	14/78
Broccoli	103/78	Potatoes	8/53
Cheese	750/480	Herring	66/254
Goat’s milk	129/106	Eggs, whole	54/205
Cow’s milk	118/93	Egg yolk	141/569
Cabbage	67/54	Lentils, cooked	25/119
Celery	39/28	Pecans	73/289
Pineapple, raw	17/8	Walnuts, English	99/380
Turnips	39/30	Coconut, dried	26/187
Grapes	16/12	Peas, dried, cooked	11/89
Butter	20/16	Peanuts	69/401
Carrots	37/36	Brazil nuts	186/693
Tofu	128/126	Barley	16/189
Grapefruit, whole	16/16	Mushrooms	6/116
Cucumber	25/27	Sweet corn	3/111
Dates	59/63	Corn, dried	22/268
Lemon juice, apples	7/10	Rice, brown	37/292
Watermelon	7/10	Cashew nuts	38/373
Pears	8/11	Millet	20/311
Apricots, fresh	17/23	Sunflower seeds	120/837
Plums	12/18	Rye, wheat	37/380
Orange juice	11/17	Oats	53/405
Onions	27/36	Torula yeast	424/1713
Peaches	9/19	Brewer’s yeast	210/1753
Sardines	430/575	Bran, rice/wheat	100/1300
Eggplant	11/21	Wheat germ	72/1118
Peas, fresh	62/90	Pumpkin seeds	51/1144
Parsnip	50/77	Meat average	10/200
Tomatoes	13/27	Liver average	15/540

Dolomite supplies calcium and magnesium in the generally desirable ratio of 2:1. However, if you are already routinely using magnesium chloride or only additional calcium is required, then you can use powdered eggshell, powdered shell grit, or commercial calcium carbonate for neutralizing. A 50 ml amount of neutralized vinegar carries about 800 mg of calcium into the body.

If we took calcium carbonate or dolomite powder instead of dissolving it, we would use up our precious supply of gastric acid to neutralize and dissolve these carbonates. Many individuals have mineral deficiencies precisely because of a lack of gastric acid. If, on the other hand, we use an inorganic mineral supplement, such as calcium or magnesium chloride, then we can absorb the calcium or magnesium, but it will not reduce any overacid condition of the body.

The Calcium-Phosphorus Ratio: In a person of good health, the ratio of calcium to phosphorus in the blood is 10:4. If there is a glandular imbalance, especially in regard to the parathyroid glands (in the neck), then this ratio will be maintained at a different level, causing long-term health deterioration. In particular, a high ratio of phosphorus to calcium sensitizes the body and increases acidity and inflammatory tendencies.

In addition to this regulation by the parathyroid glands, the calcium-phosphorus ratio is affected by our food choices. If we consistently eat foods high in phosphorus and low in calcium, such as meat, oats, wheat, and corn, then this tends to make the body overacid, depletes it of calcium and other minerals, and increases the tendency towards inflammation. By selecting suitable foods, these effects can be minimized. You can judge from Table 1-2 whether you might have an imbalanced -calcium-phosphorus ratio in your blood or your food selections. A low phosphorus intake is especially helpful with cancer and leukaemia.

As you can see from the foods listed in Table 1-3, it is easy to obtain sufficient phosphorus. In fact, if the blood calcium level is high, more magnesium rather than phosphorus is generally required.

You may wonder how the body can maintain the blood ratio in favour of calcium despite its relative scarcity in our foods. The reason is that normally about 1 g of phosphorus but only 150 mg of calcium are expelled daily with the urine. In metabolic imbalance, this excretion ratio is changed. In addition, using more phosphorus leads to the excretion of more calcium together with the excess phosphorus.

An excess of up to three to one in favour of phosphorus might still be regarded as metabolically neutral. In addition to the ratio, the actual amount of the excessive mineral is important. Therefore, to an extent the generally recommended daily calcium allowance is meaningless if it is not geared to the levels of phosphorus and magnesium. With low blood pressure, your calcium intake should be about double the intake of magnesium. However, with high blood pressure it is better to have equal amounts of magnesium and calcium or sometimes even to have more magnesium. The higher the phosphorus intake, the higher your combined calcium and magnesium intake should be.

* * * * * * * * * * *

Step 7

LIVING WATER

Living water has a great healing potential - learn how to use it.

Water, air, and food are the three main essentials to sustain life. For remaining healthy, these should be as clean and life-enhancing as possible. All three are highly polluted in our present environment, and we need to make an extra effort to improve them. Food is the most difficult aspect and is explained in later chapters. Individually, we cannot do much to improve our air, besides living in the countryside or buying an air purifier and ionizer. This leaves water in need of our attention.

Water has the ability to attract and accumulate bioenergy. Bioenergy or life-force energy is known in all traditional cultures under many different names, such as chi, ki, od, orgone, or prana. It can be seen or felt by many psychic or sensitive individuals. Water also stores an energy memory of harmful or helpful vibrations to which it was exposed in the past. This is the basis of homeopathy. Our chlorinated tap water is polluted with non-biological chemicals and negative energy imprints. It is also devoid of bioenergy and so may be regarded as being badly polluted as well as "dead."

Contaminated water can be filtered, distilled, or treated with reverse osmosis, but it remains dead unless it is energized to make it a "living" water again, one that can improve our health. While polluted and dead tap water can contribute to the deterioration of our health, living water is one of the greatest healers.

Ideally, healthy, living water should have the following properties:

- Purity: free of unhealthy chemical and biological contamination as well as negative vibrations or frequencies
- Minerals: well mineralized with calcium, magnesium, and trace minerals
- Low surface tension: due to the small cluster size of water molecules, making it a microwater
- Antioxidant activity: an abundance of negative hydrogen measurable as a negative oxidation-reduction potential (ORP)
- Vibrations: carrying positive, health-promoting vibrations
- Bioenergy: strongly charged with life force

All of these health-giving properties are naturally present in cascading, sunlit mountain streams. While it is easy to produce water meeting several of these requirements, there are some problems getting all of them in the same water. A main problem is the present lack of reliable measuring devices for bioenergy and the quality and strength of health-promoting vibrations.

Purity: This is easiest to achieve with readily available filters, distillers, and reverse osmosis. I have no strong preference for any particular method as long as you feel assured that your purified water is free of chlorine, fluoride, heavy metals, pesticides, microbes, as well as agricultural, industrial and pharmaceutical chemicals. However, most of these methods take out, in addition to the nasties, the beneficial minerals, which I recommend to put back into the water afterwards. However, there are also filters available that remove only pollutants and leave beneficial minerals in the water.

Normal water filters do not remove added fluoride but water ionizers, especially on high settings, channel fluoride into the acid water fraction so that the alkaline water becomes safe to drink. Also water high in calcium lowers the fluoride level by forming insoluble calcium fluoride.

Minerals: The minerals in our drinking and cooking water should supply a sizable part of our overall mineral intake. In living water, minerals are largely present in an energized form such as colloids and stabilized within small clusters of water molecules. I know of individuals who became badly demineralised from drinking distilled water over a long period; reverse osmosis water and rainwater (except if stored in concrete tanks) are commonly very low in minerals.

Using water low in minerals can be beneficial as a temporary measure for "calcified" individuals with high blood pressure, those who might otherwise use chelation therapy, but it is detrimental for anyone with low blood pressure. Overall, the most important bulk mineral commonly supplied with our drinking water is magnesium. Various studies show the beneficial effect of high magnesium levels on the rate of common diseases such as cancer, diabetes, and heart disease as well as bone and dental disorders.

Therefore, remineralise your purified water and add a teaspoon of hydrated or half a teaspoon of desiccated magnesium chloride to about ten litres of water. Alternatively, you can add a spoonful of sea-mineral liquid or bitterns, which is the brine left over after salt has crystallized out of seawater. In addition to being very high in magnesium, it contains all trace minerals. Instead of or in addition to magnesium, you can add an appropriate amount of commercially available colloidal minerals.

Low Surface Tension: This is a common by-product of energizing water. Water contains organized clusters of water molecules, also called liquid crystals, surrounded by large numbers of randomly moving single water molecules. In normal water, these clusters are larger and there are fewer of them. Energized or living water contains many more liquid crystals and of a smaller average size, forming cage-like structures of about six molecules. Due to this effect, energized water is also called "microwater."

The smaller clustering causes a lowering of the surface tension of the water. It makes the water "wetter" and more easily absorbed. It also becomes more like the water in fresh fruit and vegetables and within our cells. Colloidal minerals and fat-soluble nutrients can become suspended within the water cages and more easily absorbed during digestion. Also, various biochemicals, such as lecithin, detergents, soaps, and saponins, reduce surface tension, but this is not due to the formation of microwater.

Antioxidant Activity: The principle of antioxidant activity is the availability of electrons to neutralize any so-called free radicals, which have oxidizing qualities that can damage biological systems. Living water tends to contain highly reactive loosely bound electrons that react much faster than organic antioxidants to neutralize free radicals. Most of this antioxidant activity is assumed to come from an abundance of negative hydrogen, which is also present in fresh fruits and vegetables.

Such antioxidant-rich water is now commonly produced with water ionizers as electrolytic reduced water (ERW). Alternatively, you can buy capsules of Microhydrin. This is a commercial powder containing an abundance of negative hydrogen stabilized with colloidal silica; by dissolving a capsule in water, the water acquires a high antioxidant or reducing quality.

As we age, our body structures lose elasticity; everything becomes more rigid. On a biochemical level, this increasing rigidity is due to cross-linking of structural biochemicals, which in turn is due to a loss of electrons. Providing the body with an abundance of highly reactive electrons can be expected to slow cross-linking reactions and, with this, the aging process.

All biochemical energy in our body is produced by transferring electrons from food molecules onto inhaled oxygen. Thus, having more available electrons can help us to produce more energy. This may be the reason why some individuals feel more energetic on alkaline ionized or reduced water.

A water ionizer basically consists of a cell or container with two electrodes, which are separated by a membrane to minimize mixing of the water. The electrodes commonly consist of titanium, usually plated with platinum for better resistance to corrosion. When a current is applied, one electrode becomes positive, the anode, while the cathode becomes negative and releases electrons.

Positively charged ions, such as calcium and magnesium, accumulate on the side of the cathode. They form hydroxides and make this fraction of the water alkaline. Negative ions, such as chloride and fluoride, move to the anode and, in combination with positive hydrogen ions, make this part of the water more acid. At the same time, an abundance of electrons in the alkaline water leads to the formation of hydrogen gas and negative hydrogen ions, while oxygen and active oxidizing species are formed in the acid water.

In through-flow ionizers, the amount of reducing activity in the alkaline water depends mainly on the mineral content of the water, the flow rate, and the current flow or ionizer setting. With batch ionizers, it depends on the voltage of the power source, the size of the electrodes and distance from each other, the mineral content of the water, and the duration of ionizing.

Higher mineral concentrations are desirable as they require shorter ionizing times and produce greater pH and lower oxidation-reduction potential (ORP) changes. It is advisable to use water low in calcium, otherwise the cathode soon becomes coated with calcium deposits during long ionizing periods and loses its effectiveness or frequently needs to be cleaned. While the pH can be very high in alkalinity, it is in fact very weak, because the mineral density of ionized water is rather low and easily neutralized.

The negative charge of the alkaline water declines quickly with exposure to light and air and in contact with blank metal and soft plastics. However, in closed brown glass bottles filled to the top or in a cupboard, the water can keep for days and sometimes weeks.

Aim for a high ORP of between 500 and 800 mV, which can be measured with a commercially available ORP-meter. You can ask a distributor of water ionizers or a local water treatment or testing company to check the mineral level of your water and possibly the ORP of your alkaline water. Batch ionizers are more robust and easier to repair, but they are also more labour intensive; through-flow ionizers are more convenient.

Vibrations: This is an informal term based on the observation that water seems to retain a memory of various influences to which it has been exposed. There is a close connection between vibrations and energy: we may regard vibrations as an energy imprint on the water. Retained vibrations may be due to chemical, mechanical, electrical, or magnetic influences or even to feelings, emotions, and possibly thoughts. These retained memories in turn may influence the user of the water. This is the basis of homeopathy, which double-blind studies have shown to be therapeutically valid. However, it seems that only natural ingredients can be used to make homeopathic remedies, and some scientific tests for the memory of water and the effectiveness of homeopathy failed because synthetic chemicals that cannot be energized were used.

A method to assess the vibrational purity or degree of contamination of water was developed by researchers in biodynamics. If a drop of water is added to a concentrated mineral solution such as copper sulphate, it changes its crystallization pattern. The purer or more beneficial the water, the more harmonious is the crystallization pattern and vice versa. Polluted water causes a confused crystallization pattern.

Consider the work of Masaru Emoto. In his The Message from Water and on his website (www.adhikara.com/water.html), he shows in countless pictures how the crystallization pattern of freezing water changes when exposed to human thoughts and feelings. Some of these he did by typing a positive word such as “love” or a negative one on a piece of paper and sticking it on a glass of water. The positive word induced a beautiful and symmetrical crystallization pattern, while the negative word resulted in a chaotic or confused crystallization.

As our biological body contains large amounts of water (e.g., blood, lymph fluid, cells), you may now appreciate what negative thoughts and feelings are doing to your health. In a similar way, we can improve water before ingestion by blessing it with thoughts and feelings of love, joy, and peace. We may also hold it for a moment in front of the heart while doing this. For special healing effects, you can expose water to light in coloured glass containers or cover the container with coloured cellophane. For cancer therapy, green-treated water has an overall healing and balancing effect. Yellow and orange are energizing; blue is sedating and anti-inflammatory.

Another possibility is to expose water overnight to a strong magnet. Water exposed to the south-pointing pole is most suitable for pain, cancer, and infectious or inflammatory conditions. The north-pointing pole can be used to energize weak muscles and the digestive system, provided there is no inflammation or infection. You can also combine suitable magnet and color treatments.

Various energized waters are commercially available. While I expect these to be a great improvement on dead tap water, the energies and vibrations of uncontaminated natural ground, bore, or spring water are no doubt even more beneficial for us.

You can also keep water in a pyramid or immerse a suitable crystal or precious stone or pure gold or silver in it, possibly while the water is exposed to sunlight and/or moonlight. You can attach a magnet, crystal, or other energizing device to the water pipe to improve your shower or bath water or when filling a water bed.

Some good water filtering and energizing systems impart natural rock vibrations in combination with a vortex formation. A device can be fitted to a tap, which then produces running energized water, although the water may need to be chemically decontaminated beforehand. Another energizing system is a spiralling copper device through which the water flows.

A good way of energizing water is as follows: attach two magnets on opposite sides at the bottom of a funnel so that they face each other with unlike poles. Then pour water tangentially into the funnel so that it forms a vortex before running out into a collecting bottle.

Finally, you can use homeopathic potentising. Water and any active substance dissolved in it becomes energized by shaking it with a strong downward thrust. Use any uncontaminated water and shake it in a half-filled bottle or jar about 40 times. If possible, do the shaking outdoors, preferably on a dry, sunny day. Instead of potentising only water with its naturally dissolved ingredients, you can add a small amount of a live (unheated) substance that you believe may be good for you. Commonly used are herbal remedies, colloidal minerals, natural vitamins, or fresh vegetable juice.

Bioenergy: If we eat fresh raw food, we gain additional bioenergy. If we eat precooked or processed food and drink dead water, then our body loses bioenergy by charging the ingested food or drink from its own energy store. Rainwater is energized, especially during thunderstorms and if it comes as showers on sunny days. However, the maximum benefit is gained if the rainwater is collected and used without it touching metal. It appears that good quality rainwater has a beneficial negative electric charge (negative hydrogen ions), which is instantly neutralized by contact with metals. I recommend keeping energized water away from contact with blank metals, and I recommend the same for fresh raw food and especially fresh juices.

The simplest method for energizing water is to expose uncontaminated water to sunshine for several hours. However, a stronger charge can be achieved more quickly if the water is exposed to a source of bioenergy, such as sunshine, while it is in turbulent motion leading to vortex formations. High mineral content, especially colloidal silica, helps in retaining a high charge. Another recommended way is to leave drinking water for a while in contact with some freshly cut grass or leaves before using it.

The vortex principle is used in biodynamic agriculture for energizing liquid preparations. Use a large glass or ceramic container or a plastic bucket half-filled with uncontaminated water. Vortex energies can be absorbed into the water by first creating a deep vortex in one direction and then immediately reversing its direction until another deep vortex has formed. Stir for several minutes with a wooden handle or any suitable flat piece of wood or possibly plastic to create alternating vortexes. Electrical vortex machines are available from biodynamic sources for longer charging times.

Charging the water with vortex energy will be more successful when the air is high in bioenergy or orgone energy, such as outside on a sunny afternoon in natural surroundings. The energy is low during the night, on rainy days, in wet surroundings, and inside modern, tightly-sealed buildings.

You can also extract energy from volcanic rock dust or silica beach sand. Add a spoonful of rock dust to uncontaminated water and use the biodynamic vortex method before letting it settle overnight. Use the water while it is still milky from the suspended colloidal minerals. Most of the residue settles at the bottom and can eventually be used to fertilize plants. A stronger effect can be achieved by highly heating the rock dust or beach sand, preferably until it glows red, and then plunging it into water (in an enamel pot). Alternating several times of boiling the sand for 15 minutes (no microwave) and then freezing it is another method of liberating trapped orgone energy to energize water.

Another possibility is the use of flow forms, manufactured rock, or cement shapes in a series that simulate the action of mountain streams. From a reservoir, the water cascades through several levels of spiral formations, preferably while exposed to the sun. A recently developed energizing device is “orgonite,” a mixture of about equal parts of metal particles and resin; for more information on orgonite and bioenergy see Part 7.

Preferably do not store energized water near strong electric appliances. It also loses its beneficial energy if highly heated as in cooking. It is best to use energized water for drinking, to dilute juices, for enemas, or rubbing into the skin. Preferably avoid contact between energized water and metal. However, if you store drinking water for a longer time, you can add some colloidal silver or hydrogen peroxide to keep it free of moulds and other microbes or possibly put a piece of clean copper or a silver coin at the bottom of the container.

Our most common source of energized water is the water content of fresh raw fruit and vegetables. I expect the bioenergy of organic produce to be much higher than that of commercial produce. This is one more reason to eat your fruit and vegetables fresh, raw, and organic as often as conveniently possible.

* * * * * * * * * * * *

Step 8

THE PRACTICE OF HEALTHY LIVING

Improve your environment—your house, clothes, even your teeth.

Healthy living is the solid base for any health improvement. It includes all the habits that should become part of our daily lives. Our future well-being or diseases will arise from how we think, feel, and live habitually, not from any occasional transgressions of the rules of healthy living. The most important general principle of healthy living is to live in harmony - with ourselves, our social interactions, and with the forces of nature. In regard to health and natural living, this concept encompasses our environment, housing, clothing, workplace, exercise and nutrition, the water we drink, and the air we breathe.

In the following section, we will investigate various aspects of healthy living. It is important to realize that the concepts of healthy living often overlap with specific healing methods; for example, positive thinking, nutrition, reflexology, and physical exercises may all be used for normal healthy living as well as for healing yourself. Therefore, when reading this book, keep in mind that the separation between healthy living and healing methods is fluid, even somewhat artificial.

At present, most of us live in such an artificial and unnatural environment that it is not possible to remain healthy by living in the -commonly accepted way. Instead, we must become aware of the multitude of negative influences in our environment created by modern technology and try to minimize their harmful effects on us.

Improve the Physical Aspects of Your House: When building a new home, preferentially select building materials that do not shut out or distort the natural electric, magnetic, and life-force fields of the Earth. Therefore, minimize sheet metals and any steel or wire-mesh concrete reinforcements that form a cage-like structure (called a Faraday cage). Stone, brick, tiles, shingles, wood, plywood, and combinations of fibre and cement are acceptable for building. If you live permanently in a metal-clad dwelling, such as an aluminium caravan or mobile home, it is better to sleep in a non-metal annex.

Here are some practical tips:

- New copper or plastic water pipes may leach undesirable concentrations of heavy metals into the water. This is another reason, in addition to avoiding chlorine and fluoride, to use a good filter for your drinking and cooking water.
- Make your home a low-pollution sanctuary. Use natural materials for interior use whenever feasible, especially in the bedroom or wherever you spend most of your time. By far the best mattresses are those made of latex.
- Avoid cooking and heating inside with gas or kerosene except if the combustion gases are being efficiently vented. Many people are allergic to these. Wood stoves or solar heating are preferable wherever feasible.

- If you live in a modern house, try to sanitize it by minimizing the use of synthetic items that release or outgas chemicals, such as formaldehyde from glued material (many carpets, chipboard, and plywood).
- Avoid solvents and odorous household cleaners and disinfectants, tobacco smoke, mothballs, and insect sprays. If the outside air is reasonably clean, good ventilation helps to minimize the problem of indoor air pollution.
- New synthetic materials have a tendency to accumulate static electricity that can reduce our well-being in addition to releasing toxic chemicals. Synthetic materials (as well as television sets) also lower the amount of beneficial negative ions in the indoor air. Negative ions are energetic molecules with additional electrons in the air. Generally, the more negative ions there are in the air, the better we feel. Commercial air ionizers are available to improve room air.

Consider the Health Aspects of Your Clothes: Clothes are our “second skin,” and their material and color have a direct influence on our emotions and well-being. This applies especially to garments worn directly on the skin. Except for occasional short-term use, avoid underwear, including pantyhose and stockings, made of nylon or other synthetic material. They do not absorb moisture or “breathe” as well as natural fibres and appear to encourage fungal growth; sensitive individuals may experience skin irritations. Nylon seems to be worse in this regard than polyester. Synthetic clothing can also cause discomfort by accumulating static electricity. A more acceptable human-made fibre is rayon, as it is based on cellulose. The best material for underwear and stockings is silk; otherwise, use cotton.

Generally, select colours that balance your emotional state: wear bright colours when you are low in energy and blue when you are nervous and irritable. Avoid murky and dull gray colours and be extra careful with black; these can increase depressive tendencies in susceptible individuals. Commonly, the negative effect of black is balanced by combining it with red. You can use muscle testing as described in Part 2 to discover whether a certain material or color is suitable for you.

Shoes ideally should be made of natural materials with a leather sole to allow static electricity to flow off the body and into the ground. This is prevented by rubber and synthetic soles and by nylon socks and stockings. Walking barefoot for at least several minutes each day on moist grass or wet sand will discharge static electricity into the ground and improve well-being.

For women, it is advisable to reduce the time you wear a bra. A study reported by Singer and Grismaijer in Dressed to Kill found that women who wear a tight bra all the time had a 125 times higher rate of breast cancer than women not wearing a bra. Women who wore their bras 24 hours a day had a 75 percent chance of developing breast cancer. If they wore bras more than 12 hours daily but not to bed, the risk was one out of seven. Wearing a bra less than 12 hours per day improved the risk to one out of 152, a rate reasonably close to women who wear bras rarely or never; they have a one out of 168 chance of breast cancer. Women who never wear a bra have the same low breast cancer risk as men.

While this study was not adjusted for other factors, the difference in cancer rate is certainly large enough to be significant. A tight bra blocks the lymph circulation in the breasts and is much worse than a loose-fitting bra. If you need to wear a bra in public, use a loose-fitting one without underwires and minimise using it at home. Also do some exercises to stimulate the lymph flow in the breasts, such as circling your arms or massaging the area around the breast while doing inversion exercises.

Sanitize your Bedroom: Preferably use mattresses (futons) filled with natural fibres such as kapok, wool, horsehair or cotton; latex or natural rubber may be the best material. A cheap alternative is straw or hay from a pesticide-free source. Innerspring mattress may direct confused magnetic fields into the body, depending on how the spirals have been tempered. When moving a compass across the surface of the mattress the needle should not be deflected from its north-pointing position.

Synthetic foam rubber mattresses are not good either. They emit toxic gases that may be a major factor in cot death. Wrap them in clear polythene with only a small opening at the bottom. Use natural blankets and sheets on top of the mattress. Even sheepskin is chemically treated and can cause problems. Frequently expose pillows and featherbeds to the sun to kill moulds and dust-mites. This can greatly reduce asthma problems. Alternatively, use an airtight pillowcase.

Try not to use steel-framed beds or steel-spring frames that deflect the compass needle, move the bed if there are metal beams or live wires in the wall behind the bed. Featherbeds may contribute to arthritic problems due to static electricity between the feathers. Try to sleep with the head in a northerly direction or otherwise between north and east. Our brain waves are slowest when the head points towards magnetic north. Cover your head during cold nights to minimize heat loss.

Sleep in total darkness or cover your eyes during sleep, as production of beneficial melatonin is interrupted by exposure to electric light, including street lights, although moonlight and red light are harmless. Avoid even short exposure to electric light in the first part of the night.

Minimize Electromagnetic Pollution: Strong electromagnetic fields (EMF's) of about 50 to 60 cycles per second (hertz, or Hz) and the related electromagnetic radiation (EMR) may be harmful to us. Long-term exposure can aggravate any existing health problems or diseases and may cause or intensify them; this is especially true of lack of energy or fatigue, irritability, aggression, hyperactivity, sleep disorders, and emotional instability. Increasing numbers of individuals are becoming hypersensitive to EMR; many can feel the electricity going through their body and may experience disabling symptoms such as convulsions, memory problems, and depression. I have encountered many of these problems with patients.

Chronic exposure to high levels of EMR, especially while asleep, is a constant drain on our vitality. It creates chronic stress, which interferes with the regeneration and healing that normally take place during a good night's sleep. You may compare it to always swimming against a strong current, and this may well make the difference between recovering from a serious disease and succumbing to it. For more information on health problems due to electricity and many case reports, see: www.emrsafety.8m.net.

EMR exists around power lines, power tools, electric stoves, heaters, boilers, freezers, and television sets when in use, extending several feet or yards around the appliance. Stay away from them, if possible. Using an electric iron or an electric keyboard or working with handheld power tools can quickly drain our energies. When working with electrical equipment, we can reduce harmful effects by holding our hands under running water from time to time, in addition to having a shower in the evening or walking barefoot on grass.

Try to minimize electromagnetic pollution, especially while sleeping, when the pineal gland is most susceptible. Also sleep in the dark or at least cover your eyes to produce the immune-stimulating hormone melatonin. Before going to sleep, preferably switch off all power points in the bedroom and unplug all electric leads with two-prong plugs, especially any power points or leads close to the bed. This applies also to water beds, which need to be switched off or unplugged overnight. Three-prong plugs may remain connected to power points because this allows stray electricity to flow off through the grounding connection.

Increasingly governments insist that we get maximum exposure to harmful fluorescent lighting in the form of compact fluorescent light or CFL. The flickering, even if not perceptible, electromagnetic "noise", and the unnatural frequency distribution of CFL's frequently cause health problems for sensitive individuals. Most affected are those with vision problems, autism, migraines, traumatic brain injury, vestibular problems, multiple sclerosis, muscle spasms, epilepsy, tremors, cardiac arrhythmias, memory loss, diabetes, difficulty concentrating, and lupus. For details see www.cflimpact.com.

Here are some more tips:

- If your head faces a wall with power points or other electric wiring inside the wall close to the bed, then move the bed towards the middle of the room so that you can walk around the bed. The best sleeping position is with the head towards north or otherwise somewhere between north and east.
- When using an electric blanket, warm the bed beforehand and disconnect the power before you go to bed or better still, connect the blanket to a good earth. Do not habitually remain within a few meters of a working electric appliance, and minimize fluorescent lighting, watching television, using video games, computers, and even electric typewriters and handheld electric tools.
- Stray currents and radiating fields can still be emitted from electric wires even if appliances are switched off. AC electric fields do not disappear when an appliance is switched off, only AC magnetic fields disappear. Keep live wires away from your body. Preferably do not live near high-voltage lines, microwave towers, or electric-train tracks. Alternatively, you may take extra precautions to shield your house from the effects of their energies.
- Old fashioned TV sets and computer monitors emit harmful X-rays. Preferably sit as far away as conveniently possible. Flat or liquid crystal screens are much safer. An increasing hazard is microwave radiation from ovens, mobile phones, cordless phones and keyboards and other equipment. Most dangerous are long-reaching microwave stations inside the house for phones and computers, or equipment such as a mouse or keyboard that are not connected by wires, but USB modems for wireless broadband are harmless.

Electro-Proofing Your House: The measures outlined so far will go some way towards reducing the harmful effects of EMR on your health. However, often they are not enough. There is a simple way to check whether your home and workplace are sufficiently safe. You can do this by measuring the electricity that is absorbed by your body. For this you need a digital voltmeter that can measure millivolts.

Get some insulated electric wire that is long enough to reach from most rooms to the ground outside and attach an alligator clip to each end. Then get a ground stake or earth spike (a copper pipe is good for this) and drive it into moist soil. One end of the wire is clipped to a clamp at the top of the ground stake and the other to the black probe of the voltmeter. It will not work to attach the earth wire to a water pipe or to the main electricity ground stake of the house as they may carry high voltages.

To close the circuit, hold the metal end of the red probe firmly in one hand. However, it is more convenient to hold a piece of copper pipe, which is connected through a clipped or soldered insulated wire to the red probe. Set the dial to 2 AC volt; if it shows less than 200 mV, that is good and you can switch the meter to the 200 mV setting. Preferably the reading should be below 100 mV in commonly used rooms, and on your bed, below 10 mV.

In reality the readings will be much higher and you may get a reading between 2 and 80 V while lying on your bed. This is a constant drain on your vitality. By moving the bed away from any electrical wiring, you can perhaps get it to read between 500 mV and 2 V. A relatively easy way of reducing the voltage a lot further is by using a grounded electric blanket.

The easiest way is to cut off the plug. Then you can easily solder or clip a ground wire onto all three wire strands. If you want to keep using the electric blanket for heating as well, then you can attach a plug to the ground wire. Lying on a well-earthed electric blanket can bring the body voltage down to between one-quarter and one-tenth of the previous measurement.

If the bed has a steel frame or innerspring mattress, then these may also need to be grounded. Commonly all springs are internally connected and need to be grounded only at one point, but some mattresses have individual springs, which would need to be grounded individually. However, steel frames and inner-spring mattresses can still deflect the compass needle when it is moved across the bed and that is not good either. It is healthier if the earth’s magnetic field is not distorted where we sleep. The needle should continue pointing north everywhere over the bed. Metal beams in the wall or ceiling can cause the same problem, and the bed should be moved far enough away from magnetic field disturbances.

Ground any metal parts, such as window frames, metal wall claddings, or roofing iron, that show an unacceptably high reading when touched with one hand while holding the grounded probe in the other. This is especially the case if there are high-voltage power lines, microwave towers, or transformers nearby or power lines that form a right angle at street corners.

Occasionally, there may be so much ground electricity that it shows a higher reading when you’re lying on a grounded electric blanket than on one without a ground connection. In this case, move the ground stake as far away as possible from any power sources that can cause this problem and find or create a good grounding behind the house where you can reach moist soil with a long ground stake. With a short ground stake, try to keep the soil around it moist at all times. It is also possible to get a good ground by connecting the wire to a moist part of a tree or large outside plant.

To get a true reading on a concrete floor or when outside the house, you should be insulated by wearing shoes; otherwise, the reading will be much too low.

For sanitizing an apartment or office in a large building where you cannot get a good ground, a ground connection to the water pipe may be sufficient, or you could get expert help. An upper floor of a building commonly has much higher voltages than the ground floor, as there is wiring in the floor as well as in the ceiling. In this case, you may be able to reduce the EMR considerably by placing grounded wire netting under the carpet. If this is not possible with the entire floor, then ground a smaller area where you usually sit.

In other cases, it may be advisable to shield and ground the wiring inside a wall. This is much more difficult, as wall panels need to be removed, but it may be relatively easy under the roof or if the wiring is under the house. You can simply wrap electric cables with grounded wire netting or aluminium foil or nail grounded metal strips over them.

You can also disconnect the fuse or switch in the meter box for the bedroom (which commonly includes all power points in the house) overnight. Alternatively, have a separate fuse switch installed for the bedroom only or a demand switch that supplies power to the circuit only when needed. In addition, strategically placed copper coils can deflect radiation and may be used to de-stress the whole house. Such coils are offered on the Internet.

To improve your well-being while driving, and avoid road-rage or fatigue earth the inside of your car by leading the electricity from metal frames of seats and large metal areas back to the negative pole of the battery.

Beware of Earth Rays: Some dowsers claim to have discovered harmful earth rays from fault lines and underground streams under the beds of most cancer victims. Earth rays may also aggravate many other diseases. Find someone who is able to check the location of your bed for harmful underground radiation. This can be a problem even if you sleep on an upper floor. Commonly, dowsing or muscle testing (kinesiology) are used. Geopathic disturbances (another term for harmful earth rays) can also be detected with a scientific instrument that indicates the degree of distortion of the local geomagnetic field.

If you cannot have your bed checked, then move it to a different location if you were sleeping there before a serious health problem developed. In addition, use a shielding device under the bed: place wire netting under the whole area of the bed and connect this to an electric ground. To protect against fault lines or ley lines, you can also cover the area under the bed with several centimetres of quartz sand and use additional copper coils; moving the bed is the safest option. You may be able to check the bed yourself with a shortwave radio, because over a fault line the reception deteriorates; especially dangerous are crossings of fault lines.

Table 1-4: Teeth–Organ Connections

Teeth are numbered from 1 to 8 starting at the midline of the upper and lower jaw.

1 and 2 upper and lower jaw	kidney, bladder, pineal gland (upper), adrenals (lower), frontal sinus, sacrum, coccyx, foot
3 upper and lower jaw	liver, gallbladder, hip, eye, pituitary (upper), gonads (lower)
4, 5 upper and 6, 7 lower jaw	lung, large intestine, shoulder, elbow, thymus (upper), arteries, veins (lower)
4, 5 lower and 6, 7 upper jaw	pancreas, spleen, stomach, breast, thyroid (upper), lymph system (lower), jaws, front of knee
8 upper and lower	heart, small intestines, shoulder, elbow, ear, nervous system, pituitary (upper)

Dental Problems: Another aspect of our environment is our mouths - our teeth, specifically. A major factor causing or contributing to poor health and many diseases is our teeth. Frequently, there are mercury amalgam fillings or different types of metal in the mouth that act like an electric battery, and dead teeth with filled root canals or inflammation inside the jawbone (cavitations) even after removal of teeth are also a serious health problem. These problems may not cause any local pain and so are difficult to detect, but they can cause health problems in other parts of the body. Through acupuncture meridians as well as other mechanisms, each tooth is connected with a specific organ or other part of the body as shown in Table 1-4, so health consequences of problem teeth can occur throughout the body.

Heart disease, for instance, is commonly connected with inflammations or infections in the wisdom tooth positions, while kidney problems are related to the front teeth. This has been known for a long time. A 4,000-year-old papyrus describes the dialog between the pharaoh and his physician. The pharaoh complains of severe arthritis, and his doctor replies that this is due to the bad condition of some teeth that need to be removed. Even orthodox medicine is now slowly catching up and realizing that heart disease is frequently associated with microbes originating from infected teeth, gums, or jaw bones.

Root-canalled teeth appear to be a major contributing factor in many health problems, not only cancer but also heart disease, kidney disease, and autoimmune diseases. This is due to microbes that multiply in the multitude of tiny canals or tubuli in the dentine and gradually leach out into the lymph system. Even normally harmless microbes become very dangerous under the anaerobic conditions in dead teeth.

Weston Price, a dentist and former director of research for the American Dental Association some years ago, observed that the removal of root-canalled teeth from patients with kidney or heart disease would in most cases lead to an improvement in those conditions. When he inserted a removed root-canalled tooth (human) under the skin of a rabbit, the animal would die within two days; when he implanted normal teeth, there was no adverse health effect. In some experiments, he implanted the same fragments of root-canalled teeth in succession under the skins of up to 100 rabbits and they all died within two weeks of the same disease that the human donor had!

The safest solution with dead (root-canalled) teeth is immediate removal, but problems might still persist if an infection has spread to the jawbone. The microbes in dead teeth and infected jawbones apparently cannot be eliminated with pharmaceutical or natural antibiotics. Removal of dead teeth is especially urgent if they are already dead for several years and show signs of deterioration, and also in the case of cancer, autoimmune diseases, and other serious diseases. If you cannot immediately have all your dead teeth removed, have some, and for several weeks after their removal, use natural antimicrobial therapy. You do not need to fill single tooth gaps where they cannot be seen. To fill the gap of a missing front tooth, find a dentist who can cement an artificial tooth in this place without resorting to crowns or other metal support.

A possibly safe way of keeping root-canalled teeth is to fill them with Calcium Hydroxide or Iodine instead of the commonly used Gutta Percha, but no long-term studies on health outcomes are yet available. For tooth implants Titanium is commonly used but this is a metal and can still sometimes cause problems. A safer alternative with good durability is Zirconia, the dioxide of zirconium.

Try to have amalgam fillings replaced with plastic composite, preferably by a holistic dentist. A rubber dam and suction should be used when removing old fillings. If you cannot pay for proper replacements, just have them sealed with temporary filler and do not chew any hard items. However, with serious conditions and large fillings, extraction of amalgam-filled teeth can bring much better results than simple replacement.

In a German study of multiple sclerosis patients, extraction of compromised teeth resulted in an 85 percent recovery rate versus only 16 percent for filling replacement alone. Other studies have found that recovery from serious autoimmune diseases, dementia, or cancer may require more aggressive mercury removal techniques than simple filling replacement, due to the body burden of mercury. This appears to be due to the migration of mercury not eliminated by simple filling replacement into the roots and gums. Mercury in the teeth and gums has direct routes to the brain and central nervous system.[3]

Try to find a dentist who is familiar with safe amalgam removal procedures as recommended by holistic dentistry associations. Remove mercury by using high doses of vitamin C, Alpha Lipoic Acid, L-cysteine, MSM and chlorella before and after amalgam removal. Preferably, also use homeopathic mercury 1M and higher potencies.

Injections of the mercury chelator DMPS can be administered by qualified health professionals. However, this procedure has occasionally lead to complications and the oral DMSA is less expensive and apparently safer. Nevertheless, both of these chelators can activate very high mercury levels from the bones and cause serious problems, especially in sensitive individuals and neurological diseases. Therefore, I generally prefer the slower but safer method of using careful cleansing diets together with detoxifying natural supplements.

The pink color of dentures may be due to heavy metals, possibly mercury or cadmium. If you cannot get a guarantee that the colouring is free of heavy metals, it is safer to ask for clear plastic dentures made of methyl methacrylate or flexite for partial dentures. Use plastic denture teeth rather than porcelain teeth that have a metal base. Bridges and metal partials should be changed to clear plastic partial dentures. Nickel, as in stainless steel, can suppress the immune system and is generally classified as carcinogenic.

Even dental gold can be a problem because it is commonly blended with 20 percent of the more harmful palladium. A small amount of gold as for one crown may be acceptable, but as a general rule keep your mouth free of all metals and make sure only metal-free plastics are used as replacements. Nevertheless, any metal in the mouth can lead to allergies, and on average about 30 percent of individuals are allergic to the dental metal in their mouth.

After removal of bad or dead teeth, health problems can still arise or continue from root or metal fragments or due to chronic infection and inflammation in the jawbone. An experienced dentist can discover this with panoramic X-rays. If you cannot do all of this to sanitize the inside of your mouth, just do the best you can and do not worry. Compensate for any shortcomings here with a better diet and positive thinking.

* * * * * * * * * * *

Step 9

EXERCISE

Get into the habit of exercising regularly.

Adequate physical activity is one of the main requirements for achieving and maintaining good health. Deep breathing while exercising is essential: Lack of oxygen - becoming breathless - is harmful. You have probably heard of aerobics. Aerobic exercises are those that can be maintained for long periods because the heart and lungs can supply enough oxygen to the muscles. If insufficient oxygen reaches the muscles, the exercise becomes anaerobic and causes a harmful buildup of lactic acid in the tissues; this is commonly noticeable as sore muscles. Therefore, when exercising, breathe more vigorously than required for the activity at hand; increase the intensity of your exercises only gradually.

Arm and leg muscles should be regularly exercised. Supplement cycling, walking, and jogging with push-ups; swimming in unchlorinated water is also good for this. When jogging without running shoes, touch down with the ball of the foot and use your ankle joints as cushions. By using these springs freely provided by nature you cause less damage to your joints.

Everyday Physical Exercises: Depending on your body needs and inclinations, you can perform many different exercises. The more important ones are stretching, hanging from the hands or feet, using a rebounder, circling the pelvis, head-and--shoulder stand, isometric exercises, scalp and face exercises (pulling faces, tensing and relaxing scalp muscles), shaking to loosen the body, and slow graceful and rhythmic movements with (or without) music. Hatha yoga exercises are excellent for stretching and massaging the internal organs.

During the day, frequently stretch, tense, circle, or shake your muscles. Select a time of the day or week for a more thorough workout. With advancing age, stretching your muscles becomes increasingly more important. Lie on a large bed or carpeted floor and move your limbs and spine in various positions to the limits of their range, and then stretch them a bit farther. Try to let your body move on its own without much mental direction. Do what feels good.

Rejuvenation Exercises: The following five exercises were first presented in Ancient Secret of the Fountain of Youth by Peter Kelder. They are effective for strengthening glandular activities and may help to rejuvenate the body if practiced faithfully every day. Deep breathing during and between exercises is important. Inhale and exhale during the slow movements between the two end-positions as shown in Figure1-1.

Hold your breath in each end-position and tense all your muscles with maximum effort. In one position, press the chin to the chest and pull the abdomen inwards and up; in the other position, drop the head back as far as possible.

Pull your stomach in with your exhale in the lying position (Fig. 1a); in the forward bending position (Fig. 1b); while sitting (Fig. 1c); and with the pelvis raised

(Fig. 1d). Start with a few exercises daily and gradually increase to 20 of each, but do not strain yourself.

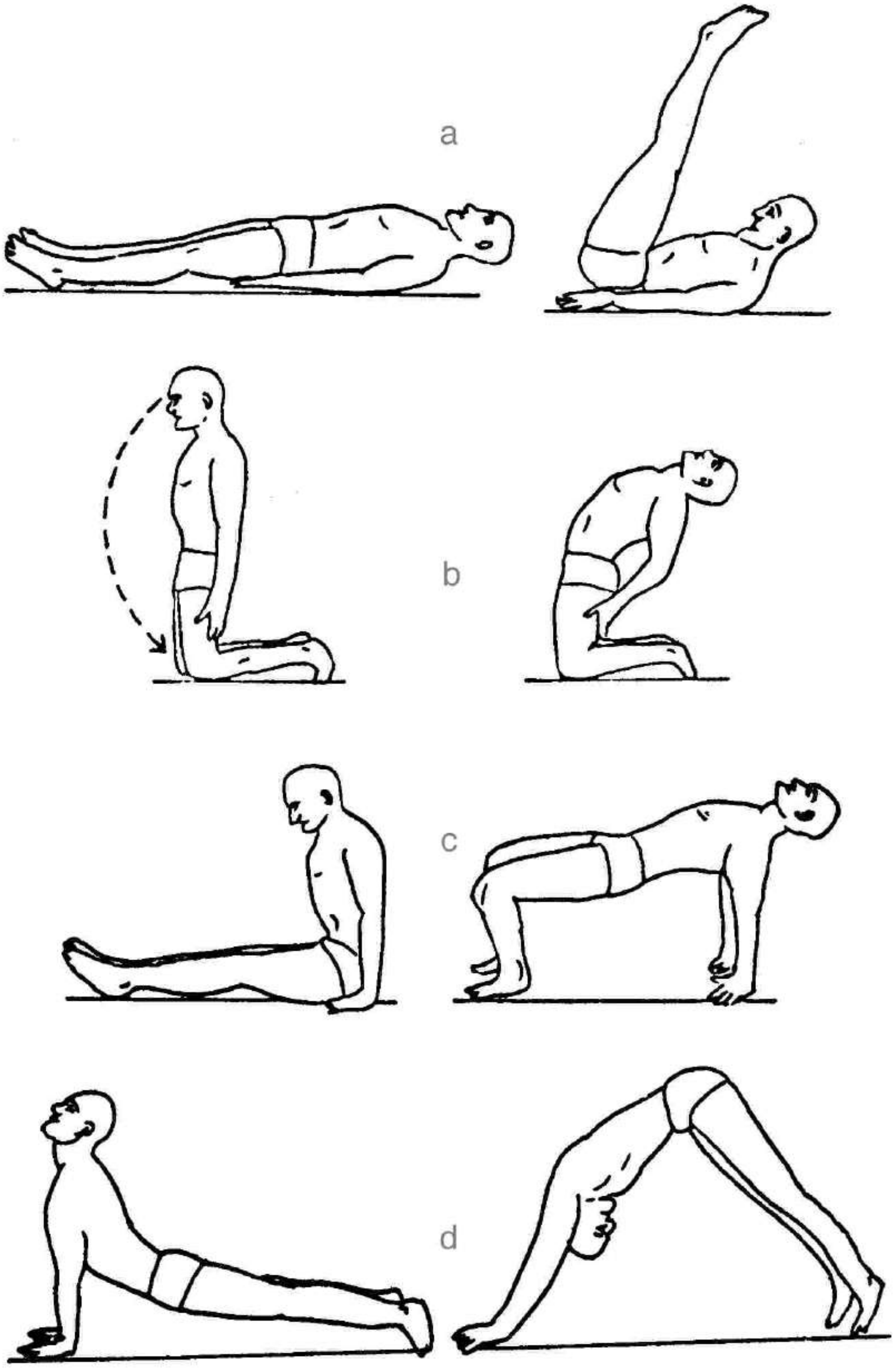

Figure 1-1: Rejuvenation Exercises

For a fifth rejuvenation exercise, spin or whirl 20 times with horizontally outstretched arms. In the Northern Hemisphere, the instructions are to spin clockwise; I have the impression that in the Southern Hemisphere anticlockwise spinning is preferable. However, you may choose to spin in whichever direction makes you feel better; use muscle testing to see what is best for you.

Inversion Exercises: Hanging upside down by your feet strengthens the internal organs, drains lymph fluid from the lower part of the body, and flushes the brain and face with blood. This exercise is especially good to delay or reverse symptoms of aging. While in the inverted position, rhythmically tense and relax the muscles of the face and neck and then the whole body. You can use inversion equipment that lets you hang straight upside down, or a slant board that puts you at a 30 to 45 degree angle. If you are sufficiently athletic, you can also do a shoulder stand, headstand, or handstand against a wall. You can do bicycling movements while in the inverted position.

Caution: Do not invert yourself if your blood pressure is high and requires medical attention.

Muscle Tensing: Tense and release one muscle after another, starting with one foot, then the other, working up towards the top of the head. Inhale deeply, each time gradually increasing the tension; hold for a few seconds, then slowly release the tension while exhaling. Finally, tense all your muscles, starting with the feet and moving up to your neck; simultaneously inhale deeply. Hold the whole body tightly tensed for several seconds; then gradually release the muscles, starting with the neck, and moving down to the feet, exhaling simultaneously.

If a particular organ or limb is weak, focus on it with repeated cycles of tension and relaxation. Frequent tensing is especially important if you have to stay in bed for a prolonged period. Do this exercise repeatedly during the day, whenever you remember it. Tensing may be done in any position such as sitting, standing, or lying down. It is easy to do while sitting at a desk all day: just take a deep breath, hold it for a few seconds while tensing the whole body simultaneously, then relax during exhalation.

Head and Neck Exercise: Slowly, but firmly, bend your head three times in each direction, forward, backward, and sideways; try to let the head drop to each side. Then move the head in circles, again slowly and firmly, three times in each direction. This is excellent for strengthening the neck and is helpful in cases of head congestion, eye and ear problems, and recurring headaches. In addition, you can circle the head up to 100 times in each direction.

Body-Mind Exercise: Slowly move a stiff joint several times to the limit of its range. Then imagine the same movement but with an extended range; repeat this several times. Then move the actual limb again and see whether its range has improved. The actual movements are easier if performed during slow exhalation.

Rebounding and Lymphacising: Rebounding may be better than jogging, especially for those low in energy. You could even bounce on top of one or two innerspring mattresses.

While you may develop a routine with various bounces, you can also use a lymphacising program instead of this or in addition to it. This is designed to greatly speed up the lymph circulation while energizing your body at the same time. While lymphacising, remain with both feet firmly on the rebounder mat. Bounce gently up and down, starting with just a few minutes and gradually increasing up to 10 or 15 minutes, three or four times daily; it is best done before meals and bedtime, and the more often the better.

Keep your hands touching body areas in need of healing, such as a tumour area, liver, kidneys, or wherever you feel discomfort. With each upward bounce, take a sniff until your lungs are filled after ten to 20 sniffs, and then exhale again in sniffs. Always try to fill and empty your lungs as completely as possible. If you do it right, you should feel energized after the exercise.

Practicing "Good Use": F. M. Alexander, the Australian formulator of the Alexander Technique, developed the principle of "good use." It means using your body efficiently and with ease, maintaining maximum balance and coordination between all of its parts. Most of us use our body poorly most of the time. By learning good use, we can harmonize our posture and movements, and with this improve our spine, muscular structure, even the functioning of our internal organs and glands. To learn good use, we concentrate mainly on the way we sit, stand up, sit down, walk, and breathe.

Start by exploring your present use in front of a mirror: Observe how you move your head from side to side, up and down, how the shoulders move, how you talk and breathe. Then try to practice the basic movements of good use: Move your head up and away from the torso, but at the same time let the whole body ease upwards in a movement that flattens and lengthens the spine.

When turning the head, moving it up or down, standing, sitting, walking, or lying down, always let the head "float up," which means away from the body and in line with the spine. This is not just a position to be maintained but part of a continuous process of easing the head up to lengthen the spine.

When you look at yourself at eye level, the neck should be straight and the chin down. Watch that the back of your head is not dropped backward where it will produce a hollow where the head joins the neck. Except when intentionally tilting the head, there should be no hollow at this point. Watch that your head remain up and your neck straight while you talk, sit down, stand up, or move in any other way. Instead of directly moving your head upwards, think and feel it floating up. Let the move be manifested through this thought and feeling.

Standing and Sitting: Another important aspect to watch is that your muscles be relaxed when at rest. When standing, feel that your eyes, jaw, shoulders, and abdomen are relaxed; when sitting, do the same for the thigh and calf muscles. Check repeatedly because these muscles have a tendency to tense when we do not watch them. Do the same when lying down: Check that all your muscles are relaxed.

When sitting down or standing up, your back and neck should remain in a straight line. Only the joints at the ankles, knees, and hips should bend. When standing up, lean forward by having your body follow your head. In this way, bend at the hips, which creates a momentum that lifts you off the seat and brings you onto your feet without any effort. Now you only need to straighten the three lower joints and you can stand in front of the chair.

Exactly the reverse procedure takes place when sitting down. The arms are not used in any way to push or swing when you're standing up or sitting down. When sitting, keep your head moving upwards in order to counteract the tendency to slump. Check yourself frequently to be sure you sit straight but relaxed. When lifting something from the floor, keep your back and neck straight and bend only at the ankle, knee, and hip joints.

Walking: When walking, touch down lightly with the full foot, not just with the heel. While the other foot is brought forward, the whole body weight now rests on the foot on the ground. Start rolling your weight off this foot by lifting your heel. In this way rise onto the ball of the foot and even onto the toes. This creates a strong momentum that propels you forward without any effort. Repeat the same process with the other foot.

The important part of walking is rising onto the toes, which gives you a light, bouncing step, as if you're walking on air. Your spine and neck remain in a straight line, your shoulders relaxed. The arms swing gently, as opposite arm and leg move at the same time. When climbing steps, rise onto your toes in the same way as for walking. (For more information, see "The Complete Guide to the Alexander Technique" and other information at www.alexandertechnique.com.)

Correct Breathing: Generally, we breathe just enough for the activity at hand. However, to prepare for action it is beneficial to supercharge your body with oxygen by breathing more deeply than you need. This makes you more energetic, helps to cleanse the body, and neutralizes overacid conditions. Practice deep breathing in fresh air. Watch that during normal breathing your lower rib cage is moving. However, during sedentary activities it is preferable to breathe as slowly and deeply as possible. This helps to relax both body and mind. Therefore, by deliberately over-breathing we create useful tension, and by under-breathing we practice relaxation. Here are some tips on breathing exercises:

Watching the Breath: This is a relaxing exercise to still the mind. Mentally follow the breath flowing in and out but without any attempt to control it.

Full Breath: Inhale for several seconds, consciously filling first the lowest part of the lungs, causing the abdomen to lift. Then expand the middle part and finally the upper part of the chest. Hold the breath for several seconds then slowly release it, starting with the upper chest; end by contracting the abdomen. After about four seconds, start a new cycle.

Abdominal Breathing: This is important for people with breathing problems. Place a hand on the abdomen and feel it rise and fall without your chest moving. Adopt abdominal breathing whenever an acute breathing problem develops.

Long Exhaling: This is good for people with obvious or hidden anxieties and for those who need to establish a proper orgasm reflex. While standing, exhale for about ten seconds in one continuous sweep. Near the end, the shoulders move forward, the buttocks press together, and the pelvis moves forward and upward in an involuntary movement. A long, gentle exhalation also helps to relax body and mind and may be used to direct energies and relax specific muscles. Inhale quickly in about one or two seconds and then exhale slowly for about ten seconds. During exhalation, focus your attention on a specific part of the body and imagine the breath energy flowing into it, making it feel warm and expansive. Use this to reduce muscle tension and pain or to relax the whole body, starting with the feet and gradually working your way upwards to the head.

Super-breathing (Hyperventilation): This is to quickly saturate the body with oxygen. Quickly and deeply—at the rate of about one or two cycles per second—inhale through the nose and exhale through the mouth. Stop when you begin to feel dizzy. A slower, but more sustained form of super-breathing, done through the nose only, can be adopted during energy healing, or before and during heavy physical work or exercises.

Hara Breathing: Sit or lie, relaxed, with the hands folded as in prayer or over the abdomen, left palm on the navel, right hand on top. Inhale gently through the nose, hold the breath, and exhale through the mouth; each period lasts five seconds. While inhaling, touch the roof of your mouth with your tongue. Accompany this with abdominal breathing. Envisage with each breath that life energy fills the abdomen; feel it as warm and tingling. Hara is a Japanese word for the centre and storehouse of vital energy located just below the navel.

Alternate Nostril Breathing: This is helpful for balancing body polarities. Close the right nostril with a finger and inhale through the left nostril, then close the left nostril and exhale through the right nostril. Inhale through the right nostril, and exhale through the left one, and so on. Several rounds of this are good at the start of meditation.

Eight–Two Breathing: This may be used for stilling the mind and emotions during meditation or spiritual healing. Inhale for a count of eight, hold for two, exhale for eight, hold for two, and so on. The flow of the breath should be smooth and harmonious.

Skin and Organ Breathing: For improving a particular part of your body, imagine and feel your breath moving through this part during inhalation and exhalation. As an example, for improving the liver, feel the breath as a cool stream penetrating your liver. When holding the breath, feel a slight pressure in the liver area and an increasing warmth.

Vocal Exercises: Sounds have a strong influence on us, and music particularly can have a profound effect on our emotions. Mantras (special words or sounds) are used in certain techniques of meditation to achieve particular effects. Much stronger is the impact of our own voice on our body. If properly pronounced, each vowel produces a vibration in a specific body part, mainly the large cavities. If you cannot feel a resonance vibration, just imagine it.

Produce the sounds as clearly as possible. Take a full breath and sing each vowel forcefully in different intonations. If your voice is low and deep, try to produce the sounds in the front of the mouth; if you have a high-pitched voice, make the vowels vibrate deeper down the throat.

Feel the resonance of each vowel in the indicated body cavity. The lower the body cavity in which you want to experience the sound, the deeper or lower you should try to make the sound. Initially, use each vowel in turn, then concentrate on the ones most appropriate to your condition. Select and use vowels according to resonance areas. A high-pitched “I” sound may be used to stimulate the brain as well as the pineal and pituitary glands. Imagine the sound vibrating from the centre of your brain.

Vowel Resonance Areas

This table shows the correct intonation of vowels and the body cavities, which they stimulate if properly performed. Try to grade the vowels with the highest note for "I" and the lowest for "U"

I	Bee	Head
E	Fairy	Throat
A	Father	Upper Chest
O	Coma	Lower Chest, Upper Abdomen
U	Zulu	Lower Abdomen

If you make the sounds properly, you will feel the vibrations in your throat and mouth and possibly in your lips and cheeks as well. When you leave a tiny gap between your teeth, you should be able to feel them vibrating strongly. This effect is usually strongest with the "m" or "om" sound. Also try to vibrate your tongue with a rolling "r" sound. If correctly done, you can feel these vibrations energizing the whole structure of the head, bones as well as tissue.

Instead of sounding "om," you can intonate "a-u-m" or "a-o-u-m" in sequence with the same breath. Gradually lower the sound, and try to feel the vowels vibrating in the appropriate resonance areas. Feel the "m" sound moving up the spine with the pitch gradually rising higher, ending at the centre of the brain.

Singing is an excellent vocal therapy, so try it often. Listen to stimulating, sedating, or uplifting music according to your needs. Lie down in a relaxed position and become fully immersed in the sounds. Feel something inside you moving, expanding, rising in resonance with the music. Try singing along.

* * * * * * * * * * *

Step10

THE HEALING CRISIS

Sooner or later you are bound to have a healing crisis, so use it to your advantage.

Even on their first clinical visit, I routinely recommended that my patients adopt a high-quality, low-allergen diet. When they came back two or three weeks later, they often said they felt much better but that a week earlier they had had a cold. When asked about the symptoms, these turned out to be mainly a profuse mucus discharge or sometimes diarrhoea, but rarely were there signs of infection. These patients had experienced their first healing crisis on their path to better health.

This shows us the general principles of a healing crisis. With each crisis or reaction, there is a temporary worsening of the condition, with either new symptoms appearing or existing ones getting worse. This is the result of toxins being mobilized for elimination by the body in what is clinically known as the Herxheimer effect. However, after the healing reaction, we reach a higher level of health than before.

This concept of a healing crisis clearly shows the opposite perceptions in drug-based medicine and holistic medicine regarding health and the healing process. In drug medicine, it is assumed that a patient who is free of disease symptoms is more or less healthy, and the aim of drugs is to achieve this condition by removing disagreeable symptoms. Frequently, alternative medicine is used in the same way; instead of toxic drugs, benign natural remedies are used. This is what most patients want, and according to their beliefs, they use either drugs or natural remedies to eliminate symptoms.

However, holistic medicine, following in the footsteps of the nineteenth-century Nature Cure movement, aims higher. Here, health is regarded not just as a temporary absence of disease symptoms, but as a state of profound physical, emotional, and mental well-being so that we cannot develop a disease.

Animals living in an unspoiled natural habitat commonly display this kind of health. If we want to come close to such outstanding health, we have to work for it by consciously minimizing the multitude of negative influences on our health and by maximizing the positive factors instead.

Experience shows we do not follow a straight line of health improvement or deterioration. Similar to the way periods of illness are interspersed amidst relative well-being on the long road to chronic degenerative diseases and death, we have ups and downs on the road to superior health. The main difference is that the road to deteriorating health slopes downhill while the road to superior health goes up.

Travelling downhill is easy; we do not need to do anything about it, but improving our health requires consistent effort, and it can seem like an uphill climb. Contrary to the lengthy periods of ill health on the downhill road, the dips on the uphill road are usually short and sharp. Such a dip is called a “healing crisis,” although I prefer the less dramatic names “healing reaction,” “cleansing reaction,” or simply “reaction.”

After each reaction, we advance to a higher level of health than we had before the reaction. Holistic therapists realize that the road to better health follows a definite pattern, once known as **Hering's Law of Cure** (from homeopathy), which says: "All cure starts from within out, from the head down, and in reverse order as the symptoms have appeared."

The emotional side of our health problems is important as well. Each disease, accident, or surgical intervention contains a strong emotional component that needs to be re-experienced during a reaction. Otherwise, the healing will remain incomplete and the problem will present itself again at a later time for healing at a deeper level.

This also means that the body selects the kind of healing crisis that is most appropriate at the time, taking into consideration its needs and abilities to have a certain area healed or improved. We can consciously influence this choice by working on a particular problem. I've noticed that the "body self" tries to select a timing that does not disable us during important upcoming events. No doubt we are guided on our healing path by our inner intelligence that has our best interests at heart.

In the beginning, our healing reactions will be mainly on the physical or biological level, but increasingly we will experience the release of emotional blocks and changes in consciousness, preparing us for greater activity on the spiritual level.

Patterns of Healing: In every healing, the sequence of our health deterioration will automatically be retraced. This means that old, long-forgotten disease symptoms may suddenly flare up again in the form of a healing crisis. The reason for this is based on the nature of disease.

At first, when the body is still vital, it reacts forcefully against pathogenic invaders; the resulting battle causes fever and inflammation. If the body is not strong enough to win completely, the acute battle symptoms will eventually subside, with the remaining invaders leading a somewhat subdued existence in a weakened gland or organ. Repeated attacks will result in the degeneration of this body part, causing chronic disease.

At some stage on the road towards health, the body will feel strong enough to start the fight again to heal the diseased organ; when it does, inflammation will result. Such healing reactions are usually short-lived but can be intense. If one supports the body during this period with rest and cleansing diets, then the organ will eventually acquire a much better level of health than before the flare-up. If, however, drugs are used to suppress the healing symptoms, then the inflamed organ will return to its subacute, chronic condition and wait for another opportunity to heal itself.

The healing pattern of other health problems that are not related to immunological processes and infections follows a similar line to this. Healing takes place on many different levels: at the biochemical within cells; in the blood circulation; in the nervous system; and especially in the flow of bioenergies (such as chi, as described by acupuncture). Healing activity will be increased on all of these levels.

At the bioenergetic level, the increased energy flow may cause congestion at various places, in the weak body part itself as well as in related acupuncture points and zone reflex areas. Excess energy in such places may be noticed as pain, heat, or tension - the common symptoms of inflammation. Eventually the body will clear the obstruction in the energy flow. We can support this process with specific healing methods or by healthy living practices in general. However, trying to suppress any acute inflammation may leave the body part permanently weak. Inflammation in itself is a healing energy release, and it is best overcome by removing the blockages in the energy flow.

Frequently there will be periods of greatly increased bowel activity or diarrhea. This is due to the detoxifying action of the liver. These toxic waste products are very irritating to the intestinal wall and the body tries to flush them out as quickly as possible to avoid intestinal discomfort and headaches. Do not try to suppress this process but you may try to ease it with frequent drinks of psyllium which help to bind the toxins and solidify the bowel content.

Skin Reactions: Another common form of healing reaction is the development of a rash or other skin problem, especially over a weak body part, such as, for instance, at a hip or shoulder. This skin reaction signifies that the energy that has been bound internally in this area in the form of muscle armouring, stiffness, and accumulated metabolic wastes is now surfacing and dissipating. Thus the affected joint or limb will become much freer and stronger. However, if the skin reaction is interfered with and made to go away prematurely, then the suppressed energy and metabolic wastes will remain in that body part, keeping it in a state of pain or weakness.

Thus there usually is no need to interfere directly with skin reactions or other healing symptoms. On the contrary, that may only drive the problem “underground” to affect internal organs and body structures. Natural healers have noticed that by artificially producing a rash over an arthritic joint, the joint greatly improves, but when a rash over a joint is artificially suppressed, then arthritic symptoms may appear in that joint.

This observation applies only to healing reactions, not to possibly infected or chronically inflamed skin areas which may be treated with antimicrobials and alkalizers in addition to nutritional therapy.

You may learn much about yourself, your body, health problems, and emotional reactions by patiently observing the healing process instead of rushing to do something about it. What you should do, however, is live healthily, cleanse the body, helping the suppressed energy to come out and dissipate, and be optimistic about it all.

Frequently, boils and increased mucus discharge, as in colds, arise during the initial stages of the healing process. Sometimes they start one or two weeks after you improve your diet and begin to take supplements. This indicates that your organs of elimination are congested. During such times, take little food, but instead consume generous amounts of herb teas, fresh vegetable juices, and pure water, and ideally go on a cleansing diet. Such healing crises may appear from time to time, separated by weeks, months, or even years, bringing you ever closer to your goal of superior health.

Listen to Your Body: Some further problems may develop during your program of health improvement. Formerly, the body response may have been dulled - it did not even react to harmful foods or drugs - but now the reaction is often immediate and forceful. Foods that you may have eaten habitually before going on a cleansing diet can now suddenly cause a gastrointestinal upset, making you feel sick and miserable for days. However, such reactions do not need to happen if you take proper precautions: Find out to which foods you are allergic and try to avoid these. If you eat in restaurants and do not know what to expect, take plenty of digestive enzymes and lecithin during the meal, and try to eat foods in correct combination; if allergy symptoms appear, use alkalizers together with sodium ascorbate to mitigate their effects.

Your body now trusts that you will not hurt it anymore in the old ways. But if you do, your body is very “offended” and lets you (your mind) know about it in no uncertain terms. This is the most effective language the body has. However, if you learn to listen, your body will first try to tell you about its requirements in a more subtle way such as in intestinal rumbling, an itch, a slight muscle weakness, a passing pain. There are numerous ways a body expresses its needs, and each body has its own language of symptoms.

If you are honest and persistent in your desire to become a “friend” to your body, it will eventually let you know its needs, its likes, and dislikes; you will rediscover your lost body instincts. This is what we should aim for: to become independent of outside expert advice and do what our body tells us is right to do, not only in nutrition, but in everyday living.

While one can usually gain relatively quick relief from distressing symptoms, rebuilding a strong and healthy body is a slow process. It requires a great deal of experimenting to find your optimal individual conditions for healing and it needs your solid determination to carry through with a program over the years.

The Progress of Healing: Generally, recently acquired chronic weaknesses are the first to flare up and become healed, while our oldest and most persistent health problems are the last to yield. Thus there appears to be an orderly sequence to our health improvement. Healing proceeds in a wave pattern on an upward slope. On the peak of a wave we feel much improved, while in a trough we experience cleansing activity or a healing reaction and feel sick. The overall movement, however, is towards greater well-being.

The body always tries to heal itself and an acute inflammation is an important tool for this purpose. An inflammation makes the walls of the blood vessels porous to allow an increased flow of lymph fluid to move immune cells in and move accumulated toxins out. At this stage it is just an acute inflammation and we can choose to turn it into a healing crisis by supporting the body mentally and nutritionally, or we can choose to suppress the acute condition and return ourselves to a chronic degenerative condition.

Therefore, anything that suppresses an acute inflammation prevents healing of the affected area. This includes not only anti-inflammatory drugs and remedies, but also strong alkalizers that reduce the release of histamine. The pain and heat in the inflamed area are the price we have to pay for healing.

However, anti-inflammatory agents and alkalizers have their role in easing the discomfort of a chronic inflammation until the body has been nutritionally and mentally prepared for a healing reaction. Further, the symptoms during a reaction may either be too severe or come at an inopportune time. You can lessen or temporarily stop them by adding cooked food to a raw food cleansing diet, by alkalizing the body, or by using anti-inflammatory remedies.

Usually, we can select a suitable time for a healing crisis by starting a strict cleanse on raw food only, especially fresh vegetable juices, or fruit or vegetable salads. A fruit cleanse is more suitable for insensitive individuals (less allergy-prone), while vegetables are much better for an overacid and sensitive body.

One of the main factors reducing the severity of any symptoms is a clean bowel, regardless of whether the reaction was triggered deliberately or involuntarily. The easiest way to achieve this is by taking a tablespoon of Epsom salt in water at bedtime or upon rising in the morning. If no solid food is used, then take a teaspoon of psyllium hulls several times daily in a glass of unchlorinated water; too little water can cause constipation (refer to Step 3 for more information).

You can further support your body with light, enjoyable outdoors activity. Prayer and meditation or guided imagery are also recommended.

As you can see, there is no simple path to completely avoiding health-related discomfort. We can only choose in which way we want to “suffer.” We can either deliberately and cheerfully endure unpleasant healing crises with the promise of better health to follow; or we can expect to suffer involuntarily and in an uncontrolled way from chronic degenerative diseases as we get older. The choice is yours.

* * * * * * * * * * * *

Part 2

NATURAL HEALING METHODS

This chapter contains a selection of natural healing methods, presented in a concise form suitable for do-it-yourself healing and experimentation. It is especially useful in the context of healing groups. You will find diagnostic methods followed by selected therapeutic methods. However, several of these overlap and can be used for both diagnosis and treatment. Experiment with those methods that appeal to you while disregarding others.

In conventional medicine, diagnosis is extremely important because the predominant use of surgery and toxic drugs directed against specific disease symptoms leaves little room for error. Holistic healing, on the other hand, aims to improve the overall health of the patient, so diagnosis is less important. Healthy living, which includes correct nutrition and mind control, in combination with various nonspecific healing methods, will be enough in many cases to lift one's health to such an extent that, even without diagnosis, diseases and their symptoms gradually disappear.

However, improvement may be faster and easier, especially in serious conditions, if you concentrate additional healing efforts on specific glands or organs. Here, diagnosis is -useful in pinpointing the areas most in need. Generally, the aim of diagnosis in holistic healing is to assess whether an organ, gland, or other part is overactive or underactive; whether energy flows are too strong or too weak; or whether there are deficiencies or harmful outside influences, attitudes, or emotions.

Worldwide, an emerging and especially active area of the healing arts is electro-medicine. This includes a multitude of diagnostic and therapeutic instruments from simple to very elaborate and computer-programmed. While some of these have a long track record, including mainstream medical use in Germany, the U.S. Food and Drug Administration (FDA) has either outlawed many of these products or actively discourages their use. In view of this, and also because of the evolving and constantly changing nature of this whole area, I recommend getting the newest information from the Internet. Look for key words such as Rife technology, BEFE (bio-electric field enhancer), electro-acupuncture, electrodermal screening, electronic blood purifier, zapper, and magnetic pulser. For the definitive publication in this area, see Bio-electromagnetic Medicine, with contributions by a wide range of scientists and medical researchers, and edited by Paul J. Rosch, clinical professor of Medicine and Psychiatry, New York Medical College, with Marko Markov of the Mount Sinai Medical School.

Here then is a selection of practical methods you can use to improve your health and that of your family and friends.

* * * * * * * * * * * *

Step 11

GENERAL EXAMINATION

Examine yourself to assess your state of health.

Testing Body Fluids: Tests in this step are recommended as a way to get to know and monitor your body rather than to replace medical diagnosis. Urine dipstick tests, available at most pharmacies, give indications of pH (acid-alkaline balance), kidney function (protein status), diabetes (glucose), fat metabolism (ketones), urinary tract infections (blood, nitrites), liver and gallbladder function, and blood stability (bilirubin, urobilinogen). Cloudiness in a fresh urine sample, for example, may be due to urates, phosphates, pus, dead bacteria, or inadequate liquid intake.

A dipstick test for vitamin C is available in some countries; otherwise you can test by adding ten drops of a five-percent silver-nitrate solution to ten drops of your urine in a test tube. Read it after two minutes. A white precipitate shows vitamin C deficiency; if beige, a small amount of vitamin C is present; if gray, your vitamin C status is good. During severe infections, sufficient vitamin C should be given to cause a charcoal-like appearance in the test fluid. From the colours of these dipsticks, you can clearly see whether the result is normal or whether there is a problem, in which case you can repeat the test from time to time or seek professional advice. (Caution: Silver nitrate badly stains skin and garments.)

Low-cost instruments are now available for measuring the level of glucose in the blood. In a standard glucose-tolerance test, 100 g of glucose, or 2 g per kg body weight, are given in water to the fasting patient. Blood and urine samples are tested every 30 minutes. A test for diabetes can be terminated after two or three hours, while for hypoglycaemia, test for five hours. Diabetes is suspected when the glucose level peaks and remains higher than normal.

Hypoglycaemia (low blood sugar) is more difficult to evaluate and may be suspected if the blood sugar curve shows a high plateau or remains flat, if it shows a steep drop, or if it falls below the fasting level. In addition, test the soft-palate pH (in the mouth) before and after the test. Greatly increased acidity shows inefficient use of glucose. Frequently, the blood sugar reaction depends on the source of the glucose or carbohydrate; that is, an abnormal blood glucose level may be caused by an allergy response to a particular carbohydrate.

Pulse and Blood Pressure: Take the pulse while sitting and standing. To find the pulse, lightly press with the pads of three fingers on the inside of the wrist of the other hand just below the pad of the thumb. Adrenal stress (predisposition to allergies) may be indicated if the rate is higher when standing than sitting. Statistically, the average pulse rate for men is 72 per minute and 80 for women. Physical activity, raised temperature, infections, an overactive thyroid gland, strong emotions, and allergies increase the pulse rate. An irregular pulse may indicate heart disease, potassium or magnesium deficiency in the heart muscle, or an allergy. Sometimes the pulse is irregular only under stress, for example, when standing or running. If there appears to be a problem, check again on subsequent days under different conditions.

Blood pressure is measured with a sphygmomanometer and is important for body balancing and assessing nutritional requirements. The normal readings are 120/80 mm Hg (millimetres of mercury). The upper value is produced when the heart contracts (systolic pressure) and mainly indicates the rigidity or elasticity of the main arteries. The lower reading (diastolic pressure) occurs when the heart dilates and indicates the amount of tension or relaxation in the small peripheral blood vessels. The difference between both readings is the pulse pressure. The normal pulse pressure is about 40. To obtain an indication of the metabolic rate (how fast or how slow the body produces energy from food) multiply the pulse pressure by the resting pulse rate. The average value is 2800 plus or minus 500. A value higher than 3300 indicates an overactive thyroid, while a result lower than 2300 indicates an underactive thyroid.

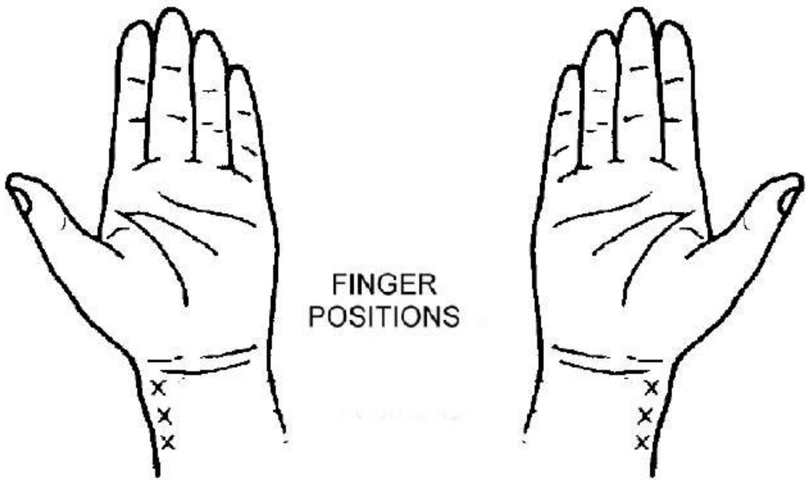

Fig. 2-1. Taking the pulse.

A fall in blood pressure when standing compared with sitting indicates adrenal stress. Sometimes a low blood pressure when sitting will rise when standing because of increased tension. The blood pressure of healthy adults does not rise with increasing age. The most common causes for high blood pressure are arterial hardening and congestion, kidney problems, and allergies.

Examination of Organs: While the person to be examined lies relaxed on the back, exert direct pressure on the organs to be tested. Inflamed or infected conditions will usually be indicated by tenderness. Of special importance is the liver; to gauge liver health, press gently under the front of the right rib cage. Press into the stomach, at the lower end of the sternum, and all around the abdomen, to check the small and large intestines. If most of these intestinal areas are tender, you can suspect an overgrowth of Candida albicans in the gastrointestinal tract.

Reflexes can be tested by striking with the edge of the hand just underneath the kneecap. Normally the leg will give a slight jerk. Absent or very strong jerks may indicate nervous system diseases and a need to seek professional help. Another test involves stroking the inside of the sole of the foot (called Babinski's reflex). In a normal condition, the toes will contract and draw together. In cases of central nervous system disease (and in newborn infants), the toes will fan out and be drawn back.

The thyroid gland is situated on both sides of the lower neck; if it is underactive, you may feel sluggish and have difficulty losing weight. To see whether your thyroid is underactive, measure the temperature under your tongue for ten minutes on three consecutive mornings before getting out of bed. Premenopausal women should do this on the second, third, and fourth day of the menses. The temperature should be

97.8°F (36.5°C) or higher. If it is consistently below this, your thyroid is likely underactive. You should then consume foods high in iodine, such as seafood and kelp, and also take selenium to stimulate your thyroid functions. Dental problems, such as dead teeth and mercury amalgam fillings, are important causative factors of thyroid problems, so use metal-free dentistry (see Step 8).

Other tests can be performed according to the skill and specialization of the examiner. The tongue, skin, hair, eyes, and nails can give valuable health information. Energy patterns can be assessed from Kirlian photography; measuring devices for acupuncture energies can reveal the meridian flows as well as the vitality of specific organs before and after treatment and at consecutive visits.

Self-examination of organs can be difficult. If you are in doubt about what you have found, consult a professional. Low-cost diagnostic instruments such as a sphygmomanometer, stethoscope, and eye, ear, nose and throat instruments can often be obtained from Chinese importers of acupuncture equipment.

Hair Analysis: For an assessment of your long-term mineral status, you can send a sample of your head hair to a specialized laboratory for analysis. (You can find an address for a laboratory on the Internet or in the advertisements of health magazines, or ask a naturo-path.) However, interpretation of the results is difficult, because mineral levels in hair are not directly related to the mineral levels in blood or cells. As the mineral content of the hair reflects the long-term mineral distribution inside the cells, it is often more useful than a serum mineral analysis.

Generally, the ratio between key elements is more important than certain high or low values taken in isolation. An abnormal ratio indicates problems with the various systems or organs. Important ideal ratios (and the body systems affected) are as follows:

- Zinc:copper = 8:1. Needed for the cardiovascular system, female reproductive system, liver functions.
- Zinc:manganese = 150:1. Muscles and skeletal system, collagen, cholesterol.
- Calcium:zinc = 3:1. Cardiovascular and kidney functions, bones and joints.
- Sodium:potassium = 2:1. Adrenal glands and glandular system.
- Calcium:magnesium = 8:1. Cardiovascular system, bones and joints, dietary imbalance.
- Iron:copper = 2.5:1. Indicates condition of blood and energy production through respiration.
- Here are some clinically significant patterns:
- Elevated levels of toxic minerals such as mercury, lead, or cadmium indicate you need to detoxify.
- Low sodium, potassium, chromium, and manganese indicate glucose intolerance.

- Elevated calcium and magnesium in the correct ratio indicate nutritionally induced hyperparathyroidism (overactive parathyroid glands) due to excess dietary phosphorus (for example, from meat and soft drinks).
- Low levels of most polyvalent elements; including copper, iron, manganese, chromium, and selenium, indicate malabsorption, with possible lack of gastric acid.

There is an interesting nutritional pattern connected with criminal behaviour. Violent, imprisoned criminals generally have one of two mineral-correlated behaviour patterns. With a high copper-to-zinc ratio, low sodium, potassium, and manganese but high lead and cadmium levels, they have extreme mood swings, poor ability to handle stress, and occasional violent behaviour. They tend to feel remorse for their actions. With the other pattern of iron, calcium, potassium, and sodium high in addition to high cadmium and lead, with low copper and zinc, such people are often violent and cruel, fascinated by fire, and exhibit no conscience or remorse. (For further information, see Diagnostic Usefulness of Trace Elements in Human Hair, by Jeffrey Bland, Ph.D.)

Face and Tongue: Face diagnosis is widely practiced in Chinese and Japanese medical traditions. Generally, the forehead corresponds to the nervous system, the middle part of the face to the circulation, and below the nose to the digestion. More specifically, the eyes are related to the liver and partly to the spleen; the nose to the lungs and heart; and the mouth and lips to the digestive functions.

To give a few examples, fluid sacs under the eyes are generally attributed to poor kidney function, especially if they are dark purple or bluish in colour, while with a puffy brown appearance they implicate the liver. Heart and lung problems may be indicated by a red or purple, swollen or bulbous nose. Soft, hanging lower lips show a weak muscle tone of the intestines. If the mouth is kept open most of the time, this shows severe metabolic imbalances and weakness.

The tongue is also a diagnostic tool. The tip of the tongue corresponds to the heart, the edges to the liver and gallbladder, the centre to the stomach and spleen; between the tip and the centre are the lungs; and at the root or back of the tongue are the kidneys, intestines, uterus, and bladder. The underside reflects the circulation in the corresponding areas. A white coating is related to congestion of the intestines; a red tongue may indicate B-vitamin deficiencies or inflammation; a bluish colour indicates coldness, underactivity, and poor circulation in the indicated areas. A yellow or brownish coating may indicate liver problems and cracks may be due to overacidity, allergies, and deficiencies.

* * * * * * * * * * * *

Step 12

IRIS DIAGNOSIS

Study your eyes to find out more about internal conditions of your body.

The iris is like a map of the body - changes in certain organs are reflected in specific parts of the iris. The right iris shows the condition of the right side of the body, while the left iris reflects the left side. The exact relationship between iris and body parts can be seen from the iris chart below. Iris diagnosis is also known as iridology.

In health, the iris is composed of densely structured fine, straight lines, radiating from the pupil to the outer rim. A close grain, similar to that of hardwood, indicates a strong inherited vitality and good recuperative powers in the case of temporary illness. If the fibres are loosely spread, as in softwood, the basic health is weak.

In poor health these lines become separated and distorted, forming various patterns, called markings. Very weak organs often show elliptically formed grey markings - so-called closed lesions resembling knots in wood. In poor health many of these closed lesions may be found in the iris, indicating areas in which the circulation is stagnating. If these lesions are not 'walled in', but open at one end or both, this indicates that despite a weakness the circulation in this area is good.

Colour Changes: Start by studying your own eyes in a mirror. Then look at the eyes of friends and relatives. Use a magnifying glass and a torch held at the side of the eye. Make a coloured copy of your own eyes or those of a friend, and compare it with the iris chart. Study the general colour pattern. Markings are much easier to detect in blue than in brown eyes. Often there will be brownish discolourations in blue or green eyes extending outward from the pupil. This area belongs to the intestines. The brown colour change indicates that there is a deterioration of the digestive system, usually associated with inherited liver and gall bladder weakness.

Frequently, the eyes of babies change from blue to brownish within days or weeks of the baby's introduction to cows' milk. Often allergy symptoms are present simultaneously, for example, eczema, respiratory and digestive difficulties. These colour changes may also occur in breastfed babies if the mother uses cows' milk or its products. Presumably this change may already occur in the foetus.

The brown colour of genuinely brown eyes comes from melanin pigments, while pathological brown colour changes originate from oxidised lipoproteins (for example, lipofuscin) and possibly from the breakdown products of blood colouring agents (for example, bilirubin). Additional discolouration may result from drug deposits.

Sometimes there is so much brown it is difficult to detect the original colour. Organ areas that border the intestinal ring where it shows strong markings are likely to suffer from reflected weaknesses. White in the iris indicates overactivity, irritation, acidity, infection, inflammation or catarrh of the corresponding body part. In some eyes the whole iris shows much white; in others it is concentrated in certain areas only.

Generally, a brown discolouration means that cows' milk products, saturated fats (all fats that are solid at room temperature), chemicals and stimulants should be avoided as much as possible. A very white discolouration indicates mucus congestion. Therefore, mucus-forming foods should be avoided - gluten, refined carbohydrates, wheat, lactose, oranges and anything mouldy.

Gradually, an organ may change from an inflamed to a chronically weak condition. The white in the iris will simultaneously change to grey; the darker the grey, the weaker the organ. During health improvement you may watch the reverse process: the grey areas becoming lighter in colour, then white and finally restored to the original colour - a process that takes many years.

Generally, white markings indicate a need for sedating, anti-inflammatory treatment of the associated organ, while grey or brown areas show a need for strengthening and stimulation.

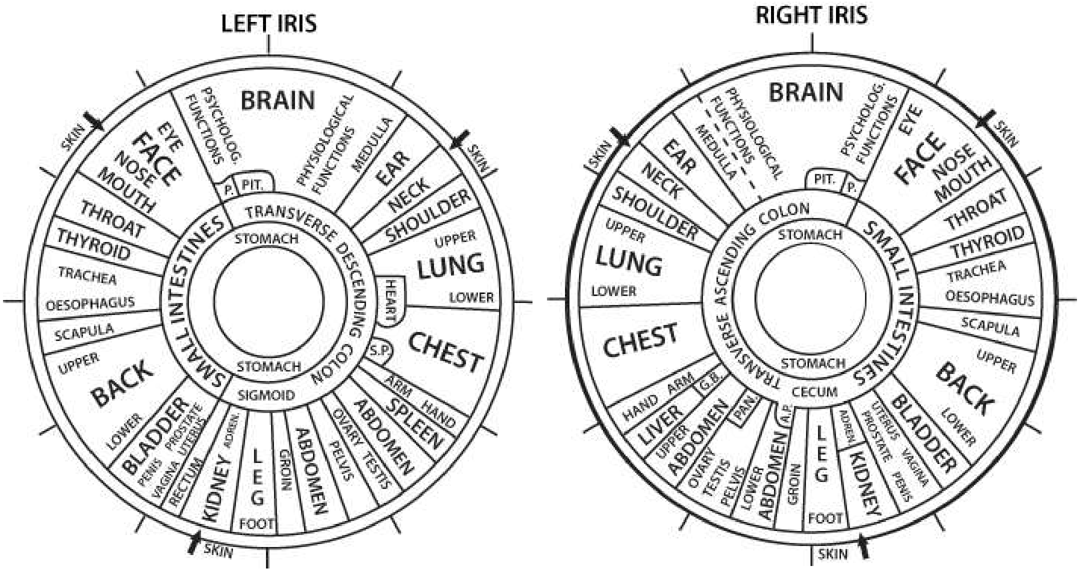
LEFT IRIS
BRAIN
PSYCHOLOG. FUNCTIONS
PHYSIOLOGICAL FUNCTIONS
MEDULLA
EAR
NECK
SHOULDER
UPPER
LUNG
LOWER
HEART
CHEST
S.P.
ARM
HAND
SPLEEN
ABDOMEN
OVARY
TESTIS
PELVIS
ABDOMEN
GROIN
L E G
FOOT
ADREN.
KIDNEY
RECTUM
VAGINA
PENIS
PROSTATE
UTERUS
BLADDER
LOWER
BACK
UPPER
SCAPULA
OESOPHAGUS
TRACHEA
THYROID
THROAT
MOUTH
NOSE
FACE
EYE
SKIN
SKIN
SKIN
P.
PIT.
TRANSVERSE
DESCENDING COLON
STOMACH
STOMACH
SIGMOID
SMALL INTESTINES
RIGHT IRIS
BRAIN
PHYSIOLOGICAL FUNCTIONS
MEDULLA
PSYCHOLOG. FUNCTIONS
PIT.
P.
EYE
FACE
NOSE
MOUTH
THROAT
THYROID
TRACHEA
OESOPHAGUS
SCAPULA
UPPER
BACK
LOWER
BLADDER
VAGINA
PENIS
UTERUS
PROSTATE
ADREN.
KIDNEY
L E G
FOOT
GROIN
ABDOMEN
LOWER
PELVIS
TESTIS
OVARY
ABDOMEN
UPPER
LIVER
HAND
ARM
G.B.
PAN.
A.P.
CHEST
LOWER
LUNG
UPPER
SHOULDER
NECK
EAR
SKIN
SKIN
SKIN
ASCENDING COLON
TRANSVERSE
STOMACH
STOMACH
CECUM
SMALL INTESTINES

IRIS CHARTS

Abbreviations: ADREN = Adrenal Gland, AP. = Appendix, G.B. = Gall-bladder, P. = Pineal Gland, PA. =Pancreas, PIT. = Pituitary Gland, S.P. = Solar Plexus.

Common Conditions:

PUPIL REFLEX: If a pupil does not contract when a bright light shines on to it, this indicates dulled nerve reflexes, weak adrenal glands or an overstimulated sympathetic nervous system, often from fear as a hidden, chronic condition. Adrenal stress is indicated if the pupils begin to expand and contract repeatedly when exposed to bright light for 30 seconds.

NERVE WREATH: A strong, white and almost circular outline of the intestinal area - the nerve wreath - indicates a good Condition of the autonomic nervous system. If this outline is weak, jagged, discoloured, or extends far towards the periphery or the pupil, we may assume the autonomic nervous system is in a poor condition. The normal position of the nerve wreath is one-third the distance between the pupil and the periphery; if the nervous system is tense and overactive, the wreath is closer to the pupil, and if the nervous system is relaxed and under-active it is closer to the periphery.

NERVE RINGS: White circles or arcs of circles in the outer part of the iris. They indicate a tense, over-reactive, irritated nervous system. The outermost iris zones represent the lymphatic and circulation system and, bordering the sclera, the skin.

DISCOLOURED STOMACH AREA: In a normal condition the stomach area is not visible, but if the stomach is irritated the area bordering the pupil will be whitish and distinct from the intestinal area. A brownish discolouration of the stomach area indicates a chronic weakness.

DISCOLOURED INTESTINAL AREA: A whitish intestinal area indicates inflammation, irritation or ulcers; avoid gluten products, lactose, spices and acid foods; use mucilaginous supplements (fenugreek, slippery elm, comfrey, linseed), brown rice, and the outer parts of potatoes. If there is a mucous congestion in the head, there is a whitish or brownish discolouration in the transverse colon area, radiating towards ear and brain. If chronically weak, this area will be dark.

RADIAL BLACK LINES: When there is a serious deterioration of the intestines, strong black lines (radii solaris) will develop, starting from the pupil and radiating towards the periphery. Organs through which these radii pass will be very weak as well.

UNUSUAL MARKINGS: Distinct markings may be coloured brown, red or yellow. Unusual colours indicate deposits of drugs or other chemicals that settle in weak organs.

WHITE OUTER RING: A heavy white ring near the outer edge of the iris points to salt (sodium) and calcium deposits. Avoid salt, drink plenty of water, possibly use potassium and magnesium supplements, and natural vitamin D; improve kidney activities.

BLUISH-WHITE FILM: A bluish-white film beginning to cover the iris from the outer rim indicates poor circulation to this area and an anaemic condition. Often this film appears in the brain area, indicating approaching senility (arcus senilis).

DARK OUTER RING: A dark ring at the outer rim (scurf rim) shows the skin is inactive with accumulated wastes, and needs frequent stimulation, better circulation and improved kidney and lung activities.

LYMPHATIC ROSARY: A series of white spots near the outer rim, the 'lymphatic rosary', indicates chronic infection and congestion of the lymphatic system. Avoid especially lactose and cows' milk products (except butter).

It may require years to become an expert in iridology, but the basic rules given here will already enable you to form a well-founded judgement about the condition of your body. You may also be able to help other people if you have this knowledge. Experiment and you will soon find iridology a valuable and fascinating tool. The most extensive information in book form on iridology is available from Bernhard Jensen, mainly his book Iridology Volume 2.

Sclerology: In sclera diagnosis or sclerology, the location and shape of the blood vessels visible in the sclera, the white of the eye, are interpreted. Generally only the problem areas are indicated, not the nature of the complaint. Sclerology may show that the function of an organ is disturbed, but not whether it is overactive or under-active. It helps if you have good light and a magnifying glass for examination.

To inspect the lower part of the sclera, let the examinee look up while you pull down the lower eyelid; to look at the upper part, the examinee should look down while you gently roll up the upper lid; examine the outside while the examinee looks toward the other eye and the inner part of the sclera while looking away from the other eye.

Generally a blood vessel pointing towards a certain organ reflex in the iris indicates that there is something wrong with that organ or part of the body. The stronger the blood vessel is visible and also the more of them are bunching together, the worse is the problem. A bluish colour of the sclera indicates under-activity.

* * * * * * * * * * * *

Step 13

MUSCLE TESTING

Test your muscle strength to see whether a food or remedy is good or bad for you.

Our muscles respond to many external as well as internal influences by becoming stronger or weaker. In muscle testing, the change in muscle strength is assessed before and during contact with a test item. This may be, for example, a food sample, a colour or a piece of material. Internal influences are our thoughts and feelings, which may either strengthen or weaken us.

The reason for the muscle changes can be explained by the influence the radiation of a tested sample has on the flow of bio-energies along the acupuncture-meridians. In the same way, our thoughts and emotions can change the flow of bio-energy. I have measured differences in muscle strength of up to 50 percent during testing.

Most suitable methods for food testing are to place a sample in the mouth - chewing it for about 10 seconds - and also to hold a sample in one hand near the navel. Most non-food items can be tested near the navel or simply by touching with a hand.

How to Test: The simplest form of muscle testing requires a tester as well as the tested person. Almost any muscle can be used for testing, but I use the following method as the most convenient. The tested person sits with uncrossed legs, one arm held out to the side, elbow level with shoulder and bent at a right angle, place the free hand over the navel.

The tester presses just above the elbow to push it down. Press the arm lightly into the shoulder as well as toward the hip. If the tested person is very strong, the pressure may be straight down, or the arm may be fully extended and the pressure applied at the wrist. Increase the pressure gradually and with measured strength; feel and remember exactly the amount 'of pressure required to force the arm down. If both arms are too weak, you may test the legs.

During testing, the tester should keep the question in his or her mind 'Is the tested sample good or bad for the patient?' but should not expect a certain result. The preferences of the tested person seem to have little influence on the result.

Now the tested person takes a sample and holds it with the free hand just above the navel. After about 10 seconds, the tester presses the arm down and carefully notes any difference in resistance. If the result is doubtful, repeat the test with and without holding the sample. Often a difference is easier to detect if the arm is pressed down three times in succession.

Sometimes the tested person feels a pain or strain in the shoulder during a weakening test, even if the tester cannot notice a difference. This should be reported. Generally the quantity of food held is not critical; it may even remain in a container or plastic bag; however, avoid a coloured container or one of semi-rigid plastic.

If the patient is difficult to test, you may strengthen the arm immediately before each test by tracing the acupuncture meridian of the lung or of the glands (triple warmer); alternatively, test a leg. Another strengthening method is to tap the thymus gland (underneath the upper end of the breastbone) several times lightly with your fingers. This strengthens the immune system and in doing so reduces allergic reactions.

What to Test: In addition to foods, testing can be done on vitamins, herbs, drugs and even garments. You will find that synthetic fibres (especially nylon), many drugs, refined foods, chemical additives and incompatible colours are frequently weakening.

Food combinations may be assessed by holding them jointly during a test. Individual items may even remain in their separate containers. Samples that are strengthening individually may test weakening if held together, and vice versa. The test results may be influenced by what you have eaten beforehand. Also, when you are tired, fatigued or sick, many foods are weakening that at other times may be strengthening. Therefore, important items should be retested from time to time under varying conditions.

Use muscle testing frequently for all your food and supplement requirements. Supplements will test strengthening as long as they are required. Sometimes, muscle testing will indicate that an item should be omitted temporarily before it is used again, or that it should be taken in a lower dosage (for example, vitamin A). For assessing the amount to be taken per day, gradually increase the number of tablets or capsules held in the hand until the arm weakens. Take about 10-20 per cent less than this amount.

You may also try to find out whether your usual sleeping location is safe. Perform a muscle test while lying in the usual sleeping position on the bed. If the arm is weaker than with a test in another part of the bed or the room then you may have to change your sleeping location.

Even thoughts and feelings may be tested. A positive thought and a good feeling may be strengthening quite generally to all muscles, although the main strengthening effect may be on one particular muscle. Negative thoughts or unpleasant feelings, on the other hand, may produce a general weakening and also a particular muscle to become very weak when tested. Mental testing may be used to explore your subconscious beliefs, feelings and body reactions.

Instead of testing the safety of your bed while actually lying on it, you may use mental testing to find out whether your bed is situated over underground radiation that is detrimental to your health. While your partner tests your muscle strength, imagine lying or sleeping in the bed. Then imagine an underground stream and harmful radiation underneath the bed. If the arm is stronger when imagining the usual sleeping arrangement but weaker with the imagined stream, then your bed is likely to be in a good spot, otherwise continue the test by imagining sleeping in a different location until the test becomes strong.

Another possibility is surrogate testing. You may use this for a small child or anyone who cannot be tested directly. If you want to know whether a certain food is good for the child, place the food on the child's abdomen; at the same time touch the child with one hand and extend the other arm for your partner to test.

Finally, you may develop a method of self-testing. One possibility is as follows: press the tips of the thumb and one finger of the left or right hand together to form a loop. With the index finger of the other hand hook behind the joined fingertips, and with a quick move try to pull it through. Adjust the pressure of the joined fingers and the force of the pulling index finger so that you cannot pull it through when the answer is 'yes' or 'good'; but when the answer is 'no' or 'bad' the joined fingers become weaker and can more easily be opened. When you can do this successfully with known answers, you can test food samples in your mouth or on your abdomen, and also try mental testing.

Vitamin and Mineral Testing: M, Grimm and W. Fishman developed the following method to assess vitamin or mineral deficiencies by holding certain contact points on the body. The tested person applies a light pressure on the contact point with a pad formed by the first three fingers of the free hand, while the other arm is tested for muscle strength. Start by testing the muscle strength without touching the body with the free hand, and then repeat the test while touching the specified contact point.

The contact points for the main vitamins and minerals are as follows.

- Vitamin A: right eyelid
- B vitamins: tip of tongue
- Vitamin C: just below left collarbone, 5 cm from mid-chest
- Vitamin D: midway point of left groin
- Vitamin F: just below right collarbone, 5 cm from mid-chest
- Calcium: just above left collarbone, 5 cm from mid-chest
- Iodine: just below Adam's apple
- Iron: midway point of the right groin
- Magnesium: middle of the navel
- Phosphorus: midway between pubic bone and left hipbone
- Potassium: hollow of right cheek
- Zinc: midway between pubic bone and right hipbone

For more information see Muscle Response Test by M. Grimm and W. Fishman.

Touch for Health: Some influences - nutrients, colours, thoughts and emotions - affect certain muscles more than others. In addition, organ functions can be tested with specific muscles. If the muscle is weak, it indicates a weak organ function. Meridian tracement may be used to strengthen a weak muscle or organ. Some useful organ and muscle tests, as used in Touch for Health by J. F. Thie and M. Marks, are given below.

Adrenals: Lie face up, with one leg up and bent so that the foot is over the other leg just below the knee; the tester presses against the heel and the top of the knee to straighten the leg.

Brain, Mental Stress: Hold an arm straight, about 15 degrees away from the body, slightly sideways; the tester presses the forearm towards the groin.

Circulation/Sex: Lie on one side, with the upper leg straight, raised as high as possible; the tester presses the ankle down.

Gall bladder: Lie face up, holding the arm straight 45 degrees above the leg, palm downward; the tester presses the forearm down.

Heart: Hold the upper arm to the side, level with the shoulder, with the elbow bent at 90 degrees and the forearm down; the tester steadies the elbow and pulls against the wrist to pull the hand forward and upward.

Kidneys: Lie face up, with one leg raised at 45 degrees, slightly towards the side, with the foot turned out; the tester presses the inside of the ankle down and inward.

Large Intestine: As for the kidneys, but with the foot turned in; the tester presses the ankle down and inward.

Liver: Hold the arm straight forward, level with the shoulder, slightly sideways, with the palm outward (thumb towards feet); the tester presses the forearm upward and outward.

Lungs: Hold the arm to the side, elbow level with the shoulder and bent at 90 degrees, forearm horizontal; the tester presses the elbow down.

Neck and Sinuses: Lie with the arms above the shoulders, head up, chin down; the tester presses against the forehead.'

Pancreas: Hold the arm straight down touching the side, with the thumb towards the body, the tester pulls the forearm away from the side. This is a good test to perform while shopping, the tested person touching a questionable food item.

Small Intestine: Sit on the floor with closed straight legs, hands on opposite shoulders, lean back slightly; pressure is against the crossed arms.

Spleen: Lie face up, with the arm straight to the side, palm upward; the tester pushes the back of the wrist off the ground.

Stomach: Hold the arm as for the liver; the tester presses downward and outward.

Generally, the tester's free hand is used to stabilise the tested person's body so that only the tested limb moves, not the whole body. Both arms and legs may be tested.

* * * * * * * * * * * *

Step 14

THE ENDOCRINE PATTERN

Simple arm and leg measurements can show the conditions.

The body has developed under the guidance of the endocrine glands. Therefore, glandular disturbances are reflected in abnormal proportions of various parts of the body. Certain measurements enable us to recognize our glandular pattern as a basis for corrective treatment.

Our endocrine system is composed of two antagonistic groups of glands. The anabolic (build-up) glands are the adrenal cortex, the posterior pituitary (in the brain), the parathyroid, and the pancreas (Islets of Langerhans). The catabolic (breakdown) glands are the adrenal medulla, the anterior pituitary, and the thyroid gland. The anabolic glands control the calcium level of the blood, while the catabolic glands control the phosphorus level.

Ingestion of sugar or sweet foods stimulates the anabolic glands, disturbs the calcium-phosphorus ratio, and leads to an excess of calcium in the blood and, consequently, to stone formation. Later, however, it leads to exhaustion of the anabolic glands.

Table 2-1: Normal Body Measurements

Ratios	**leg : arm**	**ankle : wrist**	**one-fourth**	**one-half**	**three-fourths**	**knee : elbow**
Female	1.770	1.550	1.295	1.425	1.340	1.530
Male	1.740	1.475	1.220	1.350	1.265	1.455

Interpretation:

One-fourth ratio higher than normal: underactive thyroid
One-fourth ratio lower than normal: overactive thyroid
One-half ratio higher than normal: posterior pituitary underactive
One-half ratio lower than normal: posterior pituitary overactive
Three-fourths ratio higher: anterior pituitary underactive
Three-fourths ratio lower: anterior pituitary overactive

Measurements: The following measurements can reveal the basic functional pattern of the thyroid, pituitary, and sex glands. The results for the thyroid and pituitary glands are accurate to more than 95 percent; for the sex pattern, to about 80 percent.

- Measure the length of the lower leg from the middle of the kneecap to the middle of the ankle, and the lower arm from the point of the elbow to the middle of the wrist bone.

- Divide each measurement by four and mark the quarter-points between knee and ankle, and between elbow and wrist. (This means one quarter of the distance from knee to ankle is the one-fourth point, half the distance between knee and ankle is the one-half point, and three quarters of the distance is the three-fourths point. In the same way divide the distance from the elbow to the wrist.)
- Measure the distance between the endpoints (knee to ankle, elbow to wrist) as well as between the three quarter-points in between.
- Divide all leg measurements by the corresponding arm measurements.

The measuring tape should not be tight for the circumferential measurements; contact measurement only is wanted. The foot should be at right angles to the leg during measurement. All measurements are taken for both sides of the body. Compare your measurements with the normal values shown in Table 2-1.

Table 2-2: Further Signs of Andricity or Gynicity

Signs of Andricity	**Signs of Gynicity**
The back of the hands point forward when the arms hang loosely at the sides (person must be unaware of the test)	More of the palms show when arms hang loosely at the sides
A low voice if female	A high voice if male
Pubic-hair line is triangular with apex upward	Pubic-hair line is flat on top
Aggressiveness, independence, critical thinking	Intuitive instead of critical thinking
Body in general is hairy	The other signs are opposite to those listed for andricity
Tendency to become bald	
Space between thighs	
Tends to have too much gastric acid (ulcers)	

"Andric" means more male hormones are produced than is normal for the average male or female.
"Gynic" means more female hormones are produced than is normal for the average male or female.

The anterior pituitary gland mainly produces growth and sex hormones, and it stimulates the thyroid and adrenal glands. The posterior pituitary regulates the amount of water lost in the urine; it can raise the blood pressure and stimulates labour contractions.

Assessing Male-Female Balance: Both sexes produce male as well as female hormones in a gender-specific ratio. If there are more male hormones produced than normal, the condition is called "andric"; if more female hormones are produced, it is "gynic." To give an example, a woman with a constitutionally high oestrogen-testosterone ratio is gynic, while with a low ratio she is andric.

If the leg-arm ratio is higher than normal, add the difference between actual and normal measurement to the normal ankle-wrist ratio; if the leg-arm ratio is lower than normal, subtract the difference. If the actually measured ankle-wrist ratio is higher than the adjusted normal ratio, the person is gynic; if the actual ratio is lower, the person is andric. Gynic females may be more harmed by oestrogen birth control pills than andric females. Generally, andrics need more magnesium and gynics more calcium. For further signs of andricity or gynicity, see Table 2-2.

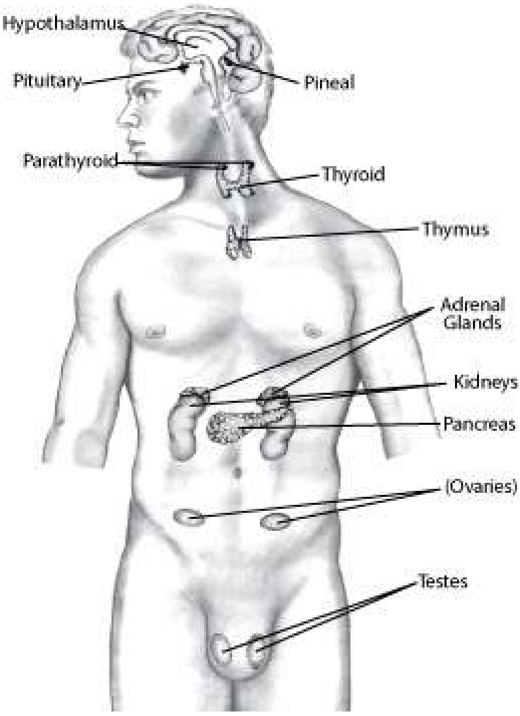
Hypothalamus
Pituitary
Pineal
Parathyroid
Thyroid
Thymus
Adrenal Glands
Kidneys
Pancreas
(Ovaries)
Testes

If one or more of the gland measurements show abnormal functions (as they usually will), treat the corresponding glands with diet, ear acupuncture, acupressure, reflexology, and guided imagery. Minute amounts of gland extracts can be prescribed. (For further information, see Body Chemistry in Health and Disease, by M. E. Page.)

* * * * * * * * * * * *

Step 15

REFLEXOLOGY

Press reflex points on your hands and feet to diagnose and treat conditions within your body.

Reflexology is excellent for self-treatment and even for diagnosis. Envisage the body divided into 10 vertical segments or zones, five on each side. Each finger and toe belongs to one of these zones. A diseased body part will show up as a sore spot in the same segment of the foot or hand. From this the alternative name 'zone therapy' is derived. Usually the feet are more sensitive and, therefore, more suitable for reflex diagnosis and treatment than the hands. In addition the ears, like the irises, contain a reflex map of the body. (We may even assume that every organ and every single cell contains a holistic reflex map of the whole organism.)

In addition, there is a reflex connection between related parts of the legs and arms. A sore knee, for example, may be relieved by pressure on the corresponding location on the elbow on the same side of the body. In a similar way, corresponding locations on the wrist and ankle, and on the hip and shoulder, influence each other. This is especially useful to know when dealing with broken or otherwise injured limbs that do not permit direct treatment.

Diagnosis: Press the soles, the toes and the sides of the feet with a knuckle, a thumb or - if the feet are not very sensitive - with a blunt probe. Overweight people and those with poor circulation usually need quite strong pressure, which must be exactly on the correct spot. However, if the lack of sensitivity is due to nerve damage, then it would be better to try and improve this condition first, as with strong sustained pressure under the back of the skull on both sides of the spine and providing any missing nerve vitamins (e.g. B1, B6, B12, folic acid).

Press the selected area in a slow circular motion until you find the exact location. Generally, the pressure should be straight down, but for the thyroid area it should be applied from the side, trying to press underneath the ball of the big toe.

Adjust the pressure to the general sensitivity of the foot. If the body is very sensitive, all foot reflexes will be extremely tender; therefore try with light pressure to locate only the most painful areas. You may mark the encountered tender spots on an examination chart, using different markings according to the extent of tenderness encountered. Concentrate later treatment on the most painful areas. The results will be better if you have someone other than yourself check your feet.

Methods of Treatment: The easiest and most common form of treatment is to apply pressure on sore points with the pad of a finger, the thumb, with a knuckle or with a blunt instrument (for example, the blunt end of a pencil or a pen). If it is quite sore, press straight down, with only light pressure at first. Gradually increase the pressure as the pain lessens. Finally, when even strong pressure can be tolerated, you may apply the pressure in a slow circular motion in order to relax the deeper muscles and ligaments.

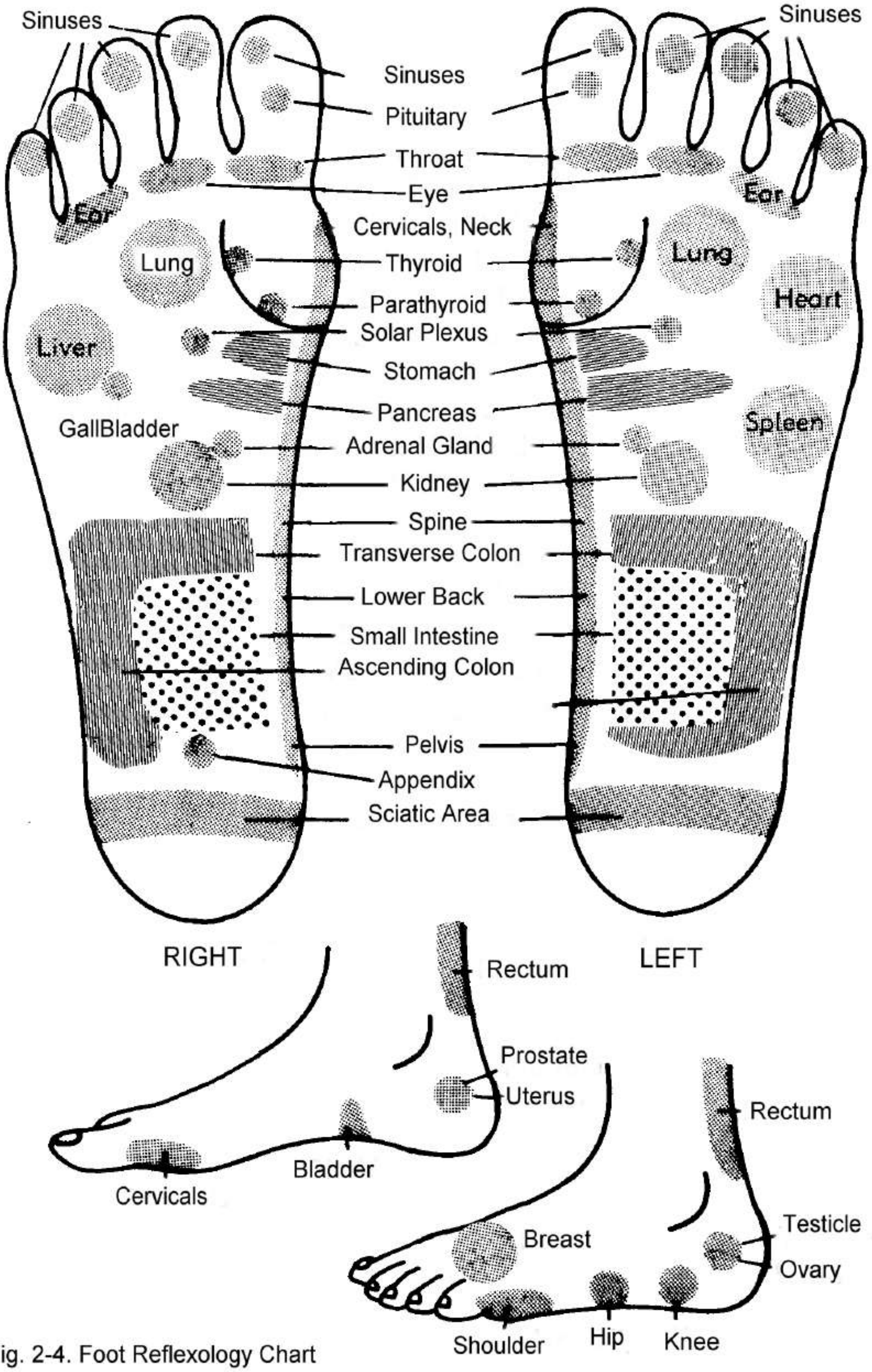
Sinuses
Sinuses
Sinuses
Pituitary
Throat
Eye
Ear
Ear
Cervicals, Neck
Lung
Lung
Thyroid
Heart
Parathyroid
Liver
Solar Plexus
Stomach
Pancreas
Spleen
GallBladder
Adrenal Gland
Kidney
Spine
Transverse Colon
Lower Back
Small Intestine
Ascending Colon
Pelvis
Appendix
Sciatic Area
RIGHT
Rectum
LEFT
Prostate
Uterus
Rectum
Bladder
Cervicals
Testicle
Breast
Ovary
Shoulder
Hip
Knee

Fig. 2-4. Foot Reflexology Chart

For self-treatment underneath the feet, it will usually be more convenient to press or roll the sore points on a golf ball, a stone or another suitable object. To press reflex areas in the palms of the hands, it is usually more convenient to use a knuckle or a suitable blunt instrument than the pad of the thumb.

Usually, the pain begins to subside after a few minutes of pressing. If this happens, increase the pressure for a while, but if the pain instead increases, stop the treatment. The effect on related body parts can be surprisingly quick: pain in the ears, shoulders or back often disappears within minutes.

If there is a serious or acute problem, use reflexology several times a day. However, this treatment may stimulate a chronically weak organ to increased activity and it may become tender for a while together with its reflex areas. Generally, deep, strong and slow pressure is sedating and pain relieving, while light, quick tapping or circling is stimulating. Select whichever is appropriate. For some points a so-called C-clamp is effective: press with the thumb from underneath the foot and with the index finger from the top. This may be used for the liver, spleen and thymus.

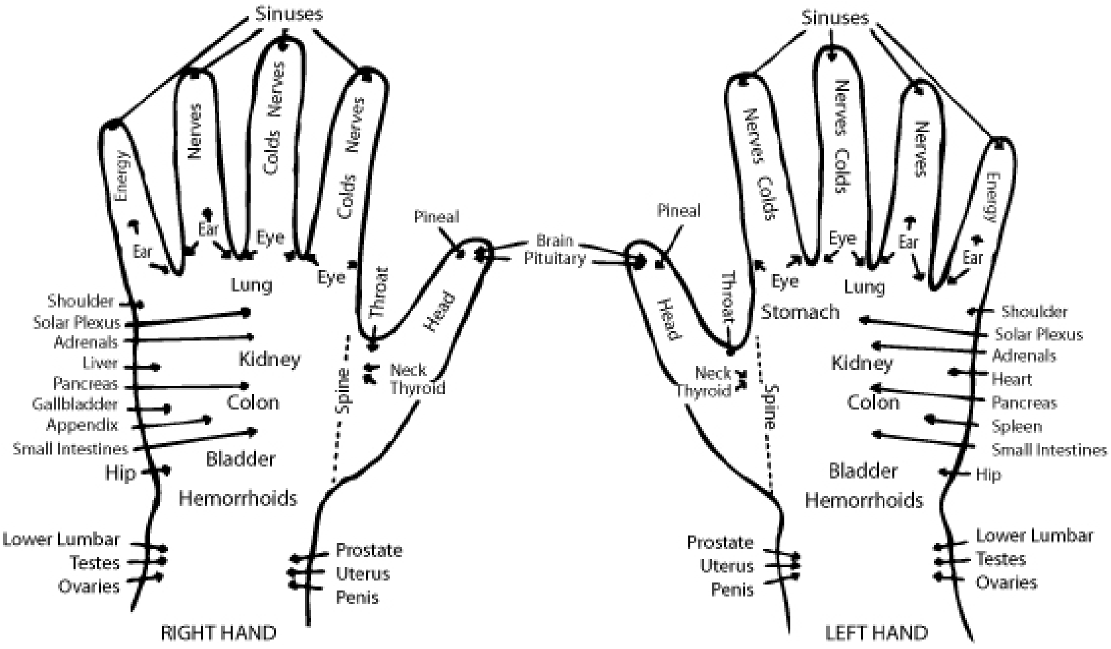

Sinuses
Energy
Nerves
Colds Nerves
Colds Nerves
Ear
Ear
Eye
Eye
Lung
Pineal
Brain
Pituitary
Throat
Head
Neck
Thyroid
Spine
Shoulder
Solar Plexus
Adrenals
Liver
Pancreas
Gallbladder
Appendix
Small Intestines
Hip
Kidney
Colon
Bladder
Hemorrhoids
Lower Lumbar
Testes
Ovaries
Prostate
Uterus
Penis
RIGHT HAND
Sinuses
Nerves Colds
Nerves Colds
Nerves
Energy
Pineal
Eye
Eye
Ear
Ear
Throat
Head
Stomach
Lung
Neck
Thyroid
Spine
Shoulder
Solar Plexus
Adrenals
Heart
Pancreas
Spleen
Small Intestines
Hip
Kidney
Colon
Bladder
Hemorrhoids
Prostate
Uterus
Penis
Lower Lumbar
Testes
Ovaries
LEFT HAND

Figure 2-5: Hand Reflexology Chart

Another method of applying pressure to fingers and toes is to put tight rubber bands around them or to use clothes pegs to clamp the digits. Rubber bands should never be left on for more than a few minutes at a time. In addition, you may firmly grasp a metal comb, pressing its teeth into the palm or the fingers. A metal comb or a brush may also be used to sweep the backs of the hands and lower arms up to the elbows in quick upward strokes. Five minutes of this will stimulate the circulation and invigorate the body.

Further effective areas for treatment are the tongue, the inside wall of the mouth, especially the palate (top), under the tongue and the back of the mouth (pharynx). You may press with the pad of a finger or thumb or use a suitable instrument. Pressure on the tongue may be applied by biting or pressing with the handle of a spoon. This is effective for treating problems in other areas of the body belonging to the same zone.

In addition, the tongue may be pulled out for several minutes by grasping it with a cloth and moving it around. Alternatively, you may poke it out as far as possible and hold it in place with your teeth. You may also work around the angles of the jaw and chin. Because of the sensitivity of the mouth, this treatment is often more effective than hand and foot reflexology, especially for all problems related to the head and throat.

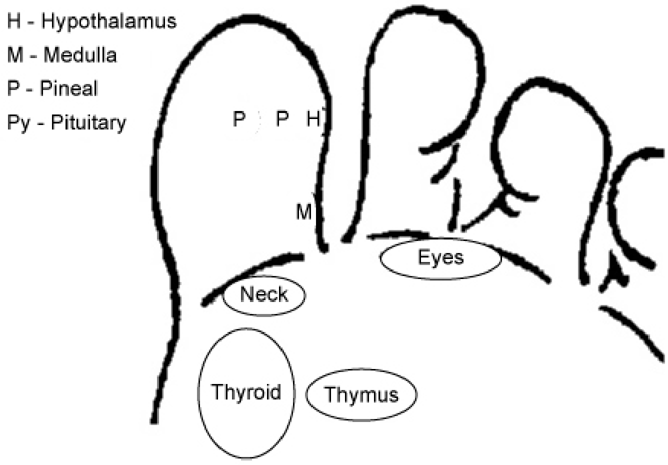

Figure 2-6: Toe Reflexology Chart

Important Areas to Treat: The foot and hand charts will give you a good indication of where to find suitable points for treating specific complaints. For best results work on all the major reflex areas: hands, feet, ears and mouth. Always treat in the same zone in which the problem arises. Your left inner ear, for example is in zone 4. Therefore, to treat it you will manipulate the left ring finger the third toe including its root on the left side, the ear points of the left ear and the outer left side of the tongue and palate.

Work frequently around the joints of fingers and toes, especially digging underneath the lowest joints of the thumbs and big toes. Rubber bands or clothes pegs may be placed for five to 10 minutes on the tips of the digits, or on or between the joints, depending on where there is more sensitivity. Lymph drainage can be improved by press-massaging the area of the upper chest on top of the feet just below the toes.

Generally, treat a sensitive area for five to 10 minutes at a time. For an acute problem, you may treat it several times daily, while chronic problems may be treated once a day. For general health improvement you may select a different area each day until you return to the same area after one or two weeks. This process may continue for years, as formerly weak areas become activated and cause increased sensitivity in their corresponding reflex areas. Walking barefoot on rough ground provides a natural foot massage - make good use of it.

The Brain: For general health improvement and regeneration work frequently on the brain, especially the pituitary gland, the pineal gland, the hypothalamus and the medulla (oblongata). The pituitary reflex often feels like a sharp bone protrusion; press it in the direction of the tip of the toe. The hypothalamus and medulla are pressed from the insides of the big toes. For general health improvement it is best to start with activating these parts of the brain and then continue with the other glands and organs.

In the system of Body Electronics as developed by Dr John Whitman Ray it is recommended to work in a certain sequence for regenerating the body. Start with the so-called STO-point: press into the valley between the muscles from both sides of the spine and then up under the skull (occiput). This is to improve the nerve transmission to all parts of the body.

Then start working on points in the following order: pineal gland, hypothalamus, pituitary, medulla, thyroid, thymus, heart, solar plexus, pancreas, adrenals, spleen, gonads (uterus, ovaries, prostate, and testicles), liver, kidneys, gall bladder, bladder, small intestine, large intestine.

* * * * * * * * * * * *

Step 16

EAR ACUPUNCTURE

Try this easy method to improve the function of any diseased or unbalanced organ.

Ear acupuncture can be a very effective method for relieving acute pains and chronically stiff joints, sometimes these have been cured within minutes. For some problems it can be more effective than body acupuncture. Ear acupuncture is easy and safe for self-treatment and in healing groups. Organ diseases and glandular disturbances may respond less dramatically, but ear needling will always be very helpful as a supportive method with other forms of treatment.

The main disadvantage of ear acupuncture is the requirement to find the most painful spot for it to be fully effective, and this is not pleasant for people who are very sensitive to pain. The best points are those that feel like an electric shock or like hitting a raw nerve! However, often the pain will be felt as being beneficial, especially because it usually decreases with continued needling. This pain may be considered as 'letting off steam' or congestive energy from a target organ through the 'safety valve' of an ear point. However, chronically weak organs may also benefit from ear needling if no pain is felt in the corresponding ear point. Generally, however, you may follow the rule 'the more sore the point, the more effective the treatment'.

Acupuncture needles are available from some medical-equipment shops, acupuncturists or from acupuncture supply companies, see the Yellow Pages of major cities for addresses. For ear acupuncture use short and medium fine needles (e.g. 25 mm No. 30 gauge or the finer 12mm long 32 gauge. Usually you buy them in packs of 10. In addition press-needles are very effective and may also be ordered in packets of 10. Press-needles are like tiny drawing pins with a 2 mm long needle. They are retained in the ear for one to two weeks, covered with adhesive tape, and stimulated from time to time by pressing between thumb and a finger pad. If an acupuncture needle is not available, you may use a bead needle or even a fine sewing needle.

The Needle Method: Some authorities suggest ear acupuncture does not work through the acupuncture meridians, but instead, like reflexology, through the press-point therapy of hands and feet. In any event, Chinese medicine says the ear is the place where all the channels meet. The ear is relatively safe and easy to use for self-treatment. The following description of using needles for ear treatment is provided only to show how it is done in some other countries or by acupuncturists. For legal reasons, I do not formally recommend using needles on the ear. Instead, you can use commercially available electro-acupuncture instruments or laser lights to treat points, or you can use press-point therapy as described below. Exercise care and caution.

For ear acupuncture, short and medium fine needles are being used in addition to press-needles. Press-needles are like tiny drawing pins. They are retained in the ear for one to two weeks, covered with adhesive tape, and stimulated from time to time by pressing between thumb and a finger pad. If an acupuncture needle is not available, a bead needle or even a fine sewing needle is sometimes used.

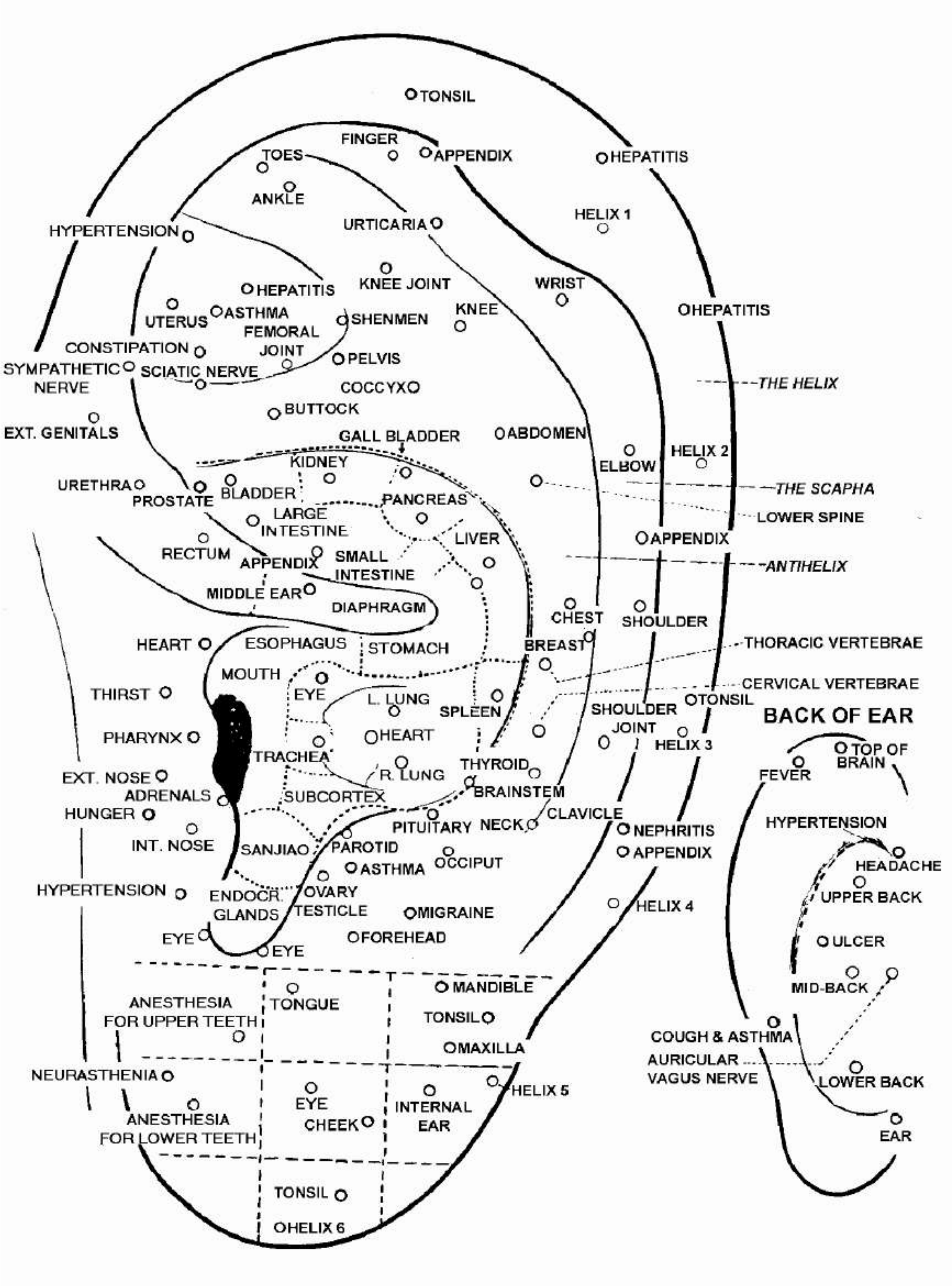
TONSIL
FINGER
TOES
APPENDIX
HEPATITIS
ANKLE
HYPERTENSION
URTICARIA
HELIX 1
HEPATITIS
KNEE JOINT
WRIST
UTERUS
ASTHMA
SHENMEN
KNEE
HEPATITIS
FEMORAL JOINT
CONSTIPATION
PELVIS
SYMPATHETIC NERVE
SCIATIC NERVE
COCCYX
THE HELIX
BUTTOCK
EXT. GENITALS
GALL BLADDER
ABDOMEN
HELIX 2
KIDNEY
ELBOW
URETHRA
PROSTATE
BLADDER
PANCREAS
THE SCAPHA
LARGE INTESTINE
LOWER SPINE
LIVER
APPENDIX
RECTUM
APPENDIX
SMALL INTESTINE
ANTIHELIX
MIDDLE EAR
DIAPHRAGM
CHEST
SHOULDER
HEART
ESOPHAGUS
STOMACH
BREAST
THORACIC VERTEBRAE
MOUTH
CERVICAL VERTEBRAE
THIRST
EYE
L. LUNG
SPLEEN
SHOULDER JOINT
TONSIL
BACK OF EAR
PHARYNX
HEART
HELIX 3
TRACHEA
TOP OF BRAIN
EXT. NOSE
R. LUNG
THYROID
FEVER
ADRENALS
SUBCORTEX
BRAINSTEM
HUNGER
CLAVICLE
PITUITARY
NECK
NEPHRITIS
HYPERTENSION
INT. NOSE
SANJIAO
PAROTID
APPENDIX
ASTHMA
OCCIPUT
HEADACHE
HYPERTENSION
ENDOCR. GLANDS
OVARY TESTICLE
UPPER BACK
MIGRAINE
HELIX 4
EYE
EYE
FOREHEAD
ULCER
MANDIBLE
ANESTHESIA FOR UPPER TEETH
TONGUE
MID-BACK
TONSIL
COUGH & ASTHMA
MAXILLA
AURICULAR VAGUS NERVE
NEURASTHENIA
EYE
INTERNAL EAR
HELIX 5
LOWER BACK
ANESTHESIA FOR LOWER TEETH
CHEEK
EAR
TONSIL
HELIX 6

Figure 2-7: Ear Acupuncture Chart.

Helix points may be used to reduce inflammation, fever, and oedema. *Sanjiao* is used for circulatory, glandular, chest and abdominal problems. *Shenmen* is used for sedating, as with nervous problems, insomnia, mental disorders; it is the main point for pain relief, as in arthritis or inflammations; also for dry cough, bronchial asthma, epilepsy and hypertension.

Sterilise a suitable needle in methylated spirits and rub the ear with a cotton bud soaked in methylated spirits (optional). Find the most tender spot in the appropriate reflex area of the ear by pressing with a toothpick or a similar probe. Insert the needle to about 2 mm deep. It must not pass through the ear. Gently move it around to find the correct depth and direction of the sorest spot.

If it is very sore, do not move the needle. When the pain subsides, start to twirl it gently between thumb and forefinger. The less pain there is, the faster you may twirl the needle. If it is not sore any more, move the needle to a different angle and twirl, and in this way gradually work around in a circle. If all or most of the pain is worked out, withdraw the needle and select another point. If it was sorer in a particular direction, insert the needle at a point 1-2 mm away from the first point in that direction and repeat the twirling and probing in different directions.

Often the pain shifts around over a reflex area and you must insert the needle and work out the pain in several points lying close together. When the ear point is sore and still in need of treatment, the needle will feel like 'sticking' when it is twirled, while it will move freely when the tension is worked out. For treating chronically weak body parts that do not have sore ear points, select the appropriate point or area from the ear chart and -twirl the needle rapidly.

Press-point Therapy: Each ear contains a complete reflex map of the body, as shown in figure 2-7. Instead of using needles, you can more safely press the ear with the head of a dull pin, toothpick, matchstick, or fingernail. Be careful not to penetrate the inner ear. Treat the indicated organ points or wherever there is a tender spot; press for a minute or more each day or whenever convenient. In addition, it will be stimulating to the whole body to massage both ears thoroughly by rubbing them between thumb and index finger simultaneously. Massage daily. Pulling the ears at different angles is beneficial as well. You can also buy an electro-acupuncture appliance, now available (and not too expensive) at many health food stores.

Each ear contains a complete reflex map of the body as shown in the chart. Instead of using needles, you may press the ear with the head of a pin, a toothpick, a matchstick or a fingernail. Treat the indicated organ points or wherever there is a tender spot; press for a minute or more each day or whenever convenient. In addition, it will be very stimulating to the whole body to massage both ears thoroughly by rubbing them between thumb and index finger simultaneously. Massage daily. Pulling the ears at different angles is also beneficial. You may also buy an electro-acupuncture appliance. These are now available from many health food shops and not very expensive.

* * * * * * * * * * * *

Step 17

COLOUR THERAPY

Use colours to balance body conditions and your emotions.

Colour has a powerful influence on our body chemistry. The strongest effect is through the eyes, but good results are also achieved by treating the skin areas over certain organs or the spine.

Colour can be used in various ways: as direct coloured light, or indirect light from the colours of walls, curtains and clothes; in foods, and by treating drinking water with coloured lights; by visualising colours with closed eyes; or by visualising that the inhaled breath or directed energy is coloured. For treatment you should be careful to select bright, clear colours; murky colours usually have unpleasant effects.

Coloured Lights: Light may be coloured by using coloured light bulbs, by filtering sunlight through coloured glass or cellophane, even by wrapping coloured cloth around a lamp. Drinking water may be treated by exposing it to sunlight in coloured-glass containers or in containers covered with coloured cellophane, or coloured light bulbs may be used. This water is beneficial for drinking, gargling in the case of throat infections, washing wounds or bathing the eyes.

Preferably, stimulating colours should be used in the morning and sedating colours in the evening. If a sequence of coloured lights is used, start with stimulating colours, proceed to blue and finish with green. When treating internal organs, shine the light at close range on to the skin area covering the organ, possibly also on to the spinal reflex area. For general treatment and especially to improve eye conditions, you may look at a brilliant, reflecting surface of the desired colour, exposed to sunlight or a strong electric light.

To treat specific parts of the body, you may cover them with a single or double layer of cellophane of the desired colour and expose them to mild sunlight or strong electric light. If using sunlight you should be careful not to over-expose as cellophane does not absorb the UV radiation. As a general tonic, you may cover the back with cellophane in the colours of the rainbow: red at the base, orange over the lower back, yellow in the middle, green at the shoulder blades, blue at the neck and violet over the head. These are also the regional chakra energy colours

Green is the general healing colour, representing harmony and balanced body polarities. In disease, there may be over-stimulation, with too much red in the system - fever, inflammation, irritability, skin rash, red eyeballs, pink fingernails; or there may be under-stimulation with too much blue - pale skin, sluggishness, bluish eyeballs, bluish nails. The over-stimulated condition needs blue colour treatment, while the under-stimulated condition requires red or orange.

The following list shows the conditions for which the different colours may be used. You may also use muscle testing or a pendulum for individual colour selection. Different parts of the body may need different colours. If you are in doubt about which area to treat, irradiate 'systemically' - over the whole body. When irradiating the back, lie on one side, not on the stomach. For best effects or professional treatments, use colour slides with a strong projector in a dark room, and irradiate for one hour or longer at a time.

The Curative Properties of Colours:

RED (BRIGHT RED): Warm and stimulating. Use in anaemia, or whenever you are feeling cold and pale, with chilled extremities; also useful for ultraviolet and X-ray burns and frostbite; red is a liver stimulant and counter-irritant (to bring internal inflammations to the surface). Nutrients: vitamin B12, acidity.

ORANGE: A respiratory stimulant; useful for weak lungs, asthma, consumption, emphysema; also for overweight and under-active people with an under-active thyroid (energises thyroid, sedates parathyroids). Orange relieves spasms and cramps; increases milk secretion; corrects rickets and soft bones, osteoporosis; stimulates stomach and digestion, and relieves flatulence and distension; also stimulates the bladder, kidneys, liver and pancreas, try with epilepsy. It is valuable for chronic kidney disease and gallstones; arthritis, gout and rheumatism during non-painful periods. Minerals: boron, calcium, copper, selenium, silicon.

YELLOW: The most luminous colour. A strong stimulant and energiser for muscles and nerves: use in all cases of paralysis and muscle disorders, for nerve building, and in nervous exhaustion. Yellow energises the alimentary tract and stimulates digestion, flow of bile, gastric juices, pancreas activity, lymphatic system, heart, eyes and ears. It generally activates all bodily functions except those of the spleen, which is sedated; use for indigestion, constipation, diabetes, depression, for skin cleansing and to destroy intestinal worms. Minerals: magnesium, molybdenum, sodium.

LEMON (YELLOW-GREEN): Stimulates brain activity and activates the thymus. It is a mild laxative and expectorant (expels mucus): use in respiratory problems involving phlegm, asthma, coughs. Lemon is useful for dry, scaly skin disorders and bone building. It neutralises overacidity, and is indicated for chronic disorders in general, including cancer (systemic), diabetes and dwarfism. Minerals: gold, iodine, iron, phosphorus, silver, sulphur.

GREEN: Stimulates pituitary gland, encourages muscle and tissue building, and breaks up hardened cell masses and clots. It is anti-congestive, germicidal, disinfectant, purifying: use for heart problems, ulcers, cancer, wound healing, skin improvement, burn pains, glaucoma, headaches, neuralgia. Elements: chlorine, nitrogen.

TURQUOISE (BLUE-GREEN): Normalising for acute disorders. It is calming, and therefore useful for headaches, for soothing inflammations, sunburn and itching. It improves, rebuilds and tones the skin: use for moist- or weeping-skin disorders. It is also valuable for acute childhood diseases and infections, acute venereal diseases, and hyperactivity. Minerals: chromium, zinc.

BLUE: Cooling, strongly anti-inflammatory, antiseptic, stops bleeding; reduces fever, pain. Use to treat all conditions marked by redness, swelling and heat, and acute phases of arthritis, gout and rheumatism. Blue is relaxing and calming, slows the pulse, can reduce high blood pressure, and is valuable for underweight and overactive people. Use to treat itching, irritation, irritability, insomnia, throbbing headaches, speedy brain; sore throat, loss of voice, burns, bruises; diarrhoea, dysentery, colitis, neuralgia, sciatica, earache, biliousness and jaundice. Also good for treating tumour areas and red or infected skin disorders. Element: oxygen.

INDIGO (BLUE-VIOLET): Sedative, pain relieving, arrests discharges and bleeding, astringent. Use for piles and varicose veins. Indigo improves the immune system during infections: use to irradiate swollen lymph glands; in whooping cough with fever and other acute lung disorders; for convulsions and nervous irritation. It stimulates the parathyroids and sedates the thyroid; use locally for goitre.

VIOLET: Relaxing and soothing. It also stimulates the spleen and builds white blood cells. Violet sedates heart and muscle activity, and calms hyperactive people. It is useful for overactive kidneys and bladder; bed-wetting; nerve disorders and overactive mental disorders; epilepsy; insomnia. It also suppresses appetite.

PURPLE (BLUE AND RED): Dilates blood vessels and reduces blood pressure: slows the pulse as with palpitations (irradiate heart area). Purple lowers body temperature, and is excellent for high fevers; it stimulates the activity of veins; is a good sleep inducer for insomnia; decreases sexual desire and pain; is anti-malarial; sedates the kidneys.

MAGENTA (RED AND VIOLET): Energising; improves and normalises the emotions, the aura and the sexual system. Magenta energises the adrenal glands and the heart muscle, normalises the heart and blood pressure, and improves the circulation. It is diuretic; useful for kidney stones; complementary to green. Mineral: potassium.

SCARLET (2 PARTS RED AND I PART VIOLET): Narrows the arteries and raises the blood pressure; energises kidneys (diuretic); use as a sexual stimulant or for frigidity, and for suppressed menstruation. Scarlet accelerates the birth process and expels the placenta.

Photochromotherapy: Absent individuals may be treated (with their consent) by exposing their photographs or even a sample of hair or hand-writing to suitably coloured lights. Following is an arrangement that reportedly gave especially good results: line a box with crumpled and partly restraightened aluminium foil. Place the photo on the bottom, covered with a glass plate to prevent heat damage. Suspend from the top a small coloured light bulb, as used in Christmas decorations or similar ways. Cover the top with more aluminium foil treated in the same way.

Another possibility is to cover the aluminium-foil-lined box with cellophane of the desired colour and expose the box to moderate sunlight. Exposure time may range from one hour to several hours, depending on the severity of the condition and the reaction of the patient. A pendulum may possibly be used for assessing colours and exposure times.

* * * * * * * * * * * *

Step 18
MAGNET THERAPY
Take advantage of the fact that both poles of a magnet have useful biological effects.

Magnets are effective healing tools. The opposite poles of a magnet have different effects on the body, so it is essential to identify the poles correctly. According to physics, if a magnet is suspended on a thread, the pole pointing towards the north is called the north pole; the other pole is the south pole. So here, north pole means north-pointing pole. Another way of finding the polarity of a magnet is to move it slowly towards the north-pointing end of a compass needle. If the needle turns away, the north pole of the magnet has been used, and vice versa. Test the polarities of any acquired magnets, as you cannot rely on the pole designations of the manufacturer.

How to Use Magnets: U-shaped magnets are not suitable for magnet therapy. Instead, flat ceramic or neodymium magnets are best; otherwise short, cylindrical, or bar magnets can be used. The north pole generally has strengthening, stimulating qualities; it gives energy. It can be used to improve weak organs (indicated by dark markings in the iris). The south pole withdraws energy; it is sedating and relaxing. It can be used for pain relief, to reduce unwanted growths (tumours) and swellings, soothe inflammation, and arrest infections (indicated by white iris markings). The magnet does not have to touch the skin to work but can be kept in a cotton bag pinned to the clothing, the specified pole facing the body.

Most therapeutic magnets are composed of ferro-ceramic material (usually called ceramic magnets) and break easily when dropped. The therapeutic strength of a magnet is a product of its gauss rating and its size. In physics, gauss is the unit of magnetic strength. The magnetic field of the earth is about 0.5 gauss; a magnet with a weak rating has several hundred gauss and a strong magnet several thousand gauss. However, even a magnet with several thousand gauss can have a weak therapeutic effect if it is small. To give an example of commonly used magnets: A strong ceramic magnet can measure three by two by one inches, and a weak magnet two by one by one-half inches, both with 4,000 gauss at the surface.

Very strong magnets are made with neodymium, a rare-earth element, and have 9,000 gauss or more. Neodymium magnets are relatively small, and their fields do not reach far into the body. Therefore, neodymium magnets are commonly used to treat areas near the surface, as in small disk or button magnets for acupuncture points. Ceramic magnets, on the other hand, can be large and heavy, their effects reaching far into the body. Therefore, with pain or a tumor deep inside, apply a strong ceramic magnet, three by two by one inches or bigger, while for near the skin use a neodymium block magnet of one by one by one-half inch.

As a general rule, apply strong magnetic fields for acute problems and weak fields for chronic problems. In both cases, you can apply a magnet either continuously or intermittently during the day until the problem is resolved. For acute problems, this can be for days and with chronic problems for weeks or months. While it is more convenient just to leave the magnet in place all day long, the effect seems to be stronger if you frequently alternate applying and removing the magnet.

If a magnet is used for too long or in the wrong place, you may feel uncomfortable or drained of energy. If this happens, remove the magnet immediately and apply the opposite pole for a while. Alternatively, place your hands on a large zinc plate or take homeopathic zinc 6X as an antidote to magnetic overexposure.

Do not use strong magnets directly on the eyes or heart, or on the abdomen during pregnancy. Preferably attach or remove strong magnets slowly. Strong magnetic fields suppress nerve signals and, therefore, either pole of a powerful magnet can eliminate pain. However, farther away from the magnet, the field weakens, and then the effects of the opposite biological pole become noticeable. Therefore, I recommend always using the correct pole.

If there is no contraindication due to pain, inflammation, or congested energies, then the preferred magnet placement for all individuals, male and female, appears to be the south pole towards the right front of the body and the north pole towards the left front. To a lesser degree, the back of the body has the opposite polarity to its front. This means that the energy of the left back is similar to that of the right front but not quite so strong. Therefore, if a north magnet is required on the right side of the body or a south magnet on the left, then it is best placed on the back.

If there is pain or any sign of inflammation in an organ, apply the south pole; if there is an obvious weakness, use the north pole. In most other cases, experiment with muscle testing to find the beneficial pole for the condition at hand. If you are in doubt about which pole to use, try one pole for some time and later the other pole, then compare results. The length of a testing period, before reversing poles, can be only a few minutes for acute cases or several days for chronic conditions.

While some people do not feel the magnet’s energy and show little response, others are hypersensitive to it. In the latter case, use only a weak magnet for a short exposure or try magnetized water instead. Do not use a magnet in a bed with an innerspring mattress, steel frame, or steel base. In this case, the steel parts can become strongly magnetized with the wrong polarity and cause deterioration of the disease condition when the treated person lies on the bed.

Magnetized water is also beneficial for general use. Place one bottle on a strong south pole and another on a strong north pole for 24 hours, then drink the contents of each bottle separately. Alternatively, place a north pole and a south pole at both sides or at the top and bottom of a non-metallic container filled with water. For specific conditions, use either north-pole or south-pole treated water on its own, for drinking or external application (for example, as an eyewash). A strong balancing effect can also be achieved by drinking north-pole treated water and south-pole treated water on alternating days. Use muscle testing or a pendulum to decide, or experiment and observe the results.

For general improvement of your well-being, place strong magnets at the hands and/or feet; select poles according to the general rules. For problems in the upper part of the body, place the right palm on one pole and the left palm on the other. For problems in the lower part of the body, place one pole under the right foot and the other under the left foot. If mainly one side is affected, the palm can rest on a south pole and the foot of the same side on a north pole. Finally, a good general placement is to have the right palm on one pole and the left foot on the other pole, or vice versa. Start using strong magnets in this way for five minutes and gradually increase to 30 minutes, or for as long as it feels beneficial.

A strong magnetic flow is sometimes beneficial for conditions of mild pain associated with weakness, such as weak shoulders that are sore when moved. Magnetic energy is often strengthening to the sex organs and lower pelvis; apply a north pole slightly to the right on the lower abdomen and a south pole slightly to the left on the lower back. For these flows, strong magnets can be kept in place for up to an hour, or for as long as it feels beneficial.

Using the Correct Pole: As pointed out, the south pole is placed towards a painful, inflamed, or infected part of the body, while the north pole is used if the body area is weak or underactive. However, there is a possibility that the weakness or underactivity is caused by a low-grade chronic infection or inflammation, which means it may still improve with the south pole and deteriorate with the north pole. Therefore, the south pole is generally safer to use.

Commonly, one side of the body has more pain, inflammation, sensitivity, or contracted, tight, or tense muscles than the other. When placing magnets under the palms or feet, place the south pole here and the north pole on the other side. Another good possibility is to lie down with a strong north pole towards the soles of the feet and a strong south pole towards the top of the head; put both magnets about four to 12 inches away from the body. You can also experiment with sleeping this way.

With caution, you can also sleep with a magnet placed directly at the crown of your head. You can move a strong magnet (e.g., 4,000 gauss, three by two by one inches) progressively closer during successive nights or sleep at first with a weak magnet (e.g., 4,000 gauss, two by one by one-half inches) and gradually increase its strength. This may be most beneficial for the elderly or those with neurological diseases. For this application, the body should lie in a north-south direction with the head pointing north. When using strong magnets on yourself while in your bed, make sure that your bed and mattress cannot be magnetized.

William Philpott, M.D., a magnetic re-searcher in Oklahoma, has described cases of increased beneficial melatonin and growth hormone production in elderly individuals sleeping with a strong south pole directly at the crown of the head. Also, symptoms of confusion, psychosis, schizophrenia, and epileptic seizures disappeared with south poles placed at the temples, Dr. Philpott has reported.

To find which pole to use, use muscle testing (see step 13). Sit down with one or both magnets placed in the questioned position. Raise a free arm and stretch it out horizontally in front of you with the palm down. Have a tester or helper gradually exert increasing pressure on the wrist of the outstretched arm to see whether it locks in position or gives way. Then reverse the poles of the magnets and test again. Either the tester or the tested person (or both) may notice that the arm is stronger with a particular placement of the magnets. This is the correct placement. The same test can be made while holding a jug with either north-pole or south-pole treated water pressed lightly against the abdomen.

Removing Congested Energy: When the south pole of a magnet, especially a strong one, has been applied to a painful or congested area (including one with unwanted growths), an effective way to remove it is to place your left hand between the magnet and the body while slowly removing the magnet. Keep the hand close to the magnet. Then hold the south pole close to a water surface or near running water for about ten seconds; rinse your hands under running water.

This process of energy removal from a painful area can be repeated several times. The south pole can be circled over the congested area, starting with a large circle close to the skin and gradually moving higher in spirals above the body. When the magnet is four to eight inches away from the skin, place your hand under it and withdraw the magnet from the body.

Examples of Correct Use of Magnets: Remember the simple rule that the north pole gives energy while the south pole withdraws it. An inflammation, a tight muscle, or a swelling are energy congestions and need the south pole; an overly relaxed muscle needs more energy and consequently requires the north pole. Here are some practical tips:

Shoulder: Frequently, one shoulder is dropped while the other shoulder is raised. The lower shoulder needs the strengthening north pole and the higher shoulder needs the relaxing south pole to balance the forces. The same rule applies to the middle part of the spine if there is any sideways distortion (as in scoliosis).

Gallbladder: In the case of a gallbladder problem, the left leg is often shortened, pulled into the hip socket by tight muscles. The south pole applied to the hip can help correct this condition. After the magnet has been in place for some time, gently pull the leg at the ankle until both ankles are exactly side by side. The right leg can be stretched in the same way, if necessary.

Tumor: Use a strong south pole over a tumor; apply for two hours twice daily or for one hour three times daily. Hold the south pole close to running water after removal. However, with pain from a tumor or cancer, you may keep the magnet in place as long as required, even for days on end.

Wounds: Fresh wounds heal better with the south pole applied initially; later, use both poles in alternation; and in the final stages of healing, use the north pole for strengthening.

Broken bones: For broken bones, the north pole should be applied above the fracture (towards the torso), while a south pole should be applied opposite the north pole and below the fracture. Alternatively, treat the same way as for wounds.

Headaches: For headaches, use the south pole wherever the centre of the pain appears to be; but for general mental refreshment, attach the south pole to the forehead and the north pole to the base of the skull. This may also improve your psychic awareness.

Stomach and duodenal pain: For stomach and duodenal pain, place the south pole over the pit of the stomach, but to strengthen the digestion after a meal, apply the north pole.

Weak heart: To strengthen a weak heart, a small magnet may be carried in the left shirt pocket with the north pole near the breastbone. If there is pain, use the south pole instead. Check your pulse before and afterwards to see whether it becomes stronger and more regular. Even if the pulse does not improve, try using it the magnet for several days (during the daytime only) to see if you feel better.

Sexual energy: To increase sexual energy, sit on the north pole of a strong magnet. This improves the energy centre at the base of the spine as well as the sex glands. However, with vaginal candidiasis, sit on a strong south pole.

Lowering blood pressure: To lower the blood pressure, apply the south pole to the right side of the neck, over the carotid sinus. This gland is located above and to the side of the Adam's apple and just below the jawbone where you feel the main artery into the head. To raise low blood pressure, apply the north pole to this spot.

Overweight: If you are overweight and underactive, apply the north pole to the thyroid gland for stimulation; if overactive, use the south pole to slow the metabolism.

* * * * * * * * * * * *

Step 19

THE ELECTRONIC ZAPPER & MAGNETIC PULSER

Use electrotherapy to keep your body free of undesirable microbes and parasites.

Electrotherapies utilize low-voltage electricity to disable a wide range of microbes, including viruses, fungi, bacteria, and parasites, and have been shown to be helpful against a wide range of infections. Various patents have been issued for such devices and scientific papers published with the results. Tests conducted with the AIDS virus in the laboratory show that the ability of the virus to infect human T cells is reduced in direct proportion to the product of the magnitude of the current and the time that the virus is exposed.

An oscillating magnetic field has been used with similar good results to kill microbes in food samples or living tissue. Numerous papers in the International Journal of Neuro-science show the positive effects of very weak pulsating magnetic fields applied directly to the top of the head on neurological diseases, such as multiple sclerosis and Parkinson's disease.4 Other researchers have used magnetic pulses on the brain to successfully treat depression.5 Even bites from poisonous snakes have been neutralized with electrotherapies without any negative after-effects.

Electronic Zapper and Blood Purifier: There are now many types of electronic antimicrobial devices available. These originated in the 1990s from the Hulda Clark "zapper" and the Robert Beck "blood purifier." Hulda R. Clark, N.D., is a Canadian microbiologist, and the late Dr. Robert C. Beck (D.Sc.) was an American electro-engineer. Both devices appear to eliminate parasites, bacteria, fungi, and viruses from the body with low-voltage, pulsed electric currents. The Clark zapper operates with 9 V at a very high pulse frequency of 30,000 Hz (Hz is the unit of electromagnetic wave frequency, equal to one cycle per second). The Beck blood purifier uses 27 to 36 V and a low frequency of 4 Hz. It appears that microbes are not killed outright by the mild electric current but are changed in such a way that they are prevented from replicating.

Evidence suggests that most of us harbor hidden microbes, and electrotherapy is an easy way to reduce or eliminate these. Therefore, I recommend electrotherapy as a basic component of a health improvement program, in the same way as I recommend intestinal sanitation and allergy testing, regardless whether there is an obvious problem or not. In addition to other measures, I regard electrotherapy as valuable in treating infections and infectious diseases.

With the Clark zapper, you hold an electrode in each hand through which a weak current flows. To establish firm contact, rub your palms with saltwater and hold the electrodes firmly. Alternatively, wrap the electrodes in absorbent paper or cloth moistened with saltwater. With weak hands, the best contact results from dipping each hand or foot in a bowl of saltwater along with a partly immersed electrode, but only if no electronics are in the handle. After the last treatment of the day, rinse your hands and/or feet under running water to remove static electricity.

If you are treating a local infection or tumor, press one or both electrodes close to the problem area; for instance, for a sore throat, press an electrode against each side of the neck. The negative electrode, commonly attached to a black wire, tends to sting sensitive skin. As with magnets, it is important in which hand you hold a particular electrode; otherwise you may feel weaker after zapping. Hold the negative electrode on the side of the body where the tumor, pain, infection, or inflammation is located, and hold the positive electrode with the red wire in your other hand. Try muscle testing to find the best electrode placement. Have someone gently press down your raised, outstretched arm while holding the electrodes one way and then the other. Use the electrodes indicated by the stronger test.

The normal treatment schedule is seven minutes, three times daily. The second zapping should follow 20 to 30 minutes after the first one, and the third one after a similar interval. (For more information, see Hulda Clark's The Cure for All Diseases; www.drhuldaclark.org.)

I have found the Clark zapper useful, but the Beck zapper works better in conditions where the blood is infested with microbes or parasites. It appears that with the Clark zapper the current travels mostly over the skin due to its high frequency, while the Beck zapper is specifically designed to electrify and purify the blood; therefore, it is sometimes called a blood purifier.

With the Beck zapper, the small electrodes are fastened directly over the arteries, so you feel a pulse on one forearm or near the ankle on each foot. My recommendation is to use the blood purifier initially for two hours daily for about two months, and then for one month once a year. I am not aware of any dangers or contraindications, but if you are unusually electro-sensitive or a problem arises, then stop treatment. As with all natural therapies, electrotherapy may start a healing crisis; in that case, interrupt treatment until you feel normal again.

Dr. Beck warned that cells may take up medical drugs and herbal medicines at a much higher rate during the zapping, due to a process called electroporation; this may cause drug overdosing. Originally, he recommended not using drugs and other potentially dangerous remedies, including raw garlic, during the weeks and months of treatment. However, in later interviews, he softened his stand and advised cautiously taking needed drugs or potentially dangerous remedies after the daily zapping rather than beforehand. There is no firm evidence that electroporation occurs with the electronic zapper, so I believe Dr. Beck may be exercising caution.

Application of the Magnetic Pulser: Dr. Beck also designed a magnetic pulser to treat swollen lymph glands and other sites of infection and tumours and other parts of the body with poor blood supply. A strong but very brief magnetic pulse induces a tiny current in the cells and body fluids. In addition to helping eliminate microbes, parasites, and cancer cells, it also appears to recharge and rejuvenate weak body cells. This technology is based on U.S. patents in which laboratory test results were submitted showing inactivation of HIV and destruction of cancer cells.

The most notable effect that I have observed in using the magnetic pulser is pain relief, be it from back or hip pain, toothache, or cancer, but it was less effective with headaches. Often the pain did not come back. Also, weak eyes and weak muscles often improved. Whether you have a disease or want to improve your health, I recommend you use the pulser frequently on any body area where you suspect an inflammation or infection. Use it on swollen lymph glands (tonsils) and any tooth or jawbone positions that are not completely healthy, e.g., where a tooth has been extracted or where there is a filling, a root pocket, or receding gums. Pulse all the lymph glands in the neck, armpits, and groin. Hold the pulser as close as possible to the treated problem area.

Commonly, a pulse occurs every two to five seconds and treatment time ranges from ten to 30 minutes. After ten to 20 pulses, you can move to the next area, treating in turn lymph glands, tumours, tender joints, teeth, organs, or any area with an infection. However, remain much longer at major problem areas, up to 30 minutes if required, as for pain relief. A recent development is a magnetic pulser with up to ten or more pulses per second. It appears to work by creating a pulsing magnetic field rather than just single short pulses. I found this more effective than the slower pulsers.

There are other types of weak magnetic pulsers available that may be helpful in stimulating cellular rejuvenation. For example, fields of 2 to 7 Hz are suitable to treat neurological disorders, but because of their much lower magnetic pulse intensity, they do not appear to inhibit microbes or parasites and may stimulate their growth. To avoid any damage to normal body cells, the Beck pulser is set at the lower end of magnetic strength to disable microbes and parasites. For this reason pulsing may have to be continued daily for weeks or months to overcome serious and long-standing health problems.

Another important aspect is to be aware of the polarity of the coil. One side acts as a magnetic north pole and the other side as a south pole. It is important to use the correct pole as explained previously in Magnet Therapy. This may not be noticeable with strong magnetic pulses, as both sides can suppress nerve signals of pain and improve general health by recharging damaged electric cell potentials. However, there are other conditions that can get worse from using the wrong pole, while using the biologically correct pole can also give better results with pain, microbes, and general health improvement, especially farther inside the body where the pulse strength rapidly declines. When holding the back of the magnetic pulse unit, called the paddle, hold it with the hand that benefits from that particular polarity.

To find out which side of the paddle is north and south, place a small flat magnet with known polarity on top of one side and watch what happens with the next pulse. Like poles repulse each other. Therefore, if the magnet jumps off, that side of the paddle has the same polarity as the side of the magnet placed on it face down. If you then turn the small magnet around, it should not jump off. Mark the polarities on the paddle. If you do not have a specific reason to use the other side, then use the south-pointing side of the paddle facing the body. However, if you do not get results within an appropriate period, then try the north-pointing pole instead.

For further information on electro and magnetic medicine, including articles by Dr. Beck and plans to build the zapper and pulser, see www.quantumbalancing.com ; www.keelynet.com/biology.htm; and www.sotainstruments.com.

* * * * * * * * * * * *

Step 20

SPINAL THERAPY & MASSAGE

Improve the activity of all your organs and glands by massaging your spine and body.

Massage is an effective and usually enjoyable healing tool. It helps one to relax, loosens up tight muscles, and removes fluid accumulations and congestions. Massaging close to the spine is especially effective to stimulate the activity of internal organs and glands. This is due to the network of nerve connections emanating from the spine and the multitude of important acupuncture treatment points along the spine.

There is a reflex connection between every organ system and the spine. The malfunction of an organ will produce tension in a specific part of the spine, while a misalignment of the spine will cause tension in a specific organ. Additionally, various acupuncture points along the inner branch of the Bladder meridian, close to the spine, are linked to specific organs, and on both sides of each vertebra (less than one inch from the midline) there are points not usually included in the standard meridian system diagrams but related to certain regions of the body.

The Benefits of Massage: A spinal massage is best given as part of an overall body or back massage. For an initial general massage, you may (or may not) use massage oil. It is helpful to know special massage techniques, but not essential. Rub, stroke, and knead as it comes naturally. A caring attitude during massage is more important than technique.

Glide slowly downward with your thumbs on each side of the spine (fingers or knuckles may be used as well). Press firmly for a moment into the muscle, then glide to the next vertebra. When you find a sore spot, ease up on the pressure, and press for several minutes with a slightly circular motion. Increase the pressure when the pain subsides. Repeat this several times, gliding down the spine, pausing at tender areas. Do not rub directly on the vertebrae.

Other body points typically in need of special attention are the base of the skull, the neck and shoulders, the tops and centres of the shoulder blades, the buttocks, the hips and the sides, and the back of each knee.

Another technique to treat tender muscle points along the spine is to apply constant but moderate finger pressure on each tender spot for about 90 seconds. Direct the pressure sideways against the spine, and try to relax the treated muscle to ease any pain caused by a strong pressure. When treating the upper part of the spine, lift the shoulder; when treating the lower part, raise the hip. For the neck, gently move the head against the finger pressure. Experiment to find the position in which the treated muscle is most relaxed. The patient should remain passive during the treatment and should not try to help move any body part.

Give your partner a massage and pressure massage once a week; this is best after a bath or shower. This is an excellent exercise for developing a harmonious relationship with a friend or partner. When using long strokes during massaging, follow the flow directions of the acupuncture meridians (see Part 7).

Massage Oils: Topical massage oils are commonly used to let the hands glide over the skin in long strokes. However, for pressure therapy, deep muscle work, and for sensing the energies, it is often better not to oil the skin. Massage oils can also be selected according to the specific needs of the massaged person. The American psychic Edgar Cayce offered the following recommendations:

- General Tonic Massage: Peanut oil, 6 parts; olive oil, 2 parts; dissolved lanolin, 1 part; rosewater, 2 parts; shake well before use.
- Peanut Oil and Olive Oil Mixture: 2 parts of each and 1 part dissolved lanolin; this is good for arthritis and rheumatism, the after-effects of anaesthesia, injuries from accidents, kidney disorders, menopausal complaints, multiple sclerosis, prostatitis, and toxaemia.
- Peanut Oil Alone: Good for apoplexy, arthritis, gallbladder inflammation, coronary occlusion, fatigue, glandular disturbances, low vitality, menopause, multiple sclerosis, palsy, paralysis, Parkinson's disease, polio, poor circulation, and ulcerated stomach. Preferably, first heat the body with an infrared lamp or by exposing it to sunshine during and after the oil rub.
- Castor Oil Alone: Good for arthritis, back pain, contractions and spasms, rheumatism, and all muscular and joint pains. Rub very warm castor oil into the affected parts, preferably using an infrared lamp or sunshine to heat the skin so as to help work it into the skin.
- Removal of Scar Tissue: Combine 1 to 2 parts camphorated oil with 1 part peanut oil; massage into the scar tissue once daily for several months. Alternatively (or in addition), use vitamin E oil.
- Cocoa Butter: This stimulates the circulation in massaged areas and strengthens the nervous system and eliminating functions. It is especially useful for massaging babies and young children gently along the spine. It helps body development and guards against head congestion. For increasing the size of underdeveloped breasts, massage around the glands under the arms and below the breasts. Direct massage of breasts with cocoa butter combined with alum water has been recommended by Edgar Cayce to decrease their size. You may use melted cocoa butter on its own or as part of a massage-oil mixture.

Add various aromatic oils for their fragrance or for any additional healing effect they offer. The olive oil should be extra-virgin, and the peanut oil cold-pressed; store both in a cool, dark place. Squeeze some vitamin E oil capsules into the massage oil. A small amount of massage oil can be lightly warmed before applying it to the body. I am not in favour of using polyunsaturated oils as massage oils as these easily become rancid.

Treating Muscle Pain: Muscle pain is commonly due to a contracted or spastic muscle. This can generate pain from pressure on a nerve or from an inflammatory condition resulting from overacidity in the tissues. The original cause may be an injury or overuse of the muscle, but this causes lactic acid accumulation in the muscle, which prevents calcium from moving out of the muscle fibres to allow them to relax. The blood and lymph circulation through the contracted muscle is greatly diminished. Therefore, lactic acid is removed very slowly by the body and oxygen supply to the muscle is low. Hence the pain.

In this condition, further use of this muscle generates more lactic acid and keeps it contracted and weak or painful. If this condition becomes permanent, the area tends gradually to calcify, such that two vertebrae may fuse or a joint may become immobile. The solution is to use suitable therapies to relax this muscle and increase the blood and lymph flow through the area, as with the following: Have a period of rest to allow the lactic acid to disperse. Heat, such as a hot pack, hot bath, or rubbing, greatly speeds up this process. Adding Epsom salt or magnesium chloride to the bath aids muscle relaxation by displacing some of the calcium in the muscle with magnesium. Adding baking soda to a bath or pack helps reduce the overacidity of the muscle. Magnesium and/or an alkalizer may be used internally; MSM, for example, taken internally as well as applied externally, generally speeds up these processes.

A quick but painful way to relax a tense muscle is to press into it for several minutes. In addition to or instead of this, a deep muscle massage greatly speeds up lymph and blood flow through the muscle; press, knead, and rub the problem area. Other supportive measures include color therapy with blue light, the south pole of a magnet, a pulsed magnetic field, reflexology, and acupuncture and meridian therapy (see appropriate steps elsewhere in this book).

Aiding Spinal Correction: Frequently, health problems are intensified by a misalignment of the spine. Hereditary factors, ingesting incompatible food, and negative or suppressed feelings can cause organ and gland functions to become overactive or underactive. This in turn weakens or tenses muscles in associated reflex areas of the back. The same may happen because of one-sided strenuous muscle activity. This causes an uneven pull of muscles on the spine, and an individual vertebra can easily become misaligned, causing pressure on nerves emerging from the spine and consequent pain or malfunction.

Therefore, most people with health problems will benefit from an initial professional adjustment of the spine. However, if the muscles are not balanced simultaneously, the correction will not last and you may become a regular visitor to a chiropractor. This is not necessary. Try to see an osteopath or chiropractor who does muscle balancing as well as adjustments. Working with a friend, you can easily balance each other. Here are some tips:

- For backache, first find out, by pressing, which side of the back has more tenderness.
- Trace the Bladder meridian of the tender side with a south-pole magnet (or the energized fingers) from toe to head several times. Trace the Bladder meridian on the other side from head to toe with a north-pole magnet. For neck, shoulder, and hip problems, trace the Gallbladder meridian in the same way (for charts of the meridians, see Part 7).
- Apply ear acupuncture, reflexology, or a strong south-pole magnet to the painful area.

- Press into the tender muscle until the pain eases, then press harder and circle the thumb, knuckle, or elbow that is applying the pressure.
- Check for other tender points in the same area and work these out in the same way. Often misaligned vertebrae will then realign themselves during subsequent back-stretching exercises or with head-and-neck exercises.

As a general rule, neck and shoulders are affected by the organs and meridians of the gallbladder, upper back by the lungs and heart, the middle by the stomach and pancreas, the lower end of the rib cage by the kidneys, and the lower back by the intestines and sexual organs. Liver problems often manifest as a pain near the lower right shoulder blade.

One of the best methods for back improvement is regularly hanging upside down, from special inversion equipment or a ladder. Traction on the neck improves the upper spine.

How to Apply a Spinal Concussion: This is an effective method for stimulating and strengthening weak organs and body parts. The general principle is to stimulate the nerves emerging from the spine with rapid soft taps to certain vertebrae. Use the edge of your hand or the knuckles to tap the selected vertebra rapidly (but not very hard) for about 30 seconds. Then rest or treat another vertebra for 30 seconds and return to the first one. Repeat this process for five to eight minutes.

For general health improvement, go the length of the spine and stimulate each vertebra in turn with this method. Tap firmly at the lower spine but only lightly at the neck. Let your patient or partner tell you how hard to tap so that it still feels invigorating and pleasant. This treatment can be repeated daily, weekly, or when feasible. If tapping produces pain, seek professional advice. If sedation rather than stimulation is required, try prolonged gentle pressure, gradually varying in intensity, on the appropriate vertebrae. Select suitable vertebrae from the following compilation.

Selecting the Vertebra Based on Its Associated Health Condition: Here are some general guidelines for this. There are seven cervical vertebrae, 12 dorsal or thoracic vertebrae, and five lumbar, plus the sacrum and coccyx.

Cervical vertebrae:

1: head, brain, pituitary, sympathetic nervous system

2: eyes, tongue; hearing problems, sinuses, allergies, fainting

2-3: for headaches, prolonged pressure between C2 and C3

3: cheeks, teeth, outer ear; acne, eczema, neuralgia, neuritis

4: nose, lips, mouth, Eustachian tube; catarrh, hayfever

5: vocal cords, neck glands; throat problems, pharynx

6: neck muscles, shoulders; tonsils, upper arm problems

7: thyroid, goitre, shoulders, elbows, contracts inner organs, contracts dilated heart, normalizes blood pressure; bursitis, nose bleeding, fainting, angina, palpitation, tachycardia (fast pulse), lung and kidney diseases, diabetes. (C7 is the most prominent vertebra at the base of the neck. Use it as a landmark for counting the other vertebrae along the spine.)

Dorsal or thoracic vertebrae:

1: lower arms, hands, oesophagus, trachea; asthma, cough, breathing problems, thymus

2: heart, coronary arteries, lungs; chest pain

2-3: for hiccough, press between D2 and D3

3: lungs, breasts; increases milk flow, contracts gallbladder, oesophagus, and pylorus; dilates heart and peripheral vessels, reduces blood pressure

3-4: tap both for developing breasts

4: contracts and empties gallbladder, increases secretions of pancreas; jaundice, hepatitis, shingles

5: liver, solar plexus, opens pylorus and empties stomach; low blood pressure, poor circulation, anemia

6: stomach; indigestion, heartburn, dyspepsia

6-7: tapping both dilates the kidneys; nephritis

7: pancreas, duodenum; diabetes, ulcers, gastritis

8: spleen, diaphragm; hiccoughs

9: adrenal glands, dilates gallbladder; allergies, biliary colic

10: kidneys; dilates pancreas and blood vessels, reduces blood pressure; nephritis, ataxia, anemia

11: kidneys, urethra; dilates heart, stomach, liver, spleen, abdominal arteries, intestines, in-creases blood supply to lungs; spasms, nervous diarrhea, angina, skin problems

12: small intestines, Fallopian tubes; contracts kidneys; affects lymph circulation; associated with enlarged prostate

Lumbar vertebrae:

1: large intestines; constipation, diarrhea, colitis, hernia

2: appendix, abdomen, upper leg; cramps, varicose veins

3: sex organs, ovaries, testicles; bladder, knee, menstrual problems, impotence, bed wetting/ incontinence

4: prostate, lower back muscles; sciatic nerve, lumbago, and backache; difficult, painful, or too frequent urination

5: lower legs, feet; contracts bladder; leg cramps, cold feet

Sacrum: hipbones, buttocks; sacroiliac conditions

Coccyx: rectum, anus; haemorrhoids, pruritus

Lymph-Drainage Massage: Danish medical doctor E. Vodder developed a method of manual lymph-drainage massage in the early 1900’s. It is designed to improve lymph flow and accelerate the transport of waste products from the tissues into the bloodstream from which it can be eliminated through the kidneys. A congested and stagnating lymphatic system is a main cause of infections, inflammations, and degenerative diseases.

The lymph flows upwards through the legs to the large lymph nodes at the groin and up the arms to the main lymph nodes at the armpits. From the groin and abdomen, the lymph collects in the thoracic duct, which flows along the midline of the body and, together with the lymph from the upper left side of the body, empties into a large vein under the left collarbone. The lymph from the upper right side of the body joins a vein under the right collarbone.

Lymph massage should start and end with the chain of glands at the sides of the neck, running from below the ears to the hollows of the collarbones. For massaging the limbs, Vodder taught light, slow, circular movements with the four fingers of the hand, not the thumb. Movement can be clockwise or counter-clockwise (whichever feels better), but do not change direction during the massage. The skin should be moved lightly over the underlying tissue. Generally, five circles are made at any one place, one circle per second. Vary the finger pressure during the circling to achieve a pumping effect. Then move about two inches along the limb and repeat the circling. During massage, a limb should be raised, while well supported and relaxed, to help the backward flow of the lymph.

When massaging an arm, start at the upper arm, near the armpit, and press-circle ten times. Then do the same at the insides of the elbow, and finally squeeze each section five times from the palm upward to the armpit. Do the same with each leg. Squeeze ten times near the groin, then the back of the knee, and finally from the foot upwards to the groin.

Along the trunk, from the groin upwards, use long, firm upward strokes towards the left collarbone; do this also along the spine and the sides of the back. However, the main benefit for the trunk lymph drainage will be derived if the patient lies on his back and makes bicycling movements in the air. The movements must be either parallel to the floor or, better still, vertical while in the shoulder-stand position. Do this exercise for one minute before and after leg-drainage massage to make room in the trunk for lymph from the legs. Varicose veins, wounds, and tumours should not be massaged. Individuals with advanced diseases should activate the organs of elimination before using this method. If in doubt, consult your health professional.

* * * * * * * * * * * *

Step 21

HYDROTHERAPY, PACKS & COLONICS

Use water to relax or alleviate painful conditions.

Hydrotherapy, the use of water to treat disease, involves practices such as drinking large quantities of water to flush out toxins, taking hot and cold baths, saunas, packs, and colonics (enemas). Preferably, use non-chlorinated water if available, as it is less irritating to the skin.

Hot Baths: Hot baths are useful for relieving aches, pains, and stiffness, for releasing toxins stored under the skin, and for cleansing the skin of fungal infections. Gradually increase the temperature of the bath water, as far as you can reasonably endure. Remain submerged, except for the face, for ten to 20 minutes. For conditions associated with magnesium deficiency, add from one to ten pounds of Epsom salt to the water; the greater the pain, nervous tension, or stiffness, the larger the amount you should use.

Become accustomed to hot baths gradually by increasing the duration and maximum temperature of the bath on successive occasions. Have someone on call, especially if you have a heart problem. Do not use a hot bath or Epsom salt in the case of low blood pressure, fatigue, distended veins, or a weak heart, and be cautious if your blood pressure is high. The body should be completely immersed, but you can cool the forehead with a cold cloth.

To improve dermatitis, place a sealed muslin bag filled with up to two pounds of bran in a hot bath; immerse yourself, as above. To relieve itching skin, add a spoonful of baking soda to the water. For arthritis, if a full hot bath is too severe for you, have a hot foot or hand bath with Epsom salt to get relief.

After your bath, cover your body with a large towel and, without drying, cover yourself up in bed to induce sweating; rest for 30 minutes, or overnight. Afterwards, a massage or a quick rub with olive or peanut oil is recommended.

Overheating Baths: The following method can be used for your general health improvement and to overcome chronic infections. However, I do not recommend it as a specific cancer treatment. It appears to work only when the blood sugar level during the overheating is low (i.e., you haven't eaten for a while) and the water temperature is high; otherwise tumor growth can be stimulated. For those without cancer, it is good occasionally to have an overheating bath followed by sweating - in other words, a bath that will overheat you. Cancer cells and viruses are weakened and start dying in environments over 105° F (40-41° C), while normal cells can withstand higher temperatures. This explains the curative effect of high fever on cancer and infectious diseases. For home treatment, a therapeutic fever can be simulated with overheating baths.

Start with a very warm bath and slowly add more hot water until the bath temperature reaches 108° F (42° C). Initially remain in the tub only up to ten minutes at the maximum temperature, but gradually extend this to 20 to 30 minutes. If you want to continue with a regular overheating treatment, then increase the temperature by 2º F (1º C) each week up to a maximum of 116º F (47º C).

If the weekly temperature increase is too severe, increase it more gradually. If you cannot reach the maximum temperature, extend the bath duration at your highest possible temperature to one hour. During and after the bath, drink several glasses of pure water to make up for fluid losses.

As a safety precaution, keep the bathroom door open and have a helper close by. You are safe while lying in the water; the danger arises when leaving the bath.

Therefore, strictly observe the following rules:

- If you must leave the bath suddenly, immediately lie down again.
- Remain in the bathtub until all the water has drained. Alternatively, add cold water until the bathwater feels only warm; stay in the cooler water for 10 to 15 minutes.
- Dry yourself only while sitting, never standing; afterwards, cover with sheets and blankets and sweat for up to one hour.
- Do not use overheating baths when you do not feel well.

Warm Baths: These are excellent for relaxation; stay in a warm bath for ten to 20 minutes; fragrant oils can be added. Staying in a hot or warm bath for longer than 30 minutes can be exhausting. If stimulation is desired, a warm or hot bath should be followed by a cold bath, a cold sitz bath, or a cold shower. You can also use hot and cold alternating showers or cold-water sponge rubs. Cold-water stimulation is especially important for low blood pressure and distended veins.

Sitz Bath: To improve intestinal elimination and the condition of the sexual glands and organs, a cold sitz bath is excellent. After warming the body by exercise or a hot bath or shower, sit for one minute in about eight inches of cold water. Gradually, the sitting time can be increased to three minutes. Keep your feet out of the water.

Showers: For general stimulation and improving the circulation, use hot and cold showers, alternating three times between hot and cold, starting with hot and ending with cold.

Foot Baths: To increase the circulation in burning or aching feet, mix some tannic acid or mustard powder into a hot foot bath. To keep your feet warm, some cayenne pepper can be added instead. Use alternating hot and cold foot baths for leg problems.

Water Packs: To make a pack, fold a cotton or woollen cloth into several layers, dip it into hot or cold water, gently press to remove excess water, and place over the selected area. Hot packs (fomentations) can be used for the relief of muscle pains but are better in combination with Epsom salt. Cold packs are effective for insomnia, high temperature, nervous indigestion and wind, headaches, sore throat, edema, and sprains. For fever, the pack can be changed every five to ten minutes, while for relaxation, it can be left in place for a long time. For insomnia and nervous indigestion, place the cold pack on the abdomen.

Castor-Oil Pack: This is excellent for improving intestinal functions and for stimulating the immune system, as well as in problems with the sexual glands and the bladder. Castor-oil packs on the abdomen should be used at the beginning of all serious attempts at health improvement, especially if there are any problems with the digestion, liver, gallbladder, or pancreas, or if you have arthritis, cancer, or any other degenerative disease. Hot castor-oil packs can be applied directly to weak, painful, or swollen parts of the body, for example, on weak eyes or eyes with cataracts or glaucoma, stiff neck, painful back, swollen knee, and on cysts and tumours. Persistent use of castor-oil packs frequently gives excellent results when other methods fail.

To make the pack, carefully warm castor oil (it burns easily); soak a woollen cloth in the oil and fold it into three or four layers. Cover the abdomen from the pit of the stomach, and including the lower rib cage, down to the pubic bone. If the cloth is not big enough, shift it on successive days of treatment so all areas get covered; favour especially the liver area under the right lower rib cage and any part of the body that causes pain or problems.

Place a plastic covering over the cloth with a hot-water bottle on top of that; finally, wrap a long towel around the torso and fasten it securely. If used in bed, place additional plastic underneath you or use old linen, as the pack might leak. Leave the pack in place for one to two hours, or overnight. Afterwards, sensitive skin can be wiped with warm water containing baking soda. Store the pack in a pot or tin for repeated use, adding more oil as required with each usage.

Use the pack for three to five days, interrupt the cycle for several days, and repeat the series as often as necessary. During or after a period of applying a pack on the liver and abdomen, take several tablespoons of olive oil at bedtime to induce emptying of the gallbladder.

Epsom Salt Pack: This is recommended for arthritis, neuritis, paralysis, back pain, muscle tension, gastritis, colitis, infections, and kidney problems. Dissolve Epsom salt in hot water to make a saturated solution; soak a cloth in it and apply it hot to the skin. Keep the pack warm with a hot-water bottle. Alternatively, Epsom salt can be spread liberally on a hot wet cloth and applied. Leave the pack on for one or two hours; repeat as required.

Rock Salt Pack: A pack of hot, coarse rock salt is effective for the relief of muscle pains, especially backaches. Cover a hot-water bottle with a moist cloth, sprinkle on a thick layer of coarse rock salt and lie with the painful area on top of this; or place this (salt-side down) on the painful area if more appropriate. Keep it as hot as you can stand it for an hour or two. Renew the salt if it is necessary to repeat the pack.

Honey Pack: The honey pack has an excellent reputation for healing infected wounds and ulcers, and it is used in alternative cancer therapy to destroy tumours and relieve pain. It is not only effective for problems close to the skin but can help problems such as pain as well. The honey must be raw and not previously heated. Smear a thick layer of honey over the affected area and cover it with linen. With long-standing problems, continue renewing the honey daily and keep the area covered all or most of the time for several weeks. It is even more effective if you mix a few drops of 35 percent hydrogen peroxide with the honey. Try to obtain honey that is not runny, as adding hydrogen peroxide makes it more liquid; crystallized honey works well.

Onion Pack: Onion packs are good for asthma, bronchitis, colds, pneumonia, and for loosening phlegm in the airways. Wrap a large onion in aluminium foil and bake it until soft. Pulp the onion in its juice and spread it over the upper chest and throat area. Cover this with a plastic sheet or cloth and keep warm for an hour or two with a hot-water bottle. Repeat as required. Do not eat the onion, as aluminium should be avoided in cooking foods to eat.

Potato Pack: Raw potato packs are excellent for irritated and weak eyes, as well as for blindness and cataracts. Grate a refrigerated potato; preferably, use the outer parts of old, organically grown potatoes. Cover the eyes for 30 to 60 minutes; you can put the potato pulp in some gauze if you like.

Cabbage Leaf Pack: The fresh outer leaves of green cabbages have healing potential when applied over weak or inflamed parts of the body. Hold one or two leaves in place over your abdomen and liver with a towel wrapped around your waist. Do this frequently. You can cut out part of the stem and roll the leaf with a bottle to flatten the ribs, especially if it is to be applied over sensitive skin areas. If no more outer leaves are available, you can also use the paler green inner leaves; wilted leaves are useless for this pack. Cabbage leaves are especially good for mastitis. For serious conditions, apply cabbage leaves 24 hours a day, renewing every four to six hours. For cancer, apply a leaf over the tumor site as well as over the abdomen and liver.

Green Leaf Pack: Other edible fresh green leaves can be applied over problem areas, either whole or as a pulp. To cover a fresh wound, for example, masticate some fresh wheat or barley grass and apply the juicy pulp.

Blistering Agents: The application of blistering agents and skin irritants produces the quickest relief for arthritis and related inflammatory conditions. The principle is to draw out through the skin the toxins and congesting energies from inside the body. Liniments, compresses, and plasters can be used. Suitable agents for this are a paste of mustard or cayenne, garlic, oil of wintergreen, or the white sap of a papaya or fig. Initially, they may cause skin reddening, which can give some relief, but if applied for a long enough time in a concentrated form, they will produce blistering, which is a much more effective healing response.

If the skin is non-sensitive, it is more effective to prick it in many places or to mechanically irritate it before applying the irritant. Cover the area with a cloth or plastic sheet until blistering has taken place. Large blisters can be drained first. Cover the blistered area with fresh cabbage leaves. For long-standing conditions, the blistering pack may have to be repeated several times.

Escharotics: The most effective way to remove skin cancer, melanoma, and tumours close to the skin, possibly including breast tumours, is to apply an escharotic. These are caustic remedies, commonly with zinc chloride and the herb bloodroot as main ingredients. These are available from some health stores and naturopaths. Skin cancers can become inflamed for a few days, then form dry pus (an eschar), and then fall out after about ten days.

To reach a tumor under the skin, such as a breast tumor, it may be necessary to leave the paste in place for two or three days and reapply fresh paste several times as soon as any developing pus has cleared. Keep covered with a Vaseline-coated pad that may require frequent changing. Commonly, this procedure causes a strong local inflammation with swelling of the breast, redness, and pain for several days. The bigger the breast and the deeper the tumor, the more difficult is it to create a “hole” deep enough for the escharotic to reach the tumor, so more applications are required.

While this can be more uncomfortable and drawn-out than surgery, one advantage is that escharotics have a reputation of not creating secondary tumours in other organs. You can ease your discomfort during this time with painkillers and anti-inflammatory measures. For more detailed information on escharotic cancer salves, see www.altcancer.com and www.cancersalves.com.

Meat Pack: A folk remedy for drawing poisons from the body is to place fresh, raw, minced meat under the soles of the feet for several hours or overnight. Replace the meat several times if required. The meat is then poisonous, so do not touch it; best bury it. You can also try this (in addition to other measures) in the case of accidental poisoning or bites of poisonous snakes or insects. Raw meat can also be placed over malignant tumours for drawing; again, do not touch it when removing it from the skin.

Infrared Heat Lamp: Instead of a hot pack, you can rub a remedy into the skin and expose the affected area to infrared heat. For arthritic pain in the knee or hip or for back pain, use peanut oil, castor oil, or a solution of Epsom salt or magnesium chloride; you can also add other helpful ingredients to the magnesium solution, such as glucosamine and MSM. Continue with the heat for 15 to 60 minutes, and rub more of the remedy into the skin whenever it becomes dry.

Bowel Cleansing: Most failures in healing programs occur because bowel cleansing is neglected. Toxins are reabsorbed from the bowels and accumulate in the blood, causing or contributing to allergies, boils, cancer, colds, eczema, fatigue, inflammations, irritability, lack of concentration, menstrual problems, mental diseases, migraine, swellings, and general toxaemia. Furthermore, constipation is a main factor in gallstones, liver problems, and probably bowel cancer.

Increasing bowel movements to two or three a day is only a first step. Even more important than the frequency of bowel movements is the speed (called transit time) with which food residues move through the intestines. To check this, eat a highly coloured food (for example, red beets or a cob of sweet corn) with an evening meal. A healthy intestine should eliminate all colouring matter before lunch the following day or at least before the next evening. If it takes longer, you are constipated. Use appropriate mineral and vitamin supplements, laxative foods, colonics, acupressure, and exercise to improve bowel conditions (see Steps 3 & 4).

Ordinary toilets encourage a sitting rather than a squatting position. However, this posture is not conducive to the complete emptying of the large intestines. A squatting position is preferable. If you cannot squat on top of the toilet, support your feet with a footstool; additionally, bend forward to exert more pressure on the bowels (for further information, see www.naturesplatform. com). An alternative way of keeping the bowels empty is by taking laxative foods, such as ground linseed (flaxseed), and periodically having an isotonic flush (a teaspoon of salt in a quart of water) or Epsom salt flush (a teaspoon of Epsom salt in a glass of water), preferably with some crushed garlic added.

Early in the twentieth century, a British surgeon became famous for “curing” all kinds of incurable diseases by surgically removing the lower part of the large intestines of his patients. We certainly can do better than that and cure or prevent many diseases simply by keeping our intestines clean and avoiding surgery. Most individuals are very particular in keeping clean on the outside, using much soap and having frequent showers, but they do not seem to mind being stuffed with stinking waste matter. While this does not apply to everyone, it is a problem for anyone with constipation and especially for meat eaters and those with fat malabsorption. Overweight individuals can easily carry more than 20 pounds of putrefying bowel content as part of their body mass.

If you prefer to be clean on the inside as well as on the outside, develop the habit of “rinsing” your lower bowel once or twice after each bowel movement and possibly again before going to bed. You can do this by sitting on the toilet, with enema bucket or bag hanging from the wall near the toilet. After a bowel movement, insert a pint of lukewarm water and expel it immediately. To let the water penetrate more deeply you may for a short time bend forward with your bottom raised, but this is not essential. Finally, rinse the outside. Preferably use some gel or oil before inserting the tube or end-piece douche, and use a plastic glove.

Habitual bowel cleansing like this not only prevents the accumulation of putrefying matter in the lower bowel, it also greatly speeds up the transit time of the bowel content. Among other benefits, this tends to increase the release of bile and toxins from the liver and gallbladder and prevents the formation and accumulation of gallstones. However, if you try it several times and find that the enema water comes out clean after bowel movements, then you do not need to do it.

Colonics: Colonics - the name is short for colonic irrigation - is an effective therapy, bringing immediate relief for many health problems. The more serious the condition, the more often colonics should be used: once a week at first, but daily use may be required for a while. After several treatments, have a colonic once a month, and later, once or twice a year. Iris diagnosis (see step 12) can guide you as to the frequency necessary for your condition. If the intestinal area in your iris shows dark or brown discoloration, or if it is extended, colonics should be used more often; this is also true if foul-smelling wind is present, as this indicates putrefactive processes under way in the bowel pockets.

The aim of colonics is to remove all putrefying matter from the bowels, clean out the bowel pockets, and remove the hard crust that often coats the walls. To achieve this, the bowels should be filled repeatedly with as much water as possible. It is best to have a thorough bowel movement before starting the colonic.

To administer a colonic, add one heaping teaspoon of salt and one level teaspoon of baking soda to two to three quarts of moderately warm water. Use a bucket enema or enema bag; place it about two to three feet above the body on a table, chair, or the rim of the bath, or hang it from the wall. A good position is to kneel on the floor with your head down. More comfortably, you can lie in a warm bath with the rectum elevated, or you can lie in bed in a similar position, or on the right side; have a plastic sheet underneath you.

Lubricate the rectal tube and insert it slowly to a depth of about eight or 12 inches. If there is any pain or discomfort, do not insert it any farther. Expel the first lot of water from your intestines when increasing pressure is felt. After that, you should be able to hold much more water. When renewed pressure is felt, stop the inflow of water and assume different positions, such as lying on your right side for a while so as to fill the ascending colon. Repeat the expulsion and intake of water until all of it has been used.

You can introduce a pint of diluted fresh juice of wheat or barley grass if available or of edible green leaves, and retain that in the intestines for up to 30 minutes. This is the standard procedure for home colonics for people with normal or acid body conditions. The addition of baking soda neutralizes the often overacid bowel condition, while salt acts partly as a drawing agent and partly prevents demineralization of the blood plasma by osmosis, which could cause shock. This is important for people with low blood pressure and poor blood circulation. Those with high blood pressure or very weak kidneys should not add salt and baking soda to the colonic water, but use highly diluted lemon juice instead. If this causes soreness, or if the body is not too alkaline, use diluted fresh juice of green-vegetable leaves.

If you cannot do a colonic at home, receive a series of professional colonics at a hydrotherapy centre. Everyone can benefit from some form of bowel cleansing from time to time. (Instead of a series of colonics, you can use the Basic Cleanse described in Step 3.)

Coffee Enema: For stimulating and detoxifying the liver in the case of serious conditions, such as cancer, a series of coffee (preferably organic) enemas can be helpful, especially if there is pain and discomfort. Boil two tablespoons of ground coffee (not instant coffee) for three minutes in one quart of water; simmer for 15 minutes and strain; rinse the coffee grounds with more water. Use this as a retention enema after cleaning the bowels with a normal enema or colonic. If you can hold in only part of the coffee, use the rest with a second enema. Retention means to hold it inside for ten to 20 minutes. Do not use this if you are allergic or sensitive to coffee.

Oxygenating Colonic: This can often improve the well-being of patients with ad-vanced diseases, such as cancer. Oxygenating colonics are available at some hydrotherapy centres.

As an alternative for home treatment, add food-grade hydrogen peroxide to a retention enema after the colon has been cleaned with an ordinary colonic. Use one to two pints of lukewarm liquid in which one teaspoon of salt has been dissolved. Initially, include one tablespoon of six-percent or two tablespoons of three-percent hydrogen peroxide. Retain this mixture inside your intestines for about ten minutes. If this is well tolerated, double the amount of hydrogen peroxide with each successive treatment until it becomes too uncomfortable and you cannot hold it any more. Then cut back to the preceding level and continue at that rate.

For cancer patients with a serious condition, under professional supervision you can start at higher dosage levels and include up to half a pint of three-percent hydrogen peroxide. This may induce the evacuation of much waste during the following days, and the patient may temporarily look and feel worse. After this, however, there should be considerable improvement. If available, lead ozone through a long tube into the rectum with an additional shorter tube as outlet and safety valve. For further information on this, see Step 31.

Vomiting Therapy: Induced vomiting, often called an “upward purge,” can be helpful for a wide range of acute and chronic conditions, such as asthma, bronchial problems, catarrh, digestive disorders and indigestion, gastritis, hysteria, infections, liver and gallbladder problems, lumbago, migraine, poisoning including snakebites, and skin diseases including eczema, gangrene, and psoriasis.

While vomiting clears out mucus and decomposing matter from the stomach, the main benefit is derived from relaxing the digestive organs and the diaphragm while stimulating the solar plexus. A need for this therapy is indicated especially by contracted intestinal reflex areas in the iris, by way of iris diagnosis (see Step 12), but also if these are whitish or brownish and especially if they have black lines.

To induce vomiting, drink five or six cups of lukewarm water and touch or tickle the throat far down. Repeat until the water comes back clear. Instead of water, an emetic herb tea can be used: adder’s tongue, bayberry bark, black root, boneset, horehound, lobelia (large doses), mustard, ragwort, or vervain. After-wards, drink peppermint or mint tea to settle the stomach.

* * * * * * * * * * * *

Step 22

URINE & UREA THERAPY

It may not be your "cup of tea", but many swear by this ancient therapy.

The use of urine in the treatment of diseases has a long history. Many cultures have used urine for medicinal purposes: It was praised in ancient Egyptian papyri and used in ancient Rome, China, India, America, and European countries. An article in the Journal of the American Medical Association stated that in “primitive medicine” there is scarcely a disease that has not been treated with the external or internal use of urine.6. Presently it is most widely used in India and Taiwan where the youthful appearance of many Buddhist monks is ascribed to their routine urine ingestion.

Thousands of scientific articles have been written about the health benefits of urine and its ingredients. A good book to get acquainted with some of this research activity is Your Own Perfect Medicine, by Martha Christy; one of many websites on urine therapy is: www.uropathy.com.

At the very least, all this past and present interest in the healing benefits of urine indicates that here is something worthwhile for the adventurous health seeker to explore. To start urine therapy the main problem we need to overcome is our social conditioning. You may also want to know how a substance expelled by the body as waste can possibly be beneficial.

Beneficial Ingredients: Urine is not a waste product full of harmful substances as is commonly believed, but instead a treasure-trove of the right nutrients that we need for our health. It is full of hormones, enzymes, vitamins, trace minerals, and other valuable biochemicals. Basically, urine is just filtered blood, and in its composition close to blood serum. The main function of the kidneys is to regulate the concentration of all these biochemicals in our blood. Surplus can be as harmful as deficiency; however, any hormone or enzyme removed as surplus may be in short supply a few hours later.

With advancing age, our hormone and enzyme production declines to suboptimal levels, while the kidneys become less and less efficient in retaining needed ingredients. Therefore, it greatly helps, especially with chronic degenerative diseases and advancing age, to recycle these valuable hormones and enzymes. This is what urine therapy does.

Another important aspect is the presence in urine of antibodies and immune-stimulating factors against viruses, harmful bacteria, or fungi we may harbour in our body. Even minute amounts of antibodies, sometimes at levels so low they cannot be detected with conventional methods, are effective in preventing and treating diseases. Curative effects of urine therapy have been shown on fungal and viral infections, such as hepatitis, poliomyelitis, and AIDS, and urine is especially effective against allergies, autoimmune diseases, and other disorders of the immune system.

When drinking urine routinely, say a cupful every morning, allergic reactions may not occur at all, thus doing away with the need to identify to which substance one has been reacting.

Benefits of Urea: Urea is an end product of protein metabolism and the main ingredient of urine. Only recently has it been discovered that the concentration of urea in the blood has a key role in regulating at least seven major pathways. Urea levels are low with cancer and many other diseases; in fact the frequency of cancer and especially of multiple tumours and metastases increases greatly if urea levels are low.

The Greek physician Evangelos Danopoulos, M.D., pioneered the use of urea in cancer treatment. Urea was notably effective in regressing liver tumours, greatly increasing survival times, and preventing metastases or secondary tumours. Dr. Danopoulos used 45 g of urea daily in six divided doses for 40 days and then 20 g in three doses for two years (45 g correspond to six rounded teaspoons, and 20 g is about three teaspoons). He also experimented with injecting urea close to tumors.[7]

Urea may not just be a waste product or the final state of protein metabolism as commonly assumed. Apparently, it can also be used as raw material to synthesize amino acids and proteins, as has been shown in patients with kidney failure. These patients deteriorate quickly on a diet containing normal amounts of protein but do much better if the diet is extremely low in protein. They show progressive clinical improvement when receiving added urea, which then becomes an important source of nitrogen for protein synthesis.[8]

Urea is the best natural diuretic and is far better than synthetic diuretics. This is important for those with oedema or fluid retention (due to kidney or heart weakness) and especially for those with fluid pressure on the brain or spinal cord or in the eye, as with glaucoma. Additional urea is important with brain tumours, stroke, meningitis, and other inflammations of the brain and spinal cord. In one case of a massive brain oedema connected with a suspected brain tumour, the patient received 256 ml of 30% urea; after two hours, the oedema and suspected tumour had completely disappeared.9

When applied externally, urea has a beneficial effect on the healing of wounds, such as infected wounds, burns, and ulcerating tumours. Urea can be directly packed into the wound or a strong or saturated solution can be applied. In this way, urea removes the foul odour often associated with an ulcerating tumour or other dead or putrefying tissue. You can obtain urea from a pharmacy. Unfortunately, most of the research into the medical use of whole urine and urea took place in the first part of the last century, while in recent years the emphasis has been on researching individual components.

CAUTION: Do not ingest large amounts of urine if you take medicinal drugs or other toxic substances, as these an be re-cycled and cause an overdose. While harmful substances are not necessarily re-absorbed and may just cause diarrhea instead, you do not know how much is safe. Therefore, either ask your health professional for advice, or use urine homeopathically, or in small amounts up to several spoonfuls daily.

How to Use Urine: There are various ways to use urine, but generally it is advisable to start with small amounts. This is not only to get accustomed to the taste of it, but also to slow down any healing reactions. These occur commonly with methods that genuinely improve our health (see Step 10).

Reactions may manifest as diarrhoea, vomiting, nausea, or a temporary flare-up of old health problems.

Start with a few drops or a spoonful. Swallow immediately; you can wash it down with some water or juice, or eat a piece of fruit.

With allergies and immune disorders, it is more effective to leave diluted urine under the tongue for a few minutes. An allergic reaction can best be stopped or neutralized by collecting the urine when the symptoms are worst; however, in practice, try to use it as soon as you notice the first symptoms. Some researchers suggest that the proper dose can be found by slowly putting single drops under the tongue until the taste and temperature of a drop can no longer be distinguished. This is then the neutralizing dose for an allergy. Each subsequent day you can double the previous dose or the number of times you take them per day until you reach your target amount or the dose with which you are comfortable.

A suggested amount for general health improvement and rejuvenation is two or three cups daily. More is advisable temporarily to overcome serious diseases, especially AIDS and other viral, bacterial, or fungal infections, as well as cancer and kidney diseases. In these cases, ingest about one quart daily, but the more the better, while for a limited period of several weeks try to drink all or most of your urine, except in the evening, so that you do not have to get up at night. However, this may only be suitable during a cleansing period, as drinking urine after a meal may cause digestive upset.

A good way to flush the kidneys is to drink one to two quarts (litres) of diluted urine before breakfast; use one or two cups of urine per quart of water. Start drinking all the urine when you first wake up, ideally at 5 a.m. Then urinate again and drink another dose 90 minutes later, then 90 minutes later again. Either drink sufficient liquids in the evening so that the first morning urine is not too concentrated, or dilute strong urine with sufficient water; about one cup to a quart of urine.

Do not use urine that is cloudy, as it often is during a healing reaction. If the body is reasonably clean, then the urine will not taste offensive or reflect the aromas of the last meal. The second morning urine, especially when drinking a fair amount of liquids after rising, is usually very mild; it is better for getting accustomed to than the first morning urine. To catch a cleaner sample, collect urine midstream.

The Urine Fast: The most powerful form of urine therapy is urine fasting, a method routinely used by John Armstrong and described in The Water of Life. Armstrong had good results with this, and I also have found it very effective, especially in cases of kidney failure. While it sounds contradictory to drink urine when one has kidney failure, the idea is to flush the kidneys with a large volume of stimulating fluid. I see this as the only logical way to rehabilitate failing kidneys; reducing the flow of fluid through will only make it worse. Diluted or weak urine appears to be most suitable, at a rate of one gallon or more daily. However, adjust intake to output; whenever there is increasing fluid retention, reduce the intake. This intake also helps to increase lymph flow as explained under lymph-drainage massage in Step 20. A main advantage of urine fasting over water fasting is the greatly increased amount of energy produced, so that the urine fast can be continued for a much longer period than a fast consisting only of water.

During the urine fast, drink all of your freshly voided urine (and initially some added water) to bring the daily intake volume to about five quarts. Gradually increase the ingested amount to this full intake before starting the fast while gradually reducing your food intake to just a few pieces of fruit or salad and fresh vegetable juices. End the fast by taking only a small amount of fresh raw fruit or vegetables for another week or two, gradually increasing the amount and variety of fresh

As with other fasts or cleanses, it is important to keep your bowels moving daily. I suggest initially taking a tablespoon of Epsom salt with a crushed clove of garlic for several mornings or evenings when starting with a greatly restricted food intake. This will help to avoid or minimize unpleasant cleansing symptoms, such as weakness, pain, or nausea. However, in many cases the ingested urine will already be overstimulating the bowels, with liquid discharges, in which case no further action is needed.

Ingest only your own freshly voided urine. Normally, urine is light yellow, clear, and sterile, especially if collected midstream, but during healing reactions it may become offensive and cloudy. In this case, use it only strongly diluted, if at all. When ingesting a large volume of your urine, check that it is not too acid. The morning urine normally is slightly acid, but during the day it should be slightly alkaline.

To check your urine, use litmus or pH paper from a drugstore. Litmus paper is red when dipped into an acidic solution and blue when alkaline; neutral or pH 7 is in between both colours. With pH paper, you get a colour chart to compare. If the urine is mostly acid, take a half-teaspoon of sodium bicarbonate in water once or several times daily, but not within three hours after meals; do this until the urine remains alkaline during the day.

Those with cancer, high blood pressure, or fluid retention should use potassium bicarbonate (from a pharmacist) instead of sodium bicarbonate. If your body is too acidic, sometimes the emotions become overly sensitive, along with a physical tendency to inflammations, mouth ulcers, and strong skin reactions.

The effectiveness of urine therapy with cancer, oedema, and infectious diseases can be increased by taking daily one to six teaspoons of urea (e.g., mixed in water or juices) in divided doses. Take three to six teaspoons when only a small amount of urine is ingested, then progressively less with larger urine intake; stop completely when nearly half or more of your daily urine is ingested.

Homeopathic Urine: It is possible to use urine ultra-diluted in homeopathic form. The simplest way to prepare it is as follows: To one teaspoon of unchlorinated water add one drop of your own fresh urine. Shake it vigorously in a small bottle about 50 times with a sharp downward stroke. Add one drop of this mixture to another teaspoon of water, shake again; and then repeat the whole procedure a third time. Place three drops of the resulting fluid under your tongue several times daily. When it is used up, wait a few days, then prepare a new homeopathic remedy.

An alternative method is even simpler and recommended as an introduction to urine therapy or for times when cupfuls of urine are not tolerated or not wanted. Dilute one or two spoonfuls of morning urine with five to ten times the same amount of unchlorinated water. Shake 50 times with a sharp downward stroke, and drink immediately or take a spoonful several times during the day. Keep it in the mouth for a minute before swallowing.

Homoeopathically energized remedies should not be exposed to strong magnetic or electromagnetic fields and should be taken with a non-metal spoon. Keep any homeopathic urine remedy only lightly covered in a cool place but not in an electric refrigerator; make it up fresh daily.

External Applications: A urine pack, with the addition of some urea, is often effective with skin problems, such as burns, wounds, gangrene, psoriasis, eczema, dermatitis, fungal problems, and aging skin. It can also be helpful for internal tumours, inflammations, and diseased organs. Keep the affected area covered with a folded cloth that is well moistened with urine or concentrated urea solution. Put another cloth or plastic sheet on top and renew several times daily.

Urine becomes more alkaline when it stands for a few hours. Armstrong regarded it as more effective than fresh urine for external applications. However, it also smells more, and anyway, fresh urine seems to be adequate in my experience. With gangrene, dead flesh, skin infections, and open cancer and ulcer sores, it is more effective to saturate the applied urine with urea; this approach also removes offensive odours. With weeping or suppurating wounds, especially if associated with foul odour, it is beneficial to cover the area with dry urea.

Urine can be concentrated by leaving it for several hours in a flat dish in the sun or in a warm place. This has been recommended by Armstrong for application in friction rubs to rejuvenate aging skin. Armstrong also recommended urine packs placed over tumour sites or diseased organs in addition to rubbing the body daily for two hours with small amounts of urine. As you might guess, the primary problem with urine packs is the offensive odour. You could try to prevent this by sealing the pack with duct tape; completely seal it and keep the pack moist by injecting fresh urine into it several times a day with a syringe.

Social Considerations: As noted, the main problem when starting urine therapy is social conditioning. You can overcome that by reading about and contemplating the expected health benefits. The taste of fresh urine is not normally offensive, except if the body is polluted or when fasting or if the urine is strongly concentrated. Therefore, dilute strong urine, start with a small amount, and gradually increase it or have a day on fruit before starting to ingest urine. Minimize salt and strong spices in your diet when on urine therapy. An easy way to drink urine or any other unpleasant fluid is by breathing only through the mouth; you can pinch the nose until your mouth has been rinsed with water or juice, or you can eat a piece of fruit.

* * * * * * * * * * * *

Part 3

NUTRIENT & REMEDIES

This chapter provides information about vital nutrients, such as vitamins, minerals, amino acids, and enzymes, as well as basic remedies including herbs and homeopathic remedies.

The more we adopt a high-quality diet, the less we need to know about individual nutrients, and vice versa. Unfortunately, most individuals presently have a less than ideal diet, and some knowledge about nutrients can help to avoid or correct imbalances and deficiencies. Furthermore, many nutrients are now used in very high doses, essentially as remedies, to treat various diseases. When used in this way, detailed knowledge is important to achieve the desired effect without exposing yourself to unnecessary danger.

While there are dangers with what I call "self-healing" and health authorities call "self-medication," these are infinitely smaller than those posed by prescription drugs. I believe there is less long-term harm in using a nutrient incorrectly than a prescription drug correctly. It is rare to encounter serious health damage from using nutrients, even in very high doses, while medical drugs are stated to be the third leading cause of death in the United States.

Nevertheless, I do not endorse the use of high-potency nutrients without a proper understanding of any benefits and dangers, and I do not endorse high-potency supplements as a substitute for a healthy lifestyle, which is the way that they are commonly used. The main emphasis of this book is on a healthy lifestyle, which in most cases will remove the need for high-potency supplements, but if you feel that you do need some, then you will find suitable information in this part.

Generally, it is best to obtain all your necessary vitamins from natural foods and the activities of intestinal bacteria. Under present conditions, however, it is rarely possible to achieve and maintain good health without resorting to special vitamin-rich foods and nutritional supplements.

In the case of poor health, absorption of vitamins and minerals is impaired, while your requirements for them are simultaneously raised. This phenomenon underlies the widespread vitamin and mineral deficiencies in our society, which result in a multitude of diseases. Well-chosen supplements can correct deficiencies, but in a good maintenance diet, the daily vitamin intake proposed by government agencies should be at least doubled.

* * * * * * * * * * * *

Step 23

VITAMINS

Find out what these essential nutrients do and if you have special requirements for any of them.

Vitamins are water soluble or fat soluble. Water-soluble vitamins are vitamin C and the B-group vitamins. Except for vitamin B12, they are easily absorbed, and deficiency is due to a deficient diet or to increased requirements. Due to genetic factors or diseases, some individuals may require up to a hundred times the usual amounts to function normally. One common problem is a thickened or hardened cell membrane, preventing nutrients from entering some organs, especially the eyes. Being water soluble, a surplus of these vitamins is easily expelled through the kidneys, and any problem from overdosing is rare.

The fat-soluble vitamins, A, D, E, and K, are more difficult to absorb. Deficiencies can easily arise even with an adequate diet, although most “balanced” diets are deficient due to the cooking methods and use of refined foods. The most difficult problem is fat malabsorption, which is relatively common and mainly due to a reaction of the intestinal wall to gluten. In this case, even a diet plentiful in fat-soluble vitamins can still lead to deficiency symptoms and may not be corrected by high-potency supplements. While symptoms can mirror individual deficiencies, a common feature of fat malabsorption is dry skin and difficulty gaining weight.

This absorption difficulty can be overcome by using water-soluble vitamin E tablets, emulsions for vitamin A, and sunshine for vitamin D. In addition, rub vitamin A and E oil (from opened capsules) into your skin and squeeze the contents of a halibut oil capsule under the tongue for slow absorption overnight. Do not use synthetic versions of oil-soluble vitamins. While these are easily absorbed, they are known to cause side effects, such as liver toxicity or interference with pregnancy.

The bacteria among our intestinal flora supply a considerable part of our B vitamins. Unfortunately, antibiotics destroy these essential bacteria and can cause vitamin deficiencies. For this reason, use fermented foods or cultures containing acidophilus and bifidobacteria and take B-vitamin supplements whenever you take antibiotics. Other widespread conditions -causing persistent vitamin deficiencies and dependencies are deprivations during foetal develop-ment and, in later life, acute and chronic infections, parasites, and inefficient metabolism.

In general, vitamin and mineral supplements may not be required for the maintenance of health or even the cure of most degenerative diseases provided the diet consists predominantly of fresh and raw foods, featuring sprouted seeds, grass and vegetable juices, as well as high-quality nutrients such as bee pollen, spirulina, and kelp.

Bowel-Tolerance Vitamin C: This means taking enough of the vitamin to produce a bout of diarrhoea to identify your specific bowel tolerance for vitamin C. This can be a successful treatment for infectious diseases, including viral infections, which do not respond to antibiotics. The dose of vitamin C is increased until the patient develops transient diarrhoea, usually one episode. Then the rate is cut back slightly until the bowel tolerates the high vitamin C intake. During serious infections, the bowel tolerance is much greater than it is in a healthy state. In normal health, for example, the bowel tolerance is about 4-15g per day; with a light cold, it might be 20-30g; with a serious cold, 60-100g. Typically, mononucleosis or viral hepatitis requires 40-60 g; gastroenteritis 60-150g; and viral pneumonia more than 150g. Other infections, including herpes simplex and herpes zoster, respond favourably to this treatment, but fungal infections do not.

When you are taking these massive amounts, it is important to select the correct form of vitamin C. If mainly unneutralized ascorbic acid is used, the body becomes too acid, while sodium ascorbate introduces too much sodium, and calcium ascorbate too much calcium. If your blood pressure is approximately normal, take a mixture of ascorbic acid, sodium ascorbate, and calcium ascorbate. With low blood pressure, use mainly sodium ascorbate with some calcium ascorbate. With high blood pressure, potassium and magnesium ascorbate combined with ascorbic acid is best. You can neutralize some of the ascorbic acid with potassium bicarbonate or magnesium carbonate at the ratio of three to one, or two parts ascorbic acid to two parts magnesium oxide. The solution should taste slightly acidic. You can dissolve the vitamin C in water, herb tea, or juices, if you like.

Depending on the severity of your condition, take vitamin C at one- or two-hour intervals. Do not continue self-treatment at bowel-tolerance level for more than a few weeks, and in serious cases you should be under professional medical supervision. The high vitamin-C intake should be reduced gradually to a daily maintenance dose of 3-10g. A sudden withdrawal of all vitamin C can cause the return of colds, allergies, or fatigue, so try to avoid this.

Bowel tolerance as a treatment is not suitable with kidney disease. Initially, treatment of life-threatening diseases or emergencies can be supported with daily or weekly injections of sodium ascorbate given intravenously in a three-percent solution. Many degenerative diseases associated with a weak immune system, for example, autoimmune diseases, cancer, and multiple sclerosis, can benefit from intermittent periods of bowel-tolerance vitamin-C treatment. This method should also be used for accidental poisoning and bites of poisonous snakes and insects.

Now let’s review all the vitamins. In the following compilation, RDA means recommended daily allowance for males as determined by the Food and Nutrition Board of the U.S. National Research Council. A second RDA value is for females. “IU” means international units. Not all of the listed deficiency symptoms are always manifested. Symptoms appear first in the weakest organs. As for measurements, 1 g = 1000 mg; 1 mg = 1000 mcg.

Fat-Soluble Vitamins: The oil-soluble vitamins are A, D, E, and K. They are soluble in oils and fats, which is where we get them in our diet.

Vitamin A: The chemical name is retinol; two molecules of retinol can be formed from one molecule of beta-carotene. The RDA is 5000/4000 IU; therapeutically 25,000 to 100,000 IU are commonly used, and sometimes up to 1,000,000 IU, especially in cancer treatment and for eye diseases. Long-term oversupply can lead to bone erosion due to a relative deficiency of vitamin D. Therefore, supply vitamins A and D together, preferably as cod liver oil, otherwise halibut liver oil. One halibut-oil capsule provides 4000 to 5000 IU. Vitamin A is toxic in very high doses over long periods and is destroyed by light and oxygen. Symptoms of vitamin A toxicity are similar to those of vitamin A deficiency. Do not take very high doses without professional supervision.

Deficiency Symptoms: Eyes are dry, infected, and sensitive to bright light; night blindness, poor dark adjustment, colour blindness, and glaucoma are other symptoms. Skin is dry, rough, horny, and scaly; there are pimples, acne, and psoriasis. Hair is dry, falling out, or dull. Nails have longitudinal ridges and are peeling and brittle. Sense of smell and taste are poor or distorted. Respiratory infections and other lung problems, ear problems, deafness, hearing noises, and cancer; thyroid is overactive and the person is underweight.

Best Sources: Fish-liver oils, liver, butter, and egg yolk. Yellow-orange vegetables and green juice are high in beta-carotene, which some individuals can efficiently convert into vitamin A.

Vitamin D: The common natural form is vitamin D3 or cholecalciferol. Vitamin D2 or ergocalciferol is synthetic and can be toxic in high doses over long periods; I do not recommend it. The RDA is 200 IU; therapeutically up to 4000 IU are used.

Deficiency Symptoms: Rickets, teeth and gum problems, muscular weakness, convulsions, cramps, tetany, nervous instability, underactive thyroid, depression, multiple sclerosis, overweight, eye problems, and weak, deformed, and porous bones and other symptoms of calcium deficiency.

Best Sources: Sun exposure, fish-liver oils, and egg yolk.

Vitamin E: The chemical name of this group is tocopherols; mainly d-alpha tocopherol is used, preferably in addition to mixed tocopherols. The RDA is 15/12 IU or mg; therapeutically, up to 3000 IU or mg can be used. Patients with heart problems or high blood pressure should increase the dose slowly. This vitamin is destroyed by oxygen, chlorinated water, inorganic iron supplements, and estrogen drugs.

Deficiency Symptoms: These can include disorders of the reproductive and circulatory systems, stroke, heart disease, leg pains, cramps, poor wound healing, pronounced scar tissue, muscle weakness, chronic fatigue, tender breasts, arthritis, cancer, eye and ear problems, gangrene, infections, hot flushes, malabsorption, multiple sclerosis, myasthenia gravis, phlebitis, skin problems, and varicose veins.

Best Sources: Cold-pressed oils (wheat germ oil), rice polishings, sprouted seeds, nuts, egg yolk, green leaves, and grass juice.

Vitamin K: The RDA is 80/60 mcg; therapeutically, 5–15 mg are used. It is destroyed by light, alkaline conditions, and strong acids. This vitamin is essential for normal blood clotting to prevent haemorrhaging. A deficiency causes frequent, prolonged, or severe bleeding, easy bruising, and menstrual clots, as well as nausea and vomiting during pregnancy. It is also needed for bone formation, or to prevent osteoporosis, tooth decay, calcium deposits, and hardening of tissues and arteries.

Injections of vitamin K derived from plant sources have been effective in the control of strong pain. This vitamin can be obtained from intestinal bacteria (in other words, made in the gut), green and especially dark-green vegetables, liver, and egg yolk. Vitamin K1 from plants is more effective than K2 from intestinal bacteria; the synthetic form (K3) can have toxic effects.

Water-Soluble Vitamins: These vitamins easily dissolve in water; with this they are easily absorbed, but also easily lost in cooking and, except vitamin B12, not stored in the body. Most of the B-group vitamins are essential components of enzyme systems.

Vitamin B1: The chemical name is thiamine. The RDA is 1.5/1.1 mg; the therapeutic dose is 50–3000 mg. It is destroyed or used up by heat, sugar, alcohol, and smoking. Despite vitamin B1 supplements, deficiency can arise from a lack of gastric acid.

Deficiency Symptoms: These include poor circulation with cold hands and feet, oedema (fluid retention), furrowed tongue, underactive thyroid, fatigue, depression, insomnia, anxiety, oversensitivity to pain and noise, nausea, poor memory, apathy, debility, weight loss, and weak and painful muscles, especially in the legs, starting with numb, burning feet, later affecting calves and thighs, and sometimes resulting in paralysis. Other conditions include severe gastrointestinal disturbances, tachycardia (fast pulse) and dyspnoea (breathlessness) on exertion, brachycardia (slow pulse) at rest, enlarged and weak heart, mental illness, multiple sclerosis, diabetes, hypoglycaemia, allergies, and addictions (to drugs, alcohol). A severe deficiency is known as beriberi.

Best Sources: Food yeast, rice bran, sunflower and sesame seeds, peanuts, millet, and grains.

Vitamin B2: Vitamin B2 is chemically known as riboflavin. The RDA is 1.7/1.3 mg, the therapeutic dose 50–1500 mg. It is destroyed by light and makes urine bright yellow.

Deficiency Symptoms: Eyes are sensitive, burning, and bloodshot (blood vessels become visible in the white of the eye) and there is blurred vision or cataracts. Skin is oily and itching; there can be eczema (especially around the nose, forehead, ears, scrotum, and vagina), redness of part of the face (rosacea), acne, bedsores, and purplish skin parts. Hair is dull and oily, with dandruff. Nails are split and the tongue is sore, burning, and purplish magenta in color. There are cracks on the lips and corners of the mouth (cheilosis), as well as allergies, anaemia, arthritis, cancer, or diabetes.

Best Sources: Food yeast, liver, kidney, almonds, sprouted seeds, and grass juice.

Niacin and Nicotinamide: Niacin is also called nicotinic acid, and niacinamide is the same as nicotinamide. As the name suggests, these are closely related to nicotine and, therefore, useful when trying to give up smoking. The name vitamin B3 is also used. The RDA is 19/15 mg; therapeutically, up to 30 g have been used in schizophrenia.

The acid form (niacin or nicotinic acid) causes flushing of the face and it is prescribed to dilate blood vessels, decrease blood pressure, reduce cholesterol, or increase circulation to the head in cases of acne or migraine. The non-acid nicotinamide or niacinamide does not have these effects. In contrast to niacin, high amounts of this vitamin can cause depression. Niacin or nicotinamide improves the oxidative energy metabolism and can be used to treat many symptoms.

Deficiency Symptoms: Schizophrenia, hyperthyroidism, protruding eyes, depression, irritability, overactivity, crying spells, suspicion, loss of humour, delusions, anxiety, insomnia, confusion, and changed sense perceptions. Tongue is strawberry-tipped or bright red and shiny, with a “raw beef” appearance; the mouth displays sores, cankers, and ulcerated corners. The skin is rough and red, with brown, often symmetrical discoloration on the cheeks, neck, and back of hands. There is dermatitis with ulcerations and pellagra (cracking of skin), which is aggravated by exposure to sun. Other symptoms and conditions include headache, backache, fatigue, anaemia, lack of gastric acid, loss of weight and appetite, nausea, vomiting, diarrhoea, painful and stiff joints, arthritis, and cancer.

Best Sources: Food yeast, peanuts, rice bran, liver.

Pantothenic Acid: This is commonly used as calcium pantothenate, also called vitamin B5. The RDA is 7/4 mg; the therapeutic amount 50–1500 mg.

Deficiency Symptoms: Tongue is beefy, enlarged, and furrowed; there is fatigue, depression, insomnia, greying hair, burning feet, arthritis, alcoholism, asthma, cancer, cataracts, epilepsy, psoriasis, stress, diseases of the digestive and nervous systems, and adrenal weakness.

Best Sources: Food yeast, rice bran and polishings, liver.

Vitamin B6: The chemical name is pyridoxine. The RDA is 2.0/1.6 mg; therapeutically, it is used from 50 to 3,000 mg. It works closely with zinc and is easily oxidized.

Deficiency Symptoms: These include an inability to recall dreams, insomnia, depression, irritability, tremors, convulsions (epilepsy), migraine, or schizophrenia. The skin is oily, with eczema and urticaria (nettle rash or hives); there is oedema (fluid retention), vomiting, halitosis, lack of gastric acid, muscle weakness, anaemia, and kidney stones. Also possible are pain, stiffness, and swellings of fingers and joints; fingers and toes become white (Raynaud’s disease); autism; caries (tooth or bone decay); diabetes; Parkinsonism; and diseases of the digestive and nervous systems.

Best Sources: Food yeast, molasses, bran, sunflower seeds, rice, liver, egg yolk, nuts, beans, and bananas.

Biotin: The RDA is 0.3 mg or 300 mcg. Biotin is necessary for protein and fat metabolism as well as for healthy hair.

Deficiency Symptoms: Digestive, nervous, and skin and hair problems; anaemia, fatigue, depression, and hallucinations.

Best Sources: Food yeast, rice bran. Usually from intestinal bacteria, except after antibiotics.

Folic Acid: The RDA is 200/180 mcg, but it is recommended to ingest at least 400 mcg, and 800 mcg when pregnant. The therapeutic dose is 5–20 mg. Folic acid is easily destroyed by light, heat, storage, oxygen, and the contraceptive pill.

Deficiency Symptoms: These are widespread; most pregnant women are deficient. Symptoms include anaemia, corners of mouth cracked, red tongue, greying hair, greyish-brown skin pigmentation, irritability, depression, fatigue, forgetfulness, lack of con-centration, insomnia, dizziness, depressed reflexes, and headaches. Also present are muscle weakness, pain, numbness in legs, difficulty in walking, cramps, digestive problems such as constipation and diarrhea, sexual and circulation problems, cervical cancer, hypoglycaemia, birth defects, gout, diabetes, and cardiovascular disease.

Best Sources: Grass juice, yeast, sprouts, dark-green leaves, and liver.

Vitamin B12: The chemical name is cobalamin, the RDA is 2 mcg, and the therapeutic dose 1000 mcg. It is essential after part of the stomach has been removed. Most patients with a lack of energy respond extremely well to vitamin B12 injections, even if laboratory tests show a normal vitamin-B12 blood level.
In poor health, vitamin B12 is not well absorbed; calcium and hydrochloric acid improve absorption. Commonly, B12 tablets are absorbed under the tongue. However, recent studies show that the best absorption takes place by putting vitamin B12 into the nostrils as drops, cream, or a crushed tablet combined with a little water to make a paste.

Deficiency Symptoms: Pernicious anaemia, chronic fatigue, debility, poor circulation, numbness and stiffness, a very red and sore tongue, emotional disturbances, mental illness, cancer, liver and nervous system diseases, nerve inflammations, and paralysis.

Best Sources: Liver, kidney, fish, egg yolk, and spirulina.

Vitamin C: The chemical name is ascorbic acid; it is commonly used as sodium or calcium ascorbate. The RDA is 60 mg; 500 to 3000 mg are recommended for good health (and may be supplied by a good diet); the therapeutic daily amount is up to 100 g (in neutralized form). This vitamin is easily destroyed by air, storage, smoking, cooking, or stress. It is useful to counteract the bites or stings of poisonous snakes, insects, spiders, and others.

Deficiency Symptoms: First signs include lassitude, weakness, irritability, and vague muscle and joint pains. Later symptoms and conditions include scurvy, starting in the limbs most used, muscle pain (especially during infections), bleeding of gums and skin, capillary weakness, fatigue, poor wound healing, and acute and chronic infections. Also possible are liver and kidney problems, senility, aging skin, thrombosis (strokes, heart infarctions), crib death, eye problems, arthritis, multiple sclerosis, diabetes, schizophrenia, anaemia, bedsores, stone formation, cancer, backache, anxiety, and allergies (manifested in reaction to bee stings, headaches, rashes, asthma, hay fever, and epilepsy).

Best Sources: Fresh fruits and vegetables, sprouted seeds, rosehips, and grass juice.

Vitamin-like Substances: These are necessary biochemical substances that have not yet been awarded full vitamin status, partly because some can be synthesized within the body, and partly because their usefulness is still disputed.

Bioflavonoids: Bioflavonoids are part of the naturally occurring vitamin-C complex. A deficiency causes capillary fragility with purplish or blue skin marks, inflammation, gum bleeding, and inner ear pain. Bioflavonoids are necessary for the proper functioning of vitamin C.

Best sources: Sprouted seeds, fresh (especially colourful) vegetables and fruits, and the residue of juiced citrus fruits; bioflavonoids are also found in flower petals. The best known members of this group are quercetin and rutin; rutin is found in concentrated form in dried buckwheat leaves and flowers.

Choline and Inositol: These two are important for the absorption and metabolism of fats and cholesterol, for the synthesis of lecithin, for proper liver and gallbladder function, and for the formation and function of the brain and nerves (transmission of nerve impulses). They are helpful in the treatment of arteriosclerosis and atherosclerosis, asthma, diabetes (inositol), glaucoma, hair problems and baldness, high blood pressure, insomnia (inositol), liver diseases, multiple sclerosis, muscular dystrophy, neuritis, and tardive dyskinesia (choline especially).

Best sources: Choline is best supplied by lecithin, while other good sources are egg yolk, liver, brain, and food yeast. Inositol is best supplied by sprouted seeds (in unsprouted seeds it remains unavailable in the form of phytic acid). Other good sources are heart, brain, liver, food yeast, cabbage, and citrus fruits.

PABA (para-aminobenzoic acid): It is a growth-promoting factor, and a deficiency contributes to white skin patches, eczema, greying hair, and fatigue. In lotion, it is excellent for protecting the skin against ultraviolet radiation. It is obtained from food yeast, liver, egg yolk, and is also synthesized by intestinal bacteria. As a supplement, the soluble potassium salt of PABA is preferable to the almost insoluble acid itself.

Vitamin B15 (Pangamic Acid): This bioflavonoid occurs in seeds, grains, and nuts, and it protects against oxygen deficiencies in a way similar to vitamin E. It has been successfully used in the treatment of alcoholism, allergies, arthritis, autism, breathing problems (asthma and emphysema), diabetes, hypoglycaemia, heart problems (cardiovascular diseases), and premature aging.

Vitamin B17 (Amygdalin or Nitrilosides): This nutrient contains a cyanide ingredient that is harmful to cancer cells. It is most effective for preventing cancer if taken regularly by eating a few bitter seeds of stone fruits (for example, apricot, bitter almond, apple). Other good sources are sprouted seeds, alfalfa, mung beans, millet, and lentils. Professionally, a purified product known as laetrile (now illegal in the U.S.) is used orally or as injections for cancer treatment. To be effective, laetrile must be used in conjunction with cleansing, supplements, and correct diet.

* * * * * * * * * * * *

Step 24

MINERALS

Find out what minerals do and if you have increased requirements for them.

Minerals are inorganic substances composed of a metal and a non-metal, both in ionic form. Metals most important for our health are calcium, magnesium, potassium, and sodium as bulk elements, and boron, chromium, cobalt, copper, iodine, iron, manganese, molybdenum, selenium, and zinc as trace elements. Essential non-metals are chloride, phosphorus, and sulphur. Harmful are aluminium and the heavy metals cadmium, lead, and mercury. While imbalanced intakes of bulk metals can cause health problems, trace elements easily become toxic in excessive doses.

Minerals are not well absorbed from grains, seeds, and nuts,10 unless these are sprouted or fermented. This fact is especially important for vegetarians who may rely on these foods. The addition of gelatine or chicken or fish broth to cooked grains improves the absorption of minerals, while cereal fibre (bran) decreases their availability. When grains and seeds are indicated as good sources of specific minerals in the following compilation, this applies only to their sprouted form or as fermented products.

In case of mineral deficiencies, take mineral supplements with meals containing gelatine (for example, fish or poultry) or, alternatively, with fresh vegetable juice or vegetables salads. Also make sure that you have sufficient gastric acid. Minerals are more easily absorbed as chelates or orotates; orotates deliver minerals directly into the cells.

Calcium (Ca): The RDA is 800 mg, but the recommended intake is 0.6 - 1.2 g daily. A deficiency is frequently due to overacidity, lack of vitamin D, magnesium, or boron, underactive thyroid and parathyroids, or an oversupply of phosphorus. The normal calcium-phosphorus ratio in blood is 10:4; if the calcium level is too high, calcifications result, such as stone formation, tartar, arthritic deposits, cataracts, and muscle tension.

Deficiency Symptoms: These include fragile and deformed bones; muscle cramps, twitching, and weakness; irritability; headaches; depression; menstrual problems such as excessive and painful menstruation; poor circulation; and tender breasts. Also likely are undue sensitivity to pain, insomnia, allergies, inflammations, low blood pressure, varicose veins, piles, distended veins and abdomen, swellings, slow wound healing, pyorrhoea, gingivitis, and eye problems (nearsightedness).

Best Sources: Bone broth, eggshells, dolomite, sardines, leaf vegetables, goat's-milk products, sesame seeds, and kelp.

Chromium (Cr): This mineral is active as trivalent chromium, and its biological value differs greatly in various foods. Aim for an RDA of 100 mcg of high bio-value supplements, such as glucose-tolerance factor or chelated chromium.

Deficiency Symptoms: Poor blood sugar regulation (diabetes, hypoglycaemia), disturbed fat/protein meta-b-olism, alcohol intolerance, impaired growth, high blood pressure, arteriosclerosis, weight loss, fatigue, and eye problems (opaque cornea, nearsightedness, glaucoma).

Best Sources: Brewer's yeast, molasses, mushrooms, seafood, and grains.

Calcium EAP for Healthy Cells

The health of our cells depends to a large degree on their ability to keep toxic materials out. This includes toxic chemicals, such as endotoxins produced by microbes inside the body, and also harmful concentrations of otherwise useful biological chemicals. Such toxic matter tends to weaken and disrupt the energy production within cells and is a major factor in disease processes such as cancer and chronic degenerative diseases.

Toxic chemicals, together with useful ones, can enter the cells only through special pores in the cell wall. In healthy cells, these entrance points are protected by special gatekeeper chemicals, mainly the calcium salt of 2-aminoethanol phosphate, also called colamine phosphate, or EAP. EAP is heavily concentrated at these entrance points and is attached to the cell wall where it keeps the calcium in a fixed position.

In addition to keeping unwanted chemicals from entering, EAP is also effective in transporting essential minerals through the cell wall and into the cell. In addition, it helps to maintain and restore a normal electric potential between both sides of the membrane. Because of its vital functions in maintaining the health of cell membranes, EAP has been called the "membrane integrity factor." With increasing age and disease-related health deterioration, we do not produce enough EAP to protect our cells. This leaves them open to attack by undesirable chemicals and toxins, an event that can cause or contribute to a wide range of diseases.

EAP salts were developed in the 1960s in Germany by Hans Nieper, M.D. Initially, calcium EAP was used in the successful treatment of multiple sclerosis, but the damaging after-effects of viral infections (e.g., chronic fatigue) could also be greatly reduced with EAP. Other diseases that responded well were ALS, allergies, asthma, blood pressure and circulation problems, capillary bleeding, cancer, chronic kidney disease, diabetes, and osteoporosis.

Varicose veins stopped deteriorating and aging skin was improved with EAP treatment. Dr. Nieper mentioned that his patients on long-term EAP therapy hardly aged. He found that a mixed formulation of calcium, magnesium, and potassium EAP was calming and harmonizing for the nerves and, with this, suitable for neurological conditions, including hyperactivity and anxiety disorders.

After evaluating the treatment of several thousand patients with multiple sclerosis over many years, Dr. Nieper found that hardly any of these patients had developed cancer. Also, in a six-year study of eight patients with repeated surgery for colon cancer, there were no further recurrences of cancer after calcium EAP therapy. The usual dose is one to three tablets three times daily, before meals. For information and orders, contact: Koehler Company, P.O. Box 11510, Prescott, AZ 86304.

Copper (Cu): The RDA is 1.5 to 3 mg, but the recommended intake is 3 to 5 mg daily. A dietary deficiency of copper is rare; inorganic copper, for example, may be oversupplied from copper water pipes. Internal copper deficiency can result from insufficient binding capacity within cells.

Deficiency Symptoms: These include anaemia, especially in infants; bone disorders; defective spinal cord, multiple sclerosis; hair greying, hair becoming fine and straight, and losing its curl; infertility; weak connective tissue, as evidenced by heart problems such as enlarged heart, weak aorta (holes, ruptures), aneurism, and varicose veins; hernias; cancer; leukaemia; arthritis; inflammations; parasites; and underactive thyroid.

Signs of Toxicity: Liver cirrhosis, jaundice, and symptoms of zinc deficiency.

Best Sources: Liver, food yeast, nuts, and oily seeds. Use supplements only in organic form, such as chelated or salicylate complexes. However, colloids and copper armbands are acceptable.

Iodine (I): The RDA and recommended intake for iodine is 150 mcg daily. Iodine is needed for the production of thyroid hormones, which regulate metabolism.

Deficiency Symptoms: These include goitre, cretinism, fatigue, overweight, mental retardation, debility, and cancer.

Best Sources: Kelp, seafood, and iodized salt.

Iron (Fe): The RDA is 10/15 mg, so the recommended intake is the RDA, but it should be less for the elderly. An iron deficiency causes anaemia and fatigue, while an oversupply in the inorganic form (tablets, bread additives) is a contributing factor in degenerative diseases, especially arthritis.

Best Sources: Kelp, liver, yeast, molasses, sesame seeds, and eggs.

Magnesium (Mg): The RDA is 350/280 mg; the recommended intake 500 - 1000 mg daily. This mineral activates many enzymes (see the section on magnesium chloride to follow).

Deficiency Symptoms: Here you find apathy, confusion, depression, disorientation, hallucinations, irritability, paranoia, and poor memory. Also typical are angina, arteriosclerosis, high blood pressure, sweating, tachycardia (fast pulse), strokes, thrombosis, or infarctions. You can also encounter convulsions, cramps, eclampsia, epilepsy, muscle twitching, numbness, nystagmus (rapid eye movements), tingling, and tremors. Further deficiency symptoms are physical and mental rigidity, stiffness, stone formation in kidney and gallbladder (but not renal failure), tartar, alcoholism, cancer, arthritis, diabetes, acute infections, intestinal malabsorption, liver cirrhosis, pancreatitis (inflammation of the pancreas), prostate problems, strong body odour, overactive thyroid, colic, and premenstrual tension.

Best Sources: Kelp, dolomite, seafood, sea minerals, grass juice, green leaves, nuts, oily seeds, molasses, and sprouted grains.

Manganese (Mn): The recommended intake is 5 to 10 mg daily; the therapeutic dose is 20 - 100 mg.

Deficiency Symptoms: These typically are muscle weakness, myasthenia gravis, drooping eyelids, poor memory, dark-red skin spots, -diabetes, hypoglycaemia, atherosclerosis, schizophrenia, epileptic convulsions, bone deformities, mucus problems, impotence and/or sterility, ataxia (muscle incoordination), poor equilibrium, abnormal inner ear, and retraction of head.

Best Sources: Grass juice, spinach, parsley, spices (cloves, cardamom, ginger), nuts, pea-nuts, and sprouted and fermented seeds.

Molybdenum (Mo): Molybdenum is an integral part of several enzymes involved in cell oxidation and carbohydrate metabolism. Daily requirements are not known, though we do know it is an antagonist to copper.

Deficiency Symptoms: These can be cancer of the oesophagus, dental decay, impotence in elderly males, asthma, and sensitivity to sulphites, MSM (see the section on MSM to follow), and other sulphur-rich foods.

Best Sources: Buckwheat, beans, fermented soy products, liver, and barley.

Potassium (K): The RDA is 2000 mg, but the recommended intake is 3 g daily. A deficiency may originate from an oversupply of sodium (salt), the use of diuretics, or adrenal weakness.

Deficiency Symptoms: Typical here are anxiety, nervousness, fatigue, diabetes, hypoglycaemia, heart strain, high blood pressure, liver problems, lower back pain, muscle weakness, multiple sclerosis, and overweight.

Best Sources: Kelp, fruits, and vegetables.

Selenium (Se): The RDA is 70/55 mg, while the recommended intake is 100 - 200 mcg daily; the therapeutic dose is up to 500 mcg taken in organic form. Selenium acts as an antioxidant together with vitamin E. It is lost in cooking and very toxic when taken at 2000 mcg or more. A deficiency causes viruses to become very virulent.

Deficiency Symptoms: These include fatigue, cancer, liver damage, pancreatic fibrosis, muscular dystrophy, weak immune system, reproductive disorders, cataracts, haemorrhaging, haemolytic anaemia, and underactive thyroid.

Best Sources: Kidney, liver, seafood, brewer's yeast, and kelp.

Zinc (Zn): The RDA is 15 mg; the recommended intake is 20 to 30 mg daily; and the therapeutic dose is up to 300 mg. This mineral is best taken together with vitamin B6. Diets low in protein and high in fibre, phytates, calcium, fat, and copper can decease the absorption of zinc. Increased zinc is needed with calcium or magnesium supplements, and during chronic infections, convalescence, and stress.

Deficiency Symptoms: These affect skin (acne, boils, burns, leg ulcers, parakeratosis [horny red skin], poor wound healing, psoriasis, stretch marks); hair (brittle, coarse, dandruff, falling out, lacking pigment); nails (white spots and bands); eyes (keratitis [inflammations or lesions of the cornea], night blindness, retinal detachment); taste and smell (lost or distorted); sexual function (male sex organs underdeveloped, prostate enlarged, delayed sexual maturity, sterility, menstruation retarded or irregular); and maturation (poor growth, dwarfism). Other symptoms and conditions include body and breath odour, mouth cankers, nausea (pregnancy), poor circulation, cold extremities, fainting, heart infarction, sickle-cell anaemia, cancer, diabetes, epilepsy, gastric and duodenal ulcers, inflammation of the intestines (Crohn's disease), joint pains (especially knee and hip), liver cirrhosis/alcoholism, toxaemia, mental problems, increased emotionalism, hyperactivity, learning disorders, autism, schizo-phrenia, loss of appetite, anorexia nervosa, and multiple allergies.

Best Sources: Oysters, herrings, sardines, kelp, seafood, oatmeal, liver, pumpkin seeds, and sprouted seeds.
Now let's look in depth at two minerals - magnesium chloride and MSM, a form of organic sulphur.

Magnesium Chloride for Health and Rejuvenation: Magnesium is nothing short of a miracle mineral in its healing effect on a wide range of diseases as well as its ability to rejuvenate the aging body. It is essential for many enzyme reactions, especially in regard to cellular energy production; it is essential too for the health of the brain and nervous system and for healthy teeth and bones. It may come as a surprise that in the form of magnesium chloride it is also an effective infection fighter. On the basis of its numerous beneficial effects, I recommend magnesium chloride as part of a healthy living program.

The first prominent researcher to investigate and promote the antibiotic effects of magnesium chloride was French surgeon and professor Pierre Delbet, M.D. In 1915, he was looking for a solution to cleanse the wounds of soldiers; traditionally used antiseptics actually damaged tissues and encouraged infections instead of preventing them. In all his tests, magnesium chloride solution was by far the best; not only was it harmless for tissues, but it also greatly increased white blood cell activity and phagocytosis (the destruction of harmful microbes).

Dr. Delbet performed experiments on the internal applications of magnesium chloride and found it to be a powerful immune stimulant. In his experiments, phagocytosis increased by up to 333 percent. This means that after magnesium chloride intake the same number of white blood cells destroyed up to three times more microbes than beforehand. Gradually, Dr. Delbet found magnesium chloride to be beneficial in a wide range of diseases. These included diseases of the digestive tract such as colitis and gallbladder problems, Parkinson's disease, tremors and muscle cramps, acne, eczema, psoriasis, warts and itching skin, impotence, prostatic hypertrophy, cerebral and circulatory problems, asthma, hay fever, urticaria, and anaphylactic reactions.

Hair and nails became stronger and healthier and patients had more energy. He also found a preventative effect on cancer and used magnesium chloride to reverse precancerous conditions such as leukoplasia, as well as hyperkeratosis and chronic mastitis. Epidemio-logical studies confirmed that topographical regions with magnesium-rich soil had a lower incidence of cancer in their population than regions with low magnesium levels.

Here is a review of other benefits of magnesium:

Magnesium for Nerves: Magnesium has a calming effect on the nervous system and is frequently used to promote good sleep. But more importantly, it can be used to calm irritated and overexcited nerves. This is especially useful with epileptic seizures, convulsions in pregnant women, and the "shakes" in alcoholism. Magnesium levels are generally low in alcoholics, contributing to or causing many of their health problems. If magnesium levels are low, the nerves lose control over muscle activity, respiration, and mental processes.

Due to its strong relaxing effect, magnesium helps not only with better sleep but is also useful in overcoming headaches and migraines. Nervous fatigue, tics and twitches, tremors, irritability, hypersensitivity, muscle spasms, restlessness, anxiety, confusion, disorientation, and irregular heartbeat - all these conditions respond to increased magnesium levels. A common phenomenon of magnesium deficiency is a sharp muscle reaction to an unexpected loud noise.

Many of the symptoms of Parkinson's disease can be overcome with high magnesium supplementation; shaking, for example, can be prevented, and rigidity can be eased. With pre-eclampsia, pregnant women can develop convulsions, nausea, dizziness, and headaches; in hospitals, this is treated with magnesium infusions.

Magnesium for the Heart: Adequate levels of magnesium are essential for the heart muscle. Those who die from heart attacks usually have very low magnesium but high calcium levels in their heart muscles. Patients with coronary heart disease who had been treated with large amounts of magnesium had a better survival rate than patients who had received drugs. This might be due to the fact that magnesium dilates the arteries of the heart and lowers cholesterol and fat levels.

Diabetics who are prone to atherosclerosis, fatty degeneration of the liver, and heart disease often have low magnesium tissue levels. They often develop eye problems, such as retinopathy; diabetics with the lowest magnesium levels have the most severe retinopathy. The lower the magnesium content of their drinking water, the higher is the death rate of diabetics from cardiovascular disease. In an American study, the death rate due to coronary heart disease was double and diabetes was four times higher in areas with low magnesium water levels as compared to areas with high magnesium levels.11

Magnesium for Healthy Bones and Teeth: Medical authorities claim that the widespread incidence of osteoporosis and tooth decay in Western countries can be prevented with a high calcium intake. However, published evidence reveals that the opposite is true. Asian and African populations with a very low intake (about 300 mg of calcium daily) have very little osteoporosis; Bantu women with an intake of 200 to 300 mg of calcium daily have the lowest incidence of osteoporosis in the world. In Western countries with a high intake of dairy products, the average calcium intake is about 1000 mg. Strangely, the higher the calcium intake, especially in the form of cow's milk products (except butter), the higher the incidence of osteoporosis.12

The average concentration of magnesium phosphate in bones is about one percent; in teeth about 1.5 percent; in elephant tusks two percent; and in the teeth of carnivorous animals made to crush bones, it is five percent. In regard to the strength of bones and teeth, think of calcium as chalk and of magnesium as superglue; the magnesium superglue binds and transforms the chalk into superior bones and teeth.

Magnesium, Cancer, and Aging: Many studies have shown an increased cancer rate in regions with low magnesium levels in soil and drinking water. In Egypt, the cancer rate was only about ten percent of that in Europe and America, and in the rural areas it was practically nonexistent. The main factor at play was an extremely high magnesium intake of 2.5 to 3 g in these cancer-free populations - ten times more than in most Western countries.

Let's see how calcium and magnesium are further linked to aging. We use our muscles by selectively contracting them; on the biochemical level, muscle contraction is triggered by calcium ions flowing into muscle cells. To relax the muscle, calcium is pumped out again. However, as we age, more calcium remains trapped in the muscles, and these muscles become more or less permanently contracted, leading to increasing muscle tension and spasms. Together with calcification of the joints, this is the typical rigidity and inflexibility we see in old age.

The higher our intake of calcium relative to magnesium, the faster we calcify and age. Further, most of the excess calcium in our diet ends up in our soft tissues and around joints, leading to calcification and potentially arthritic deformations, arteriosclerosis, cataracts, kidney stones, and senility. Biochemical stress can lead to the pathological calcification of almost any organ, which means the more stress, the more calcification, and the more rapid the aging.

The Rejuvenation Mineral: Magnesium and chloride have other important functions in keeping us young and healthy. Chloride is required to produce a large quantity of gastric acid each day, and it is also needed to stimulate starch-digesting enzymes. Magnesium is the mineral of rejuvenation and prevents the calcification of our organs and tissues that is characteristic of the aging-related degeneration of the body.

Using other magnesium salts (other than the chloride one) is less advantageous because these have to be converted into chlorides in the body anyway. We can use magnesium as oxide or carbonate, but then we need to produce additional hydrochloric acid to form chlorides so it can be absorbed. Many aging individuals, especially those with chronic diseases, who desperately need more magnesium cannot produce sufficient hydrochloric acid and thus cannot absorb the oxide or carbonate.

Among the other salts, Epsom salt is magnesium sulphate; it is soluble but not well absorbed and acts mainly as a laxative. Chelated magnesium is well absorbed but expensive and lacks the beneficial contribution of the chloride ions. Orotates are good but expensive for their content of magnesium; and both orotates and chelates lack the infection-fighting potential of magnesium chloride.

As suggested above, calcium and magnesium are opposites in their effects on our body structure. As a general rule, young women and children with a soft body structure need more calcium, while the more rigid and inflexible we become, the less calcium and the more magnesium we need. Magnesium can reverse the age-related degenerative calcification of our body structure and rejuvenate us.

How Much Magnesium to Take: Hydrated magnesium chloride contains about 120 mg of magnesium per gram or 600 mg per rounded teaspoon. It has a mild laxative effect. As a good maintenance intake to remain healthy, you can take up to one teaspoon daily in divided doses mixed with meals; it is best to use it instead of table salt as it has a somewhat salty taste. Individuals with very sensitive taste buds can start using magnesium in tiny amounts mixed with strongly flavoured food, then gradually increase the doses.

With acute infections, dissolve eight slightly rounded teaspoons in one quart of water. With children, a small glassful or four ounces can be used effectively every six hours. Adults can double this dose by drinking this amount every three hours, or even more frequently until diarrhea develops; then cut back to a maintenance intake just below the level of diarrhea, until the infection has cleared.

For daily use, it is more convenient to dissolve the magnesium chloride in water. Fill a bottle or jar halfway with the crystals and then to the top with water. Mix one teaspoon of this solution with food or drink three times daily for a daily intake of about 600 mg of magnesium. This or a more concentrated solution can also be used as a pack over tumours and infected, inflamed, painful, stiff, or calcified joints, muscles, adhesions, or scar tissue. It is also excellent as a back rub, to relax tense muscles anywhere on the body, and to rejuvenate aging skin. For sensitive skin, use it in a very diluted form. On wounds it can be used topically in a four-percent solution, that is, a level teaspoon in a small glass of water.

Non-hydrated or desiccated magnesium chloride contains twice as much magnesium as the hydrated product. If using this, use half the amounts indicated above. Magnesium chloride is produced from seawater. Instead of magnesium chloride, you can use the brine from the production of sea salt. It has the advantage of having more trace minerals than the technical magnesium chloride, but due to its high content of magnesium sulphate it is rather bitter.

CAUTION: While a higher magnesium intake is beneficial for most individuals, those with low blood pressure usually re-quire more calcium. Normal blood pressure is about 120/80, so the lower yours is the higher your daily intake of calcium should be. While those with high blood pressure benefit from ingesting twice as much magnesium as calcium, those with low blood pressure can take twice as much calcium as magnesium; both minerals are taken in relatively high amounts. Those with low blood pressure and a tendency towards inflammations should also greatly reduce their intake of phosphorus. A high level of phosphorus in the blood tends to cause magnesium as well as calcium levels to be low.

Magnesium and calcium supplementation should be avoided by people with severe kidney problems (severe renal insufficiency); extra magnesium may be weakening to those with myasthenia gravis. Too much magnesium can cause muscle weakness or cramps; if this happens stop using magnesium and use more calcium.

MSM - Organic Sulphur As a Versatile Healer: MSM (methyl-sulfonylmethane) occurs naturally in many fresh foods but is lost during processing. Its main use in the body is for collagen synthesis and forming skin, blood vessels, hair, and nails; it keeps cells, skin, and blood vessels elastic, which is the attribute of youth. That is why organic sulphur is often regarded as the “beauty mineral.” Cell walls remain permeable so that nutrients can freely flow into the cells and wastes and toxins can easily be removed.

Amino acid chains are usually linked with each other through flexible sulphur bridges. Also the oxidative energy production of cells requires reactive sulphur compounds. Thus, without sufficient organic sulphur as in the form of MSM, cells and body structures lose their elasticity and flexibility. What results are the well-known signs of aging: inelastic skin with increasing wrinkles, scar tissue, hardening arteries, varicose veins, and “hardened” lungs, causing emphysema. Sufficient intake of MSM is able to reverse these conditions to a significant degree, including emphysema. It is believed that the MSM in aloe vera is the active ingredient responsible for repairing damaged skin, a quality attributed to this plant.

Athletes use MSM to increase stamina and minimize sore muscles. Its use for this purpose is even more widespread with racehorses and greyhounds. This effect may be mainly due to the ability of MSM to increase greatly the body's ability to eliminate metabolic residues, wastes, and toxins from the cells. In a similar way, MSM appears to help those with chronic fatigue. And it helps us recuperate from severe physical and mental exhaustion; further, MSM reduces the effects of stress and the incidence of stress-related deaths in animals.

MSM tends to reduce or eliminate allergic reactions to foods, chemicals, and inhaled allergens. It also reduces reactions to the bites of mosquitoes, bees, and poisonous spiders and snakes. It lessens inflammation, pain, stiffness, and swellings due to arthritis or from other musculoskeletal system disorders, and it helps to normalize the blood chemistry in these conditions; it also reduces inflammations of the skin and mucous membranes. MSM relieves leg and back cramps and muscle spasms, be it after periods of inactivity (night cramps) or during athletic activities.

Other conditions that often benefit from MSM supplementation are hot flashes as well as discomfort due to the monthly cycle, acne, asthma, back pain, candidiasis, constipation, diabetes, diarrhea, diverticulitis, gastrointestinal ulcers, hypertension, inflammations of all kinds, itching skin, migraines, nausea, pain, stress, sunburn, and wound healing. It is believed to be helpful with other chronic degenerative or inflammatory conditions, especially Alzheimer's, cancer, Crohn's disease, myasthenia gravis, and Parkinson's.

Fingernails and hair have a high sulphur content and their condition generally improves with MSM supplementation. It also reduces parasite infections in the intestinal and urogenital tracts. MSM also appears to normalize our mental state. Individuals on MSM tend to report increased alertness, reduced mood swings, and less depression. It seems to improve the immune system and the senses of taste and smell. In addition, it neutralizes the toxicity of anticholinesterase, thereby providing protection against insecticide exposure or ingestion; it also aids the liver in the detoxification of chemicals and, because of this, it is useful in easing drug withdrawal symptoms.

In summary, MSM:

- inhibits pain impulses along nerve fibres (analgesia)
- lessens inflammation
- increases blood supply
- reduces muscle spasm
- softens scar tissue.

How to Use MSM: Unlike the related DMSO, MSM is a natural food ingredient and is free of unpleasant taste and odour. It is reported to be completely safe even in very high amounts. Because of its inert composition, it does not cause allergies or undesirable pharmacological effects; it can even be safely used to dilute blood. The body will use what it needs and remove any excess through the kidneys. The water-soluble MSM is easily absorbed and provides a therapeutically important source of organically bound sulphur.

The daily maintenance intake commonly ranges from 1 to 10 g taken in two or more divided doses. Adjust the dose according to your well-being or observed effects; initially I suggest you start with about 1 g and increase the dosage gradually.

You can also use MSM externally to carry other nutrients or remedies into the skin. With arthritis or connective tissue problems, for instance, you can dissolve glucosamine, copper salicylate, and MSM in a small amount of warm water and rub it into the affected area.

For bulk lots of MSM in the United States, see www.bulkmsm.com where you can also find more information on MSM. If your diet is already high in sulphur, as from plenty of onions and eggs, then you might not notice any benefit from taking MSM.

CAUTION: Individuals who are sensitive to sulphites, often used as preservatives, tend also to be sensitive to MSM due to a deficiency of molybdenum, which is required for liver enzymes to convert sulphur compounds into sulphates.

In this case, supplement with 200 to 1000 mcg of molybdenum and avoid a high copper intake.

* * * * * * * * * * *

Step 25
AMINO ACIDS

They're the building blocks of all body proteins.
Adequate levels are essential for good health.

There are eight essential amino acids for adults that must be supplied with the diet: isoleucine, leucine, lysine, methionine, phenyl-alanine, threonine, tryptophan, and valine. Two others, arginine and histidine, are essential in the first year of life. Some may be useful as supplements to enhance the functions described below, but commonly this is necessary only in cases of malabsorption of nutrients, with specific diseases, or with bodybuilding and sports nutrition. The specific amounts to use are usually not critical and may vary widely with practitioners. If not otherwise recommended by your health professional, you can use the amounts indicated on the product container. However, in most instances, instead of taking single amino acids, I recommend taking spirulina or chlorella instead; these are high in protein (up to 70 percent) and easily digestible.

Amino acids, like many vitamins, exist in two forms that are chemically identical but differently folded. These are named either D-amino acid or L-amino acid. Amino acids in our body have the L-form. Therefore, it is not advisable to use the synthetic D,L-amino acids, but rather the more expensive L-amino acids produced with enzymes. One exception to this is D,L-phenylalanine if used for pain relief. If the label of a supplement does not specify that it is the L-form, you can assume that it is the cheaper D,L-form, and I would not use it. Furthermore, two simple amino acids, glycine and taurine, exist only in one form and are not marked D or L.

Arginine: Arginine is the precursor of the neurotransmitter nitric oxide, which helps to relax and dilate blood vessels. It is in an antago-nistic balance with lysine. Arginine helps to channel nutrients into tissues and promotes regeneration.

- Aids in liver detoxification
- May retard growth of tumours and cancer cells
- Assists in the release of growth hormones
- Helps maintain a healthy immune system
- Important for scar tissue formation and collagen production
- Promotes muscle growth and fat utilization
- Helpful with malaria
- May help with male impotence

Note: avoid supplementation during pregnancy and lactation, and with herpes infections.

Carnitine: Carnitine is produced in a healthy liver from lysine, vitamins B1 and B6, and iron. Vegetarian diets can be deficient in carnitine, as it is found mainly in meat. D--carnitine is harmful. Acetyl-L-carnitine is the form best suited to treat neurological conditions.

- Aids transport of long-chain fatty acids to mitochondria for energy production
- Helps in weight loss
- Improves athletic ability through increased energy production
- Enhances effectiveness of the antioxidant vitamins E and C

Creatine: It is formed in the liver from the amino acids arginine, glycine, and methionine, or available from meat. Commonly 5 g are supplemented daily as creatine-monohydrate, mainly used by athletes and bodybuilders.

- Improves energy production in muscles and heart
- Helpful with muscle diseases, such as motor neurone disease
- Protects against brain injuries, and can improve memory

Cysteine: This is a high sulphur-content amino acid formed in the liver from methionine; it needs vitamin B6 to work well. It acts as a brain neurotransmitter; high-dose supplements of cysteine can interfere with brain functions in susceptible individuals.

- Aids in detoxifying the body
- Precursor to glutathione, a strong antioxidant
- Removes excess heavy metals, helps to burn fat
- Increases muscle growth
- Breaks down mucus in the respiratory tract
- Useful for burn and wound healing
- Assists in the supply of insulin

Glutamic Acid: Glutamic acid is a major neurotransmitter in the brain, so it must be avoided as a supplement as it can cause imbalances in susceptible people. Monosodium glutamate or MSG (the sodium salt of glutamic acid) is a main ingredient of food flavouring added to processed food and is variously labelled as hydrolyzed vegetable protein, soy protein extract, yeast extract, beef stock, caseinate, or natural flavouring. Evidence shows that it can kill brain cells in susceptible individuals.

- Important for neurotransmission
- Metabolism of sugars and fats
- Detoxifies ammonia (when used with glutamine)
- Together with glucose, is the brain's principal fuel

Glutamine: As mentioned, glutamine is a principle source of energy, especially for the brain, and helps to regenerate the intestinal wall. With lack of energy and difficulty gaining weight (as in advanced cancer), 10 g or more can be used daily in divided doses.

- Provides about 80 percent of the body's pool of free nitrogen
- Reduces cravings for sugar, alcohol, and other drugs

- Improves absorption of other nutrients, important with inability to gain weight (cachexia)
- Useful with impotence, allergies, senility, fatigue, and peptic ulcers
- Converted in the brain to glutamic acid and to GABA, both of which act as neurotransmitters

Glycine: This is the simplest amino acid. It has a sweet taste and acts as a nitrogen pool for the synthesis of other amino compounds.

- Prevents or slows muscle degeneration
- Needed for a healthy prostate
- Aids central nervous system functions
- Useful in epilepsy and depression
- Important for immune system, building block for nonessential amino acids

Histidine: Histidine dilates blood vessels, lowers blood pressure, and chelates minerals. It is an amino acid of special importance to the developing infant.

- Helpful for tissue growth and repair
- Important for digestion and ulcer healing
- Controls gastric acidity
- Important for production of red and white blood cells
- Possible benefits in rheumatoid arthritis

Isoleucine: An essential amino acid and one of three branch-chain amino acids (BCAAs), which form an essential part of muscle tissue protein.

- Needed for haemoglobin production
- Stabilizes blood sugar and energy levels

Note: Use in combination with leucine and valine; deficiency of isoleucine can produce hypoglycaemia-like symptoms.

Leucine: An essential branch-chain amino acid (BCAA) important in muscle protein.

- Lowers elevated blood sugar levels
- Promotes healing of bone, skin, and muscle tissue

Note: Use with correct balance of isoleucine and valine; excessive amounts can produce hypoglycaemia.

Lysine: An essential amino acid that must be derived from the diet.

- Essential building block for all protein
- Needed for proper growth and bone development in children
- Helps calcium absorption
- Maintains nitrogen balance

- Effective against cold sores and herpes viruses, possibly also beneficial with other viruses
- Increases antibody production
- Needed for collagen production
- Helps to build muscle
- Lowers serum triglycerides
- Deficiency results in loss of energy, inability to concentrate, irritability, retarded growth, and hair los

Methionine: This is an essential sulphur amino acid that must be obtained from the diet. Supplementation can be harmful with an existing vitamin B6 deficiency, so therefore use both at the same time.

- Needed for the metabolism of fat
- Aids in liver detoxification
- Involved in the synthesis of the amino acids cysteine and taurine
- Helps to reduce effects of chemical sensitivities
- Aids digestive processes
- Involved in production of choline

Ornithine: Ornithine is helpful with insomnia, possibly by removing ammonia from the brain.

- Helps release growth hormone
- Needed for immune system function
- Promotes healing
- Detoxifies ammonia

Phenylalanine: An essential amino acid, D- or D,L-phenylalanine can be used for long-acting pain relief. Do not use during pregnancy or with anxiety attacks, high blood pressure, or phenylketonuria (PKU).

- Used in treatment of depression
- Essential for the production of neurotransmitters
- Involved in the formation of the skin pigment melanin

Taurine: Taurine is a sulphur amino acid and a main component of bile salts, essential for fat absorption.

- Promotes emulsifying and absorbing fats
- Helps regulate rhythm and mineral concentrations in the heart
- An inhibitory (sedating) neurotransmitter, helps with epilepsy

Threonine: This is one of the essential amino acids.

- Maintains protein balance

- Needed for the formation of collagen and elastine
- Aids the liver and fat metabolism
- Helps control epilepsy

Tryptophan: This is another essential amino acid.

- Important for the production of the vitamin niacin or niacinamide
- Helps prevent insomnia and elevates mood
- Controls hyperactivity in children
- Needed for production of serotonin and melatonin

Tyrosine: Tyrosine is the precursor of stimulating neurotransmitters and hormones, in particular adrenaline, dopamine, and thyroxin.

- Involved in the production of melanin pigment of skin and hair
- Important for adrenal, thyroid, and pituitary gland function
- Helps control appetite and body fat levels; is an antidepressant
- May assist during withdrawal from addictive drugs
- Provides raw material for the synthesis of epinephrine (adrenaline) and dopamine

Valine: This is an essential amino acid in muscle protein.

- Important for muscle tissue maintenance and muscle coordination
- Stimulates and promotes mental vigor

* * * * * * * * * * * *

Step 26

DIGESTIVE ENZYMES

**These vital substances are needed for digestion,
but poor diet and cooking practices can deplete your supply.**

Digestive enzymes are released from our digestive organs in order to break down food into molecules small enough for it to be absorbed. Deficiencies of digestive enzymes are widespread and are caused mainly by overeating, eating too much cooked food, food allergies, and incorrect food combining. Virtually all elderly people and those with chronic degenerative diseases and allergies have digestive enzyme deficiencies.

The enzymes in raw and fermented foods help in the digestive process but are destroyed by temperatures over 120° F. Further, many heated proteins become difficult to digest and you may need more digestive enzymes to handle them. Some raw seeds, soybeans, broad beans, wheatgerm, and nuts contain enzyme inhibitors, but sprouting and cooking inactivates these.

In light of this, make a deliberate effort to improve the enzyme content and digestibility of your food. If you are in a state of chronically lowered health, it will be very helpful to use enzyme supplements. This is essential in the case of cancer and other advanced degenerative diseases and also if you have allergies, unless you are already on a predominantly raw food diet.

Most important are the pancreatic enzymes, known collectively as pancreatin and available as a supplement under various brand names. For serious diseases, take 4000 to 5000 mg of pancreatin in divided doses in the course of each meal. One tablet may contain 1200 or 1600 mg of pancreatin; this is sometimes also described as pancreatin 4NF and if so take 300 or 400 mg of it. In addition, you can take one or two tablets of an enzyme combination containing papain, bromelain, pepsin, or bile. Pepsin is indicated for protein foods if the stomach is weak, while bile is helpful if there is a problem with fat absorption. Papain and bromelain help protein digestion.

In the case of degenerative diseases and advancing age, gastric acid is often deficient. If so, you can take hydrochloric acid–pepsin tablets, especially when eating cooked protein meals. Alternatively, dilute one part commercial diluted hydrochloric acid (usually 20 percent) with nine parts water. Mix one teaspoon (use plastic) of this with meals or drink with fluids after a meal. As a rule, additional enzymes are not needed with uncooked meals.

However, digestive enzyme supplements are not required if protein-digesting enzymes can be obtained from food, such as bromelain from pineapples or papain from unheated leaves or flowers or unripe fruit of papaya. Ripe kiwi fruit is a good source of enzymes. Use any of these with protein meals. You can also predigest meat, fish, or other protein food by refrigerating it overnight wrapped or mixed with papaya leaf.

* * * * * * * * * * * *

Step 27
NUTRITIONAL SUPPLEMENTS
Concentrated nutrients in pill form can shore up nutritional deficiencies and build optimal health.

Most individuals on conventional diets with much refined and processed foods have either outright nutrient deficiencies or shortages of various vitamins, minerals, enzymes, and hormones. Generally, our intake of essential nutrients is less than optimal, making us susceptible to disease. Deficiencies are even more widespread in those who already suffer an acute or chronic disease. Sometimes individuals have had to suffer for years and have had high medical expenses for conventional treatments when a few nutritional supplements could likely have corrected the problems.

Supplements are most important for individuals with nutrient malabsorption and for those who still use processed and refined foods as part of their diet or who cannot regularly obtain vegetables and fruit grown organically in good soil. The more serious the disease or health deterioration, the more supplements are likely to be beneficial and the greater should be the dosages and range of supplements taken.

Calcium supplements are commonly used with conventional diets and yet many people are apprehensive of obtaining sufficient calcium when abstaining from milk products. This concern is unfounded. Calcium is not well absorbed from cow's milk, as can be seen in the high incidence of osteoporosis developing in people on conventional Western diets and the fact this is not a problem in milk-free Asian countries. Vegetables and sprouted and fermented seeds are high in usable calcium; the calcium content in juice from cereal grasses is higher than in milk as, after all, cows obtain their calcium from grasses in the first place.

Similar considerations apply to iron, which can easily become harmful as an inorganic supplement. Much more widespread and detrimental than calcium and iron deficiencies are deficiencies in chromium, magnesium, manganese, selenium, and zinc.

Many supplements are actually concentrated health foods, but taking these is not the same as the "pill-popping" of medical drugs. While I believe that supplements are unnecessary if one is healthy and lives on a high quality diet in an unpolluted environment, I know that it is difficult to remain healthy in present times and even more difficult to heal an "incurable" disease. Carefully selected supplements can make a great difference in these endeavours, especially for those who are just starting to improve their diet.

I am not in favour of using multivitamin tablets as they often contain synthetic forms of vitamins A, D, and E and beta-carotene. Instead, if required, it is better to use a low-potency B-complex in addition to vitamin C and cod-liver oil or halibut-liver oil capsules. Mineral supplements are often more important than vitamins and can contain approximately 15 mg zinc, 5 mg manganese, 100 to 200 mcg selenium, 200 mcg chromium, 500 mg magnesium, and up to twice this amount of calcium if one's blood pressure is low. Much higher doses, especially of selected vitamins, can be used where indicated for therapeutic effects.

In addition, you can use general supplements to increase your overall well-being and vitality. Of special benefit for those lacking energy are ginseng, coenzyme Q10, royal jelly, and freeze-dried liver. Bee pollen and spirulina or chlorella are my favourites.

As mentioned earlier, digestive enzymes and hydrochloric acid supplements can help those with weak digestion. For those in conditions of debility, amino acid and nucleic acid supplements can help, the most helpful amino acids being lysine, tyrosine, methionine, and glutamine.

Many supplements such as royal jelly, vitamin B12, and the freeze-dried animal glands are much more effective if partly absorbed under the tongue, chewed lightly, and kept in the mouth for as long as possible.

Lecithin is beneficial for most individuals, and you can mix a teaspoonful with meals or with ground linseed (flaxseed). Omega-3 essential fatty acids are universally indicated for which linseed oil or cod-liver oil and other fish oils can be taken; if malabsorption is a problem, the oil can be rubbed into the skin. These oils help to reduce inflammation, pain, and allergic swellings. On a diet low in iodized salt and seafood, we can become iodine deficient; kelp tablets or granules are excellent for supplying iodine as well as many other minerals. With infections, use high levels of vitamins C, A, and E, zinc, bee propolis, and echinacea.

Sulphur compounds, as mentioned earlier, detoxify and remove from the body heavy metals, harmful chemicals, and drugs; MSM is especially good for this. Sulphur is also a main component of connective tissue, so supplementation is beneficial with all connective tissue diseases (for example, arthritis). Even simply taking powdered sulphur has been shown to increase the formation of useful sulphur compounds. You can take a gelatine capsule filled with sulphur once a day or mix a pinch of it into each meal. The main sulphur amino acids are taurine, L-cysteine, and L-methionine; the main sulphur foods are egg and onion.

Generally, gradually increase and decrease high levels of supplements, especially vitamins, and interrupt or greatly reduce supplement intake during fasting or cleansing periods. Most vitamin-mineral supplements are best absorbed and cause less gastrointestinal irritation if mixed with the meal as powders, crushed tablets, or opened capsules. The next best way is to take tablets near the middle of the meal and avoid slow-release tablets. However, take amino acids before the meal; this applies also to calcium as it is not well absorbed and tends to form insoluble combinations (soaps) with fatty acids during digestion.

Organic or Natural Versus Inorganic or Synthetic: There is some confusion about the importance and meaning of natural or organic versus synthetic or inorganic minerals. In food production, “organic” means that food has been grown, produced, stored, and processed without the use of synthetic chemicals. However, in chemistry “organic” means that a chemical is based on carbon and has covalent bonds; “inorganic” means that a chemical is not based on carbon and has ionic bonds.

Some inorganic chemicals, such as sodium chloride, magnesium chloride, and calcium chloride, are ingredients of natural and organic foods and are also essential components of our body. Calcium in our blood plasma, for instance, is about half in ionized form while the other half is bound to proteins. In this way, taking some minerals in inorganic form can rectify ionic deficiencies of these minerals.

Further, a healthy intestinal flora can transform inorganic into organic minerals, while inorganic minerals taken with gelatine or fresh green vegetable juice can easily be incorporated into organic structures. This is of special importance with some of the trace minerals. Even so, it is safer to use these directly in organic form, best obtained from fresh, organic vegetable and grass juice, mineral-rich food or plant concentrates, or, alternatively, safely bound to amino acids or as stable complexes such as copper and zinc salicylate.

However, there is a downside to routinely adding inorganic minerals, such as table salt, to our food. We may use concentrations that are too high and irritate the intestinal wall, and we may also cause mineral imbalances. Therefore, use not only the right concentrations but the right combinations of ingredients in your supplements. It is much easier to get this right by using natural foods. Of course, the main reason we use supplements is because commercial food has inadequate mineral content; even in much of organically raised food, the mineral content is not as high as it might be.

Another way of making minerals safer and more effective is by using them in colloidal form. In this way, even tiny clusters of metal atoms can be safely and beneficially ingested. Colloidal mineral products are commercially available, made from ancient decomposed plant material. Also beneficial is the intake of minerals present in energized water.

After sufficient health improvement, you may feel that you no longer need most nutritional supplements. With a good maintenance diet, I recommend continuing indefinitely to use the following health foods, or equivalent ones: several teaspoons each of bee pollen, ground linseed, and spirulina or chlorella; half a teaspoon of kelp; and one tablespoon of cod-liver oil (emulsified with lecithin and fresh juice); also some lecithin and magnesium chloride with meals will be helpful.

Selection of Vitamin Supplements: Commercial growing and processing and traditional Western cooking methods usually reduce the vitamin and mineral content of foods to only a fraction of that found in fresh, raw, organically grown foods. As an example, methyl bromide, a widely used fumigant for grains, destroys all the pantothenic acid (a B vitamin) in cereal grains.

Unfortunately, these days many vitamins (especially vitamin C and the B vitamins) are not available in sufficient amounts in foods or natural supplements, and synthetic supplements are often necessary in order to improve or maintain our health. Vitamins available as natural concentrates (A, D, and E) should be avoided in synthetic form. Natural vitamin E is d-alpha tocopherol; the synthetic form is dl-alpha tocopherol. Highly recommended is vitamin E complex as mixed tocopherols.

Store vitamin supplements in a dry, cool, and dark place. Oil-filled capsules should be taken during a meal containing oily or fatty foods and lecithin. However, many individuals most in need of these vitamins have absorption problems and for them oil-filled capsules are not suitable. Therefore, I generally recommend using tablets of natural vitamin E rather than oil-filled capsules. If in doubt about your digestive abilities, take a halibut oil or other fish oil as a capsule at bedtime and keep it in the mouth to be slowly absorbed overnight. Other-wise, take all supplements during meals or mixed with food. Small amounts with several meals are better absorbed than a large amount in a single dose.

Some tablets have beta-carotene substituted for vitamin A. However, many individuals have difficulty in absorbing beta-carotene and converting it into vitamin A. Therefore, use tablets that contain real vitamin A and not -vitamin-A equivalents of beta-carotene.

Chewing vitamin and mineral supplements together with food produces much better results than just swallowing them. Appreciable amounts of vitamin B12 can be absorbed by keeping a tablet under the tongue, even if it cannot be absorbed in the stomach because of a lack of “intrinsic factor” (a substance usually present in the stomach that is involved in the absorption of vitamin B12).

If there are signs of increased requirements for individual nutrients, use these together with a B-complex supplement. In most conditions involving lack of energy, debility, anemia, low blood pressure, and chronic degenerative diseases, initial injections and later mouth-absorbed tablets of vitamin B12 (1000 mcg daily) are very helpful.

The acid form of vitamin C (ascorbic acid) can be taken with protein meals, especially if your system is too alkaline or lacking in gastric acid. If regularly taking more than 3 to 5 g per day of vitamin C, those with high blood pressure can partly neutralize the ascorbic acid with an equal amount of milk of magnesia, or with one-third the amount of potassium bicarbonate or magnesium carbonate. With low blood pressure and possible overacidity, take calcium ascorbate instead.

Antioxidants are of greatest importance for minimizing oxidative tissue damage and, with this, for avoiding diseases and delaying aging. In a more recent development, a high intake of antioxidants has also been found to be crucial in preventing and overcoming viral infections. The major antioxidants include the vitamins A, C, and E, the minerals selenium and zinc as components of antioxidant enzyme systems, the plant pigments beta-carotene, bioflavo-noids, and anthocyanins, as well as phenols and catechols in grape seed extract, green tea, and generally in coloured fruit and vegetables. Other powerful antioxidants are alpha-lipoic acid and negative hydrogen as in Microhydrin or electrolytic reduced water (ERW) produced with water ionizers. Whether you take antioxidants as supplements or as part of a high-quality diet, make sure you get plenty.

Are Nutritional Supplements Dangerous?: Generally, health authorities promote pharmaceutical drugs as beneficial and natural remedies as harmful. They tend to call natural remedies “drugs,” and the view has been put forward that synthetic nutrients are the same as natural ones. Research results using synthetic products, such as vitamins A, D, E, or beta-carotene, with negative or harmful outcomes are then used as “scientific proof” that natural remedies are either useless or dangerous.

While numerous trials have shown significant health benefits for individuals with higher levels of beta-carotene in their blood, clinical trials with synthetic beta-carotene and synthetic vitamin E have shown increased rates of death due to cancer.

Vitamin A has been portrayed as the most “dangerous” supplement, largely based on studies with synthetic vitamin A. There are dire warnings that it can cause death in amounts higher than the RDA (5000 IU). However, many people take 25,000 IU or more daily without problems, and as far as I can ascertain, there has never been a death recorded due to vitamin A supplements in the U.S. in the 60 years that these have been in use. There appears to be only one recorded death in all the scientific literature, that of a chemist in England in the 1950s who ingested two million IU daily of retinyl acetate for three months and died of liver cirrhosis.13

I am not aware of any pharmaceutical drug having such a good safety record as that. Even one of the safest drugs, aspirin, has been stated to cause about 100 deaths a year, and many more cases than that of toxicity.14 In comparison the reported cases of (usually easily reversible) vitamin A toxicity in the U.S. are typically five per year.15

Similar unfounded attacks have been made against vitamins C and E. For more detailed information on this subject, see the website of the Life Extension Foundation (www.lef.org); for articles on megavitamin therapy; also see www.doctoryourself.com.

Imagine the mass media frenzy if there were actually some fatalities due to vitamin A. Yet there is no concern at all about thousands of patients dying, in my view unnecessarily, due to prescription drugs. In the public, the perception has been generated that fatalities due to prescription drugs are a small price to pay for their overall benefit. This benefit has never been shown in the long term. Drug approval is based on short-term trials, and only by accident does the truth of long-term harm occasionally emerge, as in the case of hormone-replacement therapy. I believe that in the long term all drugs become a burden on the liver or kidneys by increasing our load of toxic chemicals and may accelerate our health deterioration. Therefore, with any health problem, inform yourself about all available options.

* * * * * * * * * * * *

Step 28
HERBS

You may benefit from using herbal remedies instead of pharmaceutical drugs.

Herbs are not completely harmless, but damage due to commonly used herbs is extremely rare, while the benefits are usually genuine and lasting. In contrast to culinary herbs, medicinal herbs should be used on a long-term basis only if there is a clinical reason and not just in case they might do some good. (Herbs for a wide range of health problems are discussed in part 6.) Here I will discuss a few herbs for some common conditions and show how the application can be particularly important on your road to better health.

General Herbal Remedies: A good combination to cleanse the gastrointestinal tract is a mixture of chillies and ginger root. Blend or crush both in lemon juice and olive oil and refrigerate, then use this as a salad dressing or for flavouring meals. Use an amount that is hot but not uncomfortable. Gradually increase your intake to just below your tolerance level for several weeks; then you can continue to use it in normal amounts for flavouring your food. With serious infections, inflammations, or cancer of the gastrointestinal tract, you can also add fresh garlic to this until the problem is overcome.

Peppermint tea is especially good to stimulate the digestion and uplift the spirit, while slippery elm powder soothes the digestive tract. Sage is highly esteemed. An ancient saying goes: “Why die [of a disease] when sage grows in your garden?” It can be prepared by adding two teaspoons of dried sage to one pint of boiling water; after three minutes, remove it from the heat and let it draw for another ten minutes. Sage can be combined with other herbs. Fragrant or leaf herbs such as peppermint or St. John’s Wort can be added while the herbs steep. St. John’s Wort is effective against depression.

The immune system can be strengthened with echinacea, propolis, European mistletoe (Viscum album), Easter lily, and reishi and other Asian mushrooms. Echinacea can be added to all other herb combinations. However, there are some research findings that echinacea can cause problems if taken uninterruptedly for several months, so it may be advisable to interrupt its use after a month and then to take it again a few weeks later.

Wormwood and garlic are excellent against all kinds of parasites as well as the yeast Candida albicans. A combination of goldenseal, ginger, and slippery elm powder is good for gastrointestinal inflammations and multiple sclerosis. Ginger is also good against motion sickness. In addition to being an excellent wound healer, comfrey is a good detoxifier; you can drink a cupful of comfrey tea two or three times a week.

To aid digestion, it is generally beneficial to drink half a cup of bitter herb tea after a large or cooked meal. Suitable herbs here are centaury, dandelion (not roasted), devil’s claw, and gentian; these also help to restore liver function. With any lung conditions, including lung cancer, have a daily steam inhalation (with a towel over the head) of hot chamomile tea, possibly with the addition of eucalyptus or tea tree oil or crushed onion.

Maca is a root vegetable from the Peruvian Andes, and I regard it as the best remedy for balancing the endocrine glands and hormone system. It appears to act directly on the hypothalamus and pituitary systems to regulate the thyroid, adrenals, and sex glands. It is often effective with underactive thyroid, as an alternative to hormone replacement therapy and hysterectomy, for age-related decline in the production of sex hormones, and with general lack of energy. An average daily intake is about 10 g, but start with less.

Licorice, too, has some effect on sex hormones but is most useful as a stimulant of adrenal function. With this, it can raise low blood pressure and energy levels and is useful for chronic fatigue syndrome, peptic ulcers, skin inflammations, and HIV/AIDS; it also increases pancreatic secretions. However, it causes retention of sodium and low potassium levels and should not be used with high blood pressure or kidney disease, or during pregnancy. The normal therapeutic range is from 2 - 6g daily.

Anticancer Herbs: Many herbs are useful in cancer treatment, mostly in a supportive role but sometimes as a main therapy. Frequently used are aloe vera (as juice or gel), chaparral, pau d'arco, red clover, and violet leaves. Wormwood and pau d'arco in sufficiently large amounts are especially beneficial by eliminating the fungal forms of the cancer microbe. Vinca rosea, a periwinkle, has an active ingredient that has been used in chemotherapy against leukaemia.

Milk thistle and bitter herbs, such as burdock, greater celandine, centaury, dandelion leaves and root, devil's claw, and gentian, help stimulate the liver and release bile. Celandine or greater celandine inhibits division of cancer cells and stimulates the immune system; an extract for injection is available as Ukrain. Also useful are leaves of the tropical grariola tree, commonly known as soursop.

Root herbs, such as the bitter liver herbs, need to be simmered for about half an hour, while leaves or flowers need to simmer or steep for only five minutes.
For the after-effects of radiation, wild crane's bill (Geranium robertianum) is helpful. With lung cancer use also horehound, lobelia, mullein, and St. John's Wort. Pimpernel is used partly to gargle and partly to drink in spoonful amounts for cancer of the mouth, throat, and neck glands.

Mistletoe is especially good for leukaemia and is often used as the Anthroposophic medicine Iscador or Plenosol for injection near tumours and swollen lymph glands. With prostate cancer, use also saw palmetto and/or willow herb (Epilobium, mainly parviflorum). With leukaemia, especially use herbs to strengthen the immune system.

Licorice root, taken in typical doses of 500 to 1000 mg, is strongly recommended instead of Tamoxifen with all estrogen-related cancers, as it blocks the growth-stimulating properties of estrogen. It also promotes anti-inflammatory adrenal hormones. However, it tends to keep sodium in the body while potassium is expelled. Therefore, minimize sodium and maximize potassium intake by consuming plenty of vegetable juices; do not use licorice root if your blood pressure is dangerously high or with obvious water retention.

You can select a range of these herbs that are suitable for your condition or that are locally available and use them as your special herb mixture. Alternatively or in addition, you can use a commercially available anticancer herb mixture. Best known are the Jason Winter's mixture, Hoxsey herbs, and Ojibway Indian Tea or Essiac (sheep sorrel, burdock root, slippery elm, and rhubarb root).

The Hoxsey mixture must generally be made up individually, but a local herbalist should be able to do it. The exact quantities are kept secret and may vary, but an independent analysis revealed the following quantities in 5 ml of the Hoxsey tonic: potassium iodide (150 mg), Stillingia root (10 mg), burdock root (10 mg), red clover (20 mg), buckthorn bark (20 mg), berberis root (10 mg), pokeroot (10 mg), prickly ash bark (5 mg), liquorice (20 mg), and Cascara amarga (5 mg). Cascara amarga is difficult to obtain, and the mixture can be made without it. It belongs to the buckthorn family and is also known as Honduras bark or Picramnia antidesma. This tonic is sold by the Hoxsey Clinic,16 while individual preparations commonly consist of a dried herb mixture, which is used as a tea.

CAUTION:
Use herbs for cancer treatment only in consultation with a credentialed health practitioner.

Aloe Vera: Aloe vera is excellent for all skin conditions and beneficial in the treatment of cancer, AIDS, and other infectious diseases. It is also good for gastrointestinal irritation, inflammation, and ulcers. However, not all commercial brands of aloe on the market are effective. If nothing seems to be happening with the aloe you're using, try another brand.

Preferably use fresh aloe vera gel straight from the plant, as much as you can take without getting diarrhea. Slice off the green skin and ingest just the gel, but you can also blend a small whole leaf with fruit juice or other flavouring. For wounds, it is usually better to slice part of a leaf in half and leave the skin on top. Normally avoid the base of the leaf with its yellow sap, as this is a strong irritant and laxative. Aloe vera is most beneficial for those with blood group A; for those with other blood groups, check for compatibility or observe effects.

* * * * * * * * * * *

Step 29

HOMEOPATHY

Homeopathic remedies are effective, and you can even make them yourself.

Homeopathy is a system of medicine based on the principle that “like cures like.” A substance that produces disease-like symptoms in healthy people will in properly diluted and potentized form cure a sick person with similar symptoms. For example, straight onion juice will cause watering, irritated eyes and sneezing, but in the highly diluted and potentized homeopathic form of onion (Allium cepa), it is useful in relieving symptoms of eye and nose irritation, as in hay fever. Clinical trials and the empirical evidence of two centuries of daily use show that homeopathy is effective, even though conventional scientists are at a loss to understand how this is possible.

This is because homeopathy cannot be understood as a substance-based method, such as pharmaceutical medicine or herbalism, in which healing effects appear to be due to chemical actions and reactions. A simple calculation shows that in the higher potencies not a single molecule of the original substance is present. But these high potencies are often more effective than low potencies. Therefore, homeopathy can only be understood as a system that works with energy remedies.

The explanation is that every natural substance has an etheric or subtle energy field around its molecular structure. The special shaking action or potentising used in homeopathy separates the etheric fields from their material substance. The latter is then discarded and the former concentrated. These concentrated etheric fields are much more potent in this way than when they were combined with matter. Similar etheric concentrates can be produced with radionics instruments. The healing effect of such concentrates depends on their vibrational frequency; due to this the term “vibrational medicine” has been coined for these methods.

A homeopathic medicine’s sphere of curative action is originally found by administering it in high doses to healthy human volunteers and recording their emotional, pathological, and physiological reactions to the medicine. This test is called a proving, and hundreds of substances have been proved or tested in this way.

Consequently, homeopathic remedies are effective over the whole range of mental and physical disorders. Most homeopathic medicines are made from plants, minerals, animals, or chemicals; remedies from human disease microbes are called nosodes. Sometimes a product of a disease is used in potentized form for treating the same disease; this method is called isopathy.

Homeopaths may use the classical approach or practice clinical homeopathy. The classical approach is to use only one dose of one remedy at a time and another dose or remedy later only if effects of the previous remedy have worn off. More commonly used, clinical homeopathy prescribes several remedies in combination to be taken several times daily.

Use Your Own Body Fluids: While for general treatment of common symptoms it is more convenient to use established homeopathic remedies, to overcome obstinate individual problems it is frequently more effective to prepare remedies from your own body fluids. This a form of isopathy known as auto-therapy.

Urine, for example, contains traces of almost all body chemicals, such as enzymes and hormones, as well as foreign and toxic substances and viruses. A properly diluted sample of your urine can initiate the correction of all unbalanced, unnatural body chemistry. This, then, is an excellent way to improve the balance of your glandular system, which is difficult to influence directly by any other method.

Other body fluids that can be used are a drop of diluted blood to treat blood disorders; diluted pus if there is an infected wound or a boil; sputum for mouth and throat disorders; tears for eye problems; faeces for intestinal diseases; ear discharge for ear diseases; and so on. Apart from balancing gland activities and general body chemistry, this method is known to be very effective in stimulating the body's immune system, which is essential in combating cancer and other degenerative diseases.

How to Prepare Your Own Remedies: To prepare remedies from your own body fluids, dilute a small sample of the selected fluid 1:100 with pure water. This means one part sample and 99 parts water, for instance, one teaspoon of sample and 99 teaspoons of water. The water should be purified, that is, distilled or filtered.

Shake the mixture vigorously 40 or 50 times in a sealed bottle or jar, accentuating the downward stroke. Alternatively, strike the mixing bottle (not glass) 40 to 50 times on a resilient surface. This sample is now in the first potency, described as 1C. Take one part of this, dilute it again with 99 parts water, shake as previously, and you have a potency of 2C. This process is repeated until the desired potency is reached.

If dilution proceeds in steps of 1:10 instead of 1:100, the first potency is 1X, the second 2X, and so on. A potency of 12X corresponds in dilution strength to 6C. However, the number of potencies as well as the overall dilution are important for the quality of a remedy. Start using your own diseased body fluids at a potency of 12C. If this does not bring an improvement after some time, experiment with higher and lower potencies. However, use urine in 6X potency and start foods to which you may be allergic at 4X before increasing in steps to 5X and 6X. Important: Use remedies made of your own body products only for yourself, not for anyone else.

To produce your own remedies from herbs, make a "mother tincture" by soaking a fresh or dried sample in high-grade alcohol, vodka, or brandy. After a week or two, dilute part of the mother tincture as previously described, but for the first 1:10 dilution use half alcohol and half water. With this technique, you need to buy a specific commercial tincture only once. If the bottle is almost empty, dilute its contents with either 9 parts or 99 parts water-alcohol mixture and shake for some time.

You might also try a combination of homeopathy and herbalism. Herbs often seem to be more effective when used as low-potency homeo-pathic remedies. In this way any undesirable side effects of concentrated herbs are lessened, while specific healing effects, especially cleansing actions, are greatly strengthened. You can experiment by potentising one part of an unheated herbal extract or tincture with four or nine parts water.

If you use a strong mother tincture of a potent herb, start by taking only one or two drops several times daily; keep them in the mouth for a while, and increase the dose gradually over time. If the extract is weak or of a harmless herb, then take spoonfuls or even a cupful at a time. You can make a watery extract of fresh or dried herbs in a blender, or better, by shaking the powdered or pureed herb in water. Alternatively, you can make a tincture by keeping the herbs covered with alcohol or glycerine in a jar for several days.

If the remedy is to be stored for more than a few days, use a mixture of half vodka or brandy and half water for the last step. This gives the solution an alcohol content of about 30 percent and will prevent the growth of bacteria and fungi. Alternatively, keep the remedy cold, but not in an electric refrigerator as the remedy can lose its effectiveness due to exposure to the strong electromagnetic field of the refrigerator.

How to Take Your Remedies: Take homeopathic remedies with a clean mouth, meaning free of food; the best times are before meals (about 30 minutes) and at bedtime. The usual dose is three to five drops in a little water (about two ounces) sipped slowly over five minutes. Alternatively, you can place three drops in a little water under the tongue. Generally, take one dose four times daily; for chronic conditions, twice daily; for severe or acute conditions, every ten minutes until some improvement is noted, then reduce the frequency. If the patient cannot swallow, moisten the inside of the lips and the tongue with the remedy. Try not to have the remedy come in contact with metal, use a plastic rather than metal spoon, and avoid strongly scented or aromatic drinks and toothpastes, as these may cancel out the remedy.

If a reaction to the remedy is noted (sometimes as a temporary aggravation of symptoms), this is a sign that the correct remedy was selected, but wait until the reaction has subsided before giving the next dose. Any overdose of homeopathic remedies is usually harmless, although in sensitive individuals, there is the risk of proving the remedy—getting the symptoms it is supposed to relieve. Store homeopathic remedies protected from heat and light, as well as away from magnets and strong electromagnetic fields. For best results, you should not change remedies too frequently.

If pellets are used, do not touch them, but shake a tablet into the bottle cap and from there directly into the mouth; keep it under the tongue until it dissolves. Generally, however, it is not recommended to use homeopathic remedies in tablet form (usually as sugar pills) because of the possibility of intolerance to lactose or sucrose. This is accentuated by keeping the tablets in the mouth, which provokes stronger allergic reactions than swallowing them does; however, swallowing is not desirable for homeopathic remedies as they are primarily absorbed in the mouth.

For detailed descriptions of homeopathic remedies and their range of action, see the Pocket Manual of Homeopathic Materia Medica with Repertory by William Boericke.

Which Potency to Use: The most commonly used potencies are 6X, 12X, 12C, 30C, 200C, I000C (M), I00M (CM), 500M (DM), and 1000M (MM). Sometimes the letter C is omitted, and 30 and 200 are used instead of 30C and 200C. In Europe, D is used instead of X. After selecting the correct remedy according to your symptoms, it is important to choose the correct potency as well. A general rule is that the lower potencies affect mainly physical symptoms while the high potencies work at the mental and emotional energy levels. The high-potency remedies can be used only once a day or once a week or after any reaction has subsided.

For a long-standing problem, you can start using a low-potency remedy, say 12X; after a week or two, use 30C, then 200C, and afterwards M at intervals. If you wish to use potencies closer in strength, try 12X and 30C. You may also be able to find the correct potency by using muscle testing or a pendulum. Draw a half-circle on which the various potencies are marked and see to which one the pendulum points. The M to MM potencies have been used to remove heavy metals from the body either after an acute poisoning or from chronic exposure, as in the case of chronic mercury poisoning from teeth filled with mercury amalgam.

In addition to producing low-potency remedies from allergenic foods, you can also buy M to MM remedies of foods to which you are allergic. If you have had much X-ray treatment, buy and use an X-ray remedy (see www.gdr.org/drrita.html); if you are sensitive to electricity, fluorescent lighting, or house dust, use the corresponding commercial remedy. However, you can also make these remedies yourself: Expose pure water to strong electromagnetic fields or prolonged fluorescent lighting and then potentize it, or potentize the contents of the vacuum cleaner.

Meanwhile, there are commercial remedies for various cancers (for example, bowel, lung, stomach) and for bacilli and other infectious agents; there are also extracts of glands, such as adrenals and thymus, which can be used to normalize glandular functions.

Combinations of remedies are available for the treatment of a wide range of diseases and common complaints. If someone has been bitten by a poisonous snake, spider, or other creature, try to obtain some of the poison by crushing the head of the insect in water or by sucking some of the poison out of the wound; or use some blood from the wound. Make and use a 30C potency. A healing centre might establish a collection of potentized remedies of the common poisonous creatures of the area.

Homeopathic Cell Salts: In the late nineteenth century, the German physician W. H. Schuessler originated a therapeutic system he called biochemic medicine, based on 12 essential cell salts or tissue salts. These are homeopathic preparations of mineral salts usually present in the body and commonly used as 6X tablets. The main drawback of the tablets is that they are based on lactose, to which many individuals are allergic or react to weakly in muscle testing. This weakening factor is accentuated by the common practice of keeping the tablets in the mouth until dissolved.

However, for those who should best avoid lactose, the benefits of cell salts can be obtained by using cell salts as normal homeopathic tinctures.

The 12 Schuessler cell salts and their main indications are:

- Calc. flour: over-relaxed abdominal wall, piles, varicose veins
- Calc. phos.: anemia, poor circulation, cramps, rheumatism
- Calc. sulf.: boils, catarrh, yellow pus, and scabs

Your Own Home-Remedies Kit

The following is a list of 15 basic remedies for home use and as a travel kit.

ACONITE 30C: Controls feverish symptoms that come on suddenly. Any acute condition in the early stages. Aconite is indicated when the patient is frightened and/or worse after dark.

A.G.E. 30C (ARS. IOD., GELSEMIUM, EUPATOR): A combination remedy for colds and influenza. Contains a mineral remedy and two herbal preparations to cover all stages of a cold from early onset to fully developed influenza and lingering final stages that will not clear away. One dose fortnightly will reduce the incidence of simple colds and influenza. This combination remedy is also useful for apprehensive or fearful states. Use also for stomach and bowel upset as well as aching in the bones and headaches that are better after sweating.

ARNICA 30C: This is the remedy for accidents and shock. It reduces the effects of injury, fatigue, and shock, both mental and physical. It is indicated for bruising, broken bones, and blood blisters, also for stroke and heart attack, and in cases where the patient fears being approached or touched.

BELLADONNA 30C: Covers feverish conditions that are more violent than those treated by aconite: red, dry, hot painful head and throat symptoms; also indicated for angry outbursts, and for patients whose condition is aggravated by light, noise, or jarring.

CHAMOMILLA 30C: For children during teething or in cases when the patient is irritable. Suitable for stomach upsets; indicated for children who improve on being carried.

COCCULUS 30C: Useful for nausea, vomiting, travel sickness, morning sickness, stomach upsets, and exhaustion.

HYPERICUM 30C: For injuries to nerves and any injury that is more painful than would be expected. Good for blows to the head and coccyx, crushed fingers and toes, injuries to lips, ears, or genitals; intolerable violent shooting pains along nerves; or lingering pains after injury or surgery.

HYPERICUM/CALENDULA MOTHER TINCTURE (HYPER-CAL): An antiseptic, use topically (five drops in a cup of tepid water) for all external wounds, cuts, and eruptions. This mixture can be used to rinse out the mouth after extraction of teeth. Apply to burns and press a pad soaked in this mixture to injuries requiring bandaging; use one drop in an eyebath of water for eye irritation.

IGNATIA 30C: An herbal stress reliever for the effects of bereavement, grief, tension, worry, and fright. It is suitable for hypochondria and for real illness when the patient is excessively worried about the condition; it promotes sleep when the patient is suffering from sadness or worry.

LEDUM 30C: Used for puncture wounds, cuts, and stings of insects. Follows arnica for bad bruising, especially around the eye; anti-tetanus. Think of ledum for pain of fractures and joints.

NUX VOMICA 3OC: For stomach upsets caused by overeating or indulgence in rich foods, and for the effects of too much alcoholic drink, tobacco, or drugs. Indicated for irritability and bad temper and for sleeplessness when the mind is crowded with thoughts. Useful in cases when the patient is easily chilled and cannot be uncovered. Also for stuffy head colds extending to throat or chest; and sour eructation, sweat, vomit, and bowel movements.

PULSATILLA 30C: For the effects of rich foods, earache from water in the ear, and measles. Useful for the patient who is changeable and weepy, whose pains shift, or who needs open air when the heat of the room or a car is unbearable. Indicated for many children's illnesses when aconite has given limited relief; also for menstrual problems.

RHUS TOX 30C: For simple rheumatism and for strains, sprains, and the effects of overexertion. Useful when the patient has been exposed to cold, wet conditions; also when restless and pains are better if the affected part is kept in motion. Suitable for influenza not covered by A.G.E.

RUTA GRAV 30C: Ruta is the best remedy for twisted ankles and sprained wrists. It is excellent for blows to the muscles and for injuries to the bones and accidents in which the bone sheath is damaged; it brings relief from eyestrain, neuralgia, burning eyes, dim vision, and runny eyes.

URTICA URENS 30C: For burns, rashes, stings, allergies, and any condition in which burning, stinging pains persist. Use for scalds and burns internally when using Hypercal externally. Use urtica for bladder complaints, bedwetting, hives, and swollen, painful breasts and genitals.

- Ferr. phos.: inflammations, low fever, anemia, circulation, wounds
- Kali. mur.: croup, catarrh, and skin diseases with white discharge
- Kali. phos: high fever, paralysis, sepsis, neuralgia, debility
- Kali. sulf.: catarrh, skin diseases with watery discharge
- Mag. phos.: shooting pains, neuralgia, spasms, itching
- Nat. mur.: constipation or diarrhea, cold hands or feet, drowsiness
- Nat. phos.: gout and rheumatism, overacidity, blackheads
- Nat. sulf.: liver and gallbladder diseases, dropsy, migraine
- Silicea: boils, skin ulcers, nasal catarrh, night sweats, gout

They may be used individually or in combination.

Flower Remedies: Flower remedies represent a combination of herbalism and homeo-pathy. Their main therapeutic effects are on the emotional level. Best known are the Bach flower remedies originated in the 1930s by Dr. Edward Bach, an English physician. In recent times, Australian wildflower remedies and Californian flower essences have also become available; these are sold mainly in health food stores and by various practitioners.

Flower remedies are best used together with appropriate affirmations. In clinical situations, it is recommended to use a standard questionnaire to assess a client's emotional problems to allow a more speedy and accurate selection of the appropriate remedies. This information can then be used for other -emotion-directed therapies. Of special significance is the Bach Rescue Remedy, which should be part of every first-aid kit. Individual flower remedies can be combined for treatment according to symptoms. In selecting Bach flower remedies, it is recommended to combine no more than six different remedies at one time; however, the composite Rescue Remedy can be counted as a single remedy.

The usual dosage is two to four drops of the composite preparation in a teaspoonful (plastic) of water at least four times daily, taken before meals and especially as the first and last thing daily. Hold this solution in your mouth and under the tongue for some time; store remedies in a cool place. In some cases, flower remedies can be taken for long periods for the treatment of chronic illnesses, while those for acute (short-term) problems are used only when the problem is noticed. These short-term remedies can be changed frequently, according to changing moods.

While the flower remedies are usually selected according to the problems you wish to overcome, you can also look at what you wish to achieve, how you wish to become.

For this, you can use the opposite to the descriptions given for the various remedies. The keywords for the positive attributes we wish to acquire (the negative aspects we wish to overcome are in parentheses) and some of the flower remedies indicated are as follows:

- Attentiveness (being dreamy): Clematis

Reference Guide for the Selection of Bach Flower Remedies

- Agrimony: suffering or worry hidden by smiling face
- Aspen: apprehension, vague fears of unknown origin
- Beech: intolerance, arrogance, tendency to be overcritical
- Centaury: susceptibility to influence and exploitation, weak will, timidity
- Cerato: lack of confidence, doubt, constant requests for advice
- Cherry Plum: uncontrolled temper, desperation, fear of loss of mind
- Chestnut Bud: slowness to learn, lack of observation, repetition of mistakes
- Chicory: possessiveness, self-love, self-pity, attention seeking
- Clematis: indifference, dreaminess, lack of interest
- Crab Apple: feeling of being unclean, shame, self-dislike; the cleanser
- Elm: occasional feelings of inadequacy, too much responsibility
- Gentian: discouragement, self-doubt, negative attitudes, depression
- Gorse: hopelessness, despair, despondency
- Heather: self-centeredness, love of talking, inability to listen
- Holly: envy, suspicion, revenge, hatred, jealousy
- Honeysuckle: living too much in the past, homesickness
- Hornbeam: mental and physical weariness
- Impatiens: impatience, irritability, mental tension
- Larch: lack of confidence, expectation of failure, feeling of inferiority
- Mimulus: fear of known things, shyness, timidity
- Mustard: deep gloom or depression of unknown cause
- Oak: onward struggle despite despondency/ despair
- Olive: mental and physical exhaustion or weariness
- Pine: self-blame, self-reproach, guilt, false humility
- Red Chestnut: fear and anxiety for others
- Rock Rose: terror, panic, extreme fear
- Rock Water: self-criticism, rigidity, self-denial
- Scleranthus: uncertainty, indecision, hesitation, lack of balance
- Star of Bethlehem: after-effects of shock and trauma
- Sweet Chestnut: extreme anguish, hopeless despair, loss of faith
- Vervain: over-enthusiasm, fanaticism, nervous tension, strain
- Vine: dominance, leadership, craving for power, ambition, inflexibility
- Walnut: helps in changes, link breaking, protects
- Water Violet: aloofness, reserve, pride
- White Chestnut: unwanted persistent thoughts, worry, inner mental arguments
- Wild Oat: dissatisfaction at not having found goal, uncertainty
- Wild Rose: resignation, lack of interest, apathy
- Willow: resentment, bitterness, “not-fair” attitude
- Rescue Remedy (a composite of Cherry Plum, Clematis, Impatiens, Rock Rose, and Star of Bethlehem): use for shock, terror, upsets, fright, accidents; give frequently in all emergencies; apply externally on wounds.

- Certainty (indecision): Scleranthus
- Courage (terror): Rock Rose
- Faith (doubt): Gentian
- Humility (pride): Water Violet
- Patience (impatience): Impatiens
- Peace (torment): Agrimony
- Service (self-love): Chicory
- Strength (weakness): Centaury
- Sympathy (fear): Mimulus
- Understanding (fanaticism): Vervain
- Wisdom (foolishness): Cerato

* * * * * * * * * * *

Step 30
COLLOIDAL SILVER, COPPER, ZINC & GOLD
Use these powerful remedies whenever you have an infection or inflammation.

In contrast to medical antibiotics that can be used only against specific microbes, colloidal silver has a wide range of effectiveness, eliminating bacteria as well as viruses and fungi. Unlike medical antibiotics, colloidal silver is not known to cause undesirable side effects. Colloidal zinc can be especially effective against viruses. The colloids of copper and gold, on the other hand, are strong anti-inflammatory agents; in addition, copper and zinc can be used chelated (bound) with salicylic acid for similar benefits.

At the start of your health improvement program, you can use colloidal silver to cleanse your body of undesirable microbes. With chronic or acute inflammatory conditions, experiment with copper and gold. Most diseases and painful conditions are associated with inflammations and benefit from anti-inflammatory remedies. A zinc-boron-salicylate complex is known as the Schweitzer Formula (see page 165) and is not only an anti-inflammatory, but it also generally supports and speeds up healing processes.

Colloidal Silver: Colloidal silver has a long history in medicine as a natural antibiotic. It is very effective in killing bacteria, viruses, fungi, parasites, and the cancer microbe, and no serious side effects from overdosing have been reported.

Colloidal silver is commercially available with a common solution strength of five silver atoms per million molecules of water, or 5 ppm, but it may be up to 50 ppm in very concentrated remedies. However, the size of the colloids may be more important than the strength in parts per million.

Colloid particles can have from ten to 5,000 atoms, and each particle is a fighting unit. Therefore, a colloid with the smallest size of particles is 500 times more effective than one with the largest size. Further, large colloids are not stable and easily fall out and settle at the bottom of the container; they are also not well absorbed by the body and not effective against viruses. Therefore, when choosing a commercial product or making your own, always aim for the smallest particle size.

There is no research yet on proper dosage, but it will vary according to concentration and particle size. You can use about one teaspoon three to five times daily; try this with any infectious disease or local infections; continue until symptoms of the infection subside and then continue for a few weeks longer with one teaspoon a day. It is best to keep it in the mouth before swallowing. During the height of the infection, you can double or triple the above amounts for a few days. At the beginning of your health improvement efforts and several times each year, you can take a

CAUTION:
Do not use colloidal gold, silver, copper or zinc if you suspect that you are sensitive to them or if any unfamiliar symptoms arise. If in doubt, seek professional advice.

I recommend using these remedies only for a specific purpose and a limited period, generally not more that a few months at a time.

remedial course of colloidal silver for several weeks to keep down any undesirable microbes that have arisen in your body. Have a sip or tablespoonful three times daily. Preferably, store colloidal silver in a dark glass bottle in a dark and cool place.

It is difficult to assess which of the many commercial remedies available work best or even which ones work and which do not. Therefore I recommend building your own colloidal silver generator (see sidebar).

Colloids of gold, copper, and zinc are being used experimentally with apparently good health results. They all have strong anti-inflammatory properties in addition to varying degrees of antimicrobial effects. Furthermore, zinc and copper colloids can be expected to help overcome any deficiency symptoms or conditions with increased nutrient requirements. Gold, on the other hand, has supposedly beneficial effects on the blood circulation and the glandular system. The medical psychic Edgar Cayce said that gold and silver, properly used, could almost double our present life span. It is not clear what he meant by properly used, but I assume it may involve the safe use of either colloids or vibrational medicines.

If you want to experiment, then you can produce these other colloids in the same way as described for silver colloids; use electrodes of copper, zinc, or gold instead of silver. The easiest way will be to use some gold coins or strips of copper or zinc. No recommendations are available for effective dosages and you will have to experiment, starting with small amounts and going long intervals without any colloids. Use careful self-observation and preferably muscle testing to assess your body reactions; remember you have to take responsibility for these experiments yourself.

How to Make Your Own Colloidal Silver Generator:

Commonly three 9-volt batteries are connected in series (positive to negative pole) and an insulated wire soldered to each end terminal. To each of the other ends of the wires, solder an alligator clip. Use ten inches of pure silver wire, preferably .999 fine, about 14 gauge or 1.64 mm; sterling silver is only .9275 fine. Cut the silver wire in half and attach to the alligator clips.

Immerse the electrodes about four inches deep in a glass of water. Commonly, colloidal silver is made from normal drinking water, often with the addition of a drop of saline solution. A white mist can be seen to develop almost immediately. This is mainly ineffective silver chloride. In this approach, the individual colloidal particles are quite large and not very effective.

A better way is to use distilled water without the addition of any saline drops, as this produces a very fine and much more effective colloid, though it takes a long time to start. A white mist may become visible after about 20 minutes. If you do see this fine mist near one of the electrodes, continue for five minutes longer. If you cannot see a slight mist after 20 minutes, then next time add a small amount of impure water (such as tap water) to speed up the colloid formation. The concentration will be about 5 ppm.

After each use, scrub the oxidized positive electrode until it is shiny. When making or storing the silver water, keep it protected from light, which will otherwise precipitate the silver and make it ineffective. Make your colloidal silver fresh each day or at least once a week.
For more information on making and using colloidal silver, see the website: www.quantumbalancing.com.

Copper - The Inflammation Fighter: Copper is somewhat of a problem mineral. Sometimes there is too much of the inorganic form of copper in drinking water, as is produced by acid water flowing through copper pipes. It can then gradually accumulate in the body and lead to toxicity symptoms with signs of overstimulation, psychosis, and liver damage. However, copper is excellent for reducing inflammations, strengthening connective tissue, restoring hair color, fighting parasites, and preventing cancer. Copper should not be used with existing cancers (except topically for skin cancers), as it promotes the formation of new blood vessels in tumours (angiogenesis).

Copper serum levels are elevated with inflammations and many diseases, apparently because the body mobilizes all tissue stores of copper to fight the condition. If you give animals a choice between drinking normal water and water in which a copper pipe has been immersed, reportedly they will prefer the copper water as this helps keep them free of -parasites.

The most effective anti-inflammatory agents are copper complexes, and commonly these are relatives of salicylic acid which is present in many fruits and vegetables. In addition, copper ascorbate has strong antiviral properties while copper salicylate has a better anti--inflammatory effect than cortisone and without the side effects. It also has good anticancer, anti-tremor, and anticonvulsive properties, suitable for treatment in epilepsy and possibly Parkinson’s disease; even certain brain and liver functions can improve with copper salicylate. It has inhibited tumour promotion and malignancy with animal tumors,[17] and with skin tumours a single application resulted in a 55-percent tumour reduction in 20 weeks.[18] Even short-term treatment of rheumatoid arthritis resulted in long-term remissions or improvements.[19]

The therapeutic potency and safety of the copper complexes of aspirin (acetyl-salicylic acid) and salicylic acid are much better than for aspirin itself or for inorganic copper. This means that these copper complexes are safer and more effective than either aspirin or inorganic copper. These complexes are five to eight times more effective than aspirin but less toxic, and their therapeutic index (the margin between effectiveness and toxic effects) has been stated to be significantly greater than for other anti-inflammatory drugs.

While aspirin causes or aggravates peptic ulcers and bleeding, the copper complexes have a better ulcer-healing effect than commonly used anti-inflammatory ulcer drugs. Harmful effects of aspirin and similar drugs apparently arise because they bind copper in the body and cause a localized copper deficiency in the tissues.

Unfortunately, copper salicylate or other effective copper complexes are not normally available or can be had only in very low doses, presumably because they cannot be patented. However, they are relatively easy to make for someone who wants to experiment. Buy salicylic acid at a drug store and dissolve 2 g or about half a teaspoon in a pint of hot water. Use covered glassware and distilled or deionised water with several pieces of copper with a large surface area immersed in it. Keep it warm, between 140° F and 180° F, for about ten hours or sometimes longer. Add more water as required and adjust the final volume to be one pint. When ready, it develops a green color; store the solution in a brown bottle in a cool, dark place.

During storage and sometimes near the end of the heating process, a black copper oxide starts forming and accumulates at the bottom. This is due to copper being converted from the original one-valence copper (I) salicylate to two-valence copper (II) salicylate; here the salicylic acid binds only half of the dissolved copper and the rest becomes copper oxide. The effectiveness of the solution does not seem to be affected by this and the amount of copper in the complex is not related to its potency. Either decant the solution from any settled copper oxide or filter it through tissue paper.

Copper (I) salicylate is a strong antioxidant excellent for external use in packs or rubs on sites of inflammation, or as rubbed on skin prone to skin cancer. For internal use with generalized inflammations or other indications, 60 mg of copper salicylate has been used in clinical trials, taken once or twice a day and up to four times daily. Try a teaspoonful (approximately 25 mg of copper salicylate) three times daily in liquids with meals, but preferably under professional medical supervision.

For short-term use, you can also double this amount. When it produces the desired effect, cut back to a maintenance dose of one teaspoon a day or interrupt the intake after two weeks to see what happens.

There is a possibility that you may be able to obtain copper (II) salicylate from a supplier of laboratory or fine chemicals. In this case, divide a rounded teaspoonful into 100 equal parts; each part will be approximately 50 mg. However, I want to stress that all this is only experimental and no one but yourself can take any responsibility for what you are doing, so proceed carefully.

For a potent antiviral remedy, you can produce copper ascorbate. The only problem is that ascorbic acid may become oxidized in contact with metal. Therefore, all air must be excluded. Bring some distilled or deionised water to a boil and fill a small glass or hard plastic container (holding about 4 ounces) to the top with this. Immerse a piece of copper in it and add about 2 g or half a teaspoon of ascorbic acid powder. Keep the container tightly closed and immersed in warm water for a day before you start taking a teaspoonful several times a day; keep refrigerated when not in use.

Zinc and the Schweitzer Formula: Zinc has strong anti-inflammatory and antibiotic properties but can become deficient with a high copper intake. Therefore, it is usually best to increase the intake of both minerals together. With a high copper intake, use a high zinc supplementation. This can be in the form of the Schweitzer Formula, a complex formed by zinc (oxide or carbonate), boron (boric acid), and salicylic acid. This is an excellent antibiotic, disinfectant, fungicide, anti-inflammatory, and healing remedy.

The Schweitzer Formula20 was developed in 1915 in Germany and has been sold worldwide since 1920. In addition to treating any kind of infection or inflammation, it has been used in cancer treatment to improve the immune response and blood oxygenation. Applied externally, it helps to heal injuries and skin diseases, including acne, scarring, varicose veins, and varicose ulcers.

You can easily make the Schweitzer Formula yourself. Dissolve 9.2 g of salicylic acid, 2.1 g of boric acid, and 2.7 g of zinc oxide (or 4 g of zinc carbonate) in two quarts of hot water. You can get these ingredients from a pharmacist and have the exact quantities weighed out. However, it is sufficient to use approximate amounts. Use two level teaspoons of salicylic acid and half a teaspoon each of boric acid and zinc oxide or one level teaspoon of zinc carbonate.

Use distilled or deionised water and a non-metallic container. Heat the mixture for about an hour and stir occasionally with a non-metal spoon until no more of the zinc oxide or zinc carbonate at the bottom of the container seems to dissolve. Then decant or filter into a glass container and store in a cool, dark place.

Any surplus of zinc oxide or carbonate that remains undissolved shows that all the boric acid and salicylic acid have been used up. However, any surplus of boric acid would be beneficial and supply you with additional boron. In commercial preparations, the resulting solution is very slowly evaporated to let the active ingredients form crystals; these are then dissolved before use in the appropriate amount of water.

As with copper salicylate, there are no exact guidelines on how much to take. In studies, a tablespoonful has been taken three times daily with liquid or meals for extended periods; for shorter periods, this dose has been doubled. It is also good to rub onto the skin, especially where there are any problems.

You can take a ratio of one tablespoon of Schweitzer Formula to one teaspoon of copper salicylate. For long-term use, I suggest one tablespoon of each daily. I believe that long-term use of copper or zinc should be balanced by taking the other mineral as well, be it as salicylate complex, colloid, or conventional remedy. My experience suggests that copper and zinc as complexes or colloids are safer and more effective than the long-term use of aspirin or other anti-inflammatory drugs.

After a period of use, it is advisable to interrupt using these remedies for a while and observe any effects. Remember, the use of these remedies over long periods is experimental and you must be prepared to take responsibility for any unexpected side effects.

* * * * * * * * * * *

Step 31

OXYGEN THERAPY

Oxygen, therapeutically applied, helps you quickly overcome any infection.

As an infection fighter, oxygen therapy is mainly used as ozone or hydrogen peroxide. It is beneficial for general health improvement and also for a wide range of diseases, mainly involving infectious agents. Ozone infusions are widely used in German and Mexican cancer clinics but are illegal in the United States, and hydrogen peroxide infusions have been used instead.

Therefore, most cancer victims try to help themselves with hydrogen peroxide or so-called stabilized oxygen. The latter is an unstable oxygenated chlorine compound that easily gives off oxygen. The typical amount used is ten drops in juice several times daily, but start with much less and increase gradually. You can combine taking hydrogen peroxide on rising and at bedtime with taking stabilized oxygen before meals. This is suitable for those who have difficulty digesting much hydrogen peroxide; otherwise, I regard hydrogen peroxide not only as the least expensive, but also the most suitable product for home oxygen therapy.

Another possibility is to use an ozone generator. Generators for medical or internal quality ozone can be expensive. Note that common ozone generators ozonate room air and produce potentially toxic nitrogen oxides; a suitable water ozonator (Beck type) is available for internal use. If ozone comes into contact with water, then it forms hydrogen peroxide. As this formation takes some time (about 20 minutes for all the ozone dissolved in water to be converted to hydrogen peroxide), you get a mixture of both if you drink it immediately after finishing the process.

Ozone reportedly has the advantage of releasing active oxygen without requiring enzymes in biological systems, as is required for hydrogen peroxide. Both ozone and hydrogen peroxide have an unstable single oxygen molecule that tends to attach to and oxidize or peroxidise any suitable molecule. The beneficial effect appears to be mainly the peroxidation of unsaturated fatty acids, which can then destroy microbes and malignant cells, provided these are not protected by accumulated vitamin E. A healthy immune system routinely produces hydrogen peroxide to kill microbes. Therefore, I see no advantage in using ozone at home and would rather use the simpler hydrogen peroxide.

Unheated honey has traditionally been used to treat infected wounds and chronic skin or leg ulcers. Recent research showed that the most effective honey for medicinal purposes is honey that is very high in hydrogen peroxide. Therefore, you can mix a few drops of 35-percent hydrogen peroxide with honey to improve its healing qualities. For topical packs, it is more convenient to use crystallized honey rather than honey that is runny.

Preferably, use 35-percent food-grade hydrogen peroxide, available from some natural therapists and health stores. Otherwise, use any other 35-percent product or the three-percent (ten volume) or six-percent (20 volume) peroxide from a supermarket or pharmacy. It is best to dilute the peroxide for daily use to about three percent; add one part of the 35-percent peroxide to ten parts of water. Start taking a few drops in a glass of water, then gradually increase to one teaspoon or more four times daily, before meals and at bedtime. If you develop an aversion to it, decrease the dose again to a more comfortable level. Swallow quickly instead of sipping slowly. To disguise the strong aftertaste, add to the mixture aloe vera gel, fruit flavours, herb tea, or water flavoured with cinnamon.

You can continue this program for many months, and when the specific health problem is under control, gradually reduce the intake to only once daily before breakfast. In addition, you can use hydrogen peroxide to rinse your mouth and gargle morning and evening. It helps to prevent colds, infections, and tooth decay, and it is arguably better than toothpaste. You can also use 35-percent hydrogen peroxide with a dropper; a teaspoon of three-percent peroxide is equivalent to about 12 drops of 35-percent peroxide. Be careful handling 35-percent hydrogen peroxide; in the event of accidental spills on clothes or skin, flush immediately with water or immerse in water.

It is helpful to bathe the site of an infection or external tumour daily in three-percent hydrogen peroxide or, better, cover it with a cloth soaked in it. This is especially good for problems close to the skin, and the hydrogen peroxide is preferably combined with a honey pack. Renew the pack several times daily. A pack can remain on the skin for most of the time or until the skin becomes red or sore and blisters. This will be good, as it draws out toxins. When this happens, cover the skin with a cabbage leaf instead and wash only with peroxide. You can also have a daily body rub with three-percent hydrogen peroxide, which is best done after a bath or shower. With cervical cancer or bowel cancer, douche the cervix or the bowel twice daily with suitably diluted hydrogen peroxide, retaining it inside for some time.

In the case of cancer it is advisable to use the practice of oxidative tumour destruction only initially and with a minimum of antioxidants. Activated oxygen oxidizes and peroxidizes reactive biomolecules, especially alpha- and gamma-linolenic acids in linseed (flaxseed), evening primrose, and borage oils; these in turn oxidize and destroy cancer cells. However, as mentioned, this killing effect may be inhibited by high levels of antioxidants, especially vitamins E and A, which, paradoxically, accumulate in tumours and protect them. Therefore, avoid vitamin E and high doses of vitamin A for one or two weeks before your oxygen therapy and minimize your intake of all antioxidants during the several weeks of intensive oxygen therapy. Afterwards, return to a long-term intake of high amounts of antioxidants and anti-inflammatory remedies (see step 46).

Magnesium peroxide has become available as a powder, in capsules, or as a branded product called Oxy-Mag, a powdered magnesium ozonite. The peroxide produces highly reactive oxygen radicals that are more aggressive than the oxygen molecules released by the ozonite, but I cannot say at present if either one of these is more effective. Use according to instructions on the container or enclosed literature. Sometimes simple magnesium oxide is recommended to oxidize the colon; while it may be good as a mild laxative, it does not liberate oxygen for oxygen therapy.

Of great benefit are peroxide retention enemas. After a normal enema to clean the bowels, insert into the colon a pint of water containing a teaspoon of salt and one tablespoon of three-percent hydrogen peroxide; try to hold it inside the large intestine for ten minutes or longer. The next day add two tablespoons to this mixture; continue increasing the amount of peroxide on subsequent days until you find the highest concentration that you can hold for about ten minutes. You can also try to increase the volume of water that you can retain.

If you have an ozone maker, you can use ozone in a similar way as hydrogen peroxide. You can lead it with a tube into water to produce hydrogen peroxide; insert the tube into the rectum (with another shorter tube as outlet); irradiate the skin over the diseased areas with a funnel; or put the body (but not the head) into a bag into which you introduce ozone. For information on how to obtain, construct and use a medical-grade ozone generator that does not oxidize nitrogen, see: www.quantumbalancing.com. For more information on oxygen therapies, see: www.oxytherapy.com or the books Oxygen Therapies and Flood Your Body with Oxygen, by Ed McCabe.

New Ways of Using Hydrogen Peroxide: In 1928, Dr. Richard Simmons believed to have discovered that colds and flu enter through the ear canal and start the infection process in the middle ear. More recently, it has been claimed that colds and flu can be cured within hours by putting hydrogen peroxide into the ears at the first indication of any infection. You first put a few drops of three-percent hydrogen peroxide in one ear, and when the bubbling stops, do the same in the other ear. Repeat this a few times at about two-hour intervals. In addition, gargle frequently with hydrogen peroxide for infections in the head and chest area.

My understanding is that the cold virus enters through the nose but then multiplies in mucus accumulated in the head spaces before the outbreak of symptoms. For this reason, colds can be stopped by introducing hydrogen peroxide into head spaces, but also simply by removing the mucus congestions.

A three-percent hydrogen peroxide solution can be sprayed directly into the throat several times a day and inhaled deeply into the lungs with good results. Besides breathing problems and aching muscles, this has also reportedly relieved other old-age symptoms. I assume that beneficial results with this method are largely due to the elimination of fungi and other microbes in the breathing passages.

* * * * * * * * * * * *

Part 4

HEALING FOODS

Nutrition is the main tool I have used to help patients overcome a wide range of diseases claimed to be medically incurable. Unnatural nutrition is one of the major causes of most of our diseases, while natural nutrition is an equally powerful agent in helping to undo these diseases and rejuvenate the body. My training and work experience as research chemist, biochemist, toxicologist, practicing nutritionist, and natural therapist have given me insight that would not have been possible if I had worked only as a conventional scientist or natural therapist. One of these insights is the importance of nutrition.

There is presently a trend in conventional medicine to acknowledge the importance of nutrition in causing and treating diseases. Health authorities conduct public awareness campaigns about nutrition through the mass media. While this newly found official interest is itself a positive development, the information conveyed is not likely to reduce the overall rate of chronic diseases.21 The information I present has grown out of practical experience in the prevention and reversal of most of our chronic and medically incurable diseases, and it is likely to be more useful.

Most of our commonly eaten foods contribute to our gradual health deterioration and the development of many chronic diseases. This is a main reason why there is such a high incidence of chronic degenerative disease in our society. I will explain in detail why and how our everyday problem foods help cause our diseases and how potent healthy foods can be used to undo these diseases and rejuvenate our body.

Depending on your health goals and present health problems, you can either make some gradual adjustment to your present diet or go all the way and experiment with powerful health diets designed to reverse incurable diseases and rejuvenate body and mind. The reward in terms of improved health is generally proportional to your effort.

Diet rules are designed to guide us towards healthier eating habits. There are different schools of nutrition, which do not always agree on the same principles and sometimes even contradict each other. The diet “rules” I present are based on the principles of the nineteenth-century Nature Cure movement, but modified according to the findings of modern nutrition research and my clinical experiences in overcoming diseases.

For the successful elimination of diseases, I have found that the less ideal the patient’s overall diet is, the more specific the remedies, supplements, or other therapies need to be. Conversely, on an ideal diet, specific interventions can be kept to a minimum. I have also observed that the differences in nutritional requirements between individuals begin to disappear when an ideal natural diet is adopted.

Changing our diet to an ideal natural one requires a considerable shift in eating habits and a re-education of our taste buds. It often also requires more time for food preparation. For most of us this is not easy. While some take to it enthusiastically, others need to be driven by the hope of overcoming a debilitating disease; most people are content to make minor dietary adjustments or gradual improvements as a preventive measure or to overcome an existing health problem. Therefore, I use a graded approach, with a high-quality diet for gradual health improvement and maintenance of good health and a raw food diet for maximum healing, rejuvenation, and maintenance of optimal health into old age.

I present each diet in its pure form. Most individuals probably will not be able to adopt fully one or another diet, but will move gradually towards their dietary goal in months or even years. You can modify the diet according to your special requirements, and for this, see the section on Metabolic Types and Blood Groups (Step 43) in Part 5. Ultimately, it is best to learn to trust your body to tell you what it needs.

If you are not addicted to a certain food but crave it, then it is probably good for you. The natural function of our taste buds is to guide us in the choice and quantity of our food intake. As a general rule, food that is good for us tastes good, but if we have enough of it, then it does not taste as good toward the end of the meal. If we overeat, then it may not taste good for several days or weeks. If we had too much heavy or spicy food, then fruits will taste attractive, but if we had too much fruit, then a spicy protein-fat meal suddenly tastes good.

If we eat food that the body has difficulty in handling, then it causes a stress reaction with the release of adrenal-stimulating hormones. For individuals with weak adrenal glands, this acts as a stimulation of their body functions, and they become addicted to this food as well as other stimulants. Now the taste buds are no longer a reliable guide, as other body needs are sacrificed to satisfy this need for adrenal stimulation.

As some foods frequently cause health problems - “problem” foods - so other foods generally improve our health - “health” foods. The foods with the greatest potential for healing and rejuvenation, according to my experience, are bee pollen, sprouted seeds, purple foods, freshly pressed juices of vegetables (especially of young grasses), and lactic-acid fermented foods. These form the base of my recommended diets and will be explained in this chapter.

We need to realize that we can choose to a considerable degree how healthy we will be and how long we want to live in good health or, alternatively, which diseases we are willing to accept for the comfort of satisfying our present taste preferences and maintaining potentially harmful living habits. Of course, other factors have a strong influence as well, such as our mental and emotional condition, the environment, and how we use our body (these are covered in later parts).

* * * * * * * * * * * *

Step 32

GENERAL DIETARY RULES

Understand the principles of healthy food selection, preparation, and combination.

In a nutshell, the most important dietary rule is, as much as possible, to eat your food when it's whole, alive, organic, and free of added chemicals. Other important rules are to chew it well, enjoy what you eat, combine the foods correctly, and deliberately under-eat rather than overeat.

Whole food means that there has been as little refining of it as possible. To give an example, natural whole grain brown rice is the preferred whole food. White rice has the bran and germ removed, together with most of its vitamins and minerals. The level of vitamin B1, for instance, is about 450 percent higher in brown rice than in white rice. This means that the body may not use the deficient food efficiently, and this can lead to overweight or underweight conditions and to poor blood sugar regulation; the lack of fibre in refined rice also contributes to constipation. In varying degrees, the effect of refining also applies to other commonly eaten foods.

Over thousands of years, our metabolic and enzyme structures have evolved based on whole foods and have relied on certain combinations of ingredients being supplied together. Protective or synergistic (cooperative) factors are present in whole foods. For example, vitamin C is much better absorbed and retained if taken together with bioflavonoids. Both nutrients naturally occur in the same foods. The sensitive polyunsaturated oils are protected from oxidation and structural damage as long as they remain in the whole seed. They also occur naturally with protective vitamin E. When extracted and refined, these oils are partly oxidized and harmful trans-fatty acids are formed.

Carnivorous animals do not eat just the meat but also the bones. They would quickly become mineral deficient and diseased if they did not eat the bones. The same applies to us if we eat only flesh high in phosphorus or low-mineral-content food. We then need to use mineral supplements or become deficient and prone to diseases. As relates to root vegetables, whole means using the vitamin- and mineral-rich outer parts, usually discarded as peel, as well as the inner parts, and even the cooking water. Pouring out the cooking water pours vitamins and minerals down the drain.

A story, presumably true, has been published about the crew of one of the early submarines. After a long crossing, they were fatigued and disease-prone; they had lived on tinned and dried foods. The health authorities were at a loss about the cause as there had been sufficient amounts of the known vitamins in the foods. A Nature Cure practitioner offered his advice, which, when taken, quickly corrected the problem. He told them to peel a lot of potatoes, discard the potatoes, and cook the peels; strain, discard the peels, and drink the broth. How many of us do it the other way around?

There is a downside to this with some whole foods and seeds. Seeds contain reactive proteins, called lectins, which are mainly in the outer hull or bran; individuals with certain blood types tend to react to specific lectins. Other food allergies develop mainly to the outer parts of seeds removed in refining. If seeds are not sprouted or fermented, then minerals tend to remain bound by chemicals in these outer parts and cannot be absorbed. Therefore, if you have a sensitive digestion and eat food that is not suited for you, then, ironically, refined foods will tend to cause you less immediate problems than whole foods.

Live and Organic Foods: Live food is basically the same as raw or uncooked food. Live means that the food is still high in vitality and naturally occurring enzymes. These are damaged when food is heated above 104° F and destroyed above 120° F. While milk can be heated initially in making yogurt, adding lactic-acid bacteria to it provides a new complement of enzymes to revive the food, although I prefer to make yogurt from unheated milk.

Food has different stages of “aliveness” or vitality. A dry viable seed is alive but dormant, the enzymes inactivated by inhibitors. During soaking and sprouting, the seed awakens to full life with an abundance of enzyme activity. Like an embryo, it is at the height of its vital and restorative functions.

A young, growing leaf or plant still has a high level of growth hormones and enzyme activity, but in a mature plant or fully grown leaf there is little growth hormone and reduced enzyme activity. Whole grain flour that has not been unduly heated during milling has some residual life, but inhibitors block the enzymes; enzyme activity will be enhanced and starch pre-digested when the wheat is made into sourdough. There is a saying in naturopathy: “Eat food that easily spoils, but eat it before it spoils.”

The term organic is generally understood to mean that such food also has not been irradiated and is free of genetically engineered components and any added chemicals that are normally used during growing, storing, or processing. Organically grown food means that it is certifiably free of toxic agricultural chemicals and has been grown in healthy soil without the use of water-soluble fertilizers. Very harmful are nitrogen fertilizers, especially nitrates. In one scientific experiment, hens were fed blueberries that were either unfertilized or fertilized with nitrate. Hens fed large amounts of nitrate-fertilized berries died within three days; those that received only small amounts of fertilized berries laid eggs with defective missing shells. Hens fed unfertilized berries in any quantity maintained normal health and eggs. What will a low-dose long-term intake of such fertilized foods do to us?

It is well known to organic growers that their plants and the animals fed these plants are much more resistant to pests and diseases than chemically fertilized plants and the animals fed them. It has been shown that cattle on an organic farm remained disease-free while cattle on surrounding properties were badly affected with foot-and-mouth disease. Most damaging are pesticide residues on foods or the interaction between different chemicals. Many natural therapists contend that organically grown food is an important part in the successful treatment of degenerative diseases.

Cooking Foods: Food ideally suited for human nutrition does not require cooking, such as sprouted seeds and fresh shoots, ripening seeds (sweet corn, green peas), oily seeds and nuts, sweet root vegetables, and fruits. Cooking damages many proteins; they coagulate and harden and their complement of digestive enzymes is destroyed, and this makes them more difficult to digest.

However, cooking also makes some foods more easily digestible by breaking down the cellulose in plant food and connective tissue in meat. This enables our digestive juices to get more easily to the nutrients and is especially important for people with weak digestive organs. Further advantages of cooking are the preservation of food, easier chewing, and enhanced flavours. All this means that cooked food generally is more convenient and tastes better, and for most of us that is more important than any long-term health concerns.

In this chapter I show how harmful effects of cooking can be minimized by combining the least damaging use of cooking with the more acceptable forms of raw foods consumption.

Steaming is the recommended method of cooking. Add sufficient water to cover the bottom of the pot, even without special steaming equipment. This way most of the vegetables will be steamed, and if you use the remaining water, nothing will be lost. By adding vegetables after the water has started to boil fewer vitamins are destroyed. However, I do not see a problem with simmering vegetables submersed in water as long as you also ingest this water, preferably soon after cooking.

Pressure cookers are acceptable, their main disadvantage being that they are usually made of aluminium. Microwave ovens are suspected of emitting harmful radiation while in use. In addition, they may contribute to health deterioration by scrambling the delicate structures of sensitive food molecules in unnatural ways. Therefore, avoid or minimize microwaved food. It is also safer to discard any outer part of meat or fish after grilling or broiling.

When cooking grains or legumes, always soak the seeds overnight beforehand and discard the soaking water. In this way, you will likely ingest fewer potentially harmful anti-nutrients, such as phytates which make minerals unavailable, protease inhibitors that interfere with the digestion of proteins, or lectins that cause unpleasant symptoms with some blood types. If beans cause wind, change the water once or twice during cooking.

As mentioned, the main disadvantage of all cooking is that it destroys food enzymes and some vitamins. This significantly contributes to more rapid aging and the development of degenerative diseases if the diet consists predominantly of cooked food. Therefore, try to eat only a minimum of cooked food. Most harmed by cooking are animal products and polyunsaturated oils; least harmed are starches. Eat your food, especially vegetables, as soon as possible after a short cooking time; try to minimize steam or fumes escaping from cooking food. Except for baked grains, avoid food leftover from a previous meal.

A large amount of aluminium dissolves in the water during cooking in aluminium cookware; the more acid the cooking water, the more this happens. This is also a problem with stainless steel pots. While steel is more resistant to acids, it can still release harmful amounts of nickel, and nickel is implicated in weakening the immune system and promoting cancer. Enamel cookware is generally safe, except that low-cost Asian imports reportedly may contain high levels of lead; this is also a problem with ceramic pots. Glass cookware is the safest.

Here are some important cooking rules:

- Avoid baking soda in cooking or baking as it destroys B-vitamins
- Add salt or minerals after cooking to preserve vitamins.
- Save and use the cooking water; it is rich in minerals.
- Cook for the shortest time possible.
- Do not keep food warm for extended periods; cool quickly for storage.
- Avoid aluminium and stainless steel cooking utensils; use enamel or glassware instead.
- Do not leave acid food in contact with metal surfaces.
- Do not cook what can conveniently be eaten raw.
- Do not fry or heat oils or fats.
- Do not cook with a microwave.

Chewing and Enjoying the Meal: Good chewing is of vital importance. It is necessary to break down food particles to a fine pulp and at the same time to insalivate them sufficiently for proper enzyme activity. Trace elements and vitamins are already partly absorbed through the mouth tissue, while the stomach needs adequate advance warning of the expected composition and quantity of the food. So, even if the food is already a liquid (soups, broth), chewing is important. Keep and move around such liquid food in the mouth for several seconds. Solid food needs chewing until it is thoroughly liquefied. Teach your children early in their life to chew thoroughly; possibly check their stools to see if they contain large pieces and whole seeds, a sign of inadequate chewing.

You should not only chew thoroughly, but also eat slowly and in a leisurely fashion. This improves digestion and helps prevent overeating. A certain amount of time is needed for the body to form the message that it has had enough. Therefore, if you stuff yourself within a short period, your body has no time to prevent you from overeating. Overeating can also occur if you do not pay attention to the eating itself but to a conversation, to a book, or the television. Then you cannot “hear” when your body says “enough.”

It is also important to have a peaceful mind at mealtime. Before the meal, completely relax for a minute or more with eyes closed; if able to meditate, raise your awareness to a joyful meditative state. During the meal, pay attention to the process of chewing and enjoy the various flavours as they develop.

It is beneficial to enjoy what you eat. Unpalatable food is not well digested. Neither is it good to continue eating poor food just because one has acquired a preference for it. For many, this is the most difficult problem to overcome during health improvement. The goal is to acquire a liking for what is nutritionally good.

There is no doubt that taste buds are adaptable and can be re-educated, but this comes gradually, so introduce unfamiliar foods and supplements to your diet in very small amounts only, interspersed with familiar foods. Be imaginative and experiment with new ways to prepare food. Do not eat if you are not hungry, if you are agitated or emotionally upset, if you are unwell with a fever or comparable condition, or if you are still full from the previous meal. Always prepare food lovingly.

Food Combining: Many years of experience with patients have shown me the importance of correct food combining. The fundamental biochemical considerations with regard to food combining are as follows:

- Typical protein foods (usually containing oil or fat) require strongly acidic gastric juices.
- Different types of protein foods may need differently composed gastric juices for digestion.
- Starches need B vitamins, chlorides, calcium, and magnesium for enzyme activation.
- Sweeteners or sugars as well as acids inhibit the digestion of starches.

Disregarding these facts by mixing foods indiscriminately may contribute to infections and degenerative diseases, especially allergies, arthritis, cancer, and autoimmune diseases.

Initially, most people have a reasonably strong digestive system and do not feel any ill effects, no matter what they eat or combine. This is also true if bodily responses have become dulled by being ignored for many years. However, sooner or later, the unpleasant effects and degenerative symptoms will appear.

There are three main groups of food to consider when working out correct food combinations:

1. Proteins, sweet and acid foods, sweet vegetables, fruits
2. Neutral foods such as fats and oils, non-sweet vegetables, avocados, olives, and cream cheese
3. Starches such as cereals, potatoes, and dry legumes

The neutral foods can be combined with either one of the other groups while the two extreme groups (numbers 1 and 3) should be kept apart. Observing this simple rule may prevent the more serious consequences of incorrect food combining.

In poor health and during a health improvement program, however, it is advisable to be even more careful. With few exceptions, most foods are best digested as a moderated mono-diet, which means only closely related foods are eaten together, for example, different kinds of grains. With severe digestive problems, one can temporarily adopt a strict mono-diet, eating one kind of food only at a meal, and observe its effects.

The next best possibility is to eat foods correctly combined, but still separated during the meal. This is most important with meat, fish, or nuts. These are best eaten at the beginning of the meal for proteins to receive the strong gastric juices in the pit of the stomach. The rest of the meal, if not too liquid, will remain in distinct layers on top of this. Remember this if you have proteins and starches in the same meal.

You can mix a small amount of flavouring vegetables with meat or fish, or eat a fruit together with some nuts. Generally it is preferable to have either some fresh vegetable juice or a small amount of raw vegetables before a cooked meal.

For those with a sluggish metabolism, a small amount of sweet food or fruits as dessert after a protein meal will provide energy for the digestion of slow-burning, protein-fat combinations. This is not advisable after a starch meal or for those with blood sugar problems. Glucose and maltose (as in barley sugar) do not contain fructose and therefore cause less of a problem if combined with starches. If your digestion is sluggish, a small amount of fruit acids, such as lemon juice sprinkled over meat or fish, or ascorbic acid with or after protein meals can be beneficial, as can half a cup of a bitter herb tea afterwards. Cooked subacid fruit that is low in fructose, such as apples, does not react in the same way as sweet fruit that is high in fructose. Therefore, most individuals will be fine using this as a flavouring with cereals or starches, for example, apple puree with rice.

Some factors that make precision food combining less necessary are: a strong digestive system; being physically active; chewing extremely well; and only eating small meals. The opposite factors will, of course, increase the need for correct food combining. Being able to eat just about anything, however, is not a guarantee of strong digestive powers; your system may be too weak to react to the foods. If you are in good health, you can forget about food combining and omit just one group of foods from a mixed meal - the cooked starches; most others combine reasonably well with each other. However, it is generally acceptable to eat starches at the end of a meal, separated from the protein part by a vegetable or sprout salad or by some cooked vegetables. If you are sensitive, then this middle part of the meal can consist of neutral vegetables rather than sweet ones.

Food yeast with a 50 percent protein content and also the non-oily legumes are grouped with the starches because their fat-free proteins are relatively easy to digest and require only slight acidity. For a graphic illustration of the basic food-combining rules, see Figure 4-1. Foods are classified as hard proteins, soft proteins, or starches according to the amount of gastric acid and pepsin that is needed for their digestion. The actual percentage of protein in a food is less important in this respect than the amount of fat, oil, and fibre surrounding the protein.

Meat, nuts, and cooked soybeans, for instance, are difficult to break down. They need the full strength of gastric acid and pepsin and, therefore, are called “hard proteins.” Eggs, uncooked dairy proteins, broth of meat and fish, fermented soy products, soy milk, and fine nut butters, on the other hand, need less acidity and can be called “soft proteins.” Proteins in processed cheese, in cow’s milk used as a drink, but especially in milk added to coffee or tea, are very difficult to digest and thus become hard proteins. Non-fatty fish, especially if marinated, is easy to digest and is almost a soft protein, while fatty fish is closer to meat in its gastric requirements.

Intestinal Fermentation: Combining food high in fibre with sweet food is likely to cause intestinal gas. This perception gave rise to the frequently quoted rule not to combine vegetables with fruit. In this case some of the sugar from the fruit may not be absorbed and then may overstimulate bacterial activity in the large intestine, causing excessive wind, flatulence, or bloating. Some intestinal fermentation is desirable in order to keep our beneficial bacteria happy, and this is best produced with a moderate amount of fibre in the diet. It breaks down at a slow rate and greatly helps to maintain regular soft bowel motion.

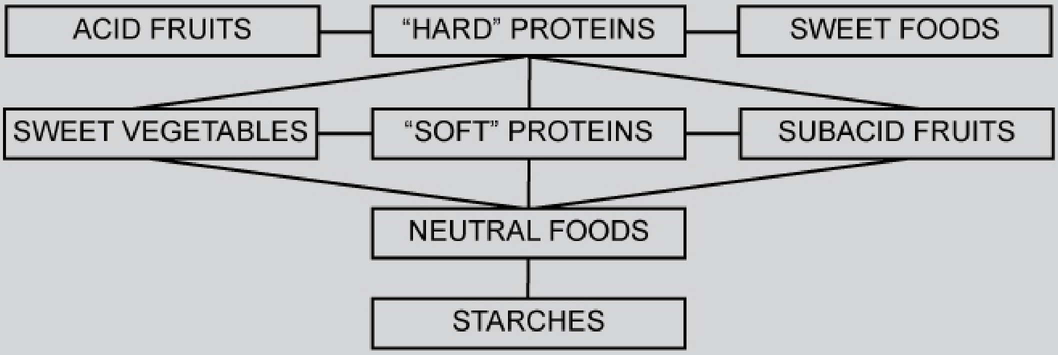

Fig. 4-1: Food Combining Chart.

Direct connections between groups indicate good combinations (e.g. starches & neutral foods). Connections between groups with one interception are fair (e.g. starches & subacid fruits). Connections groups with two interceptions are poor (e.g. starches & sweet foods).

FOOD GROUPS

"Hard" proteins: meat, fish, nuts, oily seeds, soybeans.

"Soft" proteins: eggs, milk proteins, meat/fish broth, tofu, soy milk, nut butters, fermented soy or nuts.

Acid fruit: lemon, grapefruit, pineapple, acid berries.

Sub-acid fruits: apricots, peaches, pears, bananas, cooked apples.

Sweet food: sweeteners, foods with added sugar, molasses, honey, dried fruits, oranges, figs and other sweet fruits and berries, juices of sweet vegetables.

Sweet vegetables: sweet potatoes, pumpkin, onion, carrot, turnip, beetroot, tomatoes, green peas.

Neutral foods: fats and oils, non-sweet vegetables (mainly greens), sprouted seeds, avocados, olives.

Starches: grain products, potatoes, dry legumes, food yeast, sago and taro.

Trouble brews if we ingest a large amount of sugar in any form. This is like adding gasoline to a slow, controlled burn-off of bush land that then gets out of control and becomes a raging wildfire. It is very similar in our intestines. Fibre remains in the large intestine for a long time, so excessive fermentation can happen if we have sweet food several hours after a raw salad. How much sweetness we can tolerate with fibre is individualized. The more efficiently we absorb our food, the larger the amount of sweetness we can handle without getting into digestive trouble.

For individuals with a weak sugar metabolism, it is not good to have much sweetness on an empty stomach, as that drives the blood sugar level too high, followed by symptoms of hypoglycaemia. In this case it is best to completely separate food high in fibre from sweet food and have the fibre meal after the sweet meal.

Therefore, breakfast is safest as a moderately sweet meal, for instance, by flavouring cooked rice or sago with banana or apple puree. The evening meal is then safe as a fibre meal, such as combining potatoes and other cooked non-sweet vegetables with a vegetable salad. Lunch can be a protein food with sweet vegetables such as pumpkin.

It may also be acceptable to have a piece of fruit or a fresh juice of sweet vegetables (e.g., carrots) some time before a meal that also has a vegetable salad or other high-fibre food at the end. Fresh juice does not have much indigestible fibre, so there should be no fermentation problem in combining vegetable and fruit in the same juice. However, if fruit high in fructose is combined with vegetables, then the juice is high in glucose as well as fructose and can cause a strong insulin response and hypoglycaemia in susceptible individuals.

The main rule in all this is to have the fast-digesting food first and the slow-digesting fibre afterwards. Experiment to see how strictly you have to apply this rule to avoid intestinal distress.

When eating cooked beans, you may encounter a similar fermentation problem. Beans do not need any added sugar to start a strong fermentation. They are high in long chains of carbohydrates (oligosaccharides) that are not absorbed in the small intestine, but are easily broken down by the bacteria in the large intestine. To avoid problems with beans, these can be sprouted or fermented before eating or the water in which they have been cooked can be renewed and discarded several times.

Planning Meals: Our digestive powers are strongest in the morning and weakest in the evening. Therefore, it is advisable to eat difficult foods such as meat, nuts, sweet food, and large vegetable salads during the daytime and have a light, early evening meal. This will also improve your sleep, while a heavy, late dinner often results in a restless or “drugged” sleep, with a feeling of tiredness in the morning.

If you are not normally hungry in the morning, just omit the evening meal for several nights and have 30 minutes of vigorous physical activity before breakfast. If you do not need to leave home for work, you can also spend an hour in the garden or engage in other outdoor activity before breakfast. Together with the time required to drink your juice, herb tea, or water, it will be one to two hours before you finally sit down for breakfast. It is definitely not advisable, especially with a sluggish digestion, to eat a heavy breakfast without sufficient stimulation to your circulation beforehand.

With a sluggish thyroid and weak adrenal glands, of course, it is more difficult to get going in the morning, but it is not too hard if you are determined to make the effort. Immediately after awakening, start with some positive autosuggestion. Then mentally look forward to something that you would like to do during the day. If necessary, promise yourself a treat. Lack of enthusiasm can be a main factor in suppressing your metabolic rate. Therefore, try to become enthusiastic about something - anything!

Do some deep breathing, stretching, and muscle-tensing while still in bed. Then jump out and immediately start some lively activity such as circling the arms or gently pummelling the body all over. Also stimulating are arousing music, skin brushing with rapid strokes, and cold-water applications.

Do not rush your breakfast. For many people, it is the most important meal of the day. Make it a habit to rise sufficiently early so that you have at least two hours before needing to leave the house. You can also divide the food and use a portion of your breakfast as a mid-morning snack.

A certain amount of food eaten in the morning can produce weight loss, while the same food in the evening is more likely to convert to body fat and cause weight gain. Therefore, if you want to keep your weight down, eat all your high-calorie food during the day and have only some non-starchy vegetables or a few apples in the evening. If you want to gain weight, on the other hand, then have a starch meal in the evening. There will be some, however, for whom these recommendations are not suited; you must find out under which conditions you feel best and can cope most efficiently with your daily activities.

Carbohydrates cause a greater weight gain than a protein-fat diet of equal calories. Also, the higher the amount of roughage (for example, rice bran, linseed meal, raw vegetables), the greater is the weight loss on any slimming diet. These factors should be taken into account in your meal planning.

It is desirable to have about 30 minutes of rest or light, relaxing activity after a meal before starting more heavy manual work or a more stressful activity. I agree with the advice of Edgar Cayce: "After breakfast work a while, after lunch rest some time, and after dinner walk a mile." Try to fit in at least 15 minutes of rest in the horizontal position after lunch; relax, or try to meditate or contemplate. If you are in poor health, a sleep after lunch is highly desirable. If you wake up tired from this, next time give yourself positive autosuggestions before falling asleep. Even after breakfast, you should not immediately start rushing around. Some kind of steady physical activity, such as walking or gardening, is advisable after the meal. But what is harmful is stress - the feeling of being late for school or at the office - so avoid that.

Food Balancing: The concept of food balancing arises from the fact that the various food groups have different effects on our metabolism and hormonal and nervous systems, as well as on our emotions. Food balancing is different from what is described as eating balanced meals - proteins, carbohydrates, and fats - as recommended by conventional dieticians. To give some examples: Meat stimulates the adrenal glands and the sympathetic nervous system, thus increasing tension and readiness for action. Fruits have the opposite effect. They stimulate the release of insulin, which counteracts the effect of adrenal hormones. While vegetables, especially those high in chlorophyll, make us more relaxed, fruits make us physically, emotionally, and psychically more sensitive.

Between meat on one end of the scale and fruits on the other, we can arrange our basic foods in the following order according to their sensitizing effect on body and emotions:

red meat
fowl and fish
eggs & dairy products
oily seeds
legumes
grains
green vegetables
sweet vegetables
sub-acid fruit
berries, sweet & acid fruit

Both fruits and vegetables increase our intuition. Meat, on the other hand, makes us more insensitive and possibly materialistic, and less emotional and intuitive. Red meat especially has a strong "grounding" effect and is therefore balancing for those who are oversensitive, overemotional, and open to negative psychic or occult influences. It is, however, unsuitable for those who are already too tense or insensitive either physically or emotionally or who are aggressive or overly materialistic.

Acid and acid-sweet fruits have the strongest sensitizing effect, while subacid fruits are less sensitizing and avocados are neutral. In this way, lemons and grapefruit are more sensitizing than oranges, while oranges sensitize more strongly than do mandarins.

Flesh foods are more suitable for those who cannot efficiently use glucose, such as fast oxidizers or hypoglycemics, while fruit and sweet vegetables are best for those with a slow metabolism. This means that those with an unbalanced metabolism generally should not eat a conventionally balanced diet; they become balanced by selecting foods that are either more slowly or more rapidly digested. However, it is not necessary for sensitive individuals to eat flesh foods to be in balance. They can do very well on oily seeds, legumes, grains, green vegetables, and raw sweet vegetables, but less well on cooked sweet vegetables.

Grains and legumes are neutral in their effects on our metabolism and the glandular and nervous systems. In this respect, they may be grouped in the middle between meat and fruits. They are balanced in themselves, and this shows why one can be balanced on a macrobiotic rice diet with few additions of other foods.

Someone with a normal metabolism can eat in a balanced way by having mainly the neutral grains and legumes, and balancing eggs and dairy products with vegetables. However, it is also possible to balance meat, fowl, and fish with sweet vegetables and fruits, eating very little of the in-between foods. Another possibility is to eat a greater quantity of a food group that is closer to the neutral position and replace a smaller amount of a more unbalanced food. For instance, instead of balancing meat with fruits, you can use an increased amount of sweet vegetables.

However, you do not need to worry about accurately balancing and measuring quantities. This is not an exact science but more a rule of thumb. Its main importance is to show you in which direction to move in your food selection if you feel that your health or emotions are out of balance.

* * * * * * * * * * * *

Step 33
FOOD GROUPS

Learn the health characteristics of the various food groups and apply them for healing.

Let’s start with carbohydrates. There are three groups of carbohydrates: starches, dextrines, and sugars. Sugars consist either of simple sugar molecules, the monosaccharides, or two of these linked together as disaccharides. Household sugar (sucrose) consists of one molecule each of glucose and fructose, while milk sugar or lactose contains glucose and galactose. In dextrines, up to 50 glucose molecules are linked, and in starches, several thousand are.

Starches: Starches are broken down into water-soluble dextrines by cooking, sprouting, or fermenting. Dextrines occur naturally in sweet vegetables, and in seeds during germination and ripening; examples are green peas and sweet corn. The basic starch foods are cereal grains, potatoes, sago, taro, tapioca, and to some degree the non-oily legumes. On the one hand, starches are excellent slow-digesting food, but on the other hand, our digestive systems have not yet well adapted to a diet high in grains, especially in those people with blood group O.

In our society, wheat is predominantly used. However, gluten, the protein in wheat, often acts like sandpaper on the absorption villi in the small intestine, and this is a main cause of malabsorption, intestinal inflammation, and allergies. All of us are to some extent affected by gluten and in this context wholemeal products are no better and usually worse than refined flour. Therefore, it is generally recommended that you use only a minimum of wheat products and that you also be careful with the other gluten grains, mainly rye, oats, and barley. Oats are very high in gluten and in this way not much better than wheat. Rye crispbread and rye sourdough bread are usually better tolerated than wheat products.

Rice and millet can be used as staple grains. Maize and potatoes are valuable additions, especially for those allergic to wheat. Buckwheat is generally good but causes problems in some gluten-allergic people. Potatoes should not be peeled, or the skin should be removed only after cooking. Small potatoes and the mineral-rich outer parts of potatoes are preferable to the acid-forming inner parts.

Legumes are best sprouted, as cooked dried beans often cause wind. This can be reduced if you discard the soaking water and possibly replace the cooking water after 20 to 30 minutes of cooking. Sago and tapioca are especially suited for individuals who are not doing well on grains. They are commonly available as white starch granules, in some specialty food sections of supermarkets.

Sweet Food: The term “sweet food,” as used in this book, includes sugars, sweetened food, dried fruit, sweet fresh fruit, and the juices of sweet vegetables. It does not include the sweet vegetables themselves such as carrot, onion, sweet pepper, turnip, or red beet, as their sweetness comes mainly from dextrines and they release sugars slowly. Most commonly used sugars quickly enter the bloodstream and cause a serious strain on blood sugar regulation. It is important for our well-being to keep the daily fluctuations in our blood sugar level as small as possible. For people with a poor sugar metabolism, this means restricting the intake of sweet foods to the barest minimum.

A poor blood sugar metabolism is not restricted to an inappropriate rise in the blood sugar level. An equally dangerous component is a high insulin level, especially after ingesting sucrose and, to a somewhat lesser extent, when combining glucose with fructose. A high insulin level either leads to a strong fall in the blood sugar level sometime later (hypoglycaemia) or to conversion of sugar into fat, which leads to overacidity or overweight.

A third outcome is increasing insensitivity to insulin and the development of Type II diabetes. However, maltose or barley sugar, made from sprouted barley, should have a less harmful effect if used to sweeten starches. It consists of two joined glucose molecules and does not induce a strong insulin response, but it could still cause a stronger rise in the blood sugar level than unsweetened starches. Some of the symptoms associated with a poor sugar metabolism are overweight and underweight, high or low blood pressure, heart disease, arthritis, diabetes and hypoglycaemia, dental caries, colds, allergies, weak eyes, cold hands and feet, lack of energy, oversensitivity, and overacidity.

For most individuals, sweet food is harmful because sugar is released into the system too quickly. Therefore, a sweet dessert or sweetener eaten with protein or fatty food is less harmful than sweet food such as fruit juice taken on an empty stomach or as sweetened starches (for example: sweet porridge, bread with honey, cake). Most affected are individuals with low blood pressure, sensitivity to cold and skin irritants, or those who are emotionally unstable; sugar-insensitive individuals usually benefit from naturally sweet foods. I regard artificial sweeteners such as aspartame as much worse than any natural sweeteners and suggest you avoid them completely.

One natural sugar that appears to be relatively harmless and possibly even beneficial if used in small amounts is xylitol. It occurs naturally in fruits such as plums and strawberries and tastes like normal sugar; it also helps to prevent tooth decay and possibly ear infections and is acceptable for diabetics. However, it may cause diarrhea in larger quantities, so you need to experiment to see how much is acceptable for you.

The Glycemic Index (G.I.): The G.I. lists foods according to the increase in blood glucose levels caused by their carbohydrate content two to three hours after eating. A high G.I. means a rapid increase in blood glucose levels and vice versa. It is preferable to eat mainly low G.I. foods that produce a small or slow rise in blood sugar. Here are a few key examples, with white bread as the standard with a G.I. of 100: glucose and maltodextrin, 137; sucrose (common or household sugar) and rye bread, 92; honey, 83; and fructose, 32. Legumes and nuts generally have a low G.I. From this short list, you can easily see why I have no faith in the G.I.: rye bread and sugar have the same G.I.! The reason for this is the low G.I. of fructose.

Sucrose consists of one molecule of glucose and one of fructose. Fructose produces a strong insulin response that keeps the blood sugar level low partly by converting sugar into fat and partly by channelling glucose rapidly into muscle cells. Depending on the individual metabolism, the combination of fructose and glucose in large amounts either produces overweight or overacidity. In addition, it increasingly leads to insulin resistance and Type II diabetes.

Nevertheless, conventional nutritionists argue that sugar is not harmful because its G.I. is as low as or even lower than bread. I would not be surprised if it eventually turned out that the G.I. was developed on the instigation of the sugar industry. Most foods listed in the G.I. are processed carbohydrates that I do not advise you to eat anyway. If you select your carbohydrates according to the following rules, then you do not need to be concerned with the G.I.:

- Eat mainly legumes and vegetables
- Eat (fresh) fruits on their own and not with or after meals
- Minimize sweetened food, grains, and cereals

Proteins: Proteins from animal sources are oversupplied in traditional Western diets. Initially, this causes stimulation—you feel energetic and even aggressive—but later in life it leads to enzyme exhaustion, putrefaction of bowel contents, toxaemia, breakdown of the immune system, gout, and cancer. These problems are not due to proteins per se, but rather to the cooking of proteins, which destroys their enzymes and denatures their structure. Protein deficiency, on the other hand, causes slow growth, fatigue, and debility.

Good sources of protein are grass juice (fresh or dried), pollen, spirulina, chlorella, almonds, lentils and other legumes, traditionally fermented soy products, sunflower and sesame seeds (tahini), raw egg yolk, naturally fermented and unpasteurized goat cheese and yogurt, fish, and (organic) liver and other organ meats. The less these are heated, the less digestive problems they will cause.

Use red meat from organic or grass-fed sources and avoid predatory species of fish. Nuts are generally difficult to digest, except if soaked or sprouted. Soybeans and broad beans should not be eaten raw except if sprouted or fermented. Cooked soybeans can cause indigestion, except if you discard the soaking and cooking water; processed soy foods are not recommended because of their high content of anti-nutrients and the likelihood of genetic engineering in the beans themselves, although traditionally fermented soy foods are fine.

Individual protein requirements differ, and more is needed during pregnancy, in childhood, and during convalescence. With advancing age and with degenerative diseases, use any flesh food in an easily digestible form, such as broth of fish, liver, or meat. It is good to combine legumes with starches, as, for example, lentils with brown rice. However, mixing different proteins to obtain improved amino acid composition is generally not required. Preferably, have your flesh foods no more than once a day and avoid products from feedlots or that contain growth promoters.

Vegetarianism: Most individuals can remain healthy on a diet with moderate amounts of flesh food or as vegetarians without eating any flesh food. However, a strict vegan diet without any animal products is advisable only if there are no obscure health problems, because some people require certain nutrients to be supplied from animal sources (such as taurine, carnitine, or vitamin B12).

Generally, sensitive people have weak adrenal glands and feel more energetic and emotionally balanced by using flesh foods. If these foods are excluded, such people should have a high intake of legumes and complex carbohydrates (slow-digesting food) in their diet. People with alkaline and insensitive conditions, on the other hand, benefit from a partially vegetarian diet, possibly with some seafood and poultry. A fruitarian diet, based on tree fruits, nuts, and berries, is suitable as a temporary or extended cleansing diet for alkaline, insensitive individuals.

Many vegetarian diets are not conducive to good health because of their high intake of sweet foods, milk products, wheat, oats, and inhibitors of protein-digesting enzymes in a soy-based diet. Further, minerals are better absorbed if the meal includes flesh food or gelatine. A diet high in nuts, oily seeds, and cereals can cause a deficiency in lysine (an essential amino acid) and predispose you to herpes and other viral infections; this usually can be rectified by using more legumes. Individuals with blood group O are basically meat-eater types and have great difficulty with a grain-based diet. However, on balance, vegetarians tend to be somewhat healthier than meat eaters.

Fats and Oils: I recommend using foods with a wide range of fats and oils, provided that most of them are unheated. Oils are best supplied by eating products in which they naturally occur. Most recommended is extra-virgin olive oil; otherwise use cold-pressed oil stored in brown glass bottles. Avoid “lite” oils because they are more highly refined and also oils sold in plastic bottles, as oils can leach chemicals out of plastic. Store oils in a cool, dark place and in full containers; refrigerate oil in daily use, except varieties that easily solidify.

Preferably, use only small amounts of heated saturated fats that are hard at room temperature, such as fat in hard cheese, beef, and mutton. Unpasteurized butter, like other unheated fats, is a healthy food. Avoid all chemically hardened (hydrogenated) fats, including margarine. Use mainly oils high in oleic acid, such as olive oil, almond oil, and, to a lesser extent, peanut oil. Peanut oil helps against arthritis but is not suitable for non-sensitive, alkaline people in whom it can contribute to the formation of fatty deposits in the arteries. Unheated coconut oil is beneficial for most individuals, despite being saturated and hard at room temperature. Its high content of short-chain fatty acids enables it to be easily metabolized to energy without causing undue weight gain.

Use bottled polyunsaturated oils (high in linoleic acid) sparingly; it is preferable to obtain these oils directly from ingesting oily seeds. Such oils easily oxidize if the vitamin E intake is low and can contribute to the development of cancer. However, it is recommended to use more fish oils or linolenic acid, as in linseed (flaxseed), together with adequate vitamin E.

People who are insensitive and alkaline, with cardiovascular disease, obesity, or skin and liver diseases, should be especially careful and avoid saturated and hydrogenated fats. Lecithin and vitamin A and E supplements will be especially good for them and cod-liver oil rubs are beneficial if you suspect fat malabsorption.

Fruits and Vegetables: While vegetables are good for everyone and the more the better (except if sprayed with pesticides), fruits must be treated with caution by those with overacidity and poor blood sugar regulation. Use mainly fruits and vegetables in season that are grown in your region. Include plenty of green leafy vegetables and gradually increase your intake of fresh raw vegetables as in salads.

Cook or grate root vegetables with their skins left on and use the cooking water. Among the root vegetables, red beets are highly recommended for their positive effect on cell respiration and energy production. Fruits, especially acid fruits and berries, are excellent for those with alkaline and insensitive body conditions, including cardiovascular diseases, cancer, and diabetes.

On the other hand, those who are overacid (usually with low blood pressure, allergy problems, lack of energy, and tendency to colds) easily become even more acidic on fruit. They should minimize their fruit intake and mainly use subacid varieties or, better still, neutralize acid fruit (as explained in Step 6 under “Alkalizing with Calcium and Magnesium”). Also they may better tolerate fruits that are not fully ripe; oily fruit such as avocados are usually well tolerated.

* * * * * * * * * * * *

Step 34

HEALTH DIETS

Adopt a health diet suited to the specific requirements of your body.

Diets for health improvement are different from slimming diets that do not always improve the health of the dieter. However, on a health diet, weight loss can be expected if you are overweight, in addition to an improvement in health. As a general rule, the more you follow a diet that is ideally suited for you, the better will be your health. The diets recommended in this chapter have been shown to be effective, but you may need to experiment to find individual variations that best suit your constitution and health problems. Your long-term diet together with your mental attitude and lifestyle will decide your health in future years; therefore, choose wisely.

If you are on a conventional diet at present, you can change gradually to the high-quality diet (see below), a process that could take years, especially if you have to take along a reluctant family. Others can rapidly adopt a raw food diet, especially if confronted with an advanced degenerative disease. Nearly anyone can live temporarily on raw food as a cleansing measure.

If you are unusually sensitive to raw food diets and develop digestive problems, you can eat cooked brown rice exclusively for several days or weeks as a mild cleansing diet. If even brown rice causes problems, try basmati or white rice. Use only olive oil and possibly a small amount of salt, herbs, spices, or vegetables for flavouring. Strictly observe how you feel as you reintroduce other foods.

You need not be afraid that you can never eat your favourite foods again. The long-term impact on our health comes from what we eat daily, not from occasional transgressions. When you eat out, you can eat what is available; if it does not agree with you, you will know better the next time. Often our body becomes temporarily hypersensitive to foods that we have habitually eaten before. Then our reflexes were dulled, but now the body lets us know what is good or bad for us. Try to listen to it and become friends with your body.

The High-Quality Diet: This is a low-allergy, semi-vegetarian diet based on freshly pressed vegetable juices, sprouted seeds, fermented and purple food, legumes, non-gluten grains, and a high intake of raw food. In this diet, cooked food is used more as a flavouring or to put on weight; the main part of a typical meal is a vegetable salad mixed with sprouted seeds. Use mainly grated, coloured root vegetables. If you cannot chew well, grate them finely or puree them. Examples of suitable roots are red beet, carrot, turnip, sweet potato, and pumpkin; tomato, cucumber, and celery can be added as flavouring.

As a salad dressing, you can use olive oil, lemon juice, herbs, and spices. For example, shake equal parts of olive oil and lemon juice in a jar; then add one teaspoon of lecithin, a pinch of cayenne, and some marjoram or mixed herbs. You can also blend in some unsprayed lemon or orange skin or a whole lemon. Keep refrigerated and shake before using. A small amount of cider vinegar can be used if lemon juice is not available.

Use only a minimum of commercially grown leafy vegetables and cabbage as they may have a high exposure to toxic sprays. Instead of greens, you can use sunflower and buckwheat lettuce grown in seed boxes; you might also try a small amount of very young wheat grass cut finely into a salad. If you have chewing difficulties or vegetable salads do not agree with you, have plenty of fresh vegetable juices produced with a non-centrifugal juicer.

Before cooking legumes or whole grains, soak them overnight or even longer to start the germination process. Then the enzyme phytase can start converting phytates into inositol and liberate minerals that are otherwise not available when you cook these without prior soaking. Soaking and vigorously rinsing the seeds before cooking also reduce levels of any pesticide residues and mycotoxins from fungal contamination. Use mainly non-gluten grains, such as rice, millet, and buckwheat, also sago and tapioca. Corn is fine if you've passed allergy testing for it. For individuals with blood group O, I recommend that you use mainly vegetables rather than grains as the cooked part of the diet.

Presoaked legumes need cooking only for a few minutes as they are already soft; the exception is soybeans, which are better if cooked for two hours. If cooked legumes cause wind, next time replace the water used for soaking as well as the cooking water. The cooking water for soybeans can be changed several times during cooking to avoid or reduce wind.

Lentils are recommended as a cooked addition to a meal and can be combined with rice. For sugar-sensitive individuals, it is good to have legumes with breakfast to slow the absorption of glucose from the same and even from the following meal. Fenugreek is even more effective in this regard; preferably, also use the water in which it has been soaked. It is not necessary to cook legumes in order to slow the absorption of glucose, and even more effective are partly sprouted legumes.

A breakfast based on ground linseed, yogurt, and bananas is suitable for most individuals; for details, see “Linseed Yogurt” under the recipes in step 37. For individuals with blood sugar problems it is preferable to eat sweet vegetables (e.g., carrot, pumpkin, turnip, and sweet potato) raw instead of cooked. If you cook them, do so cautiously, together with legumes or flesh food.

Flesh foods should be eaten with restriction. The best flesh foods are (organic) liver and deep-sea fish of non-predatory species or seafood from uncontaminated coastal regions. Farmed fish commonly have high levels of pollution and low levels of beneficial omega-3 fatty acids, so are not generally recommended. (For fish species with acceptable mercury levels, check: www.ewg.org). Chicken or other fowl should preferably be organically raised or otherwise free of chemicals.

Eggs in your diet should be genuinely free-range or organic and not from chickens fed any commercial pellets. Fresh, raw egg yolk is a health food that improves vitality and immune function; it can be used as part of a salad dressing. It does not increase but helps prevent the risk of heart disease. Alternatively, eat eggs soft-boiled or coddled. In view of the high incidence of egg allergy, if you are a regular egg eater, I recommend that you test yourself for egg allergy once a year after a period of one or more weeks of abstention from eggs.

The best milk products are yogurt, cottage cheese, and cheese made from goat's milk. Raw goat's milk can also be used as sour milk or clabber milk. If you are not sensitive to cow's milk, these products can be acceptable if made from fresh, raw, organic cow's milk. However, avoid milk products if you have catarrh or other mucus problems.

Almonds are excellent, but nuts in general are more difficult to digest than other foods and should be used with caution. All oily seeds are easier to digest if soaked overnight; peanuts should be vigorously washed and rinsed to reduce any fungal contamination. Nuts and other oily seeds can be used as pastes or nut butters; tahini or sesame seed paste is good, and tahini can be diluted with olive oil for food flavouring. Unfortunately, the seeds are commonly roasted before grinding, and this also applies to commercial peanut butter. Highly heating fats or oils, as is done in frying or roasting, tends to oxidize oils, making them rancid, and this contributes to clogging of arteries and so-called free radical damage. However, sesame oil is more resistant to oxidation damage than other oils.

I regard extra virgin olive oil as the best oil or fat for everyday use. Sesame and almond oils are also good, but generally expensive. Select oils with a dark color and strong flavour. "Fine, lite" oil has usually been heated during refining and lacks valuable cofactors; even cold-pressed oils may have been heated during refining, the more so the lighter they are. High-quality linseed oil is beneficial for many conditions, but as it needs extra care to keep it from getting rancid, you can store it in the freezer.

Coconut oil, on the other hand, is least prone to rancidity and appears to provide good energy and body heat without causing undue weight gain. If you have access to fresh coconuts, you may find that ingesting the flesh often causes stinking wind, with or without additional indigestion. Pieces of coconut flesh are difficult to digest and are then fermented in the large intestines; this releases smelly short-chain fatty acids. You can avoid this by blending the coconut flesh for a minute or two with a cupful of hot water and then pressing it through a strainer. Refrigerate the liquid and residue and add them in small to moderate amounts to meals, use with other foods, or use on their own.

Fruits need to be used with care. Conventionally grown fruits and tomatoes are exposed to more toxic sprays than most other foods, and their high sugar and acid contents often cause digestive problems for sensitive individuals. Minimize the use of sprayed fruit that is eaten with its skin, such as berries. You can reduce spray residues somewhat by soaking fruit in warm soapy water and by scrubbing hard-skinned fruit such as apples; afterwards, rinse well. Fruit such as apples or pears can also be peeled. Note that the core of sprayed apples and pears tends to have a higher pesticide content than the flesh.

Individuals with sensitive skin or low blood pressure are better off using only a minimum of fruit or, alternatively, neutralizing any fruit acids ingested. Some types of fruit, such as bananas, papayas, or pears, are best eaten when they just start to ripen and are less sweet. Overripe fruit is usually not well tolerated, though apples generally cause less problems and avocados are usually fine.

Individuals with insensitive skin and raised blood pressure, on the other hand, do well with fruit and may often have fruit alone instead of a regular meal, especially for breakfast. Most beneficial for them are acid fruits, such as acid berries, grapefruit, acidic oranges, and pineapples.

Individuals in a very weak or debilitated condition may not be able to eat raw vegetables. In such instances, use mainly cooked rice or other non-gluten grains with steamed vegetables and as much freshly pressed vegetable juice as can comfortably be consumed. This diet can also be used for those who temporarily become oversensitive to a wide range of food items or who want to maintain a stable basic diet while testing for food allergies or other negative reactions. If brown rice causes reactions, then try white rice or sago.

Restricted Foods: Restrict foods that have a negative influence on your health and especially if you are already in poor health. You must decide for yourself how strict you want to be and how much willpower you can muster to remain on the high-quality diet.

If you have cravings for a restricted food, it is better to give yourself a reward from time to time and eat some of it at set and well-spaced occasions rather than to snack on it daily or give up the whole diet because it is too difficult. Generally, the more severe your condition, the more strict you should try to be. Some restricted food may be harmless or even beneficial after you have sufficiently recovered.

Intake of the following foods should be initially avoided and later minimized:

- Wheat products and gluten-enriched items, commercial breakfast cereals
- Cow’s milk products (except butter and non-pasteurized cheese) and foods with added lactose
- Margarine, shortenings, and polyunsaturated oils, especially canola oil
- Sugar and other sweeteners; all artificial sweeteners, as in “diet foods”
- Pickled and smoked food and generally highly processed food with added ‘ chemicals
- Food that has been irradiated or genetically engineered
- Food cooked in a microwave oven or otherwise heated to higher than boiling temperature

However, even in this strongly restricted group, items can vary greatly in their effect on your health. Wheat grass does not in this sense count as a wheat product, and wheat-germ oil is generally beneficial except to those who are allergic to wheat. Wheat germ itself has some positive aspects, but its disadvantages are its high gluten content, strong enzyme inhibition, and the frequent rancidity of commercial products stored without refrigeration.

Sprouted wheat is a more beneficial product than cooked whole wheat, and flour products are less beneficial than whole wheat. Whole-meal sourdough bread can be beneficial for those not allergic to wheat, while white bread and cakes are devoid of vitamins and minerals and commonly contain added chemicals. However, whole grain products sometimes can cause sensitivity reactions, while the corresponding refined foods may be tolerated.

The least harmful cow's milk products are sour cream, cottage cheese, and cheese made by natural fermentation; clabber milk or sour milk made from fresh raw milk; acidophilus milk and unflavoured yogurt made from whole milk only; and butter, especially if made from sour cream. These should all remain unheated after production and before ingestion and be made from organic or otherwise unchemicalised milk.

Some of these products have various beneficial aspects while their negative aspects include the tendency to create cow's milk allergy; their contents of lactose, casein, and saturated fat are also problematic. However, even butter is better than processed cheese, pasteurized milk, and the many low-fat products with their high content of mucus-forming lactose. The most harmful milk products are generally skim-milk powder and long-life milk. Another potentially harmful aspect of fatty milk products is their high estrogen content, which can act as a growth promoter for some female cancers.

In regard to fat, butter is preferable to hydrogenated and chemicalised margarine and vegetable shortening. Avoid canola oil as it may have potentially harmful ingredients. Cold-pressed oils are preferable to conventional polyunsaturated oils, and those in plastic bottles should be avoided. Sensitive oils such as food-grade linseed oil should be extracted under nitrogen and kept in cold storage without exposure to light, for otherwise they are likely to be harmful. The only healthy way of obtaining polyunsaturated oils is by eating oil-containing seeds.

The least harmful sweeteners are barley sugar or maltose, raw honey, and soaked and ground dates. In terms of health there is not much difference between white, brown, or raw sugar, and even honey can cause health problems. Therefore, use sweeteners only on rare, special occasions and in small amounts. Avoid artificial sweeteners, especially aspartame (Nutrasweet).

Cautionary Foods

In general the following foods should be used with caution:

- non-wheat gluten grains (barley, oats & rye)
- peanut products
- yeast products
- flesh foods
- salted food
- frozen vegetables
- dried fruit & fruit juices
- soy products, except if they have been traditionally fermented.

With children or adults with cravings for sweets, a compromise agreement can be reached to abstain from sweets on all but one day of the week, when they can eat as many sweets as they like. Then they can gradually be steered away from commercial sweets to eat unsulfured dried fruit, and later fresh fruit.

The only acceptable breakfast cereals, from a health viewpoint, are those without added sugar and preferably low in gluten. Bran can have some benefit as a laxative but it reduces the absorption of minerals and I do not recommend it. Ground linseed (flaxseed) is preferable.

Pickled and smoked foods contain protein breakdown products that can trigger migraines, allergies, and neurological problems. Smoked food also contains carcinogens. Read the labels and avoid food with added chemicals unless these are vitamins, minerals, or carotenes. Beware, for some products, such as bread, are not required to have all chemical ingredients listed. Canned food is unhealthy because it is precooked and may also contain lead from the soldering and other metals from the container, especially if the contents are slightly acidic.

Excitotoxins are taste or flavour enhancers that release glutamate and other brain-active amino acids such as aspartate and cysteine. The best known example is MSG (monosodium glutamate). High blood levels can cross the normally protective blood-brain barrier and can cause brain cells to die. Excitatory amino acids cause problems, mainly when they are used either in high concentrations or in free form. In most natural foods, they are slowly released and therefore harmless. Most processed foods contain excitotoxins, notably any kind of commercial taste or flavour enhancers, such as caseinate, hydrolyzed vegetable protein, soy protein extract, yeast extract, or beef stock. These may be labelled as natural flavouring and are especially prevalent in soups, sauces, and gravies.

There are many anti-nutrients in soy products, especially inhibitors of protein-digesting enzymes, phytates, and substances that agglutinate or clump the blood. Commercial soy milk either has been highly heated to produce soy-protein extract with denatured proteins or it is made of whole soybeans and then is full of these anti-nutrients. Increasingly, the soybeans may also have been genetically modified, with unknown consequences.

The non-wheat gluten grains (especially oats with a high gluten content) can cause problems to those allergic to wheat or gluten. Rye sourdough bread is generally good, however, but many varieties of rye bread have added wheat disguised as bread-making flour. Yeasted products, such as bread, wine, and beer, can aggravate fungal problems, such as Candida infestations, which are frequently chronic with degenerative diseases. Yeast-baked bread has a lower availability of minerals than sourdough bread, while modern fast-baking processes are even worse and keep the minerals tightly bound to phytic acid and unavailable for absorption.

Peanuts contain a long-chain fatty acid that is beneficial for arthritis but can contribute to the development of atherosclerosis. Poorly stored peanuts can develop a toxic mould that may also infest peanut butter. Any such mould is generally eliminated by removing the skin and thoroughly washing or soaking the peanuts before eating. Roasted peanuts contain harmful oxidized oil, which is also present in peanut butter; therefore minimize your use. Peanut butter without oil on top is worse and usually hydrogenated - avoid it. Some health stores can make a healthy peanut butter from raw peanuts.

Flesh foods should be used sparingly. Eating meat requires that an individual have strong digestive powers, good muscle activity to properly metabolize it, healthy kidneys to discharge the high level of residues, and good bowel activity to avoid harmful putrefaction. It is often difficult to obtain chemical-free meat, depending on where you live, and it is not good to eat meat from unhealthy animals. Animals raised in feedlots and battery hens are in this category.

Processed meat is even more chemically contaminated and best avoided as is meat with a deep red color due to added nitrates and nitrites. However, you may be able to remove most of the nitrates and nitrites by soaking the meat, including diced or minced meat, in warm water for a short time and then discarding the water. Individuals with blood group O are best suited to include some meat in their diet; those with blood group A are often better on a vegetarian diet.

Predatory fish such as tuna and shellfish are often high in accumulated pesticides and mercury and are not recommended for habitual eating. While sardines are safer, they have the disadvantage of being tinned. The best choice is fresh or frozen non-predatory deep-sea fish or other seafood from unpolluted coastal areas. For more advice on pollutants in fish of different species and from various locations, see the website of the Environmental Working Group: www.ewg.org.

For most degenerative diseases and especially cancer and raised blood pressure, a high-potassium and low-sodium diet is required; therefore, food should not be salted. The absence of iodized salt in the diet may cause thyroid problems, which can be avoided by using additional kelp. However, in some conditions, when the adrenal glands are near exhaustion, such as with cystic fibrosis, glaucoma, and low blood pressure, additional salt is advisable; this is also true for individuals with dehydration, diarrhea, or excessive sweating.

By far the best source of salt is seawater. You can evaporate seawater in flat trays exposed to sunlight to make sea salt. This typically has about ten times more magnesium and trace minerals than commercial sea salt. Commercial unrefined sea salt is salt that has been crystallized in fractions, and it contains very little of the trace minerals originally in seawater. Highest in magnesium and trace minerals is the brine out of which the sea salt has been crystallized. It is sometimes available in liquid form as “sea-minerals” and is highly recommended as a mineral supplement. However, magnesium chloride, commonly derived from seawater, can be used instead.

Commercially frozen vegetables are precooked and contain sulphites or other preservatives, as do most fruit juices and dried fruit, which also tend to upset the blood sugar regulation. Dried fruit without added preservatives tends to have moulds. Preferably soak dried fruit or frozen vegetables and discard the water to minimize preservatives or moulds. Potatoes contain the toxic alkaloid solanine in any green parts, and arthritis often improves when avoiding nightshade vegetables including potatoes. Healthiest are non-greened potato peels and fresh, raw potato juice. Also avoid any green tomato pulp or skin, and preferably use red instead of green capsicum (although the latter is not as harmful as green potatoes or tomatoes).

Tea and coffee have some benefits but become a problem if used in an addictive way. The same applies to alcohol. It can be beneficial as red wine, except for wines that contain the commonly added sulphur preservatives, which can trigger breathing problems in susceptible individuals. In this case, try to obtain wine without preservatives. Some traditionally brewed beers as well as some herbal liquors can be beneficial to stimulate the digestion. However, in general it is better to minimize alcohol intake and to test yourself against any preferred brands for food sensitivity. An allergy is commonly to a raw material or additive rather than to the alcohol itself.

Raw Food Diet: From my experiences of healing as well as from reading many published reports, it is clear that a raw food diet is not only the most natural but also the most effective nutritional measure for healing and rejuvenation. While only few will be strong enough to adopt a complete raw food diet for life, it will be good for everyone to have occasional periods on raw foods only.

Here are the main biochemical and nutritional advantages of raw food as compared to cooked food:

- higher vitamin and mineral content
- minerals are largely present as biologically active colloids
- an abundance of helpful enzymes and bioenergy or life force
- proteins remain in their natural condition instead of being denatured
- there is no increase in white blood cells in the digestive tract (digestive leukocytosis) as after eating cooked foods; leukocytosis is an immune response against toxic or microbial invaders
- their polyunsaturated fatty acids and cholesterol do not become oxidized and carcinogenic nor contribute to heart disease, unlike cooked food
- glucose is absorbed slowly, protecting the blood sugar regulation
- no overweight or obesity produced
- much less food is required as counted in calories
- cleansing, rejuvenating, and anticancer properties

However, there are some risks to a raw food diet. Some raw food contains enzyme inhibitors, goitrogens (causing goitre), vitamin inhibitors, toxic substances (in some exotic plants), parasites, and bacteria (in raw meat). In addition, some people starting on a raw food diet will have strong initial reactions due to the cleansing process the foods stimulate.

There have been cases reported of severe digestive disturbances from eating large quantities of raw soybeans, broad beans, nuts, and wheat germ over long periods. This problem is due to the high content of inhibitors of protein-digesting enzymes that protect these seeds from self-destruction. Basically, all seeds have inhibitors of enzymes for the ingredients of which they are composed, but most do not have such a surplus of inhibitors as to cause digestive problems. Potatoes, sweet potatoes, and egg white also contain enzyme inhibitors.

To obtain the maximum benefit from a raw food diet, it is necessary to free most of the enzymes from their inhibitors. Cooking destroys the inhibitors, but also the enzymes. The better solution is to sprout the seeds or at least soak them to initiate the germination process; discard the water used for soaking.

Trypsin, the main protein-digesting enzyme of the pancreas, occurs in seeds. In one reported scientific experiment, the trypsin activity during germination of seeds increased from 7.5 units to 60 units after 24 hours, to 257 units after two days, and to 333 units after three days. The inhibitor activity decreased from 100 percent in dry seeds to zero percent after 24 hours of germination.

Fermentation with lactic acid bacteria also enhances the enzyme activity in seeds. This process is similar to making yogurt or sourdough. Even soaking nuts for a day before eating will greatly help. Their digestive enzymes are freed from their inhibitors to a certain degree when their tissue is damaged, and we see this when grated potatoes turn brown or mashed raw plants or fruit deteriorate within hours unless refrigerated. Similarly, raw egg white is easier to digest when it is well beaten.

Goitrogens, which are in some raw foods, interfere with the body’s use of iodine and therefore may contribute to the formation of goitre or to an underactive thyroid.

Goitro-gens are mainly found in the cabbage family, also in soybeans and the skins of red-skinned peanuts. Cooking these or adding kelp to the diet can overcome this problem. Nevertheless, with an underactive thyroid, it is better not to eat much goitrogenic food in its raw state.

Thiaminase is an enzyme in raw fish that destroys vitamin B1, but that is not a problem if food is correctly combined. Avidin in raw egg white makes biotin unavailable; this leads to a deficiency only if a large percentage of the daily diet consists of raw egg white. Furthermore, egg yolk is very high in biotin, so using whole eggs may not diminish your biotin intake.

Those who have lived all their lives on conventional food are likely to have a fair amount of metabolic wastes and toxins stored in the body. These will be released by a raw food diet and can temporarily overload the organs of elimination and cause skin problems, headaches, weakness, and other discomfort. Therefore, it is advisable to move only very gradually towards a raw food diet or to have short cleansing periods on raw foods while also stimulating the organs of elimination.

The basis of the recommended raw food diet is fresh vegetable juice as well as sprouted and fermented foods. Sprouted seeds can serve as the main part of a vegetable salad. Red beet is excellent and can be finely grated into the salad together with carrot, turnip, sweet potato, or pumpkin. Tomato and grated cucumber can be used for flavouring, with a salad dressing of extra-virgin olive oil, lemon juice, herbs, and spices. Instead of lemon juice, you can use cider vinegar, ascorbic acid, or rose-hip powder. Add lecithin as an emulsifier; shake well in a jar and keep refrigerated.

Almonds, nuts, and sunflower seeds improve through starting the sprouting process; otherwise, eat them soaked and possibly blended or, better still, fermented as nut and seed cheese (see Step 37).

While I do not recommend eating commercial eggs from hens in battery cages, even if well cooked (to kill salmonella, which is common in factory farming), eggs should only be ingested raw if they are free-range or organic. Swallowing an egg with yolk and white intact can cause indigestion in individuals with a weak digestive system. A raw egg is much easier to digest if it is well mixed with other food or drink.

Alternatively, you can beat the white and then mix the yolk back into it. You can easily detect how fresh an egg is by immersing it in water. If it lies flat on the bottom, it is fresh. The older it is, the more the blunt end will rise. If it stands on its tip, it is rather old and may be better cooked; if it floats up, then it is rotten.

If you eat flesh food, you can marinate fish or other seafood and even liver and other meat, which are best minced (see the recipes). This destroys parasites and makes raw flesh food safe to eat. You can also use raw minced meat, suitably flavoured with grated or chopped onion, tomato, or radish, and with lime or lemon juice.

Use only organic or otherwise safe meat from a reliable source. Never use meat from a feedlot; also, pork is generally not safe. The safest conventional red meat appears to be lamb. (For genuinely free-range, grass-fed beef, bison, and ostrich, see: www.mercola.com). As a further precaution, periodically use an herbal anti-parasite program (see Step 44) and possibly an electronic blood purifier (Step

19). With these precautions, raw meat should be safe and healthy, while I regard all cooked meat as more or less unhealthy, partly because of its lack of enzymes, and also because of the formation of carcinogenic chemicals when heated above the boiling point.

(For further ideas on eating meat raw and links to related sites, see: www.rawpaleodiet.org) You can also mix a free-range egg yolk with the salad; if you eat the egg white raw, it is best beaten or blended.

If you have no problems with mucus, such as occasional colds, a blocked or running nose, or respiratory problems, you can use fermented goat's milk or raw cow's milk as yogurt, sour milk, cheese, or cottage cheese. However, be careful with commercial fermented foods, such as yogurt, sauerkraut, or marinated fish. Rice bread or flat bread, preferably made from sprouted rice or otherwise made by soaking and blending the grains and possibly adding a sourdough starter, is suitable as a staple food. It solidifies to crunchy flat bread at low oven temperatures or by exposure to the sun. Sprouted rice can be blended with water, fermented milk, and possibly banana as the basis of a meal.

The most problematic food group is fruit, as I've suggested throughout this section. While meat is grounding, fruit is sensitizing. For insensitive individuals, especially with raised blood pressure, it is beneficial to have plenty of fruit, often as a meal. Especially good are acidic fruits, such as oranges or pineapple. Sensitive individuals, however, feel best with either neutralized fruit or only a small amount of subacid fruit between meals. Insensitive individuals can use fruit as their main source of energy, while those more sensitive can use rice baked at a low temperature and sprouted seeds.

Bananas are a good bulk food for most individuals on a raw food diet. Sensitive individuals can eat them in the early stages of ripening or use mainly small or non-sweet varieties. They do well in breakfast as part of the Linseed Yogurt recipe (see Step 37).

On this raw food diet there is no need to be concerned with variety or trying to have different foods for successive meals. A sprout salad contains all the required nutrients for health and healing; most individuals are happy to have the same basic salad for all meals with only minor variations depending on what is available. The only precondition seems to be that you overcome the addiction to cooked food by learning to make tasty raw salads. In cold weather, you can flavour a fresh vegetable juice with spices; carefully heat it to just below 120º F and drink as a broth.

Most raw foodists will be content to have about 90 percent of their food raw and be free to have some cooked food when eating out and at parties. Even tahini (ground, roasted sesame seed paste) and kelp powder (commonly used with raw food diets) are in fact cooked or heated. Except for food to which you may have been addicted, try to let your body or intuition decide your food selection and meal composition.

Slimming: The high-quality diet and even more so the raw food diet will normally ensure gradual weight loss for overweight individuals, while at the same time maximizing health improvement. I always put the main emphasis on the health aspects of the diet rather than its slimming potential. As we become healthier, we automatically find the right weight for our constitution. However, no diet will help if your eating problems are emotionally based, so you have to heal your emotions first.

Generally, the most important dietary aspect of slimming is to avoid all cereal or grain products, as well as sweetened food. Try experimenting with these possibilities:

- Have an appetite-reducing protein drink 30 minutes before meals with any combination of the following ingredients and in any amount that suits you: green barley juice powder, spirulina powder, food yeast, ground linseed, linseed (flaxseed) oil, or olive oil. Use this as a warm or cold drink to keep hunger away for hours.
- About 30 minutes before meals, stir one teaspoon of psyllium hulls in a glass of water and drink immediately. If necessary double the amount to two teaspoons in a large glass of water. Psyllium swells up 40-fold to fill the stomach with a soft gel, so don't overuse.
- Eat only lentils or other legumes, green vegetables, and sprouted seeds in addition to the above.
- For cooking, use mainly legumes, non-sweet, non-starchy vegetables and flesh foods; for starches choose high-fibre varieties (e.g., brown rice, rolled oats) and add ground linseed.
- Adopt a mono-diet: one day eat protein foods only, the next day fruits, and the third day sprouted seeds, vegetable salads, and vegetable juices. Repeat as often as required.
- Adopt a cleansing raw food diet with mainly vegetables and green juice, sprouts and fruits, nuts and oily seeds, possibly raw beaten egg, and marinated fish. Sensitive individuals should neutralize fruit acids (see Step 6).
- For two weeks, eat only protein and fat foods and green vegetables with less than 40 g of carbohydrates per day; ground linseed is the best protein/fat food.
- Repeat times when you rely on vegetable and grass juices in addition to vegetable salads.
- Avoid all sweet food, as it stimulates the appetite and the synthesis of fat. In particular, do not eat fructose-containing foods with starches and eat fresh fruit on its own.
- Avoid all grain products; especially helpful for people with blood group O.
- Avoid all fats and oils except fresh, unheated linseed oil, which will not put weight on you; experiment with coconut oil in your diet.
- Preferably skip the evening meal or have only a small salad or apple.
- Get some light daily exercise, such as a brisk 15-minute walk or jogging; use a rebounder or do some other aerobic exercise. The metabolic rate is speeded up for a long time after exercise. Take 1 or 2 g of carnitine before exercising to burn up more fat.
- Use a low-allergy diet combined with the correct foods for your blood group (see Step 43).

- Use mind therapy, such as guided imagery and affirmations; look at your belief systems; release negative emotions; and express your emotional needs.
- Eat very slowly; peacefully chew every mouthful until it is liquefied; savour the flavours as they develop during chewing; make every mouthful last about two minutes.

Hypoglycaemia Diet: Most people who are sensitive, are overacid, have allergies, or have a low energy level are usually hypoglycaemic to some extent and will benefit from this diet. Hypoglycaemia means low blood sugar: when a hypoglycaemic person eats sweet foods, the blood sugar level often soars too high at first, but later falls to below normal. During this fall and while the level is below normal, a variety of distressing symptoms can affect the breathing, circulation, digestion, or the emotional and mental condition.

When the blood sugar level rises after sweet food intake, the pancreas reacts by releasing insulin, a hormone that regulates the blood sugar level. In a hypoglycaemic person, the pancreas releases too much insulin, especially in response to ingesting sucrose or a combination of glucose and fructose. The excess of insulin causes glucose to enter the cells too quickly. This then creates oxygen deficiency within the cells and instead of energy, lactic acid is produced in an anaerobic process. All this makes the body tissues overly acidic and deprives them of minerals and energy.

The mainstays of the hypoglycaemia diet are:

- Gluten-free grains: rice, millet, maize, buckwheat, sago, tapioca
- High-protein seeds: chickpeas, lentils, fenugreek, peas, beans, almonds, sunflower seeds
- Animal protein: white, red, and organ meat, fish (especially sardines and salmon), eggs, goat cheese
- Sprouted seeds and vegetables: cooked, raw, and juiced; avocados, coconut oil, and extra virgin olive oil

Eat grains whole in preference to flour or meal products. However, coarsely ground grains can be beneficial if eaten raw: Immerse them in water or seed milk and let soak overnight. Rye sourdough bread and rye crispbread can be eaten after allergy testing (see Step 5). Raw egg yolk from eggs from free-range chickens can be used as a salad dressing. Other important additions are spirulina, bee pollen, ground linseed, kelp, lecithin, gelatine with vegetables, and fresh green-vegetable juices, especially grass juice or dried grass juice powder and possibly food yeast. Try coconut oil and freeze-dried liver to increase your energy level.

Most of the vitamins and minerals are just under the skin of the produce, so do not discard the peel of suitable fruits and root vegetables. Nuts should be used sparingly, except if sprouted or made into seed cheese (see Step 37). Eat cooked sweet vegetables only with a protein meal; raw carrots, on the other hand, are good any time.

Reintroduce fruits to your diet cautiously and with self-observation. Use sweet fruits only in the early stages of ripening, before they become too sweet. This applies, for example, to bananas and papaya, but do not use fruits that are sour before ripening. Subacid fruits, such as apples, cause less problems than sweet-acid fruits, such as oranges. Preferably eat fruits only sparingly and as a snack between meals, and neutralize any fruit acids (see Step 6). However, under-ripe or non-sweet varieties of bananas may also be suitable.

As a general rule, try to avoid:

- Sweet food, initially including all sweet and acid fruits and sweet vegetable juices
- Food containing lactose, especially milk products, though butter can be used
- Vinegar or other fruit acids, unless they are neutralized
- Initially no wheat, later only sparingly after allergy testing
- Tobacco, alcohol, drugs, artificial sweeteners, highly processed foods or those with added chemicals

Coffee and tea can be used in moderation, but only if you are not allergic or addicted to them. Lightly salted food may be beneficial. If possible, eat a protein breakfast and a light evening meal, and take a rest after lunch. The best protein is sprouted legumes as the main part of a salad. According to the severity of the symptoms, this diet may be followed strictly or in a more relaxed form and generally as part of the high quality diet.

* * * * * * * * * * *

Step 35

SPECIAL FOODS

Take advantage of the health-giving properties of bee pollen, sprouted seeds, fermented foods, and purple foods.

Bee Pollen and Special Foods: Various foods have general healing qualities or help to overcome specific health problems. For example, foods high in enzymes and growth hormones improve our overall health. Sprouted seeds, bee pollen, spirulina, grass juice, and unheated lactic-acid fermented foods are in this category, as are fresh edible flowers such as nasturtium, impatiens, and pumpkin. While the stamen of pumpkin flowers that bears the pollen is somewhat bitter, flower petals from the male plants are especially high in bioflavonoids, which help fight against and prevent allergies and inflammations.

Bee pollen is one of the best foods or supplements available. It is a raw food rich in enzymes, hormones, vitamins, and minerals. You can take one to three teaspoons of bee pollen several times daily or use it to flavour your freshly pressed vegetable or grass juice. To make it easier to digest, soak the pollen in liquid for some time before ingestion. It is also a good addition to yogurt or other food to be fermented.

A highly recommended protein combination is bee pollen with spirulina and ground linseed. Taking one or more teaspoons of each three times daily mixed in a juice or other suitable liquid before meals can provide up to half of your daily protein needs. The average protein content of spirulina is 65 percent; pollen has 25 percent, and linseed (flaxseed) 22 percent. Spirulina provides a link between animal and vegetable protein, and it is high enough in vitamin B12 that vegans (vegetarians who do not ingest any animal products) can live on it.

The protein in this combination (bee pollen, spirulina, and linseed) is uncooked, so it is easily digested and good as replacement for a part of the usually eaten animal protein. Another advantage is a favourable calcium-to-phosphorus ratio and a high mineral content, making it alkalizing rather than acid-forming in the body. Add to this a handful of sprouted seeds once or twice a day, and you have the basis of a healthy diet, which you can supplement with meat rather than using meat as a staple.

Foods high in sulphur can be increased to improve detoxification. Such foods include raw egg yolk, onion, horseradish, watercress, turnip, and other raw foods of the cabbage family. Most of these are also helpful in treating digestive ulcers; horseradish can be used to help treat mucous complaints. A supplement high in organic sulphur is called MSM, available in many health food stores.

Bananas are good energy providers for those with a weak digestion. However, overripe Cavendish bananas (this is the most common variety) often cause digestive discomfort or allergies in sensitive individuals. This problem is commonly caused by over gassing the green bananas with ethylene gas shortly before sale to make them ripen quickly. Ladyfinger bananas or other small varieties do not normally cause digestive problems. Sensitive individuals generally should avoid overripe fruit. Acid citrus fruits are excellent for improving liver functions and lowering high blood pressure, blood sugar, cholesterol, and fat levels in insensitive individuals.

Jerusalem artichokes consumed in frequent small quantities are beneficial for kidney function and for diabetes. Eat one the size of an egg every other day, and store Jerusalem artichokes in soil, not in the refrigerator. Also eating (organic) liver often helps those with diabetes, as does eating avocados and green beans.

Sprouted Seeds: Sprouted seeds and young shoots such as wheat grass provide a higher degree of vitality and rejuvenation than other foods. I recommend sprouted seeds as a mainstay of a good diet. Most commonly used for sprouting are legumes such as lentils, peas, mung beans, fenugreek, and chickpeas, but almonds, peanuts, and sunflower kernels can also be sprouted. Sunflower and buckwheat seeds make a tasty salad when grown in seed boxes and cut close to the soil when the sprouts are two to four inches high.

Alfalfa sprouts are high in a detergent-like saponin that can damage the intestinal wall in sensitive individuals. They have been shown to aggravate autoimmune diseases and possibly cancer. For others, they may be a good food to eat after allergy testing has confirmed you do not have allergies to them. Alfalfa sprouts should be exposed to light after the leaves emerge, while other sprouts usually taste better before their leaves develop. Alfalfa seeds easily rot in wet conditions and are best sprouted in an upturned jar covered with muslin held in place with a rubber band.

It is important to use seeds with a high germination rate. Buy a small quantity for testing, and if they sprout well, buy a larger quantity from the same source and keep them in an airtight container. Lentils and mung beans are easy to sprout. Soak the seeds overnight and then rinse several times a day, drain well, spread out lightly, and keep covered. Wash well again just before eating to remove any bacteria and fungi. If seeds are of poor quality and start rotting easily, use them when the sprouts just start to appear, otherwise wait until they are up to one inch long, which is usually within three days.

Mung beans may retain hard shell pieces after soaking that can damage your teeth. This can be avoided by pouring some boiling water over the seeds initially and letting them soak in the cooling water. Peas and lentils, on the other hand, like it cool and sprouts easily deteriorate in hot conditions. Sprouted seeds can be refrigerated in a closed container after they have reached the desired length. Individuals sensitive to fungi and microbes can wash sprouted seeds in diluted hydrogen peroxide shortly before use. If you have difficulty chewing or if sprouted seeds cause much wind, try using them as part of your fresh juice.

If seeds do not sprout well, the cause may be as follows:

- Seeds may not be viable, may be too old, or have been stored incorrectly.
- Temperature during sprouting is too high or too low.
- Seeds were kept too dry or too wet during sprouting; experiment and adjust conditions.

Fermented Foods: The use of fermented foods was widespread in former centuries. Well-known examples are yogurt, cheese, cottage cheese, beer, cider and cider vinegar, pickles, miso, sauerkraut, and sourdough bread.

The advantages of fermenting foods are several:

- Makes minerals more easily available for the body to absorb
- Eliminates enzyme inhibitors and other anti-nutrients
- Increases the amount of vitamins in the food and creates an abundance of enzymes
- Preserves the food and introduces delicious flavours
- Makes food much more easily digestible, making it essentially predigested and suitable for those with a debilitating condition and weak digestive system

Today most of these traditional fermented foods such as vinegar, pickles, cheese, cottage cheese, and baking products are commercially made by chemical means. This disuse of the fermentation process in preparing our staple foods may well be a major contributing factor in chronic diseases.

It is most important in the fermentation process that a desirable strain of bacteria develop. This depends on the bacteria already present or introduced, as well as on the temperature during fermentation. Basically, we want lactic acid bacteria that convert glucose into lactic acid. In order to ensure the development of pure bacterial cultures, the food is often heat-treated and then inoculated with the desired strain. However, I contend the product is nutritionally superior if it has not been heat-treated. Grains such as rye, oats, millet, wheat, and brown rice and vegetables such as cabbage are best fermented at a warm room temperature, while cultures containing yogurt or acidophilus bacteria do better between 85° and 105° F.

There is a distinction between the fermentation of yeast and that of lactic acid bacteria. Yeast fermentation is used in traditionally made beer, wine, and bread. Due to the overuse of antibiotics in recent decades, many individuals are now allergic or sensitive to yeast. Therefore it is advisable for sensitive individuals to restrict intake of yeast products and for others to abstain periodically or after antibiotic treatment and to test for incompatibility.

When fermentation is uncontrolled, as it is when making sauerkraut or with water in which grains have been soaked for several days, a mixture of lactic-acid bacteria and yeast usually develops. This mixture is not a problem for many, but sensitive individuals need to watch out and may have to avoid this.

In making yogurt, sour milk, sour cream, and cheeses, lactic-acid fermentation is the preferred natural method and still widely used in cottage industries. However, due to the mucus-forming potential of lactose, these products should be used with caution, especially if made from cow's milk. The curd causes less of a problem than the whey (the whey contains most of the lactose), and goat's milk less than cow's milk. Most commercial fermented milk products are heat-treated and often very high in lactose, and it is best to avoid them.

A health food seed cheese or seed yogurt can be used. This is commonly made from oily seeds such as nuts, almonds, or sunflower kernels by adding acidophilus culture to the soaked and blended seeds. Contamination of the seed cheese with yeast can be minimized by washing the seeds in diluted hydrogen peroxide before blending, adding a large amount of starter, and letting it ferment for only a short period. The same applies to sourdough bread, which is also recommended.

Do not use strongly fermenting products that have become very sour. You can discard the whey and wash the curd to make it acceptable. Nevertheless, do not use any fermented food that has a foul smell or taste. The water used for fermentation should be free of chlorine and fluoride, as these will poison the enzymes released in the fermenting process.

Fermenting or fermented products should not be kept in metal containers nor should a metal spoon be left immersed in the food as this increases your exposure to toxic metals. Plastic containers are not so good either; use ceramic or glass containers instead. Fermented foods can be refrigerated for a week or two; part of the former batch can be used as a starter for a new lot, but beware of contamination with yeast. It is especially important to eat fermented foods during and after a course of antibiotics and if you have cancer or another degenerative disease.

Purple Foods: The term purple foods as used here includes all foods that yield purple juice when cooked or pressed with their skin. The color pigment in purple foods belongs to the large group of bioactive natural chemicals called bioflavo-noids. We find them as red, blue, and purple pigments in a wide range of flowers, fruits, and vegetables, but also in the pulp of citrus fruit. Two well-known bioflavonoids are rutin and quercetin.

Another important nutrient group is the carotenoids that form the yellow, orange, and red colours in flowers and fruits. Best known among these is beta-carotene. Even more important for healthy eyes and especially to protect the retina to prevent blindness from macular degeneration are lutein and zeaxanthin. These two are found in spinach, lettuce, broccoli, peas, and corn. Lycopene appears to protect the health of the prostate gland; studies have found it protective against exercise-induced asthma. You can find it in red-hued fruits and vegetables, such as tomatoes, pink grapefruit, guavas, papaya, and watermelon.

If you have the opportunity, frequently snack on flower petals or put them through the juicer. Suitable are the big petals of male pumpkin flowers or the whole flowers of impatiens or nasturtium; nearly all flowers of edible vegetables can be used. However, if you do not know whether a flower is edible, avoid using it in this way.

The pigments of purple foods are highly unsaturated and therefore are strong antioxidants. One important group of purple pigments is the proanthocyanidins (PAC's); when many individual anthocyanin molecules are linked together they are also called oligo-proanthocyanins (OPC's). One such compound that is commercially extracted from pine bark is sold as pycnogenol; a variant with similar qualities is sold as grape seed extract. The OPC's have lost their purple color, but this may be restored when they are broken down in the body into their individual components. In nature, anthocyanidins are bound to different sugars, and during digestion, the sugars are split off to release them for absorption.

The color of black, purple, or red grapes, blackberries, blueberries, red wine, and red cabbage consists of anthocyanins, the individual molecules and active form of PAC's. The pigment of red beet belongs to a subgroup of anthocyanins, the betacyanins. Both groups have strong antioxidant and anti-inflammatory properties and improve the elasticity of blood vessels and skin. PAC's, anthocyanins, and betacyanins also help to restore normal oxidative metabolism or cell respiration.

They are important not only for normalizing cancer cells, but also for conditions of chronic fatigue and general lack of energy, as they assume the role of oxygen as electron acceptors. Purple foods ingested in sufficient amounts can improve the elasticity of skin and blood vessels. In combination with other bioflavonoids, they are beneficial with various diseases such as allergies, inflammation, infections, parasites, cancer, liver disease, vascular diseases, thrombosis (blood clotting), chemical toxicity, and coronary artery disease.

Purple foods should generally be eaten with the skin on as this is usually the most strongly coloured part of the plant. However, with purple onions the outer skins are not edible, but you can utilize them as a tea. Grapes should be well rinsed before eating to remove fungi or moulds. Sensitive individuals and especially those with Candida may find it preferable to soak the grapes for a while in hydrogen peroxide. When fresh grapes are not in season, the juice of black grapes has reportedly been used with good success in the treatment of cancer. For this, a 24-ounce bottle of dark grape juice can be sipped at intervals during the morning, taking no other food before lunch. See "The Grape Cancer Cure" on website: www.quantumbalancing/grape.htm.

Red beet is the best purple vegetable. The purple pigment in beets has been shown to increase and normalize cell respiration, the oxygen-based energy production within cells. Red beet is one of the key foods in preventing as well as curing cancer; it is equally important in the treatment of other degenerative diseases, such as chronic fatigue syndrome, characterized by reduced cell respiration.

When cellular energy is produced through the oxidation of nutrients, electrons and hydrogen ions are transferred onto the inhaled oxygen to produce water and energy. In cancer cells and in conditions of chronic fatigue, the respiratory enzymes that accomplish this transfer have been diminished or destroyed. The color pigment in beet root (and other purple food) strongly binds electrons and hydrogen and can reactivate the production of cellular oxidative energy. Researchers have shown that the energy production of cancer cells can be normalized by a combination of red beet, raw fermented food, and vitamin C.22 The multiplication of cancer cells will thus stop, and tumours become non-virulent. Clinical tests using red beet with cancer patients revealed that often tumours regressed and disappeared. In the cancer therapy with red beet, two pounds were used daily. However, if used together with other purple foods, much less beet may be needed. For instance, combine one pound of red beet with one pound of dark grapes or one pint of dark grape juice.

Use plenty of red beets grated in salads, juiced, and cooked, and remember, the residue from the juicing can be cooked. Occasionally, a small beet may leave an acrid aftertaste. Taste suspect beets before making salads or juice; cooking them is fine, and beet tops can also be cooked. Tinned beet has lost most of its pigment and is of little nutritional value.

A Checklist of Laxative Foods

For the improvement or maintenance of good health, it is desirable to have two or three bowel movements daily. A general lifestyle improvement will do much to achieve this goal, but initially it may be necessary to include some laxative foods as well. Here are some practical suggestions:

Bran: A mild laxative, however, it accentuates mineral deficiencies and weight loss. Wheat bran is not suitable for anyone who is sensitive to wheat. I do not recommend wheat bran, but you can experiment with rice bran, provided you take trace mineral supplements with bran-free meals.

Freshly fermented foods: These often have a strong laxative action.

Isotonic flush: Add 9g of salt (approx. one heaped teaspoon) to one litre of water, or use one cup of seawater diluted with 3 cups of fresh water. Drink this within a few minutes on an empty stomach and take no other food or drink for the following 30 minutes. This is excellent for cleaning the whole of the gastro-intestinal tract, but it should not be used if a severely restricted intake of sodium is indicated for other conditions.

Linseed (flaxseed): Mildly laxative and soothing for irritated bowels. Preferably use ground linseed, otherwise one tablespoon of dry seeds with water some time before meals and at bedtime; swallow whole. If the intestines are irritated, use fenugreek or slippery elm powder with the linseed.

Magnesium salts: Take one tablespoon of Epsom Salts with water; milk of magnesia is excellent if neutralised with ascorbic acid; alternatively, try increasing amounts of magnesium chloride.

Molasses: Use on its own or with yeast in warm water.

Mustard seeds: Take half a teaspoon of whole dry seeds with water twice daily, some time before a meal; do not chew. This helps to strengthen bowel function and overcome gas formation. Continue for two months or as required. Do not use mustard seeds if the intestines are irritated.

Prunes: May be eaten raw, soaked or cooked.

Psyllium hulls: These are preferable to bran. Stir a teaspoonful in a large glass of water and drink immediately about 30 minutes before meals as often as required for either laxative action or weight reduction.

Senna: Pods, powder, or leaves can be added to herb teas.

Urine: drinking a cupful of your diluted morning urine has a good laxative effect.

Water: Drinking a large quantity of pure water within a short period will cause part of it to rinse the bowels. Take two pints or more of warm water before breakfast (see Step 36).

Red beet may be available only seasonally, but you can store a large quantity in moist sand. Keep the tops exposed in a cool, shaded place with just enough moisture to prevent drying out. After a good root system has developed, you can also let them continue to grow in a sandy and well-drained soil, neither too wet nor too dry to avoid rotting or mould development, but check them frequently.

Raw Egg: Fresh raw eggs, genuinely free-range from hens fed on greens and grains, have health-giving and healing properties. However, egg allergy is widespread so if you habitually consume them, test yourself for egg allergies. Raw egg yolk does not raise the cholesterol level, and it is high in sulphur compounds that strengthen connective tissue and detoxify the liver. Fresh raw egg yolk has long been known to enhance or rejuvenate sexual energies required for having sex, but more importantly for creative activity in general and for physical fitness. Very fresh raw, beaten egg white has germicidal properties.

The cell walls of lymphocytes (white blood cells) become more rigid with aging, AIDS, and viral infections, mainly due to increased cholesterol levels in cell walls. Lipids extracted from raw egg yolk have been reported to normalize rigid cell walls and fully restore immune functions. This has led to great improvements in AIDS patients treated with egg-yolk lipids.

Linseed and Sulphur-Rich Foods: Together with fish oils, linseed or flaxseed oil is our best source of the omega-3 essential fatty acids. These polyunsaturated fatty acids are important for brain and nerve functions, for the fluidity of cell membranes, and for oxidative cellular energy production. Of great importance for this energy production are lipoproteins formed from essential fatty acids and sulphur-proteins (cysteine, methionine). These are concentrated in active tissue such as glands, liver, brain, muscles, and skin and are needed for oxidative energy production or cellular respiration.

Key enzymes formed from cysteine and essential fatty acids are typically deficient in individuals with degenerative diseases such as cancer, chronic fatigue syndrome, diabetes, and many skin diseases as well as heart, liver, and kidney degeneration. These and other diseases have reportedly been reversed with high intakes of high-quality linseed oil and sulphur-rich foods.

Linseed oil is very sensitive to oxidation (spoilage) and special precautions are needed to obtain and retain a non-rancid product. The linseed or flaxseed oil must be fresh when purchased, not more than a few months old and produced below 120° F without light and oxygen; store it in a cool, dark place and refrigerate before and after opening. It should never be heated.

As high-quality linseed oil is expensive, usually it will be sufficient to use freshly ground whole linseed instead of or in addition to the oil. Linseed contains 35 percent linseed oil and you can use several tablespoonfuls of ground linseed a day. This also has a mildly laxative effect. Grind it fresh in a coffee grinder or, alternatively, use a blender, screen it through a strainer, and add the coarse residue to the blender again with the next lot.

Ground linseed can be refrigerated for a few days but is better stored in the freezer. For individuals with malabsorption problems and for those wanting to improve their skin, linseed oil can also be rubbed on skin normally protected from the sun. In addition, I recommend unheated extra-virgin olive oil internally as well as on the skin. This is a good source of squalene, a strong antioxidant that is also important for energy production; it is also a key ingredient in the healing properties of shark liver oil.

Regarding the sulphur foods, the main sulphur amino acids are L-cysteine and L-methionine. Eggs are a good source of these, but they can easily be produced from other sulphur-rich foods such as onions. As mentioned, the best supplement form of organic sulphur is MSM. In her cancer therapy, Dr. Johanna Budwig, a fats researcher, recommends cottage cheese ("quark" in German) made from raw fermented skim milk as the main source of sulphur amino acids. However, according to the blood group diet theory (explained later in the book), this may only be suitable for individuals in blood groups B or AB, and not with hormone-related cancer.

Do not use D,L-cysteine or methionine, but use only the natural L-forms. While cysteine is very important for detoxification and energy production, new research shows that it may also act as a neurotransmitter in the brain and that high doses can create an imbalance in the brain of susceptible individuals. Therefore, it is generally best to obtain it or synthesize it from sulphur-rich foods. High doses of methionine, on the other hand, can cause a problem in individuals who are deficient in vitamin B6; problems can be avoided by supplementing it together with vitamin B6.

* * * * * * * * * * * *

Step 36

WATER AND JUICES

Heal yourself with fresh vegetable and grass juices.

Let’s start by the discussing water in your diet. It is important that we take in sufficient water for the smooth functioning of our metabolic processes, and especially to flush away their end products. In normal health and on a diet that includes fresh fruits and vegetables, an additional fluid intake of about a quart daily is usually adequate. However, in the following conditions, drinking close to two quarts of fluids daily is desirable, and more than that if you are overweight:

- If mainly concentrated foods are used in the diet
- If you’re on a diet high in protein
- If the urine is cloudy, has a strong smell, or is dark in color
- In hot conditions
- While breast-feeding
- During the initial stages of health improvement and in cleansing periods

The Benefits of Water: Your fluid intake can consist of water, diluted juices, and herb teas. Concentrated drinks such as coffee, milk, and strong tea do not count, as their water content is needed to wash out their own residues. Occasionally check that you produce about three quarts of inoffensive (in terms of odour) urine per day. In hot climates, there is a high incidence of kidney disease because many people do not drink enough water. For instance, a high incidence of oxalate kidney stones among British troops in India during World War II has been blamed on excessive tea drinking

If using large amounts of fluids, you must ingest adequate amounts of water-soluble vitamins and minerals; otherwise deficiency symptoms can develop as you’re likely to flush these nutrients out of your system. The time of drinking, too, is very important. One should be careful not to dilute the digestive juices during and after meals, although some liquid, such as in the form of herb teas, can be used after dry meals. Generally, however, it is better to drink before meals and to have the bulk of the fluid intake before breakfast. This is especially important if the digestive system is weak. I regard it as highly beneficial for the kidneys to flush them regularly with a high-volume fluid intake, such as a quart of water or diluted juice before breakfast.

After drinking water, wait for about 20 minutes before starting a meal, and after drinking juices, wait for 30 to 60 minutes, depending on the quantity. It is good to drink a glass of warm water or herb tea first thing in the morning to wash down mucus and debris.

Chlorine and added fluoride in drinking water are always harmful. However, if the water is high in calcium, then fluoride is less harmful, because calcium fluoride is nearly insoluble and not well absorbed. On the other hand, if the water is soft or rich in magnesium, iron, or sulphur, then fluoride can be harmful. Avoid chlorinated or fluoridated water for drinking or cooking; however, most of the chlorine can be removed by bringing the water to a boil and letting it cool before drinking.

As a strong enzyme poison, fluoride is especially harmful for individuals with chronic degenerative diseases. In countries where water fluoridation is common, we can suspect that all commercial liquids with added water contain fluorides. Commonly, these include all soft drinks, soy milk, and even 100 percent fruit juice, which is usually reconstituted by adding water to concentrates. However, municipal tap water also contains other added chemicals and a good water filter is recommended.

Our magnesium intake has greatly declined since the 1940s because calcium fertilizers have been overused. The death rate from arteriosclerotic and degenerative heart disease is highest in countries with the lowest levels of magnesium in the drinking water; this is notably the case in Australia. Within Australia, the death rate is lowest in Perth, which has the highest magnesium-to-calcium ratio in drinking water, while the death rate is highest in Sydney, which has the lowest magnesium-to-calcium water ratio of Australian capital cities.[23] In the U.S., in six Missouri counties with low magnesium levels in the water, the death rate from diabetes was more than four times higher than in six other counties with high magnesium levels in the water. Water high in calcium is acceptable if it is also high in magnesium; otherwise, you can use magnesium supplements.

Rainwater can be used in unpolluted rural areas. Unfortunately, galvanized iron now commonly has added aluminium with its zinc coating, and this may increase the aluminium level of water collected from modern galvanized roofs and water stored in galvanized tanks. I do not know whether the amounts of aluminium that might be dissolved are significant, but it's best to be prudent. Public water supplies are routinely purified with aluminium compounds; a study in England found that when aluminium levels in public drinking water were high, there was a correspondingly high incidence of Alzheimer's disease.

Water rich in inorganic iron, which leaves brownish residues, is harmful if used over long periods. Lead and sometimes copper accumulating in the body contributes to chronic diseases. Therefore, if you are in poor health, avoid drinking the water from lead plumbing systems and be careful with new copper pipes. Older copper pipes are coated with more resistant copper oxide. Plastic pipes can give off toxic substances as well. The best metal for drinking water systems is zinc, as in galvanized iron pipes; otherwise, ceramic can be used.

Public water supplies are usually heavily contaminated with chemicals, and I do not recommend them as a direct source for drinking or cooking. Tap water is commonly low in energy or life force. Instead use one of the great variety of filters and distillation units on the market.

Bore or spring water in densely populated or farming areas can be used safely only if a chemical analysis shows it to be uncontaminated. It should not have high amounts of calcium, copper, or iron; otherwise, a filter may have to be used or the water boiled to remove excess calcium. The very best water is fresh from a clear, uncontaminated, sunlit, shallow, and cascading stream. Preferably try to energize your drinking water by exposing it to sunlight or other sources of bioenergy (see Step 7).

I do not recommend drinking much fruit juice, except dark grape juice (as explained earlier) or neutralized acidic juices. The basic disadvantage of drinking fruit juices is the rapid assimilation of sugars, which stresses the blood sugar regulation. However, some fruit juice can be used to flavour drinking water, if so desired.

I do not recommend drinking milk, except possibly fresh raw goat's milk or cow's milk when it is still warm after milking. Use coffee and tea in moderation. While they have some benefits, they can cause problems if one becomes addicted to them. Alcohol too is best used only occasionally in small amounts. Beer drinking encourages the formation of mucus. Dry red wine is generally the least harmful and most beneficial alcoholic beverage. It can assist in the digestive process and give some protection against heart disease; grape juice offers similar protection, but has a lower concentration of protective antioxidants; the deeper the color, the better the protection. Avoid any kind of soft drinks.

Vegetable and Grass Juices: Fresh, raw vegetables are high in life force, enzymes, and other ingredients that keep us healthy. Therefore, use plenty—two pounds or more a day if possible. For many, especially the elderly, it will be difficult eating raw vegetables in -sufficient quantities or chewing well enough to extract all the nutrient goodness. That is where juicing comes to the rescue. In this way, everyone, including children and those with dentures, can have a high intake of raw vegetables.

If possible, include some fresh young grasses, leaves, and shoots grown organically in mineral-rich soil, as they are especially high in life force and enzymes. I highly recommend drinking a glass of freshly pressed vegetable juice before most meals or whenever conveniently possible.

Try using a mixed juice including grasses or leaves as well as root vegetables, such as carrot and red beet, and possibly an apple for flavouring. The chlorophyll and proteins of the leaves, while beneficial in their own right, slow down the absorption of sugars from the sweet vegetables and help to stabilize the blood sugar level. You can also include sprouted seeds, edible (vegetable) flowers, the rind of watermelons, the tops of beet root, and other unusual items in your mixed juice. Be imaginative and experiment to find out what is acceptable to your taste buds.

If unsprayed vegetables are not available, avoid sprayed green-leaf vegetables and use the less contaminated sweet vegetables. Squashes and watermelons are not normally sprayed. For individuals with a weak sugar metabolism (diabetes, hypoglycaemia, or a tendency to these conditions), it is preferable to use mainly green juice of leaves and grasses and only a smaller amount of sweet juice for flavouring. Alterna-tively, slow down the absorption of sugars from the juice of sweet vegetables by adding some olive oil and lecithin or ground linseed; then drink slowly and at intervals.

Often children and some adults can be persuaded to drink the juice only if it is flavoured with apple or pineapple. I recommend adding ginger to the juice: either use fresh ginger or ginger juice already frozen in ice cube trays. For flavouring, it is

excellent to stir a few spoonfuls of bee pollen into the juice.
Start drinking the juice immediately after you have made it, but slowly, and in sips. If your stomach is sensitive to juice, drink it with some food. If you make several juices during the day, you may need to thoroughly clean the juicer only after the last juicing, but keep the extraction unit detached and cool in the refrigerator between juicing. Alternatively, you can freeze surplus juice for later use.

If you are unhealthy when beginning such a regimen, nausea or skin reactions may temporarily occur from the release of stored toxins. In such cases, reduce the amount of juice for a while and drink lots of water and take laxative foods. Start with small amounts of grass and weed juice, possibly diluted and blended with other juices. When you get used to it, the green juice may taste fine on its own, but avoid or minimize strongly flavoured leaves such as from radish. Gradually increase to a glassful one to three times daily before meals - the more the better. While fresh juice is best, if you do not have the time to make several juices daily, you can refrigerate or freeze some of the juice for later drinking; also cool it and take it to work in a thermos.

If you have a problem absorbing fats, add some lecithin to the juice to improve absorption of beta-carotene. Sometimes, when juicing sprouted seeds, the centre of a squash, or some mucilaginous leaves, the residue seems to retain much of the goodness even after putting it repeatedly through the juicer. In this case, just add water to the residue to make it sloppy before pressing it through again. As an alternative, you can get more nutrients out of the residue by making a hot broth: Cover the residue with water, simmer for ten minutes, strain, and add some herbs and spices or other flavouring.

What Types of Juicers Are Most Suitable?: Centrifugal juicers, which are commonly available, are not suitable for grass or leaves, and even for other vegetables they are inferior. Static electricity builds up on the outside wall and discharges or inactivates enzymes and colloidal suspensions of proteins and minerals. Non-centrifugal juicers reportedly give much better healing results.

A blender can be used, and grass and other leaf vegetables liquefied with water and then pressed by hand through a strainer. Unfortunately, air beaten into the liquid leads to oxidation. You can minimize this by adding some ascorbic acid. Blend handfuls of grass repeatedly in the same water to concentrate the juice; alternatively, run the grass once or twice through a mincer or a grain mill and then press the pulp by hand through a strainer or cheesecloth.

Grass juicers, either manual or electric, are commercially available and may be sold as fruit and vegetable presses. (You can easily find information and retailers on the Internet.) This type of unit commonly has a slow-turning spiral press. Manual wheat-grass juicers are efficient and good for training the biceps, but not so good for people who do not have the energy to use them daily. Also, I do not like the juice to be in contact with so much metal; cast-iron juicers are much cheaper than the stainless-steel types.

As a general rule, the slower the speed, expressed as revolutions per minute (rpm), the better the quality of the juice. Slow-turning juicers are available as single auger or twin-gear machines. Both handle grasses and leaves very well with an rpm commonly between 70 and 120 ppm. In comparison, a medium-fast juicer has about 2,000 rpm and the common centrifugal juicers 8,000 rpm. While low-rpm juicers produce the highest quality of juice, they are also slowest in pressing a given quantity; bigger twin-gear juicers may be twice as fast as smaller single-auger juicers. (For a comparison of different brands, see www.buyjuicers.com.)

When using a juicer that easily processes grasses and leaves, there are considerable savings in having to buy fewer vegetables. You can use lots of grasses, green weeds, parsley, tops of carrots, and other leaves that cannot be easily used with other juicers. This can not only bring considerable savings, but it also provides juice of much higher quality than can be obtained from commercial vegetables. However, there is also a drawback with low-rpm juicers and growing your own wheat or barley grass: It is very time-consuming.

Juice from Grasses and Weeds: While not so pleasant to the taste buds, fresh young grasses and green edible weeds have a much greater healing potential than the vegetables commonly used for juicing. In particular, they are very much higher in their content of life force and enzymes. Instead of working hard to maintain a big vegetable garden, I believe it to be easier and healthier to let a large part of the garden overgrow with suitable grasses and edible weeds. An additional advantage of young grasses is the short time and minimal care required to grow them as compared to growing vegetables. You can harvest grasses in two to three weeks compared to several months for vegetables.

To find out which weeds are edible, ask a knowledgeable friend or neighbour, or observe what goats or poultry are eating, or chew a bit of a leaf yourself. If it is not bitter, it is not likely to be poisonous, and anyway, there are no poisonous grasses. For a very healthy juice, use grasses, green weeds, and the petals of squash flowers; flavour it with an apple, beet root, and carrot.

Juice made from young and fresh blades of cereal grass is most beneficial. These young blades have the highest vitality and are rich in enzymes and growth hormones that are missing or at low levels in mature leaves or plants, as well as in elderly humans. Kirlian photography reveals that the vitality of leaves starts to diminish soon after cutting and that most of it is lost within hours, though if refrigerated in a closed plastic bag it may keep for a day. Most nutrients will still be available from commercially dried and powdered grass, such as green barley grass (commercially available), but young fresh grass juice has a much higher vitality.

Best known is wheat grass, but I contend that all fresh, young grass grown in good soil has similar healing and rejuvenating qualities, though the flavour and toughness may differ greatly. Barley grass grows more vigorously than wheat grass and tastes good. Young ryegrass (not the cereal rye grain), grown by farmers as pasture, has a pleasant flavour and can give you several months of repeated cutting from one planting.

Experiment with different varieties of grasses to see which grow best in your climate and soil conditions, and find one you like. When using lawn grass, the main requirement is that it be grown in good soil without the use of pesticides or chemical fertilizers; preferably fence off part of the lawn for this purpose. Often there is some foam on top of the juice, especially with grass juice. This foam is high in chlorophyll and beneficial, so ingest it or rub it into your skin.

I find wheat grass juice somewhat too sweet on its own and prefer it blended with other grass. As barley grass does not taste sweet, it may be better than wheat grass for individuals with blood sugar problems. Normally, however, it is good to mix the seeds of wheat and barley and grow them together in the same garden bed or tray.

How to Grow Grasses Indoors

1. Soak the wheat or barley for planting overnight and sow very densely.
2. Cover lightly with a sandy soil or just with wet newspaper.
3. Keep moist and expose to light after leaves emerge.
4. When the grass is about two inches high, you can start using it for cutting into a salad or you can chew the grass and spit out the pulp. You can begin to juice the grass when it is about four inches high. Sometimes you can get a good regrowth after the first cutting.
5. The maximum yield of juice is available when the grass is between six and eight inches tall.
6. The higher you intend to let it grow, the less dense you should sow the seeds.
7. Compost the stubble and other organic matter for reuse in seed boxes or outside planting.
8. Add rock dust, kelp, or seaweed to mineralize the soil.

Exposure of the trays to moderate sunshine is preferable but not essential. Protect the trays from prolonged strong sun exposure and from overheating behind a window. If the trays are on a balcony, you can move them inside during cold nights or cover them with newspaper. The best growing condition is at room temperature.

When the shoots are about an inch high, it is preferable to stop watering, or water along the sides but keep the centre dry to minimize problems with moulds and rotting.

In cool climates, it may be difficult growing grass outside in winter, while in the tropics it is sometimes difficult in summer because it is either too hot or too wet and cereal grasses easily start rotting at the stem. In the city, it may not be possible anyway. In these circumstances, wheat grass or barley grass can be grown indoors behind a window or on a balcony or veranda in seed boxes (see sidebar)

Wheat produces about four times its weight as wheat grass and this in turn yields up to 80 percent juice. Tougher perennial grasses yield about 60 percent juice. The vitamin and mineral content of grasses and their juices differs greatly between different varieties, but especially with diverse soil conditions. Table 4-1, compiled from various publications, gives an estimate of the average nutrient content of the juice from grass grown in good, mineral-rich soil.

The protein content of young annual grasses is usually around four percent and of established perennial grasses about two percent. Vitamin and mineral concentrations can vary greatly and are highest in young grasses grown in soils rich in minerals and organic matter.

While grass juice provides a high amount of easily digestible protein, as well as vitamins and minerals of superior quality, even more important are the enzymes,

Table 4-1: Nutrients in Grass Juice

Protein	2–5%	Fibre	3–6%
Lipids	0.5–1%	Kilojoules	250/100 ml
Carbohydrates	6–11%		
Vitamins/100 ml		***Minerals/100 ml***	
carotene	10,000 IU	calcium	70-200 mg
vitamin C	60 mg	magnesium	50 mg
vitamin E	6 mg	phosphorous	50-100 mg
vitamin B1	0.2 mg	potassium	400-1,600 mg
vitamin B2	0.4 mg	sodium	40-150 mg
vitamin B6	0.2 mg	iron	3 mg
vitamin B12	1 mcg	zinc	0.5-2.5 mg
pantothenic acid	0.4 mg	manganese	1-10 mg
nicotinamide	1.5 mg	copper	0.3-1 mg
folic acid	150 mcg	molybdenum	0.03 mg
biotin	15 mcg	sulphur	100 mg

growth hormones, and other vital factors, including coenzyme Q10 and superoxide dismutase (SOD). In animal experiments, old rats were rejuvenated and chickens grew 15 percent faster with fresh grass, but not with other supplements. Many patients attribute their cure from "incurable" diseases to the generous intake of grass juice. Grass juice has also been successfully used as a retention enema with cancer and other debilitating conditions, as a vaginal douche, or topical rub for the skin. Juicy grass pulp is excellent on wounds and speeds up healing.

Grass juice has been shown to inhibit mutagenic and carcinogenic activity. This inhibition was stronger with wheat grass juice than with carrot or parsley juice; it has also been reported with barley grass juice. We need growth hormones not only for growing taller, but also for cell division, skin and tissue replacement, and wound healing. In growing older, we become deficient in growth hormones, and those found in grass juice can help us rejuvenate.

Most animals live on grass either directly or indirectly. It is the perfect food, not only for grazing animals, but also for humans. The juicer replaces the complicated cellulose digestion of grass-eating animals. I regard grass juice as the food with the greatest healing and rejuvenating potential; it could even sustain us during a famine.

* * * * * * * * * * * *

Step 37

HEALING RECIPES

Transform healthy foods into enjoyable meals.

I am not fond of using recipes. Most recipes seem to be designed to tickle our palate by harming the nutrients in the food. The healthiest recipe for carrots, for instance, is to pull one out of the ground, clean it, and chew it well. Any additional processing renders it less nutritious. Also, I favour experimenting, mixing this and that within the framework of the food-combining rules, and seeing how it turns out. In this spirit I offer the following recipes and cooking tips as starting points for finding ways to make healthy food tasty.

If you are willing, you can gradually change your taste preferences and come to like this new diet of healthier meals. If your life is in no immediate danger from an advanced disease, change slowly, making a gradual transition over a period of years from your present diet to the high-quality diet, and possibly to the raw food diet after that.

Baking Breads: The best method for baking is one in which enzymes in the food remain alive. This means heating bread dough to less than 120° F. It is preferable to start with whole, soaked, or sprouted seeds that are rich in enzymes rather than with commercial flours that may have had their enzymes destroyed during the milling process and may be contaminated with mycotoxins. The only practical solution I have found so far is baking with rice. After blending soaked or sprouted rice, the dough continues to absorb water and so becomes firm almost without any heat. I have not found this property in any other grain. You may have to experiment with different varieties of rice to find a good sprouting one.

Soak brown rice overnight, then rinse for two or three days until sprouts appear; otherwise use after soaking. If you are sensitive to fungi, keep for several minutes in water with added hydrogen peroxide, then wash well and blend with a minimum of water. If the blended rice does not have the consistency of a paste, add rice flour or strain off excess water. Lightly cover a tray with some rice flour or baking paper and spread the paste out flat. Preferably leave in the sun or a warm place, such as a warm oven with the heat turned off, until the dough has solidified, usually after a few hours.

You can experiment with additions, such as kelp powder, occasionally carob powder and dried fruit for children, and acidophilus culture or a sourdough starter if you want to try baking a more conventionally shaped loaf. The addition of any other kind of soaked or sprouted seed will make it more difficult for the dough to harden. You can bake flat bread in an oven at low temperatures from a mixture of various flours or meals, such as from peas, lentils, chickpeas, rice, and rye. You may add buckwheat flour to any baking mix to improve its binding qualities.

If you use sprouted seeds, it is not necessary to add acidophilus or sourdough starter to improve the nutritional quality, but only to lighten the bake. However, if you only soak the seeds and, more important still, if your main ingredient is flour, then lactic-acid fermentation will greatly improve the nutritional value as well as lighten the texture.

Any other soaked, sprouted, and blended seeds can be used for making flat bread by baking at 160° to 180° F with or without a starter. Again, it is advisable to spread the dough over a layer of flour to absorb excess moisture. You can also add other flavouring ingredients, such as banana or carrot pulp. It may take five hours or more of baking for the bread to solidify. At this temperature, the enzymes are destroyed and, unlike sun-baked rice, it is not a raw food any more. However, the protein structures generally are not damaged and there is no digestive leukocytosis (increased white blood cells in the intestinal wall) when eating this bread. You can refrigerate part of the sourdough as starter for the next bake, but if you are yeast-sensitive, it is preferable to use fresh acidophilus culture each time.

For more conventional bread, I recommend rye sourdough; the more acidophilus you add and the slower the dough solidifies, the more acid it becomes, and vice versa. Mix 1 cup of acidophilus starter with rye flour, water, flavourings (for example, caraway seeds), and 1 tablespoon of honey or molasses as food for the bacteria. Leave covered overnight in a warm place. Before adding salt, reserve and refrigerate 1 cup of this as a starter for the next baking. Add more flour, knead, shape, and cover the loaves and let them rise in a lightly warmed oven for several more hours. Then bake at a moderate heat for 90 minutes; place a pan with hot water on the bottom rack to develop steam.

Bone Broth: Use the soft bones of fowl or the bones and heads of fish. Add 1 or several tablespoons of vinegar, depending on the amount of bones you have. Simmer with sufficient water in a covered, non-metal container for at least 3 hours, or until the bones become brittle and the liquid is nearly neutral. With larger quantities and longer cooking time, you can repeatedly add more water and vinegar.

Alternatively, use a pressure cooker for 30 minutes, but without adding vinegar. When the bones have become soft, blend it all, strain (optional), and freeze in ice cube trays. Use some of the broth frequently with meals; add it to vegetable salads, as it is an excellent source of gelatine, calcium, and other minerals.

Beef Juice, Liver Broth, Liver Juice: To make liver broth, simmer pieces of (organic) liver for 2 hours; strain and mix with sweet vegetables, cooked or raw. To make beef juice, dice a pound of lean beef. Put in a jar without water, cover well, and set the jar on a piece of cloth in a pot filled with water. Boil for 3 hours. Press juice accumulated in the jar through a strainer and refrigerate or freeze it. Sip a teaspoonful 5 to 10 times daily and keep it in the mouth for some time. Make this juice fresh weekly. Occasionally, the juice can be made of liver instead of beef. Beef juice is indicated in cases of serious muscle weakness.

Butter Spread: In order to cut down on your butter consumption, lightly warm some butter and mix it with an equal amount of extra-virgin olive oil. Add lecithin, chopped onion, kelp, herbs, and spices to taste. Alternatively, or in addition to using oil, butter can be mixed with an equal amount of hot gelatine; flavour to taste and keep the butter spread refrigerated.

Marinated Fish: Dice the fish and cover with lemon juice or diluted cider vinegar or a mixture of both. Refrigerate overnight; add cooked or raw onion or herbs and spices, and possibly some juice, leaf, or green skin of papaya. Eat with vegetables or sprout salad. You can also marinate liver, minced meat, or soft cuts of meat.

Fruit Balls: These are for festive occasions. Mince any of the following: nuts, sesame, sunflower, or pumpkin seeds, fresh coconut, dried fruits such as apricots, dates, mixed peel, papaya, and pineapple. Mix well, add lemon juice to taste and also lecithin; bind with oil. Make into small balls and roll in desiccated coconut. For different flavours, add carob powder or spices to the mixture.

Hommus: This can be used as bread spread or as an addition to meals. Soak chickpeas (garbanzo beans) overnight. If seeds are viable, sprout them; otherwise, use them soaked raw or soaked and cooked for a few minutes only. Discard the soaking water to reduce intestinal gas. Puree the prepared chickpeas in a blender and mix with any combination of the following: olive oil, tahini, lecithin, cayenne, kelp, herbs, and spices. Keep refrigerated.

Hot Vegetable Juice: Normally, you drink fresh vegetable juice cold. However, in cold weather you may enjoy drinking it hot, flavoured like a broth. Use a handful of fresh green leaves, add cabbage, celery, tomato, cucumber, whatever is available, and some sliced carrot, pumpkin, or beet root. Mix this in an electric blender, together with a suitable hot liquid, for instance, herb tea, bone broth, or water. Strain and press the residue. Try to keep the temperature of the broth below 120° F.

A juice extractor can be used instead of the blender or you can mix the hot liquid with some freshly pressed juice. Flavour the drink to taste; you can use herbs, spices, miso, kelp, oil, lecithin, egg yolk, food yeast, or molasses in any combination you like. Drink the juice immediately, taking sips. Another possibility is to simmer the residue left over from juicing in water for ten minutes, strain, add some flavouring, and drink hot.

Jellies: Dissolve 4 teaspoons of white, unflavoured gelatine in 1 pint of hot water. Pour it over diced fruits or over sprouted seeds and diced or grated vegetables (for example, cucumber, tomato, carrot, or chopped onion). You can add herbs, spices, kelp, and salt. Alternatively, the gelatine can be dissolved in a smaller amount of hot water and mixed with an appropriate amount of fruit juice or fresh vegetable juice. Refrigerate for setting. Instead of commercial gelatine, a gelatinous bone or fish broth is preferable. Gelatine aids in the absorption of vitamins and minerals.

Linseed Yogurt: The following combination can be used as a snack before or between meals or instead of any meal. It is especially good as a breakfast. Mix half a cup of ground linseed with 1 or 2 tablespoons of linseed oil or olive oil or melted coconut oil, 1 tablespoon of bee pollen, 1 teaspoon of spirulina, sliced or blended bananas, lecithin granules, and enough yogurt, seed cheese, or seed milk for a smooth consistency. You can also add some chopped fruit or berries. As a liquid component, try fresh vegetable juice, grape juice, apple juice, or any other liquid.

Papaya (Pawpaw) Smoothie: This can be used as a special health food to aid the digestion, as well as in dissolving tumours or other unwanted growths. In a blender, mix mature green papaya (when it just starts turning yellow and the seeds are already black) with skin, seeds, and flesh, banana and any other fruit in season, and a sufficient amount of a suitable liquid such as a juice or yogurt. Eat on its own or as part of a meal.

Grated Potatoes: Bring 1 cup of water to a boil, keep the heat on high, add coarsely grated potato, and stir for 2 to 3 minutes. This leaves the potato semi-raw with a quite distinct flavour; add kelp, oil, and so on, and eat with vegetables or sprouts.

Protein Drink: This is good as a general energy booster, especially if you have blood sugar problems or lack of energy, and it is also excellent when consumed 30 minutes before a meal to reduce appetite if you want to lose weight. Mix in a blender 1 or more spoonfuls each of bee pollen, green barley powder, ground linseed, spirulina, and chlorella, or any combination of these in a suitable liquid. This can be seed milk or any other kind of milk or juice such as grape or apple juice.

Rejuvelac (fermented seed drink): Wash 1 cup of whole grain (preferably organically grown) and cover with 2 cups of warm water. Suitable are rice, millet, rye, and other grains. Keep in a glass or porcelain container in a warm place. Pour off the liquid the next day or when it tastes slightly sour. Use as a refreshing drink on its own, but not with or after meals. The grains can then be cooked or sprouted. Use the grain ferment only if it has a pleasant taste and smell, otherwise discard it. Rejuvelac may not be suitable for sensitive and yeast-allergic individuals.

Rice Dishes: Cook the rice until almost soft and most of the water has evaporated. Add a small quantity of apples, cover, cook until the apples are soft, and then mash them. Add cinnamon, oil, lecithin, and kelp. Eat cold as a dessert. As an alternative, add apple puree to the cooked rice. You can also try rice with a sauce based on blended raw carrots and other sweet vegetables or bananas.

Healthier than cooked rice is sprouted and blended rice. Wash and soak overnight a sprouting variety of brown rice. After sprouting, blend the rice and add other flavouring ingredients, such as banana, olive oil, kelp, carob, lecithin, or pitted dates, for a sweet meal. Blend again and eat this instead of a breakfast cereal. If too gritty, strain the rice puree after the first blending or cook it.

Sauerkraut: Use a wooden barrel or earthenware pot of any size. Place a layer of shredded cabbage 4 to 6 inches deep in the container. Sprinkle over it a small amount of salt and some herb seeds, such as caraway, fennel, or cumin; other shredded vegetables can be added for flavouring. Press the first layer down, then add another layer of cabbage and herbs, and so on. The cabbage must be completely saturated with its juice and no air pockets left. Cover the contents with cheesecloth, place a wooden cover over it, and weight it down with a heavy stone. Leave at room temperature.

From time to time, after several days, lift the cover and remove foam and mildew from the top of the mash; wash the cheesecloth, board, and stone with warm water and then put them back in place. After about 2 weeks the sauerkraut should be ready for eating. Store the container in a cool place, or put the sauerkraut in jars and refrigerate. Eat it raw and drink the juice. Raw sauerkraut may not be suitable for sensitive individuals, but it should be fine if cooked.

This recipe may occasionally fail if the barrel is contaminated. To avoid this, thoroughly clean the barrel with steam or boiling water before use. To make it easier for beneficial bacteria to develop, sprinkle organic cider vinegar into the different layers or add acidophilus culture.

Seed Cheese/Seed Yogurt: Soak oily seeds (almonds, nuts, sesame, pumpkin, or sunflower) for 8 to 12 hours. Puree in an electric blender and add acidophilus culture. If you are not yeast sensitive, experiment with adding Rejuvelac instead. Keep in a warm place for several hours until the desired degree of sourness develops. Refrigerate and use within 3 days. If it is too sour or if curd and whey have separated, strain and discard the whey; rinse the curd if necessary. Use seed yogurt as part of a salad dressing, for flavouring meals, or as a bread spread. The more sensitive you are, the less sour it should be when you use it. If it smells or tastes bad, discard it.

Seed Milk: Soak almonds, rice, or sunflower kernels overnight or for about 12 hours. The simplest way to make this milk is to change the water, blend the soaked seeds in an electric blender, and press them through a strainer. You can either drink the liquid immediately or refrigerate. You can cook the residue of the rice and add the residue of the oily seeds to any breakfast mix (possibly remove almond skins before blending). However, a much better way is to wait until the seeds start sprouting. This removes any enzyme inhibitors and provides natural sweetness and enzymes to the milk. Alternatively, use unhulled sunflower or pumpkin seeds for making milk.

Sprout Salad: Mix a variety of freshly rinsed sprouted seeds with a combination of fresh, raw vegetables, basically using whatever is available. Most suitable are sprouts of mung beans, lentils, sunflower seeds, and fenugreek, together with finely grated beet root, carrot, and turnip or radish. Tomato and cucumber (try it grated) are good for flavouring. If you have difficulty chewing, you can put all of it through a mincer, or if very weak and debilitated, you could liquefy and drink it.

The key to enjoying a sprout salad is to find a delicious dressing. Experiment. I recommend extra-virgin olive oil, lemon juice, whole blended lemon, cider vinegar, or rose-hip powder, any fresh or dried herbs or spices, and a dash of cayenne. These may be added individually or mixed beforehand and refrigerated in a jar. Flavour olive oil by mixing in a jar 1 part tahini with 3 to 10 parts oil, lemon juice, and lecithin.

For a health enhancer, add ground linseed and kelp to the salad. Start by adding very small amounts of kelp until you come to like it. Raw egg yolk is a good addition to the dressing, as is seed cheese or sour milk; you can also flavour this salad with tofu or yogurt.

Sweet Vegetables: Use any combination of the following: sliced pumpkin or squash, sweet potato, onion, turnip, beet, carrot, and tomato. Add only a minimum of water so that finally most of it has evaporated. Vegetables with short cooking times (tomato, pumpkin) can be added later to preserve their flavour. Salt, kelp, oil, curry, cayenne, herbs, and spices are best stirred in at the end of cooking.

Yogurt, Clabber Milk, Cottage Cheese: It is essential to repopulate your intestines with beneficial lactobacilli, especially L. acidophilus in the small intestines and bifidobacteria in the large intestines. Instead of or in addition to using high-potency acidophilus/bifido capsules or powders, you can use a self-made fermented product or yogurt. This can be made from fresh unpasteurized cow's milk, goat's milk, rice milk, homemade soy milk, or a mixture of ingredients based on bee pollen and honey. Commercial cow's milk acidophilus/ bifido yogurt (preferably organic or biodynamic) can be used as a starter.

The commonly used yogurt based on cow's milk has some problems in that many individuals are allergic or sensitive to some ingredients of cow's milk. It often is mucus-forming and is also high in estrogen, which is not good for various female cancers. According to the blood group diet, cow's milk is acceptable only for those with blood group B and possibly AB. However, if you are not sensitive to it, then yogurt made from fresh raw cow's milk can be beneficial, used in moderation.

You can use rice as a base for making yogurt. Cook soaked (organic) brown rice in plenty of water. After cooling, blend and strain; add starter culture and 2 teaspoons of raw honey to the strained rice water. Rice yogurt does not set, so drink it when it turns slightly sour or smells slightly fermented. You can make rice milk by blending and straining sprouted brown rice. This is sweet enough by itself, so that no added sweetener is required.

A fermented drink can be made based on bee pollen and honey. Add to 1 quart of water and starter 5 to 10 teaspoons of pollen, 2 or 3 teaspoons of raw honey, 1 or 2 teaspoons of kelp powder, and, if you like, several teaspoons of spirulina, chlorella, or cereal grass powder. Experiment and use more or less of the indicated amounts and also different ingredients. As with rice milk, this does not set and is best used when it starts frothing and tastes somewhat acid.

As an initial starter, use 1 teaspoon each of powdered acidophilus and bifido cultures, a mixed culture of these, high-potency capsules, or a commercial yogurt containing these. Do not use commercial yogurt or starter with only bulgaricus and thermophilus bacteria. Subsequently, use up to 1 cup of the previous batch as starter. Preferably, use the liquid or whey as the starter because it contains the most bacteria. The more of the starter you use and the warmer it is, the sooner the yogurt will be ready.

If you do not have a yogurt maker, mix 2 to 3 cups of any type of milk or pollen-honey liquid with 1 cup of starter and keep it in a jar standing in a container with hot tap water. It usually takes only a few hours to set, though with powdered cultures it may take 6 to 10 hours. In cold weather, renew the hot water every hour or keep the jar in a warm place. Alternatively, the milk can be kept warm by putting it in a box or some other small, enclosed space, together with an electric bulb or other source of heat. Refrigerate the yogurt when it just starts setting, as it continues to become firmer and more acid during storage. If it becomes too sour, use less of the whey as starter and eat only the strained curd or neutralize the acid.

To make sour milk (clabber milk), leave raw, unheated milk in a flat bowl in a warm place for 1 to 2 days until the milk sours and coagulates, then add a yogurt or acidophilus starter. For making cottage cheese or quark, wait until curd and whey have separated and strain or press the curdled milk through cheesecloth.

Veggie Burgers: Soak overnight 1 cup of chickpeas or lentils; next morning replace the water and put them in a blender. Soak 2 cups of rice overnight and cook. Combine the blended legumes with the cooked rice and add some buckwheat flour or an egg to bind the mixture. Flavour with any combination of the following: miso, soy sauce, fresh parsley, coriander, cumin, fresh ginger, onion, and any other herbs or spices. Form flat burgers and bake crisp in a grill or a non-stick pan.

* * * * * * * * * * * *

Part 5

PROBLEM FOODS & FOOD PROBLEMS

This part explains in more detail the reasons for the dietary recommendations in Part 4. Most problem foods are widely used basic foods that lead to health problems in many people. So-called junk foods and heavily processed and chemicalised foods belong in this category; many foods treasured by natural health enthusiasts, such as honey, dried fruits, yogurt, and whole-meal bread, also belong to this list of problem foods.

What chiefly makes a particular food a problem is the presence of one or several components that can cause metabolic or digestive problems if used indiscriminately. We use and have used problem foods in excess, so our metabolism has been weakened to such an extent that it cannot safely process even small amounts of certain foods without becoming distressed. This can be demonstrated with food muscle-testing.

However, for individuals with a suitable metabolism, most problem foods can be acceptable and sometimes even beneficial in moderation; this does not apply to heavily chemicalised food. It is advisable for nearly everyone to minimize their intake of problem foods. It is also best to more strictly avoid the problem foods described in this chapter and that you know from personal experience are a cause of your health problems.

The main problem foods are:

Lactose and cow's milk products

Gluten and wheat products

Sugar and sweet foods

Red meat and fats

Chemicalised food, stimulants, and yeast or moulds

Some problem foods frequently cause food allergies and can be the primary allergens or causative agents that sensitize us to a range of other allergens. This applies especially to the gluten in wheat and to proteins in cow's milk. When the body is still young or overacid and sensitive, it tends to respond strongly to these foods. However, problem foods cause even more serious problems when the body is older, alkaline, and insensitive, but then there are no alarm reactions such as food allergies to warn us. This leads to the development of chronic degenerative diseases such as autoimmune diseases, cardiovascular diseases, and cancer.

In addition to specific health problems being caused by specific foods, the overall impact on the adrenal glands of habitually consuming problem foods can raise our basic stress level to such a high degree that a relatively small additional stress from emotional or social problems can trigger severe reactions. These can include asthma, epilepsy, migraine, depression, irritability, hyperactivity, phobias, nervous breakdown, and mental disease. In addition, the fact that we cook too many of our foods and hence have insufficient enzymes in our diet greatly contributes to the development of chronic degenerative diseases and premature aging.

* * * * * * * * * * * *

Step 38

COW'S MILK PRODUCTS & LACTOSE

Have a closer look at cow's milk products and lactose. In some forms they can be beneficial, but as commonly used they tend to create a lot of health problems.

The three problematic ingredients in dairy products are lactose (milk sugar), butterfat, and casein (and other proteins). While lactose can be a problem in all animal milk, including goat's milk, difficulties with protein and butterfat are specific to cow's milk.

Bottle-feeding infants with cow's milk has far-reaching effects. The earlier it is substituted for human breast milk, the more damage is usually caused. The baby's digestive system is still immature and relies on enzymes provided in mother's milk. It cannot properly digest cow's milk, especially if the milk has been pasteurized and is without enzymes. In addition, in the first few weeks or months of life, the wall of the infant's small intestine is not yet fully developed and allows only partly digested proteins to pass through; this causes allergies.

In a clinical study, all infants and most older children evaluated had antibodies against cow's milk in their blood.24 This means they were allergic to it, even in the absence of obvious symptoms. However, usually "unspecific" symptoms were present, such as restlessness and crying at night, dermatitis, tender abdomen, tantrums, weak eyes, low energy, hyperactivity, indigestion, and a high incidence of colds and ear and respiratory infections. A contributing factor that makes many babies prone to infections is the absence of immuno-protective agents in bottled milk that are present in human breast milk, especially in the colostrum.

Formula-fed babies usually suffer from zinc deficiency. Zinc is essential for activating the immune system. The zinc content in cow's milk is actually higher than that in mother's milk, but in cow's milk zinc is bound to a protein from which the baby's immature digestive system cannot release it. Other trace minerals are also difficult for the baby to absorb from cow's milk; iron is especially problematic. A resulting iron deficiency in babies contributes to the development of anemia, a weakening of the immune system, and retarded mental and cognitive development.

An allergy to cow's milk and subsequent mucus congestion of the lungs, combined with zinc and iron deficiencies of the immune system, cause frequent colds and respiratory infections in babies. This, in turn, depletes babies of vitamin C. The effect of all this is a high incidence of crib deaths in bottle-fed babies, which sometimes occur shortly after immunizations. Vaccinations can further drain the already dangerously low levels of vitamin C. Archie Kalokerinos, M.D., in Every Second Child, relates that in some Aboriginal communities of Australia every second child given a vaccination died of crib death, but when fed high doses of vitamin C before and after vaccinations, not a single child died. Is it a coincidence that New Zealand has the highest rates of asthma and crib deaths in the world, but also the highest consumption of cow's milk?

Crib deaths are uncommon in breast-fed babies. However, even breast-fed babies can develop allergies if the mother has a high intake of cow's milk products or is allergic to it herself. Full-term babies in the first two weeks and premature babies in the first one or two months can become allergic to almost any substitute for mother's milk. If breast-feeding by the mother is not possible during this time and a wet nurse is not available, then fresh raw goat's or sheep's milk is the least harmful alternative.

Problems Associated with Lactose: Most adults, except for Caucasians, cannot split lactose into its two components, glucose and galactose. After the age of three, a deficiency in the lactose-splitting enzyme lactase develops, and this can cause severe indigestion and diarrhea if the diet contains appreciable amounts of lactose. This condition is known as lactose intolerance.

Lactose-intolerant people can tolerate lactose better if fermented milk is used, as in the form of yogurt or kefir, where the lactose is partially pre-split by lactic-acid bacteria. Alterna-tively, lactose-splitting enzymes are now commercially available and can be added to milk. However, lactose intolerance is only a minor problem compared to the much more serious health problems caused by galactose. Lactose intolerance appears to be a wise precaution of nature rather than a regrettable accident because it protects us from the great danger of galactose overload.

We both need and don't need galactose. Most European adults and older children who can digest lactose are unable to use galactose efficiently. Babies need galactose as an important building component of the brain, the central nervous system, and of many proteins. Thus, mother's milk is even higher in galactose than animal milk to ensure that the baby obtains sufficient galactose. In later life, very little galactose is needed and this can easily be synthesized from other sugars. Therefore, most of the ingested galactose is converted in the liver to glucose and used as body fuel, but the amount that can be converted is rather limited, even with a healthy liver.

This conversion is a slow and complex process requiring four different enzymes. One of these is sometimes missing from birth, giving rise to a condition known as galactosaemia. Continued milk-feeding leads to a buildup of galactose in the baby and can cause cataracts, cirrhosis of the liver and spleen, and mental retardation. If the liver is not healthy, it becomes less able to convert galactose. This fact is sometimes used as a criterion for a clinical liver-function test. If galactose is injected into someone with a defective liver, most of the galactose will later appear in the urine, confirming the diagnosis of liver dysfunction.

Mucic Acid and Mucus: Under normal conditions, only part of the galactose is expelled with the urine. If there is a deficiency of protective antioxidants, then the rest is mainly oxidized to galactaric acid, commonly known as mucic acid. The great health danger of mucic acid is that it is insoluble. The body cannot let it pile up in vital areas and block organ functions or blood circulation. Therefore, it forms the mucic acid into a sticky suspension in water called mucus. Thus, mucic acid is a main component of pathogenic -(disease-producing) mucus.

It is the function of the lymphatic system to remove dangerous substances, such as mucus, from areas of vital importance and transport them to the organs of elimination. Mucus is too dangerous to dispose of through the kidneys or (with bile) through the liver, but it has a special affinity to the mucous membranes that line the insides of our body openings. Of prime importance here are the lungs, the respiratory tract, the hollow head spaces, the sinuses, and the Eustachian tube (a passage between the mouth and the inner ear).

The mucus accumulates in these hollow spaces until external factors help to sensitize the mucous membranes sufficiently to allow the mucus to pass through. This is relatively easy in young individuals and those with a poor sugar metabolism, as they have high levels of histamine and inflammatory adrenal hormones. Even minor irritations of the mucous membranes, be they from cold air, dust, air pollution, pollen, or germs, will sensitize these to let some of the mucus flow out, such as through sneezing.

Such mucus cleansing may be experienced periodically as a cold, hay fever, wet cough, or running nose. In others, the accumulation of mucus, which provides a favourable breeding ground for germs, causes chronic infections in specific areas such as the sinuses, the middle ear, the respiratory tract, and the lungs. This may allow a permanent trickle of mucus through the affected mucous membrane. In addition, a dead (commonly root-canalled) front tooth may be responsible for chronic sinus problems.

With a high lactose intake, the lymph channels and lymph glands are usually congested with mucus as well. This allows influenza and other infections to spread from the sensitized mucous membranes through the mucus-filled hollow spaces into the lymphatic system, causing lymph gland swellings and inflammations. I have found that in many people the number of influenza bouts, colds, and other respiratory infections can be varied at will from none to several per year just by varying the lactose intake. Mucus congestion is also the main cause of ear infections and hearing problems, especially in children.

In most cases it is not a lactose allergy, but a galactose overload that is responsible for this excessive mucus. While in the case of those suffering from cow's milk allergy, somewhat more lactose may be tolerated when it comes from goat's milk, in most individuals the lactose in goat's milk or in tablets will be as mucus-forming as that from cow's milk.

Milk Products and Asthma: When more mucus accumulates in the lungs than can be expelled, asthma is likely to develop. Often lung congestion is combined with a strong subconscious fear element that can, for instance, result from insecurity or lack of love in early childhood. Another contributing factor is hypoglycaemia coupled with weak adrenal glands.

Many cases seem to be predominantly mucus-induced. I treated a patient who was fond of yogurt and, for health reasons, prepared it from skim-milk powder. This produces yogurt with a much higher lactose content than yogurt from full-fat milk (see table 5-1). When I persuaded her to use somewhat less yogurt and prepare it only from whole milk without additional skim-milk powder, her asthma disappeared for good. The asthma-causing skim-milk yogurt provided approximately 50 g of lactose per day, while she was asthma-free on whole-milk yogurt with about 5 g of lactose daily.

The lung irritation caused by accumulated mucus also means that the lungs are more prone to be affected by food allergies and chemicals. This could result in inflammatory swellings of the bronchial tubes. Mucus accumulating in the lungs allows bacteria to infiltrate. Some strains of these bacteria convert sugars into alginic acid, another sticky mucus. Often there is Candida albicans or fungal infestation as well, which sensitizes the mucous membranes to airborne moulds. Mucus--releasing colds in this setting can be a blessing in disguise, provided they are not treated with antibiotics. Also dead teeth can induce copious mucus due to chronic root and jawbone infections, as explained in step 8.

The mucous membranes of asthmatics, which are highly sensitized by mucic acid, react strongly to air pollutants such as smoke, pollen, and sulphur dioxide. The adrenal glands are typically weak and histamine levels are high because of a sweet diet and allergies. To overcome asthma, we need to reverse these negative conditions by using a low-allergy diet with a minimum of sweet or mucus-forming food, while the respiratory tract should periodically be cleared of mucus. Breathing exercises also help this condition.

Milk Products and Chronic Degenera-tive Diseases: A serious consequence of a congested lymphatic system full of mucus is the development of leukaemia. This happens when the immune system has been sufficiently damaged by frequent mucus-related infections combined with other factors such as toxic chemicals, a sweet diet, allergies, and -vitamin--mineral deficiencies.

It may not be a coincidence that Nathan Pritikin, famous for his much-publicized diet to cure cardiovascular diseases, developed leukaemia. The original Pritikin diet is high in skim-milk products and, therefore, imposes a severe galactose overload on the body. Leukaemia or another galactose-related degenerative disease is more likely to develop as a long-term effect of a high intake of skim-milk products. It is important to note that cardiovascular diseases can be prevented or cured nutritionally without causing other health problems such as leukaemia.

Table 5-1: Lactose Content of Dairy Products

butter	0.5%
cheese, cottage cheese	2-4%
goat's milk	4.3%
cow's milk	4.9%
yogurt and ice cream (with skim milk powder)	5-25%
skim-milk powder	52%
whey powder	70%

Leukaemia stands in between the acute mucus-related infections of childhood and the usual chronic degenerative diseases that develop with advancing age. When our metabolism slows down as we become older or if we are on an unsuitably heavy meat diet, the body gradually becomes too alkaline and the mucous membranes become more insensitive. In this condition, the release of mucus through colds and other respiratory infections becomes rare and most of the mucic acid is stored in the body.

A lactose intolerance, and thus an enforced avoidance of foods containing lactose, can also protect us from another disease: cataracts of the eyes. Even infants can develop cataracts when they cannot convert galactose to glucose, and galactose overload is an important cause of cataracts in adults. In addition, high blood glucose and fructose levels may contribute. These simple sugars are reduced to sugar alcohols that cause cloudiness in the lens. Another form of cataract is mainly caused by a chronic deficiency of vitamins A, B2, C, D, and E, and the minerals chromium and selenium. Radiation exposure or drugs can also cause cataracts.

Besides cataracts, there are other diseases that are usually considered to be typical for the aging body, but that occur in infants with galactosaemia. These include liver degeneration, oedema, and reduced memory or senility (the latter being equivalent to mental retardation in infants with galactosaemia).

Cancer (carcinoma) reportedly can result from galactosaemia. A clinical study shows that women who consumed yogurt had a higher rate of ovarian cancer than controls who consumed the same amount of lactose from other milk products.25 Because of the activity of the lactic-acid bacteria, yogurt contains more readily available galactose than other milk products.

A frequent complaint is increasing deafness with subsequent infection and inflammation because of mucus congestion of the Eustachian tube and the middle ear. In children, this condition is sometimes called “glue ear.” This is especially a problem with children of non-European background because they can still absorb lactose, but cannot easily convert galactose, especially from cow’s milk, into energy.

One Scandinavian study showed that no purely breast-fed baby developed early middle ear infection and that such babies were protected against it in later life. In contrast, early introduction of cow’s milk predisposed children to this and it was exclusively found in children who received cow’s milk before the age of six months.[26]

With lactose-induced mucus congestion, degenerative lung diseases such as emphysema can also develop. While smoking is generally considered to be the greatest hazard for lung cancer, it may actually rank equal with galactose overload; most at risk are heavy smokers with mucus-congested lungs. Sometimes the lungs fill up with mucus. A patient I know died because his lungs and breathing passage were filled with sticky mucus. He literally drowned in it. With each of his breaths, I could hear the air bubbling up through the mucus.

A combination of mucus accumulation in the lungs and the digestive system is seen in cystic fibrosis. This disease is mainly due to an overproduction of an abnormal mucopolysaccharide, a long-chain carbohydrate that normally supplies the mucus physiologically required by the body. Cystic fibrosis sufferers also may be unable to convert galactose and, as they are usually deficient in protective antioxidants, they may also produce large amounts of mucic acid. Like galactose overload, cystic fibrosis is typically a disease of Caucasians.

In part, excessive mucus formation stems from an infestation of the lungs with bacteria, which produces an abnormal amount of sticky alginic acid. The important point here is that alginic acid is synthesized from a simple sugar (mannose) that these bacteria can convert from an excess of any other sugar, such as galactose, fructose, or glucose. Therefore, mucus-forming as well as sweet foods must be avoided.

Lactose in Food: Preventing excessive mucus accumulation in the body is much easier than trying to remove it afterwards. If you are concerned about your future well-being, it is a wise precaution to reduce your intake of lactose to a minimum, starting today. Review table 5-1 for the lactose content of some common dairy products, and see which ones you can avoid.

With skim-milk powder having a lactose content of 52 percent, you may now realize how dangerous is the current fad for using low-fat ice cream, yogurt, cottage cheese, and so forth, instead of full-fat products. Such low-fat foods are made from skim-milk powder and contain three to five times as much lactose as the equivalent full-fat foods. Sometimes skim-milk powder is even added to butter. Therefore, read the product labels and avoid butter that lists “non-fat milk solids” as one of the ingredients.

Skim-milk powder is also a favourite additive to many other commercial foods, such as bread and other baked products, sausages, and margarine. The health-food industry is fond of adding lactose to many products such as soy milk and dandelion coffee. Lactose is often used as filler in white tablets; cell salts are almost pure lactose. Try to avoid white tablets if the label does not state that they are free of lactose or are low-allergy tablets.

I suspect that the average daily amount of lactose that healthy adults can handle without the danger of long-term galactose overload is less than 10 g, the equivalent of one glass of milk. However, those who have occasional mucus problems or are afflicted with a -galactose-related disease do well to have a much lower lactose intake than 10 g per day. With a lactose allergy, it is often necessary to avoid lactose completely for several months or years.

Casein, Milk Protein, and Fat: Casein is the main protein in cow’s milk and constitutes about three percent of it. Mother’s milk, on the other hand, has only 0.5 percent casein content. The high casein content of cow’s milk causes it to form a very tough, rubbery curd in the stomach; the casein then binds the calcium in milk as an insoluble salt. Thus it is extremely difficult to digest and is a frequent source of indigestion. Mother’s milk and goat’s milk, on the other hand, form finely dispersed soft curds that infants can easily digest.

Thus the protein in cow’s milk frequently is only partly digested and becomes a major source of intestinal putrefaction and toxaemia. Incompletely digested protein can pass through the wall of the small intestine and cause allergies. Symptoms of worms in children are often due to intestinal putrefaction from undigested cow’s milk. Breast milk, which contains high levels of fat-digesting lipase and other enzymes, is nearly self-digesting in the baby’s gut in contrast with pasteurized cow’s milk, which is very difficult to digest.

In a double-blind study, 24 out of 27 babies with colic became free of symptoms when put on a diet free of cow’s milk protein. The babies, when given cow’s milk protein, cried on average 3.2 hours daily and when taken off cow’s milk, cried for only one hour.[27] However, I believe that breast-fed babies who have all their needs met hardly cry at all. Allergy-prone mothers pass allergens to their babies with their breast milk. This allergenicity is often caused by beta-lactoglobulin, which is in the cow’s milk. When such mothers avoided cow’s milk (and thus its proteins), the babies’ colic disappeared.

Colic is due to an inflammation of the intestinal wall, which in turn is triggered by an allergic reaction against the protein in cow’s milk. This chronic inflammation erodes the intestines’ microvilli through which the food is absorbed, thus resulting in malabsorption. Incompletely digested proteins can also pass through the damaged intestinal wall into the bloodstream and produce various allergic reactions, such as dermatitis or brain irritation.

A general consequence of such cow's milk allergy is a weakening of the immune system, which in babies is further aggravated by a lack of protective immune factors that are normally transmitted through breast milk. This is why infants on cow's milk have frequent colds and respiratory infections; this immune weakening can also lead to sudden infant death or immune deficiency diseases in later life.

Cow's milk appears to disturb calcium metabolism. Calcium becomes trapped in undigested casein, while the long-chain saturated fatty acids form insoluble soaps with calcium. In addition, cow's milk has an unfavourable ratio of phosphorus to calcium as compared to breast milk.

Several studies show that cancer patients tend to consume more cow's milk than do matched controls. Professor Jane Plant, author of Your Life in Your Hands, may have discovered the connection between cow's milk and cancer. Her breast cancer had spread throughout her body and she had been given up to die. At that stage, she found out that the rate of death from breast cancer in China is one in 10,000 compared to about one in ten in most Western countries, and also that Chinese do not use animal milk or related products. Putting the two together, she avoided all milk products; her tumours eventually disappeared, and for 13 years now she has been free of cancer.

As Chinese have "normal" rates of other cancers, there must be a special factor that causes their low breast cancer rates. Also the largely milk-free Japanese have low breast cancer rates, but when Chinese or Japanese women adopt a Western lifestyle, their breast cancer rates start to climb, approaching the level in Western countries. Studies now link milk consumption to breast and prostate cancer.

The problem appears to be a special protein called insulin-like growth factor, IGF-1. This stimulates hormone-related division of cells and especially stimulates breast tissue to grow during puberty and pregnancy; it also affects the prostate gland. Clinical studies showed that higher levels of circulating IGF-1 in the blood were not only a strong risk factor for women to develop breast cancer, but also for men to develop prostate cancer. However, IGF-1 levels were not elevated with benign conditions.28 The crucial factor is that cow's milk is very high in IGF-1, and it is also present in the meat of dairy cows. High-yield milk production, as commonly induced with synthetic hormones, increases the IGF-1 levels in the milk.

Another protein or protein fragment is connected with juvenile diabetes (Type I). It appears that this is only a problem with milk from Frisian cows (called A1 milk), but not with milk from other, lower-yielding breeds that produce A2 milk. Most of presently consumed milk is A1 milk.

Juvenile diabetes is much higher in those who have been bottle-fed rather than breast-fed and it is lower in communities that consume less cow's milk products. Studies of newly diagnosed diabetic children revealed an immune response to this protein fragment of cow's milk protein in all of them, and found that it has the same composition as one called P69 on the beta cells. P69 is usually protected inside the pancreatic beta cells and comes to the surface only during microbial and viral infections. At those times, the immune system can mistake it for cow's milk protein and attack it and destroy the beta cells in the process. Bottle-fed infants are very susceptible to colds and respiratory and gastrointestinal infections; it is regarded as “normal” for them to have six and more infections a year, even though these are rare with breast-fed infants.

Therefore, bottle-fed infants frequently receive antibiotics that then encourage overgrowth of the intestines with undesirable microbes and a tendency towards chronic pancreatitis. One type of E. coli bacteria, though harmless in the large intestine, has the potential for causing great damage in the small intestine. That is because it produces a molecule very similar to insulin. When the immune system becomes activated against this molecule, it can then also direct its attack against related features at the beta cells.

This shows that a combination of two factors can trigger an attack on the insulin producing beta cells: one factor that brings P69 to the surface of the beta cells, and another that activates the immune system to attack them. From the first factor we have the frequent colds of bottle-fed babies, and if antibiotics are used, then this also promotes overgrowth of the small intestines with pathogenic microbes, including E. coli, as the second factor. This is reinforced with feeding cow’s milk, which intensifies the attack on the beta cells and eventually leads to juvenile diabetes.

Fermented cow’s milk products in the form of natural yogurt, kefir, and natural cheeses produce a fine curd and cause fewer health problems than unfermented milk. However, even these should be used with care and only if they do not cause allergies or mucus problems. Generally, the corresponding goat’s or sheep’s milk products are far more preferable. The least harmful milk product is cottage cheese made from fermented goat’s or sheep’s milk.

The most damaging aspect of commercial milk products, apart from oversupplying the body with growth promoters, lactose, and casein, is the destruction of natural enzymes by pasteurizing the milk. Raw butter, for instance, was formerly used to cure psoriasis, but pasteurized butter causes or aggravates it. The healing effect of raw butter is due to its high content of the fat-digesting enzyme lipase. The same is true for heart and liver problems, which are caused in part or aggravated by processed cheese and butter fat.

Such health problems did not occur to the inhabitants of the Caucasus and Bulgaria with their high intake of raw milk products. Cholesterol did not harm anyone in former times when mainly unheated milk products were used; cardiovascular diseases were almost unknown. Raw milk was formerly used to cure tuberculosis, but pasteurized milk is more likely to cause it. In general, raw butter and cream are far healthier foods than refined polyunsaturated oils or margarine.

* * * * * * * * * * * *

Step 39

WHEAT & GLUTEN

Wheat may be the staff of life, but in many people it is actually unfavourable to health.

Gluten is a mixture of two groups of proteins, gliadins and glutelins. Wheat has the highest content of gluten, especially hard wheat, which makes it possible to bake leavened bread and cakes. Under the conditions of baking, gluten forms a network of molecules, similar to a wire mesh. This molecule mesh traps small bubbles of carbon dioxide gas and prevents them from escaping. This makes the baked product light and easy to chew. However, in this way we create a problem for our digestion because the gluten network is more difficult to break down. The gluten network is only partly digested, especially if the food is not very well chewed; this is a main cause of intestinal inflammation and wheat allergy.

Malabsorption Problems with Wheat: Gluten seriously weakens the intestinal wall. Its effect on the tiny absorption villi in the small intestine can be compared to the action of sandpaper on wood. Animal experiments have shown that the intestinal absorption villi are long and slender before they come into repeated contact with wheat protein. Afterwards, they become blunt and broad, with a much-reduced ability to absorb nutrients.

Therefore, people on wheat diets absorb nutrients less well than those reared on wheat-free diets. The former greatly contributes to the widespread incidence of malabsorption and nutrient deficiencies. In such people, not only are the absorption villi blunted, but the irritation caused by the sandpaper effect of gluten produces a protective mucous coating over the intestinal wall. This makes it still more difficult for nutrients to pass through the intestinal wall to be absorbed.

Thus we find gluten, and especially wheat gluten, implicated as a cause of typical malabsorption diseases, such as cystic fibrosis, celiac disease, and sprue, but frequently also associated with autoimmune diseases, diabetes, arthritis, multiple sclerosis, and schizophrenia. The irritation caused by gluten is a main factor in causing appendicitis, colitis, and inflammation of the small intestines (e.g., Crohn's disease) as well as gastric and duodenal ulcers.

It has been estimated that about 90 million Americans suffer from gluten sensitivity, while celiac disease, a severe form of malabsorption caused by gluten, has been found to be 50 times more prevalent than previously suspected and about 1.5 million Americans are thought to suffer from it. Celiac disease occurs frequently not only in patients with gastrointestinal symptoms, but also in relatives and others with numerous common disorders even in the absence of gastrointestinal symptoms.[29] For further information, contact University of Maryland Centre for Celiac Research: see websites: www.celiaccenter.org and www.celiac.com.

The degree of damage to the intestinal wall is proportional to the amount of gluten consumed. But even so-called normal and healthy volunteers on high-gluten test diets showed a deterioration of their intestinal walls and that their ability to absorb nutrients had been reduced.

Nutrient malabsorption is even more of a problem if white bread is used because it has lost about 80 percent of its vitamin and mineral content compared to that of whole-meal bread. What makes it even worse is the modern fast-baking method; instead of fermenting the bread for four to seven hours, it is now whipped with chemicals for two minutes. Thus the few remaining minerals stay tightly bound to phytic acid and cannot be absorbed by the body.

Gluten Allergy: When the intestinal wall deteriorates, it becomes permeable and molecules larger than normal can pass through. This allows partly digested proteins and bacterial toxins to enter the bloodstream, causing allergies and a deterioration of the immune system. Thus, gluten is actually the main agent causing allergies. Without a weakening of the intestinal wall, other potential food allergens would be prevented from reaching the bloodstream, except in the case of babies with intestinal walls that are still immature.

Young babies are also unable to digest starches. Feeding them wheat products, which is very common, almost automatically leads to the development of a wheat allergy. Therefore, cow's milk and wheat are two of the primary food allergies in our society. Not only do they precede the development of other allergies, but their initial appearance makes it much easier for secondary allergies to develop.

All these problems caused by gluten are greatly intensified by the tough network of gluten molecules formed in baking with gluten-rich flour. In former centuries only low-gluten wheat was available; high-gluten wheat is a modern innovation. In addition, the amount of wheat-baking, especially with high-gluten wheat, has greatly increased in our society compared to former centuries, while at the same time our digestive powers have declined.

As this trend has existed already for several generations, most of us are by now sensitized to some degree against wheat gluten, and baked wheat products have become problem foods for most individuals. This includes whole-meal wheat bread and not just the white variety. Our ancestors used mainly millet, rye, and oats as staple grains in addition to low-gluten wheat and spelt; they also used grains more in the form of porridge or wafers than as leavened bread. However, when leavened bread was used, it was mainly as sourdough bread, except in the cities where yeast-baking gradually dominated.

Rye and barley contain less gluten than wheat, and the protein composition is slightly different in these grains; thus, even if you have a wheat allergy, you may still be able to eat these grains. However, this is often not the case, and it is preferable that you avoid these for an extended period if you have a wheat allergy, and instead use gluten-free grains such as rice, millet, and maize. Buckwheat is not a cereal grain - it belongs to the rhubarb family - but it contains a gluten-like protein that requires allergy testing before being used by individuals allergic to gluten.

Processed food, including tablets and soy sauce, often contain added wheat starch or gluten. If you have a wheat allergy and use processed food, then read the labels carefully. Wheat grass, however, is safe to use, but wheat-germ oil needs to be tested for possible allergic reactions.

Iris Diagnosis to Detect Wheat Susceptibility: Blue-eyed individuals with wheat or gluten intolerance usually have a whitish color in the area of the iris denoting the intestines (refer to Step 12). This reflex zone that surrounds the pupil indicates the irritation and mucus covering of the intestinal wall. Frequently, the white fans out to other parts of the iris, but especially to the head and brain reflex areas that are in the upper part of the iris (between the ten and two o'clock positions). This indicates gluten-induced mucus congestion of the head and irritation of the brain tissues.

Such people are vulnerable to emotional irritability, bursts of temper, and, in the case of brain allergy, to epilepsy and mental disease. They cannot relax easily, and attending emotional workshops or seeking psychiatric help will do little permanent good if the nutritional cause of the problem is not corrected. These individuals often have oversensitive sense organs and tend towards farsightedness. As long as there is no serious mucus congestion in the head, the sharpened senses provide good eyesight and excellent hearing. The disadvantages are oversensitive taste buds, an oversensitive sense of smell, and distress caused by loud noise.

Magnesium supplements will help but will not remove the basic cause. In later life, mucus congestion may lead to deafness and eye diseases. If, in addition to wheat or gluten intolerance, there is difficulty in digesting fats, the white iris areas will become increasingly yellow.

Baked wheat products, especially in combination with sugars, are the most fattening foods for susceptible individuals. The metabolism of such individuals becomes inefficient and the wheat starch is mainly converted into body fat instead of energy. Frequently, individuals allergic to wheat, gluten, and beer have a distended abdomen (potbelly).

For many individuals, gluten products, especially wheat and beer, are strongly mucus forming. Generally, the whiter the iris, the more mucus-forming is the gluten consumed. Wheat and gluten are often a problem for those with asthma and hay fever. If you have a problem with mucus, avoid all gluten products and all foods containing lactose. After sufficient improvement, you can reintroduce small amounts of rye and oats into your diet and explore how much of these your body tolerates.

Most commercial health breads contain added gluten, skim-milk powder, or dried whole grain and are worse than plain whole-meal bread. I recommend rye sourdough bread and rye crispbread; however, rye sourdough bread often contains added wheat described as whole-meal or bread-making flour, so be careful.

Oats are high in proteins and nucleic acids, while the fibre (oat bran) is useful for binding and expelling surplus cholesterol from the intestines. Barley has an even higher cholesterol-reducing effect, and both oats and barley are superior to wheat. However, oats are also high in gluten and easily cause the same problems as wheat in gluten-sensitive individuals. It may be no coincidence that Scotland has the highest rate of multiple sclerosis, with a diet traditionally high in oats and low in vitamin D due to a lack of sunshine.

* * * * * * * * * * * *

Step 40

SWEET FOODS

Your sweet tooth may be your undoing, so consider reducing your intake of sugars.

The main problem with sweet foods is that sugars are absorbed too quickly. The body tries to maintain a fairly constant blood sugar level close to 100 mg of glucose per 100 ml of blood, but this becomes difficult with the habitual consumption of sweet foods.

When we have a starch meal, it takes several hours until all the starch is converted into glucose. The absorption through the intestinal wall is gradual and the liver can easily regulate the blood glucose level by forming glycogen or fat from surplus glucose. It is equally safe to eat raw cabbage or raw carrots, which can naturally contain amounts of sugar equal to a spoonful of honey. Again, it takes hours for the glucose from raw vegetables to enter the bloodstream, while with that from honey or fruit juices, it can take less than 30 minutes.

However, when sweet food is eaten during or after a protein meal, the sugar is absorbed considerably more slowly. The same applies when sweet food is combined with fat. On the other hand, sweetened starches, such as sweet porridge, cookies, or cakes, cause digestive and metabolic problems and are not recommended for habitual use. This means, in effect, that chocolate may be safer to eat occasionally than chocolate cake.

In former centuries, only limited amounts of sweet foods were commonly available in middle and northern Europe. More sweet food was used in Mediterranean countries, and descendents from these populations generally have a stronger sugar metabolism. In addition, formerly it was much more common to work hard physically and in this way burn up the glucose as quickly as it entered the bloodstream after a sweet snack. Food was largely unprocessed and rich in those vitamins and minerals that are needed to maintain the blood sugar regulation and sugar metabolism.

Three minerals required for metabolizing sugars are zinc (a component of insulin and of several enzymes), chromium (part of the glucose tolerance factor that allows glucose to enter the cells), and manganese (an important enzyme factor). In addition to these trace minerals, we need sufficient potassium and magnesium to produce energy from sugar. The vitamins most important for producing this energy from glucose are B1 and B2.

In contrast to former centuries, many people now suffer from vitamin and mineral deficiencies that make good blood sugar regulation nearly impossible and severely weaken the energy metabolism. On top of this, large amounts of sweet food are consumed almost daily, and the only “work” that many modern people do after a sweet snack is to raise a cup or stare at the television. The combined effect of all these negative factors is an erratic blood sugar level. It rises higher and higher after the ingestion of sweet food and falls more steeply and lower shortly afterwards.

Depending on one's metabolic constitution, this has different effects on different people.

Table 5-2: Symptoms of Allergy and Blood Sugar Problems

Allergies and blood sugar problems reinforce each other and can cause the same symptoms or contribute to the same chronic diseases.

abdominal pains
addictions
adrenal exhaustion
aggressiveness
alcoholism
anemia
angina
anxiety
arthritis
asthma
autoimmune disease
backache
bloating
blurred vision
burning eyes
cancer
caries
chest pain
chronic coughs
chronic indigestion
cold hands and feet
colds
colitis
confusion
conjunctivitis
constipation
convulsions
crying spells
depression
diabetes
diarrhea
digestive problems
distended veins
dizziness
eczema
edema
emotional instability
epilepsy
exhaustion
eye weakness
fainting
fatigue
forgetfulness
gallbladder pain
gastric pain/ulcer
glaucoma
hay fever
headache
heart attacks
heartburn
haemorrhoids
hyperactivity
hypertension
immune deficiency
indigestion
infections
inflammations
insomnia
irritability
joint pains
low blood pressure
migraine
mouth ulcers
mucus congestion
muscle cramps
muscle pains
myopia
nervousness
neuralgia
neuroses
nose bleeding
numbness
obesity
palpitations
phobia
poor circulation
rashes
respiratory problems
schizophrenia
sinusitis
skin problems
sweating
tachycardia
underweight
varicose veins
vomiting
weakness

If you look at Table 5-2, you may be surprised at the variety of symptoms that can occur and the number of chronic diseases to which a weak sugar metabolism contributes. The reason that sugar can cause such a range of problems is that glucose has a central position in producing our daily energy requirement. If our cells cannot efficiently produce energy, all our organs and body functions are affected.

How Diabetes Comes into the Picture: The disease most widely associated with a breakdown of blood sugar regulation is diabetes. When diabetes develops during childhood or in young adults (Type I), the main problem is usually a deficiency of the hormone insulin. Produced in the pancreas, insulin is required to channel the blood glucose into the cells. Therefore, a deficiency of insulin leads to a rise in the blood glucose level.

Currently, accepted medical opinion does not see a connection between the overconsumption of sweet food and the development of diabetes. However, experiments have shown that the insulin sensitivity decreases with frequent use of sweet food. Starting with a raw food diet that excluded all sweet food as well as other problem foods, I have been able to help diabetics of both types to improve their blood sugar regulation so that insulin injections or drugs were no longer needed.

In adult-onset diabetes (Type II), there is usually sufficient insulin available, but its effectiveness is greatly reduced. The main reason for this is the “disaccharide effect” as well as a deficiency in chromium, which is part of the glucose tolerance factor. Insulin and the glucose tolerance factor work together like key and keyhole to let glucose into the cells, but a deficiency of chromium prevents glucose from entering the cells.

The reason for a chromium deficiency is the habitual consumption of refined and sweet foods. Refined food is deficient in chromium, while eating sweet food raises not only the blood sugar level, but also the blood chromium level; this causes more chromium to be excreted with the urine. Another effect of chromium deficiency in diabetics and others is the excessive formation of cholesterol after eating sweet food and the subsequent development of arteriosclerosis.

These same factors also lead to the formation of cataracts. A diet high in fructose is especially implicated in causing eye diseases in diabetics. Fructose and glucose are the two components of sucrose - the normal household sugar - and also of honey. Fructose (best if obtained from fruits only) can enter the cells without requiring insulin and thereby can help to normalize the diabetic metabolism, but this is risky if mineral deficiencies are not corrected at the same time, especially deficiencies in magnesium, zinc, and chromium.

In animal experiments, an excessive intake of sugar resulted in a greatly enlarged pancreas and pituitary glands, both of which are involved in blood sugar regulation. A group of Aboriginal diabetics in Australia who lived for several weeks on “bush food” all quickly returned to a normal blood sugar regulation. Similarly, Indians living in Natal (South Africa) have a high sugar consumption and high levels of diabetes and heart disease compared to Indians living in India who use very little sugar and have very low rates of these diseases.

Type II diabetics are generally treated with tablets to lower blood glucose levels. As with insulin, these hypoglycaemic drugs do not protect the patients against the various harmful complications of long-term diabetes, such as degenerative eye changes, especially involving the retina, degeneration of the peripheral nervous system, and atherosclerosis, especially affecting the legs and heart. On the contrary, studies seem to indicate that these drugs accelerate such degenerative changes.

Doctors W. H. Philpott and D. K. Kalita point out in Victory over Diabetes that the overwhelming evidence of recent studies shows a shortened life expectancy and more serious complications from using diabetic drugs. In fact, the death rate actually doubled in those taking oral diabetic drugs. Most of these same drugs are still in use today. From a biochemical point of view, this is only logical and to be expected because if sugar levels are lowered without converting them into energy, then they have to be converted into fat and cholesterol (collectively called lipids) that then cause many of these problems.

When the liver and bloodstream are already loaded with lipids, it is difficult to convert excess glucose into more lipids. Therefore, obese or overweight individuals have greatly decreased insulin sensitivity, but insulin becomes much more effective if they lose weight. Other studies show that blood sugar regulation is best maintained with a diet high in vegetable fibre, especially from legumes, and that a high intake of simple carbohydrates or sugars tends to make insulin less sensitive (active).

Sugar added to the diet of research animals or increased in the diet of healthy volunteers has been reported to disturb the glucose metabolism and cause diseases of the eyes, kidneys, and blood vessels. Even if combined with a high-fibre, low-fat diet, added sugar still adversely affects the glucose tolerance. However, short-term studies may not show the harmful long-term effects of sugar in the development of Type II diabetes. This is because household sugar or sucrose consists of one molecule of glucose and fructose. Only glucose elevates the sugar level in the normal way, while fructose affects it only slightly. Therefore, in the glycaemic index, discussed earlier, which measures the immediate effect of different foods on the blood glucose level, sucrose is listed as a good food.

Instead, the danger of fructose is in causing an exaggerated insulin response, mainly when it is together with glucose in the same meal, be it from sucrose, honey, or even starches; to some degree, this happens even when ingested on its own as a sweetener. However, fructose in whole fruits is generally fine, provided it is not ingested close to a meal containing starches.

Let’s look now at the common habit of eating sweetened starches, as in bread with jam, marmalade, or honey, cakes, cookies, muesli, or breakfast cereals. The fructose contained in the meal causes a strong rise in the blood insulin level. At the same time, a large amount of glucose from the breakdown of starches enters the bloodstream. The excess of insulin quickly channels the glucose inside muscle cells, which are now overloaded with glucose. Only a small amount is needed for energy production; the rest can be converted to lactic acid, causing overacidity, or to body fat. Gradually, cells learn to protect themselves by becoming less responsive to insulin and making it harder for glucose to enter.

Until 1980, the rate of obesity and Type II diabetes was fairly stable. However, when the health authorities in the U.S. started vilifying foods containing fats and cholesterol and recommended eating carbohydrates instead, obesity increased from 13 to 14 percent of the adult U.S. population to 25 percent within one decade and continues to rise. Type II diabetes became an epidemic as well. In addition, for the first time in history, a large number of obese children developed Type II diabetes. As a result, it is now not usually called adult-onset diabetes.[30]

While exaggerated insulin response and resulting loss of insulin sensitivity are most pronounced in obese individuals, they gradually develop also in others after prolonged use of sucrose. The damage is greater the more sucrose is eaten in a gorging pattern instead of in small meals at intervals.

Surprisingly, sucrose has a worse effect than eating its two components, glucose and fructose, at the same meal. This is called the disaccharide effect and applies also to other sugars with two components, such as maltose with two glucose molecules. A hormone in the duodenum releases more insulin after ingestion of disaccharides than after monosaccharides such as glucose or fructose.

While increased insulin levels are desirable for Type I diabetics, with Type II diabetes it just means that more glucose is being converted into fat and cholesterol. However, there is a way to increase the insulin sensitivity of muscle cells naturally, such as with regular aerobic exercise. This allows glucose to enter muscle cells without the aid of insulin. In addition, the five-carbon sugar alcohol xylitol can be used by both types of diabetics without increasing blood glucose levels, although large amounts may cause diarrhea.

Hyperactivity and ADD: Much more widespread than diabetes or hyperglycaemia is the opposite condition of low blood sugar or hypoglycaemia. The most common milder form is reactive hypoglycaemia, in which the blood sugar response after eating non-sweet foods is normal. However, when sweet food is eaten, including sweet fruits, too much insulin is released and glucose floods the cells.

Glucose cannot be stored in cells and has to be metabolized. How this happens depends on the condition of your metabolism. The first stages are easy and result in the glucose molecule being split in half; this process is anaerobic and does not require oxygen. However, oxidizing enzymes are necessary for the next stage and these are usually deficient in adults with this condition. Children, on the other hand, often still have a reasonable supply of enzymes and are able to convert most of the glucose to energy.

This creates an energy burst, whether it is needed or not, that has to be used. The result is extreme restlessness: The muscles must move to use the surplus energy, and the brain races as well. When the “straw fire” of excess energy is used up, insufficient glucose remains to sustain normal activity and attention. Mental and physical exhaustion follows, until a sweet snack lights another straw fire of hyperactivity.

This temporary speeding up of the energy cycle is exaggerated, and in other cases triggered, by various factors that cause stress on the nervous and hormonal systems. The most frequent cause of such stress is a hidden allergy to food additives and to common foods such as wheat and cow’s milk products. Usually a child’s favourite foods are also to blame. Additional triggers can be fluorescent lighting, television viewing, strong electromagnetic radiation, petrochemical fumes, tobacco smoke, Candida albicans, and emotional stress.

In adolescents and adults such energy bursts, especially when combined with exposure to dietary allergens or stimulants, can lead to periods of uncontrolled violence. Many studies, involving thousands of institutionalized male juvenile delinquents, have shown a large drop in the incidence of antisocial behaviour in those on diets low in sugar and allergens. While the average improvement rate was almost 50 percent, repeat offenders improved by more than 80 percent, and their suicide rate fell 100 percent.

Females, because of their lower adrenaline levels, tend less towards violent behaviour. Instead, the effects of the strong blood sugar fluctuations after ingesting sweet food are more internalized. In sensitive individuals, emotions closely follow the blood sugar levels, often going up and down in quick succession, alternating between elation and depression.

Problems of Hypoglycaemia: The speedy metabolism that results in hyperactivity and violent behaviour explains why the term “fast oxidizer” has been coined for those who burn sugar too quickly. “Fast oxidizer” and “hypoglycaemic” mean basically the same thing: someone who has a speeded-up glycolysis (the breakdown of glucose inside the cells), which results in a lower than normal blood sugar level.

As opposed to hyperactive children, most adult hypoglycemics have a shortage of oxidizing enzymes, and this interrupts the energy produced at the end of glycolysis. Instead of energy, lactic acid is produced in a final anaerobic stage. Thus, only 20 percent of the total energy that would be available if glucose were completely oxidized to carbon dioxide and water is produced.

This results in overacidity and a chronic lack of energy, which is typical of hypoglycemics who are on sweet diets. Additional energy is required to remove the lactic acid via the kidneys. Lactic acid must be partly neutralized and this depletes the body of positive ions, especially calcium. A further result of overacidity is a high histamine level, which makes us susceptible to strong allergic reactions and frequent inflammations, while the skin becomes oversensitive to stings and all forms of irritation.

High histamine levels together with a calcium deficiency cause low blood pressure, which contributes further to a lack of energy and poor circulation, typified by cold hands and feet. Calcium is withdrawn from the blood vessels, resulting in varicose veins. Calcium deficiency also creates menstrual problems and weakens the eyes, making us nearsighted.

Nearsightedness or myopia actually results from a deficiency of two minerals, calcium and chromium, and both are induced mainly by a high consumption of sugar. Chromium deficiency arises from eating too much refined food that lacks this mineral (which means we never get to absorb it) and also from increased excretion of chromium in the urine, which occurs after eating sweet food when the blood levels are raised in response to high levels of insulin.

The ciliary muscles in the eyes need chromium to focus the lens for near vision. If the ciliary muscles become fatigued from a combination of chromium deficiency and prolonged close-focus activity, then the eye relieves this muscle strain by increasing the intraocular pressure; this pressure in turn leads to an elongation of the eyeball and thus to myopia.

Increased intraocular pressure, however, leads to an elongation of the eyeball only in those of younger years, when the body has a high requirement for calcium. With increasing age, the tissues become more calcified and rigid. In this situation, the eyeball cannot expand with increased intraocular pressure and glaucoma may result.

Scientific studies have shown that the excretion of calcium in urine increases up to fourfold in those with a strong insulin response when given large amounts of sugar; this is because plasma calcium levels are increased after eating sugar. The most common sources of the calcium excreted are the bones and teeth; this contributes to osteoporosis and dental caries. In elderly individuals, high plasma calcium levels also accelerate irreversible bonding with collagen and elastic protein structures, leading to increased rigidity and bone deformations, as evidenced in arteriosclerosis and arthritis. A high calcium concentration in the urine leads to an increased risk of kidney stones as well as kidney damage from the calcification of the filtration mechanism.

Lactic-acid formation that occurs after eating sugar contributes to strong inflammations such as in rheumatoid arthritis and gingivitis (inflammation of the gums). Normally, the saliva is nearly neutral in pH, which is less favourable for the growth of plaque bacteria than the acidic saliva of hypoglycemics. Sugars are the only energy source for these bacteria. When we are strictly on non-sweet diets, plaque bacteria are denied energy and they cannot multiply, and this effectively prevents caries.

Giving up sweet food, however, is extremely difficult for children and many adults because they are addicted to it. This addiction is in principle the same as later addictions to stimulants and drugs. Another effect of general overacidity is an inability to alkalize the gastric content when it enters the duodenum. This makes the pancreatic enzymes, which require an alkaline medium, inefficient and in time can lead to deficiencies and digestive problems (refer to acid-alkaline problems in Step 6).

Fructose and the Disaccharide Effect: It has been estimated that about half of all carbohydrates ingested in the U.S. are in the form of sucrose, common household sugar, which typically supplies about 15 to 20 percent of the total calories of the diet. One half of the sucrose molecule consists of fructose and the other half of glucose. Fructose is metabolized more slowly than glucose and also induces a much stronger insulin response in susceptible people, especially in the presence of glucose. This means that reactive hypoglycaemia occurs mainly when fructose is part of a meal.

Another effect of habitually high insulin levels is a gradual decrease in the sensitivity of blood glucose levels to insulin. Higher insulin levels also cause an increased loss of chromium in the urine. The result is the gradual onset of Type II diabetes. The higher the insulin levels, the more are sugars converted into fat and cholesterol. Biopsy samples of the human liver showed that fructose was converted into fatty acids at rates three to 24 times faster than glucose. Fat levels after fructose consumption rise especially high in those who have a tendency towards atherosclerosis; this can lead to diabetic complications and cardiovascular diseases. These negative effects of a high fructose intake also occur after eating sucrose. Other health problems arising from the fructose component in sucrose are raised blood pressure, caries, elevated uric-acid levels, and obesity.

Regarding sugar in general, controlled experiments with various species of animals as well as with human volunteers revealed significant rises in blood pressure not only after long-term ingestion of sugar, but even after a single dose. If taken on an empty stomach, the blood pressure rose 9 to 10 mm Hg for one to two hours. After sucrose ingestion, there is also an increase in uric acid levels. Uric acid is produced in the liver from breakdown products of fructose. A raised uric acid level in the blood is a feature of gout, some forms of arthritis, and heart disease. Men generally produce more uric acid after eating than women do.

Experimental studies on animals and humans have shown that sucrose is possibly the most caries-producing substance in our diet. Plaque bacteria convert sucrose to polymers that stick to the teeth. Some individuals have inherited fructose intolerance and must avoid all fructose and sucrose; it's interesting that they have been found to have substantially less tooth decay than the general population does.

Various animal and human studies have also shown that a diet high in sucrose can lead to a greater accumulation of fat deposits in susceptible individuals, generally those with a strong insulin response. Monkeys fed a diet high in sucrose produced three times as much deposited fat than those fed an equal amount of glucose.

Surprisingly, eating a certain amount of sucrose produces a stronger insulin response than eating the same amount of mixed fructose and glucose. The same is true for maltose, which consists of two molecules of glucose and other disaccharides (disaccharides are composed of two simple sugars). The increased insulin effect of disaccharides compared to that of their component simple sugars is called the disaccharide effect, as noted previously. The reason for this effect seems to be that disaccharides liberate about twice the amount of a gastric hormone called GIP (gastric inhibitory polypeptide) that simple sugars do.

In practical terms, this means that it is harmful for most of us to have fructose and glucose at the same meal. Examples of this combination are: muesli or honey, dried fruit, or fresh fruit to sweeten cereals or starches; dessert or sweetened coffee or tea after meals. Most harmful are the sweetened starches, such as cakes, sweet cookies, or crackers, and sweetened breakfast cereals. Beware especially of mixing fructose with starches if you want to lose weight.

Problems with Natural Sweet Foods: Unfortunately, natural sweet foods such as honey, dried fruits, or fruit juices are no better for diabetics or hypoglycemics than food sweetened with white sugar. What seems to matter is how much sweet food is used and how quickly it enters the bloodstream. Fruits that are acidic as well as sweet may be even worse, especially for hypoglycemics, than plain sugar because the fruit acids reinforce the existing overacidity. However, by neutralizing fruit acids they can become beneficial, and eating the whole fruit is better than drinking the fruit juice only. Raw honey has a high concentration of the beneficial starch-digesting enzyme amylase, which is lacking in heated commercial honey; sun-dried fruit is healthier than commercial heat-dried fruit.

While fructose from fruits can be helpful for Type I diabetics, with hypoglycemics it only increases the lactic acid overload. Lactic acid overproduction from the combined effect of strong muscle activity and sweet food ingestion is also responsible for reported cases of allergic reactions after jogging. The resulting overacidity simply brings already existing hidden allergies into the open. Apart from the loss of calcium, another side effect of lactic-acid buildup is a tendency to anxiety neuroses, phobias, and panic attacks.

From a health viewpoint, it is preferable to have sweet food only occasionally and with a protein snack or meal. Removing the sweet taste from the tongue as quickly as possible helps minimize the insulin overreaction. Alternative sweeteners such as the sugar alcohols manitol, sorbitol, and xylitol should only be used in small amounts. In larger amounts, they can cause diarrhea or keep the craving for sweetness alive, which in time leads to another binge on sweet food.

The leaves of the herb stevia have long been used in traditional cultures as a sweetener. They are 15 to 30 times as sweet as sugar and an extract can be 300 times as sweet. It is not known if large amounts of stevia ingested over long periods of time cause unacceptable side effects, but I regard it as safer than aspartame or other artificial sweeteners. It is acceptable to use in small amounts but do not use it instead of sugar in an addictive way. Other sweeteners, such as saccharine and aspartame (Nutrasweet), may cause allergies or worse. Fructose is the sweetest sugar, followed by sucrose. Glucose and maltose are less sweet but somewhat safer for most people, especially combined with starches. Maltose is available as barley sugar. Preferably, use maltose instead of sucrose for sweetening starches.

As a general rule, for most individuals it is best to reserve sweet food as a rare treat for special occasions rather than to use it as an everyday food. However, if you do not have a craving for sweet food, then it is not likely to be a problem for you. Those who have a low fasting blood sugar level (feeling weak or dizzy on an empty stomach) must be more strict than reactive hypoglycemics and are advised to avoid all sweet food for several years so that the sugar metabolism has time to recover.

* * * * * * * * * * * *

Step 41

MEAT & FAT

Overconsumption of meat and fats can create serious health problems. Try cutting back on your intake.

Those who eat a lot of red meat tend to develop the opposite characteristics and health problems to those on a sweet diet. Red meat has a stimulating effect on the sympathetic nervous system and adrenal hormones. This improves the ability to "push through" and provides "drive" such that heavy meat eaters may become powerfully assertive and even aggressive. Another important effect of eating red meat is a tendency for the blood pressure to rise because of meat's effects on the adrenal hormones and the nervous system. Among the meats, beef has the strongest influence on blood pressure.

All this can be beneficial for hypoglycemics with low blood pressure, lack of drive, and weak adrenaline glands, but for others, especially stressed males, it may mean hypertension and kidney problems. Health problems associated with eating meat fall into three categories: those common to all meat, those associated with cooked meat, and those with meat produced by factory farming.

Common to all meat is a high phosphorus content with an unfavourable phosphorus-to-calcium ratio. The high phosphorus content of meat stimulates the parathyroid glands (associated with the thyroid in the neck) and raises the calcium blood level, in some cases by activating and freeing calcium from the bones. In combination with a prevailing over-alkalinity, this leads to stone formation (such as kidney stones), kidney damage, arteriosclerosis, stiff joints, ankylosis, and arthritic bone deformations.

Heavy meat eaters have been found to excrete in the urine up to four times more calcium than normal; this is a cause of kidney stones and other kidney damage. It is even worse when high meat consumption is combined with a high sugar intake. This can increase the loss of calcium and corresponding kidney damage up to tenfold above normal, while at the same time greatly accelerating osteoporosis and tissue calcifications, such as arteriosclerosis.[31] Carnivorous animals in the wild do not have this problem caused by an excessive intake of phosphorus because they also eat the calcium-rich bones. We, too, can avoid this problem by using bone broth or other foods and supplements high in calcium.

Meat, because of its high phosphoric acid content, is regarded as an acid-forming food. However, in habitual heavy meat eaters or in those with a weak digestive system, the metabolic and hormonal stimulation resulting from red meat consumption gradually declines and eventually leads to a sluggish metabolism. These individuals are called "slow oxidizers." With a slow metabolism, a deficiency in metabolic acids develops and this makes the whole body too alkaline. With over-alkalinity, histamine remains tightly bound to tissue proteins and the skin. In such people, even the emotions are sluggish and they often appear unresponsive, with little outward expression.

With low histamine levels, inflammatory responses, which are required for many self-healing actions of the body, are suppressed. Allergens entering the bloodstream will not cause a warning allergic reaction; instead, autoimmune diseases may develop from a continued invasion of allergens. Histamine dilates the blood vessels, causing the blood pressure to drop, a problem of hypoglycemics. However, lack of histamine in slow oxidizers greatly contributes to the development of high blood pressure.

A diet high in meat, together with weak kidneys, often leads to the accumulation of uric acid in the muscles, causing gout and rheumatism. General over-alkalinity leads to a lack of gastric acid in later life, resulting in poor absorption of minerals and an incomplete digestion of proteins. Intestinal putrefaction and eventually cancer may follow, aided by chronic constipation and a depletion of pancreatic enzymes.

One problem of a diet high in meat and actually of all high-protein diets is the accumulation of protein fragments in the spaces between tissue cells. This causes or contributes to water retention or oedema. In addition, capillaries become clogged, preventing efficient nutrient supply to the tissues; lymph glands congested with protein fragments are breeding grounds for bacteria and weaken the immune system. All these complications aid the development of degenerative diseases and cancer.

The indicated problems, metabolic sluggishness resulting in over-alkalinity, and accumulation of uric acid and protein fragments can happen with any diet that is too high in protein and are more prevalent with a sedentary rather than an active lifestyle. They can be eased or prevented with an increased intake of fruit acids and digestive enzyme supplements, in addition to fasting or cleansing periods with a high intake of fluids.

However, in large measure these problems are not due to eating meat, but to eating cooked meat. Raw meat contains a high level of protein-digesting enzymes (cathepsin). Left to its own devices, meat tends to digest itself. Eskimos and other traditional cultures remained healthy on a diet high in flesh foods as long as they ate it mainly uncooked.

On a diet of cooked food generally, and especially with meat eating, the weight of the pancreas has been shown to increase. The pancreas of animals on cooked diets is about 300 percent heavier than it is for those on a raw-meat diet. I take this to mean that the pancreas becomes enlarged because on a cooked diet it needs to produce 300 percent more digestive enzymes than on a raw diet. Cooked meat is difficult to digest and needs particularly strong digestive juices. An aging body or one in poor health will not be able to keep up with this excessive enzyme production and enzyme depletion and deteriorating health will be the expected consequences.

While animals on raw meat remain healthy, those on cooked diets tend to develop a variety of degenerative diseases typical of human conditions. To remain healthy, dogs and cats need a raw-meat diet that includes offal and bones. Free-living carnivores not only eat their meat raw, they also eat bones and offal, have a very active lifestyle, and efficiently eliminate metabolic residues. If we want to remain healthy on a diet high in meat, then we have to adopt similar practices.

Health problems caused by cooking and enzyme depletion are aggravated by an accumulation of toxic chemicals and hormones. This is due to the common practice of factory farming. Chemical contamination is much worse in grain-fed beef from feedlots than in grazing animals. There are numerous studies showing the harmful effects of residues from pesticides, antibiotics, and growth promoters.

If we cannot obtain the meat of healthy, chemical-free animals and eat it predominantly raw (for instance minced or marinated), or use digestive enzymes and appropriate cleansing periods, then it is better for our health to eat meat only sparingly.

Fats and Oils: Fat is an important source of energy and some of its components, the polyunsaturated fatty acids and lipotrophic factors, are essential for our health. Fats and oils are composed of fatty acids, chemically linked with glycerol (a three-carbon sugar alcohol). While in a sense all fats and oils are often grouped together as fats or lipids, commonly we speak of fats when they are solid at room temperature and of oils when they are liquid at room temperature.

Fatty acids are long carbon chains (containing from four to 26 carbon atoms) with hydrogen attached to each carbon and an organic acid group at the end. In saturated fatty acids, every carbon position is filled with hydrogen, while unsaturated fatty acids have a reactive double bond between two carbon atoms. If there are two or three such double bonds, then we call it polyunsaturated.

Naturally occurring polyunsaturated fatty acids are folded in a specific way and called cis-fatty acids; during heating their structure changes to the more stable trans-fatty acids. Cis-fatty acids are important as structural ingredients of cell membrane and to facilitate oxidative energy production, but trans-fatty acids can interfere with these functions.

Animal fats consist mainly of long-chain saturated fatty acids with a high melting point. Beef and mutton have the hardest animal fats with the highest melting points; butter is marginally softer, due to its lower melting point.

A high consumption of these hard fats poses two problems. If no attempt is made to use sufficient polyunsaturated oils that provide the necessary essential fatty acids (EFA) and lipotrophic factors (e.g., lecithin and vitamin E), then a deficiency of these is likely to develop. The other problem is that the liver may not be able to metabolize an excess of long-chain saturated fatty acids.

The EFA's and lipotrophic factors are important in forming cell walls, for the immune system, and for platelet aggregation in blood clotting to prevent bleeding and unnecessary clotting. They also form tissue hormones called prostaglandins and other important biochemicals. Deficiencies in these nutrients can cause the following: a weak immune system; blood vessel and cardiovascular diseases; blood clots causing heart infarctions, paralysis, and strokes; skin problems, often starting with scaliness and including eczema, acne, and psoriasis; poor growth; impaired reproduction; and bronchial diseases.

Many degenerative diseases have been linked to an excess of saturated fats and a subsequent deficiency in EFA's; these include atherosclerosis, nerve and brain disorders, schizophrenia, psychoses and neuroses, rheumatoid arthritis, and multiple sclerosis. However, an oversupply of polyunsaturated oils can have its dangers as well, especially if you are deficient in vitamin E. Polyunsaturated oils easily oxidize in contact with air to form oxides and peroxides that have a high potential for causing cancer and liver damage. This danger is greatest when these oils are refined and heated, as in frying.

The Problems with Saturated Fats: The liver has only a limited potential for handling fats. When the liver is flooded with a consistently high supply of saturated fats, it cannot metabolize them. Fats, by the way, do not come only from fatty foods; a surplus of carbohydrates can be converted into saturated fat. Only a small amount of carbohydrate can be stored as glycogen, so all the surplus energy from a carbohydrate meal is stored as fat.

Normally, fats are broken down into keto acids in the liver and transported by the bloodstream to other cells where they combine with products from sugar metabolism to create energy. When fats are oversupplied, there may not be enough glucose to produce energy efficiently, either because the carbohydrate intake is insufficient or because the effectiveness of insulin is reduced. A surplus of keto acids can then be resynthesised to saturated fatty acids and cholesterol (see Figure 5-2 in Step 43).

Another difficulty is that not all the fat is convertible into keto acids, and a high percentage of fatty acids and fat molecules are thus released directly into the bloodstream. The body may reduce these surplus fats by storing them in adipose tissue, forming fat deposits; this happens especially with heavy meat eaters who have become slow oxidizers, as explained earlier. Some portion of the fatty acids and cholesterol may then react with the usually elevated calcium levels of slow oxidizers to form fatty plaques, also called sludge, in the blood vessels.

Fatty deposits are called atheromas and consist mainly of smooth muscle cells filled with sludge, similar to fatty tumours. Such deposits can help cause arteriosclerosis and cardiovascular diseases. Not only animal fats, but also peanuts and peanut oil have properties that cause atherosclerosis: peanut oil contains some saturated fatty acids with unusually long chains, namely arachidic and behemic acid. However, these fatty acids seem to be beneficial for arthritic joints.

In those with sensitive skin, or especially during adolescence, the body tries to cleanse itself by removing fatty sludge through the sebaceous glands, causing acne and an unclean skin. Eating sweet food is as much to blame for this as is fat intake. Fats in the feet, for instance, have a melting point of 33° F, while fats of the inner organs melt at 114° F. If the diet contains too much saturated fat and sweet food, then the melting point of the outer fats will be raised. In cold parts of the skin, those that are exposed or with poor blood circulation, the fats will harden within the sebaceous glands and block the outflow of sludge or normal skin oils. This causes inflammation of these glands, as in acne and other skin problems.

Even worse than natural hard fats are those that are chemically hardened, such as margarine. During the hydrogenation process, non-biological fatty acids are formed called trans-fatty acids, which the body cannot properly metabolize and which interfere with the utilization of essential fatty acids.

Cardiovascular diseases and an increased incidence of cancer can be a frequent outcome of a diet habitually high in red meat and saturated fats, but so is fatty degeneration of the liver. The blood becomes more viscous or sluggish after a fatty meal. This causes a reduction in tissue oxygenation and an increased susceptibility to arthritic pain, heart pain, or shortness of breath; blood flow to the brain is reduced, producing tiredness and decreased mental alertness.

Some of the adverse effects of fatty foods result from a deficiency of lecithin in the diet. Lecithin is a natural emulsifier of fats; it can be synthesized in the liver but often a deficiency develops because naturally occurring lecithin or choline is lacking in refined food. This is a main cause of fatty degeneration of the liver, which can be reversed, however, with organ cleansing practices, lemon juice, and lecithin supplements.

Another problem with a high-fat diet is the destruction of naturally occurring lipase in cooked food. Lipase is important as our main fat-digesting enzyme, and we rely to a significant degree on lipase in natural foods to supplement our pancreatic lipase. With the destruction of food lipase by heating, on a high-fat diet we gradually develop a lipase deficiency, and this leads to the development of fat-related diseases. Formerly, inhabitants of Bulgaria and the Caucasus lived on high-fat diets without developing fat-related diseases. They even consumed an abundance of full-fat milk products, yet were famous for their health and longevity. They generally used unheated fats, which seems to make the difference.

Here is another example of the significance of natural lipase in our diet. The skin disease psoriasis is generally aggravated by fatty foods, and fats must be avoided to cure it. However, in the early twentieth century, an American doctor reportedly cured psoriasis routinely with an intake of two pounds of raw butter per week for six weeks and with reduced amounts afterwards. Modern pasteurized butter definitely aggravates psoriasis but raw butter cures it - what is the difference? Raw butter from free-ranging cows is very high in lipase and pasteurized butter is devoid of it, so the difference may be lipase.

The healthiest saturated fat may actually be unheated coconut oil. This is due to its high content of short-chain fatty acids that can easily move into cells to be converted to energy. The long fatty acids of other fats, in contrast, require L-carnitine for this process. Individuals who have added coconut oil to their diet generally seem to have more energy without putting on extra weight.

Cholesterol: Cholesterol is now portrayed as the great medical villain causing cardiovascular disease and heart attacks by clogging the arteries, especially the coronary arteries of the heart muscle, with fatty sludge. However, cholesterol is an extremely valuable biochemical. It forms the base of steroid hormones and vitamin D and it is highly concentrated in the brain, liver, adrenal glands, and nerves. Hypo-glycemics or fast oxidizers usually have low levels of this important nutrient, and recent studies link low cholesterol levels with an increased risk of cancer.

Excess cholesterol is normally excreted with the bile but may form gallstones if there is a lack of lecithin and the amino acid taurine needed to form bile salts. The cholesterol content of food is not normally a problem in itself. Most cholesterol in the body is formed in the liver from breakdown products of saturated fats, and the more cholesterol that is ingested with food, the less is synthesized. Thus the body’s overall cholesterol level normally remains stable.

You may read about HDL and LDL in connection with cholesterol and heart disease. HDL is short for high-density lipoproteins and LDL for low-density lipoproteins. Lipoproteins are biomolecules composed of proteins and lipids; they can be used to transport cholesterol between the liver and tissue cells. Those that carry cholesterol from the cells to the liver are the high-density lipoproteins, regarded as beneficial. Their cholesterol is called HDL cholesterol. Low-density lipoproteins carry LDL cholesterol from the liver to tissue cells; a low level of these is preferable.

The aim of medical intervention is to increase the level of HDL cholesterol and lower the level of LDL cholesterol in the blood. However, I am not convinced of the value of the medical strategies being used in reducing or curing cardiovascular disease. Therefore, I tend to disregard measurements of HDL, LDL, and cholesterol and specific interventions to normalize levels; instead, I focus on a natural lifestyle to improve health.

Cholesterol overproduction arises mainly from an excess of saturated fats and a deficiency of chromium. The body tries desperately to lower the keto acid and fatty levels in blood and liver, and cholesterol overproduction provides a temporary solution. Even elevated glucose and fructose levels can be reduced by their conversion into cholesterol. In this way, sweet foods contribute to the development of cardiovascular diseases. Sugar ingestion raises the insulin level, and a key enzyme that synthesizes cholesterol is actually regulated by insulin. The higher the sugar intake, the higher the insulin level, and thus the more cholesterol synthesized in the liver.

A higher intake of polyunsaturated oils in males can lead to some decrease in cholesterol levels. However, in women it has led to an increased risk of cardiovascular disease. In animal experiments, it was found that high-sugar diets were less damaging in the presence of saturated fats such as beef tallow or coconut oil. Despite huge efforts, medical research so far has not conclusively shown that a high cholesterol diet or high cholesterol blood levels cause heart disease. As we can see from the published scientific facts, it is often the contrary:

- Medically supervised trials with low-cholesterol diets were unsuccessful in significantly lowering blood cholesterol levels or reducing the risk of heart disease.
- Low blood cholesterol does not mean freedom from heart disease. Some drug treatments to lower cholesterol have resulted in increased rates of heart disease.
- About 50 percent of men under the age of 55 who die of heart attack do not have elevated cholesterol levels or any of the other risk factors such as hypertension, smoking, obesity, or diabetes.
- Some cultural groups have a low risk of heart disease despite a high intake of cholesterol; Western populations also had this before the compulsory pasteurization of milk products and the chlorination of water supplies.

Recently an evaluation was published of six major trials that lowered cholesterol in the blood in tens of thousands of individuals over many years. There was on average a cholesterol reduction of about ten percent and only a small reduction of 14 percent from death of coronary heart disease. However, the total number of deaths was significantly higher among those patients who had their cholesterol lowered compared to the control group. This included not only higher mortality from cancer but on average a 67 percent higher mortality rate from violent death, such as accidents, homicides, and suicides. This high rate of death from violent causes was found in every one of these trials.

In looking for an explanation for this increased violence, in another study researchers found that monkeys on a diet low in cholesterol and saturated fat became more aggressive. Studies on humans found lower cholesterol levels among criminals and individuals with aggressive or violent behaviour or limited self-control. This is probably due to a more erratic sugar metabolism that results if cholesterol and fats are curtailed in an inappropriate way. Erratic blood sugar fluctuations tend to induce uncontrollable mood swings, including aggressive behaviour. Criminal and aggressive individuals have been frequently shown to have an erratic or hypoglycaemic-type sugar metabolism.

After decades of habitual sugar and meat consumption, the insulin level rises permanently because of a reduced efficiency in controlling glucose levels. Then cholesterol levels also remain chronically high. Reducing cholesterol in this situation with drugs means that less sugar is converted into cholesterol. Therefore, blood sugar levels become more erratic and behaviour is more aggressive. Thus, low cholesterol diets are appropriate only if the diet is also low in sweet food.

There is a connection between cholesterol intake and atherosclerosis and heart disease. If cholesterol in food is heated in the presence of oxygen, then cholesterol is oxidized to oxy-cholesterol and this damages the blood vessels and causes atherosclerotic lesions. Once such lesions exist, normal cholesterol in combination with high calcium levels can also contribute to a further buildup of plaque. In experiments using animals, it was found that with high intakes of either cholesterol or oxy-cholesterol, atherosclerotic degeneration occurred only with oxy-cholesterol, not with cholesterol.

Oxy-cholesterol is carcinogenic. The high incidence of cardiovascular disease and atherosclerotic aging processes may be traced in part to the heating (especially frying) of meat, eggs, and milk products, which liberates oxy-cholesterol. Avoiding this and chlorinated water and using high levels of antioxidants could also ensure that cholesterol does not become oxidized inside the body.

In conclusion, to avoid cholesterol-related health problems:

- Minimize the use of sweetened food, heated fats and oils, and chlorinated water.
- Maximize the intake of unheated fats and oils, antioxidants, and fresh raw foods.
- Lower your high cholesterol levels with cleansing periods and plenty of lemon juice.

Omega-3 Fatty Acids: There are two groups of essential fatty acids, or EFA's. The common oils from warm climate seeds (sunflower, safflower) contain mainly the omega-6 series of EFA's, based on linoleic acid with two unsaturated or double bonds. Cold climate plants and fish oils, on the other hand, contain mainly omega-3 fatty acids. Linolenic acid in linseed (flaxseed) with three double bonds has the shortest carbon chain in this group. In fish oils, DHA (docosahexaenoic acid) and EPA (eicosapentaenoic acid) predominate.

The omega-3 fatty acids have generally been neglected in modern nutrition. This has led to widespread deficiency symptoms and chronic disease. Their fundamental importance can be seen in the fact that DHA is the main unsaturated fatty acid in the retina and the brain and is exceptionally high in the thymus and reproductive glands, especially the testes. DHA is obviously most needed during pregnancy and early infancy for brain and gland development. It is up to 30 times higher in breast milk than in cow's milk or formula. This may bear some relation to epilepsy and eye weakness in infants, and its deficiency has also been linked to colic; inflammation of the lungs, digestive tract, joints, and kidneys; crib death (SIDS); skin diseases; and a predisposition to frequent infections. A recent scientific study reported an eight-point higher intelligence quotient in preterm babies fed breast milk as compared to those on cow's milk formula. As explanation, the deficiency of DHA in cow's milk was cited.

Fish oils are now commonly used to prevent heart disease, while linseed oil with its 45 percent alpha-linolenic acid has been successfully used to help treat various mental diseases. Even severe cases of schizophrenia and manic-depressive illness have reportedly been overcome or even controlled with ingestion of linseed oil. One or two tablespoons of linseed oil daily can help overcome depression, but twice that amount can cause a manic episode with racing thoughts, so too much has its downside.

Generally, however, this dosage can be increased about tenfold before negative symptoms develop. Mood swings, irritability, and anxiety can be relieved, and premenstrual tension and other menstrual problems and vaginal dryness after menopause are improved; prostate enlargement, arthritis, eye problems, immune deficiency, hair loss (alopecia areata), obesity, and skin problems are also often relieved.

EFA's and especially the omega-3 fatty acids (such as DHA and EPA) regulate the fluidity or softness of the cell membranes. A deficiency is noticeable as a hardening of the skin, as with dry, scaling, and flaking skin, patchy eczema, acne, and skin sensitive to the sun. Most common is phrynoderma, with rough, hard skin on the buttocks, thighs, and arms. The colder the climate, the more omega-3 fatty acids are needed for membrane fluidity.

The omega-6 series predominantly form prostaglandins that increase inflammation and with this the pain experienced in many diseases. The omega-3 series form anti--inflammatory prostaglandins. Thus, omega-6 fatty acids aggravate allergic reactions, including asthma and rheumatoid arthritis, while omega-3 fatty acids reduce these. Another important function of both groups of EFA's is in the production of oxidative energy. Also, the "brown fat" that produces and regulates our body heat is mainly stimulated by omega-3 fatty acids. This is especially important for overweight individuals who want to convert body fat into body heat and for those who habitually have cold hands and feet.

While individuals with deficiency symptoms related to omega-3 EFA's require a higher intake for some time, it is generally estimated that a normal diet should provide two to three percent of calories from omega-3 and five to ten percent from omega-6 fatty acids. In contrast, the average modern diet provides about 0.4 percent of calories from omega-3 and ten percent from omega-6 fatty acids. You can clearly see the imbalance.

While using polyunsaturated oils instead of saturated fats has been found to reduce the cholesterol level and incidence of cardiovascular diseases, the rate of cancer has risen, presumably due to the combination of an inflammatory effect, antioxidant deficiency, and rancidity of refined oils. Biological damage can result from the peroxides and free radicals (highly reactive molecule fragments) formed from polyunsaturated fatty acids, and the common seed oils also have a strong inflammatory influence as they tend to form pro-inflammatory tissue hormones. This not only increases allergic reactions and arthritic inflammations, but also supports the growth of tumors.[32] In addition, it has been reported that melanoma sufferers had a higher rate of polyunsaturated fatty acids in their skin than non-affected individuals.

A high intake of polyunsaturated oils also requires a greatly increased intake of antioxidants, especially of vitamin E. However, refined oils as well as our normal diet are usually very low in vitamin E. This, in addition to a low intake of bioflavonoids and carotenoids, is probably a main factor in the present high incidence of skin cancer.

Health problems are further caused by the improper production of polyunsaturated oils. These oils are very delicate substances that should be produced and stored at cool temperatures and in conditions that exclude oxygen and strong light. Instead, they are typically extracted with chemical solvents or cooked and pressed hot and then refined with bleaching and deodorizing at very high temperatures. Even many cold-pressed oils are refined in this way. This causes oxidative damage to the oils and a structural change by which a high percentage of the natural cis-fatty acids form unnatural trans-fatty acids. Packaging oils in clear plastic bottles instead of brown glass bottles means they are exposed to light and can dissolve chemicals from out of the plastic, thereby contaminating the oil.

One oil, though reasonably high in omega-3 fatty acids, should be avoided. This is canola or rapeseed oil. There are medical warnings that an ingredient called erucic acid apparently can agglutinate (clump together) red blood cells and reportedly contribute to loss of vision or blindness, emphysema and other respiratory problems, anemia, and problems with the central and peripheral nervous system.

* * * * * * * * * * * *

Step 42

CHEMICALS IN YOUR FOOS

Minimize your exposure to toxic chemicals, water additives, and yeasts and moulds - your body will thank you for it.

Almost all our foods contain non-biological chemicals to some degree. It starts with water-soluble fertilizers that are overused in agriculture and cause mineral imbalances and trace mineral deficiencies in our crop plants; it continues with the liberal use of herbicides, pesticides, and fumigants. Many of these leave residues on foods we ingest. Growing our food crops using common water-soluble fertilizers causes a similar disturbance to the plant metabolism as a high-sugar diet does to a human hypoglycaemic metabolism. The cells are flooded with some nutrients, while others go lacking.

While there are now high amounts of non-biological chemicals in our food, essential nutrients in these foods are becoming rare. John D. Hamaker in The Survival of Civilization gives the following example of trace mineral deficiency induced by using fertilizers. In 1948, the highest iron concentration in cabbage was found to be 94 ppm (parts per million) and the lowest was 20 ppm. By 1963, the average value had dropped to four ppm. The story was similar for other vegetables. The general conclusion is that the average trace mineral content of vegetables in 1963 was close to or even lower than the lowest concentrations in 1948.

The latest results show that vitamin and mineral concentrations are much lower now. The food tables of the U.S. Department of Agriculture for the year 2000 show that the typical vitamin C content of broccoli, cauliflower, and spinach is only about half of what it was in 1963. Also vitamins A, B1, and B2 have declined sharply since then, and the calcium in corn has decreased by 78 percent.

The content of trace minerals can vary in the same type of vegetable by more than 100 times, depending on the quality of the soil in which it is grown. The selenium content in wheat was found to vary from 0.6 mcg/100 g in parts of New Zealand to 130 mcg/100 g in parts of Canada. The vitamin content can fluctuate to a similar degree. The vitamin and trace mineral values listed in food tables are therefore often completely meaningless, especially when the original values are decades old. The trace mineral content of our arable land is now extremely low and rapidly declining even further because of heavy cropping without replacing what has been taken out.

Trace minerals in the soil originate from disintegrating rocks. In order to remineralise the soil, crushed rock would need to be distributed over our cropland. Improved agricultural conditions must be maintained by returning to the soil whatever has been taken out, using composting and organic farming methods.

Heavy use of water-soluble fertilizers allows cropping in poor soils, but plants become deficient and offer little resistance to disease and insect attacks. This then justifies the liberal use of highly toxic agricultural chemicals, part of which remain in the plants and fruits and are eventually ingested by us. The most commonly affected are commercially produced fruits. Agricultural chemicals can then accumulate in the body, especially in fat tissue, and they can also damage the liver. Organo-phosphate insecticides, for instance, can cause weakness, muscular trembling, cold sweats (an early sign of liver damage), irritability, tightness of the chest (asthma), nausea, and abdominal pain. Organo-chlorine insecticides, which include DDT, lindane, 2.4-D and 2.4.5-T, are even more toxic to humans. They affect mainly the nervous system and can cause a wide variety of symptoms including convulsions, numbness of extremities, uncontrollable excitement, apprehensions, allergies, and birth defects.

More recently, pesticides and herbicides used on farms and in households have been increasingly linked to neurological diseases, such as Alzheimer's and Parkinson's. Develop-ment of symptoms due to low-level chronic exposure is gradual, but it has been suggested by researchers that long-term exposure to sub toxic levels of chemicals is much more likely to lead to neurological disorders than to physically based diseases. In animal experiments, chronic systemic pesticide exposure reproduced the features of Parkinson's disease.

People who use pesticides in their homes are twice as likely to develop Parkinson's disease as people not exposed to domestic pesticides. Also, a combination of different pesticides is more likely to cause greater toxicity problems than exposure to individual chemicals.

Paraquat, a common herbicide, and maneb, a common fungicide, did not affect mice that were exposed to either one of them. However, when the chemicals were used together, dopamine neurons (brain cells needed to regulate voluntary muscle movements) began to die at low exposure levels or at only a fraction of what is normally regarded as toxic. Ongoing exposure leads to progressive neurotoxicity.

There are many combinations of low-level toxic residues in our food and water, and no one knows what health effects they can produce. However, one can speculate. Only 100 years ago, cancer and Parkinson's disease were relatively rare and affected only old people. Now they are epidemic and even children and juveniles develop these degenerative diseases. Cancer and especially breast cancer have been strongly linked to pesticide accumulations in the body.

Some widely used fumigants can pose an additional hazard to our health. Organic bromides, especially methyl-bromide, destroy the pantothenic acid (an important B-vitamin) in fumigated grain and can cause serious deficiency disorders. Later on, the fumigant may not be detectable in the grain, which is then regarded as completely "safe." However, the germination power of such grains can be very poor, which shows they have been damaged.

Our unsound farming, refining, processing, and cooking methods ensure that we receive only a fraction of the vitamins and trace minerals that we would get from unprocessed raw food grown in good soil. However, these extremely low values are now regarded as "normal" by health authorities because the classical acute symptoms of serious deficiencies are still absent in most people. Apparently, no thought is given to the long-term effects of multiple subclinical deficiencies.

Meat, poultry, and eggs can contain residues of antibiotics, pesticides, and synthetic hormones acting as growth promoters. Hormones are supposed to be destroyed during digestion and therefore are regarded as harmless. However, this may not be so with synthetic hormones or when the digestive system is weak. In addition, a surprising number of nutrients can be absorbed through the mouth tissue, including giant molecules such as vitamin B12; this type of absorption may also include hormones otherwise not absorbed. For example, secondary sex changes have been reported due to the habitual consumption of hormone-treated chicken.

Besides liver degeneration, various forms of allergies, and leukaemia, chemical overload has been linked to chronic fatigue syndrome, ME (myalgic encephalomyelitis), and motor-neuron disease. Individuals with these diseases seem to have a liver that is less able to detoxify non-biological chemicals.

Chemicals Added to Foods: An even greater health hazard is the multitude of chemicals directly added to food products. These include preservatives, colours, flavours, emulsifiers, stabilizers, and so forth. Such chemicals need to be detoxified in the liver, but the volume typically ingested can contribute to liver damage. Many of these chemicals are fat-soluble and get stored in our fatty tissues. During prolonged fasting or disease resulting in weight loss, these chemicals may be released into the bloodstream in excessive amounts and cause distressing symptoms.

Nitrates and nitrites are added to most processed meat products. They can damage the blood and the thyroid gland, and in pregnant women, the foetus; they can also cause severe vitamin-mineral deficiencies and may be transformed into cancer-causing nitrosamines in the intestines. Sulphites (e.g., sulphur dioxide and metabisulphite) are widely used in processed foods, such as frozen foods, dried fruits, fruit juices, wine, beer, and prepared restaurant foods. In susceptible individuals, they can cause a swelling of hands, feet, and larynx, dermatitis, abdominal distress, and respiratory problems that especially affect asthmatics.

Food chemicals frequently cause allergies and trigger hyperactivity. This applies especially to synthetic colours. Artificial sweeteners, and especially aspartame, have been linked to a high rate of brain tumours, serious metabolic and neurological complication in diabetics, epilepsy, toxicity in the thyroid, and emotional disorders. While most of these chemicals are added only for the manufacturer’s convenience, the food industry could easily use vitamins C and E as well as lecithin and other biological chemicals as antioxidants, preservatives, and emulsifiers.

Another group of food additives called excitotoxins act like neurotransmitters (brain chemicals that transmit nerve signals) and overstimulate the brain in susceptible people. Excitotoxins are taste or flavour enhancers that release glutamic acid or glutamate. Aspartic acid and cysteine are also brain-active amino acids or excitotoxins. The best known example of an excitotoxins is MSG or monosodium glutamate, a salt of glutamic acid. High blood levels of this can cross the normally protective blood-brain barrier.

Glutamate is a neurotransmitter that is present in the extracellular fluid only in very low concentration. If levels are inappropriately raised through food ingestion then neurons fire abnormally, and at higher levels of contamination brain cells begin to die. Oxygen deficiency and lack of fuel (hypoglycaemia) interfere with the energy production of brain cells and make them susceptible to damage by excitotoxins. This may be an important factor in the development of neurological diseases.

Excitatory amino acids cause problems mainly when they are used either in high concentrations or in free form; while bound, as they are in most natural foods, they are only slowly released and therefore harmless. Most processed foods, however, contain excitotoxins, especially if any kind of commercial taste or flavour enhancer has been added, such as hydrolyzed vegetable protein, soy protein extract, yeast extract, beef stock, or caseinate; commercial soups, sauces, and gravies are usually most implicated. Beware, because on the label any of these products may be called “natural flavouring.” Aged proteins, processed meat, cheeses, and tomato purees have higher levels of free glutamate as well, but fresh tomatoes are fine. All of these products containing excitotoxins should be avoided by sensitive individuals and their use minimized by others.

Water Chlorination: Chlorination of our drinking water is yet another instance of an added substance (chlorine) that is detrimental to our health. Besides microorganisms in the water, all other organic matter present will be chlorinated. This produces a wide range of chlorinated chemicals with carcinogenic and other disease-causing properties.

Reluctantly, even health authorities now admit that there is a health problem, but they play it down by focusing only on trihalomethanes (THM’s). These are very small compounds, consisting of one carbon atom with several attached chlorine or bromine atoms. The most prevalent of these is chloroform (trichloromethane), a THM that is carcinogenic to rats and mice. You cannot protect yourself against THM’s, as they are absorbed through the skin if you swim in a chlorinated swimming pool or have a hot bath or shower with chlorinated water. A hot shower can raise the level of THM’s in the blood ten times higher than drinking a quart of chlorinated water.

While a large number of studies found an association between chlorinated water and secondary chlorination products with an increased risk of cancer, especially cancer of the bladder, I regard this only as the tip of the iceberg. The association of THM’s with bladder cancer shows that they are relatively harmless, because they are being excreted. I contend that the main health damage is caused by long-chain chlorination products that remain in the body.

The level of chlorine in treated water is about 100 times higher than the level of THM’s. There is no doubt that chlorine reacts with all organic matter, but especially with sensitive biochemicals, such as unsaturated oils, cholesterol, and vitamins. This reaction occurs when food is washed or cooked, when bread is baked, and when commercial fruit juices or soy milk are made, and it occurs within your body if you drink chlorinated water. These chlorination products have long carbon chains and remain stored in the body, just like chlorinated insecticides (e.g., DDT), to which they are closely related. In the same way that health authorities did not believe that ingesting pesticides would cause health problems, so they now believe that these long-chain chlorination products do not cause health problems. I regard this as being complacent and incompetent in protecting public health.

All or most chlorinated organic compounds that have been tested have been found to damage our health. It is also a fact that our liver is not equipped to detoxify chlorinated chemicals. Putting all of these facts together leads to the inescapable conclusion that chlorinated water damages our health to a considerable, although unknown, degree. In the next chapter, I provide more information on the connection between chlorinated water and cardiovascular disease.

Focusing only on THM's, health authorities maintain that the antimicrobial benefits of water chlorination outweigh its health dangers, and that there is no reasonable alternative to keep water supplies safe. I strongly disagree. I do not regard as "safe" water that causes cancer and very likely contributes to most other chronic and age-related diseases. There are thousands of water treatment plants in Europe that provide safe water without chlorine, using ozone, ultraviolet light, membrane filters, and slow sand filters instead. While this may be initially somewhat more expensive than using chlorine, in terms of preventing chronic diseases it will be immeasurably cheaper for individuals as well as the public purse.

In the meantime, you need to protect yourself from the dangers of chlorinated water. If you live in the country, you may have access to rainwater or bore water; otherwise use filters or reverse osmosis to purify water used for drinking and cooking, and install a shower filter. Distilled water needs an additional carbon filter to remove chlorine and THM's. For temporary use, when these are not available, you can remove most of the chlorine and THM's by briefly boiling the water and using it after it has cooled. Also, a bath can be filled with hot water and the room ventilated before getting into the bath. Skin contact with chlorinated water not only allows the absorption of THM's, it also ages the skin. You can counteract this by rubbing the skin afterwards with reduced ionized water or with a diluted vitamin C solution.

Water Fluoridation: Equally short-sighted is the official campaign to fluoridate our water supply. Modern dentistry tries to retard the growth of plaque bacteria that causes cavities by mechanical means and by poisoning the bacteria with fluoride. Fluoride is a strong enzyme poison. When present in sufficient concentration during development of the teeth, it is incorporated mainly as insoluble calcium fluoride. The lactic acid produced by plaque bacteria after the ingestion of sweet food dissolves traces of calcium fluoride from the teeth and this poisons their enzyme systems, retarding the development of caries.

However, the fluoridation of our water supply and the liberal use of fluoride toothpaste introduce dangerously high concentrations of fluoride in a very active and toxic form. This endangers our enzyme systems and greatly -contributes to a weakened metabolism in many people, and thus to the development of allergies, arthritis, cancer, and hypoglycaemia. Excessive fluoride exposure has also been linked to hip fractures, hypothyroidism, Alzheimer's disease, and reduced IQ. The bulk of scientific research shows that added fluoride does not protect the teeth of adults or children.

In natural groundwater, such as bore or spring water, fluoride sufficient for tooth protection (up to one ppm) occurs mainly in the form of calcium fluoride. While this is insoluble in high concentrations, it is sufficiently soluble in the minute amounts required for tooth protection. In this form, and also when the blood level of calcium is high, fluoride is less toxic. Fluoride causes the most damage in calcium-deficient bodies, usually children and hypoglycemics or fast oxidizers.

Seafood and especially kelp are high in a safe form of fluoride. However, using these in addition to fluoridated water and toothpaste might exceed safe fluoride intake levels. Caries, and the health problems due to artificial fluoridation, can be avoided by using kelp and restricting the use of sweet foods.

The United States is one of a very few countries that still promote water fluoridation. Most of Western Europe has outlawed it; even most of U.S. communities, when allowed to vote, have rejected fluoridation; this has happened in more than 50 communities since 1999. (For further information on the health dangers of fluoride and a good bibliography of the scientific literature on fluoride, see: www.fluoridealert.org.)

Aluminium: Another mineral harmful to our health that is added to our drinking water to precipitate impurities is aluminium. This is not a so-called heavy metal, but it causes health problems nevertheless. We can also ingest it with some brands of baking powder, free-flowing salt, or aluminium hydroxide antacids, but most commonly we absorb it through using aluminium cookware and utensils. Most dangerous to our health are acidic foods that are left in contact with aluminium surfaces. Aluminium easily dissolves in weak acids, which are present in fruits, tomatoes, cucumber, rhubarb, and beets. Even heating tap water in an aluminium pot was shown to add 1600 mcg of aluminium per litre of water, which is 3200 percent over the recommendation of the World Health Organization, which sets a limit of 50 mcg per liter.[33]

In the body, aluminium acts as a neurotoxin, causing brain disorders, though symptoms develop slowly. They may include senile dementia (Alzheimer’s disease), Parkinson’s disease, reduced memory, slow learning, motor neuron disease, overactive parathyroid glands, and fatigue. Aluminium lowers the permeability of the blood-brain barrier to other harmful chemicals and increases the incidence of epileptic seizures. You can protect yourself from aluminium ingestion by using filtered water and glass, ceramic, or enamelled cookware. I do not recommend stainless steel, as nickel can leach out during cooking, especially with acidic water or food, and nickel is known to be carcinogenic.

Mercury: Chronic mercury poisoning can originate from certain pesticides and fungicides, calomel laxatives, predatory fish, and especially amalgam tooth fillings. Such fillings release mercury from broken-off pieces that dissolve in the stomach and also from evaporation in the mouth that is caused by pressure on the fillings during chewing. Old fillings are thus much lighter than new fillings because they have lost most of their mercury. Acidic foods increase the rate at which mercury fillings disintegrate. The mercury loss from several amalgam fillings over ten years has been calculated at about 560 mg, or 150 mcg per day, and almost all of this is absorbed and stored in body fat, especially in the brain.

Symptoms of mercury poisoning include: fatigue, lethargy, depression, irritability, shyness, loss of short-term memory, allergies, respiratory infections, decreased immune response, increased salivation, gingivitis, loose teeth, mouth ulcers, nervous heart problems, digestive problems, and dermatitis, as well as changes in hearing, vision, speech, and writing. Most prominent are nervous system disorders with serious mental disturbances. Many symptoms are actually the same as in multiple sclerosis. The present epidemic of autism has been linked by some researchers to a high level of organic mercury compounds used as preservatives in some childhood vaccines.

Case reports for patients after removal of all their amalgam fillings include cure of leukaemia and Hodgkin's disease, regaining vision after 23 years of blindness, cure of moderate to severe headaches that had persisted for 50 years, cure of arthritis, and a reversal of general malaise with fatigue and depression. Minimize amalgam-related health problems by refusing any new fillings containing mercury, and contemplate replacing any old amalgam fillings; for this, see a holistic dentist.[34] For more details on dental health, review Step 8. For practical advice and support, contact: www.amalgam.org; for wide-ranging scientific information on the negative health effects of amalgam fillings and heavy metals: www.melisa.org. An independent researcher, Bernie Windham, has assembled hundreds of clinical studies on the problems caused by amalgam fillings and harmful metals at: www.home.earthlink.net/~berniew1/.

Mercury exposure from eating fish has greatly increased in recent decades. Even health authorities acknowledge that there is now a problem and advise avoiding large predatory fish, especially for pregnant women. Some studies show that on average more mercury is accumulated from fish than from amalgam fillings. The same health problems caused by dental mercury are caused by mercury from fish. Only non-predatory fish or small predatory fish are reasonably safe to eat on a regular basis. You can check the safety of commonly used fish species at: www.ewg.org.

The Pros and Cons of Cooking: Cooked food has been a hallmark of civilization. Cooking breaks down cellulose in vegetables and connective tissue in meat and thus makes these foods easier to chew and more palatable. Bacteria are destroyed at the same time, making these foods safer to eat, but only if eaten immediately after cooking. During storage, raw food remains much more resistant to microbial contamination than cooked food.

The price we pay for this convenience is a loss of vitamins and minerals, the destruction of enzymes and plant hormones, denaturing of proteins, oxidation of lipids, loss of vitality or life force, and formation of harmful chemicals at temperatures above the boiling point. Vitamins are destroyed by the combined action of oxygen and heat; minerals react to heat in ways that make them less easily absorbed; and both vitamins and minerals are lost if cooking water is discarded.

When food is heated above 180° F, protein becomes denatured or unbiological; it changes its internal structure (the way it is folded), and the immune system reacts as if it were attacked by foreign invaders. The digestive system is flooded with white blood cells after eating this food. This is called digestive leukocytosis: The white blood cells or leukocytes try to prevent food toxins from reaching the bloodstream. The higher the food is heated or the more refined it is, the stronger is the resulting leukocytosis. This defence reaction against toxic food does not take place after eating raw food. Leukocytosis is a continued strain on our immune system and is an aggravating factor in leukaemia and immune deficiency. Animals and humans evolved on raw food only; cooked food is a recent invention in evolutionary terms.

When food is heated above the boiling point, chemical reactions take place that create toxic chemicals in the food. For a long time it has been known that carcinogenic substances are produced in highly heated meat, but the more recent discovery of similar chemicals in heated cereal or grain products has come as a shock to the food industry. When foods are heated to above the boiling point, advanced glycation end products (AGE's) are formed, some known as acrylamides. These toxic products result from a reaction of carbohydrates with proteins. They are produced at a slow rate at low temperatures, but in high amounts when food is highly heated, as in frying and baking. Acrylamides are known to be cancer-causing in animals and to cause DNA damage with gene mutations.

Another scientific finding is that acrylamides are pro-inflammatory and the more these chemicals are in the diet, the more inflammation they cause. Almost all diseases are associated with inflammation, including arthritis, cancer, diabetes, and heart disease. The more inflammation, the worse are the disease and associated pain. This may be a main reason why most diseases greatly improve and chronic pain tends to disappear on raw food diets.

AGE's are yellow-brown, like another group of chemicals related to heat and oxidation, the lipofuscins, which are composed of oxidized fats and proteins. Together, these chemicals accumulate as waste material in nerve and skin cells and impede their functions. Sometimes they can be seen as age spots or liver spots.

Equally damaging is the total destruction of all the enzymes in cooked food. Raw food contains a complement of enzymes specifically designed to digest it. Starchy foods contain large amounts of the starch-digesting enzyme amylase; fatty and oily foods are provided with the fat-digesting enzyme lipase, while protein foods contain the protein-digesting enzyme cathepsin. According to the food composition, there may be mixtures of these and other enzymes.

If food enzymes are destroyed, we must produce considerably more digestive enzymes, and our overall enzyme capacity declines much more rapidly with advancing years than it would on raw food. Thus we age much sooner and develop degenerative diseases that can otherwise be avoided. Food enzymes and digestive enzymes not only break down and absorb food, but also keep the body free of inappropriate accumulations of fats and proteins and help us avoid or reverse atherosclerosis and cancer. High-dose supplements of protein-digesting enzymes are frequently used in natural cancer therapy to break down tumours.

Animal experiments show the profuse development of degenerative diseases from a diet of cooked foods and that good health and longer life result with a raw diet. There are also numerous reports of human cures from all kinds of degenerative diseases - including advanced cancer - on a raw diet of organically grown food. My own successes in healing patients in advanced and difficult conditions have often been due to raw food diets.

Modern processed foods generally combine the negative health effects of cooked, refined, and chemicalised food. Tinned food, for instance, causes even more leukocytosis than normal cooked food, indicating that it is more toxic. Equally harmful is microwaved food, as this process violently shakes and twists the delicate molecular structures. In studies, abnormal changes in the blood and immune system were noticed after ingesting microwaved food.

Fortunately, it is easy to protect yourself from the dangers associated with cooking: avoid or minimize the use of processed food; heat food only to the boiling point; and maximize your intake of raw, fresh, and organic food. Shortly before eating heated food, have some fresh, raw food, and more with the cooked meal. In this way, digestive leukocytosis can be reduced or even eliminated.

I have noticed that for many people cereals are especially difficult to give up. These cause problems for many individuals, contributing to overweight, diabetes, and hypoglycaemia. How-ever, some individuals and therapists report that raw grains do not seem to cause these problems. Therefore, if you are so inclined, you can experiment with soaked non-gluten grains as an addition to breakfast or a salad.

Harmful Eating Habits: The way we eat our food can also negatively affect our health. Commonly, food is not sufficiently chewed as our mind is busy with other matters. Thus, the food is not properly broken down and mixed with the starch-digesting amylase in the saliva. This adds to the burden of an already overworked pancreas. The composition of the required digestive juices is already signalled to the pancreas from the mouth, but if food is eaten too fast, especially cooked food, the signals are inaccurate, and we also do not receive the signal when we have had enough, so we tend to overeat.

Overeating is unhealthy because we become overweight, but also excessive amounts of metabolic waste products are formed that stress the enzyme systems and the organs of elimination. In addition, food remains partly undigested in the intestines, which encourages putrefaction and an overgrowth of harmful microbes. The overall effect can be not only a shortened life span but also early onset of degenerative disease. Many animal experiments have confirmed that life span as well as health can be greatly enhanced if the quantity of food consumed is restricted. Rats, for instance, doubled their life span on half their normal diet and earthworms lived 16 times longer on a starvation diet, according to studies.

A famous example of a long and healthy life through restricted eating is the Italian nobleman Luigi Cornaro, who lived about 400 years ago. He was born with fragile health and by the time he was 40, he had severe arthritis and many other afflictions and was close to dying. He cured himself within one year by restricting his food intake to about 12 ounces daily, consisting of bread, meat, egg yolk, and soup. He continued with this in good health until his 78th year when he gave in to the persuasions of his family and increased his food intake by two ounces. Within days, he became very sick. Again he escaped death by returning to his frugal eating habits. From then on he lived in good health, with excellent mental faculties and without fatigue until his 99th year. In his last years, he wrote a comedy and treatises on longevity and health.

Eating meat or a large meal in the evening is especially unhealthy because then our digestive powers are much weaker than in the morning. What we eat during the daytime is usually converted into energy for the activities at hand, but the evening meal is mainly converted into fat deposits and an increase in the accumulation of wastes.

Equally harmful is indiscriminate food combining. High-protein foods such as meat require different digestive juices from those needed to process a starch meal. The digestion of starch initiated by the salivary amylase is prematurely stopped if the starch is mixed with meat in a meal. Amylase is also inhibited when starches are mixed with acids and the release of salivary amylase is suppressed when eating sweetened starches. Mixing incompatible foods overloads the pancreas and our health suffers.

The Mixed Benefits and Deficits of Stimulants: Stimulants such as coffee, tea, cocoa, tobacco, alcohol, and various recreational drugs are widely used and often overused. I do not suggest that stimulants should be completely avoided but, similar to sweet foods, I contend it is much better for our health if we restrict their use.

Whether stimulants are harmful to you and to what extent depends greatly on your state of health and how frequently you use them. Someone who habitually takes a specific stimulant several times daily will be most affected by it and is likely even to be allergic and addicted to it. This applies to coffee and tea as much as to alcohol, tobacco, and other stimulating drugs. While the health dangers of alcohol and tobacco are well publicized and generally appreciated, the prevalence of allergies to coffee and tea is often underestimated. The most common symptoms are headaches, migraines, irritability, and menstrual problems.

However, coffee and tea can also make positive contributions to our health, provided that we are not addicted and/or allergic to them. Tea, especially green tea, is high in antioxidants and useful in the prevention of cancer. Black and green teas stimulate the respiratory system and are helpful with breathing problems, such as asthma. Coffee, on the other hand, acts mainly on the brain, not only to keep us awake, but it can also prevent a threatening migraine attack. Coffee drinkers appear to have a lower incidence of Parkinson’s disease. A dry red wine taken with a meal can be beneficial, and beer in moderation is fine (provided you have no gluten or yeast sensitivity). Unfortunately, modern beer and many wines are chemicalised and not beneficial.

Marijuana is probably no worse than other recreational drugs and has some useful medicinal properties along with its strongly relaxing and pain-relieving effect on the body and its stimulation of specific parts of the brain. Habitual use is most harmful for those who are already too relaxed and without much drive or enthusiasm. Generally such people have weak blood sugar regulation, lack of energy, and low blood pressure. Prolonged use of marijuana can cause genetic changes and infertility in offspring.

Signs of addiction to stimulants are a craving during periods of abstention or bingeing when a tempting opportunity arises. In order to improve your health, avoid the addictive stimulant for a very long time, preferably for several years. People in poor health, especially if their system is acid and they have a disturbed sugar metabolism, easily become addicted and allergic to any stimulant that is used for a prolonged period.

If your health is reasonably good, it will not be too damaging to use stimulants in moderation. Someone in good health rarely becomes addicted. The main thing to watch out for is that it not become a habit or a craving, and also that alcohol not be used to excess. However, smoking is almost always an addiction and has no redeeming health features at all.

If you use a stimulant regularly, more than once or twice a week, I recommend that you stop using it for one week every few months and carefully observe yourself during abstention and when it is reintroduced for signs of craving, discomfort, or allergy. This advice to interrupt stimulant use occasionally does not apply, however, to regular smoking, hard drugs, or binge drinking. For these, permanent abstention is the only solution. Pregnant and nursing women are best advised to abstain completely from alcohol, tobacco, and drugs.

Yeasts and Moulds: Yeasts, moulds, and mushrooms are all fungi. Yeasts are widely used in baking and brewing and also as concentrated yeast extracts and food yeasts. Moulds are used in some cheeses and to produce antibiotics. We all have some yeast cells resident in our intestines. Normally, these are harmless, as the intestinal bacteria keep them subdued. However, in many people the fungi have taken over, and the overuse of prescription drugs and improper dietary changes have contributed to this condition.

Yeast proliferation is stimulated by birth-control pills and consumption of sweet food, as sugars are the only energy source for yeasts, but the main reason that it has become a widespread health problem at present is the general use of antibiotics and prescription drugs that suppress the immune system. Antibiotics decimate the normal intestinal bacteria but leave the yeast cells unaffected, allowing them to proliferate and take over the space normally occupied by intestinal bacteria. From the intestinal tract, the yeast cells can invade other mucous membranes, especially those of the vagina, mouth, and lungs; they can also penetrate the intestinal wall and enter the bloodstream.

The most common form of intestinal yeast is Candida albicans, an overgrowth of which can cause gastritis, colitis, cystitis, bloating, oral (thrush) or vaginal infestation, menstrual problems, and asthma. Allergies with a wide variety of symptoms result when this yeast penetrates the intestinal wall. In addition to skin and breathing problems, severe mental and emotional reactions, especially depression, are common. Autism, neuroses, and mental diseases, including schizophrenia, have reportedly been cured with antifungal therapy. Usually in such cases, the patient has a history of Candida problems and antibiotic use. The toxicity of Candida to the nervous system is largely due to an excessive production of acetaldehyde. This substance is formed when yeast cells metabolize sugars and there is a deficiency of oxygen.

Another important aspect of systemic fungal infestation is severe damage to the immune system, leading to autoimmune diseases and immune deficiency states, including lupus erythematosus, rheumatoid arthritis and, most likely, AIDS. After the immune system has been severely damaged by this fungus, it is easy for bacteria and viruses to invade the body and cause serious and even lethal diseases.

Dietary yeasts and moulds greatly increase the difficulties of susceptible individuals, frequently causing allergic reactions and flare-ups of Candida infections. Some moulds can damage the liver, such as aflatoxins, from a mould that frequently grows on peanuts and is present in peanut butter. Moulds commonly grow on dried fruits, on poorly stored grains and nuts, on the outer leaves of cabbage, and on the skins of various fresh fruits.

A study in the U.K. in 1977 found that moulds and their poisonous mycotoxins (fungal toxins) were present in all 318 samples of flour tested. This mould contamination was believed due to insufficient drying in the process of combine-harvesting of grain. Whole-meal flour, bran, and wheat germ are even more affected by mould than is white flour; rice is also frequently contaminated.

Even if the fungi are killed during cooking, the products of their decomposition can still cause allergies. Washing whole grains, nuts, and susceptible fruits and vegetables before cooking or eating helps to remove moulds; drying them in the sun destroys fungi and prevents their development. Preferably also remove environmental moulds, such as those on walls and bathroom tiles, because they can be inhaled and cause lung and breathing problems in susceptible individuals (asthmatics).

Frequently, antibiotics are routinely added to animal feed for meat production; this applies especially to poultry and pigs. Antibiotic residues can impair the intestinal flora of the consumer of such products and thereby encourage the spread of Candida. In addition, the meat itself can be infested with Candida or other fungi because of the prolonged use of antibiotics in the animals; resistant strains of harmful bacteria may also be present. I do not recommend habitually eating such meat. As a general rule, I recommend you avoid meat from commercial animal feedlots.

* * * * * * * * * * * *

Step 43

METABOLIC TYPES, PATHWAYS, & DIETS

Know your individual metabolic type and your specific dietary requirements.

Some foods, such as meat or sweet and acidic foods, are good for certain people, and for these people they are not problem foods, while with other people their effects are adverse. To gain a better understanding of this phenomenon and insight into the food-related causes of diseases, let's look at metabolism.

Our metabolism is the total of all the millions of biochemical and enzymatic reactions that take place in our body at any given time. Some reactions use energy to build (anabolism) more complex structures from assimilated nutrients, while others release energy by breaking down (catabolism) more complex molecules.

While there are additional factors and overlapping influences, we can say generally that our metabolic pathways are most strongly influenced by our nutrition, that our glandular system is mainly affected by our emotions, and that our nervous system responds primarily to brain and mental activity. All three factors together - nutrition, glandular system, and nervous system - regulate our metabolism.

Inherited or acquired variations of these regulating factors produce a unique blend for each individual and call for different specific requirements for optimal health. However, there are some common key factors and we can distinguish several groups with similar characteristics: Each group represents a metabolic type, which I explain below. But first, two concepts must be explained: the autonomic nervous system and the tension-relaxation cycle.

The Autonomic Nervous System: The autonomic nervous system (ANS) has a great influence on our metabolism. It regulates those body processes that usually function automatically without our conscious control, such as heartbeat, digestive processes, nerve reflexes, and so forth. It has two branches with opposite characteristics: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS).

The main function of the SNS is to prepare the body for action: epinephrine pours into the bloodstream; the muscles tense; blood pressure, pulse, and breathing rate increase as well as the rate of catabolic processes in general. The digestion, however, is slowed or interrupted. The SNS is activated by emotions such as anger and fear, by the intention to take some action, but also by certain foods, and most strongly by red meat.

The PNS relaxes the muscles, improves the digestion, and lowers blood pressure, pulse, and breathing. The PNS is dominant during sleep and with relaxed body and mind conditions. The endings of the PNS nerves release acetylcholine when stimulated, while SNS nerves release norepinephrine, an epinephrine-like hormone with opposite effects to those of acetylcholine. SNS dominance leads to elevated blood levels of glucose and fatty acids (diabetes, coronary heart disease), while with PNS dominance these levels are lowered (hypoglycaemia).

The Tension-Relaxation Cycle: Overall, each body function should be in balance between the opposite forces of the SNS and PNS. However, at any given moment, one force will usually prevail; during action the SNS dominates and during rest the PNS. A healthy body is not static but in a dynamic balance, like a swinging pendulum.

Problems arise when the tension-relaxation cycle is not balanced but is one-sided. During work there may be too much tension, while the following rest period does not result in complete relaxation. If this condition becomes chronic, other tension-related symptoms can develop, such as high blood pressure or restlessness. To bring the body back into balance, remedies with a PNS influence can be applied, such as relaxation exercises, a vegetarian diet, or magnesium supplements. On the other hand, if the body generally is too relaxed, too weak, and too soft for the daily tasks, and mind and emotions cannot be aroused, then tension-producing SNS influences can be applied, such as consuming flesh foods or calcium supplements.

Basic Metabolic Types: From these considerations about our ANS, we can describe three metabolic types: the P-type in which the PNS dominates; the S-type with a dominating SNS; and a balanced type. Population groups over thousands of years developed predominantly into one of these types, depending on climate and available food. These inherited ancestral metabolic characteristics remain fairly stable during our lifetime and form our basic metabolic blueprint on which acquired metabolic traits are superimposed.

The fundamental difference between the P-type and the S-type can be illustrated if we compare the mental picture or stereotype we have of the Native American with that of a desert Arab.

The Native American is deliberate and measured in movement and speech, stoic, and rarely displays emotions. He or she has a powerful body and a broad face, is a hunter, and eats relatively few carbohydrates. This is a typical representation of the positive aspects of the P-type. We find the same characterization, only adjusted to a colder climate, in the Eskimo. The typical desert Arab, on the other hand, is agile, lively, energetic, and quick to act on feelings and emotions. He or she is lean, with a narrow face, and can (or reportedly could in former times) exist for long periods on nothing but dates and still maintain remarkable endurance. This portrait shows the positive aspects of the metabolic S-type.

The apparently unbalanced ANS of each type is actually balanced through the opposite influence of the traditional diet. Thus, the high-meat diet of the Native American stimulates the SNS and epinephrine production, while the predominantly vegetarian diet of the Arab sedates the SNS and epinephrine response. If the Native American and the desert Arab were to interchange their diets, they would become unbalanced, their energy metabolism would eventually become inefficient, and their health would decline. They would then display the negative characteristics of their metabolic types.

Other races and nations that are predominantly of the P-type are the North Europeans, Dutch, Russians, Polynesians, and New Zealand Maoris, the latter being a traditional hunter society in a warm climate. The S-type, on the other hand, is mainly grouped around the Mediterranean and the tropical and subtropical regions of Asia. Generally, the P-type originated in cold climates that offered a continuous supply of animal products but a lack of plant food in winter, while the S-type evolved with a continuous, plentiful supply of plant food. There are, however, no original populations that live entirely on animal or plant food.

Most of those with middle-European ancestors are of the balanced metabolic type, readily utilizing all three metabolic pathways: carbohydrates, fats, and proteins. However, sweet foods were relatively rare and reserved for special occasions in those metabolically formative societies.

Table 5-3: Characteristics of Metabolic Types

S-Type	***P-Type***
tendency to raised blood pressure, fast pulse and breathing	tendency to low blood pressure, slow pulse and breathing
slender, slight body frame, narrow face, flat chest	strong bone structure, broad face, rounded chest
dry mouth; ears and gums pale; skin dry, yellowish	much saliva, ears and gums pink, ruddy complexion, flushes
weak dreaming activity	much dreaming activity
sprinter type, not much endurance, fitness fanatic	long-distance runner, steady worker, good endurance
appetite diminished, digestion slow and weak	appetite above normal, digestion fast and strong
enjoys fruits and vegetables, dislikes fatty food	likes meat, starches, and fatty food
emotionally unstable, freely shows emotions	emotionally stable,
quick to make decisions	slow to make decisions, cautious
seldom sad, depressed, melancholic, or withdrawn	tendency to being sad, depressed, withdrawn, melancholic
extroverted, easily angered or irritated	introverted, rarely angered or irritated
tense, lively, overactive, jumpy	relaxed, quiet, slow or sluggish
strongly assertive, aggressive	non-assertive, nonaggressive
hate, envy, verbal abuse	resentment, guilt, worry, doubt
Slow Oxidizers	***Fast Oxidizers***
blood pressure raised or high	blood pressure low
often heavy body and hard manual worker	often overweight or underweight and low in energy
often too alkaline, insensitive to cold and skin irritants	overacid, cold hands and feet, skin very sensitive
enjoys meat and heavy food but also fruit and sour food	craves sweet food but feels better craves sweet food but feels better after protein and salty food
emotionally unresponsive, undemonstrative	emotionally unstable, irritable, outbursts, shy, resentful
rarely has colds, infections, allergies, or inflammations, susceptible to chronic infections, liver and kidney problems, cancer, cardiovascular disease, stroke	susceptible to colds, allergies, inflammations, acute infections, rheumatoid arthritis, anemia, lymph gland diseases, mucus problems, leukaemia

Table 5-3 lists the characteristics of the P-type and S-type. Those of balanced type are generally between these opposites. The positive characteristics of the P-type are described as being relaxed, dependable, quiet, cautious, and emotionally stable, with good skin circulation, a ruddy complexion, and good endurance. Pulse, breathing rate, and blood pressure tend towards the low side. The basic traits of the S-type are that individuals have plenty of drive, freely show their emotions, and are lively, assertive, and quick to make decisions. The body tends to be lean; pulse, breathing rate, and blood pressure may be elevated.

Converted Metabolic Types: Inherited weaknesses and deteriorating health can cause far-reaching changes in our metabolic characteristics, sometimes requiring a diet that is similar to that of our opposite basic type. The actual cause of such changes can be an overstimulation or weakening of the SNS or the endocrine glands; it can also be a partial blockage of a metabolic pathway because of vitamin and mineral deficiencies. Most frequently, however, our metabolism has deteriorated because we overuse problem foods.

Slow Oxidizers: Few individuals in our society still display all the positive characteristics of their basic metabolic type. Many of us live in a stimulating social environment and use stimulating foods at the same time, such as red meat, spices, alcohol, coffee, tea, tobacco, or drugs. All these factors stimulate the SNS and our endocrine glands. We become nervous, restless, sleepless, quick-tempered, and overaggressive, with raised blood pressure and permanent muscle tension. Such individuals are now negatively SNS-dominant and can be classified as negative S-types, even if they originally had a balanced or even PNS-dominated metabolism.

Other negative characteristics can develop, such as being underweight, having a dry mouth or yellowish skin, poor appetite with a weak, slow digestion, emotional instability, or an uncontrolled temper. Deteriorating liver and kidney function in combination with a high-meat diet can lead to a buildup of uric acid and cause gout and rheumatism. Irritation or ulcers of the digestive tract may be present, and sometimes migraines can develop as well. If such overstimulation persists for a long time, especially in combination with a diet high in red meat and fat, then the liver and glandular system gradually become weak and the metabolism slows down. This is the condition of a slow oxidizer.

The main characteristics of slow oxidizers are raised blood pressure and alkaline body fluids. Together with a raised calcium blood level, this frequently leads to kidney stone formations. Often, slow oxidizers originate from the P-type and are manual workers, such as farmers and builders, people with a heavy body and bone structure and who usually eat large amounts of meat.

The skin of slow oxidizers is usually insensitive to cold and irritants. Emotionally they are stable and even unresponsive and undemonstrative; they do not show or experience much in the way of strong feelings and emotions. They are slow, dependable, and reliable. If they originate from the basic S-type, they are lean and show more emotional response. Common health problems are hypertension and cardiovascular diseases, chronic infections, degenerative liver and kidney diseases, stroke, and cancer.

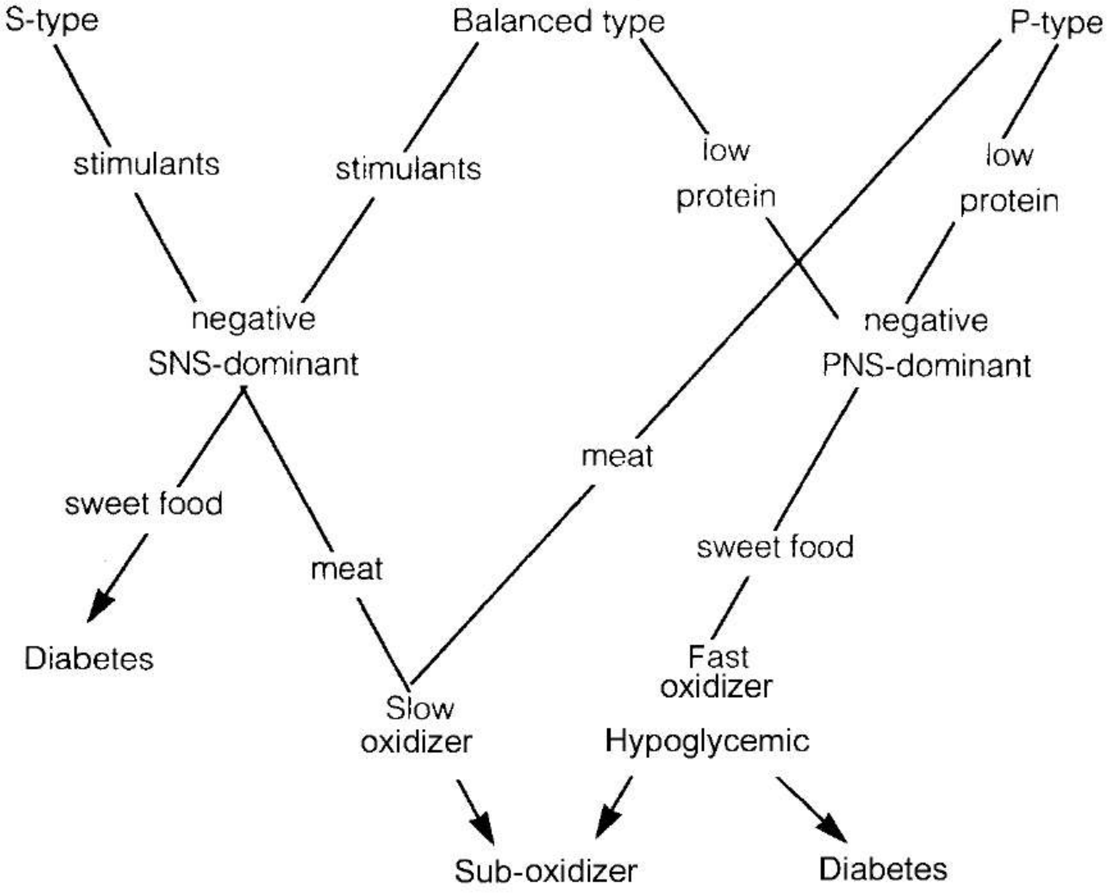
S-type
Balanced type
P-type
stimulants
stimulants
low protein
low protein
negative SNS-dominant
negative PNS-dominant
meat
sweet food
meat
sweet food
Diabetes
Slow oxidizer
Fast oxidizer
Hypoglycemic
Sub-oxidizer
Diabetes

Figure 5-1: Relationships between Metabolic Types

Fast Oxidizers: The opposite metabolic path is exhibited by those of the balanced type and the P-type who live on a low-protein vegetarian diet or who are protein deficient because of malabsorption or liver problems. They become negative P-types, displaying the traits of negative PNS dominance. They lose their normal aggressiveness, enthusiasm, and self-assertiveness, and they become sluggish, listless, and underpowered. They may become unemotional and lose interest in the happenings around them; marijuana use makes this condition worse.

This metabolic type is widespread in many underdeveloped countries, but also in alternative lifestyle groups in Western society. Women are more prone to this condition than men as they naturally have lower epinephrine levels. Other negative features that can develop are low blood pressure with a lack of energy; the weight becomes too low or too high; the individual easily becomes sad, melancholic, or depressed and has difficulties making decisions. Anemia may be present.

While S-type individuals can become diabetic on a diet high in sweet foods, P-type individuals tend to become hypoglycaemic. Glycolysis (the breakdown of glucose inside the cells) is sped up and this is why an individual is called a fast oxidizer.

Fast oxidizers usually have low blood pressure with low energy levels, especially after eating sweet foods and fruits. Often there is a craving for sweet food though they feel much better after eating proteins. Allergies, addictions, and overacidity are common, as well as susceptibility to acute infections and inflammations. The circulation tends to be poor with cold hands and feet. Respiratory tract infections and other mucous diseases are common, as are rheumatoid arthritis and a tendency to leukaemia. The fluctuating blood sugar level in combination with allergies causes emotional instability. Individuals may be irritable with occasional bouts of temper and even violence, but more often they are resentful and shy, and frequently worry or feel guilty.

Sub-Oxidizers: With advancing age or deteriorating health, both branches of the ANS tend to become weak and all aspects of the metabolism become inefficient. This is the condition of a sub-oxidizer, and is common to all metabolic types whose health has degenerated.

It appears to be rare for our health to degenerate in a straight line from a positive balanced type to the negative balance of the sub-oxidizer. It usually includes the condition of the fast or slow oxidizer and sometimes conditions of both, first as fast oxidizer and then as slow oxidizer.

Sub-oxidizers represent the typical condition of old age in our society. The endocrine system is nearly exhausted, the immune system is weak, and digestion is inefficient, with malabsorption and lack of gastric acid and digestive enzymes. The metabolism is inefficient, so the body usually is too alkaline. The blood pressure may be too high or too low. The circulation is extremely poor and there is a pronounced lack of energy and even debility. There are many chronic health problems and frequently a serious degenerative disease, such as the deformative stage of arthritis, senility, or cancer.

See figure 5-1 for a diagram of the relationship between the various metabolic types. The opposite directions to those indicated in the diagram are typically travelled during a program of health improvement. For the purpose of selecting your metabolic type, you can regard a blood pressure of 120/80 and a pulse of about 70 per minute while sitting as normal for an adult; children usually have a lower blood pressure and a higher pulse rate.

Biochemistry and nutrition are commonly taught as if the basic metabolic pathways for energy production are used in the same way by all people, except perhaps for some uncommon hormone and enzyme deficiencies. This, however, is an oversimplification that leaves the health practitioner as well as the patient without clear guidance as to how to use nutrition to overcome disease and maintain health.

A more relevant study of our energy metabolism (the way in which our cells produce energy from food) provides some fascinating insights into the reasons why certain diets and nutrients are beneficial or harmful for various diseases and for different metabolic types. While part of this study may seem technical and difficult to understand, it is worth the effort because here we have the basis of all the disease-producing and healing effects of nutrition.

Glycolysis and Citric Acid Cycle: The main energy-generating mechanism in our cells is the Krebs cycle or citric acid cycle. This is a series of enzymatic steps through which a two-carbon molecule (acetyl or acetate) is oxidized to yield carbon dioxide and water. All three main food components - glucose, amino acids, and fatty acids - can be utilized in this way.

Glucose is normally the most important fuel molecule. Through several reactions along the glycolytic pathway, it is split in half to form pyruvic acid. Niacinamide and vitamin B1, as coenzymes, are the necessary vitamins at this stage; magnesium is also required. Glycolysis proceeds anaerobically, that is, without oxygen. See figure 5-2 for a diagram of cellular energy metabolism. The next step may be in one of two directions, and it is here that most intracellular energy problems develop. Pyruvic acid can combine with carbon dioxide and become oxaloacetic acid, or it can lose carbon dioxide and form acetyl coenzyme A. Combining with carbon dioxide requires a high amount of energy as well as biotin, nicotinamide, and manganese. The step in the other direction needs oxygen and magnesium, as well as the vitamins B1, niacinamide, and pantothenic acid.

If the metabolism is in good condition, then both reactions take place in such proportions that there are always equal amounts of oxaloacetic acid and acetyl coenzyme A available. Both molecules will then combine to form citric acid as the first step in the citric acid cycle. During several more steps in this cycle, together with reactions that follow up the cycle, the acetyl component is completely oxidized into two molecules of carbon dioxide and water, while the oxaloacetate is reformed, ready for a fresh turn of the cycle.

However, in practice, usually a deficit of oxaloacetic acid will develop, because this is a very useful molecule that is required for many other reactions, such as to form amino acids or nucleic acids. The body covers this deficit by converting an appropriate amount of pyruvic acid into oxaloacetic acid. Important for the operation of this cycle are the vitamins B1, B2, and niacinamide, in addition to magnesium and oxygen.

In this way, someone of the S-type (“desert Arab”) can obtain all his energy by burning carbohydrates. The P-type (“Native American”), on the other hand, will create very little of her energy in this way. She produces energy mainly by forming acetyl coenzyme A directly from fats (requiring vitamin B2, niacinamide, and biotin) and from amino acids (requiring vitamin B6). Most of the small intake of carbohydrates will be converted to oxaloacetate.

The balanced type is able to use both main pathways efficiently and best gains energy from a mixture of carbohydrates, proteins, and fats. The third path, the conversion of pyruvic acid into lactic acid, is an emergency measure that can be used by all metabolic types. This is an anaerobic step, which in most instances is triggered by a shortfall of oxygenating enzymes, either because of strenuous muscle activity or resulting from an allergic reaction. In fast oxidizers, the buildup of lactic acid can also be caused by an accumulation of metabolic acids (e.g., citric acid or malic acid) in the cells. Lactic acid is partly neutralized and expelled with the urine and partly reconverted in the liver to pyruvic acid in the presence of sufficient nicotinamide.

Abnormal Glucose Metabolism: The P-type and the balanced type tend to become fast oxidizers and hypoglycemics on a diet high in sugars and low in proteins and fat. Most of the dietary protein is used as building blocks and little is available for energy production. The ability to convert pyruvic acid into acetyl coenzyme A is poorly developed in P-type individuals. Therefore, oxaloacetic acid is produced in excess, while lack of acetyl coenzyme A prevents the citric acid cycle from fully operating. The blockage of the citric acid cycle leads to a lack of energy and overacidity from accumulating metabolic acids and lactic acid.

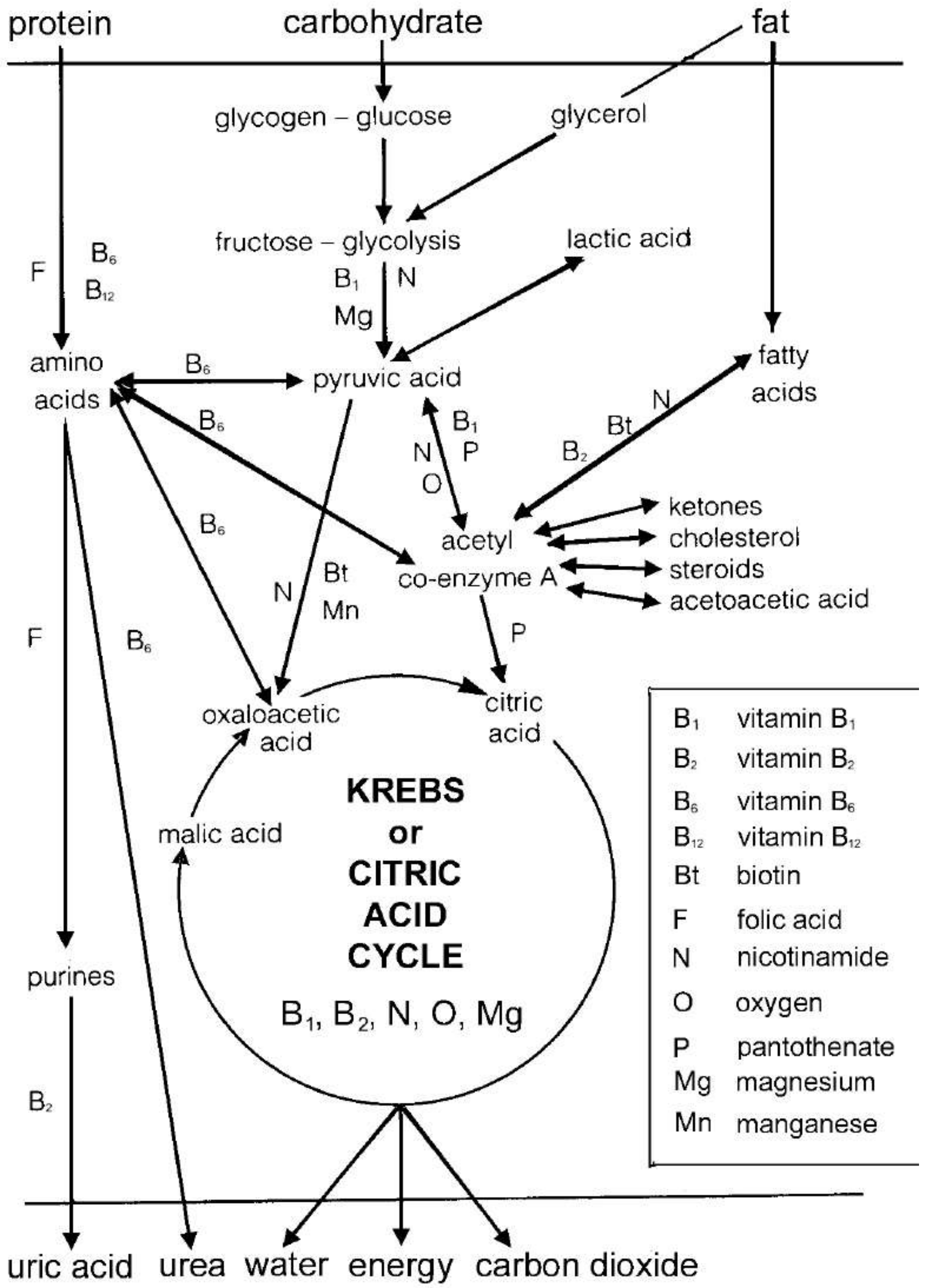
protein
carbohydrate
fat
glycogen – glucose
glycerol
fructose – glycolysis
lactic acid
B_1
N
Mg
F
B_6
B_{12}
amino acids
B_6
pyruvic acid
fatty acids
B_6
B_1
N
P
O
Bt
N
B_2
B_6
acetyl co-enzyme A
ketones
cholesterol
steroids
acetoacetic acid
N
Bt
Mn
P
F
B_6
oxaloacetic acid
citric acid
malic acid
KREBS or CITRIC ACID CYCLE
B_1, B_2, N, O, Mg
purines
B_2
B_1 vitamin B_1
B_2 vitamin B_2
B_6 vitamin B_6
B_{12} vitamin B_{12}
Bt biotin
F folic acid
N nicotinamide
O oxygen
P pantothenate
Mg magnesium
Mn manganese
uric acid
urea
water
energy
carbon dioxide

Figure 5-2: Cellular Energy Metabolism

Consuming more proteins and fats (best here is olive oil) is the solution, as fats are the most efficient source of acetyl coenzyme A. However, this will be problematic if there is malabsorption of fat. Polyunsaturated oils, on the other hand, require additional metabolic steps and are not well suited for energy production. In addition, linoleic acid may contribute to the frequent oversensitivity of fast oxidizers by forming certain prostaglandins that increase inflammatory activity.

In the negative S-type where the diet is high in meat and fat, we have the opposite problem. Initially, this provides useful additional fuel, and in combination with a raised epinephrine level makes the personality energetic. However, there is an inverse relationship between the epinephrine level and the effectiveness of insulin such that when the epinephrine level is high in response to this diet, the effectiveness of insulin is reduced.

Diabetes: Insulin regulates the speed with which glucose enters the cells. With a low insulin level, cells can be starved of glucose. The insufficient oxaloacetic acid is formed to combine with an abundance of acetyl coenzyme A, which is produced internally or by a high-fat diet. In those who are strongly SNS-dominant, this can lead to the development of diabetes. Glucose builds up in the bloodstream, raising the blood sugar level, and is spilled into the urine.

Figure 5-2 shows the nutritional solution to this problem. We must provide more oxaloacetic acid to the citric acid cycle. As a first step, we reduce all stimulating influences from the environment or from red meat, alcohol, coffee, tea, tobacco, or drugs. This will help reduce the epinephrine level and increase the effectiveness of insulin. Then we supply an abundance of zinc, which is needed to manufacture insulin, and chromium, which is part of the glucose tolerance factor that helps insulin to channel glucose through the cell wall.

Oxaloacetic acid is an unstable compound that cannot be ingested as such, but we can supply malic acid (the acid present in tart apples), which is easily converted into oxaloacetic acid. We can also supply citric acid from acid citrus fruits. This, too, can be converted into oxaloacetic acid through the citric acid cycle and it provides valuable energy.

Another food group that readily converts into oxaloacetic acid is the proteins. Most amino acids can either be transformed via pyruvic acid or directly enter the citric acid cycle. Most easily converted is aspartic acid. Only few amino acids are ketogenic (related to the fat metabolism) and yield acetyl coenzyme A: leucine, isoleucine, lysine, phenylalanine, tryptophan, and tyrosine. Incidentally, an excellent energy protein for diabetics, with almost 90 percent useful amino acids, is gelatine.

Fructose from fruits can enter muscle cells without insulin to form pyruvic acid. However, it must enter the bloodstream slowly (such as from raw food and small meals); otherwise fatty acids, cholesterol, and cataract-forming sugar alcohols may be produced. However, Type II diabetics have to be careful not to mix fructose with glucose, as pointed out in step 40 in regard to the disaccharide effect.

Sometimes glucose cannot be used and builds up in the bloodstream because the conversion of pyruvic acid into oxaloacetic acid is blocked or too slow. This conversion requires manganese, biotin, and nicotinamide. These and other vitamins and minerals required for the energy metabolism should be supplied in generous amounts as supplements.

The large buildup of acetyl coenzyme A is reduced in the liver by conversion into saturated fatty acids and cholesterol, both of which contribute to the development of atherosclerosis. In uncontrolled diabetes, a surplus of fat leads to the production of ketones and keto acids from acetyl coenzyme A (ketones are organic chemicals with an oxygen atom attached with two bonds to a carbon atom; keto acids have, in addition, an organic acid group). Ketones, such as acetone, can to some degree be used by the muscles to form oxaloacetic acid via pyruvic acid and thus keep the citric acid cycle going. However, this conversion is slow and a large amount of the acetoacetic acid accumulates and makes uncontrolled diabetics extremely acidic. The liver can metabolize glucose without the help of insulin, and high levels of glucose flooding the liver can cause a large buildup of lactic acid. Thus, an intake of sweet food contributes to general overacidity.

Intestinal sanitation, avoidance of sucrose, allergy testing, and/or a low-allergy diet are the main features in normalizing insulin production in insulin-dependent diabetes. A proper nutritional program is a second line of defence and can give good results even in resistant cases such as tumour of the pancreas.

Abnormal Fat Metabolism: The metabolic problems of slow oxidizers are similar to those described for diabetics. However, the glucose deficiency inside the cells is far less severe and, therefore, we do not have the overacidity resulting from an overproduction of keto acids. Instead, the surplus of acetyl coenzyme A is mainly converted into saturated fatty acids and cholesterol, causing atherosclerosis, cardiovascular disease, fatty degeneration of the liver, and overweight. In the slow oxidizers, overweight is mainly a result of eating too much fat, while in the fast oxidizer, overweight results mainly from eating sweet food and wheat. In both cases, the thyroid gland tends to be underactive.

The answer to these problems is to speed up glycolysis and produce more oxaloacetic acid. This can be done by supplying plenty of fruit acids as well as all the vitamins and minerals required for their sugar metabolism. Minimizing intake of fats and sugars will stop the oversupply of acetyl coenzyme A, while epinephrine levels will fall and insulin activity will rise in the absence of meat consumption and stress. This is also the key to cleaning out atherosclerotic arteries.

Sub-Oxidizers and Cancer: The sub--oxidizer has an inefficient metabolism and, in a way, combines the problems of the fast and the slow oxidizer. Many sub-oxidizers have cancer or precancerous conditions. In cancer the metabolism of all nutrients is greatly impaired. The main energy production of cancer cells is similar to that present in uncontrolled hypoglycemics: Both produce mainly lactic acid instead of energy. While this makes the tumour itself overacid, the rest of the body may be too alkaline due to a very sluggish metabolism.

Figure 5-2 on the citric acid cycle also reveals how we can starve a tumour: We must reduce all nutrients except fruit acids to the bare minimum. In this way the tumor is completely deprived of energy, while normal body cells can still derive energy from metabolizing acids together with fatty acids from the fat deposits of the body.

The time-tested method used in natural medicine for those with sufficient fat reserves is a fast, lasting several weeks, on fruits only, mainly tart varieties. The “grape cure” has become famous in this regard; tart apples and acid citrus fruits can also be used. During this time it is essential to flush the intestines daily (using Epsom salt and colonics, for example); otherwise the poisons generated by a large, disintegrating tumor can cause great distress and even death. At other times, use plenty of red beets, raw, juiced, or cooked. The red beet pigments greatly increase the oxygenating ability of the cells and normalize metabolism.

Diet Types: Understanding how your body produces energy, and how to overcome some common nutritional problems, can guide you in selecting an appropriate diet for your individual requirements. Slow oxidizers have similar dietary requirements to those of the metabolic S-type, that is, a vegetarian or semi-vegetarian diet. Fast oxidizers and P-type individuals, on the other hand, do best on a non-vegetarian diet or a high legume vegetarian diet, which at the same time should be low in sugars. The balanced type can use a mixed diet that is neither high in protein and fat nor in sweet food. Here are some dietary guidelines:

Diet for S-Type and Slow Oxidizers: Proteins should be supplied mainly from legumes, spirulina, almonds, and other oily seeds, ideally as seed cheese (as explained earlier), but the consumption of peanuts should be restricted. Fish and other seafood, eggs (not hard-boiled or fried), and some poultry and raw milk products are usually acceptable. Have plenty of fruits, especially tart varieties, frequent fruit meals, and all-fruit days. Sweet varieties of vegetables are beneficial, particularly red beets, as are sprouted seeds.

Polyunsaturated oils are mainly supplied from seafood, fermented soybeans, sunflower kernels, and freshly ground linseed. Meats, fats, and processed oils should be eaten sparingly. The evening meal should be light. Helpful supplements are lecithin, digestive enzymes, cod-liver oil or other fish oils, kelp, B vitamins, vitamin E, ascorbic acid, magnesium, potassium, chromium, copper, selenium, and zinc. The idea is to eat fast-digesting food.

Diet for P-Type and Fast Oxidizers: This diet, high in legumes, is often combined with brown rice and other grains such as millet and buckwheat (also sago, taro, arrowroot, and tapioca if available from a health food store). Have a hearty breakfast. Flesh foods are generally well tolerated. This metabolic type should use whole grains - sprouted, cooked, or baked, in preference to flour products; also use plenty of sprouted seeds. Starchy and green vegetables are best for cooking; use sweet varieties, mainly raw, with plenty of red beets, raw, cooked, or juiced. Use salted food if your blood pressure is low.

Oils low in linoleic acid should be used. Best here are olive and almond oils, but also try coconut oil. The intake of sweet food, sweet and acid fruits, and lactose products should be minimized. Neutralize fruit acids with dolomite. Useful supplements are spirulina, halibut-liver oil capsules, calcium ascorbate, low-level B vitamins, MSM, chromium, copper, selenium, and zinc. The idea is to eat slow-digesting food.

Diet for Sub-Oxidizers: People of this metabolic type may use a mixed diet of easily digestible food. Thus meat may be used as broth or stew, not fried or roasted. Nuts and oily seeds should be soaked, sprouted, or used as seed cheese. Sub-oxidizers can consume sprouted seeds and plenty of grass juice; they can also eat beef juice, freeze-dried liver, and plenty of spirulina and bee pollen. Use digestive enzymes, a hydrochloric acid supplement, or ascorbic acid with protein meals. All vitamins and minerals, especially trace minerals, should be supplemented; mix lecithin with all meals; rubbing the skin with cod-liver oil may be helpful. The idea is to eat easily digested food.

Diet for Sensitivity Regulation: Certain nutrients make the skin and sometimes the brain more sensitive and should be minimized if the skin easily reacts to irritants and

Table 5-4: Food List for Blood Type Groups

	Type A	Type AB	Type B	Type O
Almond	OK	OK	OK	OK
Apple	OK	OK	OK	OK
Avocado	OK	Avoid	Avoid	Avoid
Bacon/ham/pork	Avoid	Avoid	Avoid	Avoid
Banana	Avoid	Avoid	OK	OK
Barley	OK	OK	Avoid	OK
Beef	Avoid	Avoid	OK	OK
Beer	Avoid	OK	OK	OK
Beet root/ red beet	OK	OK	OK	OK
Brazil nut	Avoid	OK	OK	Avoid
Broccoli	OK	OK	OK	OK
Brussels sprouts	OK	OK	OK	Avoid
Buckwheat	OK	Avoid	Avoid	OK
Cabbage	Avoid	OK	OK	Avoid
Carrot	OK	OK	OK	OK
Cashew	Avoid	OK	Avoid	Avoid
Cauliflower	OK	OK	OK	Avoid
Celery	OK	OK	OK	OK
Cheese (most types)	Avoid	OK	OK	Avoid
Corn/maize	OK	Avoid	Avoid	Avoid
Cottage cheese	Avoid	OK	OK	Avoid
Cucumber	OK	OK	OK	OK
Duck	Avoid	Avoid	Avoid	OK
Egg	Avoid	OK	OK	OK
Flaxseed	OK	OK	OK	OK
Garbanzo beans or chickpeas	Avoid	Avoid	Avoid	OK
Grapes	OK	OK	OK	OK
Herring (fresh)	Avoid	OK	OK	OK
Kidney beans	Avoid	Avoid	OK	Avoid
Lamb/mutton	Avoid	OK	OK	OK
Lemon	OK	OK	OK	OK
Lentils	OK	OK	Avoid	Avoid
Lima beans	Avoid	Avoid	OK	OK
Liver (calf)	Avoid	OK	OK	OK
Lobster	Avoid	Avoid	Avoid	OK
Mackerel	OK	OK	OK	OK
Milk, cow	Avoid	Avoid	OK	Avoid
Milk, goat	OK	OK	OK	Avoid
Millet	OK	OK	OK	OK
Mung beans	OK	Avoid	Avoid	OK
Mussels	Avoid	OK	Avoid	OK
Navy beans	Avoid	OK	OK	Avoid
Oats	OK	OK	OK	OK
Octopus	Avoid	Avoid	Avoid	Avoid
Olive oil	OK	OK	OK	OK
Onion	OK	OK	OK	OK
Orange	Avoid	Avoid	OK	Avoid
Oyster	Avoid	Avoid	Avoid	OK
Papaw	Avoid	OK	OK	OK
Pea	OK	OK	OK	OK
Peach/pear	OK	OK	OK	OK
Peanut	OK	OK	Avoid	Avoid
Pineapple	OK	OK	OK	OK
Plum	OK	OK	OK	OK
Potato, sweet	Avoid	OK	OK	OK
Potato, white, red	Avoid	OK.	OK	Avoid
Pumpkin	OK	OK	Avoid	OK
Rabbit	Avoid	OK	OK	OK
Rice	OK	OK	OK	OK
Rye	OK	OK	Avoid	OK
Salmon	OK	Avoid	Avoid	OK
Sardine	OK	OK	OK	OK
Sesame seed/tahini	OK	Avoid	Avoid	OK
Soybean	OK	OK	OK	OK.
Spelt	OK	OK	OK	OK
Spinach	OK.	OK	OK	OK
Strawberry	OK	OK	OK	Avoid
String bean	OK	OK	OK	OK
Sunflower seed	OK	Avoid	Avoid	OK
Tapioca	OK	Avoid	Avoid	OK
Tofu/tempeh	OK	OK	Avoid	OK
Tuna	OK	OK	OK	OK
Turkey	OK	OK.	OK.	OK
Turnip	OK	OK	OK	OK
Walnut, English	OK	OK	OK	OK
Wheat	Avoid	OK	Avoid	Avoid
Yam	Avoid	OK	OK	OK
Yogurt	OK	OK	OK	Avoid
Zucchini	OK	OK	OK	OK

if there is emotional and mental instability. Sensitizing nutrients and factors include: acidic and sweet foods and fruits, allergens, and a diet high in phosphorus and low in calcium and magnesium. Proteins, fats, and oils, complex carbohydrates, calcium, and magnesium help to protect susceptible individuals from undue sensitivity. Insensitive individuals, on the other hand, benefit from a high intake of acidic foods and fruits, lecithin, folic acid, and a diet low in -calcium but high in magnesium. Common polyunsaturated oils and lecithin (high in phosphorus) increase sensitivity; linolenic acid (linseed), and fish oils normalize it.

Metabolic Types and Blood Groups: The naturopathic physician Peter D'Adamo, N.D., has written a series of popular books about the relationship between diet, health, and blood type, notably Eat Right for Your Type, Live Right 4 Your Type, and "The Eat Right for Your Type Complete Blood Type Encyclopaedia". He shows that individuals of different blood groups have different optimal food requirements. Special proteins called lectins found in foods not traditionally eaten frequently cause immune reactions in the intestinal wall and may eventually lead to a wide variety of diseases, he says.

The relationship between the three basic metabolic types and blood groups is straightforward. Blood group O is the oldest blood group. It originated in the hunter-gatherer area and corresponds to the P-type or non-vegetarian type. Group A resembles most closely the S-type or grain-eating vegetarian type, while B is the lacto-vegetarian type and closest to the balanced metabolic type. Blood group AB has the characteristics of A as well as B.

It is my clinical opinion that it is beneficial to know which foods are supposedly good or bad for your blood type, but it is not necessary to stick strictly to the corresponding rules. Some of the foods listed to be avoided do not actually damage our blood and have been selected for other reasons that may not apply to everyone. On the other hand, anyone not in good health should take food allergy tests and then combine the recommendations of the blood type diet with the results of the allergy testing.

If symptoms of food sensitivity show up, then I would not use a food that is supposed to be beneficial for your blood group. However, if you are in reasonably good health, allergy tests show no negative effect, and muscle testing gives a strong reaction, then you can cautiously use an "avoid" food, especially if there are other good reasons for using it. An example of this is coconut and its products, which Dr. D'Adamo advises all blood groups to avoid. However, I find coconut milk and oil beneficial for myself, and many others find it beneficial as well.

While we can expect most allergies to occur with food that is wrong for our blood type, this is not always so. What happens in our body depends mainly on the quality of our digestive system and on the quantity and digestibility of the questionable food and its interaction with other food ingredients.

It may be that we react differently to cooked foods and raw foods. Sprouted seeds appear to be more acceptable and possibly beneficial even if the same seeds should be avoided in cooked form. If all is well, then the potentially harmful lectins are broken down to amino acids and are harmless; if not, then we must improve conditions or avoid these foods. Try to develop your inner food sense and learn to listen to your body. It will then let you know by your inner urgings and taste preferences which foods to eat.

Table 5-4 is a list of some common foods in relationships to the blood groups. Basically, there are only two food classifications: foods that give a reaction and should be avoided and those that do not and may be eaten. Dr. D'Adamo also classified the "allowed" foods as either neutral/avoid, neutral, neutral/beneficial, or beneficial. Neutral means that the food does not cause any blood group related problems, but it has no other specific advantages to recommend it.

* * * * * * * * * * *

Part 6

SPECIFIC HEALTH PROBLEMS

This chapter offers guidelines for overcoming some specific health problems. However, it is essential to incorporate these specific recommendations into a holistic program for improving your overall well-being. Start with the basics: Do the earlier steps in chapter 1 as thoroughly as possible. Your disease or health problem may disappear without any further therapies being required. In some cases, where one of these basic steps is especially important, it may be mentioned for a specific disease, but all other steps need to be done as well.

In addition to any supplements specifically mentioned for your problem in this chapter, take antioxidants, spirulina, bee pollen, cod-liver oil, lecithin, kelp, or colloidal minerals, and other helpful nutrients. Also use other methods described throughout this book, even if not specifically mentioned, such as reflexology, meridian therapy, magnet therapy, color therapy, and guided imagery. The more effort you make to apply as many supportive methods as possible, the better will be the health result.

An important concept, especially with inflammatory conditions (which are present in most diseases), is chronic stress. This may be due to several individual factors, such as food allergies and chemical sensitivities, chronic infection from a dead tooth, electro-pollution, and emotional stress. Each of these factors stimulates the adrenal glands to release inflammatory hormones. Removing one or several of these inflammation promoters sometimes does not help much; when all of them are removed, you may suddenly see great improvement. If your condition is not specifically mentioned, or is rather severe, in addition to any of the recommended measures, adopt a temporary raw food diet, combined with suitable mind therapy.

* * * * * * * * * * * *

Step 44

SELF-CARE HELP FOR HEALTH PROBLEMS

Practical Tips for Relief

If you have a specific disease, this summary of typical treatment protocols can show you what to do.

I am not a doctor licensed to prescribe medications or ensure cures, but a nutritionist and natural therapist well versed in natural healing methods. This information is derived from sound nutritional and clinical research. Readers should always consult a qualified health practitioner when dealing with serious illness. I do not promise or guarantee any cures, and the reader should regard the following as general guidelines that have been shown to help improve some people with these conditions. For updates and further information, see my websites: www.health-science-spirit.com, www.healing-yourself.com and www.heal-yourself.com.au. For disease treatment protocols using nutritional therapies, see: www.lef.org and www.doctoryourself.com.

Addictions: Use Basic Cleanse (Step 4) followed by food allergy testing, and practice a diet based on grass and vegetable juice with added water for a total fluid intake of about a gallon daily. Prepare a homeopathic remedy made of the addictive substance. Use intestinal sanitation; alkalize the body; take high-potency vitamin B complex and glutamine; practice guided imagery and affirmations; have outdoor activity and group support. Do detoxification using frequent saunas or hot baths.

AIDS: Helpful is a predominantly raw food diet with plenty of sprouted seeds and fresh vegetable juices, mainly from cereal grasses and purple foods. Also try oxygen therapies alternating with high antioxidant intake, intestinal sanitation, and herbal parasite cure, and continue longer with wormwood. Colloidal silver, copper salicylate or colloidal copper, echinacea, MSM, propolis, and other natural antibiotics can be beneficial, along with a Beck-type electronic zapper and magnetic pulser, Basic Cleanse, and emotional and mind therapies. (See Infections and Inflammations and Candidiasis below.)

Allergies: Test yourself against individual foods and chemicals with an elimination diet after a five-day water fast, or after Basic Cleanse or a rice diet with only salt and olive oil for flavouring. Possibly use a rotation diet and use the same food only once every five days; eat small meals and chew well. Use digestive enzymes, hydrochloric acid, vitamin B complex, bio-flavo-noids and carotenoids, high vitamin C, glutamine, and MSM. Alkalize the body with sodium and potassium bicarbonate; neutralize fruit acids with dolomite. Avoid or minimize intake of sweet and acid food and food additives. Make a homeopathic remedy from the offending allergens.

Anemia: Consume juice or powder of cereal grasses, organic liver (underdone or freeze-dried), and sprouted seeds. Take hydrochloric acid and ascorbic acid with proteins and cooked food; minimize intake of sweet food. Vitamins B6, B12, folic acid, E, and bioflavonoids are helpful as are chelated or organic forms of iron, copper, zinc, manganese, chromium, and selenium. Do a parasite cure with wormwood, cloves, and tincture of walnut, and electronic zapper. Practice red color therapy (Step 17); get mild sun exposure and outdoor activity. Treat any excessive menstrual bleeding and avoid long-term use of anti-inflammatory drugs, especially aspirin.

Arthritis: DIET: Use the Basic Cleanse followed by allergy testing and follow a predominantly raw food diet. Use ground linseed, fish oils, cold-pressed peanut oil, olive oil, and plenty of grass juice and sprouted seeds. If your fingernails are soft, take hydrochloric acid with meals, then sodium bicarbonate three hours after hydrochloric acid. Avoid or minimize sweet and acid food intake as well as wheat products, gluten, milk, cheese, commercial cooking oils, citrus fruit, processed food, and food of the nightshade family (potato, tomato, capsicum, and eggplant).

SUPPLEMENTS: Take vitamin B6 (especially with problems of the hands) and nicotinamide in addition to multivitamins. Take 500 mg or more of magnesium, boron, zinc, and MSM; try copper salicylate for pain and inflammation, or, alternatively, copper colloids or copper bracelet. Also recommended are acidophilus, digestive enzymes, celery seeds, and yucca. Most helpful are glucosamine, chondroitin, mussel extract, or liquid bovine or shark cartilage, and broth of fish heads. (See “Alkalizing with Calcium and Magnesium” and “The Calcium-Phosphorus Ratio” in Step 6.)

ADDITIONAL THERAPIES: Use warm peanut oil rubs on affected joints; also try blistering agents, or hot Epsom salt packs and baths, or use magnesium chloride instead, followed by sweating wrapped in blankets. Dissolve MSM and glucosamine in a solution of copper salicylate and rub it into affected joints. Try blue light therapy over painful areas and orange light on stiff, calcified joints (Step 17). Try deep muscle massage; remove dead teeth and mercury amalgam fillings; do the herbal parasite cure with the electronic zapper.

Asthma: Avoid or minimize sweet and acid food, wheat, gluten, and lactose as in milk products (except butter); also avoid vaccinations, aspirin, processed foods, and the sulphites used in many foods and medications as preservatives. Practice intestinal sanitation and allergy testing. Use ground linseed, fish oils, and olive oil, and avoid other commercial oils. Halibut-oil capsules or other source of vitamin A (not beta-carotene instead), vitamins E, C, B1, B6, L-lysine, molybdenum, magnesium, manganese, MSM, selenium, and zinc are helpful. (See Alkalizing with Calcium and Magnesium and The Calcium-Phosphorus Ratio in Step 6.) Take hydrochloric acid with protein meals if your fingernails are weak, then sodium bicarbonate three hours later. Use foods high in antioxidants, bioflavonoids, and carotenoids, including green tea, grape seed extract, and purple foods.

Practice breathing exercises: inhale into the abdomen and exhale as slowly as possible; do this routinely during sedentary activities. After normal exhalation, hold your breath for as long as possible; inflate balloons as a further exercise; have a spinal adjustment. Preferably live in a dry, warm climate or higher elevation; avoid damp places, moulds, and fungi in air and food. Use natural fibre mattresses or wrap foam mattresses in polyethylene sheeting with only a small opening at the bottom.

Eliminate systemic parasites and microbes with wormwood, cloves, black walnut tincture, and use of the electronic zapper; try colloidal silver, copper, or copper salicylate. Vomiting therapy can help: Try to cough up mucus after placing hot onion packs on the chest. Make children emotionally secure; try emotional release therapy for adults with asthma. Use the Emotional Freedom Technique (see Step 63).

Some asthmatics claim to have been cured by regularly inhaling three-percent hydrogen peroxide sprayed directly into the mouth from a spray bottle, while others had success with inhaling high concentrations of negative ions from a medical-type air ionizer or with spray-inhaling a ten-percent sodium ascorbate solution. Hydrogen peroxide is probably best against fungal infestations, and antioxidant spray against inflammations and oxidative damage. Try both but generally use antioxidants as the main approach.

Autoimmune Diseases: In autoimmune diseases, the immune system attacks faulty body cells, probably due to infestation with microbes or their toxins. Common causes are dead teeth and resulting jawbone infections, overgrowth of the intestinal tract with pathogenic microbes, and some researchers suspect vaccines.

Autoimmune diseases include Type I diabetes, lupus, multiple sclerosis, myasthenia gravis, rheumatoid arthritis, scleroderma, and thyroiditis. The main aspects of treating autoimmune diseases are to avoid gluten products and substances to which you are allergic or sensitive, sweetened food, genetically modified or microwaved food, and processed food in general. By far, the best option is to adopt temporarily a raw food diet.

Use high amounts of antioxidants, in particular vitamin C, L-lysine, zinc, and other nutrients listed under viral infections in the section on Infections and Inflammations below; also use the electronic zapper and other measures listed in that section under Addi-tional Therapies. See any entries for your specific autoimmune disease.

Back Problems: Practice a low allergy diet; take vitamins C, E, and B complex, and high doses of manganese, zinc, and magnesium. Reflexology is very effective, as are Epsom salt packs. For shrunken disks, stretch the spine and have it rubbed with hypericum oil; get deep muscle massage or pressure applied at tight muscles; have a spinal adjustment. Hang upside down (inversion swing) for problems in the lower back. Try stretching exercises; sleep on a firm bed, and women should avoid wearing high heels. Improve digestive system function and any diseased organ, as back pain often results from problems in inner organs. Try other recommendations in Arthritis and see Step 20: Spinal Therapy and Massage.

Broken Bones: Fresh vegetable juices, broth of fish heads, and bone meal can be helpful as well as vitamin C, halibut-oil capsules, manganese, boron, zinc, magnesium, and comfrey root. Place the north pole of a magnet above the fracture (towards the trunk) and the south pole opposite and below the fracture (towards fingers or toes); also try a pulsating magnetic field at the site of the injury. See Osteoporosis below.

Cancer: The main aspects of a holistic cancer therapy are strengthening the immune system, controlling the cancer microbe, removing accumulated toxins and waste products, and dissolving tumours.

DIET: Start with the Basic Cleanse followed by a juice diet, grape cure, or low-protein diet; use additional spirulina and bee pollen for high-quality protein. Maximize purple foods (red beets or their juice, black or purple grapes, dark grape juice, blueberries, blackberries) with any diet. Eat at intervals during the morning one or two pounds of purple foods, or sip one bottle of dark grape juice or red beet juice; take no other food until noon. If the body is sensitive, keep the urine slightly alkaline (see Step 6). If the body is insensitive or has soft fingernails, take hydrochloric acid with meals. Structure the diet according to your metabolic type and blood group (see Step 43). Cancer-related pain tends to disappear on a raw food diet.

SUPPLEMENTS: The most important vitamins are C, 10 g or more, mainly neutralized and in five divided doses; A, best as shark-liver oil and cod-liver oil (if you have malabsorption, use A-emulsion); niacinamide, folic acid, B2, and B6; with generalized weakness, inject B12 or dissolve it under the tongue.

The most important minerals are magnesium, selenium, and zinc. Use high amounts of all antioxidants, including alpha-lipoic acid and grape seed extract. You can alternate short periods of several weeks of oxygen therapies with several months on high antioxidant supplements.

With tumours, use high amounts of the following to inhibit their formation of new blood vessels: shark or bovine cartilage (both best as liquid), zinc (30 mg, three times daily), vitamin B2, turmeric, and green tea; minimize any copper intake.

Other useful remedies are proteolytic enzymes, MSM or sulphur-amino acids (L-cysteine and L-methionine), propolis, ground linseed, aloe vera, kelp, Lactoferrin, lecithin, and freeze-dried liver. Some effective herbs are the Hoxsey herb mixture, Essiac, pau d'arco, papaw leaves, echinacea, greater celandine, red clover blossoms, and wormwood and other herbal parasite remedies. Urea helps to prevent metastases; hydrazine sulphate helps you regain weight and strength.

ADDITIONAL THERAPIES: Minimize electromagnetic pollution, and sanitize your teeth (see Step 8). Experiment with ways to increase your life force as shown in Part 7. Use an electronic blood purifier, and directly over the tumor place a magnetic pulser, strong south pole magnet, and blue light therapy for pain and tumor control. Try various types of packs over the tumor area, such as pulped grass or dark grapes, cabbage leaf, or honey (see Step 21). Avoid surgery near the full moon (increased chance of haemorrhage); premenopausal females may have a reduced rate of metastasis after surgery near ovulation time (mid-cycle) as compared to surgery near the start of the period.

Improve your mind with several hours of daily relaxation, affirmation, prayer, guided imagery, and meditation. Practice positive thinking and develop faith and a joyful mind. Nothing will cure you if you do not believe in it. Learn to feel and express your needs as well as negative emotions and become assertive. Resolve any emotional conflict - see: “Emotional-Mental Shocks” later in this part. Do not focus on fighting cancer, but rather on improving your health and enjoyment of life. Choose to live life to the fullest, have a goal for the future, and become enthusiastic about your life. Your attitude is the key to your success.

Candidiasis: This is a fungal infection caused by Candida albicans. The most common symptoms are thrush (oral) and vaginal and skin infections. However, Candida often invades the body itself, starting with the digestive tract and then entering the bloodstream, overwhelming the immune system. In this way, Candida has been found to cause or aggravate food allergies, chronic fatigue, digestive-tract diseases, all kinds of autoimmune diseases, lupus erythematosus, multiple sclerosis, myasthenia gravis, but also schizophrenia, mental depression and anxiety, impaired memory, hyperactivity, menstrual problems, and miscarriage. Basically, any part of the body can become impaired.

Antifungal remedies include garlic, golden-seal, pau d'arco, propolis, wormwood, and the oils of tea tree, cinnamon, clove, eucalyptus, mustard, and thyme. Use Intestinal Sanitation (see Step 3) combined with the hypoglycaemia diet, and avoid sweet foods, yeast, moulds, and initially fruit. For severe fungal infestation of inner organs, an internal iodine solution, commonly called Lugol's solution, can be taken for three weeks. Also try oxygen therapy, colloidal silver and copper, and an electronic blood purifier, also anti-inflammatory remedies as required.

LUGOL'S SOLUTION: This is an iodine solution taken internally, designed to eliminate Candida and other microbes, viruses, and mycoplasmas from the bloodstream. Obtain Lugol's solution, also labelled Aqueous Iodine Oral Solution, from a pharmacist. Take a test drop in juice. If no allergic reaction results, continue to take six drops four times daily in liquid or mixed with food. Iodine is an oxidant, so do not take it with antioxidants. Continue Lugol's for three weeks, but interrupt if you have a strong reaction. Be careful, for high amounts of iodine taken for more than three weeks may depress thyroid functions. You may make Lugol's solution yourself by dissolving 44 g (1.5 ounces) of iodine and 88 g (3 ounces) of potassium iodide in one quart (litre) of water. First dissolve the potassium iodide in a pint of water, then add the iodine and fill up to one quart. Shake from time to time until dissolved after about a day.

Cardiovascular Disease: Recommended is the Basic Cleanse followed by periods on fruits only, or the grape cure or juice diet. Avoid or minimize sweeteners and especially sucrose, salt, chlorinated water, smoking, peanut products, red meat, caffeine, and all commercial fats and oils except extra-virgin olive oil. Do not heat any flesh food or eggs above boiling temperature. Separate fruit and other fructose-containing foods from starches or foods releasing glucose. Use only cow's milk products that have not been pasteurized or otherwise heated.

Use plenty of fresh fruits and vegetables; seafood (if not vegetarian); some (organic) poultry; one to two tablespoons of freshly ground or frozen linseed (flaxseed) and one to three capsules or teaspoons of lecithin with most meals; fish oil capsules; oily seeds and nuts (not roasted or rancid), best as seed cheese. Papaya and pineapple are good. If your blood pressure is high, acidify your body with ascorbic acid and take hydrochloric acid supplements with meals. With acid body conditions, more histamine is released, which in turn dilates blood vessels.

SUPPLEMENTS: Take high doses of natural vitamin E (tablets), vitamin C (5 to 10g daily, in three to five divided doses, mostly as ascorbic acid), vitamin B6, folic acid, and niacin before meals to induce flushing. Take close to 1000 mg of magnesium daily as magnesium chloride or with neutralized ascorbic acid. Other helpful minerals are zinc, selenium, chromium, manganese, and copper salicylate; but males should minimize calcium. Seriously ill patients can take 5 to 6 g of lysine and 2 g of carnitine daily in four divided doses. For herbs, consider bromelain, ginger (with most meals), ginkgo biloba, and green tea; also helpful are digestive enzymes high in lipase.

ADDITIONAL THERAPIES: Remove any dead teeth and make sure there is no chronic jawbone infection in the wisdom tooth positions. Get regular exercise; have alternating hot and cold showers; and do skin brushing. Frequently press into the fleshy pad below the left thumb; pressing strongly here during a heart attack can allegedly stop the attack. A small magnet with the south pole placed over the heart can ease pain or slow the heartbeat, while the north pole placement strengthens the beat (do not use with a pacemaker). Also try the herbal parasite cure, electronic zapper, and colloidal silver. Practice the Love Cure (Step 61).

Chronic Fatigue Syndrome: In this condition, the cellular oxidative energy production is blocked by non-biological chemicals or endotoxins from a chronic microbial infection, often due to Candida present in the blood after antibiotic therapy. Techniques to try include allergy testing and a low allergy diet; consuming plenty of purple foods (beet root, red cabbage, purple onions, black/purple grapes, black currants, blackberries, blueberries, or juices of any of these), and green juice or, alternatively, barley green powder, spirulina, or chlorella. With low blood pressure, eat salty and spicy food; select foods with a favourable calcium/phosphorus ratio; and alkalize fruit acids with dolomite.

SUPPLEMENTS: Use high levels of antioxidants with 10 g or more of sodium ascorbate in divided doses, vitamin A emulsion, natural vitamin E complex as tablets, alpha-lipoic acid, grape seed extract, and Microhydrin or strong electrolytic reduced water (ERW). Also useful are vitamin B-complex, calcium orotate, magnesium, selenium, zinc, coenzyme Q10, L--glutamine, L-lysine, aged garlic, freeze-dried liver, propolis, echinacea, maca, and either L--cysteine and L-methionine or MSM. Initially try weekly vitamin B12 injections or daily vitamin B12 tablets dissolved under the tongue.

ADDITIONAL THERAPIES: Try Candida treatment, herbal parasite cure, electronic blood purifier, magnetic pulser, colloidal silver, and copper salicylate. Minimize electro-stress and sanitize the teeth (see Step 8).

Colds: Colds are beneficial clean-outs of mucus congestion, so to avoid colds, avoid mucus congestions and foods that produce them. These are mainly lactose and cow's milk products (except butter), sweet foods, moulds and mouldy foods, and any foods or chemicals to which one is allergic or sensitive. Other factors are mercury amalgam fillings, root canals, and other foci of chronic infections, as well as Candida overgrowth and intestinal dysbiosis (pathogenic bacteria).

Try to correct these conditions, and see Mucus Problems below for ways to ease mucus congestions. I have seen severe colds aborted within hours by using the Sugar Cure (see Mucus Problems). Other useful remedies are echinacea, fresh raw garlic, and very high doses, up to bowel tolerance, of vitamin C (sodium ascorbate). In addition, frequently dissolve a tablet with a small amount of zinc sulphate in the mouth. Another useful method is to dissolve one-third of a teaspoon of sodium ascorbate in two tablespoons of water and from time to time place 20 drops in each nostril.

Colds can also be stopped if, at the first sign of a coming cold, a few drops of 3 percent hydrogen peroxide are placed in each ear. Lie on one side while putting the drops into the opposite ear, wait a few minutes, then change sides and put the drops in the other ear. In addition, gargle frequently with diluted hydrogen peroxide and also swallow some. See Infections and Inflammations below.

Depression: Use a high-quality low-allergy diet. With sensitive skin and emotional instability, use mainly slow-digesting foods, and with insensitive conditions consume plenty of fruit. Avoid foods to which you may be addicted or on which you might binge, such as chocolate, coffee, tea, and alcohol. Give yourself a special treat from time to time, but without using sweet or junk food. Very important are fish oils (best as cod-liver oil) and linseed oil from ground linseed.

The most important vitamins are B6, B12, C, niacin or nicotinamide, and folic acid, and the most important minerals are zinc, magnesium, calcium, and manganese. Initially take a high-potency B-complex and mineral tablet with most meals and additional amounts of vitamin B6 and zinc (up to 500 mg daily of vitamin B6). Try hypericum or St. John's Wort and the Bach Flower Remedies Aspen, Gentian, Gorse, and Mustard. Correct constipation; take cultures of acidophilus and bifidobacteria. Have any dead teeth removed and preferably mercury amalgam fillings replaced. Avoid lead pollution. Kyolic (aged garlic), cooked garlic, and raw garlic assist in detoxification. Candida can go into the bloodstream and cause depression and other psychiatric symptoms. For treatment, see Candidiasis above.

Try amino acids: up to 2 g of L- (or D,L) phenylalanine and tyrosine, each before meals, or alternatively, up to 2 g of tryptophan. You can also try taking phenylalanine and tyrosine before breakfast only and tryptophan at bedtime only. Glutamine can be used as an additional brain fuel. Epinephrine and dopamine in the brain are formed from phenylalanine and tyrosine, while serotonin is made from tryptophan.

Depression often results from low thyroid and adrenal functions. To assess thyroid status, check your temperature under the armpit before rising in the morning; women check this in the first ten days of the menses or before ovulation. If it is below 97.8º F (36.5º C), try thyroid extract (absorb under the tongue) and iodine from kelp. In addition, selenium is required to activate thyroid hormone; take 200 mcg per day.

Seasonal affective disorder is due to insufficient sunlight exposure. Expose yourself to one or two hours of very bright full-spectrum electric light each day, preferably in the morning; if possible, be outside in sunlight and without sunglasses. In addition, use yellow or orange color therapy for all forms of depression; also look into the light source (see Step 17). Wear bright-coloured clothing. Use deep-breathing exercises; dance to lively music and shake the whole body for a minute during fast breathing.

Use prayer, affirmations, meditation, and guided imagery to uplift your feelings and solve your problems. Learn to trust in your inner or higher guidance and ask it to take care of your problems. Socialize and find a group with similar interests. Forgive yourself and others for any mistakes or misdeeds.

Diabetes: Under professional medical supervision only, go for one week on a diet of only apples, mainly tart varieties (e.g. Granny Smith) with no or only a minimum of insulin or drugs. Undertake intestinal sanitation (Epsom salt - garlic flush followed by plenty of dairy-free acidophilus cultures). Then gradually introduce non-sweet raw foods, especially sprouted seeds, lentils and other legumes, seafood, onion, and cooked vegetables. Keep a diary of the blood sugar level after each test food (a food that you ingest to see how your blood sugar reacts to it). Continue using only low-response food (a food that does not strongly elevate the blood sugar level).

Before each meal have plenty of fresh grass juice, spirulina, or chlorella, one tablespoon of freshly ground linseed, and one tablespoon of fenugreek seeds (soaked overnight; use the soaking water). Avoid all cow's milk products, wheat and gluten, saturated and especially hydrogenated fats, processed foods, and food additives. Avoid all sweet or sweetened foods, including fruit juices, and in particular all sucrose; avoid fructose and glucose at the same meal, such as having fruit after a meal, although it is generally safe to have fruit an hour before a meal. You may, however, use xylitol, sorbitol, or the herb stevia as a sweetener.

With Type I diabetes, avoid all glucose--producing foods, such as starches, grains, or cereals, sweet vegetables, and even legumes during a trial period. Use instead meat, fish, egg, goat's milk products, cod-liver oil, spirulina (one to two ounces), lettuce, cabbage, and other leafy greens, in addition to green/tart apples, lemons, and limes. If you need more carbohydrate, add fructose to the lemon juice; try fenugreek and ground linseed.

SUPPLEMENTS: Take vitamins C, E, B-group, especially B1, B6, and B12; glucose tolerance factor or chromium; manganese and zinc (chelated); 500 mg or more daily of magnesium; vanadium, if available; also MSM, fish oils, lecithin, and freeze-dried liver; continue with dairy-free acidophilus. The herb Gymnema Sylvestre is effective with both types of diabetes; take either 200 mg of extract or 1 to 2 g of dried leaves shortly before meals that are likely to elevate blood glucose levels. Also useful are aloe vera, bilberry, ginseng, goat's rue (may help to regenerate pancreatic beta cells), and various spices such as cinnamon (one teaspoon daily in divided doses), cloves, tarragon, and turmeric. Calcium arginate (calcium salt of the amino acid arginine) is effective with Type II diabetes. Also try castor-oil packs on the liver and abdomen; magnetic pulser on liver and pancreas; do the reflexology for these points on feet and hands; herbal parasite cure and colloidal silver.

Digestive Disorders: Do allergy testing and have a low-allergy diet with small meals; chew slowly and well; practice correct food combining; follow a diet appropriate for your metabolic type and blood group; consider a rice or arrowroot mono-diet for a while, with only salt and olive oil for flavouring, or alternatively a water fast (purified water only). Use the Basic Cleanse and intestinal sanitation with acidophilus and bifido cultures. If you have soft fingernails, add hydrochloric acid to your meals, taking sodium bicarbonate three hours later.

Take digestive enzymes or a papaya (pawpaw) Smoothie with meals (see the recipes in Part 4). Place castor-oil packs on the liver and abdomen; do reflexology or have acupuncture. Herbs to stimulate digestion are aniseed, caraway, fennel, fenugreek, mint, or peppermint; try half a cup of a bitter liver herb tea after the meal. With acute indigestion, induce vomiting; with diarrhea, try an alimentary canal flush with crushed garlic and a tablespoonful of Epsom salt.

With ulcers and irritable bowel (Crohn's disease), try slippery elm powder before meals; meals with arrowroot can be soothing; and use ginger root often with meals. Goldenseal and chamomile are useful as they are anti-inflammatory. Try glutamine and zinc sulphate with meals (200 mg/day); also copper salicylate, herbal parasite cure, oxygen therapy, and colloidal silver. Magnesium is best taken as magnesium chloride; take also high-quality aloe vera juice.

Use the magnetic pulser and south pole of a strong magnet on sites of inflammation or discomfort. Include raw potato in the vegetable juice for stomach ulcers and cabbage juice with duodenal ulcer; stomach ulcers usually cause pain after a meal, duodenal ulcers between or before meals. Very effective are six drops of tea tree oil diluted in a glassful of liquid, then taken orally before breakfast for three days; initially take a test drop to see that you are not oversensitive to it.

With constipation, use sufficient magnesium (as chloride or sulphate) to produce bowel movements and take one or more tablespoons of ground linseed with meals. Take long walks; have large vegetable salads; squat for bowel movements (see Step 21). For bloating or excess wind avoid any sweet food with or after vegetables or fibre-rich food; use correct food-combining with a low-allergy diet; use magnesium chloride and practice intestinal sanitation.

With hiatal hernia (with symptoms of belching, heartburn, and indigestion), sit at one end of a massage table (or something similar), your legs outstretched on the table. A helper lies across your legs or holds them down while you lean back over the edge as far as possible. Tighten your abdominal muscles to pull yourself slowly back up to the sitting position. A second helper may assist. In most instances, this will pull the stomach back to its normal position.

Epilepsy: A low-allergy diet is advisable. Initially avoid and later minimize consumption of sweet food, wheat, milk, cheese, and food additives, especially excitotoxins. Avoid contamination from copper (new water pipes), aluminium (cooking), mercury (amalgam fillings), and lead. Take vitamin B6, vitamin E tablets, and magnesium in high doses; alkalize your system with calcium and magnesium; frequently suck on a tablet of magnesium orotate; take lecithin, glutamine, MSM, taurine, black cohosh, and copper salicylate; try raw egg yolk, mussel, or oyster extract.

Have a chiropractic neck adjustment; do head and neck exercises; do the herbal parasite cure; sanitize your bedroom. Minimize emotional stress; reduce your exposure to fluorescent or flickering lights, television, and computer games. Try south-pole magnets placed at the temples.

Eye Problems: Helpful here is a low-allergy diet with plenty of fresh grass and vegetable juices. Avoid or minimize intake of dairy products, hydrogenated fats, commercial oils, and sweet foods. For all eye conditions, high-dose vitamins B2 and A (halibut-oil capsules; with suspected fat malabsorption, use vitamin A emulsion) in addition to B complex, C, and grape seed extract are advised; these also help with cataracts and macular degeneration. You need periods with no additional vitamin E intake as this can depress blood levels of vitamin A; however, with wet macular degeneration, try high-dose vitamin E tablets.

Carotenoids, especially lutein and zeaxanthin, are important for the macula; they are found in deep-green, yellow, orange, and red fruits and vegetables, as in tomatoes, squash, corn, and spinach. Have a high intake of a wide range of bioflavonoids; use the herbs eyebright and bilberry internally and as topical eye rinse. Important minerals are magnesium, chromium, manganese, selenium, and zinc.

Chromium, calcium, and a diet low in sweet food are important with myopia (near-sightedness), while magnesium is good for farsightedness. A salty diet is beneficial with glaucoma, as is sodium ascorbate up to 30 g daily (in five divided doses), to drain excess fluid from the eyes. Intensive cleansing combined with energizing or stimulating the eyes in order to unblock congested drainage channels is recommended. For red eyes or conjunctivitis, rinse the eyes with a weak solution of boric acid, possibly combined with strained eyebright and chamomile herbs as tea.

For viral infections of the eyes, make up a solution of 2500 IU of vitamin A emulsion and 250 mg of sodium ascorbate per ml of water. Use one or two drops in each eye every three to four hours. If it stings too much, then dilute as needed. Keep refrigerated and make a fresh remedy every week. Combined with sufficient vitamin B2 (up to 100 mg three times daily); this also helps with allergy-induced eye problems. For clogged arteries in the back of the eyes, try potassium iodide as explained in Cardio-vascular Disease.

Use a raw grated potato pack in gauze for irritated eyes and warm castor-oil packs for weak eyes. You can also put a drop of castor oil in each eye at bedtime for gradual strengthening. Be outside as much as possible without glasses or sunglasses; look at the blue sky, but avoid eyestrain. Minimize watching television, staring at computer screens, and reading by electric light. While reading, frequently interrupt your vision, look around, and cover eyes with palms. Preferably use natural daylight or full-spectrum bulbs; minimize use of fluorescent lights except if they are special daylight types. Press tender muscles in the back of the neck.

Do eye exercises: look up, down, sideways, diagonally in quick succession, circle, alternate looking at a near and a far object. Do head and neck exercises, reflexology, or rapidly blink at the sun when it is low in the sky. Frequently alternate looking at an orange or red light bulb of 25 watts and a blue light of 40 watts in a darkened room; then look for several minutes at a green 40-watt light.

Gynaecological Problems: Important here is a diet low in sweet food, fat, and salt, a vegetarian cleansing diet, especially in the case of heavy periods and difficult menopause. Also advisable are sprouted seeds, vegetable salads, grass juice, and a diet high in bioflavonoids and carotenoids as from coloured foods, especially yellow, orange, red, and purple. For heavy bleeding, black cherries are best; if not available, try black or purple grapes, juice of dark grapes or red beets, blackberries, and blueberries.

SUPPLEMENTS: Take B-complex vitamins, especially B6 and folic acid, rutin and bioflavonoids, vitamin E tablets (1000 to 2000 IU daily), and selenium-methionine, especially if breasts are tender and for menopausal hot flushes. Neutralize fruit acids with dolomite for a high calcium and magnesium intake. Take additional calcium if there is excessive bleeding and fatigue; take more magnesium for tension and irritability. Also helpful are zinc, manganese, MSM, Royal Jelly, bee pollen, evening primrose oil, vitamins A and C, and organic iron. With menstrual clotting, more vitamin K is needed; use plenty of dark-green vegetables, possibly injections of 15 mg of vitamin K1 (derived from plants). Fibrocystic breast disease and ovarian cysts often disappear with six to eight drops of saturated solution of potassium iodide (SSKI) taken daily in some water for two to three months. Prolonged use can inhibit thyroid functions, so use with professional supervision; see Thyroid Problems.

HERBS: For menopause: chamomile, hawthorn, red clover, and sage. Excessive bleeding: bayberry bark, red raspberry, shepherd's purse, and witch hazel. Suppressed menstruation: chamomile, catnip, fennel, pennyroyal, and peppermint. Generally try motherwort, dong quai, raspberry shoots, and chasteberry (Vitex agnus-castus); St John's Wort for depression; and black cohosh for period pain. Maca is most helpful for all conditions, helps after hysterectomy, is safer than hormonal replacement therapy, and is also recommended for age-related decline of sexual function.

ADDITIONAL THERAPIES: Take a hot sitz bath or sit on the north pole of a strong magnet for general strengthening. Press foot reflex areas on heels; use blue lighting to reduce bleeding, orange lighting to increase it. Place castor-oil packs on the lower abdomen and lower back, or Epsom-salt packs; apply pressure therapy and manipulation and adjustment of lower back; do spinal tapping of lumbar vertebrae or colonics.

VAGINAL YEAST INFECTION: This is caused by the yeast Candida albicans. Douche with either diluted hydrogen peroxide or garlic water to which may be added several drops of tea tree oil (shake well before use), tea of sage, and juniper berry, or pau d'arco. After douching, apply acidophilus in the vagina and around it; repeat as often as required and have intestinal fungicidal treatment (see Candidiasis).

Headaches and Migraines: Do the Basic Cleanse followed by allergy testing and maintaining a low-allergy diet. Minimize intake of sweet food, dairy products, wheat and other gluten products, egg, oranges, tomatoes, food additives, fatty food, chocolate, Brazil nuts, dried coconut, aged cheese, mushrooms, pickled herrings, pork, excitotoxins such as monosodium glutamate (MSG), and tobacco smoke. Avoid environmental chemical and electromagnetic pollution; sanitize your teeth (see Step 8).

Try taking magnesium if your blood pressure is elevated, or calcium and salty foods if it is low. Have a strong coffee at the onset of symptoms (if you're not addicted to it). The herb feverfew may help. Correct any constipation or intestinal inflammation. Do head and neck exercises; pressure massage the neck and shoulder muscles and around the base of the big toes; have a neck adjustment and exercise your scalp muscles. Avoid stress; rest; get fresh air; walk; and do deep breathing. During a migraine attack, imagine your arms becoming warm, possibly soak them in warm water. Riboflavin (vitamin B2), up to 100 mg four times daily, has been found to prevent most migraines; use together with B-complex.

High Blood Pressure: Most important are low-allergy diet or allergy testing, low sodium intake, and high intake of potassium and magnesium; for females also calcium. This also helps to prevent strokes. Improve kidney functions; follow tips for Cardiovascular Disease.

Hyperactivity (ADD and ADHD): Allergy testing and a low-allergy diet are advisable. Minimize or avoid sweet food, especially chocolate, cow's milk, cheese, egg, pork, citrus, soft drinks, artificial food colouring, flavouring, excitotoxins, preservatives, chemical sprays, petrochemical fumes, detergents, strong perfumes, smoke, stimulants, and aspirin. Avoid fluorescent and flickering lights and games, television, and stressful situations. Minimize electromagnetic pollution (see step 8). In regard to sweeteners, avoid sucrose or fructose in combination with glucose; see "Fructose and the Disaccharide Effect" under Step 40.

In several clinical trials, avoiding food and chemical allergens resulted in a 100 percent cure rate among test subjects. Frequently, Candida overgrowth in the intestines (commonly following a course of antibiotics) is the original cause of multiple allergies, so treat for Candida. Zinc and magnesium supplements are helpful, as are vitamins C, B6, pantothenic acid, nicotinamide, folic acid, and B12. For attention deficit, try ginkgo biloba.

Infections and Inflammations: Infections can be localized, for example, in the liver or throat, or generalized, as infectious diseases (including venereal diseases). They can be acute with sudden strong inflammations and fever, or they can be chronic and long lasting. Colds occur only if there is mucus congestion, and they are usually beneficial for removing excess mucus. Accelerate recovery by actively expelling mucus from your lungs and head spaces, including the sinuses. Colds can be avoided by using a diet that does not cause mucus congestion or by deliberately expelling mucus from the body (see Mucus Problems).

DIET: During acute conditions, semi-fast with only fresh vegetable and grass juice, spirulina and bee pollen, vegetable broth, strained rice water, and fruits (neutralize acids if the body is sensitive). Later, also consume easily digestible, high-quality food, such as sprouts, raw egg yolk, vegetable salads and cooked vegetables, rice, and bananas. Select foods with a favourable calcium-phosphorus ratio (see Step 6). Avoid gluten products, sweetened, processed, microwaved, and genetically modified food, and any food to which you are allergic or sensitive. With inflammations maximize omega-3 fatty acids (fish oils and ground linseed), and minimize omega-6 fatty acids (polyunsaturated seed oils).

SUPPLEMENTS: In acute conditions, try bowel-tolerance doses of vitamin C with sodium ascorbate; otherwise use 1 to 10 g of vitamin C daily, magnesium chloride if available, propolis, copper salicylate, and Lactoferrin. Vitamin A is especially good with respiratory infections; zinc is best as zinc sulphate lozenges. Improve the immune system with pantothenic acid, folic acid, vitamin B6, manganese, selenium, chromium, acidophilus culture, and freeze-dried thymus.

With viral infections, use a high intake of antioxidants in addition to selenium and zinc. Take 2 g five times daily or bowel-tolerance level of sodium ascorbate; 15 to 20 mg of zinc three times daily; 200 mcg of selenium; up to 2 g of L-lysine three times daily; and up to 200 mg of alpha-lipoic acid three times daily (alpha-lipoic acid is also called thioctic acid; prolonged intake over 100 mg may require extra biotin). In addition, take one tablespoon of cod-liver oil, other antioxidants such as grape seed extract, and Microhydrin or strong electrolytic reduced water; with hepatitis, also take the herb silymarin. A saturated solution of potassium iodide (SSKI from a pharmacist) is an all-round antimicrobial. As a precaution when travelling, take about ten drops in a small glass of water, and more, temporarily, with an acute infection. **Caution:** High amounts for more than a few weeks can suppress thyroid function.

With chest infections, in addition to the above, temporarily take 50,000 IU or more of vitamin A (under professional supervision). Try coughing up mucus after using a hot onion pack on the chest (see Step 21).

HERBS: These include echinacea, goldenseal, marigold, olive leaves, pau d'arco, prickly ash, and wormwood; also chamomile, elder flowers, and peppermint. With respiratory infections, try chamomile steam inhalation; gargle with diluted tea tree oil. Do the herbal parasite cure with wormwood, clove, and green hulls of black walnut. Use aromatic oils for infections: borneol, lemon, clove, eucalyptus, lavender, mint, or thyme.

ANTIBIOTICS: The main reason for a destructive fungal infection is antibiotic therapy. Candida albicans is commonly present in almost everyone, but is usually kept under control by friendly bacteria. If these bacteria are killed by antibiotics, fungi have the opportunity to spread and will seriously weaken the immune system, thereby contributing greatly to the development of the various immune-deficiency diseases, including cancer.

Other disadvantages of sustained antibiotic therapy are B-vitamin and vitamin-K deficiencies and malabsorption, caused by the destruction of intestinal flora. In addition, allergic reactions to the antibiotics can occur or boils may develop as a cleansing reaction even weeks or months later. For these reasons, if possible, antibiotics should be used only when serious infections cannot be controlled by natural therapies. With any antibiotic therapy, use vitamin and mineral supplementation, a diet high in raw food, probiotic bacterial cultures, and a natural fungicide such as fresh garlic.

FEVER: Fever is a great tool in fighting infections by killing heat intolerant germs and strengthening the immune response. Keep your body temperature up with hot baths, hot teas, and sweating. Body temperatures above 105° F can be reduced with cold packs placed around the calves, and white willow bark (aspirin).

PROPOLIS: This is an excellent natural antibiotic from the beehive. For best results, it should be applied directly to the site of the infection whenever possible. For a sore throat, keep some propolis powder constantly in the mouth (between the gum and cheek) and gargle with a propolis solution; or make it into a paste and apply it to a sore tooth. For sinus problems, sniff it up the nose; for earache, put it in the ear; for skin infections, apply it directly to the skin; for internal infections, of course, you have to swallow it. Note your reaction, as some people are allergic to propolis. You can make your own propolis tincture: Freeze-dry raw propolis (from a beehive), pulverize, and keep it in alcohol for one to two weeks exposed to the sun; shake frequently.

THYMUS: Located in the chest, the thymus is the key gland of our immune system. Most strengthening for the thymus gland are propolis, manganese, and freeze-dried thymus extract. To strengthen the thymus, frequently tap with the fingers at the upper end of the breastbone; maintain a good posture and walk with long strides and swinging arms. Apply lemon (yellow-green) color therapy on the upper breastbone (see Step 17).

COPPER SULFATE (BLUESTONE): Available from gardening supply outfits, this has strong antiseptic and anti-inflammatory properties. Collect any discharge from an infected or inflamed organ or wound in a medium-strong solution of bluestone and keep it there for several days. This may be phlegm, pus, bloodstained bandage, vaginal or nasal discharge, or urine (in bladder or kidney infections). This can promote healing by way of secondary contact and can also be used with an extracted tooth or any severed part of the body. This method works on the same principle as radionics and telepathy with a two-way exchange of information energy between the body and any separated part.

ADDITIONAL THERAPIES: Use colloidal silver or copper, oxygen therapy, and an electronic zapper, and possibly a magnetic pulser (see Step 19). You can also use a hot bath (with Epsom salt) followed by sweating; a hot foot bath with mustard powder or cayenne; cold packs placed on the abdomen and around the calves. For colds or flu, immerse hands and lower arms in hot water for 30 minutes; keep the water hot; in addition, you can chill the underside of the big toes with ice dipped in salt. Generally, blue light focused on the site of infection is helpful, as is blue-radiated water taken internally. Try also Epsom-salt purges, sweating, induced vomiting, colonics; packs of cabbage leaves or Epsom salt placed on localized infected or inflamed areas; autosuggestion and visualization or right-brain exercises: tell your immune system to produce plenty of efficient B cells and T cells.

Insomnia: DIET: Have early, light evening meals, using animal protein only sparingly; minimize sweet foods, salt, alcohol, caffeine, and other stimulants taken in the afternoon or evening; test for food allergies.

SUPPLEMENTS: Advisable are vitamins B6 and C; nicotinamide (1 g at bedtime), especially if experiencing difficulty returning to sleep after awakening; inositol; lecithin; one teaspoon of milk of magnesia or magnesium chloride at bedtime. Keep a tablet of magnesium orotate in the mouth when sleepless, take L-tryptophan (an amino acid, 500 mg) or melatonin (1 to 3 mg) at bedtime. Vitamin B12 can help to establish a normal circadian rhythm. Try the herbs passiflora, valerian, or thyme, or hops in a muslin bag placed on the chest or pillow; it is not recommended to use valerian on a long-term basis.

ADDITIONAL THERAPIES: Try a hot relaxing bath, possibly with Epsom salt, before bedtime or a cold pack on the abdomen when trying to fall asleep. The body temperature needs to drop during sleep, therefore use relatively light covers and keep the bedroom cool (but not cold). Sleep with the head in a north or north-easterly direction; use blue sheets, bed covers, curtains, and pyjamas; blue, violet, or purple lighting at bedtime; sleep in complete darkness on a natural-fibre mattress; remove drugs and chemicals from the bedroom.; and minimize electro--pollutants (Step 8). Expose yourself to bright light soon after rising (best is sunshine), but minimize bright light (including bright TV and computer screens) exposure in the evening.

To facilitate sleep, use autosuggestions when going to bed and/or relaxation exercises or visualize a peaceful scene. Do eight-two breathing (Step 9) or inhale through the nose, exhale through the mouth. Repeatedly take three deep breaths, then hold the breath as long as possible; finally continue with slow, shallow breathing to retain more carbon dioxide.

Most effective is plenty of outdoor activity, ideally done without wearing prescription glasses. Try a long slow walk in the late afternoon or early evening; alternatively, undertake other restful activities, such as listening to soft music or meditating. Do not read anything interesting; engage in no other absorbing mental activity or exciting movies; clear up any worries. Ask your lower self (the subconscious level) for cooperation: Tell it to stop thinking of a particular problem and that you will deal with it during the daytime, and then do that. Keep pen and paper beside the bed to write down any important thoughts to deal with the next day.

If you like to experiment, try the Eeman screens: Place metal wire screens (approximately 16 inches by 10 inches, from a hardware store) under your head and the base of the spine. The screens can be enclosed in a pillowcase. With an electric wire, connect the head screen to a piece of metal pipe held in the left hand, and the buttock screen in the same way to the right hand. Use the screens for ten to 30 minutes at bedtime, lying on your back with the ankles crossed.

This balances our internal bioelectric potentials and is conducive to increased tranquillity, meditation, and sleep. Basically, you can balance without screens by placing the right hand under the base of the spine with the palm facing up, and the left palm under the top of the spine. Because the arm muscles are partly contracted in this position, it is not as relaxing as with the screens.

Kidney Problems: You need a diet low in protein, salt, food additives, and drugs, along with allergy testing and a low-allergy and blood group diet. Use predominantly raw foods with plenty of sprouted seeds and fresh (mainly green) vegetable juices, ground linseed, and fish oils; try a watermelon fast. Helpful are antioxidants, spirulina, and coenzyme Q10. For kidney stones, use lemon juice and grapefruit, ascorbic acid, vitamin B6, magnesium, and zinc, but restrict your calcium intake. Useful herbs are buchu, cornsilk, dandelion, and uva ursi; as diuretics, try juniper berries, elderberries, or peach leaves. Colonics with water low in minerals delay the need for dialysis. You need sufficiently high fluid intake to produce about one gallon of urine daily. Most effective in serious conditions is urine therapy (see Step 22).

My favourite way to reactivate kidney function is drinking, slowly and at intervals, about three quarts (litres) of high-quality pure water (no chlorine or fluoride) before breakfast. To each quart, add up to a cup of fresh urine and, if available, the content of one capsule of Microhydrin or Microhydrin Plus. If already on dialysis, then use distilled or de-ionized water, but only with these additions, and increase the volume of drinking water gradually over successive days to match the output of urine. Instead of Microhydrin, you can also use strongly electrolytic reduced water (ERW). While the addition of Microhydrin or ERW is not essential, they provide strong antioxidant activity that helps to regenerate kidney function.

Liver and Gallbladder Problems: Detoxify with the Basic Cleanse, then from time to time, possibly once a month, repeat the first part of it (see Step 4). Have three days on apples only; in the evening activate the liver with a castor-oil pack; at bedtime take half a cup of olive oil with lecithin and possibly orange juice, and next morning a glass of warm water with a tablespoon of Epsom salt. This helps remove gallstones and cholesterol deposits from the liver and gallbladder.

Minimize your exposure to chemicals and drugs and sanitize your intestines. Helpful nutrients are lecithin, milk thistle or silymarin, antioxidants, B-vitamins, especially B6, B12 if you're low in energy, magnesium, selenium, and zinc. MSM or L-cysteine and L-methionine help to detoxify. Stimulate the liver with bitter teas; use half a cup after meals; or take capsules of centaury, dandelion, devil's claw, gentian, goldenseal, or rosemary. Also do the herbal parasite cure. With hepatitis, use the therapy for viral infection in Infections and Inflammations above.

With persistent problems, continue with the castor-oil liver packs. Also effective are reflexology, ear acupuncture, and acupuncture meridian therapy. Jaundice can be caused by drugs and chemicals, liver infection, gall-duct blockage, and increased destruction of red blood cells (protect with high amounts of vitamin E), so try to discover the cause of the jaundice. Treat systemically with blue color therapy (see Step 17).

Malabsorption: Malabsorption of food is frequently associated with chronic degenerative diseases, debility, and underweight or difficulty in gaining weight. This can result in floating, bulky, pale, and smelly stools and excessive wind. Common causes and solutions are as follows:

- Gluten allergy: Avoid gluten.
- Deficiency of gastric acid, pancreatin, bile, or sodium bicarbonate: Take appropriate supplements, reduce fats, use extra-virgin olive oil instead, and improve the corresponding organs.
- Inflammation of the intestinal tract: Individuals commonly have whitish inner irises as with gluten allergy; do intestinal sanitation (see Digestive Disorders).
- Overgrowth of bacteria or fungi in the small intestine: Use isotonic garlic flush followed by acidophilus and bifido cultures.
- High-fibre diet: Use white instead of brown rice.
- Worms or other parasites: Do a three-day raw food diet; take teas of chamomile, sage, or vervain, pumpkinseeds, or pumpkinseed oil. Most effective is an herbal parasite cure with wormwood, cloves, and walnut.
- Tapeworm: Use the herbal parasite cure; in addition, try a three-day water fast followed by castor oil.

GENERAL RULES: Eat small meals of easily digestible foods, for example, bone broth, meat broth or juice, vegetable and grass juices, rye sourdough bread instead of yeast-baked bread. Add pollen, spirulina, or chlorella, and possibly amino acid and nucleic acid supplements if you are debilitated. Add lecithin to all meals. The amino acid glutamine helps to improve the intestinal wall. Chew all your food extremely carefully and keep it in the mouth as long as possible, paying attention to the developing flavours of each mouthful. Frequently rub the body with cod-liver or halibut-liver oil, extra-virgin olive oil, and vitamin E oil. Emulsify oils for intended use (especially extra-virgin olive oil and cod-liver oil) by shaking in a jar with lecithin and juice.

Chew all suitable supplements and keep them under the tongue, such as freeze-dried liver, Royal Jelly, and halibut-liver oil; keep a hali-but-liver oil capsule in the mouth at bedtime. In this way, B12 can be absorbed in therapeutic quantities even if the usually necessary intrinsic factor is missing in the stomach. Most tablets can also be crushed, and capsules opened and mixed with meals. Use vitamin E in natural form as tablets, not as oil-filled capsules. In difficult situations, nutrients can be absorbed from grass juice enemas through the large intestine.

Mucus Problems: These include most respiratory diseases, recurring colds, hay fever, coughs, sinusitis, running nose, throat infections, and many ear problems. However, there may also be a mucus covering in the small intestine, causing malabsorption and, the worst mucus-related disease, cystic fibrosis. Generally, pathological mucus results either from a -lactose/galactose overload or from irritated mucous membranes. Such irritation may be due to food, such as gluten, or to food allergy or chemical sensitivity; it can result from persistent inhalation of irritants or from chronic microbial or parasitic infestations. Try to find and eliminate the cause.

DIET: Fenugreek is good; minimize sweet and fatty food; fruit can be used according to your sensitivity; however, oranges frequently aggravate respiratory problems. The main mucus-forming foods are lactose (skim milk or whey powder); milk products in general according to their lactose content; beer; mainly wheat gluten, to a lesser extent also gluten in rye, oats, and barley; egg white; and moulds and fungi or any food to which you are allergic. Neutralize fruit acids with dolomite.

Initially avoid all mucus-forming foods. After your mucus has cleared, introduce and observe the effect on your body of each of the mucus-forming foods or food groups separately for several weeks. Some foods are mucus forming even in small amounts while others can be tolerated up to a certain level. However, the mucus-forming potential of different foods may be cumulative.

HERBS: For inflamed or infected mucous membranes, use goldenseal, marshmallow, sarsaparilla, slippery elm powder, thyme, and yarrow (strengthening). To expel mucus, use coltsfoot, horehound, lungwort, mullein, nettles, vervain, and violet; most recommended is horseradish.

AROMATIC OILS: Inhale these for sinusitis: cinnamon, lemon, eucalyptus, lavender, pine, rosemary, or thyme. For inflammatory sinus and breathing problems: Inhale vapour of eucalyptus oil during the night (keep some in a flat container close to the bed). For sore throat, gargle with lemon, geranium, hyssop, sage, and thyme. As an internal antiseptic, use diluted tea tree oil. Add a few drops to a small jar or bottle half filled with water, shake vigorously for a few seconds, and then try a drop on the tongue. If this is not irritating or unpleasant, you can ingest the rest. Do not use other aromatic oils internally without competent advice that it is appropriate to do so.

ADDITIONAL THERAPIES: For nose and sinus problems or hay fever, frequently draw seawater, diluted with about three parts warm water, through the nostrils, one at a time, and spit the water out. If seawater is not available, dissolve a level teaspoon of sea salt in a pint of water; you can also add a small amount of Epsom salt. For acute conditions, sniff up a tea of the indicated herbs, preferably adding propolis solution; place castor-oil packs over the sinuses. Do head and neck exercises such as this one for clearing mucus from head spaces (e.g., sinuses): Kneel with the head to the floor and make loud snoring and snorting noises at the back of the nose. Spit out mucus; continue for ten minutes. Have any mercury amalgam fillings or dead teeth removed.

For throat problems, gargle with the indicated herb teas, adding propolis in the case of infections; place castor-oil packs and blue lighting on throat; for chronic conditions, have plenty of singing and vocal therapy. Generally, strengthen the intestines, and do colonics. Edgar Cayce recommended wearing a piece of carbon steel, preferably in the groin pocket, to improve the mucous membranes of the nose and throat and to resist colds and congestion. Use systemic lemon (yellow-green) color therapy to loosen and expel phlegm, and concentrate it on the chest and upper back.

THE SUGAR CURE: Keep a teaspoon of fine (Caster) sugar in the mouth until it is dissolved, then spit it out and take another teaspoonful. Continue with this at intervals for one or two hours and repeat on subsequent days or weeks as required. The sugar draws mucus dissolved in lymph fluid from the lymph glands and so gradually clears the head spaces. It can stop a cold within hours.

NETI: This is a yoga method to rinse the sinuses. Use a coffee or teapot with a fine snout, a neti pot, or an enema bag filled with mildly warm and salty water. Bend over a sink with the head turned to the side, insert the enema hose or the snout of the pot into the upper nostril and let the water flow out. It will rinse out the sinuses and flow out from the lower nostril. After half the water is used up, turn the head and let the water flow into the other nostril. Finally, close one nostril at a time and strongly blow out air to expel any water remaining in the sinuses. Note: You can easily obtain neti pots from yoga and meditation supply sources in the U.S.; check the Internet.

Multiple Sclerosis: DIET: Recommended are plenty of low-mercury seafood, broth of fish heads and bone broth, and possibly raw egg yolk (after allergy testing); sunflower kernels, almonds, and other oily seeds (best sprouted or as seed cheese); freshly ground linseed and possibly linseed oil; plenty of grass and vegetable juices as well as sprouted seeds (but no gluten grains). Try sprouted broad beans. Eat more vegetables that grow below than above the ground; eat raw vegetables often with added gelatine (broth of fish heads or bones). Sprout or soak lentils and beans before cooking; eat tomatoes, fresh fruits, cold-pressed sunflower oil, and cayenne.

Test for allergies and follow your blood group diet. Try to obtain organically grown food and eat small meals. Sago, tapioca, taro, and arrowroot are fine, but minimize your intake of rice, millet, and maize and avoid saturated, hydrogenated, and animal fats; cow’s milk products; gluten products; egg white; smoked, cured, and processed meat; peanuts; sweet foods; tinned or frozen fruits and vegetables; salt; and food chemicals, especially colourings and aspartame. Use a strict low-fat diet.

SUPPLEMENTS: High doses of all vitamins are good, especially C (10 to 20 g daily), E (2000 IU, daily, as tablets), B1, B6, nicotinamide, pantothenic acid, biotin, choline and inositol, and B12 (injections or sublingual), as well as grape seed extract. Take magnesium (best as chloride and orotate), zinc, copper salicylate, manganese, chromium, iodine, and selenium. Most important may be vitamin D from sunlight and fish oils, especially cod-liver oil; also try Calcium EAP 3 x 3 tablets. Take hydrochloric acid and digestive enzymes with proteins or a main meal; also take lecithin, kelp, MSM, freeze-dried liver, pollen, spirulina or chlorella, and evening primrose or borage oil (chew the capsules). The amino acid glycine (10 g four times daily) is useful, but not needed if you use a gelatinous broth of fish heads.

HERBS: Especially strong with anti-inflammatory properties are goldenseal, slippery elm powder, and ginger, also devil’s claw, ginseng, gotu kola, pau d’arco, and sage. Try suitable Bach Flower Remedies; do the herbal parasite therapy and follow protocols in Intestinal Sanitation (see Step 3).

ADDITIONAL THERAPIES: Try a strong antimicrobial therapy with electronic zapper, magnetic pulser, colloidal silver, and oxygen therapy. Also useful are spinal massage, tapping, and adjustment; deep muscle massage; frequent castor-oil packs on abdomen and lower back. Improve the liver, gallbladder, kidneys, and endocrine glands; press the foot reflex for the pineal gland.

Get frequent rest periods, avoid stress, have frequent mild sunshine on your skin. Remove any dead teeth; replace mercury amalgam fillings; guard against inorganic copper and lead intake (for example, canned food, water pipes, and car-exhaust fumes). Have frequent muscle tensing with deep breathing; do mind-body exercises; do lemon color therapy, visualization, and autosuggestion; release resentments; meditate on the feeling of love and being forceful (see Step 61).

Myasthenia Gravis: DIET: Try a low-allergy diet, select food according to your blood group diet, and include some beef as stew, broth, or juice, or liver. Follow a diet low in sweet foods, fats, and polyunsaturated oils; use mainly extra-virgin olive oil; mainly below-ground vegetables, especially red beets, and salads with gelatine (broth of fish heads). Have allergy testing for foods; use no egg white; raw egg yolk, sprouted seeds, and grass juice are acceptable.

SUPPLEMENTS: Recommended are high doses of all vitamins, with 2000 IU vitamin E tablets, folic acid, rutin, and vitamin B12 injections. Most important is manganese; initially take 20 mg of chelated manganese with each meal, then reduce after improvement. Also take zinc, selenium, calcium ascorbate (5 to 10 g), and calcium orotate (800 mg); magnesium supplements may be weakening. Also helpful are thymus extract, octacosanol, lecithin, digestive enzymes (pancreatin) or green papaya, glycine or gelatine, and acidophilus and bifido cultures.

HERBS: For muscle strengthening, use Agrimony, juniper berries, and rosemary; also use devil’s claw and tonifying herbs such as ginseng; do the herbal parasite cure.

ADDITIONAL THERAPIES: Concentrate on the thymus gland, try cabbage leaf and castor-oil packs; tap the top of the breastbone with your fingers (see thymus under Infections and Inflammations). Tape a tablet of chelated manganese in the hollow at the top of the breastbone. Do spinal tapping, especially at the base of the neck (C7 and D1). Remove dead teeth and replace mercury amalgam fillings.

Place castor-oil packs over the liver and abdomen, and over the eyes (from ear to ear) in cases of eye weakness. Practice intestinal sanitation, massage, frequent muscle tensing with deep breathing, mind-body exercises, and autosuggestion and visualization methods; meditate on being calm but forceful. Have frequent rest periods, avoid stress, keep cool, take no hot baths or hot showers, and eat no hot foods. Practice yellow color therapy systemically; use the electronic zapper, colloidal silver, and oxygen therapy.

Obesity and Overweight: The main cause of excessive weight is a diet low in fat and high in processed carbohydrates, particular sweeteners, and cereal or grain products. Avoid these and experiment with the following approaches.

- Have an appetite-reducing protein drink, cold or warm, 30 minutes before meals, with any combination of the following ingredients and in any amount that suits you: spirulina or chlorella powder, green barley or wheat grass powder, food yeast, ground linseed, linseed oil or olive oil. You can also drink this repeatedly during the day to keep hunger away for hours.
- Thirty minutes before meals stir one teaspoon of psyllium hulls in a glass of water, drink immediately. If still too hungry at mealtime, double the amount to two

teaspoons in a large glass of water. Drink some more water afterwards. Psyllium swells up to 40 times its dry bulk to fill the stomach with a soft gel.

- Eat only lentils, green vegetables, and sprouted seeds. Vegetarians can also experiment with combinations of different non-gluten whole grains and legumes.
- Adopt a mono-diet: One day eat protein foods only, the next day fruits, and the third day sprouted seeds and vegetable salads. Repeat this as often as required.
- Temporarily adopt a raw food diet (see Step 34).
- For two weeks, eat only protein/fat foods and green vegetables; ground linseed is a good protein/fat food.
- Repeat periods of consuming vegetable and grass juices and vegetable salads.
- Avoid all sweet food, as it stimulates the appetite and the synthesis of fat. In particular, keep fructose-containing foods away from starchy foods, and eat fresh fruits separately. The alternative sweeteners stevia and xylitol are acceptable; however if you have cravings for sweetness, it is preferable to overcome this by avoiding all sweeteners for some time.
- Avoid all cereal or grain products, especially if your blood group is O.
- Use a diet high in unheated fats and cold-pressed oils such as extra-virgin olive oil, linseed oil (or ground linseed), fish oils, and coconut oil; these do not normally cause weight gain. Avoid all heated fats and oils.
- Preferably skip the evening meal or have only a small salad or an apple.
- Get light daily exercise, such as a brisk 15-minute walk or jogging, or use a rebounder or other aerobic exercise. The metabolic rate is speeded up for a long time after exercise.
- Use a low-allergy diet combined with your blood group diet.
- Use mind therapy such as guided imagery and affirmation; look at your belief systems, release negative emotions, and express your emotional needs. Use the Emotional Freedom Technique in Step 63 to overcome food cravings.
- Eat very slowly and peacefully; chew every mouthful for about two minutes, until the food is liquefied; savour the food flavours as they develop during chewing.

Osteoporosis: Use allergy testing and a diet low in sucrose, acidity and phosphorus. Select foods with a high calcium-to-phosphorus ratio. Alkalize the body if overacid and neutralize fruit acids or vinegar with dolomite. Take a hydrochloric acid supplement if your fingernails are soft or if you have other indications of gastric acid deficiency. Frequently add broth of bones or fish heads to your meals; regularly drink grass and vegetable juices. The most important vitamins and minerals are vitamins C and D (cod-liver oil or halibut-oil capsules), manganese, zinc, calcium, magnesium, and boron. It’s helpful to know that 50 mg of borax contain 6 mg of boron; to measure this amount, dissolve one level teaspoon of borax in one quart (litre) of water and take a teaspoon twice daily with meals. Frequently expose your skin to mild sunlight without using sunscreen. Also, vitamin K is important and available from dark-green vegetables.

Pain: Pain is a warning signal indicating that something is wrong. Therefore it is most important to search for and correct the basic cause. The following measures are secondary to this. (See Arthritis or Infections and Inflammations if the pain is due to these conditions.)

DIET: Use a low-allergy or hypoglycaemia diet with light meals, plenty of raw vegetables and their juices, and omega-3 fatty acids, and have cleansing periods. Avoid sweet foods and dairy products (lactose); avoid or neutralize acids; do allergy testing. Pain tends to lessen or disappear while on a raw food diet; introduction of some cooked foods may increase pain again

SUPPLEMENTS: Generally, a high calcium-magnesium level and increased blood alkalinity raise the pain threshold (the level at which pain is felt; this means you feel less pain), while acids, phosphorus, and potassium lower it. Therefore, alkalize the body. You can add some additional baking soda to acids neutralized with dolomite to bring the pH of the solution to 7. Vitamins A, B1, B6, C, and E can be helpful. Vitamin B12 injections work for nerve pains, as do manganese, MSM, and zinc. D,L-phenylalanine slows the breakdown of natural pain-relieving opiates in the brain and improves other pain-relieving measures. It is slow acting and used for long-term pain control. Topically apply propolis, cayenne ointment, and oils of clove or wintergreen. Take white willow bark, cloves, wintergreen, pansy, or feverfew for headaches.

ADDITIONAL THERAPIES: With cancer, migraines, and other diseases, pain and discomfort are often the result of liver congestion. This is usually relieved by a coffee enema and other measures to cleanse the liver. Improve the lymph circulation in a painful area; see Part 1 for details.

Topically place the south pole of a magnet on the pain site and/or apply blue lighting on the painful area, along with cold or ice packs. Also make a pack of honey, cabbage leaves, or Epsom salt; add a handful of sodium bicarbonate to the bath water to alkalize the body. Practice rhythmic breathing, meditation and relaxation exercises, visualizing blue energy extinguishing the “red fireball” of pain. Concentrate on feeling the pain, its strength, extent, rhythm, and other characteristics. If you pay enough attention to your pain for a long period, it may simply disappear. Also ask the pain what message it has for you. Try to sense its message.

Try this form of energy healing: Firmly but without pressure touch the flesh in the indicated places with the tips of the thumb and the first two fingers of the right hand. Hold for about three minutes or until the patient (or you) feels tingling in the fingertips. Concentrate your mind on the painful body area with the intention of relieving the condition. Here are a few variations:

- For pain in head or neck, hold fingertips just behind and beneath the left ear on a line with the mouth.
- For pains in chest, arms, or upper back, hold fingertips to the base of the neck on the left side of the spine.
- For pain in abdomen, back, sides, or hips, put fingertips to the left side of the spine between shoulder blades.
- For pain in legs or feet, hold the left side near the centre of the spine.

Parasites: Parasite infestations combine with fungi, such as Candida, to greatly undermine your long-term health. Examples of common parasites are intestinal worms and protozoa (malaria and Giardia). Most of these parasites can be eliminated with wormwood (Artemesia absinthium), which is best combined with other herbs; wormwood also appears to be effective against Candida.

Here's how: For several weeks take up to one teaspoon of powdered wormwood once daily in juice before a meal. You may combine this with up to a quarter-teaspoon of ground cloves and chopped garlic. If any of these ingredients is too unpleasant to take, you can put it into one or more large gelatine capsules. Take a course of acidophilus and bifido cultures after finishing the course of parasite remedies. As a precaution, you can repeat a course of parasite remedies once or twice a year.

In animal husbandry, high copper levels are successful in eliminating parasites and keeping animals free of parasites. While animals seem to be able to handle copper sulphate as part of a mineral supplement, this is not suitable for humans, and I recommend using either copper salicylate or colloidal copper. During any parasite or Candida cure, maintain a high intake of liquids to flush out toxic residues. In addition to copper and herbal remedies, use an electronic blood purifier and possibly a magnetic pulser on any swollen or infected lymph glands and organs.

Parkinson's Disease: DIET: Useful here are seafood, raw egg yolk, lactic-acid-fermented food, purple food, sprouted seeds, raw vegetables (finely grated, often with gelatine) such as green beans, carrots, beet root and beet tops, cabbage, green peas (fresh, not canned or frozen), and spinach. If vegetables are cooked, use the cooking water. Preferably take fresh vegetable and grass juices before meals; have frequent interior cleanses on such juice, possibly also on fruits. Use food appropriate for your blood group and metabolic type.

Sprouted broad beans are very high in levodopa, the precursor of the neurotransmitter dopamine, lacking in people with Parkinson's. Try five to 15 sprouted beans with breakfast, and use raw or blended in juice; if cooked, also ingest the cooking water. Black currants also contain levodopa. Do allergy testing and intestinal cleansing; use organically grown foods if possible; avoid food additives, including preservatives, flavouring (excitotoxins), colouring, and aspartame. Avoid exposure to aluminium and heavy metals, especially lead, mercury, and iron.

SUPPLEMENTS: Take high doses of all vitamins, especially antioxidants and B1, B6 (but not with levodopa or sprouted broad beans), B12 injections, and natural E (1000 IU daily as tablets), and the minerals selenium, zinc, and chromium; also try copper salicylate. Take vitamin C as 5 to 10 g of ascorbic acid per day, partly neutralized with magnesium oxide or carbonate; take sufficient additional magnesium as chloride and orotate to stop any shaking and help relax muscles. Take digestive enzymes, pycnogenol or grape seed extract, lecithin, fish oil and ground linseed, evening primrose oil, kelp, MSM, royal jelly, and glutamine. Try tyrosine before breakfast and tryptophan before the evening meal. For herbs, try a combination of foti-tieng, gotu kola, licorice (but not with high blood pressure), pau d'arco, sage, and skullcap.

ADDITIONAL THERAPIES: Cleansing and strengthening of all organs and glands with all the methods described in this book are advised. Spinal tapping, spinal corrections, rejuvenation exercises, tensing and breathing exercises, and body-mind exercises are also helpful. Try regression to discover and release hidden fear. Imagine reacting appropriately in situations of panic; use emotional-release and mind-control methods. Improve and balance energy flows in acupuncture meridians, especially those of the stomach, gallbladder, and bladder and the governing meridian, by having an acupuncture treatment. Do intentional exaggerated shaking exercises with vigorous breathing. Remove dead teeth and replace mercury amalgam fillings.

Before breakfast have strong coffee and tyrosine; with breakfast have bee pollen, spirulina or chlorella, ground linseed, sprouted broad beans, ginkgo biloba, ginseng, and other stimulating measures; reintroduce levodopa drugs as late as possible in the day. If levodopa has become ineffective, have a "drug holiday" - stop using it for one or more weeks. Try breathing additional oxygen during this time (e.g., from an oxygen concentrator); however, additional oxygen is harmful while you're on levodopa. See the Love Cure (Step 61) and learn to meditate and energize the brain.

Prostate Problems: DIET: Advisable here is a mainly raw food diet based on sprouted legumes, vegetable salads, fresh fruit, and plenty of fresh vegetable and grass juice. Also good are raw egg yolk, tomatoes, seeds of sunflower and pumpkin (best as seed cheese), foods high in bioflavonoids and carotenoids, and purple foods. Have allergy testing and practice intestinal sanitation. Avoid gluten and milk proteins.

SUPPLEMENTS: Take Vitamins A (halibut-liver oil capsules and cod-liver oil), E (preferably including gamma tocopherol), C, and B6; magnesium (1000 mg daily) best as magnesium chloride and magnesium ascorbate, selenium, zinc (30 to 50 mg in divided doses), bioflavo-noids, quercetin, kelp, protein-digesting enzymes, spirulina, and bee pollen. Try Epilobium parviflorum (small-leafed willow), saw palmetto, Pygeum africanum, and stinging -nettle/nettle extract.

ADDITIONAL THERAPIES: Useful are yoga exercises; frequent long walks, and mild sun exposure; sitting on south pole of a strong magnet. Try also: reflexology of prostate points on hands and feet, tapping of lower spine, and -castor-oil and Epsom salt packs on the lower abdomen and lower spine. Have satisfying sexual activity or be satisfied with having no activity; see also Step 56: Healthy Sexuality.

Schizophrenia: Many people classified as having a mental disease are actually emotionally unstable and open to psychic influences (extrasensory perception) and, most of all, need guidance and grounding. Sometimes schizophrenic symptoms disappear during water fasts. For all types of schizophrenia, follow a fast with allergy testing, a low-allergy diet, and intestinal sanitation. If allergy testing is not possible, adopt a low-allergy diet that excludes all gluten and cow's milk protein.

From a biochemical point of view, we can distinguish between three different types of schizophrenics: The low-histamine type are slow oxidizers, also called histapenics; the high-histamine type are fast oxidizers, also called histadelics; and there are balanced oxidizers with normal histamine levels. In addition, each of these may at times display symptoms of overactivity or underactivity. If in doubt about the histamine status, under supervision of a qualified health practitioner, take (or give) 50 mg of niacin with water on an empty stomach. If a strong facial flush develops, histamine levels are high; if there is no or only a weak flush, levels are probably normal or low. Then try 250 mg of niacin in the same way; if there is still no or only a weak flush, histamine levels are low.

The Low-Histamine Type: The low-histamine type is the "typical" schizophrenic, insensitive to pain and outside influences; infections and skin reactions are rare; calcium levels are high; pH tends towards the alkaline side. Sometimes a cerebral allergy is present, causing a rapid pulse (90-100); thought processes are often overactive.

DIET: Adopt mainly a vegetarian raw food diet reasonably high in protein. Frequently use seed cheese and protein drink as described in the recipes in Part 4; use seafood, fish oils, linseed oil, ground linseed, and acid fruit, but avoid gluten and milk products. Take a long, supervised fast on fresh grass and vegetable juices.

SUPPLEMENTS: Try high-potency B-complex vitamins, high dose folic acid with each meal, inositol, and vitamin B12 (as injections or absorbed under the tongue). Also take up to several grams of niacin and niacinamide daily; up to 10 g daily of ascorbic acid, partly neutralized with magnesium in divided doses; additional magnesium chloride, manganese, and zinc; high-dosage evening primrose oil and lecithin, glutamine, histidine, and tryptophan; hydrochloric acid and digestive enzymes (pancreatin) with main meals; Bach Flower Remedies per your emotional conditions.

ADDITIONAL THERAPIES: Use blue lighting during overactive periods; at other times try orange or red. Induce the inflammation response: for example, blistering agents and overheating baths followed by sweating. Use plenty of acid fruit, such as citrus fruit.

The High-Histamine Type: The high--histamine schizophrenic exhibits a hypoglycaemic personality and is very susceptible to allergic reactions. The main symptoms include suicidal depression, obsessive rumination, blank-mindedness, periods of over-arousal and thought disorders, frequent headaches, and dizziness; blood pressure tends to be low and the pH too acid; skin is sensitive and may show dilated blood vessels (rosy cheeks).

Recommended is a strict hypoglycaemia diet that minimizes sweet and acid food intake. Have high doses of all vitamins and minerals; acid-neutralized dolomite or additional calcium, and calcium ascorbate. Copper may be beneficial, but best taken as salicylate. Also helpful are vitamin-B12 injections, nicotinamide (up to 10 g daily, no niacin), L-methionine, glutamine, folic acid, and lecithin. Alkalize the body (see Step 6). For chronic depression, use tryptophan or tyrosine; if one supplement does not help, try the other.

Stress-Induced Schizophrenia: People with stress-induced schizophrenia tend to have a better insight into their condition and a better response to treatment. In addition to the classic symptoms of schizophrenia, in this form there are signs of zinc and vitamin-B6 deficiency; also white spots on fingernails, loss of dream memory, sweetish breath odour, stretch marks, inability to tan and sensitivity to sunlight, sometimes pain in the upper left abdomen, possibly tremors, spasms, amnesia, impotence, menstrual irregularity, and anemia.

Usually a hypoglycaemia diet is helpful, along with vitamin B6 (up to 3 g daily in divided doses), zinc (10 mg) with each meal, vitamin-B12 injections, high intake of neutralized dolomite, additional magnesium, other vitamins and minerals as for the high--histamine type, and Bach Flower Remedies, as appropriate.

GENERAL RECOMMENDATIONS FOR ALL TYPES: Live with a caring group in a peaceful, natural environment; try regression therapy and emotional release therapy (see Step 63). Avoid heavy-metal contamination; sanitize your teeth; apply violet color therapy systemically; treat for parasites and Candida; use St. John's Wort as an herbal tincture during depression.

Scleroderma: Antimicrobial therapy including colloidal silver and oxygen therapy are recommended, as is the herbal parasite cure with use of the electronic zapper and magnetic pulser. Make external applications to affected skin areas as rubs, packs, or immersions using tea tree oil, aloe vera gel, magnesium chloride, a solution of copper salicylate with MSM, and blue light. To eliminate the microbes causing the condition, keep the affected part in contact with tea tree oil until the skin becomes red, inflamed, and painful; then stop tea tree oil and expose it to blue light, copper salicylate, and other sedating or anti-inflammatory measures. Internal remedies include all antioxidants, bioflavonoids, copper salicylate, echinacea, magnesium chloride, pau d'arco, propolis, wormwood, and zinc. Also advised are the Basic Cleanse, allergy testing, and intestinal sanitation.

Skin Problems: The three main causes of skin problems are accumulation of wastes and toxins, unrelieved allergies, and a deficiency of essential nutrients.

DIET: Use the hypoglycaemia diet, with freshly ground linseed, fish oils, apricot oil, almond oil, almonds, sunflower kernels, raw egg yolk (sulphur), raw cabbage, horseradish, sprouted seeds, grass and vegetable juice, raw onions, and broth of fish heads. Avoid saturated fats, margarine, peanut butter, all hydrogenated and fried foods, dairy products, gluten, and sweet foods. Do the Basic Cleanse, have allergy testing, and maintain intestinal sanitation.

SUPPLEMENTS: Take vitamins A, C, and E (up to 2000 mg daily, as tablets), also B2, B6, B12, PABA, bioflavonoids, biotin, glucosamine, hydrochloric acid (if fingernails are soft), and one spoonful of lecithin with meals. Tyrosine and kelp or iodine help as skin activators: They are cleansing, but may cause temporary aggravation of symptoms, especially with acne. Take increased dosages of magnesium, evening primrose oil (especially in the case of psoriasis), zinc, manganese, sulphur (best as MSM), selenium, silicon as in horsetail, digestive enzymes, and gelatine (broth of fish heads), especially for psoriasis.

HERBS: For internal and external use: aloe vera, burdock (for psoriasis), calendula, chaparral, echinacea, goldenseal, horsetail, pau d'arco, sarsaparilla, sassafras, St. John's Wort, and vervain.

ADDITIONAL THERAPIES: If B-vitamin deficiencies are suspected, use a paste of food yeast or bran on affected skin parts; concentrated B vitamins, especially B2 and B6, can be added to this as well. Have frequent cleansing periods; regulate bowel movements; use colonics. Drink plenty of water and green vegetable juices; use hot and cold showers; do skin brushing to activate the skin, but not on the affected areas. Include sweating after hot Epsom salt bath.

Stimulate the kidneys, lungs, circulation, small intestine (especially for psoriasis), liver, gallbladder, and endocrine glands, especially the thyroid; do lymph-drainage massage. Use only a minimum (infrequently) of an old-fashioned brand of soap, based on olive or almond oil, preferably neutralized (most soaps are too alkaline). Avoid detergents and cosmetics. Apply yellow light and measured amounts of sunshine on affected skin parts; this is especially helpful for acne. Have frequent skin massage with peanut oil, almond oil, apricot oil, and cod-liver oil (in case of malabsorption), preferably with added vitamin E. Cod-liver oil rubs (with vitamin E) are especially helpful with dry, damaged, or inflamed skin, malabsorption, moles, warts, skin cancer, sunburn, and psoriasis.

Rashes are usually caused by allergies, so use allergy testing. Chronic skin diseases, for example psoriasis, are often improved by inducing vomiting, which relaxes the digestive organs and stimulates the solar plexus. Accept your skin problems as a beneficial cleansing, removing harmful energies and impurities from inside. Send love to your skin blemishes and they are more likely to go away. Rejuvenate your skin by frequently bathing it in imaginary rose-colored energy.

Dry skin: Especially if you have difficulty gaining weight, check for fat malabsorption, which is often due to gluten allergy; stools may be pale or clay-coloured. For dry skin, frequently rub it with extra-virgin olive oil and added vitamin E and A. Take linseed oil or fish oils internally; with malabsorption, also apply it externally.

Oily skin: Use a pack of brewer's yeast (also called nutritional yeast) and bran, applied to the skin.

Acne: A mixture of equal parts of camphorated oil, witch hazel, and vitamin E oil can be rubbed into pustular areas. For severe conditions, a 0.05 percent solution of vitamin A acid (retinoic acid) can be applied for about six weeks under professional supervision. It frequently causes redness and irritation at the beginning of the treatment.

Ulcers and skin infections: Wash the area with warm diluted magnesium chloride or an Epsom salt solution and apply any of the following, alone or mixed with others: propolis as cream or solution, copper salicylate, garlic, or fresh cabbage-leaf or fresh green-vegetable juice. Aloe vera gel is excellent for sealing a wound; cover with unheated honey, mix with a few drops of hydrogen peroxide, and renew twice daily. When there is dead, proud flesh (granulation) and gangrene, apply protein-digesting enzymes (trypsin, pancreatin), but especially the inside of green papaya skin, white sap, juice of raw leaves, or urea (see also Step 22: Urine and Urea Therapy).

Boils: To draw, cover with a mixture of honey and cod-liver oil; once they are broken, treat as for ulcers. Internal cleansing is the main therapy for this problem.

Wet eczema: This is best kept dry, either exposed to air and sunshine or covered with a powder of slippery elm. For healing, after infections have cleared, apply vitamin E, fresh green juice or fresh leaves, especially cabbage.

Scar tissue: To remove, frequently rub with camphorated oil, MSM solution, and vitamin E. In addition, scars can be treated with neural therapy. An anaesthetic (Xylocain) is injected into scar tissues that are thought to induce electrical interference fields in other areas of the body. Sometimes a health problem of long standing disappears instantaneously after an injection. Alternatively, you can prick the scar with a fine needle to release the tissue tension in the area. Prick long scars in several places and leave a needle inserted for ten to 20 minutes about 2 mm deep at each end of the scar. Before and after any needling, apply a suitable antiseptic to the skin, such as tea tree oil.

Fungal infections of the skin: Use a concentrated magnesium chloride or propolis solution, tea tree oil, or other fungicidal herbs and oils. Most effective is a long-lasting pack of any of these fungicides; if one does not work, try another. With any skin-fungus problem, also use internal Candida therapy.

Itchy skin: Apply a solution of baking soda or, alternatively, extracts of borage, buckthorn bark, juniper berries, chickweed, marjoram, pennyroyal, plantain, poke root, or witch hazel; also bathe the skin with oatmeal (cooked or soaked) or baking soda. Use blue color therapy; treat the liver and gallbladder if skin is jaundiced. Take a hot Epsom salt bath; rub skin with tea tree oil or witch hazel; alkalize the body; consume no sweets or acids; test for allergies.

Warts and moles: To remove, gently rub on cod-liver oil or castor oil mornings and evenings. For stronger action, apply baking soda moistened with castor oil to the consistency of putty; leave overnight; repeat until it becomes sore. Then apply vitamin E oil, green juice, or honey for several days before continuing with castor oil. Be careful and gentle with moles. If a mole appears to turn malignant, cover with raw, crushed garlic, changing two or three times per day. When it becomes sore after several days, cover with honey instead for five days. Alternate between garlic and honey periods as long as required.

Basically, moles indicate that the systems of elimination are not functioning well. A protruding mole may also be removed by tying a hair tightly around it, then every two or three days adding another hair tying it a fraction tighter. Active moles can also be removed with an escharotics cream.

Skin cancer: If skin cancer is suspected, apply the milky juice from the stem of radiation weed, radium-weed, or milkweed (spurge, euphorbia, petty spurge). Repeat application as required; a strong skin reaction may result, followed by healing. Use this approach also for warts, moles, blemishes, and so on. For serious conditions, including melanoma and tumours close to the skin, escharotic creams based on zinc chloride and the herb bloodroot are available (see: www.altcancer.com or www.cancersalves.com). On inflamed areas and skin cancer, you can also use the south pole of a magnet and blue lighting.

Skin rejuvenation: Frequently and vigorously rub the skin with some or all of the following ingredients, dissolved in aloe vera gel or energized water: vitamin C powder, magnesium chloride, glucosamine, copper salicylate, Microhydrin, and MSM. The amounts of individual ingredients are not critical. Afterwards you can cover the area lightly with natural vitamin E oil, possibly with added vitamin A. At other times, you can make a facial mask with fresh residue from juicing vegetables or grass; moisten with some juice as required to make it stick.

GENERAL RULES FOR COLOR THERAPY FOR SKIN PROBLEMS: Initial treatment is with green; subsequent treatments for dry or scaly skin problems with lemon; for obstinate cases, use also yellow or orange. When scales drop and inner skin surface appears, treat with turquoise, as for moist or weeping skin disorders. For red, angry skin problems, use blue or indigo; for abscesses, boils, carbuncles, and furuncles, use systemic lemon and locally orange. When suppuration and throbbing begin, use yellow until the problem bursts open; then green until pus drains and a red cavity is left; finally, use turquoise and indigo. For acne and especially psoriasis, cover the area with blue cellophane (double layer) and expose to sunlight or ultraviolet light.

Thyroid Problems: For an indication of your thyroid function, measure your temperature before getting up in the morning. Measure it in the armpit or under the tongue for three consecutive mornings. For women before menopause, measure in the first half of the cycle before ovulation in the first week after the start of a new period. The temperature should be 97.8° F (36.5° C) or higher; otherwise, the thyroid is likely to be underactive. However, be sure that you do not have a raised temperature due to an infection.

If the thyroid is underactive, try the following measures: Take a supplement of the amino acid tyrosine, 500 mg three times daily before meals. Best take it together with a kelp or seaweed tablet as a source of iodine. Also helpful are freeze-dried thyroid and vitamin B complex. Selenium is needed to activate the thyroid hormone. The most helpful remedy for normalizing thyroid functions is maca, a Peruvian tuberous root, similar to wild yam but more effective. For sources, check the Internet.

Other useful measures are pressing the thyroid foot reflexes (see Step 15), using the magnetic pulser on the thyroid gland (see Step 19), and sanitizing your teeth (see Step 8). The main causes of thyroid problems, except for iodine deficiency, are dead teeth (e.g., root-canalled teeth), jawbone infections, and mercury from amalgam fillings; these can also cause overactive or toxic thyroid conditions and need to be corrected. For autoimmune disease of the thyroid, see Autoimmune Diseases.

Vaccination Damage: Despite the refusal of the conventional health establishment to investigate the long-term effects of childhood vaccinations, independent research reports have increasingly been made public. These link vaccination rates to the epidemic of childhood asthma, as well as to crib death, shaken baby syndrome, hyperactivity, attention deficit disorder, autism, and juvenile diabetes. Since the 1990's, there has been a tenfold or 1000 percent increase in autism, an increase which has been linked by some researchers to the organic mercury preservative commonly found in baby-vaccines.

A greatly increased incidence of juvenile diabetes has been correlated to specific vaccination sequences and to the number of vaccines given. In some Australian Aboriginal communities, every second child died shortly after vaccination. Also, the original explosive spread of AIDS across Africa correlates with a smallpox vaccination campaign there. Basically, most components of most vaccines are highly toxic, and these are introduced by an unbiological route directly into the bloodstream of babies, whose immune systems are immature.

While the dangers of vaccinations are greatly understated by most health authorities, the advertised benefits are greatly exaggerated. For instance, incidence of the four leading childhood killer diseases - diphtheria, pertussis, scarlet fever, and measles - had already declined 90 to 97 percent before the introduction of vaccines, due to improved sanitation and hygiene.[35]

Another controversy concerns the "shaken baby syndrome." Independent researchers claim that vaccinations can cause spot bleeding in the brain and retina that resembles the injury caused by violently shaking a baby.[36] Paralleling the increase in vaccinations, there is now an epidemic in shaken baby syndrome. A recent article in New Scientist cites a member of the U.S. child protection service as saying that there are three million abused children each year,[37] and if it becomes known that other conditions cause shaken baby syndrome (including the rare blood disease hemophagocytic lymphohistiocytosis), then some child abusers might remain unpunished. I believe that for thousands of years most parents have on occasion lightly shaken their children without causing any harm, and only now does it cause serious brain damage. There may be countless innocent parents in jail, and it would be appropriate for health authorities to investigate the possibility that some vaccines may cause capillary fragility in some babies.

The only complete protection against vaccination damage is strict avoidance of this procedure. Every state in the U.S. has a religious exemption clause; you can write a letter to exempt your child; for more details, see www.mercola.com, and go to the article page on vaccinations. If you cannot avoid vaccinations for your child, then give high doses of protective nutrients for several days before and several weeks after the event, in particular vitamin C and MSM, to aid in rapid detoxification.

Wounds and Burns: DIET: Have a diet high in easily digestible proteins (such as a protein drink - see the recipes in Part 4), fresh green-vegetable juices, foods high in arginine (such as almonds and sunflower seeds, best as seed cheese), and lentils.

SUPPLEMENTS: Take high doses of vitamins C, A, B1, and E, bioflavonoids, manganese, MSM, zinc, glucosamine, copper salicylate, -protein-digesting enzymes, and comfrey root. Use homeopathic Arnica and/or Bach Rescue Remedy, internally and topically; also burdock, chamomile, comfrey, echinacea, and yarrow, as teas and topically.

ADDITIONAL THERAPIES: Cool burns with ice; spray wounds and burns with a one-percent ascorbic acid and propolis solution; alternatively, apply weak tannic acid. After drying the site, apply vitamin E oil and cover with aloe vera gel, honey of a heavy consistency, or ointments containing propolis, vitamin E, and zinc oxide. The principle is to seal the wound or burn airtight with an antiseptic preparation until it is healed. Therefore, reapply the gel or ointment to maintain an airtight seal; renew the honey twice daily; mix a few drops of hydrogen peroxide with the honey. If the wound needs washing or cleaning, a magnesium chloride solution is best. Later, apply camphorated oil and vitamin E oil to prevent scarring. As an alternative, wounds can be covered with crushed fresh green leaves (cabbage, juicy grass pulp). Place the south pole of a magnet on fresh wounds; later, opposite poles on both sides; and finally, north pole. Use green color therapy topically; turquoise if there are signs of infection or inflammation; later lemon or yellow.

PRESSURE THERAPY: Immediately after the injury, exert strong pressure with the thumb on any cut for at least seven minutes. Partially severed parts can be pressed and held together with hand pressure or firm bandages; this usually stops bleeding and accelerates healing.

* * * * * * * * * * * *

Step 45

DEFICIENCY SYMPTOMS

Check out this list of deficiency symptoms to see if you are lacking in key nutrients - and start replacing them.

Eye, hair, nail, mouth, and skin symptoms are among the early outward warning signs of vitamin and mineral deficiencies. The following compilations can help you in diagnosing and treating these deficiencies. However, increased metabolic requirements for indicated deficiencies may persist for a long time after the outward symptoms have disappeared. Many listed symptoms can also be caused or aggravated by allergies and problems with the blood sugar and fat metabolism. Study tables 6-1 to 6-4.

The right-hand column gives the primary deficiencies first; additional treatment or the cause is given in parentheses. For each condition, only the more prominent nutrients are mentioned. However, all the typical nutrients should be supplied in increased amounts in the treatment of all conditions. Most skin conditions benefit from external as well as internal application of the indicated nutrients. EFA means essential fatty acids, mainly linoleic and linolenic acid; recommended are ground linseed (flaxseed), fish oils, and evening primrose oil. Mouth problems are often due to mercury amalgam fillings, allergies, or Candida (thrush) overgrowth.

* * * * * * * * * * *

Table 6-1: Eye Deficiency Symptoms

Symptom	*Deficiency/Treatment/Cause*
Bitot's spots (foamy patches on conjunctiva)	vitamin A
Bloodshot eyes	boric acid for fungus infection, blue light
Blurred vision	vitamins B2, B6, pantothenic acid
Bulging eyes	vitamin E, nicotinamide, iodine
Cataracts (lens becomes opaque)	vitamins B2, C, E, antioxidants (avoid lactose)
Color blindness	vitamin A
Conjunctivitis	vitamins A, B2, C (B6, zinc)
Cross-eyed	vitamins E, C, B1 (allergy testing)
Dark spots in front of the eyes	vitamins B2, B6, C, zinc (liver problems)
Dim vision (amblyopia)	vitamins B1, C, B12 (allergy testing)
Dry, hard eyeballs (xerophthalmia)	vitamin A
Farsightedness (hyperopia)	magnesium, potassium
Glaucoma	salt, high vitamin C (cleansing)
Haemorrhaging in the back of the eye (retinitis, also macular degeneration)	vitamins B6, E, C, B2, zinc, copper, magnesium, bioflavonoids, glucosamine, cartilage
Infected, ulcerating eyes (keratomalacia)	vitamin A (vitamins C, B2, B6, zinc, blue light)
Itching, burning, watery, sandy eyes	vitamin B2 *(continued)*

Symptom	*Deficiency/Treatment/Cause*
Nearsightedness (myopia)	vitamins C, E, D, calcium, proteins, chromium
Night blindness (nyctalopia)	vitamins A, (B2, B6, zinc)
Red blood vessels in the sclera	vitamin B2
Retinal detachment	zinc, vitamins B6, B2, C, E, A
Sensitive eyes, fear of strong light (photophobia)	vitamins B2, A
Tics of eyelids	vitamins B2, B6, zinc, magnesium

Table 6-2: Skin Deficiency Symptoms

Symptom	*Deficiency/Treatment/Cause*
Acne	vitamins A, B2, B6, C, niacin, biotin, zinc, cod-liver oil
Addison's disease (adrenal exhaustion): increased tanning, especially on skin-folds, scars, elbows, knees; black freckles	all B vitamins, especially pantothenic acid; vitamin C; all minerals, digestive enzymes (allergy testing, avoid stress)
Arterial spiders (fine, branching arteries on face, neck, chest)	antioxidants, glucosamine, cartilage, purple food, bioflavonoids, copper, zinc (alkalize), cod-liver oil
Bedsores	vitamins C, E
Blisters	vitamin E
Brown discoloration around small joints	vitamin B12
Brown skin spots	vitamins C, E, antioxidants (liver problems)
Dry skin	vitamins A, C, EFA (fat malabsorption)
Eczema, infantile	EFA, zinc, vitamin B6
Eczema, skin ulcers	vitamins C, B2, B6, niacin, PABA, EFA, zinc, magnesium
Edema	vitamins B6, B1, magnesium (less salt/protein)
Fingers white, numb, stiff, swellings (Raynaud's disease)	vitamin B6, zinc, cod-liver oil (also vitamin C, magnesium, calcium, alkalize)
Fungal infections (e.g., athlete's foot, ringworm)	B vitamins (tea tree oil pack, intestinal sanitation)
Gangrene	vitamins C, E, B1, magnesium chloride
Greasy dermatitis around eyes, nose	vitamin B6, zinc
Greasy skin eruptions (seborrhoea)	vitamin B2
Horny, goose-pimple-like skin	vitamin A (possibly fat malabsorption)
Horny red skin on pressure areas (e.g., knees, elbows)	zinc (B6)
Hot flushes	vitamin E, boron, maca, pollen (wild yam cream)
Ichthyosis (fish-like scales)	vitamin A, retinoic acid topically
Infant dermatitis with inflamed pustules around body openings	vitamin B6, zinc (allergy testing)
Infections of the skin (boils, cold sores, impetigo, and so on)	vitamins C, A, B6, zinc, magnesium (hot Epsom salt and tea tree oil packs, propolis)
Itching	vitamins B, C, EFA, alkalizers (allergy testing)
Jaundice	vitamins C, E, B6, B12, A, lecithin, magnesium, zinc (blue light therapy, liver cleansing)

(continued)

Symptom	***Deficiency/Treatment/Cause***
Lemon-yellow skin	vitamin B12
Oily skin, whiteheads	vitamin B2
Oversensitivity to sunlight	vitamin B6, zinc, PABA, carotenes, bioflavonoids, antioxidants, alkalizers, purple food
Pale skin	biotin, folic acid, vitamin B6, B12 (anemia)
Prickly-heat rash	vitamin C
Psoriasis	see acne; also vitamin E, lecithin, blue light, sunlight (allergy testing, avoid gluten)
Purplish or blue-black skin areas	vitamin C, vitamin B2, bioflavonoids
Rash	allergy testing, alkalizers, vitamin C, calcium
Red-brown, often symmetrical discoloration of skin exposed to the sun, later ulceration	niacin or nicotinamide (folic acid)
Red-brown or dark-red spots	manganese
Rosacea (redness of part of face)	vitamin B2
Scaly dermatitis	biotin
Scaly eczema around nose, ears, scrotum, vulva	vitamin B2, B6, zinc
Scar tissue	vitamin E, MSM, camphorated oil
Scleroderma (hardening and swelling of skin)	vitamins E, C, A, PABA, magnesium, MSM (zapper, parasite herbs, tea tree oil packs)
Sensitivity to insect bites	vitamin C, calcium, alkalizers, vitamin B1
Shingles	B vitamins, vitamins B12, C, A, E, lysine
Skin cancer	vitamins A, C, E, B6, propolis, zinc, chlorophyll, bioflavonoids, blue light, escharotics)
Skin-folds red, infected	niacin
Stretch marks	vitamins E, B6, zinc
Subcutaneous bleeding (red or purplish spots under the skin, bruises easily)	vitamin C, bioflavonoids and rutin, grape seed extract, glucosamine, cartilage, alkalizer
Swellings (e.g., face)	zinc, magnesium, vitamins B6, B12
Vaginal itching	vitamins B2, E, C (Candida treatment, avoid synthetic underwear)
Warts, moles	vitamins C, A, E, bioflavonoids, antioxidants
Weals on the skin (urticaria)	vitamin B6, zinc, vitamin C (alkalizer, allergy)
White skin patches (vitiligo)	PABA, pantothenic acid, vitamin B6, zinc
Wrinkles or aging skin	vitamins C, E, A, EFA , MSM (too much sun)

Table 6-3: Hair and Nail Deficiency Symptoms

Symptom	***Deficiency/Treatment/Cause***
Coarse, brittle hair	zinc
Dandruff	vitamins B2, B6, zinc, magnesium, biotin
Dry hair	vitamin A, zinc
Greying hair	folic acid, pantothenic acid, biotin, minerals/copper (weak circulation/adrenals, avoid sweet food, food allergies)

(continued)

Symptom	*Deficiency/Treatment/Cause*
Hair loss	zinc, vitamin B6, sulphur/MSM, selenium, biotin
Hangnails	vitamin C, folic acid, proteins
Nails opaque, white spots/bands	zinc, vitamin B6
Oily hair	vitamin B2
Peeling nails	vitamins A, C, calcium
Ridges on nails, longitudinal	vitamin A, protein (anemia, poor circulation, glandular imbalance)
Ridges on nails, transverse	fever, infection, menstrual problems, period of protein -deficiency
Scaling of cuticle or lips	biotin, vitamin B2

Table 6-4: Tongue and Mouth Deficiency Symptoms

Symptom	*Deficiency/Treatment/Cause*
Beefy, enlarged tongue	pantothenic acid
Burning, sore tongue	vitamins B2, B6, B12, niacin
Cracked lips and corners of the mouth (cheilosis)	vitamins B2, B6, folic acid
Distended, purplish-blue veins under the tongue	vitamin B2 (circulation poor, congested)
Edema or tooth-marks on tongue	niacin/nicotinamide
Furrowed tongue	vitamin B1, pantothenic acid
Gums receding, bleeding (gingivitis)	vitamin C, bioflavonoids, calcium, alkalizers
Halitosis (bad breath)	vitamin B6, zinc, magnesium, propolis (cleansing, intestinal sanitation)
Mouth ulcers, canker	folic acid, vitamin B6, zinc (overacidity, allergy)
Papillae prominent or erased	niacin/nicotinamide
Purplish or magenta tongue or lips	vitamin B2
Tongue red at tip or edges; with severe deficiency whole tongue scarlet red, sore	niacin (vitamin B6)
Tongue shiny, smooth, beefy; early sign: strawberry-red tip/sides	vitamin B12, folic acid
Tongue too small or too large	niacin, pantothenic acid
White-coated tongue	intestinal putrefaction
White patches on tongue	vitamin B12 and other B vitamins (allergy)
Yellowish-brown-coated tongue	liver or gallbladder problems

* * * * * * * * * *

Step 46

CANCER

If you have cancer, here are some useful suggestions.

You may gain a better appreciation of natural cancer treatment if I first share with you some little-known facts about orthodox cancer treatment. In cancer research, success (expressed as a five-year survival rate) is established by comparing other forms and combinations of treatment with the results from surgery alone. However, the success rate of surgery has rarely been compared with the survival rates of untreated patients. Therefore, orthodox cancer treatment is basically unscientific because it doesn't deal with all the data. The overall supposed "cure rate" is not higher than can be accounted for by removing precancerous conditions, dormant or calcified tumours, spontaneous remissions, and the placebo effect. In support of my position, I have assembled some key statements and conclusions from medical and scientific publications.

Studies appear to show that early intervention is helpful, because pre-cancerous lesions are included in early removals that frequently would not become cancerous if left untouched. It does not matter how much or how little of a breast is removed; the outcome is always the same.[38]

Researchers say it is complacent to continue subjecting at least 70 percent of women with breast cancer to surgery, a futile mutilating procedure.[39] There is no evidence that early mastectomy affects survival; if patients knew this, they would most likely refuse surgery.[40] The editor of the Lancet pointed out that despite various modifications of breast cancer treatment, death rates remained unchanged. He acknowledged that despite the almost weekly releases of "miracle breakthroughs," the medical profession with its extraordinary capacity for self-delusion (his words, not mine) in all truth has lost its way. At the same time, he rejects the view of those who now believe that salvation will come from increasing chemotherapy after surgery to just below the rate where it kills the patient. Instead, he continues, "would it not be more scientific to ask why our approach has failed?" Not too soon to ask this question, after a century of mutilating women, I would say. The title of this editorial, appropriately, is "Breast cancer: have we lost our way?"[41]

Basically, all types and combinations of conventional breast cancer treatment appear to result in the same low long-term survival rates. The only conclusion that can be drawn from this is that conventional treatment does not improve long-term survival rates. Even worse, Michael Baum, M.D., a leading British breast cancer surgeon, found that breast cancer surgery tends to increase the risk of relapse or death within three years. He also linked surgery to the accelerated spread of cancer, which it does by forming metastases in other parts of the body.[42]

This conclusion confirms an earlier finding by Ernst Krokowski, a German professor of radiology. He demonstrated conclusively that metastasis is commonly triggered by medical intervention, including sometimes even by a biopsy or surgery unrelated to the cancer.[43] Disturbance of a tumor causes a greatly increased number of cancer cells to enter the bloodstream, while most medical intervention, especially chemotherapy, suppresses the immune system. This combination is a recipe for disaster. It is metastases that kill, while primary tumours in general and those in the breast in particular are sometimes relatively harmless.

This follows earlier reports that radical surgery for prostate cancer also tends to spread the disease. Actually, prostate cancer was investigated in the first randomized clinical trials for any type of cancer. After 23 years, there was no difference in the survival rates of those who had surgery and controls who did not; those with surgery suffered more problems such as impotence or incontinence.[44] A recent epidemiological study confirmed the questionable value of conventional therapy by concluding that "medical interventions for cancer have had a negligible or no effect on survival."[45] Even the conservative New England Journal of Medicine had an article with the headline "Cancer Undefeated."[46]

Common ways to make medical statistics look more favourable are as follows. Patients who die during prolonged treatment with chemo-therapy or radiotherapy are not counted in the statistics, because they did not receive the full treatment, while in the control group everyone who dies is counted. Further, success is judged by the percentage of shrinking tumours, regardless of survival time. If survival is measured, then it is only in terms of those dying from the treated disease. It is not normally shown how many of the patients die due to the treatment itself.

The current trend is to pick up precancerous conditions very early and treat them as cancer. While this statistically increases the number of people with cancer, it also artificially prolongs survival times and lowers death rates, thereby making medical treatments appear to be more successful. However, there may also be a genuine component of improved survival, as increasing numbers of cancer patients opt for additional natural therapies.

An investigation of the records of 1.2 million cancer patients revealed that the death rate attributed to non-cancer death shortly after treatment was 200 percent higher than would normally be expected. Two years after diagnosis and treatment, this excess death rate had fallen to 50 percent. The most common cause for the excess death was listed as heart and respiratory failure. This means instead of dying several years later of cancer, these patients died from the effects of the treatment and thereby helped greatly improve the cancer statistics because they did not die strictly of cancer.[47] This misleading reporting of cancer deaths has led to demands for more honest statistics.[48]

After an analysis of several large -mammogram-screening studies found that it leads to more aggressive treatment with no survival benefits,[49] even the editor of the Lancet had to admit that there is no reliable evidence from large randomized trials to support mammography screening programs.[50] The significance of this statement goes far beyond the use of mammograms. It is openly acknowledged by the proponents of conventional medicine that they have no effective way of helping patients with advanced cancer. Until now the catch-cry always was "detect it early, then it can be cured." These mammogram evaluation studies demonstrate that it does not matter when cancer is detected; they suggest to me that the conventional methods and the whole multibillion dollar cancer industry are useless.

Here's more supportive evidence: A 13-year Canadian study of 40,000 women compared physical breast examinations with examinations plus mammograms. The mammogram group had many more lumpectomies and surgeries, and the death rate was 107 deaths in the mammography group and 105 in the physical examination group.[51] Ductal carcinoma in situ (DCIS) is a common non-invasive form of breast cancer. Most cases of DCIS are detected through the use of mammography; in younger women, 92 percent of all cancers detected by mammography are of this type. Nevertheless, on average, 44 percent, and in some states 60 percent, of these are treated by mastectomy. As most of these tumours are harmless, this greatly improves the survival statistics.[52]

In contrast, while conventional diagnosis is invasive and may help to spread the cancer, a kind of electro-dermal screening, called the Biofield test, was developed by a team from eight European hospitals and universities. The Lancet reported it as being 99.1 percent accurate in diagnosing malignancy in breast tumors.[53]

A large meta-analysis of radiotherapy results for lung cancer showed that after two years there were 21 percent more deaths in the group that had radiotherapy in addition to surgery as compared to those who had surgery alone. Researchers contend the rationale is to kill any cancer cells remaining after surgery, but it is a shame that the facts do not agree with this theory.[54]

Chemotherapy for children with leukaemia and Hodgkin’s disease is the proud showpiece of the only apparent success of orthodox cancer therapy. Now a long-term follow-up study shows that such children develop 18 times more secondary malignant tumours later in life and that girls face a 75 times (7500 percent) higher risk of breast cancer by the time they are 40.[55] A main problem appears to be the development of deep or systemic Candida albicans infections shortly after starting chemotherapy.[56] If this is not appropriately treated (see entries on Candida in Step 48), then relapses or future health problems are likely to occur.

Chemotherapy showed a clear dose dependency whereby the incidence of triggered leukaemia doubled between low-dose and -moderate-dose groups and then quadrupled between the moderate-dose and the high-dose groups. A study of ovarian cancer found that the risk of developing leukaemia after treatment with chemotherapy increased 21-fold, or 2100 percent. Other tumours commonly develop after treating malignancies with chemotherapy.[57] In a trial for multiple myeloma, no advantage was found by using chemotherapy as compared to no treatment.[58] Ralph W. Moss, Ph.D., in Questioning Chemotherapy, provides a detailed analysis of this subject; his overall conclusion is that there is no evidence that chemotherapy extends patient life with the majority of cancers.

However, even if chemotherapy extends life by a few months, what about the quality of this life? Tom Nesi, a former director of public affairs at the pharmaceutical giant Bristol-Myers Squibb, wrote in the New York Times about the successful treatment of his wife, which extended her life for three months.[59] Two weeks after the treatment, she scribbled on a notepad: “depressed - no more - please.” I am not surprised by reports that most oncologists would not use these treatments for their own families.

Why are they doing this? By “they” I am referring to what is commonly called the cancer establishment. I believe the answer was given by the eminent medical commentator and former editor of New Scientist, Dr. Donald Gould, in his article “Cancer: A Conspiracy of Silence.” The subtitle summarizes his position: “The commonest cancers are as resistant to treatment today as they were 40 or 50 years ago. Nothing is to be gained by pretending that the battle against cancer is slowly but surely being won."[60]

This truth has been deliberately concealed from the general public. According to Dr. Gould, the reason for this conspiracy of silence is money. The public must continue to see the cancer establishment as a winner to continue providing money. One of the quoted scientists said that, with tens of thousands of radiologists and millions of dollars in equipment, one still gives radiation treatment even if study after study shows that it does more harm than good. Dr. Gould says patients who could be comfortable without medical treatment until their inevitable death are made miserable with medical treatment in a pointless attempt to postpone death for a few unhappy weeks. Of course, that is when most of the money is being made. Dr. Gould says doctors poison their patients with drugs and rays and mutilate them with unnecessary surgery in a desperate attempt to treat the untreatable.

Since the early 1970's when President Nixon declared the War on Cancer, two trillion dollars have been spent on conventional cancer treatment and research, with the result that more Americans are dying of cancer than ever before.[61] While there have been many studies to evaluate the effects of various nutrients on different cancers, none of this two trillion dollars has been available for natural therapists to test holistic cancer therapies, and natural therapists have had to face a century of denigration and persecution in the U.S.

All alternative cancer clinics eventually had to relocate to Mexico. (For a list of such clinics worldwide, see: www.cancure.org) A holistic cancer approach includes superior nutrition, electro-medicine and vibrational or energy medicine, emotional healing, and mind therapy.

One of the few studies that investigate a holistic approach involved the Gerson therapy in an evaluation of five-year survival rates of 153 melanoma patients. Here, 100 percent of Gerson therapy patients with Stage I and II cancers survived, but only 79 percent did so with conventional therapy. With Stage III (regional metastases), the figures, respectively, were 70 percent and 41 percent; with Stage IVa (distant metastases), 39 percent with Gerson and 6 percent with conventional therapy.[62]

Now there is actually proof that cancer surgery is the main cause of metastasis or secondary organ tumours. About 90% of patients die from metastases while primary tumours actively suppress metastasis. Stress as from fear or most medical procedures tends to promote metastasis and this is the reason that untreated patients generally live longer than medically treated ones. Regard a tumour as a mistreated part of your body, surround it with loving care and it will thank you by preventing metastasis and gradually melt away. For more information and references see www.health-science-spirit.com/cancersurgery.htm and www.health-science-spirit.com/cancersolution.htm.

People claim to have been cured of cancer by a variety of methods and remedies. Belief or faith in the chosen method may often have been more important than the method itself. However, fundamental factors that should be addressed for a more uniform successful outcome are:

1. Alkalize the body

2. Controlling the cancer microbe
3. Normalizing metabolism
4. Cleansing the body
5. Improving the mind and emotions

When these are corrected, any type of tumor can disappear on its own or become dormant. Natural cancer therapy has a fundamentally different view of what cancer is from that of orthodox medicine. The latter believes that the tumor is the cancer and only attacks the tumours. Natural therapy approaches address cancer as a condition of the whole body and person and treat the whole person.

Alkalizing the body: If the body is sufficiently alkaline tumours do not grow. Use alkalizers, including sodium bicarbonate and potassium citrate, as shown in Step 6 to keep your urine above pH 7 most of the time.

Controlling the Cancer Microbe: The cancer microbe is the key to successful natural cancer therapy. Various researchers such as Guenther Enderlein, Wilhelm Reich, Royal Ray Rife, Virginia Livingston-Wheeler, Alan Cant-well, and Gaston Naessens have found the same phenomenon in all cancers: a microbe that changes its shape and is thus pleomorphic, “of many shapes.” It is normally present as tiny protein units, but in various disease conditions, including all autoimmune diseases, it grows into a coccus and higher bacterial forms and then mycoplasmas and fungus-like forms.

This microbe does most of the cancer damage by devastating the immune system, releasing toxins that poison the energy-producing cellular enzymes; it also produces growth hormones that result in unchecked cell divisions. If the cancer microbe in its advanced stages is cultured and injected into animals, it causes tumor growth and the wasting and sickening odour that is characteristic of terminal cancer.

All this makes it understandable that antimicrobial and especially fungicidal treatments have often been credited with having cured or reversed cancer. Kerosene, for instance, taken internally, is said to be widely used in Russia and other Eastern countries as a cancer cure, and it was even used in America a hundred years ago. A hospital in Brazil reported cancer cures with the fungicidal herb pau d’arco. The electronic zapper and the Rife machine have been used to destroy the cancer microbe electronically. Presently MMS and Lugol's iodine solution are widely used.

Normalizing Metabolism: The main reason that body cells become cancerous is the destruction of the enzymes in their energy-producing respiratory chain. These enzymes can be poisoned either by accumulated toxic chemicals, especially pesticides, or by endotoxins produced by the cancer microbe. The German cancer researcher P. G. Seeger, M.D., D.Sc.,[63] demonstrated that the virulence of cancer cells is proportional to their inability to use oxygen for energy production and that cancer cells can be transformed back into normal cells by restoring cellular respiration. This can be achieved with certain enzyme supplements and especially with the purple pigment of red beets and dark grapes.

Furthermore, Dr J. Budwig, a German fat researcher, found that flaxseed oil and sulphur compounds, especially food high in cysteine (e.g. quark, a fermented cottage cheese) were most effective in restoring a normal energy metabolism. The importance of a good diet can also be seen from a study of 535 breast cancer patients. After ten years, those with the highest insulin levels were eight times more likely to die than were patients with the lowest levels. Insulin levels are strongly elevated when ingesting common household sugar and to a lesser degree when combining glucose, as for instance from eating starches with fructose from fruit.

Cleansing the Body: Eliminating the cancer microbe and restoring cellular energy production is not always sufficient in an advanced condition because of the large amount of toxic waste produced by disintegrating tumours. Think of a dead tumor as a dead rat in your body. Patients have died after "successful" cancer cures because of congestion in their organs of elimination, especially the liver and kidneys. Cleansing is needed to remove metabolic wastes, pesticides and endotoxins that helped to cause the cancer in the first place.

Improving the Mind and Emotions: Our mind and emotions are an additional factor that makes it much easier for cancer researchers to cure mice rather than people. The orthodox cancer approach generally causes patients to be fearful, but this paralyses the immune system and makes all other body systems inefficient. Cancer treatment is likely to fail and cancer tends to return after successful initial therapy if we live in fear, especially if this fear is subconscious. Enthusiasm, joy, positive expectations, and supportive relationships have the opposite effect, and clinical trials of mind and emotional therapies have shown superior results compared to orthodox cancer therapies.

Putting It All Together: There are hundreds of supplements and remedies that might be beneficial to overcome cancer. The more effort you put into the basics, the less additional remedies you need. For some suggestions of additional supplements and remedies, see the entry on Cancer in Step 44. All the basic information that you need to reverse your cancer is in this book, for more detailed instructions see Overcoming Cancer at http://www.the-heal-yourself-series.com.

- From Step 34 use the low protein High-Quality Diet with its strong emphasis on purple foods, raw foods, and fresh vegetable and grass juices.
- Keep your urine above pH 7 as shown in Step 6.
- Do the Intestinal Sanitation and Antimicrobial Therapy as shown in Step 3.
- Do periodic fasting and cleansing as explained in Step 4, including the Basic Cleanse and if appropriate a Long Cleanse.
- Use a Beck-type electronic zapper and Reflexology as explained in Part 2.
- Release any resentment or other negative emotions and replace them with positive emotions (see Part 8).
- Practice the Love Cure (see Part 8).
- Make appropriate changes to your belief systems. Use the mind tools in Part 9, especially guided imagery..

* * * * * * * * * * * *

Step 47

EMOTIONAL – MENTAL SHOCKS

The secret to preventing illness is to undo shocks that unbalance the body.

The following information is controversial and not verified by independent scientific research, but I believe it is important for you to know. You can then make up your own mind about it.

Dr. Ryke Geerd Hamer, M.D. (born 1935), a German medical researcher, has developed an emotion-based cancer therapy. His claimed success rate of about 95 percent is by far the highest of any cancer therapist. During one of his many trials, the public prosecutor (in Wiener Neustadt, Austria) had to admit that after four to five years, 6,000 out of 6,500 patients with mostly advanced cancer who had been treated by this approach were still alive.

Dr. Hamer started his cancer research when he developed testicular cancer after the death of his son, who had been shot by the Prince of Savoy, son of the last King of Italy. He wondered if the shock caused by his son's death was the cause of his cancer. Subse-quently, he investigated and documented more than 15,000 cases of cancer and always found the following characteristics to be present, which he termed "The Iron Rules of Cancer":

1. Every cancer and related disease starts as a DHS or Dirk Hamer Syndrome (named after his son). This is a serious, acute-dramatic, and isolating conflict-shock experience for which we are emotionally unprepared. It manifests simultaneously on three levels: psyche, brain, and organ.
2. The theme of the psychic conflict determines the location of the focus, or Hamer Herd, in the brain and the location of the cancer in the organ.
3. The course of the psychic conflict correlates with the development of the Hamer Herd in the brain, and the course of the cancer in the organ.

At the moment of the conflict-shock, a short circuit occurs in a predetermined place in the brain. This can be demonstrated with a computed tomography (CT) scan and looks like concentric rings on a shooting target. Later on, if the conflict becomes resolved, the CT image changes to reflect the development of an oedema and finally of scar tissue.

The following incident shows how specific and precisely located these brain lesions are. After a professional lecture, a doctor handed Dr. Hamer the CT brain scan of a patient and asked him to explain it. From this, Dr. Hamer correctly diagnosed the patient as having a bleeding bladder carcinoma in the healing phase, an old prostate carcinoma, diabetes, an old lung carcinoma, and sensory paralysis in a specific area, in addition to the corresponding emotional conflicts.

Dr. Hamer demonstrated that at the same time as the concentric brain lesion appears, the target organ CT scan may also show such a concentric lesion. According to Dr. Hamer, this happens instantly when the psychic shock hits the subconscious level, and this same second is the start of cancer. Also, other diseases can be caused by the same mechanism. How severe a disease becomes depends on the duration and severity of the conflict-shock, while its nature and location are determined by the emotional content of the shock.

Dr. Hamer believes that the correlation between key emotional shock events, the target brain areas, and the related organs has developed as an adaptation from similar programs in the animal world. When we unexpectedly experience emotional distress, an emergency repair program is set in motion, a biological conflict program with the aim of returning the individual to normal.

Dr. Hamer gives the following example. A mother sees her child in a bad accident. As small children recover faster when they receive extra milk, the biological conflict program tries to stimulate milk production by increasing the number of breast cells. If the mother is right-handed, that will instantly cause the appearance of a Hamer Herd in a specific part of her right brain, which in turn relates to the left breast. When the child is well again, conflict resolution begins and extra milk is no longer needed. The mother gets a benign form of tuberculosis in the affected breast that breaks up the excess breast cells. However, if the mycobacteria required for this function are lacking, then the area may just calcify and remain as a dormant tumor.

The same process also applies to animals. A sheep that loses its lamb to a wolf is prone to develop teat cancer; the side depends on whether it is right- or left-footed. Commonly, the sheep resolves this conflict by bearing another lamb.

The next example illustrates the original survival value of this mechanism. Let's assume a lion chases an antelope. The antelope must immediately mobilize all its resources to survive. The sympathetic nervous system takes over and, in addition, a specific brain centre that stimulates lung activity becomes active. After the successful escape, the animal rests, and the parasympathetic nervous system becomes dominant to normalize brain and body functions.

If a human gets a cancer diagnosis, even if the diagnosis is wrong, the same biological program is set in motion by the same fear of death that helped the animal to escape. The stress level jumps and activates the brain-lung connection, but now there is nowhere to run. Until the conflict is resolved, which may take years, there will be constant stress and brain-induced stimulation to increase lung capacity by the incessant division of cells.

This process can only be stopped by switching off the trigger in the brain through defusing the original conflict shock. This happens when the patient subsequently has surgery or natural therapy, which he or she fully believes will lead to a cure. However, the same procedure in a patient who has doubts about its effectiveness will leave the conflict unresolved and the disease to progress. Thanks to Dr. Hamer's work, it is now possible to verify this hypothesis with a CT brain scan. The selection of the conflict focus occurs by subconscious association. For instance, biological conflicts involving water or other fluids, such as milk or oil, lead to kidney cancer; fear of death to lung cancer; and psychologically swallowing a bigger chunk than we can digest to stomach or intestinal cancer.

In the animal world, it is literally a big chunk of food, but for us it may be a financial over-commitment or any other obligation that we are unable to fulfil. The target focus is not determined by the event itself, but rather by the psychological significance that it has for us at the time of the event. Dr. Hamer illustrates this with another example, a woman who finds her husband in bed with another female. As a sexual frustration conflict, it can cause uterine cancer. If she instead experiences it as a partner conflict, then in a right-handed woman it leads to cancer of the right breast. If the conflict feeling is fear and revulsion, then it causes hypoglycaemia, and with lack of self-worth, cancer may develop in the pubic bone.

If the lack of self-worth had been due to a failure in sportsmanship, then the problem would have arisen in an arm or leg instead, or possibly in the fingers or shoulder required for the sporting activity. Other typical situations that can lead to biological conflicts are loss situations, of a loved one, a job, a valued possession, or a territory.

Dr. Hamer believes that most metastases or secondary tumours are caused by the cancer-fear or death-fear resulting from the patient being given the cancer diagnosis or a negative prognosis. In this case, the resulting conflict shock may not be fear of death, but anger, resentment, or a separation conflict (from partner or children), and then tumours would appear in different places. A diagnosis of colon cancer commonly leads to liver metastases because of a subconscious fear of starvation.

In regard to animal experiments, Dr. Hamer regards increased cancer rates due to exposure of carcinogens as being emotionally based. This happens when the skin is irritated with a toxic substance. He also cites early research with tobacco smoke that showed that hamsters exposed to it lived longer than non-exposed hamsters. He believes that hamsters, as burrowing animals, are not afraid of smoke and therefore are protected from its harmful effects. Mice, on the other hand, living in houses, are very fearful of smoke, which causes them to develop lung cancer when exposed to tobacco smoke.

Generally, hopelessness, despair, and meaninglessness create chronic stress and hinder the healing of cancer and other diseases, but they are not the cause. According to Dr. Hamer, the real cause of cancer and other diseases is an unexpected traumatic shock for which we are emotionally unprepared.

The emotional connections with some other diseases are as follows, according to Dr. Hamer:

Angina Pectoris: Caused by fight for territory or its contents in the active phase, followed by heart infarction in the healing phase.

Diabetes and Hypoglycaemia: A right-handed female develops hypoglycaemia from anxiety and revulsion; if left-handed, she develops insulin-dependent diabetes. A right-handed male develops insulin-dependent diabetes from a conflict of resisting or struggling against something; if left-handed, he develops hypoglycaemia.

Facial Paralysis: Fear of losing face, having been made a laughingstock.

Haemorrhoids: A right-handed woman with an identity conflict and a left-handed man with territorial anger in the healing phase may get haemorrhoids.

Multiple Sclerosis and Paralysis: Inability to escape or continue on, or to hold on to, or not knowing what to do.

Psoriasis: Separation conflict concerning mother, father, family, home, friends, or pets.

Relationships between Conflicted Emotions and Target Organs

Adrenal cortex: Wrong direction, gone astray
Bladder: Ugly conflict, dirty tricks
Bone: Lack of self-worth, inferiority feeling
Breast, left (right-handed): Conflict concerning child, home, or mother
Breast, right (right-handed): Conflict with partner or others
Breast milk duct: Separation conflict
Breast milk gland: Involving care or disharmony
Bronchials: Territorial conflict
Cervix: Severe frustration
Colon: Ugly indigestible conflict
Gallbladder: Rivalry conflict
Heart: Perpetual conflict
Intestines: Indigestible chunk of anger
Kidneys: Not wanting to live, water or fluid conflict
Larynx: Conflict of fear and fright
Liver: Fear of starvation
Lung: Fear of dying or suffocation, including fear for someone else
Lymph glands: Loss of self-worth associated with the location
Melanoma: feeling dirty, soiled, defiled
Middle ear: Not being able to get some vital information
Mouth: Cannot chew or hold it
Oesophagus: Cannot have it or swallow it
Pancreas: Anxiety-anger conflict with family members, inheritance
Prostate: Ugly conflict with sexual connections or connotations
Rectum: Fear of being useless
Skin: Loss of integrity
Spleen: Shock of being physically or emotionally wounded
Stomach: Indigestible anger, swallowed too much
Testes and Ovaries: Loss conflict
Thyroid: Feeling powerless
Uterus: Sexual conflict

Psychoses: One or more active Hamer Herds in each of the two parts of the brain.

Rheumatism: A brutal and painful separation conflict.

Vitiligo, Leukoderma: Ugly or brutal separation conflict.

In regard to AIDS, Dr. Hamer observes that no one dies of AIDS without having previously been told that they are HIV-positive or believe that they are. The implication is that it is mainly the negative perception associated with AIDS that causes its devastating effect.

Left-handed individuals develop problems on opposite sides to genuine right-handed individuals. Observing which hand is on top when clapping reveals handedness; right-handed individuals usually have the right hand on top. With right-handedness, all muscles on the right side of the body are involved with partners and other individuals, and all left-sided muscles, with children or one’s mother. These conditions may change with hormonal changes, the birth control pill, or menopause.

The Healing Phase: The start of Dirk Hamer Syndrome or conflict-shock experience is different from other conflicts that we experience in our daily lives, says Dr. Hamer. It causes a continuous stress, resulting in a tendency to develop cold hands and feet, lack of appetite and weight loss, sleeplessness, and a dwelling all the time on the conflict content. If the conflict is not resolved soon, it leads to specific disease symptoms.

When the conflict resolves, the patient is no longer occupied with the conflict content, the appetite returns, hands are warm again, and also normal sleep returns, but there may be weakness, fatigue, and a need to rest. These effects show that the parasympathetic nervous system is now in control and the healing phase has started.

During the first part of the healing phase there is water retention and inflammations, and the tumor stops growing. This eventually leads to a healing crisis, which Dr. Hamer calls an epileptic or epileptoid crisis, because it is caused by an oedema in the Hamer Herd brain lesion. It shows unique symptoms for each illness.

After this, the body starts to expel the accumulated water, the patient gradually regains strength, and body functions become normal. Now the connective tissue in the brain, the neuroglia or glia, starts repairing the Hamer Herd. This may be interpreted by conventional radiologists as a fast-growing brain tumor and treated accordingly. Dr. Hamer estimates that 99 percent of brain events, such as strokes, bleeding into the brain, cysts, and tumours are due to healing events of Hamer Herds and are temporary and self-limiting, unless there is inappropriate medical intervention. The most important support in these situations is the reduction of any brain oedema.

During the healing crisis, the patient may for a short time re-experience the original psychological conflict with cold hands and cold sweat. This serves to suppress and eliminate the oedema in the brain lesion, which then allows other body conditions to normalize. The main danger point is just before the end of the healing crisis when it will become apparent if the body is strong enough to eliminate the disease. In difficult cases, with long-standing or strong conflict, massive brain oedema can develop, for which Dr. Hamer uses cortisone injections. In natural therapy, we use diuretics, such as urea, and anti-inflammatory therapies instead.

The epilepsy crisis may be experienced as a heart attack, lung embolism, hepatitis, or lung infection. Generally, during the healing phase, the patient will have more discomfort than when the tumor was actively growing. In the first part of the healing phase, most problems are due to water retention, inflammations, and swelling of tissues that cause a lot of pain. Nevertheless, Dr. Hamer regards oedema, whether in the brain or body, as a positive sign of healing.

Tissue Repair: After the healing crisis, carcinomas are removed by fungi and mycobacteria, while the hepatitis virus may in addition help to regenerate the liver. Bacteria, viruses, and fungi that help to break down the tumours and repair damaged tissues also cause inflammation, pain, and fever.

If you find it odd that Dr. Hamer regards microbes as essential friends and helpers in the healing of cancer, he has developed an even more surprising hypothesis. In his system of tumours and cancer-equivalent diseases, he distinguishes between two opposite processes during the active conflict phase. Depending on the location of the Hamer Herd in the brain, there may be either cell proliferation or cell destruction.

The first group has cell proliferation and tumor growth during the conflict phase and then removes excess cells with the help of microbes during the healing phase. The other group shows cell destruction during the conflict phase, resulting in ulcers, necroses (dead tissues), and tissue holes, affecting, for instance, bones (osteoporosis), kidneys, spleen, or ovaries.

During the healing phase, this second group tries to fill in the created holes through cell proliferation. Bacteria repair tissue necroses and osteolyses (dissolved bone) by first forming abscesses, which are then filled in with scar tissue and later with granulating tissue to form osteosarcoma, lymphoma, fibroma, and healing cysts.

Leukaemia occurs during the healing phase after bone marrow damage from radiation, chemicals, or bone cancer.

According to Dr. Hamer, these conditions are generally self-limiting and only get out of control when additional conflict shocks occur, or the body is too old or weak, or through the methods of conventional medicine. In contrast, natural healing methods aim to support body and mind during this trying time. Most healings proceed without major problems, but about ten percent need the full support of an experienced therapist, especially at the time of the healing crisis.

Healing the Psychic Conflict: The main task in every case of cancer is to find the original emotional shock experience and make sure that it has been healed or is being healed. In many cases it will have corrected itself, and the patient suffers from an effect of the healing phase.

For instance, someone may have lost a farm or business, but has now started another satisfying venture or hobby. As an after-effect, there may now be a tumor that gradually becomes dormant or eventually disintegrates. Dr. Hamer regards about 40 percent of tumours discovered during routine medical investigations as old and harmless, dormant and calcified. However, complications can still arise if medical intervention now causes a conflict shock, such as one involving fear of death; this induces the same or another tumor to grow.

Nevertheless, in other cases the original conflict may still be active, or there may be a second active conflict. As we do not know, we have to probe in every case to find the original and any other conflicts. We need to think back, especially one to two years before our problem started, and analyse our emotional history during this time as well as before and after. Meditation or regression therapy may be useful to discover conflict-shocks.

If we still have a strong emotional response when we discover the content of the conflict, then we can be sure that it is still unresolved. If at all possible, it is best to solve it in a natural way. For instance, if it was caused by losing a partner, find someone new; if you lost a child, become pregnant again, or adopt a child, or buy a pet; with a lost job, find a new one or a hobby. However, it is not the fact of having a new partner, child, hobby, or pet that solves the conflict, but rather that it allows us to dissolve the inner conflict and emotionally get on with our life if we so choose.

If a natural solution is not possible, then use guided imagery, either on your own or with the help of a partner or a suitable therapist. In a relaxed and meditative state recreate and re-experience the conflict as intensely as you can, but then substitute a desirable or acceptable outcome. Create and experience this new outcome as vividly and in as much detail as possible; see it, feel it, and possibly even hear and smell it. The original experience may also have imprinted you with the memories of unrelated details of your ordeal (tracks) to which you may now react with allergies. You can overcome these in the same way with guided imagery

If neither of these methods is possible because you feel that you have to continue your present duties or ordeal for whatever reason, then only increased spiritual understanding and acceptance may be able to help. In either case, be aware of your vulnerabilities and avoid any further conflict-shocks, but if one does happen, get it out of your system is soon as possible.

I firmly believe that all active conflicts will be terminated and the healing phase begin when we are able to strongly feel love and forgiveness within ourselves and then radiate it to all others, but especially to anyone who we feel might have wronged us. We can further ease the healing phase by expecting it to be short and mild and lead to full recovery.

According to Dr. Hamer, animals in the wild get cancer from the same shock programs as we do. However, 80 to 90 percent survive and do not notice it much, because the healing phase can take its natural course. Those that die are mainly old animals that cannot resolve a conflict, such as regaining their territory from a rival or replacing a lost cub.

It is different in our society, as the natural healing process is routinely interfered with. It starts with getting tranquilizers or antidepressants during the active conflict phase, which prevent us from fighting back and regaining our territory. This may then lead to a cancer diagnosis that causes an additional active conflict and ends with morphine, which totally disables our healing responses.

While Dr. Hamer does not believe that health foods, remedies, cleansing, or healthy living in general can cure cancer, he says these are important to surviving the ordeals of the healing phase. Actually, Dr. Hamer regards all diseases as consisting of two phases: initially an active conflict, followed (if possible) by a healing phase that reverses the conflict program. He does not call them “diseases” but “special biological programs.” He says he has worked with more than 31,000 patients and found his theories confirmed in every single case without exception. Dr. Hamer claims that overall this approach has a 95-percent success rate with cancer.

Persecution for Therapeutic Success: Siemens, the manufacturer of CT equipment, has independently verified the existence of the Hamer Herds in the brain, which radiologists had claimed to be artefacts. Furthermore, their diagnostic significance was confirmed in 1998 at the University of Trnava in Slovakia. Nevertheless, Dr Hamer faced exceptional persecution. This included withdrawing his right to practice in Germany for not retracting his theories and swearing allegiance to the principles of orthodox medicine; assassination attempts; efforts to put him in a mental asylum; and numerous court appearances and even putting him in jail.

While many individual doctors, including professors of medicine, are alleged to have verified the principles of Dr. Hamer’s “New Medicine,” so far no German university has agreed to test them. Doctors and natural therapists in Europe who practice according to the principles of the New Medicine now face persecution. Only Spanish courts took the enlightened position that it is not their role to decide between conflicting medical theories and therapies. Dr. Hamer now lives and practices in Spain.

In 2001, a prominent neurologist, Dr. Therese von Schwarzenberg openly defended Dr. Hamer by publishing a book about the New Medicine, in which she demanded that his theories be officially tested. She declared that Dr. Hamer is right and deserves a Nobel Prize. Nevertheless, the response of prominent oncologists is still that it is absurd to assume emotions could be important in the cause and cure of cancer and, therefore, Dr. Hamer’s claims must not be tested. (For more information on Dr. Hamer and the New Medicine, see: www.newmedicine.ca.64)

Reflections on the New Medicine: My own understanding and experience leave no doubt about the primary importance of our emotions and beliefs in the cause and cure of our diseases. The writings of Dr. Hamer give the impression of a serious and meticulous researcher whose findings apparently have been confirmed by various scientific institutions and medical practitioners, including professors of medicine.

Therefore, I have no doubt about the validity of his basic premise that diseases, and in particular cancers, are associated with visible lesions in specific locations of the brain and also with specific psychological experiences. His postulated two-part disease process, in which the healing phase is a reversal of the original disease condition, is a basic concept of natural therapies.

However, Dr. Hamer appears reluctant to extend or modify his theory and take other factors into account. Basically, Dr. Hamer postulates a one-way communication from psyche to organ via the brain. I suggest instead a two-way communication from psyche to organ, as well as from organ back to brain and psyche. Assuming that an organ is chronically stressed, be it from pesticide accumulations, inappropriate growth factors, or endotoxins due to microbes, then this causes constant low-level tissue irritation of the organ and its related brain area. There are now two possibilities:

- The sensitized brain area causes us to react more strongly to any emotional conflict; therefore it is easier to develop a Hamer Herd and the related disease.
- An existing subacute conflict, combined with chronic tissue stress, may cause the gradual development of a Hamer Herd or, alternatively, its sudden appearance when a certain threshold is reached.

In this way, any environmental factors can be taken into account for creating Hamer Herds in combination with psychological conflicts. Conversely, appropriate therapy directed to a diseased organ may give a message to the associated brain area to heal the Hamer Herd and lead to a resolution of the emotional conflict.

A confirmation of the suggested interconnectedness between emotions, nutrients, brain, and organs can be found in kinesiology. Research with muscle testing has revealed a relationship between different nutrients and emotions with specific muscles and organs, transmitted through the system of acupuncture meridians. A direct involvement of the brain has been shown by placing a labelled nutrient in the back of the mouth: Within seconds, the isotope was also detected in the brain. Recently another piece of evidence in support of the New Medicine has emerged: researchers found a tumor growth promoter (GDNF) that is produced by the glial cells in the brain.

The theories of Dr. Hamer have not yet been tested by mainstream science, but neither have key components of conventional and natural medicine. If you find Dr. Hamer’s theories reasonable and do not want to wait another 10 or 20 years for them to be accepted, then I suggest that you incorporate them into your holistic healing program.

What to Do: Be aware of your vulnerabilities and avoid any further conflict-shocks, but if one does happen, get it out of your system as soon as possible. For example, if you just got a big shock, minimize or undo its potential to create an illness by using any or all of the following methods:

- Take appropriate remedies, such as Rescue Remedy or homeopathic arnica.
- Develop an inner reconciliation through prayer and meditation.
- Apply the Emotional Freedom Technique (see Step 63).
- Use guided imagery to recreate the shock experience with an acceptable outcome.
- Try to find a natural solution for your shock experience; for example, after nearly drowning, swim again as soon as possible

* * * * * * * * * * *

Step 48

CANDIDA INFECTIONS

The secret to preventing illness is to undo shocks that unbalance the body.

Lack of energy, digestive disturbances, arthritic joint pains, skin disease, menstrual problems, emotional instability, and depression are all symptoms of what I call the antibiotic syndrome. This is a condition that has greatly increased in incidence in recent years and, on further examination, more symptoms of this syndrome may be discovered. In this condition, most of the gastrointestinal tract is tender when pressed on the body's surface, especially the small intestine, liver, and gallbladder. There may have been a gallbladder operation that failed to improve the condition, sometimes worsening the symptoms. As part of this syndrome, there could be a history of thrush or oral, anal, or vaginal itching. When these are present, the diagnosis of Candida overgrowth is obvious, but it may also be present in the absence of these manifestations, and that can be confusing.

The fungus Candida albicans thrives during antibiotic treatment and I regard it as reckless negligence to prescribe antibiotics without simultaneous fungicides and replacement therapy with lactobacilli afterwards. I believe that this practice has greatly added to the numbers in our chronically sick population. However, the antibiotic syndrome is not just due to Candida. I regard it more generally as a dysbiosis in which the wrong kind of microbes inhabit the intestinal tract; it's not just Candida and other fungi, but many types of pathogenic bacteria including E. coli bacteria, which are normal in the colon but become disease-forming when they move into the small intestine.

If the problem has existed for years, there is usually a lack of gastric acid, which then allows the stomach to be colonized by microbes, causing inflammation with pain and, later, ulcers. The toxins released by the microbial overpopulation cause chronic inflammation of the liver, gallbladder, pancreas, and intestines. This can have other health repercussions. I regard it as likely that a chronic inflammation of the pancreas is a major contributing factor in the development of insulin-dependent diabetes.

Bacterial Attack: Specific types of pathogenic bacteria appear to cause or contribute to specific autoimmune diseases. One variety of coliform bacteria, for instance, produces a molecule very similar to insulin. When the immune system becomes activated against this molecule, it may then also attack related features at the beta cells of the pancreas. Another type of bacteria, Yersinia enterocolitica, induces an immune response that attacks the thyroid gland and leads to Grave's disease with a serious overproduction of thyroid hormones.

Ulcerative colitis is linked to overgrowth of pathogenic microbes, as are Crohn's disease and ankylosing spondylitis. In ankylosing spondylitis, the vertebrae of the spine fuse together, causing stiffness and pain. Other joints may in time become affected. Klebsiella, another type of pathogenic bacteria, produces a molecule that is similar to a tissue type found in people with this disease. When Klebsiella numbers in the gut decrease, related antibodies in the blood decrease and the condition improves.

Rheumatoid arthritis is linked to a bacteria called Proteus. Proteus is also a common cause of urinary tract infections. Women suffer urinary tract infections as well as rheumatoid arthritis twice as often as men, while men usually have higher levels of Klebsiella and three times more ankylosing spondylitis than women have. In addition, microbial overgrowth damages the intestinal wall so that only partly digested food particles can pass into the bloodstream, causing allergies. In this way, it appears that all autoimmune diseases can be linked to dysfunctional intestines and food allergies.

Other autoimmune diseases that have so far been linked to intestinal dysbiosis are psoriasis, lupus erythematosus, and pancreatitis. When remedies are given that bind bacterial endotoxins, these conditions usually improve. A further consequence of dysbiosis is susceptibility to food poisoning, as with Salmonella bacteria; in contrast, a healthy intestinal flora prevents these from multiplying and causing trouble.

Staphylococcus aureus causes serious infections in hospital patients. It has been found that not only S. aureus but other infections are greatly potentized when they occur with a Candida overgrowth. As Candida overgrowth is a natural outcome of the standard hospital treatment, it is easy to see why staph is so deadly in hospitals.

A similar picture emerges with AIDS. People do not die from the AIDS virus, but from bacterial infections potentized by Candida and related mycoplasmas.

I regard the antibiotic-induced dysbiosis in babies and infants as the main cause of their frequent infections called glue ear, and as greatly contributing to crib death. While it used to be uncommon for children to have more than one or two infections a year, now more than six is the norm.

In summary, in the 1940’s, Candida overgrowth was found in only three percent of bodies autopsied, but now the figure is nearer 30 percent. There are, of course, other factors that can cause dysbiosis - the contraceptive pill, steroids and other drugs, radiation treatment, chemotherapy - but the main culprit is without doubt antibiotics.

Debilitations Caused by Candida: In an Alabama hospital in 1953, Orian Truss, M.D., discovered the devastating effects of antibiotics. During a hospital round, Dr. Truss was intrigued by a gaunt, old-looking man who was obviously dying. However, he was only in his 40’s and though he had been in the hospital for four months no specialist had been able to make a diagnosis. Out of curiosity, Truss asked the patient when he was last completely well.

The man answered that he had been well until six months ago when he had cut his finger. He had received antibiotics for that. Shortly afterwards, he developed diarrhea and his health deteriorated. Dr. Truss had seen before how antibiotics cause diarrhea. It was known that Candida was opportunistic and thrived in debilitated patients, but now Dr. Truss wondered if it might not be the other way around, that Candida actually caused the debilitated condition. He had read that potassium iodide solution could be used to treat Candida infestation of the blood so he put the patient on six to eight drops of Lugol’s solution four times a day and soon the patient was again completely well.

Not long afterwards, Dr. Truss had a female patient with a stuffy nose, a throbbing headache, vaginitis, and severe depression. To his amazement, all her problems immediately cleared with Candida treatment. Some time later, he saw a female patient who had been schizophrenic for six years with hundreds of electroshock treatments and massive drug dosages. He started treating the woman for sinus allergies with a Candida vaccine. Soon she had recovered mentally and physically and remained well.

From then on, Dr. Truss treated his patients against Candida at the slightest indication of its presence. Many of his patients made remarkable recoveries from most unusual conditions, including menstrual problems, hyperactivity, learning disabilities, autism, multiple sclerosis, and autoimmune diseases such as Crohn’s disease and lupus erythematosus.

Every experienced naturopath can relate similar success stories. Ironically, antibiotics are usually not necessary in the first place. In the few percent of cases in which they are necessary, their serious after-effects could easily be avoided by using fungicides and lactobacilli. Many people doubt the effectiveness of natural therapies against apparently serious infections, but my clinical experience leads me to believe that, frequently, natural therapies are more effective, and without causing the repeated and chronic infections seen after antibiotics. I have seen patients whose long-term antibiotic treatment was unsuccessful recover within days or weeks with natural therapies.

Overcoming the Antibiotic Syndrome: The main treatment for the antibiotic syndrome is sanitation of the gastrointestinal tract combined with a low-allergy diet. In this way, most acute problems can be overcome within weeks, while degenerated joints or organs can start a slower road to recovery.

One unpleasant side effect of most methods used to reduce the pathogenic microbial overgrowth of the intestines is the Herxheimer effect, a sudden but transient worsening of symptoms due to toxins released by the dead or dying microbes. This is sometimes used as a diagnostic tool, especially for Candida: If you get the reaction, it means you had the infestation.

I recommend intestinal flushes (orally taken) combined with antimicrobial remedies to avoid or minimize unpleasant side effects. An oral isotonic flush consists of nine grams or one heaping teaspoon of salt in one quart of water, and is suitable for sensitive individuals and generally for those with low to normal blood pressure. A solution is said to be isotonic if it contains a similar amount of salt as the blood. An Epsom salt flush can contain one tablespoon of Epsom salt, more or less according to need, in a large glass of water and is useful in cases of constipation and raised blood pressure.

A crushed clove of garlic is added to the flush to reduce the microbial overgrowth. If garlic cannot be used, then up to two teaspoons of three-percent hydrogen peroxide (food-grade) or six drops of tea tree oil can be added instead. You can take another glass of water after the Epsom salt to wash down the bitter aftertaste, but not after the isotonic flush, as it would no longer be isotonic.

The flush will carry most of the dead microbes and their toxins out of the body and minimize any unpleasant reactions. However, if the flush should not be sufficient to produce a strong bowel movement within two to three hours, then another flush, preferably with Epsom salt, should immediately follow or unpleasant reactions may occur.

As a precaution after antibiotic therapy, for general body cleansing, or if there are only mild symptoms of microbial overgrowth, it will be sufficient to continue this for about three days. In severe and long-standing cases, the flush can be used for several weeks. If Candida overgrowth is suspected, any sexual partner should have at least a short flush period. If it is inconvenient to continue with a flush, as for instance with work commitments, then it may work to have a flush just once or twice and then continue for a long time with a high intake of acidophilus and bifido cultures.

Acidophilus and Bifido Cultures: This leads us to the next step: replacing the decimated pathogenic microbes with desirable lactobacilli or "friendly" bacteria. This must be done in a massive dose 30 to 60 minutes after each flush, otherwise the unwanted microbes may grow back to full strength after your next meal. Acidophilus culture is used to repopulate the freed-up spaces in the walls of the small intestines, while bifidobacteria are the protective inhabitants of the large intestines.

These bacterial cultures are available from health foods stores as high-potency capsules or powders in brown glass bottles that should be refrigerated before and after sale. You can buy a mixed culture and take several capsules or two teaspoons in water or vegetable juice or one teaspoon each of the separate cultures. If you suspect a cow's milk allergy, which is frequent in this condition, then use milk-free cultures.

After several months on the shelf of a store, cultures may have very few live bacteria present. Therefore, it is best to take powdered cultures in addition to fresh cultures found in good quality yogurt. You can test the quality of any powdered culture by using it as a starter to make your own yogurt (see Step 37). If it is of high quality, the yogurt will develop quickly. Alternatives to cow's milk yogurt do not set creamy because of their lower protein content, but that does not matter, as the beneficial bacteria are in the liquid whey.

Eat a cup of this yogurt in addition to the powdered culture after each flush, using more before or with other meals. Continue taking either the cultures or yogurt for several months after you have stopped taking a flush. Women should apply yogurt or acidophilus powder to the vagina for a few days or weeks. It is preferable initially for a few days to douche beforehand with a few drops of tea tree oil or diluted hydrogen peroxide solution at a strength of one teaspoon of three-percent peroxide to a glass of water, or more or less according to your sensitivity.

For those who do not ingest acidophilus/ bifido products regularly, take a course of "friendly" bacteria once a year as a precaution, but especially with or after taking any drugs, medical or recreational, or if you're leading a stressful life. The diet during this therapy should be free of sweeteners, sweet foods, yeast, alcohol, and initially also of fruits. Various supplements, such as caprylic acid, can be useful but are not necessary with this flush method.

If the Candida problem is severe or has existed already for a long time, chances are that the fungus has already invaded the bloodstream and internal organs. In this case, it is best to take an herbal parasite cure and at another time a course of Lugol's solution as explained under Candida and Parasites in Step 44.

If your blood is contaminated, then you may initially experience a die-off reaction of the Candida, causing weakness and possibly headache or nausea. If this happens, cut back temporarily on the amount of Lugol's solution and drink plenty of water and diluted teas or juices.

Finally, as an addition or an alternative to Lugol's solution, you can use a Beck-type blood purifier or zapper to eliminate Candida and other fungi and infectious agents from the blood (see Step 19). The magnetic pulser can be used to sanitize pockets of Candida infestations inside the mouth, vagina, or underneath the skin.

* * * * * * * * * * *

Step 49

CARDIOVASCULAR DISEASE

Managing cholesterol and improving your heart health - some practical tips.

For many years, we have been inundated with anti-cholesterol campaigns and suddenly now we hear or read of one scientific study after another telling us that cholesterol, after all, may be good for us.

The first major shock for the cholesterol - heart disease theory came in 1990 with an article in the British Medical Journal.[65] It was an analysis of six major intervention trials involving tens of thousands of people over several years. The result was unexpected. There was indeed a small reduction of about ten percent in cholesterol levels and a reduction of 14 percent in the rate of death from heart disease in the groups treated medically with drugs and low cholesterol diets.

However, the total death rate of patients with medically lowered cholesterol was significantly higher as compared to controls. This included not only higher mortality from cancer, but surprisingly, on average, a 67 percent higher mortality from violent death, such as accidents, homicides, and suicides. This high rate of death from violent causes was found in every single one of these trials, so it was unlikely to be a coincidence.

In looking for an explanation, researchers discovered that monkeys became more aggressive on a diet low in cholesterol and saturated fat, and human studies further found that criminals, on average, had lower cholesterol levels. This applied generally to individuals with aggressive or violent behaviour or limited self-control and also those involved in homicide and suicide.

Another scientific publication voiced concern over evidence that about half of the men who die of sudden heart attack do not have any of the risk factors commonly associated with heart disease, such as elevated cholesterol or blood pressure, diabetes, obesity, or smoking. This would suggest that the main cause of heart attacks is still unknown to medicine.

A Finnish study followed 1,222 Helsinki businessmen considered at high risk of heart attack for 15 years. Half were put on an intensive program of dietary regulation and other treatment while the other half served as controls. After 15 years, the low-cholesterol group had 67 deaths overall with 34 from heart disease. The control group had only half the overall death rate with 14 cardiac deaths. In light of this result, the medical director of the British Heart Foundation admitted that the existing advice to cut cholesterol was not based on “definitive research.” What, then, is it based on?

Another recent medical study confirms the general harmlessness of elevated blood cholesterol levels. Of a large group of elderly patients with high blood pressure and high cholesterol, those with the highest cholesterol levels lived longest while those with less elevated levels had the highest mortality rate. Cholesterol news that hit the headlines in 2001 showed that for old men it is best to have a “normal” blood cholesterol level because both high as well as low levels lead to an increased rate of heart attacks.

Why We Need Cholesterol: The cholesterol saga started several decades ago when researchers found that those dying from heart disease also frequently had elevated blood cholesterol levels as well as fatty plaque, called atheromas, which clog up the arteries of the heart muscle. These atheromas consist of wildly proliferating smooth muscle cells filled and surrounded by a fatty sludge containing a high level of cholesterol.

Medical authorities believed that cutting down on our cholesterol intake would lower our cholesterol blood levels and thus reduce the risk of heart disease. Biochemists, however, were generally sceptical of this idea. After all, the liver itself produces most of our cholesterol requirements. Approximately one gram daily is synthesized from the breakdown products of saturated fats and sugars and less than half a gram is absorbed from our food. The more cholesterol we absorb, the less is produced in the liver. With this, our overall cholesterol level remains fairly stable and any surplus cholesterol is normally excreted with the bile.

Cholesterol is a very valuable and useful substance. It provides the basic structure for the synthesis of steroid hormones and vitamin D and it is a major component of cell membranes; it is especially high in the brain and nervous system. In the liver it is converted into bile acids and bile salts needed for the digestion and absorption of all lipids (fats and fat-related substances). However, if we are deficient in the emulsifier lecithin or in sulphur amino acids, cholesterol can accumulate as gallstones.

So you can see that it is very important for us to have enough cholesterol. In fact, some wasting diseases, such as cancer, are associated with a cholesterol deficiency, and hypoglycemics are commonly found to have low cholesterol levels.

How We Overproduce Cholesterol: Why then do some people get raised blood cholesterol levels when the liver is supposed to keep cholesterol levels stable? Obviously, something must interfere with the regulating mechanism in the liver.

One of these causes is a high intake of saturated fats. These are broken down in the liver to keto acids that are normally used for energy production in the muscle cells. For this, the keto acids need the breakdown products of the sugar metabolism (glycolysis). If you have a habitually raised insulin level or insensitivity to insulin (e.g., diabetes), these sugar metabolites are in short supply and this can cause keto acids to accumulate. In addition, a deficiency of chromium will also interfere with the fat metabolism. The liver now attempts to get rid of the keto acids by converting them into surplus cholesterol.

There is another mechanism responsible for cholesterol overproduction involving sugar. When you ingest sugar, your blood insulin level is raised. A key enzyme of cholesterol synthesis is actually regulated by insulin and this means a higher sugar intake generates higher insulin levels and with this higher fat and cholesterol levels.

The word sugar, however, does not simply mean sucrose, our common household sugar. There are different forms of sugar. Sucrose is a combination of glucose and fructose. In susceptible individuals, fructose has a much greater influence on insulin levels than glucose. Enzymes still present in human liver biopsy samples converted fructose into fatty acids and cholesterol at rates three to 24 times faster than glucose, with the highest rates occurring in those with a disposition to atherosclerosis.

An even stronger insulin effect (stronger, that is, than by ingesting equal amounts of glucose and fructose) occurs when eating these simple sugars chemically combined as sucrose. This is called the disaccharide effect. In a study, monkeys produced three times more deposited lipids on a diet high in sucrose than with glucose.

Oxidized Cholesterol: We can now see that a diet high in sugar and fat is the main cause of a raised cholesterol level, while the cholesterol content of our diet has been repeatedly shown to have little impact. Yet this still does not prove that there is a causal relationship between elevated cholesterol levels and heart disease.

However, such a direct connection has been shown to exist with oxy-cholesterol. If cholesterol in food is heated, especially with access of air, then it is oxidized to oxy-cholesterol. Animal experiments with high intakes of either pure cholesterol or oxy-cholesterol have shown atherogenic degeneration only with the latter, not with cholesterol itself.

This would imply that a main contributing factor to our high rate of heart disease is our habit of overly heating and frying meat, eggs, and milk products. However, once the arteries are damaged, then high cholesterol levels may or may not contribute to the deposits of fatty sludge.

Further clues can be found in the history of heart disease. Heart attacks were extremely rare in earlier centuries and even at the beginning of the twentieth century. A medical textbook published in 1912 describes angina pectoris, a mild form of heart attack, as being very rare, and a famous American heart specialist did not see his first heart attack patient until 1920. Yet beginning in 1920, the explosive rise in the incidence of heart disease in Western countries was on.

As China and other less developed countries remained free of heart disease, then we may ask: What fundamentally changed the lifestyle of Western populations at that time? Two important changes took place: the pasteurization of milk products and the chlorination of public water supplies.

Pasteurization of Milk Products: Before pasteurization, some doctors recommended milk as a cure for many illnesses. Raw milk cures were famous at the beginning of the twentieth century. Formerly, people consumed plenty of cream and butter and rich meals prepared with them, and in Bulgaria and the Caucasus they set records in good health and longevity on full-fat milk products. An American doctor reportedly cured psoriasis by having his patients eat two pounds of butter a week. Despite all this fatty food, heart attacks were virtually unknown.

Now milk is no longer a cure for any disease. On the contrary, natural therapists generally regard milk as disease-forming, butter is known to aggravate psoriasis, and heart disease is rampant. What is the difference? Enzymes plentiful in raw milk are destroyed by pasteurization.

Some diseases that were formerly improved with raw milk can now be successfully treated with high-level enzyme supplements. Specifically, it is the fat-digesting enzyme lipase that works most of the healing miracles. With cardiovascular disease, there is a deficiency of lipase in the bloodstream. Lipase released by the pancreas helps to digest fats, while lipase in the bloodstream and cells helps to break down unwanted fatty deposits.

There are, however, fats and oils that arguably are much worse than pasteurized butter. These are the oils with a high content of unnatural trans-fatty acids, such as margarine, shortenings, and hydrogenated oils and fats and, to some degree, even our common salad and cooking oils. Trans-fatty acids are produced not only by hydrogenation, but also by the heating of oils. Fats and cholesterol are transported in the blood in the form of lipoproteins, and trans-fatty acids produce a more unfavourable lipoprotein profile than any other fat.

Recently I heard a cheese producer say that he submitted some of his self-produced unpasteurized cheeses for testing. They were completely free of pathogenic bacteria, unlike pasteurized products that commonly contain disease-causing microbes up to a legal limit. The reason for this superior safety of raw cheeses is their high content of lactobacilli. These kill harmful microbes, while in pasteurized products, any survivors or contaminants can multiply unchallenged. Therefore, milk products correctly fermented with lactic acid are not only healthy but also safe, unlike their pasteurized counterparts.

Chlorination of Public Water Supplies: Experimental use of chlorine to kill bacteria in public water supplies began around 1900 and was generally accepted in Western countries by the I920's. Part of the chlorine reacts with organic impurities to form organochlorines (DDT is an organochlorine), while the rest remains as residual-free chlorine in the water. It can then react with food chemicals or with parts of our digestive tract.

From 1920 onwards, the explosive increase in the incidence of cardiovascular disease and fatal heart attacks began, but only in countries that chlorinated their water supplies. These diseases remained unknown, for instance, in China, Japan, Africa, and Asia. However, Japanese citizens who emigrated to Hawaii where water was chlorinated suffered the same rate of heart attacks as the Americans, and the African-American population in the U.S. has the average U.S. rate of heart attacks, but not that of their counterparts in Africa.

In 1967 an American researcher, J. Price, M.D., performed a decisive experiment. With one group of 50 three-month-old chickens, he added one-third of a teaspoon of chlorine bleach to about one quart of water in their feed, while another group of 50 chicken served as controls. Seven months later, more than 95 percent of the chlorinated group had advanced atherosclerosis, yet none of the control group showed any such evidence. In the following years, Dr. Price repeated his experiment many times, always with the same results. Recently, researchers funded by the U.S. Environmental Protection Agency confirmed atherosclerotic-type changes in other animals, including monkeys, when exposed to chlorinated water.

Public health officials claim wrongly that water chlorination is essential to prevent microbial contamination. Public water supplies can be made safe with ozone or exposure to ultraviolet light at no greater expense than with chlorine. Even sunlight is very effective, with 99.9 percent of coliforms (pathogenic bacteria) killed within 90 minutes in one experiment. For sterilizing our individual water supply, we can simply add a teaspoon of three-percent hydrogen peroxide (food-grade) to a gallon of drinking water or expose it to sunlight in a shallow container or a clear glass bottle for several hours.

Practical Recommendations: Other contributing factors in cardiovascular disease are vitamin B6 and magnesium deficiency, high ferritin or iron levels, and a low intake of omega-3 essential fatty acids. Kilmer S. McCully, M.D.,66 discovered in the 1960s that with vitamin B6 deficiency a toxic breakdown product of the amino acid methionine occurs. This chemical, homocysteine, causes free-radical damage to the arteries. Even young children with vitamin B6 deficiency were found to die of rapidly developing atherosclerosis. As a result of this discovery, Dr. McCully lost his professorship at Harvard Medical School in 1979 and was turned down in 51 job applications in the following two years. Now he has been vindicated and his discovery is mainstream medical theory.

Moses M. Suzman, M.D., a neurologist in Johannesburg, South Africa, specialized in treating heart patients with 200 mg of vitamin B6 daily and B-complex tablets. He had remarkable success and extremely low re-infarction rates. Researchers have found that folic acid lowers homocysteine levels, especially in combination with vitamin B12; high iron levels have been strongly linked to free-radical damage in the arteries and subsequent plaque formation.

Chronic inflammation has been shown to be strongly involved with the development of arterial plaques. Often this inflammation is due to chronic bacterial infections. A high intake of antioxidants, including the various bioflavo-noids and carotenoids, as well as omega-3 essential fatty acids in linseed oil (or ground linseed) and fish oils, protects the artery walls with their anti-inflammatory and anticoagulant properties. A good anti-inflammatory herb for everyday food flavouring is ginger root.

For good heart health, avoid highly heated fatty and oily foods, especially if heated with access to air, as in frying. Not only are heated animal fats harmful when they form oxy--cholesterol, but so are oxidized and peroxidised polyunsaturated oils. Use vegetable oils that have not been heated to more than 120° F. Generally, I trust only high-quality linseed oil and extra-virgin olive oil to have retained their lipase. In addition, there are some good specialty oils, the labels of which state that the oils have been produced below 120° F. The cold-pressed oils in health food stores may or may not be completely trustworthy but are better than conventional supermarket oils.

Cut down on sweet food, in particular on sucrose, and minimize the combination of starches with anything sweet, as that causes an exaggerated insulin response and all its harmful effects. Basically, this means eat fruits on their own before or instead of any meals.

Fish oils and linseed oil are high in omega-3 essential fatty acids that are anti-inflammatory and protect our arteries from clogging. Various specific nutrients have a protective effect, such as ginger, green tea, grape seed extract, vitamins C and B6, natural vitamin E, and magnesium, preferably as magnesium chloride. All foods with color are good and especially purple foods, including red wine. In addition to an anti-inflammatory effect, these nutrients have an anticoagulant effect and discourage the formation of damaging blood clots. With seriously ill patients, the "Pauling therapy" is helpful: In addition to other recommended nutrients, use 5 to 10 g of vitamin C, 5 to 6 g of the amino acid L-lysine, and 2 g of L-carnitine daily, all in four divided doses daily.

Reflexology describes a connection between the heart and the soft pad at the base of the left thumb. If there is a heart problem, this part will be sensitive when pressed. Heart diseases and especially acute conditions often can be improved by repeatedly and for long periods pressing strongly into the base of the left thumb. Heart attacks reportedly have been stopped by doing this. In an emergency, press as hard as you can; at other times, only as much as you reasonably can stand the pain. Gradually, as the condition improves, the tenderness at the base of the left thumb will diminish. See also the entry on Cardiovascular Disease in Step 44.

* * * * * * * * * * *

Part 7
ENERGIES
Dealing with Harmful & Helpful Ones

This chapter deals mainly with the nonphysical energies, variously called life force, bioenergy, prana, chi, od, orgone energy,[67] and many other names. It is this energy level that we most easily see as the innermost aura around living objects or that we sometimes feel as various sensations such as heat or tingling in our body, especially when transferring energy, as with “laying on of hands” or Reiki. It is also the energy that fuels our sexuality and that we feel most strongly during orgasm.

This energy circulates in our acupuncture system and its vortexes form our chakra system (non-physical energy centres). It is not only at the root of most paranormal or psychic phenomena, but is the healing agent of many natural therapies. Homeopathy can only be understood as a bioenergetic phenomenon; that is why it is so fiercely rejected by conventional science, in spite of scientific evidence that it works. While bioenergy is still unknown to orthodox science, it is an everyday experience of many healers and sensitive individuals. Just as electromagnetic energies have a wide range of wavelengths or frequencies, so also have the bioenergies. On one side of their spectrum, the bioenergies border on the electromagnetic spectrum while on the other side they merge into the emotional energies.

In addition to food, water, and air, we need a balanced intake of electromagnetic energies. These include sunlight, color, the Earth’s magnetic field, and the electric field between the ionosphere and the ground. Sunlight actually is a combination of electromagnetic energy and bioenergy, and in this way we receive most of our bioenergy from the sun.

* * * * * * * * * * *

Step 50

ELECTROMAGNETIC ENERGIES

Protect yourself from the harmful effects of electromagnetic radiation and make sunlight your friend.

Strong electric fields of about 50 cycles per second (hertz, or Hz) are harmful to us. These fields exist around power lines, power tools, electric stoves, heaters, boilers, freezers, and television sets when in use and extend several feet or yards around the appliance. Stay away from them if possible. Using an electric iron or an electric keyboard or working with handheld power tools can easily drain your energies. We can reduce harmful effects of electricity on our body by holding our hands under running water from time to time when working with electrical equipment and also by having a shower in the evening. For a detailed discussion of this, review Step 8.

The DC Field of the Earth: We live in a natural DC (direct current) field. The surface of the Earth has a negative electric charge and the ionosphere high above it a positive charge of about 50,000 volts. However, the electric currents involved there are extremely small. Near the surface of the Earth, the voltage differs greatly; it is highest on mountains and in dry conditions where it can reach 1,000 volts per meter, while in low-lying and moist conditions it falls to below 50 volts. This means that the voltage difference between a point close to the ground and one meter above it can be between 50 and 1,000 volts, but the current is so low that we do not feel it. Nevertheless, it has a noticeable effect on our well-being. Populations with exceptionally long life spans always seem to live in the mountains, such as in the Andes, Himalayas, and Caucasus.

In buildings with roofs of metal, aluminium-coated insulation material, or steel reinforced concrete, the electrical field is near zero. It has been shown that this lowers our concentration span and well-being. There may even be counter-fields caused by the static electricity of plastic materials and synthetic carpets. Even worse are houses that have metal walls, whether they are of corrugated iron, coating or insulation material with aluminium, wire mesh, or reinforced concrete. Getting fresh air through open windows will improve the condition to some extent.

In a high DC field, our well-being and mental abilities are increased, while in a low field we easily become fatigued and our concentration suffers. Electrodes that produce a field of 1,000 to 2,000 volts inside offices and houses have been used with good results in Germany. If you live in a conventional modern house, have at least the bedroom or sleeping place free of shielding metals, or sleep on the veranda in summer. This is especially important for someone with a serious chronic disease. Because as a species we lived so long in this natural DC field, the natural ion distribution and electrical fields in our bodies have adjusted to it. We feel better when we are exposed to this field. This is one of the reasons why walking on fresh grass feels refreshing, while habitually wearing shoes with rubber or plastic soles is more tiring.

The magnetic field of the Earth is about 0.5 gauss, though it has been much higher in the past and is presently decreasing. Generally, life forms are bigger, more vital, and longer living in a stronger magnetic field. Most people will find it best to sleep with the head in a northerly direction, while an easterly direction is sometimes beneficial; facing towards the west, however, is often weakening. In the Southern Hemisphere, you can also try sleeping with the head to the south. Experiment to find your best direction.

Sunlight: On the level of energies, sunlight is our primary source of life. Light to moderate sun exposure is beneficial. Except in winter, do not sunbathe during the hours around midday. Preferably expose your whole body as often as possible for a few minutes in the early morning or late afternoon sun. This may be in the garden, the veranda, or even inside a room with an open window. In addition, it is beneficial to sit lightly clothed in the half-shade, as for instance under a tree, as we can absorb the bioenergies of the sunlight through our clothing.

While brief exposure to the sun is invigorating, long sunbathing, especially at the beach, is exhausting and ages the skin and can cause skin cancer. It can also cause an overactivity of the parathyroid glands. Vitamin supplements can protect us from these dangers: In addition to PABA and other B vitamins, we need plenty of antioxidants, bioflavonoids, and carotenoids for sun protection. Except for zinc cream or a solution of PABA, I am not in favour of using sunscreen lotions; instead, use suitable clothing for protection. Sunscreens contain unbiological chemicals, and it has been reported that most commercial sunscreens kill skin cells and so may inflict long-term skin damage.

Research shows that there is a strong negative correlation between available sunlight and breast cancer death rates, that is, living in a sunny area is associated with lower cancer rates. Even skin cancer is inhibited by regular low-level sun exposure; only sunburn, or high-level sun exposure, is a strong skin cancer promoter. It has been stated that trends in the epidemiological literature suggest that approximately 30,000 U.S. cancer deaths could be averted yearly by the widespread public use of regular mild sun exposure. A study on electricity workers showed that those with the most sun exposure had the lowest rates of skin cancer and melanoma. Other studies show a strong link between long-term exposure to fluorescent lighting and melanoma; melanoma often occurs on areas of the skin that have not been exposed to sunlight.

Sensitive individuals, especially with low blood pressure, lack of energy, and a tendency to have food allergies and chemical sensitivities, should expose themselves to the full sunlight for only very short periods; preferable is semi-shade, as under a tree. Insensitive individuals, on the other hand, with insensitive skin and often high blood pressure, benefit from a longer exposure that tends to sensitize them.

The eyes are important entrance points for sun energy. John N. Ott, in his Health and Light, demonstrated with his research how important it is for us to get the full spectrum of the sunlight into our eyes. This does not happen if we sit behind windows or wear glasses; these filter out essential frequencies of the ultraviolet part of the spectrum. Deficiency of this ultraviolet radiation caused by our living behind glass is a contributing factor in the development of degenerative diseases such as arthritis, cancer and leukaemia, eye diseases, hyperactivity, and sexual problems. These harmful effects are reinforced if the glass is inappropriately tinted, in this way blocking out a part of the visible spectrum as well.

Ott demonstrated that the primary effect of light entering our eyes is on the endocrine glands of the brain. He also showed that coloured lights have a strong influence on the ratio of male to female offspring when breeding animals. Especially unhealthy are most kinds of fluorescent lighting, more so if they have a pink hue. Even more harmful is light from the color screens of television sets and computer monitors; this effect is due to the artificial composition of these colours.

However, if one has sensitive eyes, one should continue to wear sunglasses if the light is very strong and glaring and only gradually reduce their use when the eyes improve. The best sunglasses, if required, are of a special composition of gray so that they reduce the intensity of all colours to the same degree. Most plastics, by the way, do not screen out ultraviolet rays, so in this way one obtains full sunlight by living behind clear plastic windows. However, a disadvantage of this is that one can get sunburned behind these plastic windows.

Lack of sunlight in winter or if one is bedridden makes us susceptible to depression. You can cure this seasonal affective disorder by looking into a bright light in the morning, preferably sunlight or lights with the full daylight spectrum.

Sunlight is also required to convert cholesterol into vitamin D. Multiple sclerosis, for example, is inversely related to the amount of sunshine in the region, so that the closer to the equator, the lower the incidence of it. This is similar to the rate of cancer that also declines the closer we get to the equator. Another important aspect of sunlight is our indirect intake of it in the form of sun-drenched fruits and vegetables. Such foods absorb bioenergy as well as light frequencies from the sun, but these beneficial energies are gradually lost if these foods are stored in the dark. In this case, it is beneficial to expose them again to sunshine for a short period before eating.

Coloured Light: Exposing ourselves to light of a certain colour has a definite effect on our emotions and body. The most balanced colour is green, and looking out over green meadows and at the green of trees has a balancing effect on body and mind. Blue and red, on the other hand, are the main opposites that help to adjust the body if it is out of balance. Red and especially orange are used for stimulation when we are fatigued and need strength. Too much of it, however, will make us nervous and irritable. Blue, of course, is sedating, even more so than green, not only for the mind, but also for the body if there is pain, fever, or inflammation.

There are many different ways in which we can treat ourselves with colours. The most natural way is to look at distinctive colours in our surroundings. In addition to looking at meadows and trees for their balancing green, we can look for long periods at the blue sky to gain peace and quiet. For stimulation, on the other hand, we can focus on a red rose or any other red flower bathed in the sunshine.

Another possibility is to paint or wallpaper a room in appropriate colours and to pay attention to the colours of our carpets and curtains. Then we can select suitable colours for our clothing. Usually we express through the colours of our dress our inner feeling at the time, for instance, dressing in black or gray if we are depressed, thereby reinforcing the negative mood. However, if we want to improve our emotional condition, select the opposite colour to our negative feeling. To ward off depression, for instance, wear pink or another bright and stimulating colour.

We can use differently coloured light bulbs, coloured lampshades, or coloured cellophane. However, most one-sided colour exposure should be only temporary, for as long as it has a beneficial effect on our condition. Too much blue, for instance, can push a susceptible individual from being peaceful to becoming depressed, or blue. Generally, it is best to use a mixture of colours inside the house and for our clothing, with possibly one dominant colour that balances our basic condition. For further information, see Colour Therapy in Part 2.

* * * * * * * * * * *

Step 51
BIOENERGY

Bioenergy is the true energy of healing, so for optimal health learn as much about it as you can.

Bioenergy is the basis of all biological manifestations. It is also called prana, ki or chi, od, orgone, ether, bioplasma, or simply life force. Our main intake of this life energy is during deep sleep, through the chakra system, breathing, from raw food, as well as sunshine and energy transfer from other people. In the air, it is commonly called prana, while within our body we usually refer to it as bioenergy; for the Chinese, it is the chi that circulates in the acupuncture meridians throughout the body. Bioenergy generally appears as a bluish-gray haze; its movement is slow and wavelike; it can be accumulated and conducted by organic material and water; and it can be separated into two opposites with a bluish and a red-yellow color. The air is low in bioenergy on rainy days and at night. The planet is enveloped in a field of orgone energy and on most days this field can be observed moving from west to east.

Our physical body is permeated with bioenergy. Part of this energy circulates through the acupuncture meridians while the surplus radiates out, forming an etheric field, or the inner aura, and protects us from harmful environmental radiation or microbes. For the body, bioenergy holds the blueprint that guides the growth and maintenance of all body structure; it is the “glue” that holds the cells together, the energy of organization.

Acupuncture points are like transformers or inflow points of bioenergy from the meridians into the tissue structure. If this flow along a meridian and into the muscles is blocked, a limb becomes lifeless and paralysed, even if the blood circulation and nerve impulses are intact. We can direct and modify this stream of bioenergy with our mind to heal ourselves. Generally, we can increase our inner reservoir of bioenergy by living in an area with a high environmental potential of bioenergy. We can find this where there is plenty of sunshine, abundant vegetation, especially trees, and an elevated position - ideally a tree-clad hilltop or mountain plateau in the subtropics.

Bioenergy in Healing: In 1973, researchers at the University of California at Los Angeles succeeded in replicating a Russian experiment in which part of a fresh leaf was cut off and Kirlian (high-frequency) photography revealed the faint outline of the whole leaf, including the missing part. This can be regarded as the first visible scientific proof of the existence of an aura around living objects. Before that, Professor Harold Saxton Burr of Yale University had made extensive measurements of the electric fields around living things and shown that these “L-fields” (Life-Fields) pre-exist as in a fertilized egg and guide the development of the related entity.

I contend that prana, bioenergy, or life energy is a major factor in healing despite the denial of its existence by orthodox medicine. Chances of recovery are greatly increased by being regularly exposed to high concentrations of prana.

For example, Wilhelm Reich in The Cancer Biopathy described his research into the effect of prana on cancer development and treatment. With a powerful microscope (4,000-times magnification), he could observe how cancer cells originated from body cells deficient in prana and how cancer cells were destroyed by red blood cells highly charged with prana. Healthy red blood cells have a strong, blue, and pulsating field, he said. His research led him to believe that cancer is basically due to chronic sexual starvation or, more precisely, to blockage of sexual energies due to emotionally induced muscle tension or armouring. This is just another way of saying that cancer is primarily caused by a lack of bioenergy or a blockage in the circulation of bioenergy in the body.

Prana is high in green surroundings and especially tree-clad hilltops and mountains. It is high in the air on sunny days, but low on dull days or at night. It is exceptionally concentrated in rainwater collected after a heavy shower on a sunny day. Use such highly charged rainwater for washing and drinking. In polluted areas, purify it with a jug-type carbon filter before ingestion. In addition, fresh rainwater, and especially rain that fell during thunderstorms, has a beneficial negative electrical charge, but this is discharged and lost if the rainwater comes in contact with a metal.

Prana is much enriched in cascading sunlit streams and even in artificial sunlit cascades. Water can also be improved by exposing it to sunlight before use. Fresh fruit and vegetables stored in the dark and especially in a refrigerator have a healthier benefit for us if they are exposed to sunlight before eating.

Revitalizing may be the most important task with chronic diseases (and especially with cancer). As Reich suggested, healthy blood charged with bioenergy readily destroys cancer cells and disease germs. To vitalize, use fresh raw juices, walk barefoot on dewy grass, do deep breathing exercises, meditation, energy healing, and energy cultivation, and use devices such as magnets, cones, and pyramids.

Freshly cut grass emits a strong radiation of bioenergy. In a similar way, a “hay cure” was formerly used with good results as a folk remedy for many diseases. The patient was immersed for hours in freshly cut hay or sometimes in fresh leaves. Kirlian photography shows that cut leaves and grasses lose most of their prana within a few hours. This outpouring energy can be used to radiate a tumor or a weak organ with a strong charge of prana by keeping a pack of freshly cut grass over the area in contact with the skin. The pack should be renewed every few hours.

Morning dew on fresh grass is another source of highly concentrated prana. Walk on it barefoot for ten to 20 minutes; with your hands, rub the dew onto your bare skin all over the body; if it’s warm enough, roll with bare skin in the dewy grass. Psychics can see our aura expanding when we walk or stand on fresh grass.

It is equally important to conserve our bioenergy by living in or visiting surroundings high in bioenergy. Generally, these are sunny locations with luxurious plant growth, but also mountains and the seashore. Some individuals claim to have lived on prana in the mountains but were not able to do this in the lowlands or a city. Males can lose much bioenergy through ejaculation and should minimize such loss when low in energy, as with cancer, AIDS, and other debilitating diseases. Instead, they can try Karezza-type (caressing with slow movements) lovemaking or other forms of spiritual sex as explained later in Part 7.

Basically, we get bioenergy from:

- Recharging our etheric body during deep sleep
- Ingesting fresh raw food, juices, and charged water
- Breathing prana-rich air
- Meditation and guided imagery
- Stepped-down energy entering directly through the chakra system

A high body charge of prana, visible to psychics as a strong and clear aura, is a main requirement for remaining healthy and overcoming chronic diseases. Therefore, protect and improve your health by exposing yourself to high concentrations of prana in any or several of these indicated ways. If you have the opportunity, visit known natural power spots or high-energy areas.

There is also negative and potentially dangerous bioenergy. This is associated with diseased and polluted conditions and also with radioactivity. It appears that healthy bioenergy moves or spins at certain frequencies. When it stagnates, as with energy blocks in our body, it becomes more dense with a slower vibration. This negative energy not only acts to maintain a diseased state, it can also be absorbed by an unsuspecting healer and can then cause health problems in her body. Flowing water is a great cleanser for this and can be used in different ways to remove negative energy.

Vivaxis Energies: Vivaxis is a term coined by Canadian researcher Frances Nixon in the 1960s for etheric energy beams that link us to specific locations. The most important one of these is our birth vivaxis, linking us to key energy points near the place of our birth in a two-way circuit. Perhaps migrating birds and insects find their distant destinations by being able to sense and follow these vivaxis beams. This would also explain how dogs sometimes cover hundreds of miles of unknown territory to turn up unexpectedly at their owner's house or how turtles can return to their beach of birth after several decades to lay their eggs. In addition, there are Earth beams circling our planet, said Nixon. These contain the frequencies of all the elements necessary for life and are related to our individual energy fields by our vivaxis energies.

Health problems may arise if our energy absorption from these essential energy beams is disturbed, Nixon proposed. This may happen as a general planetary condition after nuclear testing or, on a more individual scale, if we live in dwellings encased by metal (aluminium coating or insulation, steel-reinforced concrete, wire mesh). Further, our internal energy receptors may become disturbed after exposure to X-rays or other strong electromagnetic radiation, by wrongly applied strong magnets, prolonged mental stress, marijuana, vibrating tools, infrasonic waves, environmental imbalances, and pesticides as well as toxic chemicals in various drugs, foods, and cosmetics.

Foreign vivaxes to which we may become linked can be created by hypnotism, strong magnets, electrical acupuncture treatment and pulse-wave machines, accidentally received electric shocks, and even contact healing, reflexology, and acupressure, especially if a therapist exerts strong pressure on bilateral points. Some of these foreign vivaxes can disturb our natural-force flows before they gradually disappear. On a certain energy level, we seem to remain linked to everyone we ever touched.

You can reduce the possibility of health problems arising from disturbed or blocked flows of these subtle energies by living as naturally as possible. Especially minimize pollution from non-biological chemicals and strong electromagnetic fields; frequently cleanse and recharge yourself in natural surroundings, by walking barefoot on grass or swimming in the ocean or an unpolluted lake or river. Preferably sleep in a part of the house with a minimum of metal shielding (as from insulation material) in the walls and roof. For more information, see Judy Jacka's “The Vivaxis Connection” (Hampton Roads, 2000).

* * * * * * * * * * * *

Step 52
CULTIVATING THE ENERGIES
Learn to sense subtle bioenergies and use them to energize your body.

For increasing or maintaining our health and vitality, it is advisable to work directly with our life force, cultivating it, as the Chinese masters say. There are two main aspects to this energy cultivation. First we learn to accumulate bioenergy or prana and then use it by directing its flow.

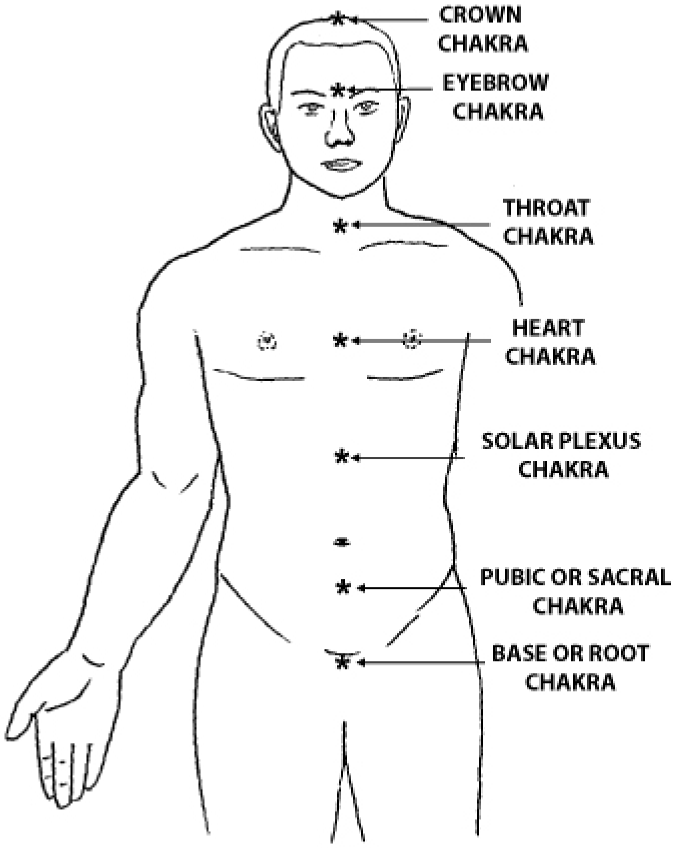

Figure 7-1: Distribution of energy centres or chakras

Intake through the Chakra System: Our main bioenergy intake and distribution system is the chakra network, subtle energy centres along the spine from groin to head. The seven main chakras are comparable to central power stations, while the acupuncture meridians can be seen as power transmission lines. Chakras are vortex-like energy centres. Commonly, seven major chakras are mentioned, six of which are rooted along the spine and one at the top of the head, as shown in figure 7-1. The chakra vortexes suck in energy from the universal energy field that is all around us.

Each chakra modifies the vibration rate of the energy and that gives it a certain color. Further, each chakra energizes a specific region of the body and is related to a major nerve centre and endocrine gland. The energy then flows along a multitude of lines (called nadis) to form the etheric body. Major energy flows are along the acupuncture meridians that, in turn, nourish the internal organs and muscles.

A health problem in a certain area of the body is usually associated with a disturbed function of the chakra governing this region. Normalizing and harmonizing our chakra functions is important for the healing of specific aspects as well as for increasing our available energy in general and for becoming whole and balanced.

The characteristics of the seven main chakras can be seen in table 7-1. In addition, there are many more secondary chakras. Those in the palms of the hands are commonly used for energy transfer. Each of the higher or finer energy bodies, such as the emotional, astral, and mental body, has its own chakra system; different energy bodies have different chakra colours. The corresponding chakras in the different bodies are interconnected. The more we spiritually develop, the greater becomes the exchange of energy between the different bodies, with the energy generally flowing from the finer to the denser bodies. Excellent books on energy healing, energy bodies, and the chakra systems are “Hands of Light” and “Light Emerging”, by Barbara Ann Brennan.

You can experiment with two complementary methods of increasing your energy reservoir. In one, imagine energy entering through the crown chakra; in the other, imagine you retain the prana inhaled with the air. With some practice, you can combine both exercises and do them at the same time. The main rule in all of this is that energy flow follows our awareness and intentions.

Table 7-1: The Main Energy Centres or Chakras

Location	***Sanskrit name***	***Gland***	***Color***	***Vertebra***	***Function***
Root/Base	Muladhara	Adrenals	Red	Sacral 4	Will to live, vitality, grounding
Pubic	Svadisthana	Gonads	Orange	Lumbar 1	Sexual energy, physical creativity
Solar plexus	Manipura	Pancreas	Gold-yellow	Dorsal 8	Emotional energy, feeling safe, protected, connected
Heart	Anahata	Thymus	Green	Dorsal 1	Spiritual feelings, love, compassion
Throat	Vishudda	Thyroid	Blue	Cervical 3	Communication, creating with words and sounds
Brow	Ajna	Pituitary	Indigo	none	Higher mentality and creativity, intuition, visualization
Crown	Sahasrara	Pineal	Violet/white	none	Spirituality, connection to Higher Self and divinity

Chakra Exercises: Sit or lie in a relaxed position and focus on one chakra at each session. Start with the crown chakra. Keep your awareness centred at the top of your head until you feel a light pressure, warmth, tingling, or other sensation developing in that area. Whatever you can feel, intensify it. However, it is preferable to replace any feeling of pressure with a feeling of warmth as soon as that can be done, in order to avoid an uncomfortable pressure from building up.

You can also imagine a stream of white light entering the top of the head or visualize an energy vortex in this position. Let it spin clockwise in the Northern Hemisphere and counter-clockwise in the Southern Hemisphere. Alternatively, imagine feeling it as a pulsating energy field. You can combine all or several of these imaginations and use whatever is most suitable for developing a feeling awareness of this chakra.

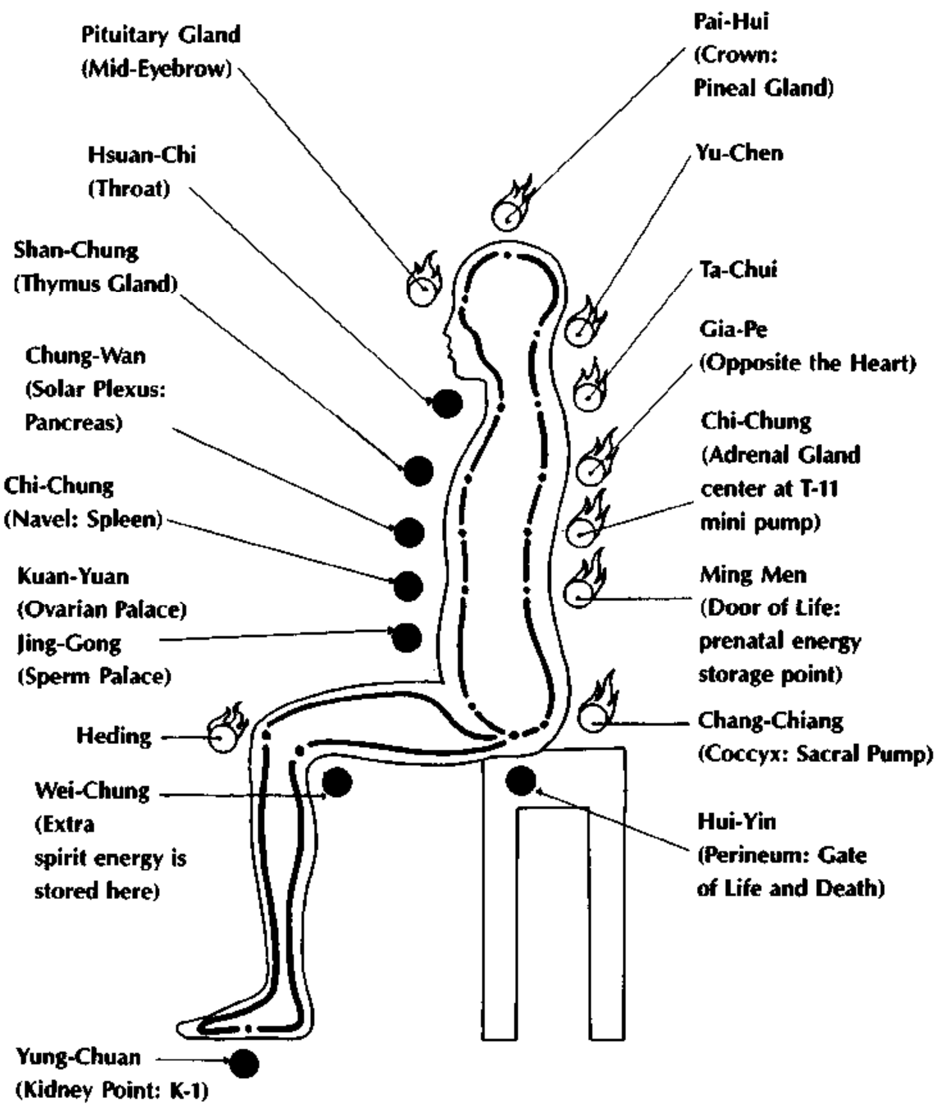
Pituitary Gland
(Mid-Eyebrow)
Hsuan-Chi
(Throat)
Shan-Chung
(Thymus Gland)
Chung-Wan
(Solar Plexus:
Pancreas)
Chi-Chung
(Navel: Spleen)
Kuan-Yuan
(Ovarian Palace)
Jing-Gong
(Sperm Palace)
Heding
Wei-Chung
(Extra
spirit energy is
stored here)
Yung-Chuan
(Kidney Point: K-1)
Pai-Hui
(Crown:
Pineal Gland)
Yu-Chen
Ta-Chui
Gia-Pe
(Opposite the Heart)
Chi-Chung
(Adrenal Gland
center at T-11
mini pump)
Ming Men
(Door of Life:
prenatal energy
storage point)
Chang-Chiang
(Coccyx: Sacral Pump)
Hui-Yin
(Perineum: Gate
of Life and Death)

Figure 7-2: The Stations of the Microcosmic Orbit, from “Bone Marrow Nei Kung”, by Mantak Chia.

Repeat, focusing on the same chakra in subsequent sessions until you can easily feel its presence. Then move to the forehead chakra. In addition to trying to feel a light pressure or warmth, it is helpful to close your eyes and focus them as if looking at the inside of the centre between your brows. You can also imagine the white light flowing into the top of your head becoming violet in the crown centre, then continuing to flow into the forehead centre and illuminating it with an indigo colour.

With the throat chakra, you can use similar imagery but make it a blue colour. Feel the warmth of your outgoing breath first in your nose and then transfer this feeling to the throat. Intensify the warmth and combine it with the blue in a pulsating rhythm.

For the heart chakra, imagine that your breath moves in and out of the chest through the middle of your chest bone. Also try this imaginary breathing through skin and bone with any of the other chakras, but it is easiest to learn it at the heart centre. In this way, learn to sense the lower chakras - the solar plexus, pubic, and base chakras. Imagine the base chakra as an energy vortex that points to the base of the spine. With this, the crown vortex and the base vortex are like an hourglass separated by the length of the spine, while all the other chakra vortexes are perpendicular to the spine.

Storing Prana: To accumulate prana from the air and store it, imagine that the inhaled breath consists of two parts: physical air molecules and prana. During inhalation, the lungs automatically separate the prana from the air molecules. You exhale only the “empty” air while the prana flows to where you lead it with your conscious attention. The body uses the abdomen with its high water content as a storage reservoir of excess bioenergy.

To energize the abdomen, feel the breath energy moving directly into the abdomen during inhalation. When exhaling, imagine pressing the energy down so that it becomes compressed at the bottom of the abdomen. With each successive breath, more compressed energy accumulates until the target area feels sufficiently energized. You may feel this as warmth, tingling, pressure, or any other sensation.

If you can visualize colours, see energy being stored as a golden-yellow fluid that gradually fills the whole abdomen. Initially, you may also see it filling the whole abdomen like a gas with a very faint color that becomes more intense with each breath. However, it is not enough just to visualize it; it is essential also to feel the energy.

Another important consideration is the actual prana content of the air you breathe. You gain the most energy when the air is high in prana, such as outdoors in natural surroundings on a sunny day. However, most of the time, you will be doing this exercise in far less than ideal conditions. Neverthe-less, persevere with it while trying to maximize the prana in your inhaled air by opening windows, exercising on a sunny day, and minimizing negative elements, such as electromagnetic forces and synthetics in your room.

In contrast to accumulating prana through breathing, there is no such restriction when taking in energy through the top of the head, as this comes from another dimensional level. But as both energies have different frequencies and complement each other, it is beneficial to work with both methods.

Circulating the Energy: A Chinese exercise to increase our energy and well-being involves circulating life-force energy in what is called the “microcosmic orbit.” Start by accumulating energy either through the crown centre or by condensing breath energy and lead it into the abdomen.

When you have a good feeling sensation of the energy in the abdomen, imagine this sensation flowing into the pubic centre. Intensify the sensation until you can feel the pubic chakra, then let the energy flow into the perineum and the base of the spine. Feel the energy there in any way you can; you may envision it having a red color in this area.

Now imagine a valve opening at the tailbone and, with each inhalation, lift part of the energy to a point in the spine behind the pubic centre. When you can feel it well, be it as pressure, warmth, tingling, or any other sensations, lift the energy with your breath to a point behind the solar plexus centre. If you find that you can lift the energy better during exhalations or by paying no attention to the breath at all, that is fine as well; experiment and use the method that works best for you.

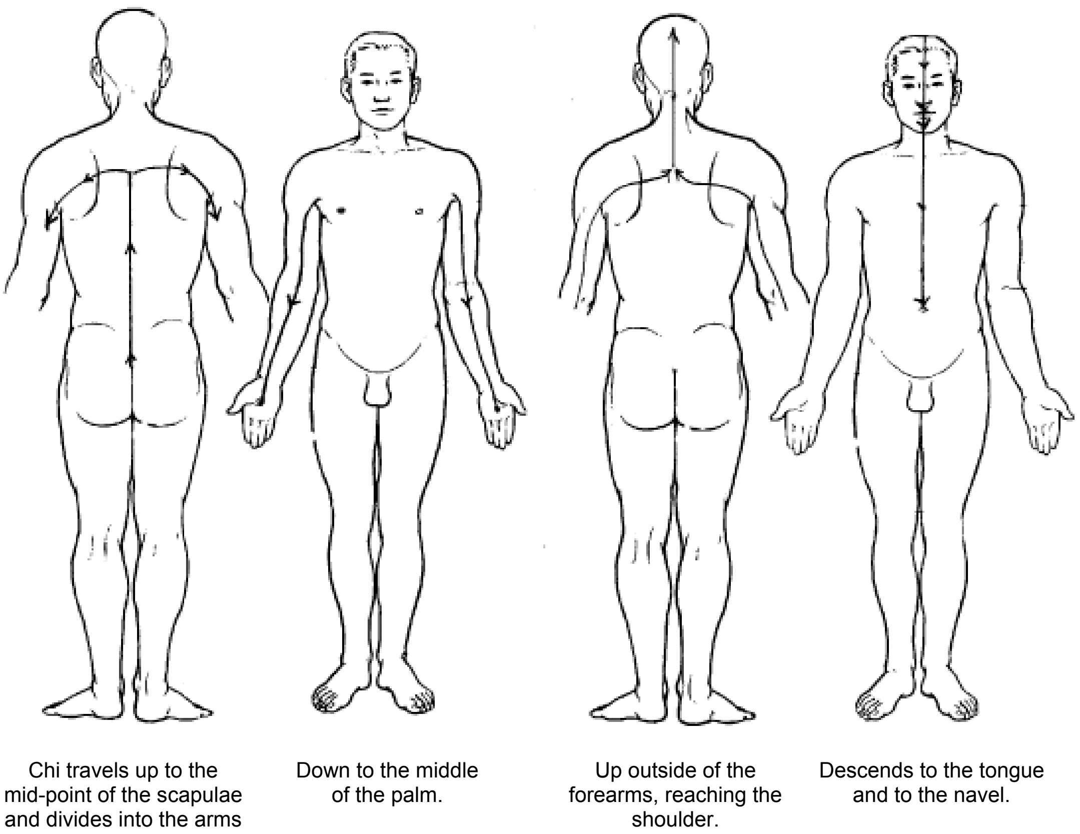
Chi travels up to the mid-point of the scapulae and divides into the arms

Down to the middle of the palm.

Up outside of the forearms, reaching the shoulder.

Descends to the tongue and to the navel.

Figure 7-3: The large heavenly cycle – the arm routes, from *Awakening Healing Energy Through the Tao* by Mantak Chia.

However, if you experience unusual heat or any discomfort while trying to lift the energy up the spine, then stop immediately or imagine that a cool energy flows downwards. The idea is to increase the flow of energies gradually, but not to open the floodgate of the Kundalini energy at the base of the spine. That might be likened to feeding high-voltage power into a system wired for low voltage.

As you continue to lift the energy to the back of the heart centre and then to the back of the neck, imagine that your spine is a straight hollow tube through which the energy flows following your intentions. Next concentrate the energy at the back of the head and finally at the top of the head. Reinforce it with the energy entering the top of the head and then lead it down into the forehead centre. It helps during the whole exercise, but especially at this time, to touch the roof of the mouth with the tip of the tongue. This makes it easier for the energy to flow farther down into the blue throat centre and then into the heart centre.

To complete the exercise, lead the energy farther down into the natural storage area for excess energy behind the navel. Imagine that it flows towards the centre of a spiral where it is condensed and stored as a ball of energy in the abdomen, ready for future use. When needed, this energy can be reactivated with your intention and used for healing yourself or others by directing it with your mind where you want it to go. It can also be used for tasks with high-energy requirements and is the same energy that masters of the martial arts aim to cultivate and use.

In addition to circulating the energy in this “microcosmic orbit,” you can sometimes include the extremities in a “macrocosmic orbit” or “the large heavenly cycle.” After circulating several times in the microcosmic orbit, stop at the perineum and divert the energy down the back of the legs. Pause at the back of the knees and try to feel the two smaller energy vortexes there. Then continue guiding the energy down the back of the calves and around the outer ankles. Feel another set of minor chakras in the centre of the soles of your feet, and finally lead the energy up again along the insides of the legs. Circle around the knees and back to the perineum to continue the microcosmic orbit.

At the top of the back, between the shoulder blades, divide the energy again and lead it under the armpits and down the inside of the arms. Focus for a moment on the inside of the elbows, then go to the middle of the palms and feel another pair of minor chakras there. Let some of the energy flow out the tips of the fingers and then move up the outside of the arms to reconnect with the microcosmic orbit where you left it.

When you can readily feel the circulating energy and the various centres on its path, you can sometimes use a short form to energize and centre yourself quickly. Imagine the energy entering the top of the head and rapidly jumping down from one centre to the next to the base of the spine. Then with one inhalation lift the energy to the top of the head and with the next exhalation drop it again down the front to the base.

Continue to circulate it in this way until you feel energized. You can, of course, switch the breathing cycle and lift the energy up the spine during exhalation if that suits you better. In a similar way, you can sometimes lead the energy very rapidly around the macrocosmic orbit with one long exhalation down the legs or arms and with the following slow inhalation back up again. In all this energy work. imagine and feel the energy moving below the surface of the skin. (For detailed instructions on exercises to accumulate prana and use it for healing, see: www.universal-tao.com and also books by Mantak Chia, such as “Awaken Healing Light of the Tao”.)

Energizing the Body: In another method, imagine your spine to be a hollow tube. While inhaling, imagine the prana flowing up the tube to the top, and during the subsequent exhalation, let the energy again drop down the spinal tube to the base. Until you are well advanced in energy work, do not do more than ten such breathing cycles at any one time.

With all energy exercises involving the spine, try to keep it straight, either sitting up straight (but relaxed) or lying stretched out on the back. However, you can practice these and other energy exercises anywhere, anytime, such as when driving, waiting somewhere, walking, or doing housework.

Another recommended way of cultivating the inner energies is through slow movements while focusing on the sensations that the awakened energy flows may produce. This is done with exercises such as Tai Chi. However, you may not need to learn the structured Tai Chi sequence and can experiment with your own slow, improvised movements. Widespread, especially in China, is the practice of qigong. You draw energy into the body with a combination of breathing, movements, and guided imagery. This is especially effective if done in an energized outside area, such as standing barefoot on grass, in a park, at the beach, or in the mountains. There are many books and websites giving instructions in qigong

A relatively easy way of learning to feel bioenergy is to shake your body, especially the arms and hands, rapidly for about a minute. At the same time, breathe deep and fast and then suddenly stop. Remain completely still with the arms slightly raised in front of your abdomen or chest as if holding a large beach ball. Concentrate on the sensation in your hands and arms. Then start slowly moving your arms and your body. Try to intensify whatever energy sensation you feel. When you can feel the energy strongly, you can mentally lead it to an area of the body where you have a particular health problem. Alternatively, you can radiate the energy out of the palms of your hands to where you want to energize yourself. It is good to do this as a regular daily exercise.

Finally, you can directly energize a weak or diseased organ or part of the body by feeling or breathing into it. If you keep your attention focused on a particular area of your body, then you can feel it becoming warm and possibly tingling after some time. You can then use imagery to intensify the warmth or tingling. Also, imagine that your breath or the prana from your breath flows into this area and accelerates its healing. Even without a specific problem, you can focus on one organ after another and in this way energize the whole body. Once you can easily feel the warmth in a chosen target area, you can develop a sequence by energizing first the chakras, starting with the crown chakra, then circulate the energy a few times, and finally energize individual organs or areas of the body. Also, during your daily activities try to remain aware of the energies in various parts of your body.

For further information on chakras and energy work, see the Love Cure (Step 61). The best and safest way to develop the chakra system is by living appropriately in all life situations with a strong flow of a variety of positive feelings. Most healthy individuals have their three lower chakras reasonably well developed. Therefore we need to concentrate mainly on developing the four higher chakras.

* * * * * * * * * * * *

Step 53
ENERGY ACCUMULATORS
Use orgone accumulators, pyramids, cones, or water to concentrate bioenergy.

Various shapes such as pyramids and cones have been shown to concentrate bioenergy or prana. Some healers have patients sit inside small-scale pyramids with good results. I have experimented clinically with Reichian orgone accumulators, small pyramids, and cones.

Bioenergy can be generated not only by certain shapes, such as pyramids and cones, and materials such as crystals and precious metals, but also by metals and non-metals arranged in alternating layers. This is the principle of the orgone accumulator invented by Wilhelm Reich, M.D., in the 1940's. He built large boxes for patients to sit in or smaller boxes, called shooters, from which the energy was directed through a flexible hollow pipe to the body of the patient.

In one instance, I connected a wire screen placed over the tumor area of a patient to the top of a two-foot-high pyramid. After only one treatment, the patient developed a high fever during the night, which is excellent for destroying tumours. In another experiment, the energy radiated from a cone over a short distance produced a strong reddening of the skin as from sunburn, showing that these energies are real.

Even water can be effective. On tumours and areas of infection and inflammation, you can use an energy suction tube to draw off excess energy. Place one end of a metal tube, about two to five feet long, in a bucket of cold water or another large volume of water. Lightly press the upper end against the area of overactivity and keep it in this position for ten to 30 minutes, without holding the tube with the hands. Do this daily until the condition is normalized.

My understanding of the principle of energy accumulation is as follows. Our planet is enveloped in a deep layer of prana (bioenergy) generated by the sun. This prana envelope has been observed to move from west to east most of the time. When this prana flow encounters a suitable material shape, it forms a vortex around and within this obstruction, similar to flowing water or air forming vortexes under similar conditions. This increases the prana concentration within the structure. In Reichian orgone accumulators, the prana flow is being condensed by the specific arrangement of the materials.

The Orgone Accumulator: Orgone is the term for bioenergy coined by Dr. Reich, and the accumulator he designed consists of alternating layers of metal and non-metal (e.g. natural fibre). Commonly, three alternating layers each of metal and non-metal are used; originally, sheet iron and plywood were used, but now more commonly, steel wool and sheep wool or other fibre are used. The outermost layer must be a non-metal and the innermost one a metal. The basic principle is that the organic or non-metal layer tends to attract and hold the bioenergy, while the metal layer radiates or reflects it. This causes the energy to drift from the outside of the box towards the centre, which is where you sit.

As an objective measurement of the higher energy charge inside the accumulator, there is a measurably higher temperature inside the box as compared to the outside. But it does not even have to be a box. You can wrap alternating layers of a metal and a non-metal in what is called a “power pack.” The easiest and least expensive way is to wrap alternating layers of aluminium foil with paper or plastic foil. However, various healers came to the conclusion over time that the energy generated with aluminium is not beneficial for our health and that it is best to use iron as a metal, either as galvanized iron or steel wool.

CAUTION: Orgone accumulators can be “poisoned” by radioactivity. Do not use if radioactive clouds from atomic tests are suspected. Do not take a wristwatch inside. Ground the accumulator with a steel pipe connected to a stream, underground water, or a bucket of water that can be renewed from time to time. Disconnect the grounding when using the accumulator. According to some orgone research sources, presently, the orgone energy in cities is so polluted that it is not safe or advisable to use orgone accumulators in these locations. Preferably use it outside city areas in a less polluted and natural environment.

If a person is to sit inside, the inside wall should be reasonably close to the body, but not touching it. A window and air vent can be put in the door. Treatment can last 30 to 60 minutes daily. Dr. Reich reported good success using this box with cancer patients.

As with pyramids, the energy can be conducted out of the accumulator and used externally on a person or for experimenting. The charge within the accumulator varies according to the bioenergy content of the surroundings; it is high on sunny days and low at night and on rainy or cloudy days.

You can also place large crystals or precious stones inside the accumulator and try to sense if that improves the quality of the energy coming out of a shooter. For more information and interesting articles on orgone energy, see the writings of Wilhelm Reich or the Public Orgonomic Research Exchange (PORE) at http://orgone.org.

Bion Packs: Dr. Reich experimented with other ways of generating orgone energy that are not susceptible to poisoning. He found that strong orgone energy could be liberated from beach sand. After heating sand to incandescence in a Bunsen burner, he plunged it into water. Within minutes, he could observe under the microscope the blue orgone “bions” (individualized units of bioenergy). When bion-water was kept in a culture medium for several weeks, the water and the darkroom became highly charged and filled with the characteristic bluish glow of orgone energy.

While working in this room in winter and without being exposed to the sun, Dr. Reich developed a deep tan all over his body and felt “extremely vigorous,” as he commented. He also used bion solutions experimentally to treat various disease symptoms; when bion solutions were injected into experimental animals, they killed or immobilized pathogenic bacteria and cancer cells, he reported.

Dr. Reich also made bion packs by boiling beach sand or clay for 15 minutes several times and then freezing it. Here’s how to do that: Fill some heavy fabric, such as a sock, with sand like a sausage, tie the open end, and boil it for 15 minutes (no microwave). Then wrap the “sausage” in additional waxed paper or plastic and freeze it solid. When using a pack for the first time, alternate the boiling and freezing several times. Before placing it over a diseased part of the body, let it sufficiently cool and drain off excess water.

Use the pack when it is as warm as you can tolerate; use additional cloth insulation, which can gradually be removed as the pack cools. It continues radiating to some degree after cooling off. Therefore, there is no time limit for how long you can keep it on, although I assume that it is more beneficial when it is warm or hot. The pack can be reactivated on subsequent days simply by boiling it again. Let the pack dry out when not in use and preferably expose it to the sun for several hours between each application.

Orgonite: This is a new and promising concept for generating bioenergy. It is a mixture of equal parts of metal particles and an organic component, such as polyester resin or beeswax. It is a development of Dr. Reich's methods, which accumulates and regenerates orgone energy. Tests have shown that orgonite does not become poisoned, but has the ability to convert the "deadly orgone" or "Dor" as Reich called it, which is produced by radioactivity and nuclear plants, into healthy bluish orgone energy.

While aluminium embedded in polyester or epoxy resin is more convenient to use for outdoor applications, for inside the house and bodywork I recommend other metals, such as copper and brass, with the metal particles preferably embedded in beeswax. It is also possible to make ceramic orgonite by mixing clay, especially polymer clay, with metal particles before firing. Crystals, precious stones, magnets, and various spirals are often added to enhance the power or healing vibrations of the orgonite. Copper pipes, crystals, magnets, spirals, and specific shapes, such as cones and pyramids, are used to point the outflow of the orgone energy in a specific direction.

You can easily build an orgone generator yourself by following the instructions on www.cloud-busters.com. You can use different versions of it to improve the life-force energy inside your house, to charge your water with life force, or to treat the body directly.

There are a variety of commercial Reichian-style energy devices available. Research the Internet using key words such as chi generators, purple plates, tachyon or tachion products, energy pyramids, and coils.

Pyramid Energy: The pyramid shape concentrates bioenergy which can be used for healing or recharging by sitting or lying inside a pyramid. The strongest force is one-third up from the base. There appears to be an energy vortex spiralling upwards to the top (apex) and another from the apex downwards, but spiralling in the opposite direction. Both energy streams seem to have stimulating properties, but on different levels. A sedating energy field is created in an imaginary, inverted pyramid below the base of the physical pyramid.

In the pyramid shape, the energy moves as a vortex, entering at the base and flowing out at the apex. Therefore, apex energy is used to charge and stimulate while an area below a pyramid (shaped as a mirror pyramid) sedates. Inside, the effect appears to be slightly sedating at the base and stimulating higher up.

Therapeutic pyramids are commonly small-scale replicas of the Great Pyramid of Giza in Egypt. One of the four sides of the therapeutic pyramid should face magnetic north. The measurements are 1 unit (height), 1.5708 units (base), and 1.4945 units (sides), measured along the midline, not along the edges. The sides form an angle of 51° 51' to the base. However, research indicates that many different pyramid-like and cone-like shapes such as a three-sided pyramid have the potential to accumulate energies.

Avoid metal in the pyramid's construction, especially sheet metal or wire mesh (except for copper or precious metal sheeting, especially at the apex). The walls do not appear to be essential, which means you need only construct a wooden or plastic frame. However, a pyramid-shaped capstone at the apex and a bottom plate will considerably increase the energy generation. The capstone can be in the form of copper sheeting of the top ten percent of the pyramid or a small copper pyramid as the apex. Keep unnecessary metal (e.g., watches) out of the pyramid when using it and place the pyramid away from strong electrical fields. One side of the pyramid should point towards magnetic north.

For easy shifting and storage, you can build two sides of an open frame and hook them together at the base with two base-length pieces of timber, or you can build one or two small cardboard pyramids. You can then place one -pyramid under your bed or (wooden) chair, the apex pointing to the base of the spine or any part of the body that you want to strengthen. Then suspend the other pyramid from the ceiling with the base over a diseased area or over the head.

The most effective way to use a pyramid shape therapeutically may be to have a pyramid outside in a dry, sunny location and to lead the energy inside the house through an insulated electric wire coming through the apex. At the end of this, attach a wire screen covered with a pillowcase and sit on it or keep your feet on it or place it where the body needs healing. The energy generated by the pyramid varies according to the strength of the prana field in the local environment. It is generally high in sunshine, elevated positions, and amidst green surroundings, and low at night, when it rains, and inside modern well-sealed houses.

Generally, therapeutic work with pyramid energies is still in the experimental stage and you should approach these energies carefully and with self-observation. Start experimenting with small pyramids before building a big one. For treatment inside the pyramid, the direction in which the head points when lying down seems to be important, and the best position may differ for each person, time, and condition. Become sensitive, and experiment.

Energy Cones: Cones seem to have properties similar to those of pyramids. You can experiment with a metal cone or a cardboard cone covered on both sides with aluminium foil. The cone can be one to two feet high, and the wide end is best sealed. The most important requirement is a right angle formed by the sides at the apex. If the angle is wider, the energy of the cone will be focused, as with a lens, while a more acute angle seems to disperse it. A beam of energy with apparently stimulating properties comes out of the tip of the cone and can be projected over a long distance. The energy can be led off the apex by a wire (as described above for the pyramid). A stronger force is generated if the cone is more than three feet above the ground.

Covering the head with an open cone or a pyramid-shaped hat can be done to use the sedating force below these forms during meditation or to increase psychic awareness. Perhaps this is the reason why sorcerers and magicians are often depicted with cone-shaped hats. However, when trying to use cones for health improvement, it is preferable to use galvanized iron rather than aluminium, due to a possibly detrimental health effect of aluminium-conducted energies, as noted above. You can, for instance, just spray a cardboard cone with copper paint.

Crystals, Gems, and Metals: Crystals and precious stones can be worn for their beneficial radiation. Crystals should be naturally grown, not manufactured, and their best effects are achieved if worn directly on the skin. To regenerate the healing power of a gem, keep it in a silk bag when it is not worn. Use muscle testing, a pendulum, or other psychic means to assess which crystal, stone, or metal suits you, where to wear it, and for how long. Alternatively, make your selection according to the rules of color therapy.

Some helpful recommendations by Edgar Cayce are as follows:

- For attunement to healing: amethyst, bloodstone, chrysolite, or pearl
- Attunement to infinity: diamond, lapis lazuli, or ruby
- Greater stamina: agate and amethyst
- Meditation and psychic abilities: lapis linguis
- Strengthening mucous membranes: carbon steel worn close to groin
- Strengthening the body: lapis lazuli, bloodstone, moonstone, pearl, and topaz

Gems have also been used for healing in the form of powders and irradiated solutions. Gem therapy, as this healing method is known, is still very much at an experimental stage. For experimentation with solutions, you can use the following guidelines.

Select a suitable gem or precious metal by muscle testing while holding the investigated item or use psychic testing. Write the name of the gem to be tested on a slip of paper and hold it near your navel while simultaneously concentrating mentally on this name; then do a muscle test. Alternatively, if you can work reliably with a pendulum, use that.

Try to obtain the gem or metal that is most strengthening. Immerse it in a shallow bowl of unpolluted water, cover with light, fine gauze as an insect and dust screen, and expose this to sun and moonlight for several days; then pour the solution into a bottle.

Take a sip once a day for up to six months. Use close self-observation and/or muscle testing and stop if any unwanted symptoms arise. If in doubt, it may be safer to take a 3x or 6x homeopathic dilution of the irradiated water, especially in the case of precious metals. If you lack energy, experiment with a ruby. Edgar Cayce recommended alternating the radiation of gold and silver because gold improves the circulation and silver the nervous system. According to Cayce, they may together almost double the present life span. Use pure gold and silver (preferably at least 99 percent pure).

* * * * * * * * * * * *

Step 54
ENERGY HEALING

Give energy or remove excess energy to facilitate healing yourself and others.

The term “energy healing” refers to the transfer of energy from a healer to a patient and to the removal of negative unhealthy energies from the patient. I imagine negative energy to be dense and stagnating, somewhat like cold, damp air. Its effect on an unprotected healer can be noticeable; she may develop the pains or other disease symptoms that have just been removed from the patient. Energy healing is also called pranic healing, magnetic healing, or laying on of hands; yet another form, using special symbols, is called Reiki. While almost anyone can practice energy healing from a distance, its use when actually touching the patient should be restricted to those practitioners who already feel in good health and have an abundant supply of energy. Here’s a technique:

Slowly move your more sensitive hand, usually the left hand in right-handed individuals, a few inches from and over the patient’s body to try sensing unusual energy spots. This can be done on a fully-clothed person. Over a diseased area, your hand might suddenly feel warm or tingling. After completing the check, you can return to the area over which you felt the strongest sensation. Now you can either give energy to strengthen an organ or the whole person, or you can draw out energy from the body as in the case of pain and inflammation (described below).

With all energy work where you touch a patient or remove energy, try to protect yourself with your intention or expectation of being immune. In addition, use guided imagery, such as visualizing a protective white light or shield around you, or with your imagination allow the patient’s energy to flow through your body and out again through your feet or through one of your hands.

Giving Energy: Before and after any contact healing, such as just described, practice energy accumulation (with breathing exercises, meditation, or guided imagery) and preferably use a prayer or affirmation. Then direct the energy into your hands and feel them becoming warm and tingling. Place your hands on or over the patient where the body requires healing, especially strengthening, and feel the energy streaming out; or hold in your mind the intention of being a channel for giving healing energy, or focus your attention on your crown chakra on the top of your head.

Continue for five to ten minutes and repeat as required. After each session, rinse your hands in cold water and do some deep breathing. If you feel drained the following day, do more energy-accumulation exercises and less energy-giving in future sessions, and especially avoid or minimize touching patients.

For general strengthening, it is often beneficial to place the right hand on the patient’s navel and the left one on their forehead. For most people, the right hand has more energy-giving qualities and the left one has predominantly energy-withdrawing qualities. However some people, especially left-handed individuals and sometimes women, have the opposite polarity. You can practice energy-giving in a group situation such that each one in turn is the “patient.”

Another method is to place the positive or energy-giving hand over the weak body area and hold the left hand over the person's head with your palm facing upwards, the fingers lightly curved and the heel of the hand preferably pointing towards the north.

You can be healer and patient at the same time by radiating energy from your hands into your own body. You can do this wherever you have a weakness or want to improve conditions; this is explained below under Beaming. Move your energized hands slowly from the top of your head over your face and down the sides and front of your body; smile inwardly while doing this.

Removing Excess Energy: Excess energy can be associated with an overactive organ or a painful area of the body, but also with unreleased and densified emotions in the form of stagnating, depressed energy. This depressed energy is often stored around the shoulders, neck, and head. Commonly, drawing out excess energy is more helpful than giving energy. The reason for this is that energy needs to flow to be useful and drawing it out improves the energy circulation in a way that's similar to walking barefoot on fresh moist grass.

Energy can be directly removed from painful or overactive areas with cupped hands and then flicked into a corner. Afterwards, seal the person's aura with smoothing, sweeping strokes downward along the body area from which the energy has been removed. At the arms or legs, make the sweeps towards the fingers or toes.

For a general energy cleansing, the patient can be seated while the healer moves his cupped hands slowly upwards close to the person's body, starting just below the level of the seat. This should be done from all sides, lifting the excess energy off the top of the person's head and flicking it off into a corner, out of a window, or into a sink. At the end of the cleansing, seal the person's aura.

The healer must have the intention and expectation of removing excess energy. With enough sensitivity, you can actually feel depressed energy as a heaviness in your cupped hands. Try rapid or deep breathing during this work with the patient. These whole-body cleansings are especially effective if done by a group of healers at the same time, on one person.

Another drawing-out method is as follows: Place your receptive or sensitive hand with the palm down a few inches above the painful or diseased area. Hold your positive or energy-giving hand, also with the palm down, over a bucket of fresh water. Treat for five to ten minutes, or less if pain disappears sooner. Then move to another part of the body if required and treat again for five to ten minutes. You may be able to increase the effectiveness of this procedure by having the fingers of the positive hand point south.

Instead of using your hands, you can draw out excess energy with a suction tube: Solder approximately five feet of an insulated wire to one end of a one-to-two-foot iron tube. Remove the insulation from the free end of the wire and spread the strands of wire in a large container of cold water. With the free end of the tube, trace a meridian for sedation, or place it over a painful area.

This implement produces a powerful suction for removing excess energy. Use it repeatedly for five to 15 minutes. However, avoid draining energy from the hand by holding the tube when in operation. Preferably fasten it to a stand in the desired position, otherwise wrap it well with a dry cloth before holding. An even stronger suction is achieved by using a three foot long galvanized-iron or stainless-steel pipe in a fixed position without attached wire; one end reaches into a water container, while the free end is directed against a treated area.

Healing at a Distance: As a possibility for distant healing in which you do not touch the patient, start by audibly or silently saying a prayer in which you ask that your patient be healed. Then have the patient lie down or sit in a relaxed position with closed eyes while you enter a meditative state. It will be helpful to use eight-two breathing or a similar slow-breathing technique (see Step 9). Then mentally radiate love and healing energy to the patient, or visualize the patient as being healed.

Very effective is feeling a strong flow of energy entering the top of your head and moving down your arms and out of your palms. Hold your palms a foot or two away from the patient. Alternatively, you can beam the energy out of your forehead or heart. Another possibility is to evaluate in which part of the body the source of the problem is located. Then energize your own chakra related to this part of the body and let this energy radiate into the corresponding chakra of the patient. Finish with a thanksgiving prayer and a blessing for the patient. Distant healing is much safer for the healer than that done through direct touch.

With distant healing, the patient is in the same room, while with absent healing the patient can be anywhere in the world. For absent healing, use a similar procedure: Visualize the face of the patient or concentrate your healing energy on a photo of the patient. You can also pray together at a prearranged time.

Beaming: Beaming is an excellent cooperative method for improving our health and speeding up our spiritual development. It helps us to release negative mental, emotional, and etheric energies while activating our energy centres and circuits at the same time. Beaming can be done alone but is more effective with a friend or partner or in a group setting.

Before starting, ask your God Presence or Higher Self to make you a channel for the healing light. (In the channelled information revealed in the book I Will Arise, by Paola Hugh, it is suggested to “ask in the name of Jesus Christ to become a channel for the healing Light of the Father’s Love.”) Then imagine a spiritual white light entering the top of your head and flowing down your arms. You can use the chakras in the palms of your hands like lenses to focus this healing light in concentrated form onto areas of the body in need of healing and cleansing. Have the mental attitude that this light originates from a higher source.

Generally, hold your hands eight to 12 inches from the body and point the palms towards the organ to be healed. If you feel warmth or tingling in your palms, this means that there is an obstruction in the person’s body; continue beaming until these disappear, then move to another area. Initially beam for ten to 20 minutes, but with growing experience go longer. The farther away from the body you hold your palms, the deeper the beamed light energy penetrates.

For overall health improvement and spiritual cleansing, start with the person’s head. Beam from all sides, but mostly with one hand over the top and the other towards the base of the brain from the back. After several sessions of beaming the head, move to the kidneys and the intestines, one palm directed towards each kidney.

At another session, treat the spleen with one hand and the liver with the other. Jointly, spleen and liver are responsible for the purification of the blood. Do the same with the heart and the throat or the thyroid. If helping others, beam their kidneys, spleen, liver, and heart from the back. On yourself, you can treat your front or back depending on your ability to beam your own back.

To establish inner energy circuits, beam simultaneously on the heart and the back of the head, on both adrenal glands (located on top of the kidneys), on one adrenal and the thyroid, then on the other adrenal and the thyroid.

Another practice is beaming light of different colours from the thumb pads into the open eyes. Use a distance of two inches and about one minute for each imaginary colour. Start by visualizing the cosmic violet colour of transmutation. See it streaming out of the thumb tips and through the eyes into the centre of intellect at the base of the brain. Then visualize other colours and body areas. Beam emerald green into the eyes for the adrenals, grass green for the heart, pale blue for the reproductive organs and for the thyroid, light electric-blue for the liver, indigo for the spleen, deep indigo-blue for the ears, and orange for the lymphatic system. Violet is for the teeth and gums, and you beam this into the open mouth.

* * * * * * * * * * * *

Step 55

MERIDIAN & ACUPRESSURE THERAPY

Balance the flow of energies along your acupuncture meridians by tracing their circuits.

The human body has a bioenergy circulation system similar to the blood circulation system, as shown in figure 7-4. The bioenergy flows along 12 main meridians, or channels. These are on each side of the body, each pair being related to a specific organ and its processes. In addition, there are two single meridians following the midline of the body, one at the front and the other along the spine. Many points are distributed along these 14 meridians. These points are essentially transformer stations through which bioenergy flows into the muscle structure to allow organized activity to take place; they are treatment points for acupuncture. In disease, energy imbalances develop in these channels, causing some of the points to become irritated or congested. This results in pain or weakness in the surrounding muscles.

Try to find tender points by pressing with a thumb or finger in a circular motion around the area indicated for each point on the charts of the meridians (see figure 7-4). You do not need to know and find the exact location of a point. If there is a problem, all of a small muscle and much of a large muscle will be tender around the point so that it is easy to find. Press into the general area of the selected point and gradually work your way toward the centre of the pain. For treatment, concentrate on the most painful spots, pressing them frequently as explained for reflexology in Part 2. Generally, the pain decreases after some time of pressing, but should it increase instead, then interrupt the treatment for several days.

Tender pressure points may stem from either a local condition or a problem in the main organ associated with that particular meridian; also, the pain may be referred from a more distant body part connected to the same meridian. There is, of course, an interaction between these different areas.

When the main organ is inflamed, most points along its meridian will be tender. The points especially useful for diagnosing and treating the organ directly are located around the ankles and wrists. They are called source-points; further source-points are situated along the Bladder meridian close to the spine, and also as local-points on various meridians near the target organ.

Rules for Selecting Treatment Points: The following general rules are useful for selecting points for treatment.

1. Points above the knees or elbows reflect mainly local conditions, while those below these joints reflect local as well as distant conditions at other parts of the meridian.
2. Generally, the closer the points are to the fingertips or toes, the closer to the other end of the meridian will be the related sphere of influence.
3. For painful, acute conditions, treat mainly -distant-points; for chronic conditions, add local-points.

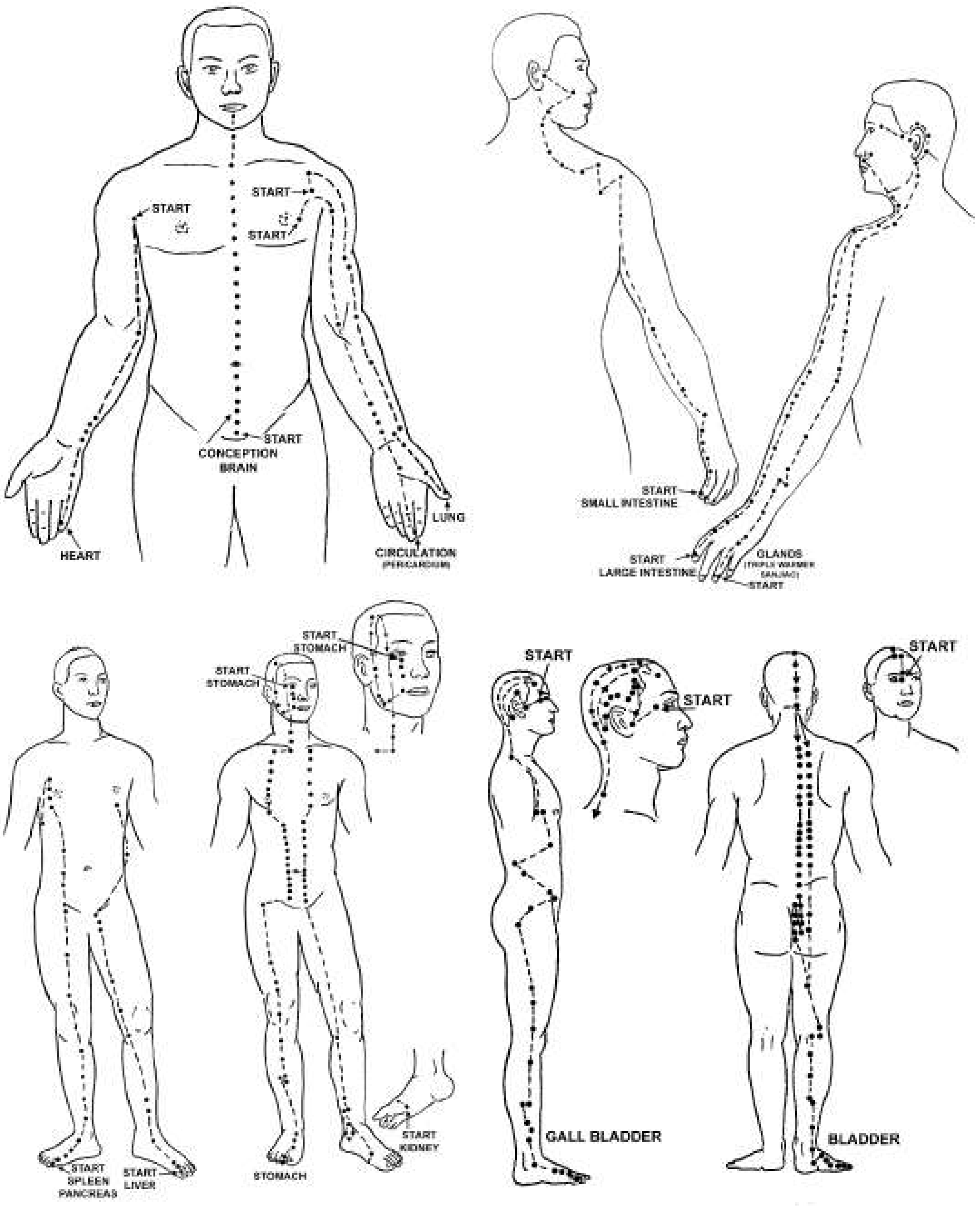
START
START
START
START
CONCEPTION
BRAIN
HEART
LUNG
CIRCULATION
(PERICARDIUM)
START
SMALL INTESTINE
START
LARGE INTESTINE
GLANDS
(TRIPLE WARMER
SANJIAO)
START
START
STOMACH
START
STOMACH
START
SPLEEN
PANCREAS
START
LIVER
STOMACH
START
KIDNEY
START
START
GALL BLADDER
START
BLADDER

Figure 7-4: Chart of Acupuncture Meridians. Begin tracing the points marked "start". Dots indicate acupuncture or pressure points. Apply finger pressure when tender. The Governing Meridian (not shown) starts at the tailbone and follows the spine in a straight line over the top of the head down the nose and ends between nose and upper lip.

This means, for example, that if you wish to treat an acute eye or ear pain, select points near the toes and fingers, while to treat shoulders and hips, move close to the elbows or knees. For a more chronic condition, you press in addition the tender spots around the ears and eyes, and hips and shoulders. You can also add points of other meridians that have a direct relationship to the problem. Eye problems, for example, are often associated with poor kidney and liver function. Therefore, some points on the meridian associated with these organs can be treated as well. Further, the points at the shoulders, the shoulder blades, and the base of the neck control the circulation to the arms and hands, while the points at the hips, the buttocks, and around the base of the spine control the circulation to the legs and feet. Therefore, treat poor circulation, numbness, and pain in the extremities by pressing the points in the indicated control areas.

Frequently, the pressure points of a pair of meridians will be more sensitive on one side of the body than on the other side. This indicates an imbalance in the energy flows. The meridian with the more tender points requires sedation, relaxation, and withdrawal of energy, while its counterpart on the other side needs stimulation. Instead of pressing acupuncture points, you can place the north pole of a small magnet on a point for stimulation or the south pole for sedation.

Meridian Therapy: A main problem in healing is the balancing of the energy flows. Usually, this balancing occurs unrecognized as a by-product of other successful treatments. However, healing can be initiated and accelerated by deliberately treating the meridians with the express purpose of balancing.

A useful method for this is to follow the outline of a meridian with one finger or with several fingers held closely together. Following a meridian in the normal direction of flow is strengthening, while tracing in the opposite direction will weaken and sedate the meridian. The results of these tracings can be checked with muscle testing. Preferably energize your hands before a series of meridian tracements. Do this by vigorously shaking your hands for a few seconds or imagining energy streaming out of your fingers, or both.

In order to perform a tracement, pause with your fingers for a moment at the starting point; then follow the meridian in a fluid motion. The fingers should be close to the body but do not need to touch the skin. It is enough to remain within about two inches of the actual course of the meridian. Related meridians on both sides of the body can be traced simultaneously. In repeated tracements, move the hands back to the starting point in a wide sweep to avoid following the meridian in the reverse direction.

You can trace all the main meridians once or several times daily, or concentrate on the meridians most in need of improvement. Important meridians can be traced repeatedly during the day for 20 times or more. The strengthening effect may be increased if, after several tracements, you do a muscle-tensing exercise. Take a deep breath while tensing your whole body and then relax it during exhalation.

The meridian system allows us to understand why many disease symptoms appear in certain body parts when the originating cause is a weak or inflamed organ in another location. The diagrams (see figure 7-4) show only the surface courses of the meridians; inside the body, each one is connected to its associated organ. In this way, the deranged energy pattern of an organ is easily transmitted to other body parts along the course of the meridian.

Diseases of the ear, migraine, and hip problems can now be seen in terms of energy associated with the gallbladder. Arthritis of the hip, therefore, is often a gallbladder problem, while arthritis of the shoulders may have its cause in the intestines. Gout, affecting the big toes, stems from the liver, while swollen ankles are related to the kidneys. See table 7-2 for the relationship of different diseases or health problems to the meridian system.

Table 7-2: Pathological Symptoms of the Meridians per 24-hour Cycle

Meridian with time of maximum daily activity	***Associated symptoms***
Liver, 1-3 a.m.	liver problems, lumbago, vomiting, hernia, urination problems, pain in the lower abdomen and along the course of the meridian
Lung, 3-5 a.m.	respiratory diseases, sore throat, cough, common cold, pain in the shoulder and along the meridian
Large Intestine, 5-7 a.m.	abdominal pain, constipation, diarrhea, sore throat, toothache in the lower gum, nasal discharge and bleeding, pain along the meridian
Stomach, 7-9 a.m.	stomach problems, abdominal pain and distension, edema, vomiting, facial paralysis, sore throat, toothache in the upper gum, nose bleeding, pain along the meridian
Spleen, 9-11 a.m.	problems of the spleen and pancreas, abdominal distension, jaundice, general weakness and sluggishness, tongue problems, vomiting, pain and swellings along the course of the meridian
Heart, 11 am-1 p.m.	-heart problems, dryness of the throat, jaundice, pain along the meridian
Small Intestine, 13 p.m.	pain in the lower abdomen, sore throat, swelling or paralysis of face, deafness, pain along the meridian
Bladder, 3-5 p.m.	bladder, neck, and back problems, headache, eye diseases, pain along the back of the leg
Kidney, 5-7 p.m.	kidney problems, lung problems, dry tongue, lumbago, edema, diarrhea, constipation, pain and weakness along the course of the meridian
Circulation, 7-9 p.m.	poor circulation, angina, palpitation, diseases of the sexual glands and organs, irritability, pain along the course of the meridian
Gland Meridian (Triple Warmer), 9-11p.m.	diseases of the thyroid and adrenal glands, ear problems, sore throat, abdominal distension, edema, swelling of cheek, pain along the meridian
Conception Meridian	diseases of the urogenital system, hernia, cough, breathing difficulties, breast problems
Gallbladder, 11 p.m.-1 a.m.	gallbladder problems, ear diseases, migraine, hip problems, dizziness, pain along the meridian
Governing Meridian	spinal problems, mental disorders, fever, nose problems, headaches

Each organ-meridian system, except for the conception and governing meridians, has a two-hour period of maximum activity and a period of minimum activity 12 hours later over the course of the day. An inflamed organ is often more painful during its active period, while a weak system should not be stressed during low periods.

While pulse diagnosis (read at the wrist) is the expert way of determining the energy balance of the meridians, you often can judge this by the kind of diseases or symptoms present. Generally, overacidity, pain, and inflammation along the course of a meridian or in its target organ require sedation, while coldness and weakness call for stimulation.

According to Chinese medicine, some interesting characteristics are associated with the meridians of the kidneys and liver. Weak kidneys are associated with a lack of will and sexual impulse, and with negativity, unease, timidity, and impatience. In the case of overactive kidneys, the energy level is abnormally high and you cannot stop working. It is similar with the liver. Weakness shows itself in dizziness, a clumsy walk, easily tired eyes, and a short temper, while an overactive liver causes excessive excitability, crying moods, and a compulsion to continue working.

Chinese medicine says that the main organs and their meridians are also associated with different emotions. An excessive amount of these emotions can damage the associated organs, while a disease of these organs may, in time, lead to excessive displays of these emotions. Accordingly, anger is associated with the liver; excessive laughter and gayness, but also fright, with the heart; worry and emotional tension with the spleen and pancreas; grief, sorrow, and negativity with the lungs; and fear and timidity with the kidneys.

Magnet Meridian Therapy: Magnets can be used for tracing meridians. Move the appropriate pole of a magnet close to the skin along a meridian. This can have a much stronger balancing effect than using the fingers. For strengthening use the north pole, for sedating the south pole.

With muscle testing, you can often detect a strong preference of a meridian for one of the magnetic poles. This can be used to diagnose the condition of the associated organ (if weak or inflamed). If muscle testing is not possible or successful, use the south pole on the body side that shows more inflammation or more tender acupressure points; trace the opposite meridian with the north pole.

Some individuals feel immediately which pole and direction of tracing is beneficial, while the opposite pole and direction may feel unpleasant. For a sedating and pain-relieving effect, it can be very effective to trace the south pole of a strong magnet against the meridian flow. Use a round magnet, one or more combined button magnets, or a small elongated ceramic magnet.

If muscle testing shows that tracing the opposite flow direction of a meridian with the south pole towards the skin is strengthening, there may be an infection or inflammation somewhere along the course of the meridian or at its associated organ. Weak meridians or their related organs, on the other hand, test stronger when tracing is done with the north pole in the correct flow direction. If the tracing is stronger with the south pole in the correct flow direction, but weak when traced in the opposite direction, there is only a moderate amount of overactivity in the meridian or its related organ.

The tracing of meridians associated with painful conditions 20 to 50 times with a strong magnet has sometimes provided almost immediate relief from pain, while similarly quick results have sometimes been achieved using the north pole for improving the mobility of impaired limbs.

Instead of using magnets, meridian tracement can be done with energies from a small pyramid or orgone accumulator. Use a metal tube, preferably galvanized iron or stainless steel wrapped in dry cloth, for the actual tracement; this can be connected to a pyramid by an insulated wire. For sedation, try a metal tube (copper, galvanized iron, or steel) from which an insulated wire leads into a container of water.

Mental Tracing of Energy Channels: Often, effective meridian therapy can be achieved simply by tracing a meridian mentally, moving your awareness along it. Try to feel a warm or orange energy stream moving in the normal flow direction for stimulation; for sedation, feel a cool or blue stream moving either in the normal direction or against it. If in doubt, move in the normal direction and imagine the meridian being traced with the correct pole of a strong magnet. You may not actually need to know which pole it is.

Meridian therapy is especially important where scars cross a meridian line. Untreated scars tend to restrict the energy flow. With meridian therapy, I have seen on some occasions pain, persisting for days or weeks after an operation, disappear within minutes and stiff or painful shoulders became normal. Even scoliosis (sideways curvature of the spine) has been corrected in a short time in this way with the addition of dietary improvement and pressure applied to tense muscles. Scoliosis can also be successfully treated with color therapy and magnet therapy using stimulation (orange color or north pole) on the weak side of the back and sedation therapy (blue and south pole) on the tense side towards which the spine is being pulled.

While prolonged pressure applied to a tense muscle is usually the quickest way to disperse pain or stiffness, it is also painful, so it helps greatly to trace the nearest meridian for several minutes with a magnetic south pole beforehand, possibly against the meridian flow. Refer to table 7-2, which shows the health problems that can be associated with the different organ meridians. Problems due to overactivity, such as inflammation, are commonly most pronounced during the time of maximum activity, while underactivity or weakness may become more normal at that time.

* * * * * * * * * * * *

Step 56

SEXUALITY

Channel your sexual energies into pleasure rather than frustration.

Strong emotions and energy flows are an integral part of sex, which is why sex has the potential to cause enormous problems if these emotions and energies are blocked or misdirected. Due to our social conditions, unimpeded sexuality is rare and sexual repression is the norm, so much so that most of us do not even realize that something is wrong. We do not know what natural, free-flowing sexuality is. I see two main factors that led to this situation in our society. One is our competitive, ego-centred, and male-dominated social structure and value system; the other is the cumulative result of our restrictive religious teachings and social traditions.

The Natural Development of Sexuality: Sexuality is much more than having sex. It involves a wide range of feelings and emotions at all ages. In the foetus, sexuality is experienced as total security, the reassuring heartbeat of the mother, the gentle rocking motion of its fluid bed, and the sucking of its thumb. For the baby, sexuality means lying against a warm, soft body, feeling its love and auric energies, gently touching and suckling at a nourishing nipple. The growing infant retains this need for loving touch and gentle body contact. Later, exploring one's own genitals and using them to generate pleasurable feelings is as natural and important as exploring aspects of the environment. Of special importance is an intimate relationship, involving bodily contact with the parent of the opposite sex or with his or her substitute.

With puberty, sexuality bursts into the open and produces powerful longings. The awakening sexual feelings need an outlet. This is provided for the girl through romantic ideals, hero adoration, and in a practical way involving religious and charitable service. Sports and nature activities provide an outlet for both sexes.

Teenage boys long for adventure and need to develop and test their abilities by following or imitating a chosen hero or role model. These romanticized adventure activities channel the sexual energies into safe outlets conducive to inner and outer growth and social development without suppressing sexual feelings or activities. Ideally, the relationship of children and teenagers with their parents should be such that all sex-related subjects can be freely discussed and possibly even practical instructions given by the parent of the same sex.

Romantic relationships with the opposite sex remain for many years just that, romantic. This is the flowering time of erotic love with ever-increasing levels of intimacy, but with penetrative sex ideally happening only in the later teens within the context of a loving relationship. The challenge now is to transform the straw fire of erotic love into mature lasting love for the partner. Erotic love is based on sexual attraction, while mature love is a communion in spirit. This transformation of love will allow the sexual relationship to mature more fully.

To recapitulate, we can say that during pregnancy and birth the most important aspect for the foetus is being wanted. During babyhood, breastfeeding and body contact are central. The growing child needs to be immersed in a loving and harmonious relationship with the whole family unit with much touching and experiencing of feelings. Then comes the time of romantic idealism for the teenager and finally the challenge to transform Eros into mature love and sex. This is the ideal of our normal and natural sexual evolution.

The sexual reality for most of us is so different from this ideal that it has left an important part of our sex-related emotional development greatly impaired. Instead of this gradual, guided growth into a sexual relationship, we are expected to be asexual, with severe repression of sex-related feelings and impulses, until we are grown up and then overnight, without a suitable role model, become responsible and accomplished sexual partners. That is just not possible.

When Things Go Wrong: The importance of several of these factors for normal sexual development has been demonstrated in an interesting experiment with monkeys. Immedi-ately after birth they were separated from their mothers and raised with two substitute “mothers.” One “mother” was a wire structure containing a food bottle with nipple and the other had terry cloth over the wire and was heated by an electric bulb inside. The young monkeys adopted the cloth-covered structure as their mother substitute and were distressed when it was temporarily removed. The bottle structure was just a feeding machine to them. All these monkeys became disturbed emotionally as well as in their behaviour when growing up, and they were unable to perform the sexual act successfully on maturity. Touch was emotionally more important to them than food.

Most of us became blocked in our sexual-emotional development at several stages of our upbringing. One of the worst aspects for a baby is the feeling of not being wanted and especially if it experiences hostility from one or both parents. Regression shows that a baby or even a foetus knows the attitude of its parents and that this remains a lifelong part of the subconscious programming of that individual.

A difficult birth, especially with forceps and glaring lights, is traumatic for the infant; so is surgery and a negative attitude by those present at the birth. Even worse is the prolonged separation of the baby from the mother as practiced in most hospitals. More than anything else, babies need persistent contact with the mother or a suitable substitute for emotional security.

It is so sad that many parents take great pride in a fanciful nursery room when the baby only longs for contact with its mother. This is perceived by the baby as a rejection, being left alone with feelings of insecurity for the rest of its life. Such insecurities surface again later, in a sexual relationship. The more secure infants feel, the sooner they will venture off on their own to explore the world around them and the more secure adults they will become.

Further problems frequently arise in the “anal” phase of development, as Freud called it, with too early or too severe toilet training. Instead of acquiring voluntary control of the sphincter muscle, children permanently contract the gluteal muscles and those of the pelvic floor. In later life, this can lessen their sexual energies and feelings.

Another conflict often arises between the ages of two and six, when parents typically suppress any manifestations of infantile sexuality. These can be in the form of touching, masturbating, or displaying the genitals, or erotic contact with the parent of the opposite sex. Such body contact is pleasurable for the child and necessary for identification with its body and development of pleasurable erotic feelings for the opposite sex. Problems arise when the parent either rejects close body contact or manipulates the child for his or her own sexual desires.

Psychologists speak of an Oedipal situation at the age of five or six, when the child must choose between its sexual feelings and retaining the love of the parents by being a “good” child. Most children choose to be good and suppress their sexuality. This then leads to frustration with the opposite sex in later life. Sexual feelings, like other feelings, are suppressed by tensing the muscles, especially in the pelvic area, making the body rigid.

During our school years, examination anxieties greatly contribute to our general insecurities and make us more nervous and neurotic. As teenagers, we often lack suitable outlets for our creative energies and romantic ideals. If our overflowing sexual energy is not transformed into creative energy in the form of sports, art, or adventure, then masturbation remains as the only outlet, aside from erotic and wet dreams. This may lead to premature penetrative sex as an experiment or because of peer pressure, without feelings of love and sharing, so necessary for maturing sex.

Sexual Neurosis: The emotional problems of our parents led to most of us facing severe emotional stresses in our childhood, and now as adults we are neurotic to varying degrees. I use “neurotic” to mean that we have emotional disturbances and inhibitions that prevent us from finding fulfilment in life and specifically in our sexual relationships. Neurosis manifests in all aspects of our personality, feelings, emotions, character structure, body structure, and body functions. If we have sexual problems, we also have problems in our other social relationships. Our childhood experiences guide us in our choice of sexual and marriage partner, in the kind of sex that we like or dislike, and in what we feel and how satisfied we are in life.

Symptoms of sexual neurosis are manifold, including shyness, stuttering, difficulty urinating if someone looks, fear of nudity, premature ejaculation, impotence, frigidity or generally failure to achieve full orgasm, lack of fulfilment, nymphomania, and a tendency to rape. Neurosis also contributes to many diseases and especially to those of the female sexual organs.

As adults we usually are completely unaware that the emotional scars from our childhood are still subconsciously active and even dominating our behaviour and decision-making. Our lower self, however, remembers fully because that is what it is made of, for memories are the basis of our inner life. If we felt deprived of something important as an infant, we will continue to feel deprived as an adult. If we felt fearful of or hostile towards our mother or father, we later feel subconsciously hostile towards all men or women or are afraid of them. How can we have a normal, fulfilling sexual relationship or any other close relationship with someone we fear or hate subconsciously?

It is said that we can only fully love a man or a woman if we basically love all men or women. If we are subconsciously fearful of the opposite sex, then we develop a fear of penetration, become impotent, or become susceptible to homosexuality. Strong hostility towards the mother can later lead to an inability for a man to have sex with a woman unless he humiliates her, possibly in the form of violent rape.

Pornography is a fantasy outlet for many forms of sexual neurosis, but it is only a poor substitute for uninhibited sexual functions.

Sexual Types: Body-oriented psychologists such as Alexander Lowen, John Pierrakos, and Barbara Brennan have devised various systems linking our childhood experiences with our sexual problems as adults. They distinguish the following sexual types: schizoid, oral, psychopathic, masochistic, and rigid. A normal type does not seem to be part of their experience.

The schizoid type: This type experienced emotional trauma already around birth and, therefore, did not fully connect with the body. When things become difficult, schizoids just withdraw from reality and live in fantasy. The resulting lack of body feeling and identification with the "self" also leads to confused sexual identification, and schizoids easily become homosexuals or prostitutes. Many passive lesbians are of the schizoid type as, like deprived children, they crave warmth and human comfort; their orgasmic potency (the ability to reach a full orgasm) is usually rather low.

The oral type: This kind is widespread because of the general abandonment of breast-feeding in our society. The baby felt abandoned and sex is sought mainly for closeness and contact, similar to the schizoid type. The breast is worshipped and oral sex enjoyed. Sexual relations as well as sexual functions remain immature. Anal sex is also related to the oral phase. Sensuality, superficial feelings that make us feel good, are used as a substitute for orgasmic sexuality, but cannot achieve a deep emotional energy release.

The psychopathic type: This kind evolved from a child that felt it had to manipulate the parent of the opposite sex to get what it wanted. In later life, the need to be taken care of is achieved either through bullying or seduction. Sex is often hostile with much fantasy. Feelings of contempt for the partner cover suppressed feelings of inferiority. Sexuality is used to control, conquer, and dominate, while pleasure is secondary. Often the behaviour in a relationship is irresponsible.

The masochistic type: This type evolved from domineering parents who crushed the child's creativity and self-respect. This lack of self-worth is expressed in a sexual relationship as a need to be humiliated. Sexual excitement can be gained from inflicted pain. The masochistic male may have difficulties with ejaculation.

The rigid type: While most neurotic individuals have various rigidities, these are especially prevalent in the rigid type. The child suppressed sexual and other tender feelings in response to a perceived rejection by the parent of the opposite sex. This was typically an emotionally cold parent who did not tenderly touch or cuddle the child. The child learned not to cry or show emotions. The rigid adult chooses sex over love and is a sex performer without deep feelings and with greatly decreased orgasmic potency. Sexual activity is compulsive and mechanical; performance provides ego satisfaction; emotional warmth and physical closeness are avoided.

Commonly, we combine features of several types in our sexual makeup, but with one type being dominant. However we resolved the Oedipus complex as a child (the relationship with the parent of the opposite sex), we assume corresponding sexual roles as an adult. We may either act as a son or daughter, as a sister or brother, as hero, sweetheart, or romantic ideal, as a father or mother figure. However, if we resolved the problem in a natural way, we act as a mature adult, encompassing all roles and playing each one as appropriate.

If the relationship of a girl with her father remains emotionally unresolved, then as an adult she tends to act mainly on the daughter level, looking for a father figure and easily becoming just a sex object in danger of drifting into prostitution. If the girl channelled her sexual feelings into sympathy for the father, she may play the role of sister and see herself as companion and helpmate of her man, wanting to share everything, but sexual contact remains relatively unimportant and unexciting.

If the sexual feelings for the father have not been suppressed as with the sister type, but just held back, the romantic type emerges, emotionally the eternal virgin, teasing but not wanting to be possessed by a man. She likes flirting, as she needs romantic excitement, but has a problem combining love with sex. Because of resulting emotional conflicts, this type is psychologically also called the “hysterical type.”

The girl who matures too early by replacing sexual feelings for her father with motherly instincts later becomes a mother type with little sexual interest. She becomes attracted to immature males of the son type whom she can mother.

Each type has an unresolved conflict: the daughter type between accepting and rejecting herself; the sister type between passivity and aggression; the romantic type between surrender and resistance; and the mother type between submission and dominance.

In the male, the conflict of the son type is between accepting and rejecting himself; he is the playboy type, immature and irresponsible. The brother type assumes the role of protector. Here the conflict is between his aggression to protect the woman and passivity or inability to satisfy his own needs. The conflict of the Casanova-like hero or knight type is between emotional surrender and resistance to the female, while the authoritative father type is torn between dominance and submission.

Orgasm: There is a popular misconception that if a man ejaculates or a woman has pleasurable feelings at the height of the sexual act, that this is an orgasm. It is not necessarily so, at least in the view of bioenergetic psychology. Usually there is just a climax with more or less of an orgasmic component.

A full orgasm can be described as follows: With climax approaching, the formerly voluntary body movements suddenly become involuntary with a strong increase in genital sensations. Just before the discharge, melting sensations are felt in the pelvis. Orgasm starts with a strong contraction around the base of the penis and the opening of the vagina. A flood of feelings rushes downwards while the whole body convulses in unison with each involuntary forward swing of the pelvis. Now the whole body is engulfed in melting and streaming sensations. Ejaculation takes place in pulsating squirts. There may be a sensation of flying or spinning or something similar at the height of the orgasm. Afterwards there is a full relaxation, with an inner glowing and a feeling of utter satisfaction and contentment.

In order to achieve such a full orgasm, there must be a sufficient buildup of sexual energy beforehand. If the body is generally low in energy, if the energy flow is blocked through armouring, or if there is insufficient foreplay or feeling awareness, then there will be only a low charge of sexual energy in the pelvis and consequently a weak discharge of energy and feelings.

A strong field of sexual energy in one partner will help the other to achieve a greater charge; if both are fully charged, they will greatly increase each other's orgasm. Con-versely, with a weakly charged partner or with masturbation, our orgasmic feelings will be diminished, usually also with oral sex, as close contact of the fully charged genital areas is required for maximal potential.

A strong sexual charge consists not only of sexual energies, but also of feeling energies. Normally, tender, loving feelings towards each other greatly help to increase our sexual charge. As an emergency measure when such tender feelings are blocked, some can achieve the necessary emotional excitement with sadistic or masochistic visual measures. Also various forms of pornography are used as a substitute, usually an inadequate one, for tender, loving feelings.

With strong, loving feelings, on the other hand, an almost irresistible desire to melt into each other can build up. The sexual charge and physical union then follow as a natural consequence. This automatically leads to a true and full orgasm. For a woman who has difficulty coming to an orgasm, it may be made easier if her partner positions himself on top somewhat higher than usual so that their navels meet. With this, the shaft of the penis will rub or press on the clitoris during movement inside the vagina.

A full orgasm is possible after overcoming bodily rigidities and emotional inhibitions, and it is a natural occurrence in societies living without sexual inhibitions.

The most common and obvious signs of only a partial orgasm are:

- A clitoral rather than a deep vaginal climax
- Feelings confined to the sexual organs rather than full body feelings of melting
- Inhibition of movements, especially of the involuntary convulsive movements
- Inhibition of spontaneous sound expressions
- Ejaculation without pulsating squirts
- Premature ejaculation or continuing sexual activity after ejaculation
- Lack of complete relaxation and satisfaction after the climax

As a general rule, we can say that the less the climax is associated with pleasurable streamings and whole body feelings and the more it is confined to the sexual area only, the less it is a real orgasm but rather just a low-energy mechanical discharge. Wilhelm Reich, M.D., was the first to describe the negative consequences of blocked sexual energies and the true nature and purpose of the orgasm as a discharge of excess bioenergy with the additional liberation of feeling energy.

Sexual Healing Exercises: Healing our sexuality means freeing up our blocked energy flows and overcoming subconscious negative conditioning by replacing it with positive feelings and attitudes about our sexuality. In order to remove the energy blockages, we need to relax our permanently tensed or contracted muscles, especially in the thighs and the pelvic area, but also in the neck and jaw.

Bioenergetic exercises are excellent for this purpose. These include shaking and vibrating the legs, the pelvis, and the whole body; stretching exercises, circling the hips, pelvic thrusts, kicking and bicycling exercises, the teachings in the book Ancient Secret of the Fountain of Youth, and other yoga exercises (see Step 9).

With moving exercises, such as circling, bicycling, and pelvic thrusts, the aim is not to train and strengthen these muscles, but rather to move them to the point of exhaustion, which will induce relaxation when stopping the exercise. During the exercise, we try to break through the pain threshold and may then be rewarded with previously unknown pleasurable sensations in the pelvic area.

Exhale through the mouth during these exercises and express any spontaneously forming sounds, such as moaning, crying, or shouting. Other important measures are deep tissue massage and pressing into tense muscles; also reflexology and tracing acupuncture meridians can help. For some approaches, such as pressing into tight muscles, deep muscle massage, and emotional therapies, the cooperation of a friend may be required; alternatively, professional help can be used.

A good exercise that can be done anytime, sitting, standing, or lying down, is rhythmically squeezing the buttocks. At other times, just squeeze the pelvic floor between anus and penis or vagina ever so lightly while concentrating your awareness on the developing pleasurable feeling. You can also do this during slow walking or while driving a car.

Improving Your Feelings: Feelings are what a sexual relationship is all about. The more we are able to feel and express our feelings, the more vigorous is our sexual activity and our life in general. If we are fearful or shy about some aspects of sex, we will be fearful or shy in other aspects of our life as well. Our sexual feelings reflect our emotional health in general, and vice versa.

Often our most basic subconscious feeling related to sex is fear. Out of this can arise anger, hate, and hostility. All of these negative feelings normally remain suppressed and subconscious. After all, it was not acceptable to feel this way as a child towards a parent and it is not acceptable now with a sexual partner.

It would have been easy to have open feelings of fear or hostility towards a parent if we had been obviously mistreated. However, what is a young boy to do if his mother is loving and sacrificing, and he has sexual feelings for her? He wants to be cuddled and touched, but his mother is always too busy caring for the family. He has to suppress his longing for caressing tenderness. He may then come to fear these sexual feelings for his mother because they remain unfulfilled and make him unhappy. This can later translate into a subconscious fear of close emotional contact in a relationship.

On the conscious level, we are kind and loving to our partner, as we were to our mother or father. However, subconsciously we may feel the opposite because a man’s lower self now equates the sexual memories concerning his mother with his feeling for all sexually attractive women. In the same way, a woman may equate the memory of the sexual feelings for her father with her feelings for all sexually attractive men.

This deep emotional split between our conscious and our subconscious mind will result in an emotional or organic dysfunction. We can thus develop frigidity or fear of penetration, premature ejaculation, impotence, and any other form of sexual disorder. This split explains the observation that “nice men” are often poor lovers.

In order to heal our split emotions and sexual feelings, we must break through the superficial layer of being “nice” and contact our suppressed core feelings. This is not easy but can be done with methods such as emotional release therapies involving intentional kicking, hitting, screaming, crying, or shouting. Even just pressing into tense muscles can produce strong emotional reactions. In addition, regression therapy or rebirthing can be helpful, as well as the services of an experienced psychic healer or bioenergetic therapist or attending body-mind workshops. However, the best way to heal our sexual self may be within a healing group as part of our overall healing and spiritual growth.

After most of the suppressed emotional pain has been contacted and released, positive attitudes and feelings can be generated with guided imagery, meditation, creative refocusing, and other methods to heal the emotional wound. An understanding partner will greatly speed up this process. If, on the other hand, we use meditation and other methods to generate positive feelings without releasing the suppressed core, then we will just reinforce the superficial layer of “niceness.” This, in turn, can then intensify our inner emotional split and, along with that, our organic disabilities and sexual dysfunction.

To raise our children without this emotional split we must be loving, honest, and open with our sexual feelings and fulfil the child’s emotional needs appropriate to its age. The baby needs plenty of body contact, including sleeping in contact with the mother. Breast-feeding should be extended more than six months and may continue more sporadically for a long time afterwards.

Why not let infants watch the sexual act? That is how they learn. The growing child continues to need a lot of cuddling and touching. Teach the child that there is nothing dirty or shameful about a naked body or about sexual organs. However, we also need to let the child know the difference between what is appropriate in public and in private.

Food for Sex: We cannot have a good sex life if we are malnourished. Certain nutrients improve our sexual well-being in general, while others are lost to the male in the seminal fluid. Therefore, good nutrition for good sex is usually more important for the male than for the female.

High-quality proteins are part of the seminal fluid and a high-protein diet is sexually stimulating. In our culture meat is preferred, while in Asia chickpeas are used to improve sexual performance. Free-range raw egg yolk, especially fertilized and very fresh, supplies sulphur compounds needed for the semen. More of a short-term effect is achieved with cooked onion, which is rich in sulphur compounds as well; alternatively, use MSM supplements. Bee pollen, spirulina, and chlorella are excellent strengthening foods for both sexes.

Histamine is produced from histidine, an amino acid. Individuals with insensitive skin are low in histamine and may have difficulty achieving ejaculation (male) or orgasm (female), as the sexual organs are less sensitive than normal. High histamine levels have the opposite effect. They sensitize not only the skin, but also the sexual organs and may lead to sexual overstimulation, premature ejaculation, and vaginitis.

Histamine levels can be elevated by making the body more acid, as well as by supplementation with folic acid and niacin, both B-group vitamins. Histamine levels can be reduced by making the body more alkaline, and also by avoiding foods and chemicals to which you are allergic. Therefore, you can acidify or alkalize your body according to your sexual requirements.

The semenal fluid is also high in zinc and requires much vitamin B6 and omega-3 fatty acids which are found in fish oils and linseed oil. Deficiency of these is widespread, especially in teenage boys who typically have frequent loss of seminal fluid. Zinc deficiency leads to underdeveloped male sexual organs and, later in life, to enlargement of the prostate gland. By pointing out this connection, I do not condemn masturbation, but rather seek only to improve teenage nutritional awareness.

In recent years, it has been discovered that the body forms nitric acid from the amino acid arginine as a signal for the blood vessels of the penis to relax. This is necessary so that the penis can fill with blood and have an erection. Also, magnesium supplements help us to relax. Stress, on the other hand, or being emotionally uptight, prevents this necessary relaxation.

Dietary Aids for Female Sexuality: Women have their share of trouble with their sexual organs. Foremost of these are menstrual problems and pelvic infections. Both problems have much to do with food and especially constipation. The constipated colon and in particular the rectum are a cesspool of toxic de-com-position products with a high content of harmful microbes. Toxins and microbes can penetrate the wall of a weakened colon and because of the close proximity to the uterus and ovaries, these are in greatest danger of being contaminated. The main effect of toxic contamination is heavy menstrual bleeding with pain and other discomfort. Microbial contamination leads to infection and inflammation of the pelvic organs. Both conditions are not conducive to a healthy sex life.

Meat consumption, combined with constipation, can produce toxic bowel conditions and heavy periods. The obvious solution is a clean colon, achieved by three easy methods. First have plenty of dietary fibre from whole grains and fresh, raw fruits and vegetables. Add sufficient ground linseed to meals or drink with fluid to achieve ideally one bowel movement for each meal. Second, take a course of cultures of acidophilus and bifidus bacteria as part of an intestinal sanitation; and third, squat or have a high footstool in front of the toilet in order to achieve an easier and more complete bowel evacuation.

As I suggested earlier in the book, the typical menstrual bleeding can be caused by an impure bloodstream, which delays healing of the wound caused by discarding the lining of the uterus. The higher mammals do not have such obvious periodic bleeding nor do women who live habitually on a mainly raw food vegetarian diet; they just reabsorb the shed uterus lining. Typically, these women do not have any menstrual or menopausal discomfort, but ovulate and conceive normally.

A clinical study found that consuming plenty of ground linseed, linseed oil, or fish oils helps postmenopausal women have better vaginal lubrication. Folic acid, in combination with antioxidants, is most important for preventing and reversing precancerous and cancerous conditions of the cervix and uterus.

Selenium, the amino acid methionine (e.g., as selenium yeast), and vitamin E, taken in combination, are effective against periodic breast pain and swelling. Vitamin B6, zinc, and magnesium together are good against premenstrual tension, especially if salt, sugar, and fatty foods are also minimized or avoided and hidden food allergies corrected. Boron raises the estrogen level in postmenopausal women. Taking good care of our bodies will increase the pleasure that we derive from our sexuality.

Another common problem is vaginal Candida infections, resulting in thrush and potentially painful intercourse. In this case, the woman as well as her partner need to go on a strict anti-Candida program (see Part 6). When making love while there is still some Candida problem in and around the vagina, lubricate liberally with oil, and if not using a condom, squirt some acidophilus yogurt into the vagina beforehand. Minimize strong movements and focus instead on feeling the energies with only gentle movements and pressure changes.

Making Love: “Making love” is commonly just a polite expression for having sex without feeling love. One can have a loving sexual relationship for just one night or a lifelong sexual relationship without love. For maximal emotional satisfaction, try to feel and express love with each sexual encounter. It is said that the most important sexual organ of a woman is her brain, but I believe that the heart may be even more important.

Loving foreplay is part of the romance, stimulating the heart as well as the right side of the brain. This is not only important for women, but increasingly also for men once they are no longer filled with youthful vigor. Not only women, but also many men enjoy and need to be played with to “get into the mood.” Increasingly, they come to resent the perception that they are expected to “perform.” They feel that they have to do all the work, first enticing a woman into bed, excitedly undressing her, then stimulating her to get her ready, then “hopping on,” and finally helping her climax. Women, on the other hand, resent being made to feel like sexual objects.

There is often much hidden fear and misunderstanding on both sides. Therefore, take it easy, there is no need to hurry or perform. Talk to each other about what you like and how you like it. Make each other comfortable and dispel any notion that he has to perform or she has to have an orgasm. This is important even if you are already in a long-term relationship together.

Not every time you have sexual play do you need to penetrate, and not every time you have penetrative sex do you need to aim for an orgasm. If you do want an orgasm, there is nothing wrong with having a clitoral orgasm, but at other times also try to aim for a vaginal orgasm.

To increase the flow of energies and feelings, do not suppress pelvic movements but rather enhance them. If you are normally quiet, then intentionally make expressive sounds wherever this can be done in private. When approaching ejaculation, it helps for the male to be “grounded” by pressing the feet against a board or wall or pressing them into the mattress.

While different positions can offer different experiences, they are less important than achieving a union of feelings. Do not try to increase your pleasure with fantasies, go for the real thing instead. Focus your full attention on your own feelings and on feeling united with your partner. To fantasize having sex with someone else is just another form of masturbation.

For some men, it may be preferable to delay ejaculation to either extend the pleasure or give their partner time to build up towards the climax. This may be done by playful interruptions, by slowing the breath, or temporarily even holding the breath, by concentrating the feeling awareness on the heart or, with the penis fully inserted, keeping the pubic areas in close contact, increasing and decreasing pressure in unison while not moving the penis.

A somewhat drastic but nevertheless very effective method of preventing ejaculation is to tense the whole body while also holding the breath. Alternatively, or in addition, strong finger pressure on the perineum just before ejaculation will keep the ejaculate inside the body. This may be advisable at times to avoid the generally weakening effect of ejaculation, especially on the elderly male, but it will require some practice, as will all of these techniques. However, if the climax is definitely approaching, do not try to suppress it but start moving vigorously for a full orgasm. Basically, I am not in favour of tensing the body, as this suppresses feelings and leads to increased muscle tension.

The frequency of intercourse is, of course, a highly individual matter. Normally, I see no reason to restrict sexual play if it does not involve either ejaculation or suppression of sexual impulses. However, the orgasmic potential is usually much stronger after a period of total abstention or separation.

From a bioenergetic point of view, females are much better able to have several lovers than males who would either have to neglect one or become worn out. However, I believe that even for females moderation leads to greater well-being. The difficulty of many women to achieve vaginal orgasm may have something to do with our habit of mating all year round just for pleasure. Female animals have days when they are in heat and strongly attract males for mating. The same may be natural for human

Around ovulation time, a healthy woman who abstained for a month will have a strong orgasmic potential and her biochemistry and the whole body are geared to receive a male. If there are no strong emotional blockages, she may find it easier to have a vaginal orgasm, even after only short but vigorous penetration.

Frequent intercourse can cause even more problems for men. While an occasional full orgasm can be strongly invigorating, frequent ejaculation is decidedly weakening. If a man is sexually overstimulated, he should better remove the nutritional and biochemical causes and find other creative outlets for his energy. It is possible that gluten allergy, often associated with a whitish inner iris, can sometimes cause overstimulation of the sex centre in the brain; synthetic food colourings and possibly other chemicals may do the same.

Generally, it is much healthier to have fewer ejaculations with strong orgasms than frequent ejaculations with weak orgasms. In teenagers, frequent ejaculation may impair their mental and physical development, while in later life sexual overactivity makes us susceptible to diseases and premature aging. What is frequent depends on the individual’s constitution and commitments. What is normal or beneficial for a healthy and strong young man might be too much for someone older or with a weak constitution. Wilhelm Reich saw the proper function of the orgasm as the release of excess energy that might otherwise lead to irritability, tension, and energy blockages. Therefore, if you have excess energy, go for it, but if you lack in energy, be careful.

In a long-term relationship, the ideal may be in having mainly sensuous and spiritual sex for most of the month and orgasmic intercourse close to ovulation, as often as desired. A female in touch with her body will feel when the time is right and let her partner know. However, the best time is always when both partners feel like it.

Alternatives to Sex: Those without a permanent partner may find an occasional short-term partner or pay for sex. If all goes well, there is no problem with this. For those without even a temporary sexual partner, the obvious choices are masturbation or abstinence, although there are some less obvious alternatives, as shown below. For masturbation, the same principles apply as for making love. That is to aim for fewer climaxes but with stronger orgasmic releases. For both sexes, this means waiting with masturbation until one feels a strong charge of sexual energy and using pelvic movements, sound, and feeling awareness to ensure a strong orgasmic discharge.

For celibates, an easy solution is a low-protein diet with few spices and condiments. This may be suitable if there is sexual overstimulation due to too much histamine from overacidity or food allergy, but otherwise I am not in favour of this low energy approach. Rather, I believe in making ourselves as energetic as possible, including having plenty of sexual energy, but then to learn to channel this sexual energy into creative outlets, such as sports, charity work, or the arts.

An alternative is the following rejuvenation rite, which is from a Tibetan monastery in which, reportedly, rejuvenation of the body was successfully practiced (see Ancient Secret of the Fountain of Youth by Peter Kelder). This rite was used in combination with several other rites as shown in Step 9. Also a daily raw egg yolk is recommended for those of us who are no longer youthful in order to increase our sexual energies.

When sexually excited, bend over with your hands on your knees and push all the air out of your lungs. Then straighten up and with the hands on the hips push up the shoulders. Pull the stomach in and up and hold this position for as long as possible. Then forcefully inhale through the nose and exhale immediately through the mouth. The arms may now hang to the sides while you take further deep breaths through the nose or the mouth. If you are still sexually excited, repeat this procedure as often as required.

What many single people and especially women miss even more than sex is touching and hugging. Some try to satisfy this need with regular massage therapy. However, there may be a simple solution. If you feel deprived of touching and hugging, just form a local “hug club.” Either discreetly spread the word around or put a notice in a suitable place asking for interested individuals to contact you. You can then arrange a weekly meeting where you just hug for an hour or two. As a follow-up step, you might also give each other massages. You may be surprised by the amount of interest there is out there in hugging. Alternatively, you can form or join a healing group where hugging can be practiced together with massage, reflexology, and the multitude of other self-healing techniques available.

Eventually, you may also want to experiment with some meditative forms of hugging as described later in this chapter. Sexual energy can be used in spiritual sex to energize the higher energy centres with guided imagery or meditation exercises, possibly leading to what is called “whole body orgasm” or “brain orgasm.”

Getting to Know Your Partner: Sensitive individuals are often apprehensive about the first sexual encounter with a new partner. The man may have anxieties about erection, premature ejaculation, and satisfying his partner, while the woman may fear penetration, getting emotionally hurt, or not coming to an orgasm. This can make the encounter tense and inner fears may become self-fulfilling.

Why not proceed slowly and be intimate without intercourse until you are fully relaxed with each other? This initial period of a relationship is precious, as it is commonly the most exciting and romantic time. Enjoy it as long as you can. Let your partner know that you feel this way. In a situation that might normally lead straight to sex, tell your partner beforehand that you enjoy becoming intimate gradually and do not intend having intercourse that night. You might be surprised by how relieved and agreeable your partner is to this. A whole mountain of tension is suddenly removed.

A good way to start the intimate part of your relationship is to have a shower or bath together and then give each other a sensual massage. Continue exploring each other's likes and dislikes and responses to touching and cuddling. You might even spend the night cuddled into each other. If, contrary to your original intentions, you should be comfortable enough with each other to proceed to full intercourse, nothing is lost, and it will probably be a better experience than it might have been in a tense atmosphere.

Such intimacy together without penetrative sex can be quite satisfying for a considerable length of time, especially in special situations, such as with elderly couples or with physical or emotional handicaps or religious, spiritual, and adolescent celibates. A sexual relationship does not need to be an all-or--nothing proposition. The "macho" encounters commonly portrayed on TV and in the movies are not an everyday reality. A loving sexual encounter without expectations cannot result in dissatisfaction. It does not matter whether there was penetration or orgasm; it is the loving feelings that matter most.

Sex Therapy: It is not generally realized that many health and emotional problems can result from unsatisfactory sexual relationships. These include migraine and tension head-aches, premenstrual tension and menopausal problems, depression, irritability, discontent, violence, anxiety, disturbances of digestive and sexual functions, psychotic symptoms, and high blood pressure, and even cancer has been linked to blocked sexual energies and tensions. All of these may, of course, have other causes and contributing factors as well.

Problems arise mainly when sexual energies are stimulated without being appropriately used or released. This is the same as with other emotional energies. Commonly, this happens if the male ejaculates before his partner reaches orgasm, but it also happens if the orgasmic release is weak or the encounter too short. The signs of a full orgasm are strong involuntary movements, a blissful discharge followed by complete relaxation, and an afterglow of satisfaction and contentment.

Ejaculation in itself or clitoral orgasm do not indicate a full release of sexual tension. There may then be the desire either for repeated or frequent sexual activity or avoidance of sexual activity in order to avoid the resulting tension and dissatisfaction. However, orgasm is only part of a successful sexual union; another important part is the exchange of bioenergy and emotional energy between the partners.

Accumulation of sexual tension with its resulting health and emotional problems can be prevented even without orgasm, while orgasm alone may not always be sufficient. The key to a successful sexual union is the duration, because it takes time, usually about half an hour, for these energies to be exchanged by the partners. This kind of “sex therapy” is not only effective, but also highly pleasurable, especially for sensitive individuals.

Sex therapy requires two willing partners. Preferably there should be an atmosphere of affection already well before the planned union. This should start with an extended foreplay of caressing with full skin contact, possibly including a sensual massage, but without any clitoral stimulation. However, if this kind of union is routinely practiced as a love meditation, then such elaborate preparation may not be required

When both feel ready, they place themselves in such a position that the woman lies on her back with knees drawn to the chest, while the male lies on his left side crosswise to her, so that his penis touches the entrance of the vagina. She now drops her legs and he places his right leg between her legs. With this, her left leg is between his legs while her right calf rests on his torso. In this way, both can remain completely relaxed with freedom of movement which is essential for success. Sides can be reversed.

Initially, and also if there is a problem with erection, natural lubrication, or premature ejaculation, both partners remain more or less motionless in this position for about half an hour. During this time, the tip of the penis, whether soft or hard, remains outside, just touching the inner entrance of the vagina. Both concentrate all their attention on the contact areas between them, but especially on the sensation in their sexual organs and the exchange of subtle energies between them.

Often, after half an hour there is an increase in sexual arousal and the penis may now penetrate. However, this is not essential. If there should be no erection or the vagina still dry, just remain in this position for another half-hour while gently caressing each other with full attention on your feelings and exchanging energies. If penetration is desired, but an erection does not come on its own, then start gently rubbing the head of the penis between the lubricated entrance and the clitoris.

If this is not sufficient, it is relatively easy in this position to insert a soft penis. This may be a solution for men with erection problems. If you firmly hold or squeeze the shaft of the soft penis, then any blood is pressed towards the head of the penis and insertion will be easier. Once inside, gentle movements by both partners increase the chance of a partial or full erection.

Even without an erection, focusing the awareness on the contact within the vagina combined with slight pressure changes at the pubic areas can be sufficiently satisfying and may eventually lead to more. However, gently liberating and feeling your sexual energies and feeling at one with your partner does more for your well-being than chasing an elusive erection or orgasm. The main precondition for a successful outcome in resolving any difficulties is a freedom from any kind of expectations on both sides.

Even after penetration, continue to move slowly and concentrate on the sensations in the sexual organs and the flow of energy and feelings between your bodies. If both partners are sufficiently aroused after a prolonged foreplay, the vagina is well lubricated, and the male able to delay ejaculation, then the half-hour waiting time outside the vagina is not necessary and immediate penetration is acceptable.

Should ejaculation occur early, then just remain close together with the soft penis inside the vagina until about half an hour has passed from the time of penetration.

With such a complete exchange of male and female energies, combined with a full orgasm, there is usually no desire for another sexual union for several days, and it is actually best to wait for about five days until the full potential of sexual energies is restored. The following union will be more pleasant and satisfying than with more frequent encounters. It is, however, relaxing to sleep together with bodily contact.

The alternative to orgasm is a deliberate choice to use this union as a love meditation, also called spiritual sex or Karezza, from the Italian word carezza, signifying its caressing aspect. You can just continue moving slowly while caressing each other for hours, while concentrating on feeling the energies and radiating feelings of love onto each other. If this Karezza type of union is practiced routinely, then an elaborate foreplay to build up a strong sexual charge is not necessary, but may be used if so desired. Condoms inhibit the exchange of sexual energy, while lubricating oils or jellies may slow it down, but this is not a problem when remaining connected for a long time.

It may take some time and require several attempts to master this method and achieve a maximum of enjoyment, but it is well worth the effort in terms of improved health and personal relationships.

* * * * * * * * * * * *

Part 8
HEALING YOUR EMOTIONS

We all would like to live a happy, healthy, and fulfilled life, yet few seem to be able to do that. Why must we have so much suffering, failure, and disappointment? We want to be secure in a loving relationship and a satisfying job with a good income and enjoy ourselves. Instead, our relationships turn sour, we have a job that we do not like or we do not have one at all, there is never enough money, and generally there is not much joy to be found anywhere. As a result we feel resentment, hopelessness, and depression, perhaps also anger and hatred. Why did it go wrong?

I believe that there is a way out of this depressing jungle of negativity and disappointments and that we can make a decision to have a happy, fulfilled life, and then do it. What I cannot promise is that it will be easy. It needs willpower and determination and the help of good friends. But you have nothing to lose by trying, and each step in the right direction is likely to bring you some reward.

We can be happy, or at least accepting, in adversity, and unhappy when we seem to have it all. What determines how we react and feel? It is no secret to psychologists that we are programmed since earliest childhood by everything that went on around us, but especially by the way our parents talked to us and to each other, by the way they felt and reacted, and by our interactions with our siblings. We became programmed by obeying, observing, and imitating our role models.

If we were lucky and grew up in a happy and loving family, we probably have an inner program that makes it easy for us to lead a happy life in a loving relationship. If, on the other hand, there was much worry, anger, resentment, and negativity in our childhood, chances are we have a hard time being happy and loving as adults. That probably applies to most of us, the products of negative programming during childhood.

If we want to change, there is only one thing to do: change our subconscious programming from a negative outlook on life to a positive one. Unfortunately, we cannot simply decide that from now on we will be happy and loving. On their own, such conscious decisions have little influence on our subconscious programming. In fact, if there is a conflict between our subconscious programming and our conscious mind, the subconscious will always win.

Therefore, we must be clever and beat the subconscious at its own game. We must throw the old program out and devise a new one that allows us to be as we want to be now. However, it is not sufficient just to work with our feelings and emotions because these do not arise out of a vacuum. Our emotions are conditioned by our beliefs. We react more to our beliefs than to facts or anything else, so we must adopt an appropriate belief system.

Where do we find the beliefs we want? It is not sufficient to try to believe that we will be happy ever after. Our proposed new beliefs must be plausible to be acceptable to our conscious and subconscious minds. They must give meaning to our life and make sense of it all. Meaning is the opposite of the inner emptiness that many of us try to fill with pleasure seeking and power games, causing so much emotional trouble in the end and depression if we fail. My proposed solution is a spiritual philosophy of life that turns emptiness into meaning.

There is also a strong influence of physical factors on our emotional life, such as nutrition and tense muscle structures. Therefore, in order to heal our emotions properly, we must work on all four levels of our being: the biological, emotional, mental, and spiritual.

* * * * * * * * * * *

Step 57
UNDERSTANDING EMOTIONS AND DISEASE

Learn how you came to be so out of touch with your feelings, what emotional problems and diseases this causes, and how this is expressed in your body.

We experience ourselves in the world with our feelings and emotions. They are the driving force, the power, and the motor of our life. Without feelings and emotions, we would be like robots, computers operating bodies with the help of electric signals.
It is the feelings and emotions, our likes and dislikes, that give our life meaning, that make us fulfilled or dissatisfied, and to a large degree decide our course of action and health. Diseases make us feel unwell and negative, and suppressed emotions are a major contributing factor to these diseases. When we are young, we are full of feelings, experience them strongly, and are sensitive and react immediately and directly to our social and physical environment. When we are old, our emotions tend to be a distant memory and diminished, and pain is often the only feeling left.

Many scientists at present are so divorced from their feelings that they believe these originate in the brain, like thoughts. They come to this conclusion because they do not feel the energy of anger in their body or the love in their heart. They just think anger or think love in the brain and act accordingly without feeling anything, or their feelings are only weak and diffuse. Of course, feelings are registered in the brain, but the feeling doesn't originate there.

Feelings provide us with the greatest pleasures in life, but also with the greatest suffering. Suffering is the key word to describe our loss of feeling with advancing age. We do not want to suffer, so we intentionally diminish our feelings in order to diminish the amount of emotional pain we feel. As an unintended side effect, this reduces the amount of pleasure we can feel. Hand in hand with our reduced overall feelings go a reduced enjoyment of life, reduced vitality, and an increased susceptibility to chronic degenerative diseases. This is a high price to pay for reducing the suffering that we might feel. In fact, we have exchanged bouts of intense short-term suffering for more low-key long-term suffering.

Helpful Definitions of Emotions: You may be unsure about the difference between feelings and emotions, so I will define the terms commonly associated with the world of feelings as they are used in this book.
I distinguish between three groups: sensations, emotions, and feelings. Sensations are body reactions due to physical or energetic causes: our awareness of hot and cold, touch or pain; during meditation or bodywork, we may feel warmth or tingling or a streaming sensation.

Emotions are inner states of experience due to nonphysical causes such as expectations, beliefs, and memories; they affect the body through brain and hormonal changes. Emotions commonly have a positive or negative connotation about someone or something: we are angry with someone, afraid of something, in love with someone.

Feelings in a wider sense are what we register as sensations or emotions. We say “I feel pain” or “I feel anger.” With this, we focus on the subjective inner experience regardless of the cause. We use the term “feelings” in a more specific way to describe inner states not caused by sensations or emotions, such as a feeling of well-being or a feeling of unconditional love during meditation. These feelings are emotionally neutral and not about someone or something. Another possibility is a dispassionate feeling about something as in an impartial judgment. At a crossroads, for instance, we may feel a preference for one road over the other. In most instances, however, this is not a proper feeling, but an intuition or hunch.

Moods are generalized feelings usually beyond our conscious control, and often with a somewhat negative connotation. Sentiments are more tender feelings, but usually about something and can be grouped with the emotions; desires are about something and are emotions. Passions are generally regarded as strongly felt and expressed emotions.

How We Lost Our Feelings: The widespread emotional misery in our society has much to do with our diminished capacity for feelings, especially for tender, loving feelings. If at all, we seem to experience these only for a short period in our life when we fall in love, and then yearn for them ever after.

The repression of our feelings seems to come largely from the male-inspired cultural priority placed on the intellect in our society, with a corresponding contempt for soft and tender feelings that are regarded as feminine. It is easier to gain power, to dominate, and to build an empire if one is not hindered by sentimental feelings. With their present bid for equal power, even many women cannot afford any more to be vulnerable and they steel themselves against feminine softness.

Self-control is highly valued, especially in the Anglo-Saxon cultures. We are not supposed to show anger or even displeasure, but rather be outwardly polite, though we may be boiling inside. Only in recent years has it become more acceptable to show tenderness in public, but even so, cuddling and touching, except in a ceremonial way, are still largely constricted to those who identify with the New Age movement. Conventional members of our society are as rigid and inhibited as ever.

We can say that our emotional dilemma arises from two sources. One is the suppression of our feelings enforced by the standards of our society, and the other is the lack of role models for the development of tender feelings in our childhood. Suppression of our feeling starts as infants when we are trained not to cry when we are unhappy. We may be just left to cry without response until we realize it is useless to express our frustration. Especially suppressed are any expressions of childhood sexuality, such as playing with the genitals and even cuddling the parent of the opposite sex or generally adults of the opposite sex.

Such contact is important to us later for developing mature emotional relationships and adult sexuality. Unfortunately, there is now a perception to regard such sexually open adults as potential child molesters; fearing this, they intentionally hold back, depriving themselves and the children of this essential expression of affection.

Also at an early age, we are trained in the use of social lies and social conventions and are discouraged to express our true feelings. It is even worse at school and university where we are overwhelmed by intellectual activity, usually of an unimaginative kind. This inhibits our creativity and intuitive abilities. Many great inventors and innovative scientists did not make their discoveries by thinking but through intuition. For the rest of his life, Einstein reportedly tried to understand on the mental level what he had perceived in an intuitive flash.

Creativity is the active outlet for the meaning we give our life, and intuition is an important link between both. This combination gives satisfaction, perpetual joy, and happiness to our life, but it is crushed by our educational system in order to prepare us for the later institutionalized conformity in our workplace. A more suitable alternative would be an educational system based on the model of Steiner, Waldorf, or Montessori schools.

Emotional Role Models: The presence of positive role models for our emotional development is most important during infancy because then we are most impressionable and learn the fastest. At that age, we learn to feel by absorbing feelings from our surroundings like a sponge. This activates our feeling centres. If we are radiated with love, our capacity to feel love develops; if we are exposed to anger, then that is what we learn. I believe that because of their own love-deficient upbringings, most parents act lovingly towards their children from a mental level without actually feeling it. The most common exceptions are non-intellectual, soft, and “feminine” mothers, especially as personified by the typical Polynesian woman.

In many third-world countries, it is common for whole families to sleep in the same room, where children may be aware of the sexual activity of their parents. This is natural, as it is in the animal kingdom, and provides role models for subconscious programming. But it is essential that the sexual interaction, like any other interaction of the parents, be a loving one. This imprinting is most important for infants and does not need to interfere with our more inhibited Western morals for older children who, unlike infants, prefer their own rooms anyway.

The lack of tender feelings combined with an exposure to negative emotions from the parents in early life has much to do with the present culture of violence in our society. This is constantly reinforced by the bombardment of violence on television and movie screens. With their tender feelings undeveloped, many do not feel the suffering of others; they simply cannot feel compassion. That may start in children with cruelty to animals and lead to individualized or institutionalized cruelty and torture when adult.

A national football coach is quoted as saying that to be successful in a contact sport you must build up hate for the opponent before the game. He is right in that hate releases stored negative energy and it may even be good that it does. Even the many onlookers can release some of their negative energy. But is this the kind of society we want, where our role models require hatred to perform? There must be a better way!

Emotions and Disease: There are many reports that demonstrate the strong influence of our feelings and emotions on health and disease. One example is a study of the survival rate of women with breast cancer. After ten years, 70 percent of those who reacted to the diagnosis with a fighting spirit were still alive; those who reacted with denial had a 50 percent survival rate; those with stoic acceptance had a 25 percent survival rate; and of those who felt hopeless and helpless only 20 percent survived.

Other findings show that cancer frequently is diagnosed about a year after a traumatic event, such as losing a spouse. The negative effect of mental depression on our immune system is well known, while we feel and to some degree are invincible when we are in love. When we are upset or stressed, our digestive juices are inhibited, and fear or apprehension can even trigger asthma attacks.

Another problem is with suppressed negative emotions. There is always the possibility that under provocation these may be released explosively in an act of violence. This pattern is now very common in our society. On the other side of the fence are those sensitive souls who are full of bottled-up feelings and emotions, but unable to release and express them. This drains their vitality and they suffer from poor circulation with low blood pressure, cold hands and feet, and lack of energy.

Generally, we can distinguish between acute or immediate influences of strong emotions and the long-term or chronic effects of unexpressed emotions. The immediate reaction is due to the direct influence of our feelings on the hormone-producing endocrine glands and on the nervous system. More insidious, however, are the long-term effects of muscle tension caused by suppressed emotions.

At the physical or body level, this leads to poor blood circulation and a reduced supply of nutrients to the affected area, while metabolic wastes and toxins tend to accumulate, similar to sediment in a slow-flowing part of a river. In addition, permanently contracted muscles generate a great strain on associated joints. A combination of these factors makes us susceptible to the development of arthritis. Constriction of the chest and diaphragm commonly leads to respiratory diseases.

Muscle Armouring: Muscle armouring is a concept described by Wilhelm Reich, M.D., an eminent American psychiatrist and at one time the apparent intellectual heir to Sigmund Freud. Dr. Reich found that the emotional disorders of his patients were to a strong degree reflected in their body structure. In particular, patients with repressed feelings commonly had hard, rigid, and permanently contracted muscles in certain areas of the body. Different kinds of negative emotions seemed to be associated with specific muscles. Reich compared these rigid muscle structures to the armour of a medieval knight and called the process of their formation “armouring.” While the steel armour of the knight had the purpose of protecting him against physical aggression, the muscle armour protects us against emotional aggression.

It may sound far-fetched that contracted muscles should have anything to do with how we feel. But we can easily observe that we become tense when we are apprehensive and that our muscles relax when we are at ease. When we expect someone to cause us pain, such as sticking a needle into us or hitting us, we tense our muscles and we may hold our breath by contracting the diaphragm.

These are automatic body reflexes designed to diminish the expected feeling of pain. However, we can use the same mechanism if we want to diminish feelings for other reasons. As an infant, we may have been afraid of being separated from our mother, of being left alone in a strange surrounding, so we tensed up and held our breath. Eventually, we had to continue breathing, but we did it in a shallow way, with contracted diaphragm and chest muscles.

What Body Language Tells Us About Our Emotions

This summary is a condensation of ideas from Body-Mind, by Ken Dychtwald.

Legs: How we move through life.

Weak, underdeveloped: not well grounded, no firm stand in life

Massive, overdeveloped: rigidly grounded, needs to explore, let go

Fat, sluggish: to move through life needs enthusiasm, jogging

Thin, tight: moves energetically through life, but often erratically, not gracefully, needs to develop tranquillity

Pelvis: The condition of our sexuality

Front tipped downward causing hollow back (lordosis): usually string sexual energy, but full flow is blocked through constant self-control, not able to 'let go' – develop faith in higher guidance.

Front tipped upward causing flat low back: lessening of sexual focus, lack of tender feelings in the lower part of the body – learn to lower attention from chest to lower abdomen, develop tender emotions

Belly: The centre of emotions as they relate to oneself

Enlarged in upper half: rugged, outgoing, masculine – develop more tender feelings

Enlarged in lower half: blocked energy flow to pelvis and legs especially if the abdominal wall is hard – let go

Moderately enlarged: if belly is soft and back not very hollow, good contact with body vitality, possibly emotionally too soft

Overall enlarged and obese: usually poor contact with vital energies

Overall flat: too much mental control, no "gut feelings", strong emotional blocks (fear, anxiety) – develop tender emotions, have faith, let go

Chest: Modifies our emotions as they relate to our intentions in reference to ourselves

Narrow, contracted: feeling of inferiority, lack of power, unexpressive – learn to communicate, to give, chest breathing

Wide, expanded: feeling of superiority, power, expressive – learn to experience tender feelings within, especially in pelvis, learn from others, abdominal breathing

Shoulders: How we carry our burdens in life

Bowed, rounded: feel overburdened – develop power, chest breathing, have faith

Raised: chronic fear – anxiety-releasing therapy, chest breathing

Square: carries responsibility – relax

Forward hunched: fear of being hurt, self protection – develop power, chest breathing

Pulled back, retracted: forceful control or suppression of unwanted emotions, especially anger – let go, express yourself in a suitable way

Narrow: cannot shoulder responsibilities – become more powerful

Right side lower: interacts in a predominantly masculine way

Left side lower: interacts in a predominantly feminine way

Arms: How we express ourselves in physical actions

Weak, underdeveloped: lack of initiative and physical expression – learn to communicate with your arms

Massive, over muscled: insensitive, forceful interactions, lack of grace – learn to be gentle

Thin, tight: inability to hold on to anything – become more peaceful, settled

Fat, underdeveloped: sluggishness in expressing yourself – become stimulated, animated

Upper back: A channel for the expression of forceful or violent emotions

Soreness, hump: repressed anger or hitting reflex – let go; hit a pillow or sandbag

Neck: Reflects tension between body emotions and mental control *(continued)*

Bent forward: explores the world first in a rational way, emotional exhaustion – develop your ‘gut feeling’

Bent to the right: arrogance, defiance – become centred

Bent to the left: playful attitude

Long, graceful: proud attitude

Heavy, short: forceful attitude

Throat: A channel for vocal expression of emotions

Tight, sore, weak: blocked verbal expressions of emotions and tears – let go, speak out, weep

Jaw and Chin: A channel for verbal expression, biting

Receding: frozen, suppressed verbal emotions – learn to speak out

Protruding: determined

Strongly protruding: defiance, arrogance – relax

Clenched: forceful, self-control, suppressed anger – let go, relax

Face: An outside mirror of our emotions

Changing expressions: show how we wish to appear to the world

Chronically tense muscles: show conflict between what we show and how we really feel

Eyes: The “windows of the soul,” reflecting health and emotions

Large, round: warm, loving personality
Protruding: reaching out forcefully (thyroid problems)

Deep-set: critically observing, withholding expression

Wide-open baby eyes: tries to hold, to draw close, not fully matured

Nearsightedness: frozen fear (early childhood), focus on immediate problems, introspective, rational - release fear, look into the future

Farsightedness: suppressed anger; focused outward, extrovert - release anger, develop your inner self.

With repeated fear responses, these muscle contractions can gradually become permanent. The child can grow to develop a narrow chest with weak lung functions, or if only the diaphragm remains contracted, a barrel chest can result. In both instances, breathing remains permanently shallow and the child is susceptible to lung infections and asthma. In babies, this may be a contributing factor in crib death or SIDS. The advantage of the armoured chest or permanently contracted diaphragm is that we now do not feel our fear anymore, as it remains subconscious. However, we gradually may develop substitute outlets for our fear, such as being afraid of heights or of public speaking, and so forth.

Another example is the suppression of anger because we are told that it is not socially acceptable to show it openly. We may initially feel an angry emotion rushing upwards from the abdomen to the shoulder with the impulse to hit or to the throat with the impulse to shout. When we suppress these impulses, the emotional energy of the anger becomes stuck in the shoulder or in the throat and tenses the muscles there. If this tension is not released by other means, it may become permanent, and if we continue our suppressive behaviour pattern, these muscles become permanently contracted. In joints surrounded by contracted muscles, we easily develop arthritis, while contracted throat muscles give us a weak voice and possibly stuttering and other speech and throat problems.

The muscle armouring becomes ever stronger with advancing age because we tend to repeat our set behaviour patterns. This then forms our distinctive facial features, body structures, and our increasing rigidity. There are, of course, other factors that contribute to shaping our body and making it more inflexible, such as heredity, nutrition, and occupational muscle use.

Dr. Reich found that his cancer patients had severely inhibited sexual energies caused by strong armouring in the pelvic and abdominal areas. Those patients who could most successfully free their sexual energies had the best chances of recovering. He regarded cancer as a “shrinking biopathy” of the total energy field of the patient. This shrinking was seen as a forerunner and not as a consequence of the disease. Tumours are only the final stage of the shrinking process, Dr. Reich said. When he succeeded in shrinking the tumor or making it disappear without revitalizing the general energy flow, the patient would die anyway. Spontaneous remissions, on the other hand, were in some instances due to a revitalized energy flow, Dr. Reich said.

Another aspect of armouring is the generation of pain. The resistance of a contracted muscle to the flow of energy produces pain similar to the heat produced by the resistance of a thin wire to the flow of electricity. Short-term muscle contraction uses energy; therefore, it causes no pain and can be used as a defence against expected emotional or physical pain.

However, if a muscle remains permanently contracted, the reduced blood and energy flows cause pain from accumulating toxins, oxygen deficiency, pressure on nerves, energy congestion, and overactive acupuncture points. Such pain may be present all the time or felt only as tenderness when the muscle is pressed. It can be stopped immediately by pressing into tense muscles or relaxing them with other methods. In a permanently contracted muscle that has become like a rope or sheet, the energy flow to the area is so diminished that there is no pain, even when pressed. Repeated deep muscle massage can eventually restore energy flow and temporary pain to the muscle. Some individuals can clairvoyantly perceive these energy flows.

Body Language: Our emotions have a strong influence not only on our glands and inner organs, but also on our body structure. Certain emotions are traditionally linked with problems in certain organ functions, as noted in part 7. Anger, for example, damages the liver and, conversely, irritability and quick temper are partly caused by liver problems. In a similar way, grief, negativity, and anxiety are linked to the lungs; fear to the kidneys and intestines; excessive laughter or lack of joy to the heart; and worry to the spleen.

Emotions, if not released in outward action, “solidify” by causing muscle contractions. The stronger the energetic charge of the emotion, the stronger the muscle contraction. Other parts of the body may be more or less blocked off from the flow of emotional energies, and these parts will become weak and start degenerating. By examining our body, we can get a reasonably good idea of the kind of emotional problems that have helped to shape it and of the corrective measures to be taken in order to improve ourselves (see Table 7-3 above).

* * * * * * * * * * * *

Step 58

HEALING RELATIONSHIPS

Try out these tools to heal your personal, family, and social relationships.

We live in three different forms of relationships - our higher relationships with God and the spiritual dimension; our relationships with our planet and its life forms; and our interpersonal relationships in society.

In holistic living, we move towards greater harmony on all three levels. Meditation and following a spiritual path are our attempt to heal our relationship with our spiritual source, while an attitude of loving care towards our planet and its creatures will heal our relationship with our biological roots.

Our social relationships, in a way, are the most difficult to heal, for they are the testing ground for what we have learned and how far we have spiritually advanced. With much effort and by using emotional release, reprogramming, and meditation, we may feel that we have made much progress in healing ourselves. But have we? We can only find out in our social interactions.

Are we upset if someone makes a mistake? Do we suppress our annoyance, or do we speak out, and, if so, in a suitable way? Do we instead feel like pointing out what has been done wrong and how to correct it, or do we prefer to let the other discover this for himself? These are all different possibilities for responding to the same problem, and the way we respond can show us how much we are already healed and what we still have to do.

To make it more complicated, there is no correct response that suits every situation. Sometimes one response is the best for a given situation and at other times another response would be more suitable. This means we need to become flexible and respond to each situation with an open mind and open heart, rather than rely on past patterns of responding. Further, no matter how well we fare in these tests of everyday life, can we be fully happy and satisfied if we encounter on the streets or on television the suffering of our underprivileged brothers and sisters? I do not think so. With our increasing sensitivity and compassion, this causes us to suffer ourselves and we feel drawn to help others. Again, there are different ways in which we can respond to ease our common suffering. Foremost may be to give hope and healing by our living example.

We can also donate money, goods, and services to the needy. We can single out an individual for special support or we can work through an organization; we can use the political system or in other ways try to change our whole structure of social injustice. What will it be? Increasingly, we will be drawn to do something because we feel and know that we are all one. Whatever our choice, it is not a question of right or wrong, but of trying to live according to our highest ideals, using the opportunities as they arise.

Intimate Relationships: Our intimate relationships, more than anything else, are a testing ground for our feelings and emotions. They can make us feel on top of the world or in the depths of depression. How do we cope with these strong feelings and emotions? Our intimate relationships are a mirror of our general level of consciousness; they show us how well we are able to manifest our ideals in everyday life.

There are different kinds of intimate relationships. In the best sense, they all will be love relationships, lovingly relating to a partner, be it in marriage or without a legal status, short-term or long-term. Preferably, we try to find a partner with similar or compatible ideals or we try to improve our existing relationship by finding and harmonizing our ideals. Our common ideals, more than anything else, can help to ensure a happy and fulfilled long-term love relationship. A relationship based mainly on sexual attraction or a need for material or emotional security, on the other hand, does not have a solid foundation and is likely to fail.

In our society, it is fashionable to make romantic love the deciding factor for marriage. However, this is the kind of love of which it is rightly said that love makes blind. With this, it combines the two worst factors for marriage: It does not last and it blinds us to the shortcomings of our partner.

After a few years, the erotic appeal has been swept away and the relationship faces the test of reality. In many instances, the couple stays together only for the sake of the children; in others, it is the need for emotional and material security. A new romance is now sought outside of marriage. Sometimes this leads to a string of marriages and divorces based on erotic straw fires.

The only hope for such a marriage to provide a happy and satisfying relationship is the transformation of erotic love into mature love. Sometimes this just happens on its own with compatible individuals, others have consciously to work towards this goal. While it is said that one cannot make oneself love someone, one can work, nevertheless, on diminishing one’s ego, which is the greatest obstacle to loving the other as a person.

I see the greatest chance of successful transformation into mature love when both partners become interested in inner growth and self-responsibility and start walking the spiritual path hand in hand, healing themselves and each other. You could also join with other couples who have similar interests.

Making a Relationship Work: One of the most destructive elements in many present relationships is the demanding of rights. Traditionally, males have had most of the rights, and a relationship can develop in this way today without any effort on anyone’s part. This commonly leads to the emotional deprivation of and increasing resentment by the female partner. The other side of the coin is the growing trend of Western and especially American females to be assertive and demanding in a relationship. This tends to push the male partner away and results in the male emotionally withdrawing; or the relationship becomes an arena for competition and fighting. The solution depends on a cooperative approach.

This may start by one of the partners realizing that the competitive approach does not lead to happiness and makes the relationship unpleasant. The main thing that this individual needs to realize is that he cannot change his partner by demanding or expecting change; each can only change themselves. Depending on the nature of the relationship, the partners may talk over how to change track or the partner who realizes this may take the first steps alone.

An example of this is the current movement in the U.S. of “surrendered wives,” which started as a counterbalance to the demanding of rights by “liberated” females.68 This certainly is a recipe for creating happier relationships, as long as it does not lead to resentment on the woman’s part. With this unconditional first step, many women will transform their partner; he will become much more considerate and pleasant to live with. However, it may not work in every case and occasionally leads to increased abuse and separation. Nevertheless, this is a spiritual solution to turning a difficult relationship around and trying to make it more rewarding for both partners.

In most instances, females want a partner who embodies their ideal of masculinity, while males want their ideal of femininity. I am not referring here to physical appearance, but to emotional qualities. Therefore, try to find within yourself and offer to your partner the best male or female qualities that you are capable of. Do this with a loving heart and without demanding or expecting an instant reciprocal change from your partner. In most cases, your partner will begin to change for the better of your relationship.

Things that your partner did not do or did only with resentment when you demanded them may now be offered as gifts. Instead of nagging or demanding what you want your partner to do, the right attitude is to think of what you can do to make your partner happy. However, if your partner is unwilling or unable to change in return, and repeated discussions do not lead to a satisfactory result, then separation appears to be the best solution. It is not a sign of spirituality to let yourself be emotionally abused and suffer in silence. If you believe that circumstances force you to remain with an uncooperative or abusive partner, then use this opportunity to learn and practice unconditional love and forgiveness.

The keys to making a relationship work are goodwill and communication. We must learn to share our inner life with our partner. It is also important how we share. We let our partner know that we feel hurt by something he or she did or did not do, but we do not accuse. We speak of our own feelings and may say, “I feel hurt by what you did,” but we do not say in an accusing voice: “You did this to me.”

Initially, it may be good to agree on set times for sharing. Ask your partner if something annoyed him or her and then let your partner speak without interruption. If this direct approach is too difficult in the beginning, keep diaries about your hurts and suggestions and exchange them from time to time for discussion. However, the same basic rules apply: Do not accuse or verbally attack your partner and give each other uninterrupted time to speak.

Infidelity: Infidelity is possibly the most common cause of severe relationship problems. From a spiritual (not a religious) point of view, the real problem is not that our partner shared love or sex with an outsider, but that he or she broke the stated (or unstated) trust in the sexual exclusivity of our relationship. One partner then feels resentful and the other guilty. This is not a recipe for a happy relationship, and how soon and well it can be mended depends mainly on the spiritual quality of forgiveness on the part of the hurt partner.

Looking at this from a spiritual point of view, it is quite natural for a sexual attraction between males and females to develop in suitable conditions, while on the other hand it is not good for our spiritual and emotional well-being to suppress strong longings or desires. This leads to a typical dilemma: We have a choice either to harm ourselves through emotional suppression or through feelings of guilt with a possible deterioration of our relationship.

A spiritual way out of this is to learn to transmute our sexual feelings for an outsider into unconditional love and to beam this back onto this person. Another possibility is instead of sharing sex to restrict an intimate encounter to a (fully clothed) meditative cuddle.

If your relationship is sufficiently open, you might also discuss if and under what conditions a sexual encounter with an outsider is acceptable. You may realize that the demand for exclusive mating rights is due mainly to a combination of our social conditioning and fear of losing our partner. This is not a solid foundation for a spiritual relationship. Instead, if sexual fidelity is important for your relationship, then it should be freely agreed to by both partners because that is how you feel about each other and that is what both really want.

However, emotionally mature partners will not make this a relationship law because they realize that we are not perfect. If one partner is not able to keep this agreement, then the spiritual solution is to ask the other partner for forgiveness and to obtain it without any damage to the relationship. If both are secure in their love for and commitment to each other, then they may not even wish to emotionally restrict or deprive their partner. However you work it out in your relationship, try to find solutions that are based on openness, understanding, and forgiveness.

An Example of a Caring Tribal Society: We may sometimes wonder what it would be like to live in a genuinely loving and caring society. Surprisingly, such societies have existed, even in modern times. In 1929, the anthropologist Bronislaw Malinowski published an account of his two-year stay with the Trobriand Islanders near the east coast of New Guinea.

All the people there were happy and caring, he reported. They shared with each other and liked to work. Children were always good-natured and conformed effortlessly to the rules of their society; and they had one freedom that appears unbelievable to us: There was not the slightest restriction on their sexual activities. There was no emotional suppression.

Children watched their parents and others making love and when growing up started imitating them. Younger children would just tenderly play with each other, while teenagers had unrestricted penetrative sex. However, there was one important social convention: Pregnancies must only occur in marriage. These islanders actually did not know that sex had anything to do with getting babies. Everyone believed that babies were a gift of the gods given about nine months after the wedding, because it always happened that way. Without any other precautions or protection than this belief, babies out of wedlock were unknown.

After loving experiences with many different partners, the youngsters eventually found their marriage partners. Marriage was desirable because it gave babies and status. The marriage lasted as long as the couple loved each other, usually for many years or a lifetime. While married, there was never any unfaithfulness or even a desire for it; they did not even have a word for it.

Marriage breakup was no disaster for the children. The father always remained a good friend of the mother and the children. Furthermore, as they did not have the concept of a biological father, the maternal uncle always played the role of the father. This was a matriarchal society and all property rights went through the mother to the children.

As these people could observe all varieties of loving sexual activity, they had no unhealthy fascination or obsession for them, as do many people in our society. Consequently, abuse of children, sexual perversion, rape, or other forms of violence were unknown to them. Sexual activity to them was as natural and open as eating is to us. No doubt, if eating were regarded as “dirty” and restricted in the same way as sexuality is, and children and others were not allowed to see us eat, most of us would have neuroses associated with eating disorders. This association of sexual freedom with a peaceful society can also be observed in the animal kingdom with the bonobos, smaller relatives of chimpanzees.

Another Example - the Continuum Concept: In the 1970’s, the American Jean Liedloff lived for several years with the Yequana Indians in the jungles of Venezuela. She discovered that these Indians lived in a continuum of unbroken traditions in close harmony with their inner and outer nature. This appeared to be the reason for their unusually positive qualities.

There was no trace of aggressiveness towards each other amongst the Yequana, Liedloff found, even if they were all drunk at an occasional party, and they never blamed anyone for anything. Even in difficult circumstances, they remained in a composed and happy state of mind. There was no competition in their interactions, only cooperation.

They were completely fearless, surefooted, agile, and enduring. At rest, they were totally relaxed, but could be fully alert in an instant. To illustrate this in a funny way, Liedloff related that sometimes one of the men would wake in the middle of the night, remember a joke, and start telling it. Instantly, all the others were wide-awake, roared with laughter, then seconds later were sound asleep again.

They loved to work, and for them it was a pleasure to use their bodies. They did not even have a word for work; they just did cheerfully whatever needed to be done. The children were always well behaved, obeyed happily and instantly, never fought among themselves, and were never punished. How they raised their children was, of course, the secret to their success as human beings and as a society.

Until the babies indicated on their own that they wanted to start crawling, they were always carried, usually by the mother. During sleep, the baby remained in body contact with the mother. There was no danger of injuring the baby when rolling over, as subconsciously, the mother always remained aware of the position of the baby.

After the baby started crawling and later creeping, the mother was always available when the baby desired body contact. However, the mother would not tell the baby what to do or not to do. There were sharp knives lying around, which the baby might pick up by the wrong end to play with. There were deep pits or dangerous drops near which it might play without an anxious glance from the mother. It might even pick up a burning branch from a fire and stumble around the thatched huts. There were no warning calls and no accidents.

Neither were the children told when or what to eat or when to sleep or which role models to follow. Boys quite naturally picked up bows and arrows from other boys and started shooting with sharp arrows without being shown where or what to shoot or any safety precautions. They might start doing this at the age of 18 months.

Girls at that age, on the other hand, might join a group of women and girls and practice grating roots. At the age of three or four, a girl might be in charge of an infant; play and work are the same to them. An adult might need something, give a short command or request, and the child would instantly comply.

The parents trusted the innate protective instincts of their children. A baby might appear to be oblivious to danger when playing at the edge of a cliff, but its subconscious or body self protected the baby much more effectively than the mind could, just as it does a sleepwalker. If we warn an infant of a danger, its mind starts interfering with the body self and it loses its instinctive protection. If, on top of this, the parent is afraid that the child will fall, then it will try to fulfil this expectation and fall.

Small children are psychic and always try to meet the expectations of their parents, especially of the mother. The Yequana children never fail because their parents have unwaveringly positive expectations. Children are neither blamed nor praised for anything, just confidently expected to do the right thing.

Further, babies in that society are deliberately exposed to the widest possible range of sensory stimulation. From this and by observing adults and older children, they learn to judge unfailingly their ability to cope with each situation, and their parents completely trust their judgments. They know that if the child is not confident to judge a situation, it will ask a parent for help.

While these men were much tougher than men in our society, if one genuinely felt strong pain, such as when cleaning out a deep wound, he would not hesitate to cry out loud, comforted without words by physical contact with his wife or mother. If either a child or an adult lost their natural family, they were adopted into another family. This was important even for adults for emotional and social support. To show that the amazing abilities and behaviour of these children were not inherited, but the result of their upbringing, Liedloff reported there was one problem child. This child belonged to a Yequana couple who had lived in contact with our Western civilization for some time; but this child also learned (belatedly) to fit in.[69]

Emotional Security in Children: We all are likely to benefit from increased levels of emotional security. Emotional security is a nebulous concept that includes how we feel about ourselves, how good we feel in relationships with other people who are important to us (e.g., family), and how confident we are that things will turn out well for us.

Probably all of us feel some doubt in some of these areas, and especially children, if they are able to think in these terms. One of the main functions of a family unit is to foster feelings of emotional security in children, and often much of the irritating or negative behaviour seen in children can be removed through actively raising the child's levels of emotional security.

This is an easy claim to make but often a harder thing to achieve. Here are some techniques that I recommend to try to raise the level of emotional security in children:

1. Set aside three to five minutes every night for the child. After the child has gone to bed, Mom or Dad (take turns) goes in, turns off the light, and sits on the side

of the child's bed in the dark. The child is in the security of its own bed, with a little body contact with the parent, and no eye contact, because it is dark. Then the parent should just talk about their own day, share some of the things that happened; no questions asked nor asking the child to talk. When the child learns that this is a regular occurrence, he will initiate things and feel free to share, and topics involving emotional stress are likely to come out.

2. When the child makes a statement to you (e.g., "I hate school"), try not to block it or answer the comment; extend it instead. The child may not realize the feeling behind this statement, and you may never find out unless you can get the child to clarify it. Hand it back to the child in a questioning way: "You hate school?" To which the reply may be, "No, just the teachers." So you say, "You hate all the teachers?" "Well, not all, mainly Mr. Jones."
3. The question "Why?" rarely achieves anything when asked of little children. Seldom can they provide an adequate answer to such a question, and so, effectively, they are cornered. Try to avoid asking "why."
4. Avoid the words "no" or "wrong," especially when the child is attempting to do something (e.g., reading). "Almost right" or "not quite but getting better" are likely to keep the child interested and keen to try. The words "no" and "wrong" are likely to make a child give up as a failure.
5. Let your children know that they are good at things, that they are nice people, and that you like them. Generally, we tell our children when they fail, when they annoy us, or when we feel let down by them, but we don't let them know the good things. Many children thus get the impression that they are failures and develop a poor self-image.
6. Right-handed children like to sleep on the right side or on the stomach with the head to the left shoulder. Left-handed children generally sleep facing the other way. Try to place your child's bed so that in the natural sleeping position (according to handedness) she sleeps facing the wall. This tends to give the child added security and often has the effect of eliminating problems with light sleepers as well as nightmares and bed-wetting. Placing a child's bed at right angles to a wall, extending out into the room, should be avoided with children who are light or restless sleepers, as it provides little or no security to the sleeping child.
7. If a child needs a night light, try a blue or green bulb rather than a normal white or red one. Blue and green are pacifying colours, whereas red is stimulating. Also position it so that the child's face remains shaded, so as not to interfere with melatonin production during sleep.
8. Accept your child's reality. If the child is upset or scared about something, irrespective of how irrelevant or trivial it may seem, accept that this is the real feeling of the child. Rather than dismissing the complaint or saying that it does not matter or not to be silly, ask what the child is feeling; then go through these feelings so that the child can either accept or work around the worrisome feeling.

9. Set aside a short amount of time on a regular basis in which your children can have your undivided attention; this could be perhaps ten minutes in the evening, or while doing the dishes. This can help avoid the repetitious “in a minute” response that we constantly find ourselves giving to our children.
10. Read to your children. Younger children (two to six years old) enjoy and benefit from favourite books being read to them numerous times; they can learn the whole story by heart and can “read” it back to themselves just by looking at the book. Older children (seven to ten) benefit from having something interesting read to them by a parent, so that they feel they are sharing a common interest with the parent; suitable may be historic stories or children’s encyclopaedia stories.
11. Fool around with your children. Let them see that adults can laugh, play, and be silly.
12. Be consistent. Consistency on the part of adults is of prime importance. If you act consistently, the child will know where he stands. If not, the child will be confused and become unpredictable as well. Always do as you say. Do not threaten punishment unless you are willing to carry it out, otherwise you lose credibility. The same applies to offering rewards.

Redoing Our Social Relationships: As compared to the Yequana, we live in a primitive society when it comes to our social relationships. I believe that future generations will see our society as severely unbalanced, as an overdeveloped technocracy built on a socially sick base. We can learn from the Yequana how to heal our society: It starts with our relationships with our children. For a long time to come, we will not be able to rely on authorities and experts in this endeavour; instead, idealistic individuals and groups must take the initiative.

Parents, preferably within the framework of a support group, must start experimenting on how to raise their children according to continuum principles as explained above. This is not easy in our situation. It involves not only a child’s constant close body contact with the mother or other carers for the first six months and a trusting attitude with positive expectations, but also the presence of other children in a wide range of ages as positive role models. Two-year-olds learn best from three-year-olds and they in turn from four-year-old children.

It can be done. I have known mothers who kept body contact with their babies until they wanted to start crawling. I have seen a three-year-old boy without close supervision or need for safety concerns use a sheath knife, hatchet, electric drill, and a blowtorch, then he immediately and happily ran home when his parents called. I believe that children, if they feel secure and loved during their early childhood, if they have a wide range of relevant sensory stimulation and positive role models, can be trusted even in our modern technological society.

Obvious dangers, such as knives, staircases, or pools, should not be a problem. Babies easily learn to float. I see less of a chance of a secure child drowning in an unfenced pool than an insecure one drowning in a fenced pool. Make hidden dangers known to your children through your own behaviour. Children imitate; therefore, do not do in their presence what you do not want them to do.

Antisocial behaviour is a sign of emotional deprivation, especially in early childhood. Give your existing children and especially those with antisocial or neurotic tendencies the opportunity to experience the missed phases of their emotional development. Explain to them in simple terms what you intend to do and why; invite your children, possibly one at a time, to sleep in your bed whenever they feel like it. Get them to do it even if they are initially reluctant; cuddle them a lot and possibly massage them, or bathe or shower together.

For infants, all this will be natural and enjoyable, but with older children emotional barriers may need to be broken down gradually. While loving sexual activity in the presence of infants is desirable to provide a role model, for older children it might be upsetting to be exposed to it unprepared. Use your own judgment. Include the children in the tender relating of the whole family, but in a way that does not intentionally cause sexual stimulation.

Healing Relationships to Produce Social Change: Emotionally deprived children grow into adults who have antisocial or asocial tendencies. On one side of the scale, this manifests as criminal and destructive behaviour and, on the other, as shyness, lack of confidence, and depression. A further result is the ego-centred individualism so prevalent in Western societies. It brings forth political, business, and community leaders who are more interested in personal power and wealth than in the overall good for society.

In contrast to our present system, which I would describe as one of adversarial competition, a society of emotionally mature individuals would be based on benevolent cooperation. Elected representatives, for instance, would fully cooperate to find the best solutions for the common good, and the most suitable citizens would be chosen to be these representatives.

Business ventures would operate according to cooperative principles, both in matters between different ventures and within each, and without a sharp division between capital and labour. This was the social structure of the Yequana Indians and the Trobriand Islanders. However, to be successful in modern society, such a system must be based on spiritually and emotionally mature individuals.

A first step towards reforming the present business climate would be a legal and community expectation that the main duty of care for corporations is to the whole of society rather than to their shareholders. The reason for the existence of companies would be seen as service to society; a company would appreciate its duty of care to its employees, shareholders, and the environment. A further step could involve government incentives to form cooperative ventures.

The judicial system would be based on conciliation and rehabilitation. The more emotionally mature the society, the less would conflict arise. The health care system would largely become redundant as people learned to live in harmony with their biological, emotional, mental, and spiritual needs. Any necessary therapy would be a combination of the best in modern and natural medicine.

Agriculture, food processing, and energy production would be sustainable and in harmony with planetary requirements. A society of self-responsible citizens would be based on a maximum of individual freedom and a minimum of bureaucratic interference.

No system of laws and law enforcement will keep a society of deprived and ego-centred individuals operating harmoniously, yet emotionally satisfied and responsible individuals will function well in any system with a minimum of rules. The main problem, therefore, is not to devise better political and social systems and improve our laws and wealth distribution, but to provide the opportunity for everyone to heal their emotions and start on the road to self-responsibility. A harmonious society will then evolve from this.

Cooperative Solutions Based on Emotional Fulfilment: Here I offer a few examples to show how a caring society based on emotional fulfilment, as outlined in this step, might solve some contemporary issues.

The most important aspect would be to raise all children in a caring and loving way, following for example, the principles of the Yequana continuum concept, to ensure a harmonious society in the future. The school and education system would concentrate on developing cooperation, creativity, intuition, and holistic concepts similar to Rudolf Steiner, Montessori, or Waldorf schools.

There would be no political parties or electioneering. In one possible model, electors would write the name of a community member of their choice on the ballot. Leaders would be chosen for their experience, maturity, spiritual insights, and wisdom. The elected leaders from several neighbourhoods would form local councils of wisdom; these would then elect someone from their midst or possibly an outsider to represent them at the next higher level. In this way, it could continue upwards to create a world council of wisdom.

Very suitable as community leaders will be women who have harmoniously raised a family because they will be skilled in facilitating just and harmonious social relationships. The legal system will be very simple, observed in spirit rather than letter. Together with other decisions affecting individuals, communities, and regions in their care, councils of wisdom on each level will also be responsible for any legal disputes and arbitrate in a spirit of cooperation.

Initially, the unemployment problem could be solved with flexible working hours. This means, instead of an unemployment rate of ten percent, the standard hours of work per week would be reduced by ten percent. Conversely, if a shortage of workers developed, the standard working week would be lengthened. Standard hours might be different in different professions or ventures. The government might just provide a basic safety net. Present government regulations generally make it more convenient to extend overtime hours for existing workers rather than hire additional staff. This is simply a problem of community attitude and government competence. Eventually, work will become a desirable outlet for creativity.

Based as it is on the adversary system, our society causes so much friction, conflict, waste, and incompetence that I estimate our present lifestyle could be maintained with less than a quarter of the present work input. However, this percentage would initially be higher in order to bring underdeveloped nations up to world standard and restore the planetary environment to an appropriate level.

In a mature society, only a fraction of present resources would be required for the legal, law enforcement, and prison systems, for the defence and pharmaceutical industries, and for the hospital system. Most of these could eventually be phased out. On the other hand, much more would be channelled into sustainable agriculture and health-friendly food production, into education, training, and community facilities, including an appropriate public transportation system.

In regard to health care, I see a unified system of holistic medicine in which surgery and drug medicine are branches, like nutrition, herbalism, or homeopathy. Extensive clinical trials would determine the best combination of procedures for specific diseases. However, in an increasingly enlightened society, diseases as we know them will become increasingly rare. Surgery, for instance, would only be required for accidents. Most of the healing work in such a society would be concerned with promoting inner and outer harmony and spiritual growth.

Sexual Initiations: In regard to a more natural sexual development of children and teenagers, I see two possibilities. One is complete openness in sexual matters, letting children watch the loving sexual activity of adults without any restriction and imitate it as they feel ready to do so. This is nature’s way as practiced in the animal kingdom and in human societies that still live close to nature. It does not necessarily lead to unwanted pregnancies, as shown by the Trobriand Islanders; other methods of birth control might have to be taught. How it would work out in the end depends on the customs of society and the examples provided by its individual members.

The other possibility is to transform juvenile sexuality into a striving for high ideals within the context of a loving and caring society. Infants, but not older children, would be exposed to lovemaking of the parents. At a certain age or level of development, girls would be ceremonially and lovingly initiated by experienced and spiritually evolved men, while boys would be taught by women with the same qualities. The teenagers might choose the initiators themselves or leave this selection to their parents. As a variation of this possibility, teenage couples could receive practical instructions before or during their first sexual encounter. This, too, could be done ceremonially. However, the main problem for teenagers is not to learn the mechanics of sexual union, but rather to become emotionally competent.

My preferred solution in a society with strong spiritual values would be to teach teenagers graded forms of spiritual sex in a ceremonial way as an initiation, and also the art of sexually relating in a loving way, again graded according to age. Different solutions will be suitable for different societies, depending on their degree of spiritual evolution, but almost anything would be better than our present system.

Of course, in our present situation all this sounds utopian. Unfortunately, I do not believe that a transformation into a more caring and mature society will take place gradually and in a coordinated way. I am afraid the present system will run into an acute crisis and more or less self-destruct. If and when this happens, it will be important to have mature leaders in all areas who then can pick up the pieces and show the survivors the way towards building a caring society.

* * * * * * * * * * * *

Step 59

HEAL IN GROUPS

Join a group of like-minded individuals with the goal of healing your feelings and helping others do the same.

The small nuclear family as the basis of our social structure is an expression of the ego-centred and individualizing nature of our society. It leads to social isolation, difficulty in relating and cooperating, and a disruption of the age-old continuum of traditions.

Formerly, and presently still in underdeveloped countries, humans enjoyed a continuum of learning experiences. Step by step, through observation and imitation, children learned how to care for babies, do the various forms of housework, repairs, and other skills, make love and relate in a loving way, care for sick people, give birth, and deal with death and dying. Grandparents made themselves useful and passed on their experiences. Now most of these skills are only incompletely learned from books or left to experts, and the art of relating and cooperating in groups is at an all-time low. Recent explorations in communal living are an attempt to return to the security and emotional closeness of the extended family and a more traditional lifestyle.

Activities in various clubs and special interest groups may lessen our social isolation, but this is only a bandaid. I suggest forming neighbourhood healing groups with the specific aim of healing ourselves, others, and our social relationships.

A healing group can provide many functions not normally available in our society. In a way, it can act as an extended family with an unlimited variety of support facilities, such as baby-sitting, child-minding, kindergarten and preschool, and possibly even alternative schooling. Members can help each other in all stressful life situations, be it childbirth, sickness, bereavement, unemployment, or building activity. There can be common fruit and vegetable production, a food co-op, communal library, leisure facilities, and shared expensive working appliances.

The main aspect of a healing group is healing activity on all levels. For healing the body, experiment with natural and holistic methods. Suitable communal healing modalities include massage, reflexology, acupressure, spinal therapy, meridian therapy, muscle testing, and allergy testing. Many different forms of massage can be used, such as stimulating massage, deep muscle massage, sensual massage, or energy distribution massage.

Individuals can share their personal experiences with the group in areas such as nutrition, color therapy, packs, or flower remedies. A member of the group may participate in a healing workshop and then pass on the experience to the others. The group can also invite natural therapists and other health professionals and healers for demonstrations, lectures, and teaching of specific methods.

There can be group evenings in which you exercise together, while on others you discuss interesting topics; someone may read a good book and report on it; or there can be meditation evenings, or all of these in varying combinations. You can use group therapy for emotional release, do rebirthing and other emotional growth work together, and do a lot of hugging. Buy some books about emotional therapies for the communal library, and experiment.

A healing group can operate on two levels. On one level, it can be a neighbourhood support group with food co-op, shared child-minding, communal leisure activities, and so forth. On another level, individuals dedicated to healing can meet weekly for this purpose. A large garage could be converted into a group meeting place while smaller groups might meet in the living rooms of participants.
Here are a few of many possible group activities for developing feelings and relationships. In group exercises, wherever possible, try to pair off with a partner of the opposite sex, or when forming a circle, alternate the genders.

I SEE THE LOVE: The women sit in a circle, possibly on floor cushions or on the carpet. A man sits opposite each woman. If there is an excess of either gender, pair off with a member of the same sex. For a minute, look silently but tenderly into each other's eyes, trying to express and radiate love. Then, with a signal from the group leader, the men move to the next partner and again both meet each other in a tender gaze. In this way, the men move around the circle until they sit again opposite their original partner.

Now, while again looking into each other's eyes, the men say, "I see the love in your eyes." They can either repeat this phrase from time to time or a variation on it, such as: "Your heart is filled with love and it shines out of your eyes" or "My heart is filled with love; it radiates from my eyes."

After a while, the men can in addition start tenderly touching and stroking the hair and the face of their partner. After about two minutes, on a given signal, roles are reversed and the women express tender feelings in words and by touching and stroking the hair and face of their man. Finally, on the signal, the men move to the next partner and repeat the same procedure and so forth until they are back with the original partner. Now, one after the other, participants can describe their feelings and how it felt different being either the active or the passive partner.

THE SANDWICH: This is to recreate the security, love, and warmth of the foetus in the womb or the baby in the mother's arms and at her breast. Someone stands in the middle, front gently pressed to the front of a partner of the opposite sex, and with the front of another partner of the opposite sex against the back. The middle person has the arms over the shoulders of the front partner and the one at the back also stretches out around the middle person to the one in front.

Stay like this for several minutes, possibly lightly swaying if that feels better. In addition, you can have two more people lightly press against the sides of the one in the middle. See if it feels better if these additional "huggers" are of the same or the opposite sex as the one in the middle.

For a woman in the middle, it may also feel better to have another woman in front; just experiment with different combinations. You may, for instance, have four people of alternating sex standing as closely as possible with their fronts together, or the whole group can form a ring, each one standing pressed with the front against the back of the one in front. The ring may stand still or slowly move. Some of these formations can also be tried while lying down.

LEARNING TO TRUST: After pairing off, couples walk around slowly, intermingling with the others, but with one partner of each pair having their eyes tightly closed or blindfolded. After some time, you swap places.

In another exercise, form a circle with seven people and one person in the middle with arms crossed in front of the chest. Those in the circle lightly touch the middle person with their hands. The one in the middle has the eyes closed and after a moment the body relaxes and starts swaying. It will lean towards one pair of hands and is then gently passed around the circle. After a while you can make the circle bigger by adding more participants or just spacing yourselves out so that the one in the middle is inclined at a greater and greater angle. His or her body should remain straight with the feet in the middle of the circle. Those forming the circle can put one foot back to be able to support a greater weight.

In a variation of this exercise, one person stands with closed eyes and the back turned towards a mattress on the ground. Several helpers stand behind the person, with outstretched hands, and catch that person as he or she falls backwards with a stiff body. Gradually extend the range before catching.

HEALING PARTNERSHIPS: If there is no healing group that you can join, or sometimes even in addition to it, you can find a healing partnership. Ideally, that is with your present partner in a long-term relationship. However, many are not so lucky and often prefer to live on their own rather than in a stressful relationship. In this situation, you can find someone with the same interest in health improvement and spiritual growth, but without either one of you wanting a love or sexual relationship. You can then meet at set times, and do all the things that might be done in a healing group. This will probably be more suitable for mature-age individuals. Eventually, others may join in and then you have your proper healing group.

* * * * * * * * * * *

Step 60

LEARNING TO FEEL

Feelings build the body; they are the glue that holds body and soul together.

Tender feelings make us open and vulnerable, but in order not to get hurt, typically we prefer to close up and not feel. This has the added advantage of making us stronger in our career because we do not need to take our feelings or the feelings of others into consideration.

Each time we suppress a feeling or do not express it in a suitable way, the generated energy solidifies into muscle tensions and armouring, as we learned earlier through the work of Dr. Reich. Eventually, this leads to widespread muscle armouring, a permanent state of muscle contraction; this closes off the circulation of blood, lymph, bioenergy, and, equally important, the flow of feeling energies. Once we are in an armoured condition, we cannot feel anymore, even when we want to. This applies especially to our tender feelings and is a great loss. Our conscious control is then out of touch with the body - its needs, wisdom, and the pleasure it could give us.

Thus we are equally out of touch with the requirements of others, with the forces of nature, and with our higher guidance. Everything is wrong if we are out of touch with ourselves and with everyone and everything else, and we have to rely exclusively on our ego-controlled mind. To heal our emotions, we need a fourfold approach:

1. We need to “let go” and release the accumulated negative emotions and associated tension.
2. We need to become aware again of how we feel and express our emotions appropriately.
3. We need to learn how to feel good about ourselves and others and to generate positive feelings.
4. We need to live in our daily lives what we have learned in our exercises.

Let Go of Negative Emotions and Tension: Feelings that we have not expressed and often not allowed ourselves even to feel have accumulated within us in the form of repressed emotions. They choke our emotional body in the same way as accumulated metabolic residues obstruct our physical body. In our relationships and social interactions, we react emotionally mainly with reactivated hurts within us and only to a lesser degree with true feeling to the actual situation itself. This causes endless frustration, misunderstandings, and disappointments in our daily lives.

In order to free ourselves of these emotional obstructions from the past, we need emotional cleansing periods in which we feel safe to release and express our suppressed emotions. This will already partly be achieved and made easier as a result of bioenergetic exercises, deep muscle massage, and other methods to relax our muscle armouring. Here are three helpful release techniques:

Forgiveness: The cornerstone of emotional healing is forgiveness. As long as we cannot unconditionally forgive ourselves and others, we remain trapped in past negative emotions. This prevents us from fully loving ourselves and others.

A good way to develop forgiveness is to make a list of everyone who has hurt you in the past or whom you may have hurt. Recall especially relationships with your parents, siblings, relatives, and partners. Think of incidents that may have caused anger, resentment, disappointment, sadness, fear, or insecurity. Then mentally go through each incident and note if there is still any trace of that hurt feeling in you. If there is, go deeply into it, reach the bottom of it, examine it from all sides; it may then evaporate.

If there is still some hurt, decide to free yourself of this. Do deep breathing and, with each exhalation, imagine blowing the pain into a balloon. When you have transferred all the hurt into the balloon, close it up and let it float off into the blue sky for the universe to take care of. Imagine yourself in the presence of the one who hurt you. Formally and lovingly forgive this person. A very important person to forgive in this way is yourself. Forgive yourself for all the distress and hurt that you caused other people. If you are not ready to forgive everyone unconditionally, come back to this step after you have worked more on your belief systems and adopted a spiritual philosophy of life. Realize you are not doing a favour to others by forgiving them, but to yourself by freeing yourself from destructive emotional toxins. This is similar to cleansing your body of chemical toxins.

An alternative is to write a letter to everyone against whom you still feel resentment and whom you are not able to forgive. Write exactly what the other one did and how it affected you. Then formally forgive them and express your love and appreciation. Give this letter to the recipient, or you can burn it and hand the matter over to the universe.

Intentional Direct Release: A simple and efficient method of emotional cleansing is possible in a secure relationship with an understanding and cooperating partner. When you feel angry, sad, or upset, tell your partner there is something coming up that you want to release.

It is often necessary to exaggerate our expressions in order to get to the old emotions. When you are angry during an intentional release, show that you are angry: Throw yourself on the bed or the floor, kick and punch a cushion, pillow, or mattress, wring a towel, scream, shout, or cry. Similarly with hatred, sadness, grief, or any kind of frustration, bring each out as forcefully as you can. Your partner can help you to deepen the feeling, telling you to get into it, to hit harder, cry louder.

Even in more restricted situations with other people, try to experience and express your immediate feelings in an appropriate way. When you are angry, say so and possibly hit with the fist on the table; when you have tender feelings, acknowledge them with a loving smile and, if appropriate, with a loving touch.

Re-enactment: In many instances an understanding partner may not be available for immediate release or the situation is inappropriate. Then it is best to re-enact a recent hurtful or otherwise frustrating experience, as long as it is still fresh in the memory and the feelings are easily aroused.

Do this at a convenient time in the privacy of your bedroom or in a car parked with closed windows near a busy highway (where nobody will hear you). Recall the details of the scene and your feelings; bring out these feelings as much as you are able to. A similar situation from the past may pop into your mind, a time when you felt hurt in the same way yet did not react to it. Relive and exaggerate the memories of this past experience to get to the old emotion behind your present feeling.

Learn to do this habitually. Whenever you encounter an emotional situation to which you cannot react in an appropriate way because of inhibitions or conventions, re-enact it and respond to it later and elsewhere in an exaggerated form. Do this for anger and hurts and also gratitude and tender feelings you were unable to express at the time.

EMOTIONAL THERAPY FOR CANCER: Many individuals and especially cancer sufferers are inhibited in their emotional expression. Often they are unselfish, quiet, and dependable, the caring and enduring partner in a relationship. However, they also tend to harbor resentments as they hold back in expressing their anger or their own needs. For them, family and partner come first, while they sacrifice their own desires and dreams. This then inhibits their life force, their inner drive, and vitality. In cancer lore, this type of behaviour has been called the Type C personality as compared to the Type A personality, which has the opposite character traits and a susceptibility to heart disease. Most cancer patients have a so-called Type C personality or behaviour. Type C personalities typically:

- Rarely or never feel or express anger or other negative emotions, such as fear or sadness
- Are patient, cooperative, appeasing, “nice” individuals, non-assertive and compliant with authorities
- Are extremely self-sacrificing and overly concerned with meeting the needs of others rather than their own
- Tend to hold resentments, instead of speaking out

Another emotional cause of cancer is an unexpected loss. This may be a partner or close relative, a job, business, financial security, or anything with which the individual has a strong emotional bond. In individuals with a Type C or cancer personality, such loss often causes a severe depression of the emotions as well as of the immune system, and any hitherto dormant or slow-growing tumours explode in a sudden growth spurt. These emotional cancer causes need to be rectified in order to give any additional therapies a chance to work. If you have the opportunity and financial resources, consult with someone who specializes in emotional therapies; otherwise work with your partner or a friend.

Write in as much detail as possible about the worst experience of your life; share it with a trusted person. Search your mind and write and talk to someone about your other resentments, hurts, disappointments, losses, and shattered dreams. Practice expressing yourself forcefully, even exaggerating gestures and body movements. Remember incidents that you did not like and induce and release anger by kicking and punching something; cry or scream if possible, then afterwards forgive yourself as well as everyone else for past hurts and mistakes.

Practice expressing emotions and feelings with your body. Read poetry aloud with feeling; express feelings through dancing, arm and hand movements, or pantomime. Find another dream, something worthwhile and exciting to do after you have recovered from your disease. Be assertive, especially with medical and other authorities; make sure that your emotional needs are met. Meet other cancer sufferers in weekly sessions for emotional group therapy and support; share information, meditate, and visualize together. Take control of your therapy program.

I believe that your emotional transformation through learning to express your emotions is the key to winning the battle with cancer or any other disease. In summary:

- feel your negative emotions and express them appropriately
- feel your emotional needs and move towards fulfilling them
- uncover and release all resentments
- forgive yourself and others for anything you or they may have done
- speak out, become assertive
- trust your own knowledge and intuition
- find a dream that you would love to make come true in coming years
- become enthusiastic about your present life and future

Keep these points in mind and move towards changing your behaviour in small steps in your daily life whenever an opportunity presents itself. Use mind improvement methods, such as meditation, affirmations, and guided imagery or visualization, to shape yourself and your future as you want it to be.

Overcome Fear: A disease such as cancer provides us with a great opportunity to learn to care for our body, to heal our mind and emotions, and to become spiritual. Some of the inner causes of cancer are long-standing resentments, habitual worries, inner hurts or grief, guilt, repressed sexuality, and a lack of joy and spontaneity. All of these are expressions of fear and we need to overcome them by developing unselfish love.

One of the most basic fears to overcome is the fear of dying. This can best be done by seeing yourself as an immortal soul who temporarily inhabits and operates a biological body to gain certain experiences. Numerous near-death experiences show that death or shedding the biological form is generally a liberating and pleasant event for the soul; you may have similar experiences to this during meditation.

Another fear is fear of pain and dependence on others, especially in an impersonal and technological hospital setting. By using holistic medicine, you can remain in charge. Even if you are too old or start holistic therapy too late, you can expect to spend your last days in a dignified way. Instead of trying to cling to the body or being drugged into a semi-conscious state, this should become the spiritual high point of your life, as you prepare in prayer and meditation for the liberation of your soul. After the shocking revelation that you have cancer, you must gradually progress to hope, belief, and faith, faith in your spiritual guidance and that all will be well. Do your best; the rest is in God’s hands, so there is nothing to be afraid about.

This inner transformation can be achieved with the indicated mind tools. Preferably develop a daily routine that you start with prayer, affirmations, and relaxation, followed by guided imagery and meditation.

Become Aware - Experience Body Sensations: Even without the goad of having cancer, we need to become aware of our feelings and emotions when they arise. An easy way is to start with simple body sensations. Try to increase your awareness of body sensations. Whenever a sensation arises naturally, be it warmth, cold, tingling, hunger, thirst, the tastes during chewing, or any other body sensation, focus your attention on it. Soon the sensation will intensify and gradually you will learn to pick up on even faint sensations that you normally would not have felt.

Unpleasant sensations will usually recede after a while and may disappear completely. This is especially the case with hunger or pain. In order not to feel hungry during a fast, tell your body beforehand that you are doing a cleanse to make it feel better and that it will get food afterwards; then ask for its cooperation. If you forget to do this beforehand, you can still explain it when you feel hungry. Usually the craving for food disappears after this.

It is similar with pain. By focusing your attention on it, the pain can tell its message, the reason why it is there, and then it can withdraw, helped by the muscle relaxation induced by your attention and non-resistance to it.

One of our most frequent sensations, both important and pleasant, is our food taste. Use the excellent opportunity provided at each meal to savour the flavours that develop while placidly chewing. This increases and helps your digestion. For healing a specific problem, imagine that powerful nutrients in that food are now on their way to the problem area to improve it.

At other times, try to feel the various parts of your body as you move, stand, sit, or lie down. Sense which muscles are tight and which ones relaxed; warmth may develop wherever you focus your attention, yet you can also feel a refreshing coolness if you concentrate on that. An easy way to do this is by mentally transferring the warmth of the breath felt during exhalation or its coolness during inhalation to a designated part of the body.

Slightly move an arm, leg, or your head and experience how each move feels. Focus on the beating of your heart; feel the pulse at the sides of the neck. Put a hand on your chest to feel the movements caused by the heartbeat and the breathing activity. Shake your whole body or one part of it and feel the energy flow when you stop. Lightly touch various objects, your skin, that of others, tree bark, a stone, glass, and feel the differences.

Slowly move your lightly cupped hands close together, but without touching, then move them apart. Repeat for a minute or two, moving them slowly towards each other and away again. Can you feel a resistance developing when the hands are brought closer together, something like an air cushion? Do the same exercise after shaking your arms and hands for a minute. Move the tips of your outstretched fingers close to those of another person. Try with the same hands and also with opposite hands. Can you feel an energy, a tingling?

Allow Spontaneous Feelings: Focus your attention on any emotions or moods that arise. When you are angry, do not just remain angry in your head. Scan your body to detect if you can feel it there. It may be in the form of an energy rushing upward from the lower chakras. Is it activating your arm or leg muscles, so you want to hit or kick? Intensify any body feeling. Hit the table or a wall with the padded side of the fist; kick a cushion or the air. How does it feel now?

If you are frustrated, where in your body do you feel it, apart from your head? Ask the same with other feelings, such as fear, anxiety, worry, jealousy, hatred, longing, compassion, love, devotion. With tender feelings, focus your attention on the heart. Do you feel a stirring there, something moving? When you feel an emotion in the body, try to stay with it and intensify the sensation, even if it is anger or another negative emotion. Try to express whatever you feel, but in a suitable way. Certain muscles may want to move; intensify the movement.

If you feel fine, sense that well-being all over the body. If you are dissatisfied, unhappy, or discontented for no specific reason, relax and focus your attention on feeling dissatisfied, unhappy, or discontented. Feel it in the body: Where is its centre? What does it want to tell you? Some of our strongest feelings develop during sexual intercourse, so go into these feelings. Try to spread them out from the sex organs into the pelvic area and throughout the body. Radiate them out to your partner. Do not let yourself be distracted by fantasies; stay with the feelings.

Sentic Exercises for Emotional Clarity: Manfred Clynes, author of Sentics: The Touch of Emotions, developed a set of “sentic exercises.” Clynes defines sentic as the brain state and its corresponding experience generally associated with the word “emotion.” He found that the same emotion in people all over the world produces the same kind of muscle response. He used the pressure of a finger on a measuring device to record the intensity and direction of the pressure generated by each emotion.

Even more important for us, this process can be reversed: A certain finger pressure tends to generate its associated emotional feeling. At the same time, this feeling is expressed through the pressure in a way that is satisfying for the body. You can press with one or more fingers or even with the whole hand against the surface on which the fingers or hand rest.

Negative feelings are best generated and expressed with the fingers pointing straight down to produce a kind of stabbing muscle response. Positive feelings are more effectively produced if the lightly cupped hand rests on the surface and the pressure is mainly expressed with the fingertips. Both hands may be used simultaneously.

To do these exercises, keep the executing arm in a relaxed position. If you sit on a chair, the jabbing pressure can be produced on the seat of a second chair or even on your thigh. For the whole hand to rest on a surface, use a table, your thigh, your other arm, or your chest. Eyes can be open or closed.

The pressure patterns of the seven key emotions are as follows:

- Anger: Make a jabbing movement away from the body, like hitting out. The duration is much less than a second and is accompanied by a sharp exhalation and a corresponding sound; the gaze is slightly downwards.

- Hate: The pressure is stronger than in anger, but develops more and is sustained somewhat longer; it ends abruptly, as in anger. The direction is away from the body and occurs during exhalation.
- Grief: This can be expressed with the cupped hand. The finger pressure is straight down and increases slowly for half a second. There follows a period of passive weakness and immobility during which the pressure gradually eases. The head can be tilted forward and to the right (for right-handers). Breathing is slow and shallow, pausing after the exhalation. The whole action lasts three to five seconds.
- Love: Pressure is gently towards the body, increasing and relaxing very slowly in a deep, smooth movement. Exhale slowly with a sensual sound; the head remains level.
- Sexual arousal: Push down and slightly away from the body; release quickly, but finish with a gentle pressure towards the body. Breathing is rapid, panting-like puffs during exhalation; the head is slightly down. The action is longer than for hate but shorter than for love.
- Joy: A quick moderate pressure straight down, immediately followed by an upwards bounce with less finger pressure on the surface than before the action, resulting in a floating feeling until the pressure gradually increases again to normal. It is like jumping for joy and floating down. The action is best performed during inhalation, head and gaze slightly raised.
- Reverence: Very gently increase pressure nearly vertical or slightly away from the body; very slowly return to normal pressure in a smooth movement without any body tensions. Breathing is extremely slow with pause after inhalation; head and gaze are slightly up.

To increase your ability to feel and express your emotions, practice these exercises for about 30 minutes daily. This will help release the emotional tension that has accumulated during a lifetime and enable you to respond in an appropriately feeling way to present and future emotional challenges.

Preferably start with a non-emotional expression. While relaxed, repeatedly press straight down, with one or more fingers or the whole hand, in a mechanical way, like hitting a typewriter key. Then express each of the emotions just listed approximately 30 times. The emotion tends to build up gradually for several minutes and then reach a plateau before easing off again. Follow the order as listed, starting with anger and finishing with reverence.

The emotional buildup is strongest if there is a certain time interval between individual actions. This interval is different for each emotion. For anger, Clynes found this to be 4.8 seconds; for hate, 5.3; grief, 8.2; love, 7.4; sex, 4.9; joy, 5.2; and reverence, 9.8 seconds. There may be timing tapes commercially available; otherwise, you can initially experiment with the timing and develop a feeling for it.

Later, when you feel free of the desire to express negative emotions or you do not derive any more benefits from it, continue with those exercises you like. Sentic exercises can be combined with other feeling exercises. Such exercises can show you that it is not necessary to be a victim of unpleasant feelings. You do not need to wait for positive outside influences to produce pleasant feelings within you. Instead, with continued practice, you can choose how you want to feel at any given moment, how you want to respond emotionally in any situation.

Exercises for Coming Alive: When we are conscious only in the head, without being permeated with our feelings, it is almost as if the body were not really alive. It is then more like a machine, a mechanical device operated by the brain. This has been the lifelong condition of most of us, so we do not know what we are missing until the body gradually awakens, vibrating with life-giving feelings and not just pain.

We can grow even further, learning to choose our feelings, not just during an exercise, but during our daily activities. In order to feel good, to feel love or joy, we do not need to wait for the rare occasions when others or circumstances temporarily ignite such uplifting feelings within us. We are not our feelings and emotions; instead these are energies we can use and enjoy according to our needs and preferences.

The three exercises described below can gradually be extended in a way that the generated feelings become habits in our daily lives instead of isolated instances of feeling good. If you do not already feel alive and vibrating before starting a feeling exercise, energize your body for a few minutes with deep breathing then shaking, especially of the arms and head.

MIRROR: Sit or stand in front of a big mirror and smile at yourself. Look lovingly into your eyes and talk to your mirror image in a soft voice. You can say something endearing, such as: “I see love shining out of your eyes. It comes from your heart. Your heart is filled with love. It shines out of your eyes.”

Repeat this over and over again, slowly, with all the feeling you can express. Pause and smile at yourself; smile with your eyes, a gentle, loving smile. While continuing with the mirror exercise, you can also use sentic hand pressure to increase the feeling in your heart centre. Sense how it feels to make passes with your hands around your head and body, slowly moving downwards with palms facing the body. Instead of addressing the mirror image as another person, say “my” instead of “your” during your endearments. Experiment to see what feels best.

As an extension of this, when you have the opportunity, practice with a partner or friend making eye contact and smiling at each other. Take turns lovingly praising each other.

MUSIC: Music is excellent for inducing feelings. For heart feelings, use soft music with an uplifting, melting quality. This can accompany any of the other feeling exercises. You can also use it during heart-feeling meditations while being quiet or during repetitive activities; at other times, select stirring or uplifting music according to the feeling you want to amplify.

Jazz can be useful for raising the energies, especially in the morning and during shaking exercises. Some blues are good for inducing tender feelings. Rock music, on the other hand, is usually weakening to all muscle systems and should be avoided. This applies also to any music that conveys chaos and disharmony instead of harmony.

There is now a wide range of New Age music available. Most of it is designed for relaxation with sounds of waves and brooks, and some is suitable to generate an atmosphere of reverence and emotional uplift. In most instances, however, it provides just a pleasant, unobtrusive background for meditation and feeling exercises. It would be excellent if a series of tapes could be obtained for inducing various defined states of feeling.

DANCE: Dancing is a moving expression of our feelings. We can even dance without music if we want to express a feeling. However, in order to generate feelings, it is preferable to dance to suitable music. You do not need to move your legs when dancing; you can dance with your head and body, but mainly with your arms, even while sitting. Most expressive are the hands. Dance vigorously to lively rhythms in the morning to stimulate the energy flow. In the evening, as a prelude to meditation, dance softly to uplifting melodies. Dance often, simply to feel and express the joy of being alive. Try to express the various feelings through dancing, such as anger, sadness, compassion, joy, love, ecstasy, and so forth.

* * * * * * * * * * * *

Step 61

TAKE THE LOVE CURE

Lead a radiant life filled with love and joy, and practice the Love Cure.

Love may be the missing factor for our emotional well-being and for overcoming disease. Love is the opposite of fear. With cancer, for instance, psychological therapy has been shown to have a far greater success rate than any other properly evaluated therapy, while fear has a major impact on cancer and AIDS patients. There is also a close relationship between fear, often present subconsciously, and breathing problems, as well as heart disease.

Emotional and psychological factors may be why various therapies work with some patients but not with others. Releasing past emotional hurts and resentments is helpful with any kind of healing, but it may not be enough if positive feelings are missing. The ideal is a combination of appropriate therapies on other levels and emotional therapy.

After many years of working and experimenting with a wide range of therapies, I conclude that some of the most important factors in restoring or improving our health and well-being are love and joy. For our emotional body, regular doses of these and other positive feelings are as essential as vitamins for the physical body; otherwise, we develop “emotional deficiency” symptoms, which can eventually disable biological functions and lead to illness.

Most of us can remember a time when we were madly in love. Love, then, was not just a mental concept, but a bodily sensation and it made us feel invincible. I believe that we can become invincible like that in regard to disease by learning to feel this way again as a way of life. Love and joy can rejuvenate the body and can be more effective than even the best nutritional supplements.

Learning to Love Again: Obviously it is not possible to fall in love with someone on a regular basis, but it is possible to experience the feeling of love as a conscious exercise. Love has traditionally been associated with the heart. Of course, it is not the heart muscle itself that makes us feel love, but the heart chakra centred at the middle of the breastbone. In order to feel love and increase our capacity to love, we need to stimulate this heart centre by focusing our attention on it. Diseases in the chest area should respond readily to this therapy and this applies especially to breast cancer, which has a strong emotional and relationship component.

For most of us, learning to feel love as a way of life does not come easily and requires a commitment to an exercise program. It will be especially difficult for most males, while many women will take to it quite naturally. Below you will find some useful exercises for this. Experiment and continue to practice those that you find best to generate love or joy in your body.

Feeling is energy and energy follows thought or consciousness. This means the most basic requirement for activating our heart centre is to keep our attention focused on it. Later, when it is easy for us to feel, we can just keep a small part of our attention there while going about our daily business. Before using these exercises, practice the chakra exercises earlier outlined in Step 52.

1. Breathing into the Heart: Becoming conscious of your breath is one of the easiest ways to generate a feeling sensation inside your body and to keep your attention focused. Sit or lie quietly and mentally follow your breath as it gently moves in and out. Breathe through the nose. Become aware that the out-flowing breath is somewhat warmer than the incoming breath. After a while, begin to pause for a second or two between inhalation and exhalation. Feel a light pressure in the middle of your chest during this pause. When you can distinctly feel it, focus again on feeling the warmth of the outgoing breath.

Now combine this warmth with the light pressure felt in the middle of the chest. This is the location of the heart centre. Keep your attention focused on it and gradually intensify the feeling of warmth. Spread it all over the chest, but especially towards the location of any health problem, such as a breast tumor. When you feel the warmth, disregard the pressure. This warmth inside your chest is very pleasant and makes you feel cosy, cared for, and protected.

Intensify any of these feelings if they arise. Try to create them deliberately by using your imagination and recalling other situations when you felt this way, possibly when relaxing in a warm bath while listening to soft music. You may find that you can perceive feeling sensations more easily and with greater intensity if you exhale very slowly and gently.

When you can do this exercise successfully, you can also use it to keep your attention on the heart centre during your daily activities. You can quickly learn to keep part of your attention focused inside your chest while performing routine tasks, although you are likely to forget it during more demanding mental activities. That does not matter. If you ask your subconscious mind to remember doing the exercise whenever conveniently possible, it will eventually oblige and may tell you a hundred times a day to refocus. If you are somewhat tired, it may even be easier to generate a feeling sensation during activities rather than by lying down and closing your eyes, which may cause you to drift off to sleep.

If it helps you to focus, imagine your incoming breath being rose-colored or bright green. You may see this color intensify in the chest area with each successive breath. Another possibility is to put one hand on the chest to feel a light pressure or to press lightly with one or several fingers at the middle of the breastbone. Experiment with a different breath imagination. Imagine your breath moving through the middle of the breastbone, and again, feel the slight pressure and then the developing warmth. For some, this exercise produces stronger sensations. When you are somewhat tired, it will be better to do these exercises with your eyes open and possibly even while walking slowly.

2. Feeling Love: After successfully practicing the breathing exercise for some time, experiment with the following suggestions to see which one suits you best. Eventually, you may use all of them at one time or another, depending on opportunities.

It will be easier to feel a love sensation if you are already in a loving relationship, a mother with a baby, or an animal lover with puppies or kittens to care for. If that is the case, look lovingly at the chosen object while doing the breathing exercise, and you may find it easy to convert the warmth in your chest into a feeling sensation of love. If you are not so lucky, recall an incident when you felt love. Vividly imagine this time and then intensify the sensation of love when it arises; then forget the actual incident and only hold fast to that feeling inside your chest.

Look lovingly at something grand or beautiful: a tree, flower, picture, clouds, the sky, or a person. Soak in the beauty and appreciate it. Then send that feeling of appreciation back to the adored object; all the time, keep focusing your attention on the breath in the chest centre and intensify any arising positive feeling. You can do the same by listening to some soul-stirring music. Eventually, you can start feeling and sending a silent appreciation of love to anyone you come in contact with and over a distance also to someone with whom you have a disagreement or a past problem. When you tend to plants, watering, transplanting, or sowing them, project onto them your feeling of love or appreciation. You can do this anytime in any activity; your only problem will be to remind yourself to do it.

If you have difficulty generating a feeling sensation of love, then feel compassion in the heart chakra. Look at or imagine pictures of suffering individuals. When you can feel and radiate compassion, try to change it to a feeling of love or a combination of love and compassion. The feelings of love and compassion can be combined with a feeling of joy, which can also be experienced in the heart centre, as well as in the forehead centre.

Or try this: Whenever you meet someone or think of someone, smile and silently say, “God bless you,” while feeling a wave of love or compassion radiating from your heart centre. Or think of someone and extend your arms with palms forward. Again, smile and feel the love or compassion in your heart. During the day, try to remember this feeling of love in your heart and centre on it.

3. Feeling Joy: Joy is the twin of love in helping us to live a healthy and fulfilled life. As with love, it is not easy to feel joy as a body sensation, especially if we do not have anything to be joyful about. There is an area of the body where we can more easily generate and feel joy and even bliss: the pituitary gland behind the middle of the forehead. Concentrating on it during meditation can generate a feeling of bliss; this can be felt as joy when focusing on the pituitary gland or forehead centre during a breathing exercise.

As the breath does not flow into the brain, imagine the prana is separated from the inhaled air and flows upwards from the nose into the pituitary gland. If you can visualize colours, see the forehead centre as filled with indigo-coloured prana.

First try to feel a light pressure partly due to the focus of your attention and partly to the accumulating prana. Again try to feel warmth or tingling during the slow and gentle exhalation. To focus on the correct point, it helps greatly if you close your eyes and turn them inwards and upwards as if looking into the brain at a 45 degree angle. When you can easily do this, you may even be able to just imagine looking upwards and inwards and feeling a light pressure in the pituitary area while you have your eyes open and are engaged in other activities.

When you are able to feel the pressure or warmth, start smiling inwardly and direct this smile to the pituitary gland. As you keep smiling, you may gradually feel a sensation of joy and happiness spreading from your head into your body. If you have already learned to feel a love sensation in your heart centre, combine this love feeling with the joy spreading from your head into a unified feeling of elation.

4. Feeling Connected and Peaceful: We can learn to overcome the stresses of daily life and live instead in a condition of profound inner peace by frequently focusing our attention at the top of our head. This will be invaluable with all stress-related diseases and conditions and also helps to energize the body in a pleasant way.

Relax, close your eyes, and imagine a stream of white light entering the top of your head. Try to feel as well as see it. However, feeling it is more important than only seeing it. It may help at first to focus on feeling a light pressure and warmth at the top of the head, possibly aided by leading the breath prana to this point. Exhale very slowly and gently, and imagine feeling peaceful, a deep peace beginning to fill and surround your body and mind. Watch for a feeling of being energetically connected to your Higher Self, the Universe, God, or whatever you want to call it. If you can feel or imagine feeling something like this, try to intensify the sensation.

When you have a feeling sensation at your crown centre, you can intensify it by generating a feeling of love in your heart centre and then sending it straight up through the top of your head to your Higher Self or God Self. Immediately feel this love offering coming back to you through the top of your head, greatly intensified. See and feel the incoming white light enlarging and intensifying to surround you now in a pillar of light. Feel how it energizes your whole body. You may also see the white light blending into other rainbow colours that are beneficial for you such as a golden-yellow energizing color, the rose color of love, or the violet flame of purification and transmutation.

When you are able to feel the peace and connectedness to your God presence during the exercise, focus part of your attention on the top of your head and feel this connection during your daily activities. Feel at peace and connected. Keep part of your attention on the top of your head while you try to feel the joy behind the forehead and the love at the heart centre. Alternatively, focus in sequence on the top of the head, the forehead, and the heart. Eventually, you may be able to focus on all three simultaneously and feel peaceful and connected, joyful and loving at the same time.

Living Your Feelings: If you have a disease or health problem in a specific organ or area of the body, you can now supply it with all the energy and emotional support that it needs for healing, in addition to the nutritional and other support that you may provide. On a regular basis, fill yourself with love and then mentally send it with an inner smile to the organ in need of healing. Breathe more energy into it or lead it there from the storage area or chakra behind the navel. Feel it getting warm or tingling or just feeling good. At the same time, see it filled with a brilliant white healing light. Intensify the vision and the feeling and hold it for a long time. Know that you are being healed.

Tender feelings are a daily requirement for your body and soul. We need them as much as we need vitamins and minerals, so make sure you obtain plenty of “emotional nutrients” each day, as there is nothing better to rejuvenate an aging body-mind. Bathe all your cells and organs in tender, loving, or joyful feelings. Focus on each part of the body in turn and concentrate longer on those areas most in need of improvement. Try to feel the love everywhere, in your fingers as well as your toes. Let all your cells feel that you love them.

The breath will help you stay in touch with your body. Make it a habit to breathe into your heart centre, and keep part of your attention there as often as possible. Alternatively, keep part of your attention focused on the crown or forehead centre. Better still, let the peace and connectedness of the crown centre fuse with the joy of the forehead centre and the love of the heart centre to a symphony of uplifting feelings.

Do not let your mind drift idly to produce repetitive mental chatter. Whenever you notice that the internal chatter has started again, renew your focus on the heart centre. Daydreaming, however, can be restorative and creative and is often superior to mental chatter.

When you are tired or exhausted, it is difficult to arouse feelings even if you try hard, but at other times gentle feelings may spontaneously arise and float through the body. This happens most easily when we are aware of something beautiful, something which uplifts the soul, a landscape, a painting, music, flowers, a smile. You can then magnify your feeling by radiating it back onto the landscape, the painting, and so forth.

When you massage someone, greet someone with an embrace, whenever you touch someone - always remember to radiate loving feelings. It is much more difficult to maintain and radiate positive feelings during conversations, but this will gradually improve. You will find it a great help if you can remain relaxed and smile instead of becoming totally absorbed by the mental aspects of the conversation.

However, it is most difficult to retain any kind of feeling awareness during intense mental activity. Usually, we completely forget the body. This will also gradually improve, especially if you can train your lower self to automatically switch back to the heart and crown awareness when the mental challenge has passed.

Simplify your life and carry an inner smile with you wherever you are. See the beauty in everything. Gradually, on rare occasions at first, a feeling of deep inner peace may arise, clothed in a feeling of being loved and cared for. If you persist with your attempts to remain focused on the heart and crown centres, then this deep inner peace will intensify and begin to permeate your whole being.

Out of this tangible feeling of inner peace will then arise the other feelings, thoughts, and activities as required and appropriate for each occasion.

* * * * * * * * * * * *

Part 9

HEALING YOUR MIND

We all would like to live a happy, healthy, and fulfilled life, yet few seem to be Our conscious decisions are made with our mind. All healing and improvement in our living conditions start at this level, and it is here that we must take the first step with a conscious decision that we want to improve conditions by following a suitable program. If we see the progression of our lives as a creative process, then we can see the mind as the architect.

In our society, rational thinking and the intellect are worshipped, while feeling and intuition are neglected. The reason for this imbalance can be found in the distorted goals and ideals of our society, which has lost sight of inner values. Wisdom, devotion, patience, and compassion are replaced by material riches; inner purity is forsaken and traded in for excessive external cleanliness and hygiene; success is measured by monetary wealth and dominance instead of self-mastery.

Thinking is the most important tool for achieving such external success, while tender feelings disadvantage us for success in our society. However, thinking is neutral and we can use it to our advantage in healing our body and emotions. Proper thinking is required to discard unsuitable beliefs and replace them with appropriate ones. If we accept that negative thoughts and beliefs are a cause of our illness and social problems, then our mind is in need of healing and we must replace disease-forming negative beliefs with health-giving positive ones.

I regard the Self as our total nonphysical entity. The Self and the body together are the whole entity. To better understand inner processes of consciousness, we can distinguish between different parts of our Self. I use a fourfold division:

- The higher Self or super-conscious level provides spiritual guidance as inner knowledge and intuition. It has further divisions, variously called soul, Christ Self, over-soul, or God Self.
- The middle self represents our normal consciousness, that of which we are aware, mainly our mind and mental body, the thinker.
- The lower self or subconscious part is the master of our memories and emotions, including long-forgotten beliefs and a Pandora’s box of suppressed feelings and emotions; it is also called the “inner child.”
- The body self is an elemental (a life-force being) that looks after the biological functions of our body; this is the level of consciousness that is still active when we are in a deep coma.

We can say that the Higher Self operates at the spiritual level, the middle self at the mental level, the lower self at the astral or emotional level, and the body self at the etheric or life-force level. The middle self can to some degree become aware of the other levels, but its centre of consciousness is based at the mental level.

Depending on our degree of spiritual understanding, this may be the concrete mind or lower mental level; it may be the higher mental level concerned with spiritual, creative, and intuitive thinking; or it may be somewhere in between.

* * * * * * * * * * * *

Step 62

HARNESS THE POWER OF THE MIND

Appreciate the power of your subconscious mind.
It is the real power behind the throne of the ego.

Our mind has awesome powers. It makes us sick and unhappy or healthy and happy. It is a powerful tool and it is up to us in which way we want to use it. In clinical trials, mind therapies have been shown to be much more effective than conventional therapies in reversing many cancers. To illustrate the overriding importance of the mind, I relate an example.

An elderly male with huge tumor masses all through his body had been given less than three months to live. At that time, trials with the alternative remedy Krebiozen were started. Despite not being eligible to participate because of his short life-expectancy, his enthusiasm about the remedy was so overwhelming to his doctor that he gave him the remedy outside the trial.

A few days after his first injections, his tumours had already halved in size and continued to reduce until they had virtually disappeared and the man felt healthy and well. However, two months later, he read a report in a newspaper that the trial was a failure and Krebiozen was useless. Immediately, he fell ill and relapsed, with his tumours quickly regrowing to their previous size.

His doctor was so astonished by these strong responses that he decided to make an experiment. He told the patient that the first trial had not worked because the remedy was too old. Now he was waiting for a fresh supply that would be double-strength. A few days later the man was given the injection; this time the tumours disappeared even faster than the first time and again the patient was completely healthy and well.

However, instead of any drug, the doctor had just injected water. Again, a few months passed and then the newspapers carried stories that the FDA had declared Krebiozen to be completely useless and a fraud. As rapidly as he had recovered, the patient deteriorated once more and this time his doctor let him die.[70]

One might think that such a strong mind effect must be very rare, but consider this: In a chemotherapy trial, one-third of the placebo patients lost their hair. In this rare trial, only half the participants had received chemotherapy and the other half a harmless substance, which they believed to be an active drug. This means that the hair loss in one-third of the placebo patients was entirely due to their belief.

This finding is confirmed by one of those rare placebo trials in surgery. After “real” operations for cardiovascular disease, 32 percent of patients had satisfactory results. However, 43 percent of a control group that had only pretend surgery reported subjective and objective improvement! This means that basically all the improvement came from their belief, but the traumatic effect of real surgery reduced the belief-based success by 11 percent.

One-third is a figure that corresponds with results from other trials as the approximate size of the placebo effect. I take this to mean that typically one-third of patients will either die or recover because of their strong belief. This applies equally to conventional medicine and natural therapy. Those who believe they will get well will do so regardless of the therapy used, and those who doubt or are fearful will receive little benefit from any therapy. I am convinced that this is the main reason why one patient is cured by cancer surgery, while another with the same condition dies afterwards.

Therefore, perhaps the main concern for patients with “incurable” diseases should not be to find an effective therapy, but rather to find a way to harness the power of the mind and become a believer with an unshakable faith - in effect, to put the placebo effect to work.

Before you decide to believe in eating a piece of cake every Sunday as your preferred cancer cure, there is another factor to take into account. Whatever you do must be totally convincing to your subconscious mind. Just believing something on the conscious mental level is not enough; if your subconscious mind is not convinced that something will cure you, then it will not. The subconscious mind is most easily impressed by something that inspires awe, by authority, ritual, the unusual, and especially enthusiasm. Whatever impresses you as a potential cure, embrace it enthusiastically. Anything that you just try to see if it works is not likely to work.

Get to Know the Lower Self: In order to heal our mind and use it more efficiently, we need to have a basic understanding of our lower emotional self. This operates on a different level of consciousness from our normal consciousness (the middle self) and has many of the characteristics of a separate entity. When working with it, it is best to regard and treat it as a reliable servant.

It is advisable to be on good terms with our lower self because it is the power broker of the body. The middle self is like the captain of a ship, the Higher Self is the shipping line and employer of the captain, and the lower self is like the union foreman. If the crew has any grievances, the boss may call a strike or otherwise make life difficult for the captain. The body self is comparable to the crew and cooperates very closely with the lower self.

Here’s what we can say about the lower self and its body functions:

1. Influences the body self and the physical body (except for the voluntary muscles)
2. Is the seat of our emotions
3. Receives all sense information
4. Keeps a record of all sense information, thoughts, and feelings
5. Operates our memory bank
6. Operates our psychic abilities
7. Is the means of communication with our Higher Self

The real power of the body rests with the lower self, but it has one shortcoming: It cannot think logically. It has only an elementary ability to think, similar to an intelligent domesticated animal or a small child. Therefore, it is dependent on the middle self to tell it what to do and, as the faithful servant that it normally is, it willingly obeys the middle self.

However, because it is simpleminded, the lower self is dogmatic. It may take our early childhood programming, especially our religious and sex-related teachings, and anything someone in authority may have said, as gospel truth. It can be moralistic and regard itself as the keeper of any important commitments that the middle self may have made. To illustrate these characteristics of the lower self, here are a few examples from The Secret Science behind Miracles, by Max Freedom Long.

A young man with a strong religious upbringing had the urge to enter the ministry. He took a job in a furniture factory instead. There the paint fumes made him sick. When transferred to the woodworking department, he developed asthma from the sawdust. He tried several other occupations, but each time he became sick from something connected with the job. In the end, a psychologist unearthed his earlier intention to become a minister. As nothing else seemed to help, he advised the man to enter the ministry. The man did and had no more illness.

A religious Hawaiian man had an affair. His wife found out but forgave him. Within a year it happened again, but this time his wife did not find out. This seemed to make it even worse for the man who gradually lost his strength and will to live until he was very close to dying. A kahuna (native healer) was called and discovered the truth. He induced the wife to forgive her husband once more and the kahuna performed a ceremony in which all the man's sins were washed away. The patient immediately started a quick recovery.

A young woman had been brought up to regard dancing and drinking alcohol as sins. After her marriage she moved in different circles and gradually started dancing and having the occasional cocktail. Soon she twisted her ankle slightly during a dance. Normally, she would have been fine after a day or two, but instead the leg became gradually worse and a deep running sore developed below the ankle. A kahuna convinced her that she had not really sinned against God and in addition forgave her and washed away all guilt of any kind. The ankle quickly recovered to its full strength.

However, the young woman neglected the kahuna's instructions to continue affirming that if she had not hurt anyone, there was no sin. Again she danced and drank a little and eventually the ankle sore reappeared. However, this time the kahuna explained that he could not help, as the lower self had now become too much entrenched in its conviction of sin. The only permanent cure was to give up drinking and dancing for good.

These examples show how our lower self holds on to childhood beliefs that as adults we have long forgotten. In contrast, the lower self never forgets. Therefore, if we suspect a conflict between the beliefs of these two selves, we have to convince the lower self that the old belief is no longer appropriate and patiently and in simple words explain the new situation instead.

The Power of Beliefs: We see and interpret the world through the coloured glasses of our belief systems. We even create our personal world according to our beliefs. We project selected ideas, thoughts, and emotions into our surroundings and with these attract or repel people and events. Most notably, we manifest what we fear. Negative thoughts harm no one more than us. When a negative thought comes into our mind, we can either accept it or immediately block it and replace it with a positive thought.

It is advisable to begin our healing and growth venture by looking closely at our basic beliefs to see if we still want them. Do they present a positive outlook and are they in harmony with our goals and our new philosophy of life? One way of finding out is to write down your beliefs about the basic questions and problems of your life. What do you believe about yourself, your friends, your work, your health problems, your body and your future, about death, Afterlife, and so forth. Then cross out whatever does not meet your new standards and rewrite in a more positive form what you would prefer to believe.

As an example, you may believe that most people are selfish. As long as you believe this, you will attract mostly selfish people or interpret their motives as being selfish, which may not always be true. Therefore, start reprogramming yourself by writing: “I attract more and more unselfish people into my life.”

Instead of believing “I will become infirm with advancing age,” you can now write: “My health and well-being are improving with every day in every way.” Yet just writing all this down is not enough, because by now the lower self will share many of the beliefs of the middle self, and the lower self is very stubborn. Most effective for influencing the lower self are autosuggestion methods, such as affirmation, self-hypnosis, and visualization, as well as studying and gradually accepting a truly spiritual philosophy of life.

However, even if you are successful in changing your beliefs, that is not the end of the road. It is just a precondition to attracting the necessary knowledge or conditions that will make an improvement possible. You still have to do whatever is necessary on one of the other levels, be it a change in diet or asking for forgiveness and sending kind feelings to someone you have hurt. Be careful with psychic predictions about yourself. Some are amazingly true in minute details; however, others can be wide of the mark. Therefore, believe only in those predictions that you would like to come true and ignore the rest. Choose your beliefs carefully because they determine what will happen to you.

Worry, Resentment, and Desire: Three of the most common obstacles to good health, inner peace, and happiness are worry, resentment, and desire.
Worry is an outcrop of fear and is part of the price we pay for a lack of faith in our guidance. For someone who is in harmony, in tune with the Higher Self, there is nothing to worry about, neither from the past, nor in the present, nor in the future.

Resentment is an expression of ego-centeredness and results from a lack of self-responsibility. We blame others for what happens to us and we try to make them pay for it by withholding our love and affection from them. A change of attitude towards self-responsibility will free us from the self-destructive effects of resentment and enable us to become tolerant and understanding of our own shortcomings as well as those of others. If someone does something that we do not like, then we should either speak out or accept and forget it. Resentment is the worst of our choices.

Desire tends to rob us of our peace of mind and destroy our inner harmony. We must, of course, tend to the needs of our body. A desire for food when we are hungry, for rest when we are tired, for sunshine, fresh air, and so forth, are natural. It is also good to desire being creative. In addition to these natural desires, most of us have numerous desires related to pleasure and social achievements.

Even such materialistic desires are no problem, provided we are happy to pay the price for their fulfilment, and usually they help us to learn or grow. Still, the more we improve our belief systems, the more the need for desire fulfilment and achievements will disappear and the easier it will be to feel peace and harmony. Actually, it is not so much the desire itself that causes us trouble, but our attachment to the outcome of our actions. It is excellent to have the desire to be creative, but if instead we focus our desire in the form of expectations on the result of our creativity, then we are heading for disappointment.

Even our path towards better health, fulfilment, and happiness will be more successful if we do not have too great a desire to achieve this goal quickly. Instead, we can walk this path for the sake of the pleasure and satisfaction provided by the path itself, by being creative, playful, having realizations, and doing something meaningful. Generally, the fun is in doing, in action, rather than in the achieved result.

When, for some reason, we are confronted with a difficult task, we must channel all our energies into this purpose and have a “one-track mind” in our desire to succeed. However, as long as we want something badly for mainly selfish reasons, we may not get it, or if we do, the price may be too high and we may not appreciate it any more.

We can overcome this obstacle by doing what needs to be done or acting for the good of our family, community, nation, or for mankind, without desiring the fruit of our actions for ourselves. In this way, we can never be disappointed and always get our reward from the pleasure it gives to do well what needs to be done. Remain aware that we are just actors in a play, with Divine guidance as the director, and whatever the outcome of the play, all is well.

Our world covers every view and every condition possible under the given circumstances. Due to our cultural, social, and personal programming, we select a few views and conditions out of all that is. We focus on these, identify with them, and move towards manifesting and even becoming them ourselves.

What we call good and evil, happiness and suffering, love and hate, joy and despair, are part of this world. From these we select, usually subconsciously, that with which we resonate; we attract it and then manifest it in our lives. However, we can consciously influence this internal selection process and focus intentionally on what we want to manifest in our lives or how we want to become. We do this by maintaining a positive mental attitude or expectation.

Develop an attitude that remains in all situations positive, playful, tolerant, kind, and humorous, laughing more about you and less about others. In no way, however, should we just pretend these attitudes when we feel differently inside; always express or acknowledge in a suitable manner whatever you feel.

It is okay to feel angry, upset, or sad. Go into it. However, in the back of your mind, remain aware that you can change the situation and laugh any time you want. The more you learn to take yourself less seriously, the easier it will be to laugh about your own mistakes, until it simply becomes impossible to be angry or upset. Even annoyances created by others may eventually cause you amusement instead of ulcers. It is all a question of attitude.

Fear of Death: The fear of death and dying is widespread in our society, due mainly to a materialistic belief system. Early childhood programming about eternal punishment in hell may still have an influence on some of us. Most of the time, this death fear is kept out of our conscious mind, unless the death of a close friend or relative brings it to the surface. At other times, it can manifest as an inner insecurity, a fear of disease, a general anxiety, or inner emptiness. Mourning, in this condition, is more or less an expression of self-pity, although very helpful and necessary with this belief system.

We can free our lives from all these negative influences by changing our belief about death and the Afterlife. If you follow an established religion, you can actively program yourself to believe more strongly in the positive aspects of the Afterlife as taught by your religion.

If you do not follow a specific religion, you can get acquainted with some books on near-death experiences or reincarnation, especially with studies using hypnotic regression. However, do not look for "proof." There is no concrete proof available in spiritual matters that will convince everyone. To our normal consciousness, we accept the spiritual world on faith and belief. Of two people confronted with the same facts, one may believe in them and the other reject them. You can choose what you want to believe.

After starting to believe in a spiritual truth, be it reincarnation, karma, divine guidance, or the Afterlife, you may find more evidence to reinforce your belief. With this, an inner knowledge will grow, which is the only true form of knowledge. Eventually, you will know from the experience of your own life and will not be interested anymore in the "proof" someone else may be able to offer. Nevertheless, it may help you to review the book "A Lawyer Presents the Case for the Afterlife" (available at: www.victorzammit.com). In it, Victor Zammit, a retired attorney of the High Court of Australia, has assembled all the known evidence of survival after death. He came to the conclusion that before a court of law, the evidence taken as a whole would constitute overwhelming evidence and irrefutable proof of the existence of the Afterlife.

The acceptance of the Afterlife as a reality may remove the fear of death in our society and reduce the mourning, sadness, and loss that we feel at the passing of a loved one. Instead, it may become a celebration, in which we congratulate the departing ones on a job well done and in which we are happy that they can now receive their well-earned rest and reward and go on to greater things. We know that we will catch up with them sooner or later and that it will be a great reunion.

I have in my mind the picture of a large group of deep-sea divers working under difficult conditions in heavy steel suits at the bottom of the ocean. Their technology is such that they can stay below for years. From time to time, one of the divers finishes his assignment and can come back to the surface. The other divers are envious and would like to accompany him. However, it is also important to them that they finish their own assignments, so they give a little party for their lucky colleague, congratulate him, wish him well, and arrange for another bigger party when they are all back together at the surface. When the successful diver finally reaches the surface, there is another big celebration with all his friends who had remained at the surface or had returned before him. What a relief getting out of that heavy diving gear after all those difficult years and seeing the sun again!

This is how I imagine the soul must feel when it can finally discard the body after a hard life. It is understandable to feel sad at being left behind by a departing loved one, yet the anticipation of the eventual reunion should make it easier to bear. Even in the case of a young child dying apparently without purpose, the departed soul may soon try to incarnate again, perhaps even with the same mother.

As we spiritually mature, we will come to know our time of departure in advance, and we will be able to depart without requiring a disease to free our soul. With this, we will also stop our efforts to prolong our life, to live as long as possible. We will be happy to leave when we have finished the job that we came to do or had the experiences we wanted.

So with this in mind, reprogram yourself into believing that you are only an actor in a temporary role in your present life and that other roles in different lives and perhaps different worlds are part of your greater existence. Then it may be much easier to take it easy, to be playful about serious matters, enjoy your role in this life, and to look at your problems as opportunities to experiment and see what comes of them.

Self-Responsibility: By aiming for health improvement on all levels of our existence as a conscious expression of our attitudes and beliefs, we show that we are prepared to take responsibility for our life. It is common to blame someone or something else for our problems, yet it rarely occurs to us to look to ourselves for the source of our difficulties. We eat unhealthy food, create harmful emotions, adopt negative attitudes, and when this makes us sick, we blame it on outside influences and want someone else to fix it.

By accepting self-responsibility, we realize that there is never anyone or anything else responsible for our problems, only us. When we become sick or have an accident, we do not blame fate or germs, but understand that we have transgressed a law of nature or moved against our life blueprint. We try to discover what we did wrong and we correct the situation and guard against making this mistake again. If our disability is of a nature that it cannot be healed, then we accept it cheerfully as a test given to us for the good of our soul and we make the best of what we have.

It is similar with social, marital, or financial problems. We assume that it is the essential purpose of our existence to learn the spiritual and biological laws, to direct our feelings and emotions into beneficial patterns, and to make the right decision at the right time in the right place. Every problem we encounter offers us the opportunity to learn. The way we learn most is by making mistakes and then becoming aware of them. This is how we explore our limits. When we have learned our lessons, we will be able to handle life’s problems without difficulty.

By adopting this principle of self-responsibility, we also adopt a position of power. We are no longer victims of anyone or anything, but rather the masters of our destiny.

We realize that we have to learn anyway, by the long, hard road of suffering caused by our ignorance or by an attitude of willing cooperation in our scheme of existence. The road of self-responsibility is so much more pleasant than the road of suffering. It is an exciting venture in self-realization. Every unpleasant happening or health problem not only mirrors what we have to learn but also shows us the solution - if and when we are able to see.

Self-responsibility in regard to health care does not mean that we should rely entirely on ourselves without outside help. Helping each other and learning to cooperate are important parts of self-responsibility; we can seek the services of a competent healer whenever that feels right. Our attitude towards a health professional need not be passive submission to authority; it can be helpful and optimistic cooperation in a joint venture in healing in which we retain the overall responsibility. Expect your healer to help you, but not to do your work; you still need to bring about the necessary inner and outer changes.

While the ultimate decision about our healing is made at or above our soul level, we have to accept self-responsibility to prepare the way and remove any obstacles for healing to become possible. I see the relationship between our Higher Self and our personality as that between parent and child. While the child cannot make the decision where to go on holiday, the parents will take the wishes of the child into consideration. If the parents are generally pleased with the child and there are no important obstacles, then the parents are likely to agree with the wishes of the child. In the case of a so-called terminal disease, the Higher Self may grant the wish of the person to be healed and retain the present body, especially if that may lead to further spiritual growth.

Changing Our Attitude: Our attitude is the key to travelling the long road from suffering to happiness. Our attitude is our mental disposition towards persons and events as well as towards ourselves. It is greatly influenced by our belief system and our subconscious programming. It being close to the surface of our consciousness, we can easily observe our attitude and work on our shortcomings. This provides an excellent opportunity for inner change. Monitor your attitude and correct it instantly whenever it does not meet your new standards.

For example, we may have the habit of evaluating passing strangers in a negative way. Looking at someone, we might think, “Her legs are much too fat,” or, “He has mean eyes.” Whenever you notice yourself doing this, think of something positive about this person or just send an inner greeting of love and peace.

You do not deny that there are negative factors in others - as seen from your point of view - but you just ignore them. Everything has two sides, which we may call positive and negative. The positive side is in agreement with our own values and the negative side is in disagreement. We tend to manifest what we concentrate on. Therefore, it is best for us to concentrate on the positive factors whenever possible and take the negative aspects into account only when we are required to do so for specific reasons.

The same applies to negative attitudes in an intimate relationship. You may think on occasions that your partner is unreasonable. When this happens, try to understand his or her position and, like a good defence lawyer, consider all the factors in your partner’s favour. He or she may not have received much love and security during childhood or had to fight hard for his or her rights, and so forth. Do this as a conscious mental exercise whenever you have negative thoughts about people or events.

In our relationships, be it in the business world or with friends and relatives, the correct attitude for us is one that aims for a fair deal for all concerned. We may be elated about an agreement that appears rather advantageous for ourselves, but what good will it do us if the other party is not happy with it? We will have created a source of resentment and trouble. Sensitive individuals will be aware of this resentment against them and cannot be happy about a deal that the other party regards as unfair.

When you have a mishap, breaking some crockery, dropping an open can of paint, or hurting yourself, pretend that you are an observer and try to see the funny side of it. When it is too serious to be funny, ask yourself why it happened, what it wants to teach you. Perhaps you simply need to be more attentive, perhaps it has deeper reasons; try to find out.

By training your sense of humour in this way or by looking immediately for the reason of the accident, you may not even remember to feel upset about it; though if you do, do not suppress it. Afterwards, realize that it happened to teach you a lesson.

Conscious Activity: In order to change undesirable habits and improve various activities, it is necessary to become conscious of these habits and activities. A good starting point is just to observe our mind, our thoughts, and feelings as they arise, without any value judgment in regard to right or wrong, good or bad. Have a part of your consciousness sitting in the back of your mind as an impartial observer, registering everything without a comment. If done consistently, this in itself can induce all the necessary changes to lead us towards a happy and fulfilled existence and draw us closer to full self-realization.

For those who prefer a more active way of changing the mind, various additional possibilities exist. As an example, you may discover that your favourite food causes allergy or mucus congestion and you decide to avoid it. You can either be unhappy about giving up your favourite food or you can be happy about coming closer to good health. It is similar with accidents or any other problem. It could always be better or worse. Your attitude decides whether you look at the positive or the negative side of the happening, whether you make yourself happy or unhappy. You have the choice.

We can program ourselves in such a way that something in us will call “Stop!” whenever we start performing in the old ways, and then we can consciously do it in the improved fashion. Another important aspect of mind improvement is to slow down or stop the constant silent chatter in our mind and replace it with an awareness of feelings arising in various body centres, especially in the heart centre. Aim to use the brain only when you want to think and switch off any idle and repetitive chatter at other times.

Two bodily activities usually in need of improvement are our posture and our breathing. These can be monitored until we perform them automatically in the correct way. Something we should always do consciously is eating, keeping our attention on the food flavours that develop while placidly chewing.

We cannot become aware of all of these activities at the same time. Therefore, we start with only one, for instance monitoring our attitude for several weeks until we are reasonably successful with it. Then we can concentrate on the posture changes and later on the breathing. It is different with the awareness of chewing and food flavours. As this is for a limited time span only, we can start it immediately and continue indefinitely. Concentrating on body sensations and feelings may be our final challenge.

Initially, when we become aware of a new activity, we may remember it for only a few minutes at a time and then quickly forget it again. However, when we give ourselves the autosuggestion to remember watching a habit, we will again and again be reminded of it. A hundred times a day, we may slump and then instantly remember to assume the correct posture again. Do not give up. It will soon become a new habit; then you can relax your attention on it. Gradually, become more and more aware of your whole body and everything you do.

* * * * * * * * * * *

Step 63

HELPFUL MIND TOOLS

Use relaxation, regression, reprogramming, prayer, meditation, and inner communication for a more fulfilled life.

Our thoughts, feelings, attitudes, beliefs, and emotions form the fabric of our inner world. They are the main factors that determine whether we are healthy or unhealthy, happy or unhappy. They have a direct influence on our body by changing or maintaining our inner blueprint and by directing, blocking, or enhancing energy flows and glandular activities. They also act in an indirect way by determining which foods we eat, which environmental factors we choose, which kind of knowledge we acquire and how we use it.

So to improve our condition, whether on the social, mental, or emotional level or with reference to our health, we must first change our inner world. Here are some recommended steps:

1. Examine your present condition on all levels. Make a list of what you do not like and another one of how you would like to be instead.
2. Acknowledge and release emotional blocks and past negative or unwanted conditioning. This may be the most difficult step in your improvement program because most of these problems exist on the subconscious level.
3. Use mind-control methods to reprogram your subconscious mind; participate in workshops.
4. Adopt a holistic and spiritual philosophy of life.

Mind-improvement therapy for acute emotional problems needs to be accompanied by nutritional improvements such as stabilizing the blood sugar level, avoiding allergens, and using appropriate supplements. In this way, the nutritional component of emotional problems is easily rectified and you can make much better progress with the remaining deep-seated problems. In this step, we’ll see how eight mind tools can help your health.

1. Relaxation as a Tool for Mental Well-Being: Deep relaxation is a precondition for being successful with most autosuggestion methods as well as with meditation and all other methods to contact the inner self. You can experiment with the methods listed below and see which one works best for you. Here are some general guidelines:

- Sit or lie comfortably and close your eyes. Focus your attention for ten to 20 seconds on your right foot. Feel it becoming warm and heavy. Then move to the left foot, the right calf, the left calf, and so on, moving upwards to the head. Include the chin, jaw, eyes, and the back and top of the head. When using this method regularly, the focusing time on each area can be gradually shortened, and related areas, such as a whole leg or a whole arm, can be relaxed at once.
- Imagine yourself in a red-coloured room on an upper level of a high-rise building. After a while, walk to an elevator and feel yourself going down and

down until you reach an orange room; then go further down to a yellow room, a green one, then blue, indigo, and finally violet.

- Imagine yourself walking down a woodland path to the bottom of a waterfall; feel the peace there.
- With each slow exhalation, feel yourself drifting lower and lower down a long tunnel, or imagine diving towards the bottom of the sea.
- Slowly count backwards from 50 to one. With each count, imagine yourself sinking deeper and deeper into a black void.
- Inhale for a count of eight, hold for two, exhale for eight, hold for two, and so forth. Focus your attention on your breath; notice how it quietly flows in and out. You can also focus on your slow inhalations and exhalations without paying attention to the timing.
- Combine several of these methods in one exercise to achieve deeper levels of relaxation.

AUTOGENIC TRAINING: This body-and-mind relaxation technique was developed by J. H. Schultz.[71] It requires diligent practice for several weeks or even months to master the technique, but it is worth the effort, especially for those who are tense or who have difficulty contacting their body.

Lie comfortably on your back, arms loosely alongside the body, palms down, or hands folded over the abdomen. Close your eyes and breathe slowly into the abdomen for a few minutes. At the same time, focus your mind on the formula: “I am peaceful.”

Afterwards, repeat slowly in your mind: “My right arm feels very heavy.” Feel your arm becoming limp and heavy like a leaden weight. If you are left-handed, use the left arm. After a few minutes, bend and stretch the arm energetically and repeat mentally: “My arm moves firm and free.” When you can readily feel the heaviness after several days of practice, include the other arm, then the legs and torso. At the end of each session, give the suggestion that the whole body is now light and free.

For the next step, add the focus: “My right arm is streaming warmth.” After you can easily feel the warmth in the arm, include also the other arm, the legs, and torso. The warmth does not need to be taken back at the end of a session, only the heaviness.

For someone new to meditation and relaxation exercises, it may take about three weeks until the heaviness and warmth can be readily felt together. Now you can move to the next step. Support your right elbow with a pillow so that your right hand lies over your heart. Feel it beating and hold the mental focus: “My heart is peaceful and strong.” When you can experience your heart, you can later keep the right arm in a usual position, relaxed along your side or hands folded over your abdomen.

Add the focus: “It breathes me.” Feel the breath flowing in and out without mental control. After a week or two, include the solar plexus (inside the upper abdomen) and say: “My solar plexus radiates warmth” or “My solar plexus is streaming warmth.”

While exhaling, imagine the warmth of the breath accumulating in the stomach and abdomen and heating them up. Finally affirm: “My forehead is light and cool.” You can mentally transfer the coolness felt in the throat from the incoming breath onto the forehead.

Each time, first go through all the previous steps, such as heaviness, warmth, and peaceful heart, and then focus on the next step you want to learn. Practice each new step for one or two weeks. After several months, it will be easy to relax the body completely in a very short time. Body and mind will be still and peaceful, with the body feeling warm and heavy and the head cool and light. You may be in an altered state of consciousness but should be inwardly alert and definitely not drowsy. This is an excellent preparation to do before any of the autosuggestion methods and other forms of communication with the inner self.

2. Regression Therapy: Regression therapy is a special technique of re-enactment that can be employed professionally, in a group situation, or with a partner or friend. You go backwards in time until you remember an emotionally charged incident. Professionally, regression may be induced by hypnosis, but it also works well with self-hypnosis or simply by suggestions given by a friend. However, the person leading the regression should have some experience or at least common sense in guiding emotionally distressed individuals.

Decide which emotional problem you want to work on. Even if you do not have a specific problem, you may have noticed that you have sometimes acted out of proportion in certain situations; this can be chosen as a theme. Lie down comfortably on your back, close your eyes, and relax in response to your own relaxation exercise or to given suggestions. Now your helper asks you to remember a recent emotional incident and to talk about it. He will encourage you to go deeper into your feelings in that situation while keeping verbal contact with you. Then the helper will suggest that you slide back in your memory until you encounter a similar situation or feeling in the past.

After a short pause, you will be asked to describe what you now hear or see through your inner senses. You may be asked how old you are in the remembered situation, if anyone else is present, and what is happening. The questions should be formulated in the present tense. When you stop talking or reacting for a while, your helper will gently lead you back into your feelings. Release your feelings as much as possible while reliving the past incident. Finally, the helper will suggest that you relive the situation once again, but now change it in such a way that you are happy with the outcome. You invent your own version of a happy ending and try to feel it.

Another starting point for a regression is to find an unusual phrase that you sometimes use or one that expresses an inappropriate belief about yourself, preferably in connection with an emotional problem. You may sometimes say or think, “I am too old (or too fat or too shy) to find a suitable lover,” or, “I am afraid of crowds (or cancer or confined spaces).” Then, after relaxation, repeat this phrase audibly over and over again and feel into it until a vision or memory develops, which you report to your helper. Prompting questions will be used, as with the other regression method, to lead you deeper into your emotions.

Another method, my favourite, goes as follows: Assume that this time you are the helper or therapist to regress a friend or patient. Have the friend lie relaxed on their back while you give instructions in a soothing voice. The initial aim is to induce as deep a relaxation as possible. You can give the suggestions for a normal relaxation exercise or suggest deep and slow breathing.

Then say: “Now you feel light and begin to float upwards. You rise higher and higher into the blue sky.” After reinforcing this suggestion for a while, you can say: “You look down but cannot see the Earth because of a dense cloud cover . . . You continue to fly over the clouds to look for a place and a time that hold an important clue for understanding your present condition . . . When you have found this spot, dive through the clouds and descend to Earth . . . After landing, look around and tell me what you see or hear or feel.” Pause after each suggestion, as indicated by the ellipses.

If the friend remains quiet, after a while ask what is going on. If the friend describes a landscape, ask them to find a house or some people and observe or talk to them. Always keep verbal contact. Whatever the friend reports, try to deepen the experience. Some may report vivid scenes and others just vague feelings. The related incident may be an actual forgotten experience or it may symbolize the problem.

Doing this exercise, a patient with cancer of the mouth and throat, for instance, related to me that he had landed on a beach. Some distance away he saw a group of people amicably gathered at another beach. I asked him to join the group to see what was going on, but he found it impossible to approach them. He then realized that he had felt an outsider all his life.

He was a good talker, but he talked to or at people, not with them in an open and personal way, and he had felt separated and excluded from group activities. This caused the energies in his mouth and throat area to stagnate and allowed a tumor to develop specifically in that area.

Another patient related an incident. He saw himself as a young man in a fishing boat with friends. They were drinking moderately. While all the others had fun, this patient became ill and vomited. He said that this usually happened when he drank alcohol. This shows that he had inherited a weak liver that he damaged still further by trying to keep up with his mates.

In another regression, a woman in her 80’s recalled that as a baby she was nearly suffocated by a cat lying across her face. All her life she had been afraid of cats, even if only seen on television. After I suggested that she imagine herself as the baby jerking her body and throwing off the cat, she immediately lost her fear of cats.

In many instances, memories and scenes come up that are obviously from a previous life, even if a person doesn’t believe in this. Most common is the experience of a violent death as a cause of serious present problems. Someone with Parkinson’s disease was so afraid of fire that she became rigid merely watching a fire in a movie. During regression, she relived the experience of being burned as a witch. After suggesting that she burst her ropes and run away, she was much improved. However, after some weeks, she started smelling smoke when there was none. Another regression saw her being burned to death in the great fire of London in 1666. After this, she was never again afraid of fire.

Instead of inducing the regression as indicated here, you may find other suitable methods and invent your own. To end a regression session, give some positive suggestions for coming back into the here and now. If memories become too disturbing and emotional, the session can also be ended with positive suggestions and repeated another time.

Repeated regressions can probe more deeply into subconscious problems or explore different situations. When investigating a panic-laden situation, such as a previous death experience, initially suggest the regressed person remain an impartial observer above the scene and describe it with detachment. In repeated regressions, you can then suggest becoming more and more closely involved and starting to identify emotionally with the experience.

If a strong negative experience arises, always ask the friend or patient to reinvent that experience with a happy or emotionally satisfying ending. One can also continue to work on unearthed problem areas with affirmations and guided imagery. As with all of these methods, it is good to start with a prayer or affirmation, stating your goal for the session, and ask your higher guidance for protection and your inner self for cooperation.

3. Rebirthing: Rebirthing can be used to release emotional traumas that have accumulated since conception. Preferably have one or more initial sessions with a professional rebirther or with a rebirthing group. Find a helper with whom you feel emotionally secure and who has some experience or a good measure of common sense in emotional guidance.

Lie on your back, lightly covered and without restricting clothing. The helper can initially sit at the head end and touch your head for reassurance. Breathe deeply in a fast, regular rhythm, without a pause between inhalation and exhalation. The helper watches that the lower as well as the upper rib cage moves strongly. During inhalation, press shoulders and buttocks into the mattress, while during exhalation, let both relax. Breathe through the mouth and with an open throat to make a sound during exhalation and preferably also during inhalation.

Let go of any emotions that arise, with the helper assisting by pressing, stroking, cuddling, and giving verbal encouragement to bring them out. If no feelings well up, then try to remember an incident when you where irritated, angry, fearful, worried, or resentful and deepen that feeling. After first working on more recent problems, try to relive your birth experience and rectify any emotional upheavals associated with it. A surprising number of people can actually come to remember their own birth in amazing detail.

Have plenty of big cushions or additional mattresses around so that you can throw yourself about. Keep your eyes closed. Preferably use powerful, stirring music with a strong, regular beat during the first part of the session, becoming quieter towards the end and finishing with harmonious, uplifting qualities. Let your feelings follow the music. The session may last for one to two hours.

Try to find a room or open location where you can shout and scream without attracting undue attention. Before you start relating your experience, use crayons to draw a picture of it, the various feelings represented by appropriate colours. You can then change places and your friend can lie down for a session. Repeat from time to time as required to release emotional negativity.

4. Reprogramming: Autosuggestion or reprogramming is one of the most effective tools for changing our belief systems and reprogramming our body and mind. It can be practiced in various forms. At the level of normal consciousness, we can use affirmations. However, suggestions given in an altered state of consciousness are much more effective; we then speak of it as self-hypnosis. We can also use suggestions during sleep. Suggestions can be given in the form of inner pictures and even motion pictures, and then we call it visualization, guided imagery, or reverie. Reprogram-ming is much more effective if we have a strong positive feeling or emotional outpouring at the same time.

EMOTIONAL FREEDOM TECHNIQUE (EFT): Before positive affirmations can become effective, we must first free ourselves from fears and the emotional content of negative memories. Otherwise, these continue to cause distress and block the free flow of energies. One of the best methods that I have found is the Emotional Freedom Technique. This method, mainly developed by therapist Gary Craig, is very effective in overcoming negative emotions and disturbing memories and their harmful effects on body and mind. You may be able to eliminate some disturbing problems within minutes.

The basic principle is to combine consciousness with energy flows. The fundamental discovery of this therapy is stated as: The cause of all negative emotions is a disruption of the body's energy system. It is claimed that single aspect problems, such as fear of public speaking, of heights, of snakes or spiders, can often be overcome within minutes. Also, many physical problems seem to yield to the same formula. This is not entirely surprising as most physical problems have an emotional component. However, in addition, we can expect a direct beneficial effect on the physical body by balancing the energy system, especially if this is combined with conscious awareness.

You can select a health problem or a memory that still evokes a negative emotion when you think of it or feel into it. Craig used the following formula: "Even though I have this [state the problem], I deeply and completely love and accept myself." For instance, if you have been negatively affected by the diagnosis of cancer, then you may say: "Even though I have this fear of cancer, I deeply and completely love and accept myself." You can also try variations, such as: "Despite my fear of cancer, I fully expect to overcome it." Therefore, you first state the problem that you want to overcome and then your expected goal or solution. Preferably say the affirmation aloud, at least in the initial tapping sequence (explained below). In repeated tapping sequences, you can shorten the selected affirmation or say it just in your mind.

The more specific you can be in selecting your affirmation statement, the more effective it will be. Your fear of cancer may be a composite of multiple aspects, such as the fear of dying, a fear of pain or suffering, a fear of having to leave your family behind, perhaps just fear of the unknown, the emotional impact of the original diagnosis, apprehension of financial hardship, of becoming incapacitated and a burden on your family, and many others. You need to address each of these different aspects separately and say, for instance, "In spite of my fear of dying . . ." or "Although I am shocked by my cancer diagnosis . . ."

In order to make the selected affirmation effective, you tap on selected acupuncture points. This temporarily normalizes the energy flow of the tapped meridian. Each time, while saying your affirmation, tap each selected point about seven times in the order given in the list. Do this firmly with the tips of the index and middle fingers of either hand on one side of the body or better still, use all slightly curled fingers on both sides of the body. The meridian points to tap are:

- at the top of the head
- at the inner end of each eyebrow
- on the bony ridge at the outside of the eyes
- on the bone under the eyes
- between the bottom of the nose and the top of the upper lip
- midway between the point of the chin and the lower lip
- where the breastbone, collarbone, and first rib meet
- on the side of the body, level with the nipple of a man or the bra strap of a woman
- on the inside of each wrist

At the beginning of the procedure, and again after a series of tapping, assess the emotional intensity of the stressful memory or fear by imagining the corresponding situation. Assign to it a number from zero to ten, whereby you do not feel any emotional content at zero but it is overwhelming at ten. If the intensity has lessened but not returned to zero, repeat the series with an adjusted affirmation. Now you can say, for instance: “Although I still have some of this fear of cancer, I deeply and completely love and accept myself.” Later you can adjust this still further and say: “Even though I still have some of this remaining fear of cancer . . .”

If your negative memory is not easily put into words, then you can just imagine it as a movie scene lasting for several seconds and rerun it when tapping each point. This may be suitable in an abuse situation with many different aspects to address. In addition to performing the tapping on yourself, you can also work on a friend or relative. If you do not get results, then you may need to reformulate your affirmation to address a different aspect of the problem, or make your affirmation more specific.

With some problems, especially of a physical nature, it may help to repeat the procedure many times during the day for an extended period. To overcome a strong addiction, such as for drugs or smoking, it may need to be repeated every waking half-hour for several days.

Some individuals may also be blocked by what have been called energy toxins. These may be due to electromagnetic pollution, chemical sensitivity, food allergy, and especially medical drugs. If no progress can be made with different affirmations, then try to perform the procedure in a different and preferably unpolluted place. If this does not help, clean yourself thoroughly in a bath or shower but without any soap and then test (using muscle testing, as explained earlier in the book) before putting clothes on. Try standing on a tile or hardwood floor instead of a carpet. If that does not help either, try an elimination diet, as described in part 1. The most common energy toxins are often the foods that we love most, in addition to perfume, herbs and spices, wheat, corn, sugar, coffee, tea, caffeine, alcohol, nicotine, and dairy products. Try muscle testing to find the source of your problem.

For more information see: www.emofree.com and www.emotionalrelief.org . Dr. Mercola has a good EFT protocol for beginners at: www.mercola.com .

As an extension of this method, I suggest that after freeing yourself of negative memories, you also experiment with positive affirmations, especially for improving body conditions. While you may start out formulating your phrase as given above and say: “Even though I still have some remaining weakness in my immune -system, I deeply and completely love and accept myself,” you may eventually rephrase it as: “Every day in every way my immune system is stronger and stronger.” You can augment any other positive affirmations in the same way.

AFFIRMATIONS: Look through your list of beliefs that you want to change and select one or two that have the most immediate importance. Often this will be a health problem or an emotional difficulty. Then work out a short statement to describe your goal.

With breathing problems, for instance, you can state: “My breath flows freely and easily at all times.” Write it down at least ten times and then read it aloud as often as you wrote it. Do this several times daily. While making each statement, try to feel it at the same time; in our example, feel your breath flowing freely.

Be aware of your inner reaction to the affirmation. Does it sound convincing or does it ring phony? Ask your inner self to participate in formulating the affirmation, so it leaves a good feeling when you use it.

Initially, you may perhaps state, “I am happy, healthy, and fit,” but something inside you may suggest instead, “You are happy, healthy, and fit” or “We are happy, healthy, and fit,” the “we” referring to your different selves. Try out such variations and different formulations of the same problem and then select the one that feels best.

Another good general affirmation was used extensively by Emile Coue, the originator of autosuggestion therapy: “With every day in every way I am better and better.” Note the rhythm and poetic sound of this suggestion. In this way, the right half of our brain, which requires artistic stimulation, can accept a verbal suggestion that it otherwise might ignore. The right brain is intimately linked to our subconscious mind. Making the message acceptable to the right brain is the key to successful reprogramming. I also like to say: “My days are filled with love and joy.” You can say this right after the Coue affirmation. Feel the joy and love while you say it.

Another way in which this can be done is by singing our affirmation or saying it in rhythm or rhyme or with suitable music. Most effective appears to be baroque music with about 60 beats per minute, preferably featuring string instruments. You can also experiment by using different colours in writing down your affirmations, or incorporate the written affirmation in a painting, or even express your affirmation as a painting.

Preferably work on only one important problem at a time, be it for days or weeks or months, until you have made sufficient progress. You can, however, use a general formula, like the one by Coue, in addition to an affirmation for a specific problem.

Affirmations are statements of goals or intentions that you imprint on the mental level with the aim of getting them accepted by the inner self. When this is achieved, when the inner self believes your statement, it is as good as done. As with prayer, the key to success is your faith that your request has already been granted at the spiritual level and just takes some time to become manifest in your existence. In order to demonstrate this faith, it helps to stop using the specific affirmation after some time and live as if it has already come true.

5. Self-Hypnosis: Instead of clothing a suggestion in an artistic dress to make it acceptable to the right side of the brain, we can achieve a similar result by switching off the analytical left side of our brain. We can do this by entering a state of deep relaxation or meditation. We are naturally in this state just before falling asleep or upon awakening. In this condition, the alpha brain waves dominate. Experiment with the various relaxation and meditation methods already described in this book.

When fully relaxed or in an altered state of consciousness, repeat mentally, or just slightly audibly, an affirmation which you have selected beforehand. Mentally wander through your body and give a positive suggestion to each organ and system, and send a strengthening flow of energy into each.

Start with the head and say mentally or aloud: “Brain healthy and balanced, hair natural color, eyes strong, scalp relaxed, forehead relaxed, neck muscles relaxed,” and so forth, working your way down to the toes. Then mention the various systems and functions, such as “nervous system strong and balanced, glandular system healthy and balanced, digestive system strong and efficient.” Mention also the blood circulation, the immune system, reproductive organs, and the lymph circulation.

Repeat each suggestion several times or for about half a minute. Concentrate for a longer period on problem areas. Tell your body to cleanse itself of toxins and metabolic residues. Feel the toxic energies leaving your body through the feet and with the breath.

Do not think while giving these suggestions. Let them casually drift into your mind and hold them leisurely for a while. Keep your attention on feeling the various body parts or the effects of any other suggestions. You are in the correct state of mind if you startle at an unexpected outside noise.

Before opening your eyes after awakening or when finishing your meditation or a relaxation exercise, give yourself a positive suggestion: “When I count to five, my eyes will open and I will feel refreshed, energized, and joyful.” Then count slowly and try to feel the sensation of being refreshed, energetic, and joyful. Give yourself helpful suggestions for a restful sleep or for solving specific problems just before falling asleep.

Children, and adults with their consent, can be given helpful suggestions during their sleep. Start by telling them to continue sleeping, that all is well, and that you love them. Then repeat the special message ten times in simple words. With children, this may involve overcoming problems of bed-wetting, behaviour, or school activity. For yourself, you can also use tape-recorded messages that automatically switch on and off.

6. Guided Imagery: For many problems, visualization or mental imagery is the most powerful autosuggestion method. After entering a state of deep relaxation, visualize your goal or the desired solution to your problem, and hold the picture in a vivid form for as long as possible. Try to feel it at the same time. Alternatively, you can produce a “mental movie,” also called a guided reverie. Visualization is usually more effective if you do not try to hold on to a still picture, but make it in some way dynamic, with changes in color, size, and intensity of feeling.

Picture various scenes eventually leading to the desired solution. See yourself as you would like to be, with natural hair color, smooth skin, ideal weight, joyfully dancing, overflowing with energy while running up a hill, and so forth. For improving and toning your body, visualize and feel each organ and gland in turn, see it in a healthy shape and color; you can refresh your memory from an anatomical atlas beforehand. In addition, picture streams of healing energies circulating through the body; use yellow or orange colours for toning, green for balancing, and blue for sedation.

Visualize your health problem. For instance, if there is a tumor inside your body, visualize it as a black or gray mass. Then do breathing exercises: With each inhalation, see your body being flooded with a golden-yellow energy. See how this energy gradually dissolves and washes away the dark tumor mass; with each breath, see it become smaller. The dissolved mass is then expelled with the outgoing breath.

Make this exercise last 15 minutes or longer until the tumor has completely disappeared before your inner eye. If you finish much sooner, then start again dissolving the full-sized tumor. Repeat this exercise daily or even twice daily in serious conditions, best in combination with prayer and meditation.

As an additional exercise, see your immune cells charging into the growth and destroying the cancer cells. Invent a picture with a strong emotional impact. You may, for instance, see your immune cells as white mice nibbling away at a cheese that represents the tumor, or you may see them as star-wars spaceships shooting down ugly enemy spaceships that symbolize individual cancer cells.

For different conditions, invent your own visualizations. For cataracts, see a gray film over your eyes slowly disappearing. If there is a severe pain, see it as a fireball that is gradually extinguished by a flood of deep-blue, water-like energy, drawn in with the breath. For a paralysed body part, see an orange or fiery red stream of energy enter the limb with each breath, causing it to become more and more hot, tingling, and gently moving with each renewed burst of energy. Use your imagination to create a mental picture suitable for your specific problems. There is no limit to what you can achieve.

The following exercise may show you the power of imagination:

- Sit relaxed, then turn your head without moving the shoulders as far as possible to each side, and note how far you can see the wall behind your back from each side.
- Perform a short relaxation exercise and then imagine that your head turns effortlessly almost 180 degrees to each side so that you can see the whole wall at your back.
- Repeat this imaginary turning of the head several times to each side, but without moving the head physically.
- Open your eyes and turn your head to each side and note how far you can now see behind your back.

BODY TALK: Guided imagery is even more effective if you combine it with talking to your body. Imagine that the body has many different levels of consciousness. Your mind usually works at the mental level while your subconscious mind is in charge of the emotional level. Below that is the life-force level, and that is where your body self operates. However, I believe that each organ in our body and even each cell has its own consciousness. You were already enlisting the cooperation of all these individual levels of consciousness with your guided imagery. Now imagine that you can talk to them as well and that they are willing and eager to comply with your wishes.

In a relaxed state, talk to any organ that you want to improve. Talk to the "body self" that is in charge of all the different organs and body parts. Give it a name, as for instance, your unused middle name, and talk to it as you would to a child with a problem. Tell it repeatedly what you want it to do, such as shrink a tumor, put on or lose weight, strengthen the immune system, remove toxins, and so forth. Give it enthusiastic encouragement and convince it that it is able to do it. If, for instance, you have a breast tumor, then ask it to heal the breast and to support the breast in its efforts to heal.

Then in the same way talk to the affected breast. Tell it what you want it to do. Tell it that it is strong enough to heal the tumor and heal any lesions. Ask it to turn to the body self for help. Assure your breast that you love it and promise frequently to send it love energy and any other support that it may need. Then do not forget to do as you said. Tell your body self as well as your breast to let you know what you can do better to support them, and in a relaxed state listen for an answer. A helpful idea may unexpectedly cross your mind.

Talk audibly and frequently to your body, at least once a day. After the initial contact, you do not always need to be in a relaxed state but can talk whenever you feel like it. You may find it very effective talking in front of a mirror. You can also directly talk to your army of immune cells, giving them a pep talk as their commander in chief, or ask the red blood cells to carry more oxygen or the lymph fluid to remove waste products more efficiently.

CHANGING YOUR REALITY: Let's say your reality is that you have been diagnosed with cancer and that you have various physical signs and symptoms to prove it. However, the crucial point is that you cannot change your reality by focusing on what it is. This is so because we create our own reality by what we focus on. If we focus on having and fighting cancer, then we reinforce having cancer as our reality. The harder we fight it, the harder it fights back. Resist and it will persist.

Therefore, to change your reality you need to change your focus. Focus on what you want instead of what you have. That is magic. Focus on what you want to be your reality and fill it with your energy. To do this, you can use various mind tools such as affirmation, visualization, guided imagery, prayer, and meditation. But most important, you must live in harmony with the new reality that you want to manifest. You must live as if - as if the healthier and more joyful reality has already started to manifest. It is not enough to expect tomorrow to be better or upon awakening in the morning to expect to have a better day. It is good to expect that, but expectation is not enough. During the day, you actually need to act according to your new reality by embracing life to the fullest.

Do whatever you need to improve your health. You may occasionally have a cleansing period; you may take certain supplements; or make juices and eat certain foods. But do it with the focus of improving your health rather than fighting the cancer and, whatever you do, do it willingly and joyfully. Make your day as interesting and joyful as you can. The more you succeed in this, the more you will forget that you have cancer and the less you will have it and the less it will have you.

Change your focus, deny your present reality your emotional and mental energy, and it will gradually vanish, allowing your new reality to manifest to the degree that you invest it with your energies. But in our physical reality, that takes time. You need to be persistent with your new focus. Any period of doubt or negativity may invalidate a long effort of positive refocusing and you may have to start all over again. Therefore, avoid people and situations, mainly of a medical nature, that tend to generate doubt and negativity. Train your mind to see only the good in everyone and everything. Even your life-threatening illness was good, by helping you to refocus on a new and better life.

CAUTION: Some individuals may be a little unstable, but may not know it. This is often due to adverse nutritional or chemical factors. During meditation, this can open the door to undesirable influences. In this condition, people may hear voices telling them to act in a manner that may harm others. Therefore, be on your guard and follow any inner voices only if their advice is in harmony with your ideals and with the signs of true spirituality. Especially do not act on any advice that may be harmful to another person.

7. Prayer and Meditation: In prayer you ask for guidance, give thanks, or offer yourself for service, while in meditation you listen for the answer. The two methods can be combined in what we can call "inner communication."

PRAYER: Prayer is a means of making our requirements known to a higher guidance. This may be a request for help for ourselves or others; it may be a request for advice; we may express our gratitude; or we may offer ourselves as vehicles for greater service to mankind. You can continue to address your prayer in the accustomed way or however it feels best. When you pray to God, Jesus, or a saint, assume that the prayer passes from the lower self to the Higher Self and from there to any other "address" as may be required.

When you need help with an important problem, formulate your request in a short, precise form and preferably write it down. Then enlist the help of the lower self to pass on your prayer to the Higher Self. This can be done through devotion, invocation, chanting, or in any way that produces an emotional uplifting, a warm emotional tide coming from your heart or solar plexus. This creates a "carrier wave" that can now be modulated by clearly and precisely stating your request.

Best say it aloud, in the same way as it is written and with the same words on successive days. In addition, create a mental picture or impression of the desired result. It is important that such prayers be performed with a deep inner conviction that the request will be granted.

When you live in close contact with your Higher Self and readily receive guidance, you may proceed in a less formal way and communicate with your Higher Self in whatever way it deems appropriate. The most effective prayer is a thanksgiving for the granted request, knowing that it has already started to manifest on the spiritual level and will in due time appear in your reality.

Here a summary of some requirements for a successful prayer:

- Contemplate and formulate your message.
- Relax or enter a meditative state.
- Create an emotional opening or outpouring.
- Say your message in simple words, preferably aloud.
- Include thanksgiving for what you are about to receive.
- Have complete faith that the prayer is successful.
- Repeat your prayer on successive days.

MEDITATION: During meditation, you still your mind and raise your awareness towards your guiding consciousness. Christian saints used to call meditation "silent prayer." In meditation, you remain in a heightened state of consciousness without thinking.

During meditation, we should be wide-awake. Therefore, if we feel somewhat tired, our meditation can be improved by raising our energies beforehand with a few minutes of deep breathing and whole-body shaking. Then proceed with a relaxation exercise and preferably an invocation or devotional service in which you state the goal of your meditation in a prayer or affirmation.

Assume a physical position in which you can be comfortable for an extended period. Preferably keep the spine straight. You can sit erect, with both feet on the ground, in a chair with a straight back, or sit cross-legged in a lotus or semi-lotus position. For many, it may be more comfortable to sit on the heels with a cushion under the feet. You can also lie flat on your back if you are able to remain wide-awake in that position.

An easy method is to start with eight-two breathing (as explained in step 9) and then gradually let the breath move on its own. Focus your full attention on the breath; feel how it flows in and out smoothly and rhythmically. Do not try to regulate or guide the breath; just follow it mentally and feel it. When thoughts arise, let them drift past without paying any attention to them; the same with outside noises, just ignore them.

At other times, you can concentrate on your heart area and feel it become warm and filled with love and compassion. Alternatively, focus on the middle of your forehead, feel it as a centre of happiness or even bliss. If the eyes are closed, it helps if you turn them up and inwards as if looking at the forehead centre. In addition, feel connected to your Higher Self through a beam of energy entering the top of your head.

Hold these sensations as long as you can. If possible, unite the sensations of the head and the heart and expand them throughout the body. Do not be disappointed if you cannot feel the warmth and love radiation the very first time, just keep on practicing. Meditate for 30 minutes or longer, the longer the better. Preferably make meditation a regular habit at the same time and in the same location each day.

A Healing and Love Meditation

- Sit or lie relaxed but with a straight spine.
- Take some slow, full breaths; mentally follow each breath as your breath continues to flow quietly in and out on its own.
- Feel how your body gradually becomes immersed in cosy warmth - you can imagine being in a warm bath.
- Concentrate on the idea that with each inhalation vital energy or prana begins to fill the body. Direct this energy into your abdomen and feel it becoming very warm.
- Direct the pleasant-feeling energy stream to those areas of the body that are especially in need of healing.
- Concentrate on the middle of the chest where you create a feeling of intense love and compassion.
- Initiate this feeling by remembering an actual situation or person that invoked such feelings in you. After a while, forget the actual memory and feel love and compassion in an abstract way.
- Imagine that you radiate these feelings, like a sun, radiating love and compassion instead of light and warmth. Radiate these to all your body areas in need of healing as well as to the outside, filling the room, the house, and finally all the world. Realize that you are not the originator of this powerful energy flow, but just a channel through which it emerges from higher dimensions and is transformed by you to the human qualities of love and compassion.
- At other times, radiate joy and happiness instead. Hold on to the impression of being immersed in an ocean of beneficial healing energy and for the rest of the day feel an afterglow of the love and compassion or joy and happiness you experienced.
- If an unpleasant condition develops during the day, try to recall mentally the ocean of warm contentment enjoyed during meditation.

CENTERING: During our daytime activities, especially on important occasions, it is beneficial to focus our feeling awareness on an inner energy centre. We may feel this as a deep inner peace, as love or compassion, depending on our focus and the requirements of the occasion. This allows us to gather our energies and gives us poise and serenity. This is called centering - moving through life in a meditative state.

There are various methods of centering. The Japanese, for instance, prefer to centre themselves in the hara, the storehouse of vital energies behind the navel. This helps one conduct daily activities in a more focused and powerful way. If we prefer to interact in our relationships and daily life with love and compassion, then we use the feeling awareness in the heart chakra as the method of centering. Focusing on the forehead chakra may be more appropriate during times of intense mental activity, such as discussions and meetings.

If we aim to act from our level of spiritual consciousness and like to feel connected with and guided by our Higher Self, then we may centre on the crown chakra. Finally, with some practice, we may be able to fuse our awareness of all of these chakras and remain centred within this integrated feeling awareness.

When we are centred, we are not only mentally and emotionally poised and balanced, but at the same time alert and intuitive. In this way, we can remain in touch with our spiritual guidance continuously during our waking activities. When we are active while being centred, both sides of our brain work in a balanced way and this is reflected in the balance of our whole personality.

It is much easier to remain centred during manual activity than while thinking, reading, or writing, but gradually we can also learn to remain centred during mental tasks. Eventually, we may be able to remain indefinitely in a semi-meditative state by focusing part of our awareness on our inner centre while another part deals with our external problems.

8. Inner Communication: In order to receive direct guidance from your Higher Self, combine relaxation, visualization, prayer, and meditation. Start with a relaxation method that gives you the feeling of drifting lower and lower as to the bottom of a lake. Or imagine sinking down a long, black tunnel towards the core of your being. The farther you sink, drift, or swim downwards, the more you leave the noise and activity of the outer world behind. Regard any remaining noise as ripples on the surface of the lake while you remain at the bottom in peace and stillness.

Instead of going down, you can mentally create a special room that you enter only when you want to meet your higher guidance, either by telepathic communication or seeing it as a wise and kind guide. Another possibility is mentally walking down a shaded path to a waterfall and meeting your guide there.

Whether you see a guide or not, address your presumed guide with reverence and ask him to start guiding you on your spiritual path, in health matters, and with other problems, by communicating with you. Ask a question and listen for the answer. You may receive an immediate telepathic reply, such as an inner voice or a thought popping into your mind. Alternatively, you may develop a certain feeling in reply to a question rather than getting a mental response and this may indicate the direction in which to proceed. Maintain a blank mind after asking your question, otherwise you may not hear the answer or start thinking of an answer yourself. You can ask several questions and hear answers as in a dialogue. Prepare your questions in advance, before starting the relaxation, so that you do not need to think while in meditation.

INTUITIVE MEDITATIVE THINKING: This is related to the traditional practice of contemplation. Suppose you want to find an answer to a problem. This can be a health problem, an invention, or a spiritual matter. Start by finding out relevant facts and information, and then examine it from different perspectives with your mind. In this way, you can identify the key question that needs to be answered in order to find the solution to your problem. Keep this key question in the back of your mind and enter a meditative state without specifically thinking about anything. Watch any thoughts drifting by. If you have thought hard enough about the problem before and now maintain a reasonably blank mind but with the question still in the background, then the answer may now drift into your consciousness as an intuition.

FOCUSING: Lie down or sit comfortably and induce a relaxed condition. If you have a pain, hurt, disease, or any problem in your body, focus your attention on it. Feel into it and ask it what it has to tell you, or visualize your problem in a simple, symbolic way, then mentally sit back and let the pictures or symbols change of their own accord.

You can do the same thing with an uneasy feeling such as created by a social, financial, or other problem. Contact that uneasy feeling, possibly by feeling it as a pressure somewhere in your body. If you have a multitude of problems, ask yourself which one is the worst and try to feel this one first.

After listening to your body and feeling the tension it creates somewhere, a certain impression may form within you as an answer to what it is all about or you may see a symbol as an answer. Then ask that tension or problem in your body whether this is the right answer and feel what it does. If your answer is correct or partly correct, the problem will shift; the tension or pain may lessen or you may feel relieved. If the answer is wrong, the feeling of unease will not change. Try again. This process may have to be repeated many times over several days or weeks to come to the bottom of deep-seated emotional problems.

SLEEP OR DREAM SOLUTIONS: Before falling asleep, pray to be shown a solution to your problem overnight or tell your subconscious mind repeatedly in simple words that it will provide an answer to the problem at hand, either in a remembered dream or as an insight during the following day. Visualize your problem in a simple scene before you drift to sleep. If you are not already in contact with your subconscious mind, this procedure may have to be repeated several times.

MAKING SACRIFICES AND AMENDS: Sometimes it may not be possible to get to the deep-seated cause of a personal problem. Especially in cases associated with feelings of guilt or having done something wrong, it is often helpful to make a sacrifice in order to regain the cooperation of our subconscious mind or lower self.

The sacrifice must be enough to hurt you to some extent, for example, selling the best part of your stamp, coin, or art collection and using the proceeds for a charitable purpose. Alternatively, give a free service or care for a stranger in need - anything that will satisfy the subconscious. Make amends with reference to past faulty mental, emotional, or dietary habits. Promise improvements and definite actions to your lower as well as to your Higher Self. However, a broken promise will lead to further health problems, so be careful and persistent with your promises. Do not resent your sacrifice, but do it in a cheerful mood.

MOVING WITH THE FLOW: With or without inner communication, we can gradually adopt an attitude towards the events of our daily lives that may be described as “moving with the flow,” as in the flow of life. This means we flow through life instead of battling it when the going gets tough.

We can assume that our Higher Self guides us in a certain direction. When we follow this direction, life will run more smoothly, events fall into place, we meet the people we need, we find information that helps us along, and we obtain the necessary resources to continue on our way. If we go too far in a wrong direction, however, things will become tough, we make mistakes, lose out, do not obtain what we want or need, and so forth. Therefore, when we adopt the basic attitude of moving with the flow and we run into difficulties, we realize that we may need to change direction. While we remain true to our spiritual goals and ideals, with the goals of daily life, and especially in the way we intend to achieve them, we try to be very flexible. Rigidity in daily life easily leads to pain and disappointment.

We can compare our stream of life to a newly formed watercourse that seeks its way down a gentle mountain slope. Sometimes it winds its way steadily downhill, while at other times it encounters a barrier, a hollow or depression in the ground. Then it has to wait patiently for a while until the hollow is filled up. It may probe in various directions until finally, quite naturally, and without a problem, it finds the overflow, the lowest point in the wall of resistance, and happily it continues on its downhill way.

This, then, shows us how we can follow our course of life in an easygoing, pleasurable way, without undue stress or hardship. When we encounter a resistance, we probe and wait patiently until we find a natural opening. All we need are flexibility, patience, and faith. Faith that we are being guided, that we meet the resistance for a purpose, and that we will learn and grow by overcoming it or finding a way around it. Faith that all will be well and that an interesting and pleasant future awaits us.

ASK FOR HELP: If problems seem to mount and you feel pulled down by it all, relax and ask your guidance for help. State your situation, that you do not know what to do or that there is apparently nothing that you can do. If appropriate, acknowledge that you made a mistake and got yourself into trouble because of a thoughtless or selfish action. Ask your guidance to take over your problems and provide a solution that is fair to all concerned. Then, whenever you notice your mind straying towards a worrying thought, affirm silently that now your guidance has taken over and is providing an equitable solution.

EXPLORE POSSIBILITIES: Sometimes you have to make a choice between several possibilities without a clear preference for any one. Like water when it meets a resistance in its flow, so you can probe in various directions, exploring the different possibilities with your mind, while at the same time closely observing your feelings.

Let us assume you have several options for your future. All have advantages and disadvantages, and you do not know what to do. Doing nothing or changing nothing may have undesirable consequences that need to be evaluated. Start by imagining that you follow one course; visualize or mentally explore the possibilities arising from this and observe how you feel about it. Explore the second choice and then a third choice in the same way. Usually one of these mind-plays will generate a better feeling and this is the one to adopt.

Whenever problems seem to mount and you start to worry, become aware of your spiritual philosophy and ideals and ask yourself: “Does it really matter?” Remain centred and all your problems will soon melt away in the warmth of your radiating love and a deeply felt inner peace.

* * * * * * * * * * * *

Part 10
SPIRITUALITY AND HEALTH

In order to become more whole and healed on all levels of our being, we need to improve ourselves on these levels. Complete healing is not possible by working on just one level. As an example, assume that you are overweight. Mustering willpower, you start a diet and are successful in losing much weight. The problem is that after some months it slowly creeps back up again and after a year your weight is where it started.

Eventually, you come across some emotional release workshops and find that these, together with a sensible diet, are more effective for losing weight and keeping it off. Nevertheless, you find it difficult to keep your emotions under control because you still continue to react in undesirable ways. Now you discover that you are influenced by negative beliefs that make life difficult for you.

You realize that you need an appropriate belief system in order to manage your emotions in an appropriate way. You start looking for a philosophy of life that helps you to become and remain healthy, happy, and fit. This is where the spiritual dimension comes in. You will find that the more you adopt spiritual principles in your life, the more life will become an adventure and a pleasure rather than a struggle.

Most of us have developed our belief systems haphazardly, drawing from parents, school, and social contacts as well as books and mass media. These beliefs are often irrational and detrimental to us.

So to heal our emotions and other levels of our being, it is essential to adopt a suitable philosophy of life to guide us. I call this a spiritual philosophy. The various religions appear to have different spiritual philosophies. These differences are most pronounced at the mental level of dogmas and doctrines, while they tend to disappear at the level of the highest ideals of each religion. In all major religions, some followers have aimed for these highest ideals. They are the mystics who left us in their writings a spiritual philosophy that transcends and unites all religions.

This last chapter offers an outline of the spiritual dimension, existing in a continuum with our physical world and not separate or in opposition to it. Out of this understanding, we can accept a spiritual philosophy of life as the basis for our beliefs, which determine our decisions and emotional reactions in life. Then I offer a sketch of a modern spiritual path that uses health and emotional problems to develop the spiritual master-ship that gives us the ability to lead a happy and fulfilled life.

* * * * * * * * * * * *

Step 64

THE SPIRITUAL DIMENSION

Enter the spiritual dimension of life and make it your home.

The concept of spirituality means different things to different people. In this book, it means deliberately manifesting or aspiring to manifest a higher, purer consciousness. We do this by developing the attributes of spirituality, such as compassion, discernment, faith, high ideals, harmony, joy, inner knowledge, intuition, kindness, openness, patience, self-responsibility, serenity, tolerance, wisdom, and a gentle love for All That Is.

To be spiritual is not the same as to be religious, which involves following the creed of a proclaimed religion. Someone can be spiritual without being religious and vice versa. Those who are religious become spiritual by following the highest ideals of their religion, while those without a religion follow their own highest ideals and their higher guidance. On the higher levels of consciousness, religion and spirituality merge. In Christian terms, for instance, we are advised to "seek first the Kingdom of Heaven" and all else will be given us. In spiritual terms, this means to seek first and follow our higher guidance to manifest spirituality, then we can expect health and happiness to result as a natural by-product.

Spiritual development should not be confused with psychic development. Psychic abilities are no measure of spirituality. However, psychic abilities will develop on their own and in a natural way during spiritual awakening, according to the need for their use. While neither good nor bad in themselves, psychic abilities without spirituality are like nuclear power without prudence; both can lead to disaster.

Neither is spiritual development a quest for altered states of consciousness, though these may occur as natural by-products of deep relaxation and meditation. They can be useful and provide us with insights as well as beautiful and uplifting inner experiences. However, our main task is at the level of our normal consciousness, and such inner experiences at elevated levels need to be transformed into improved living habits.

We become more spiritual by living according to our highest ideals, and by consciously improving our attitudes and feelings. Every moment in our day-to-day interactions provides an opportunity to become more spiritual. We are aided in this endeavour by asking for help from the highest guiding part of our consciousness, which we can call the Higher Self or, in Christian terms, the Holy Spirit.

Our Spiritual Growth: Living spiritually is our day-to-day, minute-to-minute effort to live according to our beliefs. Our body with its problems is an important partner in this; it shows us where to start and what to do and helps us become more aware of all aspects of ourselves.

Healing our body, mind, and emotions is part of our spiritual growth. Even if you do not otherwise intend to live spiritually, you do so anyway by healing yourself on these levels. I am convinced that by living in harmony with the biological, emotional, mental, and spiritual laws that govern us, we can be close to perfect; we might even personify the risen Christ. Therefore, all imperfections, usually manifesting as problems, show us what to do to become more spiritual and, with this, more happy and fulfilled.

Starting at the biological level, we may have a health problem, possibly a specific disease. This disease has one or more causes, usually on several levels. Just trying to ease symptoms with palliative care is not a spiritual solution, although we can use the disease to practice cheerfulness in adversity. With our spiritual quest, we try to discover and overcome the causes for the disease. We experiment with nutrition, cleansing, and other natural healing methods; we improve our energy flows by softening armoured muscle structures; we release suppressed negative emotions; and we may use regression and reprogramming, guided imagery, and meditation. Chances are the disease will more or less disappear as, at the same time, we have progressed on the spiritual path.

The same is true for any of our other imperfections and problems: a rigid and aging body, a lack of meaning, unsatisfactory relationships, worry, or resentments. Keep part of your mind in observer mode and be aware of what you are doing. If you catch yourself worrying about something, that is your problem to work on; worry is a negative attitude, and you need to change your approach. Do what you can to get the problem fixed and then leave it to faith; use prayer and meditation to strengthen your faith.

If you catch yourself having an unkind thought about someone, immediately make an effort to send a kind thought to that person. If you believe someone has harmed you, realize that this person is a human who needs your help. Be angry if you feel that way, but let it out in a suitable manner, and when you have calmed down, forgive and send love.

Learn to laugh about yourself when you make a mistake or break something. You would laugh if it happened to someone else, so try it for yourself; it makes life much easier. Taking ourselves too seriously shows that we are dominated by our ego.

It is said that no one is an island, that we all are in it together. We interact all the time, and helping our fellows is the most important part of interacting. We cannot help by preaching or trying to convince somebody. The most effective way is to be an example of your spiritual beliefs. Be forgiving, be kind, try to do the right thing at the right time, but if you mess up, do not worry; you are not expected to be perfect, so try again.

As your consciousness expands, you may realize that your job or your partner are no longer in harmony with your beliefs. Visualize and pray for a solution and watch.

You do not need to go out of your way to look for opportunities to help others. Just be aware and do your best whenever you come across something; your guidance will arrange the opportunities for you to learn and help. Do not blame yourself or anyone else if something goes wrong; making mistakes is the best way to learn. We need to make mistakes and learn from them - that is the spiritual path.

Specific ways of life have been devised through the ages to facilitate the growth of a spiritual body. Foremost in these are prayer and meditation, especially as part of monastic life. In a related form, we have the yoga of devotion and the yoga of action in which we dedicate our life to unselfish social work or healing. Rituals, such as performed by the Orthodox or Catholic church, metaphysical societies, and New Age groups, offer spiritual food. When we deliberately set out to awaken and build a body for the inner Christ, we are said to be on the spiritual path.

The Structure of Manifestation: Here is a brief outline of a theoretical basis of spirituality as I see it. Starting at the fundamental level, I regard consciousness as the basic ingredient of everything in existence. We can compare consciousness to air, which can be compressed to varying degrees of density in containers of different shapes and sizes. In a similar way, consciousness becomes compressed and can then aggregate to form more complex structures.

Different densities accumulate at separate levels to form an additional dimension based on density of consciousness. I call this the dimension of "space-density." A certain density produces atomic and subatomic particles, while lesser degrees of density manifest as etheric energy, bioenergy, or prana, as feelings and emotional energy, as mental energy and thought forms, and finally as spiritual awareness. All of these represent different space--densities that can interpenetrate each other to occupy the same three-dimensional space.

Therefore, there is no fundamental difference between the consciousness of an atom and a human; it is only a difference in complexity and range of densities. The universe is basically consciousness acting on itself. Less dense fields of consciousness build, change, discard, and rebuild forms from denser fields of consciousness in order to experience and express themselves. The main activity of consciousness is creativity; its highest manifestation on our planet is unconditional love.

We can assume two streams of consciousness. The "stream of creation" leads towards greater density and less complexity to form atomic particles and the physical universe, while the "stream of evolution" flows from dense towards less dense structures with greater complexity towards the spiritual universe.

As part of the stream of evolution, consciousness uses the various life forms, like water uses a riverbed to flow to the ocean. A life form, be it a species or an individual form, is useful only as long as it allows the consciousness to flow through freely and expand. If rigidity develops and consciousness becomes stagnant, then the form has reached old age and will soon disintegrate.

Physics cannot adequately and logically answer the question how all the energy and mass of our universe could have been concentrated in a single point at the start of the Big Bang and what was before. A spiritual theory of creation would say that originally there was no energy at the physical level, that it emerged instead from the etheric level, not so much as an explosion, but rather like the crystallization of a supersaturated liquid starting from a seed point. A spiritual theory also answers the question of the origin of consciousness that biology cannot answer: Our physical universe is a condensation of a less dense etheric universe. Atomic particles are the centres of etheric vortexes, comparable to atmospheric high and low pressure areas.

Planetary Evolution: I see the evolution of consciousness on our planet as a multidimensional cooperative venture. Consciousness in the form of the growth force is most active in the plant kingdom and works to transform the mineral kingdom.

In the animal kingdom, consciousness becomes dominant at the level of feelings and emotions; it is at this level where most of the experimentation and creativity are present. The emotional level channels the growth force into more individualized forms than those in the plant kingdom. A price for this individualization is a more solidified growth force with less regenerative capacity.

Insects are close to plants, with only rudimentary feelings, while primates and possibly dolphins are closest to the humans with highly evolved complex feelings. Similar to the evolution in the plant kingdom, which leads to the beginning of feelings in the highest plant species, so the higher animal species acquire a rudimentary capacity for thinking.

By having its emotional energies guided with the help of the emerging mental level, the animal kingdom culminates with the primates channelling their life force into a nearly perfect form for our planetary conditions. With this, the planetary evolution of the growth force has been further completed and humanity works now on the perfection of the emotional level with the help of the mental level. We use beliefs to experiment with our emotions.

The animal kingdom required a rudimentary mental level for the growth force to be channelled by the emotional level into a perfect form. In a similar way, humans need a level of consciousness beyond the mental to perfect our emotions. We call this next higher level “spiritual consciousness.” It evolves through the programming of our higher mentality by the incarnated soul. We can regard our different and evolving group belief systems as the human equivalent of different animal species.

The Nature of God: God can be defined in two ways. We can equate God with consciousness and say God is everything. The consciousness manifested in the various universes - the physical, etheric, and so forth up to the spiritual universe - are the energy bodies of God, while the non-manifested part might be called the mind of God. Alternatively, we can only regard the non-manifested universal consciousness as God, and this may or may not be focused into what we can regard as a personified being at certain levels of consciousness. For the evolution of our own consciousness, the details of what we believe is God do not matter very much at this stage. As part of a spiritual philosophy, we can assume that the more consciousness becomes condensed, the less it is aware, especially of being a part of God or All That Is.

The Oversoul is the creator of our soul, and as our personal God or God Self, it is our individualized representative of the universal aspect of God. Any communication from this divine level can reach our normal consciousness only by penetrating our mental and emotional levels. These, however, act as filters that distort and interpret any messages according to our beliefs and emotions.

Therefore, a devout Catholic will receive messages clothed in Catholic symbols, while a Hindu will experience divine revelations according to his religious beliefs. The less strong and structured our beliefs are, the less will any higher communication be distorted. The more we expand our consciousness, the closer we are to this higher guidance and the easier it will be to communicate. Disbelief, on the other hand, will block any communication.

A Spiritual Philosophy: A spiritual philosophy is our basic belief system about the nature of our existence, its purpose, and goal. From this we try to understand our role in society and our relationship with guiding forces, and it gives us the strength to follow the spiritual path. Everyone has a philosophy of life, even if they are not consciously aware of it. For some, this is just the notion to get as much pleasure as possible out of life while trying to avoid suffering. Others, like various Christian saints and yogis, have deliberately self-inflicted suffering and denied themselves pleasures as part of their philosophy and as a path to their God.

To have a spiritual philosophy means to see ourselves as part of a higher purpose, of a universal blueprint, and we consciously try to cooperate with its unfolding and manifestation. A good spiritual philosophy needs to give us guidance and support on the spiritual path and in our daily interactions and relationships.

In the same way that we can deliberately choose to believe in and follow a certain religion or political party, we can also deliberately adopt a spiritual philosophy that exactly suits our purpose. Normally, this adopted philosophy can be in harmony with our present religion or political views. Just look for the highest ideals and replace any negative aspects with positive beliefs.

The centre point of any philosophy of life is the question of meaning or purpose. If we see no purpose in life and especially in our own life, then it is natural to live for the sake of sense and ego gratification, but we can become depressed by the apparent lack of meaning. In a spiritual philosophy, we believe in a higher purpose; we may not know the exact nature of this purpose and it may not even matter. We can say we just want to manifest our own life blueprint, or fulfil the purpose for which we incarnated, or do God’s will, or become one with God.

The Middle Path: By now, you may be apprehensive. Basically, all you want is to be healthy, have a good life and some fun, and generally feel fulfilled and satisfied. You do not want to become a mystic or a saint to achieve this. You do not need to. I painted the big picture to show in which direction we have to move. How far or fast you move is up to you. The important part is to move in the right direction, however slowly. Suffering comes from moving in the wrong direction, no matter how far advanced we are spiritually. The saint will suffer if he moves in the wrong direction, and the sinner will feel good if he moves in the right direction.

Indeed, this is a middle path between pleasure-seeking and intentional or unintentional self-inflicted suffering. In this, it is like the path of the Buddha. For a long time, the Buddha tried to reach enlightenment through asceticism. He meditated until he left a mark like the footprint of a camel where he had been sitting. Finally he realized that he was on the wrong path, that he needed his body as a partner and had to fulfil its needs, that he had to be firm but kind to it. He called this the middle path, and many seekers after him found it to be the right path for them.

In the same way, our spiritual philosophy should help us to feel good in body, mind, and spirit. Suffering is only for the ego. If instead we identify with our Higher Self, then life will be easier and more fun.

Our basic premise is that we attract or become what we believe in. Assuming that the universe contains all possibilities, all shades of good and bad, right and wrong, desirable and undesirable, it is more desirable to believe in a benevolent universe than in a vindictive one, in a God of love rather than fear. This spiritual philosophy of the middle path says that we can choose and have what we want, but we must firmly believe in it.

* * * * * * * * * * *

Step 65

LIFE ON THE PATH

The spiritual path is the crowning glory of human endeavour, and by following the 65 steps you are already on it.

The concept of spirituality means different things to different people. In this Presently, travellers on the spiritual path have two advantages compared to those in former centuries. We now have published information on the nature and variety of spiritual paths, so we no longer need to find our way alone. Second, our capacity to understand and communicate has greatly increased. What formerly was incomprehensible mysticism, we can now understand and present in a rational way.

There is now so much information available on yoga, working with energies, healing the emotions and the mind, and so forth, that your problem is not to find the information, but to discern what is right for you. You can do this by keeping your highest goals in mind and asking your inner guidance.

Beware of getting sidetracked by psychic abilities or happenings. Developing psychic abilities can be fascinating but is no more spiritual than material, emotional, or mental activities. However, psychic activities can be more dangerous than other activities. Most psychic happenings take place on the astral level or open us to astral influences. Entities operating on this level are commonly less spiritually developed than someone on the spiritual path and often give misleading or mischievous information. They also love to imitate or pretend to be a high spiritual entity or even God.

Awakenings: Do not worry about whether you have had a spiritual awakening or are on the spiritual path. Your interest shows that you are on the path. Just start or continue searching for spiritual information. You will be attracted to what is important and right for you. While in the beginning, it is helpful and necessary to read different sources to get a good overview, later it is better to read as little as possible. By continuing to read a lot, you remain at the intellectual level, and that hinders the development of intuitive abilities. If you do not control yourself, you could become "addicted" to reading about spiritual matters, instead of doing it.

If you are a practicing member of a religious denomination, do not be concerned that the spiritual path may be incompatible with it. Instead, expect that it will increase your faith. The spiritual path does not have doctrines; writings about it are only opinions you do not need to accept. Walking the spiritual path means finding your own insights and relationship with God.

Purification does not need to be a time of misery and suffering as it once was for mystics. Instead, it can be a time of adventure and discovery. By removing biological, emotional, and mental blockages, you become healthier and happier. Life will become more satisfying. The fun is not the arrival, but the journey. You do not need to make a sharp distinction between working on the biological, etheric, emotional, and mental levels. They are all related, though it is easier to talk about one level at a time. By concentrating on one level, you affect all the others. It is easier to start with the biological level and then focus upwards.

The Biological Level: In former centuries, spiritual seekers ate only organically grown food and their water and air were free of chemical pollution. Today we have to work hard at purifying our body to the level where previous seekers started. Therefore, we start with sanitizing the biological body and for this I recommend:

- Basic Cleanse: This can be repeated from time to time, especially if you want to lose weight; also do any other cleanse you prefer; see Fasting and Cleansing in Part 1.
- Test for food allergies and chemical sensitivities at the end of the Basic Cleanse; see Allergy Testing in Part 1.
- Pay attention to the acid-alkaline balance of your body as explained in Part 1.
- Take measures to free your body of parasites and harmful microbes and use anti -inflammatory measures as required; see Parts 2, 3, and 6.
- Eat plenty of raw food, including fresh green and purple juices; see the High-Quality Diet in Part 4.
- Select food suited for your metabolic and blood type; see Metabolic Types, Pathways, and Diets in Part 5.
- Have regular physical exercise and outdoors activity, also yoga, stretching, shaking, walking, jogging, or swimming.
- Sanitize your teeth, bedroom, and workplace; see the Practice of Healthy Living in Part 1.

The Etheric Level: The main features here are the etheric body and the chakra and acupuncture systems (see Part 7). This level of subtle energy determines the health of our body and how energetic we are; this is the energy of life, of healing and manifestation.

To manifest something, be it health or a physical or social condition, we use imagery. Hold a clear visualization of the desired situation, fill it with energy, and continue to expect it to manifest. An important part of our work at this level is to accumulate and direct etheric energy.

In energy work, energy follows thought. However, mental and etheric energies exist on different density levels and do not affect each other directly. Therefore, they require the assistance of the emotional or feeling level, sandwiched between the two. For learning to accumulate and direct etheric energies, practice the energy-circulation (Part 7) and the Love Cure exercises (Part 8). When you are able to feel the energy flows, use them in guided imagery for healing and other manifestations. An excellent book on this subject is The Laws of Manifestation, by David Spangler.

By practicing the Love Cure, you can improve weaknesses and imbalances in your chakra system. This will have favourable effects on your health. If you are so inclined, use guided chakra meditations available on tape or CD from New Age outlets.

For improving your acupuncture system, see Meridian and Acupressure Therapy in Part 7, and after you have learned to feel the energy flows, use mainly mental tracing of the channels. For further practical information on working at the etheric level, see books on kahuna; the modern kahuna teachings started with books by Max Freedom Long (The Secret Science behind Miracles and The Secret Science at Work).

The Emotional Level: For most, this will be the most difficult and possibly most painful level to sanitize, but it will also be the most rewarding. I suggest that you read again and work through Part 8. The important steps in healing emotions are:

1. Emotional cleansing to remove blockages and traumatic memories.
2. Overcoming addictions and cravings.
3. Free-flowing emotions - learn to feel and express emotions freely without hurting others.
4. Generating positive feelings - use meditation and guided imagery to create positive and uplifting feelings.
5. Reacting positively - use appropriate beliefs and mental awareness to react positively to events.

1. EMOTIONAL CLEANSING: Experiment with some of the methods described in Part 9 under Mind Tools, especially with techniques for emotional release and regression; these are best practiced in healing groups (Part 8). An integrated way of healing the emotional level within a spiritual context is provided by the system of Body Electronics devised by Dr. John Whitman Ray (for more information, see: www.bodyelectronics.net).

Similar to Reichian therapy and bioenergetics, Body Electronics works mainly on the body, using sustained pressure to release memories of emotional events that have been lodged in muscles. Practice Body Electronics as a group activity, possibly after an initial assessment by a practitioner; Reichian therapy and bioenergetics are mainly practitioner-based. However, you can also practice bioenergetic exercises at home or in a group; for instructions, see The Way to Vibrant Health, by Alexander Lowen, M.D.

A simple way to contact and release emotional traumas is by using present emotional upsets. If something strongly affects you emotionally, try to feel into it, possibly combined with a relaxation or meditation exercise, to discover why it upsets you. If necessary, reinforce it so as to feel it even more strongly. Most likely, forgotten memories will come up about a similar incident in your younger years. Whatever the method, when you become conscious of an emotional event from the past, it ceases its detrimental effects or it can now be dealt with.

Some of the destructive emotions that we need to uncover and release are fear, guilt, and resentment. These block the flow of life-force energy and make us susceptible to diseases and aging. The main remedies, after making a suppressed emotional memory conscious, are unconditional love, understanding, and forgiveness.

On this level, you do not need to clean out all hidden negative emotional memories, as this would probably be impossible. Aim to uncover and remove all those factors that interfere with the flow of energies in your body, with your emotional health and enjoyment of life, and with your ability to make decisions.

2. OVERCOMING ADDICTIONS AND CRAVINGS: For many, overcoming addictions and cravings may seem to be the hardest task. Addictions to foods, drugs, and stimulants are best overcome with a combination of guided imagery, prayer, affirmation, and cleansing periods with a very high fluid intake. Homeopathic doses of the addictive agent can be helpful.

We can also be addicted to or have cravings for certain activities or objects, such as sex, gambling, entertainment, expensive dresses, jewellery, or antiques. Again, the activity or object is not the problem, but rather the attachment to it, the desire for it that controls us rather than we controlling it. The most important step is in acknowledging that we have a problem. Apart from prayer, meditation, guided imagery, and affirmations, the most helpful approach would be to live for a sufficiently long period in a simple way in natural surroundings.

3. FREE-FLOWING EMOTIONS: For most of us, freely expressing our emotions is contrary to our social conditioning. A first step then is learning to feel our emotions when they arise instead of suppressing them; we can then decide to express them immediately in a harmless way or re-enact the situation later. In our imagination, we can relive the situation and kick or punch a pillow or do or say, possibly in front of a mirror, what we felt like doing at the earlier time. Nevertheless, it is preferable to develop the skills to react appropriately whenever the situation arises because then we can also be an example to others.

Marriages and other close relationships are ideal opportunities for learning to communicate. If free expression is not always possible, then arrange a fixed time at the end of the day or once a week when you each can freely express your thoughts and feelings. It is important that one person speak without interruption until everything has been said, and then the other reply, also without interruption.

4. GENERATING POSITIVE FEELINGS: After emptying ourselves of negativity and learning to acknowledge and express our feelings and emotions, it is equally important to fill and surround ourselves with positive and uplifting feelings. Ideally, this will come as a natural by-product of habitual meditation and through daily activities that make us feel good.

It is beneficial to practice feeling exercises as explained in the Love Cure in Part 8. Look at a beautiful object, a tree, flower, cloud, or landscape, and send it appreciation, admiration, joy, and love. Do the same in thinking of an individual. Send feelings of forgiveness, appreciation, and love to anyone you may have hurt or who hurt you.

Learn to carry into your daily activities the loving feeling that you can now generate in special situations. Train yourself to carry this loving feeling in your heart wherever you go or whatever you do. Whenever you see or meet someone, embrace that individual with the warm feeling. This is especially effective for improving relationships with people who do not seem to like or appreciate you. Never again do you need to feel bored when having to wait in an office or anywhere else. Just start generating loving and joyful feelings and send them to anyone in sight or in your mind.

Try changing one feeling into another. Start with any emotion you experience at the time or generate a warm feeling by breathing into the abdomen or by imagining an appropriate situation that would evoke it. Intensify the feeling, then imagine how another emotion might feel. In this way, you can experimentally convert love into anger and then back into love. This will show you that you do not need to be at the mercy of emotions arising as a reaction to outside events, but that you can make a decision as to how you will feel.

5. REACTING POSITIVELY: After learning to remain centred in self-generated positive feelings, we need to learn to react positively and appropriately to the many frustrations in our lives. This is possible when we are guided by spiritual beliefs. After adopting a spiritual philosophy, it is easier to see everything that happens to us as a test of our progress.

Every day we have many little tests. It is all right to feel a burst of anger if we break a valuable item, but we can also just laugh about our clumsiness or look for a reason why it happened. It does not matter too much how we react, as long as it is appropriate. If we feel anger, then it is appropriate to feel it and not rationalize it. After having felt the anger, we should not remain stuck in it, but use our mind to get out of it, by seeing the funny side of it or by trying to learn from it.

The basic idea is to feel whatever emotion comes up, but then to modify our expression according to our understanding of the situation from a spiritual point of view. Our emotional reactions are a reflection of our beliefs, and the more spiritual these become, the more positive will be our reactions. The ideal is to be able to express our feelings naturally in a way that has a positive influence on others.

The Mental Level: Our mind is the most important level to heal because here we make the conscious decisions that affect all other levels. Our aim is to free ourselves of negativity and develop positive attitudes. In order to free yourself of negative beliefs and thoughts, you must first recognize them when they occur. Develop the habit of having a small part of your awareness in the background as an observer. Whenever you have a negative thought, your observer will ring an alarm bell so you can realize what you have done. Then, as a mental exercise, find the corresponding positive thought.

As an example, you may watch someone passing by and observe that she is overweight. If you notice having a critical thought about her weight, mentally apologize to her and either send her a mental compliment about some other aspect of her appearance or just wish her well. It is all right to observe, just refrain from uninvited judgments, especially of a negative kind. The habit of critically evaluating others is the opposite of an invited critical and constructive assessment in order to help someone.

Another important aspect in controlling our mind is to refrain from talking about other people, especially mutual friends, relatives, or acquaintances, except when exchanging some relevant factual information. Anything more is gossiping, judging, or criticizing and will delay progress on the spiritual path.

Practice discernment as well as creative and intuitive thinking. Creative thinking means finding answers for spiritual or theoretical questions, or solutions for practical problems; intuitive thinking means we focus on a question or problem in a meditative state and then translate the realization into words.

Helpful Tips: Any kind of worry is detrimental to your health. Worry is a form of fear and that is the opposite of positive expectation and faith. To a large degree, we shape our own future by what we expect. Our inner self tries to manifest what we expect. Our expectation acts like a magnet to attract and draw the expected event into our outer reality. If we worry about anything, then we expect a negative outcome and help to make it come true.

Therefore, our key lesson to learn at the mental level is to know what we want and then, without wavering, expect it to manifest at the appropriate time. Except where timing is essential, it is best not to set a definite time limit. If we want to use "magic" to help our expectation to manifest, use energy imagery: Visualize the desired result and fill it with emotional energy. The critical factor then is your unwavering faith until the expected event appears.

This brings us to the next important step, to know what to manifest. It is said that we should be very careful with our wishes because they might come true. That means that we are usually ignorant about the hidden implications of a fulfilled desire. Often there is a price to pay. We may delay our spiritual progress or even take on a burden of another individual whom we try to help. We may still have certain tests to pass in this lifetime, and it would be detrimental to our spiritual progress to delay or avoid them

A good way of overcoming our weaknesses is by acting as if we have already done so. With every doubtful or important decision, we can ask ourselves what a spiritual master would do in this situation and then just do it.

Remember the advice given to Jonathan Livingston Seagull by his teacher: To be instantly in another place, just imagine and know that you are already there! If you habitually act like a spiritual master, then you are a spiritual master. Your "Inner Master" is there all the time waiting to guide you. Never mind a few mistakes or wrong decisions; there is nothing better to learn from than our own mistakes. You still have many years ahead of you to practice doing the right thing most of the time. When you have reached a level where your motives are unselfish, you do not need to be too concerned about making the right decision. Just trust your intuition or inner guidance. Whatever you choose will usually be right.

Remember the Mind Tools: Prayer, meditation, and studying scriptures were the original mind tools in the West, while in the East visualization, guided imagery, and mantras were commonly used. All of these are still useful.

Develop a daily routine in which you combine several of these methods. This may, for instance, start with prayer and/or affirmations, followed by relaxation, then guided imagery, and finally meditation. During the activities of the day, practice centering. A good way to remain centred is to use a mantra. Another powerful exercise is to imagine being purified by a violet flame. Imagine that the violet flame burns up your remaining or hidden impurities, especially at the emotional and mental levels.

You can also make an effort to become aware of your Christ Self or God Presence, at first in meditation and as guided imagery, but gradually in your daily life. Imagine your God Presence as an intense white spiritual light over your head, or feel or sense it as a protective and loving presence. Initially imagine sending an offering of love or devotion through the top of your head straight up to your God Presence or, if lying down, send it up through the forehead or from the heart centre. Then imagine a strong beam of white light together with an intensified feeling of love and protection entering your top centre and proceeding through the forehead and throat centres to the heart centre. At the same time, feel and see your whole body surrounded by a greatly expanded aura with beautiful rainbow colours. Frequently connect with your Christ or God Presence in this or a similar way, especially before any prayer or healing work.

Cultivating a Healthy Outlook: We make progress on the spiritual path if in our daily life, and especially in difficult situations, we display spiritual qualities such as compassion, devotion, faith, forgiveness, harmony, high ideals, joy, kindness, oneness, openness, patience, self-responsibility, serenity, tolerance, unconditional love, understanding, wisdom, and a gentle love for All That Is.

Another way of assessing our progress is by how well we are able to unite opposite characteristics within our personality and how well we manifest the appropriate one for the situation at hand. For example, are we able to be a good leader and follower, master and servant? Can we display the “male” characteristics of assertiveness, aggressiveness, and power as well as the “female” ones of gentleness, kindness, and yielding? Do we know how to be strong in our beliefs and open to change?

After years of diligently working to purify and improve yourself, you may wonder when you will have your mind-blowing illumination. That expectation is counterproductive. The illumination experience is an act of grace, a gift of your spirit that will come when you least expect it. It may be during a meditation or in the middle of city traffic. Waiting for it is a sure way of postponing it, because patience is one of the early virtues we have to learn.

When it finally does happen, enjoy it. You do not need to look actively for good deeds to do or for your future field of selfless service. Just continue doing whatever needs to be done at each moment, and especially continue purifying yourself at deeper and deeper levels and uplifting yourself to higher and higher levels. At the right time, your mission will be revealed to you. When you have become a master of yourself, you will be suitable to become a servant of humanity.

Looking back on the path that you travelled so diligently, you may ask yourself “What path?” and wonder if there ever was a path or only a figment of your imagination. In reality, there was nowhere to go and nothing to do to get there. All you did was to change some of your perceptions and attitudes. There is a Buddhist saying: “Before enlightenment he gathered wood and carried water; after enlightenment he gathered wood and carried water.”

* * * * * * * * * * * *

A Review of the 65 Steps

The concept of spirituality means different things to different people. In this If you decide to follow the steps outlined in this book, you will have months of interesting work and play ahead of you. Try to make it fun by involving relatives and friends. You may find it helpful to keep a diary to record your experiments, reactions, and insights.

For a solid foundation of your health, get the basics right with Part 1. Start with sanitizing the intestinal tract, cleaning it out, replacing pathogenic microbes with healthy bacteria, and then keeping it all clean. Then you clean out the rest of the body in a series of fasts or cleansing periods. It will be good to continue with a cleanse once a year for the rest of your life. Then test your foods for allergy and sensitivity reactions. Use clean and preferably energized water, adjust your body acidity, and start improving your living conditions. Sanitize your house and minimize electromagnetic pollution. Get into the habit of doing exercises.

These steps provide the basic physical foundation for your health. Take time to do them thoroughly. This can cause occasional healing reactions, but these are beneficial; just take it easy for a few days.

Part 2 explains commonly used healing techniques. The diagnostic techniques show where your problems are and what to do to become more balanced. Muscle testing is easy to learn and very useful for determining whether a food or remedy is good or bad for you. Reflexology is one of my favourites; it is quick and effective for pains and problems in all parts of the body. Magnet therapy has the advantage of being a pain-free treatment. You will find various types of packs useful for many different conditions.

Part 3 contains information on nutrients and remedies. If you sense you require additional nutrients and remedies, then you need a basic understanding of their functions and interactions and how to use them, which this chapter provides.

Part 4 shows you what a good diet is like: Eat as much as conveniently possible of fresh, raw, and organic foods. Change gradually to give your body and taste buds time to adjust.

Part 5 tells you why the foods you probably like are not good for you. Do your best; the prospect of improved well-being will make it easier to give up a favorite food. Finding out if you need a diet high or low in protein and fat is a key requirement for your health.

Part 6 gives natural healing methods for a wide range of health problems; use any of these methods in addition to a healthy lifestyle.

Part 7 introduces you to the subtle energies. Most of us can learn to feel energy, commonly as warmth and tingling. For our health, it is especially valuable to use breathing exercises to activate these energies in our body and direct them to organs in need of healing.

Part 8 handles another key subject: healing emotions. The goal is to live life in a radiant way, filled with love and joy.

Part 9 shows how to appreciate the power of your subconscious mind and make it your friend, how to combine a positive expectation with the help of your subconscious friend so you can heal almost anything. Here you will find many valuable mind tools that make it possible.

Part 10 shows you how to use spiritual guidance to overcome health and emotional problems as part of a spiritual path that provides the framework for your healing venture and gives meaning to your life.

* * * * * * * * * * *

CHINESE MEDICAL QIGONG THERAPY:

A Comprehensive Clinical Text

中国气功医疗

DR. JERRY ALAN JOHNSON PH.D., DTCM, DMQ (CHINA)

CHINESE MEDICAL QIGONG THERAPY:

A COMPREHENSIVE CLINICAL TEXT

JERRY ALAN JOHNSON, PH.D., D.T.C.M., D.M.Q. (CHINA)

EDITED BY
JAMPA MACKENZIE STEWART, D.O.M., L.AC.
AND
MADELEINE H. HOWELL, M.F.T., M.Q.T.

First published in May 2000 by:
The International Institute of Medical Qigong
P.O. Box 52144
Pacific Grove, CA 93950
U.S.A.

© 2000 Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)

All rights reserved under the International and Pan-American copyright conventions. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher (The International Institute of Medical Qigong). Reviewers may quote brief passages.

ISBN# 1-885246-08-0
Printed in the United States of America.

Disclaimer:

Qigong medicine is not intended to replace orthodox medicine, but rather to complement it. The meditations, practices, techniques and prescriptions described herein are powerful and may be too mentally and physically demanding for some people. The readers should therefore use their own discretion and consult a doctor of Medical Qigong therapy, an acupuncturist, medical doctor, or mental health professional before engaging in these exercises and meditations. The author, the International Institute of Medical Qigong, and the publishers are neither liable or responsible to any person or entity with respect to any loss or damage caused, or alleged to be caused, directly or indirectly by reading or following the instructions for any condition, or interpreting information provided in this text. The treatments mentioned in this book are not meant to be used as symptomatic prescriptions. The treatment of specific organs, channels, channel points, and prescriptions must always be selected based on the thorough understanding of the origin of the patient's disease. If an ailment is severe, or if symptoms persist, please consult a medical professional immediately.

Throughout the text I will refer the doctor to prescribe herbs for certain conditions along with Medical Qigong therapy. All Medical Qigong prescriptions, exercises and "homework" assigned to patients in Section X "Treatment of Internal Diseases with Qi Emission Therapy" and Section XI "Specialized Qi Emission Therapy for Pediatrics, Geriatrics, Gynecology, Neurology, Psychology, Oncology, and Surgery" should also be included with herbal prescriptions, as well as the regulation of the patient's diet and living environment. Herbal prescriptions will vary according to the patient's constitution, condition and specific illness, and must be treated accordingly only by a doctor or herbalists qualified to prescribe Chinese medical herbs.

Throughout the text I have used the term *doctor* when referring to professional practitioners who use energetic medicine to treat patients. The word "doctor" means "to teach". I believe that the foremost duty of any doctor of medicine (Western or Chinese) should be as educator, to teach his or her patients knowledge and skills for the prevention and treatment of disease and injury. Currently, the official title, *Doctor of Medical Qigong Therapy* is only licensed and used in China.

TABLE OF CONTENTS

FORWARD XXIII

PREFACE XXXI

ACKNOWLEDGMENTS XXXV

INTRODUCTION XXXVII

SECTION I

FOUNDATIONS OF CHINESE ENERGETIC MEDICINE 1

CHAPTER 1 INTRODUCTION TO MEDICAL QIGONG 3
- THE FIVE DOMINIONS OF ENERGY 3
- DEFINING THE ENERGY OF YIN AND YANG 5
- UNDERSTANDING THE CONCEPT OF QI 5
- DIFFERENT SCHOOLS OF QIGONG 5
- MEDICAL QIGONG DEFINED 6
- MEDICAL QIGONG TRAINING IN CHINA 7
- MEDICAL QIGONG TRAINING IN THE UNITED STATES 8
- TRADITIONAL CHINESE MEDICINE AND MEDICAL QIGONG THERAPY 14
- ACUPUNCTURE THERAPY 15
- HERBAL THERAPY 18
- CHINESE MASSAGE THERAPY 19
- MEDICAL QIGONG THERAPY 22

CHAPTER 2 ENERGETIC FORMATION OF THE HUMAN BODY 25
- ENERGETIC EMBRYOLOGICAL DEVELOPMENT 25
- ENERGY, MATTER, AND SPIRITUAL INTERACTIONS 26
- UNDERSTANDING FETAL TOXINS 26
- DEVELOPMENTAL SEQUENCE 27
- THE TEN LUNAR MONTHS OF CREATION 29
- POSTNATAL ENERGY DEVELOPMENT 41
- ENERGETIC EMBRYOLOGICAL OVERVIEW 41
- THE ETERNAL SOUL 42
- SUMMARY 51

CHAPTER 3 PHYSICAL DEVELOPMENT AND STRUCTURAL FORMATION 53
- FASCIAL DEVELOPMENT AND ENERGY FLOW 53
- YIN AND YANG STRUCTURAL FORMATION 53
- THE YIN AND YANG OF BODY ASPECTS 55
- THE FIVE ELEMENTS 56
- CONGENITAL CONSTITUTIONS 57
- CLASSIFICATION OF THE FIVE CONSTITUTIONS 57
- THE WOOD CONSTITUTION 58
- THE FIRE CONSTITUTION 60
- THE EARTH CONSTITUTION 61
- THE METAL CONSTITUTION 62
- THE WATER CONSTITUTION 63

COMBINED CONSTITUTIONS 64
PERSONALITY CONSTITUTIONS OF THE EIGHT EXTRAORDINARY VESSELS 65
THE YAO IMAGE AND PHYSICAL ANATOMY 66
IDENTIFYING INTROVERTED AND EXTROVERTED STRUCTURES 71
CONGENITAL AND ACQUIRED CELLULAR PATTERNS 73
SUMMARY 73

CHAPTER 4 THE FIVE ENERGIES OF THE HUMAN BODY 75
SOUND ENERGY RESONANCES 75
LIGHT ENERGY 77
MAGNETIC ENERGY 82
GENERATED HEAT 84
ELECTRICITY 86

CHAPTER 5 THE THREE DANTIANS AND THE TAIJI POLE 89
THE THREE DANTIANS 89
THE LOWER FIELD OF ELIXIR 90
THE MIDDLE FIELD OF ELIXIR 95
THE UPPER FIELD OF ELIXIR 98
THE FUNCTIONAL ASPECTS OF THE THREE DANTIANS 102
THE DANTIAN'S YIN AND YANG ENERGETIC CHAMBERS 104
THE DOCTOR'S PROJECTED AURA FIELDS 105
THE TAIJI POLE 105
THE TWELVE GATES OF THE CHAKRA SYSTEM 109
TAKING PRECAUTIONS 115
THE TWELVE EARTHLY BRANCHES AND THE TWELVE CHAKRA GATES 115

CHAPTER 6 THE TWELVE PRIMARY CHANNELS AND ORGANS 117
INTRODUCTION 117
CLASSIFICATION OF CHANNELS 118
THE CHANNELS' RELATIONSHIP TO QI AND BLOOD 121
FUNCTION OF THE CHANNELS 122
INTERNAL AND EXTERNAL CHANNEL FLOW 122
ORGAN AND CHANNEL FUNDAMENTALS 123
THE GALL BLADDER CHANNELS 124
THE LIVER CHANNELS 126
THE LUNG CHANNELS 128
THE LARGE INTESTINE CHANNELS 130
THE STOMACH CHANNELS 132
THE SPLEEN CHANNELS 134
THE HEART CHANNELS 136
THE SMALL INTESTINE CHANNELS 138
THE BLADDER CHANNELS 140
THE KIDNEY CHANNELS 142
THE PERICARDIUM CHANNELS 145
THE TRIPLE BURNER CHANNELS 147
UNDERSTANDING CHANNEL PATHOLOGY 151

CHAPTER 7 THE EIGHT EXTRAORDINARY VESSELS AND COLLATERALS 155
THE EIGHT EXTRAORDINARY VESSELS 155
THE FUNCTION OF THE EIGHT EXTRAORDINARY VESSELS 155
THE EIGHT EXTRAORDINARY VESSELS AND MEDICAL QIGONG THERAPY 157

CLINICAL USE OF THE EIGHT EXTRAORDINARY VESSELS 157
ENERGETIC FUNCTIONS OF THE EIGHT EXTRAORDINARY VESSELS 158
THE GOVERNING AND CONCEPTION VESSELS 158
THE GOVERNING VESSEL OR DU VESSEL 159
THE CONCEPTION VESSEL OR REN VESSEL 162
THE THRUSTING VESSELS AND THE BELT VESSELS 164
THE BELT VESSEL OR DAI VESSEL 166
THE YIN AND YANG HEEL VESSELS OR QIAO VESSELS 168
THE YIN HEEL VESSELS 168
THE YANG HEEL VESSELS 169
THE YIN AND YANG LINKING VESSELS OR WEI VESSELS 171
THE YIN LINKING VESSEL 171
THE YANG LINKING VESSELS 172
SUMMARY OF EIGHT EXTRAORDINARY VESSEL PATHOLOGY 173
THE FIFTEEN MAJOR COLLATERALS (CONNECTING VESSELS) 175
THE TWELVE DIVERGENT CHANNELS 180
THE TWELVE MUSCLE/TENDON CHANNELS 186
THE TWELVE SKIN ZONES 197
THE CONNECTIVE TISSUE OF THE TWELVE SKIN ZONES 197
CHANNEL AND COLLATERAL THERAPY 200
CONTROLLING QI EXTENSION THROUGH THE CHANNELS 201

CHAPTER 8 THE BODY'S ENERGETIC POINTS 203
INTRODUCTION TO ENERGETIC POINTS 203
THE FOUR WAYS POINTS ARE FORMED 203
SCIENTIFIC RESEARCH OF ACUPOINTS 203
THE FOUR CATEGORIES OF ENERGETIC POINTS 204
THE TWO GATES OF ENERGETIC POINTS 204
THE THREE FUNCTIONS OF ENERGETIC POINTS 205
POINT NAMES 207
CLASSIFICATION OF ENERGETIC POINTS 207
SUMMARY OF POINTS 212

CHAPTER 9 THE EXTRAORDINARY YANG ORGANS 215
THE UTERUS 215
THE BRAIN 217
THE GALL BLADDER 218
THE MARROW 218
THE BONES 218
THE BLOOD VESSELS 221

SECTION II
THE THREE OUTER FORCES: HEAVEN, EARTH & MAN 223

CHAPTER 10 INTRODUCTION TO THE THREE OUTER FORCES 225
THE NATURAL POWER OF THE THREE OUTER FORCES 225
THE THREE OUTER FORCE'S NINE TREASURES 225
THE THREE OUTER FORCES AND CHINESE ALCHEMY 225
HEAVENLY QI 226
EARTHLY QI 228
MAN'S QI 229

Chapter 11 The Three Treasures of Heaven: Sun, Moon, and Stars 231
- Defining Heaven Qi 231
- Influences of Heavenly Qi on the Human Body 232
- Body Rhythms and Energetic Patterns 232
- Cycles of the Sun 234
- The Twelve Pi Hexagrams 234
- The Four Principal Time Periods of Qigong Practice 236
- The Four Seasonal Peak Transitions 237
- The Six Principal Seasons of Transition 239
- The Daily Qi, Blood, and Heat Cycle 239
- Gathering The Sun's Essence Meditation 240
- Cycles of the Moon 242
- Prenatal Eight Trigrams and the Moon 244
- Choosing the Proper Direction for Training Qi 244
- Cycles of the Stars 248
- Earthly Branches, Heavenly Stems, and Nine Palaces 254
- Summary 255

Chapter 12 The Three Treasures of Earth:Earth, Water, and Wind 259
- Introduction to the Three Treasures of Earth 259
- Gathering Energy from the Earth's Energetic Fields 260
- Techniques for Cultivating Earth Qi 260
- Gathering Qi From Trees, Bushes, and Flowers 261
- Gathering Qi From Mountains, Valleys, and Deserts 263
- Gathering Qi From Oceans, Lakes, and Streams 263
- Gathering Qi from the Six Directions 264
- Additional Techniques for Training Earth Qi 265
- The Earthly Treasure of Food 266
- Understanding the Yin or Yang Aspect of Food 266
- Understanding the Yin or Yang Aspect of Herbs 266
- Energy, Food, and Diet 267
- Using The Five Flavors of Food 268
- Guidelines for the Five Elemental Tastes 270
- Overview of the Five Elemental Tastes 271
- Summary 271

Chapter 13 The Three Treasures of Man: Jing, Qi, and Shen 273
- Prenatal and Postnatal Jing, Qi, and Shen Transformations 274
- Jing: Essence of Life 276
- The Yin and Yang Aspects of Jing 276
- Prenatal Jing 276
- Postnatal Jing 277
- The Interaction of Prenatal and Postnatal Jing 277
- Qi: Vital Life-Force Energy 278
- Eighteen Classifications of Qi 279
- The Six Functions of Qi 281
- Qi and Energetic Medicine 283
- Qi and the Body's Energetic Structure 283
- Shen: The Spirit of Life 290
- Shengong Schools of Meditations and Training 290
- Prenatal and Postnatal Shen 291

The Five Spiritual Aspects and Yin Organ Components of Shen ... 291
The Liver Stores the Essence Spirit "Hun" ... 292
The Heart Stores the Essence Spirit "Shen" ... 295
The Spleen stores the Essence Spirit "Yi" ... 296
The Lungs Store the Essence Spirit "Po" ... 296
The Kidneys Store the Essence Spirit "Zhi" ... 298
The Five Yin Organ Cultivation of Shen ... 298
Shengong Cultivation Methods ... 300
The Six Transportation's of Shen ... 301
The Eight Supernatural Powers ... 304
Soul Projection, Spirit Projection and Energy Projection ... 306

Section III

Dao Yin Training ... 311

Chapter 14 Introduction To Dao Yin Training ... 313
Static Qigong Training ... 314
Dynamic Qigong Training ... 314
Principles of Yin and Yang in Dao Yin Training ... 314
Three Foundational Methods of Dao Yin Training ... 315

Chapter 15 Postural Dao Yin Training ... 317
Static Posture Dao Yin Training ... 317
Choosing the Proper Posture ... 318
Primary and Secondary Postures ... 318
The Eighteen Rules of Proper Medical Qigong Structure ... 321
Dynamic Posture Dao Yin Training ... 330
Using the Energy of the Hands to Lead Qi ... 332
Medical Qigong Walking Therapy ... 334
Walking Combined with Color Therapy ... 338
Walking Combined with Sound Therapy ... 338
The Therapeutic Effects of Medical Qigong Walking Therapy ... 338
Three Stepping Methods Used to Treat Cancer ... 342
Ending and Closing the Walking Therapy ... 344

Chapter 16 Respiratory Dao Yin ... 349
The Purpose of Respiratory Dao Yin ... 349
The Role of the Lungs in Qi Cultivation ... 349
Deep Inhalation and Exhalation Breathing Methods ... 349
Seven Dao Yin Breathing Techniques ... 350
Respiratory Guiding and Regulating ... 353
Respiratory Tonification and Purging Methods ... 354
Basic Essentials of Respiratory Training ... 354
Breathing Patterns for Collecting and Regulating Qi ... 355
The Four Progressive Stages of Respiration ... 356
Duration Time for Respiration Therapy ... 357
Emotional Effects on Breathing Patterns ... 357
Respiratory Dao Yin Cancer Prescriptions ... 357

Chapter 17 Mental Dao Yin Training ... 359
Introduction ... 359
The Prenatal and Postnatal Mind ... 360

BENIGN AND MALIGNANT THOUGHT PATTERNS 361
THE IMPORTANCE OF MIND CONCENTRATION 361
REGULATING THE MIND 362
THE THREE CATEGORIES OF MENTAL CONCENTRATION 363
ESSENTIALS OF TRAINING MIND CONCENTRATION 364
CONCENTRATIVE MEDITATION AND INSIGHT MEDITATION 365
THE THREE STAGES OF MENTAL TRANQUILITY 366
THE THREE PROCESSES OF THE MIND 366
CREATING THE BRIDGE OF LIGHT 367
REPROGRAMMING THE MIND TO HEAL 369
USING THE IMAGINATION 369
TWELVE STAGES TO TRANSCEND AND TRANSFORM 370
SUMMARY 374

SECTION IV
QI DEVIATIONS IN QIGONG TRAINING 375

CHAPTER 18 INTRODUCTION TO QI DEVIATIONS 377
THE MAIN CAUSES OF QI DEVIATIONS 377
CORRECTING QI DEVIATIONS 379
SHEN DISTURBANCES AND EMOTIONAL DYSFUNCTIONS 386
UNCONTROLLABLE MUSCLE DEVIATIONS 391
RECTIFYING QI DEVIATIONS USING TAIJI RULER 391

CHAPTER 19 DISORDERS AND DEVIATIONS OF THE SOUL AND SPIRIT 393
INTRODUCTION TO SPIRITUAL DIMENSIONS 393
EVIL STATES AND MENTAL DELUSIONS 395
PROJECTED SPIRIT DEMONS FROM THE SUBCONSCIOUS MIND 397
ACTUAL ENCOUNTERS WITH SPIRIT DEMONS 397
OVERVIEW 399
MULTIPLE PERSONALITY DISORDER (DISSOCIATIVE IDENTITY DISORDER) 400
SOUL RETRIEVAL AND MEDICAL QIGONG THERAPY 401
WOUNDING AND CLOSING THE SPIRIT 403
HEALING THE SPIRIT 405
EMOTIONAL REBIRTHING PATTERNS 406
SOUL RETRIEVAL MEDITATION 406
THE DOOR TO THE CHAMBER OF MEMORIES 411
FINDING THE LIFE PURPOSE 411

SECTION V
DIFFERENTIAL DIAGNOSIS OF ENERGETIC PRINCIPLES 413

CHAPTER 20 INTRODUCTION TO DIFFERENTIAL DIAGNOSIS 415
DIAGNOSIS IN ENERGETIC MEDICINE 415
ETIOLOGY IN MEDICAL QIGONG THERAPY 415
THE GROWTH AND PEAK CYCLES OF JING, QI AND SHEN 416
DISORDERS OF THE HUMAN BODY 417
THE DIFFERENTIATION OF SYNDROMES 419
THE CATEGORIZATION OF SIGNS AND SYMPTOMS 420
DIAGNOSIS AND TREATMENT OF THE PATIENT 420
UNDERSTANDING THE THEORY OF ENERGETIC COUNTERACTION 422

CHAPTER 21 DIAGNOSIS ACCORDING TO THE EIGHT ENERGETIC PRINCIPLES 423
- INTERNAL AND EXTERNAL PATHOGENIC FACTORS 423
- HOT AND COLD 428
- EXCESS AND DEFICIENT 430
- YIN AND YANG THEORY 431
- DIAGNOSING YIN AND YANG IN THE BODY 432
- SIGNS OF YIN AND YANG 435
- YIN AND YANG CATEGORIZATION OF TRAINING 436
- PULSE DIAGNOSIS OF THE EIGHT ENERGETIC PRINCIPLES 436
- PULSE QUALITIES AND CLASSIFICATIONS 437
- THE CLASSICAL PULSE DIAGNOSTIC POSITIONS AND DEPTH 438
- OVERVIEW 439

CHAPTER 22 DIAGNOSIS ACCORDING TO QI, BLOOD, AND BODY FLUID DYSFUNCTIONS 441
- INTRODUCTION 441
- DISORDERS OF THE QI 441
- LEAKAGE AND LOSS OF QI 442
- EXCESS QI DISORDERS 444
- QI STAGNATION DISORDERS 444
- REBELLIOUS QI DISORDERS 445
- TURBID QI 445
- RECTIFYING QI DISORDERS 446
- DISORDERS OF THE BLOOD 447
- DEFICIENCY OF BLOOD 447
- THREE DISORDERS OF BODY FLUIDS 448
- TYPES OF PHLEGM 450

CHAPTER 23 DIAGNOSIS ACCORDING TO THE FIVE-ELEMENT THEORIES 451
- INTRODUCTION TO THE FIVE ELEMENTS 451
- THE FIVE ELEMENTS AND YIN AND YANG THEORY 451
- THE CREATIVE OR GENERATIVE CYCLE 453
- THE CONTROLLING OR RESTRICTIVE CYCLE 453
- THE INVADING OR OVERACTING CYCLE 454
- THE INSULTING OR COUNTERACTING CYCLE 455
- THE FIVE ELEMENTAL TONIFICATION CYCLE 456
- FIVE YIN ENERGETIC ORGAN MANIFESTATIONS 456

CHAPTER 24 OTHER DIAGNOSTIC SYSTEMS AND TECHNIQUES 459
- DIAGNOSIS ACCORDING TO THE SIX STAGES 459
- DIAGNOSIS ACCORDING TO THE FOUR LEVELS 460
- DIAGNOSIS ACCORDING TO THE TRIPLE BURNERS 462

CHAPTER 25 CLINICAL EXAMINATION AND DIAGNOSIS 465
- SENSORY, INTUITIVE AND PERCEPTUAL DIAGNOSIS 465
- SENSORY DIAGNOSIS 465
- CLINICAL INTAKE, EXAMINATION AND EVALUATION 465
- EXAMINATION FOR DIAGNOSIS 467
- DEVISING A TREATMENT PRINCIPLE 468
- TREATMENT STRATEGY 468
- TREATMENT 469
- AFTER THE TREATMENT 469

ALTERNATIVE TREATMENT STRATEGIES 470
INTUITIVE AND PERCEPTUAL DIAGNOSIS 470
THE METAPHYSICAL DIMENSIONS OF MEDICAL QIGONG HEALING 470
THE AWAKENING 471
LEARNING INTUITIVE AND PERCEPTUAL DIAGNOSIS 471
RECEIVING AND INTERPRETING THE PATIENT'S FIELDS OF BIO-INFORMATION 472
ACCESSING THE KNOWLEDGE STORED WITHIN THE WUJI 473
UNDERSTANDING PAST-LIFE REGRESSION 474
LEARNING MEDICAL QIGONG TECHNIQUES 476
PERCEPTUAL COMMUNICATION TRAINING 477
APPLICATIONS OF INTUITIVE AND PERCEPTUAL QI DIAGNOSIS 479
DIAGNOSIS BASED ON FLAT-PALM DETECTION (HAND SENSING) 479
DIAGNOSIS BASED ON KINESTHETIC AND EMPATHIC QI ABSORPTION 480
HAND DIAGNOSIS BASED ON BODY ASSOCIATION 481
DIAGNOSIS BY OBSERVING AURA FIELDS 483
DIAGNOSIS BASED ON INTENTION 487
LONG DISTANCE MEDICAL QIGONG THERAPY 487
FAITH PROJECTION AND MEDICAL QIGONG THERAPY 490
AUXILIARY MEDICAL QIGONG HEALING MODALITIES 492
USING MEDICAL QIGONG TO HEAL BY PROJECTING CHINESE CHARACTERS 492
USING MEDICAL QIGONG WITH SUPPLEMENTARY APPARATUSES 492

SECTION VI

ESTABLISHING THE MEDICAL QIGONG CLINIC 497

CHAPTER 26 TREATING THE PATIENT 499
SETTING A SPACE/TIME FOUNDATION 499
ESTABLISHING THE MEDICAL QIGONG TREATMENT ROOM 500
ESTABLISHING A RAPPORT WITH THE PATIENT 500
THE FOUR MAJOR PRIORITIES OF CLINICAL QIGONG HEALING 503
PREPARING THE SELF TO TREAT THE PATIENT 504
TREATING THE PATIENT 509
MEDICAL QIGONG TREATMENTS 515
TREATING MULTIPLE PATIENTS 519
ENDING THE TREATMENT 520

CHAPTER 27 COMBINING MEDICAL QIGONG THERAPY WITH OTHER HEALING MODALITIES .. 523
COMBINING MEDICAL QIGONG THERAPY WITH TCM MODALITIES 523
COMBINING MEDICAL QIGONG THERAPY WITH WESTERN MEDICAL MODALITIES 525

CHAPTER 28 PRECAUTIONS AND ETHICAL CONSIDERATIONS FOR THE QIGONG DOCTOR 527
AVOIDING ENERGETIC DEPLETION 527
PREVENTING THE INVASION OF TURBID QI 528
GUIDELINES FOR THE MEDICAL QIGONG DOCTOR 529
WU ZANG (FIVE ORGAN) MEDITATION FOR ENERGETIC PROTECTION 530
CREATING A PROTECTIVE ENERGETIC FORCE FIELD 533
CLINICAL ETHICS 534
CONSCIOUS AND SUBCONSCIOUS CLINICAL INTERACTIONS 535
ESTABLISHING DOCTOR AND PATIENT AGREEMENTS 536
POWER DYNAMICS AND ETHICAL BEHAVIOR 537
PITFALLS 538

CLINICAL BURNOUT AND MEDICAL QIGONG THERAPY 538
ENERGETIC BOUNDARY INVASION 539
STRENGTHENING THE DOCTOR'S ENERGETIC BOUNDARIES 540
CAUSES OF SPIRITUAL DRYNESS 540

SECTION VII
THE TREATMENT PRINCIPLES OF MEDICAL QIGONG THERAPY 543

CHAPTER 29 PRINCIPLES AND PARAMETERS OF MEDICAL QIGONG THERAPY 545
THE FIVE METHODS FOR TREATING PATIENTS 545
THE NINE STAGES OF TREATMENT 545
THE THREE WONDERS OF MEDICAL QIGONG THERAPY 546
CONTRAINDICATIONS FOR MEDICAL QIGONG THERAPY 548

CHAPTER 30 BALANCING THE BODY'S ENERGETIC SYSTEM 551
PURGING, TONIFYING, AND REGULATING 551
THE THREE TREATMENT MODALITIES 551
PURGATION TECHNIQUES 552
TONIFICATION AND PURGATION METHODS 552
TONIFICATION AND PURGATION GUIDELINES 554
UNDERSTANDING MEDICAL QIGONG REGULATION 554
QI REGULATION AND GUIDING PRINCIPLES 555
ENERGY BALANCING PROCEDURES 555
CLINICAL OBSERVATIONS OF PATHOLOGICAL ENERGETIC PATTERNS 556
LOCATIONS OF ENERGETIC BLOCKAGES 556
TYPES OF ENERGETIC COMPLICATIONS 559
PATIENT'S ENERGETIC FACIAL DISTORTION 562
PATIENT RESISTANCE TO ENERGETIC THERAPY 562
UNBLOCKING THE FIVE PASSES 562
FURTHER COMPLICATIONS 564

CHAPTER 31 THE MAGIC SQUARE OF THE YELLOW RIVER 565
THE MAGIC SQUARE AND EIGHT TRIGRAMS 568
THE MAGIC SQUARE AND THE EIGHT EXTRAORDINARY VESSELS 569
UNDERSTANDING THE METAPHYSICAL ASPECT OF NUMBERS 569
THE FIVE ELEMENT ENERGETIC FLOW OF THE MAGIC SQUARE 570
CROSSOVER ENERGY PATTERNS 570

CHAPTER 32 HEALING THE PATIENT'S EMOTIONAL TRAUMAS 573
INTRODUCTION 573
THE PHYSIOLOGICAL EFFECTS OF BENIGN AND CHRONIC EMOTIONS 573
FIVE LEVELS OF FEELING AND THEIR GOVERNING ORGANS 573
STAGES OF EMOTIONAL HISTORY AND RESOLUTION 576
ASSISTING THE PATIENT IN ESTABLISHING HEALTHY BOUNDARIES 578
REJECTION, BETRAYAL, AND ABANDONMENT 578
EMOTIONAL SECURITY 579
HEALING EMOTIONAL TRAUMAS 580
PULLING OUT THE PAIN MEDITATION 581
UNDERSTANDING WHY A PATIENT WON'T HEAL 582
FACING THE DEATH OF A PATIENT 583
SUMMARY 585

SECTION VIII
QI EMITTING METHODS 587

CHAPTER 33 ENERGY EXTENSION AND QI MANIPULATION TECHNIQUES 589
ENERGY PROJECTION 589
TYPES OF ENERGY EXTENSION 589
CONTROLLING ENERGETIC PROJECTIONS 591
HAND POSTURES FOR QI EMISSION 592
THE EXTENDED FAN PALM 592
THE SWORD FINGERS 594
THE ONE FINGER SKILL 595
THE INVISIBLE NEEDLE PALM 597
THE DRAGON'S MOUTH PALM 598
THE KNEADING TIGER PALM 598
THE FIVE THUNDER FINGERS 599

CHAPTER 34 QI EXTENSION, GUIDANCE, AND REGULATION TECHNIQUES 601
QI EXTENSION TECHNIQUES 601
GUIDING THE LINEAR FLOW OF ENERGY 601
CIRCULATING THE QI IN A CIRCLING PATTERN 603
SPIRALING ENERGY FLOW 605
ENERGETIC CUPPING 607
GUIDING THE STATIONARY FLOW OF ENERGY 607
GUIDING THE HOT AND COLD FLOW OF ENERGY 610
CLINICAL APPLICATION OF YIN AND YANG ENERGY EXTENSION 612
GUIDING THE ENERGETIC FLOW OF THE FIVE ELEMENTS 612
SUMMARY OF QI EMITTING METHODS 616

CHAPTER 35 THE INVISIBLE NEEDLE THEORY AND APPLICATION 619
INTRODUCTION 619
PREPARATION 619
ANGLE OF INSERTION FOR TONIFICATION OR SEDATION 620
THE HAND POSTURES OF THE INVISIBLE NEEDLE 621
TIME AND DURATION OF TREATMENT 622
REMOVING THE INVISIBLE NEEDLES AND ENDING THE TREATMENT 622

CHAPTER 36 ENERGETIC POINT THERAPY 623
CHANNEL POINT TREATMENTS AND PRESCRIPTIONS 623
CONTRAINDICATIONS 627
CHANNEL POINT MEDITATION 627

CHAPTER 37 MEDICAL QIGONG MASSAGE 631
INTRODUCTION TO TISSUE FUNCTION 631
TWO APPROACHES IN TREATMENT 632
TISSUE REGULATION THROUGH MEDICAL QIGONG MASSAGE 633
THE FIVE ELEMENTAL QI MASSAGE THERAPY 633
YIN AND YANG ENERGETIC POINT MANIPULATION 636
TREATING THE TISSUE AREA OR ENERGETIC POINT 637
ADDITIONAL GUIDELINES 637

Section IX

Medical Qigong Exercises and Prescriptions 639

Chapter 38 Introduction to Medical Qigong Prescriptions and Regulation Exercises 641

History of Medical Qigong Prescriptions 641

The Function of Medical Qigong Prescriptions 641

Medical Qigong and the Immune System 642

The Five Stages of Healing 645

The Three Categories of Medical Qigong Prescriptions 645

Choosing Prescriptions According to the Patient's Qi Circulation 646

Medical Qigong Prescriptions for General Syndromes 647

Number of Breaths Required in Medical Qigong Prescriptions 647

Chapter 39 Specific Organ Self-Massage Prescriptions 649

Introduction 649

Heart Massage 649

Liver Massage 651

Lung Massage 652

Spleen & Stomach Massage 652

Kidney Massage 653

Chapter 40 Healing Sound Therapy and Prescriptions 657

Introduction and History 657

Sound Volume Resonation 657

High-, Middle-, and Low-Pitched Sound 658

Breath and Mind Control 658

Audible and Inaudible Sound Resonation 659

The Six Healing Sound Prescriptions 660

The Proper Number of Breaths for Healing Sound Practice 665

Combining Multiple Organ Sounds 666

Old Man Searching for the Reflection of the Moon at the Bottom of Tide Pool 666

Contraindications for the Six Healing Sound Therapy 668

Sound Resonation Therapy for Tumor and Cancer 668

Preparation for Tumor and Cancer Sound Therapy 668

The Proper Amount of Tone Regulations 669

Specialized Prescriptions for Tumors and Cancer 669

Contraindications for Tumor and Cancer Sound Therapy 673

Chapter 41 Medical Qigong Regulation Exercises 675

Introduction 675

Starting and Ending the Prescriptions 675

Medical Qigong Cultivation Exercises 676

External Qi Cultivation 676

Internal Qi Cultivation 677

Regulating the Body's Qi 678

The Fire Path of the Microcosmic Orbit 678

The Water Path of the Microcosmic Orbit 681

The Wind Path of the Microcosmic Orbit 681

Medical Qigong Prescriptions 682

Descend the Qi and Cleanse the Organs 682

GATHERING QI IN THE UPPER AND LOWER DANTIANS 684
ROLLING THE BALL WITH BOTH PALMS 686
OPENING AND CLOSING THE UPPER, MIDDLE, AND LOWER BURNERS 687
SUN AND MOON ROTATING TECHNIQUE 689
EXPANDING AND CONTRACTING THE RINGS 691
DESCENDING THE YANG AND ASCENDING THE YIN TECHNIQUE 691
TURNING AND WINDING THE BELT VESSEL TECHNIQUE 693
OPENING AND CLOSING THE HEAVEN AND EARTH TECHNIQUE 695
COLLECT THE QI AND RETURN TO THE LOWER DANTIAN TECHNIQUE 697
QIGONG SELF-HEALING MASSAGE AND POINT THERAPY 698
WASH THE FACE 698
MASSAGE THE HEAD 698
BEATING THE HEAVENLY DRUM 699
PRESSING THE EARS 699
RUBBING THE EYES 699
MASSAGING THE NOSE 700
CLICKING THE TEETH 700
WAGGING THE TONGUE 701
SWALLOWING THE SALIVA 701
WHIRLING THE NAPE AND TURNING THE NECK 701
MASSAGING THE CHEST AND ABDOMEN 702
MASSAGE TAPPING THE HANDS (LI-4) 702
MASSAGE TAPPING THE FOREARMS (PC-6) 702
MASSAGE TAPPING THE ELBOWS (LI-11) 702
MASSAGING UNDER THE ARMS (HT-1) 703
MASSAGE TAPPING THE UPPER SHOULDERS (GB-21) 703
MASSAGING THE ARMS 703
MASSAGE TAPPING THE LOWER DANTIAN (CV-6) 703
MASSAGE TAPPING THE BACK OF THE WAIST (GV-4) 704
MASSAGE TAPPING THE SIDES OF THE HIPS (GB-30) 704
MASSAGE TAPPING THE THIGHS (GB-31) 704
MASSAGE TAPPING THE OUTSIDE OF THE KNEES (GB-34) 704
MASSAGE TAPPING THE INSIDE OF THE KNEES (SP-9) 705
MASSAGE TAPPING THE BACK OF THE KNEES (BL-40) 705
MASSAGE TAPPING BELOW THE KNEES (ST-36) 705
MASSAGE TAPPING ABOVE THE MEDIAL MALLEOLUS (SP-6) 706
MASSAGE TAPPING THE HEELS (KD-1) 706
TREMBLING THE KNEES 706
CIRCLING THE ABDOMEN 706
COMPLETING THE EXERCISE 707

CHAPTER 42 DAOIST QIGONG REGULATION EXERCISES 709
DAOIST FIVE YANG ORGAN REGULATION EXERCISE 709
DAOIST FIVE YIN ORGAN TONIFICATION AND REGULATION EXERCISES 710

SECTION X

TREATMENT OF INTERNAL DISEASES WITH QI EMISSION THERAPY 723
INTRODUCTION 724

CHAPTER 43 QI EMISSION THERAPY FOR LIVER DISEASES 725
INTRODUCTION 725
LIVER COMPLICATIONS AND SYMPTOMS 726

GENERAL TREATMENT FOR LIVER DISEASES 726
MEDICAL QIGONG THERAPY FOR LIVER QI DYSFUNCTION 726
LIVER QIGONG PRESCRIPTIONS 728
MEDICAL QIGONG THERAPY FOR SPECIFIC LIVER DISEASES 730
DISORDERS OF THE BILIARY TRACT 730
GALLSTONES 731
CIRRHOSIS 732
CHOLECYSTITIS 733
HYPOCHONDRIAC PAIN 734

CHAPTER 44 QI EMISSION THERAPY FOR HEART DISEASES 735
INTRODUCTION 735
HEART COMPLICATIONS AND SYMPTOMS 735
GENERAL TREATMENT FOR HEART DISEASES 736
MEDICAL QIGONG THERAPY FOR HEART QI DYSFUNCTION 737
HEART QIGONG PRESCRIPTIONS 737
SPECIFIC MEDICAL QIGONG THERAPY FOR HEART DISEASES 738
HYPERTENSION 738
PALPITATION 742
RHEUMATIC HEART DISEASE 743

CHAPTER 45 QI EMISSION THERAPY FOR SPLEEN/STOMACH DISEASES 745
INTRODUCTION 745
SPLEEN COMPLICATIONS AND SYMPTOMS 745
GENERAL TREATMENT FOR SPLEEN DISEASES 746
MEDICAL QIGONG THERAPY FOR SPLEEN QI DYSFUNCTION 747
SPECIFIC MEDICAL QIGONG THERAPY FOR SPLEEN DISEASES 749
DIABETES 749
MEDICAL QIGONG THERAPY AND STOMACH DISEASES 750
STOMACH COMPLICATIONS AND SYMPTOMS 750
GENERAL TREATMENT FOR STOMACH DISEASES 751
SPECIFIC MEDICAL QIGONG THERAPY FOR STOMACH DISEASES 751
ABDOMINAL PAIN 751
CHRONIC GASTRITIS 754
GASTRIC AND DUODENAL ULCERS 755
CHRONIC CONSTIPATION 756

CHAPTER 46 QI EMISSION THERAPY AND LUNG DISEASES 759
INTRODUCTION 759
LUNG COMPLICATIONS AND SYMPTOMS 759
GENERAL TREATMENT FOR LUNG DISEASES 761
MEDICAL QIGONG THERAPY FOR LUNG QI DYSFUNCTION 761
LUNG QIGONG PRESCRIPTIONS 761
SPECIFIC MEDICAL QIGONG THERAPY FOR LUNG DISEASES 762
TREATING THE COMMON COLD AND INFLUENZA 762
CHRONIC BRONCHITIS 764
ASTHMA 764
BRONCHIECTASIS 766
SINUSITIS 767

CHAPTER 47 QI EMISSION THERAPY FOR KIDNEY DISEASES 769
INTRODUCTION 769
KIDNEY COMPLICATIONS AND SYMPTOMS 770
GENERAL TREATMENT FOR KIDNEY DISEASES 770
MEDICAL QIGONG THERAPY FOR KIDNEY QI DYSFUNCTION 771
KIDNEY QIGONG PRESCRIPTIONS 772
SPECIFIC MEDICAL QIGONG THERAPY FOR KIDNEY DISEASES 773
TINNITUS AND DEAFNESS 773
LUMBAGO 774
INSOMNIA 775
PYELONEPHRITIS 776
IMPOTENCE 777

CHAPTER 48 QI EMISSION THERAPY FOR MISCELLANEOUS DISEASES 781
INTRODUCTION 781
DEFICIENT DANTIAN QI 781
PRESCRIPTIONS FOR REGULATING THE THREE DANTIANS 781
IMPROPER YIN AND YANG ENERGY BALANCE 782
DISORDERS OF THE QI 783
REBELLIOUS QI DISORDERS 784
ARTHRITIS 784
VERTIGO 786
HEADACHES 787
MIGRAINE HEADACHES 789

SECTION XI
SPECIALIZED QI EMISSION THERAPY FOR PEDIATRICS, GERIATRICS, GYNECOLOGY, NEUROLOGY, PSYCHOLOGY, ONCOLOGY, AND SURGERY 793

INTRODUCTION 794

CHAPTER 49 QI EMISSION THERAPY FOR PEDIATRICS 795
INTRODUCTION 795
EARLY STAGES OF CHILD GROWTH AND DEVELOPMENT 795
PATHOLOGY 796
DEFICIENT LUNGS 796
DEFICIENT SPLEEN 796
DEFICIENT KIDNEYS 796
LIVER YANG EXCESS 796
HEART (SHEN) IMBALANCE 796
HOT AND COLD EXOGENOUS FACTORS 797
THERAPY 797
ENERGETIC MOVEMENT AND FLOW 797
MEDICAL QIGONG EXERCISE PRESCRIPTIONS FOR CHILDREN 797

CHAPTER 50 QI EMISSION THERAPY FOR GERIATRICS 801
TREATING THE PATIENT'S EMOTIONAL STABILITY 802
ENERGIZING THE PATIENT'S QI CIRCULATION 802
TAPPING THE KNEES TO PREVENT SENILITY 803
CULTIVATING THE PATIENT'S YUAN QI BY ABSORBING ENERGY FROM NATURE 803
FURTHER OBSERVATIONS AND INSIGHTS ON GERIATRIC PATIENTS 804

CHAPTER 51 QI EMISSION THERAPY FOR GYNECOLOGY 807
INTRODUCTION 807
MENSTRUAL FLOW 807
QI AND BLOOD REGULATION 807
ETIOLOGY AND PATHOLOGY OF GYNECOLOGICAL DISEASES 808
VAGINAL DISCHARGE 809
MENSTRUAL COMPLICATIONS 810
DYSMENORRHEA 810
MENOXENIA 812
CYSTS AND TUMORS 812
BREAST CYSTS AND BENIGN TUMORS 812
LEVEL-1 BREAST CYSTS (BENIGN LUMPS, NODULES OR FIBROADENOMA) 813
LEVEL-2 BREAST CYSTS (MASTITIS OR CARBUNCLE) 813
LEVEL-3 BREAST TUMORS 813
OVERVIEW OF TREATMENT 814
OVARIAN CYSTS AND UTERINE TUMORS 815
UTERINE COMPLICATIONS 816
MISCARRIAGE 816
UTERINE COMPLICATIONS DUE TO ABORTIONS 818
UNDERSTANDING ENERGETIC ABORTIONS 819

CHAPTER 52 QI EMISSION THERAPY FOR NEUROLOGY (STROKE, COMA, PARALYSIS, AND MULTIPLE SCLEROSIS) 821
INTRODUCTION 821
STROKE (WIND-STROKE) 821
COMA 823
FACIAL PARALYSIS (BELLS PALSY) 824
MULTIPLE-SCLEROSIS (MS) 825

CHAPTER 53 QI EMISSION THERAPY FOR PSYCHOLOGY 827
INTRODUCTION TO ENERGETIC PSYCHOLOGY 827
THE EMOTIONAL ENERGY STORED WITHIN THE FIVE YIN ORGANS 828
EXCESSIVE ANGER INJURES THE LIVER (WOOD) 829
EXCESSIVE JOY INJURES THE HEART (FIRE) 830
EXCESSIVE WORRY INJURES THE SPLEEN (EARTH) 831
EXCESSIVE SADNESS AND GRIEF INJURES THE LUNGS (METAL) 832
EXCESSIVE FEAR INJURES THE KIDNEYS (WATER) 834
FIVE YIN ORGAN EMOTIONAL RESPONSES 835
EMOTIONAL DISHARMONY 836
EMOTIONAL FREQUENCY RESONATION 838
DIFFERENTIATION OF PSYCHO-EMOTIONAL DISORDERS 839
SELF-TREATMENT PRESCRIPTIONS FOR PSYCHO-EMOTIONAL DISORDERS 840
TREATING EMOTIONAL ANXIETIES, PHOBIAS, AND ADDICTIONS 840
TREATING PHOBIAS AND DEPRESSION USING "THE WINDOWS OF THE SKY" POINTS 844
TREATING PSYCHO-EMOTIONAL DISORDERS RELATED TO STRESS 848
TREATING STAGNANT LUNG QI AND THE REPRESSION OF GRIEF 850

CHAPTER 54 QI EMISSION THERAPY FOR ONCOLOGY 853
HISTORY OF CANCER TREATMENTS 853
CATEGORIZATION OF TUMORS 853
THREE FORMATIONS OF CANCER GROWTH 854

ETIOLOGY, THE SIX PATHOGENIC EVILS THAT CAUSE CANCER 855
EMOTIONAL COMPONENTS OF CANCER FORMATION 856
CANCER FORMATION AND THE PATIENT'S MIND, EMOTIONS, AND SPIRIT 857
STRESSFUL LIFE ACTIVITIES AND THE IMMUNE SYSTEM 858
USING MEDICAL QIGONG TO STRENGTHEN THE IMMUNE SYSTEM 859
CANCER TREATMENT METHODS 860
SOCIAL ONCOLOGY—ESTABLISHING A CANCER SUPPORT GROUP 861
SOCIAL ONCOLOGY WITHIN THE HOSPITALS 861
CANCER RESEARCH 862
MEDITATIONS FOR CANCER PATIENTS 862
SUMMARY 864
TREATING TUMORS AND CANCER WITH MEDICAL QIGONG THERAPY 864
BRAIN TUMORS 864
BREAST CANCER 865
UTERINE CANCER 866
CERVICAL CANCER 867
PROSTATE CANCER 869
LUNG CANCER (PULMONARY CARCINOMA) 871
LIVER CANCER 871
ESOPHAGEAL CANCER 872
STOMACH CANCER 873
PANCREATIC CANCER 873
COLON AND RECTUM CANCER 874

CHAPTER 55 QI EMISSION THERAPY FOR RADIATION AND CHEMOTHERAPY 877
MEDICAL QIGONG THERAPY AND RADIATION THERAPY 877
MEDICAL QIGONG THERAPY AND CHEMOTHERAPY 877
RELIEVING SYMPTOMS ASSOCIATED WITH RADIATION AND CHEMOTHERAPY 878

CHAPTER 56 QI EMISSION THERAPY FOR SURGERY 879
MEDICAL QIGONG AND PREOPERATIVE THERAPY 879
USING MEDICAL QIGONG FOR ANESTHESIA 880
MEDICAL QIGONG THERAPY AND SURGERY 882
MEDICAL QIGONG AND POSTOPERATIVE THERAPY 882
MEDICAL QIGONG THERAPY AND RECREATING ORGAN ENERGY 883
MEDICAL QIGONG THERAPY AND ORGAN TRANSPLANTATION 885
FOLLOW-UP AND REMEDIAL PRESCRIPTIONS 885
PHANTOM PAINS OF AMPUTATED ORGANS 887
MEDICAL QIGONG THERAPY AND THE TREATMENT OF SCAR TISSUE 887

SECTION XII

MEDICAL QIGONG APPENDICES 889

INTRODUCTION TO THE APPENDICES 890

APPENDIX 1A MEDICAL APPLICATIONS OF QIGONG 891
ABSTRACT 891
INTRODUCTION TO MEDICAL QIGONG 891
EXAMPLES OF MEDICAL APPLICATIONS OF QIGONG 892
THERAPEUTIC BALANCING OF THE CHANNELS AND FUNCTIONS OF THE BODY 892
HYPERTENSION 893
PREVENTION OF STROKE AND MORTALITY 893

IMPROVEMENTS IN HEART FUNCTION AND MICROCIRCULATION 894
IMPROVEMENT IN SEX HORMONE LEVELS 895
INCREASED BONE DENSITY 895
CANCER 896
REVERSING SENILITY 897
MIND-BODY REGULATION 898
BRAIN WAVES 898
BLOOD FLOW TO THE BRAIN 898
EFFECT OF RAPID AND LARGE CHANGES IN ALTITUDE 898
CARDIAC FUNCTION 898
MICROCIRCULATION DISORDERS 899
STUDY OF LUNG FUNCTION 899
SUMMARY AND CONCLUSIONS 899
REFERENCES 900

APPENDIX 1B ANTI-AGING BENEFITS OF QIGONG 903
INTRODUCTION 903
CLINICAL STUDIES INDICATING THE ANTI-AGING BENEFITS OF QIGONG 903
THERAPEUTIC BALANCING OF THE CHANNELS AND FUNCTIONS OF THE BODY BY QIGONG 904
BLOOD PRESSURE 904
MORTALITY AND STROKE 904
IMPROVEMENTS IN HEART FUNCTION AND MIRCROCIRCULATION 905
ENHANCED ACTIVITY OF ANTI-AGING ENZYME SOD 905
CARDIOVASCULAR FUNCTION 905
COMBINATION OF MEDICAL QIGONG AND DRUGS THERAPY 906
CONCLUSIONS 906
REFERENCES 907
ENDNOTES 907

APPENDIX 1C THE EFFECT OF QIGONG ON THERAPEUTIC BALANCING MEASURED BY ELEXTROACUPUNCTURE ACCORDING TO VOLL (EAV): A PRELIMINARY STUDY 910
ABSTRACT 910
INTRODUCTION 910
MATERIALS AND METHODS 911
RESULTS AND DISCUSSION 911
CONCLUSIONS 913
ACKNOWLEDGEMENTS 913
REFERENCES 913

APPENDIX 2 SCIENTIFIC INVESTIGATIONS INTO CHINESE MEDICAL QIGONG THERAPY 917
INTRODUCTION 917
QIGONG MASTERS AND ENERGETIC MEDICINE 918
THE SCIENCE OF CHINA 918
QIGONG'S LONG HISTORY 919
SIMULATION OF EMITTED QI 920
SCIENTIFIC PLAUSIBILITY OF QI, THE HUMAN ENERGY FIELD 920
REFERENCES 921
THE FUTURE OF QIGONG 921
INFRASONIC QIGONG RESEARCH 922
INFRASONIC SIMULATION OF EMITTED QI FROM QIGONG MASTERS 922
SUMMAY OF RESEARCH 922
DOUBTS ABOUT EMITTED QI 922

BACKGROUND 923
THE BRAIN AS A DETECTOR OF EMITTED QI 923
SCIENTIFIC CONTROLS 924
INFRASONIC VS. ELECTROMAGNETIC INTERFERENCE 925
SUMMARY 925
SECONDARY ACOUSTIC BIOLOGICAL RESPONSE 925
MEASUREMENT AND ANALYSIS OF INFRASONIC WAVES FROM EMITTED QI 927
NEURO SCIENCE—NERVE IMPINGEMENT AND EMITTED QI 928
INFRASONIC TREATMENT OF ASTHMA 929
TREATMENT METHOD 930
OBSERVATION CRITERIA—FOUR GRADES 930
THERAPEUTIC RESULTS 930
LABORATORY RESULTS 930
THE INFRASONIC SIGNAL FROM THE QIGONG PERSPECTIVE 931
MECHANISMS OF THE QGM 931
RELIEVES PAIN AND EDEMA 931
STRENGTHENS IMMUNE FUNCTION AND REDUCES INFLAMMATION 931
NORMALIZES AUTONOMIC NERVOUS SYSTEM 931
RELIEVES SOME SYMPTOMS OF HEART DISEASE 932
EMITTED QI TRAINING INCREASES INFRASONIC SOUND EMISSION OF HEALERS 932
SCIENTIFIC QIGONG RESEARCH 933
EFFECTS OF THE EMITTED QI ON HEALING OF EXPERIMENTAL FRACTURE 934
MEASURING QI IN THE BLOOD 935
EFFECTS OF QIGONG ON NEUROTRANSMITTERS 936
PHYSICAL CHARACTERISTICS OF EMITTED QI 937
RESULTS 938
SUPERCONDUCTING MAGNETOMETER MEASURES CHANGES DURING QIGONG 938
QIGONG AND CANCER 938
THE EFFECT OF THE EMITTED QI ON CARCINOMA CELLS 939
THE EFFECT OF EMITTED QI ON THE GASTRIC ADENOCARCINOMA CELLS: 939
SUMMARY: 940
QIGONG AS A THERAPEUTIC AID FOR ADVANCED CANCER PATIENTS 940
THE EFFECT OF EMITTED QI ON LEUKEMIA IN MICE 941
EFFECTS OF QIGONG ON MALIGNANT TUMOR 942
EFFECTS OF PATIENT TUMORS ON KIRLIAN IMAGES OF HEALTHCARE PROVIDERS 943
CHAOS AND CONSCIOUSNESS 945
LOW DIMENSIONAL CHAOS IN PRACTICING QIGONG 945
QIGONG TRAINING AND COHERENCE OF EEG 946
VIBRATIONAL TREMBLING, CHAOS, AND EEG 946
LABORATORY TESTING 947
CHAOS AND THE COLLEGE ENTRANCE EXAM 947
SEARCHING FOR JING, QI, AND SHEN IN WESTERN SCIENCE 948
CONCLUSION 949
THE ROLE OF CHAOS IN DISEASE AND HEALING 950
THERAPEUTIC CHAOS 951
IMPLICATIONS OF SCIENTIFIC QIGONG RESEARCH 952

APPENDIX 3 THE PSYCHOLOGICAL AND ENERGETIC MECHANISMS TRIGGERED IN THE HUMAN SYSTEM BY THE SELF-APPLIED HEALTH ENHANCEMENT METHODS QIGONG, TAIJI, AND YOGA 957
OXYGEN 957
LYMPH SYSTEM FUNCTION 960
BRAIN AND NERVOUS SYSTEM FUNCTION 964
ENERGETIC MECHANISMS 967
CONCLUSION 968
REFERENCES 970

GLOSSARY OF TERMS 973

BIBLIOGRAPHY 1003

CLINICAL REFERENCES 1009

INDEX 1011

ABOUT THE AUTHOR 1029
MEDICAL BACKGROUND 1029
MARTIAL ARTS BACKGROUND 1031

ADDITIONAL BRANCHES OF THE INTERNATIONAL INSTITUTE OF MEDICAL QIGONG 1033
BRANCHES WITHIN THE UNITED STATES 1033
INTERNATIONAL BRANCHES 1035

FORWARD

Dr. Jerry Alan Johnson's book on Medical Qigong and its clinical practices provides an important and necessary guide for students and advanced practitioners of Medical Qigong and Traditional Chinese Medicine.

The relevance of Dr. Johnson's Medical Qigong textbook can be seen against the larger back drop of alternative medicine. Alternative Medicine (also called complementary medicine) has become a catch word in the West and an increasingly important factor in health care. An article authored by David Eisenberg and others of the Harvard Medical School reported that in 1993, 30% of a group of people used some kind of Alternative Medicine for which they paid out of their own pockets. Extrapolated to the United States population, this group spent about $14 billion dollars, which was more than the amount spent on conventional Western care that year.

Among alternative medical therapies, Medical Qigong has blossomed explosively in Western countries during the past few years. Awareness of Medical Qigong in the United States was heightened by two international conferences on Qigong in 1990. In 1997, there were three conferences on Medical Qigong, in San Francisco, Minneapolis, and New York. Another catalyst was Bill Moyer's television documentary, "Healing and the Mind," which had a segment on Medical Qigong and Chinese medicine. In more recent years, dozens of books have been published on Qigong in English and presently there are at least 15 web sites on the World Wide Web devoted to Qigong. At this writing, many articles on Qigong have appeared in the newspapers and magazines, and several documentaries have been videotaped and others are in progress.

In China, almost everyone knows about Qigong and about 100 million people are said to practice it daily. Qigong dates back thousands of years. Many believe that it was responsible for the early development of Chinese medicine, which now includes therapy by acupuncture, herbs, massage, and nutrition.

Medical Qigong therapy was a secret art that was passed on from a master to selected persons. In the 1970s, Qigong became more accessible to the people when the Chinese government ordered that Medical Qigong be made public and serve alongside Western medicine to improve national health care. During the Cultural Revolution, Medical Qigong was suppressed because the government suspected that Qigong had religious and superstitious roots incompatible with communism. In the early 1980s, scientist in China were encouraged to carry out experimental and clinical studies into almost every aspect of how Medical Qigong affected the human body and contributed to healing. Starting in 1986, research findings were reported at international conferences on Medical Qigong held mostly in China but also in Japan, Canada, and the United States. The substance of the research studies are available in English in the Computerized Qigong Database, which was developed by the Qigong Institute. This database has provided a basis for the preparation of at least five books, three dissertations, and several research projects on Medical Qigong.

In 1992, the United States Congress recognized the potential of alternative/complementary medicine by mandating that the National Institutes of Health develop programs to evaluate the validity of selected alternative therapies. Since that time, the Office of Alternative Medicine has funded many research programs and disseminated valuable information to medical practitioners, scientists, and the public.

The swell in the interest in Medical Qigong and the recognition of its health and healing benefits have been recognized by schools of Traditional Chinese Medicine, many of which mandate that Qigong be a required course. The Qigong taught at these schools is still on a basic level and needs to be expanded to include the medical aspects of Qigong, such as described in this book.

The future of Medical Qigong and Chinese medicine appears bright. However, the need for well-trained and qualified teachers and practitioners is paramount to meeting the growing demand. Dr. Johnson's book provides the first guide in English for teachers, students, and practitioners.

Kenneth M. Sancier, Ph.D.
President Qigong Institute

FORWARD

In 1998, the Academy of Oriental Medicine at Austin gave me the opportunity to establish one of the first formal Medical Qigong training programs in the country to be offered within the context of a college of Oriental medicine. My greatest difficulty was in finding a reliable textbook for my classes. Of the numerous books on Qigong available in English, most devote only a few pages, at best, to Medical Qigong. Over the past few years, several books from China on Medical Qigong have been translated into English. Although presenting much new and valuable information, they are poorly translated and fraught with errors, making them unreliable as textbooks.

Therefore, it has been a great pleasure for me to have collaborated with Dr. Jerry Alan Johnson to some small extent on this book. Jerry Alan Johnson is one of the world's foremost authorities on the subject of Medical Qigong. This work distills several decades of his personal study, practice, and clinical experience in the United States and the People's Republic of China into an in-depth practical manual for all levels of Qigong healers.

Currently China is in the midst of what can only be called a Qigong renaissance. Today, more people are studying and practicing Qigong than in any previous time in Chinese history. In the not too distant past, Qigong was only available to monks and priests, nobles and the wealthy. Now Qigong is taught in China's primary schools and colleges, parks, temples, clinics, and hospitals to anyone who wants to learn.

China's literacy rate has increased exponentially since 1949, and now there are thousands of books, tapes, videos, and magazines on every aspect of Qigong, systems both ancient and modern, all available for general public consumption. Many regional, national, and international Medical Qigong associations have been formed, with an unprecedented exchange of information.

Although Medical Qigong has been a part of Traditional Chinese Medicine since the time of the Yellow Emperor, it has only been since 1982 that Medical Qigong hospitals and clinics have begun to spring up all over China. In the early days after the Communist Revolution of 1949, Qigong was more or less a state secret, and the Medical Qigong clinics were, for the most part, only available to high party officials. Over the past two decades, China has also devoted a great deal of effort to Qigong research and education. There are now large Medical Qigong hospitals in all of China's major cities, some of which have over 300 beds and 40 or more Qigong specialists and nurses on staff. There are hundreds of medium-and small-sized Medical Qigong clinics throughout the country, some treating inpatients and others treating only outpatients.

Before the revolution, the miracles of modern Western medicine, with its wonder drugs and surgical procedures, had caused many Chinese to see their own native traditional medicine as old-fashioned and outdated. The government encouraged the development of Western medical schools and hospitals. After the revolution, Chairman Mao recognized that with China's huge population they would need to take advantage of every form of medicine available even to begin to meet the nation's fundamental health care needs.

China is the only county in the world where Medical Qigong has been truly integrated and enthusiastically accepted as part of the national health care delivery system. According to current estimates, Medical Qigong doctors are now meeting 15% of China's national medical needs. That percentage is still growing. For decades, Chinese hospitals have successfully offered Traditional

Chinese Medicine alongside of Western medical services. Patients may receive Qigong or acupuncture anesthesia during Western surgical procedures and be treated by Medical Qigong, acupuncture, and Chinese herbal medicines during recovery. Chronic patients may be diagnosed by Western physicians, sent to Medical Qigong clinics for treatment, and have their Western doctors monitor their progress. In this respect, China has truly set the standard for the twenty-first century medicine.

Medical Qigong training and practice takes many forms in China. Some practitioners/healers are thoroughly trained in both Traditional Chinese Medicine and Western medicine, in addition to Medical Qigong. Others are heirs of family Medical Qigong systems, some dating back over a thousand years. Not all of these family practitioners have received additional training in either traditional Chinese medicine or Western medicine, yet they have been very efficient in treating disease and disorders that these conventional medicines have given up on. Still other practitioners come from martial arts backgrounds, such as Shaolin Gongfu or Daoist martial Qigong, with their own effective and unique blend of warrior training, spiritual discipline, Qigong and healing arts. Each of these categories contributes to the richness and diversity of the art of Medical Qigong.

Jerry Alan Johnson presents for us the art of Medical Qigong as it is being taught in two of the major Medical Qigong training centers in China today: the Xi Yuan Hospital of Traditional Chinese Medicine and its associated Hai Dian Medical Qigong College. Although each Qigong hospital in China defines its own program of training and practice, the material Johnson includes in this work represents the mainstream curriculum of study and practice widely accepted nationwide as a high standard of competence in Medical Qigong.

In addition to sharing the high-level protocols of Medical Qigong training in China, Johnson also draws from the reservoirs of his own decades of personal training and teaching in the internal martial arts to further enrich this work. Martial Qigong is generally practised to improve one's martial arts prowess. It includes Qigong exercises for strengthening the body; increasing power of one's strikes by emitting Qi; mind and will training; learning the vulnerable points to strike on an opponent's body to cause pain and paralysis; learning to cause serious injury or death by disrupting or destroying an opponent's internal organs, Qi and blood vessels, joints or bones; and toughening the body to withstand strikes and blows.

In what might at first seem a case of strange bedfellows, the martial Qigong tradition also has its medical side. Martial artists need to learn how to set bones and dislocations, as these injuries are commonly encountered in training and combat. Martial artists who train in the more secret arts of tien xue (vital point striking) and dim mak (death touch) are also taught how to heal the injuries that they have caused, both to heal themselves if struck by someone and to heal their opponents if the opponent is penitent. Furthermore, the martial artist who learns the methods for destroying the internal organs by emitting Qi can effectively adapt the same method to destroying pathogens and tumors in a patient for healing. Internationally recognized by his peers for his mastery of the most esoteric aspects of the internal martial arts, Johnson is eminently qualified to present its healing traditions, which he generously and openly does in this work.

I am grateful to know Dr. Jerry Alan Johnson as a colleague, teacher, and friend, and am happy to recommend this work to all students of Medical Qigong with confidence. Johnson's contribution contained herein represents the most thorough and comprehensive presentation of the art and science of Medical Qigong in the world to date. It is sure to become a modern standard of reference for anyone seeking knowledge of Medical Qigong.

Jampa Mackenzie Stewart, D.O.M.
Dean of Medical Qigong
Academy of Oriental Medicine at Austin

FORWARD

In the summer of 1995, I was sitting on the floor of Dr. Johnson's studio waiting for the start of the intermediate class in Medical Qigong. David Nakahara leaned over to introduce himself and asked me if I was a healer. Wow! It was one of those seminal moments in one's life. I considered all the implications of this question and the only relevant reply I, as a medical doctor, could muster was an ironic chuckle.

That question speaks to the heart of the current crisis in health care. Allopathic medicine lends itself to standardization of care and computer algorithms, since the structure of the clinical visit is based on protocol-derived diagnosis and treatment. The challenge is the diagnosis, but the treatment, especially high-tech treatment, gets all the press. It is incredibly satisfying to take care of people in this manner since the issues are clear and everyone gets better, unless they fail treatment; in this paradigm, the patient fails the treatment, the treatment does not fail the patient.

However, over time, you accumulate a number of patients for whom your therapeutic protocol fails, and I think that all thoughtful physicians recognize that it is these patients that teach us the most. There are women who have serious difficulty with PMS, infertility, and menopause. The cancer patients often wonder what caused the cancer to occur, what supports it, and what the meaning of this potentially fatal illness is for them. As physicians, once we emerge from the narcosis of fatigue and busyness that accompanies the daily practice of medicine in the late twentieth century, we also ask these same questions.

And we seek answers. Each physician develops his or her own rationale for treatment failures. They may rely on an unshakable faith in the scientific method, their religious upbringing or in their cultural ideals, or perhaps close themselves off entirely. I inadvertently found my answer within the martial arts.

It was clear to me from my Aikido practice that the superior martial artist must understand Qi in all its manifestations. This quest for deeper knowledge and experience of Qi led me to study Qigong with Dr. Johnson.

In my earliest lessons with Dr. Johnson, I was told that a Qigong therapist must observe and treat patients' *three treasures*: the physical body (jing or essence), their energy (Qi), and their mind/spirit/intention (shen). These three form the whole of a human being, and the parts can only be understood within the context of the whole. Furthermore, I learned that in life these three are so interrelated that it is difficult to tell where one leaves off and the other begins. Over 2,000 years ago, the Chinese philosopher Lie Zi described this by saying, "The mind leads and the Qi follows; the Qi leads and the body follows."

As a scientist, I could not help but draw the parallel between these fundamental Qigong principles and the revolutionary views of advanced modern physics: that at the deepest sub-atomic levels, matter, energy, and the person viewing them are so inextricably intertwined that they are seen as virtually one and the same, merely different manifestations of the same pattern. No longer is energy defined as a wave while matter is composed of particles. In quantum physics they have coined the term *wavicle* beause particles act as waves in one situation and as particles in another, as "probabilities" rather than concrete fixed entities and as subjective reality rather than objective.

The application of the discoveries of quantum physics to biology is just beginning. It is starting not only in the laboratory but also in many consulting rooms across the United States, as physicians search for the answers to help their patients.

I think that quantum physics is the only "hard" science that you can research just by thinking about it. As one observes the processes of nature, the thoughtful observer comes to the conclusion that there must be something else that accounts for the medical journey of the exceptional patient. The early practitioners of Ayurvedic and Chinese Medicine were careful observers of nature and through observation formulated the basic principles of Indian and Asian medical systems that are also corollaries of the principles of quantum physics.

Western awareness of the potential for the integration of the East and West began in 1971 when Nixon first went to China. James Reston, a *New York Times* correspondent covering the historic visit, developed acute appendicitis. After his emergency appendectomy, he was successfully treated with acupuncture for post-surgical pain.

Despite the dynamic nature of the penultimate Western nonbeliever confronting a totally novel experience ("only Nixon could go to China"), nothing much shifted or changed in medical practice at that time. What was missing was the readiness and ability to experience a paradigm shift in the conceptual basis of the causation of disease. Paradigm shifts in basic research have been eloquently discussed in Thomas Kuhn's landmark book, *The Structure of Scientific Revolutions*. Kuhn's brilliant analysis of the development of knowledge has recently been applied to medical science in a new book, *From Doctor to Healer: The Transformative Journey*, by Robbie Davis-Floyd and Gloria St. John. After 30 years, we are finally mature enough to begin to reassess the structure of our medical belief systems and to incorporate both world medicine and the principles of quantum physics. I have no doubt that this will take decades, but there is no doubt that the shift is beginning to occur.

Dr. Jerry Alan Johnson's book represents a major advance in this journey. This textbook is the first comprehensive treatise on the principles of Chinese energetic medical practice available to the Western medical practitioner. It also forms a bridge for the Western trained practitioner of Traditional Chinese Medicine and the allopathic (Western) practitioner. Dr. Johnson's book represents decades of research and study, and a nonjudgmental openness to any healing modality that is in the patient's best interest. I have no doubt that this book will represent a significant advance in that long awaited paradigm shift. I look forward to that time because it will address the deeper human values that have been neglected in protocol-based medicine and will bring a new depth of engagement and satisfaction to patient care.

Stephanie S. Taylor M.D., Ph.D.

Footnotes:

Kuhn, Thomas S., *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, Illinois 1962.

Davis-Floyd and Gloria St. John, *From Doctor to Healer: The Transformative Journey*. Rutgers University Press, New Brunswick, N.J., available March 1998.

FORWARD

In the Western nations at the turn of the millennium, following a century of remarkable medical breakthroughs, there still remains a major list of unsolved medical mysteries and health care challenges. Chronic disease has not been alleviated (asthma, arthritis, etc.); stress has been recognized as a major cause in many diseases; tumors and cancers are being diagnosed in exponentially growing numbers; AIDS is a world wide epidemic; negative (Western) drug interactions have grown to become the fourth leading cause of death, and medical error has emerged as the eighth cause of death. According to the Department of Health and Human Services, over 70% of medical visits are for symptoms and illnesses that are preventable, yet the health care system is still primarily focused on treatment of the symptoms rather than health maintenance or prevention.

The public, the government, and health care providers all agree that there is a crisis in Western medicine and health care. Fortunately, at the same time, there is increasing support for innovative and alternative solutions. Among the many powerful strategies for solving this crisis, Chinese Medical Qigong (Chi Kung) and its offspring Taijiquan (Tai Chi Chuan), meditation, and focused breath practice, are probably the most comprehensive systems of self-care and healing available to compliment both conventional Western and alternative medical therapies.

The public hunger for natural healing, safety from medical side effects, lower medical costs, and health self-reliance is triggering powerful and profound changes in health care. It has become obvious that in the near future the benefits of Chinese medicine will be as fully integrated here in the U.S., as the benefits of Western medicine are in China. Through this global exchange of health care wisdom and practical applications, we are entering an unprecedented era of new possibilities to resolve suffering and foster well-being.

This definitive work on Medical Qigong, the remarkable healing art from Chinese traditional medicine, will play a significant role in the immense and profound transformation in health care that is now occurring with tremendous social and economic force.

Like the revolution in training and licensing acupuncturists that quickly unfolded in the final years of the 20th century, the momentum of training and licensing Medical Qigong Therapists is rapidly increasing in the beginning of the 21st century. Clinical Qigong, including both Medical Qigong Prescriptions and Medical Qigong Therapy, will arise in America as it has done so robustly in China.

There are a number of excellent books on various aspects and methods of Qigong. However, there has not been, in English, a comprehensive exploration of Medical Qigong. Dr. Johnson has created a breakthrough work on Medical Qigong, which is a clear and useful revelation of the Medical Qigong curriculum at the Hai Dian University Medical Qigong College of Beijing, China, and an excellent synthesis of Medical Qigong theory from throughout China. This textbook will very likely remain the definitive compendium of Medical Qigong in the West for many years, and become the foundation from which the field of Medical Qigong will evolve in Western society.

Roger Jahnke, O.M.D.
Chair, Department of Medical Qigong
Santa Barbara College of Oriental Medicine

PREFACE

As we enter the new millennium, a new era of medical therapy is beginning to blossom. Alternative medicine is now capturing an estimated $14 billion in out-of-pocket health care revenues from Americans each year, according to a 1993 study published in the *New England Journal of Medicine.* By 1997, according to the *Journal of American Medical Association,* that figure had more then doubled. What this trend means in real terms, is that despite the historical lack of official recognition by the American Medical Association, despite the lack of endorsement and coverage by Medicare and the majority of health insurance plans, men and women, in ever increasing numbers, are going to acupuncturists, energetic healers, herbalists, chiropractors, massage therapists, ayurvedic specialists, homeopathic doctors, and other traditional "healers" to meet some portion of their health care needs.

This growth of public reliance on alternative medicine has caused the Western medical establishment to sit up and take notice. In fact, some of the most vocal proponents of combining alternative medical traditions with Western medicine are medical doctors. Visionary physicians such as Deepak Chopra, Andrew Weil, Larry Dossey, Dean Ornish, and Bernie Seigel have led the way toward creating a new climate of respect for ancient medical philosophies and modalities. They have pioneered the advent in the West of health care facilities where Western medical and alternative health modalities are available under the same roof, with the goal of providing patients with the best of both worlds. This combination of ancient and modern medical traditions has been dubbed *integrative* or *complementary* medicine.

In comprehending the full implications of this unlikely marriage, one must understand the divergent approaches to healing between conventional Western medicine and traditional Eastern medicine.

Contemporary Western medicine grew out of the scientific revolution of the seventeenth century. The philosophy of science, rooted in Aristotle's "empirical materialism," was given a new spin by the French mathematician, Descartes. Viewing reality as that which could be substantiated materially, Descartes applied an analytical reductionist logic to penetrating the secrets of nature, including biology. These views were echoed in the physics of Sir Isaac Newton, applying a linear cause-and-effect model to explain the workings of a material universe.

Man was seen as being separate from nature, mind was seen as separate from body, and all of these processes, in nature and in humans, were seen as similar to the workings of a machine composed of discreet parts. Mechanical laws were seen to govern all processes. Structure determines function; therefore, the physician's role developed into that of a mechanic: repairing, removing, transplanting, and replacing broken down-parts. Diseases had isolated causes, which need to be removed from the rest of the parts. Because of this approach, Western medicine has the most highly developed pharmaceuticals to kill specific organisms and the finest surgical procedures in the world today.

By contrast, Eastern medicine grew out of the empirical observation of nature, beginning at least 4,700 years ago. Oriental philosophy, from the Vedas of India to the Yellow Emperor of China, views reality as an interdependent whole. This "prescientific" understanding equates to the broader view of modern quantum physics and the general systems theory. Rather than limiting reality to that which is material, the Eastern philosophers recognized the interdependence of mind and body, the nonlinear nature of time and space, and the interweaving patterns of relationship between man and nature; in fact, they believed, we *are* nature. Anatomy, the study of human structure, takes a back seat to physiology, the study of human function. Thus, the Orientals formulated a general systems theory, in which the patterns of change that exist in nature are the same patterns that govern human biology, wherein function is viewed from a holographic perspective, and each part reflects the whole. Rather than being fixed and stable, the whole is in a dynamic process of constant change. For the whole to function harmoniously, every part must remain in balance. Therefore, the role of the Oriental physician is more similar to a gardener, following the patterns of change, diagnosing functional disharmony and restoring overall balance. Because of this approach, Oriental medicine has some of the most highly developed procedures for preventative medicine and for treating chronic diseases in the world today.

Western medical science, with its fundamental distrust of subjective diagnostic reliability, has progressed toward developing more and more expensive high-tech laboratory tests and diagnostic equipment. Thus we have the modern miracles of x-rays, MRI's, and ultrasound. Eastern medicine, trusting in human capacity, has progressed in a low-tech direction toward ever deeper training of the physician's sensory and spiritual diagnostic tools. Thus, we have the miracles of pulse diagnosis, tongue reading, and Qigong hand scanning.

With such fundamental divergencies in philosophy and technique, it is almost inconceivable that these two medical systems could ever operate together in the same setting. The fact is that they do function together, and quite effectively, too. We can thank Mao Zedong for the union of these two unlikely bedfellows.

Recognizing that there were far too few Western trained physicians and nurses to meet the primary health care needs of China's vast population, from the outset of his leadership Mao advocated the systemization of Traditional Chinese Medicine (TCM), and advocated its implementation alongside Western medicine in China's hospitals and clinics. The results of this integration have been astonishing, as witnessed in the effective use of acupuncture anesthesia during surgery.

Nonetheless, in his efforts to create a "modern" Chinese medicine, Mao shunned some of the traditional theoretical aspects of Chinese medicine, such as the concept of Qi, which he considered as feudalistic and counterrevolutionary. For this reason, he actively discouraged Medical Qigong practice as superstitious. It was not until the end of the Cultural Revolution in 1975 that Qigong reclaimed its rightful place as one of the major branches of Chinese medicine.

A high party official in Beijing was suffering from an "incurable" disease. Both Western medicine and TCM had failed to alleviate his suffering. In desperation, he went to one of the few Medical Qigong clinics operating in the country and was cured. The official then lent his support to the promotion of Medical Qigong for the benefit of the Chinese people. Before long, there were hundreds of Medical Qigong hospitals and clinics throughout China.

Since that time, Qigong has experienced an unprecedented growth in China. Qigong is now available to the general populace for the first time in history. Qigong is now taught in the public education system, beginning at the elementary school level. Qigong departments have been added to large urban hospitals. Colleges of Traditional Chinese Medicine have developed Medical Qigong training programs. According to one recent survey, one third of the population of Beijing, China's capital city, practices Qigong daily.

It has taken acupuncture and Chinese herbology nearly 20 years to develop into a respected profession in the United States. As this book goes to press, 37 states now have legislation licensing professional acupuncture practice, with an additional 10 states in which legislation has currently been introduced. Medical Qigong is still at the very early stages of public recognition, understanding, and acceptance by comparison. Oriental medical schools around the United States, as well as independent Medical Qigong masters, are only now beginning to establish comprehensive

Medical Qigong training programs. I am heartened by this development.

Traditional Chinese medical therapy is divided into four main branches: acupuncture, herbs and diet, massage therapy, and Medical Qigong. It is important for TCM practitioners to have exposure to all four branches to be able to understand the relative strengths and limitations of their particular field of expertise, so that they will be able to select the most effective and appropriate treatment modality for their patients. Until very recently, most schools in America have been relatively unaware of the extent to which Medical Qigong therapy has developed in China, and have been lacking in presenting Medical Qigong as a significant part of their training programs for students of Oriental medicine.

The purpose of this book is twofold. The first purpose is to apprise the reader of the professional standards of knowledge and skill required in contemporary China for licensing a "Doctor of Medical Qigong Therapy" (D.M.Q.). The second purpose is to describe how to effectively diagnose and treat patients with Medical Qigong therapy as set forth by the Hai Dian Medical Qigong College of Beijing and to provide a textbook for training to that standard under the guidance of a qualified instructor. Hai Dian is recognized as one of the top Medical Qigong colleges in China.

Although this book is primarily written for students and practitioners of Oriental medicine, it is also my hope that practitioners and students of Western medicine and other forms of alternative healing will benefit from the information contained herein, and that it may serve to enrich their practice in the healing arts.

This book is divided into twelve sections. The first four sections focus on building a comprehensive foundation in energetic medicine. The second four sections focus on guidelines for establishing a Medical Qigong clinic, as well as for diagnosing and treating problems of Qi deviations. The last four sections focus on treatment modalities, Medical Qigong prescriptions, Medical Qigong regulation exercises and scientific research in Medical Qigong clinical therapy.

I have done my best to present the knowledge and wisdom of this ancient Chinese art as it was passed onto me personally by my respected teachers, to whom I owe undying gratitude. Qigong has survived nearly five thousand years of growth and refinement and is now available to you, the reader, for your own personal and professional benefit. I sincerely hope that this book may serve to further bridge the partnership between Eastern and Western medicine, that all humanity may be the beneficiary of the interchange between these two great schools of healing. May all doctors and healers, Western and alternative, return to the "heart" of medicine, that each may view the patient as a complete integration of body, mind, emotion, energy, and spirit. May we support each other's skills and techniques in alleviating our patients' suffering.

Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
Dean of Medical Qigong Science
Director of Medical Qigong Clinic
Five Branches Institute, College & Clinic of TCM
Santa Cruz, California
May, 2000

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my many friends and colleagues who helped encourage and support me in writing this book: Roger Jahnke, O.M.D.; Hu Jiting, D.T.C.M.; Zhang Jingchuan, D.T.C.M.; Xu Zhuoqun, D.T.C.M.; Guo Xianhe, D.T.C.M.; Wan Taowang, D.M.Q.; William H. Lewington, L.Ac., M.Q.T.; Adam Atman, L.Ac. M.Q.T.; Tim M. Haun, D.C.; Kerrean D. Smith; Geoffrey Greenspahn M.Q.T.; Michael Winn; Mark Johnson; and my good friend Joseph Crandall.

I am also indebted to Dr. Pang Donghui, the Executive Deputy President of the Hai Dian Qigong College of Beijing, China, and to Li Fudong, D.M.Q. the university's chief director of Medical Qigong Science, and his assistants, Lu Guohong, D.M.Q. and Niu Yuhua, D.M.Q. for their constant support and encouragement, and for sharing their innermost secrets on Medical Qigong therapy.

I am very grateful to Dr. Teng Yingbo, D.M.Q., president and secretary general of the Beijing Western District Qigong Science and Research Institute, for his constant support and openness in sharing the knowledge of the Institute's clinical modalities.

A special thanks to Doctor Bi Yongsheng, D.M.Q.; Yu Wenping, D.M.Q.; and the Shandong Provincial Qigong Association for their wealth of knowledge and enthusiastic support, in sharing their clinical Qigong modalities, and to Dr. He Sihai, D.M.Q., of the Zhe Jiang Qigong Hospital for his openness and kindness in sharing his knowledge.

A special thanks to Qigong Master Zheng Zhanding, D.M.Q., for being my mentor, taking me under his wing, and teaching me advanced clinical modalities of energetic diagnosis and treatments, and to Qigong and Taiji Master Zhang Yufei for teaching, supporting, and encouraging my energetic and healing practice.

I am also indebted to Professor Lu and doctors Xu Hongtao, D.M.Q.; Ren Shuntu, D.M.Q.; Xu Zongwei, D.M.Q.; and the directors, teachers, interpreters and staff at the Xi Yuan Hospital of Traditional Chinese Medicine for their help and support.

I am very grateful to Sun Shuchun, D.T.C.M., the assistant professor and dean of the Beijing Academy of Acupuncture, Orthopedics, and Traumatology and the Hu Guo Si Hospital of Traditional Chinese Medicine for his assistance and support.

I would like to thank Doctor Meng Xiantong and Tara Peng, D.M.Q., of the Beijing Chengjian Integrated Traditional Chinese Medicine and Western Medical Experts Clinic for sharing their clinical modalities and their support.

I am very grateful to Madame Wang Yan of the China, Beijing International Acupuncture Training Center, the Acupuncture Institute of China Academy of Traditional Chinese Medicine, and the World Health Organization Collaborating Center for Traditional Medicine for supporting and believing in me.

I would also like to thank President Yue Licui, D.T.C.M., of the World Academic Society of Medical Qigong and all my friends and colleagues at the Beijing College of Traditional Chinese Medicine for their help and support.

A special thanks to Stephanie Taylor, M.D., Ph.D., whose encouragement and support has enabled Chinese Medical Qigong therapy to expand far into the terrain of conventional Western medicine, bridging the gap between these two schools with complementary modalities of healing.

A special thanks to Jason Gill, L.Ac., for his unselfish heart of giving and his invaluable help and technical input on Traditional Chinese Medicine, and his own personal experiences with Medical Qigong therapy.

A special thanks to Kenneth M. Sancier, Ph.D., and the Qigong Institute of Menlo Park, CA. Dr. Sancier's extensive research in Medical Qigong therapy, and the validity of its effect as a clinical modality is disclosed in Appendix 1 at the back of this text. Dr. Sancier is a council board member of the World Academic Society of Medical Qigong, sanctioned in Beijing, China.

A special thanks to Richard H. Lee. and the China Healthways Institute of San Clemente, CA. Mr. Lee's extensive research in the scientific investigations into Medical Qigong therapy and the validity of its effect as a clinical modality is disclosed in Appendix 2 at the back of this text. Mr. Lee is also a council board member of the World Academic Society of Medical Qigong, sanctioned in Beijing, China.

I greatly appreciate the assistance of Robert Collier, Jed Friedland, Stephanie Hensey, Terri Vandercook and Rose Mary Stewart, who spent many painstaking hours typing and deciphering my personal notes and course manuals. I would especially like to thank Jampa Mackenzie Stewart, D.O.M., L.Ac., Gaeir L. Dietrich, and Jean Vlamynck, M.Q.T., L.Ac., for their invaluable help and contribution in editing this textbook.

A special thanks to Madeleine H. Howell, M.F.T., as the senior editor of this manuscript for publication. Her keen eye for detail and constant encouragement to rewrite, explain and expound on the contents within this textbook, has resulted in a more comprehensive resource for future Medical Qigong students and Western medical professionals.

A special thanks to Robert Blaisdell and Tomo Y. for their excellent artwork and graphics, and to Y. S. Lim for his many painstaking hours of Translating Chinese Medical Qigong texts, which enabled me to draw from more comprehensive and clinical resources.

I am grateful to Willis Campbell and Niki Norrell, L.Ac., Dipl.Ac. (NCCA), M.T.C.M. for taking several of the photos.

A special thanks to Irene Morris for her excellent work on designing the front and back cover of this textbook.

Above all, I would like to express my love, gratitude, and respect to my parents Antoinette and Lt. Cdr. (Rt.) Perry E. J. Johnson for their years of love, understanding, encouragement, and support.

Finally, to my three beautiful daughters Laura, Leah, and Hannah, whom I will love forever.

INTRODUCTION

The Eastern approach in explaining medical concepts is generally expressed through a three dimensional convergence, seeing the physical body as an energetic hologram, and observing the physical, mental, emotional, energetic and spiritual reality of the tissues. By stimulating any one of these five energetic matrices, you affect the other four, and influence the body to initiate either health or disease. Diagnosis and treatment is therefore approached in a nonlinear progression, working towards the health and healing of the whole person.

The Western mind, however, is taught to view the physical body on a linear progression of cause and affect, separating the reactions of the tissues from the effect of the mind and emotions, as well as the universal and environmental influences. In order to help the Western mind understand the concept of Chinese energetic medicine, this Medical Qigong textbook was written with the goal of instructional comprehension and practical application of Chinese energetic medicine. It embraces the concepts of traditional Chinese medicine, in particular Medical Qigong therapy, from a progressive linear format. Beginning with Energetic Embryology in Section 1, it provides a basic understanding of complex energetic structure, theory, and practical application, ending in advanced clinical treatments in Section 11. The final chapters of this textbook, consists of scientific research collected from various doctors and research scientists which validate the claims made of Medical Qigong therapy being an effective clinical modality, in Section 12.

Chinese medical terminology is extremely metaphoric and is used to describe the many aspects of the human body through physical, mental, emotional, energetic, and spiritual domains of existence. Many technical terms in Traditional Chinese Medicine have numerous meanings, depending on the context of the subject.

For clarity, the Chinese terms in this book are capitalized along with the English words for which Traditional Chinese Medicine assigns a special meaning. Such words include, but are not limited to, the following: Gall Bladder, Small Intestine, Spleen, Pericardium, Bladder, Liver, etc. There is one exception to this rule; the word "energy" meaning Qi, is capped only in the first few chapters of this book for the sake of easier legibility, as it appears so often within each chapter.

When you see words such as Blood, Heart, and Marrow capitalized, assume that their meaning differs from that ascribed by Western medicine. In Traditional Chinese Medicine, the word Marrow, for instance, does not refer to bone marrow as it is traditionally recognized in the West, but rather describes the substance which is the common matrix of bones, bone marrow, the brain and spinal cord. Non-capitalized terms retain their traditional Western meanings.

One decision I have made when writing this textbook is to use the term "divine" when expressing the energy of the Dao or God. In China, many of the Qigong doctors and masters that I have trained with asked me to keep these spiritual theories in confidence for fear of governmental reprisal. Therefore, having decided to openly share these ancient energetic theories concerning the interactions of the spirit, soul and divine, I have done so without revealing my sources.

Finally, many of the energies used to describe the natural forces of nature (existing within Heaven, Earth, and Man) were historically changed into "gods" or "spirits" in order to compete with the colorful deities imported into China from India. Therefore, when reading this textbook, the reader should not become confused by certain terminologies describing "deities" which govern specific energetic principles. For example, when reading the section on the Eternal Soul, the title "Lords of the Three Dantians" implies the fact that the soul has three primary, separate, yet interconnected, energetic properties, rooted, divided and sustained within each of the Three Dantians.

Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)

SECTION I

FOUNDATIONS OF CHINESE ENERGETIC MEDICINE

CHAPTER 1
INTRODUCTION TO MEDICAL QIGONG

More than 5000 years ago, Chinese physicians came to understand that everything is composed of the same energetic substance called Qi (pronounced "chee"). These ancient masters concluded that there is a oneness and wholeness in all existence, and that energetically everything is interconnected as one body, although energy may appear to take on many different forms. All things in nature and, in fact, all things in the universe are woven together so that we are, quite literally, all symbiotically one with the universe through the system of Qi. Qi is always in motion within all things, and is the catalyst for everything to relate and interrelate within the universe.

In modern times, the laws of physics have demonstrated that matter and energy are interchangeable, and that matter is simply another form of energy. Matter is constantly vibrating in the form of tangible solids and intangible gases, and is constantly altering, being affected by, or interacting with energy. Energy is inherent in the living human body, and the human body is sustained by energy (Figure 1.1).

The ancients mastered techniques to balance the body's energy (Qi) in order to live in harmony with the environmental (Earthly) Qi as well as the universal (Heavenly) Qi. Traditional Chinese Medicine maintains that when living things start to lose their Qi, they lose their vitality. An ancient Chinese saying states, "Life comes into beginning because Qi is amassed; when Qi is scattered, the person dies."

Qi is stored within the body in the form of pools, creating the structures of the internal organs. From these internal pools, the body's life-force energy flows in the form of rivers and streams. These energetic rivers and streams form the body's vessels, channels, and collateral systems.

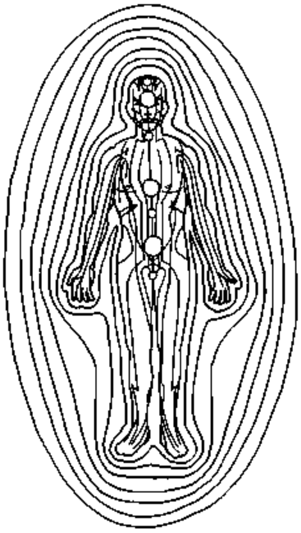

Figure 1.1. The Body's External Qi Field

THE FIVE DOMINIONS OF ENERGY

The ancient masters observed that Qi can be divided into five manifestations of matter and energy: mineral, plant, animal, human, and divine. Each form draws on the energy of the next, resonating and interacting with the divine through the form's relationship in Wuji (infinite space). The five manifestations of matter and energy are explained as follows.

1. The mineral's energetic field is considered the densest (i.e., the slowest) or lowest form of energetic vibration. The disintegration or division of the mineral's particles combine with the elements of air and water to form the Earth's soil. Every particle in the soil still retains the original primordial energy force of the mineral, which interacts with the energy of the divine.
2. The plant's energetic field is considered the next higher form of energetic vibration. All of

the Earth's vegetation (trees, bushes, flowers, herbs, etc.) absorbs a part of its life-energy from the mineral's energetic field, increasing and multiplying its energetic potential. The plant's energetic field is considered the next higher step in energetic evolution towards the divine energetic field.

3. The animal's energetic field is considered the next higher form of energetic vibration. The animal consumes and absorbs the energy from the plant's energetic field, further increasing and multiplying its energetic potential, bringing it one step closer towards the divine energetic field. Within each higher frequency of vibration there is also an increase in consciousness and level of awareness.
4. The human energetic field is considered the next highest form of energetic vibration. Mankind stands between Heaven and Earth, partaking of both energy fields. Through diet, Qigong practice, prayer, and meditation, humans can further refine and multiply their energetic potential.
5. The divine energetic field is the highest vibrational expression of energy known. As it envelops and becomes active within the human body, it further increases and multiplies the body's energetic potential, allowing man to attain divine consciousness.

All these energetic fields originate from one source, and all contain the vibrations of the one divine life-force. Likewise, with an attitude of deep respect for plants and animals that give up their life-force energy for our consumption, it is possible to enhance the nutritional value of the substances they provide us with. The blessing of food, and food prepared with a loving attitude, allows for the absorption of not only the vitamins and minerals contained therein, but also the absorption of the higher vibrations of the one divine energy inherent in all things. This is why many ancient cultures, often referred to as "primitive," prayed before hunting so that the animal spirit would willing give itself for sacrifice. Prayers were also given after the kill to free the animal's spirit so that it could return back to the divine.

Yang .	**. Yin**
active .	. passive
creative .	. receptive
masculine .	. feminine
front .	. back
left .	. right
fire .	. water
hot .	. cold
dry .	. wet
hard .	. soft
light .	. heavy
bright .	. dark
heaven .	. earth
sun .	. moon

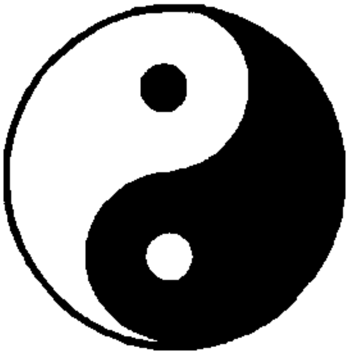

Figure 1.2. The table above shows some characteristics of Yang and Yin. Below is the Yin/Yang symbol: white represents Yang and black represents Yin. The small circles, one white and the other black, symbolize the fact that Yin is always transforming into Yang and Yang into Yin. (For more on Yin and Yang, see Chapter 3.)

Once individuals becomes aware of the divine energetic field, they begin to experience the refined vibrational energy fields of minerals, plants, animals and human beings. This increased awareness of the divine life-force energy strengthens the awareness of one's own energetic fields and that of others. This in turn can deepen the conscious and unconscious energetic connections between ourselves and others, be they human, animal, plant or mineral.

DEFINING THE ENERGY OF YIN AND YANG

Each of the five energetic fields can be further divided into Yin and Yang aspects. In Traditional Chinese Medicine (TCM), the theory of Yin and Yang energy represents the duality of balance and harmony within the body, as well as within the universe (Figure 1.2). Earth energy is Yin, while Heaven energy is Yang.

Yin exists within Yang, and Yang within Yin. Yang manifests as active, creative, masculine, hot, hard, light, and bright. Yin manifests as passive, receptive, feminine, cold, soft, and dark. The dynamic balance of Yin and Yang always changes and transforms the body's life-force energy. (See Chapter 3 for more on Yin and Yang energy.)

Successful practitioners in balancing the body's Yin-Yang energies were considered masters or "immortals," able to harmonize the body with the mind, the mind with the will, the will with the breath, the breath with the spirit, the spirit with motion, and finally, motion with the surrounding environment (Earth), the universe (Heaven), and the divine (Dao).

UNDERSTANDING THE CONCEPT OF QI

While the concept of Qi may seem complicated, it is actually very simple. Matter progresses to energy and energy to spirit. Qi is the medium, or bridge, between matter and spirit. Once we become aware of the reality of Qi, it becomes easily recognized.

Through observation and study, Chinese Qigong (pronounced chee-gung) masters discovered that each organ in the human body has a different function and a different speed of energetic vibration. By tracing the pathways (channels) Qi takes through each organ and observing the effects on bodily functions, the Chinese developed the basic theories upon which Qigong practice is founded. For thousands of years, Chinese medicine has successfully cured serious illnesses by stimulating the body's energy in very specific ways.

Through the study of Qigong, anyone wishing to cultivate awareness of the energy vibrations and their own individual pathways can learn to influence and even control them. Qigong practitioners use these skills to heal and strengthen the immune system, and to improve the functioning of various organ systems within the body. China Healthways International estimates that in Beijing alone more than 1.3 million people practice some form of Qigong every day, whereas, in China as a whole, around 80 million people practice Qigong.

DIFFERENT SCHOOLS OF QIGONG

Qi means "life-force energy" and gong means "skill," so Qigong is the skillful practice of gathering, circulating, and applying life-force energy. In China today, Qigong practice is divided into three main schools: medical, martial, and spiritual. The three schools are all based on the same philosophical system and share many of the same meditations and techniques. The schools differ primarily in focus. Students choose a school based on the use to which they want to put their Qigong training. Briefly, each school focuses on one of the following specialties:

1. The medical school trains doctors and healers in special Qigong methods for health maintenance and longevity, disease prevention, and the diagnosis and treatment of diseases and disorders. The three primary techniques of Medical Qigong therapy include the following.
    a. Purging to detoxify the body of pathogens,
    b. Tonifying to strengthen the body's internal organs and systems, and
    c. Regulating to balance the body's internal energy.
2. The martial school trains martial artists to

build their strength and power for performing martial arts applications. The three primary techniques of martial Qigong training include the following.

   a. Obvious Power (Ming Jing) techniques emphasize the training and conditioning of the muscles, strengthening the bone structure, and increasing the individual's overall stamina. This school also includes such techniques as pounding the body (arms, hands, legs, and torso) to strengthen and toughen the tissues.
   b. Hidden Power (An Jing) techniques emphasize stretching and twisting the tendons and ligaments (known as Reeling and Pulling the Silk) to cultivate resonant vibration within the body for striking and issuing power.
   c. Mysterious Power (Hua Jing) techniques emphasize the training and conditioning the mind's imagination and intention, to project and utilize the power of the individual's Shen (Spirit).

3. The spiritual school trains practitioners who seek spiritual transformation and enlightenment (Daoism, Buddhism, and Confucianism each have their own unique techniques). Their techniques include meditations for fusing, as well as releasing the Three Ethereal Souls (Hun). These souls can best be understood as personifications of moral qualities (or archetypes). When the Hun are fully developed, the practitioner acquires certain extraordinary powers and abilities, such as soul travel. The goal however, is to achieve transformation and a state of enlightenment, and not be led astray by the glamor of extra powers. The three primary techniques of spiritual Qigong training include the following:
   a. Nourishing the Spirit (Shen), to strengthen and refine the power of the individual's Shen,
   b. Housing the Shen by disciplining both thoughts and emotions, to relax and tranquilize the individual's Shen, and to become more receptive to divine energy and guidance, and
   c. Combining the Shen with the Qi, to coordinate the breath and intention for directing the spirit to guide the body's life-force energy.

Qigong training involves all of the individual's physical senses. The concentration is focused on breathing, hearing, visualizing, and muscle relaxation. Massage, and movement are also used to develop and control the body's intrinsic energy. Studying Qigong requires not only comprehending the immeasurable wisdom gathered for medical, martial, or spiritual development but also studying the ancient Chinese culture inherent within these systems.

MEDICAL QIGONG DEFINED

All living bodies generate an external field of energy called Wei Qi (pronounced "whey chee"), which translates as "protective energy." The definition of Wei Qi in Medical Qigong is slightly different than that of Traditional Chinese Medicine (TCM). In classical TCM texts, the Wei Qi field is seen to be limited to the surface of the body, circulating within the tendon and muscle tissues. In Medical Qigong, however, the Wei Qi field also includes the three external layers of the body's auric and subtle energy fields. This energy originates from each of the internal organs and radiates through the external tissues. There the Wei Qi forms an energy field that radiates from the entire physical body. This field of Qi protects the body from the invasion of external pathogens and communicates with, as well as interacts with, the surrounding universal and environmental energy fields.

Both internal and external pathogenic factors affect the structural formation of the Wei Qi. The internal factors include suppressed emotional influences (such as anger and grief from emotional traumas); The external factors include environmental influences when they are too severe or chronic, such as Cold, Damp, Heat, or Wind, etc. Physical traumas also affect the Wei Qi field.

Any negative interchange affects the Wei Qi by literally creating holes within the matrix of the individual's external energetic fields. When left unattended, these holes leave the body vulnerable to penetration, and disease begins to take root in the body. Strong emotions, in the form of toxic energy, become trapped within the body's tissues when we hold back or do not integrate our feelings. These unprocessed emotions block the natural flow of Qi, thus creating stagnant pools of toxic energy within the body.

Medical Qigong consists of specific techniques that use the knowledge of the body's internal and external energy fields to purge, tonify, and balance these energies. Medical Qigong therapy offers patients a safe and effective way to rid themselves of toxic pathogens and years of painful emotions that otherwise, can cause mental and physical illness. This therapy combines breathing techniques with movement, creative visualization, and spiritual intent to improve health, personal power, and control over one's own life.

MEDICAL QIGONG TRAINING IN CHINA

There are numerous colleges of Traditional Chinese Medicine throughout China today that focus on Medical Qigong training. The majority support the scientific study and expansion of Medical Qigong applications and Traditional Chinese Medicine treatments.

According to Qigong master and doctor of Traditional Chinese Medicine, Professor Zhou Qianchuan, all of the most famous Chinese doctors of acupuncture and moxibustion, herbal medicine, bone setting, and massage therapy, either practiced Qigong or incorporated Qigong into their clinical practices.

Major traditional Chinese medical colleges in China offer comprehensive, government-sponsored, three-year programs in Medical Qigong therapy. Programs include classes, labs, and seminars on traditional Chinese medical theory. These studies include: The foundations of Chinese medicine for internal diseases according to *the Yellow Emperor's Inner Canon, Spiritual Axis, Essential Questions,* and *the Canon of Perplexities*. The Medical Qigong classes also include energetic anatomy and physiology, diagnosis and symptomatology, energetic psychology, Qigong pathology, Medical Qigong therapy, as well as a survey of other related medical modalities. The other related modalities include: a comprehensive understanding of herbal medicine, acupuncture therapy, and Chinese massage. Classes of Western anatomy and physiology, Western internal diseases, and health and recovery, are also required.

During the certification program, three to five training hours a day accompany the standard six-day-a-week classroom curriculum. Course content, personal mastery of energy extension, and diagnosis techniques are rigorously tested each week. Upon completing the required courses and passing the final exams, the student receives a certificate of completion. Next, a six-month to one-year internship is required at a program-affiliated hospital or clinic. Upon successful completion of this internship, the new doctor is licensed as a doctor of Medical Qigong therapy by the People's Republic of China's Bureau of Scientific Technology.

Each internship program is assigned a separate wing in the selected Chinese hospitals. Both inpatient and outpatient facilities are available to the public. Each branch has specific approaches to healing a patient, with its own unique set of ground rules for diagnosis and treatment.

There are three distinct supervisory levels working within each clinical branch of Traditional Chinese Medicine in China. The first and lowest position is that of a "doctor of Medical Qigong," who is responsible for the treatment of all clinical patients (including patients in both the inpatient and outpatient clinics). The positions of Medical Qigong doctor are generally filled by the medical college graduates who have spent four to five years in clinical study and practice. The next level is called a "physician or doctor in-charge," and denotes a senior position within the clinic. This individual is responsible for the supervision of all the Qigong doctors' clinical procedures. This position is usually obtained after spending a minimum of five years as a Qigong doctor. The final and highest level is called a "director or professor;" this position requires overseeing the doctors in-charge, as well as teaching, treating, and training of other doctors to pass on Qigong clinical knowledge to future generations. This position is usually obtained after spending a minimum of five to six years as a doctor in-charge.

The licensing is reviewed and issued by either the People's Republic of China's Bureau of Scientific Technology (that issues a license in local city hospitals) or by the Ministry of Scientific Technology (that licenses to practice in any clinic or hospital throughout China). The Qigong doctor's skills are tested through oral, written and practical examinations, and a license is issued accordingly. In China today, there are five positions available for a doctor of Traditional Chinese Medicine. These five positions are described as follows.

1. A Doctor of Acupuncture Therapy (D.Ac.) specializes in the five main modalities of Chinese acupuncture.
2. A Doctor of Herbal Medicine (D.H.M.) specializes in the five main modalities of Chinese herbology.
3. A Doctor of Massage Therapy (D.M.T.) specializes in the five main modalities of Chinese massage and tissue regulation.
4. A Doctor of Medical Qigong Therapy (D.M.Q.) specializes in the five main modalities of Chinese Medical Qigong.
5. A Doctor of Traditional Chinese Medicine (D.T.C.M.) is a doctor who has trained in all four branches of Traditional Chinese Medicine (acupuncture, herbs, massage, and Medical Qigong).

MEDICAL QIGONG TRAINING IN THE UNITED STATES

The Five Branches Institute, College and Clinic of Traditional Chinese Medicine became the first TCM/Acupuncture College in the United States to open a Medical Qigong Clinic on April 7th., 2000. This clinic not only serves as an invaluable teaching resource for the Five Branches acupuncture students, but also serves as an affordable alternative medical clinic for the general public.

Founded in 1984, the Five Branches Institute has been known not only for its leadership in TCM education, but also as an academic innovator as well. It was the first, and remains the only, TCM College to establish a Neurology Center (founded in 1995) for the specialized treatment of paralyzed patients, under the directorship of the world-famous Chinese physician Doctor Ming Qing Zhu, L.Ac., Dipl. Ac. (NCCA), M.D. (China).

THE FIVE BRANCHES MEDICAL QIGONG CLINIC

The Five Branches Medical Qigong Clinic is currently divided into 3 treatment areas. The main area (Clinic A) is used to treat patients with less severe conditions; the second area (Clinic B) is used specifically for the treatment of cancer patients. The third treatment area is designed for video taping the patient's Medical Qigong prescription exercises and meditations.

Clinic A is the primary care facility of all patients. The patients are first assessed, and then assigned to a treatment team. Each team is comprised of at least one senior intern and several clinical students. Each team diagnoses, treats, and then prescribes Medical Qigong exercises and meditations for each patient. Some patients with potentially terminal illnesses (cancer, tumors etc.) may require specialized treatments, and are referred to the Clinic B.

The main goal of the Clinic B is to specialize in Oncology, and to work closely with the Academic Dean and Clinical Medical Director Dr. Joanna Zhao, L.Ac., Dipl. Ac. (NCCA), D.T.C.M. (China), who prescribes herbal therapy to complement the Qi Emission therapy and Medical Qigong prescriptions.

The Director of the Medical Qigong Clinic is Dr. Jerry Alan Johnson, whose duty is to oversee the "Doctors in Charge" of the 3 treatment areas, as well as regulate the patient load between Clinics A and B.

The Doctors in Charge of the Medical Qigong Clinics are: Jean Ruth Vlamynck, L.Ac., Dipl.Ac., M.T.C.M., M.Q.T., and Adam Atman, L.Ac., Dipl.Ac., M.T.C.M., M.Q.T. Their duty is to oversee the Medical Qigong doctors on staff within the three treatment areas, supervise the current interns, and make themselves available for consultations when needed.

The front entrance of the Medical Qigong College at the Hai Dian University in Beijing, China.

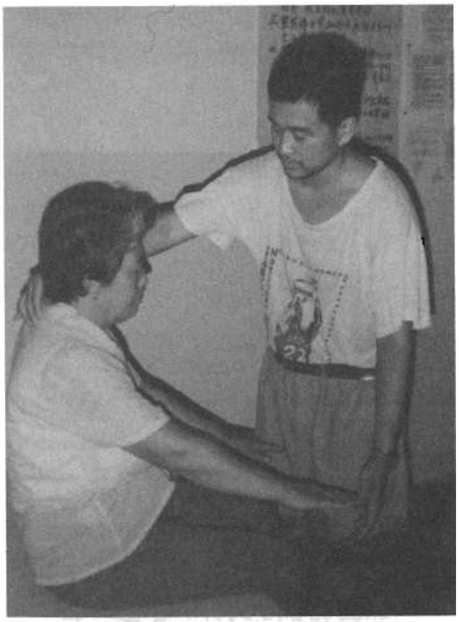

Dr. Lu Guo Hong, Director of the Hai Dian Medical Qigong College, treats a patient suffering from hypertension by dispersing Excess Heat from her body.

Right: The graduate students of the Hai Dian Medical Qigong College intern at the Xi Yuan Hospital in Beijing, China, for a period of 6 months to a year. They are licensed through the government's Medical Qigong Science and Research Institutes, which monitor the student programs.

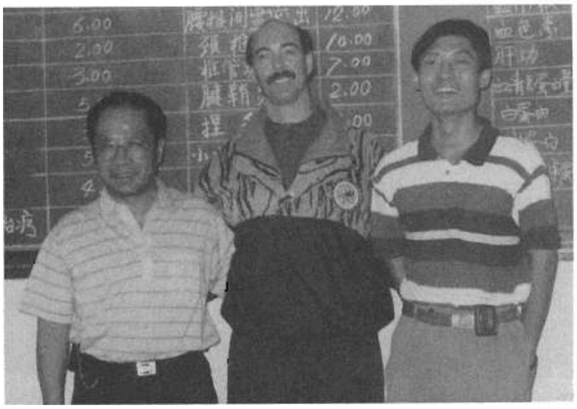

Left to right: Dr. Li Fu Dong, Associate Professor and Chief Director of Medical Qigong Science, Dr. Jerry Alan Johnson the author and director of the International Institute of Medical Qigong, and Dr. Pang Dong Hui, Executive Deputy President of the Hai Dian University.

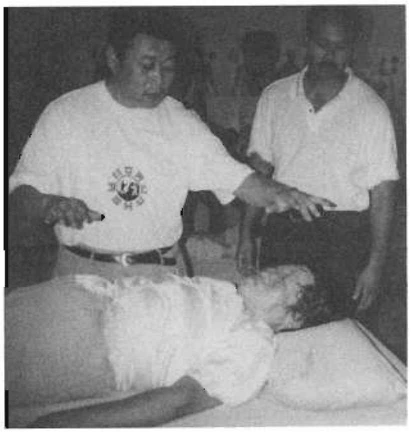

Dr. Niu Yu Hua, Assistant Director of the Hai Dian Medical Qigong College, treats a patient by regulating the Qi of her Upper and Lower Dantian.

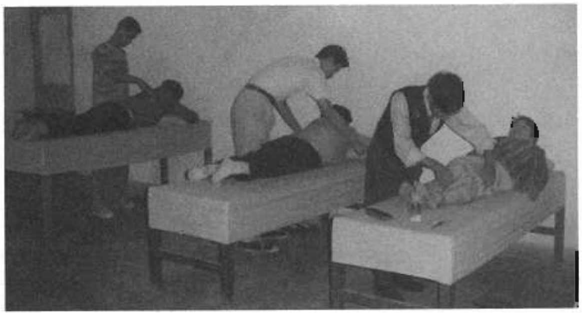

The Xi Yuan Hospital in Beijing, China, is the equivalent of the Stanford or Harvard Medical Center in the United States; it specializes in five distinct branches of healing therapies: Medical Qigong, acupuncture, herbs, Chinese massage, and Western medicine.

Three Americans completed their internships at the Xi Yuan Clinic in 1995. They are shown here posing with Dr. Xu Hongtao, the Xi Yuan Hospital staff supervising physician. *Left to right:* Dr. Seth Lefkowitz, the author Dr. Johnson, Dr. Xu Hongtao, and Dr. Arnold Tayam.

Right: The China Beijing International Acupuncture Training Center is a facility in which doctors from around the world culminate their training in Chinese medicine with Chinese doctors who are experts in their medical specialties.

The Xi Yuan Hospital is equipped with both inpatient and outpatient Qigong clinics. Here, Dr. Xu Hongtao administers "distance" Qigong therapy to patients in one of the many Medical Qigong wards.

Left to right: Qigong Master Zheng Zhanding, Jonathan Liu, and the author discussing advanced Medical Qigong theories and modalities.

CHINA BEIJING INTERNATIO
ACUPUNCTURE TRAINING CE

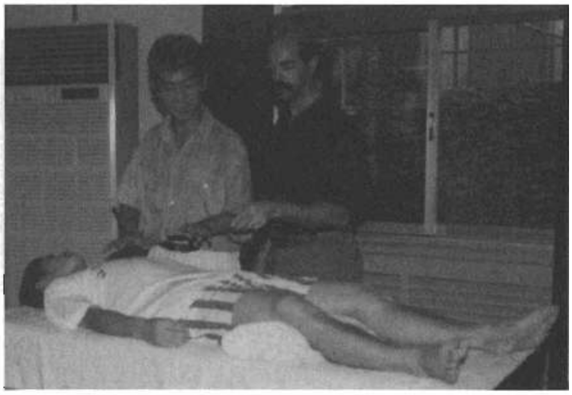

Dr. Teng Yingbo *(left)*, the president and secretary general of the Beijing Western District Qigong Science and Research Institute, with the author treating a patient.

Left: The International Institute of Medical Qigong is one of the few American Medical Qigong colleges that maintains the same strict standards as the Medical Qigong Universities in China. The initial training programs generally require three to four years of training.

The 1999 graduating class of the International Institute of Medical Qigong (*left to right*):

Back row: Matthew B. Weston, William H. Lewington, William H. De Groat, Paul E. Miller.

Middle row: Katy Reed, Madeleine H. Howell, Brooks M. Fiske, Diane de Terra.

Front row: Michael J. Finch, Jean R. Vlamynck, Dr. Johnson, Dennis M. Earnest.

Not pictured: Todd Mathew Gedryn, Luc Arnauld Logan, Geoffrey Greenspahn and Rose Mary Stewart.

WORLD CONFERENCE ON MEDICAL
WORLD ACADEMIC SOCIETY OF MEDICAL QIGONG

The author lecturing at the World Academic Society of Medical Qigong 1996 World Conference, on the psychophysical manifestations associated with tumor formation.

The 1997 graduating class of the International Institute of Medical Qigong (*left to right*):

Back row: Dr. Stephanie Taylor, Pamela Lee Espinoza, Anne Elderfield, and Dr. Carole Marie Kelly.

Front row: Dr. Arnold E. Tayam, Dr. Johnson (director of the institute), Shannon K. Brown, and Dr. Seth Lefkowitz.

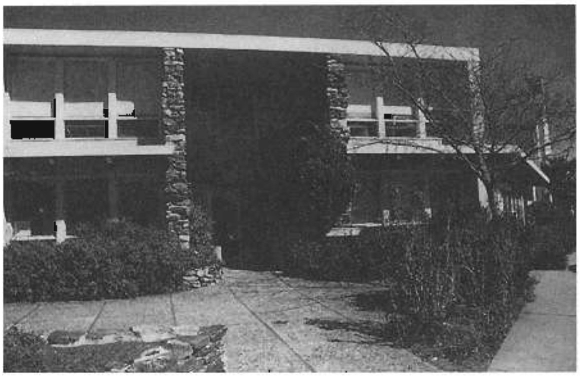

The Five Branches Institute, College and Clinic of Traditional Chinese Medicine became the first TCM/ Acupuncture College in the United States to open a Medical Qigong Clinic on April 7th., 2000.

In the main treatment area (Clinic A), patients are categorized according to the severity of their condition. For serious conditions (e.g., multiple sclerosis) patients are encouraged to lie on the treatment tables.

The resident Doctors in Charge of the Five Branches Medical Qigong Clinic are Adam Atman L.Ac., Dipl.Ac., M.T.C.M., M.Q.T. (positioned on the left), and Jean Ruth Vlamynck L.Ac., Dipl.Ac., M.T.C.M., M.Q.T. (located in the middle). They are seen here is discussing the current patient load with Dr. Arnold Tayam D.M.Q. (China).

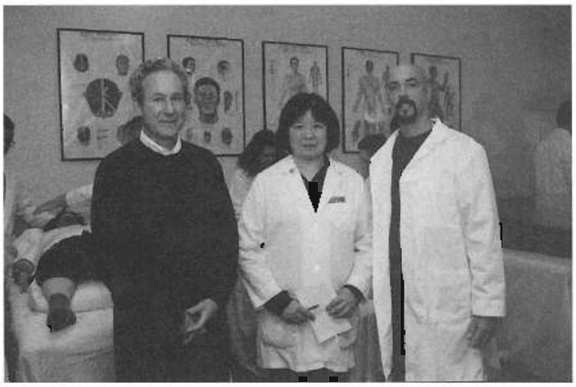

Pictured from left to right are the Five Branches President and CEO Ron Zaidman, M.B.A., M.T.C.M.; the Academic Dean and Clinical Medical Director Dr. Joanna Zhao L.Ac., Dipl. Ac (NCCA), D.T.C.M. (China); and the Dean of Medical Qigong Science and Director of the Medical Qigong Clinic Dr. Jerry Alan Johnson.

In Clinic A of the Five Branches Medical Qigong Clinic, patients with less serious conditions (e.g., hypertension and migraine headaches) are encouraged to relax on one of the treatment chairs.

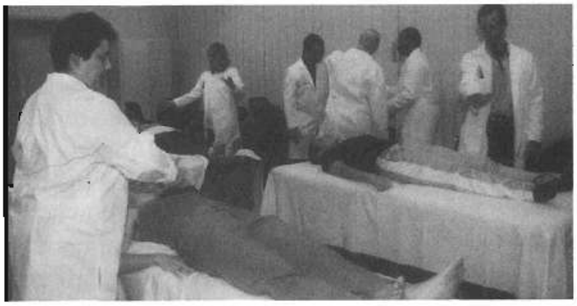

The second treatment area (Clinic B), is used specifically for the treatment of tumor and cancer patients.

The Five Branches Medical Qigong students are also taught Energetic Psychology and how to deal with the emotional discharges coming from their patients.

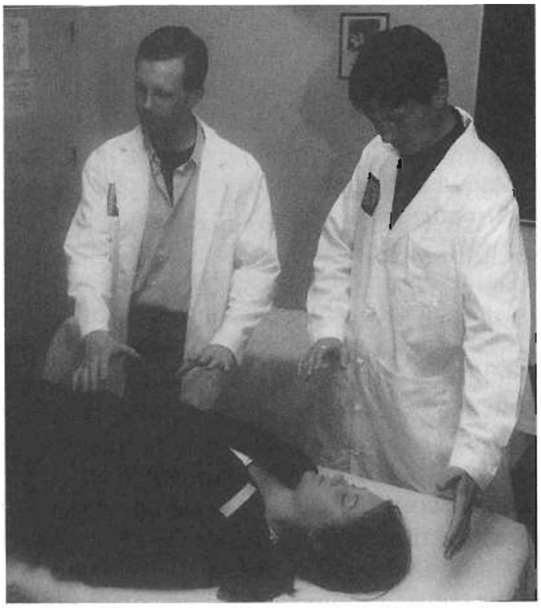

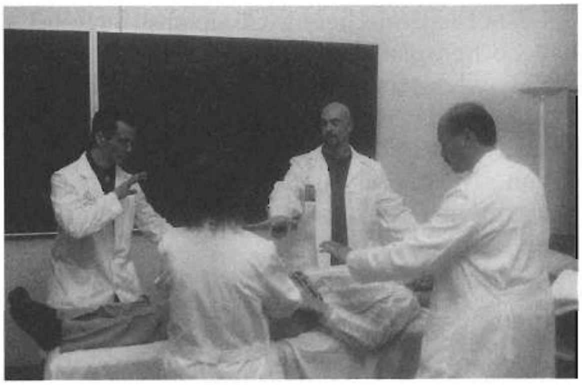

Doctors of the Five Branches Medical Qigong Clinic combine their energies to treat a specific type of cancer, observed here, in Clinic B.

Within the first semester, the Medical Qigong students at the Five Branches TCM College have learned and experienced basic applications of Medical Qigong purgation, tonification and regulation exercises. They have also been taught the underlying principles for each exercise and meditation. Furthermore, they are able to discriminate when to assign these prescriptions to patients, and when not to.

In the first phase of internship, the Medical Qigong students work as part of a Qigong treatment team, assisting a senior intern.

The Medical Qigong Clinic also includes a psychotherapist, Madeleine Howell, L.M.F.T., M.Q.T., who is trained in Medical Qigong therapy. She is available to patients who may need professional consultation after undergoing Medical Qigong treatments, as these may bring painful emotions to the surface as part of the healing process.

On the first official day of the clinic, 15 Qigong interns and doctors from the International Institute of Medical Qigong performed Qi Emission through Distance therapy on more than 80 patients. Treatments ranged from a general treatment (purging excess, tonifying deficiencies and dispersing stagnations) to the treatment of several potentially terminal cancer patients.

THE FIVE BRANCHES MEDICAL QIGONG COURSE AND CLASSES

The Medical Qigong students at the Five Branches college learn and experience basic applications of Medical Qigong purgation, tonification and regulation exercises during the first semester. Classes, include lectures and labs. They are also taught to understand the underlying principles of each Medical Qigong exercise and meditation, and learn when and when not to assign these exercises in cases for clinical application. The students are also taught Energetic Psychology and how to deal with the emotional discharges coming from their patients.

In the second semester the Medical Qigong students are led through a progressive series of advanced Shengong meditations and exercises, for developing advanced intuitive diagnostic skills. The students also learn basic Medical Qigong treatment protocols, and begin the initial phase of clinical internship. During this phase they assist the senior interns.

From the third semester on, the students begin an in depth study in Medical Qigong theory and its application. The goal is to complete a three year certification program in Medical Qigong Therapy.

MEDICAL QIGONG CLINICAL THEATRE

Conducted in a small class setting, the Medical Qigong Clinic gives the students exposure and first hand experience at Medical Qigong diagnoses and treatments. This is a hands on experience where the different aspects of energetic dysfunctions and symptom manifestations are studied, treated and observed by the Medical Qigong students. Through the careful guidance of the instructor, the Medical Qigong students are allowed to assist in all phases of diagnosis and treatment.

After completing the Clinical practicum, Medical Qigong students at the Five Branches TCM College will be qualified to diagnose and treat patients, as well as document and maintain accurate clinical records.

CLINICAL INTERNSHIP

In the first phase of internship, the Medical Qigong students work as part of a Qigong treatment team, assisting a senior intern.

In the second phase of internship, the Medical Qigong students work with their own patients, under the supervision of the instructor.

By the end of their Clinical Internship, the Medical Qigong therapists will receive their Masters degree, and will be fully qualified to diagnose and treat patients without supervision.

TRADITIONAL CHINESE MEDICINE AND MEDICAL QIGONG THERAPY

Medical Qigong therapy is the oldest of the four branches of Traditional Chinese Medicine and provides the energetic foundation from which acupuncture, herbal healing, and Chinese massage originated. It is through the understanding of Qigong that the other branches of Traditional Chinese Medicine are elevated to a spiritual path of self-realization and internal transformation. Doctors of Traditional Chinese Medicine address the patients' physical, energetic, and spiritual needs simultaneously. According to the principles of Traditional Chinese Medicine, the root cause of all disease can be traced to a critical imbalance within the body's vital energies. Therefore, the best way to prevent or cure disease requires establishing a healthy energetic balance and harmony between the body's energy field and the forces of nature and the cosmos.

Traditional Chinese Medicine is divided into four branches of clinical healing: Acupuncture

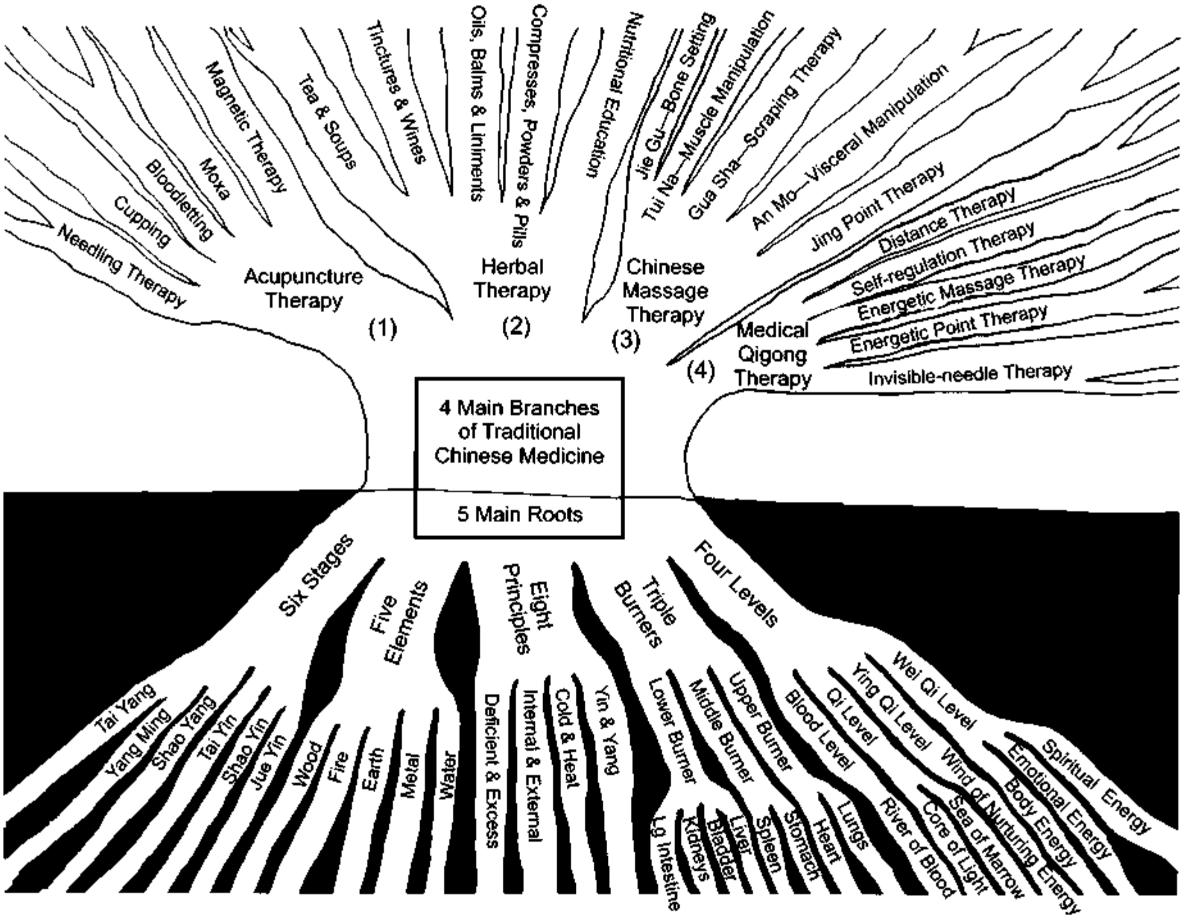
Needling Therapy
Cupping
Bloodletting
Moxa
Magnetic Therapy
Tea & Soups
Tinctures & Wines
Oils, Balms & Liniments
Compresses, Powders & Pills
Nutritional Education
Jie Gu—Bone Setting
Tui Na—Muscle Manipulation
Gua Sha—Scraping Therapy
An Mo—Visceral Manipulation
Jing Point Therapy
Distance Therapy
Self-regulation Therapy
Energetic Massage Therapy
Energetic Point Therapy
Invisible-needle Therapy
Acupuncture Therapy
(1)
Herbal Therapy
(2)
Chinese Massage Therapy
(3)
Medical Qigong Therapy
(4)
4 Main Branches of Traditional Chinese Medicine
5 Main Roots
Six Stages
Five Elements
Eight Principles
Triple Burners
Four Levels
Tai Yang
Yang Ming
Shao Yang
Tai Yin
Shao Yin
Jue Yin
Wood
Fire
Earth
Metal
Water
Deficient & Excess
Internal & External
Cold & Heat
Yin & Yang
Lower Burner
Middle Burner
Upper Burner
Lg Intestine
Kidneys
Bladder
Liver
Spleen
Stomach
Heart
Lungs
Blood Level
Qi Level
Ying Qi Level
Wei Qi Level
River of Blood
Core of Light
Sea of Marrow
Wind of Nurturing Energy
Body Energy
Emotional Energy
Spiritual Energy

Figure 1.3. The Four Main Branches of Traditional Chinese Medicine

therapy, Herbal therapy, Chinese Massage therapy, and Medical Qigong therapy. All of the four main branches are built on the same foundation of energetic diagnosis known as the Five Main Roots of Traditional Chinese Medicine. The five main roots are used for internal organ diagnosis according to the Six Stages, Five Elements, Eight Principles, Triple Burners and Four Levels theories (Figure 1.3).

ACUPUNCTURE THERAPY

By inserting very thin metal needles into specific points, the acupuncturist manipulates the patient's energy (Qi) to achieve an overall physical balance. In an acute or emergency situation, a patient's symptoms are treated first, after which the focus is directed to rerouting Qi Deviations or removing the original cause of the symptoms (e.g., stagnations).

The energetic points are specific areas both on and in the body where Qi emerges from deep within the body's organs and tissues, or submerges to travel deep within the body. The Qi moves along specific pathways known as channels and collaterals. Collaterals are the smaller streams of Qi that branch off from the main ener-

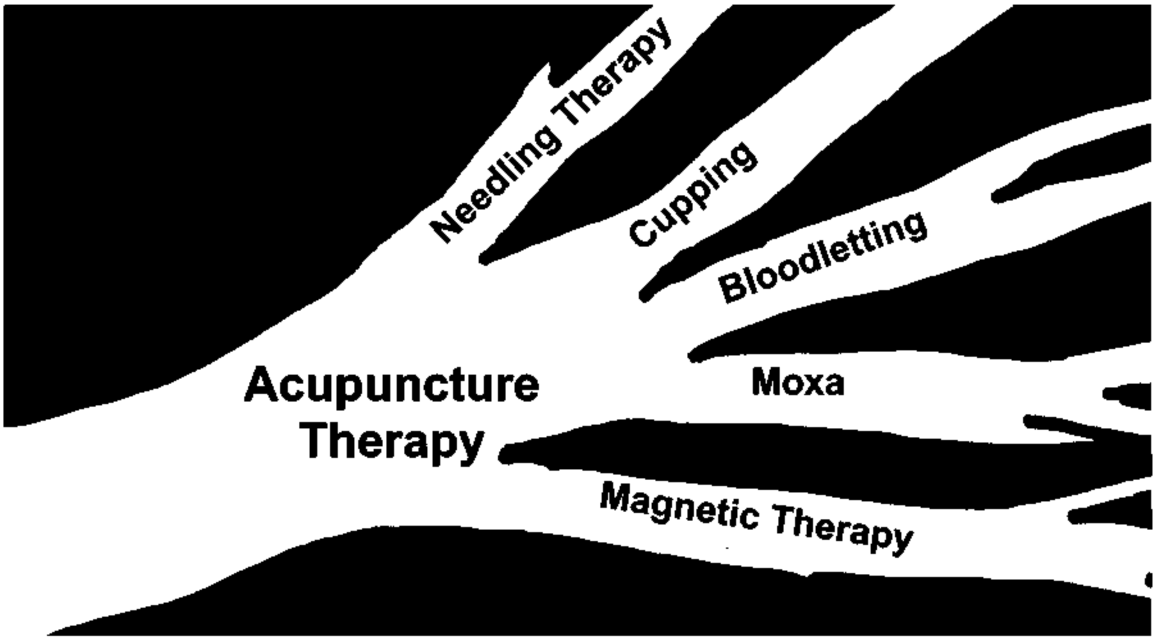

Figure 1.4. The Five Main Branches of Chinese Acupuncture Therapy

getic rivers (channels). Each Primary Channel takes its name from one of the six Yin or six Yang organs to which it corresponds. Each organ has two channels, one on each side of the body.

Often in Chinese medical diagrams and procedures, you will see abbreviations referring to specific energetic points. Traditional Chinese Medicine names the points based upon the channel on which the point lies and its distance along the channel. The first points on each side of the Gall Bladder Channels (GB), for example, are GB-1. Some channels contain over 60 points, each numbered sequentially from beginning to end. These points are also given descriptive names according to their location and energetic affect upon the body when treated. The GB-1 points, for example, are located next to the eyes, on the outer canthus, level with the pupils, and are called the Pupil's Seam.

Acupuncture therapy includes five major treatment techniques (Figure 1.4): needling, cupping, bloodletting, moxa burning, and magnet healing.

1. Needling uses acupuncture needles of various sizes that are inserted into channel points. These points are tiny areas where the Qi pools along the streams of an energy channel (Figure 1.5). The needles stimulate the nerves and energy flow to tonify or reduce Excess Qi.
2. Cupping uses wooden, clay, or glass cups that adhere to the patient's skin by suction (Figure 1.6). This suction drains, or removes, pathogenic Qi from the body's pores. This technique can also be used to tonify specific areas of the body. This modality of treatment has been successfully combined with bloodletting to treat acute sprains accompanied by Blood stagnation.
3. Bloodletting is done with instruments such as blood needles or seven- and five-star hammers to remove Toxic Qi, Blood stagnation, Heat, and other pathogenic factors (Figure 1.7). The hammer has five to seven sharp projections that pierce the skin and cause slight bleeding. The acupuncturist diagnoses and then monitors the patient's condition by the

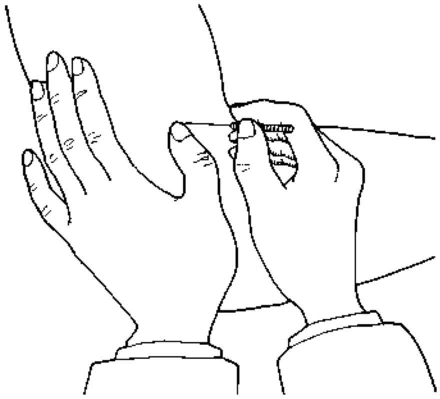

Figure 1.5. A doctor is inserting an acupuncture needle into the patient's Channel Point. Acupuncture needles of various sizes are used to stimulate nerve and energy flow for the tonification of Deficient Qi or the reduction of Excess Qi.

different shades of the patient's Blood. Trapped or diseased Blood is released until a healthy color is observed. This therapy is considered useful for treating disorders of the nervous system, physical trauma, and extremely serious febrile diseases.

4. Moxa Burning employs lighted herbal cones or sticks that are inserted on top of acupuncture needles (or held over specific channel points) to infuse heat and Qi into specific body areas for tonification (Figure 1.8). This technique is also used to expel Cold in order to disperse Blood stagnation.
5. Magnets use magnetic patches or strips that are attached to various channel points of the patient's body to stimulate a response in the electromagnetic field (Figure 1.9). Magnet therapy has been used since the Tang Dynasty (618–907 A.D.). The magnets are applied to specific points for a period of 3–5 days, removed for one day, then reapplied. Whether used for tonification or sedation, this therapy facilitates constant treatment of the channel point.

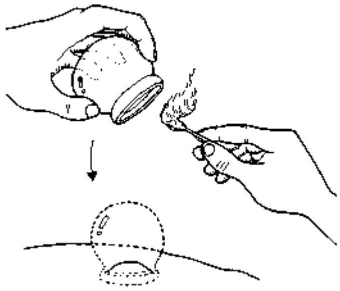

Figure 1.6. In Cupping, heated air is directed into wooden, clay, or glass cups, which are then placed on the patient's skin. The cooling air creates suction that drains, or removes, pathogenic Qi from the body's pores.

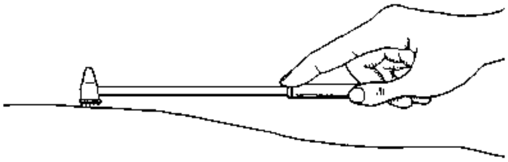

Figure 1.7. The doctor is lightly tapping with the Five-Star Hammer that has five sharp projections to pierce the skin and cause slight bleeding. The color of the patient's Blood provides the doctor with information about the patient's condition.

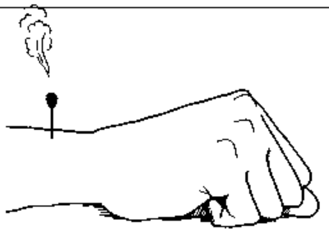

Figure 1.8. The doctor has inserted a needle with burning Moxa (herbal cones/sticks) on top of it. Burning moxa may also be held over specific channel points or placed directly on the skin.

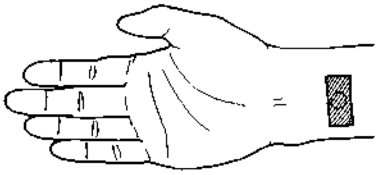

Figure 1.9. A Magnetic Patch is being worn on the Wrist. Magnets may be attached to various parts of the body or may be worn in shoes or wristbands.

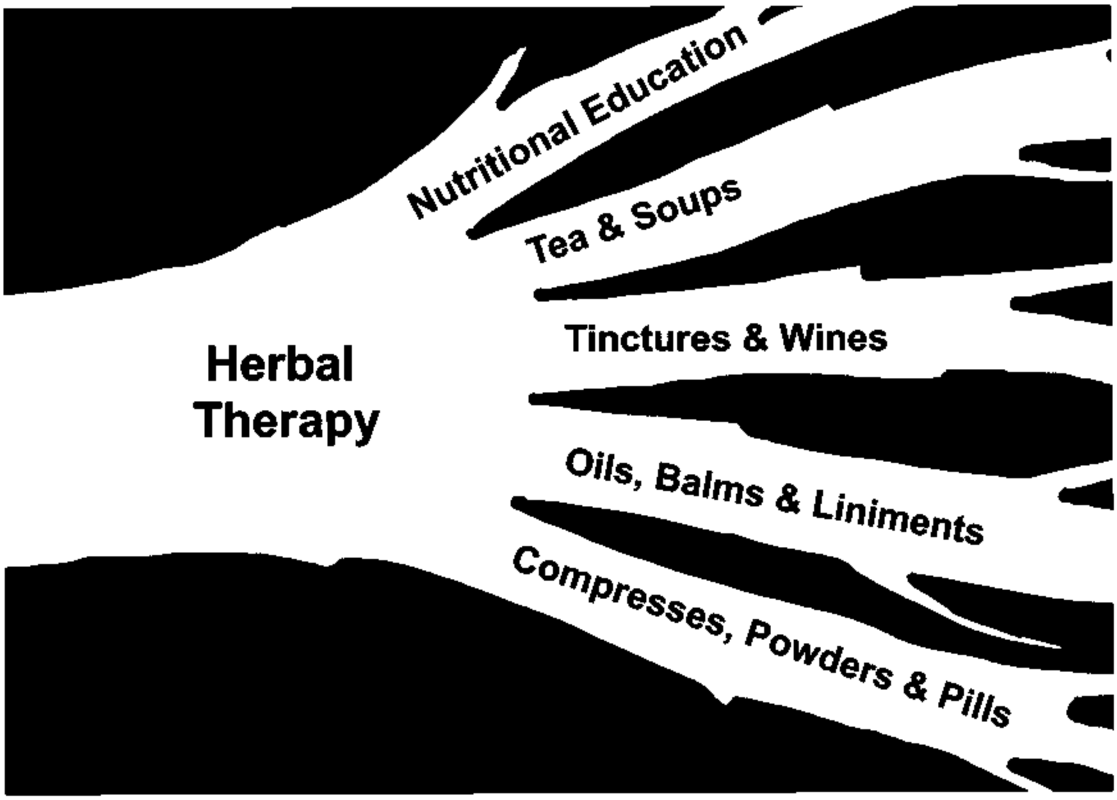

Figure 1.10. The Five Main Branches of Chinese Herbal Therapy

HERBAL THERAPY

Herbal formulas have been used successfully to treat a variety of illnesses for over 5000 years. Historically, herbal medicine has been the worldwide basis for pharmaceuticals used in most cultures, prior to modern times. Today, herbs provide the source for many of the pharmaceuticals used in contemporary Western medicine, especially for the treatment of viral and bacterial diseases, pain, tumors, chronic diseases, internal and external tissue regeneration, and many other medical problems.

Herbology is both a science and an art. An herbalist spends many years studying the herbs used to create herbal formulas. The herbalist must understand the effects of individual herbs, as well as their synergistic effects when combined.

Herbs are used for tonifying, purging, dispersing, warming, cooling, nourishing the Yin, nourishing the Yang, and clearing Heat, as well as moving Qi, Blood, Phlegm, and Fluids within the body. They cause the Qi in the body to either ascend or descend, affecting the upper or lower parts of the body.

Chinese medicine prescribes specific parts of plants (leaves, roots, bark, etc.) for particular medicinal purposes. Parts of trees, shrubs, herbs, vines, and flowers are selected for their specific properties (Hot, Cold, Warm, Cool) and taste (Sweet, Bitter, Pungent, Sour, and Salty). These properties either tonify or disperse Qi and Blood.

Chinese medical herbology not only includes the cultivation and gathering of seeds, fruits, flowers, leaves, barks, stems, and roots but also prescribes non-herbal components (such as minerals, animal or insect parts) that are sometimes added to enhance the herbs' healing effect.

Chinese herbs cure energetically by moving Qi in the channels. Different herbs enter different channels and affect different internal organs. Herbs are extremely powerful. The herbalist uses herbs to tonify (strengthen) and move Qi and Blood as well as eliminate Heat from the patient's Blood. When tailored to an individual's constitution or combined into a formula for specific symptoms, herbs can greatly help the body; however, taking the improper herbal formula can have deleterious effects.

Chinese herbal therapy includes five major clinical applications (Figure 1.10): nutritional education (food and diet); teas and soups (tang); tinctures and wines (jin); oils, balms, and liniments (you and gao); and compresses, powders (san), and pills (wan).

1. Nutritional education is stressed to assist patients in choosing foods for the body's nourishment and optimum health, as well as for the treatment of disease. Foods have many similar properties to herbs. An old Chinese saying asks, "Are herbs food or food herbs?", thus stressing the importance of a good diet
2. Teas and Soups are water-based herbal formulas traditionally prepared from raw or processed herbal ingredients. These are traditionally ingested for the treatment of internal and external disorders of both acute and chronic natures.
3. Tinctures and Wines are both alcohol-based herbal formulas. Tinctures are a concentrated alcohol-based formulas prepared from raw herbs that are used similarly to teas and soups. Wines are traditionally applied externally to alleviate pain, or ingested as a tonic, depending on the specific formula and the disease being treated.
4. Oils, Balms, and Liniments are oil-based herbal formulas usually applied externally for the treatment of muscle, tendon, and ligament trauma, to alleviate pain, disperse Excess Qi, or to draw Qi into specific areas for tonification.
5. Compresses, Powders, and Pills consist of herbs that have been pulverized into a paste, powder, or poultice and are then applied externally for the treatment of acute or chronic injuries. They can also be ingested for internal organ tonification and the rebalancing of the body's energy system. Pills are herbs specifically prepared from traditional formulas, rolled into little balls, and orally ingested for the treatment of Internal disorders.

CHINESE MASSAGE THERAPY

Chinese Massage Therapy is a generic term used for all tissue manipulation techniques currently used in China. This particular branch of Traditional Chinese Medicine consists of five popular systems that include manipulation of not only external skin, muscles, tendons, joints, nerves, and inner fascia but also the internal organs and organ systems.

This ancient therapy is used as a preventative treatment as well as a healing modality. By applying specific methods of tissue manipulations, obstructions in the channel's pathways can be removed, promoting and increasing both Qi and Blood circulation.

This therapy focuses on improving the structural alignment of the body and on healing soft-tissue injuries. It also corrects any deviant functions of the internal organs, nerves, and joints. Chinese bodywork and tissue therapy are the foundational source for modern Swedish massage, myofascial trigger point therapy, reflexology, and therapeutic neuromuscular therapy.

Chinese massage therapy is divided into five different schools of instruction: Jie Gu, Tui Na, Gua Sha, An Mo, and Jing Point therapy (Figure 1.11). Jie Gu, Tui Na, and Gua Sha employ external tissue manipulations. These three external manipulations are used to treat the bones, muscles, ligaments, and tendons, and also to treat fevers. An Mo and Jing Point therapy utilize soft-tissue manipulation. These treatment modalities are similar to those used in chiropractic, osteopathy, Western physical therapy, and massage therapy.

1. Jie Gu Therapy is used for bone setting and

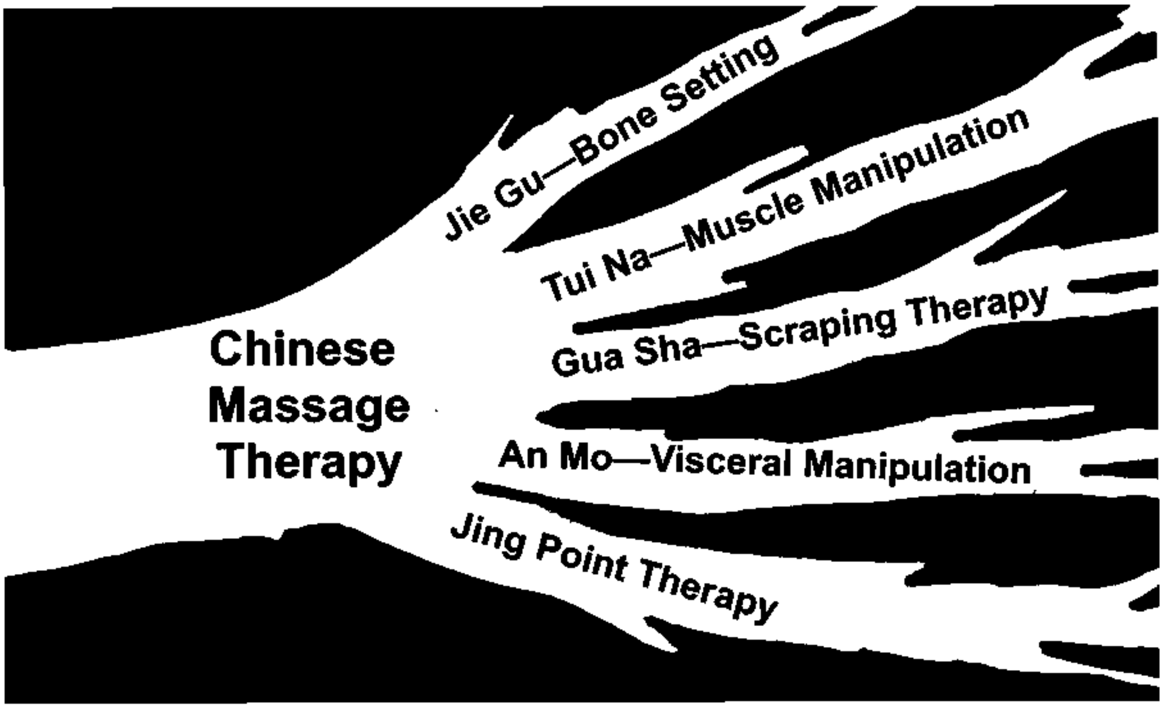

Figure 1.11. The Five Main Branches of Chinese Massage

to adjust the patient's body alignment (Figure 1.12). The literal translation for Jie Gu is "knotted bone," which describes the art of manipulating the bones and ligaments to unravel the trapped junctions of Blood and Qi channels in the patient's joints.

2. Tui Na Therapy focuses on external tissue manipulation and adjustment of the muscles and tendons to adjust abnormal Qi circulation within the body's muscular system (Figure 1.13). The translation for Tui Na is to "push and grasp." It was developed primarily for correcting the misalignment of the body's bones and muscles due to traumatic physical injuries. Reflexology is historically rooted in the use of Tui Na therapy in Chinese pediatric care.
3. Gua Sha Therapy is used to regulate febrile conditions, such as flu, cholera and malaria, and to treat musculoskeletal conditions. The translation for Gua is "to scrape or scratch" and Sha is defined as "cholera," or sand-like maculae (referring to the red discoloration that is raised on the skin by the application of scraping). This therapy focuses on external surface tissue scraping, usually around the neck and thorax areas (Figure 1.14). It is commonly used for promoting Qi and Blood circulation, removing toxins, clearing Heat, cooling the Blood, removing stagnation, and dissolving masses. A jade scraper (coin, bowl, or spoon) is used for purifying the Qi and transforming the Shen (Spirit). A water buffalo horn is commonly used for pulling Heat and toxins from the patient's body (occasionally ceramic is used, but never glass or plastic).

The two internal manipulations are used to treat the organs and nerves.

4. An Mo Therapy is used for internal organ

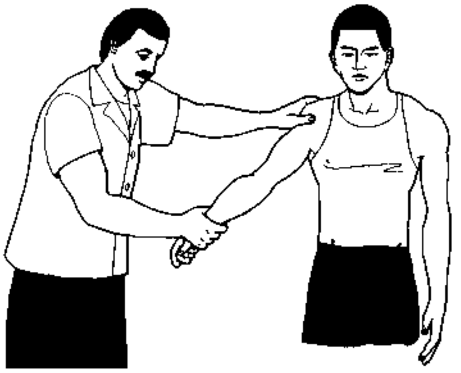

Figure 1.12. Jie Gu Therapy is used to set the bones and ligaments in order to unravel the trapped junctions of Blood and Qi channels at the patient's joints.

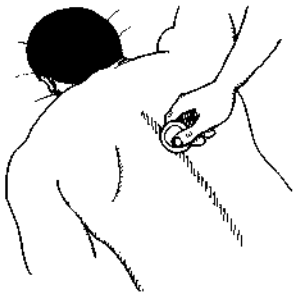

Figure 1.14. Gua Sha Therapy is commonly used for clearing Heat, cooling the Blood, removing stagnation, and dissolving masses.

regulation. An Mo focuses primarily on Qi extension and soft-tissue and internal organ manipulation (Figure 1.15). Although the literal translation means to "press and rub," this therapy focuses primarily on internal visceral regulation, concentrating directly on the treatment of specific Internal diseases.

5. Jing Point Therapy is employed for channel and internal organ regulation. Jing point therapy uses pressing, pinching, clapping, and tapping techniques on specific energetic points and energetic channels. These techniques are employed to promote Qi and Blood circulation, balance the body's Yin and Yang energy, tonify weak organs, dredge the channels, and expel pathogenic factors (Figure 1.16).

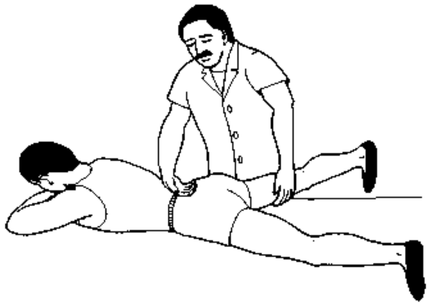

Figure 1.13. Tui Na Therapy is used to adjust the muscles and focuses on external tissue manipulation and adjustment of the muscles and tendons in order to correct abnormal Qi circulation within the body's muscular system.

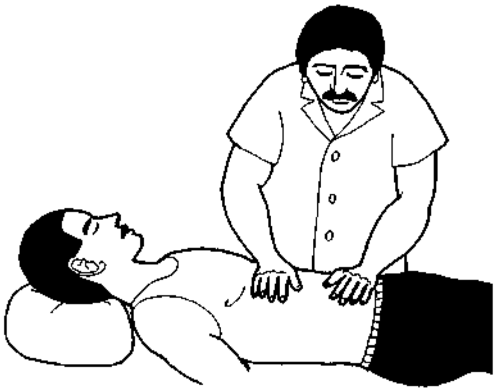

Figure 1.15. An Mo Therapy allows organ manipulation for internal visceral regulation and concentrates directly on treating specific Internal diseases.

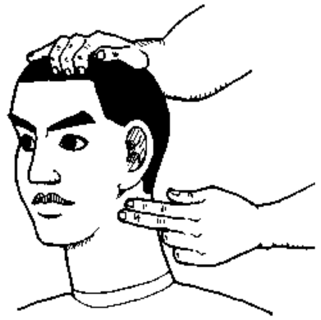

Figure 1.16. Jing Point Therapy is used to promote Qi and Blood circulation, balance the body's Yin and Yang energy, tonify weak organs, dredge the channels, and expel pathogenic factors.

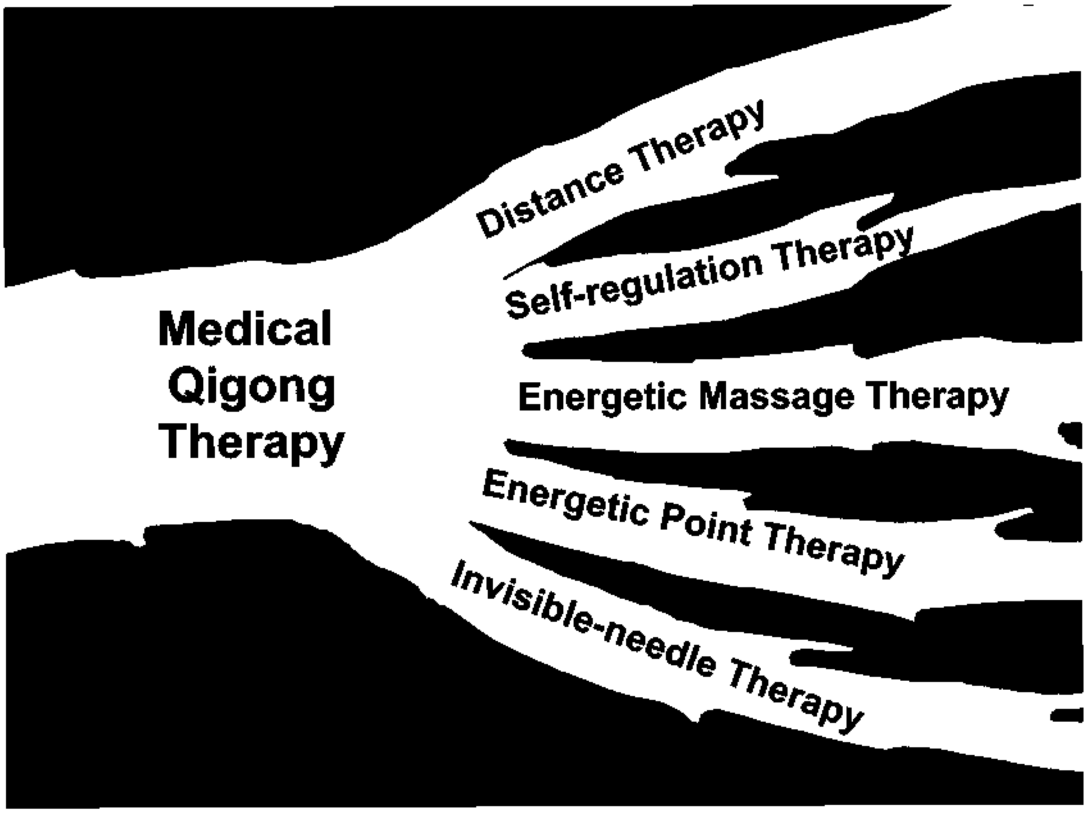

Figure 1.17. The Five Main Branches of Chinese Medical Qigong Therapy

MEDICAL QIGONG THERAPY

The objective for healing disease in Medical Qigong training is threefold:

- First, eliminate Internal pathogenic factors (the accumulation of Excessive emotions such as anger, grief, worry, fear, etc.) as well as External pathogenic factors (the invasion of Cold, Hot, Damp, etc., from the environment).
- Second, increase or decrease the patient's Qi as needed to counteract the Deficient or Excess condition within the internal organs and channels.
- Third, regulate and balance the patient's Yin and Yang energy to bring it back into harmony.

This unique therapy consists of regulating the body's three external Wei Qi fields (physical, mental/emotional and spiritual), and the four Internal fields of life-force energy (Ying Qi, Sea of Blood, Sea of Marrow, and the Taiji Pole). Some of the most common diseases treated in Medical Qigong clinics are: diabetes, arthritis, high blood pressure, breast and ovarian cysts and tumors, migraine headaches, fibromyalgia, insomnia, acute abdominal pain, irritable bowel syndrome, deep tissue obstruction, muscle atrophy, brain tumors, stroke, coma retrieval, and certain types of cancer. The medical treatment focuses on relieving pain, detoxifying the body of toxic emotions (e.g., excessive anger, fear, worry, etc.), correcting internal organ dysfunctions, and balancing Excess or Deficient Qi and Blood conditions.

Medical Qigong therapy uses five major clinical modalities (Figure 1.17): Distance therapy, Self-

Regulation therapy, Qigong Massage therapy, Energetic Point therapy, and Invisible Needle therapy.

1. Distance Therapy (also called Qi Emission) requires the Qigong doctor to manipulate a patient's Qi by focusing on the energetic properties of the patients' channels, collaterals, and points, as well as internal organs, from a distance of several inches, several feet, or even several miles away (Figure 1.18).
2. Self-Regulation Therapy (also called Qigong Prescriptions and Patient Homework) are Qigong exercises (postures, movements, sound vibrations, visualizations, etc.) given to patients by a doctor (Figure 1.19). Patients can use these Qigong techniques to regulate their own health, using various lying, sitting, moving, and standing postures. The patients may also use their own spiritual belief system as a healing tool.
3. Qigong Massage Therapy, a soft-tissue regulation technique, differs from Tui Na or An Mo (Chinese External Massage Therapy), in that the doctor's hand skims the patient's body as lightly as a feather, never exceeding the pressure one would place on an eyeball (Figure 1.20). The light skimming action is used to dredge the patients' external channel Qi, causing energy to be released from the internal channels themselves, which serve as pathways for Qi transference.
4. Energetic Point Therapy is used by the doctor to extend Qi into specific internal and external areas of the patient's body to lead and direct the Qi. This type of therapy requires the doctor and the patient to focus their attention onto a specific energetic point (Figure 1.21). This therapy demonstrates the power of the mind as an active tool in healing, and is used for purgation, tonification, and regulation.
5. Invisible Needle Therapy involves the visualization of imaginary needles of light being inserted into specific points on the patient's body. The needles of light are used to stimulate and direct the patient's Qi (Figure 1.22).

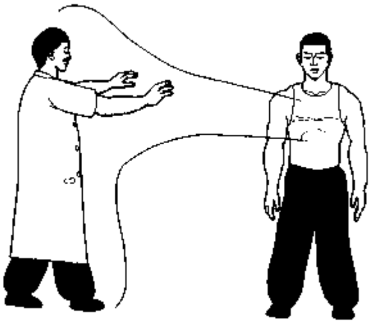

Figure 1.18. In Distance Therapy, the Qigong doctor manipulates a patient's Qi by focusing on the energetic properties of the patient's external channels, collaterals, and points from a distance of several inches, several feet, or even several miles.

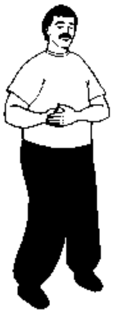

Figure 1.19. In Self-Regulation Therapy, the patients are required to self-regulate by performing Qi Qigong "prescriptions" or "homework" (postures, movements, chants, visualizations, etc.). Here the patient regulates his own *Liver Qi*.

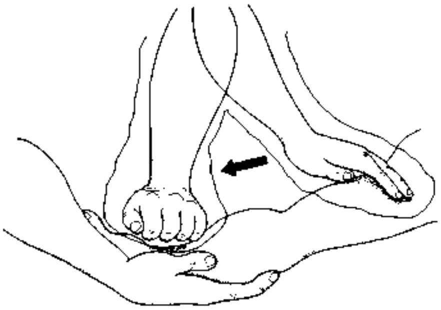

Figure 1.20. In Qigong Massage Therapy, the doctor softly dredges the patient's external channels in order to release energy from the internal channels themselves, which serve as pathways for Qi transference.

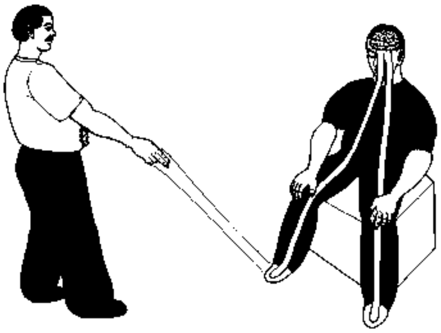

Figure 1.21. In Energetic Point Therapy, the doctor extends energy into the patient's body as both the doctor and the patient focus their attention on a specific channel point, in this case the Kd-1 point.

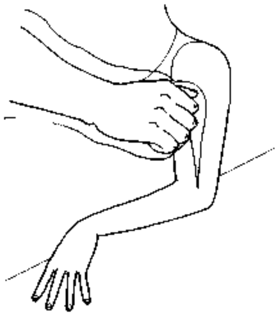

Figure 1.22. In Invisible Needle Therapy, the doctor visualizes inserting energetic acupuncture needles into the patient's channel points in order to stimulate the patient's Qi.

CHAPTER 2
ENERGETIC FORMATION OF THE HUMAN BODY

In 1993 when I began my internship at the Xi Yuan Hospital in Beijing, China, I became aware that most Qigong doctors had a limited understanding of basic anatomy and physiology as we teach it in the West. When I asked why more attention was not paid to the subject, I was told, "The perspective and priority we place on gross physical anatomy and physiology is quite different from yours -- Westerners only study the dead -- we study life, the living pools, rivers, and currents of life-force energy that can only be found in the living body."

Eastern medicine believes that when the body dies, the substance, or energy, that gave the body life returns to its original source (Heaven and Earth) and all that remains of the individual is a mass of lifeless tissue. They believe that the tangible and the energetic exist in a twofold, cohesive relationship. A Qigong doctor must understand this concept of energy in order to comprehend and prescribe appropriate Medical Qigong therapy for the patient.

The concept of the energetic formation of the human body is new to Western thought with its primary focus on the physically tangible. The philosophical foundation of Traditional Chinese Medicine, on the other hand, includes studying the whole human being in all of his or her aspects, physical, mental, emotional, energetic, and spiritual. The conception results not only in the tangible physical form of a new human being but also in various contributing energies and energetic fields. Medical Qigong studies these energies, as well as the interplay between the Heavenly and Earthly influences.

ENERGETIC EMBRYOLOGICAL DEVELOPMENT

The study of the embryological development of the body's inner fascia and connective tissues explains many of the traditional Chinese medical ideas about energy, health, and disease. The cells, tissues, and organs of the human body interrelate as a result of the "energetic patterning" that occurs at conception.

Heaven (which the ancient Chinese considered a state of subtle and non-material energies and beings) and Earth both have energetic fields that influence life. If we consider Heaven Qi (the universal energy related to the sun, moon, and stars) to be electro-positive and Earth Qi (the environmental energy related to the earth, water and wind) to be electro-negative, we can begin to understand the duality of the energetic fields and their pull on the human body.

Think of the body as being suspended between two enormous fields of energy (Figure 2.1): the sun extends its energetic field down, and the Earth's small "sun," i.e. its core, extends its energetic field upwards through the Earth's crust. Man is, therefore, suspended between the electrically positive energy field of Heaven and the electri-

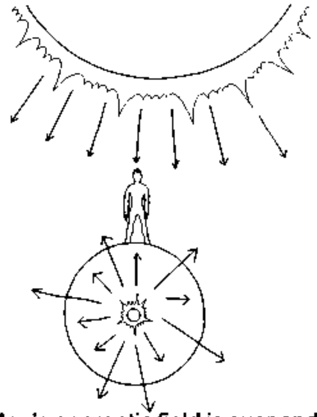

Figure 2.1. Man's energetic field is suspended between the two energy fields of Heaven and Earth.

cally negative energy field of Earth and is affected by both celestial and environmental phenomena.

ENERGY, MATTER, AND SPIRITUAL INTERACTIONS

The *Yellow Emperor's Classic of Internal Medicine* states that the woman's physiology is dominated by Blood, which serves as the foundation of her menstrual cycle, fertility, conception, pregnancy, and childbirth. A man's physiology is dominated by Qi, which serves as the basis of his physical interactions.

At conception, a myriad of energetic interactions begin. The mother, father, Heaven, and Earth all contribute energetically to conception. These four energies condense upon impact and form the nucleus of the energetic cellular patterning of the new life (Figure 2.2). Every part of the resulting human body is affected by this energetic patterning.

An old Chinese saying in Medical Qigong states that, "When people are born, Heaven gives them Jing (Essence) and Shen (Spirit) -- which align to form the mind -- and Earth gives them bones and shape, which unite to form the body. Joined together, these sources of energy cause human beings to develop. When people die, their Essence and Spirit return to Heaven, and their bones and shape go back to Earth."

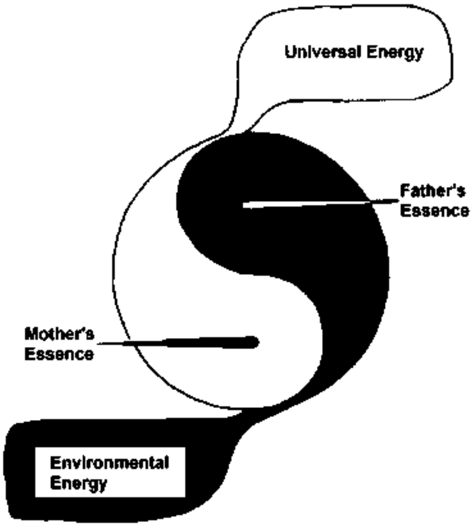

Figure 2.2. The blending of Heaven (universal energy) and Earth (environmental energy) as well as father and mother energies creates a fusion of Yin and Yang energies within the body's tissues and cells during creation. These energies are responsible, on a psychophysical level, for transferring talents and traits from generation to generation.

Pairs of channels from the mother's internal organs carry Qi that creates and nourishes the baby at each stage of development. As the baby develops, sustained through the umbilical cord, the child's navel, Kidneys, and lower abdominal area become the collection points for Prenatal Qi (energy stored within the baby's body before it is born).

Every life begins with inherent strengths and weaknesses. It is believed that when conception occurs, if the mother achieves orgasm, the child's inherent energy will become very strong. However, if the mother does not achieve orgasm during conception, the results will be a normal or weak energetic constitution. Prenatal care is mandatory for the healthy formation of the embryo. Before cell division, the DNA mass must be duplicated exactly in order to transfer normal genetic characteristics to the next generation. Although heredity plays a large part in the transference of both parents' genetic history, a weakness in the mother's channels can result in congenital problems, or toxins, that the fetus can acquire during one of the corresponding stages of development.

UNDERSTANDING FETAL TOXINS

The External invasions of pathogenic toxins (known as fetal toxins) can penetrate the zygote with latent Heat that can cause diseases during early childhood development. It is, therefore, important for both parents, especially the mother, to take responsibility for being strong and healthy at the time of conception and for the duration of the pregnancy. Toxins can be transferred into the embryo in utero in one of two ways:

- First, from either the mother or father at the moment of conception. Toxins transferred from the parents can create an inherited toxicity due to a retention of Hot Evil stored from within either of the parents' Essence and Blood; or
- Second, from Internal Heat generated by the

mother during pregnancy due to improper diet or life-style.

During pregnancy, a fetus is aware of light and sound and of the mother's reaction to the surrounding influences of her environmental energy fields. The fetus is strongly influenced by its mother's physical activities, as well as her mental, emotional, and spiritual states.

Regulating the mother's behavior to improve her child's physical, emotional, and mental health is called "fetal education" in Traditional Chinese Medicine, and is important in the development of the child's Prenatal Essence, Energy, and Spirit. This viewpoint is based on the fact that the mother's Heart and uterus are connected via the mother's internal channels, allowing Qi and Blood to flow into the uterus. Anything that influences the mother's mind, emotions, and spirit affect her Heart, which in turn, affect the fetus via the internal channels.

DEVELOPMENTAL SEQUENCE

The following description of the body's developmental sequence is but one of many theories used in order to explain Chinese energetic embryology. The ancient Chinese did not have a conception of the body's cells and cellular division; however, modern theories have postulated that the channels were formed at the earliest stages of cell division, creating an energy matrix for the developing fetus. This particular theory is being taught at the Hai Dian Medical Qigong College in Beijing, China.

Three important energies (Figure 2.3) combine in the developing fetus: Jing (Essence), Qi (Energy), and Shen (Spirit). Jing is the body's foundational substance, responsible for nourishing the tissues. Qi emerges out of Jing circulating through the tissues and promotes the body's metabolism. Shen governs the body. (see Chapter 11, "The Three Treasures of Man," for more on Jing, Qi, and Shen.)

Both parents contribute energetically to the baby's conception (Figure 2.4). The father's sperm and the mother's egg consist of Jing (Essence), Qi (Energy), and Shen (Spirit). This combination of Jing, Qi, and Shen is referred to collectively as fa-

Prenatal Energies—The Three Treasures		
Substance	**Physical Function**	**Energetic Prenatal Function**
Shen (Spirit)	governs the body	the Original Spirit
Qi (Energy)	promotes metabolism	body's Energy extending into Infinite Space (also called Wuji, or the Void)
Jing (Essence)	nourishes the tissues	body's innate true Essence, the foundational substance

Figure 2.3 Qi is the medium between matter (Jing) and Spirit (Shen) and is responsible for the embryo's transformational processes.

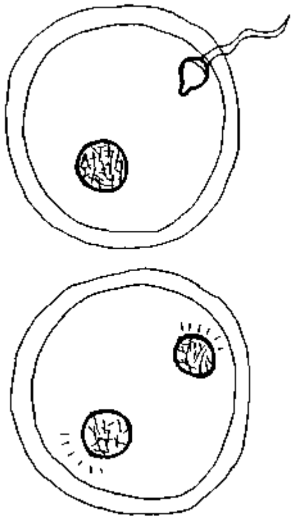

Figure 2.4. The human body is composed primarily of water. The Essence of man (sperm) and woman (egg) unite in the uterine sea to form the fetus. The ova is polarized at the entry point of the sperm, creating the original polar axis (the *Taiji Pole*) that determines the complex pattern of cellular division that occurs along the polar axis throughout development.

ther Qi and mother Qi. The combination of father Qi and mother Qi is known as Yuan Qi or Original Qi.

The quality and quantity of Original Qi that the baby receives at birth depends on three main factors:

- First, the purity and potency of both parents' genetic plasma (sperm and egg);
- Second, the condition of both parents' health and vitality and the state of their physical, mental, emotional and spiritual relationship at the time of conception; and
- Third, the spiritual factors surrounding the conception (i.e., karma that is brought into this life by the incoming spirit/soul).

As the sperm (containing the father's Qi) fertilizes the egg (containing the mother's Qi), Heaven (universal) Qi and Earth (environmental) Qi blend together within the zygote. The swirling and blending of these four energies form energetic pools (which will later evolve into organs), rivers (which later evolve into channels), and streams (which later evolve into collaterals).

When the sperm enters the ovum, it produces a polar axis that creates an energetic vortex. This vortex not only forms the central Taiji Pole, but also draws Qi from Heaven and Earth and the Eternal Soul into the fetus's body. This polarization also determines a ventral and dorsal surface, which become the embryo's Conception and Governing Vessels, respectively, at the first cell division (Figure 2.5). The Governing Vessel controls the cell division that eventually forms the back of the body, while the Conception Vessel controls the cell division of the body's front. This first cell division also establishes a right and left side. The Heel Vessels control the balance of Yin and Yang energy development in the two sides of the body.

The Belt Vessel and Thrusting Vessels form at the time of the second cell division. The four vessels formed at this point (Governing Vessel, Conception Vessel, Thrusting Vessels, and Belt Vessel) are interlinked for the production, circulation, and regulation of the body's Jing-Essence. The body's entire energy system becomes established and maintained within these four vessels when cell division occurs (Figure 2.6).

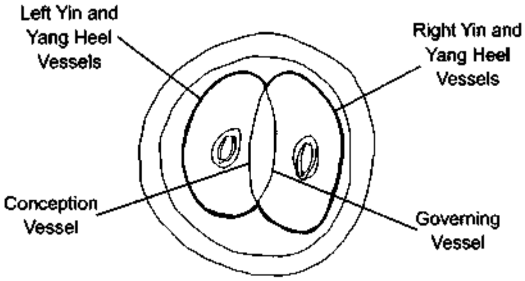
Left Yin and Yang Heel Vessels
Right Yin and Yang Heel Vessels
Conception Vessel
Governing Vessel

Figure 2.5. The electrical polarity and field of energy in the polar axis is related to the Original or Yuan Qi from which the Conception and Governing Vessels form the seas of Yin and Yang energy during the first cellular division. The Yin and Yang Heel Vessels are also established, forming the left and right sides. The exterior of the egg is determined by the Yang Linking Vessels. The interior of the egg by the Yin Linking Vessels.

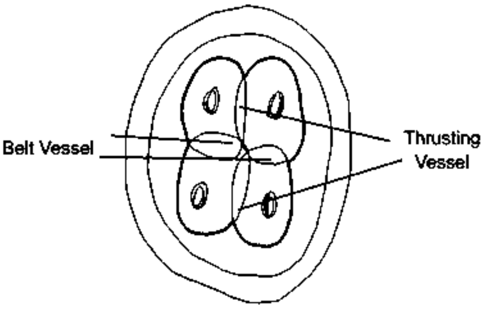

Figure 2.6. The second cellular division is caused by the development of the Belt Vessel and the Thrusting Vessel. The Eight Extraordinary Vessels have now formed.

While the embryo is forming, both the Yang and Yin Linking Vessels are respectively responsible for the exterior and interior development of the embryo (see Chapter 3 for more on Yang and Yin energy). Each of the Eight Extraordinary Vessels has a specific role in the development of the embryo.

1. The Governing Vessel (Yang) controls development of the body's back.
2. The Conception Vessel (Yin) controls development of the body's front.
3. The Thrusting Vessels carry energy through the center of the body and controls the body's center core.

4. The Yang Heel Vessels control the development of the body's right and left Yang energy.
5. The Yin Heel Vessels control the development of the body's right and left Yin energy.
6. The Yang Linking Vessels control the development of the exterior of the body (and correlates to Heaven energy).
7. The Yin Linking Vessels control the development of the interior of the body (and correlates to Earth energy).
8. The Belt Vessel binds all the channels together.

These eight vessels—Governing, Conception, Thrusting, Yang Heel, Yin Heel, Yang Linking, Yin Linking, and Belt—are also known as the Eight Extraordinary Vessels or Eight Prenatal Vessels.

The Eight Extraordinary Vessels form a vortex of energy at the center of the embryo's body—from the area between what will become the Kidneys. The Taiji Pole and Thrusting Vessels are at the center of this vortex and will form the Sea of Five Yin and Six Yang Organs, the Sea of Twelve Primary Channels, and the Sea of Blood (see Chapter 13). From the Taiji Pole and Thrusting Vessels, the body's Qi and Blood are distributed at the energetic level through small channels, or rivers of energy. This energetic vortex creates the energy for the growth of the embryo's physical form.

After the initial cell division is complete, the embryo's ten Yang channels and ten Yin channels begin the development and formation of the embryo's tissues and organs. These twenty channels are divided into two separate groups of energetic rivers known as the Eight Extraordinary Vessels and Twelve Primary Channels. As the embryo develops into a fetus and continues to grow, the twenty channels also continue to develop.

During the formation of the embryo, nine Yang channels begin to flow out of the Governing Vessel (Sea of Yang Qi) and form the nine Yang rivers known as: the Bladder Channels, Gall Bladder Channels, Stomach Channels, Small Intestine Channels, Triple Burner Channels, Large Intestine Channels, Yang Heel Vessels, Yang Linking Vessels, and Belt Vessel.

Consecutively, the nine Yin channels begin to flow out of the Conception Vessel (Sea of Yin Qi) and form the nine Yin rivers known as: the Kidney Channels, Liver Channels, Spleen Channels, Heart Channels, Pericardium Channels, Lung Channels, Yin Heel Vessels, Yin Linking Vessels, and Thrusting Vessels (see Chapter 8).

THE TEN LUNAR MONTHS OF CREATION

The following description of the sequence of embryological development was established by the late Chinese Medical Qigong expert Dr. Chao Yuan Fang during the Sui Dynasty, around 610 A.D. The months that Dr. Chao refers to are the ten lunar

Lunar Month	**Mother's Channels**	**Channel's Activity**	**Energetic Activity**
Month 1	Liver Channels	stop menses, begin embryonic growth	energies combine at conception, channels develop
Month 2	Gall Bladder Channels	saturate embryo with Jing	embryonic Qi transforms into amniotic fluid
Month 3	Pericardium Channels	mother's Qi and Body Fluids purify and cleanse the fetal Shen	Hun and Po establish in the fetus's internal organs
Month 4	Triple Burner Channels	the Yin organs develop	Water Jing accepted into fetus's body
Month 5	Spleen Channels	the Five Agents enter the Five Orbs (Yin organs)	Fire Jing accepted into fetus's body
Month 6	Stomach Channels	Six Pitches established within the Six Storage Areas	Metal Jing accepted into fetus's body
Month 7	Lung Channels	Seven Essential Stars open orifices to the light from Heaven and Earth	Wood Jing accepted into fetus's body
Month 8	Large Intestine Channels	Heart is harmonized and breath is quieted	Earth Jing accepted into fetus's body
Month 9	Kidney Channels	Kidney Channel controls the energetic intake from the umbilicus	energetic boundaries completed
Month 10	Bladder Channels	spatial cavities surrounding major organs are completed	Heaven & Earth Qi settle into baby's Lower Dantian and birth begins

Figure 2.7. Prenatal Energy Development

months. These months compose the 40 weeks of a normal pregnancy (Figure 2.7).

MONTH ONE

The first lunar month of pregnancy is traditionally called the Beginning of Form. This stage covers conception and early cell division.

At conception, the mother's Liver Channels stop her menses and begin to nourish the growth of her embryo (Figure 2.8). During pregnancy, the mother's Blood is transformed into Jing-Essence, that nourishes the mother's body as well as the embryo's. The mother's Liver Channels cause Essence and Blood to coagulate in her womb. This Blood coagulation continues after the initial cellular division.

At this stage the mother's Shen (Spirit) becomes part of a threefold activity:

- First, the mother's Shen projects through the umbilical cord like a light reflecting off a prism, sustaining and energizing the production of Jing (see Chapter 13), Qi, and Blood for the embryo (Figure 2.9).
- Second, the mother's Shen influences the embryo's Qi and Blood, and the embryo's Shen (Spirit) is stimulated, bringing it into active being (see Chapter 12).
- Third, the embryo's Original Jing combines with the embryo's Original Qi to create the embryo's Original Shen (Prenatal Spirit), which appears as multicolored light and contains the inherited knowledge of the fetus's ancestors, including talents, skills, and natu-

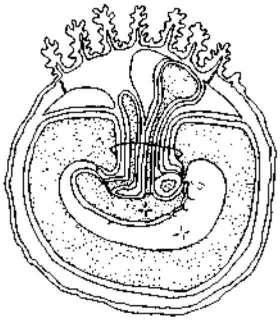

Figure 2.9. Sustained through the umbilical cord, the embryo absorbs the mother's Blood, Essence, Energy, and Spirit.

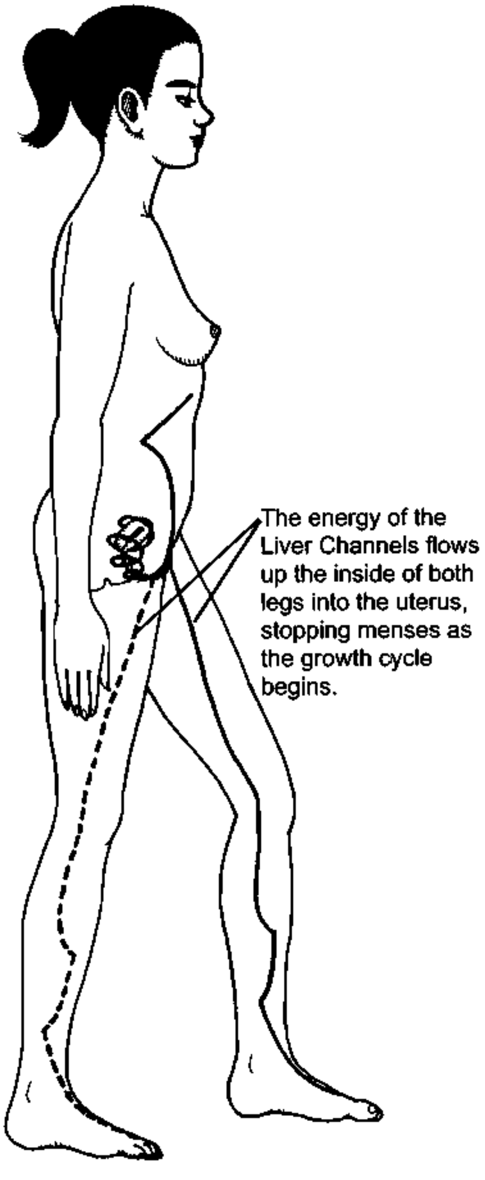
The energy of the Liver Channels flows up the inside of both legs into the uterus, stopping menses as the growth cycle begins.

Figure 2.8. The mother's and father's Essence (Jing), Energy (Qi), and Spirit (Shen) blend with the Heaven and Earth energies during the fusion of the sperm and egg. During the first lunar month, the mother's Liver Channels stop her menses and begin the embryonic growth cycle. After the initial cell division, the ten Yin and ten Yang Channels develop within the embryo.

ral abilities. This knowledge is stored deep within the fetus's cells, tissues, and consciousness.

MONTH TWO

The second lunar month of pregnancy is traditionally called the Beginning to Gel stage. During this period, the responsibility of the mother's Gall Bladder Channels is to create the environment for the development of the mesenteric membrane sac (Figure 2.10). The Gall Bladder Channels will also saturate the embryo, uterus, and placenta with Jing, causing the embryonic Qi to become denser, until it transforms into a thick liquid (amniotic fluid). The embryonic fluid will regulate the embryo's Body Fluids. At this stage, the embryo begins to take shape inside the uterine lining (Figure 2.11). The Lungs, Liver, Kidneys and major blood vessels are forming.

With the formation of the major blood vessels, the Yin and Yang energies begin to occupy the embryo's channels. As the Yin and Yang energies actively balance themselves:

- Energy that will later coalesce into Lung Qi moves to the upper part of the body;
- Original Qi (Yuan) of the Kidneys begins to collect deep in the center of the body;
- Earth Qi (the energy absorbed by the embryo from the mother's exposure to the outside environment) begins to collect in the lower front and upper back areas of the embryo's body.

All of these energies seek their own location

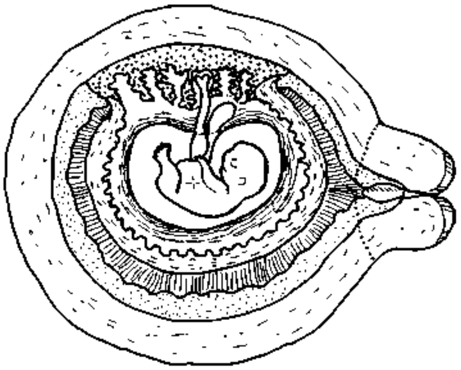

Figure 2.11. During the second month, the embryo begins to take shape as the energetic boundaries begin to form.

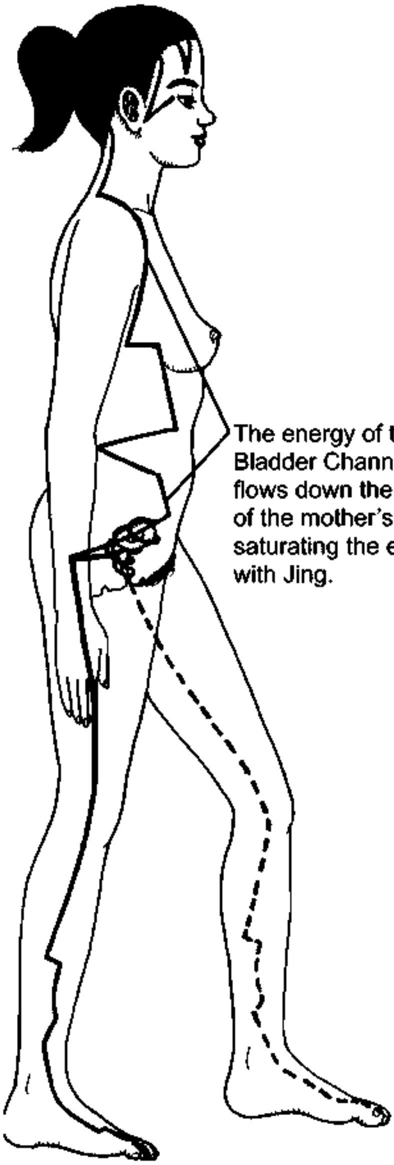

The energy of the Gall Bladder Channels flows down the outside of the mother's body, saturating the embryo with Jing.

Figure 2.10. The mother's Gall Bladder Channels saturate the embryo with Jing during the second lunar month, causing the embryonic Qi to transform into amniotic fluid. The embryo begins to take shape as the energetic boundaries, pools, spatial cavities, and energetic channels and collaterals create internal and external form.

within the embryo's developing body, creating their own pools of Qi with their own discrete energetic boundaries. These boundaries will later form the spatial cavities that surround the major organs. The areas where the energetic pools settle and begin to create a balance within themselves, will be called the organ's place of origin. This flux and movement creates tiny energy currents, eddies, and whirlpools that flow within the body. As the energy shifts, seeking balance, the larger pools of energy begin to form the major organs. This process continues, creating the brain, bones, and skin.

Once the energetic pools and rivers have formed, all of the currents and eddies evolve to form the energetic channels and collaterals, through which energy will flow continuously. These energetic currents will move in accordance with the mother's energetic respiratory patterns and will later (after birth) follow the rhythmic patterns of the child's respiration. As the energy currents continue to spiral within the channels, energetic points are established, following the body's energetic blueprint. Some of these areas spiral outward to form energetic exit points, while others spiral inward to form entry points.

MONTH THREE

The third lunar month of pregnancy is traditionally called the Beginning of the Pregnant uterus. During this period, the embryo becomes a fetus and begins micro-movement. Its heartbeat can now be detected.

The mother's Pericardium Channels control the presence and amount of Jing and Shen in the fetus's vessels, channels, and collaterals (Figure 2.12). The Jing and Shen that flow from the mother are ultimately rooted (firmly established) in the energy of her Blood. The combination of the mother's Qi and Body Fluids purifies and cleanses the Shen of the fetus, which was formed at conception. This purifying action transforms into Heat, causing the Yang energy to arouse the Hun-Three Ethereal Souls (the energy of the fetus's spirit), into life. The Five Agents, energies that stem from the Three Ethereal Souls, are currently in a state of awakening and will later reside in the organs.

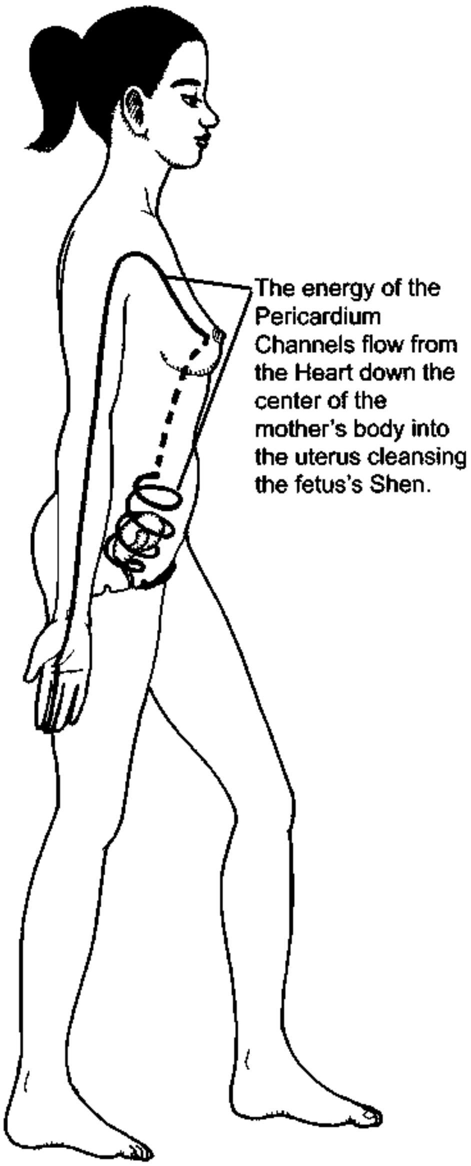
The energy of the Pericardium Channels flow from the Heart down the center of the mother's body into the uterus cleansing the fetus's Shen.

Figure 2.12. The mother's Pericardium Channels control the third lunar month of creation. The mother's Qi and body fluids purify and cleanse the fetus's Shen (Spirit). The Hun from Heaven and the Po from Earth are established within the fetus's internal organs.

The Five Agents are energies that are linked to a person's moral qualities and can be categorized as the five virtues of the Five Yin Organs. The Five Agents are also categorized into Five Elements. The psycho-emotional components of the Prenatal Five Agents and Postnatal Five Emotions are both stored within the body's Five Yin Organs, and are expressed as the child grows into adulthood (Figure 2.13). When one of the internal organs is stimulated, a Yin or Yang psycho-emotional reaction is created. Psycho-emotional energetic interactions are feeling manifestations expressed through either the positive moral qualities of the congenital Five Agents (Yang - kindness, order, trust, integrity, and wisdom) or the negative developed emotional experiences of the Acquired Five Emotions (Yin - anger, joy, worry, grief, and fear).

The Five Agents are connected to the Wu Jing Shen-Five Essence Spirits (Hun, Shen, Yi, Po, and Zhi), and stored within the energetic elemental nature of the body's Essence (Jing) of Wood, Fire, Earth, Metal, and Water. The Five Elemental Energies however, encompass not only the body, but all of the myriad phenomena of the body and nature, combining and recombining in infinite ways to produce manifested existence.

1. The Wood Agent (the Virtue of Kindness) represents benevolence and compassion. This agent is connected to the Hun-Three Ethereal Souls, and stored in the Liver. Wood affects the energetic channels' flow, tendons, ligaments, small muscles, peripheral nerves, iris of the eyes, vision, tears, bile, nails, and external genitalia. After birth, the Liver will also store the emotions of anger, irritability, blame, rage, resentment, and jealousy. When excess anger is eliminated, benevolence, compassion, and love for others is allowed to flourish.
2. The Fire Agent (the Virtue of Order) represents peace and boundary setting to foster social harmony. This agent is connected to the Shen-Spirit, and stored in the Heart, affecting the energetic channels' flow, blood vessels, complexion, perspiration, and the tongue. After birth, the Heart will also store the emotions of shock, nervousness and excitement. Eliminating excess nervousness allows love, forgiveness, and joy to be experienced. The environment is then conducive for peace, contentment, and orderliness, which allow self-esteem to grow.
3. The Earth Agent (the Virtue of Trust) represents faith, honesty, openness, acceptance, and truthfulness. This agent is connected to the Yi-Intention (thoughts and ideas) and stored in the Spleen, affecting the energetic channels' flow, large muscles, lymph and saliva secretions, mouth, lips, and taste. After birth, the Spleen will also store the emotions of worry, remorse, regret, obsessiveness, and self-doubt. Eliminating excess worry allows trust and peace of mind to exist.
4. The Metal Agent (the Virtue of Integrity) represents righteousness and dignity. This agent is connected to the Po-Seven Corporeal Souls (the material aspect of the spirit) and stored in the Lungs, affecting the energetic channels' flow, skin and mucous membranes, body hair, nose, and the sense of smell. After birth, the Lungs will also store the emotions of grief, sorrow, anxiety, sadness, shame, disappointment, and guilt. Once excess sorrow is relieved, a deeper sense of righteousness, integrity, dignity, and social responsibility exists.

The Five Elements	Congenital Agents	Acquired Emotions
Wood	Kindness	Anger
Fire	Order	Joy
Earth	Trust	Worry
Metal	Integrity	Grief
Water	Wisdom	Fear

Figure 2.13. The Five Agents

5. The Water Agent (the Virtue of Wisdom) represents rationality, clear perception, and self-understanding. This agent is connected to the Zhi-Will (mental drive and determination) and stored in the Kidneys, affecting the energetic channels' flow, brain, inner ear, hearing, spinal cord, cerebrospinal fluid, bones, bone marrow, ovaries, testes, head and pubic hair, anus and urethra, and sexual fluids. After birth, the Kidneys will also store the emotions of fear, loneliness, and insecurity. Once excess fear is eliminated, the mind becomes rational and, therefore, wise.

The energy of the Three Ethereal Souls (Hun) is composed of three separate parts and flows with the Blood as it moves. The energy of the Corporeal Souls (Po) is composed of seven separate parts and follows the Jing as it moves. The Hun and the Po are characterized by their movements. The Po are also referred to as the supernatural Ling (A Moving Spirit). The Hun are also referred to as the supernatural Shen (Mind / Spirit). The Shen is the "active impulse" and Ling is the "active, enabling mover." As the Jing and Blood combine, the fetus's Shen continues to be created.

At the end of the third month, the internal organs, limbs, and external sex organs of the fetus are fully formed, and the nails have developed.

MONTH FOUR

During the fourth lunar month, the mother's Triple Burner Channels, which are connected with the Yang organs, stabilize the fetus's blood vessels (Figure 2.14). The Water Jing is beginning to be accepted by the fetus's body allowing the Yin organs to develop normally.

Beginning in the fourth month and continuing throughout the ninth month, each of the Five Element's energetic nature and the specific characteristics of each element's Essence will be progressively activated and developed within the fetus's body. The first to enter is the Water Jing.

The Water Jing energy supervises the genetic developmental phase of the fetal growth. This energy encompass the fetus's unconscious reservoir of innate and intuitive intelligence, will, and life-force energy, relating to divine love, power,

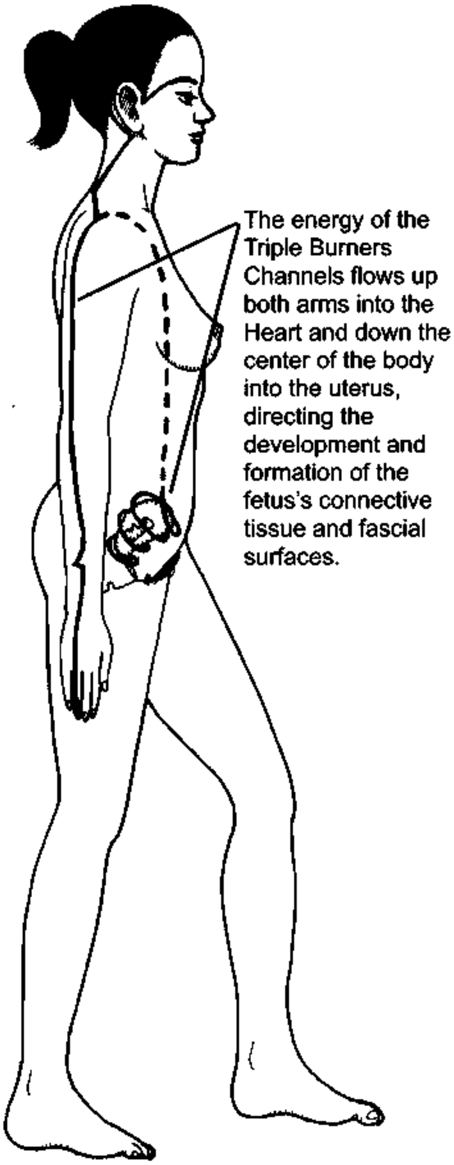
The energy of the Triple Burners Channels flows up both arms into the Heart and down the center of the body into the uterus, directing the development and formation of the fetus's connective tissue and fascial surfaces.

Figure 2.14. The mother's Triple Burners Channels are responsible for the changes in the fourth lunar month of creation. The Water Jing (Essence) is beginning to be accepted by the fetus.

and spirit. Any faltering of this energy (due to the influence of fetal toxins) is associated with both pervasive and subtle neurological disorders, and a predisposition to severe psychological disorders (e.g., schizophrenia).

As the fetus develops, the mother's Triple Burners will direct the development of the connective tissues and fascial surfaces. This process is referred to as "the development of Blood and Qi penetrating to the ears and eyes and circulating throughout the fetus's channels and connecting vessels."

Through the later stages of fetal development the embryo's seemingly homogenous tissues transform into the fetus's differentiated tissues of muscles, bones, and organs (an important part of which is connective tissues). A very large part of the body consists of connective tissues and membranes functioning in such a way as to hold together and maintain the body's external and internal structures.

From a gross anatomical level, the body's structures connect not only through the fascial planes but microscopically through the connective tissues. This internal network facilitates the body's intercellular communication. These structures, both energetic and physical, connect the fetus's body, forming a vast reservoir capable of regulating and transferring the body's Jing-Essence, Qi-Energy, and Shen-Spirit.

MONTH FIVE

During the fifth lunar month, the Spleen Channels become responsible for completing the development of the four limbs (Figure 2.15). The fetus begins its own respiratory movement along with the mother's respiration. The Fire Jing is accepted into the fetus, creating Internal Qi, that stabilizes the fetus's Five Yin Organs' Qi.

The Fire Jing energy generates and controls, protects and integrates, divides and harmonizes the fetus's internal energies to promote emotional/spiritual well-being. Any faltering of the Fire Jing energy is associated with problems of right (Yin) and left (Yang) brain communication (e.g., the correct balance of male/rational and female/intuitive energies).

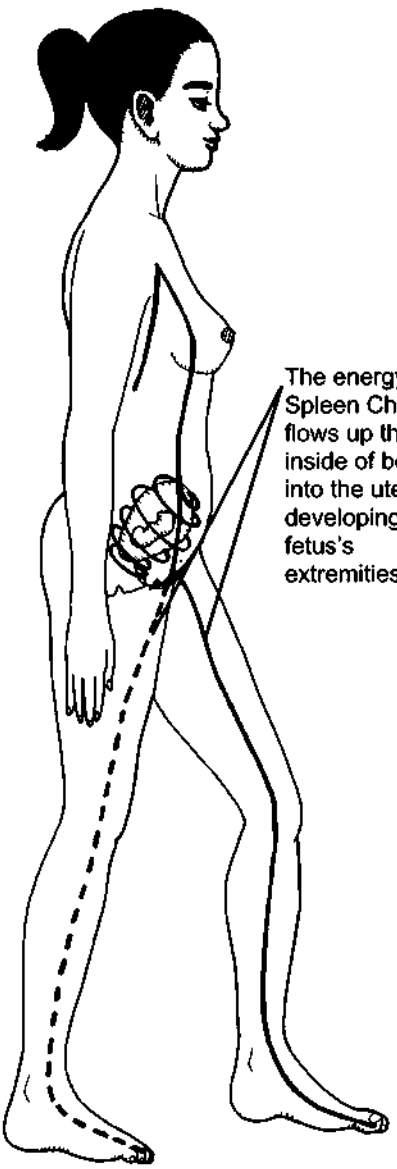

The energy of the Spleen Channels flows up the inside of both legs into the uterus, developing the fetus's extremities.

Figure 2.15. The mother's Spleen Channels are responsible for the fifth lunar month of creation. The development of the fetus's four limbs is completed. The Five Agents are distributed within the fetus's Five Orbs (Five Yin Organs). The Fire Jing is beginning to be accepted by the fetus.

At this stage of development, the Five Agents are distributed to the Five Orbs. These Orbs are energetic spheres of influence or energetic spatial cavities (i.e., internal organ tissue chambers). These Orbs pertain to the Essence, Blood, Qi pools and channels of the Five Yin Organs. The Five Orbs also pertain to and encompass all of the areas in the body that the Five Yin Organs influence. The Five Agents are distributed as follows: Kindness to the Liver, Order to the Heart, Trust to the Spleen, Integrity to the Lungs, and Wisdom to the Kidneys. This distribution causes the Hun to stabilize within the fetus's organs, which contributes to keeping the fetus's Shen at peace.

At the end of the fifth month the fetus's body systems develop rapidly. Its head is less disproportionate to the rest of the body, and its spontaneous muscular movements are commonly felt by the mother.

MONTH SIX

In the sixth lunar month of creation, the mother's Stomach Channels create the fetus's muscles. The Metal Jing is established in the fetus's body, stabilizing the sinews and connective tissues (Figure 2.16).

The Metal Jing energy is also responsible for fetal formation and the ability to form and maintain emotional bonding with others. Any faltering of the Metal Jing energy is associated with problems of emotional attachment (e.g., autism).

During the sixth month the Yang organs are in the process of developing. The Essence of the Yang organs receives, moves, transforms, digests, and excretes substances. The Six Storage Areas of the body's Yang organs constantly fill and empty, and include the Bladder, Gall Bladder, Stomach, Large Intestine, Small Intestine, and Triple Burners.

1. The Bladder controls, stores, and releases the urine. This Yang organ is responsible for receiving, storing, and releasing urine at the proper time.
2. The Gall Bladder stores and secretes bile. This Yang organ is responsible for storing and releasing bile at the proper time.
3. The Stomach stores the food. This Yang or-

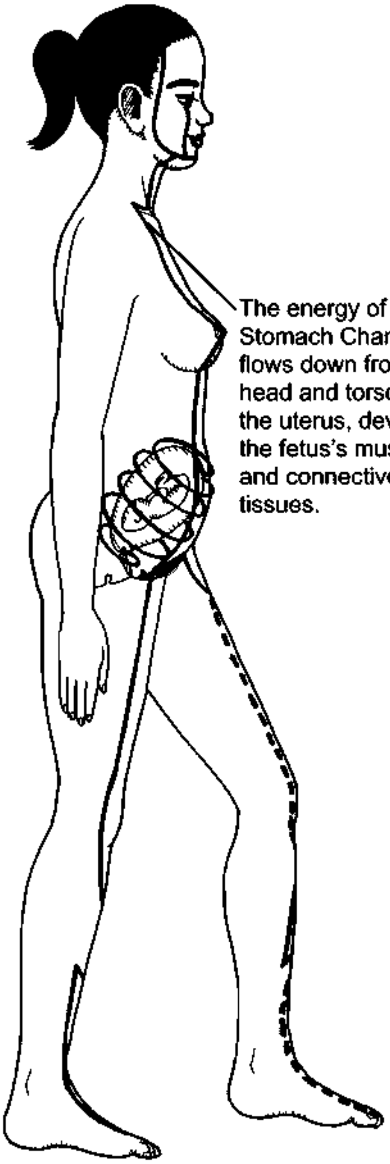

The energy of the Stomach Channels flows down from the head and torso into the uterus, developing the fetus's muscles and connective tissues.

Figure 2.16. The mother's Stomach Channels are responsible for the sixth lunar month of creation. The Six Pitches are established within the body's storage areas. The Metal Jing is beginning to be accepted by the fetus's body.

gan is responsible for the receiving, storing, rotting, and ripening of food at the proper time.

4. The Large Intestine stores the solid waste. This Yang organ receives, stores and absorbs food and releases waste at the proper time.
5. The Small Intestine stores and transforms liquid matter. This Yang organ receives, stores, transforms, and digests food, and releases its waste products at the proper time.
6. The Triple Burners store the body's Qi. This area of the body receives, stores, absorbs, and moves Qi.

At this stage of development, the Six Pitches which support and stabilize the Lower Burners (see Chapter 6) and nourish the Qi are established in the organs known as the Six Yang Organs, also called the Six Storage Areas.

The Six Pitches are six specific tone resonances (notes) that vibrate within the body's internal organs and stimulates specific organ and tissue areas. These Six Pitches relate to the Five Prenatal (Heaven) Elemental Sounds of Jue-Wood, Zhi-Fire, Gong-Earth, Shang-Metal, and Yu-Water. The sixth note Xi relates to the Postnatal (Earth) Fire Element and corresponds to the Pericardium and Triple Burners. The ancient Chinese used these Six Pitches for specific clinical treatments. The sound "Yu" for example, spoken in a low tone, will vibrate the lower abdominal area and is used for the treatment of Kidney and Bladder problems.

In the sixth lunar month of formation, the fetus's eyelids separate and eyelashes form, and its skin is wrinkled.

MONTH SEVEN

During the seventh lunar month, the mother's Lung Channels create the bones, skin, and hair (Figure 2.17). The Wood Jing is beginning to be accepted by the fetus's body.

The Wood Jing energy supervises the assertion and direction of the fetus's emotional/spiritual aspects. Any faltering of the Wood Jing energy is associated with severe psychological problems (e.g., passive-aggressive personality disorder).

In the seventh lunar month, the fetus's Stom-

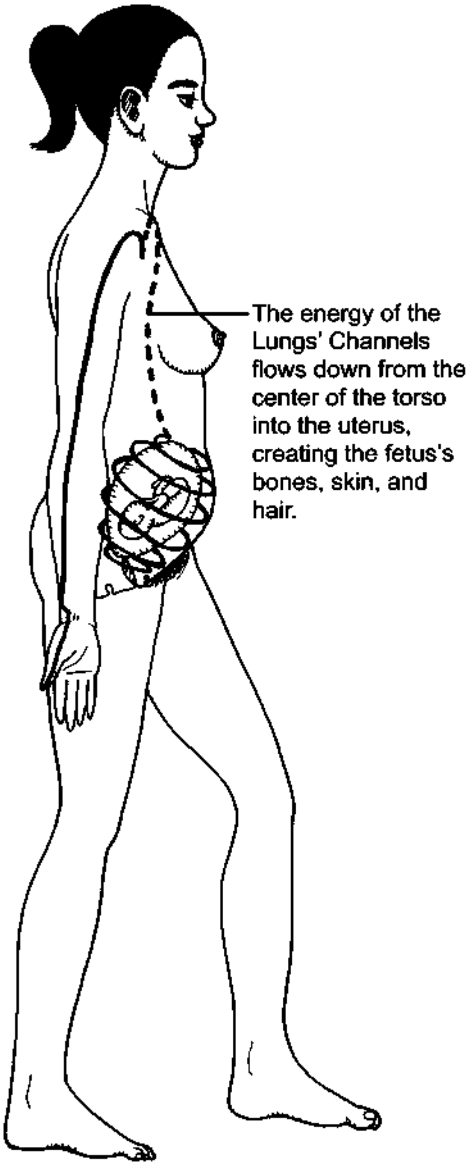
The energy of the Lungs' Channels flows down from the center of the torso into the uterus, creating the fetus's bones, skin, and hair.

Figure 2.17. The mother's Lung Channels are responsible for the seventh lunar month of creation. The Seven Essential Stars open the orifices to let in the light from Heaven and Earth. The Wood Jing is beginning to be accepted by the fetus's body.

ach and Intestines are stabilized, and the Seven Essential Stars open the body's orifices to absorb the light from Heaven and Earth. These stars consist of the sun, moon, and five planets: Mars, Venus, Mercury, Saturn, and Jupiter. Each star is associated with one of the body's orifices: eyes, ears, nostrils, mouth, anus, and urethra. These orifices serve as receiving and projecting energetic portals for Jing-Essence, Qi-Energy, and Shen-Spirit. They also serve as messengers of the body's Five Yin Organs. Thus the Liver receives messages through observation, the Heart through speech, the Spleen through taste, the Lungs through smell, and the Kidneys through hearing.

These energetic messages are received by the body's Wu Jing Shen and emotionally and energetically interact through the body's Six Openings.

1. The eyes absorb images into the Liver which affect the Hun (The Three Ethereal Souls).
2. The ears absorb sounds into the Kidneys which affect the Zhi (Will Power).
3. The nose absorbs smells into the Lungs which affect the Po (The Seven Corporeal Souls).
4. The tongue absorbs tastes into the Spleen which affect the Yi (Intent/Intellect).
5. The physical body absorbs sensations into the tissues which affect the Shen.
6. The spirit (along with the physical body) absorbs sensations into the Heart which also affect the Shen (Spirit).

Physically, during this lunar month, there is a substantial increase in the fetus's weight and its head and body are more proportionate. The fetus can survive if born prematurely (between 27 and 28 weeks); however, its hypothalamic temperature regulation and the Lungs' production of surfactant (a phospholipid substance important in controlling the surface tension of the air-liquid emulsion present in the lungs) are still inadequate.

MONTH EIGHT

In the eighth lunar month, the mother's Large Intestine Channels complete the formation of the fetus's skin, harmonizing the Heart (the Shen), and quieting the breathing. As the fetus receives the Zong Qi-Essential Qi (energy collected from Heaven and Earth and accumulated within the

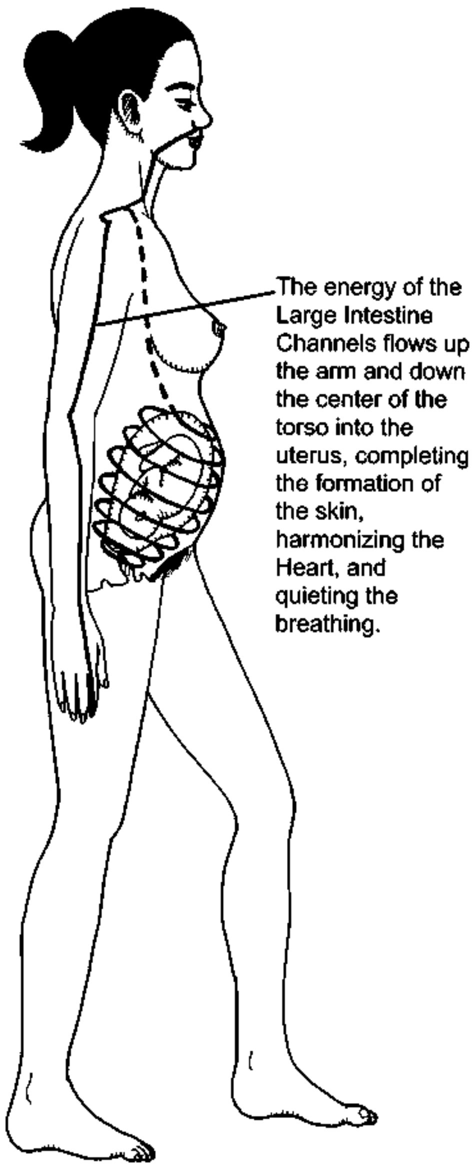
The energy of the Large Intestine Channels flows up the arm and down the center of the torso into the uterus, completing the formation of the skin, harmonizing the Heart, and quieting the breathing.

Figure 2.18. The mother's Large Intestine Channels are responsible for the eighth lunar month of creation. The Earth Jing is beginning to be accepted by the fetus's body.

chest) from the mother's Spleen, the Earth Jing begins to be accepted by the fetus's body, completing the formation of the skin (Figure 2.18)

The Earth Jing energy supervises the quality and maturation of the fetus's emotional and spiritual bonding and boundaries. Any faltering of the Earth Jing energy is associated with problems of severe psychological disturbances (e.g., schizophrenia). These psychological disturbances may be evident at birth or develop later in life.

The mother's Large Intestine Channels control the fetus's orifices. At this stage in development, the fetus's muscles and flesh are formed. The formation and consolidation of the fetus's Jing is now completed, along with the fetus's developing Zhen Qi-True Qi (energy that circulates in the body's channels and collaterals which nourishes the Yin and Yang organs and fights disease).

At the end of the eighth month, the bones of the fetus's head are soft, its skin is less wrinkled, and there is subcutaneous fat deposited throughout its body. If it is a male child, its testes will now descend into the scrotum.

At this stage, the fetus will normally assume an upside-down position to prepare for its birth descent. If the fetus is born prematurely, its chances for survival are now much greater.

MONTH NINE

During the ninth lunar month, the mother's Kidney Channels control the amount of energetic intake the fetus absorbs through its connection to the umbilicus (Figure 2.19). Qi, Blood and food nutrition are absorbed into the fetus's body, after they pass through the umbilical vein to the fetus's Liver. From the Liver, the nutrients are processed and absorbed into the Blood to be distributed throughout the fetus's body.

The flexible structure of the umbilical cord is attached to the placenta. The placenta, which formed on the uterine wall after the first week of pregnancy, consists of tissues from both the mother and the embryo. The function of the umbilical cord is to carry away the baby's waste products and pass food, energy, and oxygen from the mother's blood stream to the embryo. The mother's Kidney Channels will regulate the re-

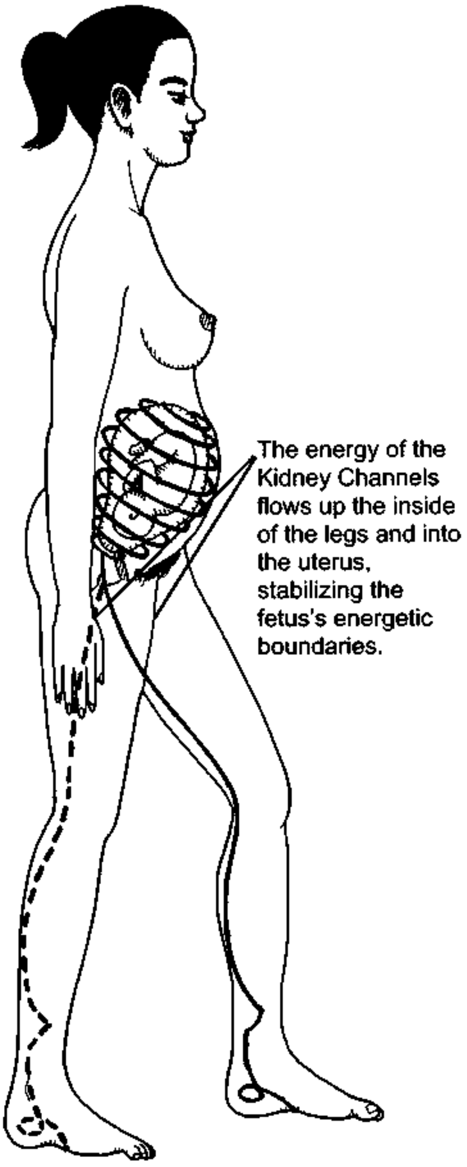
The energy of the Kidney Channels flows up the inside of the legs and into the uterus, stabilizing the fetus's energetic boundaries.

Figure 2.19. The mother's Kidney Channels are responsible for the ninth lunar month of creation. All the fetus's spatial cavities and energetic boundaries are now established.

lease and absorption of Qi and Shen flowing into the fetus's Liver through the umbilical cord.

After the umbilical cord has been severed, the baby's umbilical veins still remain. These umbilical veins eventually become the ligamentum teres that connect from the umbilicus, up along the interior surface of the abdominal wall, through the free margin of the falciform ligament, to the right and left lobes of the Liver. This maintains the baby's connection between its Liver and its Lower Dantian (navel).

At this stage in development, all the fetus's energetic spatial cavities (internal organ tissue chambers) and energetic boundaries are arranged to keep the fetus safely prepared for its birth journey.

Also in this month, the Internal Palaces and Nine Dantian Chambers (the nine internal cavities established within the fetus's Three Dantians (the body's three energetic reservoirs) are arranged and securely established to keep the fetus's Jing-Essence safe.

In the ninth month, additional subcutaneous fat accumulates throughout the fetus's body. Externally, the fetus's fingernails will extend to the tips of the fingers and sometimes beyond.

MONTH TEN

At the tenth lunar month, the mother's Bladder Channels control all Five Yin Organs (Liver, Heart, Spleen, Lungs, and Kidneys) and Five Yang Organs (Gall Bladder, Small Intestine, Stomach, Large Intestine, and Bladder) (Figure 2.20). The baby's energetic reservoirs (the Three Dantians)

Figure 2.21. Heaven and Earth Qi settle into the baby's Lower Dantian and the process of birth begins.

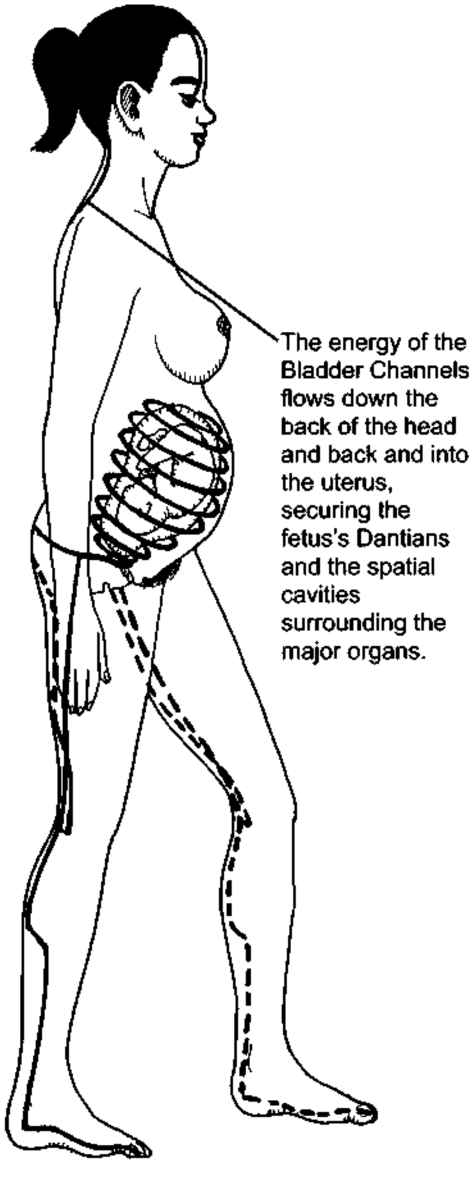
The energy of the Bladder Channels flows down the back of the head and back and into the uterus, securing the fetus's Dantians and the spatial cavities surrounding the major organs.

Figure 2.20. The mother's Bladder Channels are responsible for the tenth lunar month of creation. The baby's Dantians, and spatial cavities surrounding the major organs are completely developed to maintain the safety of the baby's Jing. Heaven and Earth Qi settle into the baby's Lower Dantian and birth begins.

and the internal spatial cavities that surround the major organs are developed to maintain the safety of the baby's Jing. The child is ready to be released into the world. The process of birth begins when Heaven Qi and Earth Qi settle into the baby's Lower Dantian (Figure 2.21).

The energy of the Seven Corporeal Souls (Po) is responsible for the first physiological processes after birth, allowing the child's eyes to see, ears to hear, and Heart to perceive. The Po are also responsible for the movements of the hands and feet and the breathing pattern.

The purpose for the Qigong doctor's study of the ten lunar month developmental process, is to create a foundation for understanding the patient's physical development. This knowledge of the patient's physical development establishes a comprehension of the structural formation of the Five Elemental Constitutions. The Five Elemental Constitutions are described in the next chapter.

POSTNATAL ENERGY DEVELOPMENT

Throughout prenatal development, the Eight Extraordinary Vessels (see Chapter 7) were responsible for transporting, transforming, and producing Qi and Blood for the fetus, while the Twelve Primary Channels (see Chapter 6) were still in the process of gradual development. The focus of energetic activity generated from the fetus's Lower Dantian and Eight Extraordinary Vessels, resonates throughout the fetus's body creating tissue development.

Once the umbilical cord is severed, the primary focus of energy shifts, becoming diverted from the Lower Dantian to the Middle Dantian (located in the baby's chest area). After birth, Gu Qi (food energy) flows into the baby's Stomach while it nurses and begins the movement of Blood and Qi within the Twelve Primary Channels. These Twelve Primary Channels will now assume the responsibility of circulating Qi and Blood throughout the baby's entire physical and energetic structure. The Eight Extraordinary Vessels shift their function to regulating the baby's channel Qi. The Taiji Pole (which has been in the process of dropping from the fetus's Mingmen area to the Huiyin area) is now stabilized in the perineum.

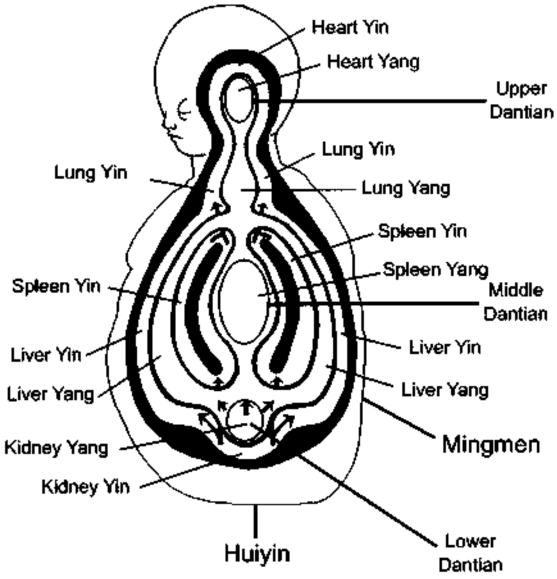

Figure 2.22. As the child develops, the energetic rivers and pools regulate the formation and production of his or her growth.

The natural resonant vibration of the baby's Taiji Pole (energetic core) causes an energetic expansion and contraction, projection and reception that simultaneously affects the baby at five distinct levels: physical, mental, emotional, energetic and spiritual. Balanced physical growth will only occur when all five levels are in equilibrium. The physical body is generally the slowest realm to respond to physical growth. Matter does not develop at the same rate as energy, mind, or spirit. This being the case, the energy, mind, and spirit must wait patiently for the physical body to evolve before progressing as a whole (Figure 2.22).

ENERGETIC EMBRYOLOGICAL OVERVIEW

Understanding the energetic process of fetal development provides the Qigong doctor with an overview of the energetic factors in tissues, organs, and channel function. This understanding is essential for diagnosing the origin and causes of disease as being either congenital or acquired. Unlike Western medicine, Chinese medicine treats the root of the illness, not just its symptoms. In Medical Qigong therapy, cases of congenital insufficiency of Qi or congenital disorders of Qi activities, are treated by either tonifying (strengthening) the Eight Extraordinary Vessels or purging (eliminating) the pathogenic energy from the Eight Extraordinary Vessels. This consolidates and regulates the patient's Prenatal Qi.

For acquired diseases, the patient can only be cured when a selection of points and the methods of emitting Qi for purgation or tonification are determined according to the imbalances of Qi circulation in the Twelve Primary Channels. Both congenital and acquired factors should be considered in every case because both the origination and development of diseases may result from either source, or a combination of both sources.

THE ETERNAL SOUL

In Chinese energetic medicine, the Original Spirit differs from the body's Eternal Soul. The Original Spirit (Yuan Shen) is rooted in the combined Jing, Qi, and Shen of the mother and father. Later the Yuan Shen is integrated and sustained through the embryo's own Five Spiritual Essences or Wu Jing Shen (see Chapter 13), which are created from the energy of the Five Yin Organs. The Eternal Soul, however, is absorbed into the mother's egg at the time of conception, as the sperm enters.

The Eternal Soul is rooted into the Heart and Middle Dantian area by a silver cord, which is additionally rooted within the body's Taiji Pole. When the Eternal Soul becomes rooted within the body's Taiji Pole, it is described as "emitting the spark of the supreme fire" (Light of God), and is considered a divine fragment of God, or the Dao. It is from this divine resonating light that the body's Taiji Pole becomes energized and all Three Dantians become interconnected. The rooting of the Eternal Soul is therefore considered the first emanation from the divine source within the human body.

To travel, and receive spiritual guidance and insight, the Eternal Soul can leave the body in the form of an "Astral Body," departing through the Baihui point at the crown of the head. The silver

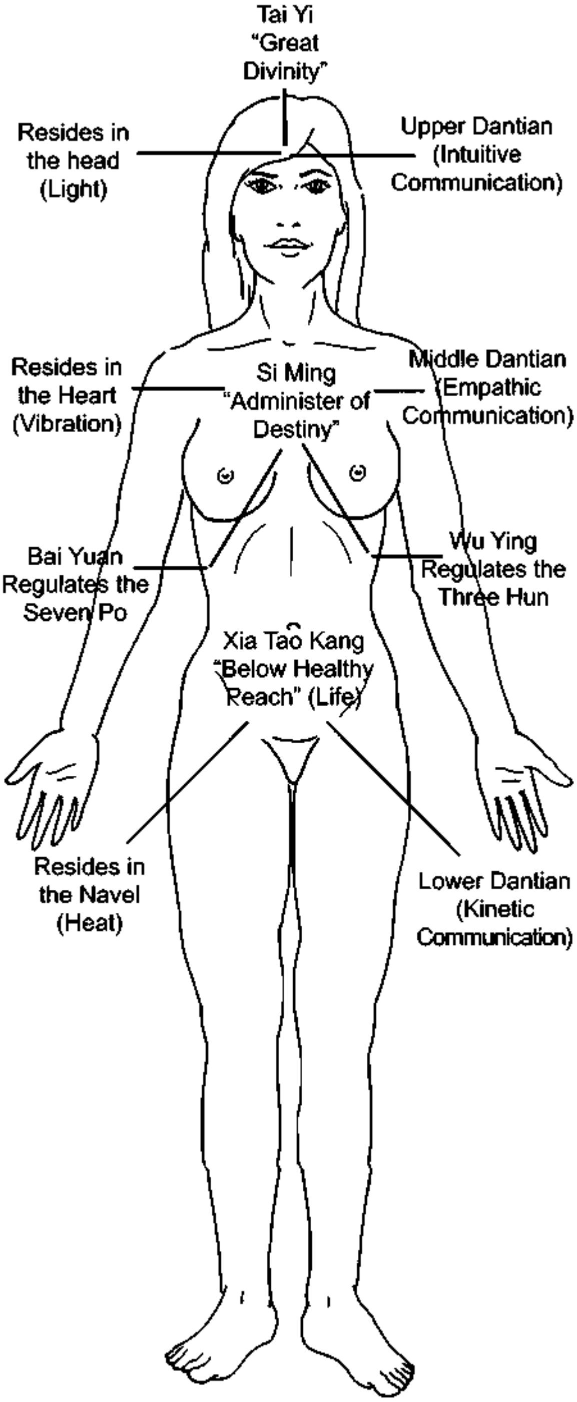
Tai Yi
"Great
Divinity"
Resides in
the head
(Light)
Upper Dantian
(Intuitive
Communication)
Resides in
the Heart
(Vibration)
Si Ming
"Administer of
Destiny"
Middle Dantian
(Empathic
Communication)
Bai Yuan
Regulates the
Seven Po
Wu Ying
Regulates the
Three Hun
Xia Taŏ Kang
"Below Healthy
Reach" (Life)
Resides in
the Navel
(Heat)
Lower Dantian
(Kinetic
Communication)

Figure 2.23 The Lords of the Three Dantians: Tai Yi, Si Ming, and Xia Tao Kang

cord, which is connected to the Astral Body, is located at the fifth and sixth thoracic vertebrae. This physical area is the place on the spine where the silver cord lifts off when an individual astral travels. This area also corresponds to the back of the Heart center, between Ling Tai GV-10 (Spirits Platform) and Shen Dao GV-11 (Spirits Gate). The astral travels of the Eternal Soul are generally to one of the nine higher spiritual planes.

The Eternal Soul becomes the intermediator between the divine, the Original Spirit (Yuan Shen), the Acquired Spirit (Zhi Shen), and the body's Soul Extensions (various traits and characteristics). These Soul Extensions manifest in the energetic field, and influence behavior and perception.

Once the Eternal Soul has established its residence, it separates its Yin and Yang spiritual energy into three spiritual energies which are the energizing forces of the body's Three Dantians, and are called Tai Yi, Si Ming, and Xia Tao Kang. These three spiritual energies are referred to as the Lords of the Three Dantians, and are named according to their function: Tai Yi translates to mean Great Divinity, Si Ming translates as The Administrator of Destiny, and Xia Tao Kang translates as Below Healthy Peach (Life).

The Chinese medical terminology, which describes the Eternal Soul as consisting of three spiritual energies called the Lords of the Three Dantians, is extremely metaphoric and is used to describe the many energetic aspects of the human soul.

The Eternal Soul radiates through the physical, energetic, and psycho-emotional domains of human existence (Figure 2.23). Whole-body consciousness is the main characteristic of the Eternal Soul. Without the interactive process of the Eternal Soul, energy would have no specific direction and would remain in meaningless activity. The Eternal Soul radiates energy in all directions throughout the body and is responsible for life and health, as well as the formation and growth of the body.

In the scope of energetic manifestations, the Eternal Soul expresses its innate qualities as energetic movement, functioning through the physical form. It is led by the Shen and is connected to all parts of the body. The light of the Eternal Soul is revealed through the light or energy of the eyes. Through the influence of the Eternal Soul, all of the body's energetic processes seek wholeness. The Eternal Soul knows exactly what is needed in every situation for survival and health preservation.

The distinction between the Eternal Soul and the Yuan Shen is that the Eternal Soul, although spiritual in nature and connected to the Wuji (infinite space), is still considered individuated. The Yuan Shen, however, although individual in its conditioned state, can also be universal. The connection between you and others, for example, is considered a spirit connection (kindred spirits uniting). The connection between you and the divine is considered a soul connection (your true innate nature reconnects with its origin), and relates only to the connection between one's self and the divine.

When the Qigong doctor meditates, the Eternal Soul consciously enters a spiritual realm wherein the doctor perceives the ideal of things rather than the things themselves. The doctor no longer depends on the senses, but upon a clear inner-vision that perceives the whole picture. It is here, within this spiritual state, that the higher knowledge of things unfolds through divine intelligence and all things reveal their true nature.

When conserved and amplified through practice, the energy of the doctor's Eternal Soul reveals a bright radiant quality, which permeates and penetrates the cells, tissues, organs, and external energy fields. These energetic fields create for the Qigong doctor a strong luminous shield of energy, capable of guarding the body against an attack of hostile energy by enabling it to:

- ward off the invasion of External pathogens and harmful electromagnetic fields,
- ward off the negative influences of malevolent spirits,
- ward off ill intentions and spiritual malice projected unconsciously or deliberately by others, and
- ward off the parasitic influences of demonic forces.

At conception, when the sperm enters the ovum, a polar axis that forms the embryo's Taiji Pole is created. The energetic vortex thus created draws spiritual energies into the body from Heaven and Earth. After birth, the degree to which an individual can draw upon, and absorb, spiritual energy depends on his or her karma (previous thoughts, actions and level of consciousness), and course of destiny.

During conception, three spiritual energies are given to each individual from the divine to create and maintain the existence of the Eternal Soul. These three spirit energies, called the Lords of the Three Dantians, reside in the innermost subtle aspect of the body and are described as follows:

1. The Tai Yi (Great Divinity) resides in the head and Upper Dantian, and is considered the Lord of the Nei Wan (Innermost Palace). It governs a multitude of the body's spirits, causing man's Yuan Shen to shine externally. The Tai Yi facilitates awareness of the Three Ethereal Souls (Hun), and advocates for personal spiritual enlightenment.
2. The Si Ming (The Administrator of Destiny) resides in the Heart and Middle Dantian, regulates the body's Qi, and is the source of the mind and its emotional connections. The Si Ming is considered the Eternal Soul (some believe that this soul reincarnates). The Si Ming controls the spirits Wu Ying and Bai Yuan, that directly affect the body's Jing, Qi, and Shen. The Si Ming challenges our reactions to various internal and external obstacles.
    a. The Spirit Wu Ying (Without Excess) occupies the left side of the body, and regulates man's Three Ethereal Souls (Hun), which are called:
        (1) Tai Guang (Eminent Light),
        (2) Shang Ling (Pleasant Soul) and the Five Agents, and
        (3) Yu Jing (Hidden Essence).
    b. The Spirit Bai Yuan (Pure Origin) occupies the right side of the body, regulates the Seven Corporeal Souls (Po), which are called:
        (1) Flying Poison
        (2) Unclean Evil
        (3) Stinking Lungs
        (4) Corpse Dog
        (5) Fallen Arrow
        (6) Yin Bird
        (7) Devouring Robber

Both the Hun (that come from Heaven) and the Po (that come from Earth) are established within the fetus's internal organs at conception but lie dormant until the third month. This is because, at the third lunar month, the fetus's Orbs are sufficiently formed. One belief maintains that the Hun and Po, although residing in the energetic form of the fetus, frequently leave and return to gather and absorb universal and environmental Qi. The Hun will connect with divine spiritual beings, as well as with God. The Po will connect with surrounding environmental spirits.

Note: The Three Ethereal Souls are traditionally referred to as the "Hun" with the assumption that this term depicts all three Hun, thus in Traditional Chinese Medicine the Hun are sometimes referred to in the singular case. The Seven Corporeal Souls are traditionally referred to as the "Po" with the assumption that this term depicts all seven Po, thus in Traditional Chinese Medicine the Po are sometimes also referred to in the singular case.

3. The Xia Tao Kang (Below Healthy Peach/Life) resides in the navel and Lower Dantian and preserves the root of the body's Jing-Essence.

Upon the death of the body, the Hun return to Heaven, the Po return to Earth, and the spiritual energies of Tai Yi and Xia Tao Kang combine with the Eternal Soul stored in the Si Ming (Administrator of Destiny). These three spiritual energies (Tai Yi, Xia Tao Kang, and Si Ming) blend together, combining into one energy that completes the integration of the Eternal Soul. This soul leaves the body and returns back through the tunnel of light to the divine light for judgement, or wanders the Earth becoming a Gui (ghost).

In Western culture, our internal dialogues are sometimes associated with encounters with good or evil spiritual influences. In Chinese medicine the internal dialogues that influence the patient's psyche are also divided into good and evil; these spiritual/emotional influences are considered to be specific good and evil characteristics in man's Hun and Po. The spiritual components and influences of the Hun and Po are considered to be separate entities, or archetypes, acting upon man's Shen, which in turn affects the Eternal Soul.

These archetypes are regarded as spirit souls that can exert a positive or negative influence on a individual's life depending on the nature of the individual's Eternal Soul. The good internal influences manifest through the Hun. The evil internal influences manifest through the Po. The Hun or Po can motivate personal growth, or can hinder it and cause illness, and even the demise of the body.

THE THREE ETHEREAL SOULS (HUN)

The Three Ethereal Souls are the spiritual part of man that ascends to Heaven upon the death of the body. The Three Ethereal Souls are composed of the Three Hun, that originate from Heaven, reside in the Liver, and resonate from the Three Dantians. The Hun have the following associations: light, Yang, Heavenly soul, Shen, as well as positive emotions and feelings (Figure 2.24). The Three Ethereal Souls are the Tai Guang, Shang Ling, and Yu Jing. They are described as follows:

1. The Tai Guang resonates within the Upper Dantian, and is situated in the cranial cavity, just below the Baihui GV-20 (Meeting of Yin) point. This Hun's name means "eminent light," and it is considered the ultimate balance of pure Yin and Yang energy in harmony. It is connected energetically with the Upper Dantian and Heaven, and always strives for physical, mental, emotional, and spiritual purity.
2. The Shang Ling resonates with the Middle Dantian and is situated in the Heart and corresponding vessels and is linked to the body's Five Agents. This Hun's name translates to "pleasant soul," and it is considered changed (or transformed) Yin energy. It is connected with the Middle Dantian and is a soul that is concerned for others. It is associated with the Five Agents, and produces our desire to be involved in a diversity of social interests and responsibilities.
3. The Yu Jing resonates with the Lower Dantian. Translated, the name means "hidden essence." This Hun is considered mixed (or combined) Yin energy. It is connected with the Lower Dantian and is associated with the Earth, producing our desire for enjoying life's pleasures and comforts, as well as the pure passions of life.

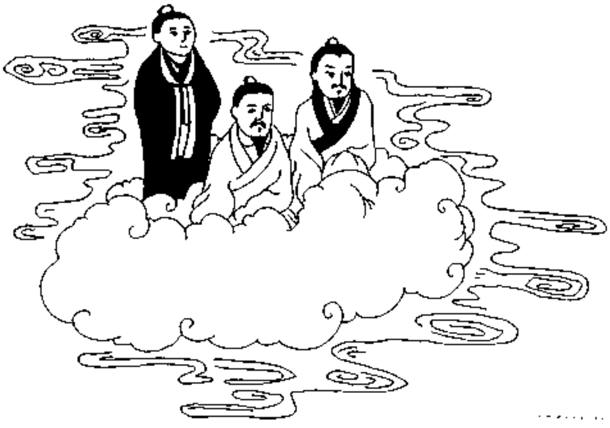

Figure 2.24. The Three Ethereal Souls are composed of the three Hun, that originate from Heaven and reside in the Liver. The Hun represent spiritual consciousness, provide the energetic movement of the mind, and are associated with Heaven Qi and the Five Agents. The Hun are the spiritual part of man which ascends to Heaven after death. The Hun are associated with positive emotions and feelings, light, Yang, Heaven, and Shen.

The Three Ethereal Souls are named (1) Tai Kuang, (2) Shang Ling, and (3) Yu Jing.

The Hun are classified as Yang spirits and are to be cultivated and refined. Imagination, visualization, and positive affirmation in the form of prayer and meditation are needed to awaken and establish an active relationship with the Three Ethereal Souls.

The Hun can be accessed through the Hunmen Bl-47 (Gate to the Hun) point on the back of the body (below the shoulders). It is used clinically to spread Liver Qi and harmonize the Middle

Burner. The Hun respond to Heaven's energetic grids (universal energetic fields). The stars and planets within these Heavenly grids exert an influence on the Hun causing each individual's body to react to certain astrological configurations. The positive or negative reaction is based on the affinity of the vibrational rate of the Hun and the energies of a particular astrological alignment.

THE SEVEN CORPOREAL SOULS (PO)

The Seven Corporeal Souls are closely linked to our body's Jing-Essence. The Po manifest the body's Essence in the form of hearing, sight, and tactile sensations.

The Po pertain to the animal nature of man's instincts and drives and are considered inferior souls to the Hun (Figure 2.25). The Po are passionate and advocate experiencing life in its fullest measure. They are, however, in a constant state of dying. The Po will return to the Earth with the body after death. The Po have the following associations: heavy, Yin, Earthly Spirit, Essence, and negative emotions and feelings.

The Po are a composite of Seven Corporeal Souls, which originate from Earth, reside in the Lungs, and resonate from specific areas in the body. The Po can be accessed through the Pohu Bl-42 (Door to the Po) point on the back of the body (between the shoulders). This point is used clinically to treat energetic imbalances of the Lungs. The Po respond to the Earth's energetic grids (see Chapter 10) and cause each of us to resonate in harmony or disharmony to certain ecological configurations. An individual's attraction, or feeling of not belonging, depends on the vibratory affinity, or lack thereof, between the body's Po and the environmental energy of a particular area.

When the fetus begins its movement, its Yin energy tranquilizes the Po, which act as guardians of the fetus's body. The Po's nature is one of survival, and their energies can be directed towards self-preservation or self-destruction (devouring and robbing the body of life-force energy).

Figure 2.25. The Seven Corporeal Souls are composed of the Seven Po, which originate from Earth and reside in the Lungs. The Po are regarded as the animal nature or drive. They provide energetic movement to the body's Essences, and are considered inferior Souls. The Po will return to the Earth with the body after death, and are associated with negative emotions and feelings, heaviness, Yin, Earth, and Jing.

The negative thoughts and emotions of the Seven Corporeal Souls are named (1) Flying Poison, (2) Unclean Evil, (3) Stinking Lungs, (4) Corpse Dog, (5) Fallen Arrow, (6) Yin Bird, (7) Devouring Robber.

The Seven Po are located along the line between the Huiyin CV-1 and the Baihui GV-20 points, embracing the body's Taiji Pole (also called the Center Thrusting Channel) at the following locations:

1. The Po of Essence (Soul of the Five Element Earth) is located at the midpoint of the Taiji Pole. This Po is the counterpart to the Hun's Earth Agent (the Virtue of Trust). This Po manifests through the acquired emotions of worry, regret, remorse, obsessiveness, and self-doubt.
2. The Po of Qi (Soul of the Five Element Wood) is located below the diaphragm. This Po is the counterpart to the Hun's Wood Agent (the Virtue of Kindness). This Po manifests through the acquired emotions of anger, irritability, blame, rage, resentment, and jealousy.
3. The Po of the Spirit (Soul of the Five Element Metal) is located posterior of the Tanzhong (CV-17) point, inside the mediastinum, near the Middle Dantian. This Po is the counterpart to the Hun's Metal Agent (the Virtue of Integrity). This Po manifests through the acquired emotions of grief, anxiety, sadness, shame, disappointment and guilt.

4. The Po of Yin (Soul of the Five Element Water) is located posterior of the navel, in front of the Taiji Pole. This Po is the counterpart to the Hun's Water Agent (the Virtue of Wisdom). This Po manifests through the acquired emotions of fear, loneliness, and insecurity.
5. The Po of Yang (Soul of the Five Element Fire) is located anterior of the Mingmen (GV-4) point behind the Taiji Pole. This Po is the counterpart to the Hun's Fire Agent (the Virtue of Order). This Po manifests through the acquired emotions of nervousness, shock and excitement.
6. The Po of Sex (Soul of Heaven) is located below the Baihui (GV-20) point in the Ni Wan (Upper Dantian) area.
7. The Po of Life (Soul of Earth) is the only one not located on the Taiji Pole, but is located at the bottom of the feet in the Yongquan (Kd-1) points and is considered a neighbor of the Soul of Essence.

The Seven Corporeal Souls (Po) are also called the "seven animals," or "sentient souls of the body," and "seven turbid demons." When afflicted or restless, the animal nature of the Seven Po quickly becomes hostile, and their names change to express the different negative thoughts and emotions expressed from each Po soul. Each Po has a favorite mantra that it whispers inside a person's mind to influence the individual. The Po's Seven Turbid Demon Natures are described as follows.

1. The Flying Poison (or Quick Evil) suddenly explodes with rage and venomous thoughts of evil intention (e.g., "I'll kill you for that!" "How dare they!").
2. The Unclean Evil (or Shame) entices by tempting and luring an individual into a place or situation from which escape is difficult, and then creates distress in the form of guilt, which generates shame. This spirit creates a feeling of being discredited, dishonored, or disgraced (e.g., "I'm worthless;" "No one has or will ever love me.").
3. The Stinking Lungs (or The Smell of Death) destroys hope, which can lead to a sense of despair (e.g., "I just want to die;" "Just kill me and get it over with!").
4. The Corpse Dog (or Being Scorned) is like a dog that has been beaten and starved, the presence of this spirit intensifies depression by despising itself and holding itself in contempt to a point of no return (e.g., "I'm not worth keeping alive;" "Leave me alone to die!").
5. The Fallen Arrow generates the foregone conclusion that attempting goals is fruitless (e.g., "I'll never be able to accomplish that." "Why even try? I'll never make it.").
6. The Yin Bird (or Night Tormentor) harasses the individual, causing him or her to experience extreme pain and severe anguish (e.g., nightmares and restless sleep).
7. The Devouring Robber (or Sipping Thief) steals the individual's life-force energy by devouring it through negative emotions such as jealousy, envy, and bitterness (e.g., "You love him/her more than you do me;" "I'll never forgive you!").

The Seven Po help in developing the growth of consciousness by providing obstacles to test the individual's faith and devotion. Many times the Seven Corporeal Souls will work in conjunction with outside spiritually hostile forces to test the individual's spiritual endurance.

After the child has been born, enlightened parents can assist the child in facing his or her own internal demons. Once the child is self-aware, he or she can be taught to restrain and control the Seven Po so as to prevent them from taking control. To minimize the effects of these souls on the body, one can learn special meditations, involving visualization, imagination, positive affirmation, and mantras (literally "protections of the mind" or "mind protectors").

A mantra is a word, phrase, or sound, repeated silently or aloud, that one can consciously use as a tool to guide one's own mental, physical, and emotional states. People naturally repeat both positive (self-affirming) mantras and negative (self-destructive) mantras as part of their internal self-talk. When positive mantras are used by an

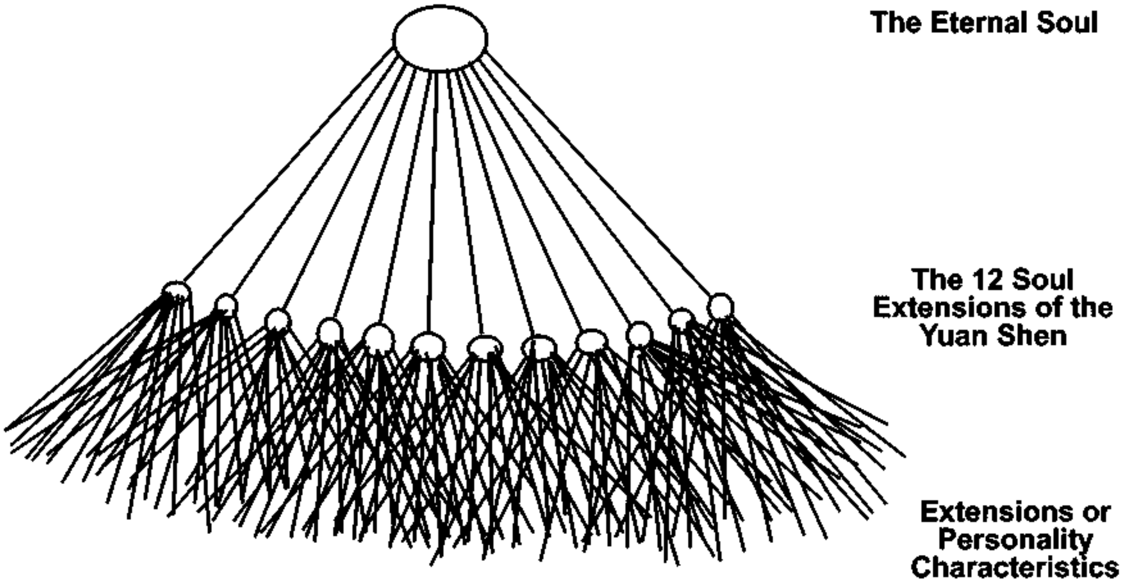
The Eternal Soul
The 12 Soul Extensions of the Yuan Shen
Extensions or Personality Characteristics

Figure 2.26. The Eternal Soul, Yuan Shen and 144 Soul Extensions

individual who holds them in high regard, they enable that individual to combat the Seven Corporeal Souls' mantras. These mantras also facilitate connecting the individual to the divine. Used consciously, mantras and/or prayers can uplift and maintain the individual's self-esteem.

The Chinese view meditation as a deep spiritual contemplation that results in openness to the divine. They view prayer as energy focused towards the divine. Individuals choose what to focus their attention on; therefore, any chronic detrimental thoughts can be countered with positive affirmations to change self-destructive habits.

THE ORIGINAL SPIRIT (YUAN SHEN)

The Original Spirit (Yuan Shen) radiates and extends from the interior of the Eternal Soul. This spiritual light crystallizes a spiritual body, gradually becoming consciously effective and moving into action. It envelops the body and communicates through innermost thoughts, sensory feelings, and emotions. Although the body's Jing and Qi will degenerate, the Original Spirit is immortal, as it is the expression of the Eternal Soul.

The Shen develops and contains Twelve Soul Extensions. These Twelve Soul Extensions contain the body's different personality characteristics (Figure 2.26). Although Si Ming controls the residence of the Eternal Soul, the body's Shen is free to make decisions that affect life and health based upon the individual's free will (the interaction between the Yuan Shen and the Zhi Shen), which then manifests through the Twelve Soul Extensions.

THE EFFECT OF THE HUN AND PO ON THE YUAN SHEN

The Yuan Shen is a manifestation of the Eternal Soul and is primary to the Hun and Po. The combination of the Seven Corporeal Souls (Yin Souls) and the Three Ethereal Souls (Yang Souls) creates the sustaining energy for the body's Original Spirit (Yuan Shen). The Yuan Shen also controls and organizes the psycho-emotional aspect of the body's Five Yin Organs, called the Wu Jing Shen or Five Essence Spirits.

The ancient Chinese viewed the Shen as an Emperor, seated within the Heart; the Hun are viewed as a Loyal Minister, seated within the Liver; and the Po are viewed as a Violent General, seated within the Lungs. If the General (Po), is left in control (being only concerned with the survival of self and the body), he will start to dominate (a condition referred to as a rebellious General and a weak Emperor). At the point when the General (Po) dominates, the individual's acquired mind takes over and the individual becomes concerned only with his or her own survival, and thus becomes self-absorbed. The goal for spiritual cultivation is to control the rebellious General, making him into a servant. Once the inner government is orderly, the strong and violent nature becomes tame. Then, through the wise council of the Loyal Minister (Hun), the individual can walk a path of virtue. Virtue is the path that leads the individual's Shen so that even his or her human consciousness is dominated by the Hun and the Yuan Shen.

When the Hun control the energy body and are nourished by the virtues, the energy body then becomes a vehicle for the Heart's Shen, which is non-local and non-linear in terms of time and space.

When the Heart's Shen is no longer dominated by the Yin and Yang souls (Hun and Po) and the Five Element energies of the lower energetic plane (Wu Jing Shen), the individual returns to an awareness of his or her connection to the universe (Wuji) and the divine. This state of consciousness is sustained through prayer and meditation.

When beginning to develop the Heart's Shen, the first separation of the individual's energetic spirit body or "dream body" from the physical body generally leaves the Shen in a weakened state (referred to as Yin Shen). The weakened Yin Shen must be protected. This Yin Shen is part of the Hun, and leaves the body naturally whenever the patient is weak, sick, in shock, or asleep. It may also leave during the early stages of Qigong meditation or Taijiquan practice. The astral travels of the Yin Shen are generally confined to the lower spiritual planes, and it sometimes needs to be reclaimed through "soul retrieval" (see Chapter 19). Once the Shen has been cultivated, refined, strengthened, fortified, and controlled through the development of the energy body, it is then referred to as a Yang Shen. When the energy body spiritually travels, the developed Yang Shen can transverse the Nine Levels of Heaven or the Nine Levels of Earth in an instant.

The Nine Levels of Heaven are nine spiritual planes which exist within Heaven's energetic grids. Within the lower levels there exists enlightened beings coexisting with other spiritual entities. The Nine Levels of Earth are nine spiritual planes which exist within Earth's energetic grids. Within these levels of Earthly spiritual planes there exists various dimensions of animal, vegetable and mineral powers, as well as Earthly spiritual beings.

YIN AND YANG ASPECTS OF THE ETERNAL SOUL

The human soul has an endless source of vitality from the infinite high-frequency force of the divine. This divine energy descends from the Heavens into the body's Taiji Pole, taking up residence and interacting on the physical plane. The body serves as a medium of transformation for the Eternal Soul. The soul's activity extends and expresses its essence through the physical form and through consciousness. As we grow and change, the Eternal Soul maintains stability by providing the energetic blueprint for orderly development. On the spiritual plane, when referring to the energetic dynamics of the Eternal Soul, there are two basic movements: (1) Yin - the downward movement and (2) Yang - the upward movement.

1. The Yin aspects of the Eternal Soul spirals downward. The spiralling movement of the energetic soul results in the energy transforming from the highest divine energetic frequencies to the lowest. Within this spiritual transformation process, there is a parallel gradual decrease in levels of consciousness. The lowest frequency creates matter, and is the bases for all the prenatal transformations. At the higher frequencies, the principles by which Qi operates are difficult to perceive because our lower state of consciousness does not resonate in harmony with the information being

Spiritual Energy	Resides in...	Associated Dantian	Purpose	
Tai Yi Great Divinity	the head as light	Upper Dantian (intuitive communication)	Governs the body's Spirits, makes life shine forth, and encourages man to know the spirits of his Ethereal Soul	
Si Ming Administrator of Destiny	the Heart as vibration	Middle Dantian (empathic communication)	The Eternal Soul—regulates the prime energy of life (Qi) and is the source of the emotions and mind Controls the Wu Ying and Bai Yuan Spirits	Wu Ying (Without Essence), occupies the left side of the body and regulates the Three Hun (Ethereal Souls), associated with the Liver
				Bai Yuan (Pure Origin) occupies the right side of the body and regulates the Seven Po (Corporeal Souls), associated with the Lungs
Xia Tao Kang Below Healthy Peach (life)	the navel as heat	Lower Dantian (kinetic communication)	Preserves the root of the body's Essence (Jing)	

Figure 2.27. The patient's soul resides in the body's Taiji Pole, connected with Si Ming, the Administrator of Destiny. Upon the death of the physical body, the Three Hun return to Heaven, the Seven Po return to Earth, and the energy of Tai Yi and Xia Tao Kang envelope the Eternal Soul (stored in Si Ming) becoming one energy. The Eternal Soul leaves the body through the specific gate (solar plexus, third eye, or top of the head) associated with the patient's degree of spiritual evolution, returning back through the tunnel of light to the divine.

presented. As energy slows down its resonance, it is more readily perceived by our five senses; this gives matter the quality of hardness, form, and smell, and allows us to observe its energetic patterning.

2. The Yang aspects of the Eternal Soul spirals upward. The spiraling movement of the energetic soul results in energy transforming from the lowest level of consciousness to the highest divine energetic frequencies during postnatal transformation. Spiritual evolution involves a progressive opening up of all of the body's energy centers (Three Dantians) along the Taiji Pole, until a state of absolute unity of body, mind, emotion, energy, and spirit is achieved (known as enlightenment).

In the evolutionary process of the soul, both stability and change become as one when there is a deep connection to wholeness. This soul serves as a pivotal point for the body's Jing, Qi, Shen, Body Fluids, connective tissues, and emotions, all of which express wholeness. Without this connection the patient experiences a serious psychic splitting of his or her energies. The part that is open to change will experience resistance from the part that demands stability.

THE FINAL EXIT OF THE ETERNAL SOUL

As the body begins to die, the Yuan Shen begins to prepare the Eternal Soul for its final journey home. This change of residence for the patient's spiritual energies is crucial. At the time of death the Eternal Soul must disconnect from the Heart and Middle Dantian area, and the soul must exit the dying patient through one of the Nine Orifices. The lower orifices include the anus and urethra, while the upper orifices include the eyes, ears, nostrils, mouth, as well as the Yintang and Baihui areas (Figure 2.27). It is believed that the area from which the patient's soul exits the body determines the state of its spiritual evolution and to which area the patient is most attached (i.e., if the patient is attached to vision, they are more inclined to leave through the eyes, etc.).

When the Medical Qigong doctor is assisting a terminally ill patient, treatment focuses primarily on purging stagnant spiritual energy from the patient's body. Spiritual purging is required to purify and cleanse the patient's Eternal Soul. Once the patient's soul has been cleansed, the patient becomes peaceful as all material, emotional, and spiritual attachments to this life are released. The patient attains a sense of completion and welcomes the final adventure of going home. The last rites performed by various clergy and ministers in Western culture serve a similar purpose.

The Qigong doctor then guides and encourages the patient's Eternal Soul into leaving the body through the top of the head (Baihui). Souls that are more highly evolved exit the body through this area (at the top of the head), while souls of lesser evolution exit the body through lower portions of the physical structure.

The formation of a Gui or ghost is said to derive from the untransformed energy of the Eternal Soul (see Chapter 19).

SUMMARY

One of the unique aspects of Traditional Chinese Medicine is that it addresses all aspects of the body's energy and spirit. It explores in great detail the spiritual transformations in the body's prenatal state, their relationship to the formation of the body, and to the dying and after-death states. Traditional Chinese Medicine further addresses the potential of spiritual involvement as a cause or cure of specific psychophysical diseases. It is through the complete study of the body's physical, mental, emotional, energetic, and spiritual aspects that doctors of Medical Qigong therapy comprehend when and how to treat their patients.

CHAPTER 3
PHYSICAL DEVELOPMENT AND STRUCTURAL FORMATION

FASCIAL DEVELOPMENT AND ENERGY FLOW

According to the *American Medical Association Encyclopedia of Medicine*, "fascia is the fibrous connective tissue that surrounds many structures in the body. One layer of the tissue, known as the superficial fascia, envelopes the entire body just beneath the skin. Another layer, the deep fascia, encloses muscles, forming a sheath for individual muscles, and also separates them into groups. The deep fascia also holds in place the soft organs, such as the Kidneys. The thick fascia in the palm of the hand and sole of the foot have a cushioning, protective function."

In Traditional Chinese Medicine, the fascia is referred to as Huang (meaning any membranous tissue). Fascial development is divided into two stages: prenatal (congenital) and postnatal (acquired). The prenatal fascia is fixed and determined by the combination of the parents' Jing, Qi, and Shen. The postnatal fascia is formed through diet, exercise, and environmental stresses.

Qi is stored within the tissues and inner fascia layers, where it envelops and protects the internal organs. Energy (that has been stored within the body's organs) is available through the fascia for transformation to nourish the patient's body, mind, emotions, energy, and spirit.

The body consists of three tissue layers:

1. The first and innermost layer is made up of the internal organs, which produce and transform Qi.
2. The second layer consists of the body's fascia, tendons, ligaments, and bones, which assist in transporting Qi to the extremities.
3. The third and outermost layer of tissue consists of the muscles and skin, where Wei (Protective) Qi circulates to protect the organism from invasion of external pathogens (Cold, Heat, Wind, and Dampness, etc.).

After being generated within the body's internal organs, Qi is distributed throughout the body's entire energetic network through the channels and collaterals, via the fascia. Each organ has its own layer of weblike fascia, that covers, connects, protects, and nourishes the tissues. The fascia forms the energetic chambers of the body's organs and channel systems. Qi within the body flows between the fascial sheaths and along the channel system. Through trauma, infection, surgery, disease, or chronic muscular tension, areas of the fascia can become stuck together, inhibiting the flow of Qi and Blood.

Our constitutions are formed, in part, by conditions from the environment that affect the innermost layers of fascial development. Traditional Chinese Medicine divides the observation and diagnosis of the physical developments of these inner fascia into Yin and Yang structures and Five Elemental Constitutions.

YIN AND YANG STRUCTURAL FORMATION

Chinese philosophy teaches that the universe is composed of a pair of opposite forces or energies -- Yin and Yang. Everything in the natural world contains both Yin and Yang. Yin and Yang are interdependent; without Yang, Yin cannot grow; without Yin, Yang cannot develop. The energetic transformation of Yin into Yang and Yang into Yin produces the various observable changes of form and matter (see Chapter 21).

Yin and Yang are the principles governing all things within the human body. Life and death originate from the energies of Yin and Yang, and they are the forces that create all physiological change. In Medical Qigong therapy, the physical structure of the human body is divided into Yin and Yang organs and substances, as well as Yin and Yang energetic functions (e.g., Qi that expands and flows outward is Yang, while Qi that contracts and flows inward is Yin).

Yin and Yang are in a constant state of waxing and waning (Figure 3.1). If this waxing and waning exceeds the body's normal energetic limits and loses its dynamic equilibrium, Deficient or Excess Yin and Yang will occur, leading to the development of abnormalities and illness.

Yin Qi naturally ascends (e.g., Spleen Qi ascends the Clear Qi). When Yin Qi becomes pathological or destructive, it moves downward, descending like cascading water (e.g., descending Spleen Qi causes diarrhea or prolapse of the viscera). Another example of the effects of pathogenic Yin Qi descending would be edema.

Yang Qi naturally descends (e.g., Stomach Qi flows downward). When Yang Qi becomes pathological or destructive, it ascends like the flames of a fire, moving upward (e.g., Rebellious Stomach Qi ascends, causing nausea and vomiting). An-

	Yang	Yin
Attributes	active, creative, masculine, Fire, Hot, Heaven, light	passive, receptive, feminine, Water, Cold, Earth, heavy
Organs	the six hollow organs (Gall Bladder, Small Intestine, Stomach, Large Intestine, Bladder, and Triple Burners)	the five solid organs (Liver, Heart, Spleen, Lungs, and Kidneys), along with the Pericardium
Respiration	exhalation	inhalation
Time of Day	the 12 hour cycle from 11 p.m. (before midnight) to 11 a.m. (before high noon)	the 12 hour cycle from 11 a.m. (before noon) to 11 p.m. (before midnight)
Seasons	spring and summer	autumn and winter
Normal Movement	outward	inward

Figure 3.1 Yin and Yang are the Principles Governing All Things Within the Body

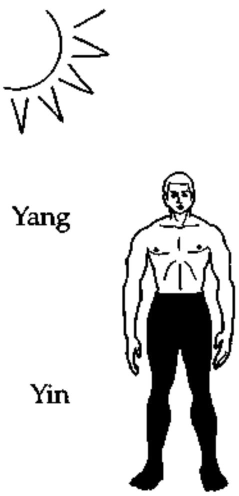

Figure 3.2. The superior aspect of the body's structure is Yang; the inferior portion is Yin. Think of the sun shining on the upper torso, illuminating the body from the waist to the head, while from the hips downward the lower torso is in the shade.

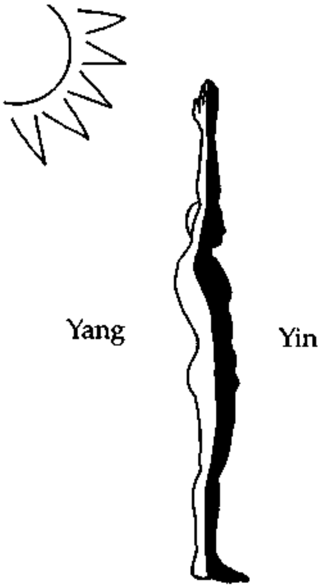

Figure 3.3. The posterior portion (back side) of the body is Yang; the anterior portion (front side) is Yin. Think of the sun shining on the back, while the abdomen is in the shade.

other example of the effects of pathogenic Yang Qi ascending would be the temperature of a rising fever.

THE YIN AND YANG OF BODY ASPECTS

Comparable to Western anatomy, Traditional Chinese Medicine also divides the body into aspects or planes. In TCM, these classifications are divided into Yin and Yang physical planes and anatomical directions. These divisions assist the Qigong doctor in defining and categorizing the patient's external tissues, and in determining the collection and movement of Qi (i.e., energy moving into and coalescing in the patient's Yin or Yang areas of the body). The structural aspects are categorized as follows.

- The cranial or superior aspect (towards the head) of the body's structure is considered Yang; the caudal or inferior portion of the body (towards the feet) is considered Yin (Figure 3.2).
- The posterior or dorsal portion (back side) of the body is Yang; the anterior or ventral portion (front side) is Yin (Figure 3.3).
- The superficial aspect (exterior) of the body is Yang; the deep portion (interior) of the body is Yin (Figure 3.4).
- The left side of the body is Yang; the right side is Yin (Figure 3.5).

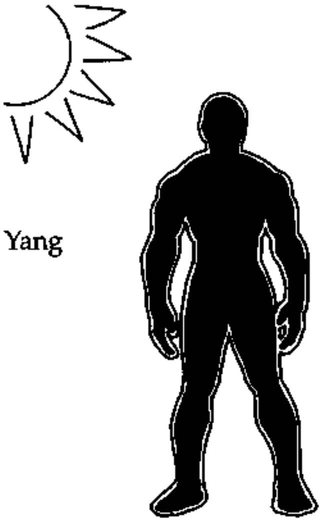

Figure 3.4. The superficial aspect (exterior) of the body is Yang; the deep portion (interior) of the body is Yin. Think of the sun shining on the external surfaces of the body, while the internal organs remain in the dark.

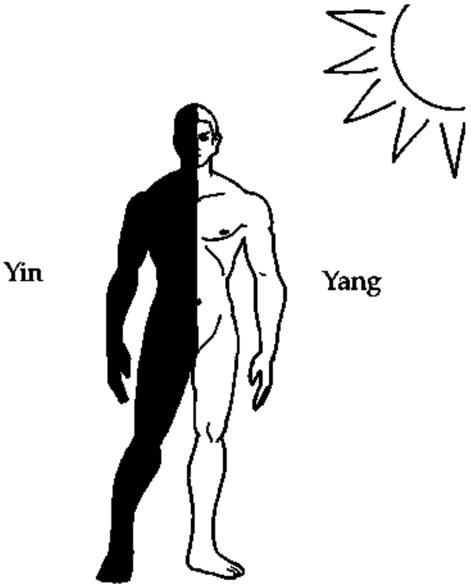

Figure 3.5. The left side of the body is Yang; the right side is Yin. Think of facing south in the morning, while performing medical Qigong practice. The sun will rise in the east, illuminating the left side (Yang) while the right side remains shaded (Yin).

Aspect	Yang	Yin
upper/lower surface	superior (upper)	inferior (lower)
front/rear surfaces	posterior (back)	anterior (front)
position from midline/middle	lateral (away from center)	medial (middle)
near to/far from the surface	superficial (external)	deep (internal)
sides	left	right

Figure 3.6. A Summery Diagram of the Yin and Yang Aspects of the Body

The Seven Emotions			
Chinese Word	English Translation	Organ Association	Qi Manifestation
Xi	joy, happiness, excitement, pleasure	Heart	slows down and relaxes the Qi
Nu	anger, irritation	Liver	causes Qi to rise
Si	meditation, contemplation, pensiveness, worry	Spleen	stagnates the Qi
You	anxiety, sorrow, grief	Lungs	obstructs the Qi
Kong	fear, extreme anxiety	Kidneys	causes Qi to descend
Jing	shock (sudden intense fear), fright	Heart then Kidneys	scatters and deranges the Qi
Bei	grief, affliction, sadness	Lungs	disperses and consumes the Qi

Figure 3.7. The Seven Emotions and Their Effects on the Organs

- The lateral aspect (further from the center) of the body is Yang; the medial portion (middle) is Yin.

The aspects are summarized in Figure 3.6.

As Yin and Yang are in a constant state of change, always waxing and waning, two points of reference alone are not enough to aptly classify the many phases and stages of transformation; for example, day (Yang) can be further split into dawn (waxing Yang), midday (peaking Yang), and late afternoon (waning Yang). To fill this descriptive gap, the Chinese developed the theory of the Five Elements.

THE FIVE ELEMENTS

The Five Elements (Wu Xing), also translated as the Five Phases, are Wood, Fire, Earth, Metal, and Water. These Five Elements paint a clear and natural image of the changing cycles of Yin and Yang (see Chapter 23).

Furthermore, each of the Five Elements can also be divided into Yin and Yang, making a total of ten subdivisions. Wood, for example, can be subdivided into Yin Wood and Yang Wood.

The Chinese use this framework of the Five Elements to understand and categorize many different areas of knowledge, from the movements of the Heavens and the land forms of Earth to the workings of human anatomy, physiology, and psychology. Earlier in this chapter, it was noted that our organs and their associated energy channels are divided into Yin and Yang. In addition, each of the organs is also related to one of the Five Elements. The Liver and Gall Bladder relate to the Wood Element; the Heart and Small Intestine to the Fire Element; the Spleen and Stomach to the Earth Element; the Lungs and Large Intestine to the Metal Element; and the Kidneys and Bladder to the Water Element. The organs help to regulate the functional aspects of the Five Elements within the human body as they interact with each other to promote and maintain life.

The Yin organs, in particular, are said to store the Jing of the Five Elements. They share this Essence with their elementally paired Yang organs to assist with the Yang organs' functions. The Yin organs also store the Five Agents (the Five Virtues), and are more readily disrupted by the Seven Emotions (anger, joy, worry, grief, sadness, fear, and shock) than are the Yang organs.

The emotions are said to be the primal force behind energy transformation; however, each individual will have a separate reaction to specific emotional changes, based on his or her dominant elemental pattern. Five internal organ emotions are associated with the energy of the Five Elements:

1. Anger - Liver - Wood
2. Joy - Heart - Fire
3. Worry - Spleen - Earth
4. Grief and Sorrow - Lungs - Metal
5. Fear - Kidneys - Water

In this way, the energies of the Five Elements shape all aspects of a human being: physical, energetic, mental, emotional and spiritual (Figure 3.7).

CONGENITAL CONSTITUTIONS

Most individuals are born with a dominant element, which will determine the individual's physical, energetic, and psycho-emotional constitution. When in Excess, the body's emotions can create an energetic imbalance, which in turn, can cause a destructive physical pattern within the body. An example of this process is excessive anger leading to hypertension.

The previous chapter on energetic embryology described how, from month four to month eight, the different Five Elemental Jings (Essences) enter the fetus from each of the mother's organs. Depending on the relative strength or weakness of the mother's internal organ energies during her pregnancy, the fetus will inherit a greater or lesser degree of elemental Jing from the mother. If the mother's Liver energy is dominant, then the child will have a tendency to develop a Wood constitution; if the Kidney energy is dominant, then the child develops a Water constitution, and so forth.

In addition, traditional Chinese astrology states that the various Heavenly influences at the time of birth influence the relative Five Elemental strengths and weaknesses in the newborn's constitution. Years, months, days, and hours all occur in separate cycles of five, corresponding to the Five Elements. Thus both the astrological influences at birth as well as the prenatal maternal transmission of Jing, Qi, and Shen, determine each person's congenital constitution.

After birth, the congenital Jing is cultivated and sustained through prayer, meditation, physical exercise, and sleep. In medical Qigong theory, the health of the congenital Jing determines the health and balance of the overall constitution; it determines the level of vitality and resistance to disease, as well as the mental and emotional nature of the individual. The congenital Jing, with its unique balance of the Five Elements, becomes the foundation for an individual's Qi and Shen.

The congenital influence is, however, not absolute. Postnatal factors, such as diet, life-style, environmental factors, and internal belief structures, can also influence the constitution, for better or for worse. Thus, patients with congenital deficiencies can still improve their health by prayer, meditation, medical Qigong, adequate sleep, proper diet and exercise, herbs, medicines, stress management, and other therapeutic modalities.

On the negative side, patients with strong inherited constitutions may develop serious constitutional imbalances through intemperate lifestyles such as: excess stress, overwork, excessive sexual activity, and poor eating habits. Therefore, the Qigong doctor must consider both prenatal (congenital) and postnatal (acquired) constitutional factors to understand and differentiate between the patient's innate constitution and his or her current condition.

CLASSIFICATION OF THE FIVE CONSTITUTIONS

Over the centuries, much has been written in Chinese medical literature about the Five Elemental Constitutions. The following interpretation was developed by Dr. Zhou Qianchuan, a Daoist master from Qi Cheng Shan and reputed by some to be the master of *The Yellow Emperor's Classic of Internal Medicine* (a 2500 year old three-volume text that is still one of the pivotal texts of Chinese medicine).

Dr. Zhou perceived that each of the Five Elemental Constitutions has both a characteristic physical appearance and a psycho-emotional profile. Through observing these aspects, the Qigong doctor determines a patient's Five Elemental Constitution, and from this diagnosis, the doctor discovers much about the balance or imbalance of the patient's physical and mental health.

In Traditional Chinese Medicine, somatic-build type and psyche are closely intertwined. Physical aspects, such as coloring, proportion, dominant tissue (muscle, fat, sinew, etc.), tone, movement, posture and holding patterns, as well as general vitality all express the patient's internal energetic and psycho-emotional matrices. The body type reveals to the trained Qigong doctor not only the patient's emotional history but also his or her innate personality and acquired personality characteristics.

Dr. Zhou subdivided each of the Five Elemental Constitutions into three categories: balanced nature, Yin nature, or Yang nature. The ideal is to have a balanced nature, regardless of one's elemental type. Note that the treatment addresses the mental, emotional, and spiritual balance of the five constitutions. Since the mind is so closely tied with Jing and Qi, it is not enough to simply tell patients how to change their personality to restore balance. The Qigong doctor must work with patients on all levels (physical, energetic, mental, emotional, and spiritual) to help them restore balance. Specific healing methods are discussed at length in later chapters. In general, the Qigong doctor selects treatment principles based on the Five Elements' Creative (Figure 3.8) and Controlling (Figure 3.9) Cycles to help restore balance (see Chapter 23).

For instance, a patient with a Yin Earth nature would tend to have weak or Deficient Spleen energy. Therapeutically, the Qigong doctor would choose one of the three following treatments.

1. Direct tonification of the Earth Element,
2. Tonification of the Fire Element in accordance with the principle, "To nourish the child, strengthen the mother"), or
3. Disperse or reduce the Wood Element (Wood, the "grandmother", may be overcontrolling Earth, the "grandson").

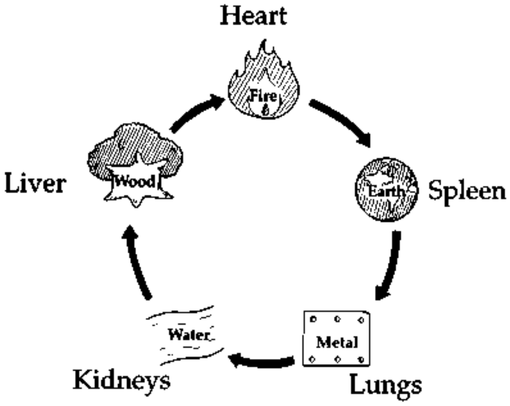

Figure 3.8 The Five Elements' Creative Cycle

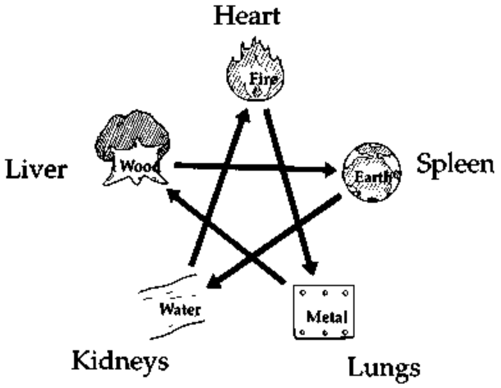

Figure 3.9. The Five Elements' Controlling Cycle

In the instance of Yang Earth nature, the patient would tend to have Excess Earth. In this case, the Qigong doctor could:

1. Use a method to reduce the Earth Element directly; or,
2. Strengthen the Wood Element (the weak "grandmother" may be unable to restrain an overly strong "grandson" without help).

THE WOOD CONSTITUTION

WOOD OUTER APPEARANCE

Wood constitution types usually have a tall and slender body with a small head and a long face. They also have small, nicely formed hands and feet, broad shoulders, and a straight, flat back (Figure 3.10). By nature, they enjoy spring and summer but dislike autumn and winter. This preference is due to the vulnerability of the Wood constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include strong sinews and green-blue facial color. Their note is "jue."

WOOD PERSONALITY

The Wood personality can be divided into three subdivisions.

1. A Balanced Wood nature predisposes individuals to be confident, strong, independent, and intuitive, with a clear understanding of themselves and their goals. They are patient, able to allow things to develop naturally, and express their personality in a relaxed harmonious way. They are kind when communicating with others, creative, free-flowing in self-expression, and display merciful and unselfish traits when dealing with the needs of others.
2. A Yin Wood nature predisposes individuals to feel externally insecure, and to be always cautious. They have a tendency to worry and have a weak sense of their own abilities and potential. They are unassertive and are unsure of their identities and life purpose. They also have difficulty expressing their egos and have weak boundaries. They are timid, lack confidence, and display considerable doubt.

   Treatment for Yin Wood nature individuals involves establishing a sense of inner strength and security by teaching them self-trust and building self-esteem. These individuals need to strengthen their boundaries to avoid the intrusion and domination of others. They need to trust in their intuition, develop a greater degree of confidence, enhance their personal power, and find a stronger sense of spiritual growth.
3. A Yang Wood nature predisposes individuals to manifest their internal insecurity through acting irritable and impatient. Such individuals are intolerant, rude, stubborn, and selfish, and tend to expand their egos without consideration for others. They are domineering, angry, aggressive, and generally known as

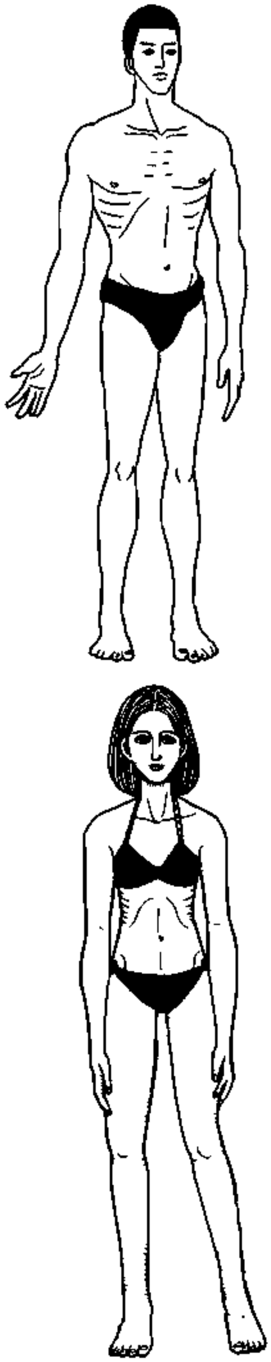

Figure 3.10. The Wood Constitution in Males and Females

overachievers; however, they are easily frustrated and depressed. Their self-esteem is elevated by acquiring higher political (influential) positions or bullying others.

Treatment for Yang Wood nature individuals should focus on cultivating a discipline of inner peace, following the Dao, to harmonize with life. In situations of conflict, they need to learn to relax, slow down, and act out of stillness. They need to surrender to a higher will, relying on their spiritual intuition and inner direction. They also need to learn to respect others.

THE FIRE CONSTITUTION

FIRE OUTER APPEARANCE

Fire constitution types usually have broad paravertebral muscles and well-proportioned shoulders, upper back, and thighs. They have small pointed heads, pointed chins, small hands, and feet, and curly, or no hair, on top of their head (Figure 3.11). By nature, they enjoy spring and summer but dislike autumn and winter. This preference is due to the vulnerability of the Fire constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include a strong circulatory system and a red facial color. Their note is "zhi."

FIRE PERSONALITY

The Fire personality can be divided into three subdivisions.

1. A Balanced Fire nature predisposes individuals to be trusting, open-minded, complacent, social, unconcerned about wealth, and fond of beauty. They love themselves and others, and are very expressive of their affection. They are calm, peaceful, happy, lively, spontaneous, funny, and fun to be with.
2. A Yin Fire nature predisposes individuals to be solemn and depressed. They tend to lack interest in life, and have a tendency to become isolated, feeling unloved and unlovable.

Treatment for Yin-Fire nature individuals requires teaching them how to store and conserve their energy, to use moderation in what

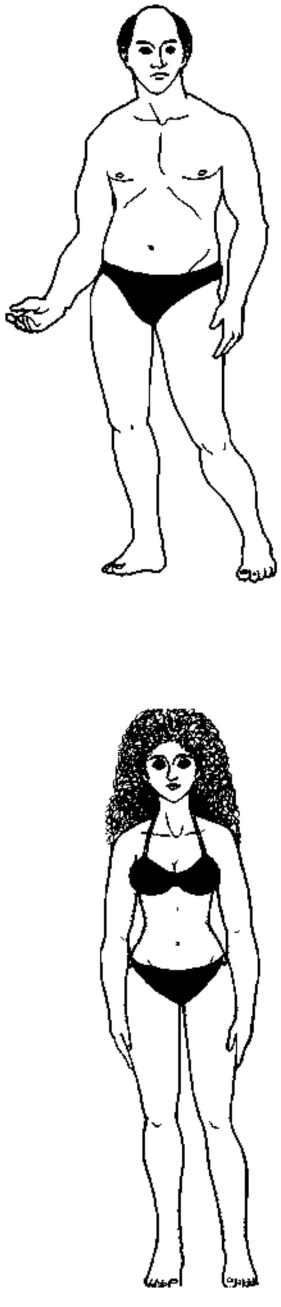

Figure 3.11. The Fire Constitution in Males and Females

they do, and to avoid extremes. They also need to engage in activities that create personal enjoyment, to find simple pleasures that awaken their affection, and to learn how to express their feelings, wants, and needs.

3. A Yang Fire nature predisposes individuals to be arrogant, ignorant, and troublesome. They are restless, excitable, and talk excessively. They always exaggerate and are overenthusiastic. They are socially and sexually overactive and seek every opportunity to assert themselves. They are foolish and display careless behavior. They are overconfident, slightly lazy, irresponsible, and less than truthful. They can be manic, get exhausted, then burnout and become suicidal (e.g., bipolar disorder).

   Treatment for Yang-Fire nature individuals requires teaching them to learn how to stop, slow down, and look for their contentment from within. They should balance their feelings of love with contemplation and wisdom, avoid over-enthusiasm, and allow their inner spirit to radiate through in a more sober way.

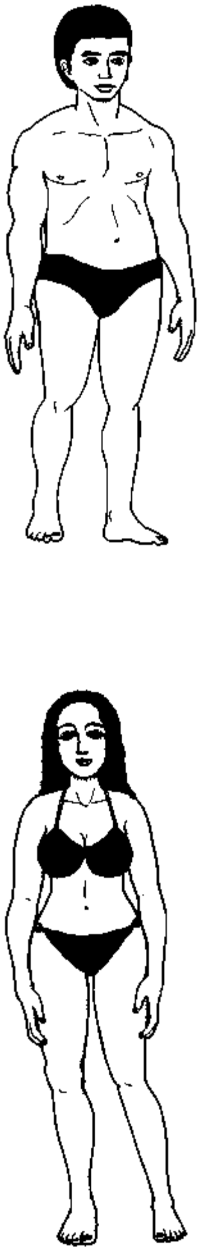

Figure 3.12. The Earth Constitution in Males and Females

THE EARTH CONSTITUTION

EARTH OUTER APPEARANCE

Earth constitution types usually have a somewhat large body, large head, large belly, strong thighs, round face, and wide jaw (Figure 3.12). They are categorized as having excessive flesh, with upper and lower limbs mutually well-proportioned. By nature, they enjoy autumn and winter but dislike spring and summer. This preference is due to the vulnerability of the Earth constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include strong muscles and yellow facial color. Their note is "gong."

EARTH PERSONALITY

The Earth personality can be divided into three subdivisions.

1. A Balanced Earth nature predisposes individuals to lead quiet, stable, and peaceful lives, unconcerned about fame or wealth. They are always at ease, calm, generous, forgiving, sincere, and unambitious. They usu-

ally have open minds and rarely live in fear, or make excessive demands. They have analytical minds and are logical and practical; they use sound reasoning to convince others of their opinions. They are quite adaptable to changing situations, and rarely use coercion to achieve power. They are kind and gentle with an earnest and well-mannered attitude. They are pleasant, sweet, sympathetic and caring, and are able to maintain boundaries.

2. A Yin Earth nature predisposes individuals to worry endlessly, becoming suspicious and self-centered. They think too much, with not enough follow-through action. Because they feel empty inside, they find it difficult to be nurturing to themselves and others.

   Treatment for Yin Earth nature individuals requires teaching them to learn how to let go of their inner feelings of worry and defensiveness. These individuals need to examine their assumptions, connect with their physical bodies, come out of their shells, and begin to live in the real world. They need to replace their negative thought patterns with positive affirmations, and realistically look at their true potential and abilities for powerful actions.

3. A Yang Earth nature predisposes individuals to cling to others; they are co-dependent, pushy, and possessive. Such individuals try to dominate in a passive-aggressive way while limiting the independence of others.

   Treatment for Yang Earth nature individuals consist of teaching them to develop their inner strength to control the fear, insecurity, and feelings of inner emptiness that make them want to hold on to others. They need to create and establish love as a source of security from within themselves. They need not depend on, or build their lives exclusively on, the praise or presence of others. They must learn to become emotionally independent.

THE METAL CONSTITUTION

METAL OUTER APPEARANCE

Metal constitution types usually have a triangular face with broad, square shoulders, a strong voice, and a strong, muscular build (Figure 3.13). By nature, they enjoy autumn and winter but dislike spring and summer. This preference is due to the vulnerability of the Metal constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include strong Lungs and white facial color. Their note is "shang."

METAL PERSONALITY

The Metal personality can be divided into three subdivisions.

1. A Balanced Metal nature predisposes individuals to process their grief, letting go of their past emotional baggage. They believe that they cannot be in the present without cleaning up and being released from their past. They are able to gain knowledge and wisdom by gathering and releasing their emotional bonds while learning and growing from each emotional attachment. They participate in life and form new bonds without fear of loss. They are also generous, just, and bright.

2. A Yin Metal nature predisposes individuals to have a difficult time creating lasting bonds because they are fearful of establishing new relationships due to past losses and emotional trauma. They avoid joining in with others and withdraw from active participation. Living in past remorse, they grieve over lost opportunities. They are also haughty and coldhearted. They are jealous, cunning, sneaky, and furtive. They become angry when they do not have things that others possess. Because they are covetous, they only obtain happiness by possessing certain objects.

   Treatment for Yin Metal nature individuals requires teaching them how to strengthen their physical body, as well as the energy of their Heart, Spleen, and Lower Dantian. They also need to strengthen their abilities to form close emotional bonds with people, reduce their fears of rejection and abandonment, and gain the strength and courage to let go of their past hurts. They need to come out from inside of themselves emotionally and experience the warmth and compassion of life.

3. A Yang Metal nature predisposes individuals

to suppress their emotions and hold on to their grief. They are generally considered whiners and complainers, talking to others about their complaints in order to unload their grief. They use new relationships as an emotional bandage to avoid the unprocessed grief of their past relationships. They are also meticulous, independent, and strong-willed.

Treatment for Yang Metal nature individuals requires teaching them how to let go of the need to control and suppress their emotional feelings, and how to genuinely grieve over past wounds. They need to face the truth and be honest about their feelings, instead of selfishly using people and new relationships to cover up their grief. They also need to learn to be sympathetic with the pain and sorrows of others, and put their feelings of grief into perspective.

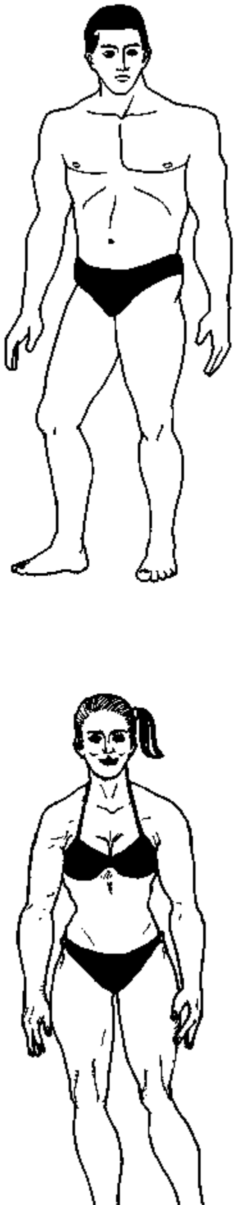

Figure 3.13. The Metal Constitution in Males and Females

THE WATER CONSTITUTION

WATER OUTER APPEARANCE

Water constitution types usually have a big round face, head and body (Figure 3.14). They have a long upper back, and unbalanced or uneven physical features. They, by nature, enjoy autumn and winter but dislike spring and summer. This preference is due to the vulnerability of the Water constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include their digestive system and black or dark facial color. Their note is "yu."

WATER PERSONALITY

The Water personality can be divided into three subdivisions.

1. A Balanced Water nature predisposes individuals to be skilled negotiators, that are not discouraged by difficulty, and do not take foolish risks. They are sympathetic and loyal to their employer and friends. They have a clear perspective and are sensitive and intuitive. They are powerful, tender, and soft. They have a firm will and know their boundaries and limitations. They are known for their inner strength and strong faith in themselves.

2. A Yin Water nature predisposes individuals to lack spiritual, emotional, mental, and physical energy. They give up on life and surrender the control of their own destinies. They lack the determination to achieve their goals, do everything halfheartedly, and become easily discouraged by difficult challenges.

   Treatment for Yin Water nature individuals requires teaching them how to conserve and strengthen their energy by not attempting to do what's beyond their capacity. They should complete their projects, however, once they have started them, and not procrastinate or leave tasks unfinished. They need to find the strength to overcome their fear of failure by learning to take action through the completion of their goals.
3. A Yang Water nature predisposes individuals to be ambitious overachievers and to live under great stress. They lack consideration for others, and can be reckless and foolhardy. They can also be greedy, ruthless, and cold-blooded. Seemingly modest, they are actually insidious and sinister, concealing their true emotions by suppressing their fears. Because they fear loss of control, their safety lies in dominating others. They only perform duties that are self-serving and blame others for their problems.

   Treatment for Yang Water nature individuals requires that they learn how to act from inner stillness and gain strength and courage from their inner true self. They also need to emotionally and physically slow down, and to balance their activities with rest. They need to learn to gain consideration for themselves and others, as well as to open up their heart and begin to love.

COMBINED CONSTITUTIONS

In ancient China, The Yellow Emperor's physician Shao Shi taught that the Five Elemental constitutions consist of Five Elemental forms, which can be further divided into five additional shapes depending on the strength or development of the individual's Jing, Qi, and Blood. After establishing the patient's physique (Wood, Fire, Earth,

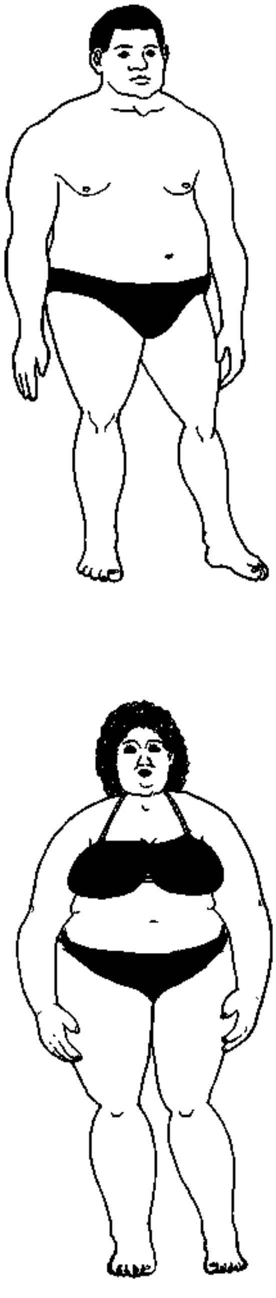

Figure 3.14. The Water Constitution in Males and Females

Metal, or Water), the patient is then classified in terms of five separate colors, and differentiated according to the five prenatal (Heavenly) notes. This system recognizes 25 constitutional types.

The Five Elemental Constitutions are therefore, guidelines for the Qigong doctor to evaluate the patient. These guidelines are not fixed, as the doctor is most likely to encounter combinations of two or more of these elements within one person.

When two constitutions combine in an individual, the combination may be congenital or acquired. The Qigong doctor must evaluate the patient, differentiating between his or her congenital and acquired tissue formations.

A woman who has large hips (Water) and small chest and shoulders (Wood), for example, probably inherited this structure from her parents. This woman's physical structure and personality would then be considered a combination of both Water and Wood (depending on which internal organ characteristic dominates).

A female patient who swims, on the other hand, may be congenitally Wood but may have developed a dominant Metal upper physique through years of physical training. If, through intense physical activity, someone acquires a Metal physique, he or she may also acquire a Metal personality, depending again on his or her psychological format, which is rooted within the internal organs.

The external physical structure serves as a visual guide to the interaction and balance of the internal organs.

PERSONALITY CONSTITUTIONS OF THE EIGHT EXTRAORDINARY VESSELS

Another system of identifying personality traits is practised by observing the four personality constitutions of the patient's Eight Extraordinary Vessels.

Through observation and inspection of the patient's personality and overall energetic and emotional demeanor, the Qigong doctor can further determine which organ and organ system, is Deficient or has an Excess condition and set up a treatment plan.

To begin, the Qigong doctor will first observe the patient's dominant Yin or Yang emotional characteristics and determine which of the Eight Extraordinary Vessels are governing his or her present state of feelings. These personality traits may appear in different combinations, or oscillate from one type of characteristic to another, due to the crossover of acquired and congenital organ energy. This crossover can change with the patient's age (maturation) or situation (environment).

THE PATHOLOGICAL PERSONALITY TYPE OF THE GOVERNING AND YANG HEEL VESSELS

The Governing and Yang Heel Extraordinary Vessels are located on the back and lateral sides of the body and affect the Bladder, Stomach, and Gall Bladder Channels. To "open" and increase the flow of energy within these channels, the Qigong doctor stimulates the patient's SI-3 and Bl-62 points simultaneously (see Chapter 8).

The pathological personality types of the Governing and Yang Heel Vessels are divided into Yin and Yang emotional characteristics. These characteristics are as follows.

1. Patients with Deficient Qi in the Governing and Yang Heel Vessels tend to be weak-willed and spineless. They tend to be overly submissive, and easily give up their personal power. They lack determination, courage, clarity, and inner strength, and have no control over their lives. They participate as little as possible in life for fear of failure.
2. Patients with Excess Qi in the Governing and Yang Heel Vessels tend to be tense, overpressured, stiff, rigid, inflexible, stubborn, and awkward. They are narrow-minded and overcontrolling. Fearful of letting go, they overcompensate and attempt to restrict reality.

THE PATHOLOGICAL PERSONALITY TYPE OF THE CONCEPTION AND YIN HEEL VESSELS

The Conception and Yin Heel Extraordinary Vessels are located on the front and inside of the body. They affect the Lung and Kidney Channels, which in turn affect the Kidneys, Lungs, and Heart organs. To open these channels, the Qigong doctor stimulates the patient's Lu-7 and Kd-6 points simultaneously. These character diagnoses are as follows.

1. Patients with Deficient Qi in the Conception and Yin Heel Vessels tend to be weak and depressed. They live in the past, lack interest in life, have no ambition, and also lack interest in sex. It is hard for them to form new relationships because they fear forming emotional bonds due to past losses, failures, and emotional traumas. They constantly daydream and live in their own world of past memories.
2. Patients with Excess Qi in the in the Conception and Yin Heel Vessels tend to participate in life, but hold on to, and suppress, their grief. They fail to let go of emotional attachments and fear being alone. The female patients usually develop Qi stagnations resulting in breast cysts, tumors, and cancer, as well as uterine fibroids and cancer.

THE PATHOLOGICAL PERSONALITY TYPE OF THE BELT AND YANG LINKING VESSELS

The Belt and Yang Linking Extraordinary Vessels are located around the waist and on the outside of the body. They affect the Gall Bladder and Triple Burner Channels, which in turn, affect the Kidneys, Liver, and Gall Bladder organs. To open these channels, the Qigong doctor stimulates the patient's GB-41 and TB-5 points simultaneously. These character diagnoses are as follows.

1. Patients with Deficient Qi in the Belt and Yang Linking Vessels are considered weak, indecisive, and unproductive. They fear criticism and are touchy, snappy, irritable, and hypersensitive. They have low self-esteem and also lack interest in sex.
2. Patients with Excess Qi in the in the Belt and Yang Linking Vessels are aggressive, angry, resentful, bitter, and vindictive. They are also opinionated, inflexible, intolerant, domineering, selfish, and frustrated, with a sex life based on anger or rage.

THE PATHOLOGICAL PERSONALITY TYPE OF THE THRUSTING AND YIN LINKING VESSELS

The Thrusting and Yin Linking Extraordinary Vessels are located within the center of the body (internally) and on the medial and anterior aspect of the body (externally). They affect the Kidney and Stomach Channels, which in turn, affect the Kidneys, Spleen, and Heart organs. To open these channels, the Qigong doctor stimulates the patient's Sp-4 and Pc-6 points simultaneously. These character diagnoses are as follows.

1. Patients with Deficient Thrusting and Yin Linking Vessels are physically and emotionally weak. They quickly become exhausted and depressed. They easily have their feelings hurt and take a long time to recover from emotional upsets. Their internal world of emotional existence is made up of fear, anxiety, worry, and tension. They do not connect well with the outside world or gain much pleasure from life or relationships.
2. Patients with an Excess in the Thrusting and Yin Linking Vessels tend to participate in life, but have difficulty expressing their affections and communicating their needs in relationships. They are inclined to have stagnant Qi and Blood in their chests, epigastrium, or uterus areas due to fear, anxiety, worry, and sorrow. In their personal relationships, they fear loss of control and cannot make a commitment or a deep emotional connection with others. They also fear surrendering and letting go in their love and sexual relationships.

THE YAO IMAGE AND PHYSICAL ANATOMY

Sometimes because of surgery, congenital deformity, or traumas, it is not possible, nor is it recommended, to directly treat certain areas of the patient's body. The Qigong doctor must then choose an indirect method of treatment to ensure better results. One such method available to the doctor utilizes the understanding of the Yao image and the patient's physical anatomy. In order to better comprehend the complexity of this diagnostic and treatment modality, the Qigong doctor must first understand the energetic origin of the Yao images.

All matter is composed of different proportions of Yin and Yang energy. Within the Void (infinite space or Wuji) both Yin and Yang energy gather or disperse to balance the forces of nature. This interaction causes the Yin and Yang energies

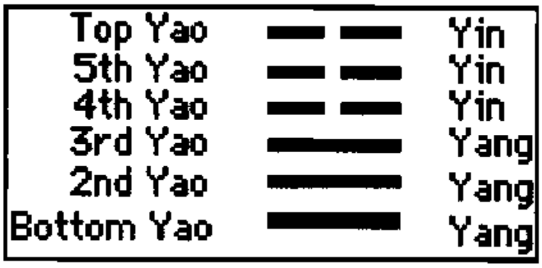

Figure 3.15. This diagram shows a hexagram, composed of an upper and lower trigram. In this particular example, the upper trigram is all Yin lines (symbolized by broken lines), the lower all Yang (symbolized by solid lines).

to develop and transform themselves into four phases of energetic powers or stages: Great Yin, Small Yin, Great Yang, and Small Yang. These four phases can be explained as follows.

1. Great Yin (Tai Yin), is affiliated with midnight and the new-moon phase. Modern physicists associate the Great Yin with a weak nuclear force.
2. Small Yin (Shao Yin), is affiliated with the sunset and the waning-moon phase. Modern physicists associate the Lesser Yin with a heavy force, and gravity.
3. Great Yang (Tai Yang), is affiliated with high noon and the full-moon phase. Modern physicists associate the Strong Yang with a strong nuclear force.
4. Small Yang (Shao Yang), is affiliated with the sunrise and the waxing-moon phase. Modern physicists associate the Lesser Yang with a light force and electromagnetism.

These four solar and lunar energetic phases transform themselves into the energies of the eight foundational trigrams. These trigrams further combine to form sixty-four hexagrams, and manifest as the ancient binary system known as the Yi Jing (I Ching). The formation and pattern of these energetic powers or stages vary according to the composition of Yin and Yang, and can be expressed through the energetic symbols of the Yao.

A Yao is a line that represents either Yin or Yang energy. The lines fall into two categories:

- The negative Yao is symbolized by broken lines (Yin energy).
- The positive Yao is symbolized by solid lines (Yang energy).

When these lines are stacked in combinations of three, they form what are known as trigrams (a group of three Yao lines). When pairs of trigrams are joined, hexagrams (six lines) are formed. The lines are arranged from bottom to top, with the first Yao being on the bottom (Figure 3.15).

The Yao trigrams and their relationship to the body's physical and energetic structure can be utilized by the Qigong doctor for diagnosis and treatment. The doctor applies the Yao image as a template and arranges its pattern alongside the physical form of the patient's body. This arrangement of Yao images is designed by the doctor to assist him or her in recognizing the physical and energetic imbalances within the patient's body. This type of physical Yao image diagnosis has been used in China for centuries.

Knowing that each different region of the human body corresponds to one of the six Yaos in the hexagram, as well as a different aspect of Qi, the ancient masters explained the energetic Yao and physical body interaction as follows.

1. The two lowest Yaos align with the feet and legs and correspond to the body's interaction with the Earth's energy.
    a. The 1st Yao extends from the patient's feet, ankles, and shins, to the knees.
    b. The 2nd Yao extends from the knees to the upper thighs.
2. The middle two Yaos align with the lower and upper abdominal regions, and correspond to the body's interaction with human energy, or "Man's energy" (emotional responses).
    a. The 3rd Yao extends from the perineum to the navel.
    b. The 4th Yao extends from the navel to the xiphoid process in the diaphragm.
3. The upper two Yaos align with the upper chest, neck and head, and correspond to the body's interaction with Heaven's energy.
    a. The 5th Yao extends from the xiphoid process of the diaphragm to the top of the manubrium, at the base of the throat.

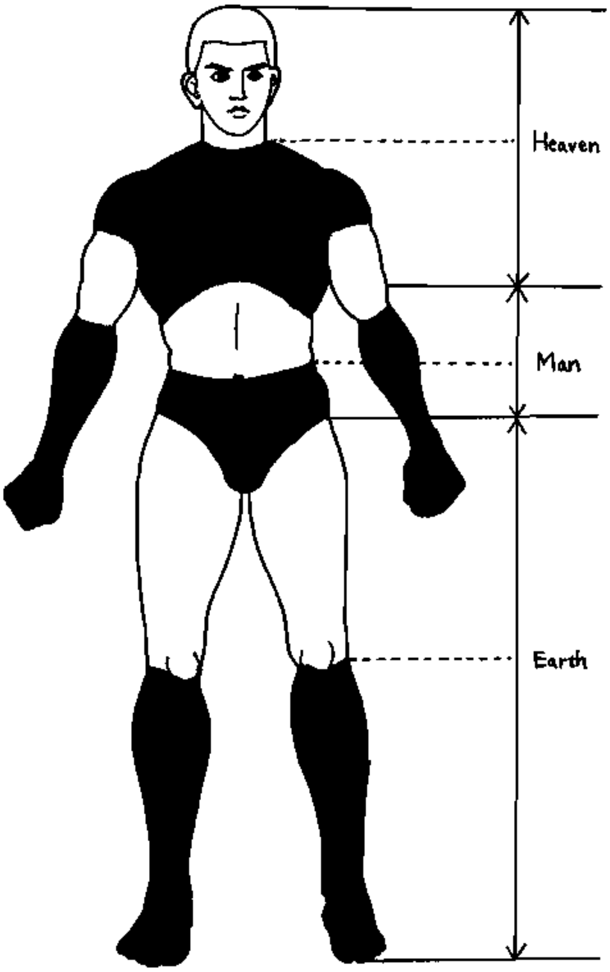

- The top or 6th Yao is from the neck to the top of the head. (Heaven)
- The 5th Yao is from the diaphragm to the neck. (Heaven)
- The 4th Yao is from the navel to the diaphragm. (Man)
- The 3rd Yao is the lower perineum to the navel. (Man)
- The 2nd Yao is from the knee to the top of the thigh. (Earth)
- The bottom or 1st Yao is from the foot to the knee. (Earth)
- The Categorization of the arms are as follows: From the fingers to the elbow represents the Bottom or 1st Yao; From the elbows to the base of the shoulder represents the 2nd Yao; From the base to the top of the shoulder represents the 3nd Yao.

Figure 3.16. Yao Image and the Human Body

b. The 6th Yao extends from the manubrium, at the base of the throat, to the top of the head.

The doctor further divides the patient's body into six upper and lower Yaos, to determine the dominant energetic and emotional patterns (where the patient's Qi is gathering, collecting, and stagnating). Through observation and study, the Qigong doctor is able to determine the energetic association the patient has with the natural environment. This diagnosis is achieved by observing the body's relationship to the three energetic Yao divisions of Heaven, Earth, and Man (Figure 3.16). When observing the body, for example, if the upper Yao is out of harmony with the patient's middle and lower Yaos, the patient's energetic balance may be compromised, and he or she may experience a spiritual disconnection with his or her body. This can result in either a Yang condition (heaviness, restlessness) due to an Excess spiritual "Heaven" state, or a Yin condition (tired, dizziness) due to a Deficient spiritual "Heaven" state.

In addition to diagnosing the hexagram for the body's Yin and Yang relationship to Heaven, Earth, and Man, the ancient Chinese Qigong masters also studied two other hexagrams for clinical evaluation. These hexagrams were known as the Upper and Lower Hexagrams.

UPPER AND LOWER HEXAGRAMS

In the Ming Dynasty (1368-1644), Tang Rongchuan stated in his book *Detailed Explanations of Application of the Book of Changes to Medicine*, "In the clinical environment, the body may be divided into two different sets of hexagram constructions [the upper and the lower hexagram partition] for the purpose of making a diagnosis, detecting Qi, and for performing Qigong therapeutic treatments." This relationship of Yao positioning is suitable to both the channel system and the nervous system.

Physical disorders located within these hexagrams are generally expressed as symptoms manifesting in the extremities. If stagnant Qi or other abnormal conditions occur in any particular part of the body, an imbalance of both Internal and External Qi may be detected in either that specific area of the body or its corresponding area.

The ancient Qigong masters used the body's hexagrams to explain the etiology, pathology, clinical manifestations, and treatment principles of disease.

The body's upper torso, arms, neck, and head are divided into what is known as the upper hexagram partitions, while the body's lower limbs and torso are divided into the lower hexagram partitions. The six Yao of the upper and lower hexagram partitions interlock at the chest and abdomen. By using these sets of hexagrams, the Qigong doctor has a template or guiding pattern with which to understand the patient's body.

The Upper Hexagram Construction

The Upper Hexagrams range from the base of the lower perineum to the top of the patient's head. These six Yaos are organized as follows.

1. The 1st or bottom Yao encompasses the pubic symphysis to the navel. Reproductive, digestive, and urinary tract diseases are ascribed to this area.
2. The 2nd Yao encompasses the navel to the xiphoid process. Malfunctions in digestion, elimination and the transportation of nutrients are assigned to this area, as well as diseases of the Kidneys, adrenal glands, Liver, Spleen and pancreas.
3. The 3rd Yao encompasses the xiphoid process to the supraclavicular notches. Respiratory and circulatory diseases are ascribed to this area.
4. The 4th Yao encompasses the supraclavicular notch to the tip of the nose. Teeth, jaw, and thyroid diseases are ascribed to this area.
5. The 5th Yao encompasses the tip of the nose to the eyebrow. Sinus problems and headaches, as well as ear, nose, and eye diseases are manifest in this area.
6. The 6th or top Yao encompasses the eyebrows to the top of the head. Several types of headaches, brain tumors, and upper cranial dysfunctions are ascribed to this area.

The Lower Hexagram Construction

The Lower Hexagrams range from the bottom of the patient's feet to the base of the patient's throat. These six Yaos are organized as follows:

1. The 1st or bottom Yao encompasses the bottom of each foot to the ankle. Foot, toe and ankle dysfunctions are ascribed to this area.
2. The 2nd Yao encompasses the ankles to the knees. Shinsplints and calf and ankle dysfunctions are ascribed to this area.
3. The 3rd Yao encompasses the knees to the pubic symphysis. Thigh, knee, and quadriceps dysfunctions are ascribed to this area.
4. The 4th Yao encompasses the pubic symphysis to the navel. Reproductive, digestive, and urinary tract diseases are ascribed to this area.
5. The 5th Yao encompasses the navel to the xiphoid process. Digestion, elimination, and the transportation of nutrients, as well as diseases of the Kidneys, adrenal glands, Liver, Spleen and pancreas are ascribed to this area.
6. The 6th or top Yao encompasses the xiphoid process to the supraclavicular notches at the base of the throat. Respiratory and circulatory diseases are ascribed to this area.

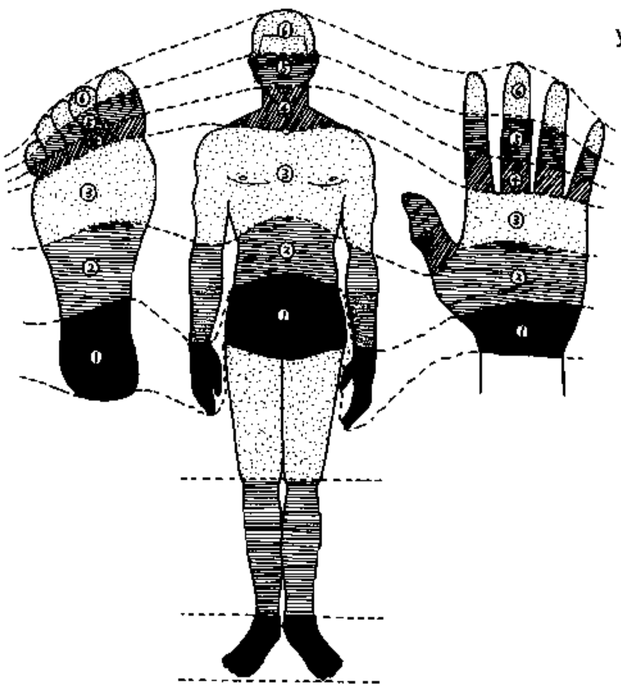

The diagram at left depicts the correspondence of the yaos of the body to the yaos of the hands and feet.

6 The second joint of the toes or fingers pertains to the top Yao and the energy of Heaven.

5 The first joint of the toes or fingers pertains to the 5th Yao and the energy of Heaven.

4 The area below the first joint of the toes or fingers pertains to the 4th Yao and the energy of Man.

3 The ball of the foot or upper palm pertains to the 3rd Yao and the energy of Man.

2 The instep on the midpalm with the thumb pertains to the 2nd Yao and the energy of Earth.

1 The heel or lower palm pertains to the bottom Yao and the energy of Earth.

Figure 3.17. The Therapeutic Use of the Yaos and the Upper and Lower Hexagrams

THERAPEUTIC USE OF THE YAOS

The entire body can be treated by focusing on one small area. When Excess, Deficient, or stagnant Qi occurs in one part of the body, the imbalance of Internal and External Qi is detectable in that area or its corresponding region on one of the body's extremities. This ancient Chinese modality of treatment is popularly called reflexology in the West.

The Yao images can be superimposed onto the patient's body, and the patient diagnosed according to the energetic symptoms expressed through either the patient's extremities (the hands, forearms and arms; or the feet, shins, and thighs), or the six divisions located on the patient's torso and head. A Yao hexagram can additionally be visualized on specific locations of the patient's body such as the face, nose, ears, hand or foot for diagnosis. The base of the palm's heel, for example, corresponds to the lower abdominal area and reproductive organs, which relate to the 1st, or lower, Yao in the hexagram of the hands. These six Yao areas of the hand provide a basis for diagnosis, for treatment of certain diseases, and even for perception and prediction of certain future diseases (Figure 3.17).

As an example of how the Qigong doctor can use the six Yaos of the body for emitting Qi therapy, consider the following situation: An elderly patient with a problem in his prostate area visits the Qigong doctor. The prostate is located in the lower abdominal area, which is located in the 1st Yao of the upper hexagram. Since the six Yao positions can be transferred to the hands, feet, etc., the doctor knows that this particular Yao is congruent to the 1st Yao of the foot. Therefore, as an energetic entry point into the patient's prostate area, the doctor may focus his or her attention on a specific area on the foot.

Dividing the foot into six Yaos, the doctor knows that the 1st Yao of the foot is the heel area. In order to treat the prostate gland, the doctor will

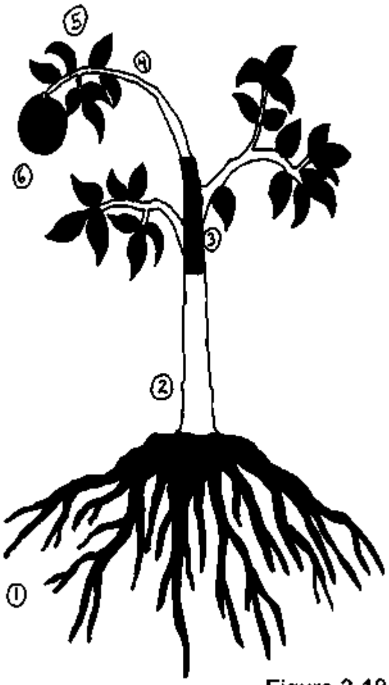

6 The fruit or flower of the herb pertains to the top Yao and the energy of Heaven.

5 The leaf of the herb pertains to the 5th Yao and the energy of Heaven.

4 The branch of the herb pertains to the 4th Yao and the energy of Man.

3 The upper stem of the herb pertains to the 3rd Yao and the energy of Man.

2 The stalk of the herb pertains to the 2nd Yao and the energy of Earth.

1 The root of the herb pertains to the bottom Yao and the energy of Earth.

Figure 3.18. The Yao Image Applied to Plants

focus his or her treatment on the patient's heel and ankle area. Alternatively, the doctor could also have chosen the 1st Yao area of the hand, which is considered the 1st Yao position of the arm. The 1st Yao of the hand corresponds to the prostate and urogenital area.

HERBS AND THE YAOS

Affirming the ancient Chinese belief that the energy of Heaven, Earth, and Man combine as "one," and exist in the same world of energetic dynamics, herbs are also structured and prescribed in accordance with the six Yao hexagram concept. With plants, the following correspondences apply (Figure 3.18):

1. The roots correspond to the bottom or 1st Yao.
2. The stalk corresponds to the 2nd Yao.
3. The upper stem corresponds to the 3rd Yao.
4. The branches correspond to the 4th Yao.
5. The leaves correspond to the 5th Yao.
6. The flowers and fruit of the plant correspond to the top or 6th Yao.

When plants are utilized as a medicine, the part of the plant corresponding to the number and position of the Yao may manifest its energetic properties, affecting the ascending, descending, floating, or sinking action of the patient's Qi. The achyranthes root (Niu Xi), for example, corresponds to the bottom Yao, and leads healing Qi downward, to heal the patient's knees and joints; whereas the chrysanthemum flower (Ju Hua) corresponds to the top Yao, and leads healing Qi up into the patient's head and eyes.

IDENTIFYING INTROVERTED AND EXTROVERTED STRUCTURES

Qi flows through the matrix of the connective tissue and adapts to the postural and structural demands of the body by changing the density and direction of the body's connective tissues and inner fasciae. A postural habit will create an imprint or tissue memory in the supporting connective tissue, as well as an energetic memory in the supporting field of Qi. These imprints result in introverted as well as extroverted postural structures.

Each introverted (Yin) and extroverted (Yang) postural structure has a predictable impact on the

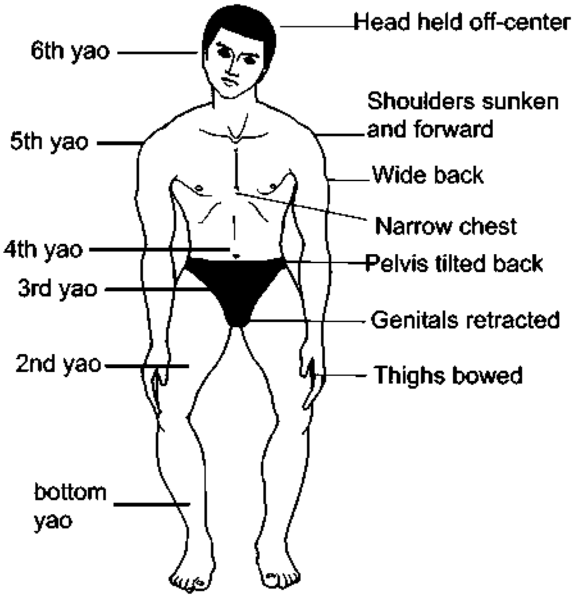
Head held off-center
6th yao
Shoulders sunken and forward
5th yao
Wide back
Narrow chest
4th yao
Pelvis tilted back
3rd yao
Genitals retracted
2nd yao
Thighs bowed
bottom yao

Figure 3.19. The Introverted Structure

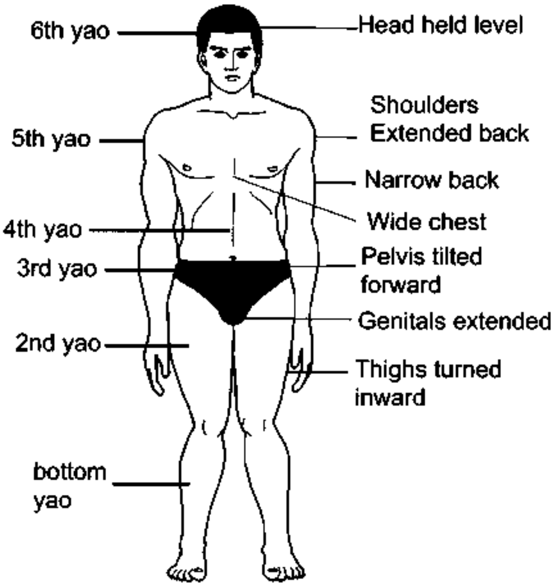
6th yao
Head held level
Shoulders Extended back
5th yao
Narrow back
Wide chest
4th yao
Pelvis tilted forward
3rd yao
Genitals extended
2nd yao
Thighs turned inward
bottom yao

Figure 3.20. The Extroverted Structure

myofascial webbing, as well as a natural emotional response. Emotions respond to changes in breathing patterns and vice versa. Both the prevalent breathing pattern and the patient's general mood affect the body's posture and eventually the body's structure. All chronic stress creates postural imprinting or tissue memory, and a corresponding energetic memory in the supporting field of Qi.

INTROVERTED STRUCTURE

When observing the introverted type of structure, the doctor will notice that the back (Yang) area of the patient is expanded, and the muscle structure is usually well-developed. The front of the body (Yin), especially around the chest and Heart area, is retracted, resulting in a narrow chest. The patient's head is generally held off-center, shoulders are sunken and forward, and the pelvis is tilted back, with the genitals retracted (anterior tilt). This condition is commonly referred to as "extended Yang" and "retracted Yin" of the upper torso (Figure 3.19).

EXTROVERTED STRUCTURE

When observing the extroverted type of structure, the doctor will notice that the back (Yang) area of the patient is retracted, while the front of the body (Yin) is expanded and overly exposed. The head is usually held vertically, the shoulders are extended back (military posture), the chest is wide, and the pelvis is tilted forward, with genitals extended (posterior tilt). This condition is commonly referred to as "extended Yin" and "retracted Yang" of the upper back (Figure 3.20).

COMBINED STRUCTURAL FORMATIONS

As with the Five Elemental Constitutions, the introverted and extroverted structures are guidelines for the Qigong doctor to evaluate the patient. These guidelines are not fixed, as the doctor is most likely to encounter combinations of both introverted and extroverted structures within one person.

When two structures combine in an individual, the combination may be due to a congenital or acquired formation. The Qigong doctor must evaluate and assess the patient, differentiating between his or her congenital and acquired tissue formations.

CONGENITAL AND ACQUIRED CELLULAR PATTERNS

The Qigong doctor evaluates the physical condition according to the congenital or acquired patterns manifested within the patient's tissues. The doctor may ask questions to determine if the condition was acquired through exposure to environmental pathogens, diet, trauma, occupation, or inherited. This information aids the Qigong doctor in understanding and assisting the patient in reprogramming any congenital or acquired cellular patterns, as well as internal disease functions.

1. Congenital Cellular Patterns and disease predispositions are locked within the patient's congenital Jing and can be released like a biological time bomb. This is a similar concept to the Western medical concept of genetic congenital disease. The goal of the Qigong doctor is to alter (if possible) the Toxic Qi patterns and help the patient's body recognize that the formation of the disease is a mistake. In order to reprogram these patterns, the Qigong doctor often assists the patient in using guided meditations that employ vivid imagination, colorful visualizations, and positive affirmations. These images are used to encapsulate the disease and return it to the divine light. While in the presence of the divine light, the Toxic Qi (Xie Qi) will be energetically transformed and recycled back to the patient as healing energy. Patients with a direct family history of cancer, for example, can benefit from this type of medical Qigong imagery.
2. Acquired Cellular Patterns and disease functions arise from either external exposure to pathogens, or the suppression of extreme emotions. The goal of the Qigong doctor is to first assist the patient in discovering the origin of the disease. The next step is to teach the patient to alter the toxic patterns and help the patient's body recognize that the development of the disease is a mistake. In order to reprogram these patterns, the Qigong doctor will assist the patient in learning meditations that employ healing mental imagery, colorful visualizations, and positive affirmation. The patient must encapsulate the disease and return it to the divine light, to be energetically transformed and recycled. A patient who has been experiencing severe headaches, for example, could benefit from this kind of treatment given that most headaches are stress related.

SUMMARY

Through the observation of the patient's physical and structural development, a clear understanding can be obtained as to the strength and weakness of the patient's internal organs and immune system.

The four patterns of physical development and structural formation (The Five Elemental Constitutions, The Yao Hexagram Formations, The Yin or Yang Structures, and The Congenital and Acquired Cellular Patterns) are determined as follows.

1. The doctor first begins to diagnose the patient's physical body according to the Five Elemental Constitutions and the dominant element. This helps the doctor to understand the dominant condition of the patient's internal organs.
2. The doctor further divides the patient's body into six upper and lower Yaos, to determine the dominant energetic and emotional patterns (where the patient's Qi is gathering, collecting, and stagnating).
3. Next, the doctor evaluates the patient's external structure according to his or her body's Yin and Yang characteristics, to understand whether the patient has a dominant introverted or extroverted physique.
4. Finally, the Qigong doctor determines whether the patient's condition is congenital, acquired, or a combination of both. This information helps the doctor to understand the innate cellular patterns that dominate the patient's physique.

CHAPTER 4
THE FIVE ENERGIES OF THE HUMAN BODY

The study of various forms of energy and their transformations from one form to another have led to one of the greatest generalizations in physics -- the Law of Conservation of Energy, which states, "Energy cannot be created or destroyed. It may be transformed from one form into another, but the total amount of energy never changes."

Physics defines four types of energy: mechanical, electromagnetic, chemical, and thermal. All four types of energy are related to and convert from one another. Electricity, for example, can be converted into sound, light, mechanical energy and into heat. Sound, light, mechanical energy and heat can all be reconverted into electricity.

Western Science is still for the most part, founded on the principles of Newtonion Physics, that views energy as an impersonal and inanimate force. The Traditional Chinese view of energy is much more akin to the theories of Quantum Physics. Both Chinese energetic theories and Quantum Physics hypothesize that energy cannot be studied without taking into account the mind's influence over it. Furthermore, energetic behavior cannot be studied independently of matter.

From a medical Qigong perspective, the entire human body is suffused with energy, which converts within the human organism. Light energy, for example, is absorbed, stored and converted into the various energies of the body, the same way that plants convert absorbed and stored light into chemical energy.

The human body possesses five energies which resonate within the very core of one's being. Sound, light, magnetic fields, heat, and electricity envelop and permeate our very existence; these energies sustain, govern, control, and determine our psychology as well as our physiology. In Medical Qigong therapy, these energies are considered the spiritual reality that governs the physical reality. These energies are observed in both diagnosis and treatment of illness. Research in China confirms the fact that the body emits sound, light, magnetic energy, electricity, and heat.

SOUND ENERGY RESONANCES

When the body experiences any type of sound, the cell tissues respond to the tone frequency and have either a negative or positive automatic response. The human body both receives and generates sound energy. These subtle sounds resonate from three distinct physical energetic actions. The following subtle sounds are natural tones that resonate as a result of three physical actions.

1. Breathing creates respiratory sound resonances that follow the breath in and out of the lungs mouth, and nose.
2. Muscle movements create somatic and visceral sound resonances that include the clicking sound of the joints, beating of the Heart, gurgling of the Stomach and intestines, etc.
3. Qi activities can produce energetic sound resonances (i.e., when Liver Yang, Liver Fire, or Liver Wind rise, a high-pitched tone is heard in the ears. When the Kidney Yin is Deficient, this sound is like rushing wind, etc.).

SOUND THERAPY THROUGH EMITTED QI

As sound waves vibrate through the body, crystalline structures within the tissues transform the vibration into pulsed currents. These currents are then conducted to the various corresponding organs and glands, depending on the frequency and amplitude of the incoming wave signal. This tissue transformation (due to wave vibrations) automatically changes the function and flow of energy in the body. Thus sound vibrations have a profound effect on both human psychology and physiology. Sound or tone resonances, have been used for centuries as an effective healing tool and are currently used as an adjunct to modern Qigong Medicine.

For centuries in China, patients have used healing sound therapy to resonate certain parts of the body, stimulating the flow of Blood and Qi to and from specific internal organs to cure disease. These healing sounds specifically relate to the Five Agents and the Five Elemental aspects of the patient's internal organs (see Chapter 2). According to the Five Elements' Cycle, each of the Five Elements has a note:

- Gong - C = Earth
- Shang - D = Metal
- Jue - E = Wood
- Zhi - G = Fire
- Yu - A = Water

The Qigong doctor can project resonant sound vibrations by first focusing on all three Dantians at the center of the Taiji Pole. This focused concentration is used first to vibrate the doctor's specific internal organs and then to treat the patient. The Qigong doctor can use vibrating sound therapy by:

1. Audibly speaking the healing sounds to fill the Qigong doctor's energetic field with healing sound vibration. This energetic field of sound is then projected onto his or her patient. Audible sound resonation is considered a Jing to Qi level energy projection, and is generally used when the patient is armored and not energy sensitive.
2. Inaudibly speaking the healing sounds and projecting these mental vibrations into the patient. In this case, the Qigong doctor will begin by internally focusing his or her mind on the healing sound. As this internal sound fills the Qigong doctor's energetic field, it is then emitted into the patient's body. To increase the intensity of energy sound projection, the doctor needs only to increase the mind's intention and the pressure of his or her exhalation. Inaudible sound resonation is considered a Qi to Shen level energy projection, and is generally used when the patient is energy sensitive.
3. Both the doctor and patient simultaneously resonate the sound to enhance the synchronization of both their energy fields. As the Qigong doctor resonates his or her own internal organs, both the doctor and patient can modulate the sound and energy patterns as needed.

THREE MODALITIES OF SOUND THERAPY

The resonant sounds produced by Qigong doctors are effective because the patients' energetic matrix is sensitive to the sound resonances. The modulated, yet unpredictable, nature of sound pulsation affects the vibrational structures which shape and maintain the patient's physical structure. Once the patient's vibrational structures have been disturbed, their body must immediately begin its reconstructive process, thereby creating new energetic patterns to restructure and heal the tissues.

There are three modalities used for sound therapy:

1. Characteristic Sound Therapy refers to the sound from musical instruments (flute, guitar, piano, drum, etc.) and their effect on the subconscious mind. This type of musical energy creates an emotional, physical, and spiritual release of energy within the listener.
2. Extraordinary Sound Therapy refers to the vocal resonances used in silent chanting, prayer and singing. Such techniques include the Daoist six-word healing sound method, Zhuang Zi's breath listening method, Lao Zi's sound-voice method, etc. This type of therapy uses the patients' energy to produce the sound and inner vision to direct the vibration within the body. It is also practised in conjunction with breathing techniques that are guided by the Yuan Shen (the intuitive consciousness of the spirit). All of these techniques are combined to achieve physiological and psychological healing.
3. Infrasonic Sound Therapy refers to the low frequency sound waves emitted from the hands of Qigong doctors. These chaotic, low frequency sound waves, inaudible to the human ears, are naturally produced by the hands

of experienced Qigong doctors when emitting Qi. This infrasonic sound resonance affects the central nervous system, changing the body's neurophysiological functions. Extensive research performed by Richard H. Lee of China Healthways Institute in Los Angeles, California, as well as research performed by the Beijing College of Traditional Chinese Medicine, and the Department of Natural Science in Beijing China, confirms that all humans have a very high degree of acoustic activity in the subsonic range below 20 Hertz (infrasonic). This subsonic activity is similar to the alpha rhythm of an EEG (see Appendix #2).

Qigong doctors may combine the characteristic, extraordinary, and infrasonic sound therapies into one treatment modality to facilitate a deeper and more thorough transformation within their patients. The choice of sound prescriptions vary according to each patient's cultural upbringing, religious beliefs, emotional temperament, and comprehension of energetic healing modalities.

Sound therapy can be prescribed to relieve a patient's feelings of uneasiness or extreme depression. Music, for example, has always been acknowledged as a powerful medium for emotional enhancement. Sounds and tones stimulate the cerebral centers creating a wide range of results from enhancement of memory retention, to facilitating certain emotions, and creative inspirations.

Music therapy specifically created for the harmonizing the Five Elements within the Yin and Yang organs is also used in conjunction with Feng Shui training to regulate any imbalances of the body and mind stemming from the changes of seasons. During autumn, for example, when dryness is prevalent and there is a transition from warm to cool, music can help to stabilize the body's response to these external conditions.

WESTERN SOUND THERAPY

Sound therapy is actively being used in Western medicine. The following are examples of current clinical modalities:

- music therapy for postoperative healing,
- ultrasound therapy for sore muscles and back injuries, and
- shock-wave lithotripsy therapy for kidney stones and calcified gallstones.

LIGHT ENERGY

All matter produces, radiates, transports, or transmits light, and therefore has properties of color vibration. Matter and light are fundamentally inseparable. Matter can be considered a condensation of photons. Its transformation can be viewed as a result of light interacting with atoms and molecules. All life on Earth, whether plant, animal, or human, is dependent upon light and its miraculous qualities for existence.

All living cells emit units of light called "biophotons." Biophoton emissions radiate from the internal organs, as well as the body's surface tissues, to create the body's auric field.

This interaction of light on the body's tissues and within the body's internal organs has a profound healing effect, especially when activated by the Qi emission from a Qigong doctor. When muscles or nerves are activated through medical Qigong therapy and / or exercises, the intensity of the biophoton emission increases. Studies performed on the healing potential of light and color therapies demonstrate that it is clinically effective. Like sound resonances, light therapy is also being researched further as an adjunct to modern clinical medicine.

The body both absorbs and projects light energy. The divine light that enters at the time of conception continues to reside in our Taiji Pole throughout our lives. This light responds and reacts to the doctor's emitted therapeutic light energy. Emitted light and color resonance are absorbed into the patient's body, causing the energies of Heaven and Earth to fuse inside the tissues; this facilitates the healing process.

In Traditional Chinese Medicine, the early diagrams illustrating the body's internal organs, channels, routes of Body Fluids, and arteries were called Charts of the Hall of Light. These detailed maps of Qi and Blood flow illustrated the energetic organs, as well as the body's Jing and Shen (which are rooted in the Blood). The Shen flows inside of the body, transforming into light and radiating from inside of the tissues outward. The

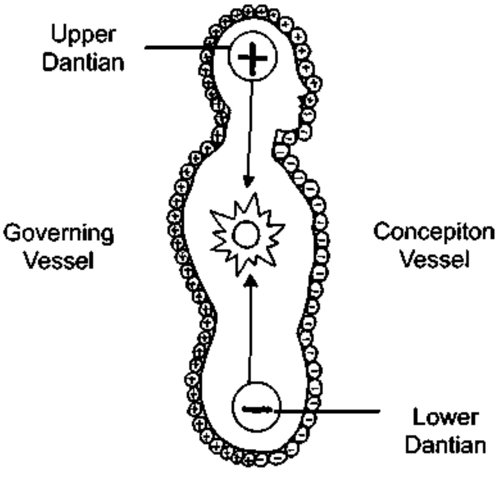

Figure 4.1. Light resonates within the body's Taiji Pole. From the body's Yang Governing Vessel flows a sea of electropositive energy. From the body's Yin Conception Vessel flows a sea of electronegative energy.

light radiating from the physical body spontaneously interacts with the mental, emotional, and spiritual light energy released from other beings.

Light energy can also be accumulated in the storage chambers of the Three Dantians. The Upper Dantian, is called the "peak of Yang energy," and is considered electropositive; it is the doorway to the chamber of Heavenly or divine light. The Lower Dantian, is called the "peak of Yin energy," and is considered electronegative, and the direct root to the Earth.

In time, with Qigong training, the postnatal Heat circulating up the Governing (Yang) and down the Conception (Yin) Vessels can awaken the Prenatal Qi stored in the brain, causing it to unite with the body's nervous system. This is considered a full integration of Yin and Yang energy and manifests as a shimmering white light glowing in the center of the Upper Dantian.

White light is stored in and released from the Upper Dantian, where the spiritual energy is stored. Once white light energy manifests, it should be drawn down the Taiji Pole into the center of the Lower Dantian. By using inner vision to focus the eyes on the center core of the body, the Qigong doctor can concentrate on the white light energy vibration and fuse this energy into one column of light, stemming from the Upper to the Lower Dantian. When the Microcosmic Orbit is completed (i.e., when the Governing and Conception Vessels are energetically fused), the first center column of light can be accessed via the body's five Center Thrusting Channels into the Taiji pole. This column of light runs through the center of the body, connecting all Three Dantians together (Figure 4.1). Accumulated Qi is stored as a golden glowing ball in the Lower Dantian.

COLOR VIBRATIONAL PROJECTIONS

Qigong doctors project light of different colors by first drawing the divine light energy into their Taiji Pole, then focusing on the light energy and releasing it either through the arms and out the hands into the patients, or through a visualization of divine light traveling from the Yin Tang (Third Eye) area into the patients.

When Qigong doctors connect with the divine healing white light energy, they absorb massive quantities of this energy into their Taiji Pole. This white light energy prisms into six color projections of light and energy vibration. As the white light energy transforms into multicolored beams of light, the colors can be either absorbed to strengthen the internal organs, or projected out for Qi emission.

The Qigong doctors use inner vision techniques to connect with the correct color for energy projection. One such technique used by doctors allows the color to naturally develop by mentally focusing both eyes to the back of the head. The color that is observed in the back of the mind is projected into their patients' body. This visualization technique is only performed after connecting with the patients' energetic field.

LIGHT THERAPY THROUGH EMITTED QI

The human body produces light which is visible to both clairvoyants and Qigong doctors (Figure 4.2). In China, doctors of TCM have discovered that light therapy is effective in treating certain diseases because of the light field's ability to penetrate the tissues, and interact with the patients' energetic fields.

Extending the image of the color is an important part of the Qigong doctor's ability to treat

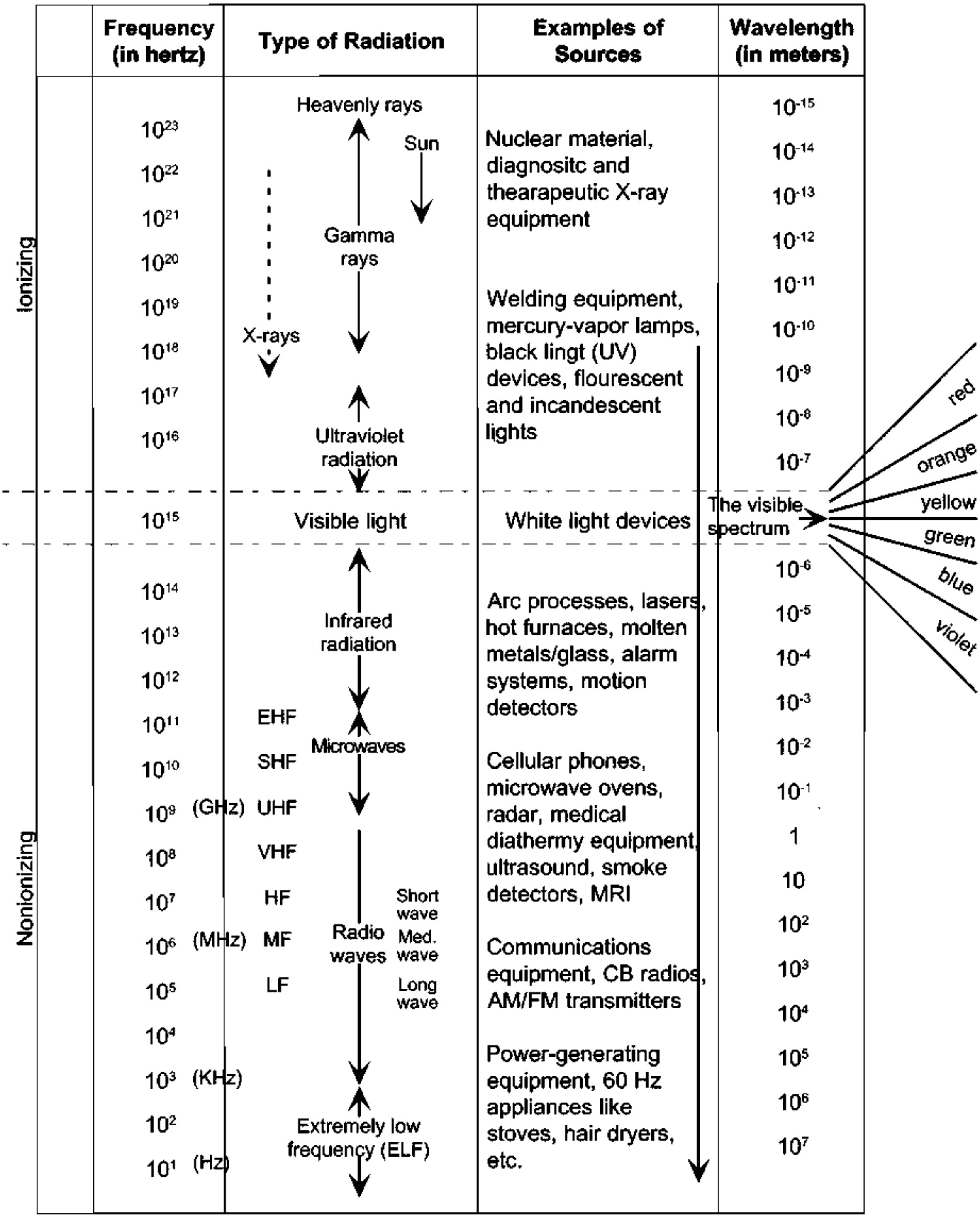
Frequency (in hertz)
Type of Radiation
Examples of Sources
Wavelength (in meters)
Ionizing
Nonionizing
10^{23}
10^{22}
10^{21}
10^{20}
10^{19}
10^{18}
10^{17}
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
10^{11}
10^{10}
10^{9} (GHz)
10^{8}
10^{7}
10^{6} (MHz)
10^{5}
10^{4}
10^{3} (KHz)
10^{2}
10^{1} (Hz)
Heavenly rays
Sun
Gamma rays
X-rays
Ultraviolet radiation
Visible light
Infrared radiation
EHF
SHF
UHF
VHF
HF
MF
LF
Microwaves
Radio waves
Short wave
Med. wave
Long wave
Extremely low frequency (ELF)
Nuclear material, diagnositc and thearapeutic X-ray equipment
Welding equipment, mercury-vapor lamps, black lingt (UV) devices, flourescent and incandescent lights
White light devices
Arc processes, lasers, hot furnaces, molten metals/glass, alarm systems, motion detectors
Cellular phones, microwave ovens, radar, medical diathermy equipment, ultrasound, smoke detectors, MRI
Communications equipment, CB radios, AM/FM transmitters
Power-generating equipment, 60 Hz appliances like stoves, hair dryers, etc.
10^{-15}
10^{-14}
10^{-13}
10^{-12}
10^{-11}
10^{-10}
10^{-9}
10^{-8}
10^{-7}
The visible spectrum
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
1
10
10^{2}
10^{3}
10^{4}
10^{5}
10^{6}
10^{7}
red
orange
yellow
green
blue
violet

Figure 4.2. Light Spectrum Chart

patients to establish a more gentle healing session.

5. The flame blue color cools the body and is beneficial in soothing and calming a patient's anxiety. Blue can be used to stimulate the Bladder, Kidneys, the reproductive system, skeletal system, and the Water Qi of the body. The color blue is also excellent for relieving inflammation and fever and can be used as a mild anesthetic.
6. The indigo or dark blue color has a deep penetrating property and is used for knitting bones and creating "energetic casts" (used for enveloping wounds and specific internal organs). The color indigo can also be used to stimulate the Bladder, Kidneys, skeletal system, reproductive system, and the Water Qi of the body.
7. The reddish-blue color of violet is known for its purifying force, which facilitates the rapid healing of difficult infections, e.g., pneumonia.
8. White energy extension indicates a fusion of all colors. White energy is the most neutral and can be used when one is not sure which other color is appropriate. White is useful for calming the mind and placing a protective field around an organ or around the body. White stimulates the Large Intestine, Lungs, the respiratory system, and the Metal Qi of the body.

USING HEALING COLOR IMAGERY

Specific color meditations can be prescribed to assist patients in their healing. The patients are encouraged to visualize specific organ colors while the Qigong doctor projects that same color into their body and energetic fields (Wood-Liver-Green, Fire-Heart-Red, Earth-Spleen-Yellow, Metal-Lungs-White, and Water-Dark Blue/Indigo-Kidneys). This treatment technique is successful for tonifying, regulating, or purging all the body's major organs and is often combined with healing sounds.

When using healing color imagery, the patients are encouraged to first visualize the diseased organ as being dull, dark, and impure in color. In cases of Liver Heat, patients might imagine a dull, turbid brownish-green, tinged with red. For a Heart imbalance, the red color may be first imagined as a dull, weak or darkish red, tinged with brownish-red or purplish-red (like the color of stagnant Blood). Patients begin the meditation by concentrating on exhaling the toxic, pathological colors. As the patients inhale, they imagine a pure, vibrant ruby-red color flowing into their body from the Heavens or Earth, energizing, cleansing, and replacing the toxic energetic color of the diseased Heart.

Note: The image of the vibrant ruby-red color is prohibited in cases of Excess Heart Fire.

EXPOSURE TO EXTERNAL COLOR FIELDS

In this type of color therapy, patients are required to sit and meditate in rooms painted in the specific color relating to their condition. A healthy color is chosen in accordance with either the Five Elements' Creative Cycle or a Controlling Cycle, depending upon the nature of the patient's disease and which particular Yin organ (or organs) are involved. One example is using the Five Elements' Creative Cycle to tonify a patient's Deficient organ. The theory of the creative cycle is to stimulate and energize the "mother" in order to strengthen its proceeding "child" organ. For a Liver imbalance, for example, the patient can be placed in a blue room for tonifying the Kidneys (the Liver's "mother" organ). The Kidney color indigo (the mother) is used to nourish the Liver organ (the child). For a Liver Excess, the Qigong doctor would use the Five Elements' Controlling Cycle and place the patient in a white room, allowing the Lungs (the grandmother) white Metal color to control the Liver Wood (the child) organ.

Also specific colors can be worn by the patient, or the patient can surround him or herself with certain natural colors to initiate healing (e.g., emerald green for Liver conditions).

FOOD AND COLOR

Another treatment method requires the presentation of food (diet) as a form of color and light therapy. In this treatment the doctor makes sure that the Five Elemental colors are present within the patient's food. Each color will start a resonance within the patient's organs (i.e., Green-Liver, Red-Heart, Yellow-Spleen, White-Lungs, and Dark Blue/Indigo-Kidneys).

When eating the food the patient focuses his or her intention on ingesting the color, light, and energy of the food being eaten.

USING DISTANCE AND RANGE

When using color projection, an important factor is the doctor's ability to maintain focused intention. If the doctor becomes distracted or looses visual concentration, the color resonance being emitted becomes degraded, and the projected color transforms back into non-differentiated white-light energy. Since it is the vibratory rate of the specific color that promotes healing, a strong connection between the doctor and patient is vital in distance color therapy projection.

WESTERN LIGHT THERAPY

Light therapy is currently being used in Western medicine through the following modalities:

- Laser-light therapy surgery,
- Full-spectrum light therapy for SAD (seasonal affective disorder),
- Color-light therapy for eye problems and various other illnesses, and
- UV-light therapy for "blue babies" with hyperbilirubinemia.

MAGNETIC ENERGY

Magnetism can influence energetic flows of energy current within the body. It is difficult to separate the energetic properties of magnetism (Yin) and electricity (Yang), as they are almost two aspects of the same energetic field. The movement of electrical currents generate magnetic fields. By establishing different magnetic polarities, the Qigong doctor can influence the electrical flow in the body.

The body's cells carry electromagnetic fields that both attract and repel each other, resulting in a magnetic pull between the body's tissues, organs, and extremities, as well as the body's electromagnetic interaction with the Earth. Clinical research in China demonstrates that the iron in the hemoglobin of the Blood is attracted to the magnetic pull of the body's tissues, as well as the magnetic pull of the Earth.

The Earth, like the body, has both a Yin (magnetic) field and a Yang (electrical) charge. These energetic fields are interdependent. The magnetic

Figure 4.4. The Earth is enveloped with electromagnetic currents, affecting both weather and vegetation.

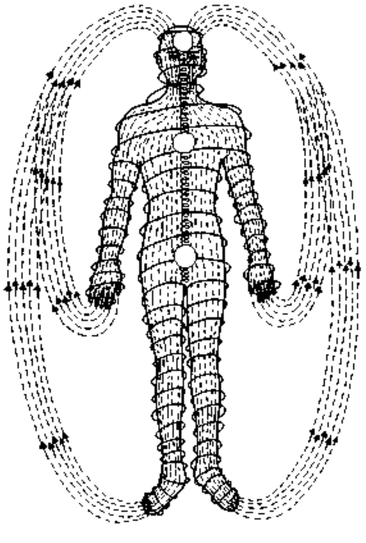

Figure 4.5. The body is enveloped with electromagnetic currents, affecting both internal and external organ functions.

field is caused by the movement of charged particles and spans the circumference of the Earth creating a strong electrical field (Figure 4.4).

Some modern researchers believe that the physical body's channel system (see Chapter 6) charges its field of magnetic flux through movement. It is further believed that the channel system creates an electrical field that attracts electrons into the body's Taiji Pole.

Richard Lee of the China Healthways Institute found that the Yin magnetic substances stored within the body's water molecules align with the electrically conducting structure of the body's channels (see Appendix 2). When the body's magnetic energy field is low, these channels become weakened and the conductivity is lowered. This magnetic energy can be replenished either from ingesting and transforming food, air and water, or gathered directly from the Earth's magnetic field.

The body's magnetic field conforms to the Earth's magnetic field through the energetic exchange within the body's electromagnetic fields (Figure 4.5). The body maintains this electromagnetic connection through the energetic resonance of the Taiji Pole. Within the structure of the body's Taiji Pole are located two major conductive poles of energy. These two conductive poles are located at the opposite ends of the Taiji Pole, positioned at the top of the head and the base of the perineum. The purpose of these energetic poles is to absorb Qi from universal and environmental fields, connecting and integrating the energy into the body's Three Dantians (see Chapter 5).

The electromagnetic lines in the body's force field begin from the at top of the head (traditionally considered the south pole) where the Heaven Qi flows into the body, and end at the at the base of the perineum (traditionally considered the north pole) where the Earth Qi flows into the body (Figure 4.6). Each of these two magnetic poles (the Lower and Upper Dantian) have a different energy influx. The energy originates and converts in the Lower Dantian and eventually flows to the Upper Dantian. The bottom pole, located in the Lower Dantian, converts Jing (Essence) into Qi (Energy) and increases the body's overall life-force energy. The upper pole, located in the Upper Dantian, converts Shen (Spirit) into perceptual insight and spiritual light.

There are several ways in which the Qigong doctor can manipulate the electromagnetic field which surrounds the body. The doctor can directly absorb Heaven and Earth energy from the bottom of his or her feet, top of the head, and through both palms (called Absorbing Qi from the Five Gates). The doctor can also, through creative visualization, gather and loop the different types of environmental energy around his or her body, increasing the thickness and power of the electro-

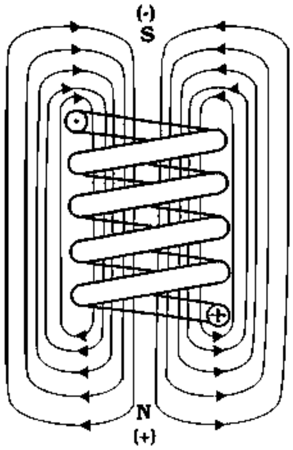

Figure 4.6. Magnetic field around a spiral coil

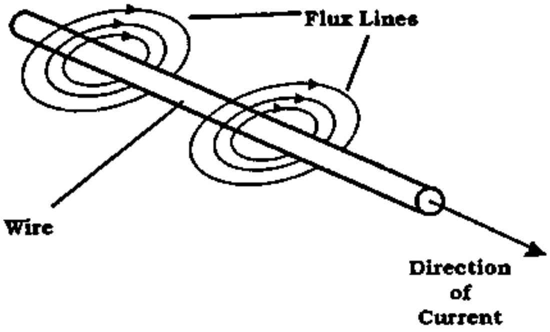

Figure 4.7. Flux lines around current-carrying conductor

magnetic field (the energetic boundary that surrounds the doctor's body). This looping action is similar to wrapping a magnet with wire to increase its magnetic field potential (Figure 4.7). Stronger lines of magnetic force will be produced around the body, thus creating a stronger "energy bubble" (Figure 4.8).

The direction of electromagnetic wrapping should be done in accordance with the natural flow of the surrounding environmental fields. Each individual should remain sensitive as to which technique works better for them, as there are individual variations used in accordance to the doctor's intention and belief.

MAGNETIC THERAPY THROUGH EMITTED QI

Qigong doctors are trained to manipulate their patient's electromagnetic field by extending emitted energy into the patient's tissues. This stimulation of the body's magnetic field is extremely effective for regulating the flow of channel energy, as well as stimulating the activity of the nervous system.

Magnetic therapy is also used as a companion to acupuncture therapy. Acupuncturists use magnetic patches, placing them on various channel points on the patient's body, to induce energy flow and tissue stimulation.

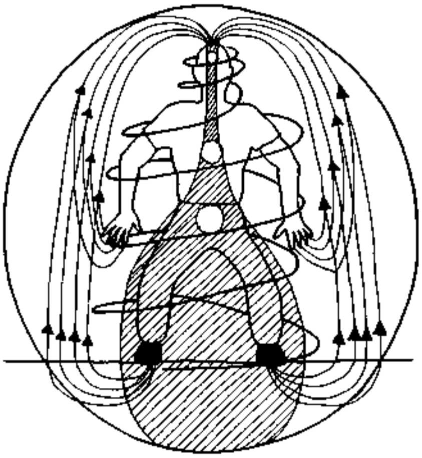

Figure 4.8. Qigong doctors can create a strong energetic field by absorbing Qi into their Lower Dantian. This energy extends into the Earth (like an anchor), and the Wei Qi field surrounding the doctor's body automatically increases and expands.

The negative north magnetic field pole of the magnet is placed in contact with the patient's skin in treating conditions due to Excess Heat and stagnation. The negative north pole is found to have a cooling and calming effect on tissues. Herbal teas used for purgation, dispersing, anti-inflammatory or sedative properties can be augmented by placing the tea on the north pole of a magnetized surface for a minimum of an hour before drinking.

The positive south magnetic field pole of the magnet is placed in contact with the patient's skin in treating Deficiencies, Cold and blockages. The positive south pole is said to have a stimulating and warming effect on tissues. Herbal teas used for warming and tonification can be augmented by placing the tea on the south pole of a magnetized surface for a minimum of an hour before drinking.

WESTERN MAGNETIC THERAPY

In 1990, Dr. Arthur Trappier submitted a paper on "Evaluating Perspectives on the Exposure Risks from Magnetic Fields" to the *Journal of the National Medical Association*. In his article, Dr. Trappier explained that exposing cancer cells to a negative magnetic field discourages the growth of the cancer, while exposing cancer cells to a positive magnetic field encourages their growth. Magnetic therapy is currently being used in Western medicine through the following clinical modalities:

- Permanent magnet therapy for localized pain and inflammation relief,
- Super-magnet therapy for stimulation of the thymus for immune enhancement in cancer patients, and
- Pulsed magnetic therapy for the treatment of arthritic joints.

GENERATED HEAT

Electricity passing through any substance will produce heat. The amount of heat that will be generated depends upon the resistance of the substance and the density of the current's flow. Heat is generated at the electron level by the friction created through molecular motion. By increasing the motion of the molecules, more heat is generated. Molecules in living organisms are in constant motion, increasing or decreasing their rate of acceleration according to the environmental temperature. In order for heat to have any effect on the body, enough Qi has to be transferred to increase the movement of the molecules.

The heat within the body causes increased electron movement, resulting in more heat being generated on a cellular level. Heat is also caused by cellular metabolism. The body's metabolism produces 75% of the energy being created in the form of heat. Cellular activity is increased through electrical, magnetic, heat, sound, and light stimulation.

From a Medical Qigong perspective, Heat in the body is generated from the accumulation of Three Fires which emanate from specific locations within the body. Within the body's chest cavity is the Heart Fire, located within the abdominal cavity is the Kidney Fire, and all through the body flows the Bladder Fire. When the Heart Fire first awakes, the Kidney Fire responds to it, and when the Kidneys Fire moves the Bladder Fire follows it.

When these Three Fires follow their normal course of energetic movement, they issue and lead the body's life-force energy, creating and sustaining life. The Three Fires are responsible for regulating the Yin and Yang energy of the body by fusing the Five Elemental energies (stored within the body's Wood, Fire, Earth, Metal, and Water organs) with the energy of the Three Dantians. As the Shen from the Heart Fire is drawn into the Lower Dantian, the Bladder Fire fuses with the Kidney Fire, creating the body's True Fire. This action causes the body's Jing to create Qi in the Lower Dantian and then transform Qi into Shen in the Middle Dantian. Once this fusion is obtained, the mind, breath and body connections all become regulated.

The Three Fires also represent the regions of vital Heat responsible for the circulation of energy that sustains the Eternal Soul. Accessing the energy of the three Fires is used for cultivation and spiritual liberation, and is brought about through Qigong practice, prayer and meditation.

1. The Heart Fire, also called the Commanding or Emperor's Fire, is located in the center of the chest. The Heart Fire is responsible for transforming the body's energy into Shen.
2. The Kidney Fire or Mingmen Fire, is located in the back of the body, just below the last floating rib. Qigong masters in ancient times regarded the Mingmen Fire as the motivating force of the body and paid special attention to its training during Qigong exercises. A Deficiency of the Mingmen Fire may lead to decreased sex-drive, hypogonadism and impotency. Conversely, if the Mingmen Fire is in Excess, increased sex-drive or sexual obsession and hypergonadism will occur.

   Dr. Zhao Xianke, an expert on medicine during the Ming Dynasty period, states that "the Mingmen Fire dominates all Twelve Primary Channels. Without it the Kidneys would be weak, the Spleen and Stomach could not digest food, the Liver and Gall Bladder would not give any energy to think or plan, the urine and feces would not be moved, and the Heart would malfunction causing dizziness and endangering life."
3. The Bladder Fire, also called the Common People's Fire, is located in the lower abdominal area by the perineum and is responsible for evaporating water.

The Three Fires are not the Triple Burners, as the purpose of the Triple Burners is to regulate the major internal organs and is a completely different energy system. The physical locations of the Three Fires however, coincides with the locations of the Triple Burners.

HEAT THERAPY THROUGH EMITTED QI

Stimulating the body's tissues through Heat therapy is extremely effective for treating Deficient and Cold syndromes. In China, Qigong doctors use the extension of Heat through their emitted field of electromagnetic energy to stimulate their patients' tissues. This therapy helps to regulate the flow of the patients' channel energy, and tonifies the Blood and Qi.

Heat therapy is also used in the form of moxibustation as a companion to acupuncture therapy. Acupuncturists will use moxa sticks and/ or cones, lighting and placing them on or over various channel points of the patients' body. This is practiced to induce Qi flow and tissue stimulation.

WESTERN HEAT THERAPY

Radiant and conductive heat therapy are currently being used in Western medicine for vasodilation and pain relief. These heat therapies can further be broken down and categorized into dry-heat therapy and moist heat therapy.

Dry-heat therapies include the following:

- Dry packs
- Hot-water bottles
- Heliotherapy-sun therapy
- Ultraviolet-heat therapy
- Infrared-heat therapy
- Diathermy therapy

Moist-heat therapies (hydrocolators) include the following:

- Hot bath packs
- Hot wet packs
- Hot foot baths
- Fomentations
- Poultices
- Vapor and paraffin baths

ELECTRICITY

A flow of electrons is called a current. Just as a current produces a magnetic field, a magnetic field, when it moves in relation to a conductor, induces an electrical current. In Medical Qigong therapy, the body's channels are also considered electrical circuits, and the points existing within each channel can be considered booster amplifiers (or step-up transformers) that maintain the current's strength. According to Dr. Robert O. Becker, research scientist and author of the book *Cross Currents*, acupuncture needles have the capacity to act as antennae, drawing charged particles (ions) from the atmosphere into the body. The acupuncture needle delivers a low-level electrical stimulation to the channel points and can be used to charge up, or decrease the energetic potential of these "step-up transformers," affecting the current along the channel.

The human body is an excellent conductor of electricity and contains both alternating and direct electrical currents. The alternating current is responsible for the transformation of the body's magnetic field into the body's electrical field. The direct current is responsible for the body's positive Yang and negative Yin flow of electrical charges.

The friction produced by rubbing the feet on a carpet creates a charge that is stored within the body until touching another conductor (metal, another person, etc.) to release the charge. This normal and common occurrence demonstrates the storage, conductance, and discharging of electrons from the body. Any time electrons travel, heat, or thermal radiation is produced.

According to Swedish Physician and medical research scientist Dr. Bjorn Nordenstrom, the biological circuits of the body are driven by the accumulated charges, which, unlike a battery, oscillate between positive and negative. The body's system of channels and blood vessels act as insulated cables, while the blood plasma acts as the conductor. In the permeable tissue, the fluid between the cells conducts ions. A key component of the body's electrical circuit is the natural electrodes in the capillary walls, known as the lipid bilayers of the electron transport chain.

A myriad of electrical forces work within the body. Every human thought and action is accompanied by the conduction of electrical signals along the fibers of the nervous system. In fact, life would not exist at all without a constant flow of ions across the membranes of cells.

The electromagnetic energy in the body's cells is continuously being generated through the biochemical transformation of food, and air, and is circulated by the electromagnetic fields being generated within the tissues. Perineural cells, or nerve sheaths, carry the direct current of the body's electricity. These cells are responsible for motivating the body to heal, regenerate, and repair itself. Healing is always affected by a change within the body's electromagnetic field. The rate and efficiency of healing is based on the strength and polarity of the body's field of energy. This electricity is one of the primary energy sources responsible for maintaining life itself.

Dr. Becker's research demonstrates that bones are "piezoelectric": when stressed, mechanical energy is converted into electrical energy that produces an increased electrical current. Dr. Becker discovered that running a minute electrical current through a fractured bone will stimulate the reproduction of the cells, creating a healing current similar to the body's natural healing mechanism.

According to Richard Lee of China Healthways Institute, electrostatic waves are observed in the body through EEGs (electro-encephalograms) and EMGs (electro-myograms). Different types of therapies affect the EEGs (sound therapy, meditation therapy, self-regulation Qigong therapy, light therapy, etc.). By controlling their thoughts, Qigong doctors can affect the current in their body and thus affect the EEG measurements. These thought patterns will in turn influence the Qigong doctors' electrical field. Experiments show that EEG measurements in test subjects receiving energy from healers tend to synchronize. Qigong doctors, when treating patients, can produce voltages as high as 190 volts, 100,000 times greater than regular EEG voltages.

The body's natural ability to gather, store, and move Qi increases the body's abilities to gather, store, and move electrical charges. The gathering of these electrical charges can cause a gradual electric buildup within the tissues and internal organs, developing into an Excess Qi condition within the body. Excess Qi conditions often manifest mood swings and are responsible for the energetic circuit overloads that occur spontaneously within the body. Excess Qi manifests as a sudden "explosion" or "release" of emotions such as rage, fear, grief, worry, fright, anxiety, and joy. An excess electromagnetic charge, internally combined with the Qi from a patient's on-going emotions can be drawn into an organ's tissue area and absorbed by the internal organ that is in the most elevated state of energy conversion. This energetic reaction can cause the already overcharged internal organ to spontaneously release emotions. Sometimes the release is physiological, e.g., sudden sweating, blushing, twitching or jerking, yawning, stretching, sighing, burping, passing gas, etc. These can all be ways that the organ seeks to regain hemostasis.

ELECTRICAL THERAPY THROUGH EMITTED QI

The idea that an electric current can stimulate bodily repair, alert defence mechanisms, and control the growth and function of cells is not a new concept in Western medicine. In fact, the use of bio-electromagnetism dates back at least 200 years. Electrotherapy is found to be very useful in relieving pain by signaling the brain to activate and alter the body's neurochemicals. The insertion of electrically stimulating needles into a patient's body, for example, causes a release of endorphins, and is used for pain management.

Medical Qigong, acupuncture, and Chinese massage likewise, stimulate the peripheral and cutaneous nerves that carry sensory information, via the spinal cord, to the brain. This stimulation of the cutaneous nerves activates the brain's opiates (endorphins - endogenously generated morphine) and facilitates the closure of the body's pain relay gates, killing pain, and is the basis for anesthesia. Because medical Qigong therapy also has an analgesic affect on the body's cutaneous tissues, it is being increasingly used in hospitals for pre- and postoperative procedures.

WESTERN ELECTROTHERAPY

Electrotherapy is currently being used in Western medicine through the following modalities:

- Giga-TENS therapy for stimulation of healing,
- TENS therapy for pain relief,
- CES-cranial electro-stimulation therapy-for depression and substance abuse, etc., and
- Micro-stimulation therapy for micro-current stimulation below the threshold of awareness, to stimulate nonspecific healing, the reduction of inflammation and the harmonization of tissue polarity.

CHAPTER 5
THE THREE DANTIANS AND THE TAIJI POLE

THE THREE DANTIANS

According to Chinese energetic physiology, humans have three important energy centers, located in the center core of the body, that store collected energy in much the same way as a battery does. These three centers are called the Three Dantians (Figure 5.1). The word Dan literally translates to mean "cinnabar," while the word tian means "field." In terms of Medical Qigong tian is referred to as a field of energy. Together, the characters for Dantian literally translates as "cinnabar field" or "field of elixir."

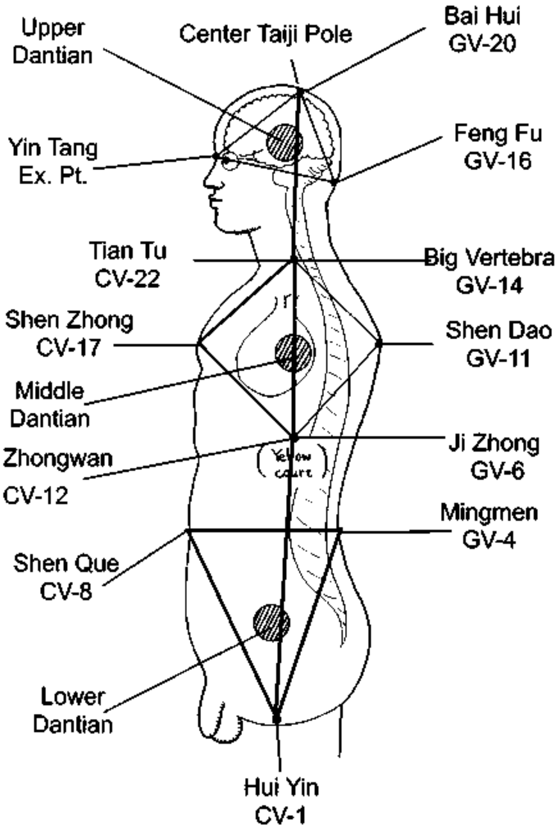

Figure 5.1. The Anatomical Locations of the Three Dantians and Center Core (Taiji Pole)

Cinnabar is mercury sulfide, a bright red or vermillion mineral, which is the principal ore of mercury. Cinnabar was an important mineral in ancient China, the source material needed to manufacture vermillion ink, which could be used only by the emperors. Cinnabar was -- and is today -- also used in Chinese medicine to sedate the Heart and calm the Shen, but always in small doses for short periods due to its highly toxic nature.

Cinnabar was also a vital elixir in Daoist alchemy, as it was discovered to be very balanced in its Yin and Yang properties. Daoist alchemy, like Western alchemy, was practiced in two ways: outer alchemy (Wai Dan) and inner alchemy (Nei Dan). Outer alchemy was the ancestor of modern chemistry. Outer alchemists set up laboratories and experimented with many substances from mineral, animal, and plant sources with the goal of discovering how to turn base metals into gold. Secretly, they were also seeking to discover an elixir, or drug, that would confer immortality or, at the least, greater longevity. Along the way they made many important discoveries, such as gunpowder, medicines, and many other substances vital to the world today.

Inner alchemy was concerned with purifying human nature and transforming the spirit into its most pure and radiant potential without the use of outer agents. Instead, they used Qigong and meditation to circulate and gather the "inner elixirs" of sexual energy, Qi, and consciousness at various locations within the body. The Three Dantians, or "elixir fields," in addition to storing energy, act like an alchemist's crucible, or cauldrons in the role of gathering and transforming vital substances, energies, and various elements of awareness. The ingredients of the inner elixir are Jing, Qi, and Shen. The inner alchemists kept their work secret by using mineral code words such as "gold," "lead," and "cinnabar" to describe the movement and transformation of energetic and spiritual substances within the body.

The goal of the inner alchemists is immortality, a complete transformation of Jing, Qi, and Shen. Jing, Qi, and Shen are collectively called the Three Treasures and are the three fundamental energies necessary for human life. To accomplish this transformation, alchemists first gather and transform Jing into Qi in the Lower Dantian. They then gather and transform Qi into Shen in the Middle Dantian. Next, they transform Shen into Wuji (the absolute openness of infinite space) in the Upper Dantian. Finally, they merge Wuji into Dao (divine energy). The Three Dantians serve as the inner crucibles for these transformations.

These transformations can be compared to the changes of water (the consistencies and not the temperature), which when heated, can change from solid ice, to a liquid, and then to vapor.

The Three Treasures (Jing, Qi, and Shen) are also connected with the Three Outer Forces or Powers known as Heaven, Earth, and Man. Jing (reproductive essence) is the most substantial and therefore the most Yin of the three, and it is closely linked with Earth Qi. In Medical Qigong practice and Daoist inner alchemy, the Earth energy is gathered in the Lower Dantian and is associated with heat. Qi is a mixture of Yin and Yang and is closely connected with the atmospheric energy, which is a blend of Heaven and Earth Qi and gathered into the Middle Dantian, (which is associated with vibration). Shen (Spirit) is the most insubstantial and, therefore, the most Yang of the three; it corresponds with Heaven Qi, and is gathered in the Upper Dantian which is associated with light.

The Three Dantians are connected to each other through the Taiji Pole. The Taiji Pole acts as a passageway for communication between the Three Dantians, and as a highway for the movement of the various life-force energies. The Eternal Soul is drawn into the body at the moment of conception through the Taiji Pole, and departs through it at death. The Taiji Pole also serves as a transport for the body's Hun.

Regardless of whether or not a modern Qigong practitioner is concerned with spiritual transformation, from a Medical Qigong standpoint the Three Dantians are still vital centers for the cultivation of energy. They are important areas for diagnosis and self-healing, as well as for projecting Qi to patients.

The following is an exploration of the Three Dantians and their individual relationships to the Three Treasures of Jing, Qi, and Shen.

THE LOWER FIELD OF ELIXIR

The Lower Dantian is the Dantian most familiar to martial artists and Zen meditators. It is regarded as the center of physical strength and the source of stamina. Called the "Hara" in Japanese, it is located in the lower abdomen, in the center of the triangle formed by drawing a line between the navel, Mingmen (lower back), and perineum. These three points form a pyramid facing downward. This configuration allows the Lower Dantian to gather the energy from the Earth.

The Lower Dantian is the major storage area for the various types of Kidney energies. The Kidney energies are, in turn, closely linked with our prenatal energies and provide the foundation for all other types of Jing, Qi, Yin, and Yang energies in the body.

The Lower Dantian is connected to the first level of Wei Qi (Protective Qi) circulating outside the body, extending from the body's tissues to about one inch. As the Lower Dantian fills with Qi, the Wei Qi field naturally becomes thicker.

THE LOWER DANTIAN AND JING

The Lower Dantian collects Earth energy and represents the body's physical energy of Jing. The Earth energy that is transformed in the Lower Dantian is a dense, full energy with a thick quality to its texture. In our analogy with the transformations of water, the energy in the Lower Dantian is dense, like ice.

The Lower Dantian is closely linked to the Jing Gong (Essence Palace) located in the perineum, which serves as a reservoir of Jing. Our Prenatal Essence (Yuan Jing), determines our constitutional strengths and vitality, and is stored in the Lower Dantian. It interacts with the Kidney energies to form Kidney Jing. The Kidney energies are all closely intertwined: Kidney Jing, Kidney Qi, Kidney Yin, Kidney Yang, and Kidney Fire. The Mingmen Fire, also called Kidney Yang, helps transform the Jing into steam (Kidney Qi).

Kidney Jing circulates throughout the body via the Eight Extraordinary Vessels, in particular, the Governing, Conception, and Thrusting Vessels, all of which originate in the Lower Dantian.

Kidney Jing controls the reproductive energies and life cycles in the body. Some of the ancient alchemical texts describe the Lower Dantian in women as being located in the Bao or uterus, and it is related to the function of Jing in a woman's body. In men, the reproductive essence is located in the Jing Gong (Essence Palace) or prostate and seminal vesicles. In some Chinese medical, Daoist, and Qigong literature, the term Kidneys is used as a synonym for the testes and ovaries.

The location of the Jing Gong is affected by the different anatomical locations of the male and female reproductive organs. In men, this area is located in the center of the body, at the level of the superior border of the pubic bone, posterior to the Qugu CV-2 (Crooked Bone) point. The Jing Gong area in women is located higher, centered in the uterus, about an inch above the superior border of the pubic bone, posterior to the Zhongji CV-3 (Utmost Center) point.

This difference in location, in turn, affects the storage of Jing (i.e., the testicles in the male cause the transformation of energy to occur lower in the body than in females, due to the higher position of the woman's ovaries).

Jing is the most physical, material form of Qi within the body (corresponding to Yin and Earth energy). The Lower Dantian is the place where Qi of the Earth is drawn into the body and transformed by heat.

The Lower Dantian acts as a reservoir for heat and energy and is associated with the Kidneys. The Kidneys control the Water element in the body. Jing is said to be like the water in the cauldron. Through focused concentration and meditation, the Jing (Essence) in the Lower Dantian is refined and transformed to produce Qi (Energy). When heat is generated as a result of the Heart and Mingmen Fire mixing with the Kidney Water, the alchemical transformation of Jing in the Lower Dantian area transforms Jing into the steam of Qi (the character for Qi is composed of steam rising from a cooking pot). This alchemical transformation is known as "changing Jing into Qi," and takes place within the Lower Dantian.

THE LOWER DANTIAN AND QI

The Lower Dantian is often called The Sea of Qi. It is the place where Qi is housed, the body's Mingmen Fire is aroused, the Kidney Yin and Yang Qi is gathered, and the Yuan Qi is stored. Also called Source Qi, the Yuan Qi is the foundation of all the other types of Qi in the body. The Yuan Qi is closely linked with the Prenatal Essence (Yuan Jing). Together, the Yuan Qi and Yuan Jing determine our overall health, vitality, stamina, and life span.

The Yuan Qi is the force behind the activity of all of the organs and energies in the body. It is closely related to the Mingmen and works to provide body heat. The body's Yuan Qi is the catalytic agent for transforming the food we eat and the air we breath into Postnatal Qi. It also facilitates the production of Blood.

Yuan Qi is housed in the Lower Dantian, and it also flows out to the internal organs and channels through the Triple Burners. Yuan Qi is said to enter the Twelve Primary Channels (the body's twelve major energy pathways) through the Yuan points (sometimes called Source points) in acupuncture theory.

Of the Three Dantians, the Lower Dantian is closest to the Earth, is the most Yin, and has the strongest ability for gathering Earth Qi. In Medical Qigong, once students have learned to conserve and circulate their own Qi, they can increase it by connecting to the unlimited reservoirs of Qi in the natural environment. Earth energy is the first form of external Qi with which the Qigong practitioner connects. This energetic connection to the Earth is important for several reasons.

1. First, Qigong practitioners need the Yin grounding power of the Earth Qi to counterbalance the gathered active Yang energy from Qigong exercises. Without grounding in Earth Qi, many Qigong practitioners develop Qi deviations in the form of Excess Heat.
2. Second, each person's supply of Qi is limited. When Qigong doctors extend their Qi to heal others, they deplete their personal supply of Qi unless they are able to simultaneously replenish their supply from outside sources.

Even people who do not practice Medical Qigong draw Earth Qi into their Lower Dantians. By practicing Qigong and using conscious intent, the amount of Earth Qi drawn into the body is vastly increased.

THE LOWER DANTIAN AND SHEN

The Lower Dantian itself is considered a center of consciousness. The consciousness of the Lower Dantian is more kinesthetic or physical given its Yin nature and close proximity to the Earth.

The body's Jing is connected with the Po (Seven Corporeal Souls). The Po control our survival instinct and the subconscious physical reflexes associated with survival. For this reason, Oriental martial artists spend many hours cultivating their lower elixir field to have the integration of Jing, Qi, and Shen needed for the split-second clarity of focus demanded in life-and-death struggles.

THE LOWER DANTIAN AND KINETIC COMMUNICATION

Not only is the Lower Dantian the center of physical strength and the source of stamina, but it is also considered the "house" of physical (kinetic) communication, awareness and feelings. The level of awareness referred to as "the intuition of the physical body," or kinetic communication, is stimulated by the subconscious. The subconscious mind picks up many signals from the environment that are not processed by the logical mind. These signals can motivate spontaneous body movements and responses and are sometimes referred to as gut feelings.

Kinetic means "characterized by movement." It is usually felt as a movement in the body or of the body. It is in this kinetic state of awareness that allows Qigong doctors to naturally feel the patient's internal resonant vibrations within their own body. When the doctors' body suddenly feels hot or cold, starts shaking or trembling, this may indicate that their subconscious mind is trying to communicate the location and condition of the diseased area within the patient's tissues.

Most of the time, the feelings experienced in the Lower Dantian are very subtle, but Qigong doctors, with a heightened degree of awareness of their own body, are able to pick up subtle variations of energetic shifts within themselves and others. When doctors collect energy in the Lower Dantian, an increased awareness and sensitivity naturally occurs. Cultivating this ability simply requires practice in paying attention to the physical body. A high level of awareness of the physical body, the surrounding environment, and the relationship between the two is required to maximize kinetic communication. When awareness is increased, perceptual feeling and kinetic body movements happen naturally. These are subtle senses that allow Qigong doctors to feel, smell, or hear energetic phenomena as they are released from the diseased tissues of patients.

According to research conducted by Dr. Michael Gershon, a professor of anatomy and cell biology at Columbia Presbyterian Medical Center in New York, the Lower Dantian sends and receives impulses, records experiences, and responds to emotions. Its nerve cells are bathed in and influenced by the same neurotransmitters as the brain. The Lower Dantian's "brain," known as the enteric (intestinal) nervous system, mirrors the body's central nervous system and is a network of 100 million neurons (more then the spinal cord contains), neurotransmitters, and proteins that can act independently of the body's brain, and can send messages, learn, remember, and produce feelings.

Dr. Gershon explains that active within the neural system of the lower abdominal area are major neurotransmitters like serotonin, dopamine, glutamate, norepinephrine, nitric oxide, enkephalins (one type of natural opiate), and benzodiazepines (psychoactive chemicals that relieve anxiety). The lower abdomen also has two dozen small brain proteins called neuropeptides. Dr. Gershon's research results provide modern scientific verification of what Eastern wisdom has taught for millennia -- that centers of consciousness exist at places in the body besides the brain and that the abdomen is one of the body's major centers of awareness.

ANATOMICAL LOCATION OF THE LOWER DANTIAN

The Lower Dantian is centered below the umbilicus, inside the lower abdomen, forming a downward pointing triangle. It occupies the first three lower Chakra gates. Chakras are small energy centers that originate from the Taiji Pole and have one or more gates.

1. The lowest point of the Lower Dantian extends to the Huiyin CV-1 (Meeting of Yin) point at the perineum. The name refers to the area on the body responsible for gathering and absorbing the Earth energy. This area is responsible for gathering the Yin energy into the body and Lower Dantian area via the three Yin leg channels (Liver, Spleen, and Kidney). This area is sometimes known as the Lower Gate of the Taiji Pole, or the Bottom Gate of the Lower Chakra.
2. The front area of the Lower Dantian is located posterior to the Shenque CV-8 (Spirit's Palace) point at the navel. The name refers to the place where the mother's Qi and Shen enter the embryo during fetal development. This area is sometimes known as the Front Dantian, or the Front Gate of the Second Chakra.
3. The back area of the Lower Dantian is located at the Mingmen GV-4 (Gate of Life) point on the lower back, anterior to the second lumbar vertebra. The Mingmen occupies the place between both Kidneys. It is the root of Yuan Qi, and therefore determines life and death. The Mingmen provides the body's True Fire; it supplies the heat for the Triple Burners, and is responsible for stabilizing the Kidneys and Lower Dantian area. This area is sometimes known as the Back Dantian, or the Back Gate of the Second Chakra.
4. Medical Qigong schools in China differ in their belief as to where the center of the Lower Dantian is located. Some schools teach that the center of the Lower Dantian is affected by the different anatomical locations of the male and female reproductive organs. In these particular schools, the students are taught that in men, the center of the Lower Dantian is located posterior to the Guanyuan CV-4 (Gate of Original Qi) point. The center of the Dantian area in a women is said to be located internally higher, and posterior to the Qihai CV-6 (Sea of Qi) point. This area is sometimes called the Middle of the Dantian, referring to its position between the navel and Mingmen areas.

THE NINE CHAMBERS OF THE LOWER DANTIAN

The human body is viewed as a microcosmic replica of the power of Heaven. Just as Heaven is said to be divided into nine different levels, each containing various palaces, our bodies also are said to contain a large variety of palaces and chambers. The nine stars of the Big Dipper are said to correspond to the Nine Chambers of each of the Three Dantians.

The functional aspects of the body's psyche were described by ancient Qigong masters as "spirits" that lived within the nine chambers of the Lower Dantian. These "spirits" linked the body's energetic channels and vital internal organs into an organic harmony of life-force energy. Qigong masters believed that as each Dantian became energized, it would initiate specific reactions within the body's energetic system, causing certain energetic/spiritual awareness to manifest within the practitioner's psyche.

Each of the nine chambers is several inches in diameter and is numbered in accordance to its en-

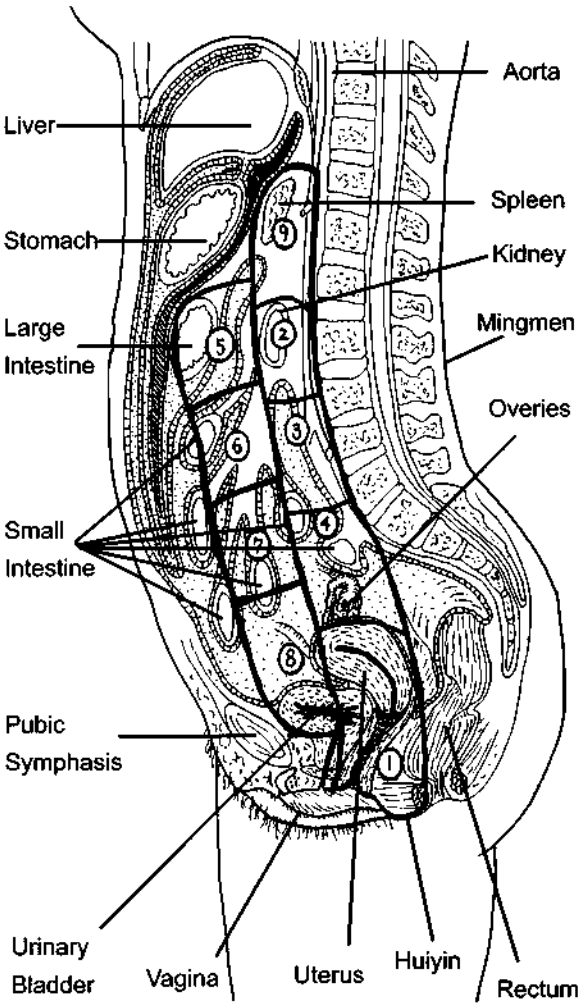
Liver
Aorta
Spleen
Stomach
Kidney
Large Intestine
Mingmen
Overies
Small Intestine
Pubic Symphasis
Urinary Bladder
Vagina
Uterus
Huiyin
Rectum
1
2
3
4
5
6
7
8
9

Figure 5.2A. The Nine Chambers of the Lower Dantian are shown here in the female body. Each number encompasses the entire chamber.

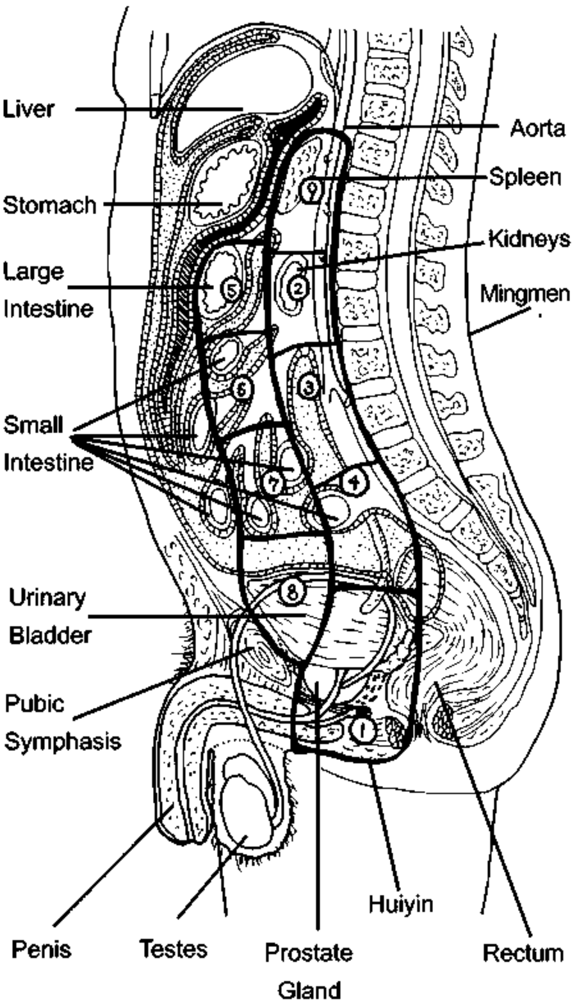
Liver
Aorta
Spleen
Stomach
Kidneys
Large Intestine
Mingmen
Small Intestine
Urinary Bladder
Pubic Symphasis
Penis
Testes
Prostate Gland
Huiyin
Rectum
1
2
3
4
5
6
7
8
9

Figure 5.2B. The Nine Chambers of the Lower Dantian are shown here in the male body. Each number encompasses the entire chamber.

ergetic stimulation (Figure 5.2 A-B). This vertical abdominal set of Lower Dantian chambers relates to many cavities of the body's internal viscera. The Nine Chambers of the Lower Dantian are named as follows.

1. The Palace of Jade (Jade Stem or Jade Cave)
2. The Official Health Monitor (Kidneys)
3. The Minister of the Orchard Terrace
4. The Chamber of Moving Pearls
5. The Minister of House Cleaning (Large Intestine)
6. The Palace of Mystical Spirits (Small Intestine)
7. The Chamber of Mysterious Elixir
8. The Spirit of the Jade Court (Bladder)
9. The Spirit of the Yellow Court (Spleen)

ENERGETIC FOCUS OF TRAINING THE LOWER DANTIAN

All Qigong training begins with the focus on the Lower Dantian. In the beginning stages of Medical Qigong training, the doctor will encourage students to focus their mind and breath on the Lower Dantian. The purpose of this training is to gather the body's Yuan Qi into the Lower Dantian (called "returning to the source"), to strengthen the foundational root for the body's energy.

It is dangerous for Medical Qigong students to bypass the discipline of Lower Dantian cultivation training to progress more quickly to the more advanced intuitive and psychic training of the Upper Dantian. Such an approach to training may lead to Qi deviations (see Chapter 18) and cause emotional instability.

THE MIDDLE FIELD OF ELIXIR

The Heart is the primary organ related to the Middle Dantian; the second organ is the Lungs. In Medical Qigong, the thymus gland is also of primary importance to the Middle Dantian.

In children, the thymus gland is quite large. As the child matures into adulthood, the thymus gland shrinks in size. Until recently, Western biologists thought that the thymus gland became vestigial and inactive in adults. Beginning in the 1980's, however, with the advent of the AIDS epidemic and the increase in cancer cases, intensive new research was launched into the immune system. As a result, scientists discovered that the thymus gland plays a major role in educating and maturing the white blood cells to become immunocompetent. This thymus function continues throughout one's life.

The Middle Dantian collects Qi and represents the body's energetic reservoir for mental and emotional vibrations and energy. The energy of man that is transformed in the Middle Dantian has a fluid quality—like water.

A refining process also takes place in the Middle Dantian, transforming the fluid energy into more steam-like energy that is then transferred to the Upper Dantian. The Middle Dantian transforms Qi into Shen by bringing the transformed Qi into the Heart Fire. This alchemical process is commonly called "changing Qi into Shen" and refers to kinetic energy transforming into spiritual consciousness.

The Middle Dantian is connected to the second level of Wei Qi, circulating about one-and-a-half feet outside the body. As the Middle Dantian fills with Qi, the colors of the student's middle field of Wei Qi change, becoming even more pronounced. The reason for this change is that the Middle Dantian is connected to the Five Agents, which in turn govern the Five Yin Organs and organ emotions. As the students begin to experience various stresses and emotional releases their aura (resonating from the internal organs, throughout the second energetic field) changes its colors.

THE MIDDLE DANTIAN AND JING

The Heart is related to the Fire element. The Heart derives its Yang Fire from the Kidneys. Modern research in Chinese medicine equates the function of the adrenal glands to the traditional function of Kidney Yang. In Western physiology, the adrenal glands help to regulate the pace of the heart.

To keep the Heart Fire in balance, the Heart also needs Yin. Heart Yin is derived from Kidney Yin (Jing is one aspect of Kidney Yin).

In traditional Chinese physiology, the Heart is said to govern the Blood. Not only is the Heart responsible for the circulation of Blood, but in the Chinese view, the transformation of Gu Qi into Blood takes place in the Heart. Food Qi (energy derived from the consumption and transformation of food matter) is a form of postnatal Jing, derived from the Spleen and Stomach. Blood is composed of Nutritive Qi (Ying Qi), Jing and Fluids (see Chapter 22). The Kidneys also send prenatal Kidney Jing to the Heart to make Blood. Therefore, Jing - particularly postnatal Jing - is vital to the Heart's function of governing Blood.

THE MIDDLE DANTIAN AND QI

Similar to the Lower Dantian, the Middle Dantian is also considered to be a Sea of Qi. The Qi of the Middle Dantian is called Zong Qi. Zong Qi is translated as Ancestral Qi, Gathering Qi, Genetic Qi, or Essential Qi. In English translation, it is sometimes confused with the Original Qi (Yuan Qi), but they are not the same. The Zong Qi is a form of postnatal Qi, whereas the Yuan Qi is housed in the

Lower Dantian and is a form of prenatal Qi. Zong Qi and the Yuan Qi assist each other to maintain the healthy function of the Heart and Lungs.

The Zong Qi nourishes both the Heart and Lungs, controls the speech and the strength of the voice, and interacts with the Kidneys to aid in respiration. According to Traditional Chinese Medicine, the Kidneys assist the Lungs in holding and stabilizing the breath when inhaling.

Qi and Blood are closely related. In Chinese medicine, it is often said, "Qi is the master of Blood; Blood is the mother of Qi." Qi gives the Heart and blood vessels the strength to circulate Blood and gives life to the Blood. Blood, on the other hand, houses and carries Qi to all the cells in the body. When one loses Blood, one also loses Qi. Therefore, one should consider Qi and Blood to be inseparable.

Qi is also inseparable from the mind and spirit. According to the teachings of ancient Tibetan Qigong masters, the channels are, metaphorically speaking, the road, the Qi is the horse, and the mind is the rider. Through refining the Qi, the mind and spirit are refined and purified. The Middle Dantian is the main focal point for this refinement of Qi into spirit.

THE MIDDLE DANTIAN AND SHEN

Classically, the Chinese locate the Mind in the Heart. In Chinese, the word for "mind" (Xin) is also the Chinese word for Heart. In Medical Qigong, a distinction is made between the Acquired Mind (Ren Xin) and the Original Mind (Yuan Xin).

The Middle Dantian is said to house the Shen and control all of the other functions of Shen that are attributed to the other Yin organs. Thus the Heart is often referred to as the Heavenly Emperor.

Throughout the world, people relate the heart to emotions and feelings. Emotions and feelings are one important aspect of the spirit. Any type of emotion will have an effect on the Shen.

The negative emotions are sometimes called "the five thieves" because even though negative emotions are necessary for life, chronic states of negative emotions drain the Qi.

The human mind easily falls under the influence of the Po (the Seven Corporeal Soul), that are concerned with survival. When the Po dominate the Heart, their overexaggerated self-concern gives rise to a chronic state of fear, sadness, worry, anger, and defensive arrogance.

The redeeming virtue of the Heart is a sense of propriety and discriminating awareness. The Hun (the Three Ethereal Souls) control the smooth flow of Qi throughout the body and are nourished by the Five Virtues of kindness, order, trust, integrity, and wisdom. These Five Virtues give peace and clarity to the Heart and allow the higher qualities of the Yuan Shen to overrule the Po.

An important relationship regarding the Middle Dantian and Shen is found in the Heart's role of governing the Blood. The ancient classics state that the Shen also resides in the Blood and pervades the body through Blood circulation. This relationship between Blood and Shen is one reason why anemic patients are often restless and suffer from insomnia. Through nourishing the Heart Blood, many forms of spiritual unrest can be treated.

According to Dr. Candace Pert's information on neurotransmitters (stated in *Psychoneuro Immunology*), the brain and white blood cells both contain the same neurotransmitters and biochemical constitutes that are prerequisite for consciousness awareness to exist. These same neurotransmitters and biochemical constitutes which are linked to consciousness are synthesized and created by the white blood cells. This similarity indicates that not only do the brain and abdomen have their own consciousness and nervous system but so does the Blood. This similarity also implies that consciousness is possible anywhere in the body, substantiating the ancient Chinese understanding that consciousness is pervasive throughout the body via the Shen, which resides in the Blood.

THE MIDDLE DANTIAN AND EMPATHIC COMMUNICATION

The Middle Dantian is also considered the "house" of emotional (empathic) communication, awareness, and feelings. Emotional communication is analogous to the feelings of empathy felt within the Heart. Empathy is the means by which the Qigong doctor will most frequently get in touch with the emotional components of the patient's energetic blocks and imbalances.

Empathic communication is felt as an emotion and originates in the Heart and Middle Dantian area. When Qigong doctors focus on the Middle Dantian area, a line of communication is created with their higher self. We are all born with this ability, but as we grow older, we tend to override this type of emotional communication with our logical mind. These impressions slowly atrophy, eventually causing us to lose this natural empathic ability of communication. We disconnect from this higher perception due to the negative and double messages received from our parents and from society. The way to reconnect with the intuitive self is to look inward and become one with our true self, which is connected to the divine.

ANATOMICAL LOCATION OF THE MIDDLE DANTIAN

The Middle Dantian is shaped like a tetrahedron (four points): one area points toward the Upper Dantian and the Heavens, one area points toward the Lower Dantian and the Earth, one area points toward the front, and one area points to the back .

1. The front lower point of the Middle Dantian is located at the Zhongwan CV-12 (Middle Stomach Cavity) point on the midline of the abdomen, just below the sternum. This area is the master point of the Middle Burner and is sometimes known as the Front Gate of the Third Chakra.

   This emotional storage area is also called the Yellow Court because it reflects the emotions stored from the Heart, (the Heart was sometimes referred to as Suspended Gold in ancient China).

   The back lower point of the Middle Dantian is located on the Jinzhong GV-6 (Middle of the Spine) point. The name refers to the point's location. This area is sometimes known as the Back Gate of the Third Chakra.
2. The front center point of the Middle Dantian is located at the Shanzhong CV-17 (Central Altar) point, on the middle of the sternum at the level of the fourth intercostal space. The name refers to the "place of worship" where the Shen resides. This area is sometimes known as the Front Gate of the Fourth Chakra.

   The back center point is located two inches up from the shoulder blades, at the Shendao GV-11 (Spirit Path) point located at the hollow between the fifth and sixth thoracic vertebrae. The name refers to the easy accessibility into the patient's Shen residence (within the Heart) through this particular area on the back. This area is sometimes known as the Back Gate of the Fourth Chakra.
3. The upper front point of the Middle Dantian is located at the Tiantu CV-22 (Heaven's Chimney) point, at the throat. The name refers to the visceral cavity area at the base of the throat responsible for "pooling" escaped Heaven Qi from the Lungs. This area is sometimes known as the Front Gate of the Fifth Chakra.

   The upper back point is located on the Dazhui GV-14 (Big Vertebra) point on the back. The name refers to the point's location below the seventh cervical vertebra and above the first thoracic vertebra. This area is sometimes known as the Back Gate of the Fifth Chakra.
4. The center of this Middle Dantian area is located in the right atrium of the heart, centered between the SA (sinoatrial) and the AV (atrioventricular) nodes. The center of the Heart is considered the seat of all emotions.

THE NINE CHAMBERS OF THE MIDDLE DANTIAN

The Middle Dantian regulates the body's Heart—Mind connection. The Heart-Mind connection includes both the all pervasive consciousness of the entire body and the Shen. It distributes the flow of Yuan Qi into the Yin and Yang aspects of the body's mental and emotional functions. These functions within the psyche are controlled by the interaction of the body's Wu Jing-Shen or Five Essence Spirits (Hun, Po, Zhi, Yi, and Shen). The energy of the Middle Dantian is also

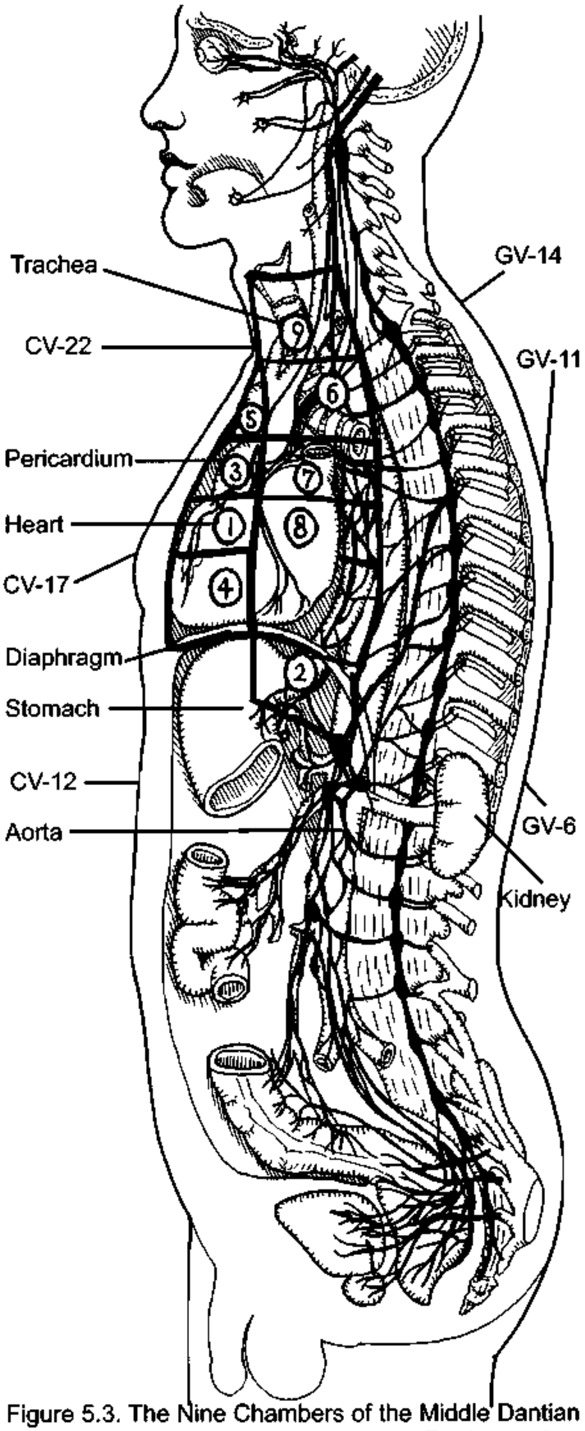

Figure 5.3. The Nine Chambers of the Middle Dantian are identical for men and women. Each number encompasses the entire chamber.

associated with the Spleen and Stomach energies. The vertical thoracic set of nine chambers (Figure 5.3), located within the Middle Dantian, originate around the auricles and ventricles of the heart (especially the pericardial and pleural cavities). The Nine Chambers of the Middle Dantian are named as follows.

1. The Chamber of Mysterious Elixir
2. The Lower Court of the Heart
3. The Chamber of the Ultimate
4. The Chamber of Splendor
5. The Heaven's Cover
6. The Twelve Storied Chamber
7. The Chamber of Government
8. The Purple Chamber
9. The Heaven's Chimney

ENERGETIC FOCUS OF TRAINING THE MIDDLE DANTIAN

In Medical Qigong training the doctors encourage their students to focus their mind and breath on the Middle Dantian to regulate the Heart. Techniques are used for treating Deficient conditions by drawing Qi into the Heart and Middle Dantian area, then regulating the body's energetic fields. For treating Excess conditions, the students are encouraged to lead and purge the Excess Qi from the Heart and Middle Dantian area through the body's extremities. The purpose of this training is to release the toxic Excess Qi gathered in the patient's Heart and Yellow Court areas. These areas of Excess tend to cause disease.

THE UPPER FIELD OF ELIXIR

The Upper Dantian is the collector of Heaven Qi and represents the spiritual aspect of man and his connection to the divine. The Heaven energy that is transformed in the Upper Dantian is of a thinner more ethereal quality—like vapor.

The Upper Dantian is connected to the third level of Wei Qi, circulating several feet outside the body. As the Upper Dantian fills with Qi, the spiritual intuition and psychic perception of the individual increases.

In Chinese physiology, the brain controls memory, concentration, sight, hearing, touch, and smell. These senses stay in close communication with the Heart and Shen.

THE UPPER DANTIAN AND JING

The Jing and Qi form the material foundation for the Shen. In Chinese, the term Jing-Shen means mind or consciousness. Jing-Shen may also mean vigor, vitality, or drive. In China both Western and Chinese doctors use the term Jing-Shen Bing to refer to all types of mental illness. The term Shen is rarely used by itself in a medical context without the concept being connected to the close relationship between the power of the mind and spirit.

The term Wu Jing Shen is used in Medical Qigong to describe the body's Five Essence Spirits (Hun, Po, Zhi, Yi and Shen). These five spirits combine the energetic essence of the Five Yin Organs in order to create the body's innate spiritual consciousness.

The Jing itself is considered the basis for, and ruler of Marrow (a substance derived from the Kidneys that nourishes the brain, spinal cord and forms bone marrow). The brain, is one of the six Extraordinary Organs and is called The Sea of Marrow, and is considered to be a form of Marrow. The six Extraordinary Organs are hollow Yang organs that store Yin Jing. Deficiency of Jing may lead to poor concentration, poor memory, dizziness, and absentmindedness. Deficiency of Prenatal Jing is related to mental retardation and attention deficit disorder (ADD) in children.

In some styles of Qigong, the Jing is intentionally conserved and its energy is drawn upwards from the Lower Dantian through the spine to nourish the brain. Such nourishment benefits the mind and spiritual consciousness.

THE UPPER DANTIAN AND QI

The head is the most Yang part of the body since it is the closest part of the body to Heaven. The Qi that operates in the Upper Dantian is, therefore, Yang in nature. The Spleen and Kidneys send the Clear Yang Qi (pure, light, and insubstantial) upwards to the brain to facilitate mental clarity and activity.

The Upper Dantian is also the place where we connect with the Yang Qi of Heaven. Qigong practitioners consciously absorb Heavenly Qi through the upper doorway, Baihui GV-20. The Heaven Qi is composed of the Qi from the celestial bodies: the sun, moon, planets, and stars.

The Upper Dantian is located in the center of the brain, in an area that encompasses the pineal, pituitary, thalamus, and hypothalamus glands. The pineal gland and hypothalamus have been shown to be extremely sensitive to the influence of light. In his book, *The Body Electric*, Dr. Robert Becker cites experiments with bees and several species of birds that indicated that these animals navigate by the light of the sun. Birds have disproportionately large pineal glands. He also discovered that birds seem to have a backup system of navigation based upon a sensitivity to the electromagnetic fields of the earth.

There are three forms of Qi to which the brain responds and interacts: light, electricity, and magnetism. These particular forms of energy stimulate the pineal, pituitary, thalamus, and hypothalamus glands, affecting and influencing the emotional responses.

THE UPPER DANTIAN AND SHEN

We have already discussed the relationship of the Jing and the Shen with the Upper Dantian. Of particular interest to the alchemists is the opening of the center of the Upper Dantian, called the Crystal Room, where psychic perceptions and intuitive knowing take place. Higher communications, a sense of great bliss, and perceptions that transcend the seeming bonds of time and space are associated with the Upper Dantian. These experiences are particularly valuable to Qigong doctors, who use these heightened perceptions to diagnose illness. Their efficacy is well documented in the works of such authors as C. Norman Shealy and Caroline Myss, who use the term "medical intuitive" to describe this paranormal ability.

The Upper Dantian is also the place where the Eternal Soul connects with the Wuji, and with the Dao. The awareness associated with this union is beyond description, as its unity supersedes the differentiation inherent in conceptual thought.

Although the Upper Dantian is responsible for intuitive and psychic perceptions, the combined energetic properties of all three Dantians establishes the foundation for all psychic perceptions. The steam-like quality of the energy with the Upper Dantian fuses within the light now existing in this center. As this energy disperses, it travels out into the Wuji, returning back to the divine. This interaction is also responsible for what the Chinese call "receiving the message," which describes the ability of connecting with the patients' subtle energy field to acquire the hidden knowledge stored within the patients' tissues.

While in a state of tranquillity and inner peace, the Qigong doctors' Upper Dantian will also intuitively process information from the environment and universe. This intuitive knowledge provides Qigong doctors with a greater ability to explore both their own consciousness and the patients'. The ancients called this ability "knowing without knowing."

THE UPPER DANTIAN AND INTUITIVE COMMUNICATION

As the Shen is developed and the Upper Dantian is opened, spiritual communications may reveal themselves in a flash of an image or a vision in the mind's eye. These images and visions are sometimes very abstract and short lived. Interpreting images takes practice because the images streaming from the Yuan Shen cannot be interpreted easily by the logical mind.

The Qigong doctors must be able to distinguish between true and false messages reflected through their visions. True visions are received from the divine connection to the Dao or Wuji; false visions reflect the messages from the subconscious. The ability to accurately separate these visions is an another example of "knowing without knowing."

Although communication from within usually makes itself felt as a strong impulse, Qigong doctors must learn to keep the logical mind from interfering by practicing spiritual meditations. These meditations involve the forging of links in the chain of communication with the higher self. They should be practiced again and again until this connection becomes a natural, recurring phenomenon. The more one practices at stilling the logical mind and circumventing the ego, the easier it will be to get a clear communication from the higher self. When the higher self initiates a communication, it does not demand or impose itself, and if ignored it will usually stop sending signals.

Five Elements must be in place before the communication lines become fully open and operational:

1. Purity in intention,
2. No hidden agendas,
3. Surrender to the divine will,
4. Complete trust and faith in success (believe and expect), and
5. A quiet and receptive stillness of mind.

The lines of communication with the higher self are severed by the logical mind through doubt, fear, and disbelief. Strong faith is required to open this line of communication. Faith requires no logical proof; if proof is needed, then doubts interfere and breed failure. The logical mind cannot know absolute faith and, therefore, faith must come from deep within one's true self. Any form of cynicism will lead to the death of spiritual growth, for it strikes at the root of faith itself.

Faith is not something that can be forced. Even after practicing Medical Qigong for many years, Qigong doctors may still have to battle their own questions. However, through successful practice, the seed of faith is established, allowing it to grow and blossom. The opposite of faith is doubt combined with fear. Suppression and denial of fear builds and armors the ego, which leads to further pain. If the fears are accepted and acknowledged, they can be overcome by faith.

ANATOMICAL LOCATION OF THE UPPER DANTIAN

The Upper Dantian is located in the head, approximately three inches posterior to the Yintang point (between the eyebrows). It is shaped like a pyramid, facing upward to gather the energy from Heaven. This pyramidal reservoir houses light.

1. The front point of the Upper Dantian is the Yintang (Hall of Impression) point. The name refers to the ancient tradition of placing a red mark or "seal" over the Bright Hall, or Entrance of the Spirit, this point represents wisdom and enlightenment. This area is sometimes known as the Front Gate of the Sixth Chakra, or the Third Eye point.

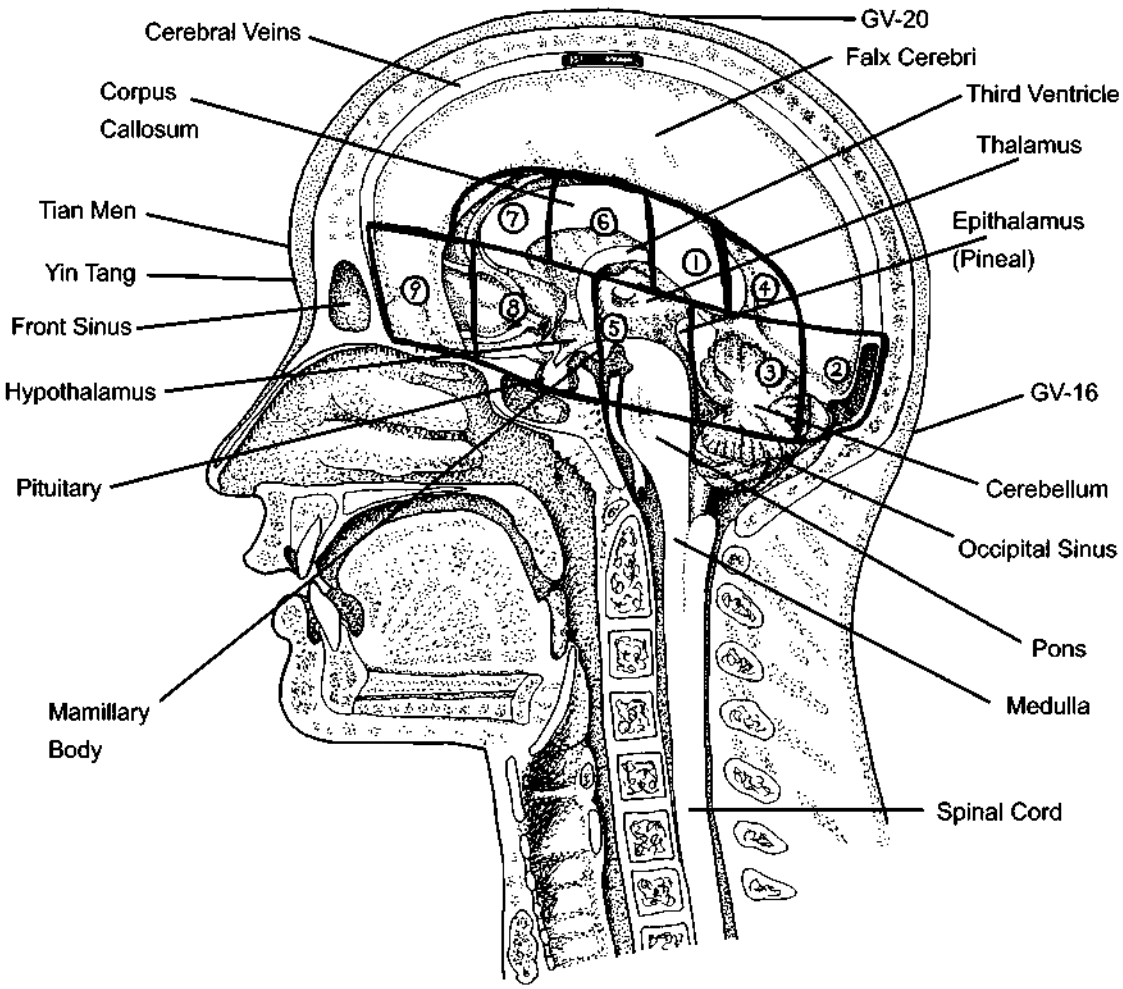
Cerebral Veins
Corpus Callosum
Tian Men
Yin Tang
Front Sinus
Hypothalamus
Pituitary
Mamillary Body
GV-20
Falx Cerebri
Third Ventricle
Thalamus
Epithalamus (Pineal)
GV-16
Cerebellum
Occipital Sinus
Pons
Medulla
Spinal Cord
①
②
③
④
⑤
⑥
⑦
⑧
⑨

Figure 5.4. The Nine Chambers of the Upper Dantian are portrayed in this drawing. Each number encompasses the entire chamber.

2. The back point of the Upper Dantian is located below the external occipital protuberance, on the Fengfu GV-16 (Wind Palace) point. This area is also connected with and surrounded on both sides by the Bl-10 (Heavenly Pillar) point. The energy field connected to this point may be likened to an antennae receiving messages; it allows Qigong doctors to regulate their state of consciousness. The Governing Vessel 16 point is also a Sea of Marrow point, which is used to affect the flow of Qi and Blood to the brain, and a Window of Heaven Point (one of eleven points used for treating Shen disturbances), as well as one of the thirteen points identified by the famous Daoist physician Sun Simiao as a Ghost Point (points used for treating spirit possession). It has been my personal observation that students with a more prominent occipital protuberance tend to see auras more easily and receive psychic intuition faster. This area is sometimes known as the Back Gate of the Sixth Chakra.
3. The highest point of the Upper Dantian is lo-

cated on the vertex of the crown, on the Baihui GV-20 (One Hundred Meetings) point. The name refers to the ancient tradition that one receives divine messages and spiritual intuition from this point into the Upper Dantian's Chamber of Mysterious Elixir, located within the third ventricle of the brain. This area is sometimes known as the Upper Gate of the Taiji Pole or the Upper Gate of the Seventh Chakra.

It is also said in Traditional Chinese Medicine that all of the body's major channels send a connecting vessel to the Baihui so that at death the Qi can leave the body through this upper doorway and ascend to the Heavenly realms.

4. The center of the Upper Dantian is located in the pineal gland. It is the area where the Shen transcends and merges with the Wuji and then progresses on to the Dao. The Upper Dantian is also considered the house of spiritual (intuitive) communication, awareness, and feelings.

THE NINE CHAMBERS OF THE UPPER DANTIAN

The horizontal cephalic set of nine chambers, located within the Upper Dantian (Figure 5.4) is based on the different ventricles of the brain. The Nine Chambers of the Upper Dantian are named as follows.

1. The Chamber of Mysterious Elixir (Mystical Medicine)
2. The Palace of Jade Emperor
3. The Palace of Moving Pearls
4. The Chamber of Splendor
5. The Hall of the Upper Dantian (Medicine Field)
6. The Chamber of the Ultimate Truth
7. The Cover of Heaven (Palace of the Heavenly Court)
8. The Chamber of Government
9. The Entrance of the Spirit (Bright Hall)

ENERGETIC FOCUS OF THE UPPER DANTIAN

In Medical Qigong training, the Upper Dantian pertains to gathering spiritual intuition and light to advance the doctor's intuitive and psychic ability (see Six Transportations of Shen, Chapter 13).

The Qigong doctor may also absorb universal and environmental Qi into the Upper Dantian through the Yin Tang (Third Eye region) and the Tian Men areas (located in the center of the forehead), and emit healing Qi to patients through either one or both of these points. The Shen can both exit and enter the body from the area in the Upper Dantian by way of the Baihui, Yin Tang, and Tian Men (Heavenly Gate) areas.

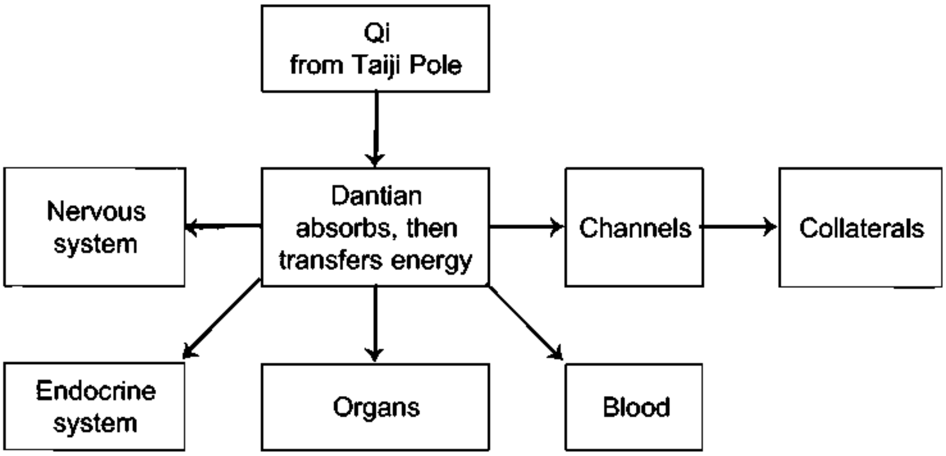

Figure 5.5. Qi Dispersion From the Center Taiji Pole Through the Dantians into the Body

THE FUNCTIONAL ASPECTS OF THE THREE DANTIANS

The three energy reservoirs of the Dantians are linked externally through the Governing and Conception Vessels and internally through the Thrusting Vessels and the Taiji Pole.

The center of each Dantian is penetrated by and attached to the Taiji Pole, which exits from the Baihui at the top of the head to the Huiyin point at the perineum. Each of the Dantians corresponds to anatomical locations that are centers for magnetic and electrical vibration and charge. The body's intensity of vibration and charge is dependent on the mental intention, the posture, and the respiration.

Qi moves into the body's Dantians through the body's Taiji Pole. The energy is then absorbed into the body's major organs and surrounding tissues as it flows out of the Dantians and into the body's internal and external channels and collaterals.

Each Dantian acts like a reservoir, collecting energy and redistributing it into all of the internal organs. This energy projects through the surface of the body into the Wei Qi field. The same energy also projects inward flowing through the energy channels, along the nervous system and endocrine glands, and then through the Blood, to nourish the body (Figure 5.5).

You can visualize this energy transformation as follows: Qi flows into the body like rainwater flowing into a lake (the body absorbing and collecting Qi into the Dantians). The rainwater is then absorbed into the surrounding soil, foliage, and root systems (skin, tissues, and cells) before it gathers, collects, and pools into deep artesian wells (the Dantians). Pressure begins to build up as these artesian wells fill with the rainwater, and eventually overflow, pouring into smaller pools (the organs) before combining with the rushing action of underground streams (the channels).

Another popular analogy is to consider the Dantians as batteries, the body's Taiji Pole as a magnetic bar connecting the batteries together, the channels as the wires, and the Wei Qi fields as the electromagnetic fields manifesting from the energy contained within the structure.

Mental and emotional awareness of a specific tissue area can be heightened through increasing the flow of energy to that location. When energy fills the tissues, a cellular release of energy causes the tissues to either store or release emotions, depending on the body's overall Excess or Deficient Qi.

If an increase of Qi is focused into the Lower Dantian, the result is a more powerful and stable feeling in the body. When an increased amount of Qi is focused into the Middle Dantian, the result

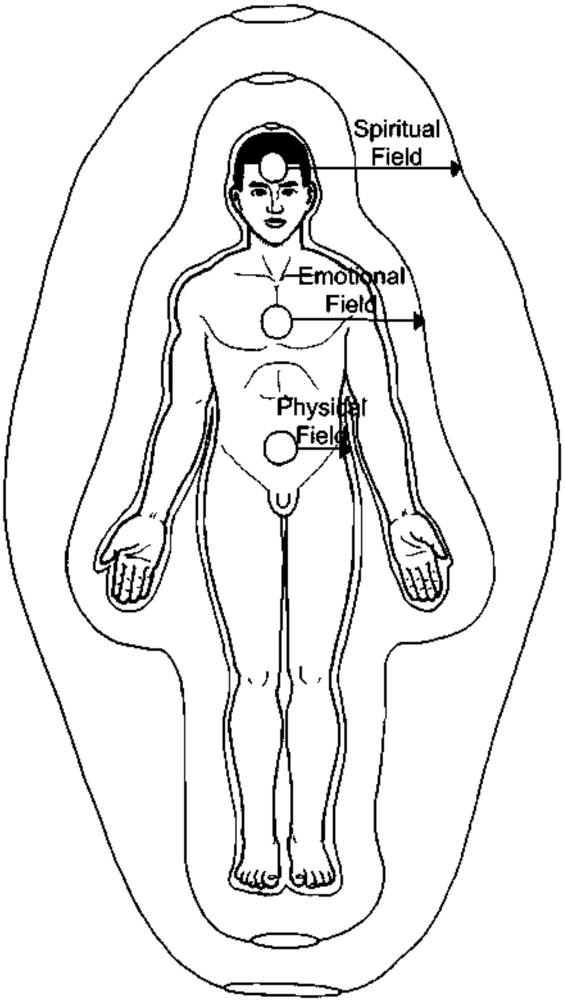

Figure 5.6. The Three Dantians and Their Relationship to the Three External Wei Qi Fields

is a heightened feeling of emotional perception and awareness. If an increase of Qi is focused into the Upper Dantian, a heightened spiritual awareness and sense of connection to the divine occurs. The greater these energies flow, the healthier the individual becomes, and the stronger his or her energetic fields become (Figure 5.6).

THE DANTIAN'S YIN AND YANG ENERGETIC CHAMBERS

Each of the Three Dantians can be divided into Yin and Yang energetic chambers. The Yang chambers relate to each Dantian's upper chambers and the spiritual aspects of the Hun. The Yin chambers relate to each Dantian's lower chambers and the carnal aspects of the Po. The chambers and their energetic potential are described as follows.

THE LOWER DANTIAN

1. The Yang Energetic Chamber of the Lower Dantian relates physically to the upper quadrant of the abdomen, located within the small intestine in men and the uterus in women. When influenced by the Hun, quiescence and serenity are enhanced.
2. The Yin Energetic Chamber of the Lower Dantian relates physically to the lower quadrant of the abdomen, which includes the reproductive organs and the Bladder, urethra, and anus. This chamber is responsible for reproduction and sexuality. When influenced by the body's Po, raw physical power and sexual vitality are aroused.

THE MIDDLE DANTIAN

1. The Yang Energetic Chamber of the Middle Dantian relates physically to the upper quadrant of the Heart, which includes the left and right atrium, the upper portions of the left and right ventricles, and the atrioventricular node. This chamber is responsible for spiritual attitudes and virtues related to the influence of the body's Hun. The Hun are responsible for:
   - gathering and transmitting divine inspirations and spiritual insights,
   - giving and receiving unconditional love, and
   - motivating spiritual growth through prayer, devotion, and commitment.
2. The Yin Energetic Chamber of the Middle Dantian relates physically to the lower quadrant of the Heart, which includes the left and right ventricles. This chamber is responsible for sensual passions, conquests, and conditional love. It deals with actions which are often based on hidden agendas. These emotions are related to the influence of the body's Po. The influence of both the Po and the Yin Energetic Chamber of the Upper Dantian activates, energizes and enhances:
   - biological drives (for food and self preservation,
   - sexuality, lust and desire, and
   - sensuality.

THE UPPER DANTIAN

1. The Yang Energetic Chamber of the Upper Dantian relates physically to the upper quadrant of the brain, often referred to as the third ventricle or higher brain centers. When stimulated, the higher brain centers manifest spiritual intuition and divine insight, related to the influence of the Hun. Such insight is responsible for spiritual growth and maturation. Insights eventually lead to the emergence of extrasensory perceptions (ESP), e.g., clairvoyance, clairaudience, telepathy, psychokinesis, and spiritual enlightenment.
2. The Yin Energetic Chamber of the Upper Dantian relates physically to the lower quadrant of the brain, often referred to as the "reptilian" brain. When stimulated, the reptilian brain activates the thalamus, hypothalamus, cerebellum, and cortex, which awaken the body's intuition, as well as animalistic and primordial instincts of survival. When the Yin Energetic Chamber dominates, the sensory, animalistic nature of the Po emerges full force. This phenomenon sometimes occurs when coma patients begin to recover. As energy begins to fill the lower chambers of the Three Dantians, their initial reactions are basic and carnal in nature, i.e., to either engage in sexual activity or strike out in violence. As the Qi begins to fill the upper chambers of the Three Dantians, the energy balances in the patient's

Taiji Pole and the patients' impulses return to normal.

THE DOCTOR'S PROJECTED AURA FIELDS

In China, Qigong doctors are tested and categorized according to the predominant color of their Qi emissions. The healing color will depend upon which of the three Dantians is dominant (corresponding to either the Lower Dantian, Middle Dantian, or Upper Dantian), and the energetic strength and potential of the individual being tested. The emitted color is observed in the visible-light spectrum projected from the Qigong doctor's hands.

YELLOW EMITTED COLOR

Individuals who have just begun their training will emit a yellow glow around their external energy field. The Qi will naturally overflow from the Lower Dantian area and the color yellow will dominate the aura. This is actually a normal aura color observed from most beginning Qigong doctors, indicating that the energy field is still too weak to treat serious diseases.

RED EMITTED COLOR

Qigong doctors who have mastered their body and mind will emit a red radiant glow around their external energy field. The Qi will also overflow from the Lower Dantian area, and the color red will dominate the aura. This is actually a normal range of aura color observed from the many Qigong doctors who have just graduated from Medical Qigong colleges, indicating that the energy field is at the intermediate stage.

PURPLE EMITTED COLOR

Qigong doctors who have evolved to the next level of mastering the emotions will emit a purple radiant glow around their energy field. These doctors' Qi will naturally overflow from the Middle Dantian area, and the color purple will dominate the aura. This aura color is observed in strong and advanced Qigong doctors.

BLUE EMITTED COLOR

Qigong doctors who have further refined their energy will emit a blue radiant glow around their energy field. These doctors' Qi will also overflow from the Middle Dantian area, and the color blue will dominate the aura. This aura color is observed in very advance Qigong doctors, and is developed only after many years of clinical practice and personal spiritual transformations.

WHITE EMITTED COLOR

The Qigong doctors who have mastered their spiritual life, as well as surrendered their body, mind, emotions, and spirit to a divine or higher calling will emit a white radiant glow around their energy field. These doctors' Qi will naturally overflow from the Upper Dantian area, and the color of the aura is considered the strongest and most powerful.

CLEAR-OPALESCENT EMITTED COLOR

The divine energy is emitted through a clear-opalescent color. This Qi emission is different then the white radiant glow, and is a sign of divine intervention and healing. These times of divine intervention can occur in any stage of the doctor's energetic development, and are a testimony to the healing virtue of the divine.

Although every Qigong doctor uses a combination of all Three Dantians when projecting Qi into patients, the color of the aura surrounding the doctor's body reflects which Dantian's reservoir is predominantly used. Through time, patience, and much practice, the Qigong doctor will be able to transform from a yellow to red, purple, blue, and finally to white-light healer.

THE TAIJI POLE

The Taiji Pole can be observed as a vertical column (or pole) of brilliant white light, full of vibration and energetic pulsation. This energetic channel flows from the Baihui point at the top of the head, through the center core of the body, connecting the Three Dantians. The Taiji Pole roots its energy at the base of the perineum at the Huiyin area.

In Chinese energetic embryology, as the father's sperm enters the mother's egg, it creates the upper or Heavenly vortex point of the Taiji Pole. At this stage of energetic development, the Heavenly vortex creates and connects the Three

Dantians together, rooting the column of white light energy into the Kidneys and Mingmen area. As the fetus continues to grow and develop, the Taiji Pole slowly descends, shifting its energetic root to the base of the perineum. After birth, the Taiji Pole can be observed as a vertical column of brilliant white light, surrounded by a veil of golden spiraling light.

HEAVEN, EARTH, AND MAN RELATIONSHIPS

The Taiji Pole of the body is comparable to the central axis of the Earth, which is known as the Taiji Pole of the Earth. At the end of the Earth's Taiji Pole are the North and South Poles, which are compared to the body's Baihui at the top of the head and Huiyin at the base of the perineum. The Earth's Taiji Pole (or central axis) aligns to the North Star.

The North Star is seen as the Taiji Pole of Heaven and is sometimes called the Pole Star or Taiji. It is sometimes said that the original Chinese character for Taiji means Yin and Yang. While the original meaning of the Chinese character for Yin is the shady side of the mountain, and Yang depicts the sunny side of the mountain, the mountain's peak is considered a Taiji, where both Yin and Yang meet (or the center of Yin and Yang).

The center Taiji is also called the Still Point, because although all of the changes of Yin and Yang occur in the Heavens within the course of a day, and all of the stars seem to make a 360 degree rotation (as the Earth rotates) the Pole Star remains stationary as Heaven's Still Point. In a similar manner, the body has the Taiji Pole as the "still point of man." A great stillness is experienced when a Qigong practitioner gathers Qi into the Taiji Pole.

INTERNAL ENERGY INTERACTIONS

Extending from the body's Taiji Pole are the Three Dantians, Five Jing-Shen, and the Eight Extraordinary Vessels. These vital energetic systems feed all the major organs and the body's Twelve Primary Channels. As the Three Dantians absorb energy from the body's Taiji Pole, they also extend, or "feed," and energize the body's Yin and Yang organs, Twelve Primary Channels, and Eight Extraordinary Vessels. Each individual Dantian is connected to the body's Taiji Pole and is responsible for creating its own alchemical transition in the process of converting Jing into Qi, Qi into Shen, Shen into Wuji, and Wuji back to the Dao (see Chapter 13).

In my clinical experience, the closer I work to the patient's Taiji Pole the more powerful the energies become. Accessing the patient's Taiji Pole, Three Dantians, and Eight Extraordinary Vessels for whole body tonification and regulation is faster and less draining on my energetic reservoirs then emitting Qi into the patient's Twelve Primary Channels to accomplish the same type of treatment.

THE FIVE THRUSTING CHANNELS

Five internal Thrusting Channels connect to and transverse the body's center Taiji Pole. These five internal channels begin at the perineum and spiral up through the center of the body's core, flowing into the upper crown of the head. They connect the Baihui area at the top of the head to the Huiyin area at the bottom of the perineum. These five internal channels absorb the Five Elemental energies of Heaven and Earth (Wood, Fire, Metal, Earth, and Water) into the body.

The Taiji Pole is responsible for absorbing the energy from Heaven and Earth and distributing the collected Qi into the body's major internal organs. Similar to the Earth's central axis, the Taiji Pole is divided into two main energetic polarities: The Five Portals of the Heavenly Yang Gate, located at the top of the head, and the Five Portals of the Earthly Gate, located at the perineum (Figure 5.7).

When the body's internal and external energetic wave patterns become synchronized through prayer or meditation, a rhythmic pulsation occurs within the Taiji Pole. This pulse begins deep within the center core of the body and vibrates outward towards infinite space connecting the Wuji with the Dao.

SPIRITUAL MANIFESTATIONS

When doctors penetrate the outside veil surrounding the center core of a patient's Taiji Pole and extend their intention into the light's center, they experience the sensation of falling into space

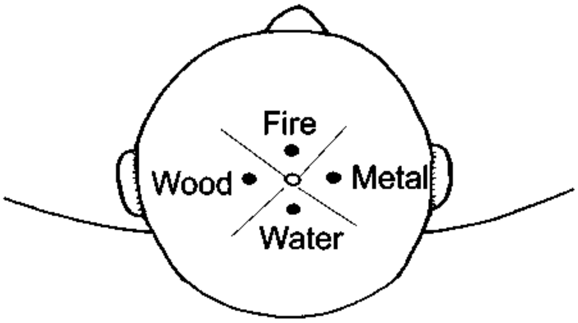

Top of Head

Figure 5.7. The Five Portals of the Heavenly Yang Gate are portrayed here as seen looking down at the top of the head. The Universal Energy is absorbed into the Taiji Pole through the Five Portals of the Heavenly Yang Gates, located at the center of the Baihui (One Hundred Meetings) and Sishencong (Four Spirits Hearing) areas, at the top of the head.

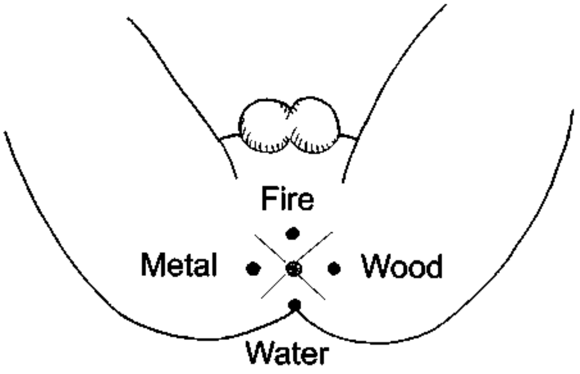

The Five Portals of the Earthly Yin Gate as seen from the perineum. The environmental energy is absorbed into the Taiji Pole through the Five Portals of the Earthly Yin Gates, located at the center of the anal sphincter, between the Huiyin (Meeting of Yin) and the Changqiang (Long Strength) areas, at the base of the perineum.

or shooting through space. This sensation is followed by seeing flashing colors and moving shapes, as the surrounding core dissolves into infinite space itself. As the Qigong doctors continue to extend their intention to reach this area, time and space feels as if they stretch into eternity. This experience is considered the true connection with the Wuji returning back to the Dao. The potential for this experience exists within everyone and can be accessed through deep prayer and meditation. It is quite a normal phenomenon within deep spiritual practices and is sometimes known as accessing the river of God.

ENERGETICALLY OPENING THE TAIJI POLE

When we are born, we are connected to great spiritual wisdom and power through our center core. During the maturation process, this connection slowly fades away, yielding to masks and defense mechanisms, that seek to shield us from our pain and self-hatred (shame). In childhood, every time we repressed the flow of painful feelings connected to an event, we froze that particular event in both energy and time, and locked it within our muscles. Since the external Wei Qi is composed of energy consciousness, a block of frozen psychic energy is formed within the Wei Qi at the moment we inhibit the pain. By walling off our wounds, we also wall off our connection to our deeper core self (Figure 5.8). We pretend not to feel the pain anymore, we block off the memory of the event. Thus, we prevent a natural resolution to the pain and the trauma of the event. Through denying our true feelings we disconnect from our core self and put on masks. We smile when angry, for example, or pretend indifference to our hurts. Since creative inspiration comes from our core, we lose our creative ability. When we stop the negative experiences of anger, pain, or fear, we also stop the positive healing of the physical, mental, and emotional aspects of the experience.

During Medical Qigong treatments, patients regress layer by layer through the pain and fear associated with the blocked energy that gives rise to the disease. Even though the memories may be very strong and frightening, as the emotions

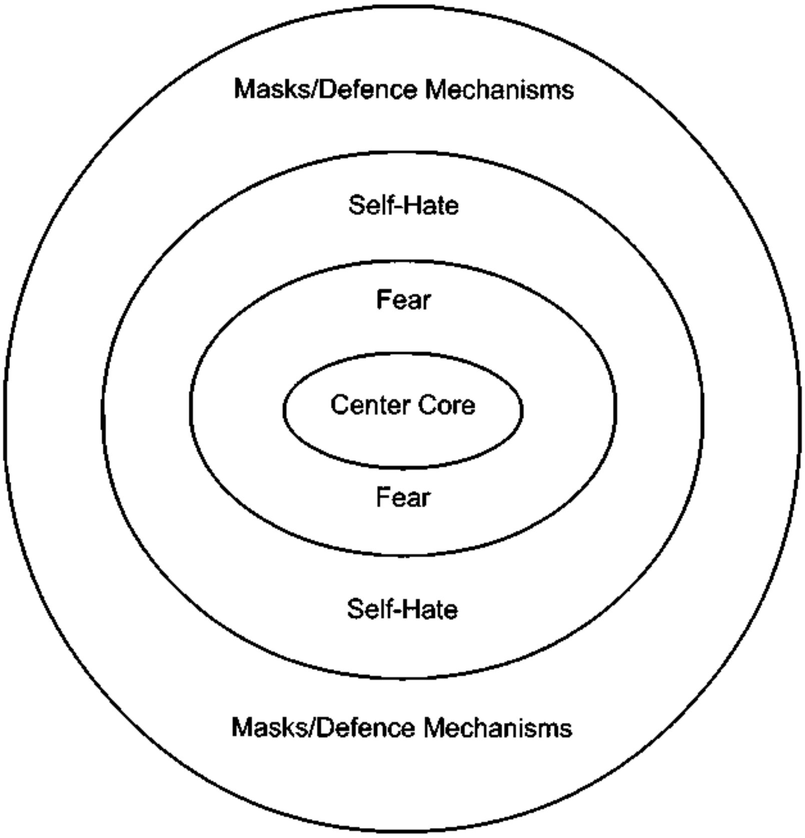

Figure 5.8. The True Self is Contained Within the Core Self

release, the pain decreases with the dispersal of the original trauma.

I have found in my clinical practice that most of the patients' pain comes not from the original trauma but from the unconscious belief system established to protect them from the original trauma. In other words, more pain and illness is created in our bodies by avoiding working through the original trauma (through our habitual defense and avoidance patterns) than was present in the original trauma. It requires enormous amounts of energy to suppress feelings, and each time we do so, we create further injuries to ourselves.

SPIRITUAL AWAKENING

When patients become aware of their spiritual connection to the divine, it is called an "awakening." This change can be observed through the spiritual transitions experienced within the patients' center core through their Taiji Pole. When the center Taiji Pole (which connects all three Dantians) begins to awaken, certain physical, emotional, and energetic reactions may occur within the patients' body. These transformations occur because the tissues of the energetic fascia (that connect to the major organs of the body) are enveloped within the energetic structure and webbing of the Taiji Pole.

THE TWELVE GATES OF THE CHAKRA SYSTEM

Although the body's major internal energy centers are contained within the interacting network of the Three Dantians, the Three Dantians also connect to the body's major energy doors (located on the Governing and Conception Vessels), known as the Twelve Gates of the Chakras. These Chakra gates are superficial, on the surface of the skin, and are not as deep as the Chakra Cores and Three Dantians which are all connected to the Taiji Pole (Figure 5.9).

The word Chakra is Sanskrit for "wheel." The Chakras look like small, colored disks, usually about the size of a silver dollar. Energetically the Chakra gates look like funnels or vortices of Qi. Qigong doctors and sages have for centuries described their perceptions of these gates as resembling energetic wheels or vortexes within the subtle energy body. Each funnel extends and expands its energetic vortex out into the body's Wei Qi field. As energy travels up and down the Taiji Pole it creates an energetic pulse. This energetic pulse resonates out from the body through the Twelve Chakra Gates, and can be felt several feet from the body. The bottom Chakra is located at the perineum and only has one gate, as does the top Chakra located at the top of the head; the second, third, forth, fifth, and sixth Chakras have two gates each, one on the front of the body and the other on the back of the body. These Chakra Gates serve the following different energetic functions:

1. The top or Crown Chakra gate is responsible for absorbing Heavenly Qi and light into the body, as well as energizing the center core's Taiji Pole.
2. The front Chakra gates are responsible for the patient's feelings and emotional activity. They are related to the patient's Shen and the Heart Fire energy. They are connected to the Conception Vessel.
3. The back Chakra gates are responsible for the patient's willpower and determination. They are related to the patient's Zhi and the Kidney Water energy. They are connected to the Governing Vessel.

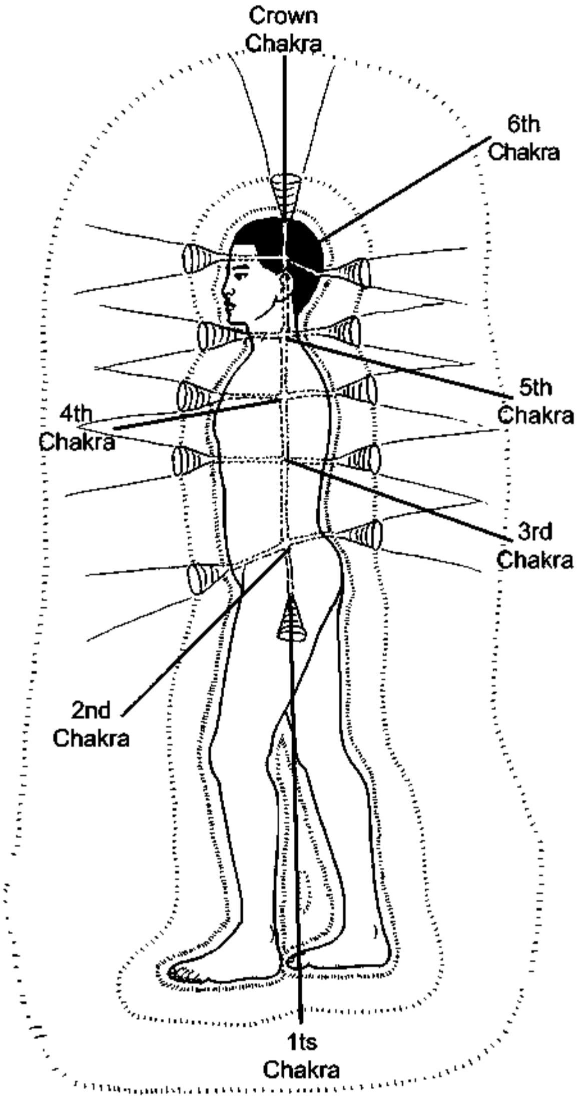

Figure 5.9. The body's Twelve Chakras Gates extend from the Taiji Pole and expand outward into the three external Wei Qi fields.

4. The center of the Chakras are actually deep inside the body's center core, located within the middle of the Taiji Pole.
5. The bottom or Base Chakra gate is responsible for absorbing Earthly Qi and heat into the body, as well as energizing the Taiji Pole. Generally, when treating a patient, the Qigong

doctor extends energy into the patient's back Chakra gates to stabilize the emotions and to access control over the patient's will and determination. The doctor can, for example, connect with the Heart Chakra back gate, located on the patient's Shendao, GV-11 point, to access the emotions stored within the tissues of the Heart. The front Chakra gates are used to remove Excess emotions from the patient's tissues. Although the treatment approach differs, the Three Dantians system and Twelve Chakra Gate system, mutually support each other, both in the location of the energy centers, and in their physical manifestations of energetic function.

The Yin and Yang polarity of the Chakra gates are opposite in men and women (Figure 5.10). The Twelve Chakra Gates serve as subtle energy distributors, that help absorb and distribute environmental Qi to the patient's organs, tissues, and major nerve plexus areas closest to each gate. Each Chakra gate is connected to its own potential of psychic perception, interfaces with the body's nervous system, and is associated with a different endocrine gland. Any energy extended from the Qigong doctor towards a patient's Chakra gate will affect the patient's physical body. As the patient begins to energetically open the Chakra gates, the external energetic Chakra wheels begin to spin in half-circle rotations, opening and closing with the core's energetic pulse. The energetic action is similar to the centripetal (closing) and centrifugal (opening) action of the body's energetic channels.

Each opening of a Chakra gate may result in a spontaneous emotional release. Some of these releases can involve painful memories that have been dislodged from the energetic filter (or veil) that envelops the Chakra gates. This energetic filter prevents external emotional traumas from entering into the body's center core.

A Chakra gate can become stuck open, closed, or may also tilt out of alignment with the Taiji Pole, causing Qi Deviations that may result in a distorted or obstructed flow of energy. This obstruction may cause physical, as well as psychological, stress or trauma. The release of these emotional

Polarity		
Chakra	**Men**	**Women**
7th	+ Yang	– Yin
6th	– Yin	+ Yang
5th	+ Yang	– Yin
4th	– Yin	+ Yang
3rd	+ Yang	– Yin
2nd	– Yin	+ Yang
1st	+ Yang	– Yin

Figure 5.10. The energetic Yin and Yang polarity of the Chakra Gates are opposite in men and women. The Chakra Gates of a man start from a Yang (+) energetic field (at the Lower Chakra Gate) and change to Yin (-) at the second Chakra Gate, alternating from Yang to Yin up the body, ending at Yang in the Upper Chakra Gate. Women display the opposite energetic polarity in each Chakra Gate.

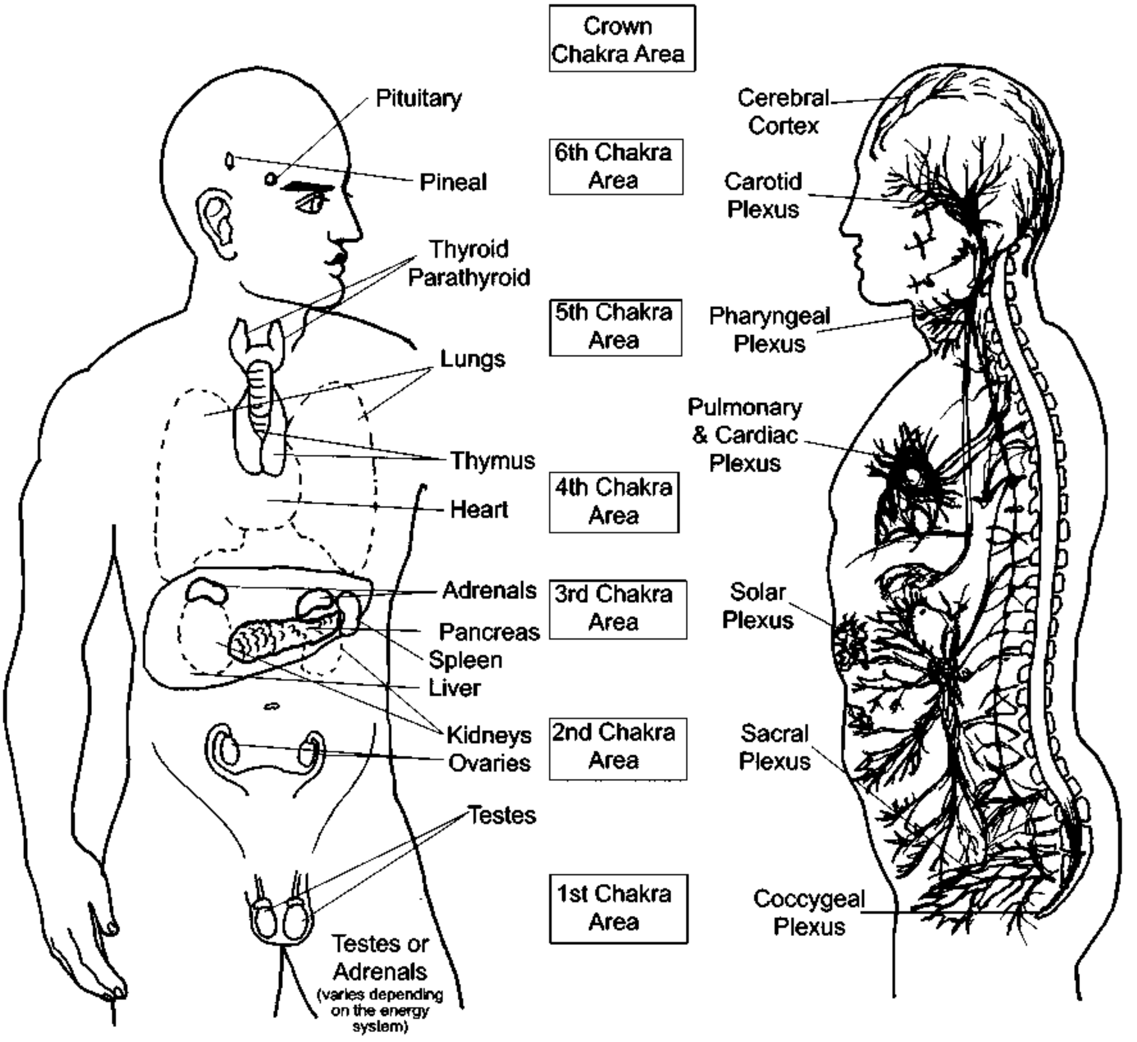
Crown
Chakra Area
Pituitary
Pineal
6th Chakra
Area
Cerebral
Cortex
Carotid
Plexus
Thyroid
Parathyroid
5th Chakra
Area
Pharyngeal
Plexus
Lungs
Pulmonary
& Cardiac
Plexus
Thymus
Heart
4th Chakra
Area
Adrenals
3rd Chakra
Area
Solar
Plexus
Pancreas
Spleen
Liver
Kidneys
2nd Chakra
Area
Ovaries
Sacral
Plexus
Testes
1st Chakra
Area
Coccygeal
Plexus
Testes or
Adrenals
(varies depending
on the energy
system)

Figure 5.11. The body's Chakra system is connected by the Taiji Pole and extends outward into the body's nerve plexus, affecting the glandular system and both anterior and posterior external energetic fields.

traumas can be overwhelming and painful if the patient has not prepared for, or does not understand the nature of, these transitional energetic malfunctions. These manifestations usually occur when the patient's system has been energized and the body is actively seeking to release and get rid of toxic emotions from certain areas. Qigong doctors should be aware of these strong emotional reactions to assist their patients in facing and working through their fears and pain as the Qi Deviations along the Taiji Pole are corrected. Sometimes these experiences unwind at a rapid rate, stimulating the patient's central and anterior nervous system; this can release a flood of mental and emotional images and sensations, that are sometimes accompanied by shaking, thrashing, and other unusual movements of the body (Figure 5.11).

THE BOTTOM CHAKRA GATE

The bottom or first Chakra gate is sometimes called the Meeting of Yin (referring to the Earth energy), and is located near the perineum, in front of the anal sphincter on the Huiyin CV-1 point. This energy center controls the reproductive system and the urogenital organs and intersects with the Bladder and Kidney Channels. It is associated with survival instincts (the "fight or flight" response), security, tribal power (group acceptance), and the "grounding" or "rooting" of the body.

The first Chakra is associated with the Lower Dantian. Its energetic elemental nature relates to Earth, and its color is red. It connects to the Taiji Pole at the base of the perineum and exits at the top, or Crown Chakra.

The bottom Chakra is associated with the kinesthetic, tactile, and proprioceptive senses. It supplies energy to the spinal column, adrenal glands (in some systems), testicles/ovaries, and Kidneys. This Chakra gate is also connected to both the Conception and Governing Vessels.

The bottom Chakra gate will immediately open in any emergency situation, or when there is danger, to release stored information to ensure survival. Energetic malfunctions caused by stored emotional trauma in this Chakra can cause a massive amount of fear for one's life to be released. The retention of massive amounts of fear, or the sudden energetic release of panic, can cause the following symptoms: lower back pain, sciatica, constipation, hemorrhoids, rectal tumors and cancer, knee problems, and varicose veins.

THE SECOND CHAKRA GATES

The second Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Spirit's Palace Gate and is located at the navel around the Shenque CV-8 point. The back gate is sometimes called the Gate of Life and is located on the lower back at the Mingmen GV-4 point.

This second energetic area is considered the "gateway of the Lower Dantian." Its energetic elemental nature relates to Water, and its color is orange. It is the energy center through which other people's emotions are perceived (kinesthetic perception), allowing one to be sensitive to other people and to sense dangerous situations.

The second Chakra is the energy center for sexuality, sensuality, the hunger for power, and financial issues. It supplies the sexual organs (ovaries and testicles) and the immune system with energy.

When the second Chakra gate opens the patient feels the release of sexual desires, waves of orgasm, and sexual fantasy. Because this area is also associated with the desire for power, sometimes the desire to destroy is released. Energetic malfunctions caused by stored emotional trauma in this Chakra can cause lower back problems, frigidity or impotence. Second Chakra dysfunction can also cause Kidney, uterine, Bladder, and urinary problems.

THE THIRD CHAKRA GATES

The third Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Solar Plexus Center, in this system of Chakra diagnosis it is located between the navel and the xiphoid process (at the level of the pancreas), at the Zhongwan CV-12 (Middle Stomach Cavity) point. The back gate is called the Middle of the Spine and is located on the middle of the back at the Jizhong GV-6 point.

The third Chakra area is associated with the Middle Dantian and is known as the Yellow Court in Medical Qigong therapy. Its energetic elemental nature relates to Fire, and its color is yellow. It is the body's distribution point for psychic energies (gut instinct and intuition), personal power, and self-image. It is the area where the body's Qi transforms into Shen. It is also a storage chamber for severe emotional pain and anger.

This area supplies energy to the Liver, Gall Bladder, Stomach, Spleen, pancreas, adrenal glands, and nervous system. The third Chakra is the center for personal power-storing issues of responsibility, self-esteem, personal honor, and fear of rejection.

When the third charka's energy gates open, feelings of power and waves of anger, rage, fear, greed, jealousy, judgment, and criticism can be released. Energetic malfunctions caused by stored emotional trauma in this Chakra area can cause ulcers, hypoglycemia and diabetes, as well as Liver and / or adrenal problems.

THE FOURTH CHAKRA GATES

The fourth Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Heart Center and is located at the center of the breastbone, at the Shangzhong CV-17 point. The back gate is sometimes called the Spirit Path and is located on the Shendao GV-11 point on the back, between the scapula.

The fourth Chakra area is associated with the Middle Dantian. Its energetic elemental nature relates to Wind, and its color is green. This energy center is associated with love, compassion, empathy, clairsentience, and intuition. This area supplies energy to the Heart, Lungs, circulatory system, thymus gland, vagus nerve, diaphragm, and upper back.

When this energy center opens, it can be physically painful (the patient may experience feelings of a weight on the chest, heart attack sensations, or the release of tears) as the patient becomes aware of being attached to the personal ego. The feelings associated with the fourth Chakra have to do with our internal world, while feelings associated with the third Chakra have to do with the external world. Energetic malfunctions can cause Heart and Lung diseases.

THE FIFTH CHAKRA GATES

The fifth Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Throat Center and is located just above the hollow of the throat, at the Tiantu CV-22 point. The back gate is sometimes called the Big Vertebra and is located on the back, at the base of the neck on the Dazhui GV-14 point.

This energy center is associated with sound vibration, personal and divine will, and communication. The fifth Chakra area is associated with the Middle Dantian. Its energetic elemental nature relates to sound, and its color is bright blue. It is the center for psychic clairaudience, and the power source for the Qigong doctor's "inner voice" guidance. This area supplies energy to the throat, thyroid and parathyroid glands, neck vertebrae, mouth, Lungs, and alimentary canal (the digestive tube from the mouth to the anus).

When the fifth Chakra energetic center opens, a compulsive urge to sing and chant arises. This area is also where Heat from the Liver Fire sometimes becomes entangled with Phlegm (a viscid substance that is a product and a cause of disease) from the Lungs, forming what is known as "plum pit Qi" (an energetic knot within the throat). Energetic malfunctions caused by stored emotional trauma in this Chakra area can cause stiff neck, sore throat, throat "plum pit Qi" symptom, thyroid problems and swollen glands. The patient sometimes experiences a release of unexpressed grief and anger, resulting in the physical reactions of excessive swallowing, coughing, or laryngitis.

THE SIXTH CHAKRA GATES

The sixth Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Third-Eye Center and is located at the middle of the forehead at the Yintang or Hall of Impression point. The back gate is located on the back of the head, between GV-16 (Wind Palace) and GV-17 (Brain's Door), at the external occipital protuberance.

The sixth Chakra area is the center for psychic intuition, clairvoyant seeing (inner vision), and energy projection. The sixth Chakra area is associated with the Upper Dantian. Its energetic elemental nature relates to light, and its color is indigo (deep violet blue). This area of the body is responsible for enabling the Qigong doctor to see auras, Chakras, and other energetic images. This center is also responsible for mental telepathy, allowing the doctor to sometimes know the patient's thoughts and feelings. This area of the body supplies energy to the pituitary gland, lower brain, ears, nose, eyes, and nervous system.

When the sixth Chakra energy center opens, energetic malfunctions can result in headaches, eye strain, pain around the eyes, distorted or blurred vision, and sometimes blindness. Inaccurate interpretations of events, projections of personal fears and other emotions and motivations onto others may sometimes occur. Reoccurring nightmares and misguided fantasies may also be incurred.

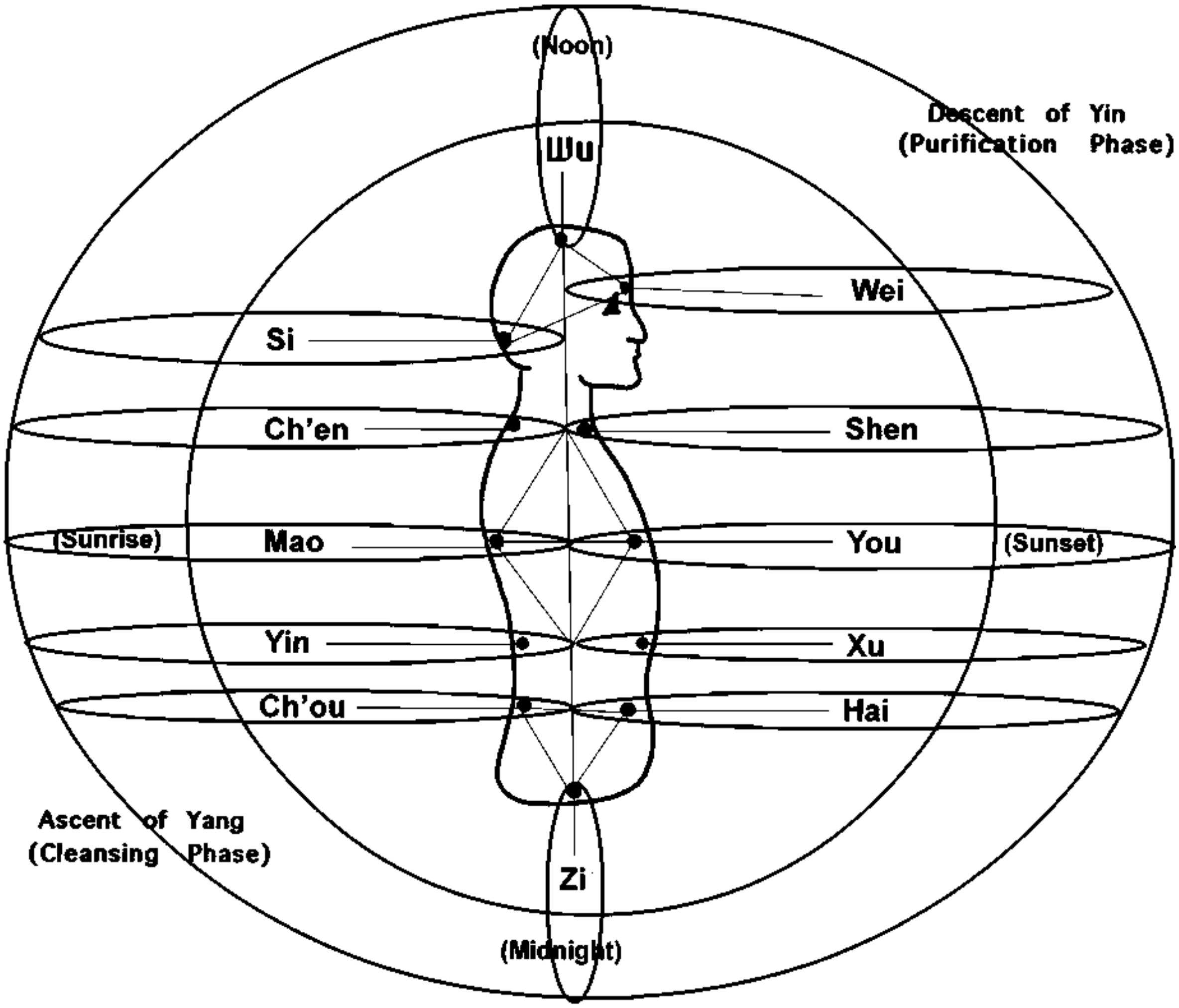
(Noon)
Wu
Descent of Yin
(Purification Phase)
Wei
Si
Ch'en
Shen
(Sunrise)
Mao
You
(Sunset)
Yin
Xu
Ch'ou
Hai
Ascent of Yang
(Cleansing Phase)
Zi
(Midnight)

Figure 5.12. This figure shows the Twelve Earthly Branch Relation to the Twelve Chakra Gates. The Governing and Conception Vessels flowing through the middle of the back and front of the body correspond to the ecliptic path of the sun along the Twelve Chakra Gates, and are related to the Twelve Earthly Branches.

THE UPPER CHAKRA GATE

The upper or top Chakra gate is sometimes called the Crown Center and is located on top of the head at the Baihui GV-20 point. This energy center is associated with higher knowledge, understanding, pure intuition, and ecstasy. The upper Chakra area is associated with the Upper Dantian. Its energetic elemental nature relates to infinite space (Wuji), and its color is violet. This area supplies energy to the pineal gland and upper brain.

When the upper Chakra energetic center opens, it is like a golden flower opening, accompanied by a dazzling bright light and a strong, powerful connection with the divine. Energetic malfunctions result in confusion, apathy, alienation, boredom, depression, and a state of incomprehension.

TAKING PRECAUTIONS

When any of the energetic reactions to opening the Chakra gates take place, it is important for the student or patient to realize that these are only transitional stages and are normal phenomena which occur as a sign of energetic and spiritual transformations. The student or patient should take certain precautions, however, like being under the care of a qualified Qigong doctor when beginning any type of energy training.

It is also important not to allow patients to experience these emotional transitions alone. Because these energy transitions, malfunctions, and deviations are new to the Western mode of thinking, they can easily be misdiagnosed by doctors unfamiliar with energetic medicine. Patients should be encouraged to find or establish a support group consisting of advanced practitioners of energetic meditation. Having a support group of experienced practitioners allows these emotional and spiritual transitions to occur in a safe environment, where other experienced practitioners can monitor the patient's feelings, if and when, unfamiliar emotions start to emerge.

THE TWELVE EARTHLY BRANCHES AND THE TWELVE CHAKRA GATES

The body is viewed as a small and complete universe unto itself. The internal organs are influenced by the celestial movements of the sun, moon, planets, and stars. The Governing and Conception Vessels, where the Twelve Chakra Gates are located, are also affected by the Heavenly cycles.

In ancient China, the day was divided into twelve separate time divisions. Each time division corresponds to two hours of the day and was named after one of the Twelve Earthly Branches. These twelve time divisions were further organized into months and seasons (see Chapter 11). The ancient Qigong doctors discovered that the body's Qi and Blood mirror the Earth's seasonal ebb and flow, rising and falling like the lunar tides.

Each of the time periods corresponding to the Twelve Earthly Branches is regarded as having a specific influence on each of the Twelve Gates of the body's Chakra system. The rhythmic variations of the waxing and waning of Qi and Blood is associated to the waxing and waning of Yin and Yang energy, as well as the circulation of Qi following the Microcosmic Orbit (Fire) cycle.

Each of the Twelve Chakra Gates relates to one of the Twelve Earthly Branches following the Microcosmic Orbit (Fire) cycle (see Chapter 41). These Twelve Chakras Gates extend their energy from the patient's Taiji Pole outward, along the center of the body, through the anterior and posterior fields of Qi. Beginning at the bottom Chakra gate (the "Zi" Branch, representing midnight), the energy follows the Fire Cycle of the Microcosmic Orbit, traveling up the Governing Vessel, following the "ascent of Yang." After the Yang Qi reaches its peak, the Yin begins to grow. Starting at the upper Chakra gate (the Wu Branch, representing noon time), the energy travels down the Conception Vessel, following the "descent of Yin" (Figure 5.12).

CHAPTER 6
THE TWELVE PRIMARY CHANNELS AND ORGANS

INTRODUCTION

According to Traditional Chinese Medicine, a system of channels exists that integrates all the body's separate parts and functions into a unified organism. The understanding of these channels and their energetic function and interactions provides the Qigong doctor with the basis for understanding the relationships among the various physiological, pathological, diagnostic, and therapeutic relationships in Traditional Chinese Medicine.

The study of Chinese medicine rests on, and is inspired by, the circulation of Qi, as many diseases follow a predictable course of development. The pathology in an internal organ will often manifest itself in certain external or systemic symptoms; therefore, understanding the circulation of Qi flow allows the doctor to control each organ's energetic function. By stimulating specific sites on the body's surface, diseases in both the superficial tissues and internal organs can be treated.

A doctor of Medical Qigong uses various techniques that facilitate the opening of the major energetic connectors of the body's internal organs. These connectors, called channels and vessels, are responsible for connecting the flow of Qi within the whole body to transfer Qi both internally and externally.

The channels are also known as meridians or Jing-luo. The Chinese character for this type of Jing means "to move through," while the translation of the character luo means "a net" (the body's inner fascia). Along these channels are major trunks and lesser branches that connect internally with the vital organs and externally with the major channels, limbs, sensory organs, and orifices.

The smaller branches are known as collaterals. The tiny areas along these channels and collaterals where the Qi pools are called points. These points are the spots where the patient's spirit and energy enter and leave the body. When stimulated, these channel points (sometimes also called "acupuncture points"), cause an energetic response from within the internal organs and channels, resulting in internal Qi flowing from organ to channel, from channel to channel, or from point to point along the same channel.

The channels unite the body as one unit. Channel theory is interrelated with organ theory. Traditionally, the internal organs have never been regarded as simply independent anatomical entities. Rather, Traditional Chinese Medicine focuses on the functional and pathological interrelationships between the channel network and the organs. The internal organs and their systems are sometimes known as "orbs" or "spheres of influence."

Qi travels throughout the physical body along the channels and collaterals much as water flows through rivers and streams. An ancient Chinese medical text explains this concept as follows: "Heaven is covered with the constellations, Earth with the waterways, and man with channels." Qigong doctors of ancient China paid much more attention to the Body Fluids and energies circulating through the body than to the physical anatomical structures. The ancient Chinese considered the Body's Fluids and energies much more fundamental.

The channels serve as the link between the energies, such as Jing, Qi, Shen, and the ingredients, such as Blood and Body Fluids, that feed and enliven the tissues. The human body's main energetic rivers are the Twelve Primary Channels and Eight Extraordinary Vessels (see Chapter 7).

CLASSIFICATION OF CHANNELS

The channels are classified into four main categories: Jing, Luo, Muscle Tendon, and Skin Zones.

1. The Jing Channels are the Primary Channels. These channels include the Eight Extraordinary Vessels, the Twelve Primary Channels, and the Twelve Divergent Channels.
2. The Luo Channels are the Collaterals. These collaterals include the Fifteen Major Collaterals, the Minute Collaterals, and the Superficial Collaterals.
3. The Muscle/Tendon Channels consist of twelve channels that serve as external connections to the major channels, flowing through the body's muscles, tendons, and ligaments.
4. The Cutaneous Regions are skin zone areas where the channels surface on the external tissues.

JING CHANNELS

The Jing Channels are the major trunks of the circulatory tree. They generally flow vertically through the body through relatively deep tissues. They also transverse the limbs peripherally and penetrate the body cavities to connect with the internal organs (Figure 6.1). They are the body's main interior and exterior rivers of Qi.

1. The Eight Extraordinary Vessels connect with, and regulate, the Qi and Blood of the Twelve Primary Channels by either absorbing any energetic "runoff" in times of channel Excess, or by replenishing energy in times of Deficiency.
2. The Twelve Primary Channels are bilateral and symmetrical. They can be identified in three specific ways: (1) according to the corresponding Yin or Yang organ to which they are connected, (2) by the arms or legs in which the channels originate or end, and (3) according to the six divisions of Yin or Yang Qi to which the channels relate (i.e., Tai Yang, Yang Ming, Shao Yang, Tai Yin, Shao Yin, and Jue Yin).

   The Yin channels run along the medial and anterior aspects of the body. They are associated with the solid Yin organs and connect with the hollow Yang organs. The Yin channels include the Kidneys, Spleen, and Liver Channels that flow from the feet to the torso, a well as the Heart, Pericardium, and Lung Channels that flow from the torso to the hands.

   The Yang channels run along the lateral aspect of the body. They belong to the hollow Yang organs and connect with the solid Yin organs. The Yang channels include the Small Intestine, Triple Burners, and Large Intestine Channels that flow from the hands to the head, as well as the Bladder Gall Bladder and Stomach Channels that flow from the head to the feet. The interaction between the internal/external, Yin/Yang relationship of the organs and channels results in an alchemical transference between energy and matter.
3. The Twelve Divergent Channels branch off the Twelve Primary Channels. They are mainly distributed on the chest, abdomen, and head. The Divergent channels have the energetic functions of connecting internally and externally related channels, strengthening the connection of the Twelve Primary Channels to their related organs, and serving as extensions of the Twelve Primary Channels.

LUO OR COLLATERALS

The Twelve Primary Channels are connected by collaterals, which are small, interlinking streams. These collaterals are the connecting branches of the energetic circulatory system. They generally flow superficially, in horizontal and vertical energetic directions. They are the body's secondary streams of Qi that form an intricate network that traverses the body's surface and interconnects the main rivers, connective tissues, and cutaneous regions (Skin Zones).

1. The Fifteen Major Collaterals (Luo) transfer Qi and Blood from the Twelve Primary Channels to all parts of the body and link the body's interior with its exterior, connecting the body's internal and superficial channels, as well as the Governing and Conception Vessels and the Great Luo of the Spleen (see Chapter 7).

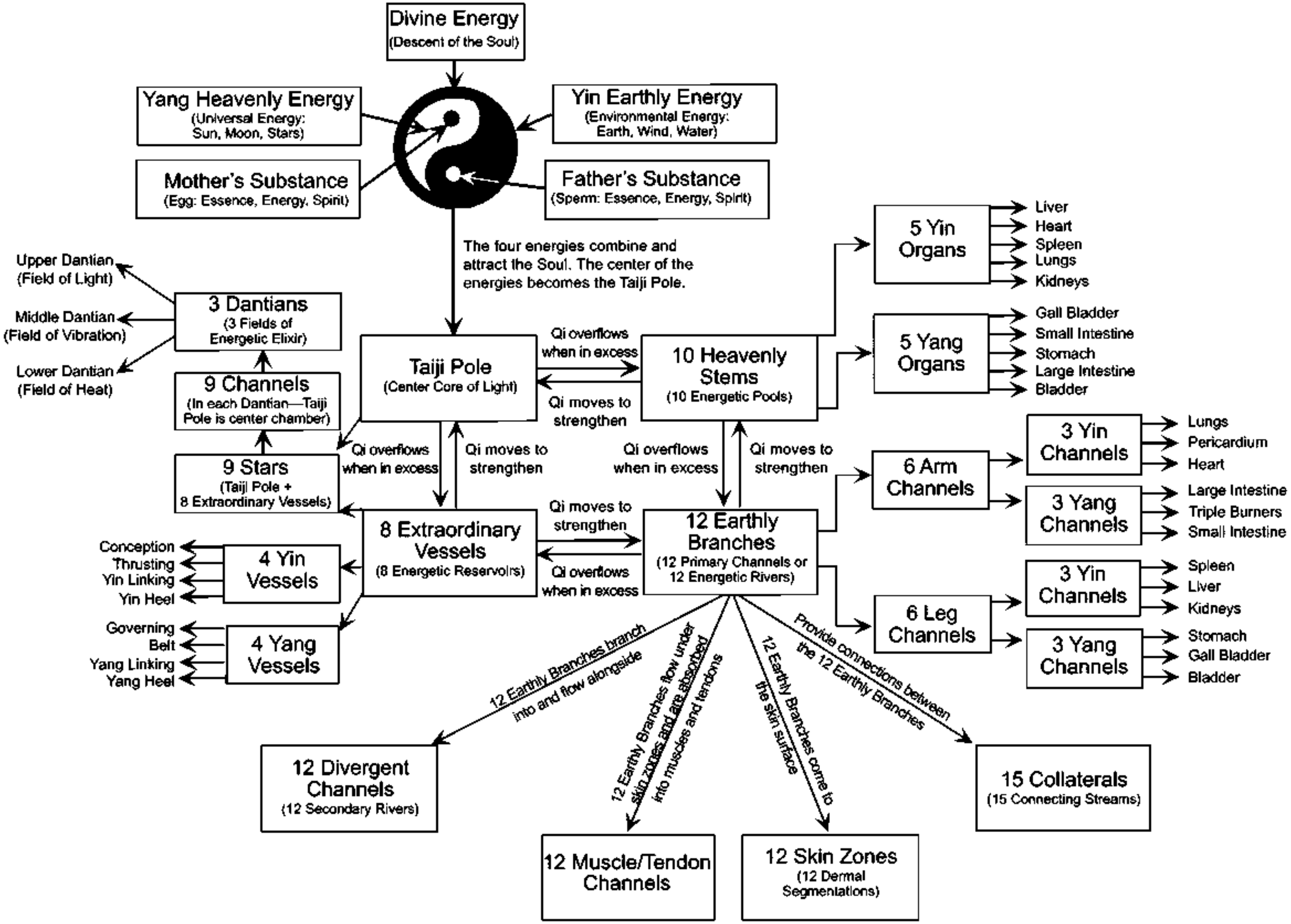
Divine Energy
(Descent of the Soul)
Yang Heavenly Energy
(Universal Energy: Sun, Moon, Stars)
Yin Earthly Energy
(Environmental Energy: Earth, Wind, Water)
Mother's Substance
(Egg: Essence, Energy, Spirit)
Father's Substance
(Sperm: Essence, Energy, Spirit)
The four energies combine and attract the Soul. The center of the energies becomes the Taiji Pole.
Upper Dantian
(Field of Light)
Middle Dantian
(Field of Vibration)
Lower Dantian
(Field of Heat)
3 Dantians
(3 Fields of Energetic Elixir)
9 Channels
(In each Dantian—Taiji Pole is center chamber)
9 Stars
(Taiji Pole + 8 Extraordinary Vessels)
Taiji Pole
(Center Core of Light)
Qi overflows when in excess
Qi moves to strengthen
10 Heavenly Stems
(10 Energetic Pools)
5 Yin Organs
Liver
Heart
Spleen
Lungs
Kidneys
5 Yang Organs
Gall Bladder
Small Intestine
Stomach
Large Intestine
Bladder
Qi overflows when in excess
Qi moves to strengthen
Qi overflows when in excess
Qi moves to strengthen
6 Arm Channels
3 Yin Channels
Lungs
Pericardium
Heart
3 Yang Channels
Large Intestine
Triple Burners
Small Intestine
8 Extraordinary Vessels
(8 Energetic Reservoirs)
Qi moves to strengthen
Qi overflows when in excess
12 Earthly Branches
(12 Primary Channels or 12 Energetic Rivers)
Conception
Thrusting
Yin Linking
Yin Heel
4 Yin Vessels
Governing
Belt
Yang Linking
Yang Heel
4 Yang Vessels
6 Leg Channels
3 Yin Channels
Spleen
Liver
Kidneys
3 Yang Channels
Stomach
Gall Bladder
Bladder
12 Earthly Branches branch into and flow alongside
12 Earthly Branches flow under skin zones and are absorbed into muscles and tendons
12 Earthly Branches come to the skin surface
Provide connections between the 12 Earthly Branches
12 Divergent Channels
(12 Secondary Rivers)
12 Muscle/Tendon Channels
12 Skin Zones
(12 Dermal Segmentations)
15 Collaterals
(15 Connecting Streams)

Figure 6.1. Channel and Organ Energy Flow

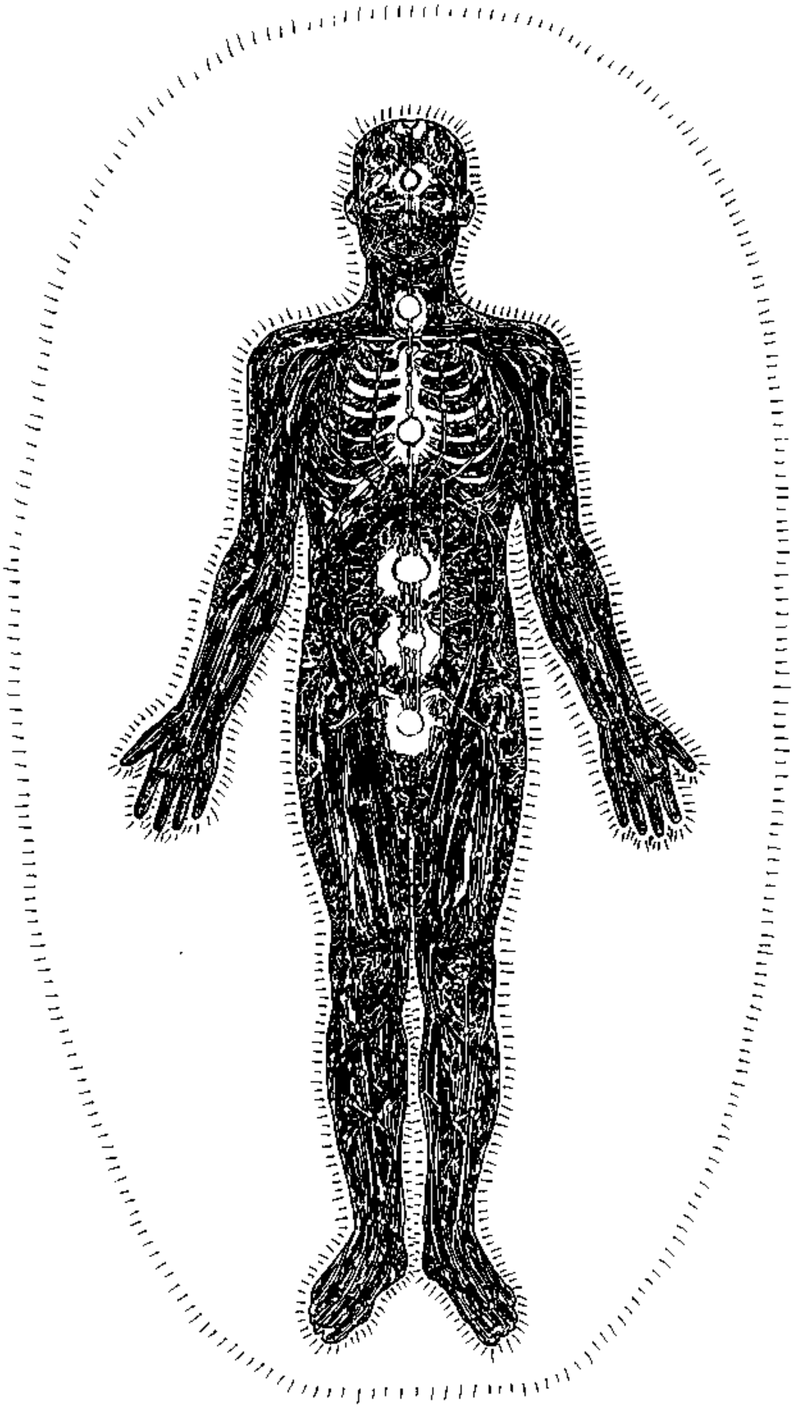

Figure 6.2. The Human Body's Energetic Matrix

2. The Minute Collaterals are smaller branches of the Fifteen Major Collaterals. They are countless in number.
3. The Superficial Collaterals are smaller branches of the Minute Collaterals. They are also countless in number.

MUSCLE TENDON CHANNELS

The Twelve Muscle/Tendon Channels are the external connections of the major channels. They generally flow superficially, traversing the body's surface and serve to join the main rivers, connective tissues, and Cutaneous Regions. The Twelve Muscle/Tendon Channels are also regions of the body where the Qi and Blood of the Twelve Primary Channels nourish the muscles, tendons, and ligaments.

CUTANEOUS REGIONS

The Twelve Cutaneous Regions (also called Twelve Skin Zones) are areas of the body where the Blood and Qi of the Twelve Primary Channels surface and connect to the body's skin tissues. These cutaneous regions have a continuous and direct contact with the external environment.

The body resonates in a continuous interplay of Yin and Yang harmony, balancing the channels' and organs' energy flows. The channels help create the body's external Yin and Yang polarities and create the energetic matrices around which the body's energetic fields can flow. Together these energetic matrices establish the foundation of the body's energetic tissues (Figure 6.2).

THE CHANNELS' RELATIONSHIP TO QI AND BLOOD

The channels transport Qi and Blood to nourish, moisten, and vitalize the whole body. Healthy bodily function depends on the balanced circulation of Qi and Blood. Qi is Yang and provides the energy or force necessary for the body's functional activity (i.e., movement and transportation, warming, containing, transforming, and defending). Blood is Yin and is the source of the body's moistening, nourishment, and lubrication.

The Blood, Qi, and heat circulate through the Twelve Primary Channels every two-hours, ebbing and peaking in energetic flow. The maximum peak of a channel's Qi, or "high-tide," occurs at the time of day when the Qi in that channel is at its fullest. The minimum period of a channel's Qi, or "low tide," occurs twelve hours from the peak time (Figure 6.3).

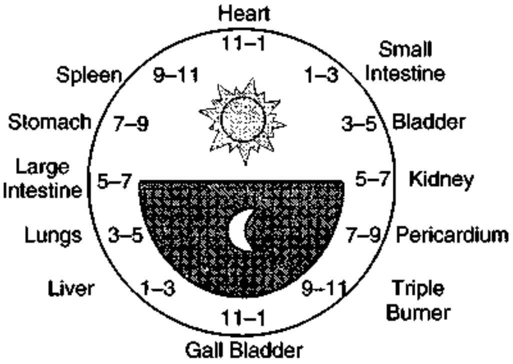

Figure 6.3. The Body's High and Low Energetic Tides

The connective tissues of the body's fasciae transmit a variety of biological energies. Think of the inner fascia as lubricated linings that cover the muscles and internal organs, allowing the tissues to slide and move easily. Within the lining of the superficial fasciae, the channels are webbed; this webbed network conducts energy in the form of electron, proton, and ion transmission, and is influenced by light and sound.

According to Traditional Chinese Medicine theory, each of the Twelve Primary Channels has its own individual resources of Qi and Blood; however, the energetic quality varies in accordance with the amount of Blood and Qi available in each organ and its particular channels. From a traditional Chinese medical perspective, the Spleen plays two important roles in regard to Qi and Blood. First, it converts food into Gu Qi which, when further refined by the Liver and Kidneys, becomes Ying (Nutritive) Qi; and second, it manages the Blood by keeping it in the channels. Ying Qi secretes Body Fluids that enter into the blood vessels and are transformed into Blood by the Heart. The Heart both creates and governs the Blood; whereas, the Liver stores it and spreads Qi throughout the body. Blood is considered to be a denser, more material form of Qi. Qi and Blood

Organs & Channels	Qi	Blood
Liver	less	more
Lungs	more	less
Large Intestine	balanced	balanced
Stomach	balanced	balanced
Spleen	more	less
Heart	more	less
Small Intestine	less	more
Bladder	less	more
Kidney	more	less
Pericardium	less	more
Triple Burners	more	less
Gall Bladder	more	less

Figure 6.4. The Body's Organ and Channel Relationship to the Proportions of Qi to Blood

are interdependent: Qi moves the Blood, but the Blood nourishes Qi; both flow together.

The chart in Figure 6.4 shows each of the Twelve Primary Channels and the proportion of Qi and Blood each channel carries.

FUNCTION OF THE CHANNELS

The channels are responsible for responding to any malfunction in the body. They can also, however, be disseminators of energetic dysfunction and disease. Disease passes in to the internal organs, or from one organ to another, via the channels. Exogenous diseases progress from the body's skin pores to the tiny collaterals, then to primary channels, and finally, to the internal organs. The channels are affected by disease in a predictable fashion.

1. Diseases and symptoms manifest along the channel pathway.
2. Points along the channels become tender, painful, tight, or flaccid.
3. When in a deficient state, the electrical resistance and heat tolerance diminish on the skin at the diseased point(s).
4. The body's sensory organs (sight, smell, taste, hearing and touch) are affected through channel relationship.

From a traditional Chinese medical perspective, the functions and flow of the channels and collaterals, as well as the internal organs' health can be affected by Medical Qigong, acupuncture, Chinese massage, and herbal medicine. Each system of Traditional Chinese Medicine disperses Excess, moves stagnation, stimulates and nourishing the Qi to tonify Deficiency, and stimulates the Wei Qi to fight External pathogens via the channels. In fact, each time the body's channels and collaterals are trained or regulated, they become thicker with Qi.

INTERNAL AND EXTERNAL CHANNEL FLOW

As the Qi of an internal channel flows into or out from its "primary organ," it generates a current that is similar to a river rushing back into the sea. The moment that energy flowing through the internal channel permeates the primary organ, an energetic connection and reaction is immediately created in the organ tissues. When the Qi of an internal channel flows into its "associated organ," however, the energetic reaction is quite different. Instead of immediately permeating the organ, the channel's energy "spirally wraps" the organ's tissues through the internal fascia that envelops the associated organ. The associated organ tissues then absorb the energy from the surrounding fascia. This difference in energetic penetration and absorption rate is an important factor when working with the body's internal energetic channels. When Qigong doctors extend their energy into a patient's channels, they must be aware that the absorption rate will be immediately felt in the channels pertaining organ, and will have a gradual effect on the organ's associated organ.

Channel Name	Yin or Yang	Associated Organ	Element	Blood Heat Cycle	Qi Blood Cycle	Abbreviation
Gall Bladder	Yang	Liver	Wood	11 p.m.– 1 a.m.	more Qi less Blood	G.B.
Liver	Yin	Gall Bladder	Wood	1 a.m.– 3 a.m.	less Qi more Blood	Liv.
Lungs	Yin	Large Intestine	Metal	3 a.m.– 5 a.m.	more Qi less Blood	Lu.
Large Intestine	Yang	Lung	Metal	5 a.m.– 7 a.m.	balanced Qi and Blood	L.I.
Stomach	Yang	Spleen	Earth	7 a.m.– 9 a.m.	balanced Qi and Blood	St.
Spleen	Yin	Stomach	Earth	9 a.m.– 11 a.m.	more Qi less Blood	Sp.
Heart	Yin	Small Intestine	Fire	11 a.m.– 1 pm.	more Qi less Blood	H.
Small Intestine	Yang	Heart	Fire	1 p.m.– 3 p.m.	less Qi more Blood	S.I.
Bladder	Yang	Kidney	Water	3 p.m.– 5 p.m.	less Qi more Blood	Bl.
Kidney	Yin	Bladder	Water	5 p.m.– 7 p.m.	more Qi less Blood	K. or Kid.
Pericardium	Yin	Triple Burner	Fire	7 p.m.– 9 p.m.	less Qi more Blood	P.
Triple Burner	Yang	Pericardium	Fire	9 p.m.– 11 p.m.	more Qi less Blood	T.B.

Figure 6.5. The Twelve Primary Channels are divided into Yin and Yang. The Yin channels are connected to the solid organs, and the Yang channels are connected to the hollow organs.

ORGAN AND CHANNEL FUNDAMENTALS

The purpose for energetic anatomy and physiology is to understand the fundamental actions of each organ and each channel's unique energetic characteristics. This understanding includes the learning of differential diagnosis to categorize the various symptoms into meaningful patterns of disharmony. This ability to categorize the various symptoms will enable the doctor to analyze and recognize specific pathologies indigenous to specific organs and channel energies (Figure 6.5).

In Chinese medical science, the study of energetic physiology is called Zhang Xiang Xue Shou, which literally translates as "the theory of the phenomena of internal organs." Here, the word "phenomena" means visible external manifestations. The organ's physical structure is of little consequence; whereas, its function in the body is of primary importance. While Western anatomy and physiology are primarily concerned with the physical body in its most concrete form (investigating the structures of the major body systems and organs), energetic anatomy and physiology focuses mainly on the underlying energy that enlivens and sustains the physical form.

In Chinese medical science, the body's internal organs are responsible for creating and distributing Qi, Blood and Body Fluids. Traditional Chinese Medicine divides Body Fluids into two distinct categories: Jin, which are the body's clear, light fluids, and Ye, which are the body's thick and heavy liquids.

1. The Jin Body Fluids are the Yang Fluids of the Upper Burner that move with Wei Qi. They are fast, light, and clear fluids, such as: sweat, tears, saliva, and mucus. The Lungs control the Jin.
2. The Ye Body Fluids are the Yin Liquids of the Middle and Lower Burners that move with Ying Qi. These are slow, heavy, and thick liquids. These liquids include those that moisten the joints, spine, brain, and Bone Marrow and lubricate the eyes, ears, nose, and mouth. The Spleen and Kidneys control the Ye.

This internal action of creating and distributing Qi, Blood and Body Fluids is manifested throughout the body's tissues through the Blood's Heat Cycle, which actively follows a peak (high-tide) and ebb (low-tide) energetic time period. Each organ and channel has either a Yin or Yang characteristic and is grouped and paired according to its Five Element nature.

This chapter focuses on the Twelve Primary Channels and their relationship to the organs. Most Traditional Chinese Medical texts begin the sequence of these channels with the Lung Channels. In this text, however, due to the energetic waxing and waning of the Yin and Yang Yao Cycles, and the Twelve Pi Hexagrams (see Chapter 11), I will start with the Gall Bladder Channels. The Twelve Primary Channels are described as follows: Gall Bladder, Liver, Lung, Large Intestine, Stomach, Spleen, Heart, Small Intestine, Bladder, Kidney, Pericardium, and Triple Burner.

THE GALL BLADDER CHANNELS

The Gall Bladder Channels are Yang channels. The external branches flow from the head to the feet on both sides of the body (Figure 6.6). The internal branches of the Gall Bladder Channels flow from the head to the Gall Bladder. The internal and external branches of these two rivers originate from the outer canthus of each eye. The external branches zigzag around the side of the head, flowing down the sides of the torso. The internal branches leave the outer canthus of the eye and descend the neck, entering the supraclavicular fossa, where they meet the main channels at the St-12 and Pc-1 points, and pass through the diaphragm, spirally wrapping the Liver, and permeating the Gall Bladder. From there they encircle the genitals and enter deeply into the body to emerge at the sacrum.

The external rivers continue to descend the lateral aspects of the torso and legs, ending at the lateral sides of the tips of the fourth toes.

THE GALL BLADDER ORGAN

The Gall Bladder's associated organ is the Liver; its element is Wood. The Gall Bladder's function of storing and excreting bile depends on the Liver's function of maintaining a smooth flow of Qi throughout the body.

The Gall Bladder also helps the Liver control the sinews, tendons, and ligaments. The Liver provides Blood to the sinews, while the Gall Bladder provides Qi to the sinews in order to ensure proper movement and agility.

PSYCHO-EMOTIONAL ASPECTS

The Gall Bladder is responsible for making decisions and judgments, as well as providing courage and initiative. This organ is sometimes called The Court of Justice or The General's Advisor. Although the Kidneys control drive and vitality, the Gall Bladder provides the capacity to turn this drive and vitality into decisive action.

The Gall Bladder has an influence on the quality and length of sleep. If the Gall Bladder is Deficient, the patient will often wake up suddenly, very early in the morning, and be unable to fall asleep again.

Patient's who are timid, indecisive, and easily discouraged by slight adversity, are said to have a weak Gall Bladder; conversely, decisive and determined patients are said to have a strong Gall Bladder.

GALL BLADDER'S ENERGY FLOW

The Gall Bladder Channels are often used to drain off the energy from an overheated or Excess Liver condition. Energetically, the Gall Bladder Channels store more Qi than Blood, acting more on energetic and nervous functions than on physical substances and Blood functions. At the

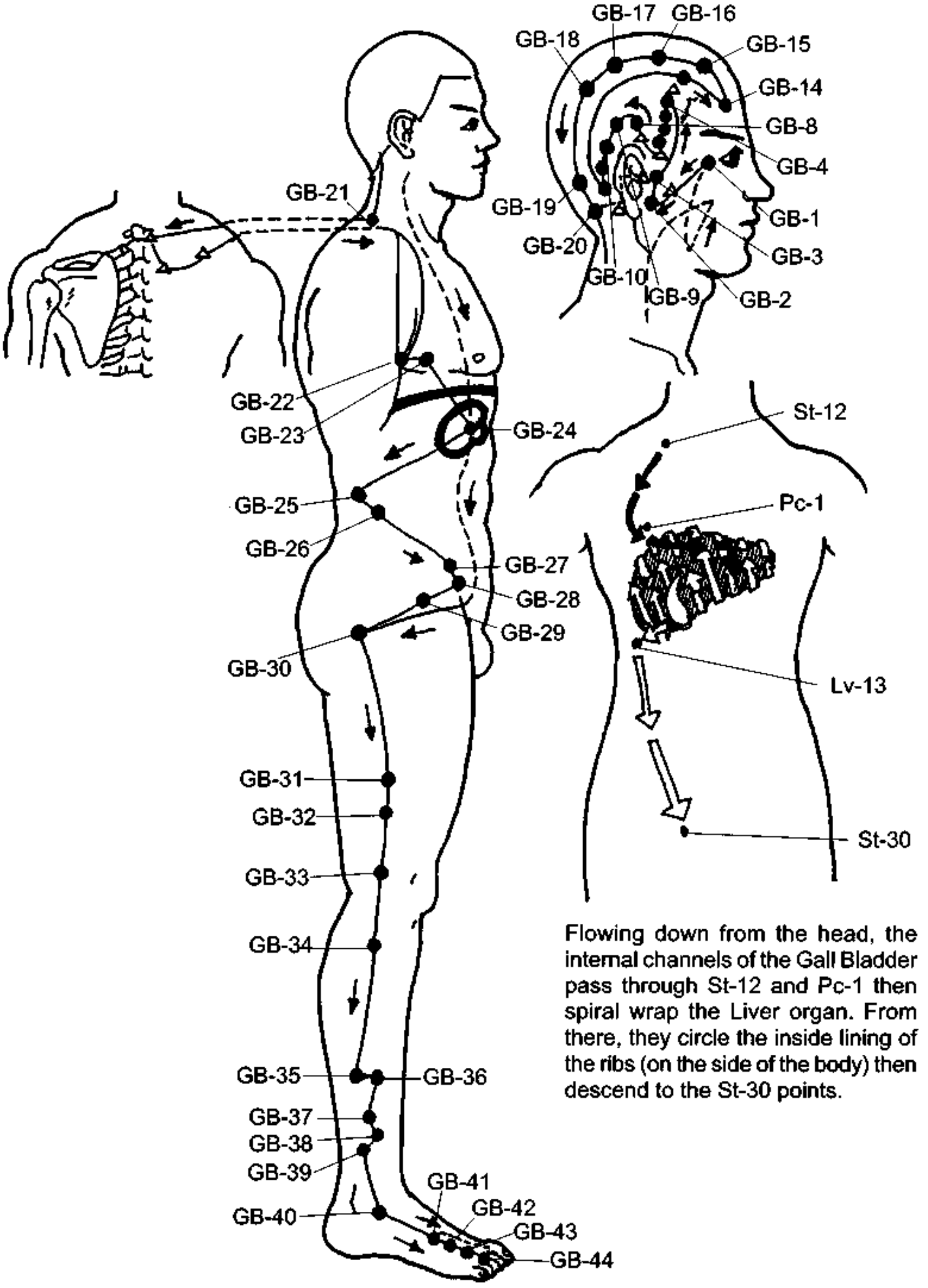
GB-17
GB-16
GB-18
GB-15
GB-14
GB-8
GB-4
GB-21
GB-19
GB-1
GB-20
GB-3
GB-10
GB-9
GB-2
St-12
GB-22
GB-23
GB-24
Pc-1
GB-25
GB-26
GB-27
GB-28
GB-29
GB-30
Lv-13
GB-31
GB-32
St-30
GB-33
GB-34
GB-35
GB-36
GB-37
GB-38
GB-39
GB-41
GB-40
GB-42
GB-43
GB-44

Flowing down from the head, the internal channels of the Gall Bladder pass through St-12 and Pc-1 then spiral wrap the Liver organ. From there, they circle the inside lining of the ribs (on the side of the body) then descend to the St-30 points.

Figure 6.6. The Internal and External Qi Flow of the Gall Bladder (GB) Channels

high-tide time period (11 p.m. to 1 a.m.) Qi and Blood abound in the Gall Bladder, and so the Gall Bladder organ and channels can more easily be dispersed and purged. During low-tide (11 a.m. to 1 p.m.) they can be easily tonified. The Gall Bladder Channels' energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The gall bladder stores then releases bile through the cystic duct into the common bile duct that flow into the small intestine to aid in the digestion of fats.

PATHOLOGICAL MANIFESTATIONS

The main symptoms of the Gall Bladder Channel diseases are:

- pain in the right-and left-upper quadrants of the abdomen,
- disorders of the head, which include the eyes, ears, and face, and
- disorders of the external sides of the legs along the channels' pathways.

THE LIVER CHANNELS

The Liver Channels are Yin channels that flow externally from the feet to the torso (Figure 6.7). These two rivers originate externally from the inside of the big toes and flow upward on the inside the legs, to circle the groin. From there they continue to ascend externally to the lateral aspects of the thorax, where they enter internally and permeate the Liver. They then connect to and spirally wrap the Gall Bladder. From there, they flow to the Lungs, ascend internally into the thorax, and connect with the eyes, cheeks, and inner surface of the lips. They emerge from the forehead and connect with the Governing Vessel at the Baihui GV-20 point (Figure 6.8).

THE LIVER ORGAN

The Liver is responsible for the "free and easy wandering" of Qi in the body, i.e., the Liver makes the Qi flow smoothly in and around the body. The Liver governs the Belt and Thrusting Vessels, as well as the body's Qi flow and circulation. The smooth flow of Qi ensures normal mental and emotional activity, as well as normal secretion of bile. Impairment of this Liver function leads to a "binding depression of Liver Qi," associated with impatience, hasty decisions impulsive actions, and a hot temper.

The Liver's associated organ is the Gall Bladder; its element is Wood. The Liver energy opens externally at the eyes and manifests externally through the fingernails and toenails. The Liver stores the Blood, and serves as a reservoir of Blood to regulate the circulation of Blood volume. In women it assists the uterus and regulates the menses. The Liver regulates the function and control of the muscles, tendons, and ligaments and is the source of the body's physical strength. It also stores the emotion of anger.

PSYCHO-EMOTIONAL ASPECTS

The Liver is responsible for planning and creativity, as well as instantaneous solutions or sudden insights; it is therefore considered The General in Charge of Strategy. The Liver houses the body's Hun and governs fright.

Its positive psycho-emotional attributes are kindness, benevolence, compassion, and generosity; its negative attributes are anger, irritability, frustration, resentment, jealousy, rage, and depression.

The Liver is also called the "root of resistance to fatigue." Whenever the Liver is not functioning properly (stagnate or excessively Hot due to suppressed emotions) the patient can experience fatigue as well as physical weakness.

THE LIVER'S ENERGY FLOW

The Liver Channels move the Qi to the whole body, storing and distributing the Blood. The Liver Channels contain more Blood than Qi, thus they affect physical substances more than they affect energetic functions. At the high-tide time period (1 a.m. to 3 a.m.), Qi and Blood abound in the Liver, and so the Liver organ and channels can more easily be dispersed and purged.; whereas, during low tide (1 p.m. to 3 p.m.), they can be easily tonified. The Liver Channel's energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The liver stores vitamins A, B-12, D, E, and K, and regulates blood volume. It is one of the main

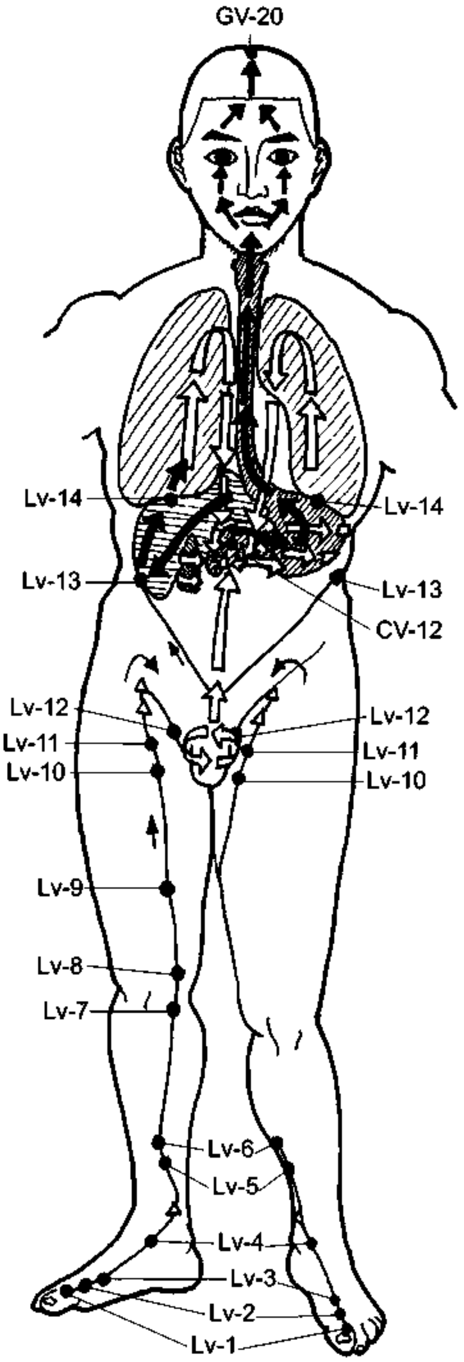
GV-20
Lv-14
Lv-14
Lv-13
Lv-13
CV-12
Lv-12
Lv-12
Lv-11
Lv-11
Lv-10
Lv-10
Lv-9
Lv-8
Lv-7
Lv-6
Lv-5
Lv-4
Lv-3
Lv-2
Lv-1

Figure 6.7. The Internal and External Qi Flow of the Liver (Lv) Channels

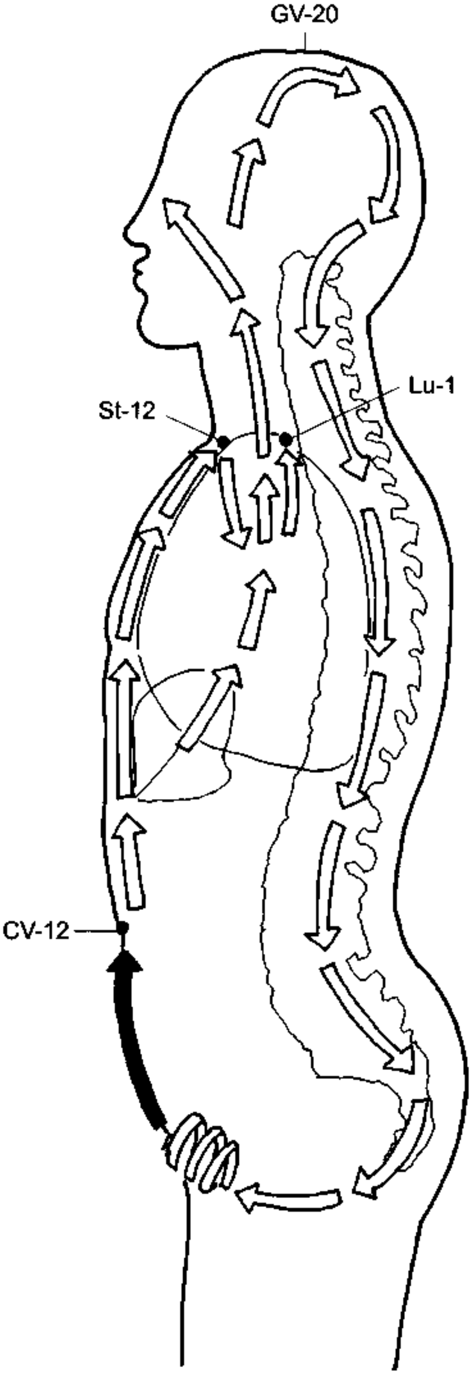
GV-20
Lu-1
St-12
CV-12

Figure 6.8. Side View of the Liver Channel's Internal Flow of Energy

sources of the body's heat. The liver manufactures cholesterol and functions in lipid metabolism and controls bile secretion.

PATHOLOGICAL MANIFESTATIONS

Diseases of the Liver Channels cause swelling and a distended sensation of the hypochondrium. The Liver Channels also correspond to diseases of the lower abdomen and genital organs.

Since the Liver supplies the tendons with nutrients to develop physical strength, Liver impairment causes numbness, tremors, or spasms of the muscles, and sluggishness in joint movements.

The Liver and Kidneys are mutually dependant upon each other. The Liver stores the Blood that nourishes the Kidney Jing; whereas the Kidneys store Jing that helps produce the Blood. Deficient Kidneys may lead to Blood Deficiency, and Deficient Liver Blood may cause weakness of the Kidney Jing due to lack of nourishment from the Blood. The hair on top of the head is also nourished by the Blood. When the hair turns grey, it is often alleged to be caused by insufficient Blood stored in the Liver, as well as a Kidney Jing weakness.

THE LUNG CHANNELS

The Lung Channels are Yin channels and flow externally from the torso to the hands. The main river originates internally from the Middle Burner, in the middle of the chest, and descends downward, connecting with and spirally wrapping the Large Intestine. From there, it ascends along the upper surface of the Stomach passage and through the diaphragm, where it branches and permeates both Lungs (Figure 6.9).

From the Lungs, the rivers of Qi merge and ascend into the pit of the throat, where they separate again into two channels that transverse below the clavicle. These two main channels then surface externally to descend down the arms, and end on the outer side of each thumb. A small stream of energy branches off each wrist at the Lu-7 point and runs directly to the radial side of the tip of the index finger, where it connects with a branch of the Large Intestine Channels.

THE LUNG ORGAN

Since the Lungs are the uppermost organs, they are compared to canopies that shelter and protect all other internal organs. The Lungs' associated organ is the Large Intestine; its element is Metal. The Lung Qi opens externally at the nose and manifests itself through the voice (i.e., when the Lung Qi is in Excess the voice is too loud; when Deficient the voice is too soft). The Lungs govern the body's Qi and respiration. They regulate the metabolism of Water passages (sweat and Body Fluids), the opening and closing of the pores, the skin and the hair's texture.

The Lungs also control the circulation of Qi in both vessels and channels, as well as the dispersing or spreading of the body's Wei Qi (protective energy). They control the descending and holding functions of the body's Qi and Body Fluids.

All blood vessels lead to the Lungs, i.e., all Blood within the body must pass through the Lungs.

PSYCHO-EMOTIONAL ASPECTS

It is said that the Lungs are "the priest" or The Minister of Heaven and are responsible for establishing the foundation of Qi for the entire body (see Chapter 12). The Lungs house the body's Seven Corporeal Souls (Po) and are responsible for self-protection and self-preservation.

The Lungs positive psycho-emotional attributes are righteousness, dignity, integrity, and high self-esteem; their negative attributes are disappointment, sadness, grief, despair, anxiety, shame, and sorrow.

THE LUNGS' ENERGY FLOW

Functionally, the Lung Channels' energy acts on the Lungs, bronchi, throat, and larynx. If the Lung Qi is combined with Liver Qi and becomes stagnate in the throat area, a condition known as a "plum pit," or knot in the throat, develops. The Lung Channels store more Qi than Blood, thus they have a greater affect on energetic and nervous functions than on physical substances and Blood functions. At the high-tide time period (3 a.m. to 5 a.m.), Qi and Blood abound in the Lung organs and channels. They can therefore be more

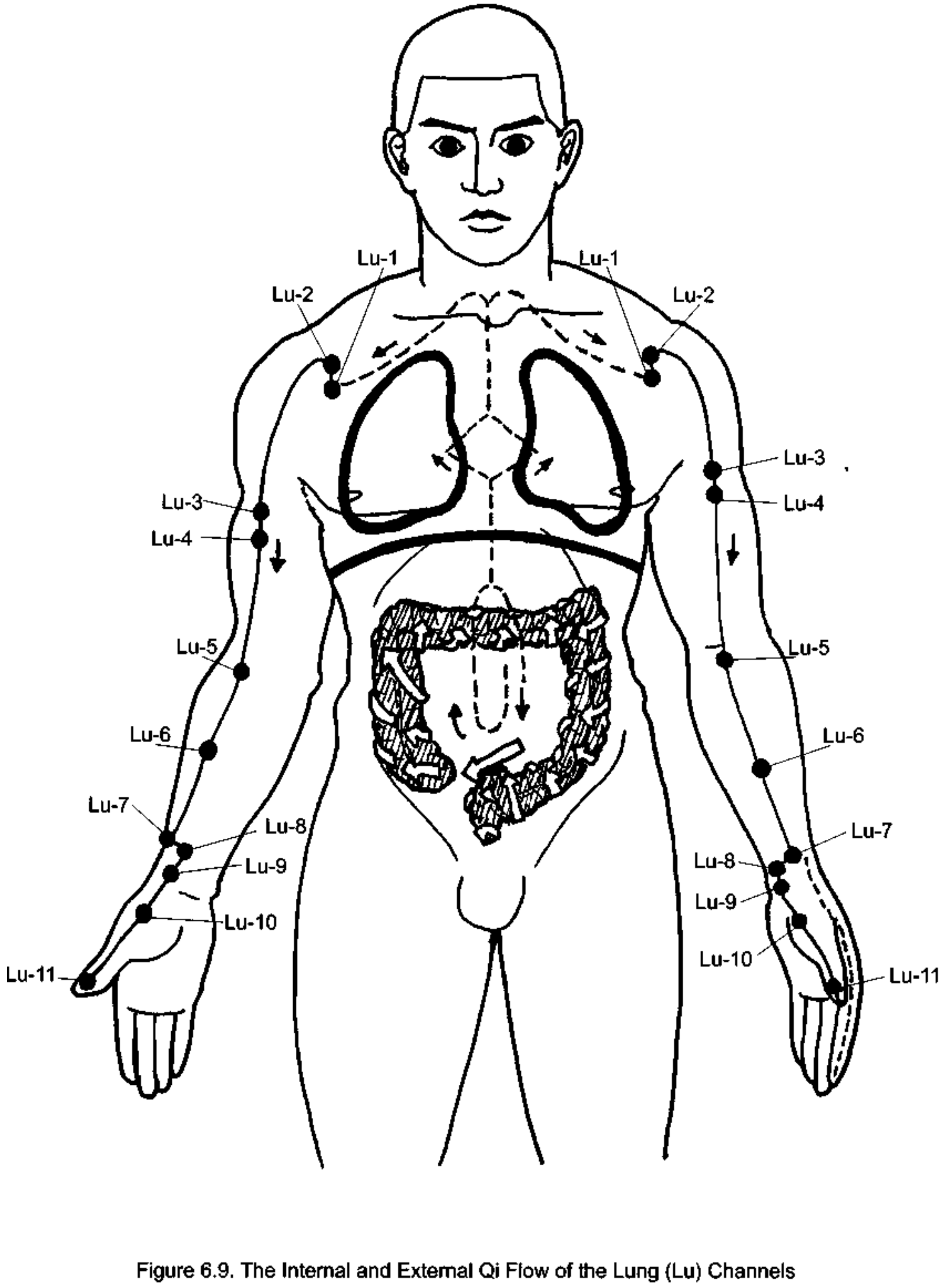
Lu-1
Lu-1
Lu-2
Lu-2
Lu-3
Lu-3
Lu-4
Lu-4
Lu-5
Lu-5
Lu-6
Lu-6
Lu-7
Lu-8
Lu-9
Lu-7
Lu-8
Lu-9
Lu-10
Lu-10
Lu-11
Lu-11

Figure 6.9. The Internal and External Qi Flow of the Lung (Lu) Channels

easily be dispersed and purged at this time period; whereas during low-tide (3 p.m. to 5 p.m.), they can more readily be tonified. The Lung Channels' energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The action of inhaling air introduces oxygen into the lungs; exhaling expels carbon dioxide from the lungs. The heart pumps blood into the lungs. Blood passing through the lungs receives oxygen, and is then delivered back into the heart, where it is circulated throughout the body via the circulatory system.

PATHOLOGICAL MANIFESTATIONS

The Lung Channels' dysfunctions can result in chest and Lung diseases, as well as diseases on the radial side of the upper arm and palmar area of the hand.

The Lungs keep the pathway of air unobstructed and disseminate vital Qi throughout the body. If these functions are impeded, obstructions of the nose, coughing, dyspnea, and fullness of the chest may occur.

The Lungs cleanse the inhaled air and keep the Qi flowing downward. If these functions are impeded, coughing, asthma, oliguria (diminished amount of urine formation), and edema may occur.

Since the Lungs have their external orifice at the nose, stuffy nose, nasal discharge, and impairment of smell are common symptoms when the Lungs are being attacked by a pathogenic invasion of Wind and Cold.

THE LARGE INTESTINE CHANNELS

The Large Intestine Channels are Yang channels and flow externally from the hands to the head. These two rivers originate externally from the tips of the index fingers, ascend up the arms, and cross the shoulders, where they connect with the 7th cervical vertebra and then split into two branches. One set of branches (one branch from the left Lung and its corresponding branch from the right Lung) descends internally and spirally wraps the Lungs before moving through to permeate the Large Intestine. The other set continues ascending externally through the neck and cheek to the gums of the lower teeth. It curves around the upper lip and flows to the opposite side of the nose, where it connects with the Stomach Channels (Figure 6.10).

THE LARGE INTESTINE ORGAN

The Large Intestine's associated organ is the Lungs; its element is Metal. The main function of the Large Intestine is to receive food essence from the Small Intestine. The Large Intestine organ is said to govern humor (any fluid or semifluid substance in the body). It absorbs fluid that was separated by the Small Intestine and controls the transformation and transportation of solid waste in the body. The Large Intestine also controls the throat, teeth, and the drainage of the nose.

PSYCHO-EMOTIONAL ASPECTS

The Large Intestine relationship to the Lungs makes it equally affected by the emotions of sadness, grief, and worry. An energetic imbalance in the Large Intestine can result in physical weakness and provoke emotional introversion, accompanied by feelings of depression, irritability, discouragement, distress and apathy.

Strong emotions of fear or panic can produce an energetic-stool reflex reaction in the Large Intestine resulting in a spontaneous defecation.

THE LARGE INTESTINE'S ENERGY FLOW

The Large Intestine Channels have equally abundant amounts of Qi and Blood, acting equally on the body's energetic and nervous functions, as well as on physical substances and Blood functions. At the high-tide time period (5 a.m. to 7 a.m.), Qi and Blood abound in the Large Intestine organ and channels and can therefore be more easily dispersed or purged at this period; whereas during low-tide (5 p.m. to 7 p.m.), they can be more readily tonified. The Large Intestine Channels' energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The large intestine, sometimes called the "colon," is divided into four quadrants; the ascending, transverse, descending, and sigmoid or pelvic colon. The function of the large intestine is to reabsorb the fluid from the contents of indigest-

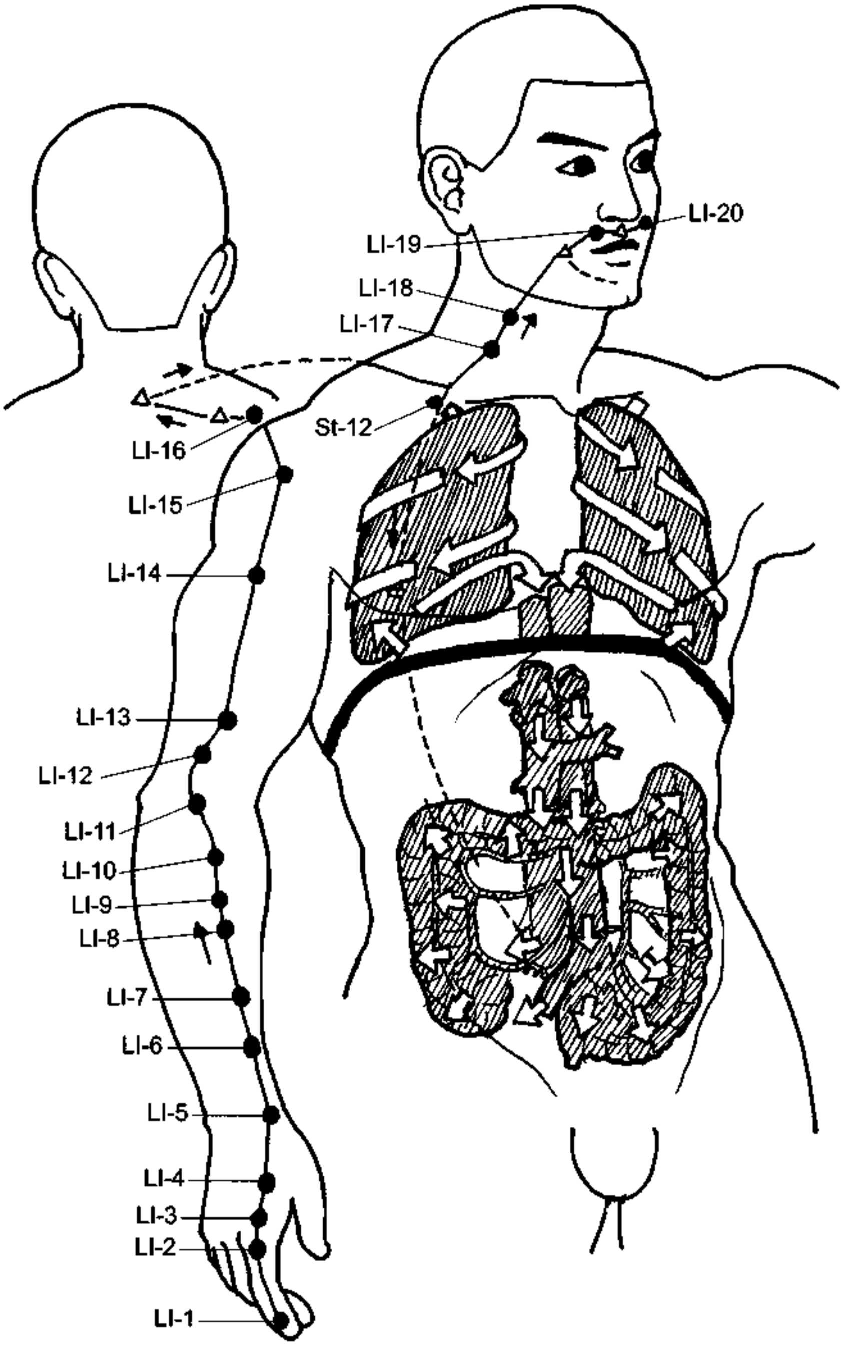
LI-20
LI-19
LI-18
LI-17
St-12
LI-16
LI-15
LI-14
LI-13
LI-12
LI-11
LI-10
LI-9
LI-8
LI-7
LI-6
LI-5
LI-4
LI-3
LI-2
LI-1

Figure 6.10. The Internal and External Qi Flow of the Large Intestine (LI) Channels

ible food residues moving through the intestines and to eliminate them from the body in the form of semisolid feces.

Whatever digestion takes place in the large intestine results from bacterial action. The fermentative bacteria (found in the middle part of the large intestine) and putrefying bacteria (found in the lower part of the large intestine) may produce toxic products, such as indole and skatol; however, if absorbed into the body, such bacterial decompositions undergo detoxification in the liver. In intestinal obstructions these bacterial decompositions are absorbed and eliminated in the urine in the form of indican.

PATHOLOGICAL MANIFESTATIONS

Dysfunctions in the Large Intestine Channels can result in diseases of the lower part of the face (including the nose, oral cavity, and teeth), throat, and front part of the neck, as well as disease of the back and radial sides of the upper extremities.

Tonification of the Large Intestine can be used to eradicate eye pain, toothache, earache, and to prevent hemorrhages, as well as to greatly reduce excessive menstrual bleeding.

Because of its relationship to the Lungs, the Large Intestine can be purged to treat coughing and asthma caused from Excessive Heat in the Lungs; whereas replenishing the vital Qi of the Lungs can cure constipation in debilitated patients.

The Large Intestines' energetic functional patterns generally relate to disturbances in bowel movements. Disturbances of the Large Intestine's energetic function usually gives rise to one of the two following complications:

1. Excess Conditions of the Large Intestines can result in symptoms of: Heat, Heat obstruction, Damp Heat, or Cold invading the Large Intestine.
2. Deficient Conditions of the Large Intestines can lead to an invasion of Cold or Dryness, and in severe cases of Deficiency, to the Collapse of the Large Intestine.

THE STOMACH CHANNELS

The Stomach Channels are Yang channels, and flow externally from the head to the feet (Figure 6.11). These two rivers originate externally from the lateral sides of the nose (LI-20), ascending upward to the base of the eye and bridge of the nose, where they meet the Bladder Channels. They then descend under the eyes, down the face, winding along the angle of the mandible, then ascend in front of the ears, following the anterior hairline till they reach the forehead. From the St-5 point, the primary branches on each side descend the neck and torso. Externally, these branches continue down the torso and legs and end at the lateral tips of the second toes. A second set of branches separate from the St-12 points and internally permeate the Stomach before spiral wrapping the Spleen and joining the primary channel branches at the St-30 points.

THE STOMACH ORGAN

The Stomach organ's associated organ is the Spleen; its element is Earth. The Stomach is considered the origin of the Body's Fluids. Its main function is to accept and decompose food. It separates the "clean" (pure) and "turbid" (impure) parts of ingested food and transfers the clean, energetically usable portion upwards to the Spleen, while sending the turbid unusable portion downward to the Small Intestine, to be further processed. Both the Stomach and the Spleen are considered the roots of Postnatal Qi and are often called The Ministers of Food Storage.

PSYCHO-EMOTIONAL ASPECTS

The Stomach influences the mental state; an Excess condition can agitate the mind and cause mental symptoms such as: mania or hypomania, confusion, severe anxiety, and hyperactivity.

THE STOMACH'S ENERGY FLOW

The Stomach Channels have equally abundant amounts of Qi and Blood, acting equally on the body's energetic and nervous functions, as well as on physical substances and Blood functions. At the high-tide time period (7 a.m. to 9 a.m.), Qi and Blood abound in the Stomach organ and channels, and can therefore be more easily dispersed or purged at this period; whereas during low-tide (7 p.m. to 9 p.m.), they can be more readily tonified. The Stomach Channels' energy acts on the skin, muscles, and nerves found along their pathways.

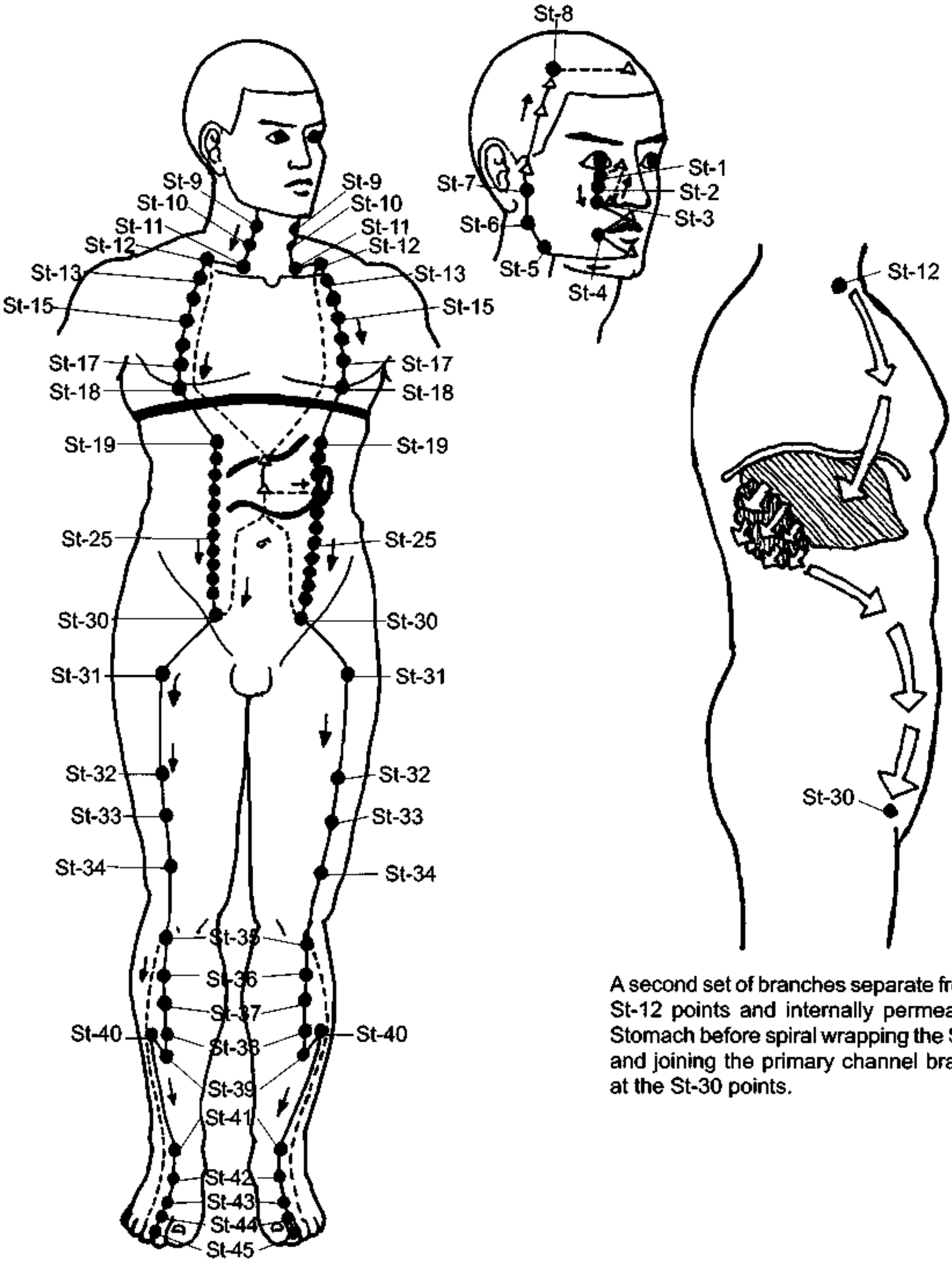
St-8
St-1
St-2
St-3
St-7
St-6
St-5
St-4
St-12
St-9
St-10
St-11
St-12
St-13
St-15
St-9
St-10
St-11
St-12
St-13
St-15
St-17
St-18
St-17
St-18
St-19
St-19
St-25
St-25
St-30
St-30
St-31
St-31
St-32
St-32
St-33
St-33
St-34
St-34
St-35
St-36
St-37
St-40
St-38
St-40
St-39
St-41
St-42
St-43
St-44
St-45
St-30

A second set of branches separate from the St-12 points and internally permeate the Stomach before spiral wrapping the Spleen and joining the primary channel branches at the St-30 points.

Figure 6.11. The Internal and External Qi Flow of the Stomach (St) Channels

WESTERN MEDICAL PERSPECTIVE

The stomach secretes gastric juice for food digestion. The proteins are converted into peptones; the stomach then regulates the passage of food to the other digestive organs.

PATHOLOGICAL MANIFESTATIONS

Diseases of the Stomach Channels include diseases of the face (nose, oral cavity, and teeth), throat, front of the neck, the abdomen, the frontal part of the legs and gastrointestinal area. Stomach disease can also induce certain psychiatric diseases.

Diseases of the Stomach organ are primarily caused by improper diet. To determine the root of the disease, the following five areas are considered:

1. The patient's Five Elemental Constitution and present state of health,
2. The type and energetic quality of the foods ingested (Hot or Cold), and whether the food is in season or out of season,
3. The scheduling of meals (This not only includes eating meals at regular times, but also eating a balanced meal, not too fast, and not too late at night),
4. The Yin/Yang balance of the types of food digested (This includes balancing the foods Five Tastes and Five Colors), and
5. The emotional factors surrounding meal time. This includes having the proper mental attitude (being relaxed and calm) and avoiding emotional distress (such as feeling rushed, upset, and so on).

The Stomach easily suffers from Excess patterns (e.g., Fire or Phlegm Fire), which in turn agitate the Shen. When the Shen is disturbed, it can cause manic symptoms such as: inappropriate laughter, pressured speech, unconscious talking, laughing, or singing, violent or otherwise inappropriate behavior (e.g., taking off one's clothes in public). In milder cases, symptoms may include mental confusion, severe anxiety, hypomania (a milder form of mania), and hyperactivity.

Since the Stomach has the function of sending the preliminarily digested food downward, an impairment of this function often causes vomiting.

THE SPLEEN CHANNELS

The Spleen Channels are Yin channels and flow externally from the feet to the torso (Figure 6.12). These two rivers originate externally from the medial tips of the big toes, ascend upwards along the inner thighs, then enter the abdomen (connecting to the CV-3 and CV-4 points) and further up the upper torso (connecting to the CV-12 point). The external Spleen Channels meet at the CV-12 point, enter the body, where the left channel flows into the Spleen, and then spirally wraps the Stomach organ. An internal branch ascends from the CV-12 point into the Heart. At the CV-12 point both left and right main channels meet and separate again to ascend through the diaphragm, up the torso and throat, alongside the esophagus, to reconnect at the base of the tongue. From the Sp-20 points external branches descend along the lateral aspects of the torso and terminate on the mid-axillary line in the seventh intercostal space.

THE SPLEEN ORGAN

The Spleen's associated organ is the Stomach; its element is Earth. In Traditional Chinese Medicine, the pancreas (a gland that produces insulin and is connected to the Spleen) is considered part of the Spleen/Stomach function. The Spleen organ is responsible for transforming food received from the Stomach into Gu Qi. The Gu Qi is then sent to the Liver and Kidneys to be "cooked" and further refined. The Spleen's main function is to govern or oversee this transportation and transformation of Gu Qi. It also controls the Blood by keeping it in the blood vessels, and governs the muscles, flesh, and limbs. The Spleen organ's energy opens externally at the mouth, controlling taste, and connects externally by the lips. The Spleen is sometimes called The Minister of Grains, because it is responsible for distributing what the Stomach has stored. Energetically, the Spleen controls the body's central cavity, and maintains the internal organs in their places.

PSYCHO-EMOTIONAL ASPECTS

The Spleen houses the body's thoughts and intentions (Yi), and is responsible for analytical thinking, memory, cognition, intelligence, and ideas. The Spleen is responsible for directing

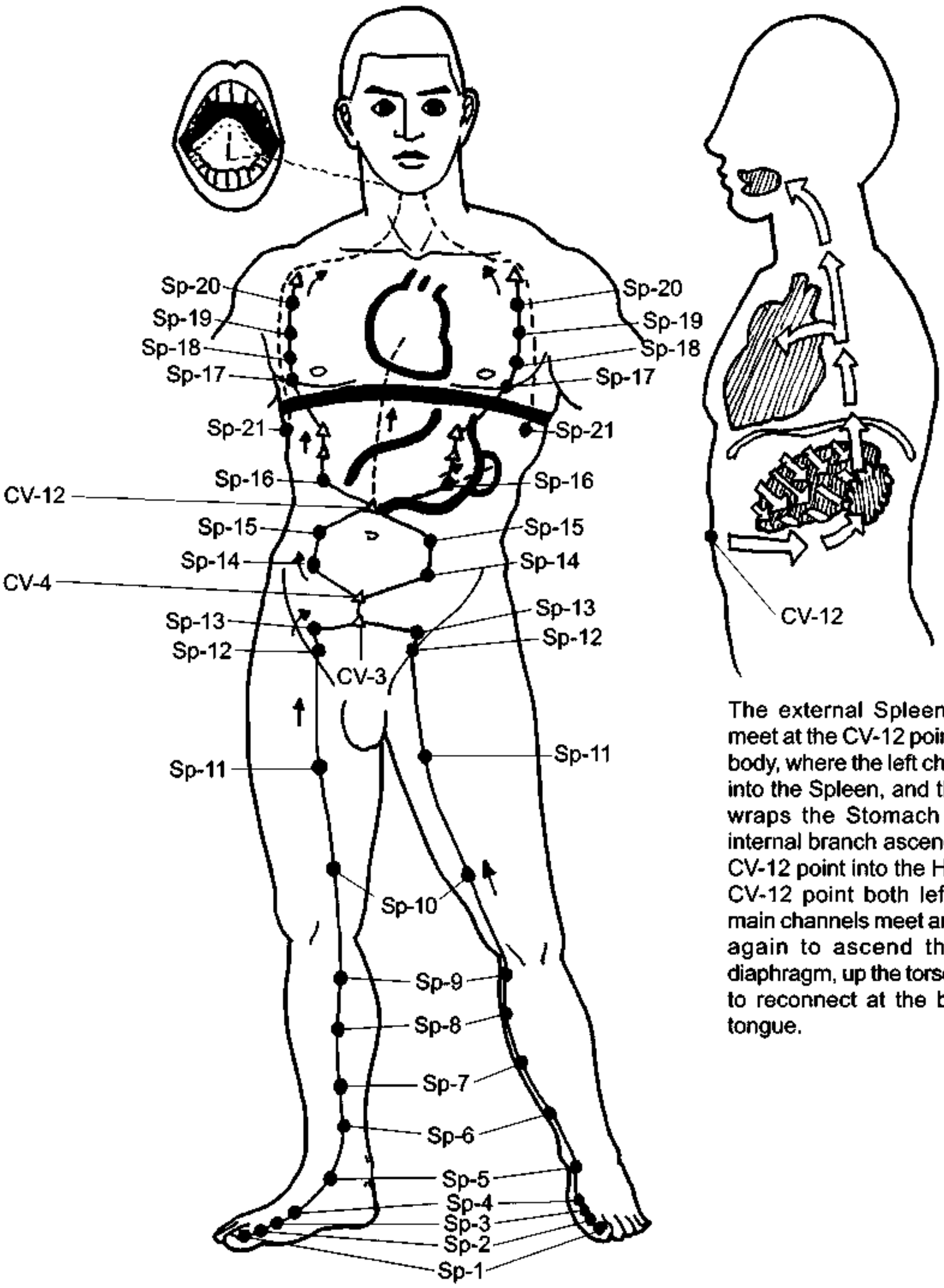
Sp-20
Sp-19
Sp-18
Sp-17
Sp-21
Sp-16
CV-12
Sp-15
Sp-14
CV-4
Sp-13
Sp-12
CV-3
Sp-11
Sp-10
Sp-9
Sp-8
Sp-7
Sp-6
Sp-5
Sp-4
Sp-3
Sp-2
Sp-1
Sp-20
Sp-19
Sp-18
Sp-17
Sp-21
Sp-16
Sp-15
Sp-14
Sp-13
Sp-12
Sp-11
CV-12

The external Spleen Channels meet at the CV-12 point, enter the body, where the left channel flows into the Spleen, and then spirally wraps the Stomach organ. An internal branch ascends from the CV-12 point into the Heart. At the CV-12 point both left and right main channels meet and separate again to ascend through the diaphragm, up the torso and throat to reconnect at the base of the tongue.

Figure 6.12. The Internal and External Qi Flow of the Spleen (Sp) Channels

memories to the Kidneys for short-term storage. The Kidneys will later transfer these memories to the Heart for long-term memory storage (see Chapter 12).

The Spleen's positive psycho-emotional attributes are trust, honesty, openness, acceptance, equanimity, balance, and impartiality; its negative attributes are worry, excessive thinking, pensiveness, obsessiveness, remorse, regret, obsessions, and self-doubt.

THE SPLEEN'S ENERGY FLOW

The Spleen Channels store more Qi and less Blood, acting more on energetic and nervous functions than on physical substance and Blood functions. At the high-tide time period (9 a.m. to 11 a.m.), Qi and Blood abound in the Spleen organ and channels, and can therefore be more easily dispersed and / or purged at this period; whereas during low-tide (9 p.m. to 11 p.m.), they can be more readily tonified. The Spleen Channels' energy acts on the skin, muscles, and nerves found along their pathway.

WESTERN MEDICAL PERSPECTIVE

The spleen is responsible for blood formation, filtration, and storage. In the embryo, all types of blood cells are formed by the spleen, but in the adult, only lymphocytes and monocytes are formed.

PATHOLOGICAL MANIFESTATIONS

The main diseases of the Spleen Channels include gastrointestinal dysfunctions (disturbances of digestion and absorption of food), and diseases of the tongue and throat. Spleen disorders also affect the inner side of the lower extremities along the Spleen's Channels.

Since the Spleen has the function of sending food essence upwards (to the Lungs), if this function is impeded, diarrhea or prolapse of the viscera may occur.

The Spleen also has the function of keeping the Blood flowing in the blood vessels; impairment usually leads to chronic hemorrhagic diseases.

The Spleen has the function of nourishing the flesh (muscles). A person with a healthy Spleen will usually have a healthy figure and a toned body; a diseased Spleen makes the patient lose muscle definition.

The Spleen nourishes the limbs. Strength of the limbs depends upon the nourishment guaranteed by the normal functioning of the Spleen. A diseased Spleen usually causes weakness of the limbs.

THE HEART CHANNELS

The Heart Channels are Yin Channels and flow externally from the torso to the hand (Figure 6.13). Three rivers originate internally from the Heart on each side of the body. The first bilateral set ascends and connects to the tissues connecting the eyeball. The second set descends, first to permeate the Pericardium, then descends further to connect with and spirally wrap the Small Intestine. The third set flows up into the Lungs, emerges externally at the armpits, then descends the medial aspect of the arms and ends on the inside of the little fingers.

THE HEART ORGAN

The main function of the Heart organ consists of controlling the Blood and vessels, regulating the Blood flow, perspiration, and housing the Shen. The Heart organ's energy opens externally at the tongue and is expressed through the complexion.

PSYCHO-EMOTIONAL ASPECTS

The Heart's associated organ is the Small Intestine; its element is Fire. Long-term memory, thinking, emotions, intimacy, cognition, intelligence, and ideas are all dominated by the function of the Heart. The Heart is sometimes called The Emperor, or "supreme controller of all Yin and Yang organs" (see Chapter 12). The Heart houses the body's Spirit (Shen). The Heart dominates sleep; if the Heart is strong the patient will fall asleep easily and sleep soundly. If the Heart is weak, the patient's mind will "float," resulting in an inability to fall asleep, disturbed sleep, or excessive dreaming.

The Heart's positive psycho-emotional attributes are love, joy, peace, contentment, propriety, insight, wisdom, orderliness, forgiveness, and courtesy. Its negative attributes are hate, guilt,

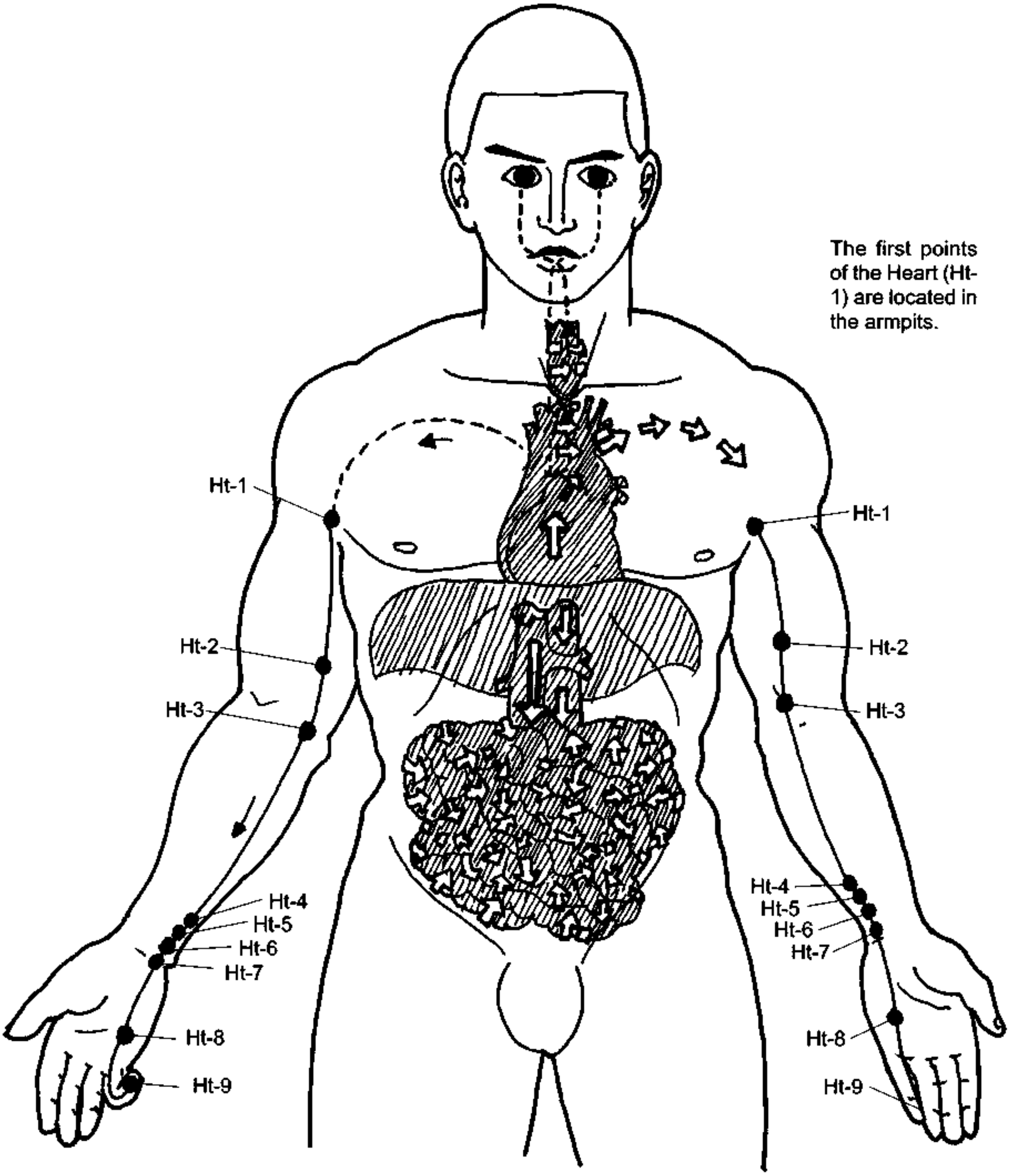
Ht-1
Ht-2
Ht-3
Ht-4
Ht-5
Ht-6
Ht-7
Ht-8
Ht-9
Ht-1
Ht-2
Ht-3
Ht-4
Ht-5
Ht-6
Ht-7
Ht-8
Ht-9

The first points of the Heart (Ht-1) are located in the armpits.

Figure 6.13. The Internal and External Qi Flow of the Heart (Ht) Channels

shock, nervousness, excitement, longing, and craving.

THE HEART'S ENERGY FLOW

The Heart Channels store more Qi than Blood; that is to say, its dominant action affects the energy. This energetic action controls the morale, the spirit of enterprise, and provides the energy required for respiration. At the high-tide time period (11 a.m. to 1 p.m.), Qi and Blood abound in the Heart organ and channels, and can therefore be more easily dispersed or purged at this period; whereas during low-tide (11 p.m. to 1 a.m.), they can be more readily tonified. The Heart Channel's energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The heart provides the propulsive force for circulating the blood throughout the vascular system. The blood circulation distributes the nutrients and oxygen needed for cell and tissue survival. The heart's propulsion also aids in the removal of wastes and other substances through the interconnecting blood vessels.

PATHOLOGICAL MANIFESTATIONS

The main diseases of the Heart Channels include diseases that exert pressure on the brain, eyes, or pharyngeal wall, as well as diseases of the Heart itself, and the lateral side of the chest. The Heart also relates to diseases of the ulnar palmar side of the arm.

The Heart is in charge of mental activities, including consciousness and thinking. Dysfunctions of the Heart can lead to insomnia, impairment of consciousness, amnesia, and psychosis.

The Heart and Kidneys have a mutual energetic relationship of supporting and checking each other. The Heart controls the body's Fire, while the Kidneys control the body's Water. Normally, the Fire of the Heart is sent down to warm the Kidneys, and the Water of the Kidneys is sent up to irrigate the Heart. If this balanced relationship breaks down (especially when the Kidney Water is insufficient to check the Heart Fire), a series of Fire symptoms of the Heart, such as hypertension, hyperactivity, palpitations, and insomnia may result.

Since the Heart has its external opening in the tongue, the condition of the Heart is reflected in the tongue:

1. A dark purple tongue indicates Blood Stasis of the Heart.
2. A pale tongue reveals Deficient Blood of the Heart.
3. An ulcer on the tongue reveals Excess Fire of the Heart.

THE SMALL INTESTINE CHANNELS

The Small Intestine Channels are Yang channels and flow externally from the hands to the head. These two rivers originate externally from the lateral tips of the little fingers (ulnar side) then ascend the arms to the shoulders, where they divide into two sets of branches. One set of branches descends the supraclavicular fossa internally, spiral wrapping the Heart, then continues down the esophagus, diaphragm, and Stomach, before permeating the Small Intestine (Figure 6.14).

The other set of branches ascends externally up the sides of the neck, then divides again into two more sets of branches at the cheek; one set will end at the ears, the other at the lateral sides of the nose and inner canthus of the eyes.

THE SMALL INTESTINE ORGAN

The Small Intestine temporarily stores partially digested food, absorbing the essential substances of the undigested food and a portion of the liquid content. It later transfers the residue (with a considerable amount of liquid) to the Large Intestine, and that is why it is said that the Small Intestine "governs liquid."

The Small Intestine is sometimes called "The Official in Charge of Separating the Pure from the Impure, because it separates the "clean" (reusable) food essence (Gu Qi) from the "dirty, turbid" food essence. The clean Gu Qi is transported throughout the body to nourish the tissues. The turbid is transported to the Large Intestine and Bladder organs to be further processed.

The Small Intestine's associated organ is the Heart; its element is Fire.

PSYCHO-EMOTIONAL ASPECTS

The Small Intestine influences the patient's

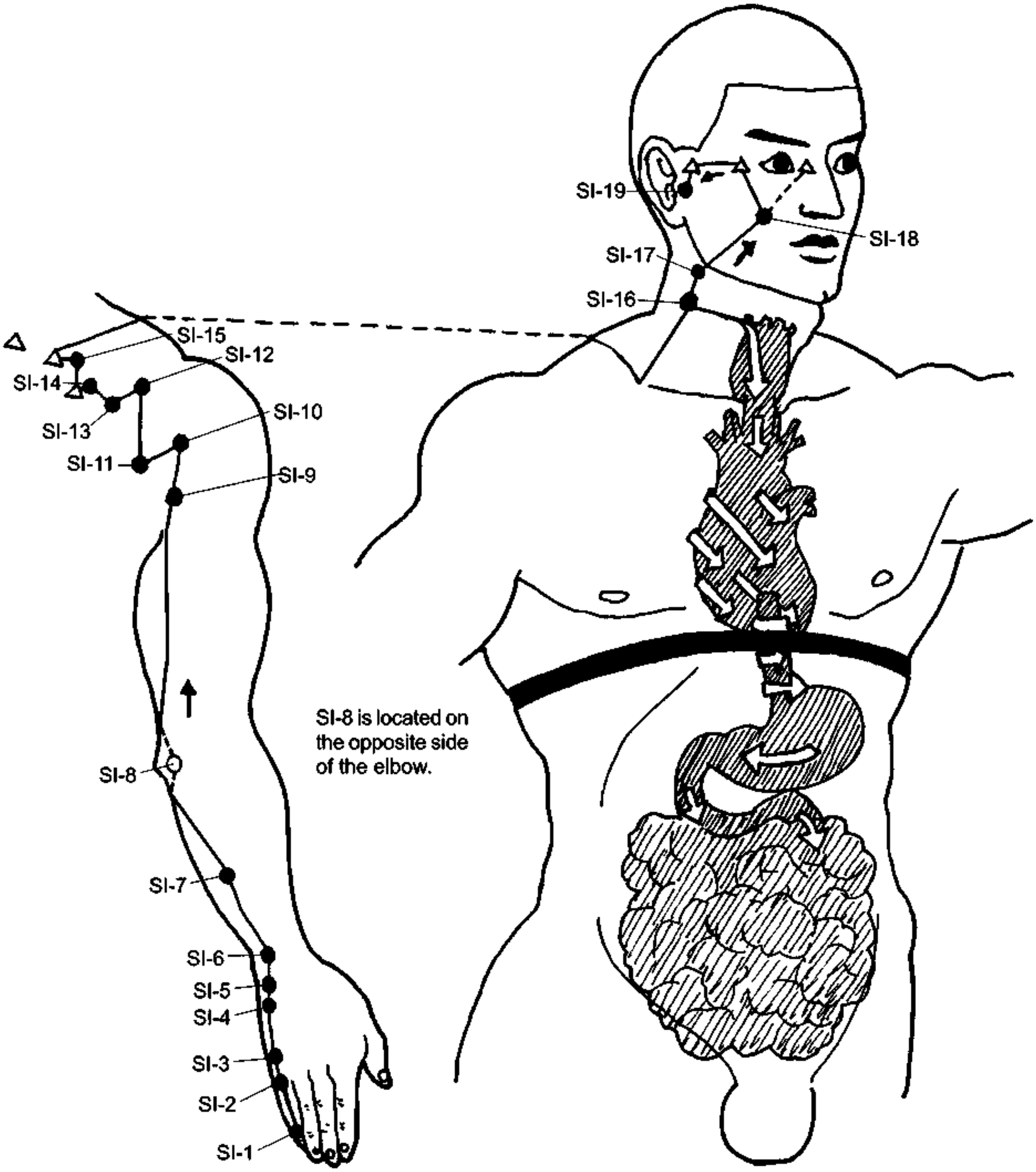
SI-19
SI-18
SI-17
SI-16
SI-15
SI-12
SI-14
SI-13
SI-10
SI-11
SI-9
SI-8 is located on the opposite side of the elbow.
SI-8
SI-7
SI-6
SI-5
SI-4
SI-3
SI-2
SI-1

Figure 6.14. The Internal and External Qi Flow of the Small Intestine (SI) Channels

mental clarity, judgement, and powers of discernment. The ability to distinguish relevant issues with clarity before making a decision is attributed to the Small Intestine.

THE SMALL INTESTINE'S ENERGY FLOW

The Small Intestine Channels store more Blood than Qi, and act more on the physical substances and Blood functions than on energy. At the high-tide time period (1 p.m. to 3 p.m.), Qi and Blood abound in the Small Intestine organ and channels, and can therefore be more easily dispersed and purged at this period; whereas during low-tide (1 a.m. to 3 a.m.), they can be more readily tonified. The Small Intestine Channels' energy act on the skin, muscles, and nerves found along their pathway.

WESTERN MEDICAL PERSPECTIVE

The small intestine's function is to receive the food mass from the stomach through the pylorus, the bile from the liver and gall bladder, and the pancreatic juices for digestion and absorption. The small intestine then passes what it doesn't absorb into the large intestine.

PATHOLOGICAL MANIFESTATIONS

Diseases of the Small Intestine Channels include diseases of the face, ear, cheek, lower jaw, the neck, and throat, as well as the dorsal ulnar side of the upper extremities.

In TCM, food and herbs are divided into Yin (Cold) or Yang (Hot) properties. The Small Intestine is easily affected by the type and temperature of food eaten. An excess consumption of "cold" and raw foods, for example, can create a Cold condition within the Small Intestines, while an excess consumption of "hot" foods can create a Hot condition.

Disturbances of the Small Intestine's energetic function usually gives rise to one of the following complications:

1. Excess Conditions of the Small Intestines can result in symptoms of: Full Heat, Qi pain, Knotted Qi, and the infestation of worms in the Small Intestine.
2. Deficient Conditions of the Small Intestines can manifest as: Deficient and Cold Small Intestine.

Pathologic Heart Heat may be transmitted to the Small Intestine, resulting in urodynia (painful urination) and hematuria (blood in the urine). Herbal prescriptions used to dispel Heat from the Heart are also effective in treating these urinary symptoms, which are alleged to be caused by dysfunctions of the Small Intestine.

THE BLADDER CHANNELS

The Bladder Channels are Yang channels and flow externally from the head to the feet. Two rivers originate externally from the inner canthus of the eyes, ascend upward over the head to join the Governing Vessel at the GV-20 point, where they divide into two additional branches that flow into each temple, and into the brain. From there, they then divide into two sets of branches (at the Bl-10 point), which descend the back, and internally connect to and spiral wrap the Kidneys, before they descend to permeate the Bladder. They continue externally down the thigh, the popliteal fossa at the back of the knee, calf, and foot, and end on the lateral sides of the tips of the little toes (Figure 6.15).

THE BLADDER ORGAN

The function of the Bladder is to remove water by Qi transformation. The Bladder receives the "impure" parts of the fluids separated by the Small Intestine and temporarily stores and transforms these fluids into urine, to discharge it when the Bladder is full. The Bladder and the Small Intestine work together to remove Body Fluids from the Lower Burner.

The Bladder Channels activate the Yang aspect of the Kidney's function and are sometimes called The Water District Officials or "the controllers of the storage of waste water." The Bladder's associated organ is the Kidney; its element is Water.

PSYCHO-EMOTIONAL ASPECTS

An imbalance in the Bladder can cause such psychological symptoms as habitual fear, lack of decision making capability and a diminished moral character. If the imbalance becomes chronic, it results in such emotional responses as jealousy, suspicion, and a holding on to long-standing grudges.

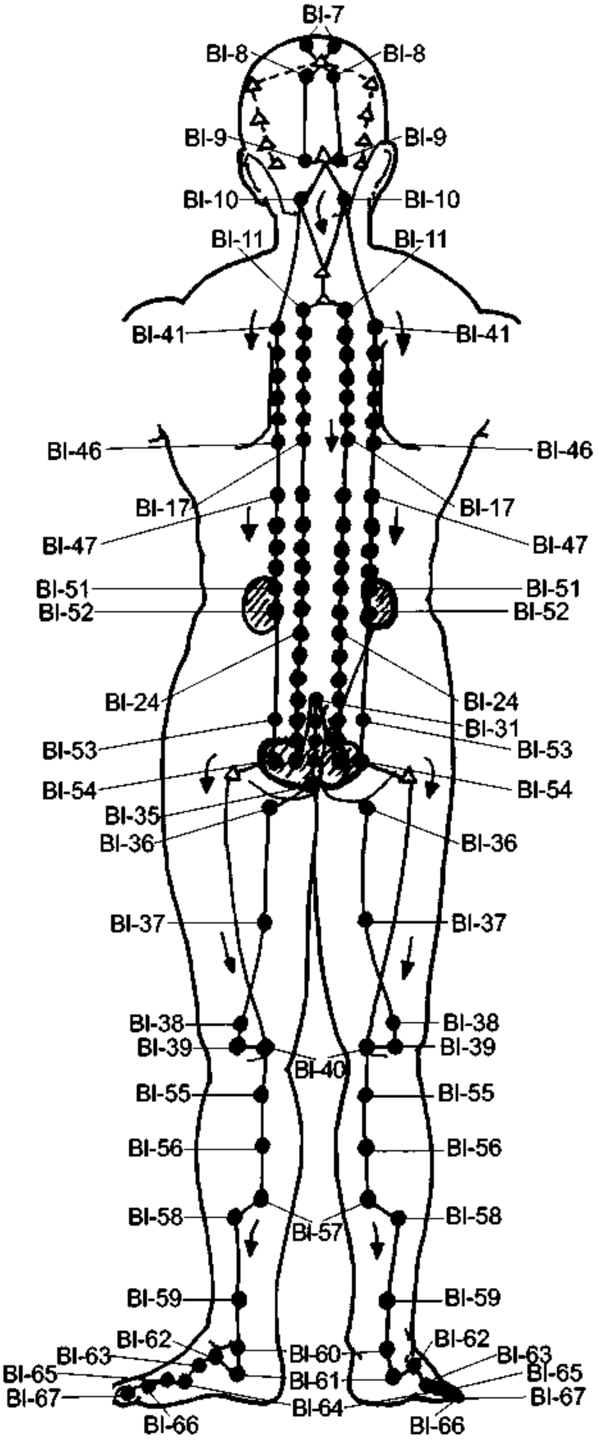
Bl-7
Bl-8
Bl-8
Bl-9
Bl-9
Bl-10
Bl-10
Bl-11
Bl-11
Bl-41
Bl-41
Bl-46
Bl-46
Bl-17
Bl-17
Bl-47
Bl-47
Bl-51
Bl-51
Bl-52
Bl-52
Bl-24
Bl-24
Bl-31
Bl-53
Bl-53
Bl-54
Bl-54
Bl-35
Bl-36
Bl-36
Bl-37
Bl-37
Bl-38
Bl-38
Bl-39
Bl-39
Bl-40
Bl-55
Bl-55
Bl-56
Bl-56
Bl-58
Bl-57
Bl-58
Bl-59
Bl-59
Bl-62
Bl-63
Bl-65
Bl-67
Bl-60
Bl-61
Bl-64
Bl-66
Bl-62
Bl-63
Bl-65
Bl-67
Bl-66

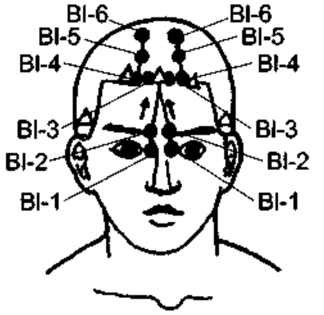
Bl-6
Bl-6
Bl-5
Bl-5
Bl-4
Bl-4
Bl-3
Bl-3
Bl-2
Bl-2
Bl-1
Bl-1

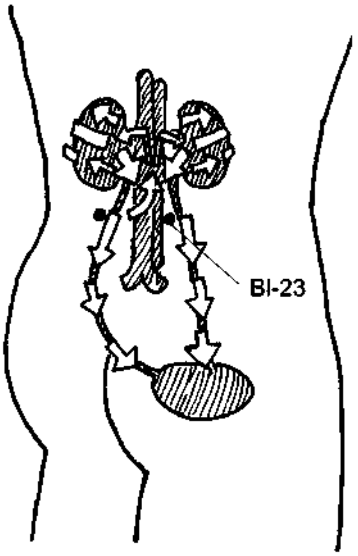

The internal flow of the Bladder Channels rushes into the lower back and lumbar area, passing through the renal arteries, to spiral wrap the Kidneys. From the Kidneys, the energy flows downward, following into the ureter, to permeate the Bladder.

Figure 6.15. The Internal and External Qi Flow of the Bladder (Bl) Channels

THE BLADDER'S ENERGY FLOW

The Bladder Channels regulate the functions of the Kidneys. The Bladder Channels store more Blood than Qi, acting more on the physical substances and the Blood functions than on the energy. At the high-tide time period (3 p.m. to 5 p.m.), Qi and Blood abound in the Bladder organ and channels, and can therefore be more easily dispersed and purged at this period; whereas during low-tide (3 a.m. to 5 a.m.), they can be more readily tonified. The Bladder Channels' energy acts on the skin, muscles, and nerves found along their pathway.

WESTERN MEDICAL PERSPECTIVE

The bladder is a reservoir for the urine received from the kidneys; it discharges urine from the body through the urethra.

PATHOLOGICAL MANIFESTATIONS

The main diseases of the Bladder Channels include:

- Diseases of the top of the head,
- Brain disorders,
- Disorders of the neck and back, especially the lumbar and sacral regions,
- Disorders of the back of the legs and thighs, and
- Disorders of the lateral side of the feet.

Diseases of the Bladder Organ manifest in changes in urine and urination; these changes will reflect a Deficient or Excess condition of the Bladder.

1. Deficient Conditions are attributed to Deficient Kidney Qi which affects the Bladder's ability to transform Qi. This dysfunction causes frequent urination, dribbling, or enuresis (involuntary discharge of urine).
2. Excess Conditions are attributed to Damp Heat in the Bladder, resulting in symptoms such as:
   - heat and pain during urination,
   - the short release of murky or reddish urine,
   - frequent difficulty in urination,
   - pus, or Blood in the urine, and
   - Bladder stones which cause a urinary block and painful distention of the lower abdomen.

The most common pathological factor in the diseases of the Bladder organ is due to the accumulation of Dampness (e.g., Damp Heat or Damp Cold). Other pathological conditions include: emotional fears, anxieties and insecurities, excessive sex, and Deficient and Cold Bladder.

THE KIDNEY CHANNELS

The Kidney Channels are Yin channels and flow externally from the feet to the torso (Figure 6.16). These two rivers originate externally from underneath the little toes, circling the inside of the heels, ascending through the medial aspect of the legs, where they merge and enter into the coccyx and lower lumbar vertebrae. Dividing again, one branch permeates the Kidneys, while the other branch continues ascending into the cerebral cortex. From the Kidney organs, two additional sets of channels internally emerge. One set descends and spiral wraps the Bladder. The other set ascends into the Liver, diaphragm, and Lungs, then spiral wraps the Heart and travels up the throat, stopping at the root of the tongue (Figure 6.17).

THE KIDNEY ORGANS

The Kidneys store Jing substances, produce Marrow, and control the development of the bones. They also control respiration, reproduction, growth and development, and govern the Body Fluids. The Kidney Jing is the biological basis for the woman's menstrual blood. The Kidneys are the root of the Yin and Yang Qi of all the body's organs. The Kidneys energetically open externally through the ears and are sometimes called The Minister of Ingenuity and Vitality, or "the controller of water" (see Chapter 12).

The Kidney organs' associated organ is the Bladder and it's element is Water. Kidney Yin Qi flows to the Liver, Heart, and Lungs. It is responsible for the body's fluid-like essences and rules the birth, growth, maturation, and reproduction cycle. Kidney Yang Qi flows to the Spleen, Liver, Heart, and Lungs. It supports the Yang of all the body's organs via the Mingmen.

PSYCHO-EMOTIONAL ASPECTS

The Marrow produced from the Kidney Jing flows into the brain. The thinking ability is strengthened when Qi and Blood in the cerebral

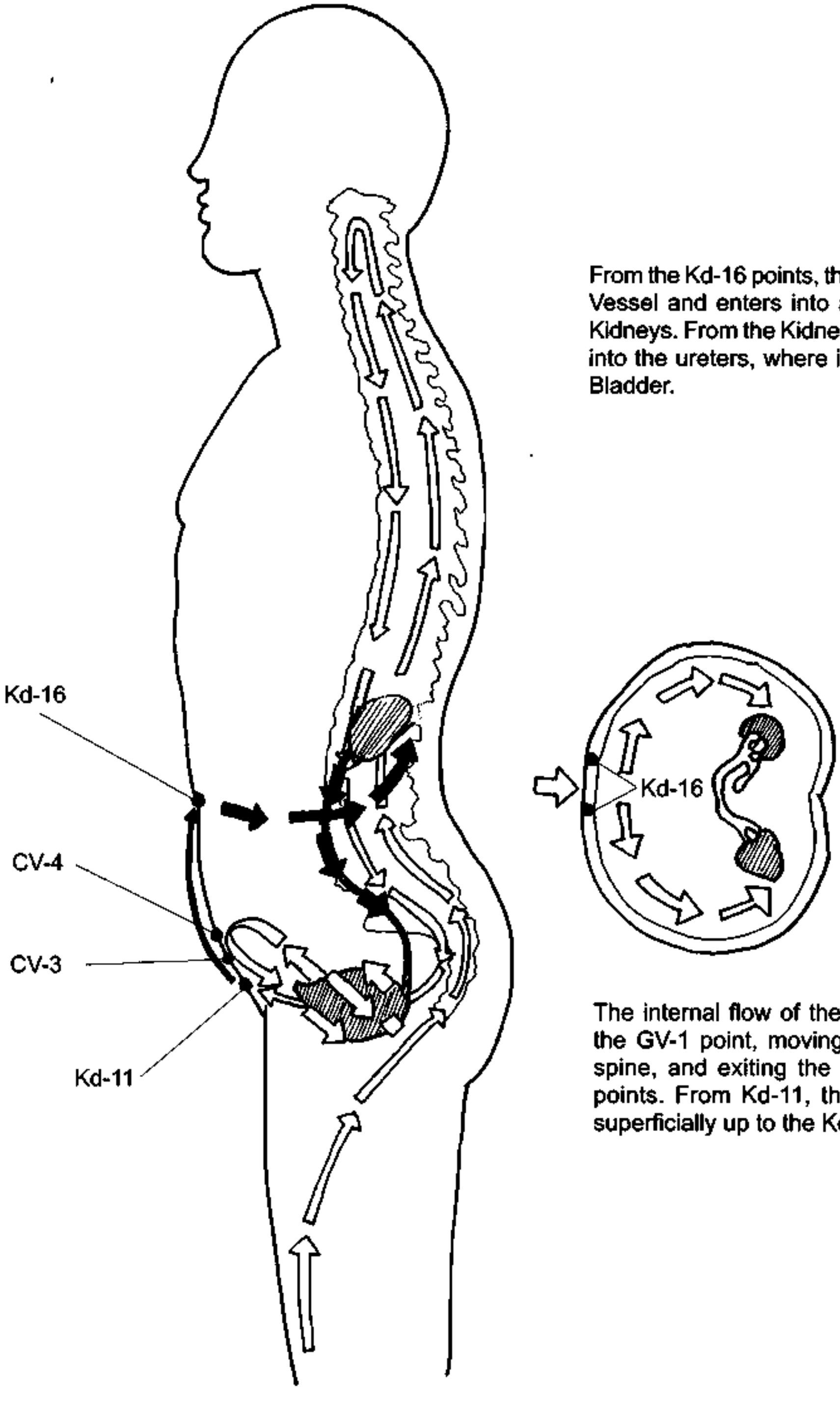

From the Kd-16 points, the Qi follows the Belt Vessel and enters into and permeates the Kidneys. From the Kidneys, the Qi descends into the ureters, where it spirally wraps the Bladder.

The internal flow of the Kidney Qi passes the GV-1 point, moving up and down the spine, and exiting the body at the Kd-11 points. From Kd-11, the energy ascends superficially up to the Kd-16 points.

Figure 6.16. The Internal Flow of the Kidney Qi

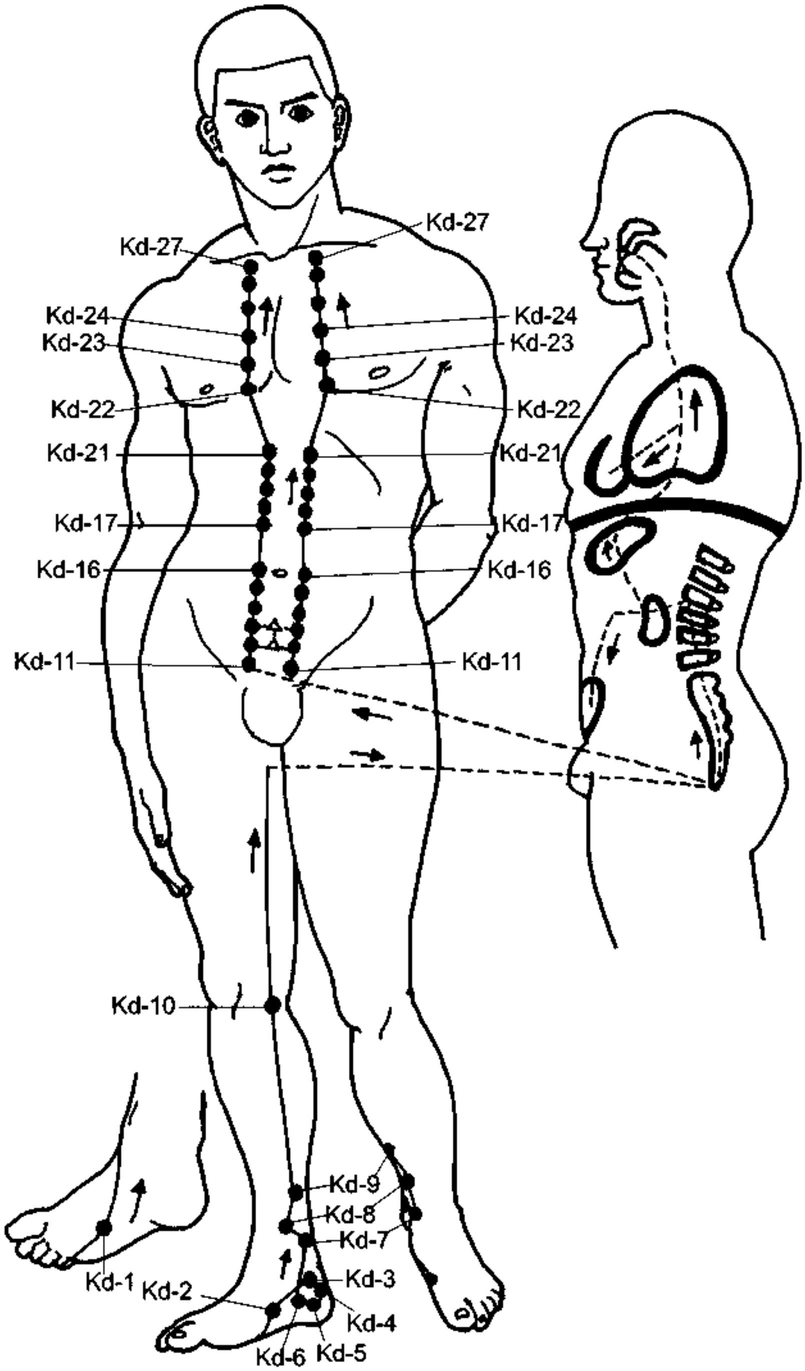
Kd-27
Kd-27
Kd-24
Kd-24
Kd-23
Kd-23
Kd-22
Kd-22
Kd-21
Kd-21
Kd-17
Kd-17
Kd-16
Kd-16
Kd-11
Kd-11
Kd-10
Kd-9
Kd-8
Kd-7
Kd-3
Kd-1
Kd-2
Kd-4
Kd-6
Kd-5

Figure 6.17. The Internal and External Qi Flow of the Kidney (Kd) Channels

cortex are abundant. With the increase of thinking activity, a circle of light develops within the body's Taiji Pole. The individual's innate intelligence (Yuan Shen) is represented by the intensity of light. The degree of accumulated spiritual energy is reflected by the number of light circles developed within the Taiji Pole. These circles of light can be best observed when first waking. By placing slight pressure on the external eye lids the inner light of the Taiji Pole is projected onto the optic nerves, reflecting an image of the circle of light. If the circle of light is complete, it reflects a strong, healthy condition. If the circle of light is dark within its center (similar to a doughnut), this reflects a Deficient condition. If the circle of light is broken or interrupted, it reflects an extreme Deficiency (Figure 6.18).

The "memory zone," as well as the "thought center" are also located in the cerebral cortex and will not develop until the Kidney Channels travel through the spine, along with the Liver Channels, to reach the cortex. When the Qi of these two channels is abundant, the memory function is keen.

The Kidneys house the body's will power (Zhi). They control short-term memory and store data. The Kidneys provide the capacity and drive for strength, skill, and hard work. A patient with strong Kidneys can work hard and purposefully for long periods of time. Consequently, when the Kidneys are in a state of disharmony, the patient can sometimes be driven to a state of excessive-compulsive working habits (a workaholic). A patient with weak Kidneys will lack strength and endurance.

The Kidney's positive psycho-emotional attributes are wisdom, rationality, clear perception, gentleness, and self-understanding. The negative attributes are fear, loneliness, insecurity, and shock (which attacks the Heart first then descends into the Kidneys to become fear).

THE KIDNEYS' ENERGETIC FLOW

The Kidney Channels store more Qi than Blood, acting more on energetic and nervous functions than on physical substances and Blood functions. At the high-tide time period (5 p.m. to 7 p.m.), Qi and Blood abound in the Kidney organ and channels, and can therefore be more easily dispersed and purged at this period; whereas during low-tide (5 a.m. to 7 am.), they can be more readily tonified. The Kidney Channels' energy acts on the skin, muscles, and nerves found along their pathways.

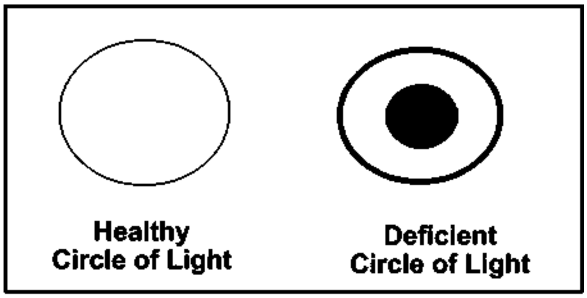

Figure 6.18. The individual's Yuan Shen is represented by the intensity of the light, and the degree of accumulated spiritual energy is reflected by the number of light circles developed within the Taiji Pole.

WESTERN MEDICAL PERSPECTIVE

The function of the kidneys is to regulate water and electrolytes and to process the body's urine.

PATHOLOGICAL MANIFESTATIONS

Kidney Channel diseases cause general deterioration of the entire body, weakness in the lower extremities, lumbar pain, and hot sensations deep inside the feet.

The Kidneys open through the urogenital orifices and the anus. The energetic condition of the Kidneys can be, partially reflected by the condition of the patient's urination and defecation; in males, this includes the ejaculation process.

Since the Kidneys are responsible for concentration and memory retention, poor concentration and loss of memory are common symptoms of Kidney hypofunction.

THE PERICARDIUM CHANNELS

The Pericardium Channels are Yin channels and flow externally from the torso to the hands. These two rivers originate internally from the center of the chest, flowing out of the Pericardium. They descend through the center of the body, spi-

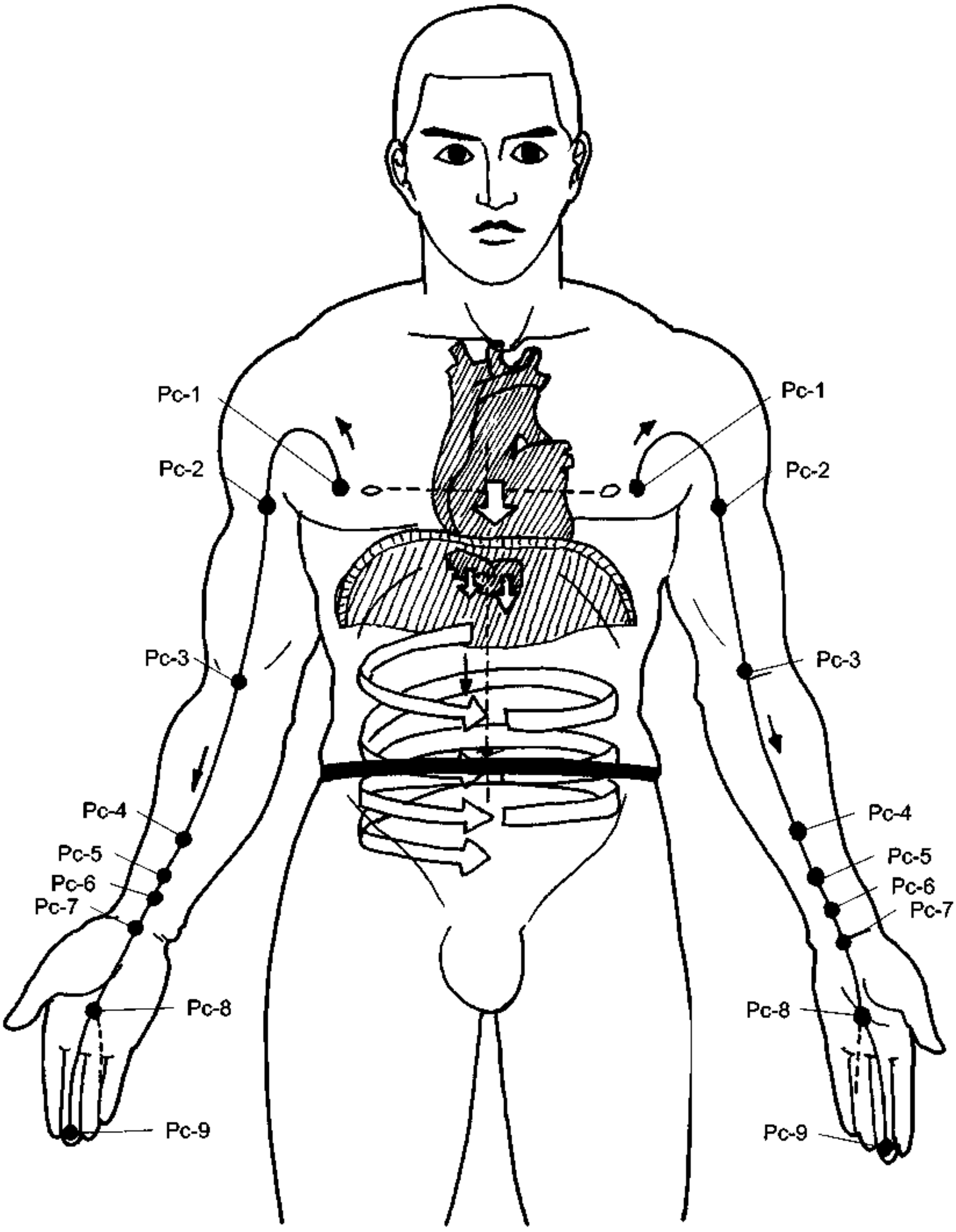
Pc-1
Pc-2
Pc-3
Pc-4
Pc-5
Pc-6
Pc-7
Pc-8
Pc-9
Pc-1
Pc-2
Pc-3
Pc-4
Pc-5
Pc-6
Pc-7
Pc-8
Pc-9

Figure 6.19. The Internal and External Qi Flow of the Pericardium (Pc) Channels

ral wrapping the Upper, Middle, and Lower Burners. Two rivers surface and branch externally by the nipples to proceed down the center of each arm, ending at the tips of the middle fingers. A second branch arises from the center of each palm at the Pc-8 point and flows into the ring fingers to connect with the Triple Burner Channels (Figure 6.19).

PERICARDIUM ORGAN

The Pericardium is associated with the Triple Burners; its element is Fire. The Pericardium is not considered an independent organ, as much as a protective covering for the Heart, that protects the Heart from pathogenic factors. The Pericardium activates, energizes and controls the Yin channel's distribution of the Kidney Yang Qi to the Yin organs. Along with the Heart, it governs the Blood, and houses the Shen. The Pericardium is considered the Prime Minister, or the official who protects the heart.

PSYCHO-EMOTIONAL ASPECTS

The Pericardium has a powerful influence on the patient's mental and emotional states. Its goal is to "create feelings of joy and/or pleasure for the emperor (Heart)."

THE PERICARDIUM'S ENERGY FLOW

The Pericardium Channels have an affect on the body's circulation of Blood. They are considered the "mother of Yin" and are connected to the Mingmen. The Pericardium Channels store more Blood than Qi, acting more on the physical substance and Blood functions than on the energy. At the high-tide time period (7 p.m. to 9 p.m.), Qi and Blood abound in the Pericardium organ and channels and can therefore be more easily dispersed and purged at this period; whereas during low-tide (7 a.m. to 9 a.m.), they can be more readily tonified. The Pericardium Channel's energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The function of the pericardium is to provide a protective covering for the heart. It contains a thin serous fluid that allows the heart to work in a relatively friction-free environment.

PATHOLOGICAL MANIFESTATIONS

The main diseases of the Pericardium Channels include diseases of the Heart, front of the chest, major blood vessels, and diseases on the midline to upper palmar side of the upper extremities. Mental abnormalities may also manifest.

THE TRIPLE BURNER CHANNELS

The Triple Burner Channels are Yang channels that flow externally from the hands to the head. These two rivers originate externally from the tips of the ring fingers and ascend along the lateral aspect of the arms, over the shoulders to the clavicles, then branch internally and spiral wrap the Pericardium. They circle the diaphragm and permeate the Upper, Middle, and Lower Burners. A set of branches originate internally from the chest CV-17 point at the Heart, or center of the Middle Dantian. Flowing upwards, these branches emerge from the supraclavicular fossa at the St-12 points to ascend up the neck, head, and then circle the ears. Another set of branches flows downward to the cheek, terminating in the infraorbital region; the other set ends above the ears, by the outer canthus of the eyes.

THE TRIPLE BURNERS ORGAN

The Triple Burners' associated organ is the Pericardium; its element is Fire. The Triple Burners are not regarded as an independent organ, but are assigned to specific energy areas, and are a part of the function of the Yin and Yang organs. The function of the Triple Burners' energy is to produce Heat, and to regulate the body's temperature. This Heat is increased through such meditative disciplines as the Microcosmic Orbit (which connects the Governing and Conception Vessels), and the Macrocosmic Orbit (which connects all Twelve Primary Channels with the Governing and Conception Vessels). The main function of the Triple Burners (Figure 6.20) is to regulate the general ingestion and digestion of food and fluids throughout the body.

The Triple Burners are known as The Official of Balance and Harmony, because they govern Water metabolism, control the production of Wei

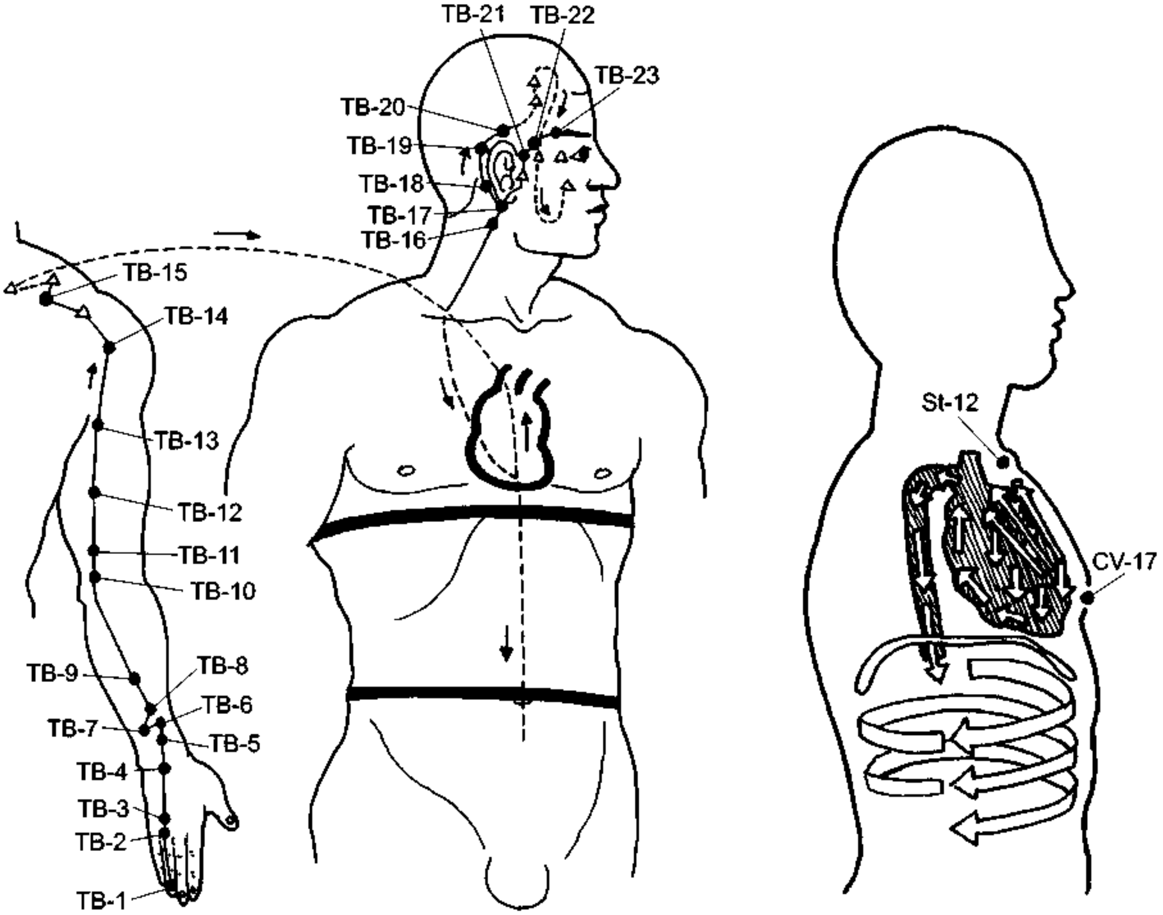
TB-21
TB-22
TB-23
TB-20
TB-19
TB-18
TB-17
TB-16
TB-15
TB-14
TB-13
TB-12
TB-11
TB-10
TB-9
TB-8
TB-7
TB-6
TB-5
TB-4
TB-3
TB-2
TB-1
St-12
CV-17

The internal flow of the Triple Burner Channels moves from St-12 down to CV-17, where it disperses into the chest and envelops the Pericardium. From the Pericardium, as the energy flows down through the diaphragm, it circulates through the chest and permeates the Triple Burners (a continuing branch flows upward to the St-12 points).

Figure 6.20. The Internal and External Qi Flow of the Triple Burners (TB) Channel

Qi, and transport Yuan Qi from the Kidneys to all the other organs of the body. This name refers to the Triple Burners' role in regulating metabolic functions, as well as their production of Qi, Blood, Body Fluids, and waste.

The Triple Burners' energy is composed of Zong Qi (Essential Qi). The Zong Qi assists the Heart in circulating the Blood and the Lungs in respiration. The Zong Qi also assists in conveying the Ying Qi (Nourishing Qi) to nourish the Blood, organs, and tissues of the body, as well as strengthens the Wei Qi (Protective Qi) that protects the external body (see Chapter 14).

1. The Upper Burner

   The Upper Burner is formed from the Fire created from the combined energies of the Heart, Pericardium, and both Lungs. The Upper Burner's energy is housed from the area of the head, throat, and upper chest, to the diaphragm. The Upper Burner is responsible for respiratory and cardiac functions. It moves the body's finer energy, circulating and distributing nutrients and Qi throughout the body like a "mist."

2. The Middle Burner

   The Middle Burner is formed from the Fire created from the combined energies of the Stomach, Spleen, pancreas, and Gall Bladder. The Middle Burner's energy is housed from the upper abdomen (the diaphragm) and the umbilicus.

   The Middle Burner is responsible for digestion, fermentation, and transformation of food and drink into nutrients for distribution. In the Middle Burner area, the Small Intestine connects downward to the Bladder. The Small Intestine transforms waste, that is then sent down to the Large Intestine, and also distills the body's fluids, sending them down to the Bladder.

   The Ying Qi (Nourishing Qi) of the Middle Burner receives its nourishment from the Stomach and Spleen. The Spleen extracts Gu Qi from the food "prepared" by the Stomach, churning the food essence into "foam." The Spleen refines this energetic foam and sends the processed Gu Qi to the Lungs. The Lungs further refine the Gu Qi, sending the impure part to the Kidneys (for additional refinement of the Gu Qi). The Kidneys return the Clean Qi to the Lungs and send the turbid part to the Bladder to be expelled from the body. The Qi from the air mixes with the Gu Qi in the Lungs to produce Zong Qi. The Lungs circulate the refined Gu Qi in the form of vapor or mist which is housed in the Upper Burner. The Heart then produces Blood from this vapor.

3. The Lower Burner

   The Lower Burner is formed from the Fire created from the combined energies of the Liver, Kidneys, Bladder, Intestines, and genitalia. The Lower Burner's energy is housed in the area which starts just below the umbilicus and extends to the lower perineum.

   The Lower Burner is responsible for the reproductive functions and for the filtering and elimination of waste products (Figure 6.21). It moves the body's coarser energy, acting as a "channel for water."

The Triple Burners contribute to the process of the three stages of transformation and ais in the distribution of Ying and Wei Qi throughout the body as follows:

- After the Gu Qi (food energy) has been separated into Clean and Turbid Qi, the Upper Burner releases the body's clean Wei Qi, directing it to the Lungs;
- the Middle Burner releases the body's clean Ying Qi, directing it to all of the body's organs and tissues; and
- the Lower Burner releases the body's Body Fluids, directing the turbid part to the Bladder.

PSYCHO-EMOTIONAL ASPECTS

The Triple Burners are considered the Ambassadors or "intermediaries" for the body's Yuan Qi. On a psychological level, they can be used to move Qi and lift depression derived from stagnation of Liver Qi.

When the Triple Burners, which regulate the consciousness, are full, the consciousness becomes stable and the Mind's intent is benevolent and kindhearted. The Triple Burners are also linked

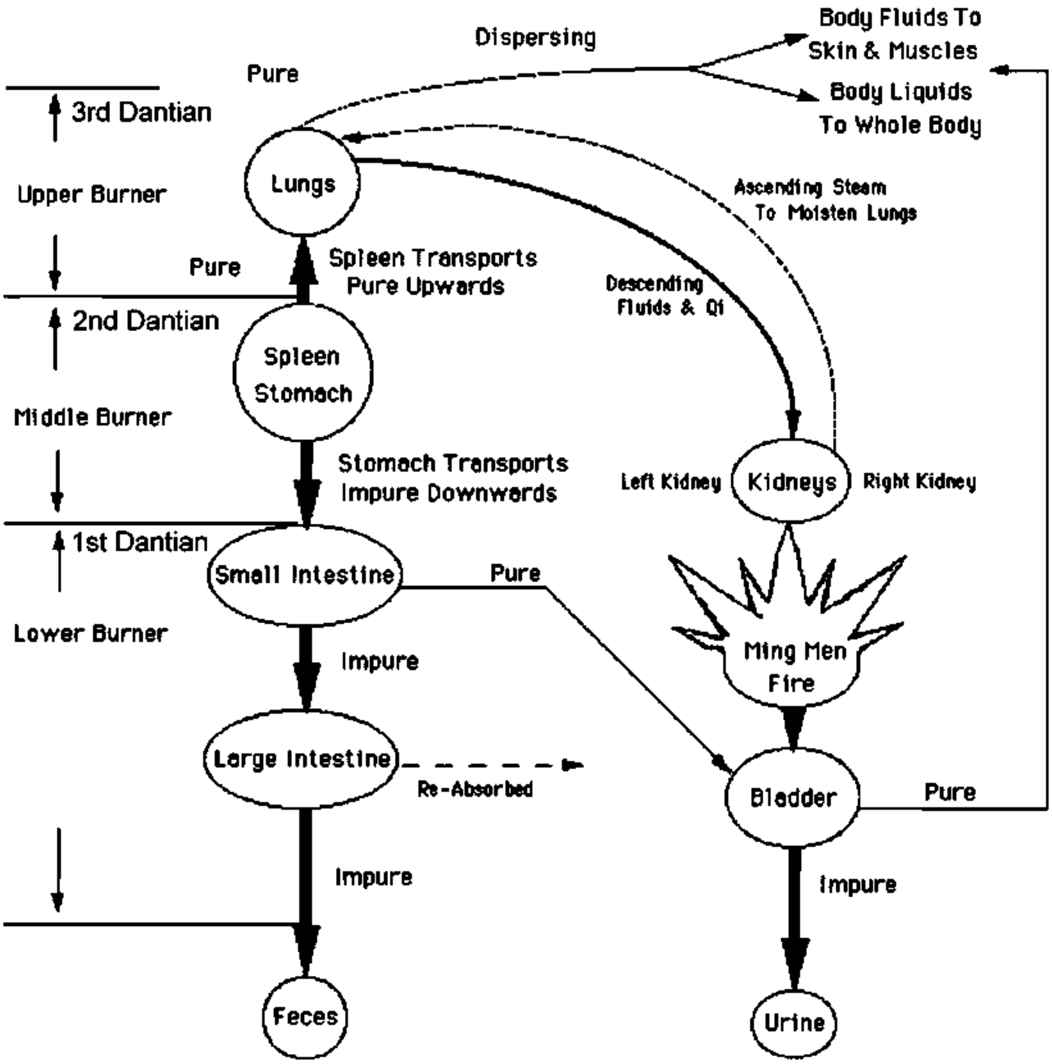
Dispersing
Body Fluids To Skin & Muscles
Body Liquids To Whole Body
Pure
3rd Dantian
Upper Burner
Lungs
Ascending Steam To Moisten Lungs
Pure
Spleen Transports Pure Upwards
Descending Fluids & Qi
2nd Dantian
Spleen Stomach
Middle Burner
Stomach Transports Impure Downwards
Left Kidney
Kidneys
Right Kidney
1st Dantian
Small Intestine
Pure
Lower Burner
Ming Men Fire
Impure
Large Intestine
Re-Absorbed
Bladder
Pure
Impure
Impure
Feces
Urine

Figure 6.21. The Metabolic Functions of the Triple Burners are divided into three parts or functions: the Upper Burner, Middle Burner, and Lower Burner. These different functions of the Triple Burners control the movement of various types of Qi in the three stages of energy production:

Receiving—The Upper Burner, also known as mist, includes the organs from the diaphragm upwards (the Pericardium, head, throat, Lungs, and Heart). The Upper Burner assists the Spleen in directing the body's clean fluids upwards and helps the Lungs in dispersing and scattering the fluids to the skin and muscles.

Rotting and Ripening—The Middle Burner, also known as the muddy pool, and includes the area between the diaphragm and umbilicus (Stomach, Spleen, and Gall Bladder). The Middle Burner assists the Stomach in the transformation and transportation of Fluids and Liquids and in directing the impure part downwards.

Excreting—The Lower Burner, also known as the drainage ditch, includes the organs below the umbilicus (Liver, Kidneys, Intestines, and Bladder). The Lower Burner assists the Small Intestine, the Blood, and the Kidneys in transforming and transporting liquids and wastes.

with the Heart and Pericardium and are affected by the emotion of joy.

When the energy of the Heart is strong and pure (without guilt), and the desires and thoughts of an individual are at peace, then the energy of the body's sexual essence (Jing) will spread into the Triple Burners, and the Blood will flourish within the individual's vessels. If the "fire of desire" is allowed to Heat and combine with the energy of the Triple Burners, the energy of the individual's sexual essence will overflow, mixing itself with the energy of the Mingmen and will leave the body via the reproductive organs and tissues. This leads to Jing and Qi depletion.

THE TRIPLE BURNERS' ENERGY FLOW

The Triple Burner Channels store more Qi than Blood, and act more on the energetic and nervous functions of the body than on physical substance and Blood functions. At the high-tide time period (9 p.m. to 11 p.m.), Qi and Blood abound in the Triple Burner Channels, and can therefore be more easily dispersed and purged at this period; whereas during low-tide (9 a.m. to 11 a.m.), they can be more readily tonified. The Triple Burner Channels' energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

Western medicine recognizes no translation of, or reference to the Triple Burners.

PATHOLOGICAL MANIFESTATIONS

The main diseases along the Triple Burner Channels involve the face, ear, cheek, larynx, and neck. Diseases of the Triple Burner Channels also include disorders of the back of the upper extremities from the midline to the upper arm and forearm.

When diagnosing problems due to dysfunctions of the Triple Burners, the Qigong doctor considers the following.

1. A blockage of the Wei Qi located within the Upper Burner causes an impairment of the Lungs' dispersing function. This can result in the invasion of the Lungs by External Evils (e.g., Wind and Heat, etc.), that penetrate the Pericardium (which corresponds to the initial stage of externally contracted Heat diseases).
2. A blockage of the Ying Qi located within the Middle Burner causes an impairment of the Spleen's transporting function. This can result in gastrointestinal Heat stagnation, and cause Spleen and Stomach Damp Heat (which corresponds to the second stage of externally contracted Heat diseases).
3. A blockage of the Body Fluids located within the Lower Burner causes an impairment of the Bladder's transformation function. This results in the deep penetration of Toxic Evils, weakening the body's Kidney Yin. This in turn can cause Deficient Liver Blood, and Wind stirring due to Empty Yin (this corresponds to the advance stages of externally contracted Heat diseases).

UNDERSTANDING CHANNEL PATHOLOGY

Channel pathology is the oldest of all the modes of pathological pattern classification, dating back to the Spiritual Axis from the *Yellow Emperor's Classics on Internal Medicine*. In understanding a channel's pathology, the Qigong doctor takes into consideration several aspects of the patient's energy flow (i.e., the channels are viewed as exterior, whereas the organs are viewed as interior). Channel pathology is often related to organ disturbances, but it can also be distinct from organ pathology (Figure 6.22).

The disease of one channel may cause disease in other channels and organs. Likewise, tonification of one channel may cause a tonification of other channels and organs.

THE FOUR CAUSES OF CHANNEL PATHOLOGY

1. The Invasion of External Pathogens such as Cold, Wind, Heat, and Damp leads to channel pathology. The exogenous pathogenic factors often settle in the joints, causing Bi Syndrome (painful obstruction). Channel pathology is quite closely related to joint pathology. In Traditional Chinese Medicine, joints are important to the circulation of Qi and Blood. Qi and Blood gather and concentrate in the joints. Qi enters and exits the channels at the joints. The Five Source Points, also called Shu Points (see Chapter 8), are usually located on

of near the joints. It is at these points that the External pathogenic influences may enter the body and settle in the joints. These External pathogenic influences upset the Yin and Yang balance and block the flow of Qi and Blood, causing pain and swelling. The body's joints are also affected by Deficient Qi and Blood, causing local weakness and pain from lack of movement.

2. Physical strain or the overuse of certain joints can cause local stagnation that creates pain and weakness.
3. Injuries, trauma, and strain can cause local stagnation resulting in stiffness, bruising, and pain.
4. Internal organ disharmonies may also affect the channels.

DIFFERENTIATION BY CHANNEL FULL/EXCESS AND EMPTY/DEFICIENT

When a channel is too full, its pattern characteristics are that of intense pain, stiffness, contractions, and cramps. A red color along the flow of the channel indicates Heat; a bluish color along the flow of the channel indicates Cold.

When a channel is Empty, its pattern characteristics are that of a dull ache, weakness in the muscles, numbness, and muscle atrophy along the channel.

DIFFERENTIATION OF CHANNEL PATTERNS BY SPECIFIC CHANNELS

This clinical overview of the body's channels was originally expressed through the knowledge found in the Spiritual Axis (for more in-depth dissertation please see *The Foundations of Chinese Medicine* by Giovanni Maciocia).

1. The diseases of the Gall Bladder Channels may cause: Alternating chills and fever, headache, acute onset of deafness, pain in the hip and the sides of the body, pain along the lateral sides of the legs, and pain and distention of the breasts.
2. The diseases of the Liver Channels may cause: Headache, pain and swelling of the eyes, and cramps in the legs.
3. The diseases of the Lung Channels may cause: Fever, aversion to cold, stiffness in the chest, pain in the shoulders, clavicle, and arms.
4. The diseases of the Large Intestine Channels may cause: Sore throat, toothache, nosebleed, runny nose, swollen and painful gums, swollen eyes, and pain along the channels.
5. The diseases of the Stomach Channels may cause: Pain in the eyes, nosebleed, neck swelling, facial paralysis, cold legs and feet, and pain along the channels.
6. The diseases of the Spleen Channels may cause: Vaginal discharge, weakness of the leg muscles, and a cold feeling along the channels.
7. The diseases of the Heart Channels may cause: Pain in the eyes, pain along the scapula, and pain along the inner side of the arms.
8. The diseases of the Small Intestine Channels may cause: Pain and stiffness in the neck, and pain along the lateral side of the scapula, elbow, and arms.
9. The diseases of the Bladder Channels may cause: Fever and aversion to cold, headache, stiff neck, pain in the lower back, pain in the eyes, pain in the backside of the leg along the channels.
10. The diseases of the Kidney Channels may cause: Lower back pain and pain in the sole of the feet.
11. The diseases of the Pericardium Channels may cause: Stiff neck, contraction of the elbow and hand, and pain along the course of the channels.
12. The diseases of the Triple Burner Channels may cause: Alternating chills with fever, acute onset of deafness, pain and discharge from the ear, pain at the top of the shoulders, pain in the elbow, and pain along the course of the channels.

SUMMARY OF TWELVE PRIMARY CHANNEL PATHOLOGY

The clinical significance of studying the Twelve Primary Channels is evident through the observation of certain pathological manifestations that indicate specific patterns peculiar to that channel and its internal organ. The quality, quantity and proportions of Qi and Blood circulation

Twelve Primary Channels	**Pathologies**
Bladder	• Diseases of the top of the head, brain, neck, back, lumbar and sacral region • Diseases of the back of the leg and thigh and the external side of the foot
Gall Bladder	• Pain in upper right- and left-quadrants of abdomen • Diseases of the head, face, eye, ear • Diseases of the external sides of the lower extremities
Heart	• Diseases that exert pressure on the brain, eyes, or pharyngeal wall • Diseases of the Heart and lateral side of the chest • Diseases of the ulnar palmar side of the upper extremities • Insomnia, impairment of consciousness, amnesia, and psychosis
Kidneys	• Diseases that cause general deterioration of the entire body • Weakness in the lower extremities and lumbar pain • Hot sensations deep inside the feet
Large Intestine	• Diseases of the lower part of the face, nose, oral cavity, teeth, throat, and front part of the neck • Diseases of the back and radial side of the upper extremities
Liver	• Swelling and a distended sensation of the hypochondrium • Diseases of the lower abdomen and genital organs
Lungs	• Chest and Lung diseases • Diseases on the radial side of the upper arm and palmar area of the hand
Pericardium	• Diseases of the Heart, front of the chest, major blood vessels • Diseases of the midline to upper palmar side of the upper extremities • Mental abnormalities
Small Intestine	• Diseases of the face, ear, cheek, lower jaw, neck, and throat • Diseases of the back ulnar side of the upper extremities
Spleen	• Diseases of the tongue and throat • Gastrointestinal diseases (disturbances of digestion and absorption of food) • Diseases of the inner side of the lower extremities
Stomach	• Diseases of the face, nose, oral cavity, teeth, throat, and front of the neck • Diseases of the abdomen, the frontal part of the lower extremities and gastrointestinal area • Certain psychiatric diseases
Triple Burners	• Diseases of the face, ear, cheek, larynx, and neck • Diseases of the back of the upper extremities from the midline of the torso to the upper arm and forearm

Figure 6.22. Pathologies of the Twelve Primary Channels

through each of the Twelve Primary Channels has a vital impact on all the body's organ systems. These Qi and Blood flow concepts define the reciprocal relationships existing between the body's internal organs, channels, and energetic points situated throughout the body.

Figure 6.22 describes some of the most common Twelve Primary Channel pathologies, and their clinical manifestations along the "root and branches" of these channels.

The term "root and branch," is used to describe the difference between the medial and lateral, upper and lower orientations of the channels and their points of origin. They are also used to describe the progression of a disease, as well as the priorities and sequence of the Qigong doctor's treatment.

Specifically, the word "branch" is used to describe the channels' flow of Qi and Blood and the progression and direction of a disease. The word "root" is used to described the channels' internal organ, or the origin of a disease.

CHAPTER 7
THE EIGHT EXTRAORDINARY VESSELS AND COLLATERALS

THE EIGHT EXTRAORDINARY VESSELS

Extending from the body's Taiji Pole are the Three Dantians, Five Jing Shen, and the Eight Extraordinary Vessels. These energy systems feed all the vital organs and the body's Twelve Primary Channels. The Eight Extraordinary Vessels are so called because of they differ in energetic function from the Twelve Primary Channels. These eight vessels have an extraordinary ability to regulate the deeper energetic reservoirs of the body. They are the first vessels (or channels) to form in the developing fetus, and are sometimes called the Eight Ancestral Channels, Eight Prenatal Channels, Eight Preheaven Channels, or Eight Psychic Channels.

These vessels represent the merging of the mother's and father's energy, and the linking of the body's prenatal and postnatal energies. They functionally connect to the Twelve Primary Channels and circulate the Jing Qi (Essence Energy) throughout the body.

The energetic expansion and contraction of the Eight Extraordinary Vessels affects the creation and development of the baby's tissues up until the time of birth. In energetic embryology, the three-dimensional space of the impregnated egg is viewed as an object that can be described by eight different surfaces along four axes. Each of the Eight Extraordinary Vessels corresponds with one of these directions:

- anterior/posterior—Governing and Conception Vessels
- superior/inferior—Yin and Yang Linking Vessels
- right/left—Yin and Yang Heel Vessels
- interior/exterior—Thrusting and Belt Vessels

THE FUNCTION OF THE EIGHT EXTRAORDINARY VESSELS

The Eight Extraordinary Vessels have neither a direct connection nor an internal/external relationship with the internal organs. Similar to the Three Dantians' function of distributing the body's energy, these Eight Extraordinary Vessels are also reservoirs that regulate the distribution and circulation of Jing and Qi inside the body. They are the foundation of the body's energy, bridging the Yuan Qi (Original Prenatal Energy) with the body's postnatal energy.

The Eight Extraordinary Vessels serve as reservoirs of Qi. When these reservoirs become full, the energy overflows into the center channel or Taiji Pole. This stimulation of the Taiji Pole expands consciousness and increases perceptual intuition.

The Eight Extraordinary Vessels have five main functions: They serve as reservoirs of Qi, store and circulate Jing Qi, circulate the Wei Qi, regulate the body's life cycles, and integrate the six primary Yang organs with the Six Extraordinary Organs and with the Kidneys.

1. They Serve as Reservoirs of Qi. If the Qi flow of the Twelve Primary Channels becomes Excessive, they overflow into the Eight Extraordinary Vessels, which act as Qi reservoirs, receiving the Excess Qi for storage and distribution, while at the same time regulating the energy flow. If the Qi flow of the Twelve Primary Channels becomes Deficient, they can draw from the reservoirs of energy stored within the Eight Extraordinary Vessels which again regulate the body's energy flow, bringing it back into balance.
2. They Store and Circulate Jing Qi. The Eight Extraordinary Vessels draw their energy from the Kidneys and are responsible for storing and circulating the body's Jing Qi throughout the tissues, particularly to the skin and hair and to the Six Extraordinary (Curious) Organs, also known as the Six Ancestral Or-

gans. The Six Extraordinary Organs are the Brain, Bone, Marrow, Blood, and Gall Bladder (also included is the Uterus in women).

3. They circulate the Wei Qi (Protective Energy) to protect the body against invasion by External pathogens. The Governing, Conception, and Thrusting Vessels are primarily responsible for circulating the body's Wei Qi over the thorax, abdomen, and back.

   The Eight Extraordinary Vessels provide the link between the Kidneys Jing and the Wei Qi. Although the Wei Qi is circulated by the Lungs, its root is in the Kidneys. This illustrates how the Kidneys play an important role in supporting the Lungs function of circulating the body's Wei Qi. The Kidneys function on the Wei Qi field and therefore the immune system can become vulnerability to pathogenic factors, as well as in causing allergic diseases, e.g., asthma. It also demonstrates why the Kidneys are always at the root of latent Heat syndromes (chronic fatigue), caused by Kidney Deficiency.
4. The Eight Extraordinary Vessels regulate the body's Life Cycles. In the first chapter of *the Yellow Emperor's Classic of Internal Medicine,* the life changes in women (every seven years) and men (every eight years) are governed by the Conception and Thrusting Vessels. These life cycles are related to the body's Jing and the Eight Extraordinary Vessels' role in moving and circulating the body's Jing (see Chapter 13).
5. They Integrate the Six Extraordinary (Curious) Organs with the Six Primary Yang Organs and the Kidneys. The Eight Extraordinary Vessels provide the link between the body's brain, uterus, blood vessels, Gall Bladder, Marrow and bones, and the body's internal energy flow.
   a. The Brain is Regulated by the Governing Vessel and the Yin and Yang Heel Vessels.
   b. The Uterus is regulated by the Thrusting and Conception Vessels.
   c. The Blood Vessels are regulated by the Thrusting Vessels.
   d. The Gall Bladder is regulated by the Belt Vessel.
   e. The Marrow is regulated by the Thrusting Vessels.
   f. The Bones are regulated by the Thrusting and Conception Vessels.
6. They Integrate the Four Seas. The *Magic Pivot (Ling Shu, Hai Lun)* states, "People have Four Seas... the Sea of Marrow, the Sea of Blood, the Sea of Qi, and the Sea of Grain and Water." The Eight Extraordinary Vessels provide the link between the body's Four Seas and the body's internal energy flow.
   a. The Sea of Marrow is the brain, and it is related to the Governing Yin and Yang Heel Vessels. The Sea of Marrow points are located on the Governing Vessel GV-16 and GV-20. When the Sea of Marrow is Deficient, there will be headaches and dizziness.
   b. The Sea of Qi is located in the chest of the center, and is regulated by the Conception Vessel CV-17. Some Medical Qigong schools maintain that there are two reservoirs of Qi: the Middle Dantian, being the Sea of Postnatal Qi, and the Lower Dantian, being the Sea of Prenatal Qi (which is regulated by the Qihai CV-6 point). When the Sea of Qi is in Excess, there is a feeling of fullness in the chest, dyspnea (urgent breathing), and a red complexion. When the Sea of Qi is Deficient, there is weak energy and insufficient speech.
   c. The Sea of Nourishment (also known as the Sea of Grain and Water) is the Stomach, and it is regulated by the Thrusting Vessels, which are accessed by the St-30 points. When the Sea of Water and Grain is in Excess, there is a feeling of fullness in the abdomen. When the Sea of Water and Grain is Deficient, there is a feeling of hunger with an inability to eat.
   d. The Sea of Blood (also known as the Sea of the Twelve Primary Channels) is related to the Thrusting Vessels, the Liver, and Xuehai Sp-10 points. The points can

also be accessed through Bl-11, St-37, and St-39. When the Sea of Blood is in Excess, the body feels big; When the Sea of Blood is Deficient, the body feels small.

THE EIGHT EXTRAORDINARY VESSELS AND MEDICAL QIGONG THERAPY

The energetic pathways, functions and use of the Eight Extraordinary Vessels and points in Medical Qigong are often quite different than those taught in acupuncture colleges, even though they may sometimes bear the same names.

Most acupuncturists, with the exception of some Japanese researchers, often pay little attention to the Eight Extraordinary Vessels in diagnosis and treatment. One Medical Qigong doctor from China has claimed that a study done in China showed that the Eight Extraordinary Vessel treatments were far more effective than those of the Traditional Chinese medical protocols. The results, however, were not released because officials did not want to disturb the TCM institutional operations. The reason Medical Qigong makes use of the Eight Extraordinary Vessels, is that they are much easier to access through Qi emission and Medical Qigong self-regulation prescriptions than through acupuncture therapy.

The Eight Extraordinary Vessels have been of special importance to Medical Qigong practitioners for thousands of years, and are viewed as the foundation of the body's energy, the bridge between the Yuan Qi (Prenatal Qi) and the Postnatal Qi. These Eight Extraordinary Vessels affect the body on the deepest level of our basic constitutional energy. In Medical Qigong regulation exercises, the focus on opening the flow of energy through the Eight Extraordinary Vessels is a prerequisite for opening the energy flow in the Twelve Primary Channels. The major purpose of opening the Eight Extraordinary Vessels is to provide a container for storing the Three Forces: Heaven Force (absorbing universal energy), Earth Force (absorbing environmental energy), and Man's Force: Jing, Qi and Shen (Essence, Energy and Spirit).

It is essential for the Qigong doctors to open their Eight Extraordinary Vessels to attain mastery of their energetic body. The Microcosmic Orbit is introduced for purgation, fusion, tonification, and regulation of the Governing and Conception Vessels in the first stages of training. This exercise restores a healthy flow of energy throughout all of the channels and vessels creating a healthy and balanced body.

Through the refinement of their energy, Qigong doctors receive more Qi, and the quality of their Blood changes, i.e., its energetic potential changes to contain more Qi and less Blood. At this advanced state of transformation, Qigong doctors rechannel the flow of energy through their body's Three Dantians, including the "mystical pass" (located within the Upper Dantian chambers of the brain). When the flow of energy is directed to stimulate the Mystical Pass, the Qigong doctors acquire greater awareness and control over their own bodies (physical, mental, emotional, energetic and spiritual), and are then able to enhance their perception and communication with all the other energetic planes of the world (mineral, plant, animal, and human), and with the Dao (or divine).

CLINICAL USE OF THE EIGHT EXTRAORDINARY VESSELS

The Eight Extraordinary Vessels form and establish the foundational energetic pattern for the developing fetus's tissues, as well as the adult's entire body. The clinical use of the Eight Extraordinary Vessels can be divided into several different approaches of energy manipulation to change the energetic patterns within the patient's tissues.

One approach divides the Eight Extraordinary Vessels into four pairs of Yin and Yang vessels. Each of the vessels is grouped according to the same polarity, two pairs of Yin and two pairs of Yang. When paired this way, each set of the Yin and Yang Extraordinary Vessels has a common range of energetic action in terms of the patient's body area.

1. The Conception and Yin Heel Vessels affect the energetic flow to the patient's abdomen, chest, Lungs, throat, and face.
2. The Governing and Yang Heel Vessels affect the energetic flow to the back of the patient's legs, as well as the back, spine, neck, head, eyes, and brain.

3. The Thrusting and Yin Linking Vessels affect the energetic flow to the inner aspect of the patient's legs, as well as the abdomen, Stomach, chest, and Heart.
4. The Belt and Yang Linking Vessels affect the energetic flow to the outer aspect of the patient's legs, as well as the sides of the body, shoulders, and sides of the neck.

ENERGETIC FUNCTIONS OF THE EIGHT EXTRAORDINARY VESSELS

The Eight Extraordinary Vessels can also be categorized is according to their energetic functions.

1. The Conception, Governing and Thrusting Vessels are considered the source of all the other Extraordinary Vessels These three vessels affect the patient's energy at a deep constitutional level. They originate directly from the Kidneys and are connected to the body's Jing.
2. The Yin and Yang Heel Vessels are complementary in that they both flow from the legs (controlling the state of the muscles of the legs) into the eyes (controlling the muscles that open and close the eyes).
3. The Yin and Yang Linking Vessels complement each other by linking the body's Yin and Yang Channels.
4. The Belt Vessel is the only horizontal vessel in the body, encircling the main channels. Its energetic influence is in its ability to affect the circulation of energy in the body, especially within the legs.

Each of the Eight Extraordinary Vessels is also connected to one of the Eight Trigrams (or Eight Natural Manifestations), and can be energetically regulated through the use of the Magic Square. For further treatment and clinical protocol see Chapter 31 regarding the Eight Extraordinary Vessel and the Magic Square.

THE GOVERNING AND CONCEPTION VESSELS

The Governing and Conception Vessels are the main rivers of the body's Yin and Yang energies. They are polar aspects of the body, perfectly complementary, like midnight and midday. They are responsible for the formation of the holoblastic cleavage and the first cellar division of the fertilized ovum in embryological development (see Chapter 2).

In China, Dr. Li Shi Zhen was the first to believe that the Governing and Conception Vessels are two branches of the same source, an inseparable Yin-and-Yang, front-and-back duality. These vessels connect the uterus with the Kidneys, Heart, and Brain.

They originate externally at Huiyin CV-1 point, ascend the front and back of the torso, and form a small circle when the tip of the tongue touches the upper palate in the mouth and the anal sphincter is squeezed. Not only does this action complete the balance of Fire (Heart) and Water (Kidneys) energy, it also increases the body's protective Wei Qi.

Regulating the Conception and Governing Vessels is a priority in Medical Qigong practice. Along these vessels, the Qigong practitioner draws the Yang Fire and Yin Essence up and down his or her body, fusing the Water and Fire energies together. This fusion facilitates a Yin and Yang balance throughout the body.

The Governing and Conception Vessels each have two energy flows on the anterior and posterior vertical midline of the body. Each vessel's pathway is complete, being composed of an ascending energetic flow and a descending energetic flow. The duality of these two medial lines joins at the extremities (the head and perineum), forming one complete circle of energetic current (Figure 7.1).

Both vessels are superimposed on each other, with the energy of the Governing Vessel being predominant up the back and inferior down the front (behind the Conception Vessel); the energy of the Conception Vessel is predominant up the front and inferior down the back (behind the Governing Vessel).

The energetic flow of the Conception Vessel corresponds to Yin, negative polarity, the female aspect, and responds to bass tones; while the energetic flow of the Governing Vessel corresponds to Yang, positive polarity, the male aspect, and responds to treble tones.

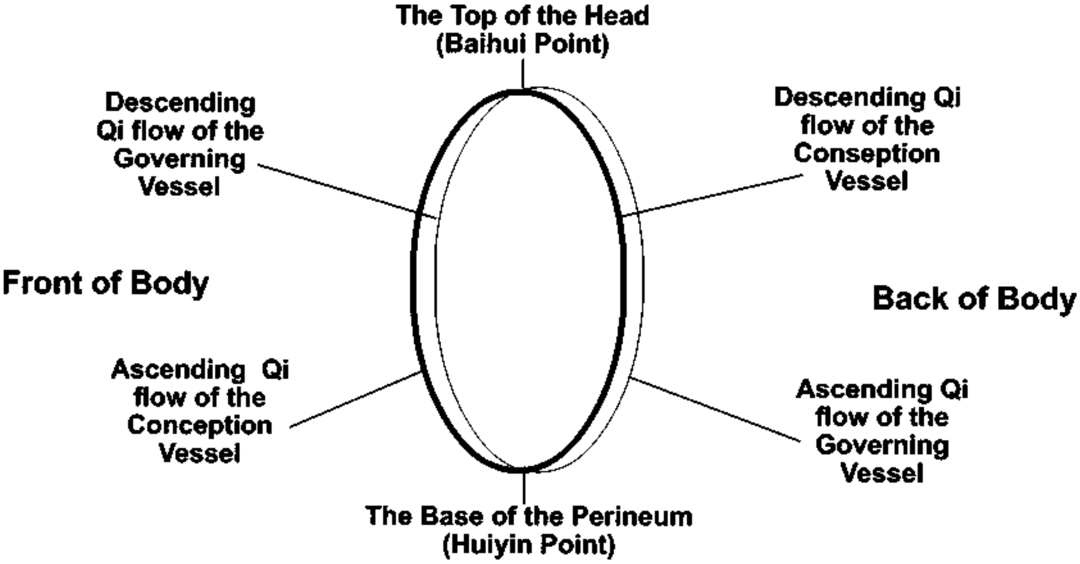
The Top of the Head (Baihui Point)
Descending Qi flow of the Governing Vessel
Descending Qi flow of the Consеption Vessel
Front of Body
Back of Body
Ascending Qi flow of the Conception Vessel
Ascending Qi flow of the Governing Vessel
The Base of the Perineum (Huiyin Point)

Figure 7.1. The Energetic Flow of the Conception and Governing Vessels

It is important to understand the energetic potential of these two currents flowing in both an ascending and descending direction on the posterior and anterior vertical midline in the body. The energetic movement of these two currents explains why there are two opposite directions of focused concentration used in energy cultivation meditations. One direction follows the Microcosmic Fire cycle along the Governing Vessel (up the spine and down the chest) to stimulate the emotional regulation of the acquired mind (Zhi Shen); the other direction follows the Microcosmic Water cycle along the Conception Vessel (up the chest and down the spine) to stimulate spiritual intuition and activate the perceptions of the Yuan Shen.

THE GOVERNING VESSEL OR DU VESSEL

The word Du translates as "governing," and refers to a general, someone who controls and is in charge. During the development of the embryo, the Governing Vessel is responsible for the development and formation of the medulla oblongata and cerebrum. The Governing Vessel controls all the Yang Channels in the body, and is called the Sea of Yang Channels. It is responsible for nourishing the brain and spinal cord and for consolidating the Yuan Qi in the Kidneys.

The Governing Vessel originates in the Lower Dantian in both men and women. It is composed of many energetic branches, emerging externally at the perineum. The primary branch ascends externally on the midline of the back, over the head and ends in the frenulum of the upper lip. As the Qi flows up the Governing Vessel one branch enters the interior of the spinal column and enters the Kidneys. Another set of branches enters the brain at the Fengfu GV-16 point and ascends internally to the vertex at the Baihui GV-20 point.

From the lower perineum, another set of branches encircles the anus and envelops the external genitals, ascends past the navel, Heart, and throat to circle the mouth. After circling the mouth, it further ascends to the eyes, emerging just below the middle of the eyes. From the inside of the eyes (at the Jingming Bl-1 point), a set of branches follows the Bladder Channels along the forehead, converging at the vertex of the Baihui GV-20 point and enters the brain. From the brain, the Vessel emerges at the Fengfu GV-16 point, and divides into two additional branches which descend the

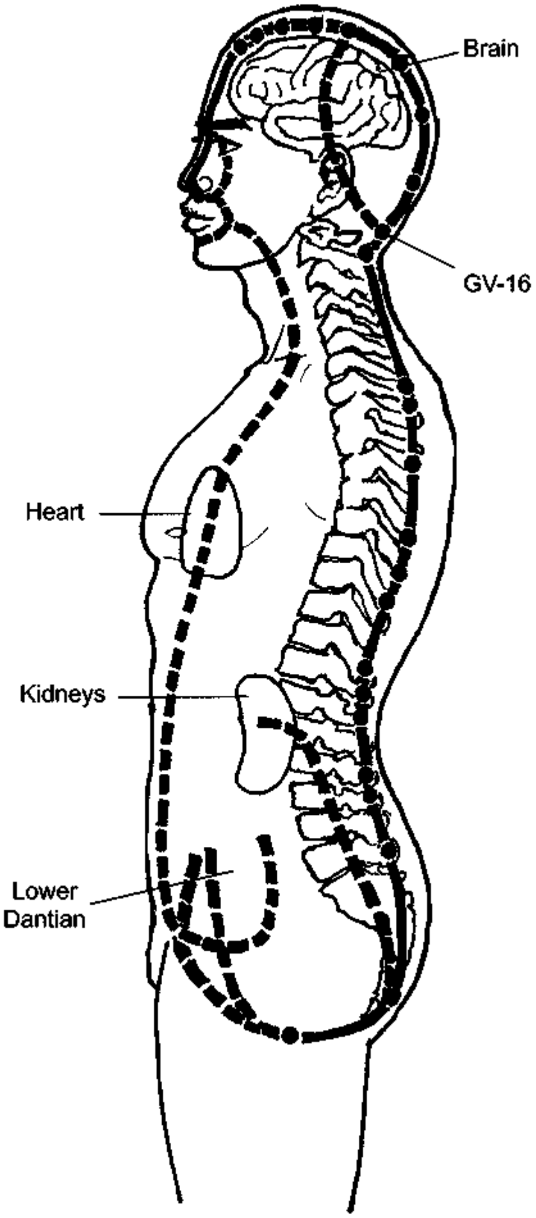

Figure 7.2. The Governing Vessel (Side View)

back, passing through the Fengmen Bl-12 points along the sides of the spine and enter into the Kidneys (Figure 7.2).

The Governing Vessel tonifies the Kidney Yang and strengthens the spine and back. When the Governing Vessel is excessive, the back becomes stiff; when it is depleted, the head becomes heavy, unstable, and shaky (Figure 7.3).

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Governing Vessel are as follows.

1. To strengthen and control all the Yang Channels of the patient's body (particularly in relation to the Kidney Yang and the brain);
2. To tonify the Kidney Yang and strengthen the patient's spine and back (especially in cases of chronic lower back pain due to Kidney Deficiency);
3. To purge External Wind from the patient's body when symptoms are present, such as runny nose, head ache, fever, and stiff neck;
4. To purge Internal Wind from the patient's body when symptoms are present, such as dizziness, tremors, convulsions, epilepsy, or the condition of Wind Stroke;
5. To strengthen and nourish the patient's brain and Marrow, and for treating such symptoms as poor memory, dizziness, and tinnitus; and
6. To treat depression, due to the Governing Vessel's connection to the body's Jing (Kidneys, lack of willpower), Qi (Heart), and Shen (Brain).

Note: In men, the Governing Vessel is generally treated by itself, and in women, its treatment is combined with treating the Conception Vessel.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include diseases of the head (apoplexy, aphasia, epilepsy, headaches, tetanus, etc.), the back, neck, and Kidneys (pain and stiffness in the spinal column), spastic muscle movement of the extremities, night sweating, and Jing Shen disorders (hyperexcitability, hallucinations, etc.).
2. Symptoms of Deficiency result in shaking of the head along with a feeling of heaviness, instability, and an inability to concentrate. The patient lacks physical and mental stamina, displays weakness in character, and may also experience hemorrhoids, sterility, and impotence.

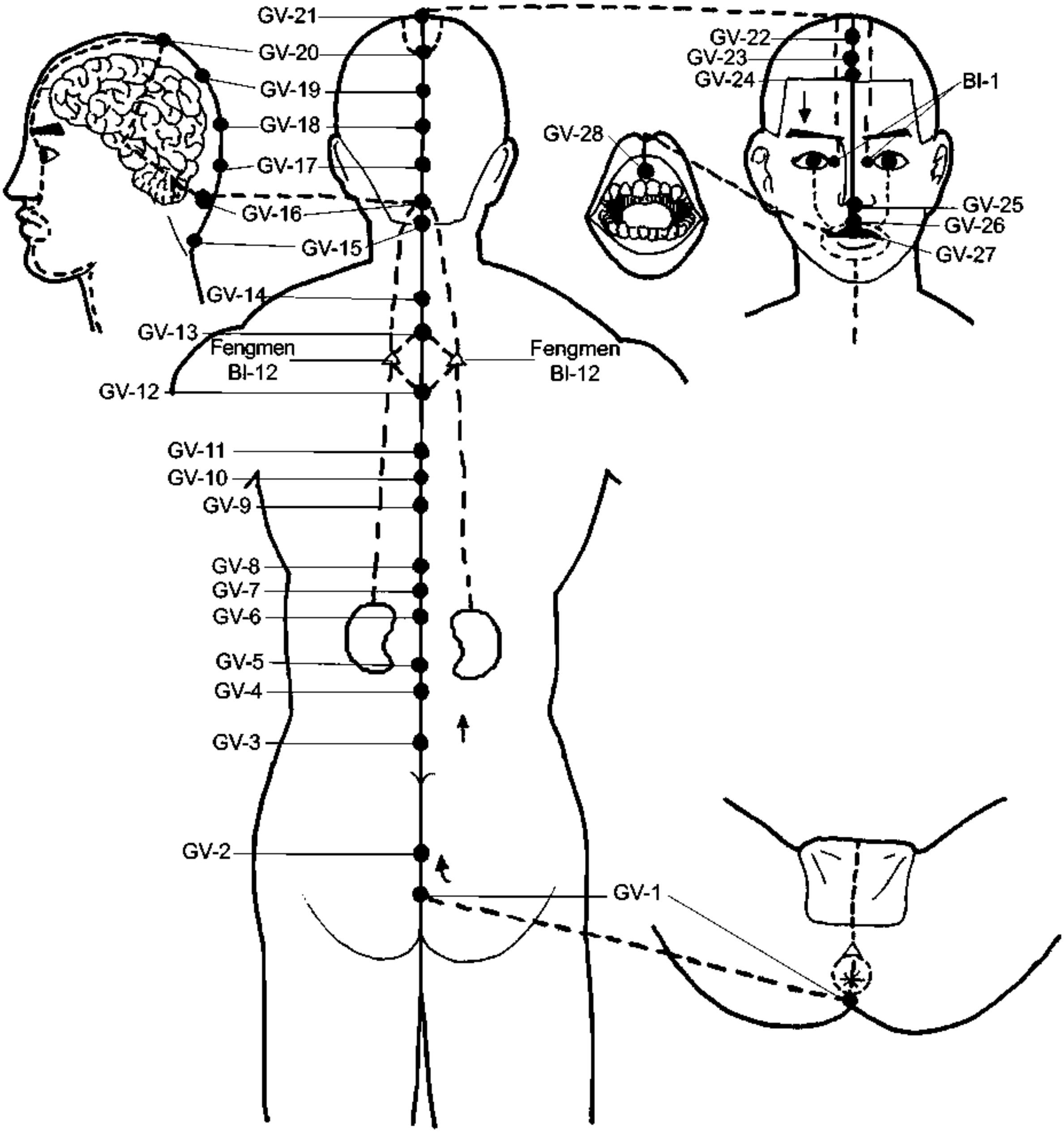
GV-21
GV-20
GV-19
GV-18
GV-17
GV-16
GV-15
GV-14
GV-13
Fengmen
Bl-12
GV-12
Fengmen
Bl-12
GV-11
GV-10
GV-9
GV-8
GV-7
GV-6
GV-5
GV-4
GV-3
GV-2
GV-1
GV-22
GV-23
GV-24
Bl-1
GV-28
GV-25
GV-26
GV-27

Figure 7.3. The Governing Vessel (GV)

THE CONCEPTION VESSEL OR REN VESSEL

The word Ren translates as Conception, and refers to pregnancy, responsibility, or obligation. It can also mean "to accept." The Conception Vessel governs all the Yin Channels of the body and is called the Sea of Yin Channels. It nourishes the Yuan Qi of all Five Yin Organs. The Conception Vessel originates from the Lower Dantian, and specifically in the uterus in females (Figure 7.4).

Functionally, the upper third of the Conception Vessel, on the sternum, controls respiratory functions; the middle third, on the epigastrium, controls digestive functions; and the lower third, on the abdomen, controls the urogenital functions (Figure 7.5).

In women the Conception Vessel is primarily responsible for nourishing the uterus and the genital system and determines the seven-year life cycles. It links the Yin energy with all aspects of conception and reproduction. The Conception Vessel, along with the Thrusting Vessels, have an important relationship with obstetric diseases related to the development of the fetus, delivery, and menstruation.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Conception Vessel are as follows.

1. To strengthen and nourish the Yin energy of the patient's body (especially in women after menopause) and harmonize the Lungs and Kidneys;
2. To regulate the energy of the reproductive system, tonifying the Blood and Yin, and to reduce the effects of Heart Empty-Heat symptoms developed from Kidney Yin Deficiency after menopause (i.e., night sweating, hot flashes, anxiety, mental irritability, insomnia and dizziness, etc.);
3. To promote Blood supply to the uterus and regulate menstrual disorders (i.e., dysmenorrhoea, menorrhagia, amenorrhoea, and metrorrhagia);
4. To move the Qi in the patient's Lower Burner and treat abdominal lumps, as well as myomas, fibroids, and carcinoma in the woman's uterus, and hernia in men; and
5. To stimulate the energetic interaction of the Lungs Channels descending Qi function, and the Kidneys' function of receiving and holding the Lung Qi, which can, when malfunctioning, result in asthma.

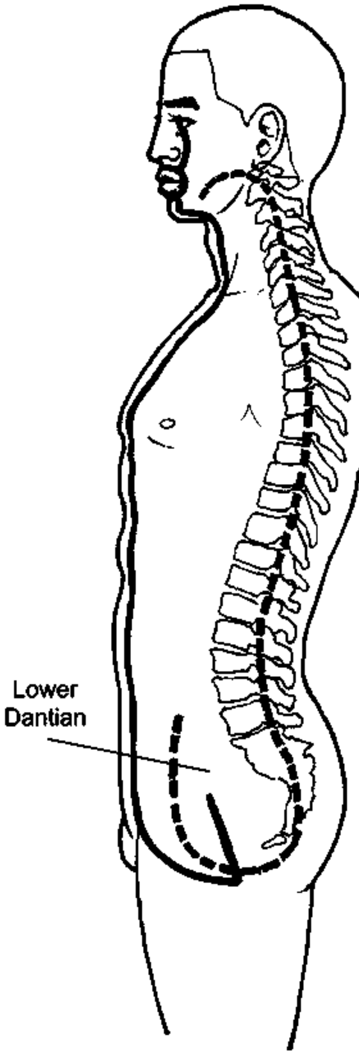

Figure 7.4. The Conception Vessel (Side View)

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include diseases of the reproductive and gastrointestinal systems (hemorrhoids, diarrhea, decreased urination, etc.). In the male, problems in the Conception

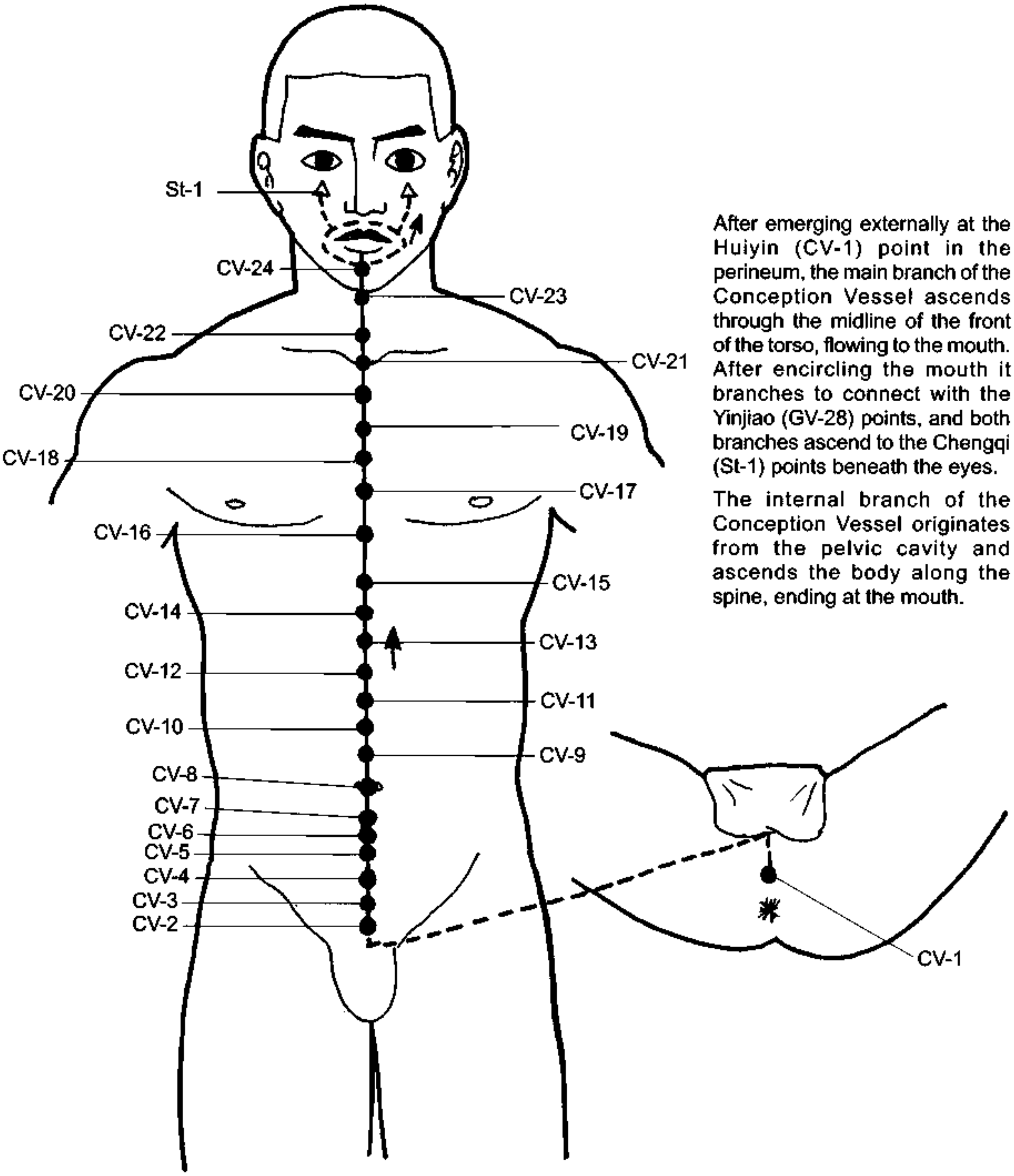
St-1
CV-24
CV-23
CV-22
CV-21
CV-20
CV-19
CV-18
CV-17
CV-16
CV-15
CV-14
CV-13
CV-12
CV-11
CV-10
CV-9
CV-8
CV-7
CV-6
CV-5
CV-4
CV-3
CV-2
CV-1

After emerging externally at the Huiyin (CV-1) point in the perineum, the main branch of the Conception Vessel ascends through the midline of the front of the torso, flowing to the mouth. After encircling the mouth it branches to connect with the Yinjiao (GV-28) points, and both branches ascend to the Chengqi (St-1) points beneath the eyes.

The internal branch of the Conception Vessel originates from the pelvic cavity and ascends the body along the spine, ending at the mouth.

Figure 7.5. The Conception Vessel (CV)

Vessel produce sterility. In the female, the Conception Vessel can cause menstruation problems such as leukorrhea and dysmenorrhea. Other female problems include breast pain, paralysis after delivery, emaciation, and sterility. Other problems of the female reproductive system include the external genitalia and vulva, vagina, and cervix disorders.

2. Symptoms of Deficiency include abdominal pains (hernia), pruritus, and a heavy feeling in the hips, lower ribs, and lumbar area.

THE THRUSTING VESSELS AND THE BELT VESSELS

The Thrusting and Belt Vessels balance the external tissues and internal organ energies, controlling an Excess or Deficient condition by regulating the upper and lower, as well as left and right, energetic quadrants.

The Belt Vessel is responsible for the second cellular division of the fertilized ovum in embryological development (see Chapter 2).

THE THRUSTING VESSELS OR CHONG VESSELS

The word Chong translates as "a street" and expresses the idea of passing or penetrating through something. This passing through action refers to its function as the vital pathway for alchemical transformation, used to produce energetic change (Jing to Qi, Qi to Shen and vise versa). The Thrusting Vessels, also called the Penetrating Channel, and are regarded as the Sea of Blood and the Sea of the Twelve Regulating Channels. They regulate both the Qi and Blood of all Twelve Primary Channels and extend to the anterior, posterior, upper, and lower parts of the body. During the development of the embryo, the Thrusting Vessels are responsible for the development of the adrenal glands and the cortex.

The Thrusting Vessels originate in the Lower Dantian. The Qi of the Thrusting Vessels travel to the head and face to flow into the Chong and penetrate the lower limbs, irrigating the body's Yin (Figure 7.6). According to some Chinese doctors, all energetic points that have Chong in their names relate to the Thrusting Vessels (i.e., Qichong St-30, Taichong St-42, and Shaochong Ht-9).

Along with the Conception Vessel, the Thrusting Vessels are responsible for regulating the changes in life cycles that occur every eight years in men, and every seven years in women. Abnormalities of the Thrusting Vessels during the beginning of pregnancy result in the mother's body aborting the fetus. Abnormalities of the Thrusting Vessels during the end of pregnancy result in an inability to expel the placenta.

The Thrusting Vessels control the woman's menstruation in all aspects, influencing the supply and amount of Blood in the uterus, and nourishing the woman's Jing. They also flow along the Kidney Channels and are related to the muscles of the abdomen and the penis (and are responsible for tightening the abdominal muscles).

THE ENERGETIC PATHWAY OF THE THRUSTING VESSELS

The classical description of the Thrusting Vessels' energy flow begins inside the uterus in women (the corresponding lower abdominal area in men), continuing down to the perineum. From the perineum, one branch ascends inside the spinal column, while four other branches flow externally.

Two external energetic rivers ascend up the front of the body alongside the Kidney Channels and enter into the chest. They then ascend through the throat, encircling the lips, and end at the corner of the eyes. When the Thrusting Vessels are full and overflowing (through meditation), they radiate a sparkling white-light energy, which can be observed to extend from the upper chest to the eyes.

From the perineum, two other external energetic rivers descend down the inner aspect of the thighs (one on each leg), along the Spleen Channels. They divide again at the medial malleolus. One set of branches flows into the big toes, a second set of branches terminates at the bottom of each foot.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Thrusting Vessels are as follows.

1. To strengthen, nourish, and regulate weak

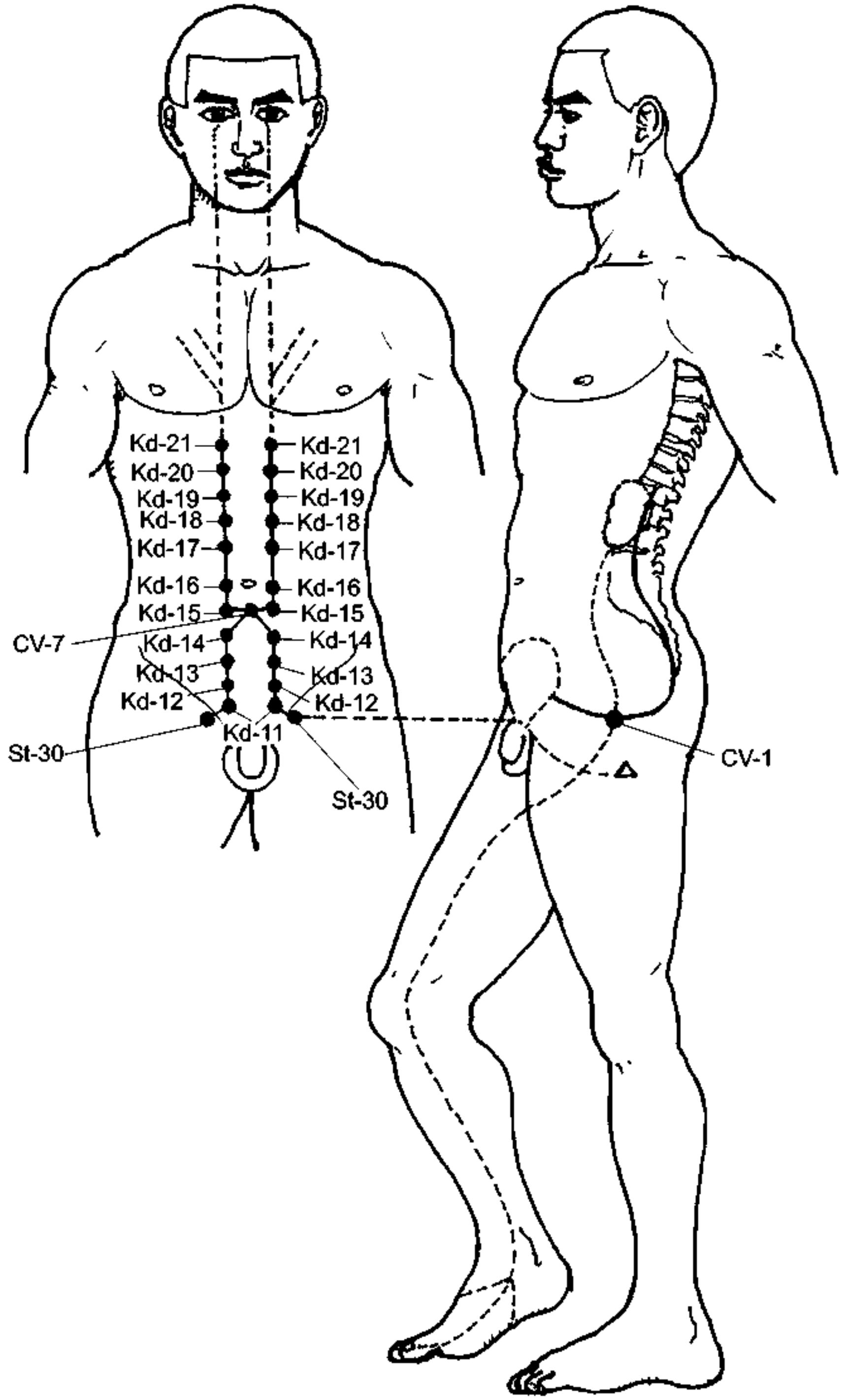
Kd-21
Kd-20
Kd-19
Kd-18
Kd-17
Kd-16
Kd-15
CV-7
Kd-14
Kd-13
Kd-12
St-30
Kd-11
Kd-21
Kd-20
Kd-19
Kd-18
Kd-17
Kd-16
Kd-15
Kd-14
Kd-13
Kd-12
St-30
CV-1

Figure 7.6. The Thrusting Vessels

constitutions with digestive symptoms (e.g., abdominal distension, poor appetite, and poor assimilation of food);
2. To move the Blood of the Heart to relieve symptoms of pain and stiffness of the chest (the Thrusting Vessels control all the Blood in the connecting channels);
3. To purge Qi and Blood stagnation in the abdomen and chest;
4. To treat feelings of anxiety (within the chest) caused by Rebellious Qi; One of the Thrusting Vessels' most important pathologies is Rebellious Qi (Qi that moves in the wrong direction, going upwards instead of downwards). Feelings of anxiety that arise in the patient's abdomen and ascend to their chest are especially indicative of this malfunction;
5. To treat gynecological problems (hot flashes), when the Qi of the Thrusting Vessels rises upwards, and causes the patient's hands and feet to get cold, the face to get hot, and a feeling of fullness in the chest. The treatment is to regulate the Thrusting Vessels and subdue the Rebellious Qi in the Chong Vessel. If there are accompanying emotional problems and Liver Qi stagnation, then also treat the patient's Lv-3 points; and
6. To treat Rebellious Qi caused by the energy of the Thrusting Vessels rebelling upwards; the symptoms are oppressive feelings in the chest, as well as dizziness, nausea, and vomiting.

PATHOLOGICAL MANIFESTATIONS

The main diseases associated with the patient's External Thrusting Vessels include diseases of the Heart, fullness in the chest and abdomen, gastritis, abdominal pain, and convulsive diseases. If a woman's Thrusting Vessels are Deficient or Empty she may develop such conditions as amenorrhoea, scanty periods, or late periods. If there is stagnant Qi and / or Blood in the Thrusting Vessels, she may experience dysmenorrhoea. Abnormalities of the Thrusting Vessels also result in the mother aborting the fetus.

The Thrusting Vessels work with the Qi of the Kidney Channels to control the brain's physiological functions. If the energy from the Thrusting Vessels to the cortex becomes Deficient, this can inhibit the development of the brain.

THE BELT VESSEL OR DAI VESSEL

The word Dai translates as a "belt" or "girdle," and refers to the action of supporting something. The Belt Vessel is the only horizontal vessel in the body. Beginning at the Mingmen, this vessel encircles the waist like a belt, dipping down into the lower abdominal region anteriorly and running across the lumbar region posteriorly. It connects with the Gall Bladder's -26, -27, and -28 points and crosses the Conception Vessel at CV-4. Many sources say that it connects with the Liver -13 (at the free ends of each 12th rib). It binds, joins, and controls all the channels of the body, exerting an influence on the energetic circulation of the body's Governing and Conception Vessels (Figure 7.7).

Medical Qigong schools teach that the Belt Vessel circle wraps the entire body, like a enveloping cocoon, flowing from the feet to the head. The waist is considered the "hub" of the energetic wheel, and the access points of the Belt Vessels center channel (which circles the waist) is used to control the entire vessel. As the energy of the body increases (through Qi cultivation), the Qi within the entire Belt Vessel increases, circle wrapping the tissues from feet to head, increasing the body's Wei Qi fields (Figure 7.8).

In addition, the Belt Vessel keeps both the Yin and Yang Channels under control, connecting all of the leg Yin and Yang channels as they traverse the body's trunk. As a result, the Belt Vessel assists in regulating the circulation of Qi in those channels, especially from the waist down.

The Belt Vessel has an important influence on the body's physiology by encircling the leg channels, influencing the circulation of energy to and from the legs, as well as influencing the energetic actions of the genitals, waist, and hips. This vessel not only restrains the flow of the body's Liver and Gall Bladder Qi, but also harmonizes the ascending and descending flows of energy from the Kidneys and Spleen through its connection with the Kidney Divergent Channel.

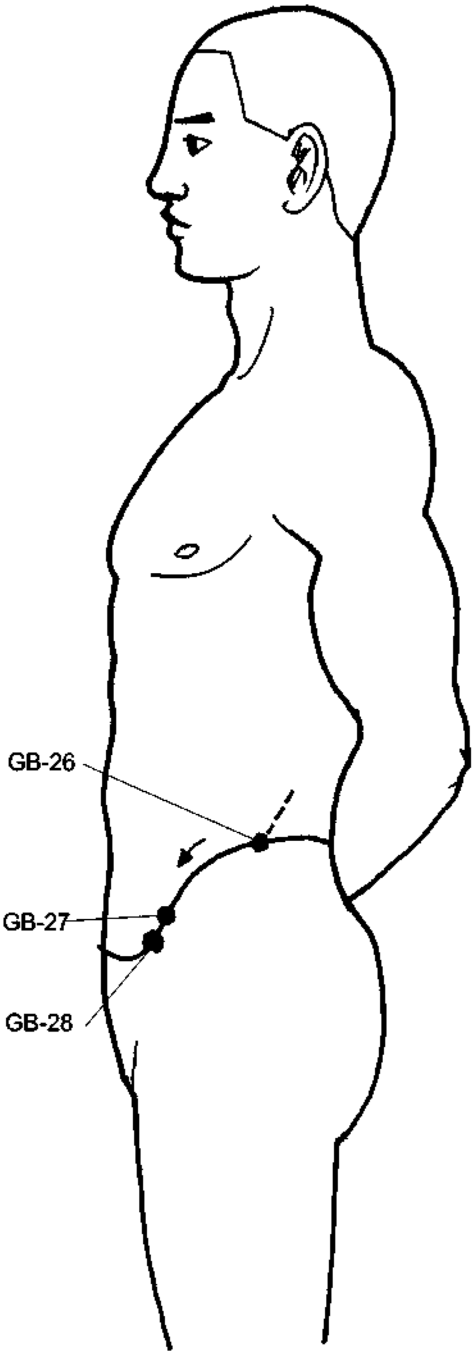

Figure 7.7. The Center of the Belt Vessel

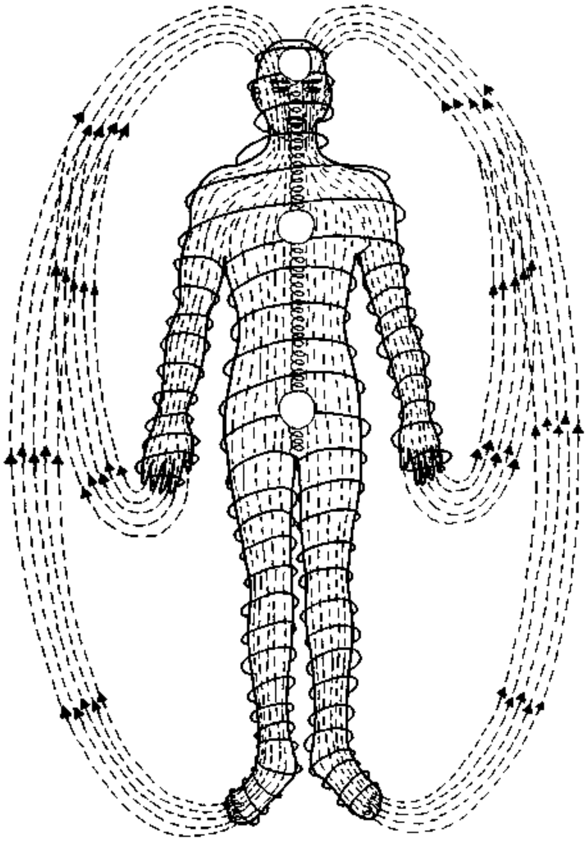

Figure 7.8. The Belt Vessel winds up the body like an energetic coil.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Belt Vessel are as follows.

1. To treat impaired circulation of Qi in the leg channels resulting in such symptoms as cold legs and feet, or tense leg muscles (gastrocnemius and tibiales). This condition is due to Liver Blood not moistening the sinews of the legs;
2. To treat impaired circulation of Qi and Blood that cause numbness, weakness, atrophy, or motor impairment of the leg muscles, due to a Deficiency of energy in the Stomach and Spleen Channels;
3. To tonify and harmonize the circulation of

Liver and Gall Bladder Qi due to Excess patterns of Liver energy;
4. To disperse Damp Heat in the genitals that has resulted in symptoms such as difficulty or burning during urination;
5. To treat hip pain caused by Deficient Liver Blood (leading to sinew and joint malnourishment and Excess of Liver Yang; and
6. To regulate the lower abdomen due to a Belt Vessel imbalance, causing symptoms such as a sagging waist or bloated abdomen.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include pain in the back (lumbar region), and sides of the lower abdomen, as well as weakness in the shoulders, upper extremities, and lower limbs. Symptoms may also include weakness in the opposite sides of the body (e.g., eye, breast, ovary, etc.), and a feeling of heaviness in the body and abdomen (as if carrying 5,000 coins) due to exposure to Dampness.
2. Symptoms of Deficiency have physical sensations similar to that of "sitting in cold water" up to the waist. This description generally refers to pain, weakness, and a cold, heavy sensation in the patient's lumbar and sacral regions. Other symptoms include umbilical, abdominal and lumbar pain, as well as a feeling that something like a stick is pressing against the groin. There can also be abdominal fullness and distention. In women, there may be an abnormal white vaginal discharge, and a prolapse of the uterus.

THE YIN AND YANG HEEL VESSELS OR QIAO VESSELS

The word Qiao translates as "the heel" or "to stand on the toes," and refers to the action of kicking one's foot as high as possible. The Yin and Yang Heel Vessels flow along the medial and lateral aspects of the lower legs and torso, connecting at the inner canthus (by the eyes). They link the channel energy of the body's Yin and Yang Channels and regulate the movement of all four limbs. They also control the amount of energy being used by all the other channels in the body. Once these vessels are full, they relax the tissues, enabling the limbs to become more dexterous.

The Yin and Yang Heel Vessels are sometimes called Bridge Channels, because they act like a bridge linking the stored Qi in the body with the areas in need of Qi. When any channel uses more than its share of energy, other channels become Deficient. Thus the Heel Vessels seek to ensure that energy is always distributed in a balanced way.

The Yin Heel Vessels are an offshoot of the Kidney Channels at the front of the body, while the Yang Heel Vessels are an offshoot of the Bladder Channels at the back of the body. Together, the Yin and Yang Heel Vessels can be used to treat structural imbalances and to harmonize the right and left sides of the body.

Because the Heel Vessels cause the motor nerves to develop during the formative stages of the embryo, the Chinese believe the Yang Heel Vessels cause little boys to be more physically active (running, jumping, etc.), while the Yin Heel Vessels cause little girls to be less actively inclined.

THE YIN HEEL VESSELS

The Yin Heel Vessels control the Yin of the left and right sides of the body. The Yin Heel Vessels influences the male and female reproductive system, as well as the lower abdomen in women.

When Yang energy is slowed down in the Heel Vessels, the Yin energy moves more rapidly. The Excess Yin causes the following problems: sleepiness, the inability to keep the eyes open while trying to stay awake, hypotension, choking, painful urination, stomach rumbling, vomiting, diarrhea, difficult bowel movements, and unconsciousness. Excess Yin in women also causes a difficult labor. Also, when the Yin Heel Vessels are in Excess, the inner leg muscles are tight, while the outer leg muscles are loose. Weakness in the Yin Heel Vessels can cause the feet and ankles to invert, whereas an Excess condition can cause the feet to turn outwards.

The Yin Heel Vessels begin at the superficial part of the ankles, just below the medial malleolus of the tibial bones at the Kd-2 points. They continue up the front of the body, ending at the inner

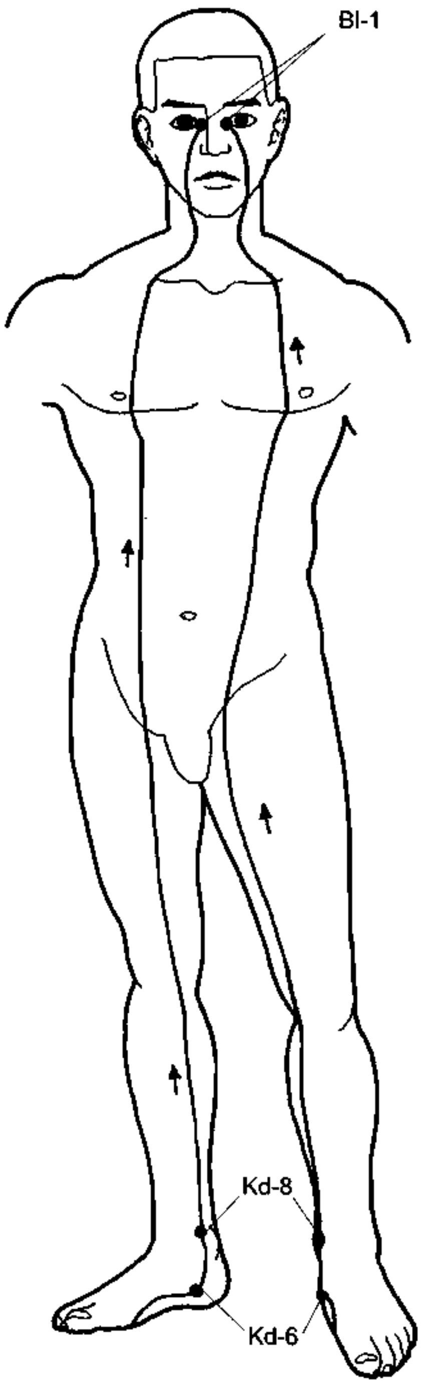

Figure 7.9. The Yin Heel Vessels

corners of the eyes, where they join the Yang Heel Vessels (Figure 7.9).

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Yin Heel Vessels are:

1. To treat symptoms of Excess conditions in the Lower Burner in women (e.g., abdominal distension, difficult delivery, or the retention of the placenta, abdominal masses, lumps, and fibroids);
2. To structurally balance the left and right sides of the body;
3. To treat disturbances of sleep (e.g., insomnia or somnolence); and
4. To treat certain cases of atrophy (e.g., when the muscles of the inner aspect of the leg are loose and the outer leg muscles are tight).

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess of the Yin Heel Vessels include: lower abdominal pain, vomiting, difficult bowel movements, and spasms on the medial side of the legs. The Yin Heel Vessels also control diseases of the eyes (watery eyes, heavy sensations of the eyelids or an inability to open the eyes), migraines, congestive headaches, and hypersomnia. Abnormalities of the Yin Heel Vessels cause pregnant women to have difficult labor.
2. Symptoms of Deficiency include: Aggravations of symptoms during the evening time, nocturnal headaches, insomnia, cramps, or convulsions.

THE YANG HEEL VESSELS

The Yang Heel Vessels control the Yang of the left and right sides of the body. Abnormalities of the Yang Heel Vessels in newborns cause vomiting of milk.

When Yin energy is slowed down in the Heel Vessels, the Yang energy moves more rapidly. The Excess Yang causes the following problems: insomnia, difficulty in closing the eyes, painful eyes, hypertension, stiff back and waist (inability to bend down), thigh tumors, bad colds, spontaneous sweating, headaches, paralysis of the arms and legs, deafness, epilepsy, nose bleeding, swelling of the body, pain in the joints, and head sweating. Excess Yang can also cause the vomiting of milk in infants.

Weakness in the Yang Heel Vessels can cause foot and ankle eversion, while Excess can cause foot inversion. Also, when the Yang Heel Vessel is in Excess, the inner leg muscles are loose and the outer leg muscles are tight.

The Yang Heel Vessels begin on the outer side of the ankles below the lateral malleolus, at the Bl-62 points. From there, they continue to ascend the outside of the thigh and the back to end just below the eyes at the Bl-1 points, where they communicate with the Yin Heel Vessels near the eyes (Figure 7.10).

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Yang Heel Vessel are:

1. To treat acute Excess conditions of the lower back (e.g., aches due to spasm or invasion of Cold) and pain along the Bladder Channels of the legs;
2. To subdue Internal or External Wind from the head (e.g., facial paralysis, severe dizziness, and aphasia); and
3. To purge Wind-Heat and Wind-Cold manifesting in symptoms, such as sneezing, head ache, runny nose, and stiff neck.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include: Stiffness of the back and waist, lumbar pain, spasms on the outer side of the legs, tumors of the thighs, nocturnal epileptic seizures, and insomnia or restless sleep. The Yang Heel Vessels are also involved with diseases of the eyes such as dry and itchy eyes.
2. Symptoms of Deficiency include: Fatigue, lassitude and weakness during the day, and aggravations of symptoms from stress during the day. These symptoms improve as the night progresses.

According to some Traditional Chinese Medical Classics, if epilepsy occurs during the daytime, Qigong and moxa are given on the Yang Heel Vessels; however, if epilepsy occurs at night, Qigong and moxa are given on the Yin Heel Vessels.

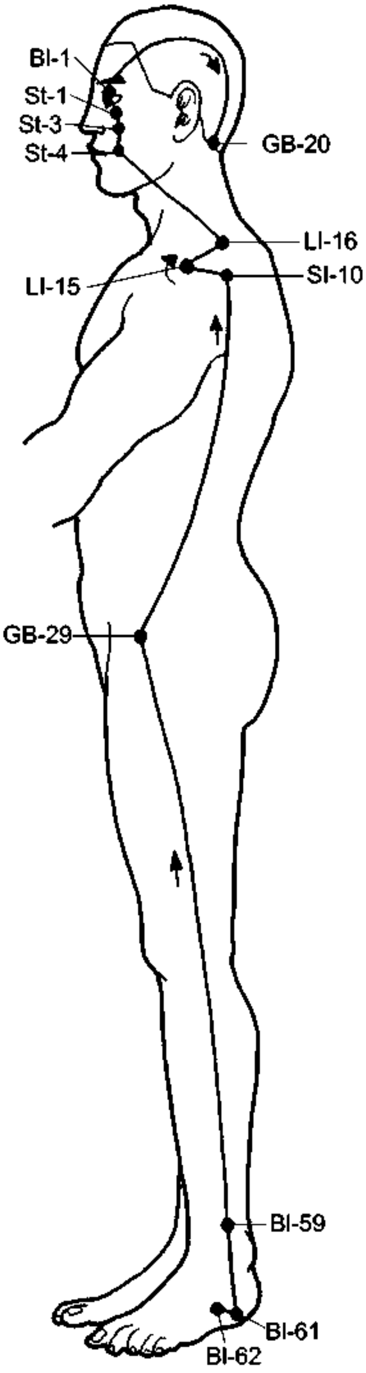

Figure 7.10. Side View of the Left Yang Heel Vessel

THE YIN AND YANG LINKING VESSELS OR WEI VESSELS

The word Wei translates as "to link" or "bind" and refers to a rope that is tied around something, pulling it down and securing it. The Linking Vessels are sometimes called the Regulator Channels. The Linking Vessels are divided into Yin and Yang energetic pathways.

Both Yin and Yang Linking Vessels start at the lower legs and flow upwards to the head, along the medial and lateral aspects of the lower legs and torso. When the Linking Vessels reach the neck and back of the head, they join the Conception and Governing Vessels. Secondary Vessels, called the Yu (surplus) Vessels, branch away from the energetic flow of the major Vessel, connecting the energetic flow of each Yin and Yang Linking Vessel to the hands.

Instead of serving as streams transporting Qi and Blood, these two vessels act as lakes that store Qi and Blood that overflows from other vessels. Together, these four vessels regulate the circulation of Qi and Blood for the whole body, store the overflowing Qi and Blood, and release the Qi and Blood into the channels in the event of insufficiencies.

THE YIN LINKING VESSEL

The Yin Linking Vessels lie on the medial axis of the body. They help maintain the connection to all the Yin Channels. They begin at the Kd-9 points at the inner side of the lower legs and ascend up the sides of the abdomen and chest. They move toward the front of the body, along the sides of the chest to the tips of the nipples. At the Ruzhong (St-17) Breast Center points they split into two sets of branches. One set of branches connects with the Conception Vessel at the neck. The other set of branches becomes the Yin Yu Vessels, and extends from the chest, over the shoulders and down the inside of the arms following the route of the Pericardium Channels. Just above both wrist folds on each arm, each branch energetically pools at the Neiguan (Pc-6) Inner Pass points, before it flows into the palms. The Pc-6 points are therefore considered the Master Points for the Yin Linking Vessels (Figure 7.11).

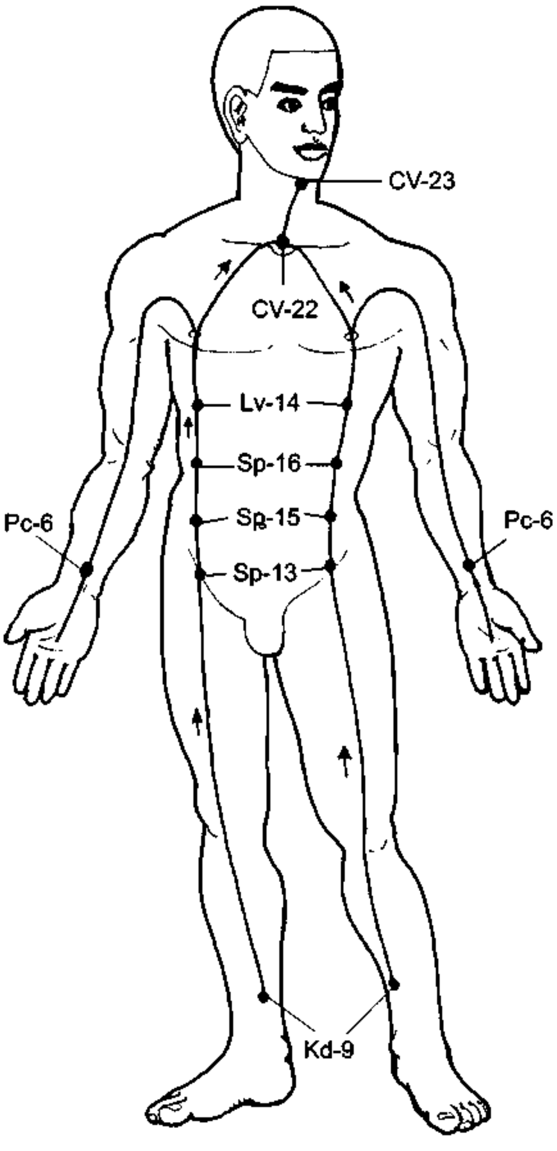

Figure 7.11. The Yin Linking Vessels

The Yin Linking Vessels are responsible for moving the Yin energy and regulating the Blood and inner parts of the body. They connect with all the body's Primary Yin Channels: Liver, Heart, Spleen, Lung, Kidney, and Pericardium. When the Yin energy is slowed down in the Linking Vessels, the Yang energy moves more rapidly. Therefore, if the Yin Linking Vessels become unbalanced, the Excess Yang condition can cause the patient to suffer from Heart pain.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Yin Linking Vessels are:

1. To tonify the Heart (for symptoms of pain, stiffness, tightness and oppression in the chest, as well as mental depression, anxiety, apprehension, and nightmares);
2. To treat Deficient Yin and Blood conditions, especially if accompanied by psychological conditions, such as mental restlessness, anxiety, and insomnia; and
3. To treat headaches located in the back of the neck due to Blood Deficiency.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include diseases of the Heart such as hypertension, delirium, nightmares, etc., as well as, cardialgia (tightness and oppression in the chest), difficulty in swallowing, convulsive diseases, contracted feeling in the Lungs, and dyspnea. Diseases of the Yin Linking Vessels also include prolapse of the rectum and diarrhea.
2. Symptoms of Deficiency include timidity or fear, apprehension, nervous laughter, mental depression, hypotension, and weak respiration.

THE YANG LINKING VESSELS

The Yang Linking Vessels lie on the lateral aspects of the body. They serve to maintain and communicate with all the Yang Channels in the exterior portion of the lateral aspects of the body. They begin at the external part of the ankles, just below the lateral malleolus at Bl-63 points, ascend the sides of each leg, up each side of the back of the body, through the shoulder areas to the Naoshu (SI-10) Scapula's Hollow points. The SI-10 points

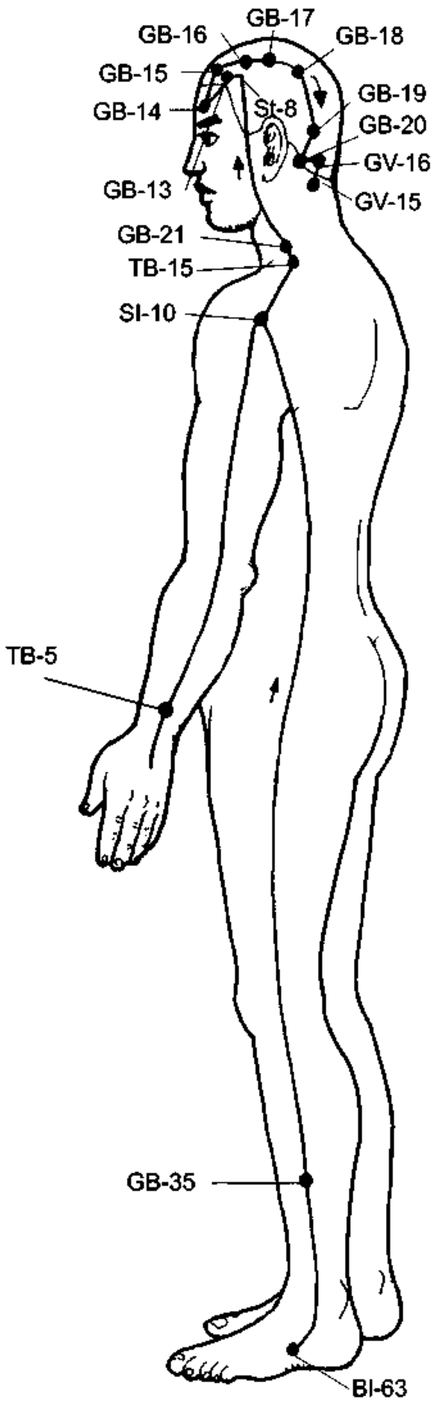

Figure 7.12. Side View of the Left Yang Linking Vessel

are the intersecting points of the Yang Linking and Yang Heel Vessels located on the Small Intestine Channels. From the SI-10 points, both Yang Linking Vessels split into two sets of branches. One set of branches flows up the sides of the neck and head connecting with the GB-20 points (at the lower occipital part of the head), where it communicates with the Governing Vessel. The other set of branches becomes the Yang Yu Vessels. The Yang Yu Vessels extend from the shoulders, down the back sides of the arms, following the route of the Triple Burner Channels. Just above both wrists on the outside of each arm, each of the branches pools at the Weiguan (TB-5) Outer Pass points, before flowing into the back of the hands. The TB-5 points are therefore considered the Master Points for the Yang Linking Vessels (Figure 7.12).

The Yang Linking Vessels are responsible for moving the Yang energy and controlling the Protective Qi, regulating the Wei Qi's resistance to external infections, and regulating the external parts of the body. They connect with all of the body's Primary Yang Channels: Gall Bladder, Small Intestine, Stomach, Large Intestine, Bladder, and Triple Burners. When the Yang energy is slowed down in the Linking Vessels, the Yin energy moves more rapidly; therefore, if the Yang Linking Vessels become unbalanced, the Excess Yin condition may cause the patient to catch colds and fevers more easily.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Yang Linking Vessels are:

1. To treat ear problems due to the rising of Liver Fire, resulting in conditions such as tinnitus and deafness;
2. To treat ear diseases due to a Gall Bladder disharmony;
3. To treat hypochondriac pain;
4. To treat sciatic pain in the lateral aspects of the legs (along the Gall Bladder Channels);
5. To treat intermittent fevers that alternate between chills and fever.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include: Alternating chills and fever; pain on the lateral sides of the neck, trunk and legs, as well as pain in the head. Symptoms also include pains and skin problems during weather changes, sensitivity to changes in climate (aching muscles, skin rashes, etc.), swelling, pain and fever in the joints, diarrhea and night sweats.
2. Symptoms of Deficiency include: Coldness and lack of body heat, loss of energy, and physical strength. Symptoms also include cold knees, stiffness, and fatigue (especially during cold or rainy weather).

SUMMARY OF EIGHT EXTRAORDINARY VESSEL PATHOLOGY

The Eight Extraordinary Vessels link all of the Yin and Yang Channels in the body and regulate the flow of energy in these channels to maintain a state of energetic balance. The clinical significance of the Eight Extraordinary Vessels is manifested through certain pathological indications that are particular to the vessels' intersection with their specific Primary Channel(s). Their symptomatology is therefore not distinct from, but rather a composite of, the pathological symptoms associated with their joining Primary Channel(s).

Figure 7.13 describes some of the Eight Extraordinary Vessels pathologies and the clinical manifestations indicated along their root (beginning) and ending branches.

The Eight Extraordinary Vessels' energetic pathways used in certain esoteric Qigong meditations are somewhat different from the ones seen in Traditional Chinese Medical texts and acupuncture charts. The reason for this uniqueness is because their purposes are different.

The goal of Chinese acupuncture is to restore sick people to health. The energetic points being treated by an acupuncturist must be along the superficial channels, so that they can be activated by acupuncture needles.

Medical Qigong exercises and meditations aim to maximizing health, and to take the individual beyond mere physical health to spiritual enlightenment. In Medical Qigong training the channels and points can be deep within the body, since the energy is guided by the mind or by postures and movements, rather than by needles.

Eight Extraordinary Vessels	**Pathologies**
Belt	• Pain in the back, lumbar region, and sides of the navel • Weakness in the lower limbs
Conception	• Diseases of the reproductive and gastrointestinal systems (hemorrhoids, diarrhea, decreased urination, etc.) • In the male, sterility • In the female, the menstruation problems such as leukorrhea and dysmenorrhea; also breast pain, paralysis after delivery, emaciation, and sterility; all reproductive system problems, including internal and external genitalia (vulva, vagina, and cervix)
Governing	• Diseases of the head (apoplexy, aphasia, epilepsy, headaches, tetanus, etc.), back, neck, and Kidneys • Stiffness in the spinal column, spastic muscle movements of the extremities, night sweating, and circulatory disturbances around the anus
Thrusting	• In women, amenorrhoea, scanty periods or late periods, dysmenorrhoea, spontaneous abortion, inability to expel the placenta, menopause problems • Diseases of the Heart, fullness in the chest and abdomen, gastritis, abdominal pain, convulsive diseases • Brain dysfunction of physiological origin
Yang Heel	• In newborns, vomiting of milk • Stiffness of the back and waist, lumbar pain, spasms on the outer side of the legs, and tumors of the thighs • Diseases of the eyes
Yang Linking	• Diseases from Cold— fevers resulting in a sensitivity to changes in climate; cold knees; stiffness and fatigue; swelling, pain, and fever in the joints and extremities; and night sweating
Yin Heel	• Lower abdominal pain, vomiting, difficult bowel movements, and spasms on the medial side of the legs • Diseases of the eyes • Difficult labor in women
Yin Linking	• Diseases of the Heart (cardialgia), tightness of the chest, difficulty in swallowing, convulsive diseases • Diseases also include prolapse of the rectum and diarrhea

Figure 7.13. Pathologies of the Eight Extraordinary Vessels

THE FIFTEEN MAJOR COLLATERALS (CONNECTING VESSELS)

The Fifteen Collaterals are also known as the Fifteen Luo. Luo is translated to mean a net or web, and in Traditional Chinese Medicine it refers to the Connecting Vessels. These vessels are the major "passage ways" for the circulation of the body's channel energy, emerging out of the Luo (pathway) points on the Twelve Primary Channels, plus the Governing and Conception Vessels. The Luo points are located below the elbows and knees and provide an additional energetic barrier to keep Evil Winds from affecting the Primary Channels, being somewhat deeper then the Muscle/Tendon Channels. The Collaterals are the streams of energy that connect the paired Primary Channel rivers.

The Fifteen Collaterals are superficial and much thinner in formation than the Primary Channels running through the interior aspect of the body. The collaterals distribute Qi and Blood to those areas not directly traversed by other channels. Flowing out of each Primary Channel's Luo point are two Luo vessels, which are counted as one:

1. The Transverse Luo Vessels connect to the Source Points on the Yin and Yang coupled Primary Channels. They act as a safety valve to maintain balance between Yin and Yang channels by diverting Excess energy from one channel to tonify Deficiencies in the Orb (internal organ and energetic field) of their paired Primary Channel. When a channel is Deficient (Empty) and its paired channel is in Excess (Full), for example, the tonification of the Luo point on the Deficient channel is enough to replenish it, while normalizing the channel previously in Excess.
2. Longitudinal Luo Vessels flow out of the Luo points, but do not connect with the coupled Primary Channels. The Luo Vessels usually flow proximally toward the channel's organ.

The Luo vessels' main function is to transfer Qi and Blood from the Primary Channels to all parts of the body to nourish the tendons, the bones, the skin, and the five sense organs (nose, eyes, ears, lips, and tongue). The Luo vessels also link the body's interior with its exterior, connecting the internal and superficial channels (Figures 7.14 through 7.19).

The Fifteen Major Collaterals have smaller branches, known as the Minute and Superficial Collaterals, extending from them. The Superficial Collaterals are sub-branches similar to capillaries serving to transport Qi and Blood to the surface of the body. Both Minute and Superficial Collaterals, which can be seen beneath the surface as blood vessels, are called Blood Luo Vessels.

PATHOLOGY

The pathology of the Fifteen Major Collaterals are categorized in terms of Excess and Deficient syndromes.

1. Excessive Conditions of the collaterals are due to exogenous invasion if (1) the organ associated with the Luo is in Excess or (2) the body's Wei Qi is weak. Evil Winds enter the body through Jing Well or Wind points, and start moving up the channel. The Longitudinal Luo Vessels provide a route for the diversion of Evil Winds. The Longitudinal Luo Vessels have more Wei Qi then the Primary Channels and can better fight pathogens. Sometimes with a Wind Cold invasion, a blue color is visible along the path of the Longitudinal Luo Vessel. If the invasion is due to Wind Heat, there may be a red color along the vessel.
2. Deficient Conditions of the collaterals are due to: (1) the organ associated with the Luo being Deficient or (2) the patient's Qi is Deficient due to exogenous factors.

POINT LOCATION OF THE FIFTEEN MAJOR COLLATERALS

1. The Foot Taiyin-Spleen 4 Point is located on the Spleen Channel of each foot, on the medial side, just posterior to the base of the first metatarsal bone. This Lou's energy flows downward to connect with the Stomach Channel on each foot. A second branch flows upward along the medial aspect of the inner thigh, passing the abdomen and connecting with the Stomach and Small Intestines (Figure 7.14).

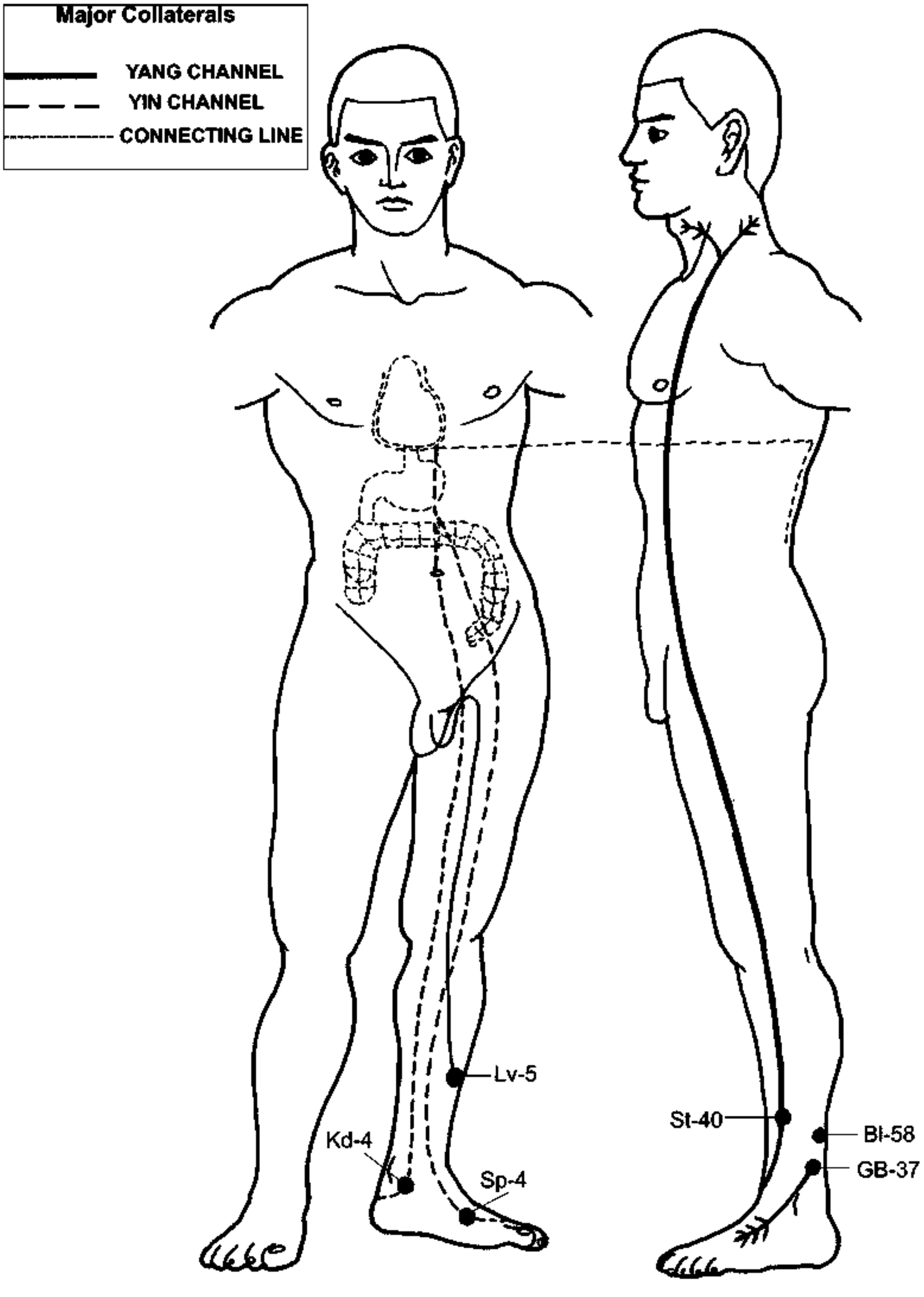
Major Collaterals
YANG CHANNEL
YIN CHANNEL
CONNECTING LINE
Lv-5
Kd-4
Sp-4
St-40
Bl-58
GB-37

Figure 7.14. The figure at left shows the three Luo points of the Yin Collaterals of the left foot located on the Primary Channels of the Kidneys, Spleen, and Liver. The figure on the right shows the three Luo points of the Yang Collaterals of the left foot located on the Primary Channels of the Stomach, Bladder, and Gall Bladder.

   a. Symptoms of Excess include sharp intestinal pain, vomiting, and diarrhea.
   b. Symptoms of Deficiency include abdominal swelling.
2. The Foot Shaoyin-Kidney 4 Point is located on the Kidney Channel of each foot, just posterior to the medial malleolus. The Lou's energetic flow runs downward, crossing the heel, and connects with the Bladder Channel on each foot. A second branch flows upward along the medial aspect of the inner thigh, following the Kidney Channel. It passes the abdomen and connects with the pericardium of the Heart then flows laterally to connect with the lumbar vertebrae (Figure 7.14).
   a. Symptoms of Excess include enuresis, emotional irritability, and depression.
   b. Symptoms of Deficiency include lower back pain.
3. The Foot Jueyin-Liver 5 Point is located on the Liver Channel of each foot, on the medial side, several inches above the medial malleolus. The Lou's energetic flow connects with the Gall Bladder Channel on each foot and flows upward along the medial aspect of the inner thigh. It connects with the genitals and terminates at the penis in males, and the clitoris in women (Figure 7.14).
   a. Symptoms of Excess include swelling of the testicles.
   b. Symptoms of Deficiency include itching in the pubic region.
4. The Foot Taiyang-Bladder 58 Point is located on the Bladder Channel of each foot, on the lateral side, several inches above the external malleolus. The Lou's energy flows downward to connect with the Kidney Channel on each foot (Figure 7.14).
   a. Symptoms of Excess include nasal congestion, headache, and back pain.
   b. Symptoms of Deficiency include clear mucus nasal discharge and nosebleed.
5. The Foot Shaoyang-Gall Bladder 37 Point is located on the Gall Bladder Channel of each foot, on the lateral side, several inches above the external malleolus. The Lou's energy flows downward to connect with the Liver Channel on each foot, then continues downward to disperse over the dorsum on each foot (Figure 7.14).
   a. Symptoms of Excess include fainting.
   b. Symptoms of Deficiency include weak and flaccid muscles of the feet.
6. The Foot Yangming-Stomach 40 Point is located on the Stomach Channel of each foot, on the lateral side, several inches above the external malleolus. The Lou's energy flows downward to connect with the Spleen Channel on each foot. A second branch flows along the lateral aspect of the tibia, upward to the top of the head where it divides, with one branch converging with the other Yang Channels on the neck and head, and the other branch connecting with the throat (Figure 7.14).
   a. Symptoms of Excess include epilepsy and insanity.
   b. Symptoms of Deficiency include pharyngitis, sudden aphasia, and flaccid or atrophied muscles in the legs or feet.
7. The Hand Taiyang-Small Intestine 7 Point is located on the Small Intestine Channel of each hand, on the dorsal side, several inches above the wrist, on the ulnar side. The Lou's energy flows upward past the elbow and connects with the Large Intestine 15 point (Figure 7.15).
   a. Symptoms of Excess include fever, headaches, and blurred vision.
   b. Symptoms of Deficiency include atrophy of the muscles in the elbow and arm, and a looseness in the joints.
8. The Hand Yangming-Large Intestine 6 Point is located on the Large Intestine Channel of each hand, several inches above the dorsal side of the wrist on the radial side. The Lou's energy flows upward on each arm to the jaw and pours into the area of the teeth. Another branch ascends into each ear, connecting with the Thrusting Vessels which supply energy to the head (Figure 7.15).
   a. Symptoms of Excess include deafness and toothache.

   b. Symptoms of Deficiency include a sensation of coldness in the teeth, as well as fullness and congestion in the chest.
9. The Hand Shaoyang-Triple Burner 5 Point is located on the Triple Burner Channel of each hand, located just above the dorsal transverse crease of the wrist. The Lou's energy flows upward past the arm and over the shoulder, dispersing into the chest and connecting with the Pericardium Channel (Figure 7.15).
   a. Symptoms of Excess include muscle spasms of the elbow.
   b. Symptoms of Deficiency include flaccid muscles in the arm and elbow joint.
10. The Hand Taiyin-Lung 7 Point is located on the Lung Channel of each hand and arises from the cleft of the tendons and bones on the radial side of the wrist, where it flows down into the palm, spreading through the thenar eminence (Figure 7.15).
    a. Symptoms of Excess include hot palms or wrists.
    b. Symptoms of Deficiency include enuresis and shortness of breath.
11. The Hand Shaoyin-Heart 5 Point is located on the Heart Channel of each hand, just above the transverse crease of the wrist. The Lou ascends along the Heart Channel and enters the Heart then continues up the chest into the head, flowing into the root of the tongue, then ascends to connect with each eye (Figure 7.16).
    a. Symptoms of Excess include fullness and pressure in the chest.
    b. Symptoms of Deficiency include aphasia.
12. The Hand Jueyin-Pericardium 6 Point is located on the Pericardium Channel of each hand, just a few inches above the medial transverse crease of the wrist, between the two tendons. The Lou's energy flow follows the Pericardium Channel, connecting with the Heart (Figure 7.16).
    a. Symptoms of Excess include chest pain.
    b. Symptoms of Deficiency include irritability.
13. The Governing Vessel 1 Point is located on the

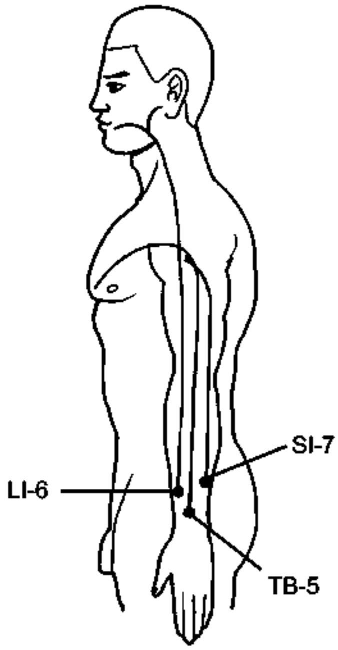

Figure 7.15. The three Yang Collaterals and their Luo points located on the hand and on the Primary Channels of the Large Intestine (LI), Triple Burners (TB), and Small Intestine (SI).

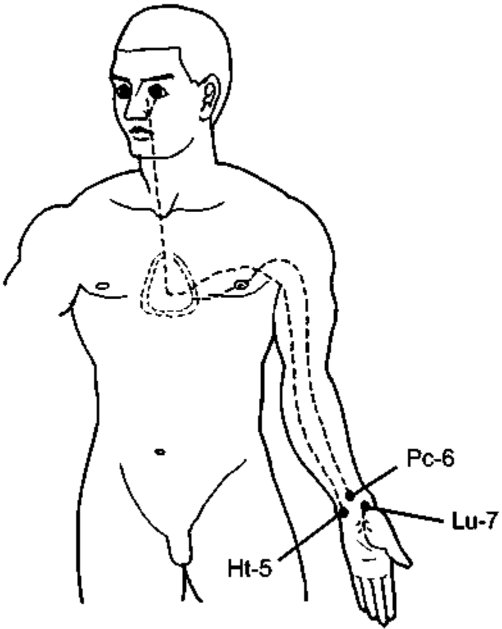

Figure 7.16. The three Yin Collaterals and their Luo points located on the hand and on the Primary Channels of the Heart, Pericardium, and Lungs.

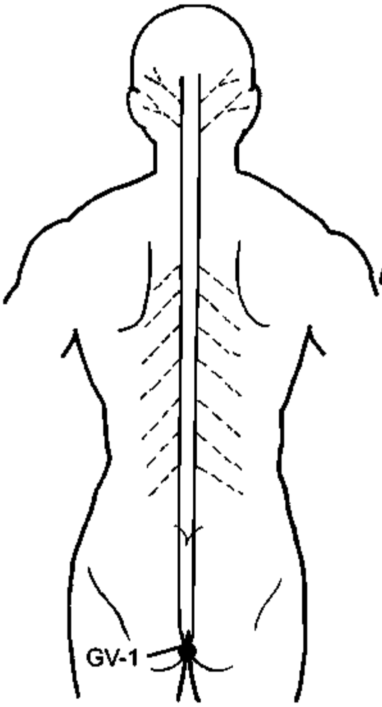

Figure 7.17. The main collateral and Luo point of the Governing Vessel is located on the Governing Vessel (GV) 1 point.

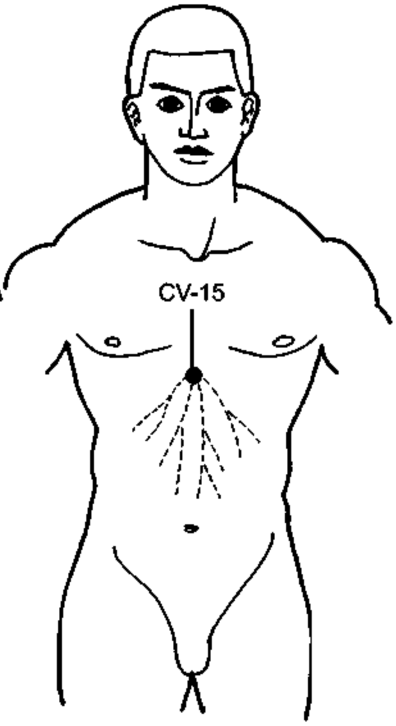

Figure 7.18. The main collateral and Luo point of the Conception Vessel is located on the Conception Vessel (CV) 15 point.

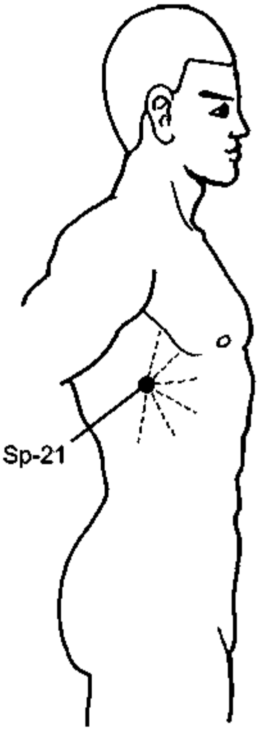

Figure 7.19. The major collateral and Luo point of the Spleen is located on the Spleen (Sp) 21 point.

base of the Governing Vessel, at the perineum. The Lou's energy flows upward along both sides of the spine to the nape of the neck, spreading to the top of the head. It continues flowing into the scalp regions on both sides of the head and connects with the Bladder Channel and merges with the spine (Figure 7.17).

   a. Symptoms of Excess include stiffness along the spine.
   b. Symptoms of Deficiency include dizziness or heaviness in the head.

14. The Conception Vessel 15 Point is located on the Conception Vessel of the chest, just below the xiphoid process. The Lou's energy flows downward, pouring over the abdomen (Figure 7.18).
    a. Symptoms of Excess include pain on the surface skin of the abdomen.
    b. Symptoms of Deficiency include itching on the surface skin of the abdomen.
15. The Major Luo of the Spleen 21 Point is located on the Spleen Channel of the chest, just below the auxiliary fold of each arm. The Lou's energy flow spreads through the chest and hypochondriac region, gathering the Blood like a net (Figure 7.19).
    a. Symptoms of Excess include general aches and pains throughout the entire body.
    b. Symptoms of Deficiency include weakness in the muscles of the limbs and joints.

THE TWELVE DIVERGENT CHANNELS

The Twelve Divergent Channels comprise an important part of the body's channel system. These channels are branches off the Twelve Primary Channels and share the energetic function of circulating Qi throughout the body. The energetic field of the Twelve Divergent Channels forms an enormous web of complex interconnections within the body's Twelve Primary Channels.

Because the area over which the energy of the Twelve Divergent Channels is distributed is quite extensive, they are considered a separate component of the channel system. They are secondary streams that parallel the primary rivers, yet each has its own functional characteristics and clinical application independent of the Primary Channels. The Twelve Divergent Channels run deeper in the body, supplementing the areas along the pathways that the Primary Channels do not reach. They have no points of their own, although there are intersection points where they cross the major channels (Figures 7.20 through 7.25).

One of the primary functions of the Twelve Divergent Channels is to integrate all parts of the body with the Twelve Primary Channels. There are areas in the body which are not traversed by the pathways of the Twelve Primary Channels, as well as internal organs that are otherwise unconnected, or only remotely connected, by the Primary Channels. These areas are more securely linked by the energetic flow of the Twelve Divergent Channels, which strengthen the bonds between the Twelve Primary Channels and the physical areas that are connected to, or adjoining, their pathways.

Another primary function of the Twelve Divergent Channels is to facilitate the connection between pairs of Yin and Yang Primary Channels and Organs. Therefore, all primary Yin and Yang organs are connected by the Divergent Channels. Both Yin and Yang Divergent Channels ultimately connect with the body's Yang Primary Channels.

1. The Yang Divergent Channels complete a cycle of leaving the primary channels (e.g., Primary Gall Bladder Channel) and entering their associated organs (e.g., The Gall Bladder Organ) before they resurface on the neck and reconnect with their original channels (e.g., The Primary Gall Bladder Channel).
2. The Yin Divergent Channels leave their primary channels (e.g., Primary Liver Channel), then join their associated Yang Divergent Channels (e.g., The Divergent Gall Bladder Channel), which then join with the Yang Primary Channels (e.g., The Primary Gall Bladder Channel).

Similarly, within the body's cavities, most of the Twelve Divergent Channels first join with their pertaining Yin or Yang organ and then connect with the associated organ (belonging to the associated channel) in the Yin / Yang pair. Through this interaction, the connection between paired Yin and Yang organs and channels is strengthened.

An understanding of the relationship between the paired Yin and Yang interactions is important to the Qigong doctor, as sometimes a disease affecting a Yang channel can be treated by selecting certain areas on its associated Yin channel, and vice versa. The same theory holds true for diseases of the internal organs.

The Twelve Divergent Channels are paired into six confluences according to their internal and external relationship.

1. Bladder Divergent Channel (Foot Tai Yang) and Kidney Divergent Channel (Foot Shao Yin)
2. Stomach Divergent Channel (Foot Yang Ming) and Spleen Divergent Channel (Foot Tai Yin)
3. Gall Bladder Divergent Channel (Foot Shao Yang) and Liver Divergent Channel (Foot Jue Yin)
4. Small Intestine Divergent Channel (Hand Tai Yang) and Heart Divergent Channel (Hand Shao Yin)
5. Large Intestine Divergent Channel (Hand Yang Ming) and Lung Divergent Channel (Hand Tai Yin)
6. Triple Burners Divergent Channel (Hand Shao Yang) and Pericardium Divergent Channel (Hand Jue Yin)

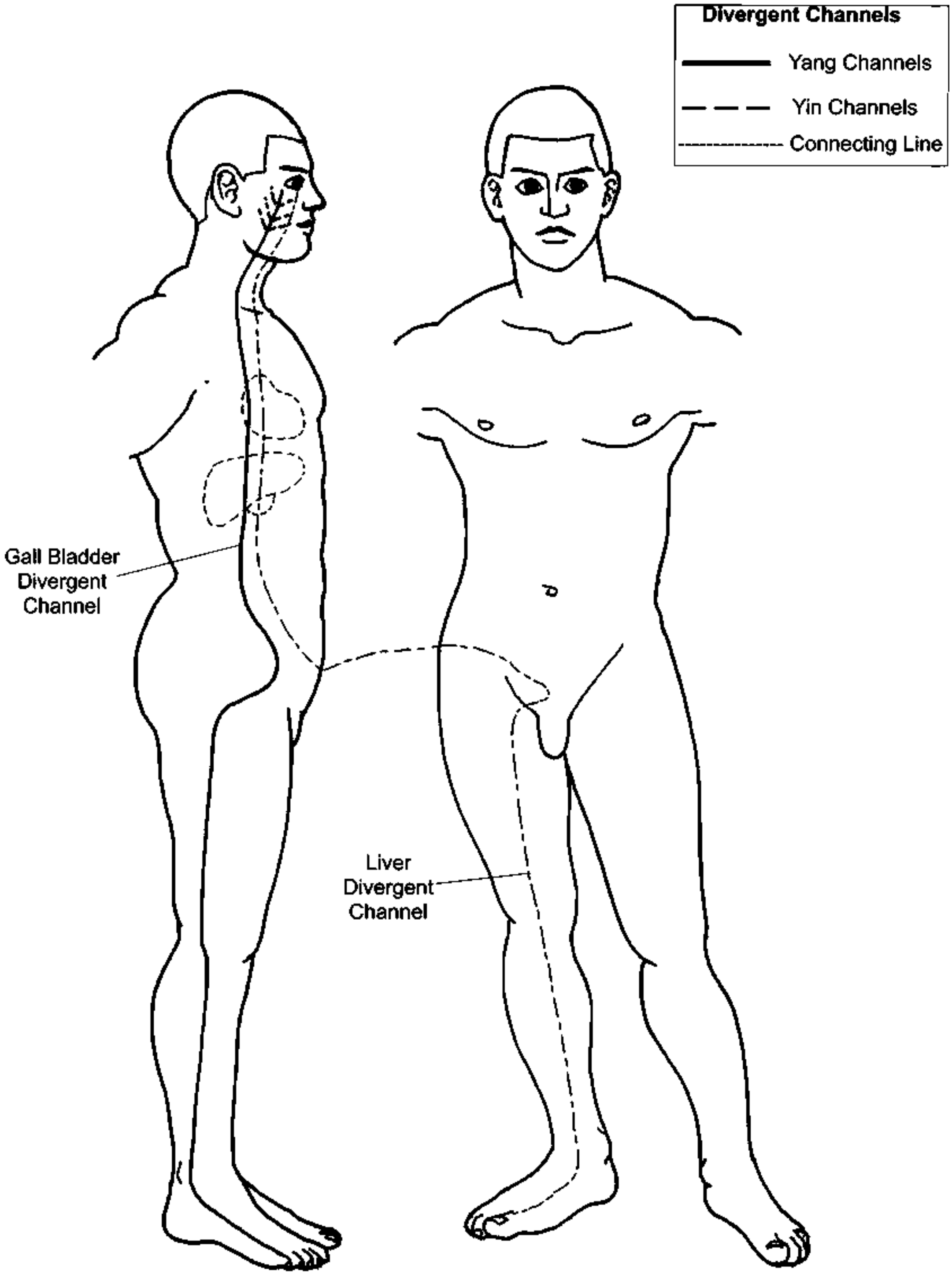
Divergent Channels
Yang Channels
Yin Channels
Connecting Line
Gall Bladder Divergent Channel
Liver Divergent Channel

Figure 7.20. The Divergent Channels of the Gall Bladder and Liver

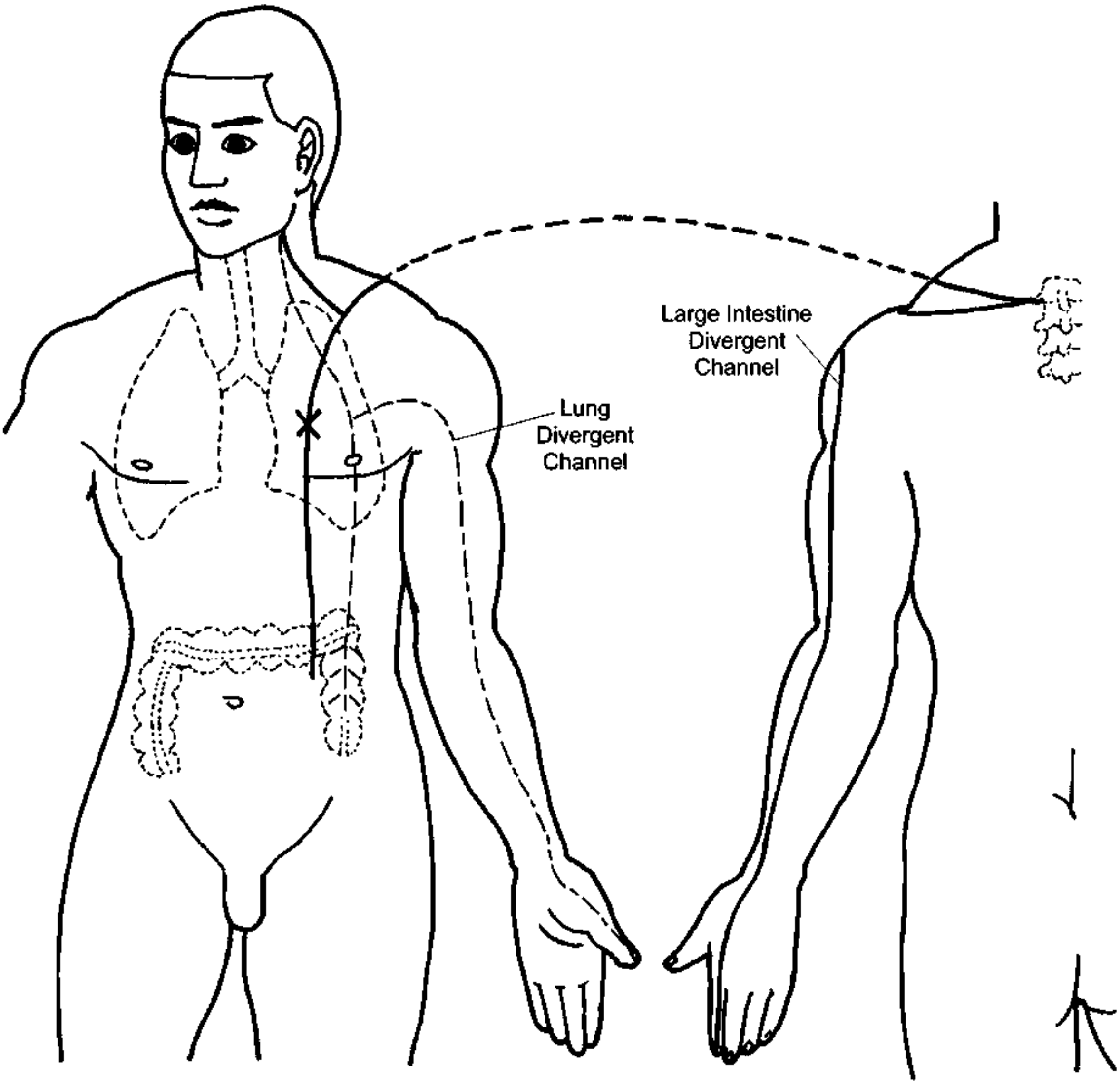
Large Intestine Divergent Channel
Lung Divergent Channel

Figure 7.21. The Divergent Channels of the Lungs and Large Intestine

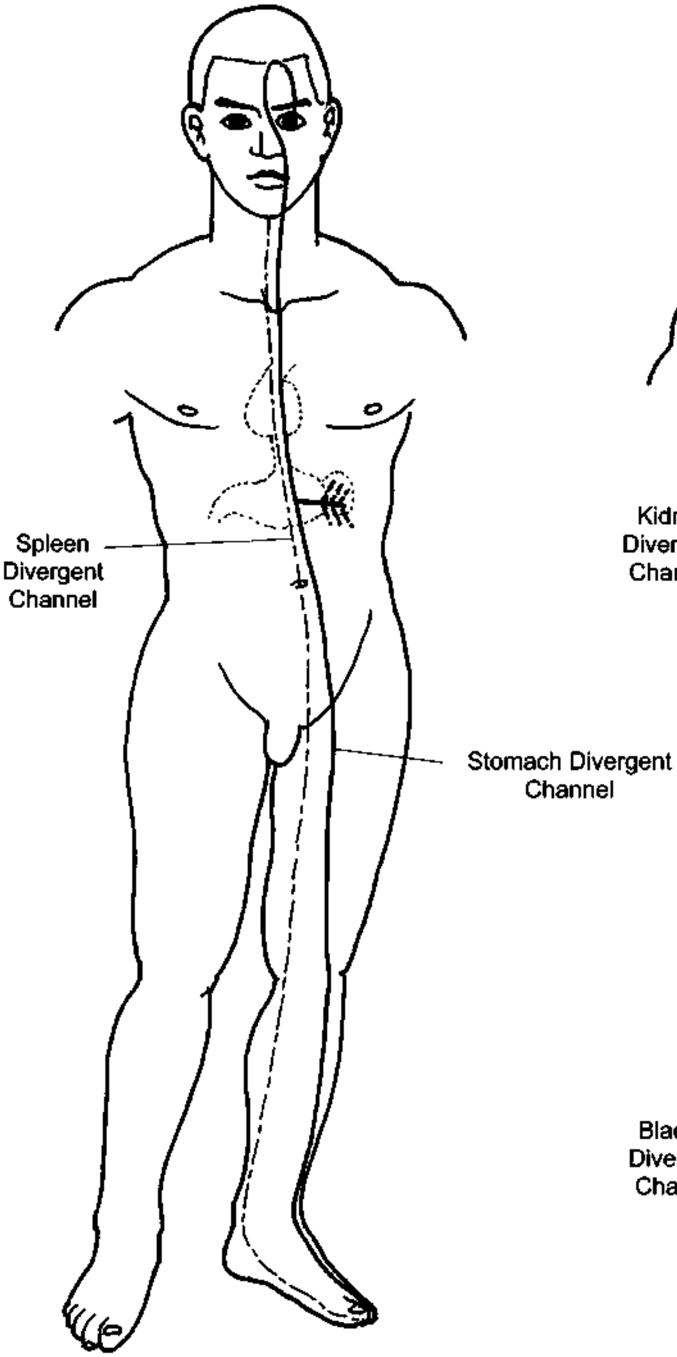
Spleen Divergent Channel
Stomach Divergent Channel

Figure 7.22. The Divergent Channels of the Stomach and Spleen

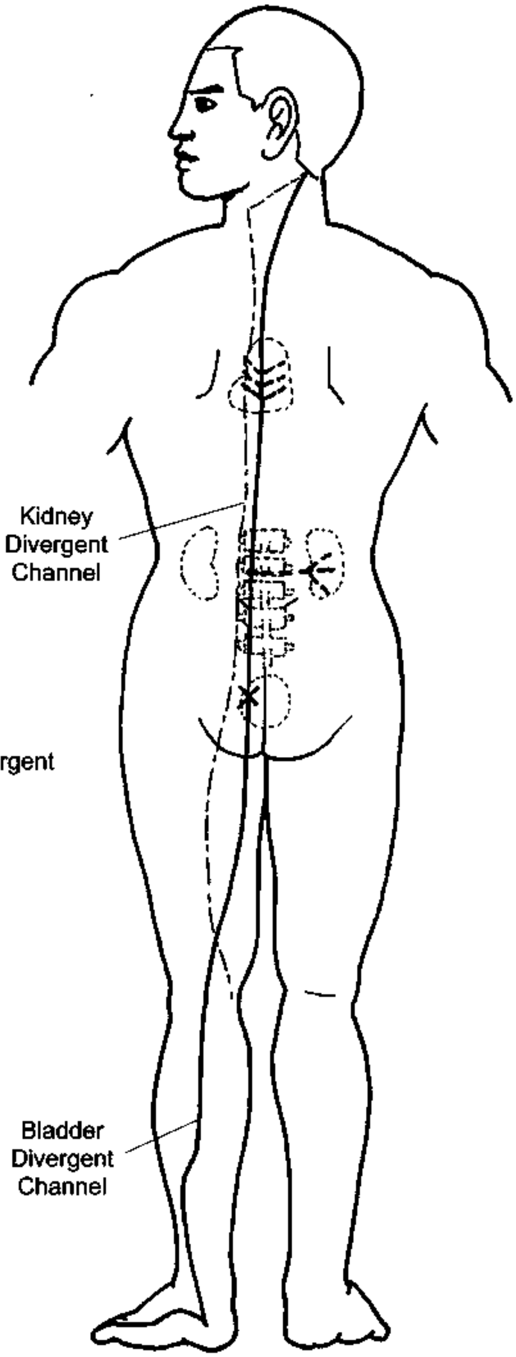
Kidney Divergent Channel
Bladder Divergent Channel

Figure 7.23. The Divergent Channels of the Bladder and Kidneys

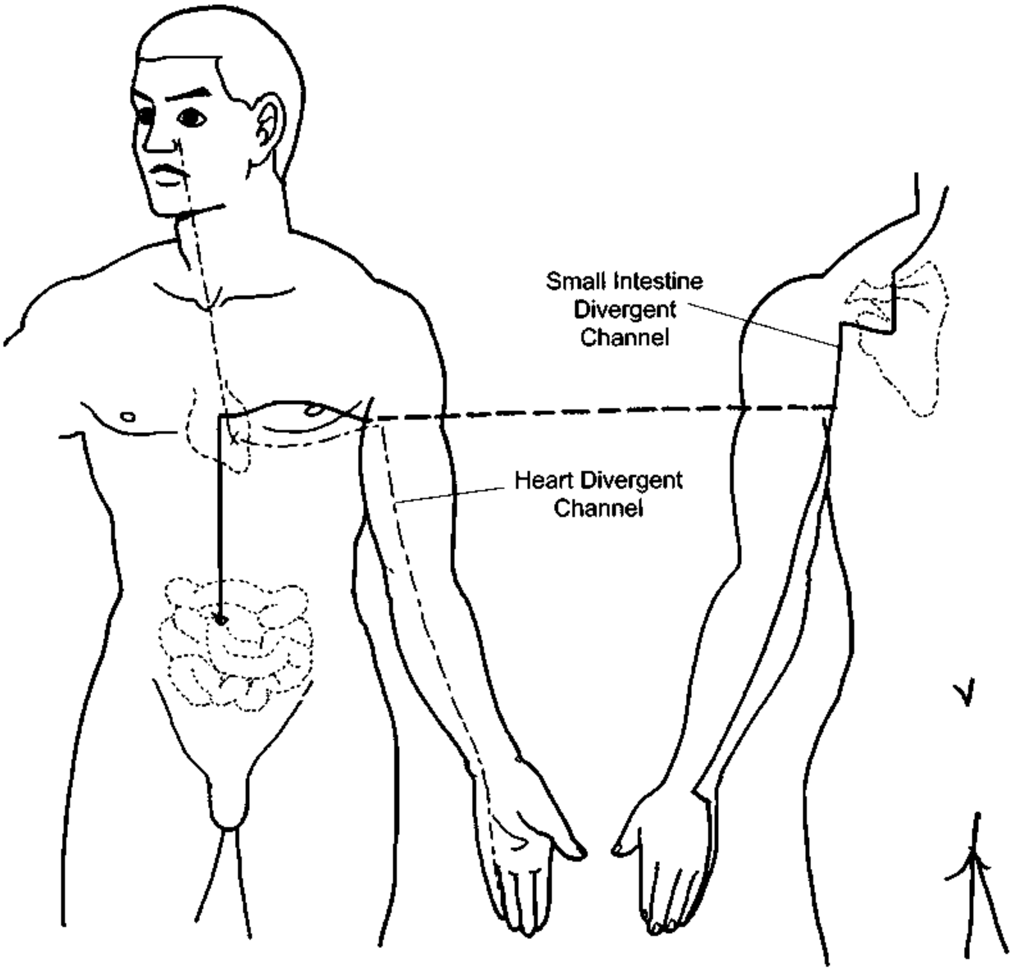
Small Intestine Divergent Channel
Heart Divergent Channel

Figure 7.24. The Divergent Channels of the Heart and Small Intestines

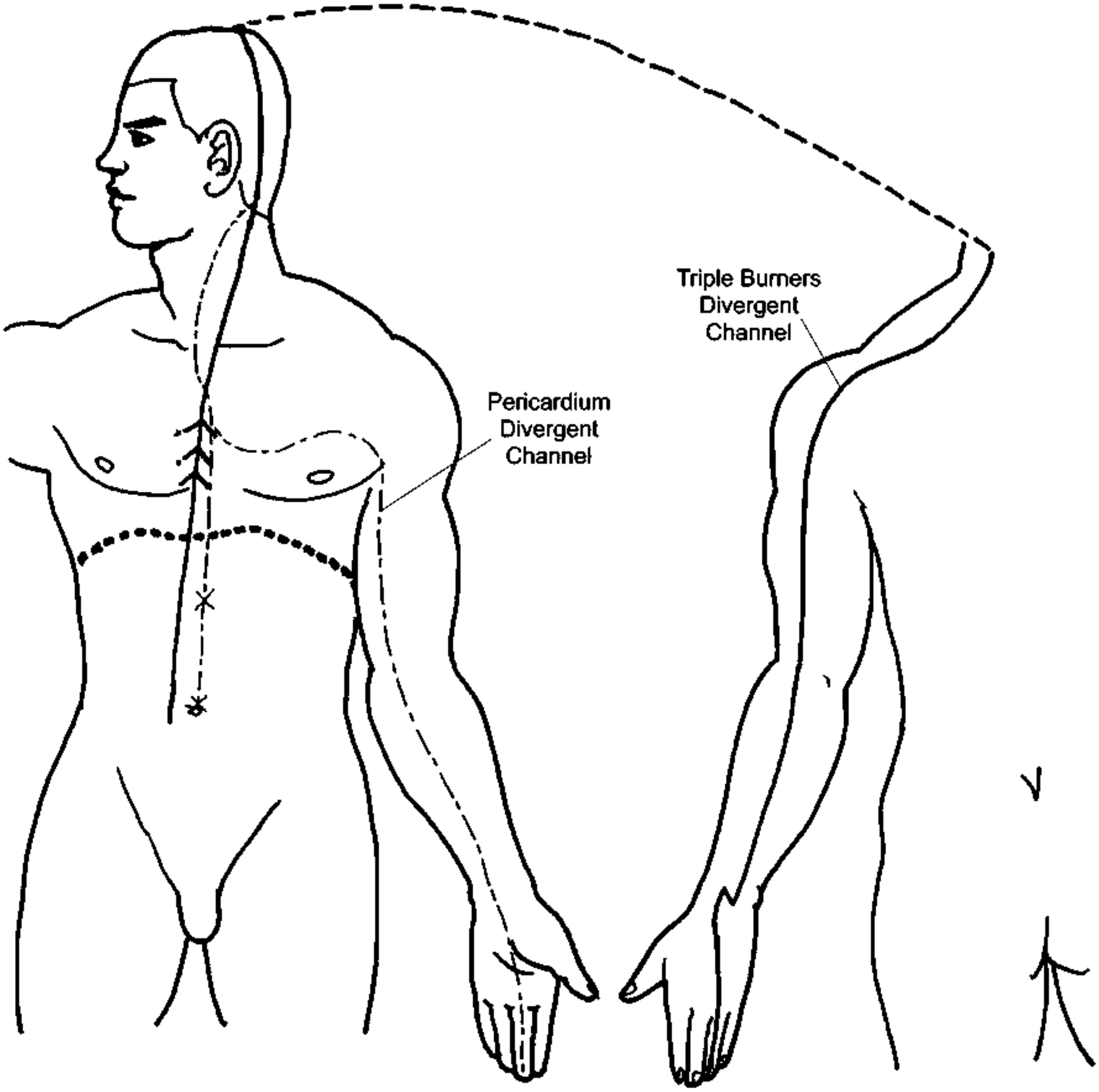
Triple Burners Divergent Channel
Pericardium Divergent Channel

Figure 7.25. The Divergent Channels of the Pericardium and Triple Burners

THE TWELVE MUSCLE/TENDON CHANNELS

The Twelve Muscle/Tendon Channels are the muscle and soft tissue regions of the body where the Qi and Blood of the Primary Channels nourish the skin, muscles, and tendons. They are very superficial in relationship to the Primary Channels and form a capillary network that travels in the depressions and planes between muscles and tendons. The Twelve Muscle/Tendon Channels, subsequently spread over the whole of the epidermis, through its close connection with the cutaneous tissues.

Originating in the extremities, these channels ascend to the head and torso. Physiologically, they do not enter into the internal organs but coordinate the movement of the bones and limbs. The Twelve Muscle/Tendon Channels are connected to the inner fascia of the body's muscles, tendons, and ligaments, as well as other connective tissues. They are affiliated with the channels and collaterals (Luo) systems on the exterior of the body, and serve as mediators between any energetic reactions (trauma, stress, etc.) that vibrate from the body's exterior surface to the deep internal organs.

These channels are found along the four extremities on the surface of the body, as well as along the head, neck, back, chest, and abdomen. The name of each Muscle/Tendon Channel comes from the Twelve Primary Channels whose external energy flow they follow. They also receive Blood and Qi nourishment for their functional activity from and through the Twelve Primary Channels. There are points of connection, as well as intersecting points along these channels (Figure 7.26 through 7.37).

Treatment at these conjunctive points consistently responds successfully to stimulation; therefore, the Qigong doctor can effectively apply Jing Point Therapy to specific areas and points along the Muscle/Tendon Channels.

These channels are responsible for extending and flexing the muscles, tendons, ligaments and joints; their pathology is reflected in symptoms of impaired movement (i.e., pulled, twisted, strained, cramped or atrophied muscles, spasms, etc.).

In Traditional Chinese Medicine the muscles and tendons are described in characteristics of Yin (flexion, contraction, internal rotation, etc.) and Yang (extension, expansion, external rotation, etc.) which balance each other. When the Yin and Yang action of the muscles and tendons fail to balance and regulate each other, Muscle Channel Dysfunction results; for example, when exposed to cold, the muscles/tendons become tense and over-contract; when exposed to heat, muscles/tendons become loose and overextend.

The pathology of the Muscle/Tendon Channels also reflects the symptoms of corresponding groups of muscles and other connective tissues. The connective tissues are divided into three groups: the large, the small, and the membranous connective tissue.

Therapeutically, a local Muscle/Tendon Channel symptom can be treated by stimulating an area located next to the origin of the pain; for example, if the area is Yang (lateral) and overactive, then the Yin (medial) will be underactive and vice versa. Treatment is directed towards restoring the balance between the Yin and Yang Muscle/Tendon Channels.

Being superficial, the Twelve Muscle/Tendon Channels contain Wei Qi; thus, they provide the body's third line of defence against any unfavorable exogenous influence (the first being the body's Wei Qi field, and second being the Wei Qi stored within the skin). Only after overcoming the resistance of the body's Muscle/Tendon Wei Qi can the Evil Exogenous pathogens travel down the length of the channel to penetrate the corresponding Primary Channel at the Jing-Well point. Jing-Well points are points of energetic union, connecting Primary Channel points together with the points where the Muscle/Tendon Channel have their origin.

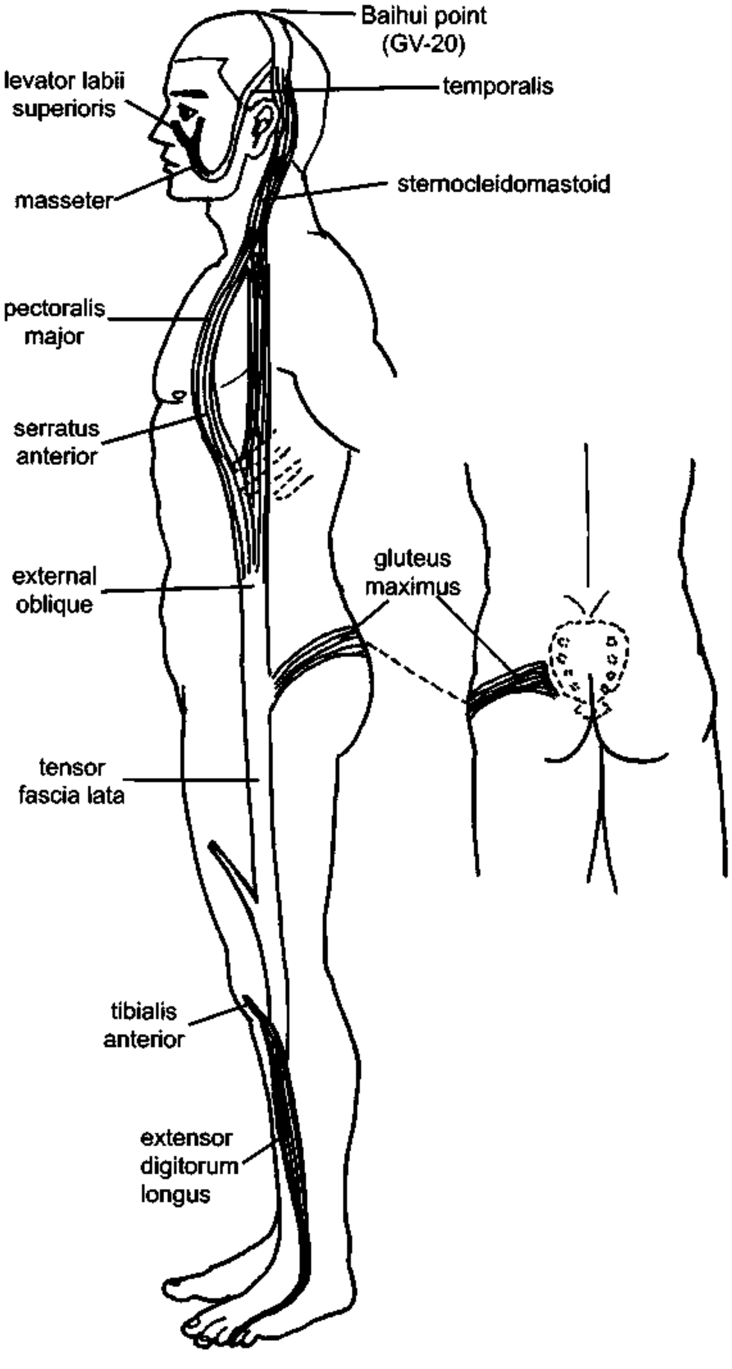
Baihui point
(GV-20)
levator labii superioris
temporalis
masseter
sternocleidomastoid
pectoralis major
serratus anterior
gluteus maximus
external oblique
tensor fascia lata
tibialis anterior
extensor digitorum longus

Figure 7.26. The Muscle Region of the Gall Bladder Channel: High Tide is 11p.m.–1 a.m. Pathological symptoms include strained muscles from the fourth toe to the knee upon lateral rotation, with an inability to bend the knee; muscle spasms or stiffness within the popliteal fossa; strained muscles of the sacrum, pelvis, and lower ribs; pain in the hypochondria, chest, and clavicle region; and an inability to turn the eyes to the left or right.

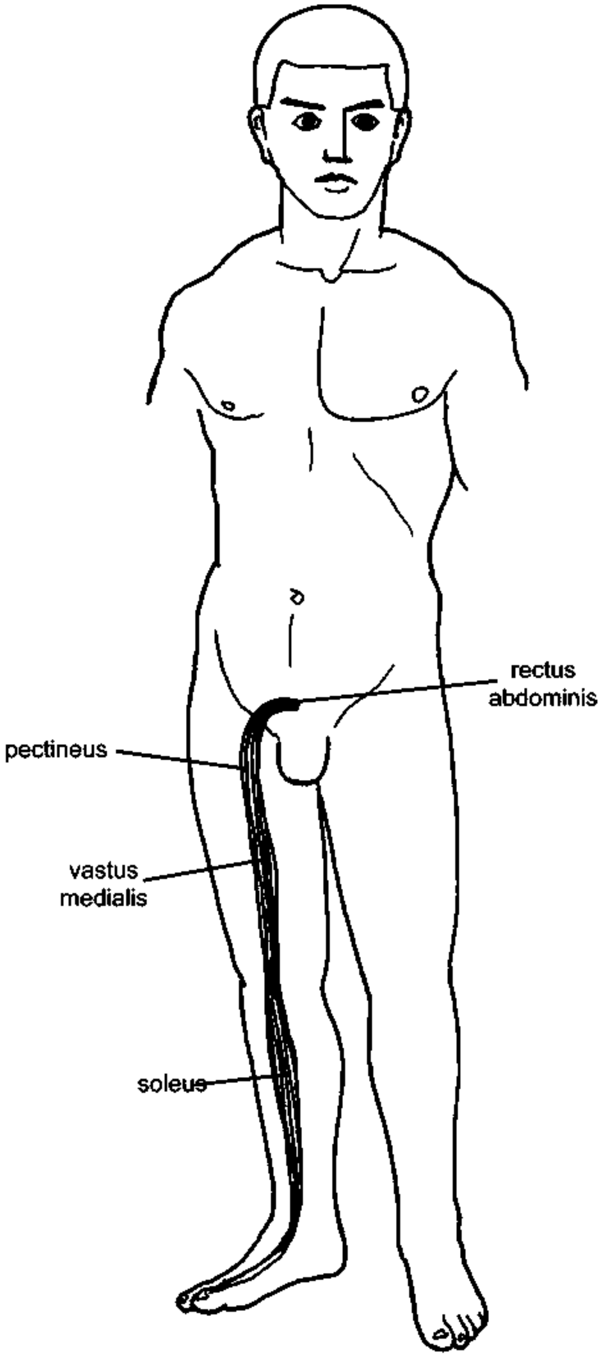
rectus
abdominis
pectineus
vastus
medialis
soleus

Figure 7.27. The Muscle Region of the Liver Channel: High Tide is 1 a.m.–3 a.m. Pathological symptoms include strained muscles of the big toe; pain in the anterior internal malleolus of the ankle; pain at the medial aspect of the knee and thigh; and dysfunction of the reproductive organs, i.e., impotence.

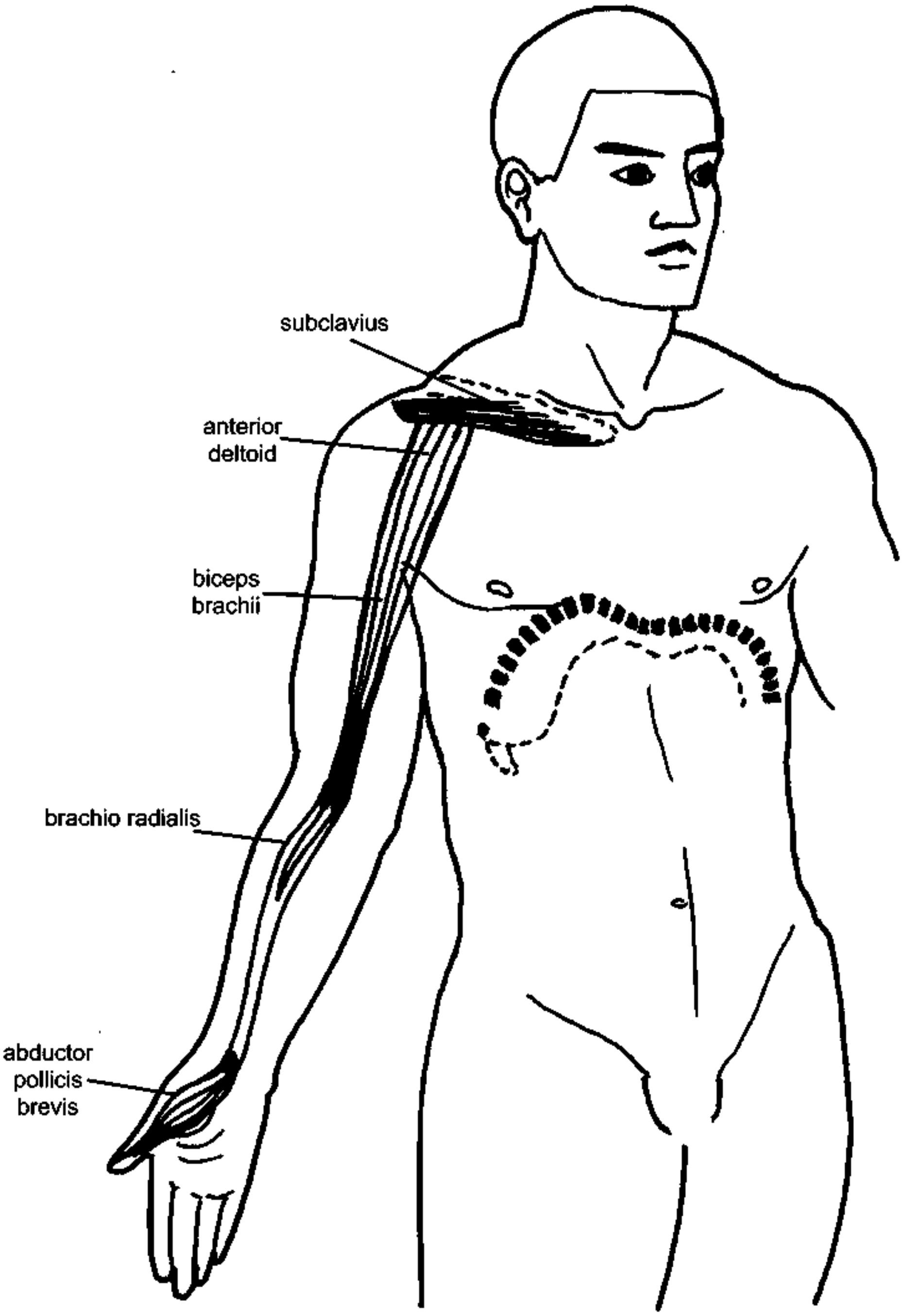
subclavius
anterior deltoid
biceps brachii
brachio radialis
abductor pollicis brevis

Figure 7.28. The Muscle Region of the Lung Channel: High Tide is 3 a.m.–5 a.m. Pathological symptoms include strained muscles of the thumb; stiff, strained or muscle spasms, and/or pain along the course of the Lung Channel. In more serious cases, there will be muscle spasms over the rib area and spitting of blood.

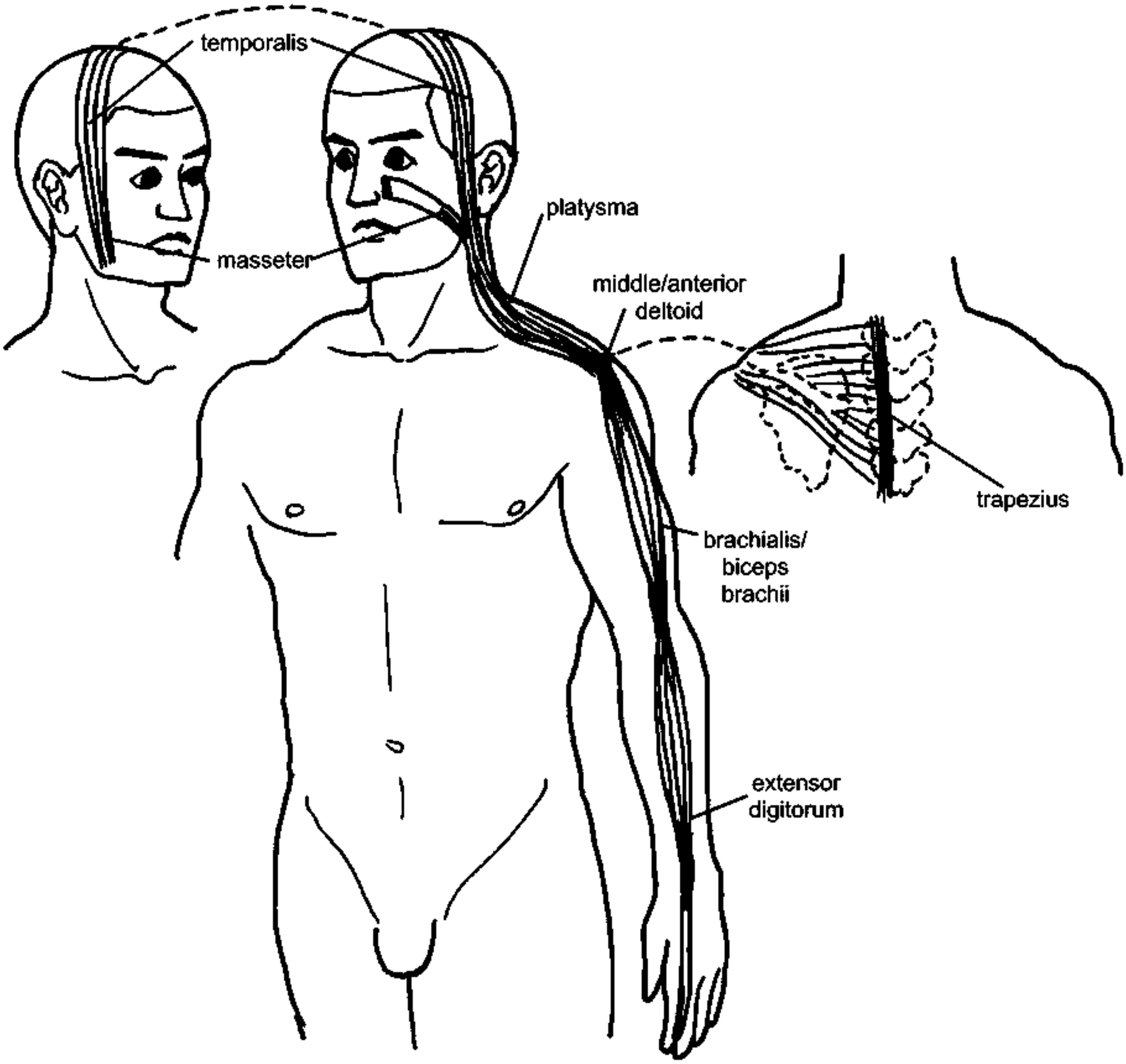
temporalis
platysma
masseter
middle/anterior deltoid
trapezius
brachialis/ biceps brachii
extensor digitorum

Figure 7.29. The Muscle Region of the Large Intestine Channel: High Tide is 5 a.m.–7 a.m. Pathological symptoms include strained muscles of the index finger; stiffness, strained, or muscle spasms along the course of the Large Intestine Channel, resulting in frozen shoulder; and an inability to rotate the neck from side to side.

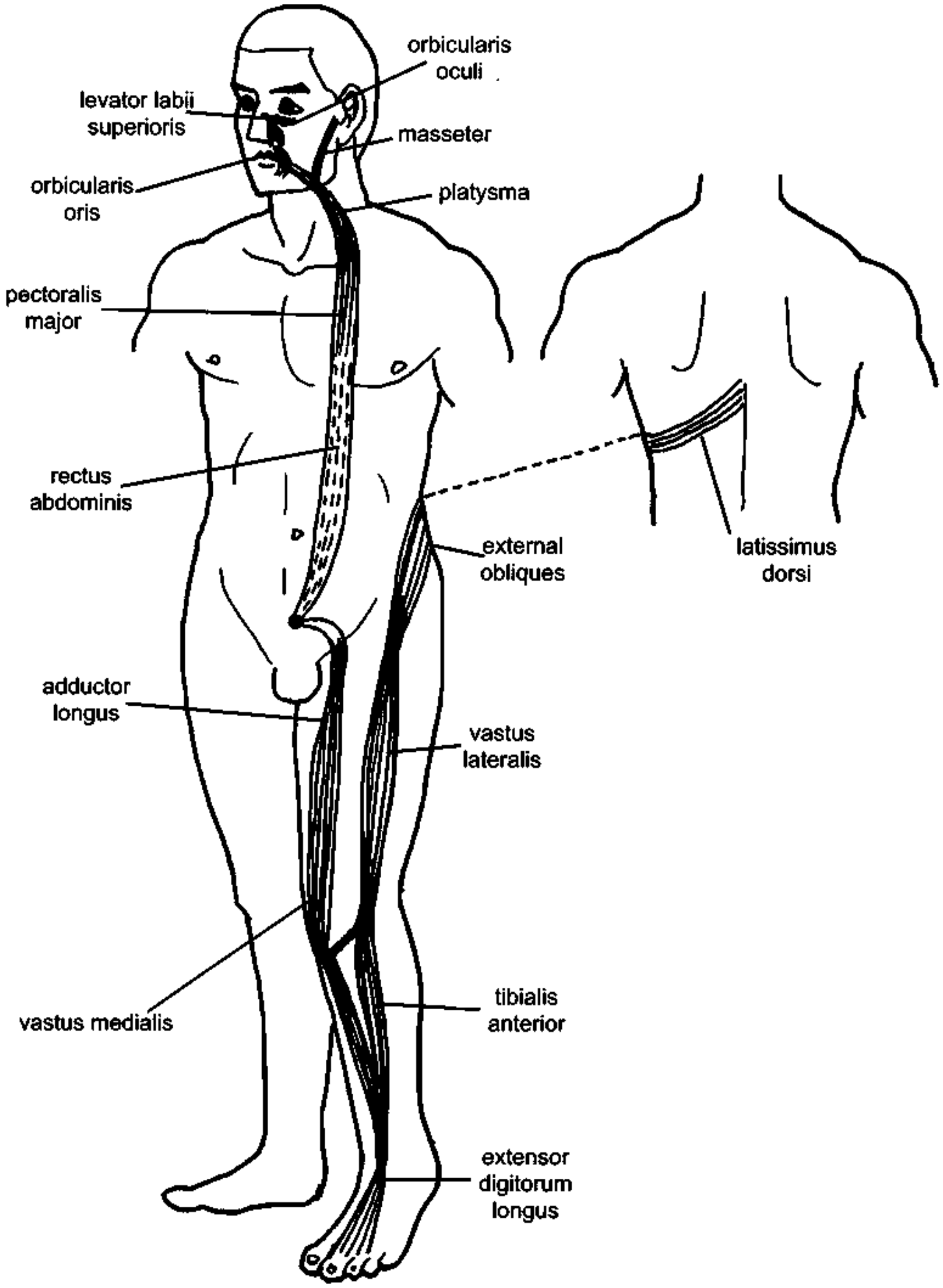
orbicularis oculi
levator labii superioris
masseter
orbicularis oris
platysma
pectoralis major
rectus abdominis
external obliques
latissimus dorsi
adductor longus
vastus lateralis
vastus medialis
tibialis anterior
extensor digitorum longus

Figure 7.30. The Muscle Region of the Stomach Channel: High Tide is 7 a.m.–9 a.m. Pathological symptoms include strained muscles of the big toe; spasms or hardening of the muscles in the foot; knotted or twisted muscles in the lower leg and thigh; swelling in the anterior pelvis region; hernia; spasms of the abdominal muscles; spasms or stiffness of neck and cheek muscles; and eye spasms.

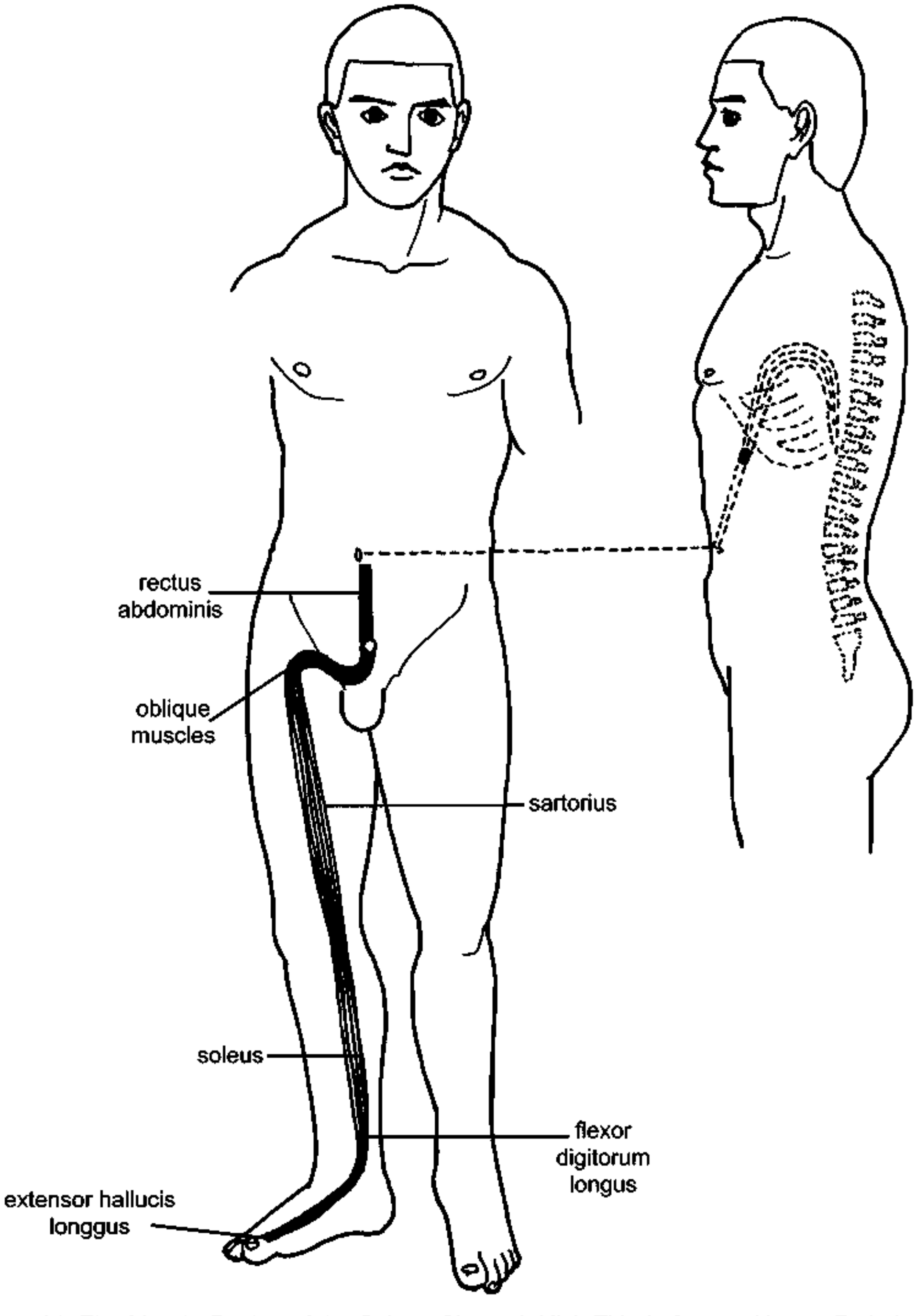
rectus abdominis
oblique muscles
sartorius
soleus
flexor digitorum longus
extensor hallucis longgus

Figure 7.31. The Muscle Region of the Spleen Channel: High Tide is 9 a.m.–11 a.m. Pathological symptoms include strained muscles of the big toe; pain in the internal malleolus of the ankle upon rotation; pain along the medial aspect of the knee and adductor muscles of the thigh; groin strain; and pain due to strained upper abdominal muscles and mid-thoracic vertebrae.

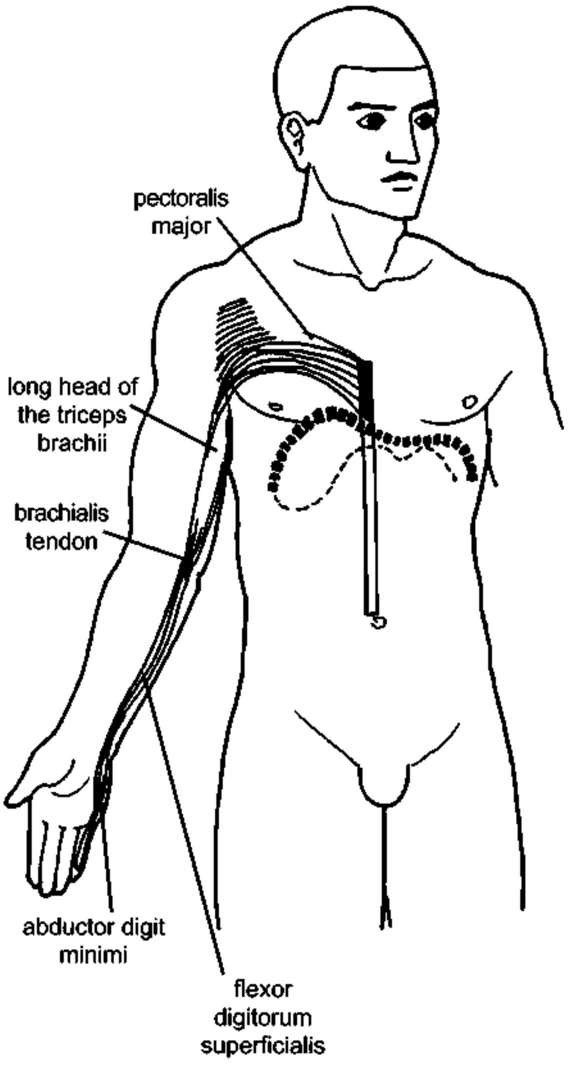
pectoralis major
long head of the triceps brachii
brachialis tendon
abductor digit minimi
flexor digitorum superficialis

Figure 7.32. The Muscle Region of the Heart Channel: High Tide is 11 a.m.–1 p.m. Pathological symptoms include strained muscles of the little finger; stiff or strained muscles with spasm and/or pain along the course of the Heart Channel, including internal cramping within the diaphragm and upper abdominal area.

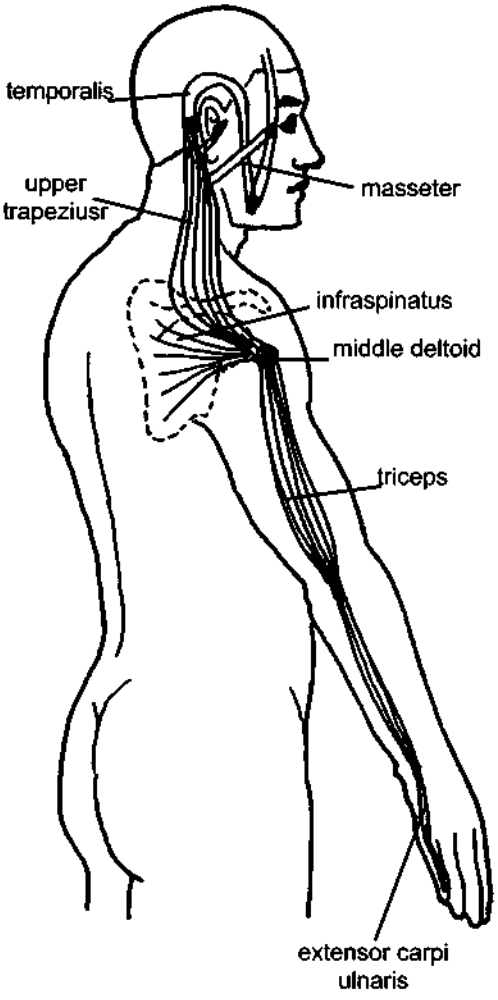
temporalis
upper trapeziusr
masseter
infraspinatus
middle deltoid
triceps
extensor carpi ulnaris

Figure 7.33. The Muscle Region of the Small Intestine Channel: High Tide is 1 p.m.–3 p.m. Pathological symptoms include strained muscles of the little finger; pain along the medial and posterior aspects of the elbow; pain in the posterior aspect of the axilla, neck, and scapula region; tinnitus related to ear ache; and poor vision.

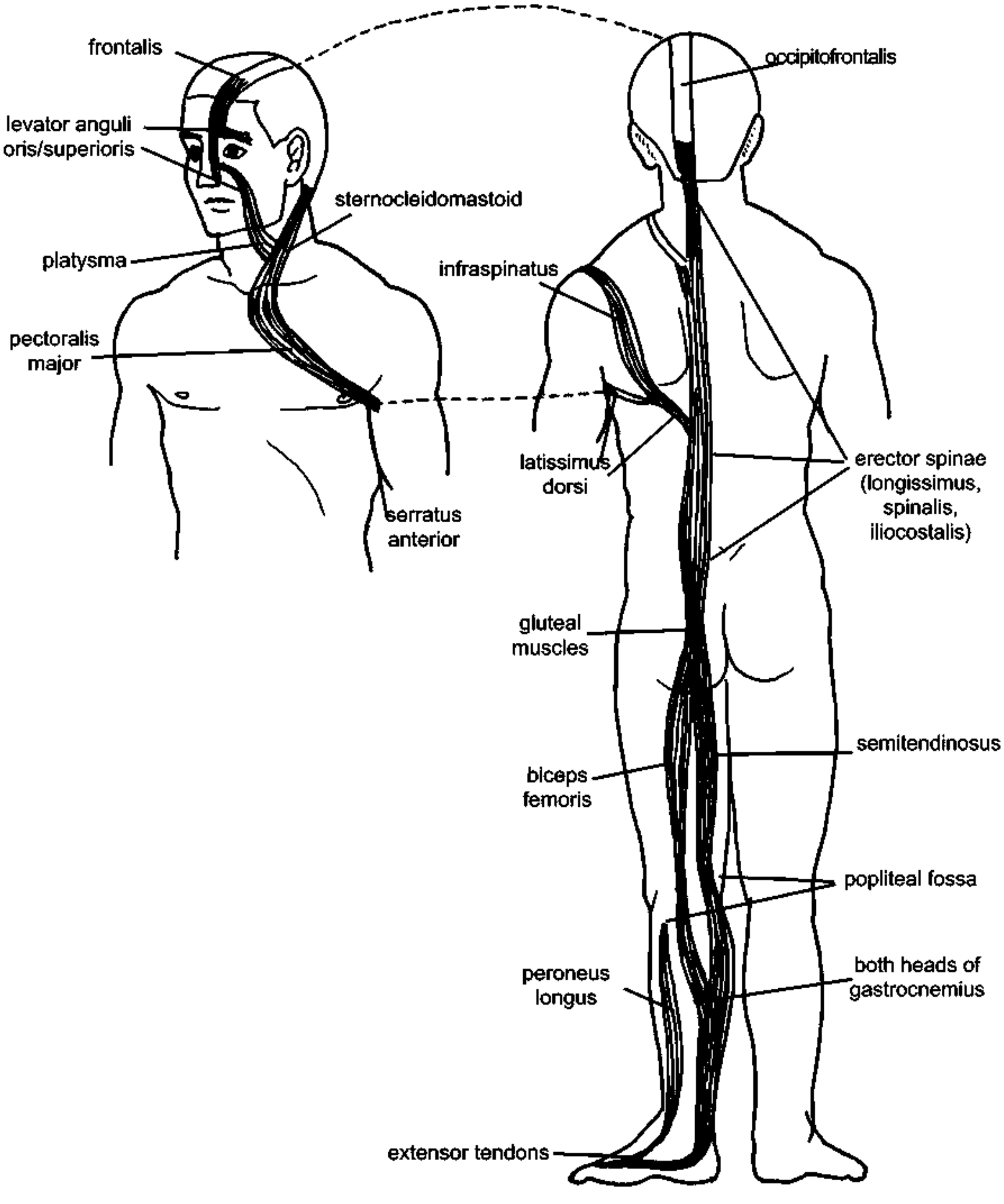
frontalis
levator anguli
oris/superioris
sternocleidomastoid
platysma
pectoralis
major
serratus
anterior
occipitofrontalis
infraspinatus
latissimus
dorsi
erector spinae
(longissimus,
spinalis,
iliocostalis)
gluteal
muscles
semitendinosus
biceps
femoris
popliteal fossa
peroneus
longus
both heads of
gastrocnemius
extensor tendons

Figure 7.34. The Muscle Region of the Bladder Channel: High Tide is 3 p.m.–5 p.m. Pathological symptoms include strained muscles of the big toe; swelling and pain in the heels; stiffness or spasms along the spine and back area; frozen shoulder; stiffness or spasms in the axillary and clavicle regions.

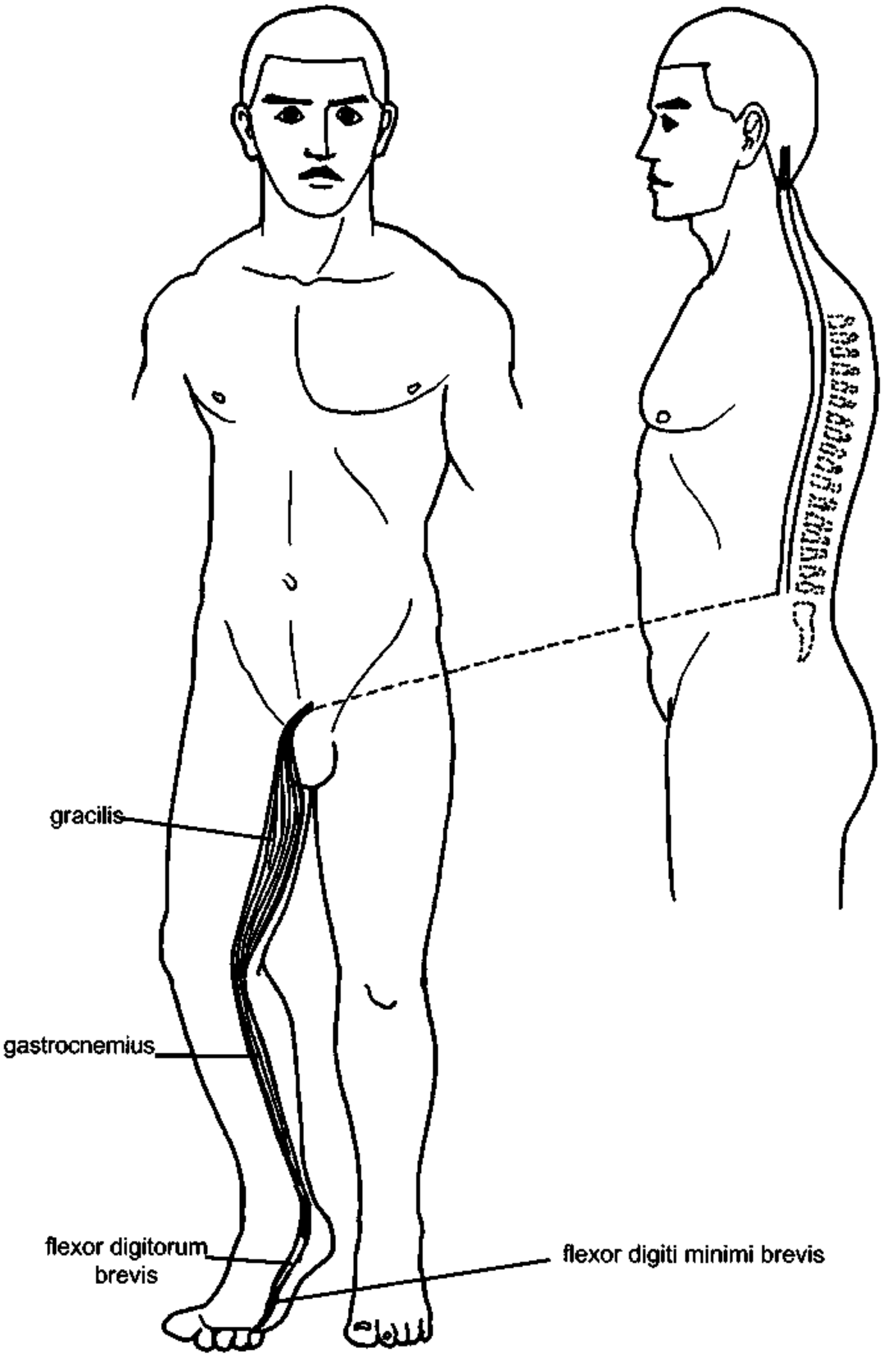
gracilis
gastrocnemius
flexor digitorum brevis
flexor digiti minimi brevis

Figure 7.35. The Muscle Region of the Kidney Channel: High Tide is 5 p.m.–7 p.m. Pathological symptoms include strained muscles on the bottom of the foot; spasms or stiffness along the Kidney Channel, resulting in an inability to bend forward (Yang disorder) or backward (Yin disorder), with difficulty in flexing or extending the head.

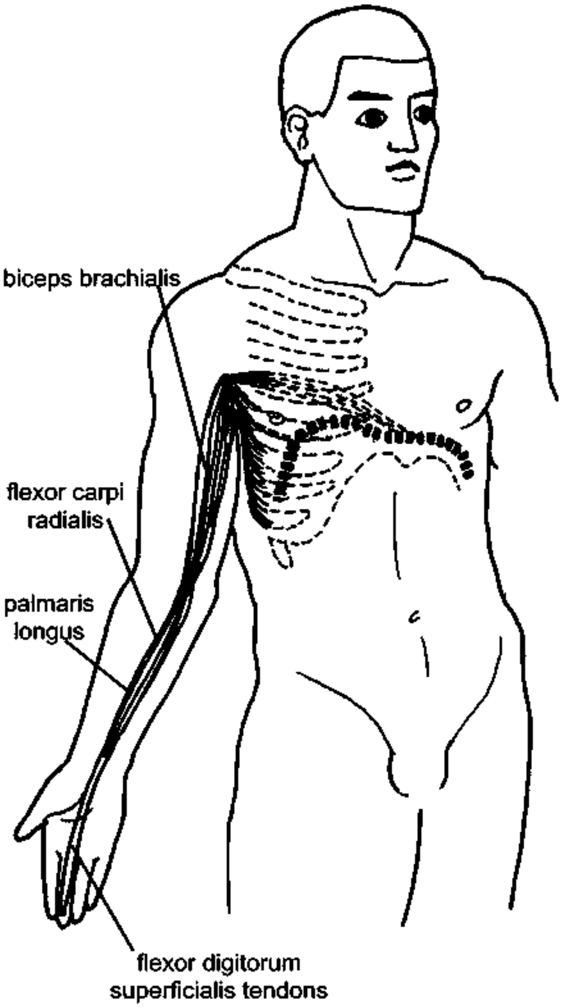
biceps brachialis
flexor carpi radialis
palmaris longus
flexor digitorum superficialis tendons

Figure 7.36. The Muscle Region of the Pericardium Channel: High Tide is 7 p.m.–9 p.m. Pathological symptoms include strained muscles of the middle finger; stiff or strained muscles, or spasms and/or pain along the course of the Pericardium's Channel; and chest pain and spasms.

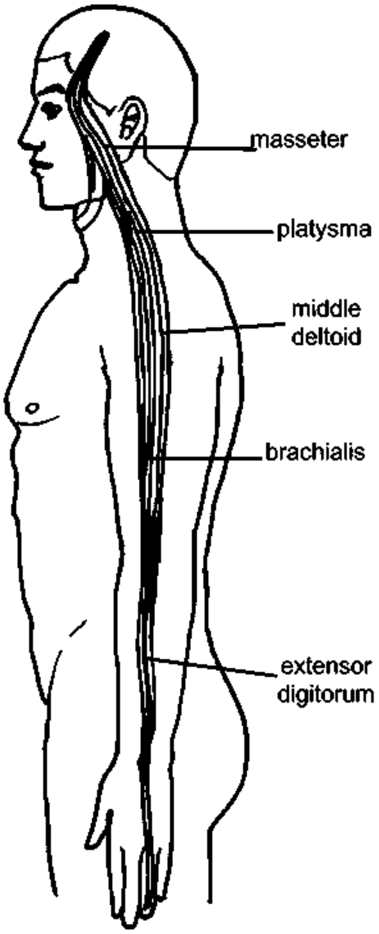
masseter
platysma
middle deltoid
brachialis
extensor digitorum

Figure 7.37. The Muscle Region of the Triple Burner Channel: High Tide is 9 p.m.–11 p.m. Pathological symptoms include strained muscles of the ring finger; stiff or strained muscles, or spasms and/or pain along the course of the Triple Burners' Channel.

THE TWELVE SKIN ZONES

Based on the Twelve Primary Channels, their collaterals, and the body's Muscle/Tendon Channels, the surface tissue is divided into Twelve Skin Zones. These Skin Zones are the surface contact areas for the body's channel and collateral systems. These twelve regions (sometimes called the Twelve Cutaneous Channels) are located in the superficial layers of the skin and have continuous and direct contact with the external environment. They are, therefore, the most sensitive to climactic changes and must adapt to protect the body from External pathogenic factors.

Both the Qi and the Blood at the Twelve Skin Zones rely on the collaterals for their nourishment. The Twelve Skin Zones circulate Wei Qi, which in addition to its defensive function, is also in charge of opening and closing the pores. The skin relies primarily upon the strength of the Wei Qi for resistance to the invasion of External pathogenic influences, as these harmful influences must first penetrate the skin before they can affect the body's internal tissues and organs. Since the Lungs rule the skin, weak Lung Qi can allow pathogens to penetrate the skin and affect the Lungs.

Pathological symptoms associated with the Twelve Primary and connecting Channels manifest along the surface of the Twelve Skin Zones, before progressing deeper into the body's connective tissue. The early stages of disease are called Exterior conditions; if the exterior Wei Qi is strong enough, pathogens will be stopped at the external level. The order of pathogenic progression is as follows.

1. The Body's Wei Qi is the body's first line of defense. A healthy body will project a strong, protective, external energetic field, which will maintain a capable defensive boundary. If the patient becomes weak, tired, or stressed, however, External pathogens may advance onto the patient's skin and begin attacking his or her health.
2. The Skin is the body's second line of defence. If the body's Wei Qi is not strong enough to resist an External attack and the skin is not capable of warding off the pathogens, then the pathogens will attack the skin, causing the sweat pores to open, thus allowing the pathogens to advance towards the patient's collaterals.
3. The Collaterals are the third line of defense. If the body's collaterals are unable to redirect or purge the advancing pathogens, the pathogens will then advance into the patient's channels.
4. The Channels are the last line of defense before the advancing External pathogens invade the patient's internal organs. If the body's channels are unable to redirect or purge the pathogens, and the pathogens are allowed to continue their progression, they will advance further into the body's internal organs.
5. The Internal Organs are affected once the pathogens have breached all the outer defences and disease can now begin to establish its residence.

Pathologies associated with the Primary Channels manifest diagnostically through the body's skin, channels, and points. The observation of pimples, moles, discoloration, and changes in electro-conductivity are all valuable signs of pathogens, or obstructions, invading the skin.

A disease of external origin that first lodges within the patient's skin can be treated by the Qigong doctor. The doctor stimulates the Wei Qi in the affected skin zone through external Qi projection before it progresses further into the body's connective tissue. An alternate treatment would be to apply herbal ointments and/or moxa therapy.

THE CONNECTIVE TISSUE OF THE TWELVE SKIN ZONES

The connective tissue always responds to the outside stimulation of the Twelve Skin Zones. The body's connective tissue is just below the skin, and is regarded as one of the largest, most extensive tissue organs in the body. Its function is to support, connect, contain, and transmit. It is a continuous network of structure that binds tissues into their organ shape. It also supplies the internal organ's life support with vessels and ducts, and securely fastens each organ within the body

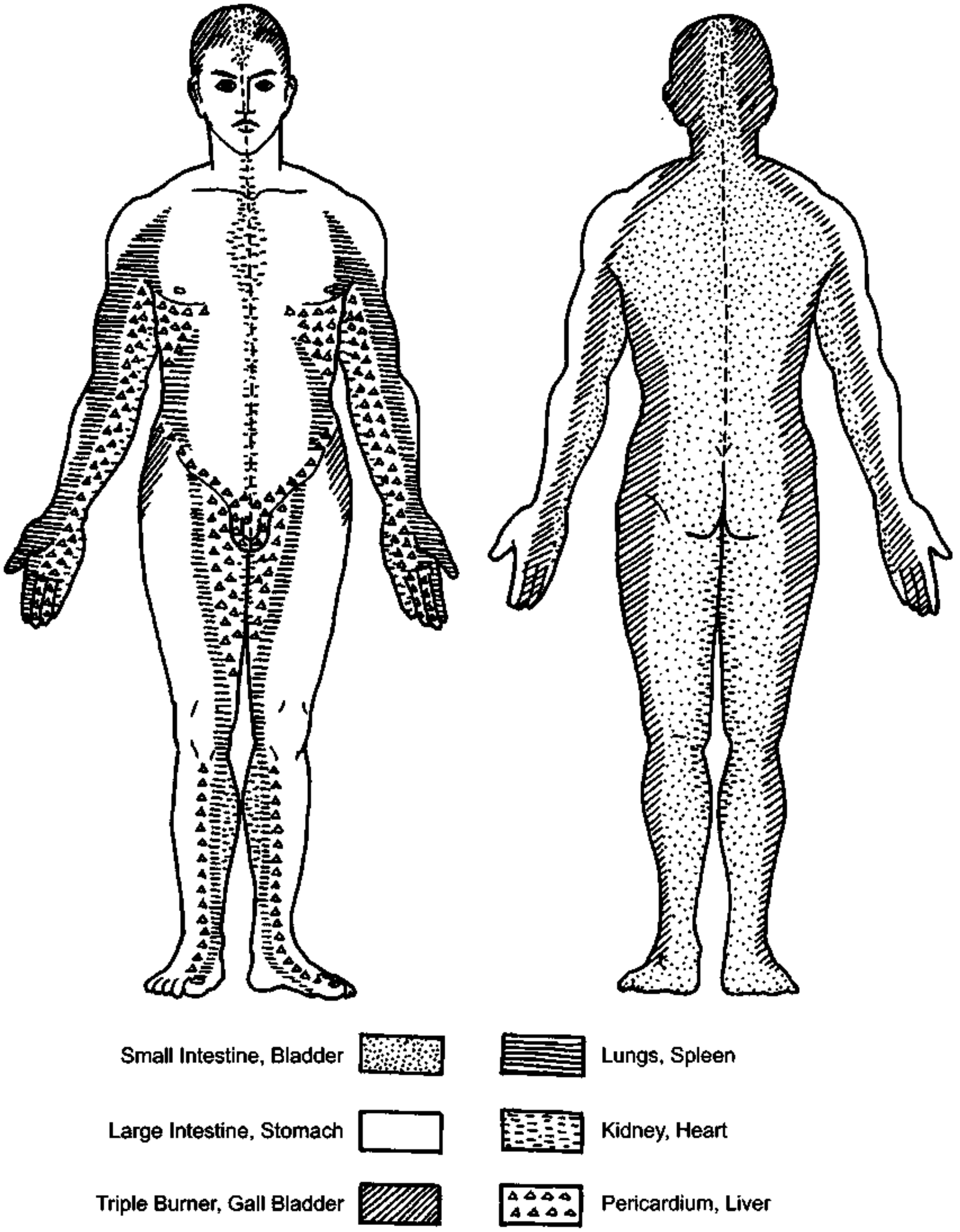
Small Intestine, Bladder
Lungs, Spleen
Large Intestine, Stomach
Kidney, Heart
Triple Burner, Gall Bladder
Pericardium, Liver

Figure 7.38. The Body's Twelve Skin Zones, Based on the External Flow of Qi from the Twelve Primary Channels and Their Collaterals

cavity. This connective tissue surrounds and anchors the vessels within the muscles, bones, and organ tissue. Its fluid nature supports the entire body structurally by hydrostatic pressure. Any stimulation on one of the body's Twelve Skin Zones will directly affect the body's connective tissue, and will stimulate the corresponding internal organ associated with that particular zone.

The superficial fascia (the connective tissue just under the skin) divides into a top and bottom layer. The top layer is the fatty layer, which constitutes the main fatty tissue of the outer surface of the body and fascia. This fatty layer acts as an insulator, helping to maintain a constant body temperature. It is metabolically active, functionally storing fat as fuel, releasing it in response to nerve and hormonal stimuli. It corresponds to the greasy layer where the Wei Qi circulates throughout the body's tissues. The deeper layer of the superficial fascia envelopes the nerves, veins, arteries, lymph vessels, and nodes.

From a Traditional Chinese Medical perspective, the Qi circulating within the body's inner fascial connective tissue can be stimulated through the external energy stimulation of the Twelve Skin Zones. Heat and movement help maintain the body's connective tissue's base fluids, facilitating movement and the conduction of energy. Any obstruction occurring on the body's surface tissues can result in the binding or thickening of the connective tissue, thus creating adhesions. This obstruction slows the metabolic process and compromises the body's immunity (Figure 7.38).

If the integrity of the connective tissue is compromised, the body's immune system declines. Compartments of connective tissue influence the spread of toxins, diseases, infections, and tumors. The fibrous walls, as well as chemicals in the fluid of the connective tissue, prevent the spread of pathogens from one area of the body to another.

CLINICAL DIAGNOSIS AND THE TWELVE SKIN ZONES

Clinically, the diagnosis and treatment techniques relating to the use of the Twelve Skin Zones is quite extensive. In diagnosis, the skin's surface area along these zones is examined for evidence of discoloration or tissue obstruction. Changes in skin color and tissue formations are observed and categorized as follows.

1. A darkish hue reflects obstruction of Qi and Blood.
2. A bluish-purple color indicates local pain.
3. Change in color from yellow to red shows evidence of Heat.
4. A white or pallid skin tone signifies Deficiency or Cold.
5. Boils, pimples (especially on the back), hives, and eczema, as well as hard lumps or nodules beneath the surface of the skin, indicate diseases associated with the Twelve Primary Channels (manifesting through the Twelve Skin Zones).

A palpable lump can develop from a variety of sources, including congealed or stagnant Blood. There are generally eight types of lumps observed in the clinical environment:

1. The Shrimp Lump, is shaped in the form of a shrimp.
2. The Turtle Lump, is named for its shape and multiple lump formations.
3. The Blue Lump, derives its color from the superficial blood vessels on which it forms.
4. The Single Lump, is diagnosed as opposed to multiple lump formations.
5. The Dry Lump, indicates an obstruction.
6. The Blood Lump, relates to congealed Blood.
7. The Abdominal Lump, is named for its physical location.
8. The Yellow Lump, is named because of the yellow pallor in the specific region of the lump.

PALPATING THE PATIENT'S SKIN

Before palpating and diagnosing the skin, the Qigong doctor first scans the patient's body to feel the temperature, moisture, and texture of the patient's tissues.

1. The doctor scans and palpates the patient to diagnose the condition according to the temperature:
   a. If the skin feels hot to the touch it often indicates a presence of a Damp-Heat condition.

   b. When first touching the patient with light pressure if the skin feels immediately hot, and as the pressure is maintained the feeling of heat increases, this indicates an invasion of Exterior Wind-Heat (with a pathogenic factor still on the exterior surface).
   c. If the skin over a blood vessel feels hot on medium pressure, it indicates Interior Heat in the Middle Burner or Heart area.
   d. If the skin feels hot on deep, heavy pressure, it indicates an Empty-Heat condition from a Yin Deficiency.
   e. If the skin feels cold to the touch, it often indicates a Cold condition. This condition is often manifested in the lower back and lower abdominal region, indicating a Deficiency of the Kidney Yang.
2. The doctor scans and palpates the patient to diagnose the condition according to the amount of moisture:
   a. A moist feeling of the skin may indicate an invasion of the Exterior by Wind-Cold or Wind-Heat.
   b. If the skin feels moist due to spontaneous sweating, it indicates a Deficiency of the Lungs' Qi (in the absence of exterior symptoms).
3. The doctor scans and palpates the patient to diagnose the condition according to the texture:
   a. If the skin is scaly and dry, it indicates an exhaustion of the Body's Fluids.
   b. If the skin feels dry, it indicates either a Yin condition of the Lungs or a Blood Deficiency.
   c. If the skin is swollen and an indentation is left visible after pressing, it indicates Edema (called Water Swelling).
   d. If the skin is swollen and no indentation is left visible after pressing, it indicates a retention of Dampness (called Qi swelling).

CHANNEL AND COLLATERAL THERAPY

Because the Yin and Yang organs' Qi flows within the body's internal and external surfaces, all the internal and external channels and collaterals, five tissues (tendons, blood vessels, muscles, skin, and bones), and five sense organs (eyes, ears, nose, mouth, and tongue), together create an energetic network. Through practicing sitting meditations, patients as well as Qigong practitioners become aware of the circulation of Qi along the body's channels and collaterals. The awareness of what the Qi feels like enables the meditators to feel the flow and function of each channel, as well as the Blood and heat cycle in each Skin Zone. This makes it possible to control the Qi circulation through intention and imagination.

Balancing the Conception and Governing Vessels is of the utmost importance in the practice of any Qigong regulation. Along these two Vessels the Qigong practitioner draws the Yang Fire and Yin essences up and down the body, thus fusing the Water and Fire energies. This fusion is utilized by the doctor in every Qigong treatment, and is responsible for balancing the patient's Yin and Yang energy. This balance is created through the discipline of the Microcosmic Orbit meditation, which connects the Governing and Conception Vessels.

The students' awareness of Qi flowing through the body usually develops in three distinct stages:

1. First, they begin to feel the energy flow along the surface channels of the body, especially within the areas of the extremities. Usually at this stage, the students will feel heat and tingling sensations within the body's surface muscles.
2. Next, they begin to feel the energy flow deeper within their tissues, especially along the tendons, deeper muscles, and visceral organs. Usually at this stage, they begin to feel mild electric shocks and vibrations within their muscles, bones, and visceral organs.
3. In the final stage, they feel the Earth and Heaven energy penetrate through their body's

outside channels and inner organs, connecting into their Taiji Pole, and then turning outward again. Usually at this stage, they will feel their entire energized body vibrate. This vibration may be triggered through either universal or environmental changes, as the Medical Qigong practitioner becomes hypersensitive to any form of external energetic disturbance.

CONTROLLING QI EXTENSION THROUGH THE CHANNELS

The Qigong doctor must first be able to control his or her own energy circulation, causing it to flow in or out, expand or contract at will, before beginning to extend energy for the treatment of any patient. This energetic control is gained through specific imagination and visualization techniques. The fundamental theory for these techniques is "The imagination leads the Mind, the Mind leads the Qi." The doctor must be able to utilize energy from the natural environment (trees, waterfalls, ocean, desert, etc.) in order to replenish and facilitate his or her own energy.

By drawing in Heavenly Qi and combining it with his or her body's connection to Earthly Qi, the doctor may mobilize and activate the Qi of the patient. This combining method establishes an energetic healing field of Qi, which responds to the Qigong doctor's own Qi circulation, as well as to that of the natural environment, which includes the patient. This ability must be achieved in order to activate the patient's Qi and then regulate it by purging the Excess Qi, replenishing the Deficient Qi, dredging the channel Qi, and guiding the Qi back to its origin.

CHAPTER 8
THE BODY'S ENERGETIC POINTS

INTRODUCTION TO ENERGETIC POINTS

The Chinese word for point is Xue or Xue Wei, meaning "cavity place, cave, den, hole, or hollow." Although the Chinese word Xue is usually translated into English as "point" or "acupoint," it is valuable for the English-speaking student to understand the word's different connotations. Points are almost always located in hollows or depressions, which are physically palpable and found in over 1,000 locations all over the body. Many, but not all, are located along the channels and vessels. Sometimes when touching a point, it feels as if your fingertip has found the entrance of a small cave or opening hidden by the skin covering it.

The word Xue also implies a den or lair: thus, it is not only a cave but a home for some type of powerful life form. In the human body, our living Qi gathers and pools within these protected hollows (Figure 8.1). For this reason, some writers have preferred to translate Xue as "vital hollow," rather than "point." Because of its popular acceptance, I have chosen to use the term "point" throughout this book.

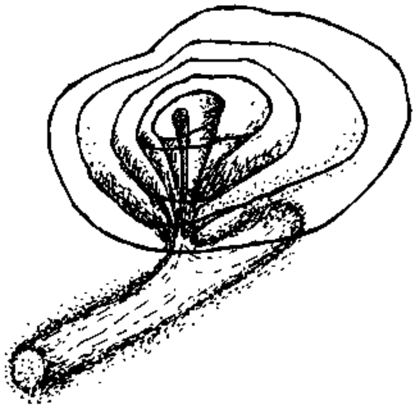

Figure 8.1. The energetic points are considered the body's small energetic pools.

THE FOUR WAYS POINTS ARE FORMED

Generally, points are formed in four ways. The first type of point occurs when two muscles, tendons, or ligaments intersect or overlap in a way that creates a small depression at the area of convergence. To form the depression, the two muscles either come together, separate, overlap, or run parallel to each other. Usually the channels lie under a protective layer of muscle, but at the aforementioned junctures, the channels may be exposed, and thus closer to the surface.

The second type of point occurs when a nerve is exposed in an area without much muscle tissue to cover and protect it. Because the nerve is exposed, such points are especially sensitive, and may cause a sharp pain or an electric shock sensation when stimulated.

The third type of point lies under a protective superficial layer of muscle. It can however, still be stimulated due to its natural sensitivity and because the overlying layer of muscle is thin.

The fourth type of point lies in depressions on the surfaces of the bones and cartilage. These depressions may take the form of grooves, pits, hollows, indentations, fissures, or crevices.

SCIENTIFIC RESEARCH OF ACUPOINTS

In the early 1970's, shortly after President Nixon's historic visit to China, Dr. Robert O. Becker, M.D., a Syracuse University orthopedist, received a grant from the National Institutes of Health to research how acupuncture works. Dr. Becker reasoned that the channels were electrical conductors (independent of the nervous system yet communicating with it) that carried an injury message to the brain, which responded by sending back the appropriate level of direct current (DC) to stimulate healing in the troubled area."

Becker also suggested in his book *The Body Electric,* that "any current grows weaker with distance, due to resistance along the transmission cable. The smaller the amperage and voltage, the faster the current dies out. Electrical engineers solve this problem by building booster amplifiers every so often along a power line to get the signal back up to strength. For currents measured in nanoamperes and microvolts (such as those generated in the human body), the amplifiers would have to be no more than a few inches apart—just like the acupuncture points."

Dr. Becker and his assistant, Maria Reichmanis, a young biophysicist, designed special equipment to measure and map the electrical conductivity along the channels. They found the predicted electrical characteristics along the channels and at half the points. These same points showed up on all the people tested. They suggested that the other traditional points may simply be weaker, or of a different kind, than the ones that their instruments detected.

Each point that they found was electrically positive compared to its surrounding tissue, and each point had an electrical field surrounding it with its own characteristic shape.

Later research (see appendices) in China, France, and the United States has shown not only that these points and channels have electrical properties but that they also emit and absorb light and sound, both of which are vital to healthy biological functions.

THE FOUR CATEGORIES OF ENERGETIC POINTS

Points are divided into four categories (Figure 8.2) and are described as follows.

1. Channel points are the Primary points that are distributed along the Twelve Primary Channels and the Conception and Governing Vessels. There are 365 channel points.
2. Extra points are not regarded as having a specific origin in the fourteen main channels, although they are named and are in definite locations. These points are also called "miscellaneous points."

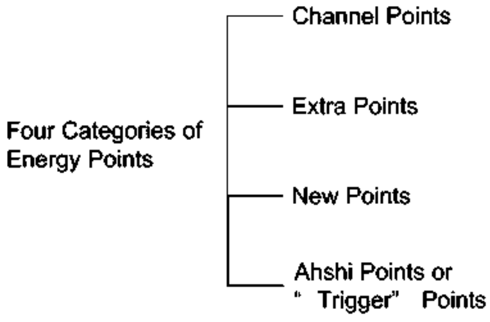
Four Categories of Energy Points
Channel Points
Extra Points
New Points
Ahshi Points or " Trigger" Points

Figure 8.2. The Four Categories of Energetic Points

3. New points are a fairly large group of points discovered since the Communist "Liberation" of 1949, in many cases through electric point detection. These points generally have no energetic description, but do have indications. Combined with the Extra points, there are more than 700.
4. Ahshi Points are tender or painful spots near a diseased or injured area and can be anywhere, on or off of traditional points and channels. They function like trigger points, as they are points of tension or pain. Ahshi points are not given specific names or definite locations as they are mostly used for pain syndromes.

THE TWO GATES OF ENERGETIC POINTS

Another way of categorizing the points is into the two general types commonly used by the Qigong doctor.

1. The Outer Qi Gates are points through which the patient's Shen and Qi enter and leave the body. The Qigong doctor or patient may gather healing Qi from the outer environment through these gates. In terms of pathology, these points are especially vulnerable to the Six External Pathogens, i.e., Cold, Hot, Wind, Damp, Dry, and Fire invasion (see Chapter 21), which may enter the body through these gates. These same pathogens may also be ex-

pelled through these gates. For this reason, many of these outer Qi gates have the word Feng (Wind) in their names, such as Fengchi GB-20 (Wind Pool).

2. The Inner Qi Gates are points through which the Qi of the Yin and Yang organs and channels is transported back and forth internally and externally, deeply and superficially. These points are more involved with the movement of Qi within the body and are less involved in the exchange with the outer environment. In this way, the Inner Qi Gates enable the Qi to communicate between the different parts of the body.

These two types of points (or gates) are specific areas where the Qigong doctors concentrate their intention and focus their energy extension. Point locations have a greater amount of accumulated Qi, and are used for dispersing or tonifying the body's organ energy, to promote the correct balanced circulation of Qi and Blood.

THE THREE FUNCTIONS OF ENERGETIC POINTS

The points can be used for three clinical functions: to manipulate Qi and Blood, for diagnosis, and to serve as pathways for the elimination of disease.

1. Qi and Blood can be manipulated to purge Excess conditions, tonify Deficient conditions, or for regulation.
    a. To disperse Excess conditions, the Qigong doctor emits Qi over a specific point (or points) and begins to purge, or disperse, the energy in the area (to purge is to remove toxins directly from the organ, whereas to disperse is to lead the toxins out through the channels). The goal of this treatment is to move any Excess, particularly stagnant Qi and Blood. If only Qi is stuck, the pain will move and feel dull. If only Blood is stuck, the pain will not move and will be sharp. This distinction, however, is more important in Chinese herbology than it is for medical Qigong, acupuncture, or massage because, as stated in the classics, "Qi is the commander of Blood"; therefore, moving the Qi in the latter therapies will also move stagnant Blood.
    b. To tonify a Deficient condition, the Qigong doctor emits Qi over a specific point or points, and begins to tonify or strengthen the patient's Qi, Blood, Yin, Yang, or specific organs and tissues. The Qigong doctor uses varying techniques according to the patient's needs (e.g., in order to prevent feeling discomfort, the doctor will tonify weak patients slowly).
    c. To regulate, the Qigong doctor emits Qi over a specific point or points, and allows the body's natural homeostasis to take over and restore healthy function. This type of treatment is also used when a patient has a combined Excess/Deficiency syndrome (such as stagnant Qi in the chest due to a Qi Deficiency in the Lungs and Kidneys).

The healing potential of point stimulation can be accomplished through various modalities, such as Energy Projection, Energetic Point Therapy, Jing Point Therapy or Invisible Needle Therapy.

2. Diagnosis through point palpation and inspection is an important diagnostic tool in all the branches of Chinese medicine. If a point hurts when touched with light pressure, is hard or swollen, or is purple, black and blue, or red, this indicates an Excess condition. If the point hurts on deep pressure or is soft to the touch, lacks resilience, or is sallow in color, it indicates an underlying Deficiency.

The skin over points often feels sticky, in contrast to the slippery or smooth quality of the skin surrounding the point. When passing the fingers over a point, the doctor generally feels the energetic pulse of the channel. When touched, the point energetically responds like an echo, vibrating along the channel into the organ and then back again to the

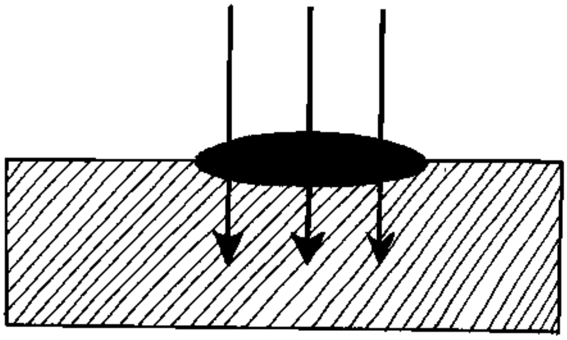

Figure 8.3. Weak or Deficient tissue will naturally pull and absorbe Qi into the surrounding areas.

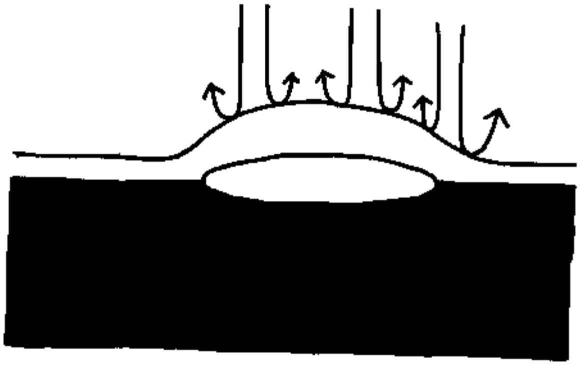

Figure 8.4. Strong, Excess, or Armored tissue will naturally repel Qi from the surrounding areas.

doctor's finger. The doctor can use this energetic response for diagnostic evaluation by sending a pulse along the flow of the channel or against it (into the organ itself). Once sufficient stimulation is applied to the point, its organ or region is "reminded" of its normal function. In this way, diagnosis and treatment are accomplished simultaneously.

Once the location and energy flow of the channel to be treated is diagnosed, the doctor must take into consideration that each patient will react differently in accordance with the severity of symptoms, as well as the individual's constitutional type. Before beginning treatment, the doctor should extend Qi between the energetic point's boundary (the space separating the Wei and Ying energy levels), until the soreness or distention is felt on the tips of the doctor's fingers or palms (which act like a barometer). This form of diagnosis determines the condition of the channel and the area needing treatment. Although the feelings and sensations will vary with each patient, energetic truths remain constant. For example:

a. Weak or Deficient areas will feel like an empty hole or deep well. Deficient Blood or Qi will actually pull the energy into the area and will feel something like a subtle vacuum suction. If the area that the doctor is treating attacks his or her Qi and begins to absorb it, the doctor should tonify that area immediately. Think of this reaction as a dry house plant starving for water. The minute the water touches its soil, the plant immediately absorbs it into its roots (Figure 8.3).
b. Strong or Excess areas will repel the doctor's Qi. If the diseased area repels the doctor's Qi, he or she must immediately dredge the Excess. First, the doctor should focus on the region surrounding the Excess to drain the Excess energy away from the damaged points, then dredge the Excess energy away from the original trauma area. This technique is excellent for damaged or sprained joints. The doctor should focus on the patient's breathing when dredging. As the patient inhales, the doctor exhales and vice-versa. As the patient exhales, he or she releases pathogenic factors; by inhaling when the patient exhales, the doctor facilitates this process, and prevents depletion of his or her own Qi (Figure 8.4).

3. Disease transmission may occur due to the External invasion of pathogenic Qi into the body through the points. Some points are particularly vulnerable to exogenous invasions, especially Wind invasion. Wind-susceptible points usually have the word Feng in their names.

POINT NAMES

Traditionally, each point has a name, which either describes its location, its energetic function, or both. In addition, modern acupuncture texts have assigned a name and number for each point, according to its channel and its order along the natural course of the channel.

Sometimes a point may have several different traditional names, varying according to whether they are being used by medical Qigong doctors, acupuncturists, martial artists, or religious practitioners; for example, GV-1 is called Changqiang (Long Strength) in Traditional Chinese Medicine, and Weilu (Tail Gate) in medical Qigong and Daoist inner alchemy.

The history of point naming and the different terminologies used can be traced back to the different Buddhist and Daoist temples in ancient China. Each religious sect had its own specific names given to certain points, in accordance with the temple's spiritual and energetic needs. Giving the points different names was a form of secret code, used to keep the system pure and to prevent the knowledge from falling into the hands of the uninitiated or unscrupulous.

Sometimes the same name may refer to different points or areas on the body; for example, Yuzhen (Jade Pillow), located in the occipital region on the posterior base of the skull, is the name commonly used for Bl-9 in Traditional Chinese Medicine. However, in medical Qigong and Daoist inner alchemy, Yuzhen instead refers to an entire area the size of one's fist placed between the bilateral Bl-9 points.

CLASSIFICATION OF ENERGETIC POINTS

The points are further classified according to their energetic potential. This classification will assist the doctor of Traditional Chinese Medicine in choosing the right point or points to treat. The following is a list of energetic point classifications.

THE FIVE ANTIQUE SHU POINTS

Like the internal organs and channels, some points have a more powerful influence on the body then others. Each point has an action on the quantity of energy of an organ. Along each of the Twelve Primary Channels lie five specific points below the elbow or the knee called Antique Shu points (also called Transporting points and Command points). These points belong to the oldest classification of points, and are described in water images, comparing each section of the channel to the course water takes as it emerges from the ground and makes its way to the ocean. They are identified as follows: Jing (Well), Ying (Spring), Shu (Stream), Jing (River), and He (Sea). These five points are situated along the flow of energy from the extremities to the torso (running from the fingers to the elbows and from the toes to the knees), occupying the same progressive location along the patient's channels (Well, Spring, Stream, River, and Sea).

Centrifugal and Centripetal Energy Flow

All of the body's channel points energetically progress in two opposite currents according to the direction and flow of the channel's energy. These two energy currents are described as follows.

1. The Channel's Outward Centrifugal Flow of Qi passes from the organ, through the channel, to the point, and is expressed or intensified through the patient's exhalation. The outward centrifugal flow of energy is manifested in channel points and in areas of the body where there are conditions of Excess.
2. The Channel's Inward Centripetal Flow of Qi passes from the point, through the channel, to the organ. The inward centripetal flow of energy is manifested in channel points and areas of the body where there are conditions of Deficiency.

Through stimulation, channel points fulfill a double function. Centrifugally, they inform the doctor of impending internal disorders by discharging Excess energy. Centripetally, they transmit the energetic action placed on them, absorbing energy to supply the body's internal insufficiencies.

Yin and Yang Channels and the Five Elements

Each of the Five Antique Shu points also corresponds to the Five Elements, in the progression of the Creative Cycle. Each Yin channel begins

Points	Yin Channel Elements (Ethereal Soul—Hun)	Yang Channel Elements (Corporeal Soul—Po)
Jing-Well Points	Wood	Metal
Ying-Spring Points	Fire	Water
Shu-Stream Points	Earth	Wood
Jing-River Points	Metal	Fire
He-Sea Points	Water	Earth

Figure 8.5. The Five Antique Shu Points and their Correspondence to the Five Elements

with the Hun progression through the Five Elements, starting with Wood, then Fire, Earth, Metal, and Water. Each Yang channel begins with the Po progression through the Five Elements, starting with Metal, then Water, Wood, Fire, and Earth (Figure 8.5).

Similar to the eyes relationship to Wood (Hun) and Metal (Po), the beginning points on the Twelve Primary Channels also have a relationship to either a Wood (Ethereal Soul) or Po (Corporeal Soul) connection. Hence, in meditation exercises, when specific finger tip points are connected, the blending of the Hun and Po energies are established. The practice of connecting specific finger tip points is sometimes referred to as the connection (or fusion) of the body's Ethereal and Corporeal Soul.

The Five Antique Shu points energetically flow from superficial to deep and are susceptible to external pathogens and climatic changes. The distal points on the feet tend to be more powerful than those on the hands. After the manipulation of the patient's distal points, the local channel points are used according to their feeling of tenderness. The Five Transporting points are explained as follows.

1. Jing-Well points are the first points along the extremities of the channels and are located at the tips of the fingers and toes. These points are where the energy of the channel pools and is at its most superficial and thinnest.

   These points are where the energy of the channels leaves the body moving centrifugally outward. At the end of the fingertips and toes, the channel's energetic polarity changes from Yin to Yang or vice versa. Due to this shift in polarity, the energy at the channel's extremities tends to be unstable, and therefore more easily influenced.

   Due to their outward movement of energy, the Well points can be used by the Qigong doctor to eliminate the patient's pathogenic factors quickly, especially when the Yin organs are affected. The Well points will also have a strong effect on the patient's mental state, quickly changing the patient's mood. These points are generally used for acute disorders, expelling exogenous pathogens, and mental disorders. They are used as "revival" points for loss of consciousness from fainting and heat stroke, convulsions, a feeling of fullness below the Heart, and local neuropathy (tingling, numbness, burning pain). In mental disorders, they are commonly used for irritability, mental restlessness, anxiety, hysteria, mania, and insomnia.
2. Ying-Spring points are the second points along the channels and, in all cases, are located in the second position of the channels' energetic progression up the arms or legs, and are located next to the Well points. These points are where the channel's energetic flow quickens its progression of Qi into the body.

At the location of the Spring points, the Qi in the channels is very dynamic and powerful. Slipping and gliding like the swirling movement of cascading water, the energy in these points can change quickly. The Spring points are generally used to eliminate Internal and External pathogenic factors (especially Heat) from the patient's body. They are also used when a disease effects a color change in the patient's complexion.

In the Yin channels, this point is always a Fire Element. Purging the Spring point in the Yin channels reduces Heat and clears Fire. In the Yang channels, the Spring point is always the Water Element point. The doctor can reduce Heat in the patient's Yang channel by tonifying this Water point.

3. Shu-Stream points are the third points along the channels and are located in the third position in the channels' energetic progression up the arms or legs, next to the Spring point (except for the Gall Bladder's Channel where it is the fourth point). These points are where the channel's energy rapidly pours through and slightly deepens its flow into the body.

   At these points, the external pathogenic factors penetrate deep into the channels and can be "transported" into the body's interior. These points are used to clear Wind and Dampness from the patient's channels or when a disease manifests intermittently.

4. Jing-River points are the fourth of the Five Transporting points; they are, however, not always located on the fourth point along the channel. The points are, however, always located between the wrist and elbow on the arms and between the ankle and knee on the legs. These points are where the channels broaden, and the energy flow continues to increase.

   At these points, the energy current flows wider, deeper, and more irregularly, like a large river, directing the invasion of external pathogenic factors towards the body's tendons, joints, and bones. These points are generally used when there is stagnation of Qi and Blood in the patient's channels, as well as for coughing, asthma, dyspnea, sore throat, and upper respiratory diseases.

5. He-Sea points are the fifth points of the Five Transporting points and, in all cases, are located at the elbows and knees. These points are where the energy of the channels is vast and plunges deep into the body.

   The Qi at the He-Sea points moves centripetally inward, flowing much slower, joining, collecting, and fusing with the general circulation of the body's energy (like a mighty river flowing into the sea). The He-Sea points are generally used for all Stomach and Intestinal diseases and for Internal disorders of the organs.

THE FIVE ELEMENTAL POINTS

Each of the Five Elemental points correspond to the Five Antique Shu points. The Five Elemental points are aligned in the progression of the Creative Cycle. Each Yin Channel begins with the Three Ethereal Souls (Hun) progression through the Five Elements, starting with Wood, then Fire, Earth, Metal, and Water. Each Yang channel begins with the Seven Corporeal Souls (Po) progression through the Five Elements, starting with Metal, then Water, Wood, Fire, and Earth.

THE SIXTEEN XI CLEFT (ACCUMULATION) POINTS

These points are used for purging acute conditions of Excess in the channels and organs. There is one point on each of the Twelve Primary Channels, plus one point on the Yin and Yang Heel Vessels, and one point on the Yin and Yang Linking Vessels. The Yin Channel Accumulation points are Lu-6, Pc-4, Ht-6, Sp-8, Lv-6, Kd-5, Yin Heel Kd-8, and Yin Linking Kd-9. The Yang Channel Accumulation points are LI-7, TB-7, SI-6, St-34, GB-36, Bl-63, Yang Heel Bl-59, and Yang Linking GB-35.

THE TWELVE FRONT MU (ALARM, COLLECTING) POINTS

The energy of the organs collect at these points; therefore, they can be used for both diagnosis and treatment. These points are palpated for diagnosis and may feel tender to the patient, either spontaneously or upon pressure. If they respond to light pressure or feel tight or swollen, the patient has an Excess condition; if the points respond to deep pressure or feel hollow or deflated the patient has a Deficient condition. The Twelve Front Alarm points are: Lungs Lu-1, Pericardium CV-17, Liver Lv-14, Spleen Lv-13, Gall Bladder GB-24, Kidneys GB-25, Heart CV-14, Stomach CV-12, Large Intestine St-25, Triple Burners CV-5, Small Intestine CV-4, and Bladder CV-3.

THE EIGHTEEN BACK SHU (ASSOCIATED, TRANSPORTING) POINTS

These points are all located on the Bladder Channel and are used for both diagnosis and treatment. These points transport Qi very effectively and directly to the internal organs. They can be used for treating the corresponding sense organs. The Eighteen Back Transporting points are: Lungs Bl-13, Pericardium Bl-14, Heart Bl-15, Governing Vessel Bl-16, Diaphragm Bl-17, Liver Bl-18, Gall Bladder Bl-19, Spleen Bl-20, Stomach Bl-21, Triple Burners Bl-22, Kidneys Bl-23, Sea of Qi Bl-24, Large Intestine Bl-25, Gate to the Yuan Qi (Guan Yuan) Bl-26, Small Intestine Bl-27, Bladder Bl-28, Sacrum Bl-29, and Anus Bl-30.

THE FIFTEEN LUO (CONNECTING) POINTS

One of these points is located on each of the Twelve Primary Channels, as well as the Governing and Conception Vessels and the Great Luo point of the Spleen. These points are used to treat channel pathology. The Fifteen Collateral points are divided into Yin and Yang channel points. Yin Channel points include: Lu-7, Pc-6, Ht-5, Sp-4, Lv-5, Kd-4, CV-15, and the Great Luo Channel Sp-21. The Yang Channel points include LI-6, TB-5, SI-7, St-40, GB-37, Bl-58, and GV-1.

THE TWELVE ENTRY POINTS

These points are where the energy from a Primary Channel enters its coupled channel in the Horary Cycle. The entry point is the first point on each channel in all cases, except for the Large Intestine's Channel in which it is LI-4.

THE TWELVE EXIT POINTS

These points are where the Qi leaves the Primary Channel to flow into its coupled channel in the Horary Cycle. The Yin Channel Exit points include: Lu-7, Pc-8, Ht-9, Sp-21, Lv-14, and Kd-

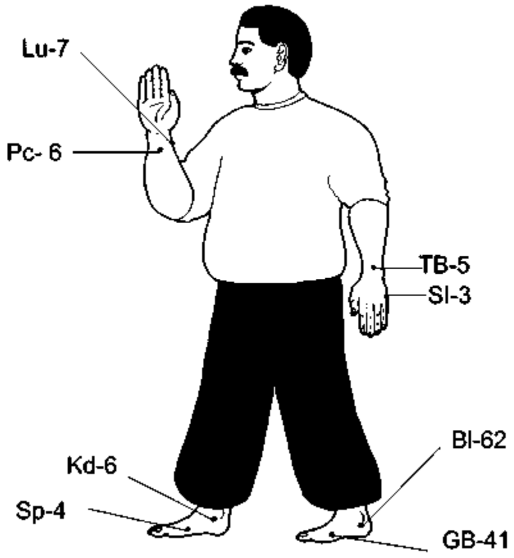

Figure 8.6. Locations of the Body's Master and Couple Points

Channel	Open with Master Point	Complete with Couple Point
Governing	Hand (SI-3)	Foot (Bl-62)
Conception	Hand (Lu-7)	Foot (Kd-6)
Thrusting	Foot (Sp-4)	Hand (Pc-6)
Belt	Hand (TB-5)	Foot (GB-41)
Yin Linking	Hand (Pc-6)	Foot (Sp-4)
Yang Linking	Foot (GB-41)	Hand (TB-5)
Yin Heel	Foot (Kd-6)	Hand (Lu-7)
Yang Heel	Foot (Bl-62)	Hand (SI-3)

Figure 8.7. The Vessels with their Associated Master Points and Couple Points

22. The Yang Channel Exit points include: LI-20, TB-23, SI-19, St-42, GB-41, and Bl-67.

THE EIGHT INFLUENTIAL POINTS

These points have specific effects on their respective organs, substances, and tissues. They affect the body's Qi, Blood, sinews, blood vessels, bones, and Marrow, as well as the Yin and Yang organs. The Eight Influential points are as follows: Yin Organs Lv-13, Yang Organs CV-12, Qi CV-17, Blood Bl-17, sinews GB-34, blood vessels Lu-9, Bones Bl-11, and Marrow GB-39.

THE EIGHT CONFLUENT POINTS

These eight points communicate with the Eight Extraordinary Vessels. The Confluent points are divided into Master and Coupled points. The Master point is the primary point chosen for treatment, its Coupled point is the connecting or secondary point used to open the specific Vessel. Each of the Eight Extraordinary Vessels has both a Master and Coupled Confluent point on the upper or lower limbs.

All of the Eight Extraordinary Vessels can be treated in various combinations according to their specific Master and Coupled point locations (Figure 8.6). If, for example, the Qigong doctor causes Qi to flow from the patient's Master point on the hand (SI-3) to its Coupled point (Bl-62) on the foot, the Governing Vessel will open. To complete the treatment, the doctor must first remove stimulation from the Coupled point then disconnect from the Master point (Figure 8.7).

The Eight Confluent points are as follows: Governing Vessel SI-3, Conception Vessel Lu-7, Belt Vessel GB-41, Thrusting Vessels Sp-4, Yang Heel Vessel Bl-62, Yin Heel Vessel Kd-6, Yang Linking Vessel TB-5, and Yin Linking Vessel Pc-6.

THE TWELVE YUAN-SOURCE POINTS

Each of the Twelve Primary Channels has a Yuan-Source point, where the body's Original Qi surfaces and pools. On the Yin Channels, the Yuan-Source points are always the Shu-Stream points. On the Yang Channels, however, the Yuan-Source points are separate points located between the Shu-Stream and the Jing-River points. These points are usually the fourth point from the distal end of the channel, except in the case of the Gall Bladder Channel, where the Yuan-Source point is the fifth point.

The Yuan Qi originates and resides in the Lower Dantian area and is dispersed to the Yin and Yang organs and then to the limbs by the Triple Burners.

The Yuan-Source points can be used clinically in diagnosis for the Yin organs. The Qigong doctor examines the points by both palpation and visual diagnosis.

The Yuan-Source points are the primary points on the Yin channel for tonifying and regulating their respective Yin organs. On the Yang Channels, the Yuan-Source points are quite different in terms of their importance and energetics. These points generally have a negligible tonifying effect on their related Yang organs and comparatively little Qi regulating effect. Their main functions are to expel various pathogenic factors, and to treat disorders along their channel pathways. The Twelve Yuan points are as follows: Gall Bladder GB-40, Liver Lv-3, Lungs Lu-9, Large Intestine LI-4, Stomach St-42, Spleen Sp-3, Heart Ht-7, Small Intestine SI-4, Bladder Bl-64, Pericardium Pc-7, and Triple Burners TB-4.

THE POINTS OF THE FOUR SEAS

The human body has four seas: the Sea of Qi, Sea of Blood, Sea of Marrow, and the Sea of Nourishment (literally Water and Grain). To access these seas, the Qigong doctor stimulates the following points:

1. The Sea of Qi points include: Bl-10, St-9, CV-17, GV-14 and GV-15. When the Sea of Qi is in Excess, symptoms include fullness in the chest, flushed complexion, and dyspnea. When the Sea of Qi is Deficient, symptoms include fatigue or low energy and an inability to speak.
2. The Sea of Blood points include: Bl-11, St-37, and St-39. When the Sea of Blood is in Excess, symptoms include anxiety, uneasiness, and unrest, and the entire body feels big. When the Sea of Blood is Deficient, symptoms include the body feeling too small for no apparent reason.
3. The Sea of Marrow points include: GV-15, GV-

16, GV-17, and GV-20. When the Sea of Marrow is in Excess, symptoms include sensations relating to that of Excess energy (lightness, feelings of strength and vitality). When the Sea of Marrow is Deficient, symptoms include fatigue, vertigo, dizziness, tinnitus, pain in the lower legs, impaired vision, and a desire to sleep.

4. The Sea of Nourishment points include: St-30 and St-36. When the Sea of Nourishment is in Excess, the abdomen feels distended. When the Sea of Nourishment is Deficient, symptoms include hunger with an inability to eat.

UPPER (UNITING) HE-SEA POINTS

There are three points on the upper arm that have a powerful effect on the Yang organs. The affected organs and their points include the Stomach LI-10, Large Intestine LI-9, and the Small Intestine LI-8.

LOWER (UNITING) HE-SEA POINTS

Each of the Yang organs has a uniting He-Sea point on the leg, which can be used to treat Yang organ problems. The three Yang channels of the feet also correspond to the three Yang channels of the hands. The Lower Six Lower He-Sea points are as follows: Stomach St-36, Large Intestine St-37, Small Intestine St-39, Triple Burners Bl-39, Bladder Bl-40, and Gall Bladder GB-34.

THE ELEVEN GHOST POINTS

In ancient times, these eleven points were traditionally used for spirit possession and have a specific effect on the mind's psychological functions. In traditional Chinese medical clinics today, these points are commonly used to treat emotional disorders. The Eleven Ghost points are as follows: GV-26, Lu-11, Sp-1, Pc-7, Bl-62, GV-16, St-6, CV-24, Pc-8, GV-23, and LI-11.

THE TEN WINDOW OF THE SKY POINTS

These ten points also have a specific effects on the mind's psychological functions and are also commonly used to treat emotional disorders. The Ten Window of the Sky points are as follows: Lu-3, LI-18, TB-16, Bl-10, St-9, CV-22, GV-16, SI-16, SI-17, and Pc-1.

SUMMARY OF POINTS

Understanding the origin and function of the body's energetic points, as well as the Qi flow along the energetic channels, allows the Qigong doctor to emit energy through these energetic cavities to easily access the tissues and internal organs of the patients body. It is through this understanding that the doctor can also lead and purge the Toxic or pathogenic Qi (which has become stagnant or detrimentally active within the patient's body) away from the diseased organs, through the channels and out the body.

The following graph (Figure 8.8) is a basic categorization of the body's points and their descriptions.

Category	Points	Description
Five Antique Shu Points	Jing-Well, Ying-Spring, Shu-Stream, Jing-River, He-Sea	Transporting and Command Points
Sixteen Xi Cleft Points	Yin: Lu-6, Pc-4: Ht-6, Sp-8, Lv-6, Kd-5, Yin Heel Kd-8, and Yin Linking Kd-9 Yang: LI-7, TB-7, SI-6, St-34, GB-36, Bl-63, Yang Heel Bl-59 and Yang Linking GB-35	Accumulation Points
Twelve Front Mu Points	Lungs Lu-1, Pericardium CV-17, Liver Lv-14, Spleen Lv-13, Gall Bladder GB-24, Kidneys GB-25, Heart CV-14, Stomach CV-12, Large Intestine St-25, Triple Burners CV-5, Small Intestine CV-4, and Bladder CV-3	Alarm, Collecting Points
Eighteen Back Shu Points	Lungs Bl-13, Pericardium Bl-14, Heart Bl-15, Governing Vessel Bl-16, Diaphragm Bl-17, Liver Bl-18, Gall Bladder Bl-19, Spleen Bl-20, Stomach Bl-21, Triple Burners Bl-22, Kidney Bl-23, Sea of Qi Bl-24, Large Intestine Bl-25, Gate to the Original Qi (Guan Yuan) Bl-26, Small Intestine Bl-27, Bladder Bl-28, Sacrum Bl-29 and Anus Bl.30	Associated, Transporting Points
Fifteen Luo Points	Yin: Lu-7, Pc-6, Ht-5, Sp-4, Lv-5, Kd-4, CV-15, the Great Luo Channel Sp-21 Yang: LI-6, TB-5, SI-7, St-40, GB-37, Bl-58 and GV-1	Connecting Points
Twelve Entry Points	The first point on each channel except for the Large Intestine Channel in which it is LI.4	Qi enters Channel
Twelve Exit Points	Yin: Lu-7, P-8, Ht-9, Sp-21, Lv-14 and Kd-22 Yang: LI-20, TB-23, SI-19, St-42, GB-41 and Bl-67	Qi exits Channel
Eight Influential Points	Yin Organs Lv-13, Yang Organs CV-12, Qi CV-17, Blood Bl-17, Sinews GB-34, Blood Vessels Lu-9, Bones Bl-11 and Marrow GB-39	Influence Internal Organ Function
Eight Confluent Points	Governing Vessel SI-3, Conception Vessel Lu-7, Belt Vessel GB-41, Thrusting Vessels Sp-4, Yang Heel Vessels Bl-62, Yin Heel Vessels Kd-6, Yang Linking Vessels TB-5 and Yin Linking Vessels Pc-6	Master and Couple Points
Twelve Yuan-Source Points	Gall Bladder GB-40, Liver Lv-3, Lungs Lu-9, Large Intestine LI-4, Stomach St-42, Spleen- Sp-3, Heart Ht-7, Small Intestine SI-4, Bladder Bl-64, Pericardium Pc-7, and Triple Burners TB-4	Source of Original Qi Points
Four Seas Points	Sea of Qi Points: Bl-10, St-9, CV-17, GV-14 and 15 Sea of Blood Points: Bl-11, St-37 and 39 Sea of Marrow Points: GV-15, 16, 17, and 20 Sea of Nourishment Points: St-30 and 36	Give Access to the Four Seas
Upper He-Sea Points	Stomach LI-10, Large Intestine LI-9 and the Small Intestine LI-8	Affect the Yang Organs
Lower He-Sea Points	Stomach St-36, Large Intestine St-37, Small Intestine St-39, Triple Burners Bl-39, Bladder Bl-40, and Gall Bladder GB-34	Affect the Yang Organs
Eleven Ghost Points	Governing Vessel 26, Lungs 11, Spleen 1, Pericardium 7, Bladder 62, Governing Vessel 16, Stomach 6, Conception Vessel 24, Pericardium 8, Governing Vessel 23, and Large Intestine 11	Affect the Mind, Emotion, and Spirit
Ten Windows of the Sky Points	Lung 3, Large Intestine 18, Triple Burners 16, Bladder 10, Stomach 9, Conception Vessel 22, Governing Vessel 16, Small Intestine 16, Small Intestine 17, and Pericardium 1	Affect the Mind, Emotion, and Spirit

Figure 8.8. Point Graph

CHAPTER 9
THE EXTRAORDINARY YANG ORGANS

Besides the six Yin organs (Liver, Heart, Pericardium, Spleen, Lungs, and Kidneys) and six Yang organs (Gall Bladder, Small Intestine, Triple Burners, Stomach, Large Intestine, and Bladder), the body has six Extraordinary Yang organs, also called the "curious organs." These Extraordinary organs function like Yin organs (storing Yin Essence, i.e., Blood, Marrow, or Kidney Essence) but look like Yang organs (they are hollow). These six curious organs include the Uterus, Brain, Marrow, Bones, Blood Vessels, and Gall Bladder. The body's Eight Extraordinary Vessels integrate the Extraordinary Yang organs with the main six Yang organs and the Kidneys.

THE UTERUS

In Chinese medicine, the uterus encompasses the fallopian tubes and the ovaries and has the function of regulating menstruation, conception, and pregnancy (Figure 9.1). The uterus connects to the Kidneys (which provide the uterus with Jing), the Conception Vessel (which provides the uterus with Qi and nourishes the fetus), and the Thrusting Channel (which provides the uterus with Blood). Only when the Jing of the Kidneys is sufficient can the menstrual period occur regularly, the woman become pregnant, and fetal growth become possible. The Qi and Blood of the Twelve Primary Channels pass into the uterus through the Thrusting and Conception Vessels, affecting the amount of menstrual flow and its cycle.

The three Yin organs of the Heart, Liver, and Spleen also energetically connect to the uterus through their relationships with the Blood. The Heart governs the Blood. The Liver stores the Blood and regulates the volume of circulating Blood, which is responsible for normal menstruation. The Spleen controls the Blood.

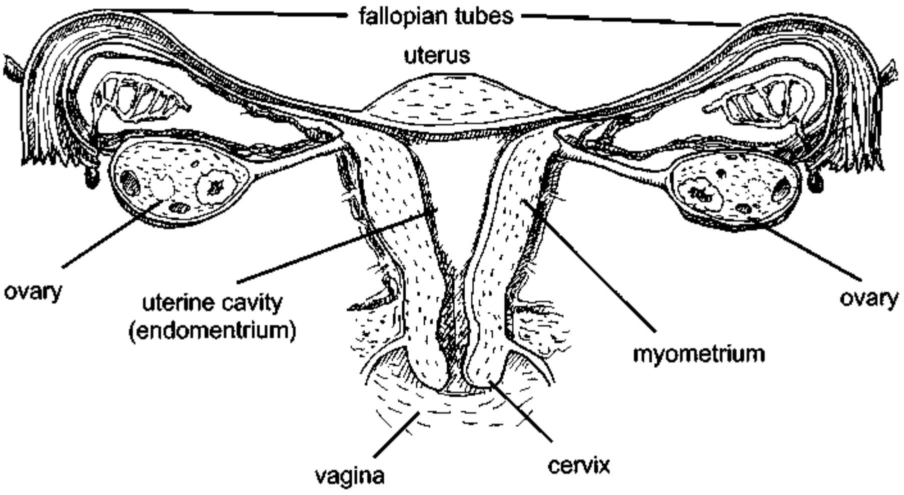

Figure 9.1. The Female Reproductive Organs, including the Uterus, Fallopian Tubes, and Ovaries

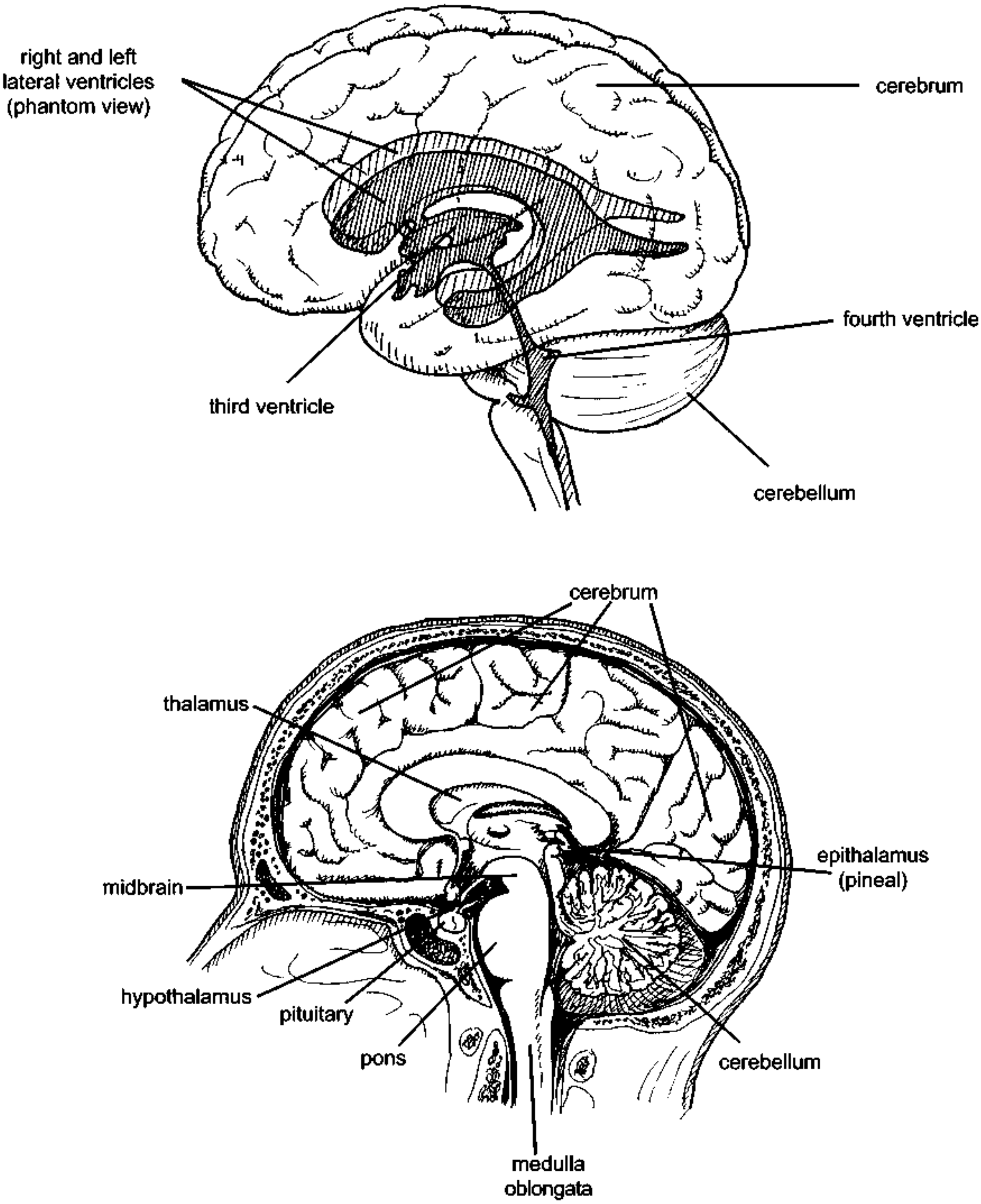
right and left lateral ventricles (phantom view)
cerebrum
fourth ventricle
third ventricle
cerebellum
cerebrum
thalamus
epithalamus (pineal)
midbrain
hypothalamus
pituitary
pons
cerebellum
medulla oblongata

Figure 9.2. The Human Brain

THE BRAIN

The Brain is considered the house of the Yuan Shen, and the seat of mental function. It is believed that the senses and control of the body's physical movements are related to the Brain. The Brain is functionally related to the Kidneys and controls memory, concentration, hearing, touch, sight, and smell. The Brain is also called the Sea of Marrow, extending from the top of the head to the GV-16 Wind Palace point, just below the external occipital protuberance. The Kidney Essence produces Marrow, which fills the spinal cord and Brain. The Brain also depends on the Heart's Blood for its nourishment. Because the Kidneys store the Jing and the Heart governs the Blood, the Brain depends on the balanced state of the Heart and Kidneys for vitality. The Brain is considered a transformation chamber of both Prenatal and Postnatal Essence (Jing) and Energy (Qi). The Brain is regulated by the Governing Vessel, as well as the Yin and Yang Heel Vessels.

This viewpoint is quite different from the Western approach, which views the brain's anatomy as consisting of five regions (Figure 9.2), including:

1. The cerebral hemispheres or cerebrum (the neocortex (neopallium),
2. The thalamus, hypothalamus, and epithalamus (pineal),
3. The midbrain/mesencephalon (the colliculi and cerebral peduncles),
4. The pons and cerebellum, and
5. The medullae oblongata.

In Western medicine, the oldest part of the Brain (situated at the top of the Brain stem) is often called the "reptilian Brain" because its anatomical formation is similar to that found in lizards, alligators, and turtles. Surrounding it is the mammalian Brain, which is called the limbic system. The third part of the Brain is called the neocortex, which wraps itself around the limbic system. All three parts of the Brain (reptilian, mammalian, and neocortex) are viewed as biologically distinct, both in their chemistry and in their structure. The older formations of the Brain are responsible for the autonomic nervous system, whereas the neocortex is responsible for thinking and voluntary movement.

In Chinese energetic medicine, each selective set of the human brain cells absorbs and records sets of vibrations. The Brain receives these vibrations through the skin, eyes, ears, nose, and mouth, and records them in its cells. The mind interprets these incoming vibrations as images, sounds, smells, and sensations, and organizes these into memories. The extent to which the mind interprets the incoming data depends upon the level of consciousness and awareness of the individual at the time of the recording. When a Qigong doctor, attunes his or her vibrations to the patient's energetic field and emits Qi, the brain cells responsible for storing the vibrations are stimulated. When these vibrations are stimulated through the energy projected by a Qigong doctor, the patient's tissues respond and reproduce an energetic pattern similar to the original one used to record the patient's physical, mental, emotional, energetic and spiritual experiences. This energetic tissue reaction occurs only because the patient's cells are held directly to their same consistent vibrational patterning.

One set of selective brain cells receives, records, and maintains the vibrations of the thoughts, actions, movements, and pictures that other forms project. These vibrations can also be reproduced and projected, being received by the Qigong doctor as images from within his or her own mind through inner-vision. The Qigong doctor can arrange his or her cells' energetic pattern in such a way as to reproduce the words and motions of these forms or objects, even the thoughts of those that sent them. Through these cells, the Qigong doctor can assist his or her patient, as well as him or herself, in controlling their thoughts, as it is through these cells that accidents and disease are brought into existence; for example, someone either observes or imagines a specific thing happening. The corresponding vibration is fixed in the cells, sent out to be imprinted on the corresponding cells of anothers' Brain, then projected back until the incident is so fixed that the occurrence actually happens (self-fulfilling prophecy).

Another set of selective brain cells receives, records, and maintains the vibrations of the thoughts and activities of the divine. The divine energy vibration pervades every substance and is always sending out divine knowledge and true wisdom. Stored within the body are the cells that receive and project the vibrations of the divine Mind. If the doctor separates him or herself from the divine in thought, he or she will also separate from the divine in manifestation.

THE GALL BLADDER

The Gall Bladder stores the bile, which is considered a "pure" fluid. The function of the Gall Bladder is to aid digestion, as well as drain off Excess Heat from the Liver (Figure 9.3). The Gall Bladder is regulated by the energy of the Belt Vessel.

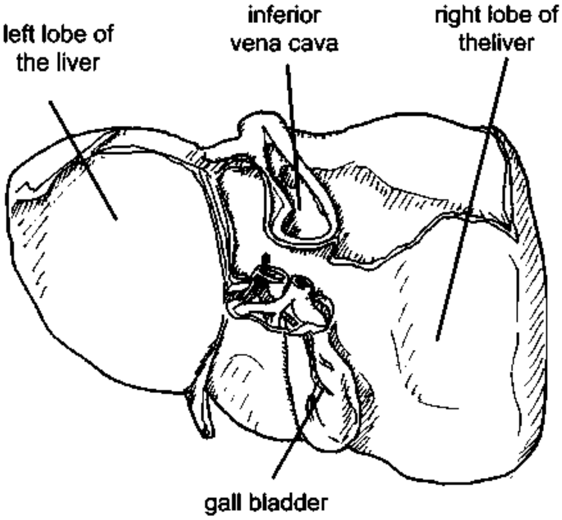

Figure 9.3. The Gall Bladder

THE MARROW

In Chinese medicine the term Marrow (Sui) is different from bone marrow as defined in Western medicine. The Marrow is the substance which is the common matrix of bones, bone marrow, the spinal cord and the Brain. The Marrow is rooted in the Jing and is connected to the Lower Dantian through the Governing Vessel. It is regulated by the energy of the body's Thrusting Vessels.

The Kidney Jing is the origin of the Marrow. The Marrow functions to form the bone marrow, as well as nourish the Brain and spinal cord. When the Mingmen warms and nourishes the body, the Marrow becomes full (which sustains the body's vertical posture and gives strength to both the Brain and the bones). When the Marrow is full, thinking is clear. Too much thinking leads to Excess Fire, which burns the Brain, causing dizziness, blurred vision, and tinnitus.

Both red and white blood cells are produced in the Marrow of the bones. The red blood cells (which circulate oxygen and eliminate carbon dioxide) are produced within the body's long bones (humerus, femur, tibia, etc.). The white blood cells (vital to the body's immune system) are produced within the body's flat bones (skull, sternum, scapulae, pelvis, etc.) The type of blood cells that are produced depends on the type of Marrow that predominates.

When all of the Postnatal energy acquired from the combination of physical (diet, food, and drink), mental/emotional (thoughts and emotions), and spiritual (prayer, meditation, and sex) cultivations is circulated through the Microcosmic Orbit (Fire Cycle), the Prenatal Qi stored in the bone marrow and Brain become stimulated, awaking the spiritual consciousness. This combination and circulation of physical, mental, emotional, and spiritual energy is why the brain marrow is considered to be constructed from the finest and most subtle essences, manifesting the original and hidden power of the Kidneys.

THE BONES

The bones are related to the Kidney's Jing. The bones support the body's structure, strength, and mobility, as well as store the bone marrow (Figure 9.4). The bones are extremely porous and are always "breathing," they are regulated by the energy of both the body's Conception and Thrusting Vessels. The pores of the bones allow Qi and Blood to be absorbed and released, the same way a sponge absorbs and releases water.

There is a natural union between the energetic function of the Marrow and the bones. The Marrow, held within the bones, assures the power and

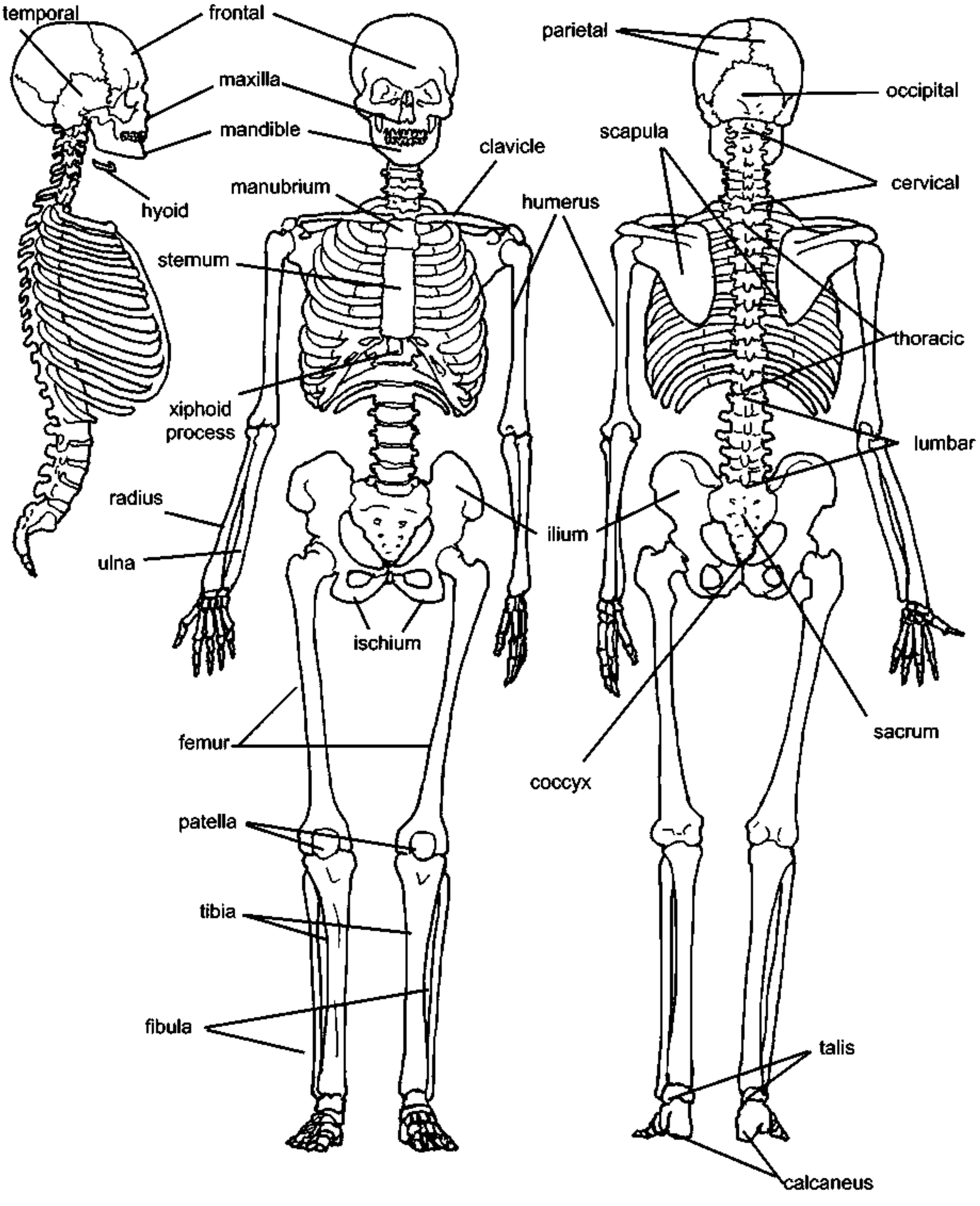
temporal
frontal
maxilla
mandible
hyoid
manubrium
sternum
xiphoid process
radius
ulna
clavicle
humerus
parietal
occipital
scapula
cervical
thoracic
lumbar
ilium
ischium
femur
patella
tibia
fibula
coccyx
sacrum
talis
calcaneus

Figure 9.4. The human skeleton consists of 200 bones.

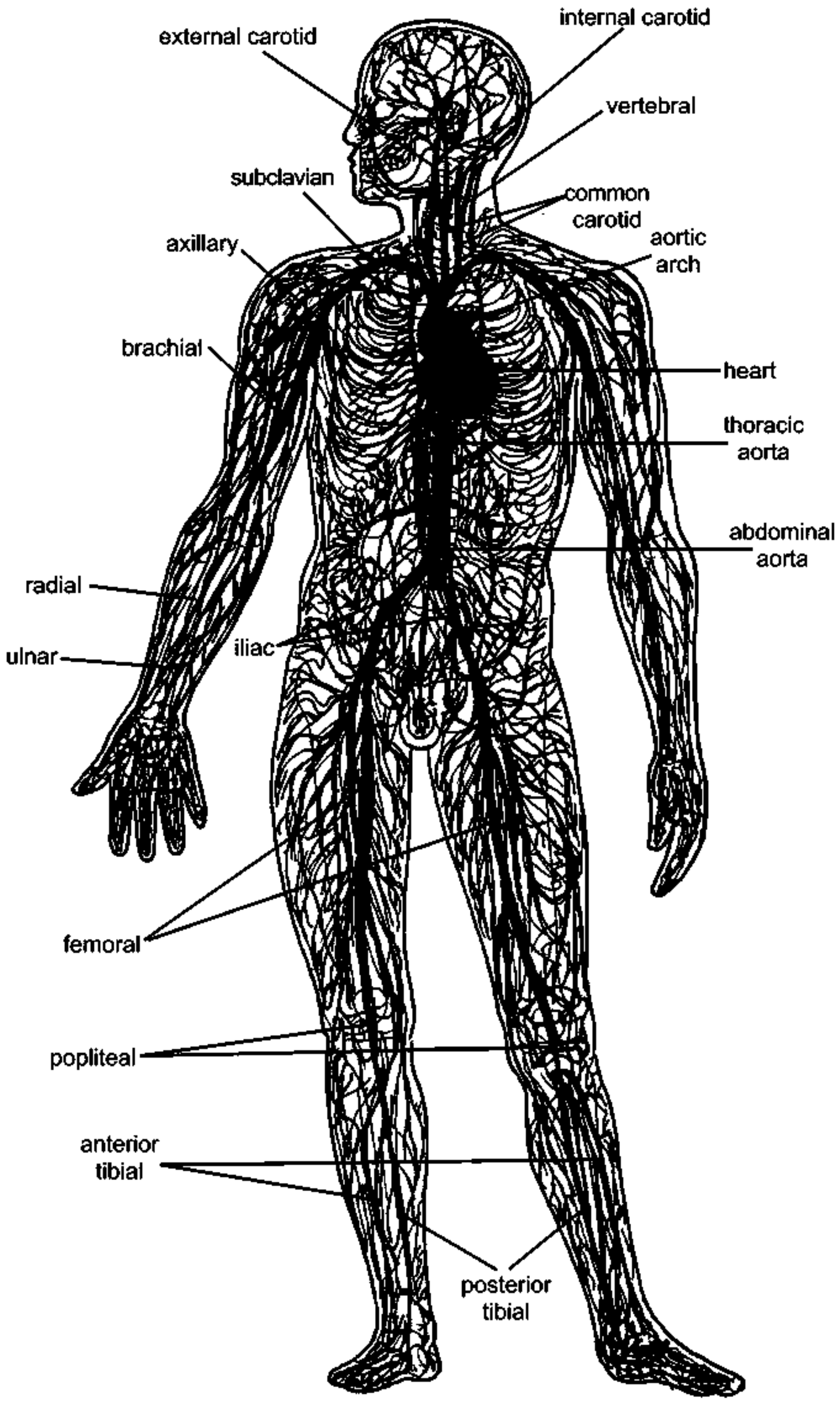

internal carotid
external carotid
vertebral
subclavian
common carotid
aortic arch
axillary
brachial
heart
thoracic aorta
abdominal aorta
radial
iliac
ulnar
femoral
popliteal
anterior tibial
posterior tibial

Figure 9.5. The blood-vascular system of the human body consists of the heart and blood vessels.

smooth suppleness of the bones. The bones, in return, prevent the dissipation of the essential richness of the Marrow.

The bones are the only solid crystalline substance in the body capable of generating piezoelectric charges. These electromagnetic charges generate fields of energy that receive and send impulses to the blood cells, tissues, organs, and channels. The Brain, nervous system, Heart, and lower abdomen also generate electromagnetic fields that resonate with the bones, as well as other crystalline-like structures. The crystalline structures of the bones amplify, radiate, and transmit energy, as well as bio-information, to the rest of the body. The rhythmic oscillation of bone-generated electromagnetic fields is released through the transmission of the bones acting as tuning forks. These living tuning forks vibrate the Qi throughout the body's entire physical structure, extending outward into the body's external Wei Qi fields.

THE BLOOD VESSELS

In Chinese medicine, the concept of Blood is different in characteristic and function from that of Western medicine. Blood originates from the transformation of food and drink by the Spleen, which then transfers the refined food energy to be further enhanced by the Heart Qi, Lung Qi, and the Qi of the air.

Qi and Blood flow together, with Qi being the active force that makes the Blood circulate and keeps it within the blood vessels (Figure 9.5). Both Qi and Blood flow within the channels and blood vessels, continuously circulating throughout the body to nourish, maintain, and moisten the tissues. Qi is an energetic form and is considered a Yang substance, while Blood is a liquid form of energy and is considered a Yin substance.

The blood vessels contain Blood and are indirectly related to the Kidneys, because the Kidney Essence (Jing) produces Marrow, which contributes to the production of Blood. The function of the blood vessels is to transport Qi and Blood throughout the body for nutrition and regeneration. The blood vessels are regulated by the energy of the body's Thrusting Vessels, and the Blood is contained in the blood vessels by the Spleen.

SECTION II

THE THREE OUTER FORCES: HEAVEN, EARTH & MAN

CHAPTER 10
INTRODUCTION TO THE THREE OUTER FORCES

THE NATURAL POWER OF THE THREE OUTER FORCES

The Three Outer Forces (Powers) of Heaven, Earth, and Man represent the sum total of all natural forces and energetic factors, which affect the human body on all levels of existence (physical, mental, emotional, energetic, and spiritual). In ancient times, the relationship between the waxing and waning of Heavenly (universal) and Earthly (environmental) energy, as well as the corresponding shifts, changes, and adjustments occurring within the human body, provided the Chinese sages with the theory of the Three Outer Forces. These ancient Chinese masters discovered that the human body's internal network of systems is directly affected by the influences and variations of energetic cycles of the sun, moon, and stars. We now know that these natural influences also affect atmospheric pressures, causing high and low temperatures, which in turn, affect the body's tissues. Recently, Ukrainian physicists found that human DNA vibrates at the same frequency as solar energy.

The Heaven and Earth Powers are responsible for the energetic field changes that cause adjustments and shifts within man's Jing, Qi and Shen (known as the Three Treasures), stored within the internal organs and tissues. Together the combination of Heavenly and Earthly influences affect the functional status of the body's internal organs and the distribution of Qi and Blood throughout the body.

Each of the body's cells, tissues, organs, and organ systems emanates its own specific electromagnetic energy field. This energetic field pulsates at its own particular frequency, regulating the body's internal energy currents and radiating an energetic field that extends outside the body. The body's energetic field (Wei Qi) interacts with and is influenced by the larger energetic fields surrounding it (i.e., power lines, crowds of people, geological formations in the Earth, and the cosmic fields of the solar system).

THE THREE OUTER FORCE'S NINE TREASURES

Each of the Three Outer Forces manifest through a secondary triad of Treasures. These Three Treasures complete the energetic matrix that encompasses the Outer Forces unified energetic field.

1. The Three Treasures of the Heavenly Power are the energy fields of the sun, moon, and stars;
2. The Three Treasures of the Earthly Power represent the energy of the Earth (soil and rock), wind, and water;
3. The Three Treasures of Man are contained within man's Essence (Jing), Energy (Qi), and Spirit (Shen).

These three intricate networks exchange their energetic influence and affect the functional status, as well as the distribution of the body's Qi and Blood.

THE THREE OUTER FORCES AND CHINESE ALCHEMY

Each of the Three Outer Forces and their Three Treasures also constitute the basis of Chinese internal alchemy, whereby Jing (Essence) is transformed into Qi (Energy), Qi is transformed into Shen (Spirit), and Shen is transformed into Wuji (infinite space), and then released back to the Dao (or divine).

The Powers of Heaven and Earth are not only interlinked, but also exist as a microcosm or "small universe" within man's physical body. The human body is an active interaction of two basic elements of Yin and Yang polarity. Yang energy ascends and is expressed through expansive, active, hard, and aggressive energetic movements. Yin energy descends and is expressed through contractive, passive, soft, and yielding movements. All matter, from the smallest molecular particles to enormous planets and stars, is composed of energy and bound into various patterns of vibration. These matter formations are held in place by electromagnetic and nuclear forces interconnected by the power of Yin and Yang polarity, and are expressed through the Three Outer Forces. The seasonal cycles are an example of Yin Earthly movements, while the solar and lunar cycles are expressions of Yang Heavenly energy.

The creation and manifestation of matter is composed and regulated by what are known as the Five Elemental patterns. In traditional Chinese understanding, it is believed that perfect health and longevity depend largely on man's ability to harmonize with the transitions and five seasonal changes of Heaven and Earth, in accordance with the Five Elemental patterns. In the winter, for example, there is less sunlight, and the temperature can be very cold (Yin), and the Kidneys (Water Element) can easily be overworked; therefore, man should endeavor to lighten his work load and rest more in winter, conserving his Qi to avoid energy depletion.

The principles of Chinese cosmology, Yin and Yang Theory, Five Elements Theory, the *Book of Changes (Yi-Jing),* and Traditional Chinese Medicine are all incorporated into the study of Medical Qigong. This culmination of knowledge reveals the basic structures of energetic dynamics relating to the universal and environmental energy gathered from the Heavens and Earth, and their relationship with man.

HEAVENLY QI

The first of the Three Outer Forces is known as the Universal Force, Original Force, or Heavenly Energy. This force manifests as the energy of the entire Cosmos. Each life born into human form is not only developed and structured but is also influenced by its own unique group of stars within the Heavenly realm. This particular group of stars (including the sun, moon, planets, and constellations), controls or influences the persons life-force energy (Figure 10.1).

Heaven is the realm that predetermines the energetic cycles of birth and death, as well as the seasons and times of planting, fertility, and harvest. It is the source of the Yuan Shen and the Virtue of Dao, as it manifests through universal awareness (divine inspiration, enlightenment and spiritual communion), and free will. Heaven is the residence of spiritual deities and the Jade Emperor (or Divine Creator), and it manifests its prenatal power in man's life through the mysterious forces of fate, destiny, karma, and reincarnation.

According to the ancient Chinese, we exist in a three-dimensional world that exists within the relative parameters of time and space. Heaven represents time, which is symbolized by the circle. Because time moves in circles, hours and days are defined by the circular rotation of the Earth on its axis; months are defined by the circling of the moon around the Earth, and years are defined by the Earth's orbit around the sun.

PHILOSOPHICAL CORRELATION

Man is constantly absorbing energy from the sun, moon, planets, and constellations. In order to explore the patterns of Heavenly energy in detail, the ancient masters developed the science of Chinese astrology, which includes the study of Twelve Animals and Five Elements, as well as the interactions of the Twenty Eight Constellations, Twelve Earthly Branches, Ten Heavenly Stems, Five Planets, and Lunar and Solar Energy Cycles.

The ancient sciences of astrology and divination evolved as a means to analyze, interpret, and predict how the various forces of Heaven influence and guide man's life on Earth. The study of Chinese astrology involves four elements of comprehension: the study of the Signs, the study of the Five Elements, the study of Yin and Yang, and the study of the Sixty-Year Cycle.

1. The Study of the Signs involves the twelve animals of the Chinese zodiac (meaning circus of animals) that encircle the Earth and are divided into twelve equal parts, called "signs." Each sign of the zodiac has certain

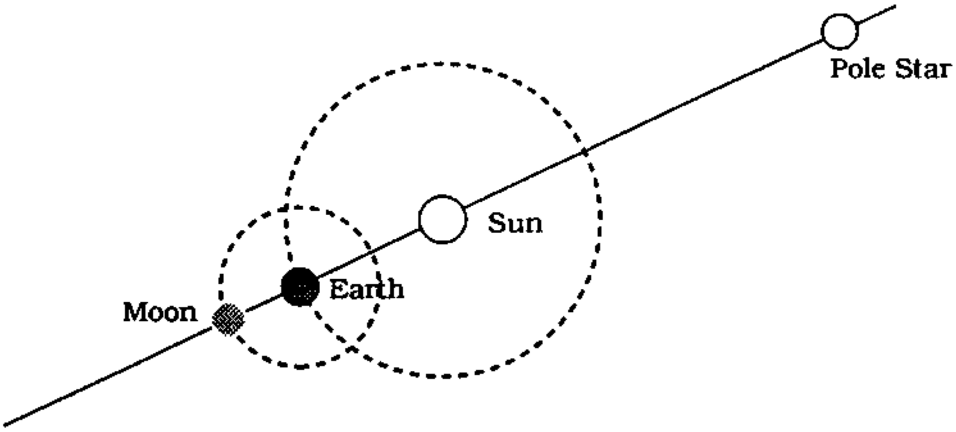

Figure 10.1. The Three Outer Forces of Heaven Consists of the Energy Released from the Sun, Moon and Stars.

characteristics that are determined by a particular planet and constellations. The signs affect the character of a person born under them. Each person is given three animal characteristics: one according to the year, the second according to the month, and the third according to the hour of his or her birth. You might, for example, be a Snake according to the year of your birth, a Boar according to the month, and a Dragon according to the hour. The animal governing your birth year represents your external self (the part you show others) and is said to exert a profound influence over your life. The month animal (known as the lunar animal) represents the emotional and psychological aspects of man's relationships. The hour animal represents the inner self (the part hidden from the rest of the world). The combination of the three determines the uniqueness of each individual.

2. The Study of the Five Elements is the next division involved in the study of signs. Each of the twelve animal signs is further differentiated into one of the Five Elements (Wood, Fire, Earth, Metal, and Water). The year of birth determines the element of the sign. Each element rules for a two-year cycle (e.g., January 27, 1952-February 13, 1953 was the year of the Dragon under the element of Water; February 14, 1953-February 2, 1954, was the year of the Snake under the element of Water. The next Water cycle of Dragon and Snake will be in 2012, based on the sixty-year cycle). These Five Elements display an energetic force that affects the formation of the internal organs in the growing fetus, as well as influences the physical, mental, emotional, energetic, and spiritual changes after the child is born.
3. The Study of Yin and Yang is the third division of the study of signs. Each of the twelve animal signs is also governed by a Yin (passive) or Yang (active) direction. The Yin and Yang cycles alternate each year, with Yang being the even years, and Yin the odd years.
4. The Study of the Sixty-Year Cycle is the final division of the study of signs. The yearly correspondence of the twelve animals of the Chinese zodiac also relates to the Twelve Earthly Branches, and is based on a sixty-year cycle. Every sixty years, the sun, moon, Earth, and the North (Pole) Star are in alignment, and a new sixty-year cycle begins. A sixty-year cycle also completes the rotation of the twelve animals through each of the Five Elements. The Chinese consider sixty years a "century" in terms of the human life cycle.

PHYSICAL CORRELATION

Heavenly Qi relates to the congenital strength and formation of the body's internal organs, as determined by the Universal Energetic Fields (which include the energy emitted from the sun, moon, planets, and Twenty-Eight Constellations).

HEAVEN'S ENERGETIC GRID

Similar to the grid system imprinted on the body's energetic field, there exists a massive interconnected Heavenly grid extending throughout the entire universe. This Heavenly grid connects the sun, moon, stars, planets, and constellations throughout our universe, and extends to other solar systems outside our celestial sphere. There are pockets of Heavenly energy (like energetic points), as well as subtle frequency pathways (similar to energetic channels), that follow a progression of universal order.

EARTHLY QI

The second of the Three Outer Forces is known as the Environmental Force, Acquired Force, or Earthly Qi. This force manifests as the energy of the entire planet. Each life born into human form is not only developed and structured but also influenced by its own unique environment within the Earthly realm. The Qi of the individual's particular geographic location includes the energy of the plants, soil, water, animals, and all natural formations responsible for controlling or influencing the person's life-force energy.

The Chinese believed that the Earth is the realm that predetermines the abundance of harvest, as well as the temperance of man's physical stamina. Both harvest and man's physical stamina is determined through the cyclic changes of the five seasons and their relationship to the Five Phases (birth, maturation, peak, decay, and death). The Earth is the source of the natural forces and rhythms of nature, as well as the basic elements and energies that compose the material world of soil and water, mountains and valleys, oceans and rivers, plants and animals. The Earth is the foundational base of the Five Elements through the cyclic transformations of nature, which in turn, shape the human body's physiological form and govern the biological functions of the tissues.

The ancient Chinese viewed the Earth as the center, and as the ground on which we stand looking outward toward the four directions. Just as Heaven defines our movement through life in terms of time, Earth defines the space in which we live, providing the energetic field for our unfolding, within the boundaries of time.

In Chinese symbology, the Earth is symbolized by the image of the square, whose walls face the four cardinal directions. This concept of the energetic interaction of Heaven (symbolized by the circle) and Earth's (symbolized by the square) was so fundamental to ancient Chinese thought that they even designed the coinage of the imperial realm to reflect this interconnection (a round coin with a square center).

PHILOSOPHICAL CORRELATION

Earthly Qi relates to the study of Feng Shui (Harmony of Wind and Water). Feng Shui is divided into two prominent schools of study, the Form School and the Compass School.

1. The Form School is the original school of Feng Shui, originating in Southwest China. This school focuses primarily on developing environmental harmony, by observing the shape and form of the terrain in conjunction with the Five Elemental animals.
2. The Compass School is the second school of Feng Shui. It focuses primarily on developing universal and environmental harmony through the directional orientation of the compass, in relationship to the Ten Heavenly Stems, Twelve Earthly Branches, Twenty-Four Solar Compass Directions, and Twenty-Eight Constellations, as well as the Magic Square and the Bagua (Eight Trigrams).

PHYSICAL CORRELATION

Earthly Qi relates to the "acquired" strength (muscular and skeletal development after birth), as well as the formation of the body's internal organs. Tissue development is determined by the environmental energetic fields (which include the energy emitted from the Earth, Water, and Wind/ weather). The Earth energy also expresses its power in man's life through the basic instincts for survival and the primal drives for sex and reproduction.

EARTH'S ENERGETIC GRID.

Similar to the Heavenly energetic grid system imprinted in the cosmos, there exists a massive interconnected Earthly grid extending throughout the entire planet. This Earthly grid connects the planet's eight energetic directions into one energetic field. There are pockets of Earthly energy (like energetic points), as well as subtle frequency pathways (similar to energetic channels), that follow a progression of environmental order. One example of the Earth's axial grid system in Western folklore is the "Ley lines," which are energetic pathways that connect energy vortices or places of power on the planet. This understanding of an Earthly energetic grid system is also reflected in American Indian and South American traditional understanding.

MAN'S QI

The third of the Three Outer Forces is known as the Human Force, or Man's Energy. This force manifests as the energy of the entire body (Figure 10.2). Each life born into human form is not only developed and structured, but also influenced by the individual's adaptations to his or her own physical, mental, emotional, energetic, and spiritual energies.

It is believed that within man, the realm of the Dao or divine conscious spirit exists, wherein both the energies of Heaven and Earth coexist in harmonious balance. Man is formed from the cosmic particles of exploded stars that have descended onto the Earth. These particles are absorbed into the soil, water, and foliage, creating sustenance for human life. This constant internal connection, absorption, and regeneration of vibrating energy from Heaven and Earth is the reason that human life is considered to be the highest physical manifestation of the divine energetic field (see Chapter 1).

PHILOSOPHICAL CORRELATION

Man's Qi relates to the study of the Yi-Jing (I-Ching), or Book of Changes. The Yi-Jing is composed of two sections: the Book of Oracles and the Book of Commentaries.

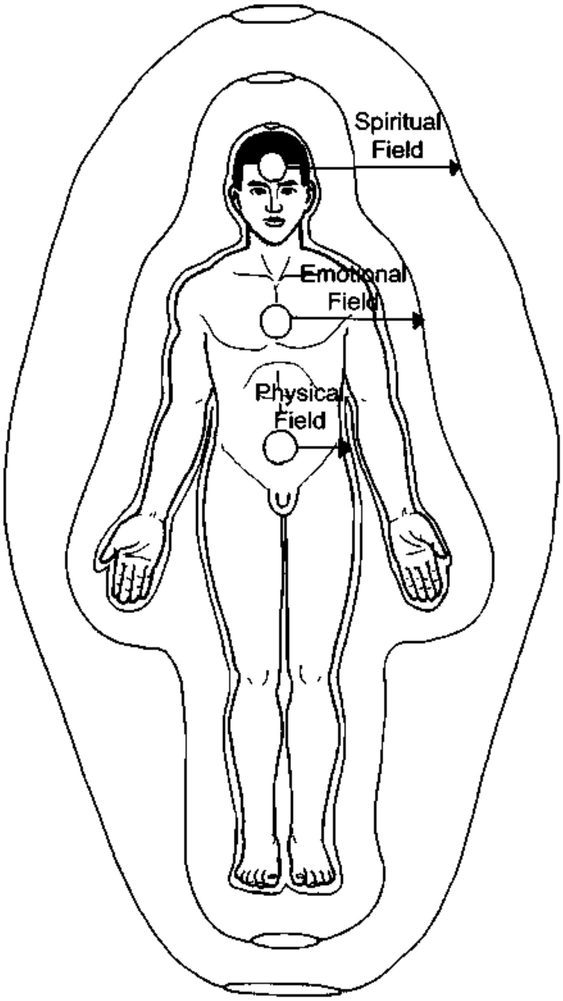

Figure 10.2. The Three Outer Forces of Man's Energy is composed of the Energetic Fields Released from Jing, Qi, and Shen.

1. The Book of Oracles is the original text by King Wen and the Duke of Zhou. This book is based on the formation of eight trigrams and their various energetic attributes. These eight trigrams combine to form the sixty-four hexagrams of the Yi-Jing, which were later used for divination.
2. The Book of Commentaries is based on the commentaries of Confucius and his followers, and is also called the Ten Wings. The commentaries are called "wings" because each

wing helps the reader to soar higher in comprehension of the original text.

PHYSICAL CORRELATION

Man's Qi relates to the waxing and waning of man's physical, mental, emotional, energetic, and spiritual biorhythmic cycles.

A Qigong doctor must become proficient at understanding, respecting, and adopting these laws of natural variations between nature and man. Through this understanding the doctor can attain a deeper comprehension of the principles and application of Medical Qigong Dao Yin training (mind concentration methods, respiration regulation, and postural manipulation).

MAN'S ENERGETIC GRID

Similar to the Energetic grid system of Heaven and Earth, the human body is imprinted with a massive interconnected grid that extends throughout the body's tissues into the environment. The body's energetic grid is responsible for transforming energy frequencies from cell to cell, tissue to tissue, and from the body's internal energetic field to its external energetic field. These are known in Traditional Chinese Medicine as the 'Jing Luo' channels and collaterals (or the meridian system). The body's energetic grids are also responsible for maintaining the production, growth, and development of energy and tissue formation, and are influenced through intention, thought, and emotion.

CHAPTER 11
THE THREE TREASURES OF HEAVEN: SUN, MOON, AND STARS

DEFINING HEAVEN QI

The Outer Force (Power) of Heaven's Three Treasures (the Sun, Moon, and Stars) manifests as electromagnetic radiation from the sun, reflected light and energy from the moon, starlight energy, and the gravitational forces and interactions of the above mentioned heavenly bodies. The Outer Force of Heaven's Three Treasures also include the twenty-eight constellations, five major planets, and their energetic and gravitational forces.

Heaven Qi is the energy emitted from the universe outside of the realms of Earth and man. The sun and moon, first five planets (Jupiter, Venus, Saturn, Mars, and Mercury), North Star, Big Dipper, and the twenty-eight constellations are the most easily observed natural Heavenly energies, which have a direct connection to life on Earth. The diagram below (Figure 11.1) shows the Five Elements relationship between the universal treasures of Heaven, the environment of Earth and Man's internal organs.

The sun and moon are the primary Heavenly manifestations of Yin and Yang energy. The sun is Yang and is considered the Spirit Soul (Hun) of the sky. The moon is Yin and considered the Corporeal Soul (Po) of the Earth. This interaction of Heavenly Yin and Yang energy affects and governs the Earthly Yin and Yang energy, as well as the Qi within the human body. The sun and moon each have their own cyclic rhythms and energies, changing their relative degrees of Yin to Yang at different times of the day, month, and year.

To cleanse, purify, and strengthen the body, Qigong doctors absorb the essence of the Heavenly Qi by guiding the star's energy into certain areas of the body. The energy of the stars may also be used to reinforce the circulation of Wei Qi for

Organ	Element	Time of Day	Direction	Planet	Associated Animal
Liver	Wood	Early Dawn	East	Jupiter	Green Dragon
Heart	Fire	Midday	South	Mars	Red Phoenix
Spleen	Earth	Midafternoon	Center	Saturn	Yellow Phoenix
Lungs	Metal	Late Dusk	West	Venus	White Tiger
Kidneys	Water	Night	North	Mercury	Indigo Tortoise

Figure 11.1. Natural Heavenly Transformations

added protection against external pathogenic invasion.

INFLUENCES OF HEAVENLY QI ON THE HUMAN BODY

The body's physical fluctuations work in harmony with the cycles of the sun, moon, and nature. The science that deals with the study of these biological clocks is known as chronobiology. The levels of corticosterone and plasma adrenocorticotropic hormone (ACTH), for example, follow a cyclical pattern during the course of a day. Generally, the body's emotional and menstruation cycle in women follow the changing patterns of the moon (i.e., approximately every 28 days). Additionally, approximately every 23 days, the male body releases stored toxins, changing the smell of the sweat in accordance with the body's physical biological cycle. The smooth functioning of natural biological rhythms can at times be disrupted. The "jet lag" phenomenon, for example, is due to the interruption of circadian rhythms in the body.

There are daily, weekly, monthly, and annual solar and lunar cycles to which the human body responds. Other human physical, emotional, and mental cycles are strongly influenced by the types of life-style chosen, and other environmental factors involved. The body's circadian rhythms influence not only the immune system, but also the Yin and Yang energies of the internal organs. The activities of the cells, glands, Kidneys, Liver, and nervous system are coordinated with one another and with the day-night rhythm of the environment. Hunger, excretion, and body temperature are influenced by the body's ability to respond to the micropulsations of the subtle energy fields within the internal organs and tissues, as well as the body's biological clocks.

In China, clinical scientist Liu Bing, from Beijing Medical University, notes in his research that the human body undergoes rhythmic variations within a 24-hour time period. These physical changes include temperature, oxygen consumption, blood pressure, pulse rate, content of hemoglobin, blood sugar, blood amino acids, blood adrenocortical hormone, hepatic glycogen, metabolic rate of fat absorption, and pace of cell metabolism. The natural clockwork of the body can also be altered by food, drink, drugs, and abnormal sleeping patterns.

BODY RHYTHMS AND ENERGETIC PATTERNS

The prenatal and postnatal energetic patterns of the body are also determined by the time and place of conception. The time and place of conception establishes the patient's spiritual energy and ancestral traits (food preferences, manner of dress, preferences of art, spiritual beliefs, and so on). The time and place of birth also determine the patient's energetic biorhythms (physical, emotional, and intellectual cycles).

PRENATAL ANCESTRAL TRAITS

The time and location of conception determine the nature of the ancestral spiritual influence on the patient. As the divine infuses the fetus's soul, the energy of the Eternal Soul combines with the environmental spiritual influences of the geographic location (e.g., the Orient, Europe, North America, etc.). This infusion of geographic spiritual energy creates within the patient a predisposition towards specific ancestral traits and cultural attractions.

At conception and throughout pregnancy, the mother absorbs the natural environmental energy through respiration (breathing through the mouth, nose, and pores), ingesting food (grown in the soil), and visual/auditory ingestion (observing and experiencing the cultural environment). The energetic history of each cultural environment exists within the memories stored in the natural energetic fields contained within that environment. Therefore, someone conceived within a particular geographic location in the Orient may find him or herself, unconsciously drawn to the social and cultural influences of that particular Asian population. This subconscious attraction is considered a natural phenomenon due to the geographic influences of the country's spiritual/ancestral power, causing an effect, indirectly, on the individual.

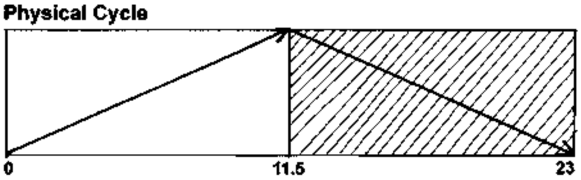

The Physical Cycle is 23 days long. The first 11 1/2 days are the positive side of the cycle. One experiences a feeling of good physical strength and endurance. The second 11 1/2 days are the negative side of the cycle, marked by having less endurance and a tendency towards fatigue.

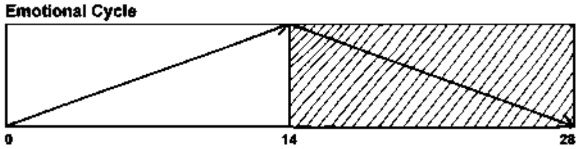

The Emotional Cycle is 28 days long. The first 14 days are the positive side of the cycle. One feels optimistic, cheerful, and cooperative. For the second 14 days, a negative stage of the cycle results in tendencies to be moody, irritable, and pessimistic.

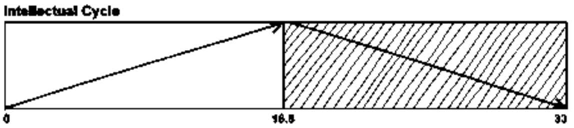

The Intellectual Cycle is 33 days long. The first 16 1/2 days are the positive side of the cycle. One has greater success in learning new material and pursuing creative, intellectual activities. The next 16 1/2 days are the negative side of the cycle; one is encouraged to review old material rather than attempting to learn new concepts.

Figure 11.2. The physical, emotional, and intellectual cycles begin on the day of birth.

1. The natural environment of the location at conception exerts a strong influence, such as a preference, or attraction to similar geological formations throughout the individual's life. People conceived by the ocean, for example, may find within themselves an unconscious need to live by the ocean. Likewise, people conceived in the mountains, valleys, tropics, deserts, etc. may find an energetically calm feeling of peace envelop them when visiting such places.
2. The time of conception also exerts a strong influence. The energetic formation, as well as strength or weakness, of the fetus's internal organs is also determined by the positions of the sun, moon, and stars. These Heavenly energies begin their effect on the fetus at the time of conception and continue influencing the fetus's formation throughout the entire pregnancy.

BIRTH-POSTNATAL ENERGETIC PATTERNS

The time and location of birth determines the biorhythmic patterns of each individual. These biorhythms are developed according to the influence of the Heavenly energy (position of the sun, moon, and stars), as well as the Earthly energy (i.e., geographic location, e.g., mountains, valleys, by the ocean, etc.). This infusion of the Earth's geographic energy and Heaven's constellation energy creates within the individual a predisposition towards specific psychological traits.

1. The location of birth strongly influences the individual's desire to seek out and/or surround him or herself with the familiar environment.
2. The Time of Birth may also have an unconscious influence on his or her living patterns, as certain individuals born at night naturally become "night people," while others born in the morning function better during the morning hours.

The body's biorhythms, which react to universal and environmental energetic vibrations, are divided into three distinct cycles and energy flows (Figure 11.2). Each rhythm is cyclical, waxing and waning, creating and contributing to times of physical, emotional, and intellectual triumph, as well as times of reflection and withdrawn behavior. These three cycles begin at the moment of birth and continue with absolute regularity until death. The conditions of the cycle are divided into positive (the first half) and negative (the second half) attributes.

The body's biorhythms are but one example of the effects that Heavenly energy has on the physical, emotional, and intellectual cycles of mankind. By understanding the cycles of the sun, moon, and stars and their energetic potential, the Qigong doctor can utilize the universal energy of Heaven to regulate and balance the patient's Qi. The biological rhythms influence the time at which many illnesses occur or worsen, as well as how fast a medication takes effect and how long the effect lasts. Therefore, Medical Qigong therapy, as well as herbal therapy are prescribed at specific times to enhance their effect on the patient's body.

CYCLES OF THE SUN

The Earth's rotation around the sun causes seasonal changes and instigates cyclic metabolic patterns in the human body. The body's entire metabolic function is influenced by its ability to respond to the seasonal micropulsations of these subtle energetic fields.

The Chinese use both a solar and a lunar calendar. The Heavenly cycle of the moon has 12 lunar month divisions, which correspond to man's 12 internal organs and channels: Lungs, Large Intestine, Stomach, Spleen, Heart, Small Intestine, Bladder, Kidneys, Pericardium, Triple Burners, Gall Bladder and Liver. The Heavenly cycle of the sun has 12 periods of the day, which correspond to man's 12 sections of the body. These 12 internal cycles are sometimes called the 12 Blood / Heat Cycles because they affect the Qi, Blood and temperature within the Twelve Skin Zones (see Chapter 7).

The Earth's rotation and the rhythm of the five seasons (the usual four plus late summer) influence the body's internal organs and channels, affecting their energetic activity. In each of the five seasons, a specific pair of internal organs reaches an energetic peak.

1. In the summer, the Heart and Small Intestine are at their energetic peak.
2. In the late summer, the Spleen and Stomach reach their energetic peak.
3. In the autumn, the Lungs and Large Intestines reach their energetic peak.
4. In the winter, the Kidneys and Bladder reach their energetic peak.
5. In the spring, the Liver and Gall Bladder reach their energetic peak.

THE TWELVE PI HEXAGRAMS

The Twelve Pi Hexagrams symbolize the seasonal changes, which cause the waxing and waning of Yin and Yang energy. The trigrams and hexagrams are used to illustrate the movement of energy in Man (see Chapter 5) as well as Heaven. The Twelve Pi Hexagrams were also used to study and comprehend the interactive energetic fields of man and nature that govern growth and development. The interactive energetic fields engender the creation of energy channels and influence the circulation and transformations of Qi, which in turn, create and maintain the body's internal organs. The ancients believed that the human physique reflects the structure of the universe, and that prenatal Heaven and Earth energy fuse with postnatal Heaven and Earth energy at birth (see Chapter 2). The trigrams reflect the twelve hours of the day, the daily fluctuations of Yin and Yang energy, and the Qi of the Five Elements (Figure 11.3).

The Twelve Pi Hexagrams represent "the laws of hardness and softness," domination and yielding of Yin and Yang, the relationship between the seasons, time of day, major organs, channels, and the transformation and circulation of Qi.

THE YANG YAO CYCLES

When studying the Twelve Pi Hexagrams, it is important to examine the first six (from Fu — 11 p.m. to Qian — 11 a.m.), that represent the gradual increasing of the energetic potential of the Yang Yao (see Chapter 5) and the gradual decreasing of the Yin Qi. The Yang Qi rises and expands during this time period; it is the best time for training techniques that emit external Qi and for treat-

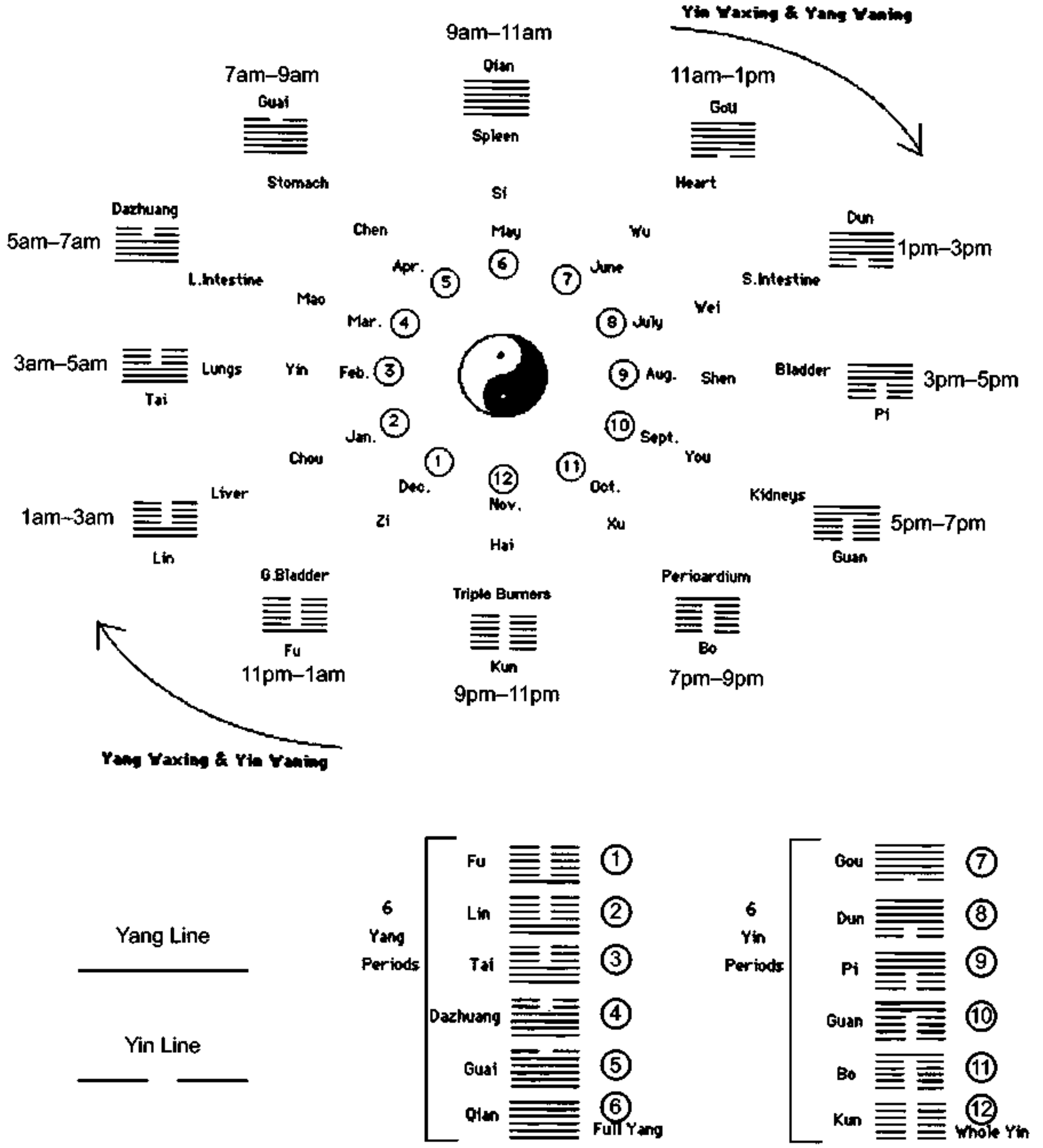
Yin Waxing & Yang Waning
9am–11am
Qian
Spleen
7am–9am
Guai
Stomach
11am–1pm
Gou
Heart
Si
Dazhuang
5am–7am
Chen
May
Wu
Dun
1pm–3pm
L.Intestine
Apr.
5
6
7
June
S.Intestine
Mao
Mar.
4
8
July
Wei
3am–5am
Lungs
Yin
Feb.
3
9
Aug.
Shen
Bladder
3pm–5pm
Tai
Pi
Jan.
2
10
Sept.
Chou
1
12
11
You
Dec.
Nov.
Oct.
Liver
Kidneys
1am–3am
Zi
Xu
5pm–7pm
Lin
Hai
Guan
G.Bladder
Pericardium
Triple Burners
Fu
Bo
11pm–1am
Kun
7pm–9pm
9pm–11pm
Yang Waxing & Yin Waning
Yang Line
Yin Line
6
Yang
Periods
Fu
1
Lin
2
Tai
3
Dazhuang
4
Guai
5
Qian
6
Full Yang
6
Yin
Periods
Gou
7
Dun
8
Pi
9
Guan
10
Bo
11
Kun
12
Whole Yin

Figure 11.3. The circular figure depicts the Heavenly cycle of the sun and the Twelve Pi Hexagrams. Solid lines in the trigrams and hexagrams represent Yang, and broken lines represent Yin. The hexagrams on the left depict the Yang Yao Cycle (from Fu to Qian), with the gradual increase in Yang and decrease in Yin, so that the last hexagram, Qian, is full Yang. The hexagrams on the right depict the Yin Yao Cycle (from Guo to Kun), with the gradual increase in Yin and decrease of Yang, so that the last hexagram, Kun, is full Yin.

ing patients who suffer from insufficiency of Yang Qi. This time period is called the "time of the living breath," and represents the beginning of energetic harvesting. When training in the morning time, the practitioner should progress from still, relaxed, quiescent Qigong meditations to active forms of dynamic Qigong movement.

THE YIN YAO CYCLES

The last six hexagrams (from Gou — 11 a.m. to Kun — 11 p.m.) indicate a gradual increasing of the energetic potential of the Yin Yao and the gradual decreasing of the Yang Qi. This time period is called the "time of the dead breath," and represents the beginning of energetic planting. The Yin Qi increases and expands in this time period; it is the best time for patients suffering from an insufficiency of Yin Qi to train, nurture, and restore the Kidneys' Yin Essence, which in turn will facilitate the preservation of Yang Qi. When training in the afternoon, the practitioner should progress from active forms of dynamic Qigong movement to gradually slower, relaxed, quiescent Qigong meditations.

When prescribing Medical Qigong for patients, the time for practice should be selected in accordance with the philosophy of Yin and Yang theory, the Five Elemental theory, the circadian rhythms of Qi and Blood circulation, as well as the severity of the illness.

USING THE YANG AND YIN YAO CYCLES

Each of the broken lines represents a Yin Yao and can be used to symbolize an inhalation, while a straight line represents a Yang Yao and symbolizes an exhalation. A hexagram consisting of three bottom Yang Yaos and three top Yin Yaos, for example, represents a balanced respirational pattern. When the Qigong doctor writes out a prescription for the patient, the doctor can adjust the patient's respiration in accordance with the Yao hexagrams, taking into consideration each Yao line as representing an inhalation or exhalation. The basic theory of Yin and Yang Yao cycles is also used to determine the time and direction for Qigong practice, as well as the methods of respiration and mind concentration required to establish proper Yin and Yang balance.

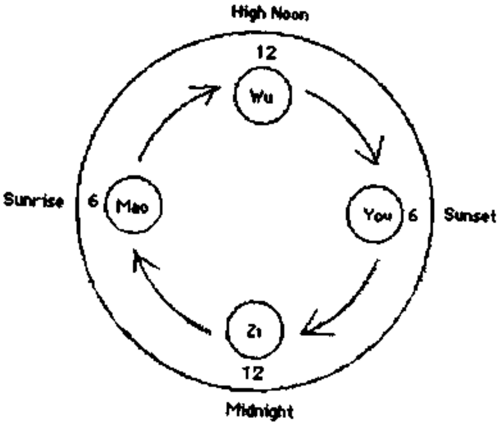

Figure 11.4. The Four Principle Time Periods

When practicing Medical Qigong at the hour of the Fu Hexagram, for example, the Yao indicates there is more Yin than Yang, so the practitioner should regulate his or her body with the intent to nourish the Yang through longer inhalation (Yin) and shorter exhalation (Yang).

At the stage of the Gou Hexagram, the Yao indicates there is more Yang and less Yin so the Qigong practitioner should perform regulation through shorter inhalations (Yin) and longer exhalations (Yang).

THE FOUR PRINCIPAL TIME PERIODS OF QIGONG PRACTICE

Practicing Medical Qigong during the "four principal time periods" (Zi, Mao, Wu, and You) facilitates the vigorous growth of internal energy in harmony with the changes of energy in nature (see Figure 11.4). To nourish energy, practice during the sunrise and/or sunset (Mao and You) periods. To stabilize energy, practice during the periods of midnight and/or high noon (Zi and Wu).

Due to the differences between individual constitutions (Excess or Deficient Yang or Yin), the selection of the proper time for practice and the number of Yin or Yang respirations varies between patients. The following is a brief discussion and explanation of the four principal time periods.

1. Zi Time Period corresponds to the Fu Hexagram, and has one Yang and five Yin lines or Yaos. The Zi time period is at midnight (11 p.m. – 1 a.m.). It is related to the beginning of the Yang cycle and is perfect for gathering energy to store the Yuan Qi (Prenatal Qi). The Yuan Qi is the root of life, formed through the divine combination of the primordial Yin and Yang; it is stored primarily in the Kidneys' Water energy. The Zi period of time corresponds to the Fu Hexagram. During this time the Earth energy and the Kidney's Yang Qi are in full harmony, continuously combining and accumulating energy. Practice during this period of time results in twice the effect with half the effort.
2. Mao Time Period corresponds to the Dazhuang Hexagram, and has four Yang and two Yin lines or Yaos. The Mao time period is at sunrise (5 a.m. – 7 a.m.). As the sun rises, warming the air, and natural Yang progresses. At this time of the day, the Yang Qi of the body is in full bloom. Practice during this period of time aids the vigorous growth of Yang.
3. Wu Time Period corresponds to the Gou Hexagram, and has one Yin and five Yang lines or Yaos. The Wu time period is at high noon (11 a.m. – 1 p.m.), and is regarded as the Heart's Fire. At high noon, the Heart Channel (considered the Supreme Yang) will naturally manifest the energetic peak of Yang energy. Because the first line is a Yin Yao, however, the Yang energy Yaos tend to wane rather than grow. Therefore, practicing during this period helps the growth of Yin energy and suppresses the hyperactivity of Yang energy.
4. You Time Period corresponds to the Guan Hexagram, and has four Yin and two Yang lines or Yaos. The You time period is at sunset (5 p.m. – 7 p.m.). At this time, the energy of the environment turns from clear and radiant to dark. Qigong practice at this time nurtures the increase of the primordial Yin energy and the conservation and nourishment of Yang energy.

THE FOUR SEASONAL PEAK TRANSITIONS

Another theory maintained by Qigong doctors is that the sun's light during the day represents Yang, and the energy from the moon's light at night represents Yin. The observation that light coincides with heat and that darkness coincides with cool, led to the division of the year into four seasonal peak transitions (which also correspond to the four principal time periods of the day) (see Figure 11.5).

At the time of the equinoxes and solstices, the sun passes through transitional regions: The Palace of Eternal Frost in the North is entered in the winter; The Golden Gate in the East is entered in the spring; The Palace of Universal Yang in the South is entered in the summer; and The Gate of the Moon in the West is entered in the autumn. The four seasonal peak transitions in the northern hemisphere are described as follows.

1. The Palace of Eternal Frost at the Winter Solstice (December 22–25) indicates the time period when the sun is furthest south from the equator. This is a time of maximum Yin or darkness in the northern hemisphere, with the shortest days and longest time of cold and darkness. This energetic transition corresponds to "midnight" within the four principal time periods.
2. The Golden Gate at the Spring Equinox (March 21) indicates the time period when the sun's center crosses the Heavenly equator, creating an energetic balance between the days and nights. Although light and darkness and Yin and Yang energies are in balance, this seasonal time period is a transitional harmony, as Yang energy is growing. The weather is also moderate. This energetic transition corresponds to the "sunrise" within the four principal time periods.
3. The Palace of Universal Yang at the Summer Solstice (June 21–23) indicates the time period when the sun is furthest north from the equator, a time of maximum Yang or brightness, with the longest days and times of light and warmth. This energetic transition corresponds to "noon" within the four principal time periods.

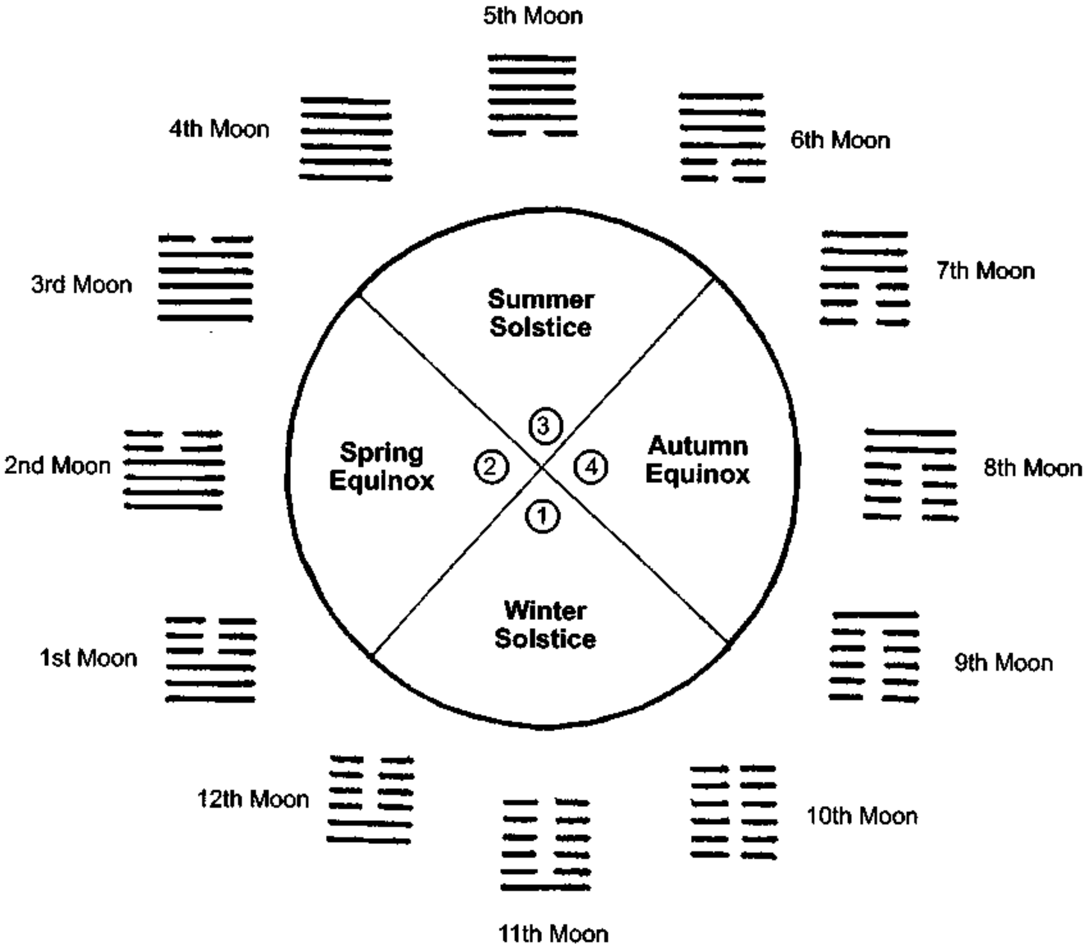
5th Moon
4th Moon
6th Moon
3rd Moon
7th Moon
2nd Moon
8th Moon
1st Moon
9th Moon
12th Moon
10th Moon
11th Moon
Summer Solstice
Spring Equinox
Autumn Equinox
Winter Solstice
③
②
④
①

Figure 11.5. Four Season Peak Transitions

4. The Gate of the Moon at the Autumn Equinox (September 22) indicates the time period when the sun's center crosses the Heavenly equator, creating an energetic balance between the days and nights. Although light and darkness and Yin and Yang energies are in balance, this seasonal time period is in transitional harmony, as Yin energy is growing. The weather is also moderate. This energetic transition corresponds to the "sunset" within the four principal time periods.

The energy of these four principal time periods enters the practitioner's Taiji Pole during meditation and regulates the "auspicious powers" located within the five regions of space inside and outside of the doctor's body:

- Front—energy of the Heart
- Back—energy of the Kidneys
- Right—energy of the Lungs
- Left—energy of the Liver
- Center—energy of the Spleen

THE SIX PRINCIPAL SEASONS OF TRANSITION

In Traditional Chinese Medicine, the year is divided into six primary periods of seasonal transitions (wind, heat, damp, fire, dry, and cold). Each season represents the ruling Qi, or expected weather, dominant in that particular season (Figure 11.6).

These six periods delineate the differences between and transitional transformations of Heaven energies (through the Ten Heavenly Stems) and Earthly energies (through the Twelve Earthly Branches). In terms of weather, the first half of the year is governed by the Qi of Heaven, while the second half of the year is administered by the Qi of Earth. The Ten Heavenly Stems are the ten energies of Heaven that rule the changes of the Five Elemental Seasonal Transitions, and are represented in the human body as the Yin and Yang aspect of the Five Elements (represented as the ten major internal organs). The Twelve Earthly Branches are the twelve energies of the Earth that determine the six Qi factors of the seasonal transitions (represented in the body as the Twelve Primary Channels). In accordance with the Creative Cycle of the Five Elemental Seasonal Transitions: summer creates late summer, late summer engenders autumn, autumn gives rise to winter, and winter gives birth to spring.

THE DAILY QI, BLOOD, AND HEAT CYCLE

These four principal time periods and seasonal peak transitions affect the body's Blood, Heat, and Qi Cycles. The daily cycle of Heat, Blood, and Qi flow along the body's Twelve Primary Channels, organs, and tissues, according to the two hour Yin and Yang time table are:

- Gall Bladder (GB) - 11 p.m. to 1 a.m. - (Yang)
- Liver (Lv) - 1 a.m. to 3 a.m. - (Yin)
- Lungs (Lu) - 3 a.m. to 5 a.m. - (Yin)
- Large Intestine (LI) - 5 a.m. to 7 a.m. - (Yang)
- Stomach (St)- 7 a.m. to 9 a.m. - (Yang)
- Spleen (Sp) - 9 a.m. to 11 a.m. - (Yin)
- Heart (Ht) - 11 a.m. to 1 p.m. - (Yin)
- Small Intestine (SI) - 1 p.m. to 3 p.m. - (Yang)

Season	Date
1. Spring Begins	Feb. 5
Rainwater	Feb. 20
Insects	March 7
Spring Equinox	March 22
2. Clear and Bright	April 6
Grain Rains	April 21
Summer Begins	May 6
Grain Filling	June 7
3. Grain Full	June 22
Summer Solstice	June 22
Slight Heat	July 8
Great Heat	July 24
4. Autumn Begins	Aug. 8
Limit of heat	Aug. 24
White Dew	Sept. 8
Autumn Equinox	Sept. 24
5. Cold Dew	Oct. 9
Frost Descent	Oct. 24
Winter Begins	Nov. 8
Slight Snow	Nov. 23
6. Great Snow	Dec. 7
Winter Solstice	Dec. 22
Slight Cold	Jan. 6
Great Cold	Jan. 21

Figure 11.6. Six Principal Seasons of Heaven and Earth Energy Transition (1-6), and the Twenty Four Diagram Chart

- Bladder (Bl) - 3 p.m. to 5 p.m. - (Yang)
- Kidneys (Kd)- 5 p.m. to 7 p.m. - (Yin)
- Triple Burners (TB) - 7 p.m. to 9 p.m. - (Yin)
- Pericardium (Pc) 9 p.m. to 11 p.m. - (Yang)

The body's active high and low heat, Blood, and Qi time periods, reflect its relationship with the Twelve Primary Channels. For instance, every morning during the time period of 3–5 a.m., the strongest energy streams down along the Lung Channels, making the Lungs quickly responsive to treatment.

GATHERING THE SUN'S ESSENCE MEDITATION

The ancient Qigong masters used this meditation to absorb the essence of the sun. By gathering pure Yang energy to tonify the body's Deficiencies, as well as to consolidate the Zheng Qi (Gathering energy), they were able to eliminate pathogenic factors and prevent premature aging. This particular exercise is considered part of the foundational base from which the ancient Qigong doctors replenished any depleted life-force energy.

The gathering the Sun's Essence Meditation must only be practised on a clear day, when the sun is not obstructed. Avoid practising this meditation in times of fog, smog, thick clouds, and solar eclipses, as cultivation during these times will cause the unnatural cultivation of Turbid Qi.

1. While outside on a bright sunny day, face the sun, and assume a Wuji standing posture, hands by your sides, with the feet slightly wider then shoulder-width apart; stand in a relaxed and tranquil state, breath evenly and naturally, and rid your mind of any stray thoughts (Figure 11.7).
2. If the sun is rising above the horizon, slightly close your eyes; you should still be able to see soft, gentle, reddish sunlight. If the sun has already risen, close your eyes and use your inner vision to visualize the sun.
3. Inhale through the nose and imagine filling your mouth with energy from the sunlight's essence. Still facing the sun, with your eyes half open, "infuse," or absorb into the eyes, the energy from the horizon. Blend the energy

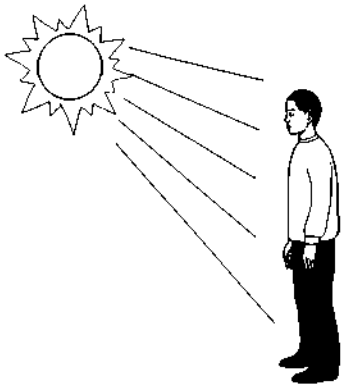

Figure 11.7. Gathering the Sun's Essence Meditation. Step one, stand and face the sun.

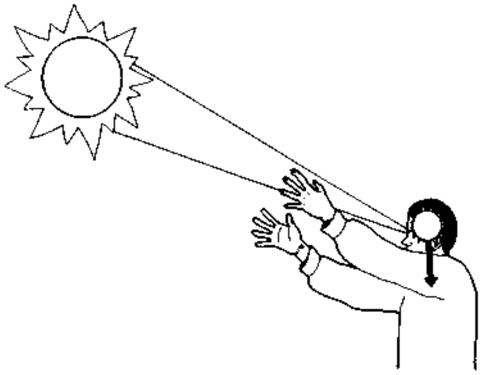

Figure 11.8. Step two, imagine embracing the sun.

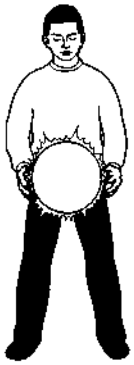

Figure 11.9. Step three, imagine embracing the sun and gently placing it in front of your Lower Dantian.

absorbed from your eyes into your Upper Dantian. Close your eyes and imagine blending the sun's golden light energy with the Qi you inhaled through your nose. Hold your breath and focus the mind, mixing both energies; swallow the energies slowly while exhaling, and send the energy down to the Lower Dantian. Repeat this process for several minutes.

4. Stretch your hands towards the sun and imagine embracing it (Figure 11.8). Point your index fingers toward the center of the sun, and imagine pulling the sun down in front of your Lower Dantian. Keep it positioned there for several minutes (Figure 11.9).
5. Imagine the sun gently rolling and turning itself in front of your lower abdomen (Figure 11.10–11). Keep your elbows at your sides and gently shift, or sway from side to side, imagine that it is the sun's movement that causes both your arms and body to move. Using this image keeps the sun in an energetic power position.
6. Place both hands around the sun as if embracing it (Figure 11.12). Next, imagine the sun slowly moving into your navel. Allow both hands to follow the sun into the navel, ending with your left palm on top of your right palm (opposite for women) (Figure 11.13). With one deep breath inhale and absorb the sun's energy into the Lower Dantian while imagining the sun (beneath the navel) transforming into a luminous ball of golden energy. With each inhalation, this golden ball shines brightly; with each exhalation, the sun's rays extend, spreading throughout the entire body.
7. To end, relax the mind and imagine Heavenly Qi melting down the body, pouring through the front, back and center of the body like warm oil, into the Earth. This ending exercise (sometimes called Pulling Down The Heavens), allows the Qi to sink into the Lower Dantian, and the body, mind, and spirit to come into a state of peaceful rest and tranquillity.

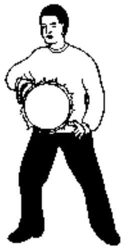

Figure 11.10. Step four, imagine the sun gently rolling and turning itself in front of the Lower Dantian.

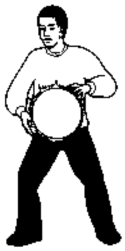

Figure 11.11. The energy of the sun moves the entire body rather than the arms.

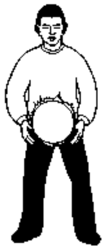

Figure 11.12. Step five, embrace the sun and absorb it into the navel and then into the Lower Dantian.

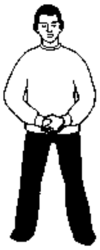

Figure 11.13. Each breath transforms the sun's energy into luminous golden Qi.

CYCLES OF THE MOON

The ancient Qigong masters believed that the sun has a Yang nature, emitting enormous amounts of hot, thermal energy, whereas the moon has a Yin nature and emits a much cooler type of energy, by reflecting the light of the sun. The energy of the moon is used to replenish the body's Yin Qi, and follows a comprehensive cycle every 28 days (Figure 11.14). Following the cycle of the moon, the energy of the body completes a full rotation, flowing along the Governing Vessel to the top of the head, and down the Conception Vessel to the base of the perineum. When the moon is full, the body's Blood and Qi reach a crest, filling the Baihui point on the top of the head. At the new moon, the body's Blood and Qi gather at the Huiyin point, in the perineum.

Since the lunar cycle at its extreme (full moon and new moon) activates the top and bottom points of the Taiji Pole, these are said to be the most productive times to perform Taiji Pole Qi practice.

The appearance of the moon during the first half of the lunar month symbolizes the waxing of Yang and waning of Yin Qi. As Yang grows and Yin declines, regulation of breathing at this period of time should be aimed at absorption of the Moon's Essence. The second half of the lunar month symbolizes the waxing of Yin and waning of Yang Qi. As Yin grows and Yang declines, regulation of breathing at this period of time should be aimed at replenishing Yang to help the Yin nourish the Yin Jing-Essence (Figure 11.15-20).

As the moon increases the amount of light it reflects, the body's Qi and Blood become more nourished; when the moon is full, the Qi and Blood become substantial. As the moon begins to decrease the amount of light it reflects, the muscles start to lose their energetic substance. This decrease in energy causes the body's Eight Extraordinary Vessels and Twelve Primary Channels to empty themselves to support the muscles that are slowly becoming Deficient. Qi and Blood are influenced by the gravitational pull of the sun and moon, just as the tides of the sea are influenced. Therefore the ancient Qigong masters described the energetic tonifying and purging principles of Qigong regulation in these specific terms: when treating a patient, perform no tonifying when the moon is full, or when the Blood and Qi are sufficient. Perform no purging when the moon is dark, or when the Blood and Qi are Deficient. This observation explains the effect of the "full-moon phenomenon" on patients who are hypersensitive to energy and become restless when the moon's Qi is "full Yang in Yin."

The waxing and waning of the sun and moon affect the adjustment time for Medical Qigong practice and the methods of breathing regulation. In order to facilitate the intake of Heavenly Es-

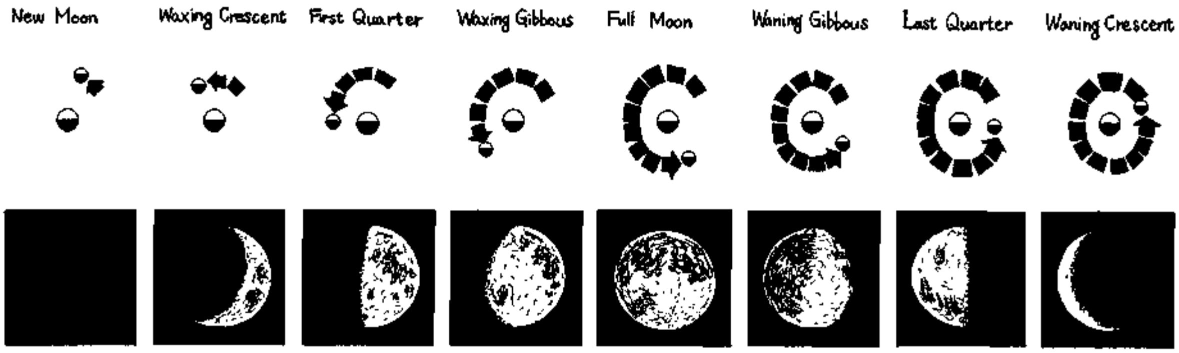

Figure 11.14. The waxing and waning cycles of the moon correspond to the moon's orbit around the Earth.

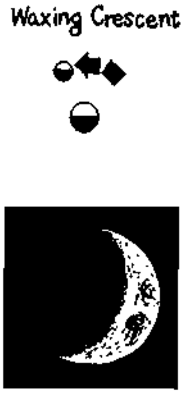

Figure 11.15. From the 1st to the 5th night of a lunar month, more Yang absorption and Yin nourishment can be obtained by performing breathing regulations while facing the west. The crescent surface of the Moon is reflecting less sunlight.

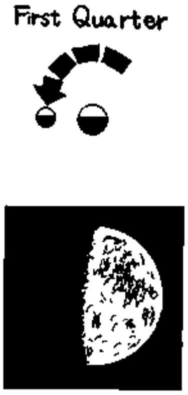

Figure 11.16. From the 6th to the 10th night of a lunar month, more Yang absorption and Yin nourishment can be obtained by performing breathing regulations while facing the south. The surface of the moon is turning from a crescent to half-full Yang, and its surface is reflecting more sunlight.

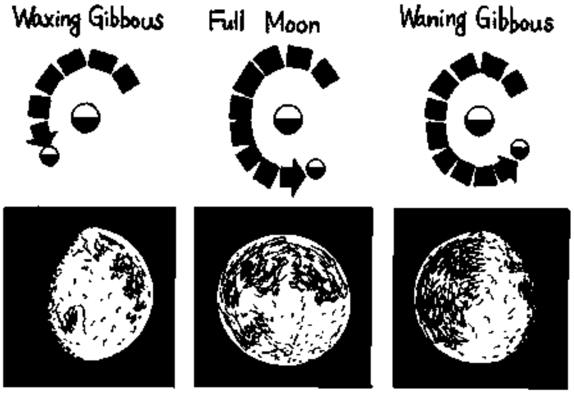

Figure 11.17. From the 11th to the 15th night of a lunar month, equal amounts of Yang absorption and Yin nourishment can be obtained by performing breathing regulations while facing the direction of the moon, as both Yang and Yin energies are plentiful at this time. At this point, the moon is full with its surface reflecting the most sunlight (called "Full Yang in Yin").

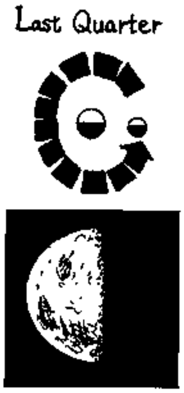

Figure 11.18. From the 16th to the 20th night of a lunar month, the Yin grows and Yang declines. Regulation of breathing at this period of time should be aimed at replenishing Yang to help the Yin nourish the Yin Essence, while facing the west. At this time, the crescent surface of the moon is reflecting less sunlight. The reflected sunlight on the moon's surface decreases as Yang wanes and Yin waxes.

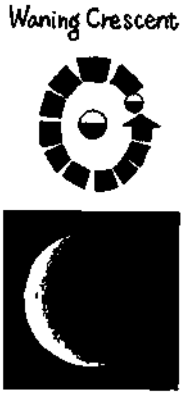

Figure 11.19. From the 21st to the 25th night of a lunar month, Yin continues to grow and Yang declines. Regulation of breathing at this period of time should be aimed at replenishing Yang to help the Yin nourish the Yin Essence, while facing the south. At this time, the crescent surface of the moon's shape is turning to half-full Yin, and its surface is reflecting less sunlight. The reflected sunlight on the moon's surface continues to decrease as Yang wanes and Yin waxes.

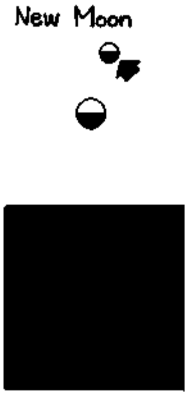

Figure 11.20. From the 26th to the 30th night of a lunar month, no absorption of Yang is possible when the moon appears in the east at nightfall and its surface is dark, leading to the period of "Pure Yin." Breath regulation should be performed facing east.

sence to nourish the body, the ancient Qigong masters used a system wherein each month is divided into 6 separate sections (a total of 5 days was assigned to each section).

PRENATAL EIGHT TRIGRAMS AND THE MOON

The symbols on the Prenatal Eight Trigrams represent the appearance of the moon, while the Ten Heavenly Stems represent the positions of the moon. Each trigram embraces one or two of the Ten Heavenly Stems. The following chart of Prenatal Trigrams, Ten Heavenly Stems, moon and tide regulation, as well as the principles of purgation and tonification of Qi and Blood is a reference for Qigong doctors for clinical treatments (Figure 11.21).

In this chart, the Yang energy of the Earth gradually increases from the 11th to the 15th. Influenced by this action, the spring tide occurs on the 15th when the moon is full, inducing both the Qi and Blood to become extremely active. This is a good time for Qigong patients who suffer from energy Deficiencies to replenish their Qi. However, if the patient's energy is abundant, he or she should not try to replenish at this time, or an Excess condition of Qi will occur.

CHOOSING THE PROPER DIRECTION FOR TRAINING QI

Since ancient times, Qigong doctors have placed great emphasis on facing certain directions when cultivating energy. Observing that energy training and practicing should be adapted to nature, each routine is selected according to the season, month, and time of day, as well as the physical conditions of the practitioner. The four directions, as well as the sun (solar cycle), moon (lunar cycle), and stars (four sets of seven constellations, each one grouped in the N, S, E, or W), have a direct influence on the human body.

One example of directional training is in setting up treatment schedules according to the energy emitted from the Heavens. It is generally advised when conducting energetic regulations that:

- Patients with Yang Deficiencies should face east or south, to develop Yang Qi;
- Patients with Yin Deficiencies should face west or north, to help develop the Yin Qi;
- Patients with a Kidney Deficiency should face the north;
- Patients with a Liver Deficiency should face east;
- Patients with a Heart Deficiency should face south;
- Patients with a Lung Deficiency should face west;
- Patients with a Spleen Deficiency should focus their attention on the center, i.e., the Earth.

Generally, when first beginning Qigong exercises and meditations, practitioners are taught to face the south or southeast. When practicing specific cultivation exercises, however (such as Gathering the Sun's Essence), the practitioners are taught to face the direction of the sun. In the evening time, unless practicing specific cultivation exercises (such as Gathering the Moon's Cream), it is best to have the patient face the corresponding direction for the Deficient organ. These cultivation methods train the Yang energy to fuse with the Zhen Qi (True Qi), as well as with the Heavenly Qi and Environmental Qi. The timing and methods of training are based on the increase and decrease of Yin and Yang energy.

It is important to note that although these training methods are effective, the most experienced Qigong doctors often choose the place and direction according to their own intuition. To select the appropriate direction, the doctors start by facing south (sitting or standing). As they begin to feel the body filling with Qi, they may feel that they are being pulled towards a specific direction, and that they cannot stand or sit comfortably without facing that direction. The doctors feel more rooted and stable by following their intuition. Through understanding the proper directions, concentration, breath regulation, and assuming the appropriate postures, Qigong doctors are able to help regulate and increase their patients' energetic activities, while avoiding many Qi deviations (see Chapter 18).

GATHERING MOON'S CREAM QIGONG MEDITATION

Qigong doctors use this particular exercise in conjunction with the Gathering the Sun's Essence meditation for replenishing depleted energy. The vital essence of moon is called the Cream or "cool light" and is used to nourish the Kidneys and strengthen the Jing, Marrow, and brain, while contributing to longevity. It is practiced only nine days a month, three days before, during, and after the full moon, which are the brightest nights of the month. It is prohibited during new moon transitions because there is too little vital essence of the moon to gather.

Gathering the Moon Cream meditation must only be practiced on clear nights, when the moon is not obstructed. Avoid practicing this meditation in times of fog, smog, thick clouds, and lunar eclipses, as cultivation during these times will cause the unnatural cultivation of Turbid Qi

1. Stand outside in the fresh air, while facing the moon. Stand in a relaxed and tranquil state, breath evenly, and rid your mind of any stray thoughts (Figure 11.22).
2. Lower your eyelids until you only faintly see the moonlight. Inhale through the nose and imagine softly "sipping" one mouthful of the moon cream. At the same time, with your eyes half open, visualize "inhaling" the energy of the moon cream with your eyes into your Upper Dantian. Close your eyes and imagine mixing the moon cream with the energy you inhaled through your nose. Slightly hold your breath and concentrate on feeling the moon cream's silvery energetic essence. Swallow this essence slowly, sending it down to the Lower Dantian.
3. Stretch your hands towards the moon (Figure 11.23). Imagine embracing the moon and pulling it out of the sky, placing it just above your Upper Dantian, at your Baihui point (on top of your head). Hold this image for several minutes (Figure 11.24).
4. Next, imagine a second moon forming out of the first moon, and bring it downward with both hands until level with, and in front of, the Middle Dantian (Shanzhong CV-17 point) at the center of the chest. Hold this image for several minutes (Figure 11.25).
5. Finally, imagine a third moon forming out of the second moon, and bring it down to the level of the Lower Dantian, suspended in front of the lower abdomen. Hold this image for several minutes (Figure 11.26).
6. Continue to imagine all three moons suspended (two in front, and one above your body) and inhale, raising both of your arms over your head. Imagine pressing the first moon into your head through the Baihui point, filling the Upper Dantian with the moon cream's silvery white light. As both hands press the moon into the head, allow the left hand to remain on top of the right hand (opposite for women). As you exhale, visualize the light of the moon filling the entire cavity of the Upper Dantian (Figure 11.27).
7. Next, bring both hands to the front of your body. Inhale and press the second moon into your chest through the Middle Dantian, imagining the moon cream's silvery white light filling the chest. As you exhale, visualize the light of the moon filling the entire cavity of the Middle Dantian (Figure 11.28).
8. Finally, bring both hands in front of your lower abdomen. Inhale and press the third moon into your Lower Dantian, imagining the moon's silvery white light filling the abdomen. As you exhale, visualize the light of the moon filling the entire cavity of the Lower Dantian (Figure 11.29).
9. Allow both hands to rest by your sides then slowly raise your left hand (opposite for women) and tap the top of your Baihui point softly nine times. As you tap, imagine that all three moons fuse into a brilliant white-light energy throughout your Taiji Pole, connecting all three Dantians together (Figure 11.30).
10. Relax and perform Pulling Down The Heavens, letting your body, mind and spirit sink into tranquility.

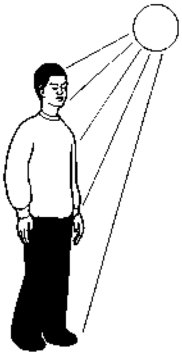

Figure 11.22. Stand facing the moon.

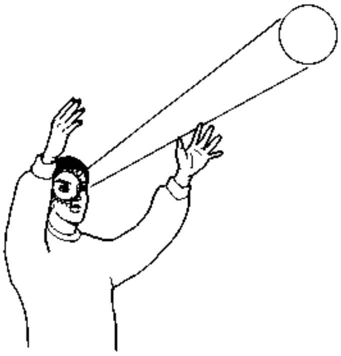

Figure 11.23. Stretch your hands to embrace the moon.

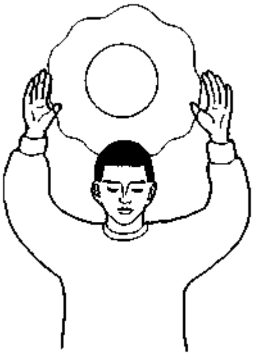

Figure 11.24. Suspend the moon above your Upper Dantian.

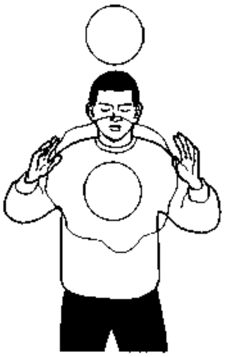

Figure 11.25. Suspend a second moon in front of your Middle Dantian.

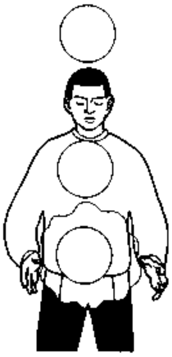

Figure 11.26. Suspend a third moon in front of your Lower Dantian.

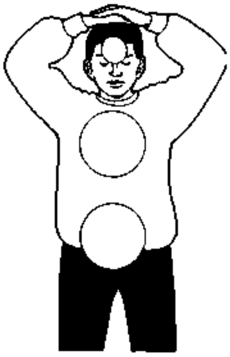

Figure 11.27. Absorb the first moon into your Upper Dantian.

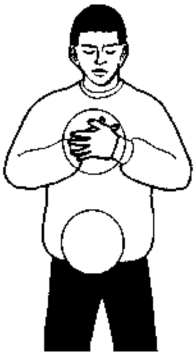

Figure 11.28. Absorb the second moon into your Middle Dantian.

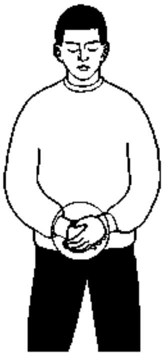

Figure 11.29. Absorb the third moon into your Lower Dantian.

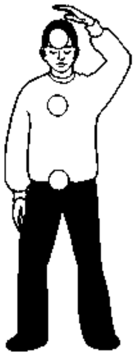

Figure 11.30. Tap lightly until all three moons flow into your Taiji Pole.

CYCLES OF THE STARS

The most influential Heavenly Treasures, other than the sun and moon, are the stars composed of the Big Dipper, Polaris, and the five planets, existing within the twenty-eight constellations.

The North Star is believed to be the unchanging area of Heaven, from which the Heavenly order began (often called the Axis of the Vault of Heaven), and so it was named The Divine Center of the Earth and Sun Cycles. It is also believed that from this star the entire universe expands outward. By placing the North Star in the center of a circle of the Heavens, the circle can be divided into four directions, each containing seven constellations that have a different color energy and sound frequency.

In ancient China, the seven constellations were divided into four sections or directions (north, south, east, and west). In the region of the north was believed to exist the energy of the Tortoise-Snake, in the south was the Red Phoenix, in the east was the Green Dragon, and in the west was the White Tiger. Each of these regions was further divided into twenty-eight constellations or posts (xiu).

According to traditional Chinese belief, the moon travels twenty-eight days from west to east, resting at each post as it circles the Earth. In other words, the twenty-eight days are further divided into twenty-eight regions which are courier stations or travelling houses for the moon.

It was further believed that the Pole Star axis stemming from the North Star traveled downward from the center of Heaven, connecting with the Earth's center core (through the north and south poles). The north pole was regarded as the peak (or Baihui) of the Earth, and was believed to be comprised of seven stars (making it a fifth constellation). The energies of each of the four animals and their seven constellations moved around the Earth's central Pole creating the Five Elemental changes in weather (see Chapter 11). If the weather changes progressed in their expected order, they were considered normal; however, if the weather transition occurred either earlier or later than expected, the result was considered an environmental Evil Qi.

Therefore, the ancient Chinese believed that one could predict the changes of weather by studying the interrelationships between the Five Elements and the cycles of the sun, moon, Polaris, the Five Planets, and the twenty-eight constellations.

THE ENERGY OF THE BIG DIPPER

As the Earth orbits the sun, the angle at which the Big Dipper can be viewed changes on a daily, monthly, and yearly basis (Figure 11.31). These changing angles correspond to the different cycles of the moon as they are arranged in the Twelve Earthly Branches and repeat themselves after completing an orbit (360° rotation of twelve moon cycles). These twelve lunar cycles determine the twelve seasonal periods (Figure 11.32). The beginning part of each month marks the beginning of the seasonal periods of the Twelve Earthly Branches and the beginning of the energy cycle of that month. When the Big Dipper points in the direction of the "lunar month," facing the direction of the Big Dipper will invigorate the practitioner's energy. The Big Dipper is related to the energetic transformations of germinating, beginning, and growing.

There are three kinds of practice centered around the Big Dipper:

1. The first consists of invoking its power for protection;
2. The second consists of ascending into it and pacing its rotation;
3. The third involves gathering the star's essence or star cream into the body.

In ancient times, the Big Dipper was considered the bridge between the energies of the sun and the moon, and the source of "10,000 things." The sun and moon are considered a bipolar energy field, whereas the Big Dipper is considered unipolar.

The sun corresponds to the Heart and Fire energy, which is related to the front area of the body. The moon corresponds to the Kidneys and Water energy, which is related to the back area of the body. The Big Dipper corresponds to the Spleen, which is related to the Taiji Pole.

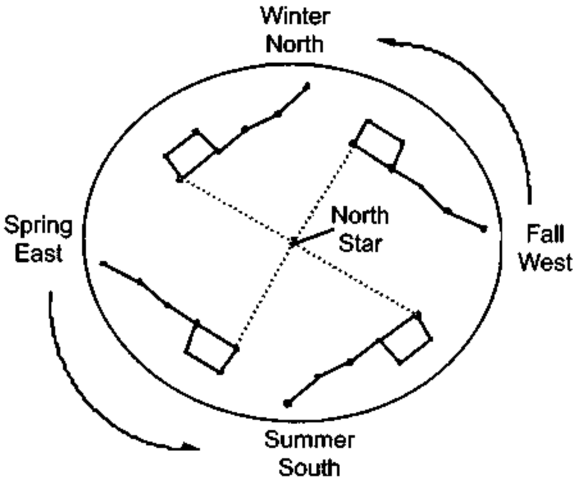
Winter
North
Spring
East
North
Star
Fall
West
Summer
South

Figure 11.31. The handle of the Big Dipper can be used to determine the season: (1) in winter, the handle points to the northern horizon, (2) in spring, the handle points to the eastern horizon, (3) in summer, the handle points to the southern horizon, and (4) in fall, the handle points to the western horizon. The North Star is positioned in the center of the diagram.

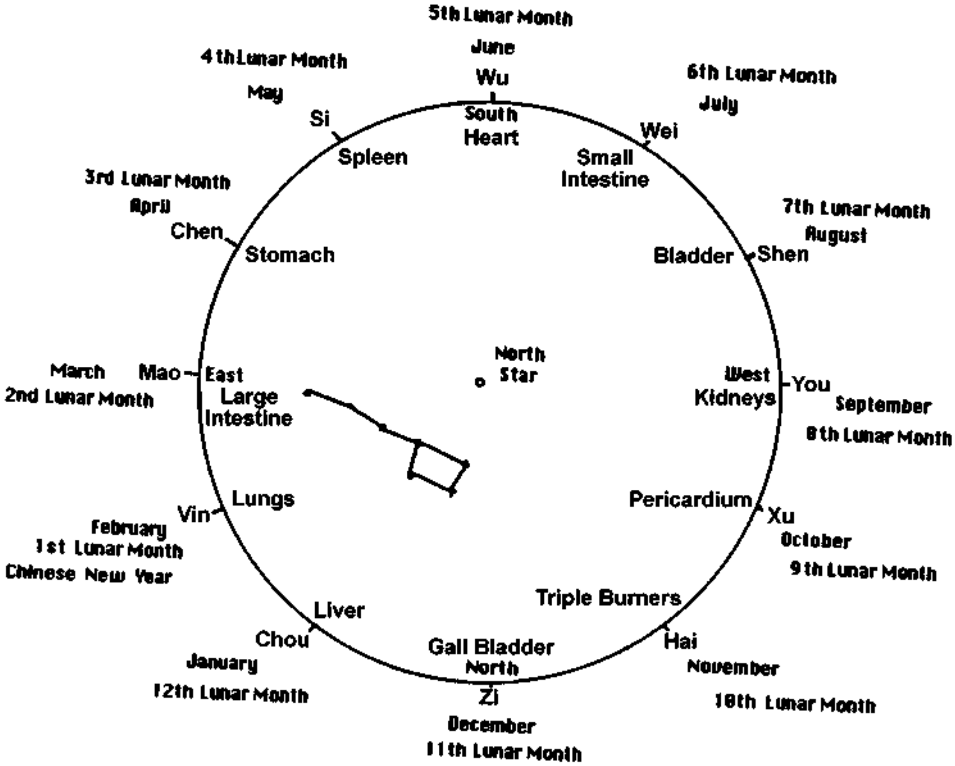
5th Lunar Month
June
Wu
South
Heart
4th Lunar Month
May
Si
Spleen
6th Lunar Month
July
Wei
Small
Intestine
3rd Lunar Month
April
Chen
Stomach
7th Lunar Month
August
Bladder
Shen
North
Star
March
2nd Lunar Month
Mao
East
Large
Intestine
West
Kidneys
You
September
8th Lunar Month
Lungs
Vin
February
1st Lunar Month
Chinese New Year
Pericardium
Xu
October
9th Lunar Month
Triple Burners
Liver
Chou
January
12th Lunar Month
Gall Bladder
North
Zi
December
11th Lunar Month
Hai
November
10th Lunar Month

Figure 11.32. The beginning of the seasonal periods of the Twelve Earthly Branches and the twelve moon cycles is determined by the angle of the Big Dipper.

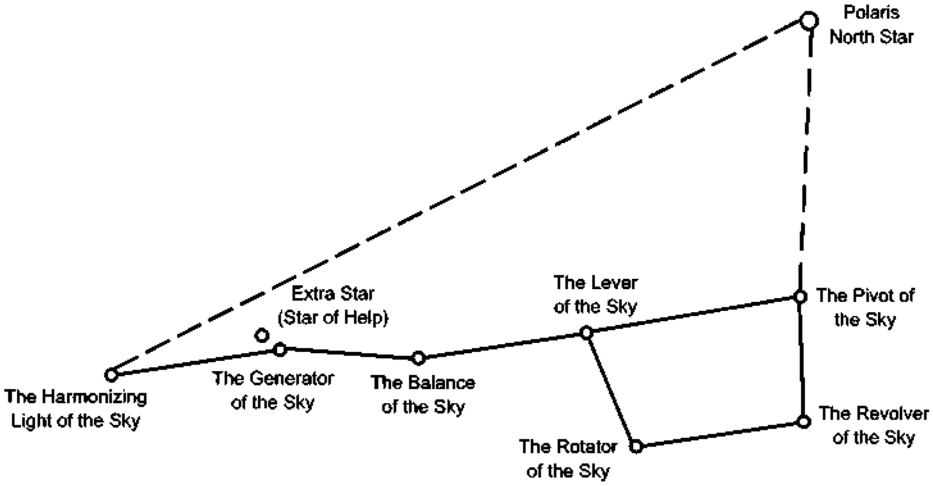

Figure 11.33. The Big Dipper is considered a reservoir of heavenly energy, collecting Qi from all of the other stars and constellations as it moves through the sky.

THE STARS OF THE BIG DIPPER

The Big Dipper represents the foundational base for the Gate of Access in Chinese mysticism, which is considered the spiritual passageway between life and death. Through this gate the individual may obtain knowledge and insight concerning past, present and future events.

The energy from the stars of the Big Dipper is used to balance the body's Five Agents (stored within the body's Five Yin Organs) and can be arranged internally to enhance internal perception of time and space (Figure 11.33).

One example used in ancient China for invoking the powerful protection of the Big Dipper energy is for the Qigong doctor to cover himself - or herself - with the stars of the Big Dipper by visualizing the stars descending and arranging themselves in a specific order around the doctor's body The stars can be arranged in four ways:

- above the doctor's head,
- in front of or behind the doctor's body,
- placed within the doctor's Three Dantians, or
- placed with the bowl of the Dipper in the Lower Dantian and the handle along the spine.

The doctor imagines placing the stars of the Big Dipper into the center of his or her body. The star's energetic function is to govern the front, back, right, and left of the body, thus regulating the doctor's Yin and Yang energy. The star's energetic function can also be used for absorbing energy from the four seasons into the doctor's body.

The stars of the Big Dipper are also considered responsible for opening the Seven Orifices of the embryo's body and for giving the embryo life. These stars also correspond to the Nine Chambers in the Three Dantians. Usually, meditations dealing with gathering energy from the sun and moon are followed by meditations involving the Big Dipper to further stimulate and energize the body's energetic fields.

THE FIRST FIVE PLANETS

The five planets (Saturn, Jupiter, Mars, Venus, and Mercury) are responsible for the internal Yin organ energy formation within the body during fetal development. These planets are the focus of several Qigong meditations because of their connection to color resonations within each internal Yin organ.

1. Of the first five planets in our solar system, Saturn is coupled with Earth and is considered the central force of planetary energy. It takes twenty-eight years for Saturn to make a complete revolution around the sun. Each year, when Saturn enters into another constellation, it is called "Saturn holding authority over that constellation for one year."

   Saturn's position in the sky relates to the Taiji Pole. The planet Saturn represents the Stomach and Spleen (Earth Element) and emits a golden-yellow light energy.
2. Jupiter represents the easterly direction and the Liver (Wood Element), it emits a green light energy; its position relates to the left side of the Taiji Pole.
3. Mars represents the southerly direction and the Heart (Fire Element). It emits a red light energy and relates to the front side of the Taiji Pole.
4. Venus represents the westerly direction and the Lungs (Metal Element). It emits a white light energy and relates to the right side of the Taiji Pole.
5. Mercury represents the northerly direction and the Kidneys (Water Element). It emits a dark blue/indigo light energy and relates to the back side of the Taiji Pole.

GATHERING ENERGY FROM THE FIVE PLANETS

To absorb energy from the five planets, start by quieting your mind while standing in a Wuji posture and facing the direction of the Big Dipper. Visualize yourself absorbing the energetic essence and color of each planet, one by one, into your body, filling each organ's entire Orb (the organs complete system, channels and tissues). Imagine filling the body with each planet's Heavenly energy from the top of your head, through the center Thrusting Channels.

Begin with the energy of Saturn, imagine it flowing into and filling your body through the Taiji Pole. Next, imagine the energy of Mars flowing into your front Center Thrusting Channel. Now imagine the energy of Mercury filling your body through the back Center Thrusting Channel, Venus's energy filling the body by descending into the right Center Thrusting Channel, and Jupiter's energy filling the body by flowing into the left Center Thrusting Channel. Gather the energy from their corresponding Orbs and lead the Qi downward into the Lower Dantian. After collecting and storing the Heavenly Qi, end the meditation with the Fire Cycle of the Microcosmic Orbit meditation.

THE TWENTY-EIGHT CONSTELLATIONS

Since the year 2000 B.C., the ancient Chinese Qigong Masters observed and used the movements and energy of the twenty-eight constellations, the sun and moon, and the first five planets with the Chinese zodiac, to forecast auspicious times for matters of state and war, and to predict weather and natural disasters. Each of these energetic systems has its own cyclic pattern as well as a direct and indirect influence on human life.

The Chinese practised gathering energy from the Heavenly stars, believing that the study of astrology, like alchemy, united the body's inner world with that of the outer. According to ancient divination practices, the celestial bodies exert forces and exhibit personalities that influence people and events on Earth. A person's character and destiny throughout life is based upon the positions of the planets and constellations at the exact time and place of birth. These influences can be determined by mapping the positions of the twenty-eight constellations in the sky at various times.

So important was this belief, that in ancient China, the Emperor was considered the high priest of the Heavens and frequently made sacrifices to the stars to maintain harmony with the universe. The four corners of the Emperor's palace represented the cardinal points in space (the equinoxes and solstices), and he and his family would move from one corner to another as the seasons changed.

GATHERING ENERGY FROM THE CONSTELLATIONS

At night time, when the moon is not suitable for absorbing Qi (new moon), the Qigong doctor can absorb the energy of the twenty-eight constellations into his or her body.

The energy of all five planets and twenty-eight constellations can be drawn into the body's internal organs and tissues through the five energetic Thrusting Channels which envelop and support the body's Taiji Pole. These Channels flow through the center of the doctor's body, connecting the top of the head to the perineum.

Externally, the twenty-eight day constellation cycle can be arranged externally on the body's Governing and Conception Vessels. The "ecliptic" path of the sun follows the Fire Cycle of the Microcosmic Orbit, which correlates to the individual's emotional/mental and physical health.

There are several meditations used in order to absorb the energy of the five planets and twenty-eight constellations into the doctor's body; one of which is known as The Heavenly Five Yin Organ Meditation.

HEAVENLY FIVE YIN ORGAN MEDITATION (TIAN WU ZANG)

This meditation is very popular in China. It focuses on fusing the energy from the five planets and the constellations into the body through the Five Thrusting Channels that surround the core center of the body's Taiji Pole. These Five Thrusting Channels should not to be confused with the TCM perspective of the Five Thrusting Vessels (as seen in Figure 11.34). Each of these Five Thrusting Channels relates to a specific organ energy and element (Figure 11.35).

When practicing this meditation and gathering the energy from the stars, it is important to first absorb the energy into the Five Thrusting Channels, then spiral the energy in a clockwise downward and counterclockwise upward direction through the body's Taiji Pole. This spiralling action allows the energy to balance itself and stabilizes the body's internal organs (Figure 11.36).

In order to perform the Heavenly Five Yin Organ Meditation and gather the Qi of the twenty-eight constellations, assume a Wuji posture, with your arms suspended by your sides. Face the direction of the North Star and Big Dipper (Figure 11.37). This exercise is performed for thirty-six breaths in each direction.

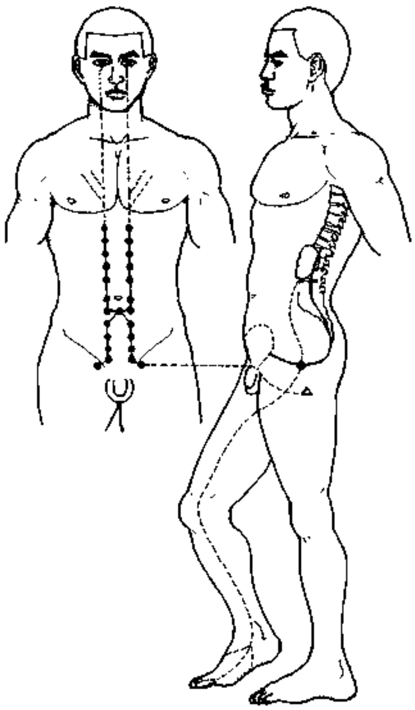

Figure 11.34. The Thrusting Vessels

1. To begin, imagine golden energy pouring out of the Big Dipper and flowing into the center of the Baihui area, at the top of your head. As you inhale, this golden energy fills the center core of your body, and flows into the extremities. As you exhale, this golden energy gathers into your Spleen.
2. Next, imagine drawing in a red energetic mist from the seven constellations in front of your body. This red mist flows into your body from the front side of the Baihui area, filling your entire body as you inhale. As you exhale, the energy coalesces into your Heart.
3. Now focus your attention on the back of your body. Inhale and imagine drawing in the dark blue energy from the seven constellations behind you into the back side of the Baihui area, filling your entire body as you inhale. This energy gathers into and fills the Kidneys as you exhale.
4. Next, focus on the right side of your body and imagine drawing in white energy from the

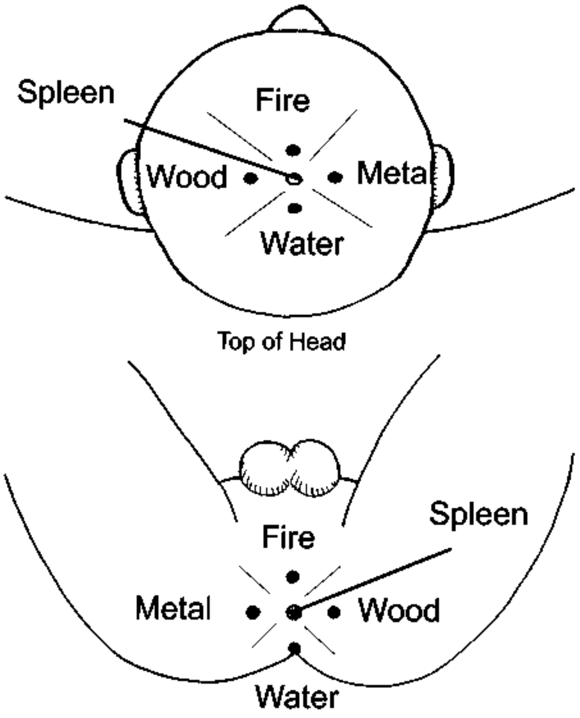
Spleen
Fire
Wood
Metal
Water
Top of Head
Spleen
Fire
Metal
Wood
Water

Figure 11.35. Each of the Center Thrusting Channels relates to a specific organ energy and element. Each channel has its own entry point that is positioned either on the top of the head, or on the base of the perineum.

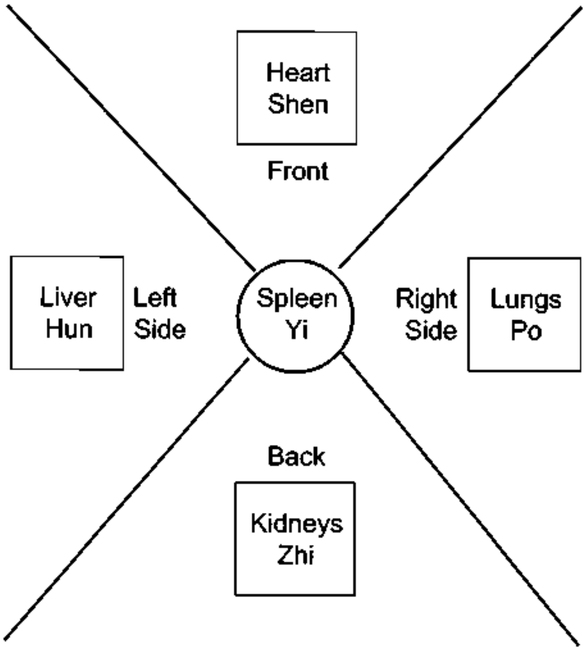
Heart
Shen
Front
Liver
Hun
Left
Side
Spleen
Yi
Right
Side
Lungs
Po
Back
Kidneys
Zhi

Figure 11.36. The body's Yuan Spirit is stored within the body's Five Yin Organs. The placement of the Five Energetic Houses is maintained throughout many Medical Qigong meditations and is used to tonify, strengthen, extend, and project the body's Shen.

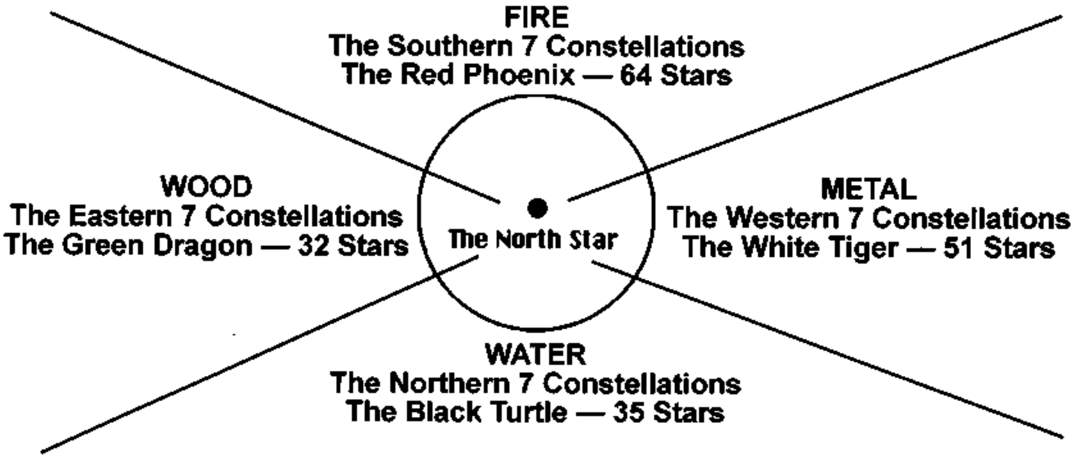
FIRE
The Southern 7 Constellations
The Red Phoenix — 64 Stars
WOOD
The Eastern 7 Constellations
The Green Dragon — 32 Stars
The North Star
METAL
The Western 7 Constellations
The White Tiger — 51 Stars
WATER
The Northern 7 Constellations
The Black Turtle — 35 Stars

Figure 11.37. The energy of the Twenty-Eight Constellation Stars is absorbed into the body through the same structural pattern as the placement of the Five Energetic Houses of the body's Shen.

seven constellations to your right into the right side of the Baihui area, filling your entire body as you inhale. This energy gathers and fills the Lungs as you exhale.

5. Finally, focus your attention on the left side of your body and imagine drawing in green energy from the seven constellations on your left into the left side of the Baihui area, filling your entire body as you inhale. This energy gathers and fills the Liver as you exhale.

EARTHLY BRANCHES, HEAVENLY STEMS, AND NINE PALACES

There are many relationships between the macrocosm and the microcosm of man, through the Twelve Earthly Branches, the Ten Heavenly Stems, and the Nine Palaces outlined in the Chinese Classics on Internal Medicine. The Twelve Earthly Branches are Earth energies and are represented in Man as the Twelve Primary Channels. The Ten Heavenly Stems relate to the energies of Heaven and are represented in man as the Yin and Yang aspects of the Five Element organs of Wood, Fire, Earth, Metal, and Water (as seen in Figure 11.38). The Nine Palaces are related to the Eight Extraordinary Vessels (corresponding with the Eight Postheaven Trigrams) in addition to the body's energetic Taiji Pole.

The Twelve Earthly Branches, Ten Heavenly Stems, and the Nine Palaces are interwoven into almost all aspects of Chinese culture, arts and medicine. In Chinese divination, as well as Chinese medicine, the temporal cycles of the Earthly Branches and Heavenly Stems form the foundation for these practices. They can be used to explain the physiology, pathology, diagnosis, and treatment of each clinical case. For each year, month, day, and hour there is a Stem and Branch association. By studying the Stem and Branch association, the Qigong doctor can identify patterns (i.e., determine the cycles and diagnose the patterns of the patient's disharmony) and design treatment protocols, depending on the time of year and the time of day.

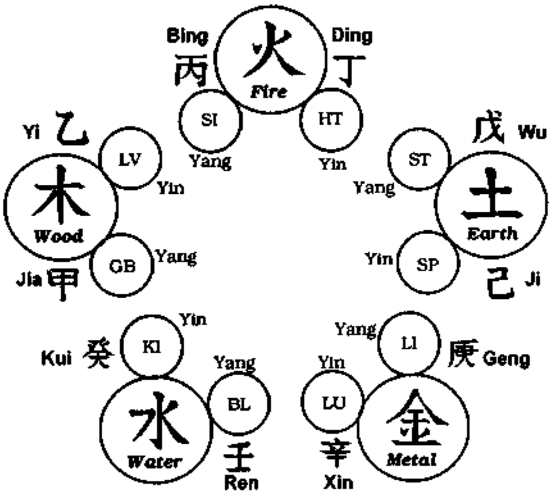

Figure 11.38. The Five Element Chart and Its Relationship to the Ten Celestial Stems

The yearly cycles of the Stems and Branches are based on a sixty-year cycle, when the sun, moon, Earth, and North Star are aligned (Figure 11.39). The last sixty-year cycle began in 1984. The exact origin of the Ten Heavenly Stems and Twelve Earthly Branches is not known. They were used, however, to mark the hours, days, years, and planetary cycles as early as 1500 to 2000 B.C. Some historians attribute their origin to the legendary emperor Huang-Di (as early as 2600 B.C.). Each cycle was regarded by the Chinese as being a "century" in the life span of the average man. Those who lived past the age of sixty were regarded as living a "second life." During this second life, the person would experience similar occurrences as in the "first life," but would be able to apply the wisdom and maturity gained from his or her first life. The individual can then further his or her spiritual growth and use the knowledge to benefit society as a whole.

The twelve animals of the Chinese zodiac correlate with the Twelve Earthly Branches and are arranged as follows.

- Zi branch corresponds to the Rat
- Chou to the Ox
- Yin to the Tiger
- Mao to the Rabbit

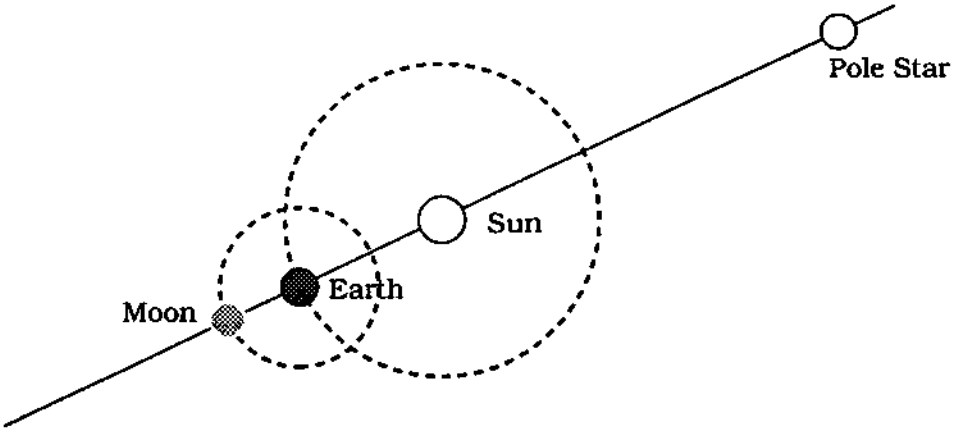

Figure 11.39. The beginning of the sixty-year cycle is based on the alignment of the sun, moon, earth, and North Star.

- Chen to the Dragon
- Si to the Snake
- Wu to the Horse
- Wei to the Sheep
- Shen to the Monkey
- You to the Rooster
- Xu to the Dog
- Hai to the Boar

The first year of the sixty-year cycle is associated with the first Branch and the first Stem. The first Branch, "Zi," corresponds with the Rat; the first Stem, Chia, corresponds with the element Wood. Thus, the first year of the sixty-year cycle is the year of the Wood Rat in Chinese astrology.

The second year of the cycle represents the second Branch and second Stem, and so on. When you reach the 11th Branch, there are no more Stems and, thus, the 11th year of the cycle represents the eleventh Branch and the first Stem. The cycle continues in this manner until it has transitioned through six cycles of Stems and five cycles of Branches and returns to the first Branch and first Stem paired together.

Figure 11.40 shows these yearly cycles; for example, 1997 was the year of the Fire/Ox, so the year 1998 is the year of the Earth/Tiger.

The Ten Heavenly Stems represent Yang, and the higher form of Heaven's vitality (Figure 11.41). The Twelve Earthly Branches represent Yin and the lower form of Earth's vitality (Figure 11.42). Together they manifest the transformation of all natural events expressed through the growing, thriving, declining, and dying of all living things.

1. Each lunar year the Earthly Branch time divisions are divided into 12 separate month divisions.
2. Each day is divided into 12 two-hour Earthly Branch time divisions during a 24-hour cycle.

SUMMARY

In comprehending the energetic potential of the Three Treasures of Heaven, the Qigong doctor evaluates the patient's condition and treats the disease according to the patient's energetic rhythm and flow, which is blended with the Heavenly and Earthly energies. The treatment also includes prescribing homework (the Medical Qigong prescriptions needed to support and maintain the doctor's treatments).

Although mankind is terrestrial, there can be no separation from the influence of the celestial. For mankind is born on Earth in physical form, created through the integration of both celestial (Heavenly) and terrestrial (Earthly) energies. According to statements recorded in the *Canon of Internal Medicine,* ancient doctors had already rec-

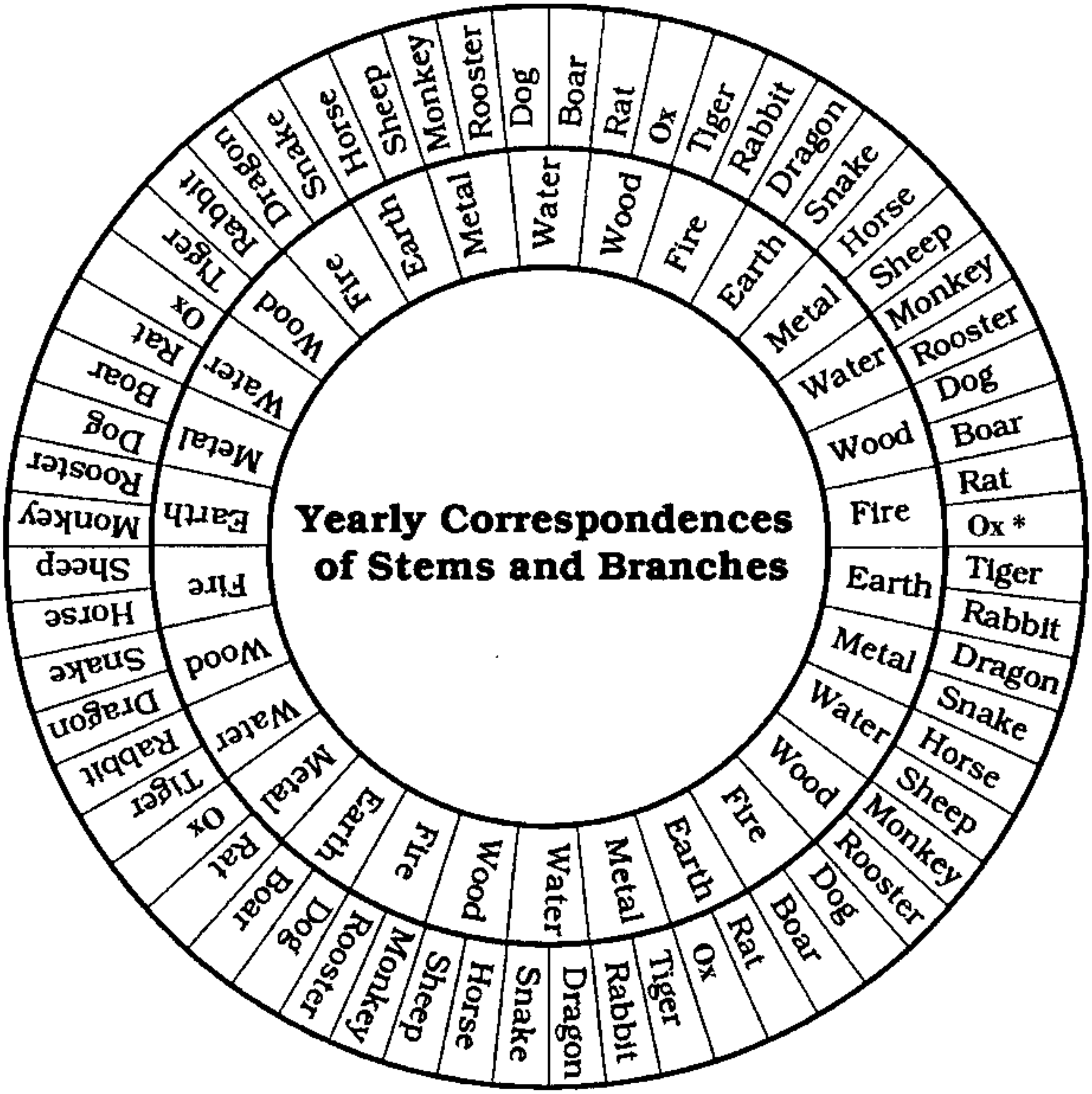
Yearly Correspondences of Stems and Branches
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Dog
Boar
Rat
Ox
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Boar
Rat
Ox *
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Boar
Rat
Ox
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Boar
Rat
Ox
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Boar
Rat
Ox
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster

Figure 11.40. The asterisk (*) shows the year 1997, which was the year of Fire/Ox. Each year, one of the twelve animals of the Chinese zodiac flourishes and brings its energy into action. Every two years, a new Element (Fire, Wood, Water, Earth, Metal) dominates in the cycle of creation. The chart moves clockwise, so 1998 was the year of Earth/Tiger.

Character	Name	Number	Animal	Month	Time	Branch	Channel	Image	Element
子	Zi	1	Rat	Nov. 22–Dec. 21	11-1 am	**Little Yang**	Gall Bladder	Reproduction of Life	**Water**
丑	Ch'ou	2	Ox	Dec. 22–Jan. 20	1-3 am	**Decreasing Yin**	Liver	Initial Growth	**Earth**
寅	Yin	3	Tiger	Jan. 21–Feb. 19	3-5 am	**Great Yang**	Lungs	Breaking Ground	**Wood**
卯	Mao	4	Rabbit	Feb. 20–Mar. 20	5-7 am	**Bright Yin**	Large Intestine	Thick Vegetation	**Wood**
辰	Ch'en	5	Dragon	Mar. 21–Apr. 19	7-9 am	**Bright Yang**	Stomach	Full Awakening	**Earth**
巳	Si	6	Snake	Apr. 20–May 20	9-11 am	**Great Yin**	Spleen	Preparation for Ripening	**Fire**
午	Wu	7	Horse	May 21–June 21	11-1 pm	**Little Yang**	Heart	Peak of Growth	**Fire**
未	Wei	8	Sheep	June 22–July 21	1-3 pm	**Great Yin**	Small Intestine	Taste of Full Ripeness	**Earth**
申	Shen	9	Monkey	July 22–Aug. 21	3-5 pm	**Great Yang**	Bladder	Time of Harvest	**Metal**
酉	You	10	Rooster	Aug. 22–Sept. 22	5-7 pm	**Little Yin**	Kidneys	Resting after Harvest	**Metal**
戌	Xu	11	Dog	Sept. 23–Oct. 22	7-9 pm	**Decreasing Yang**	Pericardium	Moving Inward	**Earth**
亥	Hai	12	Boar	Oct. 23–Nov. 21	9-11 pm	**Little Yin**	Triple Burners	Seed Awaiting Growth	**Water**

Figure 11.41. Environmental Energetic Interactions of the Twelve Earthly Branches

ognized the close relationship between the body's Qi and Blood and the influences of the natural environment. In ancient China, the Twelve Earthly Branches represented time units, each of them corresponding to a two-hour period. They are primarily used to represent the 12 months in the Lunar Calendar.

The Ten Heavenly Stems are made up of the Five Elements, each of which is divided into pairs of corresponding Yin and Yang, resulting in a ten-step system of numerology. Using the Ten Heavenly Stems allows the Qigong doctor to determine the ebb and flow of the Qi of Heaven and its corresponding relationship to the Qi of Man.

Both the Ten Heavenly Branches and Twelve Earthly Stems represent the characteristics of growing, thriving, declining, and dying of all living things in the universe, as well as the development and transformation of all natural phenomena.

The energy of the Ten Heavenly Stems can be observed within five energetic movements (front, back, right, left, and center), as well as within the elemental energy of the body's major internal organs.

CHAPTER 12

THE THREE TREASURES OF EARTH: EARTH, WATER, AND WIND

INTRODUCTION TO THE THREE TREASURES OF EARTH

The Outer Force (Power) of Earth's Three Treasures consists of the Yin and Yang interaction of energy and light, emitted, absorbed, and reflected back from the land, water, and wind. The Three Earthly Treasures are brought about by the waxing and waning of the five Yin and Yang climatic transformations (the five seasonal growth cycles), which in turn are caused by the sun, moon, and stars. Each seasonal change brings a life transforming gift that affects the mind, body, emotions, and spirit (Figure 12.1). The wind (weather) reflects the various conditions and transformations of energy in the sky, and is viewed as an Earthly manifestation of Heaven's moods.

1. Earth Qi consists of the Yin and Yang interactions of energy and light, manifesting as hot and cold energy originating from the Earth's surface.
2. Water Qi consists of the Yin and Yang interactions of energy and light, manifesting as hot and cold energy originating from the vast oceans, lakes, and rivers of the world.
3. Wind Qi consists of the Yin and Yang interactions of energy and light, manifesting as hot and cold air circulation created from the interaction of the sun's heat on the surface of the planet. This interaction includes all cloud and barometric pressure formations (e.g., tornadoes, hurricanes, lightning storms, etc.).

UNDERSTANDING HEAT AND LIGHT INTERACTIONS

Heat and light are manifested both externally on and from the Earth's surface, and internally within the Earth.

1. Externally, the greater the intensity of the sun's light absorbed into the surface of the Earth, the more heat is manifested from the Earth's surface. Cloudy days diminish the

Natural Earthly Transformations					
Organs	**Element**	**Season**	**Elemental Factor**	**Temperature**	**Transition**
Liver	Wood	Spring	Wind	60– 70° F	Germination
Heart	Fire	Summer	Heat	above 90 F	Growth
Spleen	Earth	Late Summer	Damp	75– 90° F	Transformation
Lungs	Metal	Autumn	Dry	60– 70° F	Gathering
Kidneys	Water	Winter	Cold	below 60° F	Storage

Figure 12.1. The Natural Earthly Transitions

sun's thermal interaction and tend to bring coolness.

2. Internally, heat and light, as well as other natural forces, are contained within the Earth itself. The sun's intense energetic fields draw the heat and light from within Earth, which are then reflected back to the Earth by the atmosphere.

The Earth's air extends only a comparatively short distance out from the Earth's surface, and the effects of the heat rays decrease as you leave the Earth's surface and ascend toward the outer regions of the atmosphere. As the air becomes less dense, there is less reflection. The higher you ascend, the cooler the air becomes. When you have reached the limit of air, you have reached the limit of the Earth's heat.

The heat and light form a protective energetic field around the Earth's surface similar to the body's Wei Qi field. Once you have reached the limit of heat and light, you reach what is called the Great Cold. This cold is considered far more solid than steel, and it presses down upon the Earth's energetic field and atmosphere with an almost irresistible force, holding them together.

UNDERSTANDING THE EARTH'S YIN AND YANG INTERACTIONS

Between Heaven and Earth there is an on-going interaction of Yin and Yang energy exchange. When the Earth Qi is in balance, plants grow and animals thrive. Power, in the form of energetic fields, is created by the integration of Yin and Yang. This power directly and indirectly affects the body's electromagnetic field, as well as the regulation and formation of the body's tissues. Visualize the Earth's energy as Qi resonating and emitting from the Earth itself. This energy includes the electromagnetic fields, subterranean radiations, and thermoluminescence (light and heat emitting from the center of the Earth).

The closer the Qigong doctor bonds physically, mentally, emotionally, and spiritually with the Earth and nature, the easier it is to tap into the Earth's energy and vibration. The development of Earth Qi in the body and the body's connection to the energy of the Earth are mutually supportive. The first step in internal cultivation involves focusing the mind on the Lower Dantian (with the aid of breathing techniques) to enhance the mind-body connection; this interactive connection develops a relationship with the Earth Qi.

GATHERING ENERGY FROM THE EARTH'S ENERGETIC FIELDS

The ability to gather energy from nature and the environment is extremely important to the Qigong doctor. The energy of the Earth stays within the structural components of the planet, flowing like mighty rivers of Qi. It is absorbed into the terrain and vegetation, as well as the water that we drink. Environmental Qi is specific to each geographic area. It is the energy that has been collected from the structural components of the Earth itself (i.e., soil, fields, deserts, plants, trees, forests, mountains, streams, rivers, lakes and oceans, etc.). The energetic potential of the Earth's Qi changes from Yang to Yin energy with the rising and setting of the sun, and therefore directly affects the environmental energy.

Although food and water are excellent sources for gathering energy, the Qigong doctor cannot continually ingest food, or drink herbal teas while treating patients, without overtaxing his or her digestive system. It is, therefore, important for the doctor to find an alternative source of energy, to replenish his or her system once it becomes depleted. By studying the waxing and waning of the Earth's Yin and Yang energies the doctor can optimize the absorption of Yin and Yang energy.

TECHNIQUES FOR CULTIVATING EARTH QI

Generally, several meditations are used to cultivate and absorb energy from the Earth by utilizing the energy of trees, bushes, flowers, lakes, etc. The primary objective of these meditations is to first cleanse the body of pathogenic factors, then cultivate and absorb the environmental Qi to replenish the body's energetic field. Practitioners and patients should visualize themselves immersed in a focused environmental field of energy, filled with its vibration, color, and light. They should allow this energized field to envelop their entire body.

CLEANSING THE BODY

One of the most popular cleansing techniques used in the Xi Yuan Hospital's Qigong Clinics begins from a lying, sitting, or standing posture (usually in a Wuji posture) while in front of a natural object, such as a tree.

To cleanse the body's energy, extend both hands towards the tree, as if to embrace it (it does not matter if the tree is actually being touched). Upon inhalation, begin to guide the energy from the tree into your torso through the Laogong (Pc-8) points at the center of your palms, and the Baihui (GV-20) point at the top of your head. Absorb this energy deep into your body. Once you feel that the body is fully saturated with Qi, exhale through the mouth and guide the Turbid Qi down the legs and out of the body through the Yongquan (Kd-1) points at the bottom of the feet. Continue to guide the Turbid Qi down through the Earth into the root system of the tree. Then inhale through the nose, as you absorb the purified energy from the upper part of the tree leading it down your body through the top of your head. Continue this process until you feel totally cleansed. Remember to synchronize breathing in through your nose as you focus on absorbing the Earth Qi, and to exhale through the mouth, as you focus on dispersing and releasing the Turbid Qi.

For tonification, extend your Qi and intention deep into the root system of the tree and imagine absorbing its natural energy through the Baihui point at the top of your head (Figure 12.2).

GATHERING QI FROM TREES, BUSHES, AND FLOWERS

Earth Qi and environmental energy saturate all types of foliage, allowing the Qigong doctor to select, gather, and absorb Qi from many botanical sources. Throughout history, the Chinese have used trees, bushes, and flowers for healing and medicine. Trees and plants absorb air, light, energy, water, and minerals. When Heavenly Qi is combined together with the environmental energy of the tree or plant, the result is a powerful invigorating energy source.

In China, I noted that every Medical Qigong clinic I visited had its own garden. Each garden was full of healthy trees, bushes, and beautiful flowers from which the patients could draw energy. The patients could choose from different plants and bushes, each type of foliage having its own unique strengths and benefits. Any combination of healthy plant energies can be absorbed by the patients to improve their health.

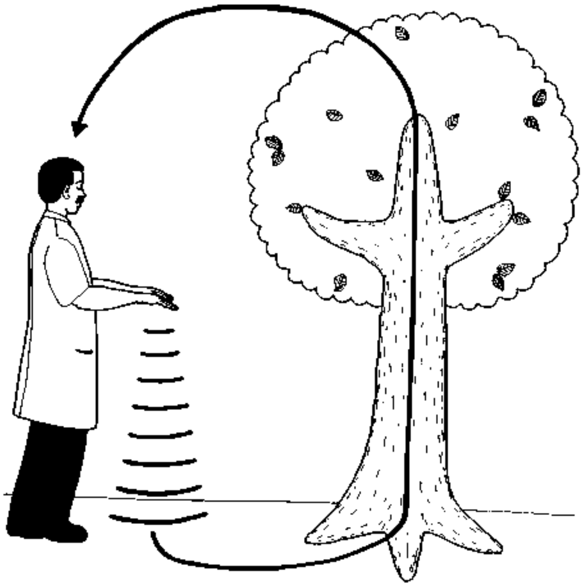

Figure 12.2. A Qigong doctor can extend Qi and intention deep into the root system of the tree and absorb its natural energy through the Baihui point at the top of the doctor's head.

The best time for absorbing energy from trees, bushes, and flowers is between the hours of sunrise (Mao: 5 - 7 am.) and noon (Wu: 11 a.m. – 1 p.m.).

It is important to note that, within the differing regions of a specific country, each type of tree and plant will have either minor or major variances regarding its energetic potential. This difference in energetic potential is due to the diverse environmental influences affecting each plant. We in the West, for example, have many different varieties of plants and trees similar to those in Asia, however, because of the energetic continental differences, the quality and quantity of Qi in each plant or herb can vary. This energetic difference is brought about by the various Qi fields produced by each region's altitude, the mineral content of the soil and water, and the amount of exposure to the energy of the sun. Even herbs of the same name, but grown in different countries, can have different affects on the body's energetic system, slightly altering their clinical potential (e.g., American Ginseng is sweet, slightly bitter and cold, and enters the body affecting the Heart, Lung and Kidney Channels; Siberian Ginseng is sweet, slightly bitter and slightly warm, and enters the body affecting the Lung and Spleen Channels).

GATHERING ENERGY FROM TREES

Trees are tremendously powerful plants and are commonly used by both doctors and patients to absorb and transform the negative pathogenic Qi into clean, healthy, life-giving energy. Pine trees tend to be one of the most powerful, as they radiate a great deal of Qi. Once an individual has connected with a tree, the primary goal is to:

1. remove stagnation from the channels,
2. tonify the internal organs,
3. stabilize and replenish any depleted energy within the body,
4. nourish the Blood, and
5. strengthen the nervous system.

The following is a brief description of various trees, their Yin or Yang potential, and their energetic affect on the body's internal organ system:

- Apple - slightly Yin - St and Sp
- Bamboo - Yin - Ht, Lu, GB, and St
- Birch - Yin - St
- Cassia (Cinnamon) - Yang - Sp, Kd, and Bl
- Cherry - Yang - Lv, Lu, Kd, and St
- Chestnut - Yang - Sp, St, and Kd
- Cranapple - neutral - Lv, Lu, and Ht
- Elm - neutral - St, SI, and LI
- Fig - neutral - Lu, LI, and Sp
- Ginkgo - neutral - Lu, Kd, and Ht
- Hawthorn - slightly Yang - Sp, St, Lu, and LI
- Magnolia - Yang - Lu, St, GB, LI, and Sp
- Maple - neutral - Sp, Kd, and LI
- Mulberry - Yin - Lu, LI, Sp, and Kd
- Oak - slightly Yang - LI, Sp, Kd, and Ht
- Pine - Yang - Ht, Lu, LI, and Kd
- Willow - Yin - Ht and Sp

GATHERING ENERGY FROM BUSHES

Bushes are an extremely effective source for gathering and ingesting different forms of healing energy. Similarly to trees, each bush has its own energetic properties and causes a specific reaction in the body's energetic field. Bushes, however, are not as powerful an energy source as trees for gathering massive quantities of Qi.

GATHERING ENERGY FROM FLOWERS

Flowers are unique in their effect on stimulating the nervous system. Different colors, shapes, and sizes can affect the emotions, causing the individual's spirit to open. Each color can be absorbed into the body in order to facilitate the stimulation of internal organ energy. The gift of flowers has had a long history in soothing emotional pain and bringing relief to patients.

CULTIVATION AND REGULATION TECHNIQUES

Techniques used in the Qigong clinics for cultivation, absorption and regulation of energy generally begin by having the patient face a tree, bush, flower, etc. Although the hands and body postures are similar to the last meditation, the mind's intention is quite different. In this meditation the patient absorbs Qi from the tree, by inhaling it through his left palm, to draw the energy into his body. The energy is guided down the patient's chest, into his or her Lower Dantian. From there the patient circulates this gathered energy through his or her Microcosmic Orbit. As the patient exhales, the Qi exits the right palm into the tree, and begins the cycle again. This meditation is practiced to replenish the body's energetic field.

Patients should be monitored regularly to prevent Qi deviations, which may occur while absorbing Qi through their pores. For best results, choose a quiet, safe, and healthy environment. Select trees, bushes, and flowers with a stable root, or thick trunks. The color is also very important, as the leaves and flowers should be bright, full, and healthy.

Because the body's internal energy matches that of the external environment, it is important to avoid selecting any area which is unpleasant or unhealthy. Avoid any tree, bush, or flower that is sick, dying, has lost its color, or has just been pruned. Using trees, bushes, or flowers for cleansing the body, as well as to absorb energy, is forbidden at night, because their oxygen production diminishes after the sun has set.

PRECAUTIONS

It is important not to meditate in front of any tree, bush, or flower that has parasites, or has been poisoned or polluted, since such vegetation will induce an impure energetic resonation within the Qigong doctor's body.

GATHERING QI FROM MOUNTAINS, VALLEYS, AND DESERTS

The mountains, valleys, and deserts absorb and release light, energy, and heat from the sun. These energies are absorbed and released from the structures of the land quicker than from the oceans, lakes, and streams. Environmental energy is also gathered and stored by the geological formations and local ecology. The ancient Chinese Qigong masters viewed the Earth as a living entity, with rivers and pockets of energy similar to that of the human body. Certain altitudes, densities of colors, and structural formations can affect the Earth's energetic potential.

GATHERING ENERGY FROM MOUNTAINS

Mountains are extremely powerful conduits of energy. They act as peak areas, or collection points for energy. The higher the mountain, the more the air is charged with electromagnetic potential. In ancient times, for prayer and meditation, Taoist and Buddhist temples, shrines, and caves were constructed on energetic pockets high within the mountains of China. In fact, all high, mountains that are far from human civilization are regarded as superior places for energetic cultivation, due to the clarity of light and extraordinarily potent energies which include a very high negative ion count in the air.

GATHERING ENERGY FROM VALLEYS

The Earth's valleys act as channels for Earth energy, collecting and carrying energy throughout the natural terrain. Systems of valleys extend through the plains, hills, and mountains, and generally flow towards the ocean. By standing at the "floor" of the valley (surrounded by its walls), energy can be easily absorbed into the body for cultivation.

GATHERING ENERGY FROM DESERTS

The deserts are considered seas of dry heat energy. They are excellent environments for meditations when used to gather Qi into the body, especially for combating diseases pertaining to Wind, Cold, and Damp invasion (which give rise to many diseases).

CULTIVATION TECHNIQUES

Generally there are several meditations used to cultivate and absorb energy from mountains, valleys, and deserts. The primary objective of these meditations is either to cleanse the body of pathogenic factors, or to cultivate and absorb Qi. When cultivating energy from mountains, valleys, or deserts, practitioners should visualize themselves immersed in the Earth's field of energy, filled with vibration, color, and light. The practitioners should allow this energized field of Earth Qi to envelop them completely.

For best results, select mountains, valleys, and deserts with a stable foundation. The color is also very important, as the Earth, soil, rocks, and surrounding area should be clean.

Also, there are certain locations on the planet known as "power spots" which can produce remarkably powerful effects in the body's energetic fields.

Patients should be monitored regularly in order to prevent Qi deviations, which may occur while absorbing Qi through their pores.

PRECAUTIONS

Because the body's internal energy matches that of the external environment, it is important not to meditate in front of any mountain, valley, or desert area that has eroded, is dying, has lost its color, or is polluted. Also prohibited are areas where there is ongoing seismic, or volcanic activity, as such areas will induce unstable resonation within the Qigong doctor's body.

GATHERING QI FROM OCEANS, LAKES, AND STREAMS

The oceans, lakes, and streams retain and release the sun's light, energy, and heat slowly and are, therefore, important in energy cultivation. The body instantly absorbs the light, energy, and resonant vibrations stored within water. Energized water (sometimes in the form of herbal teas) is often used by Qigong doctors in clinics throughout China. This energetically transformed water provides the patient's body with the basic components necessary for restoration and regeneration, and is considered the "elixir of life" by many Qigong doctors.

GATHERING ENERGY FROM OCEANS

The energy from the ocean has a strong cleansing and purifying effect on the body's Qi. Meditating by the ocean is used extensively for dispersing negative emotions and regulating the body's internal organs. The active rhythm and sound of the waves allows the patient to harmonize his or her internal energetic rhythms with that of the tide, thus quieting the nervous system.

GATHERING ENERGY FROM LAKES

The energy from a lake has a strong calming affect on the body's energetic field. The quiescent solitude of a lake can be used to sedate active emotions, balance any Excess or Deficiency, and calm the body's Shen.

GATHERING ENERGY FROM STREAMS

The energy of a stream can be used to replenish and restore depleted Qi, sedate active emotions, balance any Excess or Deficiency, and also to calm the body's Shen.

CULTIVATION TECHNIQUES

Generally there are several meditations used to cultivate and absorb energy from oceans, lakes, and streams. The primary objective of these meditations is to either cleanse the body of pathogenic factors, or cultivate and absorb Qi to replenish the body's energetic field. When cultivating energy from oceans, lakes, or rivers, patients should visualize themselves immersed in a pool of water, filled with energy, color, and light. The patient should allow this energized water to absorb, cleanse, and invigorate the entire body.

Patients should be monitored regularly to prevent Qi deviations, which may occur while they are absorbing Qi through their pores. For best results, select oceans, lakes, and streams with a calm, stable current. The color is also very important, it should be clear and clean. The water and the surrounding areas should also be clean.

PRECAUTIONS

Because the body's internal energy matches that of the external environment, it is important to avoid selecting any oceans, lakes, or streams that are turbulent, have become dull in color, polluted, or stagnant.

GATHERING QI FROM THE SIX DIRECTIONS

Absorbing Qi from the six directions (Heaven, Earth, and the Four Winds), focuses on gathering environmental energy into the center core of the body (Taiji Pole) from: Heaven, Earth, and the four directions: South corresponds to the front of the body, North to the back of the body, West to the right of the body, and East to the left of the body. The Taiji Pole is considered the central axis, with two directions, above and below. This is a tonification exercise that is used for gathering and absorbing the Earth's environmental Qi, to energize a weakened or Deficient condition.

CULTIVATION TECHNIQUES

There are several meditations used to cultivate and absorb energy from the four directions of the horizon, with the fifth and sixth directions being at the top and bottom of the Taiji Pole. The primary objective of these meditations is to first cleanse the body of pathogenic factors, then to cultivate and absorb environmental Qi. Each direction of energy (N, S, E, W, Heaven, and Earth) has its own unique field of power and vibration, and is assigned a specific color. Each vibration and color stimulates the energy field of one of the five major Yin organs through the body's Heaven and Earth connection (i.e., the Taiji Pole).

When cultivating energy from the horizon, practitioners should face the direction of the south and visualize themselves surrounded in an energetic mist. This mist contains six different colors and light vibrations. From the four directions (N, S, .E, and W) emanate the corresponding following four colors: ruby red, brilliant white, indigo and emerald green. Each of these colors are associated with a particular Yin organ: ruby red with the Heart, brilliant white with the Lungs, indigo with the Kidneys, and emerald green with the Liver. From the Earth itself arises a golden yellow mist; this color is associated with the Spleen. From Heaven descends a silvery white mist entering the body at the top of the head; this color is associated with the Taiji Pole. These six different colored mists permeate and envelop the body with their healing light vibrations.

To perform this meditation, stand in a Wuji posture as you visualize:

1. A silvery white mist descending from the center of the Heavens to penetrate and envelop your Taiji Pole. This silvery white mist enters your body through the top of the head at the Baihui (GV-20) point, and flows into and fills up the Taiji Pole, saturating and absorbing into your center core.
2. A golden yellow mist, ascending from the center of the Earth, entering the body through the bottom of your feet at Yongquan (Kd-1) points. This golden yellow mist saturates your Spleen as it penetrates and envelops your center core;
3. A ruby red mist, flowing from the horizon in front of you enters your Heart, as it saturates and envelops the front of your body.
4. An indigo or dark blue mist flowing from the horizon behind you enters your Kidneys, as it saturates and envelops your back.
5. A brilliant white mist flowing from the horizon on your right side enters your Lungs, as it saturates and envelops the right side of your body.
6. An emerald green mist flowing from the horizon on your left side enters your Liver, as it saturates and envelops the left side of your body.

Patients should be monitored regularly to prevent Qi deviations which may occur while the patient is absorbing Qi through his or her pores. For best results while absorbing environmental Qi, the color of the energy should be clean, clear and bright.

PRECAUTIONS

Because the body's internal energy matches that of the external environment, it is important to avoid selecting any areas that are windy, turbulent, polluted, or stagnant. Also avoid cultivating energy in storms, during hurricanes, sweltering heat, or blistering cold. Avoid any type of extreme weather condition.

ADDITIONAL TECHNIQUES FOR TRAINING EARTH QI

The following exercises at one time were secretly practiced throughout China, as a means for Medical Qigong doctors and Qigong masters to practice advanced methods of environmental energy manipulation. The purpose of these Qigong exercises is to allow the doctors the chance to practice connecting and infusing their energetic field with the environmental energetic field.

PULLING CLOUDS OUT OF THE SKY

This exercise requires the Qigong doctors to first root and immerse themselves with the vibrational resonance of the Earth (this can be accomplished by practicing the One through Ten Meditation, described in Chapter 26). Next, the doctors extend an energetic line from the Earth into the sky, connecting to and enveloping a specific cloud. Through focused intention, the doctors imagine sucking the cloud out of the sky and down into the Earth, rooting its energetic form into the Earth's energetic field. As the cloud's energy is absorbed into the Earth, it dissolves from the sky.

Once the doctors have rooted the cloud into the Earth's energetic field, they have two choices: either to release the cloud back into the sky by attaching its energetic form onto an already existing cloud, or to allow the cloud to disperse into the energy of the surrounding environmental field.

DIVIDING A CLOUD IN HALF

After connecting and rooting into the Earth's energetic field, the Qigong doctors extend their intention into the sky, connecting with a particular cloud formation. Next, the doctors condense their upward intention, and like a laser beam, begin dividing and separating the cloud into two halves. After the cloud has been divided into two separate formations, the doctors can either continue to further divide and dissect the cloud or reconnect its structure.

THE EARTHLY TREASURE OF FOOD

The Earth's natural energetic forms and transformations affect the flow of the body's life-force energy. Food and water are examples of a natural form of Earth Qi, and are viewed as Earthly Treasures. Although the air we breath is representative of Heaven Qi, due to the influence of environmental energetic fields, it is also viewed as an Earthly Treasure. The objective of self-cultivation work is to continually purify the energy within the Microcosm of Man, and increase the awareness of the pure energy within the Macrocosm of the universe.

Because of its connection to the Earth's Yin and Yang energy fields, the body's life-force energy can be replenished through the consumption of food and herbs. Both food and herbs have Yin or Yang properties which, when consumed, help the body in moving, tonifying, purging or regulating the life-force energy.

UNDERSTANDING THE YIN OR YANG ASPECT OF FOOD

When trying to determine whether a food is more Yin or Yang, several characteristics should be evaluated. Yin foods take less time to grow, are more watery or wet, and are colder, and sweeter. Yang foods take more time to grow, are drier, hotter, and less sweet. The more Yin the food, the more expanded it will become (e.g., a pumpkin); the more Yang the food, the more contracted or dense it will become (e.g., a turnip).

Diet directly influences the body's ability to generate Qi. The body's ability to effectively digest food plays a big part in how efficiently the body utilizes the nutrients it receives.

UNDERSTANDING THE YIN OR YANG ASPECT OF HERBS

Food is a powerful healing tool. Medicinal herbs are considered "special foods," and are divided into Yin and Yang aspects throughout China. They are prescribed to be taken in soups (tang), teas, or as additives to other dishes; for example, Chinese Qigong doctors, as well as many martial arts masters, frequently drink herbal teas to help regulate their body's energies during seasonal changes. The herbal teas are consumed according to the following sequence.

1. The summer is considered the season of maximum Yang; green herbal tea is drunk in order to cool down the body and regulate the internal organ systems.
2. The autumn is the season when Yin and Yang energies balance; however, the Yin energy is increasing, so flower tea is drunk to help stabilize the body's Yin and Yang balance.
3. The winter is considered the season of maximum Yin; black or red tea is drunk to warm the body and regulate the internal organ systems.
4. The springtime is the season when Yin and Yang energies balance; however, the Yang energy is increasing, so flower tea is drunk to help stabilize the body's Yin and Yang balance.

Teas are frequently prescribed for their particular healing properties. Each herb has a specific nature. Herbs that are considered Hot or Warm are used to treat Cold conditions, and herbs that are considered Cold or Cool are used to treat Hot conditions. Herbs are also categorized according to their flavor (sour, bitter, sweet, pungent, or salty) and are used to affect the body's internal organ energy. Herbs help the Qi ascend or descend in order to purge or tonify an internal organ or channel. Herbs are also used in moving the Blood. Specific herbs are also chosen for prescription, based on the Differential Diagnosis according to the Eight Energetic Principles (see chapter 21), the Five Elements (see chapter 23), and the Six Stages (see chapter 24) theories .

Generally, because all herbs have medical uses and contraindications, they are collected and divided into three categories: superior herbs - which consist of tonics, common herbs - which are slightly toxic, and inferior herbs - which are toxic. To gather herbs, the doctor must have an extensive knowledge of the environment (season, place, time, etc.), and the specific herb (roots, bark, leaves, flowers, fruits, etc.) The doctor must also possess the harvesting skill required for gathering the herb. Once the herbs have been gathered, the doctor must then begin the procedure of processing them. The process work is important as some herbs are poisonous and need to be detoxified, while others have their properties enhanced through the extracting process.

ENERGY, FOOD, AND DIET

Many generations have asked the same question, "Is food medicine or medicine food?" An old Chinese proverb states, "To tonify with food is better than with medicine, but to tonify with Qi is better than food." Several decades of nutritional, epidemiological, and clinical studies point to a powerful link between disease and dietary habits.

For centuries, Chinese healers have studied food, discovering the properties of specific foods and the secrets of using them to achieve health and longevity. Chinese nutrition focuses on the energetic properties of food and their qualitative actions on the body as a whole, as well as their influences on the various internal organs and channels. Unless the patient eats correctly, Qigong, acupuncture, Chinese massage, and even herbal therapy will not be as effective, or long standing. In the Tang Dynasty, the great physician Sun Simiao wrote the medical classic *One Thousand Ounces Of Gold*, which described the dietary treatments of several diseases, including night blindness and goiter.

What we eat can either aggravate or enhance the healing of a particular condition. When Gu Qi (food energy) is being generated, strong surges of energy from the Spleen and Stomach (in the Middle Dantian) cause the body to generate more Yang Qi. As the body's Yin aspect weakens, its energetic nature becomes dryer and wanes; the Yang is no longer held in check and an overexuberant Yang begins to generate Heat, which produces, and sometimes releases, emotions.

Emotions begin with, and stem from, the energetic properties of Postnatal Qi. When ministering to patients, the doctor should ensure a healthy physical and social environment (e.g., clean air and a loving support system), as well as monitor the quality and quantity of food consumption. Hostile surroundings interfere with the proper digestion of food and cause toxic Qi. Food, air, and water become the fuel that generates energy in the body. When life-force energy becomes impure from poor diet, Heat accumulates in the organs, creating Excess Yang, or Excess Fire Qi.

One essential skill for harvesting the Earth's Treasures is controlling the intake of food and diet. It is important that a patient not eat foods that are excessively Cold, Hot, spicy, or greasy. If food is ingested while the food is too Yin (Cold) or out of harmony with the seasons, an external pathogenic factor can be created. When a Cold or Yin factor is created (a condition where Cold and Excess Yin consumes the Kidneys Yang, preventing the body's Yang Heat from warming the body), the Qi becomes sluggish, thus blocking the channels or collaterals. These blockages cause pain and damage to the Stomach and intestines and affect the Heart and Lungs.

If food is ingested while the food is too Yang (Hot) or out of harmony with the seasons, a Yang factor is created, causing internal energy to be released, which damages both the Blood and Qi. Because of its turbid nature, food that is too greasy may damage the Stomach Qi, causing boils and pyogenic infections, along with ulcerous skin diseases. Excessive eating may cause too much Gu Qi to develop, generating an excessive amount of power. This Excess power produces an upward reaction of Stomach Qi, which is detrimental to both the Spleen and Stomach, causing the breath to become obstructed (mother/Spleen affecting child/Lungs), and the psychic centers (Shen within the Heart) to become blocked. Many postnatal illnesses result from eating the wrong kinds of food in the wrong season, not eating enough of the right kind of food, overeating, or all of the above.

Balancing the diet in accordance with the cycles of nature and the principles of Yin and Yang

Balancing the Diet					
	Liver	**Heart**	**Spleen**	**Lungs**	**Kidneys**
Excess	Sour	Bitter	Sweet	Pungent	Salty
Controlled By	Pungent	Salty	Sour	Bitter	Sweet
Counteracts	Sweet	Pungent	Salty	Sour	Bitter

Figure 12.3. The Five Flavors and their Controlling and Counteracting Cycles

should be an integral part of the daily practice. In some situations, the body can compensate for an improper or unbalanced diet. To counteract this imbalance, however, the body will expend a great quantity of energy. Eating similar foods all the time results in the accumulation of toxins from those foods and can cause mildly allergic reactions. A balance of the different food groups is the key to dietary success.

USING THE FIVE FLAVORS OF FOOD

The energetic properties of the Five Flavors of food (sour, bitter, sweet, pungent, or salty) can be used to balance and control the energy of each organ. Any excess or deficiency of one flavor affects not only the organ to which it corresponds, but all the other organs as well (Figure 12.3). In the clinics in China, the theory of the Five Flavors or tastes is combined with the theory of the Five Elements to describe the interaction of the different "flavors" on the body's internal organs and energetic systems (Figure 12.4). The Five Flavors and their affect on the body are described in the following sections.

SOUR TASTE, ELEMENT WOOD

Some examples of Sour foods include: lemons, pickles, liver, vinegar, wheat, salad oils, and sour cream. Sour foods affect the Liver, eyes, and tendons. They are also excellent for stimulating the energy of the Liver and Gall Bladder.

1. Sour foods are astringent (causes shrinking), and can be prescribed to achieve the following treatment goals:
   - to prevent or reverse the abnormal leakage of Qi and Fluids,
   - to counteract diarrhea and excessive perspiration by slowing the movement of fluids,
   - to drain the Liver of Excess Qi, thereby indirectly strengthening the Lungs, and
   - to correct an irregular, slow heartbeat.
2. An excess of sour foods can cause the following dysfunctions:
   - injures the muscles,
   - hardening and wrinkling of the skin, and
   - stiffness spreading along the tendons that are controlled by the Liver. As the Liver controls the tendons, patients with Liver trouble should also eat less sour food.

BITTER TASTE, ELEMENT FIRE

Some examples of Bitter foods include: coffee, black and green teas, asparagus, celery, turnips, leeks, cabbage, broccoli, and cauliflower. Bitter foods affect the Heart, tongue, and Blood. They are also excellent for stimulating the energy of the Heart and Small Intestine.

1. Bitter foods are used for draining and drying, and can be prescribed to achieve the following treatment goals:
   - to stimulate the energy of the Heart, control the Heart Fire,
   - to reducing fever and Excess body Heat,
   - to stimulate digestion,

Five Elemental Cycle					
Element	Fire	Metal	Wood	Earth	Water
Season	Summer	Autumn	Spring	Late Summer	Winter
Organ	Heart	Lungs	Liver	Spleen/ Pancreas	Kidneys
Bowel	Small Intestine	Large Intestine	Gall Bladder	Stomach	Urinary Bladder
Superficial Organ	Blood Vessels	Skin	Nerves	Muscles	Bones
Release Point (Sense Organ)	Tongue	Nose	Eyes	Mouth	Ears
Taste	Bitter	Pungent	Sour	Sweet	Salty

Figure 12.4. The Five Flavors in Relationship to the Body's Internal Organs, Seasons, and Elements

- to drain Excess Spleen Qi, and
- to remove obstructions in the upper respiratory tract.

2. An excess of Bitter foods can cause the following dysfunctions:
   - dryness and congestion of the Spleen and Stomach; and
   - hyperactivity of the Heart Fire, and the consumption of Kidney Yin fluids.

SWEET TASTE, ELEMENT EARTH

Some examples of Sweet foods include: honey, sugar cane, orange juice, pecans, snow peas, corn, milk, and sweet potatoes. Sweet foods affect the Spleen, mouth, and muscles. They are also excellent for stimulating the energy of the Spleen, Stomach, and pancreas.

1. Sweet foods are used for tonifying, regulating and sometimes moistening, and can be prescribed to achieve the following treatment goals:
   - to stimulate the digestive Fire,
   - to tonify the Spleen, Stomach and pancreas, and
   - to drain the Heart of Excess Qi.
2. An excess of sweet foods can cause the following dysfunctions:
   - an imbalance in the Kidneys,
   - an aching in the bones,
   - hair loss from the head,
   - muscle dysfunctions, and
   - injury to the Spleen and Stomach. The Sweet flavor spreads through the muscles, so patients with diseases of Spleen, Stomach, or muscles (diabetes, hypoglycemia, systemic candida, fibriomyalgia, chronic fatigue, etc.) should limit their intake of Sweet foods.

PUNGENT TASTE, ELEMENT METAL

Some examples of Pungent foods include: garlic, onions, ginger, black and white pepper, mustard, and chilies. Pungent foods affect the Lungs, nose, skin, and body hair. They are also excellent for the stimulating the energy of the

Lungs and Large Intestine. Because the Lungs control the Qi, the Lungs are very quick in absorbing the energy of Pungent foods.

1. Pungent foods are used for dispersing and moving, and can be prescribed to achieve the following treatment goals:
   - to induce perspiration,
   - to improve circulation of Qi and Blood,
   - to stimulate the digestion,
   - to activate the Lung energy and promote the Lung's function,
   - to expand the Lung Qi,
   - to drain the Lungs of Excess Qi, and
   - to increase Fluids and saliva secretions.
2. An excess of Pungent foods can cause the following dysfunctions:
   - great loss of Lung energy and damage to the Lung organs,
   - knots in the muscles, and
   - withering or decay of the fingernails and toenails.

SALTY TASTE, ELEMENT WATER

Some examples of Salty foods include: shellfish, cheese, soy sauce, salt, ham/pork, and margarine. Salty foods affect the Kidneys, ears, and bones. They are also excellent for stimulating the Kidney Qi and Bladder Qi.

1. Salty foods are used for purging and softening, and can be prescribed to achieve the following treatment goals:
   - to soften hard nodules such as cysts,
   - to purge inflamed lymph glands, and
   - to soften knotted muscles.
2. An excess of Salty foods can cause the following dysfunctions:
   - cause damage to the Kidneys and Bladder,
   - cause the Blood to coagulate or clot,
   - cause Blood diseases,
   - harden the pulse, and
   - cause changes in the complexion.

GUIDELINES FOR THE FIVE ELEMENTAL TASTES

During each season, the five separate extremes in flavors serve as a catalyst to draw Qi into the Deficient organ. Overexposure to the seasonal elements (Cold, Wind, Damp, etc.) is responsible for external pathogenic factors that invade the body and affect the body's Jing, Qi, and Shen. The development of disease is due to an imbalance in the struggle between the body's health preserving Qi and the pathogenic influences. If the pathogenic factors are excessive, they can obstruct the body's normal physiological function, creating illness. If, however, the body's Qi is strong, the pathogenic factors will have a difficult time causing any energetic imbalances.

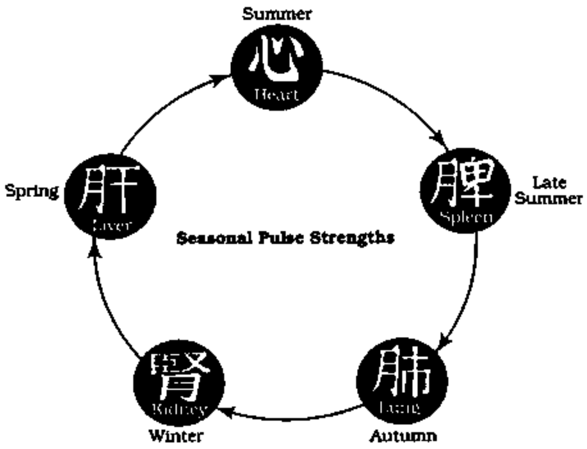

Figure 12.5. The quality of the body's pulse is constantly changing according to the rhythm of the five seasons.

Traditional Chinese Medicine divides the effects of these tastes into five seasons, each with its own unique function and flow of energy affecting the human body. One example of the effects of nature's rhythms on man is that the pulse quality in the viscera constantly changes in accordance with the five seasons (Figure 12.5). The Liver pulse is relatively strong in spring, compared to the other organs; in the summer the Heart is stronger, in the late summer - the Spleen, in the autumn - the Lungs, and in winter - the Kidneys.

Basic guidelines are developed for eating in accordance with the flow of the Five-Element cycle. These guidelines will help train the body to recognize what foods it needs. The food types and tastes are divided into five categories, which relate to the five seasons of the year, the Five Elements, and the Yin and Yang organs (Figure 12.6).

To achieve overall balance, food and energy must be equally balanced. If the Liver is weak, for example, eating Sour food will help correct it and Salty food will also help. This prescription is based on the Five Element theory, which focuses primarily on two cycles—the Generating Cycle and the Controlling Cycle (Figure 12.7).

OVERVIEW OF THE FIVE ELEMENTAL TASTES

The overall understanding of the Five Elemental tastes empowers the Qigong doctor to accurately assist in prescribing better eating habits for his or her patients. By understanding the seasonal eating patterns (i.e., in the winter eating root food, which is grown under the ground, and in the summer eating food that is grown via branches and flowers) the doctor can observe whether the patient is further compounding his or her disease condition through improper diet.

SUMMARY

In comprehending the energetic potential of the Three Treasures of Earth, the Qigong doctor will be able to evaluate the patient's eating and living conditions and their affect on the patient's disease. This also includes prescribing the Medical Qigong exercises, meditations, diets, and eating habits needed in order to support and maintain the doctor's treatments.

Through monitoring the patient's connection to their environment and eating patterns, the Qigong doctor is able to understand the Earthly influence on the patient's physical, mental, emotional, and spiritual being.

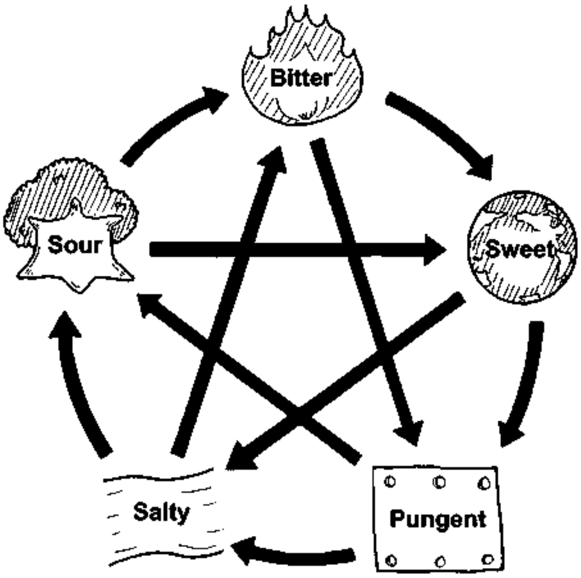

Figure 12.6. The Five Elements' Interaction of the Five Flavors

In the Creative or Generating Cycle—		
Salty Kidney (Water)	generates	(Wood) Sour Liver
Sour Liver (Wood)	generates	(Fire) Bitter Heart
Bitter Heart (Fire)	generates	(Earth) Sweet Spleen
Sweet Spleen (Earth)	generates	(Metal) Pungent Lung
Pungent Lung (Metal)	generates	(Water) Salty Kidney

In the Controlling Cycle—		
Salty Kidney (Water)	controls	(Fire) Bitter Heart
Bitter Heart (Fire)	controls	(Metal) Pungent Lungs
Pungent Lungs (Metal)	controls	(Wood) Sour Liver
Sour Liver (Wood)	controls	(Earth) Sweet Spleen
Sweet Spleen (Earth)	controls	(Water) Salty Kidneys

Figure 12.7. The Creative (or Generating) and Controlling Cycles of the Five Flavors

CHAPTER 13

THE THREE TREASURES OF MAN: JING, QI, AND SHEN

The Outer Force (Power) of the Three Treasures of Man are expressed through Jing (Essence), Qi (Energy), and Shen (Spirit). The human body is made up of vibrational frequencies that interact to create energetic fields and substances, or energetic materials ranging in structure from less dense to more dense (Figure 13.1). In order of increasing density, these energetic materials manifest as Spirit (Shen), Energy (Qi), Essence (Jing), Blood (Xue), Body Fluids (Ye/Jin), Marrow (Sui), and Bone (Gu). Of these energetic substances, Essence (Jing), Energy (Qi), and Spirit (Shen) are regarded as the most important. These three materials perform the most important functions in the human body, and are known as the Three Treasures of Man or San Bao.

The energy of the human body, or Man's Qi, is created from and through different interacting rates of vibration and energy fields that combine to form the body's cells, tissues, and organs, creating a living organism. In Chinese philosophy, the Three Treasures of Man also represent the interaction of Heaven and Earth within the human body.

Man is composed of these combined Heaven and Earth energies:

1. The Jing is the body's tangible essence, composed of a morphogenic field of energy that creates and sustains the body's physical form. It is the lower vibrational frequency of the human energy matrix. During the healing process, the patient's Jing directs the proper cells to multiply so that the body can heal rapidly and completely.
2. The Qi is the life-force energy that maintains the body's Jing. It is the electromagnetic body of the channel system and distributes vitality to the body through energetic transference. It

Natural Human Transformations

Yin Organ	Yang Organ	Sense Organ	Taste	Color	Body Smell	Sound	Tissue	Positive Emotion	Negative Emotion
Liver	Gall Bladder	Eyes	Sour	Green	Rancid or goatish	Shout	Nerves & tendons	Kindness, benevolence	Anger
Heart	Small Intestine	Tongue	Bitter	Red	Scorched or burned	Laugh	Blood vessels	Love, joy, respect	Elation, arrogance, impatience
Spleen	Stomach	Mouth	Sweet	Yellow	Sweet, fragrant	Sing	Muscles & fascia	Fairness, openness	Worry, anxiety
Lungs	Large Intestine	Nose	Pungent	White	Tidepool, fishy	Weep	Skin & hair	Righteous, courageous	Sadness, grief
Kidneys	Bladder	Ears	Salty	Blue	Putrid, rotten	Groan	Bones	Gentleness, wisdom	Fear

Figure 13.1. This diagram shows the natural transformations of man's internal organs, energy, and matter, as well as their natural manifestations.

is the middle vibrational frequency of the human energy matrix. The vibrational field of Qi sustains and contains the thoughts and emotions, and interacts between the vibrational fields of Jing and Shen.

3. The Shen is the spirit and psyche which maintains the body's Qi. It is the higher vibrational frequency of the human energy matrix. It is also the energetic field which sustains the vibration of the Mind (the process of awareness and consciousness located throughout the body, not limited to the brain), interfacing between the Mind and Qi.

These three substances resonate at different frequencies and are mutually dependent upon each other for maintenance and survival. Low frequencies are attributed to the resonance of Jing, medium frequencies to the resonance of Qi, and high frequencies to the resonance of Shen. Metaphorically speaking, Jing can be compared to a dense, solid material such as ice, Qi to a flowing material such as water, and Shen to a vapor.

In Chinese medicine, the Qigong doctor uses the Three Treasures of Man for assessing and diagnosing the patient. Observing the state and interaction of the patient's Jing, Qi, and Shen allows the doctor to determine the patient's prognosis and chances for successful treatment. The doctor can assess the Jing, Qi, and Shen by:

1. Observing the general vitality of the patient's Jing through pulse reading and tongue diagnosis,
2. Noting the general vitality of the patient's Qi through observation of the Wei Qi field, and
3. Observing the general vitality of the patient's Shen by the quality of glitter in the patient's eyes.

Within this model, Jing, Qi, and Shen are viewed as layers of energetic fields between the body and the Mind. Jing is the innermost energetic layer, the blueprint which holds the physical body to the desired form. Shen is the outer energetic layer of the body through which the Mind experiences and directs the activity of the body. Qi is the intermediate energetic layer, the medium which allows the Jing and Shen (the energetic fields of the body and spirit) to communicate.

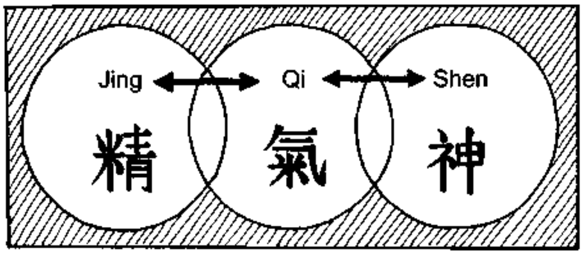

Figure 13.2. The Three Treasures of Man

PRENATAL AND POSTNATAL JING, QI, AND SHEN TRANSFORMATIONS

Jing is the body's foundational substance, responsible for nourishing the tissues. Qi emerges out of Jing circulating through the tissues. It promotes the body's metabolism. Shen governs the body. The body contains both prenatal and postnatal versions of Jing, Qi and Shen.

1. Prenatal Jing (Yuan Jing) is the body's innate and true Original Essence, which determines the constitutional makeup, strength, and vitality. It is not the reproductive Essence (sperm and ovary essence), which is considered Postnatal Jing.
2. Prenatal Qi (Yuan Qi) is the body's Original Energy, extending into the Wuji (infinite space or Void). It is the motivative energetic force linking the Prenatal Jing with the Prenatal Shen. It is not inhaled air (oxygen), which is considered Postnatal Qi.
3. Prenatal Shen (Yuan Shen) is the Original Spirit, that is able to perceive and intuit truth and knowledge (knowing without knowing). It is not cognitive thinking or consciousness, which is considered Postnatal Shen.

Through meditation, the body's alchemical processes of internal energy transmutations can be cultivated. Jing can be used to create or increase Qi; likewise, Qi can be used to create or expand the Shen. This creative cycle also works in reverse order by transforming Shen to Qi, and Qi to Jing (Figure 13.2).

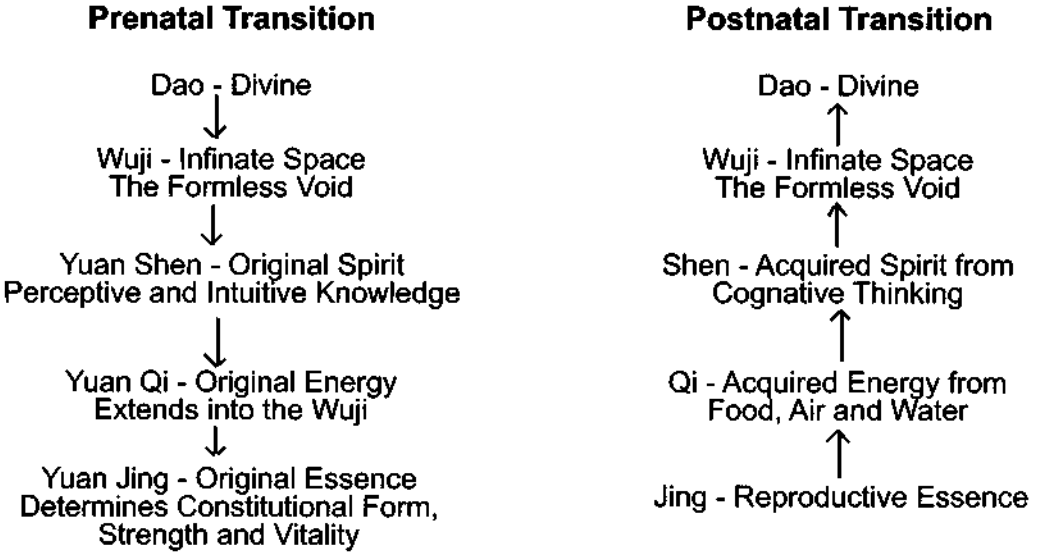
Prenatal Transition
Dao - Divine
Wuji - Infinate Space
The Formless Void
Yuan Shen - Original Spirit
Perceptive and Intuitive Knowledge
Yuan Qi - Original Energy
Extends into the Wuji
Yuan Jing - Original Essence
Determines Constitutional Form,
Strength and Vitality
Postnatal Transition
Dao - Divine
Wuji - Infinate Space
The Formless Void
Shen - Acquired Spirit from
Cognative Thinking
Qi - Acquired Energy from
Food, Air and Water
Jing - Reproductive Essence

Figure 13.3. The Prenatal and Postnatal Alchemical Transitions of Jing, Qi, and Shen

PRENATAL TRANSFORMATIONS

The following transformations describe the origin and creation process of the Three Treasures within the body of the fetus (Figure 13.3). Through comprehending these transformations, the Qigong doctor can teach individuals how to cultivate their Jing. The body's prenatal energies, which support the body's postnatal Jing, are sustained by prayer, meditation, and sleep. This quiescence state is necessary for the body to reenergize its Prenatal Qi.

1. Divine Energy (Dao) transforms into infinite space (Wuji). At this beginning stage, the spirit of the God envelops infinite space, including the energetic fields of the Heavenly universe and the environmental Earthly fields.
2. Infinite space (Wuji) transforms into Spirit (Yuan Shen). At this transition, the spirit of God descends to permeate, unite, and dwell within the Original Spirit (Yuan Shen) of the forming fetus.
3. Spirit (Yuan Shen) transforms into Energy (Yuan Qi). At this stage of transition, the Original Spirit of the fetus transforms into Original Energy (Yuan Qi).
4. Energy (Yuan Qi) Transforms into Essence (Yuan Jing). At this final stage, the Original Energy of the fetus is transformed into Original Essence (Yuan Jing).

POSTNATAL TRANSFORMATIONS

The following transformations describe the origin and creation process of the Three Treasures within the body after birth. Through comprehending these transformations, the Qigong doctor is able to achieve a harmonious balance between the three energies through Postnatal Jing cultivation. The body's postnatal energies are supported and sustained first of all through food, drink, and air.

1. Essence (Jing) Transforms into Energy (Qi). At this beginning stage, the body's Jing is Heated within the Lower Dantian to create Qi. This transition is similar to ice melting into water.
2. Energy (Qi) Transforms into Spirit (Shen). At this transition the body's Qi is steamed within the Middle Dantian, to be transformed into Shen. This transition is similar to water transforming into vapor.
3. Spirit (Shen) Transforms into infinite space (Wuji). At this stage, the body's Shen fuses with the resonant light of the Upper Dantian

and is transformed and released into the Heavenly and Earthly fields of energetic space. This transition is similar to vapor diffusing into space.

4. Infinite Space (Wuji) Reunites with the Divine (Dao). At this final stage, the Shen is reunited with the divine.

The Shen directs the increased amount and flow of Qi during the transformation process of Jing into Qi and Qi into Shen through conscious intention. Through this soft focused intention an alchemical cycle of transformation is created whereby Jing, Qi, and Shen can then be gathered, refined, and trained.

TRANSFORMATIONAL ATTAINMENT MODALITIES

Each of these prenatal and postnatal transformations include training exercises and meditations that are either developed in low, medium, or high attainment approaches. The low attainment approach is known as the physical attainment approach; the medium attainment approach is known as the mental attainment approach; and the high attainment approach is known as the spiritual attainment approach. Qigong regulation encompasses all three attainment modalities in order to facilitate the ultimate in energetic transformation. The three attainment approaches for increasing to maximum capacity the body's energy are described as follows.

1. The Physical Attainment Approach concentrates on the development and cultivation of the body's energy by focusing on Dynamic (energetic movement) and Quiescent (stillness) Qigong meditations to fully realize one's energetic potential.
2. The Mental Attainment Approach concentrates on increasing mental power by focusing on symbolic drawings, reciting prayers, or mantras to manifest fully one's mental potential.
3. The Spiritual Attainment Approach concentrates on the development and cultivation of the body's Original Spirit (Yuan Shen) by focusing on Shengong meditations (meditations that focus on spiritual cultivation) to harvest one's spiritual potential.

JING: ESSENCE OF LIFE

Jing refers to the indispensable bioenergetic substance of all living things (i.e., the energetic matrix of the body's cells, tissues, etc.). In Medical Qigong therapy, Jing is considered the most Yin form of Qi. Jing, as a pure Essence of living matter, is more dense than Qi but more refined than Blood or Fluids. When Jing is cultivated and allowed to arise naturally within the body, the outer surface of the body's tissues becomes harmonized and glows. Deep within the tissues, the Jing pools like a deep spring. The greater the degree of harmony between the individual's physical, mental, energetic, emotional, and spiritual fields, the more Jing is transformed into Qi.

THE YIN AND YANG ASPECTS OF JING

Jing can be divided into the functional aspects of Yin and Yang properties. The Yin aspect of Jing provides the material basis for growth, development and reproduction, and is the substratum for the formation of the materials associated with the Marrow and Blood. The Yang aspect of Jing activates transformations, growth, development and reproduction, and is associated with the Yuan Qi, Blood and Body Fluids (Figure 13.4).

As a subtle energy, Jing has a tendency towards downward movement, flowing like water, it fills the Lower Dantian, adrenals, Kidneys, and urogenital organs. Essence originates from two sources and can be divided into Prenatal and Postnatal Jing.

PRENATAL JING

Prenatal Jing is also called Original Essence (Yuan Jing). It is inherited from the Jing of both parents and is considered the root of life. It is created upon conception through the blending of both parents' sexual energies and substances. Prenatal Jing constitutes the original substance of the body, and serves as the material base for growth, development, and reproduction. With the help of the Qi and Blood acquired from the mother, the Jing is responsible for the nourishment and development of the embryo's and fetus's brain, bones, muscles, tendons, skin, and hair. The Pre-

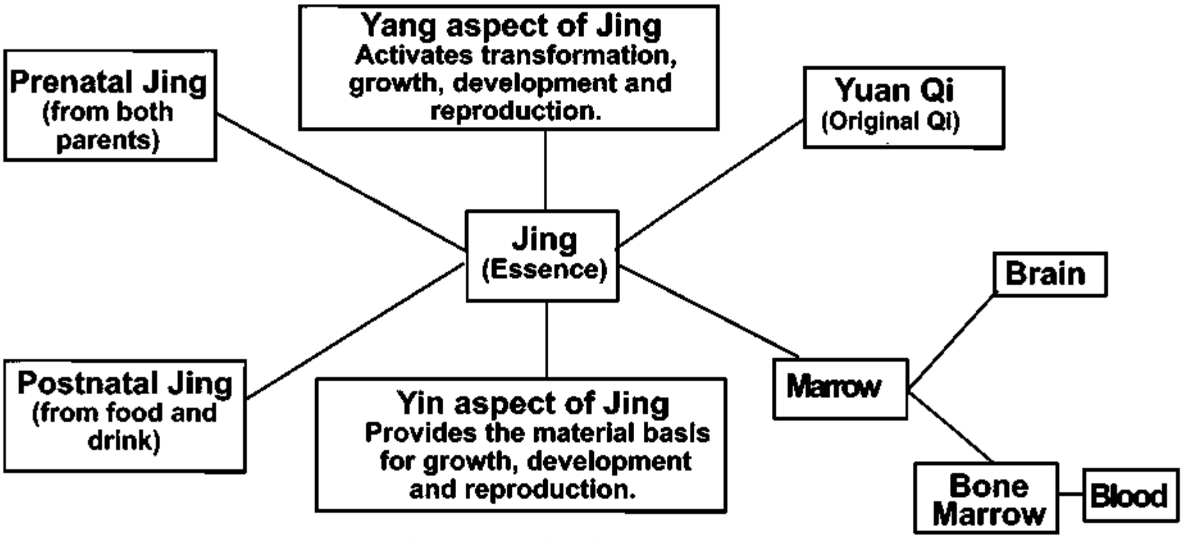
Prenatal Jing
(from both parents)
Yang aspect of Jing
Activates transformation, growth, development and reproduction.
Yuan Qi
(Original Qi)
Jing
(Essence)
Brain
Marrow
Postnatal Jing
(from food and drink)
Yin aspect of Jing
Provides the material basis for growth, development and reproduction.
Bone Marrow
Blood

Figure 13.4. The Formation of Jing

natal Jing determines each person's basic constitution, strength, and vitality.

POSTNATAL JING

Postnatal Jing is also called the Acquired Essence. It is the Essence of the body's main organs, derived from air and Gu Qi (produced from food and drink by the transformation process of the Spleen and Stomach) which is transported through the channels and vessels to nourish the body's Yin organs, and to irrigate the Yang organs. Postnatal Jing is the material basis for the functional activities of the body's internal organs and metabolism. It is transported (through the function of the Lungs) to all the Yin and Yang organs where it becomes reproductive Yin and Yang Jing. The Kidneys are responsible for storing any surplus Jing to be distributed when needed.

THE INTERACTION OF PRENATAL AND POSTNATAL JING

The Prenatal and Postnatal Jing are interdependent and contribute equally to the growth and development of each other. Postnatal Jing is continually being used by the body and replenished with food and drink. Only with the nourishment of the Postnatal (Acquired) Jing can the Prenatal (Congenital) Jing be enriched and function optimally. Without the function of the Prenatal Jing, the Postnatal Jing cannot be transformed into Qi.

Postnatal Jing is continually being used and replenished through the Spleen's production of Gu Qi from food and drink. This is why so much emphasis is placed on the first postnatal alchemical transition of changing Jing into Qi. Through Medical Qigong exercises, the body is able to stimulate, nourish, and preserve its Jing.

Both Prenatal and Postnatal Jing are related to the Kidneys, which generate bone growth as well as nourish the brain, memory development, and body stamina. The *Yellow Emperor's Classic of Internal Medicine* states that women are on a seven year developmental cycle, while men are on an eight year cycle. Problems with the body's Jing result in developmental problems (i.e. improper maturation, sexual dysfunction, infertility, and premature aging).

The Kidney Jing is responsible for extending the life-force energy up the Governing Vessel into the brain. The Kidney Jing also transforms into Marrow and is responsible for the production of bone marrow, which fills up the spinal cord and brain. When the Jing becomes exhausted, the memory begins to fail and the spirit becomes fatigued. It is important to remember that all types of Jing do not exist as separate substances, but support, interact with, and transform into one another.

The Kidney Jing also works in conjunction with the Qi to establish the body's state of mind and emotional health, which together determine the development of the Five Elemental Constitutions.

QI: VITAL LIFE-FORCE ENERGY

Qi is the energetic substance from which the entire universe was created. Through the interaction and transformation of Yin and Yang Qi, the various substances of form and matter are produced. This vital life-force energy comprises both material and functional aspects of the body.

Qi is the medium between matter (Jing) and Spirit (Shen). Chinese medicine describes many different kinds of Qi. Qi can assume an infinite number of forms according to its state of condensation or dispersion. When Qi condenses, matter is created, and when Qi disperses, energy is created. Although Qi in the body can assume different forms, that have different functions, it is essentially all one energy. The patient's Jing, for example, may be ready to receive a new energetic pattern for accelerated healing, and the Shen may be ready to send this required pattern. But, if the body's Qi is weak or turbulent, it will not transmit this information from the Shen to the patient's Jing, thus delaying or impeding the healing process.

The body's Qi can be classified into two primary forms of energy: Prenatal and Postnatal Qi.

PRENATAL QI

Prenatal Qi (or Xian Tian Zhi Qi, which translates as, "before the baby sees the sky energy") is also called Yuan Qi, or Original Qi. This is the energy that the baby inherits from his or her parents during conception. It is stored in the sexual glands and adrenal cortex, and it is essential for the growth and development of the fetus.

It is Jing in the form of energy, that is cultivated through Qigong practice. Yuan Qi is the innate or inborn energy that is the original force which maintains normal growth and development, and promotes the functional activities of the internal organs. Its role is to bring into existence the vital mechanisms of the human body according to their genetic pattern, conforming them to the models transmitted by each person's lineage.

Prenatal Qi contains elements of all of one's predecessors and gives rise to the Prenatal Jing, and is seen as a means of transmitting the Deoxyribonucleic Acid (DNA) code. Prenatal Qi relies on nourishment from Postnatal Jing. Through Prenatal Qi, each individual is energetically connected with everything in existence. Prenatal Qi also relies on Postnatal Qi for nourishment.

Methods of training Qi and regulating it include counting respirations as the Mind focuses on rooting the Qi into the Lower Dantian, Mingmen, or Kidneys' area.

There are three types of prenatal energy regulations used for circulating and harmonizing the body's internal energy or Dantian Qi:

1. The first type of Qi regulation focuses on moving energy through the body's main channels and torso. This is the most common type of Prenatal Qi circulation, directing the energy to flow along the course of the body's Conception and Governing Vessels, i.e., the "small" Heavenly cycles (Microcosmic Orbits). This type of Qigong exercise is commonly referred to as an Inner Cultivation and Regulation method.
2. The second type of Qi regulation focuses on moving energy through the body's channels and extremities. In this type of Qi circulation, the body's Prenatal Qi is directed to flow only along specific channels (i.e., being led from one channel out from the torso into an extremities and then back into the body through another channel). This type of energetic balancing is commonly seen in the "large" Heavenly cycles (Macrocosmic Orbits) and bone marrow regulation.
3. The third type of Qi regulation focuses on

moving the energy of Heaven (Universal Qi), Earth (Environment Qi) and Man (Qi of the physical body) through the channels. In this type of circulation, these three types of energy are integrated or merged into one energy and then regulated. Upon inhalation, the individual imagines Prenatal Qi circulating from Heaven through the body's channels into the Earth. Then upon exhalation, the energy is imagined flowing from the Earth, through the channels, back into Heaven.

POSTNATAL QI

Postnatal Qi (or Hou Tian Zhi Qi, which translates as, "after the baby sees the sky energy") is also called acquired or Post-Heaven Qi. This is the energy that the baby derives from air, food, and drink after it has been born. Postnatal Qi relies on Prenatal Qi for development.

Prenatal and Postnatal Qi form the foundation for the body's vital energy. In Traditional Chinese Medicine it is believed that the parents' health and their state of Jing, Qi, and Shen at the time of conception determines the general health of the child. If the parents are sick, or their health has is compromised by drinking, smoking, or drug use, the baby's health will suffer. If the parents are old, or have suffered an emotional shock during the pregnancy, this too will negatively impact the baby's health.

Methods of training Postnatal Qi include deep, tranquil abdominal respiration. This method trains the respiratory system to improve both the respiratory and digestive systems, enhancing both the Lung and Spleen Qi.

EIGHTEEN CLASSIFICATIONS OF QI

The distribution and function of life-force energy can be further divided and classified according to the source and function of the Qi. Below is a description of eighteen transitions of Qi and their energetic effects on the body.

1. Prenatal Qi (Yuan Qi), is often referred to as Original Qi, and includes the Original Yin and Original Yang energies of the body. Yuan Qi has its root in-between the Kidneys and Mingmen area, it is distributed throughout the entire body via the Triple Burners.
2. Postnatal Qi, is an external energy derived from the consumption of food and drink (from the Earth), and air (from the Heavens), and is a term used to describe the Qi which is cultivated after one is born.
3. Heavenly Qi (Tian Qi), is energy that has been developed from the energies of the sun, moon and stars. Tian Qi is absorbed into the Lung Qi as Air Qi (oxygen) through respiration.
4. Earthly Qi (Di Qi), is energy that has been developed from the energies of the soil, water and wind. Di Qi is absorbed into the Stomach and Spleen where it is converted into Gu Qi (food energy); Di Qi is also directly absorbed through the pores from the environment.
5. Food Qi (Gu Qi), is the first stage of energy transformation derived from food and drink. It is also called Qi of the Water and Grain in the digestive system. After food essence is cooked in the Stomach, the Spleen transforms it into Gu Qi. At the first stage of development Gu Qi is a coarse, unusable form of food essence. The Spleen divides the energy it derives from the food essence into Pure (Clear) Yang Qi and Impure (Turbid) Yin Qi. The Clear Yang essence of Gu Qi is transported by the Center Qi (Zhong Qi) via the Middle Burner, sending it upwards into the chest, first to the Lungs where it combines with Heavenly (Tian) Qi to form Gathering (Zong) Qi, then to the Heart where it combines with the Yuan Qi (from the Kidneys) to create the Blood. The Turbid Yin essence of Gu Qi is transported downwards through the gastrointestinal tract to be expelled as waste.
6. Clear Yang Qi (Qing Qi), is the clean, pure energetic essence of Gu Qi, transported by the Spleen, via the Middle Burner, up towards the Upper Burners, and chest area. While in the chest, it is combined with Gathering (Zong) and Yuan Qi to form True Qi.
7. Turbid Yin Qi (Zhou Qi), is the clouded, impure energetic essence of Gu Qi, transported by the Spleen, via the Middle Burner, down

towards the Lower Burner, Small Intestine, and Bladder organs, to be further refined. It is a more coarse, unrefined, polluted energy.

8. Center Qi (Zhong Qi), is the energy generated from the Middle Burner, derived from the Spleen and Stomach. Its primary function is to transport the Gu Qi into the chest, to be blended with the Heart and Lungs' Qi.
9. Gathering Qi (Zong Qi), is also called Respiratory Qi. Gathering Qi is considered the "essence of life," manifested through the conversion of the purest and most potent forms of the body's Jing (particularly sexual fluids, hormones, and neurochemicals). It is essential for the formation of the fetus and is the initial motivating power of life. It promotes vitality and increases stamina; it also enhances the immune system, promotes mental clarity and prolongs life. Gathering Qi combines the Heavenly Qi (inhaled by the Lungs) and the Gu Qi (derived by the Stomach and Spleen), and accumulates these energies within the chest in order to assist the Heart in circulation the Blood, and the Lungs in respiration. It controls the speech and strength of the voice, as well as the Blood circulation to the extremities. Gathering Qi gathers in the chest and is controlled by the Middle Dantian.
10. True Qi (Zhen Qi), is also called Anti-pathogenic Qi. It originates in the Lungs and is the last stage of refinement and transformation of Qi. Gathering Qi and Yuan Qi combine their substances to form the True Qi in the chest area. True Qi fills and nourishes the body as it is further broken down into Nourishing (Ying) and Protective (Wei) Qi. True Qi is the energy that circulates inside and outside of the body, in the channels and collaterals, nourishing the Yin and Yang organs. It is called True Qi because it is the most refined and pure, and is in harmony with the environment. It is opposed by Evil Qi (Xie Qi), also called Toxic Qi, which is disharmonious and thus untrue to the body's essential nature.
11. Nourishing Qi (Ying Qi), is another form assumed by the True Qi and is the Yin aspect of the Channel Qi. Ying Qi has two major functions: to nourish the Blood, organs and tissues, and to link the mind and intention to the body. It flows within the blood vessels, as well as the channels, and is the fuel of the body's metabolism. Its potency depends upon the quality of food and water consumed, the purity of air breathed, and the efficiency of the body's digestive, respiratory, and circulatory systems. Since it is Yin, it is considered the more refined form of True Qi, as it flows in the body's inner layers and internal organs. Being directed by thoughts, it is thereby closely related to the emotions.
12. Protective Qi (Wei Qi), is a second form of True Qi and is the Yang aspect of the channel Qi. It protects the Blood, organs, and tissues from invasion of external pathogenic factors such as: harsh weather conditions, microorganisms, and harmful emotional influences. It also protects the individual from evil spiritual forces. It is the coarser Yang aspect which flows through the outer layers of the body's skin. Its texture is slippery in nature and cannot enter into the channels. It therefore circulates under the skin and in between the muscles. The Wei Qi vaporizes in between membranes, and diffuses over the chest and abdominal area to form a kind of energetic armor. It also warms, moistens, and contributes to the nourishment of the skin and muscles. The Wei Qi also helps regulate the perspiration by opening and closing the pores.

Wei Qi has its root in the Lower Burner (Kidneys), is nourished in the Middle Burner (Stomach/Spleen), and spreads from the Upper Burner (Lungs) to the outside of the body. Wei Qi is said to be located at the superficial level of the body during the day, in the muscles, skin, hair, etc. At night it sinks to a deeper level to circulate through the viscera. If the Wei Qi is obstructed in some manner from moving inward, insomnia may result. Also, Wei Qi becomes thick and extends outside its normal realm during Medical Qigong practice and may, therefore, take longer to move inward at night, causing some practitioners to experience difficulty falling asleep after evening Qigong practice.

13. Channel Qi (Jing-Luo Qi), is also called Meridian Qi. This energy flows throughout the body's deep internal and superficial channels, as well as through the collaterals and blood vessels like rivers and streams.
14. Organ Qi (Zang and Fu Qi), is the energy of the physiological activity of the body's internal organs and manifests as a major aspect of its physiological function. Individual pools of Qi are formed within the pairs of Yang-Fu (hollow) bowels and Yin-Zang (Solid) viscera. The Yang organs are responsible for transforming food and drink to produce Qi and Blood. The Yin organs are responsible for storing vital substances. Each organ can be additionally divided into pairs of Yin and Yang energetic aspects.

    Each of the body's organs has its own energy, which is governed by, and corresponds to, one of the Five Elemental energies. These Five Elemental energies respond to the external energetic fields drawn from the universal and environmental energetic fields. The internal processes of thinking and feeling, as well as the physiological process such as metabolism and hormonal production influence the Organ Qi.
15. Evil Qi (Xie Qi), is also called Pathogenic Qi, and Toxic Qi. This usually refers to any harmful external pathogenic factor (especially the Six Climatic factors). However, it also refers to the evil spirits and stuffed toxic emotions, which affect the individual's physical, energetic, and emotional balance.
16. Righteous Qi (Zheng Qi), is the defensive aspect of the True Qi, which protects the body from Evil Qi. This is not actually another type of energy, but a term used to indicate the True Qi's function of protecting the body from invasion by external pathogenic factors. The occurrence and development of disease is directly related to either the hyperactivity or hypoactivity of the Righteous Qi.
17. Spiritual Qi (Ling Qi), is the supernatural energy associated with the most highly refined energy in the human body manifesting through the Eternal Soul. This type of energy enhances spiritual awareness and constitutes the foundation for advanced levels of spiritual conditioning, and the attainment of an enhanced energetic body, and enlightened states of mind and spirit.
18. Divine Qi (Shangdi Qi), is also called Holy Energy or God's Healing Light. It is the supernatural energy associated with the Divine (Dao). The divine energy is the true source of spiritual awareness and constitutes the foundation for advanced levels of spiritual healing, as well as attaining enlightenment.

THE SIX FUNCTIONS OF QI

A Qigong doctor will primarily focus on a patient's combination of Original Qi (Yuan), Gathering Qi (Zong), Nourishing Qi (Ying), and Protective Qi (Wei) when treating or prescribing homework. Initially, Qi has six main functions in regulating the body. Although these functions may seem to overlap, their individual responsibilities are distinct.

1. Qi transforms substances into energy (one example of its transforming nature is changing food into energy).
2. Qi transports the substances it creates.
3. Qi holds and contains energy, Blood and organs, etc., within the energetic and visceral structures.
4. Qi protects against both external and internal pathogenic factors.
5. Qi raises and elevates things (for example, Blood, associated Fluids, and the body's temperature).
6. Qi keeps the body warm (for example, strong Qi will warm the body, while weak Qi makes the body cold).

The normal physiological activity of Qi in the body is a constant harmonious movement, similar to the ebb and flow of a tide. When the body's Qi is in constant motion, it moves in eight primary Yin and Yang directions: ascending and descending, gathering and dispersing, expanding and contracting, and entering and leaving. If however, there develops a Qi obstruction, or a Deficiency of Qi, the Qi will deviate from its natural course, and may cause Rebellious Qi, disharmony and disease.

Qigong doctors carefully examine the energetic ebb and flow of Qi to comprehend the complexity of the energetic hologram in which we exist as a living energy. As the human body absorbs, stores, utilizes, and releases energy, its natural function of transforming energy promotes health and allows us to live in harmony with the natural environment. However, once an internal organ or organ system develops a Rebellious, Excess, Deficient, or stagnant condition within the body's energetic flow, disease develops. By understanding the energetic pattern and function of each energetic organ and organ system, the Qigong doctor can diagnose the disease by retracing the energetic dysfunction to the root cause. The following graph denotes the function and flow of the body's life-force energy patterns (Figure 13.5).

QI AND ENERGETIC MEDICINE

The utilization of the Three Regulations (posture, respiration and mental Dao Yin concentrations) enable Qigong doctors not only to disperse any Toxic Qi acquired from patients, but also to regulate their own channels, through gathering energy, and increasing their own energetic fields. Through the proper use of the Mind (whole body consciousness and awareness), the doctor can simultaneously influence the energy flows within the patient's body, as well as his or her own Wei Qi fields.

As the doctor begins to focus on specific goals (i.e., dispersing Toxic Qi from a particular organ, tissue area, or channel), the patient may experience a surge and release of strong emotions and memories that had been stored in that area. The patient's energy field then shifts and moves throughout the body according to the specific emotion(s) elicited and released. This energy field is closely related to the thinking activities of the brain. As the Mind begins to release certain specific memories, the field of energy shifts and moves through the body in accordance with the specific emotions that are activated and released. During treatment, the shape, flow, rhythm, nature, and thickness of this energetic field can be perceived by the doctor who has become highly sensitive to perception of energy through "inner-vision" (see Chapter 25).

By training in the Three Dao Yin Regulations (see Chapter 14), internal energy can be guided to surge not only in the natural direction of flow within the channels, but also against it. Internal energy can also be guided along several channels at the same time, in any given direction or to specific areas of the body. By directing Qi against the flow of a channel, the doctor arrives at the source of the river (or energy). This technique is sometimes used by the Qigong doctor to collect or tonify the Qi within the patient's internal organs. This technique for storing and collecting energy is called "following the river to reach the sea." An example of this technique would be extending Qi from the patient's St-36 points, up the Stomach Channels to root, strengthen and stabilize a Deficient or Rebellious condition within the Stomach organ itself. Conversely, by directing Qi with the flow of the stream, one can disperse energy (purgation) and pull Qi out of an area leading it away from the source (e.g., pulling Liver-Stomach Heat down and out the Stomach Channel to cool an overheated Stomach).

QI AND THE BODY'S ENERGETIC STRUCTURE

Similar to the biomechanics of the respiratory, circulatory, and digestive systems, the body's energetic system has its own organization and established rules of action for regulating organ and tissue functions. The adaptability of Qi to the outside environment enables the body to absorb not only healthy energy, but also to expel Toxic Qi.

The nature of Qi is to endlessly gather and disperse, expand and contract, to flow inward and outward, to rise and fall, as it circulates. When a Qigong doctor begins to treat a patient's channels or points, he or she stimulates and activates the patient's Qi, which may arouse muscle movement and sensations which can last for quite some time.

If the frequency and nature of the doctor's Qi is similar to that of the patient's, both energetic resonance and synchronization will synergize, allowing the patient's Qi activities to be regulated and balanced at a rapid rate.

Because people vary in sensitivity to energy, some patients can be treated with Medical Qigong therapy immediately. Sensitive patients are usually taught Medical Qigong exercises allowing them to ingest energy from trees, plants, and other things of the natural world. Patients with low sensitivity may have difficulty in perceiving energy and often dissociate from their body during treatment. These patients should be given Shengong meditations to restore conscious awareness of the toxic emotions trapped within the body, thereby reestablishing a harmonious relationship with the Mind (see Chapter 32). The patient's insensitivity to the doctor's energy extension is an indication that the patient's spirit has gone out of the body in order to avoid feeling the emotions which arise from Medical Qigong therapy.

A patient insensitive to energy projection can still be in touch with basic surface feelings (i.e., hot, cold, tired, etc.); however, the deeper emotions are hidden, or energetically walled off through denial. The better the relationship between the patient and the doctor, the greater the patient's trust and surrender to the doctor's energy.

When observing the patient's Qi and energetic structure, the doctor is able to perceive and enter the patient's five energetic fields. These fields form the matrix of the patient's energetic structure, which constitute a blueprint for the body's physical structure. These energy fields have their own biological components and degree of fluidity.

These five energetic fields form the body's Sea of Energy. The term Sea of Energy (Qi), in this context, refers to the description used to explain the internal and external currents of life-force energy which surround, permeate, and function within the physical body. These energetic fields are described as follows.

1. The External Fields of Wei Qi (Protective Energy),
2. The Internal Current of Ying Qi (Nourishing Energy),
3. The Internal Sea of Blood (Xue Hai),
4. The Internal Sea of Marrow (Sui Hai), and
5. The Center Core of Light (Taiji Pole).

Within these five energetic fields, the protective and nourishing energetic fields of Wei and Ying Qi expand and contract, move and adjust their energetic flow according to the body's respiration patterns. The Sea of Blood resonates within the pulse rhythm established by the heartbeat. The Sea of Marrow undulates from the rhythm established within the body's micropulsation of the sacrum and cranium bones' flexion and extension. The energetic pattern of the Taiji Pole vibrates at the same frequency as divine light (the Dao).

THE THREE EXTERNAL FIELDS OF PROTECTIVE ENERGY (WEI QI)

There are three external energy fields that create the body's Wei Qi field. The Wei Qi field's energy flows from around the regions of the body's external muscles, skin surface, channels, collaterals and channel points, and extends several feet out from the body's tissues. All energetic forms of the physical body, including the internal organs, blood vessels, nervous systems, etc., can be accessed and treated through the three extended fields of Wei Qi.

These three fields are distinguishable from each other by gradations of density, vibrational rate, light, and temperature. All three energy fields communicate with each other, interacting so as to form the body's subtle energetic boundaries.

Enveloped within these energetic boundaries are the body's external "energetic grids," spanning the body's energetic structure like a giant web. These energetic grids are the body's external energetic field patterns, etched through the three fields of Wei Qi like a three dimensional webbed hologram (similar to a three dimensional blueprint).

It is important to note, that any disease within the body's internal structure is also imprinted in this external energetic grid. Disharmony of the body's tissues often appears first in the Wei Qi fields, before manifesting within the body. If the energetic grid is not transformed (through emitted light, sound, and resonant vibration), the disease cannot alter its destructive course, because it follows the patterns dictated within the grid. Thus, for example, a tumor surgically removed often quickly reappears. Trauma, be it physical, mental, emotional, or spiritual, as well as drug and alcohol abuse, distorts, damages, and destroys the delicate energetic web which constitutes the body's external energetic grids, making the patient more susceptible to external environmental influences.

The external energetic grids function to protect the body from external pathogenic factors, as well as to receive and transmit environmental energy. Thus, for example, a person can receive, absorb, and digest other people's emotions. The five senses are constantly receiving on outside environmental influences, which directly and indirectly affect the body's major organs. This constant influx of data, changes the body's emotional patterns, and forces individuals to adjust their emotional perspective frequently.

Researchers conducting tests in San Antonio, Texas, for example, determined that when subjects are being stared at, galvanic skin responses are considerably increased. These increases occurred even though the subjects were not consciously aware of being observed. The research determined that the body has its own expansive field of subtle awareness, which reacts even when the individual is not consciously aware of the body's energetic extension.

By reconnecting the spirit and emotions with the mind and body, the Qigong doctor is able to perceive the subtle Messages that the patient is unconsciously conveying (see Chapter 25).

THE THREE ENERGETIC BARRIERS

The Wei Qi fields are experienced at three different levels. These levels are differentiated by energetic barriers, which protect the body's boundary system. These barriers not only protect and contain the body, they help establish the individual's sense of reality, and sense of self. The third energetic field or "outer spiritual barrier" is depicted in the illustration as the outside speckled field of energy, and affects the second energetic field or "middle emotional/mental barrier." The second field of Wei Qi is depicted in the illustration as the middle white field of energy, which in turn influences the first energetic field or "physical barrier," seen as the black field of energy in the illustration (Figure 13.6).

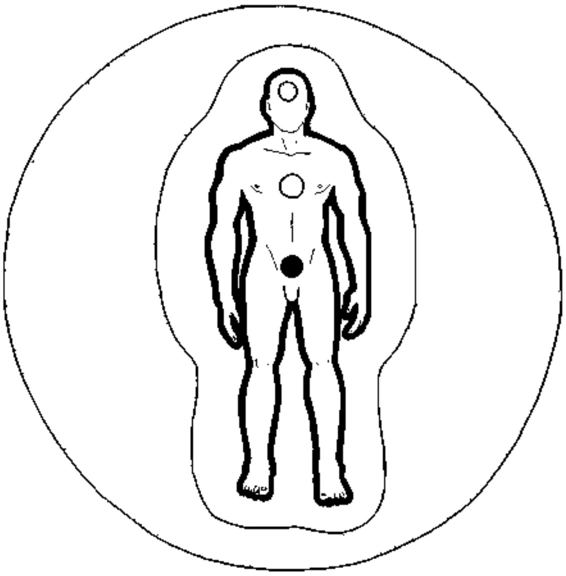

Figure 13.6. The Three External Fields of Wei Qi

These three Wei Qi barriers are always interesting for the doctor to feel, because they change so much from patient to patient, and sometimes from moment to moment. Some patients have very strong energetic fields, they are very palpable and full. Emotions strongly affect this type of energy field. The Wei Qi field of a patient who is armoring out of fear, reticence, or self-protectiveness, can actually repel the doctor's hand back, if their field and intention is strong enough. This repelling reaction is helpful because it lets the doctor know how to pace the treatment, and encourage the patient to let go of the armoring. This allows the patient to consciously focus on relaxing and sensing the areas in the body that he or she was trying to protect.

The energetic boundaries, act as gateways to the next level of energetic field energy. To perceive the patient's three fields of Wei Qi, the patient's energetic boundaries must be gently palpated. These boundaries are more easily sensed due to a thickening of energy which serves as a barrier to external pathogenic factors. Sometimes the doctor may have to move back into the patient's third or "outer" energetic field, and telepathically let the patient know that he or she will not be violated, and is safe. The doctor waits until the patient exhales (thereby expelling some tension) before entering and palpating the outer Wei Qi field. Once a release in the energetic field is felt, the doctor moves forward into the next energetic field. Because this same process occurs at all three levels, the Qigong doctor responds accordingly, waiting for the exhalation that allows them to proceed deeper into the next level of the patient's Wei Qi.

THE FIRST EXTERNAL FIELD OF WEI QI

The first external field of energy is confined to about one to two inches outside the physical body, and composed of lower etheric matter. It is related to the Lower Dantian and manifests as a holographic energy template for the physical body. It provides the spatial guidance for morphological development during the creation and formational stages of the embryo, as well as the morphological energetic patterns that assist in the repair of damaged tissues. Its energetic boundary is responsible for armoring the body against invasion of external pathogenic factors. This physical energy body is connected to the somatic tissues and cells, thus any physical-energetic interaction stimulates the body's external channels and points. This first layer is very dense and most Qigong doctors get the impression of physically contacting the patient several inches before actually touching the patient's material body. Its resonant vibrational pattern is very broad and slower than the other Wei Qi fields.

THE SECOND EXTERNAL FIELD OF WEI QI

This energy flows through the regions of the internal organs and tissues. It extends about a foot or more outside the body's tissues, and is composed of astral matter. It is related to the Middle Dantian and the emotional energy body. It contains a form of disembodied consciousness that can astral travel, and at times leave the physical body, as in near death experiences. This second layer of Wei Qi interfaces with the limbic system of the brain (the neurobiological seat of emotions) and is also linked with the sympathetic and autonomic nervous system. The second Wei Qi field's energetic boundary is responsible for protecting the body from negative emotions, destructive feelings and criticism from others. It is automatically activated when the patient is placed in a position where a lack of trust exists, or there is unwanted physical touch. The emotional energy body is connected to the visceral organs. Its main function is to receive, interpret, and verify an individual's emotions, feelings, desires, impulses, and thought patterns. It is through this energetic field that the patient's aura is typically observed as colors and light patterns. The second layer is more subtle then the first layer. When physically palpating, the sensations are similar to a magnetic attraction and repulsion. Its resonant vibrational pattern is faster than that of the first layer.

THE THIRD EXTERNAL FIELD OF WEI QI

This field extends a few feet to several hundred yards depending upon the spiritual evolution of the individual, and is composed of ketheric matter or substance. It is related to the Upper Dantian and the spiritual energy body. It is associated with intuition, inspiration, creativity, and visionary insights. Its energetic boundary is responsible for establishing security by informing the body of impending encounters, conflicts, or environmental transitions. This spiritual energy body is connected to the body's Taiji Pole, and is a container for karmic-related illnesses within the physical body. The Taiji Pole is known as the "root with many branches," as it is a repository of the soul's ancestral memories and experiences. Its function is to receive the subtle energy of the finest and fastest vibrations. It senses and interprets the data received from its environmental awareness and universal connection to the divine. The third layer is most subtle in physical sensations and requires delicate, intuitive contact. Usually the first impression when palpating this energetic field, is one of contact without a material component (like wind). It is sometimes cold, sometimes warm but always "breezy." Its resonant vibrational patterns are the highest and fastest of all the Wei Qi layers.

INTERNAL CURRENT OF NOURISHING ENERGY (YING QI)

The Ying Qi level of internal energy flows within the underlying tissues of the body and sweeps through the torso like a fast wind, expanding and falling with each respiration (Figure 13.7). The Lungs act as a great energy pump; each breath is like a gust of air flowing in and out, in a cycle of expansion and contraction through the entire body. Ying Qi swirls like a breeze; its function is to nourish the Blood, organs and, tissues. It flows within the blood vessels, as well as the channels and links the mind and intention to the body. It is responsible for the organization, construction, completion, and reconstruction of the vitality of the individual. It is integrated with the breath, and extends outward to the body's physical boundaries (skin and body hair).

The internal current of Ying Qi is most palpable when the doctor is gently touching the patient's surface tissue with his or her finger tips, and it feels almost like electromagnetic energy. Feeling the wind of Ying Qi can be likened to falling into "space" after passing through the denser outer field of Wei Qi. The action of Ying Qi, in most cases, is free flowing and connected with the breath. When listening to certain patients, the Qigong doctor may feel the Ying Qi flowing through the patient's channels, wrapping itself around the whole body like a vapor. It feels light, thin, and fairly cool compared to the next River of Blood level.

To access the "current of Ying Qi," the Qigong doctors begin by suspending their hands above the patient's first level of Wei Qi. They imagine dipping their hands into water as they allow their hands to slowly make contact with the patient's body. Once contact is made, the doctors extend their intention under the patient's skin. The doctors should match the patient's respiration and stay connected to the body until they feel the patient relax. Matching their respiration almost always slows the patient's respiratory rate, moving the patient into an altered state of consciousness. Sometimes at this point, age regression can be sensed, as the patient's body memory is triggered.

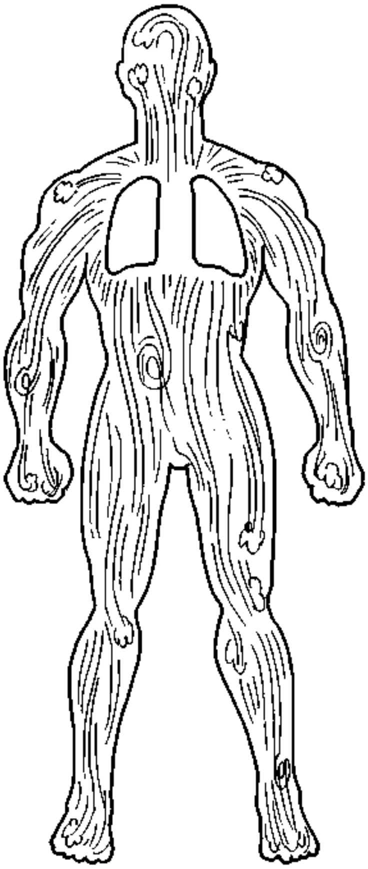

Figure 13.7. The internal current of Ying Qi flows from the Lungs and center of the chest through the entire body like a mighty wind.

SEA OF BLOOD (XUE HAI)

The Sea of Blood and its pathways constitute the body's internal network for circulating and maintaining the individual's life-force, under the direction of the Wu Jing Shen (the five separate fields that encompass the psycho-spiritual energetic transformations of the Five Yin Organs). The

Sea of Blood's energy flows through the blood arteries, vessels, capillaries, and the tissues of the major organs (Figure 13.8). By extending further into the body, the Qigong doctor perceives the patient's Sea of Blood; it feels like a warm watery pool in the center of the Heart, and flows through the entire body. The transition from Ying Qi to Xue Qi (Blood Qi) is easy to detect.

The energetic sensation of Blood is more substantial than the Wei or Ying Qi. The Blood being a liquid, is a denser form of Qi, and flows through the vessels, organs, and channels propelled by the Ying Qi. It is thick (this thickness varies from person to person), warm and somewhat "slick." The Heart's energy itself feels warm, substantial, and very alive.

As doctors gently touches the patient's surface tissue with his or her fingertips, the doctor extends intention deep into the patient's Heart. Upon contacting the Xue Qi, the doctor immediately becomes connected to the heartbeat of the patient. In some cases, the pulsing will connect and merge with the doctor's own heartbeat. Once the doctors encapsulates the patient's Heart, the strong emotions emanating from patients will usually subside. These feelings may be replaced by sporadic spurts of emotions such as relief, peace, or comfort, which sometimes invite the doctor further into the patient. On the other hand, some patients may react with feelings of fear, intense emotional pain, or even panic causing the doctor's hand to "bounce" away.

If the doctor's energy is "bounced back," he or she should immediately disperse the energetic charge into the Earth for ground purposes. The doctor should then either move to a more superficial energy level within the patient's body, or extend Qi into the patient's thorax or extremities. This technique allows the patient to relax after experiencing a Heart energy discharge. Since the pulse is simply another form of energy, circulating, expanding, and contracting with each heartbeat, the doctor should be able to ride the pulse out into the patient's extremities and viscera by listening to the Heart. This is similar to the listening ability of a master acupuncturist who, when reading a patient's pulse, extends his or her energy and intention through the patient's channels and into the organ of the Heart to diagnose the strengths and weakness of its congenital formation and current condition.

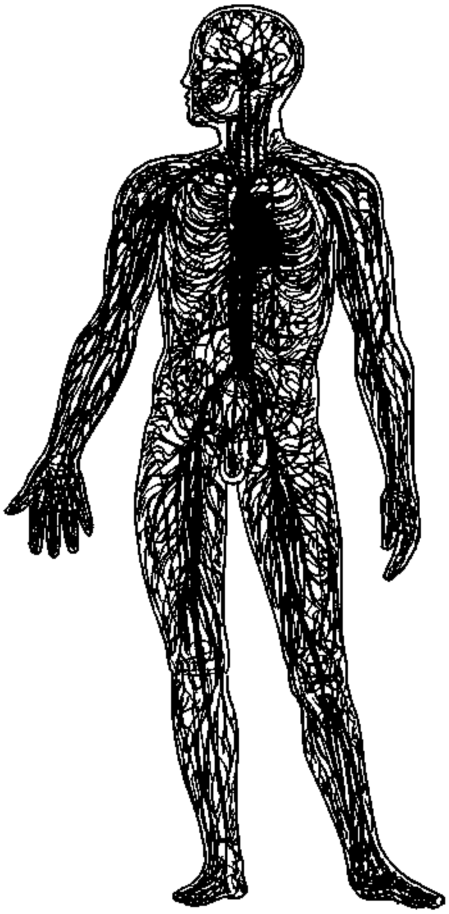

Figure 13.8. The Sea of Blood flows from the Heart through the body like a dense form of Qi.

SEA OF MARROW (SUI HAI)

This energy flows through the central core of the spine into the brain and enters the center core of the bones. In Traditional Chinese Medicine, the Marrow is not the same thing as bone marrow, but the stage prior to becoming bone marrow. Marrow is the pre-substantial Jing that forms the matrix for bone marrow, the spinal cord, and the brain (Figure 13.9). This flow of Sea of Sui energy is much slower than that of the River of Blood. It flows from the Kidneys' Jing, into the center of the spine, brain, and bone marrow. It has sometimes been called the "cranio-sacral rhythm" in Western medical terminology, and can be accessed from anywhere in the body. The Marrow energy generally feels cool and slick or slippery, with a texture like soft silicone gel. Once contact is made with the Sea Of Sui, it is very easy to utilize this pathway and travel throughout the patient's body. Its rhythm is slow yet manifests as a strong energetic pulse.

As the doctor gently touches the patient's surface tissue by dipping in with fingertips, into the Sea of Marrow, he or she can project intention into the center of the patient's spinal column. This connects the doctor with the rhythmic pulse of the Sea of Marrow. It is from this pulse that the doctor can determine the strength of the patient's Original Essence (Yuan Jing), Energy (Qi), and Spirit (Shen).

THE CENTER CORE OF LIGHT (THE TAIJI POLE)

This energy flows through the very center core of the body, uniting the three reservoirs of life-force energy (the Upper, Middle, and Lower Dantians). The fifth and final level of energy feels like a still, quiet place deep inside of the center core of the body's Taiji Pole structure. In China, Qigong doctors describe it as a holy place that almost defies description. It is observed as a vertical column or pole, of brilliant white light, full of pulsating vibrations. This energetic pole is surrounded by a vale of gold spiralling light which travels from the perineum to the top of the head.

When both the body's internal and external energetic wave patterns become synchronized through prayer or meditation, a resonate pulsation occurs within the Taiji Pole. This pulse begins deep within the center core of the body and resonates outward towards space. The Qigong doctors may also feel a powerful stream of energy leaving the body from the top of the head, connecting the Taiji Pole with the energy of the divine. The center core is a very reverent and special place, as it contains the essence of the patient's Eternal Soul, devoid of all ego, masks, or defense mechanisms (the patient's True Self).

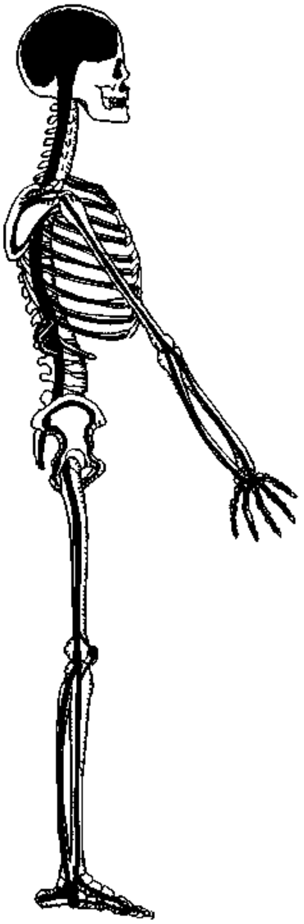

Figure 13.9. The Sea of Marrow flows from the kidneys through the spinal cord, brain, and bone marrow.

If the doctor penetrates the outside veil surrounding the patient's Taiji Pole, and extends his or her intention into the light's center, the doctor experiences the sensation of falling into space, or stretching out into eternity. Sometimes the sensation of shooting through space is followed by flashing colors and moving shapes, as the surrounding core dissolves into the Void (Wuji) itself. This same sensation and experience can also be personally accessed through deep prayer and meditation, and is quite a normal phenomena within deep spiritual practices (Figure 13.10).

SHEN: THE SPIRIT OF LIFE

The Chinese character ideogram for Shen is formed to symbolize "that which drops from the sky," and "that which passes through the body." In Medical Qigong therapy, Shen is considered the subtle energy of the spirit, associated with the Heart. It is associated with the spiritual energy and has a radiant nature.

When energized, Shen flows upward like fire, extending through the body's Eight Extraordinary Vessels creating a healthy nervous system. When Shen is gathered in the Upper Dantian, it can be projected as light and energy through the eyes, and the Yin Tang (Third Eye) point. According to Traditional Chinese Medicine, all illness has spiritual roots. The Shen (Spirit) is regarded as an integral part of the diagnosis and treatment process. An old Chinese saying states, "When the Shen is abundant there will be cheerfulness, when it is insufficient there will only be dullness and sorrow."

Shen is developed from Jing and Qi. According to Medical Qigong theory, when the patient's Shen is strong the individual is able to lead energy to an injured area and speed up the healing process. This strong Shen energizes the Qi, enabling it to increase cellular production. When Shen is scattered, the functions of the human body are weak from fatigue, the Shen becomes unsettled, and the mind restless. When the Shen and Qi combine together they form a solid connection of light and vibrational energy. This union is utilized to infuse energy and transfer thought through intention. This can be achieved by either conscious or unconscious intent by the doctor.

SHENGONG SCHOOLS OF MEDITATIONS AND TRAINING

The Shengong (Spirit Skill) meditations emphasize the transformations of consciousness, as well as the development of perceptual insight and the birth of wisdom. From an ancient Chinese energetic perspective, the three primordial qualities of true wisdom, or divine universal consciousness are:

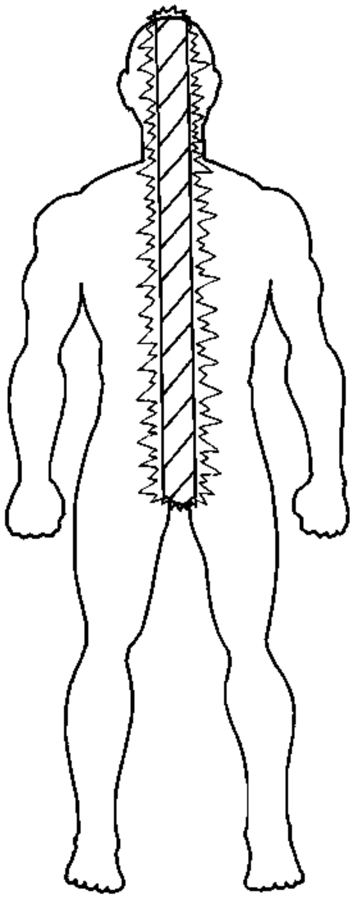

Figure 13.10. The body's center core Taiji Pole spreads from the base of the perineum to the top of the head.

1. The experience of emptiness within the Wuji (the Void), corresponds to the manifestation of the primal virtue of Wisdom.
2. The light manifested from the Shen's luminosity, corresponds to the primal virtue of Love.
3. The life-force energy inherent in all things, corresponds to the primal virtue of Power.

In China, there are three major schools of Shengong training. These spiritual schools of Qigong are described as follows.

- The Daoist School focuses on strengthening the body and mind equally with an emphasis on contemplating nature. Religious Daoism has its roots in Siberian Shamanism.
- The Buddhist School focuses on tempering the mind while giving little thought to the body.
- The Confucianist School focuses on regulating the mind to reach a state of quiet rest and peace through sincerity and the cultivation of moral character.

PRENATAL AND POSTNATAL SHEN

The body's Shen is the Supreme Yang aspect of the psycho-emotional components, the Heavenly Emperor or "master controller." It can be divided and classified into two primary forms through which the Spirit manifests: the Prenatal Spirit (Yuan Shen) and the Postnatal Spirit (Zhi Shen).

1. Yuan Shen is the body's spiritual element derived from Prenatal Jing (Yuan Jing) and Prenatal Qi (Yuan Qi). This aspect of the body's Shen is considered to be the intuitive Mind of the Dao. It is unborn and undying. It shines brightly and clearly. The Yuan Shen transcends conceptual thinking (which is the Postnatal aspect of the mind). It is immortal, unconditioned, primordial, and constitutes the mind's Yang nature.

   The Yuan Shen is produced and developed by the blending of both the male and female elements at conception (Body Fluids, Jing, Qi and Shen), and develops further with the nourishment from food and water. The Yuan Shen dominates all of the body's vital activities (heartbeat, digestion, motor function, etc.) and is also responsible for all intuitive and perceptive insight. The Prenatal Shen constitutes the true spiritual aspect of the Mind.

   By returning to a state of inner quiet and peace, through prayer, meditation, and adequate sleep, the Yuan Shen and the body's health are restored and maintained.
2. Postnatal Spirit (Zhi Shen) is the spiritual element derived from Postnatal Jing and Postnatal Qi. This aspect of the body's Shen is considered to be the conditioned mind or "will," also known as the discriminating, or acquired mind, and constitutes the mind's Yin nature.

   The Postnatal Shen is produced and developed by mental stimulation through contact with the environment after birth. The Postnatal Shen dominates all mental activity and is responsible for thinking and learned behavior. The Postnatal Shen constitutes the true analytical aspect of the mind.

   Prenatal and Postnatal Shen usually interact intermittently with each other in maintaining the body's health. Continual exhaustion of the Postnatal Shen impairs and suppresses the Prenatal Shen. Excessive studying or sleep depravation, for example, can lead individuals to mistake the projection of their deepest fears and desires for genuine intuitive perceptions.

THE FIVE SPIRITUAL ASPECTS AND YIN ORGAN COMPONENTS OF SHEN

Prenatal Shen affects the spiritual disposition which presides over man at the moment of conception. Closely tied to the Yuan Shen are five important, although seldom discussed outside of China, energies or "spiritual entities" called the Five Agents (see Chapter 2). The function of the Five Agents establishes the foundation for the psychological aspect of the Five Element theory used in clinical diagnosis throughout China today.

Each of these Five Agents is connected to, or said to "reside in" one of the Five Yin Organs. The Three Hun reside in the Liver and embody the virtue of Kindness, the Shen resides in the Heart and embodies the virtue of Order, the Yi resides in the Spleen and embodies the virtue of Trust, the Seven Po reside in the Lungs and embody the virtue of Integrity, and the Zhi resides in the Kidneys and embodies the virtue of Wisdom (Figure 13.11). Whereas each of the Five Agents embodies a virtue, each of the Five Yin Organs embodies the energetic qualities of one of the primordial elements. The Liver embodies the element of Wood, the Heart embodies the element of Fire, the Spleen embodies the element of Earth, the Lungs embody the element of Metal, and the Kidneys embody the element of Water.

Each of these Five Agent energies are connected to the five separate yet functionally interactive Yin organs, and thus create complex, multidimensional energy fields (Figure 13.12). These energy fields encompass emotional, mental, and spiritual shifts, as well as energetic transformations. The fluctuations depend upon which of the Five Agents is dominating from moment to moment. The psycho-spiritual attributes engendered are sometimes known as the Wu Jing Shen (Five Essence Spirits). As the universal energy flows through the Taiji Pole, these Five Agents are continuously being energized. The cultivation of the five virtues of these Five Agents nourishes the Eternal Soul.

The Qigong doctor needs to comprehend these multidimensional interactions to trace the complex spiritual, mental, emotional, and physical manifestations of Excess or Deficient conditions to the source or "root." Although Jing and Qi form the physical basis for the Shen, the Five Elemental Yin organs form the components for the Shen and the Five Essence Spirits. The energy of the Five Agents continually circulates through the patient's body, and are described below.

Five Agents		**Yin Organs**	
志	**Zhi**	腎	**Kidneys**
意	**Yi**	脾	**Spleen**
魄	**Po**	肺	**Lungs**
魂	**Hun**	肝	**Liver**
神	**Shen**	心	**Heart**

Figure 13.11. The Five Agents and Yin Organs

THE LIVER STORES THE ESSENCE SPIRIT "HUN"

The Liver stores the Three Ethereal Souls, also called the Hun. The Hun or Three Ethereal Souls are rooted in the Liver Yin (which includes the Liver Blood). The Liver organ is responsible for the free and easy flow of Qi throughout the body.

The Hun are typically characterized as the Yang, brighter, Heavenly souls. It is through the influence of the Hun that the five virtues of the Five Agents manifest. The Hun leave the body, ascending back to Heaven at the time of death. The ideogram for the Three Hun has two parts. One part is the character for Earthly spirits-Gui (Ghost), represented by a head suspended above a vaporous form of a body, with an appendage (symbolizing the whirlwind that accompanies the movements of the Earthly spirits). The other part of the character is the image for clouds that are seen as vapor rising from the Earth and gathering in the Heavens. The Hun move within the body as freely as clouds, following the Yi (Will) of the "Heavenly breath, within the celestial vault" (following the will of the Eternal Soul stored within the Heart, and originating in the Taiji Pole).

From the ideogram we also get a distinct picture of the spirit rising to the Heavens. The Eternal Soul is different from the spirit in the context of classical Chinese theology, in that the soul is seen as the more personal of the two, whereas spirits are seen as more universal temperaments or as archetypes. The ancient medical classics say that there are Three Ethereal Souls (Hun) and Seven Corporeal Souls (Po) that symbolize different attributes of the human being. The Hun's spiritual energy is said to be able to leave the body and then return, thus indicating a relationship with out of body travel into the spirit world.

The Hun and Po are expressions of the body's "true spirit." When disorder exists (physical, mental, emotional, or spiritual), the Hun may sometimes fly away (like startled birds in a yard), and the Po will either stir about thoughtlessly, in the absence of effective control, or become animalistic in nature and attack.

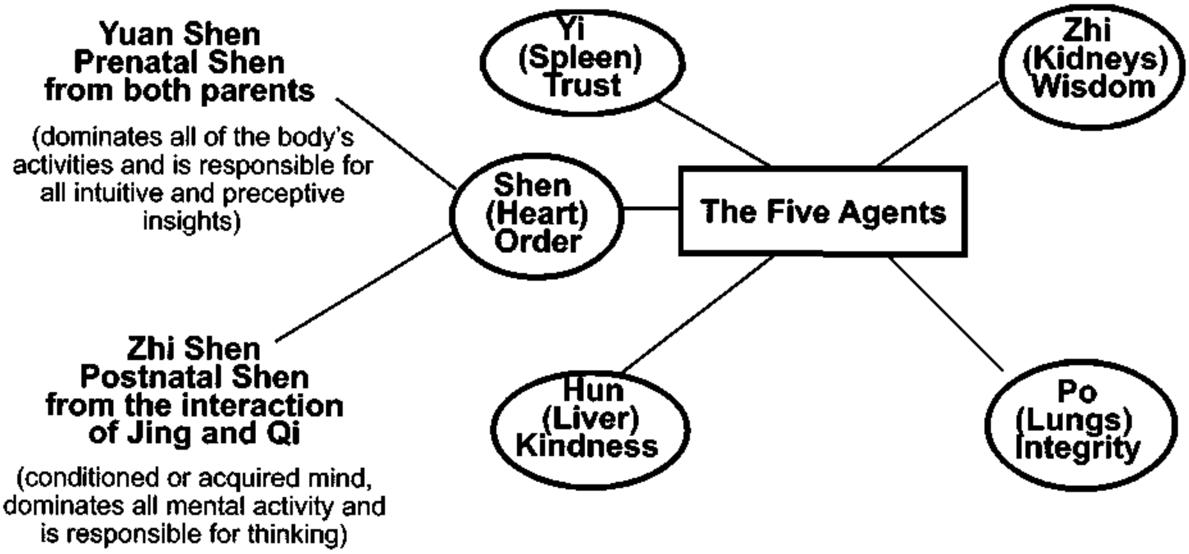
Yuan Shen
Prenatal Shen
from both parents
(dominates all of the body's activities and is responsible for all intuitive and preceptive insights)
Zhi Shen
Postnatal Shen
from the interaction
of Jing and Qi
(conditioned or acquired mind, dominates all mental activity and is responsible for thinking)
Shen (Heart) Order
The Five Agents
Yi (Spleen) Trust
Zhi (Kidneys) Wisdom
Hun (Liver) Kindness
Po (Lungs) Integrity

Figure 13.12. Formation of Shen

FUNCTIONS OF THE HUN

1. The Hun control sleep and dreaming. They reside in the eyes during the day and lodge in the Liver at night. When residing in the eyes, they see; when they lodge in the Liver, they dream. Dreams are the roaming of the Three Ethereal Souls. It is the nature the Hun to wander, and it happens easily. At night the Hun must be anchored or rooted in the Liver, so the Liver Blood and Liver Yin must be strong. If they are not, the Hun wanders, and the person dreams too much or has unpleasant dreams. Patients who suffer from severe Deficiency of Yin, may experience a floating sensation just before falling asleep. This condition is due to the Hun not being rooted in the patients' Yin.

   Dreams are an example of the information gathered during the Hun's traveling. The ancient Chinese Qigong masters understood that the body's Hun can traverse the Nine Levels of Heaven or Nine Levels of Earth instantaneously (see Chapter 2).

   The Hun store the sum total of past experiences. The expressions of the Three Ethereal Spirits are manifested through images, symbols and ideas from the divine and Wuji. These images, symbols and ideas emerge into the patient's mind, affecting his or her spiritual life. Without this interaction, the patient's mental and spiritual life would be deficient in images, ideas, and dreams.

   The Hun also controls dreams and daydreaming, as well as the ideals, aims, and direction in life. The absence of these objectives and goals results in feelings of depression. If the body's Liver Yin is depleted, the Hun are deprived of their residence, resulting in such conditions as: fear, excessive day dreaming, insomnia, and a lack of sense of direction or purpose in life (one of the main features of depression).

   Traditionally, sleepwalking is believed to be a function of the Hun. In sleepwalking, the body's Shen is not active and functioning, however the Hun are moving the individual. This is why when treating a patient for sleepwalking, the doctor will emit Qi into the patient's Hunmen Bl-47 (The Door of the Ethereal Soul). In modern research, however, some Qigong doctors have begun to theorize that the Po may take an active role in sleep walking, especially when the sleep state is related to incoherent trauma (patients who thrash about, hurt people, or destroy things when sleeping).

When using hypnosis or "creative visualization" in psychotherapy (the therapist may suggest a specific image or scene for the patient), the therapist is trying to bypass the patient's acquired mind and bring psychic material from the Hun, to avoid the judging aspect of the Postnatal Shen.

With regard to sleep and herbal medicine, if the patient does not sleep well because the Hun are wandering, the Qigong doctor can prescribe sour and astringing herbs (i.e., Bai Shao, Mu Li, Suanzao Ren) which will encourage the Hun back into the body.

2. The Hun assist the Shen in mental activities. This relationship is very important, as the energy of the Shen and the Hun must continually be coordinated. The Shen is related to rational thinking and inspiration. The Hun gives the Shen a sense of direction, projecting outwards, relating to other people. The Shen needs to direct the Hun, to "gather the Hun." The Hun give the Shen its movement and direction, encouraging the Shen to relate with people, to socialize and bond. The Hun, however, also need to be gathered and restrained by the Shen. Otherwise if the Shen is weak and does not control the Hun, then the Hun move about too much and have lots of ideas, but never accomplish anything, leading to frustration.

   It is important that the Shen allow the Hun to move in and out of the body, but in so doing must give the Hun direction and purpose. The Hun inspire and give us dreams. That is why the words "movement," "coming and going," and "swimming" are often used in connection with the Hun. In fact, many doctors say that the Hun are the coming and going of the body's Shen. In mediumistic trances, when the spirit of one person enters another's body, it is the Hun that come in; the host's Shen is disabled at that time.

   Because knowledge is dependent on the awareness of the Three Ethereal Souls, the Hun are therefore considered the collectors of information. The Hun however, do not interpret, rationalize or analyze the knowledge, they just pass the information on to the Shen, which is responsible for rational thinking, intuition, and inspiration. The Shen helps distinguish between the useful and non-relevant information.

   There is an interesting correlation with the Hun controlling sleep and dreaming, and the Hun being rooted in the Liver Blood and Yin. The Hun's moving in and out of the Liver governs the free flow of Liver Qi and vice-versa; the free flow of Liver Qi is a manifestation of the swimming energy of the Hun, moving in and out of the tissues.

3. The Hun maintain balance in one's emotional life, under the leadership of the Shen.

   Everyone experiences emotions in life; that is normal. The Hun are responsible for keeping a balance, so that the emotions do not become excessive and thus become the cause of disease. The Hun have a regulatory function, closely related to the balance between Liver Blood (Yin), and Liver Qi (Yang). The Hun's function here is the mental equivalent of the Liver's emotional regulating and harmonizing function.

   Due to its acquired nature, the Postnatal Shen discriminates, while the Hun does not. The relationship between the Shen and the Hun is very similar to concepts of consciousness and unconsciousness in Jungian terms. The Hun are a repository of images and archetypes, connecting the (personal) subconscious mind and the collective unconscious. If the Hun are unsettled, then the Shen (consciousness) is cut off, confused, isolated, aimless, sterile, and without dreams. The movement of ideas within the body has to be controlled or it gets out of hand, resulting in madness. This madness occurs to the degree that the Shen doesn't control the Hun anymore, and the patient receives an uncontrolled amount of emotional, mental, and spiritual input. All the energies and symbols coming through the Hun have to be integrated and assimilated. Otherwise there can be serious mental illness and possibly psychosis (except in young children who are continually full of ideas and have active imaginations). In a child, the Hun are very active and the Shen is not as restraining, so there is a continual flow of energies streaming from the unconscious world of symbols, without the result of insanity.

4. The Hun are responsible for the eyes. When the Hun wander in through the eyes, the eyes can see. The Hun give us vision, both mentally and spiritually.
5. The Hun influence a person's courage. If the Hun are not strong, the person is timid and fearful. If the Hun are strong, the person is fearless, can face up to the difficulties in life and dares to take action. A patient with weak Hun will have difficulty gathering information, making decisions, will lack courage, and become easily discouraged and apathetic.
6. The Hun control planning with the aid of the Shen and the sense of direction. The mental and spiritual confusion about one's role in life, what to do, what goals to set, can be compared to the aimless wandering of the Hun. This is a strong feature of major depression. If the Liver is strong and the Hun are strong and rooted, the person has a strong sense of direction.
7. The Hun control Spirit Travel. By housing the Shen for "spirit travel" (the spirit journeying outside of the physical body), it is possible for the Qigong doctor to consciously direct the Hun's travelling. This is different than Astral Travel, which is the Eternal Soul journeying outside of the physical body, connected to the Middle Dantian by a silver "cord of life." The Three Ethereal Souls, accompanied by the individual's consciousness, act as one unit which is sometimes referred to as the Spirit-Soul. The Spirit-Soul allows the doctor to know the exact location of the "spirit routes" travelled, as well as which path and direction it takes when it leaves the body. Otherwise, when the Hun wonder, the doctor "Shens out" and has no recollection of where he or she has been in the spiritual realms.

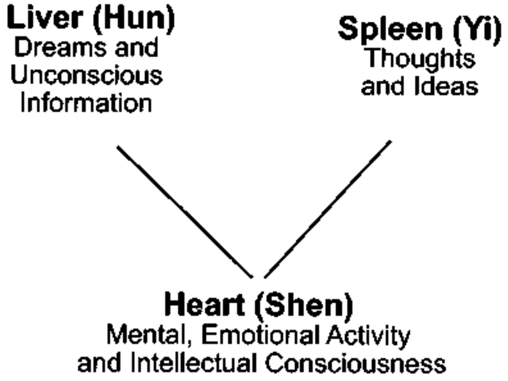
Liver (Hun)
Dreams and Unconscious Information
Spleen (Yi)
Thoughts and Ideas
Heart (Shen)
Mental, Emotional Activity and Intellectual Consciousness

Figure 13.13. Two Ways to Feed the Shen

THE HEART STORES THE ESSENCE SPIRIT "SHEN"

The Heart stores the Shen and is responsible for mental and emotional activity, intelligent consciousness, long term memory, and sleep. It is responsible for organizing mental concepts and activities, and has the capacity to judge.

The Shen is divided into two kinds of energies, Prenatal Shen and Postnatal Shen.

The Prenatal Yuan Shen is associated with the congenital spiritual energy that descends from Heaven. It is the universal divine spirit that is implanted in each person, uniting man to the spiritual realm (that which exists beyond the physical and mental reality). It has been compared to the "Christ within us," or "higher self." This is the energy Qigong doctors attempt to become filled with during sitting meditation practice and prayer.

The Postnatal or Acquired Shen is associated with the wisdom of the five senses and the acquired knowledge of the individual's experiences.

There are two ways to "feed" the Shen. One way is through dreams or unconscious information coming from the Hun of the Liver; the other way is through thoughts and ideas from the Yi of the Spleen (Figure 13.13). A person who has a well-developed, strong, and abundant Shen comprehends things immediately and is able to apply what he or she has learned.

The Hun give the Shen movement allowing the mind the capacity of self insight and introspection, as well as the ability to project outwards from the body and envelop, as well as relate to, other people.

THE SPLEEN STORES THE ESSENCE SPIRIT "YI"

The Chinese term Yi, is typically translated as mind, thought, opinion, idea, sentiment, inclination, intention, intellect, scholar mind, analytical thinking, and memorization in scholastic learning. Because the Spleen stores the Yi, it is responsible for the transformation and transportation of all thoughts and ideas on an intellectual level, through study, concentration, and memorization. This is different from the Hun's creativity. It is the interaction between the Yi of the Spleen and the Shen of the Heart that allows man to place how he thinks, speaks, and acts into sounds, thoughts, and actions.

Concerning memory, there is a lot of overlap between the functions of the Spleen (responsible for memory in terms of concentration and studying), the Kidneys (responsible for storing things on a day to day basis), and the Heart (responsible for storing past events).

The Postnatal Qi and Blood form the physiological basis for the intellect. A person with well-developed Yi has total recall of events experienced in their personal past, and is able to memorize things easily. A patient with an Excess of Yi will be obsessed with the past, while a patient with Deficient Yi will be absentminded, suffer memory loss, be inattentive, and have problems in maintaining concentration and mental focus. Although all physical pain is registered by the Po, and all psychological pain is registered by the Hun, only the memory of the pain is registered by the Yi (Figure 13.14). Therefore, a Deficient Yi condition can often lead to the phenomenon of transference and countertransference. Transference involves the incorrect attribution of certain emotions and intentions onto another person (usually onto the doctor by a patient). Countertransference involves the same type of projection of one's feelings, colored by one's own expectations in response. In countertransference, the doctor or therapist is doing the projecting.

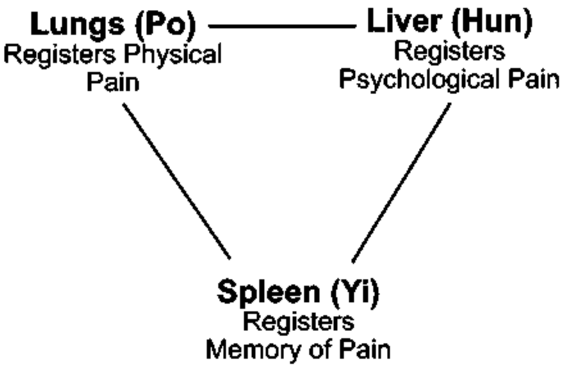

Figure 13.14. Three Ways the Body Registers Pain

Yi also has the meaning of divine purpose when applied to the Yuan Shen energy. Whenever the physical energy is purified by self-cultivation and through connecting to the divine, the intention of the Shen and the divine purpose become one and the same. It is said when the Yi is conserved, it will help build the Zhi (or willpower). The Yi is filled with information from the past along with current knowledge and sensations.

THE LUNGS STORE THE ESSENCE SPIRIT "PO"

The Lungs store the Seven Corporeal Souls (Po), which are physical in nature and are attached to the body's Jing and Qi. The word Po is defined as vigor, animation, or life. There are two parts to the Po ideogram; one is the character for Gui, the spirits of the Earth, and the other represents the color white. Thus, the Po are linked with a descending movement of energy and with the Jing. The Seven Corporeal Souls are also said "to come and go, enter and exit," in association with the body's Essences.

In Medical Qigong therapy, the body's Eternal Soul is seen as being strongly influenced by two main divisions of internal spiritual energies, the Hun and the Po. The Hun are the Three Ethereal Souls and represent the positively charged aspects of the Eternal Soul, while the Po are the Seven Corporeal Souls and are considered the physical, negative, Yin, heavy, and Earthly aspects of the Eternal Soul. The Po are the counterpart of the Hun and can be described by its following six manifestations.

1. The Po are the somatic expressions of the Eternal Soul, related to the reflexive nervous system and limbic system (the "reptilian brain"). They manifest through the body's sensations of feeling, hearing, and seeing. They have an impulsive tendency towards action and correspond to the deep animal instincts within the Mind (personal subconscious) and cells (our reflexes are a Po reaction). The Po also provide us with the animal strength and resources necessary to mobilize the body and perform incredible feats of power. The "animal within" is driven by the Po. They are the manifestation of the body's Jing in the sphere of sensations and feelings. Just as the Three Hun provide the individual with the energetic movement for the Shen, the Po provides the individual with the energetic movement of the body's Jing.
2. The Po are responsible for all physiological processes in childhood. In the beginning of life, the Po are responsible for the sensations of pain and itching. The Po serve as the intermediary between the Jing and the body's other vital substances. At conception the interaction of the body's Jing not only forms the embryo, but also establishes the Po within the body. Although the Po are stored in the Lungs, they stay attached to the body until death, there upon returning their energy back to the Earth within a few days.
3. The Po are related to weeping and crying. The interconnection between the Po and Lungs is very important from an emotional point of view. When the Po's movement in the Lungs is constricted, grief and sadness are suppressed in the chest through shallow breathing.

   In the morning time, when waking, if the patient feels dull and depressed, it is a sign that his or her Shen is clinging to the body, i.e., the patient's Shen is clinging to the excessive energy of the Po.
4. The Po are closely linked to breathing. Because the Po reside in the Lungs, all meditation systems, breathing exercises, and forms of breath control are methods used to regulate the Po, calm the Shen, and access the higher Hun states. Breathing is the pulsation of the Po. Each emotional change the body experiences is related to a shift in the body's respiration rhythm. Because breath control relies on the Lungs which influence the sympathetic and parasympathetic nervous systems, the quality of Qi and its circulation are dependent upon the method, speed, and quality of the breath. To support the greatest longevity possible, it is important to breathe with a long, slow, and even quality to the breath. The breath (air from Heaven) interacting with the Po in the Lungs, plays a significant role in the Heaven-Man-Earth concept for balancing the emotions.

   The Lungs' virtues are righteousness and courage. These virtues give a person the drive and strength to do the "right thing" when the need arises. These virtues manifest and promote good health when a correct energetic and balanced alignment with the Po is achieved (between the Lungs breathing pattern and the Po).
5. The Po are connected to sexuality on the sensation level. All instinctive sexual reactions and passions come under the authority of the Po. In human life, the Po are linked to the basic instincts and perceptions of the body. The Po provide the fundamental biological energy, and are the source of biological needs and impulses. They are driven by basic instincts and urges, and their sole concern is the immediate gratification of biological needs and impulses (emotional and physical survival, reproductive urges, etc.). Most self destructive behavior, such as an attraction to unhealthy and dangerous life-styles, is due to the Po.
6. Because of the Corporeal Souls' relationship with the Lungs and Large Intestine, the anus is considered the "Po Men" or "the door of the Po." It acts as a doorway for the elimination of the waste products for the Five Yin Organs, through draining off impure liquids and waste.

THE KIDNEYS STORE THE ESSENCE SPIRIT "ZHI"

The Chinese term Zhi (translated as "will") is the prenatal spiritual entity associated with the Kidneys, and is not the personal will of the ego that is driven by our desires. The term Zhi has two meanings. It means memory, but it also means willpower. Both are primary features of the Zhi. The word Zhi can also be used to mean the Mind (whole body consciousness and awareness). In connection with the five mental aspects of the Mind, the Three Ethereal Souls, Seven Corporeal Souls, the Shen, the "intellect," and the "willpower" are sometimes referred to as the Five Zhi.

MEMORY

Memory is defined as the ability to remember information when studying or learning a particular subject or pattern. The Kidneys keep a determined focus on our goals, not forgetting where we are going and what we are working to achieve; they maintain a vital mindfulness. The Kidneys relate to short term memory, whereas, the Heart governs the body's long term memory (Figure 13.15). This is why elderly patients, whose Kidneys are declining, often cannot remember what day or year it is, but can remember events long past.

WILLPOWER

This is the most important aspect of the Zhi, because it is responsible for supplying the mental drive, willpower, determination, and single minded pursuit of goals and aspirations. It enables the realization of ambitions by providing the focussed energy necessary to carry ideas to fruition. A powerful Zhi creates the magnetism and charisma necessary to manifest and materialize our dreams.

Even though a person has acquired all the information available via the Hun, Po, and Yi, without the Zhi there can be no action. Although the Hun give us the goal (thinks of something), the Zhi is needed to accomplish it (decides and acts on it). A person with well-developed Zhi demonstrates perseverance, determination, and a tenacity to complete personal goals. Patients with Deficient Zhi become indecisive and fearful. Patients with Excess Zhi have a blind obedience to authority, or tend to be fanatics.

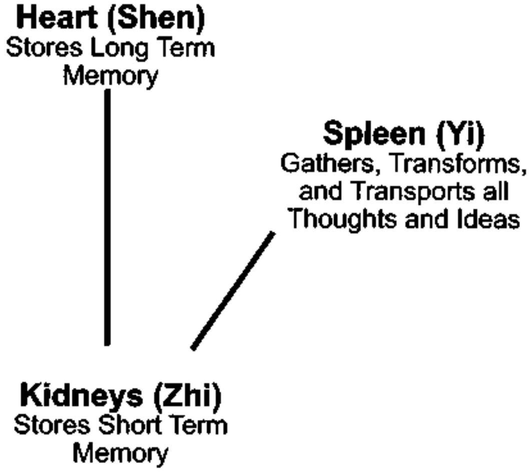

Figure 13.15. Three Ways the Body Stores Memory

The goal of the Qigong doctor is to reach such a level of attainment that his or her personal will merges with Heaven's will (Tian Zhi) becoming one and the same. Tian Zhi is considered the movement of the divine expressed in man as his personal virtue (De).

A man's virtue is defined as his spiritual righteousness, the authenticity of his heart and actions. It is through the acquisition of virtue that man finds and possesses his true nature. If man's virtue is initiated into the "mysterious" (embracing Heaven's will), his intuitive evolution will give way to limitless perception.

Tian Zhi is a divine inner prompting that guides us on our spiritual quest, if we are open to its message. Tian Zhi is what we can call the divine will and carries within it man's purpose in life.

THE FIVE YIN ORGAN CULTIVATION OF SHEN

The Shen indicates the vitality level, and the state of physical, mental, emotional, energetic, and spiritual being. A flourishing vitality is a manifestation of the Jing, Qi, and Blood inherent within the Five Yin Organs. These five components of the Shen are expressed through psychological manifestations (Figure 13.16). They form the bases of the patient's emotional characteristics. Each of the emotional components of the Five Yin Organs contribute to the patient's spiritual manifestation; adding or detracting from the natural harmony of the patient's spiritual nature.

The activity of conscious thinking, insight, and intelligence are the foundational aspects of the body's Shen, activated by the patient's Heart (wherein the Shen resides). The Heart is also responsible for the long term memory of the body's sensory memories (the perceptions, feelings and sensations related to hearing, seeing, smelling, tasting, and touching) and sleep. These sensory activities carry over from the other internal organs, due to their relationship with the Five Yin Organs. Emotional disturbances that affect the Heart can cause energetic changes in the patient's body, influencing health, toward recovery from diseases or illness by impeding or facilitating bodily functions (see Healing the Patient's Emotional Pain, Chapter 31).

When cultivating the Shen, it is important for the Qigong doctor to keep in mind that the spirit is indivisible from matter, and that an individual's emotional and spiritual well-being and development depend upon the healthy functions of the internal organs. When the Shen leaves the body (through shock, trauma, stress, depression, guilt, shame etc.), the Qigong doctor can observe a diminishment of glow in the patient's complexion, lack of luster in the eyes, along with changes in the respiration rate and state of mind.

Once the patient's Shen is disturbed, the Qi is affected. This in turn leads to the body's Jing becoming weakened. Patients with this condition are generally taught Medical Qigong exercises and prescriptions to not only tonify and strengthen their body's internal organs to improve their immune system, but also to improve the overall condition of their emotional/spiritual state. Even a low vitality of spirit can cause emotional disturbances that can scatter or stagnate the body's Qi.

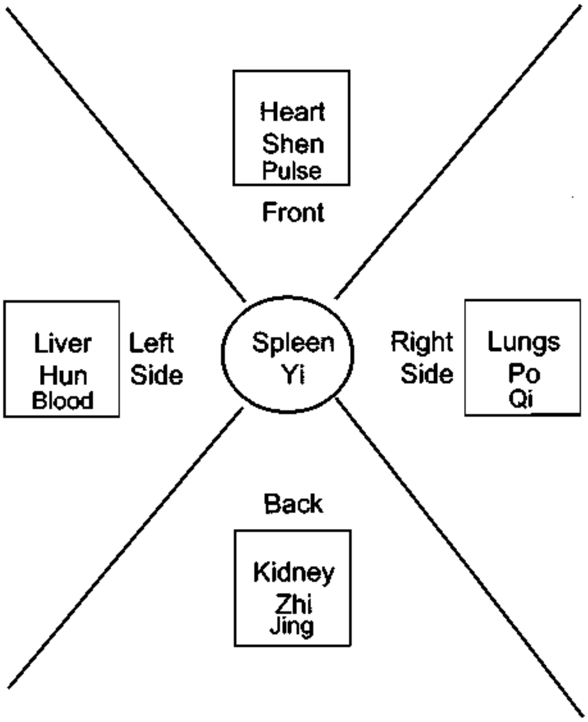

Figure 13.16. The body's Shen is physically housed in the Five Yin Organs. Each Yin organ houses a particular psycho-spiritual and energetic aspect of the human being. For example: The pulse is stored in the Heart, which is the residence of the Shen (Spirit), and placed in the front of the body, the Blood is stored in the Liver, which is the residence of the Hun (Three Ethereal Souls), and placed on the left side of the body, the Jing is stored in the Kidneys, which is the residence of the Zhi (Willpower), and placed on the back, and the Qi is stored in the Lungs, which is the residence of the Po (Seven Corporeal Souls) and placed on the right side of the body.

The Yi (Intention) resides in the Spleen and is placed in the center of the body, governing the Qi of the four aspects of spiritual manifestation, distributing them along the body's Taiji Pole. When Jing and Blood combine Shen is born. The Hun follow the Blood along in all its movements. The Po follow the Jing along in all its movements. The state of Qi and Blood of each organ can influence the Mind and the Spirit, any alteration of the Mind or Spirit affects one or more of the internal organs.

Advanced cultivation methods utilize the energy of the Five Yin Organs, flowing to and from the body's Taiji Pole. It is from the Thrusting Channels along side the Taiji Pole that the energy of the Shen is dispersed into the body's Yin and Yang Organs and then into the extremities (Figure 13.17 - 13.18).

SHENGONG CULTIVATION METHODS

The key to enhancing the Qigong doctor's development of energy, is through learning how to control the vast amount of Qi within the body, without compulsively "Shening out" (allowing the spirit to leave because the body feels uncomfortable). To avoid disconnecting from the body, the Qigong doctor practices several specific training methods of Shengong cultivation. When focusing on Shengong cultivation and regulation, there are four traditional disciplines the doctor must adhere to: Nourishing and Strengthening the Shen, Housing the Shen, Combining the Shen with Respiration, and Combining the Shen with Qi (Figure 13.19).

NOURISH AND STRENGTHEN THE SHEN

The Qi is able to nourish and strengthen the Shen through proper Qigong regulation. This is brought about through a form of coherent light meditation, that focuses the doctor's Shen on a small circle of light (generally about the size of a marble or pearl). The doctor's focus of concentration can extend either externally onto a specific point outside the body, or internally on a specific point inside the body (e.g., the Upper Dantian). Through this focused "coherent light" (light with wave patterns that are all in step or in phase), it is possible to build the Shen's strength and refine its power. Through focused intention on a specific area, the Qi and Shen condenses together forming a more powerful type of energetic projection.

HOUSING THE SHEN

By disciplining (i.e., quieting) both the emotions and thoughts the Shen becomes relaxed and tranquil, and remains in its residence (the Heart). The Shen must not be allowed to become attracted to outside emotional disturbances. The goal is to allow the Shen to be energized but not excited.

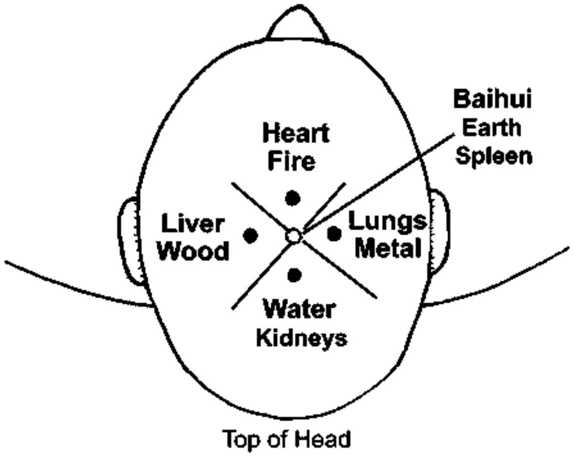

Figure 13.17. The Five Thrusting Channels as seen from the top of the head. Located within the Five Thrusting Channels are the "Five Portals of the Heavenly Gate." These gates include four entry points which surround the Baihui (One Hundred Meetings) point. These four points are known as the Sishencong or "Four Spirit Hearings" and are the access points for the upper spiritual matrix surrounding the body's Taiji Pole, and connect the Five Thrusting Channels of the upper gates with the Five portals of the Earthly Gates.

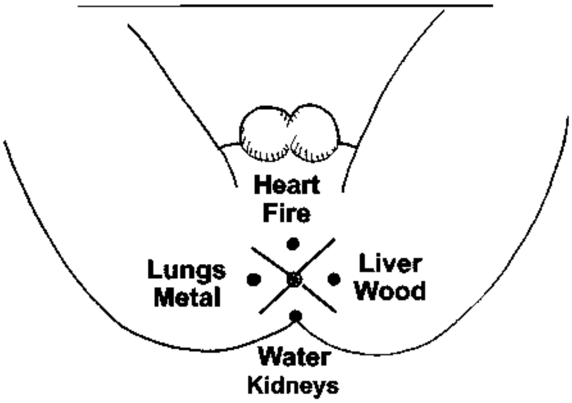

Figure 13.18. The Five Thrusting Channels as seen from the perineum. Located on base of the Five Thrusting Channels are the "Five Portals of the Earthly Gate." The Fire point is at the location of the Huiyin point on the Conception Vessel (CV-1), the Water point is at the location of the Changqiang point on the Governing Vessel (GV-1), the Wood and Metal points are located just below the ischial tuberosity, lateral to the anal sphincter. These points are known as the access points for the lower spiritual matrix surrounding the body's Taiji Pole.

COMBINING THE SHEN WITH THE RESPIRATION

Once the Shen has been nourished, strengthened, and housed, it is then connected with the doctor's breath. Once the Shen is combined with the respiration, maximum healing results can be obtained by releasing Healing Sounds and Qi projection simultaneously.

COMBINING THE SHEN WITH THE QI

The final stage is for the doctor to be able to direct the Shen in coordination with the life-force energy being emitted into the patient. Since Shen projects as light and Qi projects as vapor, this combination of Spirit and Energy projection is very powerful, especially when combined with healing colors and sounds.

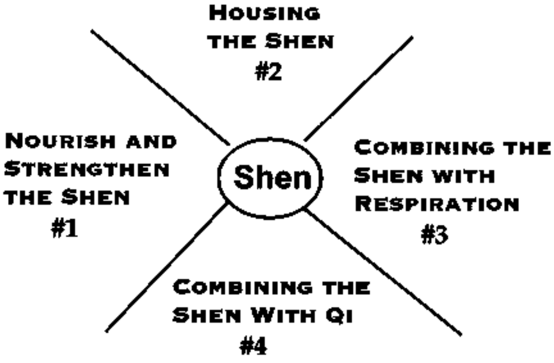
HOUSING THE SHEN #2
NOURISH AND STRENGTHEN THE SHEN #1
Shen
COMBINING THE SHEN WITH RESPIRATION #3
COMBINING THE SHEN WITH QI #4

Figure 13.19. Four Ways to Cultivate the Shen

THE SIX TRANSPORTATION'S OF SHEN

The psychic abilities developed in Medical Qigong training result from a well-developed intuitive faculty, brought about through Shengong meditations. Through these meditations the Yuan Shen can transcend the space-time continuum, as both space and time are multidirectional and interconnected.

Shengong meditations cause altered states of consciousness that differ substantially in subjective and objective observation from the general norms for that individual. As the Qigong doctor dissolves his or her Shen into the Wuji, the subconscious will be able to see forward and backward in time, unhampered by distance.

The conscious mind acts as a filter, seeking to analyze every perception, and attempts to identify and categorize each perception. Once the conscious mind is engaged, the flow of perception usually stops because the perceptions arising from the subconscious are usually vague (often with multiple messages). Qigong doctors must train themselves to first root (ground / stabilize) the acquired mind, in order for their perceptions to come through in a pure form, void of thoughts and judgments.

The ancient Qigong masters explained the ability to access the manifestations of Shen as follows: "by rooting the acquired mind (Zhi Shen), the Heart (Yuan Shen) opens up to 10,000 voices (Wuji)." Meaning that, when the "chattering mind" is anchored into the Earth through focused intention, the Heart is able to perceive the spiritual realm (the 10,000 voices), or be aware of its connectedness to the multidirectional space-time continuum.

Extraordinary abilities that develop from the cultivated spiritual consciousness include: telepathy, clairvoyance, clairaudience, precognition, time travel into the past, levitation, and teleportation.

The Qigong doctor experiences different psychological adjustments when opening to the deep spiritual transitions of Shengong meditations. The adjustments the doctor will experience include:

1. Time distortion,
2. Deep mystical states of perception and intuition,
3. Deeper and more powerful levels of self healing,
4. A more spontaneous reaction to movement and sound, and
5. Difficulty communicating verbally.

CULTIVATING THE SPIRITUAL CONSCIOUSNESS TO OBTAIN THE SIX TRANSPORTATION'S OF SHEN

In Shengong meditations, it is through the use of imagery and creative visualization that the doctor's altered states are developed, and a connection is forged with the Yuan Shen and Eternal Soul.

After the doctor's Shen has entered into a quiescent state, his or her altered state of spiritual conscious can be initiated through choosing one of the following five images.

1. The doctor is instructed to go inward into the center of the Taiji Pole to connect with his or her Yuan Shen and Eternal Soul.
2. The doctor is instructed to go into an imaginary temple or sacred place located within the astral plane to connect with his or her Yuan Shen and Eternal Soul.
3. The doctor is instructed to float upwards into the Heavens to connect with the eternal source of his or her Yuan Shen and Eternal Soul.
4. The doctor is instructed to ascend out of the body (after reliving the death of a past life experience) and receive the knowledge learned from that experience collected from by the Yuan Shen and Eternal Soul.
5. The doctor is instructed to allow the Yuan Shen to directly connect with the Eternal Soul and transmit information.

The Higher Self and Eternal Soul are usually considered as the same thing, and are often regarded as the "Master Within." In Jungian psychology, the collective unconscious (Wuji) is accessed through the subconscious mind (Shen) in the form of dreams or directly through "visions." The personal subconscious (Shen) has several elements:

- the anima, which is the wise man (influenced by the Hun)-for males, or
- the animus, which is the wise woman (influenced by the Hun)-for females, plus
- the shadow, which is the individual's dark side (influenced by the Po).

Once the Qigong doctor has cultivated the Shen into a high level of effective energy (absorbing and projecting energy from the universal, as well as the immediate environment), the result is a gradual unveiling of six supernatural diagnostic powers known as the Six Transportations of Shen.

These six metaphysical abilities not only enable the Qigong doctor to accurately evaluate a patient's state, but also to predict the probable future progressions of the patient's life as well as of his or her disease. The Six Transportations of Shen are: Observing the Present, Comprehending the Past and Observing the Future, Knowing a Person's Thoughts, Perceiving a Person's Destiny, Hearing the Sounds of the Universe, and Examining the Universe (Figure 13.20).

OBSERVING THE PRESENT

The Qigong doctor must first quiet the Zhi Shen's (acquired mind) analytical chattering by housing its energy within the Five Yin Organs. As the Mind becomes rooted, the doctor's breathing slows and the Lung Qi harmonizes with the Heart Qi. Once the Seven Po are quieted in their residence and the Zhi Shen is rooted, the Yuan Shen's messages can come through clearly without interference. The Yuan Shen is always connected to the Wuji where all knowledge is stored. The Yuan Shen is always at the "center of time" where it can gather and absorb the patient's past and present experiences, emotional reactions and thought patterns. It is from the Qigong doctor's intention and focus to be at the "center of time" that he or she can observe the present state of "now."

Observing the past is the ability to analyze and accurately discern the actual state of present situations (the here and now). By keeping away from conscious intervention, the Yuan Shen is free to truly listen and perceive as an "observer." The doctor's ego is not engaged and his or her mind's subconscious observer receives and reviews the patient without bias.

This enables the doctor to thoroughly analyze and comprehend events and situations clearly and unambiguously. It is through this state of perception that the Qigong doctor can accurately discern the true emotional problems which disturb his or her patient's energetic balance.

COMPREHENDING THE PAST AND OBSERVING THE FUTURE

Comprehending the past and observing the future is the ability to analyze and understand the thought patterns of the patient, the reasons these patterns take place, and to accurately predict the outcome of future events based upon those pat-

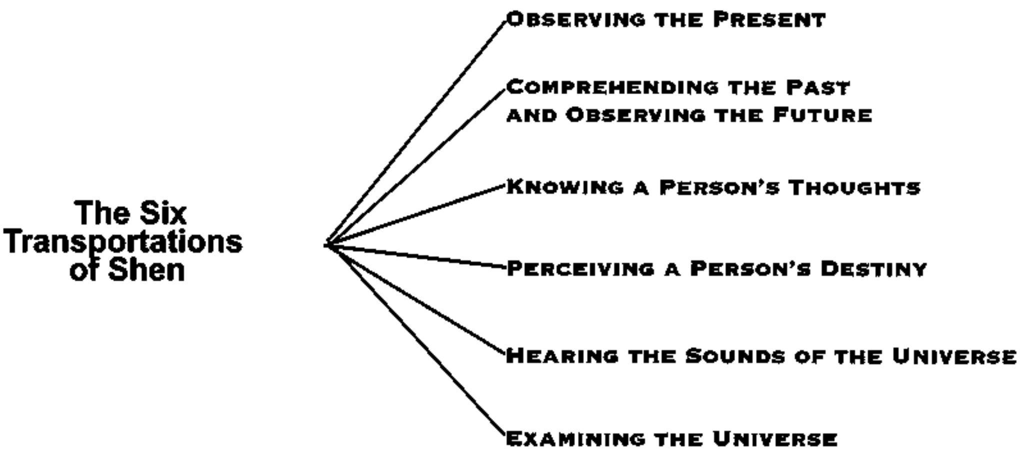
The Six Transportations of Shen
Observing the Present
Comprehending the Past and Observing the Future
Knowing a Person's Thoughts
Perceiving a Person's Destiny
Hearing the Sounds of the Universe
Examining the Universe

Figure 13.20. The Six Transportations of Shen

terns. By observing the patient from the active state of being in the "present," the Qigong doctor is free to travel within the infinite space (Wuji) enveloped within the patient's tissues. The doctor's Yuan Shen has no need to astral travel to obtain this information. The doctor's energetic connection to the patient's energy body gives him or her direct access to the personal information needed. It is from this internal space that the Qigong doctor can begin to access the memories existing within the patient's body (stored within the cells of the tissues).

At the subatomic level, the body and its cells contain a great deal more energetic space than matter. The doctor can directly access the internal Wuji enveloped within the patient's tissues and cells to retrieve the patient's past memories and emotions. Through tracing the patient's emotional, mental and spiritual development a greater understanding of the patient's present condition is achieved. After the doctor's Yuan Shen has gathered this information, it can be further analyzed by the doctor's Zhi Shen, and organized to accurately determine the cause and effects on the patient's present state.

Through understanding past patterns of thought and emotional reactions, the doctor can then predict future patterns of change and transition. These predictions are quite accurate, since people are prone to repeat programmed patterns indefinitely, unless there is an internal spiritual transition that frees them from this subconscious patterning.

Knowing a Person's Thoughts

This is the ability to sense the thoughts, judgments, fears, and emotions of another person, knowing their innermost fears and subconscious thoughts. The subconscious mind of the doctor communicates with the subconscious mind of the patient. This is a nonverbal, telepathic communication. It is therefore important for the doctor to monitor his or her thoughts and feelings, as they are automatically being projected into the subconscious mind of the patient. After the treatment is over the doctor may discuss with the patient any important information received that would be helpful to the patient.

Perceiving a Person's Destiny

This is the ability to predict future events and upcoming transitions in an individual's life based on the doctor's understanding of the patient's past energetic patterns. Through the Yuan Shen's ability to perceive the past and present, the doctor is able to understand the patient's mind, thoughts, acquired characteristics and personality. Above all, however, the doctor is able to perceive the individual's true nature that exists behind all of the patient's external masks. The doctor's Zhi Shen (postnatal mind) is then able to analyze and categorize the individual's patterns. Through deductive thinking, the doctor is able then to perceive the probable outcome and development of these repetitive patterns. This enables the doctor to help the patient to identify the patterns that require restructuring and give added encouragement and advice, when needed.

This spiritual counsel or guidance is extremely important for healing transitions since old patterns established in the three energy matrixes (Jing, Qi, and Shen) must be changed together to achieve a more permanent healing. Otherwise, the illness is likely to reappear. The spiritual matrix sets the foundation for the energetic matrix, which in turn sets the foundation for the physical matrix.

HEARING THE SOUNDS OF THE UNIVERSE

This is the ability to hear sounds, music, and voices not audible to the normal hearing. This is not to be confused with schizophrenia -- a mental illness that causes a person to hear voices; the difference is that the schizophrenic cannot think clearly, the mind is very confused and unordered.

The ability to hear the sounds of the universe is also the ability to be aware of, and receive, divine inspiration or messages (clairaudience). These abilities allow the doctor to affect the lives of those they know or soon will know. The doctor's own inner-voice can guide, instruct, teach, and lead the doctor to ever higher states of enlightenment.

Through inner hearing, the Qigong doctor is able to listen to, as well as understand, the resonant sounds generated by life-force energy and spiritual entities. This enables the doctor to tap into an unlimited source of divine knowledge and wisdom, which in turn allows the doctor to be present for the patient's highest good.

EXAMINING THE UNIVERSE

This is the ability to actually perceive current objects, events, or people that may not be discerned through the normal senses. It is the ability to perceive both time and space on a spiritual dimension (clairvoyance), to understand and acknowledge one's part, and to accept one's divine calling in life. It is also the ability to understand how one's spiritual calling affects both the cosmos and mankind. Through inner-vision the Qigong doctor is able to see inner-visions of divine enlightenment, observe and exchange information with celestial beings, and find holy places for spiritual sanctuary.

The metaphysical abilities of the Six Transportations of Shen are spiritual manifestations from the Eternal Soul. Once the Qigong doctor regulates his or her spiritual life, these abilities become a natural course of experience, giving way to the development of eight supernatural powers. A chart describing the interaction of the body's Eternal Soul with the energetic functions of the Three Dantians and Five Yin Organs' Wu Jing Shen (Five Essence Spirits) is depicted in Figure 13.21.

THE EIGHT SUPERNATURAL POWERS

Once an individual has obtained the Six Transportations of Shen, his or her psychic abilities develope into eight supernatural powers. These powers can be utilized in the individual's personal spiritual practice. Through these eight supernatural powers the doctor's Yuan Shen can allow him or her the ability to transcend all physical boundaries and invoke certain spiritual manifestations. The eight supernatural powers are described as follows.

1. The ability to make the one's physical body light.
2. The ability to make the one's physical body heavy.
3. The ability to reduce one's energetic body to the size of an energetic particle. This allows an individual the ability to astral travel into the subtle energetic and spiritual realms of existence.
4. The ability to expand one's energetic body to an infinite size. This allows an individual the ability to connect to, through, or envelop objects, as well as energetically protect people, places and things.
5. The ability to call into existence and attain one's specific needs or desires. Due to the individual's strong connection to the divine,

he or she may manifest through faith that which is desired.

6. The ability to experience and enjoy one's desires fulfilled.
7. The ability to gain authority over one's situations. First and foremost, the individual will obtain the spiritual authority over a specific situation, then the authority over the emotional, mental and physical realms follow.
8. The power to attract and control things. This allows an individual the ability to engender help and security (in the form of outside council, finances, and spiritual and emotional support) in order to exercise power over specific situations.

SOUL PROJECTION, SPIRIT PROJECTION AND ENERGY PROJECTION

The phenomena of Soul Projection, Spirit Projection and Energy Projection is known and accepted as a natural part of existence in almost every culture in the world. The holy men and women, shamans, and energy healers of many primitive societies were required to prove their powers by exercising these spiritual and energetic abilities.

Tibetan Buddhism (influenced by the energetic practices of Indian cosmology) contains elaborate doctrines about Soul Projection (also known as Astral Projection), Energy Projection, the Astral Body, and also describe the feats individuals can perform while out of their physical bodies. Some of these doctrines are quite similar to the traditional Daoist teachings which are evident in advanced Daoist Shengong training.

All Chinese systems used for inducing Shen and Qi projections use visualization, concentration, willpower and the individual's innate skill. The three types of external projections are: Soul Projection, Spirit Projection, and Energy Projection.

- Soul Projection occurs when the body's Eternal Soul is externally projected out of the body, while still attached to the Middle Dantian through a Silver Cord;
- Spirit Projection occurs when the body's Shen is externally projected outside of the body as an energetic vehicle of light;
- Energy Projection occurs when the body's Qi is externally projected out of the body like an energetic vapor (Figure 13.22).

In Soul Projection, Spirit Projection and Energy Projection, the effects of time are not acknowledged outside of the physical realm. Within the Astral Plane, time is omnipresent, there is no past or future, only the individual's perception of it.

The body has an established relationship to both the physical and spiritual world. The physical world of the living is considered a Yang state of existence, while the supernatural-spiritual world is in a Yin state of existence (see Chapter 19). As the Qigong doctor's Shen reaches a high state of sensitivity, he or she can transcend the ordinary senses and normal conscious perceptions. These abilities enable the doctor to observe, communicate and even travel within the "higher" and "lower" spiritual Yin world. There are nine energetic levels of Heaven (i.e., the higher astral planes), and nine energetic levels within the Earthly spiritual realm (or lower astral plane).

It is important to note that the term "higher" and "lower" astral realm has nothing to do with the actual space, but with the energetic and spiritual vibrational levels. The higher and lower astral planes overlap and exist within the same space as the physical plane. This blending between the higher planes, lower planes and the physical plane creates energetic dimensions (or places) where the separation between these worlds is thin. In these energetic dimensions, lower beings can reach and attach onto an individual – even though the individual is not directly within the astral regions.

The energetic frequency of each astral being determines the spiritual level wherein it exists. Beings of higher frequency, for example, vibrate at a faster rate and reside in the higher, less dense astral dimensions; whereas beings of lower energetic frequency can only dwell in the dense lower astral levels.

Negative feelings (e.g., worry, fear, anger, etc.) produce a lower vibrational rate that can not only keep an individual's spiritual vibration at a lower level, but can also attract lower level astral beings to that individual.

Soul Projection	The Eternal Soul is projected out of the body, attached to the Middle Dantian through a Silver Cord.
Spirit Projection	The Shen is projected outside of the body as an energetic vehicle of light.
Energy Projection	The Qi is projected out of the body like an energetic vapor.

Figure 13.22. Soul Projection, Spirit Projection and Energy Projection

SOUL PROJECTION

Soul Projection is the skill of sending the body's Eternal Soul (at will) to other places, times and dimensions (or worlds) in order to retrieve knowledge and experiences that assist the individual in his or her spiritual evolution. It is a nonphysical world in which time and space have no meaning or influence. The astral dimensions form a parallel universe that exists both within and outside ourselves.

The Eternal Soul is a fluid energetic form that can disengage from the physical body and travel about as a disembodied apparition taking the human consciousness along with it. It can pass through walls, travel to higher dimensional worlds, and encounter nonhuman creatures and mystical masters. Many Qigong masters use Soul Projection to seek spiritual knowledge, travel through time to view the past, explore other spiritual dimensions, engage in astral sex, or view the "infinite knowledge" (i.e., Messages) contained within the Wuji.

When the Eternal Soul begins to leave its energetic chamber, the brain wave patterns slow to a theta pulse. The body may feel paralyzed, and begin to vibrate as if an electrical current were passing through it.

There is also an odd roaring sound heard in the ears before the Eternal Soul separates from the body's tissues, or sometimes buzzing noises, musical notes, humming sounds, or chattering voices can be heard. When leaving the physical body, sometimes a swaying sensation can be felt as the astral body separates from its physical shell. The astral body can either spin out or float out of the physical body. When traveling in the astral body individuals may observe streaks of light flashing behind them as they travel.

As the Eternal Soul travels, a silver energetic cord is attached from either the top of the head (at the Baihui area) or at the Lower Dantian (at the umbilical area). This energetic cord is the media through which an individual retains his or her awareness of the physical self while projecting. This energetic cord is constructed of subconscious spiritual energy and maintains the individual's protective perceptual awareness of his or her physical body. It is believed by some that if this cord is severed the individual will die.

When the Eternal Soul (in its astral form) approaches someone, there is a feeling of electricity and coolness that is general felt. While in its phantom form, the Eternal Soul can display telekinetic powers (e.g., knocking on walls or tables, closing a half open door, etc.).

Warnings and Contraindication

Do not practice Soul Projection just before, or during, thunderstorms as the electromagnetic buildup in the atmosphere can hamper your ability to leave or return to the body. Also avoid approaching any kind of power-line during Soul Projection, there is a risk of getting caught by the power-line and becoming "stuck" (like a fly in a spider web).

1. A Body of Light	This energetic form resembles an energetic duplicate of the individual's physical self. It involves creating a near exact energy body duplicate of oneself (often with improvements, e.g., no wheelchair, or the elimination of other physical challenges).
2. An Animal Form	This energetic form resembles that of an animal (known as Shape Shifting). It is an ancient Daoist Shamanistic skill used to transcend the astral plane, and to take on the consciousness and energetic behavior, skills and perceptions of the animals the shamans assumed.
3. A Ghost or Saint	This energetic form resembles that of a Ghost or Saint (also known as an Invocation). It involves energetically taking on the image, form, and personality characteristics of a departed individual or spiritual deity.
4. The Immortal Observer	This energetic form resembles that of an All Seeing Eye (also known as The Astral Eye). It involves creating an energetic ball of light in the Upper Dantian at the Yin Tang (Third Eye) area. The large eye surrounded by a sphere of light is released from the Yin Tang area to gather psychic impressions via a long energetic cord attached to the Upper Dantian. As the eye returns, its energetic form is absorbed back into the individual's Upper Dantian. These gathered images are then analyzed and the information is assimilated before the energy is dispersed into the individual's Lower Dantian (and then down the legs and into the Earth).
5. The Element Body	This energetic form usually resembles that of different elements of nature, such as wind (as vaporous mist), Water, Earth (the terrain) or Fire. The ancient Daoist would use these elemental forms of nature as a disguise in order to obtain information or spy on hostile intruders.

Figure 13.23. Types of Spirit Bodies Used in Spirit Projection

Do not practice Soul Projection while recovering from an acute sickness (the drain on your energetic reserves can make you more susceptible to further illness).

SPIRIT PROJECTION

Spirit Projection is the skill of sending the body's Shen (at will) to other places, times and dimensions (or worlds) to retrieve knowledge and experiences to assist the individual in his or her spiritual evolution.

Spirit Projection (also known as mind projection, remote viewing, mind travel or traveling clairvoyance) can entail a combination of "genuine" observation with an extrasensory perceptional ability to gather information. When individuals perform traveling clairvoyance in the form of Spirit Projection, they can send their consciousness anywhere in the world. Awareness is not physical in nature, it is experiential, totally nonmaterial and cannot be limited to, or fully contained within the body.

Spirit Projection is practised through the creation of an energetic vehicle for the spirit body, such as a body of light, an energetic form of an animal, or a ball of light. An energetic shape is formed through the imagination and intention. The individual then transfers the energetic consciousness from his or her physical body into the energetic vehicle. This energetic form can resemble various patterns shapes and sizes to serve the individual's purpose (Figure 13.23).

As the individual thinks about a person or place, his or her body will be pulled to that object like a magnet, followed by a flood of images. The spirit body can fly over various terrains as it travels (e.g., oceans of water, deserts, mountains, forests, etc.). Once the spirit body arrives at its destination, the individual is able to feel and experience the area's environmental Qi (e.g., the humidity, wind, cold, heat, light, darkness, etc.).

After traveling in the spirit body the individual must reabsorb the energetic body back into his or her physical body to retain a clearer memory of the experiences.

Warnings and Contraindication

Do not perform too many Spirit Projections within a relative short time span. Excessive Spirit Projection has a tendency to weaken the Kidneys and drains the body's Wu Jing Shen.

While traveling, be careful of what your intentions are, as like begets like. Any negative projection, emotion or evil intention brought into the astral plane will quickly rebound back to the individual, or may attract negative entities that feed off of these lower vibrations.

ENCOUNTERING ASTRAL BEINGS

Within the subtle realm of the spiritual world, the existence of spiritual entities and ghosts are divided into Yin and Yang. While traveling in the astral realm, an individual can encounter either type of spiritual entity.

1. The Yang Beings native to the higher astral realm vibrate at a faster energetic rate, and are associated with more advanced, highly evolved beings, such as angels, guides, beings of the light, and enlightened masters.
2. The Yin Beings native to the lower astral realm vibrate at a slower energetic rate, and are associated with lower, less evolved beings, such as devils, demons, spirits of darkness, evil spirits, and other similar negative spiritual manifestations.

In all truth, the astral body cannot be hurt the same way that the physical body can. The astral body can, however, carry back to the physical body mental, emotional, and spiritual scars, or hitchhiking negative astral beings, all of which can negatively affect the physical body.

All actions and reactions on the astral plane, whether productive or destructive, are a product of the individual's belief system (you bring about each condition according to your faith).

ENERGY PROJECTION

Another type of life-force energy that the body can project outside its physical structure is Qi. Qi Projection is emitted outside the body like a fine vaporous mist, enveloping and sensing everything that it touches. Similar to Soul Projection and Spirit Projection, Qi Projection is directed according to the individual's intention, imagination, will, and faith. Distance therapy, pertaining to Qi emission, can be observed as energy leaving the doctor's hand like heat waves emanating towards an external object. Within this emitted life-force energy is contained the doctor's Shen (Spirit), as well as the energetically encoded "Messages" that the doctor sends to his or her patient. This constitutes the energetic foundation for Medical Qigong therapy.

SECTION III

DAO YIN TRAINING

CHAPTER 14
INTRODUCTION TO DAO YIN TRAINING

Qigong developed into a systematic healing art for health preservation during the "Warring States" period in Chinese history (476–221 B.C.). Records from that time contain documentation of symptoms, treatments, and specific points of attention given within the body and energy channels, by doctors practicing Medical Qigong, as well as details of the effects of these Qigong exercises.

The Huang Di Nei Jing (*The Yellow Emperor's Classic of Internal Medicine*) states, "to treat diseases it is important to use both Dao Yin and An Qiao." These are two names given to Medical Qigong therapy in ancient times.

"Dao" refers to physical movements, which are guided by the strength of the mind to stimulate the body's Qi and bring about a state of relaxation. "Yin" refers to the fact that with the aid of physical movements the body's Qi can reach the extremities (feet, fingers, and head). Dao and Yin together mean to lead and guide the energy through both motion and quiescence. This therapy includes mental, respiratory, and postural exercises.

"An" has two meanings, "massage" and the "up and down movement of the wrists." "Qiao" means "to lift up." An Qiao refers to the methods of the postural Dao Yin exercises. The basic hand and body methods of postural Dao Yin training are as follows:

- Kai (Opening)
- Fen (Separating)
- He (Closing)
- Tui (Pushing)
- Rou (Rubbing)
- Xuan (Rotating)
- An (Pressing)

These are the foundational movements a Qigong doctor will use in order to purge, disperse, gather, tonify, emit and guide Qi. The training of energy is achieved and implemented primarily through a combination of dynamic (active) and static (quiescent) exercises and meditations. Although Medical Qigong schools are diverse and complex in their form and content, they can all be divided into two main branches of energetic cultivation: Static Qigong and Dynamic Qigong (Figure 14.1).

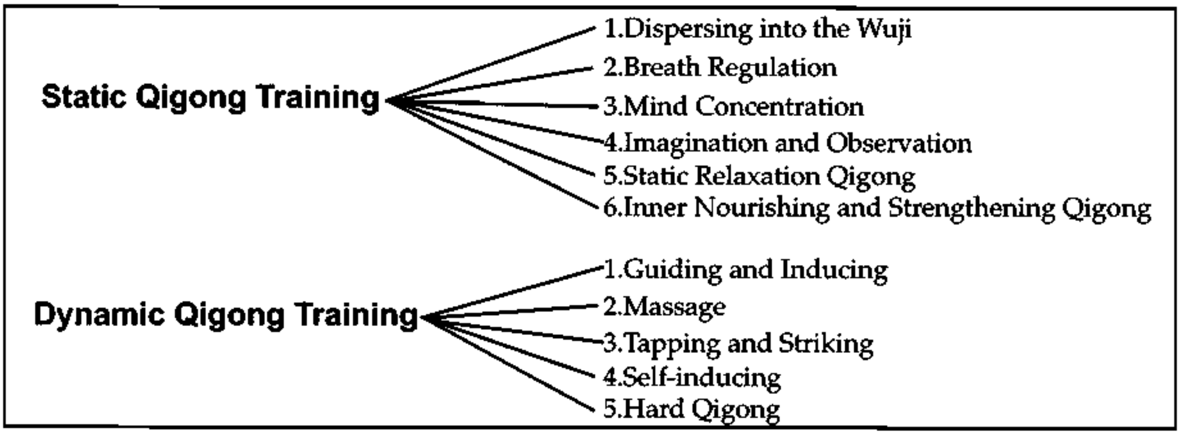
Static Qigong Training
1.Dispersing into the Wuji
2.Breath Regulation
3.Mind Concentration
4.Imagination and Observation
5.Static Relaxation Qigong
6.Inner Nourishing and Strengthening Qigong
Dynamic Qigong Training
1.Guiding and Inducing
2.Massage
3.Tapping and Striking
4.Self-inducing
5.Hard Qigong

Figure 14.1. Two Main Branches of Energetic Cultivation: Static Qigong and Dynamic Qigong

STATIC QIGONG TRAINING

This training involves bringing about and keeping a quiescent state of mind and breath. In fixed postures (lying, sitting, or standing) the practitioners can train to accumulate substantial Qi in their Lower Dantian, while circulating it through their bodies. Static Qigong training is further divided into six different schools of practice: Dispersing into the Wuji, Breath Regulation, Mind Concentration, Imagination and Observation, Static Relaxation Qigong, and Inner Nourishing and Strengthening Qigong.

1. Dispersing into the Wuji requires the practitioner to dissolve his or her energy into infinite space, or the Void, without focusing or directing their attention on any object from the external environment.
2. Breath Regulation requires moving Qi internally or externally while inhaling, exhaling, or swallowing the breath, etc.
3. Mind Concentration requires concentrating the mind on one fixed internal or external point.
4. Imagination and Observation require disciplining the imagination with vivid, colorful visualizations.
5. Static Relaxation Qigong requires the relaxation of both body and mind, keeping them in a constant state of quiescence.
6. Inner Nourishing and Strengthening Qigong require the use of multiple Static Qigong modalities (for example; Breath Regulation combined with Imagination and Observation).

DYNAMIC QIGONG TRAINING

This involves training the body, mind and Qi simultaneously, to promote the functional activities of Qi and to enhance its free flow throughout the body's channels. These exercises often require standing or walking, and affect muscular tension, weight distribution, Blood and Qi circulation, as well as the functional activities of the major organs. These exercises are differentiated by various kinds of body movements and can be divided into five different schools of training: Guiding and Inducing, Massage, Tapping and Striking, Self-inducing, and Hard Qigong.

1. Guiding and Inducing requires physical movements such as dancing, mimicking animal movements, and gymnastics.
2. Massage focuses on both the manipulation of specific points and channels, as well as on the knowledge of muscles, tendons, and bone structure.
3. Tapping and Striking focus on using patting, slapping, tapping, and striking techniques with the fingers, palm, fist, wooden stick, or material in a cloth bag.
4. Self-inducing training encompasses small to large physically strenuous Qigong movements.
5. Hard Qigong focuses on martial arts training in Iron Palm Training, Iron Body Training, etc.

The incorrect employment of dynamic or static principles of Qigong exercises and meditations can cause energetic deviations which are capable of either hindering, or harming, the individual's health. There are two purposes for Medical Qigong development: To Protect and Strengthen the Body's Health, and To Treat Diseases.

1. To protect and strengthen the body's health: Medical Qigong exercises are prescribed in accordance with the individual's physical constitution, state of health, emotional constitution, climate, and different seasons.
2. To treat diseases: Medical Qigong exercises are prescribed in accordance with the individual's symptoms, pathological condition, and degree of severity, complications, and the constitution of the patient.

PRINCIPLES OF YIN AND YANG IN DAO YIN TRAINING

Both static and dynamic Qigong methods supplement each other when combined equally. No matter which static or dynamic method is practiced, one of the four principles of Yin and Yang will always be adhered to: Quiescence in Motion, Action within Motion, Motion in Quiescence, and Stillness within Quiescence.

1. Quiescence in Motion is referred to as a Yin within Yang technique, and requires being quiet, calm, and still on the inside while being active on the outside.

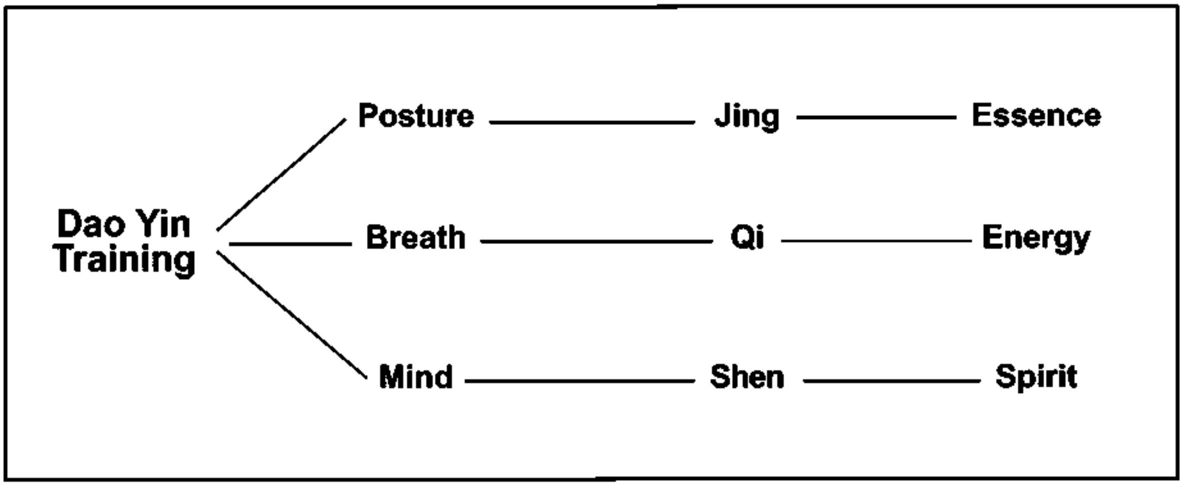

Figure 14.2. Dao Yin Chart

2. Action within Motion is referred to as a Yang within Yang technique, and requires being active both on the inside and on the outside.
3. Motion in Quiescence is referred to as a Yang within Yin technique, and requires being active on the inside while being quiet, calm, and still on the outside.
4. Stillness within Quiescence is referred to as a Yin within Yin technique, and requires being quiet, calm, and still both on the inside and on the outside.

THREE FOUNDATIONAL METHODS OF DAO YIN TRAINING

The diverse methods of practicing Medical Qigong exercises fall within three categories of Dao Yin training: Posture Dao Yin training, Respiratory Dao Yin training, and Mental Dao Yin training.

1. Posture Dao Yin consists of posture training for Medical Qigong therapy. It is divided into six specific types of training: walking, standing, sitting, kneeling, lying, and massaging.
2. Respiratory Dao Yin consists of breathing exercises for Qigong training, and therapy that includes various techniques of inhaling, exhaling, aspirating, blowing, and holding the breath.
3. Mental Dao Yin training requires that the thoughts be concentrated on one object, placing the cerebral cortex into a special inhibitory state. In various Medical Qigong circles in China this method is commonly referred to as "the internal keeping of the mental state."

The Dao Yins are taught in a postnatal energetic progression of Jing (posture) to Qi (respiration), to Shen (mind and emotion). In this way, the body regulates and supports the production of energy, which in turn regulates and supports the production of Spirit (Figure 14.2).

Despite the fact that many postures and movements may vary according to the specific styles of each Medical Qigong school, the primary benefit of each exercise will:

- improve and strengthen the patient's metabolism,
- smooth the Qi of the channels and collaterals,
- strengthen the body and mind,
- emotionally detoxify the internal viscera, and
- facilitate the increased flow of Righteous Qi (the Qi that fights the pathogenic invasion of Evil Qi) within the body.

CHAPTER 15
POSTURAL DAO YIN TRAINING

The goal of Postural Dao Yin training is to promote the regulation and circulation of Qi. Part of this training involves learning the techniques to guide the internal energy to flow using the movements and changes in the electromagnetic fields of the hands and body.

Success in Medical Qigong training is directly related to the correctness of the body's alignment. Correct posture regulates the Heart, and results in a calm mind; with a relaxed mind it is easy to lead the flow of Qi downward, to affect the Blood circulation and pulse. When the body is incorrectly aligned, it disturbs the smooth flow of Qi; energy then becomes obstructed or deviates, resulting in a restless Mind (both the thinking mind and the whole body consciousness).

The proper sequence of relaxation and energy circulation within the body is as follows:

1. the muscles are first to relax,
2. followed by the tendons and ligaments,
3. next, the nerves relax, and
4. finally the bones.

Once the muscles, tendons, and ligaments relax, the Qi can circulate freely through the channels and collaterals. The skin and muscles are generally the first level of awareness cultivation. The sensations of Qi and Blood circulation are usually experienced by the practitioner as heat, tingling, fullness, and expansion. The Chinese call these sensations the Eight Energetic Touches, which are known as the first of three transformations of energy (also known as the Subtle Wonders, see Chapter 29).

The next level of awareness involves the deeper flows of Qi within the nerves; this feels like an electric vibration, or an energetic current traveling through the body. Sometimes, due to an energetic surge released from within the body's organ system or channels, an electric shock is felt through the extremities or through the entire body.

The final level of Qi circulation awareness involves the Qi circulation within the bones and the Marrow. In the beginning stages it sometimes feels cold and arthritic in the joints (especially within the extremities). These sensations will later develop into a vibrational resonance which is felt deep within the bones.

The main point of Postural Dao Yin is to seek a state of physical and energetic balance. Postural Dao Yin is also divided into two distinct methods of treatment: Static Posture and Dynamic Posture training.

STATIC POSTURE DAO YIN TRAINING

This method includes quiet movement while in the position of lying, sitting, standing, or walking, and is used to accumulate and strengthen the practitioner's internal Qi. Before performing the static posture forms for Medical Qigong therapy, it is important to note that static posture training is divided into three basic modalities: Choosing the Proper Beginning Posture, the Inhaling and Exhaling Methods, and Opening and Closing the Lower Dantian.

1. The purpose of the choosing the proper beginning posture is to calm the Heart and relax the mind, and focus on a specific goal for training the body's life-force energy.
2. When the energy begins to flow freely, the practitioner focuses his or her attention on each inhalation and exhalation for the specific goal of tonification, purgation or regulating the life-force energy.
3. The purpose of the "opening and closing" of the Lower Dantian ending postures is to bring the body's energy back down into the Lower Dantian and to root the Qi. Rooting the Qi in the Lower Dantian after Qigong practice

avoids creating Excess conditions within the body's internal organs.

CHOOSING THE PROPER POSTURE

Medical Qigong meditations and prescriptions are not limited to sitting practice alone. Each of the different lying, sitting, standing, and walking postures has its own specific form of meditation.

The lying, sitting, and standing postures create the foundational base for Static Posture Dao Yin training. Each static posture can induce the body's energy to descend quickly into the Lower Dantian. The selection of the appropriate technique and body position is determined by each practitioner's body type (structural formation) and the condition of the individual's physical health. The length of time that a practitioner focuses on the Medical Qigong prescription is also determined by the individual's specific disease as shown in the following examples.

- Patients with digestive problems, gastroptosis, or prolapse of the internal organs should begin their training while lying supine (on their back), with their knees pointing upward, their buttocks elevated 4 inches off the ground, and their feet flat on the floor.
- Patients with gastrointestinal diseases should begin their training lying flat on their backs. As the condition improves, they should combine lying and sitting meditations to promote the digestion and absorption of food.
- Patients with Heart disease, hypertension, or nervous disorders should begin training from a sitting position. Later, they should combine sitting and standing meditations in order to direct their Qi and Blood downward.
- Patients with Lung diseases, asthma, chronic bronchitis, emphysema, or cardiopulmonary function disorders should begin their training from a semi-reclining position (usually a 45 degree angle) to help respiration flow smoothly and to alleviate stress on the Lungs.

  In the hospital, each pillow that the patient uses for elevation represents the degree of strain that is placed on the Lungs. The patient is said to have strong Lungs, for example, if he or she is reclining and breathing comfortably on one pillow. If however, it takes three pillows before the patient is reclining comfortably, the Lungs are considered weak and congested (e.g., in cases of dyspnea).
- Patients with severe physical weakness (e.g. chronic fatigue syndrome and fibromyalgia) should begin their training from a lying or sitting position. Later, they can combine sitting and standing meditations as their constitution improves.

PRIMARY AND SECONDARY POSTURES

Patients are taught a primary and a secondary posture to prevent fatigue, and increase the healing effect of the Medical Qigong prescriptions. A primary standing posture is performed, for example, in the morning when the patient is strongest, and the energy is cleanest. A secondary sitting posture is performed in the afternoon, or evening when the patient is tired, with a focus on quiet regulation.

After a patient has learned a particular lying, sitting, or standing meditation, he or she can be taught the secondary posture without the risk of depleting his or her Qi.

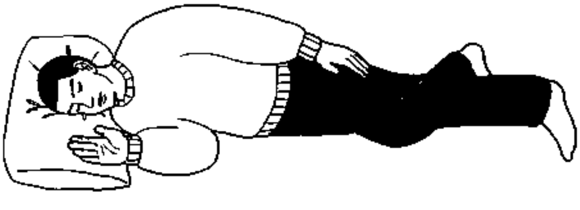

Figure 15.1. Sleeping Dog Posture

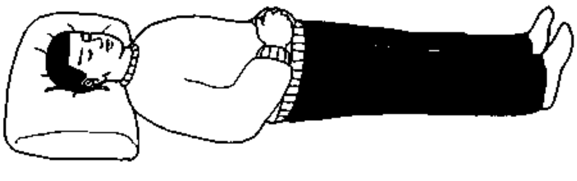

Figure 15.2. Supine Posture (Hibernation Posture)

LYING POSTURES

The Chinese saying, "The body should lie like a bow," is a perfect description of the curled position of the patient's body, while in the Sleeping Dog posture (Figure 15.1). This particular position is beneficial for aiding digestion and the absorption of food.

In prescribing the Sleeping Dog posture, the patient should lie on their right side with the left knee bent and the right leg extended straight. To avoid compressing the Heart, do not allow the patient to lie on their left side. Keep the patient's head raised slightly to allow the Blood to flow downward into the torso. Have the patient bend the right arm at the elbow.

When the patient is chronically ill, however, an exception is made. In this case it is important that the patient lie on the "healthy side" and not the side of the diseased organ (e.g., in cases of Liver disease, patients might lie on their left side). Remember to always adjust the patients' postures as needed and encourage them to practice "natural breathing."

Another popular posture is called the Hibernation posture. By lying supine in the Hibernation posture, the patient will regulate the functional activities of the central nervous system (Figure 15.2). Lying supine is also suitable for patients with serious diseases who are debilitated (this is also a sleep inducing method of training).

SITTING POSTURES

The Chinese saying, "The body should sit like a bell," describes the secure (or rooted) position of the patients' body, especially the lower legs and torso. Sitting is the most commonly used position for Medical Qigong therapy. This position helps the patients to relax and focus their mind's intent.

There are several sitting postures used in Medical Qigong therapy, and are prescribed according to the patient's flexibility: sitting upright on a chair (Figure 15.3), a cross-legged posture called the Full Lotus, another called Half Lotus, or with the legs loosely crossed (Figure 15.4-6). Make sure the patients sit on a raised cushion in order to keep their buttocks slightly higher than their knees.

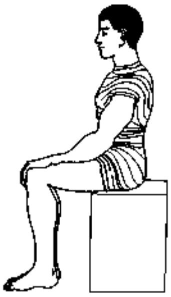

Figure 15.3. Upright Sitting Posture

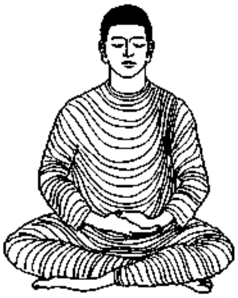

Figure 15.4. Cross-Legged Zi Wu Posture

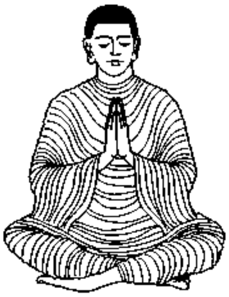

Figure 15.5. Cross-Legged Buddhist Greeting Posture

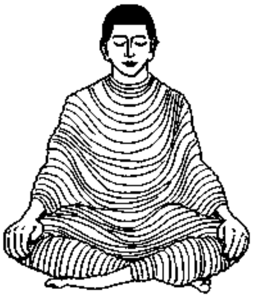

Figure 15.6. Cross-Legged Buddha's Warrior Attendant Posture

One of my Medical Qigong instructors in China suggested elevating the hips at least four inches off the floor when meditating, to avoid undue stress on the lower back, spine, and leg circulation. This elevation of the hips forms a triangle between both knees and the Lower Dantian, establishing a firmer connection to the Earth energy. This allows the Earth Qi to easily flow from the patients' knees and coccyx into their body.

When sitting in a chair meditating, the feet should touch the ground firmly, instead of being suspended in midair. This allows the Earth Qi to flow easily from the patients' feet into their body. Otherwise patients may develop swollen feet and lumbago due to Blood stagnation, or even hallucinations.

Keep the patients' spine straight by aligning the tip of the nose on a vertical line with the navel, to release pressure and tension on the cervical nerves.

STANDING POSTURES

The Chinese saying, "The body must stand like a pine tree," describes a relaxed position of the body's structure, standing with the bones stacked on top of each other, from the ground up. Standing meditations are particularly tonifying in nature, and are absolutely essential for accumulating Qi. The advantage of the standing position is that the flow of energy can be easily regulated and promoted. This posture helps to normalize blood pressure, and calm the Heart. Standing meditations also produce a unique psychophysiological change within the patients, allowing them to shed energetic obstructions, and to create a healthier body, mind, and spiritual awareness. Standing meditations are also essential for the Qigong doctors, as it facilitates the accumulation of Qi that is necessary for the emitting of energy.

Standing meditations, however, are not suitable for patients who are extremely ill or in poor health. Therefore, exercises in the standing postures should only be prescribed for minimal durations during convalescence, to build up the patients' constitution, promote health and to enhance their strength slowly.

The body should be held in an upright posi-

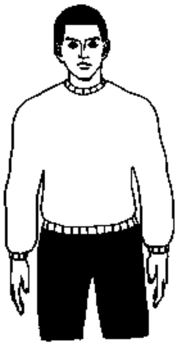

Figure 15.7. Wuji Posture

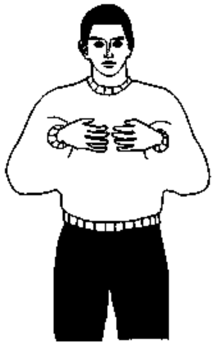

Figure 15.8. Immortal Post Posture

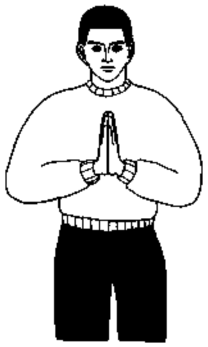

Figure 15.9. Buddhist Greeting Posture

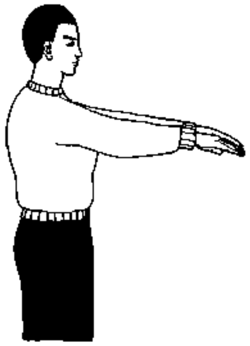

Figure 15.10. Extending Palm Posture

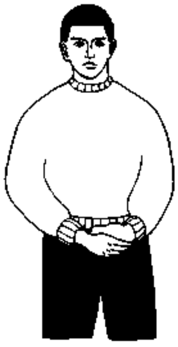

Figure 15.11. Embracing Dantian Posture

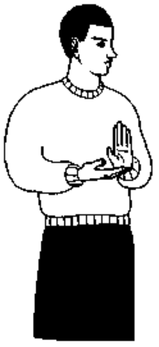

Figure 15.12. Square-Round Posture

tion, with three points connected in a straight line: Baihui GV-20 point at the top of the head, Huiyin CV-1 point at the bottom of the perineum, and the Yongquan K-1 point at the bottom of the feet. The patient's posture should be kept relaxed, still, solid, and rooted like a tree. The patient should also imagine their feet growing roots deep into the ground, to absorb the Earth's energy from the soil.

Standing exercises are usually accompanied by a variety of hand positions (Figures 15.7–15.12). The most popular are (1) Wuji: hands hang down by sides, (2) Immortal Post: hands embrace an invisible ball, (3) Buddhist Greeting: both hands press together at the solar plexus level, (4) Extending Palm: both hands press downward, (5) Embracing the Dantian: folding the hands and resting them on lower Dantian, and (6) Square-Round: left palm's heel rests inside the right palm's center.

In each posture, whether lying, sitting, or standing, the function and flow of body energy shifts and changes in relationship to, and as a manifestation of, the patient's mental belief structure.

If, for example, the patient while standing, shifts his or her weight and mental focus onto the balls of the feet, the energy will naturally flow through the front of the body via the front Yin channels. If the patient shifts his or her weight and mental focus onto the back on the heels, the energy will naturally flow through the back of his or her body via the back Yang channels. The center of the feet connect the patient's flow of energy to the Taiji Pole and can be used either to transfer the energy forward or backwards.

Generally, before beginning standing meditations, patients are encouraged to relax, sink, and shift their body weight back and forth between their heels and their toes. This practice massages their internal organs and disperses any blockages by stimulating the corresponding organ areas at the bottom of the feet as described in the therapeutic use of the Lower Hexagram Yaos (see Chapter 3).

Patients should inhale while shifting their weight backwards and exhale while shifting forwards. Their attention should be placed on their internal organs and the body's surrounding field of energy. The patients' awareness is also focused on the body moving back and forth, like the rhythm of an ocean tide.

Women are generally encouraged to practice meditation in the standing position (especially around the time of their menstruation), as the sitting position tends to decrease energy circulation, and may cause energy blockages in their lower abdominal area.

THE EIGHTEEN RULES OF PROPER MEDICAL QIGONG STRUCTURE

The most important basic techniques of Medical Qigong training are guided by eighteen rules of proper form and structure. The main point of Postural Dao Yin training is to relax and seek quiescence while in the various postures. Tension in any area of the body restricts the whole structural system, since the body seeks to balance its structure naturally by shifting its energy and weight. The general function of the muscles is to guide the flow of energy through the channels.

The following is a list of the Eighteen Rules for proper standing postures, these rules apply to every school of energy cultivation:

1. Stand with the feet flat,
2. Bend the knees,
3. Relax the hips,
4. Round the perineal area,
5. Close the anal sphincter,
6. Pull in the Stomach,
7. Relax the waist,
8. Tuck the chest in,
9. Stretch the back,
10. Relax the shoulders,
11. Sink the elbows,
12. Hollow the armpits,
13. Relax the wrists,
14. Suspend the head,
15. Tuck the chin,
16. Close the eyes for inner vision,
17. Close the mouth and turn your hearing inwards, and
18. Touch the tongue to the upper palate,

Each of these eighteen rules is described in detail as follows.

RULE 1—STAND WITH THE FEET FLAT

Stand with the feet flat, parallel to the outside of the shoulders. Weight should be distributed evenly on both feet about an inch and a half inward (anterior) from the heels, and perpendicular to the tibia. Focus on dropping the weight down behind the legs, not down the front. This action takes the pressure away from the knees and places it in the center of the feet (Figure 15.13).

Stretch the feet and let the toes softly grasp the ground to keep the body firmly rooted as you tuck the sacrum under (Figure 15.14). Relaxing is important. Rigid, contracted feet disrupt the flow of energy from the Earth into the body. The Kidney energy flowing within the bone can be accessed by shifting the weight onto the heels. The Heart energy can be accessed by shifting the weight onto the balls of the feet. The weight in the center of the feet accesses the Taiji Pole.

While training, the feet may vibrate or feel hot, as if on fire. This is a normal reaction to correct postural training and is beneficial because it dissolves the calcium deposits stored within the extremities of the feet. If, however, the Fire in the feet is too intense (due to Excess Liver Fire), hit the heels on the ground and rub the lumbar spine, using the mind's intention to send more of the Kidney Yin (Water) down to the feet to cool the Fire.

RULE 2—BEND THE KNEES

The knees should be slightly bent and facing the same direction as the feet. The knee caps should be directly aligned and extended no further than the tips of the middle toes. Do not allow the knees to twist inward.

To secure the structure, imagine holding a large beach ball between the knees. Many patients develop knee problems because they extend their knees too far forward or allow their knees to twist to the side, rather than aligning them with the toes. If pressure is felt on the sides or front of the knees, check the positioning. Knees that are unable to point in the direction of the toes are most often the result of tight hip joints (Figure 15.15). If the energy in the legs is tightly constricted around the knee area, Qi flow will be severely limited. A relaxed knee-joint, on the other hand, will increase

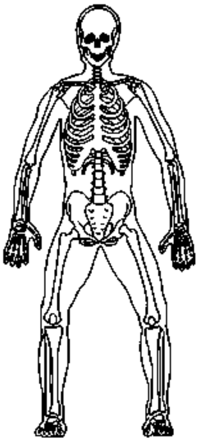

Figure 15.13. Stack the bones on top of each other from the bottom of the feet to the skull.

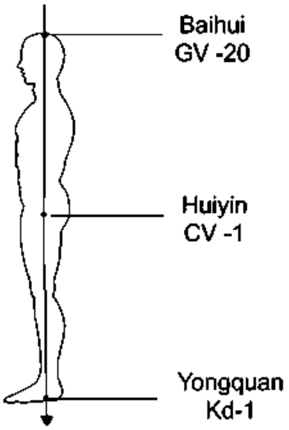

Figure 15.14. The three points (Baihui, Huiyin, and Yong quan) can be connected by a straight line.

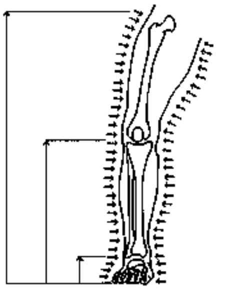

Figure 15.15. Keep the center of the body's weight aligned to the center of the knees, to relax the connective tissue surrounding the knees.

the Qi and Blood flowing through the legs, making it easier to relax the hip and to round the perineal area.

Relaxing the knees also permits free flow of energy through the three descending Yang channels on the outside of the leg, and the three ascending Yin channels on the inside of the leg (Figure 15.16). The techniques for relaxing the knees will vary according to the emotional and structural pattern of each individual, as emotions of fear, etc., have a tendency to gather and pool behind the knees.

RULE 3—RELAX THE HIPS

Drop the buttocks slightly while sinking the Qi from the upper torso into the Lower Dantian, and relax the hips. Keep the hips facing forward. The buttocks should be gently tucked under the body to facilitate straightening of the spine and the keeping the Qi in the Lower Dantian.

Imagine a weight of a thousand pounds hanging from the coccyx (Figure 15.17). As the weight pulls down, tilt the sacrum under the body. When the hip-joints are relaxed, the lower limbs will be able to move freely.

RULE 4—ROUND THE PERINEAL AREA

There are three parts to rounding the perineal area. First, set the knees apart, turning them both outward. This allows the groin and hips to shift to an anterior tilt. Second, bring the knees back together and turn them inward, while also relaxing the hips. This allows the groin and hips to shift to a posterior tilt. Third, suspend the perineal area to lift the perineum and the anus slightly. This allows the groin and hips to come to a center balance and to sink. Rounding the perineal area keeps the perineum free from pressure, yet sealed. This ensures better functional activities of the Qi, and allows the "lower door" (the anus) to remain closed to avoid leakage of vital energy.

RULE 5—CLOSE THE ANAL SPHINCTER

The anus is called the Lower Bridge and it is where the Yang and Yin channels meet and combine. The anus is considered the door of the Seven Corporeal Souls (Po). Although the anal sphincter is held closed, it is important that the perineum remain relaxed. There are three stages in controlling the anal sphincter:

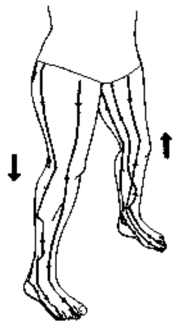

Figure 15.16. Relaxing the knees also permits the free flow of energy through the ascending Yin channels and descending Yang channels of the legs.

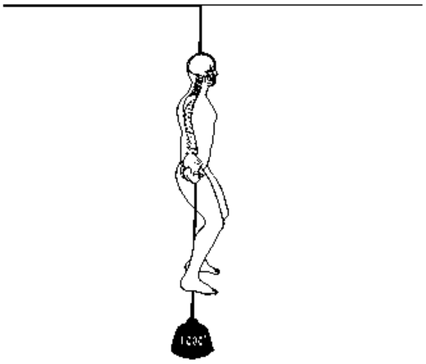

Figure 15.17. Imagine a 1000-pound weight hanging from the coccyx and a heavenly cord suspending the head.

1. In the first stage, the anus is contracted with slight muscle control, and the energy is directed into different energy centers along the Microcosmic Orbit. When first beginning, inhale and expand the abdomen while imagining that the anus is connected to the navel and is pulling up towards it. On the exhalation, contract the abdomen and draw the energy from the navel back down towards the anus. Next, inhale and create a connection between the anus and the Mingmen. On the exhalation, draw the energy from the Mingmen back down towards the anus. It is important to synchronize the drawing up of the anus towards the navel and Mingmen area with the expansion and contraction of inhalation and exhalation. This will facilitate a

sacral-cranial rhythm throughout the body.

2. In the next stage of training, the constriction of the anus is used to lead the energy up the Governing Vessel in the Microcosmic Orbit (Figure 15.18). Upon inhalation, draw the mind's intention from the navel towards the anus and then down to the tip of the spine (the coccyx), allowing the Qi to flow up the spinal column towards the Mingmen, past the Shendao to the base of the skull. Upon exhalation, roll the eyes back into the head and use the intention to meet the energy at the base of the skull. Next, lead the energy over the Baihui at the top of the head and down the Conception Vessel and back down to the naval.
3. At the third stage, imagine twisting the energy of the anal sphincter, spiraling the Qi up the center of the body, to connect with the Baihui area at the top of the head. Close the anal muscle gently, inhale and imagine the energy vibrating and spiraling in a counterclockwise direction up the body's Taiji Pole. This connects the Lower, Middle, and Upper Dantians through the Taiji Pole. Upon exhalation, lead the vibrating Qi down the Taiji Pole in a clockwise spiral ending within the Lower Dantian area. Continue this meditation until the entire body pulses with the rhythm of the Taiji Pole.

The anus is divided into five regions: Front, Middle, Back, Left, and Right. By contracting the anus in different parts, you can bring more Qi to the various organs and glands.

1. The Front region opens the gate to the sexual organs, as well as to the Conception Vessel when it is contracted.
   - For Men, this will affect the prostate gland, Bladder, Small Intestine, thymus gland, thyroid gland, parathyroid gland, mouth, nose and front part of the brain (Figure 15.19).
   - For Women, this will affect the vagina, uterus, vena cava, aorta, Stomach, thymus gland, thyroid gland, parathyroid gland, mouth, nose, pituitary gland, and front part of the brain (Figure 15.20).
2. By contracting and pulling the middle region of the anus, more Qi is drawn into the Thrusting Vessel and Taiji Pole, as well as to the genitals, aorta, vena cava, Stomach, Heart, thyroid gland, parathyroid gland, tongue, pituitary gland, pineal gland, and Baihui area (Figure 15.21).
3. By contracting the back region of the anus the Qi is connected to the Governing Vessel, as well as the sacrum, lumbar, thoracic and cervical vertebrae, and the cerebellum (Figure 15.22).
4. By contracting the left side of the anus the Qi connects to the left side of the body.
   - For Men, this affects the left testicle, Large Intestine, left Kidney, left adrenal gland, Spleen, the left side of the Heart, left Lung, left ear, left eye, and left hemisphere of the brain (Figure 15.23).
   - For Women, this affects both ovaries, both Kidneys and both adrenal glands, Spleen, Liver, Heart, both Lungs, thyroid gland, parathyroid gland, and both hemispheres of the brain (Figure 15.24).
5. The Right Side. By contracting the right side of the anus the Qi connects to the right side of the body.
   - For Men, this affects the right testicle, Large Intestine, right Kidney, right adrenal gland, Liver, Gall Bladder, right side of the Heart, right Lung, right ear, right eye, and the right hemisphere of the brain (Figure 15.25).

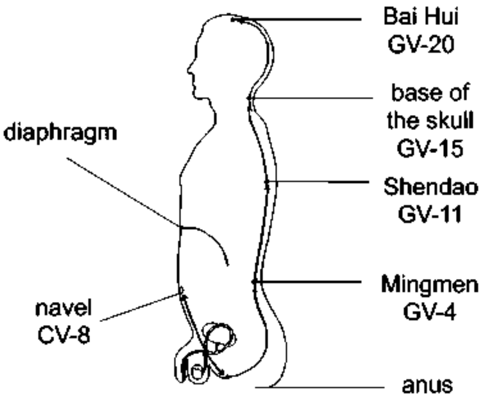

Figure 15.18. The anus is used to draw the Qi from the navel into the Lower Dantian and lead the Qi up the Governing Vessel.

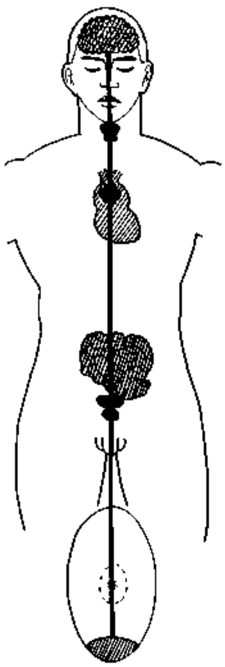

Figure 15.19. The Front of the Anus

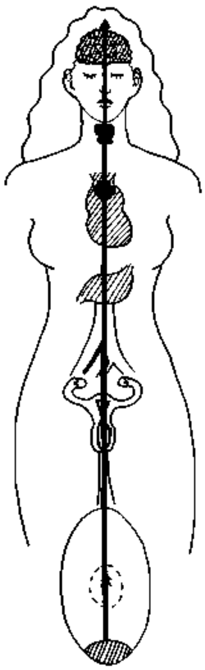

Figure 15.20. The Front of the Anus

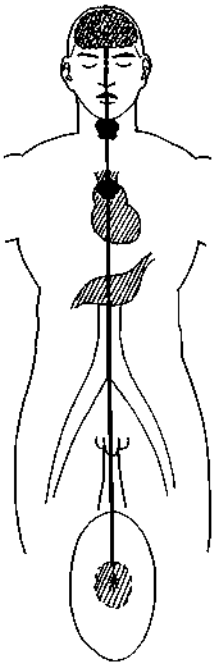

Figure 15.21. The Middle of the Anus

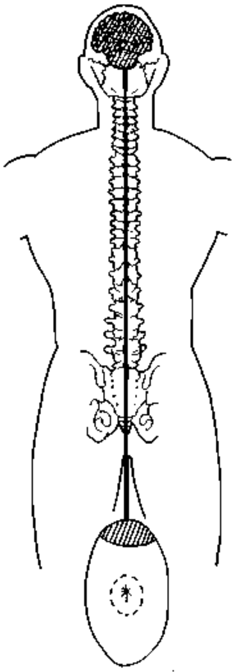

Figure 15.22. The Back of the Anus

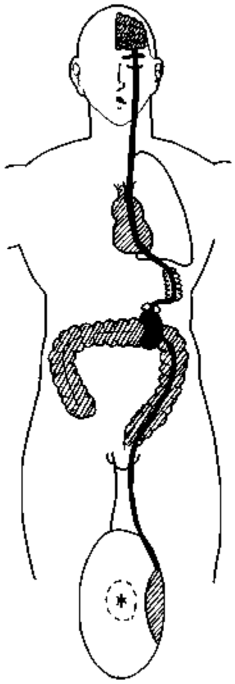

Figure 15.23. The Left Side of the Anus

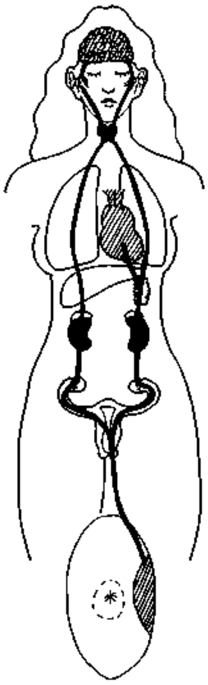

Figure 15.24. The Left Side of the Anus

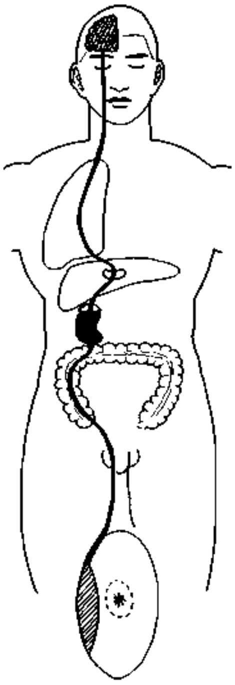

Figure 15.25. The Right Side of the Anus

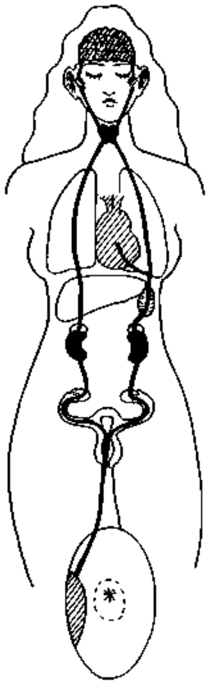

Figure 15.26. The Right Side of the Anus

- For Women, this affects both ovaries, both Kidneys and both adrenal glands, Spleen, Liver, Heart, both Lungs, thyroid gland, parathyroid gland, and both hemispheres of the brain (Figure 15.26).

RULE 6—PULL IN THE STOMACH

Pull the stomach in above the pubic bone, and avoid tensing of the abdominal muscles. This helps to restrain the Yuan Qi, thereby increasing the internal pressure and promoting the flow of internal Qi throughout the whole body.

RULE 7—RELAX THE WAIST

If the waist is not relaxed the Qi can not sink back to the Lower Dantian. Relax the waist by relaxing the hips and bending the knees. When the waist is relaxed, the spine will stand upright, yet remain relaxed.

The waist and the abdomen are two important areas in training and guiding Qi. The abdomen is considered to be the furnace for refining Qi. The waist is the residence of the Kidneys and Mingmen Fire and is an important junction for Qi and Blood circulation.

The waist is considered the master of the body and must be kept sunken and relaxed (Figure 15.27). As the buttocks tuck under and the waist sinks, small pockets should form on the inner thighs where the pelvic bones meet the femur bones of the legs. Most people usually keep the pelvis area tight and immobile, as it is associated with shock, fear, and sexual trauma (when this area becomes loosened, such emotions are often released).

The specific actions to relax the waist are as follows: lift both shoulders in order to stretch the back, then release them downward immediately as you exhale, relaxing the waist. To achieve genuine waist relaxation, you must also relax and tuck the sacrum, this action opens the sacrum. When the sacrum is opened, the Lower Dantian will expand and the storehouse of Yuan Qi will be enlarged. Be sure that the waist and hips face forward in the same direction as the knees and feet.

RULE 8—TUCK THE CHEST IN

The chest should relax inward so that air flows freely and the Qi sinks down into the Lower Dantian. Fill the abdomen with air while relaxing the chest inward. The back should be open and expanded. Relax the scapulae and allow them to sink to their lowest point, this will help the shoulders to sink. Rounding the scapulae (pushing to the sides and forward) will help open and expand the back.

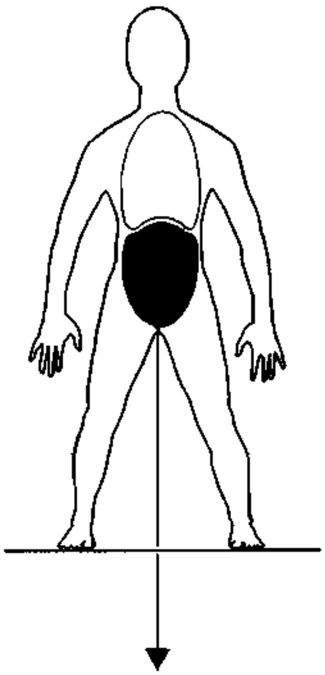

Figure 15.27. Relax and sink the waist, allowing the hips to drop.

Do not thrust the shoulders inward, nor intentionally close the thoracic cavity. The intentional tucking in of the vertebral area requires the outward opening of the ribs and expansion of the intercostal muscles. The action of opening the chest outward, then tucking the chest inward, helps to relax the thoracic cavity and avoids tension in the thoracic muscles.

RULE 9—STRETCH THE BACK

The key point to stretching the back is to hang the shoulders and "open the scapulae." When the two shoulders drop evenly, the lumbar vertebrae straighten out. The function of tucking the chest and stretching the back calms the Heart and Lungs.

The Heart controls mental and emotional activities, and also directs Blood circulation. When the thoracic cavity is spacious and the back is upright, the Heart and mind will be calm, and the Lungs will be clear and free from Heat. Tucking the chest and stretching the back allows the chest and back to relax, permitting the Conception and Governing Vessels to communicate with each other and improve energy movement.

Imagine a weight hanging from the coccyx and a rope pulling up on the crown point of the head. Feel the elongation of the spine with each inhalation and exhalation. Remain relaxed (Figure 15.28), do not hold yourself unnaturally erect and avoid contracting the abdomen to straighten the lower back. The abdomen must remain relaxed. When sinking the body's weight, try to feel the spine press into the heels of the feet.

RULE 10—RELAX THE SHOULDERS

If you relax the shoulder joints and let them hang naturally, the shoulders and neck will relax. If you elevate your shoulders when you are standing or sitting, the energy will not sink and you will begin to feel tired, the shoulders will be stressed, and the flow of Blood and Qi will become blocked.

When the shoulders are encouraged to relax and are allowed to drop and roll forward, this action facilitates the hollowing of the chest and the expanding and opening of the back, allowing the energy to flow into the arms (Figure 15.29). Try to feel your arms link up with the scapulae as the clavicles press downward. Your shoulders should remain very loose and free.

RULE 11—SINK THE ELBOWS

When practicing the Immortal Post posture (refer back to Figure 15.8), and the hands are positioned by the Heart, it is important to bend the elbows slightly at the elbow joint, holding them lower than the hands.

While in a Wuji posture, imagine that a weight is hanging from the elbows, or a sensation of compression will be felt at the shoulder and elbow joints. The elbows should be slightly bent and relaxed, remaining structurally correct and not weak. This will help to roll the shoulders forward

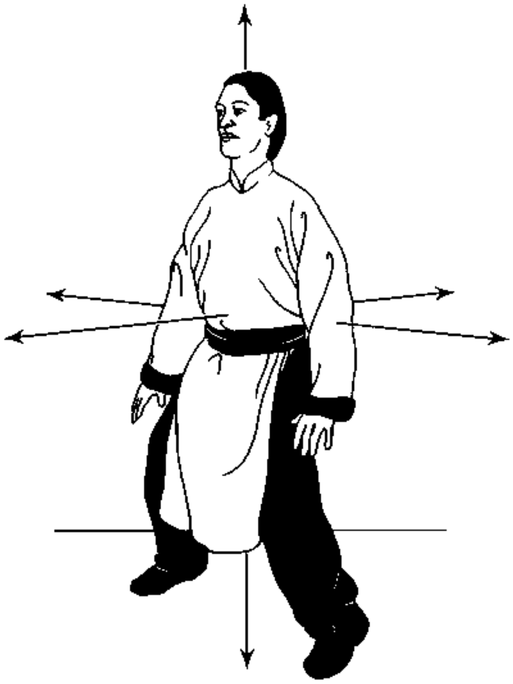

Figure 15.28. By imagining a rope pulling on the head and coccyx, the spine is allowed to naturally stretch, expanding the body's Qi in all directions.

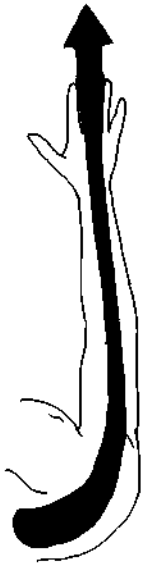

Figure 15.29. When the shoulder (here viewed from above) relaxes, the Qi flow into the arm increases.

and open up the back. If the elbows are not bent, there will be decreased energy flow (Figure 15.30).

RULE 12—HOLLOW THE ARMPITS

Hollow the armpits as if each one is holding an egg, this allows the Qi and Blood to flow freely in the upper limbs. To relax the shoulders, move the tips of the elbows outward to the left and right. Avoid sloping or squaring the shoulders.

When practicing the Immortal Post Posture (refer back to Figure 15.8), the arms must embrace inward as if holding a large ball or barrel. Do not push the elbows out so far that you loose the curve in your arm, this is of extreme importance. Your arms should not bend inward more than 45 degrees. The energy flowing in your arms is like water running through a hose. If you bend it too sharply, much of the supply will be cut off.

RULE 13—RELAX THE WRISTS

This includes hollowing the palm, and the comfortable curling of the fingers. Relaxation of the wrist allows the Shu-Stream points in the wrist to communicate with the fingers (see the Five Antique Shu Points in Chapter 8). Hollowing the palm and curling the fingers allows for the natural stretching and curving of the hand, which prevents the Qi from diffusing. It is important to avoid being rigid and purposefully bending the fingers, as all hand postures effects the brain and the body's Qi flow.

Relaxing the wrists helps to regulate the three Yang and three Yin channels of the hands so that Qi and Blood can flow freely. The wrist must form a line with the shoulders. Do not bend the wrists, but hold them straight so that the middle finger forms a line with the forearm, allowing the hands and fingers to hang relaxed but not limp.

RULE 14—SUSPEND THE HEAD

The classics say that "the head is held aloft not by force, but as if it were placed there by a ruling power." Hold the head as if it were suspended, keeping the Baihui point pointed to Heaven (Figure 15.31). This produces space in the body by elongating the spine. When suspended, the head must be kept upright, positioned in the middle of the chest and shoulders. This helps to set the whole body in the upright position, and

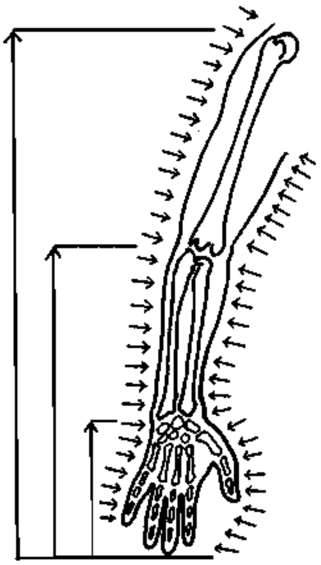

Figure 15.30. Allow the bones in the arms to feel heavy, naturally stretching the tissues and increasing Qi flow to the extremities.

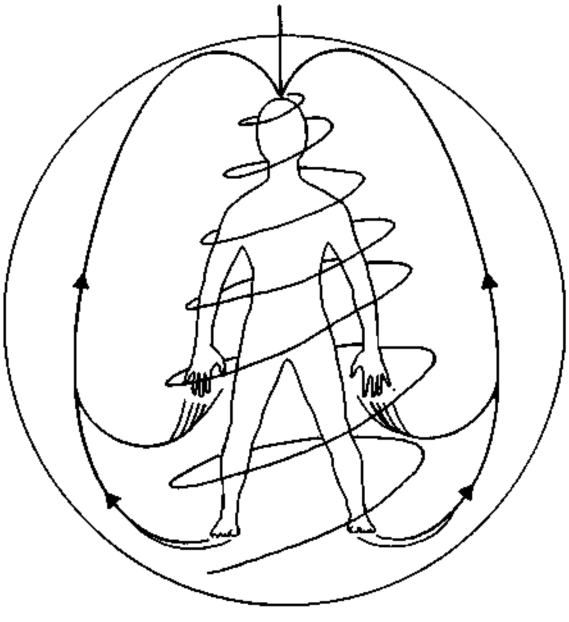

Figure 15.31. Suspending the head allows the body to become sensitive to the surrounding environment and external fields of Qi.

also directs the Qi and Blood to flow upward to nourish the brain and the mind. The Shen will be reinforced while the Jing and the Qi will be invigorated (it is the Shen which controls all the physical activities). An effective visualization is to imagine your head lifting high into space, your waist positioned in the clouds, and your feet sunk deep into the Earth.

RULE 15—TUCK THE CHIN

The head cannot be properly suspended if the chin is not tucked in. Only then can the Baihui point on the top of the head be correctly positioned and the nasal respiration kept free.

Tuck the chin and slightly lean forward to stretch and straighten the spine. You should feel the stretch all the way down into the coccyx. Next, slowly shift the weight backwards, stacking each vertebra on top of the next, starting from the bottom of the coccyx and sacrum and moving upwards.

RULE 16—CLOSE THE EYES FOR INNER VISION

The Chinese have a saying: "close the curtain." This refers to dropping the eyelids to facilitate inner-vision, and focusing on any specific areas where Qi is being trained or circulated. The eyes are of great importance in Medical Qigong exercises. It is said that the functional activities of Qi are determined by the eyes, and that the eyes are the messengers of the mind, and the mind is the residence of the Shen.

Completely closed eyes induce an absolute Yin state which is generally used in meditations for tonifying any Deficiencies. However, when used excessively (with all Yin and no Yang energy), the Qi can become too inactive.

Completely open eyes induce an absolute Yang state which is generally used in meditations for dispersing any Excesses. However, if overused, or if the patient has an Excess Yang condition, this type of meditation can make the practitioner too restless.

When the goal is to regulate the body's energy, slightly open the eyes in order to leave a small thread of light inside the eyes. This will keep the Yang (light) within the Yin (dark), so that the Qi will be more balanced. Because energy leaves the body through the eyes, after performing various Medical Qigong exercises, or prescriptions, allow your eyes to close halfway to continue the cultivation of life-force energy.

Tightly shutting the eyes causes tension in the upper and lower eyelids. Closing the eyelids gently, however, will restrain the vision, which helps to calm the Shen and the mind.

When meditating, if the eyes look to the distant horizon, they should not be focused but stare into infinity, this technique is also called the One Thousand Yard Stare, and it helps in regulating the venous pressure in the cranial cavity.

Strive for balance. Each eye has its own specific responsibility and energetic function.

- The right eye is generally used to project intention and extend emotion, energy, and spirit out from the body. The right eye relates to Water, the Kidneys, Qi, moon, Yin, Earth, Metal, and the Po.
- The left eye is generally used to receive, drawing in images, emotion, energy, and spirit into the body. The left eye relates to Fire, the Heart, Shen, sun, Yang, Heaven, Wood, and the Hun.
- If both eyes look upwards towards the Upper Dantian, the body's energy will draw up into the infinite space of the Wuji and merge with the divine.
- If both eyes look towards the tip of the nose, the Middle and Lower Dantians are linked together for transforming Yuan Qi.
- Crossing the eyes results in the union of the Wood and Metal, the elements of alchemic transition, for increasing the development of psychic energy. This intensification of psychic development is due to the union of the Hun and Po within the body, and is said to connect the meditator with his or her Yuan Shen.
- After meditation, store the Qi of the eyes behind the posterior orbital surface of the eyeballs themselves. This prevents the eyes from overheating and drying out.

RULE 17—CLOSE THE MOUTH AND TURN YOUR HEARING INWARD

In order to enhance the Yang during Medical Qigong exercises, close the lips with the teeth gently touching. Closing the lips will prevent the internal Qi from flowing out.

The Chinese use the expression: "Stopping the ears," which means to focus the hearing inward so as to be free from outside interference.

RULE 18—TOUCH THE TONGUE TO THE UPPER PALATE

When the tongue touches the upper palate, it forms what is called the Bird's Bridge. The connection of the tongue links both the Governing Vessel (which ends at the nose and upper palate) and the Conception Vessel (which ends in the mouth). As the tongue touches the upper palate, the body's field of Wei Qi expands. This can help to eliminate diseases and prolong life. The belief is that when someone has a disease, the Governing and Conception Vessels are to some extent disconnected.

Some Qigong doctors test their patients' progress by having them meditate with their tongue suspended in their mouth. Upon command, they connect the tongue to the upper palate. The suddenly increased expansion and strength of the patients' Wei Qi, observed by the doctor, indicates the degree of their progress.

1. The technique called "Gathering the Saliva," begins when the tongue touches the upper palate, causing the saliva secretion to be increased. Saliva contains many enzymes, which not only aid digestion, but also benefit the patient's physiology. In ancient China, this practice was called Gathering the Immortal's Water or the Juice of Jade to create the Immortal Pill.

   There is an exchange of Yin and Yang energy within the saliva, as Heaven Qi and Earth Qi are gathered within man's Fluid Essence in order to create the Immortal Pill.

   Before swallowing the saliva, it is important to tilt the head slightly forward (after inhaling). Swallow the saliva three times with three big gulps, using the breath as a cork to enclose the saliva. Send the Immortal Pill, i.e., the combined saliva and air down the Conception Vessel, bathing the Five Yin Organs, and ending at the Lower Dantian. Next, draw the energy up the Governing Vessel into the brain to replenish and nurture the Upper Dantian.
2. There are Five Elemental Tongue Positions. Each tongue position is related to one of the Five Yin Organs and can be used to direct the Qi to that specific organ. Sometimes, while practiscing Medical Qigong prescriptions, if too much Heat is created, the patient will be instructed to "change the tongue position" as well as meditation. These Elemental tongue positions are described as follows.
   a. The Earth (Spleen) position requires that the tongue be placed on the bottom of the jaw at the base of the lower palate, beneath the teeth and gum line (Figure 15.32).
   b. The Metal (Lungs) position requires that the tongue be placed between the maxilla and mandible bones, suspended, behind the teeth (Figure 15.33).
   c. The Fire (Heart) position requires that the tongue be placed on the hard palate, at the front of the upper palate, behind the teeth and gum line (Figure 15.34).
   4. The Wood (Liver) position requires that the tongue be placed on the hard palate, at the middle of the upper palate at the center of the roof of the mouth (Figure 15.35).
   e. The Water (Kidneys) position requires that the tongue be placed on the soft palate, at the back of the upper palate (Figure 15.36).

DYNAMIC POSTURE DAO YIN TRAINING

While Static Qigong exercises accumulate and strengthen the Qi internally, Dynamic Qigong exercises train the channels, collaterals, muscles, and bones externally. The more dynamic movement a patient makes, the more the energy will be converted and transformed inside the body. When dynamic movement is discontinued, some of the accumulated energy will begin to dissipate, while the remainder will flow through the channels and increase the patient's Qi circulation.

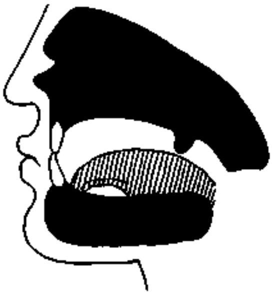

Figure 15.32. For the Earth position, which is related to the Spleen, the tongue is placed on the bottom of the jaw at the base of the lower palate, beneath the teeth and gum line.

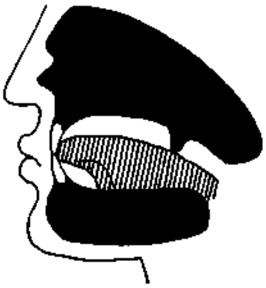

Figure 15.33. For the Metal position, which is related to the Lungs, the tongue is placed between the maxilla and mandible bones, suspended, behind the teeth.

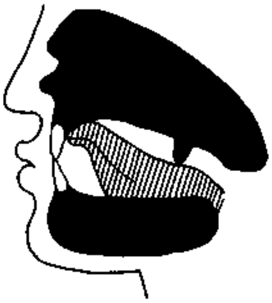

Figure 15.34. For the Fire position, which is related to the Heart, the tongue is placed on the front of the upper palate, behind the teeth and gum line.

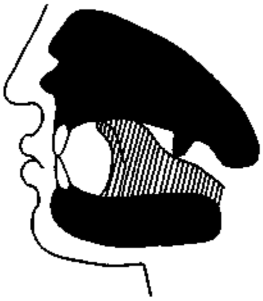

Figure 15.35. For the Wood position, which is related to the Liver, the tongue is placed on the middle of the upper palate at the center of the roof of the mouth.

The lymphatic system is pumped by normal muscular contraction. When a patient performs Dynamic Postural Movements, the lymphatic system is flushed, thus freeing the body of toxins.

Dynamic Qigong is divided into two methods of training: The Yun (Yin) Method and the Dong (Yang) Method.

1. The Yun Method is a more Yin approach, and employs slow, even, graceful movements. The focus of the mind's intention is placed on the internal organs, resulting in emotional relaxation.
2. The Dong Method is a Yang approach, and employs intense physical movements. Focus is placed on the movement of the body's ex-

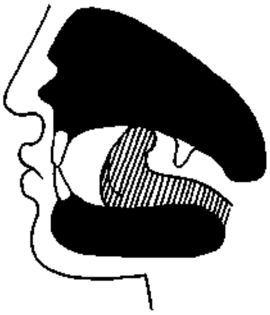

Figure 15.36. For the Water position, which is related to the Kidneys, the tongue is placed on the soft palate at the back of the upper palate.

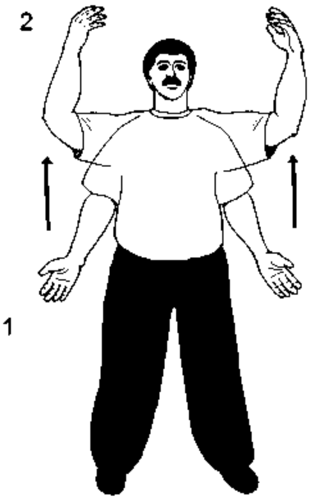

Figure 15.37. Hands Rise, Leading to the Ascending of the Qi.

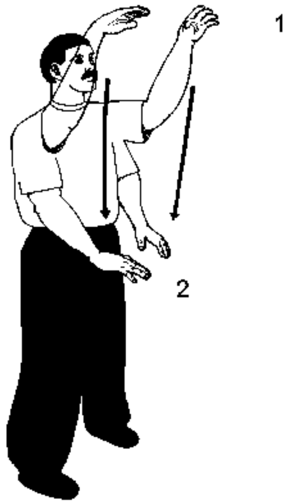

Figure 15.38. Hands Fall, Leading to the Descending of the Qi.

tremities, resulting in tissue regulation.

Both Yin and Yang (Yun and Dong) training methods must be balanced when prescribing proper Medical Qigong exercises. With the Yin method, the patient will experience external quietness with internal energetic movement. With the Yang method the patient will develop internal quietness with external, physically active movement.

USING THE ENERGY OF THE HANDS TO LEAD QI

Dynamic Postural training is performed by the extensive movements of the hands and torso. Because the palms have a stronger electromagnetic field than any other part of the body, the hands are used for leading and guiding the Qi. These movements are in harmony with the rotating, bending and stretching movements of the head, neck, torso, and limbs. Any change of the body's position or movement can change the direction or position of the electromagnetic fields of the hand. Likewise, any change or movement of the hands or palm positions, also changes the direction and flow of the electromagnetic field of the body. In Medical Qigong therapy the Dao Yin Hand Techniques have the following four functions:

- Hands Rise—Qi Ascends (Figure 15.37)
- Hands Fall—Qi Descends (Figure 15.38)
- Hands Open—Qi is Spread (Figure 15.39)
- Hands Close—Qi is Restrained (Figure 15.40)

The positions of these hand techniques can be further divided into three ranges: upper, middle, and lower. These ranges can relate to either the position of the body's Triple Burners, or the body's three Dantian areas, depending on the specific Medical Qigong prescription, and the doctor's intention for the exercise. The directions of these hand movements can further divide a posture into four front and four oblique types of Qi regulations. Moreover, depending on the arm movements, each posture can be divided into Yin (palms facing down) or Yang (palms facing up), as well as half Yin or half Yang regulations (Figure 15.41).

The actions of these methods supplement, regulate, smooth, and reinforce the vital Qi. Using these techniques will improve Blood and Qi circulation, as well as sharpen the sensory nerves. Since all hand postures relate to the brain and the function and flow of the body's energy, they also have a psychological and physiological affect on the nervous system.

When training, posture and Qi movement are

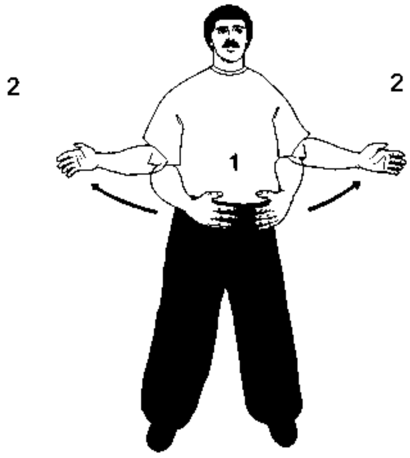

Figure 15.39. Hands Open, Leading to the Spreading of the Qi

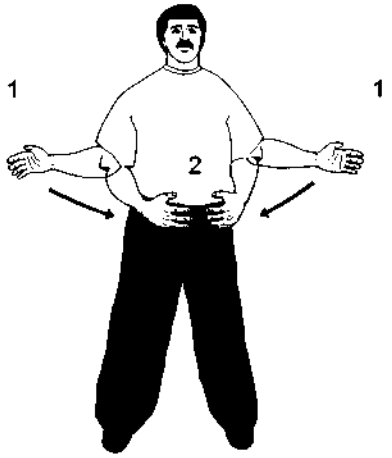

Figure 15.40. Hands Close, Leading to the Restraining of the Qi

sometimes in unity, and at other times proceed in opposite directions. Raising the arms, for example, causes the energy to rise; if the torso also rises, the rising power of the energy increases. This is an example of synergistic promotion of Qi regulation.

The following are three examples of the use of each separate hand action, and demonstrate how the body's reaction to energy can be used to treat patients with Medical Qigong therapy.

1. If you have hyperactive ascending Qi (e.g., hypertension) place your hands, palms facing downward, at the side of your body. Inhale quickly and move your hands quickly upward. Allow the middle fingers on each hand to touch each other, connecting the Pericardium Channels. Exhale, while slowly moving the hands downward along the front and side of the body. Purge and guide the Toxic Qi to descend the Liver and Gall Bladder Channels down the

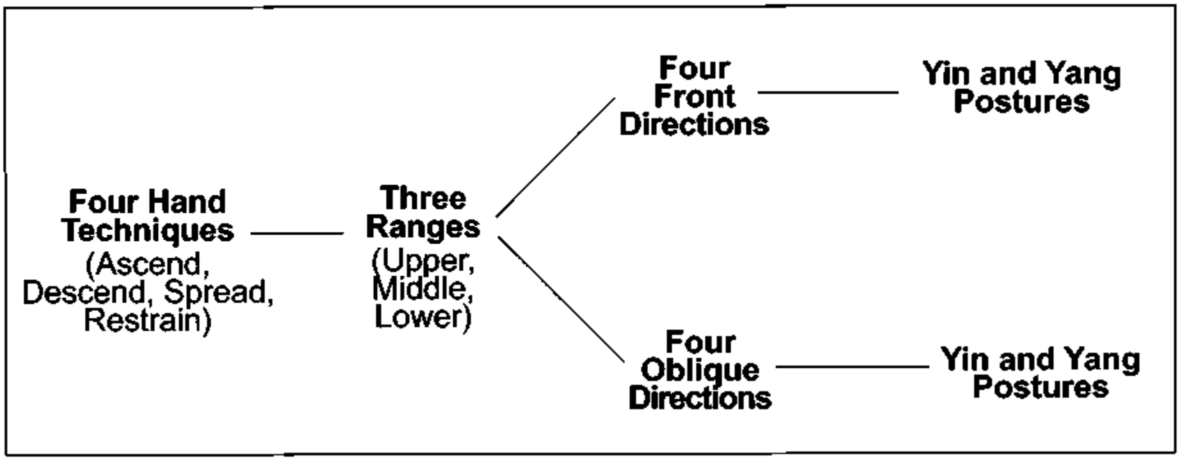
Four Hand Techniques
(Ascend, Descend, Spread, Restrain)
Three Ranges
(Upper, Middle, Lower)
Four Front Directions
Yin and Yang Postures
Four Oblique Directions
Yin and Yang Postures

Figure 15.41 Hand Position Techniques

torso to the hips (GB-30), and then to the outside of the legs (Figure 15.42).

2. If you have a Yin and Yang imbalance (e.g., chronic illness) place your palms facing your Lower Dantian and slowly raise them upwards to the chest. At the level of the Heart, turn the palms away from the body as you move your arms outward. Once the hands are extended away from the body begin moving them downward. Regulate the speed of motion of the hands and breath in order to guide and induce the Clear Yang to ascend (as you raise your arms), allowing the Turbid Yin Qi to descend (as you lower your arms). This balances the body's Yin and Yang Qi (Figure 15.43).
3. If you have a Deficiency of ascending Qi (e.g., hypotension or prolapse of the viscera) begin with the palms facing upward, in front of your Lower Dantian. Slowly inhale while gradually raising the palms close to your body. When exhaling, move the hands downward quickly away from the body to avoid any excessive downward flow that would counteract the tonifying upward flow (Figure 15.44).

MEDICAL QIGONG WALKING THERAPY

Dynamic postural training can be performed by keeping the feet stationary and rocking the weight between the heels to the balls of the feet, or by walking. The Chinese saying, "The body must walk like the wind," describes the relaxed placement of the feet and supple stepping techniques used in Medical Qigong Walking therapy. Walking, or Toe-Raised Stepping, as it is commonly called in China, is excellent in treating the diseases of the Five Yin Organs. It can stimulate the Qi of the Spleen and Liver (due to their channels originating on the big toes), and can cause the body's internal Qi to rise and flow vigorously, thus reinforcing the function of Qi and Blood.

With each step of the Walking Meditation, the patient should practice tranquil mind and body balancing. This is the same imaginative technique as is used in standing, sitting, or lying meditations. The inner stillness and outer motion are in harmony. Touching the heel to the ground, with the

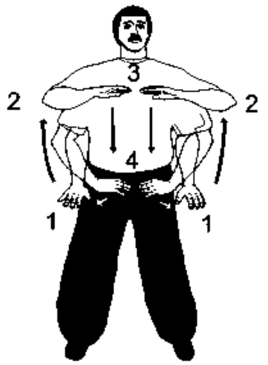

Figure 15.42. With palms facing downward (1), move the hands quickly upward (2). As the hands move slowly downward (3), allow the middle fingers on each hand to touch (4).

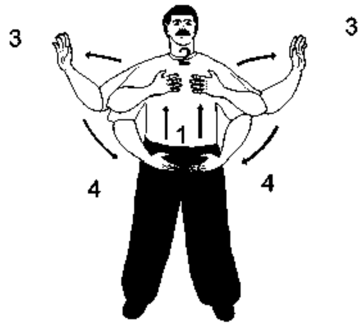

Figure 15.43. With palms facing the body (1), move the hands upward (2). Face the palms away from the body (3) as the hands move downward (4).

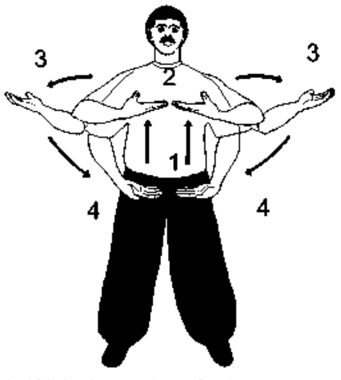

Figure 15.44. With the palms facing upward (1), raise the hands slowly close to the body (2). When moving the hands downward (3) in order to start again, quickly move them away from the body (4).

toes raised can also activate the Yin and Yang Heel Vessels (which originate from the center of the heel), and allow the energy to flow between both vessels so that the internal energy, exiting from the Yin Vessel, can enter the Yang Vessel, and vice versa. This allows the vessels to be dredged and the Yin and Yang energies to be regulated.

The heel is also stimulated when stepping, thus increasing the flowing action of the Bladder and Kidney Qi (due to their channel placement on the ankle). While stepping, focus the intention on the Lower Dantian to guide downward any abnormal rising of Lung Qi, which can cause shortness of breath, and affect the Heart and blood pressure. Each step is placed deliberately, with intention, and without collapsing of the feet. This stepping action relieves any Excess Qi syndrome of the upper torso and coordinates the balance of both the upper and lower portions of the body's Qi. Medical Qigong walking therapy consists of three training methods.

1. The Preparation (before walking), consists of the following steps:
   - The Opening and Closing of the Three Dantians, and
   - The Stationary Qi Strengthening and Regulating Exercises.
2. The Walking Therapy, consists of one or several prescriptions of the Toe Raised-Stepping techniques.
3. The Ending (after walking), consists of the following steps:
   - Leading the Qi back into the Lower Dantian,
   - Kneading the abdomen, and
   - Closing the Dantians and self-massage.

THE FUNCTIONS AND BENEFITS OF THE PREPARATION STAGE (THE OPENING AND CLOSING OF THE THREE DANTIANS)

Before beginning Medical Qigong Walking Therapy, the Opening and Closing of the Three Dantians must be practiced. In Medical Qigong training there is a saying, "When the body is Open, the Evil Qi will be expelled. When the body's surface is free from Evil Qi, it is better to Close it." The terms "opening" and "closing" refer to the purification of the body by means of Opening (purging) and Closing (sealing) specific areas of the body. The Opening and Closing of the Three Dantians can regulate either an Excess or Deficiency syndrome that is caused by an imbalance between the upper and lower, or interior and exterior parts of the patient's body. The purpose of the preparation stage is to:

- Eliminate the body's Evil Qi,
- Descend the Turbid Yin Qi, and
- Circulate the Clean Yang Qi through the organs and channels before walking.

The focus of this particular technique opens and closes the Upper Dantian (at the Yintang point), the Middle Dantian (at the Shanzhong CV-17 point), and the Lower Dantian (at the Qihai CV-6 point).

The Opening and Closing technique utilizes four specific disciplines:

- Using the Mind's Intention,
- Dividing the Yin and Yang Energy (ascending the Clear Yang Qi and descending the Turbid Yin Qi),
- The Hand Postures for Opening and Closing the Dantian, and
- The Focus on the Breath.

The Opening and Closing of the Three Dantians also requires multiple soft focusing that becomes easier with each practice session. Arm movements and the breath need to be perfectly coordinated; the expelling of Toxic Qi from the internal organs needs to be visualized; and the abdomen needs to be contracted while inhaling, and expanded while exhaling. The mind's intention must assist in the Opening and Closing of each Dantian, the coordinated movements alone do not suffice.

1. When focusing the mind's intention while performing the Opening and Closing techniques, imagine a dark Turbid energy from deep inside the body exiting through the body's pores. As the energy leaves, seal off the tissues to prevent it from returning to the body. Imagine the dark Turbid energy being absorbed into the ground, and being purified by the Earth.
2. The Dividing the Yin and Yang Energy refers

to separating the Yin and Yang energy within the body for healing. The Yin Qi stores the substance of vital energy, and the Yang Qi guards against pathogenic Evils and protects the body.

   a. Ascending the Clear Yang Qi. When the Upper Dantian is Opened and Closed it allows the Clear Yang Qi to ascend to the upper orifices. The increased circulation of Clear Yang Qi through the channels and organs of the body stimulates and energizes the body's surface, as well as strengthens the four extremities. When the Clear Yang Qi moves upward its pure nature helps to tranquilize the Shen.
   b. Descending the Turbid Yin Qi. When the Middle and Lower Dantians are Opened and Closed, it allows the Turbid Yin Qi to descend and facilitates the storing of the Clear Yang Qi.

3. The hand postures for Opening and Closing the Dantian is described as follows. The action for the Opening of Qi is directed by the outward movement of the hands and arms, and the mind's intent on Opening (or expanding) the abdomen.

   The Closing of Qi is directed by the inward movement of the hands and arms towards the abdomen, and the mind's intent on Closing (or contracting) the abdomen.

   The function of the Dao Yin Posture training, while regulating the breath and reinforcing the body's Qi, relies primarily on the movement of the hands and arms to arouse and invigorate the Qi circulating in the Large Intestine and the Lungs' channels. The action of opening and closing the Upper, Middle and Lower Dantians causes the Turbid Yin Qi to descend and be expelled, while the Clear Yang Qi ascends to the upper orifices of the body.
4. The focus is placed on the breath. The Opening and Closing exercise uses Reversed Breathing, i.e., the abdominal muscles contract while inhaling and relax and expand during exhalation (Figures 15.45 and 15.46).

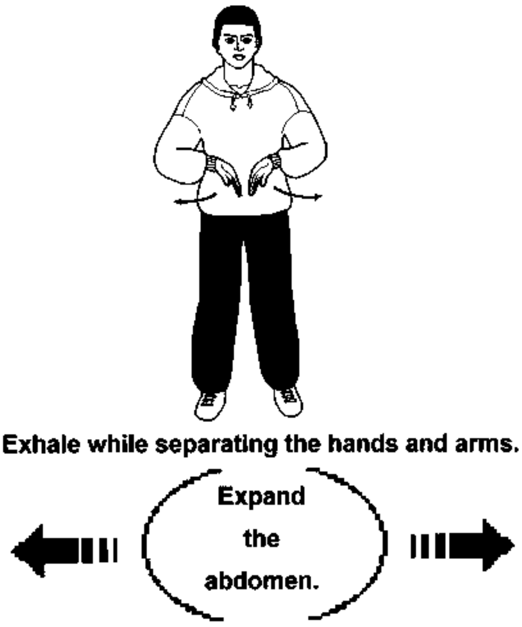
Exhale while separating the hands and arms.
Expand
the
abdomen.

Figure 15.45. The Opening of Qi is directed by the outward movement of the hands and arms and an opening (or expanding) intention in the abdomen.

This method of breathing is also referred to as Daoist breathing. The action of abdominal compression acts as a bellow to purge any Toxic (Yin) Qi from the body. These contractions and expansions of the abdomen are combined with the mental focus on the Lower Dantian. The nerve plexuses in the Lower Dantian area are considered, in modern Medical Qigong theory, to connect the Tai Yang Channels (Major Yang channels). If the mind is concentrated on the Lower Dantian for a long period of time, a conditioned reflex will occur that causes the nerve plexuses to produce bioelectricity, which spreads into the surrounding areas of the Tai Yang channels. This reflex causes the small blood vessels and capillaries to dilate. This reaction reduces the pressure in the blood vessels and improves Blood circulation to the Heart.

- Contraindications:

  It is important to note that Reversed Breath-

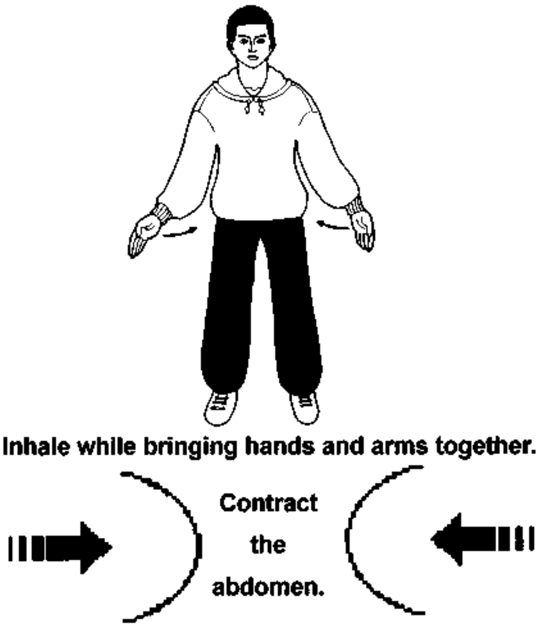
Inhale while bringing hands and arms together.
Contract the abdomen.

Figure 15.46. The Closing of Qi is directed by the inward movement of the hands and arms toward the abdomen for a closing (or contraction).

ing is contraindicated for young children, older patients and for people suffering from high blood pressure, stroke, or Heart diseases. In these cases, the Reversed Breathing techniques should be replaced with Natural Breathing. When practicing with Natural Breathing, the patients imagine Heavenly Qi descending into their body, filling and expanding the Lower Dantian, as they inhale. This expansion of the abdomen (upon inhalation) allows their Lower Dantian to Open, causing the Heavenly Qi to naturally purge the body of Toxic Qi. Upon exhalation, the patients contract their lower abdominal muscles to Close and seal their Lower Dantian.

THE OPEN AND CLOSE THE THREE DANTIANS METHOD

Normally an exercise begins with an equal number of inhalations and exhalations receptively, as the arms move toward the body, and away from the body. In this particular exercise, however, when approaching each separate Dantian, the first Open and Close action is to be performed by inhaling through the entire movement, without any exhalation. This action stimulates the opening of the Dantians. For the second through fourth Open and Close movements, the patients exhale when Opening and inhale when Closing. Practice the Opening and Closing of the exercise, step by step.

- Begin from Wuji posture. Raise your arms and hands to face the Upper Dantian. As you inhale, sperate your arms outward away from the body as you contract the abdominal muscles, then while still inhaling, bring the arms back to the Dantian. Exhale as you separate the arms away from the Dantian, allowing the abdomen to expand, inhale and contract the abdomen as you bring the hands back together. Repeat this action two more times.
- Place your arms and hands in front of your Middle Dantian. As you inhale, separate your arms outward away from the body as you contract the abdominal muscles, then while still inhaling, bring the arms back to the Dantian. Exhale as you separate the arms away from the Dantian, allowing the abdomen to expand; inhale and contract the abdomen as you bring the hands back together. Repeat this action two more times.
- Place your arms and hands in front of your Lower Dantian. As you inhale, sperate your arms outward away from the body as you contract the abdominal muscles, then while still inhaling, bring the arms back to the Dantian. Exhale as you separate the arms away from the Dantian, allowing the abdomen to expand; inhale and contract the abdomen as you bring the hands back together. Repeat this action two more times.

STATIONARY REGULATING EXERCISES (PREPARATION FOR WALKING)

This next preparation exercise is good for the patient's general tonification and for increasing respiration for Lung cancer patients.

- Start from a Wuji posture, inhale and exhale through the nose. Place your palms facing your lower abdominal area and begin extending Qi into the naval. Imagine the energy flowing into your navel, extending deep into your Mingmen and Lower Dantian area (Figure 15.47).

- Next, face your palms toward the ground and absorb the Earth Qi into your Lower Dantian (Figure 15.48). Absorb the Earth's energy for a few minutes.
- Turn your palms upwards, towards the Heavens, and begin absorbing the Universal Qi for several minutes (Figure 15.49). Finally, begin to close the exercise by allowing your palms to turn back towards your navel, absorbing and rooting the Qi back into the Lower Dantian

WALKING COMBINED WITH COLOR THERAPY

In China, to help visualize specific colors for therapeutic use, patients will carry brightly colored silk ribbons while walking. The colors and their affect on the internal organs are matched in accordance with the Five Elemental colors.

WALKING COMBINED WITH SOUND THERAPY

In China, in order to help the patients step with a structured rhythm, coordinate breathing patterns, and improve mental intention, Medical Qigong practitioners beat large wooden drums, click wooden sticks, and clang large metal cymbals. These rhythmic wooden and metal sounds resonate the walking area with an energetic field vibrating with Yin and Yang Qi. The patients, while walking in a circle, naturally absorb the Yin and Yang energy resonating from the instruments and seek to naturally regulate their internal organs.

THE THERAPEUTIC EFFECTS OF MEDICAL QIGONG WALKING THERAPY

The purpose of "raising your toes when stepping" is to invigorate the Yin and Yang channels and vessels of the legs. The energetic nature of the Spleen necessitates walking slowly during this exercise. The Spleen is the source of Postnatal Energy and is responsible for the activity of the muscles and extremities. The movements of the four limbs therefore strengthen the Spleen; this is why "toe-raised walking" and the "hand swaying" methods have both been adopted into the exercises and postures of most Medical Qigong therapies.

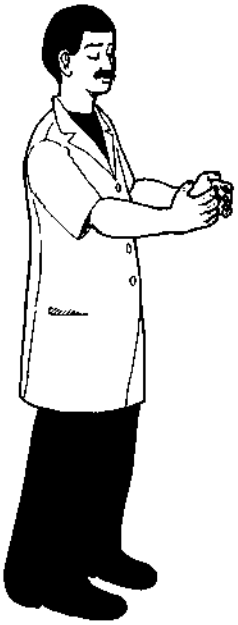

Figure 15.47. Begin by placing the palms toward the navel, in order to stimulate the Lower Dantian Qi.

Waist rotation plays a leading role in regulating the Liver Channels, as well as invigorating the circulation of Kidney Qi. This is achieved through all activities that initiate the actions of dropping, bending, relaxing, and turning the waist.

The opening and closing actions, combined with the rising and falling movements, and the swaying action of all four limbs, promote a pumping action within the muscles. This action is further intensified with the bending and leaning of the upper body, squatting and crouching of the lower body, and the alternate weight shifting when stepping. This pumping action greatly benefits the circulation of the lymphatic system.

Clinically, the Medical Qigong walking exercises result in increased appetite, strengthening

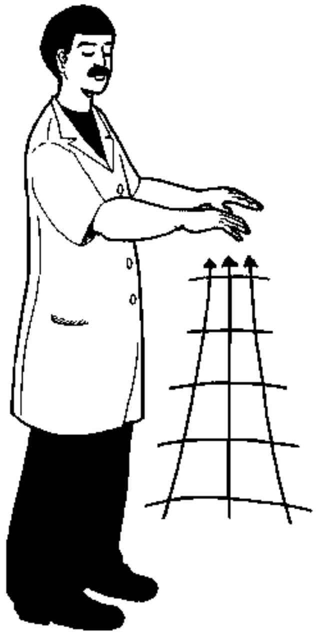

Figure 15.48. Next, begin to absorb the Earth energy into the Lower Dantian and Mingmen areas.

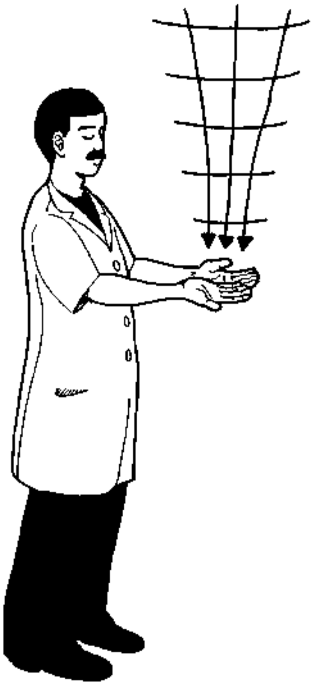

Figure 15.49. Turn the palms up to absorb Heaven Qi into the Lower Dantian Area.

of the patient's body, tonification of the Kidneys, Lungs, Liver, and Spleen, and the strengthening of Heart function.

Cancer patients need to regulate the function of their tissues and Yin and Yang organs to increase their healing potential. The initial goal in treating cancer with Medical Qigong, is to have the patient focus on the condition of the tumor, and its collection of energy. The patient imagines that the tumor area is beginning to soften, and the tissues are gradually transforming back into insubstantial energy. This technique is used to gradually dissolve tumors and certain types of less aggressive cancer.

The following are examples of Toe-Raised Slow Walking therapeutic exercises and prescriptions used in Medical Qigong therapy:

- Dynamic Regulation Walking Therapy
- Lung Tonification Walking Therapy
- Kidney Tonification Walking Therapy
- Liver and Spleen Tonification Walking Therapy
- Heart Tonification Walking Therapy
- Cancer-treatment Prescriptions

The explanation of these walking and breathing exercises and prescriptions are described as follows:

1. Dynamic Regulation Walking Therapy is a good exercise for regulating the Yin and Yang Qi of the patient's entire body. It can be prescribed for general tonification and nourishing of the patient's body, as well as strengthening the "anti-pathogenic" or True Qi. The Dynamic Regulation Walking Therapy is the primary walking exercise used to treat cancer patients.

   In every exercise the patient starts and ends with the Opening and Closing the Lower Dantian method. When stepping, the toes should be raised with the heels touching the ground (rocking back and forth), stimulating the Yin and Yang Heel Vessels of the legs, to strengthen the Kidney and Bladder Channels. As you step forward with the heel of the right foot, the Kidney and Bladder Channels open up. Swing your left hand up to the center of your Heart (CV-17 point) while your right hand swings to your right hip around and near the GB-30 point (Figure 15.50).

It is important to utilize the mind's concentration in conjunction with the breath. The breathing pattern is as follows:

- First step—inhale through the nose
- Second step—inhale through the nose
- Third step—exhale through the mouth
- Fourth step—hold the breath

The same details of stepping and hand movements apply to stepping forward with the left foot.

2. Lung Tonification Walking Therapy is a good exercise for tonifying the Qi of the patient's entire body. It can be prescribed for nourishing the patient's skin, as well as strengthening the True Qi. It is used for treating Lung Qi Deficiencies, bronchitis, tuberculosis, asthma, emphysema, and sinusitis.

The heel stepping pattern is similar to the Dynamic Regulation Walking exercise, however, when stepping, the hand postures will now switch from a normal relaxed position to the thumbs and first fingers touching at the tips (Figure 15.51).

Some Medical Qigong hospitals encourage the patients to rub their thumb and index fingers together when practising the Lung Tonification Walking Therapy. This rubbing action causes friction to develop between the Shaoshang Lu-11 and the Shangyang LI-1 points, which in turn stimulates and invigorate the flow of Qi along the Lung and Large Intestine Channels, strengthening the Lungs.

The breathing pattern is as follows:

- First step—inhale, inhale
- Second step—inhale, inhale
- Third step—exhale
- Fourth step—no breath

3. Kidney Tonification Walking Therapy. This exercise will purge the Liver Fire and tonify the Kidney Yin. It can be prescribed for patients who suffer from Kidney Qi Deficiencies, edema, blood diseases, gynecological diseases, and cancer of the Kidneys.

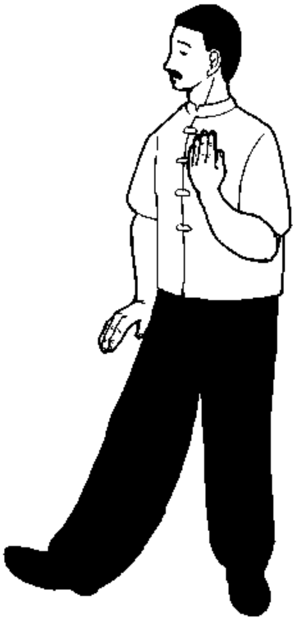

Figure 15.50. With each step, the opposite hand swings up to the Heart and Middle Dantian area.

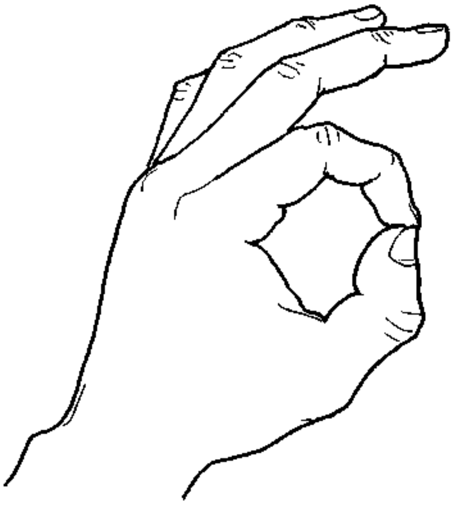

Figure 15.51. Tonification Hand Position for the Lungs

Begin this exercise with the palms embracing an imaginary ball at the naval and Lower Dantian area. This walking therapy causes the Qi to tonify your Lower Dantian.

Step forward so that the toes of the right foot touch the ground first, followed by the heel. This stepping action will open up the Yang Channels on the outside of your foot. Next, step with the left heel and point the toes up. This stepping pattern should be toe, heel, then step with heel (Figure 15.52).

On the first step, inhale through the nose while imagining the Qi filling your thorax. On the second step, sound the tone "Xu" (pronounced Shu) as you exhale. This particular sound is used to purge Toxic Liver Qi out the patient's body, allowing the Kidneys to fill with the rising Earth Qi.

- First step—inhale
- Second step—exhale "Shu"

The breathing pattern is as follows: inhale for the toe and heel, step and exhale on the heel step (all three actions equal one set). After walking in a circle for nine complete rotations, pause and begin again, starting with the left foot as the downward foot that makes the toe heel step as you exhale.

While stepping, keep both palms on your Lower Dantian. Regulate your breathing, while drawing the Earth Qi upward from the Yongquan Kd-1 point at the bottom of each foot. Press the Mingmen area backwards while breathing in the Earth Qi; this reinforces the Kidneys' Yuan Qi.

When ending the Kidney exercise, place both hands at the center of your chest at the CV-17 point. For men the right hand is on the outside, the left hand is on the inside, touching the chest. The hand position is opposite for woman. Make the sound "Shu" three times. End with the Opening and Closing of the Lower Dantian.

4. Liver and Spleen Tonification Walking Therapy activates the Liver's function of storing Blood and the Spleen's function of transforming Blood. It is used to treat digestive and blood diseases, hepatitis, Liver cancer, sclerosis of the Liver, and Gall Bladder stones.

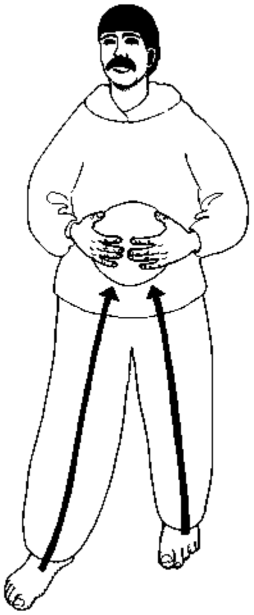

Figure 15.52. The Kidney Strengthening Step is as follows. While stepping, place both palms on the Lower Dantian, regulate your breath and draw the Qi upward from the bottom of each foot. Press your Mingmen area backwards while breathing in the essence of the Qi.

The starting position is similar to the exercise for tonifying the Kidneys with the toes pointing down, except that you step forward with the left toes. Step forward so that the toes of the left foot touch the ground first followed by the left heel. While this step is taken inhale twice through the nose. Then step out 45 degrees with the right heel while inhaling twice more through the nose. Next, step out with your left heel 45 degrees, exhale through your nose twice, then step out with your right foot with the toes pointed down as you exhale once.

- First step—inhale, inhale
- Second step—inhale, inhale
- Third step—exhale, exhale
- Fourth step—exhale

When walking with the toes raised, let the big toe touch the ground to stimulate the Yinbai Sp-1 and the Dadun Lv-1 points on the big toe. This promotes the ascension of the Qi flowing upward along the Liver and the Spleen Channels. The internal energy of these two channels is then reinforced, and the Liver and Spleen are strengthened.

After walking in a circle for nine complete rotations, pause and begin again starting with your right foot. Spleen and Liver Channels are activated when the toes pull up and the heel steps down.

The hand postures are as follows:

a. For Spleen problems the thumbs touch your ring fingers only when you step out to the fourth step (Figure 15.53).
b. For Liver problems the thumbs continuously touch your ring finger.

5. The Heart Tonification Walking Therapy calms the mind while tonifying the Qi and Blood. It is used for treatment of coronary Heart diseases, arthritis, palpitations, and angina pectoralis.

The starting position is the same as for the Liver and Spleen Walking therapy. The stepping is exactly like the last exercise, with the exception of the hand positions. In this exercise, the hands are opened and relaxed until they reach the chest (CV-17 point) at which point the middle fingers of each hand touches the Pc-8 point at the center of each palm (Figure 15.54).

When breathing, focus your Mind on inhaling the Qi upward from the Earth. When exhaling, imagine that the Qi within your chest is flowing downward.

- First step—inhale
- Second step—exhale

This walking exercise is practiced at a slightly slower speed then the previously described exercises. On the third step, the middle finger of each hand touches the center of your palms. This stimulates the Zhonchong Pc-9 point on each finger and is used to invigorate the flow of internal energy along the Pericardium Channels.

THREE STEPPING METHODS USED TO TREAT CANCER

These supplementary Medical Qigong Step-

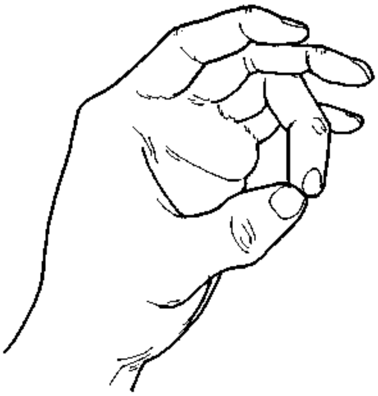

Figure 15.53. Hand Position for Tonification of the Liver and Spleen

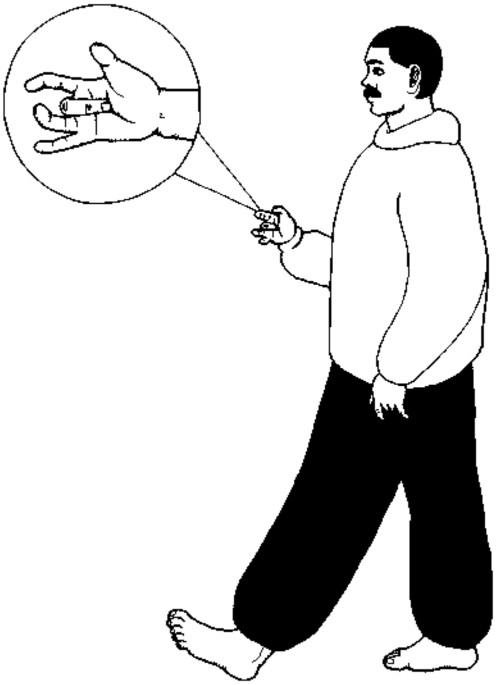

Figure 15.54. The Heart Strengthening Step is practiced at a slightly slower speed. On the third step, the middle finger of each hand touches the center of the palm.

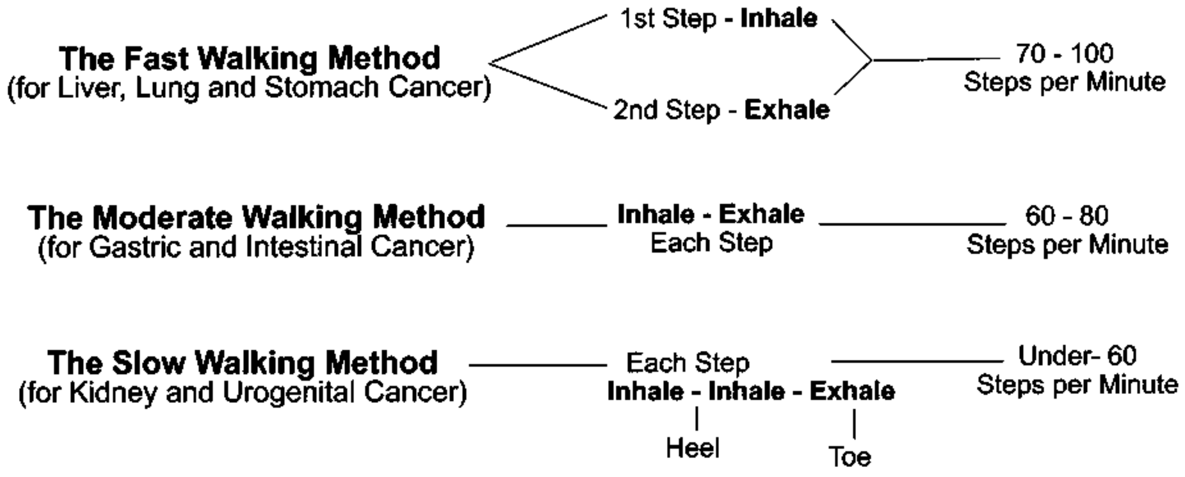
The Fast Walking Method
(for Liver, Lung and Stomach Cancer)
1st Step - Inhale
2nd Step - Exhale
70 - 100
Steps per Minute
The Moderate Walking Method
(for Gastric and Intestinal Cancer)
Inhale - Exhale
Each Step
60 - 80
Steps per Minute
The Slow Walking Method
(for Kidney and Urogenital Cancer)
Each Step
Inhale - Inhale - Exhale
Heel
Toe
Under- 60
Steps per Minute

Figure 15.55. Cancer Walking Therapy

ping exercises are given to cancer patients for breath regulating and Qi reinforcement. These breathing methods are applied with the Dynamic Regulation Walking therapy (the first walking exercise taught in this section). These walking methods use the "exhaling method" to strengthen the Kidneys, increase oxygen intake and enhance the patients' immune system.

The stepping methods are performed at three different speeds depending on the patients' symptoms (Figure 15.55). The patients are encouraged to walk at least once a day for 15-20 minutes. These walking exercises are classified into three kinds of prescriptions according to the following breathing rates.

THE FAST EXHALING METHOD

This walking method is mainly prescribed for patients in the early stages of cancer and is especially suitable for patients with Liver, Lung, and Stomach cancer. This exercise is practised using even breathing, inhaling for one step, followed by exhaling for another step. The patient steps out with the left foot, while inhaling through the nose. Then steps out with the right foot, and exhales through the nose. Both hands swing naturally up to the chest, while walking.

The main point to remember while stepping, is that the head leads the waist in turning and swinging the arms. The speed should be 70-100 steps per minute. When practicing the Fast Exhaling method it is important not to overexert yourself, as this would be counterproductive, and would lead to depleting the immune system.

THE MODERATE EXHALING METHOD

This walking method is mainly prescribed for patients in the early stages of cancer and is especially suitable for patients with gastric and intestinal cancer. The patient breathes in and out through the nose with each step taken. When taking each step, focus is on the respiratory rhythm. The speed should be between 60-80 steps per minute.

THE SLOW EXHALING METHOD

This walking method is mainly prescribed for patients with Kidney and urogenital cancer. The patient steps out with the left foot first and then with the right foot. Each step is accompanied by two inhalations and one exhalation. The patient steps out with the heel while taking two inhalations and then touches the ground with the ball of the foot, while exhaling. Because one step involves three breathing movements, the walking speed should be a bit slower, under 60 steps per minute.

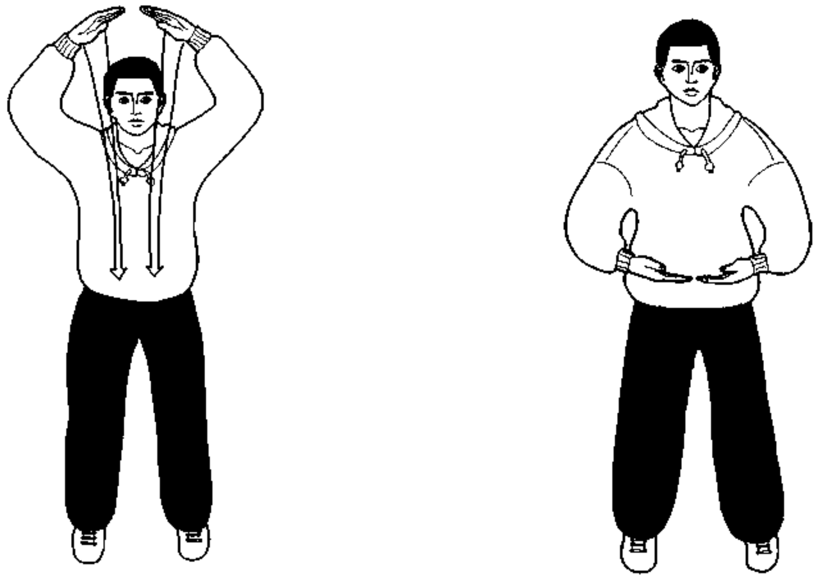

Figure 15.56. Secure the Upper and Middle Dantians in order to come to a tranquil state of relaxation.

ENDING AND CLOSING THE WALKING THERAPY

To end the Medical Qigong Walking therapy, the following exercises are taught to the patients in order to bring their Qi back into their Lower Dantian and to return to a state of quiescence. The Ending and Closing procedures are divided into three forms of practice: (1) Leading the Qi back into the Lower Dantian; (2) The official Closing of the prescription; and (3) Self-massage Regulation. These three ending procedures are described as follows.

LEADING THE QI BACK INTO THE LOWER DANTIAN.

This particular exercise is divided into four separate stages: (1) Securing the Three Dantians, (2) Kneading and Rooting the Lower Dantian, (3) Circle Massaging the Lower Dantian, and (4) Rooting the Qi.

1. The Securing the Three Dantians exercise has three purposes: to initiate a tranquil state of relaxation, to secure a harmonious balance of the body's Yin and Yang organs, as well as to bring the Qi back into the Lower Dantian.

   From a standing Wuji Posture, imagine embracing a ball of energy at the Lower Dantian (hold this posture for several breaths). After exhaling, separate the hands without inhaling, moving, and massaging the Qi along the hub of the body's Belt Vessel (at the waist).

   Inhale and begin raising the palms above the head. Imagine inhaling the Earth Qi into the body while the palms are facing the ground. At hip level, rotate the palms to face the sky, and begin inhaling Heaven Qi into the body. Continue raising the palms until they are positioned above the Baihui area on top of the head (left hand on top, opposite for women).

   Exhale and imagine the combined Earth and Heaven Qi flowing into the Taiji Pole, connecting all Three Dantians together. Descend the palms in front of the body, facing the Yintang, throat, Heart, upper abdomen (at the Yellow Court), and navel, then end by embracing the Lower Dantian. Begin again and repeat this exercise nine times (Figure 15.56).
2. Kneading and Rooting the Lower Dantian fills the Liver and Spleen with Blood and Qi, helping to further detoxify these organs.

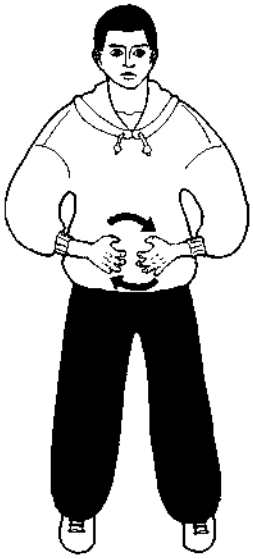

Figure 15.57. Knead the abdomen 36 clockwise then 24 counterclockwise.

After performing the Securing the Three Dantians exercise, secure the Qi in the Lower Dantian and knead the abdomen until it is sufficiently stimulated.

3. Circle Massaging the Lower Abdomen combines the body's organ energy together, thus forming the One True Qi. This allows the body to alleviate any Excess or Deficient conditions which may be caused from the excessive accumulation of Qi, due to the previous Qigong Walking exercises.

   Begin at the navel and start to circle massage the lower abdominal area thirty-six times in a counterclockwise direction (the movement should be directed against the flow of Qi in the Large Intestine). Begin with small and end with large circles, completing the movements at the base of the Lower Dantian. Next, reverse the direction of Qi flow by circle massage twenty-four times in a clockwise direction (with the flow of the Large Intestine), starting with big and finishing with small circles that end at the base of the Lower Dantian (Figure 15.57).

4. Rooting the Qi is sometimes referred to as "leading the Qi back to its origin." Rooting, extending, and stabilizing the energy into the Lower Dantian will facilitate the regulation of the Conception Vessel, Lower Burner, and Water pathways. Rooting the Qi also tonifies the Kidneys, Qi, Blood, and Yuan Qi, and allows the Toxic Qi to be easily expelled from the body.

   Leave the hands on the Lower Dantian and push the Mingmen backwards while inhaling the Qi and heat from the hands into the Lower Dantian (right hand on outside, opposite for women).

THE CLOSING

The closing procedure should be initiated after each practice period. After three deep breaths, click your teeth together thirty-six times (to stimulate the Kidney Qi), then circle the tongue eighteen times in each direction (to stimulate the Spleen Qi). Circle thirty-six times from the bottom of the teeth to the back of the throat. Tilt your head forward slightly, then swallow the saliva three times while imagining the Clear Qi of Heaven and Earth mixing with the saliva and flowing down to the Lower Dantian.

SELF-MASSAGE REGULATION

This method shifts the attention slowly off the area of concentration, while leading the Qi back down to the Lower Dantian. After the mind disengages from the specific point, relax the body, slowly open the eyes and perform self-massage.

Self-massage includes rubbing the hands, bathing the face (rubbing the face with the palms), combing the hair with the fingertips, and dredging the Twelve Primary Channels.

1. After rubbing the hands to heat the palms, place both hands on the face, cup the eyes, and draw the heat in through the eyes, into the Upper Dantian.
2. Next, comb the hair with your fingertips and drain the Excess Qi from off the head and neck area (Figure 15.58).
3. Then, rub from the feet to the abdomen to dredge the three Yin channels of the feet (Figure 15.59).

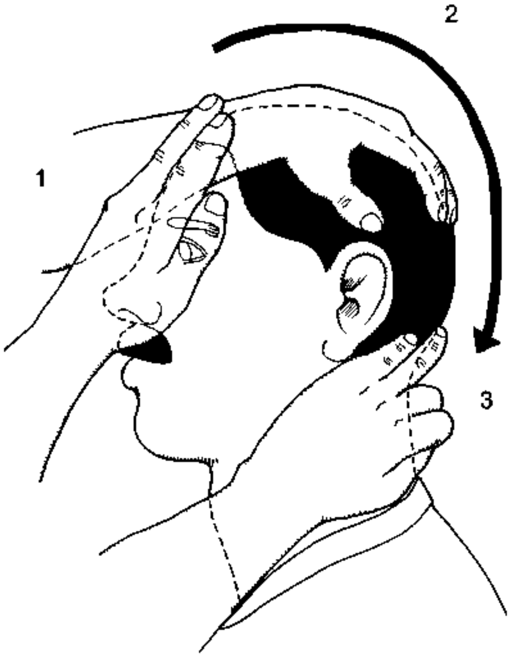

Figure 15.58. After rubbing the hands to heat the palms, place both hands on the face and draw the heat in through the eyes and into the Upper Dantian (1). Next, comb the hair with the fingertips (2), and drain the excess Qi from the head and neck area (3).

4. Rub from the chest to the hands to dredge the three Yin channels of the hands (Figure 15.60).
5. Rub from the hands, shoulders and lateral sides of the head, down the sides of the chest and abdomen to dredge the three Yang channels of the hands (Figure 15.61).
6. Rub from the waist and hips to the feet to dredge the three Yang channels of the feet (Figure 15.62).

These closing and self-massage procedures should be performed seven times each, followed by stretching to end the exercise.

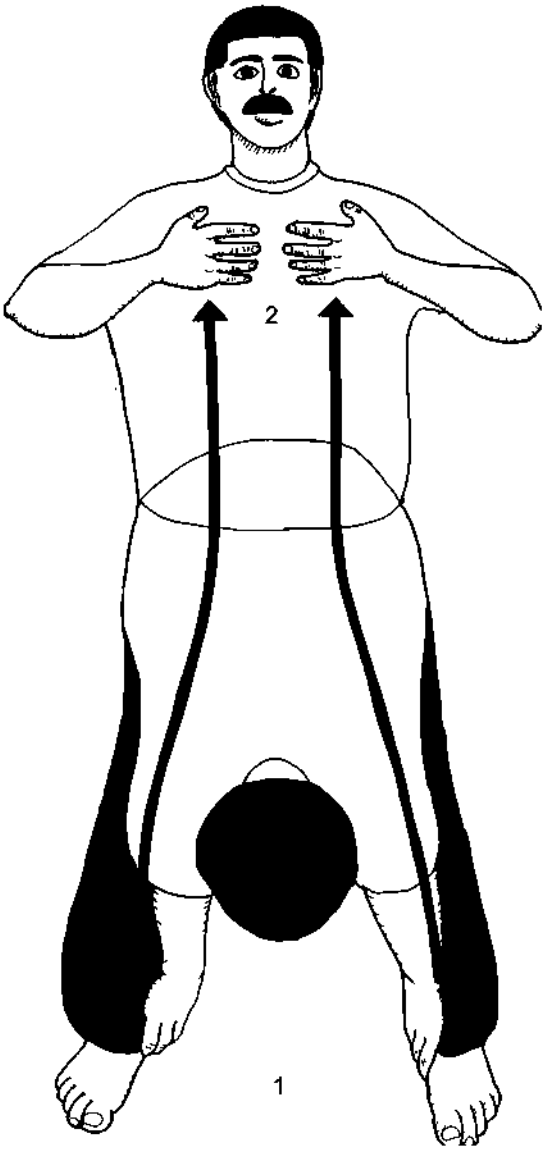

Figure 15.59. Rub from the feet to the abdomen (1–2) to dredge the three Yin channels of the foot.

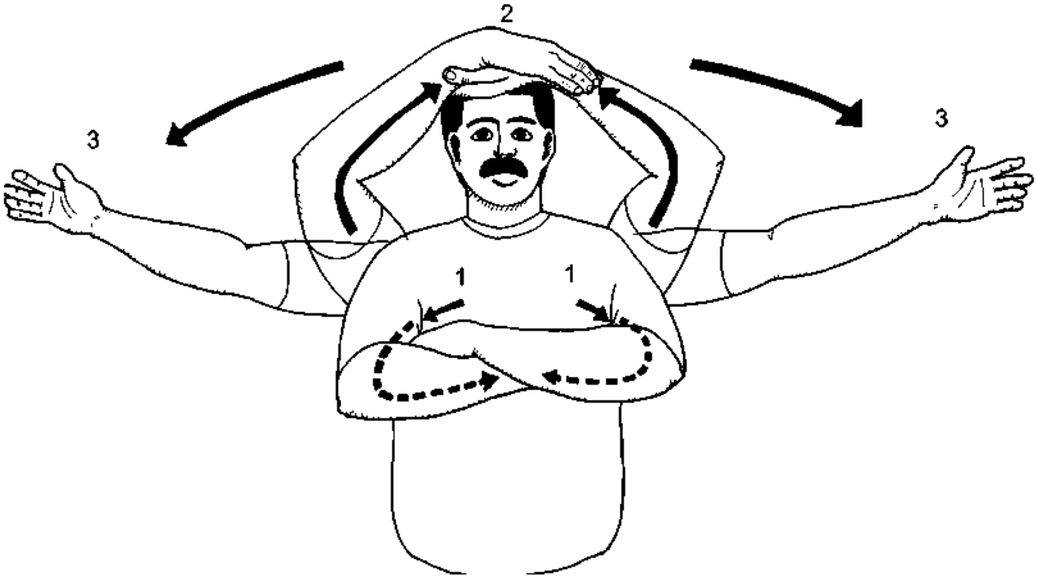

Figure 15.60. Rub from the chest (1) to the hands (2-3) to dredge the three Yin channels of the hand.

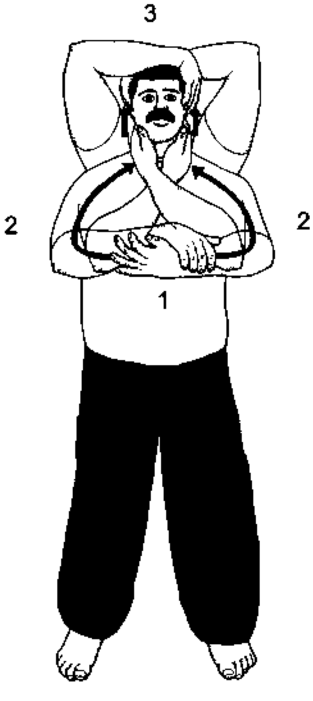

Figure 15.61. To dredge the three Yang channels of the hand, rub from the hands (1), shoulders (2), and lateral sides of the head (3).

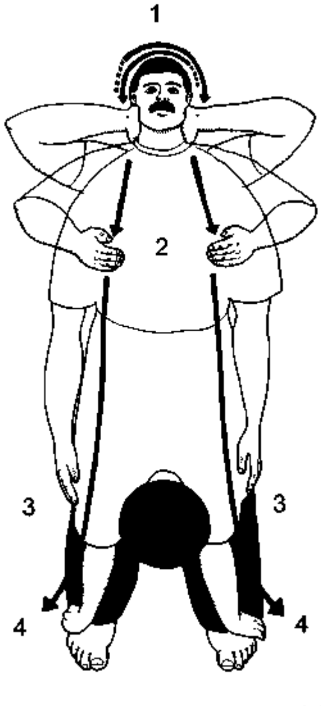

Figure 15.62. Continue down the sides of the chest (1) and abdomen (2). Next, rub from the waist and hips (3) to the feet (4) to dredge the three Yang channels of the foot.

CHAPTER 16
RESPIRATORY DAO YIN

THE PURPOSE OF RESPIRATORY DAO YIN

Respiratory Qigong therapy is often called Tu Gu Na Xin which means "expelling the old, drawing in the new." The purpose of Respiratory Dao Yin is to increase the intake and absorption of Clear Qi through regulating the breath. The extra absorption of Qi enhances the body's vitality, and harmonizes the Blood and Qi to promote health and eliminate disease.

Because the breath is also the link between the body and mind, by controlling the pace and quality of the breath, individuals can affect deep changes in their physiological functions. Breathing exercises have proven to be effective in reducing anxiety, depression, irritability, muscle tension, and fatigue. These techniques are also used in the treatment and prevention of agoraphobia, hypertension, hyperventilation, shallow breathing, and cold hands and feet.

THE ROLE OF THE LUNGS IN QI CULTIVATION

The Lungs control the body's Qi, although the Kidney Qi is responsible for aiding the Lungs by regulating the respiration (i.e., upon inhalation, as the Qi descends into the thorax, the Kidneys hold down and stabilize each breath). The Yuan Qi stored in the Kidneys must be continuously supplemented by the Gathering Qi derived from air, food, and water.

The internal absorption of air, food, and water rely on the smooth circulation of the Lungs' energy. Because of the significant role the Lungs play in absorbing universal and environmental Qi, the ancient Qigong masters placed a great deal of importance on breathing exercises. They believed that a patient's health and emotional condition were deeply affected by their breathing patterns, and by the amount of oxygen consumed in proportion to the amount of carbon dioxide released.

Through the study of these ancient Qigong observations, modern research confirms that relaxed and natural breathing patterns cause every tissue and cell to decrease its consumption of energy, while increasing its storage of Qi.

The ancient Qigong masters recommended that "vigorous breathing," in the form of Dynamic Breath Regulation, should be used when training Qi, and "gentle breathing," in the form of Quiescent Breath Regulation be used after training practice for nourishing and replenishing the Qi. Each breathing regulation exercise has long been considered a powerful prescription used to balance the patient's health.

DEEP INHALATION AND EXHALATION BREATHING METHODS

In breathing regulation exercises and prescriptions, the patient's inhalation is used to gather universal and environmental Qi into the body for tonification. Exhalation is used to eliminate Turbid or Toxic Qi from the body through purgation. Once the patients are in the prescribed posture, relaxed and free from distractions, they should, for example, exhale and imagine releasing Turbid Qi from out the mouth, nose, and pores of the body.

Clinical studies show that the Deep Inhalation of Medical Qigong prescriptions can have the following effects on the patient's body:

- increase stimulation of the sympathetic nervous system,
- cause the blood vessels to contract,
- raise the blood pressure, and
- increase the pulse rate.

On the other hand, the Deep Exhalation methods of Medical Qigong prescriptions can also have the opposite affects on the patient's body:

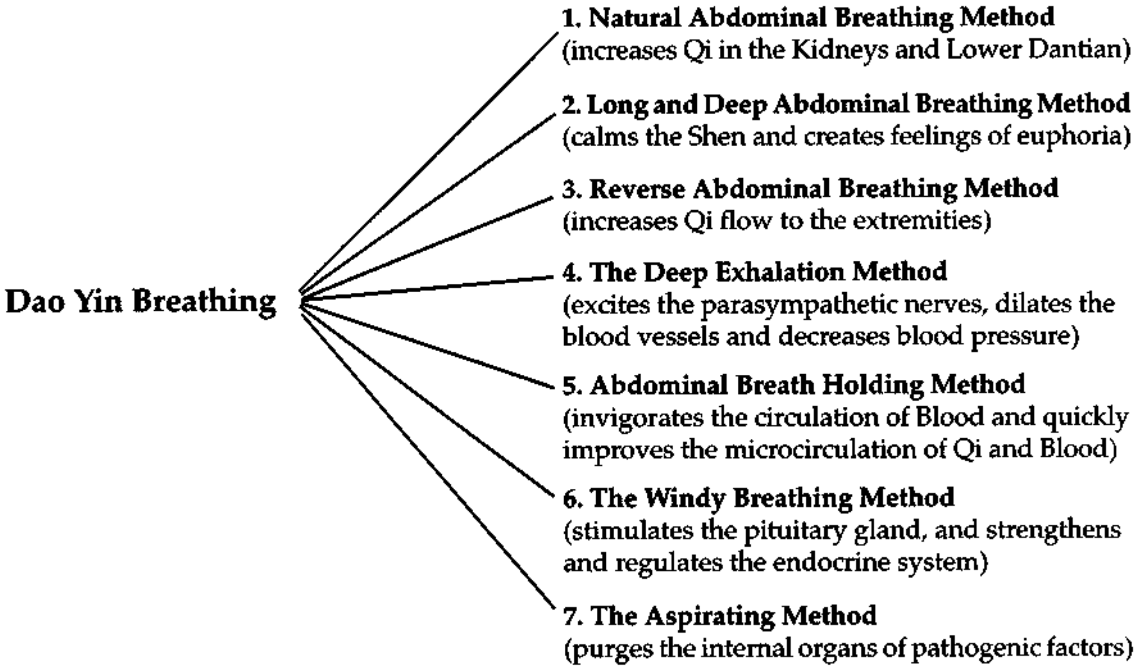

Dao Yin Breathing
1. Natural Abdominal Breathing Method
(increases Qi in the Kidneys and Lower Dantian)
2. Long and Deep Abdominal Breathing Method
(calms the Shen and creates feelings of euphoria)
3. Reverse Abdominal Breathing Method
(increases Qi flow to the extremities)
4. The Deep Exhalation Method
(excites the parasympathetic nerves, dilates the blood vessels and decreases blood pressure)
5. Abdominal Breath Holding Method
(invigorates the circulation of Blood and quickly improves the microcirculation of Qi and Blood)
6. The Windy Breathing Method
(stimulates the pituitary gland, and strengthens and regulates the endocrine system)
7. The Aspirating Method
(purges the internal organs of pathogenic factors)

Figure 16.1. The Seven Dao Yin Breathing Techniques

- increase the stimulation of the parasympathetic nervous system,
- dilate the blood vessels,
- lower the blood pressure, and
- decrease the pulse rate.

Deep exhalation helps to keep the body's Yin energy moving downward, and so has a downward regulating action on diseases caused by Yang Qi domination. Deep inhalation helps to keep the body's Yang energy moving upward, and has the effect of sustaining the Yang and tranquilizing the mind.

Students beginning Medical Qigong practice should focus their mind's intention primarily on the method of exhalation, to simultaneously stimulate the parasympathetic system and to relax. Once health is regained, patients will begin to alternate between deep inhalation and deep exhalation. This alternation produces a diastolic and systolic action of the blood vessels that also enhances the elasticity of the blood vessels' walls.

SEVEN DAO YIN BREATHING TECHNIQUES

The Seven Dao Yin Breathing Techniques are intended to regulate the symptoms caused by the Excess of Yang and Deficiency of Yin, which are characterized by too much energy in the upper part of the body and a weakness in the lower portion of the body.

Today, in China's Medical Qigong clinics, the breathing regulation exercises employ seven types of Dao Yin breathing techniques (Figure 16.1). Normally, all inhalation is done gently through the nose and exhalation is released slowly through the mouth. Inhalation and exhalation through the nose is also important, but should only be used when a higher state of "stillness" is obtained by the patient. This is to prevent the patient from falling back into his or her old shallow breathing patterns, until the Deep Abdominal Breathing method has been perfected.

1. The Natural Abdominal Breathing Method is also called Natural Breathing, and is used in

conjunction with mental concentration. In this method the abdomen expands upon inhalation and contracts upon exhalation. It naturally increases the body's peristaltic action, massages the internal organs, invigorates and increases the movement of Qi from the Kidneys into the Lower Dantian. Because of the differences in energetic physiology between men, women, and the breathing habits of each individual, Natural Breathing is further divided into Natural Thoracic Breathing, Natural Abdominal Breathing, and a combination of the two.

2. The Long and Deep Abdominal Breathing Method is also called Wen Huo Breathing. It is a gentle breathing method that consists of natural, deep, and full breathing. This results in a relaxed and comfortable feeling of euphoria. This method of breathing is performed as follows:
    a. At the beginning of the breath, inhale all the way down into the perineum, feeling the lower abdomen expanding in all six directions:

- from the bottom - at the base of the perineum,
- from the front - to the pubic bone and navel,
- from the back - to the coccyx and Mingmen,
- from the right - to the hip and lower ribs,
- from the left - to the hip and lower ribs, and
- from the top - at the base of the diaphragm.

    b. While continuing to inhale, feel the oxygen expanding and filling the upper thoracic cavity completely. The entire torso, from the perineum to the collar bones is expanded in six directions:

- from the bottom - at the top of the diaphragm
- from the front - to the solar plexus, Heart, and collar bones,
- from the back - to the base of the ribs, spine and top of the shoulders,
- from the right - to the rib area,
- from the left - to the rib area, and
- from the top - to the throat and neck area (Figure 16.2).

    c. Upon complete expansion, tuck the pelvic bowl under to increase the capacity of storing more energy. On the inhalation the pelvis performs a slight posterior tilt, on the exhalation the pelvis tilts slightly in the anterior direction. These actions will naturally ripple the spine. Exhale the breath from the upper thoracic cavity downward, to the lower abdomen.

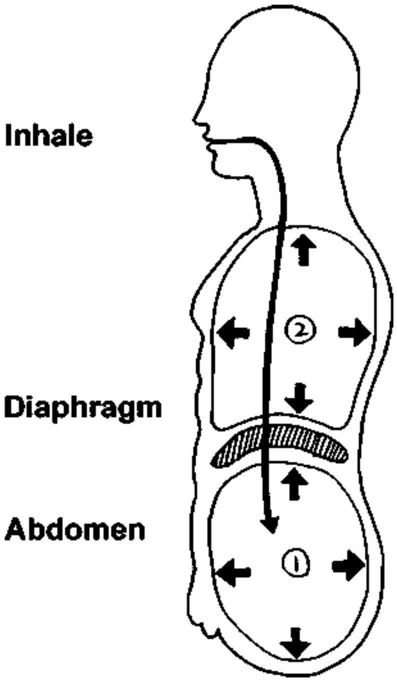

Figure 16.2. Inhale all the way down to the perineum. Make sure that the anal sphincter is closed in order to keep the Qi full and expanded in the Lower Dantian.

3. The Reverse Abdominal Breathing Method is an example, and subcategory of Long and Deep Abdominal Breathing. This method is performed by contracting the abdomen (and anus) when inhaling and expanding the abdomen upon exhaling. This is the primary breathing method used in Qigong training for guiding and emitting Qi, because the quality of air pressure and exertion of energetic pressure is greater. The Reverse Abdominal Breathing method results in superior efficiency in leading Qi to the extremities and has a much greater effect in raising the Qi from the legs into the brain.

However, when prescribing Medical Qigong therapy to treat diseases and alleviate physical stress, the Reverse Abdominal Breathing Method is prohibited except for very specific conditions. The reason it is not encouraged in Medical Qigong prescriptions is because patients with clinical hypertension or Heart disease often practice Reversed Breathing automatically. Scientific studies confirm that 90% of these patients practice reverse breathing chronically, as do most individuals with anxiety problems. Reverse Abdominal Breathing can also cause a constriction of the blood vessels and thus exacerbate problems such as cardiac diseases (i.e., angina) and migraine headaches.

4. The Deep Exhalation Method is often referred to as Wu Huo, which means "vigorous breathing." It is conducted with strong conscious intent. It can excite the parasympathetic nerves, dilate the blood vessels and decrease blood pressure.

   When practicing the Deep Exhalation method, the patient expels the Qi through the mouth, followed by a shallow inhalation.
5. The Abdominal Breath Holding Method invigorates the circulation of Blood and reduces swelling within the tissues, quickly improving the microcirculation. This method is also used to deepen the breath by stopping or holding the respiration for a short period of time.

   One Abdominal Breath Holding Method, known as Two Breathings, consists of inhaling through the nose, deep into the Lower Dantian, with one breath, pausing, then exhaling out through the mouth. The pause should be sustained for as long as possible without feeling tension (although a feeling of slight pressure building up is normal). After practicing deep abdominal breath holding, the breath will begin to develop its own stopping patterns. This stage involves the discipline and interdependence of the Heart and mind (Emotion and Spirit) with the breath.
6. The Windy Breathing Method consists of inhaling and exhaling through the nose. It is often referred to as "Windy" because it employs shallow breathing which resonates in the nasal passages. This kind of breathing purifies and warms the air as it passes through the nasal tract, resulting in air oscillation which stimulates the pituitary gland, thus strengthening and regulating the endocrine system. This method is designed to be used by patients with Lung disease because "the Lung energy opens at the nose."
7. The Aspirating Method refers to the method of drawing out by suction. This popular method can be used with healing sounds (or "tones") to purge pathogenic factors, or to regulate the body's internal organs. The pronunciation is based on resonating the sounds to stimulate specific organs such as the Heart, Spleen, Lungs, Kidneys, Liver and Triple Burners (Figure 16.3). There are two primary

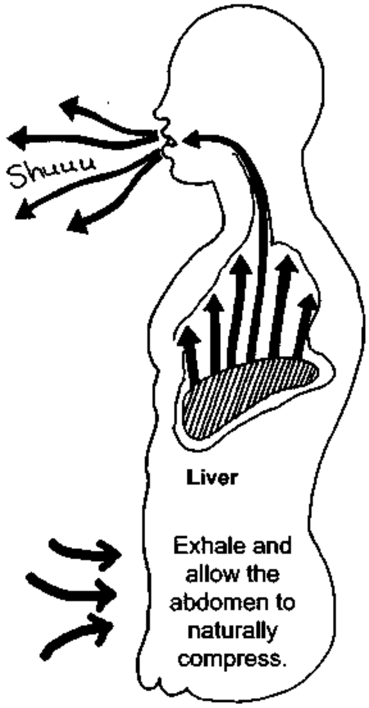

Figure 16.3. An example of tone resonation by aspirating the body's Excess Heat from the Liver is by using the sound "Shu."

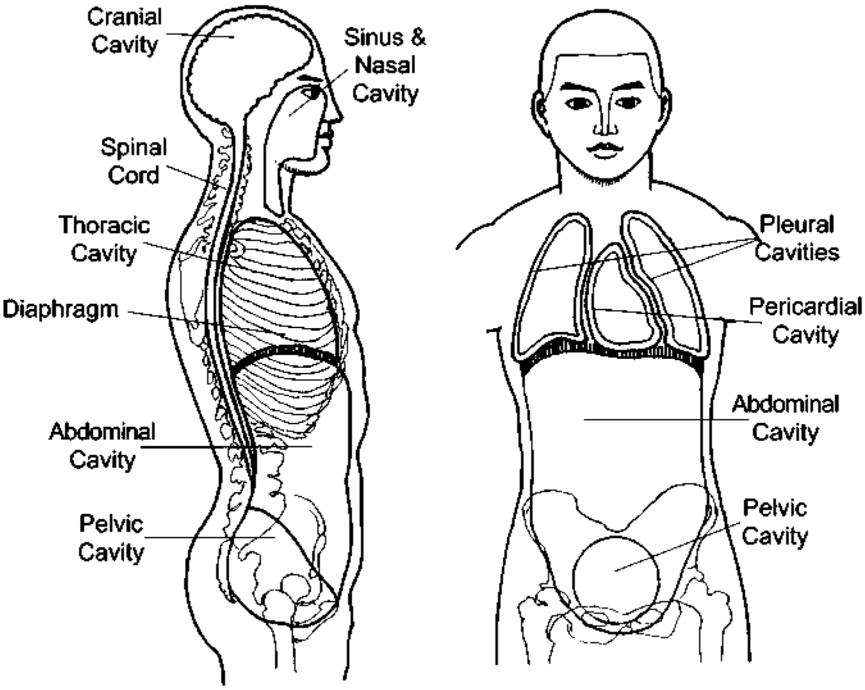

Figure 16.4. The body has seven main spatial cavities, as well as individual spatial cavities that surround the Yin and Yang organs.

sound systems of tone resonation used in Medical Qigong therapy, The Six Healing Sounds and the Tumor and Cancer Tone Resonations. These two healing sound methods are described as follows.

- The Six Healing Sounds are used to regulate, strengthen or detoxify the body's Yin and Yang Organs. This healing sound system is generally prescribed for patients with acute, or less serious conditions such as indigestion, insomnia, or Excess Heat in the organs, etc.
- Tumor and Cancer Tone Resonation Therapy is used to vibrate and destroy malignant tissue and cell growth in patients with chronic and serious organ diseases (see Chapter 40). This healing sound system is prescribed for diseases such as cysts, tumors, and cancer.

The choice of organ, and the number of times to pronounce each sound or character, is determined by the physical and energetic constitution of each individual. Each sound vibrates a specific organ or organ system, and surrounding set of involuntary muscles. Through repetition, the muscles acquire a kinesthetic memory of the vibrational sound. The doctor, or patient, can then trigger this memory by mental concentration alone to produce the desired results. As the patient performs the Sound Therapy, not only are the channels and organs cleansed, but also the spatial cavities (or Orbs) which surround the specific organs (Figure 16.4). For more information on Sound Therapy, see Chapter 40.

RESPIRATORY GUIDING AND REGULATING

There are four Respiratory Dao Yin methods the Qigong doctor can prescribe to regulate the

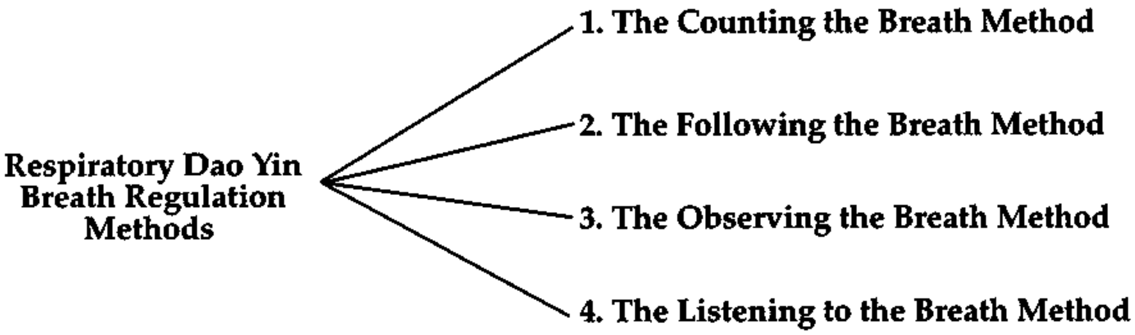
Respiratory Dao Yin Breath Regulation Methods
1. The Counting the Breath Method
2. The Following the Breath Method
3. The Observing the Breath Method
4. The Listening to the Breath Method

Figure 16.5. The Four Methods to Prescribe Respiratory Dao Yin For Breath Regulation

patient's breath (Figure 16.5): Counting the Breath, Following the Breath, Observing the Breath, and Listening to the Breath.

1. The Counting the Breath Method can be utilized to eliminate distractions of the mind during Walking Qigong Therapy by having patients silently count the number of their breaths. This method is also performed silently to regulate the body. In this particular method, one inhalation and one exhalation equals one breath.
2. When practicing the Following the Breath Method, the patients will contemplate each inhalation and exhalation of breath without counting. The patients mind, free from distractions or thoughts, follows each breath into and out of the body.
3. When practicing the Observing the Breath Method, the patients will observe their body's tissues conforming to the tranquil mental images from their subconscious imagination (e.g., relaxing, sinking, etc.).
4. When practicing the Listening to the Breath Method, patients are encouraged to listen to the "sounds" of their breathing (used in the Windy Breathing Method) to achieve a state of tranquillity.

RESPIRATORY TONIFICATION AND PURGING METHODS

Medical Qigong Respiration Therapy has its own unique methods and techniques for tonifying and purging the body's Qi. The primary technique used is focused concentration on the breathing. Through breath regulation, syndromes of Excess and Deficiency may be brought into balance either through tonification or through purgation.

1. Inhalation is used for tonifying the Qi in cases of Deficiency. The duration of the patient's inhalation should be longer than that of the exhalation. Tonification or Qi reinforcement patterns are created by applying the following methods:
   - more inhalation, and less exhalation
   - long inhalation, and short exhalation
   - focusing the mind on the inhalation
2. Exhalation is used for purging and reducing the Qi in cases of Excess. The duration of the exhalation should be longer than that of the inhalation. Purgation or Qi reducing patterns are created by applying the following methods:
   - more exhalation, and less inhalation
   - long exhalation, and short inhalation
   - focusing the Mind on the exhalation

BASIC ESSENTIALS OF RESPIRATORY TRAINING

Medical Qigong Respiratory training consists of the regulating, tonifying or purging exercises practiced in accordance with the doctor's treatment and specific homework prescriptions. There are four basic essential guidelines used to assist the patient in achieving maximum results in respiratory therapy, these guidelines consists of: Ensuring the Correct Posture Before Beginning, Purifying and Cleansing the Body's Qi, Tonifying and Purging the Qi, and Regulating the Body's Qi.

1. Before beginning Respiratory Dao Yin training, it is important that the doctor make sure the patient's posture is correct. When prescribing dynamic and static Medical Qigong therapy, it is best to start with a solid foundation and proper structure. Breath training should only begin when the patient becomes proficient in Postural Dao Yin, otherwise adverse effects may occur, (i.e., emotional trauma, headaches, chest stuffiness, or respiratory distress).
2. After their posture has been corrected, have the patients open their mouth and focus on their exhalation. As the patients exhale, they imagine all of the obstructed channels and collaterals being simultaneously purged of Turbid Qi. Each time the patients exhale they are to relax their body, more and more.

   Next, the patients close their mouth, and focus on inhaling clean, fresh Qi in to their body (through the nose). Have the patients practice this method using Natural Breathing. Each time the patients inhale and exhale they are to continue to relax the body.
3. After the body's Qi has been cleansed, the patients will focus their intention on any and all internal organs that specifically need to be purged. Breathing exercises used for purgation include focusing the patients' intention on the exhalation, as well as exhaling specific healing sounds. The patients continue to purge their body in this manner in accordance to the doctor's instructions.

   Having completed the purgation exercises, the patients are free to begin their tonification prescriptions. Breathing exercises used for tonification include the focus of the patients' intention on the inhalation, as well as visualizing specific colors being ingested into certain internal organs.
4. After practicing the tonifying or purgation prescriptions, the final goal of respiratory training is to achieve a long, natural, even and deep respiration. Urgent respiration causes the body's internal Fire to escape into the pores, resulting in uncontrolled Heat, feverish conditions and irritability. Normal breath regulation consists of even inhalation and exhalation patterns. The patient must avoid speeding up their breath, as this will also interrupt the breathing rhythm and lead to deviations of Qi.

BREATHING PATTERNS FOR COLLECTING AND REGULATING QI

The following are examples of exercises used for regulating and collecting Qi in the Lower Dantian. The doctor or patient will perform the breathing patterns as follows:

1. Collecting Qi in Lower Dantian is used to tonify and increase the collection of Qi gathered in the Lower Dantian.
   a. Inhale through the nose, filling the lower Dantian (expanding the abdomen), pause, then exhale through the mouth. Repeat and continue this breathing pattern for three times then proceed to b.
   b. Take three partial inhalations through the nose, filling the lower Dantian, pause, then complete one exhalation through the mouth. Next, inhale 50% of the breath into the Lower Dantian (expanding the abdomen), hold for three heartbeats, inhale remaining 50% into the Lower Dantian and hold for five heartbeats, relax and slowly exhale. Repeat and continue the pattern for thirty minutes a day (for three months), or until it becomes effortless.
2. Regulating Qi in Lower Dantian is used to stabilize the energetic balance of the Lower Dantian.
   a. Inhale through the nose, filling the lower Dantian (expanding the abdomen), exhale through the mouth, and

then pause. Repeat and continue this breathing pattern for three times then proceed to b.

   b. Inhale through the nose, then perform three partial exhalations through the mouth, pause, then one complete inhalation through the nose. Next, exhale 50% of the breath out of the Lower Dantian (contracting the abdomen), hold for three heartbeats, exhale remaining 50% out of the Lower Dantian and hold for five heartbeats, relax and slowly inhale. Repeat and continue the pattern for thirty minutes a day (for three months), or until it becomes effortless.

3. The Mysterious Pass is the pause between the inhalation and the exhalation. The ancient Qigong Masters believed that in the pause was the realm of infinite space (Wuji), existing between the Earthly Heaven (the congenital form of energy) and the Later Heaven (the acquired form of energy) where time and space stands still.

   For clinical practice, the Qigong doctors train in the method of accessing the Mysterious Pass in order to extend Qi through the entire body and to access the energy of the Wuji (Infinite Space). In the clinic, the training is divided into two stages (a and b).

   a. Begin this training by holding the breath from three to twelve counts (known as the "Small Count Breathing Regulation").
   b. Eventually increase to the "Big Count Breathing Regulation," which consists of 120 counts. This method enriches the Yuan Qi by breathing in more and exhaling less, and is sometimes used in Medical Qigong prescriptions to cure specific types of illness. When used with a strong intent, breath holding can constrict the blood vessels and raise the blood pressure; it is therefore sometimes prescribed for patients with hypotension. This Qigong method is, however, contraindicated for patients or doctors with hypertension.

THE FOUR PROGRESSIVE STAGES OF RESPIRATION

There are four progressive phases of difficulties that patients may experience during the beginning stages of their Medical Qigong Respiratory training: The Wind Phase, The Gasp Phase, The Unbalanced Air Phase, and The Quiescent Phase.

1. During the Wind Phase, the breathing is audible and may prove distracting to a beginner. The goal of focused concentration is to allow the energy to become gathered and collected. If the goal of the breathing prescription is to focus on a Deficient internal organ or specific tissue area, for example, and the patient's concentration becomes diverted (shifting to the noise of their own breath), it may disrupt the patient's mind. This can result in the patients collected Qi being dispersed instead of gathered.
2. During the Gasp Phase, the breathing is no longer audible, but any stagnant air, or obstructed air in the throat may disrupt the mind's concentration. Focused concentration on the stagnant air or obstruction may cause a sense of physical obstruction and mental anxiety, resulting in Qi stagnation in the throat (known as Plum Pit Qi).
3. During the Unbalanced Air Phase, the breath remains silent and there is no sensation of obstructed air in the throat. The patient's breathing, however, may be uneven, or unbalanced. Focused concentration at this phase may cause strain in the patient's mental awareness. If the mind becomes overstrained (trying to balance the inhalations and exhalations) the result will be fatigue.
4. During the Quiescent Phase, patients are able to achieve a long, natural, even and deep respiration without sound or stagnation, resulting in a state of extreme quietness. Concentration on this phase will result in a restful peace of mind.

DURATION TIME FOR RESPIRATION THERAPY

Patients in relatively good health should practice for about five minutes a day in the first week, when seeking to regulate their body's internal energy. In the second week patients should increase to ten minutes a day, and in the third week, fifteen minutes a day. Thereafter, five more minutes each day should be added until the patients reach a minimum of twenty minutes and a maximum of forty minutes per day for optimal Qi development.

When treating patients (for general tonification) using multiple standing meditation postures, have them begin by counting up to sixteen breaths per each posture. As the patients progress, the respiratory rate during meditation begins to slow down naturally. Normally, within a few months of training, respiratory rate drops from an average of sixteen breaths per minute, to three or five breaths per minute. This naturally increases the time spent in each posture, extending each posture from one to five minutes in duration.

Qigong doctors, on the other hand, should practice a minimum of forty minutes of standing posture each day to recharge and increase the efficacy of their treatments.

EMOTIONAL EFFECTS ON BREATHING PATTERNS

Breathing patterns are directly related to both the patients' thoughts and emotions. Any shift in one will affect the other. Each of the five primary emotions is associated with a specific breathing pattern as follows.

1. An angry patient's exhalation will be much stronger than the inhalation.
2. A grieving or sad patient's inhalation will be much stronger than the exhalation.
3. A fearful or scared patient's breathing will be fast, held high in the Lungs, and shallow because of the Kidneys' inability to hold the Lung Qi down.
4. A happy patient's inhalation and exhalation will be moderate, and irregular, with quick bursts.
5. When worried or deep in thought, the patient's inhalation will be shallow and weak, sometimes held for long periods of time, followed by a long gulping burst of inhalation and exhalation. The patient may also experience spontaneous sighing for prolonged periods of time.
6. When a patient experiences fright or shock, the breath immediately stops as the Qi attacks the Heart.

RESPIRATORY DAO YIN CANCER PRESCRIPTIONS

Respiratory prescriptions such as the Deep Relaxed Breathing Method described below, are used in treating cancer, and require the patient to combine the use of two methods of exercise and treatments:

1. The Opening and Closing Methods of Qi regulation (see Chapter 41), and
2. The Deep Relaxed Breathing methods for purgation, rather than tonification purposes.

The purpose of these treatment methods is to purge Excess Qi from the body, remove Blood stasis, and disperse Qi stagnations.

THE DEEP RELAXED BREATHING METHOD

The function of this exercise is to purge the channels and collaterals and to promote the circulation of Qi. It is also used to harmonize the ascension of Yang and the descending of Yin. Calming the Yin and suppressing the Yang promotes the flow of Qi and Blood through the organs, and has the effect of eliminating Evil pathogenic factors, and supporting healthy energy.

The Deep Relaxed Breathing method requires the patient to inhale through the nose and exhale through the mouth. The main points to remember in this method are as follows:

1. The patients inhalation and exhalation should be gentle, thin, even and long. They should focus their mind on the Hibernation Breathing Method. This method requires inhaling and exhaling through every pore on the body's surface while lying supine. The patients should not focus on their lower abdomen or the expansion of the upper chest and

Lung region. Instead, they should imagine inhaling through the pores into the Taiji Pole, and exhaling out from the Taiji Pole through the tissues and out the pores.

2. When breathing, the patients' whole body should stay relaxed; the key part to relaxing the whole body is to relax the waist.

Through training in these respiratory methods, the patients' immune system is enhanced and their vital Qi is nourished.

CHAPTER 17
MENTAL DAO YIN TRAINING

INTRODUCTION

There are two divisions of the Mind (Yin and Yang), each consisting of three different levels. The Yin part belongs to the energy of the Earth and is a more body-oriented type of mind. The Yang part of the Mind belongs to the energy of Heaven and is a more consciousness-oriented type of mind. The three levels of the Mind are as follows.

1. The first and deepest level exists within the Jing and the Shen, and represents the Original (Yuan) and primordial form of the mind.
2. The second or middle level exists within the Seven Corporeal Souls (Po) and Three Ethereal Souls (Hun) and represent the moving and active aspects of the mind (e.g., body movements, reflexes, instincts, drives, and astral projection - when accompanied by the Yuan Shen).
3. The third and superficial level exists within the Will (Zhi) and Intention (Yi) and represents the mind's everyday function (i.e., cognitive thinking). The Zhi is considered the "thinking body," while the Yi is considered the "thinking mind" (for more information refer to the Five Yin Organ components of Shen in Chapter 13).

These different levels of energies create the body, as well as activate the mental and emotional transitional states of the Mind. All three levels of the Mind are interactive and interdependent (Figure 17.1).

Energetic Division (Qi)	**Yin**	**Yang**
Energy of the Mind	Earth	Heaven
First Level of the Mind	Essence (Jing) Kidneys	Spirit (Shen) Heart
Middle Level of the Mind	Corporeal Soul (Po) Lungs	Ethereal Soul (Hun) Liver
Third Level of the Mind	Will (Zhi) Kidneys	Intention (Yi) Spleen

Figure 17.1. Mental and Emotional Transitional States of the Mind

Traditional Chinese Medicine believes that the Heart stores the Spirit (Shen). The Shen is further divided into Yuan Shen (the intuitive congenital spirit) and Zhi Shen (the analytical acquired spirit). The Shen is responsible for feeling and controlling the body's mental and emotional activities. The Zhi Shen is generally strong and can be stubborn and suspicious, therefore, it must be led by a master; the Yuan Shen must be that master, and must lead the Heart's Shen as one unit.

The brain is considered the house of the Yuan Shen, and the seat of mental function. It is believed that the senses and control of the body's physical movements are related to the brain. If the Heart has any functional disorder, the mental activities of the brain, and the emotional activities of the Heart also become disorganized. The Heart, which is the seat of all emotions, controls not only the Shen, but also the blood vessels, therefore Blood circulation is also influenced by the emotions.

The preservation of Jing depends on the cultivation of Qi, which in turn depends on the mental focus of the Shen. The mental focus of the Shen is to Qi as a mother is to her child. Mind concentration must combine with Qi; without the Shen

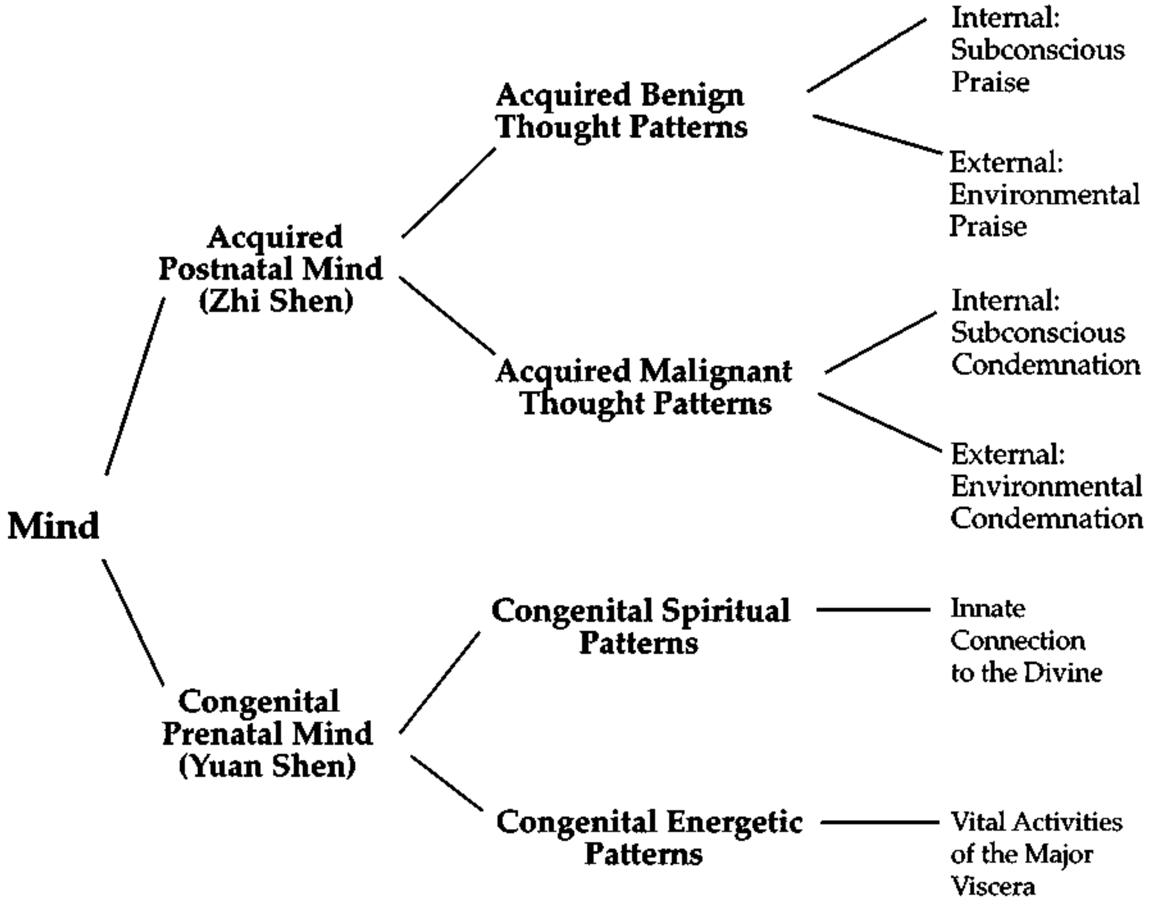
Mind
Acquired Postnatal Mind (Zhi Shen)
Congenital Prenatal Mind (Yuan Shen)
Acquired Benign Thought Patterns
Acquired Malignant Thought Patterns
Congenital Spiritual Patterns
Congenital Energetic Patterns
Internal: Subconscious Praise
External: Environmental Praise
Internal: Subconscious Condemnation
External: Environmental Condemnation
Innate Connection to the Divine
Vital Activities of the Major Viscera

Figure 17.2. The Differentiation Between the Prenatal and Postnatal Patterns of the Mind

guiding it, the body consumes the energy. The concentration of the Shen and Yi (intention) causes Qi to consolidate, while any distraction to the Shen and Yi causes Qi to disperse.

This concept is considered similar to that of regulating the emission of light, because Qi is emitted similar to the wave patterns of rippling water, and Shen is emitted similar to the wave patterns of light. When the Mind begins to focus concentration on the projection of Qi, both Qi and Shen unite, enabling the Qigong doctor to regulate the emitted Qi, changing its form from an expanded energy field to a finely honed beam.

THE PRENATAL AND POSTNATAL MIND

In Traditional Chinese Medicine, it is believed that the cerebral activity of the mind falls into two distinct categories: the Prenatal (Yuan Shen) Mind, and the Postnatal (Zhi Shen) Mind.

1. The Prenatal Mind (Yuan Shen) is inherited from the Jing of both parents and dominates the vital activities of the major viscera, as well as the function of the body's entire energetic organism and spiritual matrix.
2. The Postnatal Mind (Zhi Shen) is developed through interaction with people and the en-

vironment after birth. It dominates thought and interaction, and engages in unlimited mental activity.

The Yuan Shen and Zhi Shen are interactive and interdependent. If the Zhi Shen is active and chattering (sometimes called the Monkey Mind), it is difficult for the Yuan Shen to convey its perceived information. Regulating the Mind is therefore needed to harmonize the analytical and intuitive aspects of the Mind.

BENIGN AND MALIGNANT THOUGHT PATTERNS

Mental activity and thought processes can also be categorized into two clinical syndromes: benign and malignant thought activity (Figure 17.2).

BENIGN EMOTION AND THOUGHT ACTIVITY

This is a gentle or favorable belief structure of the subconscious mind, resulting in relaxed, pleasant thoughts. This thought pattern will both improve and regulate the excitation and inhibitory processes of the cerebral cortex. These positive thoughts and beliefs help to restore strength to the Yin and Yang organs, and thus promote health. Benign emotional thought activity can be further divided into internal and external influences:

1. Internal benign thought activities stem from the inner Hun's influences over the subconscious mind. These positive thought patterns tend to enhance self-esteem, by complementing and encouraging our thinking and actions.
2. External benign thought activities consist of praise and encouragement received from others (parents, friends, co-workers, etc.), which have been accepted and believed. These positive thought patterns promote confidence and self-esteem.

MALIGNANT EMOTION AND THOUGHT ACTIVITY

This is a harmful belief structure causing excited, nervous, stressful, and unpleasant thought patterns that interfere with the performance of the cerebral cortex, resulting in internal organ dysfunction and disease. Malignant emotional thought activity can be further divided into internal and external influences.

1. Internal malignant thought activities are the thoughts stemming from the subconscious mind, which tend to destroy self-esteem, cause obsessions, and discourage productive thoughts and actions. This thought pattern extends from the influences of the Po.
2. External malignant thought activities are internalized negative cynical beliefs and criticisms that originated from the surrounding environment (work, parents, teachers, friends, mate, etc.). External malignant thought activities cause thought distortions that destroy our self-esteem and discourage us.

THE IMPORTANCE OF MIND CONCENTRATION

An essential requirement for training the Qi is the training of mind concentration. If the concentration is forced, the Qi from the Lower Dantian rises up to the head causing the Stomach to become nauseated. If the Qi continues to ascend to the head, the result will be dizziness, headaches, and Qi deviations.

Thoughts and mental activities are reflexes of the brain and cerebral cortex. A Mind (whole body consciousness and awareness) agitated or depressed over a long period of time causes a long-term disorder to the whole nervous system, and leaves the body vulnerable to various kinds of chronic diseases. Long-term anger, for example, leads to the contraction of the blood vessels, sclerosis of the arterioles, gastric ulcers, and hypertension. Long-term emotional disturbances disrupt the endocrine system, and lead to a weakened immune system.

Psycho-neuro immunological research confirms that the Mind (whole body consciousness and awareness) can both cause and cure diseases. Healing visualizations sometimes act as a catalyst, increasing the rate of chemical reaction within the tissues. Improper concentration, however, can transform good Qi into Toxic Qi, and can lead to chemical imbalances that perpetuate the diseased condition. This is especially true when guiding and emitting energy.

According to medical statistics in China, with certain types of cancer, patients whose worry and anxiety were the main cause of their disease made up about 52% of the patients. These patients had been emotionally upset for half a year before becoming sick. The focus of the patients' mental concentration was directed on detrimental thoughts. It is a fact that chronic diseases are closely related to the patients' thoughts, emotions, and mood swings. To treat chronic diseases, it is necessary for the Qigong doctor to first relieve the patients' depression, irritability, and resentment and then place the patients, when possible, in a quiet comfortable environment.

It is important for the patients to relax their Mind (i.e., mind and body). Relaxation can be achieved through tissue manipulation (massage), or by inducing meditations, or trance states. To relax the Mind the Heart must be regulated. To regulate the Heart means to adjust the agitated or depressed Mind and place it into tranquility. One of my teachers in China, Qigong Master Zheng Zhan Ding would often say, "When the spirit, like a lake, is undisturbed, it mirrors the Heavens perfectly." Meaning that, the clarity of the Mind emerges gradually (or slowly) when entering into stillness, much like mud slowly settling in water. This allows ones true inner nature to be revealed.

According to research published by Dr. Karl Lashley, specific memory is not located in any one place in the brain. He found that destroying a portion of the brain does not destroy the memory assigned to that specific area, and also that memory could not be located in specific brain cells. His observation is that memory is distributed all over the brain as an energy field. Other researchers have made similar observations that support the conclusion that the human brain functions as a hologram, collecting and reading information from a holographic universe.

REGULATING THE MIND

Mental Dao Yin training involves regulating the Mind. This requires the Qigong doctor to diminish the mental activities (or judgements) of the Zhi Shen to prevent interference with the intuitive understanding of the Yuan Shen. The diminished mental activity allows for true relaxation, peace, and inner tranquility.

If the Zhi Shen is allowed to become unbridled through excessive internal chatter, it can become injured by the Excess accumulation of energy attributed to the Seven Internal Emotions: anger, worry, joy, fear, fright, sorrow, and grief; as well as the Four Desires that become pathogenic when in Excess: sex, money, fame, and power. These Seven Internal Emotions and Four Desires engage the mind, robbing the body of its life-force energy by depleting the Yin and Yang organs, disrupting their balance, and causing obstruction in Qi and Blood circulation that results in disease.

All thoughts carry within them emotional reactions, these in turn have a significant physiological as well as psychological effect on an individual's health. The degree of intensity as well as the frequency of returning thoughts determines the extent of the internal transitions experienced in the individual's body.

To illustrate this difficulty, when I began my initial Qigong training, my instructor informed me that the internal training would keep me warm in the winter and cool in the summer. Excited about the possibility of possessing my own internal air conditioner, I practiced diligently.

After three years of difficult training I had not yet developed this ability. I slowly began to lose trust in my instructor, and approached him to inquire about the promised skill. He informed me that it was the "imagination" that changes our thoughts and belief structures, enabling the body's tissues to feel what the mind directs. Stating that, "The Imagination leads the mind, the mind leads the Qi." My teacher further stated, "Whenever training in the summer when the heat is unbearable, focus your mind on the coolness of the sweat and imagine the air around your body as a cloud of cold air. In the winter time, when the cold is severe, focus your mind on the heat that your body is producing and imagine the air around yourself as a circle of hot fire." After applying these principles to my training, I experienced both the warmth and cool sensations that I was trying to achieve, and regained trust in my teacher.

Thoughts and emotions create electromagnetic waves in the brain. Your breath can also create or control these waves and cause your mind to take

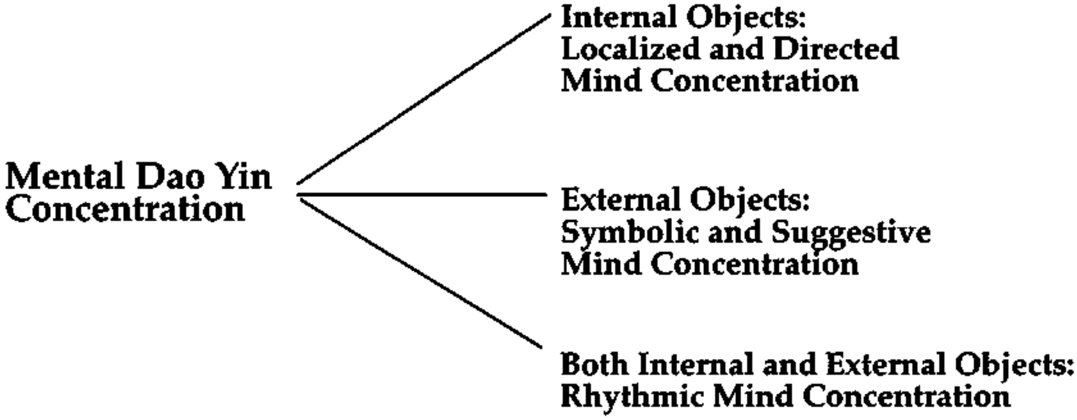
Mental Dao Yin Concentration
Internal Objects: Localized and Directed Mind Concentration
External Objects: Symbolic and Suggestive Mind Concentration
Both Internal and External Objects: Rhythmic Mind Concentration

Figure 17.3. The Three Categories of Mental Concentration

on the thought wave patterns of specific emotions. When the mind is tranquilized after certain Medical Qigong exercises, all disturbances stop, there is a quiet and comfortable feeling sensation (produced by the deep inhibitory state of the cerebral cortex) that facilitates the expansion of consciousness and awareness throughout one's entire being. This state of being is referred to as the Mind.

THE THREE CATEGORIES OF MENTAL CONCENTRATION

Mental Dao Yin training can be summarized by the following three foundational modalities: concentrating on internal objects, concentrating on external objects, and concentrating on both internal and external objects (Figure 17.3).

These three modalities of mental concentration encompass regulation of both the benign and malignant emotional thought activities which influence our everyday lives. People focus their mental concentration everyday on either benign or malignant internal objects (inner thoughts and feelings), external objects (outer thoughts and feelings), or engage in both internal and external thoughts and feelings through conversation.

CONCENTRATING ON INTERNAL OBJECTS

In China, the method of concentrating on internal objects is called "localized" and "directed" mind concentration. In this type of training, the mind is absorbed in concentrating on a specific area of the body. Patients, for example, are directed to focus their eyes downward along the nose in the direction of their navel, while simultaneously using inner-vision to concentrate on their Lower Dantian. Patients can also focus their concentration on other external or internal areas of the body such as:

1. The palms or feet to lead Excess Qi away from the head or torso.
2. A specific channel point (called "energetic point therapy") or pathway, to direct Qi into or away from specific areas of the patient's body, or
3. A Yin or Yang organ to direct Qi into or away from specific internal organs.

The ancient Qigong Master Li Shizhen once said, "When the mind is concentrated upon the inner channels, the practitioner is able to look within his own self." Qigong doctors and patients who are energy sensitive can feel the channels along which their own vital energy flows. This sensitivity is experienced in the more advanced stages of tranquility.

CONCENTRATING ON EXTERNAL OBJECTS

In China, the method of concentrating on external objects is called "symbolic" and "suggestive" mind concentration, and is focused through vivid imagery. In this type of training the patient's mind is absorbed in concentration on some form of object outside of the body. As the mind enters a state of tranquility, it focuses through vivid imagery, or energizes through sound vibrations. Some excellent examples of external objects are: music, the sound of rain or water flowing, a calm ocean, the sky, clouds, light, sound vibrations, mantras, a golden or white glowing ball, specific colors, trees, etc. In this type of concentration, the patient's internal vibration connects with and matches the external form of some object in nature or some specific symbol (such as a lit candle).

When focusing on external objects, it is important for patients with high blood pressure to choose low lying objects, such as a meadow, quiet stream, flower, etc. Patients with low blood pressure should choose objects that are eye level, such as a small pine tree, to balance the energy of the Heart.

When practicing this particular method of mental concentration it is important to focus on external objects that can be physically observed without straining. This way if distracting thoughts interfere with the concentration, the patients can return to the real image, or sound, to reestablish their concentration. The focus should never be forced or too intense. Dr. Xu, of the Xi Yuan Hospital in China, once stated, "In the state of emptiness you will see an image through your imagination. This image will arouse your emotions. These emotions help you to form a more colorful image. From these colorful images you will obtain a deeper state of emptiness. This is the state of Mind sought after in order to obtain emptiness and maintain tranquility." In this particular case the patient's mind is like still water, free from expectations or desires. In quietness you become the observer of images, allowing what will unveil itself to occur naturally.

CONCENTRATING ON BOTH INTERNAL AND EXTERNAL OBJECTS

These methods employ "rhythmic Mind concentration" and are the primary methods used in Medical Qigong therapy. They are simple, easy, safe, and effective. In this type of mental training, the patient's body responds to their thoughts. Patients focus their attention on the breath as it flows in and out of the body, listening to the sounds of their breathing and sensing the expanding and contracting movements of their abdomen. During this type of mental concentration the patients are sometimes given a mantra, such as thinking the word "quiet" during the inhalation and "relax" during the exhalation, or they are instructed to silently count numbers. Sometimes, the patients are also directed to imagine the feeling of quietness, like a mist, flow from their feet into their head (absorbing quiescent Earthly Qi), or from their head down into their feet (absorbing quiescent Heavenly Qi).

When first teaching patients the methods of mind concentration, keep the images simple. When the patients have recovered from their illness, they can begin to concentrate on either an internal or external object to further strengthen their health.

ESSENTIALS OF TRAINING MIND CONCENTRATION

These three mental Dao Yin techniques can be utilized to their full potential once the Qigong doctor understands the essentials of training the mind, the three stages of mental tranquility, and the skill of regulating the mind.

Mental activities should be coordinated naturally with respiration and posture. In Dynamic Qigong Dao Yin exercises, for example, mental activities must be adapted to the posture and the lifting, opening, and closing manipulations of the hands. When using an active (Yang) posture (if the goal is to regulate the patients' Qi), the inner concentration should be motionless (Yin). The ultimate goal is to combine both physical and mental activities that promote harmony in the body's energetic fields. The Qigong doctor should monitor their patients' mental concentration, making sure that they comprehend the following three concepts: Relaxing the Mind, Being Self-Confident, and Being Mentally and Emotionally Stable.

1. The purpose of relaxing the mind is to allow all mental activities to be carried out (without straining) in a composed state of mind. The mind should be kept clear and all distracting thoughts expelled gently. During concentration, the mental focus should never be forced. Mental concentration can be either strong (active) or mild (motionless).
2. Self-Confidence is a prerequisite, for without it, the training of mind concentration cannot happen. No matter what kind of mental activity the patients train in, they should be confident that they can reach their goal. This establishes a strong faith, which is the root of all mental Qigong projection abilities. Patients should also be realistic about not expecting quick results. Although miracles do happen, healing often takes time. The dynamic or active postures and exercises train the body's Qi, while quiescent or motionless exercises sustain and nourish the patients' Qi.
3. The purpose of being mentally and emotionally stable is to assist patients in not becoming overjoyed or frightened if something unexpected happens, or is perceived, during the Qigong exercise. It is important that patients take advantage of this time to keep their mind concentrated and the Qi consolidated. If patients feel weary, remind them that sleep and food can reinforce their mind and Qi.

CONCENTRATIVE MEDITATION AND INSIGHT MEDITATION

When Medical Qigong practitioners begin their meditative practices, they are generally encouraged to keep their mind focused and under control, this is known as "concentrative meditation." Concentrated power is manifested through silence; when the mind is noisy (distracted by physical sensations and chatty), the internal power becomes diffused. When Qigong doctors reach the place of silence in their Mind, they can connect with the divine power, a place of true power, and a place where "all is one." Through concentrated intention, Qigong doctors can bring all of their energies to bear on one point of focused power or force. At the same time, they contact and connect with the divine in silence (the hookup), and become one with the divine healing power.

This transition and conscious union with the divine occurs only when the Qigong doctors turn from the chaotic external world of sensory input (sight, hearing, smelling, etc.), to the quiescent state of silence. This divine energy is then manifested as power, substance and true intelligence, flowing through the Qigong doctors for healing.

Contact with the divine is only obtained in the deepest part of silence. Within oneself, and within this silence abides a state, or dimension, where the true "Gate of Heaven" can be opened and enlightenment obtained. Ideas conceived in this divine state of consciousness come into expression and can manifest as form. All Jing, Qi, and Shen can be transformed and transmuted into form, through this change in consciousness. When the doctor's Intention (Yi) and Will (Zhi) are aligned with the Dao (or divine will) all things are possible. The doctor is in direct contact with the universal Mind, and can manifest in physical form, that which is needed through faith. The divine energy resides within the body's Eternal Soul as power, substance, and intelligence and is brought into form and expression through consciousness. The degree to which the doctor can tap into and utilize the consciousness imbedded in the infinite Mind of the divine, is determined by the concept, or belief, that is held in the doctor's as well as the patient's conscious and subconscious mind.

In the silent state of quiescence, when the Qigong doctor sees a mental pattern or mold, a template is created into which will flow the substance (Qi) needed to bring it into being. The pattern will manifest into the form intended by the doctor's consciousness. Through the power or process of faith and thought, the doctor (or patient) can transmute and evolve the body or outer conditions and surroundings, by recognizing the divine consciousness within themselves.

As the practitioners advance in ability, they are encouraged to focus their attention indiscriminately on all sensual stimuli they receive, this is known as "insight meditation." In practicing insight meditation, practitioners are encouraged to reevaluate their experiences and world views according to the cause and effect of their disease(s). Patients are also taught to use this type of insight to observe their physical and mental processes. This self-observer is used as a vehicle through which to observe, scrutinize, and carefully examine the fundamental energy transformations within the body for spiritual growth.

THE THREE STAGES OF MENTAL TRANQUILITY

Before the mind surrenders to a state of tranquility, it must first transition through three stages. The Qigong doctor must experience all three of these stages in his or her effort to become effective when treating patients. In China, Qigong masters have a saying, "When you root the Mind, the Heart will open up to ten thousand voices." This means that when the Mind is removed from all distractions and excess chatter, the Yuan Shen is free to receive intuitive perceptions. This state of "tranquility" or "stillness" also dissolves time perception, allowing time to disperse into nothingness (Wuji). The three stages of mental tranquility are described as follows:

STAGE 1

In the first stage, the doctor's four extremities and Lower Dantian may experience heat and tingling as the mind slowly eliminates distracting thoughts and begins to withdraw from the external world. The doctor's mental focus and concentration, however, is not consistent. While emitting Qi, distracting thoughts still emerge, and the doctor must patiently wait for the chattering mind to subside.

In this stage of training, especially in male Qigong doctors, the cognitive, analytical thinking and speaking may sometimes become sporadically impeded. This is due to the thin energetic field within the corpus callosum that separates the right (intuitive) from the left (analytical) hemispheres of the brain. This makes it more difficult for men to access both hemispheres simultaneously. Men should therefore train to rely more on their intuitive perceptions instead of analyzing (they must feel the energy, and not think it).

This however, is not a problem in female Qigong doctors, as the energetic field which separates the corpus callosum is much wider and already active. This allows female Qigong doctors to easily access both sides of their brain simultaneously, enabling them to draw upon the perceptive, intuitive as well as cognitive, analytical thinking and speaking hemispheres of their brain.

STAGE 2

In the second stage, there is an increase in heat to all four of the doctor's extremities from the Lower Dantian. Through intention, the connection to the Lower Dantian is strengthened and the body's Qi feels like a rushing wind flowing throughout the doctor's torso.

While diagnosing and treating, the doctor experiences one or several of the patient's "eight manifestations of Qi" (also called the Eight Energetic Touches). These manifestations consists of the following physical and sensory phenomena of energy: shaking, rippling, heat, coldness, sinking, floating, itching, and heaviness. As the doctor focuses on extending energy, he or she will be able to remove internal distracting thoughts, but will still hear the distracting sounds from the external world.

STAGE 3

In the third stage, all of the doctor's sensory input (seeing, hearing, tasting, touching, smelling, and perceiving) is sealed from outside distractions. The doctor's Yuan Shen and Zhi Shen are now relaxed and in a state of mental tranquility. The Qigong doctor can now approach the patient free of chattering thoughts, judgments and all outside distractions. This enables the doctor to intuitively perceive and sense the patient's energetic traumas, stagnations, deviations, and energetic clusters.

THE THREE PROCESSES OF THE MIND

The mind can be divided into three separate but interdependent processes, which are responsible for our everyday actions, responses and health (Figure 17.4). Thoughts, ideas, and feelings are constantly being implanted, influencing and affecting the spiritual, emotional, mental, energetic, and physical process of healing. The heal-

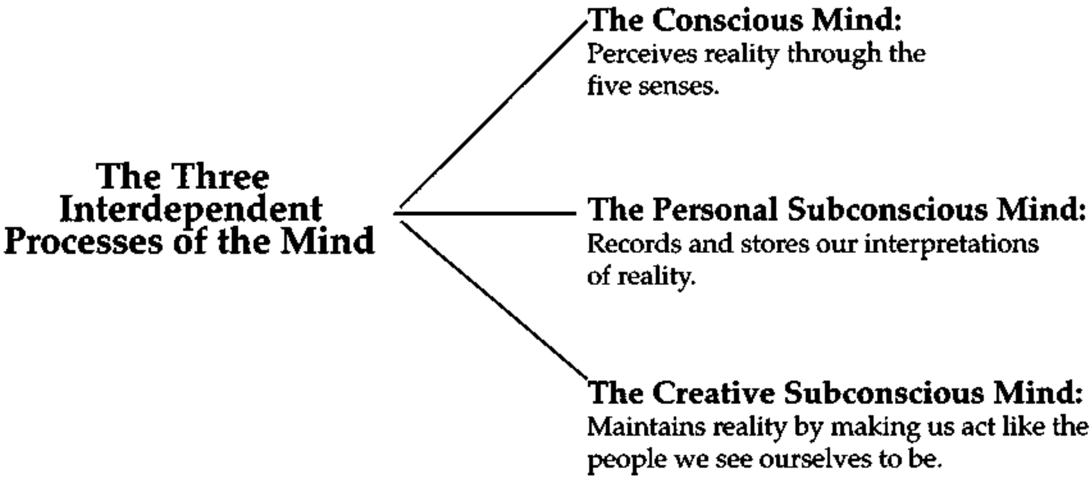

The Three Interdependent Processes of the Mind

The Conscious Mind:
Perceives reality through the five senses.

The Personal Subconscious Mind:
Records and stores our interpretations of reality.

The Creative Subconscious Mind:
Maintains reality by making us act like the people we see ourselves to be.

Figure 17.4. The three interdependent processes of the Mind, which are responsible for our everyday actions.

ing process includes reprogramming and regulating the conscious, personal subconscious and creative subconscious mind.

1. The Conscious Mind perceives reality through the five ordinary senses. It evaluates reality by investigating and interpreting the data through the filter of the past experiences. The conscious mind makes decisions based on personal goals, and manifests conscious thinking patterns. It is considered to be the Postnatal Mind (related to the Zhi Shen).
2. The Personal Subconscious Mind records and stores the conscious mind's interpretations of reality. It consists of the personal, perceptual data that were received through our five senses. The personal subconscious mind shares responsibility for how we think, what we say and imagine about ourselves and our emotional and behavioral reactions to those experiences. It is also responsible for the automatic function of the living mechanism (i.e., heartbeat, breathing, digestion, etc.) through its energetic connection to the autonomic nervous system. It is considered to be influenced by the spiritual aspects of the Wu Jing Shen (especially the Hun and Po).
3. The Creative Subconscious Mind maintains our reality by making us act like the people we see ourselves to be. It solves problems and provides the drive and energy to succeed or fail, heal or become sick. It is the center core energy of our innate spirit or soul consciousness. It is considered to be influenced by the spiritual aspects of the Yuan Shen.

CREATING THE BRIDGE OF LIGHT

The Bridge of Light is an energetically patterned wave frequency which consists of three main components of the body's energetic field. Its energetic connection to the body's tissues functionally envelops the conscious, personal subconscious and creative subconscious mind, and is considered to be the bridge between the three process of the mind. It consists of three separate but intertwining threads: the thread of life, the thread of consciousness, and the thread of creativity.

1. The thread of life comes directly from the Eternal Soul and is rooted in the Heart during conception. It is connected to all feelings that stem from the divine higher energy field and is considered the "seat of life."
2. The thread of consciousness also comes directly from the Eternal Soul and is rooted in the pineal gland within the brain. It embod-

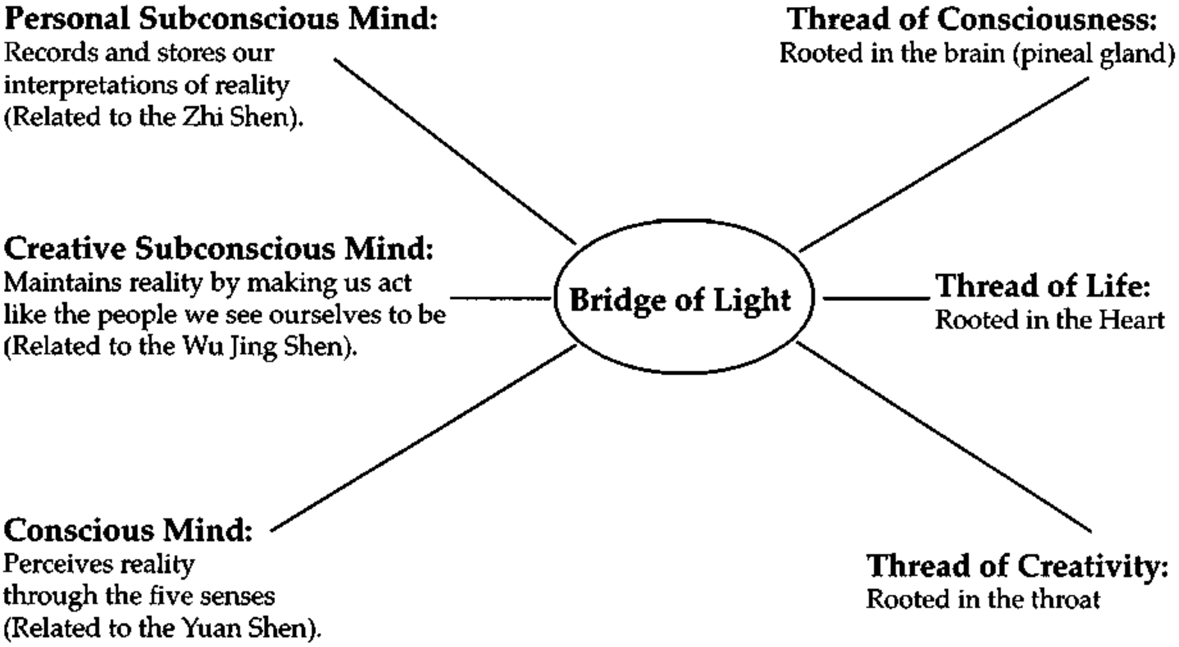
Personal Subconscious Mind:
Records and stores our interpretations of reality (Related to the Zhi Shen).
Thread of Consciousness:
Rooted in the brain (pineal gland)
Creative Subconscious Mind:
Maintains reality by making us act like the people we see ourselves to be (Related to the Wu Jing Shen).
Bridge of Light
Thread of Life:
Rooted in the Heart
Conscious Mind:
Perceives reality through the five senses (Related to the Yuan Shen).
Thread of Creativity:
Rooted in the throat

Figure 17.5. The Bridge of Light

ies portions of the energies of consciousness and is considered to be the "seat of consciousness."

3. The thread of creativity is rooted in the throat and is unique in that it is created and specifically constructed by each individual. It is also an extension, or synthesis, of the thread of life and thread of consciousness.

   The thread of creativity is in itself made up of three main components. These three components, or additional threads, intertwine as one unified thread of creativity and are explained as follows.

   a. The first portion of this thread extends from the physical Heart to the Spleen, and is connected from the body to the first external field of Wei Qi.
   b. The second portion of this thread extends from the solar plexus to the Heart and eventually extends and connects from the body to the second external field of Wei Qi.
   c. The third portion of this thread extends from the Yintang (Third Eye) through the brain, ascending through the Baihui region and beyond. This energy field is connected from the body to the third external field of Wei Qi.

After all three threads of creativity have been activated, energized, and developed, the next goal is to align and integrate the Qi of the thread of life and the thread of consciousness, with the frequencies of the thread of creativity.

Because the thread of life is connected to the Eternal Soul, the integration of the three threads allows an individual the direct connection to his or her core self (Figure 17.5). This enables the individual the ability to consciously access his or her true inner voice (the voice of his or her soul) and receive information and guidance from the divine. The goal in accessing the Bridge of Light is to unite these three energetic threads into one powerful and functional energy field.

When all three major threads (which compose the Bridge of Light) are connected as one harmonious cord of light (extending upward through the Taiji Pole), a spiritual gateway is opened that enables access to the Pure Heavenly Sound (also known as Hearing the Sounds of the Universe) (see the Six Transportations of Shen in Chapter 13).

REPROGRAMMING THE MIND TO HEAL

The mind holds specific energetic patterns which maintain a patient's physical, mental, emotional, and spiritual belief structures. These patterns are maintained by the dictates of the personal subconscious mind and can only be changed when the creative subconscious mind is engaged.

To change specific destructive patterns, the personal subconscious mind must be reprogrammed. There are three things that are needed in order to reprogram the personal subconscious mind: imagination, visualization, and positive affirmation. Imagination combines with vivid pictures, sounds, sensations, etc., and creates a new reality on a creative subconscious level. It is the awakening to this new reality that reconnects the individual to the creative subconscious mind.

If patients do not use imagery and affirmations to change their present personal subconscious representation of reality, the patient's personal subconscious mind will automatically seek to correct any newly made change. The personal subconscious mind views any new deviation from the normal energetic pattern as a mistake, and will automatically resist, or sabotage, any change of pattern, returning the patient back to his or her original state of mind and body sickness.

It is a simple fact that "we bring about what we think about," and act in accordance with the truth that we have come to believe. If we act out of a particular new state of mind and feeling over a sufficient time period, it will become a permanent reality of the self. In using the mind to make transitions, it is important to note that words, images, and sensations have tremendous power, and the spirit behind the intent is the key to any permanent transformation.

Because patients trust their doctor, it is the doctor's responsibility to direct the patients into a healing mental state for the purpose of restoring health. The doctor must never be guilty of "clinical hexing." Clinical hexing occurs when a patient (who completely trusts the doctor) is told by that doctor, that there is no hope. The patients, believing the doctor, may give up their desire to change and heal. Thus, through accepting the inevitability of death, they suppress their desire to live. Qigong doctors are therefore taught to always encourage their patients on their recovery, and to never destroy their hope.

Through positive affirmation and reinforcement, it is possible for the patients to turn around the progression of their disease. Scientific studies in China have proven, for example, that through imagination, visualization (which include using all of the five senses) and positive affirmation, the rate of phagocytosis (the ability of the immune cells to engulf and destroy invading germ cells) greatly increases and enhances the body's immune function.

USING THE IMAGINATION

The Yi (which consists of the imagination and intention) leads the Shen (composed of the thought, emotion and spirit) which in turn leads the Qi (the body's life-force energy). The power of the imagination has a strong influence on health, and can be used to help cure diseases, or to further complicate the condition. If, for example, patients imagine that their immune cells are effectively destroying the cancer cells, Western medical treatments such as chemotherapy or radiation can become more effective. Research from China reveals that patients with a positive belief structure produce certain physical transformations, which manifest in the increased production and release of hormones and immune cells. The imagination is the inner form of all things, the primal inspiration from which reality later manifests. In a sense, patients bring about what they think about. While treating disease, Qigong doctors encourage their patients to imagine healing energy from the divine being directed into the diseased area. Once the treatment is finished, the patients are instructed to imagine that their disease has become completely healed.

During the state of relaxed meditation, the foundational aspect of the patients' disease, which is normally hidden from consciousness, reveals itself in the form of images and impressions. After a Medical Qigong treatment, or meditation, the patients may describe to the doctor any unusual thoughts, images, or feelings that they may have experienced while in the deep state of relaxation. The doctor then helps the patients find the origin of the image (a past incident, unhealthy belief structure, or trauma, etc.) through the doctor's ability to trace the images and emotions associated with it to the patients' internal organ or organ systems involved in the creation of the illness. Once the connection is made to the patients' disease, and the Medical Qigong homework is prescribed.

Medical Qigong prescriptions initiate energy movement. This energetic movement brings about the release of trapped emotion and energetic feelings. As the emotions unravel, a new level of awareness surfaces that allows the patients to become cognizant of certain mental functions that were previously unconscious. This awareness in turn allows patients to access their creative subconscious Mind to change the conditioned programming of their Zhi Shen (Figure 17.6).

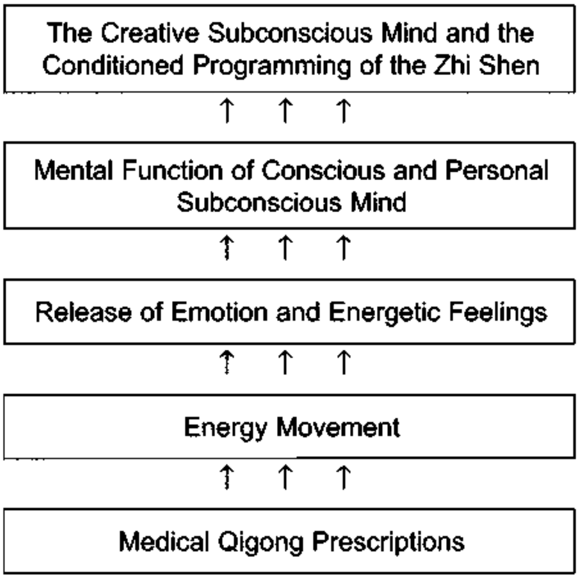
The Creative Subconscious Mind and the Conditioned Programming of the Zhi Shen
Mental Function of Conscious and Personal Subconscious Mind
Release of Emotion and Energetic Feelings
Energy Movement
Medical Qigong Prescriptions

Figure 17.6. Five stages of transition and energy awareness allow patients to access their creative subconscious energy in order to reprogram their initial belief structures.

Through creative imagination, the patients' spirit is expressed and the patients' body is spiritualized. The imagination is an energetic world founded on the interrelation of various levels of spiritual, imaginative and physical realities. These energetic dimensions lie between what the Chinese call the "realm of the unfathomable hidden mystery" and the "world of animated physical forms." The imagination is an energetic and spiritual world that exists between the reality within oneself, and the reality outside of oneself. This energetic and spiritual world seems imaginary, but in fact it is very real, in that the patients live in it, and through it create a new belief system and self.

TWELVE STAGES TO TRANSCEND AND TRANSFORM

During the Ming Dynasty there began to surface twelve specific poems accompanied with twelve pictures, describing the twelve stages to enlightenment through contemplation. In this series there are ten pictures which describe the first ten stages of mental, emotional, and spiritual transcendence, and two additional pictures which describe the last two stages of mental, emotional, and spiritual transformation. These pictures and poems use the image of the ox (water buffalo) to describe man's animal nature, which in the first stage of spiritual training is identical to man's spiritual nature. The image of a young man is used to depict human intention and will. The interaction of the ox and young man reflects the idea that by coming to terms with our own animal nature, we can move toward transformation and arrive at a new way of being.

The images of the ox changing from dark, Turbid Yin, to bright, white Yang, represent the field of consciousness transforming through the controlled deliberate intention of one's Zhi (Will). Such metaphors as capturing and controlling the ox represent an explanation of the difficulties and dangers of transforming one's self, as well as the inner changes of alchemical purification and the reconciliation with one's darker side. These pictures suggest the combination of both the sacred and the natural within Man.

These stages of controlling the mind are described as follows.

1. In the wild, the first stage of training is divided into three separate phases:
    a. The Searching for the Ox phase depicts the state in which the young man is unaware of his own true nature. Although he is not sure what he is seeking, he has begun searching for something he is desperately missing within his life. He is in a state of omnidirectional searching, without an instinctive direction.
    b. The Finding the Tracks phase depicts the young man undergoing the first traces of mental, emotional and spiritual transformation. Realizing that other people have walked along this path, the young man begins to rediscover his own spiritual instincts. He is in the beginning phase of achieving focus and direction.
    c. The First Glimpse of the Ox phase depicts the young man experiencing his first "awakening." He has the first glimpse of the true union of nature and spirit in the form of a spiritual vision. The animal nature, however, is still uncontrollable and in the wild state (due to its connection to the painful memories of the past). The young man must chase after and entice the ox, which displays its horns, bellows aloud, runs away, and is overshadowed by a dark cloud (anger, despair, grief and fear). It tramples the wheat seedlings (beginning moments of spiritual insight) wherever it goes. In this phase the ox is unruly, Yin, and pure black in color (Figure 17.7).
2. The Initial Training Begins with this next stage. The ox is controlled by a rope through its nose, and runs swiftly under the young man's whip. The young man struggles hard to control the ox's willful temper. He begins talking and listening to the ox, matching its consciousness with his own. At this stage the ox is still Yin, and pure black (Figure 17.8).
3. The Ox Comes Under Control through con-

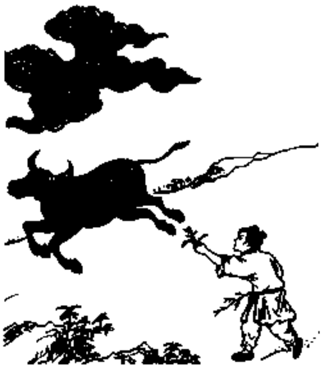
Figure 17.7. The Ox is in the Wild

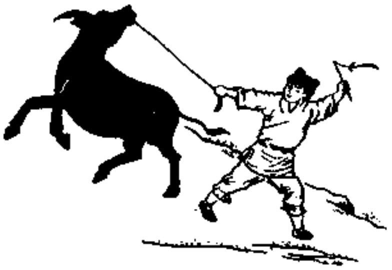
Figure 17.8. The Initial Training Begins

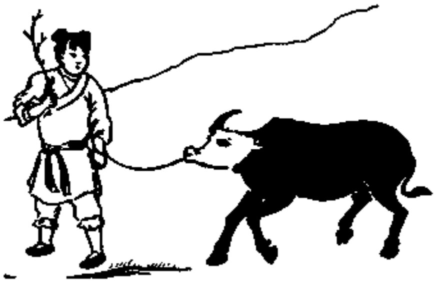
Figure 17.9. The Ox Comes Under Control

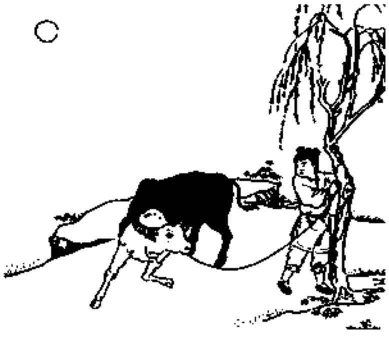
Figure 17.10. A Time of Transition

stant training. The ox stops its dashing about, and begins to follow the young man. In spite of his fatigue, the young man tends the ox all day long, not daring to loosen his control of the whip or his grip on the rope. Gaining control of the ox requires the young man to also become more vulnerable and honest, using less force to control the situation. In this stage only the ox's head, which is under control, is Yang and white, the rest of his body is Yin and black (Figure 17.9).

4. During the Time of Transition, the young man feeds and takes care of the ox with respect. He also does not abuse it. At this stage, the natural strength of the ox harmonizes with the will of the young man. A long time has passed and the restless temper of the ox gradually has turned gentle. However, still not trusting the ox, the young man maintains control of the rope. The ox is one third Yang and white and two thirds Yin and black (Figure 17.10).
5. The Ox Is Now Tamed and moves in harmony with nature. It follows the young man, who no longer uses a rope to lead the ox, but still maintains control of the whip. The ox is half Yang-white and half Yin-black (Figure 17.11).
6. The Ox Is Freed of Worldly Hindrance. The ox no longer needs the whip for control. The young man can begin to experience and enjoy the inherent rhythm of life, because he now has the ox's undivided attention. The ox is three quarters Yang-white and one quarter Yin-black (Figure 17.12).
7. The Ox Is Under Complete Control. After a long stage of struggling against its instincts, the ox is now tamed and free from worldly distractions. It drinks when it is thirsty and eats when it is hungry. The young man can now enter into deep sleep. If he so desires, he can sit atop the ox and play his flute, because the ox knows where its going and the young man does not have to direct it. The ox is now completely Yang-white (Figure 17.13).
8. The Young Man and the Ox Unite With Heaven. The white ox wanders among the heavenly clouds. The young man is free from

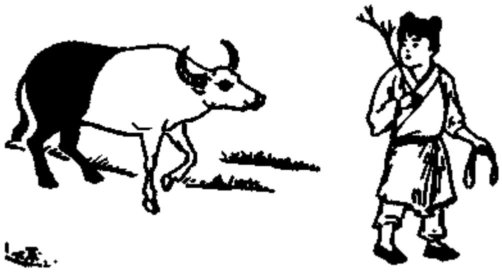

Figure 17.11. The Ox Is Now Tamed

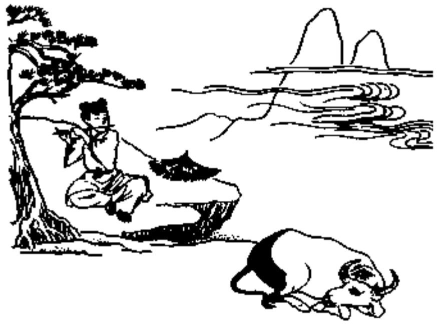

Figure 17.12. The Ox Is Freed of Worldly Hindrance

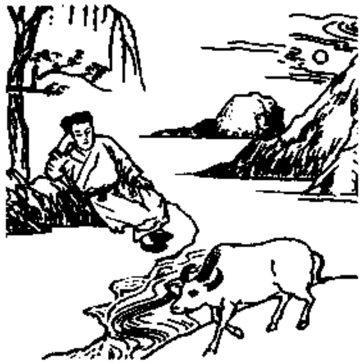

Figure 17.13. The Ox Is Under Complete Control

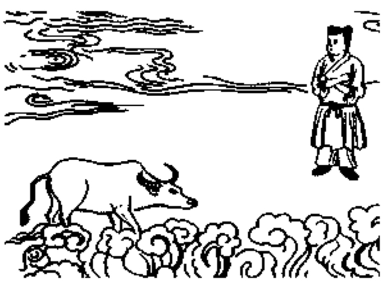

Figure 17.14. The Young Man and the Ox Unite With Heaven

worldly concerns, and so is the ox. Penetrated by moonlight, the Heavenly clouds grow whiter and drift away (Figure 17.14).

9. The Ox Transcends and A Single Light Remains. The clouds drift away and the ox vanishes. The young man finally enjoys his leisure time, singing under the moon (Figure 17.15).
10. Both the Ox and Young man Transcend and Return to the state of Wuji. As the moon continues to illuminate the vast void, both the young man and the ox vanish into Wuji and are nowhere to be found. Everything is formless, quiet, and pure. Existence is nonexistence and vice versa (Figure 17.16).
11. The Young Man and Ox Transform and Return to the Source. There is a difference between transcendence and transformation. Transcendence is a temporary or periodic experience of unity within the divine, the heavens and the earthly environment. It is a state, or condition of consciousness. Whereas transcendence implies a unifying experience that informs one of a new way of being, transformation is a process and a means to arrive at a new way of being. Mature transformation leads to an all-inclusive way of being that embraces the physical, mental, emotional, energetic, and spiritual existence of everything. In this stage man returns back to the original source of prenatal understanding, and is able to harmonize with Heaven and Earth (Figure 17.17).
12. The Young Man and Ox Transform and Live in the World. This stage depicts what happens after the transformation experience, and the reentry through nature into the everyday world. In this stage, man reenters the world as a transformed person with a new connection to his center core. He is able to become a guiding light to others. When needed, he can instantly return to a deeper state of enlightenment for additional instruction and personal guidance. Living in the world he can be in the seductions and temptations of the world, without being in any way affected by them (Figure 17.18).

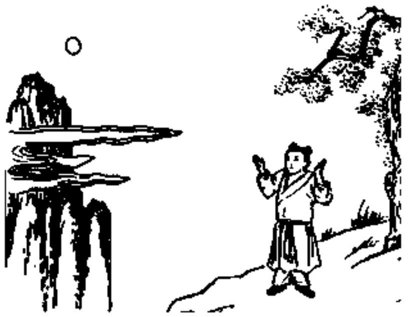

Figure 17.15. The Ox Transcends and A Single Light Remains

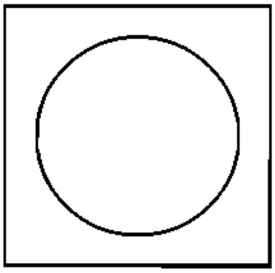

Figure 17.16. Both the Ox and Young Man Transcend and Return to the State of Wuji

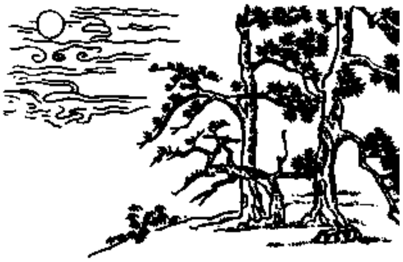

Figure 17.17. The Young Man and Ox Transform and Return to the Source

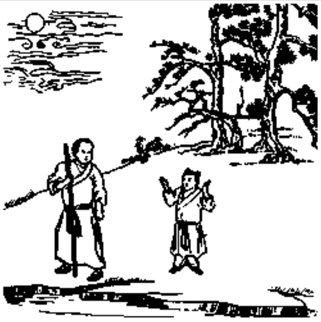

Figure 17.18. Transformation and Living in the World

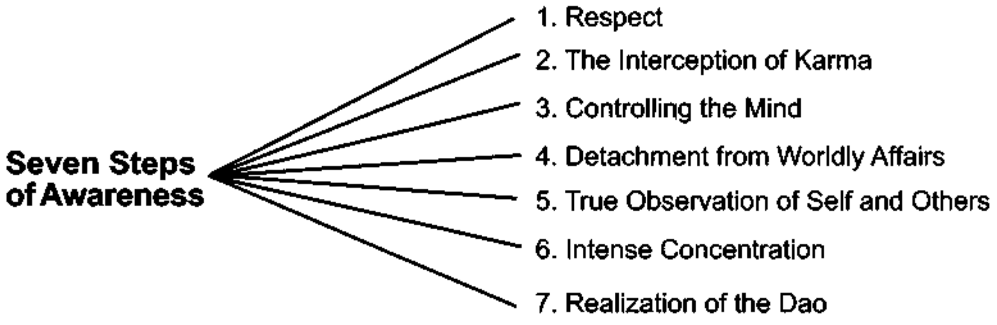
Seven Steps of Awareness
1. Respect
2. The Interception of Karma
3. Controlling the Mind
4. Detachment from Worldly Affairs
5. True Observation of Self and Others
6. Intense Concentration
7. Realization of the Dao

Figure 17.19. The Seven Steps Needed to Transform the Mind

SUMMARY

During mental, emotional and spiritual transformations, individuals undergo stages of physical and psychological purification, as they begin to experience the active, conscious introspection into their mind, body and soul. These individuals are able to eventually identify and connect their own personal will to that of the divine, and begin to lose their attachment to the ego, developing a new and more expansive identity. These individuals will also begin to see themselves as truly spiritual beings that are merely housed in fragile, physical frameworks, which are subject to all the transformations that their spirit must transcend. Generally, when using Medical Qigong as a vehicle for transformation, individuals will evolve through seven steps to awaken the spirit and reach a stage of enlightenment: Respect, The Interception of Karma, Controlling the Mind, Detachment from Worldly Affairs, True Observation of Self and Others, Intense Concentration and Realization of the Dao (Figure 17.19).

1. The first stage, Respect, encompasses the self, others and the divine. Through the cultivation of respect a stronger faith develops.
2. During the second stage, The Interception of Karma, individuals begin to accept accountability for all their actions, deeds, and thoughts. At this stage individuals receive instant karma for all their actions.
3. During the third stage, Controlling the Mind, individuals begin to discipline their conscious actions (Taming the Ox).
4. The fourth stage begins the process of Detachment from Worldly Affairs. At this stage individuals begin to develop and maintain a spiritual priority.
5. The fifth stage of True Observation of Self and Others, enables individuals to perceive themselves and others clearly. At this stage they begin perceiving the universe and environment from a deeper, more spiritually evolved understanding.
6. The sixth stage involves the development of Intense Concentration. At this stage individuals transform thought, through faith, into manifestation. They can now initiate into reality what they focus their intention on.
7. In the final stage to transformation, Realizing the Dao, individuals become aware of their relationship with the divine, not only believing in God, but personally knowing God or the Dao.

SECTION IV

QI DEVIATIONS IN QIGONG TRAINING

CHAPTER 18
INTRODUCTION TO QI DEVIATIONS

Energetic deviations occur naturally through the course of everyday life, however, the body's energetic constitution quickly corrects any disharmony in energy flow. The difficulty when cultivating Qi is not in the development of large amounts of energy, but in the creation of a constitutional system strong enough to use this energy without being damaged by it. This is one reason why Qigong doctors initially spend a lot of time developing the safety measures and internal reservoirs necessary to ensure that their body will not be damaged by creating too much energy too fast. These alterations of energetic patterns can result in feelings that can range from simple discomfort to an abnormal susceptibility to disease. Through proper Medical Qigong therapy, Qi deviations can be rectified.

Qi deviations can be precursors to disease. In recent years, medical researchers in China have reported that some patients admitted to the Medical Qigong clinics had developed mental as well as physical disorders as a result of Qi deviations. A deviation refers to abnormal phenomena occurring within the patient or Qigong practitioner. The clinical symptoms include abnormalities in perception, thinking, and behavior.

If the Qigong doctor has not been adequately trained in internal purging techniques, his or her body's increased energetic field can act as a magnet and begin to attract the toxic energy from the surrounding environment (including the patient's pathogenic Qi). Once toxic energy enters into the body, it can immediately create a Qi deviation. Qi deviations should be treated immediately if possible. If Qi deviations occur during Qigong exercises or meditations, the practitioner should stop the practice immediately and guide the Turbid Qi (and bad feelings) out from the body, discharging the toxic energy out through the four extremities and into the ground.

The process of Medical Qigong meditations and exercises is said to cause an alchemic transformation of Shi Shen (Turbid Spirit) into Yuan Shen (Original Spirit). This transformation occurs when first beginning Qigong practice, at the "laying a foundation" stage of the workout. If the Turbid Qi is not removed at the beginning of practice, the Qi concentrated in the Dantian areas (or other parts of the body) will also become Turbid. This may bring about internal disturbances when the practitioner encounters unfavorable external factors such as suddenly being startled or becoming angry. Too much concentration and hard focus on the exercises and meditations can also lead to Qi deviations. Turbid Qi can create Turbid Shen.

When Turbid Qi travels along the Governing Vessel to the head, there is a feeling of heavy pressure in the head. In severe cases, psychosis can occur. Turbid Qi that escapes into the body's channels can result in distending or numbing of the body. To avoid such side effects, it is extremely important to establish an equilibrium between the Five Yin Organs. No matter which Medical Qigong system is followed, the primary task is always to dredge or purge the channels of pathogenic Qi, to regulate the Yin and Yang organs and to cultivate the body's Jing, Qi, and Shen.

THE MAIN CAUSES OF QI DEVIATIONS

Mental disorders and personality disorders may exist in some individuals prior to their study of Medical Qigong. Sometimes, individuals with a family history of psychosis and certain other mental disorders (such as bipolar disorder and depressive disorder) may be more predisposed towards these diseases. The practice of Medical Qigong meditations and exercises may trigger the onset of symptoms in predisposed individuals, especially if Qi deviations are not properly addressed.

1. A Week Constitution	From severe disease and weak respiration
2. Improper Dao Yin Qigong Regulation	From forced practice, hastily and indiscriminately training
3. Disbelief and Suspiciousness	From a closed mind, superficial mind, or preoccupied mind
4. Too Much Mental Concentration	From concentrating too hard, resulting in Qi and Blood to stagnation

Figure 18.1. The Main Causes of Qi Deviations

The main causes believed to contribute to Qigong deviations, according to research gathered by Dr. Wei Lin Shen of Shanghai, China, include the person's constitution and pathological state, improper Dao Yin Qigong regulation, disbelief in and distrust of the Qigong exercises and meditations, or too much mental concentration (Figure 18.1).

1. A week constitution can lead to a pathological condition. Patients with a weak constitution should not be required to sit or stand for long periods of time. Patients with severe disease and weak respiration incur Qi deviations if they are allowed to stand too long.
2. Improper Dao Yin Qigong regulation can result in Qi deviations. Qigong deviations may also occur as a result of forced practice (when the state of illness does not allow relaxation), or when individuals hastily and indiscriminately train in the techniques that circulate internal Qi. Lack of proper instruction from a qualified master of Medical Qigong therapy often results in a failure to train in the correct way. The failure to correctly apply the three Dao Yin principles and methods (proper regulation of the body's Jing, breath's Qi, and mind's Shen) in training practice, for example, leads to Qi deviations. Creating new techniques, training recklessly before mastering the primary techniques, or changing the techniques at random can also result in Qi deviations. For this reason the practitioners should be patient when practicing Dao Yin regulations. The goal is to achieve a tranquil, quiet, and empty state.
3. Disbelief, and suspiciousness can induce Qi deviations. Disbelief or an inability to adopt a correct attitude toward Medical Qigong effects (such as extreme suspiciousness), results in mental tendencies that will become deviations over the course of time. Qi deviations can result from: A Closed Mind, A Superficial Mind, and A Preoccupied Mind.
    a. A closed mind, full of fear, pride or bitterness may prevent the patients from connecting with their True Self.
    b. A superficial mind, or lack of commitment and belief in energetic healing, may cause the patient to fail in taking the Medical Qigong prescriptions seriously.
    c. A preoccupied mind may cause the patient's spirit (Hun) to "wander off" while practicing Medical Qigong therapy.
4. Too much mental concentration is the most frequent cause of Qi deviations. The strong focus of mental attention on the flow of Qi can distort its path and lead to Qi deviations. It can cause too much concentrated Qi and Blood to flow into the brain. Concentration on internal objects with too intense a focus can actually cause a stagnation, or Excess of

Qi in the area (e.g., concentrating too hard on a tumor can actually cause it to grow instead of dispersing it). Mild concentration is always preferred when practicing Medical Qigong exercises and meditations. A relaxed mind and body allows the Qi to flow freely.

CORRECTING QI DEVIATIONS

To correct any Qi deviation, it is important to understand the root cause of the deviation. The Medical Qigong doctor should check for the following causes of Qi deviations.

1. Postural Dao Yin deviations result from incorrect postural changes.
2. Respiratory Dao Yin deviations result from improper breathing methods.
3. Mental Dao Yin deviations result from improper mental activities and emotional disturbances. These emotional disturbances (due to the accumulation of toxic emotions within the internal organs) may manifest as spontaneous emotional outbursts.
4. Stressful life activities resulting from an improper balance of work and play, lack of sleep, or an improper diet may likewise cause Qi deviations.
5. Environmental disturbances during Medical Qigong practice such as the phone or doorbell ringing, may startle the practitioner and cause Qi deviations.

To correct the Qi deviation, the Qigong doctor addresses each possible cause separately, beginning with the patient's posture.

POSTURAL DAO YIN DEVIATIONS

To diagnose Qi deviations due to incorrect postural alignment it is important to observe whether:

- relaxation is being achieved,
- the posture is structurally correct to facilitate relaxation, and
- the appropriate tonifying, purging, or regulating technique is being used.

When a deviation occurs, the patient should examine it in accordance with the basic laws of postural alignment. A beginner is more prone to some disorders due to incorrect postures that cause headaches, dizziness, and a stiff neck.

MANIFESTATIONS OF POSTURAL QI DEVIATIONS

If postural deviations occur, first review the Eighteen Rules of Proper Medical Qigong Structure in relationship to the sitting or standing positions, and then review the other Dao Yin Qigong patterns (respiratory and mental Dao Yin training). The following are examples of specific Qi deviations related to improper postural alignment.

1. Eye distention results from excessively staring at external objects during the meditations and exercises. To prevent this type of Qi deviation, encourage the patients to always use a soft focus.
2. A stiff or painful neck is usually caused by forcefully attempting to suspend the head. To prevent this type of Qi deviation, have the patient relax the neck muscles and sink the shoulders.
3. Shoulder and back pain can occur when patients fail to relax the muscles of the shoulders and back, stretch the back excessively, hollow their chest too much, or hold an unfamiliar posture too long. To prevent this type of Qi deviation, have the patients relax and sink their shoulders and imagine the back melting down into the hips and legs.
4. Waist and hip pain (Lumbago) come from several factors.
    a. Tension in the hip muscles causes pain in the hips that extends to the waist.
    b. Twisting the waist without first relaxing the hips can result in both hip and waist pain.
    c. Forcefully twisting the waist may also cause pain.

    To prevent this type of Qi deviation, have the patients relax their waist and hips and sink the energy into the Earth via the feet.
5. Abdomen and lower extremity distention or flatulence results from the forceful pulling-in of the Stomach, or too much concentration of the mind on the lower extremities. To prevent this type of Qi deviation, have the patients relax their abdomen and lower extremities, allowing the Qi to sink naturally.
6. A swollen and painful anus, or dry stool, can

result from raising the anus too forcefully. Also, hemorrhoids can develope from holding the breath too forcefully while focusing on the lower abdominal area during the Abdominal Breathing Holding method. To prevent this type of Qi deviation, have the patients close their anal sphincter softly and relax their buttock muscles.

7. Sore knees are a result of crouching too low, over-bending the knees, or extending them past the toes. To prevent this type of Qi deviation, have the patients keep their knees in proper alignment and guide the pressure of the body weight to flow behind the legs and not in front.
8. Heel pain can be caused by shifting the body's weight excessively onto the heels during walking or standing practice. Heel-ache can also occur if the patients focus too much attention on their heels instead of the center of their feet for balance. Knocking, stomping, or excessively dropping the heels against the ground, as well as sexual Excess, or a Kidney Deficiency can all cause pain in the heels. To prevent this type of Qi deviation, have the patients relax their legs and waist and imagine that their feet are kneading the ground. This squashing and kneading action is used to encourage the free flow of Yin Qi from the Earth into the body.
9. Improper hand movements occur when patients fail to distinguish between Yin and Yang energetic properties, or apply either too much or too little strength in the raising and lowering of the arm movements during Qigong practice. Both the alignment and energetic function of the patients' arms, hands, and fingers should be checked.
    a. If the arm movements are too dynamic, they will cause an abundant amount of energy to become either gathered or dispersed; this results in Qi deviations.
    b. While performing a tonification exercise, if the palms of the hands are turned outward, away from the body (instead of inside facing the body) this can easily cause diarrhea or a Qi Deficiency.
    c. If the fingers move up to point at the chest, the energy released from the hands can cause tightness in the patients' chest. If the patients' hands point obliquely at their neck and face, the emitted Qi can cause dizziness, nausea, and edema of the face in those patients who are sensitive to the energy movement within their channels and collaterals.

RESPIRATORY DAO YIN DEVIATION

Deviations in breathing often result from the incorrect practice of the deep inhalation and exhalation methods, when first starting Qigong meditations. Both inhaling and exhaling should be gentle, thin, even, and long. When practicing, the patient should breathe properly and naturally to regulate the breath and should avoid holding the breath.

1. Deviations can be due to improper exhalation. Determine whether the exhalation is long enough for relaxation and quiescence when sinking the breath. Excessive sinking of the breath, however, can cause pain in the legs by depleting Qi flow to the upper torso, and by increasing the weight and gravitational pressure within the tissues of the lower torso. Exhalations that are too long or too deep can cause shortness of breath, headaches, tightness in the chest, discomfort in the Heart, and abdominal distention.
2. Deviations can be due to improper inhalation. Determine whether the inhalation is long enough for relaxation and quiescence. For example, when practicing the Abdominal Breath Holding technique (the Two Inhaling and One Exhaling Method, which employs Reverse Breathing), the two inhalations should not be connected, for tightness in the chest will occur. Pause, therefore, between each breath and allow the exhalation to be fast and short rather than sustained and drawn out. Otherwise, the condition will cause general weakness and fatigue.

To rectify this condition, when you inhale twice through the nose, simultaneously pull in the navel and perineum area as if to lift your body. When you exhale through the nose, guide the Qi down into the Middle Dantian.

If this method is applied, the symptoms such as tightness in the chest, shortness of breath, and abdominal distention will disappear (The Rectifying Respiratory Qi Deviations exercise can also be used). When the breath is well regulated there will be no difficulty sinking the Qi down to the Lower Dantian.

3. Deviations can be due to improper tone resonation. Another Respiratory and Mental Dao Yin deviation can be created through the improper use of healing words or tone resonations. The improper use of tones (with too much breath or mental concentration placed on the rising and falling sound) may cause symptoms such as dizziness and slight fever, tightness in the chest, weakness in the legs, foul breath, and bloody stool.

These symptoms cannot be relieved by Western medicine. Some patients, however, have succeeded in rectifying such deviations by practicing the Opening and Closing the Three Dantians exercise (see Chapter 15). This exercise is used to regulate the Excess Yang, and to restore the body's energetic equilibrium.

Some patients are given the prescription of uttering one pitch sounds (known in China as the first or straight tone). This straight sound has a steady tone, and a rising volume, which can easily cause the energy to ascend. After the patients have completed their prescription and ended the exercise, if the energy continues to rise, the Qi deviation will cause dizziness and nausea. In this particular case, the patients should stop the exercise immediately. To rectify this condition, have the patients utter the sound in the third tone (the descending and ascending tone), and by deepening his or her voice, then raising the pitch. This increasing volume and deepening tone relaxes the body and causes the Qi to descend.

This example illustrates the fact that if the patient is allowed to select any prescription at random for mental concentration, without understanding the theory of Medical Qigong, he or she can easily get into trouble.

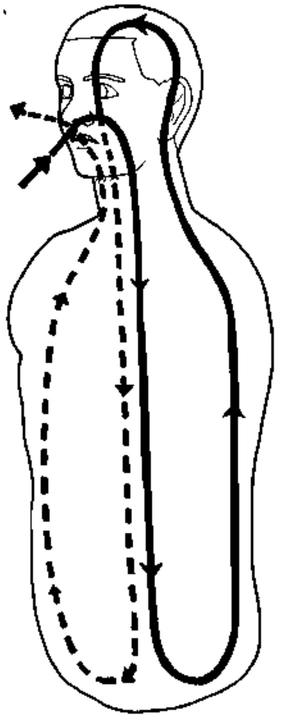

Figure 18.2. Rectifying Respiratory Qi Deviations

4. To rectify respiratory Qi deviations, the following exercise can be used. It will amend respiratory deviations and balance the Yin and Yang Qi, as well as clear any stagnant Qi.

When the practitioner feels uncomfortable after Medical Qigong exercises, practice this method for 20 minutes. This is an alternate nostril breathing exercise and requires both inhaling and exhaling through the nose. The abdomen expands upon inhalation and contracts upon exhalation.

From a seated posture, begin by imagining that there are three channels that start at your tailbone and travel up the body. The first channel intersects with the right nostril, the second channel joins the Baihui GV-20 point and the third channel intersects with the left nostril (Figure 18.2).

Place your left hand on your Lower Dantian, with the right hand resting on the nose. It is important to focus the mind on the energetic flow of the breath. Inhale through the left nostril and allow the Qi to flow down the Conception Vessel into the Lower Dantian and tailbone area. Hold the breath, pinch the nose and draw the Qi up the Governing Vessel into the Baihui area, and then circle the energy down to the Conception Vessel to the coccyx to complete the Microcosmic Orbit. Exhale through the right nostril. Repeat this sequence, alternating the focus of the mind's concentration to follow the beginning flow of energy from the left to right nostril and visa versa.

When training in Medical Qigong prescriptions, it is equally as important to relax the mind as the body. Therefore, this exercise should be practiced in a leisurely manner, with a carefree attitude, and a smile.

MENTAL DAO YIN DEVIATIONS

Deviations in mental activities are primarily caused from excessive thinking and too much focused concentration. These type of mental deviations should be rectified by controlling one's mental activities. It is important to fully understand the foundational principle of the Shen controlling all mental activities that determine the physical, emotional, and energetic activities. Any mental deviation should be regulated and controlled through intention.

It is sometimes difficult for patients who have just started Medical Qigong prescriptions to become absolutely tranquil; this may lead to mental Qi deviations. One way mental Dao Yin deviations can be rectified is by a mantra, or counting breaths (using one thought to replace many thoughts). After this technique is mastered the patients can advance to concentrating their mind on an external object in the surrounding area.

1. The cerebral cortex plays an important role in Qi deviations of the mind. Medical Qigong regulates the cerebral cortex to restore balance to the energetically unbalanced parts of the brain. If the mind is agitated or depressed for a long time, it can cause a long term disorder of the whole nervous system and, thus, cause various kinds of chronic diseases. A healthy nervous system takes the leading role in adjusting, regulating, and unifying all functional activities. It is responsible for maintaining a dynamic equilibrium.

If deviations occur during Medical Qigong practice, the cerebral cortex loses control of the nervous system. Abnormal phenomena may occur and mental and emotional activities may become uncontrollable. If patients are not careful, they can actually develop meditation induced diseases, or meditation induced psychosis known as Tsou Huo Ru Mo, which translates to "the Fire becomes Excessive and the demons enter." In other words, the patient's Heart and Liver become overheated, causing the Hun to leave and Po to take control of the body. This in turn can lead to chronic psychoses or demon possession.

Because Qi deviations can have serious consequences, the Qigong doctor must have an extensive knowledge of Medical Qigong principles to prevent and correct Qi deviations, stagnations, and adverse Qi flow. The mind should, for example, always be free from over-concentration or "distortions of consciousness" during practice. If a beginning patient feels a hot sensation in certain parts of the spine, and tries to force its energetic movement up through the Five Passes of the Governing Vessel (see Chapter 30), this can cause "distortions of consciousness" and mental confusion, which may result in hallucinations.

A deviation caused by excessive focused mental concentration can also lead to a loss of self-control and produce such disorders as uncontrollable head-shaking, shoulder shrugging, hand and leg quivering, stumbling, staggering, uncontrolled crying or laughing, or even stiff and twisted extremities and opisthotonosis. To avoid this loss of self-control, do not allow the patients to use their mental intention and focused concentration at random, but rather have them focus their attention on quiescence to tranquilize the mind and recuperate their health.

Failure to properly lead and guide energy during Medical Qigong practice may also cause adverse deviations of Qi in the patient's Middle Dantian. In the Middle Dantian, the patient's Lung Qi rises and returns through the opening and closing actions of the Shan Zhong CV-17 point. If the Qi within the Middle Dantian begins to deviate, it can condense and cause stiffness, pain, and a feeling of suffocation in the chest region. To correct this condition the Qigong doctor extends energy into the patient's Middle Dantian to regulate and tonify the patient's Qi, and especially the Gathering (Zong) Qi. This treatment also regulates and tonifies the patient's Lungs and Upper Burner, as well as diffuses any Rebellious Lung Qi (see Chapter 22) as it unburdens the emotions of the Heart.

2. There are three rules for determining proper mental focus. The following three rules determine the focus of the patient's mental activities and are implemented to help avoid Qi deviations while practicing Mental Dao Yin training:
    a. Concentrate the mind on a motionless object. A patient suffering from hypertension can, for example, concentrate on still lake water, but not on a stormy sea.
    b. Concentrate the mind on a near object. A patient with Liver trouble can, for example, concentrate on a nearby pine tree. If the patient is at a park, absorbed in concentration on a pine tree that is far away, the patient must mentally trace and locate the pine tree with painstaking effort. This is achieved through expanding and extending the imagination. If the patient suddenly becomes mentally distracted or preoccupied, the purpose of the meditation is defeated. Once distracted, it will be very difficult for the patient to reconnect to the pine tree because of its distance. As a result, the patient's mind wanders and the patient may become discouraged.
    c. Concentrate the mind on a familiar object. Concentrate on something which creates a peaceful, emotional connection. The emotional connection will facilitate the proper attitude and diminish any likelihood of mental strain.

    When deciding on which familiar object to focus the attention on, patients must not choose a person, their own work or creations, or obscene things. Objects that have been used to do harm or evil should also be avoided. Patients are also forbidden to shift their thoughts from one thing to another or back and forth without stopping. Remember that it is the individual's Shen (Spirit) which guides all mental activity and determines the body's energetic actions, therefore the mental concentration should have purpose.

Adjust each technique according to the specific case. When prescribing Medical Qigong therapy, it is important for the doctor to adjust each meditation and exercise according to the patient's specific case. In all cases, the Yi (imagination and intention) must lead the Shen and the Shen must lead the Qi.

The following examples demonstrate how to adjust prescriptions for different patient's needs.

- Patients with hypertension should concentrate on a low place instead of a high place, or on an object below eye level.
- Patients with hypotension (including other Deficient conditions such as anemia) should concentrate on a high place instead of a low place, or on an object placed above eye level.
- Patients with Lung problems such as tuberculosis should concentrate on white and light objects (e.g., white clouds).
- Patients with Spleen and Stomach diseases should concentrate on yellow objects (e.g., yellow chrysanthemum, dahlia).
- Patients with Liver problems like hepatitis should concentrate on green objects (e.g., pine and cypress trees).
- Patients with a Heart Deficiency should concentrate on red objects. Patients with an Excess conditions of the Heart, however, should

never concentrate on the color red, and should only choose pink objects instead. A pure red color is too energetic for patients with Excess conditions and may overexcite their Heart, making it difficult for the patient to become relaxed.

- In Medical Qigong, the color black is sometimes assigned to the Kidneys. Focussing on this color can, however, easily cause the Kidney energy to sink downward, creating an emotional state of depression or anxiety. Therefore, when treating patients with Kidney diseases, it is preferable to choose things of purple, indigo and vibrant blue color.
- Mentally depressed patients (in a severe cases) should not practice any dynamic Qigong exercises until their depression has passed. Otherwise, they may develop headaches, tightness in the chest, or in severe cases, shock. To rectify this condition, the Qigong doctor must lead the toxic and stagnant Qi downward along the channels and out of the body.
- Patients with uncomfortable light or heavy sensations should concentrate on the opposite side on their body instead of the afflicted area. If they feel that their lower limbs are too heavy to lift during the exercises, for example, they can correct this symptom by focusing on the Baihui GV-20 point at the top of the head. If the patients feel that their body is too light and unsteady, they can focus on the Yongquan Kd-1 point located on the soles of the feet.

STRESSFUL LIFE ACTIVITIES

Stressful life activities create a flood of adrenaline within the body. This occurs not only from the core of the adrenal gland in the medulla (which responds to stress by releasing adrenaline in to the blood stream), but also from the endings of the vertebral lumbar division at the core segment of the major internal organs. Adrenaline spreads throughout the body and activates not only the body's sympathetic nervous system, but also the adrenal glands. The Heart rate, respiratory rate, blood pressure, and Blood flow to the muscles are all increased as they respond to the "fight or flight syndrome."

When this happens, the immune system is called upon to balance the whole stress mechanism and is, therefore, called away from its normal activities which include monitoring cellular activity and fighting the growth of disease. A life of constant stress, therefore increases susceptibility to illness. Meditation takes the patient out of the stress mode, thereby shutting down the adrenal glands, and allowing the immune system to return to its normal functioning.

Chronic stress wears down the mind and body, thereby producing anxiety or depression. Stress begins with an overexcitement of the nervous system and then slowly works its way deep into the body like wet cement, causing the shoulders and upper back to become tense and rigid. Happy occasions (such as weddings) and other normally pleasant events, also produce stress, which can be detrimental when the patient's life style is not properly balanced. An unstable life style is very stressful and can lead to depression, anxiety, and Qi deviations. The capacity to bear stress depends upon the individual's constitution, diet, support system (or lack thereof), the nature of the stress, and the congenital strength of the nerves. When stress levels surpass the nervous system's capacity to handle them, "the nerves break down", resulting in all sorts of physical, mental, and emotional disturbances. Certain neurotransmitters are depleted, leading to a chemical imbalance which can be aggravated by self medication (e.g., alcohol or drug abuse). These disturbances, if unheeded can eventually lead to organ malfunctions, cysts, tumors, cancer, and premature death. In order to avoid such states, a balance between work and rest must be maintained.

At a conference of the International Society for Neuro-immunomodulation (entitled *The Effects of Stress and Depression on Physical Disease*) Dr. Philip Gold, of the National Institute of Mental Health, stated that stress and depression cause the body to secrete high levels of hormones; (for example, cortisol) into the bloodstream. These hormones, if unregulated, can destroy the appetite, cripple the immune system, shut down the processes that repair tissue, inhibit sleeping patterns, breakdown bone density, aid the onset of infection, and even instigate cancer cell growth.

An individual's emotional past determines his or her present state of health, both physical and emotional. Past emotional belief structures are responsible for both creating and healing diseases. Just as thoughts drift through the mind but do not manifest until they reside within a belief structure, so emotional wounding resides within the tissues and cells waiting to be activated by emotional upheavals. Our thoughts materialize via choices and manifest within the body.

In American society, sharing traumatic wounds is often the first act of intimacy and bonding. It provides a common ground for relationships and supports the "pain structure" that deepens the bond. People can become addicted to empowering their wounds which, in turn, support and empower the disease. This exchange of painful personal traumas is reinforced by receiving sympathy and strong empathy, but does not fully release the pain from the tissues.

Although talking about past traumas can begin the process of healing, problems can emerge when these past traumas manifest through the following belief structures.

1. The individuals believe that they can only receive affection and attention when they suffer, and use their "neediness" to keep their significant other(s) from leaving or becoming more independent.
2. The individuals become stuck in one emotion -- anger, grief, fear, worry, or hurt -- and are unable to express a healthy range of emotions.
3. The individuals do not have an adequate support system.
4. The individuals do not have the incentive to heal themselves or the belief that they can be healed.
5. The individuals bond to anyone, out of fear of being alone, as long as they receive the attention they crave.
6. The individuals have low self-esteem and do not seek help or work out their emotional issues; they have difficulty trusting anyone to understand their pain.

The desire for sharing ones pain in order to feel loved can lead to a habitual way of relating to those closest and dearest. When such individuals feel taken for granted, or unappreciated, they may immediately resort to past traumas. If this fails to illicit the desired response, a new trauma or wounding may be unconsciously created to get back the attention and love and need. The individuals may, for example, physically injure themselves in an accident, or manifest a physical illness. This in turn can become an addiction, empowering old wounds and creating new ones, and may lead to chronic illness. The Shen becomes divided to support the new belief structure, and the Qi deviates from its assigned job of protecting the body, thus creating illness. This is not a conscious process, but an unconscious one, which needs to be brought into the patient's awareness to be healed. Then, and only then, can the Yuan Shen and Zhi Shen act as one unit directing the body's Qi towards its proper function of healing.

Because energetic healing occurs in the present and not in the past, victims of past traumas will never heal if their energy continuously regresses to the armored traumatic belief structure. Both the doctor's and the patient's Yi (Intent) must therefore, focus on forgiveness (of self, others and the specific tragic situations) to free the patient from toxic wounds. The patient's desire and determination to be well must reach the degree that they are willing to do anything to get well, including giving up on the "secondary gains" of illness (i.e., risking abandonment, or loss of attention received during the illness). Patients must be encouraged to find new ways to fulfill their needs, and must be encouraged and supported through this difficult transition. This point is illustrated effectively in the following traditional Chinese story.

> There once was a palace official named Cheng Guang Wei who owned a servant named Song Li. As the pressures of life wore on Director Cheng, he began to sink into a deep depression. Soon he began to lose all hope in life and desperately sought relief. Not content with the wine and festivities at hand, his attention soon turned to his servant Song Li, who always seemed to be happy. Song Li was plowing the fields, up to his ankles in

mud, singing and seemingly enjoying his life. Upon observing this, Director Cheng walked up to the edge of the field and called out, "Song Li, why are you so happy? What is the secret of your contentment?" To which Song Li responded, "Sir, come into the field with me, and I will tell you." The director, not wanting to get himself muddy said, "No, tell me from where you are." To which Song Li replied, "I cannot explain everything to you if you do not come into the field." Director Cheng, aggravated by this reply, answered, "Very well, since you refuse to tell me, tomorrow you will work in the stables." Song Li answered, "If that is your wish, sir," and went about his work contentedly.

Two weeks went by, and the pressures of the palace were starting to take a serious toll on Director Cheng. All of his problems, it seemed, were insurmountable. Locked in turmoil, he began to roam his property seeking peace until he heard singing coming from one of the stalls in the stables. Director Cheng went to investigate and sure enough, it was Song Li, now up to his knees in horse manure and still singing as beautifully and joyfully as a bird. At first Director Cheng was enraged at Song Li's contentment, but then became perplexed, because the duties of Song Li's new job were much more difficult than the previous work.

Director Cheng approached the edge of the stables and called, "Song Li, why are you so happy? What is the secret of this contentment you possess?" To which Song Li responded, "Sir come into the stall with me, and I will tell you." Repulsed by the stench of the horse manure and filth, not to mention feeling protective of his royal robe, Director Cheng answered, "Song Li, tell me the secret of this joy and contentment that you possess even in the midst of a barn full of stench and dung." Song Li answered, "I will tell you sir, but first you must come in to where I am." "If that is your answer, then tomorrow you will work in the hog pens!" shouted Director Cheng. To which Song Li answered, "If that is your wish, sir," and went about his work contentedly.

Well, things went from bad to worse for Director Cheng in the palace; seeing that he was about to lose everything, he contemplated taking his own life. Making one last round of his property, he again heard singing, this time from the hog pens. Realizing it was Song Li, Director Cheng slowly approached. Sure enough, there was Song Li, up to his thighs in pig dung, slop and filth, and still singing. Leaning against the fence, Director Cheng called out, "Song Li, why are you so happy? What is the secret of your contentment and joy?" To which Song Li responded, "Sir, come into the pen with me, and I will tell you." Director Cheng, now at the end of his rope, answered, "Very well, I will come in with you," and began to climb the fence to enter the hog pen. "Stop!" shouted Song Li, "I will come over to you, it is not right that you defile yourself on my account. But now you will truly listen to what I have to share and are ready to hear the words I speak. For now you are willing to sacrifice all, even your physical comfort, to receive spiritual knowledge and insight."

The moral of the story is clear: unless we are ready to abandon all masks (pretenses and defence mechanisms) of the emotional ego, it is extremely difficult to free ourselves from the toxic emotions that steal our life-force energy and direct our lives. If we are not intimate with our emotions, we cannot perceive the dynamics behind those emotions. After all, emotions are only energy currents which flow in, out, and through our bodies. Because our emotions reflect our intentions, any awareness of unprocessed emotions will lead to an awareness of deep-seated intentions. It is through this understanding that Qigong doctors train their emotions and mental thinking to regulate the Shen.

SHEN DISTURBANCES AND EMOTIONAL DYSFUNCTIONS

All Shen disturbances relate to emotional energetic dysfunctions. Generally, Shen disturbances can be divided into two categories: Yin disturbances and Yang disturbances (Figure 18.3). These

Yin and Yang disturbances affect the patient's thoughts and emotions.

A thought is energy that has been shaped by consciousness. Every experience, or change in experience, reflects intention, desire, and will. Any discrepancy between intention and emotions leads to a splintering reaction of the energetic self, which can cause a breakdown of the body's life-force energy and ultimately lead to disease.

Emotional energetic currents have different frequencies, resonating between low (fear, anger, worry, etc.) and high (love, forgiveness, joy, compassion, etc.). Lower frequencies deplete the physical body, draining it of its precious reserve of life-force energy. This deterioration begins to manifest in the spiritual fields of energy surrounding the body, and percolates down to the emotional, mental, and finally the physical level. Conversely, high frequency emotional energy raises the body's energetic frequency and begins to radiate outside the physical body. This does not mean that the patient must learn to suppress feelings of anger, hurt, etc. These are natural emotions that serve a positive function in protecting the individual from harm. It is only when emotions are suppressed, denied, and accumulated that they become highly toxic to the body. When this happens, new anger is fueled by old anger. The new hurts are suddenly blown out of proportion (related to the event), fueled by ancient emotional pain, until the patient's physical structure is overwhelmed.

Patients may become overactive and overemotional when under stress, when their energy becomes disrupted. Every emotional disturbance alters the breathing pattern, changing from slower to faster respirations (each breathing pattern varying according to the internal organs involved).

According to *The Yellow Emperor's Inner Cannon* (Spiritual Axis), the Blood, Ying, Jing, Qi, and Shen are stored in the Five Yin Organs (Liver, Heart, Spleen, Lungs, and Kidneys). An Excess condition will cause any of these components to leave their respected organs. This will:

- deplete the patient's Jing (Essence),
- cause the patient's Hun (Ethereal Soul) and Po (Corporeal Soul) to become unsettled,
- cause the patient's Zhi (Will) and Yi (Intention) to become muddled, and
- cause the patient's Shen (Spirit) to leave the patient's body.

When emotionally upset, energy literally rises upward, affecting the organs and releasing additional emotions stored within the tissues. To become free from these emotional reactions, patients must be taught to relax and release the energetic charge of emotions. This will allow stagnant Qi to become energetically discharged from the body. By cultivating a healthy attitude, and not holding onto, or suppressing, the emotions, the body will begin to seek its energetic balance naturally. This is initiated by having the patient sink his or her Mind and breath deep into the Lower Dantian to rebalance the life-force energy.

Shen Disturbances	
Yin	**Yang**
Qi is compressed.	Qi is expanded.
Patient is depressed, withdrawn.	Patient is manic, expressive, impulsive, volatile.
Patient is too quiet.	Patient is too loud.
Patient avoids feelings.	Patient inappropriately expresses feelings.
Patient is too introverted.	Patient is too extroverted.

Figure 18.3. Shen Disturbances

TREATMENT TECHNIQUES

When treating patients with Shen disturbances, the Qigong doctor categorizes the emotional and thought dysfunctions into Yin and Yang Shen disturbances.

1. In Yin Shen disturbances the patient's energy compresses inward. These patients have a tendency to avoid feelings, becoming quiet, withdrawn, and depressed. They also tend towards an introverted personality and posture.

   When treating patients with a Yin Shen disturbance it is important for the Qigong doctor to redirect the patient's Shen towards specific actions, and encourage them to use spiritual mantras and affirmations to gain inner strength and confidence.
2. In Yang Shen disturbances the patient's energy expands outward. These patients have a tendency to express their feelings, are loud, impulsive, volatile, and manic. They also tend to be extroverted in personality and posture.

   When treating patients with a Yang Shen disturbance it is important for the Qigong doctor to redirect the Shen (with a "reality check" towards the patient's actions) and set parameters for their boundaries. A sudden awakening to the responsibility of personal actions will generally send the patient into emotional shock.

   One treatment technique commonly used in the clinic to calm and restore harmony in the patient's Shen (due to a Deficient condition) is for the Qigong doctor to extend Qi into the patient's Middle Dantian and Yellow Court areas. Once this area has been sufficiently balanced, the doctor then ends the treatment by rooting the patient's Qi, by drawing it into the Lower Dantian.

PRESCRIPTIONS AND HOMEWORK

Several of the most recommended meditations used for the purpose of calming the patient's Shen are listed as follows.

1. The Energy Melting Meditation is used to relax the body and reduce stress. Relax and imagine melting from the top of the head to the bottom of the feet, like ice melting into water and water flowing downward into the Earth. This image centers, grounds and roots the Mind.

   If the patients are lying down, have them feel the upper layers of their body melting through the lower layers and into the Earth. Ice melts into water and pours into the Earth.
2. The Exchange of Fire and Water Meditation unifies the energies of the Kidneys (Water) and the Heart (Fire) and is often called the Fusion of Kan and Li. It is used to unify opposite principles within the body, to balance the energies of the mind, body, and Prenatal and Postnatal Shen for creating wholeness. It also strengthens the immune system and the bones, thus slowing or preventing osteoporosis.

   Sit and begin by breathing naturally, while inhaling and exhaling through the nose. Visualize the Lower Dantian as an ocean of water, and the Heart as a ball of fire. Imagine the waters of the Lower Dantian beginning to flow up the center of the body, reversing their natural flow. Now imagine the Fire of the Heart be-

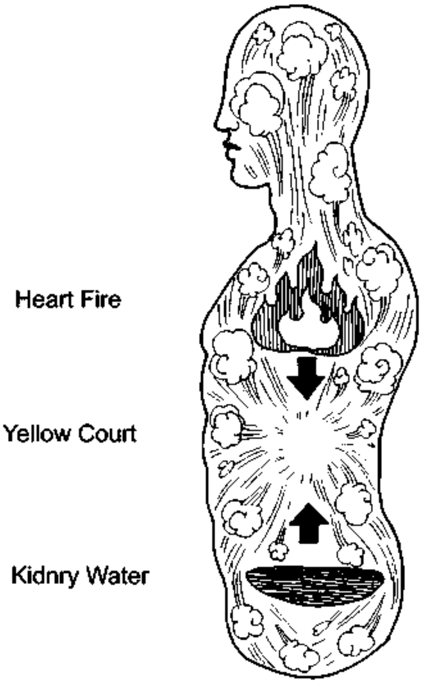

Figure 18.4. During the exchange of Fire and Water meditation, the fusion of Heart Fire and Kidney Water creates steam.

ginning to descend the center of the body, reversing its natural flow. As the Water and the Fire join in the Yellow Court (located in the lower part of the solar plexus), the Water vaporizes and becomes steam (Figure 18.4). This hot mist travels throughout the entire body. It should feel warm and pleasant.

Next allow the hot mist to enter into your bones and flow through the Marrow. Imagine this mist beginning to solidify and harden making your bones solid and radiant with white light. Upon completion of the meditation, relax and allow the images to disperse as you sink into the Wuji.

These following exercises are a series of meditations used in the clinic for detoxifying emotional stasis.

3. The Sun and Moon Meditation is used to detoxify emotional stagnation and establish physical, emotional, and spiritual clarity.

   Begin from a sitting or Wuji posture (Figure 18.5). Inhale and exhale through the nose. Imagine the sun over the left eye, and the moon over the right eye, appearing side by side over the head. Feel their presence over the Baihui point, at the top of the head. Imagine that the sun (on the left side of the body) pours golden light, and the moon (on the right side of the body) pours silver light. Both streams of light meet at the Baihui point on the top of the head, combining into white light energy. This white light energy pours into the body, filling the entire body from the feet to the top of the head (like a pitcher of water filling a glass). Once the body is completely full, the energy begins to spill out of the pores. At first a thick, black, sticky energy pours out of the pores, and flows down to the ground. This energy contains all the physical, mental, emotional, and spiritual toxins that are being released from the body. As more and more of these toxic substances leave the body, they start turning from black to gray, then from gray to white. Feel the body radiate this white energy in all six directions (front, back, right, left, up, and down) filling the entire room. After several minutes of purging and tonifying the body, relax the mind and focus your attention on to the Lower Dantian, before ending the meditation.

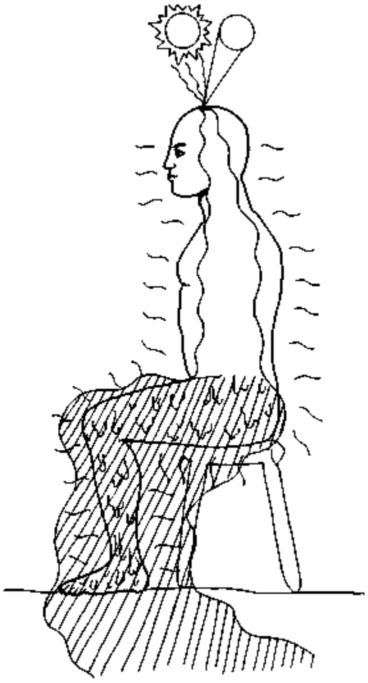
Figure 18.5. During the Sun and Moon Meditation imagine that the white light energy that is pouring down the body, is purging toxins from the internal organs, releasing the Evil Qi out from the pores.

4. The Releasing Emotional Blockages and Energetic Armoring Meditation is used to dissolve emotional blockages from the internal organs. When these blockages are released, the emotions are restored to balance. Begin from a Wuji posture, and inhale through the nose and exhale through the mouth.
   a. Direct your attention to your Liver. Every time you inhale, visualize and feel the emotions of kindness and compassion entering into your Liver. As you exhale, release anger or jealousy. Repeat for 18 breaths.
   b. Direct your attention to your Heart. Every time you inhale, visualize and feel the emotions of joy, contentment, and tranquillity entering into your Heart. As

you exhale, release nervousness and excitement. Repeat for 18 breaths.

c. Direct your attention to your Spleen. Every time you inhale, visualize and feel the emotions of trust, openness, and sincerity entering into your Spleen. As you exhale release worry, obsessiveness, and self-doubt. Repeat for 18 breaths.
d. Direct your attention to your Lungs. Every time you inhale, visualize and feel the emotions of dignity, integrity, and courage entering into your Lungs. As you exhale release grief and anxiety. Repeat for 18 breaths.
e. Direct your attention to your Kidneys. Every time you inhale, visualize and feel the emotions of self-confidence and inner strength entering into your Kidneys. As you exhale release fear and loneliness. Repeat for 18 breaths.
f. Finally, inhale, visualize and feel the emotion of a full-bodied state of peace and bliss. As you exhale, ripple and resonate this feeling into the room, filling the surrounding energetic space with this emotional enlightenment. Repeat for 18 breaths.

DISTURBANCES DURING QIGONG PRACTICE

The patient should carefully choose a safe and quiet environment for Medical Qigong exercises and meditations. Most often a patient is already in a state of relaxation when they become startled by disturbances generally caused by an unexpected sound or event. This sudden disturbance can disrupt the Heart Qi. When training, if you become startled, do not open your eyes. If you open your eyes, the disturbed energy will be thrown into disarray, causing a blockage of vital Qi (usually in the Heart). This reaction will cause deviations which can be difficult to rectify.

If instead you continue exercising with your eyes shut as if nothing had happened, you can restore your mind to a normal state of quiescence, by guiding the Qi up the Governing Vessel and down the Conception Vessel (i.e., the Microcosmic Orbit). Even if you were shocked by the un-

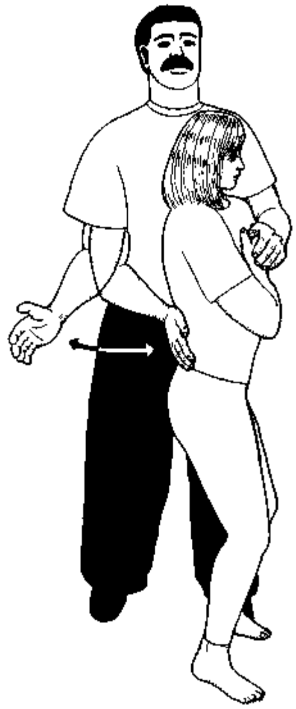

Figure 18.6. The doctor strikes the patient's Mingmen area and focuses his or her attention to lead the disruptive Qi downward into the Earth. This helps the patient stop the uncontrollable muscle deviation spasms.

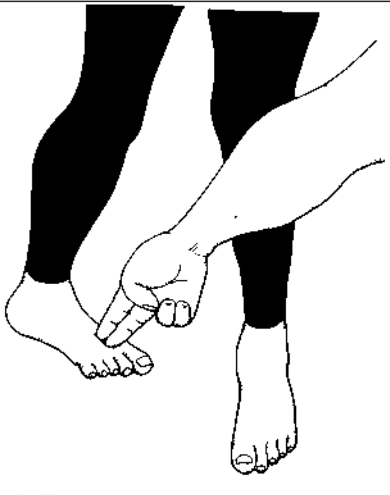

Figure 18.7. The doctor directs energy into the patient's foot, through the Lv-3 point, rooting the patient's Rebellious Qi into the Earth.

expected sound or event, restore the mind to a normal state of calmness by immediately performing the Opening-and-Closing the Two Dantians technique. The two Dantians refer to the Upper Dantian (Yintang point) and the Lower Dantian (Qihai point) areas (see Chapter 20).

To perform this technique, raise both hands to the lower jaw and breathe in with a hissing sound as if inhaling a mouthful of cold air. After the inhalation, the hands will move up to the Upper Dantian, and perform the opening exercise while exhaling slowly. Next, exhale while closing slowly. After four repetitions move the hands to the lower abdomen and perform the Opening and Closing of the Lower Dantian four times.

Usually the emotions will calm down after performing this exercise four times. If you still feel uncomfortable, you can perform the exercise eight times or more, until you become calm.

Note: It is good to ensure a quiet, safe environment in which to practice, preferably a place outside, with fresh air, among flowers, grass or thick, healthy green trees. Because the inside of the body matches the outside environment, when searching for specific directions to face while practicing Medical Qigong, it is important to avoid training in or near:

- extremely intense direct hot sunshine,
- a cold strong wind,
- a thunder and lightning storm,
- old withered trees,
- old graves or tombs,
- electric generators, power lines or power plants,
- dump sites,
- polluted sites, rivers, etc., and
- do not practice shortly after earthquakes, tornadoes, hurricanes, hailstorms, etc.

UNCONTROLLABLE MUSCLE DEVIATIONS

Sometimes a patient's body begins to automatically quiver, shake, tremble, or vibrate, revealing an energetic re-patterning causing muscle movement which cannot be stopped. When this happens, part of the body often keeps moving for several hours, causing exhaustion. Sometimes Qi in the legs and Heart rushes into the head, compounding the problem.

For older patients this is considered a dangerous symptom because their lower extremities are generally weaker. In mild cases the patients feel energy rushing upward into their head. In severe cases the whole body trembles because the Qi leaves the Conception and Governing Vessels and empties into the body itself. In these extreme cases energy muscle patterning should be controlled in the following manner.

1. Advise the patient by saying, "It is time to relax and calm down. You need to end your exercise and rest." This method is usually very effective when repeated over and over, again. If the patient is unable to stop, perform the next method.
2. Direct the patient to place their left hand on the Middle Dantian and the right hand on top of the left hand. Tell the patient that you will help them stop by striking their Mingmen area (Figure 18.6). Press their right hand with your left hand, to compress their Middle Dantian, while using your right palm to strike the patient's Mingmen three times, directing your attention down to their feet. This can help them stop moving.
3. If the above two methods do not help, direct the patients to turn their right foot inward with the big toe pointing at the heel of the left foot. If they cannot do it without help, take their left hand in your right hand and with your foot push the big toe of the patient's right foot to point at the left heel. That should stop the movement.

   Next, touch the patients' Taichong Lv-3 point, between the big toe and second toe, with the "sword fingers" Qi emitting technique. In this technique, the doctor uses the index and middle fingers to emit Qi into the patient and guide their Toxic Qi down into the ground (Figure 18.7).

RECTIFYING QI DEVIATIONS USING TAIJI RULER

If patients are experiencing too much Heat, they are concentrating too much and need to change the position of their tongue and perhaps

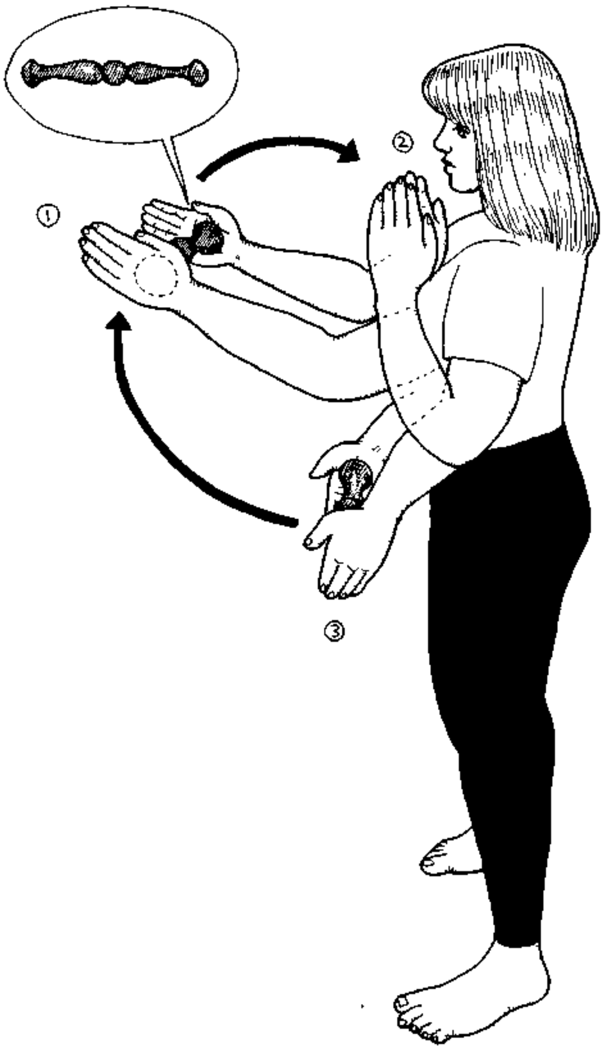

Figure 18.8. The Qi regulation exercises of the Taiji Ruler system are excellent for correcting Qi deviations.

switch meditations. The Taiji Ruler is very effective in rectifying deviations caused by inappropriate tonification or purgation, and can be used to dredge the channels and collaterals to promote the flow of Qi and Blood.

When performing this exercise, the rotation of the Ruler upward and outward (away from the body) is generally used for leading and increasing the energetic flow of Qi and Blood in the upper extremities. The downward and inward rotation of the Ruler is generally used for leading and increasing the energetic flow of the Qi and Blood in the lower extremities.

The purgation of energy blocks are usually cleared through downward movements (from Heaven to Earth). Tonification is implemented through upward movements (from Earth to Heaven). Using these methods, the patients can amend their deviations and promote the increase of Qi (Figure 18.8).

CHAPTER 19

DISORDERS AND DEVIATIONS OF THE SOUL AND SPIRIT

INTRODUCTION TO SPIRITUAL DIMENSIONS

An accepted reality in Medical Qigong therapy is the body's relationship to both the physical and spiritual world. The physical world of the living is considered a Yang state of existence, while the supernatural-spiritual world is in a Yin state of existence. Once the Qigong doctors' Shen has reached a high state of sensitivity, they can transcend their normal conscious perceptions and senses, to observe, and even communicate with the spiritual Yin world (see Six Transportation's of Shen, Chapter 13).

Within the subtle realm of the spiritual world, the existence of spiritual entities and ghosts are divided into two divisions of Yin and Yang (Figure 19.1). There is a significant difference between spiritual entities and ghosts.

1. A spiritual entity belongs to the Yang category of the Subtle Realm, and comes from the formation of nature. Spirits are considered the ethereal beings of the universe, who depend on the absorption of natural energy and are affected by the energetic cycles of the Earth. They are conscious beings, able to transform themselves into any size, shape, animal, or being (shape-shifting). A spiritual entity can affect changes in electrical current within the body's energy fields, or control certain areas on the body's nervous system. Spiritual entities can be separated into two divisions: Yang (good) and Yin (evil) influences.
    a. Yang spiritual entities are commonly referred to as angels, guides, spirits of the light, enlightened masters, and similar positive beings. These spiritual entities assist mankind (communicating through the body's Hun) in their quest for enlightenment. They assist individuals to make beneficial decisions that promote spiritual maturity.

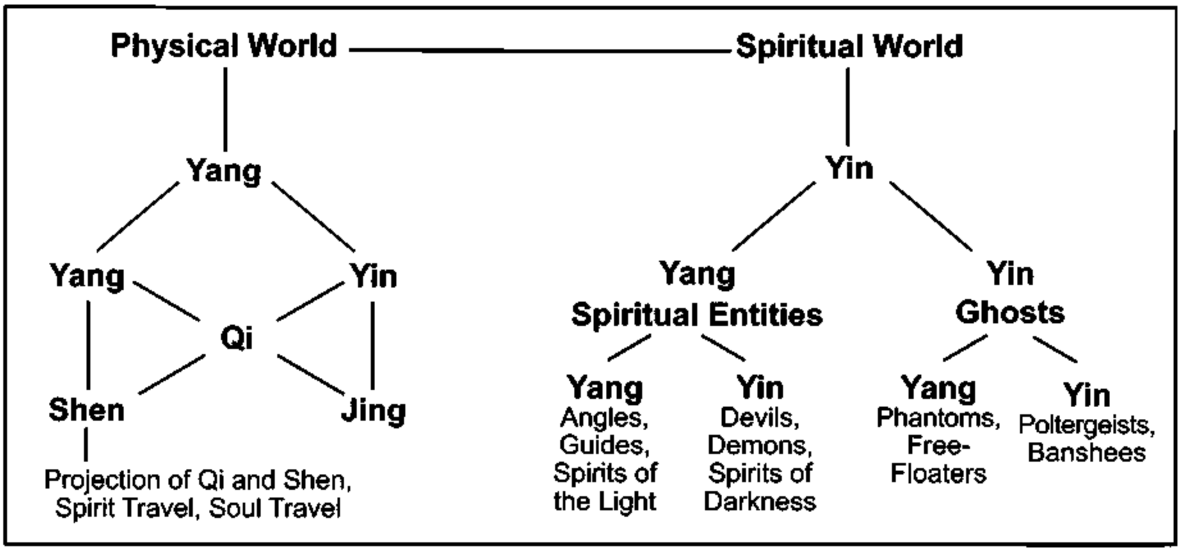
Physical World
Spiritual World
Yang
Yin
Yang
Yin
Qi
Shen
Jing
Projection of Qi and Shen, Spirit Travel, Soul Travel
Yang Spiritual Entities
Yin Ghosts
Yang
Angles, Guides, Spirits of the Light
Yin
Devils, Demons, Spirits of Darkness
Yang
Phantoms, Free-Floaters
Yin
Poltergeists, Banshees

Figure 19.1. The Spirit World

   b. Yin spiritual entities are commonly referred to as devils, demons, spirits of darkness, evil spirits, and other similar negative spiritual manifestations. Yin spiritual entities challenge mankind (communicating through the body's Po) by confusing the human spirit. If given the opportunity, these spiritual entities can possess an individual and become extremely territorial.

2. Ghosts or Gui, as they are commonly called in China, belong to the Yin category of the subtle realm. Ghosts are developed from the subtle form of the combined spiritual essence of an individual's Eternal Soul. When the body dies, the energies of the Corporeal Soul (Po) enter into a state of rest, returning back to the Earth, thus allowing the Ethereal Soul's (Hun) to return back to Heaven. The individual's Eternal Soul then fuses with the energies of the Three Dantians (within the Taiji Pole) and is projected outside the body, entering the tunnel of light to return back to the divine. Sometimes, however, because of unprocessed emotional issues, traumatic death, unfinished business, or extreme attachment to people, places, or possessions, an individual's Eternal Soul can begin to wander the Earth, becoming a Gui or ghost. Generally, it is believed that it takes three days for the Corporeal Soul to withdraw their energy completely from a deceased body. Ghosts are considered Eternal Souls that have departed from human life and are unable to detach from their life experiences; they are bound by their unprocessed issues. Ghosts can also be separated into two divisions of Yang (good) and Yin (evil).
   a. Yang Ghosts are commonly referred to as phantoms and free floaters. These disembodied souls can either be lost, confused, delusional, or purposely return in order to guide and protect the living.
   b. Yin Ghost are commonly referred to as poltergeists, banshees, or other names varying by location. These disembodied souls are angry, malicious, mischievous, and destructive.

It is difficult, if not impossible, to be immune to the influence, or encroachment, of the spiritual world. Spiritual manifestations are usually evident through audible sounds, unusual unexplainable smells, extreme cold, and the displacement of objects. Other phenomena include tactile sensations, visual images, voices, and the apparent psychokinetic movement of objects. Spiritual entities seem to move through solid matter and can appear or disappear abruptly. They can also cast shadows and be reflected in mirrors; some seem corporeal, while others are luminous, transparent, or ill defined. Many spiritual entities have jerky or limited movements, while others are lifelike in movement and speech. Encounters with spiritual entities can be divided into five main categories.

1. Spiritual manifestations of the dead generally occur within a short time after the death of the person. Visits from the departed are usually to bring comfort, to announce their role as a guardian, or to complete unfinished business. If an individual's life ends in an untimely manner, or so violently that the soul does not realize the body is dead, the disembodied soul often goes about its daily routines. Occasionally these souls will manifest as mischievous entities, attempting to get the attention of the living, but they lack the malevolence of an evil poltergeist manifestation.
2. Deathbed phenomena are generally of divine beings, religious figures, or luminous apparitions. Visions of previously deceased loved ones, who come to guide the dying soul into the next level of existence, are common occurrences.
3. Manifestations of impending tragedy are visual images of an entity which usually appears in a "waking vision" or in dreams. When they appear, these spiritual entities either communicate about the death or severe crisis of a loved one, or to give warning of an impending tragedy.
4. Reincarnating spiritual entities sometimes appear in a dream to a member of the family into which they will be born. Such dreams are referred to as "announcing dreams."

5. Manifestations of poltergeists are generally mischievous and often malevolent. They may be either disembodied souls, or demon-like entities (see Chapter 18). These disembodied souls were often extremely dissatisfied and bitter while living, and may be unable to reintegrate into the Earth after death, or to ascend into the Heavens. This may cause spiritual hauntings, either around the location of their death or around a person, or persons, that the disembodied souls blame for their life's misery. They make their presence known by assaulting the living, either human or animal. The most common phenomena include: battering with a rain of small stones or sand, throwing and moving of objects, loud noises and shrieks, strange lights and vile smells.
6. The ghostly images of the astral body can be projected and manifested in distant locations as a disembodied apparition. Sometimes the astral body can travel to other locations and actually appear in holographic-like form to others. This image is not actually considered a ghostly apparition (although it may appear so), but the extension and material form of an individual's Eternal Soul.

   One example of this type of phenomenon was demonstrated in France in 1908 by scholar Hector Durville, who was researching "traveling clairvoyance." In one series of tests, Mr. Durville and a colleague hypnotized a female psychic to facilitate her astral projection to another location. An observer was placed in another part of the house. The experimenter's instructed the female psychic's to touch, hit, or pull at the observer's body while in her astral body. The observer had no idea what to expect, but felt the touches, strikes and pulling on his body by invisible hands. The astral body was even visible as a whitish apparitional figure to some of the people Durville used as witnesses.

EVIL STATES AND MENTAL DELUSIONS

Sometimes the "spirit world" itself can affect the physical body through a subconscious emotional connection with the individual's Shen. These affected subconscious emotions can never be totally repressed because they will continually surface as dreams, visions, flashes of images, etc., and can only be redirected into either positive or negative actions.

Seeing a spiritual entity or ghost is a normal part of Medical Qigong training and is accepted as a natural energetic phenomena; however, disorders such as schizophrenia and psychosis (delusions and hallucinations, etc.) are considered Qi deviations and can be caused by improper Medical Qigong practice. These mental states are known as "evil states" and are difficult to correct. Evil states also include: infatuation, delusions, obsessions, encounters with seductive spirits, demon oppression or possession, and multiple personality disorder. This does not mean, however, that the individuals suffering from these disorders are themselves evil.

INFATUATION WITH THE TREATING DOCTOR

Infatuation is culturally regarded in China as an evil state. This generally refers to the occurrence of deceptions of the subconscious mind during, or after, Medical Qigong healing, which lead to mental derangement. While under the influence of these self-deceptions, patients display intensely amorous affection towards the doctor. As the doctor's energy becomes more intensified within the patients' nervous system, the patients experience a type of biological euphoria which gives way to feelings of unexpressed emotions, suppressed fantasies, and symptoms of eccentric disposition, such as sluggishness, apathy, or elation.

During treatment, patients become extremely open and vulnerable, as their boundary systems completely dissolve under the influence of the doctor's emitted Qi. The Qigong doctor's ability to channel divine light and compassion may easily be mistaken for displays of personal love toward the patients. Sometimes this divine love has never before been experienced by these patients, making them infatuated with the doctor, who is the apparent source of this love. This is an illusion which can usually be dispelled by the doctor's explanation. When the doctor is unable to dispel this illusion by conversing with the patients, this indicates an underlying, often long standing, mental disorder. Such patients should be referred out to a mental health practitioner for professional counseling.

ENCOUNTERS WITH SEDUCTIVE SPIRITS

On rare occasions, individuals who do not have a strong enough connection with the divine may become susceptible to encounters with seductive spirits. These spirits are known as Incubus (male) or Succubus (female), and tend to manifest in dreams. These dreams can range from the romantic and seductive, to rape and sodomy. Although these encounters occur during the individual's sleep, physical sensations are always a large part of the encounter. Often, once the victim has awakened, the physical sensations continue.

DELUSIONS

This is the occurrence of delusions of the subconscious mind during or after Medical Qigong treatment or training, which lead to mental derangement. A delusion is a false belief brought about without appropriate external stimulation (seen most often in psychoses), and is inconsistent with the individual's own knowledge and experience. Patients suffering from delusions display unsociable and eccentric dispositions, sluggishness, apathy, and trance. Some patients may become very disillusioned and depressed. They may lose confidence in life, and think about committing suicide; others may have symptoms similar to those of psychosis, such as persistent visual or auditory hallucinations. These symptoms occur because of preexisting emotional problems, which the patient has not worked through, or has consistently avoided facing.

In all truth, the practice of Medical Qigong does not create these problems; Medical Qigong can, however, occasionally allow patients to become extremely sensitive to the preexisting unresolved problems that continue to surface into their subconscious mind due to their overactive denial system.

SPIRIT DEMONS

An encounter with the phenomenon known as demon oppression or demon possession rarely occurs during clinical practice, but never the less, it must be addressed. Demon oppression or possession describes the state of mind attributed to the patient who has come under the influence of the "spirits of the objective world." These spirits of the objective world (also called "fallen angles," "spiritually hostile forces," or "demons") can either influence, or inhabit, a patient's thoughts and body. They can affect anyone who becomes open to their influence. Any practice or situation which opens the individual's subconscious mind may increase the person's vulnerability to such states. Spirit demon oppression or possession can occur from:

- misguided or improper use of meditations,
- misguided or improper use of hypnosis, or self-hypnosis,
- indulging in the practice of astral projection,
- meditations that cause adverse psychological side-effects,
- general anesthesia,
- drug and alcohol abuse,
- indulging in the practice of channeling, or Ouija board playing,
- deep emotional or physical traumas,
- indulging in the practice of deviant psychosexual encounters (Tantric sex with the wrong individual)
- exposure to places that are haunted by evil spirits when the individual is very stressed out, or deeply fatigued and depleted, and
- indulging in the practice of "black magic."

When the patient's connection to their Shen becomes extremely Deficient, to a point that they surrender their Hun, the patient's Hun may leave their body and wonder, making them vulnerable to the possibility of spirit oppression or possession.

One of the first things a demon seeks, once it enters a body, is to fulfill its sensual appetite (be it sexual pleasures, rage, or emotional addictions, etc.). A disembodied spirit lacks the ability to feel physical sensations due to the disconnection with its bodily senses (touching, feeling , smelling, tasting, etc.). Thus, it can only satisfy these cravings through someone else's living body. Consequently, if a patient is disassociated from the divine center core, or has little or no boundary system, a condition of oppression or spirit possession may occur. In China this condition is commonly described as, "the Spirit (Hun) leaves and the demon enters, uniting with the body's Po" and must be rectified with spiritual atonements and soul retrieval.

Although such encounters in the clinic are rare, it is still important to recognize and treat these conditions when they occur. While treating patients in China, I was fortunate enough to share my experiences with demon oppressed or possessed patients with certain doctors at the Xi Yuan Hospital. It was refreshing to talk with another doctor about this particular subject, as in America it is difficult, if not impossible, to find other individuals who can discuss such a topic in a clinically sound manner.

In understanding the phenomenon of spirit demons, it is important to realize that in the clinics in China, encounters with these types of spirits are divided into two categories: Projected Spirit Demons from the Subconscious Mind, and Actual Encounters with Spirit Demons.

PROJECTED SPIRIT DEMONS FROM THE SUBCONSCIOUS MIND

Sometimes through meditation, the subconscious mind will release visions into the conscious mind in the form of ghosts, phantoms, spirits, demons, etc. These energetic visions are subjective projections of part of the self, released through emotional arousal. Each color image represents certain feelings and repressed memories within the organ which the subconscious mind is trying to communicate to the conscious mind. This also applies to dreams and nightmares in which the patient is continually being haunted by certain spirits, demons, or monsters. For example:

- A green demon represents the deeply suppressed emotions of anger, rage, and irritation which are connected to the Liver.
- A red demon represents the deeply suppressed emotions of excitement, panic, and even anxiety which are connected to the Heart.
- A yellow demon represents the deeply suppressed emotions of worry and pity which are connected to the Spleen.
- A white demon represents the deeply suppressed emotions of sadness, grief, sorrow, guilt, anxiety, distress, and heartache which are connected to the Lungs.
- A black demon represents the deeply suppressed emotions of fear, paranoia, horror, panic, and terror which are connected to the Kidneys.

The patient's subconscious demonic images may tend to be culturally influenced. Each patient may find himself or herself observing something quite different, but with the same underlying emotion. In modern Western culture, individuals are more likely to dream of thugs and assassins that are threatening and pursuing them, as well as monsters.

ACTUAL ENCOUNTERS WITH SPIRIT DEMONS

These spiritual encounters have an objective reality, as they are true spiritual entities which exist outside an individuals thoughts or feelings. For obvious reasons it is extremely important for Medical Qigong doctors to have a strong spiritual foundation. Without a powerful connection to the divine, the doctor risks absorbing the patients' pathogenic Qi, and may become more vulnerable to "evil states" that may place him, or her, at the mercy of spirit demons who seek to dominate, control, and feed off of the negative emotions of the human body.

The spiritual state of the energetic field controls the emotional field, the emotional field in turn controls the mental, and the mental controls the physical field. This is the spiritual template used in the Medical Qigong clinics in China for initiating the energetic laws of dominance and control over the body's tissues. When there is a spiritual interaction involved, the emotions manifest the outcome. There are two ways a spirit demon can attach itself to the body; through spirit oppression, or through spirit possession.

1. Demon or spirit oppressed patients tend to first have had severe emotional trauma, resulting in extremely low self-esteem, depression, fear, rage, and so on (these are the emotional energies on which the evil spirit feeds). The spirit demon becomes attached to the patient's second field of Wei Qi, drawing energy and sustenance from the patient's chronic release of negative emotions.

   These spirit demons are external energy feeders. They are spiritual parasites which provoke and increase negative emotional reactions. Once the patient discharges these negative emotions into their second field of Wei Qi, the spirit demons then ingest and absorb this energy as food. If, for any reason, the spirit demons leave or abandon the patient's second field of Wei Qi, they will, like a predator, continue to look for their next feeding ground. This is why, in certain families, oppressive dynamics are passed from generation to generation, due to certain congenital spiritual patterns.

2. Demon or spirit possession occurs when the patient's Spirit / Mind (i.e., the focused power of their consciousness) becomes too attached to something; it may leave the patient (becoming a wondering spirit) and not return. The vacuum that is left by the original consciousness leaving its residence can become filled by the spirit of other beings. These alien energies tend to enter the patient's body, replacing their original consciousness (Yuan Shen), which is now lost. The resulting state is one of demon or spirit possession. Because the spiritual quality and stability of the patient's mind is lost, the patient will become increasingly dependent on the alien spiritual consciousness for orientation towards others, and the outside environment.

When either a demon or spirit entity possesses a patient's body, it tends to be extremely territorial and protective. This is because, as the demonic spirit surrounds the very core of the patient's energetic body, it initially has a strangle hold on the patient, feeding and distorting the patient's reality. This results in wild mood swings, and sometimes, in displays of extraordinary intuitions, perceptions, and incredible physical powers.

In most cases the spirits causing the pathology are not of a high order of intelligence; they are more on the level of spiritual bacteria, viruses, or parasites. In some cases, however, spirits can be of a high order of demonic intelligence, such as depicted in the movie "The Exorcist."

EXAMPLES FROM PERSONAL CLINICAL EXPERIENCE

Through my personal clinical practice, I have treated several patients who were actually demon possessed, or demon oppressed. One example occurred in 1992, when a patient came to me with a tumor located on her pituitary gland. The Western doctor who was treating her had suggested that she receive immediate surgery, so as a last resort she came to me.

Knowing that she had written several books that were currently being used in a major university as a standard text, I was quite impressed and even intimated by her reputation as an intellectual. She, being the mentor of several of my students who were psychotherapists, did not ease my apprehension. Although she was elderly, she was very strong willed, and quite opinionated. She requested that I lower the treatment table to the floor before starting the therapy. I thought that this was a little strange, but complied never the less, and so we began the initial treatment with the patient lying supine on the table, placed on the floor.

I always begin each treatment with a connection or "hookup" to the divine After connecting with the divine, the energy of the treatment room changes, as the room's energetic field transforms into a divine healing energetic field. Immediately upon "hookup" something in the patient shifted. Although her eyes were closed, the patient began to thrash about on the table like a wild animal, literally snarling, and gnashing her teeth. The power radiating from her body was quite incredible. I knew and believed, that if she wanted to, this elderly, frail woman, could now lift my body up, and toss it across the room like a rag-doll. I was quite aware of the supernatural power and spiritual demonic form that I was now facing, and knew that my 27 years of martial arts training would not serve me in this situation. To my advantage, I had previous exposure to such spiritual demonic phenomena, and had been successful in dealing with demon possessed individuals. Believing that the current healing work was part of my ministry and life purpose, and having participated in several demonic exorcisms, I was aware of the procedure which needed to transpire.

I began dredging and casting out the demons

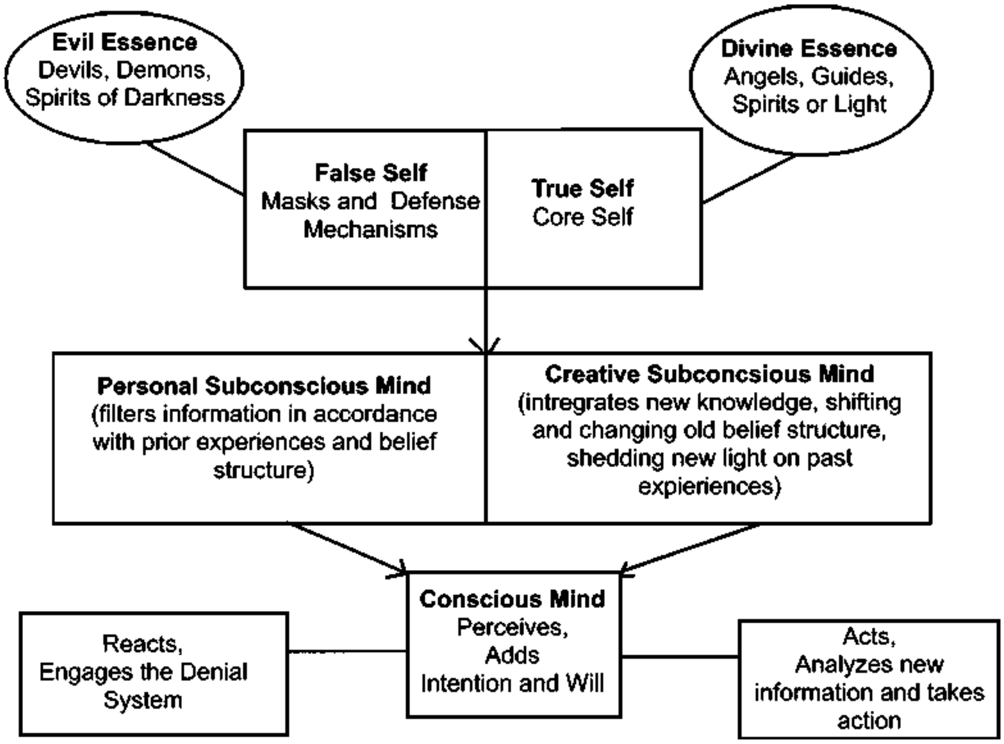
Evil Essence
Devils, Demons,
Spirits of Darkness
Divine Essence
Angels, Guides,
Spirits or Light
False Self
Masks and Defense Mechanisms
True Self
Core Self
Personal Subconscious Mind
(filters information in accordance with prior experiences and belief structure)
Creative Subconcsious Mind
(intregrates new knowledge, shifting and changing old belief structure, shedding new light on past expieriences)
Conscious Mind
Perceives,
Adds
Intention and Will
Reacts,
Engages the Denial System
Acts,
Analyzes new information and takes action

Figure 19.2. All human beings are subject to divine and evil spiritual influences. These influences can affect the subconscious and conscious mind, as well as the sensory perceptions and physical performance.

one by one, claiming each portion of her body for the glory of Christ (my divine authority). This procedure continued for six weeks, each treatment involved reclaiming certain areas of her body. Beginning at her head, I continued treatment down the patient's body. On the sixth week, when I reached her Lower Dantian, the last spirit demon left her body, and she immediately curled up into a fetal position and began to cry. Her Hun then returned to her body, and for the first time in many years she could feel loving emotions again.

It is important to note, that while the treatments were being administered, the patient was not consciously aware that anything out of the ordinary was happening. She had been so completely detached from her body, that she had no conscious recollection of her own emotional-spiritual field of existence, or its affect on her physical body. She had literally hid within the confines of her intellect, where she felt safe. She was also not aware of being demon possessed, or even that her body thrashed about during each treatment. In effect, she had disassociated from her Yuan Shen during childhood, and thus became dissociated from her emotions. After six weeks of treatment her brain tumor had completely dissolved. But, far more remarkable, was the change in her personality, for she became both friendly and courteous.

OVERVIEW

As mentioned before, once the Qigong doctors' Shen has reached a high state of sensitivity, they are able to transcend their conscious perceptions to sense, observe, and communicate with the spiritual world (Figure 19.2). This ability to see disembodied spirits or ghosts is a normal part of Medical Qigong training and a natural part of energy observation.

These spiritual presences can be seen or felt for either brief moments in time (usually as a hovering image of light, or a dark shadow), or for extended periods of time (usually an encounter). Some of these spiritual entities can also take the form of animals.

Some doctors view these energetic entities as either individuals who are astral-projecting (from the past, present, or future), or forms of energy that can either be ascended spirits of light (divine guides or angels), or fallen spirits of light (demons and hostile forces). In any case, encounters with such entities of the spiritual world should not startle or scare the Qigong doctor, but be expected as a part of the existence of the spiritual domain.

A final important note to make is that, although demon possession is real, the location, or type of the patient's disease, has nothing to do with demonic possession. Demonic possession is related to the disassociation from the Yuan Shen and the absence, or suppression, of the patient's Hun. It is not caused by obstructed tissue formations or physical disease.

After helping the patient uncover any severe emotional trauma, the Qigong doctor should assist the patient in "soul retrieval" meditations, in order to return the patient's Hun back into its residence (the patient's physical body) and to facilitate a strong reconnection to the Yuan Shen. During soul retrieval meditations, patients sometimes talk about viewing certain traumas from different visual perspectives, such as: watching the incident from above their body, or watching from a different corner of the room. The reason for this change in visual perspective is due to the displacement of the patient's Hun once it leaves the body. Although the Hun are outside the body, they still observe, listen, and record. This accounts also for the numerous cases of astral projection during surgery, as well as during near death experiences. Under anesthesia, many patients travel out-of-body and observe in fine detail what the doctors and nurses are saying and doing. They accurately recall the doctors' and nurses' conversations, and the experience of viewing the operation from above the body. These experiences are becoming more documented in recent years partly due to modern techniques in resuscitation, that have made such experiences more frequent. First hand accounts of out-of-body experiences abound in "near death experience" literature.

Similar experiences have been recounted by many physically and sexually abused patients, who, as children, learned to dissociate from the body, and observe the abuse from the ceiling, or corner of the room. In most cases, these adults, and children, do not suffer from multiple personality disorder. Only in the most severe cases does the personality fracture and become multiple.

MULTIPLE PERSONALITY DISORDER (DISSOCIATIVE IDENTITY DISORDER)

Occasionally the Qigong doctor may encounter a patient with a multiple personality disorder. The condition of a multiple personality disorder is generally due to extreme sexual, physical, and emotional abuse during childhood; in some cases, demon possession may also occur. This is a condition in which the patient has developed a split in their consciousness, resulting in the development of two or more distinct identities or personalities that recurrently take control of the patient's behavior and consciousness, one at a time.

It is believed that the original personality splinters during childhood (usually around the age of 6 or 7 years, or as early as 3) creating several different personalties, or personality states, that are disassociated from the patient's original whole personality. The different personalities created can number as few as two, or upwards of one hundred. They are often referred to as alternate identities or simply 'alters.' These alters, in fact, often themselves give birth to more fractured alters, each with its own gender, age (corresponding to the fracturing trauma) and function (or role). Some alters may be very aggressive and protective, others will be more passive and submissive.

Some alters will have different medical conditions from the others, which inhabit the same body. As the patient changes personalities (alters), the internal and external energetic fields likewise undergo changes. Each personality can differ in age and stage of development. In most cases, the personalities are usually unaware of the words, actions, and feelings of the other personalities, especially at the beginning of the treatment. Each personality can exhibit a different set of symptoms, as well as diseases. One personality may have chronic asthma and allergies, but when switching to another personality, the patient may suddenly display healthy lungs but display nearsightedness or high blood pressure.

Usually one personality becomes a helper to the doctor, while another personality may be psychotic, indifferent, and so on. Some of the alters may have access to highly advanced transpersonal states of consciousness, i.e., a finely honed and active psychic realm which may have provided a haven, or sanctuary during the abuse.

TYPES OF ALTERS

Over time, an alter's role in the patient's personality changes. There is a difference between an established personality which maintains both an internal and external role and just a fragment personality (e.g., washes the dishes only). Also, there will usually be a number of child personalities included, as well as adult personalities. The patient's past history of violence is the best indicator of whether there are dangerous alters. Some of the most prevalent personality types of alters are described as follows.

1. The first personality encountered by the doctor, or psychologist, is usually the patient's host personality, not the patient's original personality.
2. The demonic or spiritual personality type is generally found in very religiously oriented patients.
3. The persecutor personality type is generally responsible for the patient's attempts at self mutilation (a very common type syndrome with multiple personality disorder) or suicide. They tend to be children or teenagers. It is believed that these types of personalities are either interjections of the patient's original abuser, or have evolved from "helper" personalities.
4. The helper personality type is generally responsible for providing the patient with the emotional support to counteract some of the self-destructive behavior of the patient's persecutors.
5. The administrator or obsessive-compulsive personality type is often responsible for allowing the patient to earn a living. They generally appear cold and aloof.
6. The imitator, or impostor, personality type generally mimics other personalities and is responsible for handling certain situations that are difficult for the patient to bear.
7. The autistic or handicapped personality type generally emerges when the patient feels that he or she is under intense scrutiny, or feels that he or she is being controlled or confined. They may manifest autistic or catatonic states or become functionally deaf or blind.
8. The promiscuous personality type is generally responsible for the patient's overactive or uncontrollable sexual urges (e.g., nymphomania).

While treating patients with multiple personality disorder, the therapeutic goal is usually to integrate all the different personalities, merging them together through creative visualization (usually using hypnosis). This is achieved by having the patient imagine each personality and subset "alters" merge into one.

Some patients, however, opt to remain multiple, as merging incurs the "death" of others. In such cases, contracts are made with the different alters to keep each other informed of events occurring during their "coming out," and some of the more responsible personalities keep watch over the more self-destructive ones. Due to the complexity of treating multiple personality disorder, and due to the high potential of suicide in multiple personality cases, it is recommended that the Qigong doctor refer these type of patients out to specialists who are familiar with such conditions and their recovery.

SOUL RETRIEVAL AND MEDICAL QIGONG THERAPY

According to the Medical Qigong perspective, one of the major causes of illness is "soul loss" (the loss of parts or memories of the Eternal Soul). On the spiritual plain, the Eternal Soul is the manifestation and sum total of our energetic associations and the energy of the Wu Jing Shen (see

Chapter 2). The Eternal Soul serves as a middle point between the source and organization of all life-force energy. The Eternal Soul is considered the seat of all emotions, feelings, and sentiments. As a crystal grows around a central matrix, so the body grows around the Eternal Soul. In revue, the Eternal Soul radiates energy in all directions, like the sun, and is responsible for life, health, and the formation and growth of the body. In the scope of energetic manifestations, the Eternal Soul expresses its innate qualities as energetic movement, functioning through the physical form (led by the Shen). It is connected to all parts of the body, and its light or energy is reflected through the eyes via the Hun. It is through the influence of the Eternal Soul that all of the body's energetic processes (organs and organ systems) seek wholeness. It knows exactly what is needed in every situation for survival and health preservation, and takes action via the Po. Whole body consciousness is the main characteristic of the Eternal Soul. Without the process of the Eternal Soul, energy would have no specific direction and remain in meaningless activity.

"Soul loss" is not demon possession, but a spiritual illness that causes emotional, mental, and physical disease. When disturbed, the Hun leave the body and the Mind (Shen) wanders off on its own. When the Hun leave and the Shen wanders off for very short periods of time, it is sometimes called, "dissociating" or "spacing out." When the Hun leave for extended periods of time, it is called soul or spirit travel, which generally happens when the patient is asleep. It is believed that if the Hun completely vacate the body, the patient will die. A coma is one example of a chronic physical state where the Hun have left, but are still attached to the body, through the Hun's connection with the patient's energetic Qi field. When a patient is in a coma, the Hun and Shen abandon their residence, wandering about in the astral plane of existence.

The Hun and Po create the sustaining energy for the Wu Jing Shen, which provides the energy for the Yuan Shen. If the Qigong doctor can retrieve the lost parts of the patient's Eternal Soul (the suppressed memories attached to the Hun and Po), then the individual's Shen can be restored back to harmony and health.

REASONS THE ETHEREAL SOUL LEAVES THE BODY

In ancient times, disease was attributed to the Hun having strayed away, or having been stolen. It was believed that the Hun would leave the body through three main incidents: they could be stolen (through deep infatuation), frightened away, or simply wander off. In addition to soul theft or "rape of the soul," some Qigong doctors relate Hun loss to the interference of Gui (ghosts, demon spirits), as well as other human beings (i.e., a broken heart or severe trauma).

Soul retrieval is an essential part of Medical Qigong therapy; the body and mind without the Hun, or the spiritual connection of the Yuan Shen, remain in a suspended state of spiritual and emotional shock. When the spiritual needs of the Eternal Soul are neglected, the Shen begins to disassociate, and symptoms of addiction to drugs or alcohol, obsessions, violence, or deep depression take over (see Chapter 32).

Because we are born understanding hidden spiritual realities, patients will inevitably succumb to various mood swings; these are due to deep emotional pain and experiences that stay within their memory and touch their Heart. A spiritual life of some kind is absolutely essential for psychological health. Traditionally, the Hun are said to suspend themselves midway between the conscious and the unconscious, and that their perceptual modality is neither the mind's nor the body's, but the imagination's (i.e., the Creative Subconscious).

From a clinical perspective it has been noted that the Hun leave the body during severe emotional traumas (i.e., surgery, the death of a loved one, miscarriages, abortions, incest, molestations, physical and emotional abuse, the end of a deep relationship, and the stress of combat). Whenever we experience trauma, our Hun, which are part of our vital spiritual essence, separates from us to survive the experience and to escape the full impact of the pain (Figure 19.3). When patients be-

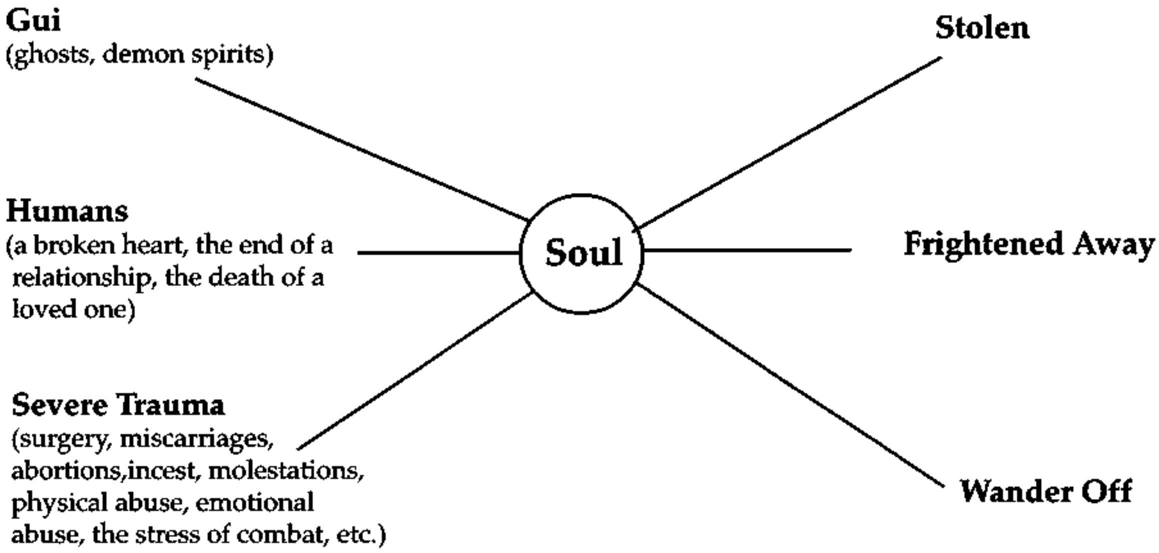
Gui
(ghosts, demon spirits)
Stolen
Humans
(a broken heart, the end of a relationship, the death of a loved one)
Soul
Frightened Away
Severe Trauma
(surgery, miscarriages, abortions,incest, molestations, physical abuse, emotional abuse, the stress of combat, etc.)
Wander Off

Figure 19.3. Reasons for the Ethereal Soul Leaving the Body

come dissociated, aspects of their personality separate from the mainstream of their consciousness, resulting in gaps of memory loss. An extreme example of personality separation occurs in multiple personality disorder.

When a patient has been violated or wounded in some way and remains in an emotionally present state of mind, within the body, the result usually encompasses feelings of anger from past hurts, grief from present pain, or fear of future harm. If an over-stimulation of the energetic organs occurs, it may wound the patient's Shen, causing the patient to close off from the Yuan Shen, to shut off all emotions. This action is similar to that of a sea-anemone, whose natural response (for the purpose of survival) is to immediately close itself after being touched.

THE PURPOSE OF SOUL RETRIEVAL

The purpose of Medical Qigong soul retrieval therapy (called age regression therapy in Western psychology) is to bring the awareness of emotions back into areas of the physical body, allowing the patient to express and release the feelings and reactions that are the cause and symptoms of his or her disease. This will help the patient transform the trapped, or frozen energy (stagnated Qi), back into its energetic potential to be reintegrated with the Shen.

A Qigong doctor will accompany and tend to the patient's Eternal Soul in times of emotional transformation, including times of crisis, illness, and emotional death and rebirth. This allows the patient to remain in the energetic present time mode. This also results in the relief from symptoms of illness and the return of the patient's personal power, creating the potential for fulfillment in work and rewarding relationships. Some patients may come to the realization that they must quit their job and find another that is more spiritually rewarding. They may also find themselves discarding old friendships and making new ones, as their energetic, emotional, and spiritual fields undergo growth and change.

WOUNDING AND CLOSING THE SPIRIT

The body's Yuan Shen is the main expression of the Eternal Soul as it moves from its inactive state to a state of dynamic activity. The Yuan Shen is the medium or active environment of the Eternal Soul. The Yuan Shen also expresses its energetic influence through a continual interchange between the personal-self and the higher-self through the act of breathing. The Eternal Soul expresses itself specifically through the Yuan Shen, and the Yuan Shen manifests itself through the physical body.

If a person's Yuan Shen becomes wounded (feels attacked, violated, taken for granted, etc.), it begins to close itself off from its environment. When the Yuan Shen "closes", the emotions begin to shut down, and the Hun wander. Depending upon the patient's set patterns, this entire process can occur within seconds. Once the Hun wander, the acquired (analytical) Zhi Shen and surviving Seven Corporeal Souls (Po) take over the body, and the patient responds like a machine (i.e., the patient can still think and function; reflexes remain intact, but he or she is void of interpersonal feelings). Patients whose Shen have closed down cannot make emotional attachments, and live in a state of perpetual hunger. These patients go through three stages of isolation before their spirit leaves.

1. The first stage involves anger and grief.

   Once a patient has been hurt (before his or her spirit begins to close), his or her natural survival energy immediately affects the Liver and Lungs, causing emotional reactions that are a natural response to feeling attacked. Hurt, sadness and anger are natural responses to feelings of rejection, betrayal and abandonment. These emotions reflect a spiritual, emotional and mental protest to feeling insignificant, excluded, and dis-empowered.

   a. The patient's congenital emotions of kindness and compassion, stored within their Liver, instantaneously shut down as the Liver Qi expands violently outward. This energetic reaction gives way to releasing the patient's acquired feelings of anger and rage, stored within the Liver. The energy of anger and rage is projected outward to protect the patient's emotional boundaries. If the patient has deep seated emotional problems, however, the energy can also implode inward creating self-hate.
   b. If this protective emotional reaction is not effective in obtaining the needed results (being heard, honored, valued, etc.), the patient's energetic field rebounds inward affecting the Lungs. Once the congenital emotions of integrity and dignity stored within the patient's Lungs become effected, the Lung Qi becomes obstructed, and the acquired emotions of grief and sorrow are released. The energy of grief and sorrow is encapsulated internally to encase the patient's wounded spirit.

2. During the second stage anger and grief transform into depression and despair.

   As the patient begins to lose hope, the Shen begins to close, causing two reactions.

   a. The patient's Liver Qi becomes obstructed and stagnant. This in turn transforms the emotion of anger into depression;
   b. At the same time the patient's Lung Qi transforms grief into despair.

   The patient slowly begins to feel helpless, hopeless, and empty inside. Because of the increased depression and despair, the patient becomes physically, mentally, emotionally, energetically and spiritually tired, and gives in to feelings of insignificance, unfulfillment and loss of control over his or her life.

3. During the third stage depression and despair transform into indifference and detachment.

   In the final stage, the patient has lost all hope of resolution and gives up entirely. The Shen closes down, and the Hun wander. The patient feels emotionally detached from people, the surrounding environment, and his or herself. The patient wanders around "dead to the world," isolated, internally unfilled, and alone. This creates what is known as a "depersonalization disorder."

   A depersonalization disorder involves the persistent, recurring episodes of depersonalization, characterized by a feeling of detachment, or estrangement from one's self. The individual may feel as if he, or she, is living in a dream, or movie, and responds to life like

Three Stages of Emotional Transitions before the Spirit Closes

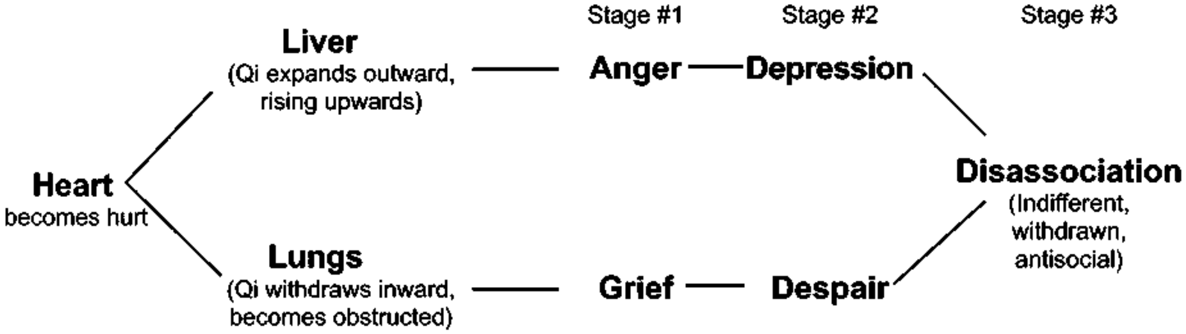
Stage #1
Stage #2
Stage #3
Liver
(Qi expands outward, rising upwards)
Anger
Depression
Heart
becomes hurt
Disassociation
(Indifferent, withdrawn, antisocial)
Lungs
(Qi withdraws inward, becomes obstructed)
Grief
Despair

Figure 19.4. There are three stages of emotional transition the patient must experience before their Shen closes and their denial system engages. After these stages are experienced the patient's defense mechanisms take over.

a robot. They may also have the sensation of being an outside observer (to their mental processes, body, or specific parts of the body). This feeling is often accompanied by sensory anesthesia, a lack of affective response, and a feeling of lack of control over their actions (including speech).

Voluntary induced expressions of depersonalization or derealization that are meditative, or cultural trance practices, should not be confused with this involuntary disorder. It is believed that approximately 50% of the population may experiences a single, brief episode of depersonalization, usually precipitated by severe stress. Approximately one-third of individuals exposed to life-threatening danger incur a brief episode of depersonalization. Depersonalization episodes may be accompanied by visual distortions, such as the shape or size of objects, and other people may appear unfamiliar or mechanical. Other common associated features include: anxiety or depressive symptoms, obsessions, somatic concerns, and time distortion. Depersonalization and derealization are some of the symptoms of panic disorders (not due to drugs or medication), which are usually caused by long-standing anxiety and stress. Panic disorder can also be brought on by emphysema and asthma due to the difficulty in breathing.

When the Eternal Soul becomes wounded, the first two stages in closing down the Yuan Shen are caused by the patient trying to repress feelings of anger (stored within the Liver), and sadness and grief (stored within the Lungs). This reaction is incurred because the Liver stores the body's Three Ethereal Souls (Hun), while the Lungs store the body's Seven Corporeal Souls (Po). These two emotions are expressions of the patient's natural protest against the lack of love that their Yuan Shen (which encompasses or contains both the Hun and Po) is experiencing (Figure 19.4).

HEALING THE SPIRIT

The first stage of healing in soul retrieval is to open the patient's Yuan Shen by giving them hope and encouragement. Hope helps the patient's perspective to be positive and restores meaning to life. Once the patient has closed his or her Yuan Shen, it cannot be opened through argument or logic. Instead the Qigong doctor must use a gentle, soft tone of voice expressing compassion and high regard for the patient. The patient must feel honored and valued by the doctor to feel safe enough to release the hurtful and negative emotions.

Praise must be used carefully and gradually, since some patients may be as yet, unable to accept a higher concept of themselves. Too much praise may cause the patient to feel "mocked" or "lied to", given his or her very low self-esteem. This type of healing only happens when the patient's spirit is engaged, stimulating the inner emotions and not just the intellect.

Once the patient's Yuan Shen begins to open, it is important for the doctor to increase the patient's hope for interpersonal relationships. This is initiated by using silent mental affirmation, or prescribing positive affirmations in conjunction with vivid visualizations. The patient's imagination can reengage his or her feelings and return them to the body's organs and tissues. Instilling hope and giving encouragement allows the doctor's influence on the patient to be increased.

As the patient's Shen begins to open, it is important for the Qigong doctor to take the time to energetically and emotionally travel with the patient back into the memories of the past, to encounter the precise moment in which the patient's Hun left the body. The choice to return the emotional spirit back into the body must come from the patient's desire to retrieve his or her original feelings, no matter how painful the process. The patient must be ready to reexperience and release these feelings. It is important that the patient's "adult observer" be present in retrieving these painful memories. The "adult observer" refers to the patient's ability to suspend his or her judgemental attitude towards his or herself, and to observe the events that are being released with compassion (i.e., as if they were happening to a friend to whom loving support can be given). This allows the patient to experience true healing with the help of a compassionate companion (his or herself).

Tears often come as the memory of the original trauma releases from the patient's body or flows back into the Mind. Many patients notice that after these experiences their dreams change significantly, becoming more vivid. Also, their life environment as well as their relationships tend to change drastically.

EMOTIONAL REBIRTHING PATTERNS

Often, when patients experience deep emotional and spiritual traumas, they also experience physical sensations which reflect the trauma of their biological birth. These psychophysical sensations are due to the physical, energetic, and spiritual connection each individual has with the maternal host while inside the womb. Although these experiences are natural, they can sometimes be disturbing and even frightening to patients unfamiliar with deep emotional and spiritual transitions. The following are but a few examples of physical rebirthing patterns, and the emotional reactions related to these types of spiritual transformations.

- In patients born with the umbilical cord wrapped around their neck, each time they experience deep emotional and spiritual changes it will feel as if they are being strangled.
- In patients born through C-section, each time they experience deep emotional and spiritual changes it will feel as if they are being torn asunder.
- In patients born through the natural birth cannel, each time they experience deep emotional and spiritual changes it will feel as if they are falling through space with nothing to support them.
- In patients born breech, each time they experience deep emotional and spiritual changes it will feel as if they are being crushed in by all sides and suffocated.
- In patients with mothers who were heavily sedated, each time they experience deep emotional and spiritual change they will feel mentally disoriented and unstable.

SOUL RETRIEVAL MEDITATION

The following comes from my clinical experiences in dealing with patients who are very eager to heal, but reluctant, or unable to tolerate rapid change within themselves. The patients' family dynamics, as well as the quality of traumatic and nourishing experiences they received while growing up, play an important role in shaping their individual personalities. It is an accepted fact within modern psychology that an individual's present emotional life is shaped by the events occurring in the "formative" years (especially the years before the individual can articulate their thoughts and feelings). These experiences sink deep into the individual's psyche, hidden from consciousness through repression and denial. This repression is generally due to the inability to understand and express in words the traumas they suffered.

This particular Shengong (which literally means working with spirit) technique involves creative visualization (in a very relaxed, quiescent state of mind), and is used to gain access to the patient's unconscious past emotional experiences. The goal of this meditation is to allow the patient to free him or herself from the repressed emotional material which systematically surfaces during daily life. These programed emotional responses systematically affect the patient's behavior, causing neurotic symptoms, slips of the tongue, associated dreams, and fantasies.

Many patients believe they are ready for this experience, but in reality they are not yet able to deal with such intense pain. These patients are likely to unconsciously choose the least painful memories in order to test how the doctor will handle their feelings, or fail in the attempt to retrieve the lost parts of the soul. Gradually, as trust grows, these patients will be able to access more painful and relevant memories (this is similar to peeling an onion layer by layer). Sometimes patients try very hard to access the memories and obtain nothing, or very little information, leaving them feeling frustrated. This frustration must then be processed and resolved.

The doctor must instill hope and dispel this frustration by explaining to the patients that they have made an important step forward by seeding the unconscious mind with the request, or concept, that past memories must be recalled. In time, this seed will sprout and give forth results. It just needs a little time to incubate. It is extremely important that the doctor never show disappointment or take away the patients' hope.

In such cases, the patient must be reassured that their higher-self (the observer self) senses that the timing is not right. There may be current issues or events, that must first be worked through for a successful soul retrieval to take place. To have done so now would have resulted in more trauma, or more precisely, re-traumatization. They may need to strengthened and develop more self-trust before trying again. Counseling may be recommended, and in time they will succeed.

Some patients may need several sessions before feeling comfortable enough with just entering and surfacing from trance states, before undergoing soul retrieval. It may be necessary to first retrieve pleasant memories that build self-trust and self-confidence before re-attempting soul retrieval (i.e., learning to fly a kite, etc.).

The Qigong doctor should have some training in hypnosis and hypnotic trances before giving this type of trance inducing meditation to their patients. The doctor should know what to do and be able to identify whether or not the patient is able to connect with their Yuan Shen or "higher self." As the doctor induces trance in the patient through this meditation, the following changes in the patient may be noted:

- a change of color in the complexion,
- a relaxation of the facial muscles,
- a change in breathing patterns, and
- sporadic eye movement (behind closed eyes).

If none of the above signs are noted, the patient may need more time to relax deeply before continuing with the meditation. Suggestions of imagining and feeling the body's tissues melting into the Earth may be given, along with suggestions of taking deep relaxing breaths to calm the patient's Shen.

While performing this meditation the patient may experience all the sensory, emotional, and physical sensations of these repressed memories, causing specific core memories (that have been repressed from infancy and childhood) to be brought to the surface and relived. As the patient experiences these memories, his or her facial expressions and demeanor, will often change, regressing to reveal the appropriate age and period of time during which the original trauma occurred.

This particular method of soul retrieval is divided into two stages. The first stage involves locating the body's three storage chambers of past, present, and future emotions. The second stage focuses on accessing the location of, and coming in contact with, the patient's unconscious repressed emotional traumas.

This meditation must first be understood in depth before it is given to the patient. It is very complex, and is not meant to be simply read through to the patient. Also, the doctor should set aside sufficient time to process what the patient experienced after ending the meditation.

STAGE 1. ACCESSING THE ENERGETIC CHAMBERS

The purpose of this meditation is to locate the patient's three storage chambers which access his or her past, present, and future emotions. This meditation also allows the patient to acquire a starting point, or outlet, for removing these painful experiences.

In this meditation, the patient first concentrates on feeling his or her entire body. Begin by having the patient sit comfortably in a chair, both feet on the floor, hands resting on the thighs, eyes closed, tongue on the upper pallet, and breathing naturally through the Lower Dantian. It is important that the patient relax, feel, and experience the body completely. The patient should also be encouraged to accept and feel everything as one unit, to connect with the Yuan Shen or "higher self."

The Storage Chamber of Past Emotional Memories

In order to find the chamber of past emotional feelings, the patient begins to focus attention on remembering and feeling a painful childhood experience. Once the patient has retrieved the memory and is experiencing the painful trauma, he or she points to the tissue area of the body where the feelings are located. This area is considered to be the current storage chamber, where the patient stores the energy of past emotional memories.

The storage areas change over a period of time, given the fact that the patient continues to process emotions throughout life. Each emotional healing causes the previous energetic holding pattern to die, and a new pattern is to be born within the patient's energetic structure. The chamber of the patient's past usually coincides with a tissue area where the patient is already experiencing disease or discomfort.

The Storage Chamber of Present Emotional States

In order to find the chamber of the present emotional states, the patient focuses attention on feeling the doctor's voice within his or her body. As the doctor continues to speak, the patient feels where the sounds are resonating. Once the patient experiences and locates this tissue resonation, he or she points to that area. This area is considered to be the current storage chamber, where the patient stores the energy of present emotional states.

The Storage Chamber of Future Emotions

To find the chamber of future emotional states, the patient focuses on a specific project that he, or she, is planning to do in the near future. Once the patient experiences slight physical sensations, he or she points to the area where the feelings are located. This area is considered to be the current storage chamber, where the patient stores the energy for future endeavors and future expectations. The patient's current ambitions may change as the patient begins to heal from past emotional wounds and makes life-style changes.

STAGE 2. RETRIEVING THE LOST MEMORIES

The final part of this meditation explores the patient's unconscious feelings and reprograms the destructive subconscious patterning. It is used to access the repressed memories stored within the patient's individual Personal Subconscious, which is hidden from through their denial or defense mechanisms. Sometimes after practicing this meditation, patients will not only gain access to the repressed memories of their infancy and childhood, but will also connect to their life within the womb, as well as their birth-process.

1. Begin by having the patient sit comfortably in a chair, both feet should be placed on the floor, hands resting on the thighs, eyes closed,

tongue on the upper pallet, anal sphincter lightly closed. The patient should breathe naturally, expanding the Lower Dantian with each inhalation, contracting the abdomen with each exhalation.

2. The patient focuses attention on the area of the body where the energy of the past emotions is stored. The doctor directs the patient to imagine opening the chamber of the past emotions, allowing the energy, like steam, to release itself from this area. This steam is gathered to the left side of the body. Once all of the energy has been released, it begins to form into a cocoon shaped cylinder of past emotional memories.
3. Next, the patient focuses attention on the area in the body that stores the energy of the present emotions. The patient imagines releasing this energy from out of the body like steam, creating a second cocoon shaped cylinder. The patient places this second cocoon in front of the body, next to the first cylinder.
4. The patient then focuses attention on the area in the body that stores the energy of future emotions. The patient imagines releasing this energy from out of the body like steam, forming a third cocoon shaped cylinder. The patient places this third cocoon to the right side of the body, next to the second cylinder.
5. Next, the patient imagines that all three separate energetic cocoons are joined together. The energetic cocoons are to be connected end to end, in a straight line, with the past on the left, the present in the middle, and the future on the right side of the patient's body.
6. The patient then focuses attention on his or her Taiji Pole, imagining the energy of the Upper Dantian descending and combining with the energy of the Middle Dantian (where the Eternal Soul is stored). Next, the patient imagines the collected energies of the Upper Dantian, Middle Dantian, and his or her Eternal Soul, descending and combining with the energy of the Lower Dantian. The patient uses the collected energies of the Three Dantians to create an "energy bubble" within the Lower Dantian to surround the Eternal Soul. While focusing on the center of the energy bubble, the patient imagines the Eternal Soul transforming into a small energetic form of his or herself.
7. Encapsulated within the energetic bubble, the small energetic form of the patient's self manifests within the physical body. The patient's Eternal Soul, now in small energetic form, begins slowly ascending the Taiji Pole, leaving the body through the Baihui GV-20 point. As the patient's Eternal Soul leaves, it is still connected to the physical body through a silver energetic cord.
8. The patient's Eternal Soul floats over the three energetic cocoons, hovering over the cocoon shaped cylinder on the left side of the body (to access the past emotional memories). Next, the patient observes the black, oval shaped cylinders, imbedded within the structures of the energetic cocoon of past emotional memories. These dark cylinders are energetic emotional clusters. They contain the memories of hurtful feelings and specific times in the life of the patient when he or she felt humiliation, degradation, shame, rejection, or emotional deprivation.
9. The patient descends into the energetic cocoon of past memories and allows the Eternal Soul to find the particular cylinder which currently needs addressing. In choosing a specific area to work on, the patient's Yuan Shen automatically selects the most relevant material from the patient's unconsciousness.
10. The patient then removes his or her Eternal Soul from the energetic cocoon, bringing with it a specific dark cylinder. The cylinder is removed from its webbing within the walls of the energetic cocoon and placed in front of the hovering soul. The cylinder is still connected to the energetic cocoon through an attached energetic cord. Through vivid concentration the patient focuses on the color of the cylinder emitting light from the Yin Tang (Third Eye) point of the little energetic self. This projected light pulsates and changes the dark cylinder, from black to white. As the cylinder pulsates and changes its color, it begins

to open and reveal the lost memories and unresolved feelings. These memories of emotional traumas are brought to the surface. Some of these experiences may be memories of events where the survival or integrity of the patient's physical body was threatened.

11. Next, the patient connects with the divine (God or "higher power") and begins the process of feeling, expressing, understanding, forgiving, and letting go of the emotional charge attached to the memories. This process of emotional purging is connected to the emotions stored within the patient's Five Yin Organs, and are expressed through the following stages:
    a. Anger, blame, and resentment are stored in the patient's Liver.
    b. Hurt, pain, sadness, and disappointments are stored in the patient's Lungs.
    c. Fear and insecurity are stored in the patient's Kidneys.
    d. Remorse, regret, and responsibility for keeping the painful memories and feelings alive are stored in the patients's Spleen.
    e. Love, understanding, and forgiveness are stored in the patient's Heart.
12. As the emotions begin to release themselves from the energetic cylinder, the energetic debris transforms into a vaporous cloud. After understanding and resolving the emotions contained within the energetic cloud, the patient divides the released energy into either a receiving or discharging cloud of energy.
13. At this point, the patient focuses on the energetic cloud and begins separating the pain and hurtful memories from the knowledge and wisdom gathered from the experience.
    a. First, the patient places those memories gathered from the knowledge and wisdom of the experience on the right side of the room, imagining them as sparkling gold, silver, and white light energy.
    b. Second, the patient places those memories gathered from the pain and hurt of the experience on the left side of the room, imagining them as thick, dark, black energy.
14. The patient focuses on the right side of the room, imagining this energy as bright, pure, and illuminating. The patient then inhales, absorbing the knowledge and wisdom gathered from the past experience into his or her small energetic body. Since this is actually the energy of the patient's soul, the patient's body immediately feels the emotion, energy and spirit becoming empowered, as it can now review these past experiences without the feelings of hurt, pain, and judgment.
15. Next the patient focuses on the left side of the room where the feelings of hurt and pain are stored. The patient imagines this hurt and painful energy as a dark, turbid cloud. Because the cylinder has a direct life line connected to the patient's past, the energetic cord attached from the cylinder to the energetic cocoon must be severed. Once the cord connecting the cylinder to the patient's past is severed, the patient is then truly released from the spiritual and energetic connection of the memory. This is initiated by imagining an angelic being, with the "sword of truth," severing the energetic life line connected to the patient's cylinder. The sword of truth represents the patient's decision to regain his or her power by deciding to no longer suppress and carry the painful memories. Once the energetic cord is severed, it causes the cylinder to vaporize, forming an energetic cloud. The patient then imagines severing any last connection or attachment to these toxic emotions, and releases this cloud of hurt and pain, visualizing it floating up out of the room, through the ceiling, out into space.
16. The patient then imagines that God (the patient's divine healing power) is seated in the Heavens, with ripples and pulsations of love and mercy emanating from the center of this divine presence. The patient visualizes one of these waves of love and mercy form the divine descending towards the Earth. As the wave of love and mercy touches the as-

cending dark cloud of hurt and pain, it immediately explodes and destroys the dark cloud. The moment that this dark cloud explodes, the patient sees the cloud's negative energy transforming into blue crystal drops of divine healing energy.

17. Next, the patient visualizes these bright, fluorescent blue, drops of pure, clean energy descending from the heavens and pouring down like rain on his or her small energetic self, saturating it completely. This energy immediately transfers into the patient's physical body. The patient inhales and absorbs this energy flowing into every tissue of the body, allowing the divine healing light to penetrate and radiate throughout his or her entire being.
18. Next, the patient observes the changes created within the three energetic cocoons. The energy in each cocoon will have shifted slightly, revealing an energetic alteration of the patient's present state of mind, and a change in the patient's future. By transforming the past, the patient has altered the present and initiated a change in the future course of events.
19. Finally, the patient returns his or her energetic bubble (containing the small energetic self) back down through the Baihui point into the Taiji Pole, drawing the energy back into the Lower Dantian. From the Lower Dantian the patient's soul energy will return back into the Middle and Upper Dantians, as well as the surrounding tissues. Finally, the patient reabsorbs the energetic cocoons back into the body, and returns to a quiescent meditation.

THE DOOR TO THE CHAMBER OF MEMORIES

A simple visual meditation can also be used to access the patient's suppressed past memories. This exercise is a form of guided meditation, in which the doctor leads the patient into the imagined hidden parts of his or her own soul.

Begin by having the patient lie supine. Ask the patient to close the eyes and imagine walking down a flight of stairs. Lead the patient through three separate doors, descending three separate staircases (this is performed in order to arrive at the patient's "chamber of memories").

After entering through the last door, the patient descends one final staircase until he or she reaches the chamber of memories. This is where all of the memories of the patient's past traumas have been kept, hidden deep within the subconscious mind.

Inside the Chamber of Memories, the patient discovers dark cylinders of past emotional pain. Each cylinder represents a specific trauma experienced and stored within the patient's energetic field. The patient then makes a decision to heal and free the body of the enormous energetic field used to suppress certain memories. To heal and dissolve the energetic cyst, the patient follows the same process of eliminating past memories (repeating section 8 through 15 from the previous meditation).

Once this process is completed, the Qigong doctor guides the patient back up the flight of stairs, through the three doors, and back to consciousness. Because these meditations have a profound affect on the patient's constitution (affecting their physical, mental, emotional, energetic, and spiritual nature) it is important for the patient to be referred to a psychotherapist while undergoing this type of Qigong therapy to further facilitate the patient's healing transformation.

FINDING THE LIFE PURPOSE

Another aspect of Soul Retrieval consists of an exercise that assists the patient in finding his or her purpose in life. Finding the purpose of the patient's life encourages interpersonal healing, and fills the empty space deep within the patient's heart. This in turn increases the healing potential by providing the patient with hope.

1. The first step is to ask patients to write down on paper what is was they wanted to be when they were very young. Have them list the people that they imagined or fantasized being while they were still children (e.g., a super hero, doctor, fireman, etc.). This exercise allows patients to access their emotional past, and to feel and recover pleasant, lost memories, and phantasies.

2. Next, have the patients link all the peak experiences of each imaginary character (e.g., a doctor and fireman possess power to save lives, and can be considered heroes, as they are both admired for their skills, etc.). This knowledge informs patients regarding the underlying nature of each imaginary character and may also reveal hidden talents, passions and ideas.
3. Then, have the patients write down a list of their personal priorities. This list should contain the patients' present state of active priorities, and their current value system.
4. Then, have the patients write down their own "code of conduct," i.e., their conscious value system. The patients' code of conduct represents the ethical standards by which they strive to act. This code of conduct influences feelings, thoughts, and beliefs. It affects the patients' personal priorities and reflects the quality and nature of their interpersonal relationships. It is from this personal code of conduct that they accept, or condemn, themselves and others. By physically writing down these moral conducts, patients can emotionally connect with new found feelings of moral integrity, as well as a establish a stronger connection with their True Self.
5. Finally, have the patients write down a "life purpose" statement, linking together the collected emotions gathered from the peak experiences of their childhood and their code of conduct.

It is important that patients know and feel that what they are writing is a true and accurate account of their thoughts and emotions. The patients' live purpose statement can always be modified and changed, as they undergo personal transformation and spiritual growth through experience and expanding knowledge.

SECTION V

DIFFERENTIAL DIAGNOSIS OF ENERGETIC PRINCIPLES

CHAPTER 20
INTRODUCTION TO DIFFERENTIAL DIAGNOSIS

DIAGNOSIS IN ENERGETIC MEDICINE

Chinese Medicine developed into a comprehensive and extremely effective diagnostic system, over many centuries of accumulated clinical experiences. The symptomatology to identify disease patterns and their underlying disharmony follows the way of nature in its relationship to cause and effect. To identify the patterns of disharmony, the Qigong doctor combines diagnosis of pathology and treatment principles into one continuum. The identification of the disease pattern is not only developed through categorizing a list of symptoms, but also through observing the pathogenesis of the disease. A symptom is a feeling, or reaction, that is a sign of disease, resulting from certain disorders (physical, mental, emotional, energetic, and spiritual).

Treatment must be aimed at finding and treating the cause of the disease. This is one of the most fundamental principles of Traditional Chinese Medicine in terms of differential diagnosis and treatment. The fundamental purpose of diagnosis is to find:

- the cause of the disease,
- the root of the disease, and
- the location of the disease.

Finding the primary cause (or root) of the disease requires understanding its nature, and allows the doctor to focus treatment on the most important etiological and pathological aspects of the disease. When diagnosing a disease, emphasis is placed on observing the energetic movement (or lack thereof), as well as energetic transformation, and not on the present set structures of the existing tissues. The tissues's structures are temporarily generated by the existing energetic changes and are therefore of secondary interest to the changes themselves.

ETIOLOGY IN MEDICAL QIGONG THERAPY

The etiology, or study of the causes of diseases in Medical Qigong therapy can be divided into what is commonly called in Traditional Chinese Medicine as the "three periods of life." These three periods encompass the developmental activities of the patient's Jing, Qi, and Shen formations during the transitional periods of their life, i.e., within the womb, childhood, and adulthood. Each patient's constitution results from the continuous interaction between his or her inherited form and the environment.

Within each period of life Three Stars develop and determine some of the major characteristics and traits of each person. Each group of Three Stars accounts for one third of what, and who, the person is. They combine together to form and complete a Nine Star System of life evaluation.

ETIOLOGY WITHIN THE WOMB

Problems arising during the Three Star Period within the womb (and during birthing) encompass the formative time span of 40 weeks. This time period can be influenced by the following factors.

1. The state of the patient's Jing prior to conception. If the Jing of one or both parents is deficient, the sperm and/or ovum will be weakened.
2. The mother's mental state during pregnancy has a deep affect on fetal formation, and directly affects the developmental activities of the fetus's Jing, Qi, and Shen. If the mother is malnourished, ill, emotionally disturbed, or taking alcohol or drugs during pregnancy the development of the fetus will be affected.
3. If the birth is abnormal, difficult and/or prolonged, it can cause an emotional shock within the newborn.

Diseases may originate from one or more of these factors causing the baby to be born with an illness or a predisposition to illness. This predisposition can be manifested as a particular type of emotional or mental imbalance, a diseases of particular Yin and Yang organs (or organ systems), or a general weakness as a whole. These conditions are commonly observed as "prenatal shock," and are indications of active fetal toxins. Prenatal shock can also be indicated when newborn babies are restless sleepers, continually opening and closing their eyes, and/or having fevers.

These problems are sometimes difficult to diagnose; simple observations however, can help the Qigong doctor to determine if the patient has experienced prenatal shock. A bluish color on the forehead of the patient, for example, is one indication of this condition.

ETIOLOGY DURING CHILDHOOD

Problems arising during the Three Star Period of childhood encompass the formative time span of 18 years. These internal problems can result from abuse, malnutrition, trauma and excess sex (during the teenage years). The subconscious mind of a child is extremely impressionable to both positive and negative influences.

If, as children, the patients experienced any physical, emotional, or sexual abuse, suffered a traumatic loss (of a relationship with either or both parents, siblings, peers, etc.), then they may become predisposed to illness later in life, or develop a disease during this time period. These traumas can affect the adolescent developmental activities of the patient's Jing, Qi, and Shen formations, as this is the child's greatest time of physical growth and development.

ETIOLOGY DURING ADULTHOOD

Problems arising during the Three Star Period of adulthood (from the age of 18 onwards), encompassing the time span of 60 years. These problems are caused by the patient's internal emotions and reactions to surrounding environmental factors. Environmental factors include: diet, relationships (be they social, work related, family or other), frequency of sex, and so on, and will affect the continual developmental activities of the patient's Jing, Qi and Shen formations. An adult, although less impressionable than a child, is less resilient than the child to trauma and illness.

PROGNOSIS

If the patient's constitution has been poor since birth, or if a specific disease has been present since birth, the prognosis is generally poor. If, however, the specific disease is acute, or has only developed during the patient's adult time period, the prognosis is more favorable.

There are always energetic interactions between these Three Periods of life. If, for example, a young female patient has a hereditary imbalance between her Conception and Governing Vessels, and also has severe emotional strain during puberty, she may suffer mental problems later on in her life.

THE GROWTH AND PEAK CYCLES OF JING, QI AND SHEN

By studying the physical, mental, emotional, spiritual, and energetic qualities inherent within the human body, the Qigong doctor is able to observe and predict the various changes according to the Nine Star System of life evaluation. It is believed that the body is given specific time periods in which to flourish during the three periods of life. The following chart depicts the body's natural time frame of each physical, mental, emotional, spiritual, and energetic peak (Figure 20.1).

- From birth to age ten represents a time of growth and development on each dynamic level. The child is in a formative stage, growing physically, mentally, emotionally, spiritually, and energetically.
- From age ten to age twenty represents a time of peak physical growth. Within this time frame the child physically enters into puberty, and is able to reproduce; the growing process is stabilizing.
- From age twenty to age thirty represents a time of peak mental growth. Within this time frame the adult actively acquires knowledge for survival sake. He or she is now able to work, produce results and begins planning for the future.

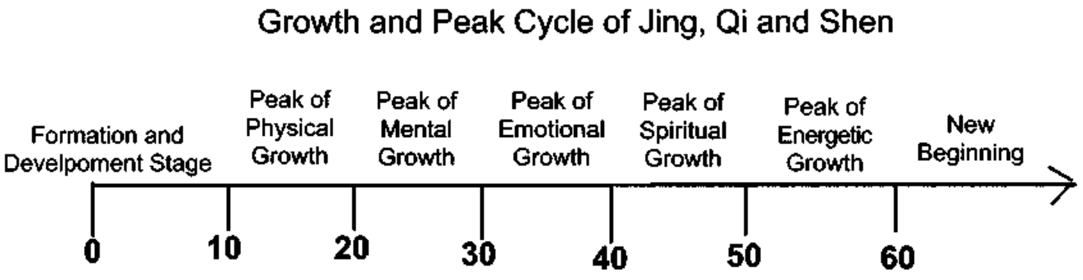

Figure 20.1. Chart of Physical, Mental, Emotional, Spiritual, and Energetic Peak

- From age thirty to age forty represents a time of peak emotional growth. Within this time frame the adult actively acquires knowledge for emotional comprehension of his or her surrounding environment. The adult is now able to creatively express his or her emotions in a civil manner and deal well with emotional disruptions.
- From age forty to age fifty represents a time of peak spiritual growth. Within this time frame the adult actively acquires knowledge for spiritual growth. Now comprehending his or her mortality, the adult begins to actively seek spiritual peace.
- From age fifty to age sixty represents a time of peak energetic growth. Within this time frame the adult's previous experiences come into fruition. If the adult has successfully acquired physical, mental, emotional, and spiritual harmony, the energetic peak will prepare the adult for the rebirth. If the physical, mental, emotional, and spiritual energy is not in harmony, the life-force is terminally exhausted, and the patient dies.

It is commonly noted that when martial arts masters practice their training excessively or incorrectly, it is within this energetic time frame that they die. The cause of death is generally due to the over exertion of the master's Liver and Lungs, and the weakening of the Spleen, Heart, and Kidneys. This energetic disharmony stems from an improper balance of combat training with little or no spiritual training.

- From age sixty on is a time of new beginnings. This time frame represents new challenges of leadership within the family and the community.

DISORDERS OF THE HUMAN BODY

Disorders of the human body are generally due to three factors: constitutional failure of the body's Prenatal or Postnatal Qi, pathogenic factors and unrestrained Evil Qi, or one or more of the Eight Miscellaneous Factors.

CONSTITUTIONAL FAILURE OF PRENATAL AND/OR POSTNATAL QI

Prenatal and postnatal constitutional failure results in a constitutional Qi Deficiency, disease, or stagnation of the vital energy.

The Prenatal Constitution is determined by the strength of the parents when a child is being conceived, or during the pregnancy. If the parents are weak, sick, drunk, or on drugs during pregnancy or conception, the parents' Qi, Blood, and Body Fluids will be deranged and the child will be affected.

In uterus, if the mother is sick, the energy of the child will be affected. Also, negative emotions, poor nutrition, smoking, drugs, and alcohol used by the mother can affect the child within the womb and predispose the child to illness after birth (e.g., prenatal alcohol syndrome and crack babies are born addicted and with mental and physical defects). The prenatal constitution is inherited from both parents and is unchangeable.

The Postnatal Constitution is determined once the child is born. Proper nutrition and a nurturing environment enhances his or her constitution. While it is not possible to entirely erase the effects of the patient's prenatal constitutional patterning, it is possible to enhance it through a balanced life-style and development of their postnatal constitution.

PATHOGENIC FACTORS AND UNRESTRAINED EVIL QI

Pathogenic factors and Evil Qi are derived from Six External Factors and Seven Internal Factors.

The Six External Factors arise from the six climatic changes (Wind, Summer Heat, Damp, Dryness, Cold, and Fire). When the patient's energy system is not well balanced, or the Qi is Deficient, the body is unable to adapt to climatic factors. These climatic factors can then invade the patient's Wei Qi fields and penetrate deeply into the channels and internal organs of the body.

The Seven Internal Factors arise from excessive internal emotions (anger, joy, worry, grief, sadness, fear, and shock). These feelings overwhelm the patient's mental state as the individual responds to social, emotional, and environment interactions. Under normal circumstances, emotions in and of themselves do not cause problems, if they are allowed to be expressed and released. The excessive accumulation of one or more emotions can, however, cause Qi deviations, affecting the corresponding organ or organs of the body.

EIGHT MISCELLANEOUS CAUSES OF DISEASE

Sometimes a patient's disease is caused by what is known in Traditional Chinese Medicine as the Eight Miscellaneous Factors (Diet, Overexertion, Sex, Excessive Child Bearing, Exposures to Poisons, Parasites, and Iatrogenic Disorders). The Eight Miscellaneous Factors offset the patient's balance of Yin and Yang energy and are described as follows.

1. An improper diet can throw the patient's body and mind out of harmony. Nutritional problems can be divided into three main categories:
   a. The excess intake of certain particular foods, as well as overeating food in general.
   b. Unwise eating habits such as:
   - eating foods that are inconsistent with the seasons,
   - eating a diet of too many Cold, Damp, Dry, Hot, greasy, spicy or raw foods,
   - eating irregular meals,
   - oscillating from feasting to fasting, or
   - eating while hurrying or emotionally upset.

   c. Malnutrition, which is developed from a deficient intake of the basic requirements needed to maintain a balanced diet. This condition can be caused from ignorance, poverty, or an impaired digestion, absorption and metabolism.
2. Overexertion consumes the Qi and results in energetic deviations. This condition pertains to both mental/emotional and physical exhaustion, and can stem from loss of sleep, inadequate rest, etc.
   a. Mental and emotional overexertion can stem from a dissatisfaction with one's work (or the lack of work), or a dissatisfaction with one's life. It can lead to stress, frustration, boredom, apathy, or depression.
   b. Physical overexertion varies according to the individual's occupation, as each job has its own characteristics hazards to the patient's health. Problems arising in this area can range from structural damage to organs and tissues (caused from Qi and Blood stagnation), to sensory impairment.
   c. Lack of sleep and inadequate rest can also be either a contributing factor or direct cause in the development of Qi deviations.
3. An unbalanced sex life can cause disease. An individual's sex life is innately connected to the personal spiritual qualities of his or her Hun and Po. Therefore, the majority of sexual problems are rarely of purely physical origin, and reflect the energetic balance of each individual's emotional and spiritual passions.
   a. An Excessive sex life: what is excessive for one individual may not necessarily be excessive for another. The condition of Sexual Excessiveness should therefore be based on the individual's con-

stitution. When an individual exceeds his or her body's normal sexual function, the excessive sexual act can burn up the body's Yin, Jing, and Kidney Qi, causing Deficient syndromes (this condition is more prone in men then women). The act of sex itself is considered a Hot and Damp state.

   b. A Deficient sex life: when an individual abstains from sex, his or her Jing production begins to increase. If this increased energy buildup is not transformed into spiritual energy, it can overflow into the Sea of Marrow affecting the brain. The prolonged lack of sex sometimes creates an emotional need and dependency towards addictions to replace the need for intimacy.

4. Excessive childbearing injures the mother's Blood and weakens her Qi. Women suffer loss of Jing during childbirth, therefore excessive childbirth can cause a Deficient condition.
5. Traumatic accidents and injuries pertain to physical traumas which disperse or congeal the body's Qi and Blood, resulting in energetic deviations. Even after the physical trauma has healed, a site for potential weakness often remains. As the patient ages, the site becomes vulnerable to further injury due to old age, physical exhaustion, stress, and exposure to cold. All traumatic accidents and injuries also pertain to emotional trauma as they are energetically inseparable.
6. Exposure to poisons pertains to any substance taken into the body by ingestion, inhalation, injection, or absorption that interferes with the patient's normal physiological functions. These poisons include both chemical as well as environmental toxins (e.g., pesticides, carbon monoxide, electromagnetic fields, and radiation).
7. Parasites include infections and infestations, and pertain to the numerous organisms capable of living within the body.
8. Iatrogenic disorders pertain to any adverse mental, emotional, or physical condition induced through medical treatment.

THE DIFFERENTIATION OF SYNDROMES

The essence of Chinese medical diagnosis is holistic and based on the concept that all things and events are interrelated. Thus the Chinese physician takes into account not only the physical symptoms of the disease, but also spiritual, mental, emotional, and environmental factors or events, that both interrelate and interact to form "patterns of disharmony" within the patient. The doctor studies and seeks to understand these patterns of disharmony, called "syndromes," to provide the framework for treatment.

Syndromes can include External or Internal factors, congenital or acquired disorders, and symptoms of stagnation, Excess, and Deficiency, or traumatic injuries to the organs and tissues. The Qigong doctor must first determine whether the patient's disease is primarily due to an energetic disorder (e.g., an Excess or Deficient condition) or a physical disorder (i.e., structural damage). Once this determination has been made, the goal or goals of treatment can be set, and the means of treatment chosen. In the treatment of the lower back, for instance, if lumbar pain is due to a subluxation of the vertebra, caused by external injuries, the doctor should first reposition the vertebra through hand manipulations, by using Tui Na therapy, then remove the Qi stagnation. If the doctor is unfamiliar with such techniques, it is better that he or she first refer the patient out for spinal adjustment, and then remove the Qi stagnation. Only in this way can a satisfactory curative effect be obtained. If, on the other hand, the back pain is due to abdominal distention or a Kidney malfunction (i.e., primarily an energetic disorder) then treatment must proceed through Qi emission.

Several methods can be used singly, or in combination, to diagnose syndromes. They are described as follows.

1. Diagnosis according to the identification of the Eight Energetic Principles.
2. Diagnosis according to the identification of Qi, Blood and body fluid distinctions.
3. Diagnosis according to the identification of

the Five Elemental patterns.

4. Diagnosis according to the identification of the Six Stages.
5. Diagnosis according to the identification of the Four Levels.
6. Diagnosis according to the identification of the Triple Burners' System.

THE CATEGORIZATION OF SIGNS AND SYMPTOMS

In order to categorize symptoms into syndromes, the Qigong doctor must determine: The Onset, Location, Aggravating or Relieving Factors, Type of Pain or Sensation, and Course Since Onset.

1. The Onset includes the date the symptoms first appeared, whether they appeared suddenly, or had a gradual onset, and the order of their appearance.
2. The Location specifies whether the symptoms remain fixed, or migrate throughout the body.
3. The Aggravating or Relieving Factors indicate the specific conditions that affect the symptoms in a positive or negative way. Aggravating or Relieving Factors may include:
   - if the symptoms feel better or worse during daytime or nighttime,
   - if the condition is better with application of warmth or cold,
   - if the condition is better with pressure or worse with pressure,
   - if the condition is relieved by movement or aggravated by movement, and
   - if the condition is relieved by eating or made worse by it.
4. The Type of Pain or Sensation describes the nature of the symptoms, such as:
   - sharp or dull
   - pounding
   - stiffness
   - burning
   - itching
   - intermittent (coming and going)
   - electrifying
   - pulling or shooting
   - nauseated
   - distended
5. The Course Since Onset includes the incidence, progress, and the effectiveness of the treatments received.

The syndromes do not equal the patient's disease, but are the patterns of the illness. In determining the patient's syndrome, the Qigong doctor takes into account that these patterns can occur simultaneously, and vary in degree of severity. The purpose for assigning the patient's disease a syndrome is not merely to classify the disease, but to understand its process. Not all the symptoms listed for a particular illness need to be present in order to properly determine the patient's syndrome.

DIAGNOSIS AND TREATMENT OF THE PATIENT

The assessment and evaluation of the patient generally proceeds in eight stages. The first seven of the eight stages are performed prior to the initiation of the actual treatment. The final stage is undergoing the treatment plan and the periodical reevaluation of the course of the disease since the beginning of the treatment (Figure 20.2). During these eight stages, the Qigong doctor assess: The Whole Person, The Patient's Constitution, The Signs and Symptoms of the Disease, The Syndrome, The Patient's Organs, The Etiology of the Disease, The Treatment Goal, and the Treatment Plan.

1. The Whole Person must first be evaluated according to his or her whole being. This process involves assessing the patient's:
   a. physical appearance (e.g., is the patient disheveled or appropriately groomed),
   b. energetic demeanor (is the patient very nervous, calm, manic, etc.),
   c. structural build:
   - according to the physical Yao formations,
   - introverted or extroverted features, and
   - Yin or Yang appearance.
2. The Patient's Constitution is assessed according to the Five Elemental Constitutions and their subdivisions of balance, and Yin or Yang predominant element.
3. The Signs and Symptoms of the Disease are then classified as:
   a. Objective Signs that are apparent to the doctor,

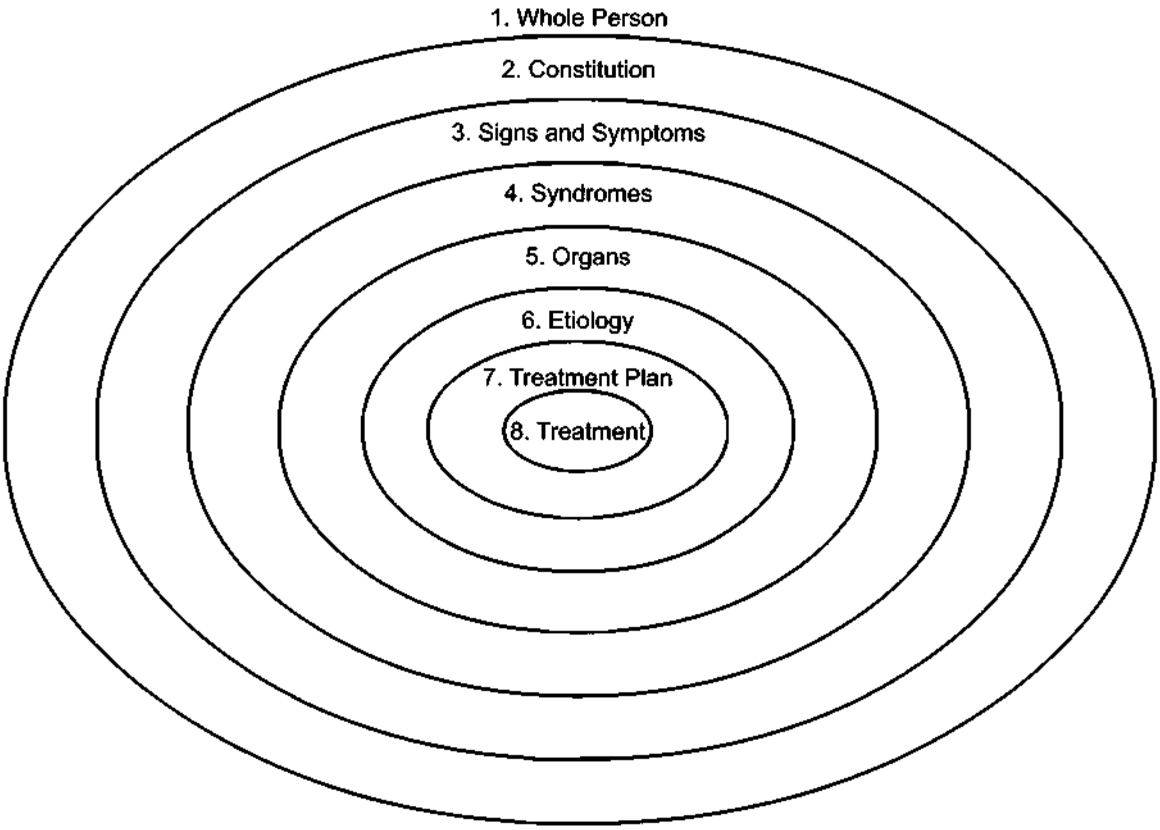

Figure 20.2. The Qigong doctor begins to diagnose the patient by first observing the patient as a whole. Next the doctor observes the patient's constitution, symptoms, syndrome, health and state of the internal organs, and the cause of the energetic dysfunction, before treating the disease. All the acquired knowledge is placed against the different diagnostic templates of the Eight Parameters, Five Element Theory, Qi, Blood and Body Fluid Dysfunctions, etc. This entire process is initiated in order to assist the Qigong doctor in prescribing the correct treatment. The treatment focuses on the specific cause of the disease and how it affects the patient's internal organs. The primary goal being to relieve the patient's symptoms, strengthen the constitution, and return the patients to a state of wholeness.

---

   b. Subjective Symptoms that the patient reports to the doctor,
   c. Cardinal Symptoms and Signs that are used for differential diagnosis of disease, and
   d. Constitutional Symptoms and Signs that are indicative of a systemic disorder.
4. The Syndrome (a pattern of disease) is then determined according to the patient's symptoms and signs of disordered function, by assessing their relationship to one another.
5. The Patient's Organs are then evaluated according to any pathological changes occurring within the internal organs.
6. The Etiology (or root causes) of the syndrome is determined to complete the comprehensive diagnosis.
7. The Treatment Plan is then formulated, taking into account all the data gathered by the doctor. The Qigong doctor devises a treatment goal and determines a strategy of treatment

that addresses the patient's syndrome or disease, in accordance with the agreed upon goal and needs of the patient.

8. The Treatment is implemented and prescriptions are assigned to initiate the healing process. Over time, the treatment plan is reevaluated, as the patient improves or suffers setbacks (for various reasons). Prescriptions, likewise undergo changes in accordance with the patient's changing condition.

Once treatment begins, the process of the development of the disease is usually reversed, or the symptoms are alleviated. After the initial treatment begins, the causes of the disease are addressed, thus affecting the organs, which in turn changes the syndrome and the symptoms. This process strengthens the constitution and establishes a dynamic balance of health and well being.

UNDERSTANDING THE THEORY OF ENERGETIC COUNTERACTION

When treating patients it is important for the Qigong doctor to understand the theory of Energetic Counteraction. This theory describes the body's natural regulation, wherein no two specific symptoms (affecting the patient's general constitution) can occur at the same time, for any considerable period, within the same system.

Generally as a disease progresses, the old symptoms will give way to the newer ones. Consider, for example, a patient who has asthma, which is brought about by Wind Cold invading the Lungs (resulting in a dyspnea, wheezing, coughing, and severe mucous). If this patient unexpectedly suffers from an attack of gout (a painful inflammation and swelling of the joints), caused from the Wind Damp invading the joints and channels, the sudden attack of gout will cause the patient's asthmatic symptoms to disappear. When the gout subsides, the asthmatic symptoms will return. The new disease naturally prevails over the original condition.

CHAPTER 21

DIAGNOSIS ACCORDING TO THE EIGHT ENERGETIC PRINCIPLES

The method of diagnosis according to the Eight Energetic Principles was formulated by doctor Cheng Zhong Ling during the early Qing Dynasty. The Ba Gan or Eight Principles of Differential Diagnosis is a system using four pairs of opposite symptoms viewed in Traditional Chinese Medicine as being essential to the understanding and treatment of disease. In most cases, knowledge of the diagnostic systems and treatment principles of general Yin and Yang conditions is enough to give the Qigong doctor some good guidelines for steering the patient's training program. Similar principles can be found in most ancient healing traditions.

The four opposite pairs are Yin-Yang and the three subdivisions of Yin and Yang into: Internal-External, Hot-Cold, and Deficient-Excess.

1. Internal and External Symptoms inform the doctor of the origin of the pathogenic condition.
2. Hot and Cold Symptoms inform the doctor of the disease's onset (how it came on) and whether the patient's temperature has been altered due to a viral infection or a metabolic change.
3. Excess or Deficient Symptoms inform the doctor whether the condition is either to be tonified or purged, and the strength of the patient's constitution.
4. Yang and Yin Symptoms inform the doctor how to establish a physical (external), as well as energetic (internal) balance.

A detailed knowledge of the Eight Energetic Principles is required to pinpoint more subtle patterns of disharmony. This understanding allows the Qigong doctor to unravel complicated patterns and identify the basic contradictions within them. The main purpose for applying these principles is to understand the etiology of the disease and the nature of the dysfunctions; not simply to categorize the illness. Only then can a treatment pattern be initiated.

INTERNAL AND EXTERNAL PATHOGENIC FACTORS

Diagnosis according to internal and external pathogenic factors is not based on the etiology but on the basis of the disease's location. If a disease is caused, for example, by an "external" pathogenic factor but is currently affecting the internal organs it is classified as an Interior condition. An Exterior condition affects the patient's skin, muscles and channels, known as an "exterior pattern" of pathogenic factors. An Interior condition will affect the bones and internal organs, known as an "interior pattern" of pathogenic factors.

EXTERIOR CONDITIONS: SIX EXOGENOUS PATHOGENIC FACTORS

The external pathogenic factors—Wind, Cold, Dampness, Fire, Heat, and Dryness—are the six "evil" factors that come from the environment outside of the body. Under normal circumstances they are good for the health, but can be very harmful when they become abnormal.

Unseasonable weather (weather that is too cold or too hot for the body to tolerate) can make people ill by impeding the normal flow of internal energy thus making the body vulnerable to disease.

Each external factor, once it enters the body can manifest as an internal factor (i.e., External Wind can penetrate and become Internal Wind, etc.). It is also important to know that an external invasion once it penetrates the organ can transform into another internal factor (i.e., an External Cold invasion can give rise to Internal Heat).

These external factors are seen as not only the cause of disease, but also as manifestations of the disease. A disease may be due to any one of these Six External Pathogenic Factors, or simply manifest as an external "evil," even though that pathogenic factor was not present at the time the patient got sick (e.g., Damp-Heat). Treatment is therefore based on how a disease manifests itself, rather then the specific cause. However, it is wise to have a sense of the etiology of the disease to determine where to intervene and what might be expected next.

The Six External Pathogenic Factors are categorized in China according to the seasons, temperature, and climate. These transitions are explained as follows.

1. The Heat Season is characterized by temperatures that are usually above 90 degrees Fahrenheit. The Heart, Small Intestines, and blood vessels are the most vulnerable at this time. This is most evident in summer.

   Heat is a Yang pathogenic factor that rises and spreads excessively, consuming and diminishing the Prenatal (Yuan) Qi, Body Fluids and Lung Qi. As Heat rises, the body's Qi is drawn upward and dispersed, causing the Yin energy to descend. Symptoms can range from restlessness, anger, delirium, loss of consciousness to even coma. This condition is most evident during the summer. The Heart organ benefits from bitter tasting foods which directly stimulate and enhance the Heart's Qi.

   a. Mild Heat that has been created within the body, causes an ascending action of Qi. This ascending action causes the body's Yin energy to become disturbed creating an energetic imbalance.
   b. Fire (Extreme Heat) is a Yang pathogenic factor and spreads excessively, consuming and diminishing the Yuan Qi, Body Fluids, and the Lung's Qi. It produces an explosive reaction, causing the Qi to flow up into the head and disturb the Blood. This action causes symptoms ranging from nosebleeds, skin infections, swelling, and skin eruptions, to hemorrhaging and vomiting Blood.

2. The Damp Season most often affects the Spleen, Stomach and muscles that are at their most vulnerable time. The Spleen organ benefits from sweet tasting food. The temperature is usually between 75 and 90 degrees Fahrenheit. This is evident in the late summer, during the rainy season.

   Dampness is a Yin pathogenic factor that easily obstructs the body's production of Qi and impairs the Yang. It is considered a Yin pathogenic factor because, as the invasion of Dampness moves downward, it manifests as a heavy and turbid energy, resulting in a full chest, nausea, vomiting, and a sticky sweet taste in the mouth. During the damp season, an invasion of Dampness may affect the urinary organs, causing stagnation of Qi. Once pathogenic Dampness enters the body, it becomes foul in nature. If the body becomes obstructed, the results will become evident in skin diseases with abscesses, and oozing ulcers with turbid, cloudy urination. Disease also manifested through infections, in combination with Heat.

3. The Dry Season is characterized by temperatures that are usually between 60 and 70 degrees Fahrenheit. The Lungs and skin are the most vulnerable at this time. This is most evident in autumn, because of the lack of humidity.

   Dryness is a Yang pathogenic factor because as the Heat rises, dryness consumes the body's Yin Qi, especially in the Lungs. This results in dry, chapped skin, cracked skin, and dry cough with sputum, throat pain, or asthma. The Lungs themselves benefit from pungent tasting foods, which directly stimulate and enhance the Lung's Qi.

4. The Cold Season is characterized by temperatures that are usually below 60 degrees Fahrenheit. The Kidneys, Bladder, and bones are most vulnerable at this time. This is most evident in the winter.

   Cold is a Yin pathogenic factor characterized by astringency of the tissues. During the Cold season the Yin pathogenic factors consume the Yang Qi. Cold blocks the channels and collaterals, causing pain and damage to the Heart, as well as consuming the Kidney Yang. Pathogenic Cold also causes the channels and collaterals to contract, resulting in the slowing down of Qi and Blood circulation. As the Yin energy moves downward consuming the body's Kidney Yang, the body's Yang is unable to generate heat causing chills, pallor, shivering, diarrhea with undigested food, stiffness, pain, numbness, and clear urine. Internal Cold is characterized by conditions of overall stagnation and contraction. The Kidneys themselves benefit from salty tasting foods which directly stimulate and enhance the Kidney's Qi.

5. The Wind Season affects the Liver, Gall Bladder, and tendons that are at their most vulnerable time. The beneficial taste is sour. The climate in the Wind season, depending on the area, is usually between 60 and 70 degrees Fahrenheit. This is most evident in the spring. Wind is Yang in nature and tends to injure the patient's Blood and Yin. It is divided into both internal and external pathogenic Wind conditions. External Wind is called "the mother of one thousand pathogenic factors," and is responsible for the propagation and growth of the pathogenic environment, because it combines with other pathogenic factors to create Wind Heat, Wind Cold, Damp Wind, Dry Wind, etc.

   In the Wind season the Yang pathogenic factors are in the upper portion of the body, weakening the Wei Qi, and causing derangement of the opening and closing of the pores on the body's surface. The pathogenic Wind tends to move constantly causing abnormal motion and rigidity in the trunk and limbs. External Wind affects the upper part of the body, shifting and moving the pain like blowing a leaf. Once External Wind invades the body, the symptoms are stiff neck, itchy sore throat, watery eyes, headache, nasal obstruction, facial puffiness, aversion to Wind, irritability, abnormal sweating, migrating joint pain, tremoring, convulsions, rigidity, and facial paralysis.

   Internal Wind is manifested through the same symptoms. Because of its serious nature, it can also cause tremors, convulsions, and paralysis throughout the body (with the exception of the face, which is due to External Wind invasion).

INTERNAL PATHOGENIC FACTORS

An interior disharmony is diagnosed when the internal organs have become affected by pathogenic factors, creating an Interior syndrome. Interior syndromes are the pathological conditions which have resulted from one of three factors: (1) the invasion of the channels by external pathogens, which can be transmitted to the body's Yin and Yang organs; (2) the direct attack on the Yin and Yang organs by external pathogens; and (3) sudden emotional changes due to improper diet and stress which directly affect the body's Yin and Yang organs leading to functional disturbances.

1. A Persistent Exterior Invasion by one of the Six Exogenous pathogenic factors can penetrate deeper into the body creating an Interior syndrome, if it is not purged from the patient's body.
2. A Direct Invasion of the Yin and Yang organs can also occur from an Exterior Invasion of pathogens. In such cases, the affected organ will manifest as an Interior Excess or Deficient condition depending on the syndrome.
3. Emotional Factors can likewise cause damage to one of the Yin organs, which in turn can cause damage to the other Yin and Yang organs. People encounter emotional upheavals in everyday activities. Under normal circumstances, the Seven Emotions are good for health (see Chapter 21). Appropriate anger helps to disperse stagnant Liver Qi, preventing stasis of Liver energy and helps the patient to establish a healthy boundary system. Rage, on the other hand, leads to the abnormal increase and ascension of Liver Qi, resulting in a flushed face and dyspnea (shortness of breath). If the pattern continues (i.e., the person is a rage-aholic), the violent and uncontrolled anger may lead to serious disorders of the Liver, including eye injury, fainting due to inadequate Blood flow to the brain,

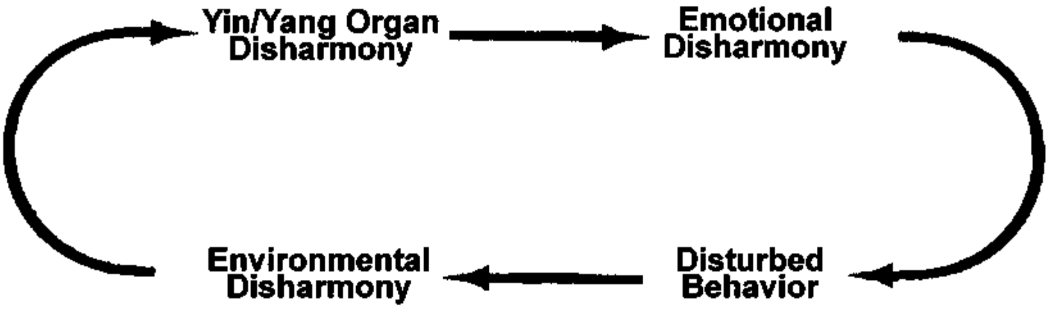

Figure 21.1. Cycle of Disharmony

cerebral hemorrhaging, or even death. It is believed therefore, that Internal conditions can arise from either an Excess emotional condition caused by emotional suppression, or a Deficient emotional condition caused by excessive emotional discharging.

It is also believed that a person's personal emotional history determines his or her biology (life processes). Each person creates this reality through his or her belief system. Hence the patient creates the disease by embracing traumatic past wounds and formulating belief structures that support these traumas. These energetic clusters are then fed via the major viscera by the excessive emotional energy. This pathological emotional process can result in the formation of cysts, tumors, cancer, etc.

Any imbalance of the Seven Emotions leads to the unnecessary consumption, or blockage of, the internal Qi flowing through the channels of the viscera. In either instance, this results in abnormal Qi activity and creates a functional disorder of the cerebral cortex.

Emotional disharmony induces a vicious cycle which affects the mind, body and spirit. Any factors (Yin and Yang disharmony, emotional disharmony, disturbed behavior, or environmental disharmony) can trigger the cycle (Figure 21.1). When caught in this cycle, emotional energy can deplete the physical body.

Anger, for example, generally causes the Qi to rise; however, when anger turns inward it can transform into depression. Mental depression may obstruct the functional activities of Qi and this stagnation may cause Fire. The Fire syndrome is caused by the disorders of the Seven Emotions and may burn up the fluid of the organ related to the predominant emotion.

In Western culture, people are generally not encouraged to acknowledge painful emotions; feelings are usually disregarded and conquered by a strong will. The storage of emotional energy in the muscles and organs of the body leads to tension, stress, and illness. The Qi cannot circu-

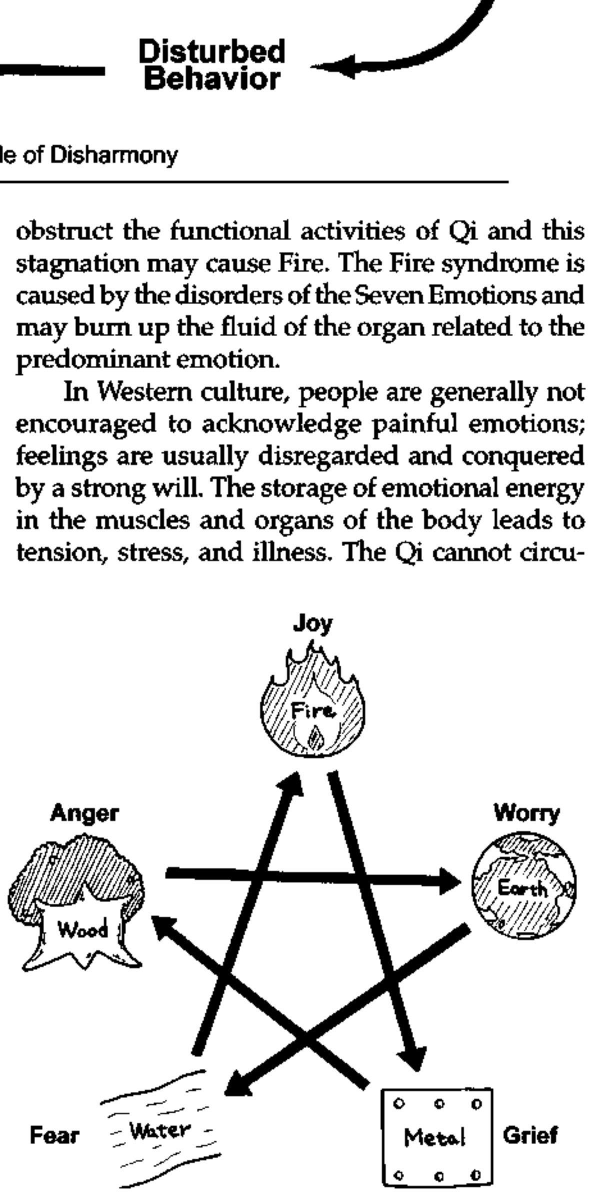

Figure 21.2. Five Elemental Controlling Cycle

late with emotional blocks. Many specialists in Western medicine today agree that a large percentage of diseases being treated are, indeed, caused or aggravated by stress.

The Five Elements have a creating (generating) and restrictive (controlling) relationship. The sequence of the creative relationship of the Five Elements is: Wood creates Fire; Fire creates Earth; Earth creates Metal; Metal creates Water; Water creates Wood. This circle is endless.

The sequence of the controlling relationship of the Five Elements is: Wood controls Earth; Earth controls Water; Water controls Fire; Fire controls Metal; Metal controls Wood. This circle is also endless.

The controlling relationship can be applied to problems with emotions. Anger, for example, belongs to the rising of Excess Liver energy and can be counteracted by descending the Lung's Qi (Figure 21.2). The Lungs control sadness, and sadness may lead to the consumption of abnormal rising Liver Qi (Metal controlling Wood).

Excessive sadness may lead to the extravagant consumption (and depletion) of Lung Qi, which damages the body's Yin and the Heart. The consumption of the Lung Qi can be stopped by the Heart's energetic release of the emotion of joy. Joy causes the rapid rising of the Heart's Qi. It can be regulated by fright and terror, which are controlled by the rapid descending energy of the Kidneys. Terror leads to the abnormal falling of Qi, it therefore counters the abnormal rapid rising of Heart's Qi. Excessive fright and terror may lead to the abnormal sinking of Gathering Qi, which damages the Kidney's Qi (Figure 21.3).

Because over-thinking can cause Qi to gather, it can be used as a tool to help bring back dispersed energy. Over-thinking and anxiety, however, may lead to the depression and stagnation of Qi, which weakens the transporting and converting ability of the Spleen's energy, and causes a lack of appetite. Mild cases of this condition can manifest as dyspepsia (indigestion), abdominal distention and diarrhea. In severe cases, the Qi

The Seven Emotions			
Chinese Word	English Translation	Organ Association	Qi Manifestation
Xi	joy, happiness, excitement, pleasure	Heart	slows down and relaxes the Qi
Nu	anger, irritation	Liver	causes Qi to rise
Si	meditation, contemplation, pensiveness, worry	Spleen	stagnates the Qi
You	anxiety, sorrow, grief	Lungs	obstructs the Qi
Kong	fear, extreme anxiety	Kidneys	causes Qi to descend
Jing	shock (sudden intense fear), fright	Heart then Kidneys	scatters and deranges the Qi
Bei	grief, affliction, sadness	Lungs	disperses and consumes the Qi

Figure 21.3. The Seven Emotions Chart

and Blood stagnate in the chest and abdomen, causing a feeling of fullness in the chest or upper abdomen. This stagnated, ever-sinking Spleen energy can be smoothed and aroused by anger (the Liver's energy expanding upward and outward). In short, the hyperactivity of Yin is treated with Yang to check the unnecessary consumption of internal energy, and to promote a mild Yin energy. Then the Qi and Blood can flow without interruption and disease can be prevented.

One effective technique to avoid an emotional and energetic overload or depletion of a Yin organ's Qi, is to apply the Five Elemental counter-emotion to restore emotional balance (see Chapter 53).

OVERVIEW OF EXTERNAL AND INTERNAL CONDITIONS

The internal and external diagnostic parameters locate the depth of the condition of the disharmony and are fairly easy to distinguish, relative to the other pairs of opposites. Thus, determining the internal or external nature is usually the first step in diagnosis of a disorder.

Internal conditions are usually caused by internal disharmonies, such as excessive emotional energy in the form of anger, excitement, sorrow, anxiety, fear, and mourning, or other imbalances associated with the Seven Emotions. Excessive thinking and stress can also lead to an imbalance that is characterized as internal. Internal conditions are often chronic and tend to have a gradual onset and a longer duration.

Conditions of External disharmony are typically caused by the external influences on the body, known in Chinese medicine as "pernicious influences." They are Wind, Cold, Heat, Damp, Dryness, and Summer Heat. Pernicious influences usually invade the body when the body's Wei Qi is weak.

Typical External diseases are colds, influenza, skin eruptions, or diseases caused by injury or shock. Excessive amounts of eating, drinking, working, and sexual activity can also lead to both Internal and External disharmony. External conditions are often acute, with sudden onset, and of short duration.

Many times Internal conditions will develop when an External condition moves inward. This represents a further weakening of Wei Qi and the body's natural immune system. Before the condition has developed into a full Internal sickness, it will go through a stage of being half-external and half-internal.

Examples of conditions which may be considered as being half-internal and half-external are: alternating fever / chills, heaviness in the chest, restlessness, nausea, vomiting, no appetite, dry throat, irritation of the mouth, and dizziness. A bounding pulse is also an indication of this type of condition.

This is one reason why treatment of an External disease (colds and flu) requires sweating therapy to encourage the disease to remain external. The contractile mechanism of the skin is a specific external representation of the body's Yang immune system at work.

When treating extreme cases, such as internal injuries and shock, conserve the patient's Yang Qi by keeping them warm.

HOT AND COLD

Hot and Cold describes whether the condition is the result of a Hot or Cold state and is determined by hot or cold sensations. The nature of this clinical manifestation depends on whether it is combined with a Full or Empty condition.

HOT CONDITIONS

Hot conditions usually arise when there is excess activity of the body's Yang functions, or if there is an insufficiency of Yin Qi or Fluids in the body. A Hot condition can be further divided into either Full Heat or Empty Heat.

1. A Full Heat conditions manifest as the following physiological and mental / emotional symptoms.
    a. Physical symptoms include:
    - increased energy and metabolism, i.e., physiological hyperfunction
    - fever with thirst, and desire for cold liquids
    - dry stool and constipation
    - dark yellow and reduced urine
    - flushed face, red eyes, hot red skin
    - warm extremities
    - aversion to warmth
    - a rapid, full pulse and a red tongue with yellow fur

b. Mental and emotional symptoms include:

- irritability
- restlessness
- loud voice, talkativeness
- extroverted and aggressive behavior

Inflammations, ulcers, and infections are typical of Full Heat diseases. There are many other symptoms depending upon which organ is affected.

These Full Heat symptoms arise when there is an Excess of Yang energy in the body. Excess Yang energy can be caused from overconsumption of Hot foods (which cause Stomach or Liver Heat), or long-standing emotional problems (which cause Liver or Heart Heat) due to Qi stagnation. Full Heat can also develop from the invasion of external pathogenic factors which have transformed into Heat.

2. An Empty Heat conditions manifest as the following physiological and mental/emotional symptoms.

a. Physical symptoms include:

- dry mouth and throat (at night)
- night sweats
- afternoon fever
- a feeling of heat in the chest, palms and bottoms of feet (Five Palms Hot)
- dry stool
- scanty, dark urine
- a floating, empty and rapid pulse, with a red, peeled tongue

b. Mental and emotional symptoms include:

- fidgeting and mental restlessness
- vague anxiety
- insomnia

There are many other symptoms depending upon which organ is affected.

These Empty Heat symptoms arise from a Deficiency of Yin. If the patient's Yin becomes Deficient (usually due to a Deficiency of Kidney Yin), this affects the Yin of the Liver, Heart, and Lungs and causes a relative Excess Yang condition.

COLD CONDITIONS

Cold conditions usually arise when there is excess activity of the body's Yin functions, or if there is an insufficiency of Yang Qi in the body. A Cold condition can be further divided into either Full Cold or Empty Cold.

1. A Full Cold conditions manifest as the following physiological and mental/emotional symptoms.

a. Physical symptoms include:

- diminished physiological function and lowered immunity
- decreased energy
- aversion to cold and craving for warmth
- cold extremities
- lack of thirst
- clear urine, watery stool
- desire for hot drinks
- white face
- a slow pulse and a pale, white tongue with moist fur

b. Mental and emotional symptoms include:

- timid behavior and soft speech
- slow movement
- a lack of motivation

Generally, discomfort is increased by cold and reduced by warmth. Cold symptoms may also appear at later stages of fevers and chronic peptic ulcers, denoting a weakening of the body's ability to overcome the disease.

These Full Cold symptoms arise when there is an excess of Yin Qi in the body. Excess Yin Qi can be caused from the direct invasion of External Cold into the body's interior. In this particular condition, the External Cold can either invade the Stomach causing vomiting and epigastric pain, invade the Intestines causing diarrhea and abdominal pain, invade the uterus causing dysmenorrhea, or it can invade the Liver Channels causing pain and swelling in the scrotum. One of the main manifestations of Interior Full Cold is abdominal pain caused from the Cold constricting and obstructing the circulation of Yang Qi. Full Cold can also develop from the invasion of external pathogenic factors which have transformed into Cold.

2. An Empty Cold conditions manifest as the following physiological and mental/emotional symptoms.

a. Physical symptoms include:

- cold limbs and chills
- pale face
- fatigue
- no thirst
- loose stools
- clear, abundant urination
- sweating
- pale tongue with thin, white coating and a deep, slow or weak pulse

b. Mental and emotional symptoms include:

- tiredness and lack of motivation
- slow movement

An Empty Cold disharmony usually arises when the Yang Qi of the body is insufficient, and fails to warm the body. This is caused by a Deficiency of Spleen Yang, Kidney Yang, or Heart Yang (sometimes Lung Qi Deficiency). The most common cause is related to the Spleen Yang Deficiency, which fails to warm the muscles and thereby causes chills. If the Spleen does not receive heat for its function of transforming food, the result will also be loose stools.

COMBINATION OF HOT AND COLD SYMPTOMS

While distinguishing Hot diseases from Cold diseases may appear to be fairly straightforward, the doctor may find that there are many circumstances where both Hot and Cold symptoms appear simultaneously. The use of sensory input is just a general indicator, however, because it provides the doctor with only partial data; it can sometimes be misleading if other considerations are not taken into account. The doctor needs to consider the other symptom characteristics according to Yin-Yang or Deficient-Excess patterns.

1. An External Cold with Internal Heat syndrome manifests when a patient has a preexisting internal Heat condition and is subsequently invaded by External Wind Cold. This condition can also occur in attacks of latent Heat combined with the invasion of Wind Cold. Symptoms can include fever with an aversion to cold, body aches, irritability, and thirst.
2. An External Heat with Internal Cold syndrome manifests when a patient with a preexisting Cold condition is attacked by exterior Wind Heat. Symptoms can include fever with an aversion to cold, chills, headache, sore throat, and thirst.
3. A Heat Above-Cold Below syndrome manifests when a patient's pathogenic Heat rises resulting in Heat above and a Cold syndrome below. Symptoms can include thirst, bitter taste, sour regurgitation, irritability, Herpes Simplex I, borborygmus, loose stools, and pale and profuse urination.

Qigong patients must be aware of these signs of Hot and Cold disharmony. If a patient's training routine is too Yang, this will burn the body's Fluids, thus diminishing the Yin Qi's ability to contain the Excess Yang, which in turn produces a Hot condition. An overactive metabolism can also induce a Hot condition, while a sluggish metabolism can produce a Cold condition. Less extreme conditions are classified as warm, cool, or neutral.

EXCESS AND DEFICIENT

The capacity of an individual to maintain resistance to the invasion of pathogens is relative to the strength of the disease and the strength of the individual. An Excess condition is characterized by the presence of a pathogenic factor (interior or exterior) and by the fact that the body's Qi is still functioning normally. The battle against the pathogenic factors results in the symptoms and signs of an Excess condition. A Deficient condition is characterized by a weakness of the body's Qi and the absence of a pathogenic factor. If the body's Qi is weak, but the pathogenic factor continues to attack, the condition is considered Deficient and is characterized by an Excess.

EXCESS CONDITIONS

An Excess condition will usually occur when a bodily function becomes overactive or Qi accumulates unnaturally due to a blockage. It is characterized by strength, buoyancy of spirits and often acute symptomatic reactions. Disharmonious conditions which begin and end suddenly are due to Excess. Yang Excess can be quickly improved, in most cases, with a diet of predominately vegetables and fruits.

It is not unusual for patients whose diets are heavy in meat proteins, to suffer from Excess diseases. Many Naturopaths advocate the predominant use of raw fruits and vegetables and have experienced outstanding results due to the elimination of Excess Yang (from meat toxins) from the body.

Such a diet is not good therapy for individuals who eat a relatively balanced vegetarian diet, because it serves to make them more Yin and threatens to throw them off balance. This is a good reason to avoid "fad" diets. A diet or method that works well for one person may have the opposite affect for someone else; each case must be considered separately.

Some indications of an Excess condition are as follows: flushed face, loud voice, coarse and full breathing, acute pain, extreme menstrual pains, a tender abdomen, which worsen with the application of heat. A strong pulse can also be a sign of an Excess condition.

In cases of Excess, to expel pathogenic evils, the patient imagines that the Toxic Qi is being expelled from specific channel points upon exhalation. The results are always more effective when Qigong prescriptions are given in conjunction with the treatments.

Any interior pathogenic factor (Heat, Cold, Damp, Wind, Fire, and Phlegm) can give rise to an Excess condition, including the stagnation of Qi and stasis of Blood.

DEFICIENT CONDITIONS

Although Deficiency is characterized by insufficient Jing, Qi, Shen, Blood, and Fluids, it does not generally mean a lack of some nutrient as it does in Western medicine. It is more a statement about the body's inability to find, or produce, what is necessary for its immune function to provide a functional integrity in the body. Deficiency, therefore, is taken in a more general sense and denotes the overall condition of the patient, or of the disease.

Conditions which linger are usually due to Deficiencies. The general indications are: great weakness, inconsistent energy, shallow and rapid respiration, pain characterized by soreness, weak or inconsistent volume of voice, moderate menstrual pains, withered face, low spirits, and abdominal pain. A Deficient condition will generally respond well to heat and pressure. A weak pulse can also be an indication of a Deficient condition. There are four types of deficiencies: Deficient Qi, Deficient Blood, Deficient Yin, and Deficient Yang.

1. Deficient Qi is the first and least severe stage of Deficiency. Most symptoms arise from a weakness of the Lungs' Qi (which fails to control the breathing) and a weakness of the Spleen's Qi (which fails in its function of transforming and transporting).
2. Deficient Blood is caused from a dysfunction of various internal organs. Most symptoms arise from a weakness of the Liver, Heart, and Spleen.
3. Deficient Yin is caused by a dysfunction of various internal organs. Most symptoms arise from a deficiency of Yin within the patient's Liver, Heart, Stomach, Lungs, and Kidneys.
4. Deficient Yang is caused by a dysfunction of various internal organs. Most symptoms arise from a deficiency of Yang in the patient's Liver, Heart, Spleen, Lungs, and Kidneys.

In cases of Deficiency, when replenishing Qi is required, the best effect can be obtained when the patient cooperates with the doctor's extension of Qi through mental concentration and focused inhalation.

COMBINED EXCESS AND DEFICIENT CONDITIONS

Sometimes Excess and Deficiency conditions can occur simultaneously. Although Excess is one of the characteristics of a Yang disease, this condition can also give rise to certain symptoms of Deficiency. These are mostly a result of Qi being blocked in the utilization of its own vital nutrients.

YIN AND YANG THEORY

Yin and Yang energy is usually represented through the image of the Taiji symbol. Yang Qi is represented through the color white and Yin Qi is represented through the color black. Through the image of the Taiji symbol the interrelationship of Yin and Yang represent the stages of waxing and waning of energy (Figure 21.4). In Traditional Chi-

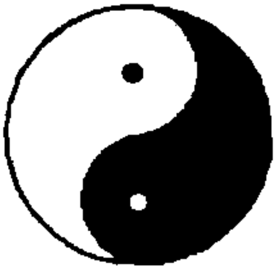
Yang - hot, bright, strong, upward, outward, clear, movement. Qi (non-substantial matter) function of an organ
Yin - cold, dark, weak, downward, inward, muddled, stillness. Blood (substantial matter) substance of an organ.

Figure 21.4. In Traditional Chinese Medicine the theory of Yin and Yang represent the dynamic duality of balance and harmony within the body, as well as within the universe itself, and as represented through the image of the Taiji Symbol. There is always Yin within Yang and Yang within Yin. These two energies are always changing and transforming the body's life force energy.

nese Medicine, the theory of Yin and Yang is also used to explain both the psychological and pathological phenomena of the body. It is considered a major principle for diagnosing and treating diseases.

Generally speaking, an Excess of Yang may lead to a Deficiency of Yin, and vice versa. A Hot syndrome can be the result of either Excess Yang or Deficient Yin. A Cold Syndrome can be the result of either Excess Yin or Deficient Yang (Figure 21.5).

The Qi activities of these Yin and Yang aspects unite and regulate the body's internal organs and tissues. The direction and energetic nature of Qi (Excess or Deficient, strong or weak, evil or vital) can be detected at corresponding body surfaces so that information for diagnosis and treatment can be obtained. For example:

A strong, dense, dry, and hot sensation of Qi in the head area may be a symptom of Excess Yang in the upper portion of the body and a Yin Deficiency in the lower portion. This would require the Qigong doctor to guide the Yang Qi to descend and the Yin Qi to ascend to restore the balance of Yin and Yang energy.

To treat patients with an overabundance of Yang Qi and hyperactivity of Fire, the Qigong prescription meditations should be practiced during the Yin periods (11 a.m. to 11 p.m.) while facing north, with emphasis placed on exhaling to purge the Heat. The doctor should adhere to the principle of replenishing the patient's Yin to regulate the Yang, or leading the Yang to descend to supplement the Yin.

Once the nature of Yin and Yang disequilibrium is determined, the proper method of treatment can be selected to regulate the Excess or Deficiency. The doctor always treats Yin for Yang diseases and Yang for Yin diseases. A Yang syndrome, for example, should be treated with Yin herbs, and a Yin syndrome should be treated with Yang herbs. The treatment should be aimed at purging the Excess, replenishing the Deficiency, and regulating the Deficient Yin or Yang to restore a normal balanced state of energy with in the body.

In Medical Qigong therapy, the principle of "using Yin for treatment of a Yang disease and using Yang for treatment of a Yin disease" must be strictly adhered to. It is important to dredge or purge the Excess before replenishing and tonifying any Deficiencies. The intake of cold foods (fruit, salad, ice cream, etc.) or drinks, for example, may lead to an Excess of Yin, which gives rise to a Cold syndrome of the Spleen and Stomach. This manifests as pain, diarrhea, aversion to cold, and cold extremities.

To treat this Excess, the doctor guides the Cold Qi out of the body by way of St-36 and Sp-9 points, then projects Hot Yang Qi towards specific channel points on the patient's body, such as CV-12 and St-25.

DIAGNOSING YIN AND YANG IN THE BODY

Every person has both Yin and Yang elements, but will tend to be predominantly one or the other in terms of personality, physique, life-style pref-

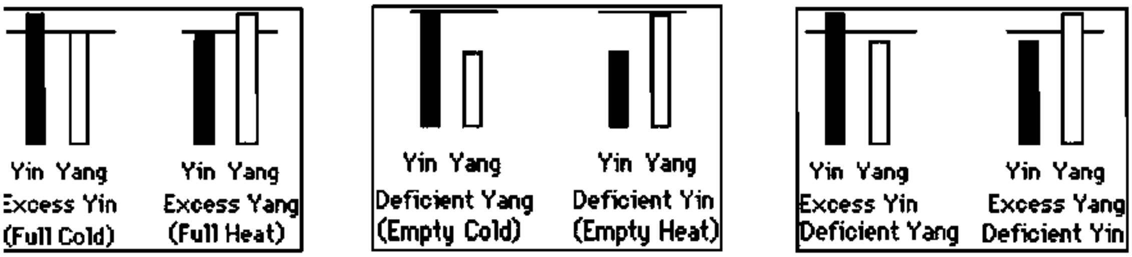

In cases of Excess, the doctor should Purge.

In cases of Deficiency, the doctor should Tonify.

In cases of Excess and Deficieny combined, the doctor should Regulate.

Figure 21.5. Yin and Yang Energetic Diagnosis

erences, speech patterns, mannerisms, etc. Within a general constitution there will be more subtle fluctuations within the Yin/Yang continuum as the body reacts to external and internal energetic movements. In designing the patient's exercise program, diet, or herbal remedies, it is necessary to first determine whether the patient is predominantly Yin or Yang in constitution.

YIN/YANG CONSTITUTION CHART

There are general indications that will help the doctor diagnose a patient's condition as predominantly Yin or predominantly Yang. The patient's basic Yin/Yang constitution, and an analysis of the current Yin/Yang balance of the body will play an important role in determining the type of exercise and the time of day that it will be performed. In most cases, a Yang condition should be balanced by a Yin exercise and vice versa.

There are varying levels of disharmony in the patient's body. Yin and Yang are only relative terms, and they are often not enough to adequately describe the character of the body's condition. Other factors such as Hot - Cold, Deficient - Excess, Internal - External must be considered in more detail. Yin and Yang are never in a permanent state; there is always dynamic movement. All relationships based on Yin/Yang are relative, and their mutual interaction must be considered.

Yin and Yang theory is valuable because it provides a tool for understanding the patient's constitution at a given point in time. This information enables the doctor to determine the best course of treatment, including exercise, food, or herbal formulas that the patient's body needs. Consideration of the patient's inherited and acquired attributes (developed through diet and life-style, environmental influences, natural energy cycles, and other miscellaneous influences such as sleep, work conditions, and relationships) should be based on observation and contemplation of the patient's physical, mental, emotional, and spiritual state of mind.

Designing a program to establish balance and harmony requires a keen awareness of energetic and mental states. In evaluating the patient's condition the doctor takes into account the location and nature of the disharmony, the severity of the condition and the overall physical, mental, and emotional condition of the patient (Figure 21.6). Furthermore, the doctor may notice a combination of both Yin and Yang symptoms or any of the other contrasting pairs be they Hot/Cold, Internal/External, Excess/Deficient. A combination of such symptoms creates a challenge when diagnosing a patient, since symptoms and indicators tend to be more extreme in a person who is ill (Figure 21.7).

YIN CONDITIONS

Primary Yang generates primary Yin. Everything in the universe needs support from Yin. On a basic level, Yin conditions in the body tend to be Cold, internal, Deficient, and degenerative in nature. If a condition becomes overly Yin, it may express some Hot or Yang symptoms. Extreme cold in the form of snow, for example, can cause frostbite. This is called a "false Yang" condition because the condition created is manifested as Yang instead of Yin.

Type of Sign	Yin	Yang
General	• Cool skin temperature • Pale, dark, dull, or yellow complexion • Cold hands and feet • Aversion to cold • Discomfort relieved by warmth and pressure • Senses dull • Curled lying posture	• Warm skin temperature • Red, oily, shiny complexion • Warm hands and feet • Aversion to heat and pressure • Five senses strong • Stretching posture in sleep
Diet	• Prefers warm food and drinks • Poor appetite • Poor sense of taste • Indigestion	• Prefers cold drinks and cold food • Strong appetite • Thirsty, dry mouth
Energy	• Tired • Weak	• Forceful, restless • Strong, stout
Behavior	• Underactive • Quiet, withdrawn	• Overactive • Noisy, outgoing
Breathing & Voice	• Weak, shallow breathing • Weak voice, quiet	• Deep, heavy breathing • Loud, coarse voice, talkative
Excretion	• Copious clear urine • Watery stool	• Scanty dark urine • Hard solid stool
Reproduction	• Sexually underactive • Scanty, pale menses • White leukorrhea	• Sexually overactive • Profuse, red menses • Yellow leukorrhea
Pulse	• Deep, slow • Empty, weak • Fine	• Superficial, rapid • Full, strong • Flooding
Tongue	• Pale, fat body • Moist, thin white coat	• Red, solid, sharp body • Dry, thick yellow coat

Figure 21.6 Signs of Yin and Yang Conditions

Yin conditions are typically characterized by the following: watery stool, clear and copious urine; depression; pale, dull, and yellow complexion; weak, tired, and thin constitution; cool skin, cold hands and feet; dulled senses; curled lying posture when sleeping; soft voice; silent, shallow breathing; preference for warm food; and a tendency towards indigestion. A Yin condition will also be reflected in a deep, weak, fine, slow, or sluggish pulse and a fat, pale, moist, and smooth tongue with possibly a white coating.

YANG CONDITIONS

Yang conditions tend to be Hot, External and Excessive. A Yang disease is generative. Under cer-

Attribute	Progression from Extreme Yang to Extreme Yin
Spirit	Aggressive, outgoing, joyful, sad, fearful, anxious, depressed (The degree of outward expression determines the yangness of the spirit, not so much the emotion itself.)
Body Type	Large and muscular, normal build, thin, emaciated
Posture	Stiff, erect, relaxed, hunched over, limp
Activity	Very animated, normally active, little movement, still
Respiration	Heavy and loud, loud sighing and stretching, breathing heavily through mouth, shallow light breathing, soft intermittent sighs
Voice	Loud, rough, high pitched, regulated and moderate, soft, murmuring, almost inaudible
Skin Cast	Red, yellow, green, pink, white
White of Eye	Red, yellow, blue-green, pink, white
Lip Color	Red, yellow, blue-green, pink, white
Moistness of Lips	Cracked, dry, moist, overly wet
Sputum	Yellow and thick, white and thick, thin, clear
Smell of Breath	Strong, mild, faint
Body Odor	Strong, mild, faint
Urine	Thick, yellow, smelly, thin, whitish, no odor
Feces	Dark, strong smelling, soft, loose, light color

Figure 21.7. The Continuum of Yin and Yang Constitutional Signs and Symptoms

tain conditions, an overabundance of Heat (Yang) may transform into Cold (Yin) to bring about a condition known as "false Yin." Extreme heat and sun, for example, can cause sunstroke. This is called a "false Yin" condition because the condition created manifests as Yin instead of Yang. If an acute Yang condition persists, it will become Yin.

Yang conditions are typically characterized by the following: hard, solid, dry stool; thick, yellow urine; outwardly excited appearance; strong, stout, muscular constitution; restless, active, irritable in character; warm hands and feet and overall warm skin and temperature; strong senses; tendency to stretch frequently; loud voice, talkative; heavy breathing; thirsty, dry mouth; and a preference for cold drinks. A Yang condition will also reflect a floating, flooding, fast, or strong pulse, and a red or yellow, solid, dry tongue with cracks and little or no coating.

SIGNS OF YIN AND YANG

Determining true from false Yang may seem confusing. Yin Deficient people who are very thin and emaciated are often quite hypertensive, this can be seen in their excessive loud talking, rest-

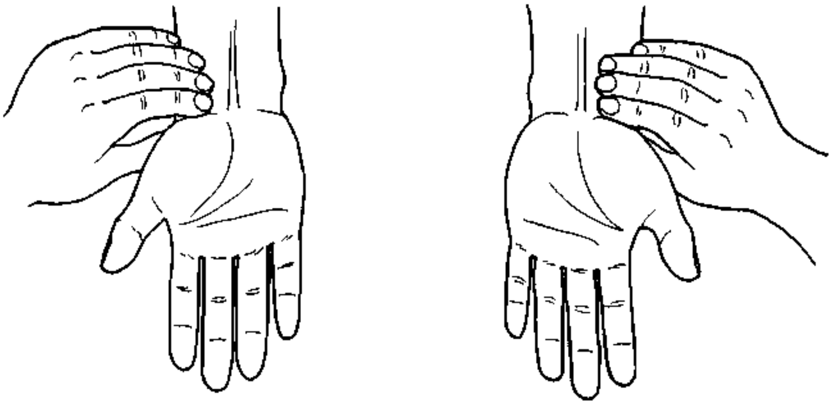

Figure 21.8. The Nine Regions of Pulse Diagnosis

lessness and insomnia. Such individuals can also develop erratically feverish symptoms which usually change from Hot to Cold.

Generally, a true Yin or Yang condition is determined by three or more symptoms that simultaneously manifest either one of those conditions.

False Yin symptoms occur when the Yang becomes overly strong, causing a weakening of the Yin. (Again, the overall condition should be assessed before making a differential diagnosis.) False Yin symptoms include weakness, coldness, wetness, and other Yin symptoms in an otherwise Yang condition. One should bear in mind that there are seldom purely Yin or Yang symptoms because Yin and Yang are so interdependent on each other.

Furthermore, stages of Yin and Yang may change during the course of the day, or over the course of an illness. If a Yang stage is allowed to progress, it can degenerate into weakness described as a Yin stage. Therefore, if a Yang condition changes to Yin it is probably not a good sign, but if a Yin condition changes to an acute Yang condition it may be a positive sign of movement towards overall balance. The focus is to understand that it takes active energy (Yang) to manifest an acute condition, and the stronger the manifestation, the stronger the energy.

YIN AND YANG CATEGORIZATION OF TRAINING

Chinese herbs and foods are typically classified in terms of Yin and Yang, as are Qigong exercises and meditations. With exercises and meditations the state of mind and level of relaxation during the exercise or meditation is critical in determining the Yin or Yang categorization. It is important to be aware of this when developing a patient's program to provide the proper energy adjustments to maintain a healthy balance. If a disharmonious condition develops and the wrong prescription is given, the patient can be thrown further off balance, making it more difficult to restore good health.

PULSE DIAGNOSIS OF THE EIGHT ENERGETIC PRINCIPLES

The body's pulse in considered another form of energy manifestation. The pulse is imagined as waves of various Fluids flowing within the body, originating from within the Heart, where the Shen resides. Pulse diagnosis can give the Qigong doctor detailed information on the state of the patient's internal organs, as it reflects the patient's flow of Qi and Blood, as well as Yin and Yang energy. The pulse is felt at the radial artery, which is divided into three areas (front, middle, and lower), and at three different levels (superficial, middle, and deep), creating "nine regions" of pulse diagnosis (Figure 21.8).

The three areas of the wrist relate to the energetic manifestations of the Triple Burners. The front position corresponds to Heaven and reflects the diseases from the head to the chest; the middle position corresponds to Man and reflects the diseases

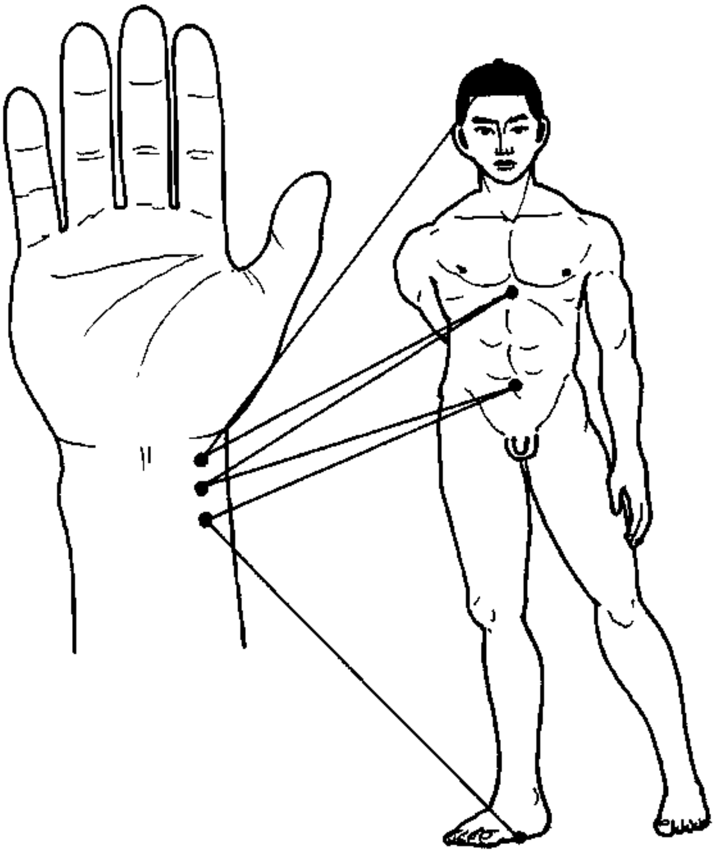

Figure 21.9. Pulse Correspondence with the Triple Burners

from the diaphragm to the umbilicus; the lower position corresponds to Earth and reflects the diseases from the umbilicus to the feet (Figure 21.9).

The three different levels relate to the energetic manifestations of the body's diseases. The superficial level corresponds to exterior diseases, the body's Qi and Yang organ energy; the middle level corresponds to the condition of the Blood, as well as the energy of the Stomach and Spleen; the lower level corresponds to interior diseases, the body's Yin energy and Yin organ energy.

The important thing is to diagnose how the body's energy is flowing, taking into consideration the relationship of the Yin and Yang on the pulse, i.e., whether it is superficial or deep, fast, or slow, Excess or Deficient. The quality of the pulse may also vary according to the patient's:

1. General Constitution and Body Build, with bigger frame patients generally having a stronger pulse then smaller frame individuals,
2. Level of Activity, with patients who engage in heavy physical work having a stronger pulse then those who engage in mental work, and
3. Exposure to Seasonal Weather, as a deeper pulse is generally felt in the winter and a more superficial pulse is felt in the summer.

PULSE QUALITIES AND CLASSIFICATIONS

In China, attention is placed primarily on the overall quality of the pulse. While reading the superficial pulses, it is important to note that they indicate the condition of the patient's Qi and Yang Organs. The deep pulses indicate the condition of the patient's Jing, and Yin Organs. Pulses can be divided into three main classifications, which are related to either Yin and Yang qualities: The Depth, The Speed, and The Strength (Figure 21.10).

1. The Depth at which the pulse can be felt is considered first.
   a. Superficial/Floating Pulses are felt with a light touch, barely touching the artery. They indicate an Exterior syndrome (e.g., Wind-Cold or Wind-Heat), chronic illness, or a general weakness. This informs the doctor that there is an Excess at the patient's Qi level (or a Deficiency at the Yin level).
   - Superficial and Weak (Empty) pulses are felt in Yang Excess diseases (superficial diseases), when the patient's Wei Qi is fighting the pathogenic factors at the superficial level.
   - Superficial and Forceful (Full) pulses are felt in interior diseases due to Yin or Jing Deficiency.

   b. Deep Pulses are felt on heavy pressure, near the bone. They indicate an interior syndrome (problems within the Yin Organs), and relate to the patient's Jing.
   - Deep and Weak (Empty) pulses are felt in diseases due to a Deficiency of Qi and Yang.
   - Deep and Forceful (Full) pulses are felt in deep diseases (when the disease has penetrated into the patient's internal organs). They confirm an Excess at the Yin level, or Deficiency at the Qi level, and point to a stasis of Qi or Blood, and to a Cold or Heat condition.
2. The Speed of the pulse is considered next.
   a. Slow Pulses are less then 4 beats per the patient's breath; they indicate a Cold syndrome.
   - Slow and Weak (Empty) pulses indicate Empty Cold and Yang Deficiency.
   - Slow and Forceful (Full) pulses indicate Full Cold and Yin Excess.

   b. Rapid Pulses are more then 5 beats per patient's breath and indicate a Heat syndrome.
   - Rapid and Weak (Empty) pulses indicate Empty Heat and Yin Deficiency.
   - Rapid and Forceful (Full) pulses indicate Full Heat and Yang Excess.

	YANG	**YIN**
DEPTH	Superficial — Exterior	Deep — Interior
SPEED	Rapid — Hot	Slow — Cold
STRENGTH	Forceful — Excess	Weak — Deficient

Figure 21.10. The Eight Energetic Principles are manifested through the body's pulses.

	LEFT HAND **Superficial - Deep**	**RIGHT HAND** **Superficial - Deep**
FRONT POSITION	Small Intestine - Heart	Large Intestine - Lungs
MIDDLE POSITION	Gall Bladder - Liver	Stomach - Spleen
LOWER POSITION	Bladder - Kidneys' Yin	TB & Pc - Kidneys' Yang

Figure 21.11. Left and Right Hand Correspondence

3. The Strength of the pulse is considered last.
   a. Weak (Empty) Pulses are weak and disappear on heavy pressure; they indicate a Deficient syndrome.
   b. Forceful (Full) Pulses are forceful and are felt on deep pressure; they indicate an Excess syndrome.
   - Forceful (Full) and rapid pulses indicate Full Heat.
   - Forceful (Full) and slow pulses indicate Full Cold.

THE CLASSICAL PULSE DIAGNOSTIC POSITIONS AND DEPTH

Traditionally the best time to take the patient's pulse is when the patient is calm and relaxed. There are several schools of pulse diagnosis, each one with its own specific methods of approach to understanding the patient's symptoms. The most important aspect in listening to the patient's pulse is to feel the strength, quality, and spirit of the pulse. Only then can the doctor receive accurate information about his patient's condition (Figure 21.11).

Pattern	Signs	Pulse	Tongue
Internal	Vomitting, high fever, gradual onset, chronic, long duration	Deep	Changes in texture and coating
External	Cold, flu, skin eruptions, sudden onset, acute, short duration	Superficial	Changes in coating
Cold	Diminished physiological function, decreased energy, lowered resistance, lowered immunity, white face, aversion to cold, cold extremities, lack of thirst, craving for warmth, clear urine, watery stool, desire for hot drinks, diarrhea, coolness of limbs, timid behavior, soft voice, slow movement, lack of motivation	Slow, tight	White coat, pale body, moist fur
Heat	Physiological hyperfunction, increased energy, increased metabolism, acute pains, fever, thirst, desire for cold liquids, flushed face, red eyes, restlessness, irritabiliity, dry stool, dark yellow urine, reduced urine secretions, aversion to warmth, hot red skin, constipation, warm extremities, loud voice, talkative, extroverted manner, aggressive behavior	Rapid	Red tongue with yellow fur
Deficiency	Great weakness, low syptomatic reactions, inconsistent energy, shallow and rapid respiration, less severe pain, pain characterised by soreness, weak or inconsistent voice, dull menstrual pain, withered face, low spirits, abdominal pain	Empty, weak	Little or no coat
Excess	Flushed face, loud voice, coarse and full breathing, sharp and acute pain, extreme menstrual pain, tender abdomen, condition worsens with apllication of heat	Full	Thick coat

Figure 21.12. Signs of Internal, External Cold, Hot, Deficient and Excess Conditions

Usually, in men, the pulse on the left (Yang) side of the body should be slightly stronger then that of the pulse on the right (Yin) side, and vice versa for women. Also in men, the front (Yang) positions should be slightly stronger, while in women the lower (Yin) positions should be stronger.

OVERVIEW

In the clinic we have a saying, "Where the Qi goes, the Blood flows." Although a patient's pulse is produced by the movement of Blood in the arteries, the activity of that movement is initiated by the force of the patient's Qi. By feeling the movement of Blood in the vessels, the activity of the patient's Qi, Blood, and Body Fluids are diagnosed. This internal communication is summarized in the above chart (Figure 21.12).

CHAPTER 22
DIAGNOSIS ACCORDING TO QI, BLOOD, AND BODY FLUID DYSFUNCTIONS

INTRODUCTION

This method of diagnosis identifies patterns on the basis of disharmonies according to Qi, Blood, and Body Fluids. It is used in clinical practice (especially for internal diseases) for diagnosing according to the pathological changes occurring in the internal organs.

The essential cause of disease is due to an imbalance of Qi, Blood, or Body Fluids. When one or more of these substances are unbalanced, the body's normal physiological function is disrupted and sickness results. This can be caused by the effect of the Six Exogenous Pathogenic Factors, Seven Emotional Pathogenic Factors, and is influenced by the patient's general constitution. Each of the patient's organs and channels have sets of symptoms or syndromes of these types of imbalances, which can then be identified for treatment. The Qigong doctor's ability to identify problems based on the pathological manifestations of Qi, Blood, and Body Fluids is founded on clinical observations of Excess, Deficient, Stagnant, and Rebellious patterns (Figure 22.1). These patterns can be further defined as predominantly Yin or Yang. These patterns also described as the various disorders of Qi, Blood, and Body Fluids, and sometimes overlap, for example:

- Excess conditions can give rise to Rebellious Qi and also create stagnation;
- Deficient conditions can give rise to Rebellious Qi and also create stagnation.

DISORDERS OF THE QI

Whenever energy movement is impeded, disorders in Qi functions occur. Generally, these dis-

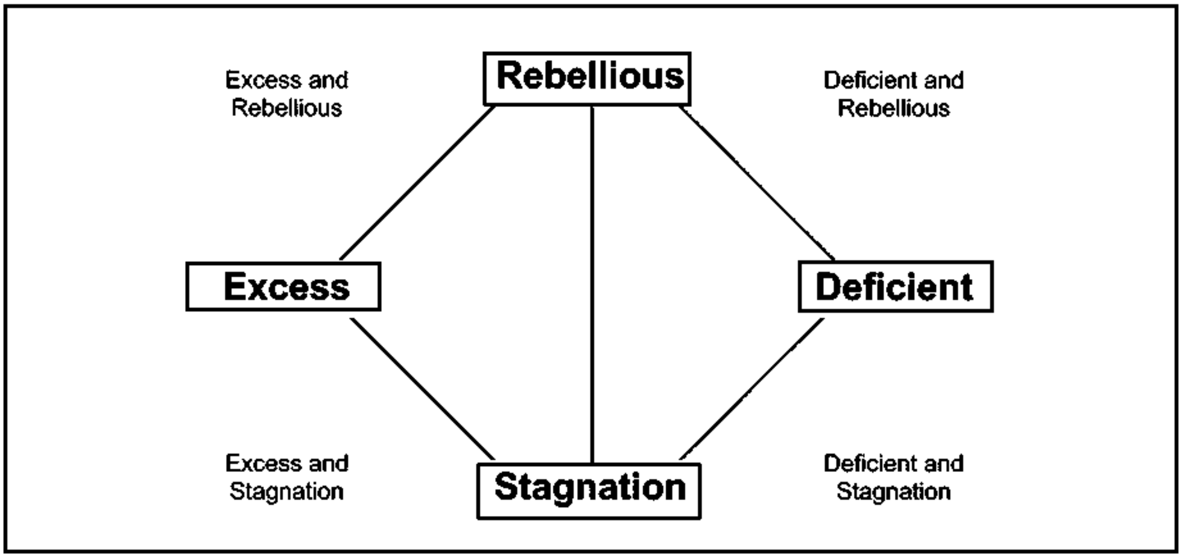

Figure 22.1. Pathological Manifestations of Qi, Blood, and Body Fluids

Fire		Water
Constraint of Qi	First Stage	Stasis of Qi
Arrogant/Agitated	Second Stage	Fluid Builds Up
Full on Fire is Created	Third Stage	Phlegm is Created
Heart Fire is Created	Final Stage ——> Mania <——	Phlegm Fire is Created

Figure 22.2. Disorders caused from the obstruction of Qi flow result in either Fire or Water energetic dysfunctions.

turbances cause either disorders of Fire, or disorders of Water (Figure 22.2).

The cause and progression to these disorders can be further broken down and classified according to their energetic dysfunction (Figure 22.3). There are several types of Qi disorders observed in the clinic and are generally described as follows.

LEAKAGE AND LOSS OF QI

During or after the training practice, the patient may feel that Qi is continuously and uncontrollably leaking out of the body from the genitalia and anus or anywhere else on the body. This loss of Qi can lead to Qi Deficiencies, and manifest through the following symptoms: emaciation, weakness of the limbs, grayish and dull complexion, nervousness, mental disturbances, distractibility, failing memory, spontaneous sweating, night sweats, seminal emission, insomnia, lassitude, and sluggishness.

DEFICIENT QI DISORDERS

Deficient Qi exists when the entire body, or particular organ, is not being sufficiently nourished by the Qi. This results in there being too little energy to maintain proper organ function. This condition is usually caused by a weakness of the Lungs, Spleen, or Kidneys, by chronic illness, or a weak constitution, poor diet, or fatigue. Deficient Qi is unable to:

- transform, resulting in symptoms such as diarrhea,
- transport, resulting in symptoms such as stagnation of Qi,
- hold, resulting in symptoms such as bruising,
- warm, resulting in symptoms such as cold extremities,
- protect, resulting in symptoms such as sickness, and
- raise, resulting in symptoms such as prolapse.

Deficient Qi is divided into three subcategories from which the patient can suffer. These subcategories are Empty Qi, Sinking Qi, and Qi Collapse. Each of these subdivisions can further be divided into Yin and Yang types.

1. Empty Qi is considered the first type of Deficiency. If not treated, Deficient Qi may progress to a condition known as Empty Qi. Diseases that are classified as Empty Qi are characterized by a serious weakness or Deficiency of the body's Qi. Empty Qi, especially Empty Yang Qi manifests as a weakness

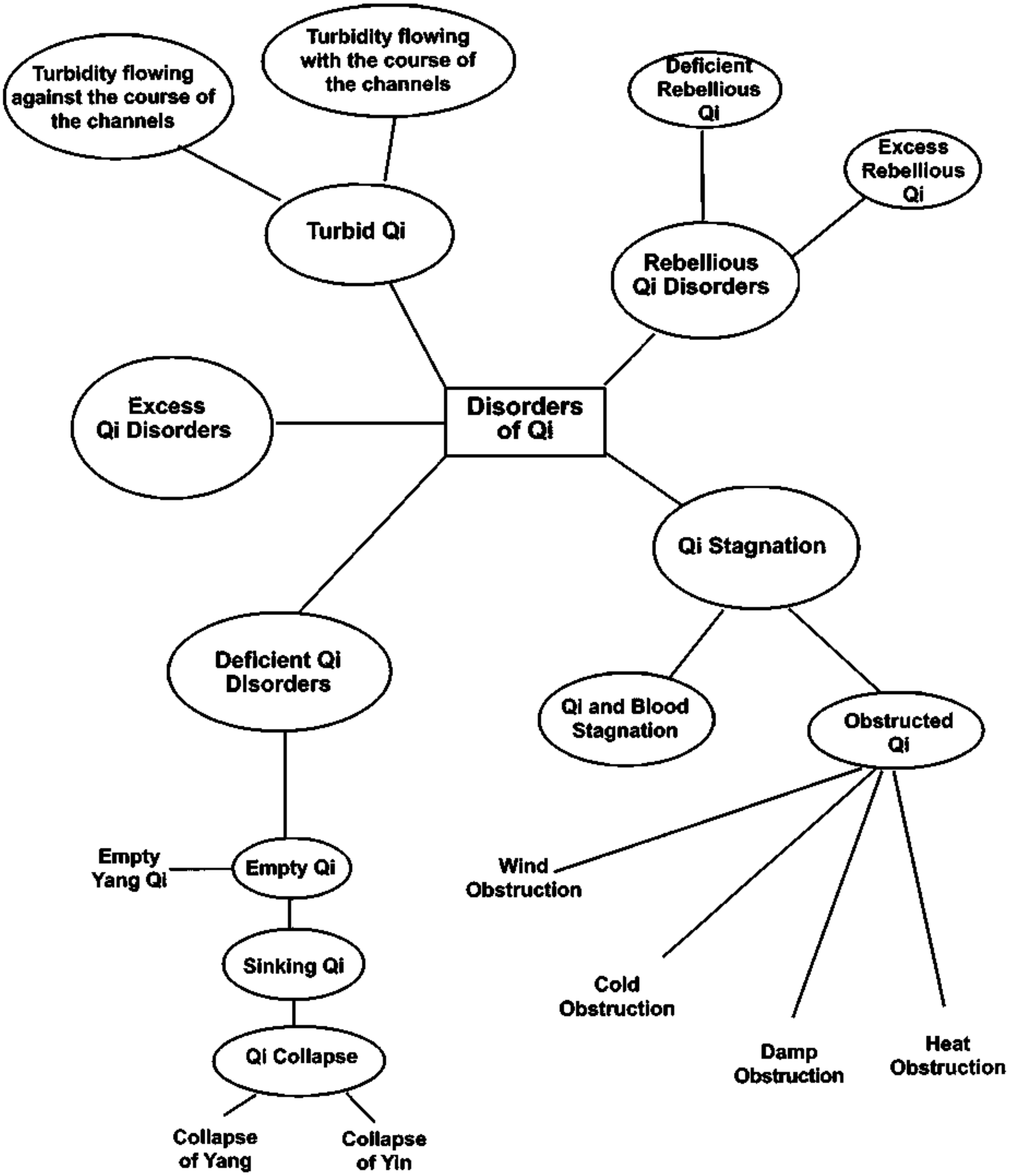
Turbidity flowing against the course of the channels
Turbidity flowing with the course of the channels
Turbid Qi
Deficient Rebellious Qi
Excess Rebellious Qi
Rebellious Qi Disorders
Excess Qi Disorders
Disorders of Qi
Qi Stagnation
Deficient Qi Disorders
Qi and Blood Stagnation
Obstructed Qi
Empty Yang Qi
Empty Qi
Wind Obstruction
Sinking Qi
Cold Obstruction
Qi Collapse
Damp Obstruction
Heat Obstruction
Collapse of Yang
Collapse of Yin

Figure 22.3. Whenever energy movement is impeded or in excess, disorders in Qi function occur, resulting in Qi Deviations.

in the function of holding, moving, and transforming Qi.

   a. Empty Qi originates primarily from the patient's constitution and manifests as malnutrition, fatigue, premature signs of old age, and an excess of extreme emotions.
   b. Empty Yang Qi belongs to the category of Yang. Because Yang Qi has the function of warming, Empty Yang manifests as an inability to adequately warm the energy, resulting in Cold symptoms. Empty Yang is generally caused by a weak constitution, poor diet (excessive raw, cold food and drink), overwork, old age, and excess sex.

2. Qi Sinking is also a subcategory of Deficient Qi, and is considered the second type of Deficiency. If not treated, Empty Qi may progress to a more serious Deficient condition known as Qi Sinking. This condition results in symptoms such as tiredness, listlessness, and mental depression, and may eventually lead to the prolapse of one or more of the organs.
3. Qi Collapse is a subcategory of Deficient Qi and is considered the third and most severe type of Deficiency. A Qi collapse exists when the Qi is so Deficient that it can no longer hold the internal organs in place. This Qi depletion not only causes a weakness of the body's Righteous Qi, which leads to the prolapse of the organs (Stomach, uterus, intestines, rectal area, vagina, or Bladder), but also induces prolonged diarrhea, upper abdominal distention, and dizziness.
   a. The Collapse of Yin results in pure Fire. The main manifestations are abundant perspiration and external tissue, which is hot to the touch.
   b. The Collapse of Yang results in pure Cold. The main manifestations are chills and external tissues which are cold to the touch.

EXCESS QI DISORDERS

An Excess Qi condition is caused from too much Qi, which overacts on an organ, affecting its function, and usually produces too much transformation and Heat in a given organ. This can be due to over-energizing the organs and channels, from excessive Qigong exercises and meditations, or too much energy stimulation during a Qigong doctor's treatment.

QI STAGNATION DISORDERS

Qi stagnation exists when the Qi is not flowing smoothly, and becomes sluggish when moving within the body's channels, organs, or tissues. Qi stagnation can be caused by many different factors such as: suppressed emotions, trauma, contraction of microorganisms, poor diet, and weak digestion. Qi stagnation, over time, results in localized pain, distention, cysts, tumors, etc. Qi stagnation disorders have two subcategories: Qi and Blood Stasis, and Obstructed Qi.

1. Qi and Blood Stasis refers to the sensation of local pain, heaviness, soreness, distention, and compression, which are caused by the stagnant Qi and Blood in certain areas of the body. These sensations do not disappear by themselves. This stasis may occur during or after Medical Qigong training practice from Qi disorders and visceral malfunctions.
2. Obstructed Qi is also considered a subcategory of Qi stagnation, and exists when the Qi becomes so stagnant that it no longer flows. At this stage the Qi can not move the Blood. This can be caused by trauma, surgery, and acute Qi stagnation.

When Qi becomes obstructed inside the body, certain clinical manifestations result, such as:

- Cold Obstruction causes Qi to become fixed with local numbness that improves with heat;
- Damp Obstruction causes Qi to become fixed with local swelling;
- Heat Obstruction causes Qi to become inflamed, red, and swollen;
- Damp Heat Obstruction is a combination of Damp and Heat. It is called "summer Damp Heat," and results in poor appetite, dizziness, diarrhea, heaviness of the head, and sensations of chest suffocation;
- Wind Obstruction causes migrating pain that moves from joint to joint.

Organ	Normal Flow	Rebellious Flow	Symptoms
Liver	Upwards	Excessively Upwards Horizontally to Stomach Horizontally to Spleen Horizontally to Intestines Downwards	Dizziness, Headache, Irritation Nausea, Belching, Vomiting Diarrhea Dry Stool Burning Urination
Heart	Downwards	Upwards	Insomnia, Mental Restlessness
Spleen	Upwards	Downwards	Diarrhea, Prolapse of Organs
Stomach	Downwards	Upwards	Hiccup, Nausea, Belching, Vomiting
Lungs	Downwards	Upwards	Coughing, Asthma
Kidneys	Downwards	Upwards	Asthma

Figure 22.4. Chart of Rebellious Qi Flow

REBELLIOUS QI DISORDERS

Rebellious Qi disorders manifest as Qi flowing in the wrong direction (a direction different from its normal physiological progression); this causes a disturbance within the harmonious flow of energy and affects the organs' functions. The symptoms have to be analyzed according to the internal organ which is involved. There are two types of Rebellious Qi, the Excess and Deficient type. Rebellious Qi is categorized as follows:

DISORDERS OF EXCESS REBELLIOUS QI

Disorders of Excess Rebellious Qi occur when the vital Qi increases and forces the pathogenic Qi to move; leading to local sensations of pain, soreness, distention, heaviness, cold or heat, which can develop during a patient's Qigong practice (Figure 22.4).

DISORDERS OF DEFICIENT REBELLIOUS QI

Disorders of Deficient Rebellious Qi refers to the disorderly flow of Qi and Blood during or after the training practice, which causes dizziness, fright, a heavy feeling in the chest, shortness of breath, shaking limbs, trembling of the hands and feet, or fainting. Generally, energy sensitive patients know the location and direction of the disorderly flow of Qi in their tissues.

TURBID QI

While transforming and transporting energy, the body divides ingested and absorbed Qi into Clear and Turbid energy for distribution (keeping and redistributing the clean, while disposing of the turbid). Turbid Qi refers to any dark, murky energy which can originate within, or from outside of the body's organ system, and sometimes travels throughout the body's channel system. Turbid Qi attaches to the body's wounds (physical, emotional, or spiritual), manifesting as dark, discolored clouds of energy. This energy can adversely affect the body's physical, emotional, and energetic balance, especially when combined with any pathogenic internal or external factor.

PATHOGENIC SENSING OF TURBID QI

Feeling the sensations of Turbid Qi is referred to as "pathogenic sensing." This skill is acquired by the doctor while projecting energy to his patients. While emitting energy, the Qigong doctor can sense the patient's Yuan Qi, as well as the Turbid Qi inside the patient's body. This enables the doctor to diagnose and treat their patients' diseases according to the various Qi sensations, and to categorize the patients' Turbid Qi. The Turbid Qi is also called Evil or "pathogenic Qi." The flow of Turbid Qi within the patient's channels is characterized as follows:

- Turbidity flowing with the course of the channel indicates transference of Turbid Qi towards the body's interior.
- Turbidity flowing against the course of the channel indicates the beginning production of Turbid Qi.

Turbid Qi differs from the pathogenic factors of infectious diseases referred to in Western medicine. According to clinical experience, the pathogenic sense of Turbid Qi can be classified as: Dry or Hot Sensations, Cold or Chilly Sensations, Soreness or Numbing Sensations, Impure or Foul Sensations, The Sensations of The Seven Emotions, and The Sensations of The Six Climate Conditions.

1. Dry or Hot Sensations are considered Yang signs or symptoms of Heat. This type of Turbid Qi feels feverish and flaccid; it can cause perspiration, irritation, etc. The sensations of dry-heat on the body or hands of the doctor often cause the doctor to fidget, as if being scorched by a fire. It may also serve as an indicator of syndromes of Excess Heat and hyperactivity of Fire due to a Yin Deficiency.
2. Cold or Chilly Sensations are considered Yin signs or symptoms of Cold. This type of Turbid Qi feels cold, rigid, and stagnant. It may be so cold that the doctor's fingertips turn cold immediately. This coldness transmits from the doctor's fingertips upward, causing shivering and contractions of the sweat glands ("goose-bumps"). This gives the doctor a particular feeling of cold and discomfort. Cold feelings often serve as a message of Deficiency and Cold syndromes such as rheumatism, rheumatoid disease, Deficiency of the Kidney Yang, and certain types of cancer.
3. Soreness or Numbing Sensation often indicate syndromes of hyperactivity of the Liver Yang and Excess of Wind, Cold, Dampness or Phlegm, as well as arthralgia syndrome and cancer. The doctor may experience an uncomfortable feeling of local soreness or pain, heaviness, numbness, tingling and itching.
4. Impure or Foul Sensations can be felt when the doctor is standing opposite the patient, or when extending energy towards the patient. It is an indescribable unpleasant feeling that is often felt when the open palm and fingers are used to detect a disease.
5. The Seven Emotions (joy, anger, worry, anxiety, fear, shock, and grief) in excess will cause discomfort to the doctor. During treatment, the doctor may detect and feel the different natures of these emotional pathogenic Evils flowing through the doctor's body. The doctor may, for example, have a specific feeling of anger if the disorder of the patient is due to excessive anger. Once these pathogenic emotions are discovered, the doctor must purge or disperse them, as well as determine their origin, and cause.
6. The Six Climate Conditions (Wind, Cold, Summer Heat, Dampness, Dryness, and Fire) will also cause discomfort to the doctor. During treatment, the doctor may detect and feel the different natures of these pathogenic Evils. The doctor may, for example, have a specific feeling of slippery wet ooze if the disorder of the patient is due to Dampness.

RECTIFYING QI DISORDERS

The primary treatment principles for rectifying Qi Disorders are as follows:

- for Qi leakage - tonify
- for Deficient Qi - tonify
- for Empty Qi - tonify
- for Qi sinking - raise and tonify
- for Collapsed Qi - tonify
- for Excess Qi - reduce
- for Stagnant Qi - move
- for Obstructed Qi - move
- for Rebellious Qi - calm and subdue
- for Turbid Qi - remove, dredge, and purge

These are often found in combination with other patterns. Here are some examples of treating combined patterns:

- for Excess and Rebellious Qi, reduce Excess and calm Rebellious Qi;
- for Deficient and Rebellious Qi, tonify Deficient and calm Rebellious Qi;

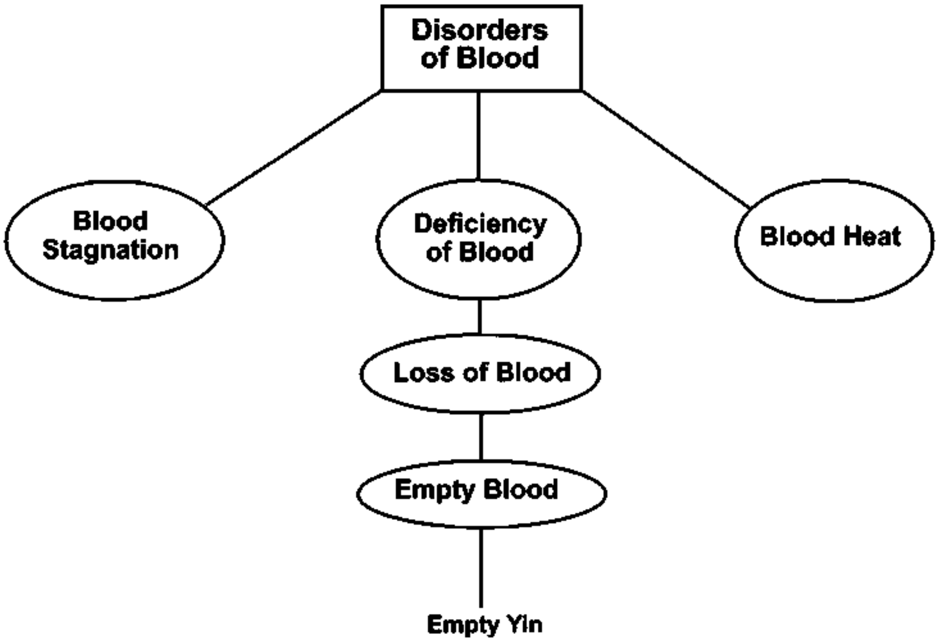

Figure 22.5. Whenever Blood movement is impeded or in excess, disorders in Blood function occur.

- for Excess and Stagnant Qi, reduce Excess and move the Stagnant Qi;
- for Deficient and Stagnant Qi, tonify Deficient and move the Stagnant Qi;
- for Excess and Deficient Qi, drain Excess and tonify Deficient Qi; and
- for Rebellious and Stagnant Qi, calm Rebellious and move the Stagnant Qi.

DISORDERS OF THE BLOOD

To understand Blood disorders it is important to understand the concept of Blood from the perspective of Traditional Chinese Medicine. In Chinese Medicine, Blood is different in concept, characteristics and function from that of Western medicine. Blood originates from the transformation of food and drink by the Spleen, which then transfers the refined food energy (Gu Qi) to be further enhanced by the Heart, Lungs, and air energy. Blood is therefore considered a transformation of constructive energies.

Disorders of the Blood can relate to disorders of the Qi. Qi and Blood flow together, Qi being the active force which makes the Blood circulate and keeps it within the blood vessels. Both Qi and Blood flow within the channels and blood vessels, continuously circulating throughout the body nourishing, maintaining, and moistening the tissues. Qi is an energetic form and is considered a Yang substance, while Blood is a liquid form of energy and considered a Yin substance. The cause and progression to Blood disorders can be further broken down and classified according to their energetic dysfunction (Figure 22.5). There are several types of Blood disorders observed in the clinic and are generally described as follows.

DEFICIENCY OF BLOOD

A Deficiency of Blood exists when the entire body, or a particular organ, is insufficiently nourished by the Blood, or not enough Blood is present to nourish the organs and channels, to support the Qi. Blood Deficiency is usually caused from Blood loss, chronic illness, poor diet, or Spleen Qi Deficiency, which weakens the Qi. This in turn affects the Liver (causing symptoms of dizziness, blurred vision, and numbness) and Heart (causing symptoms effecting the mind, memory, and Shen).

LOSS OF BLOOD

Loss of Blood is considered a subcategory of Deficient Blood. Loss of Blood is caused by Spleen Qi failing to control the Blood, resulting in the Blood spilling out of the vessels and organs, etc. Loss of Blood is divided into two main types:

- Deficient loss of Blood is caused by an insufficiency of Qi to hold the Blood.
- Excess loss of Blood, is caused by either Blood-Heat pushing the Blood out of the vessels, or by Blood stagnation and Yin Deficiency which causes the Blood to pool.

EMPTY BLOOD

Although Empty Blood is a subcategory of Deficient Blood, it is considered a serious Deficient condition, mostly due to the Blood of the whole body becoming Deficient. Empty Blood corresponds to a Deficiency of Qi, caused by:

- Inadequate nutrition,
- Loss of Blood through accident or disease,
- Loss of Body Fluids through excessive sweating, urination, vomiting, or diarrhea,
- Extreme emotions (e.g., rage, terror and shock),
- Blood stagnation, which leads to poor Blood circulation.

Since the Empty Blood fails to nourish the tendons and muscles, it results in shaking in the extremities and head, dizziness, and blurred vision.

Empty Yin

Blood belongs to the category of Yin. Because Yin Qi has the function of cooling and moistening, when an Empty Yin condition is present, the symptoms manifest as Excess Heat and Dryness. Empty Yin is caused by a weak constitution and a poor diet (overcooked foods and hot liquids). The organs most likely to be effected by the Empty Yin condition are the Kidneys, Lungs, Heart, Liver, and Stomach.

BLOOD STAGNATION

Blood Stasis exists when the Blood is unable to flow smoothly, and becomes obstructed or congeals. Blood stasis is usually caused by an impact trauma, Deficient Blood, stagnant Blood, Blood Heat, Blood Cold, or stagnant Qi. It results in localized pain which worsens with pressure, swelling of the internal organs, cysts, and tumors. The organ most often affected by Blood stasis is the Liver.

BLOOD HEAT

Blood Heat is mostly due to Liver Heat and External Wind. When toxic Heat enters the Blood, the Blood may break out of the vessels and cause: nose bleeding, hemorrhaging, and broken blood vessels. Heat consumes the Blood and Yin, injuring the tendons and muscles, causing spasms and rigidity. The Blood Heat also injures the Governing Vessel, Pericardium, and the Shen.

The Four Main Treatment Principles

- for a Deficiency of Blood, tonify
- for Empty Blood, tonify
- for Blood Stasis, move Stasis
- for Blood Heat, cool down

THREE DISORDERS OF BODY FLUIDS

To understand disorders of the Body Fluids it is important to understand the concept of Body Fluids, from the perspective of Traditional Chinese Medicine. Body Fluids are a denser form of Qi, and depend on Qi for transformation and transportation. Qi occupies and directs the movement of Blood and Body Fluids. The function of the Fluids is to moisten and nourish the body's tissues, as well as thin the Blood to prevent stasis. Body Fluids are formed from ingested food and drink, and are then transported by the energy of the Spleen to the Lungs and Small Intestines for distribution.

Disorders of the Body Fluids can relate to disorders of the Blood. There is a constant interchange between Fluids and Blood, such as, for example, a Deficiency of Fluids can be developed from a heavy loss of Blood. Body Fluids are divided into two categories:

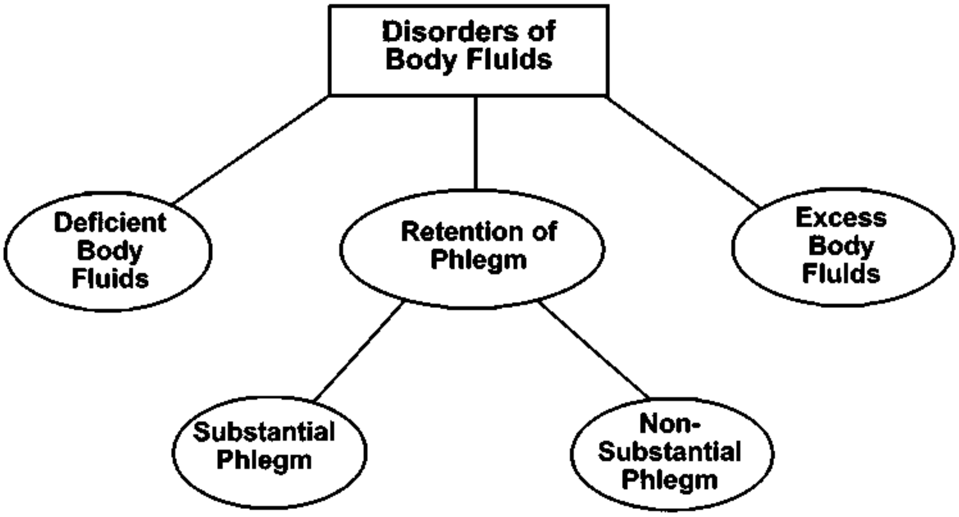

Figure 22.6. Whenever Body Fluid movement is impeded or in excess, disorders in Body Fluid function occur.

1. Jin refers to the light, thin, watery, clear, quick-moving Fluids and is considered Yang. These Fluids are under the control of the Lungs which circulate them with the Wei Qi. These Fluids also permeate and moisten the skin and have the function of warming and nourishing the muscles. The Upper Burner is responsible for controlling the Jin transformation and moving it towards the skin. The Jin is manifested in the body in the form of sweat, tears, saliva, mucus, and parotid serum.
2. Ye refers to the heavier, thicker, turbid, dense, slow moving Fluids and is considered Yin. These Fluids are under the control of the Middle and Lower Burners, which circulate the Ye with the Ying Qi. The Spleen and Kidneys are responsible for controlling the Ye transformation. This Fluid lubricates the joint cavities (synovial fluid), nourishes and lubricates the brain and spinal cord (cerebrospinal fluid), bone marrow and the orifices of the sensory organs (eyes, ears, nose, and mouth).

There are three primary disorders of Body Fluids observed in the clinic, they are described as follows (Figure 22.6).

1. Deficient Body Fluids exists when the entire body, or a particular organ, is insufficiently nourished by Body Fluids. This is generally due to insufficient Yin to create or nourish the body's organs or joints. Deficient Body Fluids are usually caused by prolonged loss of Fluids (sweating, diarrhea, and vomiting), prolonged illness, a Heat condition, or a Yin Deficiency which leads to Dryness. A Deficiency of Body Fluids usually affects the Lungs, Kidneys, Stomach, and Large Intestines.
2. Excess Body Fluid accumulation is usually caused by an Excess Yin and one of three factors, or a combination of all of them. They are:
   - A Lung Qi Deficiency, which affects the top of the body, face and hands,
   - A Spleen Qi Deficiency, which affects the middle part of the body, and abdomen,
   - A Kidney Qi Deficiency, which affects the lower part of the body, legs and ankles.

If there is a Deficiency within one or all three of these organs, and the Body Fluid is not able to transform properly, it will overflow the channels and settle under the skin.

3. Retention of Phlegm is due primarily to a Deficiency of the Spleen failing to transport

and transform the Body Fluids. If Phlegm is retained for long periods of time, disease of the organs (Substantial Phlegm), as well as channels, skin, and joints (Non-Substantial Phlegm), can form.

The Three Main Treatment Principles

- for Deficient Body Fluids, moisten;
- for Excess Body Fluids, dry or drain;
- for Retention of Phlegm, resolve.

TYPES OF PHLEGM

The concept of Phlegm is an important pathological and etiological factor in Traditional Chinese Medicine. If Phlegm is allowed to accumulate and becomes retained within the body, it causes disease.

The main cause of Phlegm formation is Spleen Deficiency. If the Spleen becomes weak and fails to transform and transport the Body Fluids, these will accumulate and transform into Phlegm. However, Phlegm formation can also occur if the Lungs fail to disperse and lower Fluids, or if the Kidneys fail to transform and excrete Fluids.

In Traditional Chinese Medicine, there are two types of Phlegm: Substantial Phlegm and Non-Substantial Phlegm.

1. Substantial Phlegm is described in ancient texts as "having a form." It can be observed as the thick mucus that collects within the Lungs and spat out during bronchial infections and other Lung diseases. As the Lung disease progresses, the patient's mucus changes color, progressing from clear watery to yellow and thick, then to green and pussy, and finally brown and red.
2. Non-Substantial Phlegm is described in ancient texts as "having no form." This type of Phlegm can accumulate within:
   - the body's organs forming stones,
   - the joints forming arthritic bone deformations,
   - the skin forming lumps or cysts, and
   - the channels causing stagnation of Qi and Blood, resulting in numbness.

Both Substantial and Non-Substantial Phlegm can assume different forms of disease, developing in accordance to its associations with Heat, Cold, Damp, Wind, Qi, and Fluids. For example:

1. Phlegm Heat is a type of Substantial Phlegm that affects the Lungs, Stomach, or Heart. It manifests as yellow sticky phlegm, a red face and a red tongue with sticky yellow coating.
2. Cold Phlegm is a type of Substantial Phlegm that affects the Lungs and Stomach. It manifests as white watery phlegm expectoration, cold feelings in the back and extremities, and a pale tongue with a white and wet coating.
3. Damp Phlegm is a type of Substantial Phlegm that affects the Lungs. It manifests as a profuse white sticky phlegm, feelings of stuffiness in the chest and epigastrium region, and a sticky tongue coating.
4. Wind Phlegm is a type of Non-Substantial Phlegm and is seen in Wind Stroke. It causes dizziness, nausea, vomiting, coughing up of Phlegm, numbness of the limbs, rattling throat, and aphasia.
5. Qi Phlegm is a type of Non-Substantial Phlegm which affects the throat. It manifests as a "plum pit" syndrome within the throat. Its formation is related to emotional problems, created when the Liver Qi rises and stagnates within the throat. Its condition manifests as a feeling of swelling in the throat, difficulty in swallowing, and stiffness in the chest and diaphragm.
6. Phlegm Fluids is a type of Substantial Phlegm which has a fluid or watery nature, and is called Yin in Chinese medicine. There are four types according to their location:
   - Phlegm Fluids located in the Stomach and intestines manifest as abdominal fullness, distension, and vomiting;
   - Phlegm Fluids located above the diaphragm manifest as coughing, asthma, edema, and dizziness;
   - Phlegm Fluids located within the hypochondrium manifest as distension in the hypochondrium area, with pain and coughing;
   - Phlegm Fluids located in the four limbs manifest as heaviness and pain in the muscles.

CHAPTER 23
DIAGNOSIS ACCORDING TO THE FIVE-ELEMENT THEORIES

INTRODUCTION TO THE FIVE ELEMENTS

This method of diagnosis is based on the interpretation of clinical manifestations according to the creative, controlling, invading, and insulting sequences of the Five Elements. The Five Elements are five phases of transition used in Chinese Medicine and are sometimes referred to as the Wu Xing. Wu translates to mean the number five, Xing translates to mean a process of movement or walking. The concept of the Five Elemental theory began to first appear in documents in China during the Zhou Dynasty (from 1000 to 770 B.C.). This theory classified tangible and intangible substances into five categories for observation and study, as well as diagnosis and treatments.

1. The Wood Element is associated with the active functions of birth in nature's growth stage of development.
2. The Fire Element is associated with maximum functional activity in nature's growth stage of development.
3. The Earth Element is associated with the function of stabilization in nature's growth stage of development.
4. The Metal Element is associated with nature's declining functions of the growth stage of development.
5. The Water Element is associated with the maximum state of rest in nature's growth stage of development.

The Five Elemental theory is used to explain the classifications and characteristics along with the laws of universal cycles (creating, controlling, invading and insulting). These interactions are studied today in Traditional Chinese Medicine to assist the doctor's diagnosis in comprehending the growth and development of the body's energetic anatomy, physiology, disease processes, and symptom development (Figure 23.1).

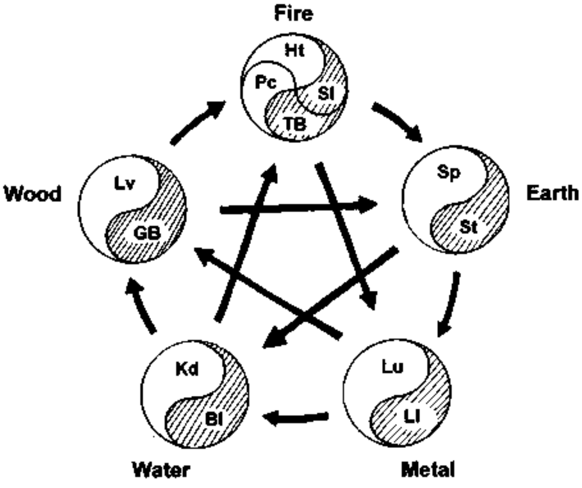

Figure 23.1 The Five Elements

THE FIVE ELEMENTS AND YIN AND YANG THEORY

It is important to understand the Five Elemental theory's energetic origin and its interaction with Yin and Yang Qi to appreciate its energetic categorizations.

The ancient Chinese Qigong masters believed that the creation, development, and declining transitions of all things throughout the universe were the result of the interaction of Yin and Yang Qi. The Yin and Yang theory was therefore used to describe the opposing, interdependent, as well as waxing and waning transitions of all things, especially the interactions of the Five Elements (Wood, Fire, Earth, Metal, and Water). Fire and Wood are considered active Yang elements, while Water and Metal are considered quiescent Yin elements. The Earth is considered a balance point between Yang and Yin.

Qigong doctors assign the Five Elements of Traditional Chinese Medicine to the body's organs and tissues. The Five Elements are considered the Jing of the body's Yin and Yang energy pools. Imagine each element organ as a pool of liquid energy. The energetic movement, much like water, flows from element to element. This flowing movement of liquid energy is considered to be the body's energetic pulse, moving Qi to and from these pools.

The rivers of Qi cannot exist without the pools of Jing to nourish their flow and function; and the pools of Jing cannot act without the rivers of Qi stimulating their action. Therefore the principle action of the Five Elements is the circulation of the Yin and Yang Qi.

The human body is like a living Taiji symbol (see Chapter 21). Its back channels are considered Yang; its front channels are considered Yin; and its center channel is contained through the energetic actions of the Taiji Pole. The Taiji Pole produces Yang Qi by the inhalation of life-force energy through the breath. When Yang Qi reaches its peak, it rests, while the Yin Qi begins to increase. When the Yin Qi reaches its peak and comes to rest, then the Yang Qi begins to increase once more. This continuous interaction of Yang energy with Yin energy produces and enhances within the body, the Five Elemental organ energies of the Liver (Wood), Heart (Fire), Spleen (Earth), Lungs (Metal), and Kidneys (Water).

The first stage of Qigong exercises and meditations create an energetic transformation which regulate, refine, and reduce these Five Elemental organ energies into two primary Yin and Yang energies known as Fire and Water Qi. The second stage of transformation consists of combining the Water and Fire energy into one union of Three Human Treasures (Jing, Qi, and Shen). This energetic transformation is responsible for "bringing 10,000 things into being" (Figure 23.2). In China, these meditation stages are used for exploring the stimulation of the energetic functions of the central nervous system, the peripheral nervous system, cardiovascular system, digestive system, reproductive system, etc.

The Five Elemental theory assists the Qigong doctor in examining the energetic correspondences for clinical evaluation. This is only one phase of clinical observation however, and should always be applied with the Yin and Yang theory.

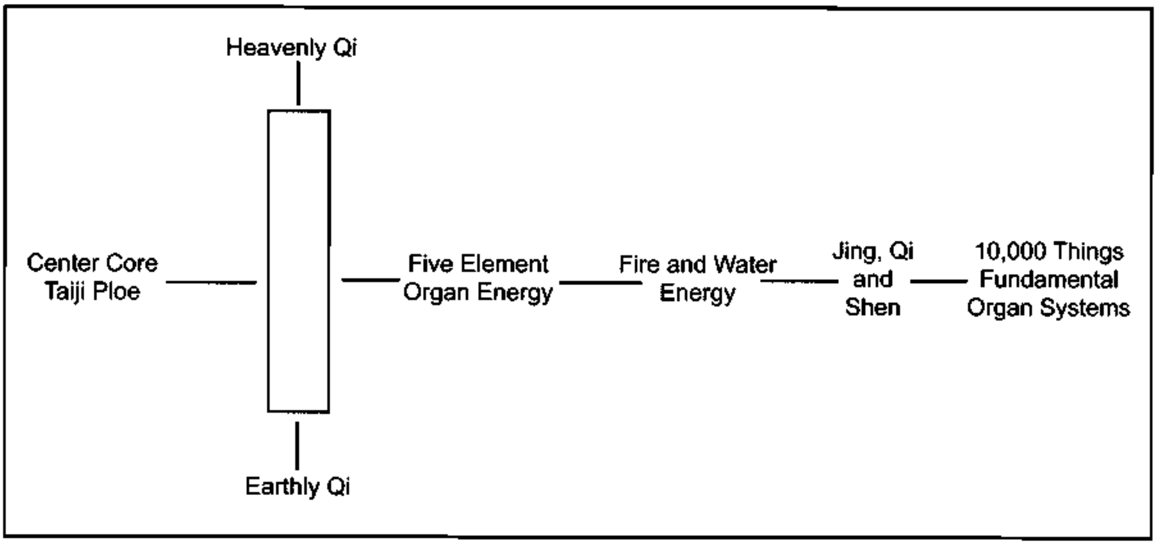

Figure 23.2 Stages of Energetic Transformation

THE CREATIVE OR GENERATIVE CYCLE

Within every Qi cycle each element has two main aspects: it is either (1) creating or being created, or (2) restricting and/or being restricted. In the process of growth and development, both creative and restrictive patterns are needed. Without creation, growth would have no origin, and growth without restriction can create harm. The creative element is called the Mother, and the element created is called the Child. This is the "Mother and Child" relationship referred to in Traditional Chinese Medicine. The Creative Cycle is generally used for tonification, as the stimulation of one organ enhances the function of the next (Figure 23.3). The Five Elemental Creative Cycle and its organ correspondences are as follows.

1. The Liver's element is Wood, which is used to create the Heart's element Fire. The Liver is the Mother of the Heart, which is the Child. The Liver stores the Blood and Blood houses the Shen. If the Mother (Liver) becomes Stagnant, the Child's Heart (Shen) will suffer. Clinical manifestations include indecision, timidity, lack of courage, palpitations, and early morning insomnia.
2. The Heart's element is Fire, which is used to create the Spleen's element Earth. The Heart is the Mother of the Spleen. The Heart's Qi pushes the Blood, thereby helping the Spleen in its function of transporting. If the Mother (Heart) becomes stagnant, the Child (Spleen) will suffer, creating a Spleen Yang Deficiency. This is due to failure of Heart Fire in providing Heat to the Spleen. Clinical manifestations include chills, weakness in the limbs, and loose stools.
3. The Spleen's element is Earth, which is used to create the Lungs' element Metal. The Spleen is the Mother of the Lungs. The Spleen Qi provides Gu Qi (food energy from the Earth) to the Lungs where it combines with air (energy from Heaven) to form Gathering (Zong) Qi. If the Mother (Spleen) becomes Deficient, the Child (Lungs) will suffer, creating Phlegm which obstructs the Lungs. Clinical manifestations include Phlegm in the chest, coughing, and general weakness.
4. The Lungs' element is Metal, which is used to create the Kidneys' element Water. The Lungs are the Mother of the Kidneys. The Lungs send Qi down to meet the Kidney Qi, as well as Fluids to nourish the Kidneys. If the Mother (Lungs) becomes stagnant, the Child (Kidneys) will suffer, due to the Kidneys not receiving enough Qi. Clinical manifestations include breathlessness, loss of voice, coughing, and asthma.
5. The Kidneys' element is Water, which is used to create the Liver's element Wood. The Kidneys are the Mother of the Liver. The Kidneys' Yin nourishes the Liver Blood. If the Mother (Kidneys) become Deficient, the Child (Liver) will suffer (i.e., Liver Yin Deficiency). Clinical manifestations include headaches, blurred vision, dizziness, and vertigo.

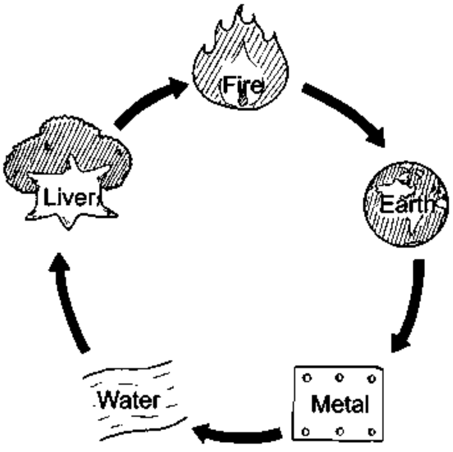

Figure 23.3 The Five Elemental Creative Cycle

THE CONTROLLING OR RESTRICTIVE CYCLE

This is a condition in Traditional Chinese Medicine where one element brings under control, or restrains, a corresponding element, thus ensuring that a dynamic balance is maintained among the Five Elements. The element that restricts is considered successful, while the element being controlled is in check. It is commonly referred to as "the strong controlling the weak," and is generally used for sedation, to ensure that the Creative Cycle does not over-create and cause imbalances. The element prior to the Mother is commonly referred to as the Grandmother. In the Controlling Cycle, it is the Grandmother who controls the Child (Figure 23.4). The Five Elemental Controlling Cycle and its organ correspondences are explained as follows.

1. The Wood absorbs nutrients from the Earth, i.e., the Liver (Grandmother) controls the Spleen and its paired organ the Stomach (Child). Two of the Liver's energetic functions are to help the Stomach "rot and ripen" the food and the Spleen to transform and transport the Gu Qi.
2. The Earth restricts Water, i.e., the Spleen (Grandmother) controls the Kidneys (Child). Both the Spleen and Kidneys control the transformation of Body Fluids.
3. The Water extinguishes Fire, i.e., the Kidneys (Grandmother) control the Heart (Child). The energetic aspect of the Kidneys' Water controls the energetic balance of the Heart's Fire.
4. The Fire melts Metal, i.e., the Heart (Grandmother) controls the Lungs (Child). The Heart governs the Blood, while the Lungs govern Qi (both Qi and Blood mutually assist and nourish each other).
5. The Metal chops Wood, i.e., the Lungs (Grandmother) control the Liver (Child). If the Lungs' Qi is weak it will not descend thus allowing the Liver Qi to rise.

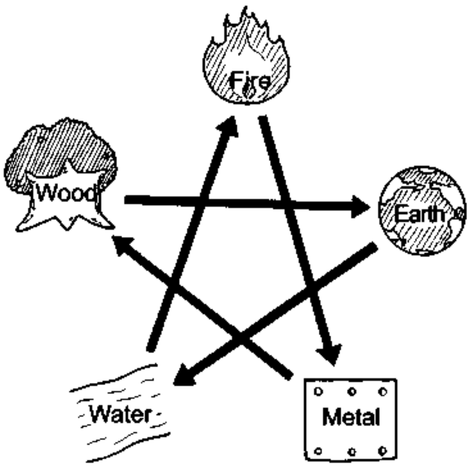

Figure 23.4. The Five Elemental Controlling Cycle

THE INVADING OR OVERACTING CYCLE

This is a condition in Traditional Chinese Medicine where an element overacts, overcontrols and invades, the element it is supposed to control (Figure 23.5). Instead of the role of checking and restraining, one element severely attacks another element, stealing Qi from its energetic pool. This is generally observed in cases of Excess or Deficiency conditions of the organs. The Five Elemental Overacting Cycle and its organ correspondences are as follows.

1. The Wood invades the Earth, i.e., the Liver (Grandmother) overcontrols the Spleen (Child) and its paired organ the Stomach. In the clinic, this is commonly called "Liver invading the Spleen." Clinical manifestations include hypochondriac and epigastric pain and distension, irritability, poor appetite, loose stool, and a green face color.
2. The Earth invades the Water, i.e., the Spleen (Grandmother) overcontrols the Kidneys (Child). This occurs when a Deficient Spleen fails to transform and transport Fluids. These Fluids accumulate and obstruct the Kidneys' function of transforming and excreting Fluids. Clinical manifestations include edema, difficulty urinating, and a yellow face color.
3. The Water invades the Fire, i.e., the Kidneys (Grandmother) overcontrol the Heart (Child). This pattern is almost never seen in clinical practice since the Kidneys are rarely in Excess.
4. The Fire invades the Metal, i.e., the Heart (Grandmother) overcontrols the Lungs (Child). This results in Full Heat in the Lungs.

Pulmonary disease can be caused by cardiac disorders, this is known as "Fire Overacting and Invading Metal." Clinical manifestations include cough with yellow sputum, a feeling of Hot sensations, and a red face color.

5. The Metal invades the Wood, i.e., the Lungs (Grandmother) overcontrols the Liver (Child). Clinical manifestations include irritability, a feeling of distension, fatigue, and a white face color.

THE INSULTING OR COUNTERACTING CYCLE

This is a condition in Traditional Chinese Medicine where the elements reverse their restrictive cycle and the Child insults the controlling element, the Grandmother (Figure 23.6). This is generally observed in cases of Excess or Deficiency within the Five Elements. The Five Elemental Insulting Cycle and its organ correspondences are as follows.

1. The Wood dulls the Metal ax, i.e., the Liver (Child) insults the Lungs (Grandmother). If the Child (Liver Qi or Liver Fire) becomes stagnant, it can obstruct and prevent the Lung's Qi (Grandmother) from descending. Clinical manifestations include coughing, a feeling of distension in the chest and hypochondrium area, and asthma. If pulmonary disease is due to the impairment of the Liver, it is explained as "Wood Insulting or Counteracting Metal."
2. The Metal controls the spread of Fire, i.e., the Lungs (Child) insult the Heart (Grandmother). If the Child becomes Deficient, it can cause the Heart Qi (Grandmother) to become Deficient. Clinical manifestations include palpitations, breathlessness and insomnia.
3. The Fire dries up the Water, i.e., the Heart (Child) insults the Kidneys (Grandmother). If the Grandmother (the Kidneys' Yin) becomes Deficient, then the Heart (Child) can become excessive, giving rise to Heart Empty Heat. Clinical manifestations include dizziness, lower back pain, night sweating, insomnia, malar flushes, and dry mouth (at night).
4. The Water erodes the Earth, i.e., the Kidneys

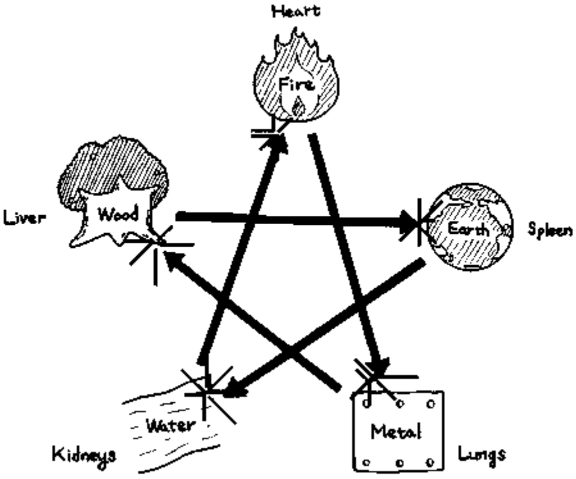

Figure 23.5 The Five Elemental Invading Cycle

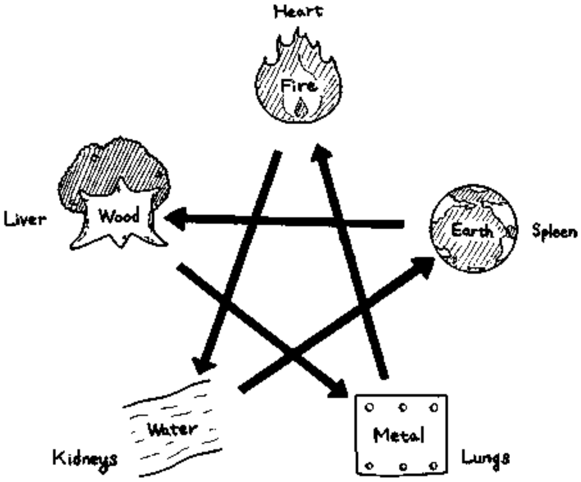

Figure 23.6 The Five Elemental Insulting Cycle

(Child) insult the Spleen (Grandmother). In this case, the Child (the Kidneys' Yang) becomes Deficient causing the Grandmother (Spleen) to become Deficient. Clinical manifestations include weak limbs, fatigue, edema, and loose stools.

5. The Earth resists the Wooden plow, i.e., the Spleen (Child) insults the Liver (Grandmother). If the Spleen (Child) fails to transform Fluids, a Damp condition can result. This Dampness can accumulate and begin to obstruct the flow of the Liver's Qi (Grandmother). Clinical manifestations include hypochondriac pain and distension, bile obstruction, and jaundice.

THE FIVE ELEMENTAL TONIFICATION CYCLE

The Five Elemental tonification cycle can be used for balancing the life-force energy of the body in conjunction with nature. Wood corresponds to spring, Fire corresponds to summer, Metal corresponds to autumn, Water corresponds to winter, and Earth corresponds to the late stage of each season (late spring, late summer, late autumn and late winter). The element Earth is centrally placed, associated with the 18 days of transformation at the end of each of the four seasons. The object of using the Five Elemental Tonification Cycle is to support and strengthen the next organ in the cycle (Figure 23.7).

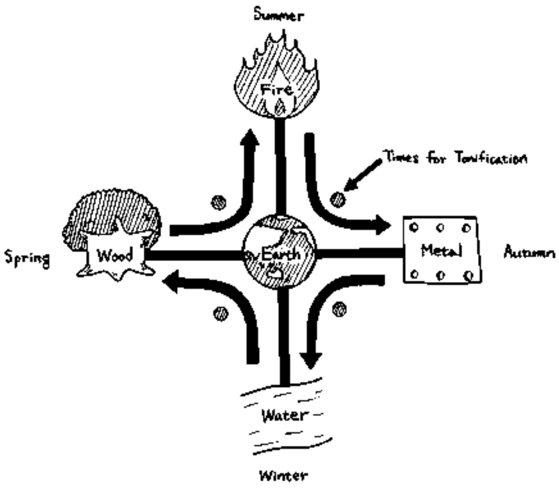

Figure 23.7 The Five Elemental Tonification Cycle

FIVE YIN ENERGETIC ORGAN MANIFESTATIONS

The Five Elemental energies are manifested through five different directions of movement. The Qigong doctor studies the movements and actions of a patient while they are in a meditative state (usually a sitting or standing posture), to diagnose their energetic disposition. After being in a meditative state for awhile, the patient's organs will overflow with Qi. When the organs overflow with Qi, certain physical reactions are evident, indicating which particular element has a problem.

1. Wood or Liver Qi is responsible for the dispersing effect of Qi, and is manifested by expansive outward movements in all directions, causing the body to sway with gentle and soft motions. When the Liver's energy is overabundant the torso makes a gentle, soft movement like a pine tree swaying in the wind. This is caused by the sprouting and branching manifestation of the Liver Qi's rising warm current.
2. When Fire or Heart Qi is overabundant, it is evident through the physical manifestation of the upward ascending actions of the torso caused by the flaring up of Yang Heat. This affects the energetic body and is manifested by upward movements and sometimes jerky actions (like a flame crackling in a fire).
3. Earth or Spleen Qi is generally manifested by stable movements, however, when overabundant, it is evident through movements that are both smooth and swift. This allows the body to progress in a limber, flexible, and free moving action (think of the body movement similar to a spinning ball).
4. When Metal or Lung Qi is overabundant, it is evident through the twitching actions of the torso and is manifested by contracted, inward movements, causing the body to violently jerk from side to side (think of the body as a metal bell being violently struck).
5. When Water or Kidney Qi is overabundant, it

is evident through the clumsy, awkward, heavy downward moving action of the torso (think of a waterfall descending onto rocks).

When the patient's body, unintentionally or unconsciously, rocks or sways gently as a result of a meditative state, the patient does not feel tired upon completing the meditation. However, when a patient consciously tries to force relaxation or augment the movement of his or her body while meditating, he or she will be tired by the end of the meditative session.

Note: While interning at the Medical Qigong Ward at the Xi Yuan Hospital in China, I found it fascinating to observe meditating patients displaying the results of both energetic organ manifestations or consciously induced augmentation of movement. All the above listed energetic manifestations were evident.

CHAPTER 24
OTHER DIAGNOSTIC SYSTEMS AND TECHNIQUES

DIAGNOSIS ACCORDING TO THE SIX STAGES

This system of diagnosis was formulated by Dr. Zhang Zhong Jing in the early Han Dynasty. The application of the six stages of diagnosis provides the key link in treating febrile diseases which are due to an attack of Cold. The six stages are described as Tai Yang, Shao Yang, Yang Ming, Tai Yin, Shao Yin, and Jue Yin.

THE THREE YANG SYNDROMES

The signs and symptoms of these pathological changes result from attacks by exogenous factors, especially cold weather. The first three syndromes describe the pathological changes occurring in the six Yang organs and channels (Small Intestine, Bladder, Large Intestine, Stomach, Triple Burners and Gall Bladder), and are called the Three Yang Stages. In the Three Yang Stages, the pathogenic factors predominate, but the body's energy is still relatively strong. The major symptoms are signs of Heat, and the treatment is directed at eliminating the pathogenic factors. These are the external conditions which can progress from the superficial Tai Yang Stages to the deeper Shao Yang Stages.

1. The Tai Yang or Great Yang Stage affects Qi on the exterior surface of the body and all three Wei Qi levels. This energetic disturbance affects the Small Intestine and Bladder Channels and organs that govern the exterior of the body. Although they are the most superficial channels, they are very rich in Wei Qi.
   - Clinical Manifestations of an Attack of Cold include: Aversion to cold, stiff neck, headache, no sweating, fever, body aches, floating-tight pulse, and shortness of breath.

   The focus of the Qigong treatment is to first open the pores to eliminate pathogenic factors by purging and dispersing, followed by tonification of the Lungs.
   - Clinical Manifestations of an Attack of Wind include: Aversion to wind, stiff neck, sweating, body aches, floating pulse, and fever.

   The focus of the Qigong treatment is to first open the pores to eliminate pathogenic factors by purging and dispersing, then to relax the muscles for diaphoresis (profuse sweating) in order to regulate the Ying and Wei Qi, and finally to prescribe herbs (or refer out for herbal treatment).
2. The Shao Yang or Lesser Yang Stage affects the interior and exterior levels of the body equally. This energetic disturbance affects the Triple Burners and Gall Bladder organs and channels. The patient experiences an alteration of fever and chills with a fullness of the costal and hypochondriac regions, blurred vision, loss of appetite, irritability, and wiry pulse with a white-slippery tongue coating.

   The focus of the Qigong treatment is on dispersing Heat with purging techniques and promoting the production of Body Fluids using herbs.
3. The Yang Ming or Bright Yang Stage affects Qi at the interior level of the body. This energetic disturbance affects the Large Intestine and Stomach organs and channels. The pathogenic factors transform into Heat.
   - Clinical Manifestations of the Bright Yang Channel Stage include: Pure Interior Heat known as the Four Bigs (fever, thirst, sweating, and pulse). In addition the patient also experiences an aversion to heat and has a red tongue with a yellow coating.

   The focus of the Qigong treatment is on dispersing Heat with purging techniques and on promoting the production of Body Fluids through herbs.
   - Clinical Manifestations of the Bright Yang Or-

gan Stage include: Interior Heat and the Four Bigs with constipation, profuse sweating, high fever which worsens in the afternoon, fullness and pain in the abdomen (which becomes worse if you press it), and a red tongue with a dry, yellow or black coating.

The focus of the Qigong treatment is on eliminating pathogenic factors with purging and dispersing techniques, and relieving constipation.

THE THREE YIN SYNDROMES

The second three syndromes are those of the Three Yin Channels. They are based on the pathological changes of the Six Yin Organs and channels (Lungs and Spleen, Heart and Kidneys, Pericardium and Liver). These syndromes are the reflection of pathological changes within the Yin and Yang organs, channels, and collaterals. In the three Yin stages, the pathogenic factors are still present but diminishing. The body's Qi has become weakened, and there are Deficient signs of Cold. The treatment is focused on strengthening the body's Qi. These are internal conditions that progress from the Tai Yin stages deeper into the body until they reach the Jue Yin stage.

1. The Tai Yin or Great Yin Stage affects the Yin organs, as the penetration of pathogenic factors flow deep into the energetic layers. This energetic disturbance affects the Lung and Spleen organs and channels, resulting in a Deficiency and Cold syndrome. The clinical manifestations are: Vomiting and diarrhea, abdominal fullness, absence of thirst, and a slow, deep pulse with a pale tongue.

   The focus of the Qigong treatment is on heating and energizing the Middle Burner and dispersing the Cold using herbs. It is also important to tonify the patient's organ Qi.
2. The Shao Yin or Lesser Yin Stage affects the Heart and Kidney organs and channels. Clinical manifestations are as follows:
   - The transformation of Cold Deficiency to Kidney Yang Deficiency: Aversion to cold, chills, cold limbs, diarrhea, listlessness, lethargy, abundant-pale urine, deep-fine pulse, and pale tongue.

     The focus of the Qigong treatment is on tonifying the patient's organ Qi, preventing the Yang from collapsing and treating with herbs.
   - The transformation of Heat Deficiency to Kidney Yin Deficiency: insomnia, fever, irritability, dry throat and mouth, scanty-dark urine, fine-rapid pulse, and a red tongue without coating.

     The Qigong treatment should focus on tonifying the patient's organ Qi, nourishing the Yin, and dispersing Fire using herbs.
3. The Jue Yin or Terminal Yin Stage affects the Pericardium and Liver organs and channels. This is the final stage of pathogenic invasion before the life-force energy vanishes altogether.

   Clinical manifestations are as follows: cold limbs, diarrhea, vomiting, pain, feeling of heat and energy expanding in the chest, and feeling of hunger with no desire to eat.

   The focus of the Qigong treatment is on purging, tonifying and regulating the patient's organ Qi. Applying Cold and Hot herbal medicines simultaneously can be used to drive out the patient's pathogenic factors and aid in purgation. Herbal tonics can also be used to increase the patient's resistance to pathogenic diseases and aid in tonification (Figure 24.1).

DIAGNOSIS ACCORDING TO THE FOUR LEVELS

This system of diagnosis was formulated by the famous Chinese physician, Dr. Ye Tian Shi in the late 1600's. These syndromes are used to explain the etiology and pathogenesis of febrile diseases. The differentiation of syndromes pertaining to the Four Levels of Diagnosis is concerned with febrile diseases caused by externally contracted Heat. Because exogenous diseases are always developing and changing, by studying the transmission of pathological changes the Qigong doctor is able to deduce the prognosis. These four levels are identified by four types of energy, and are classified as follows: The Wei Qi Level, Qi Level, Ying Qi Level, and Blood (Xue) Level.

1. At the Wei Qi Level (Protective Energy), the

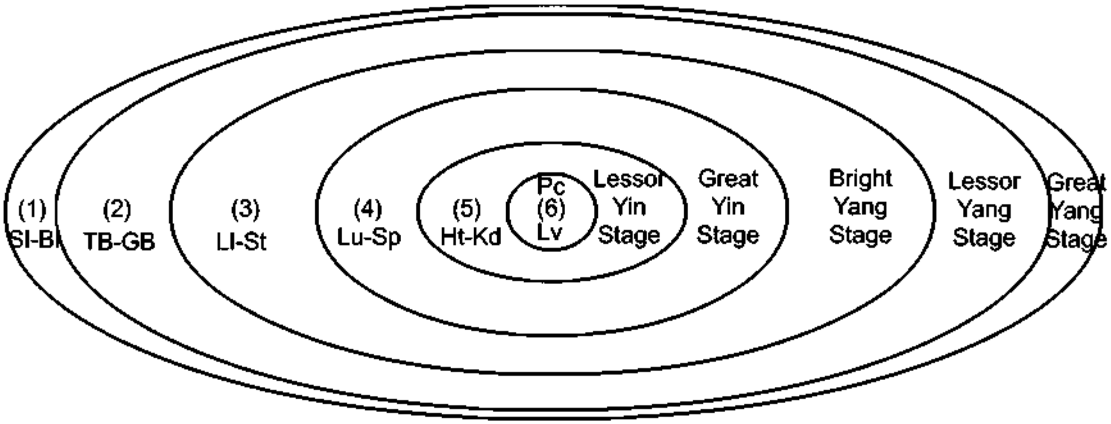

Figure 24.1. The Six Stages of Pathogenic Invasion are illustrated here. The invasion of external pathogenic factors follows a systematic progression through the body's Six Levels of Qi.

1. The Great Yang Stage (Tai Yang)

Affects the Small Intestine and Bladder Organs and channels. At this stage the pathogens are at the superficial level of the body's Wei Qi, caused by an attack of Wind and/or Cold Invasion. This stage is considered an External Cold Invasion at the Wei Qi level.

2. The Lessor Yang Stage (Shao Yang)

Affects the Triple Burners and Gall Bladder Organs and channels. At this stage the pathogens are half Exterior and half Interior, flowing in and out of the body, causing alternating chills and fever. This stage is also considered an Internal condition at the Qi level.

3. The Bright Yang Stage (Yang Ming)

Affects the Large Intestine and Stomach Organs and channels. At this stage the pathogens absorb deeper into the body's tissues, transforming into Heat affecting the internal organs and/or channels. This stage is considered an Internal condition at the Qi level.

- Bright Yang Organ Stage: Results in Interior Heat with constipation.
- Bright Yang Channel Stage: Results in pure Internal Heat with "Four Bigs" (fever, thirst, sweating, and pulse).

4. The Great Yin Stage (Tai Yin)

Affects the Lungs and Spleen Organs and channels. At this stage the penetration of pathogenic factors flow deep into the energetic layers affecting the Yin organs, invading the Middle Burner (affecting the Spleen) causing a Deficient condition. This stage is considered an Internal condition at the Ying Qi level.

5. The Lesser Yin Stage (Shao Yin)

Affects the Heart and Kidney Organs and channels. At this stage the pathogenic factors transforms either into Cold or Heat, causing further Deficient conditions. This stage is also considered an Internal condition at the Ying Qi level.

- Pathogens will transform into Cold due to a Deficiency of Kidney Yang.
- Pathogens will transform into Heat due to a Deficiency of Kidney Yin.

6. The Terminal Yin Stage (Jue Yin)

Affects the Pericardium and Liver Organs and channels. At this stage the pathogenic factors continue to deplete the body's life-force energy. This is the final stage of pathogenic invasion before death occurs. This stage is also considered an Internal condition at the Blood level.

pathogenic Heat invades the outer level of Wei Qi, causing the human body to manifest certain disorders which are clinically observed as fever, aversion to wind and cold, excessive or absence of sweating, sore throat, coughing, headache, red tongue with thin coating, and a floating and rapid pulse.

The Qigong treatment should focus on dispersing pathogenic factors and tonifying the exterior by utilizing pungent tastes and cooling herbs. Points are generally selected from the Lungs, Large Intestine, and Bladder Channels, as well as from the Governing Vessel.

2. At the Qi Level, the pathogenic Heat continues to progress deeper into the body, attacking and affecting the Qi system. At this stage, the body system is still strong, but the normal functions of the Yin and Yang organs are impaired. Because of the different functions of the Yin and Yang organs the manifestations differ. The three conditions associated with this stage of pathogenic invasion are as follows:

- Heat in the chest and diaphragm is the first stage and is not considered a severe syndrome. The symptoms manifest as fever, burning sensation in the epigastrium, thirst, and a yellow coated tongue.

  The Qigong treatment should focus on dispersing pathogenic Heat accumulated in the chest and diaphragm.
- Heat in the Stomach Channel is the second stage, and is considered a more severe Heat syndrome. These symptoms manifest as Heat, thirst, a dry tongue with yellow coating, and a slippery, rapid pulse.

  The Qigong treatment should focus on purging pathogenic Heat, and tonifying the body's Fluids.
- Heat in the Lesser Yang Channel is considered the final stage of Qi invasion. This syndrome is similar to the Lesser Yang Stage of the Three Yang Syndromes (according to the Six Stages), except that there is Dampness.

  The Qigong treatment should focus on purging pathogenic Heat from the Lesser Yang Channel, dispelling Dampness and dispersing Phlegm.

3. At the Ying Qi Level (Nutritive Energy), the pathogenic factors have penetrated deeper to the level of the Ying Qi and the organs, the Blood, blood vessels, and the Heart are therefore all affected. The symptoms in this stage are generally caused by the burning up the Yin. Clinical manifestations include: irritability, insomnia, mental restlessness, fever that worsens at night, dry mouth, absence of thirst, faint skin eruptions, a deep red tongue with little or no coating, and a fine-rapid pulse.

   Focus the Qigong treatment on purging pathogenic Heat from the Ying Qi level, using dispersing techniques. As an auxiliary method, bleeding may be performed by pricking the Heart and Pericardium Channels, as well as the Governing Vessel.
4. The Blood (Xue) Level, is the deepest and final level of pathogenic invasion. This stage contains the most serious manifestations that the disease can cause. Clinical manifestations include: high fever, skin that is hot to touch, Bloody stool, vomiting of Blood, Blood in urine, delirium, skin eruptions of purple or black color, a deep red tongue, a fine-rapid pulse, and in severe cases convulsions.

   The Qigong treatment should focus on purging pathogenic Heat from the Blood using bloodletting techniques (points are mainly selected from the Heart, Pericardium, Liver, Stomach, and Large Intestine Channels, along with the Governing Vessel), Cool the Blood, relieve convulsions, calm the mind, and reduce pathogenic factors with herbs.

DIAGNOSIS ACCORDING TO THE TRIPLE BURNERS

This system of diagnosis was developed by the famous Chinese physician, Dr. Wu Ju-Tong in the late 1700's. This system is concerned with febrile diseases caused by externally-contracted Heat, invading the body's Triple Burners. This Heat syndrome generally proceeds from the Upper Burner to the Middle and ends up at the Lower Burner. However, in cases of diseases due to Damp Heat, the pathogenic invasion of Heat begins at the Middle Burner in the Spleen.

1. The Upper Burner encompasses the body from the top of the head to the diaphragm. This area includes the Heart, Lungs, Pericardium, throat and head, and is responsible for respiratory and cardiac functions. The Upper Burner moves the body's Clean Qi, circulating and distributing nutrients and Qi throughout the body like a mist.

   Clinical manifestations of diseases in the Upper Burner correspond to those of a pathogenic invasion of the Lungs and Pericardium.
   a. When Wind Heat invades the Lungs, it is either manifested at the Wei Qi level or deeper in the Lungs themselves. The symptoms include: fever, sweating, coughing, stiffness and pain in the chest, wheezing, thirst, a red tongue with a yellow coat, and a rapid pulse.
   b. When pathogenic Heat invades the Pericardium, it is manifested at the Wei Qi level. These symptoms include: fever, coldness of the limbs, delirium, aphasia, burning sensation in the epigastrium, and a deep red tongue with spots.

   The Qigong treatment should focus on purging pathogenic Heat and dispersing Phlegm.
2. The Middle Burner encompasses the area from the diaphragm to the umbilicus. This area includes the Stomach, Spleen, pancreas, and Gall Bladder and is responsible for digestion, fermentation, and transformation of food and drink into nutrients for distribution. It moves the body's energy, circulating it like a swamp.

   Manifestations of diseases in the Middle Burner from pathogenic Heat are divided into two syndromes: Heat in the Bright Yang and Damp Heat in the Spleen.
   a. Heat in the Bright Yang include: symptoms of interior Heat and the Four Bigs (severe excess fever, sweating, thirst, and pulse) with constipation, profuse sweating, high fever which worsens in the afternoon, fullness and pain in the abdomen (which becomes worse when pressure is applied), and a red tongue with a dry, yellow or black coating.
   b. Damp Heat in the Spleen include: symptoms that are identical to the early stages of diseases caused by externally-contracted Heat.

   The clinical manifestations include: aversion to cold, fever that worsens in the afternoon, a heavy sensation in the head, body and limbs, fullness of the chest and epigastrium, nausea and vomiting, a white, sticky tongue, and floating or slippery pulse.

   The Qigong treatment should focus on Heating the Middle Burner, tonifying any Deficiency, purging Dampness, and prescribing herbs.
3. The Lower Burner encompasses the area from the umbilicus to the feet. This area includes the Liver, Kidneys, Bladder, intestines, external genitalia for men, and the uterus for women. The Lower Burner is responsible for filtering and eliminating waste products and for reproductive functions. It moves the body's Turbid Qi, circulating it like a drainage ditch.

   Manifestations of diseases in the Lower Burner are caused from pathogenic Heat attacking the Kidneys or Liver and are divided into two syndromes:
   a. Invasion of the Kidneys by pathogenic Heat dries up the Kidneys' Yin causing Heat symptoms due to a Yin Deficiency. Clinical manifestations include: lingering fever, hot hands and feet, dryness of the mouth, deafness, lassitude, a red-peeled tongue, and a rapid, floating or empty pulse.

   Focus the Qigong treatment on tonifying Yin, and purging the Heat.
   b. Invasion of the Liver by Heat occurs after the Kidney Yin has been exhausted. A condition known as Liver Wind is created. Clinical manifestations include: low-grade fever, cold limbs, dry and cracked lips, convulsions and trembling, a dry, reddish-purple tongue, and a rapid-deep pulse.

The Qigong treatment should focus on tonifying Yin, and purging the Heat and Wind.

CHAPTER 25
CLINICAL EXAMINATION AND DIAGNOSIS

SENSORY, INTUITIVE AND PERCEPTUAL DIAGNOSIS

Clinical diagnosis can be categorized into two primary methods of energetic examination and treatment: Sensory Diagnosis and Intuitive and Perceptual Diagnosis. The clinical methods of Sensory Diagnosis are quite prevalent throughout China, especially within the acupuncture, massage, and herbal clinics. The methods of Intuitive or Perceptual Diagnosis, however, have been kept secret for many years, and are primarily found only within the Medical Qigong clinics. The latter methods (Intuitive or Perceptual Diagnosis) will be explained later in this chapter.

The root of any disease can be drawn from its complex symptoms and signs, by using the two main methods of Sensory and Intuitive and Perceptual Diagnostic principles. These two main methods of diagnosis serve as guidelines for the Medical Qigong doctor in clinical evaluation. Once the cause and extent of the disease is known, a treatment plan can be developed. There are four main types of Qigong treatment, which are chosen according to the treatment goals:

- a routine treatment (basic or nonspecialized),
- a specialized treatment (designed for the specific patient and his or her condition),
- a treatment aimed at etiology, or
- a treatment focused on alleviating the symptoms and signs.

It is therefore important to determine and assess the cause (root) and the manifestations (branches) of the patient's disease.

SENSORY DIAGNOSIS

Before treatment begins, the Qigong doctor performs a clinical intake, examining and evaluating the patient's condition, using various methods of sensory diagnosis. This clinical format involves taking a history of the patient's complaints. Next the doctor inquires about the patient's medical history. The doctor then conducts an examination for diagnosis, and devises a treatment principle on which the treatment plan is based, before implementing the treatment.

CLINICAL INTAKE, EXAMINATION AND EVALUATION

It is important that the doctor clearly identify the patient's problem before any treatment may be given. Never treat a patient without first making a diagnosis. The diagnosis is developed from the doctor's sensory examination and is evaluated according to symptom observation.

Both the root (the source and cause of the disease) and its branches (the symptoms or manifestations of the disease) should be examined carefully before treating the patient. In the case of an emergency, the doctor may choose to treat the urgent symptom first before treating its source.

The diagnosis and treatment of each patient traditionally begins by observing the patient as a whole, followed by an evaluation of the patient's individual constitution. Next, the doctor observes the symptoms and syndromes of the individual's organs. From these observations the cause of the disease is determined and treatment is begun. The treatment usually focuses on addressing the cause of the disease, to return the organ or organ system back into harmony with the rest of the patient's body. This affects changes within the patient's constitution and returns him or her back to a normal state of health.

Sensory examination for diagnosis is divided into five primary methods: diagnosis by visual observation, diagnosis by listening, diagnosis by smelling, diagnosis by palpation, and diagnosis by questioning the patient. By looking, listening, smelling, touching, and asking the patient specific questions, a Qigong doctor is able to diagnose the patient's syndrome (this includes the etiology of the disease).

DIAGNOSIS BY VISUAL OBSERVATION

1. Observe the patient's Shen, overall appearance, spirit, emotional nature, energy field, and composure. Look at the Shen of the patient's face; a healthy patient should have a brightness or vibrant shininess to his or her countenance. The Shen of the face should be checked against the state of the patient's eyes.
2. Observe the patient's eyes, especially the spirit in their eyes -- the glitter or dullness, and the expression of a controlled or uncontrolled emotional state. The eyes manifest the Essence of all the organs, and are the messengers of the Heart. There are two things to look at in the eyes -- glitter and focus.
    a. Are the patient's eyes bright and do they glitter? The less the glitter, the more long-standing the emotional and mental problems. It is said that there are three emotions that make the eyes lack glitter: sadness, grief, and shock. If the patient's eyes are dull, it looks as if he or she is seeing through a mist or film.
    b. Are the eyes focused, is the attention sustained and penetrating or uncontrolled? Uncontrolled attention is actually worse then having no glitter. If the patient has unfocused eyes it means that there is an obstruction of the patient's Shen and mind, generally due to guilt, preoccupation, or anxiety. It is said that Excess joy makes the "eyes uncontrolled."
3. Observe the color of the patient's face, complexion, eyes, ears, nose, mouth, tongue, skin zones, and ask about the discharges (mucus, vaginal, urine, and stool). Anger manifests as a greenish tinge on the cheeks or forehead (if it is affecting the Stomach); worry manifests as a grayish color, pensiveness as a sallow complexion, fear as a white complexion (unless combined with Kidney Yin which turns the cheeks red). Shock creates a bluish tinge on the forehead. A change in the complexion often indicates a deeper and longer-standing psycho-emotional problem. Also, if the eyes show no Shen but the face does, it indicates a recent problem.
4. Observe the patient's body and any outstanding structural features. Are they thin, emaciated, obese, or paralyzed? Are there deviations or obstructions in or on the body, face, head, mouth, teeth-gums, eyes, nose, and ears? Are there blemishes on the skin? Observe the appearance of the nails, hair and so on.
5. Observe the way the patient moves. Do they shake and tremor? Are they rigid or loose? Notice if the movements are fast or slow, and so on. When you shake hands upon meeting for the first time, are his or her hands clammy or dry; is his or her grip strong or weak?
6. Observe the patient's tongue. Is it cracked, serrated, fat, thin, or thick? If you observe a central crack, the deeper it is, the more long-standing the emotional problem. A red tipped tongue can also indicate a Shen disturbance.
7. Observe the patient's channels and points. Are there darkened red blotches, blemishes, swelling, sweat, etc.?

DIAGNOSIS BY LISTENING

1. Listen to the patient's speech for slurring or clarity.
2. Listen to the patient's tone. Is it high, low, loud, or quiet?
3. Listen to the patient's tempo. Is it fast or slow?
4. Listen to the patient's words. Do they make sense? Are his or her thoughts well ordered, or does the patient change subjects abruptly and frequently (if so, the patient may be suffering from a severe mental disorder).
5. Listen to the patient's respiration. Is it loud, strong, weak, or wheezing?
6. Listen to the patient's cough. Is it loud, dry, or wet?
7. Listen to the patient's Stomach and Intestines. Are they gurgling or quiet?

DIAGNOSIS BY SMELLING

1. Does the patient smell strong, weak, or odorless? Does the patient smell of medications, or alcohol?
   - For Liver problems, the patient's body or breath will smell like a goat.
   - For Heart problems, the patient's body or breath will smell like something has been burnt.
   - For Spleen problems, the patient's body or breath will smell fragrant or sweet.
   - For Lung problems, the patient's body or breath will smell like fish or a tide pool.
   - For Kidney problems, the patient's body or breath will smell rotten.

DIAGNOSIS BY PALPATION

1. How does the patient's pulse feel? Is it strong or weak? The pulse shows more about the patient's Qi than the state of his or her Shen. Because the pulse reflects the short-term state of the patient's Qi, it should always be integrated with tongue and complexion diagnosis.
2. How do the patient's channels and points feel? Are there tender areas, movable nodules (Qi stagnation), or fixed nodules (Blood stagnation) along the patient's channels?
3. How does the patient's abdomen feel?

DIAGNOSIS BY QUESTIONING

1. Ask about the history of the patient's main complaint (illness, disorder, or injury).
   - When did it begin?
   - What other signs or symptoms accompany it?
   - Ask about the course of the illness, is it improving or getting worse?
   - What makes it better or worse (i.e., pressure, warmth, cold, and so on)?
   - What other treatments has he or she tried?
2. Ask about the patient's general medical history.
   - Has the patient had any surgeries?
   - Has the patient experienced any past or recent physical or emotional traumas?
   - Ask if the patient is on any medication(s).
   - Ask about self-medication with alcohol or drugs.
   - Ask about his or her current coffee, soft drink and sugar intake.
   - Ask about the patient's extremities, do they feel hot or cold?
   - Ask about his or her sweat.
   - Ask about his or her thirst.
   - Ask about his or her appetite and diet (how often does the patient eat, and how regular are the meals?).
   - Ask about his or her digestion (gas, belching, etc.).
   - Ask about his or her stool (loose, well formed, hot, smelly, cold, fishy, color, etc.).
   - Ask about his or her urine (frequency, urgency, and color).
   - Ask about his or her sleeping patterns (difficulty falling asleep, remaining asleep, or recurring nightmares).
   - Ask for an energetic self-evaluation on a scale of 1 to 10.
   - Ask about current emotional and mental problems or concerns.
   - Ask about his or her current physical exercise program.

EXAMINATION FOR DIAGNOSIS

There are certain general principles used by all doctors of Traditional Chinese Medicine to become proficient in diagnostic skills. In Medical Qigong therapy, becoming skilled and accurate in clinical diagnosis requires the Qigong doctor to become proficient in combining perceptual and mental skills with spiritual understandings and insights. The following is a thorough and systematic approach for arriving at an accurate diagnosis.

1. Record all information observed and gathered as meticulously and impartially as possible. When first encountering a patient, be open minded, but note all first impressions.
2. Organize all the information according to the Eight Energetic Principles or Five Elemental Theories.
   - When diagnosing according to the Eight Energetic Principles, the doctor should first determine whether the condition is Yin or Yang (chronic or acute), then classify the patient's

symptoms according to the condition's general location (External or Internal), followed by the disease's symptomatic appearance (Hot or Cold), and finally, the physio-energetic nature of the disease (Excess or Deficient condition).

Next, the Qigong doctor focuses on the Yin or Yang aspects of the patient's body, in terms of the physiological properties of the patient's disease, noting which is more affected, the Qi and Blood, Yin and Yang organs, etc.

The Qigong doctor then determines the precise depth and severity of the patient's disease, observing whether it is the patient's surface channels which are being affected, or the patient's internal organs. If it is determined that the condition is an internal problem, the Qigong doctor then differentiates the energetic level of penetration into the patient's body. This level of pathogenic penetration is diagnosed according to either the Six Stages, Four Levels, or the Triple Burners method of diagnosis.

- If the doctor is diagnosing according to the Five Elemental Theory, he or she should first determine the condition of the patient's pulse to assess which organ or organs are Excess or Deficient. The doctor must then decide whether the imbalance is attributed to either an individual internal organ or a combination of internal organs. If the patient's pulse does not match with the information gathered through the examination, the doctor must reconsider whether the patient's disorder will in fact conform to a Five Elemental pattern.

3. Evaluate all information: The doctor should evaluate the gathered information according to the patient's signs and symptoms, determining what is of primary and secondary importance. It is in this stage of evaluation that the Qigong doctor determines the "root and branch" of the patient's disease. If the disease is External in origin, the doctor determines what pathogenic factors are involved. If the disease is Internal in origin, the doctor determines whether the cause is emotional, hereditary, dietary, or an energetic charge remaining from a previous external disruption.

   Finally, the doctor should determine whether the cause of the disease is still active within the patient's life. If it still is, the doctor must take this factor into consideration before treating the patient and prescribing homework.

DEVISING A TREATMENT PRINCIPLE

In devising a treatment principle, the Qigong doctor should always include the Condition of the disease, Location of the disease, and Possible Cause of the disease.

1. The Condition of the Disease refers to the categorization of the disease, and may be diagnosed according to a combination of the Eight Energetic Principle's criteria. The patient's condition could, for example, be considered Excess or Deficient, Hot or Cold, Internal or External, or a combination thereof, such as: Excess Heat, Internal Heat, Deficient Heat, etc.
2. The Location of the Disease refers to the physio-energetic level of the patient's body that the disease is affecting (whether it is in the Upper, Middle or Lower Burner, or affecting certain Yin or Yang organs, channels, or muscles, etc.).
3. The Possible Cause of the Disease is either a result from a congenital or acquired illness. If congenital, consider the patient's DNA patterns, and all hereditary and ancestral traits. If acquired, consider environmental exposure to toxic poisons, diet, trauma (physical or emotional), electromagnetic fields, etc.

TREATMENT STRATEGY

Once the working diagnosis and treatment principle has been established, it is time to determine a treatment strategy. A treatment strategy consist in applying the treatment principle. It is the foundation from which the Qigong doctor chooses the approach in treating the patient's disease (Figure 25.1). The doctor will choose to treat the internal organs, channels, collaterals, points, and tissue areas which are the primary cause of the patient's disease. These areas are specifically chosen in order to facilitate a rapid affect on the patient's disease, to purge, tonify, or regulate the tissue area.

TREATMENT

The treatment usually focuses on the specific cause of the disease and how it affects the patient's internal organs. The primary goal is to relieve the patient's symptoms, strengthen the constitution and return the patient to a state of wholeness. This process may involve restructuring the patient's belief structure, to emotionally release old traumas.

AFTER THE TREATMENT

After treating the patient's condition, the Qigong doctor explains to the patient in clear, comprehensive terminology, what the doctor has discovered, and his or her approach in rectifying the condition. The more the patient understands the

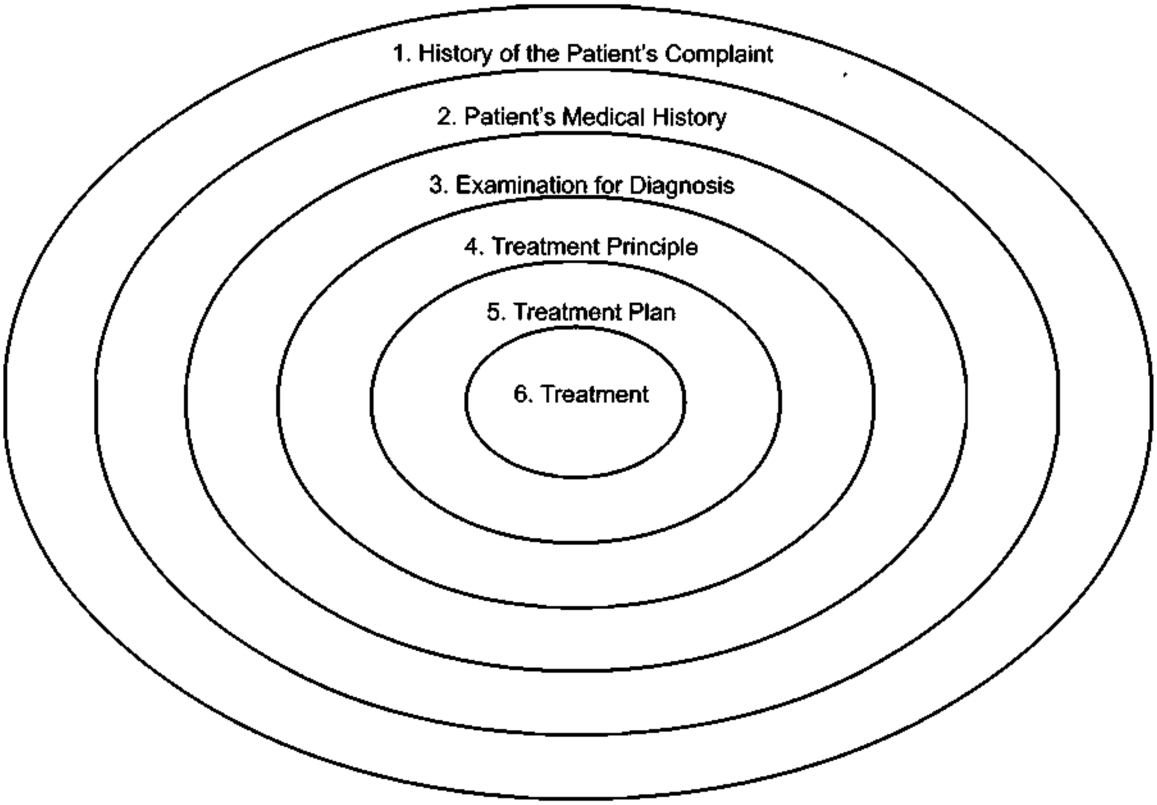
1. History of the Patient's Complaint
2. Patient's Medical History
3. Examination for Diagnosis
4. Treatment Principle
5. Treatment Plan
6. Treatment

Figure 25.1. Examination, Diagnosis, and Treatment: The Qigong doctor begins diagnosing the patient by taking a "history of the patient's complaint." Next the doctor inquires about the "patient's medical history." Then, the doctor begins the "examination for diagnosis," observing the patient's constitution, symptoms and signs and determines the syndrome(s), health and state of the patient's internal organs, and the cause of the energetic dysfunction. Next, all the acquired knowledge is assessed in accordance with the different diagnostic templates of the Eight Parameters, Five Elemental Theory, Qi, Blood and Body Fluid Dysfunctions, etc. and the doctor "devises a treatment principle." Before treating the disease the doctor "develops a treatment plan." The treatment focuses on the specific cause of the disease and how it affects the patient's internal organs. The primary goal is to relieve the patient's symptoms, strengthen the constitution and return the patient to a state of wholeness. This entire process is initiated to assist the Qigong doctor in prescribing the correct treatment.

doctor's diagnosis and approach in addressing the treatment, homework prescriptions, and suggestions for life-style modifications, the greater the likelihood of a successful healing.

After the initial treatment, it is important for the Qigong doctor to establish a treatment schedule. This establishes a realistic goal and time frame for the therapy. With each treatment, the doctor monitors the patient's progress, and may choose to modify the treatment approach in accord with the patient's changing condition.

As the patient's symptoms improve, the doctor lengthens the time between treatments. This continues until the symptoms are completely relieved, and the doctor is satisfied that the disease will not return.

ALTERNATIVE TREATMENT STRATEGIES

As the human body is in a constant state of change, the doctor's diagnosis and treatment approach will sometimes have to be altered. The doctor may also choose to modify each treatment according to the patient's changing symptoms or disease development. When choosing alternative treatment strategies, it is important that the doctor choose adjunctive therapies that enhance the primary goal of purging, tonifying, and regulating the patient's condition.

INTUITIVE AND PERCEPTUAL DIAGNOSIS

THE METAPHYSICAL DIMENSIONS OF MEDICAL QIGONG HEALING

The dictionary describes "metaphysical" as "the study of psychic phenomena beyond the limits of ordinary or orthodox psychology." There are other dimensions of reality that are very tangible, and access to them is only limited by fear and ignorance. Since theories on the nature of the mind and spirit cannot be easily verified or disproved by intellectual analysis or scientific experiment, sometimes certain kinds of knowledge are disbelieved, ridiculed, and dismissed. The advanced energy theories that are presented in this textbook are based upon the five thousand years of cultivation and experience of Chinese medical healing. The energetic form of Qi is neither good nor evil, it is simply another aspect of matter. The ancient Chinese Qigong masters refined the ability to analyze and diagnose these different forms of energy into an effective healing modality.

The existence of energy and its multidimensional patterns is embraced and accepted by many other cultures and societies, with the exception of Western conventional thinking, which is unfamiliar with this unorthodox approach to healing.

These energetic realms, which form the reality of the Medical Qigong doctor's clinical healings, are actually normal perceptional skills that the doctors have developed and trained within themselves. These doctors have simply rediscovered and trained their sensitivity to energies that they were open to during childhood. These energies encompass more then just the physical world. These perceptive skills and abilities of Shen are actually our natural, energetic "birthright," and have been with us since birth. Shen Gong skills, and all psychic powers, lie within the Eternal Soul of an individual, although in most people these skills are undeveloped and largely unused. Often we fear the most the things we understand the least. It has also been my observation that individuals are taught to ignore these skills because of one or more of the following reasons:

1. They have become frightened of them, because as children they were told that these abilities do not exist;
2. They were told that such abilities were evil, or of the 'devil';
3. Unable to understand them, they do not know how to integrate them into their personal lives;
4. They may fear the emotional and spiritual responsibility of knowing the truth about themselves and others; or
5. The known is familiar and therefore more comfortable then the unknown and new.

Sometimes even facing our own human potential alarms us so greatly (fear bordering on terror) that we are willing to deny empirical evidence that these paranormal skills and abilities are real.

We live, therefore, with a paradox. These paranormal abilities seem to be part of our nature, yet they can disturb us so greatly that we often violently reject them. Medical Qigong doctors are healers who have learned to accept these paranormal abilities and direct their energy, spirit and psychic skills towards the focus of alleviating pain and suffering. They are "wounded healers" who have learned to face their fears, and have the courage to walk through and overcome them.

THE AWAKENING

For the Qigong doctor to understand spiritual insights, three phases of energetic transformation must first be experienced. These three phases are collectively known as "the awakening" and are described as follows.

1. The awakening describes the beginning phase or "opening" of the Qigong doctors' Mind and Spirit to their energetic potential. This acceptance becomes rooted within the doctors' intention, emotion, and thought, and sets into motion the correct vibrational field for accessing their higher consciousness.
2. Entering the inner sanctuary describes the phase wherein Qigong doctors begin to receive greater clarity of mind and an elevated consciousness. The doctors receive information on a multidimensional level and are now able to access, as well as better understand their inner core being. This sets into motion the Qigong doctors' understanding of their interconnectedness to everything in the universe. At this point, the doctors must accept responsibility for their actions and non-actions.
3. Accessing the Wuji through the Creative Subconscious Mind describes the Qigong doctors' ability to access and reprogram their Mind and core vibration. This sets into motion the doctors' personal contact with their higher-self, or Eternal Soul.

LEARNING INTUITIVE AND PERCEPTUAL DIAGNOSIS

Before beginning the primary methods of Intuitive and Perceptual Energy Diagnosis, it is important to understand the initial training needed to access and train these skills. Without an understanding of these subtle energies, the risk of misdiagnosis of the microprojections of life-force energy is high. The subconscious mind communicates within ourselves and the outside world through the Three Dantians, by way of three distinct mechanisms:

- Physical or kinetic communication is established through the Lower Dantian.
- Emotional or empathic communication is established through the Middle Dantian.
- Spiritual or intuitive communication is established through the Upper Dantian.

These three modalities of energetic communication must be finely honed into effective, receptive listening tools.

Memories, thoughts, emotions, and knowledge are stored in the patient's tissues throughout the body at the pre-atomic and pre-cellular levels. They are arranged and structured similar to the way sounds and images are recorded on the magnetic molecules of a video cassette tape. Not only are emotions and sounds stored, but tastes, smells, and touch can be invoked simultaneously in vivid detail as the cells are stimulated. The patient feels that they are actually reliving the event from the past exactly as if it were occurring now.

The Qigong doctor must be ready to recognize, uncover, and be able to feel his or her patient's pain when interceding as an empath. Sometimes the doctor will not only stumble upon pockets of unresolved pain, but also suppressed memories hidden deep within the patient's tissues. When this occurs, the doctor diagnoses the patient's toxic emotional symptoms through his or her own body and spirit. To perform this type of diagnostic technique successfully, doctors must be secure in their established emotional boundaries. Otherwise they might project their own feelings onto the patient, or absorb the patient's pathogenic energy into their own body, where it may become trapped.

The goal is to absorb and receive the patient's symptoms briefly for diagnosis, and then release them by exhaling out the mouth (away from the patient). This expels the Toxic Qi into the ground, and allows the patient's feelings to pass through the doctor's body without harm.

The state of consciousness in which the doctor listens and feels the patient is more important than the technique or modality used. This is why there are so many diagnostic tools available to assist doctors in their evaluation. The patient's body maintains the energetic pattern of trauma until it is released and dispersed. The amount of energy required to neutralize this traumatic charge and return the tissues back to their normal state is in direct proportion to the intensity of the original trauma.

RECEIVING AND INTERPRETING THE PATIENT'S FIELDS OF BIO-INFORMATION

The human body emits several "bio-fields" of energy, varying in density and frequency and resonating within a holographic field. Like radio waves and other modern technological wave transmissions, these fields are encoded with information. The key to what we absorb and radiate lies in the energetic qualities of the rate or frequency of vibration, amplitude, and wave length of these bio-fields.

The physical world as we know it is made up of energetic fields. These energetic fields hold our universe together (Figure 25.2). Matter is simply the temporary expression of these energetic fields as they interact. The human body consists of Shen, Qi, and Jing actively resonating within the cells and tissues. Qi, as energy, is considered both substantial and insubstantial. Qi creates matter and acts as a medium between matter and spirit; it carries their mutual interactions in the form of waves. Because matter manifests as a distinctive type of consciousness, it operates at discreet levels of awareness, differing only in its complexity of structure and degree of freedom. Matter also manifests itself at every level of organization, from fundamental particles in physics, to biological organisms within the human body. The cells are governed by these bio-fields which determine their growth.

It is the conscious aspect of energy that dictates how, where and in what form energy is manifested within the body. This energetic consciousness releases bio-information which transmits encoded information about the patient both internally and externally. Form, force, and medium are three aspects of energy held in dynamic patterns by consciousness. The bio-information released from the patient's consciousness is read and diagnosed by the Qigong doctor before treatment begins.

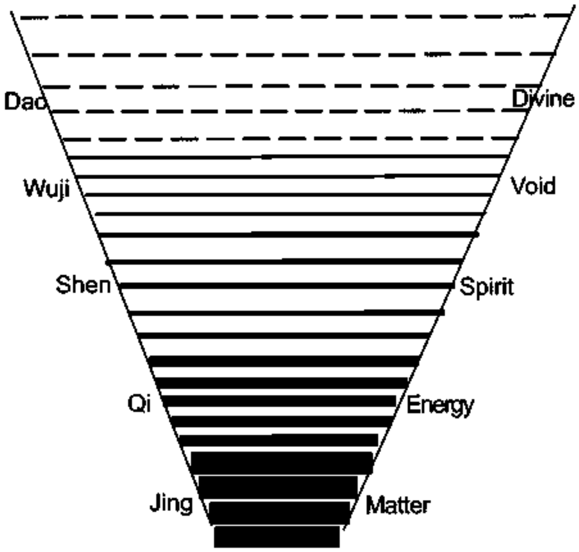

Figure 25.2. The Energetic Fields

The body's bio-information resonation Xin Xi or the "message" as it is called in China, is not limited by space or time. This Message refers to the energy fields that emanate to and from the Wuji or Void (infinite space embodied in between matter and energy).

The Wuji is the state of no boundaries, of pure openness, of complete oneness. Its essence is emptiness (as depicted by the empty circle), which relates to the awareness aspect of our Yuan Shen. When it begins to manifest within the relative universe of Qi, Yin, and Yang (referred to as Taiji), its nature appears as clear light which dispels darkness. When it manifests as a pure and tangible physical form, it appears as an enlightened being (saint or immortal) or as the "divinity" inherent within each one of us.

All of these levels exist within each individual, on both a pure level and pathological level. These levels can be detected and interpreted by the trained Qigong doctor as bio-information resona-

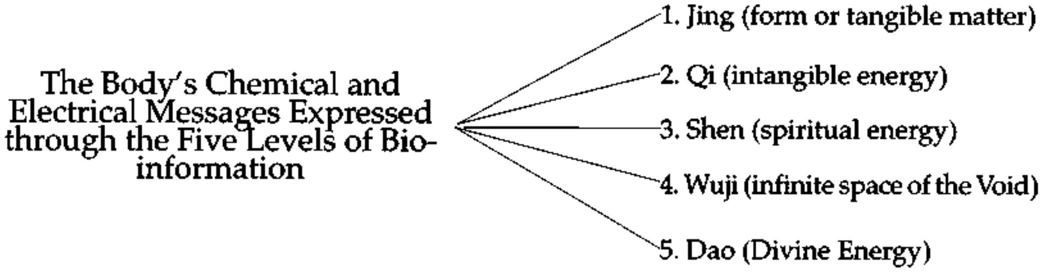
The Body's Chemical and Electrical Messages Expressed through the Five Levels of Bio-information
1. Jing (form or tangible matter)
2. Qi (intangible energy)
3. Shen (spiritual energy)
4. Wuji (infinite space of the Void)
5. Dao (Divine Energy)

Figure 25.3. The Five Levels of Bio-information

nce. Bio-information resonance is subdivided into five levels of expression:

1. Jing (form or tangible matter) contains the least amount of encoded information which is emitted at the lowest frequency.
2. Qi (intangible energy) contains more encoded information which is emitted at a higher frequency than Jing.
3. Shen (spiritual energy) emits an even higher frequency containing still more encoded information than Qi.
4. Wuji (the infinite space of the Void) contains even more encoded information than Shen which is emitted at a still higher frequency.
5. Dao (Divine Energy) emits the highest, most complete encoded information or Message and is the most powerful of all the emissions.

Medical Qigong views congenital disease as a form of distorted bio-informational vibration. The body's cellular systems internally regulate themselves via communication through these coded Messages. Messages are sent and received both on the energetic level (Shen, Qi, and Jing) and on the biochemical-electrical level. These two different levels interact continuously. Distortions on the energetic level interfere with the necessary transformations of Jing into Qi, Qi into Shen, and vice versa. Failures in energetic transformations also affect and distort chemical and electrical Messages between the different types of cells in the body (blood cells, tissue cells, nerve cells, and so on).

The maturity of the physical body depends on the transference of these Messages as the source of its evolution. Each time the egg cell divides (morula, embryonic, and fetal stages) the Message is transcribed into the newly-generated cells. Some encoded Messages are transformed but, due to the adverse evolutionary conditions or pathogenic factors, some become latent. Germ cell Messages, however, are not restricted by these evolutionary conditions since they multiply, evolve, and transform at an accelerated pace.

Information of the body's disharmonies is received as a distorted, biological resonation, which indicates the location and severity of the patient's condition. All diagnostic scanning techniques utilize this type of perception. The body's cellular systems internally regulate themselves via communication through coded bio-informational signals. This biological information exists in energetic, electrical, and chemical forms which are constantly being transformed from one form into another (Figure 25.3). The function of Medical Qigong therapy is to provide the correct bio-informational instructions (or Message) to the patient's body in the form of Qi emission (or coded energetic bio-informational signals) to initiate healing. By receiving Messages from the patient, the Qigong doctor can project energy while receiving and diagnosing the patient's energetic patterns.

ACCESSING THE KNOWLEDGE STORED WITHIN THE WUJI

When the Qigong doctors emit healing energy to the patient, they are sending healing Messages encoded within the Qi. When these Messages are received by the patient's cells the healing process is initiated. Qigong doctors access the knowledge needed to treat patients, by remaining receptive to the Messages stored within the patient's Jing, Qi, Shen, and surrounding energetic space (Wuji or Void).

Within the energy of the Wuji is the knowledge stored throughout time. In Medical Qigong practices, accessing the knowledge of the Wuji is the Chinese equivalent of the Sanskrit concept of accessing the Akashic Records. "Akasha" is a Sanskrit term used to describe the all-pervasive space of the universe. There are two kinds of Void: the first type of Void is internal, limited by the material plane (or matter) and can be associated with the enveloping of the internal human form. Matter can be described as consisting of more space then actual physical matter and, therefore, is considered infinite in its internal division of energetic properties. Within this Void forms the foundation for the matter or the body's personality by way of physical forms, sensations, perceptions, mental formations, and consciousness.

The second type of Void is external, unlimited, beyond all description, unbound by the material, yet contains all things material. It is the vehicle for all life, resonating through sound and light, permeating everything in the universe. In the practice of Medical Qigong the Wuji is one of three universal principles, along with Qi and Shen, which form a trinity of sources of intuitive and psychic power. These three universal principles also manifest the Eternal Soul, allowing divine thought to infuse matter.

The knowledge of the Wuji records the vibrational resonance of every action (all thought and emotion), as well as light and sound. These energetic impressions are stored within the astral plane. These records exist as impressions in the astral/spiritual dimension, providing a sort of accessible filing system for those who wish to receive information about past history or past lives, or even for the examination of their own spiritual progress. Admittance to this sea of knowledge and wisdom can be triggered in the hypothalamic limbic system of the brain, and is accessible through the energetic stimulation of the third ventricle of the Qigong doctor's brain.

The body is composed of literally billions of energetic molecules. Each molecule is a hologram of ancestral particles, knowledge, and experiences existing throughout time, spanning our entire history as we know it. These molecules gather together to form and create matter. They function for specific purposes of life transitions and energetic interaction, and then dissolve, and transform back into Qi and Shen. Each molecule stores its energetic experiences, to be later accessed through spiritual intention.

As the molecules gather to form a fetus, energy as well as ancestral history is stored within the tissues and cells of the child via the environmental, universal, mother and father's energetic fields.

Once the Qigong doctor has fused with his or her patient, he or she will be able to access specific information about the patient through the internal connection to the energetic impressions of the patient's Void. This allows the doctor to study and learn about the patient's past history of disease formations, as well as the information contained within the molecular structures of the patient's tissues.

UNDERSTANDING PAST-LIFE REGRESSION

Although seldom talked about, and much less understood, the clinical significance about past-life regression is quite important. Psychiatrists Sigmund Freud and Carl G. Jung both believed that the patients' worst fears, pain, and traumas are buried deep within their unconscious mind. Sometimes the patients' emotional traumas are so deeply entrenched that they can only be expressed through the energetic presence of a different time-space reality. One metaphor used by patients to describe this other form of reality is through a "past-life trauma."

During Medical Qigong treatments, patients will sometimes describe traumas or situations that do not relate to present life experiences. These experiences might be explained through past life theories. There are six major theories used to explain past-life recall (Figure 25.4). These theories are described as follows.

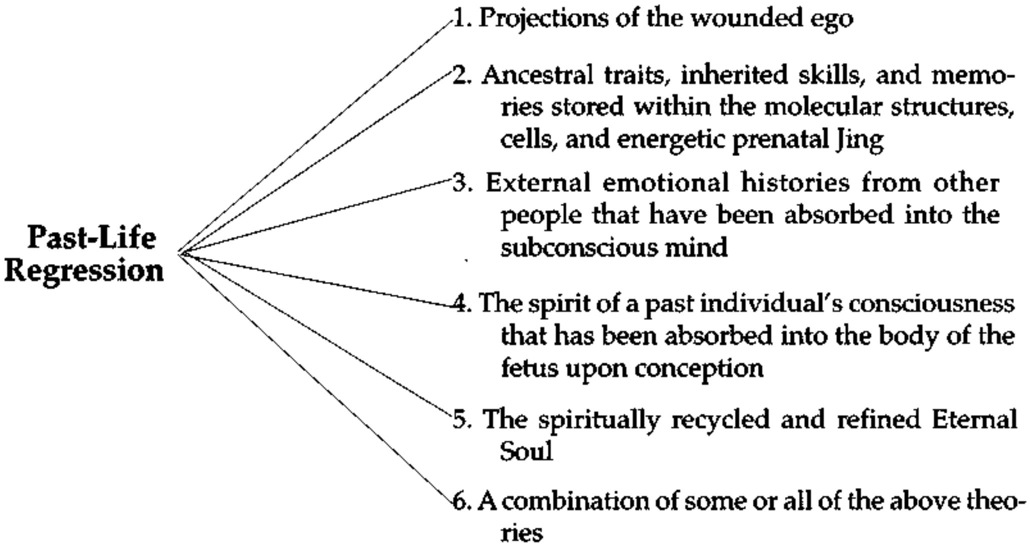

Past-Life Regression

1. Projections of the wounded ego
2. Ancestral traits, inherited skills, and memories stored within the molecular structures, cells, and energetic prenatal Jing
3. External emotional histories from other people that have been absorbed into the subconscious mind
4. The spirit of a past individual's consciousness that has been absorbed into the body of the fetus upon conception
5. The spiritually recycled and refined Eternal Soul
6. A combination of some or all of the above theories

Figure 25.4. The Six Major Theories of Past Life Regression

1. Some Qigong doctors contend that these past-life memories are the projections of the patient's wounded ego, used by the subconscious mind to assist the patient in dealing indirectly and symbolically with life's present painful experiences. These are painful memories that are consciously avoided by the patient, but are necessary for healing to take place.
2. Other doctors believe that these past-life phenomena are the ancestral traits, inherited skills, and memories stored within the molecular structures, cells, and energetic Prenatal Jing of the patient, that periodically surface from time to time and generation to generation.
3. Some doctors believe that apparent memories of past-life regressions, as well as of past-life traumas, are external emotional histories from other people that have been absorbed into the patient's subconscious mind. The patient responds like an energetic empath to sensations of telepathy, registered by the patient's hypothalamic limbic system. These external memories are triggered when the patient accesses the knowledge of the Wuji.
4. Other doctors believe that the spirit of a past individual's consciousness can be absorbed into the body of the fetus upon conception. This means that wherever a child is conceived, the environmental energy, as well as any native hovering spiritual entity, can be implanted into the forming energetic field of the fetus.
5. Other doctors believe that one's Eternal Soul is constantly being recycled and refined through generation after generation, constantly being reborn in order to evolve into a higher plane of consciousness.
6. Still other doctors believe that past-life memories can be a combination of some or all of the above theories.

Some patients feel that they experience spontaneous past-life recall through the treatment of Medical Qigong therapy. While receiving treatments or performing Medical Qigong prescriptions these patients may experience intuitive flashes or visions. Whether the patients' memories actually are of their own historical past lives, or are reconstructions from subconscious material, is a matter of controversy. The primary objective is for the doctor to accept the patients' reality, supporting them in this belief system to facilitate their healing process.

It is the doctor's responsibility to respect the patient's reality and not impose his or her own interpretations on everything the patient says. By accessing the knowledge of the Wuji, the doctor can assist the patient to connect with and benefit from these past life recollections or "reconstructions" to facilitate the healing process. Ultimately it matters not if the memories are reconstructions of subconscious material, true past life events, or even telepathic experiences. What is important, however, is that these memories are processed, learned from, and integrated to promote a better balance within the patient.

PAST-LIFE HEALING THERAPY

Clinically, past-life regression therapy has been effective in treating such problems as chronic fears and phobias, detrimental habits and cravings, suppressed hostility and guilt, depression, the fear of death, sexual dysfunctions, chronic pains and headaches, as well as many physical disorders. It is also used for understanding relationship problems, since the theory maintains that we often meet the same souls in different lives (and bodies) until the problems are worked out.

The Qigong doctor is not qualified to perform past-life regression therapy. This requires a specialist or at least supervision by an expert. The Qigong doctor can, however, facilitate healing by accepting the patient's experience and validating it to prepare the way for the healing process. In Medical Qigong therapy, the doctor accepts the patient's belief that the causes of his or her present physical or psychological problems can be traced to past-life traumas or death experiences. The improvement attained through past-life memories is said to be achieved through the release of the energetic charge that is trapped in the patient's body from a former life. This energetic charge keeps the patient's emotions prisoner, chained to the old feelings that are associated with the painful past-life experience. It is important for the Qigong doctor to realize that the same therapeutic benefits as the retrieval of current-life real memories can be achieved by allowing the patient to relate any images, impressions and so on, that may or may not stem from his or her present life experiences. The foundational base of the emotional trauma will often be rooted in feelings associated with rejection, betrayal, and abandonment. The psychological and physiological reactions to the patient's past-life memories, whether they be real or contrived, will have a profound affect on the patient's healing.

LEARNING MEDICAL QIGONG TECHNIQUES

According to Qigong Master Zheng Zhan Ding from Beijing, China, when learning Medical Qigong techniques, the doctors experience what is known as the Three-Part Wisdoms.

1. The Upper Dantian Wisdom is achieved through divine calling. When the full potential of the Upper Dantian is tapped, the doctor suddenly finds him or herself able to utilize skills and techniques that were previously completely beyond his or her ability. This is a major energetic breakthrough for the doctor, enabling him or her to perform supernatural feats that he or she immediately recognizes and accepts as part of a personal spiritual calling.
2. The Middle Dantian Wisdom is achieved through inspirational observation. Upon seeing energetic techniques performed, the doctor immediately establishes an emotional belief that he or she possesses that same energetic ability. This is a kind of "if they can do it, I can do it" observation and belief structure.
3. The Lower Dantian Wisdom is achieved through practice. This is a type of "practice makes perfect" method of transition which allows the doctor time to accept and integrate new beliefs and abilities through long hours of constant practice, observation, and tutelage.

All Qigong techniques are based on the imagination and spiritual belief. It is necessary to create an image (a spiritual belief) within the mind, to create a positive environment for healing transformations (the spirit transforms energy, and energy transforms matter). Sometimes the Qigong doctor will want to utilize a certain image of his or her belief, but because of certain doubts, is unable to bring it into existence. In order to successfully cultivate this image, the doctor must use creative thinking to focus his or her life-force energy on this specific objective, accept it as true and know it will happen. To create this image the doctor must use his or her imagination, emotional feelings, unwavering belief, visualization, affirmation, and acceptance.

Energy (Qi) is the bridge between the material and the spiritual, and is the medium through which consciousness interacts with the world. Through intention, Qi is gathered, focused, and used to transform light, sound, and other resonations. Through intention, the Qigong doctor is able to change structural formations, and to transform tangible and intangible forms and vice versa, to dissolve tumors, cysts, and so on. Of the many methods of healing available in Medical Qigong, it is important for the Qigong doctor to explore the healing skills suited to his or her body and personality, to become most effective. Once the doctor has learned and mastered these healing techniques, he or she will no longer need to think about the methods, or when they need to be employed, they will become second nature.

PERCEPTUAL COMMUNICATION TRAINING

Subconscious communications are divided into physical, emotional, and intuitive forms of interaction and perception. The internal-vision techniques required to utilize energetic perception in a clinical setting are expressed and enhanced through these types of communication. These types of communications are realms of perception that can be explored, wherein, the Qigong doctor is so present in "the moment" that his or her attention begins to dissolve the perceptual realities of all five senses, and begins to perceive and diagnosis patients on a spiritual level. At this level, all feelings, sounds, colors, or any combination of the doctor's senses take on new dimensions of reality.

DIAGNOSIS THROUGH DANTIAN COMMUNICATION

Any treatment, exercise, or meditation the Qigong doctor prescribes must connect with the energetic resonation of the patient's disease. This connection is established through the doctor's ability to communicate with the patient's body via the Three Dantians.

1. The Upper Dantian's perceptional awareness is received through a type of spiritual/intuitive communication. Through this intuitive subconscious communication the doctor is able to interact and perceive the patient's spiritual and energetic disturbances. This allows the doctor to perceive the patient's spiritual condition without bias.
2. The Middle Dantian's perceptional awareness is received through a type of emotional/empathic communication. Through this empathic subconscious communication the doctor is able to interact and perceive the patient's feelings, actively empathizing, listening, and hearing what patients communicate about their pain and problems. This allows the doctor to perceive the patient's true emotional condition by bypassing the person's ego.
3. The Lower Dantian's perceptional awareness is received through a type of kinesthetic communication. Through this physical subconscious communication the doctor is able to interact and perceive the active condition of the patient's physical tissues. This allows the doctor to perceive (through his or her hands and physical body) the patient's physical symptoms, such as pain, Heat, and so on.

UTILIZING INTERNAL-VISION AS A CLINICAL MODALITY

The brain has the ability of receiving and recording the vibrations of any object that the senses see, hear, smell, taste or feel. The vibrations of light, sound, smell, color, and object's shape are all recorded. This also holds true for thoughts, words, and actions. The brain also has the ability of reproducing these vibrations and projecting them out again through the use of inner vision and the imagination.

Before beginning internal Qi diagnosis it is important that the Qigong doctor have a thorough understanding of the internal energy's function and flow within his or her own body. Dr. Li Shi Zhen pointed out during the Ming Dynasty that, "The internal organs and channels can be perceived as unobstructed and free from disease only by those who can see internally."

This is achieved by performing the Nei Guan or Internal Viewing Meditation. The purpose for internal-vision is to carefully examine, through conscious introspection, the energetic flow and structure of the body and mind. Most doctors in the Ming Dynasty studied the Nei Guan Meditation as part of their clinical practice to develop the light for internal viewing. It is believed that light enters the body through the Baihui Point (at the top of the head), the Yin Tang Point (the Third Eye area) and through both eyes. Since internal-vision uses light that has a higher vibration than visual light, it is able to penetrate through the skin deep into the tissues.

Many years ago I was truly skeptical of this phenomenon, although I met many Qigong masters who possessed such skills. After acquiring this skill, however, I was able to demonstrate its effectiveness quite thoroughly while interning in the Qigong hospitals of Beijing, China. It is actually a simple matter of consistent practice and focused concentration which allows this sense of perception to become tangible. It is important to remember that to know something, one must first think it, then feel it in order to integrate the knowledge fully. Always watch as an observer, void of emotional attachment, noticing what is actually there.

Nei Guan—Internal Viewing Meditation

The best way to practice the Nei Guan Meditation is through relaxed, quiescent breathing. Begin from a Wuji posture. After quieting the mind, relax and focus the imagination. Imagine that the left eye is a bright radiant sun, while the right eye is a bright and luminous moon. Both the sun and moon join together at the Yin Tang (Third Eye point). As their energies come together, they form a bright white ball. This light actually gathers around the pituitary gland, illuminating the optic chiasm, filling the occipital lobes (for internal perceptual vision) and the thalamus (for oculomotor control) (Figure 25.5).

As the light expands it stimulates the corpus callosum and penetrates the pineal gland (which acts as a projector for internal vision) (Figure 25.6).

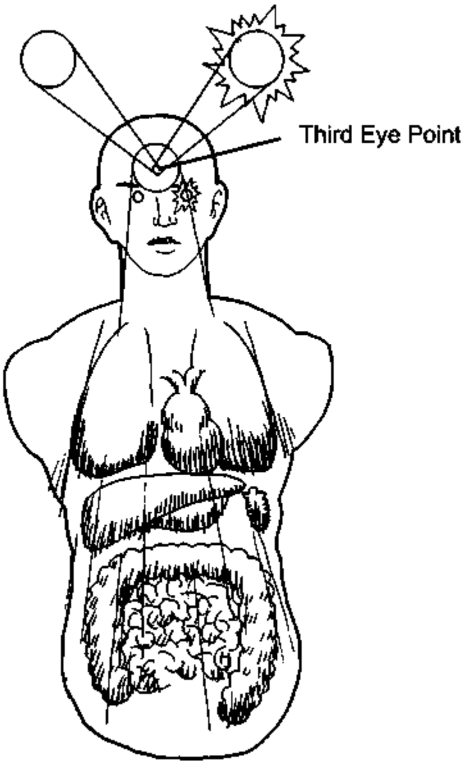

Figure 25.5. The Internal Viewing Meditation

Next, focus your intention on directing this ball of white light, allowing it to shine down into the body, illuminating all the internal organs. This enables you to see directly inside your own body and view each organ. Observe and feel the energetic strengths and individual weaknesses of the bones, organs, and tissues. Allow both energy and information to gather and expose itself through the Third Eye region of the Upper Dantian. After a while, relax and return to Wu Wei (i.e., No Mind), allowing all of these images to settle.

After a relatively short time of constant practice Qigong doctors learn to discern the state of their own organs, and are able to regulate their own physical and energetic body. This internal diagnostic ability sets the foundation for doctors to extend their Internal-viewing capability outside their own body to use on patients.

Once this occurs, doctors find it possible to effectively diagnose patients while scanning and using the Flat Palm Detection technique. The patient's points and areas for diagnosis are revealed to the doctors through this modality of energetic extension.

When using internal-vision, extend your intention as deep into the patient as possible, to slowly scan the patient's tissues, layer after layer, observing any cellular changes and distortions. Use the internal light resonating from inside the patient's body to identify the pathogenic factors. Then, look deeper into the energetic currents to find the root and origin of the patient's disease.

When you use your hands to read the patient's internal organs through Flat Palm Detection or any other diagnostic modality, it is extremely important to use internal-vision to determine the severity of the disease. You may observe a specific color, texture, or feel pools of emotional energy. According to Dr. Zheng Zhang Ding, doctors will be able to determine the condition of the diseased organ by its shade of color. The internal-vision colors most commonly observed in internal-vision diagnosis are as follows:

- A red or yellow color represents a healthy or normal organ;
- A white color signals the beginning of energetic dysfunction, but is not serious enough for treatment;
- A gray color signifies sickness; and
- A black color denotes, tumors, cancer, etc.

APPLICATIONS OF INTUITIVE AND PERCEPTUAL QI DIAGNOSIS

There are a variety of methods concerning the application of intuitive and perceptual diagnosis of diseases. The specific physiological methods such as remote sensing, hand detection and physical body observation are very popular in the Medical Qigong clinics in China today. Five of the most favorite modalities of intuitive and perceptual Qi diagnosis are described as follows.

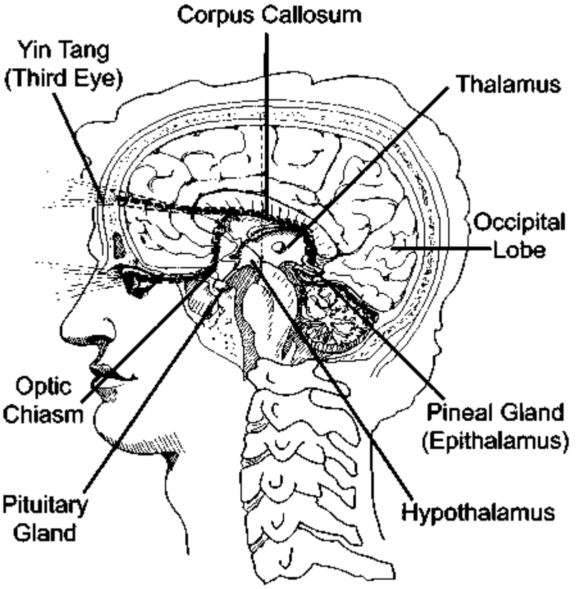

Figure 25.6. The inner light gathers at the pituitary gland, illuminating the optic chiasm, filling the occipital lobes and thalamus, stimulating the corpus callosum and penetrating the pineal gland.

DIAGNOSIS BASED ON FLAT-PALM DETECTION (HAND SENSING)

Because the body's energetic field reflects the state of health or dysfunctions of the physical body, any changes occurring in the tissues and organs are detectable through sensing this energetic field. Detection of these changes with the palm is termed Flat-Palm Detection.

To perform this technique, the doctor begins by asking the patient to take a standing, sitting or lying posture. The patient's whole body should be relaxed, the mind focused on breathing naturally. The doctor positions his or her palm several inches from the patient's energy field, with the Laogong Pc-8 point or fingertips facing the direction of the patient's body.

The doctor then emits a little energy, slowly, followed by Pushing, Pulling and Leading manipulations used to detect the Qi of the patient's energetic body. The patient's surface tissue is divided into upper and lower, left and right aspects. The doctor focuses on the surface tissue area first, specifically the area corresponding to the region of the patient's disease. The doctor then determines the energetic nature of the corresponding regions by making comparisons between the sensations gained in different areas to find where energy feels abnormal. Gradually, the doctor's emitted energy is extended deeper into the body to purge, disperse, tonify or regulate the patient's organs and organ systems (Figure 25.7).

The Qigong doctor then analyzes these energetic sensations, comparing how they feel from the different areas of the body (both on the superficial and deep layers). The doctor categorizes the pathological findings (the location, nature, and severity of the disease) in accordance with the Eight Principles of Diagnosis and the energetic patterns of The Five Elemental Theory.

Feeling a great difference in energy intensity between the two sides of the Stomach Channels, for example, indicates a Yin and Yang imbalance of energy circulation between the left and right sides of the body. The Deficiency and Cold syndrome of the Spleen and Stomach may be manifested by a cold and contracted sensation in the epigastric region.

When an acupuncture needle is inserted into an obstructed point the energy grabs the needle and vibrates it. Once this energy block is removed, the needle ceases to vibrate because the channel is now open. This same phenomenon occurs when the palm is placed on an area of stagnant Qi. The palm vibrates until the obstructed energy is dispersed.

Energy blocks are commonly cleared by dredging in a downward direction from the head (Heaven) to the feet (Earth). When treated this way, patients will experience a temporary feeling of fatigue as the stored toxins are released from the body. Tonification is facilitated by bringing the energy current up the patient's body from the feet to the head (Earth to Heaven). This action will sometimes cause patients to have psychic experiences and see visions, as the energy enters the patient's Upper Dantian region.

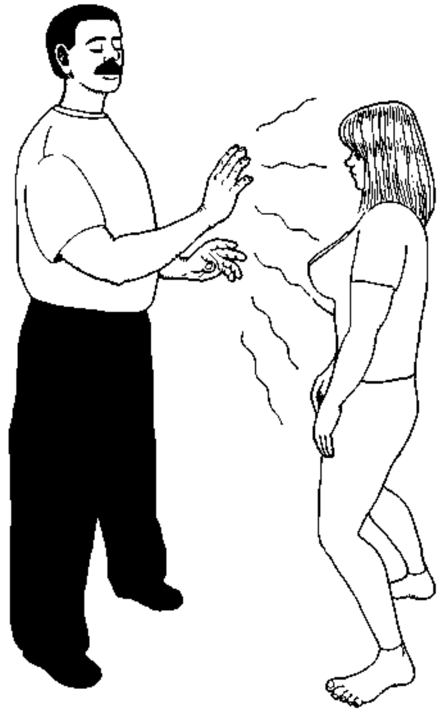

Figure 25.7. In Flat Palm Detection, the Qigong doctor uses the entire palm to sense the condition of his or her patient's internal organs and tissues.

DIAGNOSIS BASED ON KINESTHETIC AND EMPATHIC QI ABSORPTION

There are some individuals who tend to easily absorb the feeling and emotions of other people, even without trying. Sometimes when treating a patient, a Qigong doctor will inadvertently ingest a small amount of their patient's energy. As this energy absorbs into the doctor's body, it infuses throughout his or her entire energetic field. This energetic fusion sometimes causes the doctor to experience abnormal feelings in the locations that correspond to the pathogenic areas of the patient's body. From these feelings, the doctor can diagnose the location and the nature of the patient's disease.

To actively perform this method of Kinesthetic and Empathic Qi Absorption, the Qigong doctor stands facing the patient (Figure 25.8) and projects white healing light energy toward the top of the patient's head (Baihui GV-20 point). The patient is thus enveloped by the white healing light energy, from the top of the head, to the bottom of the feet. Next, the doctor directs the Qi to flow from the bottom of the patient's feet into the bottom of his or her own feet, absorbing the patient's energy. This energetic action forms a complete cycle of energy between the doctor and the patient's body. The Qigong doctor then begins to carefully feel the sensations of the patient's energy in his or her own tissues and organs. The locality of abnormal feelings, their shape, and/or prominent characteristics reflect the locality and nature of the patient's illness.

Note: It is generally encouraged to avoid this type of body detection on cancer patients, because the risk of pulling in the patient's excessively Hot pathogens into the doctor's body is too high, especially if the doctor's energy has become depleted.

HAND DIAGNOSIS BASED ON BODY ASSOCIATION

A small percentage of energy absorbed from the patient can suffice to diagnose the whole energetic system. Because the law of bio-holograms holds that one part of the body contains information about the whole (including various parts of the body, such as the viscera, Qi, and Blood, etc.), information obtained from certain isolated areas can reflect the condition of the whole body. Such parts may include the hand, foot, eye, ear, nose, face, etc. The location on these parts that represent different areas and organs are termed as "points" or "zones:" they include hand points, foot points, ear points, and so forth.

Hand diagnosis (called Shou Zhen) is used to evaluate the physical predispositions and ailments of patients. This diagnostic technique first appeared in Chinese medical literature during the Sui Dynasty (618-581 B.C.) from the clinical writings of Dr. Chao Yuan Fang, who authored the "General Treatise on the Origin and Symptoms of Various Diseases." To begin diagnosis through hand-body association, the Qigong doctor first neutralizes his or her left palm (by releasing the energy into the Earth) to balance and stabilize its sensitivity. The left hand is generally used for reviewing energy and diagnosing diseases.

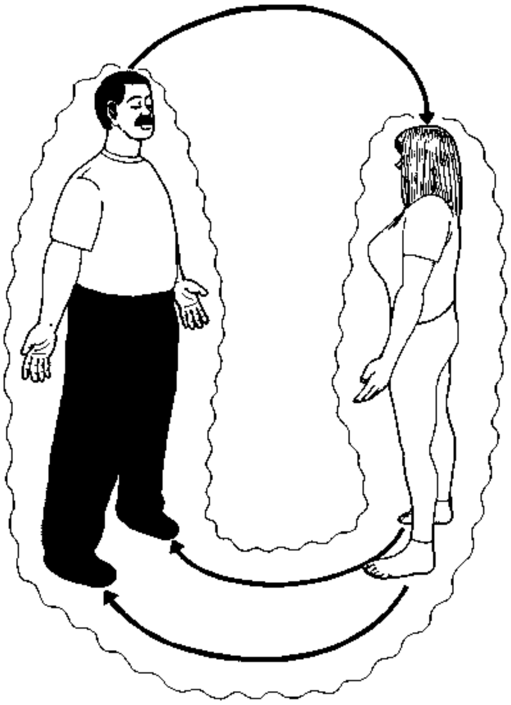

Figure 25.8. In Diagnosis based on Kinesthetic or Empathic Absorption of the patient's Qi, the Qigong doctor temporarily absorbs the patient's energy field into his or her own body.

Once the left palm has been neutralized, the doctor uses his or her intention to gather the patient's Qi into the left palm, distributing the absorbed energy evenly throughout the hand. Focused concentration on any one specific area of the hand should be avoided. The doctor imagines the patient's energy enveloping the left hand and fingers, like an energetic glove. The doctor carefully begins to sense the sensations of the patient's energy on his or her palm. The center and front of the doctor's palm is considered the center and front of the patient's body, while the back of the hand is considered the back of the patient's body. The doctor feels various kinds of sensations on specific areas of the left palm, which indicate the location of the patient's disorders. These different sensations,

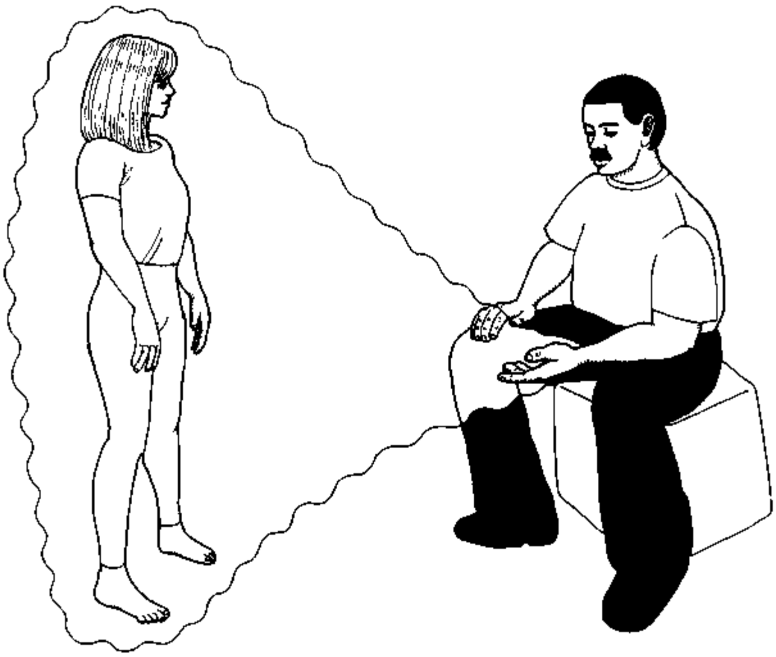

Figure 25.9. Diagnosis through Hand/Body Association

such as distention, heaviness, numbness, heat, cold, itching, and throbbing correspond with the features of the disorders (Figure 25.9).

While in Beijing, China, I was fortunate to study under Qigong Master Zheng Zhan Ding. Master Zheng had a remarkable gift for diagnosis and was compassionate enough to share his particular skills and training with me and two of my senior Medical Qigong students.

There are many schools of teaching for hand diagnosis. I have used several methods with great success in diagnosing and treating patients in both China and North America. The primary requirement for correct diagnosis is the use of the mind's intention. The imagination sets the foundation for the "energetic blueprint" of the patient's body (Figure 25.10). The method for employing this technique is described as follows.

1. Begin to scan the left palm and feel for Hot (Excess) or Cold (Deficient) areas. These areas will relate to specific locations of illness. The doctor begins to treat these regions of disease by using one of the following three methods:
    a. The doctor uses intention to tonify or purge the troubled areas inside his or her own left palm;
    b. The doctor uses his or her right hand (usually with the index and middle finger) to tonify or purge the energy from his or her left palm; or
    c. The doctor extends and treats the patient's body with his or her right hand while reading and diagnosing with the left palm.
2. Another variation of hand diagnosis is practiced by first having the patient breathe evenly into the Lower Dantian. Next, the patient focuses on his or her entire body, while placing the right palm (facing outward) towards the direction of the doctor. The patient is encouraged to relax while keeping his or her palm steady.

The doctor slowly moves the index or middle finger of his or her right hand towards the patient's right palm. While continuing to focus on the patient's energetic field (emitted from the patient's extended right palm), the

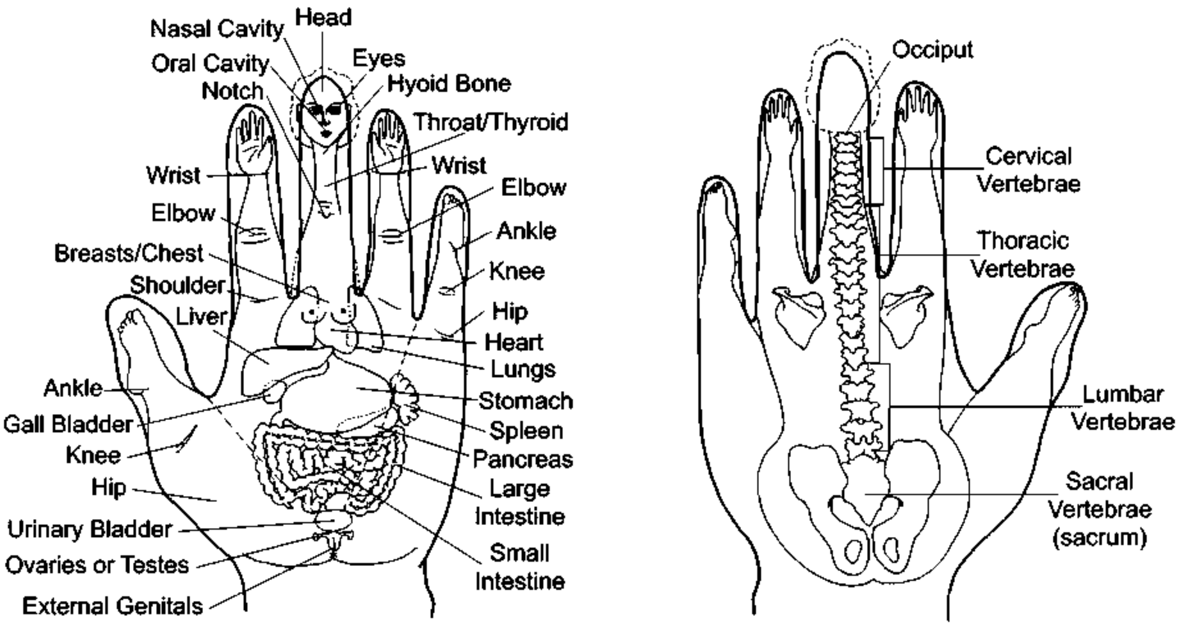
Nasal Cavity
Head
Oral Cavity
Eyes
Notch
Hyoid Bone
Throat/Thyroid
Wrist
Wrist
Elbow
Elbow
Ankle
Breasts/Chest
Knee
Shoulder
Hip
Liver
Heart
Lungs
Ankle
Stomach
Gall Bladder
Spleen
Knee
Pancreas
Hip
Large Intestine
Urinary Bladder
Ovaries or Testes
Small Intestine
External Genitals
Occiput
Cervical Vertebrae
Thoracic Vertebrae
Lumbar Vertebrae
Sacral Vertebrae (sacrum)

Figure 25.10. Front and Back View of the Hand/Body Association

Qigong doctor gently points to where he or she feels a particular energetic sensation (Figure 25.11). This special feeling or sensation that the doctor is receiving, represents the location of an energetic problem or disease. While scanning the patient's right palm, each different feeling or sensation represents a diversity of energetic manifestations related to the patient's disorder.

DIAGNOSIS BY OBSERVING AURA FIELDS

Some Medical Qigong doctors use aura readings to determine the physical, mental, emotional, energetic, and spiritual health of the patient. Aura colors will sometimes appear in the form of waves or other energy patterns (for example: evenly layered, blotchy, mixed together, etc.). Aura colors are in a state of constant flux, shifting their colors and patterns with each shift of the patient's moods and emotional thoughts. These colors merge and blend as they fluctuate, weaving in and around the patient's field of Wei Qi.

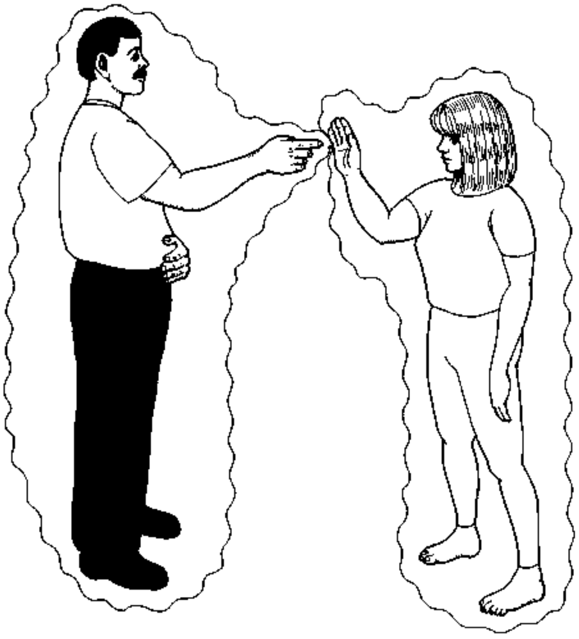

Figure 25.11. Variation of Finding the Patient's Disease Area by Diagnosing the Patient's Hand

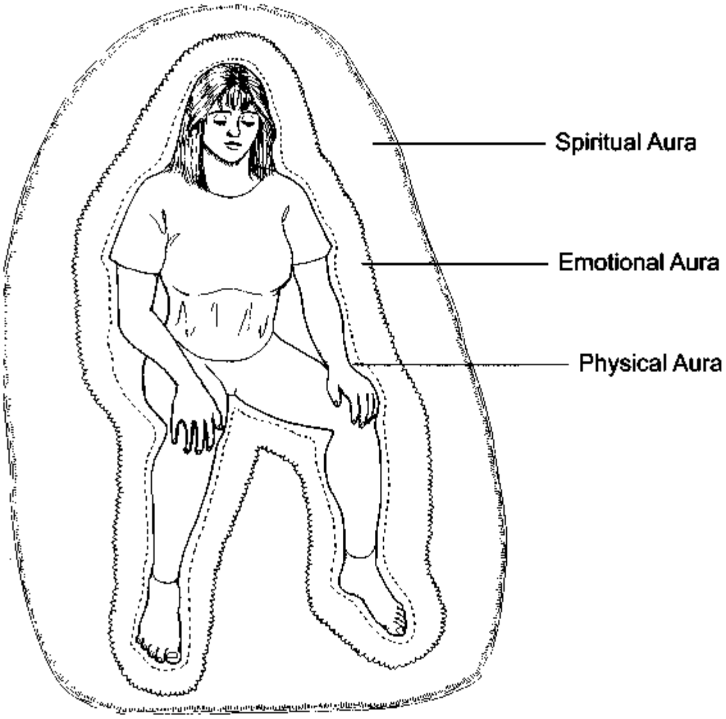

Figure 25.12. Diagnosis by Observing the Patient's Aura Fields.

To begin observing auras, ask the patient to relax his or her whole body, while sitting or standing in front of you. The patient should breathe evenly, to release any distractive thoughts, and then concentrate on the location of the specific disorder. The Qigong doctor focuses his or her attention on the space surrounding the patient's body, while simultaneously concentrating on his or her own Upper Dantian.

Focusing on the Upper Dantian to view the patient's body facilitates seeing the patient's aura field. The patient's aura field may be observed in the form of transparent or opaque Qi flow, or masses of different densities and colors. This may be used as a basis for determining the location and features of diseases (Figure 25.12). Each Yin and Yang organ will have its own color vibration.

When the energy of an organ is healthy, the Qi color is clean, clear and lucid. When an organ is diseased its Qi becomes dark gray and turbid. These colors reveal whether an organ is in an abnormal state and to what extent it has been affected.

COLOR VIBRATIONAL CLASSIFICATIONS

The variety of shifting colors can be interpreted according to their density and tone. The following chart displays the color frequency correlation in nanometers. All living organisms vibrate at a frequency between 300-2,000 nanometers. When analyzed, specific colors emanate consistent frequencies and wave forms. The slower frequencies register in the infrared light spectrum. The highest vibrational frequencies register in the ultraviolet light spectrum (for more information see Chapter 4).

Red	700 nm	Infrared
Orange	600 nm	
Yellow	550 nm	
Green	500 nm	
Blue	450 nm	
Violet	400 nm	Ultraviolet

SEEING AURAS AND EMOTIONAL ENERGY

Seeing auras is a visual phenomenon which occurs after many hours of regular Qigong training. This ability to see the external energy field is not limited to Medical Qigong training; many people are born with this special gift. The word "aura" literally means "breeze," manifesting as shimmering layers of luminous, colored energy that circulate around and penetrate into the physical body. These colors are part of the body's biophoton emissions (radiant light energy that emanates from all living systems). These colors are constantly in motion, reacting to thoughts, feelings, emotional patterns, and environmental influences.

Through the stimulation and physical dilation of the occipital lobe of the brain, the doctor begins to develop this unique perceptual ability. In the beginning stages the doctor may see energy coming off the patient's body like steam.

Later, brilliant, luminous colors become visible. The doctor is actually observing the infrared and ultraviolet radiation color patterns that the patient's body emits. The infrared spectrum is just below the average body's visual spectrum, while the ultraviolet is just above the average visual spectrum. By stimulating and dilation the occipital lobe, the visual spectrum expands to include the infrared frequency, allowing the doctor to observe from a much greater range of colors.

It has been my personal observation that students who possess a large bump on the back of the cranium, specifically between the channel points GV-16 and GV- 17, begin to see auras at a very early stage. These points are below and just above the external occipital protuberance and surround the internal branches off the occipital arteries, veins and nerves. When energy fills this area, the stimulation and dilation of these nerves and arteries causes the visual receptors of the brain to observe certain phenomena at a faster rate, which causes time to seem to slow down (Figure 25.13).

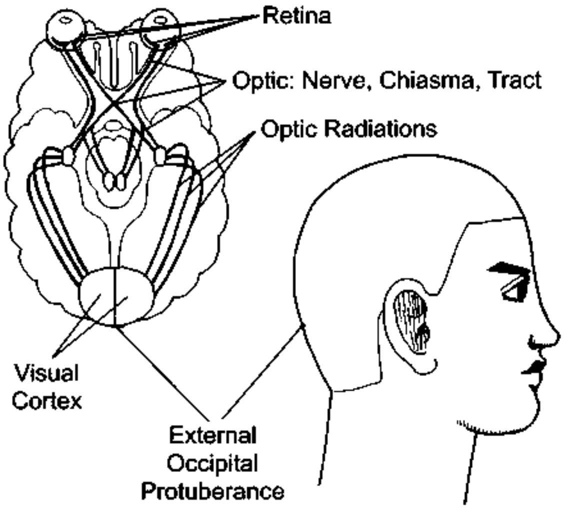

Figure 25.13. As the occipital lobe fills with Qi, the visual spectrum changes.

It is very important to note that these observations are all possible due to the occipital membranes being dilated. To maintain this dilation the doctor must remain calm and relaxed. Any tension brought about by stress only diminishes this ability. Auras are generally seen through the peripheral field of vision.

If the doctor begins to stare at any particular object, focused concentration will contract the occipital membranes, causing tension and pressure that inhibit the dilation process. The secret in maintaining this altered state of observation is to anchor the mind deep into the ground. By focusing the mind on a specific area, focus can be shifted to receiving and observing auras in addition to extending energy without distraction.

The following meditation can be used to stimulate the occipital membrane and enhance the perception of auras.

Seeing Aura Meditation

This meditation can be practiced either sitting or standing. The mind must be free of all distractions and tensions in order for the energy to permeate the occipital lobe. Draw an imaginary line from the top of the head (Baihui) down the spine, down the back of the legs, out the heels and deep into the Earth. This is your mental anchor line which is used to drain your mind of all thoughts (Figure 25.14).

By continuing to drain your thoughts deep into the Earth, you free your Yuan Shen to rely on intuitive perceptions that allow for this altered enhanced peripheral field of vision. This meditation should be practiced for a period of 15 to 30 minutes each day to establish a pattern of relaxation.

As you scan a patient, look to the sides of the patient's body, observing the silhouette of the outer shell of the patient's Wei Qi. It is important not to stare, but just observe the patient's external Wei Qi field (e.g., observing where the field is large, full, thin, or broken). Sometimes different lighting will affect the dilation of the occipital lobe, enhancing visual receptivity.

Next, allow the silhouette to become different colors. Usually, in the beginning, you can see only two or three different colors. With time and practice you will be able to differentiate more colors, with finer detail.

DIAGNOSIS BASED ON THE BODY'S EXTERNAL CHANNELS AND AURA COLORS

Each of the five emotions are related to both the Yin and Yang organs and their Yin and Yang natures. These five emotions are also related to five major colors, both inside and outside of the body's physical structure. By comprehending the colors and understanding their connection to the emotions of the body, the Qigong doctor can successfully analyze the abundance or depletion of energy in his or her patient's major organs. Not only will this reveal the patient's present psychological state of being, but will also expose which internal organs are weakened.

The color chart (Figure 25.15) is used to correlate the patient's energetic fields. These colors are similar, yet have their own unique variations from the physical body's tissue colors.

Think of these colors as extensions from the original organ pools. The mist of the organ pools flows into the body's aura fields, and can be observed in the patient's external energy fields. They are generally very vibrant colors and veil themselves around the first two layers of the Wei Qi. Observe their location around the body, whether they are static or pulsing. If a particular body organ malfunctions, or if any disease is present, then certain parts of the aura may become dim, murky, or produce a dark void.

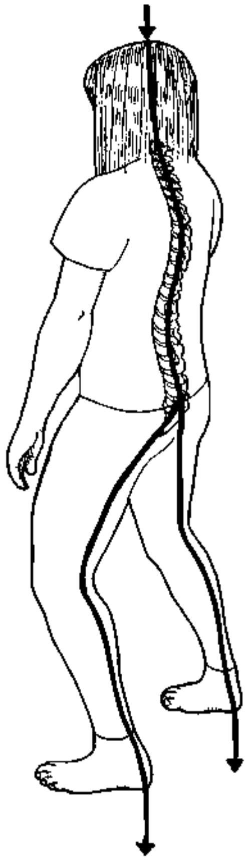
Figure 25.14. Root the Mind deep into the Earth.

It is important to remember that each patient has his or her own color schemes. When observing the color spectrums of the patient's energy field the most important characteristic is the clarity of the colors. Observe if the colors are sharp, murky, or too concentrated.

The Yang patient's colors will vary in tone and brilliance from the color spectrums of Yin patients. Each aura adjusts to match the patient's mood and temperament. People also often choose for their environment colors that reflect the characteristics of their aura and emotional spectrum (e.g., clothing, furniture, house, and room color, etc.).

When using color energy for healing, the Qigong doctor chooses which specific color to emit, e.g., using blue to cool the Qi or red to heat the Qi (see Chapter 4). Generally, when treating patients, the Qigong doctor will either wear all white (which is the combination of all colors, and is used to naturally project energy), or all black (which is the absence of all colors, and is used to naturally absorb energy), but never the color red (the color red is too intense or "Hot" for Qi extension).

DIAGNOSIS BASED ON INTENTION

The doctor's own developed inner vision can provide a diagnosis through observation of a mental image of the patient. The doctor begins by observing an image of the patient as a complete and whole energy system. Next, the doctor imagines the energetic form of the patient's internal organs (Figure 25.16).

If any part or organ is dark or gives a feeling of turbidness, this indicates the location of the disease. When using this type of energetic diagnosis, the sense of smell, the sense of taste, direct perspective, and remote sensing are also helpful diagnostic skills to enhance detection.

LONG DISTANCE MEDICAL QIGONG THERAPY

Long Distance Medical Qigong Therapy has been with us for centuries. One of the ministers of Huang Di, the Yellow Emperor and patriarch of Chinese medicine, was a shaman named Zhu You. Zhu You was a famous doctor who advocated exorcistic prayer over the use of acupuncture needles and herbs to treat illness. Some scholars believed that Zhu You practiced Qi emission healing at the same time that he prayed for his patients. So effective was this healing approach that The Yellow Emperor's Classics states that, in ancient times, most illnesses were treated according to the methods of Zhu You. In fact, at one time, professional "prayer healers" were once widespread throughout China.

There are several schools of thought on the modalities of long distance healing. Each school

Internal Organ	Aura Color
Lungs	Pure White
Large Intestines	Off-white
Stomach	Deep Yellow
Spleen	Orange Yellow
Heart	Red with Slight Blue Tint
Small Intestine	Pink
Bladder	Deep Blue
Kidneys	Light Flame Blue
Pericardium	Purple Red
Triple Burners	Orange Red
Gall Bladder	Yellow Green
Liver	Deep Green

Figure 25.15. The internal organs are each associated with a specific aura color.

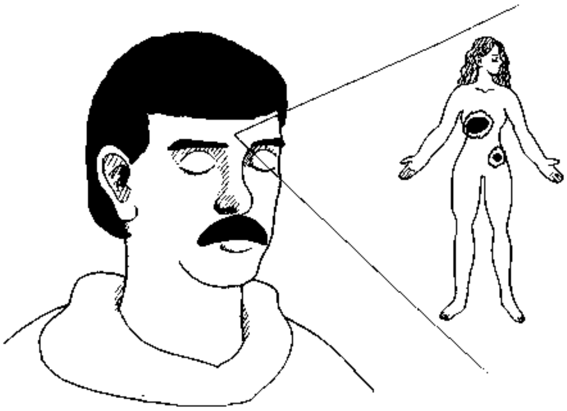

Figure 25.16. The Qigong doctor can diagnose the patient's body through energetic observation and intention.

has its own unique, yet extremely effective application of this esoteric skill.

Although I believe in the effectiveness of prayer, at one time I sincerely doubted the existence of long distance healing. I was thoroughly convinced that it was a matter of projected influential visual and auditory input, similar to hypnotherapy. I expected the doctor to talk to the patient and have him or her close the eyes to facilitate an emotional and mental healing over the telephone. The difficulty in this kind of healing is that if the patient does not trust you, his or her consciousness will not cooperate and will sabotage any healing energy extended in his or her direction.

As fate would have it, one of my Medical Qigong instructors called me one day. As we were conversing, my Kidneys began to heat up. This heat was quite intense, and I was extremely startled because no prior warning of this particular kind of treatment had been given. Yet, I could not dismiss the overwhelming searing heat that I felt in my lower back, Kidneys, and Mingmen area. This healing was done without my conscious consent or foreknowledge! My instructor laughed and explained that as she was talking to me, she was holding a pillow (which represented my body) and visualizing treating my Kidneys with her right hand.

At that point, my beliefs surrounding long distance Medical Qigong healing began to change radically. I began to inquire as to the qualifications needed to treat with that modality of therapy. I learned from my instructor that long distance energy transformation is similar to close range energy transformation. The only difference is that in long distance healing, the doctor's faith and psychic faculty has to be extremely developed and accurate.

To become proficient at long distance therapy, the Qigong doctor must extend his or her thought vibration into the atmosphere, which connects with the thought vibration of the distant patient. Both vibrations collect and blend, drawing together at a selected point on, or within, the patient, creating and reproducing the patient's energetic fields, organs, and tissues.

Zhu You's long distance healing through prayer may be compared to the effectiveness of prayer groups today. Larry Dossey, M.D., for example, documents one double-blind experiment involving nearly 400 patients admitted to the coronary care unit of San Francisco General Hospital for heart attacks or suspected heart attacks. These patients were evenly subdivided into two groups. Both groups received "state of the art" medical care; the only difference was that one group was prayed for by Protestant and Catholic prayer groups throughout the country. These prayer groups had only the patients' first names and sketches. The patients in the experimental group had no idea they were being prayed for; yet when the study was concluded, they experienced significantly fewer complications and deaths than did the non-prayed for control group.

SPACE/TIME AND ENERGY

To study long distance energetic healing, we need to look at a three-dimensional energetic pattern which includes Positive Space/Time, Negative Space/Time, and Energy.

1. Positive Space/Time Energy includes commonly accepted forms of energy (such as electromagnetism, nuclear energy and gravity), and all vibrations ranging from extremely slow to the speed of light. It includes lower vibrations (from inanimate objects) and the higher vibrations of thoughts.
2. Negative Space/Time Energy includes all vibrations which travel faster then the speed of light (super-luminal), and accounts for such phenomenon as telepathy, psychokinesis, precognition, and levitation. In negative space-time, particles have a negative mass and are known as antimatter. When activated to super-luminal speeds, these particles appear to dematerialize, and produce a levitation force which balances the gravitational effects of positive space/time. This also accounts for the Qigong doctor's ability to transmit energy through all known force shields (excluding mirrors, due to their ability to reflect light-energy).
3. Energy and matter form a continuum as does space/time. Energy is the medium between

the positive and negative forms of space. It is the medium between which matter and spirit interact with the space/time continuum, through which heat, light, sound, and resonant vibrations travel.

Matter is composed mostly of space with a minute amount of mass, bound in a particular pattern. This pattern is unique for every form of matter, including the human body. Particles of matter in the human body are constantly vibrating in intermediate stages of energy interaction. Consciousness (the qualitative and directional aspect of energy), acts as a medium for the Qigong doctor's intention to project into the time/space energy field. Consciousness functions as the energy source for resonant vibrations which create the physical world.

To perform long distance healing, the Qi in the Lower and Middle Dantians (the body's Fire and Water energy) must overflow and transform into Shen. The Shen is then united with the divine Qi and extended outside the body for long distance healing of the patient. This allows the energy to cross both spatial and temporal boundaries, thus enabling the Qigong doctor to heal patients in the absence of physical contact.

The first priority in long distance healing, is to establish a connection between all three Dantians through the absorption of divine Qi into the Taiji Pole; this process is called the "empowerment with divine Qi," or simply, the "hookup." This action surrounds the doctor's Wei Qi with Tian Qi (Heavenly Energy), which then is transformed into Tian Shen (Heavenly Spirit).

The doctor then visualizes a long stream of white light extending from the Yin Tang (Third Eye) point, which connects to the patient's body (Figure 25.17).

Once the doctor feels this connection, he or she may now begin the treatment. The physical manifestations of long range healing are the result of the Shen affecting the Qi, and the Qi affecting the body's Jing.

LONG DISTANCE SCANNING EXERCISE

Long distance scanning can be develop by practicing the following mediation exercise.

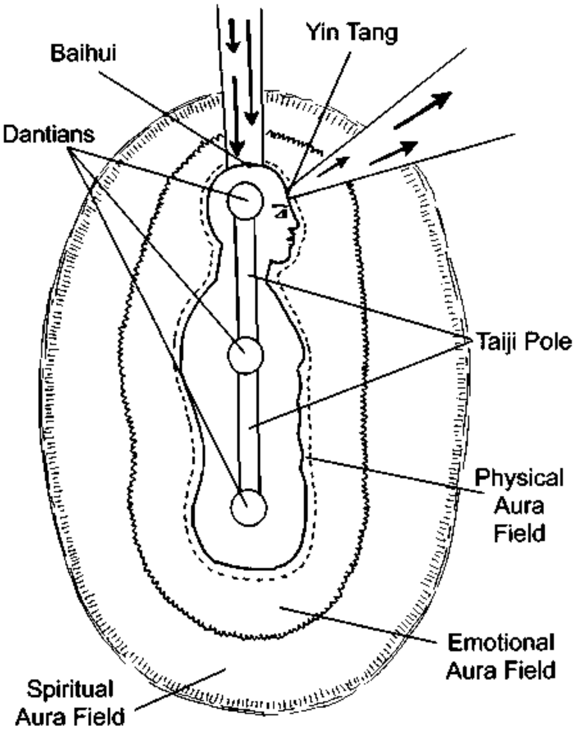

Figure 25.17. The Qigong doctor absorbs divine Qi through the Baihui area, filling the center Taiji Pole, then projects the divine energy out the body through the Yin Tang or Third Eye area.

- The patient sits in front of the doctor. After connecting with the divine, close your eyes and start to scan the patient's external and internal energetic body. Try to feel the patient's different shades of light and different color spectrums, as well as the patient's temperature, projections, resonant vibrational essence, and varying qualities of texture in the patient's electromagnetic fields.
- Next, open your eyes and scan the patient, comparing and contrasting any differences from your initial readings. Continue to practice this modality of distant scanning until your reading and diagnosing abilities become very accurate. Perfecting your long distance scanning abilities may require several months of training.

LONG DISTANCE THERAPY

To dredge the channels, tonify and regulate the patient's energetic fields using long distance energy transformation, the Qigong doctors sometimes practice the two following meditation exercises.

Long Distance Dispersing Meditation

The doctor begins by placing his or her patient in the next room. After the doctor connects with the divine, he or she imagines his or her body surrounded by a ball of brilliant white light. The doctor then extends a stream of this white light energy out from the top of his or her head and onto the top of the patient's head, enveloping the patient's entire body.

As this white light energy descends down the patient's body, the doctor visualizes the patient's diseased energy dispersing and melting away into the ground. The patient's pathogenic energy is extended into the center core of the Earth, where it is transformed by the Earth's center core fire and then recycled back into the patient as pure, refined Earth Qi.

Next, the doctor focuses on specific areas of the patient's body. The doctor visualizes any area which needs specific attention such as dark areas in the patient's body. The doctor imagines these areas becoming brighter until all the dark areas are dispersed. After energizing the patient's diseased area, it is important for the doctor to disconnect from the patient and withdraw his or her projected Shen (Spirit) back into his or her body.

Long Distance Purging Meditation

The doctor begins by sitting on the edge of a chair with both feet firmly on the ground. Next, the doctor brings both palms up in front of his or her body, at throat level (middle fingers pointing straight upwards), and imagines a ball of white light suspended in-between the palms.

The doctor now pulls the Heavenly Qi towards the center of the ball via the fingers. As the ball fills completely, the doctor imagines the Excess Qi beginning to overflow into the center of his or her palms (Laogong points), descending down the arms and spine, and collecting into his or her Kidneys and Lower Dantian area. The goal is to first collect and store Heavenly Qi in the doctor's Lower Dantian and Taiji Pole (Figure 25.18).

To facilitate long distance healing, the doctor imagines the patient in the center of the ball of light. The doctor focuses on the Excess energy no longer overflowing into the palms, but flowing into his or her patients body, cleansing, filling, and healing the disease (Figure 25.19).

FAITH PROJECTION AND MEDICAL QIGONG THERAPY

This technique is a combination of both psychosomatic and visceral healing. It entails speaking to the patient and establishing a curative relationship through belief modalities, similar to those of "faith healing" used in the West.

It is a form of mind projection; the stronger the Qigong doctor's faith and intention, the better the healing will be. It is important for the patient to have an extremely strong faith in the doctor and to feel secure in the healing environment. The stronger the patient's and doctor's faith, the better the healing.

One popular example of this faith projection ability comes from the famous Qigong Master Shen Chang. Master Shen, lecturing at a conference sponsored by the China Ministry of Broadcast, Film and Television (with hundreds of witnesses present), extended Qi into a woman patient (in her 50's) with a three inch tumor on her leg. As Master Shen began his transmission he began to shout "Gone!" Instantly the tumor started to shrink. After shouting "Gone!" two more times, the tumor had completely disappeared. To the astonishment of everyone, Master Shen then shouted "Grow!" and caused the tumor to reappear. After this amazing demonstration, he then removed the tumor, and completely healed the patient.

Perhaps the body manifests what the mind believes and the Heart feels. It then adjusts itself according to the belief structure and forms a new holding pattern (both physically and energetically). Usually, the doctor will say an instructional phrase four times to the patient, intensifying his

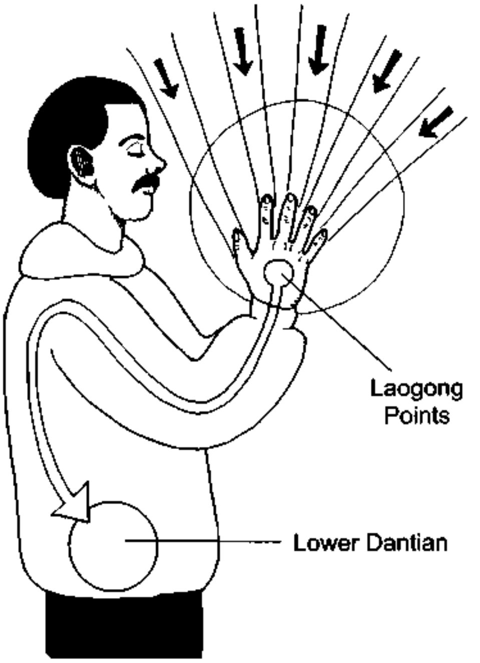

Figure 25.18. The Qigong doctor imagines creating a ball of divine healing white light, suspended between his or her palms.

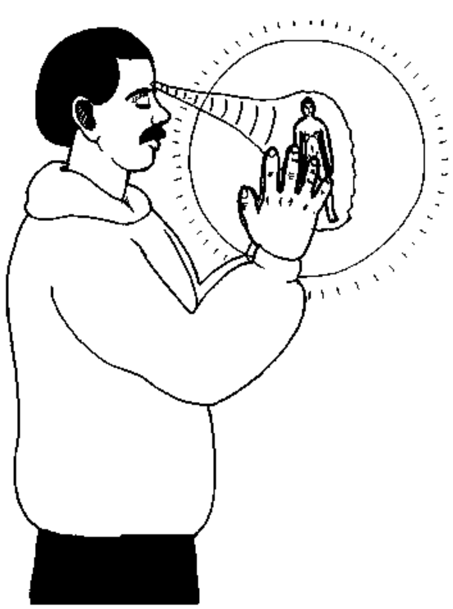

Figure 25.19. The Qigong doctor imagines that his or her patient is positioned in the center of the ball of divine healing light.

or her intention and instruction each time. The doctor's intention begins to reprogram the holding patterns of the tissues by shocking the system into a new and healthier pattern; a kind of "Be Healed Now!" approach. It is important to know that the energetic field surrounding the patient is strongly affected by either positive emotions and support, or negative emotions, suspicion and disbelief.

To illustrate this fact, I would like to recall an incident that took place in a Santa Cruz bible college in the late 1980's. One morning a colleague was lecturing in an auditorium full of 250 young seminary students. Suddenly, the back door of the auditorium swung open, and a young woman carrying a small baby came walking in. She was crying, and asked for help. She told my colleague that "God" had spoken to her in a dream and told her that if she came to this particular bible college and asked for help, that he would heal her child of a serious disease.

My colleague was both startled and shocked. Not knowing what to do, or even how to handle the situation, he quickly turned the meeting over to a local spiritual elder who was monitoring the lecture. This elder, seasoned in these types of spiritual phenomena, again asked the woman what the problem was. After hearing her story a second time, he informed her that while "God" had not spoken to him directly about the situation, he would be obedient to whatever "God" wished.

Next he looked up to the 250 young seminary students sitting in the auditorium and stated, "If anyone here does not believe that God can and will heal this child instantly, please leave now." All but three seminary students left the auditorium. The elder called to the three remaining students to join him with my colleague in praying for the baby. The child was instantly healed (a follow-up visit confirmed this fact), and the mother left, smiling. My colleague, perplexed by the whole event, turned to the elder and inquired as to how he knew what to do. The elder responded, "Even Jesus, who was the greatest healer of all time, couldn't heal in his own home town because of everyone's disbelief. Here in this auditorium it would have been five of us who had the faith in healing, versus 247 who were skeptical and full of doubt. It is better to have a few with great unwavering faith than to have many with little faith at all."

AUXILIARY MEDICAL QIGONG HEALING MODALITIES

The following are auxiliary Qigong modalities used by various Qigong doctors in hospitals and clinics, as well as their personal practices throughout China.

One of the methods used for these Medical Qigong therapies is the projection of Qigong Messages into specific material objects (i.e., a liquid, food, a scroll of calligraphy, a talisman, a cloth, a necklace, etc.). Once the object contains the Qigong doctor's Message it will have a powerful effect on the patient's body. This is a form of "transmission of intention" which is a form a Qi and Shen projection. Qi and Shen projection transcend time and space, lasting an indefinite amount of time depending upon the belief of the Qigong doctor and patient.

Certain Qigong doctors can transfer their energy or "thought message" through works of art. I experienced this technique quite accidently while lecturing at the Third World Conference on Medical Qigong in Beijing, China in 1996. I was in the back of the lecture hall sketching while one speaker was sharing his dissertation. Another Qigong Master seated in front of me noticed my artwork and asked to see it. Curious as to what he would do, I gave him my notebook. He placed his hand above my drawing and began projecting Qi into the picture. Immediately energy flooded my being from head to toe. I was literally vibrating all over, being filled with light and sound resonation (that felt similar to a strong electrical current). Since my original piece of art contained a direct connection to my physical energy, the Qigong doctor was able to extend energy into my body, by simply projecting his Qi into my artwork.

The doctor can transmit healing energy through any object which has had contact with or was created by the patient. This technique is possible because creative artwork, or any object touched or created by hand, is imprinted with the person's energy. This energy contains a unique blend of the individual's encoded energy patterns. This information regarding the individual can be retrieved through psychometry (the act of sensing the thoughts, images and so on, with which the object has been imprinted). A connection to the person who created or held the object can thus be established through the Wuji. This object can then serve as a focusing point to send healing to that person, through the doctor's intent.

Energy, in the form of waves, is expressed in the abstract form of thoughts, accessible to the vibrational medium of space and time. This allows for the energetic transformation of the individual's emotional patterning.

USING MEDICAL QIGONG TO HEAL BY PROJECTING CHINESE CHARACTERS

There are other Qigong doctors who draw specific Chinese characters in the air and send them encapsulated in a bright white light ball of energy into their patient's body to initiate healing (Figure 25.20). This projected energy is created through the doctor's intention and is absorbed into the patient's body, through the patient's receptive attitude and desire to heal.

USING MEDICAL QIGONG WITH SUPPLEMENTARY APPARATUSES

There are certain Qigong doctors who will use specific objects ranging from such supplementary equipment as rubber acupuncture dolls to pillows, in order to perform long distance healing. To utilize these objects for long distance healing, these Qigong doctors place the object, a doll for example, on the table in front of them and begin to draw their patient's Essence (from many miles or even across other countries) into the doll. The doctors then treats the patient by draining off the pathogenic excess from the doll (Figure 25.21). Once the doctors feel that the patient has been sufficiently cleansed and the disease dispersed, they begin to tonify the patient's weak organs. While

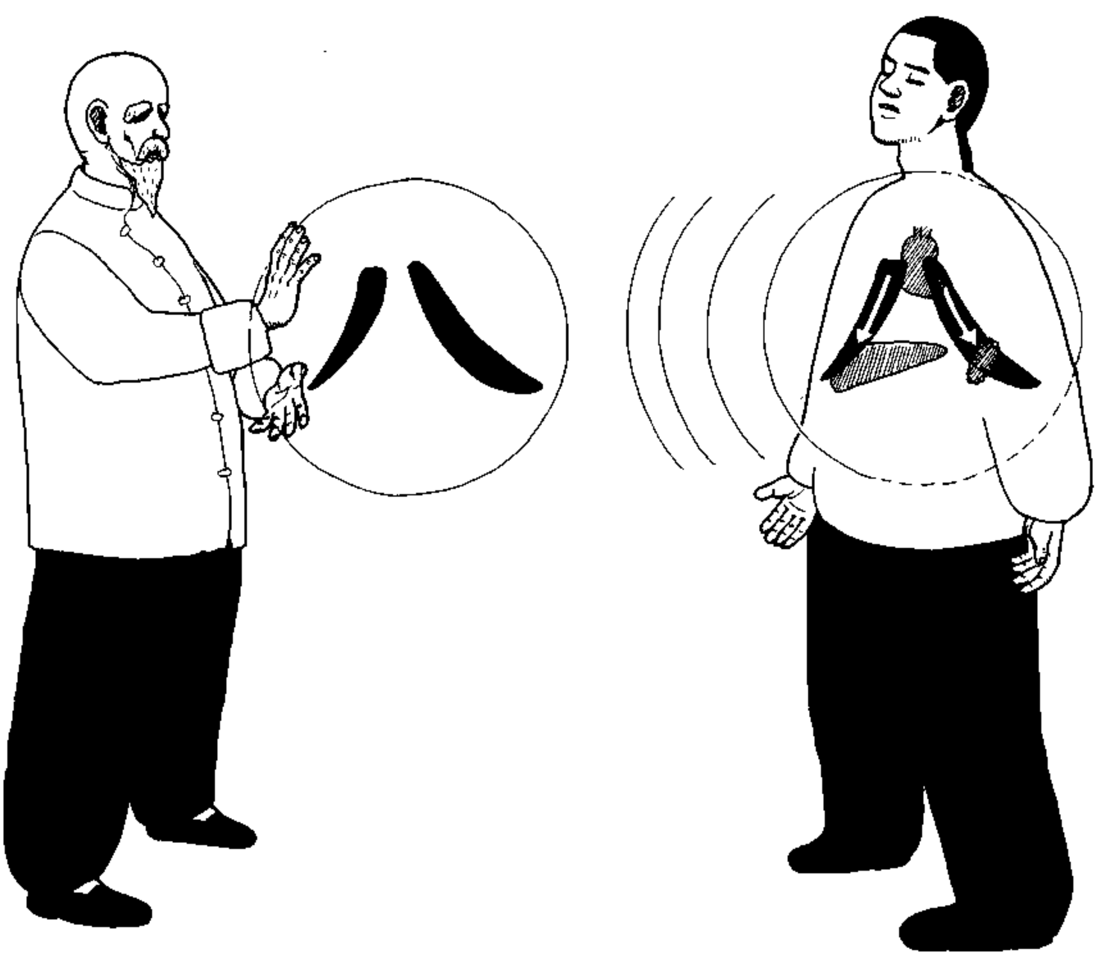

Figure 25.20. Another skill utilized by Qigong doctors, is to create different colored balls of energy to encapsulate specific Chinese characters (or Messages) within them, which are then projected into the patient's body as a form of treatment.

the Qigong doctors work on a patient, you can actually feel hot and cold spots emitting from the rubber acupuncture doll. Several of my students have been able to successfully implement this modality of training and find it fascinating.

USING MEDICAL QIGONG TO ENERGIZE HEALING CLOTHS

Some Qigong doctors extend Qi into pieces of paper or cloth, to be placed on the patient's body to invoke a healing. This is similar to the "prayer

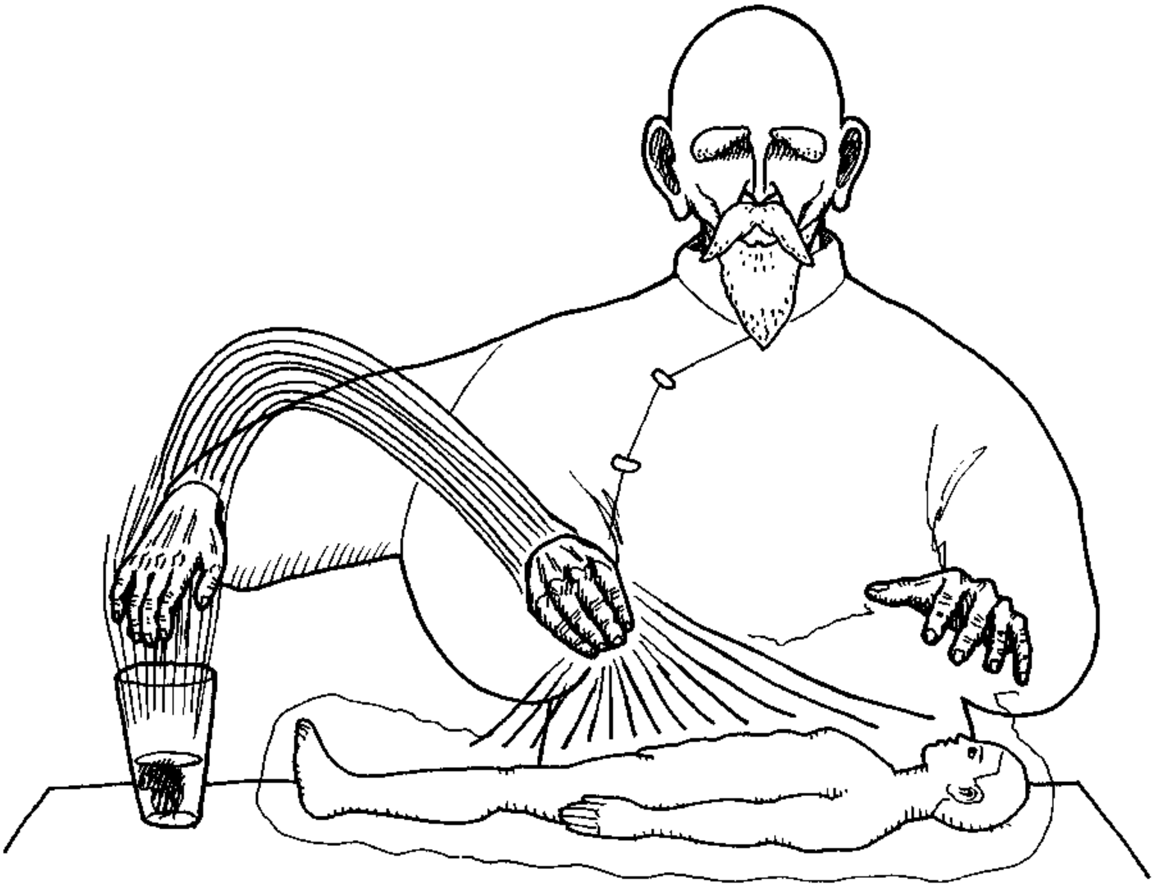

Figure 25.21. Some Qigong doctors use a rubber acupuncture doll (man or woman) as an energetic medium, while treating patients at a distance. The Qigong doctor is removing pathogenic Qi and throwing it into a glass of salt water (used to neutralize the Turbid Qi).

cloths" and other techniques used today in the Pentecostal Churches. The energy stored within the cloth or paper, will last indefinitely, however, the most important component of this type of healing is the patient's belief structure (Figure 25.22-23).

USING MEDICAL QIGONG WITH CRYSTALS

The energy in the human body produces a polarity similar to that present in crystals. Because crystals emit a very strong white light energy, they can be used for energetic surgery, allowing the

Qigong doctor to cut into the patient's energetic field like a laser beam. Crystals can additionally be used for scooping energetic toxins, cleansing and purging the patient's body, and for cutting loose pathogenic Qi within the patient's tissues and external fields of Qi. Crystals can also be used for tonification, by enhancing the doctor's energetic extension and projection abilities.

USING MEDICAL QIGONG TO ENERGIZE LIQUID COMPONENTS

Other Qigong doctors extend their energy into herbal teas, wines, and I.V. fluids, as well as simple components such as water, to energize and utilize these liquids for healing. These doctors send the energized liquid substance to the patient, who ingests the medicine to tonify deficiencies. Water is an excellent repository for electromagnetic energy, and absorbs vibration better then any other substance (Figure 25.24).

USING MEDICAL QIGONG TO ENERGIZE FOOD

There is an ancient Chinese proverb which poses the question, "Is food medicine or medicine food?" Because a basic principle of Traditional Chinese Medicine states that anything ingested is transformed into energy (Gu Qi) by the body, some Qigong doctors extend their energy into the patient's food before it is ingested. This belief is not isolated to the Orient, but is also part of the Western culture, hence the practice of praying and blessing over food before eating it.

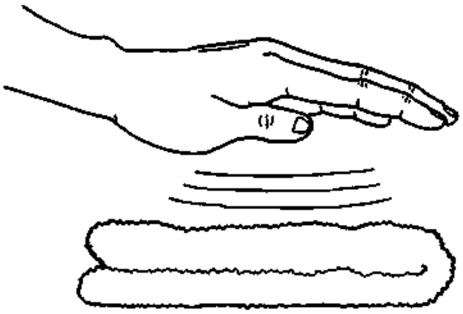

Figure 25.22. Emitting Qi into Cloth

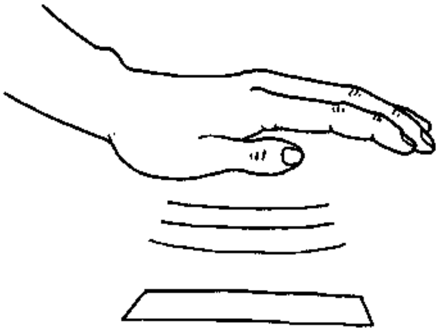

Figure 25.23. Emitting Qi into Paper

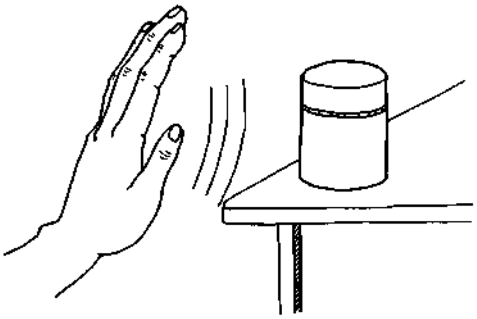

Figure 25.24. Emitting Qi into Liquid

SECTION VI

ESTABLISHING THE MEDICAL QIGONG CLINIC

CHAPTER 26
TREATING THE PATIENT

SETTING A SPACE/TIME FOUNDATION

Before treating patients, and to establish a sacred healing space, the Qigong doctor must first bring his or her true self into what is called the center of space and the center of time before performing the divine hook-up.

This establishes a foundation for treating patients, and allows the doctor to fuse with the patient, while accessing the energies of the divine, as well as the universal and environmental energies. Once this is achieved the doctor can allow him or herself to act as an observer, viewing the patient from a non-judgemental position.

FINDING THE "TRUE SELF"

The Qigong doctor's "true self" is defined as his or her pure innate spiritual existence, void of all ego masks and personalities. The "true self" is spiritually connected to, and rooted in, the energy of the divine. It is the part of the soul that unconditionally accepts self and others, and is more in the role of a compassionate observer than a judge. The "true self" allows the doctor to observe the patient's condition (spiritual, emotional, mental, energetic, and physical), void of any personal emotional biases. The "true self" is a state where unconditional love and acceptance is fostered and supported by the divine.

To discover the "true self", the Qigong doctor is trained through special Shengong (Spirit skill) meditations. One such meditation asks the question "Who are You?" This is repeated over and over until the doctor realizes and experiences his or her "true self". Who you are is not determined through expressing a correct answer, but by actually experiencing the spiritual quest. It is through the realization of this innate truth, which is inherent within one's own enlightenment, that the doctor is freed from the ego (known as the "Death of the Ego Fear"). Once the doctor realizes who he or she truly is, the answer is reflected from the light within the doctor's eyes and not through any specific mental concepts. There is a saying from Lao Zi, written in the *Dao De Jing* (The Way and Its Power) that explains this phenomenon: "The Dao that can be named is not the true Dao."

FINDING THE CENTER OF SPACE

Before treating a patient, the Qigong doctor must also place his or her body, mind, emotions, energy, and spiritual essence into a place of infinite existence called the center of space. The center of space is actually a spiritual and energetic concept, wherein the doctor imagines him or herself as being the actual center of the universe. This concept and image maintains that all energy, light, and energetic fields extend from the middle of the doctor's center core, expanding outward into infinite space (Wuji). This energy continues to extend and envelop the entire universe and all the energetic dimensions beyond.

To access the center of space the doctor must place his or her "true self" into the Taiji Pole, connecting the core self with the divine. It is from the orientation of the core self that the Qigong doctor accesses the deeper, hidden, spiritual, and energetic fields of the patient's body.

FINDING THE CENTER OF TIME

The final image that the Qigong doctor visualizes before the divine hook-up is to place himself or herself into the space known as the center of time. The center of time is the place of "Now." It is a dimension where infinite future is forever flowing into the doctor's center core, while the infinite past is forever rushing away.

To grasp the center of time concept, think of time as a linear progression of light and energy. The concept of future time is rushing directly into the front of your body. Within a split second, the Yang changes into Yin and the future transforms into the past, and continues rushing out the back of your body. You must find the still point, inside of your body, where Yang transforms into Yin, and the future transforms into the past - this is the center of time. From the center of time the Qigong doctor brings the patient into the present moment. This process enables the doctor to access the hidden past memories trapped deep within the patient's tissues.

ESTABLISHING THE MEDICAL QIGONG TREATMENT ROOM

Before receiving patients it is important to establish an environment that is conducive to healing. The density and quality of energy in the healing room is a very critical factor. Filling both the treatment room and clinic with the energy that flows from the Heavens to the Earth sets up an energetic vortex. Through the creation of this energetic vortex the patient's pathogenic factors will automatically flow into the ground once they are dispersed from the patient's body. The more energy that the healing room contains, the more successful the treatment will be. It is important to cleanse the healing room before and after each patient. This cleaning is achieved through the doctor's divine hook-up.

When establishing a treatment room it is also important to have adequate lighting and clean air ventilation. Avoid fluorescent lighting; if this is impossible, then use a full spectrum light.

It is important to also keep the patient's living quarters and bed energetically cleansed, as pathogens have a habit of collecting in areas of sickness and disease. The patient is encouraged to burn salt with a little alcohol to cleanse the treatment room (at least once a day) and fill the room with fresh air and light. Often neglected is the living space of the patient. Energetically cleaning the entire house or apartment often, during and after there has been an illness, allows the toxic energy to be purified from the patient's home.

One example of how energy stagnates is observed in hospital wards. Each ward has an energetic "feeling," associated with the different types of diseases or conditions treated there. The cancer ward, for example, feels entirely different from the maternity ward, emergency room, etc.

The treatment room must also be kept uncluttered, as energy stagnates in cluttered environments. Also, after each treatment, always change the bedding and energetically clean the clinic or treatment room.

WHEN TO AERATE, ILLUMINATE, OR SMUDGE

It is important to daily cleanse the energetic field of the treatment room. This facilities a fresh uncontaminated environment, allowing the next patient to feel safe and protected from any previous contaminants which may have been left from a prior patient. The three most common ways to cleanse the clinic before and after each patient (besides changing the table coverings), are:

1. Aerate by ventilating the clinic, making sure that the air circulation is fresh and clean (free from pathogens, Toxic Qi, dust and pollens, foul orders, smoke, etc.).
2. Illuminate by exposing the clinic to bright sunlight, or divine healing light (through prayer).
3. Smudging is performed by burning fragrant herbs in the form of certain incense (such as, sandalwood, myrrh, frankincense, or sagebrush, etc.). Remember to aerate after smudging.

The Qigong doctor may wish to incorporate either one or several modalities of cleansing. Each particular technique offers its own advantages and disadvantages, depending on the size and location of the clinic and its surrounding environment.

ESTABLISHING A RAPPORT WITH THE PATIENT

It is important to establish a foundation of trust between the doctor and patient before the treatment begins. This trust is enhanced by: Talking with the Patient to Establish Treatment Goals, Explaining and Reaching Agreements on the Treatment Modalities to be Used, Honoring the Patient's Spiritual Belief System, Receiving the Patient's Consent Before Treating, Matching the

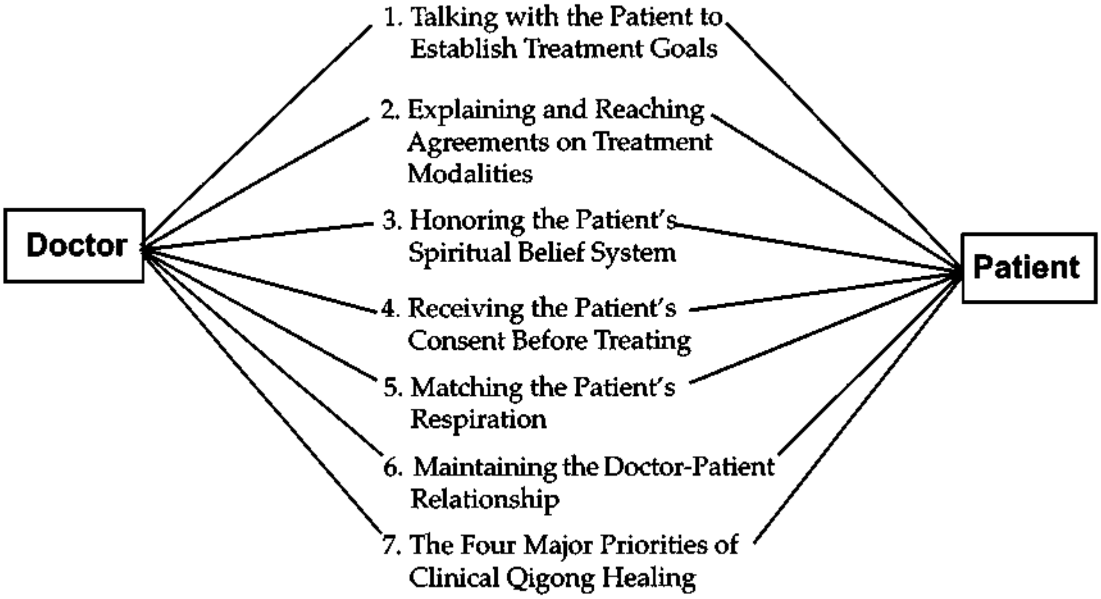
Doctor
1. Talking with the Patient to Establish Treatment Goals
2. Explaining and Reaching Agreements on Treatment Modalities
3. Honoring the Patient's Spiritual Belief System
4. Receiving the Patient's Consent Before Treating
5. Matching the Patient's Respiration
6. Maintaining the Doctor-Patient Relationship
7. The Four Major Priorities of Clinical Qigong Healing
Patient

Figure 26.1. Establishing a rapport with the patient begins with trust.

Patient's Respiration, Maintaining the Doctor-Patient Relationship, and The Four Major Priorities of Clinical Qigong Healing (Figure 26.1).

TALKING WITH THE PATIENT TO ESTABLISH TREATMENT GOALS

After completing the diagnosis, establishing the treatment principle, and determining the treatment strategy, the Qigong doctor discusses with the patient the treatment goal (i.e., which internal organs, channels, and tissue areas are affecting the primary cause of the patient's disease, and how to address it). The treatment areas are specifically chosen to facilitate a rapid affect on the functioning process of the patient's disease. The doctor explains the intended goal of using the purging, tonifying, or regulating techniques, and how they will affect the patient's body.

EXPLAINING AND REACHING AGREEMENTS ON THE TREATMENT MODALITIES TO BE USED

The Qigong doctor explains, in clear, comprehensive terminology, what he of she has determined as to the best way of approaching and rectifying the patient's condition (e.g., having the patient in a lying, sitting or standing posture), and what treatment modalities the doctor will be using (i.e., through Qi Emission, Qigong Massage, Sound Therapy, Light Therapy, etc.).

The more the patient understands the doctor's approach in addressing the condition, the greater the likelihood of the patient relaxing during the treatment. This explanation also assists the doctor and patient in developing trust, which will encourage the patient when practicing the homework prescriptions, or in receiving suggestions on life-style modifications.

HONORING THE PATIENT'S SPIRITUAL BELIEF SYSTEM

When treating patients, it is important that the Qigong doctor not become distracted by the illusion of solid tissue matter, remembering that we are spirit, and that healing begins in this fifth dimension. The primary belief in Medical Qigong therapy is that tangible and intangible energy, through faith, will merge at the highest level of cultivation. Prayer is an invocation that extends the mind, emotion, and spirit to the divine, while meditation is a state of mental, emotional, and spiritual receptivity for divine inspiration, guidance, and manifestation. Both actions (prayer and meditation) connect the doctor and patient to the divine to bring patients back into awareness of their body and to initiate divine healing.

In China, Qigong doctors have a saying, "train your virtue and morality to increase your energy, allow your Spirit to lead your Qi and the Qi will flow pure." Meaning, the stronger the individual's virtue has been developed, the more powerful his or her spiritual intention will flow, and Qi move.

The healing modality choosen by the doctor must be congruent with the patient's spiritual belief structure. The patient must be connected to his or her own unique interpretation of universal consciousness (God) in order for divine healing to take place.

When the patient disbelieves in a Higher Power – be it God or a universal consciousness – the doctor can still help the patient heal, providing the patient has established sufficient trust and hope in the doctor. If the patient feels understood and listened to by the doctor, this can empower the patient to heal, regardless of his or her own conscious belief system.

If the doctor, on the other hand, does not believe that a true healing can take place unless the patient believes in a Higher Power or God - then no healing will take place, for the doctor will lack confidence, hope, and trust. The therapeutic bond between the doctor and patient will be broken and replaced by a battle of wills.

For patients who lack faith, applying a hands on Qigong technique may be more effective. Gradually, over the course of time, the doctor may incorporate other healing modalities as greater trust is built.

The Qigong doctor should always begin treatment in harmony with the patient's spiritual belief system (that may change in time), through touch, verbal conversation and unconditional love (empathy). The doctor should also communicate mind to mind, for this will facilitate trust at the subconscious level.

RECEIVING THE PATIENT'S CONSENT BEFORE TREATING

The treatment modalities should be decided together by the doctor and the patient. It is important for the Qigong doctor to provide the patient with enough information about the type of therapy he or she will receive, so that informed consent can be given for the treatment. This empowers the patient and allows relaxation and trust to occur.

MATCHING THE PATIENT'S RESPIRATION

Another technique to relax the patient is called "matching and pacing." The Qigong doctor matches the patient's breath and breathing patterns, and gradually shifts his or her own breathing to a slower and deeper pace. It is important to note that the patient's physiology affects his or her energy, which in turn affects his or her emotions and respiration. This technique of matching the patient's and doctor's breath and leading it into a state of quiescence is very often used by Medical Qigong doctors to relax their patients. It not only connects the doctor with the patient on an energetic and emotional level, but also provides a strong spiritual connection.

Note: Do not match an asthmatic patient's breathing, as this is disruptive to the flow of Qi within your own body.

MAINTAINING THE DOCTOR-PATIENT RELATIONSHIP

Whenever the doctor begins to judge or formulate negative opinions of a patient, he or she is no longer operating in the divine "present" state of spiritual intuition, but has shifted into the "past" (as all judgements are based on past experiences and emotional encounters). The doctor is then in danger of absorbing the discharged energetic pathogens from the patient, or projecting subconscious biased opinions onto the patient. Whenever the doctor relates, relives, or orients from his or her own personal past (the feelings can be either positive or negative), the doctor is at high risk of absorbing the patient's energetic pathogens.

When the patients come to the doctor, they are giving their consent for the doctor to learn about personal things which the patient has often withheld from their mate, family, or even themselves. They are further giving the doctor permission to utilize their combined energies to affect changes within their physical, mental, emotional, energetic, and spiritual fields and body.

Each treatment will be different for the patient, as each experience will vary according to the patient's present state of mind and emotional spirit. The patient's willingness to allow the Qigong doctor's emitted energy to penetrate deeper into his or her body with each session will depend upon the level of trust established, as well as any outside influences that may have a positive or negative effect on the patient's ability to be open and vulnerable. Patients who experience ridicule or criticism from family and friends for their choice of a non-Western form of treatment, may temporarily shut down their receptivity in self-defence against others. At times, patients may also close-off as a result of an event that has a strong impact on their emotional and spiritual well-being, such as a death of a friend, a physical trauma, or other upsetting news.

The patient always has the final control over the doctor's emitted energy, and can choose to either absorb and utilize the energy, or disperse and sabotage the doctor's treatment. It is for this very reason that the doctor-patient trust factor, as well as the goal orientation, must be securely established before the treatment begins.

Be sure the patient's arms and legs are not crossed, as this can impede, or "short-circuit," any energetic flow, causing Qi obstructions. The patient should close his or her eyes, but should not be allowed to internally guide or direct the Qi unless specifically instructed to do so by the doctor. A patient's unguided imagery can cause the body's energy to increase, alter, or obstruct the flow of the doctor's energy, which may cause Qi deviations or aggravate an existing condition. Usually the patient is given the image of dissolving his or her tissues into space, which encourages relaxation and freedom from distractions. If the patient is lying down, he or she is given certain visualizations, such as floating on a sea of warm water (or basking in the sun on the beach or shore of a lake). These images relax the patient and enable the Qigong doctor to disperse the patient's armoring. Once the patient's "street armor" is down, the doctor will have easier access into the patient's tissues.

After the treatment, the doctor and patient evaluate each session, discussing any and all feelings, observations, transitions, and sensations that both the doctor and patient experienced. Homework, in the form of Medical Qigong prescriptions are then generally given, in addition to follow-up treatments (depending on the patient's case and condition).

THE FOUR MAJOR PRIORITIES OF CLINICAL QIGONG HEALING

According to China's prominent Qigong Grand Master Pang Ming, clinical Qigong healing is based on four basic priorities. In establishing a clinical environment it is important to inform the patient about the structures of these four priorities: A Strong Faith, Group Healing, Medical Qigong Therapy, and Medical Qigong Prescriptions and Exercises.

1. A strong faith that Qi or life-force energy can heal the patient's specific ailment is built up by the patient listening to testimonials. These testimonials are given by other patients who have recovered from the same type of disease, as well as those patients who are experiencing improvement through Medical Qigong therapy.

   Qigong deals directly with the psychological aspects of diseases. It is well known that two side effects of cancer are depression and fear. A very popular theory maintains that depression (being the suppression of anger) can cause cancer. Even before the cancer patients begin their treatment program, their confidence is bolstered by the positive experiences of more advanced students of Qigong, who at one time also suffered from these particular kinds of cancer. In addition, as part of the Qigong exercises, the new patient's mind is focused on pleasant images or memories. All of this creates a positive mental attitude in the patient.

Sometimes patients come to the doctor for healing, yet at a deeper level they feel that they need or deserve the illness. Sometimes the secondary gain of illness (for example, to keep the wife or husband underfoot) outweighs the benefits of healing completely. If a patient is subconsciously unwilling to be healed, his or her attitude or disbelief will sabotage all clinical modalities of healing (both Western and Eastern). You cannot make someone better against his or her own will. Illness, like health, is often a matter of choice.

2. Before a group of patients begin Qigong, the doctor verbally synchronizes the thinking of the group by absorbing Heavenly Qi (the divine hook-up), and bringing it down into the healing energy field which surrounds everyone including the doctor. The healing effect is enhanced because the group is acting as one unit, with one faith.
3. The doctors initiate the actual Qigong healing by bringing divine healing energy to each individual patient.
4. The patients are taught how to easily follow Medical Qigong exercises and meditations, practicing them over and over again.

When a patient enters a Qigong hospital in China, they are diagnosed by a doctor and then assigned a specific Qigong class for a twenty-four day treatment period. The patient spends eight hours a day practicing Qigong without television, newspaper, or telephone. This is required to help the patient avoid regression into a previously established belief structure or subconscious pattern. Those who are able, practice Qigong from a standing posture. Other patients with limited mobility practice Qigong sitting or while lying in bed. After the twenty-four day treatment program is over, the patient is again diagnosed by a Qigong doctor. The diagnoses is classified into one of four categories for statistical purposes.

1. The patient has been cured: symptoms have disappeared and appropriate instruments (e.g., EKG, ultrasound, etc.) register normal tissue.
2. The treatment has been very effective: symptoms have almost disappeared and allopathic tests indicate great improvement.
3. The treatment has been effective, in that the patient is now able to eat, sleep, and feel good.
4. The treatment has been noneffective: there has been no change, or the condition has worsened.

PREPARING THE SELF TO TREAT THE PATIENT

This section prepares the Medical Qigong doctor to treat patients. The "preparation" sets the energetic ambiance for the clinical treatment and secures the foundation for dispersing pathogens. The preparation includes: The One through Ten Meditation, Creating a Sacred Healing Space, The Divine Hook-Up, Rooting to Prevent Energetic Shock, and The Energetic Circle Drawing Pattern (Figure 26.2).

THE ONE THROUGH TEN MEDITATION

The following is called the One through Ten Meditation and is used to prepare the Qigong doctor for clinical work. It allows the Qigong doctor to relax, sink, and root the mind. This meditation allows the doctor to extend his or her energetic field before connecting with the divine. This meditation is used to create a sacred healing space and is taught in a rhyming one through ten counting format, described as follows.

1. One Is Fun, therefore take pleasure in your work. Imagine a smile melting down the front of your body like warm oil, melting all stress and tension into the Earth. A second smile melts down the back of your body, dissolving all stress and tension down into the Earth. A third and final smile melts down the center of your head, relaxing all thoughts, judgments and feelings, and flows down your shoulders and arms and out your hands. The melting continues to flow down the center of your torso, like warm oil, melting all emotional blocks and tension, then flows down your legs and into the ground. At this point your mind and body should feel content, relaxed, and peaceful.
2. Two Is Shoe, thus imagine your feet melting into the Earth, like ice melting on a hot side-

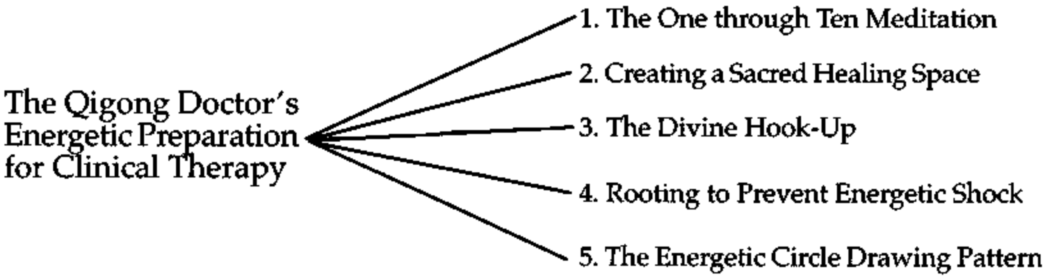
The Qigong Doctor's Energetic Preparation for Clinical Therapy
1. The One through Ten Meditation
2. Creating a Sacred Healing Space
3. The Divine Hook-Up
4. Rooting to Prevent Energetic Shock
5. The Energetic Circle Drawing Pattern

Figure 26.2. The Qigong Doctor's Preparation for Treating the Patient

walk. The energy of the feet fuses with the Earth, melting in five directions (forward, backward, right, left, and down). This connects your mind and body with Earth energy.

3. Three Is Tree, therefore feel yourself extending tree-roots twice your body's height deep into the ground. The roots expand in five directions (forward, backward, right, left, and down), securing your rooting and connection with the Earth energy.
4. Four Is Core, thus imagine your roots extending deep into the core of the planet. The Earth energy flows into these roots, ascending up the legs, spine, over your head, down the chest, and enters your Lower Dantian. This action allows you to draw Earth energy into your body and circulate it through the Microcosmic Orbit. A solid connection to the Earth roots your physical, mental, emotional, and spiritual energies, and fills up your Lower Dantian with Earth energy.
5. Five Is Alive, therefore feel the Earth energy building up within your Lower Dantian. The energy increases to the point that your Dantian can no longer contain its power. It rushes up the center of your chest like a mighty river, dividing into two streams which flow out each palm into the earth. This action releases and emits the energy stored within your Lower Dantian.
6. Six Is Thick, thus imagine the room or environment where you are standing being filled with energy. This energy becomes so thick, it is as if you are submerged under water. This fuses the environmental energy with your energy, and the Earth's energy into one dynamic field of energy.
7. Seven Is Heaven, therefore allow your Baihui point to open up and draw in the divine healing white light into your Upper Dantian. Let the divine light fill your head and the entire Upper Dantian area till it floods and shines down from the Upper Dantian into your chest, filling up the entire Middle Dantian. Finally, the light shines down from the Middle Dantian filling up the lower abdominal area and the entire Lower Dantian. This light connects you with the Heavenly healing white light energy from the divine, filling up all three Dantians and your Taiji Pole. Thus Heavenly and Earthly energies fuse within your body.
8. Eight Is Open the Gates, therefore imagine opening every pore on your surface tissues, and begin to draw the environmental energy into your Taiji Pole upon inhalation. Upon exhalation imagine your center core vibrating and glowing like a neon light. This visualization energizes your center core, harmonizes your energy and breath, and prepares you for Qi and Shen (energy and light) projection.
9. Nine Is Shine, therefore imagine that your center core becomes completely full of energy and overflows by shining (from the Taiji Pole outward) through your pores until it fills up the entire room, like a bright phosphorus flare. This visualization fuses Heaven Qi, Earth Qi, and Man's Qi with the external clinical environment.

10. Ten Is Begin, with all things now in order, you are physically, mentally, emotionally, energetically, and spiritually in harmony, with your body, the Heavenly energy, and Earthly energy. You can now begin treating patients.

CREATING A SACRED HEALING SPACE

After completing the preparation, the Qigong doctor creates a sacred healing space in which to treat patients. Creating the sacred healing space allows the doctor and patients to interact in a safe environment without outside distractions.

The Qigong clinic should be a place of complete sanctuary, a consecrated place of refuge, where the patient feels safe, warm, and cared for (spiritually, emotionally, mentally, energetically, and physically). The patient should experience the peaceful ambiance of the treatment room as being enveloped in unconditional love and security.

To create the sacred healing space, the Qigong doctor performs the divine hook-up. Only after completing the divine hook-up should the Qigong doctor treat his or her patients.

THE DIVINE HOOK-UP

Once the doctor has established energetic grounding, the next step in establishing a clinical environment is to connect with his or her spiritual center through the divine hook-up. From this spiritual center the Qigong doctor observes the patient, void of personal thoughts, feelings, and judgments. This spiritual state of open receptivity allows patients to feel secure and to release their guarded emotions. Lacking a safe environment, patients may feel the need to repress emotions or to be distracted from experiencing and letting go of the original trauma.

The Qigong doctor, once healing has begun, enters into a state of "oneness" with the patient and the divine, in which the doctor's will and intention envelops and fuses together with the divine will and wisdom. Both doctor and patient begin to communicate through deep subconscious interactions. The doctor's success or effectiveness as a Qigong healer depends on a willingness and ability to connect with the patient and the divine healing light. Drawing from this interpersonal communication with the patient, and the deep spiritual connection with the divine, the doctor begins the treatment.

According to Dr. Meng Xian Tong of the Beijing Chengjian Integrated Traditional Chinese Medicine and Western Medicine Experts Clinic, when Qigong doctors perform the divine hook-up, they increase their own body's Righteous Qi, which then extends the energetic Message encoded in their Qi into the patient's body. The doctor's Righteous Qi is then absorbed by the patient's body and fused together with his or her Qi. This fusion empowers the patient to heal his or her disease.

Before contact with any patient the doctor first connects with the divine to purify and stabilize, not only him or herself, but also the treatment room (with a constant flow of Heavenly Qi pouring into the room). This divine hook-up enables the Qigong doctor to not only establish a sacred healing space, but also to access both spiritual insights and divine healing power.

ROOTING THE DIVINE HEALING LIGHT

Rooting the divine healing light is accomplished by visualizing a beam of healing light (the doctor can also choose specific colors) streaming into his or her body via the Baihui and Sishencong points (at the top of the head), and entering into the Taiji Pole. The doctor holds this frequency until the divine healing light becomes rooted or "anchored" within the Lower Dantian.

Next, the doctor brings the healing light vibration up the Taiji Pole to energize each Dantian until the healing light infuses and harmonizes all of the body's tissues. As the doctor feels the healing light reach the Middle Dantian, he or she visualizes the light intertwining with the "bridge of light," which connects the doctor's Heart (Middle Dantian), throat and Yintang (Upper Dantian) areas together. This energetic connection fuses at the Upper Dantian, becoming one unified cord. The healing light is then stored within the doctor's body and released during Qi emission.

USING THREE DIVINE INVOCATIONS

To assist the doctor in rooting the divine healing light, I encourage my students to begin each treatment with a small prayer and three invocations, for example:

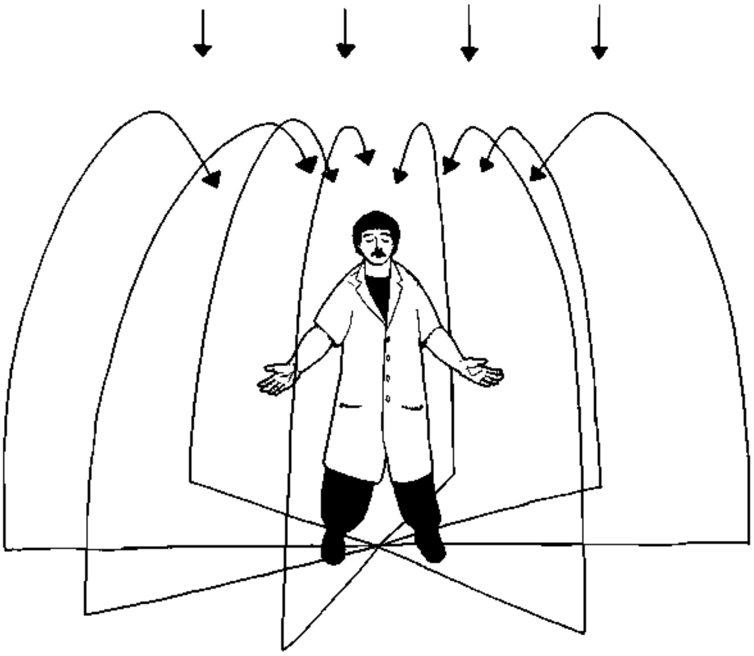

Figure 26.3. As the divine healing Qi absorbs into his or her body, the Qigong doctor expands this field to envelop the entire treatment room.

1. "Lord God, fill me with your Holy Spirit and divine healing power. Increase my skill, power, knowledge, and wisdom and use me for Your Glory." The first invocation connects the doctor to the healing energy of the divine and roots the doctor deep into the Earth. This prevents the patient's pathogenic Qi from entering into and contaminating the doctor's body.
2. "Fill this room with your Holy presence and healing virtue, and let your Glory shine through me into this room." This second invocation connects the doctor and the treatment room to the divine and secures a sacred healing space (or sanctuary) for the patient. It also establishes an expansive divine energetic field by causing the doctor's Wei Qi field to overflow with divine healing light, filling the treatment room. This divine energetic field produces a dynamic field of healing energy (Figure 26.3).
3. "I lift (the doctor will mention the patient's name) before Your Holy Presence, and ask that you may heal these wounds and free him or her from these chains of despair, agony and disease. Thank you, God. Amen." After this final invocation, the doctor envelops the patient with an energetic bubble of light and love.

Next, the doctor roots the patient's energetic bubble deep into the Earth, and creates an energetic vortex underneath the patient. The energetic vortex will spiral clockwise, deep into the center of the Earth. This energetic rooting technique is practiced to purge and drain the patient's Toxic Qi into the Earth, where it is cleansed and then recycled back into the patient's body.

ROOTING TO PREVENT ENERGETIC SHOCK

Rooting (sometimes called "energetic grounding") is an important technique used by the Qigong doctors to establish and maintain a firm physical, mental, emotional, energetic, and spiritual connection to the Earth. Connecting to the Earth is essential to prevent energetic shock (a jolt of energy which flows from the patient into the doctor's body), which can result in the doctor ab-

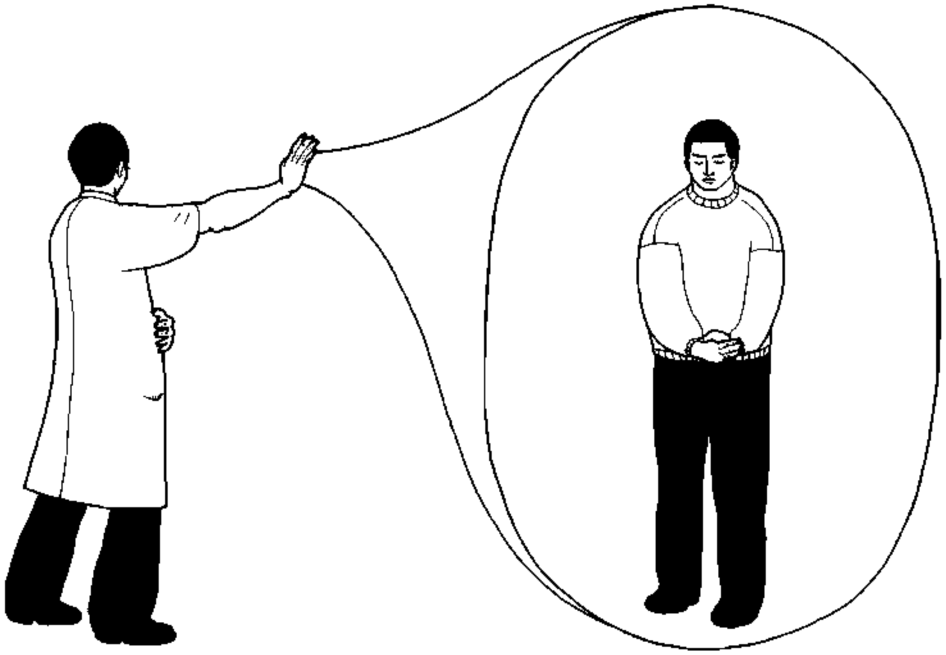

Figure 26.4. The Qigong doctor envelops the patient in an energetic bubble (or healing circle) before beginning the treatment.

sorbing the patient's energetic pathogenic disease. To prevent energetic shock from happening, Qigong doctors go through several steps to prepare themselves and the clinical environment before treating patients. Energetic rooting is utilized in all phases of meditation and allows doctors to maintain a strong physical connection to the Earth. It is especially used when the doctor is absorbing and emitting environmental and universal energy.

- From a Wuji posture, imagine an energetic cord extending from your Lower Dantian, down the Taiji Pole and out the Huiyin point (at the base of the spine) deep into the center of the Earth. This energetic cord will root your energy into the Earth. It is also affected and maintained by your will and intention. The energetic cord follows the doctor's movements smoothly and easily.
- If the patient is lying, sitting, or standing in front of you, extend your energy to envelop the patient within an energetic cocoon called "energetic circle drawing."
- Imagine extending an energetic cord from the patient's Lower Dantian through this energetic cocoon, to root the patient's body deep into the Earth. This process energetically grounds the patient's body, mind, emotion, and spirit.

THE ENERGETIC CIRCLE DRAWING PATTERN

Once the divine hook-up is completed, the doctor grounds the patient by drawing an energetic circle around the patient's body. This is used to charge the patient's energy field. This energetic projection also roots the doctor's intention around the patient and strengthens the doctor's concentration. This circular field changes the patient's static field to a divine dynamic field and enables the Qigong doctor's energy to enter the patient's body with ease.

The enveloping of the patient's body allows the patient's body and energetic fields to become even more accessible to the Qigong doctor, while keeping the patient's pathogens isolated from the doctor's body (Figure 26.4). As the doctor steps into the energetic bubble, he or she is able to treat the patient's tissues, feeling absorbed and enveloped within the Messages expressed from the patient's energetic fields.

The energetic bubble is created by emitting Qi towards the patient, and enveloping and circle wrapping the patient in a clockwise energetic rotation. This energetic bubble forms a shield, maintaining protection for the patient and the doctor. Next, the doctor creates an anchoring line to root the patient's energetic field deep into the ground. Finally, the doctor creates an energetic vortex (spiraling in a clockwise direction) underneath the patient, to pull any dispersed pathogens deep into the Earth (this helps when the doctor is dredging and purging the patient's energetic fields).

USING THE ENERGETIC CIRCLE DRAWING PATTERN TO PURGE PATHOGENS

The energetic circle drawing pattern helps to facilitate energetic cleansing. Energetic purging facilitates the absorption of Clear Qi and also reduces the risk of contamination within the body's channels and collaterals by expelling energetic toxins and other pathogenic factors.

After creating the energetic circle, the doctor begins using dredging and purging techniques to expel pathogens from the patient's body. Next, the doctor treats the patient with tonification and regulating techniques, ending the treatment by closing the diseased area and sealing the patient's three energetic fields (three Wei Qi fields). This ending technique will securely close the energetic holes in the patient's outer field of Wei Qi (see Chapter 13). Without this sealing, the healing process is hindered due to the patient's Qi having a tendency to leak or disperse itself (even after tonification). Once the sealing is completed, any Excess Qi will be evenly distributed throughout the patient's body, establishing a Yin and Yang balance within the visceral organs.

USING THE ENERGETIC CIRCLE DRAWING PATTERN TO PURGE TOXIC HEAT

Before treating patients who have received radiation or chemotherapy, the Qigong doctor should wrap the patient with an energetic bubble and then purge the Toxic Heat from the patient's body. The Toxic Heat is dispersed into the energetic vortex created under the patient's body, and then released deep into the Earth. The Qigong doctor can then step into the energetic bubble and begin combing, dredging and purging the patient's tissues.

USING THE ENERGETIC CIRCLE DRAWING PATTERN TO PURGE TOXIC COLD

When using the circle drawing method to purge pathogenic Cold, the doctor stands outside of the energetic bubble, and extends energy deep into the patient's Taiji Pole. Once the doctor feels the patient's center core Qi, he or she draws the pathogenic Cold or Hot energy out from the patient's tissues (in all directions), filling the energetic bubble. The doctor, standing outside the energetic bubble, observing the patient's energetic field, waits for an energetic shift which is immediately followed by Toxic Cold (or Hot) Qi filling up the energetic bubble. Once this occurs, the Qigong doctor must drain the pathogenic Cold (or Hot) Qi down into the Earth to rid the patient of the Toxic Qi.

TREATING THE PATIENT

After the Qigong doctor has "circle wrapped" the patient, the treatment begins. First, the doctor dredges the patient's external energetic fields, channels and collaterals (from head to feet). The doctor senses the patient's Qi with his or her hands, changing the depth of penetration of the touch by extending intention deeper into the patient's tissues. Once the patient's pathogenic Qi has been reached, the doctor feels various energetic sensations (vibration, warmth, thickness, etc.) and removes any pathogenic energy from around the patient's body.

The doctor continues the treatment by extending his or her intention deep into the patient's body, purging and dispersing the toxic pathogens from out of the patient's internal organs and channels. It is usually during this deep internal purging that the patient experiences profound emotional discharges, observes flashes of light, patterns of colors, see visions, or relives deep seated past emotional traumas. This purging releases toxic emotions which are generally the cause or contributing factor to the patient's disease.

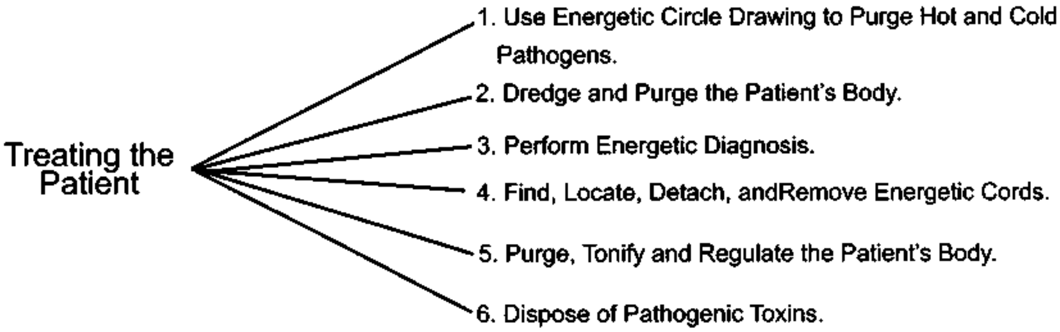

Treating the Patient

1. Use Energetic Circle Drawing to Purge Hot and Cold Pathogens.
2. Dredge and Purge the Patient's Body.
3. Perform Energetic Diagnosis.
4. Find, Locate, Detach, andRemove Energetic Cords.
5. Purge, Tonify and Regulate the Patient's Body.
6. Dispose of Pathogenic Toxins.

Figure 26.5. The Different Steps Involved in Treating a Patient

After the patient has been sufficiently purged and cleansed of pathogenic energy, the Qigong doctor begins to tonify the patient's internal organs, channels, collaterals, and Righteous Qi. This is done to help the patient build a stronger immune system, and fight or avoid any future complications caused by an Excess or Deficient internal organ condition (Figure 26.5).

DREDGING AND PURGING TECHNIQUES

When treating a patient, the doctor purges an internal organ and dredges a channel. The technique of dredging is a particular type of purging method. The energetic circle drawing techniques are always performed with dredging techniques (using the pulling and stroking methods of energy manipulation), to drain the energy into the Earth. These techniques can be initiated with the patients either sitting, standing, or lying.

It should also be noted that the diseased area of a patient's body must be thoroughly dredged and prepared before any kind of tonification or energizing is performed. This technique is similar to a Western doctor cleaning a wound before operating.

If the Qigong doctor does not dredge and purge the pathogens from the patient's body before treating, the patient can suffer adverse reactions when the doctor's clean energy fuses with the patient's Turbid Qi.

DREDGING AND PURGING THE PATIENT

There are two stages of clinical purgation which are described as follows.

1. The doctor dredges and purges the surface pathogens from the patient's body. The doctor moves to the right lateral side of the patient's body and energetically combs the patient's body from head to toes using the Tiger Kneading and/or Extended Fan Palm techniques (in this particular example the patient is lying supine on a treatment table). The combing technique is performed to energetically evaluate the patient's reaction to the doctor's emitted Qi, while dredging the patient's three energetic fields of Wei Qi.
2. The doctor purges from the bottom of the patient's feet. The Qigong doctor stands by the patient's feet and continues to dredge the patient's body by energetically combing the patient from head to toes. This action further purges the patient's body of pathogenic energy.

ENERGETIC DIAGNOSIS

After performing the clinical purgation, the Qigong doctor begins the energetic diagnosis. Some examples of energetic diagnosis are described as follows.

- The doctor can begin energetic diagnosis by first connecting with patient's heels (either by touching or not touching). This particular method and approach to diagnosing allows the Qigong doctor to energetically connect with the patient's inner-fascia. It is through this connection that the Qigong doctor feels the energetic flow of Qi moving through the patient's internal organs. By extending energy from the patient's heels up to the head (like a rippling ocean wave or sonar), the doctor can "listen" for obstructions or deviations of Qi flow.

- Next, the doctor can use the Flat Palm diagnosing technique to scan the patient's entire body for any Excess or Deficient areas within the patient's Wei Qi fields. This second type of diagnosis serves to verify the obstructions or deviations of Qi flow that the doctor discovered while first connecting with the patient's heels.

PURGING, TONIFYING AND REGULATING THE PATIENT'S BODY

After diagnosing the patient's body, the Qigong Doctor purges, tonifies, or regulates the energetic fields, channels, tissues, organs, and organ systems, according to the etiology and prognosis of the disease.

While dredging and purging the patient's body, the Qigong doctor may discover several energetic cords attached to the patient's tissues and internal organs. These energetic cords may sometimes contribute to the cause of the disease.

UNDERSTANDING THE BODY'S ENERGETIC CORDS

The human body attaches, extends, absorbs, and collects energetic cords. The size and shape of these cords can be big, small, thick, or thin, depending on the amount of emotion invested in their attachment. These cords are located within the Three Dantians and Chakra Gates, and form the spiritual, emotional, mental, and physical connection of the patient to the outside world. Generally, what makes the patient sick is not his or her genetic predisposition, but the negative, co-dependent bonds the patient creates in his or her relationship with others. It is the spiritual, emotional, mental, and physical reactions to these relationships that either promote health or disease. Energetic cords are also formed towards objects, pets and places that have special sentimental or emotional value (be they positive or negative) to the patient. Energetic cords are also sometimes used to feed the formation of the preexisting energetic clusters and cysts, which later transform into the patients' diseases.

Energetic cords are the psychophysical connection between what the patient feels, believes and knows to be true; they represent the connection to the patient's personal reality. This does not mean that the patient's reality is valid, only that the patient has come to believe it to be true. By removing these cords an energetic shift is initiated, allowing the patient to reprogram his or her feelings, thinking processes, and behavior, by becoming proactive instead of reactive. Energetic cords are full of light and energy and can be categorized into two distinct divisions: prenatal (congenital) and postnatal (acquired) energetic cords.

- Prenatal Energetic Cords originate in the fetus, when the baby is still in the womb. They form the strong congenital attachment and permanent bond the child has with the mother and father. Because of the mother's energetic/spiritual connection with her egg, and the father's energetic/spiritual connection with his sperm, when the egg and sperm unite, the energetic cords of each parent fuse together with the universal and environmental energy creating the energetic structure of the child. It is through the parents' energetic cords that inherited talents and defects are passed along to the infant.
- Postnatal Energetic Cords are acquired after the baby is born through interaction with siblings, relatives, and other individuals (in some cases animals or places are involved). These cords will either form the emotional attachments which can help the child through life by being a means of support, or set the foundation for disease by depleting the child of his or her life-force energy. They represent the emotional bonding (positive or destructive) inherent within each one of us.

In a clinical environment, the attachment of these cords (prenatal and postnatal) will affect the healing relationship of both the patient and the doctor.

- The Patient's Energetic Cords are still connected to the patient after the pathogenic Qi has been removed from the patient's body. These small cords are the energetic attachments to part of the patient's old belief structure (which created the disease or emotional attachment). Although there is still an energetic connection to the departed pathogenic Qi, these energetic cords will soon disperse once the patient's Creative Subconscious is able to reprogram the individual's belief structures. Only then can the physical body, as well as the mental, emotional and spiritual energetic Wei Qi fields maintain the new energetic pattern. This new belief structure creates an energetic transformation, allowing the patient to release his or her hold on to the painful past and begin to heal. The patient is encouraged to emotionally disconnect from old emotional belief structures through creative visualization, and by affirming the new belief structure. Through affirmation and positive emotional, spiritual visualizations, the patient is able to "sever" any energetic cords that connect to the pain and disease.

- Energetic Cords between the doctor and patient are formed during healing sessions. These new energetic cords are passed back and forth between the Qigong doctor and the patient. Sometimes this occurs without either party being aware of this energetic connection. Usually, as the doctor steps away from the patient and reconnects with the divine, an energetic purging is automatically initiated, freeing the doctor from the attachment to the patient's energetic cords. If, however, there has been an emotional connection between the doctor and the patient, these cords can remain attached even when the patient leaves the office.

  The doctor should make every effort to disconnect from the emotional attachment of the patient's cords, otherwise the increased energetic pull on the doctor's center will affect the doctor on every level. Staying connected to the patient through an energetic cord can cause not only emotional, but physical and spiritual influences on the doctor. Any remaining cord can be used by the patient to unconsciously project thoughts and feelings onto the doctor. This can result in the doctor being depleted of energy either through direct draining of his or her Qi, or through obsessing about the patient's well-being, long after the treatment has ended. The drain of the energetic cord can lead to either clinical "burnout" from absorbing excessive negative emotions, or deep emotional bonding, which can lead to unethical sexual indiscretions.

FINDING, DETACHING, AND REMOVING ENERGETIC CORDS

Energetic cords manifest through their effects on behavior and emotional responses of the patient. They are relationship based patterns, located within the patient's tissues. Energetic based patterns can evoke certain reoccurring emotional responses, such as attractions, addictions, etc.

1. To assist the patients in removing the toxic energetic cords, the Qigong doctor must first locate the cords.
   a. Through the emission of Qi, the doctor may elicit (from the patient) either an attraction or repulsion response. This response occurs when certain memories are stimulated within the patient's body, indicating that there is an energetic cord attached within the tissues.
   - The attraction response causes the patients to experience deep emotional release through laughing, crying, shouting, etc.
   - The repulsion response causes a physical jerking of the patient's muscles, anywhere within the patient's body (observed by the wincing and grimacing of the patient's face). Often the patient will "Shen-out" (emotionally and spiritually leave their body) to avoid feeling the painful surfacing emotions.

   b. The doctor locates the origin and entry point of the energetic cord through Flat Palm detection. The root of an energetic cord acts as an open receptor to a specific psychophysical problem that the patient is experiencing. When stimulated, these cords may manifest physical symptoms such as nausea, hot flashes, vertigo, dizziness, headache, and muscle contractions. The patient may also experience emotional reactions such as momentary panic, fear and so on.

2. Before detaching the energetic cords, the doctor requests the patient's permission to do so either verbally, or silently through mind to mind communication.
3. To remove the energetic cord, the doctor dredges and purges the patient's body, while silently communicating through the Creative Subconscious mind. The doctor encourages the patient to let go of, and fully release, the toxic cords along with the memories that sustain them. To ensure against their recreation, the doctor discusses with the patient any images, emotions or sensations that surfaced during the removal of these cords; through this process the doctor brings the patient's conscious awareness of the fears, pain and major unresolved issues that resulted in the disease. The patient thus gains a clear identification of his or her traumas to facilitate the reprogramming of his or her dysfunctional relationships through affirmations and creative visualizations.

If the patient is not ready for such a transition (feeling scared, unprepared, undeserving, etc.), his or her subconscious mind will sabotage any effort made in that direction, and the energetic cord, emotional cluster and all energetic attachments to the disease will be recreated and reintegrated. It is important in cases such as these, that the Qigong doctor refer the patient out to a psychotherapist for additional support.

In Western society, the Twelve Step Programs (such as Alcoholics Anonymous) are but one example of a pattern based reprogramming of energetic cord attachments. Alcohol and Narcotics Anonymous groups, for example, help detach energetic cords toward illegal substances, whereas, CODA meetings help reprogram the energetic cord attachments to co-dependant relationships. In a codependent relationship, both parties rely too heavily on one another, instead of being interdependent. One person usually dominates the other, as both parties base their self-esteem on the partner's approval, or disapproval, thus avoiding the personal responsibility of their own actions.

Energetic reprogramming encourages the participants to take responsibility for filling and replacing the energetic gap that the toxic cord has left with a connection and attachment to the divine, or "Higher Power."

DISPOSING OF PATHOGENIC TOXINS

One problem a doctor may confront when dredging and purging is where to dispose of the patient's toxic and noxious Qi. To maintain a clean treatment room it is essential for this pathogenic energy to be disposed of properly. This will keep both the doctor and other patients from contamination. If the doctor's Qi becomes toxic or contaminated during treatment, this turbid energy may be transmitted to other patients.

Because the patient is in a state of energetic shock (after the therapy), attention must be placed on the patient avoiding contact with the contaminated Qi. Otherwise, there is a possibility of the diseased Qi returning into the patient's body via the energetic cords of attachment connected to the Toxic Qi.

Imagination and visualization are two of the most powerful and widely used techniques employed in healing today for disposing of Toxic Qi. Since we experience reality through creative imagination, pictures, words, and feelings are extremely powerful techniques which can be used to restructure our physical, mental, emotional, energetic, and spiritual foundations. This also holds true for the disposing of the patients' pathogenic Qi.

There are several methods for disposing of the patients' pathogenic energy in the treatment room. These methods include purification by either sound, light, or the Five Elements, in conjunction with disposing of the toxic energy through the energetic vortex.

PURIFICATION BY SOUND AND LIGHT

- Sound penetrates any substance and causes profound shifts in energy, quickly breaking up clusters of stagnant energy. Some examples of sounds used for purifying the treatment room include: music, toning, chanting, singing, wind chimes, gongs, singing bowls, bells, drumming, and clapping the hands.

- Light can be used to cleanse and purify the treatment room of toxic energy, due to the radiant heat emanating from the light source. Some examples of light being used to purify the clinic include: exposure to sunlight, candles, and imagining the divine healing light incinerating any energetic toxic debris.

PURIFICATION BY THE FIVE ELEMENTS

Some Qigong doctors utilize the energy stored within the Five Elements to cleanse the treatment room of toxic debris. This works best in conjunction with the light and sound purification methods. When using one or more of the elements, it is best to consider how much debris has accumulated from the previous treatment and choose accordingly.

1. Dispose of the Toxic Qi through Fire. By burning fire-water (alcohol with salt), lighting a candle, or releasing the Toxic Qi into the Earth's molten core, Toxic Qi can be transformed and purified.
   - To create fire-water, the Qigong doctor places a small amount of alcohol in a round dish and sprinkles a cone of salt (or camphor) in the center of the dish. Next, the doctor lights the salt and allows the smoke to purge the treatment room, making sure that the doors and windows are open to secure clean ventilation.
   - Other Qigong doctors may imagine the center of the Earth as an incinerator. This image is used to purge the patient's Toxic Qi by directing it to the Earth's molten core. Once the patient's Toxic Qi is incinerated (at the speed of thought), the transformed / purified energy is reabsorbed back into the patient's body.
2. Dispose of the Toxic Qi through Water. Water can also be used to absorb toxic energy and for healing and purifying. Some examples of water usages include: Holy Water and salt water. Some doctors use a container of salt water as their disposal unit for the patient's Toxic Qi (Figure 26.6). Once the patient's toxic energy reaches the container of salt water (usually placed underneath the treatment table), it is immediately absorbed due to the molecular structure of the salt crystals. Sometimes it is advisable for the Qigong doctor (after a full day at the clinic, or at the end of a busy week) to submerge him or herself in a baking soda or take an Epsom salt bath in order to cleanse and purify the body.
3. Dispose of the Toxic Qi through Wind / Air. Wind and air can be used to purify the room of toxic energy. Some examples of this include open air ventilation along with the placement of indoor plants.
4. Dispose of the Toxic Qi through Divine Healing Light. Some doctors dispose of the patient's pathogens into the air, where it is immediately incinerated by a divine healing light. Because of the preestablished divine healing field of energy, once the patient's toxic energy leaves the doctor's hand, it is immediately absorbed by the divine healing light and is instantly transformed into purified energy and reabsorbed back into the patient's body.
5. Dispose of the Toxic Qi through Wood. Wood can also be used to purge toxic energy from the treatment room. Some examples of the Wood Element used to cleanse the room are: the burning of incense or herbs such as myrrh, frankincense, cedar, sage (an old American Indian method), sandalwood or camphor.
6. Dispose of the Toxic Qi through the Earth. Certain stones can be used in the clinic because of their ability to absorb Toxic Qi. Strategically placed in the clinic, these Earth Elements can be used as a natural filtering device. Some examples of this include: amethyst, salt, clear and colored quartz, as well as many other minerals or gemstones. These crystals require cleaning at the end of each treatment day to dispose of the toxic Qi they have absorbed, and should be cleansed by placing them in salt water till they are clean, or burying them in the ground for three days.

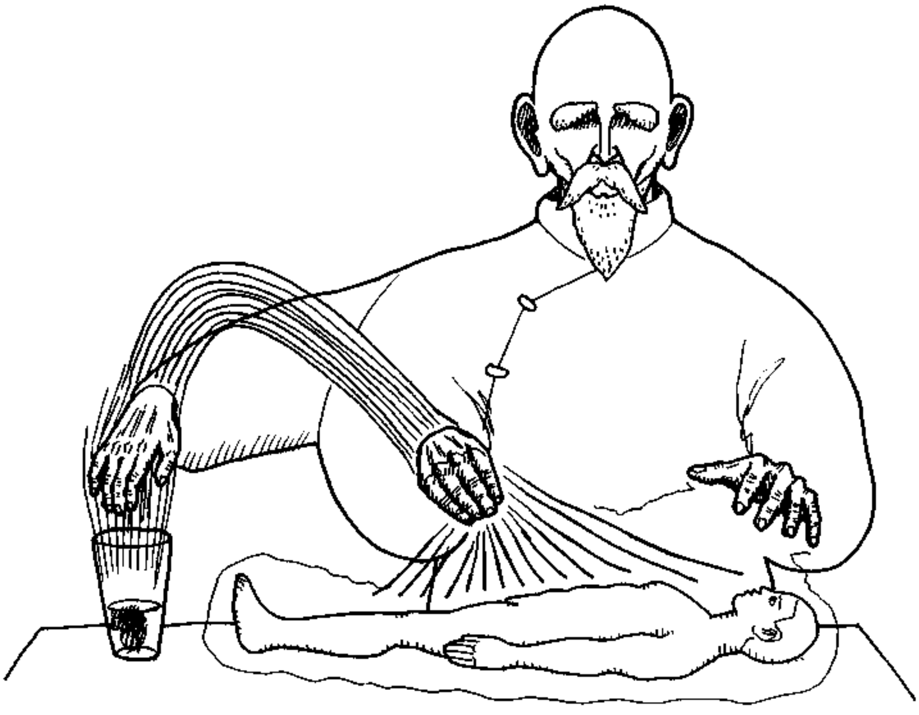

Figure 26.6. One example of a Qigong doctor using a rubber acupuncture doll for treating and disposing of the patient's toxic energy into salt water.

MEDICAL QIGONG TREATMENTS

The following is a clinical procedure that a Medical Qigong doctor will use when treating a patient. The clinical procedure known as the "preparation" remains the same for all Medical Qigong treatments (i.e., the hook-up, dredging and purging the patient, as well as energetic diagnosing). However, the initial treatment will vary depending on the patient's constitution, disease, and personal goals.

After scanning the patient the doctor has a solid understanding of which areas or organs in the patient's body are currently in a state of Excess or Deficiency. The doctor then chooses which area to begin dredging and purging in accordance with the treatment plan. Only after purging all areas of Excess should the doctor then move to tonifying those areas that are Deficient.

TONIFYING A PATIENT'S BODY

After completing the diagnosis, if you have determined that the patient's body is not in an Excess condition, but is instead constitutionally weak or depleted due to Kidney exhaustion, the treatment may be continued as follows.

1. Begin by extending energy through the patient's Yongquan Kd-1 point at the bottom of each foot. Imagine that the Qi is ascending up the patient's legs, via the Yin channels, filling the Lower Dantian, Kidneys, thorax and spinal column into the brain. Hold this position for several breaths, then proceed to the next area.
2. Physically embrace the area of the patient's ankle with one palm, while the other palm embraces the Kd-1 point on the same foot. Start with the right ankle first and then the

left. Project energy into the foot and ankle, by holding the mind's intention on that specific area for several breaths, then begin circling the energy up the Yin channels and down the Yang channels. Hold this position for several breaths, then proceed to the next area.

3. Energize each of the patient's knees (back and sides) by holding the mind's intention on those specific areas for several breaths, then connect each knee to the Yongquan Kd-1 point at the bottom of each foot (the right foot first, then the left). Connect and emit energy into the bottom of each foot to the knee, circling the Qi up the Yin channels and down the Yang channels. Hold this position for several breaths, then proceed to the next area.
4. Energize each of the patient's hips, by holding the mind's intention on that specific area for several breaths, then connect each hip to the Yongquan Kd-1 point at the bottom of each foot (right first, then the left). Connect and project energy into the bottom of each foot to the hips, circling the Qi up the Yin Channels and down the Yang Channels. Hold for several breaths, then proceed to the next area.
5. Energize the patient's Lower Dantian by holding the mind's intention on that specific area for several breaths, allowing the Qi to completely fill up the energetic chamber. Place your left hand on the patient's Mingmen GV-4 point. Your right hand is placed on the patient's navel (CV-8) with your fingers extending down to the patient's Qihai CV-6 point. Mentally connect the Lower Dantian with both Yongquan Kd-1 points on the bottom of the patient's feet and imagine the energy simultaneously traveling down the outside of both legs into the feet. Imagine the energy from the feet continuing to travel along the inside of both legs back into the patient's Lower Dantian area. Hold this intention for several breaths, then proceed to the next area.
6. Energize the patient's Yellow Court by holding the mind's intention there for several breaths. Your left hand is placed on the Jizhong GV-6 point, your right hand should be placed on the Juiwei CV-15 point. Mentally connect the Yellow Court with the patient's Lower Dantian, allowing this energy to connect and fuse both areas. Hold this intention for several breaths, then proceed to the next area.
7. Energize the patient's Middle Dantian by holding the mind's intention there for several breaths. Place your left hand on the Shendao GV-11 point, while your right hand is placed on the Shanzhong CV-17 point. Mentally connect the Middle Dantian area with the patient's Lower Dantian. Hold this position for several breaths, then proceed to the next area.
8. Energize the patient's throat area by holding the mind's intention on that specific area for several breaths. Your left hand is placed on the Dazhui GV-14 point, while your right hand is placed on the Tiantu CV-22 point. Mentally connect the throat area to the Middle Dantian area, then to the patient's Lower Dantian. Hold this intention for several breaths, then proceed to the next area.
9. Energize the patient's Upper Dantian area by holding the mind's intention there for several breaths. Your left hand should be placed to stimulate the occipital region at the Naohu GV-17 point. Your right hand is simultaneously stimulating the Yintang (Third Eye) and Baihui (GV-20) areas. Mentally connect the Upper Dantian area to the Middle Dantian first, and then to the patient's Lower Dantian. Hold this intention for several breaths, then proceed to the next area.
10. Regulate the patient's Microcosmic Orbit, to balance the Heart Fire and Kidney Water energy. Remove your hands from the patient's body, yet still remain energetically connected through your intention. Having stepped away from the treatment table, continue treating the patient with external Qi emission, regulating the patient's Microcosmic Orbit (extending energy up the Governing Vessel and down the Conception Vessel). Hold this intention for several breaths, then proceed to the final step in the treatment.

11. Envelop and seal the patient's three external fields of Wei Qi by connecting the patient's body with the divine energetic field. Hold this intention for several breaths, then proceed to the "closure."
12. Disconnect yourself from the patient and close the healing session. Be sure to disconnect physically, mentally, emotionally, energetically, and spiritually from the patient by willfully detaching all energetic connections to the patient.
13. After disconnecting from the patient, center yourself, then regulate your own energetic fields to balance, recharge, and replenish any energetic depletion.
14. Finally, after you have completed centering yourself, finish the cleansing by washing your hands to purge any of the patient's lingering pathogens which may still be attached to your body.

GENERAL THERAPEUTIC TREATMENT

The following is an example of a clinical routine used in the Medical Qigong Clinic for "emotional purging." The following purging routine is separate from the previous clinical modalities for tonifying a patient. It is important to note that when treating a patient, the Qigong doctor usually purges the patient's Liver and Lungs before tonifying the Spleen and Kidneys.

After completing the diagnosis, if you have determined that the patient has both Excess and Deficient conditions, you may decide to continue the treatment as follow.

1. Begin by dredging, purging, and activating (energizing) the patient's Upper Dantian area to reconnect the patient to the divine, to activate the patient's spiritual/emotional presence.
    a. Start by purging the patient's Upper Dantian.
    b. Then regulate and energize the patient's Upper Dantian by performing a small Microcosmic Orbit within the patient's head.
    c. Finally energize the patient's Taiji Pole through the Baihui area.
2. Purge and dredge the patient's Lungs, to disperse the emotions of grief and despair caused by "rejection, betrayal, and abandonment issues" which are stored in the patient's Lungs.
    a. Start by dredging the patient's neck (trapezius, clavicle, and manubrium) of grief and despair. Begin by standing on the right side of the patient's body, and use the Kneading Tiger Palm and Dragon's Mouth Palm hand techniques (see Chapter 33) to pull the Toxic Qi from the patient's body. Switch from the right to the left side and continue purging.
    b. Then roll the patient's shoulders back to open the Lungs and to facilitate their purging of any deep seated grief. Start on the left Lung first. This area is responsible for storing the emotions related to the patient's mother.
    c. Next dredge the inside Yin Channels of the patient's left arm by using Qi Massage techniques (see Chapter 37).
    d. Regulate and treat the patient's body through "hand diagnosis and treatment" (based on energetic hand/body association, see Chapter 25) by focusing on the patient's left hand.
    e. Finally, begin the entire Lung purging procedure again, this time focusing the attention on the right side of the patient's body. This side reflects the patient's issues with his or her father.
3. Regulate and tonify the patient's Heart to calm the Spirit.
    a. Your left hand connects and extends energy deep into the patient's Shendao GV-11 point, the area in-between the patient's shoulder blades. Your right hand connects to the Shanzhong CV-17 point above the patient's Heart, purging emotional toxins.
    b. Finally, regulate the Heart by circulating Qi through the patient's upper Microcosmic Orbit (between patient's Upper and Middle Dantian). This reconnects the divine light emanating

from the Upper Dantian back with the spiritual emotions (unconditional love, devotion, etc.) stored within the upper chamber of the Heart (see Chapter 5).

4. Purge and dredge the patient's Yellow Court to rid the body of toxic emotions stored within the tissues of the Yellow Court. Your left hand remains secure on the patient's back, at the Shendao point to maintain contact with the Heart Qi. The patient's true feelings are stored within the Yellow Court and expressed through the patient's Heart. Place your right hand on patient's Juiwei CV-15 point, located below the patient's xiphoid process. Sometimes you may see images of the patient's past as traumas are released from the tissues.
5. Your left hand remains secure on the patient's back, at the Shendao point while your right hand purges the patient's Liver organ of anger and depression caused through the storing of "rejection, betrayal and abandonment issues." As you dredge the Liver's Hot toxic emotions from out of the patient's body, purge the pathogens out the patient's right Gall Bladder Channel.
6. Keep your left hand on the patient's back, at the Shendao point while you tonify and regulate the patient's Spleen with your right hand.
7. With your left hand on the patient's back, at the Shendao point, tonify and regulate the patient's Kidneys with your right hand.
8. Maintain contact with the patient's back, at the Shendao point, as your right hand connects with the patient's Mingmen (GV-4) area to regulate all three Dantians. This is initiated through the Microcosmic Orbit (extending energy up the Governing Vessel and down the Conception Vessel), to balance all of the patient's energetic reservoirs.
9. Then slowly remove your hands from the patient's body, but remain energetically connected through intention. Step away from the treatment table, as you continue to regulate the patient's Microcosmic Orbit. Hold this intention for several breaths, then proceed to end the treatment.
10. Envelop and seal the patient's three external fields of Wei Qi, by connecting the patient's body with the divine energetic field. Hold this intention for several breaths, then proceed to the "closure."
11. Disconnect physically, mentally, emotionally, and spiritually from the patient to close the healing session. It is important to shake or whip your palms to completely disconnect from the patient's energy. The shaking or whipping action frees the doctor from any energetic attachment which may have occurred while treating the patient.
12. After disconnecting from the patient, center yourself, then regulate your own energetic field. This balances, recharges, and replenishes any energetic depletion you may have suffered during the treatment session. If you are affected by any pathogenic Qi, immediately expel it from your body.
    a. Next, focus your attention on your Taiji Pole, imagining both the universal and environmental energy collecting, absorbing, and resonating deep within your Lower Dantian and center core. This image draws Qi into your body like a magnet, replenishing any energetic depletion you may have experienced. By placing the fingertips together in front of the Lower Dantian, you will further facilitate the replenishment of the energy.
    b. During this time the patient is allowed to continue relaxing on the treatment table and to slowly come out of the trance state induced by the treatment. The patient should then allow the Yuan Qi to settle back to its origin (the Lower Dantian), by relaxing and calming the mind, breathing, and posture. Instruct the patient not to wash the area that was treated for several hours because water absorbs some of the Qi that was projected into the diseased area.
13. Finally, wash your hands, using tepid water (hot water allows turbid Qi to enter the

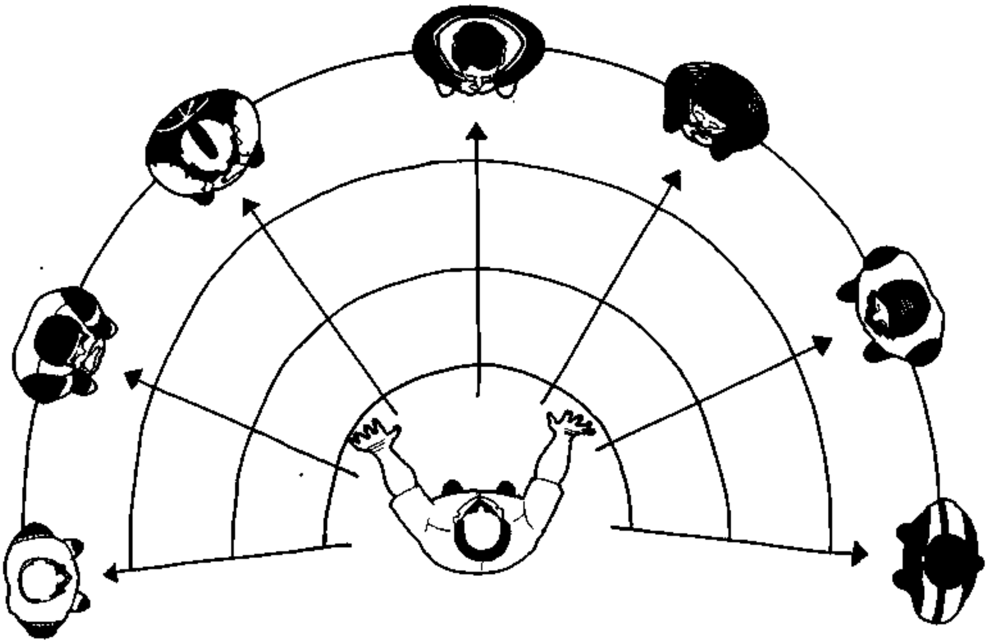

Figure 26.7. As the patients practice their own individual Medical Qigong prescriptions, the Qigong doctor extends energy into the patients, helping to facilitate a stronger therapeutic reaction.

doctor's channels, cold water closes off the channels) to purge any of the patient's lingering pathogens.

ADJUSTING AND FORTIFYING THE PATIENT'S ENERGETIC GRIDS

At the end of the treatment the doctor regulates the patient's energetic fields by adjusting and fortifying the patient's energetic grids (the webbing within the patient's energetic fields). This keeps the patient's energy in harmony, allowing both his or her internal and external fields to operate at their maximum potential for health and healing.

The Qigong doctor adjusts and fortifies the patient's energetic grids, to initiate a permanent energetic transformation for healing within the patient's energetic fields, tissues, and cells. The body's energetic grids are responsible for tissue formation and for maintaining the production, growth, and development of energy. These grids are influenced by thoughts and emotions.

The creative subconscious mind maintains the body's energetic grid formations and causes the physical body to act (grow stronger or weaker) in accordance with the energetic "blueprint" encoded within the structure or webbing. After the Qigong doctor has changed, or corrected, the energetic grid, the patient is given Medical Qigong prescriptions (homework) in the form of visualizations and affirmation meditations, as well as Qigong exercises.

TREATING MULTIPLE PATIENTS

Once an energetic field in the treatment room is established, the patients can begin their individual Qigong regulations. The doctors position themselves according to the number of patients needing treatments; for example, one doctor may be positioned in front of several patients and extend Qi into the patients' energetic fields (Figure 26.7). The doctors' postures and energy extension varies according to their individual skill and energy projection level. It is important for Qigong

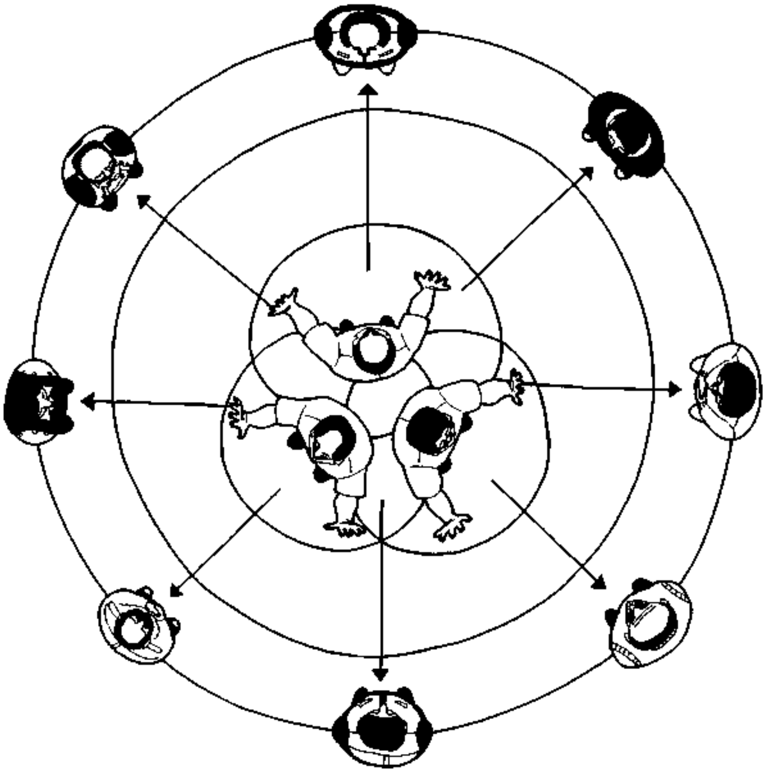

Figure 26.8. The Qigong doctors will position themselves back to back, in the center of a circle, in order to induce energetic movement within the multiple patients' tissues.

doctors to set the pace for treatment. The patients' rhythm is based on the pulse and vibration of the Qigong doctors.

In China, doctors usually rotate every fifteen minutes when treating patients. This prevents the doctors from fatiguing while allowing the patients to absorb new and fresh Qi every fifteen minutes.

If only one or two Qigong doctors are available when treating several patients, then the doctors will position themselves back to back in the middle of the room, with their patients surrounding them in a circle (Figure 26.8). After several minutes the doctors will slowly begin to rotate the center circle extending their energy, and treating each patient individually. The doctor's emitted Qi combines and synergizes with the patient's Qi to produce a much stronger energetic healing field.

By establishing a treatment room that facilitates the regulation of the patients' Qi, the treatments tend to become more effective with faster results. In China, patients are treated in a group treatment room environment three times every day (sunrise, sunset, plus one time before bed).

ENDING THE TREATMENT

When the treatment is over, encourage each patient to relax and then return the patients' Qi back to its origin (the Lower Dantian). The duration of the treatment time, and proper recuperation after Medical Qigong treatment will be addressed next.

DURATION OF TREATMENT TIME

The duration of treatment time varies according to each patient's disease, condition, constitution, and age. A long treatment is usually between twenty and thirty minutes and is generally performed on cancer patients, senior citizens, or for severe traumatic or chronic illnesses. A short treatment is generally between three to fifteen minutes and is usually administered for athletic sprains and ligament dislocations.

PROPER RECUPERATION AFTER MEDICAL QIGONG TREATMENT

Proper recuperation is very important for the replenishment of strength and energy after the treatment session (even after the patient's own prescription practice). This may include having the patient change his or her life activities. The patient must incorporate a balance of work and rest, and acquire regular healthy eating habits (with reasonable nutritional goals). Establishing adequate sleeping patterns is also necessary. Of equal importance is an open-minded positive attitude by the patient, for this reason, a supportive network can be invaluable.

After the Qigong treatment the patient is also encouraged to avoid cold showers, the consumption of alcohol, excessive sex, and recreational drugs. Also prohibited is anything that will emotionally shock the patient, as remaining in a quiet state of mind and spirit is essential.

CHAPTER 27

COMBINING MEDICAL QIGONG THERAPY WITH OTHER HEALING MODALITIES

COMBINING MEDICAL QIGONG THERAPY WITH TCM MODALITIES

Traditional Chinese Medical treatments in conjunction with clinical Qigong modalities are generally broken down into exclusive Medical Qigong Therapy, Medical Qigong and Chinese Massage Therapy, Medical Qigong and Acupuncture/Moxa Therapy, and Medical Qigong and Herbal Medicine Therapy.

EXCLUSIVE MEDICAL QIGONG THERAPY

Medical Qigong therapy can be administered by the Qigong doctor (Figure 27.1) to tonify and promote energy circulation or to sedate the body's energetic channels and collaterals. Qigong prescriptions can also be given for the patient to practice alone. Medical Qigong therapy by itself is excellent for the treatment of ailments such as:

- Insomnia
- Neurasthenia (Chronic Fatigue)
- Hypertension
- Stroke
- Acute abdominal pain
- Injuries to the nerves
- Hemopleura (Blood in the walls of the thorax and diaphragm, along the pleural space)
- Myopia
- Migraines
- Muscle atrophy
- Sprains
- Cancer
- Tumors
- Cysts

MEDICAL QIGONG AND CHINESE MASSAGE THERAPY

Medical Qigong therapy in conjunction with Chinese Massage therapy is a powerful form of combined clinical modalities (Figure 27.2). It enhances tissue response as well as the activation of psychophysical reactions within the body. These therapies can be combined to clear the joints before setting the bones (as in Jie Gu therapy), to purge and dredge the channels to adjust the Qi circulation before regulating the body's muscular system (as in Tui Na and Gua Sha therapies), or to disperse toxins before regulating the body's internal viscera

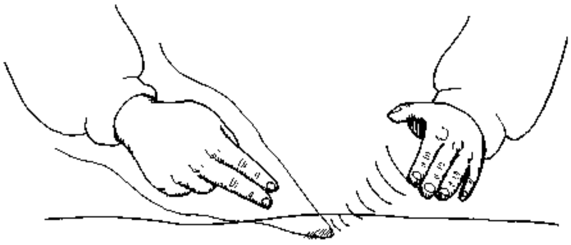

Figure 27.1. Medical Qigong Therapy

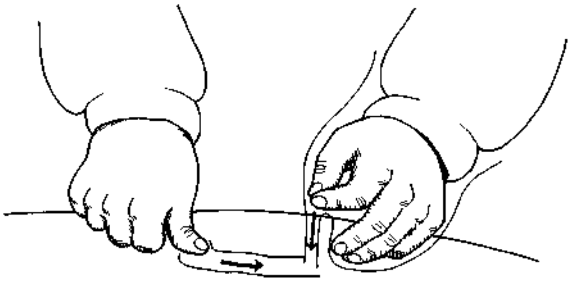

Figure 27.2. Medical Qigong Therapy with Chinese Tissue Regulation Therapy

(as in An Mo and Jing Point therapies). Qigong and Chinese Massage therapy are most commonly used in the treatments of patients with:

- Deep tissue obstruction
- Stiff muscles
- Frozen shoulder
- Low sensitivity to emitted energy

MEDICAL QIGONG AND ACUPUNCTURE/MOXA THERAPY

Acupuncture needles are inserted into the patient's body before the doctor extends Qi into the patient to facilitate healing (Figure 27.3). This also applies in the treatment of Moxa sticks in which the doctor extends energy into the patient through the burning herb. The combined energies (the Moxa heat and doctor's own Qi) initiates increased Qi flow within the patient's channels resulting in a stronger tonification. Moxa is especially beneficial in the treatment of arthritis. Qigong and Acupuncture therapy are combined in tonification or purgation treatment.

MEDICAL QIGONG AND HERBAL MEDICINE

Medical Qigong Therapy in combination with herbal therapies (or with pharmaceutical prescriptions) are becoming more popular. There is ample evidence that the combination of self-applied Medical Qigong therapy and drug therapy is far superior to that of drug therapy alone. In China, this data is reported in several studies of patients with hypertension as well as in cancer patients.

Medical Qigong therapy relaxes the body, promotes the flow of Qi, Blood, oxygen, and nutrients to all cells of the body, as well as promotes the removal of waste products from the cells. Medical Qigong also promotes drug uptake by the cells and tissues by means of increased microcirculation within the body's microcirculatory system. This increased Qi flow through the body's microcirculation nourishes diseased or stressed tissues.

One diagnostic herbal technique I observed at the Xi Yuan Hospital was performed by a Qigong doctor who absorbed into his body the effects of a specific medicine by touching the bottle that contained the herbs. He then extended energy to his patient and began absorbing the patient's Qi to see if the patient's body was affected by the herbal prescription (Figure 27.4).

The Qigong doctor is encouraged to prescribe herbs along with certain of the Medical Qigong exercises. If the Qigong doctor is not qualified to prescribe Chinese medical herbs, then a referral can be made to an acupuncturist/herbalist. The herbal prescriptions will vary according to the patient's condition, constitution and illness.

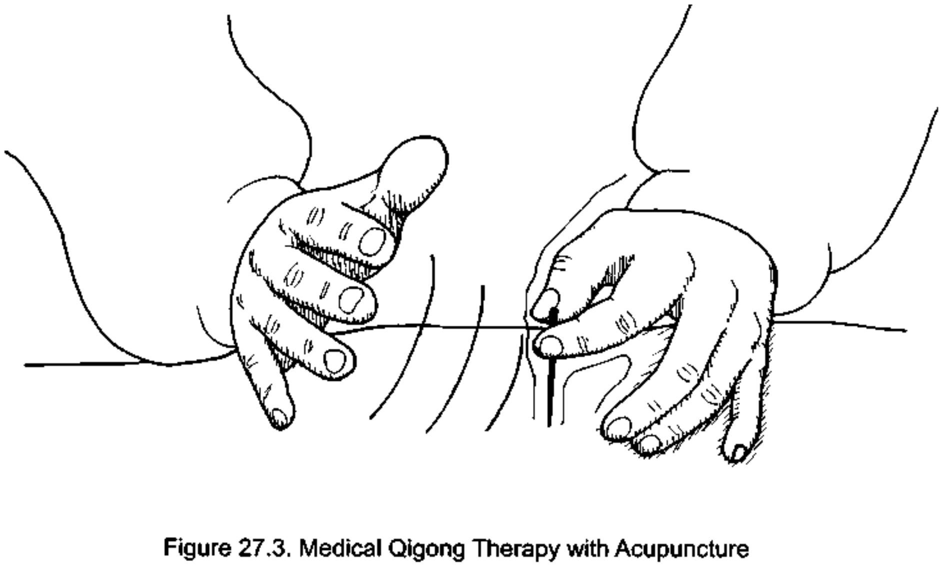

Figure 27.3. Medical Qigong Therapy with Acupuncture

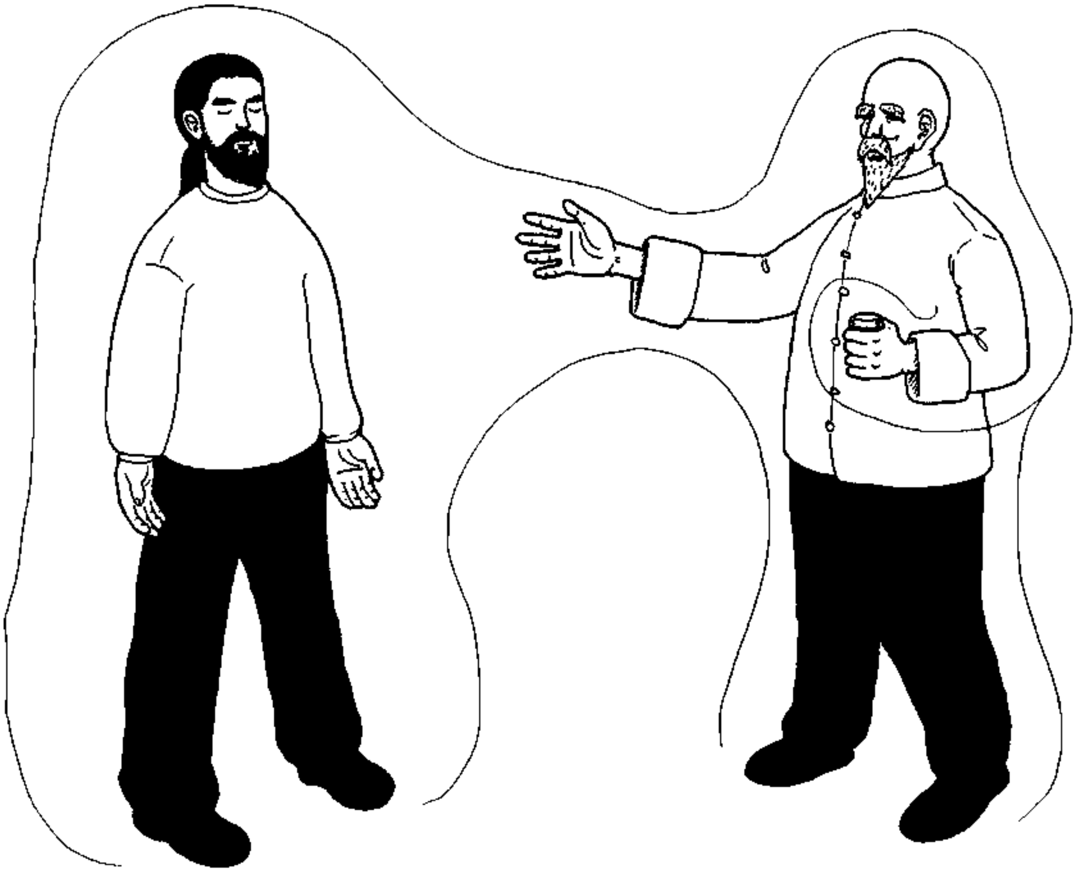

Figure 27.4. A Qigong Doctor Testing the Efficacy of a Specific Herbal Medicine for his Patient

Qigong and herbal therapies are most commonly used in the treatments of:

- Tumors
- Stomach cancer
- Lung cancer
- Hepatitis
- AIDS
- Stroke

COMBINING MEDICAL QIGONG THERAPY WITH WESTERN MEDICAL MODALITIES

Traditional Chinese Medical Qigong serves as a powerful tool in aiding Western medical approaches in alleviating patients' suffering. Because of its emphasis on mental, emotional, energetic, and spiritual approaches to healing, Medical Qigong therapy has been successfully combined as a complementary source of medical treatment with the following Western professional approaches:

- Pediatrics
- Geriatrics
- Gynecology
- Neurology
- Psychology
- Oncology
- Surgery

CHAPTER 28

PRECAUTIONS AND ETHICAL CONSIDERATIONS FOR THE QIGONG DOCTOR

AVOIDING ENERGETIC DEPLETION

Because Medical Qigong therapy consumes energy, as long as the Qigong doctor's quantity of energy is higher than that of the patient's, the doctor can transmit energy into the patient with no ill effect. If, however, the doctor is very tired or depleted and tries to treat a patient, the energy that the doctor is able to produce may become weaker than that of the patient.

It is important to note that energy currents flow from high to low (much like water). Therefore, the doctor may, if in a weakened state, absorb the Turbid Qi or disease energies from the patient. Any imbalance in the doctor's organ system can increase the doctor's vulnerability to foreign pathogens. These pathogenic sensations may also be transmitted into the body of healthy people, or other patients as well. Individuals who have not practiced Qigong, however, and whose energy circulation and channel points are not open, are not as prone to interference by these sensations. These individuals have a natural barrier, or to put it plainly, they are not sensitive to energetic fields of transmission and because of their denial system, cannot perceive energy sensations.

If the Qigong doctor already has a tendency towards a particular illness, exposure to pathogens may instigate or intensify the disease. If the doctor is suffering from any energetic Deficiencies, for example, the end result may be that the doctor performing the treatment may experience Qi deviations due to absorbing the patient's Toxic Qi. This means that more Qi and energy have been extended outside of the doctor's body than can be replenished, weakening the doctor's energetic constitution.

When treating patients the doctors receive and diagnose information from their Six Openings (also called the Six Windows of the Body). These Six Openings are described as follows:

- The eyes -- open for seeing energy patterns and colors,
- The ears -- open for hearing energy patterns and tones,
- The nose -- opens for smelling different odors,
- The tongue -- opens for tasting or flavor sensing,
- The body -- opens for energetic feeling sensations, and
- The Spirit (Heart/Mind) -- opens for energetic perceptions of emotions, thoughts, and spiritual insight.

While using the Six Openings for diagnosis, the doctor naturally depletes some life-force energy. Since all five organs (eyes, ears, nose, tongue, and body) are directed by the doctor's Spirit (Shen), the expenditure of energy is considerable. Any imbalance due to organ depletion will cause the doctor to absorb Turbid Qi from the patient. This Turbid Qi can move within the doctor's body causing Qi deviations, especially if it mixes with the doctor's Clear Qi.

1. If this happens within the doctor's Upper Dantian, the symptoms may include mental fatigue, dizziness, headaches, etc.
2. If this happens within the doctor's Middle Dantian, the symptoms may include emotional fatigue, discomfort in the Liver (hepatic region), etc.
3. If this happens within the doctor's Lower Dantian, the symptoms may include physical fatigue, cold in the extremities, discomfort in the lower abdomen, etc.

This is why Qigong doctors must constantly be refilling and rebalancing themselves energetically. It is important to note that if the Qigong doctor has indeed ingested the patient's pathogens, and Qi deviations have resulted, no Western doctor, acupuncturist, or herbalist can help him or her. Only the doctor with the Qi deviation (or with the help of another Qigong doctor) can change the Turbid Qi back to Clear Qi, by using divine energy and purging the body's Shen, Qi, and finally the Jing. This supports the doctor's Righteous Qi and fights the invasions of the patient's pathogens.

Emission of Qi should also be avoided if the doctor becomes sick, tired, hungry, overfed, deeply grieved, indignant, or drunk. Also, if the doctor is unable to stay in the present mind-frame he or she will inevitably fall asleep or "Shen out." To avoid this state of depletion, the Qigong doctor should:

- Cut back on the number of patients,
- Decrease treatment time—from 25 to 20 minutes maximum,
- Maintain a consistently healthy diet with supplements,
- Receive frequent exposure to fresh air and sunlight, and
- Perform the Microcosmic or other regulating Qi meditations in between patients.

Note: When treating patients with Medical Qigong therapy the doctor should never wear the color red. The color red is too hot and has a dispersing affect on Qi, which tends to scatter the body's energetic fields, distorting the Qigong doctor's Qi emission.

PREVENTING THE INVASION OF TURBID QI

When Turbid Qi enters the doctor's body, it interferes with normal Qi circulation, causing disorders in part of, or in all of the energy circulatory system. In mild cases, Qi may stagnate in a certain areas of the body (i.e., shoulder, arm, chest, or back), causing physical and mental symptoms.

- Physical symptoms may include: tingling, pain, cold, contraction, heaviness, soreness, and distention and stuffiness in the chest; and
- Mental symptoms may include: interference with mental activity, causing dizziness, headache, heaviness in the head, vexation, and restlessness.

In severe cases, the doctor may experience the same symptoms as the patient, but upon physical examination there is no verification of any existing disease. The doctor should therefore be on the alert to avoid any danger of developing symptoms from the patient after Turbid Qi is perceived.

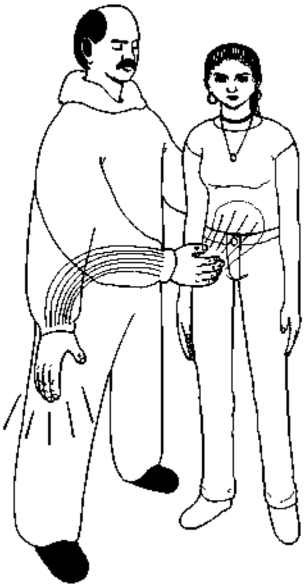

Figure 28.1. In order to disperse the patient's Evil Qi, the Qigong doctor exhales while slinging the toxic energy into the Earth.

It is very important for doctors who treat patients with Medical Qigong therapy to possess the ability to prevent and expel Turbid Qi so that it will not disturb their own energetic activities. Qi deviations are most often seen in those individuals who treat patients after they have gained some knowledge of Medical Qigong but have no experience, though they can occur even to some veterans of Medical Qigong practice. The interference by Turbid Qi is often an important factor in the the doctor's health.

When the interference by Turbid Qi is perceived, the doctor should use proper hand manipulations and readjust energy activities to expel the Turbid Qi immediately. If the Turbid Qi invades the doctor's fingers, or penetrates into certain channels or points, the doctor should guide the Qi to the specific points, channels, and infected areas and then relax and shake the hands to discharge the Turbid Qi while exhaling.

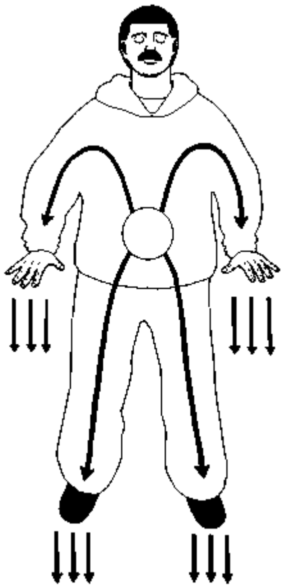

Figure 28.2. The Qigong doctor gathers energy into his or her Lower Dantian and expels any and all pathogenic invasions out through the hands and feet.

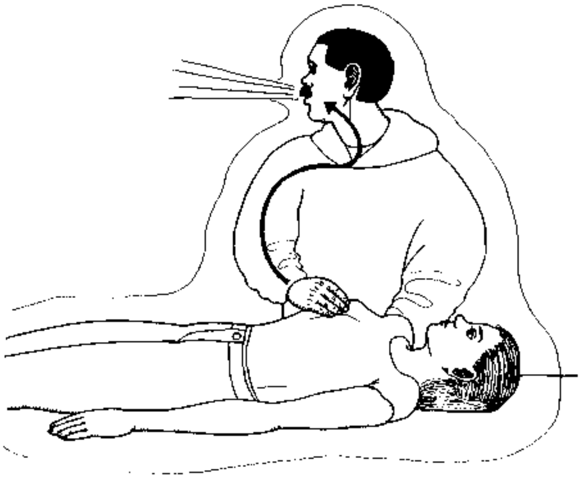

Figure 28.3. To avoid absorbing the emotions from an energetic discharge (released from dissolving an energetic cluster), the Qigong doctor turns his or her head and expels the patient's toxic emotions.

The doctor should be able to stop the Turbid Qi before it reaches the Dazhui GV-14 point on the upper back, Tiantu CV-22 point at the base of the throat, Quepen St-12 points on the front of the shoulders, and the Fengfu GV-16 and Fengchi GB-20 points on the back of the head. It is important to expel the Turbid Qi when it has reached the fingers, wrists, elbows, and at most the shoulders.

When this occurs the doctor should drive the Evil Qi out and regulate his or her Qi circulation as soon as possible. There are several ways to do this:

- Exhale while swinging the arms toward the ground, driving the Evil Qi deep into the center of the Earth. When treating tumors, be careful to disperse the pathogens out your fingers and do not allow the toxins to stick to your palms (Figure 28.1).
- Exhale while "blowing Qi" (extending Qi) out of the palms (Pc-8) and feet (Kd-1) (Figure 28.2).
- Exhale out of the mouth (but never over the patient, or towards anyone else). A shift in breathing is a common reaction when the emotions are discharged, or when an energetic cluster is reached (Figure 28.3).

GUIDELINES FOR THE MEDICAL QIGONG DOCTOR

1. The Qigong doctor must avoid practicing Medical Qigong in wet clothes. If the doctor perspires a little when training, he or she should avoid standing in the wind; if the doctor perspires profusely, practicing should be discontinued as it is liable to damage the doctor's Yin Qi.
2. The doctor should not go to the toilet within half an hour after practicing in order to avoid developing a conditioned reflex. This will help prevent the leakage of vital energy.
3. Patients suffering from cancer may affect the doctor's Qi circulation more seriously. Care should be taken when treating these patients, or avoid treating them completely, until sufficient energy has been cultivated, and sufficient experience has been gained.
4. The doctor must always wash his or her hands

with soap after each patient's treatment to prevent the energetic transference of diseases. It is important to note that washing the hands in cold water immediately after Qi emission is harmful to the doctor's energetic field because of the sudden shock to the system. Washing with hot water immediately after treatment, opens the channels on the doctor's hands and arms, sending the Toxic Qi deep into the doctor's body. To avoid these complications, the doctor should wait several minutes after completing the treatment, before washing his or her hands.

While waiting, the doctor should touch his or her finger tips together and focus on the Lower Dantian, allowing the Qi to settle back into its origin (minimum of three minutes). The doctor then washes up to the elbows with water that is room temperature to reduce the possibility of absorbing the patient's Turbid Qi.

If the doctor is not able, during treatment, to expel the Turbid Qi right away because of some distraction, or if the doctor's internal energy is not substantial, the Turbid Qi may enter his or her body through the fingers traveling up the wrists, elbows, or entering the Baihui GV-20, Tanzhong CV-17, Fengchi GB-20, Yintang (Extra Point) and Yongqauan Kd-1 points. If this begins to happen, the doctor should not continue to extend energy, but should stop the treatment immediately and perform a "closing," expelling the Turbid Qi by flicking the fingers, wrists, and shaking the arms until all discomfort has been eliminated. If any Turbid Qi still lingers, the doctor should practice Medical Qigong exercises to regulate the energetic activities.

5. It is important that doctors, as well as patients, avoid strenuous activities (such as running immediately after Medical Qigong exercises) to prevent the wrong flow of vital energy or pain in the legs. It is always better for the doctors and patients to be active first (running, swimming, forms training, etc.) and then cool down with the Medical Qigong Quiescent exercises (unless otherwise prescribed).

WU ZANG (FIVE ORGAN) MEDITATION FOR ENERGETIC PROTECTION

Even after the treatment of diseases, the doctor must perform certain Medical Qigong exercises to prevent and expel any undetected pathogenic Qi. Otherwise the doctor may develop the symptoms of the illness of the patient in corresponding locations. For instance, if the patient has a headache or pain in the hepatic area, the doctor may also acquire a feeling of discomfort, numbness, itching, and pain in the head and hepatic region. Therefore, in clinical practice, a deep understanding and correct judgment of the pathogenic Evils is compulsory.

Today in China, the Wu Zang meditation is still taught to Qigong doctors for clinical use to prevent invasion from the patient's pathogenic Qi. This meditation requires the doctor's Shen to guide the Qi of the Five Yin Organs (also known as the Wu Jing Shen, see Figure 28.4), to make the body's Qi substantial. The ending of the meditation, which causes the doctor's Qi to be reabsorbed into the Taiji Pole, allows the transformed Qi to flow into the appropriate organs, strengthening the doctor's energetic fields.

When a Qigong doctor enters the space of a diseased patient, the potential for absorbing the patient's pathogenic Qi is high. Noting this potential for harm, the ancient Chinese Qigong masters developed certain meditations to protect themselves from the patient's diseased or Turbid energy.

The Yellow Emperor's Classics on Internal Medicine states that before treating patients of any kind, the doctor should first perform the Wu Zang Energy meditation. It encouraged doctors to practice this meditation before treating patients in a clinical environment and before entering the treatment room, because it surrounds them with a strong field of protective energy (Wei Qi).

The Wu Zang meditation focuses on six directions (north, south, east, west, Heaven and Earth). It is very powerful and can be used to gather environmental Qi at the beginning of each season. The focus of the meditation is to extend each of the five organs' energy far into the horizon to gather the Qi. When practicing the Wu Zang

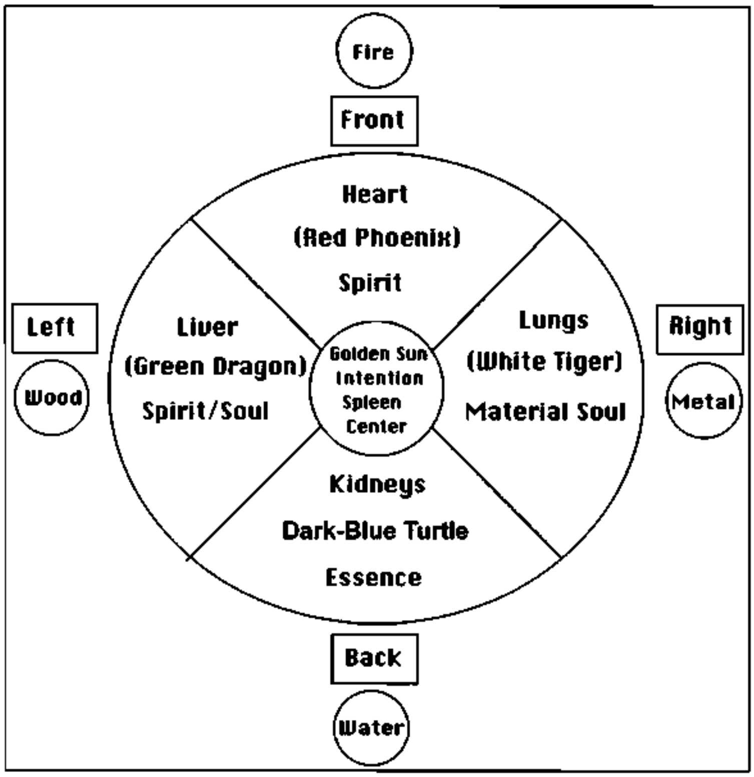
Fire
Front
Heart
(Red Phoenix)
Spirit
Left
Wood
Liver
(Green Dragon)
Spirit/Soul
Golden Sun
Intention
Spleen
Center
Lungs
(White Tiger)
Material Soul
Right
Metal
Kidneys
Dark-Blue Turtle
Essence
Back
Water

Figure 28.4. The Wu Jing Shen Relationship of the Wu Zang Meditation

meditation the Qigong doctor should:

- face east during the spring equinox;
- face south on the day of the summer solstice;
- concentration on the center of the Earth during late summer, before the fall equinox;
- face west on the day of the fall equinox;
- face north on the winter solstice.

Start by focusing on the center of the body. Imagine opening up the Baihui point (at the top of the head) and begin to pull in the Qi from the Heavens. Imagine the divine light as a bright shining white light, illuminating and filling your entire body through the Baihui. Feel the body radiating this divine white light energy. Next, imagine this white light energy coalescing into the body's center core, forming an energetic tube of Heavenly energy that extends from the Baihui point to the Huiyin point. The center core vibrates and resonates with the divine white light energy.

1. Now, imagine a golden yellow mist of Qi arising from under the Earth and filling your body and connecting with the Spleen organ. Next, feel this golden light Earth energy envelop the center core of divine white light energy, merging together and synergizing. This represents the energy of your Yi (Intention), to root and stabilize your power.
2. Begin to focus on your Heart and imagine a portal opening and the Qi flowing out of your Heart like a red swirling wind in front of you, full of power, protecting you with your Shen and the fire of a red phoenix. This represents your innate spirit, alive, graceful yet powerful.
3. Focus your attention on your back and the Kidneys and Mingmen area. Imagine a portal opening and the Qi flowing out of your Mingmen behind you like water. Out of this water grows an enormous dark blue turtle,

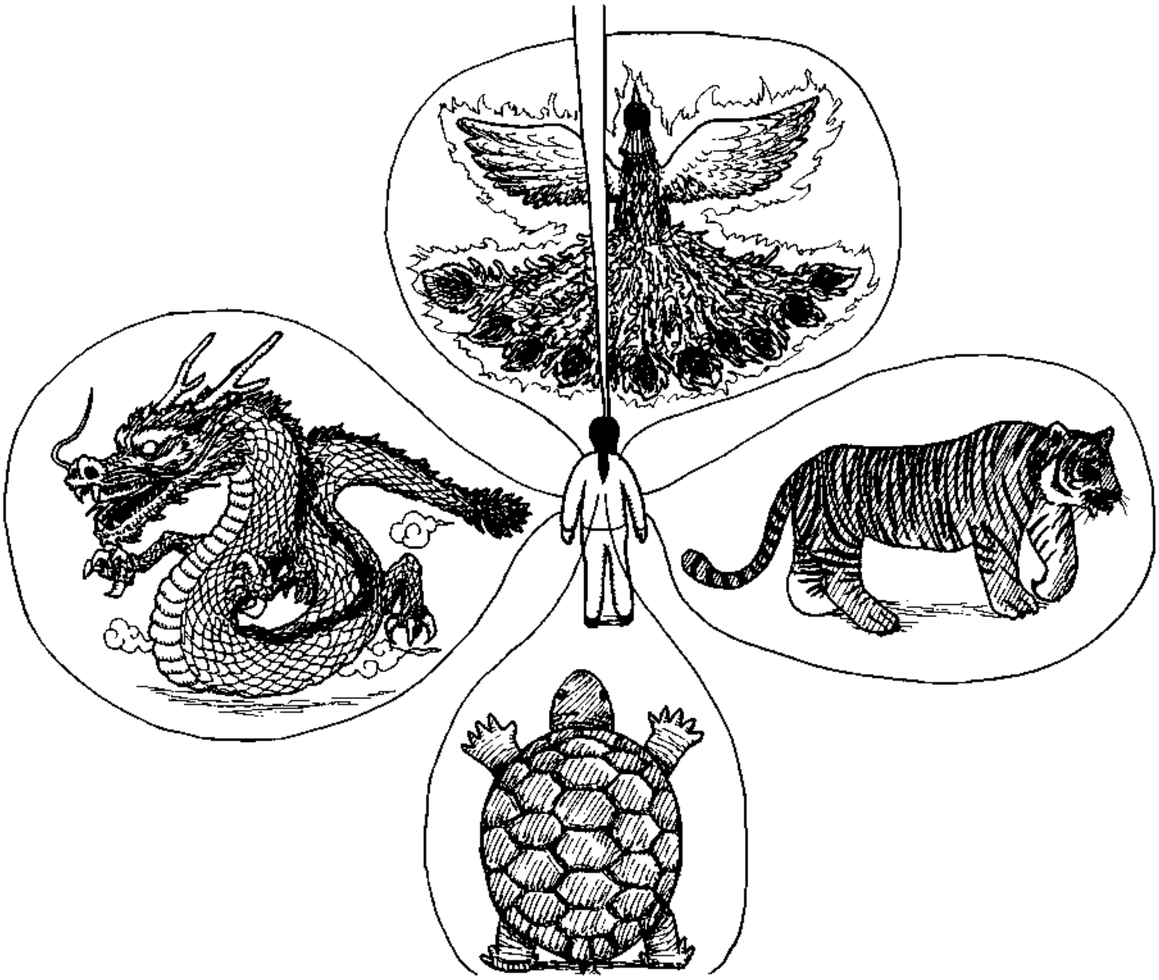

Figure 28.5. The Wu Zang Meditation

whose shell protects you like a mighty shield. This represents the energy of your Zhi (Willpower), and the Jing and Zhi of all your ancestors, backing you and supporting you.

4. Place your attention on your Lungs and visualize a portal opening on your right side of your body, under the right ribs. Imagine the Lung Qi flowing out of to the right side of your body like steam, forming a white tiger, as strong as steel. This represents the Po and the body's animal nature, that guards and protects you with an animal passion for survival.
5. Next, place your attention on your Liver and visualize a portal opening on your left side of your body, under the left ribs. Imagine the Liver Qi flowing out of to the left side of your body like steam, forming a green dragon, sinewy and resilient as bamboo. This represents the Hun and the body's divine nature, guarding and protecting you with a spiritual passion for victory.
6. Each animal begins to rotate to the left, protecting, stalking, and defending the previous animal's position (Figure 28.5). Slowly begin

to circle these energies, then increase their speed (like a mighty wind) counterclockwise around your body, blending these colors together, to form an energy bubble.

After forming a protective rainbow hue around your body, draw the energies back in through the Baihui point and return the energy of each organ color back to its origin (the color red returns back to the Heart, dark blue to the Kidneys, white to the Lungs, and green back to the Liver). As the colors return back to their organ of origin, imagine steam (white light) flowing out of the pores and filling up the energy bubble created by the animal rotation. This forms a solid connection between the body's internal organs and the body's external field of energy.

7. When the protection of the body with the Five Elements is done, imagine that there are many sparkling lights like the Big Dipper above your head. Then you can enter the room of the patient, safe and protected.

Keep in mind that you can send the five organ energy outside of your body instantly, any time you need protection. Some schools teach the doctor to extend his or her energy like mist out through the eyes, enveloping the body with the five organ colors (green, red, yellow, white, and dark blue/indigo).

CREATING A PROTECTIVE ENERGETIC FORCE FIELD

This second meditation also establishes a powerful energetic force field, needed to protect the doctor from any attack or attachment of external pathogenic Qi. It consists of wrapping the doctor's external fields of Wei Qi to increase the energetic field's power.

This is similar to wrapping a magnet with coils of wire to increase its magnetic field (Figure 28.6).

1. Begin from a standing Wuji posture. Focus on all the body's energy collecting into the Lower Dantian. Concentrate the Qi in the Lower Dantian, condensing it into a dense ball of white light energy. Imagine this ball of white light energy exiting the body through the navel and circling the body, spinning in a counterclockwise direction nine times, wrapping the Belt Vessel, Mingmen GV-4 and Shenque CV-8 points.
2. Move the energy ball down to the perineum and continue circling the lower body in a counterclockwise direction nine times, spinning and wrapping the Huiyin CV-1 and Changqiang GV-1 points.
3. Move the energy ball down to the knees and continue circling in a counterclockwise direction nine times, spinning and wrapping the ascending Yin channels and descending Yang channels of the legs.
4. Move the energy ball down to the ankles and continue circling in a counterclockwise direc-

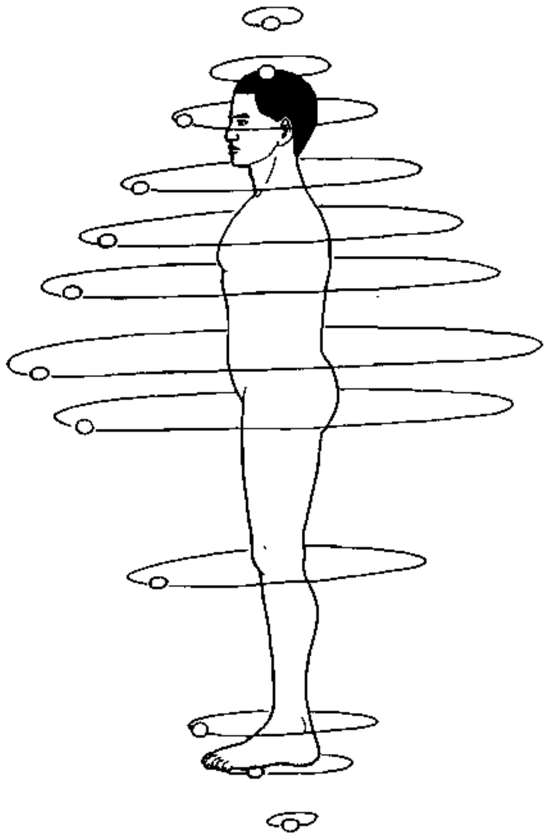
Figure 28.6. Creating a Protective Energetic Forcefield

tion nine times, spinning and wrapping the ascending Yin channels and descending Yang channels of the legs.

5. Move the energy ball down to the bottom of the feet and continue circling in a counterclockwise direction nine times, spinning and wrapping the Yongquan Kd-1 points.
6. Move the energy ball several feet into the ground and continue circling in a counterclockwise direction nine times, spinning and wrapping the body's Earth Transpersonal point. This energetic point (under the ground) is responsible for establishing the body's rooted connection to the Earth and for absorbing the Earth energy into the body's internal organs, channels, and collaterals.
7. Next, imagine the energy ball slowing down and stopping its rotation, pulsing for several heartbeats, then reversing its energetic rotation, and beginning to spin in a clockwise direction nine times. The direction and flow of the energetic ball now reverses its orbit, ascending and circling the same areas and points nine times, until it returns back to the navel.
8. From the navel move the energy ball up to the solar plexus area and continue circling the body in a clockwise direction nine times, spinning and wrapping the Jiuwei CV-15 and Zhiyang GV-9 points.
9. Move the energy ball up to the Middle Dantian at the center of the chest, and continue circling in a clockwise direction nine times, spinning and wrapping the Heart, Shanzhong CV-17 and Shendao GV-11 points.
10. Move the energy ball up to the throat and continue circling in a clockwise direction nine times, spinning and wrapping the Tiantu CV-22 and Dazhui GV-14 points.
11. Move the energy ball up to the Upper Dantian at the Third Eye area, and continue circling in a clockwise direction nine times, spinning and wrapping the Yintang and Naohu GV-17 points.
12. Move the energy ball up to the top of the head and continue circling in a clockwise direction nine times, spinning and wrapping the Baihui GV-20 point.
13. Move the energy ball several feet above the head and continue circling in a clockwise direction nine times, spinning and wrapping the body's Heavenly Transpersonal point. This energetic point (above the head) is responsible for establishing the body's extended connection to the Heavens and for absorbing the divine energy into the body's internal organs, channels, and collaterals.
14. Next, imagine the energy ball slowing down and stopping its rotation, pulsing for several heartbeats, then reversing its energetic rotation, and beginning to spin in a counterclockwise direction nine times. The direction and flow of the energetic ball now reverses its orbit, descending and circling the same areas and points nine times, until it returns back down to the navel. The body should now be encircled, above and below, with several haloes of white light energy.
15. Finally, imagine the white light energy ball absorbing back into the body through the navel. As the energy ball descends back into the Lower Dantian, imagine energetic steam filling the entire body.
16. Once the body is completely full, the steam begins to exit the body through the pores and begins connecting and filling the space outside the body with white light energy, forming an energetic cocoon. This cocoon creates an energetic force field which protects the Qigong doctor from pathogenic invasion.

CLINICAL ETHICS

The study of clinical ethics refers to the morality of the Qigong doctor and the ability to set boundaries. The study of morals in the therapeutic relationship involves reflecting upon the ideals the doctor assumes as a professional care giver and the enforcement of these professional standards in his or her actual behavior. Both the morality and the self reflection are engaged, both consciously and unconsciously, to align the doctor's values with his or her actions.

Ethical behavior stems from the doctor's internal values and actions (emotionally, cognitively, and spiritually) and his or her sense of external connection (physically, existentially, and socially). The Qigong doctor's behavior should be always congruent with high ethical standards and in harmony with his or her values, intuition, knowledge, and feelings.

The importance of establishing ethical standards for the doctor-patient relationship is to avoid those behaviors which prove to be counter therapeutic, and encourage those that are most effective therapeutically for the patient. The patient, as well as the doctor, rely upon each others' trust and boundaries. It is the Qigong doctor's responsibility to create a context for the patient, in which the patient can make a change in his or her life in safety.

When patients are treated energetically, they relax into an altered state of consciousness. This causes the patient's boundaries to become relaxed or nonexistent. Because Medical Qigong therapy often reveals intimate knowledge of the patient's life and life-style, there is a danger of enmeshment, i.e., of co-dependent relationships forming (see Chapter 19).

When working with patients, doctors sometimes encounter intense physical, mental, emotional, and spiritual reactions from the patients. These experiences tend to disperse the patients' boundaries and can confuse the roles between the doctor and patients, intensifying transference and triggering countertransference.

When both the doctor and patient have a personal relationship with a divine higher power, it naturally brings them into a state of divine healing and spiritual wisdom. The acknowledgment and honor of this divine relationship should establish ethical boundaries. This is especially true as the doctor begins to see the bigger picture of how his or her intentions and actions (in relationship with patients) affect others in a rippling outward motion. It is important that doctors also see the effect on themselves when they take certain actions towards others, including their patients.

CONSCIOUS AND SUBCONSCIOUS CLINICAL INTERACTIONS

Clinical experiences are divided into two states of interactions: the body/mind and emotional/spiritual.

- The body/mind interaction deals with ordinary states of consciousness, such as: talking, thinking, analyzing, working, interacting with the material world and other human beings. These different qualities of interactions help the doctor to define and categorize, educate, and establish his or her solid belief structures.
- The emotion/spirit interaction deals with subconscious states of communication, such as: perceiving, sensing, intuiting, interacting with the spiritual/metaphysical world, as well as with energetic fields. Every day people go into states of light trance (daydreaming, spacing-out, etc.) allowing their focus of attention, thoughts, feelings, sensations, and intuitions to interconnect with their environment and that of the divine. This allows access to information and facilitates the healing and understanding of themselves and their life experiences.

Because most wounding involves an emotional/spiritual trauma at the time of the incident, the patients must relive this trauma during their healing process. Through the emotional/spiritual interaction between the Qigong doctor and his or her patients, ingrained habits of thought, feelings, and perceptual understandings begin to recede, diffuse and break down. This causes the patients to find a new understanding of certain traumas and allows them to reclaim disconnected parts of their past (soul retrieval).

Each time patients transcend to a new level of understanding of their fears and belief structures, they change, mature, and expand to adjust to the new mental, emotional, and spiritual growth. The moment that patients break out of their old boundary structure, feelings of emotion rush through their body/mind, followed by a sense of either expansiveness or disorientation as they find themselves in unfamiliar territory.

It is not useful for patients to relive their emotional traumas without the awareness, understanding, and insight of the process they are undergoing. To relive the trauma without this awareness leads to re-traumatization. That is why the Qigong doctor coaxes the patients through the "five stages of emotional healing" (see Chapter 19). The patients can then experience their past emotions in the present time, and become free of them through present awareness, rather then repeating the past through a continuous loop of repression and denial. In the clinic, this is known as "emotional time traveling" and refers to the practice of allowing the patients the experience of reliving past traumas through their presently increased state of awareness and knowledge. At this point of transition, the patients will not only experience the emotions and sensations from their past, but may also experience the motivations and emotions of the perpetrators as well.

The empathic knowledge of the perpetrator's emotions is due to the victim's identification process with the abuser. This phenomenon is well recognized in helpless hostage situations, and is a major contributor to childhood victims of physical, sexual or emotional abuse "acting out" their repressed fears, angers, and pain, in similar ways to the manner in which they were abused.

ESTABLISHING DOCTOR AND PATIENT AGREEMENTS

For the protection and integrity of both the doctor as well as the patients, it is important that a description of the clinical procedures that are to be used, their physical and energetic boundaries, as well as their limitations, be firmly established. The following seven principles are used to establish the foundation of clinical ethics.

1. The Qigong doctor will cause no harm (physical, mental, emotional, or spiritual). The Qigong doctor will maintain a high standard of skill, knowledge, and professional conduct, offering his or her services without favoritism, prejudice, or discrimination with regards to race, religion, nationality, gender, or sexual orientation.

   The Qigong doctor is committed to the patients' education and personal development of practicing Medical Qigong methods that relieve pain, suffering, and other symptoms of illness. The doctor will exercise his or her best judgement to determine if the therapeutic techniques prescribed for patients are contraindicated, dangerous, or not in the best interest of the patients.

   The doctor will never make negative suggestions, criticize, or tell the patients that they will get worse or are in bad shape, as that destroys the patients' hopes and is counterproductive to healing (this is known as "clinical hexing"). Clinical hexing will cause patients, who trust the doctor, to sabotage their own healing potential by succumbing to the doctor's particular beliefs about their conditions.

   One example of trusting a doctor's diagnosis to ruin, happened last year in 1999, when two patients' lab tests were accidently switched. The healthy patient, whose lab test had been switched, was informed by the Western doctor that she had terminal cancer; and the patient with terminal cancer, on the other hand, was given a clean bill of health. The healthy patient died several months later as a result of believing the doctor's diagnosis of having a terminal condition; whereas the patient with terminal cancer (believing that there was nothing wrong) resumed a normal, active life-style, and the patient's cancer went into remission.
2. The Qigong doctor will keep all sessions in confidence. The Qigong doctor will keep all patient information strictly confidential, in keeping with ethical and legal standards held by other health care professionals. This allows patients to experience the treatment in safety and give honest feedback.

   Exceptions to the rule of confidentiality are as follows: when the patient is a danger to self (intending suicide), or others (intending homicide), when the patient has a life threatening emergency, in cases of suspected child abuse (sexual, physical, or neglect), and elderly abuse (sexual, physical, neglect or fiduciary).

3. The Qigong doctor will first receive informed consent from the patient before starting the treatment, or informed consent of the patient's legal guardian. Before treatment, the Qigong doctor explains to the patient the therapeutic goal of the treatment, and what the patient may energetically and emotionally expect to experience. The doctor will respect the patient's boundaries at all times, before, during, and after each treatment. At times the doctor may have to impose boundaries for patients who cannot do so for themselves, in order to avoid total enmeshment, which would endanger the professional relationship.
4. The Qigong doctor and the patient will tell the truth to each other. No part of the Medical Qigong diagnosis should ever be hidden from the patient, yet the truth should always be delivered with hope, since no one can be certain of the future.

   The Qigong doctor should never make unrealistic, inaccurate, or false claims about Medical Qigong therapy, or give the patient false hope. A realistic assessment of the patient's condition should always be conveyed, and if need be, the patient should be referred for additional medical or psychological modalities of treatment.
5. The Qigong doctor will honor all agreements established with the patient. It is important for the patient to be able to rely on the doctor's trust. Because this is a critical factor, the doctor is never allowed to borrow money, cars, etc. from patients.
6. The Qigong doctor and the patient will not act sexually or romantically with each other. The Qigong doctor will not instigate or tolerate sexual advances while interacting with his or her patients. Although there are times that the patient, or doctor, may experience certain attractions towards each other, acting out on these feeling will cause an emotional / sexual entanglement which will be detrimental to the patient's healing.
7. The Qigong doctor and the patient will agree on the time, place, duration of treatment, and fee. The Qigong doctor must be consistent in maintaining an established clinical treatment time and location. This instills a stable security for the patient and doctor. If this boundary is not maintained, patients without strong boundaries will take advantage and hurt both doctor and themselves. The Qigong doctor is responsible for giving adequate advance notice (two weeks) to the patient of any and all treatment or fee changes. If, for example, the original agreement was for six or fewer sessions and the doctor changes fees after the first session, the original agreement must be upheld (new patients however, will be charged the new fee).

POWER DYNAMICS AND ETHICAL BEHAVIOR

In any clinical setting there is an active power dynamic at work. Generally it is the doctor who maintains the power over the patient, as the purpose of the patient's visit is for help or assistance in specific healing. As the patient has come to the doctor with expectations, it is important for the doctor to be congruent with these expectations. These expectations set both the standard of the patient's healing and the power dynamics of the relationship. Through the power dynamics of the doctor-patient relationship there can develop three classical problems: Transference, Counter-transference, and Emotional-Sexual Entanglement.

1. Transference is the process whereby a patient unconsciously transfers feelings, thoughts, beliefs, and patterns of behavior that had been previously experienced towards others onto the doctor.
2. Counter-transference is the process whereby a doctor loses his or her objectivity and unconsciously transfers feelings, thoughts, beliefs, and patterns of behavior to the patient.
3. Emotional-Sexual Entanglement is the process whereby the doctor and patient both lose their objectivity and transfer emotional-sexual feelings, thoughts, beliefs, and patterns of behavior onto each other.

In order to prevent these types of enmeshment and co-dependent relationships from happening, it is important for the doctor to constantly maintain his or her boundaries. The doctor and patient both must remember that it is the patient who actually does the healing in energy work, the doctor merely creates the situation in which this energetic transformation can transpire.

PITFALLS

There are certain pitfalls the Qigong doctor should be aware of and may experience, if a boundary has already been, or is beginning to be violated. These experiences are described as follows.

- The doctor begins to identify with the patient's problem and feels the need to discuss his or her own similar problems with the patient.
- The doctor begins to become emotionally attached or attracted to a patient sexually and feels the need to act on it (i.e., asks for a date, inappropriate touching or sexual contact).
- The doctor begins to feel the need to go out of the way above and beyond the normal course of responsible treatment to meet a patient's needs.
- The doctor begins to treat the patient for free.
- The doctor begins to find him or herself becoming overprotective, or worrying about a patient at inappropriate times.
- The doctor begins to need validation from his or her patient (this approval can be either as a healer, emotionally, or sexually).
- Arranging an exchange of services with a patient can often lead to the patient feeling abused or hurt.
- Allowing the patient to build up a large bill to be paid back later is unethical and may lead to legal action.
- Disclosing too much personal information regarding self (the doctor) in order to help the client can make the patient feel responsible for the doctor's well-being. Self-disclosure can be used very sparingly when the patient can benefit from solutions that worked for the doctor. It is safer, however, to pretend that these situations came from someone else.
- The doctor does not listen carefully to what the patient is saying, thinking that he or she knows what the problem is in advance.

It is important for the doctor to remember that the patient's emotional / spiritual safety and health must always come first. To avoid sinking into one of these pitfalls, the Qigong doctor is encouraged to constantly monitor his or her own actions and motivations. This is a broad and important topic that is often neglected in professional training and education. Because it is well dealt with in Kylea Taylor's book entitled, *The Ethics of Caring*, I encourage the reader to refer to this work for in-depth treatment of this subject.

CLINICAL BURNOUT AND MEDICAL QIGONG THERAPY

One classic problem all healers encounter is clinical burnout. The condition known as clinical burnout results from chronic job stress, and may sometimes occur to the Qigong doctor. This condition is characterized by extreme frustration, disappointment, physical and emotional exhaustion, and sometimes physical illness. The result of this condition is the doctor's loss of concern for the patients, or the loss of fulfillment with his or her performance as a healer. The Qigong doctor is susceptible to experiencing clinical burnout, especially if he or she feels that:

- There are too many or too few patients;
- Too little time is spent for recreation;
- There is too much stress in his or her life;
- Not enough attention is placed on self care and personal psychological growth; and
- Not enough attention is placed on a personal spiritual path and practice.

If the Qigong doctor's life revolves primarily around the clinic (which is usually the case), he or she will disproportionately come into contact with the patients' pain and emotional traumas, as well as be chronically bombarded with toxic energetic discharges released from the patients.

Sometimes patients, in order to heal their intense emotional traumas, require an enormous amount of spiritual, emotional, and mental support from the doctor. If the doctor does not carefully monitor his or her own energetic reservoir, this may result in Qi depletion.

If Qigong doctors expend all their spiritual/ emotional energy from intense personal interaction and enmeshment with their patients during treatment, the toxic interaction will deplete the doctors, resulting in complete physical and mental exhaustion. If Qigong doctors become depleted in the middle of a treatment, they may abandon their Upper Dantian (psychic) connection with patients (spiritually, emotionally, and/or energetically), and begin to suddenly express feelings of criticism, anger, disappointment, or grief. This reaction occurs when the Qigong doctors' own internal organ energy, depleted and worn, becomes susceptible to absorbing the patients' powerful emotional discharges. The problem of clinical burnout usually stems from energetic boundary invasion. Remember, Medical Qigong healing requires that the doctor heal him or herself first.

Often, a Qigong doctor's clinical burnout is primarily related to the unresolved personal issues and not the number of patients on the case load, or the amount of environmental stress the doctor experiences.

ENERGETIC BOUNDARY INVASION

In understanding clinical burnout due to energetic boundary invasion and its relationship between the doctor and patient interaction, it is important to review seven important factors which control and transform the body's energetic fields.

1. Energy exists on different planes. The body is like a complex hologram, composed of physical, mental, emotional, and spiritual planes of energy. The spiritual field of energy controls the emotional and mental energetic field, which in turn controls the physical energetic field. These energetic planes affect and control the energetic formation of the body's cells, tissues, organs, and organ systems.
2. Because the body produces heat, light, electricity, magnetic energy, and resonate vibrations, these natural energetic releases create a basic Yin (negative) and Yang (positive) polarity, which is found in all manifestations of energy.
3. Everything in the body is in vibratory motion. Various forms of tissue develop from different energetic vibrations. When the body's energy begins to slow down, it begins to take on a denser form of resonant vibration. The impediment of this resonant energetic vibration can stem from trauma or shock to the system (physical or emotional), resulting in energetic stagnations and disease.
4. The body's internal organs collect and store emotional energy, both positive and negative. As any suppressed, or "stuffed," emotional energy begins to culminate, the body will naturally (and quite frequently) disperse these internal emotional charges. They are generally released externally through: the Respiration, the Posture, and the Mental Dialogue.
   - The respiration, (which includes sighing and laughing), is released from the Heart, shouting released from the Liver, groaning from the Kidneys, crying from the Lungs, and singing from the Spleen.
   - The posture includes the body's physical structure adapting or conforming to an Expanded-Yang or Contracted-Yin energetic influence.
   - The mental dialogue includes both benign and malignant thought patterns.
5. The body has three primary energy centers known as Dantians. Each Dantian resonates at its own distinct vibration and is responsible for specific modes of interacting with others in interpersonal communication.
   - The Lower Dantian is considered the most physical energetic base. It is the most dense and has the slowest vibratory rate of the Three Dantians and communicates kinetically with the outside environment.
   - The Middle Dantian is considered the emotional and mental base. It has moderate resonances and communicates empathetically with the outside environment.
   - The Upper Dantian is considered the spiritual base and has the fastest and most delicate of the vibrational interactions. The Upper Dantian functions at an intuitive level.

6. When communicating, the body will interact from all five levels: physical, mental, emotional, energetic, and spiritual. This interactive communication will sometimes energetically fuse with the external environment, including people.
7. Our external and internal thoughts and feelings, as well as tissues and cells are all byproducts of energetic interactions, stemming from communication from within ourselves and our environment.

After reviewing the following facts, the Qigong doctor can understand why every time he or she connects with a patient, the patient's energetic fields will automatically be energetically ingested by the doctor. The doctor does this consciously in order to analyze the patient's emotional past, as well as present emotional outbursts. Although the doctor and patient's energies fuse together, the fields of energy around the doctor's internal organs generally protect the doctor from absorbing negative emotions. These energetic boundaries also extend out into the doctor's Wei Qi field.

Both the patient's and doctor's Wei Qi fields are fused during interaction and treatment. It is therefore imperative that the doctor's energetic boundary system maintain its structure (both internally and externally), yet be permeable while treating the patient. The doctor must not lose sight of his or her purpose. Any criticism aimed at the doctor should not be taken seriously, nor should any inappropriate behavior by the patient. The doctor's Wei Qi field should be at maximum strength. If not, the cords of energetic attachment flowing from both the doctor and the patient allow energetic emotional transference to happen. If the doctor, for example, has a preexisting Liver Heat condition, it will be further compounded by the inflowing of the patient's released of anger and rage. If and once this happens, the patient may feel relieved, but the doctor will become quite agitated.

STRENGTHENING THE DOCTOR'S ENERGETIC BOUNDARIES

It is important for the Qigong doctors to allow the patients' emotions to constantly flow through their own bodies, feeling, understanding, then releasing these feelings. If at any one point in this transition, a doctor (because of personal issues) suddenly stops the flow of emotions and begins to focus, or dwell on, a particular type of energetic charge, the patients' discharging emotions will invade the doctor's body.

To avoid energetic boundary invasion the doctor should maintain a regular Qigong meditation and exercise program. This program should include the following exercises and meditations.

1. Practice Qi dredging and emotional dispersing exercises and meditations to release your own emotional traumas.
2. Perform Qi and organ strengthening exercises and meditations to strengthen both your internal organ energies, as well as the three fields of Wei Qi. Since the emotions are stored within the body's internal organs, strengthening these organs will facilitate a stronger control of the body's emotional energy.
3. Practice Qi regulation and balancing exercises and meditations to balance the body's Yin and Yang, as well as internal and external energetic channels, organs, and organ systems. This helps to control the energetic surges released from internal emotional discharges.
4. Maintain emphasis on the energetic grounding and the divine hook-up meditations before treating patients. These meditations are important for establishing an emotional as well as spiritual boundary, and will assist you in avoiding clinical burnout and energetic boundary invasion.
5. Maintain emphasis on personal time, including any and all spiritual, social, and recreational times away from the clinic.

CAUSES OF SPIRITUAL DRYNESS

Sometimes clinical burnout can result from "spiritual dryness." This type of burnout can result from experiencing a "dry period" of spiritual insight and perception. It can occur during clinical observations and treatments, resulting in frustration for the doctor. Spiritual dryness is commonly due to one or more of the following five conditions.

1. If the doctor is tired, hungry, or sick, the fo-

cus will tend to be on the doctor's own particular needs instead of the patient's.

2. A disconnection from the divine is caused by guilt, lack of forgiveness (self-blame), shame, etc., resulting in a loss of spiritual vision. It is important for the Qigong doctor to have an open and honest connection with the divine to receive the more subtle spiritual insights. Without this type of transpersonal relationship, the Qigong doctor's ability to dissolve into the Wuji surrounding the patient's tissues can be hindered.
3. Rushing personal meditation time is caused by failing to relax during the preparation meditation practice (the divine hook-up). This condition is normally caused by worries or preoccupations due to hidden agendas, that disturb the doctor's Shen prior to meditation. It is important for the doctor to relax and return to a state of quiescent peace, in order to truly listen to the divine.
4. Getting into a rut is caused by the doctor performing repetitive Shengong meditations in a mechanical way (i.e., the same way an individual will routinely brush his or her teeth), thus losing the emotional and spiritual connection with the core self. The meditation then becomes meaningless and powerless (just words and no spirit). To avoid getting into a rut, the doctor should try to experience each meditation as a time of new beginnings and enlightenment, and not as a mandatory routine.
5. It is important for the doctor to be able to share emotional and spiritual insights with other colleagues in the same field of work, and when appropriate, with the patient. This sharing of spiritual insights, observations, fears, and triumphs renews the doctor's confidence in his or her spiritual gifts, and maintains a strong faith in the healing potential of Medical Qigong therapy.

SECTION VII

THE TREATMENT PRINCIPLES OF MEDICAL QIGONG THERAPY

CHAPTER 29

PRINCIPLES AND PARAMETERS OF MEDICAL QIGONG THERAPY

The foundational factors of the patient's state of health originate from their actions, thoughts, and feelings. A Qigong doctor, studying a patient's physical problems, understands that the origin and solution to these problems are to be found within a patient's recurring thought patterns, both conscious and unconscious.

The brain processes data consistent with what it is familiar with; patients will generally experience and accept information that coincides with their own belief system. Their consciousness then, creates its own reality, resulting in either health or disease.

Good health is a result of bringing the patient's own unique spiritual and emotional essence back into consciousness and increasing the awareness of the physical body. Both health and disease are created through this process which is the foundational base for clinical Qigong treatments.

THE FIVE METHODS FOR TREATING PATIENTS

Qigong doctors generally use five methods for treating their patients' diseases (Figure 29.1). These methods initiate an energetic restructuring within the patient's body. The five methods are described as follows.

1. The doctor uses his or her own Qi to initiate the healing.
2. Qi is drawn, by the doctor, from the surrounding natural environment (from the energetic fields of the Earth) to be emitted to the patients.
3. Qi is drawn, by the doctor, from the surrounding universe (from the energetic fields of the Heavens) to be emitted to the patients.
4. Divine energy is channeled through the doctor to treat the patients.
5. Homework is prescribed in the form of self-regulation exercises and meditations to support the healing process.

THE NINE STAGES OF TREATMENT

Using these five methods of treatment, the Qigong doctor initiates nine stages of healing within the patient's body. The healing process will always vary according to the belief structure of the patients, their acceptance and performance of the Medical Qigong prescriptions, and their willingness to change. The Qigong doctor's nine specific stages in treating a patient are described as follows.

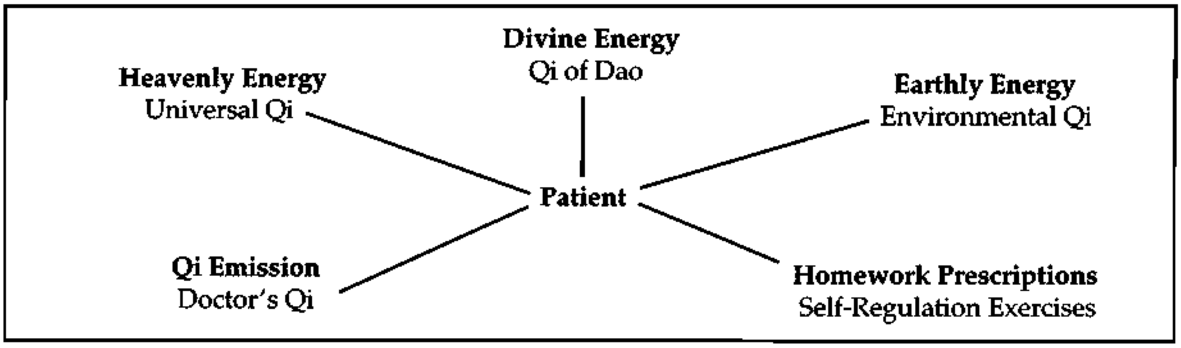

Figure 29.1. Five Methods Used For Treating Patients

The Subtle Wonders	Expensive or Contracted, Heavy or Weightlessness, Cold or Hot, and Tingling, or Vibrating Sensations
The Mysterious Wonders	Flashing Light Sensations, Sound Sensations, Smell Sensations, and Fainting (Syncope) Sensations
The Incredible Wonders	The patient will experience various colored lights, sounds, smells and other sensations according to the doctor's intention.

Figure 29.2. The Three Wonders

1. Activate, connect, and energize the patient's energetic field;
2. Dredge the patient's channels and collaterals;
3. Purge and eliminate any energetic Excess;
4. Tonify any energetic Deficiencies;
5. Regulate the Qi and Blood content of the patient's Yin and Yang organs (harmonize the organ Jing);
6. Regulate the patient's Prenatal and Postnatal Qi (harmonize the Qi of the Three Dantians and Taiji Pole);
7. Balance the patient's Yin and Yang Shen (harmonize the Heart/Mind and Emotions);
8. Seal the patient's external field of energy;
9. Give prescriptions.

THE THREE WONDERS OF MEDICAL QIGONG THERAPY

The cultivation of Qi and Shen through Medical Qigong therapy generates what is described as the Three Wonders. These Three Wonders consist of the three separate transformations of energetic abilities that the Medical Qigong doctor and, sometimes, the patients experience (Figure 29.2). The Three Wonders are explained as follows:

THE SUBTLE WONDERS

The Subtle Wonders consist of subtle changes inside and outside the body, known as the Eight Energetic Touches. These interactions of tissue and energy form the physical sensations and manifestations of Yuan Qi. When the doctor emits Qi to treat diseases, involuntary phenomena can occur instantly, or gradually, in local areas of the limbs, or the patient's whole body. In some cases the muscles stir slightly, while in other cases the limbs, or the body, will perform extreme, harsh movements. This results from the patient's external energy (Wei Qi) inducing spontaneous movements.

When the Qigong doctor emits energy to treat diseases, some patients produce, in local areas, electromagnetic, or stimulating, sensations of Qi similar to those felt in Qigong circulation/cultivation exercises and meditations. This includes: cold or hot sensations, feelings of compression or tugging, impressions of creeping or tingling, experiences of heaviness or lightness, and of floating or sinking, as well as other sensations. The Subtle Wonders result from energy in the channels circulating and responding to the doctor's intention.

Generally, the effects of Qi on the body are manifested in physical sensations and sensory phenomena during Qigong exercises. These phenomena are all manifestations of the free movement of Qi in the channels, collaterals, and inner body circulation.

The doctor may also experience these same sensations while extending energy into the patient. The doctor should be able to sense the patient's Qi circulation, density, direction, as well as his or her own energy. The sensations known as the Eight Energetic Touches of the Subtle Wonders are: Expansive, Contracted, Heavy, Light or Weightlessness, Cold, Hot, Tingling or Itching, and either Vibrating, Shaking, Rippling or Moving.

1. When the Qi is circulating freely within the body, the capillaries expand as the body's energetic field expands, resulting in the sensation of being very large, tall, and expanded.
2. When the Qi enters from the outside of the body to gather in the Lower Dantian, the patient experiences the sensation of feeling very small or contracted.
3. When the Qi sinks down (upon exhalation), the body feels heavy, condensed, and solid as a rock.
4. When the Qi rises upward (overflowing the channels), the body feels light and weightless.
5. When the Qi circulates through the Microcosmic Orbit (balancing the Qi of the Heart and Kidneys), the Kidney-Yang becomes sufficient to allow the Kidney Yin to well up; this results in a pleasant cool sensation.
6. When the Qi is vigorous and thermal energy is gathered, the parts of the body where the Qi passes will feel hot.
7. Once Qi circulates vigorously and passes through the body's obstructed channels, collaterals and minute collaterals, the result is itching or tingling of the skin and scalp. This is an energetic manifestation of the body's obstructed channels, collaterals, and minute collaterals being cleansed.
8. When the Qi passes through the channels, collaterals, and parts of the body it has opened, the patient feels electrical and creeping sensations in the extremities as well as the body, causing the tissues to vibrate, shake, ripple, or move, etc.

These are all normal phenomena of Qi cultivation and circulation. As long as the patient remains calm and present, these sensations will appear and disappear as spontaneously as they were formed.

THE MYSTERIOUS WONDERS

Patients may also experience elusive as well as dynamic transitions inside and outside of the body, called the Mysterious Wonders. These energetic changes are sometimes present, but are rare, and may be confusing to the patient. The sensations known as the Mysterious Wonders are: Flashing Light Sensations, Sound Sensations, Smell Sensations, and Fainting (Syncope).

1. Flashing Light Sensations sometimes occur in some patients during Qigong treatment. The photoelectric phenomena is due to the interaction of light with matter. In some cases, patients feel electric sensations in the limbs and body, and in other cases light patterns may be seen in different forms, most of which present as round, sheet light, or lightning patterns.

   One of the manifestations of the Qigong doctor's cultivation through Shen Gong meditations is a flashing light sensation which is seen and felt within the doctor's head and body. This glow may sway back and forth, pulse like a white neon light, or pierce into the doctor's brain, and illuminate the spinal column. It may also appear as a ball of white light rotating at high speed through the Microcosmic Orbit.

   This ball or beam of light may exist continuously at the Qigong doctor's Baihui GV-20 point throughout the duration of the meditation and several hours afterwards. The degrees of light sensations and their colors will vary in accordance with the doctor's skill level. Also, it is this beam of light that the Qigong doctor will later utilize for inner vision diagnosis.
2. Sound Sensations are sometimes experienced during treatment, such as a rustling, rumbling, or high-pitched sound.
3. Specific Smell Sensations may occur, such as the sweet scent of sandalwood, or fragrances of various flowers, or incenses.
4. Fainting (Syncope) can occur during Qigong treatment. A few patients may suffer from perspiration and accelerated heart rate, fol-

lowed by syncope (fainting). When this occurs, the doctor lightly presses the patient's channel points: Baihui (GV-20), Mingmen (GV-4), Jiangjing (GB-21), and Yintang (Extra Point). The doctor then grasps the shoulders and presses down on the GB-21 points with force. Finally, the doctor emits Qi along the Governing and Conceptual Vessels and leads the patient's Qi to the Lower Dantian. This allows the patient to recover consciousness.

THE INCREDIBLE WONDERS

The Qigong doctor's ability to have complete conscious and automatic control over the patient's reaction to energetic projections constitutes what is known as the Incredible Wonders. When the Qigong doctor takes control over the patient's energetic projections, he or she can willfully initiate the patient to experience various colored lights, sounds, smells and other sensations according to the Qigong doctor's will. This procedure is commonly known as Cultivating the Spirit to Generate Wonders.

When the Qigong doctor emits energy towards patients to treat their diseases, most patients will either instantly or gradually feel some form of energetic sensation from the increased life-force energy. The degree of energetic sensation felt by patients is in accordance with their emotional/spiritual sensitivity and degree of connection to their own physical body.

The sensations of Yuan Qi are the most common reactions felt by both the Qigong doctors and their patients. In rare instances some patients display no physical effects and experience no Qi sensations during treatment. These patients can still achieve good therapeutic results by improving the degree of their emotional/spiritual connection to their body.

In some cases, patients may even feel miserable or uncomfortable. When this occurs, they should continue training so that their Righteous Qi can overcome their Evil Qi. This ongoing battle will also affect their spiritual and emotional being. Fear usually arises at this point causing them to question their ability to continue in their practice. This is a normal transition and can be resolved through encouragement, patience and continual practice.

CONTRAINDICATIONS FOR MEDICAL QIGONG THERAPY

There are several contraindications that the Qigong doctor should be aware of before treating patients, or even prescribing Medical Qigong exercises. The following is a list of different personality types that may be encountered in the clinical settings that require special treatment modifications.

1. The very "nervous" type of patient should not practice Medical Qigong, or must be under strict supervision due to the patient's emotional instability.
2. The "extreme temperament" type of patient who has a fanatical type of personality should also avoid Medical Qigong training or remain under strict supervision. Patients of extreme temperament generally suffer from Excess disorders and compulsive overtraining. They are therefore prone to Qi deviations.
3. The "extremely introverted" type should be careful when practicing Medical Qigong meditations because strong emotions may be set into motion, and even certain mental disorders may be violently released that the patient may have been unaware of.
4. The "narrow minded" type will generally condense Qi into smaller areas of the body causing stagnations. Such patients should be instructed to relax and let go of their stress to allow the condensed energy to disperse.
5. Patients with the following conditions should be restricted to very specific Medical Qigong therapy: serious mental disorders, congestive Heart failure and severe organ failure, severe hemorrhaging, menstruation, nerve suppression, bone fracture, pregnancy, geriatric patients, young children, and cancer patients (Figure 29.3).

- Patients with mental disorders such as schizophrenia and bipolar (previously called manic depression) disorders (or individuals with a family history of such mental diseases), should refrain from training. The Qigong exercises and meditations could induce symptoms of these said disorders, or could enhance

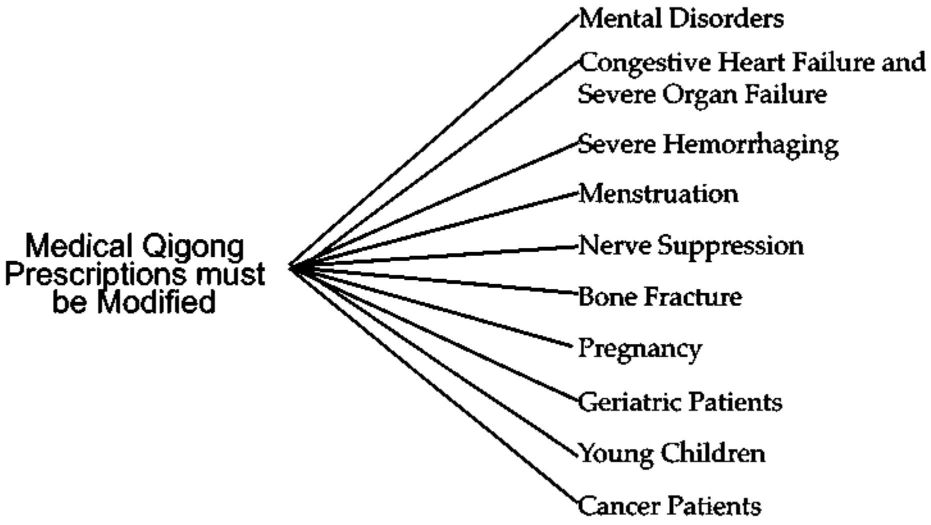
Medical Qigong Prescriptions must be Modified
Mental Disorders
Congestive Heart Failure and Severe Organ Failure
Severe Hemorrhaging
Menstruation
Nerve Suppression
Bone Fracture
Pregnancy
Geriatric Patients
Young Children
Cancer Patients

Figure 29.3. Restricted Conditions for Prescribing Medical Qigong Exercises and Meditations

the symptoms due to the patient's mental and emotional instability.

- Patients with congestive Heart failure and severe organ failure (any internal organ that has experienced an energetic collapse), require that the treatment be specifically oriented towards this condition.
- Patients with severe hemorrhaging require prescriptions that prevent the leakage of Blood and Qi, since Blood and Qi flow together.
- Women during their menstruation cycle should not store Qi in their Lower Dantian. Any Blood and Qi stored in the uterus will purge from the woman's body naturally, drawing with it any Qi stored within the Lower Dantian.
- Patients suffering from nerve suppression (cervical spondylosis) may experience a worsening of their condition. In severe cases, the additional pressure on the nerve roots may complicate the patient's condition, especially when the body's energy field expands causing the Marrow within the spinal column to ascend.
- Patients with bone fractures must first have the bones set correctly and immobilized to prevent further injury. The subtle involuntary movements of the limbs during Qigong treatment or practice can disturb the bone's healing process.
- Pregnant women must be treated by a Medical Qigong doctor who specializes in obstetrics, otherwise the risk of involuntary abortion (miscarriage) or other complications can arise.
- The treatment of geriatric patients prohibits Dynamic Qigong because of the delicate nature of their internal tissues. All exercises and meditations should be kept slow, quiescent, and tranquil.
- The treatment of young children requires that all exercises and meditations be age appropriate, and be specifically oriented toward healing his or her disease. Some Qigong exercises, meditations and treatment methods may interfere with the natural development of the child's internal organs and tissues.
- The treatment of cancer patients requires modified prescriptions in accordance to the specific disease. With certain cancers, Dynamic Qigong is contraindicated to prevent the cancer from further metastasizing.

CHAPTER 30
BALANCING THE BODY'S ENERGETIC SYSTEM

PURGING, TONIFYING, AND REGULATING

In Medical Qigong therapy, successful results can be achieved only when the doctor adheres to the principle of "treating Deficient syndromes with tonification and Excess syndromes with purgation." Tonification supplements the insufficiency and strengthens the body's resistance, while purgation reduces the Excess and expels pathogenic factors. The final aim is to balance Yin and Yang to regulate the physiological functions of the body.

Medical Qigong Therapy adheres to the traditional Chinese medical explanations of illnesses and general principles of healing. Only the techniques used during treatment differ. Qigong doctors Purge the patient's body, then emit Hot or Cold Qi in conjunction with the Five Elemental energies, using various hand manipulations, including: pushing, pulling, rotating, leading, and shaking.

Through energy extension, the doctor can reinforce the patients body's resistance to pathogenic factors, replenish the Yin and Yang organ energy and enhance mental clarity.

THE THREE TREATMENT MODALITIES

Medical Qigong therapy is divided into three treatment modalities (Figure 30.1). These treatment modalities enable the Qigong doctor to purge, tonify, and regulate the patients' body. These techniques are described as follows.

1. Medical Qigong Purging techniques are used by the doctor to treat the patient's Excess syndromes, and to detoxify the body of pathogenic Qi. The "lower aspect" of classical Chinese medicine governs the treatment of the patient's illness and corresponds to Earth. It involves purging pathogenic influences and dispersing stagnations in order to cure the patient's illnesses. This allows the patient the ability to remove Excess Qi from organs and tissues.
2. Medical Qigong Tonification techniques are used by the doctor to treat the patient's Deficient syndromes, as well as strengthen and stabilize the organs and organ systems. The "middle aspect" of classical Chinese medicine governs the nourishment of the patient's nature and corresponds to Man. It involves helping the patient prevent illness by tonifying

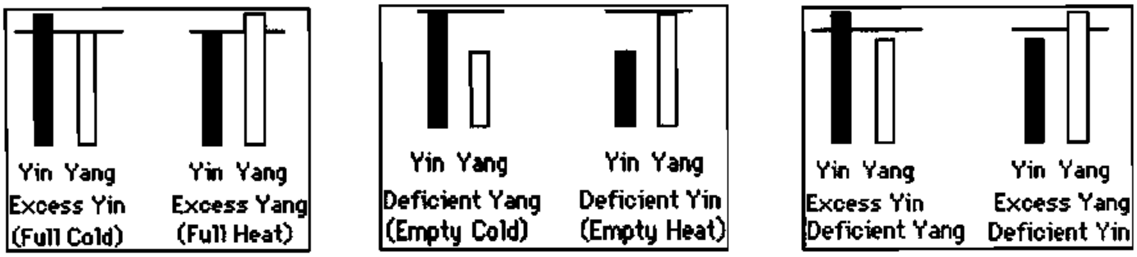

In cases of Excess, the doctor should Purge.

In cases of Deficiency, the doctor should Tonify.

In cases of Excess and Deficieny combined, the doctor should Regulate.

Figure 30.1. Purgation, Tonification, and Regulation

Heaven —— Regulation ——— Balance Qi and Blood —— Nourish the Patient's Destiny

Man ——— Tonification —— Treat Deficient Conditions — Strengthen the Body to Prevent Illness

Earth —— Purgation ——— Treat Excess Conditions —— Detoxify the Body to Treat Illness

Figure 30.2. Ancient Chinese Perspective on Purgation, Tonification, and Regulation

organ energy Deficiencies. This allows the patient the ability to replenish organ energy depletions and tissue emaciation.

3. Medical Qigong Regulation techniques are used by the doctor to balance the patient's Yin and Yang energy, as well as organ Qi. The "highest aspect" of classical Chinese medicine governs the nourishment of the patient's destiny and corresponds to Heaven. It involves regulating the patient's Three Dantians, Taiji Pole, Yin and Yang organs and energy fields. Regulation techniques promote a balance of Jing, Qi and Shen for the optimal development of all the aspects of man, i.e., body, mind, emotion, spirit and energy. Thus, in China, Medical Qigong regulation is said to "nourish the patient's destiny" (Figure 30.2).

PURGATION TECHNIQUES

The patient's body can be dredged and purged by using two different expelling methods. The first method is the direct removal of pathogens from the patient's tissues and organs. The second method is the indirect purging of the patient's tissues and organs through the channels and points of the patient's body. These two methods serve two different purposes: The Direct Removal of Pathogenic Factors, and The Purgation of Channels and Points.

1. The Direct Removal of Pathogenic Factors is achieved through extending Qi into the tissues and organs where they are lodged, and dredging the Toxic Qi directly out of these locations. This method is used to expel pain due to Excess Heat or Cold. The various techniques used in this type of purging employs stationary Qi-guiding manipulations. The various hand manipulations used for the direct removal of pathogenic factors consists of such techniques as: The Five Thunder Fingers Hand Posture, Shaking and Trembling Hand Posture, Bellows Palm, Vibrating Palm, Dragon's Mouth Palm, Tiger Kneading Palm, and Spiraling Palm (see Chapter 33).
2. The Purgation of Channels and Points techniques are used to expel External pathogenic Evils (such as the invasion of Wind, Fire, Heat, Dry, Damp, Cold, and environmental toxins, etc.), or Internal pathogenic Evils (such as fear, anger, and grief, etc.) from the organs and tissues of the body. It is also used to redirect Rebellious Qi (e.g., Stomach Qi). The various techniques used in this type of purging employs dynamic-linear Qi-guiding manipulations such as the Pulling, Leading, and Guiding methods (see Chapter 34).

TONIFICATION AND PURGATION METHODS

The doctor should always purge before tonifying the patient's body. In the clinical setting, these two methods are often used together in combination with other methods in treatment. These combined methods are as follows.

TONIFICATION AND PURGATION WITH RESPIRATION

Tonification occurs as the doctor exhales while guiding and directing Qi into the channels, points, and the internal organs and tissues of the patient's body. Purgation happens when the doctor inhales while guiding, emitting, and leading the patient's Qi, or pulling the pathogenic Evils out of the patient's body.

During treatment, it is important that the patient cooperate with the doctor by matching (if possible), per the doctor's instruction, his or her breathing rhythm to the doctor's treatment. The patient should assume a proper posture (lying, sitting or standing), relax, and concentrate his or her mind on the location being worked on. When the doctor exhales and emits Qi toward the patient, the patient should inhale to absorb the doctor's Qi into the designated location. When the doctor performs purgation, the patient should exhale to release Toxic Qi out of his or her body.

TONIFICATION AND PURGATION ALONG AND AGAINST THE CHANNEL FLOW

With regard to acupuncture therapy, the classics say that, "puncturing along the direction of the channels yields a tonifying effect, while going against it induces purgation." In Medical Qigong, however, due to the Qigong doctor's nature of emitting, leading and guiding the energetic flow of Qi, certain modifications are used. The following are a few examples of tonification and purgation techniques used along and against the energetic flow of the channels:

- Causing Qi to flow up the Yin channels of the legs and into their associated organs in the torso is considered tonification (think of following the river to the sea). One example of this type of treatment therapy is emitting energy into the base of the patient's feet (at the Kd-1 points); this causes the energy to flow into the torso to strengthen the Kidneys.
- Moving Qi from the torso down the legs against the flow of the leg Yin channels is considered purgation (following the river from the sea). One example of this treatment therapy would be pulling Excess Qi from the Liver organ, down the right Liver channel and out the feet to sedate the Liver Fire (the Gall Bladder Channels can be used to purge Excess energy from the Liver organ down the leg Yang channels).
- Moving Qi against the flow of the arm Yin channels into their associated organs is considered tonification (think of following the river to the sea). One example of this treatment would be emitting energy from the base of the thumbs, at the Lu-11 points, into the torso to strengthen the Lungs.
- Moving Qi against the flow of the arm Yang channels into their associated organs is considered purgation (following the river from the sea). One example of this treatment would be pulling Qi against the flow of the Large Intestine Channels down and out the hands to purge Excess Qi in the head.

TONIFICATION AND PURGATION POINT THERAPY WITH THE HORARY CYCLE

When prescribing Jing Point therapy as a means of self-regulated homework, the Qigong doctor encourages the patients to follow the formula of tonification and purgation according to the energetic flow of the Horary cycle. According to ancient Chinese medical theory, the ebb and flow of Qi and Blood along the different channels is related to designated days and hours (see Chapter 11). This theory applies to two specific tissue manipulations: Holding the Point for Tonification, and Closing the Point for Purgation.

- Holding the point for tonification requires the patient to use the Massage Tapping method (see Chapter 37) on specific points two hours prior to its opening or peak flourishing time (high-tide).
- Closing the point for purgation of the internal organs or channels requires that the Qigong doctor drain the diseased area at high-tide (when it is most full of Qi and Blood). When the patient is applying Jing Point therapy for homework prescriptions, however, the application is quite different. This technique requires the patient to use the Massage Tapping method on specific points two hours after its opening time.

The patients may also be instructed in self-massage using the Tapping method for self-healing. If the patient has a Kidney disease, for example, the specific points should not be tapped during its peak time (5–7 p.m.), but rather two hours prior to its peak time (3–5 p.m.). This will improve the Kidneys' function.

TONIFICATION AND PURGATION GUIDELINES

When the Qigong doctor purges and tonifies a patient, he or she must remember to adhere to the following guidelines carefully.

GUIDELINES FOR THE DOCTOR

When treating the patient, the doctor must always consciously direct and be aware of the movements of each hand.

- Generally, the right hand is used for tonification, while the left hand is for purgation. Diseases that require purgation are treated by draining Toxic Qi with the left hand. Diseases that require tonification are treated by extending Qi through the right hand.
- Turning hands towards the body is used for tonification, while turning the hands away from the body is used for purgation.
- Moving the hands away from the body and upward can regulate the Yang energy, causing it to rise, and is inclined to tonify. Moving the hands downward pulls the Yang energy down, and is inclined to purge.
- Clockwise drilling (the palm or extended fingers move in a clockwise direction over an area) is used for tonification and increasing Qi; counterclockwise drilling is used for purgation and decreasing Qi.

GUIDELINES FOR THE PATIENT

The patient, (like the Qigong doctor), must be aware of hand movements when practicing Qigong prescriptions. In addition, the patient must be aware of his or her posture, respiration, and mental attitude.

- The eyes are open for purgation; closed eyes are used for tonification. The patient should apply these principles according to the symptoms of the disease.
- Inhalation is used for tonification, exhalation is used for purgation. During purgation treatments, the patients assume the proper postures, relax, and concentrate their mind on the disease area while exhaling the pathogenic Qi. This exhalation technique is used not only to promote the removal of the surfacing toxins, but is also used to enhance the doctor's efficacy in purging the diseased area.
- The audible and silent tone resonations (see Chapter 40) also have the function of either tonification and purgation. Speaking the silent tone when inhaling can be used for tonification; while exhaling, the tone is used for purgation.

UNDERSTANDING MEDICAL QIGONG REGULATION

The Qigong doctor must master three main techniques of Qi regulation to facilitate clinical healing through Qi emission. These techniques include: Training the Qi, Guiding and Inducing, Tonifying and Purging the Qi, and Emitting the Qi.

1. Training the Qi exercises and meditations are used to gather Qi and to promote its circulation within the body. The Qigong doctor must have an extensive comprehension of several Medical Qigong techniques used to gather and circulate energy throughout the body. This includes a complete understanding of Postural, Respiratory, and Mental Dao Yin modalities, as well as the static and dynamic clinical modalities. These training techniques provide the Qigong doctor with the ability to regulate and balance the patients' Qi, and set the foundation for the patients' Medical Qigong prescriptions.
2. Guiding and Inducing, Tonifying and Purging the Qi are used to circulate Qi within the external field of energy (Wei Qi). These techniques require a complete understanding of guiding and inducing (leading the Qi), and tonifying and purgation techniques, to being able to rectify Qi deviations and remove energy blocks.

These training techniques provide the Qigong doctor with the foundation for the patients' Medical Qigong therapy.

3. Emitting Qi techniques must be learned and perfected by the doctor. The Qigong doctor must have an extensive understanding of the multiple techniques used to extend Qi into the patient's body. These include electromagnetic energy, light, sound, and hot and cold Qi projections. These training techniques and abilities also set the foundation for the patients' Medical Qigong therapy (see Chapter 34).

QI REGULATION AND GUIDING PRINCIPLES

Qi regulation is considered a combination technique for reinforcing and reducing. It is used to readjust and balance the Yin and Yang energy. If the patient's Qi rises adversely, the doctor should guide and secure it downward. The same principle is applicable for any Deficiency or Excess of Qi in the left or right sides of the patient's body.

Qi regulation also includes purgation after tonification, and tonification after purgation. This method of guidance is used when Qi stagnation, stasis of Blood, or disorders of the channel Qi develop. Under pathogenic conditions, Qi of the viscera and channels may become weak and stagnant. This is manifested by symptoms such as listlessness, general weakness, and the hypofunction of certain organs or tissues. Stagnation of Qi and stasis of Blood can also obstruct the channels causing pain. Pain may also be caused by outside pathogenic factors such as trauma, inflammation, muscular spasms, and adhesive degeneration. In these circumstances, the channels are dredged to promote Blood circulation, and to remove the stasis in order to relieve spasms, and dispel pain.

Every organ in the body has its own Qi that circulates and interacts with the rest of the body. When the normal function of the organ is affected, the circulation of Qi may become disturbed. This disturbance of Qi in one organ may influence the functional balance of the whole Qi system.

The doctor regulates the patient's channel Qi by dredging the channels and collaterals. This facilitates Qi circulation and leads Qi back to its origin. In the case of a weak and hypofunctional Lung Qi, for example, emitting Metal-Qi plus Pushing, Shaking and Leading manipulations (see Chapter 34) can be adopted to activate the Lungs' Qi. The combination of these techniques causes the channels' Qi to overflow freely, thus improving the function of the Lungs.

ENERGY BALANCING PROCEDURES

When extending energy into the patient the main emphasis is placed on balancing the patient's energetic fields. The doctor must have both physical and emotional releases from the patient's body, to detoxify and liberate the trapped Qi.

It is important to not let the patients avoid facing their problems; keep them centered around the troubled area, focused on the awareness of what they are feeling. Basic energy balancing procedures are as follows.

1. If the treatment area is Hot, cool it down. The doctor must continue treating the patient until the Hot area has cooled. A strong and quick moving hand gesture is best used to treat a Heat syndrome.
2. If the treatment area is Cold, heat it up. For a Cold syndrome, use a longer treatment to warm up the channel in the area treated, to eliminate the Cold.
3. If the tissue's energy is dark or cloudy, change it to light and clear.
4. If you feel tingling on any area or organ, remain there until the tingling subsides and the organ or area completes its energetic unwinding. Tingling in an area indicates that the patient's body is releasing an energetic cluster. The patient may or may not be aware of any feelings or sensations when this occurs.
5. If the area is emitting electric shocks, dampen or sedate it by connecting it with its associated organ. Connecting Yang (positive) into Yin (negative) will ground the energy.
6. If you sense pressure or Excess, dredge the area, and circulate the Qi.
7. If you feel intense pulsation, decrease the intensity and make it rhythmical.
8. When treating an energetic field that is De-

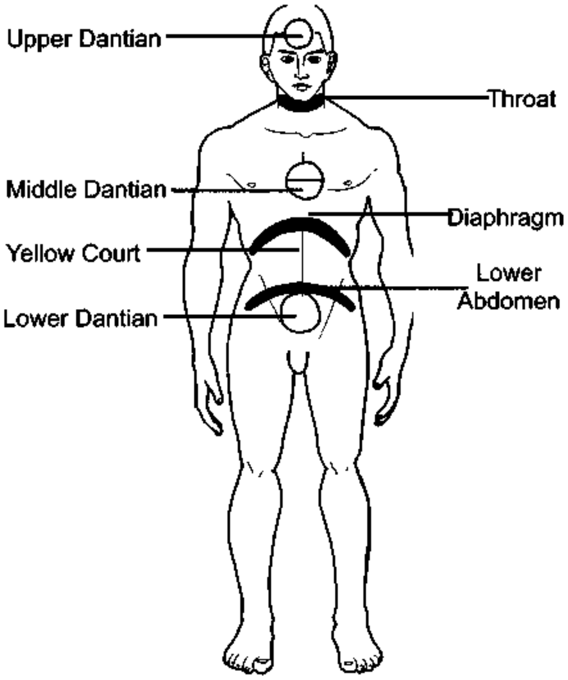

Figure 30.3. Locations of Energetic Blockages

pressed (sinking or contracting inward), use an "ascending method" of Qi manipulation for tonification (i.e., move and emit Qi from the lower areas of the body upwards). Some examples of this type of ailment are:

- Lower Burner Depression
- Middle Burner Depression
- Immune System Depression and Breakdown
- Autoimmune Diseases (Lupus and Fibromyalgia)
- Chronic Fatigue Syndrome
- Heart Failure
- Kidney Failure
- Prolapse of the Stomach

9. Assign Medical Qigong prescriptions for patients, including meditations and exercises, as needed.
10. Treat patients with severe diseases by:
    - Stopping and controlling the spread and development of the disease,
    - Stopping the patient's Excess consumption of Qi,
    - Treating the cause and root of the disease, after the patient's condition is under control, and
    - Assigning homework.
11. For chronic diseases, use the method of purging and dredging stagnation to move any impacted Qi.
12. For sprains, use Qi compression (using the Qi to press the tissues) around the local area. Close the channels surrounding the sprain first, isolating the trauma. Then pull out the stagnant Qi. If the doctor immediately treats the local area, he or she will disturb and further complicate the injury. It is important to close off the surrounding channels around the wound first, leaving only the affected channel open, before pulling out the stagnant Qi. This controls the flow of energy in to the stagnated area.

Check the patient after seven treatments to determine if the therapy has been effective. Reevaluate each patient according to each individual's constitution.

CLINICAL OBSERVATIONS OF PATHOLOGICAL ENERGETIC PATTERNS

When the Qigong doctor begins to treat patients, certain energetic observations and phenomenon will be experienced while diagnosing the energy inside the patient's tissues. These energetic experiences are quite different from the Eight Energetic Touches which focus on the feelings of the body after meditation practice (see Chapter 29). The Internal conditions of the patients' disease, its energetic nature, as well as its clinical manifestation, are expressed through the following: Hot, Cold, Deficient, Excess, moving, stagnant, Damp, Dry, Clear (Clean), Turbid (Impure), open (accessible), closed (armored), and leaking.

LOCATIONS OF ENERGETIC BLOCKAGES

In clinical practice we generally observe six main areas in the patient's body where energy becomes trapped and blocked. Each energetic blockage is an expression of the patient's ability to armor and protect him or herself from emotional wounding. These six main areas in the patient's body can be the foundational base and the root cause for the patient's disease (Figure 30.3). These six main areas are described as follows.

THE AREA OF THE HEAD

The area in the cranium that is most commonly obstructed is within the base of the occiput, at the back of the head. This obstruction is usually noted within patients who suffer from migraine headaches and brain tumors. This barrier is often created through chronic stress patterns which cause the Liver's Heat to rise up the patients' back and become trapped inside the head. This condition is usually caused by intense concentration. Too much hard focusing can also cause Qi to condense and stagnate. The location of Qi obstructions due to physical trauma vary according to each type of trauma, and the and the severity of the injury.

THE AREA OF THE THROAT

This area separates the Upper Dantian (the spiritual/perceptive center) from the Middle Dantian (the emotional/empathic center). Messages of divine inspiration and awareness descend downward from the Heavens through the Baihui area seeking to stir the emotions of the Heart. If the patient is not receptive (due, for example, to a disconnection or even a resistance to spiritual growth) a barrier is sometimes formed within the patient's throat, dividing the Upper Dantian from the Middle and Lower Dantians.

This barrier can also be created to stop emotions from flowing through the Heart and rising upward to be released through the windows and caverns of the soul (the eyes, ears, nose, tongue, body, and spirit). If, for example, the teeth become clenched and the throat tightens from feelings of grief or anger, these emotions can become trapped in the throat, forming a "plum seed" or "knot." When the Lungs' Qi congeals in the areas of the upper chest, shoulder blades, neck and the back of the occiput, this indicates chronic suppression of these emotions. It is also evidence of the mind and body struggling to reconnect and express the emotional and spiritual self.

THE AREA OF THE HEART

The Heart is energetically divided into two main chambers. The upper chamber houses and responds to spiritual emotions, while the lower chamber houses and responds to sensual emotions.

Patients will tend to suffer not only from repressing their basic biological drives (the sensual emotions based in the lower chamber of the Heart), but also by repressing their emotional connections with the divine (by failing to accept their higher spiritual calling or spiritual enlightenment associated with the upper chamber of the Heart). The higher self is the focal point of the superconscious realm; it is all-inclusive and at the very center of our being where individuality and universality blend (center core).

Patients who have chronically armored their Heart tend to exist in a state of guarded suspicion. This pattern of emotional suppression tends to lead to Qi stagnation and the creation of disease. When patients armor their upper chamber, they tend to hide in the emotions of their lower chamber and visa versa. Patients who armor the upper chambers of the Heart, for example, may become involved in relationships of extreme sensual passions related to the Heart's lower chamber, to avoid painful unresolved issues hidden within the past. The deeper and more energetically charged the issues, the stronger the sensual passions. Whereas, patients who armor the lower chamber of the Heart, might become involved in relationships of extreme spiritual, or religious passions related to the Heart's upper chamber to hide from their painful unresolved issues. The deeper and more energetically charged the issues, the stronger the spiritual or religious passion.

THE AREA OF THE DIAPHRAGM

The diaphragm area separates the Middle Dantian (the emotional center) from the Lower Dantian (the physical/sexual center). When a barrier in this particular area is created, it separates the emotional Heart from feeling the sexual passion which rises up from the Lower Dantian and the Kidneys' area.

It is interesting to note that patients with a

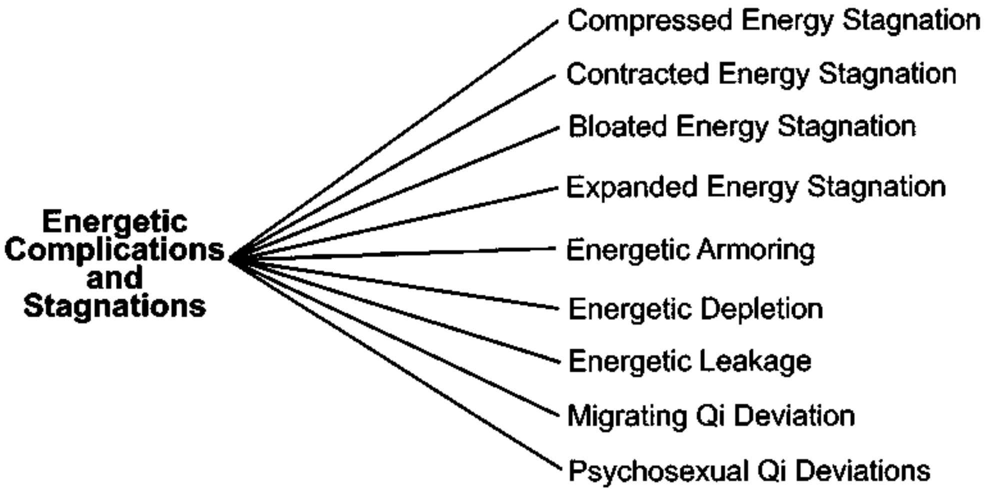
Energetic Complications and Stagnations
Compressed Energy Stagnation
Contracted Energy Stagnation
Bloated Energy Stagnation
Expanded Energy Stagnation
Energetic Armoring
Energetic Depletion
Energetic Leakage
Migrating Qi Deviation
Psychosexual Qi Deviations

Figure 30.4. The Body's Energetic Complications and Stagnations

barrier in the diaphragm sometimes enjoy the physical sensations of sexual intercourse, but are separate from their spiritual/emotional connection to true bonding or "oneness" with their partner. Again, the connection between the emotional spirit and body/mind has been severed.

THE AREA OF THE YELLOW COURT

The Yellow Court (CV-12) is an area located in the center of the chest, just below the diaphragm and xiphoid process of the sternum. Its function is that of being the access area to release the body's internal organs' emotional memories. Its location is also attributed to the Third (solar plexus) Chakra.

To access the patient's stored emotional memories and release them from the Yellow Court, the Qigong doctor supports the patient's Shendao GV-11 (Spirit Gate) with the left hand while energetically manipulating the patient's Yellow Court area with the right hand. If the patient is not energy sensitive, the doctor uses a slight Compression/Release technique (see Chapter 37) to disperse the patient's armoring system.

The patient should be in a supine position, with knees pointed upward, feet flat, arms by his or her sides, mouth open, inhaling deeply into the Lower Dantian. As the patient exhales, the doctor lightly compresses the patient's Yellow Court until a deep groaning sound begins to resonate from the patient's thorax. The doctor circulates the energy from the patient's Yellow Court into the Shendao, filling up the patient's Heart with energy. The doctor then moves this energy back into the patient's Yellow Court. This process continues until the patient's suppressed emotions surface and are released.

Note: It is important to first purge and release the Lungs of grief before attempting to access the patient's Yellow Court.

THE AREA OF THE LOWER ABDOMEN

This area also separates the Lower Dantian from the Middle Dantian, insulating the sexual feelings from the middle and lower torso. This energetic freezing of the pelvis is usually found in victims of rape, sexual molestation, or in women who have had one or more abortions. The Qi flowing into the lower pelvis and legs becomes blocked due to stored fear, shame, guilt, and sorrow. Sometimes the patient's physiology in that particular area may become obese; the abdomen, thighs or hips may be heavily padded and armored due to Qi stagnation. Often times, victims of childhood sexual abuse become overall obese to avoid attracting the opposite sex, or to block and pacify these hidden traumatic memories.

TYPES OF ENERGETIC COMPLICATIONS

Each organ and channel has the potential to carry or create several types of energetic complications and stagnations. The doctor should observe and feel the shifts in the patient's energetic balance to determine the specific conditions of the disorder (Figure 30.4).

COMPRESSED AND CONTRACTED ENERGY STAGNATION

Compressed or Contracted energy stagnations are sometimes considered a Yin condition due to the type of energetic compression involved in its formation.

- Compressed Stagnation is caused by the patient's energy pressing downward (externally moving inward) and will feel armored and hollow (Figure 30.5). When this type of energetic stagnation disperses it does not energetically unwind, but disperses and dissolves into the surrounding tissue areas. Because the patient's surrounded consciousness has armored the tissue area, the energetic dispersion can sometimes cause spontaneously released emotions or behavior such as crying, laughing, shouting, screaming, sighing, coughing, etc.
- Contracted Stagnation is caused by the patient's energy pulling inward, and will generally feel armored and solid (Figure 30.6). This type of energetic stagnation can be formed by an emotional core, or focused concentration, that draws additional Qi into its center, compacting the belief structure, layer after layer with energy. When this type of energetic stagnation disperses it can sometimes begin energetically unwinding throughout the entire body. This energetic unwinding sometimes creates involuntary muscle twitches and physical undulations.

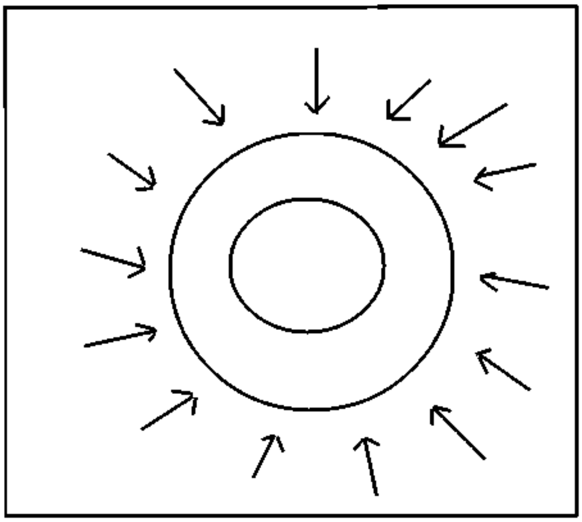

Figure 30.5. Compressed Stagnation

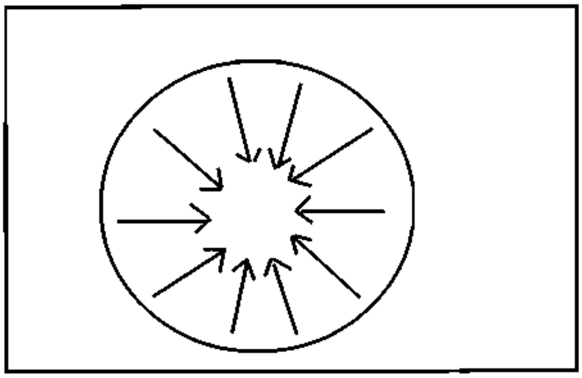

Figure 30.6. Contracted Stagnation

Body tissues and muscles accumulate emotional energy, becoming hard and tense in the adjacent areas. Denseness indicates the accumulation of pathogenic stagnant Qi. If this energy block remains, disease can manifest as ovarian cysts and tumors, or even in extreme cases as cervical cancer. Breast cysts, tumors, and cancer can result from stagnant Qi in the Lungs, or from a combination of Lung and Liver Qi stagnation. This type of stagnant energy feels very Hot and volatile because of the accumulated anger and rage.

Treatment Principals for Compressed and Contracted Energy Stagnation

The Qigong doctor disperses and purges the patient's stagnant Qi using the Thunder Palm technique, Tiger Kneading techniques, and counterclockwise Circular Spiraling (see Chapter 34).

BLOATED AND EXPANDED ENERGY STAGNATION

Bloated and Expanded energy stagnations are sometimes considered a Yang conditions due to the type of energetic expansion involved in their formation. Qi stagnation, with a characteristically expansive or bloated appearance, can be caused from an accumulation of Phlegm and Body Fluids (in addition to Qi and Blood) in the adjacent tissue areas of the body (Figure 30.7 and 30.8). This type of stagnation usually occurs in the Lower Dantian and abdominal areas. These are examples of Excess syndromes of stagnant energy and manifest as: distending, turbid, leaking, tingling, and dense Qi.

If this type of energy stagnation remains obstructed, the resulting disease can manifest as edema within the tissues, colitis, or angina pectoralis due to unprocessed despair and grief (in the Lungs) with anxiety (in the Spleen). The energy of grief and despair feels thick, heavy, and slimy like phlegm or mucus. Often this energy is felt in the Lungs, as well as in the Spleen.

Treatment Principals for Bloated and Expanded Energy Stagnation

The Qigong doctor disperses or dissolves the patient's stagnant Qi by using the Tiger Kneading, Vibrating Palm, and Shaking Palm techniques (see Chapter 34).

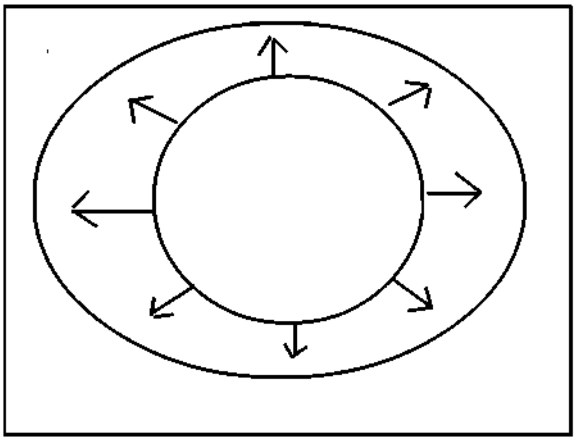

Figure 30.7. Bloated Stagnation

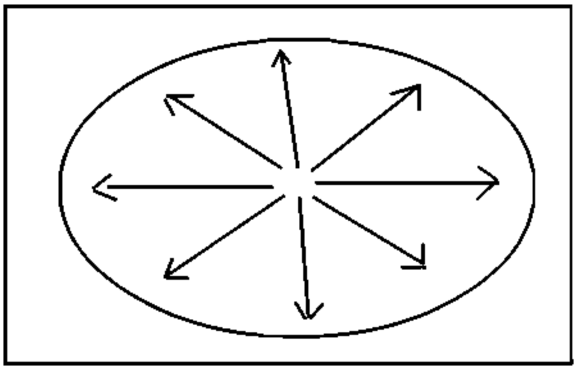

Figure 30.8. Expanded Stagnation

ENERGETIC ARMORING

This condition can result from the patient protecting specific tissues, organs, or areas of the body. Energetic armoring is initiated when the patient freezes certain emotional feelings to maintain the denial system. Patients may be unaware of this armoring condition because they dissociate from the original trauma, or minimizing its severity. Emotions attributed to energetic armoring are generally manifested as anger, fear, guilt, and shame. The angry patient's denial system is maintained through physically or emotionally attacking others; the anxious patient, full of fears, guilt, and shame, copes by "Shening out." Both of these reactions shield the patients in emotional situations from their own feelings, while reinforcing the energetic armoring.

Armoring results in tense organs and tissues. When patients cannot feel their body, other complications can arise, such as compressed, contracted, bloated or expanded energetic stagnations.

Treatment Principals for Energetic Armoring

The Qigong doctor can assist the patient through "soul retrieval" to unlock the feelings logged deep within the body's tissues (see Chapter 19). The primary goal is to first help the patients acquire a sense of security within their own body, then address the root of the disease.

ENERGETIC DEPLETION

Whole body depletion results from the following conditions:

- Not receiving enough Postnatal Qi for restoration, resulting in improper internal organ functioning,
- From suppressing emotions (e.g., worry or fear),
- Experiencing a damaged energetic field due to trauma, or
- Being born with a Deficient prenatal constitution.

Generally, energetic depletion can result in physically undeveloped areas in the body. In some cases, emotional traumas can cause specific organs or tissues to become energetically depleted to the degree that their natural development cannot take place.

Treatment Principals for Energetic Depletion

The Qigong doctor, if qualified, can assist the patient by prescribing herbal tonics. The patient should also be encouraged to eat regular, nutritional meals and to balance work with relaxation time. If the patient's energetic field has been damaged, the Qigong doctor can repair it by using Flat Palm and Circle Spiraling techniques (see Chapter 34).

ENERGETIC LEAKAGE

Energetic leakage is a result of the patients releasing Qi from out of their joints, due to injury or unconscious sabotage. When energy drains from the patients' joints, the specific areas along the channels become depleted, resulting in a Deficient Qi which can cause an atrophy syndrome. Any type of energetic thinness along the Wei Qi field indicates an insufficiency of Qi. A Deficient energetic field can result in an empty, numb, or weak feeling of Qi within the tissue areas.

Treatment Principals for Energetic Leakage

The doctor envelops the patient's depleted tissue area (which is leaking) with Spiral Wrapping techniques to create an energetic cast that allows the patient's energetic field to completely heal (see Chapter 34).

MIGRATING QI DEVIATION

Migrating Qi Deviations can result from either the patient's suppressing and avoiding feelings, or from Wind invasion (either Internal or External). In Traditional Chinese Medicine this condition is known as Wind Bi (Wind Pain) or Wandering Bi (Migrating Pain). Once patients feel cornered by a specific memory, they subconsciously shift the toxic energy to another part of their body; this energetic shifting requires the treatment to be restarted. When treating these patients, it is important to first connect the original trauma to the Heart to prevent these energetic shifts, and to stabilize the condition. Once stabilized, disperse the Wind Invasion and tonify any Deficient organ.

Treatment Principals for Migrating Qi Deviation

The Qigong doctor first envelops the primary area of complaint, isolating the pain, and then begins to internally Spiral Circulate the patient's internal Qi. Once the energy is in motion, the doctor leads the Toxic Qi out of the patient's body.

PSYCHOSEXUAL QI DEVIATION

Psychosexual Qi Deviations can result from an immediate energetic tissue over-stimulation. Patients with this condition experience intense sexual undulations and orgasms when being treated in a safe clinical environment. This reaction is due to their avoidance issues, brought about by an energetic trigger mechanism. This psychosexual reaction usually takes place when touching any part of the torso or extremities. Because the energy seeks balance (much like water), when it flows into the genital area, the sexual neurons become overstimulated, activating and releasing the energetic turbulence already stored therein. This condition is mostly found within patients who have either been raped or molested as children.

Treatment Principles for Psychosexual Qi Deviation

The doctor is encouraged to refer the patient to a psychotherapist who specializes in such trauma. Meanwhile, the doctor continues to treat the patient by rooting the patient's Heart (emotional chamber) to the patient's Lower Dantian (sexual chamber). Once this connection is established, the doctor reconnects the patient's Lower Dantian to the patient's Upper Dantian (spiritual chamber), forming a complete body connection within the patient's Taiji Pole. By reconnecting the sexual chamber with the spiritual chamber, the doctor empowers the patient, allowing the patient to reclaim his or her true sexual / spiritual identity.

PATIENT'S ENERGETIC FACIAL DISTORTION

Sometimes, when treating a patient, the Qigong doctor will notice the patient's facial features energetically distort, changing their facial pattern. These facial distortions can appear and disappear rapidly, or maintain their illusion for several seconds. The cause for the patient's energetic facial distortion can be brought on by several factors: The Patient Reliving A Past Trauma, The Ancestral Patterning of the Patient's Jing, or The Energetic Manifestations of the Patient's Active Spiritual Guardians.

1. The patient reliving a past trauma can release an energetic distortion that expresses itself on the energy field of the patient's face. This projected image reflects the age and time frame wherein the patient received the initial trauma. Often, as the doctor connects with the energetic root of the patient's disease, the patient's face will regress to reveal the delicate features of the patient as a young child. This visual facial pattern of energetic age regression reveals the exact age the patient suffered the initial trauma.
2. The ancestral patterning of the patient's Jing can also be released as energetic distortions that express themselves on the energy field of the patient's face. Occasionally, these projected images can reflect the patient's ancestry, or reveal an ancestral guide. Multiple projections of racially identifying features; images of men, women, and sometimes children will appear to shadow the patient's features, much as a projected image over an original form.
3. The energetic manifestations of the patient's spiritual guardians are believed to project their image onto the face of the patient during treatment. Some cultures accept that these guides manifest in the form of animals. The animal form will appear to flicker across the patient's face and then dissipate. This phenomenon is the most rarely seen of all the energetic facial distortions.

PATIENT RESISTANCE TO ENERGETIC THERAPY

Sometimes, in order to avoid the painful memories, the patient will subconsciously sabotage the treatments. Resistance can manifest in one of four ways.

1. Falling asleep is a form of physical escape.
2. Excessive talking is a mental form of avoidance behavior.
3. Irritability or sexual response is a form of emotional escape.
4. "Shening out," also called "spacing out," is a form of spiritual escape.

When observing these patterns, the doctor encourages the patient back into the body to feel the emotions and memories trapped in the particular areas being treated. It is important for the doctor to let the patient know the cause and effect of the patient's chosen defense mechanism. These personal conversations are necessary for the patient's growth and emotional healing. In cases of sexual response, tact and appropriate professional boundaries are especially important.

When emitting Qi, the doctor can encourage the patient's Shen to stay in its residence by asking the patient to focus his or her breathing on the tissue areas being treated. This focused attention causes the patient's Po to settle and calm themselves within the patient's body, allowing the patient's Hun to communicate the feelings surfacing within the patient's Heart.

UNBLOCKING THE FIVE PASSES

The Wu Guan or Five Passes refer to the five important gates on the Governing Vessel. These gates, or passes, are sometimes called Tie Bi, which translates as the "iron wall," and are the areas of the body where it is most difficult for the energy to pass through when circulating the Microcosmic Orbit. Free passage through the Governing Vessel is important, especially when tonifying the patient's body or regulating the patient's Qi. To circulate the Qi up the Governing and down the Conception Vessels, these five gates must remain open.

Opening these Five Passes are essential for the treatment of diseases in the neck and lumbar-sacral regions. In Medical Qigong therapy, disease is sometimes an indication that both the Governing and Conception Vessels have been, or are, in a state of disconnection. If disorders of Qi, along with Qi obstruction, develop along the Governing and Conception Vessels, the doctor should emit Qi into one of the Five Passes (depending upon the affected area) to facilitate the regulation of Qi (Figure 30.9).

THE LOWER SPINAL PASS

The Lower Spinal Pass or "Coccyx Pass" (Wei Lu Guan) is located on the lowest segment of the spine just posterior to the anus, near the Chang Qiang GV-1 point. If Qi becomes blocked in this area the patient suffers a dull persistent pain by the coccyx, along with feelings of heaviness.

To treat, have the patient direct energy up the Lower Spinal Pass, using mild or gentle intention; have the patient position the tongue against the upper palate and inhale deeply while gently contracting the anus and abdominal muscles. These gentle contractions of the muscles along with the position of the tongue on the upper palate create an energetic sacral pump, which facilities the release of the energetic block.

THE MIDDLE SPINAL PASS

The Middle Spinal Pass consists of the Mingmen (GV-4) area and two points (Jia Ji Guan) located on the lateral sides of the Mingmen. If Qi becomes blocked in these areas the patient suffers a dull persistent pain in the middle of the spine, or on one or both sides of the Mingmen. This obstruction causes a physical sensation that feels like a spinal rupture, break, or herniated disc near the waist area when Qi circulates through these points.

To treat this disorder, the doctor extends Qi into the patient's Middle Spinal Pass, while leading the patient's energy upward along the Governing Vessel.

THE UPPER SPINAL PASS

The Upper Spinal Pass is located between the shoulders on the Shendao point (GV-11) Spirit Gate. If Qi becomes blocked in this area, the patient suffers a dull persistent pain on the midline of the upper back below the fifth thoracic vertebra. This obstruction causes a physical sensation that feels like a spinal rupture, or causes intense Heart pain when Qi circulates through this point.

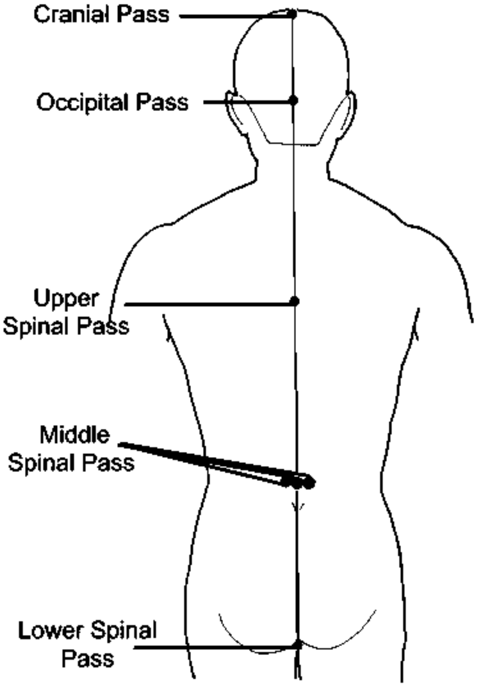

Figure 30.9. The Five Spinal Passes

To treat this condition, the doctor extends Qi into the patient's Upper Spinal Pass, while leading the patient's energy upward along the Governing Vessel.

THE OCCIPITAL PASS

The Occipital Pass (Yu Zhen Guan) is located just inferior to the occipital bone where the brain originates. If the Qi becomes blocked at this area, the patient suffers a dull persistent pain around the occiput, stiff neck, feelings of heaviness, as if there is something were stuck inside the back of the head.

To treat this problem, the doctor should instruct the patient to close the eyes, raise the head slightly and look upward, while gently concentrating on the Baihui GV-20 point. Have the patient imagine the energy block as a cube of ice. The ice melts into water, the water dissolves into vapor, and travels upward along the Governing Vessel into (and through) the Baihui area, then proceeds down the Conception Vessel toward the Lower Dantian.

THE CRANIAL PASS

The Cranial Pass is located at the vertex on the midline of the skull. If the Qi becomes blocked at this area, the patient suffers a dull persistent pain around the eyes and the top of the head, as well as feelings of heaviness, as if there is something were stuck inside the top of the head.

To treat this condition, the doctor instructs the patient to close the eyes, raise the head slightly, while gently concentrating on the Lower Dantian. Have the patient imagine the energy block as a cube of ice. The ice melts into water, the water dissolves into vapor, and travels downward along the Conception Vessel toward the Lower Dantian.

FURTHER COMPLICATIONS

In balancing the patient's energetic body, the doctor first locates the Excess or Deficient condition, determines if the tissues should be tonified, purged, or regulated, then ascertains which energetic complications have initiated the patient's disease.

The doctor must also determine if the patient is sensitive to treatment or resistant to energetic therapy. Below is a summary listing several energetic complications and suggestions on how to rectify them.

1. For Compressed Stagnation, break up the stagnation, purge the pathogens, then tonify and regulate the patient's Qi.
2. For Contracted Stagnation, break up the stagnation, purge the pathogens, then tonify and regulate the patient's Qi.
3. For Bloated Stagnation, break up the stagnation, purge the pathogens, then tonify and regulate the patient's Qi.
4. For Expanded Stagnation, break up the stagnation, purge the pathogens, then tonify and regulate the patient's Qi.
5. For Energetic Armoring, access the patient's Heart (through the Shendao GV-11 area), and induce "soul retrieval" (see Chapter 19).
6. For Energetic Depletion, tonify and regulate the patient's Qi.
7. For Energetic Leakage, envelop the patient's tissues creating an energetic cast to stabilize the energetic fields.
8. For Migrating (Wandering) Qi Deviations, purge the pathogens and then isolate and stabilize before tonifying and regulating the patient's Qi.
9. For Psychosexual Deviations, root the patient's Heart, then regulate the patient's Taiji Pole with divine healing light.

CHAPTER 31
THE MAGIC SQUARE OF THE YELLOW RIVER

The configuration known as the Lo Shu (Lo Writings) or Magic Square of the Yellow River, is shown below. It is said to have been derived from markings on the shell of a "spiritual turtle" that crawled out of the River Lo when the Emperor Yu was draining off the floods. It is considered a mathematical model of the universe. The relationship between the numbers presented in this chart may be used to diagnose the energetic relationship between universal and environmental phenomena, as well as the various pathologies within the occurring internal organ correspondence.

It corresponds to the Later Heaven sequence of the trigrams of the Yi-Jing (I-Ching) and is arranged in accordance with the directions of the compass, with the number "5" at the center of the compass, in an arrangement known as the "Magic Square."

The numerals 1–9 are assigned to each position in accordance with the Lo Scroll markings and represent the Acquired Essences (Postnatal Jing).

Understanding the concept of balance is paramount in Qigong medicine. When Yin and Yang energies in the body are balanced, health is established. The Chinese have given numerical relationships to Yin and Yang properties representing this dynamic connection.

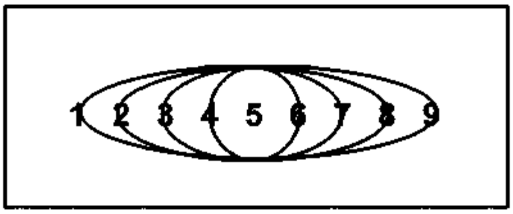

Figure 31.1. Yin and Yang Number Balance

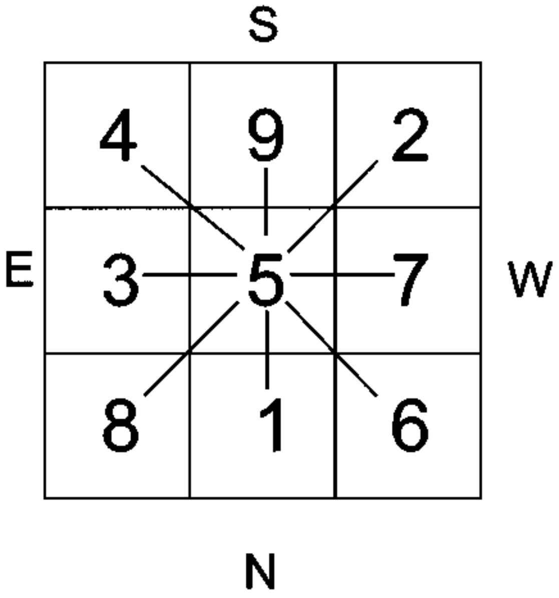

Figure 31.2. The Magic Square

Yang energy is represented by all the odd numbers (1, 3, 5, 7, and 9). Yin energy is represented by all the even numbers (2, 4, 6, and 8). The cycling of Yang (odd numbers) and Yin (even numbers) indicates the law of motion of Yin and Yang in all things. In Chinese medicine the numerical configuration is based on the sum totaling ten (Figure 31.1).

The numbers arranged together create what is known as the Magic Square. Investigation will uncover several numerological correspondences (Figure 31.2). The sum of any two numbers opposite each other in a row, column, or diagonal (with the number "5" between them) equals 10; thus the sum in any of these rows, columns, or diagonals equals 15. Fifteen is the number that represents "Man," who lives between Heaven above (the number being 20, which equals the sum total of the four primary directions) and Earth below (the number also being 20, which equals the sum total

SW =	2 }		SE =	4}		
NE =	8 }		NW =	6}		
	10	+		10	=	20 Yin Earth Qi
S =	9 }		W =	7}		
N =	1 }		E =	3}		
	10	+		10	=	20 Yang Heaven Qi

Figure 31.3. Heaven and Earth Numbers

of the four secondary directions). The numbers in the four primary directions (north, south, east and west) are odd and represent Yang Heaven Energy, while the numbers on the corners are even and represent Yin Earthly Energy (Figure 31.3).

EARTHLY YIN ENERGETIC FLOW

The numbers on the corners of the Magic Square are "terrestrial" in design. All the terrestrial numbers are even numbers. They represent the Earth's energetic field and flow counterclockwise in their circular progression. These terrestrial cycle phases are expressions of Earthly Yin energy and are designed in progressions of "times two." For example, (2) Fire: 2 × 2 = (4) Metal: 4 × 2 = (8) Wood: 8 × 2 = (16) Water (which is reduced to 6, as the ten position is not counted): 16 × 2 = 32 (which is reduced to 2). This completes one cycle in the terrestrial progression, which then returns to the element Fire and begins all over again in the same fashion, continuing in an endless progression of "times two." (Figure 31.4).

Through understanding the Yin energetic progression, the Qigong doctor rotates the patient's energy in a counterclockwise direction to purge Excess conditions and expel pathogenic Qi (called "Yin destroying Evils").

HEAVENLY YANG ENERGETIC FLOW

The numbers on the sides of the Magic Square are "celestial." All the celestial numbers are odd numbers. They represent the Heaven's energetic field and flow in a clockwise circular progression. These celestial cycle phases are expressions of Heavenly Yang energy and are designed in progression of "times three." For example, (1) North: 1 × 3 = (3) East: 3 × 3 = (9) South: 9 × 3 = (27) West (which is reduced to 7, as the ten position is not counted): 7 × 3 = 21 (which is reduced to 1). This completes one cycle in the celestial progression, which then returns to the North and begins all over again in the same fashion, continuing in an endless progression of "times three."

In understanding the Yang energetic progression, the Qigong doctor will rotate the patient's energy in a clockwise direction to tonify. This technique facilitates the continuous circulation of Yang Qi and is used in order to help the tonification of Qi and Blood and replenish Deficiencies (called "Yang promoting growth").

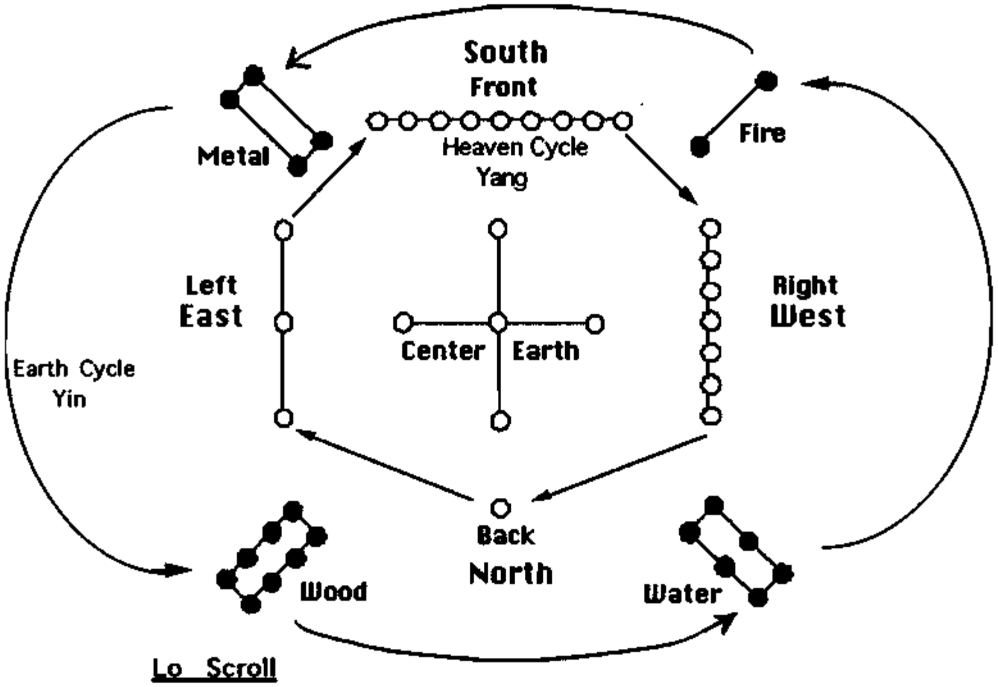
South
Front
Heaven Cycle
Yang
Metal
Fire
Left
East
Right
West
Center
Earth
Earth Cycle
Yin
Back
North
Wood
Water

Lo Scroll

	South East	South	South West	
	☴ (GB-41) 4 Belt Vessel	☲ (Lu-7) Heart 9 Fire Conception	(Kd-6) ☷ 2 Yin Heel	
East	Liver (TB-5) ☳ 3 Wood Yang Linking	5 Spleen Earth	(SI-3) Metal Lungs 7 ☱ Governing	West
	Yin Linking 8 ☶ (Pc-6)	Yang Heel (Bl-62) Kidneys 1 Water ☵	Thrusting 6 (Sp-4) ☰	
	North East	North	North West	

The Magic Square

Figure 31.4. The reason the South is positioned on the top of the page is because, in China, maps were specifically designed for the convenience of the Emperor, whose throne was always positioned facing South, the direction of Yang energy. As the magistrates of his court unrolled the maps before him, the East would be positioned on his left, the West to his right, and the South furthest away at the top of the map (whereas, in the Western culture North is at the top).

THE MAGIC SQUARE AND EIGHT TRIGRAMS

The following Eight Trigram Hand chart represents the terrestrial aspect of the body's acquired structures, evolution, and exchange of Yin/Yang energy.

It should be noted that if the trigrams associated with the numbers shown in the Magic Square are placed in their correct positions, the postnatal or "Later Heaven" trigram arrangement will form. The middle number "5" does not correspond to a trigram because it is the unifying number placed in the center.

In order to use the Eight Trigrams in conjunction with the Magic Square, the patient's palm is divided into nine sections. The Eight Trigrams form the exterior of the palm, while the heart of the palm (known as the Bright Hall or Ming Tang) aligns the Magic Square to the patient's internal organs. The Qigong doctor observes the color, luster, sinking or bulging of skin, as well as any prominent blue veins appearing on the surface of the patient's skin. These observations are compared to the organs and functions associated with each trigram that rules that particular area of the patient's palm (Figure 31.5).

1. The Trigram Kan, located at the base of the palm's heel, is associated with the patient's Kidney and Bladder organs, as well as the function of their urogenital and reproductive systems.
2. The Trigram Kun, located on the palm just below the little finger, is associated with the patient's left Lung organ and the left side of the chest (the left breast in women).
3. The Trigram Zhen, located on the upper part of the patient's thenar eminence, is associated with the function of the patient's Liver and Gall Bladder organs.
4. The Trigram Sun, located on the palm, just under the index finger, is associated with the patient's right Lung, and the right side of the chest (the right breast in women).
5. The Bright Hall (Ming Tang), located on the center of the palm is associated with the patient's Heart, blood vessels, and mental health.

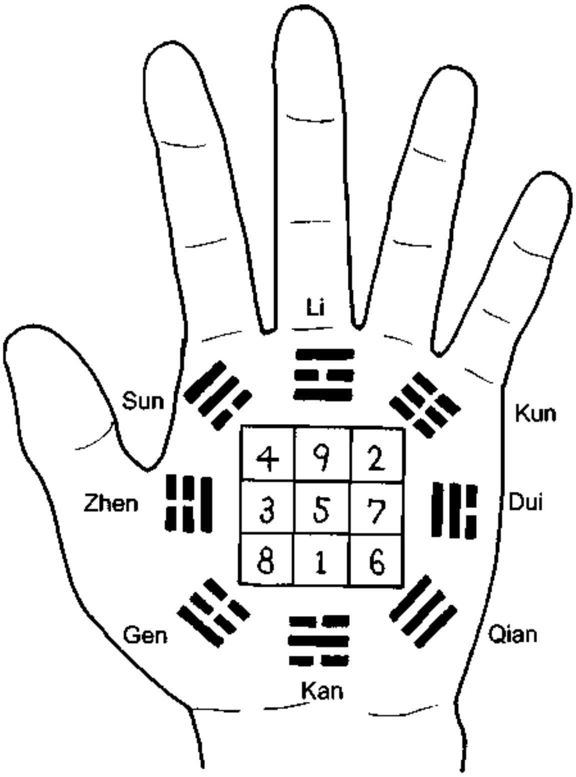

Figure 31.5. The "Magic Square" can be superimposed on the human body for observation, diagnosis, and treatment. For example:

2 and 4 represent the location of the Lungs

3 and 7 represent the location of the Liver and Spleen

8 and 6 represent the location of the Kidneys

9 represents the location of the Upper Burner

5 represents the location of the Middle Burner

1 represents the location of the Lower Burner

6. The Trigram Qian, located on the lateral edge of the palm's heel, just above the wrist crease, is associated with the patient's descending and sigmoid colon.
7. The Trigram Dui, located on the lateral edge of the palm's heel, midway between the little finger and the wrist crease, is associated with the patient's Spleen, pancreas, and Stomach organs.
8. The Trigram Gen, is located on the lower half of the thenar eminence, is associated with

Trigram Correspondences

Trigram	Name	Meaning	Channel	Shu Point	Direction	Number	Season
☲	Li	Fire (Clinging)	Conception	Lu-7	South	9	Mid Summer
☷	Kun	Earth (Receptive)	Yin Heel	Kd-6	South West	2	Start of Fall
☱	Dui	Lake (Joyful)	Governing	SI-3	West	7	Dividing of Fall
☰	Qian	Heaven (Creative)	Thrusting	Sp-4	North West	6	Beginning of Winter
☵	Kan	Water (Abysmal)	Yang Heel	Bl-62	North	1	Absolute Winter
☶	Gen	Mountain (Keeping Still)	Yin Linking	Pc-6	North East	8	Start Of Spring
☳	Zhen	Thunder (Arousing)	Yang Linking	TB-5	East	3	Dividing of Spring
☴	Sun	Wind (Gentle)	Belt	GB-41	South East	4	Start of Summer

Figure 31.6. The Eight Trigram Correspondence

patient's appendix and ascending colon.

9. The Trigram Li, located on the palm between the middle and ring finger, is associated with the patient's blood circulation, Heart and eyesight.

THE MAGIC SQUARE AND THE EIGHT EXTRAORDINARY VESSELS

In Chinese medicine, the Lo Writings and their correspondences form the model which is the basis for the Ling Gui Pa Fa (The Eight Techniques of the Mysterious Turtle) system of acupuncture. In this system there are Eight Confluential Points that are associated with the Eight Trigrams (Figure 31.6). These points are where the Eight Extraordinary Vessels and Twelve Primary Channels intersect each other.

This method describes the theory that all the points on the body are dominated by the Shu Points, which are in turn dominated by the Eight Confluential Points. The confluential points and their trigram correspondences are Kd-6 (Kun Gua), Lu-7 (Li Gua), GB-41 (Sun Gua), TB-5 (Zhen Gua), Pc-6 (Gan Gua), Sp-4 (Qian Gua), SI-3 (Dui Gua), and Bl-62 (Kan Gua).

UNDERSTANDING THE METAPHYSICAL ASPECT OF NUMBERS

The Chinese medical text of the Nei Jing explains numerology and its mystical significance, as well as the practical utilization of the correspondences between numbers and natural phenomena. Man's physical, mental, emotional, energetic, and spiritual form manifest as part of the microcosm of these nine correspondences, and are explained as follows.

1. Number One corresponds to Heaven, the Center Taiji Pole, and represents the skin.
2. Number Two corresponds to Earth and represents the muscles. The number two also cor-

responds to the two primary forms of Yin and Yang.

3. Number Three corresponds to Man and represents the blood vessels. The number three also corresponds to the Triple Burners, Three Powers, Three Treasures, Three Dantians, and Three External Fields of Wei Qi.
4. Number Four corresponds to the four seasons and represents the muscles corresponding to the body's four extremities, the Four Earthly Quadrants and the Four Divisions of Heaven (N.S.E.W.), and the Four Prenatal Trigram Essences (Great Yang, Small Yang, Great Yin, and Small Yin).
5. Number Five corresponds to the Five Notes of the Chinese pentatonic scale and represents the voice. The number five also corresponds to the Five Elements, Five Colors, Five Tastes, Five Planets, and the Five Elemental Constitutions.
6. Number Six corresponds to the Six Divisions (Tai Yang, Shao Yang, Yang Ming, Tai Yin, Shao Yin, Jue Yin) and represents the union of Yin and Yang. The number six also corresponds to the Six Yao lines and their correspondence to the body, and the Six Accomplishments (intelligence, humanity, holiness, sincerity, moderation, and benevolence).
7. Number Seven corresponds to the Seven Luminaries (sun, moon, and the five planets), which represent the Seven Openings of Man (two eyes, two ears, two nostrils, and one mouth). The number seven also corresponds to the Seven Emotions, and the Seven Stars of the Big Dipper.
8. Number Eight corresponds to the Eight Extraordinary Channels from which the body's prenatal energy flows and represents the breath. The number eight also relates to the Eight Trigrams and their various correspondences.
9. Number Nine corresponds to the formless Void or Wuji, as well as the Nine Heavens (one is centered, and the others are in the eight separate divisions of the Lo Pan compass) and the Nine Lo River Chart (Magic Square). The number nine also represents the nine openings of man (two eyes, two nostrils, two ears, one mouth, one anus, and one urethra).

THE FIVE ELEMENT ENERGETIC FLOW OF THE MAGIC SQUARE

The number formations on the Magic Square are not only considered the energetic ebb and flow, evolution and exchange of Yin and Yang Qi, but also correspond to the energetic harmony of the Five Elements. The energetic harmony of the Five Elements are arranged as follows.

- The Kidneys are positioned in the North; their element is Water and number is 1.
- The Liver is positioned in the East; its element is Wood and number is 3.
- The Heart is positioned in the South; its element is Fire and number is 9.
- The Lungs are positioned in the West; their element is Metal and number is 7.
- The Spleen is positioned in the Center; its element is Earth and number is 5.

These five numbers are arranged in accordance to their energetic progression and are used as a templet for observation and diagnosis of the patient's body, according to their Five-Elemental postnatal structure.

CROSSOVER ENERGY PATTERNS

According to the theory of energetic harmony, when one area of the body lacks Qi, there is generally an Excess energy accumulation in another area. Because the body is symmetrical, one can naturally look on the opposite side for the misplaced Qi. The Qigong doctor may dredge the full side in order to drain the Excess and support the Deficient area.

If the imbalance is not on the opposite side, the doctor can scan for Excess Qi using a figure eight pattern (Figure 31.7). Classically, the Crossover Energy Pattern is one of the most powerful pathways by which Qi flow in the body can be stimulated, and is excellent for balancing internal body energy. The figure "8" pattern can also be used in order to balance all sides of the body (top and bottom, as well as left and right).

When the Qigong doctor must balance the energy in the left and right sides of the brain, the patient may be assigned homework that requires them moving Qi in a figure "8" pattern sideways (i.e., drawing the infinity symbol). This exercise prescription stimulates, strengthens, and balances the right and left hemispheres of the patient's brain. It is used successfully in the treatment of patients with dyslexia.

When patients have internal problems that affect the function of the whole body, it may be necessary to treat two areas simultaneously. The Spleen and Stomach areas, for example, may be tonified along with the Lung areas to strengthen and increase the patient's resistance to External pathogenic factors.

Because organ imbalances are deeper, and tend to be the result of years of energetic deviations, they may take longer to reverse.

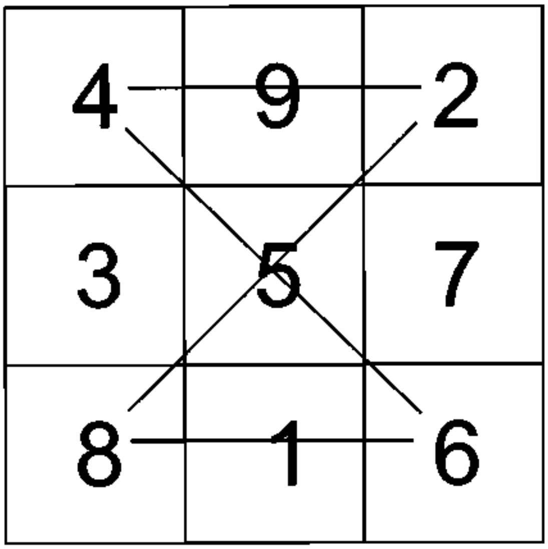

Figure 31.7. In this Crossover Energy Pattern, area 4 is treated using area 6 and so on.

CHAPTER 32
HEALING THE PATIENT'S EMOTIONAL TRAUMAS

INTRODUCTION

In my personal practice I feel the main problem for patients is in dealing with their own "hurt," which may have been inflicted upon them several years prior to the manifestation of their disease. It is a recognized truth that when people isolate themselves from each other, their environment, and their spiritual life, it is because they feel the need to guard themselves.

They guard and protect themselves and their belief structures from invasion, criticism, exploitation, or attacks, whether real or imagined, allowing interchange only if it supports their beliefs.

People build energetic walls, and when the walls are penetrated they often experience anger, self-hatred, pain, fear, and shame. Consequently, masks and defence mechanisms are created to hide and support these hidden belief structures.

How the patients develop and maintain their energetic patterning will determine which clinical modality the Qigong doctor will utilize to initiate healing. To free themselves from personal hurt the patients must reconnect with their anger, pain, fear, or worry. Eventually, with the unconditional loving support of the doctor, patients are able to reintegrate the hurt and anger, and accept parts of themselves they had rejected for so long. This acceptance of painful feelings and memories affects the Heart and frees the patients from the emotional traumas of the past.

To heal and release these pathogenic emotions, the patients go through a five-stage detoxifying process during which they feel, experience, and release the emotional trauma that initiated their disease. This is a natural progression in the process of Medical Qigong therapy and healing.

THE PHYSIOLOGICAL EFFECTS OF BENIGN AND CHRONIC EMOTIONS

Within the clinic it is observed that benign emotions (love, joy, peace, etc.) bring about a calm and restful condition within the patient's physiology. Chronic emotions however, create a detrimental stressful physiology, adversely affecting the quality of the patient's life-force energy. As discussed previously, emotional internal and external environmental factors can and do affect the quality of the body's Qi.

1. Excessive joy from the Heart slows down the Qi.
2. Excessive worry and thinking from the Spleen stagnates the Qi.
3. Excessive sadness from the Lungs obstructs the Qi.
4. Fear from the Kidneys descends the Qi.
5. Excessive anger from the Liver ascends the Qi.

FIVE LEVELS OF FEELING AND THEIR GOVERNING ORGANS

To complete their initial healing, the patients go through five stages of emotional resolution. These stages act as a template for the patients' emotional healing, allowing them to progress through, and detoxify each of their Five Yin Organs. The Five Yin Organs store the patients' emotional history in their tissues. Each of the Five Yin Organs also affects different functions and parts of the body (Figure 32.1).

1. The Liver governs the emotions dealing with the past. The Liver stores the negative emotions of irritability, jealousy, anger, rage, blame, and resentment. These emotions and mental states affect the flow of Qi in the channels, influencing the patient's tendons, ligaments, nails, small muscles, peripheral nerves,

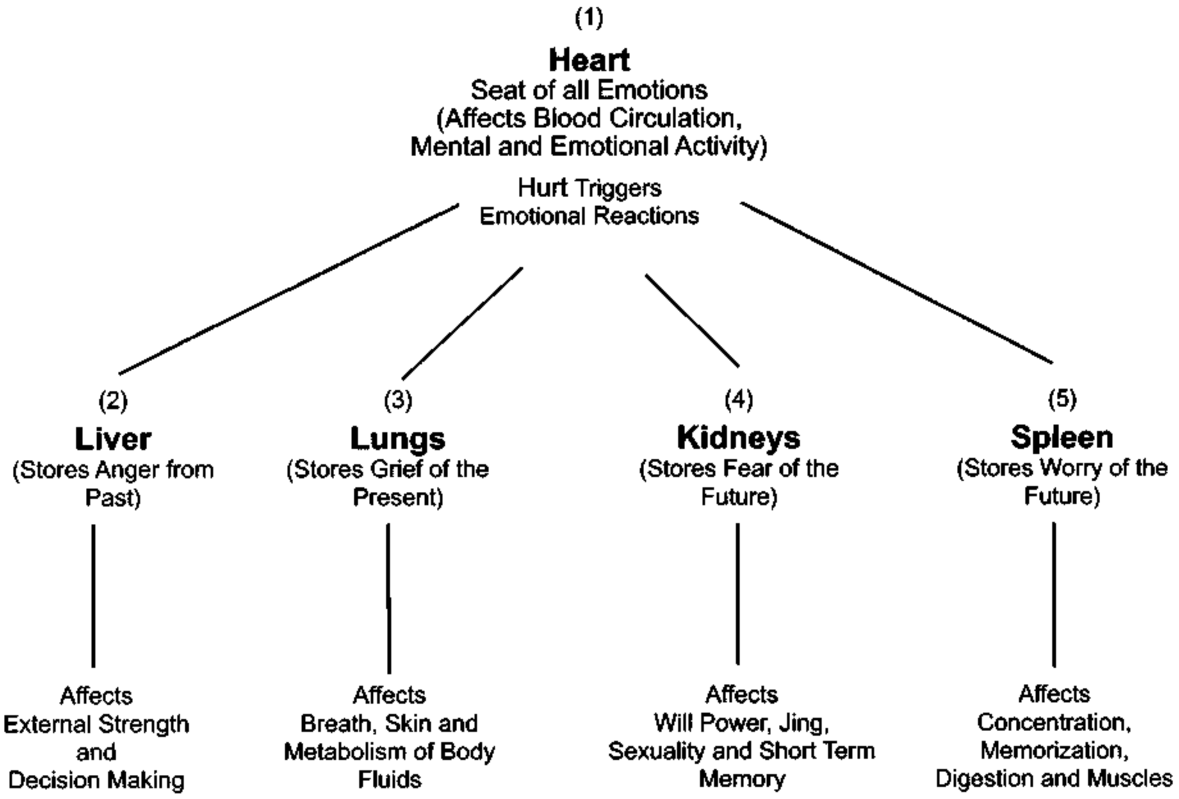
(1)
Heart
Seat of all Emotions
(Affects Blood Circulation, Mental and Emotional Activity)
Hurt Triggers Emotional Reactions
(2)
Liver
(Stores Anger from Past)
Affects External Strength and Decision Making
(3)
Lungs
(Stores Grief of the Present)
Affects Breath, Skin and Metabolism of Body Fluids
(4)
Kidneys
(Stores Fear of the Future)
Affects Will Power, Jing, Sexuality and Short Term Memory
(5)
Spleen
(Stores Worry of the Future)
Affects Concentration, Memorization, Digestion and Muscles

Figure 32.1. Diagram of Five Elemental Emotional Transitions

external strength, the iris of the eyes, vision, tears, the Gall Bladder, bile, decision making, the external genitalia, and the Hun (Three Ethereal Souls).

2. The Lungs govern the emotions dealing with the present. The Lungs store the negative emotions of grief, sorrow, sadness, anxiety, shame, guilt and disappointment. These emotions and mental states affect the flow of Qi in the channels, as well as the patient's breath, skin, Large Intestine, mucous membranes, the metabolism of the Body Fluids, body hair, smell, and the Po (Seven Corporeal Souls).
3. The Kidneys govern the emotions dealing with the future. The Kidneys store the negative emotions of fear, loneliness, and insecurity. These emotions and mental states affect the patient's energetic channel flow, as well as the Jing, bones, brain, bone marrow, inner ear, hearing, head hair, pubic hair, spinal cord, cerebrospinal fluid, ovaries, testes, anus and urethra, sexual fluid, reproduction, short term memory, and the Zhi (Will Power).
4. The Spleen governs the emotions dealing with resolution and responsibility. The Spleen also stores the negative emotions of worry, remorse and regret, which can create obsessions. These emotions and mental states affect the flow of Qi within the channels, as well as the patient's large muscles, lymph and saliva secretions, the mouth, lips, taste, the Stomach, digestion, concentration, the ability to memorize, and the Yi (Intention).
5. The Heart governs the emotions dealing with

shock, nervousness and excitement, as well as the final stages of healing (i.e., peace, love and forgiveness). These emotions and mental states affect the patient's energetic channel flow, as well as the patient's blood vessels and circulation, complexion, perspiration, mental and emotional activity, long term memory, Small Intestine, and the Shen (Spirit).

In the process of forgiveness, the anger and resentment are traversed to reach through to the levels of original hurt, sadness and grief. Only after reaching this level, and working through the pain, can the patient begin the process of letting go of the anger, blame, and resentments that protect them from feeling the full impact of the original trauma. Forgiveness begins only after the anger and crying ends. There are three stages of forgiveness: forgiving oneself, accepting the situation, and forgiving the abuser.

1. The forgiving oneself stage deals with the energy involved in healing oneself, and is divided into two specific stages:
   - Forgiving oneself for allowing the hurt to occur in the first place (even if the patient had no choice in being victimized), or for not fighting back strongly enough, if at all, and
   - Forgiving oneself for hanging on to all that anger and pain for so many years.
2. The accepting the situation stage deals with releasing the energetic emotional charge within the environment, or the unwanted consequences connected to the original trauma (e.g., forgiving the accident, injury, handicap, illness, trauma) within the patient.
3. The forgiving the abuser stage is the most difficult final step. Many factors can interfere with this process, such as the patient's fear of re-abuse or continued abuse if forgiveness is genuinely given. Another major factor is the negative, toxic, external energy projected by the abuser upon the patient, sometimes over many years. This energy is absorbed by the patient over time and often leads to the patient identifying with the abuser. The fear of being, or becoming, just like the abuser can be overwhelming. Anger at the abuser may seem to be the only thing standing between them; and allowing any feelings of love, compassion, or understanding may appear to be an acceptance of "evil," not only within the abuser, but within the patient's own self. An acceptance of a dark side within each of us is necessary to forgive the abuser. Often victims of abuse cannot tolerate the idea that they are anything like their abuser. Only the conscious mind, working in concert with the patient's spirit, can reintegrate this essential part of the self that has been utterly rejected and denied.

Children who have experienced abuse may be especially hard pressed to accept anger as a natural, integral part of themselves without which they cannot feel whole. They frequently view all and any anger as not only dangerous but an "evil" force. Thus, forgiveness of the abuser requires restructuring of belief systems, and the acceptance of one's own dark side, as well as the abuser's.

Through conscious intent and creative subconscious intuition, it is possible to come to an understanding that the "evil" experienced stems not from the devil, but from ignorance and pain. In most cases the abuser's intent was not to harm and destroy, but ironically, to teach and discipline for the good of the patient. The abuser did the best that he or she could at the time, given all the unresolved pain and anger within the self.

When patients experience extreme abuse, or prolonged abuse, they may lose the ability to distinguish the difference between wrong actions and the abuser. The abuser is not evil incarnate though evil deeds were done. This often leads victims to see the world only in terms of black and white with no shades or colors. The true healing requires acquiring the ability to see anew (as a little child), to reawaken to the amazing diversity of life.

These skills sometimes require teaching the patient and may necessitate referring the person out to a professional, especially if the patient suffers from a character disorder (e.g., borderline personality disorder, schizo-affective disorder, and so on).

Generally, after performing specific Qigong exercises and meditations which release the patient's pent-up, toxic emotions, the patient begins to come out of denial and awaken to genuine spontaneous emotions and feelings. This transition can be very painful at first. I usually tell the patients that this transition is similar to having their arm frozen, and then thawed out.

Imagine if the arm is placed into ice cold water, and left there. The first thing the body would feel is shock, followed by pain. As the feelings slowly begin to leave, the pain begins to diminish, and the arm becomes numb. As the mind shifts its attention away from the feelings in the arm, it begins to establish a denial system, ignoring, minimizing, or suppressing the condition in which the arm has been placed. When first removing the arm from the ice water, the tissues are still numb. After awhile, however, as the numbness begins to wear off, feelings begin to return to that particular area of the body, and pain is felt. The body is no longer in shock and is now feeling the pain and the memory of the past experience.

Our lives are the same way, in regard to physical, mental, emotional, and spiritual pain. While growing up, we are subject to constant trauma on all four levels of existence. Through an instinctive ability, we adapt and learn to become numb to our true feelings in order to survive. When we begin to awaken, we are suddenly thrust into shock, then pain, as we begin to face the source and cause of our trauma.

From a Traditional Chinese Medical perspective, pain is a means by which the body communicates its needs to the mind. In cases of chronic pain, the patient generally learns to divert awareness away from the area of pain through distracting thoughts and emotions. This helps the patient to cope with the pain, but also interferes with the healing process. If the patient reduces conscious awareness of the painful area(s) in the body, the mind/body connection becomes blocked and slows or prevents healing. The initial clarification of the "Message" regarding the disturbance can cause both a sensation of increased pain and an activation of the healing process.

By comprehending the five levels of feeling and their progressive transitions, the patient gains understanding and healing from these past traumas.

STAGES OF EMOTIONAL HISTORY AND RESOLUTION

After the patient understands the five levels of feeling and how they affect their respective organs, the Qigong doctor can assist patients in the cognitive restructuring of their disease-causing belief structures. This emotional re-patterning is initiated through the unveiling of several stages of painful emotional history, which affect the energetic healing abilities within the patient's body.

It is important for the Qigong doctor to remember that all physical matter is conscious energy activated and focused by consciousness. The patient's perceptions create the physical presence of the disease. Since all matter and energy is unified, all matter down to the smallest subatomic particle has awareness, makes choices, and communicates simultaneously with all other matter in the universe. Choices are initiated in the patient's body through the active fields of environmental and universal energy.

Understanding these subatomic communications allows the Qigong doctor to assist the patient in discharging toxic energy trapped and stored within the tissues. The experience of the emotional trauma, and its destructive emotional energetic patterning, is clinically explained and described in the following stages.

1. The patient is hurt (physically, mentally, emotionally, and/or spiritually). The patient receives the original trauma through either an aggressive, direct attack (i.e., being struck, screamed at, belittled, or shamed, etc.), or by a passive, indirect attack (i.e., being abandoned, rejected, ignored, betrayed, etc.). In this stage of development, the patient's trusting connection to life is violated.
2. The patient registers the hurt, feels the pain, and naturally expresses this trauma in an outburst of emotion (i.e., screaming, crying, laughing, groaning, etc.). In this stage the patient recognizes the wounding.
3. The patient seeks attention and looks for a

human emotional response, seeking validation and self-affirmation. In this stage, if the patient does not seek and receive attention for healing the trauma or, if the individual is not free to express his or her real feelings, healing will not take place.

4. The patient's hurt remains unresolved and becomes history. At this point in the transition it is important that the patient feel the hurt, understand the circumstances surrounding the hurt, then discharge the pain. If the patient's hurt does not become completely discharged from the body, it becomes stored within the patient's tissues. Thus, the patient's hurt becomes internalized; the emotions of unexpressed anger, fear, pain, and shame begin to accumulate in the body. The trauma remains as an active charge of unexpressed emotional energy.
5. The patient's whole body stores the distressful emotional memories. This distress gets restimulated in the present (through various trigger mechanisms), and alters or blocks the patient's flow of energy and thinking.
6. The patient learns defense mechanisms to survive. The patient develops strong defense mechanisms instead of healing from the emotional wounds (known as a pathological recovery). At this point the patient's denial system is reinforced, and kept active to protect the patient from any further hurt, in order to ensure survival and maintain at least the semblance of a normal, functional life.
7. A dynamic polarity arises that seems to take on a life of its own. The patient swings back and forth between overcontrolling emotions and loosing control over them completely. The patient becomes obsessed with the painful feelings, either hiding from them or becoming overwhelmed by them. The patient then chooses either one of two options:
   - To become isolated, emotionally removed and distant, or
   - To become needy, co-dependent and attention hungry.
8. The patient develops addictions to continually hide from the pain. The patient's actions become based on the unhealed feelings resulting in four overlapping ways to hide:
   - Physically through sports, excessive exercise, food, sleeping, etc.,
   - Mentally through compulsive behavior (e.g., studying, being a work-a-holic, compulsive book reading, etc.).
   - Emotionally through either chronic rage, grief, obsessive worrying, and chronic fear, or through a compulsive search for love, joy, and excitement (sex/love addiction), etc., and
   - Spiritually through fanaticism of various types, or extreme spirituality.
9. The patient begins to recover from the trauma. To heal and recover, the patient must cease concentrating on the pain and redirect his or her full attention towards setting healthy boundaries and positive affirmations. Some patients' boundaries are too solid and immovable (i.e., they won't allow for any meaningful interaction with others). These patients do not really need to set boundaries, they need to allow their boundaries to become more permeable. Patients' boundaries need to be flexible in order for them to be healthy. Patients must learn when to open and when to close their boundary system, and not just constantly shut everything out (people, relationships, responsibilities, etc.) or let everyone and everything in.

It is important for patients to understand that perpetrators of abuse are in a state of denial of their own shame and imperfection; thus they inflect pain and suffering onto anyone present at the time. Some perpetrators single out one family member as a recipient for their pain. Sometimes the chosen victim reminds the perpetrator of his or her younger innocent self; thus all the accumulated self-hate is projected onto this one child through no fault of the victim. The resemblance may be real (slight physical similarities for instance) or imaginary (the child's vulnerability, for example, or a singled out personality trait that triggers the abuser's anger). The one common factor among abusers is their inability to see their victims for who they really are. Abusers make no effort to know the victim, to really see the whole person; instead they see a distorting mirror which reflects their own image back to them, magnifying their own flaws and fears. Through setting boundaries and affirming their own self-worth, the patients will be able to release years of toxic emotions stored within their tissues. This can initiate health and healing on every level of their being (physical, mental, emotional, and spiritual).

ASSISTING THE PATIENT IN ESTABLISHING HEALTHY BOUNDARIES

The patient is generally informed that in order to make it through the spiritual, emotional, and mental transitions brought about by Medical Qigong therapy, it is important to establish an active but flexible boundary system. The patients are then taught to divide their relationships into three categories: True Friends, Acquaintances, and Toxic Relationships (Figure 32.2).

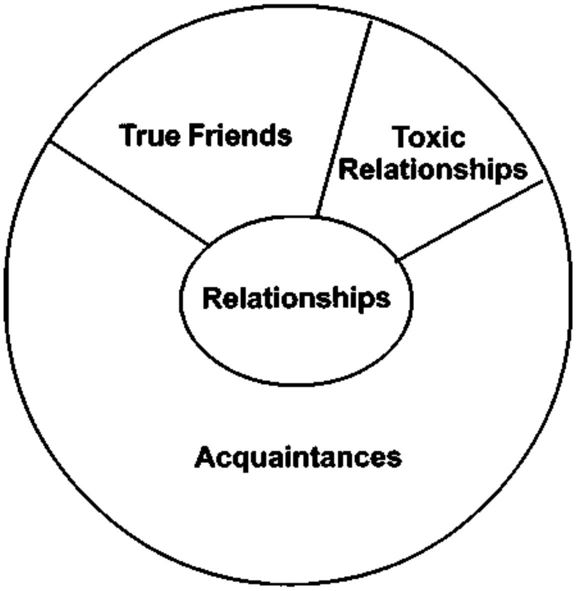

Figure 32.2. Types of Relationships for Setting Flexible Boundaries

1. True Friends consist of people with whom the patient can share and expose his or her most intimate hidden secrets and emotional feelings, without worrying about being judged. In the presence of true friends, the patient feels emotionally accepted and loved unconditionally. The boundary system should be down, and the patient should feel completely open. It is safe to be completely vulnerable.
2. Acquaintances consist of people with whom the patients should share only limited amounts of information regarding their personal/emotional life experiences. The most hidden secrets could be judged, misconstrued, and gossiped about. All conversation with these individuals should, and must be, kept on a fairly superficial level. The patient's boundary system should be up, and the patient should monitor the conversation, as it is not safe to be totally vulnerable.
3. Toxic Relationships consist of people with whom the patient cannot and should never share any knowledge of their personal/emotional life. These individuals chronically abuse, mistreat, belittle, and attack others (whether they are members of the patient's own family, work, or neighborhood). All conversation with these individuals should be kept to a minimum. The patient's boundary system should be on full armor, and whenever possible, these individuals should be avoided.

REJECTION, BETRAYAL, AND ABANDONMENT

From a young age, people are introduced to the multidimensional experience of being hurt (physical, mental, emotional, and spiritual). This affects the individual's center core and initiates a coping response due to feelings of rejection, betrayal, and abandonment. The normal reaction to such emotions results in experiencing emotions ranging from fear to frustration. These individuals begin to develop masks, that warp and scar their developing personality. As these individuals continue to grow, the illusions of these masks demand more Qi to sustain their form, which in turn drains vital energy from the growing body, and weakens the immune system. This denial system, fortified by the individual's emotional patterning,

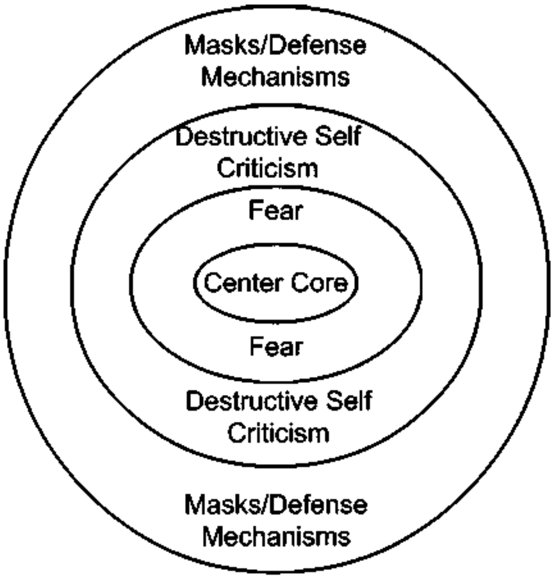

Figure 32.3. The True Self is contained within the Core Self.

sustains him or her until they can heal and re-pattern the no longer beneficial belief system.

To re-pattern their belief systems, patients and Qigong practitioners must first be made aware of their initial programming patterns and the reasons for which these were created in the first place. The following "spiritual template" is applied to all patients, regardless of culture, race, or religion.

To access the patient's center core, the Qigong doctor must help the patient identify the masks which have taken on a life of their own and created the false "ego self." Next, the patient must address personal self-hate issues, which have driven the patient to develop the masks and defense mechanisms. Finally, the patient must face the foundational fear that has severed his or her connection with their "true self" contained within the center core (Figure 32.3).

In the Medical Qigong clinics in China there is a saying, "Patients must first stop running from their fears, and face the "Paper Tiger" (i.e., the illusions and fantasies that have been chasing them); in the West, it is considered "stopping and facing your internal demons."

MASKS AND DEFENSE MECHANISMS

To armor the ego, individuals resort to what Sigmund Freud labeled as "ego defense mechanisms." There are nine strategies; and with the exception of the last one listed, i.e., "sublimation," they serve to mask the core self. The nine defense mechanisms and their strategies are: Repression, Displacement, Projection, Intellectualization, Regression, Fixation, Denial, Reaction-Formation, and Sublimation.

1. Repression is the pushing down of unwanted ideas and emotions into the unconscious;
2. Displacement is the shifting of impulses aroused by one person, or situation on to a safe target;
3. Projection is the attribution of unacceptable impulses within oneself to other people;
4. Intellectualization is an elaborate rationalization of a naked impulse, to justify it;
5. Regression is the return to an earlier childhood stage of behavior to reduce the demands on the ego;
6. Fixation has the same result as Regression, but the person becomes fixated at a particular stage of mental and emotional development;
7. Denial is a conscious refusal of an impulse-evoking fact, feeling or memory;
8. Reaction-Formation is the conversion of one feeling into its opposite, typically seen in love turning into hate;
9. Sublimation is the channeling of unacceptable impulses into acceptable, refined social forms and is the only defense mechanism considered to be a healthy reaction.

The ego mediates between the "id" (i.e., the primordial survival drives and impulses) and the "super ego." The super ego consists of the ideals and conscience of the individual.

EMOTIONAL SECURITY

Each individual has basic emotional needs that must be met to feel secure. Without the comfort of these needs being met an individual will sometimes lash out and react in an antisocial behavior. This reaction is generally due to the individual fostering feelings of rejection, betrayal, and abandonment. The three personal needs required

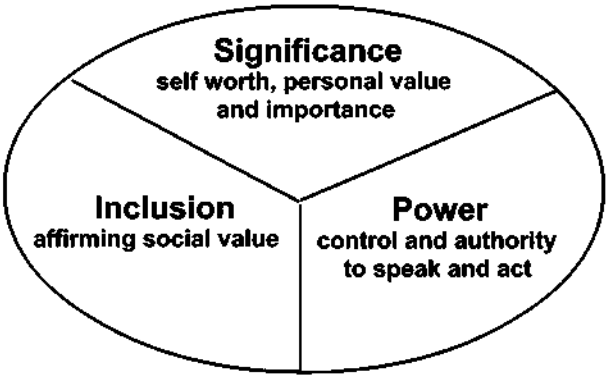
Significance
self worth, personal value and importance
Inclusion
affirming social value
Power
control and authority to speak and act

Figure 32.4. The Three Personal Needs That Foster Emotional Security

to foster emotional security are as follows: Significance, Inclusion and Power (Figure 32.4).

1. For an individual to feel significant, he or she must be given self worth, personal value and importance. Without personal value the individual will feel insignificant, unimportant, and worthless.
2. For an individual to feel included, he or she must be given affirming social value (knowing that they play an important role in the family, relationship, or society as a whole). Without personal acknowledgment the individual feels isolated, segregated, and separated.
3. For an individual to feel powerful, he or she must be given the ability to control and the authority to speak and act. Without personal power the individual feels powerless, weak, and incapable.

In addressing the patient's disease, it is important to sometimes address his or her living and work environment, as both can contribute and influence the patient's healing potential.

HEALING EMOTIONAL TRAUMAS

There are several techniques and methods used to assist patients in healing through emotional traumas (Figure 32.5). Such meditations as Pulling out the Pain are often prescribed as homework along with other treatment methods of Medical Qigong therapy. Some of these methods used to heal emotional trauma are described as follows.

1. Medical Qigong Distance Therapy affects the body, mind and emotions. The Qigong doctor alters the patient's energetic tissues and fields to re-pattern the patient's emotional responses.
2. Medical Qigong Massage Therapy and Bodywork involves lightly touching and stimulating the patient's tissues, to release the trapped emotions so that they may be appropriately reprocessed.
3. Medical Qigong Respiration Therapy unites the body, mind, and emotions by moving the energy through specific breathing techniques. By changing the breathing pattern, the Qigong doctor changes the patient's energetic emotional patterns, releasing trapped memories deep within the patient's tissues.
4. Medical Qigong Dynamic Postural Therapy expands and stretches the patient's tissues. This active adjustment of the patient's body alignment increases the flow of Qi and Blood through his or her internal organs, facilitating an emotional release (e.g., the Dry Crying exercise).
5. Medical Qigong Imagery and Meditation Therapy uses awareness meditations and images to sedate the patient's reactive mind in order to connect with the higher self-identity. The doctor interrupts the patient's emotional

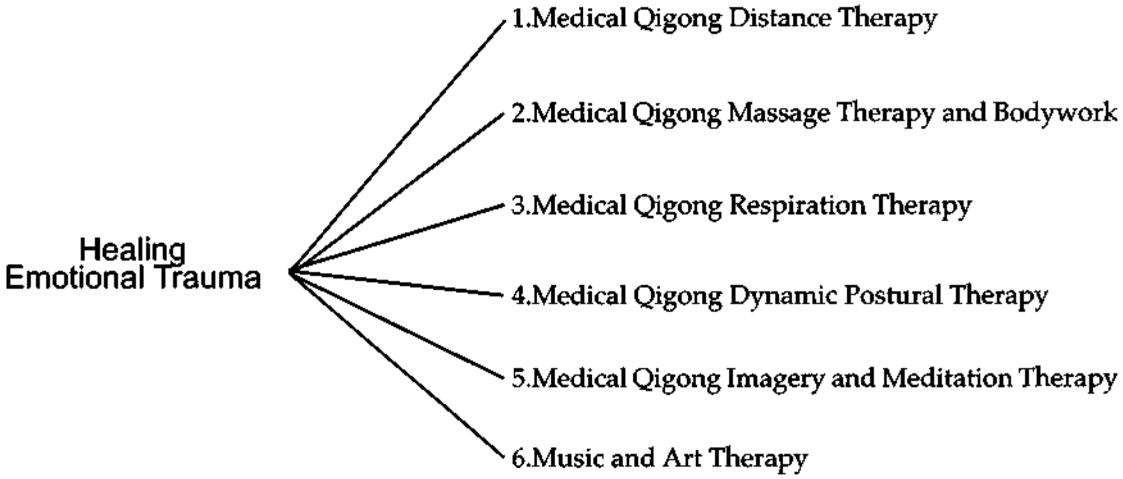
Healing Emotional Trauma
1.Medical Qigong Distance Therapy
2.Medical Qigong Massage Therapy and Bodywork
3.Medical Qigong Respiration Therapy
4.Medical Qigong Dynamic Postural Therapy
5.Medical Qigong Imagery and Meditation Therapy
6.Music and Art Therapy

Figure 32.5. Six Methods Used For Healing Emotional Trauma

thought patterns by changing the patient's sequence of sensations, word pictures and feelings. This emotional reforming changes the meaning of the patient's interpretation and challenges his or her self-limiting beliefs.

6. Music and Art Therapy are sometimes prescribed to change the patient's mood and awareness. Through playing a musical instrument, dancing, singing, or creative free-association drawings, sculptures, pottery, and so on, patients can access their Creative Subconscious Mind to repattern old belief structures and enhance the body's energy. Some patients will resist this form of therapy as they do not consider themselves to be good artists. It is vital to explain that these creative endeavors are not for the purpose of creating art, but for expressing feelings and emotions. These art forms need not be pretty, perfect, or shared with anyone if the patient does not desire to do so. They need not even be kept so long as the patient has released some feelings and emotions. The process itself is the healing. This type of therapy is used to help the patient recreate, identify, and release trapped emotions deep within the subconscious mind.

PULLING OUT THE PAIN MEDITATION

This meditation uses part of the guided imagery used in the Soul Retrieval Meditation (found in Chapter 19). It releases and transforms the patient's conscious feelings and helps to change or remove the Personal Subconscious toxic holding patterns.

After the meditation, patients should feel as if they have been cleansed and purified of toxic emotional debris. This is, however, only the first level. Patients can only release and transform the surfacing emotions one level at a time. Like peeling an onion, the patients must continue to work through each level until reaching their core issues. These core issues surround the suppressed feelings of hurt and pain.

The doctor instructs and guides patients through this meditation matching the rhythm of his or her words to the patient's exhalations. Patients draw their painful memories and traumas out of their body in the form of dark smoke; they then separate the distilled wisdom and knowledge from this dark cloud, reabsorbing the healing light into every pore and cell of their body. The dark cloud must then be released with the help of an angelic being who severs the energetic cord(s) still attaching the patients to their pain and suffering. When the patients reach this sixth step, the doctor assists them by clapping his or her hands as the patients visualize the "sword of truth", held by this holy being, severing the dark energetic cord(s). The sword of truth represents the patients' decision to regain their power by letting go of their attachment to their illness. This attachment creates very real energetic cords that bind the patients to their painful memories and toxic emotions. Severing these cords, or cord, symbolizes the end of the patients' investment in their disease, and the beginning of true healing. This releases all the pain and despair, i.e., the black cloud ascending into the Heaven into the hands of God or a "higher power," in accordance with the belief structure of the patients. As the cloud reaches the first wave of love and compassion emanating from God or a "higher power," a mighty explosion transmutes the black despair and pain into divine healing light. The doctor then assists the patients in accepting this healing, loving light back into their body to experience, perhaps for the first time in their lives, true forgiveness and a state of grace.

1. Begin by sitting comfortably in a chair, with both feet on the floor, hands resting on your thighs, eyes closed, tongue on the upper palette. Breathe naturally through the Lower Dantian.
2. Focus your attention on the Yellow Court area below the diaphragm where you store painful memories and trauma. Imagine opening up this area to release the toxic energy out of your body as a stream of dark steam releasing from a pot.
3. As the dark steam flows out of your body you are releasing feelings and memories of guilt, anger, rage, humiliation, abandonment, degradation, rejection, insecurity, and sorrow. These toxic emotions have, until now, prevented your healing through creating a deep distrust of self and others.
4. Focus your attention on this energetic cloud, and begin separating the pain and hurtful memories from the knowledge and wisdom gathered from these experiences. Visualize this knowledge and wisdom in the form of golden, white, and silver light energy gathering on the right side of the room. Continue to drain the dark cloud of misery and pain, now occupying the left side of the room, until you have extracted all the new insights from it. As you do so, you will notice the dark cloud becoming heavier and darker.
5. Focus on the right side of the room. Through your intention, begin to inhale and imagine this bright, illuminating energy flowing back into your body. Absorb this knowledge and wisdom gathered from past experiences, void of any feeling of hurt, pain, and judgement into every cell of your body. This distilled knowledge and wisdom empowers you to heal from your wounds on a physical, mental, emotional, energetic, and spiritual level.
6. Next focus your attention on the dark black cloud containing all your pain and suffering. Imagine an angelic being, righteous and holy, holding a "sword of Truth," standing by your side, ready to sever the dark energetic cord(s) still connecting your physical body to this dark cloud of despair. The sword descends and severs the cord(s) (*the doctor claps his or her hands as the sword severs the cord(s)*.
7. The cloud of darkness begins to float up, through the ceiling, through the sky into space. Far in the distance you begin to perceive God or a "higher power" emanating waves of compassion, love, and mercy descending towards the Earth. One of the waves touches the dark cloud of hurt and pain, exploding it into a brilliant light. The cloud is immediately transmuted into fluorescent crystal blue drops of light. This pure, clean energy descends from the Heavens like gentle rain. Breathe in this crystal blue healing energy, let it penetrate deeply into every pore, every tissue and cell, saturating your body completely. Feel the light cleansing and radiating throughout your being.

UNDERSTANDING WHY A PATIENT WON'T HEAL

Normally, in a clinical setting, 80% of the patients will respond favorably to Medical Qigong therapy. In some cases, complete remission of the disease is achieved. An additional 10% will improve, but stabilize in their condition, while a final 10% will not respond at all to the treatments and will progressively get worse. When the patient does not respond to the treatments, the reason can usually be related to the following conditions: Failing to Carry Out the Prescriptions, The Pathogens are Just Too Strong, and Divine Will Guides the Outcome.

- Failing to carry out the prescriptions can be caused by the patient subconsciously sabotaging the healing. Refusing to practice is usually due to deep unresolved emotional issues, e.g., feeling unworthy to heal, fear of abandonment by spouse or others, and so on.
- The pathogens are just too strong to be healed by Medical Qigong therapy alone; when an illness is long-standing, additional complementary treatments, i.e., herbs, Western medicine, or surgery may become necessary to achieve healing.
- Divine will guides the outcome in some cases. The patient's time in the physical realm may be complete, and the soul is preparing to return back to the divine.

Sometimes the healing process takes more time then expected. I explain to patients that permanent transitions can sometimes be instantaneous, but are more likely to be gradual. I explain that they are like a sea captain, standing at the helm of a great ocean liner. If the captain decides that the present course is unsafe, and destructive to the ship and its crew, he quickly orders an immediate change in the course of the ship's direction. With a flick of a switch and a hard spin of the steering wheel, the direction of the rudder is immediately changed. However, because of the perpetual inertia of the massive ship, it will take some time to observe any alteration in the course of the ship's current direction. Although the eyes of the captain can be focused on the new direction, the ship still requires enough time to effectively move in the new direction. It is the patient's spiritual intentions and focused goals that eventually direct the patient's emotional belief system, which in turn, redirects their thoughts and body's physical reactions.

Sometimes, toxic emotions and memories are released through diligent Qigong practice and treatments. With some patients, however, although major changes have already occurred in their conscious attitude, belief structure, and spirit, it may take longer for the unconscious reprogramming to take hold and stop reinforcing the old fears, and traumas. This impediment of unconscious reprogramming is due to the patient's conscious mind not being able to fully integrate the new reality and belief structure. A cancer patient may, for example, have been diligently practicing the prescriptions and the disease may now be in remission, but, the patient still visualizes the cancer as all-invasive. Rather then starting with an image of less cancer in the body, the patient continues to start the meditations with the cancer condition visualized at its worst peak. This sends a double message to the subconscious mind (heal and don't heal).

The same thing applies to observing the patient's transitions of chronically programmed patterns. New behaviors and attitudes cannot become habits overnight. It simply takes time, endurance, and constant encouragement in order to make progress. It is important for the patients to keep their eyes focused on their intended goal and not give up hope.

It is also important not to force mental imagery when feeling angry, depressed or upset. To be effective, creative visualization requires a quiescent mind and soft focus; it should not be hard work. Time allotted to creative visualization can be brief (only a couple of minutes), several times a day, for it to be effective.

FACING THE DEATH OF A PATIENT

Because human beings are created from the energies of Heaven and Earth, they are subject to the transitions of nature, death being part of that transition. Sometimes, no matter who the doctor is, or what his or her specific training has been, there is nothing that can be done to physically heal the patient. Because patients have made attachments during their stay on Earth (physical, men-

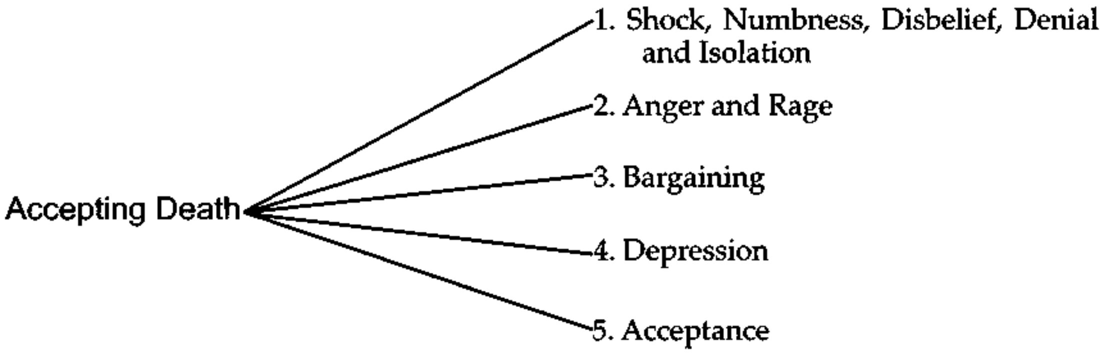

Figure 32.6. The Five Stages and Transitions of Accepting Death

tal, emotional, and spiritual), it is important for the doctor to assist them in their final transitions of releasing and letting go. The fear of death often causes patients to feel loss of control of their life, facing and conquering this fear allows them to regain a sense of control.

According to Dr. Elisabeth Kubler Ross, there are five stages of dying that facilitate closure for patients, especially in the life and death transition (denial, anger, bargaining, depression, and acceptance). These transitions also occur when facing the death and closure of any personal relationship or job, as well as the termination of one's life (Figure 32.6). Although these stages were originally believed to follow each other in an orderly fashion, further research has indicated that these stages need not, and most often do not occur one at a time. The different "stages" can occur at any time going from "denial" to "acceptance," back to "bargaining," etc., (sometimes in just a few minutes). Every individual reacts differently. Some never accept death and struggle to the very end. Medical Qigong therapy describes these transitions as follows: Shock, Numbness and Disbelief, Denial and Isolation, Anger and Rage, Bargaining, Depression, and Acceptance.

1. Shock, numbness, and disbelief occur when an individual is faced with the fact that his or her own death is near. Because of shock, the mind will generalize, rationalize, delete, or distort information to make the information acceptable to the patient's belief system. When these initial feelings of shock, numbness, or stoicism begin to disappear, most patients then experience disbelief. They can not believe or will not accept what is happening to them.

   Some patients bypass this stage all together. They feel great relief at learning of their impending death. They look forward to the cessation of pain and rejoining those close to them who have passed on.
2. Denial and isolation is the first semi-conscious choice the patient makes after hearing the news. The patients' disbelief allows them to establish their own denial system. This denial system functions as a shield to protect them, as they emerge from the shock. Denial and isolation further allows these patients time to gather additional knowledge, hoping that the original diagnosis was incorrect.
3. Anger and rage arise as the denial system breaks down. Patients experience anger, rage, envy, and resentment, as they view their terminal progress as an injustice. The patients' discomfort and anger are usually displaced and projected towards anything and everyone (usually people and God). Patterned, fear-based behaviors get stored in the body and mind. These emotions affect the patients' future perceptions, diminishing their belief system in personal survival.

4. Bargaining is the next stage in the process of adjusting to the verdict. Patients offer counterproposals (usually with God), hoping to change or alter the course of events and postpone the inevitable from happening.
5. Depression replaces hope and the patient experiences a deep sense of loss. Still resisting the outcome, patients sink into depression, sometimes lingering for quite some time before death. Depression is divided into two different categories, each one is different in nature.
   - Regretful depression is caused by guilt and shame for not having lived life to the fullest. The patients regret having given in to fears and having held on to past hurts (both real and imaginary) for so long, that many dreams were abandoned. Opportunities for healing and enjoyment were missed, and old sorrows and conflicts never resolved. Patients with this type of depression often seek to voice their regrets. Many ask to let go of the bitterness and open up to a spiritual healing, rather than a physical healing.
   - Preparatory depression is not caused by regrets, but by mourning the loss of projected future experiences. These patients are usually quiet and reflective, as they face the process of losing everything and everyone that they love.
6. Acceptance is the final stage before death. Having nowhere else to go, patients evaluate what was learned in life. They are now able to express their feelings of envy for the living and the healthy, as well as anger at those who do not have to face their fate so soon. They finally accept the inevitability of death. Depression gives way to a time of rest before the divine journey home. The patients, having faced their life's work, fears, and feelings, now have a new outlook on life and are prepared for the spiritual transition of death.

Facing death is an important transition in life and must be approached with a reverent respect for spiritual transformation. When a patient is dying it is extremely important for the doctor to accept the "higher order" or "divine will" for the patient, and not harbor a win or lose attitude. No matter what stage of transition the patients are in, they will almost always maintain some glimmer of hope, even though they may claim otherwise, to the very last moment of their life. It is the doctor's responsibility to accept and support patients to their inevitable end and not cease to treat them for reasons of personal fears, sorrows, or regrets. Listen to your patients, and allow them to release their fears. This continued support, acceptance, and encouragement for the patient allows them to die in peace and with dignity.

SUMMARY

Death is considered merely a change of residence for the patient's spirit. The ailing body is sometimes viewed as "a house with rotting walls." Once the dying body becomes unusable, it is best to abandon it, and look for another place to stay. The spirit-body is immortal and will exists forever in accordance with the will of the divine. This is why one is taught to be unattached to the physical transformations of this world and accept and embrace death as part of the natural process of existence.

Think of the transition of death as similar to that of a baby chick which is about to be hatched. Although it has lived and existed within its shell all its formative life, as the baby chick continues to develop, it has slowly begun to outgrow its place of existence. The baby chick is now becoming harshly compressed and crushed within the security of its own shell. As the little chick exhausts itself, trying to get free from what has now become its prison, it seemingly faces death. With its apparent last breaths, the little chick shatters the surrounding shell. Immediately, an intense light appears, followed by a whole new world that reveals itself in new wonderment. The baby chick is now reborn into a fresh new realm of existence, with new room to grow and the ability to experience true freedom.

Once, while in China, I was talking to Dr. Wong of the Xi Yuan Hospital, in Beijing. He made the profound statement that everyone he had ever treated had gotten healed. I have never heard such a bold statement being made by a Qigong master, or by any doctor for that matter, so I immediately began to inquire as to what his special techniques were. Dr. Wong responded, "Everyone I treat gets healed; some get healed spiritually, others emotionally, some mentally, and some physically. Not all get cured, but all do get healed. Some patients that come to me for therapy are supposed to die. It is their time of passage, and I help them accept death with dignity and honor. This is achieved by helping the patients heal their spiritual, mental and emotional wounds before they pass on. Other patients that come to me are in an active stage of energetic transition, and I help these patients to see where they've become stagnant, Excessive, or Deficient in their bodies and release their pathogenic Qi." Life is a pathway of transition and change, and so, in the field of energetic medicine, all patients are moving toward or away from healing their relationship with themselves and others.

SECTION VIII

QI EMITTING METHODS

CHAPTER 33
ENERGY EXTENSION AND QI MANIPULATION TECHNIQUES

ENERGY PROJECTION

The body's energetic constitution is a combined matrix of sound, light, heat, and electromagnetic energy. Studies on the effects of Qi projection conducted by Chinese scientists concluded that energy released by a Qigong doctor into a patient carries the properties of infrasound, electromagnetics, static electricity, infrared radiation, gamma rays, particle and wave flows, organic ion flows, and light. These properties produce the physiological changes in the patient's tissues during the Qigong treatment.

A Qigong doctor can emit energy during a treatment with or without touching the patient. The treatment with the doctor's hand moving at a distance from the patient's body is considered Qi Emission. If the doctor's hand touches the patient's body, it is considered Qi Massage, which combines both tissue manipulation and energy flow. Functionally, it is also helpful if the Qigong doctor also has a thorough grasp of the tissue manipulation skills of Chinese massage therapy (Jie Gu for bone setting, Tui Na for traumatology, and An Mo for Internal diseases and visceral manipulation), as well as acupressure and TCM theory, to ensure a highly proficient quality of treatment and to avoid misdiagnosis and erroneous treatments.

Qi is projected outside the body and travels in ripples and vibrational pulses similar to the ripples of an ocean wave. The frequency of this energy wave can be modulated into dense or dispersed resonating patterns depending on the Qigong doctor's needs.

Shen is projected outside the body as light and travels in light beam particles. This particle light extension can change in color and intensity depending on the doctor's intention. When Qi and Shen are combined, the light and wave frequencies facilitate maximum energy projection.

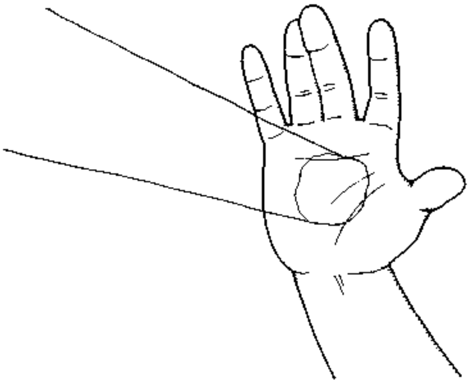

Figure 33.1. Palm Qi Extension

TYPES OF ENERGY EXTENSION

There are nine types of energy extension and projection skills that are divided into three categories based on the source of the projected energy. The first three are projected from the extremities: palms, fingers, and soles of feet. The next category of Qi projection originates from the three Dantians: Upper, Middle, and Lower Dantian. The third category includes Qi emission from the eyes, throat (sound resonation), and thought projection.

ENERGETIC PROJECTION FROM THE EXTREMITIES

The first three energetic projections are released from the extremities (palms, fingers, and the soles of feet) in accordance with the doctor's imagination and intention.

- Palm extension is the ability to release emitted energy through the center of the doctor's palm (Pc-8). It is considered the primary hand posture used by all Medical Qigong doctors for emitting Qi (Figure 33.1).
- Finger extension is the ability to release emitted energy through the extended fingers (usually through the index and middle fingers).

The finger extension techniques allow the Qigong doctors the ability to finely hone their energetic projections into a smaller, condensed energetic beam (Figure 33.2).

- Soles of feet extension is the ability to release energy through the bottom of the feet (Kd-1); this method is usually used to surround and envelop the treatment room as well as the patient (Figure 33.3).

ENERGETIC PROJECTION FROM THE DANTIANS

The next three energetic projections are released from the Upper, Middle, and Lower Dantians. These energetic projections are based upon the doctor's kinesthetic, empathic and intuitive abilities.

- Upper Dantian (Yin Tang) extension from the Qigong doctor's Third Eye area can be used in conjunction with the eyes for intensifying the doctor's Shen projection towards the patient.
- Middle Dantian (CV-17) extension from the Qigong doctor's Heart area can be used in conjunction with projection from the doctor's hands or throat, for vibrational resonation. This technique intensifies the emotional connection between the doctor and the patient.
- Lower Dantian (CV-6) extension from the Qigong doctor's abdominal area can be used in conjunction with Qi emission from the feet for expansion or extension of Wei Qi. This method intensifies the projected External field of Qi around the patient (Figure 33.4).

ENERGETIC THOUGHT PROJECTION OF QI AND SHEN

The last three energetic projections are released from the eyes, throat and mouth (sound resonation) and thought projection. These techniques are a form of long-distance thought projections, based upon the doctor's ability to project Qi and Shen.

- Projection from the eyes releases energy through the eyes and is generally used to increase the doctor's intention and focus of Qi (Figure 33.5).
- Sound resonation projects the voice to tonify or disperse the patient's energy through visceral resonations (Figure 33.6).

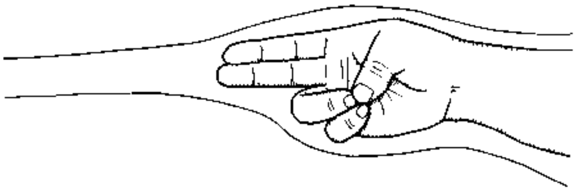

Figure 33.2. Finger Qi Extension

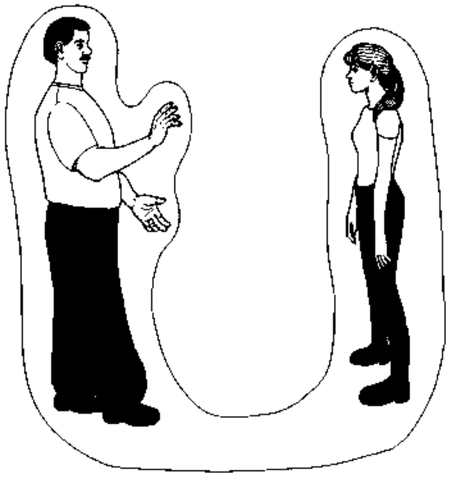

Figure 33.3. Extending Energy from the Soles of the Feet

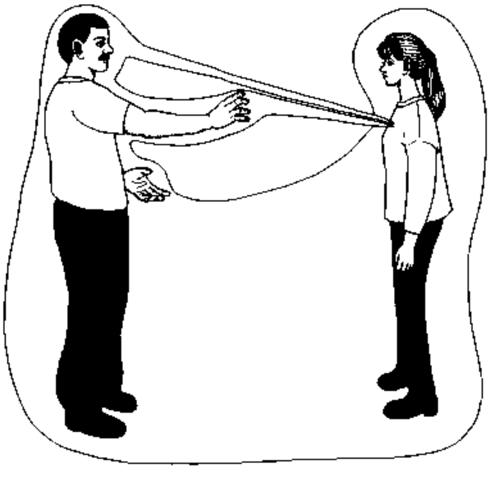

Figure 33.4. Qi Extension. When treating a patient, all Three Dantians extend their energy into the patient.

- Thought projection is another way to release Qi. It is used for long distance energy therapy and is a form of transmuting Qi into Shen, which is transmuted through the Wuji into the patient's Yuan Shen (Figure 33.7).

CONTROLLING ENERGETIC PROJECTIONS

The problem with energy extension lies not in releasing life-force energy but in controlling the power and flow of its projection. If the doctor is distracted or unstable in his or her projection of energy, the doctor may induce more problems in the patient's body. Qi projection can be stabilized in two specific ways:

- First, by visualizing and projecting specific colors (blue, red, violet, etc.) with the energy extension. This causes the Qi to take on a dense and fuller projection, which facilitates the rooting and stabilizing of the energy into the patient's tissues.
- Second, by using intention to mentally focus the size and shape of the projected energy (like a laser), to intensify the Qi projection.

When treating a patient, the diseased tissue must be sufficiently energized. Inadequate energizing results in only slight improvement or a slower rate of healing; over-energizing may cause Qi congestion.

The degree of curative effect depends on the competence and versatility of the doctor, as well as the responsiveness of the patient. If the patient is not responsive, repeated stimulation and regulation of Qi is needed to raise their sensitivity.

Sometimes the patient's emotion and spirit are outside and disconnected from their body and mind. Although this is a perfect adaptation for a military environment, it is not conducive for an energetic healing environment. The patient must undergo a series of Shen Gong training sessions to retrieve their emotional spirit. This training is commonly referred to as "soul retrieval" because the patient's soul is in a state of shock (or suspended animation) due to severe emotional trauma, and must be returned to its Yuan Jing within the patient's tissues. (See Chapter 19).

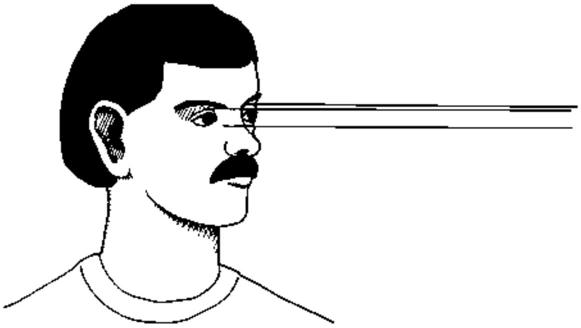

Figure 33.5. Energy Projection from the Eyes

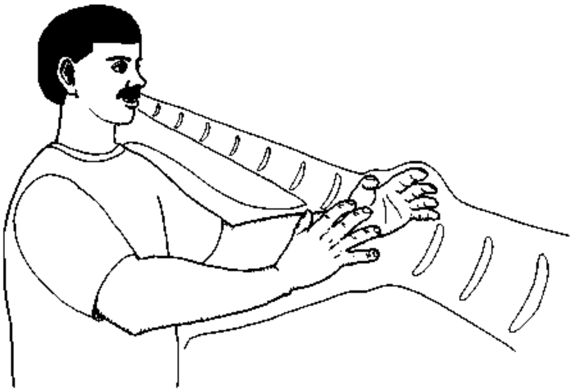

Figure 33.6. Sound Resonation Projection

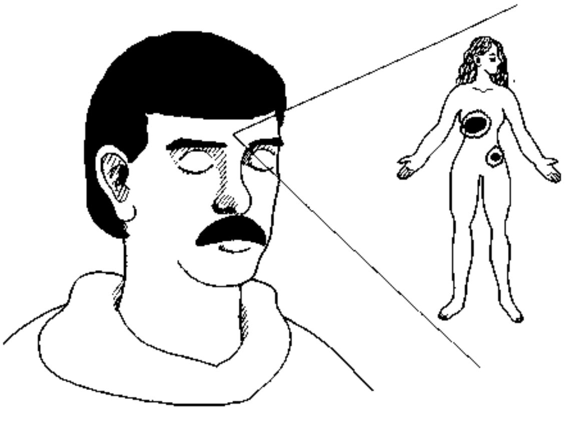

Figure 33.7. Thought Projection

In addition to Shen Gong training, the Qigong doctor must make the patient's Qigong meditations and prescriptions more active so that a satisfactory effect can be achieved. However, a desirable effect can only be attained when the patient is cooperative with the doctor and is consistent with the assigned homework.

HAND POSTURES FOR QI EMISSION

There are many hand postures used in emitting and drawing in energy. Each variation depends on the skill and particular background of the doctor. Generally the doctor uses the entire hand to extend energy, constantly changing the intensity of the Qi being extended. By expanding and contracting the hand and fingers, the "beam of Qi" being emitted into the patient's body will expand and contract. Presented here are six hand postures most preferred by myself and many other Medical Qigong doctors in clinics throughout China.

THE EXTENDED FAN PALM

This first hand posture is the most popular in terms of Qi emission. It releases the widest field of Qi emission. In the Extended Fan Palm hand posture, all five fingers stretch naturally, separate, and open like a Chinese fan (Figure 33.8). There should be no tension, and the focus of the energy should be in the center of the palm (Pc-8). This Extended Fan Palm is the primary hand posture for extending energy because it can be easily used for emitting either Hot or Cold energy. It can also be used in conjunction with other techniques as a reinforcing method to strengthen the flow of energy extension.

Note: Before beginning these exercises, it is important not to eat a heavy meal (at least one and a half hours before practicing).

PALM BREATHING ENERGY EXERCISE—STAGE 1

1. From a Wuji posture, bring both hands up in front of the body, with the palms facing the Lower Dantian. Breathe naturally while concentrating on the Lower Dantian. Next, place

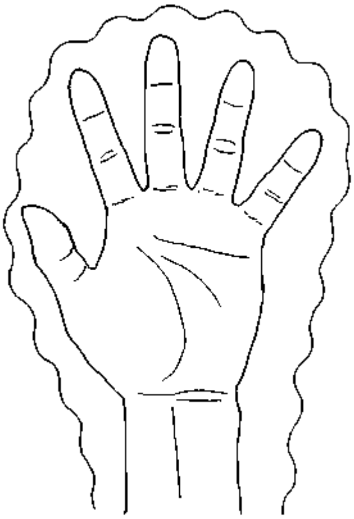

Figure 33.8. The Extended Fan. This palm is the primary hand posture used for emitting Qi.

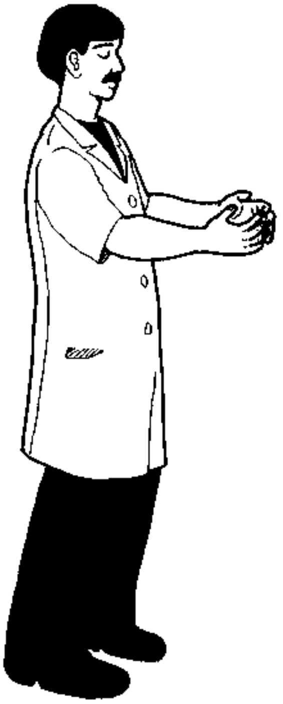

Figure 33.9. The Image of Holding a Ball of Energy

the left palm facing the right palm, as if holding an imaginary ball. Allow the energy between both hands to build and increase its energetic field (Figure 33.9).

2. Press the Mingmen backwards and feel the energy absorbing through the arms into the Lower Dantian. Next, release the pressure from the Mingmen and feel the Qi return back into the center of the palms.
3. Shift the attention to the center of the right palm, and begin to extend energy into the center of the left palm. Next, shift the attention from the center of the right palm, and begin extending energy into the center of the left palm.
4. As a feeling of pressure begins to build between the hands, perform the "Pushing" (energy extension) and "Pulling" (energy retraction) manipulation technique. This technique is practiced by pushing the Qi of the right palm into the center of the left palm. Hold it there, feeling the pressure mount, then withdraw the Qi back into the center of the right palm. Repeat for several minutes, then move to step five.
5. After several minutes switch the mind's focus to the opposite hand and perform the exercise again. Do this exercise twice a day for 15-30 minutes each practice. Be careful not to let the hands touch when running the energy back and forth between the palms. After 100 days go to Stage #2.

When training to increase their own energetic field, the Qigong doctors must strive to overcome the Pituitary Pain Syndrome. This is a state of transition during which the body is overcome by pain to the extent that the mind gives up and the doctor quits practicing (e.g., feeling fatigue when practicing a Holding the Ball posture). It is important that as the doctor become stronger, the training posture be maintained longer, so that the pituitary gland is challenged to activate the hypothalamus. The hypothalamus causes the body to produce endorphins such as ACTH, which is a natural form of cortisone that helps to escalate the doctor's healing potential. This physical reaction is similar to a runner who must overcome physical cramping when racing to obtain a second wind.

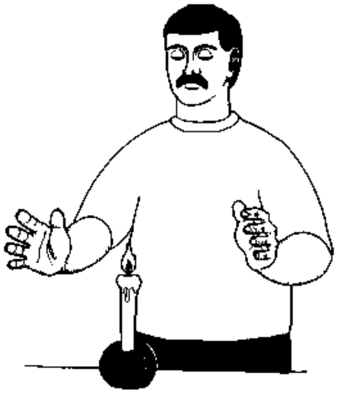

Figure 33.10. Posture for Palm Breathing Exercise

PALM BREATHING ENERGY EXERCISE—STAGE 2

1. Light a candle and place it on a table. From a Wuji posture place the center of the right palm several inches to the right of the tip of the flame (Figure 33.10). The left palm should be placed on the opposite side of the candle, facing the right palm. The Laogong points and the tip of the candle should be in a straight line with each other.
2. Breathe naturally while concentrating on the Lower Dantian. Shift the attention onto the center of the right palm, and begin extending energy into the center of the left palm.
3. Once a feeling of pressure begins to build between the palms, begin the Pushing and Pulling energy manipulation technique. Push the Qi of the right palm into the left palm. Hold it there for several minutes, feeling the pressure mount, then draw the Qi back into the center of the right palm. Practice this exercise for several minutes.
4. Next place the left palm on the Lower Dantian while continuing to extend and retract energy from the right palm. After several minutes, switch hands. Notice that as the Qi is being extended out of the palm that the flame bends away, and as the Qi is retracted back

into the palm the flame bends towards the center of the palm. Practice this exercise twice a day for 15-30 minutes each time.

Note that the Yin Tang area (the Upper Dantian's projection point) can grasp the flame and hold it stationary. Since this technique (grasping and rooting) is used for rooting and securing the patient's Qi, its application can be counter productive when trying to move the flame tip. To avoid this condition, the doctor should relax the mind and use the imagination when projecting Qi.

THE SWORD FINGERS

In this second hand posture, the expansive field of Qi being released is reduced and condensed into a more solid beam of light. In the Sword Fingers hand posture, the index and middle fingers of the right hand are extended and joined together, the ring and little fingers curl into the center of the palm (Figure 33.11). The thumb should rest on the nails of the ring and little fingers forming a circle where Qi is gathered.

The index finger indicates Wood or Liver's Qi, and the middle finger indicates Fire or the Heart's Qi. Together, both fingers activate the Qi of Fire and Wood, which is a catalyst for the Yang energy needed for guiding Yang Heat with the intention (Yi). When extending energy, the Qi is gathered first in the palm and the circle between the thumb and last two fingers, then it is released through the tips of the index and middle fingers.

SWORD FINGER QI EXTENSION—STAGE 1

1. From a Wuji posture, extend both hands to form an imaginary circle of energy in front of the Lower Dantian. As soon as the Dantian is activated and full, bring the ball of energy in front of the body, level with the Middle Dantian (Heart / solar plexus level). Direct the Qi from the Lower Dantian into both hands (Figure 33.12).
2. Point the middle and index fingers of the right hand towards the left palm (Pc-8) and feel the energy press into the tissues (Figure 33.13). After a while, neutralize the energy by turning the palms towards each other, then switch and feel the energy of the left sword fingers

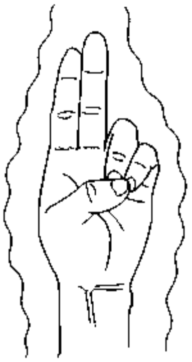

Figure 33.11. The Sword Fingers Technique

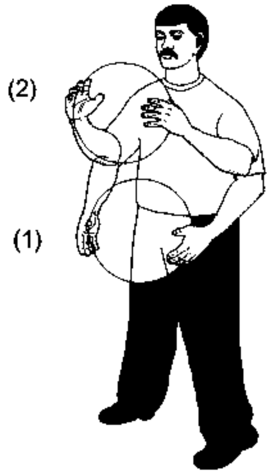

Figure 33.12. Imagine a ball of energy in front of the Lower Dantian, then raise it up to the level of the Middle Dantian.

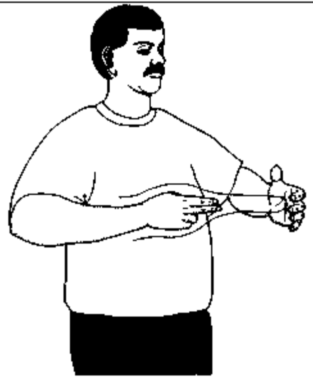

Figure 33.13. Point the Sword Fingers of the right hand towards the center of the left hand.

(index and middle finger) press into the center of the right palm. Practice this exercise twice a day for 15-30 minutes each practice. After 100 days go to the next stage.

SWORD FINGER QI EXTENSION—STAGE 2

1. Light a stick of incense and place it in a standing upright position on a table. From a Wuji posture, place the center of the left palm (Pc-8) several inches to the left of the incense. The right hand sword fingers point through the tip of the incense, extending Qi into the center of the left palm. All three points - left center palm (Pc-8), tip of the incense, and the right sword fingers - should be in a straight line (Figure 33.14).
2. Breathe naturally while concentrating on the Lower Dantian. Shift your attention onto the tip of the incense and focus your concentration there. Continue to extend Qi until you have a strong feeling of Qi in the center of your left palm. After a while switch to the right palm. Practice this exercise twice a day for 15-30 minutes each time.

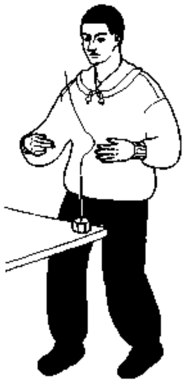
Figure 33.14. The Sword Fingers point at the tip of the lit incense.

THE ONE FINGER SKILL

In this third hand posture, the field of Qi being released is further condensed into an more solid beam of light. In the One Finger Skill hand posture, the hand forms a soft fist with one appendage (finger or thumb) extended (Figure 33.15). The energy is projected in a straight line (like a laser beam) and can be used to Heat the patient's tissues, channels, or points, as well as to lead, extend, and tonify the energetic field (Figure 33.16).

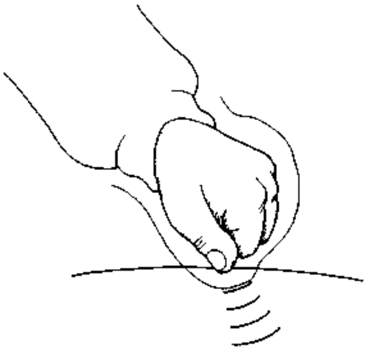
Figure 33.15. The use of the finger or thumb to emit Qi is an example of the One Finger Skill technique.

One Finger Skill can be used to extend long range energy into the patient's channels, or for tissue and channel point regulation. The former technique is regarded as an internal application, and the later technique is regarded as an external application.

- The external application requires that the doctor make physical contact with the patient's tissues, extending the mind deep into the specific areas, channels, or points to be treated. The Qigong doctor physically rotates and shakes a specific finger (or thumb) from side

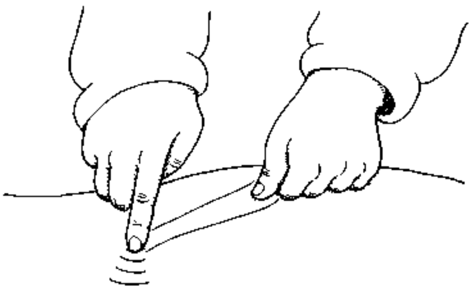
Figure 33.16. Energy is extended through the One Finger Skill technique.

to side while emitting Qi into the patient's tissues or energetic fields (Figure 33.17).

- The internal application requires that the Qigong doctor extend his or her arm, pointing a specific finger at a particular area on the patient's body. The doctor then focuses on that specific channel point, leading and directing the emitted Qi into or out of the patient's body.

ONE FINGER SKILL ENERGY EXTENSION EXERCISE

1. From a Wuji posture, focus the concentration on the Lower Dantian area and bring both arms up to chest level and form an energy ball. Once the Qi in the Lower Dantian becomes activated, direct its energy flow to the tip of the right index finger. Once a feeling of heat and expansion of energy is felt within the right index finger, point the energy beam into the heart of the left palm (Figure 33.18).
2. Once the energy is felt pressing into the left palm, begin to circle-rotate the energy beam in a clockwise direction for several minutes, then in a counterclockwise direction for several minutes.
3. Next, extend and focus the Qi from the Dantian into the left index finger, and repeat the exercise.

SOARING DRAGON TECHNIQUE

The Soaring Dragon technique is a variation of the One Finger Skill energy extension technique. If the pathogens are strong and too much Heat is being emitted from the patient's tissues (as in certain types of cancer), the Qigong doctor may decide to use this particular variation to avoid absorbing the dangerous toxins being released from the patient's body. The Qi is emitted out the doctor's middle finger into the patient's body while the other four fingers disperse the pathogenic Heat into the ground (Figure 33.19).

Before initiating the Soaring Dragon technique, it is important for the Qigong doctor to use his or her mind's intention to seal the wrist. This enables the doctor to pull the patient's Qi into the palm and to disperse it into the Earth via the other four fingers, while avoiding the absorption of the patient's pathogenic Qi.

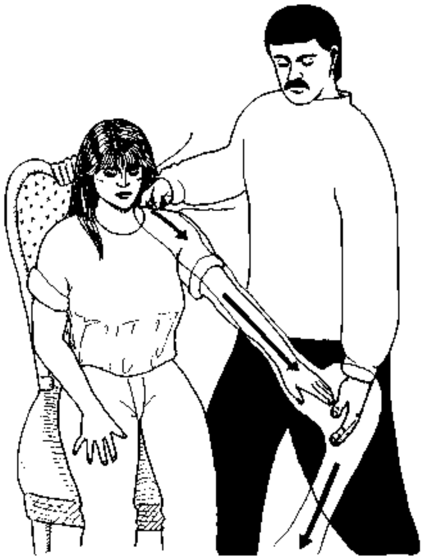

Figure 33.17. The doctor makes contact with the patient's tissues, extending his or her intention along the patient's channels, leading the Toxic Qi out her body.

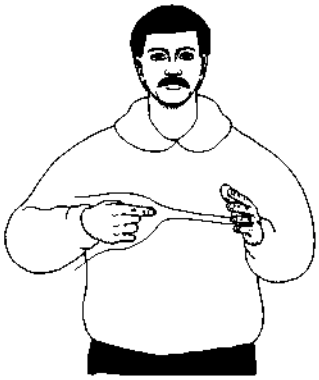

Figure 33.18. One Finger Skill Energy Extension Exercise

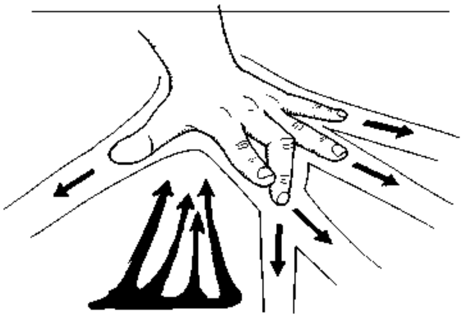

Figure 33.19. Soaring Dragon Hand Posture Technique

I have found this technique very effective in my personal practice. As the pathogenic Heat from the patient's tissues releases upward, the dispersing energy that flows from the four fingers cause the patient's toxic Heat to dissipate and be carried away from the doctor.

BELLOWS PALM TECHNIQUE

Another variation of the One Finger Skill is the Bellows Palms technique, which uses the Lao Gong Pc-8 point at the center of the palm as a pump, to expel Qi out the middle finger. Originally, while treating patients, the palm is Hot (in a Yang state). However, when pulling Qi from the center of the palm and directing it out the middle finger, the polarity of the palm changes from Yang to Yin, causing the center of the palm and the middle finger to turn white (Figure 33.20).

Because this technique is used to pull out Heat from the patient's body, before initiating the Bellows Palm it is important for the Qigong doctor to use his or her mind's intention to seal the wrist. The doctor then pulls the patient's Qi into the palm and disperses it into the Earth via his or her middle finger to avoid absorbing any of the patient's pathogenic Qi (Figure 33.21).

THE BACK BRIDGE BAR TECHNIQUE

The Bellows Palm technique is also used with the "back bridge bar" (where the trapezius muscles join both arms across the doctor's shoulders) to aspirate toxic energy from the patient's body. In this particular technique, the doctor's left palm is used as a vacuum to aspirate a specific area inside the patient's body. The toxic energy travels through the doctor's left arm, across the shoulders and back (never the front), and out through the right palm into the Earth. The right palm applies the "bellows" technique used to create the energetic suction applied by the left palm, for removing Toxic Qi. This technique is contraindicated when treating cancer patients.

THE INVISIBLE NEEDLE PALM

In this fourth hand posture, the field of Qi being released is further condensed into an intense beam of light, similar to that of a laser beam (Figure 33.22). In the Invisible Needle Palm hand posture,

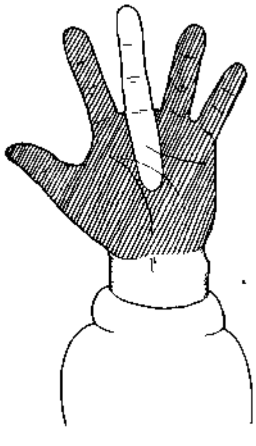

Figure 33.20. When using the Bellows Palm technique, the center of the palm and middle finger remain white in color, while the rest of the palm remains red.

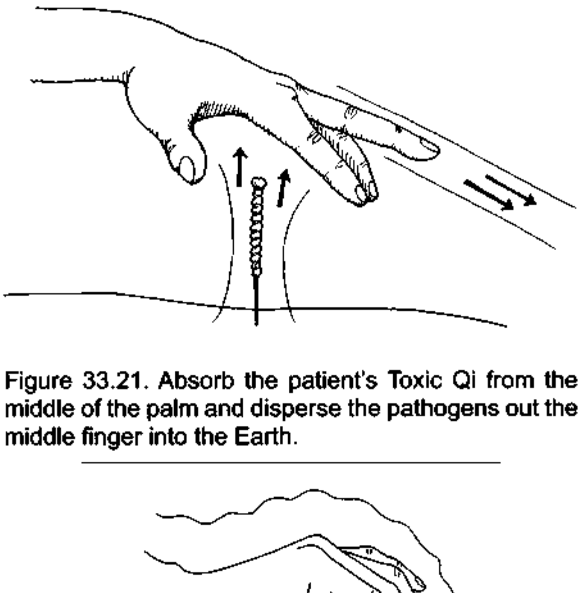

Figure 33.21. Absorb the patient's Toxic Qi from the middle of the palm and disperse the pathogens out the middle finger into the Earth.

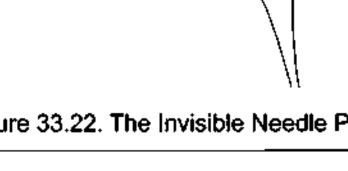

Figure 33.22. The Invisible Needle Palm

three fingers on the right hand are curled into the center of the palm with the thumb resting against the tip of the index finger (see Chapter 35 for Invisible Needle Theory and Application).

THE DRAGON'S MOUTH PALM

This hand posture is sometimes also called the Duck's Bill Palm, and is used to pull energy (like pulling a silk thread). The thumb and all four fingers perform the action of "pulling the golden thread" in order to lead or pull energy along and through the body's channels and points. The intensity of the Qi and movement between the thumb and fingers is determined by the relaxed movements of the wrist and finger joints. This hand technique is often used to purge energy away from specific points when treating areas of Excess (Figure 33.23).

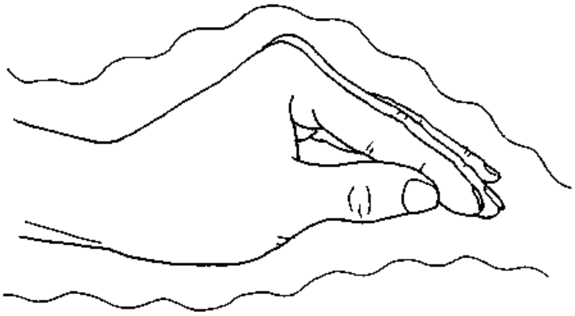

Figure 33.23. The Dragon's Mouth Palm

THE KNEADING TIGER PALM

In this hand posture, all five fingers naturally curl as if embracing a ball. The wrist action is very fluid and supple, allowing the fingers to expand open like the Extended Fan Palm, and then gently close as if kneading a cotton ball. The entire action should be similar to a jellyfish moving in the ocean. This is an excellent technique for pulling, shaking, and rotating energy (Figure 33.24).

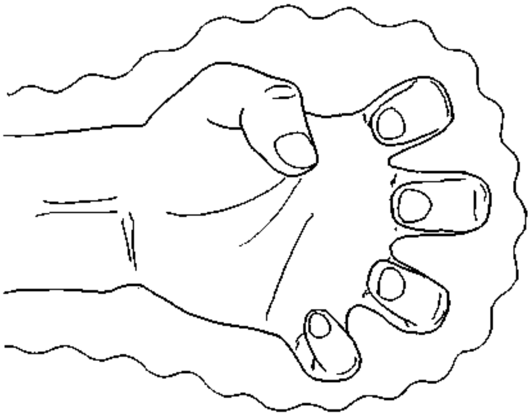

Figure 33.24. The Kneading Tiger Palm

KNEADING TIGER TECHNIQUE

This hand posture is used for dispersing stagnations and dissolving energy blocks. Once the doctor has selected the proper hand posture and the area to be treated, the doctor places his or her hands and fingers above the patient's body, performing a circular kneading action (slightly opening and closing the palm) while simultaneously extending energy into the patient's body. The spiraling energetic frequency should be synchronized with the doctor's hand kneading and the patient's pulse and respiration. Both the pressure and the strength of energy extended into the patient are determined according to the severity of the illness. The Tiger Kneading technique is used for extending energy into energetic points of the chest, abdomen, and extremities to purge Qi stagnation and improve energetic circulation.

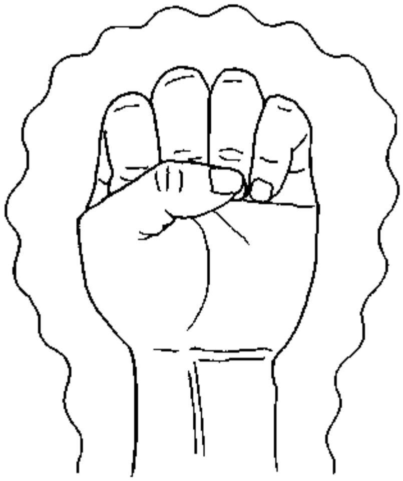

Figure 33.25. The Five Thunder Finger Palm

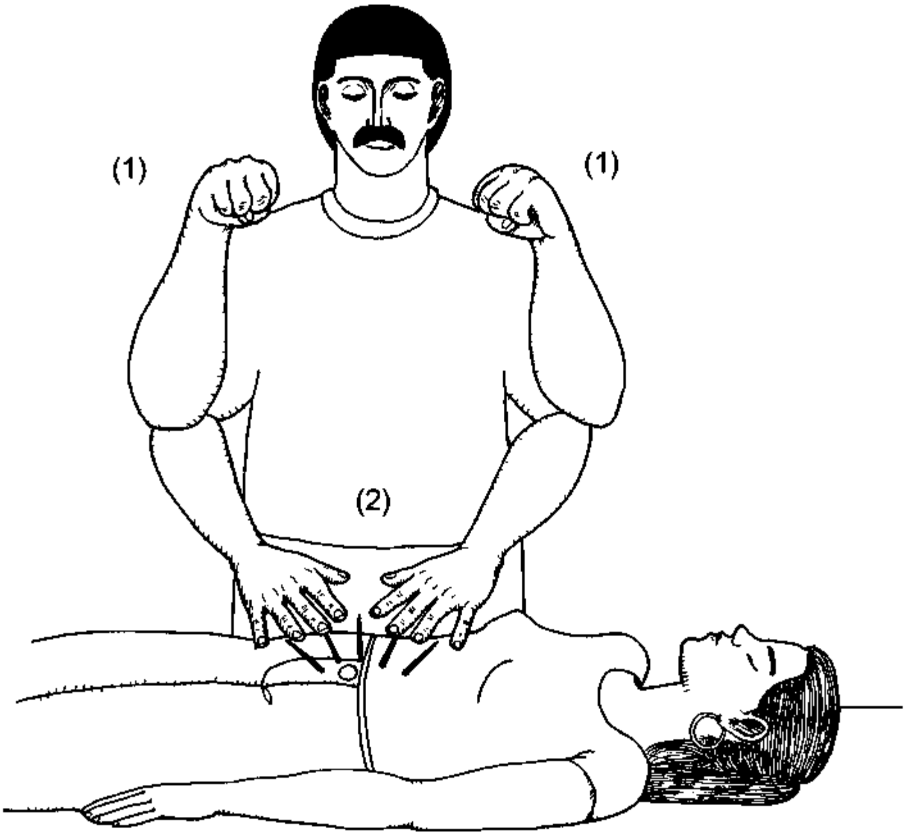

Figure 33.26. The Five Thunder Fingers. This technique is used primarly to destroy cysts, tumors and pathogens.

THE FIVE THUNDER FINGERS

This hand posture is one of the most popular techniques used for dispersing stagnations, dissolving energy blocks and treating tumors and cysts. In the Five Thunder Fingers hand posture, all four fingers curl into the center of the palm with the thumb resting against the nail of each finger. Keep the center of the palm hollow. Qi is gathered and collected there upon inhalation (Figure 33.25).

When emitting or extending Qi, the doctor exhales while suddenly stretching the palm open like "a tiger exposing its claws." After extending the energy, the doctor inhales and resumes the Five Thunder Fingers posture while gathering Qi back into the center of the palm (Figure 33.26).

Note: Because it builds up the doctor's guiding and gathering skill, this is an important technique and should be practiced consistently.

QI BLAST EXERCISE

From a Wuji posture, raise both palms upward to shoulder level. Both hands form the Five Thunder Fingers posture. Concentrate on the Lower Dantian when inhaling. Imagine drawing the Qi up through the chest, gathering the energy into both palms as the arms raise upward. Upon exhalation, sink the body's weight, and suddenly stretch out the fingers (like a claw) at chest level, striking the air. This action should look like each hand is heaving a rubber ball. Perform this exercise once or twice a day, using 24 or 48 breaths each practice time.

CHAPTER 34
QI EXTENSION, GUIDANCE, AND REGULATION TECHNIQUES

Qi extension is most effective in activating and balancing the patient's flow of Qi, since both the doctor and patient are enveloped in a strong field of Wei Qi. In Qigong massage the patient's channels are more easily dredged and their Yin and Yang energy are often adjusted much more quickly because of the stimulation produced by the doctor's hand on the patient's channels and points.

Therefore, both modalities - Qigong massage and Qi extension - are usually combined to provide the most effective and efficient means of energetic therapy.

QI EXTENSION TECHNIQUES

The clinical guiding and directing of Qi is performed by extending energy outside of the doctor's body. The energy follows the dictates of the doctor's mind as to the pattern, amount, and direction in which to flow. Qi is emitted in three basic formats: linear, circular/spiral, and stationary flows of energy. The combination of these three forms insures that the patient is receiving the maximum benefit from the treatments, and may be combined with Hot and Cold energy extension, the Five Elemental energy extension, as well as with Qigong massage. These following exercises lay a foundation for guiding and directing the patient's Qi within the channels, points, and Wei Qi fields.

GUIDING THE LINEAR FLOW OF ENERGY

This method refers to a straight line of energy being emitted either between both hands or between one of the doctor's hands (usually the right) and a certain point or area on the patient's body. This is the primary Qi guiding method used in Medical Qigong therapy to sweep over the patient's body in a straight line in the direction of the outgoing flow of energy.

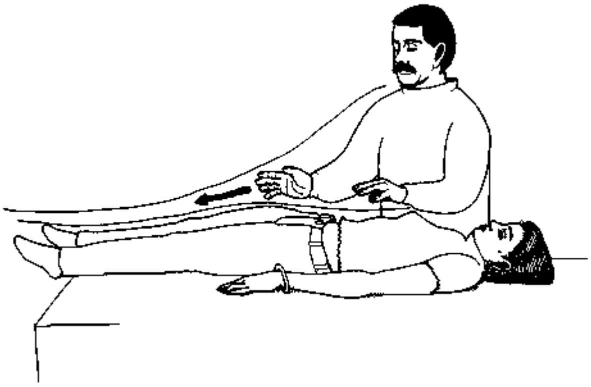

Figure 34.1. An example of a Qigong doctor using the Pushing Energy technique in a Straight Linear energy flow

Because patients vary in their sensitivity to transverse energy wave flow, the doctor should be aware of the influence of the different wave patterns affecting the patient during treatment. Linear energy is relatively mild and often gives the patient a sensation of constriction, of heat or cold, or of heaviness, pressing, tugging, tingling, or pulling. This is the basic means of inducing the channel Qi to purge Excess and to supplement any Deficiencies. Energy is directed within the body through either pushing, pulling, or leading the patient's Qi.

PUSHING ENERGY

Pushing energy is initiated by first extending energy over the surface area of the patient's body (Figure 34.1). Once the doctor has selected the appropriate hand posture above the patient (between 4 inches to 3.5 feet from the channel point or area), he or she will decide whether to push the energy in a linear or circular fashion.

Using intention, the Qigong doctor gently extends energy to the desired area and begins treatment. Pushing energy is an essential energetic

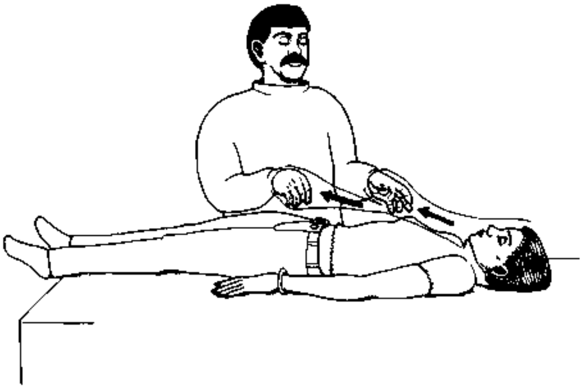

Figure 34.2. An example of a Qigong doctor using the Pulling Energy technique in a Straight Linear energy flow

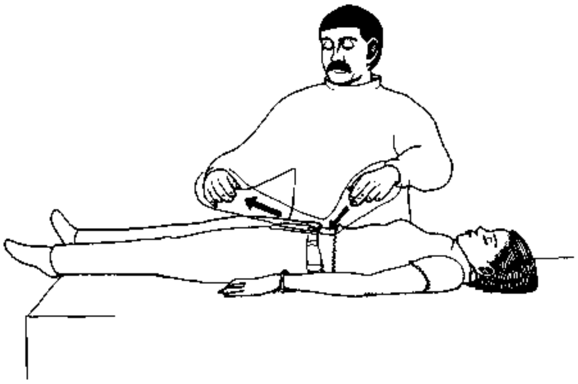

Figure 34.3. An example of a Qigong doctor using the Leading Energy technique (with the left hand) while Pulling Toxic Energy (with the right hand) from the patient's body

manipulation used to open up the points to activate or replenish the body's Qi. The patient feels pressure from the doctor's Qi, and may experience heat, tingling, condensation, expansion, swelling, or heaviness.

Sometimes when the doctor uses the pushing energy technique, the patient may spontaneously lift the stimulated body area towards the hand of the doctor. This spontaneous reaction enables the doctor to implement a combination of manipulation therapies such as push-pull, push-pull-rotate (spiral), push-pull-shake, etc.

PULLING ENERGY

Pulling energy is initiated first by extending energy over the surface area of the patient's body. After selecting the appropriate hand position, the doctor chooses to either pull the energy in a straight line or spiral it along the channels.

Using intention, the Qigong doctor usually begins to pull the patient's toxic energy along the channels while extending clean energy to the affected area (Figure 34.2). Pulling energy is an essential energetic manipulation used for dredging and expelling pathogenic factors from the patient's body, and can also be used to open up channel points to activate or replenish the patient's Qi.

During treatment the doctor actually feels the pathogenic factors being pulled out of the patient's body. This causes the patient to feel heavy, tingling, dryness, heat, cold, or chills. When combining multiple manipulation methods, the pushing and pulling actions cause the energy inside and outside of the patient's body to flow into each other, opening up the patient's channel points.

The combination of pulling and shaking techniques promotes the flow of the patient's Qi, and induces spontaneous reactions of the tissue's energetic fields.

LEADING ENERGY

Leading energy is initiated first by extending energy over the surface area of the patient's body. After selecting the appropriate hand posture, the doctor extends energy towards the affected area and begins to lead the patient's channel Qi up, down, left, right, with, or against the flow of the body's channels (Figure 34.3).

The Qigong doctor bases the decision of whether or not to lead the patient's energy (and which technique should be used) upon the patient's condition and severity of the illness. This manipulation technique is used for guiding the circulation of channel Qi, regulating any Excess or Deficiency of Yin and Yang energy and to transfer Qi back to its origin. Once the channel Qi and point Qi have been stimulated, it is important that the doctor immediately use the "leading energy" technique to guide the patient's energy in as smooth a transition as possible.

CIRCULATING THE QI IN A CIRCLING PATTERN

Emitting a flow of energy in a circling pattern is used to move and circulate stagnant Qi. The flow of energy can be moved in a continuous pattern with or against the channel flow, and is used to regulate the patient's channel and organ Qi (Figure 34.4).

This method refers to techniques of rotating Qi in a circle pattern inside the patient's body. When using circling energy, the doctor extends Qi into the patient's body and implements the circle technique with the right (Yang) hand, while rooting and discharging the patient's toxic Qi with the left (Yin) hand, which points to the Earth. There are many variations of this technique and the doctor must become familiar with each modality of healing and its potential for treating the patient.

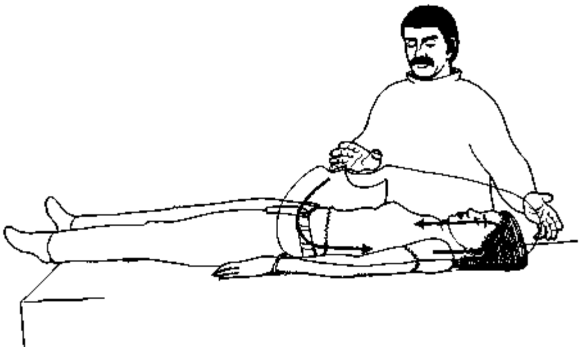

Figure 34.4. An example of a Qigong doctor using the Circling Energy technique to regulate the Heart Fire and Kidney Water, through the Microcosmic Orbit

ENERGETIC CIRCULAR PATTERNS

The following patterns are specific templates used for circulating the body's internal and external Qi. Each pattern has an advantage over the others when used for the regulation of Yin and Yang disharmony. The depth of energetic penetration will vary in accordance with the doctor's intention. Each pattern's movement will utilize the ring of energy that surrounds the center of each Dantian (Figure 34.5).

1. **The Eternity Pattern** is used for balancing the energy between the Lower Dantian and the Middle Dantian. The doctor moves the energy in a figure "8" pattern, Guiding and Leading the Qi from the Lower Dantian

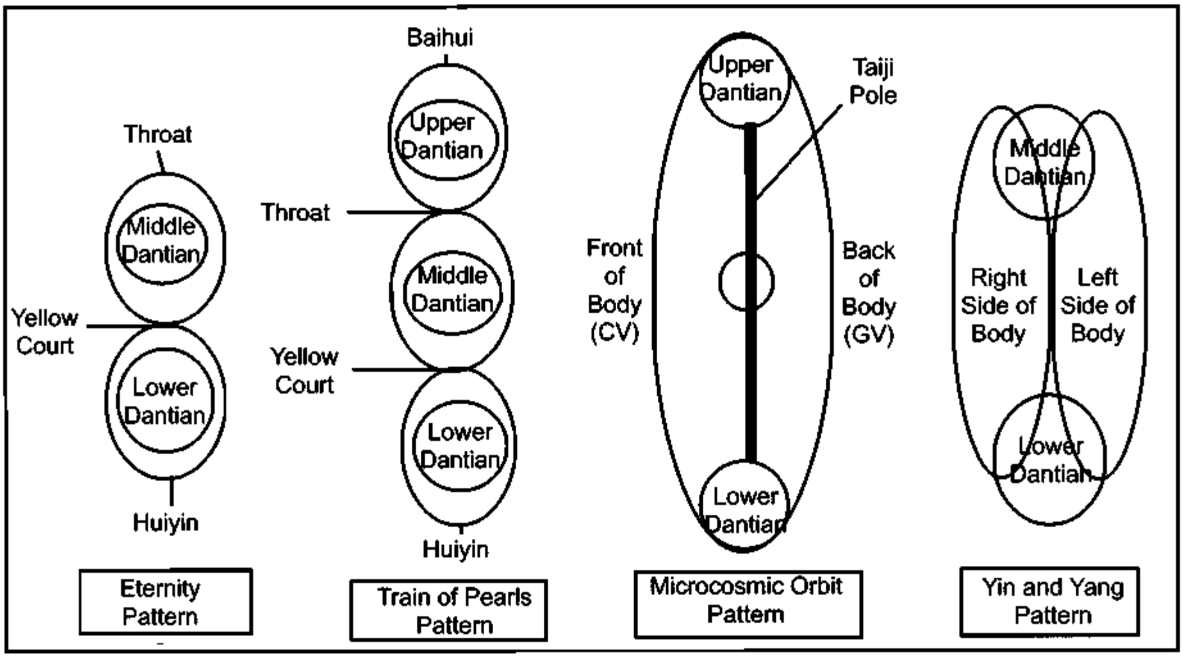
Throat
Middle Dantian
Yellow Court
Lower Dantian
Huiyin
Eternity Pattern
Baihui
Upper Dantian
Throat
Middle Dantian
Yellow Court
Lower Dantian
Huiyin
Train of Pearls Pattern
Upper Dantian
Taiji Pole
Front of Body (CV)
Back of Body (GV)
Lower Dantian
Microcosmic Orbit Pattern
Middle Dantian
Right Side of Body
Left Side of Body
Lower Dantian
Yin and Yang Pattern

Figure 34.5. A Qigong Doctor Using Circling Energy

(through the Yellow Court) to the Middle Dantian, and back again. The doctor continues this pattern until the patient's energy is regulated. If the patients are suffering from an Excess Yang condition in the upper part of the body, and a Deficient Yin condition in the lower part of the body, the doctor must regulate this condition by using the Eternity Pattern of Qi regulation.

2. **The Train of Pearls Pattern** is used for balancing the energy between all three Dantians. The doctor moves the energy from the Lower Dantian (through the Yellow Court) to the Middle Dantian (through the throat), and into the Upper Dantian, then back down again. The Qi is swept up and down the entire torso, combining and regulating the energies of the patient's Wei Qi, Channel Qi and Organ Qi. The doctor continues this pattern until the patient's energy feels regulated. If the patients are energetically disconnected from the upper and lower parts of their body due to emotional congestion and energetic stagnation in the throat and Yellow Court areas, the doctor reconnects these areas by using the Train of Pearls Pattern of Qi regulation.
3. **The Microcosmic Orbit Pattern** is used for balancing the energy between the Three Dantians, Taiji Pole, the Heart's Fire and the Kidneys' Water. The doctor moves the energy from the Lower Dantian (Mingmen area) to the Middle Dantian (Shendao Area), then back down again, completing the Fire Cycle of the Microcosmic Orbit. The doctor continues this pattern until the patient's energy is regulated. The Microcosmic Orbit Pattern is used for the complete regulation of the body's Yin and Yang energy, naturally purging, tonifying, and regulating any Excess or Deficient condition along the Governing and Conception Vessels.
4. **The Yin and Yang Pattern** is used for balancing the energy between the right and the left sides of the body. The doctor moves the energy in a horizontal figure "8" pattern, Guiding and Leading the Qi through the left and right quadrants of the patient's body. In cases of paralysis after a stroke, for example, the doctor regulates the Qi of the patient's left and right quadrants and energetically balance these areas by using the Yin and Yang Pattern of Qi regulation.

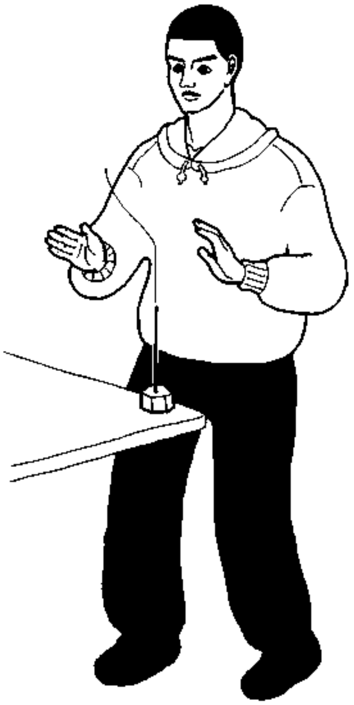

Figure 34.6. An example of a Qigong doctor moving the incense smoke, using the image of an energy ball

The Qigong doctor selects the appropriate energetic pattern according to the patient's condition.

ENERGY BALL ROTATION FOR DEVELOPING CIRCLE ENERGY FLOW

- Light a stick of incense and stand it on a table. From a Wuji posture, place the center of both palms on opposite sides of the tip of the incense. All three points, the tip of the incense, and the heart of each palm (Pc-8), should form a triangle (a tree or flower can be substituted for the incense).
- Breathe naturally while focusing on the Lower Dantian. Slowly shift the attention to lead the energy into the center of each palm to form a ball of energy (basketball size). Concentrate and

imagine the ball condensing to form a powerfully compacted ball of energy (about the size of a marble). Next, imagine a circle or rounded triangle between both palms and the tip of incense. Exhale lightly while extending the energy ball toward the tip of the incense. Imagine the ball of energy moving counterclockwise from the right palm to the incense and then to the left palm, moving in a circle. The right hand pushes while the left hand pulls. The incense smoke should move in accordance with the Qi movement. After a while reverse the flow of the circle. Practice this exercise twice a day for 15-30 minutes each time (Figure 34.6).

SPIRALING ENERGY FLOW

A spiraling flow of energy moves the energy in a spiral pattern (clockwise or counterclockwise) to drill the energy deep inside the patient's body (Figure 34.7). This "drilling" action, caused by the spiraling energy, allows for deeper penetration and is excellent for breaking up stagnations when combined with the Kneading Tiger Palm or the Five Thunder Fingers techniques.

Spiraling energy can also be used in the clinic to regulate the patient's Qi activities. Patients may feel the energy penetrating deep into their body and experience light, sound, and mild electric shock. When patients are lying supine, the doctor may choose to spiral the energy superior to the internal organs to pull toxic Qi out of the tissues of the organ; or the doctor may spiral the energy inferior to the internal organs to push the Qi back into the tissues of the organs (Figure 34.8).

Spiraling energy is an essential energetic manipulation used for activating the channel Qi, guiding the energy to spiral in, up, or down, much like a cyclone. The Qigong doctor chooses to use either the circle or spiral energy flow, based upon the patient's condition. Most often these two energetic flows are combined to achieve a more functional treatment. The Qigong doctor will, for example, use "spiraling" energy to drill deep into the body to access the toxic energy stored deep inside the patient's organs. Once the patient's toxic Qi has been reached, the doctor utilizes purging techniques to remove the energetic debris. After purging the doctor can circulate the patient's energy to move the stagnant Qi out of the organs. This may be followed by the Microcosmic Orbit to regulate the patient's energy.

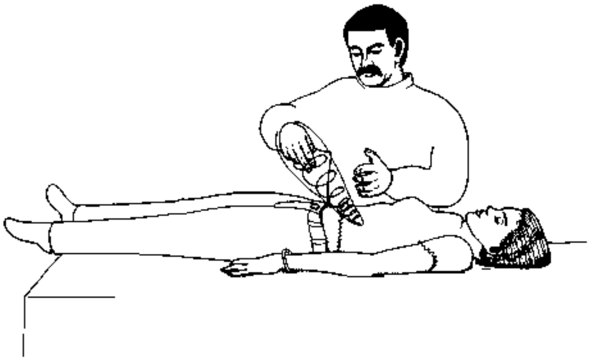

Figure 34.7. A Qigong doctor using the Spiraling Energy technique, to purge Toxic Qi from the patient's body

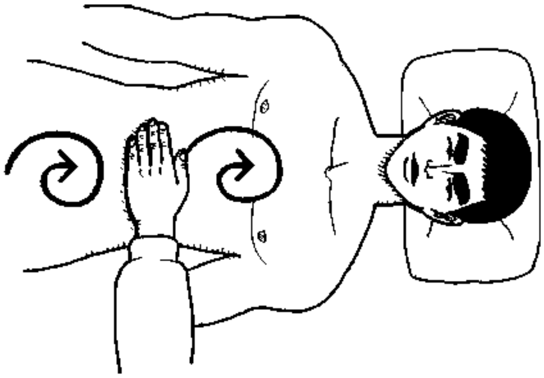

Figure 34.8. A Qigong doctor Spiraling the Qi in order to penetrate the patient's tissues

CLOCKWISE AND COUNTERCLOCKWISE ENERGY FLOW

The doctor focuses on sensing and creating an internal energy vortex within the Lower Dantian that moves up the body in a clockwise or counterclockwise spiral pattern. Only then is the spiraling Qi projected out through the doctor's right hand into the patient's body.

Using the appropriate hand position, the doctor slowly spirals the energy in either a clockwise direction to guide the patient's energy into the affected area, or in a counterclockwise direction

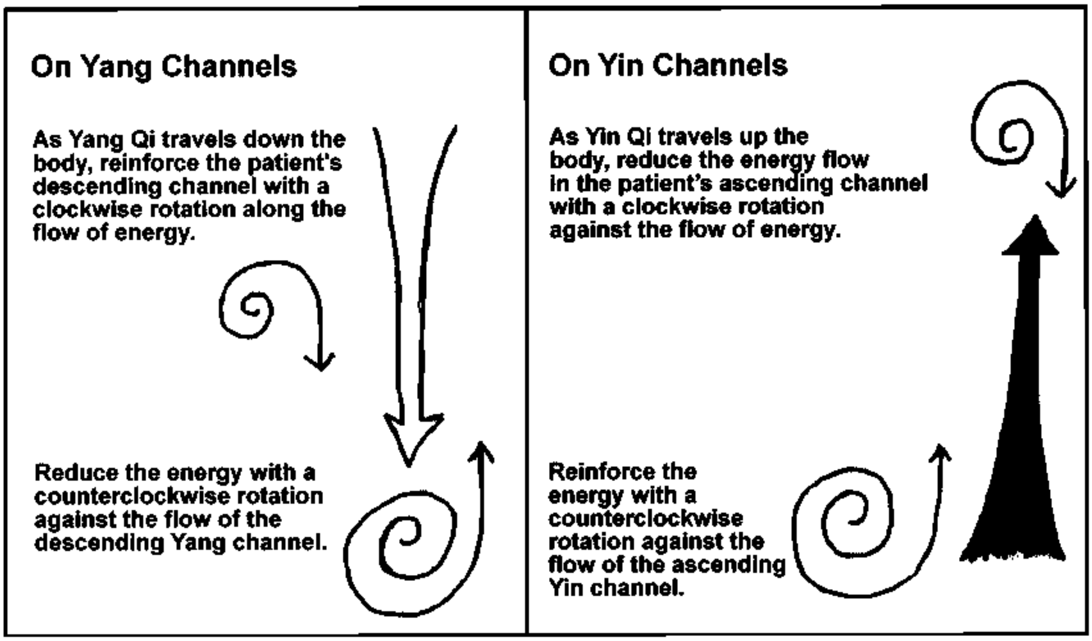

On Yang Channels

As Yang Qi travels down the body, reinforce the patient's descending channel with a clockwise rotation along the flow of energy.

Reduce the energy with a counterclockwise rotation against the flow of the descending Yang channel.

On Yin Channels

As Yin Qi travels up the body, reduce the energy flow in the patient's ascending channel with a clockwise rotation against the flow of energy.

Reinforce the energy with a counterclockwise rotation against the flow of the ascending Yin channel.

Figure 34.9. The reinforcing and reducing techniques of moving energy along the patient's channels is determined by the direction of the Spiraling Energy flow.

to remove the patient's Toxic Qi out of the diseased area.

- The counterclockwise rotation is considered Yin and is used to reduce Heat and to cool the patient. It is generally used to purge conditions of Excess (think of unscrewing the pathogens from the diseased area when purging).
- The clockwise rotation is considered Yang and is used to increase energy and warm the patient. Generally, it is used to tonify conditions of Deficiency (think of drilling and filling the Deficient area when tonifying).

The clockwise spiraling energy originating in the Lower Dantian is emitted and synchronized with the doctor's circling hand movements to reinforce the patient's Qi. To reduce the flow of the channel's Qi, the counterclockwise spiraling technique is used (Figure 34.9).

The Qigong doctor must cultivate these movements into a conditioned reflex before applying these treatment methods in a clinical environment.

SPIRALING ENERGY EXERCISE

- From a Wuji posture, place the left palm on the Lower Dantian. Slowly extend the right palm (keep elbow bent) in front of the chest, aligned with the center line at the Middle Dantian (CV-17), palm facing outward.
- Breathe naturally, while focusing on the Lower Dantian. With the mind's intention, begin to spiral the Qi in the Lower Dantian in a clockwise circle. Lead the energy up through the chest and out the extended right palm. Make sure that the spiraling of the Lower Dantian is synchronized with that of the right palm, making the inner Lao Gong Pc-8 points the center of concentration (Figure 34.10). Start off slowly, and gradually increase speed. The energy circle spirals outward with increasing or decreasing radius, depending upon the doctor's intent.

Repeat the exercise, only this time use counterclockwise spiraling.

ENERGETIC CUPPING

One popular clinical technique called "energetic cupping" is initiated by creating an energetic vortex which remains on the patient, pulling or drawing in energy, similar to that of an acupuncturist's cupping method. These energetic vortices can be used to drain or purge pathogens from a specific point or area on the patient's body. These energetic vortices can also be used to fill or tonify the patient's body. For example, when the patient is prone, the Qigong doctor can stimulate points LI-15, GV-4, and GV-14 (Figure 34.11). Then, using his or her intention, the doctor can pull with one hand while the other hand pushes, increasing the spiraling action already set in motion. This initiated energetic torque can sometimes cause the patient's body to turn, twist, bend, and sway. This modality of treatment simultaneously activates the patient's channel Qi, regulates the Yin and Yang organs, and balances the patient's Yin and Yang energies.

The doctor is advised to rely on intuition or divine guidance in choosing the right method, after careful consideration of the patient's condition (age, sex, personality, illness, etc.).

Note: If combined with the "shaking" technique, the circle/spiraling energy becomes especially effective for relieving pain.

GUIDING THE STATIONARY FLOW OF ENERGY

Stationary energy is emitted from a non-mobile position and can be used to stimulate the channel Qi, points, and Dantians, while restoring and replenishing the patient's Qi. After selecting the appropriate hand posture, the doctor positions his or her hand above the area being treated and emits energy from a fixed or stationary position. This is the most common energy projection technique, in which Qi flows like dense waves or is spaced like a chain of moving pearls flowing outward one after another (Figure 34.12). The patient often feels the sensation of vibration, tingling, or heat. It is not uncommon for the patient to develop spontaneous involuntary reactions such as full body spasms or shaking, leg jerking, etc.

Stationary energy is emitted from a station-

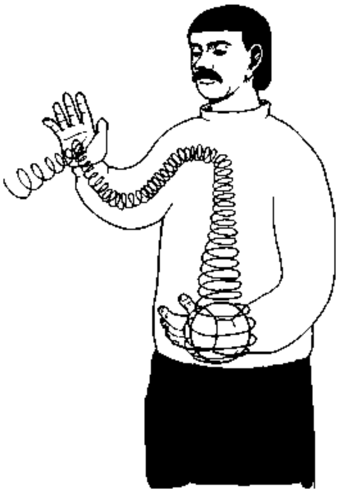

Figure 34.10. An example of a Qigong doctor practicing the Spiraling Energy exercise

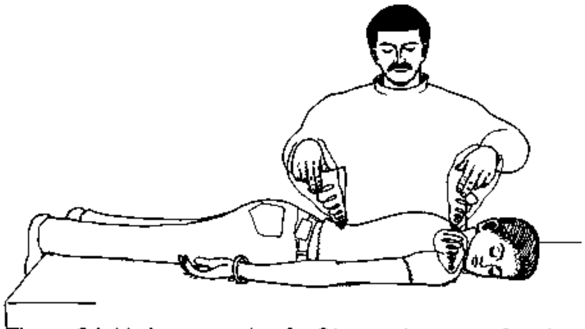

Figure 34.11. An example of a Qigong doctor performing multiple point stimulation with the Energetic Cupping Technique

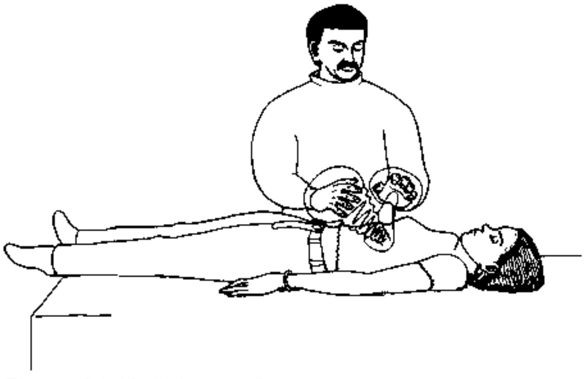

Figure 34.12. When a Qigong doctor extends energy from a stationary position, the Qi flows either like a dense wave, or is spaced like a chain of moving pearls.

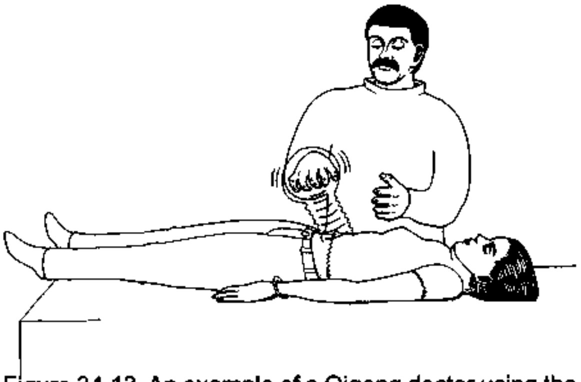

Figure 34.13. An example of a Qigong doctor using the Shaking Energy technique along with the Kneading Tiger Palm

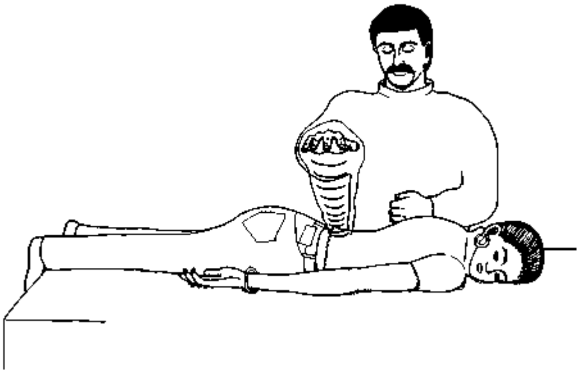

Figure 34.14. An example of a Qigong doctor using Vibrating Energy through the Extended Fan Palm technique

ary palm position. There are two main energetic pulsations utilized in the Stationary Palm techniques: Shaking Energy technique and Vibrating Energy technique.

SHAKING ENERGY TECHNIQUE

The Shaking Energy technique is initiated by the doctor first extending his or her energy over the surface area of the patient's body. After using the Tiger Kneading Palm technique to grasp the Turbid Qi, the doctor slowly begins to guide the patient's Qi out of the tissue area with a stationary manipulation. The doctor then shakes his or her hand from side to side, while continuing to pull the Qi from the area being treated (Figure 34.13).

To practice this particular skill, the Qigong doctor begins from a Wuji posture, using natural and slow breathing. The doctor imagines the waist as the axis and the abdomen as a pump and vibrates the Qi inside his or her Lower Dantian. The doctor then allows the energy to flow upward and out the palm, into the patient. It is important for the doctor to use the mind's intent to follow, guide, and direct this vibration of energy flow, instead of forcing the muscles and palm to shake.

VIBRATING ENERGY TECHNIQUE

Vibrational trembling is one way that memory is stored or processed in the physical body. The Qigong doctor's vibration penetrates the patient's field of "vibrational activity" (the body's natural vibrational rhythm), awakening this memory, allowing the patient to release trapped thoughts and emotions (Figure 34.14).

Qi naturally oscillates and vibrates in the doctor's Three Dantians, each Dantian vibrating at its own unique frequency pattern. The body's bone structure is built with natural formations similar to that of "tuning forks," positioned in both the arms and the legs. These tuning forks are connected together through the body's spinal column and are used to vibrate and increase the body's energetic resonance. Through a balanced interplay of stretching and releasing the body's tendon, ligament, and muscle fibers (in a rhythmic pattern), a natural vibration is produced. After some time of training (through naturally twisting and stretching the tendons), the Qigong doctor, using proper hand postures and intention, is able to direct and lead this vibration throughout his or her entire torso and extend it outwards through the hands.

Vibration is released through three forms of resonation:

1. Jing vibrational resonance is felt within the patient's external tissues and bones. It is related to the energetic release of Heat and the energy of the doctor's Lower Dantian.
2. Qi vibrational resonance is felt deep within the patient's internal organs. It is related to the energetic release of the patient's emotions and the energy of the doctor's Middle Dantian.
3. Shen vibrational resonance is felt within the patient's center core. It is related to the ener-

getic release of the patient's spirit and the energy of the doctor's Upper Dantian.

The release of energetic vibration is primarily used to activate the patient's channel Qi. Clinically, it is most effective when the frequency of the doctor's hand vibrations matches and synchronizes with the vibrational frequency of the patient.

While treating the patient, the doctor's palm usually gets hot and vibrates when coming in contact with Qi stagnation. If a patient's channel point becomes blocked, energy builds up causes Qi stagnation. Once the Qi stagnation is removed, the doctor's hands will cease to vibrate as the channels open and flow freely. When cleared, the patient feels fatigue as the toxins are released from the body and the energy seeks to balance itself.

To apply the Vibrating Technique, the doctor selects the proper hand posture and the area to be treated, then barely touches the patient's body. Any firm pressure should be avoided so as not to hinder the extending and directing of Qi. The doctor initiates the vibrational extension of Qi from the Lower Dantian, slowly turning clockwise or counterclockwise, while directing the vibration out through the hands. The doctor focuses on his or her own breath to adjust the frequency and amplitude of the vibration, the shape and nature of the emitted Qi (Yin or Yang, etc.), and the strength of the energy.

Some examples of hand postures and tissue regulation techniques for emitting vibrational energy include:

a. The Extended Fan Palm is used for activating and regulating Qi activities by:
    - Vibrating the Extended Fan Palm on the Baihui GV-20 point, to descend the patient's Yang Qi, tranquilize the patient's mind, regulate Qi in the Upper Burner, and induce sleep;
    - Vibrating the Zhong Wan CV-12 point to replenished the patient's Qi, to regulate the Stomach's energy and to induce sleep;
    - Vibrating the patient's Lower Dantian, to Lead the Qi back to its origin through supplementing its Qi;
    - Vibrating the Mingmen GV-4 and Da Zhui GV-14 points, to activate Qi in the Governing Vessel.

b. The Sword Finger Palm is used for directing and emitting concentrated vibrational energy into channel points, to stimulate the patient's energetic field.

c. The Invisible Needle Palm is used for directing and emitting the most condensed form of concentrated vibrational energy into the channel points, to stimulate the patient's energetic field.

VIBRATING PALM EXERCISE

The training for this kind of vibrational extension usually begins from a static Wuji posture, then gradually progresses into a dynamic state of energy projection.

- From a Wuji posture, press both palms together at the center of the chest forming a "Buddhist greeting" posture (see Chapter 15). Press the Baihui point upward and the Mingmen backward, while slightly drawing in the chest. Keep hips relaxed, knees bent, tongue on palate, and eyes closed.
- Breathe naturally while focusing the mind on the Lower Dantian. When the Heat and Qi begin to circulate, exhale while mentally guiding the Qi up the chest into both of the palms, by sending the energy through the three Yin channels of the hands.

When inhaling, mentally guide the Qi along the three Yang channels of the hands, back down the back into the Lower Dantian. Maintain natural breathing, and continue to focus attention on both the palms and the fingertips. The palms should heat up and the fingertips tingle, as they expand with Qi (Figure 34.15).

Lead the vibrating energy in the Lower Dantian up into the palms and fingertips, feeling the tissues dissolve inch by inch, cell by cell. The Qi is always centered on the inner Lao Gong Pc-8 point, gathering and collecting but not dispersing. Practice this exercise once or twice a day for 3-10 minutes.

Note: It is imperative that the doctor, when using the technique of "stationary energy pro-

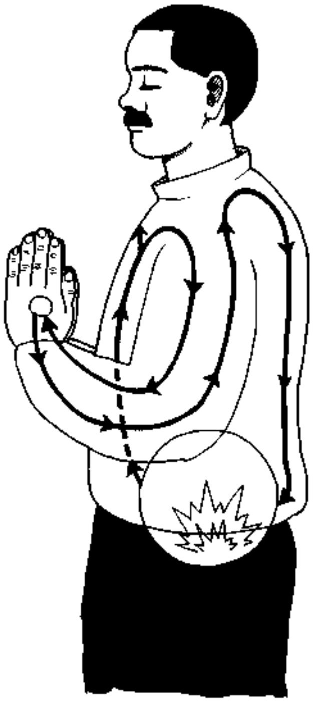

Figure 34.15. The Vibrating Palm exercise helps to increase the Qigong doctor's projected vibration.

jection," never hold his or her breath, or purposely make the hand vibrate by forcefully shaking the muscles. To do so causes Qi stagnation which results in pain and stiffness in the chest and sharp pain in the arms that feels similar to having a fractured bone or a laceration of the muscles. To ensure that there are no difficulties, the doctor should first master the Vibrating Qi method, allowing the vibrating energy to extend and project naturally.

GUIDING THE HOT AND COLD FLOW OF ENERGY

This method refers to the doctor's release of energy being emitted into the patient's body to either regulate Hot syndromes with Cold energy, or Cold syndromes with Hot energy. These two Yin and Yang energetic properties are used to develop the emission of both Hot and Cold energetic flows, through breathing, mental concentration and finger placement.

When Yin Qi is applied, the patient will feel "cold as iron," when Yang Qi is applied, the patient will feel "hot as fire." The mind plays an important role in regulating the doctor's and patient's Yin and Yang energy. When the Qigong doctor thinks of fire, the doctor's body will naturally become hot; conversely, when the doctor thinks of water, his or her body will naturally become cold.

HAND MANIPULATION OF YIN AND YANG ENERGY EXTENSION

The Yin (Cold) and Yang (Hot) energy extension utilizes specific palm and finger arrangements. The index and middle fingers together are assigned Yang properties, whereas the ring and little fingers together are assigned Yin properties. The doctor's left thumb connects with the selected two fingers, to determine the release of Yin or Yang energies. This energy is first gathered from the doctor's left hand, through the arm into the Lower Dantian, then emitted out through the right hand. The Qi that is emitted when the doctor's thumb, index and middle fingers connect, for example, releases the Heart's Yang Qi through the right hand. This technique is used to replenish the patient's Yang and regulate the patient's Heart energy.

GATHERING AND ISSUING YANG (HOT) ENERGY

Before emitting Hot energy into the patient, the doctor's Qi is first gathered into his or her left palm by imagining a small "fiery red sun" radiating heat, light, and fire. This small sun is imagined within the circle created between the thumb and first two fingers on the left hand. The hot energy of the sun is then absorbed through the doctor's left arm into the Lower Dantian. The Heat is then blended with the Lower Dantian's Qi and released out the doctor's right arm through either the tips of the index and middle fingers (the Yang Sword Fingers hand technique) for a deeper, concentrated type of penetration, or through the Extended Fan Palm hand technique for a dispersing type of penetration.

To extend Hot energy deeper into the patient, the right hand Yang Sword Finger posture is used. The doctor's index and middle fingers are extended and joined together, the ring and little fingers curl into the center of the palm. The thumb should rest on the nails of the ring and little finger, forming a circle.

The first or "index" finger indicates the Wood Element, i.e., the Liver's energy, and the middle finger indicates the Fire Element or the Heart's Qi. Together, both fingers activate the Qi of Fire and Wood, which catalyze the Yang energy needed for guiding Heat during point concentration.

Guiding Hot Yang Qi Exercise

This particular meditation is used to increase Heat and light in specific areas of the patient's body (usually for chronic, Deficient, or Cold syndromes). The following meditation, for example, can be used to Heat and tonify the Kidneys in chronic fatigue patients.

- Sit on the edge of a chair in front of a table, feet flat, eyes closed, with the left hand resting comfortably on the left thigh. The right palm is extended face down suspended over the table.
- Breathe naturally while concentrating the mind on the Lower Dantian. Imagine the energy in the Lower Dantian is a hot, burning sun expanding out and enveloping the entire body. Gather the energy back into the Lower Dantian, transforming it into heat and light. Draw the heat and light up into the chest, down the arms, and out the right palm. Concentrate and imagine the heat and light burning and extending into the table, penetrating deep into it. After awhile, switch palms. In China, this exercise is practiced while standing or sitting in the sunshine, facing south or east. Practice for 20-40 minutes (Figure 34.16).

GATHERING AND ISSUING YIN (COLD) ENERGY

Before extending Cold energy into the patient, the doctor's Qi is first gathered into his or her left palm by imagining a small, watery blue moon, radiating cool light. This small moon is imagined within the circle created between the thumb and last two fingers. The cool blue energy of the moon is then absorbed through the doctor's left arm, into the Lower Dantian. The cool blue energy is then blended with the Qi of the doctor's Lower Dantian and released out the doctor's right arm through either the tips of the ring and little fingers (the Yin Sword Fingers hand technique) for a deeper, concentrated type of penetration, or through the Extended Fan Palm hand technique for a dispersing type of penetration.

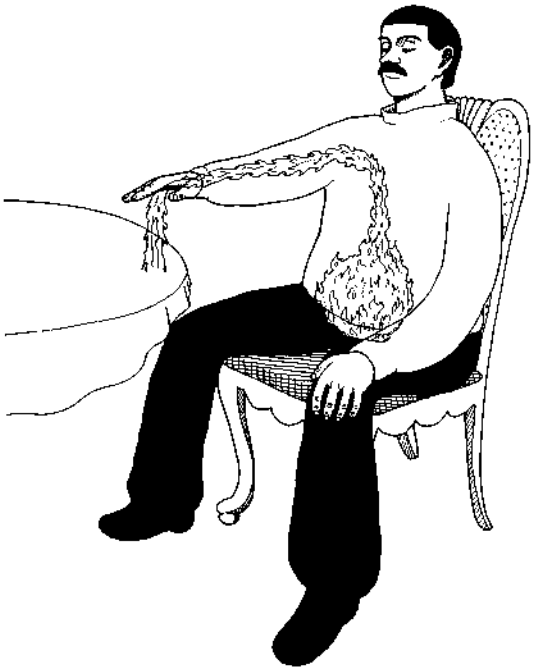

Figure 34.16. Guiding Hot Yang Qi Exercise

In extending cool blue energy into the patient, the right hand Yin Sword Finger posture is generally used. The doctor's ring and little fingers are extended and joined together, and the index and middle fingers curl into the center of the palm. The thumb should rest on the nails of the index and middle finger forming a circle.

The ring finger indicates the Metal Element or the Lungs' energy, and the little finger indicates the Water Element or the Kidneys' energy. Together, both fingers activate the Qi of Metal and Water, which is a catalyst for the Yin energy needed for guiding cool Yin Qi during point concentration.

Guiding Cold Yin Qi Exercise

This meditation is used to cool specific areas in the patient's body in acute, inflamed Excess or Heat syndromes. It can also be used, to cool down and dredge the Liver in cases of chronic depression.

- Sit on the edge of a chair in front of a table, feet flat, eyes closed, with the left hand resting comfortably on the left thigh. The right palm is extended face down suspended over a table.
- Breathe naturally while concentrating the mind on the Lower Dantian. Begin to gather energy from the heels and Bubbling Well (Kd-1) points, directing the energy through the chest (CV-17), out the arms, and into the right palm. Imagine the right palm turning as cold as ice. Focus the mind on the coldness of the right palm and imagine the cold energy penetrating deep into the table. In China, this exercise is practiced at night, facing north or west for 20-40 minutes (Figure 34.17).

Note: While practicing the Cold Yin Guiding exercise, never imagine that the whole body is cold, also do not direct and keep the cold energy into your body (only through it). Otherwise the Cold energy will affect the balance and energetic harmony of the Qi's flow. You can, however, use the image of cool blue water or vapor flowing through your body into the right hand (which is freezing cold), without causing harm to your body's energetic balance.

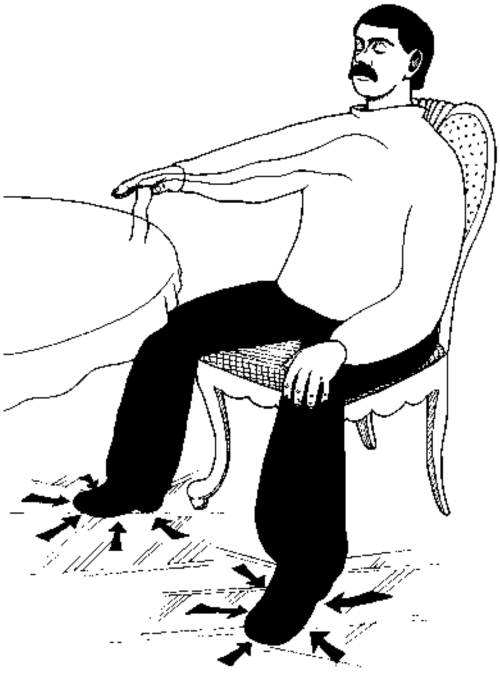

Figure 34.17. Guiding Cold Yin Qi Exercise

CLINICAL APPLICATION OF YIN AND YANG ENERGY EXTENSION

Traditionally, the Qigong doctor would emit Hot Yang Qi during the Yang period of the day (from midnight to high noon) to warm the patient's Cold areas. During the Yin periods of the day (from high noon to midnight) the doctor would apply the Cold Yin Qi to cool the Hot areas of the patient's body.

In China today, the clinical application of issuing Hot or Cold energy is no longer limited to the position of the sun and moon, but is applied according to the doctor's ability to regulate the Excess or Deficient condition of the patient's disease.

GUIDING THE ENERGETIC FLOW OF THE FIVE ELEMENTS

This particular Medical Qigong healing system is divided into two popular methods of energy extension: The Five Elemental Organ Energies Method and The Five Elemental Channel Energies Method. Both systems can be used for diagnosing the patient (using palm detection) or for treating Excess or Deficient conditions within the patient's body (using energy extension techniques). Each of these two powerful systems of treatment has its own unique advantage in application. Both systems combine healing sound resonances with energy light extension of various colors, and are used in accordance with the theory of the Five Elements.

THE SIX CHARACTER FORMULA

When utilizing either the Five Elemental Organ, or Five Elemental Channel system of energy projection, the Qigong doctor will guide the energy of the Five Elements in conjunction with the Six Character Formula. The Six Character Formula includes: The Five Elemental Energies, Time of Treatment, Direction of Energy, Position of the Thumb, Guiding Energy, and Emitting Energy and Sound.

1. The Five Elemental Energies of Wood, Fire, Earth, Metal, and Water relate to either the Five Elemental Channel system of energy projection and the Blood / Heat Cycles; or the relate to Five

Elemental Organ system and the energy of the Five Yin Organs. The decision regarding which of the two methods to choose is determined by the patient's condition.

2. The time of treatment is traditionally determined by the doctor in accordance with the patient's Blood/Heat Cycle, the time of day (sunrise, noon, sunset, and midnight) and the season (summer, winter, autumn, or spring).
3. The direction of energy emitted into the patient's body is determined by the doctor in accordance with his or her intention to purge, tonify or regulate the patient's tissues.
4. The position of the thumb is determined in accordance with the type and amount of energy the doctor wishes to absorb into his or her Lower Dantian.
5. Guiding (Pushing, Pulling and Leading) the energy is determined by the doctor in accordance with the Excess or Deficiency of the patient's Five Yin Organs: Liver (Wood), Heart (Fire), Spleen (Earth), Lungs (Metal), and Kidneys (Water).
6. The type of energy (Qi or Shen) and the choice of sound to be emitted are determined by the doctor's intention and the condition of the patient's organs and channels. Qi and sound are emitted like water, Shen is emitted like light.

EMITTING THE ENERGY OF THE FIVE ELEMENTAL ORGANS

The method of projecting the Five Elemental Organ Energies is utilized by blending a specific organ energy together with the Qi stored in the doctor's Lower Dantian. Both of these energies are combined and then issued out the doctor's right palm into the patient. The fusion of the doctor's Lower Dantian Qi with the selected organ Qi is very powerful. This synergized Qi is further combined with sound resonation and colored light to treat patients with extreme conditions of Deficiency.

The left thumb connects with either the tip of the index (Wood), middle (Fire), ring (Metal), little finger (Water) or upper pad of the palm (Neutral) to connect with the selected organ energy (Figure 34.18).

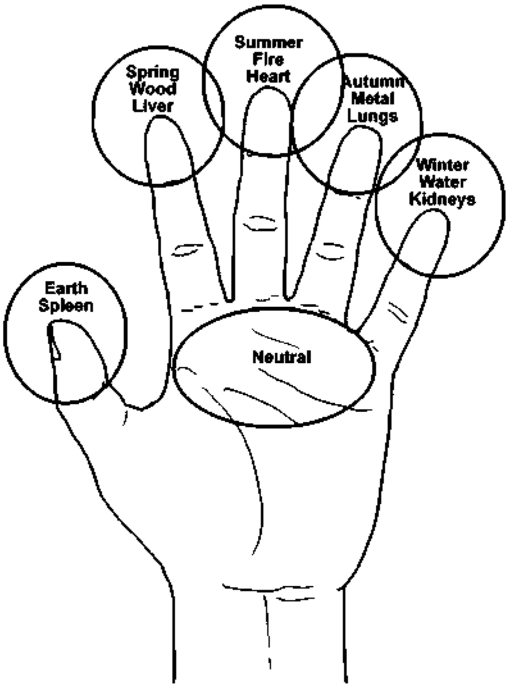

Figure 34.18. Left Palm - The Five Element Organ Energy Pattern: the left hand is used in order to absorb the organ Qi into the doctor's Lower Dantian. This arrangement corresponds to the Five Seasonal Pattern Cycle and the Five Elements.

Five Fingers	Five Elements	Five Viscera
Thumb	Earth	Spleen
Index	Wood	Liver
Middle	Fire	Heart
Ring	Metal	Lungs
Little	Water	Kidneys

Figure 34.19. Hand Chart of the Five Elements

After choosing the appropriate thumb and finger position on the left hand to tap into the selected organ's energy, the doctor descends this energy into the Lower Dantian. This energy is then blended together with the energy of the Lower Dantian and led out of the doctor's body, to be emitted into the patient using the right hand, along with the sound and specific organ color (Figure 34.19).

EMITTING THE ENERGY OF THE FIVE ELEMENTAL CHANNELS

The method of projecting the Five Elemental Channel Energies is utilized by blending a specific channel energy together with the Qi stored in the doctor's Lower Dantian. To tap into the selected channel's energy, the thumb is positioned on the corresponding finger, or palm segment (Figure 34.20). The selected channel's energy is then blended with the Qi stored in the doctor's Lower Dantian. These combined energies are then issued out the doctor's right palm into the patient. The blending of channel and Lower Dantian Qi is strong but not as powerful as the doctor's organ energy.

The resonant sound and light emitted from the combined channel and Lower Dantian energy is considered appropriate for patients with moderate Deficiency conditions. Healing sounds in combination with specific colors can be emitted by the Qigong doctor to stimulate the internal tissues. When healing sounds are directed into the patients tissues they have a tonifying action, as opposed to the purging affect which happens when the patient's practice healing tones by themselves (see Chapter 40).

The treatment is conducted according to the Five Elemental Creative Cycle. First the energy of the "mother" element is gathered into the doctor's Lower Dantian to be combined with the doctor's Lower Dantian energy. After the doctor collects and mixes the mother's elemental channel energy with the Lower Dantian's energy, the doctor switches the left thumb's position to the mother's corresponding "child" element position. This enables the mother's channel energy to support the strength of the child's channel energy, which is emitted into the patient's body (Figure 34.21).

Note: The doctor first connects to the mother element, then to the child element. Thus the mother's elemental channel energy, Lower Dantian energy and the child's elemental channel Qi are synergized and projected into the patient's corresponding "child" organ. The patient's organ, into which the Qi is emitted, has the same elemental attribute as the doctor's "child" transmitted energy.

1. To Emit Liver Energy and Guide Wood Qi proceed as follows:
   - First, from a Wuji posture, press the left thumb into the channel Water bands (9, 10). Inhale this Water energy into the Lower Dantian, mixing it together with your Lower Dantian energy.
   - Second, extend the right hand facing outward from the chest, facing the patient. Then exhale and press the thumb onto the Wood channel bands (1, 2) while uttering the sound "Shu." Guide the Wood energy out of the right palm into the patient's Liver organ or channel. In this particular technique, the doctor is using Water energy to support the extension of Wood energy.
2. To Emit Heart Energy and Guide Fire Qi proceed as follows:
   - First, from a Wuji posture, press the left thumb into the channel Wood bands (1, 2), while inhaling the Wood energy into the Lower Dantian, mixing it together with your Lower Dantian energy.
   - Second, extend the right hand facing outward in front of the chest, towards the patient. When exhaling, press the thumb onto the Fire channel bands (7, 8) with the thumb while uttering the Heart sound "Haa." Guide the Fire energy out the right palm into the treatment area. In this particular technique, the doctor is using Wood energy to support the extension of Fire energy.
3. To Emit Spleen Energy and Guide Earth Qi proceed as follows:
   - First, from a Wuji posture, press the left thumb into the channel Fire bands (7, 8) gathering this Fire energy into your Lower Dantian.
   - Second, extend the right hand in front of the

Colors		Element	Channel
1.	Green	Wood	Gall Bladder
2.	Green	Wood	Liver
3.	White	Metal	Lungs
4.	White	Metal	Large Intestine
5.	Yellow	Earth	Stomach
6.	Yellow	Earth	Spleen
7.	Red	Fire	Heart
8.	Red	Fire	Small Intestine
9.	Blue	Water	Bladder
10.	Blue	Water	Kidney
11.	Red	Fire	Pericardium
12.	Red	Fire	Triple Burner

Figure 34.20. The Twelve Channels are superimposed onto the Qigong doctor's left palm, arranged according to the progression of the Two Hour Qi and Blood/Heat Cycle.

This Blood/Heat Cycle also corresponds to the Four Principle Time Periods and can be divided into quarters and arranged according to the Midnight, Sunrise, Noon, and Sunset patterns of the Sun.

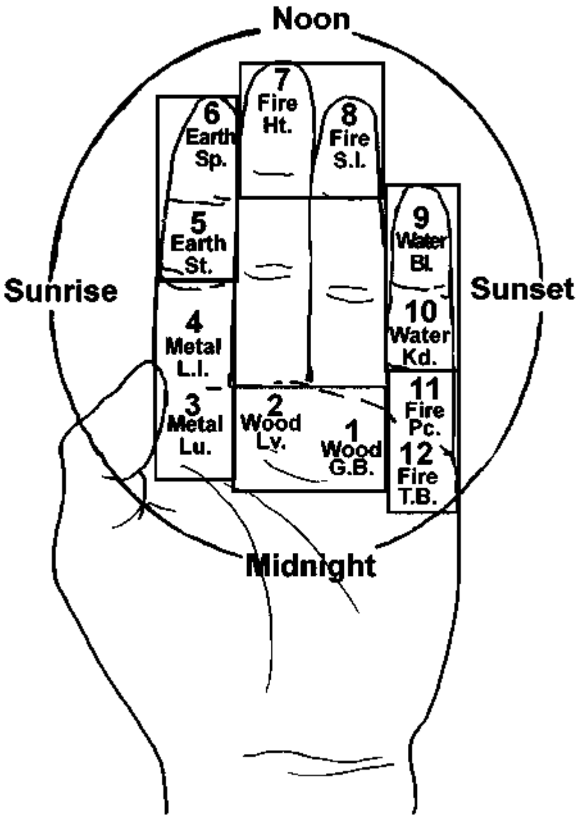

Figure 34.21. Left Palm - The Five Elemental Channel Pattern:

The left palm is used in gathering and absorbing the channel Qi of the doctor's Five Yin Organs. As the Qigong doctor's thumb touches the specific areas on the fingers, the energy is gathered into the doctor's Lower Dantian and released out from the right palm.

chest, facing outward towards the patient. When exhaling, press the thumb into the Earth's channel bands (5, 6) while uttering the Spleen sound "Who." Guide the Earth energy out of the right palm into the treatment area. In this particular technique the doctor is using Fire energy to support the extension of Earth energy.

4. To Emit Lung Energy and Guide Metal Qi proceed as follows:

- First, from a Wuji posture, connect the left thumb onto the tip of the left index finger (i.e., the Earth channel bands #5, 6), thus forming a circle. Imagine this circle full of golden light. When inhaling, draw this golden Earth energy into the Lower Dantian, while also inhaling Heavenly Qi into the Yellow Court area (CV-12) below the solar plexus. Mix all three of these energies together.
- Second, extend the right hand facing outward in front of the chest, towards the patient. When exhaling, press the thumb the Metal channel bands (3, 4) while uttering the Lungs' sound "Sss." Guide the Metal energy out of the body by way of the right palm into the treatment area. In this particular technique, the doctor is using

imagination to draw the Earth's golden light energy into his or her Lower Dantian to support the extension of Metal energy.

5. To Emit Kidney Energy and Guide Water Qi proceed as follows:
   - First, from a Wuji posture, press the left thumb into the Metal channel bands (3, 4). Inhale the Metal energy into the Kidneys, mixing it together in the Lower Dantian.
   - Second, extend the right hand facing outward in front of the chest. As you exhale, press the Water channel bands (9, 10) while uttering the "Chree" sound (in a very low pitch). Guide the Water energy out of the body by way of the right palm. In this particular technique, the doctor is using Metal energy to support the extension of Water Qi.

Once skilled in these five exercises, the Qigong doctor should continue to practice them in combination with the Linear, Circular, or Spiral Energy Guiding Methods, creating an individualized, structured technique.

COMBINING THE ENERGY OF THE FIVE ELEMENTAL TONGUE POSITIONS

The doctor can also use the Five Elemental Tongue Positions to increase the power and energetic potential of the doctor's internal organ. Each tongue position connects with the organ's Orb (the complete organ system), thereby increasing the power of the doctor's Qi projection. These tongue positions are described as follows.

1. To connect with Earth (Spleen) energy—the tongue is placed on the bottom of the jaw at the base of the lower palate, beneath the teeth and gum line.
2. To connect with Metal (Lungs) energy—the tongue is placed between the maxilla and mandible bones, suspended, between the teeth.
3. To connect with Fire (Heart) energy—the tongue is placed on the hard palate, at the front of the upper palate, behind the teeth and gum line.
4. To connect with Wood (Liver) energy—the tongue is placed on the hard palate, at the middle of the upper palate at the center of the roof of the mouth.
5. To connect with Water (Kidneys) energy—the tongue is placed on the soft palate, at the back of the upper palate.

COMBINING THE ENERGY OF THE FIVE ELEMENTAL COLORS

The doctor can also include the Five Elemental Colors to increase the power of his or her internal Qi projections (see Chapter 4). The Five Elemental Colors are described as follows.

- Earth (Spleen)—The organ color bright yellow, the channel color is soft yellow.
- Metal (Lungs)—The organ color bright white, the channel color is soft white.
- Fire (Heart)—The organ color bright red, the channel color is soft red.
- Wood (Liver)—The organ color bright green, the channel color is soft green.
- Water (Kidneys)—The organ color bright blue, the channel color is soft blue.

SUMMARY OF QI EMITTING METHODS

When the Qigong doctor emits Qi into the patient, it effects the quality of light, sound, heat, vibration, and electromagnetic energy stored within the patient's tissues and cells. There are many variations of Qi emission hand postures and techniques used by Qigong doctors in China. Each differs according to the various schools, colleges, and styles of Medical Qigong training. These changeable and diverse hand postures, as well as the various Qi extension, guidance, and regulation techniques can all be combined and grouped together to assist the Qigong doctor in understanding the clinical treatment potentials. The following is an outline of the aforementioned Hand Postures and Extension, Guidance, and Regulation techniques (described in Chapter 33 and 34), along with their applications.

HAND POSTURES USED FOR EMITTING QI

1. The Extended Fan Palm Hand Posture energy emission has the widest beam of Qi extension.
2. The Sword Finger Hand Posture energy emission has the width and beam reduced into a more condensed form of Qi extension.
3. The One Finger Skill Hand Posture energy emission is further reduced into an even more condensed form of Qi extension.

4. The Invisible Needle Hand Posture energy emission is condensed even further into a finely honed needle of light, similar to that of a laser beam.

HAND TECHNIQUES USED FOR BREAKING UP ENERGETIC STAGNATIONS

1. The Thunder Fingers Hand Technique is the strongest, most aggressive hand technique, and is used to break into the energetic boundary of the Qi stagnation. It is used as a preparation technique before purging the stagnation.
2. The Shaking and Trembling Hand Technique is the second most aggressive hand technique, and is used to purge stagnation and toxic Qi from the patient's body.
3. The Vibrating Hand Technique is a powerful, yet subtle hand technique, which can be used to expand, contract, vibrate, and loosen the patient's Stagnant Qi.
4. The Kneading Tiger Hand Technique is used to separate the Toxic Qi, and to dredge and purge it out of the patient's body.

ENERGY EXTENSION AND GUIDANCE TECHNIQUES FOR MOVING QI

1. The Pushing Technique is used to push and move the energy within the body's tissues, channels, and energetic fields.
2. The Pulling Technique is used to pull and move the energy within the body's tissues, channels, and energetic fields.
3. The Leading Technique is used to lead, guide, and move the energy within the body's tissues, channels, and energetic fields.
4. The Circling Technique is used to lead, guide, and move the energy in a circling pattern within the body's tissues, channels, and energetic fields.
5. The Spiraling Technique is used to spiral or drill energy into or out of the body's tissues, channels, and energetic fields.
6. The Energetic Cupping Technique is used to lead, guide, and move the energy within the body's tissues, channels, and energetic fields.

HAND POSTURES USED FOR PURGING, DREDGING, AND REMOVING TOXIC QI

1. The Bellows Palm Hand Posture is used like a vacuum to remove toxic Qi
2. The Dragons Mouth Palm Hand Posture is used to grasp, pull, and lead toxic Qi from the body.
3. The Sword Fingers Hand Posture is used to circle-wrap toxic Qi (like a fork wrapping spaghetti) and remove it from the body.
4. The Kneading Tiger Hand Posture is used to grasp, seize, and pull toxic Qi from the body.
5. The Shaking and Trembling Hand Posture is used in order to grasp, seize, and pull toxic Qi from the body.
6. The Thunder Fingers Hand Posture is used to shake and release toxic Qi from the doctor's body.
7. The Extended Fan Palm Hand Posture is used to dredge, pull, and purge toxic Qi from the body.

TYPES OF ENERGY RELEASED THROUGH QI EMISSION

1. Hot Qi Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.
2. Cold Qi Emission is used to cool down or sedate a specific area, organ, channel, or energetic field.
3. Color Qi Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.
4. Sound Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.
5. Five Elemental Organ Qi Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.
6. Five Elemental Channel Qi Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.

CHAPTER 35
THE INVISIBLE NEEDLE THEORY AND APPLICATION

INTRODUCTION

In China today, the use of Invisible Needles in Medical Qigong therapy is usually reserved for patients who are hypersensitive to energy flow. As the patient's sensitivity increases, the treatments are modified in order to complement their perceptive abilities.

To perform the Invisible Needle technique, the Qigong doctor must first connect with the divine energy. This is performed by drawing energy from the Heavens through the Baihui GV-20 point at the top of the head and extending it out through the hands, enveloping the patient. Some Qigong doctors imagine that needles of light, from the Heavens, descend through their Baihui point. As this light fills their body, they become empowered with a never-ending reserve of Invisible Needles to treat the patient. In order to use these needles, the doctors rotate their wrist clockwise and the Invisible Needle flows out the Pc-8 Point at the center of their palm (Figure 35.1).

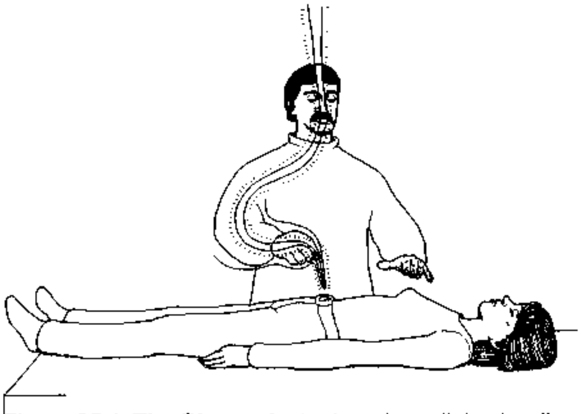

Figure 35.1. The Qigong doctor imagines divine healing light descending from the Heavens, filling the doctor's body with the energy to create the Invisible Needle.

Other Qigong doctors use a single needle image. The doctor imagines needles of light coming out of the Heavens, spiraling around the head like a golden halo. This golden halo is positioned above the head, above the Heavenly Transpersonal Point (located about a foot above the head). To use these needles, the doctors reach above their head and pluck the needles from the Heavens. These needles are then inserted into the patient's body (Figure 35.2). When using either visualization, the important factor is to focus on how deep the needles should penetrate. When treating an area deep in the major organs, a clockwise rotation reinforces and tonifies, while a counterclockwise rotation sedates the organ area.

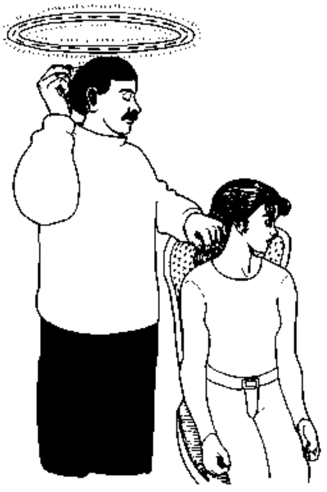

Figure 35.2. Another variation is for the Qigong doctor to imagine divine healing light descending from the Heavens and encircling the doctor's head like a golden halo. Contained within this halo are the Invisible Needles.

PREPARATION

Before inserting an Invisible Needle into the patient, it is important for the doctor to dredge the diseased area first. This is similar to swabbing an area with alcohol (or disinfectant) before inserting an acupuncture needle. As one hand extends the Invisible Needle, the other is used to reinforce its energy. The doctor works the Invisible Needle into the patient's tissues until he or she feels it inside the patient's body. Sometimes the patient may feel a cold sensation while being treated by the doctor. This is a normal reaction to the Invisible Needle insertion. The longer that the doctor keeps the Invisible Needle inserted inside the patient's body, the better the healing effect.

ANGLE OF INSERTION FOR TONIFICATION OR SEDATION

The skill of using Invisible Needles encompasses a complete system of energetics that utilizes both reinforcing (tonifying) and reducing (sedating) techniques. The polarity of the patient's channel (Yin or Yang) determines the angle at which the Invisible Needle is inserted.

- When performing reducing or sedating techniques, it is important to insert the Invisible Needles perpendicularly into the patient's points or channels, forming an energetic dam to slow the flow of Qi (Figure 35.3). The doctor removes the needle slowly, and leaves the point open (this allows the point to continue to release Qi naturally).
- When performing any reinforcing or tonifying techniques, insert the Invisible Needle at an angle, pointed in the direction of the channel's flow (provided that the channel is flowing towards the direction of the intended organ). Then, when retracting the Invisible Needle, remove it quickly and seal the point. This quick retraction method is done to prevent the Qi from escaping the area once it is stimulated, and is followed by pressing the needle hole to seal the point (Figure 35.4).

One favorite tonification technique used by certain Qigong doctors in Beijing, China is to leave the Invisible Needles inside the patient, while regulating the Conception and Governing Vessels. Once they have strengthened the patient's Microcosmic Orbit (Fire Cycle), they retract the Invisible Needles and disperse them one by one into the Earth. The insertion and manipulation of the Invisible Needles is performed with the Invisible Needle Palm technique.

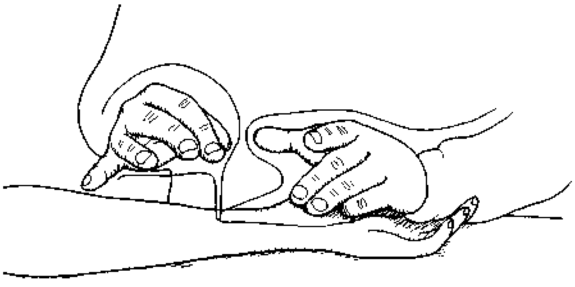

Figure 35.3. The doctor's right hand extends the Invisible Needle perpendicularly into the patient's point or channel, to reduce or sedate the energy flow. The doctor's left hand is used to reinforce the energetic purging of the patient's toxins.

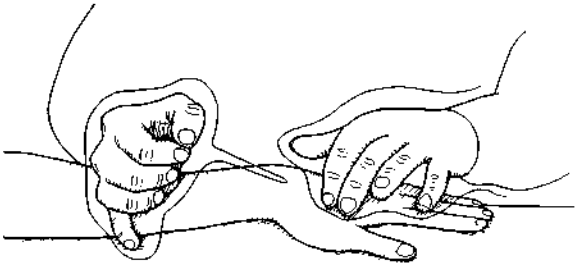

Figure 35.4. Insert the Invisible Needle at a 45 degree angle with the flow of the patient's channel to reinforce or tonify, provided the channel's flow is towards the organ.

In order to prevent the Qi from escaping, when performing any reinforcing or tonifying techniques, the Qigong doctor presses the patient's channel point after retracting the Invisible Needle.

Another favorite technique used for tonification is for the Qigong doctor to first insert the Invisible Needle deep into the patient's Kd-1 point at the bottom of the feet. Once the needle is inserted, the doctor extends his or her intention, causing the needle to grow and extend up the patient's legs and into the Kidneys. Next, the doctor emits energy up the patient's legs to further tonify the Kidneys and Mingmen area.

The Invisible Needles can also be transformed into the specific colors of the Five Elements to enhance tonification. A blue needle, for example, can be used to intensify Kidney tonification. Each needle can additionally be spiraled in a clockwise or counterclockwise direction to generate an energetic vortex used for either tonification or purgation (Figure 35.5).

THE HAND POSTURES OF THE INVISIBLE NEEDLE

The following is a description of the various hand postures used when treating patients with the Invisible Needle technique.

THE DOCTOR'S RIGHT HAND

The doctor's right hand is generally used for controlling and projecting the energy of the Invisible Needle into the patient. It is not only responsible for the formation but also the insertion of the Invisible Needle. The action and movement of the right palm projecting and stimulating Qi is sometimes referred to as the Chicken Pecking technique. In this hand posture, the thumb connects with the index finger or index and middle finger to form the shape of a bird's beak (Figure 35.6). While these fingers are joined together, the other fingers on the hand relax in an extended, straight position. The doctor's intention is focused on the image of an Invisible Needle which is projected into the patient's body.

The pricking, insertion, and manipulation of the Invisible Needle is only effective when the Qigong doctor connects and roots his or her intention deep within the patient's Jing, Qi, and Shen. This intention connects the doctor with the patient's Qi and Blood, and their animated Spirit.

THE DOCTOR'S LEFT HAND

The doctor's left hand is used to either manipulate the patient's tissues to enhance the Qigong treatment, or to collect Qi into the doctor's Lower Dantian to be projected into the patient. This technique is similar in method to collecting the Yin (Cold) and Yang (Hot) energy before emitting it into the patient. Once the thumb connects with any of the fingers, the released energy bears the nature of the associated element and its Yin and Yang properties (Figure 35.7). The Qi that is collected when the thumb and middle finger connect, for example, is the Yang Qi within the Heart.

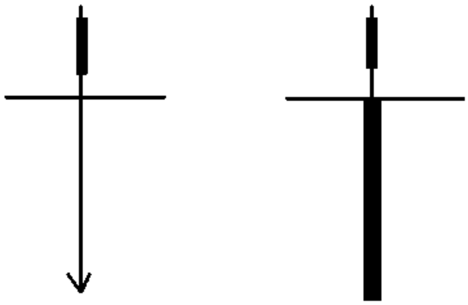

The needle can be made to grow and extend deep inside the patient's tissues or channels.

The needle can be made to grow fat inside the patient's tissues or channels to increase Qi projection.

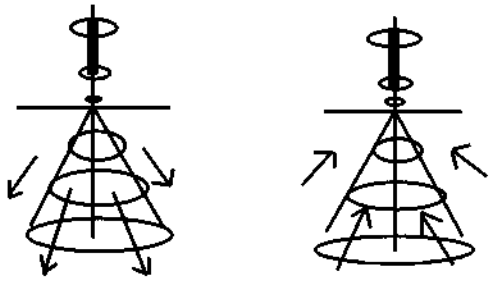

Clockwise spiraling the needle is used to create an energetic vortex for tonification.

Counterclockwise spiraling the needle is used to create an energetic vortex for purgation.

Figure 35.5. The Invisible Needle Techniques

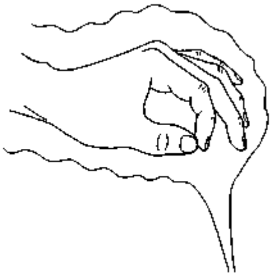

Figure 35.6. The Invisible Needle Palm

This Qi replenishes the Yang Qi and regulates the energy of the Heart (Figure 35.8). If the doctor's hand is positioned over the Pericardium Channel and energy is emitted into the patient, the patient's Shen will become peaceful. When treating with the Five Elemental organ energy, the index and middle finger stand for Fire and Wood; this combination is considered Yang. The ring and little finger stand for Water and Metal; this combination gathers and creates Yin energy.

TIME AND DURATION OF TREATMENT

In general, Medical Qigong Invisible Needle Technique and Point Therapy is applied once a day. A routine therapeutic course includes six to eighteen treatments. Patients with mild diseases may continue the therapy from six to twenty-four treatments. For chronic patients, however, treatment may last as long as one to three months depending on the patient's condition (i.e., for paraplegic patients the treatment is generally from three to six months).

As far as the time sequence of when to treat the patient, or when the patient is to treat themselves, the chronometric rate is based on the high-tide sequence of the patient's channels (i.e., 11 am to 1 pm is the Heart time, which is the best time period to treat patients with Heart disease).

REMOVING THE INVISIBLE NEEDLES AND ENDING THE TREATMENT

During tonifying treatment, the Qigong doctor removes the Invisible Needle quickly from its point of origin while the patient is inhaling. This allows the patient to keep the Qi and Shen within the body's channels and tissues. After extracting the needle, the doctor presses down on the acupuncture point with emitted Qi and lightly rubs it in a clockwise direction until the open point has closed.

During a sedating treatment, the Qigong doctor has the patient exhale to expel the Evil Qi and Heat from the patient's tissues. After extracting the needle slowly, the doctor allows the acupuncture point to naturally drain and eventually close by itself. This aids in draining the Evil Qi and Excess Heat from the patient's body.

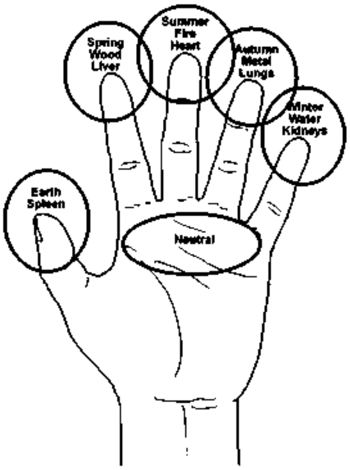

Figure 35.7. When treating patient's with the Invisible Needle technique for tonification, the Qigong doctor emits the required amount of energy into the patient, according to the specific Deficient organ and channel's condition.

The thumb, being the Earth Element and neutral, allows the Qigong doctor the ability to connect and combine the energy of the Five Yin Organs into his or her Lower Dantian, before extending the Qi into the patient.

Five Fingers	Five Elements	Five Viscera
Thumb	Earth	Spleen
Index	Wood	Liver
Middle	Fire	Heart
Ring	Metal	Lungs
Little	Water	Kidneys

Figure 35.8. Hand Chart of the Five Elements

CHAPTER 36
ENERGETIC POINT THERAPY

The importance of Energetic Point therapy as a Medical Qigong modality can best be understood when exploring the amount of information stored within the body's tissues. According to theoretical physicist Michio Kaku, the body's DNA contains over one hundred trillion times the information stored in current computers. There is an even greater amount of information stored within the three subtle energy fields outside of the human body. Most of this information is not readily translatable or decipherable, and is hidden from the conscious mind.

By stimulating certain energy points, for a specific length of time, an energetic trigger mechanism, or response reflex is created. This energetic reaction releases memories and stored information unique to each patient. To facilitate the unlocking of this stored information, the Qigong doctor must maintain a deep state of Alpha consciousness when stimulating the patient's energy points.

Energetic Point therapy involves two distinct modalities of treatment: Channel Point treatments and Channel Point meditations. Either can be used in conjunction with the Invisible Needle therapy. In these particular modalities of treatment, the Qigong doctor can either emit Qi into specific channel points on the patient's body to enhance the Medical Qigong treatment, or have the patient meditate on a specific point while the doctor administers the therapy. Both approaches can be used by the Qigong doctor to gather energy, move Qi, or release trapped energy from the patient's tissues.

CHANNEL POINT TREATMENTS AND PRESCRIPTIONS

The following treatments are generally used in the clinic by the Qigong doctor for Invisible Needle therapy, but are not necessarily limited to that modality of treatment. These Energetic Point treatments are arranged and categorized according to specific diseases. The Energetic Points are located on the patient's front and back as depicted in the anatomical graphs (Figures 36.1 and 36.2). These points are clinically used for the following treatment goals:

POINT THERAPY TO TONIFY YIN ENERGY

1. To Tonify Yin energy and to reduce a Deficient type of Fire, extend Qi into the patient's Kd-3 point.
2. To nourish the Yin energy and to reduce Fire, extend Qi into the patient's Kd-3 and Lu-10 points.
3. To treat a Deficiency of Yin of the Liver and Kidneys and to reduce the causes of Deficient Fire, extend Qi into the patient's Sp-6 and Lv-3 points.
4. To treat a Deficiency of Yin causing a dryness of the Lungs, extend Qi into the patient's Lu-1 and Bl-13 points on the front and back of the body. Regulate the respiratory tract, as well as the patient's Lu-7 and Kd-6 points, thereby tonifying the body's Yin energy and activating the descending function of the Lungs.

POINT THERAPY TO TONIFYING OR DISPERSE YANG ENERGY

1. To strengthen the Yang of the patient's Kidneys, extend energy into the patient's GV-4, Bl-23, and Kd-3 points.
2. To warm, and tonify the Yang of the patient's Spleen and Kidneys, extend energy into the patient's Bl-20 and -23 points. This treatment is ideal for eliminating Damp fluid caused from Deficient types of edema.
3. To fortify the Yang and avert a state of Yang collapse, tonify the patient's Yuan Qi by extending energy into the patient's Lower Dantian, focusing specifically on the patient's CV-4, -6, and -8 points.

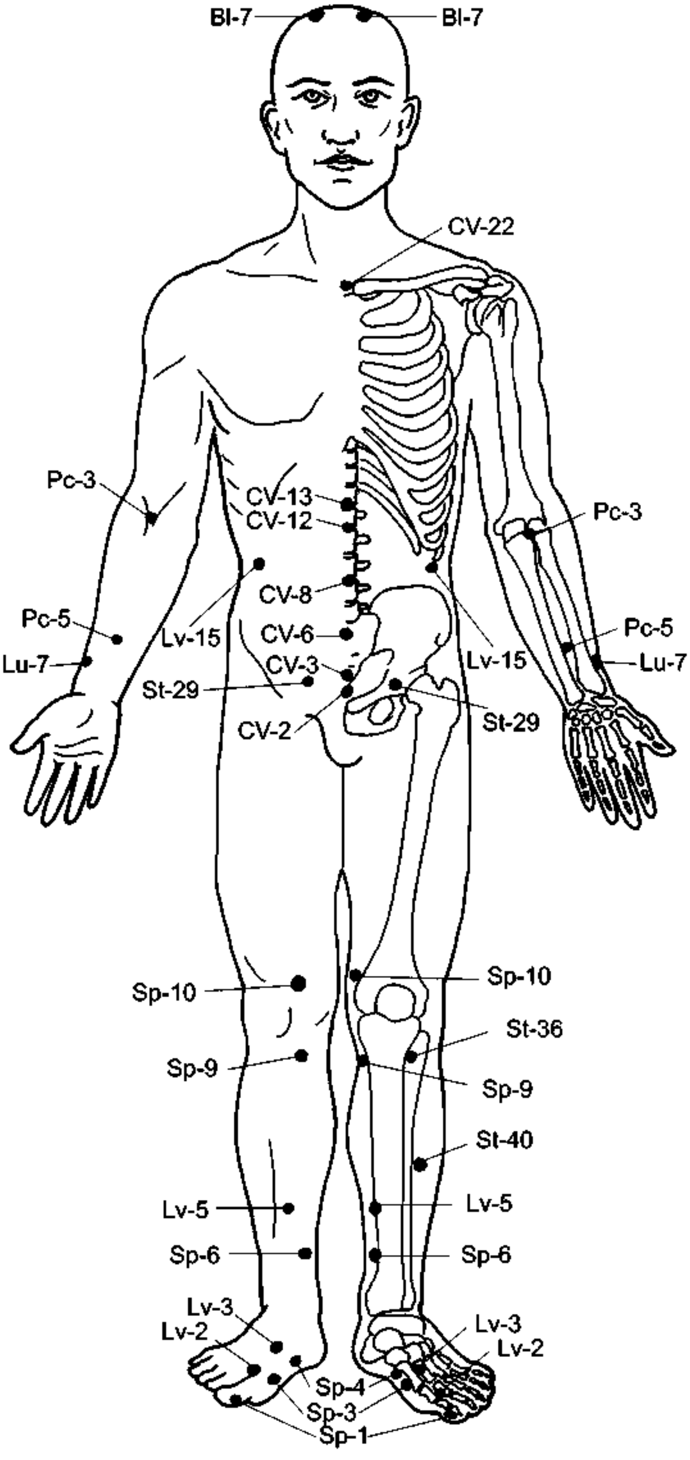
Bl-7
Bl-7
CV-22
Pc-3
CV-13
CV-12
Pc-3
CV-8
Pc-5
Lv-15
CV-6
Pc-5
Lu-7
CV-3
Lv-15
Lu-7
St-29
St-29
CV-2
Sp-10
Sp-10
St-36
Sp-9
Sp-9
St-40
Lv-5
Lv-5
Sp-6
Sp-6
Lv-3
Lv-3
Lv-2
Lv-2
Sp-4
Sp-3
Sp-1

Figure 36.1. Chart of Specific Channel Points

4. To treat a condition of Deficiency of Yang of the Spleen, strengthen the patient's Stomach and Spleen to eliminate Damp and Phlegm. This is done by extending energy and tonifying the patient's St-36, CV-12, and Bl-20 points. Then activate and warm the Qi of the patient's Lungs by extending Qi into Bl-13, and -43 points.
5. To strengthen the Yang of the Spleen, promote the Spleen's transporting function by extending energy into the patient's CV-12, Bl-20, Lv-13, and Sp-3 points.
6. To raise the patient's Yang Qi or the Yang Qi flowing within the patient's Governing Vessel, extend energy into the GV-20 point.
7. To reduce hyperactive Yang of the patient's Liver, first dredge and disperse the patient's GB-20, Bl-18, and Lv-2 points to pacify the Yang of the Liver; then tonify by extending energy into the patient's Bl-23 and Kd-3 points to strengthen the Kidneys.
8. To reduce Excess Yang within the patient's Governing Vessel which has caused a hyperactivity of Heat, dredge and sedate the patient's GB-2O point. Then purge the Heat from the patient's GV-14 and LI-4 points.

POINT THERAPY TO TONIFY OR DISPERSE QI

1. To strengthen the patient's Yuan Qi, extend Qi into the patient's Lower Dantian, focusing on CV-4 and -6 points.
2. To fortify the patient's Qi and reestablish their Yang, extend energy into the patient's GV-20, CV-6, and St-36 points.
3. To subdue the patient's ascending Qi, extend Qi into the patient's CV-22 and Bl-17 points.
4. To disperse the patient's stagnant Qi and Phlegm, extend Qi into the patient's Bl-13 point.
5. To tonify the Kidneys' Qi and Jing and to reduce any Kidney Deficiency, extend Qi into the patient's GV-4, Bl-52, and Kd-3 points.
6. To stimulate the circulation of Qi and Blood, extend Qi into the patient's LI-4 and Lv-3 points.

POINT THERAPY TO TONIFY AND NOURISH THE BLOOD

1. To activate the patient's Blood circulation, extend energy into the patient's Sp-10 point.
2. To stimulate the function of conducting Blood, extend energy into the patient's Sp-1 point.
3. To control the production of Blood, extend energy into the patient's Bl-20 point to control the Blood, as well as Bl-15 to increase the production of Blood.
4. To activate Blood circulation and remove stasis, extend energy into the patient's Sp-6 and Bl-17.
5. To strengthen the source of Blood formation, extend energy into the patient's St-36 and Sp-6 points.
6. To remove Blood stasis of the uterus, extend energy into the patient's Bl-32 and St-29 points.

POINT THERAPY TO ELIMINATE COLD AND DAMP

1. To warm the Spleen and Stomach to eliminate Cold, extend energy into the patient's Lower Dantian and CV-12 and -13 area.
2. To disperse Cold, relieve pain, pacify the Stomach and strengthen the Spleen, extend energy into the patient's Sp-4 and Bl-20 points.
3. To strengthen the Spleen to disperse Dampness, extend energy into the patient's Sp-6 and St-36 points.
4. To eliminate edema above the waist, extend energy into the patient's Bl-20 and Lv-13 points.
5. To eliminate edema below the waist, extend energy into the patient's Bl-28, LI-6, and Sp-9 points.

POINT THERAPY TO ELIMINATE HEAT

1. To eliminate Heat from the patient's body, extend energy into the patient's LI-11 points, or dredge and disperse energy from the patient's Shi Xuan points (at the tips of the fingers).
2. To move Heat downwards from the patient's body, dredge and disperse energy from the patient's Kd-1 points.
3. To disperse External Heat, dredge and disperse energy from the patient's SI-3 points.
4. To eliminate Internal Heat, dredge and disperse energy from the patient's Pc-5 points.
5. To eliminate Damp Heat, extend energy into the patient's Bl-20 and Sp-9 points.

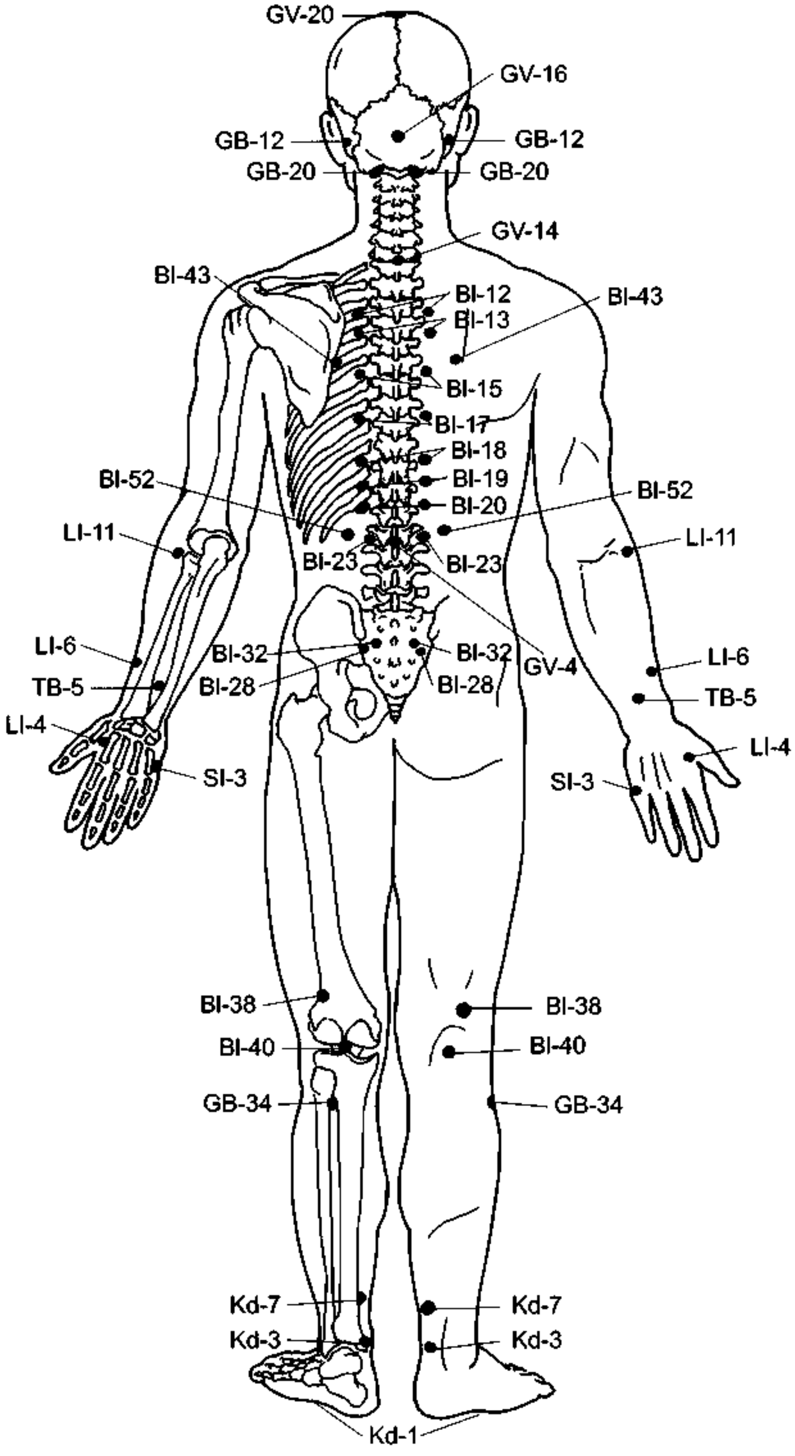
GV-20
GV-16
GB-12
GB-12
GB-20
GB-20
GV-14
Bl-43
Bl-12
Bl-13
Bl-43
Bl-15
Bl-17
Bl-18
Bl-19
Bl-52
Bl-20
Bl-52
LI-11
LI-11
Bl-23
Bl-23
LI-6
Bl-32
Bl-32
GV-4
LI-6
TB-5
Bl-28
Bl-28
TB-5
LI-4
LI-4
SI-3
SI-3
Bl-38
Bl-38
Bl-40
Bl-40
GB-34
GB-34
Kd-7
Kd-7
Kd-3
Kd-3
Kd-1

Figure 36.2. Chart of Specific Channel Points

6. To eliminate Damp Heat from the Blood, dredge and disperse energy from the patient's Bl-40 points, then extend energy into the Pc-3 points.
7. To reduce Fire of the Liver, dredge and disperse energy from the patient's CV-3 and Lv-5 points.
8. To reduce Liver and Gall Bladder Upward Fire disturbance, dredge and disperse energy from the patient's GB-12, as well as Bl-19 and -18 areas.

POINT THERAPY TO ELIMINATE WIND

1. To eliminate Wind Cold, dredge and disperse energy from the patient's GV-16, Bl-12, GB-20, Lu-7, LI-4, and Kd-7 points.
2. To eliminate Wind Heat, dredge and disperse energy from the patient's GV-14, LI-4, TB-5, and GB-20 points.
3. To eliminate Wind from the upper portion of the patient's body, dredge and disperse energy from the patients Bl-7, GV-20, and -16 points.
4. To eliminate pathogenic Wind, extend energy into the patient's Lu-7 points to activate the natural dispersing function of the Lungs.
5. To eliminate Wind and reduce Fire, dredge and disperse energy from the patient's GB-20 points.
6. To calm the Wind of the Liver, dredge and disperse energy from the patient's Lv-3 points.

POINT THERAPY TO ELIMINATE PHLEGM

1. To eliminate Damp Phlegm in the Interior, first strengthen the function of the patient's Spleen and Stomach to eliminate the Dampness, by extending energy into the patient's Bl-20 and CV-12 points, then eliminate the Phlegm by dredging both of the St-40 points.
2. To eliminate Phlegm Fire in the patient's Stomach, dredge the patient's GB-34 and St-40 points.

CONTRAINDICATIONS

Never apply Energy Point therapy to patients who are in an acute stage of inflammatory diseases (especially of the abdomen). Also avoid using Energy Point therapy on patients with hypertension, heart diseases, severe cases of pulmonary tuberculosis, hemophilia, purpura hemorrhagic, purpura allergic, and severe skin diseases.

CHANNEL POINT MEDITATION

After applying the Invisible Needle technique or Energetic Point therapy, the Qigong doctor may decide to initiate Channel Point meditation to intensify the treatment. This is achieved by having the patient concentrate on specific points along the channels. This concentration will produce two distinct results. First, it will get rid of any of the patient's distracting thoughts, and second, it will cause different fluctuations of the Qi and Blood circulation through the patient's Yin and Yang organs. This focused concentration will allow the internal organs to supply the patient's body with either: a more concentrated flow of Qi and Blood into the tissue area; or the focused attention will have a stronger dispersing effect on the patient's tissues (depending on the patient's focus).

Each time a patient switches the focus of concentration to a different point in a different area of the body, the circulation and regulation of Qi and Blood will alter to accommodate the mental transitions of intention. In giving the patient a specific point meditation as a prescription, the Qigong doctor should be careful to select the areas and points in accordance with the particular Qigong exercises that will arrest the patient's Yin and Yang organ disease.

Generally speaking, when treating an Excess condition with Channel Point meditation, the Qigong doctor has the patient imagine draining the Excess Qi out through the extremities. In the case of a Deficiency, the doctor has the patient focus on tonifying the Deficient organ(s) by absorbing energy and light into the Deficient area. Below is a list of channel point regulations and their applications.

QIHAI (CV-6) SEA OF QI

The CV-6 point, located in the Lower Dantian, is the most frequently selected point (Figure 36.3). It is chosen for normal conditions to lead the Qi back to its origin. It is this area from which Qi emanates and returns, hence its name, Sea of Qi. Focusing on this point will allow the patient to tonify the Kidneys (Yuan) Qi, and can be used to regulate the Lower Burner, Conception Vessel, and Water pathways. The Qigong doctor may also extend energy into this area to:

- restore collapsed Yin or Yang Qi,
- raise the patient's Middle Burner Qi,
- warm a patient's Yang Qi,
- warm a patient's Cold condition,
- treat mental disorders,
- treat male sexual problems,
- treat urinary problems,
- treat local pain,
- treat fatigue, and
- treat Qi and Blood Deficiency.

MINGMEN (GV-4) GATE OF LIFE

This point is also called the Palace of Essence (Figure 36.4). Located between the two Kidneys, this point is used for patients with a Deficiency of the Kidney Yang and declining Fire of the Mingmen (this condition is manifested as lumbago pain, seminal emissions, and an aversion to cold). Focus the patient's concentration on the navel first, then gradually have the patient focus deeper into the Mingmen area. This area is known as the Sea of Blood and Essence. Focusing on this point will allow the patient to tonify and stabilize the Kidney (Yuan) Qi and Jing. A Qigong doctor extending energy into this area can regulate the patient's water pathways, heat the body's Yang Qi, and lower blood pressure. This point is also used to treat lower back pain, sciatica, as well as sexual/genital problems.

SHAO SHANG (LU-11) LESSER SHANG

This point is also called Ghost Sincerity, and is considered the second of eleven ghost points (Figure 36.5). The sound "shang" is a musical note and corresponds to the Element Metal. This particular point is used for patients with a Deficiency of Lung Qi (this is manifested by asthma and coughing). It reinforces the Qi and regulates the Lungs. A Qigong doctor can pull energy from this area to dredge and clear Lung Fire, Heat, and Summer Heat, as well as dispel Wind Heat from the patient's body.

ZHONG CHONG (PC-9) MIDDLE RUSHING

This point is used for patients with a Deficiency of Heart Qi (which manifests as palpi-

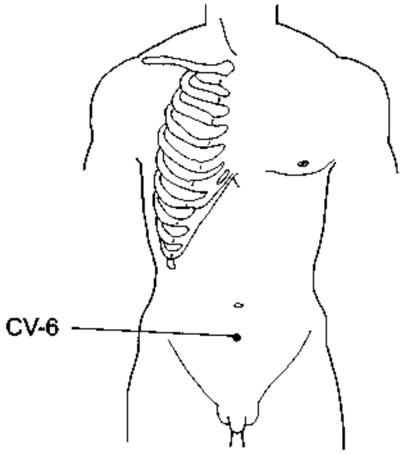

Figure 36.3. Lower Dantian (CV-6)

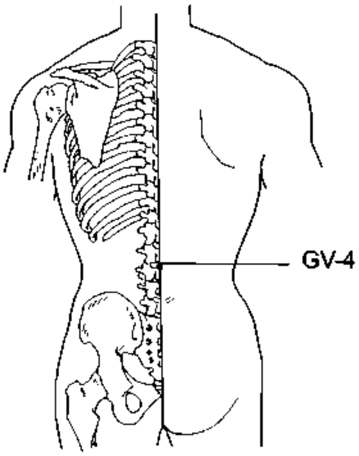

Figure 36.4. Mingmen (GV-4)

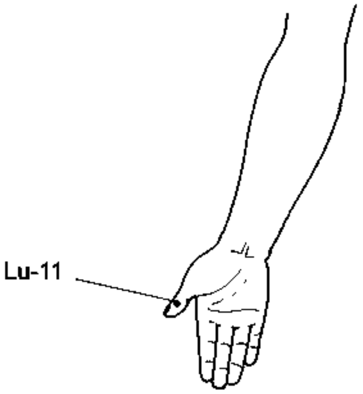

Figure 36.5. Shao Shang (Lu-11)

tations and restlessness). It reinforces the Qi and calms the Heart and mind. This area is good for regulating the Heart's Qi, reviving consciousness, and restoring collapsed Yang. It can be used to dredge and clear the patient's body of Heart Fire, Heat, and Summer Heat conditions (Figure 36.6).

ZU SAN LI (ST-36) FOOT THREE MILES

This point is used for patients with a disorder of transporting and transforming Spleen and Stomach Qi (which manifests as abdominal distention and pain). This point has a strong tonifying effect (especially for the Spleen) and regulates the Stomach. It also tonifies the Ying (Nutritive) Qi, and regulates the Lower and Middle Burners. It can be used by a Qigong doctor to dredge and reduce digestive stagnation, to redirect Rebellious Qi, to drain pathogenic influences from the Stomach, and to soften hard abdominal masses or tumors (Figure 36.7).

DA DUN (LV-1) GREAT PILE

This point is used for patients with hyperactive Liver Yang or overactive Liver and Heart Fire. It calms the Liver to treat diseases of the upper portion of the body (primarily the head). By focusing on this area, the patient can regulate and tonify the Liver Qi and Blood. The Qigong doctor may dredge this area to disperse Liver Qi and transform Damp Heat in the Lower Burner (Figure 36.8).

YONG QUAN (KD-1) GUSHING SPRING

This point is used for patients with Yin Deficiencies of the Liver and Kidney, hyperactivity of Fire due to Yin Deficiencies, or Excess in the upper torso and Deficiency in the lower torso. Concentration on this area will also calm the spirit, clear Fire and Heat from the head area, restore collapsed Yang, and transform Heart Phlegm (Figure 36.9).

BAIHUI (GV-20) HUNDRED MEETINGS

This point is used for patients with a Deficiency of Qi in the Middle Burner (manifested by a shortness of breath, dizziness, intractable diarrhea, and a prolapse of the internal organs). This point is known as the Sea of Marrow. Focusing on this area will clear the brain and calm the spirit. If the Qigong doctor treats this area, the patient's

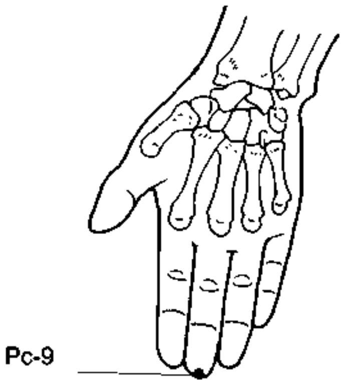

Figure 36.6. Zhou Chong (Pc-9)

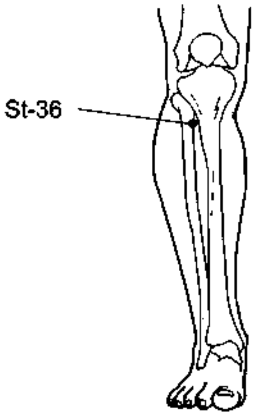

Figure 36.7. Zu San Li (St-36)

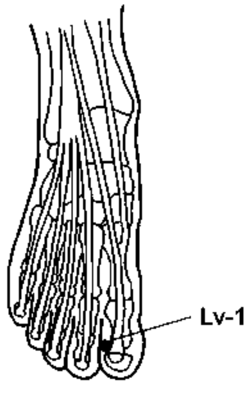

Figure 36.8. Da Dun (Lv-1)

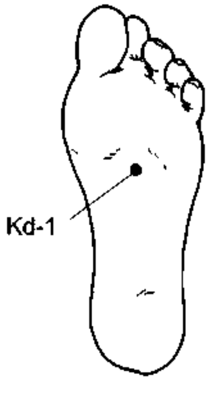

Figure 36.9. Yong Quan (Kd-1)

lower orifices can be stabilized, the Qi can be warmed and tonified, and Yang Qi can be warmed and a collapse can be prevented. This point is also used in order to treat hypertension, insomnia, seizures, dizziness, headaches, and the prolapse of internal organs (Figure 36.10).

HUIYIN (CV-1) MEETING OF YIN

This point is used for patients with a Deficiency of Lung Qi and Kidney Yin (manifested by asthma and coughing). Focusing the patient's concentration on this area stabilizes the patient's Jing and lower orifices, as well as calms the spirit and clears the brain. It can be used by the Qigong doctor to tonify and regulate the patient's Qi, to clear Heat from the patient's body, to treat irregular menses, urethritis, prostatitis, and to raise the blood pressure (Figure 36.11).

YIN TANG (EX. HN. 3) SEAL HALL

This point is used for patients to calm the spirit, and can be used by the Qigong doctor to dispel Wind and clear Heat from the patient's body. This point is also used to treat epilepsy, nausea and vomiting, insomnia, sinus headaches, dizziness, and vertigo (Figure 36.12).

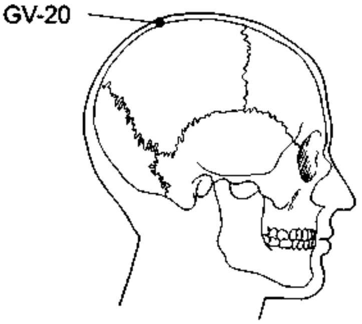

Figure 36.10. Baihui (GV-20)

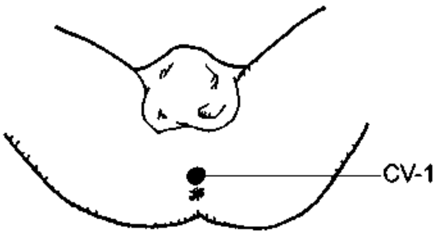

Figure 36.11. Hui Yin (CV-1)

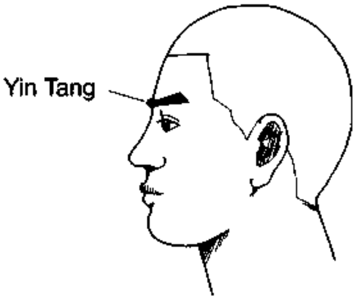

Figure 36.12. Yin Tang (Ex.Pt.)

CHAPTER 37
MEDICAL QIGONG MASSAGE

INTRODUCTION TO TISSUE FUNCTION

To understand the energetic function of surface and deep tissue stimulation through Medical Qigong Massage, it is important to first understand the reflexes and spontaneous energetic reactions within the body's tissues. The tissues of the human body are in a constant state of perpetual change. No matter where illness or damage has occurred, everything in the body is linked to the organs' and bowels' functions of producing and governing energy, and their interaction with the body's surface tissues (Figure 37.1–37.2).

In Traditional Chinese Medicine, the human body is viewed as a conglomeration of Qi, Blood, channels, tendons, fascia, bones, Marrow, bowels (Yang organs), and viscera (Yin organs). Each part is interdependent, each having inseparable Yin and Yang relationships and all of them interacting with each other. The following describes the interaction of: Qi and Blood, Tendons and Fascia, Bones and Marrow and the Bowels and Viscera.

1. Qi is considered the Blood's leader: as the Qi leads, the Blood follows. If the Qi ceases to flow, the Blood stops.
2. The tendons and fascia are used for channel energy movement. As the tendons and fascia stretch, they cause the channel Qi to increase its energetic movement and flow. When the tendons and fascia are calm, the channel Qi slows down.
3. The bones are the Marrow's residence. When the bones are hard, the Marrow is solid; when the bones are soft, the Marrow is hollow.
4. The bowels (Yang organs) are the viscera's (Yin organs') manifestations. If the bowels are strong, the organs become robust; when the bowels become weak, the viscera begin to fail.

Medical Qigong massage takes into consideration the function and flow of the body's tis-

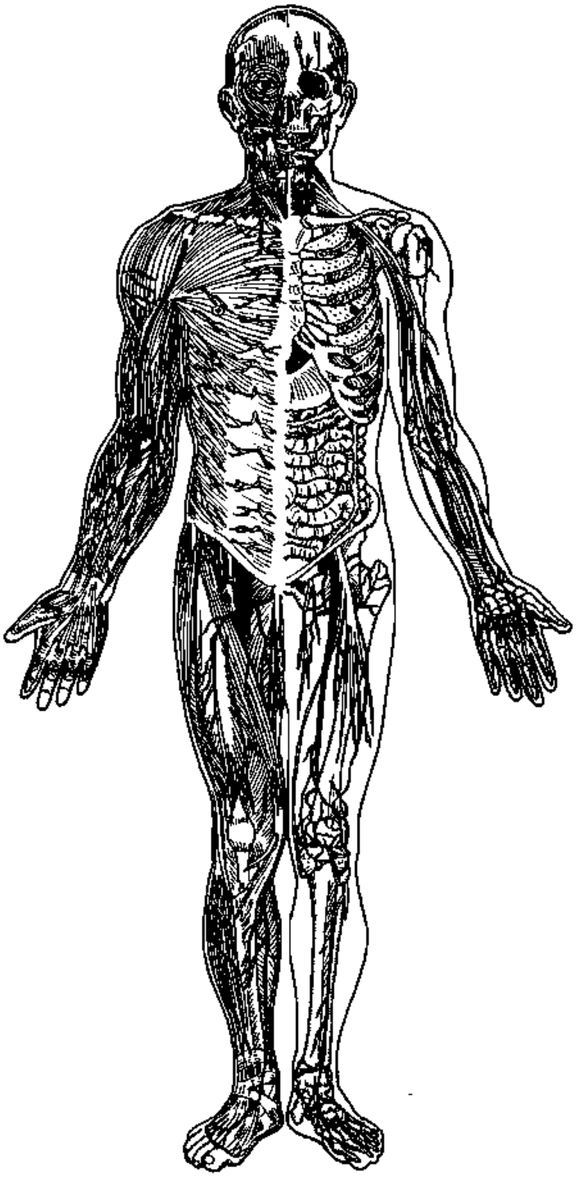

Figure 37.1. The Human Body (Anterior)

sues, in conjunction with the interacting systems of the body's energetic fields, striving to regulate both systems at the same time.

TWO APPROACHES IN TREATMENT

Medical Qigong massage can be divided into two schools of treatment modalities: Surface Tissue Massage Therapy and Visceral Tissue Massage Therapy.

1. Surface Tissue Massage Therapy utilizes the Five Elemental Qi Massage Therapy techniques in conjunction with external tissue stimulation. All external tissue manipulation techniques used in Qigong massage should be light as a feather and should not exceed the pressure one would place on an eyeball. When treating with Surface Tissue Massage, the doctor's focus must be centered on the dispersion of pathogenic Qi and the expansion of the patient's Wei Qi, while softly touching the patient's tissues.
2. Visceral Tissue Massage Therapy utilizes the Five Elemental Qi Massage Therapy techniques in conjunction with deep tissue stimulation. By matching the patient's pulse frequency and respiration with that of the doctor's, a stable, energetic resonance is secured and treatment begins. When treating with Visceral Tissue Massage, the doctor's focus must be centered deep into the patient's internal organs without touching the patient's external tissues.

After external tissue massage is used to treat the patient's tissues, it is important for the doctor to maintain focused intention on the patient's channels while slowly removing his or her hands. If a break in contact with the mind's intention is made, the patient's tissues will return to their pretreatment state. Once the doctor has physically disconnected from the patient in this manner, Distance Qi Emission and Qi manipulation techniques can be applied.

It is believed that the treatment of diseases should be carried out primarily by Qi emission as the main approach, and secondarily through Qigong massage. By combining both external and internal tissue manipulation skills with energy

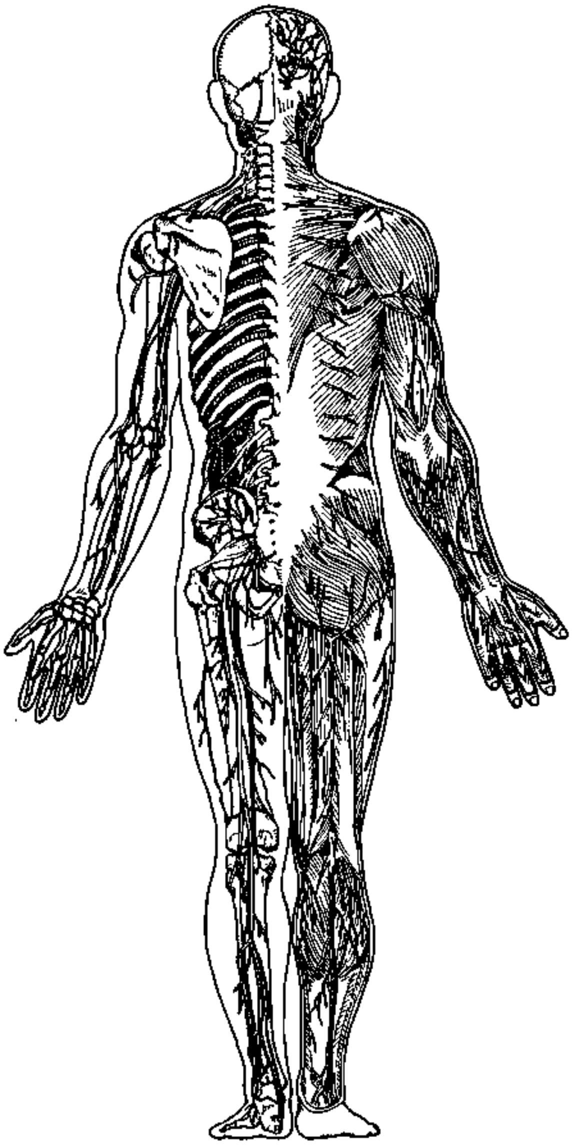

Figure 37.2. The Human Anatomy (Posterior)

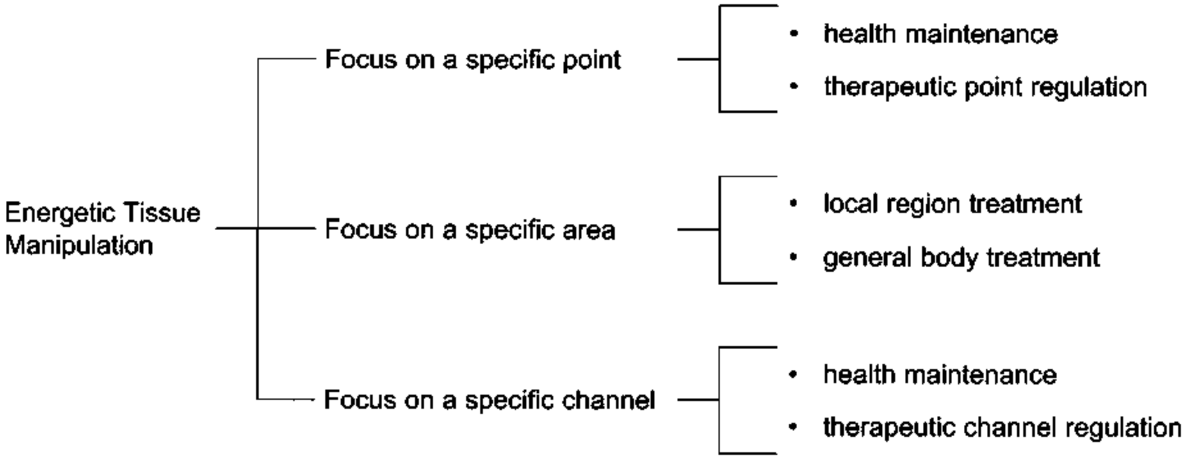
Energetic Tissue Manipulation
Focus on a specific point
• health maintenance
• therapeutic point regulation
Focus on a specific area
• local region treatment
• general body treatment
Focus on a specific channel
• health maintenance
• therapeutic channel regulation

Figure 37.3. Energetic Tissue Manipulation Chart

extension techniques, the Qigong doctor greatly improves the quality of treatments. In the clinical setting, this type of energetic tissue manipulation is utilized before or after energy extension. This facilitates the relaxation of the patient's muscles and joints and the opening or closing of their points. It also aids in dredging the channels, guiding the flow of energy, and in regulating Yin and Yang energy.

TISSUE REGULATION THROUGH MEDICAL QIGONG MASSAGE

The external treatment of the patient's tissues encompasses a wide variety of body manipulation and alignment methods to provide a complete treatment program (Figure 37.3). Used for toning the organs, glands, nervous system, and joints, Qigong massage is applied to treat soft tissue injury. The following are the six treatment goals of Qigong Massage:

- To relax the tendons, activate the channels, promote circulation of Qi and Blood, and relieve pain,
- To disperse Blood stagnation,
- To relieve muscle spasms,
- To expand the tendon sheaths,
- To treat fascia in order to move stagnations and separate adhesions, and
- To correct dislocations.

The following is a list of the five predominant external tissue manipulation techniques used by Qigong doctors in China today.

THE FIVE ELEMENTAL QI MASSAGE THERAPY

Traditionally there are five different techniques for Medical Qigong massage therapy. These five techniques are very important for reestablishing the energetic vitality of the patient's body. Each of these techniques relates to a specific action and organ associated with one of the Five Elements. The objective of these techniques is to influence the muscles, nerve fibers, and deep-lying tissues of the body in order to aid metabolism and stimulate energy flow (Figure 37.4). The applications of these five methods of energy manipulation and their affect on the body's tissues are described as follows.

THE CIRCLE-TWISTING METHOD

This technique relates to the Spleen, which rules and affects the muscles, and is implemented as a quick drilling motion into the body's surface tissues, using the fingers, palms, knuckles, or heel of the palm.

When treating a patient's tissue area, a circle-twisting movement will stimulate the Qi and Blood within the channel points. This will generate a numbing and tingling sensation that will

Circle-Twisting Method	Spleen-Earth Element: affects the muscles, stimulates and improves Blood circulation.
Compression Release Method	Kidneys-Water Element: affects the bones, and regulates the Wei and Ying Qi.
Thrusting Method	Heart-Fire Element: affects the Blood, pulse, and regulates and activates the Qi.
Grasping and Shaking Method	Liver-Wood Element: affects the tendons, ligaments and muscles, promotes, clears, and activates Qi flow, and balances the Yin and Yang energy.
Tapping Method	Lungs-Metal Element: affects the Qi and skin.

Figure 37.4. The Five Elemental Qi Massage Methods

cause the channel points to induce the energy to counterattack the foreign invasion of Turbid Qi and disease. The circle-twisting action will also stimulate and improve Blood circulation and is divided into two modalities of treatment:

- To stimulate, massage the area in a clockwise, spiraling, circular motion from the outside to the center of the point. The purpose of this action is to gather energy from the surrounding areas of the patient's body and collect it into the center point the doctor is treating.
- To sedate, massage the area in a counterclockwise, spiraling circle from the center of the point outward. The purpose of this action is to disperse the energy from the area the doctor is treating. This Excess energy will later be either moved into other organs, or dredged outside the patient's body through an adjacent channel (Figure 37.5).

THE COMPRESSION RELEASE METHOD

This technique relates to the Kidneys, which rule and affect the bones, and is performed by the doctor extending his or her intention into the patient, softly pressing the body's surface tissues, using the whole palm, parts of the fingers, or elbows. The Compression Release method is commonly used to purge or drain the tissues of pathogenic Qi.

The Compression Release techniques are generally used to regulate the Ying and Wei Qi, which travel along the channels and collaterals. Deep pressure reaches the Ying Qi, while shallow pressure stimulates the Wei Qi. Once damage has occurred in the body, the damaged channel points and organ regions cause certain changes in the Wei Qi, Ying, Qi, and Blood areas of the body, resulting in numbness, aches, pain, swelling, etc. Both Circle-Twisting and Compression Release techniques can eliminate these symptoms and create a more stable energy flow.

The Compression Release Technique

Once the proper hand posture and area to be treated has been selected, the doctor guides Qi into his or her palm and fingers using Hot, Cold, vibrating, or spiraling energy. The doctor then initiates a rhythmic compression over the patient's tissues stimulating the patient's energetic field. This action has the function of relieving stagnation or congestion and can be used to dredge the

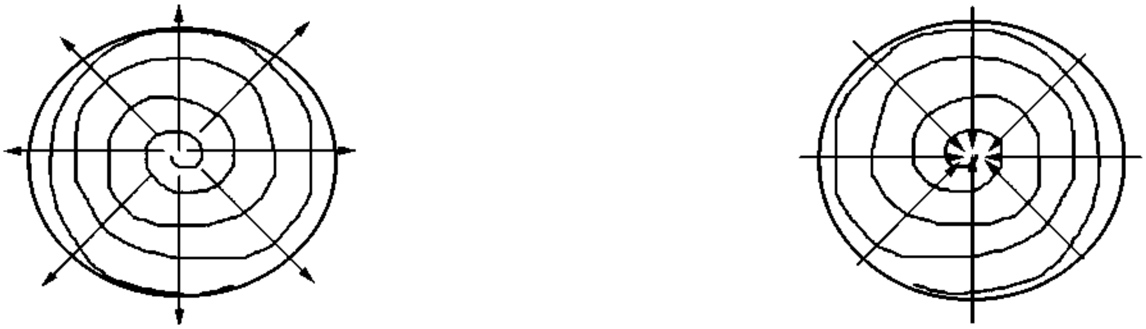

To sedate, purge, or disperse, move the energy in a counterclockwise direction.

To stimulate, tonify, or gather, move the energy in a clockwise direction.

Figure 37.5. Circle Twisting Method of Energy Therapy

channels and energetic points, as well as to relieve muscle spasms and physical pain.

The doctor must apply the correct amount of pressure for the patient's physical condition:

1. A slow Compression Release method is applied to any kind of hyperactive illness, nervous system disorder, mental disorder, or loss of Blood.
2. A medium Compression Release method is applied to illness for Spleen and Stomach disorders.
3. A fast Compression Release method is applied to hypoactive or Cold illnesses.

It is important to note that when applying the Compression Release method to tonify a specific area, the doctor should energetically press on the channel at a forty-five degree angle, slightly above where the channel flows into the injured area. This augments the flow of energy along the channel.

To disperse or drain a particular area, the doctor should apply slight pressure at an angle past where the channel flows into the injured area. This downward pressure will cause the energy to flow away from the injury. This application is used when treating Excess conditions.

THE THRUSTING METHOD

This technique relates to the Heart, which rules and affects the Blood and Blood pulses. It is utilized to gently push the tissues, extending the Qi and Blood along the patient's surface channels, using the thumb, ball of the thumb, or heel of the palm. The thrusting method can draw energy into a Deficient area or move Qi out of an Excess area.

The Thrusting Technique

Once the proper hand posture and area to be treated has been selected, the doctor guides Qi into his or her palm and fingers using Hot, Cold, vibrating, or spiraling energy. The doctor then rubs lightly or lightly touches above the patient's tissues while simultaneously emitting Qi.

The technique should be synchronized with the frequency of the patient's pulse and respiration. The Thrusting technique has the function of regulating and activating the patient's Qi, alleviating pain, and dredging the channels.

THE GRASPING AND SHAKING METHOD

This technique relates to the Liver, which rules and affects the tendons. It is administered by lightly vibrating the skin (gently at first, then slowly increasing the intensity). The Grasping and Shaking method sends a wave of vibration throughout the patient's body, and is used to increase Qi and Blood flow to the tissue area.

The Grasping and Shaking Technique

Once the proper hand posture and area to be treated has been selected, the doctor guides Qi into his or her palm and fingers using Hot, Cold, vibrating, or spiraling energy.

The doctor then lightly grasps and shakes the energetic field within the patient's tissues, while extending energy. The Grasping and Shaking technique is used for directing Qi to flow with or against the natural currents in the body's channels. It promotes, clears, and activates the channels and flow of Qi, relaxes the muscles and tendons, and balances the Yin and Yang energy.

THE TAPPING METHOD

This technique relates to the Lungs, which rule and affect the Qi and skin. This technique is administered by lightly and rhythmically tapping the patient's skin (with varying degrees of force), using the fingers, palms or fists (Figure 37.6). The Tapping method sends a pulsating ripple throughout the patient's tissues and is used to disperse stagnation from a point, channel, or organ.

YIN AND YANG ENERGETIC POINT MANIPULATION

Before treating patients with Five Elemental Regulation therapy, the Qigong doctor should be aware of the principals of Yin and Yang energetic polarity involved within the Energetic Point Manipulation. This method of treatment is used to regulate the patient's Wei Qi, Ying Qi, and Qi and Blood systems (as a whole), in order to form a governed systemic balance. In the Energetic Point Manipulation methods, the treating techniques are directly based on Yin and Yang theory.

THE YIN THEORY OF QI MANIPULATION

This technique is expressed as a passive, quiet, or motionless action (more energy and less physical movement). The energetically passive Yin movements are used to push, fill, or tonify the patient's Righteous Qi, located within the patient's tissues. It is used for treating external tissue conditions and for internal organ conditions.

- If the external tissue area is in an inactive or Yin state, apply a Yin technique. Treat a Cold area with soft, passive manipulations, while extending the mind superficially onto the patient's surface tissues to fill or tonify anti-pathogenic factors, and to revitalize the Deficient area.
- If the internal organ area is either in an active or Yang (Hot or hard) state, the doctor must apply a Yin, passive, or slow action technique, while extending his mind deep into the patient's tissues and organs.

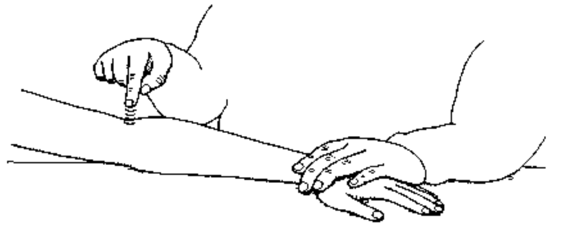

Figure 37.6. In the Massage Tapping Method, the Qigong doctor will extend his or her intention deep into the patient's tissues in order to stimulate the channel point.

THE YANG THEORY OF QI MANIPULATION

This technique is expressed as an active technique, in motion. The active energy of the Yang movements are generally used to pull or purge pathogenic factors. It is used for treating external tissue conditions and for internal organ conditions.

- If the external tissue area is in an active or Yang state, apply a Yang technique (treating an External Hot or hard area with fast dredging manipulations) while extending the mind superficially onto the patient's surface tissues. This is done to pull or purge pathogenic factors, or break up the stagnation.
- If the internal organ area is either in an inactive or Yin (Cold or flaccid) state, the doctor must apply a Yang, quick, or active technique, while extending his or her mind deep into the patient's tissues and organs.

This combination of movement (Yang) and non-movement (Yin) is embodied in both the Circle-Twisting and Compression Release Methods of the Five Elemental Regulation Techniques.

APPLICATION FOR YIN AND YANG TECHNIQUES

The following text gives examples of specific Yin or Yang hand techniques, and when to apply them:

1. The Qigong doctor uses Yin hand techniques for tonifying, filling and pushing Qi and

Blood through the patient's organs, tissues, and channels. The Yin hand actions are demonstrated through slow or inactive hand movements.

2. The Qigong doctor uses Yang hand techniques for sedating, purging, and pulling Qi and Blood. The Yang hand actions are demonstrated through fast or active hand movements.

YIN AND YANG TREATING METHODS

The following text gives examples of specific Yin or Yang hand treatment methods, and when to apply them:

1. When a patient's tissues are in an Excess Yang state (Hard or Hot), the Qigong doctor sedates the area with counterclockwise rotations, leading the Excess energy from inside the tissue area out of the body.
2. When a patient's tissues are in a Deficient Yin state (Flaccid or Cold), the Qigong doctor tonifies the area with a clockwise rotation, to bring energy into the Deficient tissue area.

TREATING THE TISSUE AREA OR ENERGETIC POINT

In treating the patient's surface tissue area or specific energetic points, some Qigong schools teach that with a male patient, one should begin with points on the left side, then on the right side; whereas for female patients, one begins working on the points on the right side of the body, then on the left.

DEEP AND SHALLOW PRESSURE

Energy compression (deep or shallow) is controlled by the doctor's release. A deep compression relates to the Ying Qi, while a shallow compression pertains to the Wei Qi. This Compression Release technique affects the body's Blood and vital energy harmony. It seeks to balance the body's energy by preventing the Qi from flowing in the wrong direction. This technique also leads the pathogenic Qi outside the patient's body and is commonly called "The Dispersing Evil Qi Technique."

If stagnant energy remains in the center of a channel point, then the doctor must employ a Compression Release method. The compression causes the channel point to contract, leading it toward a calm, stable state. When the doctor releases the compression on the channel point, it expands and becomes stimulated into an active state.

THE GRADUAL ENERGETIC PENETRATION OF TISSUES

When treating a patient with External Tissue Regulation, there should always be a gradual penetration of energy permeating the patient's tissues. The Qigong doctor first sinks his or her focused intention onto the patient's surface tissues. The intention is then slowly sunk deeper, first into the patient's muscles and inner fascia, then into the bones or internal viscera.

At no time should pressure damage the tissues or compound pain. Treatment or pressure time at each area or energetic point should not be excessive. To establish the correct amount of pressure, determine the normal energetic pulsation in accordance with the Five Elemental Qi Massage methods, then decide if the technique is to be light, medium, or heavy.

ADDITIONAL GUIDELINES

The following are specific guidelines of Qigong massage treatment methods, and when to apply them:

1. Before beginning Qigong massage, the patient must be comfortable and relaxed. Otherwise the patient's body will not be receptive to the tissue stimulation or Qi emission.
2. When first touching the patient, the touch should be slow, as if the doctor is dipping his or her hand to test the water. The movements should have an established rhythm.
3. The doctor should never hold his or her arm, or the patient's arm, rigid or fully extended; the elbows should always be bent to some degree.
4. A symptom or damaged area occurring on one area or point at the extremity of a channel may be effectively treated by stimulating the point at the opposite end of the channel (see Magic Square, Chapter 31).
5. Points which are painful with light pressure

are in a state of Yang and need to be purged and sedated.

6. Points which are painful to deep pressure are in a state of Yin and need stimulation.
7. When an organ is functioning normally, it should receive weak, short stimuli that will increase the function, action, and energetic potential of the organ.
8. Cool skin over a point or channel indicates a Deficiency in that channel and should be tonified.
9. Hot skin indicates Excess Qi accumulation. It also indicates a current or pending illness in an organ, or area associated with that channel, and should be purged immediately.

SECTION IX

MEDICAL QIGONG EXERCISES AND PRESCRIPTIONS

CHAPTER 38

INTRODUCTION TO MEDICAL QIGONG PRESCRIPTIONS AND REGULATION EXERCISES

HISTORY OF MEDICAL QIGONG PRESCRIPTIONS

Many of the earliest known Medical Qigong prescriptions were derived from the movements of animals. Dating back to the second century B.C., the Qigong Classic (Dao Yin Tu) illustrates, in manuscripts written on silk, over 45 Qigong postures with descriptions of the movements as well as the names of the diseases which they treat. Over half of these illustrated postures are animal movements.

In the ancient Chinese text called the *Spring and Autumn Annals*, it is written, "Flowing water never stagnates, and the hinges of an active door never rusts. This is due to movement. The same principle applies to the body's Jing and Qi. If the body does not move, Jing does not flow. When Jing does not flow, Qi stagnates." The soft flowing movements in the Medical Qigong prescription exercises enhance and increase the flow of the body's Qi and Blood.

THE FUNCTION OF MEDICAL QIGONG PRESCRIPTIONS

When a disease develops, certain active points of the disease correspond to different organs, muscles and limbs of the body, and a pathological reflex is formed. Trigger and reflex points which are painful to touch are perfect examples of this phenomenon. The cerebral cortex can be inhibited in specific ways by regulating the Mind and entering a state of tranquility through Medical Qigong prescriptions, thus overriding the pathological reflex pattern, allowing the body time to heal.

Although acupuncture and herbs can, in addition to relieving pain, move Qi, Blood, and Phlegm (the roots of the disease), they fail to address the patient's life-style which is often a major contributing factor. Problems such as stagnation, Excess or Deficient conditions are often exacerbated by a patient's sedentary life style. A responsible doctor of Medical Qigong therapy addresses this particular element by prescribing physical therapy, exercise, and meditations as part of the overall treatment strategy.

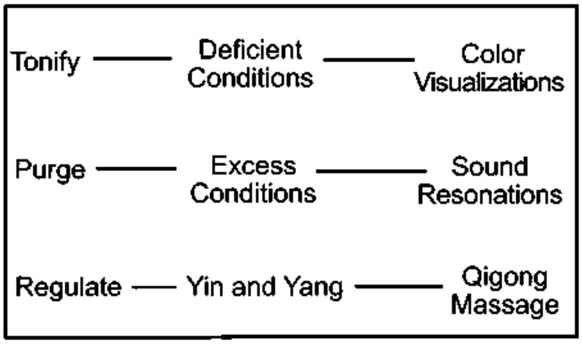

Figure 38.1. Medical Qigong Treatments

The standing rule for treatment in Medical Qigong prescriptions is to:

- Tonify the Deficient organs and organ systems with color visualization, by using the Mind's intention to focus on moving Qi into a Deficient organ, strengthening, thus illuminating its tissues;
- Purge the Excess organs and organ systems with sound resonation, by using the Mind's intention to focus on moving the Qi out from an Excess organ, resonating, and draining its tissues;
- Regulate the body's Yin and Yang organs and organ systems with Qigong massage, by using the Mind's intention to balance the energetic fields by moving the Qi up and down, right and left, and inside and outside of the tissues (Figure 38.1).

The therapeutic effects of Medical Qigong meditations and prescriptions are designed to move Qi and Blood stagnation, to tonify the internal organs and organ systems, and to enhance the autonomic nervous system. These prescriptions stimulate the patient's body, thereby gradually eliminating fatigue, as the normal body functions are restored. Another important aspect of the Medical Qigong meditations and prescriptions is the enhancement of the body's immune system functions.

MEDICAL QIGONG AND THE IMMUNE SYSTEM

The many health benefits of Medical Qigong include increased white blood cell production, stabilized blood pressure, and an improved immune function. The immune system, unlike the vascular circulation system, defies a clear anatomical description. It can only be understood through its functions and interrelationships with the mind and various other parts and systems of the body, such as the lymph nodes, Spleen, tonsils, appendix, Peyer's patches (in the ilium), thymus gland and the bone marrow.

The lymph nodes contain specialized compartments - some contain B cells (B stands for Blood), some T cells (T stands for Thymus), and some contain macrophages. The tonsils likewise contain B and T cells. The appendix and Peyer's patches are components of the lymphatic drainage system, and serve as locations where the B cells mature and where antibodies are produced by the B cells. The body's T cells mature in the thymus gland.

Both B and T cells begin their life in the Liver, when the fetus is only nine weeks old. They then migrate to the bone marrow where they form precursor cells. Afterwards, the T cells migrate to the thymus and the B cells remain in the bone marrow.

The immune system utilizes two major strategies for protecting the body from an infectious disease: The Humoral Immunity and the Cell Mediated Immunity.

- In humoral immunity the B cells play the primary role, and the T cells as well as other related cells play supporting roles.
- In cell mediated immunity the T cells play the main role, while B cells and other related cells play auxiliary roles.

Both T cells and B cells are called "Lymphocytes." They are two types of white blood cells that originate in the body's bone marrow and travel in the blood and lymph circulatory system (Figure 38.2). Both T cells and B cells are concentrated in the lymph nodes, which are found in clusters near groups of veins, and are heavily concentrated under the arm pits, in the adenoids, the Spleen and in the groin area.

When B cells encounter an antigen or foreign substance, they differentiate into plasma cells. These plasma cells secrete antibodies that fight and neutralize bacteria, or coat the antigen to mark it for destruction by other components of the immune system (such as the T cells).

T cells specialize in destroying cells infected by viruses, in attacking tumors, and in rejecting transplanted organs and tissues. The T cell, after leaving the bone marrow, travels to the thymus gland where it undergoes further development before returning to the lymph and blood circulatory systems. The T cells respond to an antigen by differentiating into several kinds of T cells, three of which are of primary importance: Killer T cells (also called cytotoxic T cells), Helper T cells and Suppressor T cells. Killer T cells bind to viral antigens displayed on the surface of infected cells and destroy them. Helper T cells bind to the antigens on the surface of a B cell that has already bound itself to the antigen. Each Helper T cell then releases Lymphokines (one well know class of lymphokines is interferon), that enhance the maturation of the B cells, or signal them to differentiate and multiply into plasma cells that secrete antibodies. The Suppressor T cells function to dampen the immune response and to signal the "all clear."

A third group of important immune cells are the phagocytes. These cells are the body's natural scavengers. They engulf antigens and destroy them. Those that specialize in cleaning up the debris are called macrophages. The macrophages can also signal B or T cells to the presence of foreign material.

Another important group of cells are called Null cells, among which the N.K. or "natural killer" cells are found. These cells are able to identify tumors and virally infected cells, and kill them without having any prior encounter with such an antigen.

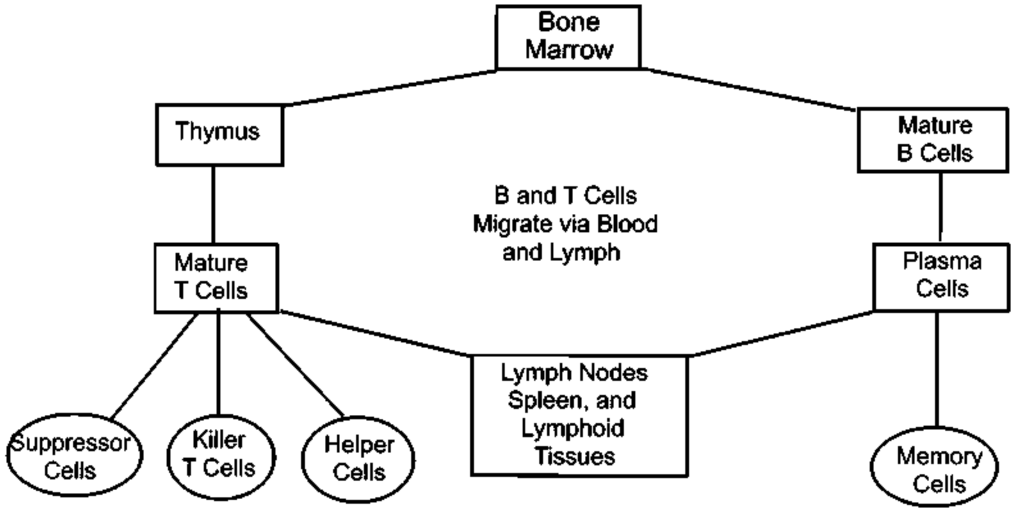

Figure 38.2. The Body's Immune System

A final important group consists of Mast cells. They contain powerful chemicals: heparin - an anticoagulant, serotonin - a neurotransmitter, and histamine - a biochemical released in overabundant quantities in allergic reactions.

These white blood cells live in the blood and lymph fluid, and are commonly known as the leukocytes. There are two main varieties of white corpuscles: The Polymorphonuclear Leucocyte (Polymorphs) and Lymphocyte.

- Polymorphs are found in the bone marrow stem cells; they destroy invading bacteria and remove dead or damaged tissue.
- Lymphocytes are found in the lymph glands, Spleen and thymus gland and circulate in the blood along with the polymorphs. They are the smallest of the leucocytes, and designated the lymphocyte because it is identical with the lymphoid cell derived from the lymphatic glands.

Both the white blood cells called polymorphs and lymphocytes move into the body's surrounding tissues as a part of the immune function. The purpose of these cells is to work together to defend the body from invasion of pathogenic factors. Their function is to identify, ingest, destroy and eliminate disease-causing organisms, as well as to remove any and all dead, damaged, or irregular cells.

Immune cells are produced in the lymphoid tissue and red bone marrow, and are stored in the lymphatic organs and blood. In order to perform their function of protecting and cleaning, these cells must circulate throughout the entire body. The capillaries and lymphatic vessels carry the lymph fluid through the lymph nodes and the lymphatic organs of the thymus gland, spleen, and tonsils, where the lymph fluid is filtered, cleansed of impurities, and infused with new lymph cells.

In order for an external pathogen to affect the body, it must first pass through the protective layers of cells. Generally, the macrophages (specialized white blood cells) present within the tissues ingest the pathogens and dead tissues. This action of the macrophages is reinforced by the action and movement of the Wei Qi. When a pathogenic intruder invades the body, the macrophages are the first to attack the intruder. If the macrophages are not strong enough, more blood and fluid, which contain more immune cells, are brought to the area. The increased fluid engorges the tissue, and brings more cellular activity. The increased

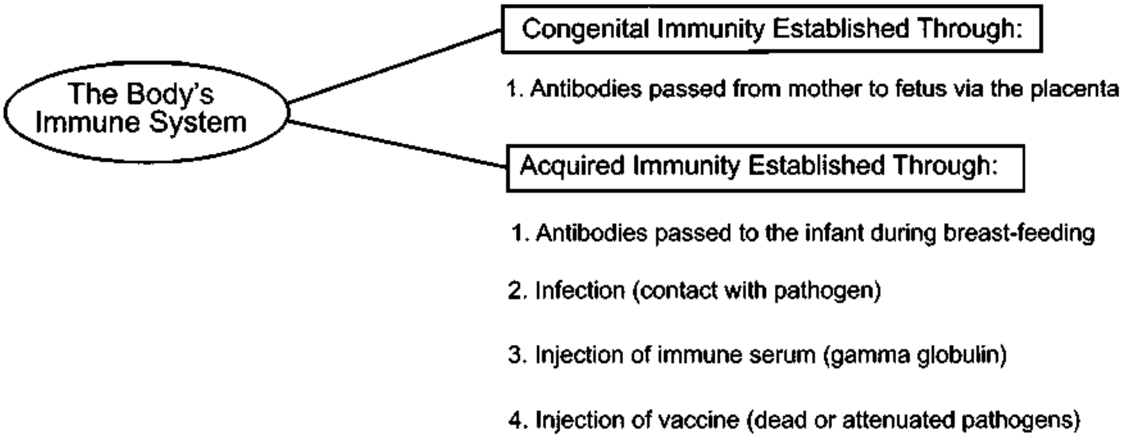
The Body's Immune System
Congenital Immunity Established Through:
1. Antibodies passed from mother to fetus via the placenta
Acquired Immunity Established Through:
1. Antibodies passed to the infant during breast-feeding
2. Infection (contact with pathogen)
3. Injection of immune serum (gamma globulin)
4. Injection of vaccine (dead or attenuated pathogens)

Figure 38.3. Congenital and Acquired Immunity

blood and cellular metabolism increases the local temperature. The immune cells try to break down and ingest the pathogen, creating a process which continues until the disease is eliminated.

Once the battle is won, some of the Killer T cells and Helper T cells will remain circulating in the blood and lymph circulatory system as Memory cells that will be able to deal more efficiently with any future antigen of the same kind. Likewise, some of the mature B cells will also become Memory cells that hover around lymph nodes which they use as a base of operation.

Because Qigong movements generate heat and soften the body's connective tissues, this encourages the flushing of the intercellular fluid throughout the body's tissues and organs. When combined with Breathing, Mental, and Postural Qigong exercises, the entire body is affected. These energetic Qigong actions increase oxygenation of the blood and improve circulation, thereby enhancing lymphocyte production and strengthening the body's immune system.

CONGENITAL AND ACQUIRED IMMUNITY

The body has two ways of establishing or maintaining its immune functions: Congenital Immunity and Acquired Immunity (Figure 38.3).

- Congenital Immunity (Prenatal) is established during gestation, through antibodies that are passed from the mother to the fetus via the placenta.
- Acquired Immunity (Postnatal) is established in the following ways: immediately after birth during breast-feeding (through the colostrum-mother's first milk), through contracting an infection (contact with external pathogens), the injection of an immune serum (gamma globulin), and the injection of vaccines (dead or attenuated pathogens).

Due to the importance of establishing a strong immune system within the body's Prenatal Jing, Traditional Chinese Medicine strongly endorses fetal education, instruction on diet and nutrition, maintaining emotional well-being during pregnancy, and Qigong exercises specifically designed to enhance the child's Prenatal Essence, Energy, and Spirit.

OVERVIEW

One specific goal of Medical Qigong treatments, exercises and meditations is to activate and strengthen the immune function. Qigong therapy can be used to not only strengthen the metabolism of the organs and cells, but also increase cellular energy, thus making the immune system stronger.

In China, even though the lymphatic function itself was unknown and unnamed, its effects were

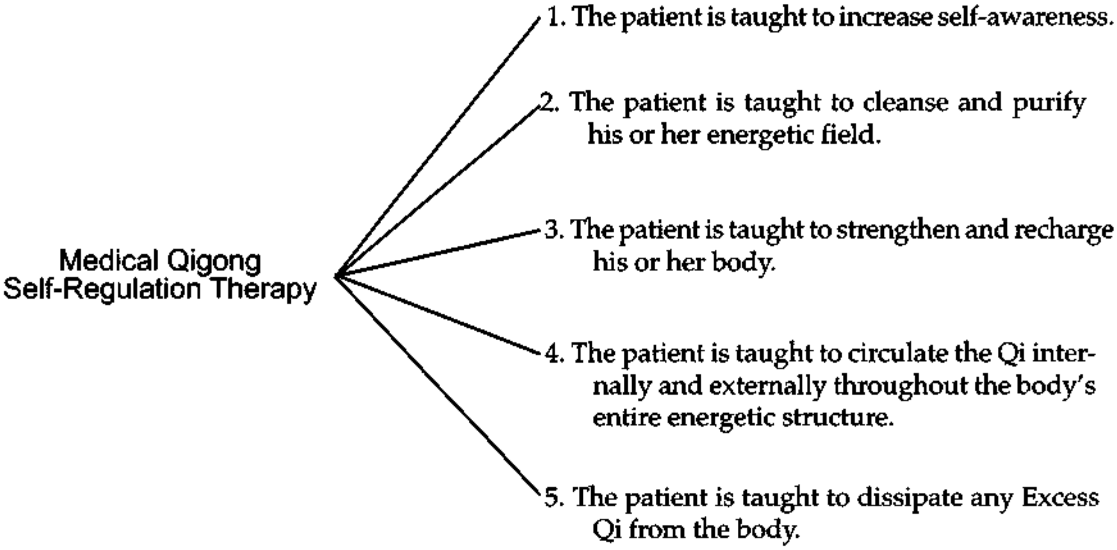
Medical Qigong Self-Regulation Therapy
1. The patient is taught to increase self-awareness.
2. The patient is taught to cleanse and purify his or her energetic field.
3. The patient is taught to strengthen and recharge his or her body.
4. The patient is taught to circulate the Qi internally and externally throughout the body's entire energetic structure.
5. The patient is taught to dissipate any Excess Qi from the body.

Figure 38.4. The Five Stages of Healing Transitions That Occur Through Self-Regulation Therapy

generally ascribed to the proper action of the Qi and Body Fluids. Therefore, elaborate methods for generating and circulating lymph were developed, and faithfully practiced through Qigong. Breath, movement and postural modification as well as meditations have specific effects on the circulation of the lymph and therefore enhance the overall function of the lymphatic system.

THE FIVE STAGES OF HEALING

When practicing Medical Qigong exercises and meditations, there are five basic stages of healing transitions which occur through self-regulation therapy (Figure 38.4).

1. The patient is taught to increase his or her awareness of the body (physically, mentally, and emotionally) and its current condition.
2. The patient is taught to cleanse and purify the body's energetic fields to rid it of Qi stagnations and toxic pathogenic factors.
3. The patient is taught to strengthen and recharge his or her body to replenish the Qi.
4. The patient is taught to circulate the Qi internally and externally throughout the body's entire energetic structure, thus moving any stagnant Qi and strengthening the body.
5. The patient is taught to dissipate any Excess Qi from the body by way of self-massage (lightly brushing the skin to release stagnant Qi).

THE THREE CATEGORIES OF MEDICAL QIGONG PRESCRIPTIONS

The Medical Qigong prescriptions are divided into the same three categories (or goals) used in Medical Qigong therapy: tonification, purgation and regulation. These three goals are achieved through both Dynamic and Quiescent Qigong exercises, meditations, prayers, and mantras. These prescriptions enable the patient to reinforce the treatments performed by the Qigong doctor, and prevent the patient's energy from reverting back to old destructive patterns.

The Qigong doctor selects the proper Qigong prescription to fit the patient's illness (to Tonify, Purge, or to Regulate) and formulates a rational program that suits the patient's case. The patient should never be allowed to practice at random, otherwise pathological changes or Qi deviations may occur.

The following are but a few examples of Qigong prescriptions, which are interchangeable, as there are certain variations of purging exercises that can be modified and used for tonification and visa versa.

QIGONG PURGING EXERCISES

Qigong Purging exercises are practiced by the patient to detoxify pathogenic Qi. Some examples are already presented in this textbook are:

- Channel Purging exercise (Chapter 15)
- Cleansing the Body meditation (Chapter 12)
- Dry Crying meditation (Chapter 53)
- Energy Melting meditation (Chapter 18)
- Internal Organ Sound Therapy exercises (Chapter 4 and Chapter 40)
- Old Man and Tide Pool meditation (Chapter 40)
- Pulling Out The Pain meditation (Chapter 28)
- Releasing Emotional Blockages and Energetic Armoring meditation (Chapter 18)
- Sun and Moon meditation (Chapter 18)
- Tumor and Cancer Tone Resonating exercises (see Chapter 40)

QIGONG TONIFICATION EXERCISES

Qigong Tonification exercises are practiced by the patient to strengthen and stabilize the internal organs and organ systems. Some examples are already presented in this textbook are:

- Channel Point meditations (Chapter 36)
- Color Visualization meditations (Chapter 4)
- Cultivating Yuan Qi meditation (Chapter 50)
- Daoist Five Yin and Yang Organ exercises (Chapter 42)
- External Qi Cultivation meditations (Chapter 41)
- Gaining Wisdom meditation (Chapter 49)
- Gathering Energy from the Five Planets meditations (Chapter 11)
- Gathering Moon Cream meditation (Chapter 11)
- Gathering Sun's Essence meditation (Chapter 11)
- Gathering Qi from Mountains, Valleys, and Deserts (Chapter 12)
- Gathering Qi from The Four Directions (Chapter 12)
- Gathering Qi from Trees, Bushes, and Flowers (Chapter 12)
- Improving the Intellectual Facilities meditation (Chapter 49)
- Internal Qi Cultivation meditations (Chapter 41)
- Nei Guan - Inner Viewing Meditation (Chapter 25)
- One Through Ten meditation (Chapter 26)
- Opening and Closing the Three Dantians exercise (Chapter 41)
- Organ Massage exercises (Chapter 39)
- Qigong Walking exercises (Chapter 15)
- Seeing Auras meditation (Chapter 25)
- Soul Retrieval meditations (Chapter 19)
- Tian Wu Zang meditation (Chapter 11)

QIGONG REGULATION EXERCISES

Qigong Regulation exercises are practiced by the patient to balance their Yin and Yang, as well as their internal organ energy. Regulation exercises usually follow Tonification or Purgation exercises, because the goal is to reestablish balance. Some examples already presented in this textbook are:

- Exchange of Fire and Water meditation (Chapter 18)
- Gathering Qi in the Upper and Lower Dantians (Chapter 41)
- Microcosmic Orbit meditations Chapter 41)
- Opening and Closing the Three Burners (Chapter 41)
- Opening and Closing the Three Dantians exercise (Chapter 15)
- Pulling Down the Heavens (Chapter 41)
- Qigong Self-Healing massage (Chapter 41)
- Taiji Ruler exercises (Chapter 18)
- Wu Zang meditation (Chapter 28)

CHOOSING PRESCRIPTIONS ACCORDING TO THE PATIENT'S QI CIRCULATION

When prescribing Medical Qigong prescriptions, the priority is to get the patient to move and circulate the Qi and Blood superficially and deeply within the body. According to Chinese physician, Professor Yan De-xin, the motion and quality of Blood flow in young and middle-aged patients is different from that in the elderly, which should be kept in mind when prescribing exercises.

- When prescribing Medical Qigong therapy for young patients, use mostly Dynamic and some Quiescent Qigong regulations.
- When prescribing Medical Qigong exercises for both adolescent and middle aged patients, it is important to increase Yin Qi cultivation while maintaining and storing the patient's Yang Qi.
- When prescribing Medical Qigong for senior citizens, focus is placed on Quiescent Qigong, with a minor emphasis on Dynamic breathing exercises, depending on their age and constitution.
- When prescribing Medical Qigong therapy for patients with very weak health or constitution, the doctor should prescribe mostly Quiescent Qigong regulations to start with.
- When prescribing Medical Qigong therapy for senior citizens with a strong constitution and health, the doctor can prescribe Qigong regulations that increase Yin Qi cultivation while maintaining and storing the patient's Yang Qi.

MEDICAL QIGONG PRESCRIPTIONS FOR GENERAL SYNDROMES

The following are some examples of the proper application of Dynamic and Quiescent Qigong prescriptions as presented by Dr. He Si Hai of the Zhe Jiang Qigong Hospital in China:

1. For Excess and Heat Syndromes it is recommended to prescribe static (quiescent) Medical Qigong exercises. After the gradual decline of the patient's fever or disease, the Qigong exercises are gradually increased in numbers and duration, and the purging training method is added (provided they have not excessively consumed their constitution).
2. For Deficiency and Cold Syndromes it is recommended to prescribe Dynamic Medical Qigong exercises. Gradually increase the exercises to include methods of lifting the limbs. If however, the pathological state is more serious, then static exercises are prescribed first.
3. For Cerebral Hemorrhage, have the patient practice static exercises such as the Hibernation Breathing exercise (quiescent lying posture, see Chapter 15) while combining it with the Yongquan (Kd-1) purging exercise (once the patient has regained consciousness). Dynamic Medical Qigong exercises that use methods of passing, circulating, or distributing Qi through the head are forbidden.
4. For Cerebral Embolism combine the training methods to include exercises that accelerate Blood flow and strengthen the function of the limbs as well as quiescent tonifying methods.
5. For Malignant Tumors (in a state of non-diffusional transference, in situ), combine an effective Dynamic exercise with a Quiescent exercise. Some examples of non-aggressive tumors / cancers are considered energetically warm and in a non-aggressive state of transition are:
   - Prostate cancer
   - Cervical dysplasia
   - Endometrial adenocarcinoma (uterine)
   - Basal cell skin cancer
   - Squamous cell skin cancer
6. For Malignant Tumors (in a state of diffusion, i.e., metastatic) prescribe Quiescent Medical Qigong exercises which are similar to the Hibernation Breathing exercise, in combination with strengthening exercises for patients with chronic diseases. Dynamic Medical Qigong exercises which accelerate the Blood flow are forbidden. Some examples of tumors that are considered energetically Hot and in an aggressive state of transition are:
   - Pancreatic cancer
   - Leukemia
   - Cervical cancer
   - Clear cell carcinoma (uterine)
   - Melanoma (skin cancer)

NUMBER OF BREATHS REQUIRED IN MEDICAL QIGONG PRESCRIPTIONS

Each Medical Qigong treatment will generally last up to three days. The Medical Qigong prescriptions that the doctors gives their patients will extend the treatment's effects indefinitely, providing the patients do all their homework.

The number of breaths required to Regulate the patient's Yin and Yang energy in a Medical Qigong prescription is often calculated in respiratory cycles, and on the basis of the number of the Yin and Yang aspects of diseases. This is often determined by the Yin and Yang energy numbers. Yang energy is represented by all the odd numbers (1, 3, 5, 7, 9, etc.). Yin energy is represented by all the even numbers (2, 4, 6, 8, etc.). These numbers are related to the Five Elements (Figure 38.5). For example: 6 is the number of the Kidneys, 7 the Heart, 8 the Liver, 9 the Lungs, and 10 the Spleen (Figure 38.6). The cycles of Yin and Yang numbers can be implemented through the utilization of the Magic Square (see Chapter 31).

The Yang number is used to calculate the respiratory cycles for Tonification and Purgation of the Yang channels; likewise, the Yin number is used for the Yin channels. This calculation should only be used as a general reference guide in treatment. Clinically, respiratory cycles should be determined based on the location and nature of the disease (if it is Yin or Yang, Exterior or Interior, Deficient or Excessive). The primary goal for using respiratory calculations is to ensure that sufficient Qi is administered when applying Medical Qigong treatments and therapeutic prescriptions. A disease due to a Deficiency of Yin, for example, should be treated with a tonification method for 18 or 64 respiratory cycles. To treat a Deficiency of Yang, a tonification method is used by the patient for 24 to 36 breaths.

	General Treatment	Moderate Treatment	Strong Treatment
Yin	6	18 or 36	64
Yang	9	27 or 49	81

Figure 38.5. Number of Yin and Yang Respirations

5 Yin Organs	Kidneys	Heart	Liver	Lungs	Spleen
Number of Respirations	6	7	8	9	10
Element	Water	Fire	Wood	Metal	Earth

Figure 38.6. Yin and Yang Five Element Respiration

CHAPTER 39
SPECIFIC ORGAN SELF-MASSAGE PRESCRIPTIONS

INTRODUCTION

The next few chapters of this book include some of China's most popular clinical therapies. It is the author's sincere desire to educate the reader as to the diversity, yet effectiveness of these prescriptions. It is important to remember that it is not necessarily the exercise or prescription's history that matters, what does matter is the initial goal of the prescription (whether to purge, tonify, or regulate) and the correct posture, breath, and mental focus.

This section begins with energy massage techniques, which are prescriptions that require the patient to regenerate his or her own body energy for self-healing, cultivation, and circulation. These exercises are designed to purge, tonify and regulate the Five Yin Organ System. They are not arranged in any particular order, and can be singled out by the doctor and given to patients for specific ailments. Each massage prescription is named after the organ it stimulates. For clinical therapy, perform these exercises three to five times a day.

HEART MASSAGE

The doctor can chose to give the patient either the One Through Four meditation (1-fun, 2-shoe, 3-tree, 4-core; from Chapter 26), or the Stacking the Bones meditation (described below) to prepare for the following meditations.

1. Begin in a Wuji posture to perform the Stacking of the Bones meditation. Imagine the energy in your body sinking and melting into the ground; the bones vertically stack, one by one, on top of each other. Each of the body's articulations stacks on top of the next section of bone, from the bottom of the feet, ankles, knees, etc., to the top of the head.

   The spine is kept straight but relaxed, the anal sphincter is closed, the tongue is touching the top of the upper palate, the Mingmen is slightly pressing towards the back, and the body is completely relaxed.
2. Perform the Heart Regulation. When prescribing these exercises for male patients, have them place the left hand over the Heart with the right hand on top of the left hand (for women, the hand positions are opposite, and the beginning placement of the hands is below the left breast). Massage in twelve circular rotations clockwise to the left, then twelve rotations counterclockwise. The hand can either lightly touch the skin (using Qi massage), or be slightly off the skin, up to several inches away from the patient's body (using Qi Distance therapy), depending on the degree of energetic sensitivity and projection skill (Figure 39.1). Focus the mind's intention on the Heart organ, allowing the energy within the Heart area to flow and circulate with the movement of the hands.

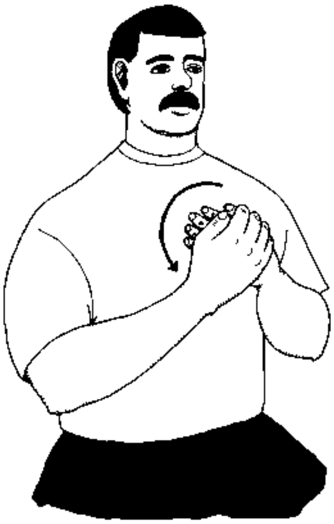
Figure 39.1. The patient uses the mind's intent to regulate the Qi of the Heart.

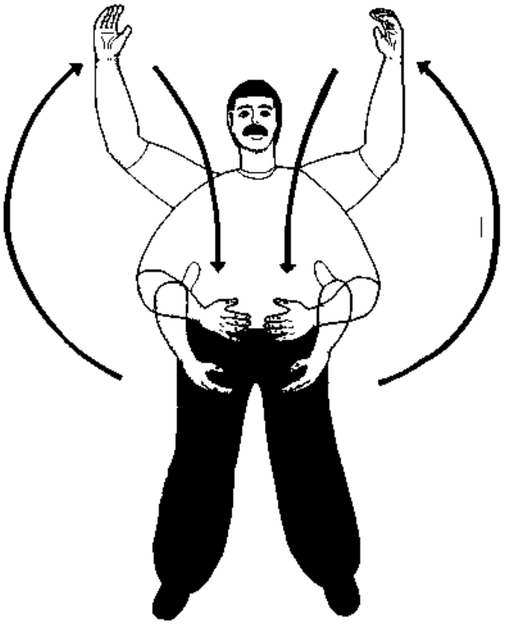

Figure 39.2. Pulling Down the Heavens allows the patient to sink the Qi, returning the excess energy back into the Lower Dantian.

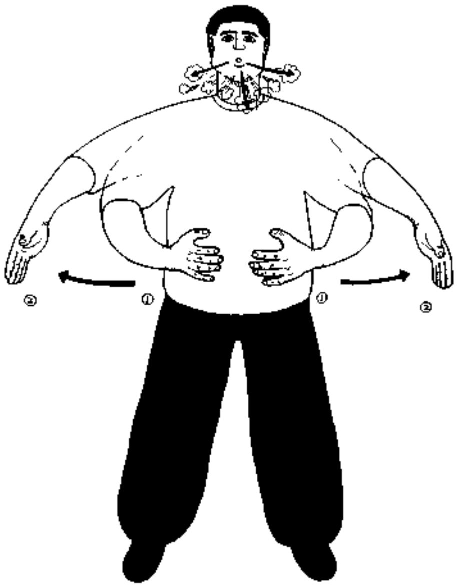

Figure 39.3. Exhaling and Opening the Lower Dantian allows the body to expel the Evil Qi from the internal organs. Expand the abdomen.

The purpose of the rotating action of the Heart's Qi is for promoting Blood circulation and dispersing Blood Stasis, while dredging the channels and collaterals.

3. Perform the Point Respiration. This technique focuses breathing and imagination into the channel points for the specific treatment area. For this particular exercise, the patient should leave the hands over the Heart, while concentrating the focused intention and breath deep into the tissue of the organ. The patient should then exhale while lightly squeezing and pressing the area of the Heart, and imagining divine healing light radiating outward from the tissues. While inhaling, the patient should lift the hands away from the body and imagine divine healing light flowing into the Heart organ area. This action is similar to that of a bellows. This exercise should be performed twelve times, then lead the Qi down into the Lower Dantian.
4. Perform the Pulling Down the Heavens method. Inhale, pause, then exhale while slowly bringing the arms down the center line of the torso, while quietly drawing the energy back down into the Lower Dantian. When lowering the arms, imagine the energy of the Heart like water, melting and flowing down the torso.

   Pause for a breath at the Lower Dantian, rooting the energy. Then slowly raise the hands above the head, while inhaling. Exhale slowly and imagine the energy in the arms and head melting down the body, beginning at the top of the head and tips of the fingers, flowing down and into the Lower Dantian.

   This action is important, because if the Qi is not brought back into the Lower Dantian after each exercise, the energy will be easily dispersed (Figure 39.2).
5. Perform the Opening and Closing of the Lower Dantian. Finally exhale and imagine the Evil Qi being expelled through the mouth. As you exhale, separate both of the hands away from the abdomen (as you contract the abdomen). This is called the Opening (Figure 39.3).

   After completing the Opening of the Lower

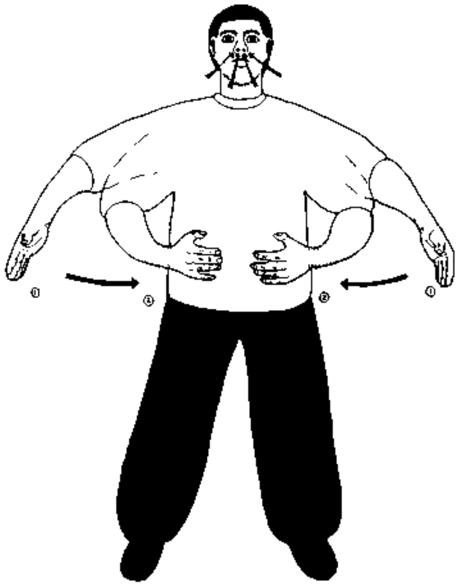

Figure 39.4. Inhaling and Closing the Lower Dantian allows the body to seal itself from outside pathogenic factors. Contract the abdomen.

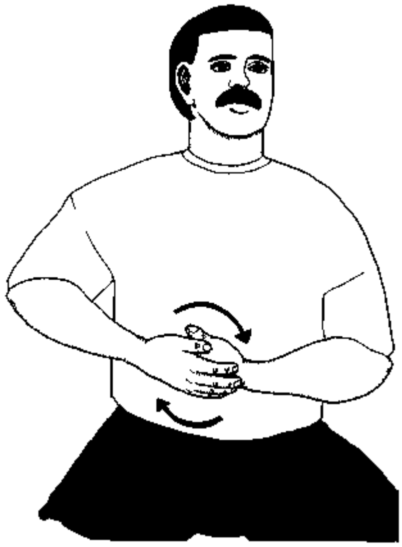

Figure 39.5. The patient will focus their mind's intent on the Liver area while massaging, allowing their energy to circulate into the organ's tissues.

Dantian, inhale through the nose while bringing both hands back to face the Lower Dantian (as you expand the abdomen). This action is called the Closing. When Closing, imagine sealing the body in a protective energetic field, so that the pathogenic factors cannot get inside (Figure 39.4). Perform the Opening and Closing method three times.

Note: The techniques 1 through 5 are one complete set. Perform a total of three sets. Three sets are necessary because the body will naturally progress through three stages. The first time the body is introduced to a new pattern it will shock the system. The second time the body begins to understand the pattern. The third and final time the body recognizes and accepts the changes the patient is attempting to reprogram.

LIVER MASSAGE

1. Begin from a Wuji posture and perform either the One Through Four or Stacking the Bones method as described in Heart Massage (1).
2. Perform the Liver Regulation. Place the left hand on the Liver with the right hand on top (opposite for women). Massage in twelve circular rotations to the left in a clockwise direction, then twelve rotations in the opposite direction. When massaging, focus the mind on the Liver. Inhale and imagine divine healing light coming down from the Heavens and filling the Liver. Exhale and purge the Liver of Turbid Qi through the mouth.

   The purpose of this exercise is to smooth and regulate the activities of the vital energy flowing inside the Liver, encouraging the hyperactive and Rebellious Liver Qi to flow downward (Figure 39.5).
3. Perform Point Respiration with both hands still over the Liver, for twelve breaths, as described in the Heart Massage (3).
4. Perform Pulling Down the Heavens by leading Qi down into the Lower Dantian, as described in the Heart Massage (4).
5. End with Opening and Closing method as described in the Heart Massage (5).

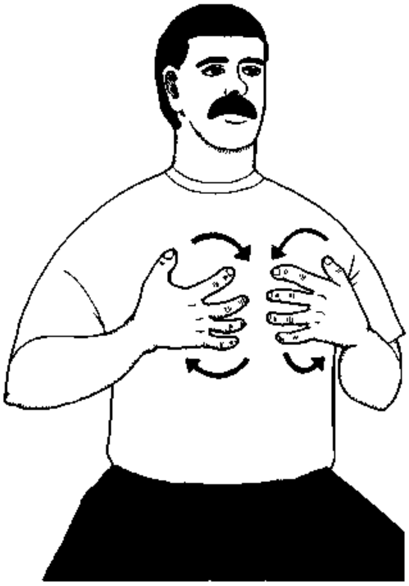

Figure 39.6. The patient focuses the mind's intent on the Lung area while massaging, allowing the energy to circulate into the organ's tissues.

Note: The techniques 1 through 5 form one complete set. Perform two more sets for a total of three to complete the prescription.

Lung Massage

1. Begin from a Wuji Posture and perform either the One Through Four or Stacking the Bones method as described in the Heart Massage (1).
2. Perform the Lung Regulation. Place both palms on the Lungs (specifically on an injured or diseased region, if known). Massage in twelve circular rotations from the bottom, up the center to the top and then to the outside of the chest. Then massage twelve rotations in the opposite direction, from the top, down the center and to the outside. When massaging, concentrate on the Lung region. Inhale and imagine divine healing light coming down from the Heavens and filling the Lungs. Exhale and purge the Lungs of Turbid Qi.

   The purpose of this exercise is to bring Blood and Qi into the Lung area, clean and purify the inhaled breath and support the Lungs' Qi flowing downward (Figure 39.6). TB patients with pulmonary holes should concentrate on abundantly filling and expanding the vital energy in the Lungs while visualizing the Qi activating the body's cells to close the holes. If the patient is unable to visualize this, or the reaction is too uncomfortable, have the patient mentally pull the Qi downward.
3. Perform the Point Respiration with both hands still over the Lungs, for twelve breaths, as described in the Heart Massage (3).
4. Perform Pulling Down the Heavens method, leading the Qi back down into the Lower Dantian as described in the Heart Massage (4).
5. End with Opening and Closing method as described in the Heart Massage (5).

Note: The techniques 1 through 5 are one complete set. Perform two more sets for a total of three to complete the prescription.

Spleen & Stomach Massage

Spleen Massage

1. Begin from a Wuji Posture and perform either the One Through Four or Stacking the Bones method as described in the Heart Massage (1).
2. Perform the Spleen Regulation. Place the left hand on the Spleen with the right hand on top (opposite for women). Massage in twelve circular rotations to the left in a clockwise direction, then twelve rotations in the opposite direction. When massaging, concentrate the mind on the energy within the Spleen area, imagining the Spleen Qi movement flowing with the movements of the hands (Figure 39.7). Inhale and imagine divine healing light coming down from the Heavens and filling the Spleen. Exhale and imagine the Spleen absorbing the divine Qi.

   The purpose of this exercise is to strengthen and increase the Spleen's function of transporting and converting Qi. This is an excellent massage for patients with splenomegaly (enlargement of the spleen) and dyspepsia (painful digestion).
3. Perform the Point Respiration with both hands still over the Spleen, for twelve breaths,

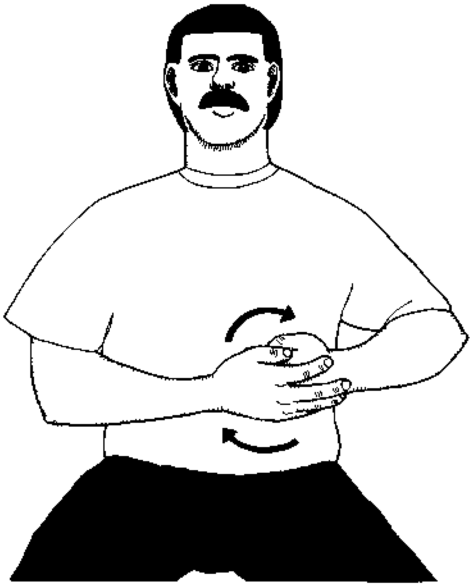

Figure 39.7. The patient focuses the mind's intent on the Spleen area while massaging, allowing the energy to circulate into the organ's tissues.

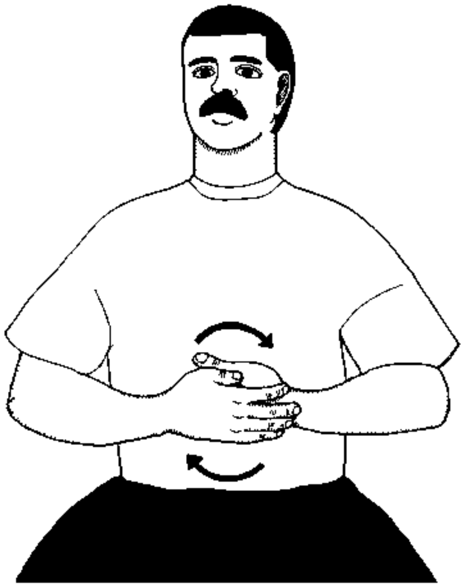

Figure 39.8. The patient will focus their mind's intent on the Stomach area while massaging, allowing their energy to circulate into the organ's tissues.

as described in the Heart Massage (3). Then move both hands to the Stomach area.

STOMACH MASSAGE

4. Perform the Stomach Regulation. Place the left hand on the Zhongwan CV-12 point at the middle of the Stomach with the right hand on top (opposite for women). Massage in twelve circular rotations to the left in a clockwise direction, then twelve rotations in the opposite direction. When massaging, concentrate on warming the Stomach area (Figure 39.8). Inhale and imagine divine healing light coming down from the Heavens and filling the Stomach. Exhale and purge the Stomach of Turbid Qi.

   This is an excellent massage for patients with thoracic depression, abdominal distention, gastritis, and gastric ulcers.
5. Perform the Point Respiration with both hands still over the Stomach, for twelve breaths, as described in the Heart Massage (3).
6. Perform the Pulling Down the Heavens method to draw the Qi down into the Lower Dantian, as described in the Heart Massage (4).
7. End with the Opening and Closing method as described in the Heart Massage (5).

Note: The techniques 1 through 7 are one complete set. Perform two more sets for a total of three to complete the prescription.

KIDNEY MASSAGE

This exercise is excellent for tonifying the Kidneys, as well as healing impotence, premature ejaculation, lumbago, lower back problems, and pelvic inflammatory disease (for PID use reverse breathing with static Qigong training).

This is an extremely important exercise for patients and for Qigong doctors to practice; it enables them to increase their Qi projection when treating patients.

1. Begin from a Wuji Posture and perform either the One Through Four or Stacking the Bones method as described in the Heart Massage (1).
2. Perform the Kidney Regulation. Close the eyes and place both hands on the back (waist

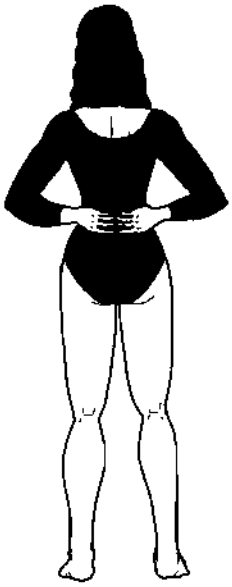

Figure 39.9. The patient will focus the mind's intent on the Kidney area while massaging, allowing the energy to circulate into and warm the organ's tissues.

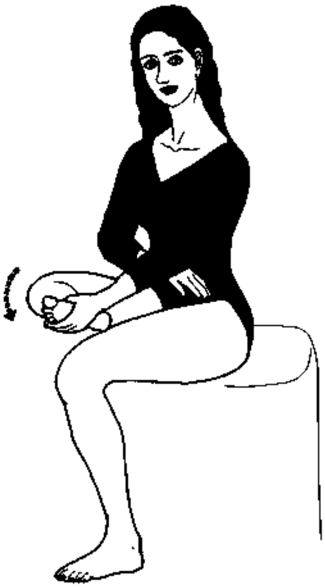

Figure 39.10. The patient will focus the mind's intent on the bottom of the foot while massaging, allowing the energy to flow through the Kidney Channel into the Lower Dantian.

level), middle fingers touching the Shen Shu Bl-23 points. Focus on the Kidneys. Begin massaging twenty-four times inward, then twenty-four times outward, until the area becomes warm (Figure 39.9).

3. Perform the Point Respiration with both hands still over the Shen Shu Bl-23 points for thirty-six breaths. Exhale as you press the palms into the Kidneys, inhale as you raise the palms from the back. While inhaling, imagine divine healing light coming down from the Heavens and filling the Kidneys completely. While exhaling, imagine the Kidneys absorbing the divine Qi.
4. Perform the Pulling Down the Heavens method, drawing Qi into the Lower Dantian as described in the Heart Massage (4).
5. Perform the Opening and Closing method as described in the Heart Massage (5).

Note: The techniques 1 through 5 are one complete set. Perform two more sets for a total of three to complete the prescription, then proceed to the Closure.

6. Perform the Closure. After completing three sets of the Kidney exercise, place the right hand on the navel and the left hand on the Mingmen. Focus on gathering Heat and Qi in the Lower Dantian. Next, place the right foot on the left knee (opposite for women). With the right hand still on the Lower Dantian, place the left hand at the bottom of the right foot and gently massage the Yongquan Kd-1 point. Massage this area 100 times in a clockwise direction, then 100 times counterclockwise. When massaging the Kd-1 point, work slowly, focusing the mind's intention through the center of the palm into the center of the foot, imagining the Qi flowing up the leg into the Lower Dantian (Figure 39.10).

Note: Some Medical Qigong Clinics in China perform this same exercise while slapping the bottom of the feet (instead of using circular massage). They begin this treatment with 100 strikes on each foot for the first day, increasing an additional 100 strikes each consecutive day (not exceeding 300 strikes) until the desired result is achieved.

7. Perform the Point Respiration for twelve breaths, with the right hand still on the Lower Dantian and the left hand still on the bottom of the right foot, as described in the Heart Massage (3).
8. Switch feet, placing the left hand on the Lower Dantian while the right hand gently massages the Kd-1 at the bottom of the left foot, 100 times clockwise then 100 times counterclockwise.
9. Perform the Point Respiration for twelve breaths, with the left hand still on the Lower Dantian and the right hand still on the bottom of the left foot.
10. End with the Opening and Closing method as described in the Heart Massage (5).

Note: Do not use this ending method (5 and 10) for patients who suffer from insomnia, as it will have an awakening effect as it invigorates the vital energy. When performed at bedtime it can be used for treating insomnia by following the simple prescription below.

TREATMENT FOR INSOMNIA

This is a very famous prescription for insomnia. I have used it in my clinical practise for many years and have found it to be extremely effective. If the insomnia is due to a Deficiency of Kidneys' Yin, or the patient is restless during the night due to the Liver's Heat rising and causing the patient's mind to wander, this prescription will help. After clinically treating patients for insomnia, the Qigong doctor can prescribe the following exercise to be performed by the patient.

This exercise is to be practiced just before the patient is going to sleep. Begin by sitting at the edge of the bed, repeating the previous Kidney exercises 2, 3, and 6.

1. Perform the Kidney Regulation. Close the eyes and place both hands on the back (waist level), middle fingers touching the Shen Shu Bl-23 points. Focus on the Kidneys. Begin massaging twenty-four times inward, then twenty-four times outward, until the area becomes warm (Figure 39.9).
2. Perform the Point Respiration with both hands still over the Shen Shu Bl-23 points for thirty-six breaths. Exhale as you press the palms into the Kidneys, inhale as you raise the palms from the back. While inhaling, imagine divine healing light coming down from the Heavens and filling the Kidneys completely. While exhaling, imagine the Kidneys absorbing the divine Qi.
3. Perform the Closure. After completing three sets of the Kidney exercise, place the right hand on the navel and the left hand on the Mingmen. Focus on gathering Heat and Qi in the Lower Dantian. Next, place the right foot on the left knee (opposite for women). With the right hand still on the Lower Dantian, place the left hand at the bottom of the right foot and gently massage the Yongquan Kd-1 point. Massage this area 100 times in a clockwise direction, then 100 times counterclockwise. When massaging the Kd-1 point, work slowly, focusing the mind's intention from the Lower Dantian through the leg into the center of the foot, imagining the Qi flowing down the leg into the bottom of the foot (Figure 39.10). As the mind begins to relax, close the eyes and fall sleep. If you become fatigued, do not complete the exercise.

An explanation for the reason for this exercises success is as follows:

1. The Kidney exercise 1 and 2 will root the body's Blood, Heat, and Heart Qi back into the Kidneys and Lower Dantian, where they will become stabilized.
2. The Kidney exercise 3 will draw the body's Heat, Blood, and the Heart Qi away from the torso and Lower Dantian into the extremities of the feet, quieting the mind and causing sleep.

CHAPTER 40
HEALING SOUND THERAPY AND PRESCRIPTIONS

INTRODUCTION AND HISTORY

As early as the Qin Dynasty (221-207 BC) healing sounds have been recorded in China. For thousands of years, Qigong masters have studied occurrences in nature and in the physical body to develop a holistic approach to health and longevity. One of the many discoveries made is that all humans produce similar sound patterns in certain situations. For instance, after a tense situation many people utter a sigh of relief. This sigh is the body's way of releasing emotional stress. Specific situations will release different emotional sounds, because emotions are stored as tone resonances in the different chambers of the Five Yin Organs.

During the Sui Dynasty (590-618 AD), ancient Qigong Master Zhi Zhuan pointed out that there are six ways of expelling toxic energy from the internal organs, using the six different sounds formula. The silently spoken six sounds are coordinated with the breathing and the movements of the body's extremities to purge and cleanse the organ cavities and particular channels of stagnant Qi. These exercises are excellent for treating muscular injuries and for the Yin and Yang organs. The impure Qi produced by external pathogenic factors (from food, air, and drink) and internal pathogenic factors (toxic emotions and negative mental attitudes) will be expelled directly from each organ.

The healing sound techniques can either be used by the patient as a form of static Medical Qigong prescription (for purgation), or projected into the patient by the doctor as a clinical form of energetic stimulation (for tonification or purgation). The six words are: Xu (Shu), Ke (Haa), Hu (Who), Si (Sss), Chui (Chree), and Xi (Shee). These sounds are also used in chanting (in spiritual Qigong). Ancient documents state that these six sounds denote the Five Yin Organs as well as the five seasons.

It is important to note that there are several healing sound systems used by different Medical Qigong schools and other cultures. These sound systems all affect the body's sympathetic and parasympathetic nerves to some degree. The main emphasis, when performing any healing sound, should be placed on the connection of the mind, breath, and imagination to the area or internal organ focused on by the patient. This will reduce the potential for causing Qi deviations. The patient must have an internal image and actually hear and feel the resonating sound inside the body. It is also important that the patient feel enveloped in the sound, vibration, energy, and light of the therapeutic tone.

SOUND VOLUME RESONATION

The pronunciations of the sounds must be correct. The resonant effects of the sounds must be brought into full play to be successfully transmitted to the diseased areas of the internal organs. Therefore, it is extremely important that the patient focus the mind and intention on the area being treated. There are three volume resonations commonly used in Medical Qigong sound therapy: soft, moderate, and loud.

1. Soft volume is spiritual in its energetic nature and is related to the projection of Shen. A soft volume is generally used when beginning sound therapy treatments. With soft volume pronunciation, the patient can easily feel where the sound waves reach and acquire resonant effect, but the amplitude of the sound wave is not wide enough to produce the desired healing effects, and its vibrational force is very weak.
2. A moderate volume of sound is emotional and mental in its energetic nature and is related to the projection of Qi. A moderate volume is

considered the standard for all healing sound treatments. Moderate volume is decided by the individual's own voice. It should be limited to the range in which the patient's voice does not strain and in which the voice can be raised or lowered effortlessly.

3. Loud Volume is related to the projection of Jing. A loud volume is used to shock the patient into feeling the diseased tissue area. This is due to the fact that loud volume is more tangible (or physical) in its energetic nature. With loud volume pronunciation, the patient can easily feel the body's internal organs and tissues vibrate; however, this sound is only used to reintroduce a desensitized patient to his or her body. The problem with prolonged use of loud volume is that the increased volume is liable to create tension and cause Qi deviations.

The volume of sound uttered also depends on the patient's current state of health. People who are physically weak can usually practice the breath regulating and Qi reinforcing exercises, but should not practice the sound pronouncing method until they feel appropriately energetic.

To practice the word pronouncing method, the patient should start from the soft volume and increase it gradually. Diseases of different parts of the body require different volumes. For example:

- Patients with Heart or Lung trouble can comfortably take the soft to moderate volume pronunciation;
- Patients with Liver, Spleen, Stomach, Kidney and intestinal diseases can take the moderate to louder volume pronunciation.

HIGH-, MIDDLE-, AND LOW-PITCHED SOUND

The healing sounds have high, middle, and low pitches with different penetrating potentials. When correctly vocalized, they cause resonant vibrations in the body, producing certain states of consciousness that have both a psychological and a physiological effect on the nervous system.

As the tones change from a high to low pitch, they regulate the depth of the tone penetration. Similar to a dimmer switch on a light bulb, the tone adjusts the depth of the vibration resonance in the body.

The three pitch sounds also affect the energy movement of the body's Triple Burners, as well as the internal organs positioned within the thorax.

1. The high-pitched sound is pronounced in a straight tone with the head slightly drooped. The high pitch focuses the energy resonation into the specific organs of the patient's Upper Burner from the sternum to the manubrium. After completing the sound resonation exercise, return back to the Wuji position and relax; regulate the body using natural breathing.
2. The middle-pitched sound is pronounced in either a rising or descending tone, depending on the patient's specific conditions. The rising tone will cause the patient's Qi to ascend, while the descending tone will cause the Qi to drop. The middle pitch focuses the energy resonation into the specific organs of the patient's Middle Burner from the navel to the sternum within patient's body. Middle-pitched sounds are only prescribed for specific conditions (organ prolapse or organ energetic overload).
3. The low-pitched sound is pronounced in a dropping, then rising tone, within one breath. This causes the energetic resonation to saturate the diseased organ from top to bottom, then from bottom to top. This focuses the energy resonation into the specific organs of the patient's Lower Burner from the navel to the perineum. The method of pronouncing the low-pitched sound is usually practiced immediately after sounding the high-pitched tone.

After pronouncing either the high-, middle- or low-pitched sounds, perform the Open and Close the Three Dantians method once, then close with the Pulling Down the Heavens exercise. This completes one set.

BREATH AND MIND CONTROL

Healing sounds produce an aspirating effect analogous to vacuuming a room. This analogy describes to the patient the desired intention needed to successfully utilize this modality of healing. Toning without proper focused intention, breath, and mind control placed on the specific internal organ is like trying to vacuum a room while waving the hose in the air. Although the hose is sucking the air, it is not performing the desired result of purging the room of dust and dirt. To become proficient and establish the desired result, the individual must focus the attention and place the vacuum hose onto the specific area that needs cleaning.

Healing sound and toning therapy work the same way. It is not just sounding the tone which allows the technique to work, but the combination of the breath and mind focused on the intended internal organ. When practicing sound therapy, the patient generally uses the Natural Breathing method, along with focused intention and color visualization.

While inhaling, the patient imagines breathing in divine, white light healing energy through the nose and absorbing this healing Qi into the diseased area, vibrating it.

When exhaling, the patient imagines breathing out the dark, diseased, Turbid Qi through the mouth. It is important for patients to maintain the healing color (i.e.: green, red, golden yellow, white, or blue) vibration inside the diseased organ, even when they exhale.

SEPARATING THE BREATH INTO YIN AND YANG

The breath can be divided into either Yin or Yang qualities. When breathing shallow and focusing the mind on the front of the throat, the breath feels cool and is therefore considered Yin; however, when shifting the mind's attention towards the back of the throat, the breath, when released, feels hot, and is considered Yang. This change in temperature is due to the front of the body being Yin in nature and the back of the body being Yang.

AUDIBLE AND INAUDIBLE SOUND RESONATION

Before projecting sound therapy in a clinical setting, the Qigong doctor must vibrate the three Dantians while focusing on his or her Taiji Pole. Once the sound vibration has filled the doctor's body, the resonate healing sounds can be projected into the patient's tissues and organs.

Another way to produce sound is to imagine that the sound is appearing as a specific color and causing a vibration within the body. Clinical sound therapy is further divided into three distinct levels of emitted energies: Obvious Sound Resonation, Hidden Sound Resonation and Mysterious Sound Resonation.

1. The Obvious (Ming) Sound Resonation is audible sound. This energy sound resonation is considered Yang and is produced through the vocal organs. When toning audibly, the healing sounds fills the Qigong doctor's energetic fields. These fields can then be transferred onto the patient to produce a very strong, physical affect on the patient's tissues and organs. The Qigong doctor connects (through intention) with the patient's internal organ, and imagines his or her voice penetrating the diseased organ's tissues. The doctor then imagines the patient's tissues vibrating like rippling water with the sound exhalation, to purge the pathogenic Qi.
2. The Hidden (An) Sound Resonation is slightly audible sound. This energy sound is considered Yin transitioning into Yang, and is produced by softly exhaling sound energy resonation into the patient's body. This slightly audible energy sound resonation affects the patient's tissues and organs more on an emotional plane, enveloping and penetrating the patient's tissues like a mist.
3. The Mysterious (Hua) Sound Resonation is inaudible sound. This energy sound resonation is considered Yin and is mentally produced inside of the doctor's body. The Qigong doctor first focuses and fills his or her body and energetic field with the vibrant sound. The doctor then focuses the mind on the healing sound like a mantra. This inner sound fills the Qigong doctor's field with resonant vibration which can be transferred into the patient's body. This energy sound resonation affects the patient's tissues and organs on a more spiritual dimension, enveloping and purging like a gentle light.

The Qigong doctor can also coordinate the projection of the healing sounds with the patient's toning to synchronize their energetic fields. All sound techniques have specific postures, breathing, and mind regulations to facilitate whole body integration.

Patients and students are usually encouraged to practice the audible sound vibrations until they can vibrate the tissues of the organs at will, before they proceed to the inaudible sound resonation.

THE SIX HEALING SOUND PRESCRIPTIONS

Before starting any organ sound exercise, there are certain grounding and cleansing techniques that the patient should practice. These preparation exercises are described as follows.

1. Always begin sound resonation from a Wuji standing posture.
2. Breathe naturally from the abdomen, inhaling through the nose and exhaling through the mouth. The body should be relaxed, and respiration should be gentle, thin, even, and extended. The mind should not concentrate on the expansion of the Lungs, but on the Lower Dantian first, and then on the treatment area. Also, breath through every pore of the body into the Lower Dantian.

   Upon exhaling, sound the healing word with a low volume pronunciation, feeling where the sound waves reach. Once the tissue area is stimulated, increase the volume. At the same time draw in the abdomen, draw up the anus, slightly knead the ground with the feet, sink the shoulders, and then shift the body weight to both heels. This will access the Kidneys' Water Qi. When inhaling, close the mouth, raise the tongue against the hard palate, and shift the body weight to the anterior soles of the feet. This will access the body's Fire energy. Exhale the specific sounds through the mouth while imagining the impurities of that specific organ being expelled.
3. Pull Down the Heavens (the ending breath regulation). Inhale and exhale through the nose. When inhaling, slowly raise both arms forward and upward from the sides of the body, with both palms facing downward until reaching shoulder level. Then turn the palms up, and continue to circle the arms over the head. When exhaling, bend the elbows to make a curve in front of the chest, and press the abdomen. Then separate the arms to the sides of the legs, and return to the Wuji posture (Figure 40.1).

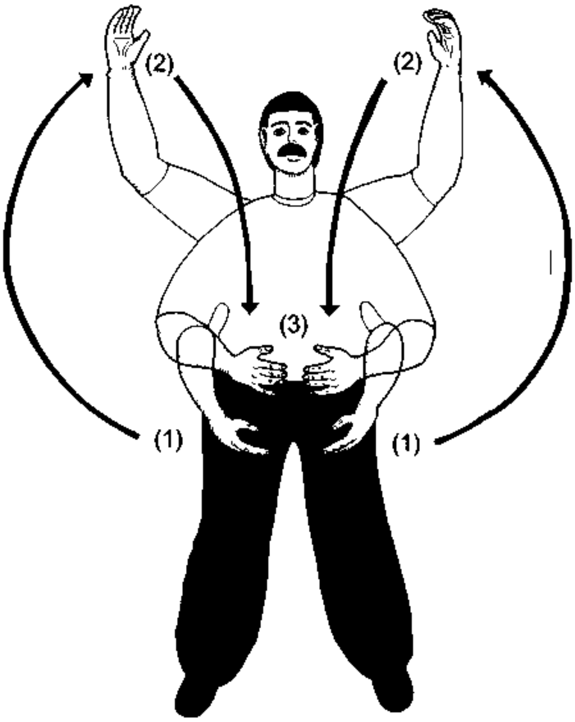

Figure 40.1. Pulling Down the Heavens allows the patient to sink the Qi, returning the excess energy back into the Lower Dantian.

Regulate the breath three times in the Wuji posture after completing the exercise. Breathing Regulation (Pulling Down the Heavens) is required after the exercise of each sound is completed and before the next exercise begins.

THE SOUND "XU" (SHU) NOURISHES THE LIVER

This sound relieves Liver stagnation, aids in correcting certain types of visual distortions, and purges the accumulation of Phlegm in the body (e.g., the condition of Lung Hot Phlegm, due to an already existing cold).

Both hands overlap the inner Laogong Pc-8 points, with the left hand on the outside for men (opposite hand position for women). The Shaoshang Lu-11 point of both thumbs presses against the navel, with the Laogong points pointing at the umbilicus.

When inhaling, expand the abdomen and close the eyes, feeling the breath envelop and penetrate the Liver organ. At the same time direct the Qi of the Liver's Channels to flow upward from the Dadun Lv-1 point on the lateral portions of the big toes. Visualize the Qi rising through the lower abdomen, linking the Gall Bladder to the Lungs, traveling to the throat, and eyes, to the forehead, and then to the Baihui GV-20 point.

When exhaling, contract the abdomen and release the pathogenic factors. Focus the attention and imagination on the Qi flowing down the face, to connect with the Lungs' Channels, flowing downward to the Shaoshang Lu-11 points on the medial portions of the thumbs. The Qi is then directed and rooted into the Lower Dantian. Open the eyes wide, and sound the word "Shu" (the lips are slightly brought together, the tongue tip stretches forward with the lateral sides slightly curling toward the middle). Exhale completely.

In this particular exercise, the breath should feel lukewarm as you exhale (Figure 40.2). Perform six times.

THE SOUND "KE" (HA) REPLENISHES THE HEART

This sound relieves Fire from the Heart and expels Heat from the body; it also improves poor circulation and reduces fevers. The respiration patterns are the same as those in the previous (Liver) sound prescription.

When inhaling, raise both hands to shoulder level as if embracing a post. Both palms should face each other with the arms making a curve in front of the chest. Expand the abdomen and close the eyes; feel the breath envelop and penetrate the Heart organ. At the same time direct the Qi of the Spleen's Channels to flow from the Yinbai Sp-1 point on the medial portion of the big toes, along the inner portions of the legs, upward to the abdomen to meet with the Thrusting Channels, and enter into the Heart.

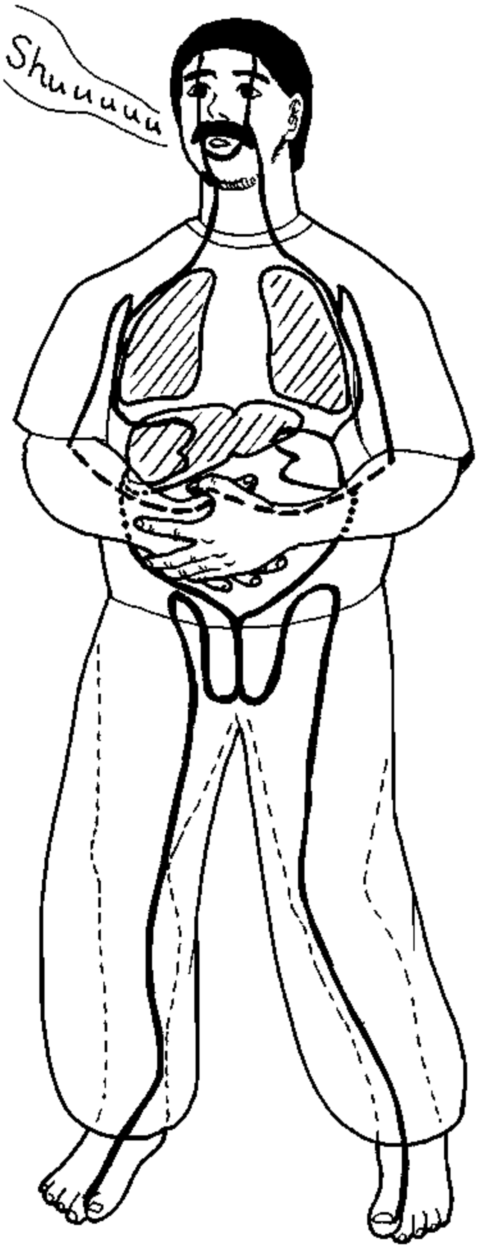

Figure 40.2. The sound "Xu" (Shu) is used to purge Liver stagnation as well as to disperse the accumulation of Phlegm in the body (Lung Hot Phlegm).

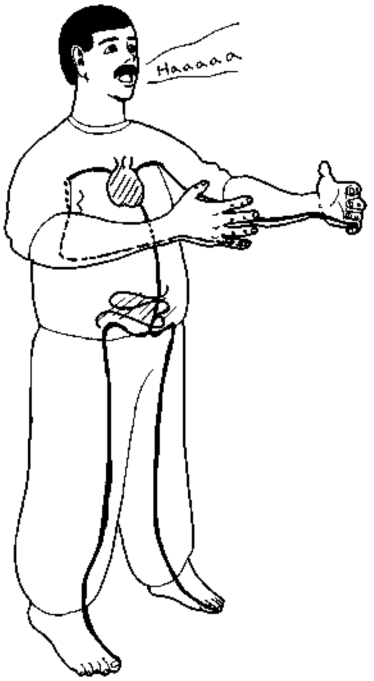

Figure 40.3. The sound "Ke" (Ha) is used to purge Heart Fire as well as expel Heat from the body.

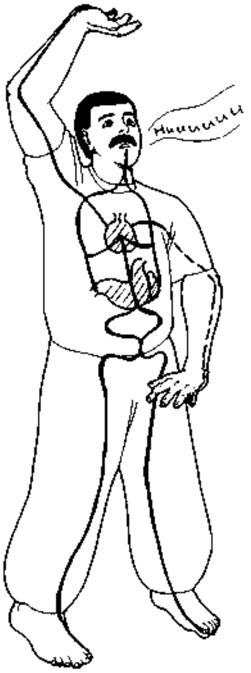

Figure 40.4. The sound "Hu" (Who) is used to treat digestive problems, by purging Stomach and Intestinal stagnation.

When exhaling, softly open the eyes and draw the Qi out from the Heart, past the Lungs, through the armpits, along the inner portions of the arms, to the Shaochong Ht-9 point on the medial portion of the little fingers. Point the fingers forward to release the pathogenic Qi, and sound the word "Ha." The mouth is half closed with the tongue placed against the lower mandible. Exhale completely. This breath should feel hot (Figure 40.3).

Return to the beginning position and perform six times.

THE SOUND "HU" (WHO) STRENGTHENS THE SPLEEN

This sound is used to correct digestive problems, especially when the Stomach or intestines feel full or sluggish.

When inhaling, raise both hands upward in front of the abdomen, as if embracing a ball in the Lower Dantian area. Both palms should face each other with the arms making a curve. Expand the abdomen and close the eyes; feel the breath envelop and penetrate the Spleen organ. Direct the Qi of the Spleen's Channels to flow from the Yinbai Sp-1 points on the medial portion of the big toes, up to the abdomen to enter the Spleen, and to link up with the Stomach.

When exhaling, the right palm turns outward, rolling over and pushing above the head, thus forming the posture of Holding the Heavens. At the same time, imagine the Qi ascending up to the throat, to the root of the tongue, and spreading under it. Imagine a secondary river of Qi flowing from the Stomach to the Heart, and ascending along the Heart's Channels to the Shaochong Ht-9 point on the medial portion of the little fingers, and exhale the sound "Who." The lips are brought together as if playing a flute, the tongue is level and stretched forward as much as possible to allow the ascending energy of the Thrusting Channels to rush out. Exhale completely while the left palm turns downward and presses down to the outside of the left hip (Figure 40.4).

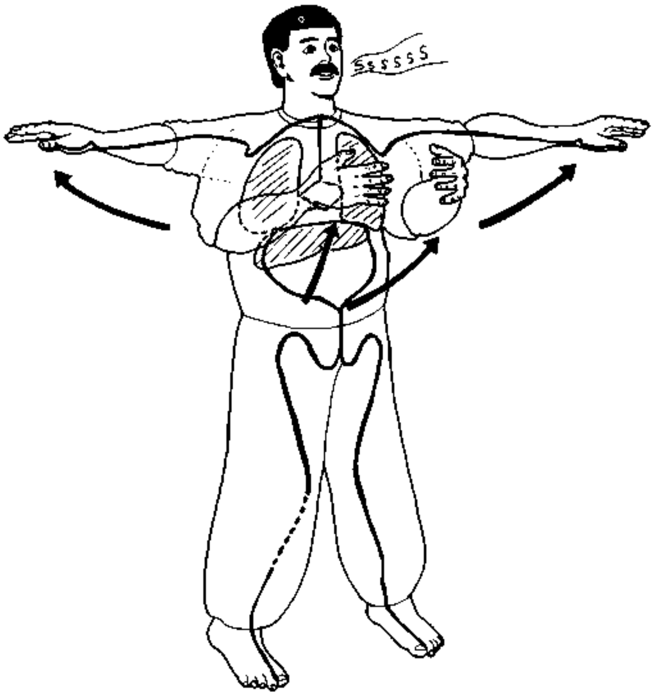

Figure 40.5. The sound "Si" (Sss) is used in order to Purge the Lungs.

Next, inhale and turn the right palm over to face the body; slowly descend the right arm in front of the body, while the left arm turns towards the body and ascends. The two hands cross in front of each other (with the ascending arm on the inside, and the descending arm outside). Now the left hand pushes up and the right hand pushes down. At the same time, exhale and repeat the sound "Who." Pressing both left and right hands form one set. Perform six times. At the end, cross both arms in front of the chest and return to the beginning position.

THE SOUND "SI" (SSSS) MOISTENS THE LUNGS

This sound is used to nourish the Lungs, treat problems such as tuberculosis, abscesses, and dermatitis.

From the beginning Wuji posture, inhale and raise both hands as if holding a ball, moving it upwards, until the hands face the Shanzhong CV-17 point at the front of the chest. Direct the Qi of the Liver's Channels to ascend from the Dadun Lv-1 point on the lateral portions of the big toes, along the inner portion of the legs, through the lower abdomen into the Lungs.

When exhaling, turn over both palms to face downward, then separate the palms horizontally to the sides. Guide the Qi along the Lungs' Channels to the Shaoshang Lu-11 points on the inner portion of the thumbs, while exhaling the sound "Sss" The lips are slightly drawn back, the upper and lower teeth are slightly brought together forming a small slit; the tip of the tongue is placed against the slit in-between the teeth to articulate the sound. Exhale completely, then allow both arms to descend to the sides of the body (Figure 40.5). Repeat six times.

THE SOUND "CHUI" (CHREE) STRENGTHENS THE KIDNEYS

This sound is used to regulate the Kidneys, expel chills, and to maintain the Internal Heat of the Lower Dantian and Mingmen Fire.

Inhale and raise both arms to embrace an imaginary ball in front of the chest (Middle Dantian area).

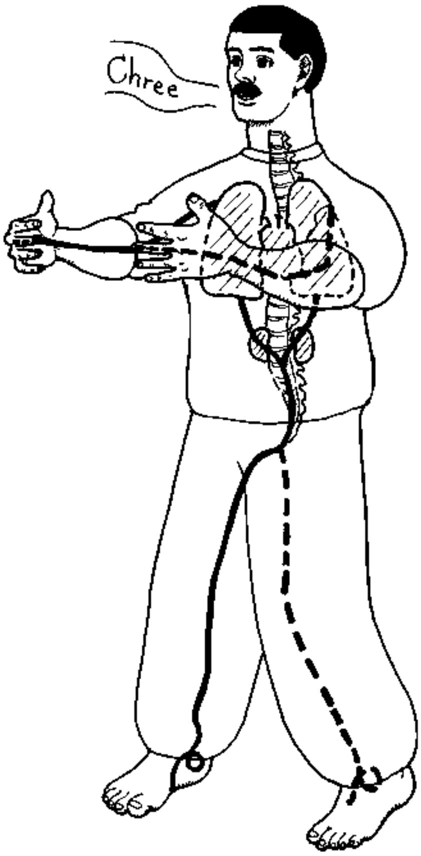

Figure 40.6. The sound "Chui" (Chree) is used to regulate the Kidneys and the Mingmen Fire, as well as to expel chills from the body.

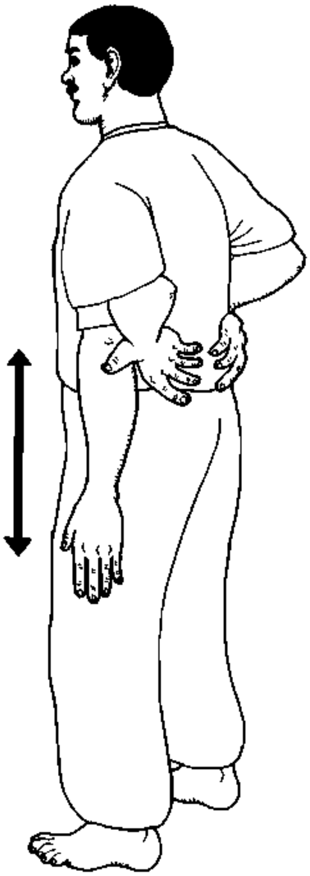

Figure 40.7. Massage the sides of the body.

The joints in between the thumb and index finger of both hands are stretched wide open. Lead the Qi to flow upward from the Yongquan Kd-1 points of the Kidney Channels, through the inner portions of the thighs, along the spinal column, and into the Kidneys.

When exhaling, hold the posture and stretch the fingers, while leading the Qi from the Kidneys into the chest and then to the Lungs. The channel Qi then transfers into the Pericardium Channels and flows down the arms to the Zhongchong Pc-9 points at the tips of the middle fingers. While exhaling, sound the word "Chree." The mouth is nearly closed with the corners slightly drawn back; the tongue moves forward. The breath should feel cold (Figure 40.6).

Slowly straighten the body as the hands circle downward, from the lower back to the thighs. Rub the lower lumbar region (Shenshu Bl-23 points) and the sides of the thighs (Huantiao GB-30 points), moving the hands upward while massaging the buttocks. Return to the holding the ball posture and repeat six times (Figure 40.7).

THE SOUND "XI" (SHE) REGULATES THE TRIPLE BURNERS

This sound is used to eliminate Excess Heat, for treating rheumatism, poor circulation, and problems of the Triple Burners. This exercise helps to recharge the body's Qi metabolism, regulate Internal Qi, and balance the eight systems of the body (respiratory, digestive, reproductive, circulatory, lymphatic, nervous, muscular/skeletal, and excretory systems).

The "She" sound of the Triple Burners is always prescribed for patients undergoing radiation or chemotherapy. This allows the patient to disperse the toxic Heat left within the tissues.

From a beginning Wuji posture, inhale and raise both hands as if holding an imaginary ball up in front of the chest at the Middle Dantian. Direct the channel Qi of the Triple Burner Channels in the arms to flow downward to the Lower Dantian, then down the legs via the Gallbladder Channels to the tips of the fourth toes of the feet at the Qiaoyin points (GB-44).

Exhale the sound "She," at the same time rotating both arms upward to form the Embracing the Heavens posture. The lips are slightly open and drawn back; the tongue suspends behind the teeth. When exhaling, direct the Qi to flow from the Qiaoyin GB-44 points on the outer portion of the fourth toes upward, along the Gall Bladder Channels, through the outer portions of the legs to the head, and transfer the energy flow into the Triple Burner Channels, directing it along the outer portions of the arms to the Guanchong TB-1 points at the ends of the ring fingers. Exhale completely.

Inhale, while turning the palms over to face downward, and begin lowering both arms along the front of the chest. Imagine the Qi flowing down from both arms into the Lower Dantian. Exhale without sounding "She," and press both palms down along the front of the abdomen to the sides of the hips, returning to the Wuji posture (Figure 40.8). Repeat the entire sequence six times.

This exercise is different from the others in that the energy circulates back and forth through the channels but does not disperse outside the body. Meaning you inhale Qi from fingers to toes, exhale Qi from toes to fingers and sound, inhale Qi from fingers to Lower Dantian, exhale Qi from Lower Dantian to toes, without sound, then begin entire sequence again.

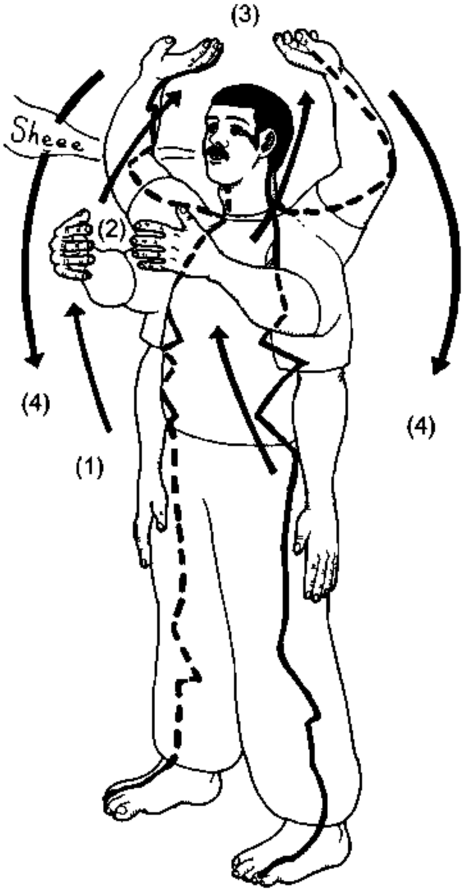

Figure 40.8. The sound "Xi" (She) is used to purge Excess Heat from the body.

THE PROPER NUMBER OF BREATHS FOR HEALING SOUND PRACTICE

When practicing all six healing sounds, each sound should be practiced a minimum of three to six times before proceeding on to the next sound. If the patient wishes to practice the six healing sounds daily, the patient should perform each sound no more than twenty-four times. If the patient is practicing one or more healing sounds for a specific therapy, then the number of times for tone resonances should be thirty-six.

Practice the six healing sounds in the sequential order of the Fire Cycle (i.e., Liver, Heart, Spleen, Lungs, Kidneys, and Triple Burners) to enhance the

Yang energetic field. Practice the six healing sounds in the sequential order of the Water Cycle (i.e., Liver, Kidneys, Lungs, Spleen, Heart, and Triple Burners) to enhance the Yin energetic field.

It is advisable to practice the healing sounds at the end of a meditation practice. This is because the free circulation of Qi is already in motion and freely accessible after meditating.

COMBINING MULTIPLE ORGAN SOUNDS

The six healing organ sounds may be performed as in the prescribed series, or as individual therapies ranging from single to multiple organ sounds. Some examples are:

1. Single organ sound regulation, like the "Haa" sound for the Heart, is used followed by swallowing the collected saliva (after Qi purgation) to relieve a fast, irregular heartbeat.
2. Multiple organ sound regulations, like the sounds of "Shu" (Liver), "Haa" (Heart), and "Sss" (Lungs), are used in sequence to quench the pathogenic Fire inside the patient's body (for mild symptoms).
3. All six healing sounds are used for regulation and balance to treat "False Cold" (a feverish sensation followed by a feeling of cold all over the body). This symptom is caused by Excess Fire. The Internal Fire drives the Yin to the body's surface, thus producing the cold sensation. To treat, the Qigong doctor guides the patient's Yin Qi back to its original organs and uses all six sounds to regulate the body.
4. Sounding according to the Five Seasons will also be beneficial to the patient's Qi regulation. The five sounds are divided according to each season, and the vulnerability of each organ. They are categorized as follows:
    a. "Shu" sounding in the springtime will benefit the Liver.
    b. "Ha" sounding in the summer will benefit the Heart.
    c. "Ssss" sounding in the autumn will benefit the Lungs.
    d. "Chree" sounding in the winter, and during all seasons, will vitalize the Kidneys and harmonize Qi and Blood.
    e. "Who" sounding in all seasons helps vitalize the Spleen and regulates the function of Postnatal Qi.
5. Combining healing sounds according to the specific prescriptions is similar to taking combinations of Chinese herbs. The Six Healing Sounds may be combined in order to target specific areas on the body's energetic system. Generally, the doctor gives the patient specific healing sound combinations as prescriptions to purge stagnation, and to increase the energetic movement within the diseased tissues or organs.

    One such combination prescription used to treat overall Shen disturbances due to emotional depression, from suppressed anger, guilt or grief is called, "Old Man Searching for the Reflection of the Moon at the Bottom of the Tide Pool." In Medical Qigong Soul Retrieval, this healing sound prescription is used to bring the patient back in touch with his or her painful emotions.

OLD MAN SEARCHING FOR THE REFLECTION OF THE MOON AT THE BOTTOM OF TIDE POOL

This particular exercise was introduced into the United Stated in the early 70's by Dr. Her Yue Wong. Doctor Wong believed that after the Qigong student has spent a sufficient amount of time practicing the Five Elemental Daoist Qigong tonification exercises, his or her deep seated emotions will begin to surface and must be purged from the body. In order to purge the toxic emotions, this healing sound prescription was given:

1. Begin this exercise prescription from a standing posture. Both feet are facing forward, shoulders width apart. Inhale, and imagine divine healing light filling the Lungs. While inhaling, separate and stretch both arms to the sides of the body, forming the shape of a "T" (Figure 40.9).
2. Lean over and begin to exhale, making the "Sh-h-h-h-h" sound. While exhaling imagine Toxic Qi from both Lungs flowing down each arm and pouring into the ground. Both arms should swing from one side to the other, crisscrossing

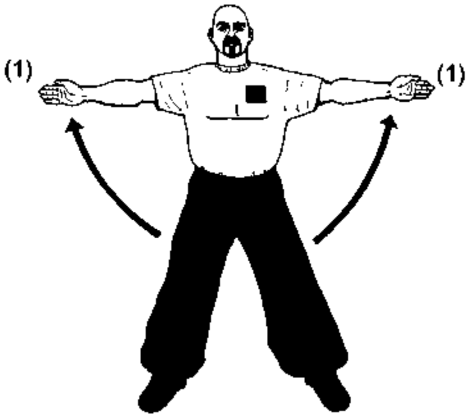

Figure 40. 9. Old Man and Tide Pool (a)

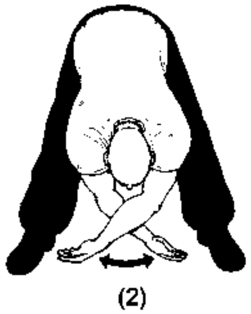

Figure 40. 10. Old Man and Tide Pool (b)

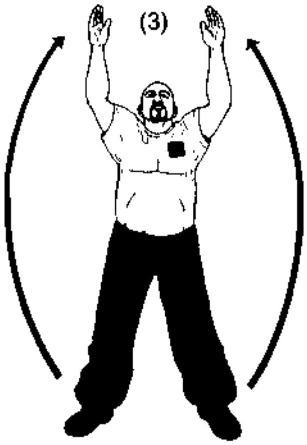

Figure 40. 11. Old Man and Tide Pool (c)

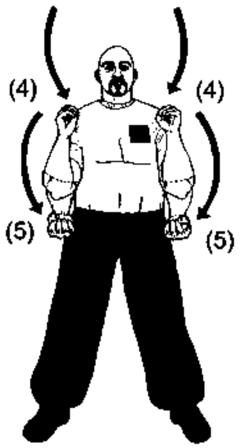

Figure 40. 12. Old Man and Tide Pool (d)

while making the purging sound (Figure 40.10).

3. Return to an upright position, while inhaling and imagining divine healing light filling the chest (especially the Heart, Liver and Spleen areas). When inhaling, look upward and raise the hands over the head (Figure 40.11).
4. When exhaling, both hands should descend to shoulder level, making soft fists (as if embracing two sparrows). Focus the mind's intention on the center of the chest and imagine the Toxic Qi releasing from the Heart while making the "Haa-a-a-a" sound. Exhale only half of your breath. Next, exhale the second half of your breath while imagining Toxic Qi releasing from the Liver and Spleen and making the "Who-o-o-o" sound. While exhaling, move both hands down to the hips while keeping the fists soft (Figure 40.12). Repeat the last two steps (c. and d.) three times. This equals one set. Practice 36 sets to complete the exercise. It should take about 25 minutes.

There is an story associated with this ancient healing sound prescription; it is described as follows.

One night, on a full moon, an old man on the southern coast of China, went looking for sea crustaceans to feed his family. Holding his lamp upward, he gazed down and saw a large beautiful pearl at the bottom of a tide pool. "Oh My!" He exclaimed, as he quickly placed his lamp on a rock. "Look at the size of that pearl! With a pearl that large, I'll be able to retire! Even my family won't have to work, and we'll all live in the lap of luxury for the rest of our lives!"

So the old man stretched his hands wide and began splashing his arms in the tide pool, trying to grasp the large pearl. Sh-h-h-h-h-h-h-h-h! went the water as the old man frantically grasped at nothing. Perplexed, the old man looked up. Reaching into the night sky, the old man's eyes widened as he saw the full moon. "Ha-a-a-a-a-a-a!" exclaimed the old man, "It's the moon!" "Whooo-o-o-o-o-o" cried the old man, "There's no money!"

CONTRAINDICATIONS FOR THE SIX HEALING SOUND THERAPY

There are several counter-indications for the use of the Six Healing Sound Therapy:

- The resonating sounds have a strong affect on the tissues, therefore patients who have broken or fractured bones are forbidden to practice sound therapy. Only after the bones tissues have mended are the patients allowed to practice the healing sounds.
- Due to the stimulating effect of the Six Healing Sound Therapy, it is advisable for the Qigong doctor not to prescribe the practice of the six healing tones in a very acute stage of illness. Vibrating the patient's internal organs in the beginning stages of a disease (i.e., a high fever) may stimulate the patient's virus or bacterial growth.
- Women are prohibited from practicing Tone Resonation therapy when pregnant, as the tone resonations have a pattern of dispersing clots and may lead to spontaneous miscarriages.
- Women are prohibited from practicing Tone Resonation therapy when menstruating, as it is important for the woman's body to be allowed to naturally Purge Toxic Blood from the body.

SOUND RESONATION THERAPY FOR TUMOR AND CANCER

Cysts, tumors, and cancer are characteristic of Excess stagnation of Qi and Blood, as well as stagnations of Heat and Cold. Chronic diseases, marked by obstruction of channel Qi, disorders of Qi and Blood, and the breakdown of Yin and Yang balance, are also included as Excess stagnation syndromes. The following sound methods involve a special series of tone resonation exercises (similar to, but stronger then the six healing sounds) which are utilized to disperse stagnations.

In order to find the correct pitch, the patient starts with a soft volume sound to experience the physical characteristics of the internal organ vibrating. While pronouncing the tone with the vocal cords relaxed, the patient gradually increases the sound to a louder volume.

While exhaling the sound, the patient focuses on an internal image of the diseased tumor dissolving. The patient must actually hear and feel the sound resonating the body's tissues, saturating the organ with sound vibration, energy, and light.

The following sounds are prescribed for the treatment of tumor and cancer therapy: the "Guo" sound is used by patients with Liver diseases, the "Zheng" (pronounced as jang) sound is pronounced by patients with Heart trouble, the "Gong" sound by patients with Spleen diseases, the "Shang" sound by patients with Lung trouble, and the "Yu" sound by patients with Kidney diseases. The relationship between the five pronunciations and the five solid organs as well as the number of pronunciations are shown in Figure 40.13.

PREPARATION FOR TUMOR AND CANCER SOUND THERAPY

Have the patient begin this exercise in a Wuji posture, and encourage the patient through the One Through Four meditation (One-Fun, Two-Shoe, Three-Tree, and Four-Core) described in Chapter 26.

Each tumor and cancer healing sound is pronounced in a straight tone, facing the left and exhaling the tone while rotating the body to the right. This allows the patient to purge the specific diseased organ. After completing the required number of straight tone exhalations, the patient faces forward and pauses.

The patient then begins the second half of the

		Elements				
		Wood	Fire	Earth	Metal	Water
Yin organs	solid organs	Liver	Heart	Spleen	Lung	Kidney
Yang organs	hollow organs	Gall Bladder	Small Intestine	Stomach	Large Intestine	Bladder
body turns left to right body turns right to left	high-pitched straight tone low-pitched descending-ascending tone	guo	zheng	gong	shang	yu
number of repetitions		8	7	10	9	6

Figure 40.13. Tumor and Cancer Sound Therapy Chart

prescription by facing the right direction and exhaling a descending tone while rotating the body to the center of the room, then rising tone while rotating the body towards the left side of the room. This application causes the patient to purge the diseased organ's channels of Toxic Qi.

Upon completion the patient performs the Pulling Down the Heavens technique to cleanse the body of any leftover energetic residue.

THE PROPER AMOUNT OF TONE REGULATIONS

The beginning patient should not be too anxious to perform the required number of pronunciations. By gradually reaching the required number, and slowly adapting themselves to the Tone Resonation method, the patients avoid getting tired. The sound "Shang," for example, is pronounced nine times for each exercise in the case of Lung disease. However, a beginner may start by pronouncing the tone three times, and then gradually progress to six times in twenty days. Then, in another thirty days, the patient can reach the required number of times. The number of repetitions is always dependent upon the patient's age, type of constitution and the severity of the disease.

In other words, the amount of times in a day that the patient performs the healing sounds depends upon the patient's general condition. The sounds used in tumor and cancer therapy, for example, will be prescribed more frequently with less volume then that of the regular six healing sound therapies. For Lung cancer, for example, the patient should perform the "Shang" sound eighteen times a day, nine breaths in each direction, right and left; for Lung congestion, perform the "Sss" sound three times a day, 36 breaths for each set.

SPECIALIZED PRESCRIPTIONS FOR TUMORS AND CANCER

The sound therapies used for treating tumors and cancer are quite different and much more powerful then the six healing sounds previously prescribed to purge and regulate the body's internal organs. Because of the dramatic difference in power and energetic potential between these two healing sound systems, the cancer patient will be given a lower number in tone respirations, but an increased number of times of practice per day.

The healing sounds used for treating tumors and cancers are as follows: for Liver/Gall Blad-

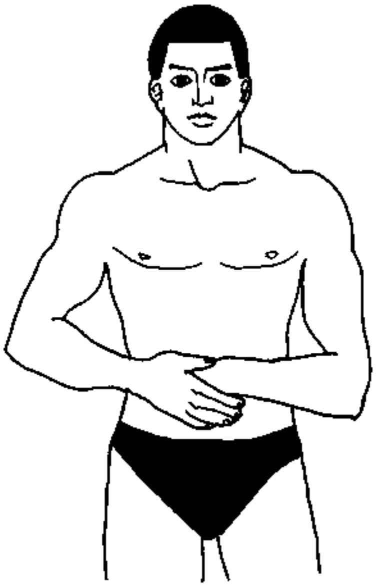

Figure 40.14. For Liver diseases the sound is "Guo."

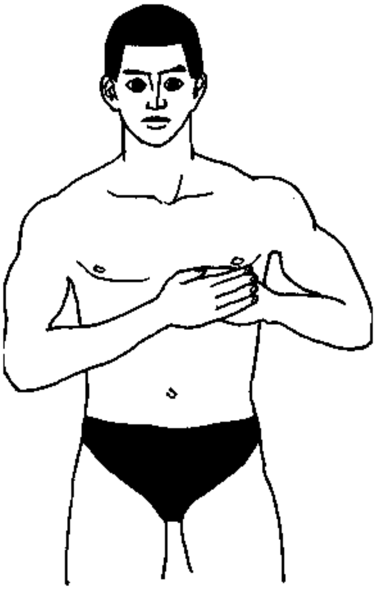

Figure 40.15. For Heart trouble, the sound is "Zheng" (Jang).

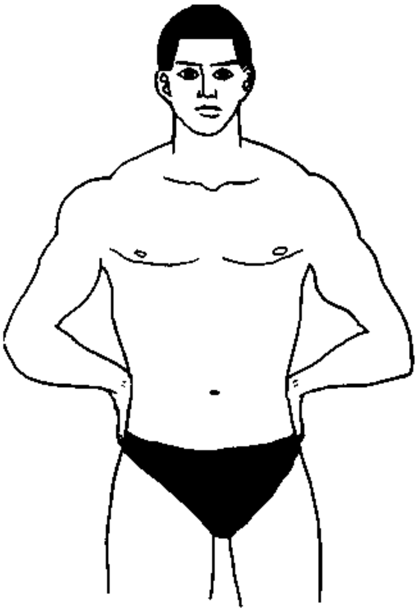

Figure 40.16. For Spleen diseases, the sound is "Gong."

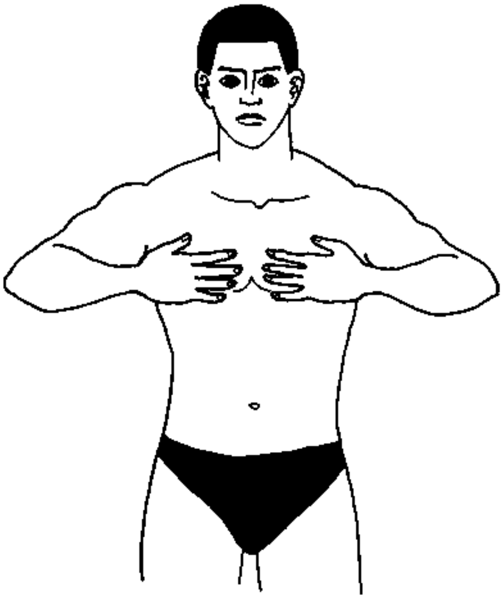

Figure 40.17. For Lung trouble, the sound is "Shang."

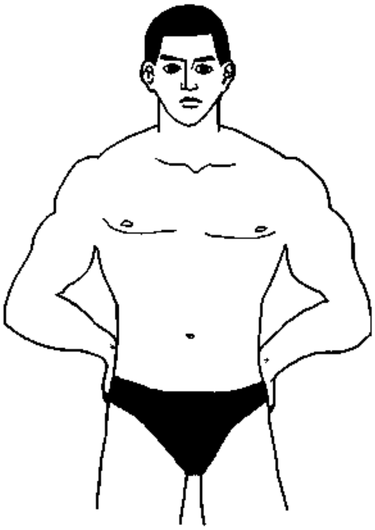

Figure 40.18. For Kidney diseases, the sound is "Yu."

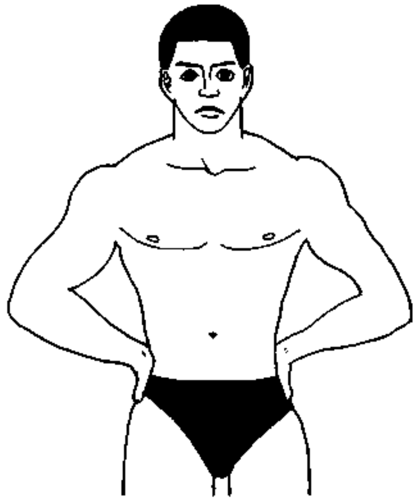

Figure 40.19. For chemotherapy, the sound is "Xi" (She).

der tumors, use the sound "Guo" (Figure 40.14); for Heart/Small Intestine tumors, use the sound "Zheng" (Jang - Figure 40.15); for Spleen/Stomach tumors, use the sound "Gong" (Figure 40.16); for Lung/Large Intestine tumors, use the sound "Shang" (Figure 40.17); and for Kidney/Bladder tumors, use the sound "Yu" (Figure 40.18). The Triple Burners' sound "Xi" (She) is used to purge Heat from the body after radiation and chemotherapy (Figure 40.19).

The choice of sound resonation must suit the patient's specific disease. Generally, in the beginning, patients with cancer should mainly pronounce the purgation sound "Haa" nine times to purge the Heart of stagnant emotions (Figure 40.20). After pronouncing the sound "Haa" for a certain time period, the patient begins to become emotionally vulnerable. The doctor then adds the specific sounds for the treatment of the diseased internal organs. The "Shang" sound, for example, is prescribed for patients with Lung cancer.

1. The tones for Radiation and Chemotherapy patients should include the Triple Burners' sound "She." This toning should be done when their blood count is low after chemotherapy or radiation therapy and is used to purge Excess Heat from the body. The sound "She" is pronounced six times. Its high-pitched sound is in the straight tone (perform six times), and its low-pitched sound is in the dropping and rising tone (perform six times).

   After the patient has stabilized and the red and white blood count has reached a normal level, the "She" sound is dropped and replace by the sound "Haa." This is done to stabilize the patient's emotions. Instead of "Haa," the doctor can also replaced the healing sound with the specific sound needed to treat the diseased internal organ.
2. The tones for Stomach cancer patients should include the sound "Dong." The Dong sound has a deeper penetrating affect than the "Gong" sound (which is usually prescribed for Stomach diseases). The "Dong" sound should be pronounced ten times (Figure 40.21). Its high-pitched sound is "Dong" in the straight tone (perform ten times), and its low-pitched sound is "Dong" in the dropping and rising tone (perform ten times).
3. The tones for Brain tumor patients (with cerebral carcinoma) should include the sound "Duo." The "Duo" sound should be pronounced ten times (Figure 40.22). Its high-pitched sound is "Duo" in the straight tone (perform ten times), and its low-pitched sound is "Duo" in the dropping and rising

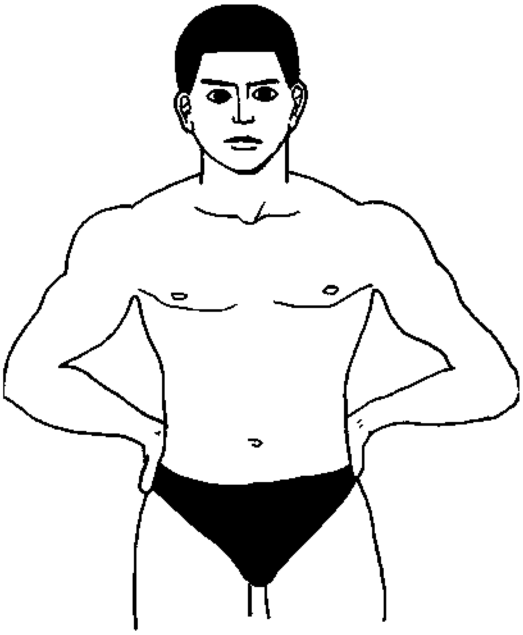

Figure 40.20. For cancer in general, the sound is "Ha."

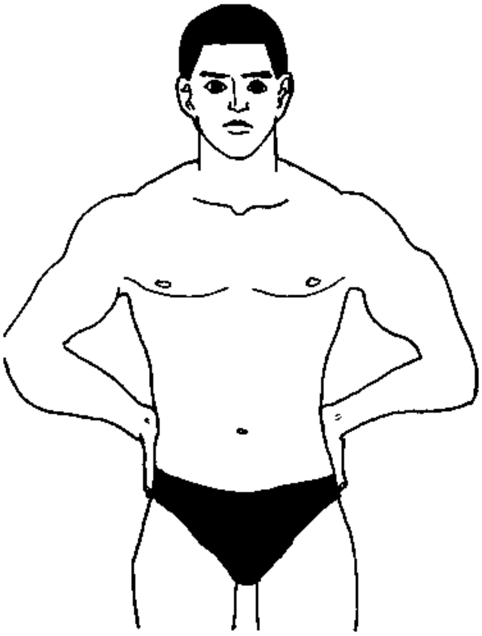

Figure 40.21. For Stomach cancer, the sound is "Dong."

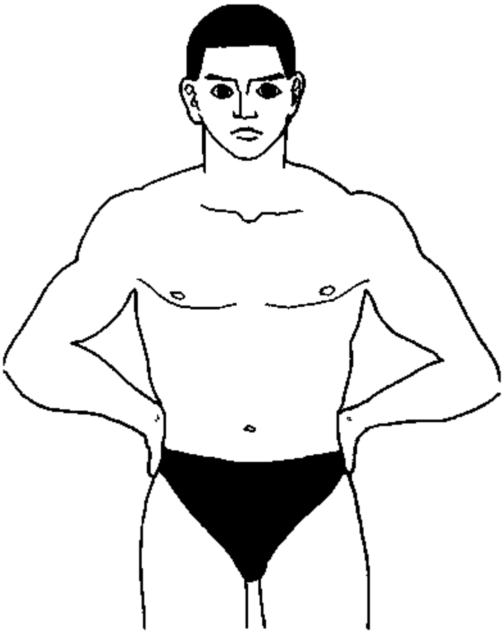

Figure 40.22. For brain cancer, the sound is "Duo."

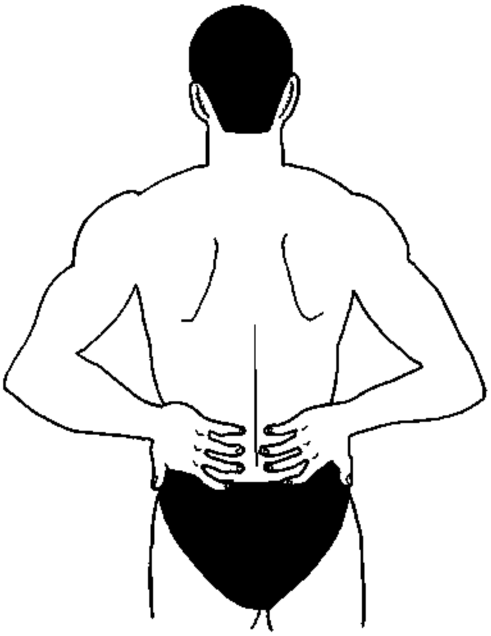

Figure 40.23. Back View for the Cancer Prescriptions

tone (perform ten times). In addition the patient should also practice techniques to strengthen the Kidneys and Heart.

4. The tones for Lung tumor patients (with carcinomas of the Lungs) should include the "Shang" sound. The "Shang" sound should be pronounced nine times. The patient should also use the Fast Breathing Method technique (2 inhalations, followed by 1 exhalation of the sound "Shang"), in addition to practicing Lung and Kidney strengthening exercises.
5. The tones for Large Intestine tumor patients (with carcinomas of the Large Intestine) should include the "Shang" sound. The "Shang" sound should be practiced nine times using the Fast Breathing Method technique, in addition to practicing Lung strengthening exercises.
6. The tones for Nasopharyngeal tumor patients should include the "Shang" sound. The "Shang" sound should be practiced nine times using the Fast Breathing Method Walking technique (see Chapter 15), in addition to practicing Lung strengthening exercises.
7. The tones for Small Intestine tumor patients (with carcinomas of the Small Intestine) should include the "Zheng" sound. The "Zheng" sound should be practiced seven times, using the Fast Breathing Method technique, in addition to practicing Heart strengthening exercises.
8. The tones for Liver Cancer and Gall Bladder tumor patients (with carcinomas of the Liver and Gall Bladder) should include the "Guo" sound. The "Gou" sound should be practiced eight times, using the Fast Breathing Method technique, in addition to practicing Lung, Liver, and Spleen strengthening exercises.
9. The tones for the Bladder and uterus tumor patients (with carcinomas of the Urinary Bladder and uterus) should include the "Yu" sound. The "Yu" sound should be practiced six times, using the Fast Breathing technique, in addition to practicing Kidney strengthening exercises.

   While in a Wuji posture, the patient places both hands on the Kidneys and presses the middle fingers at the Shenshu Bl-23 point beside the Mingmen before pronouncing the sounds (Figure 40.23).
10. The tones for Prostate cancer patients should include the "Yu" sound. The "Yu" sound should be practiced six times, using the Fast Breathing Method technique, in addition to practicing Kidney strengthening exercises.
11. The tones for Skin Cancer patients (with dermal carcinomas) should include the "Shang" sound. The "Shang" sound should be practiced nine times, using the Fast Breathing Method technique, in addition to practicing Lung, Liver, and Spleen strengthening exercises.

CONTRAINDICATIONS FOR TUMOR AND CANCER SOUND THERAPY

As previously mentioned, the tumor and cancer healing sounds are only prescribed for purging Excess conditions within the patient's diseased organs, and not for treating Deficient organ conditions. If a specific internal organ is already in a state of Deficiency and the patient is allowed to focus on that organ with one of the tumor and cancer sounds, the patient's condition will worsen. This increases the potential for rapid growth of the disease.

CHAPTER 41
MEDICAL QIGONG REGULATION EXERCISES

INTRODUCTION

The following chapters explore several systems of Medical Qigong exercises used in regulation, purgation, and tonification prescriptions. Conceived in ancient China, these exercises originate from the various medical, martial, and Daoist schools of energy cultivation. It is from these ancient energetic systems that today's modern clinical Qigong prescriptions originate.

Each Qigong system has its own unique approach to circulating, balancing, and storing the body's life-force energy. No one particular style of energetic healing is superior to the next, only different in its approach. The primary goal of each of these systems is removing stagnant pathogenic energy and replenishing and circulating clean life-force energy.

The Xi Yuan Hospital in Beijing, China is famous for its advanced Medical Qigong clinics and Medical Qigong prescriptions. These exercises utilize both Static and Dynamic Postures and are generally taught in both the inpatient and outpatient clinics. The static exercises are used to store the Qi, while Dynamic Qigong exercises are used to open the patient's channels and circulate the Qi.

Each prescription should be practiced for five to ten minutes for each organ or organ system on a daily basis; or if the patient is practicing only one set (e.g., Regulation of Spleen and Stomach Qi), they should practice for thirty minutes. Unless otherwise indicated, all breathing should be "natural."

The goal is for patients to empty their body and mind, so that their emotions can be felt and activated. In the Medical Qigong clinic there is a saying, "The patient can only heal what they feel."

STARTING AND ENDING THE PRESCRIPTIONS

Before starting any Medical Qigong prescription, it is important for the patient to understand that every Qigong exercise must start and end with specific preparations.

- Before starting the Medical Qigong prescription, the patient must always purge his or her body. It does not matter if the exercise they are about to begin is used to purge, tonify, or regulate the internal energy; they must always purge first. This is similar to washing a dirty glass before filling it with clean water.
- After ending the Medical Qigong prescription, the patient must always gather the generated Internal Qi and the return to its origin. The term "Return to its Origin" refers to a three fold process:
  - cultivating, collecting, and accumulating the prenatal energy,
  - developing and transforming it into postnatal Jing, Qi, and Shen, and
  - returning it back to prenatal Jing, Qi, and Shen.

This allows the Post-Heaven Kan and Li (Water and Fire energies) to return to their origin and restores them to Pre-Heaven Qi (Heaven and Earth energies). This fusion into one energy, often called the Hun-Yuan or Blended Originals, requires the patient's original substances to condense into one energetic substance and be merged with the patient's Lower Dantian energy.

In order to collect the energy, the patient is taught to apply the Three Natures Become One technique of focused concentration (all thinking, seeing, and listening is focused into the Lower Dantian). The Three Natures are as follows: the eyes represent the perceptive nature, the ears represent the spiritual nature, and the Heart represents the courageous nature. When all three natures unite into one, the energetic refinement is called "collecting the body's true nature." The patient can then collect the body's power.

MEDICAL QIGONG CULTIVATION EXERCISES

The body is composed of different densities of energy and matter in the form of internal organs, organ systems, channels, and collaterals. When practicing Qi cultivation methods both students and patients are instructed to train in a progressive order of Qi cultivation.

First, the students and patients are taught how to purge, tonify, and regulate their internal and external channels. Once the Yin and Yang energy of the channels are balanced, the practitioners learn how to balance the Yin and Yang energy of the internal organs. This training likewise involves purging, tonifying and cultivating methods. The directed focus of intention on specific points, channels, or areas of the body causes the Qi to gather in those locations and to be absorbed into the body's tissues through the Mind's intention (whole body awareness and consciousness). The final step of training the Yuan (Original) Qi is undertaken only after the first two steps have been accomplished.

All cultivation practices of Medical Qigong therapy can be divided into two main categories of gathering and training Qi: the External Elixir Cultivation (Wai Dan Shu) and the Internal Elixir Cultivation (Nei Dan Shu). These two Qi cultivation training methods are described as follows.

EXTERNAL QI CULTIVATION

The External Elixir Cultivation (Wai Dan Shu) involves several methods of gathering energy from the external environment (Heaven and Earth) to strengthen the body's organs, organ systems, channels, and collaterals.

HERBS AND FOOD

The first method is developed through food and herbal soups, teas, and pills to increase the body's energy. When herbs are prescribed as a tonic, they stimulate the organ's energy by increasing the flow of Qi in the patient's channels. Qigong and herbology go hand-in-hand with respect to the energetic regulation and balance of the body's lifeforce energy. Herbs can be extremely important for maintaining the physiological balance of the metabolism and restoring any depletion of the body's energetic mechanism. Herbs act as a form of energetic nutrient, allowing the Earth energy to be ingested into the body through a natural substance.

MEDITATION

The second method is developed through meditation. Qi is cultivated through the body's extremities from energy originating outside of the body's torso (i.e., both palms, both soles of the feet, and the top of the head). The External Elixir Cultivation meditation focuses the concentration on either the Heaven or Earth energy fields to draw their energy into the torso for increased Yin or Yang Qi regulation. The Postural Dao Yins in External Elixir Cultivation may be either static or dynamic.

Two examples of External Elixir Cultivations are: the Earth Yin Qi Regulation and the Heaven Yang Qi Regulation. The Earth Yin and Heaven Yang Qi Regulations are used to balance the energy along the body's ascending and descending energetic channels. To balance along the horizontal energetic channel, perform the Turning and Winding the Belt Vessel exercise described later on in this chapter.

- **Earth Yin Qi Regulation**

  The Earth Qi is considered Yin Qi or electronegative energy. Blood also pertains to Yin Qi. Through Earth Regulating Qigong meditations, the body's Blood cells can be replenished with Earth Yin energy.

  Begin the Earth Qi Regulation meditation from a Wuji posture, while using Natural Breathing. In this particular exercise, the practitioner takes advantage of the Kd-1 points at the bottom of the feet, and the Pc-8 points at the center of the palms which are sensitive to the magnetic pull of the Earth.

  After standing in the Wuji posture for several minutes, inhale and begin to absorb the

Earth Qi up from the ground through the centers of the palms and the bottom of the feet (Figure 41.1). After several more minutes of drawing the Earth energy into the Lower Dantian, begin circulating the energy through the Microcosmic Orbit's Fire Cycle. The palms should begin to expand and feel almost swollen; the red and white blotches on the palms and feet indicate that the exercise has been practiced correctly.

- **Heaven Yang Qi Regulation**

  Heaven energy pertains to the sun's Yang Qi or electropositive energy, which is used to tonify the body's Yang Qi.

  Begin from a Wuji posture, placing both palms above the head, facing upwards, like two antennas collecting the sun or Heaven's Qi into the Middle Dantian. Use Natural Breathing. Imagine the sun's golden light showering the outside of your body. Open your mouth and imagine swallowing the sun's Qi, allowing it to mix with your saliva. As you swallow, feel the energy flow down into your Lower Dantian, and then out of your pores. Once the Dantian is full, the body should feel expansive. Always end the meditation by returning the energy to the Lower Dantian to avoid any Qi deviations (Figure 41.2).

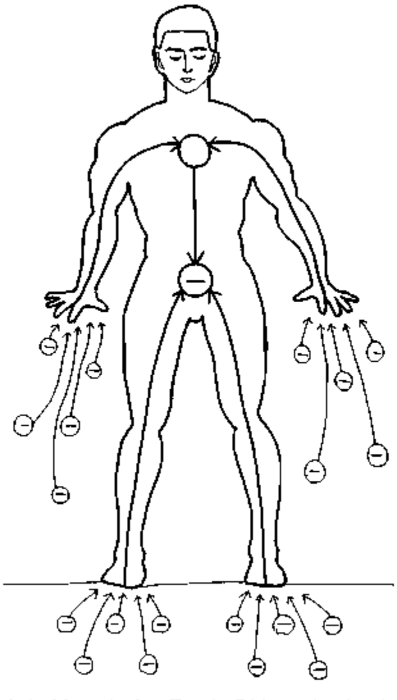

Figure 41.1. Absorb the Earth Qi into the body through both hands and feet, and store the Qi in the Lower Dantian.

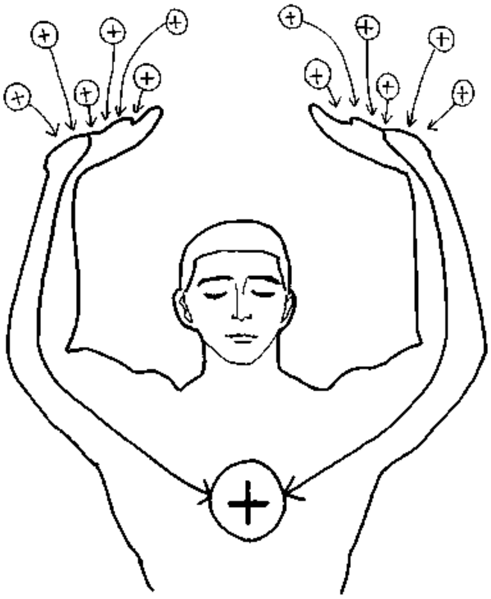

Figure 41.2. Absorb the sun's energy into the body through both palms, moving it into the Middle Dantian, and store the Qi in the Lower Dantian.

INTERNAL QI CULTIVATION

The Internal Elixir Cultivation (Nei Dan Shu) method focuses on first training the energetic substances already gathered and existing inside the torso, and then extending this accumulated energy out into the extremities. Since it first appeared in the Han Dynasty, the training and theory of Internal Elixir Cultivation has become the primary focus of many Qigong masters, working with postnatal energy transformations. Internal Elixir Cultivation focuses on transforming the body's Jing, Qi, and Shen. When practicing this exercise, ancient Qigong masters imagined the Lower Dantian as a stove, the Shen as a Fire, and the body's Jing, Qi, and Shen as the internal medicine.

At birth, the energy flows up the front of the body (Conception Vessel) and down the back (Governing Vessel). This is considered the circulation of the Yuan Qi (generated from the prenatal Jing stored in the Kidneys). It is considered the Perceptual Yin or Water Cycle of the Microcosmic Orbit. This Perceptual Yin Cycle allows the newborn child to possess powerful psychic, empathic, and kinetic abilities; it enables the child to energetically connect with the environment in order to ensure survival.

As the child advances in age, he or she begins to encounter and experience some hostile messages in the form of physical or emotional frustration and pain from both parents, siblings, and the environment. For protection, the child armors and creates a denial system, which insulates him or her from the energetic and emotional wounding. The child begins to instinctively stop the energetic flow of the perceptual Yin Water Cycle and begins to activate the acquired rational mind, trying to cope with his or her feelings and the stress of his or her surroundings. Whenever the child begins to experience painful emotions, he or she learns to avoid experiencing these hurtful feelings through certain mental and physical actions by contracting the throat and thorax. These acquired actions further create an energetic restriction and impede the flow of Qi along the Conception and Governing Vessels. This physical activity naturally causes the Yin Water Cycle to stop its perceptual flow of energy, and allows the child's body to deal with the trapped emotional experiences by storing them within the tissues. Although this is not a healthy response, it allows the child to survive and develop in a sometimes difficult environment. In cases of deep traumatic wounding (physical, emotional or sexual), the child becomes void of deep emotional connections, while still allowing for certain perceptual survival instincts.

Also, our culture emphasizes left brain thinking for educational purposes, and later to enable young adults to rival others in the competitive job market. The child thus must forgo most of his or her innate intuitive perceptions to succeed in school and to appropriately integrate society's values and belief structures. Very little encouragement is given to right brain activity; it is in fact discouraged by society at large.

REGULATING THE BODY'S QI

The first stage in circulating and cultivating the Qi is to regulate the Yang or Fire Cycle of the Microcosmic Orbit. This allows the trapped emotions within the tissues to be released. This is initiated through mental concentration, focusing the flow of the Qi up the back and down the front of the body. It is extremely important to allow the trapped energy to release itself naturally from the body through the Fire Cycle of the Microcosmic Orbit. Otherwise, trapped emotions deep within the tissues can cause Qi deviations within the body, which may lead to energetically induced psychosis.

When regulating the Microcosmic Orbit, there are three primary methods of energetic cultivation that harmonize the body's Jing, Qi, and Shen. They relate to the sun (Fire and Heaven), moon (Water and Earth) and stars (Wind and Man).

THE FIRE PATH OF THE MICROCOSMIC ORBIT

The Yang or Fire Path is the most popular method of cultivating energy along the Small Heavenly Cycle or Microcosmic Orbit (Figure 41.3). Because it joins together the body's Yin and Yang rivers of Qi, which regulate the Twelve Primary Channels, it is considered the foundational meditation for the Internal Elixir Cultivation methods. The circulation of Postnatal Qi is generated from the Acquired Essence (food, air, and water). In this cycle, the Qi moves up the Governing Vessel and down the Conception Vessel, and is considered the ascension of Yang or the opening of the Fire Channel. The initial goal is to circulate the Fire energy from Yang to Yin to convert Jing to Qi and then to Shen. The Yang (Fire) Qi supports awareness of the emotional side of consciousness.

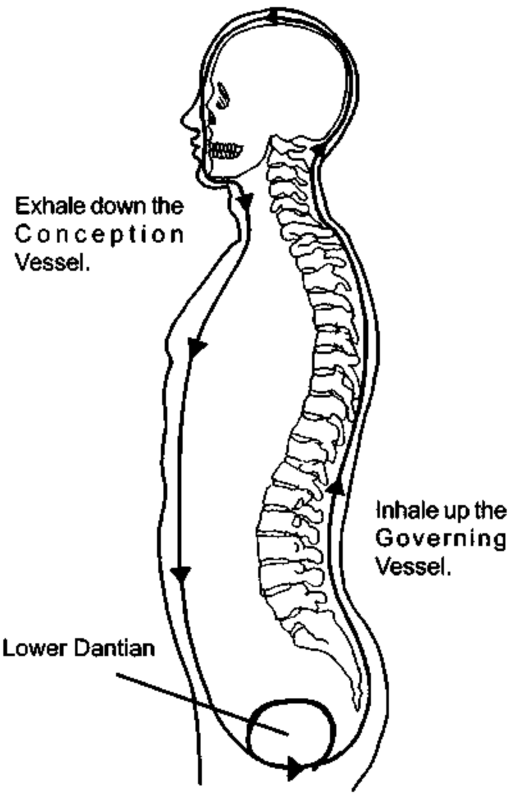
Exhale down the Conception Vessel.
Inhale up the Governing Vessel.
Lower Dantian

Figure 41.3. The Fire Path of the Microcosmic Orbit flows up the back and down the front of the body.

The Heart and Middle Dantian are considered the House of Shen; Fire and passion (Qi that manifests as emotions) and are related to the flow of the Conception Vessel. The Kidneys and Lower Dantian are considered the House of Jing, Water and pleasure and are related to the flow of the Governing Vessel. It is important to use Reverse Breathing while uniting the Heart's Fire and Kidneys' Water, to cause the alchemical transformation of Jing into Qi in the lower abdomen.

When moving the Qi along the Governing Vessel, practitioners must move the energy through the Five Gates (see Chapter 30). The energy first gathers at the Lower Dantian, then overflows into the Governing Vessel, collecting, building, and then passing through to the next gate, one gate at a time. As the energy gathers at these gates, the Qi and Blood begins to Heat and vibrate the tissues. This energetic reaction affects the patient's emotions that correspond to or are embedded in each specific gate, causing a powerful emotional release. As the Qi passes through to the next gate, a new set of emotions is experienced, then released.

THE JIAJI POINTS' EFFECT ON THE GOVERNING VESSEL'S ENERGETIC FLOW

Jia translates to mean "besides." Ji translates as "the spinal column." These points are located lateral to the lower border of each spinous process of the cervical, thoracic, and lumbar vertebrae. Each point has its own Luo pathway (energetic channel), which intersects with the spine and connects to the Jiaji point located on the opposite side of the Governing Vessel. It is believed that one pair of the ascending Governing Vessel's Luo Channel intersects and connects to all of the Jiaji points along the spine.

This being the case, when the two energetic flows from the Jiaji points (via the Luo pathways) are superimposed, the reaction causes a crisscrossing and intertwining of Qi flow, which resembles the intertwining energies of the caduceus or double snakes.

THE TWELVE EARTHLY BRANCHES' RELATIONSHIP WITH THE MICROCOSMIC ORBIT

Within the Fire Cycle of the Microcosmic Orbit, there are twelve channel points located along the Governing and Conception Vessels which relate to the Twelve Earthly Branches and the waxing and waning of the Yin and Yang cycles (see Chapter 11). The Fire phase of the Microcosmic Orbit starts at the Zi branch located at the Sheng Si Qiao area at the root of the penis in men and at the vagina in women. Moving up the back of the body, there are four phases of Qi ascent in the Governing Vessel. The four phases are: from Zi to Chou, from Zi to Yin, from Zi to Chen, and from Zi to Si. The Mao point is one of the four cardinal points (along with Zi, Wu, and You) and is a point of purification. The path up the Governing Vessel cleanses the energy through this point of purification.

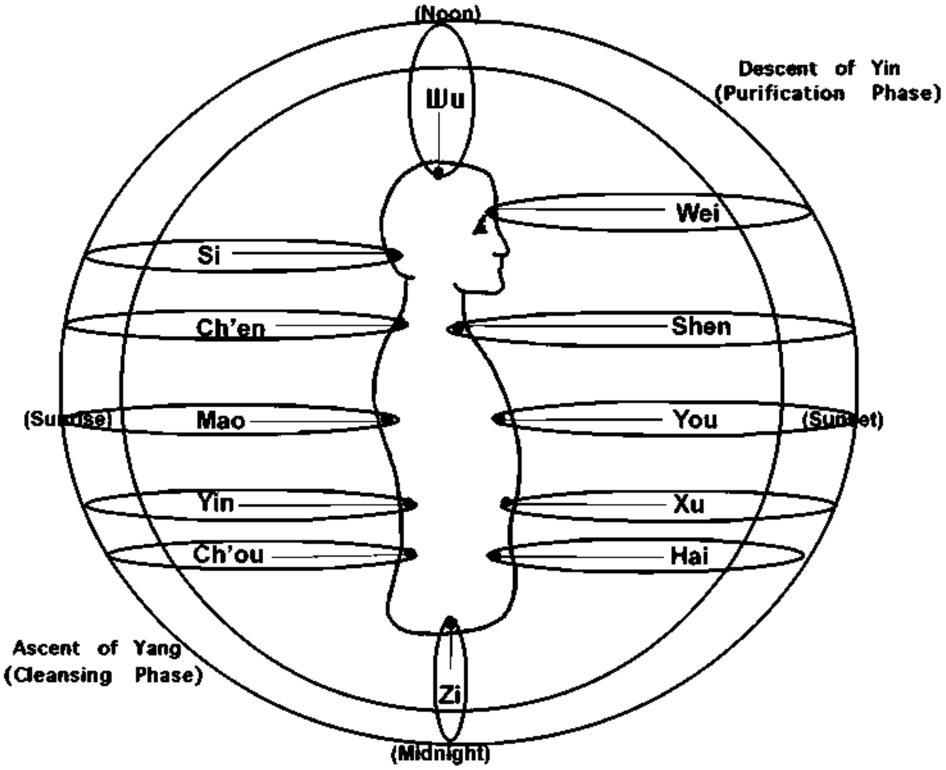

Figure 41.4. Twelve Earthly Branch Relation with the Microcosmic Orbit: the main channels of the back and front of the body correspond to the ecliptic path of the sun. These twelve special energy points are associated with the Twelve Chakra Gates.

When the energy reaches the Wu point at the top of the head (also known as the Ascending point), it stays there briefly before beginning the descent of Yin in the Conception Vessel. The descent down the Conception Vessel purifies the energy (Figure 41.4).

While at the perineum and Zi point in the orbit, the Yang Qi begins to develop as the Yin Qi reaches its zenith. Once the energy reaches the Mao point at the middle of the back, the Yang energy has developed fully and begins to prosper. At the Wu point at the top of the head, the Yin Qi begins to develop while the Yang Qi reaches its zenith. This point represents the turning point from Yang to Yin.

If the energy is to flow properly down the Conception Vessel, there must be an opening of the energy centers via a connection with the Shen. This purification of energy commences at the Wu point when the Yang Qi turns to Yin Qi. While the Earth's daily orbit around the sun has the same Yin-Yang transitions, the body absorbs the Yang energy during the hours of Zi and Wu, and expels Yin energy during the hours of Mao and You.

At the Zi hour (11 p.m.–1 a.m.), the Jing naturally collects at the Sheng Si Qiao point at the root of the penis (prostate area) in men causes a natural erection while the man is sleeping. This hour also marks the commencement of the Yang half of the day. This period of the day is the best for collecting Jing and, thus, is the best time to perform Wuji standing practice.

CONTRAINDICATIONS

Whenever the practitioners are weak, it is important that they practice Natural Dantian Breathing and avoid practicing the Microcosmic Orbit. If patients attempt to practice Circular Breathing while in a depleted state, the Qi may become trapped in their head causing Qi Deviations, and initiating more problems. Circular Breathing is achieved by guiding the Qi in a circular pattern through even inhalation (e.g., up the spine) and exhalation (e.g., down the front of the body).

THE WATER PATH OF THE MICROCOSMIC ORBIT

The Water Path is trained only after the completion of the Fire Path circulation exercise. After the patient's Prenatal Qi has been cultivated and accumulated to a substantial level, the patient may lead the Qi up the Conception Vessel and down the Governing Vessel (Figure 41.5).

This is opposite the direction of the flow of the Fire Path, and is considered the ascension of Yin or the opening of the Water Channel. The purpose for the Water Path circulation is to change, regulate, and charge the body's Yuan Shen. The Water cycle supports the perceptional side of consciousness, cools down the body's overheated Yang Qi, and regulate the Qi circulation of the Fire Path to rebalance the body.

THE WIND PATH OF THE MICROCOSMIC ORBIT

The Wind Path of the Microcosmic Orbit is the third method of circulation and requires practitioners to lead energy up the center of their body through the Taiji Pole (Figure 41.6). This energy pathway flows up through the Marrow of the spine, through the brain, and washes over the falx cerebri into the pineal and pituitary gland. It then descends down along the Conception Vessel and back into the Lower Dantian.

The purpose of the Wind Path is to increase divine perception (by accessing the energy of the Wu Jing Shen), and to regulate the body's glandular and endocrine systems, to enable the practitioner to slow down the aging process.

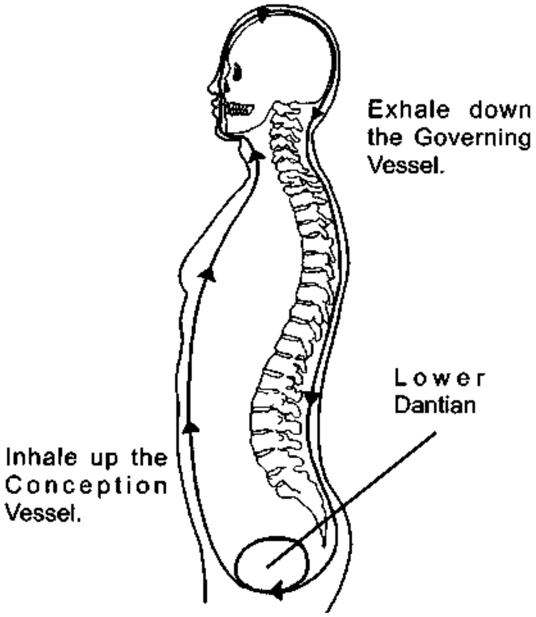
Exhale down the Governing Vessel.
Lower Dantian
Inhale up the Conception Vessel.

Figure 41.5. The Water Path of the Microcosmic Orbit travels up the front and down the back of the body, ending at the Lower Dantian.

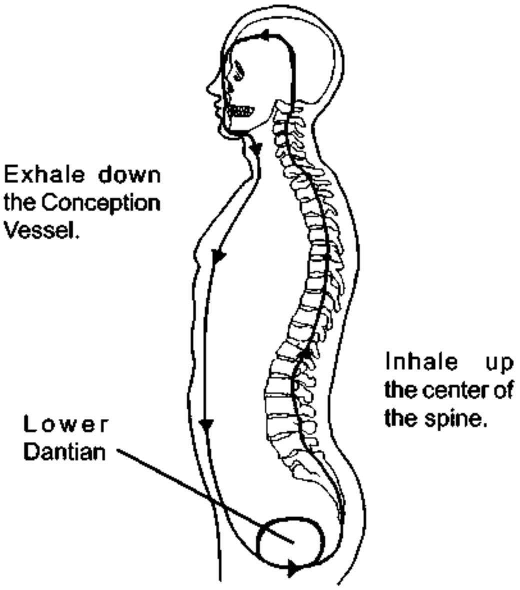
Exhale down the Conception Vessel.
Inhale up the center of the spine.
Lower Dantian

Figure 41.6. The Wind Path of the Microcosmic Orbit travels up the center of the spine into the brain and third-eye point, down the front of the body, ending at the Lower Dantian.

MEDICAL QIGONG PRESCRIPTIONS

The following exercises are taken from various hospitals, universities, and institutes of Medical Qigong. These regulation prescriptions are used throughout China.

DESCEND THE QI AND CLEANSE THE ORGANS

This exercise was developed to utilize the body's Hun-Yuan Qi (the body's internal combined organ energy) to flush and cleanse the Triple Burners and the Five Yin Organs. This is a purging exercise, used to:

- eliminate the body's Turbid Qi,
- unclog the channels and collaterals,
- wash the bones,
- transform the Marrow, and
- combine the remaining pure energy with the body's Yuan Qi.

It can be prescribed for patients with Excess conditions to descend the Liver's or Heart's Fire.

PULLING DOWN THE HEAVENS

This exercise requires the focus of energy to flow down the body in three directions (front, center, and back) to complete one set.

1. Begin in a Wuji posture, with both feet facing forward, a little wider than shoulders width apart. Both hands are at the sides of the body. The spine is kept straight, with the body relaxed and centered (see Chapter 15 for proper alignment). Stand with the eyes focused on the horizon. Pull the Qi from the horizon into the eyes and down to the Lower Dantian, filling the entire body. Hold this posture for one to three minutes (Figure 41.7).
2. Inhale through the nose. With the palms facing downward, begin to scoop the energy from the Earth into the body, gathering and absorbing the environmental energy. Focus on slowly filling the body with Earth Qi while moving the hands upwards to about waist level (Figure 41.8).
3. Continue inhaling, turning the palms to face the Heavens and drawing both hands upwards at a forty-five degrees angle. As both palms continue to move upwards, begin pulling in and absorbing the Heavenly Qi while guiding the hands above the head. Complete the inhalation with the arms above the head, palms facing the top of the head. Pause when the hands are above the Baihui area (Figure 41.9).
4. Exhale through the nose, and begin descending the palms in front of the body. Allow the palms to face the body and direct the Qi flow through the head and throat area (Figure 41.10).
5. Imagine the energy continuing to pour down through the chest and abdominal areas (Figure 41.11).
6. While exhaling, slowly lower the hands in front of the body in one continuous movement, until they reach just above the knees (Figure 41.12).
7. Inhale and begin the sequence again. Each time the hands pass in front of the body, exhale and imagine the Qi being emitted through the hands into the tissues. Three passes over the front of the body equal one set, practice 36 sets. The energetic pathway of each set is explained as follows:
    a. The first pass begins by leading the Qi down the front of the body from the Baihui area. Each palm facing the body divides the energy into two rivers, which flow from the top of the head, over the ears, and down the throat. Continue guiding the two rivers down the front of the chest, joining again at the navel. From there lead the Qi down into the Huiyin CV-1 point at the perineum, then separate them again into two rivers which continue to flow down the Yin channels on the inside of each leg, ending at the Yongquan Kd-1 point at the bottom of each foot.
    b. Start the next pass at the middle of the head above the Baihui area. Lead the Qi down the center of the body, and absorb the Heavenly Qi as it passes through all the body's internal organs, till it reaches the Lower Dantian area. From the Lower Dantian divide the energy into two rivers and descend the Qi down the center of the legs ending

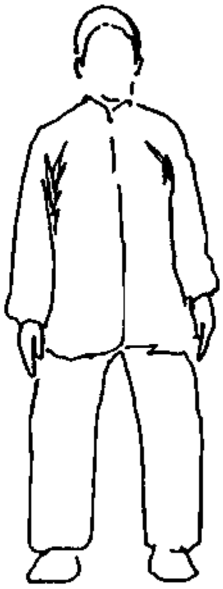

Figure 41.7. Wuji Posture

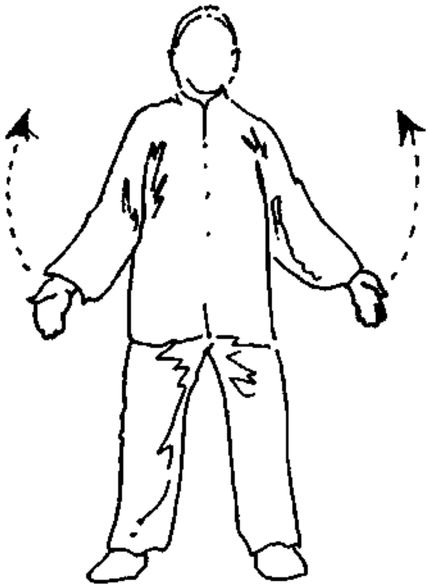

Figure 41.8. Pulling Down the Heavens (A)

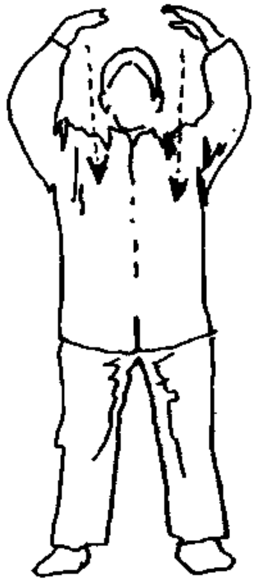

Figure 41.9. Pulling Down the Heavens (B)

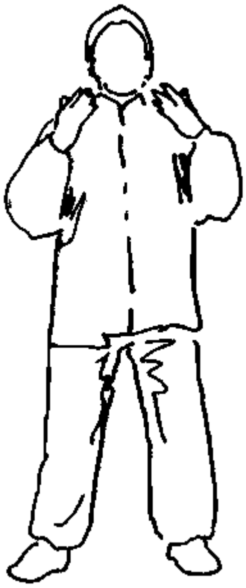

Figure 41.10. Pulling Down the Heavens (C)

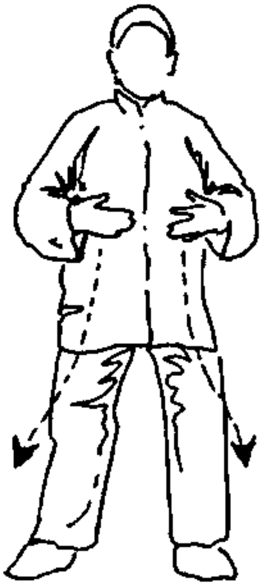

Figure 41.11. Pulling Down the Heavens (D)

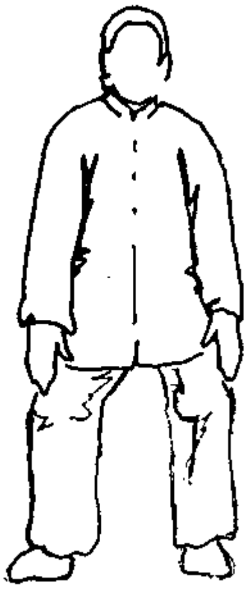

Figure 41.12. Pulling Down the Heavens (E)

at the Yongquan Kd-1 point at the bottom of each foot.

c. The final pass likewise begins at the top of the head, at the Baihui area. Imagine the energy flowing like a river, down the midline of the back of the head and back, until it reaches the Mingmen GV-4 point. From there it divides into two rivers, descending into the Huantiao GB-30 point at the side of each buttock, and continues to flow down the outside of each leg and ends at the Yongquan Kd-1 point at the bottom of each foot. As the Qi descends (on each pass), it washes the organs from the top of the head (Baihui point) to the bottom of the feet.

d. To end the exercise, relax the whole body and return to the Wuji posture, allowing the Qi to return to its origin. As the Qi descends through the body, imagine the environmental energy fusing with the body's Yuan Qi, forming a mist and dew. Imagine this mist slowly flowing down (from top to bottom) and bathing, cleansing, and purifying the entire body. As the mist descends, if a certain organ is diseased, pause for a while to bathe the area with the vapor before continuing the flow of energy. Consciously disperse the pathogens out the extremities.

GATHERING QI IN THE UPPER AND LOWER DANTIANS

This exercise is designed to strengthen, tonify, and regulate the body's internal organs and energetic fields. It is excellent for promoting and maintaining health, and can be used for tonifying the Kidneys, as well as to treat impotence, premature ejaculation, lumbago, lower back problems, and pelvic inflammatory disease (PID). When treating PID, use reverse breathing with Static Qigong training.

This is a tonification exercise, used to increase the production of Jing (sperm or ovary energy), transforming it into Qi, and to lead the Jing up the spinal cord to support the brain, as well as nurture the bone marrow. This exercise adds Fire to the Mingmen, simmering the condensation of energy within the Kidneys. When the Kidney (Water) and Heart (Fire) energies are sufficient, their energies merge to form the True Qi which cultivates one's nature, spirit, and life. This exercise also assists the Microcosmic Orbit and establishes a base for refining and gathering energy in the Upper and Lower Dantians and Huiyin.

GATHERING QI IN THE UPPER DANTIAN

Begin from a Wuji posture. Inhale, as both hands gather energy while moving up the sides of the body and embrace the Qi into the Upper Dantian. Pause and exhale while separating the hands (Figure 41.13).

Inhale and imagine the middle fingers of each hand gathering the Heaven and Earth Qi into the Upper Dantian. Imagine the body absorbing these energies through the Yin Tang point (Third Eye), into the pineal gland at the center of the brain. Imagine the light of the Upper Dantian getting brighter as the energy begins to increase (Figure 41.14). Repeat this exercise nine times.

Upon completion, focus on the Upper Dantian and begin to gather and fuse the body's Yuan Shen along with the collected Heaven and Earth energies into one energy. As this energy collects, imagine it forming into an energetic ball of white light and allow the hands to guide the light downward into the Lower Dantian.

During the process of descending the Qi from the Upper to Lower Dantian, the focus should be on feeling, guiding, listening, and seeing the energy traveling downward. As the energy of the Shen descends, the Jing energy naturally ascends. This descending and ascending action of Jing and Shen energies cultivates an abundant amount of Qi in the Lower Dantian. After the Qi has been gathered and stored there, pause a moment, allowing the energy to settle. Then proceed to regulate the Lower Dantian.

GATHERING QI IN THE LOWER DANTIAN

Both hands sink down to waist level and form an embracing posture in front of the navel. Inhale while slowly drawing the palms inward towards the navel, gathering the Qi into the center of the Lower Dantian. Pause a moment and relax, allow-

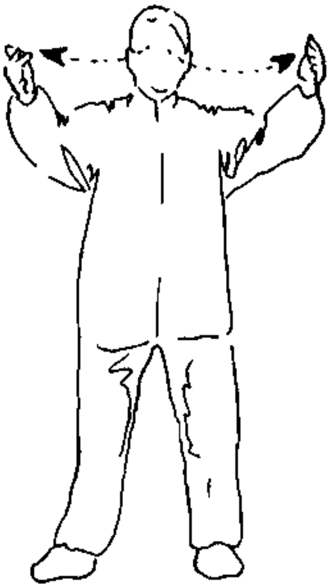

Figure 41.13. Gathering Qi in the Upper Dantian (A)

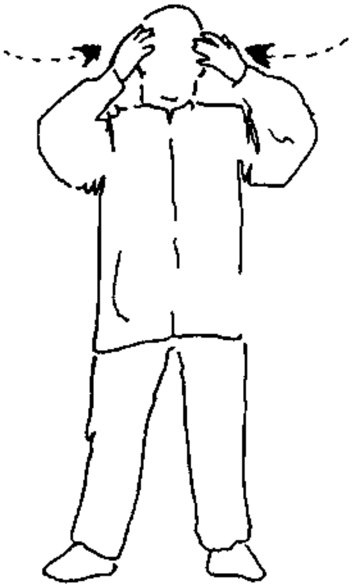

Figure 41.14. Gathering Qi in the Upper Dantian (B)

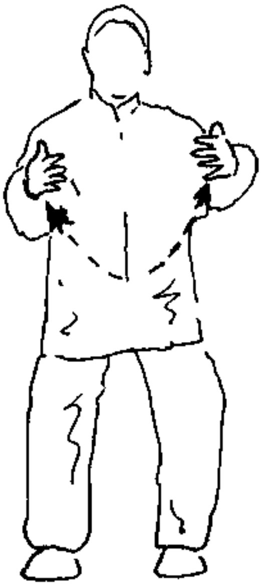

Figure 41.15. Gathering Qi in the Lower Dantian (A)

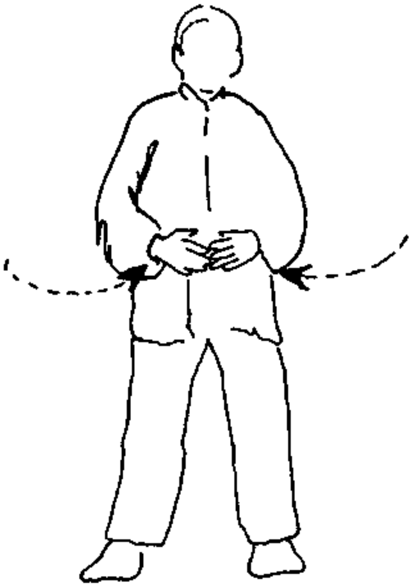

Figure 41.16. Gathering Qi in the Lower Dantian (B)

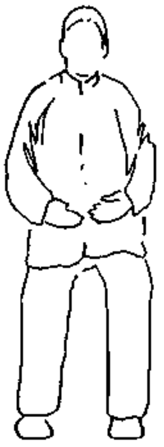

Figure 41.17. Gathering Qi in the Huiyin Area (A)

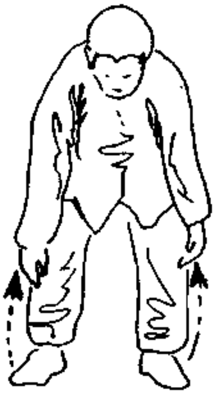

Figure 41.18. Gathering Qi in the Huiyin Area (B)

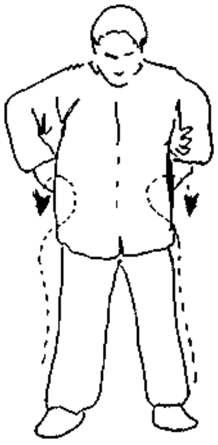

Figure 41.19. Gathering Qi in the Huiyin Area (C)

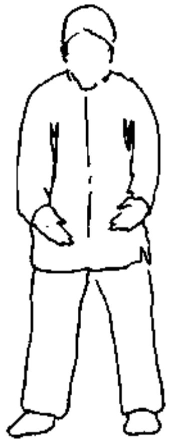

Figure 41.20. Gathering Qi in the Huiyin Area (D)

ing the energy to settle, then exhale and separate the hands (Figure 41.15).

Inhale and imagine the body absorbing the universal and environmental energies in through the navel and Lower Dantian areas (Figure 41.16). Repeat this exercise nine times.

Imagine the Lower Dantian as a reservoir, storing but never overflowing. The expansion and contraction of the abdominal muscles creates a "beating and drumming" movement which improves the abdominal muscles' elasticity while massaging the digestive system. The "beating and drumming" movement strengthens and increases peristalsis.

GATHERING QI INTO THE HUIYIN AREA

Continuing from the last posture, both hands begin to move downward from the Lower Dantian (Figure 41.17).

As you inhale, bend over and imagine scooping the Earth energy along the inside of the legs, directing the Qi through the three Yin channels into the body's Huiyin, coccyx, and Mingmen areas. While inhaling, it is important to pull up on the anal sphincter and imagine gathering the Earth Qi directly through the lower body's orifices (testicles and penis, or vagina) into the Kidneys and Mingmen area (Figure 41.18).

Move both hands from the insides of the thigh, groin, and lower abdomen backward, along the Belt Vessel, expanding the energy into the Kidneys and Mingmen area. Hold this posture and exhale, imagining the Qi flowing into the Kidneys and Mingmen areas (Figure 41.19).

Next, circle the Qi down to the Lower Dantian and Huiyin area. Pause for a moment to allow the Qi to settle and root (Figure 41,20). Repeat this exercise nine times and end this exercise.

ROLLING THE BALL WITH BOTH PALMS

The purpose of this exercise is to regulate the Yin and Yang balance of energy in the body. This exercise also causes the capillaries along the Microcosmic Orbit to open up, improving the Qi circulation and nerve stimulation in the extremities.

It is excellent for purging and regulating emotions from the Heart and Yellow Court, as well as for descending Rebellious Lung Qi downward and purging Yang Qi Rising.

SHAKING THE SUN AND MOON

Begin from a Wuji posture. Both arms raise up in front of the shoulders, palms and fingertips pointing upward. Imagine the left palm embracing the sun and the right palm embracing the moon. Vigorously shake the palms, focusing the mind's intention on the center of the palms and fingers (Figure 41.21).

Place both palms in front of the chest and Middle Dantian area, and imagine the energy of the sun and moon melting into each other, creating a huge ball of energy. Make sure the eyes focus on the center of the ball (Figure 41.22).

ROLLING THE BALL

Wait until you can feel the sensation of the energy ball before beginning the rotations. Imagine the energy ball in the hands and the energy ball in the Lower Dantian blending naturally into one energy, simultaneously rotating and rolling together both internally and externally.

Begin rolling the ball, keeping the mind's intention focused on the center of each palm. Imagine the ball is slowly growing between the hands. Roll the ball, with the hands opposite each other, (one hand is up while the other is down), eighteen times in a clockwise direction, then reverse the direction and repeat another eighteen times (Figure 41.23). After the last rotation, relax the arms and return to Wuji posture.

OPENING AND CLOSING THE UPPER, MIDDLE, AND LOWER BURNERS

This exercise is used to regulate the body's Yin and Yang energy, tonify the Yin organs, calm and harmonize the Yang organs, Regulate the Triple Burners, and help open the Conception and Governing Vessels. This is an excellent exercise for purging and regulating Excess conditions within the Triple Burners, as well as for purging the patient's Liver Fire.

There is an ancient Chinese saying, "Open, close, and come and go, one hundred illnesses will be healed." When opening, the Qi circulates into the muscles. When closing, the Qi sinks and gathers into the bones. The purpose of this exercise is to train the mind's intention, making it easier to increase the body's external field of Wei Qi.

OPENING AND CLOSING THE UPPER BURNER

Begin from a Wuji posture. Both arms raise up to shoulder level. Vigorously shake the hands, palms facing down. Before opening and closing the Upper, Middle and Lower Burner areas, the awareness must settle in the Laogong area (center of the palms) for a few minutes. Wait until heat

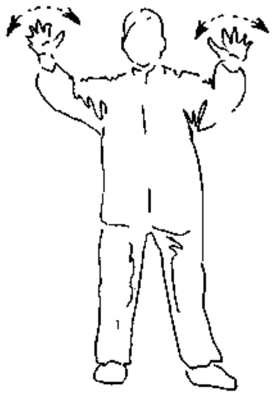

Figure 41.21. Shaking the Sun and Moon (A)

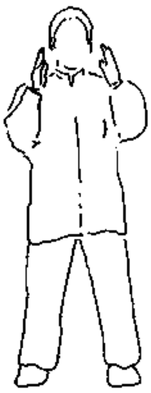

Figure 41.22. Shaking the Sun and Moon (B)

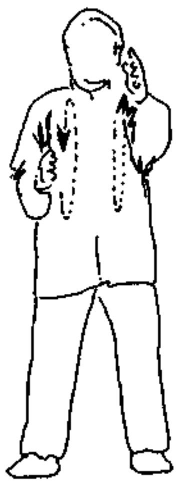

Figure 41.23. Rolling the Ball with Both Palms (A)

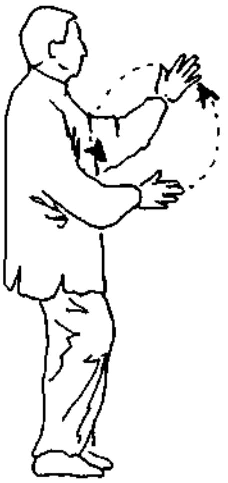

Rolling the Ball with Both Palms (Side View)

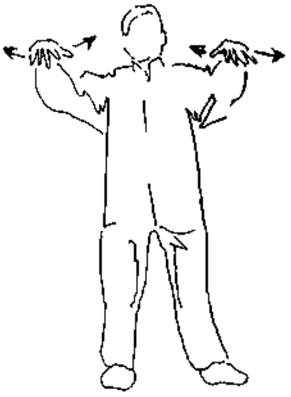
Figure 41.24. Vigorously Shake the Hands

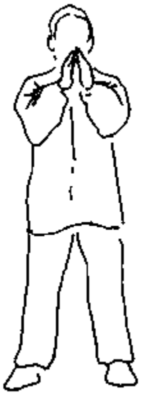
Figure 41.25. Open and Close the Upper Burner (A)

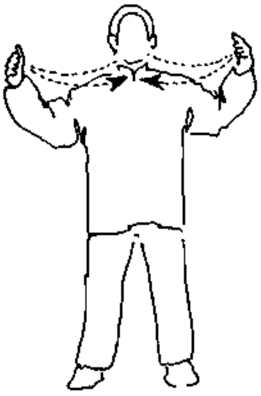
Figure 41.26. Open and Close the Upper Burner (B)

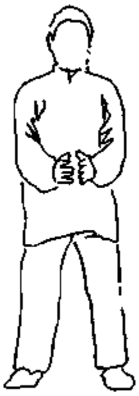
Figure 41.27. Open and Close the Middle Burner (A)

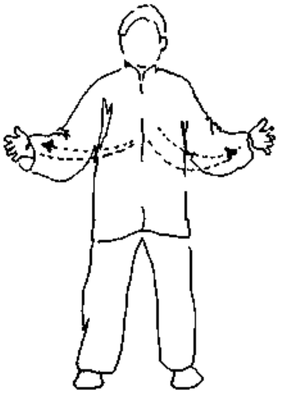
Figure 41.28. Open and Close the Middle Burner (B)

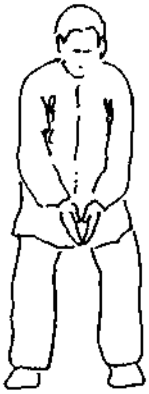
Figure 41.29. Open and Close the Lower Burner (A)

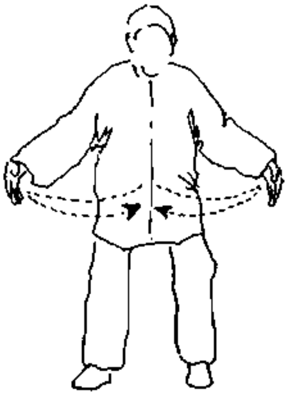
Figure 41.30. Open and Close the Lower Burner (B)

is felt in the center of the palms before beginning (Figure 41.24).

After a while, bring both hands close to each other in front of the throat and manubrium sterni area (the upper segment of the sternum). Both palms face each other, with the fingers facing upwards. Pause a moment, focusing the mind's attention on the center of the palms, forming an energetic ball (Figure 41.25).

Slowly begin to open by pulling the arms apart, while imagining the energy between the center of the palms stretching like warm taffy. When the hands separate and open, the mind and body (chest and abdomen) also open. Slowly begin to close by bringing both palms together, pushing and condensing the stretched energy to form an energetic ball. When the hands close, the mind and body close by removing the attention away from the outside world (Figure 41.26).

When inhaling, imagine extending the fingers deep into the body and the Qi pouring deep into the tissues. When exhaling, separate the hands while imagining the Excess Heat and toxic energy being purged from the body. Repeat the exercise eighteen times.

OPEN AND CLOSE THE MIDDLE BURNER

Both hands sink down in front of the diaphragm and xiphoid process (just above the navel). Pause a moment, keeping the eyes on the energy ball and an imaginary line which connects the inner Laogong Pc-8 points at the center of the palms (Figure 41.27).

Slowly begin to open by pulling the arms apart, while imagining the energy between the center of the palms stretching like warm taffy. While exhaling, separate the hands while imagining the Excess Heat and toxic energy being purged from the body (Figure 41.28).

Slowly close by bringing both palms together, forming an energetic ball as you push and condense the energy. While inhaling, imagine extending the fingers deep into the body and the Qi pouring deep into the tissues. Repeat the exercise eighteen times.

OPENING AND CLOSING THE LOWER BURNER

Both hands sink down in front of the reproductive organs and perineal area. Pause a moment, keeping the eyes on an imaginary line which connects the inner Laogong Pc-8 points at the center of the palms (Figure 41.29).

Slowly begin to open by pulling the arms apart, imagining the energy between the center of the palms stretching like warm taffy. When exhaling, separate the hands while imagining the Excess Heat and toxic energy being purged from the body (Figure 41.30).

Slowly close by bringing both palms together, forming an energetic ball by pushing and condensing the Qi. While inhaling, imagine extending the fingers deep into the body and the Qi pouring deep into the tissues. Repeat eighteen times.

To end the exercise, draw the Qi back into the Lower Dantian, and return to a Wuji posture.

SUN AND MOON ROTATING TECHNIQUE

This exercise will massage the internal organs of both the Heart and the Lungs. In Medical Qigong, the sun is sometimes referred to as the energy of the Heart, which is the considered the master of the three Yang energies. The moon pertains to the Lungs, which are the masters of the three Yin energies. Through the harmony of these dual powers the body's Yuan (Original) Qi is born. This exercise increases the elasticity of the Heart muscles and arteries and improves Blood circulation. It also flushes the Eight Extraordinary Vessels, increases vitality, and causes the upper half of the body to become lighter while making the lower half more solid.

This exercise can also be used in conjunction with the healing sound "Shang" to treat Lung diseases (i.e., asthma and chronic bronchitis), as well as breast cysts and tumors by purging Toxic Qi from the tissues of the Lung.

ROTATING THE PALMS

Begin from a Wuji posture, start by vigorously shaking the hands and then rubbing the palms together (Figure 41.31).

Place both hands on the chest, and draw the heat from the center of the palms into the chest, inhaling three times. Once the chest has absorbed a certain amount of heat, both palms begin to ro-

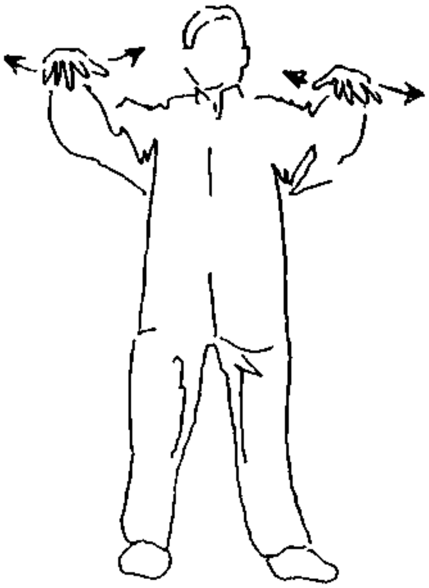

Figure 41.31. Vigorously Shake the Hands

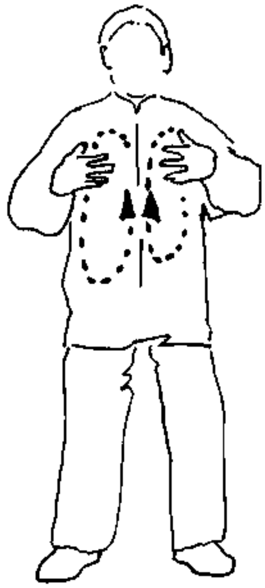

Figure 41.32. Sun and Moon Rotating Technique (A)

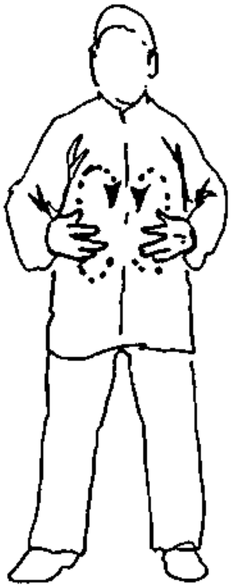

Figure 41.33. Sun and Moon Rotating Technique (B)

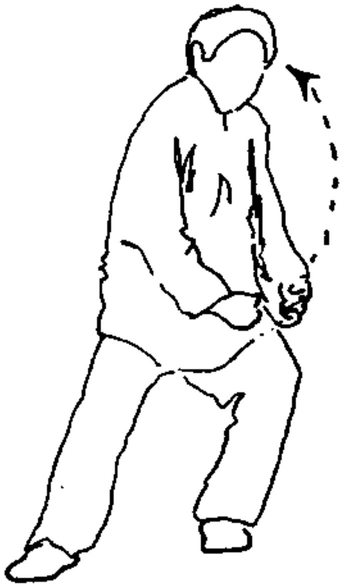

Figure 41.34. Expanding and Contracting the Rings (Forward-A)

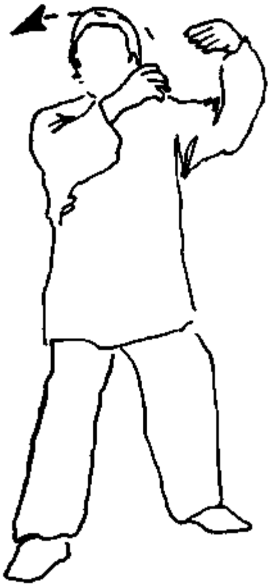

Figure 41.35. Expanding and Contracting the Rings (Forward-B)

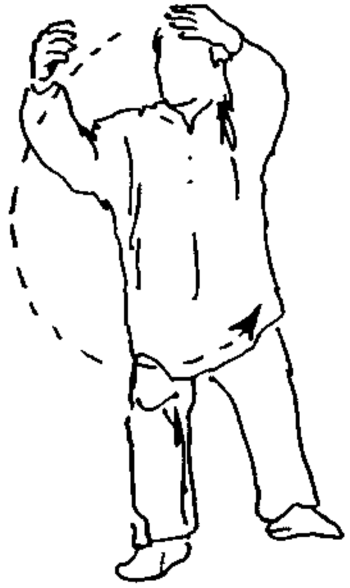

Figure 41.36. Expanding and Contracting the Rings (Forward-C)

tate up the center line of the chest. The rotation of the arms forms two circles, that circulate the energy from the inside to the outside of the breast. The focus of the intention should be on the "chest opening and the back closing" of the thorax. Repeat this exercise 18 to 36 rotations, depending on the severity of the condition (Figure 41.32).

Next, rotate the circles in the opposite direction, moving from the outside towards the center line of the chest and then down, circulating the energy from the outside to the inside of the breast. Repeat 18 to 36 rotations, depending on the severity of the condition (Figure 41.33).

When the rotations are completed, the middle fingers of both hands slightly press the Conception Vessel downward to descend the Qi into the Lower Dantian. Rub the abdomen three times and resume a Wuji posture.

EXPANDING AND CONTRACTING THE RINGS

Expanding and contracting the rings refers to the expansion and contraction of the spinal vertebra. This exercise reinforces the flow of the Microcosmic Orbit, invigorates and strengthens the Yuan Qi, aids in transforming Jing to Qi, and strengthens the entire spine. The patient's coordination and balance are also positively affected.

While performing this exercise, the body should move like a "dragon playing in the ocean waves" by appearing and disappearing, rising and sinking, and twisting and coiling. The goal is to stretch and loosen the spinal column, rib articulations, and the back. This will cause the body's Fire to descend and Water energy to ascend naturally through the body's center channels.

This exercise can be modified into a stationary position for convalescing patients or patient's confined to a wheelchair. The patient bends forward, allowing the spine to slowly stretch, shifting the body to a forty-five degree angle.

Begin this exercise from a right Bow Stance (weight is shifting 70% to the back, 30% forward) with both hands positioned in front of the Lower Dantian (Figure 41.34).

Both hands raise upward and stretch forward, creating an arc and stretching the spinal column. When extending and stretching the spine forward, imagine the Earth Qi traveling up the back leg, through the spine and Governing Vessel to the top of the head (Baihui point), and extending down to the Yin Tang and both Laogong points at the center of the palms (Figure 41.35).

While extending the energy outward, use intention to connect with the Heavenly Qi, the energy of the horizon, or a tree (Figure 41.36).

As both hands withdraw, shift the weight to the back leg, allowing the spinal column to contract. As you shift your weight back to the starting position, imagine drawing Qi up from the Earth into your body. Repeat nine times. End in the Wuji posture. Then switch directions by switching the position of the feet and repeat the entire sequence while reversing the rotation.

To perform the second half of the exercise, shift the body's weight forward and begin moving the body in the reverse circular direction (Figure 41.37).

Imagine the Heaven Qi traveling from the Baihui point down the Conception Vessel and out the front foot, extending deep into the Earth (Figure 41.38).

As the weight shifts back, imagine drawing Qi from the Heavens back into the body through the head and hands (Figure 41.39).

Sink the Qi deep into the Earth allowing the body to be cleansed and purified (Figure 41.40). Repeat the exercise nine times. End in the Wuji posture. Then switch directions and repeat the entire sequence.

DESCENDING THE YANG AND ASCENDING THE YIN TECHNIQUE

This exercise cultivates Qi in the channels and collaterals of the lower body, aids the free flow of Qi to the Lower Dantian, strengthens the Kidneys, Spleen, Stomach, and the Liver Qi, while regulating the Gall Bladder Qi.

It is an excellent exercise for pulling the Earth Qi (Spleen Qi) up while sinking the Stomach Qi, and is good for treating problems such as irritable bowel syndrome and problems caused from Excess Heat in the Middle and Lower Burners. In order to treat such conditions, the exercise should be performed for twenty minutes, twice a day.

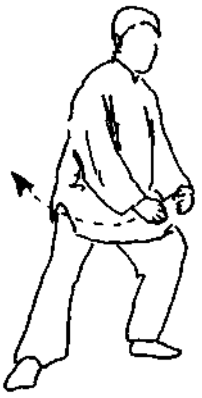

Figure 41.37. Expanding and Contracting the Rings (Reverse-A)

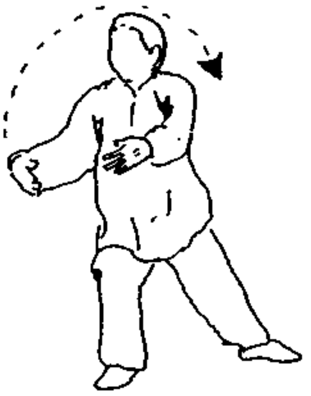

Figure 41.38. Expanding and Contracting the Rings (Reverse-B)

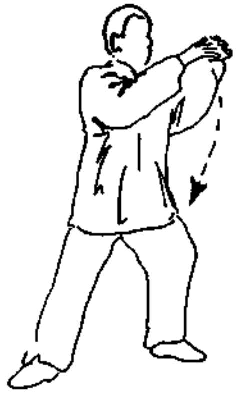

Figure 41.39. Expanding and Contracting the Rings (Reverse-C)

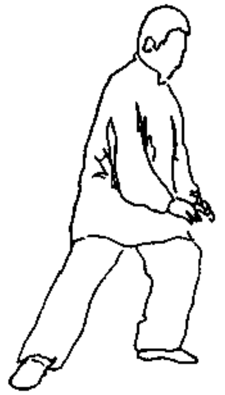

Figure 41.40. Expanding and Contracting the Rings (Reverse-D)

This exercise can furthermore be used in conjunction with the healing sound "Yu" to treat reproductive organ diseases (i.e., uterine and ovarian cysts, tumors, and cancer, as well as prostate cancer). The patient should practice a total of one to four hours each day, depending on the severity of the condition.

Begin from a Wuji posture and embrace the abdomen (Figure 41.41).

Gather Qi into the Lower Dantian and turn towards the left, both hands move into the Huantiao GB-30 point at the side of the hip (Figure 41.42).

Both hands slowly descend the three Yang channels down the outside of the left leg with a "pushing" action, ending on the outside of the left foot at the Bubbling Well point (Kd-1) at the bottom of the foot. As both hands reach downward to the foot, stretch the body and allow the waist to follow the descending Qi with a downward crouching movement (Figure 41.43).

Next, shift the hands around the foot and begin ascending the three Yin channels with the hands inside the left leg. Continue the movement all the way up into the Huiyin point, pubic bone, and Lower Dantian area with a "pulling" and "drawing" action. Pull the anal sphincter upward while drawing the Earth Qi into the body. When both hands raise and gather Qi upward, the body and waist must follow the ascending and rising movements (Figure 41.44).

Pause for a few seconds and let the Qi settle into the Lower Dantian (Figure 41.45).

Shift the body's weight, and repeat the entire sequence on the opposite leg. The turning of the waist should match the movements of the thighs. The toes should naturally grasp the ground and the soles of the feet should feel full of Qi. Repeat nine times on each leg, and end in a Wuji posture.

End the exercise by ascending and descending the Earth Qi through both legs.

From a Wuji posture, both hands embrace the abdomen to gather the Qi of the Lower Dantian and to separate the energy into two balls of energy (Figure 41.46).

Both hands then slowly descend the outside Yang channels of each thigh, all the way to the outside of the feet, ending at the Bubbling Well points (Figure 41.47).

Move the hands along the front of the feet, toward the inside of the foot, slowly ascending the inside Yin channels of each thigh all the way up the body into the Huiyin and Lower Dantian areas (Figure 41.48)

Pause and allow the Qi to settle in the Lower Dantian, then begin again (Figure 41.49). Repeat

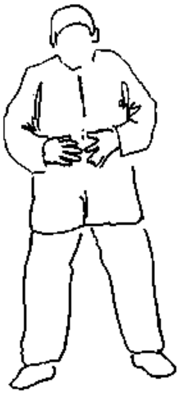

Figure 41.41. Descend the Yang and Ascend the Yin Technique (Single-A)

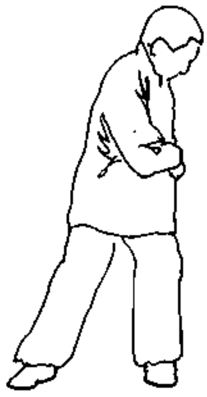

Figure 41.42. Descend the Yang and Ascend the Yin Technique (B)

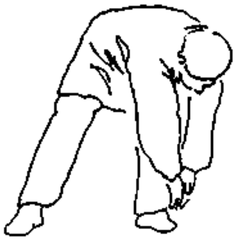

Figure 41.43. Descend the Yang and Ascend the Yin Technique (C)

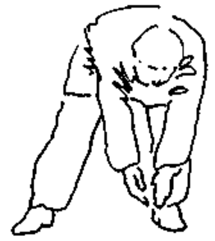

Figure 41.44. Descend the Yang and Ascend the Yin Technique (D)

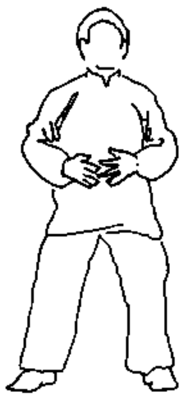

Figure 41.45. Descend the Yang and Ascend the Yin Technique (E)

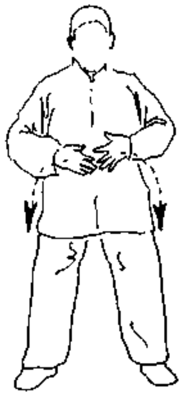

Figure 41.46. Descend the Yang and Ascend the Yin Technique (Double-A)

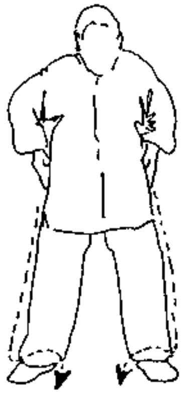

Figure 41.47. Descend the Yang and Ascend the Yin Technique (B)

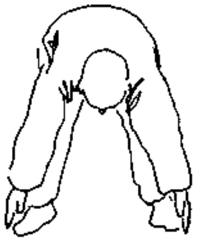

Figure 41.48. Descend the Yang and Ascend the Yin Technique (C)

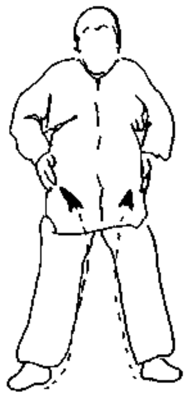

Figure 41.49. Descend the Yang and Ascend the Yin Technique (D)

this exercise 18 times, then return to the Wuji posture and end the exercise.

TURNING AND WINDING THE BELT VESSEL TECHNIQUE

This tonifying exercise is divided into two methods: the Small Turning and Winding, and the Large Turning and Winding techniques. Each method of exercise causes a different action between the rotation of the energy's circle, the form's circle, and the internal turning action of the Lower Dantian.

Practice the Small Turning and Winding technique first, then proceed to the Large Turning and Winding technique. The small method uses a 180 degree semicircular rotation in addition to the Belt Vessel and Lower Dantian circle rotation.

When practicing the Large Turning and Winding method, the patient's mind must imagine a 360 degree circle rotation of the Belt Vessel. The body moves externally in one direction, while the energy in the Lower Dantian rotates in the reverse direction.

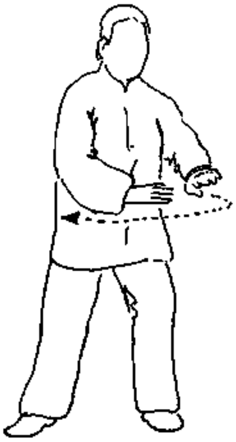

Figure 41.50. Small Turning and Winding (A)

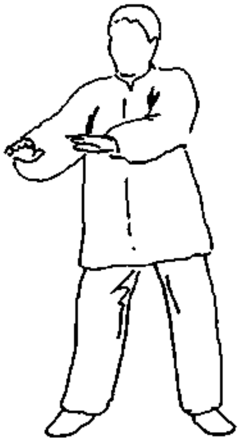

Figure 41.51. Small Turning and Winding (B)

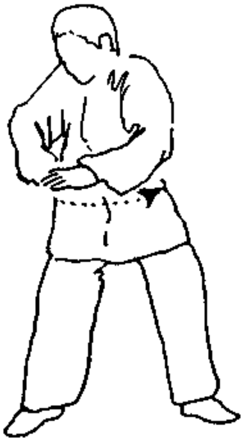

Figure 41.52. Small Turning and Winding (C)

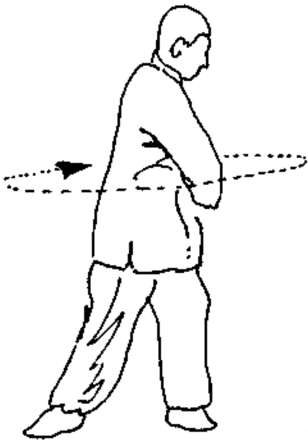

Figure 41.53. Large Turning and Winding (A)

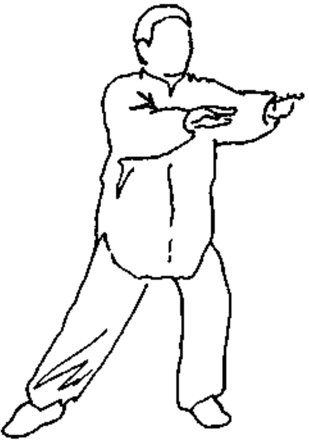

Figure 41.54. Large Turning and Winding (B)

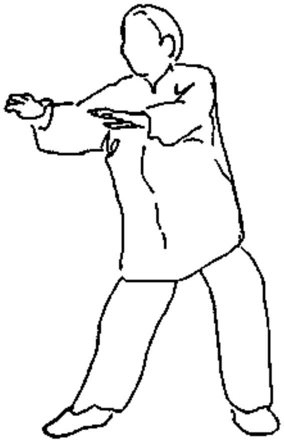

Figure 41.55. Large Turning and Winding (C)

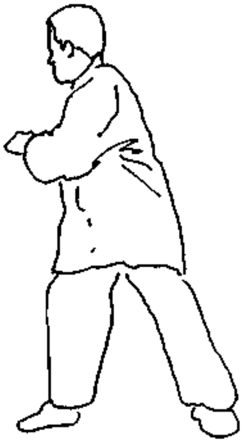

Figure 41.56. Large Turning and Winding (D)

The back and forth, side to side movements of the Turning and Winding action move the body's Qi, and stimulate the energy moving within the Belt Vessel's Qi circle. This energy circle fuses with the external rotation of the hands, allowing the energy from both external and internal Qi circles to be absorbed into the Lower Dantian. Once absorbed into the Lower Dantian, the Qi expands into the body's center core and blends with the Yuan Qi. This strengthens the waist, spine, and enables the Middle Burner's Qi to pervade through the upper and lower parts of the body. This exercise also develops strong and vigorous Kidneys and thus promotes the transportation of the Kidney's Jing and Qi.

SMALL TURNING AND WINDING TECHNIQUE

From a Wuji posture, step forward with the left foot to form a Leaning Horse stance. Raise both palms to waist level, in front of the Lower Dantian, with the palms facing the ground (Figure 41.50).

Inhale and shift the body forward towards the right direction and begin to rotate, turning and twisting the body in a clockwise direction. Imagine gathering the Earth's Qi into the Lower Dantian (Figure 41.51)

Exhale, but allow the mind to focus on absorbing the Earth Qi into the Lower Dantian. Pause for a minute, allowing the Qi to settle after filling the Lower Dantian (Figure 41.52). After nine circle rotations switch directions, moving in the opposite direction nine times.

LARGE TURNING AND WINDING TECHNIQUE

Bring the hands to the right side of the Mingmen and right Kidney area, while placing the mind's intention onto the Lower Dantian and navel area (Figure 41.53).

Inhale, as you extend both hands forward, and begin gathering the Earth's Qi (Figure 41.54).

Begin to circle rotate a 270 degree turn towards the left, while internally the energy of the Lower Dantian circles in the opposite direction to the right (Figure 41.55). Imagine the navel to be the center of the external circle and the Belt Vessel to be the internal circle. The mind's intention leads the energy which guides the hands and moves the body and the waist. The hand movements and center of weight shifting must be synchronized.

Exhale and gather the Earth's Qi into the left Kidney and Mingmen area; repeat nine times. Allow the Qi to fill the Mingmen and Kidneys, then switch directions for nine times (Figure 41.56). End in the Wuji position.

OPENING AND CLOSING THE HEAVEN AND EARTH TECHNIQUE

The purpose of this exercise is to cultivate and strengthen the Lower Dantian energy by blending it together with the Heaven and Earth Qi using the Beating and Drumming the Qi method of Qi cultivation. This cultivation method roots and stabilizes the body, causing the Five Hearts to return to their origin. The Five Hearts (sometimes called the Five Palms) are as follows:

- The Heart of the Baihui area (the Crown Chakra)
- The Hearts of each Laogong area (the palm Chakra)
- The Hearts of each Yongquan area (the bottom of each foot Chakra)

When the energy of the Lower Dantian disperses through the body, as the Mingmen pushes back, the Yuan Qi returns and gathers into the navel. This gathering and extending, back and forth action of energy of the Five Hearts transforms the body's Yuan Qi, strengthens the Yin and Yang energy, opens the body's channels, and encourages the merging of the Heart's Fire and the Kidneys' Water Qi. This method is divided into two techniques: Open the Heaven and Close the Earth, and Open the Earth and Close the Heaven techniques.

OPEN THE HEAVEN AND CLOSE THE EARTH

Begin from a Wuji posture with the left hand on the Lower Dantian and the right hand resting on top of it, both Laogong points face the navel (Figure 41.57).

Keep the mind's intention focused on the navel as the center base. Inhale as both arms raise up the center line of the body (Figure 41.58).

Extend the intention into the Heavens and begin to gather the universal Qi (Figure 41.59). Imagine Heavenly Qi descending and completely

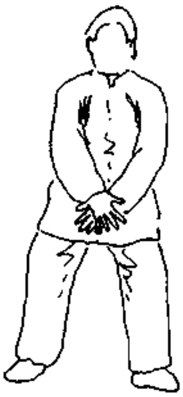

Figure 41.57. Open the Heavens and Close the Earth (A)

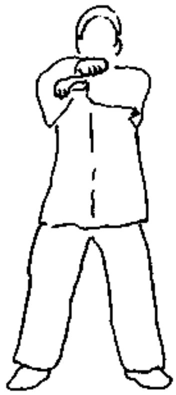

Figure 41.58. Open the Heavens and Close the Earth (B)

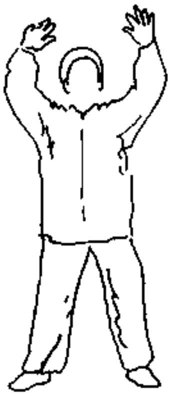

Figure 41.59. Open the Heavens and Close the Earth (C)

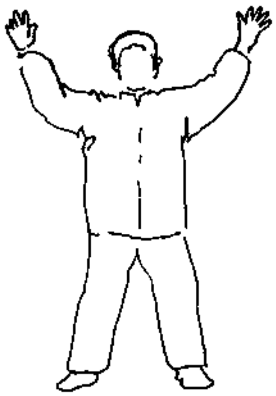

Figure 41.60. Open the Heavens and Close the Earth (D)

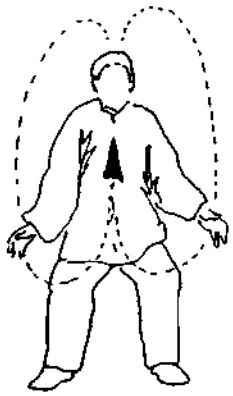

Figure 41.61. Open the Heavens and Close the Earth (E)

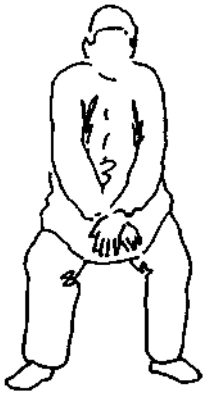

Figure 41.62. Open the Heavens and Close the Earth (F)

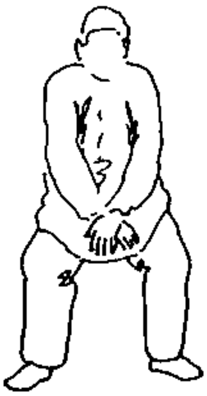

Figure 41.63. Close the Heavens and Open the Earth (A)

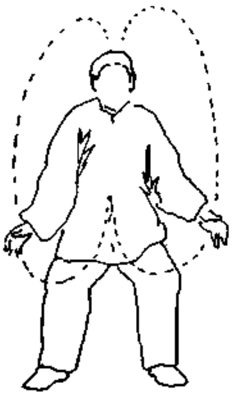

Figure 41.64. Close the Heavens and Open the Earth (B)

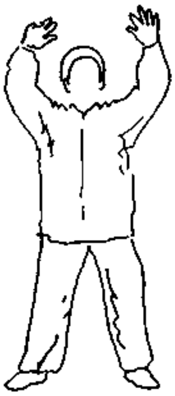

Figure 41.65. Close the Heavens and Open the Earth (C)

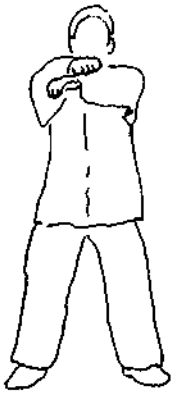

Figure 41.66. Close the Heavens and Open the Earth (D)

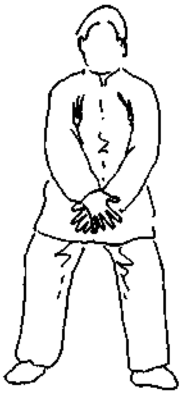

Figure 41.67. Close the Heavens and Open the Earth (E)

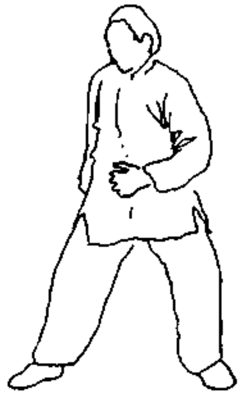

Figure 41.68. Collect the Qi and Return it to the Lower Dantian (A)

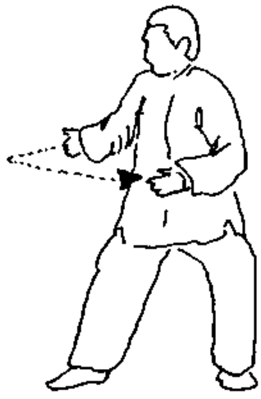

Figure 41.69. Collect the Qi and Return it to the Lower Dantian (B)

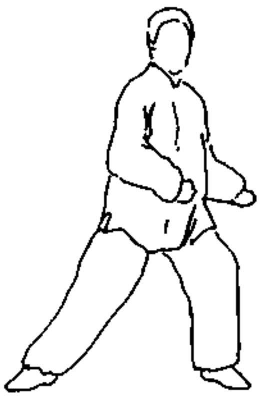

Figure 41.70. Collect the Qi and Return it to the Lower Dantian (C)

filling the body (Figure 41.60). Exhale and gather the Heavenly Qi deep into the body's center core (Figure 41.61).

Lead the energy from inside of the body's center core to the outside, imagining Heavenly Qi, like steam, releasing through the pores and filling up the body's external energy bubble like a mist (Figure 41.62). Repeat 18 times.

OPEN THE EARTH AND CLOSE THE HEAVENS

This is the exact reverse action of the Open the Heaven and Close the Earth technique. Begin from a Wuji posture with the left hand on the Lower Dantian and the right hand resting on top of it, both Laogong points face the navel (Figure 41.63).

Inhale and begin to separate the hands, moving them to the outside of the body (Figure 41.64). Imagine gathering the Earth's Qi into the body as the hands raise upwards towards the head (Figure 41.65). Exhale and gather the Earth's Qi deep into the body's center core (Figure 41.66).

Lead the energy from inside of the body's center core to the outside, imagining the Earth's Qi, like steam, releasing through the pores and filling up the body's external energy bubble like a mist (Figure 41.67). Repeat 18 times, then pause and relax, returning to the Wuji posture.

COLLECT THE QI AND RETURN TO THE LOWER DANTIAN TECHNIQUE

The purpose of this exercise is to gather both universal and environmental energy and to store it in the navel. This exercise is used for treating Deficient conditions.

COLLECT AND RETURN

Begin from a Wuji posture and step out to the right, forming a Horse Stance. The left hand swings to embrace the navel and Lower Dantian as the body shifts its weight and twists toward the right (Figure 41.68).

The body slowly begins to shift from side to side while the opposite hand gathers the universal and environmental energy, gathering and packing it into the navel and Lower Dantian area (Figure 41.69).

During the exchange of left and right hands collecting Qi, imagine that the body is swimming in water. Imagine the energetic consistency of the air, like water, following each arm as it moves, flowing into the center of a whirlpool located in the navel and Lower Dantian area (Figure 41.70). Repeat 18 times. End in Wuji posture.

QIGONG SELF-HEALING MASSAGE AND POINT THERAPY

Qigong Massage and Point therapy is usually practiced after the Qigong meditations and exercise prescriptions. It involves having the patient lightly massage or tap the skin, while extending Qi into the various channels, points, and tissues. These self-healing massage techniques are specifically designed to disperse stagnations and Excess conditions, as well as to activate and regulate the energetic flow within the body's external and internal channels, thereby increasing the energy flow.

These massage techniques can be practiced from either a lying, sitting, or standing posture.

WASH THE FACE

This exercise stimulates the skin and facial muscles. Begin from a Wuji posture. Shake the hands vigorously to remove any External pathogenic Qi. Then slap the palms and rub the hands together vigorously until they get hot. Gently place both hands over the eyes and inhale, absorbing the heat into the face and Upper Dantian. While inhaling, allow the Qi to enter the Taiji Pole.

Exhale and imagine the energy gathered in the Upper Dantian descending down the center of the body, flowing into the Lower Dantian. Perform this gathering and descending action three times.

Next, both hands begin to wash the face, in an outward circular motion. Allow the hands to massage the face, up from the nose, eyes, and forehead, down along the temples, cheeks, and across the chin and mouth, returning back to the nose. Perform this face massage technique for 9 breaths.

Shake the hands vigorously to remove any External pathogenic Qi, then proceed to the next exercise.

MASSAGE THE HEAD

Guide all ten fingers over the head and down over the back of the neck (this is called Ten Dragons Run Through the Forest). Repeat nine times. This can be performed either by touching or not touching the head (Figure 41.71).

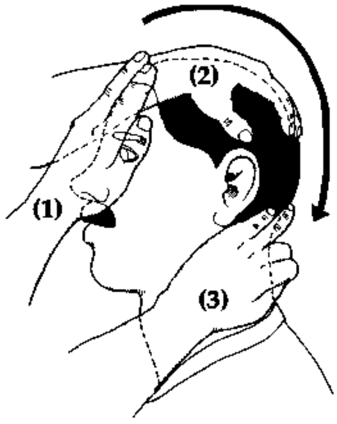

Figure 41.71. Wash the Face and Massage the Head: After rubbing the hands to heat the palms, place both hands on the face and draw the heat in through the eyes and into the Upper Dantian (1). Next, comb the hair with the fingertips (2), and drain the excess Qi from the head and neck area (3).

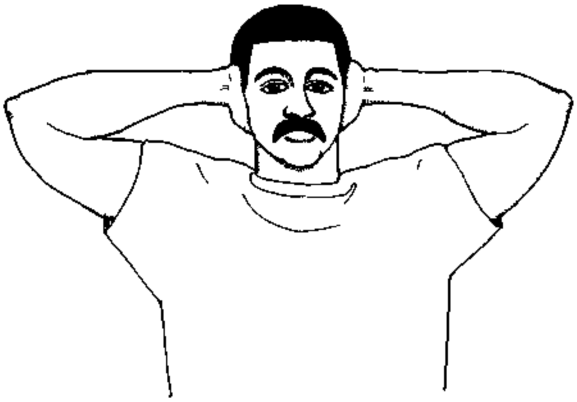

Figure 41.72. Beating the Heavenly Drum

BEATING THE HEAVENLY DRUM

This exercise balances the air pressure in the eustachian tubes and ear canals, relieves tinnitus, helps prevent vertigo, relaxes the mind, helps to improve hearing, expels stagnant Qi collected at the back of the head (within the Jade Pillow), and also stimulates the pineal gland and Kidneys' Qi.

Begin by interlacing the fingers and place both hands on the back of the head. The palm of each hand should cover the ears. While compressing the ears, close your eyes and practice the Beating and Drumming thumping method at the back of the head in the rhythmic sequence of a heart beat. This causes relaxation and creates an altered, in-utero state of consciousness. The index fingers of each hand should thump the Jade Pillow (Bl-9) at the back of the occiput pass 36 times (Figure 41.72).

After thumping the back of the head, clench the teeth, close the mouth, and bend the body over. Exhale while bending over and bring the head between the legs. The legs are straight, waist and buttocks are relaxed, and both eyes look towards the back from in-between the legs, while holding the breath (Figure 41.73).

Next, raise the body up, inhale and again perform the Beating the Heavenly Drum sequence. Practice this exercise 14 times.

PRESSING THE EARS

This exercise is used to enhance hearing, stimulate mental energy and also to stimulate the Kidneys' Qi. Use the Laogong areas at the center of each palm to compress the ears (like a plunger). Repeat this exercise nine times (Figure 41.74).

Next, move the right hand behind the head and pull the left ear (at a 45 degree angle) towards the back, with the left hand is placed on the Lower Dantian. Hold this posture for 9 breaths, then reverse for 9 breaths.

RUBBING THE EYES

RUBBING THE EYES (SEQUENCE #1)

This exercise is good for enhancing Qi and Blood circulation to the eyes for improving vision, treating eye diseases (especially glaucoma), relieving fatigue (caused by eye strain), and also for stimulating the Liver Qi.

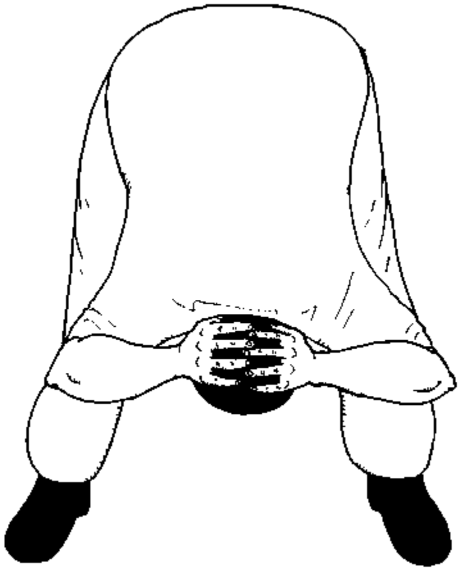

Figure 41.73. Bend the Body Over While Holding the Breath

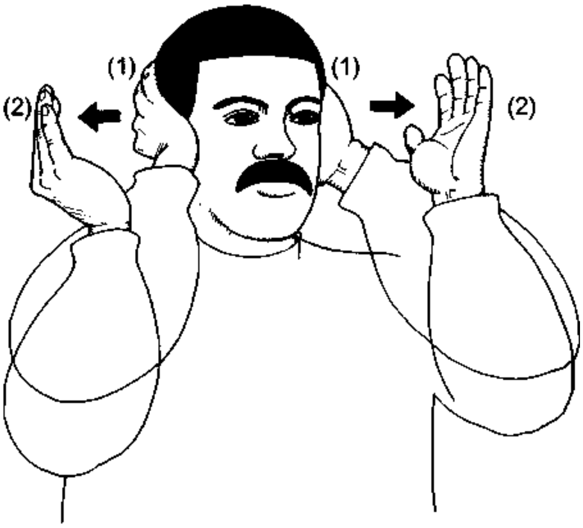

Figure 41.74. Pressing the Ears

With both thumbs resting on the cheekbones, form two soft fists and, using the edge of the index knuckles, begin to massage in a circle the top of the eyelids (from the bridge of the nose out along the eyebrows). Next massage just bellow the bottom of the eyelids from the bridge of the nose. Alternate from top to bottom 18 times (Figure 41.75).

RUBBING THE EYES (SEQUENCE #2)

With the thumbs still resting on the cheekbones, use the tip of the index fingers to massage the points around the eyes, from the inner corners of the skin, in a circular motion. Start rotating in a small circle, at the Bl-1 point. Circle nine times towards the inside, then nine times towards the outside, for a total of three breaths.

Next massage the middle of the eyebrows at the Extra Point Yu Yao. Then proceed to the outside of the eyes at the GB-1 point. (3) End at the base of the eyes at the St-1 point. Each area should be massaged 18 times (9 times towards the inside, then 9 times towards the outside) (Figure 41.76).

MASSAGING THE NOSE

This exercise is performed to treat colds, reduce swelling of the sinus tissues, eliminate mucus from the airways, stimulate the sinus nerves, open the nasal passages, and also stimulate the Lungs' Qi.

Begin by making soft fists and use the sides of the thumbs to massage the sides of the nose at the LI-20 points in an up and down motion 9 times pressing inward, and then 9 times pressing outward (Figure 41.77).

Next, use the tips of the index fingers and press deeply at the base of the nose at the LI-20 points, for a period of ten seconds, then circle-massage for 18 rotations. Press deeply midway up the nose, on both sides at the Bitong points for a period of ten seconds, then circle-massage 18 times. Finally, both fingers press deeply at the upper point at the bridge of the nose at the Shangen points for ten seconds, then circle-massage 18 times.

CLICKING THE TEETH

This exercise is good for strengthening the teeth and gums, tonifying the jaw muscles, im-

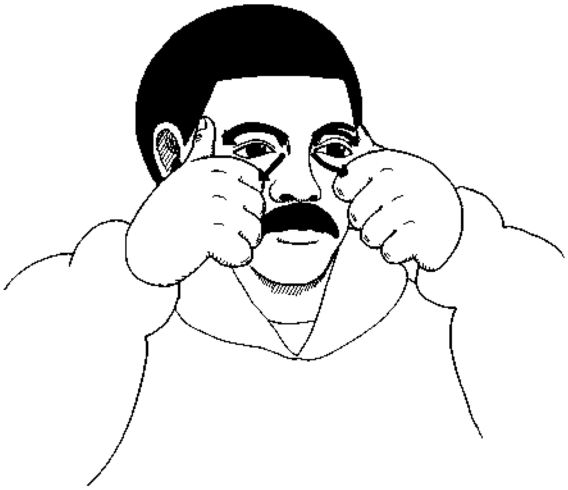

Figure 41.75. Rubbing the Eyes (Sequence 1)

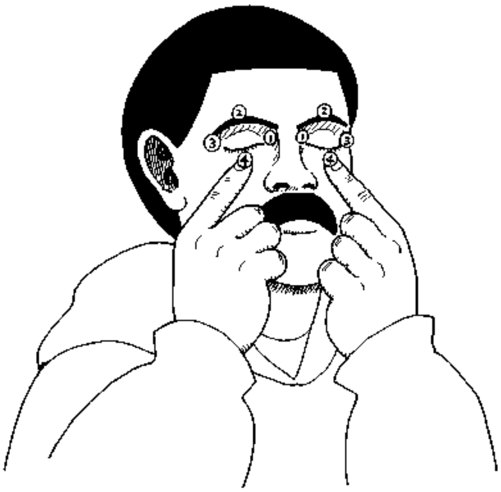

Figure 41.76. Rubbing the Eyes (Sequence 2)

proving the Qi and Blood circulation to the gum tissues, and also to enhance the Kidneys' Qi. Clamp the jaws shut so that the teeth are pressing together and the jaw muscles flex. Click the teeth 36 times.

WAGGING THE TONGUE

This exercise is good for stimulating the salivary ducts beneath the tongue, and also for stimulating the Heart's Qi. Wag the tongue on the outside gums 18 times in one direction, 18 times in the opposite direction, then 18 times from the lower to upper palate, and finally, 18 times pressing the tip of the tongue against the back of the teeth (Figure 41.78).

SWALLOWING THE SALIVA

Saliva contains many kinds of enzymes which aid in digestion. In ancient China, this practice was sometimes called the Immortal's Water or Juice of Jade, because the saliva is energized during the Qigong meditation and believed to contain special healing properties.

To begin, inhale and swallow the saliva with a big gulp (three times), using the breath as a seal or bubble to enclose the air, then send it down the Conception Vessel to bathe the Five Yin Organs, ending at the Lower Dantian. It is important to tilt the head slightly forward (after inhaling) before swallowing the Immortal's Water.

WHIRLING THE NAPE AND TURNING THE NECK

The purpose of this exercise is to massage the thyroid and larynx as well as to stretch the muscles and tendons throughout the neck and cervical vertebrae.

To begin, place the back of the hands on the Kidneys. While relaxing, allow the head to relax and naturally hang, slowly rotating its motion to the left and right nine times (Figure 41.79).

Next, suspend the head and slowly twist the nape of the neck to the right and left direction nine times. While performing this exercise, focus the mind's intention on the Lower Dantian. When finishing, focus the mind's attention on the Bubbling Well points at the center of the feet.

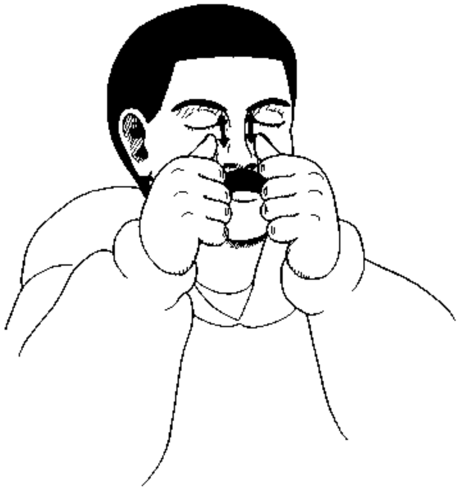

Figure 41.77. Massaging the Nose

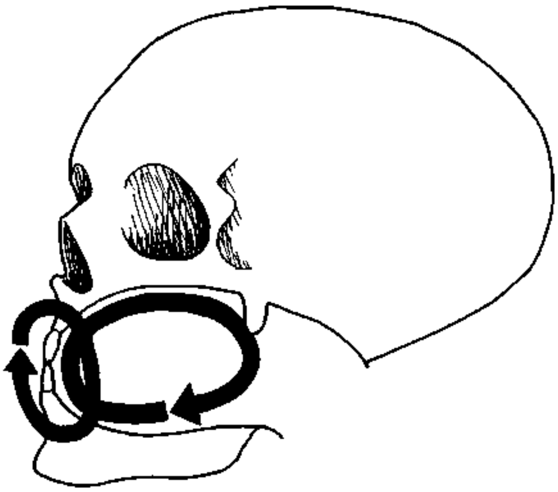

Figure 41.78. Wagging the Tongue

MASSAGING THE CHEST AND ABDOMEN

The purpose of this exercise is to disperse Excess Qi from the chest and abdominal area. This allows the body to balance the energy of the upper torso with that of the lower torso.

Beginning with the mind's attention focused on the Laogong areas at the center of each palm (Pc-8), brush the chest and abdomen downward nine times, starting from the collar bone and ending by the pubic bone (Figure 41.80).

MASSAGE TAPPING THE HANDS (LI-4)

Stimulating the Hegu or Joining of the Valleys points. The purpose of this exercise is to clear and activate the channels and collaterals, regulate, tonify and promote the circulation of Qi, to dispel pathogenic Wind, clear Fire and Heat, and to tranquilize the mind.

Begin by extending both arms in front of the body and allow both hands to collide into each other, thus tapping the Hegu points located between the thumb and first finger. Repeat nine times (Figure 41.81).

MASSAGE TAPPING THE FOREARMS (PC-6)

Stimulating the Neiquan or Inner Border Gate points. The purpose of this exercise is to regulate and tonify the Qi and Blood of the Heart and Stomach, to clear the Heart Fire, transform the Heart Phlegm, and to calm the Heart.

Begin by forming a hollow fist with the right hand and then lightly tap the Neiquan point located three fingers up from the wrist fold, on the lower left forearm area, nine times. Switch arms and repeat (Figure 41.82).

MASSAGE TAPPING THE ELBOWS (LI-11)

This exercise stimulates the Quchi or Crooked Pond points. The purpose of this exercise is to help dispel pathogenic Wind from the body, to regulate the Stomach and intestines, eliminate stasis, relax the muscles and tendons, lubricate the joints, and to cool the Heat in the Blood.

To begin, the right hand forms a hollow fist

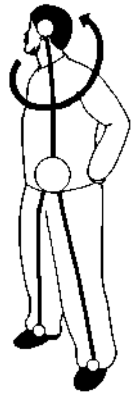

Figure 41.79. Whirling the Nape and Turning the Head

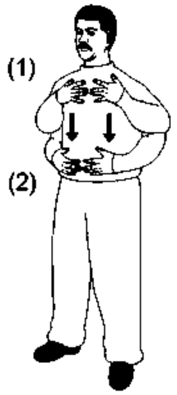

Figure 41.80. Massaging the Chest and Abdomen

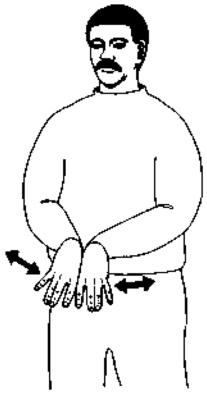

Figure 41.81. Massage Tapping LI-4

and lightly taps the Quchi point located at the end of the left elbow-fold, nine times. Switch arms and repeat (Figure 41.83).

MASSAGING UNDER THE ARMS (HT-1)

This exercise stimulates the Jiquan or Supreme Spring points. The purpose of this exercise is to regulate the Qi of the Lungs and Heart, promoting the circulation of Qi and Blood, as well as to regulate the blood vessels.

To begin, the right hand reaches under the left arm pit and circle massages nine times. Switch sides and repeat nine times (Figure 41.84).

MASSAGE TAPPING THE UPPER SHOULDERS (GB-21)

This exercise stimulates the Jianjing or Shoulder Well points. The purpose of this exercise is to regulate the body's Qi activities, disperse Liver Qi stagnation, extinguish Liver Wind, dispel Wind and Cold, redirect Rebellious Qi downward, and to dredge Excess Qi from all the Yang channels.

To begin, the right hand forms a hollow fist or soft palm and taps the left Jianjing point located on the upper ridge of the neck, while the left hand alternately taps the right Jianjing point. Both left and right tapping count as one set. Perform nine times (Figure 41.85).

MASSAGING THE ARMS

This exercise stimulates the six channels of the arms (the three Yang channels located on the outside, three Yin channels located on the inside).

Begin by placing the right palm on the inside of the left shoulder. Exhale and rub the entire arm from inside the shoulder and arm to the fingers. Next inhale and rub the entire outside of the arm from the fingers to the outside of the shoulder. Continue this movement for 18 breaths, then repeat on the other side (Figure 41.86).

MASSAGE TAPPING THE LOWER DANTIAN (CV-6)

This exercise stimulates the Qihai or Sea of Qi points. The purpose of this exercise is to energize and tonify the lower abdominal organs, as well as to stimulate the Qi of the Lower Dantian.

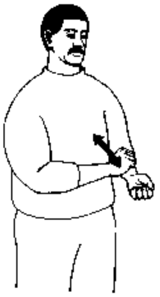

Figure 41.82. Massage Tapping Pc-6

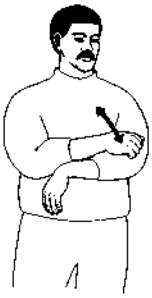

Figure 41.83. Massage Tapping LI-11

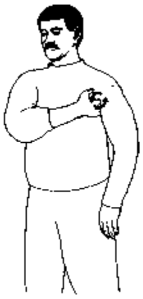

Figure 41.84. Massaging Under the Arm Ht-1

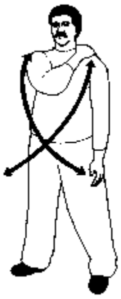

Figure 41.85. Massage Tapping GB-21

Begin by stroking the chest, gathering the body's Qi down into the Lower Dantian. Using either soft fists, palms, or fingers, both hands lightly tap the Lower Dantian and navel area for three breaths (Figure 41.87).

MASSAGE TAPPING THE BACK OF THE WAIST (GV-4)

This exercise stimulates the Mingmen or Gate of Life point. This exercise stimulates and tonifies the Kidneys, adrenal glands, and the Mingmen. It is used to regulate the Governing Vessel, excite and strengthen Kidneys' Yang Qi, to replenish the Kidneys, as well as to strengthen the lower back and bones.

Begin by slightly leaning the body forward. Using the right and left soft palms or hollow fists, lightly tap the back of the waist, Mingmen and the Kidneys' area for three breaths (Figure 41.88).

MASSAGE TAPPING THE SIDES OF THE HIPS (GB-30)

This exercise stimulates the Huantiao or Jumping Circle points. The purpose of this exercise is to dispel Wind and Cold, reduce pain in the thighs and legs, as well as to strengthen the lower back, and tonify the lower extremities.

Using the right and left hollow fists or soft palms, lightly tap the hip bone Huantiao points nine times (Figure 41.89).

MASSAGE TAPPING THE THIGHS (GB-31)

This exercise stimulates the Fengshi or Wind's Market points. The purpose of this exercise is to dispel Wind and Cold, clear Heat, tonify the legs, reduce flaccidity and numbness of the lower extremities, and to transform Dampness.

Using the right and left hollow fists or soft palms, lightly tap the upper thighs at the Fengshi Points nine times (Figure 41.90).

MASSAGE TAPPING THE OUTSIDE OF THE KNEES (GB-34)

This exercise stimulates the Yanglingquan or Yang Mound Spring points. The purpose of this exercise is to relax the muscles and tendons, acti-

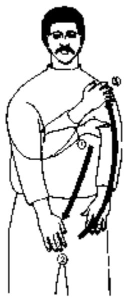

Figure 41.86. Massaging the Arm Channels

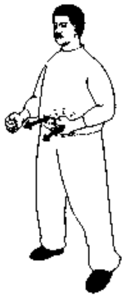

Figure 41.87. Massage Tapping the Lower Dantian

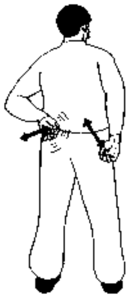

Figure 41.88. Massage Tapping the Mingmen GV-4

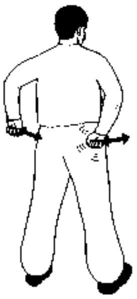

Figure 41.89. Massage Tapping GB-30

vate the flow of Qi and Blood in the channels and collaterals, relieve spasms and pain, clear away Damp-Heat, and disperse stagnant Liver and Gall Bladder Qi.

Using the right and left hollow fists or soft palms, lightly tap the outside of the knees at the Yanglingquan points nine times (Figure 41.91).

MASSAGE TAPPING THE INSIDE OF THE KNEES (SP-9)

This exercise stimulates the Yinlingquan or Yin Mound Spring points. The purpose of this exercise is to regulate and tonify the Spleen (Yang), Stomach (Yin), and Lower Burner, to resolve Dampness and Damp Heat, as well as to treat knee pain and problems with difficulty in urination.

Using the right and left hollow fists or soft palms, lightly tap the inside of the knees at the Yinlingquan points nine times (Figure 41.92).

MASSAGE TAPPING THE BACK OF THE KNEES (BL-40)

This exercise stimulates the Weizong or Entrusting Middle points. The purpose of this exercise is to relax the muscles and tendons, activate the flow of Qi and Blood in the channels and collaterals, cool the Blood to stop bleeding, transform Damp Heat in the Bladder and intestines and to strengthen the waist and knees.

Using the right and left hollow fists or soft palms, lightly tap the back of the knees at the Weizhong points nine times (Figure 41.93).

MASSAGE TAPPING BELOW THE KNEES (ST-36)

This exercise stimulates the Zusanli or Foot Three Miles points. The purpose of this exercise is to regulate the Stomach and intestines, facilitate the flow of Qi to relieve bloating and distention, to dissipate stagnation and obstructions, to promote the circulation of Ying Qi, to alleviate water retention and subdue swelling, and to activate the flow of Qi and Blood in the channels and collaterals.

Using the right and left hollow fists or soft palms, lightly tap the lower leg Zusanli points nine times (Figure 41.94).

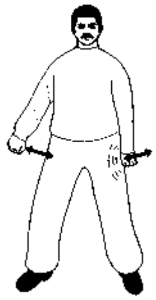

Figure 41.90. Massage Tapping GB-31

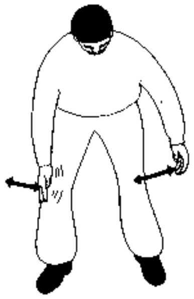

Figure 41.91. Massage Tapping GB-34

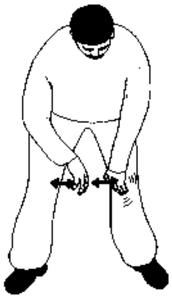

Figure 41.92. Massage Tapping Sp-9

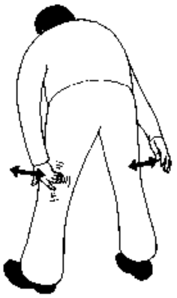

Figure 41.93. Massage Tapping Bl-40

MASSAGE TAPPING ABOVE THE MEDIAL MALLEOLUS (SP-6)

This exercise stimulates the Sanyinjiao or Three Yin Crossing points. The purpose of this exercise is to tonify and regulate the Qi and Yang of the Spleen, as well as regulate the Qi of the Stomach, Middle and Lower Burners, reduce digestive stagnation, and to facilitate Blood flow.

Using the right and left hollow fists or soft palms, lightly tap the inside of the lower leg at the Sanyinjiao points nine times. (Figure 41.95).

MASSAGE TAPPING THE HEELS (KD-1)

This exercise stimulates the Yongquan or Bubbling Spring points. The purpose of this exercise is to tonify the Kidneys' Yin and Jing, tranquilize the mind, clear away Heat, purge pathogenic Fire, calm the Liver and to stop pathogenic Wind.

To begin this exercise make sure that the knees are slightly bent. Quickly rise and fall on both heels to stimulate the Yongquan points and shake the back nine times (Figure 41.96).

TREMBLING THE KNEES

The purpose of this exercise is to disperse stagnant Qi, as well as to relax the perineum.

To begin this exercise, relax, then bend and straighten the knees quickly for nine times. Next shake them from side to side quickly in order to tremble and shake the entire body. Repeat nine times (Figure 41.97).

CIRCLING THE ABDOMEN

The purpose of this exercise is to balance the Lower Dantian Qi, stimulate the Stomach and bowels, and enhance digestive peristalsis.

To begin, place the center of both palms on the Lower Dantian, embracing the navel (for men, the left palm is on the top, opposite for women). Focus the mind's intention deep into the Lower Dantian. While using the navel as the center, connect the energy of the Lower Dantian with the center of the palms as they rotate.

Begin to circle 36 times counterclockwise for men (from the patient's point of view), whereas the rotation is clockwise for women. Allow the circle to gradually move from a smaller to a larger

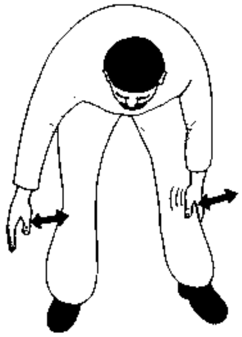

Figure 41.94. Massage Tapping St-36

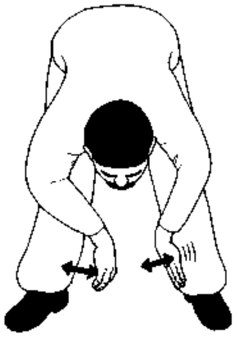

Figure 41.95. Massage Tapping Sp-6

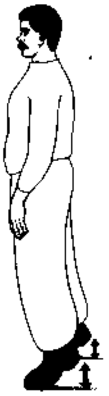

Figure 41.96. Dropping the Heels to Stimulate Kd-1

rotation (as the hands circle, imagine that they are dispersing the Qi). Pause and reverse the circle, rotating 24 times in the opposite direction, gradually reducing the circle's size. This will draw the Qi from the outer portion of the body and collect it into the Lower Dantian. Pause for a moment then return to the Wuji position (Figure 41.98).

COMPLETING THE EXERCISE

Upon completion of the exercise, close the Four Doors (the center of each palm and foot) and cultivate the Qi with stillness, allowing the energy to return to its origin. Hold a quiet Wuji postue for about 15 minutes.

The purpose for the first circle rotation is to release the Yuan Qi which has been collected and gathered in the Lower Dantian. This is to avoid any stagnation which may cause the energy flow to slow down. This also disperses and releases the Yuan Qi, allowing it to blend with the Essential Qi of the Five Yin Organs.

When performing the reverse or second circle rotation, allow all the dispersed energy to gather together and return back to the Lower Dantian. This accumulates and stores the organ energy, blending it together with the Original Qi.

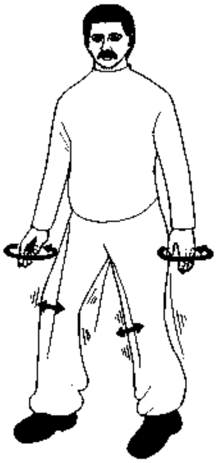

Figure 41.97. Trembling the Knees

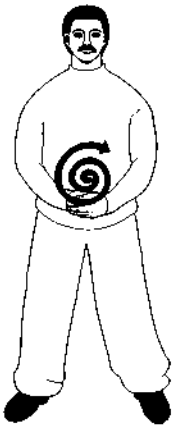

Figure 41.98. Circling the Abdomen to Finish

CHAPTER 42
DAOIST QIGONG REGULATION EXERCISES

This particular school of Daoist Medical Qigong Regulation was first introduced in the West by one of my teachers, Dr. Her Yue Wong. These techniques have become increasingly popular, as they are simple, yet extremely effective in tonifying and regulating the body's organ energy.

The interesting thing about Western healing practices is that they focus on the neuromuscular skeletal system to the exclusion of the visceral system, not recognizing that the visceral system is necessary to supply the fuel for the neuromuscular system. These following exercises focus on the visceral system and enhance the quality of life of the viscera.

The system as a whole is divided into two sections. The first section stimulates, energizes, and regulates the body's Yang organs and channels. The second section stimulates, energizes, and regulates the body's Yin organs and channels.

The exercises can be practiced as a complete system, or singled out individually and practiced as tonification prescriptions. As with any Medical Qigong prescription, the patient must purge his or her body of Toxic Qi before starting the tonification exercise and return the Qi back to the Lower Dantian upon completion.

DAOIST FIVE YANG ORGAN REGULATION EXERCISE

This exercise massages the five Yang digestive organs, strengthens the peristaltic action of the body's digestive system, as well as increases the capillary circulation through the stimulation of the autonomic nervous system. The Yang digestive organs (Stomach, Small Intestine, Large Intestine, Urinary Bladder, and Gall Bladder) are used by the body to release waste and Turbid Qi; therefore, it is important to keep the Yang organs strong and functioning smoothly.

The Beating and Drumming the Qi Breath Regulation (which is employed within this exercise) is beneficial for those patients who suffer from digestive problems and irregularity (i.e., spastic colon, chronic diarrhea, irritable bowel syndrome, etc.).

When the patient's arms swing rhythmically throughout the exercise, the nerves and Qi channels of the shoulder joints become stimulated. The swaying action of the arms increases the circulation in the energetic channels which connect the different internal organs and terminate at the hands.

Although this exercise is simple, the result in strengthening the body's peristaltic action and curing digestive illnesses is very effective. The patient is therefore cautioned not to eat prior to practicing this exercise.

SWAYING THE ARMS WHILE BEATING AND DRUMMING THE QI

Assume a wide stance with the arms suspended by the sides of the body. While inhaling, swing the arms up straight out in front of the body and then bring the hands in by the chest (Figure 42.1). Expand the abdomen fully focusing the mind's inten-

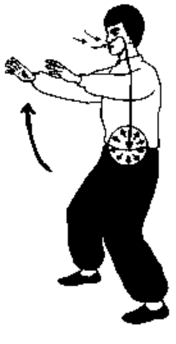

Figure 42.1. The Daoist Five Yang Organ exercise. Inhale and draw the Qi into the Lower Dantian while swinging the hands forward.

tion on filling the lower abdomen with Qi.

When exhaling, swing the arms back behind the body while compressing the abdomen down and inward directing the Qi to flow from the lower perineum, circulate back into the sacral area, then to the navel and back down to the perineum (Figure 42.2).

Continue to swing the arms back and forth at a pace that is comfortable using natural breathing. The abdomen should expand and contract without forcing the respiration. The patient must have the anal sphincter closed throughout the entire exercise. Continue this repetition for at least 50 breaths. Try to work up to 250 breaths or more for chronic conditions of abdominal obstruction and Qi stagnation within the digestive system.

The pressing and releasing action of the abdomen stimulates the energetic functions of the autonomic nervous system. The autonomic nervous system has two divisions:

1. The fight or flight mechanism of the sympathetic division -- which shuts down the digestive system due to the production of adrenaline or norepinephrine and redirects Qi and Blood flow to the brain and skeletal muscles, and
2. The vegetative, renewal mechanism of the parasympathetic division -- which draws Qi and Blood from the brain into the digestive system.

The centers for neuronal control of the autonomic nervous system and also control the digestion and tubal contraction are located in the cranium and in the sacrum. The accessory nerve and vagus nerve carries from the brain stem, its nucleus (or center), all of the information to supply the digestive system with impulses to stimulate peristaltic contraction. At the lower end of the digestive system, the sigmoid colon, rectum, Bladder and uterus are controlled by nerve fibers coming from the first, second, and third sacral foramina. The action of the cranial-sacral division of the parasympathetic nervous system is to stimulate peristaltic action (hence the movement of nutrients through the digestive system).

The movement of the abdominal cavity, diaphragm, and thoracic cavity changes the position of each organ and enhances its function. By tightening the anal sphincter and compressing and releasing the urogenital diaphragm, the reproductive organs are also stimulated, including the uterus, prostate, urethra, and testicles. Finally, the Daoist Five Yang Organ Regulation exercise also enhances the peristaltic action of the body's lymphatic system, circulatory system, and tubal systems.

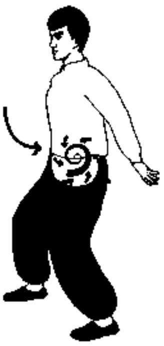

Figure 42.2. Exhale and compress the abdomen, allowing the hands to swing behind the body.

DAOIST FIVE YIN ORGAN TONIFICATION AND REGULATION EXERCISES

Each of the following exercises are designed to stimulate the specific internal Yin organ, as well as the paths of energy that flow from the channels of each organ. Since the Yin organs have the responsibility of storing energy, each exercise is named for the organ it tonifies.

The Yin exercises are performed in the sequence shown below without pausing in-between transitions. Unless specifically prescribed as a therapeutic modality, each exercise should flow directly into the next. The transitions are very important for continuity of energy flow. This particular Five Yin Organ series begins with an opening exercise and ends with a closing exercise.

OPENING, COLLECTING, AND MOVING THE QI

The opening exercise is designed to stretch and loosen the spinal vertebrae and to facilitate the energy flow through the patient's Governing Vessel. The spinal stretch also stimulates the flow of fluids and electrical impulses in the central nervous system (brain and spinal cord), which is nurtured by the cerebral spinal fluid. When the fluid starts to move, it awakens the nervous system causing more nutrients to enter the cellular structure of the nerves. The stretching, compressing, and pumping action of the spinal cord stimulates the movement receptors located in all of the ligaments and fascia energizing the body's entire proprioceptor mechanism.

The opening exercise is also designed to stretch the connective tissue, warm the body's intercellular matrix and tissues, stimulate the somatic system, free adhesions, as well as enhance the elasticity and compressibility of the ligamentous structure.

After finishing the Daoist Five Yang Organ Regulation exercise, return both hands to the Lower Dantian. Next, raise both hands above the head as you inhale (Figure 42.3).

Imagine energy (like a rushing river) flowing up from the feet, filling the legs, hips, waist, chest, arms, and head. Every square inch of the body is absorbing and being saturated with this Earth energy. When the entire body is completely full, the hands should still be positioned above the head (Figure 42.4).

Begin to exhale and feel the hands getting very heavy. The heaviness pulls the hands forward and then slowly starts to pull the body over. Slowly bend the head forward and feel the cervical vertebrae stretch. Let the hands pull the arms, which in turn will pull the shoulders and torso, followed by bending at the waist. The purpose of this movement is to feel each vertebra of the spine stretch sequentially so that a rippling effect descends down the spine. This will facilitate maximum stretch of the spine. While bending over, exhale and imagine the energy melting away from the entire body (like ice melting into water) slowly dissolving and rushing down through the feet and out into the ground (Figure 42.5).

Once completely bent over, imagine picking up a ball. Bend the knees and slowly stand up. While standing up, reverse the rippling of the spine from the coccyx, sacrum, and lower lumbar vertebrae to the base of the skull (Figure 42.6). Repeat these movements five times. Remember to inhale while raising the body, bring both arms over your head, and exhale as the body descends.

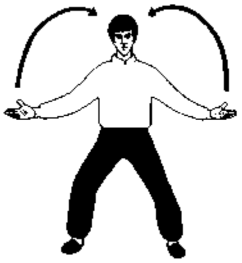

Figure 42.3. The Daoist Five Yin Organ Opening exercise. Begin by stretching the arms above the head.

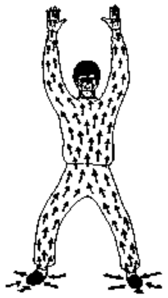

Figure 42.4. Inhale and imagine the Earth Qi like water filling the entire body.

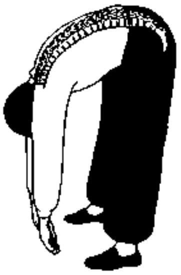

Figure 42.5. Exhale, bend the body and imagine any tension melting out the body into the Earth.

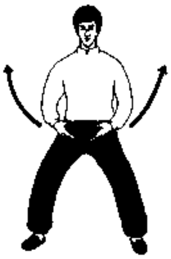

Figure 42.6. For the transition movement to the Lung Tonification exercise: inhale, bend the knees and slowly stand up.

LUNG TONIFICATION AND REGULATION EXERCISE

This exercise is designed to massage the Lungs' tissues and nerves, as well as tonify the entire Lung organs and channels. In the Lung exercise, both the sympathetic and parasympathetic divisions of the autonomic nervous system are stimulated. The compression and release action of the scapulae and thoracic vertebrae stimulate the nerves, arteries, veins and capillary system in the thoracic region.

Compressing and releasing of the muscles and blood vessels in the back and chest areas stimulates the transforming of the Qi and Blood and enhances the health of the individual tissue cells. The compression and release of the thorax also increases the stimulation of the nervous system in the Lung area, as well as stimulates the "release and return" action of the lymphatic system. All the major lymphatic ducts are located in the posterior portion of the thoracic wall.

The compression and release of the second, third, and fourth thoracic vertebrae stimulates the sympathetic preganglionic fibers from the spinal cord, which in turn stimulates the postganglionic fibers and the sympathetic outflow to the Lungs and bronchi. Patients practice this exercise to strengthen their respiratory system as well as increase the production of their Wei Qi to enhance their body's immune system.

Upon completion of the Opening exercise, the patient brings the hands above the head and places them in front of the torso at shoulder level while exhaling. The patient's palms should be facing downward parallel with the floor (Figure 42.7).

Inhale and bring the arms straight out to the sides (both hands should not go any higher than the shoulders) (Figure 42.8).

As the arms reach the sides of the body, rotate the palms until they face upward. Keep the shoulders stable and relaxed (Figure 42.9).

While exhaling, bring the arms straight out in front of the body returning to the beginning posture (Figure 42.10).

Allow the scapulas to gently push the arms forward. The sternum at this point is pushed inward so as to hollow the chest. Once the hands are aligned in the front of the shoulders, rotate the arms and

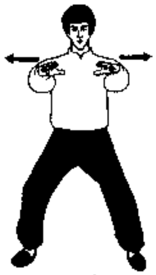

Figure 42.7. The Lungs Tonification exercise: Inhale as the hands separate.

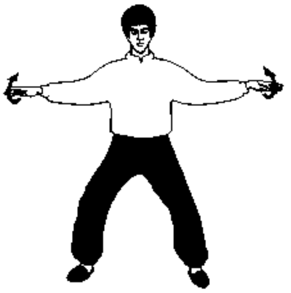

Figure 42.8. Separate the Hands and turn the palms upward.

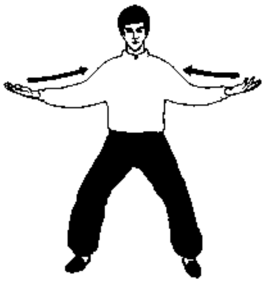

Figure 42.9. Exhale as you bring the arms toward each other.

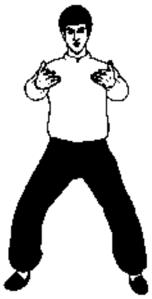

Figure 42.10. Draw the hands toward each other and turn the palms facing the Earth.

inhale to begin again. Repeat 20 times.

Note: Do not let the hands touch when bringing the arms back together towards the center line of the body. The body will short circuit itself, resulting in a mild shock or a sick feeling inside the thorax by the Heart.

KIDNEY TONIFICATION AND REGULATION EXERCISE

This exercise is designed to massage the Kidneys' tissues and nerves, as well as tonify the entire Kidney organs and channels. According to Western physiology, the Kidneys maintain the chemical, pH (potential of hydrogen), and temperature balance of the Blood. The Kidneys function like a rheostat changing the body's ability to absorb, retain, or release water according to the environment (e.g., retaining water in hot climates).

The twisting back and forth, pumping action of the exercise stimulates an opening and closing action within the Kidney organ. The Kidneys receive the major Blood supply from the abdominal aorta. In the Kidney exercise, the action of raising the torso upwards will literally suck Qi and Blood into the Kidney area.

Patients practice this exercise to strengthen their reproductive system, as well as increase the production of their Jing. It also helps balance the body's general energy condition, and can be used to treat sexual disorders and malfunctions.

Upon completing the Lung exercise, both arms are extending straight out in front of the body (Figure 42.11).

Inhale and drop the elbows and bring the hands down in front of the Lower Dantian (Figure 42.12).

Exhale and shift the weight to the left. As the left outer palm wraps around the back, resting itself on the right Kidney, the right palm extends towards the left direction (Figure 42.13).

The right palm, facing outward towards the left direction, begins to circle upwards in front of the face at eye level. As the palm begins to move, the eyes watch the back of the right palm (Figure 42.14).

Shift the weight towards the right side of the body and begin leaning towards the right direction. The eyes continue to follow the right palm as it begins to circle downwards in front of the body. Ex-

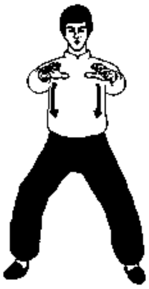

Figure 42.11. After completing the Lung exercise, center the body to prepare for the Kidney Tonification exercise.

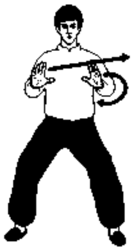

Figure 42.12. Preparation for the Kidney Tonification Exercise

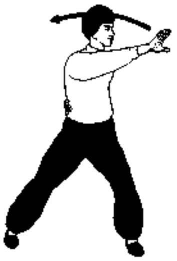

Figure 42.13. Kidney Tonification Exercise: exhale as the left outer palm wraps around the right kidney, and the right palm extends outward in a left direction.

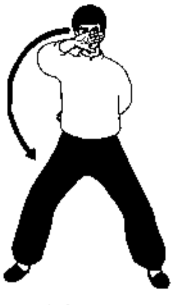

Figure 42.14. Begin to inhale as the right palm circles upwards and turns outward at eye level and begins to descend.

hale while shifting the weight (Figure 42.15).

Shift the weight, bend the knees and begin to lean over. While continuing to exhale, imagine scooping water with the right palm. Bend at the waist and scoop with the right arm. Keep the back relaxed and straight (Figure 42.16).

Allowing the right arm to follow the motion of the body, begin rising up slightly while turning forming an arc in front of the torso. As the palm raises, imagine the energy (which has been scooped into the arm) draining down the right arm across the shoulders and down the left arm into the right Kidney. Repeat this side for ten breaths (Figure 42.17).

After completing 10 breaths switch directions by shifting the weight onto the right leg (Figure 42.18 and Figure 42.19).

The right hand wraps the torso ending at the left Kidney and Mingmen area while the left palm is positioned in front of the face. Begin to inhale and repeat 10 times on the opposite side.

LIVER TONIFICATION AND REGULATION EXERCISE

This exercise is designed to massage the Liver's tissues and nerves, as well as tonify the entire Liver organ and channels.

According to Western physiology, the Liver is the body's master filter, collecting all of the nutrients that are being absorbed from the digestive system via the hepatic portal vein. The Liver functions like a series of channels. The cells of the channels are considered the processing factory for the entire body. If the body has a need for a particular nutrient, the Liver will create the enzymes for digestion, breaking down into components that which the body needs (or storing the components until needed). The Liver is also responsible for detoxifying, producing antibodies, and creating new cells to handle foreign substances recently absorbed by the body (storing substances of which it has no need).

The compression and release action of the torso causes a physical compression within the Liver organ itself, flushing Qi and Blood from the digestive system through the Liver and into the Gall Bladder where bile is stored.

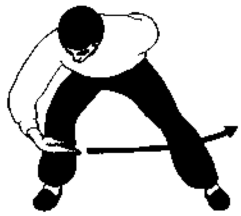

Figure 42.15. Continue inhaling as you bend over at the waist. Once your head sinks below your hips begin to exhale. Allow your right palm to circle downward to the left.

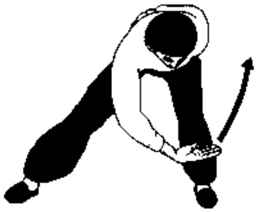

Figure 42.16. Continue exhaling as your weight shifts to the other foot and imagine the palm scooping up water as the body begins to circle upward.

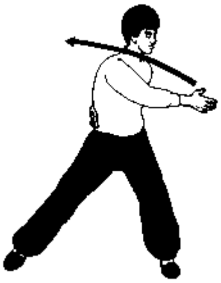

Figure 42.17. As the torso begins to arch upwards, allow the right arm to follow the body's movement. As the arm reaches the chest level, turn the outer palm towards the face and begin inhaling.

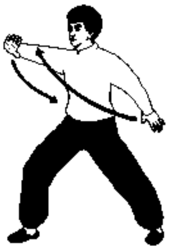

Figure 42.18. Preparation to switch hand positions: Shift the body's weight onto the right leg as the right arm begins to move to the back left Kidney and the left arm leaves the right Kidney to move forward.

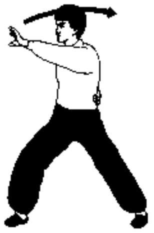

Figure 42.19. Kidney Regulation exercise final movement: begin to inhale as the left palm circles upwards and turns outward at eye level and begins to descend to the left.

Patients practice this exercise to strengthen their tendons and ligaments. This exercise helps to regulate the patient's emotional state.

Upon completing the Kidney exercise, begin to unwind the arm from behind the back and rotate the hips so they face forward (Figure 42.20).

Place the right arm straight out in front of the body, shoulder level, palm facing down. Place the left hand by the left hip, palm facing up (Figure 42.21).

Simultaneously draw the right palm back and extend the left palm forward while inhaling. The left hand moves forward passing the right hand which is moving backwards (Figure 42.22).

The hands continue in this motion extending outward from the center line of the body. Use long, slow inhalation and exhalation, inhaling and exhaling every three moves. Relax and continue to repeat this sequence for 20 breaths (Figure 42.23).

HEART TONIFICATION AND REGULATION

This exercise is designed to massage the Heart's tissues and nerves, as well as tonify the entire Heart organ and channels.

The compression and release of the skeletal muscles squeezing and releasing the cardiac muscle energetically stimulates the pericardium surrounding the heart. The twisting movement and compression of the torso also enhances the fluidity of all the connective tissues that are involved with the Heart. Every time the thoracic wall moves, expands, contracts, or twists, the tissues of the Heart are pulled, stretched, and released.

Patients practice this exercise to strengthen their circulatory system. This exercise helps to

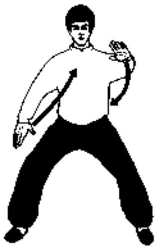

Figure 42.20. Preparation for the Liver Tonification Exercise: upon completing the Kidney exercise, begin to unwind the arm from behind the back and rotate the hips so they face forward.

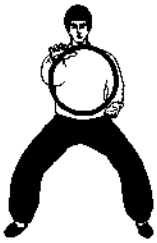

Figure 42.21. Liver Tonification Exercise: place the right arm straight out in front of the body, shoulder level, palm facing down. Place the left hand by the left hip, palm facing up.

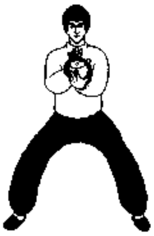

Figure 42.22. Simultaneously draw the right palm back and extend the left palm forward while inhaling. The left hand moves forward, passing the right hand which is moving backwards.

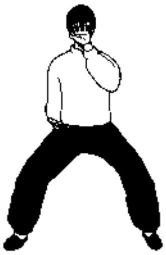

Figure 42.23. Use long, slow inhalation and exhalation, inhaling and exhaling every three moves.

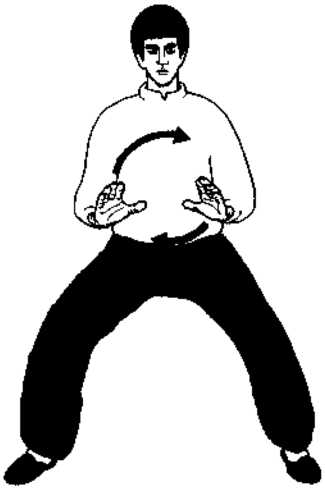

Figure 42.24. Preparation for the Heart Regulation Exercise: upon completing the Liver exercise allow both hands to sink down in front of the lower abdomen.

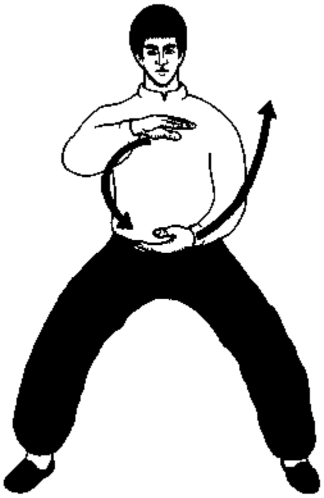

Figure 42.25. Heart Tonification Exercise: imagine the hands embracing a ball, placing the right hand on top, and the left hand on bottom: inhale.

stimulate the patient's intellectual activity and mental acuity or clarity.

Upon completing the Liver exercise allow both hands to sink down in front of the lower abdomen (Figure 42.24).

Imagine the hands embracing a ball, placing the right hand on top, and the left hand on bottom (Figure 42.25).

Exhale and twist the hips to the left, simultaneously raising the left hand up over the head while extending the right hand towards the left (Figure 42.26).

Continue pressing until both hands extend to the furthest point (Figure 42.27).

Inhale and turn the right palm upwards facing the sky as the body twists back to the center allowing the right hand to end up in front of the navel (Figure 42.28).

The left hand simultaneously turns so that the palm faces the center of the right hand following it to the center line of the body (Figure 42.29).

As the body reaches the point where it faces forward, twist to the right side and push in the same manner while exhaling. Repeat 10 times on each side (Figure 42.30).

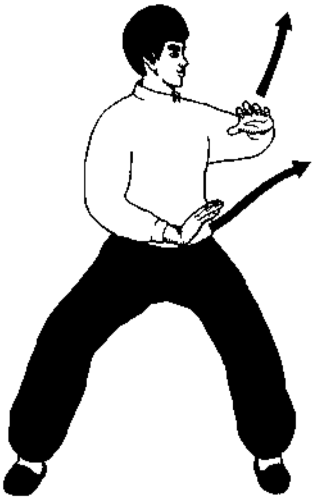

Figure 42.26. Exhale and twist the hips to the left simultaneously raising the left hand up over the head while extending the right hand towards the left.

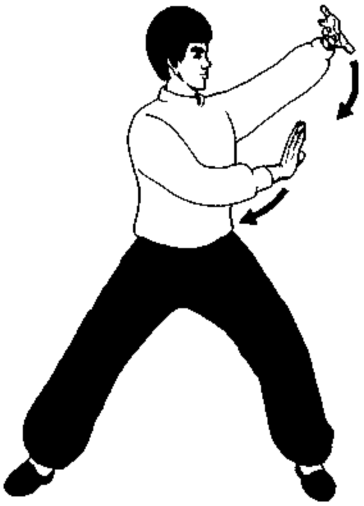

Figure 42.27. Continue pressing until both hands extend to the furthest point.

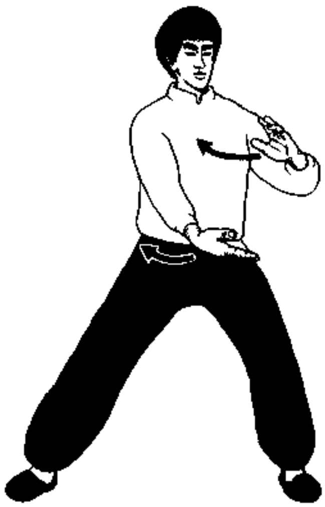

Figure 42.28. Inhale and turn the right palm upwards facing the sky as the body twists back to the center allowing the right hand to end up in front of the navel.

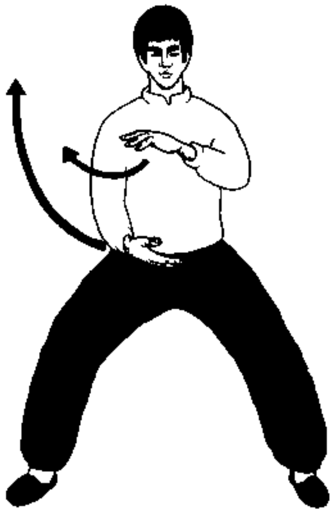

Figure 42.29. The left hand simultaneously turns so that the palm faces the center of the right hand following it to the center line of the body.

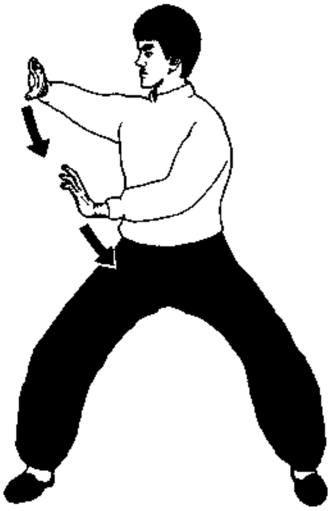

Figure 42.30. As the body reaches the point where it faces forward, twist to the right side and push in the same manner while exhaling.

SPLEEN TONIFICATION AND REGULATION EXERCISE

This exercise is designed to massage the Spleen's tissues and nerves, as well as tonify the entire Spleen organ and channels.

Raising the arms above the head stretches all of the suspensory ligaments of the Spleen making the Spleen energetically accessible by the twisting movements of the torso. The compression and release action on the connective tissue and fascia that support the placement of the Spleen allows the internal channels to become dilated facilitating a greater flow of Qi and Blood into the organ.

While looking up and turning to the right, the left brain is stimulated. Consequently, by looking up and turning to the left, the right brain is stimulated. According to Western physiology, light waves come into the eyeballs through the retina stimulating the optic nerve, chiasma, tract, lateral geniculate body (stimulating the thalamus), superior colliculi (in the brain stem where visual reflexes are initiated), optic radiations, cerebral cortex, and occipital (visual) cortex, which integrates both visual and memory impulses responsible for the perception of the image. Half of the information received from the right eye goes to the right side of the brain, while the other half of the information crosses over and stimulates the left brain.

Patients practice this exercise to strengthen their digestive system, as well as to enhance their visual equilibrium.

Upon completion of the Heart exercise, bring both hands down to the waist level, palms facing downward. Circle them above the head and perform the Pulling Down the Heavens exercise. Next, move the palms towards the outside of the knees while twisting the thumbs to point towards the back of the body (this stimulates the Luo points of the arms and squeezes the Wei Qi of the hands and arms deep into the bones) (Figure 42.31).

The thumb and index fingers touch and form a triangle as the hands slowly raise above the head (Figure 42.32).

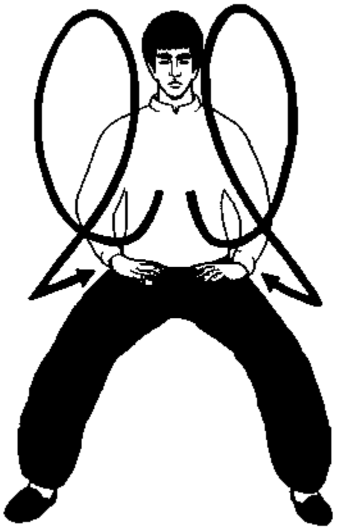

Figure 42.31. Preparation for the Spleen Regulation Exercise

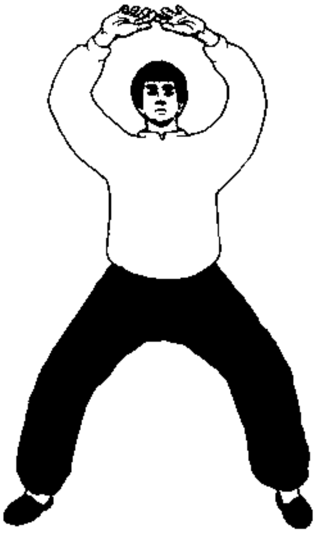

Figure 42.32. Spleen Regulation Exercise: the thumb and index fingers touch and form a triangle as the hands slowly raise above the head.

The eyes look through the center of the triangle as the upper torso is turned from side to side. It is important that the hips remain stationary and that the upper torso rotate in order to massage the digestive organs. Inhale in the center. Exhale on each side. Repeat 10 times on each side (Figure 42.33).

ENDING, ROOTING, AND STABILIZING THE QI

The purpose of the ending is to gather the Qi back into the patient's Lower Dantian rooting and stabilizing the patient's Qi. This ending exercise is very important as its initial goal is to gather any excess Qi developed from the over stimulation of the Five Yin Organs and to collect the runoff energy into the Lower Dantian (where it can be absorbed and used by the Eight Extraordinary Vessels). Otherwise, if the individual unknowingly has a preexisting Excessive condition in a particular Yin organ, the additional tonification can exasperate the organ's condition. To avoid this Excess condition, certain Medical Qigong schools will use external massage and lightly brush the energetic channels after practicing certain Qigong exercises and meditations to disperse any excess Qi.

The soft compression used during the ending exercise increases the partial pressure of oxygen in the Blood as more oxygen passes into the cells. The action of moving oxygen into the cells creates a movement of carbon dioxide out of the cells increasing the exchange rate and establishing a greater metabolic breakdown.

The breathing initiated for the ending exercise is known as "Turtle Breathing." The Turtle Breathing method requires slow respiration and stimulates the center hub of the Belt Vessel.

Upon completion of the Spleen exercise, exhale and start to separate both hands (as if embracing a ball) (Figure 42.34 and 42.35).

Bring both hands down to waist level and separate the arms, moving the hands away from the center of the body (Figure 42.36). Inhale and raise the arms up over the head. Each hand is placed above the head as if embracing a ball (Figure 42.37).

Exhale and form loose fists with the hands (Figure 42.38).

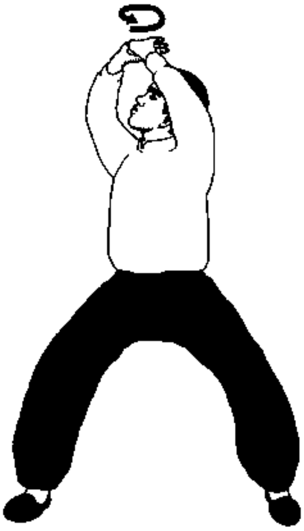

Figure 42.33. The eyes look through the center of the triangle as the upper torso is turned from side to side. Inhale in the center, exhale on each side.

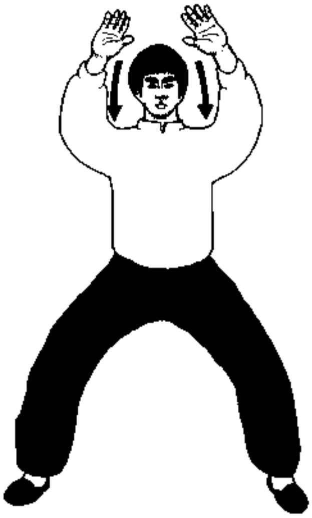

Figure 42.34. Preparation for Ending the Daoist Five Yin Organ Exercise (a)

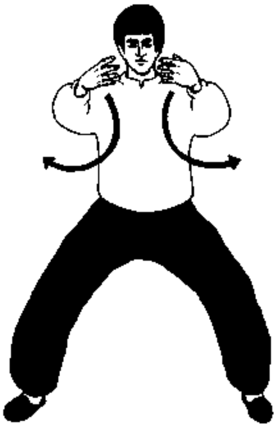

Figure 42.35. Preparation for Ending the Daoist Five Yin Organ Exercise (b)

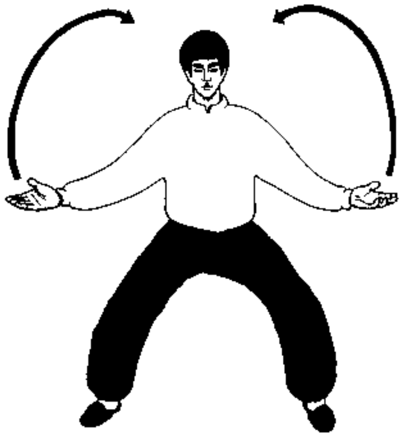

Figure 42.36. Preparation for Ending the Daoist Five Yin Organ Exercise (c)

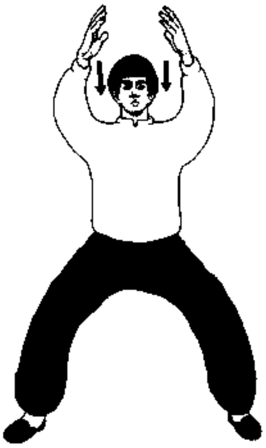

Figure 42.37. Ending the Daoist Five Yin Organ Exercise: each hand is placed above the head as if embracing a ball.

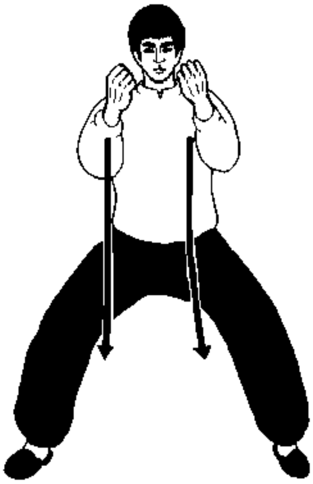

Figure 42.38. Exhale and form loose fists with the hands.

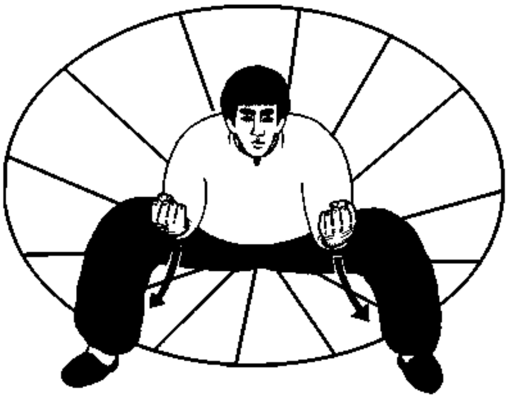

Figure 42.39. As you bend over, slowly turn the soft fists so that the curled fingers face the body and begin squatting, forming a Turtle Posture.

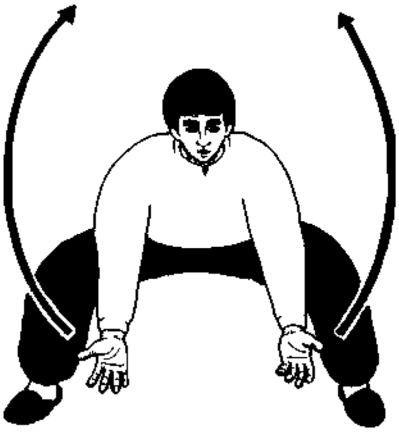

Figure 42.40. Inhale and slowly begin to raise the body up while opening the fists, and allow the hands to return back to their original position at the sides of the body.

As you bend over, slowly turn the soft fists so that the curled fingers face the body and begin squatting forming a Turtle Posture (Figure 42.39).

Inhale and slowly begin to raise the body up while opening the fists and allow the hands to return back to their original position at the sides of the body (Figure 42.40). Repeat five times and end in the Wuji posture.

When practicing the Turtle Breathing method, the mind's intention and respiration should be focused onto the Lower Dantian, navel, and Mingmen areas of the body.

The Lower Dantian is divided into eight sections (see Figure 42.41). Each section is numbered, beginning with the front of the Lower Dantian at the navel area (1), increasing in number as one continues over the left and right sides of the waist (4), ending at the Mingmen area (8).

Think of the anal sphincter as a great funnel, attached in eight different segments (like canals) which progress up the torso. As you focus your mind on each section (begin with the navel at section 1), allow any excess energy to flow downwards from that particular section to be gathered into the Lower Dantian. With your imagination, direct the breath and Qi from your torso to flow into each of the eight areas.

The breath and Qi should be slowly inhaled and guided into the Lower Dantian via the anal sphincter area for an even three counts. Then, proceed to gather Qi from the next consecutive area of the body. By pulling upwards on the anal sphincter and perineal areas, an energetic vacuum is created within the pelvic diaphragm and urogenital diaphragm. This energetic vacuum is used to gather the internal organ's excessive Qi.

Begin with collecting Qi from the front of the body, to the navel via the anal sphincter. Hold the Qi in the Lower Dantian area for three counts, then proceed to the next areas at the sides of the body absorbing the Qi into the Lower Dantian via the anal sphincter for three counts.

Next, proceed to the sides of the body continuing in progression until you reach the Mingmen area (8). Then, reverse the progression beginning with the Mingmen and working around

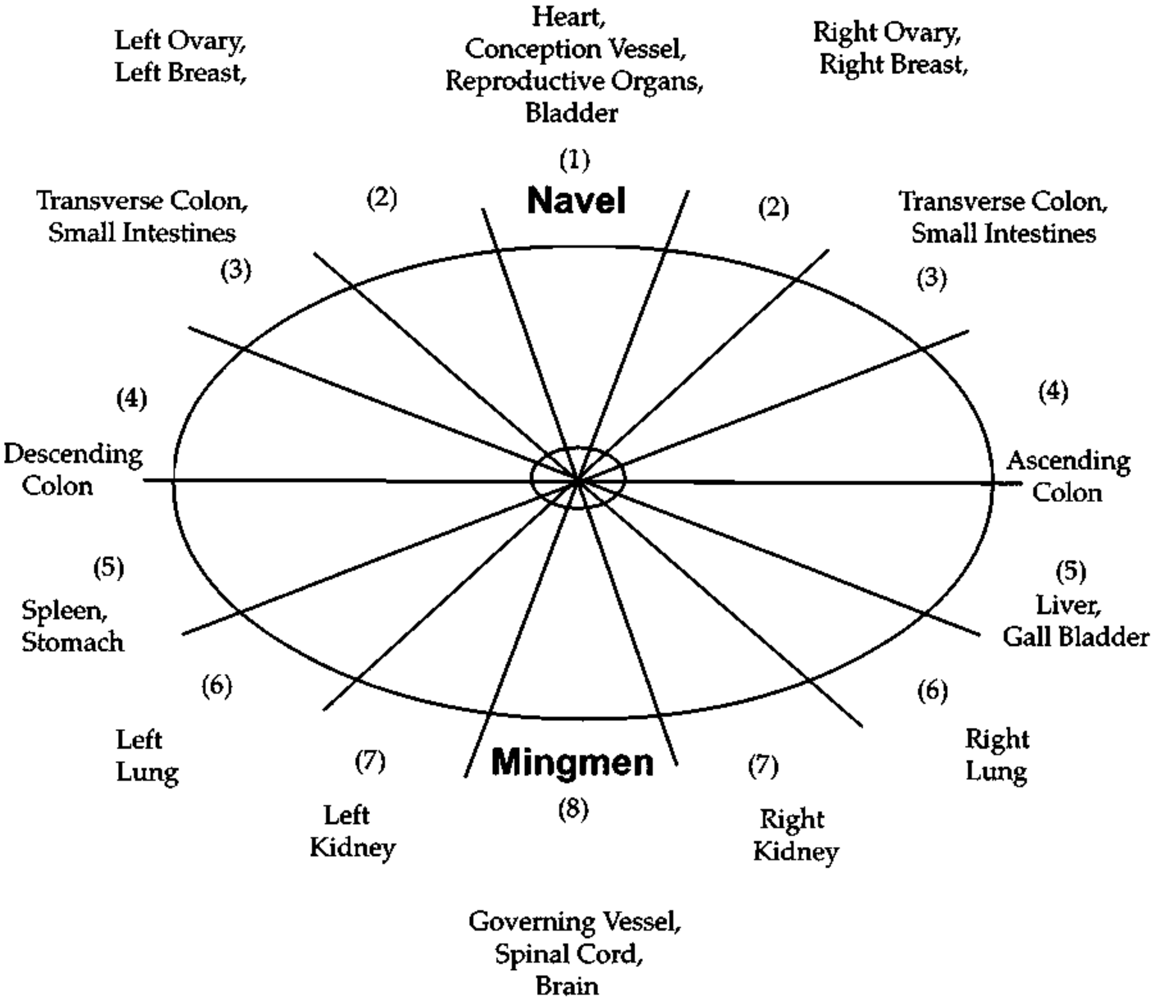
Left Ovary,
Left Breast,
Heart,
Conception Vessel,
Reproductive Organs,
Bladder
Right Ovary,
Right Breast,
(1)
Navel
Transverse Colon,
Small Intestines
(2)
(2)
Transverse Colon,
Small Intestines
(3)
(3)
(4)
(4)
Descending
Colon
Ascending
Colon
(5)
Spleen,
Stomach
(5)
Liver,
Gall Bladder
(6)
(6)
Left
Lung
(7)
Mingmen
(7)
Right
Lung
(8)
Left
Kidney
Right
Kidney
Governing Vessel,
Spinal Cord,
Brain

Figure 42.41. When practicing Turtle Breathing, inhale and slowly begin drawing the Qi from the front of the body into the navel area (1), absorbing the energy from the Heart, Conception Vessel, reproductive organs, and Bladder into the Lower Dantian. This absorbing action is created by pulling upwards on the anal sphincter in the direction of the intended energy. Next, proceed to the sides of the body. In women, the next area (2) will include the excess energy gathered from both the left and right ovaries, as well as the left and right breasts. In men, however, this area (2) is considered general and not specific to certain internal organs or tissues. Continue in the progression of circling the waist until you have reached the Mingmen area (8). Then, reverse the order of the tissue areas (8, 7, 6, etc.).

the waist until you reach the navel (8, 7, 6, etc.). Make sure that you push your anal sphincter towards the direction of the area of the body from which you are absorbing Qi.

After perfecting the Turtle Breathing, a practitioner can create the energetic vacuum and initiate the absorption of excess Qi into the Lower Dantian by moving his or her intention from the anal sphincter through the Lower Dantian and torso at an extremely fast speed.

SECTION X

TREATMENT OF INTERNAL DISEASES WITH QI EMISSION THERAPY

INTRODUCTION

Clinical records from China verify that Medical Qigong treatments are most effective on nervous system diseases, migraines, bronchial asthma, nocturnal enuresis, psychosomatic diseases, gastric ulcers, arthritic pains, sterility, and in the elimination of ovarian cysts and benign tumors.

The following chapters are designed to assist the Medical Qigong doctor in his or her professional clinical practice, by explaining the various techniques and prescriptions used in several Medical Qigong Hospitals, Institutes and Colleges throughout China.

These Medical Qigong techniques are presented as basic guidelines for the Qigong doctor to use, and need not be followed exactly to the letter. They are presented as an example to advise Qigong doctors in treatment and prescription protocols.

I have personally found these various approaches of Chinese energetic medicine to be extremely effective in my own clinical practice, and have therefore encouraged the directors of the International Institute of Medical Qigong to instruct their students in the following techniques.

This particular section of the book (Section X) deals with the treatment of Zang (Yin) and Fu (Yang) disharmonies in clinical practice, and discusses general and specific diagnosis, treatments, and prescriptions. I will begin by exploring several examples of the Five Elements' disease patterns and the various techniques used to treat them (Chapter 43-48). Starting with the treatment of Liver diseases, I will progress to the treatment of Heart, Spleen / Stomach, Lung and Kidney Diseases.

In the ending chapter in this section emphasis will be placed on the practical application and theoretical principles of "Miscellaneous" diseases and their treatments.

CHAPTER 43
QI EMISSION THERAPY FOR LIVER DISEASES

INTRODUCTION

The Five Yin Organs' interrelationship with each other, and their complex interrelationship with the energetic ebb and flow of Qi must constantly be regulated to maintain health. The constant flux of energetic flow within the body sometimes causes an Excess or Deficiency of Qi requiring specific organ therapy for restoration and balance. This therapy is designed to bring the specific organ or organ systems back into balance by using various meditations, visualizations and breathing exercises.

This section focuses on specific Liver diseases and how to rectify them. The Liver is responsible for the circulation and smooth movement of the body's internal Qi.

The Liver is a solid (Zang) organ which stores the Blood and is attributed to the element Wood. The Liver is sometimes referred to as the Green Emperor. The Liver is in charge of strategy without hesitation. It is energetically paired with the Gall Bladder (a hollow organ) which carries out the decisions made by the Liver (Figure 43.1).

The Liver houses the Hun. The Liver's association with the environment is through the eyes. Through observation, the Qigong doctor's eyes can receive images that reveal the status or intention of the patient. These images are absorbed into the doctor's Shen via the Hun. There is an old Chinese saying, which goes, "When the Hun swim to the eyes, a person can see."

In the springtime, Liver Qi becomes more active in individuals with strong Liver Qi; whereas, it may become Deficient in those with weak Liver Qi. When Liver Qi is Deficient, the patient becomes fearful; when the Liver Qi is in Excess, the patient becomes consumed by anger. The excessive eating or drinking of sour foods, and overexposure to wind depletes the Liver.

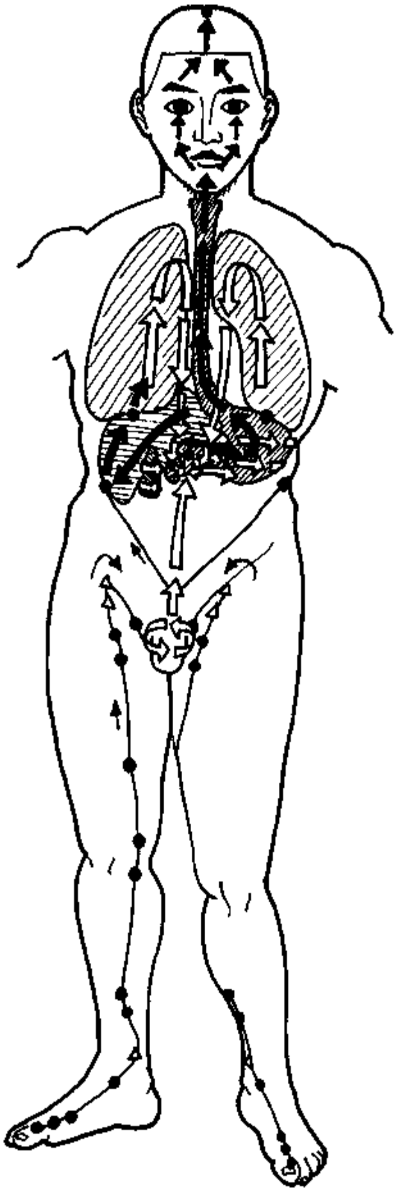

Figure 43.1. The Liver (Lv) Channel

LIVER COMPLICATIONS AND SYMPTOMS

Liver diseases often manifest with the following symptoms which are common to the organ's energetic dysfunction: headaches, dizziness, tiredness, and mental and emotional problems. When Liver disorders occur, the patient may experience Qi movement below the left side of the navel (which feels hard upon pressure), pain and distention in the hypochondria, cramps, drowsiness, blurred vision, etc. A list of Liver syndromes and their symptoms are described as follows.

1. Stagnation of Liver Qi can cause such symptoms as: headaches, mental and emotional problems, hypochondriac pain, abdominal pain and masses, diarrhea, constipation, painful urination, dysmenorrhea, and premenstrual tension.
2. Stagnant Liver Qi Invading the Stomach can result in epigastric pain.
3. Stagnant Liver Qi Invading the Lungs can cause wheezing.
4. Liver Yang Rising can cause such complications as: headaches, dizziness, and tiredness.
5. Liver Fire Blazing can cause such symptoms as: severe headaches, dizziness, breathlessness, wheezing, coughing, mental and emotional problems, insomnia, tiredness, painful or scanty urination, and constipation.
6. Liver Wind can cause headaches, dizziness, seizures, tremors, muscle spasms and tiredness.
7. Liver Blood Deficiency can cause such symptoms as headaches, mental and emotional problems, tiredness, hypochondriac pain, constipation, and premenstrual tension.
8. Liver Yin Deficiency can cause such symptoms as wheezing, mental and emotional problems, insomnia, tiredness, and hypochondria pain.
9. Stagnation of Liver Blood can cause such symptoms as headaches, mental and emotional problems, hypochondriac pain, abdominal pain, and dysmenorrhea.
10. Damp Heat in the Liver and Gall Bladder can cause such symptoms as jaundice, headaches, mental and emotional problems, hypochondriac pain, and abdominal pain.

GENERAL TREATMENT FOR LIVER DISEASES

The following is a description of a general protocol used for treating common Liver diseases. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted or Turbid Qi is removed. Once the Qigong doctor becomes proficient at locating these specific points, the treatment flows more smoothly.

The reader will notice that after comparing the various Liver and Gall Bladder therapies, certain patterns repeat themselves.

1. Generally, the Qigong doctor accesses the patient's Liver through either direct energetic insertion into the Liver's anatomical location, or through the patient's channels or external energetic fields.
2. After purging, tonifying, and regulating the patient, the Qigong doctor stimulates the back gates to the patient's Liver and the Yellow Court.
3. Next, the doctor roots the patient's Middle Burner Qi into the Lower Dantian.
4. Finally, the Qigong doctor leads any remaining Turbid Qi down the right Gall Bladder Channel and out of the patient's body (Figure 43.2).

Keep in mind that these areas are general, and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

MEDICAL QIGONG THERAPY FOR LIVER QI DYSFUNCTION

The primary therapies used to treat Liver diseases include the Windy Breathing Method (see Chapter 16) to strengthen the Liver, Spleen, Lungs, and Kidneys, and Quiescent Qi Regulating meditations (see Chapter 14).

1. Patients with Deficiency of Qi and Cold syndromes of the Liver may feel cold and are li-

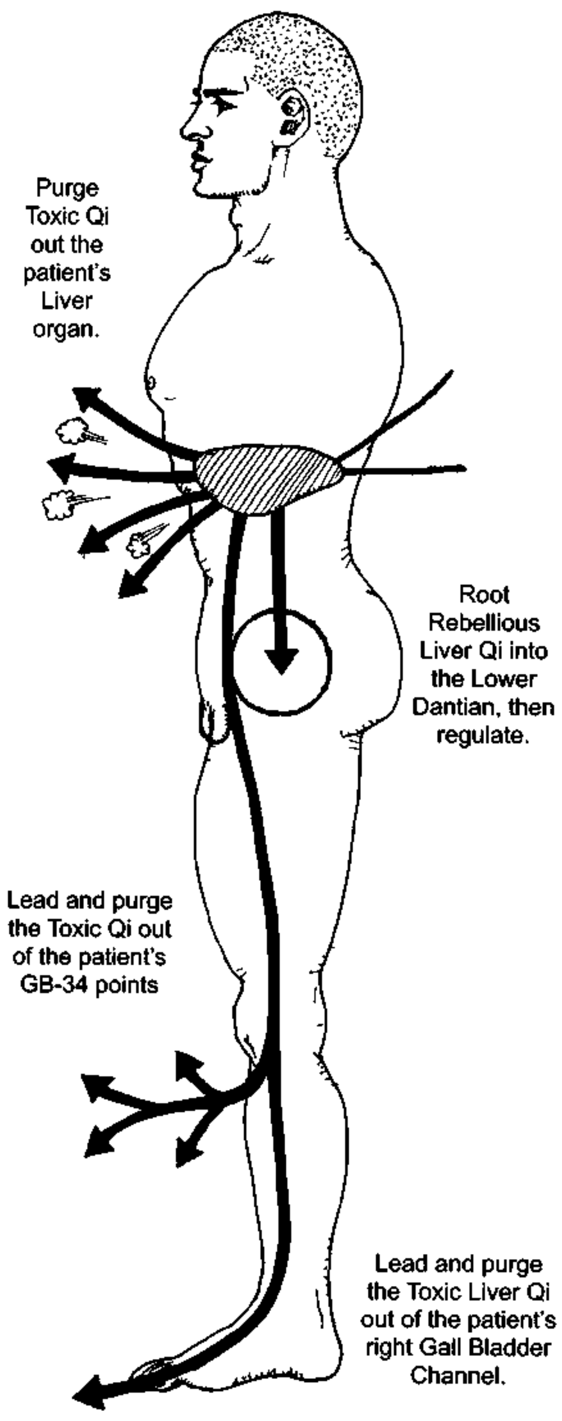
Purge Toxic Qi out the patient's Liver organ.
Root Rebellious Liver Qi into the Lower Dantian, then regulate.
Lead and purge the Toxic Qi out of the patient's GB-34 points
Lead and purge the Toxic Liver Qi out of the patient's right Gall Bladder Channel.

Figure 43.2. Basic Treatment for Liver Stagnation

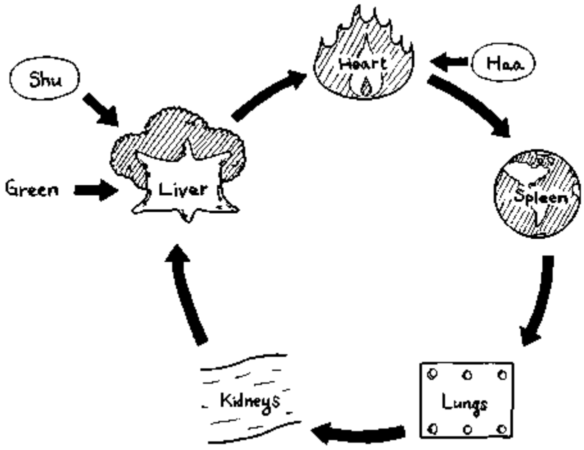
Shu
Heart
Haa
Green
Liver
Spleen
Kidneys
Lungs

Figure 43.3. Purging the Child to Treat the Mother

able to feel frightened. To treat this condition, the patient can practice the meditation of Taking in the Green Qi to nourish the Liver and also Taking in the Blue Qi to nourish the Kidneys (Formula: nourishing the mother to replenish the child).

2. Patients with an Excess of Qi and Heat syndromes of the Liver may have dry eyes, pain in the hypochondria that may radiate down to the lower abdomen, and changing moods accompanied by flushed cheeks. To treat this condition, the patient can practice uttering "Xu" (pronounced Shu) for the Liver, and "Ha" for the Heart (Figure 43.3. Formula: purging the child to treat the mother).

   Note: In treating Liver diseases, exhaling the sound "Shu" purges and sedates, while inhaling "Shu" replenishes and tonifies (see Chapter 40).

3. Stagnation of Liver Qi and Liver Fire Rising (anger or rage impairing the Liver) may cause an abnormal dispersing and impair the Liver's function. The patient may experience headaches, deafness, and swelling cheeks. To stabilize the hyperactive Liver Yang (and guide the flaring up Liver Fire to descend), the Qigong doctor guides the patient's Qi back to its origin with external energy projection. If treatment of Liver stagnation is not undertaken, there may be an organic change in the

Liver, resulting in hardening of the Liver or Liver cancer.

4. Patients with Rebellious Liver Qi resulting in hypochondriac pain and stiffness with spasms, should be treated by guiding their Qi back to the origin of the Liver with external Wei Qi therapy.

LIVER QIGONG PRESCRIPTIONS

The following system of exercises can be used to regulate the Qi and Blood of the Liver Channels, soothe and regulate Qi circulation, and suppress hyperactive Yang. These exercises can also be used to prevent and treat diseases and syndromes such as hypertension, psychoneurosis, chronic hepatitis, and hepatosplenomegaly (the enlargement of both Liver and Spleen), as well as bitter taste in the mouth, dry throat, dizziness, vertigo, and fullness in the chest and hypochondrium.

COLOR VISUALIZATION AND ORGAN CORRESPONDENCE

In China, the clinical use of color therapy is prescribed according to the Five Color Correspondence Theory. Each of the five major colors are assigned to specific organs and their channels (see Chapter 4). If an organ is diseased, it will generally appear in various shades from gray to black. To heal the disease, the patient imagines taking in the organ's clean color, while dispersing the turbid pathogenic color. This visual meditation helps to return the organ's Qi back to its origin, in order to achieve balance and health.

PRESCRIPTIONS FOR TONIFYING THE LIVER

Taking in the Green Qi can be used to tonify the Liver. The properties inherent within the color green have a persistent vibratory rate also used for tonifying the Liver and Gall Bladder.

1. Assume a standing, sitting, or lying posture. Relax the whole body, breathe naturally, and get rid of any stray thoughts. Place the tongue against the soft palate on the roof of the mouth.
2. Visualize green energy in front of you. Draw from a visual/emotional association with green trees, grass, bushes, etc. from your imagination, or if possible, actually look at a green object. Inhale the green light energy in through the nose and down into the Liver organ. As you exhale through the mouth, the dark, pathogenic Qi leaves the Liver; however, the bright, clean color remains in the organ, stimulating and vitalizing it. With each breath the organ retains more clean energy and begins to glow. Repeat for five breaths.

   Next breathe the color green in through the nose, filling the entire mouth. When exhaling, send the Qi slowly to both costal portions at the sides of the Lungs, and then down to the Lower Dantian. Practice eight to sixteen times.

PRESCRIPTIONS FOR PURGING THE LIVER

Purge the Liver by having the patient rub his or her chest while sounding "Shu." Have the patient assume a sitting or standing posture, and place both palms flatly on both sides of the chest and inhale slowly. When exhaling, the patient should sound the word "Shu," while gently rubbing both costal regions in a clockwise circular motion with both palms. Practice for ten to twenty breaths.

PRESCRIPTIONS FOR SOOTHING THE LIVER AND CONDUCTING QI

The following exercise is used to regulate and soothe the Liver Qi; it is used after the purging or tonifying exercises.

1. Stand relaxed with both arms naturally hanging down, palms facing downward, with the five fingers of each hand slightly held up. Press down with slight force and imagine Qi reaching the palms and extending to the fingertips. Press the palms down three times (Figure 43.4).
2. Lift both hands up in front of the chest, with each palm facing forward. Focus the mind on both palms. Push the palms forward three times, first drawing them closer to the body and then pressing them further away from the chest (Figure 43.5).
3. Stretch both hands horizontally out to the sides of the body (as if imitating a bird that is stretching out its wings) with all the fingers pointing upward, and the palms pushing to

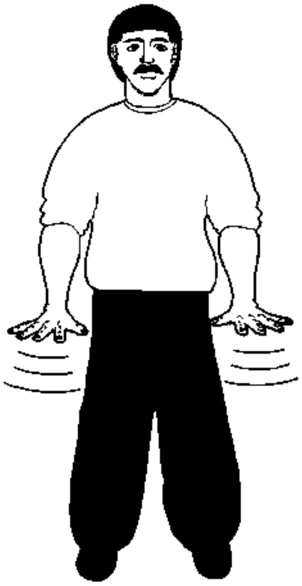

Figure 43.4. Press the palms downward three times.

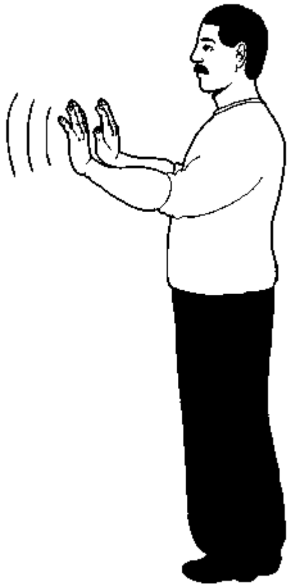

Figure 43.5. Press the palms forward three times.

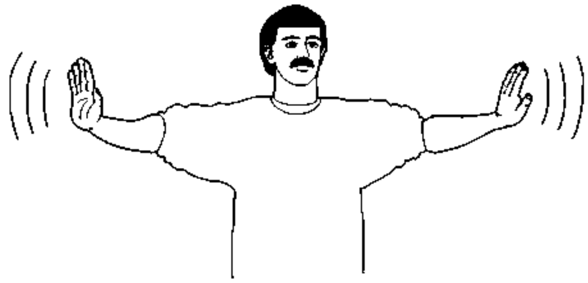

Figure 43.6. Press the palms to the sides of the body three times.

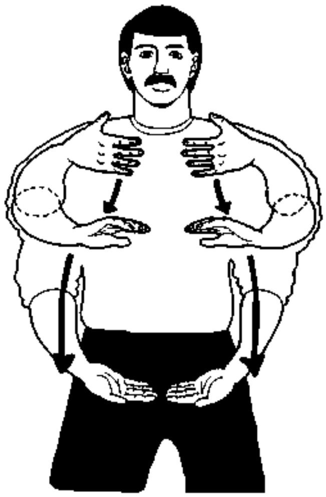

Figure 43.7. Direct the Qi to flow downward from the Middle to the Lower Dantian three times.

the sides. Focus your attention on extending the Qi out the palms to the fingertips. Press the palms out to the sides three times (Figure 43.6).

4. Draw both palms back to the front of the chest, palms facing upwards, elbows down. Rotate the palms to face downwards, and circle them down to the pubic arch, while focusing the mind on both palms (Figure 43.7). When Qi flows to the Lower Dantian turn the palms to face upward as if embracing a ball at the Lower Dantian.

Perform this exercise three times, then place both hands against the sides of the body. This completes one set. Do three sets.

MEDICAL QIGONG THERAPY FOR SPECIFIC LIVER DISEASES

The following are several Medical Qigong treatments used in China to treat specific Liver diseases.

DISORDERS OF THE BILIARY TRACT

The biliary tract includes the organs and ducts that participate in the secretion, storage, and delivery of bile into the duodenum.

ETIOLOGY

Disorders of the biliary track mainly include cholecystitis (inflammation of the Gall Bladder), cholelithiasis (formation of calcium or bile stones in the Gall Bladder), and ascariasis of the biliary tract (infestation by ascaris lumbricoides parasites). Traditional Chinese Medicine teaches that the first two diseases belong to the category of hypochondriac pain or jaundice, while the last condition is called biliary ascariasis.

Disorders of the biliary track can be caused by mental depression, stagnation of Gall Bladder Qi, the excessive intake of fatty foods, exopathogenic invasion, failure of biliary drainage system resulting from stagnation of Dampness and Heat, or obstruction by ascarid parasites. Although the causes and symptoms of these disorders are different, the Medical Qigong treatments and prescriptions are the same.

SYMPTOMS

The onset of biliary tract disorders are usually acute, and the pain is felt on the right upper abdomen and the right hypochondrium area of the thorax. Other symptoms include nausea, vomiting, rigor, high fever, yellow-colored skin pigmentation (jaundice), yellow sclera of the eyes, and whitish-grey stool.

Patients with biliary ascariasis may experience severe colic or a tearing pain below the xiphoid process. The pain is sometimes so severe that the patient experiences extreme perspiration accompanied by nausea and vomiting. If the ascarid parasites withdraw from the biliary tract, the pain is immediately relieved, but can return intermittently. If the ascarid parasites move completely into the Gall Bladder, the pain will be continuous, causing Gall Bladder distention. Symptoms such as jaundice, rigor, and high fever may appear, and a tenderness on the upper right quadrant of the xiphoid process can be felt with deep pressure.

TREATMENT

Cholelithiasis (gall stones) and cholecystitis (inflammation) are often treated by guiding the Toxic Qi downward and out of the patient's lower extremities along the Liver and Gall Bladder Channels, and by balancing the Qi activities between the upper and lower quadrants of the patient's body, as well as the right and left sides.

1. To begin treating disorders of the biliary tract, place the patient in a sitting position on the table. Stimulate the Bl-18, -19, -20, -21 points beside the patient's lower thoracic vertebrae, paying particular attention to the right side of the patient's body.
2. Next, extend Qi into the Gall Bladder area for 24 breaths (on the front of the patient's body), stimulating this area with the Tiger Kneading Palm technique, using Pushing, Pulling, and Shaking Qi manipulations.
3. To regulate the patient's Qi, and relieve epigastric distension, the doctor purges the patient's Toxic Qi, guiding it to flow out of the body downwards along the patient's Gall Bladder and Stomach Channels and out of the feet.
4. Using the Vibrating Palm technique, stimulate the patient's Bl-18, -19, -20, -21 points on the right side of the body once more, this time for 28 breaths. Then use the same Vibrating Palm technique on the front side of the body while directing the energy towards the Gall Bladder for 28 breaths.
5. Finally extend energy into the patient's Bl-19 point on the right side of the body for 12 breaths, and using the Sword Fingers technique, lead the Qi down the right Bladder Channel to the foot and out the small toe three times.

TREATMENT MODIFICATIONS

When treating the condition of ascariasis of the biliary tract, also include rubbing the abdomen 18 times in a counterclockwise direction.

PRESCRIPTIONS AND HOMEWORK

In prescribing homework, place the patient in a standing Wuji posture, with both palms placed against the hypochondria. Have the patient slowly massage the Gall Bladder area while purging the organ with the sound "Shu" for 24 exhalations.

Next, have the patient sit or lie down in a comfortable setting. As the patient inhales, he or she should meditate on the word peace; as the patient exhales, he or she should focus on relaxation. The patient should focus on releasing all tension in the head, chest, Liver, Gall Bladder, back, abdomen, and waist, for 36 breaths.

GALLSTONES

A gallstone is a Gall Bladder stone formed as a result of an excess of cholesterol in relation to the bile acids. The formation of calculi which results from this skewed relationship of cholesterol and bile is believed in Traditional Chinese Medicine to be dependent on the Liver Qi.

ETIOLOGY

Stagnant Liver Qi is believed to be a prerequisite for the formation of gallstones. If the Liver Qi stagnates, the bile will not secrete properly, leading to an accumulation of Damp Heat in the Gall Bladder. The steaming action of Heat on Dampness over prolonged periods of time, leads to the formation of gallstones.

It is generally believed that gallstones are related to cholestasis, an infection of the biliary tract, and metabolic disturbances of cholesterol. Gallstones form when the bile contains more cholesterol than it can keep in solution. The cholesterol precipitates out to form gallstones. The incidence of gallstones in women is approximately twice that in men. Any factor which increases the patient's cholesterol or decreases bile acids may lead to the formation of gallstones (i.e., obesity, high fat diet, Liver disease, and bile stasis increase the risk of developing gallstones). There are three types of gallstones: Cholesterol Stones, Mixed Stones, and Pigment Stones.

1. Cholesterol stones are usually single stones, which are composed almost entirely of cholesterol, and correspond to the pattern of Damp Heat in the Liver and Gall Bladder.
2. Mixed stones are the most common type of stones, consisting of lamellated layers of cholesterol, bilirubin, and calcium.
3. Pigment stones are less common and are always numerous. They are composed of bile pigment.

Gallstones are clinically manifested as pain in the right upper abdominal quadrant and right costal region, and are classified in Traditional Chinese Medicine as being placed in the categories of either hypochrondriac pain or jaundice.

SYMPTOMS

- Liver Qi stagnation is manifested by pain and distention in the right upper abdominal quadrant and right hypochondriac region. The pain can range from mild to severe. The patient may experience a distention and fullness in the Stomach with no obvious fever or jaundice.
- Damp Heat in the Liver and Gall Bladder is manifested by persistent pain in the right upper abdominal quadrant and right hypochondriac region. The pain manifests proximally or radiates up to the right shoulder. The patient experiences abdominal distention, fever, and jaundice.

TREATMENT

Have the patient sit on the edge of the table. Begin stimulating the patient's middle-back area, focusing on the Bl-21, Bl-20, BL-19, and Bl-18 points, on the right side of the back and drain the Toxic Qi down the torso and out the GB-34 point below the knee.

Use the Vibrating Palm hand technique and emit Qi into the patient's mid-back area for 14 breaths at each point. Then emit Qi into the painful area of the front side of the patient's body for 28 breaths.

Next, applying the Extended Fan Palm hand technique, emit Qi into the patient's Gall Bladder for 24 breaths using the Pulling and Leading manipulation skills to draw the Qi down through the Gall Bladder and Stomach Channels and out the patient's feet.

TREATMENT MODIFICATIONS

When treating a patient with a Liver Qi stagnation, use the Pulling and Leading manipulations to draw Qi down the lower limbs of the patient's body.

When treating patients with Damp Heat in the Liver and Gall Bladder: use the Extended Fan Palm hand technique and emit Qi into the patient's CV-12 and St-21 points, leading the Qi down the right Stomach Channel to the feet.

PRESCRIPTIONS AND HOMEWORK

Have the patient practice Liver Massage described in Chapter 39, while exhaling the sound "Shu" for 36 breaths.

PRESCRIPTION MODIFICATIONS

- For patients with Liver Qi stagnation, have the patient practice rubbing the hypochondrium while imagining the Toxic Qi descending down and out the body. Instruct the patient to practice the Descending the Yang and Ascending the Yin Technique (see Chapter 41).
- For patients with Damp Heat in the Liver and Gall Bladder, have the patient practice rubbing the hypochondrium while imagining the Toxic Qi descending down and out the body as he or she sounds the "Hu" sound to purge and regulate the Spleen.

CIRRHOSIS

Cirrhosis is a generalized disease marked by hepatic lesions. It is a chronic disease of the Liver characterized by the formation of dense lobular connective tissue, degenerative changes in the parenchymal cells, structural alterations in the Liver lobules, and sometimes fatty and cellular infiltration within the Liver.

ETIOLOGY

Cirrhosis refers to a series of pathological changes of degeneration, necrosis and regeneration of Liver cells, as well as the proliferation of fibrous tissue due to prolonged or recurrent damage of the Liver caused by various pathogenic factors. Cirrhosis is clinically manifested as a series of symptoms caused by a hypofunction of the Liver and portal hypertension.

SYNDROMES

1. Liver Qi stagnation and Spleen Deficiency results in lassitude, weakness, fullness in the chest and abdomen, hypochrondriac distension and pain, occasional nausea, belching, and vomiting.
2. Qi and Blood stagnation results in poor appetite, hypochrondriac distention and pain, and enlargement of the Spleen and Liver marked by spider nevus and Liver spots.
3. Water retention results in hardness and fullness of the abdomen with a severe fullness in the Stomach, distention of the abdomen or intestines (by gas), nausea, vomiting, and anorexia. Patients with an insufficiency of Spleen and Kidney Yang suffer from symptoms of cold limbs, pale complexion, and an aversion to cold. Patients with a Deficiency of Liver and Kidney Yin experience sporadic fever, irritability, and a dark complexion.

TREATMENT

Place the patient on the edge of the table and stimulate the Bl-18 and BL-20 points on the mid-back, as well as the Lv-13 and Lv-14 points (at the sides of the body), and the CV-12 point (at the center of the body). Use a Vibrating Palm hand technique to emit Qi into these points for a period of 14 to 18 breaths.

Use the Extended Fan Palm technique and emit Qi into the patient's mid-back at the Bl-18 points, as well as the sides of the torso at the Lv-13 and Lv-14 points. Purge and drain the Toxic Qi down and out the patient's body along the Gall Bladder and Liver Channels.

TREATMENT MODIFICATIONS:

When treating patients with Liver Qi stagnation and Spleen Deficiency, use the Vibrating Palm to emit Qi into the Bl-20 and CV-12 points.

When treating patients with Qi and Blood stagnation, use the Extended Fan Palm technique to emit Qi into the Lv-13 and Lv-14 points, conducting the Qi along the Gall Bladder Channel down the leg to GB-34, and out the body.

When treating patients with water retention, use the Extended Fan Palm technique to emit Qi into the Sp-21 points. Lead the Qi down the Spleen Channels into the lower limbs. For patients with a Spleen and Kidney Yang Defeciency or a Liver and Kidney Yin Deficiency, use a Vibrating Palm technique to emit Qi into the Lower Dantian and Mingmen areas.

PRESCRIPTIONS AND HOMEWORK

Have the patients practice self-regulation exercises to tonify the entire body's energetic system.

PRESCRIPTION MODIFICATIONS

1. For patients with Liver Qi stagnation and Spleen Deficiency, have the patients practice the Old Man Searching for the Reflection of the Moon at the Bottom of the Tide Pool (see Chapter 40).
2. For patients with Qi and Blood stagnation, have the patient perform the Liver Massage while exhaling the sound "Xu."
3. For patients with water retention, have them practice rubbing the Yellow Court area (CV-12) and sounding the word "Hu" to dredge the Spleen and Stomach. For patients with a Spleen and Kidney Yang Deficient, have them practice Gathering the Sun's Essence (see Chapter 11), and for patients with a Kidney and Liver Yin Deficient, have them practice Gathering the Moon Cream. (see Chapter 11).

CHOLECYSTITIS

In Traditional Chinese Medicine, both sides of the hypochondrium are related to the Liver Channels. In Western medicine, however, only the right side reflects a possible Liver pathology. Thus, it is important to have a distinction between left and right hypochondriac pain. Cholecystitis consists of the inflammation of the Gall Bladder with or without gallstones. It is related to hypochondriac pain on the right side of the body. It is created when cholesterol is maintained in the bile solution within the bile acids. Hypochondriac pain on the left side of the body extending to the epigastrium and left scapula is often related to acute pancreatitis.

ETIOLOGY

Cholecystitis is mostly caused by a bacterial infection or an obstruction of the bile duct. Clinically it is characterized by pain in the upper right quadrant of the abdomen and dyspepsia, and may be accompanied by nausea, vomiting, and sweating.

SYNDROMES

- Damp Heat manifestations are alternating attacks of chills and fever. Sometimes there is fever without chills, nausea, vomiting, loss of appetite, and pain in the right hypochondrium region.
- Qi stagnation manifestations are distending or moving pain in the right hypochondrium, belching, acid regurgitation, nausea, and anorexia.

TREATMENTS

1. Have the patient sit at the edge of the table. Stimulate the patient's mid-back, emitting Qi into the Bl-18, Bl-19, and Bl-20 points as well as the GB-34 point below the right knee.
2. Using the Vibrating Palm, first emit Qi into the patient's mid-back for 14 breaths, then directly into the painful area on the front side of the patient's body for 28 breaths.
3. Using the Extended Fan Palm hand technique, emit Qi into the patient's Gall Bladder region for 24 breaths, then purge the Toxic Qi down their right Gall Bladder Channel to the base of the foot and out of the body.

TREATMENT MODIFICATIONS

- For patients with Damp Heat, use the Extended Fan Palm hand technique to emit Qi into the patient's CV-12 area, then purge the Toxic Qi down the Stomach Channels to the base of the feet and out the body.
- For patients with Qi stagnation, use mostly purging and dredging Qi manipulations, focusing on the patient's lower torso.

PRESCRIPTIONS AND HOMEWORK

It is advisable to have the patient perform the Six Healing Sound Regulation exercise (see Chapter 40).

PRESCRIPTION MODIFICATIONS

- For patients with Damp Heat, it is advisable to have them perform the Spleen Massage (see Chapter 39) while exhaling the sound "Hu" to purge the Spleen and Stomach, allowing the Toxic Qi to flow down the Spleen and Stomach Channels into the ground.
- For patients with Qi stagnation, it is advisable to have them perform the Liver Massage (see Chapter 39) while exhaling the sound "Shu."

HYPOCHONDRIAC PAIN

Hypochondriac pain is a syndrome characterized by pain in one or both sides of the hypochondrium. The hypochondrium is located on the part of the abdomen beneath the lower ribs on each side of the epigastrium.

Chronic hepatitis, hepatolithiasis, cholelithiasis, cholecystitis and other diseases related to the Liver and Gall Bladder can be differentiated and treated in accordance with the exercises described in the prescription section.

ETIOLOGY

Hypochondriac pain may be caused by different factors such as trauma, the accumulation or blockage of Qi, stagnation of Phlegm and the Deficiency or Excess of Liver Qi (commonly seen in patients with Excess Liver Fire or Liver Qi stagnation). The symptoms of hypochondriac pain may occur in one or both sides.

SYNDROMES

- An Excess syndrome occurs when a patient suffers from Excess Liver Fire. This condition manifests with symptoms such as severe pain in the hypochondrium area with difficulty in breathing. The pain may radiate from both sides of the body, or from one side only.
- A Deficient syndrome occurs when a patient suffers from a Liver Yin and Kidney Yin Deficiency (often caused by mental depression or hemorrhage). This condition is observed clinically with symptoms such as a dull or sharp pain in the hypochondrium areas.

TREATMENT

1. Begin by stimulating the patient's Middle Dantian and Lv-14 and Lv-13 points at the front of the chest in the hypochondrium area (on both sides or the right side only). Next, stimulate the patient's mid-back around the Bl-17 and Bl-18 points and the base of the outside of the knees around the GB-34 areas.
2. Using the Extended Fan Palm hand technique, emit Qi into the patient's Lv-13 and Lv-14 points and the painful region of the chest and hypochondrium and begin purging the Toxic Qi, dredging it down the body and out of the extremities through the patient's Liver and Gall Bladder Channels.

TREATMENT MODIFICATIONS

- For patients with an Excess syndrome, extend energy into the patient's mid-back at the Bl-18 points. Then grasp and softly rock this tissue area to free the trapped Qi.
- For patients with a Liver and Kidney Yin Deficiency, emit Qi into the patient's Lower Dantian and Mingmen area using the Extended Fan Palm technique for 12 to 24 breaths.

PRESCRIPTIONS AND HOMEWORK

1. Have the patient perform the Liver Massage (see Chapter 39) and sound the healing tone "Shu" for 36 breaths.
2. If the pain is on the left side of the patient's chest, have the patient place his or her palms on the painful area and begin to purge the Toxic Qi from the midline of the thorax, down the left side of the body, following the Gall Bladder Channel to the ground, for nine breaths.
3. Finally, have the patient stand in a Wuji posture and focus the mind's intention on his or her palms. The patient then raises the arms sideways, like a bird stretching its wings (creating a "T" formation). Next, both arms embrace (as if holding an imaginary ball) at the Middle Dantian level. The patient imagines the Qi from the middle chest flowing downward and rooting into the Lower Dantian.

CHAPTER 44
QI EMISSION THERAPY FOR HEART DISEASES

INTRODUCTION

The Heart is a solid (Zang) organ and controls Blood circulation. The Heart corresponds to the element Fire and is sometimes referred to as the Red Emperor. The Heart houses the Shen and is in charge of mental activities and long term memory. It is energetically paired with the Small Intestine (a hollow organ), which is in charge of separating the pure, clean energy from the impure Qi (Figure 44.1).

The Heart's upper external connection with the outside is through the tongue, hence sensations of the tongue (from food, kissing and so on), as well as talking can travel directly into the spirit, mind, and emotion, affecting the Heart's desires and senses.

In the early summer months, the Heart energy becomes more active in individuals who already have strong Heart Qi, but can become Deficient in those who already have weak Heart Qi. Excessive eating and drinking of bitter foods, and overexposure to heat depletes the Heart.

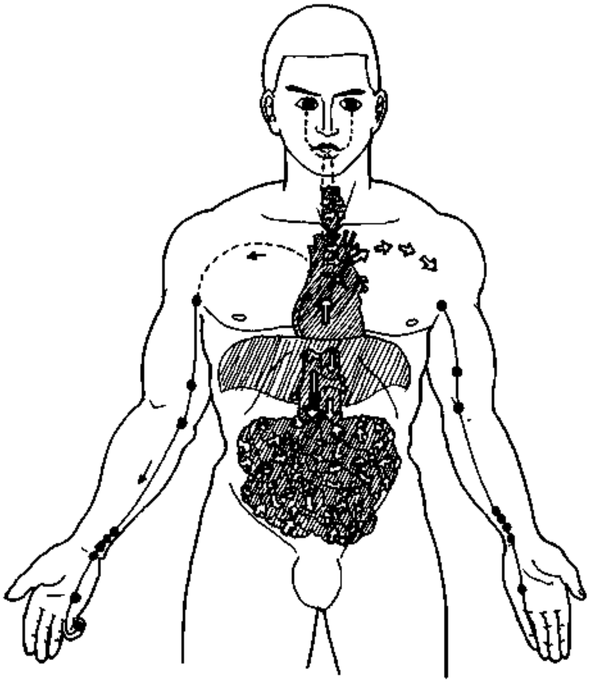

Figure 44.1. The Heart (Ht) Channels

HEART COMPLICATIONS AND SYMPTOMS

The Heart is responsible for governing the flow of Blood through the body's arteries and veins. Coronary artery disease (CAD) is a disease in which the patient's coronary arteries begin to harden, or impede adequate vascular flow to the myocardium, resulting in an insufficient supply of Qi and Blood throughout the body. This causes such diseases as hypertension, vasculitis, myocarditis, congenital Heart disease, rheumatic Heart disease, nervous malfunctioning of the Heart, organic pathological changes of the Heart, and arteriosclerosis of the brain.

In general, patients with Heart disease may have an arterial pulse above the navel which often disappears upon pressing, as well as a hot sensation in the soles of the feet and palms, dry mouth, stiffness of the tongue, and absentmindedness. A list of Heart syndromes and their symptoms are described as follows.

1. Heart Qi Deficiency can cause such symptoms as: physical tiredness, palpitations, sweating, pallor, and shortness of breath when exerting oneself.
2. Heart Yang Deficiency can cause such symptoms as: tiredness, tightening in the chest, and edema.
3. Heart Blood Deficiency can cause such symptoms as: headaches, mental and emotional problems, tinnitus, and tiredness.

4. Heart Blood stagnation can cause such symptoms as: mental and emotional problems, as well as tightness or pain in the chest.
5. Heart Yin Deficiency can cause such symptoms as: tiredness, insomnia, and mental and emotional problems.
6. Heart Fire can cause such symptoms as: insomnia, painful urination, and mental and emotional problems.

GENERAL TREATMENT FOR HEART DISEASES

The following is a description of a general protocol used for treating a patient's Heart. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted (or Turbid Qi is removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

After comparing the various Heart therapies, the reader will notice that certain patterns repeat themselves.

1. Generally, the Qigong doctor can access the patient's Heart through four methods:
   - direct energetic insertion into the Heart's anatomical location,
   - the patient's channels,
   - the external energetic Wei Qi fields, or
   - the patient's Taiji Pole (Figure 44.2).

   To avoid energetic armoring (while accessing the Qi in the front of the Heart), it is important for the doctor to already be connected to the patient's Shendao point (through the back of the Heart).
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor stimulates the patient's Heart and Shendao area, and roots the Upper Dantian Qi into the Lower Dantian.
3. Next, the Qigong doctor leads any remaining Turbid Qi down the Pericardium and Heart Channels and out from the patient's hands (Figure 44.3).

Keep in mind that these areas are general and

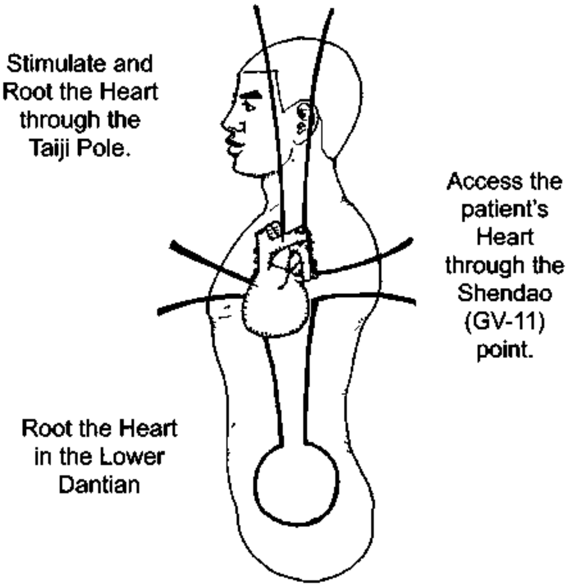
Stimulate and Root the Heart through the Taiji Pole.
Access the patient's Heart through the Shendao (GV-11) point.
Root the Heart in the Lower Dantian

Figure 44.2. Treating Heart Conditions

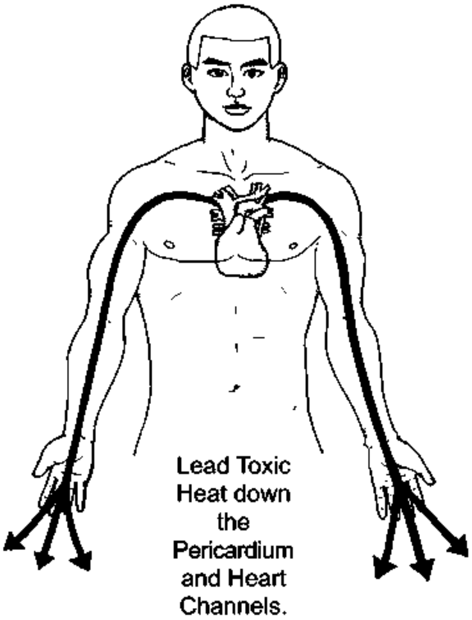

Figure 44.3. Lead the Toxic Qi out of the body through the patient's Pericardium Channels.

that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

MEDICAL QIGONG THERAPY FOR HEART QI DYSFUNCTION

The primary therapies used to treat cardiovascular diseases include: the Windy Breathing method to strengthen the Heart and Kidneys, Toe-Raised Walking, Heart Massage, and Quiescent meditation (see Chapter 15).

- Patients with an Excessive Qi syndrome of the Heart with extreme Heat may have a flushed complexion, full pulse, and excess sputum; they may feel pain in the chest, hypochondrium, back, shoulders, and arms. This syndrome can be treated by having the patient utter the "Ha" sound while guiding the Kidney Water Qi upwards to suppress the Heart Fire (Figure 44.4). Alternatively, the Qigong doctor may guide the Evil Qi out of the patient's body along the Heart and Small Intestine Channels with Pulling and Leading Qi manipulations.
- Patients with Deficient Heart Qi may dream of flames and red colored objects and have a radiating pain in the abdomen. This condition can be treated by having the patient perform the Taking in Yellow Qi meditation (see Chapter 4) which replenishes the child (Spleen), and Taking in Pink Qi meditation to nourish the mother (Heart); or, by using the Pushing and Guiding Qi manipulations to reinforce the Heart Qi and the Qi in the Lower Dantian.

In this particular case (Deficient Qi syndrome of the Heart), the reason the patient should not take in the energetic color red is because the Heart has already established a draining pattern, and any energy pouring into the Heart will only continue to drain away.

HEART QIGONG PRESCRIPTIONS

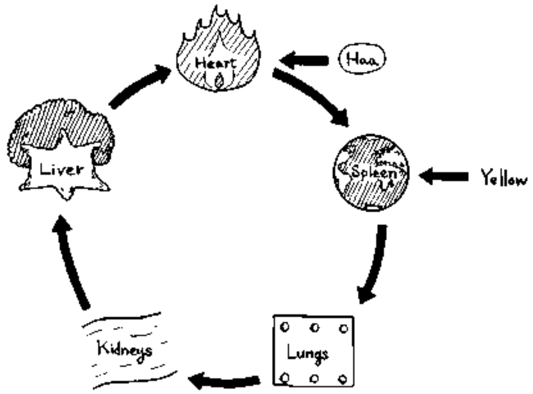

Figure 44.4. Treating Patients with an Excess Qi Syndrome of the Heart

To tonify the Heart, tranquilize the mind, promote Blood circulation, and remove obstructions in the channels, prescription exercises can be used. The following exercises can be used to prevent and treat diseases and syndromes such as coronary Heart disease, hypertension, angina, arrhythmia, rheumatic Heart disease, cardiac neurosis as well as palpitations (ranging from mild to severe), pericardial pain, and insomnia.

PRESCRIPTIONS FOR TONIFYING THE HEART

Taking in the Red Qi can be used to tonify the Heart. The properties inherent within the color red have a persistent vibratory rate used for tonifying the Heart and Small Intestine.

1. Assume a standing, sitting, or lying posture. Relax the whole body, breathe naturally, and get rid of any stray thoughts.
2. As you inhale, visualize red energy in front of you. Inhale the red light energy in through the nose and down into the Heart organ. As you exhale through the mouth, the dark, pathogenic Qi leaves the Heart; however, the bright, clean color remains in the organ, stimulating and vitalizing it. With each breath the organ retains more clean energy and begins to glow. Repeat for five breaths.
3. Next breathe the red energy in through the nose and fill the entire mouth with it. When exhaling through the nose, send the Qi slowly to the Heart, then to the Lower Dantian, to connect the energy of the Heart and Kidneys and to remove obstructions in the whole body. Perform this exercise seven or fourteen times. Return to the first step and close the training.

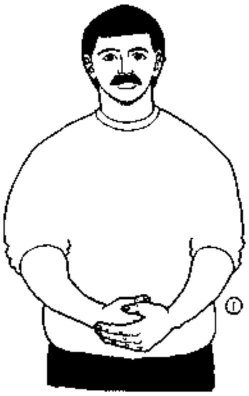

Figure 44.5. Hands rest in front of the Lower Dantian.

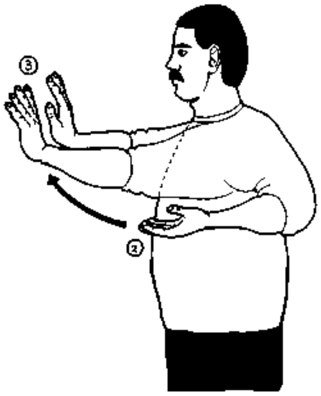

Figure 44.6. Both palms simultaneously press forward.

PRESCRIPTIONS FOR PURGING THE HEART

Purge the Heart by having the patient rub his or her chest while sounding "Ha."

1. The preparatory postures are the same as above.
2. Begin by clicking the teeth, stirring the tongue, and swallowing the saliva, then place the palms on the pectoral region of the left side of the chest (left hand on the outside for men, opposite for women). Inhale slowly, then gradually exhale the sound "Ha," focusing the mind on the Heart area. Repeat for six to twelve breaths altogether while rubbing with the palm lightly in a clockwise circle.

PRESCRIPTIONS FOR REGULATING THE HEART AND CONDUCTING QI

The following exercise is used to regulate the Heart after purging or tonifying.

1. Assume a standing or sitting posture. Relax the whole body, breathe naturally, and raise the tongue against the hard palate, just behind the teeth. Bring both palms loosely together (left over right for men, opposite for women) at the Lower Dantian; stand still for several minutes and concentrate the mind on the Lower Dantian (Figure 44.5).
2. Turn both palms outward, and push the arms along the sides of the body towards the back. Hold this posture for several minutes.
3. Turn both palms upward and lift them to the sides of the chest.
4. With both palms facing forward, stretch both hands level and forward. Focus your attention on the ends of the middle finger of each hand, while the thumb and little fingers stretch and separate (Figure 44.6).
5. Clench both hands into fists and pull them along the sides of the body towards the back, as if dragging heavy weights. Perform this movement three times (Figure 44.7).
6. Raise the right palm, as if holding a heavy weight in front of the chest, and push the palm out vertically to the right. Then, draw the right hand back to the side of the body near the hip. Repeat on the left side, and finally return to the starting posture. Perform these movements three times (Figure 44.8). This completes one set. Perform three sets.

SPECIFIC MEDICAL QIGONG THERAPY FOR HEART DISEASES

The following are several Medical Qigong Treatments used in China to treat specific Heart diseases.

HYPERTENSION

Hypertension is considered a condition in which the patient has higher blood pressure then what is determined to be normal for his or her

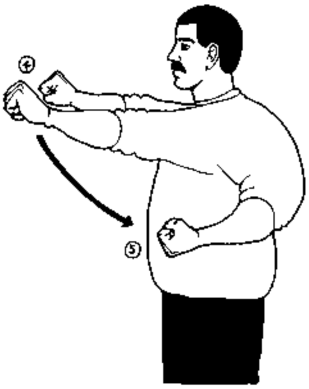

Figure 44.7. Both hands simultaneously clench into fists and pull to the hips.

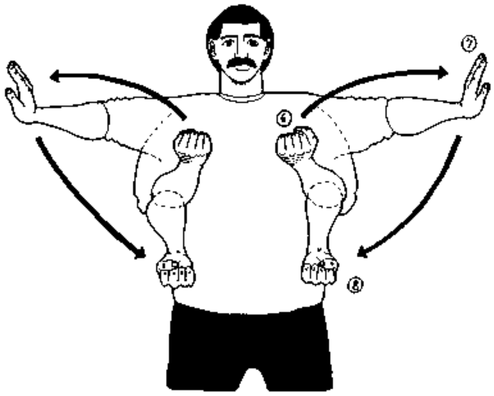

Figure 44.8. Alternate each palm as it presses towards the side of the body.

chronological age and physique. Generally, Traditional Chinese Medicine places it in the categories of vertigo and headache.

ETIOLOGY

Hypertension is one of the most common problems impacting the Heart. It is the result of an imbalance of the Yin and Yang functional aspects of Deficient Kidney Yin and Excess Liver Yang, and/or an overabundance of Phlegm and Dampness within the body.

Hypertension can be caused by the patient eating too much food or salt, by too much stress, or not enough exercise. A renin imbalance within the patient's Kidneys, increased cholesterol, a preexisting genetic predisposition, a secondary illness (diabetes, nephritis, or hyperthyroidism), or even pregnancy can also lead to hypertension. These causes of hypertension can result in one of three internal factors: high cardiac output due to stress, reduced elasticity of the vessels, and resistance to Blood flow.

1. When the body experiences emotional tension it releases adrenaline, which speeds up the Heart rate and causes the blood vessels to narrow, increasing the body's blood pressure.
2. A buildup of plaque and normal aging reduces the vessel's elasticity and can lead to hypertension.
3. Old age, certain hormonal imbalances or disorders, or a genetic predisposition can likewise cause high blood pressure.

The overall goals in treating hypertension patients are to:

1. Lower blood pressure,
2. Stabilize blood pressure,
3. Establish a quiescent state of mind (generally, meditating 20 minutes, twice a day provides the patient with a prolonged state of deep relaxation needed to combat stress), and
4. Encourage life-style changes (i.e., monitor potassium intake, decrease sodium intake, weight control, and stress reduction management). Treatment programs are established as described below.

SYNDROMES

As previously stated, there are three underlying causes or syndromes relating to hypertension: Excess Liver Yang, Deficient Yin, Excess Liver Fire, and Kidney Yin and Yang Deficiency.

1. If the body's Kidney Yin is Deficient, it fails to nourish the Liver Yin, which leads to a condition of hyperactivity of Liver Yang. The clinical manifestations are feelings of fullness in the head, daydreaming, headaches, dizziness, tinnitus, insomnia, and dysphoria with feverish sensations in the chest.
2. Excess Liver Fire can cause headaches, dizziness, restlessness, irritability, flushed face, and constipation.
3. If the body's Kidney Yang is Deficient the Yin (Water) accumulates, if the Yang does not move, the blood vessels cannot relax. This results in

obstructed Blood flow. The clinical manifestations are dizziness and tinnitus, daydreaming, palpitations, amnesia, lassitude and weakness in the legs, as well as insomnia.

TREATMENT FOR PATIENTS SENSITIVE TO ENERGY

1. Use Medical Qigong therapy, and have the patient sit in a chair with his or her feet flat on the ground. Begin the treatment by cleaning the patient's Conception and Governing Vessels, using purgation methods. Focus specific attention on clearing the patient's head area (Figure 44.9). Literally sweep with the entire hand, and comb the patient's Wei Qi fields with the fingers.

   By scanning the occipital region of the patient, you can verify that the Excess Qi in the patient's upper extremities has abated. Once these vessels are clean, you can focus on treating the hypertension.
2. Stand behind the patient, using the Extended Fan Palm technique (with both hands) extend energy from the left hand into the top of the patient's head at the Baihui GV-20 area, while your right hand extends energy into the back of the patient's head, beginning at Yuzhen Bl-9. The two streams of Qi flow downward to the Dazhui GV-14 area, where they join into one river, which continues to flow downward through the patient's Taiji Pole. Continue to visualize these two streams joining and filling the Lower Dantian, for a period of 6 to 12 breaths (Figure 44.10).
3. Next, move your left hand and extend energy into the patient's Yellow Court, while your right hand is placed above the patient's Mingmen area. Lead the Qi from the Yellow Court into the patient's Lower Dantian, continuing to root the Qi, for a period of 6 to 12 breaths.

TREATMENT MODIFICATIONS

These modifications are additional formulas which are added onto the previous treatment according to the specific cause and condition of the patient's Heart.

1. When treating patients with Excess Liver Yang and Deficient Yin, use the Extended Fan Palm

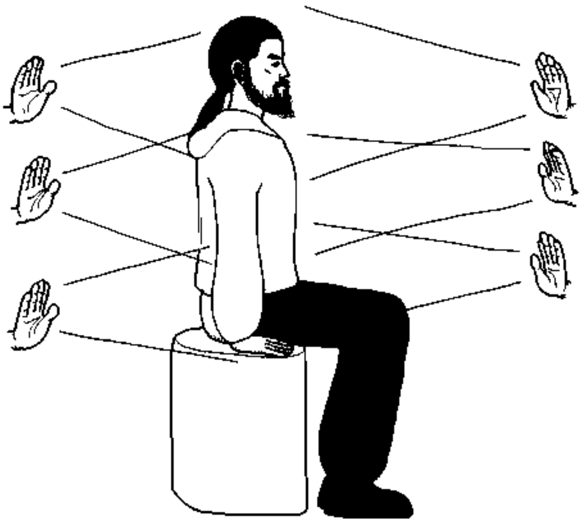

Figure 44.9. The Qigong doctor begins by dredging, purging and cleaning the patient's Conception and Governing Vessels.

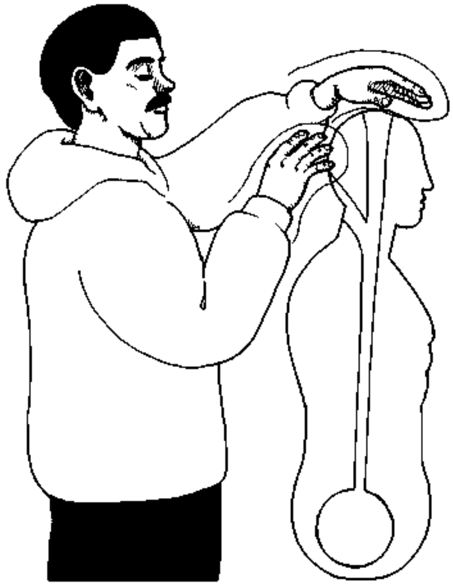

Figure 44.10. The Qigong doctor emits Qi into the patient's Baihui, Yuzhen and Dazhui points, in order to access the Taiji Pole and fill the patient's Lower Dantian.

hand technique and emit Qi into the patient's mid-back, drawing the energy down and rooting it in the patient's Lower Dantian and Mingmen.

2. When treating patients with Excess Liver Fire, it is important to treat the Lv-13 and -14 points, by dredging and purging the Excess Qi, and leading it down the patient's right Gall Bladder Channel, through the foot and into the Earth.
3. When treating patients with Kidney Yin and Yang Deficiency, use the Extended Fan Palm technique to emit Qi into the patient's mid-back, drawing the energy down and rooting it in the patient's Lower Dantian and Mingmen.

TREATMENT FOR PATIENTS NOT SENSITIVE TO ENERGY

Use Jing Point therapy in conjunction with Medical Qigong therapy (see Chapter 36). Have the patient lie prone. Using the Sword Finger technique, touch the patient's head with the left hand and extend energy into the Baihui GV-20 point. Guide this Qi down to the Lower Dantian through the Taiji Pole, while your right hand embraces and energizes the patients' Mingmen GV-4 point, thereby filling the Lower Dantian and rooting the patient's Heart Qi (Figure 44.11). Next, treat the patient's LI-11, St-36, Lv-2, and Sp-6 points on both sides of the body to lower the patient's blood pressure.

PRESCRIPTIONS AND HOMEWORK

1. A static posture can be given to convalescing patients as part of a meditative homework prescription for treatment of high blood pressure. The prescription is as follows.
    a. Have the patients lie supine (the patients may use a pillow under the head for comfort) and have him or her imagine that they are floating in warm water.
    b. The patients should perform the toning sound "Zheng" (prenounced Jang), and imagine dispersing Qi out of his or her body. The patient should use natural breathing while toning and dispersing the Toxic Qi. There are ten major areas, from the top of the head to the bottom of the feet, from which the patients must discharge the toxins. These areas include: the top of the head, Upper Dantian, throat, Middle Dantian, Yellow Court, Lower Dantian, the center of the thighs, knees, ankles, and bottom of the feet (Figure 44.12).
    c. After several minutes the patients will relax and focus their attention on the Lower Dantian or Bubbling Well points at the bottom of the feet.
2. Another meditation prescription given to patients with hypertension is to have them sit and imagine the feeling of warm water pouring down over the head, torso, extremities, and feet. This meditation should be practiced several times a day, using 36 breaths each time.
3. The patient can also be given the prescription of rubbing the hypochondrium and toning the sound "Shu" to purge Excess Heat from the Liver.

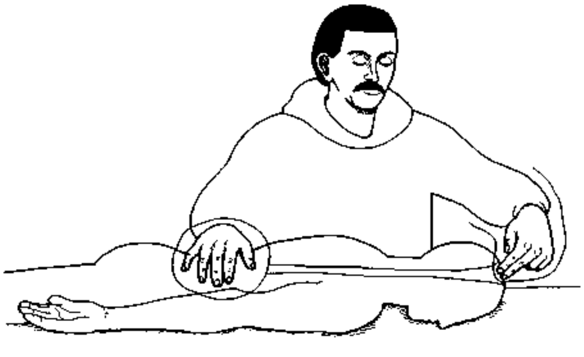

Figure 44.11. If the patient is not sensitive to energy, have them lie prone, and apply Jing Point Therapy in conjunction with Qi emission.

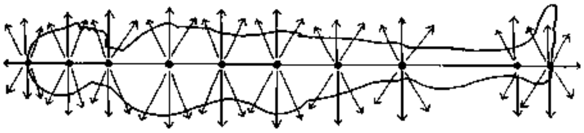

Figure 44.12. In this Static posture used to treat high blood pressure the patient lies supine and performs the toning sound "Zheng" (Jang), for dispersing Qi.

PALPITATION

A palpitation is an abnormally rapid throbbing or fluttering of the Heart. The symptom is characterized by nervous feelings or restlessness due to strong anxitey or fright. Clinically, it mostly attacks paroxysmally and is caused by emotional stress or overexertion. Symptoms such as insomnia, vertigo, tinnitus, etc. often accompany palpitations.

Diseases with palpitation as the main manifestation include arrhythmia, iron-deficiency anemia, aplastic anemia, hyperthyroidism and anxiety disorders. These diseases may all be differentiated and treated in reference to the prescriptions described in this section.

ETIOLOGY

There are several factors which contribute to palpitations. These factors and their symptoms are described as follows.

1. In patients suffering from Deficient Qi and Blood, and Deficient Heart Qi, palpitations may occur following sudden mental irritation. The patient may note unduly rapid heartbeat accompanied, in cases of Deficient Heart Qi, by pallor, restless sleep at night, dizziness and vertigo. This condition may also be caused by Deficient Heart Blood that fails to nourish the Heart.
2. Phlegm Fire disturbing the Heart may also give rise to palpitations. In cases with internal disturbances of Phlegm Fire, symptoms also include fidgeting and disorientation.
3. Another factor causing palpitations is a Deficiency of Heart Yang that leads to fluid retention. In cases of fluid retention with epigastric stiffness symptoms will also include: dizziness, excessive salivation and mental fatigue.

TREATMENT

1. Begin by extending Qi into the patient's Bl-14, Bl-18, CV-15, and CV-17 points. To open up these points, emit Qi into the tissue areas surrounding these points using the Extended Fan Palm technique.
2. Next, use the Sword Fingers hand technique and vibrate Qi into each of these points for the duration of six to twelve breaths at each point. Upon completion, the doctor will then guide the Qi back down the patient's torso and root it into the patient's Lower Dantian.
3. Finally, the doctor emits Qi into the patient's Yellow Court area using the Extended Fan Palm technique and then guide the Heart and Pericardium Channel Qi down the patient's arms using Pushing, Pulling and Leading techniques. This ending is performed to balance the Qi in the upper and lower aspects of the patient's body.
4. As with every treatment, end the procedure by regulating the patient's Microcosmic Orbit (Fire Cycle).

TREATMENT MODIFICATIONS

These modifications are additional formulas which are added onto the previous treatment according to the specific cause and condition of the patient's Heart.

1. In cases of insufficiency of Heart Blood, add Bl-17 and Bl-20 to the treatment points.
2. In cases of internal disturbance of Phlegm Fire, purge from the CV-12 to St-40 points.
3. In cases of Fluid retention, add Bl-21 and Bl-22 to the treatment points, for 14 to 24 breaths. Then, guide the Qi to flow down the patient's Gall Bladder Channels into their Bl-40 points.

PRESCRIPTIONS AND HOMEWORK

1. For treatment of palpitations due to Deficient Heart Blood, have the patient practice the method of Taking in the Red to nourish the Heart (see Chapter 4). Have the patient send the red Qi slowly down into the Heart and then to the Lower Dantian during the exhalation. Practice should be for several minutes, according to the patient's strength and constitution.
2. For treatment of palpitations due to Phlegm Fire, have the patient place both hands on the Heart and slowly inhale. As the patient exhales, he or she should pronounce the sound "Ha." This exercise should continue for several minutes.
3. For Heart regulation, have the patient practice the Heart exercise from the Daoist Five Yin Organ prescriptions (see Chapter 42).

RHEUMATIC HEART DISEASE

Rheumatic valvular Heart disease is a condition where there is chronic damage of the cardiac valve caused by rheumatosis. It clinically manifests with palpitations, dyspnea and edema.

SYNDROMES

1. Heart Blood stagnation is manifested through symptoms such as severe palpitations, bad cough with hemoptysis, lassitude and weakness, purplish-red color of the lips, nails and zygomatic facial regions, pain in the Heart, chest, or both.
2. Deficient Qi and Blood manifests through such symptoms as shortness of breath, palpitations which are aggravated on exertion, perspiration and pale complexion.
3. Heart and Kidney Yang Deficiency manifest through such symptoms as palpitations, pale or darkish complexion, edema, cough, cold hands and feet and dyspnea.

TREATMENT

1. Have the patient sit at the edge of the table, and begin to stimulate the Shendao (GV-11) and Middle Dantian areas.
2. Use the Vibrating Palm hand technique and emit Qi into the patient's Yellow court and Shendao (GV-11) area for 6 to 12 breaths.
3. Use the Extended Fan Palm hand technique and emit Qi into the patient's Heart region, leading Qi down the Heart and Pericardium Channels to the tips of the fingers (on both sides of the body).

TREATMENT MODIFICATIONS

1. When treating patients with Heart Blood stagnation, use the Extended Fan Palm hand technique and emit Qi into the patient's Middle Dantian, leading the stagnant Qi down the Heart and Pericardium Channels and out the hands (on both sides of the body).
2. When treating patients with Deficient Qi and Blood, use the Extended Fan Palm technique and emit Qi into the patient's Middle Dantian (both Shendao GV-11 and CV-17 areas), leading the Qi down the Conception Vessel to the Lower Dantian to root and stabilize the patient's Qi.
3. When treating patients with Heart and Kidney Yang Deficiency, use the Extended Fan Palm technique and emit Qi into the patient's Shendao (GV-11) and Mingmen (GV-4) for 24 breaths.

PRESCRIPTIONS AND HOMEWORK

When prescribing homework for patients with rheumatic Heart disease, have them begin with a quiescent sitting meditation, while focusing on their Lower Dantian.

PRESCRIPTION MODIFICATIONS

1. Have patients with Heart Blood stagnation perform the Heart Massage exercise (see Chapter 39), while toning the sound "Ha."
2. It is advisable for patients with Qi and Blood Deficiency to practice the method of Taking in the Yellow Qi into the Spleen and Red Qi into the Heart from the natural environment.
3. Have patients with Heart and Kidney Yin Deficiency practice Gathering the Moon Cream (see Chapter 11), as well as the method of Taking Yellow Qi into the Spleen and Red Qi into the Heart from the natural environment (see Chapter 4).

CONTRAINDICATIONS

The Heart may either have functional or structural problems. In either condition, Medical Qigong is contraindicated when working directly over a pacemaker.

Also, when treating Heart conditions, make sure the patient is under a Western doctor's care, and inquire as to any medications the patient may be taking.

CHAPTER 45
QI EMISSION THERAPY FOR SPLEEN/STOMACH DISEASES

INTRODUCTION

The Spleen is a solid (Zang) organ and it stores and nourishes Ying (Nourishing) Qi. The Spleen corresponds to the element Earth and is sometimes referred to as the Yellow Emperor. It houses the Yi, which governs the intention, thought, study, concentration, and memorization. It is energetically paired with the Stomach (a hollow organ). The Spleen is in charge of transporting, distributing and transforming nutrients and Qi into Gu Qi, as well as controlling Blood circulation (Figure 45.1).

The Spleen's external connection with the outside world is through the mouth, hence any connection with the mouth can affect the body's acquired constitution, developing a tolerance or intolerance to specific substances (allergies), as well as affect concentration and thought.

In the late summer months, Spleen Qi becomes more active in individuals who already possess strong Spleen Qi, but can become Deficient in those individuals who already have weak Spleen Qi. Excessive eating and drinking of sweet foods, and overexposure to dampness weakens the Spleen.

SPLEEN COMPLICATIONS AND SYMPTOMS

Because of the importance of both the Spleen and Stomach, this chapter is divided into two separate sections. The first section focuses on specific Spleen diseases and how to rectify them. The second section explores the energetic dysfunctions of the Stomach.

The following discussion of treatment begins with an explanation of Spleen complications and symptoms. Spleen syndromes have symptoms which are common to the organ's energetic dysfunction such as: diarrhea, tiredness, and edema. A list of Spleen syndromes and their symptoms are described as follows.

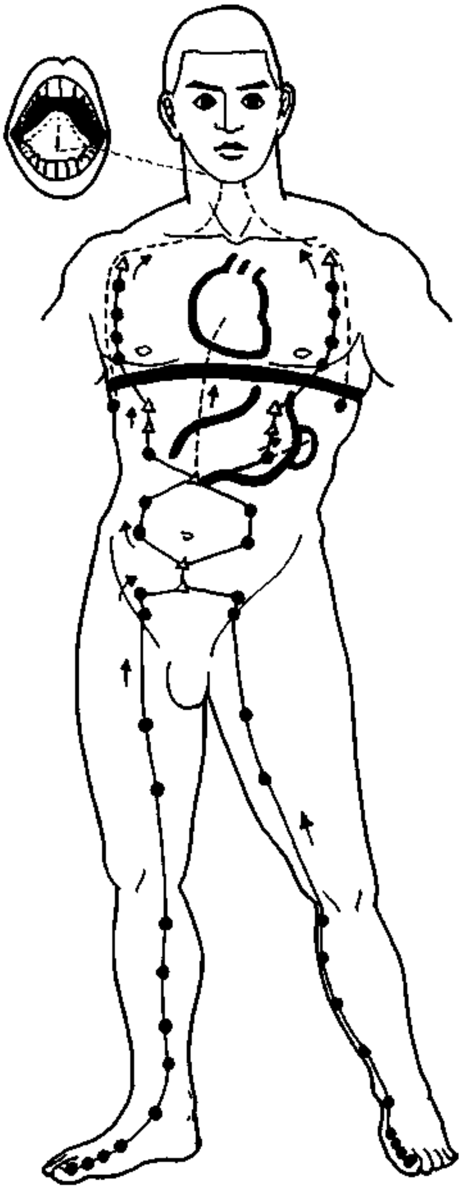

Figure 45.1. The Spleen (Sp) Channels

1. Spleen Qi Deficiency can cause such symptoms as: wheezing, tiredness, abdominal pain, diarrhea, constipation, painful urination, enuresis, bleeding and menorrhagia.
   a. Patients with a failure of normal transporting and transforming Qi activity in the Spleen may have a poor appetite, abdominal distention, loose stools, edema and retention of Phlegm.
   b. Deficient Spleen Qi may cause the Qi in the Middle Burner to sink. This results in shortness of breath, unwillingness to talk, persistent diarrhea, prolapse of the anus or uterus, or gastroptosis (downward displacement of the Stomach).
   c. Spleen Qi Deficiency can result in failure of the Spleen to control the Blood circulation and may be manifested by the following symptoms: bloody stool, bleeding from the uterus, and subcutaneous hemorrhaging.
2. Spleen Yang Deficiency can cause such symptoms as: tiredness and edema.
3. Spleen Blood Deficiency can result in fatigue.
4. Spleen Yin Deficiency can result in tiredness.
5. Damp Heat in the Spleen can cause such symptoms as: sinusitis, edema, common cold, and influenza.
6. Dampness in the Spleen can cause such symptoms as: diarrhea, tiredness, edema, and Wind Stroke.

GENERAL TREATMENT FOR SPLEEN DISEASES

The following is a description of a general protocol used for treating the Spleen. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted (or Turbid Qi removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

After comparing the various Spleen therapies, the reader will notice that certain patterns repeat

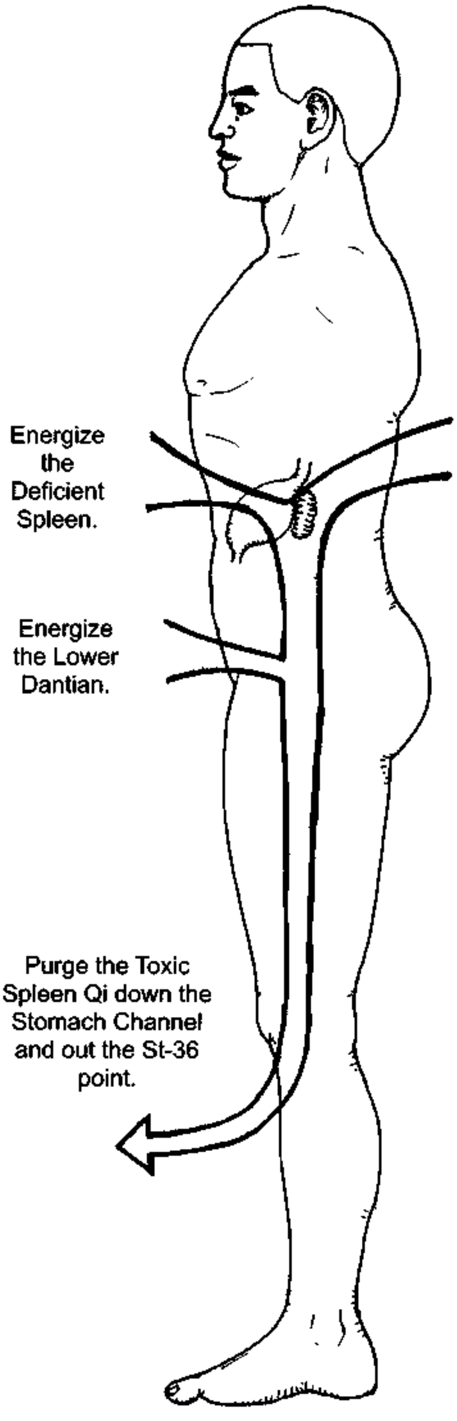
Energize the Deficient Spleen.
Energize the Lower Dantian.
Purge the Toxic Spleen Qi down the Stomach Channel and out the St-36 point.

Figure 45.2. Treating Spleen Conditions

themselves.

1. Generally, the Qigong doctor accesses the patient's Spleen through either direct energetic insertion into the Spleen's anatomical location, or through the patient's channels and external energetic fields.
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor stimulates the patient's Spleen and Yellow Court areas, and roots the patient's Middle Burner Qi into the Lower Dantian.
3. Next, the Qigong doctor leads any remaining Turbid Qi down the Stomach Channels and out of the patient's St-36 points (Figure 45.2).

Keep in mind that these areas are general and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

MEDICAL QIGONG THERAPY FOR SPLEEN QI DYSFUNCTION

The primary therapy used to treat the Spleen and gastric diseases include the Windy Breathing method (see Chapter 16) to strengthen the Liver, Spleen, Lungs, and the Kidneys.

1. Spleen diseases should be treated by vocalizing the "Hu" sound, which is effective for removing Turbid Qi accumulated in the Spleen, as well as for aiding digestion (Figure 45.3).
2. To treat insufficiencies of the Middle Burner, the doctor emits Qi into the patient's Zhong Wan CV-12 point and Lower Dantian points to reinforce the patient's Qi in the Middle Burner area.
3. Patients with an Excess Heat syndrome of the Spleen should be treated by guiding the Qi out along the Spleen and Stomach Channels with Pulling and Leading Qi manipulations to expel the Turbid Qi from both the Spleen and Stomach.

SPLEEN QIGONG PRESCRIPTIONS

The following exercise regulates the Qi and Blood of the Spleen Channels, strengthens the Spleen and replenishes Qi. It also regulates the Stomach and promotes digestion. It can be used to prevent and treat diseases and syndromes such as: gastritis, gastric and duodenal ulcers, colitis and gastrointestinal neurosis, abdominal distention, diarrhea and constipation.

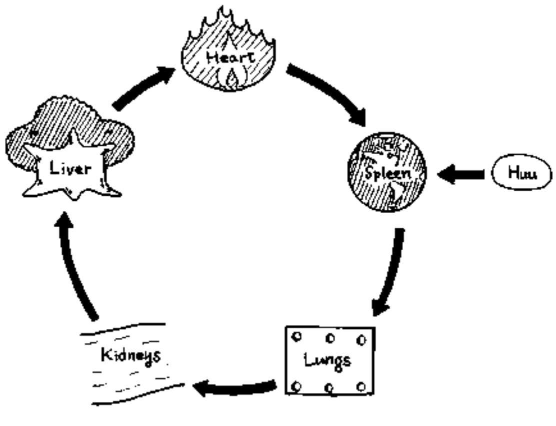

Figure 45.3. The "Hu" sound will purge the Spleen.

PRESCRIPTIONS FOR TONIFYING THE SPLEEN

The method of Taking in the Yellow Qi can be used to tonify the Spleen. The properties inherent within the yellow color have a persistent vibratory rate used for tonifying the Spleen and Stomach.

1. Assume a standing or sitting posture. Relax the whole body, breathe naturally, and get rid of any stray thoughts.
2. When inhaling, imagine yellow energy in front of you. Inhale the yellow light energy in through the nose and down into the Spleen organ. As you exhale through the mouth, the dark, pathogenic Qi leaves the Spleen; however, the bright clean color remains in the organ, stimulating and vitalizing it. With each breath the organ retains more clean energy and begins to glow. Repeat for five breaths.
3. Next breathe the color yellow in through the nose, filling the entire mouth. When exhaling through the nose, send the Qi slowly to the Zhongwan CV-12 point, then out to the four limbs, skin, and hair. Repeat five to ten times. Return to the beginning posture and close the training.

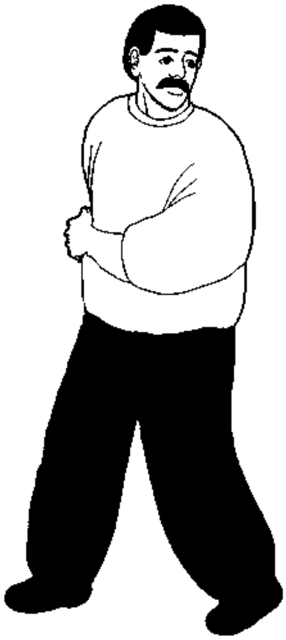

Figure 45.4. Twist at the waist, swing the arms, and look towards the opposite direction, focusing the mind's intention on the heel.

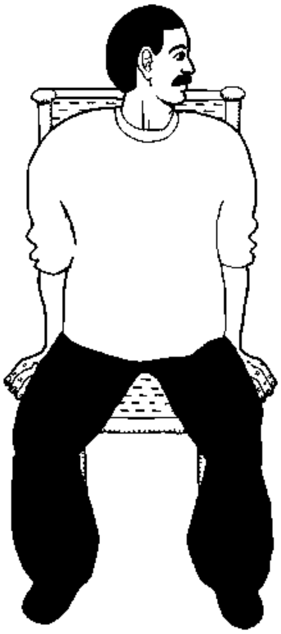

Figure 45.5. Turn and look over the shoulder with both eyes wide open.

PRESCRIPTIONS FOR PURGING THE SPLEEN

Purge the Spleen by having the patient rub his or her epigastrium while toning the sound "Hu."

Assume a sitting or standing posture. Gently place both palms (left on the outside for men, opposite for women) flatly on the Zhongwan CV-12 point on the upper abdomen. Exhale, rubbing the palms in a clockwise direction while sounding "Hu". Perform ten to twenty breaths.

Dredging the Spleen and Stomach

The following exercise is used to purge Toxic Qi from the Spleen and Stomach.

1. Assume a standing posture and breathe naturally. While relaxing the whole body, twist at the waist as you swing both arms in front of the body to the left, then to the right. Look in the opposite direction when turning the torso. As you turn your torso, your mind should be focused on your heels. Continue until you feel all tension has left your body (Figure 45.4).
2. Assume a sitting posture. While sitting on a bed, press the palms level on the mattress and meditate until you feel peaceful and relaxed. Turn the head towards the left and look over your left shoulder into the distance. Then turn the head to the right and look back over your right shoulder into the distance. Pause in the center, facing forward (Figure 45.5). When looking over the shoulders, both eyes should be wide open. Repeat five times in each direction.

PRESCRIPTIONS FOR REGULATING THE SPLEEN AND CONDUCTING QI

The following exercise is used to regulate the Spleen after purging or tonifying.

1. In a Wuji posture, begin by placing the hands on the Yellow Court and begin circle rotation in a clockwise direction for 36 rotations, ending at the Lower Dantian. Pause for 18 breaths to root the Spleen Qi in the Lower Dantian.
2. Next, rotate the abdomen in a counterclockwise direction for 36 rotations, ending at the Lower Dantian. Pause for 18 breaths and return to Wuji posture.

SPECIFIC MEDICAL QIGONG THERAPY FOR SPLEEN DISEASES

The following Medical Qigong treatment is used in China to treat specific types of Spleen disease.

DIABETES

Diabetes refers to a syndrome characterized by polydipsia (excessive thirst), polyphagia (excessive eating), and emaciation. It is a general term used to describe diseases characterized by insufficient secretion of insulin and excessive urination.

The prescriptions described in this section can also be used for the treatment of such disorders as: psychogenic polyuria, polyphagia due to hyperthyroidism, hyperhidrosis (excessive sweating), and emaciation.

ETIOLOGY

According to Traditional Chinese Medicine, diabetes is caused by disturbances in the Spleen and Stomach functions of transforming and transporting energy and the accumulation of toxic Heat in the body. These malfunctions can be due to eating too much sweet and greasy food, or drinking too much alcohol.

From the Western medical perspective, there are two types of diabetes, one type can be controlled through diet, the other cannot (usually hereditary). Diabetes is included in metabolic diseases (the Beta cells in the pancreas do not produce enough insulin) and caused by pathogenic factors relating to genetic predispositions, environmental stress, diet, obesity, and alcohol consumption.

In the Medical Qigong clinic there are three main factors that can contribute to the origination of diabetes: improper diet, emotional factors, and congenital predisposition.

1. The first factor is due to improper diet, including the excessive intake of alcohol, acrid, sour, pungent and greasy foods. The excessive intake of these types of foods can cause stagnation at the Middle Burner (which houses the Stomach and Spleen), and failure of the Stomach and Spleen to transport the accumulated Heat inside the body. This in turn can give rise to failure of the body's Yin fluids to nourish the Lungs and Kidneys, resulting in diabetes.
2. The second factor is due to a state of constant emotional upheaval. Emotional upsets can lead to stagnation of Qi which in turn impairs the body's Fluids, causing a Yin Deficiency with a hyperactivity of Evil Fire, resulting in diabetes.
3. The third factor is due to a genetic predisposition that can eventually surface if the patient is negligent with his or her diet and emotional health.

The Qigong doctor focuses on strengthening the patient's Spleen and Stomach, regulating the Liver Qi, and tonifying the Lung, Stomach, and Kidney Yin (which has been consumed by the Toxic Heat). It also is helpful to monitor the patient's eating habits to correct the disturbed metabolic rate and promote the regeneration and secretion by the patient's pancreatic islet cells.

SYMPTOMS

1. Diabetes involving the Upper Burner is characterized by excessive thirst, dry throat, dry tongue, and frequent urination.
2. Diabetes involving the Middle Burner is characterized by consuming abnormally large quantities of food during meals, emaciation, or constipation.
3. Diabetes involving the Lower Burner is characterized by frequent sweet smelling urination, or chyluria (passing of fat globules in the urine), and weakness and softness in the waist and knees.

TREATMENT

1. Begin by stimulating the Bl-23, Bl-20, Bl-13, St-36, and Sp-6 points on both sides of the patient's body.
2. Then purge and dredge the patient's CV-12, CV-4, and St-21 points.
3. Next, use the Extended Fan Palm hand manipulation with the Vibrating Palm technique to emit Qi into the patient's Bl-23, Bl-20, Bl-13, CV-12, and CV-4 for six to twelve breaths.
4. Continue emitting Qi into those points and into the Middle Dantian to regulate the Qi

activities of the Upper, Middle, and Lower Burners.

5. End the treatment by softly holding on to the patient's LI-11, LI-4, and Ht-1 points on both sides of the body, and with a gentle rocking or shaking motion allow the Qi to flow into the patient's body, settling into the patient's Lower Dantian.

PRESCRIPTIONS AND HOMEWORK

The "Inner Health" cultivation meditation is given to patients with diabetes, and is prescribed as follows.

1. Have the patient begin in a sitting posture, eyes closed, body relaxed using natural abdominal breathing.
2. While inhaling, the patient places the tongue against the hard palate, and begins to mentally repeat the mantra "I am quietly sitting."
3. While holding the breath, the patient mentally continues by saying the words "I am healthy."
4. The patient then exhales, releasing the tongue from the palate, while thinking the last phrase.

MEDICAL QIGONG THERAPY AND STOMACH DISEASES

This second section focuses on specific Stomach diseases and how to rectify them (Figure 45.6). The following discussion of treatment begins with an explanation of Stomach syndromes and symptoms. Several Stomach diseases have symptoms which are quite common to the organ's energetic dysfunction, such as epigastric pain.

STOMACH COMPLICATIONS AND SYMPTOMS

A list of Stomach syndromes and their symptoms are described as follows.

1. Stomach Heat can cause headache, epigastric pain, and bleeding.
2. Stomach Damp Heat can cause sinusitis and epigastric pain.
3. Stomach Phlegm Fire can cause epigastric pain and mental and emotional problems.
4. Stomach Deficiency can cause epigastric pain, diarrhea, atrophy syndrome, and bleeding.

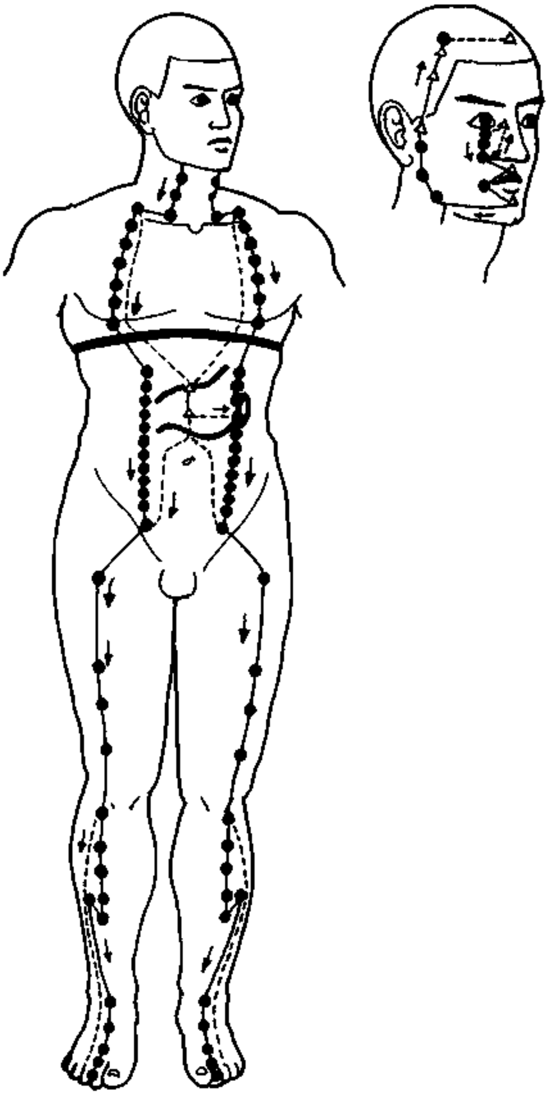

Figure 45.6. The Stomach (St) Channels

5. Stomach Yin Deficiency can cause tiredness, epigastric pain, and bleeding.
6. Cold invading the Stomach can result in epigastric pain.
7. Stagnant Blood in the Stomach can result in epigastric pain.

GENERAL TREATMENT FOR STOMACH DISEASES

The following is a description of a general protocol used for treating the patient's Stomach. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted (or Turbid Qi removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

After comparing the various Stomach therapies, the reader will notice that certain patterns repeat themselves.

1. Generally, the Qigong doctor accesses the patient's Stomach through either direct energetic insertion into the Stomach's anatomical location, or through the patient's channels and external energetic fields.
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor stimulates the patient's Stomach, Spleen, and Yellow Court areas, and then roots the patient's Middle Burner Qi into the Lower Dantian.
3. Next, the Qigong doctor leads any remaining Turbid Qi down the Stomach Channels and out of the patient's St-36 points (Figure 45.7).

Keep in mind that these areas are general and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

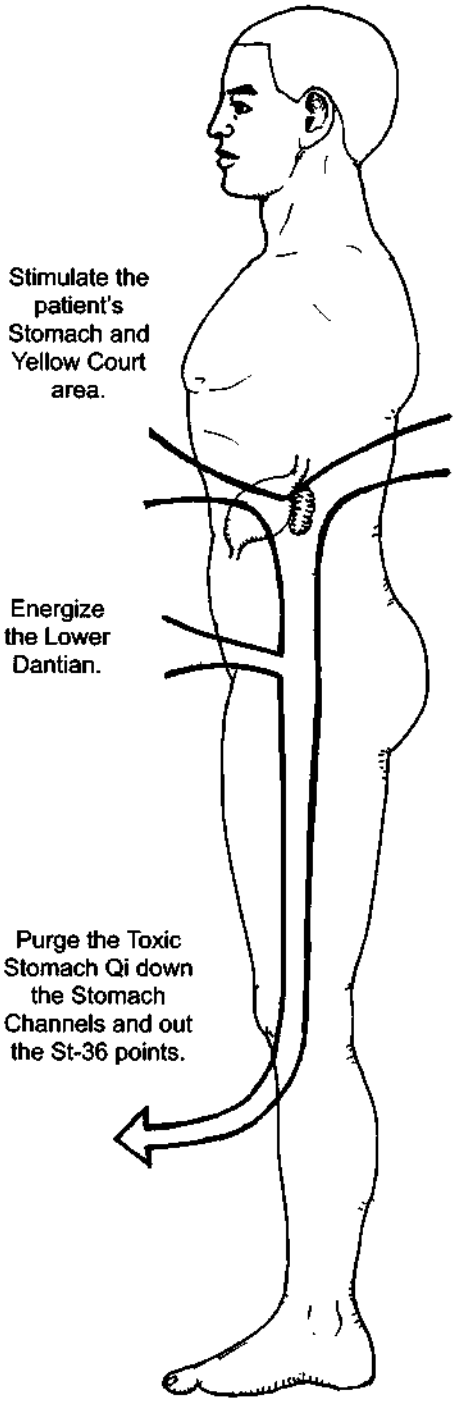
Stimulate the patient's Stomach and Yellow Court area.
Energize the Lower Dantian.
Purge the Toxic Stomach Qi down the Stomach Channels and out the St-36 points.

Figure 45.7. Treating Stomach Conditions

SPECIFIC MEDICAL QIGONG THERAPY FOR STOMACH DISEASES

The following are several Medical Qigong treatments used in China to treat specific Stomach diseases.

ABDOMINAL PAIN

Abdominal pain is an abnormal condition of the abdomen in which there is a sudden, abrupt onset of severe pain in any part of the abdomen below or just around the umbilicus.

Diseases such as acute pancreatitis, gastrointestinal spasm, stress induced abdominal pain, dyspeptic abdominal pain, and peritonitis, etc., in which abdominal pain symptoms appear, should be differentiated. Each differential diagnosis can be treated according to the exercises taught in the prescription section.

ETIOLOGY

The treatment for abdominal pain that is primarily intestinal in origin is different then that of gynecological abdominal pain, which is related to the menstrual cycle (see Chapter 51). There are four main factors which can contribute to the intestinal type of abdominal pain: abdominal pain due to the internal accumulation of Cold in the Lower Burner, abdominal pain due to dysfunction of Spleen Yang, abdominal pain due to food stagnation, and abdominal pain due to emotional stress.

1. Internal Cold in the Lower Burner may accumulate from eating too many cold foods and drinks, or can be due to the invasion of pathogenic Cold caused by exposure to the elements. Women in particular are more prone to Cold invading the abdomen during or immediately after menstruation. When Cold contacts the tissues, it slows down the movement of Qi and Blood, thus causing pain. External Cold may invade the patient's intestines directly without causing exterior symptoms, resulting in a sudden onset of pain and diarrhea.
2. Patients who are usually Deficient in Yang Qi often have a weak digestive system and a dysfunction of the Spleen Yang Qi. In this condition, pain often follows a pathogenic invasion, no matter how mild it may be. The pathogenic invasion may be due to overeating, overexertion, or brought on by eating after fasting and dieting, or simply by an Exterior Cold invasion.
3. Food stagnation is related to food intake or constipation, and is induced by Qi stagnation. The Stomach fails in digesting the food and the Spleen fails in separating and transporting the Clear and Turbid Qi.
4. Emotional stress is related to the circulation of Qi and Blood in the abdomen, and thus affects both the Liver and Spleen. Anger, frustration, and resentment can cause Liver Qi stagnation in the intestines, resulting in abdominal pain. Pensiveness and worry affect the Spleen and Lungs, impairing the transforming ability of the Spleen Qi and the downward descending ability of the Lung Qi, and likewise result in abdominal pain.

SYMPTOMS

1. Abdominal pain due to the internal accumulation of Cold, occurs abruptly, and may be relieved by defecation.
2. Abdominal pain due to Spleen Yang dysfunction, is characterized by continuous recurring pain, which may be relieved by pressure. Symptoms include loose stool, mental fatigue, and aversion to Cold.
3. Abdominal pain due to food stagnation, is characterized by gastric and abdominal distention, severe tenderness in the abdominal area, aversion to food and belching with acid regurgitation.
4. Abdominal pain due to emotional stress, is characterized by crying and vomiting.

TREATMENT

1. Have the patient lay on his or her side and begin stimulating the Bl-21, Bl-20, CV-12, St-36, and St-25 points on both sides of the body.
2. Purge the patient's Bladder Channels from the mid-back and waist area down to the feet three to five times.
3. Using the Vibrating Palm, emit Qi into the patient's CV-12, St-25, Bl-21, and Bl-20 points on both sides of the body. Purge and guide the Toxic Qi through the abdomen and down the lower extremities along the Stomach and Bladder Channels and out the patient's feet.
4. Using the Extended Fan Palm technique, extend Qi into the patient's CV-12 and both St-25 points for 14 breaths. Using Pulling and Leading manipulations, guide the patient's Qi down along the Stomach Channels and out both of the patient's feet.
5. Finally, Circle Rotate the patient's abdomen for 81 breaths in a clockwise direction.

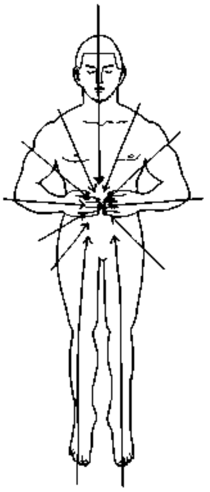

Figure 45.8. To tonify Deficient Stomach Qi, the patient must imagine drawing Qi into the body through the Baihui point at the top of the head, and the Yongquan points at the bottom of the feet. All three rivers of energy must meet to form a pool of energy within the Stomach area, which the body will absorb and store.

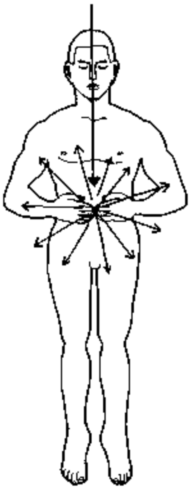

Figure 45.9. To treat Excess Stomach Qi, begin from a supine posture. While inhaling, imagine the Qi flowing down the Stomach Channels from the Baihui point ending just below the sternum at the Stomach area. Exhale and imagine the Turbid Qi dissolving and melting into the Lower Dantian, then disperse it from the body, out the pores.

TREATMENT MODIFICATIONS

1. In cases of abdominal pain due to the internal accumulation of Cold, add the method of extending Hot Qi.
2. In cases of pain due to Spleen Yang dysfunction, extend Hot Qi into the patient's Bl-21, Bl-20, and Mingmen area.
3. In cases of pain due to food stagnation, emit Qi using the Extended Fan Palm with clockwise then counterclockwise Spiraling methods into the patient's CV-12, CV-4, and St-25 points.
4. In cases of pain due to emotional stress, emit Qi using the Extended Fan Palm hand technique into the patient's Ht-7 and Pc-6 points on both sides of the body. Purge and Circle Massage the Shendao (GV-11) area in a counterclockwise direction.

PRESCRIPTIONS AND HOMEWORK

1. Have the patient place his or her right palm on the CV-12 area and begin to Circle Massage 36 times in a clockwise and then 36 times in a counterclockwise direction.
2. After completing the circle rotations, the patient pauses for a moment, and then leads the Qi down to the navel and continues the 36 rotations, in both clockwise and counterclockwise directions.
3. Next, the patient inhales and presses the fingers of both hands gently below the xiphoid process. The patient exhales and directs the Qi down along the midline of the body to the pubic symphysis area for 36 times.
4. Finally the patient overlaps his or her hands (right over the left in men and opposite in women) and places them on the midpoint of the lower abdomen to knead the tissue area clockwise and then counterclockwise 36 times.

PRESCRIPTIONS MODIFICATIONS

- In cases of Stomach pain due to Deficient Stomach Qi, have the patient lie supine. The patient must imagine drawing Qi into his or her body through the Baihui point at the top of the head, and the Yongquan points at the bottom of the feet. All three rivers of energy must meet to form a pool of energy within the Stomach area, which the body will absorb and store into the Stomach's spacial cavity (Figure 45.8).

- In cases of Stomach pain due to Excess Stomach Qi and to treat Excess Stomach Qi symptoms, have the patient lie supine. As the patient inhales, he or she is to imagine the Qi flowing down the Stomach Channels (from the Baihui point) ending just below the sternum at the stomach area. As the patient exhales, he or she imagines the Turbid Qi trapped within the Stomach dissolving and melting into the Lower Dantian. Next, the patient disperses the Toxic Qi from the body out through the pores (Figure 45.9).

CHRONIC GASTRITIS

In Traditional Chinese Medicine, this disease belongs to the category of epigastralgia. Gastritis is described as an inflammation of the Stomach. Inflammation of the gastric mucosa is the main clinical manifestation of chronic gastritis, and results in chronic epigastralgia and dyspepsia (air hunger, resulting in difficult breathing, shortness of breath, sometimes accompanied by pain).

ETIOLOGY

Generally, emotional factors such as worry, anxiety, or anger may cause stagnation of Qi. Stagnant Qi, by impeding the function and flow of the body's energy, causes Qi deviations and obstructions that result in a stomach ache.

Gastritis can be superficial (affecting only the superficial layers of the Stomach) or atrophic (causing thinning of the Stomach mucosa). Its main manifestations are epigastric pain, which occurs about one hour after eating, followed by nausea, belching, and vomiting with sour regurgitation.

SYNDROMES

1. Liver and Stomach Qi stagnation, manifests as a distending pain, fullness, and discomfort in the Stomach, which is aggravated after a meal. The pain is not fixed to any specific Stomach location and worsens under emotional stress, causing belching with acid regurgitation and nausea.
2. Stomach Yin Deficiency due to Stomach Heat manifests as a stomach ache accompanied by a scorching sensation. The pain is not fixed to any specific Stomach location and is aggravated when the Stomach is empty, and is alleviated after a meal. The symptoms are expressed as irritability, dysphoria, and excess appetite.
3. Weakness of the Spleen and Stomach manifests as a dull stomachache which is relieved by Heat and pressure. Symptoms are expressed as abdominal distention, dim complexion, lassitude, cold limbs, and anorexia. If the patient has congenital insufficiency of Yang in the Middle Burner (the functional area of the Spleen and Stomach), Cold may originate from the interior of the body.

TREATMENT

1. Have the patient lie supine and relax his or her whole body, focusing especially on the Stomach area. The patient imagines the Toxic Qi dispersing upon exhalation.
2. Begin to purge and dredge the patient's CV-15 point (just below the xiphoid process), leading the Qi down to the Lower Dantian.
3. Next, use a Vibrating Palm hand technique to emit Qi into the patient's CV-12 point and both St-21 points for 14 to 28 breaths.
4. Applying the Extended Fan Palm hand technique to the patient's CV-12 and Middle Dantian, begin drawing Qi downward along the Conception Vessel and Stomach Channels to the bottom of the patient's feet.
5. Finally, have the patient lie prone and stimulate the Bl-20 and Bl-21 points on the lumbar-dorsal region, directing the energy from the upper back down to the Mingmen area.

TREATMENT MODIFICATIONS

1. For Stagnant Liver and Stomach Qi, use the Extended Fan Palm hand technique to emit Qi into the patient's Bl-18 points, pulling and conducting Qi along the Gall Bladder and Liver Channels down the legs and out the feet on both sides.
2. For Stomach Yin Deficiency due to Stomach Heat, use the Vibrating Palm hand technique

to emit Qi into the patient's Bl-20 and Bl-23 points, leading the Qi down the patient's Bladder Channels to the feet.

3. For Spleen and Stomach Deficiency, use the Extended Fan Palm hand technique to emit Qi into the patient's Lower Dantian area.

PRESCRIPTIONS AND HOMEWORK

Have the patient practice the Descending the Yang and Ascending the Yin technique (see Chapter 41).

PRESCRIPTION MODIFICATIONS

1. For patients with Liver and Stomach Qi stagnation, have the patient practice the Spleen and Stomach Massage (see Chapter 39). Then perform the healing sounds "Hu" to regulate the Spleen and "Shu" to regulate the Liver.
2. For treating patients with Stomach Yin Deficiency due to Stomach Heat, have the patient first dredge the Spleen and Stomach and then perform the "Chui" (pronounced Chree) healing sound to regulate the Kidneys. Next, have the patient practice the tonification method of Taking in the Yellow Qi (see Chapter 4).
3. For treating patients Deficient Spleen and Stomach Qi, have the patient first practice the Daoist Five Organ Regulation exercise (see Chapter 42). Next, have the patient practice the tonification method of Taking in the Yellow Qi to strengthen the Spleen and Taking in the Red Qi to strengthen and regulate the Heart.

GASTRIC AND DUODENAL ULCERS

An ulcer is an open sore or lesion of the skin or a mucus membrane, accompanied by sloughing of inflamed, dead, or dying tissue, that causes the Stomach's lining to become more vulnerable to acids. If the sore becomes infected puss is discharged.

ETIOLOGY

A gastric or duodenal ulcer refers to a chronic ulcer seen only in the regions where the gastrointestinal tract contacts the patient's gastric juice. Its formation and development is closely related to the digestive functions of the acidic gastric juices and pepsin. Clinically this type of ulcer is characterized by chronic recurring epigastralgia, lasting for a few weeks, with pain occurring from 30 minutes to two hours after eating.

Currently, Stomach ulcers are also being treated through either herbs or antibiotics (from Western medicine). They are caused by bacterial infections resulting from either Qi and Blood stagnation or a Deficient Cold syndrome of the Spleen and Stomach.

SYNDROMES

- Qi and Blood stagnation results in a distended pain and fullness in the Stomach, tenderness, belching, and acid regurgitation, which becomes aggravated by emotional stress.
- A Deficient Cold syndrome of the Spleen and Stomach results in a dull pain within the Stomach which is relieved by heat and compression and aggravated by cold. There can also be vomiting with watery regurgitation, shallow complexion, cold extremities, lassitude, and weakness.

TREATMENT

1. Have the patient sit at the end of the treatment table, and begin extending energy into the patient's mid-back along their Bl-18, Bl-20, and Bl-21 points using the Extended Fan Palm technique.
2. Next, use the Vibrating Palm to extend energy into the patient's Yellow Court area. When the area has been sufficiently stimulated, purge the Toxic Qi out of the patient's body by leading the Qi down to the Lower Dantian for 14 to 28 breaths.
3. Finally, use the Extended Fan Palm hand technique to emit Qi into the patient's CV-12 and St-34 points for 14 breaths. Then lead the Qi down the Conception Vessel into the Lower Dantian to root and stabilize the patient's energy.

TREATMENT MODIFICATIONS

- When treating patients with Qi and Blood stagnation, use the Extended Fan Palm hand method to emit Qi into the CV-12, CV-13, and St-21 points. Next, direct the Qi along the patient's Conception Vessel and Stomach Channels, leading the Toxic Qi down the patient's legs and out of the body via the St-36 points.

- When treating patients with a Deficient Cold syndrome of the Spleen and Stomach, use the Extended Fan Palm hand method to emit Qi into the patient's Mingmen and Lower Dantian areas.

PRESCRIPTIONS AND HOMEWORK

In treating patients with ulcers, have them practice the Old Man Searching for the Reflection of the Moon at the Bottom of the Tide Pool exercise (see Chapter 53).

PRESCRIPTION MODIFICATIONS

- For patients with Qi and Blood stagnation, have them perform the Spleen and Stomach Massage (see Chapter 39) while sounding "Hu." Next, have the patient dredge the Spleen and Stomach using the sound "Shu" to purge the Liver and regulate Liver Qi.
- For patients with a Deficient Cold syndrome of the Spleen and Stomach, have them practice the method of Taking in the Yellow Qi from the natural environment.

CHRONIC CONSTIPATION

Chronic constipation refers to the prolonged retention of feces in the intestinal tract and difficulty in discharging the fecal matter. This condition is not only painful to the patient, but also increases toxicity in the body. It is clinically characterized by decreased times of fecal discharge or small amounts of feces which are considered either too dry and hard, or too difficult to be discharged.

Generally, the transitional time for food to pass through the digestive tract for humans takes 24 to 40 hours. The postponement of defecation over two days is considered constipation, although some normal patients may have regular defecation every two to three days.

ETIOLOGY

The common causes of constipation are: Diet, Febrile Diseases, Emotional Stress, and A Sedentary Life-style.

1. Diet plays a major role in causing constipation. Constipation can result from the excessive consumption of hot, spicy, or greasy foods. This excess consumption of hot and greasy foods can dry up the fluids of the Stomach and intestines, thereby drying up the stool so that it cannot properly move through the intestinal walls.

   An excess consumption of Cold foods can block the Spleen's function of transporting and thus prevent the stool from moving downwards.

   Finally, a lack of fiber in the food, lack of sufficient fluid intake or excessive alcohol consumption can cause constipation.
2. Febrile diseases can be caused by External Wind Heat invasion, which if not rapidly expelled from the body, can turn into an Interior Heat condition. The Interior Heat usually attacks the Lungs, Stomach, and intestines causing the Body Fluids to dry up. A Deficiency of Body Fluids from the retention of Heat within the intestines (even after the febrile disease has passed) can cause dry stool and constipation.
3. Liver Qi stagnation (due to mental depression, anger, resentment, or frustration) can cause a Qi obstruction in the Lower Burners that affects bowel movements and leads to constipation, abdominal distention, and pain. Excessive studying, worrying, and thinking affects the Spleen, slowing down its transportation of food in the intestines, this in turn can lead to constipation.
4. A sedentary life-style, whether from choice or due to a weak constitution from a chronic or severe illness (or surgery), can cause constipation. Patients in convalescent homes often suffer from constipation due to a lack of exercise and poor diet, which can lead to Spleen Qi and Blood Deficiency. Lack of exercise weakens the Spleen Qi and can cause Liver Qi stagnation. When Spleen Qi becomes Deficient for long periods of time, the Spleen energy fails to provide the Qi to move the stool, resulting in constipation.

SYNDROMES

- Constipation due to an Excess condition is caused by decreased times of fecal discharge (once every three to seven days), and is usually accompanied with fever, thirst, poor appetite, and emotional anxiety.

- Constipation due to a Deficient condition is mostly seen in patients who have just recovered from an illness, or have just given birth and have become Deficient in Qi and Blood. It may also be observed in patients who suffer from a congenital insufficiency of Yang Qi in their Lower Burner and an accumulation of Yin Cold, which makes it difficult for the intestines to perform their normal transporting function. Patients with a Deficient condition may have bowel movements every three to seven days. The stool may be very dry, hard, and difficult to pass. This condition is clinically manifested by a pale and dull complexion, listlessness, and cold pain in the abdomen which is relieved by heat. The patient may also experience emotional fright due to a Kidney Qi Deficiency.

TREATMENT

1. Have the patient sit at the end of the treatment table, and begin to stimulate the patient's mid-back between the Bl-23 and Bl-25 points. Then stimulate the patient's lower lumbar and coccygeal areas, dredging and purging the Toxic Qi down the patient's thighs and the feet, while stimulating the Bl-31 through Bl-34 points.
2. The doctor will then have the patient lie supine and begin to stimulate the patient's Yellow Court and Lower Dantian areas, using the Vibrating Palm hand technique for 14 breaths.
3. Next, apply the Kneading Tiger Palm with the Vibrating and Quivering hand manipulations to emit Qi into the back of the patient's head at the Bl-10 points for 12 breaths, and then onto the patient's Lower Dantian for 14 breaths.
4. After energizing the patient's Lower Dantian, emit Qi, using the Extended Fan Palm technique, into the patient's middle epigastrium and navel area, as well as into the Lower Dantian. Finally, use the Spiraling and Quivering hand manipulation in a clockwise direction to gather the stagnant Qi and to lead it down the patient's Stomach Channels and out the legs.

TREATMENT MODIFICATIONS

- When treating patients with an Excess type of constipation, use the Extended Fan Hand Palm technique to emit Cold Qi into the patient's lower abdomen. Rotate the hand in a counterclockwise direction, while purging the Toxic Qi from the patient's body. Repeat for 14 breaths.
- When treating patients with a Deficient type of constipation, use the Vibrating Palm technique to emit Qi into the patient's Lower Dantian and Mingmen area for 8 to 16 breaths.

PRESCRIPTIONS AND HOMEWORK

To treat chronic constipation, have the patient practice the Daoist Five Yang Organ Regulating exercise (see Chapter 42) for 250 breaths twice a day.

PRESCRIPTION MODIFICATIONS

- For patients with an Excess type of constipation, place them in the Wuji posture with their hands resting on their lower abdomen. Begin with a clockwise rotation, imagining the Toxic Qi draining from the patient's Spleen and Stomach areas, down into their Stomach Channels and out their feet.
- For patients with a Deficient type of constipation, have them practice rubbing their Lower Dantian to strengthen the abdominal Qi, and have them practice the Gathering the Sun's Essence meditation (see Chapter 11).

CHAPTER 46
QI EMISSION THERAPY AND LUNG DISEASES

INTRODUCTION

The Lungs are solid organs that control the cycles of Qi circulation in the body, and are sometimes referred to as the White Emperor. They correspond to the element Metal. The Lungs house the Po (Corporeal Souls) and are in charge of respiration and regulating the metabolism of the Body's Fluids (Figure 46.1). The Lungs are also responsible for keeping Qi pure and extending it downward. The Lungs are energetically paired with the Large Intestine (a hollow organ).

The Lungs' upper external connection with the outside world is through the nose, hence any fragrance or odor inhaled through the nose affects the senses and evokes emotional responses from the Po.

In the fall months, the Lung Qi becomes more active in individuals that already possess strong Lung Qi, but becomes Deficient in those who already have weak Lung Qi. The excessive eating and drinking of pungent foods and the overexposure to dryness drains the Lungs.

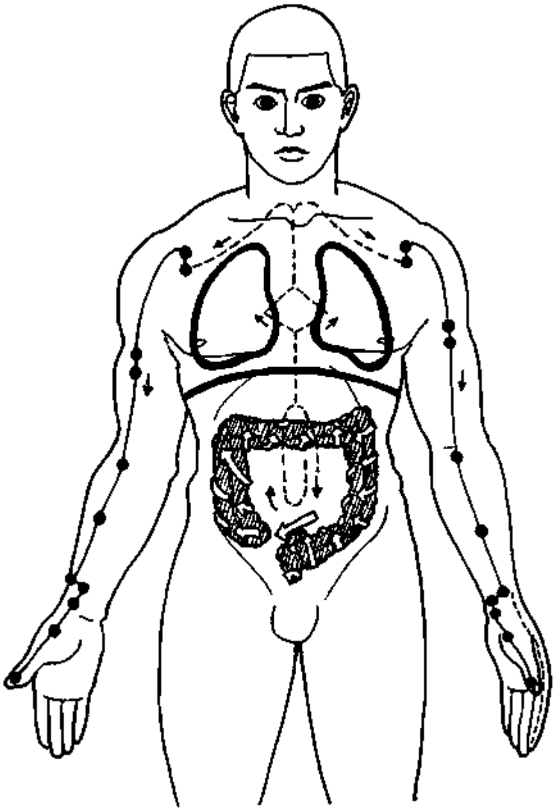

Figure 46.1. The Lung (Lu) Channels

LUNG COMPLICATIONS AND SYMPTOMS

This section focuses on specific Lung diseases and how to rectify them. The following discussion of treatment begins with an explanation of Lung syndromes and symptoms. Several Lung diseases have symptoms which are quite common to the organ's energetic dysfunction, such as headaches, cough, the common cold and influenza. A list of Lung syndromes and symptoms are described as follows.

1. Wind Cold invading the Lungs usually enters the patient through the pores of the skin and body hair (which are ruled by the Lungs). If the Lungs' energetic field is invaded by Wind Cold, it may cause the Lung Qi to become impure, ascend, and cause obstructions in the pores, resulting in a Wind Cold syndrome. This can cause the following symptoms: severe aversion to cold, headache, cough, the common cold or influenza, breathlessness, asthma, and allergic rhinitis.
2. Wind Heat Invading the Lungs usually enters the body by way of the mouth and nose; the nose being the orifice to the Lungs. If the Lungs' energetic field is invaded by Wind Heat, it may cause the Lung Qi to become impure, ascend, and impede the ability of the

pores to expel pathogenic Evils, resulting in a Wind Heat syndrome. This can cause the following symptoms: slight aversion to cold, headache, cough, the common cold or influenza, asthma, allergic rhinitis and sinusitis.

3. Wind Dampness is a type of Wind Cold syndrome. It consists of both Wind and Dampness together, invading either the skin (causing itching and rashes) or the channels and joints (causing painful obstruction syndrome). If the Lungs' energetic field is invaded by Wind Dampness, it causes the following symptoms: skin rashes, swollen joints, body aches, headaches, the common cold and influenza.
4. Wind Dryness invading the Lungs causes the following symptoms: dry, itchy sore throat, dry lips, dry mouth, the common cold and influenza, cough and bleeding.
5. Lung Heat causes the following symptoms: restlessness and a feeling of heat, common cold and influenza, cough, sinusitis, breathlessness and atrophy syndrome as the body's functions begin to shut down.
6. Lung Phlegm Heat causes the following symptoms: common cold and influenza, cough with expectoration, breathlessness and a sensation of oppression in the chest.
7. Cold-Damp Phlegm invading the Lungs causes the following symptoms: a feeling of heaviness, cough with profuse expectoration, wheezing, nausea, poor appetite and breathlessness.
8. Lung Qi Deficiency causes the following symptoms: breathlessness, wheezing, slight cough with no phlegm, asthma, allergic rhinitis, tiredness, spontaneous sweating, and enuresis or incontinence.
9. Lung Yin Deficiency causes the following symptoms: night sweating, extreme tiredness, breathlessness, asthma, dry cough in short bursts, and Five Palms Hot (a condition also known as "Five Center Heat," in which the patient feels a hot sensation in the palms and soles of the feet, accompanied by heat and agitation of the chest and/or head area).

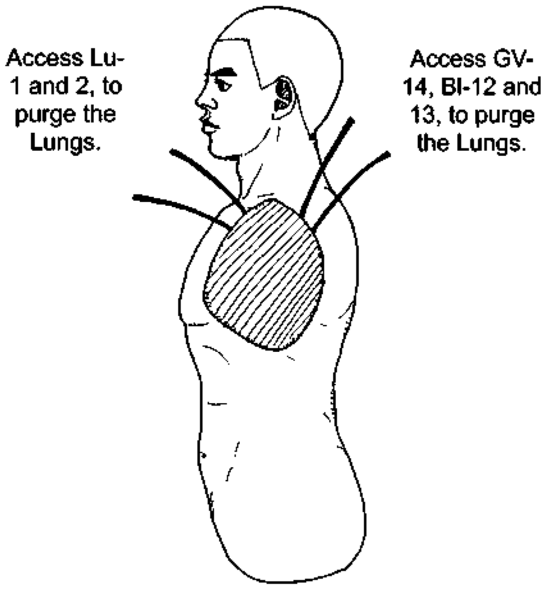
Access Lu-1 and 2, to purge the Lungs.
Access GV-14, Bl-12 and 13, to purge the Lungs.

Figure 46.2. Treating Lung Diseases

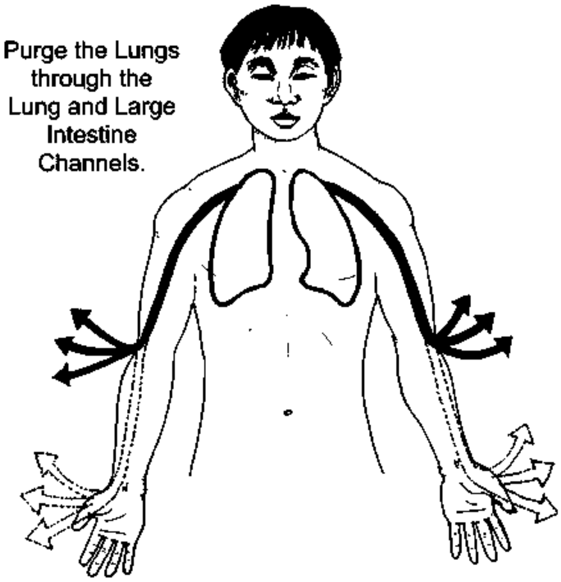
Purge the Lungs through the Lung and Large Intestine Channels.

Figure 46.3. Drain the Turbid Qi out of the Lungs, down the Lung Channels and out of the hands.

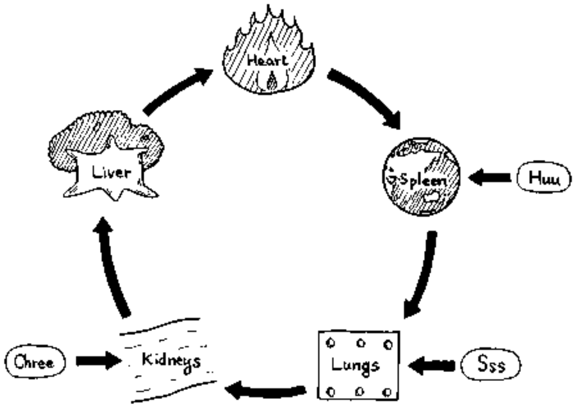

Figure 46.4. Replenish the mother (Spleen) to nourish the child (Lungs).

GENERAL TREATMENT FOR LUNG DISEASES

The following is a description of a general protocol used for treating a patient's Lungs. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted (or Turbid Qi removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

After comparing the various Lung therapies, the reader will notice that certain patterns repeat themselves.

1. Generally, the Qigong doctor accesses the patient's Lungs through either direct energized insertion into the Lungs' anatomical location, or through the patient's channels and external energetic fields (Figure 46.2).
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor stimulates the patient's Lungs and Upper Burner areas and roots the patient's Qi into the Lower Dantian.
3. Next, the Qigong doctor leads any remaining Turbid Qi down the Lung Channels and out the patient's hands (Figure 46.3).

Keep in mind that these areas are general and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

MEDICAL QIGONG THERAPY FOR LUNG QI DYSFUNCTION

Patients with Lung disease may have a pale complexion, dry hair, dyspnea (difficult breathing), hyposmia (a deficient sense of smell), stuffy nose, dryness and itching of the skin, or discomfort and pain in the chest and back. Patients may dream of the sun, moon, clouds, or of having sex. The primary therapies used to treat Lung diseases include the Breath regulation and Qi reinforcing meditations (see Chapter 16).

- Patients with an Excess Heat syndrome of the Lungs may have a flushed right cheek. To treat this type of syndrome, have the patient utter the sounds "Si" (Sss) for the Lungs and "Chui" (Chree) for the Kidneys. Guide the Evil Qi out of the patient's body along the Lung and Large Intestine Channels using the Pushing, Pulling and Guiding Qi manipulations.
- Patients with a Deficient syndrome of the Lungs may have shortness of breath and be unable to regulate their breathing. To treat this type of syndrome, have the patients inhale the sound "Hu," as they visualize absorbing the color yellow to replenish the Spleen (the mother); this will nourish the Lungs (the child). Next, the patient guides the Qi back to the Lungs (Figure 46.4).

LUNG QIGONG PRESCRIPTIONS

These exercises regulate the Qi in the Lung Channels, tonify the Lungs, replenish Qi, ventilate the Lungs, and relieve asthma, coughing, and sputum. These exercises can be used to prevent and treat the following diseases and syndromes: bronchitis, pulmonary emphysema, and bronchial asthma, as well as dyspnea and abundant expectoration.

PRESCRIPTIONS FOR TONIFYING THE LUNGS

The Taking in the White Qi meditation can be used to tonify the Lungs. The properties inherent within the color white have a persistent vibratory rate used for tonifying the Lungs and Large Intestine.

1. Assume a standing, sitting, or lying posture.

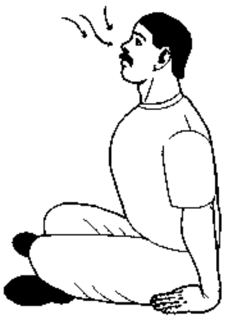

Figure 46.5. Inhale

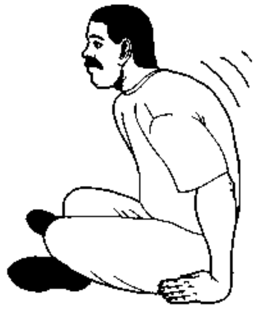

Figure 46.6. Exhale

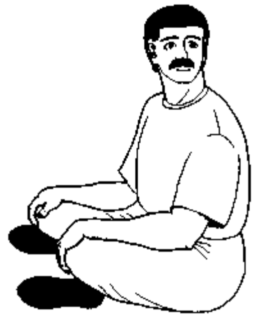

Figure 46.7. Inhale

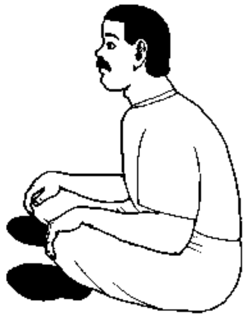

Figure 46.8. Exhale

Relax the whole body, breathe naturally, and get rid of any stray thoughts.

2. When inhaling, imagine a white energy in front of you. Inhale the white light energy in through the nose and down into the Lungs. As you exhale through the mouth, the dark, pathogenic Qi leaves the Lungs; however, the bright clean color remains in the Lungs, stimulating and vitalizing them. With each breath the Lungs retain more clean energy and begin to glow. Repeat for five breaths.
3. Next breathe the color white in through the nose, filling the entire mouth. When exhaling through the nose, send the Qi slowly into the Lungs, then down to the Lower Dantian, and then further out into the skin and hair of the whole body. Repeat nine to eighteen times. Return to the beginning posture and close.

PRESCRIPTIONS FOR PURGING THE LUNGS

Purge the Lungs by having the patient rub his or her chest while sounding "Sss."

The patient should assume a sitting or standing posture. Both palms are placed flat on the sides of the chest as the patient inhales slowly. When exhaling, the patient sounds the word "Sss" while rubbing the chest with both palms (circling up and down, from the center outward). This exercise is performed for six to twelve breaths.

PRESCRIPTIONS FOR REGULATING THE LUNGS AND CONDUCTING QI

The following exercise is used to regulate the Lungs after purging or tonifying.

1. Sit with the legs crossed, breathe in with the spine arched backwards and the hands by the sides pressing against the ground; extend the chest, and inhale until the chest is full (Figure 46.5).

   Pause, then push the upper back in a posterior direction and draw in the chest while exhaling through the nose (Figure 46.6). Repeat four to nine times.
2. Next, still sitting with the legs crossed, place the palms on the knees. Inhale with the spine erect and turn the head allowing the torso to turn slightly to the left (Figure 46.7). Then bend forward and exhale while facing the center (Figure 46.8). Next, inhale and turn to the right, then again bend forward and exhale while facing the center. Always inhale when turning to the sides, and exhale while facing the center.

SPECIFIC MEDICAL QIGONG THERAPY FOR LUNG DISEASES

The following are several Medical Qigong treatments used in China to treat specific Lung diseases.

TREATING THE COMMON COLD AND INFLUENZA

Upper respiratory tract infections belong to the category of common colds, while influenza belongs to the category of seasonal epidemic colds. Both conditions may be treated by referring to the prescriptions described within this section.

ETIOLOGY

The common cold is a generic term used to describe an Exterior condition caused from the invasion of pathogenic Wind (Wind Cold, Wind Heat, Wind Damp Heat or Wind Dry Heat). It is generally associated with coryza or inflammation of the respiratory mucous membranes.

Upper respiratory tract infection is caused by viruses or bacteria. Influenza is known as an acute, contagious respiratory infection, characterized by sudden onset. The chief clinical manifestations are stuffy nose, nasal discharge, coughing, sneezing, sore throat, fever, headache, chills, and an aversion to wind.

If the pathogenic Qi is on the Exterior level of the body, the internal organs are not affected, and only the patient's Wei Qi is involved. If the pathogenic factor becomes Interior, the patient's internal organs are affected, especially the Lungs and Spleen. If the pathogen is not cleared and purged from the patient's body, it may either penetrate more deeply into the patient's body and cause serious problems, or give rise to residual Heat which causes chronic post-viral fatigue syndromes.

SYNDROMES

- A Wind Cold syndrome has the following symptoms: chills, headache, aches and painful joints, coughs and sore itchy throat, stuffy nose, low fever, and no sweating with profuse, watery nasal discharge.
- A Wind Heat syndrome has the following symptoms: fever, cough, headache, slight aversion to cold, swollen and sore throat, thick yellow sputum, slight thirst, and perspiration. Once the pathogenic factor penetrates into the patient's interior, the body's Qi begins to fight against it, causing a fever and feeling of heat inside the body, with an aversion to cold, and shivering.

TREATMENT

1. Have the patient sit at the end of the treatment table.
2. Begin by stimulating the energy within the patient's head, purging and leading the Toxic Qi down the patient's arm and out the LI-11 and LI-4 points.
3. Then stimulate the patient's Yintang (Third Eye) and Taiyang (temple) points. Using Pulling and Leading Qi manipulation techniques, purge and guide any remaining Toxic Qi downward, along the Conception Vessel and the Stomach Channels, out the patient's feet, three to seven times.
4. Next, to balance the functional activities of the body's Qi, stimulate the tissue area on the patient's upper back and head (GV-14 and GV-16, Bl-12 and Bl-13) using the Extended Fan Palm technique, and guide the Qi downward along the Bladder Channels to the patient's feet.
5. Finally, softly grasp the patient's LI-4 and LI-11 points and gently shake the arm, stimulating the patient's Lung Qi. This technique is performed on each of the patient's arms.

TREATMENT MODIFICATIONS

1. When treating patients with Wind Cold syndrome, extend Hot Qi into the GV-14 and Bl-12 points, using the Extended Fan Palm technique, and guide the Qi downward along the Bladder Channels to the patient's feet and out the body.
2. When treating patients with Wind Heat syndrome, extend Cold Qi into the GV-14, Bl-13, and GB-20 points, using the Extended Fan Palm technique, and guide the Qi downward along the Stomach Channels to the patients' feet and out the body.
3. When treating patients with nasal obstruction, extend Vibrating Qi, using the Sword Fingers technique, into the LI-20 points for 9 breaths.
4. When treating patients with cough, chest stuffiness, and Phlegm, extend Vibrating Qi, using the Sword Fingers technique, into the CV-17, Lu-1, and Lu-2 points for 8 breaths.

PRESCRIPTIONS AND HOMEWORK

The patients should be given the Ten Dragons Run Through the Forest exercise (see Chapter 41) to sweep and purge the Gall bBladder

Channels. Most of the attention should be placed on feeling the Qi beneath the hands, with emphasis placed on exhalation to expel the pathogenic Wind from the body. Perform this prescription for 24 to 50 times.

CHRONIC BRONCHITIS

Chronic bronchitis is a condition caused by the chronic inflammation of the mucous membranes of the bronchial tubes.

ETIOLOGY

Chronic inflammation of the bronchial mucosa and peribronchial tissues is mostly due to lingering and recurrent acute bronchitis, or may occur after an infection from other diseases. The chief clinical manifestations are chronic recurring cough, expectoration and dyspnea. Bronchitis can be divided into acute and chronic conditions.

1. Acute bronchitis corresponds to the Qi level of infection within the four level identification patterns. It is the easiest to treat, and normally manifests with Hot Phlegm symptoms.
2. Chronic bronchitis takes a little longer to heal and the doctor should focus treatment on attending to the root and manifestation of the disease simultaneously (purging Phlegm and dispersing the Heat or Cold).

SYNDROMES

- For Phlegm Dampness attacking the Lungs, the main symptoms are abdominal distention, cough with copious white sputum, stuffiness in the chest and loss of appetite.
- For Lung and Spleen Deficiency, the main symptoms are cough with sputum, shortness of breath, spontaneous sweating, loss of appetite and loose stool.

TREATMENT

1. Have the patient sit at the edge of the table. Begin to purge the patient's Lungs, dredging the Toxic Qi down the Lung Channels and out of the hands.
2. Use the Vibrating Palm technique to stimulate the patient's Middle Dantian, Lu-1, -2, Bl-13, GV-16, and St-40 points. Use 8 to 12 breaths for each point.

TREATMENT MODIFICATIONS

- For Phlegm Dampness attacking the Lungs, use the Extended Fan Palm technique to emit Qi into the patient's Lungs; purge the Spleen and Stomach Channels, leading the Toxic Qi down the channels and out of the St-40 and Sp-6 points on each leg.
- For Spleen and Lung Deficiency, emit Qi using the Vibrating Palm technique into the patient's Bl-20, Bl-23, Bl-13 and CV-12 points for 6 to 12 breaths.

PRESCRIPTIONS AND HOMEWORK

Have the patient practice the Lung exercise from Daoist Five Yin Organ Tonification and Regulation exercises (see Chapter 42) to strengthen and balance the Lungs.

PRESCRIPTION MODIFICATIONS

- For patients suffering from Phlegm Dampness attacking the Lungs, have the patient perform the Sun and Moon Rotating technique (see Chapter 41), sounding "Sss" to regulate the Lungs and the sound "Hu" to regulate the Spleen.
- For patients suffering from Lung and Spleen Deficiency, have them perform the Taking in the White Qi to tonify the Lungs, and the Taking in the Yellow Qi to tonify the Spleen.

ASTHMA

Asthma is a chronic disorder of the respiration organs characterized by difficulty in breathing, wheezing, and tightness in the chest. Bronchial asthma, chronic bronchitis and pulmonary emphysema are all included in the category of this disease.

ETIOLOGY

Asthma is an Excess condition which is caused mainly by a retention of Wind, Cold, Heat or Damp Phlegm in the Lungs. When a patient is affected by invasion of pathogenic Dampness, disorders in Qi mobility develop. Phlegm mixes with Qi and obstructs the air passages, causing dyspnea (shortness of breath) and wheezing.

Asthma may also result from an obstruction of the Lungs due to Phlegm retention caused by a failure of the Spleen to transport, or from a Deficient condition of the Kidneys. The diagnosis of this disease can be divided into two types of asthma: early onset asthma and late onset asthma.

1. Early onset asthma (also called extrinsic or atopic asthma), usually starts during childhood and is divided into non-allergenic and allergenic asthma. Allergic asthma is associated with eczema and it tends to be difficult to treat because it stems from a congenital Lung and Kidney Qi Deficiency and is often hereditary. Early onset asthma can be affected by diet (e.g., lactose intolerance), emotional problems, fatigue, chronic illness and invasion by external pathogenic factors. It is related to an allergic hypersensitivity of the immune system.
2. Late onset asthma (also called intrinsic asthma) normally starts later in life and is due to bronchial hyperreactivity. In late onset asthma there is no hereditary basis. It is generally characterized by a Lung, Spleen, or Kidney Deficiency and a retention of Phlegm. In some cases, asthma can be caused by Liver Qi stagnation or Liver Fire obstructing the descending action of Lung Qi. Late onset asthma can be affected by overwork, excessive sexual activity, emotional stress and the excess consumption of sour, greasy or cold foods.

SYNDROMES

1. For Wind Cold asthma due to External Evils, the main symptoms are dyspnea, coughing, abundant, thin expectoration and clear Phlegm accompanied with external Wind Cold syndrome.
2. For asthma due to the accumulation of Phlegm Dampness in the Lungs, the main symptoms are dyspnea, cough, abundant, thin and sticky expectoration, stuffiness in the chest and in the epigastric region.
3. For asthma due to Deficient Lung and Spleen Qi, the main symptoms include dyspnea, spontaneous perspiration, aversion to Wind, an expectoration of clear, thin Phlegm, weak voice, loss of appetite and general lassitude.
4. For asthma due to Deficient Lung and Kidney Yin, the main symptoms are dyspnea, dry throat and mouth, dry cough, inconsistent fever, night sweating, and with some patients there may be hemoptysis (the expectoration of Blood).

TREATMENT

The therapeutic effect of Medical Qigong is better for patients who have had this disease for a short time period. In chronic asthma, improvement usually takes more time.

1. Begin by stimulating the patient's CV-22, CV-17 and CV-4 points along with the Bl-13 points.
2. Using the Vibrating Palm technique and emit Qi with the left palm into the patient's lower neck and upper back area, while simultaneously extending energy into the patient's Lower Dantian with the right hand for 14 to 28 breaths.
3. Purge the Toxic Qi from the patient's CV-22 and Bl-13 areas, then lead and guide the patient's Lung Qi down the Governing Vessel and Bladder Channels into the Mingmen GV-4 and Bl-23 points. Repeat for 3 to 7 respirations.
4. Using the Sword Fingers hand posture, vibrate energy into the patient's CV-22, -17, Lu-1 and Lu-2 points. Then guide the patient's Qi to flow downward from the chest to the Lower Dantian along the Stomach Channels and Conception Vessel to stabilize the patient's Qi.

TREATMENT MODIFICATIONS

1. To treat a case of Wind Cold syndrome due to exogenous Evils, purge and sweep the patient's Gall Bladder Channels, concentrating on the head area. Then extend Qi into the patient's Lu-1 and Lu-2 points. Use Pulling and Leading manipulations to draw the Toxic Qi along the patient's Lung Channels, expelling the pathogenic Qi out of the fingertips on both hands.
2. In treating cases of Deficient Lung and Spleen Qi, use the Extended Fan Palm hand posture and Vibrating Palm technique to emit Qi towards the patient's CV-12, CV-6, Bl-20 and Bl-13 points for 9 to 18 breaths.

3. In treating cases of Deficient Lung and Kidney Yin, use Extended Fan Palm hand posture and Vibrating Palm hand technique to emit Qi towards the patient's Lower Dantian, Mingmen, and Bl-13 points for 6 to 12 breaths.

PRESCRIPTIONS AND HOMEWORK

Patients are encouraged to practice physical exercises to strengthen the body and enhance their resistance to pathogenic invasion. These exercise prescriptions however, must not be too strenuous. Two prescriptions used for relieving asthma and respiratory pain are described as follows.

Exercise to Relieve Asthmatic Pain. For chronic conditions of asthma, patients are given the following prescription:

1. From a sitting or standing posture, slightly close the eyes and relax. Stimulate the energy of the chest by kneading the sternum from the sternal notch to the xiphoid process 36 times.
2. Inhale and place the right palm on the left side of the chest. Exhale, and begin dredging the left Lung Channel using a sweeping motion to purge the Toxic Qi down the arm and out through the hand. Perform this Purging exercise 10 times, then repeat 10 times on the opposite side.
3. Finally, perform the Kidney Rub exercise (see Chapter 39) for 18 breaths.

Exercise to Relieve Asthma Attacks

For acute asthmatic attacks, patients are given the following prescription.

From a sitting or standing posture, slightly close the eyes and relax the muscles of the chest and back. Take a deep breath and say "relax" silently, while simultaneously imagining the Lungs relaxing and descending the body's Qi down to the toes along both lateral sides of the body. As the Qi descends from the Lungs, imagine this energy flowing like water, down your body into the Earth. Repeat this procedure until the asthma attack has receded.

PRESCRIPTION MODIFICATIONS

- Patients with Deficient Lung Qi should practice the previous exercises in combination with Taking in the White Qi.
- Patients with a domination of pathogenic factors should practice the previous exercises in combination with rubbing their chest and uttering the sound "Sss."

BRONCHIECTASIS

Bronchiectasis is the chronic dilation of the bronchi, with a secondary infection that usually involves the lower portion of the Lungs. Dilation may involve an isolated segment or can spread throughout the entire bronchi.

ETIOLOGY

This disease is characterized by hypertrophy of the mucus-secreting glands in the bronchial tree. The dilation and deformation of the bronchi are caused by the damage of the vessel walls due to chronic inflammation of the bronchi and the peribronchial Lung tissue. In advanced cases the bronchi themselves can become inflamed with infection. The clinical manifestation is a chronic cough with an excessive expectoration of sputum. If followed by infection, there may be night sweating, fever and lassitude.

SYNDROMES

1. Liver Fire attacking the Lungs can manifest with pain and distention in the chest and hypochrondium, flushed face, cough, expectoration of yellow, thick sputum and repeated hemoptysis (Blood discharging from the larynx, trachea, bronchi, or Lungs).
2. Dry Heat scorching the Lungs manifests as a dry throat and dry cough with little sputum, which is bright red and bloody.
3. Qi and Yin Deficiency manifests through shortness of breath, lassitude, emaciation, repeated hemoptysis, cough with little sputum, erratic fever and flushed face (especially around the cheeks).

TREATMENT

1. Have the patient sit at the edge of the table, and begin to stimulate the upper chest at the Lu-1 and Lu-2 points and the upper back around the patient's Bl-13 points.
2. Using the Extended Fan Palm hand technique, emit Qi into the patient's Middle Dantian directing the energy to flow into the Lu-1 and Bl-13 points for 6 to 12 breaths.

3. Using the Leading and Guiding hand techniques, purge and dredge the Toxic Qi out the Lung Channels (i.e., down the arms and out of the hands).

TREATMENT MODIFICATIONS

1. When treating patients with Liver Fire attacking the Lungs, include the Vibrating Palm technique, and emit Qi into the patient's Bl-13 and Bl-18 points for 12 breaths. Then purge the Liver Channels down the torso and out of the legs.
2. When treating patients with Dry Heat scorching the Lungs, apply the Vibrating Palm technique to emit Qi into the patient's Bl-23 and Bl-13 points for 24 breaths.
3. When treating patients with Qi and Yin Deficiency, apply the Vibrating Palm technique to emit Qi into the Bl-13, Bl-20, and Bl-23 points for 12 breaths.

PRESCRIPTIONS AND HOMEWORK

It is advisable for patients to practice regulating Lung exercises with Natural Breathing to strengthen the Lungs.

PRESCRIPTION MODIFICATIONS

1. For Liver Fire attacking the Lungs, have the patient perform the Sun and Moon Rotating technique (see Chapter 41), while sounding the word "Sss" for 36 breaths, to purge the Lungs in addition to performing the Lung tonification and regulation exercises (see Chapter 42). Next, have the patient rub the hypochrondrium area and sound the word "Shu" for 36 breaths.
2. For Dry Heat scorching the Lungs, (from a Wuji posture) have the patient bring both hands up to the sternum and begin kneading the sternum down to the xiphoid process 36 times. Next, the patient places both palms on the chest with the fingers touching the sternum to allow the Qi to flow into the Lungs for 10 breaths. Finally, the patient rubs the hypochrondrium and purges the Toxic Qi from the body while exhaling and rubbing the sides of the chest, starting at the armpits, to direct the Qi to flow downward into the Lower Dantian.
3. For Qi and Yin Deficiency, it is advisable for the patient to practice Lung tonifying exercises and the Gathering the Moon Cream exercises (see Chapter 11), as well as the methods of Taking in the White Qi.

SINUSITIS

Sinusitis is described as the inflammation of the sinuses. Infections stemming from the common cold or influenza frequently cause secondary infections in the sinuses, these infections tend to become chronic. The body's frontal and maxillary sinuses are very prone to infection and inflammation.

ETIOLOGY

An invasion of pathogenic Wind Cold into the Lungs through the nose (the orifice of the Lungs) can bring about the accumulation of Heat and cause an obstruction of the Lungs' Qi. This condition further impairs the Lungs' dispersing and descending function, giving rise to sinusitis.

The excessive consumption of greasy, hot foods leads to a buildup of Phlegm and Heat toxins within the body and may predispose the patient to sinusitis. Greasy, hot foods may also lead to the formation of Damp Heat within the Stomach and Spleen, which can be transported upwards into the sinuses by way of the Stomach Channels.

Excess conditions, such as Wind Heat, Lung Heat, Stomach and Spleen Damp Heat, and Liver and Gall Bladder Fire, should all be purged from the patient's body, whereas chronic conditions caused by Spleen Deficiency should be tonified.

SYNDROMES

Sinusitis is characterized by consistent yellowish and foul smelling nasal discharge accompanied by cough, impairment of smell, and dull pain in the upper part of the cranium. The headache caused from sinusitis is located on the face and corresponds with either the frontal ethmoidal or the maxillary sinuses.

TREATMENT

1. Begin by stimulating the patient's frontal sinuses, focusing specific attention on the LI-20, GB-20, and Yintang points.

2. Then emit Qi into both of the patient's hands and wrist areas, focusing specific attention on the LI-4 and Lu-7 points.
3. Next, using either the Sword Fingers or Extended Fan Palm hand posture, emit Qi into the patient's LI-20 points. Begin using the Pushing, Pulling, and Leading manipulations to guide the patient's Toxic Qi down the Stomach Channels, purging it out through the St-36 point on each leg.

PRESCRIPTIONS AND HOMEWORK

The following prescription is used for relieving sinusitis.

The patient, from a seated posture, rubs his or her palms until the dorsal sides of each thumb are very hot. Then the patient places the thumbs on the sides of the nose and begins to gently rub up and down for 10 breaths. This exercise is called Bathing the Nose.

Next, the patient places the tips of the middle fingers on each side of the nose (LI-20), and begins kneading these points. This is followed by using circular rotations from the cheeks to the nose five times during inhalation, and then from the nose to the cheeks five times during exhalation. This part of the exercise is called Kneading the Apex of the Nose. This exercise is performed for 10 breaths.

CHAPTER 47
QI EMISSION THERAPY FOR KIDNEY DISEASES

INTRODUCTION

The Kidneys are solid (Zhong) organs that store the body's Jing. They correspond to the element Water and are sometimes referred to as the Black Emperor. The Kidneys house the Zhi, which is the will power, and they are in charge of life and reproduction. The Kidneys are also responsible for bone, brain and Marrow formation. They govern the Mingmen Fire and are the root of the Yin and Yang energies within all the organs (Figure 47.1). The Kidneys are energetically paired with the Bladder (a hollow organ).

The Kidneys are responsible for inhalation. Upon inhalation the Kidneys grasp the Lungs' Qi and hold it. This "holding" action secures the body's ability to maintain deep, slow inhalation. Therefore, when the Kidneys become weak (due to shock, trauma or disease), one of the clinical manifestations of Deficient Kidneys includes dyspnea (shortness of breath and gasping for air).

The Mingmen, or Gate of Life, is the root of the Yuan Qi, the residence of Water and Fire energy. The Yin Qi of the Five Yin Organs cannot be nourished without it, and the Yang Qi of these organs cannot be lifted without it. Kidney Qi flows to the brain. The ears (Yang orifice), anus and urethra (two Yin orifices) are the orifices of the Kidneys.

The Kidneys' upper external connection with the outside world is through the ears, hence any sound or tone resonation affects the core essence of the individual's spiritual and emotional will power.

In the winter months the Kidneys become more active in individuals who already possess strong Kidney Qi, but becomes Deficient in those with weak Kidney Qi. The excessive eating and drinking of salty foods and overexposure to cold depletes the Kidneys.

Kidney Yin energy flows to the Liver, Heart, and Lungs. It is responsible for the body's Jing

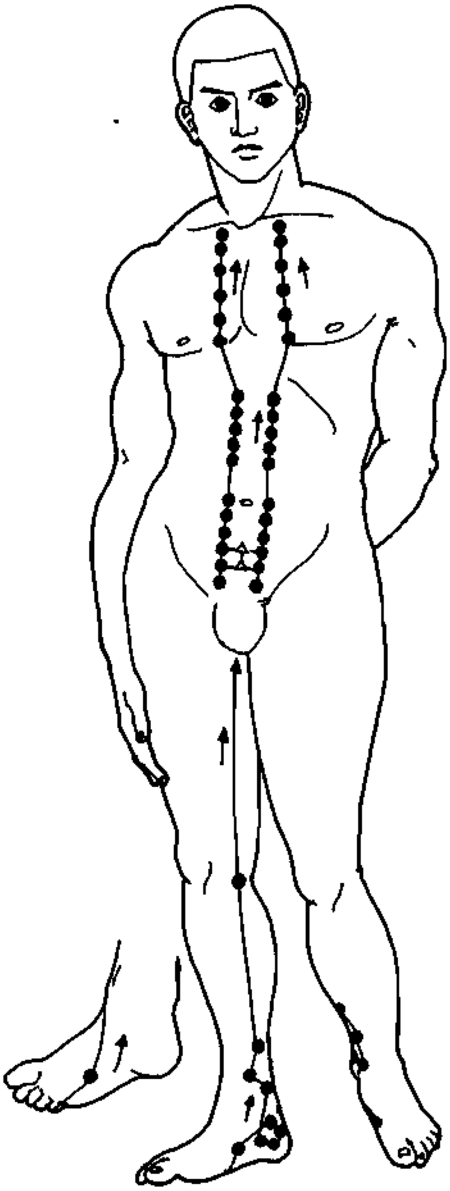

Figure 47.1. The Kidney (Kd) Channels

and produces the birth, growth, maturation, and reproduction cycle. Kidney Yang energy flows to the Spleen, Liver, Heart, and Lungs. It supports the Yang of all the body's organs via the Mingmen (Figure 47.2).

KIDNEY COMPLICATIONS AND SYMPTOMS

This section focuses on specific Kidney diseases and how to rectify them. The following discussion of treatment principles begins with an explanation of Kidney syndromes and symptoms. Several Kidney diseases have symptoms which are common to the organ's energetic dysfunction, such as mental and emotional problems, lower back pain, tiredness, and breathlessness. A list of Kidney syndromes and their symptoms are described as follows.

1. A Kidney Yang Deficiency syndrome manifests through such symptoms as headaches, breathlessness, wheezing, asthma, tiredness, mental and emotional problems, lower back pain, diarrhea, constipation, painful urination, enuresis, edema, menorrhagia and premenstrual tension.
2. A Kidney Yin Deficiency syndrome manifests through such symptoms as tiredness, chest tightness, asthma, breathlessness, mental and emotional problems, constipation, enuresis, lower back pain, bleeding, Parkinson's disease, menorrhagia and premenstrual tension.
3. A Kidney Jing Deficiency syndrome manifests through such symptoms as tinnitus, dizziness, and mental and emotional problems.

GENERAL TREATMENT FOR KIDNEY DISEASES

The following is a description of a general protocol used for treating a patient's Kidneys. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body through which Qi is emitted (or Turbid Qi removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

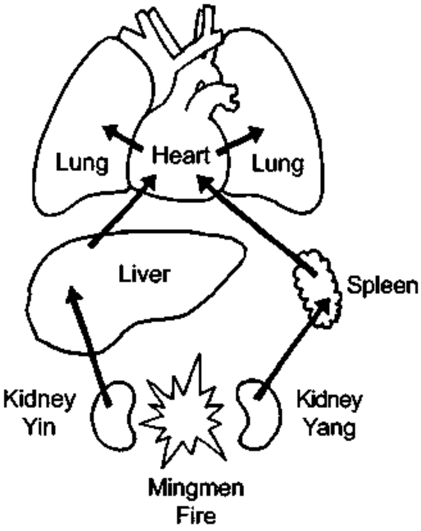

Figure 47.2. Kidney Yin flows from the Kidneys to the Liver, then to the Heart and Lungs. Kidney Yang flows from the Kidneys to the Spleen, Liver, Heart and the Lungs.

After comparing the various Kidney therapies, the reader will notice that certain patterns repeat themselves.

1. Generally, the Qigong doctor accesses the patient's Kidneys through either direct energetic insertion into the Kidneys' anatomical location on the back, through the naval and Mingmen access gates, or through the patient's channels and external energetic fields (Figure 47.3).
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor will stimulate the patient's Kidneys, Mingmen, and Lower Burner areas, and roots the patient's Qi into the Lower Dantian.
3. Next, the Qigong doctor stimulates the patient's Kidneys and Lower Dantian through the Taiji Pole and Bubbling Spring (Kd-1) points, regulating the body's Yin and Yang Qi via the Microcosmic Orbit (Fire Cycle).

Keep in mind that these areas are general and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

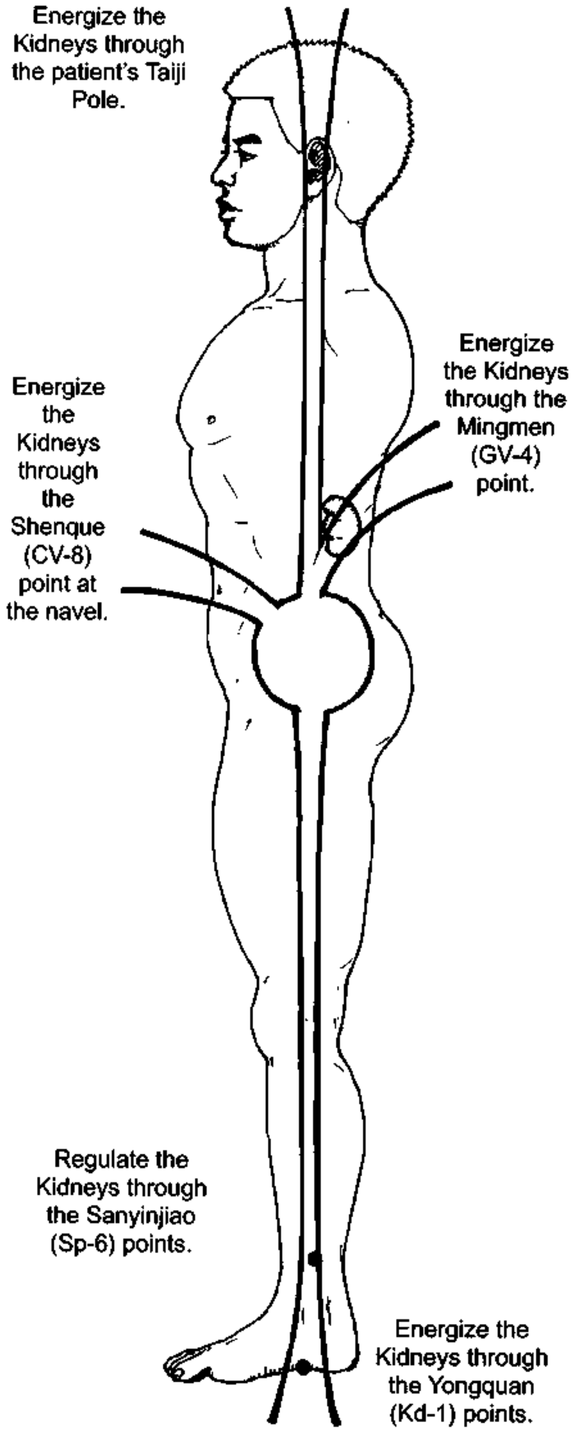
Energize the Kidneys through the patient's Taiji Pole.
Energize the Kidneys through the Mingmen (GV-4) point.
Energize the Kidneys through the Shenque (CV-8) point at the navel.
Regulate the Kidneys through the Sanyinjiao (Sp-6) points.
Energize the Kidneys through the Yongquan (Kd-1) points.

Figure 47.3. Treating Kidney Diseases

MEDICAL QIGONG THERAPY FOR KIDNEY QI DYSFUNCTION

Some symptoms of Kidney disease can manifest as dark circles under the eyes, decaying teeth, distended abdomen, dyspnea, and perspiration. The primary therapy used to treat Kidney diseases include: the Windy Breathing method to strengthen the Kidneys, Toe-Raised Walking, and Kidney massage (see Chapter 15).

1. An Excess Heat syndrome of the Kidneys should be treated by purging through uttering the "Chui" (Chree) sound, and reinforcing its effectiveness by focusing intent on the exhalation (Figure 47.4). This gradually relieves sluggishness and stagnation of the Kidney's Qi.
2. A Kidney Deficiency syndrome manifests mid-lower back pain at the waist level, and should be treated by Taking in the Blue Qi (see Chapter 4) into the Kidneys, Mingmen, and Lower Dantian.
3. A Deficiency of Mingmen Fire should be treated by emitting Qi into the patient's Lower Dantian and Mingmen area with the Pushing technique along with clockwise Qi rotation. This manipulation invigorates the Kidneys' Qi and replenishes the Fire of the Mingmen.

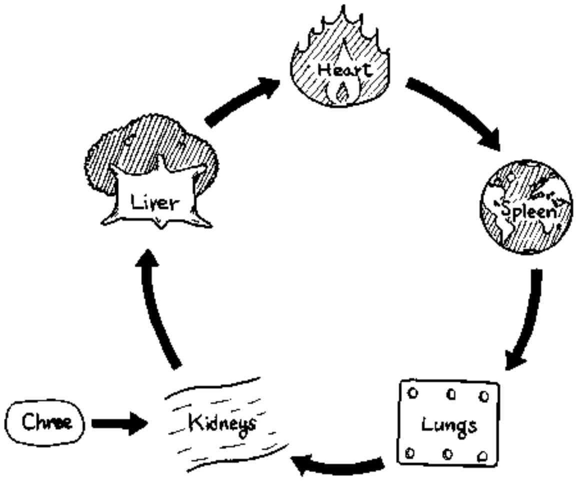
Heart
Spleen
Lungs
Kidneys
Chree
Liver

Figure 47.4. The Kidney sound is "Chui" (Chree).

KIDNEY QIGONG PRESCRIPTIONS

These exercises promote the flow of Qi and Blood of the Kidney Channels, nourish the Kidneys, strengthen Yang, and increase the Yuan Qi. These exercises can also be used to prevent and treat the following diseases and syndromes: nephritis, psychoneurosis, lumbago, rachialgia (spinal inflammation), tinnitus, deafness, frequent micturition and aversion to cold due to having a Cold condition.

PRESCRIPTIONS FOR TONIFYING THE KIDNEYS

The Taking in the Blue Qi can be used to tonify the Kidneys. The properties inherent within the color blue have a persistent vibratory rate used for tonifying the Kidneys and Bladder.

1. Assume a standing, sitting, or lying posture. Relax the whole body, place the tongue against the hard palate behind the teeth, and get rid of any stray thoughts.
2. When inhaling, visualize blue energy in front of you. Inhale the blue light energy in through the nose and down into the Kidney organs. As you exhale through the mouth, the dark, pathogenic Qi leaves the Kidneys; however, the bright clean color remains in the organ, stimulating and vitalizing it. With each breath the organ retains more clean energy and begins to glow. Repeat for five breaths.
3. Next breathe the blue color in through the nose, filling the entire mouth. When exhaling through the nose, send the blue Qi slowly down to the Kidneys, then into the Lower Dantian. Perform six to twelve times. Return to the beginning posture and close.

PRESCRIPTIONS FOR REGULATING THE KIDNEYS

Remove stagnation from the Kidneys by having the patient rub the abdomen while sounding "Chree."

Assume a sitting or standing posture. Place both palms flatly on the lower abdomen (left hand on the outside for men, opposite for women), inhale slowly, and exhale the sound "Chree" while rubbing the lower abdomen clockwise, gently with the palms. Perform ten to twenty breaths.

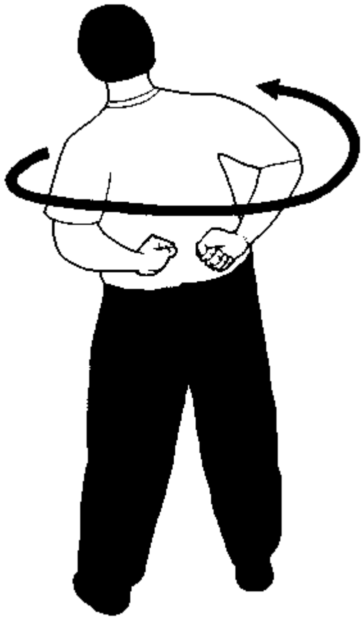

Figure 47.5. Rotate the waist in a clockwise and counterclockwise direction.

PRESCRIPTIONS FOR STRENGTHENING THE KIDNEYS AND CONDUCTING QI

The following exercise is used to tonify the Kidneys.

1. Assume a standing posture. Clench both hands into fists, and place them against the soft tissue on both sides of the Mingmen. Rotate the waist, turning towards the left in a clockwise direction for six rotations. After the six rotations pause in the center, then rotate in the opposite direction for six times (Figure 47.5).
2. Assume a standing or sitting posture. Rub the sides of the Mingmen with both hands, up and down thirty-six times, focusing the mind on the lumbar region (Figure 47.6).
3. Tuck the sacrum under (to stimulate the renal capsule), placing both palms on the groin, under the junction of the pubic bone (left hand on the outside for men, right for women).

   Support and gently pull up the area around the perineum with both hands eighty-one times (Figure 47.7). Switch hands and repeat. Note: The Kidneys are injured by irregular

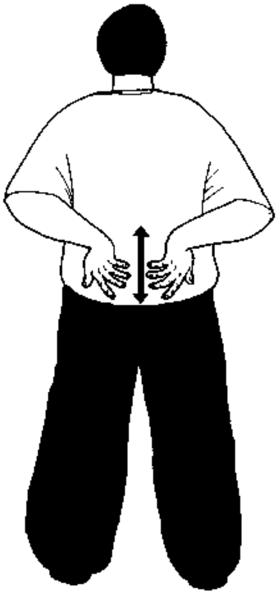

Figure 47.6. Rub the Kidney and Mingmen area.

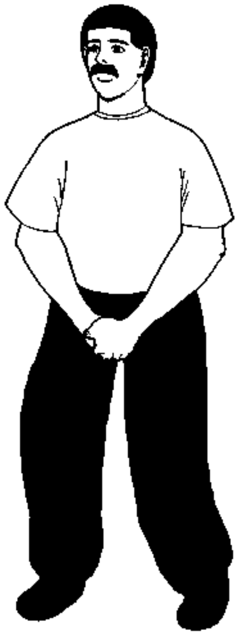

Figure 47.7. Support the groin and perineal area.

sleep patterns, fear, excessive caffeine, sex, drugs, alcohol, and smoking. Excess saltiness and Cold (climate and food) have a draining effect on the Kidneys.

SPECIFIC MEDICAL QIGONG THERAPY FOR KIDNEY DISEASES

The following are several Medical Qigong treatments used in China to treat specific Kidney diseases.

TINNITUS AND DEAFNESS

Tinnitus is a condition of a high-pitch or rushing-wind sound within one or both ears. The sound may be a consistent pitch or vary in intensity and modulation.

ETIOLOGY

Although the ears are the opening of the Kidneys, there are other influences which must be addressed when dealing with tinnitus. Tinnitus is most commonly caused by stagnation of the Triple Burner and Gall Bladder Channels. These channel stagnations are due to three adverse conditions: the rising of Liver Wind Fire which may be induced by violent rage, the obstruction of the orifices in the head due to External Wind invasion, or by failure of the Kidney Jing to reach the ears due to a Kidney Deficiency.

This disease can be divided into Full and Empty Types of tinnitus:

1. The Full Type of tinnitus is caused by the flaring up of pathogenic Wind, Yang, Fire, Phlegm, or Phlegm Fire. It is most noted in the Excess rising of Liver and Gall Bladder Fire, or Phlegm Fire flaring upwards. This type of tinnitus is experienced as a sudden onset with a loud sound and is aggravated by cupping one's hands over one's ears.
2. The Empty Type of tinnitus is caused by a Deficiency of Kidney Qi, Kidney Jing, Lung Qi, or Heart Blood that is unable to extend to the ears. It is most noted in Kidney Jing Defi-

ciency, Deficient Upper Burner Qi and Heart Blood Deficiency. This type of tinnitus manifests clinically through a gradual onset with a low sound and is improved by cupping one's hands over one's ears.

SYMPTOMS

1. The Excess Type of tinnitus manifests as a sudden deafness or distending pain in the ears accompanied with a loud-pitched ringing noise. The adverse rising of Liver Wind Fire or Gall Bladder Wind Fire causes vexation and irritability, and often manifests as a flushed face and dry mouth.

   The Excess condition caused by an invasion of External Wind also includes symptoms such as chills, fever, and headache.
2. The Deficient Type of tinnitus can either occur constantly or intermittently, and may be aggravated by overstraining and relieved by hand pressure. This condition is often accompanied by dizziness, soreness of the waist, seminal emission, or leukorrhea.

TREATMENT

1. Begin by stimulating the patient's TB-3, TB-17, SI-19, and LI-4 points on both sides of the body.
2. Emit Qi towards the patient's Lower Dantian, using the right Sword Finger hand posture and left Extended Fan Palm hand posture.
3. Next, extend energy into the patient's affected ear and focus the Qi emission around the patient's TB-17 and TB-21 point areas. Next, guide the Qi to flow down the head, neck, shoulders, and arms along the Triple Burner Channels and out of the hand 3 to 7 times.
4. Finally, extend Qi into the patient's affected ear and guide the Qi down and out of the body along the corresponding Gall Bladder Channel ending at the corresponding foot.

MODIFICATIONS

- In cases of adverse rising of Liver Wind Fire, use the Vibrating Palm to emit Qi into the patient's Bl-18 and -19 points for 14 breaths on both sides of the body.
- In cases of Kidney Deficiency, emit Qi into the patient's Bl-23 points for 14 breaths.

PRESCRIPTIONS AND HOMEWORK

In the following three prescriptions make sure the patient heats his or her hands before beginning.

1. Have the patient perform the Beating the Heavenly Drum exercise (see Chapter 16) for 24 breaths.
2. Have the patient press the ears with the palms in a compression-release fashion 10 times each. Be sure to warn the patient to avoid forceful or violent compression. The patient's intention should focus deep within the orifice of the ears.
3. Have the patient pinch the top of the auricles of the ears gently, and massage them from the top downwards 24 times.

LUMBAGO

Lumbago is a general non-specific term for dull aching pain in the lumbar region of the back.

ETIOLOGY

Lumbago is a common symptom observed in many diseases, such as acute and chronic muscular strain of the lower back, subluxation of the discs, prolapse of the lumbar vertebra, etc. Patients may suffer from lumbago with pain radiating down to their lower limbs. Lumbago pain can also be aggravated by weather changes. Traditionally, there are three general causes of lumbago: lumbago due to Wind Cold and Dampness, lumbago due to Kidney Deficiency, and lumbago due to trauma.

1. Lumbago due to Wind Cold and Dampness, is caused by retention of Cold and Damp pathogens in the channels and collaterals. This in turn leads to the obstruction or stagnation of Qi and impairs Blood circulation.
2. Lumbago due to Kidney Deficiency is caused by an over consumption of the Kidney Jing due to an overactive sex life. This results in the failure of the Kidneys to nourish the Triple Burner Channels.
3. Lumbago due to trauma occurs from an injury to the tissues and muscles (e.g., dislocated vertebra, fractures, ruptured spinal disks, etc.), and leads to stagnation of Qi and Blood. Qi and Blood stagnation leads to obstructions in the channels and collaterals.

SYMPTOMS

1. Lumbago, due to Wind Cold and Dampness, is manifested by heaviness and pain at the waist, which may radiate to the hips and down the legs, or may be accompanied by muscular stiffness. In chronic cases, the pain is intermittent with various degrees of severity. On wet and cold days the patient often feels cold at the waist and lumbar area.
2. Lumbago, due to Kidney Deficiency, is manifested by continuous, dull pain, accompanied by soreness and weakness of the waist and knees, cold limbs and nocturnal emission.
3. Lumbago due to trauma is manifested by localized pain along with tenderness and limited mobility.

TREATMENT

1. Place the patient in a sitting position on the table. Begin stimulating the patient's Mingmen area with your left hand while the right hand guides the energy down to the patient's Bl-40 points behind the knees.
2. Using the Extended Fan Palm hand manipulation emit Qi into the patient's Mingmen and Bl-18 areas, guiding the Qi to flow along the Bladder Channels down the legs.
3. Next, place the patient in a supine posture, knees pointing upwards and feet flat on the table. Position yourself over the patient allowing the weight of your chest to support and compress the physical structure of the patient's knees, lower legs and feet. Embrace the patient's hips and imagine separating the inner fascia and tissues. The purpose for this compression is to initiate a strong internal stretching action within the tendons and smaller muscles of the patient's lower back and sacral area. This technique is used to disperse stagnant Qi and Blood within the patient's lumbar vertebra, sacrum and femur articulations. At the same time, direct the Qi upwards along the patient's spine and Governing Vessel. Hold this compression for several minutes.
4. To avoid respiratory congestion (while your weight is stretching the patient's sacrum and lower lumbar), the patient is instructed to raise the arms above the head.
5. As the patient experiences the lower back compression, he or she imagines the stagnant energy (trapped within the lumbar area) melting like ice into water and flowing through the Microcosmic Orbit (Fire Cycle). With each exhalation the patient relaxes to a deeper level, imagining him or herself sinking deeper into the table.

TREATMENT MODIFICATIONS

1. For lumbago due to Wind Cold and Dampness, purge the pathogenic Qi out of the patient's body first, and then emit Heat into the channel points associated with the specific location of the patient's pain.
2. For lumbago due to Kidney Deficiency, use a Vibrating Palm technique to emit Qi towards the patient's Mingmen, Bl-23 and Bl-18 points for 14 breaths each point on both sides of the body.
3. For lumbago due to trauma, extend Qi for 28 breaths into any Ah Shi point (painful tissue area).

INSOMNIA

Insomnia is described as a sleeping disorder that disrupts the patient's normal sleeping pattern. In mild cases, the patient may have difficulty falling asleep or remaining asleep (waking numerous times during the night). In severe cases the patient may be unable to fall asleep at all. Diseases with insomnia as part of their manifestation include anxiety disorders and menopausal syndrome.

ETIOLOGY

The amount and quality of sleep depends on the state of the patient's Shen, which is rooted in the Heart (specifically, the Heart Blood and Heart Yin). If the Heart is agitated or Deficient, the Shen cannot root, and sleep will be affected. Any disharmony of the body's internal organs due to an Excess or Deficient condition affects the patient's Jing and Blood. Since the Jing and Blood root the Shen, the Shen will have no residence and insomnia may result.

Insomnia can be caused by insufficient Blood due to impairment of the Spleen Qi, or from malnourishment of the Heart Qi due to a Deficiency of Heart Blood. This condition can also result from overfatigue, anxiety or irritability. This may be due to a disharmony between the Heart and Kidneys (resulting from Deficient Kidneys), irritability (due to mental depression) or from disharmony between the energetic function of the Stomach and Spleen (resulting from an improper diet).

Since the Hun (rooted in the Liver Blood or Liver Yin) play an important role in the length and quality of sleep, Liver Blood Deficiency can cause the Hun to be deprived of their residence and to wander at night causing restless sleep.

SYMPTOMS

The main symptom is an inability to sleep, which based on the different conditions of this illness can be further classified as:

- An inability to get to sleep,
- Sleepwalking, and an
- inability to fall asleep after waking.

The symptoms vary with the causes of the disease.

1. Insomnia due to a Deficiency of the Heart and Spleen (resulting from severe worry) may cause shallow sleep, or dream disturbed sleep. It is manifested by palpitations, poor memory, vertigo, dizziness and spontaneous sweating.
2. Insomnia due to Kidney Deficiency (resulting from excessive sexual activity) is manifested by dizziness, tinnitus, soreness in the loins and back area, as well as involuntary seminal emission.
3. Insomnia due to upward stirring of Liver Yang is manifested by pain and distention of the chest and epigastrium, belching, and acid regurgitation, as well as mental hyperirritability, dizziness and headaches.

TREATMENT

1. Begin by having the patient sit at the edge of the table. Your right hand emits Qi into the patient's CV-12 point, directing the energy to flow down into the patient's Lower Dantian. The left hand extends energy into the patient's mid-back around the area of the patient's Bl-18, Bl-19, Bl-20 and Bl-21 points, and leads the Qi down the Bladder Channels to the feet on both sides.
2. Next, use the Vibrating Palm technique to extend energy into the patient's Baihui, GV-14, Mingmen, CV-12 and Lower Dantian areas for 9 to 18 breaths each.
3. Emit Qi into the patient's Baihui area and leads the Qi from the top of the patient's head down the Conception Vessel into the Lower Dantian.
4. Finally, emit Qi into the patient's GB-8 point by the top of the ears and guide the Qi down the patient's Stomach Channels to the St-36 points.

TREATMENT MODIFICATIONS

1. For treatment of insomnia due to Deficiency of Heart and Spleen, emit Qi into the patient's Bl-20 and Bl-15 points, guiding the Qi down along the Bladder Channels to the feet.
2. For insomnia due to Kidney Deficiency emit Qi into the patient's Mingmen and Bl-23 points for 18 respirations.
3. For insomnia due to emotional depression, emit Qi towards the patient's Lv-3 and Lv-4 points (on both sides of the body), guiding the Qi down to the patient's lower extremities along the Liver and Gall Bladder Channels.
4. For insomnia due to Spleen and Stomach disharmony, use the Vibrating Palm hand technique to emit Qi into the patient's CV-12, St-36, and St-21 points on both sides of the body.

PRESCRIPTIONS AND HOMEWORK

For the best results, have the patient practice Medical Qigong prescriptions one to two hours before going to bed. For chronically ill patients, the causes of the insomnia should be removed first, followed by Medical Qigong prescriptions and regular physical exercise for lasting therapeutic results.

Have the patient perform the Kidney Massage for treating insomnia prescribed in Chapter 39.

PYELONEPHRITIS

Pyelonephritis is an infectious disease caused by bacteria invading the renal parenchyma and renal pelvis unilaterally or bilaterally. It clinically manifests through the following symptoms: fever, lumbago and abnormal discharge of urine.

SYNDROMES

1. Damp Heat in the Bladder is clinically manifested through symptoms such as frequent and painful urination, fever, distention and pain in the lower abdomen, lumbago and aversion to cold.
2. Stagnation of Liver and Gall Bladder Heat is clinically manifested through symptoms such as fever with or without chills, nausea, loss of appetite, lumbago or pain in the lower abdomen, frequent urination with cloudy urine and feeling agitated.
3. Spleen and Kidney Deficiency with Heat stasis is clinically manifested through such symptoms as edema of the face, swelling of the feet, abdominal distention, thin stools and frequent urination, lassitude and weakness in the entire body (especially the legs), dizziness, tinnitus and anorexia (lack of appetite).

TREATMENT

1. Have the patient sit at the edge of the table and begin stimulating the Mingmen and Lower Dantian areas using a Vibrating Palm technique for 11 to 22 breaths.
2. Next, have the patient lay supine on the table and stimulate the patient's Sp-6 and Kd-3 areas on both sides of the body.
3. Use the Extended Fan Palm hand technique to emit Qi into the patient's Lower Dantian, guiding it down the Kidney Channels to the bottom of the feet.

TREATMENT MODIFICATIONS

1. When treating patients with Damp Heat in the Bladder, use a Sword Fingers hand posture and apply the Vibrating technique, emitting Qi into the patient's Mingmen and Lower Dantian for 14 to 28 breaths.
2. When treating patients with stagnant Heat in the Liver and Gall Bladder, use the Extended Fan Palm hand posture to emit Qi into the patient's mid-back and Liver, purging Toxic Qi down the Liver and Gall Bladder Channels and out of the body via the legs.
3. When treating patients with Deficient Spleen and Kidneys, use the Extended Fan Palm hand posture to emit Qi into the patient's mid-back Bl-20, and Bl-23 points, CV-12 and Lower Dantian areas.

PRESCRIPTIONS AND HOMEWORK

Have the patients practice the Standing Post Posture with arms suspended at the Middle Dantian, and the mind's intention focused on the Lower Dantian. Use natural abdominal breathing (see Chapter 16).

PRESCRIPTION MODIFICATIONS

1. For patients with Damp Heat in the Bladder, have them practice rubbing the costal area and sounding "Who."
2. For patients with stagnant Liver and Gall Bladder Heat, have them practice massaging the costal area and sounding the word "Shu."
3. For patients with Spleen and Kidney Deficiency, have them practice the method of Taking in Yellow Qi, as well as the Gathering the Moon Cream exercise (see Chapter 11).

IMPOTENCE

Impotence is a weakness and inability of the male to achieve or maintain an erection. Symptoms vary from premature ejaculation to partial penile erection (the penis is erect but not hard, resulting in the incapability of sexual intercourse), to a complete failure to achieve any erection at all. Impotence indicates a weakness in all Five Yin Organ energies. The combined energies must be sufficient to culminate an erection, as described below.

1. The Heart energy quality is strongly affected by the nature of the sexual encounter. Casual sex may not involve the highest spiritual energy of the Heart's upper chamber (see Chapter 5), only the basic instincts and drives of the lower chamber. The Heart may also be affected by the individual's moral values and beliefs, as well as the quantity and quality of emotional passion (i.e., with or without guilty feelings) at the time of arousal. Heart energy also pertains to the energy of the Blood.
2. The Spleen energy pertains to what the man

is thinking and his intention at the time of arousal. This also affects the man's muscles.

3. The Liver energy pertains to visual stimulation that affects a man's sexual drive. This affects the man's strength. Many aspects of man's sexuality are relate to the Liver's functions, i.e., visual images evoking emotions, Blood filling the erectile tissue, and the power of desire.
4. The Kidney's Qi pertains to the Jing energy stored within the Marrow, and relates to the man's will power and endurance. It is associated with audible stimulation and affects the man's endurance.
5. The Lung Qi is man's somatic expression of sexual passion and is stimulated through touch and smell. This energy also affects the man's endurance.

ETIOLOGY

There are two main factors ascribed to impotence: psychogenic and organic. In the first case, the patient's physical, mental, emotional, and spiritual energies are out of harmony. This can be caused by fear, excitement, performance anxiety, etc. In the second case, impotence can be caused by medications, drugs, alcohol, etc. In most organic cases, however, impotence is caused by prostate failure, cardiovascular disease or diabetes. If this condition persists long enough the man looses sexual interest and the condition becomes psychosomatic.

SYNDROMES

- Insufficiency of Kidney Yang is clinically manifested through symptoms such as impotence, lumbago, weakness in the knees, dizziness, tinnitus, aversion to cold, cold limbs and a cold sensation in the external genitalia, listlessness and shortness of breath.
- The downward flow of Damp Heat is clinically manifested through such symptoms as impotence, itching and pain around the external genitalia and scanty, dark urine.

TREATMENT

1. Stimulate the patient's Mingmen and Lower Dantian areas using the Vibrating Palm technique for 12 breaths.
2. Next, stimulate the patient's Sp-6 areas on each leg, leading the Qi up into the Lower Dantian for 12 breaths.
3. Finally, return back to the Mingmen and Lower Dantian areas using the Extended Fan Palm hand technique. Emit Qi for 24 breaths, rotating in a clockwise direction to gather the Qi in the Lower Dantian.

TREATMENT MODIFICATIONS

- When treating a patient with an insufficiency of Kidney Yang, vibrate the energy using the Sword Fingers Hand technique, while emitting Qi into the Lower Dantian for 16 to 18 breaths.
- When treating a patient with a condition of Damp Heat flowing downward, purge the Qi along the Kidney Channels down into the Lower Dantian. Use the Extended Fan Palm hand technique and emit Qi into the patient's Lower Dantian, Kd-13 and Kd-14 points on both sides of the body, as well as the navel area, for 24 breaths.

PRESCRIPTIONS AND HOMEWORK

It is advisable to have the patient perform the Deer exercise, prescribed for increasing the body's Jing.

1. Have the patient sit on the edge of the chair, rubbing the palms to create heat. This right hand will cup his testicles so that the palm completely covers them with slight pressure. The left hand is placed on the area of the Lower Dantian just below the navel. The left hand moves in a circular clockwise direction 81 times. Next, the hands are rubbed together again, reversing the hand positions so that the left hand cups the testicles and the right hand is on the Lower Dantian. The circle rubbing is repeated in the opposite direction 81 times. Concentrate on the heat and Qi filling up the Lower Dantian and genital area.
2. The patient should tighten and draw up the anal muscle while inhaling, feeling the Qi being drawn up the rectum and prostate area, filling up the Lower Dantian. He should hold the breath as long as possible, then exhale, release and relax. The anal contraction exer-

cise should be repeated beginning with 25 breaths, and eventually working up to 250 breaths.

PRESCRIPTION MODIFICATIONS

1. When treating patients with Kidney Yang Deficiency, it is advisable to have them practice Beating and Drumming the Qi with lower abdominal breathing (see Chapter 42).
2. When treating patients with a downward flow of Damp Heat, it is advisable to have the patient rub the Lower Dantian and sound the healing tones "Chree" for 36 breaths and "Shu" for 36 breaths.

CHAPTER 48

QI EMISSION THERAPY FOR MISCELLANEOUS DISEASES

INTRODUCTION

This chapter deals with miscellaneous diseases that are classified as either complicated multiple diseases, or energetic conditions which have not yet been classified. This section describes the conditions, symptoms and treatment for these syndromes. The discussion of these various treatments begins with energetic dysfunctions stemming from the Qi of the Three Dantians, and expands further to include other tissues, internal organ and energetic organ system complications.

DEFICIENT DANTIAN QI

One of the most common problems found in the body's energetic system is Deficient Dantian Qi. The symptoms vary according to the location of the Deficiency, as well as the patient's constitution.

1. If the energy is Deficient within the Lower Dantian, the symptoms and syndromes cover a large range that includes: chronic diarrhea, asthma due to Kidney Deficiency, hypertension due to Yin Deficiency, Blood Deficiency, insomnia, spermatorrhea, collapsing syndrome due to Deficiency, abandoned-type Wind Stroke, abandoned-type coma and all types of organ prolapse.
2. If the energy is Deficient within the Middle Dantian, the problems may include shortness of breath, bronchitis, asthma, chest pain, palpitations, intercostal neuralgia, all breast disorders including acute mastitis and breast abscesses.
3. If the energy is Deficient within the Upper Dantian, the problems may include frontal headaches, dizziness, vomiting, fainting due to Blood loss, insomnia, eye soreness, febrile convulsions and insanity.

PRESCRIPTIONS FOR REGULATING THE THREE DANTIANS

The following meditations are prescriptions used to regulate the heat, vibration, and light in the Lower, Middle, and Upper Dantians.

1. The Lower Dantian meditation solidifies the Qi in the Lower Dantian that allows for stable energy transference to take place. The purpose for Heat Regulation in the Lower Dantian is to cause the internal organs to steam; this transfers energy from organ to organ to nurture and regulate the body's Qi.

   Have the patient focus on breathing into the center of the Lower Dantian. Upon inhaling, the patient imagines drawing Qi into the body from the nose, urethra and anus, and directs it into the Lower Dantian. After inhaling, the patient closes the lower orifices and circulates the Qi up through the coccyx and Mingmen area, then down into the navel and into the Lower Dantian. It is important that the patient visualize a ball of energy rotating in the lower abdomen like a spinning wheel. As this Qi begins to collect in the Lower Dantian, heat and vibration begin to resonate in the lower abdominal area.
2. The Middle Dantian meditation expands the patient's Wei Qi, which in turn enhances the body's protection against an attack of Hot or Cold pathogenic factors. The patient should draw the Toxic Qi out of the Yang Channels of the arms, through the center point of the palms. This is an excellent meditation for treating diseases caused from Excess Heat, such as insomnia, headaches, hypertension, and arthritis of the joints:

   To regulate the heat of the Heart Fire and Middle Dantian area, the patient focuses the breath and mind's intention on the center of the chest (CV-17) while placing the right thumb inside the heart of the left palm (Pc-8), while resting both hands on the lap.

3. The Upper Dantian meditation not only expands the patient's Wei Qi, but also fills the body with sound and light vibration. Sound and vibration are used in the Qigong clinic to purge Toxic Qi out of the patient's body. This is an excellent meditation for balancing and energizing all Three Dantians.

   The patient begins in a standing or sitting Wuji posture, with the spine straight, Mingmen pressed back, anal sphincter closed, tongue on upper palate, and breathing naturally. This particular exercise follows the opening and closing pattern of Qi regulation. The focus is placed on each Dantian area for a period of 18 to 24 breaths. Each Dantian area needs a specific sound and light visualization to open, purge, close, and regulate its energy. The sounds and physical actions are as follows.

   - To open, energize and regulate the Upper Dantian, use the sound of "Om." The arms separate and come together above the Baihui area (above the head) in order to open and close the Upper Dantian. The focus of the mind's intention is placed on drawing in divine healing light while inhaling, to fill the Taiji Pole. When exhaling, imagine the energy rushing into the Upper Dantian while toning the first vowel sound "O." Then imagine light shinning outside of the body while sounding the consonant sound "m." On the beginning of the vowel sound, the tongue is placed on the lower palate. As the tone and breath reach its midpoint, the tongue rises to connect with the upper palate to finish the sound "Om."
   - To open, energize and regulate the Middle Dantian use the sound of "Ha." The arms separate and come together above the chest at the CV-17 point while opening and closing the Middle Dantian. The focus of the mind's intention is placed on drawing in divine healing light while inhaling, to fill the Taiji Pole. When exhaling the tone "Ha", separate the arms and imagine that the Qi rushes into the Middle Dantian and then expands to shine outside of the body. On the beginning of the tone "Ha," the tongue should be placed on the lower palate. As the tone and breath reach its midpoint, the tongue maintains its connection to the lower palate.
   - To open, energize and regulate the Lower Dantian use the sound of "Hai-riem." The arms separate and come together above the lower abdomen at the CV-8 point while opening and closing the Lower Dantian. The focus of mind's intention is placed on drawing in divine healing light while inhaling, to fill the Taiji Pole. When exhaling, imagine the Qi rushing into the Lower Dantian with the first syllable "Hai," then shines outside of the body when the syllable "reim" is pronounced. At the beginning, the tongue is placed on the lower palate to pronounce the sound "Hai." As the tone and breath reach its midpoint, the tongue changes its position to behind the teeth to tone the sound "reim."

   All three sounds are used to activate and energize the body's Three Dantians by practicing these sounds in the following two step progression:

   1. Upon inhalation, focus the mind's intention on drawing divine healing light into the body to fill the Taiji Pole.
   2. While exhaling, sound the specific tone of each Dantian, while both hands perform the opening (the expanding movements of the arms and hands) and closing (the contracting movements of the arms and hands) technique at each Dantian. The mind visualizes the light resonating and shining out from the body at each Dantian during the toning.

IMPROPER YIN AND YANG ENERGY BALANCE

Another common problem observed in the clinic is an imbalance of Yin and Yang Qi within the patient's body. Converting and balancing the Yin and Yang energy diverts the Qi before it has a chance to overload the body's system. The Qigong doctor first removes any obstruction in the patient's Governing Vessel and then harmonizes the patient's Yin and Yang energy by extending Qi into the patient's back, on the GV-13 and -14 points.

The following meditations are used as prescriptions to develop a strong electrical charge in the body, that balances the Yin and Yang currents:

Have the patient place the thumb in the heart of his or her palm (Pc-8), fingers touching the back (dorsal) side of the hand, opposite the (Pc-8) point (Figure 45.1). To energize the Lower Dantian the patient concentrates on his or her lower abdominal area. To energize the Middle Dantian the patient should concentrate on the area at center of the chest. The only difference between the Lower Dantian exercise and the meditation used to heat the Middle Dantian is the focal point of the patient's concentration. In these particular exercises, the hand postures are the same, however, the mind is focused on either the hands near the Lower Dantian, or on the middle of the chest.

- For men to develop a stronger electropositive Yang energy field, the right hand should cover the top and center of the left hand. To develop a stronger electronegative Yin energy field, the left hand should cover the top and center of their right hand.
- For women to develop a stronger electropositive Yang energy field, the left hand should cover the top and center of the right hand. To develop a stronger electronegative Yin energy field, the right hand should cover the top and center of the left hand.

Note: In generating energy within each Dantian area it is important to understand basic Yin and Yang polarities and their relationship to male and female. The female chest is considered Yang while the female genital area is considered Yin. The female Middle Dantian (or Heart area) is the focal point of female Yang (positive) energy; the genital area is the focal point of female Yin (negative) energy. Conversely, the male chest is considered Yin while the male genital area is considered Yang. The Middle Dantian (or Heart area) is the focal point of male Yin (negative) energy; the genital area is the focal point of male Yang (positive) energy (see Chapter 5).

DISORDERS OF THE QI

Whenever energy movement is impeded, disorders in Qi functions occur. Generally, these disturbances cause either disorders of Fire, or disorders of Water (see Chapter 22).

The cause and progression to these disorders can be further broken down and classified according to their energetic dysfunction. There are several types of Qi disorders observed in the clinic and they are generally described as follows.

LEAKAGE AND LOSS OF QI

During or after training practice, the patient may feel that Qi is continuously and uncontrollably leaking out of the body from the genitalia and anus or somewhere else on the body. This loss of Qi can lead to Qi Deficiencies and manifest through the following symptoms: emaciation, weakness of the limbs, grayish and dull complexion, nervousness, mental disturbances, distractibility, failing memory, spontaneous sweating, night sweats, seminal emission, insomnia, lassitude and sluggishness.

To treat this condition apply the Extended Fan Palm technique, use the Pushing, Pulling, and Shaking manipulations to emit Qi into the Mingmen; then use Vibrating and Shaking manipulations to emit Qi into the Lower Dantian.

QI STAGNATION DISORDERS

Qi stagnation exists when the Qi becomes sluggish when moving within the body's channels, organs, or tissues. Qi stagnation can be caused by many different factors such as suppressed emotions, trauma, contraction of microorganisms, poor diet, and weak digestion. Qi stagnation, over time, results in localized pain, distention, cysts, tumors, etc. Qi stagnation disorders have two subcategories: Qi and Blood Stasis, and Obstructed Qi (see Chapter 22).

1. Qi and Blood Stasis refers to the sensation of local pain, heaviness, soreness, distention, and compression, which are caused by the stagnant Qi and Blood in certain areas of the body. These sensations do not resolve by themselves. This stasis may occur during or after

Medical Qigong training practice from Qi disorders and visceral malfunctions.

2. Obstructed Qi is also considered a subcategory of Qi stagnation, and exists when the Qi becomes so stagnant that it no longer flows. At this stage the Qi can't move the Blood. This can be caused by trauma, surgery, and acute Qi stagnation.

Reconstructive Qi therapy, for conditions of Qi stagnation use the following two methods:

1. Select corresponding channel points in the region of the Qi and Blood stasis. Knead the selected channel points, then push and rub along the obstructed channel.
2. Finish the treatment by applying the Extended Fan Palm technique; use the Pushing, Pulling and Shaking manipulations to emit and conduct Qi along the channels to open them.

REBELLIOUS QI DISORDERS

Rebellious Qi disorders manifest as Qi flowing in the wrong direction (a direction different from its normal physiological progression); this causes a disturbance within the harmonious flow of energy and affects the organ functions. The symptoms have to be analyzed according to the internal organ which is involved. There are two types of Rebellious Qi, the Excess type and the Deficient type. Rebellious Qi is categorized as follows.

DISORDERS OF EXCESS REBELLIOUS QI

Disorders of Excess Rebellious Qi occur when the vital Qi increases and forces the pathogenic Qi to move. This causes local sensations of pain, soreness, distention, heaviness, coldness or hotness which can develop during Qigong practice.

To treat for conditions of Rebellious Qi, use the following two methods:

1. From the head and torso, follow the routes of the channels and collaterals out to the extremities. Pinch and lightly pull on the extremities (fingers and toes). Then press, knead, push and rub the location of the discomfort, guiding and leading the Qi along the channels away from the area of discomfort.
2. Apply the Extended Fan Palm method; use the Pulling and Leading manipulation to pull and guide the Qi, or open up the channel points to disperse the Qi.

DISORDERS OF DEFICIENT REBELLIOUS QI

Disorders of Deficient Rebellious Qi refer to the disorderly flow of Qi and Blood during or after training practice, which can cause dizziness, fright, a heavy feeling in the chest, shortness of breath, shaking limbs, trembling of the hands and feet or fainting. Generally, energy sensitive practitioners know the location and direction of the disorderly flow of Qi in their tissues.

To treat conditions of Deficient Rebellious Qi use the following two methods:

1. Open the appropriate channel points to regulate the functional activities of the Eight Extraordinary Vessels.
2. Select appropriate regions and channel points and apply the Extended Fan Palm or Sword Fingers technique. Use Pushing, Pulling, and Shaking manipulations to activate the channel Qi of that particular channel or region. Then, use the Leading manipulation to direct the Qi into specific channels or return it to the Lower Dantian.

ARTHRITIS

Arthritis is defined as inflammation of the joints, usually accompanied by pain, swelling and frequently changes in structure. This is a chronic, multiple arthropathy and can be divided into the peripheral type (pertaining to the four limbs), and the central type (pertaining to the spinal column, i.e., rheumatoid arthritis or ankylosing spondylitis).

ETIOLOGY

Arthritis is caused by pathogenic Wind, Cold and Dampness attacking the body's channels and hindering the Qi and Blood circulation. This invasion can give rise to pain, soreness, heaviness, and numbness of the tendons and muscles, as well as swelling and difficulty in moving the joints. This condition can be caused from an unstable lifestyle weakening the body's Wei Qi, overexposure to a draft after sweating, wading in water, exposure to cold weather, or laying on a damp ground for long periods of time.

SYMPTOMS

Arthritis is clinically manifested primarily as arthralgia and dysfunction of the joints and can be accompanied by muscle spasm, loss of muscle tone, loss of muscle mass and deformity. Symptoms can be divided into early and late stages.

- In the early stages of arthralgia, the local swelling usually starts from the small joints of the fingers and toes, wrists and ankles, or from the iliosacral joint and spreads upward into the body. Several weeks or months later, the patient may feel local tenderness and inflexibility of the joints with some clicking or cracking sounds. In this stage the pain is worse at night and early in the morning, and may be alleviated with gentle exercise. In the early stages of arthritis, the patient's condition can be controlled and even restored back to normal through Medical Qigong prescriptions.
- In the later stages of arthritis, the articulation of the joint movement is markedly limited, and patients may experience deformity in the joints, as well as stiffness, muscular atrophy and paralysis. The phalangeal and carpal joints of the patient's hands may become rigid and deviate to the ulnar side of the arm. The fingers may also become deformed with restricted mobility, and the elbow and shoulder joints may also be affected. In the late stages of arthritis, the treatment can only control the progression of the disease and alleviate local symptoms. The patient's locomotive function can only be gradually improved.

If the patient's lower limbs are affected, the ankle, knee and hip joints may become rigid, and he or she may experience motor impairment. If arthritis occurs in the patient's spinal column, it may cause kyphotic deformity and dyspnea, as well as restriction in neck movement. At this stage a sudden backward flexing of the neck may cause sudden death in patients with rigid forward bending neck deformity.

When treating an External invasion of pathogenic factors, the Qigong doctor observes three main symptoms of arthritis:

1. Wind or migratory arhtralgia is manifested by pain that has no particular fixed location, but moves throughout the body's extremities and joints, and is sometimes accompanied by chills and fever.
2. Cold arthralgia is manifested by localized pain of the joints, which may be relieved by heat and aggravated by cold.
3. Damp arthralgia is manifested by fixed pain and swelling in the joints and extremities, or by numbness of the skin which may be aggravated in cloudy and rainy weather.

TREATMENTS

1. Place the patient in the standing posture and begin dredging the Governing Vessel, drawing the energy into the ground.
2. Use the Extended Fan Palm hand technique and begin to emit Qi into the patient's GV-14, Mingmen, and Lower Dantian areas.
3. While emitting Qi into the painful joint areas, purge the pathogenic energy out of the patient's body, leading the Toxic Qi along the channels and out of the extremities.

PRESCRIPTIONS AND HOMEWORK

In treating arthritis it is important to prescribe exercises which lubricate the joint articulations to increase the production of Qi and Blood to the extremities. The Daoist Five Yin and Yang Organ exercises will accomplish this goal quite easily.

The patients are required to perform Qigong prescriptions which improve the body's resistance, but do not cause fatigue. Patients should also improve their diet and should be encouraged to consult a nutritionist for natural supplements that can help rebuild healthy cartilage and reduce pain. They should also make every effort to keep their body warm.

For patients with osteoporosis, stretching techniques, as well as sudden forward flexing of the neck are absolutely prohibited.

PRESCRIPTION MODIFICATIONS

1. When treating patients who suffer from arthritic pain in the upper limbs, advise them to practice shoulder and arm Dynamic Medical Qigong, moving at a slow, fluid pace.
2. When treating patients who suffer from ar-

thritic pain in the lower extremities, advise them to practice Dynamic Walking Medical Qigong (see Chapter 15), moving at a slow, fluid pace.
3. When treating patients who suffer from arthritic pain in the waist and lower back, advise them to practice the Expanding and Contracting the Rings exercise (see Chapter 41).
4. When treating patients who suffer from arthritic pain in the neck, advise them to practice slow-moving non-forceful neck rotation exercises in conjunction with stationary neck massage.

VERTIGO

Vertigo is a syndrome manifested mainly by dizziness and dim eyesight. It is defined as a sensation of moving around in space or having objects move about the person as a result of a disturbance of equilibrium. Vertigo is sometimes used as a synonym for dizziness, feeling light headed, or giddy.

Patients with mild vertigo may experience relief from this condition after a short bed rest, with their eyes closed. In severe cases, patients may suffer from a serious rotary vertigo and will not be able to stand or walk straight (e.g., sea and air sickness), sometimes followed by nausea, vomiting, perspiration, or even syncope (fainting).

Diseases that have vertigo as their main manifestation include cerebral arteriosclerosis, auditory vertigo, hypertension, anemia, and psychosomatic illnesses such as hyperventilation (which often accompanies anxiety disorders, especially panic disorders). These diseases may be diagnosed, categorized, and treated with reference to the prescriptions described in this section.

ETIOLOGY

Several factors contribute to the vertigo syndrome. In ancient times, Chinese physicians believed that vertigo was caused by Yin Deficiency and Excess Liver Wind, or by a head trauma. In Excess conditions, vertigo is caused by Phlegm Fire or Wind Phlegm. Currently, the clinical protocol for treating vertigo is divided into three conditions:

1. Hyperactive Liver Yang causing an insufficiency of Kidney Yin,
2. A Deficiency of Qi and Blood in the Heart and Spleen causing a Deficiency in the Marrow reservoir, and
3. Stagnation of Phlegm Dampness caused by a Deficiency of Spleen and Stomach Qi (which fail to transport the Phlegm).

Vertigo may be caused by a variety of factors affecting the ear, the eighth cranial nerve, and the brain stem; including middle ear disease, infectious diseases, trauma, tumors, toxemia due to food poisoning, hypertension, sun stroke, and toxic conditions such as those caused by alcohol and drugs. Over medication and malnutrition (especially in the elderly who suffer from dehydration) also cause vertigo.

SYMPTOMS

1. In cases where the patient has hyperactive Liver Yang, the syndrome of vertigo is characterized by severe dizziness, splitting headache, blurred vision, nausea, soreness and weakness of the waist and legs, mental fatigue (caused by emotional upset and anger), flushed face, constipation and tinnitus.
2. In cases of Qi and Blood Deficiency, the symptoms include rotary vertigo, blurred vision or night blindness, listlessness, disinclination to talk, palpitations, insomnia and anorexia.
3. In cases of stagnation of Phlegm Dampness, the symptoms include chest pain and distress, nausea and vomiting, poor appetite, listlessness of the body and heaviness in the head.

TREATMENT

When treating patients with vertigo, begin with the upper body working downward and from the medial aspect of the body towards the lateral part of the body.

1. Begin with the patient sitting on the edge of the table. Stimulate the patient's Baihui, the sides of the head, (GB-4, GB-8, and Taiyang areas), GV-14, Shendao, Mingmen, and the Lower Dantian.
2. Next, use the Vibrating Palm hand manipulation to emit Qi into the patient's Baihui, GV-14, Yellow Court and Lower Dantian for 12 to 24 breaths.

3. Use the Extended Fan Palm hand technique to emit energy into the patient's St-8, St-21 and St-36 points on both sides of the body.
4. With the left palm, emit Qi into the patient's Baihui point while using the right hand Sword Fingers to guide the patient's Qi from the Yellow Court down the Conception Vessel into the Lower Dantian.
5. Next, direct the Qi from the patient's St-8 points down to the St-36 points via the Stomach Channels.
6. Purge the energy of the upper torso out through the hands via the Large Intestine Channels.
7. Finally, lightly pat and knead the patient's Baihui and GV-14, softly rocking their upper limbs to end the treatment. Softly pressing and kneading the patient's tissues along their channels can push Qi and Blood downward to free the patient's mind from Toxic Qi and can improve the eyesight. The soft grasping can relieve stasis of Qi and Blood in the patient's blood vessels and promote circulation, thus reducing and relieving vertigo.

TREATMENT MODIFICATIONS

1. In cases of hyperactive Liver Yang, use the Extended Fan Palm hand method to emit Qi into the patient's Bl-18, Mingmen, and Lower Dantian. Guide the Qi to flow along the channels (or in a clockwise direction) to nourish the patient's Yin and to suppress the hyperactive Yang.
2. In cases of Deficient Qi and Blood, use the Extended Fan Palm hand method to emit Qi into the patient's mid-back (Bl-17, Bl-18, and Bl-20 points) and Lower Dantian to nourish the patient's Qi and to replenish the Blood.
3. In cases of stagnation of Phlegm Dampness, use the Extended Fan Palm hand method to emit Qi into the patient's Yellow Court. Guide the Toxic Qi down the Stomach Channels and expel the pathogens out from the patient's St-36 points.

PRESCRIPTIONS AND HOMEWORK

To treat vertigo, have the patient practice the Descend the Qi and Cleanse the Organs exercise (see Chapter 41).

PRESCRIPTION MODIFICATIONS

When treating patient's who suffer from a Deficiency of Kidney Yin and a hyperactivity of Liver Yang, prescribe the following meditation. Have the patient take a standing or sitting posture, breathing naturally, and relaxing. While inhaling through the nose, the patient imagines blue energy filling the mouth. When exhaling, the patient imagines this dark blue energy slowly pouring down the throat, like warm water, filling the Kidneys and Lower Dantian area. The patient should practice this meditation for 7 breaths.

HEADACHES

A headache is defined as a diffused pain in different portions of the head which is not confined to any one nerve distribution area. It may be acute, chronic, frontal, temporal, occipital, confined to one side of the head, or affect the region immediately over one or both eyes. The character of the pain may vary from dull and aching, to acute stabbing pain that is almost unbearable. There may be intermittent, intense pain, a throbbing pain, a pressure in the head which feels as if it will burst the head, or a penetrating pain driving through the head.

ETIOLOGY

Transient, acute headaches may vary in their etiology. Disease of the perinasal sinuses, teeth, eyes, ears, nose or throat, acute infections, invasion of Wind Cold pathogens into the channels of the head, or trauma to the head are among the most frequent causes of headaches.

Chronic headaches may also be caused by a variety of conditions including physical, emotional, psychosomatic, or psychogenic factors, including fevers, metabolic conditions (hypertension, etc.), or even exposure to toxic chemicals.

SYNDROMES

Headache syndromes can be divided into five categories which are described as follows.

1. A Shaoyang headache (migraine) is clinically manifested through symptoms such as a severe headache (splitting pain) in the temple region accompanied by Excess Heat in the

head, conjunctive congestion, and sometimes intercostal fullness or pain. Nausea often accompanies a severe migraine attack.

2. A Taiyang headache (occipitocervical headache) is clinically manifested through such symptoms as a headache and pain in the back of the head which can be accompanied by fever, aversion to cold, and rigidity at the back of the neck.
3. A Jueyin headache (vertical ache) is clinically manifested through such symptoms as a vertex headache, vertigo, insomnia, dizziness, and irritability with a flushed face.
4. A Yangming headache (frontal ache) is clinically manifested through such symptoms as a headache in the upper part of the cranium accompanied by dysphoria (depression and anxiety) with a smothering sensation, thirst, bad breath and constipation.
5. A Deficient Qi and Blood headache is clinically manifested through such symptoms as a lingering headache with a heavy sensation in the head, blurred vision, lassitude, and weakness.

TREATMENT

Begin by purging and dredging the patient's Upper Dantian, focusing on the Yintang, Taiyang, Baihui, GV-16, GV-20, and GB-4 points. Purge the pathogenic Qi down the patient's arms and out the LI-11 and LI-4 points, as well as down the Gall Bladder Channel out of the GB-41 points. The point and channel purging is specific to the location (right or left side) of the headache.

Next, using the Extended Fan Palm hand technique, emit Qi into the patient's headache region, purging it down the Conception Vessel to the Lower Dantian.

TREATMENT MODIFICATIONS

1. When treating a patient with a Shaoyang headache, apply the Extended Fan Palm hand technique to emit Qi into the GB-4, GB-8, and TB-20 points, leading the pathogenic Qi out of the Triple Burners and Gall Bladder Channels.
2. When treating a patient with a Taiyang headache, apply the Extended Fan Palm hand technique to emit Qi into the LI-11 and Bl-10 points, Purging pathogenic Qi down the Large Intestine and Bladder Channels and out of the hands and feet.
3. When treating a patient with a Jueyin headache, start at the head using the Circle Spiraling hand technique in a counterclockwise direction, purging the Toxic Qi out the patient's Gall Bladder Channels and out of the legs.
4. When treating a patient with a Yangming headache, use the Extended Fan Palm hand technique to emit Qi into the patient's Taiyang, Yintang, and St-8 points, purging the Toxic Qi down the Stomach Channels and out of the legs.
5. When treating patients with Deficient Qi and Blood, use the Extended Fan Palm hand technique to emit Qi into the patient's mid-back, at the Bl-17 and Bl-20 points, and the chest at the CV-17 point. Then emit Qi into the patient's CV-12 and Lower Dantian areas.

PRESCRIPTIONS AND HOMEWORK

1. It is advisable to have the patient practice the Ten Dragons Run Through the Forest exercise found in the Wash the Face and Massage section of Chapter 39.
2. It is beneficial to prescribe the following meditation to prevent the advancement of an acute attack. The main focus of this meditation is to direct Blood, Heat, and Qi to the body's upper extremities (the arms and hands), away from the original source of pain (i.e., the head).
   a. From a sitting, semi-reclining, or lying posture, the patient begins to relax and quiet the mind, with the eyes closed.
   b. The patient imagines that it is summertime and he or she is at the beach. Next, the patient focuses on both arms and hands, imagining them growing warmer and warmer in the hot sun. The patient imagines feeling massive amounts of heat radiating from the arms and hands, redirecting the Blood flow away from the head.
   c. Then, the patient imagines the face becoming cooler, as the pressure and pain

melts down the neck, shoulders, arms, forearms and out of the hands.

Prescription Modifications

1. For patients with a Shaoyang headache, focus the prescription meditations and exercises on treatments that purge the bilateral sides of the patient's head along the Triple Burners and Gall Bladder Channels, dredging the pathogenic Qi down and out of the patient's arms and legs.
2. For patients with a Taiyang headache, focus the prescription meditations and exercises on treatments that purge the patient's occiput and cranium, the Small Intestine and Bladder Channels, and that dredge the pathogenic Qi down and out of the patient's arms and legs.
3. For patients with a Jueyin headache, focus the meditations and exercises on treatments that purge the patient's vertex cranium, dredging the Toxic Qi down the Liver Channels and out of the legs.
4. For patients with a Yangming headache, focus the prescription meditations and exercises on treatments that purge and dredge the energy from the patient's forehead down the Large Intestine and Stomach Channels and out of the legs.
5. For patients with a Deficient Qi and Blood headache, focus the prescription on Lower Dantian Cultivation exercises and meditations.

MIGRAINE HEADACHES

This type of headache initially consists of a constriction in the head arteries, followed by vasodilation of the extracerebral cranial arteries.

ETIOLOGY

Migraine headaches are usually brought about by stress, and can also be initiated by an allergic reaction to what the patient's body considers to be a toxic invasion (e.g., cheese, chocolate, red wine, contraceptive pills, etc.). These types of stresses on the body's system result in internal organ dysfunction, which is generally caused by one or more of the following three factors.

1. The energy of the Heart and Spleen are not regulated.
2. A rising up of Liver Heat and Fire occurs.
3. A Deficiency of Kidney Yin and rising of Heart Fire occurs simultaneously.

The result of stress-related internal organ dysfunctions causes the initial constriction of the arteries in the head, followed by vasodilatation and distention of the vessels.

SYMPTOMS

The clinical manifestations of migraine headaches are hypersensitivity to light and sound, followed by severe, throbbing pain (most commonly felt on only one side of the head), nausea, dizziness, and sometimes vomiting. Some migraine patients experience seeing an aura or visual disturbances (colors, shapes, lines, flashing spots, or temporary reduction in the field of vision) before the head pain begins.

TREATMENT

For treatment, the Qigong doctor must pull the pathogenic energy out of the patient's head, from the location of the pain; then lead the Excess Toxic Qi out of the patient's extremities, and open up the channels so that Qi can rise and fall to the patient's head unobstructed.

1. In a severe case, combine Medical Qigong therapy with Jing Point therapy.
    a. Begin the treatment by pulling out (with strong intention) the stagnant Qi trapped in the patient's head. First, purge the patient's Wei Qi field, then enter the patient's Jade Pillow at the base of the head. Next, emit Qi along the falx cerebri and circulate it using your intention to create a small Microcosmic Orbit (circling from the back of the head to the nose and back). This encourages the patient's stagnant Qi to move.
    b. Next, dredge around the area of the patient's eye (from the inner corner of the eye to the side of the eye), starting at the Bl-1 point to the Yin Tang point (Third Eye), following the sphenoid arch across to the temple at the Taiyang point, then to the back of the ear at the

TB-17 point. Next, lead the pathogenic Qi down the neck and pull out the stagnant Qi through the extremities (Pc-6 and LI-4 on the hand). If the patient is feeling nausea, purge the Toxic Qi from the Yellow Court and lead the Toxic Qi down and out the body through the St-36 point on the leg and Lv-3 point on the foot.

   c. To remove sluggish Qi stagnation, squeeze the patient's trapezius muscles, draining the pathogenic Qi out of the axillary fold of the patient's arm. The patient should breathe into the painful areas, and imagine any remaining Toxic Qi descending down the arm into the fingertips when exhaling. The doctor should squeeze the patient's arm as the patient exhales, to support the descending action of the Toxic Qi.
   d. The patient should be given homework, depending on his or her constitution.

2. In treating chronic migraine headaches, combine Medical Qigong therapy with An Mo therapy, working from the branch to the root.
   a. Begin the treatment by pulling out the stagnant Qi trapped in the patient's head. This is done by first purging the patient's Wei Qi field, then entering the patient's Jade Pillow (at the base of the head) to purge and remove Toxic Qi and stagnation. Then emit and guide the Qi with your intention, to flow along the patient's falx cerebri. This forms a small Microcosmic Orbit (circling from the back of the head to the nose and back) and encourages the patient's stagnant Qi to move.
   b. Extend Qi into the patient's Bl-1 point (on the inside of the eye).
   c. Next, focus energy into the patient's Yin Tang point and begin circulating the Qi along the orbicularis oculi muscle of whichever eye is in pain (along the eyebrows to the ear). Work from the inside of the eye, out along the eyebrows (Figure 48.1).

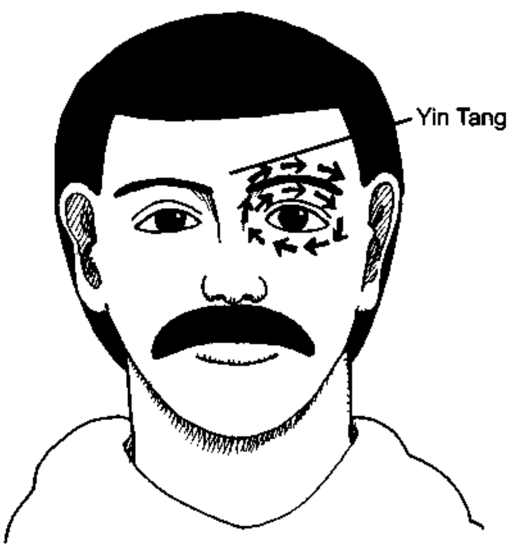

Figure 48.1. For treating chronic migraine conditions (which dominate behind the left eye), begin by extending Qi into the patient's Bl-1 point and start to dredge the Toxic Qi from the patient's Yin Tang area, circulating the orbicularis oculi muscle of the eye in a clockwise direction.

   d. After several rotations, follow the sphenoid arch to the back of the ear at the TB-17 point (if the pain is bilateral, for men treat the left side first; for women treat the right side first).
   e. Next, beginning at the TB-21 point, outline the outside of the ear, then massage the lateral aspect of the ear's ridge (upper auricle) (Figure 48.2).
   f. Hold both of the patient's ears firmly and begin to apply slight pressure at a 45 degree angle towards the sides of the patient's head. This allows the Qi trapped within the patient's head to escape. Using intention, move the Qi from the Jade Pillow (along the tentorium cerebelli) into the patient's pituitary gland. Then divide the energy into two rivers of Qi that circulate through the middle of the cerebrum and end back at the Jade Pillow.

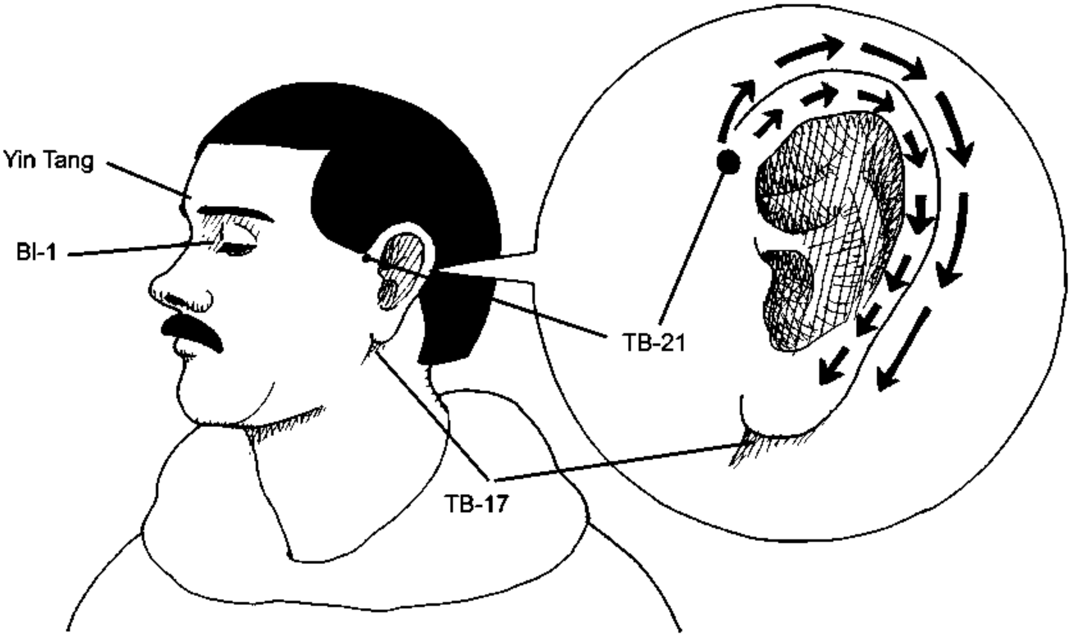

Figure 48.2. Outline the outside of the patient's ear, then massage the lateral aspect of the ear's ridge.

g. From the front of the head, at the edge of the hair, gather the Qi and lightly scratch the patient's head, as you use the Kneading Tiger technique. Visualize pulling the Toxic Qi down the front and sides of the patient's head (down the Gall Bladder Channels) to the base of his or her neck (Figure 48.3).

h. From the base of the neck, pull the Toxic Qi down and out of the arms.

i. The patient should be given seated Quiescent Qigong meditation homework, the length of meditation time will depend on his or her constitution.

Note: When treating migraine headaches, the head should always be treated first before treating the extremities.

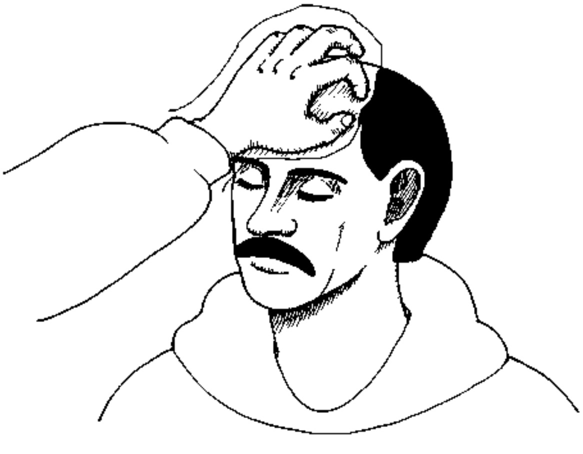

Figure 48.3. Using the Kneading Tiger technique, visualize pulling the pathogenic Qi out of the patient's head, down the side of the neck and down along the sides of his or her arm.

SECTION XI

SPECIALIZED QI EMISSION THERAPY FOR PEDIATRICS, GERIATRICS, GYNECOLOGY, NEUROLOGY, PSYCHOLOGY, ONCOLOGY, AND SURGERY

INTRODUCTION

In this section we will explore disease treatment techniques and prescriptions unique to Medical Qigong doctors who specialize in Pediatrics, Geriatrics, Gynecology, Neurology and Oncology.

The last two chapters will address Medical Qigong therapy and the treatment of radiation, chemotherapy, and surgery.

CHAPTER 49

QI EMISSION THERAPY FOR PEDIATRICS

INTRODUCTION

Pediatrics is the branch of medicine which deals with the care of children and the treatment of childhood diseases.

From a traditional Chinese medical perspective, children are different than adults in several important ways. First, the Yin and Yang energy in children is fresh and pure, and their vitality usually flourishes. Second, although the internal organs of children have begun to take shape, the organs are still in the process of developing Qi; the organs are still maturing and are not yet fully developed. Because children are still growing organs and tissues, stimulating the channels and channel points has a different effect on their body than on an adult.

Specifically, the Lungs, Spleen, and Kidneys are especially delicate in children. The Lungs and Spleen are weak, and the Kidneys are particularly vulnerable. The Lungs control the Qi of the whole body. The Spleen is responsible for the transformation of food and the transportation of Gu Qi (food Qi). The Lungs and Spleen are the postnatal foundation of health. The Kidneys are in charge of the bones and Marrow and are the prenatal foundation of health. Children are still full of Prenatal Qi; they are just beginning to use their Lungs and Spleen for the function of acquiring Postnatal Qi. Consequently, their energetic system is vulnerable and unstable.

If children are born without congenital defects or weakness, are nursed properly, and avoid contracting major diseases, the internal organs should develop normally. However, because of inherent weakness or immaturity of the digestive system, almost all pediatric diseases of children under the age of six usually begin with some element of indigestion. A child's tiny digestive system makes him or her prone to the creation of Phlegm. Since the Spleen is the root of Phlegm production, and the Lungs are the storehouse of Phlegm, then a congenitally weak Spleen can lead to the creation of Turbid Dampness within the child's body. This Turbid Dampness can congeal into Phlegm and lodge within the child's Lungs. These symptoms can be observed in young children as runny noses, ear infections, phlegmy coughs, and other respiratory difficulties.

EARLY STAGES OF CHILD GROWTH AND DEVELOPMENT

In childhood, growth and development occur rapidly. In the ancient Chinese medical classic entitled *"Prescriptions Worth a Thousand Gold Ducats"* it states, "A newborn is feeble and tender; two months later, its pupils begin to become well developed, it can smile and distinguish parents from strangers; 150 days later, its sacral bones have been well developed, it can sit independently; 210 days later, its metacarpal bones have been well developed, it can crawl; 300 days later, its patellae have been well developed, it can stand alone; at one full year of life, its shank and occipital bones have been well developed, it can walk and speak."

In the natural course of childhood growth and development, there may be fever, irregular pulse signs, or perspiration due to the imbalances of Yin and Yang. The child usually recovers spontaneously within a day or two. Such episodes are called Changing and Steaming. Changing implies the transformation of the Five Yin Organs, with the accompanying variations of emotions, and the development of the child's intelligence. Steaming implies the vaporous exchanges within the six Yang organs induced by accumulated Heat. The ancient Chinese discovered that there is one Changing every 32 days, and one Steaming every 64 days. Thus, within 320 days, there are 10

Changings, and within 576 days, there are 9 Steamings. Every Changing or Steaming is an ordinary event in the course of growth and development, and should not be mistaken for a condition of childhood disease. However, if children are not cared for properly during Changing and Steaming, they are more apt to fall ill.

PATHOLOGY

In "*Detailed Analysis of Epidemic Febrile Diseases*," Dr. Wu Tang stated that children's diseases are apt to change because the internal organs are frail and their defense mechanisms are incomplete; children contract infections easily because the skin and muscles are still tender, and the vitality is timid and weak. Propensity to illness, inclination to Excess and Deficiency, and being prone to suffer from Heat and Cold syndromes are common features of pediatric pathology. Since children are full of pure Yang, the course of disease is changeable and unpredictable. Nevertheless, recovery of health is generally easy.

DEFICIENT LUNGS

Children are apt to suffer from internal organ disorders, especially of the Lungs, manifested by symptoms of cough, dyspnea, retention of Phlegm and Dampness in the Lungs, seasonal febrile diseases, allergies, asthma, and other Lung disorders.

The Lung is a delicate organ in charge of the skin and hair. Therefore, particularly in children, junctures between the skin and muscles are not compact enough to resist the invasion of Exogenous Cold, Heat, Dampness, Wind, and other pathogens. Seasonal pathogens may attack the Lungs through these slackened junctures and through the nose, thereby interfering with the descending and dispersing functions of the Lungs. This results in the accumulation of Heat, fullness in the chest, productive cough, or dyspnea.

DEFICIENT SPLEEN

Since the Spleen is not yet matured and thus usually Deficient, it may be harmed or damaged due to improper diet, irregularity of meals, or imbalances caused by excessively Hot or Cold diets. Consequently, Clear Yang fails to rise, Turbid Yin fails to descend, and Stomach Qi becomes disharmonious. Therefore, children frequently suffer from abdominal masses, abdominal pain and distension (stomachache), belching, hiccups, vomiting and diarrhea, infantile malnutrition, and other digestive disorders. These conditions often lead to children becoming hypersensitive to the foods they consume.

According to Five Elements theory, the Spleen (Earth) generates the Lungs (Metal) and the Lungs (Metal) generate the Kidneys (Water). The elements work in coordination and influence each other. In children suffering from Spleen Deficiency, there will also be a Deficiency of Lung Qi; in children suffering from Lung Deficiency, there will also be a Kidney Deficiency.

DEFICIENT KIDNEYS

Kidney problems can manifest in two different ways: Kidney Jing Deficiency, which can cause retardation, or Congenital Qi Deficiency (i.e., Deficiency in the Sea of Marrow), which can lead to Down's Syndrome, Attention Deficit Disorder (ADD) and learning disabilities. Since the Kidneys control the lower orifices, Deficient Kidney Qi can also lead to chronic bed wetting.

If the child was improperly nursed when they were young, the Deficient Kidney Yin not only fails to nourish the bones and Marrow, but also fails to nourish tendons. During depletion of Liver and Kidney Yin, there can be atrophy, weakness of tendons, five types of developmental retardation, as well as five kinds of flaccidity.

LIVER YANG EXCESS

By contrast, Liver Yang is usually Excessive, and Liver Wind may be brought on by Heat or Fire, resulting in infantile convulsions, twisting, and even opisthotonos.

HEART (SHEN) IMBALANCE

The Heart is responsible for the child's mental stability. The Heart Fire and the Kidney Water must constantly be balanced to stabilize the child's Wu Jing Shen. An imbalance of the Heart can lead to Shen disturbances which can result in mental instability or disability.

According to the active energetic balance of the child's Wu Jing Shen, the Hun tend to naturally be overactive, and the Shen of the Heart is generally unstable. The reason for this instability is because the Hun are very active within a child. The Hun are responsible for creativity, enthusiasm, dreaming, going in and out of the body, and seeing into the Spirit world. The Shen, which is responsible for the child's discriminating mind, respect and conveying the appropriate behavior, is not fully developed. Consequently, young children tend to often be wild, hard to control, and get easily bored.

In treating a child's overactive Hun and unstable Shen, the Qigong doctor harmonizes the child's Heart Fire and Kidney Water energy, and opens the Heart's orifices (using Qi massage on the child's eyes, ears, nose, mouth and all upper sensory orifices) to stabilize the Wu Jing Shen.

HOT AND COLD EXOGENOUS FACTORS

Susceptibility to Exogenous Cold and Heat reflects the characteristics of puerile Yin and Yang. When children are invaded by Exogenous Evils, for example, there will be Excess Heat syndromes; when the child's healthy energies are depleted, there will be Deficiency Cold syndromes.

Although various defects and Deficiencies do exist in children, their Yang, Liver, and Heart are in a hyperfunctional state. Therefore, children are full of vitality and their anti-pathogenic responses are sensitive and immediate. The causes of childhood diseases are usually comparatively simple, relative to adults, and their convalescence is easier and quicker provided they receive appropriate treatment and care.

THERAPY

Because the internal organs of children are still forming and are, therefore, still frail and tender, children are much more sensitive than adults, and cannot stand strong energetic stimulation. This is why the primary modality for treating young children is gentle pediatric Tui Na, which is usually on the child's hands, feet or abdomen, or Qigong massage (lightly touching the child's body). Acupuncture is used only infrequently, and when it is, the needles are usually not retained for long - the points are quickly and lightly stimulated and then the needles are removed. Similarly, when using Medical Qigong emission therapy, a milder stimulation is applied than would be used for adults when purging, tonifying, and regulating.

Medical Qigong is useful for treating pediatric neurological disorders, and particularly for such conditions as epilepsy, hyperactivity, attention deficit disorder (ADD), and mental retardation. Effective points for tonification include Tianzhu (Bl-10), Sishencong (four extra points, located around the Baihui area at the top of the head), Baihui (GV-20), Yamen (GV-15) and Fengfu (GV-16).

When children under the age of six practice self-regulation Medical Qigong exercises (such as Qigong to increase intelligence), they frequently develop special abilities. These abilities include ESP (e.g., telepathy and telekinesis), and special Qigong healing abilities.

ENERGETIC MOVEMENT AND FLOW

The child's Microcosmic Orbit naturally flows in the direction of the Water cycle, moving up the front (CV) and down the back (GV) in order to facilitate the child's spiritual, intuitive, and psychic perceptions. This energetic movement continues in this direction until the child reaches puberty, at which time the energy reverses its direction and flows up the back (GV) and down the front (CV), in the direction of the Fire cycle. This energetic switch in direction facilitates the child's cognitive development and the ability to control emotions and impulses. The time of the Microcosmic Orbit's energetic reversal varies depending on the child's physical constitution, state of health, and his or her environment.

MEDICAL QIGONG EXERCISE PRESCRIPTIONS FOR CHILDREN

The following Medical Qigong exercise is considered a basic training method for children. This exercise is specifically developed for children of all ages and physical conditions, and is used to improve a child's physical and mental health, as well as aid in curing some common childhood diseases. This exercise can also help to cure food allergies, loss of appetite, malnourishment, a weak immune system, neurasthenia, insomnia, dreaminess, and disturbances of the Spleen and Stomach. It is believed that this exercise also helps moderate a child's stubbornness and impatience.

IMPROVING INTELLECTUAL FACILITIES QIGONG (DA-ZHI-GONG)

This particular Medical Qigong exercise is excellent for children who range from 3 to 12 years old. The time and frequency of the following Qigong exercise is not fixed. It is important, however, that the child be consistent with the practice, and that the body and mind remain relaxed while training.

It is important that the child practice the entire Qigong sequence in its proper order. The child is not allowed to practice only a part of the exercise.

Prohibited from this particular exercise are children who have weak constitutions (i.e., have difficulty in standing up) or suffer from serious diseases.

1. Begin by having the child assume a standing Wuji posture. Have him or her relax the body, with the eyes either open or closed. The child then raises both hands in front of the Upper Dantian. Boys will place the left hand facing the forehead, about three inches from the Yintang point. Girls will place the right hand facing the forehead, about three inches from the Yintang point. At the same time the opposite hand should be level with the other palm and should be facing outwards (Figure 49.1).
2. Both hands begin to make a clockwise circle rotation (from the child's point of view). The child keeps the Yin Tang point as the center of the circle, for 18 breaths. Next, the child moves both hands to the front of the right eye and continues the same clockwise movement while keeping the right eye at the center of the circle. Rotate for 18 breaths. Then, the child moves both hands to the front of the left eye and continue the clockwise movement for 18 breaths. Finally, the child returns to the front of the forehead and continues the rotation, keeping the Yintang point as the center for 18 breaths, then

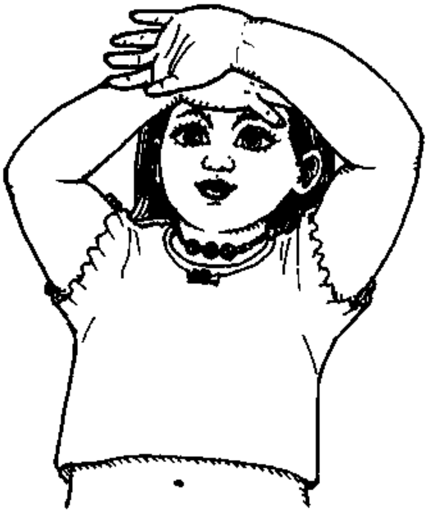

Figure 49.1. Improving Intellectual Facilities (1)

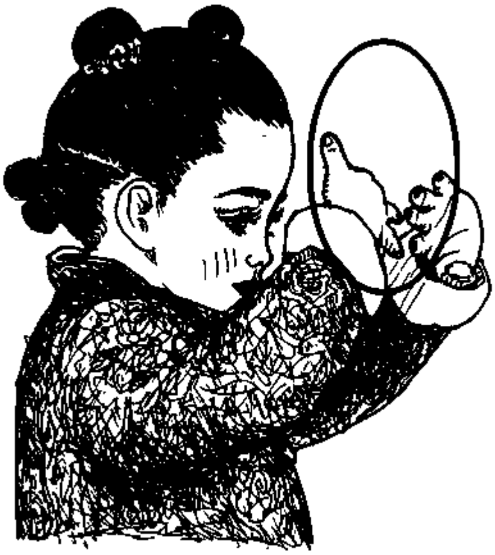

Figure 49.2. Improving Intellectual Facilities (2)

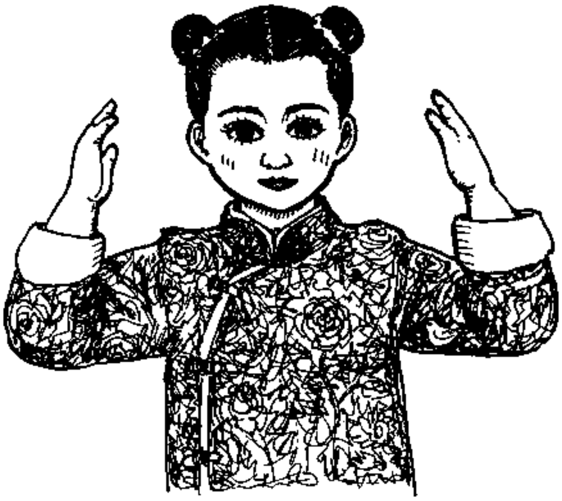

Figure 49.3. Improving Intellectual Facilities (3)

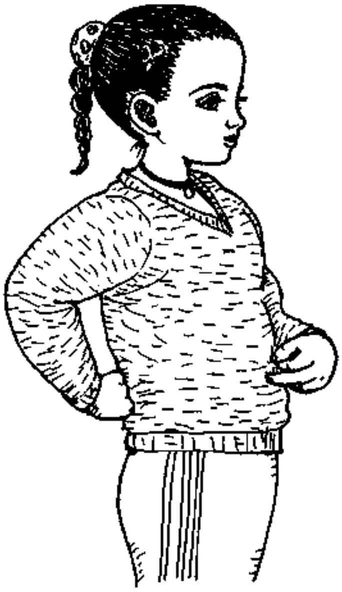

Figure 49.4. Improving Intellectual Facilities (4)

ends the meditation (Figure 49.2).

3. The child places both hands in front of the ears, with the center of the palms facing the external opening of the ear canals; breathing naturally, the child moves both hands forward and backwards while emitting Qi into the ears. Make sure the child does not concentrate on any specific thing, but simply relaxes (Figure 49.3).
4. Next, the middle fingers softly touch the navel with the left hand, and the right hand softly touches the Mingmen. Have the child imagine that there is a big red ball of fire in his or her abdomen (where the child's two middle fingers are pointing). The child's fingers should remain is this position until warmth is felt in the abdomen. If the heat becomes too hot, have the child exhale the sound "Ha" three times to release the Excess Heat (Figure 49.4).
5. To end the exercise, have the child cover his or her tailbone with both hands and jump up and down gently for several minutes. Do not allow the child to become tired from jumping. Again, make sure that the child does not concentrate on any specific thing (Figure 49.5).

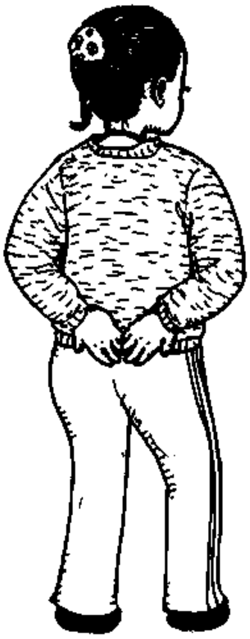

Figure 49.5. Improving Intellectual Facilities (5)

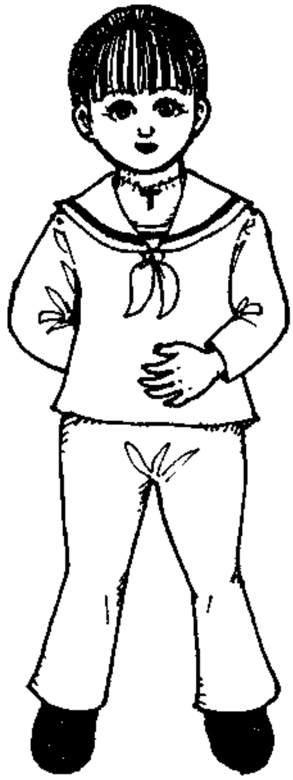

Figure 49.6. Gathering Wisdom Qigong (1)

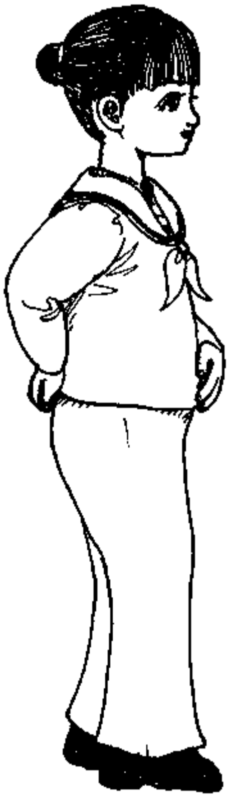

Figure 49.7. Gathering Wisdom Qigong (2)

MEDICAL QIGONG FOR GAINING WISDOM QIGONG (TONG-LING-GONG)

This particular Medical Qigong exercise is excellent for children who range from 3 to 16 years of age. The time and frequency of this particular Qigong exercise is not fixed, and may be decided by the child. Also, there are no contraindications for this particular Qigong exercise.

1. The child begins by standing with his or her feet flat, shoulders width apart. The child should place one hand on the Lower Dantian (below the navel) and the other hand at the back, below the Mingmen (opposite the Lower Dantian). Have the child close his or her eyes and relax the body and mind for about three minutes. The child is to imagine that the upper part of the body (above the waist) is extended into the blue sky above, and that below it (around waist level) are clouds (Figure 49.6).
2. Next, the child imagines that he or she has changed into an Immortal (a Super-Hero). Encourage the child's imagination to flow freely with this concept, allowing the child to move about, gently shaking the knees (Figure 49.7). After several minutes end the meditation.

CHAPTER 50
QI EMISSION THERAPY FOR GERIATRICS

Geriatrics is a branch of medicine which deals with the problems of the aging. Geriatric specialists addresses the physiological, pathological, psychological, economic, and sociological problems of the elderly. The importance of geriatrics is emphasized by the fact that the expected life-span of the average person is increasing.

Diseases of the internal organs can spread throughout the body via the channels and collaterals, and vice versa, as channel diseases can be transmitted to the internal organs. If Qi and Blood lose their harmony due to old age, and the patient's internal organs, channels, and collaterals become obstructed, then Phlegm and Stasis will develop internally. The accumulation of Phlegm, Qi and Blood stasis may lead to a chain reaction of pathological transformations, such as Hot, Cold, Excess or Deficiency Syndromes in the patient's internal organs. The following are examples of some of the pathological transformations.

1. If the patient's Qi and Blood in the Liver lose their harmony, the patient may experience fatigue and depression with dry eyes or diminished vision. The patient may also experience numbness in the extremities and muscle spasms due to a weakening of the sinews and blood vessels.
2. If the patient's Qi and Blood in the Heart lose their harmony, the patient may experience poor memory, insomnia, heart palpitations, and shortness of breath (which may be aggravated by movement).
3. If the patient's Qi and Blood in the Spleen lose their harmony, the patient may experience fatigue and general weakness in all four limbs. The patient may also experience loose or bloody stools, and there will be abdominal distention after food consumption.
4. If the patient's Qi and Blood in the Lungs lose their harmony, the patient may experience coughing with Phlegm and panting, as well as spontaneous sweating.
5. If the patient's Qi and Blood in the Kidneys lose their harmony, the patient may experience poor memory, anxiety, tinnitus or deafness, as well as a sore or weak lower back, knees, and teeth.

The Chinese have a saying, "When the Heaven (Heart Qi) connects with the Earth (Kidney Qi), all seven emotions are kept in moderation." When Qi becomes scattered the Heart Fire and the Kidney Water cannot converge. This results in an unbalanced and unstable energy flow.

The Kidneys are extremely important for health maintenance and healing in all patients, and especially in senior citizens. As one gets older, the Kidneys, lower back, and legs are the first things to be affected.

It is important to stabilize the Heart because the Heart is responsible for mental and emotional orientation. Traditional Chinese Medicine holds that, "when the Heart is moved, all the other organs will be shaken."

In teaching Medical Qigong to senior citizens, first quiet the Heart, so that the other organs can begin to relax. When the mind becomes disturbed, the body becomes disturbed. To affect one is to affect the other.

The eyes are considered the seedlings of the Heart as they receive information, gathering it and feeding it to the brain. It is important for the Heart's Qi to descend into the Lower Dantian to calm the mind. When the patient is in a hurry, the Heart is in a hurry, thus inducing the Qi to rise to the head. This results in nervous tension and anxiety. It is only through rest and quiet training that the mind and the Heart can begin to settle.

One method of quieting the mind is through prayer and meditation. This is considered "active rest." Sleep is considered "passive rest." When focusing inward and not using the eyes to see or the ears to hear, the Qi begins to converge in the middle of the body and the energy is drawn back into the organs, settling down into the Lower Dantian. This convergence of Qi nourishes and cultivates the natural energetic structure of the internal organs. Senior citizens are encouraged to sit for longer periods of energy cultivation. If they practice Taijiquan they are to avoid low postures and big motions. The arm movements should be kept small, soft, and slow.

TREATING THE PATIENT'S EMOTIONAL STABILITY

When treating the elderly, the Qigong doctor bases the regulation therapy prescriptions on two major priorities. The first is to achieve emotional stability and the second is to improve the Qi and Blood circulation.

The first priority in working with healing prescriptions is to focus on the patient's psychological profile. When teaching Medical Qigong to senior citizens, first teach them how to quiet the Heart. Once the Heart is calm the other internal organs begin to relax. Always keep in mind that when the mind becomes disturbed, the body becomes disturbed. To affect one is to affect the other.

Depression and sadness can have a profound effect on the patient's healing ability. Medical Qigong regulation can be a strong aid in releasing an elderly patient from a chronic, detrimental attitude that complicates the healing cycle. Take for example the three stages of emotional transition the patient endures before his or her Spirit closes (see Chapter 18).

If anger and grief are not allowed to be expressed by the patient (which is the first stage of energetic expression and release of emotional pain), then anger transforms into depression, and grief transforms into despair. This is the second stage of the energetic/emotional transformation due to emotional pain. If the patient's depression and despair are not alleviated, then indifference, the final stage of energetic/emotional transformation, is developed. The patient becomes withdrawn, apathetic and either antisocial, or obsessive/compulsive in his or her thinking and behavior.

When the patient's Spirit closes, they tend to be extremely difficult to work with, and can easily sabotage their own healing potential. Though the use of Soul Retrieval meditations (see Chapter 19), Medical Qigong prescriptions can be used to reconnect the patient with their disassociated emotions. This allows the patient the ability to feel and regain contact with his or her true feelings.

Pleasant emotions bring about a calm and relaxing physiology, while strong, chronic negative emotions cause Qi and Blood stagnation, as well as Excess conditions within the patient's body, effecting the quality of the body's life force energy.

ENERGIZING THE PATIENT'S QI CIRCULATION

The second priority is to get the patient to move and circulate the Qi and Blood deep and superficially within his or her body. Static Blood is considered the primary mechanism of senility. All growth, development, decline, and debility of the human body, is related to the condition of the patient's Qi and Blood.

Pathophysiologically, both Qi and Blood affect each other; if one becomes diseased, the other is affected. Due to the Kidney's influence, as seniors get older, their hearing, hair, memory, lower back, and legs become adversely affected. Traditionally, all senility-combating formulas of Medical Qigong stress treating the patient's Spleen and Kidney Deficiency by supplementing the patient's Yin and/or Yang. Then, by applying the balancing methods of activating the patient's Blood and rectifying the Qi, the doctor can also combat the patient's senility.

Senility is also frequently caused by dehydration and malnutrition. Both conditions are frequently overlooked by Western medicine. The dehydration is usually due to a breakdown of the body's thirst signal. The patient does not feel thirsty, and thus forgets to drink. Malnutrition may be caused by a lack of enzymes that aid in absorbing nutrition from foods. Overmedication likewise can cause senility. These issues should always be addressed when treating the elderly, along with Medical Qigong exercises and meditations.

One example of a Medical Qigong exercise for aiding Qi and Blood circulation for geriatric patients is the Microcosmic Orbit (Fire Cycle) which can be practiced in a sitting position.

The elderly are more subject to musculoskeletal disorders and visceral diseases, especially in the Kidneys and Liver. As the Kidney Yang rises, it begins to restrict the heat in the upper portion of the body, thereby affecting the Heart. Also, as the patient gets older the Lungs cannot sustain liquid retention; this affects the Wei Qi, and results in a thinner layer of skin covering the body.

The Qigong movements prescribed for the elderly should be kept slow and smooth. Qi regulation can be cultivated safely by focusing the patient's intention on the center of the palms while keeping the mind's intention moving downward. The patient must not exert strain when using the muscles. When the patient turns his or her body, make sure that the center of each palm (Pc-8) faces the other. Slow and even breathing will calm the patient's emotions and sedate the mind.

It is important to begin Medical Qigong regulation training for senior citizens with natural breathing. Natural breathing allows their respiration to become soft, natural, gentle and quiet. The focus of the mind should be on relaxation. In dealing with senior citizens, if the focus of their mind's intention is allowed to become too concentrated, it can cause the Qi in their body to constrict and develop into Qi or Blood stagnation. Redirect their focus of intention on their breathing by practicing the Small Heavenly Cycle meditation. The Small Heavenly Cycle meditation is used to regulate the body's Yin and Yang energy and to strengthen the patient's mind and spirit. The Large Heavenly Cycle meditation can be prescribed to increase the patient's Qi and Blood circulation, to rejuvenate the body and the sense organs, to clear the complexion, and to improve mobility.

TAPPING THE KNEES TO PREVENT SENILITY

One exercise prescription that the doctor can use to assist the patient in combating senility is to have the patient lightly tap above, below, and around his or her knees. This exercise stimulates the Kidneys' energy and can also be used as an important adjunct to maintaining health for senior citizens, especially when combined with the foot tapping (Kd-1) Kidney Tonification exercise (see Chapter 39). Combined, both of these exercises can help to strengthen the patient's Kidneys, lower back, and legs. When combining both exercise prescriptions, the Knee Tapping exercise should follow after the Kidney Tonification exercises.

To begin the Knee Tapping exercise, have the patient sit in a chair, with both knees lower then the thighs. Encourage the patient to direct the focus of his or her attention in and around the knee area. As the patient taps or softly slaps the tissue surrounding the knees, he or she should imagine the energy and vibration flowing into the Lower Dantian. After several minutes (about 36 breaths) the patient can stop.

CULTIVATING THE PATIENT'S YUAN QI BY ABSORBING ENERGY FROM NATURE

Another priority of focus for older patients is the gathering, storing and cultivation of his or her Yuan Qi. In China, in order to tonify the Yuan Qi, elderly patients practice their Qi Tonification and Regulation Exercises facing a cypress tree daily (a cypress tree is considered very powerful and full of energy). If a cypress tree is not available, then the patient can use a pine or oak tree. This exercise prescription is considered a slow and gentle Dynamic cultivation. Any Dynamic exercise for senior citizens should always be practiced slowly, and gently, in order to replenish the Yuan Qi during the Static Qigong posture (in this case, standing).

The flexion and extension of the muscles relates to the interaction of Yin and Yang within the body's energetic tissues and muscles. This flexion and extension of the muscles becomes the foundation of energy balance. Relaxation of the body results in soft, gentle movement, peacefulness and tranquillity. Muscular rigidity, on the other hand, results in hardness, restlessness and disease.

In order to practice Medical Qigong exercises, senior citizens are encouraged to keep their body's energy connected with the energy of the Earth. This connection is obtained via their energetic attachment through the bottom of their feet and the center of their palms. If the patients are confined to wheelchairs, or are bedridden, they can still focus the mind's intention on the Lower Dantian, and then extend their energy deep into the Earth.

As with all Medical Qigong prescriptions, the patients must first purge and cleanse their body before gathering, collecting, tonifying and regulating the body's energy. To purge the body of Toxic Qi, the patients imagine divine healing energy pouring down from the Heavens, filling the entire body, as they inhale. This vibrant healing energy is absorbed into their tissues and the Toxic Qi stored within the patient's body is dispersed into the Earth.

As the patients exhale, they imagine the Toxic Qi melting out their tissues and flowing downward, deep into the ground. The patients perform this purging and dispersing sequence for several breaths until they feel cleansed. Next, the patients begin to fill and regulate their body.

To fill and regulate the body, the patients imagine circulating the energy downward, deep into the Earth, and into the root system of a tree.

Next, the patients imagine the Earth's energy ascending up through the top of the tree, blending with the energy of Heaven, and then descending through their head, filling their entire body from the feet upward like water being poured into a glass and filling it up (Figure 50.1).

The inhalation and exhalation should follow the Natural Abdominal Breathing method, and the mind's intention should be focused on filling, gathering and circulating the universal and environmental Qi.

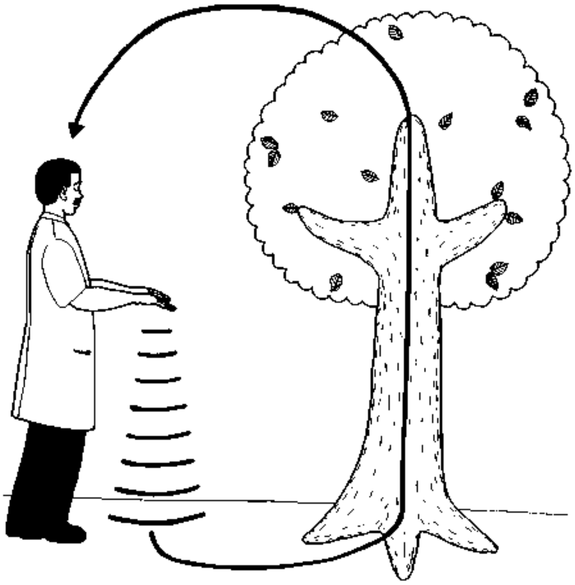

Figure 50.1. Extend Qi and Intention deep into the root system of the tree and absorb its natural energy through the Baihui point at the top of the head.

FURTHER OBSERVATIONS AND INSIGHTS ON GERIATRIC PATIENTS

According to Mark Johnson (an expert who has been treating "Seniors" with Medical Qigong Therapy and Taijiquan since 1987), any encouragement to undertake new experiences is very cathartic for older patients. Getting involved in anything creative seems to prolong the quality of their life.

Understanding the physical and emotional transitions that geriatric patients undergo in their every day life helps the Qigong doctor establish a better awareness of the patient's energetic dynamics. These physical and emotional transitions can best be understood when realizing the internal and external conflicts senior patients must constantly address.

COMPANIONSHIP

Many people tend to lose meaning in their life when they retire, lose a spouse, witness the death of countless friends, or when their children move away. Patients who lose their life purpose (or their meaning of life) usually do not live long. Elderly patients, who have lost their friends, usually have an extremely strong emotional attachment to any pet they may own. If the pet dies, this is usually as devastating as the loss of a close family member.

According to clinical observation, most older patients who become involved in spirituality, social service or voluntary work tend to be the most fulfilled and therefore most healthy. Humor and laughter as a means of internal emotional medicine for older patients cannot be overemphasized. Socializing with an understanding yet active support group also tends to bring healing to their heart.

In addressing the issue of being lonely, it has been found that the healing quality of having a pet to love can be very helpful for seniors who have become solitary or isolated.

OVERMEDICATING

Too many older patients are overly medicated. Taking up to ten pills a day is not unusual for many seniors. In a book initiated by Ralph Nader, titled "Worst Pills, Best Pills" the author emphasizes the fact that if an older person is taking over three different drugs, they are taking too many and should consult their doctor with the expressed purpose of cutting down to a maximum of only three types of medication.

The drugs given to seniors are usually tested on 30 year old individuals who are generally much stronger than most 70 year olds. Therefore half dosages are usually considered safer. In addition, there is a minimum of thirty-three drugs on the market today that can give the symptoms of Parkinson's disease. Several healers have gotten their patients off Parkinson's medication after demonstrating the fact that one of their other medications initiated the patient's symptoms.

SEDENTARY LIFE-STYLE

Too many older patients are sedentary, especially men. It is generally agreed that people living today move only 35% as much as people living 100 years ago. This has got to have a major impact on one's health, especially on older people whose circulation is not what it once was. Our bodies are designed to move; to become sedentary violates one of our basic needs. The studies conducted at Emory College confirm that the safest and most beneficial exercises for seniors is Medical Qigong, Taijiquan, walking, mild aerobics and light weight training. Seniors are encouraged to stay away from running.

IMPROPER DIET

Because of the patient's age and changing metabolic needs, diet is as equally important for seniors as exercise. Proper nutrition can be used to augment the benefits from the patient's prescriptions, as well as to increase the patient's healing potential. The diet in senior housing communities is usually atrocious. Sugar, canned vegetables and too many heavy meats are the norm. When prescribing Medical Qigong exercises for seniors, the Qigong doctor must also take into consideration the patient's diet and fluid intake, making sure that it is not sabotaging the healing effect of the Qigong prescriptions.

SLEEP DEPRIVATION

As patients gets older, their Kidney Jing becomes weak, their hair begins to turn grey, the hearing and sight becomes affected, the bone structure, mental comprehension and sleeping patterns also undergo changes. These physical transitions can and do affect the patient's emotional outlook on life. This however, is all part of the natural course of ageing.

Due to the weakening of the patient's Kidneys, sleep deprivation is more common in seniors than most people realized. More training in dealing with insomnia and changing sleeping patterns is necessary for Western doctors, as the serious impact on health by sleep deprivation is poorly understood in Western society.

Sleep is needed to invigorate and stabilize the patient's Wu Jing Shen (Five Essence Spirits). Without sleep, the patient's Hun (Spirit Soul), Shen (Spirit), Yi (Intention), Zhi (Willpower) become depleted and they lose control of their Po (Corporeal Soul). This affects the patient's intuitive perceptions, mental comprehension, drive and willpower, and can result in dementia.

UNDERSTANDING DEATH AND DYING

Classes and workshops on death and dying should be more available, and both patients as well as doctors should be encouraged to attend them. The doctor can benefit greatly by learning specific meditations being taught to facilitate this major transition, and also how to relate better to their elderly and young dying patients. The patients can benefit greatly by increasing their awareness and understanding of this process, as well as learning the numerous meditations that help resolve unprocessed emotions. Furthermore, patients can meet others with whom they can share their fears and hopes openly.

Too often the elderly do not wish to share these feelings, even with those closest to them for fear of burdening their family and friends. Patients may also be very reluctant to disclose their personal concerns for fear of being alienated, even with Qigong doctors, especially if the doctor's fear of death becomes evident through too much encouragement to get well and not lose hope.

CHAPTER 51
QI EMISSION THERAPY FOR GYNECOLOGY

INTRODUCTION

Gynecology is the study of diseases of the female reproductive organs. The earliest records of Chinese gynecological medical writings date as far back as the Shang Dynasty (1500–1000 BC). From a Traditional Chinese Medical perspective, a woman's physiology is characterized and diagnosed in accordance with her menstrual flow and vaginal discharge (also included are the energetic formations and transitions during pregnancy and childbirth). A woman's physiology is dominated by, and rooted in Blood. Her Lower Burner houses the uterus, which stores Blood, and corresponds to her Lower Dantian.

MENSTRUAL FLOW

Normally, a woman will follow the cycles of the moon and will menstruate every 28 days. Her cycles can be categorized into either new moon or full moon periods. The length of days between menstruation and the time of duration vary; four days of flow is considered average.

A woman's menstrual cycle can be categorized into four distinct time periods: Irregular Menstruation, Early Menstruation, Late Menstruation, and Prolonged Menstruation.

1. An irregular menstrual cycle is never consistent: sometimes early, sometimes late. The cause of an irregular menstrual cycle can be due to either a stagnation of Liver Qi (from suppressed anger or emotional shock), Liver Qi invading the Spleen (emotions affecting the Liver invades the Spleen), or a constitutional weakness of the Kidneys.
2. The cause of an early menstrual cycle can be due to either Heat in the patient's Blood or Qi Deficiency.
3. The cause of a late menstrual cycle can be due to either Cold stagnation of Blood, Blood Deficiency, stagnation of Qi (caused from emotions), or Deficiency of Qi.
4. The cause of a prolonged menstrual cycle can be due to either Qi Deficiency or Heat in the Blood (this is because Blood that is Hot wanders out of the vessels).

QI AND BLOOD REGULATION

Due to periodical menstruation and the complex changes that occur in the uterus and the organs (ovaries, cervix, and vagina) associated with reproduction, women tend to become especially vulnerable to Blood damage. The formation, circulation, and control of Blood depends upon the transformation, regulation, and production of the woman's Qi. This interdependence of Qi and Blood is regulated through the harmony of the body's Five Yin Organs. For example:

1. The Liver stores the Blood.
2. The Heart controls the Blood.
3. The Spleen contains or keeps the Blood in its vessels. In addition, the body's Spleen and Stomach are also known as the source of reproduction and transformation of Qi and Blood.
4. The Lungs store the Qi, which commands or moves the Blood.
5. The Kidneys store the Jing, which transforms or creates the Blood.

The woman's Sea of Blood can only flow freely when the Qi and Blood of the Five Yin Organs are in harmony. When a woman's Sea of Blood is full and exuberant, then her menstruation will be healthy and normal.

Regarding the Eight Extraordinary Vessels, a woman's Governing, Conception and Thrusting Vessels all originate within her uterus. The Governing Vessel governs the body's Sea of Yang and represents the Yang aspect of a woman's reproductive functions. The Conception Vessel governs the body's Sea of Yin and provides the Yin substances for all physiological and hormonal transitions, including puberty, conception, pregnancy, childbirth, and menopause. The Thrusting Vessels function as the Sea of the Twelve Primary Channels, and influence the supply and proper movement of Blood in the uterus, and control all aspects of menstruation.

The Belt Vessel encircles the Governing, Conception, and Thrusting Vessels, and governs, guides, and supports the energetic flow of Qi and Jing to the uterus.

Regarding the Twelve Primary Channels, their internal network is distributed throughout the uterus and is connected to both the Heart and Kidneys. In addition, the energetic connection with the uterus is further strengthened by the convergence of the Liver, Spleen, and Kidney Channels, along with the Conception Vessel, through the connecting vessels in the genitalia.

ETIOLOGY AND PATHOLOGY OF GYNECOLOGICAL DISEASES

In order for the Qigong doctor to make a correct diagnosis pertaining to a woman's menstrual flow and reproductive organ function, attention must be placed on identifying the primary etiology of the imbalance in the patient's Qi and Blood.

The Governing, Conception, Thrusting, and Belt Vessels are the most important vessels in terms of diagnosing gynecological pathophysiology. Consequently, any damage or disease of the these vessels are believed to be the main causes of gynecological disorders.

Either a direct or indirect attack to the Conception and Thrusting Vessels can cause disharmony of Qi and Blood, resulting in a loss of internal organ regulation.

DIRECT CAUSE

A direct cause of damage to the Conception and Thrusting Vessels can result from either Evil Toxin infections or from hygienically unclean sexual encounters. If the Conception and Thrusting Vessels become damaged, the result can manifest in symptoms such as profuse abnormal vaginal discharge or uterine bleeding.

INDIRECT CAUSE

Indirect causes of damage to the Conception and Thrusting Vessels can result from internal damage due to: The Seven Emotional Factors, Invasion of the Six Exogenous Factors, and An Inconsistent and Undisciplined Diet.

1. The Seven Emotional Factors tend to be the major factors in determining the root causes in most gynecological diseases.

   The suppression of anger, rage, resentment, and jealousy often results in Liver Qi depression, which eventually leads to Blood stagnation. This can result in such symptoms as delayed, painful, or blocked menstruation. If a hyperactive Liver Yang condition occurs, the patient can develop symptoms as serious as pre-eclampsia or eclampsia (coma and convulsive seizures that can occur between the 20th week of pregnancy and first week postpartum; symptoms include edema of the legs and feet, puffiness of the face, hypertension, severe headaches, dizziness, epigastric pain, nausea, sudden convulsive seizures and coma). The Liver's Blood Heat can lead to profuse uterine bleeding, leaking of Blood, hemoptysis (throwing up Blood) or epistaxis (bleeding of the nose).

   The suppression of worry and regret, or obsessiveness, can harm the Heart and Spleen. This can lead to amenorrhea (the absence or suppression of menstruation) or a fetal leakage (after conception, if a small amount of bloody fluid discharges from a woman's vagina, it is known as Tai Lou or fetal leakage).

   The suppression of fright, fear, loneliness, and insecurity can damage the Kidneys and lead to miscarriages.
2. Invasion of the Six Exogenous Factors, in particular, Cold, Heat, and Dampness, are the primary causes for gynecological diseases.

   If the Blood is invaded by Cold, it congeals, causing delayed menstruation, painful menstruation, or severe contractions and thick conglomerations of Blood.

   If the Blood is invaded by Heat, it moves, causing the Blood to flow recklessly. This reckless movement of Blood leads to early menstruation, excessive menstruation, and uterine bleeding.

If the Blood is invaded by Dampness (Damp Evils invading the Spleen), it can lead to abnormal vaginal discharge, uterine bleeding, and irregular menstruation.

3. An Inconsistent and Undisciplined Diet can lead to a Spleen and Stomach disharmony. If the Spleen and Stomach's ability to transform and transport food becomes hindered or weakened, the Sea of Blood becomes Empty. This can lead to delayed menstruation or scanty menstruation.

   If the Spleen loses its ability to restrain and contain the Blood, both Blood and Qi leak out of the vessels. This leaking of Qi and Blood can lead to excessive menstruation and uterine bleeding.

   If the Spleen's Yang does not perform its transporting function, Dampness and Turbidity will descend, resulting in abnormal vaginal discharge.

VAGINAL DISCHARGE

In Traditional Chinese Medicine, abnormal vaginal discharge (Dai Xia) is considered a disease of the Belt Vessel. If the Belt Vessel loses its restraint on the other vessels, then abnormal vaginal discharge is produced. This is usually caused by a Spleen Deficiency, Dampness, and descending Turbid Qi.

CONSISTENCY

The consistency of the patient's vaginal discharge can be divided into three categories: Turbid, Watery, and Clotting (thick and pasty).

1. A Turbid discharge refers to Blood Heat and an infection in the uterus. It also refers to signs of Cold stagnation.
2. A Watery discharge is considered a Cold Damp and/or Deficient condition.
3. Clots (thick and pasty) are considered a Damp Heat and Excess condition.

SMELL

The smell of the patient's vaginal discharge can be divided into conditions of either Damp Heat, indicated by a leathery smell, or Cold Dampness, indicated by a fishy smell. Also, if the patient's menstrual Blood has a strong, foul smell, it usually indicates a Heat condition.

COLOR

Vaginal discharge color is divided into five colors (Wu Se Dai): white, yellow, red, green-blue, and dark brown or black. The pathology of the discharge color is mostly caused from a Deficient Conception Vessel, loss of restraint from the Belt Vessel, Spleen Dampness, or Evil Wind invading the uterine vessels. These conditions create Heat, which descends into the uterine area causing disease.

Only profuse vaginal discharge, which bears an unnatural color and gives off a malignant odor, is considered pathological. The five colors of vaginal discharge are as follows.

1. White vaginal discharge (Bai Dai) is the most commonly seen type of vaginal discharge. It is a Cold condition caused by a Spleen or Kidney Yang Deficiency, or from Exterior Cold Dampness. It manifests as a white, sticky, slimy fluid discharge from inside the patient's vagina with a fishy smell. Symptoms include sore lower back, weakness, and irregular menstruation.
2. Red vaginal discharge (Chi Dai) is caused from Damp Heat in the Lower Burner or Fire Blazing through the Heart and Liver channels. The latter eventually causes a Yin Blood Deficiency. Red vaginal discharge manifests as thick and sticky, white with red mixed together, pale red, or red vaginal discharge with a foul odor. Symptoms include thirst, bitter taste in the mouth, and reddish urine. Continued red vaginal discharge may be a sign of cancer, and the patient should immediately be referred out to an oncologist.
3. Yellow vaginal discharge (Huang Dai) is caused from Heat, usually Damp Heat in the Lower Burner produced from a Deficient Spleen not transporting the Dampness. This condition causes Water Dampness to brew and bind, which then transforms into Damp Heat. The Damp Heat invades the Conception Vessel which causes the Belt Vessel to lose its ability to restrain. It manifests as a yellow, sticky, and

slimy vaginal discharge with a foul odor. A vaginal discharge that has yellow and red with white pus (after menopause) is considered a manifestation of Toxic Damp Heat in the uterus and can indicate endometrium cancer.

4. Green-blue vaginal discharge (Qing Dai) is caused by Damp Heat in the Liver or Gall Bladder Channels. A yellow-green color with pus accompanied by itching and soreness is also a sign of Damp Toxic invasion.
5. Dark brown or black vaginal discharge (Hei Xia) is seen as a form of Blood stagnation.

All gynecological disorders are closely related to the patient's whole constitution, even though the manifestations appear in the reproductive organs. The patient's general constitution, age, dietary habits, home and work environment, should all be taken into account when diagnosing clinical pathophysiologies.

MENSTRUAL COMPLICATIONS

DYSMENORRHEA

Dysmenorrhea is one of the most frequent gynecologic disorders observed in the clinic today. It is described as pain associated with menstruation and refers to cyclic pain (before, during or after menstruation) felt in the lower abdomen or lumbosacral area of the woman's body before, during, or after menstruation. Serious cases of dysmenorrhea can be accompanied by headache, nausea, vomiting, and even syncope.

ETIOLOGY

From a Traditional Chinese Medical perspective, a woman's body is more susceptible to an attack of external pathogens and injury from internal emotional disturbances during her menstrual period. Mental depression may injure the Liver and cause Qi and Blood stagnation.

The invasion of Cold into the Thrusting and Conception Vessels may also cause stagnation and coagulation of Blood, which in turn causes Qi and Blood Deficiency (leading to poor nourishment of the uterus). Excessive exposure to Cold and Dampness can cause Cold to invade a woman's uterus, causing stasis of Blood. A women is prone to Cold invasion when her uterus and Blood are in a weakened state, during and soon after her period.

There are four different phases identified during a woman's menstrual cycle: Before Menstruation, During Menstruation, After Menstruation, and During Mid-cycle.

1. Before menstruation, the woman's Yang rises in preparation to move Blood during her menses; Liver Qi also moves into the lower abdomen in preparation to move the Blood. If a woman's Liver Qi stagnates, it can cause the Blood stagnation in the uterus, leading to pain, especially before her period. This type of stagnation is the most important pathological condition causing dysmenorrhea. Liver Qi stagnation can be caused from emotional strain resulting from suppressed emotions, such as anger, resentment, hatred, and frustration.
2. During menstruation, the woman's Blood is moving. The movement of Blood relies on Liver Qi and Liver Blood. If a woman's Liver Blood stagnates, it will cause pain during her period.
3. After menstruation, the woman's Blood and Yin are Empty.
4. During mid-cycle, the woman's Blood and Yin gradually fill up the Thrusting and Conception Vessels.

According to Western medicine, dysmenorrhea is related to mental stress, emotional instability, mental depression, and fright. Dysmenorrhea can also be due to diseases of the ovaries, uterine, or endocrine glands.

The menstrual pain may appear one to two days before the onset of the menstrual period, and remain until the start of the menstrual flow. It usually lasts from just a few hours to one to two days.

SYMPTOMS

In Traditional Chinese Medicine, dysmenorrhea is classified into two types of syndromes: the Excess type and the Deficient type.

- The Excess type of dysmenorrhea can be caused by either Qi and Blood stagnation, Cold stagnation, or Damp Heat. An Excess condition manifests as distention and pain in a woman's lower abdomen and / or lumbago, occurring before or during menstruation. Symptoms also include obstructed menstruation with pale-purplish color and clots.

- The Deficient type dysmenorrhoea can be caused by either Qi and Blood Deficiency or Kidney and Liver Yin Deficiency. Symptoms of this condition manifest as pain in the lower abdomen, occurring before or during menstruation, which can be relieved by heat and compression. The patient will experience soreness and distention of the waist and legs with pale and scanty menses.

TREATMENT

1. To begin, have the patient sit at the edge of the table and begin stimulating the Yellow Court, Mingmen, and Lower Dantian areas, using the Vibrating Palm technique, for 6 to 8 breaths.
2. Emit Qi, using the Extended Fan Palm technique, into the patient's Lower Dantian, purging the Qi down the Stomach Channels to her feet.

TREATMENT MODIFICATIONS

- When treating patients with the Excess type of dysmenorrhea, emit Qi into the patient's Lower Dantian, St-21, and Lv-14 points, then purge the Stomach and Liver Channels down to and out of the feet.
- When treating patients with the Deficient type of dysmenorrhea, emit Qi into the patient's Lower Dantian, then guide Qi along the Stomach and Spleen Channels to regulate the body.

PRESCRIPTIONS AND HOMEWORK

Have the patient practice the Descending the Yang and Ascending the Yin techniques (see Chapter 41).

PRESCRIPTION MODIFICATIONS

1. For patients with the Excess type of dysmenorrhea:
    a. Have the patient sit at the edge of a chair, relax and breath naturally, with the tongue placed up against the hard palate. The patient's mind should concentrate on her Lower Dantian and navel area.
    b. The patient should imagine the navel as the center of a wheel. When inhaling, she should move the Qi upwards, rotating the energy clockwise, from the lower portion of the right side of the abdomen. The patient silently chants the phrase: "The white tiger hides in the east." Exhaling, she rotates the Qi from above the navel to the left side of the abdomen back to the beginning point, while silently chanting the phrase: "The blue dragon shelters in the west." The patient continues circulating the Qi around the navel, rotating it in a clockwise direction for 36 breaths. In this particular meditation, it is important the patient start with small circles and gradually increase the circumference of the energy circle, in order to disperse the accumulated Heat in the uterus.
2. For patient's with a Deficient type of dysmenorrhea:
    a. Have the patient sit at the edge of a chair, relax and breath naturally, with her tongue placed up against the hard palate. The patient's mind should concentrate on the Lower Dantian and navel area.
    b. Imagine the navel as the center of a wheel. When inhaling, she should move the Qi upwards, rotating the energy counterclockwise, from the lower portion of the left side of the abdomen. The patient silently chants the phrase: "The blue dragon shelters in the west." Exhaling, she rotates the Qi from above the navel to the right side of the abdomen back to the beginning point, silently chanting the phrase: "The white tiger hides in the east." She continues circulating Qi around the navel, rotating from large to small circles in a counterclockwise direction for 36 breaths. In this particular meditation, it is important that the patient start with large circles and gradually decrease the en-

ergy circle's circumference in order to gather energy into the uterus.

MENOXENIA

Menoxenia refers to the pathological changes of menstruation occurring in the cycles, such as the color, quantity, and quality. This includes extended or delayed menstrual cycles, as well as menorrhagia (excess bleeding) and scanty menstruation.

SYMPTOMS

1. Blood Heat symptoms include: prolonged menstrual cycle, profuse menstruation of bright red or purplish color, viscous in quality, and a heavy feeling in the chest.
2. Deficiency of Qi symptoms include: prolonged menstrual cycle, profuse, pale and watery menstruation, listlessness, palpitations, shortness of breath, and an empty or dropping sensation in the lower abdomen.
3. Blood Stasis symptoms include: delayed menstrual cycle, purplish and massed menstruation with pain and distention in the lower abdomen.

TREATMENT

1. Have the patient sit at the edge of the table and begin to stimulate her Yellow Court, Lower Dantian, and Mingmen areas using the Vibrating Palm, for 8 to 16 breaths.
2. Use the Extended Fan Palm hand posture to emit Qi into the patient's Mingmen and Lower Dantian using the Spiral Rotating method of Qi manipulation. Rotate the energy in a counterclockwise direction for conditions of Excess and a clockwise direction for Deficient conditions.

TREATMENT MODIFICATIONS

- When treating a patient with Blood Heat, emit Qi into her Yellow Court and Lv-14 points, drawing the energy down her Liver Channels and out of her feet.
- When treating patients with a Deficiency of Qi, use a Extended Fan Palm hand method while tonifying and filling her Mingmen and Lower Dantian areas.
- When treating a patient with Blood Stasis, use the Extended Fan Palm hand method to emit Qi into her Lower Dantian and rotate the Qi in a clockwise direction.

PRESCRIPTIONS AND HOMEWORK

In order to move the energy and allow Qi and Blood to flow smoothly in the uterus, have the patient practice the Microcosmic Orbit Fire Cycle (see Chapter 41).

PRESCRIPTION MODIFICATIONS

- For patients with Blood Heat, it is advisable to have them practice the Liver Massage method (see Chapter 39) and sound the word "Xu" (Shu).
- For patients with Qi Deficiency, have the patients perform Natural Breathing meditations, as well as the Daoist Five Yin and Yang Organ Exercise (see Chapter 42).
- For patients with Blood Stasis, have the patient practice Circling the Abdomen, for 36 breaths in a clockwise direction (see Chapter 41).

CYSTS AND TUMORS

BREAST CYSTS AND BENIGN TUMORS

A breast cyst or benign tumor is a lump or nodule found within the breast tissue.

ETIOLOGY

The fibrocystic changes that can occur in a woman's breast tissue are commonly divided into three levels of benign breast disorders. Beginning with Level-1 breast cysts (which include the development of benign lumps, nodules or fibroadenoma), progressing onto the development of Level-2 breast cysts (which include the development of mastitis or carbuncles), and ending with Level-3 breast tumors.

The Qigong doctor's treatment will vary according to the type and condition of the lump enclosed within the patient's breast tissue.

DIAGNOSIS ACCORDING TO DIFFERENT TYPES OF BREAST LUMPS

The Qigong doctor differentiates among the types of breast lumps and initiates treatment accordingly. Breast lumps can be divided into three major categories: Phlegm, Blood Stasis, and Toxic Heat.

1. Phlegm lumps are soft with distinct edges, mobile and slippery. They are not swollen, and the skin surface around the area is white.
2. Blood Stasis lumps are hard, immobile nodules. The skin surface around the surrounding area is purple.
3. Toxic Heat lumps are hard, immobile, with indistinct edges. The skin surface of the surrounding area is indented and red.

LEVEL-1 BREAST CYSTS (BENIGN LUMPS, NODULES OR FIBROADENOMA)

These are lumps that feel like hard nodules; they are usually painless, round, flat, and mobile. The lumps do not feel hot or cold to touch, the skin color is unchanged, and the size of the lump may vary depending on the patient's mood changes.

ETIOLOGY

Level-1 cysts are usually caused by the patient's Liver and Spleen being affected by anger, worry, and depression. The Spleen Qi is weakened, thus disrupting the free-flow of the Liver Qi. This leads to conditions of nodules due to both Qi and Phlegm stagnation. Long term Qi stagnation eventually leads to the accumulation of Phlegm and Stasis of Blood.

SYMPTOMS

These type of cysts are characterized by their process of development; they generally start off small and increase in size after several months.

TREATMENT

Eliminate the stagnation and remove the Phlegm using Qi purging and dispersing techniques.

LEVEL-2 BREAST CYSTS (MASTITIS OR CARBUNCLE)

This condition is a painful inflammation of the breast skin and underlying tissue.

ETIOLOGY

Level-2 cysts are usually caused by Qi and Blood stagnation, resulting from the stagnation of the Liver or Gall Bladder Qi, which is accompanied by excessive stagnation of Toxic Heat in the Stomach Channels. Breast cysts can be internally induced by suppressing the emotions, especially grief. Generally mother issues relate to the left breast and Lung area, while father issues relate to the right breast and right Lung area.

TREATMENT

1. After clearing the diseased area of Toxic Qi, the Qigong doctor uses the Sword Fingers technique to stimulate CV-17, St-18, Lu-1, Bl-18, and -21 points, to open the points along the Liver, Lung, and Stomach Channels.
2. Next, the Qigong doctor emits Qi, using the Vibrating and Shaking techniques, to disperse the cyst.
3. The doctor then applies the Dragon Mouth manipulation to dredge the patient's body, by pushing, pulling, and leading the Stagnant and Toxic Qi down the patient's Lung Channels (out the arms), and Stomach Channels (out her legs), via the St-36 points.
4. The patient is given homework, consisting of Jing Point Therapy, and of kneading the surface and surrounding tissue of the cyst at least once a day. This prescription should also include the use of the "Shang" sound tone resonation to further dissolve the cyst, and to dispel the Toxic Qi out of the breasts.

LEVEL-3 BREAST TUMORS

This type of breast cyst corresponds to a severe condition of dysplasia.

ETIOLOGY

Level-3 cysts are usually caused by Liver Fire, accompanied by Heart and Lung Qi stagnation due to the accumulation of grief, sorrow, exasperation, and being upset, or due to the derangement of the patient's Thrusting and Conception Vessels.

TREATMENT

1. After clearing the diseased area of pathogenic Qi, the Qigong doctor uses the Thunder Palm technique to disperse the tumor.
2. After several minutes of blasting the area, the Qigong doctor uses the Spiraling and Circling Energy techniques to circle the tissue areas

surrounding the tumor, drawing the pathogenic Qi down and out of the patient's nearest arm via the Lung Channel (Figure 51.1).

3. The Qigong doctor then uses energetic Pulling and Shaking techniques to release the pathogenic Qi from the patient's body and to draw it out of the arm. The doctor must then pull and snap each of the patient's fingers and thumbs, to release and disperse the Excess Heat and pathogenic Qi.
4. Next, circle rotate the energy within the patient's Lower Dantian, pulling the Qi down the lower legs and out the St-36 points.
5. Finally, fill her Lower Dantian and circle regulate her Governing and Conception Vessels using the Microcosmic Orbit to balance the Yin and Yang energy.
6. Give the patient homework consisting of Jing point therapy that stimulates the GB-21, St-36, CV-17, and Sp-6 points, once a day. If the patient is using magnetic therapy, she should place the magnets on the points for 1 hour. This prescription should also include herbs, and the use of the "Shang" sound tone resonation (straight tone and descending/ascending tone prescriptions) to further dissolve the tumor.

OVERVIEW OF TREATMENT

The general treatment for breast diseases, depending on the severity of the condition, is divided into five main components: Medical Qigong therapy, Medical Qigong massage, prescriptions, Jing point therapy, and herbal teas.

MEDICAL QIGONG THERAPY

The doctor's emitted energy can assist the patient in dispersing stagnations, as well as dissolving breast tumors and cysts. Generally, the doctor's treatment lasts for a period of three days, and can be indefinitely prolonged and increased in effectiveness if the patient is given additional Medical Qigong prescriptions as homework.

MEDICAL QIGONG MASSAGE PRESCRIPTION

This therapy is most important as it allows the patient to reconnect with her tissues and further allows her to feel and release emotions trapped

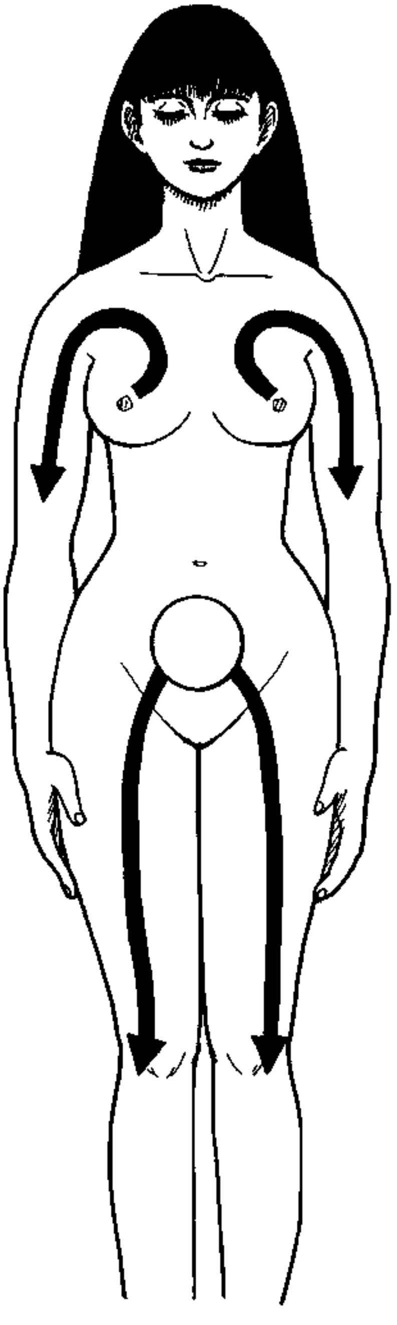

Figure 51.1. After purging the Toxic Qi from the patient's upper body, the Qigong doctor continues to disperse the patient's Toxic Qi from the lower abdomen, leading it down the patient's leg and out of the St-36 points.

within her body (remember, you can only heal what you feel). This healing process is as follows.

1. The patient begins by rubbing her hands together to get them hot.
2. Placing her hands on her breast, the patient inhales and draws the heat from her hands into the breast tissue, while imagining divine healing light flowing from the Heavens, and pouring into her body to fill up her chest.
3. The patient circle massages her breasts 36 times towards the outside of her body, concentrating on the breast tissue filling up with the divine healing energy that destroys any and all pathogens stored within her breast tissues.
4. After completing 36 times in one direction, the patient reverses directions but maintains the same visualization.
5. Next, the patient exhales the straight healing sound "Shang" for 9 breaths, toning, while turning the body from the left to the right side. The patient focuses on purging the pathogens out of her breast area through her mouth. This purges her Lungs and the surrounding breast tissues.
6. The patient then exhales the descending/ascending healing sound "Shang" for 9 breaths as she turns at the hips, toning from the right to the left side of her body, and focuses on purging the pathogens out of her breast area. This purges her Lung Channels and the surrounding energetic connections to the breast area.
7. The patient then rubs 36 times in a clockwise direction, from her abdomen to her Lower Dantian, leading the Qi downwards. This is performed to root the energy back into her Lower Dantian.
8. The patient ends the prescription in a quiescent state of mind.

MEDICAL QIGONG PRESCRIPTIONS

This therapy allows the patient to purge the organs that are found to be the root cause of the disease. In Medical Qigong, the doctor views the patient's cysts or tumors as a flower. If, through surgery, the top of the flower is plucked from its stem (and the branch, stem, and root cause of the disease is left still intact), the tumor or cyst will return again. In order to heal the patient's condition, the doctor must also address the original cause of the disease and destroy the entire support system. The purpose for these prescriptions is to destroy the disease's entire root system to kill the flower.

1. The Dry Crying prescription is given to purge the emotions of grief and sorrow from the patient's Lungs (see Chapter 53).
2. The "Guo" healing sound is given to purge the patient's Liver Fire.
3. The "Shang" healing sound is given to dissolve the cyst or tumor.
4. Prescribe Jing Point Therapy. Specific points are prescribed for the patient to enhance the doctor's treatments, and increase her own healing potential. The patient is encouraged to stimulate specific areas of her body, by emitting Qi while pressing and focusing her intention into the following points:
    a. The Small Intestine 3 point affects the patient's breast disease caused by Rebellious Qi and Qi depression.
    b. The Stomach 18 point frees Qi flow locally in the patient's breast.
    c. The Stomach Channel transverses the breast, and the Stomach 36 point regulates Qi of the patient's entire body.
    d. The Spleen 6 point is specifically chosen for treating breast cysts.
    e. The Gall Bladder 21 point (Shoulder Tapping) is prescribed to assist the patient in sinking Liver Qi Heat.

HERBAL TEAS

All herbal teas are administered to move the Qi and assist the doctor's treatment. Dandelion tea, for example, is mixed with green tea to clear Heat from the patient's body; or green orange peel is added to help move the Qi.

OVARIAN CYSTS AND UTERINE TUMORS

In an ovarian cyst, there is a closed sack or pouch containing fluid, semifluid, or solid material, usually characterized by Phlegm, Damp Heat, and Stasis of Blood.

ETIOLOGY OF OVARIAN CYSTS

Ovarian cysts can be caused by psychological upheavals which lead to enlarged follicles that fail to rupture, as well as by Liver Fire stagnation. Ovarian cysts are usually related to the anger from the Liver turning inward and attacking the body.

ETIOLOGY OF UTERINE TUMOR

A uterine tumor, or myoma, is a solid benign growth in the myometrium, and is often called a fibroid. It contains muscle tissue. Excessive uterine bleeding is a common effect of myomas. Ovarian cysts and uterine tumors are generally divided into three types of myomas:

1. The subserous type is located on the outside wall of the uterus;
2. The interstitial type is located within the wall of the uterus;
3. The submucous type is located on the inside of the uterus.

TREATMENT

The same treatment is used for both ovarian cysts and uterine tumors. To treat these conditions, the doctor uses Medical Qigong therapy in conjunction with Jing point therapy and herbs.

1. Begin with the patient in a supine position and purge and clean the diseased area of pathogenic Qi, using the Thunder Palm technique in conjunction with the Vibrating Palm, to disperse the stagnation in the ovaries or uterus.
2. Disperse pathogenic Qi out the patient's body via the Gall Bladder Channels.
3. Tonify the patient's Lower Dantian with Qi and circulate the energy through the Microcosmic Orbit.
4. Press and stimulate with Jing point therapy the Sp-6, Sp-9, St-36, GB-34, and GB-38 points (Figure 51.2).

PRESCRIPTIONS AND HOMEWORK

The patient should be given homework, in accordance with her constitution. The prescriptions should include:

1. The "Guo" sound to disperse Liver Fire.
2. The "Yu" sound to disperse ovarian cysts.
3. Prescribe Lower Dantian regulation exercises to access and regulate the ovarian energy.
4. Prescribe herbs for chronic conditions.

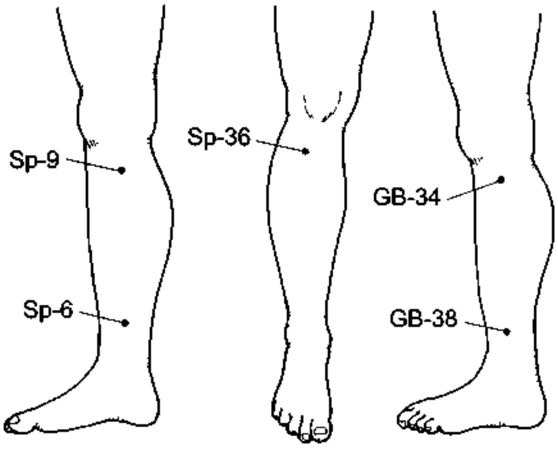

Figure 51.2. When treating ovarian cysts, use Jing Point therapy in conjunction with Qi emission, to stimulate Sp-6,-9, St-36, G.B.-34 and 38 points.

UTERINE COMPLICATIONS

MISCARRIAGE

A miscarriage is the termination of a pregnancy, at any time before the fetus has attained extrauterine viability. Miscarriage occur usually between conception and the fourth month of pregnancy.

In Traditional Chinese Medicine, a miscarriage can have several stages and levels of severity, ranging from such conditions as Fetal Leakage, Restless Stirring of the Fetus, Falling Fetus, and Small Birth. Within each of these conditions there are several patterns of disharmony.

ETIOLOGY

There is an ancient Chinese saying which states "A miscarriage is more serious than child birth." According to Traditional Chinese Medicine, there five basic mechanisms associated with the cause of a miscarriage: Empty Conception and Thrusting Vessels, Blood Deficiency, Depressed Liver Fire, Blood Stasis, and External Injuries.

1. An Empty condition of the Conception and Thrusting Vessels causes miscarriages. The mother's Kidney Qi consolidates the fetus and is responsible for constricting the anal and vaginal orifices. If the Conception and Thrust-

ing Vessels become empty, they cannot consolidate and secure the fetus within the womb. If any of the following three conditions occur singly or in concert, both the Conception and Thrusting Vessels may indeed become Empty: (1) the mother's Yin Essence is depleted or consumed; (2) the Mingmen Fire is insufficient, and (3) both the Yin and Yang Qi are Deficient. Any of these three conditions may be due to: congenital weakness and insufficiency, chronic disease, extreme or prolonged stress, unrestrained sex after conception, or the mother's age (over 40).

2. Blood Deficiency from an irregular diet, excessive worry and anxiety may deplete the Spleen and Stomach organs resulting in a miscarriage. Since the Spleen is the postnatal root for Blood generation and transformation, an Empty or Deficient condition of Spleen Qi may cause Blood Deficiency. Furthermore, as the Spleen Qi restrains and supports the mother's abdomen, a Spleen Qi Deficiency, or Empty Spleen Qi, may lead to the prolapse of the Central Qi and produce insufficient Blood to nourish the fetus. Both the Central Qi prolapse and the Blood Deficiency can present the Conception and Thrusting Vessels from consolidating the fetus and securing it in the womb.
3. Depressed Liver Fire due to suppressed emotional frustration (with mental agitation, anger, jealousy, and envy) may cause a miscarriage. The excessive Heat generated by the depressed Liver Fire may in turn attack the mother's Sea of Blood, scorching and injuring the Thrusting (Chong) Vessel. This injury, in turn, may cause the mother's Yin Blood to flow recklessly downward, leaving no place for the fetus to rest.
4. Blood Stasis may be due to three different factors: past internal injuries, iatrogenic disorders, and long term Qi stagnation. The excessive accumulation of Qi and Blood within the mother's womb during conception can sometimes aggravate a preexisting condition, and thereby cause hemorrhaging, as well as initiate energetic patterns which fails to stop her from bleeding.
5. External Injuries from accidents (sustaining a sprain or strain in the abdominal region), or battery (being hit or struck) can lead the body into a state of shock. This shock to the mother's body may damage either the Conception or Thrusting Vessels, preventing them from securing the fetus within the womb.

TREATMENT

In the Medical Qigong clinics in China, doctors are cautioned to be extremely careful when treating pregnant patients. Qi emission can cause a spontaneous miscarriage if done incorrectly. It is therefore recommended that only Medical Qigong doctors, who have specialized in the specific treatment of pregnant women, be allowed to diagnose and treat them. The interdependencies of the mother's and the fetus' energy circulation is too complex and delicate for a generalized practitioner to handle.

Treatments for miscarriage vary, based on pattern discriminations of Kidney Deficiency, Qi and Blood Deficiency, Blood Heat or physical trauma. An example of an ancient Chinese prescription requires the doctor to treat Yingu Kd-10 for miscarriages with vaginal bleeding, dark urine, abdominal distention, and fullness with an inability to lie down.

If the patient has already experienced a miscarriage, and is in the state of recovery, specific treatments vary according to the doctor's diagnosis concerning the etiology and pathology of the patient's condition.

HOMEWORK AND PRESCRIPTIONS

Each consecutive miscarriage causes further injury to the patient's Kidneys, creating or aggravating the stagnant Blood condition in the uterus. Doctors of Traditional Chinese Medicine therefore encourage their patients to take herbal medicines, and to practice Medical Qigong for six months before attempting to conceive again.

When they do become pregnant again, they are encouraged to continue to practice Medical Qigong meditations and exercises for strengthening and balancing their body's energetic fields and internal organs.

A patient recovering from a miscarriage is given specific prescriptions according to her emotional condition, organ dysfunction, and the etiology of the condition.

UTERINE COMPLICATIONS DUE TO ABORTIONS

Another important aspect of Medical Qigong and gynecology is helping a patient face the loss, or termination of a pregnancy. This can occur either intentionally (through an abortion), or unintentionally through a miscarriage, spontaneous abortion, or the death of a fetus by other causes.

When a woman has undergone an elective abortion, the Qigong doctor must not only consider the potential for scar tissue formation and physical complications within her uterus, but also the energetic and emotional complications resulting from the abortion.

While treating the patient, the doctor may discover within the uterus what appears to be the energetic remains of the aborted embryo. These phantom embryos may in fact be thought forms created by the mother's feelings of guilt and shame at having undergone an abortion. These difficult emotions are further compounded by deeper feelings of self-hate, that go unnoticed by others.

Most people feel little compassion for women who suffer a miscarriage, and discourage them from grieving. They are frequently advised by well meaning friends to believe that the miscarriage was all to the good, preventing a possible deformed or sickly baby from being born.

If women who miscarry are discouraged from grieving, those who make the painful decision to abort often face scorn and disbelief when they grieve their loss. Silently they mourn as best they can, haunted by feelings of shame imposed upon them by others and their own conscience.

Most women struggle long and hard before resorting to abortion. Afterwards, they often obsess about the living embryo, and what it felt like to be pregnant, centering all their attention on the uterus. Such intense, emotional concentration creates phantom embryos that, in fact, take on the form of ghostly embryos. These thought forms or energetic imprints are clearly visible to the doctor's inner vision. The energetic imprint contains the location, exact age and size at the time of the embryo's termination. The unprocessed energetic charge of the embryonic ghost's location creates an energetic vortex. This vortex, in turn, draws emotions such as rage, anger, guilt, and shame into itself. If this area is not purged, the resulting accumulation of negative emotions will develop into cysts, tumors, and possibly cancer.

This energetic reaction to surgical abortions has been confirmed by several Western doctors (all wishing to remain anonymous). These doctors observed that after many years of performing abortions, within a very short time period several patients (whose ages ranged from 16 to the mid 30's) developed reoccurring cysts and tumors. The location of the cysts or tumors in the majority of the patients uteri was at the exact tissue area of the embryo.

HEALING THE ENERGETIC TRAUMA

When the patient has decided to heal from her loss, she must not only address her physical condition (scar tissue formation, etc.), but also heal from the spiritual and emotional woundings as well. To assist the patient to heal emotionally and disperse the energetic charges, the doctor must help the patient transform the energy stored within her Wu Jing Shen, which is responsible for maintaining the toxic energetic pattern. Only after experiencing and releasing the emotions stored within the Wu Jing Shen (see Chapter 13) is the patient ready to heal. The healing process is divided into three distinct steps:

1. The patient is encouraged to completely forgive herself for losing the child and to reconnect with her higher power.
2. The patient is encouraged to completely forgive herself for being in the situation, and assume the responsibility of the pregnancy and the loss.
3. The patient is encouraged to completely forgive the person responsible for making her pregnant.

It is important for the patient to completely feel and experience all three stages of forgiveness in order to begin healing these energetic wounds.

Withholding the smallest bit of guilt or anger will seek to sabotage the healing process, and will also allow the toxic energetic charge to remain within the patient's tissues. Even the smallest bit of anger or guilt that remains will continue to recreate a destructive energetic cycle, storing and retaining the patient's toxic emotions.

RELEASING THE EMBRYO'S ETERNAL SOUL

The purpose and goal of releasing the embryonic thought form is to free the patient from the energetic remnants of the phantom embryo. The patient is encouraged to make peace with herself and with the energetic form of her unborn child, as well as the divine (God or higher spiritual force incorporated in the patient's own belief system). The patient does this through deep, meditative prayer, positive affirmation, and an attitude of honor and respect for the relationship between herself and the embryo's energetic remains. Even though the Eternal Soul does not remain present during the abortion, the mother may still need to atone for her decision not to bring the baby to term. This atonement may be necessary to fully release all feelings of guilt in accordance with the patient's religious belief system.

The patient must also make peace with the embryo's Eternal Soul, by enveloping the phantom embryo in a circle of divine healing light and love, to return it back to the divine. In order to achieve this, it is helpful for the patient to visualize creating a small energetic bubble, full of love and healing light, to house the departed embryo's energetic remains before returning it back to the divine. This serves to complete the emotional closure. Once closure is established, the energy of the embryo's energetic remains will leave her body, allowing the patient to feel a fulfilling sense of release, completion, calmness, and inner-peace.

UNDERSTANDING ENERGETIC ABORTIONS

The gift of life is an incredible miracle. If a difficult decision is made to not keep a pregnancy, however, then the patient should consider contacting the embryo's Eternal Soul. Through prayer, meditation, imagination, and visualization, the mother and father can connect to the developing embryo and request that the embryo's Eternal Soul return back to the light of the divine. Sometimes called an energetic abortion, this energetic transformation can take place within the first developing stages of the mother's pregnancy.

Since the Eternal Soul is ageless, it is possible to carry on a mind to mind conversation. When a baby is born, its brain must have time to develop before it can grasp complex information; the mind itself must have time to direct the brain's growth and the body's maturation. Once the Eternal Soul is freed from the body, however, it is not confined by the physical limitations of the infantile brain and should be spoken to as an equal. This method is used by psychotherapists on the leading edge today, using hypnosis to facilitate the patient's connection to the Eternal Soul through visualization.

Patients who have undergone age regression psychotherapy to the state of conception and embryonic state often describe visiting the mother's womb during the time of conception, then leaving and returning at various times during the embryonic stage of development. The Eternal Soul's residence does not appear to be permanent until the end of the first trimester (see Chapter 2). Until then, the Eternal Soul may come and go as it chooses, staying, at times, just long enough to insure the continued growth and development of its physical body.

It may be for this reason, that talking mind to mind with the embryo's Eternal Soul to request its departure, is often successful in precipitating a miscarriage. As the Eternal Soul leaves the embryo's body (sometimes observed as a small ball of white light energy), all that remains in the mother's womb is an empty, lifeless mass of tissue. Usually the mother's body will naturally abort the lifeless tissue mass, either immediately, or by the next morning.

CHAPTER 52

QI EMISSION THERAPY FOR NEUROLOGY (STROKE, COMA, PARALYSIS, AND MULTIPLE SCLEROSIS)

INTRODUCTION

Neurology is a branch of medicine that deals with the nervous system and its diseases, as well as neuropathology. Neuropathology is the study of diseases of the nervous system and the structural and functional changes occurring within them. Neurological diseases are divided into either congenital predispositions that develop over time into full-blown disorders, or those in which acquired destructive pathogenic factors act upon the brain. The latter are usually caused by toxic, traumatic, mechanical, inflammatory factors or neoplastic types of diseases (abnormal tissue formation).

The following is a description of several treatments and prescriptions used in various Qigong clinics and hospitals throughout China. Please be aware that specific treatments will vary according to each patient's constitution and the nature of his or her diseases.

STROKE (WIND-STROKE)

A stroke is a sudden loss of consciousness followed by paralysis, and may be caused by one of several different factors, including a brain hemorrhage, the formation of an embolus or thrombus in an artery, or the rupture of an extracerebral artery causing subarachnoid hemorrhage.

ETIOLOGY

Although a stroke occurs suddenly, in actuality, the causal factors build up for several years. A stroke can be caused from high blood pressure, diabetes, high cholesterol, chronic dehydration, overwork, emotional stress, irregular diet, excessive sexual activity and severe migraines. Strokes usually occur at night, or early in the morning (the same is true regarding heart attacks).

There are four pathogenic factors involved in causing stroke. These factors are: Wind, Phlegm, Fire and Stasis. The Liver is dependent on the Kidney Water (Yin Qi) to moisten and soften it. If, however, the Kidney Yin becomes Deficient, this may cause the Liver Yang energy to grow, and to generate Fire. Excess Liver Fire may stir up Internal Wind which causes the Qi and Blood to rebel upwards. Phlegm is then formed, which obstructs the cavities and vessels, causing the patient to lose consciousness. This can further develop into Penetrating Wind (i.e., stroke). Internal Wind may also develop from Heat and Phlegm caused by the patient eating too many sweet and fatty foods (irregular diet), overwork, emotional stress, excessive sexual activity, and inadequate rest.

In the clinic, conditions referred to as stroke (Wind-Stroke) include cerebral hemorrhage, cerebral thrombosis, cerebral embolism, and cerebrovascular spasm. These four conditions are termed in Western Medicine as "Cerebro-Vascular Accident" (CVA), and include: Cerebral Hemorrhage, Cerebral Thrombosis, Cerebral Embolism, and Cerebrovascular Spasm.

1. Cerebral Hemorrhage consists of bleeding, which is caused from a rupture of a sclerosed, or diseased vessel in the brain;
2. Cerebral Thrombosis is an obstruction of a cerebral artery by a thrombus (blood clot);
3. Cerebral Embolism is a condition which occurs when an embolus (bubble of air, or piece of a thrombus) detaches from a thrombus and obstructs a cerebral artery; and
4. Cerebrovascular Spasm (Ischemia) is a condition which occurs when a cerebral vessel temporarily contracts, or constricts, impeding Blood flow. Hence, the oxygen supply is obstructed sufficiently to cause symptoms. When reversible, or when there is no permanent damage, they are known as Transient Ischemic Attacks, or TIA's. The TIA's are pre-stroke symptoms, commonly known as a mini-stroke.

SYMPTOMS

In dealing with stroke patients, it is important to differentiate between the severe type (which attacks the internal organs and channels), the mild type (which attacks only the channels), and the chronic type.

Stroke manifestations can either arise independently from an attack on the channels, or as a result of an attack on the internal organs. After a severe type of stroke, the patient enters into the next stage, clinically manifesting symptoms similar to that of the mild type. The shorter the time lapse before beginning treatment on a stroke patient, the better the chances are for recovery.

THE SEVERE TYPE OF STROKE

In a severe type of Wind-Stroke, the energy attacks the patient's internal organs and channels. It is characterized by a loss of consciousness, aphasia, numbness, paralysis and/or coma. If the right side of the brain is damaged, there will be paralysis on the left side of the patient's body with accompanying memory defects and language difficulties. If the left side of the brain is damaged, there will be paralysis on the right side of the patient's body, speech and language difficulties, slow, cautious movements, and memory problems.

Clinically, severe stroke patients are further classified into two categories: Tense or Flaccid. The Tense type (also referred to as the Closed type) corresponds to the Collapse of Yin. The Flaccid type (also referred to as the Open type) corresponds to the Collapse of Yang. These conditions and symptoms are further described as follows.

1. The Tense Stroke Symptoms include: clenched hands, open eyes, clenched jaw, red face, coarse breathing, and a forceful pulse.
   - **Treatment for Tense Stroke conditions.** Rescue the patient's Yin by giving herbs, as well as extending Qi into the patient's GV-20, -26, Kd-1, Lv-3, St-40, and the 12 Jing-Well points on both hands (Lu-11, LI-1, Ht-9, SI-1, Pc-9, TB-1). The patient should be given homework upon recovering, depending on his or her constitution.
2. The Flaccid Stroke Symptoms include: limp hands, closed eyes, open mouth, weak breathing, and a weak pulse.
   - **Treatment for Flaccid Stroke conditions.** Rescue the patient's Yang by giving herbs, as well as extending Qi into CV-8, CV-6, and CV-4 points, and GV-20 points (with or without the application of Moxa), to increase the patient's Yang. Patient should be given homework upon recovering, depending on his or her constitution.

THE MILD TYPE OF STROKE

In the Mild Type of Wind-Stroke, the energy attacks only the patient's channels. This is characterized by numbness, slurred speech, and unilateral paralysis. Clinically, mild stroke patients are further classified into two categories that are determined by the type of channels that are attacked.

1. An attack of the main channels causes no loss of consciousness or coma; however, there will be unilateral paralysis (hemiplegia), numbness of the limbs and deviation of the eye and mouth. Clinical manifestations vary according to the location of the channels and collaterals that have been affected.
   a. Symptoms of an attack on the channels include: facial paralysis, hemiplegia, motor impairment and numbness of the limbs.
   b. Symptoms of an attack on the collaterals consists of a unilateral numbness of the face and extremities.
2. An attack of the connecting channels is characterized by numbness.
   - **Treatment for mild stroke conditions.** In order to treat a Mild stroke condition combine Medical Qigong therapy with acupuncture therapy and herbs, to:
     a. Remove obstructions from the patient's channels,
     b. Purge Wind and transform Phlegm,
     c. Invigorate the patient's connecting channels,
     d. Regulate the circulation of Qi and Blood in the patient's channels, and
     e. Prescribe homework in accordance with the patient's constitution and condition.

In the acute stage, stroke patients should immediately be treated in combination with Western medicine, utilizing Qigong, acupuncture and herbal therapy. Treatments should be given everyday (usually every six hours) for the first month, and every other day for the second month. The general treatment is aimed at relieving the patient's spasms, and lowering his or her blood pressure (if the condition is due to hypertension). The primary points treated are determined by the type of stroke, which internal organs were affected, and the extent of damage to the patient.

THE CHRONIC TYPE OF STROKE

This condition is characterized by slurred speech or an inability to talk or eat, and involves either a unilateral (hemiplegia) or bilateral (paraplegia) paralysis.

- **Treatment for Chronic Stroke conditions.** In order to treat a chronic stroke condition the Qigong doctor must invigorate the patient's Blood, tonify the Spleen and open the patient's channels. The specific points will vary according to the area and location where the channels are blocked. The most commonly used points include: St-4, St-6, St-7, TB-17, GB-14, LI-4, SI-18, Tai Yang (Temple), and Yin Tang (Third Eye Point).

  If the body is in a state of unilateral paralysis, the doctor also treats the opposite side of the patient's head using Jing Point Scalp Therapy, in order to move stagnation, purge any Excess, and stimulate and fill any Deficiencies.

COMA

A coma is an abnormal, deep stupor occurring during illness, as a result of the illness, or due to an injury.

ETIOLOGY

From a Traditional Chinese Medical perspective, a coma results from Qi obstruction to the brain. More then 50% of coma cases result from trauma to the head, or circulatory accidents in the brain. A coma can be initiated from several different types of conditions.

1. A trauma induced coma is caused by an injury to the head as a result of accidents and internal hemorrhages. Electrocution can also induce a coma.
2. Insufficient Blood flow to the brain can cause a coma. This can result from circulatory accidents within the brain caused from hypertension, arteriosclerosis, thrombosis, and tumors.
3. An overdose of drugs, alcohol or medications, poisoning, gases or fumes such as carbon dioxide or carbon monoxide can all induce a coma.
4. Temperature abnormalities, such as hypothermia or sunstroke can cause vomiting, diarrhea, shock and then coma.
5. Insulin shock, epilepsy, electrolyte abnormalities, endocrine problems, meningitis, uremia or renal problems, and hyperglycemia can all induce a coma.
6. An infection induced coma is caused by an acute infection and bacterial intoxications as in fevers, botulism, and other infectious diseases; or as a result from abscess formations.
7. Psychogenic comas are also known as a hysterical coma.
8. Windstroke comas are caused by tension, hyperactive Liver Yang and Phlegm Fire overheating the brain. This results in urine and stool incontinence, cerebral hemorrhaging, and thrombosis. This is often located at the union of the anterior and posterior cerebral arteries, called the Circle of Willis.

SYMPTOMS

When a patient is in a coma, the Shen is completely out of its residence and the patient cannot be aroused by external stimuli. The patient's Shen is inactive, yet the Hun and Po are still there. For death to occur, it is not enough for the patient's Shen to leave, but the energy of the Hun and Po must also leave.

TREATMENT

In treating coma patients, it is mandatory to use methods that will activate the patient's Qi, remove the Blood stasis, and return the Shen back into the body. In the acute stages of treatment, use Medical Qigong therapy in addition to the Invisible Needle technique (see Chapter 35). It is most important to treat the coma patient as soon as possible. I have trained with three doctors of Traditional Chinese Medicine who were proficient in coma revival. Below are two accounts of coma revival cases from Dr. Lu Guo Hong's clinic at the Hai Dan Medical Qigong College in Beijing, China.

Case Study #1. Dr. Lu's patient had been in a coma for four hours prior to treatment. The patient regained consciousness within fifteen minutes during the treatment. Dr. Lu treated his patient using the Vibrating Palm technique (on points GV-20, Yin Tang, GV-15, CV-17, Bl-15, Pc- 6, TB- 5, and the extra points Shi Xuan at the tips of the fingers). By squeezing the Shi Xuan points, the Excess Heat was allowed to disperse (Figure 52.1).

Case Study #2. A female patient apparently incurred a series of Transient Ischemic Attacks (TIA) over a period of two years prior to coming to Dr. Lu. As the condition progressed, she developed a thrombosis on the left side of the brain, which paralyzed the right side of her body. The patient then apparently incurred thrombosis of the right side as well.

Because there were thrombosis on both sides of the brain, the patient was expected to die within a month. Dr. Lu, assisted by two other doctors, treated the patient for three days using the Extended Fan Palm technique, Invisible Needles, and Vibrating Palm technique. In this particular case, the doctors used the same points as in the previously mentioned case to cultivate the patient's Yuan Qi, adding points Bl-23, GV-4, CV-4, and CV-8. The patient was treated twice a day for 15 minutes, both morning and evening.

On the fourth day, the patient regained consciousness, and the treatment was continued once a day for a month using the Conception and Governing Vessel circulation points (to move the Qi along the Microcosmic Orbit), in addition to points on the arms including LI-4, LI-11, LI-15, TB-5, and TB-14 points.

When treating the lower limbs, he used GB-30, GB-31, GB-34, Sp-6, Sp-10, and St-41 points. The patient was given prescriptions and homework, specific to her constitution, state of consciousness, and motor control.

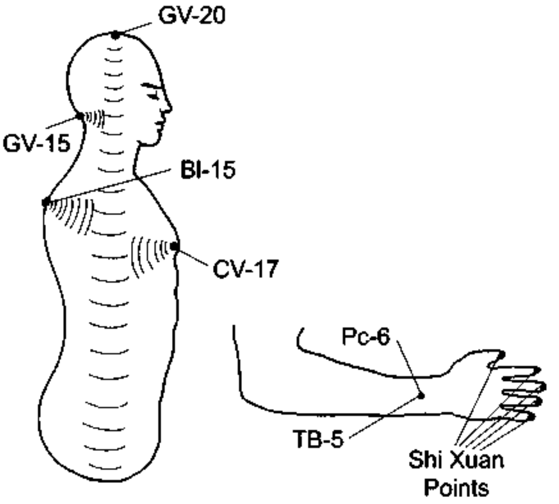

Figure 52.1. Common Points used for Coma Revival

The patient gradually regained movement in her limbs. Once the patient could move her limbs, she was given Dynamic Qigong to increase the flow of energy to her extremities. This included meditations and exercises that used guided imagery to regain muscle cellular memory. After two months, the treatment time was changed to every other month until the patient was fully restored.

FACIAL PARALYSIS (BELLS PALSY)

This condition is known as the sudden onset of unilateral facial paralysis.

ETIOLOGY

The patient experiences the swelling of the seventh facial nerve due to the External Wind and Cold invasion.

SYMPTOMS

The disease is characterized by hemiparalysis of the face which manifests as numbness of the affected area, deviation of the eye and mouth, distorted mouth with the corners tilted on the healthy side, and a half closed eye with dacryorrhea (excess tear flow).

TREATMENT

For this type of paralysis, it is important that the Qigong doctor treat the patient as soon as possible. The shorter the duration time from the initial onset and the treatment, the more successful the therapy tends to be.

1. Begin by purging the Toxic Qi out from the patient's face and head area, dredging the Toxic Qi down the arm and out the LI-4 point.
2. Using the Extended Fan Palm hand technique, emit Qi into the affected area, guiding the Qi to circulate in the Microcosmic Orbit from the anterior to the posterior of the patient's head.
3. Finally, initiate the Fire Cycle Microcosmic Orbit to regulate the patient's Qi.

TREATMENT MODIFICATIONS

- For treatment of the first stage of paralysis, the doctor is advised to apply gentle Pressing and Kneading techniques to the patient's face before using the Extended Fan Palm hand technique.
- To treat patients whose paralysis has remained in a chronic state, the doctor is advised to use deep pressing and kneading techniques before using the Vibrating Palm technique.

PRESCRIPTIONS AND HOMEWORK

1. Have the patient massage the affected side of the face from the corner of the mouth to the ear and forehead, down over the side of the ear and down the side of the neck 81 times.
2. The patient then focuses on specific areas on the face, circle massaging with the tips of the thumbs for 24 times.
3. Finally, the patient is given the Wash the Face and Massage the Head exercise to perform for 18 times (see Chapter 39).

MULTIPLE-SCLEROSIS (MS)

In Chinese medicine, multiple sclerosis is considered a type of Atrophy Syndrome. It is an inflammatory disease of the central nervous system in which infiltrating lymphocytes (predominantly T cells and macrophages) deteriorate the myelin sheath of nerves. Its pathology consists in the partial destruction of the myelin sheath around the spinal cord, brain and optic nerves.

ETIOLOGY

Multiple sclerosis is suspected to be an autoimmune disease, somehow linked to a viral infection originating from an invasion of External Dampness (especially in the beginning stages). As the External Dampness invades the body it begins to obstruct the channels, causing heaviness, numbness, and tingling in the limbs.

Diet is another contributing factor. The excessive consumption of greasy fried foods, dairy foods, or Cold foods can impair the Spleen, leading to the formation of Dampness.

In the later stages of MS, excessive sexual activity can also be a contributing factor for weakening the body and escalating the disease's progression. Excessive sexual activity can weaken the patient's Kidneys, contributing to such symptoms as dizziness, blurred vision, urgency or hesitation of urination, and extreme weakness of the legs. If excessive Liver Yang develops, symptoms will also include stiffness of the legs, severe vertigo, and vomiting. In the later stages of development, Liver Wind can cause tremors and severe spasms of the patient's legs.

Clinically MS is categorized as either a passive aggressive type of disease, in which episodes of neurologic dysfunctions are followed by times of remission, or a steady progression of neurologic dysfunction.

SYMPTOMS

The symptoms of MS were once thought to develop gradually; however, this is not necessarily true. In about 40% of clinical cases, the onset may occur in less than a few hours. In about one half of the patients, at onset the symptoms include weakness, or numbness in one or more limbs; later, the tendon reflexes become hyperactive.

In young patients the first obvious symptom is often retrobulbar neuritis (inflammation of the nerves behind the eyeball) causing blurred vision. In older patients the most common symptom is weakness of the legs.

Patients may experience vertigo, retrobulbar or optic neuritis, double vision, unstable walking, vomiting, and difficulty urinating. Also included are numbness or tingling of the extremities, as well as periodical electric shocks radiating through the limbs. These symptoms may be due to optic nerve, brain stem, cerebellar and spinal cord involvement.

Traditional Chinese Medicine divides these symptoms into two categories of clinical manifestations: Damp Phlegm with Spleen Deficiency, and Liver and Kidney Deficiency.

Clinical manifestations in patients with Damp Phlegm and Spleen Deficiency include feeling weak and tired, dizzy, as well as numbness, tingling and feeling of heaviness in the legs.

Clinical manifestations of patients with Liver and Kidney Deficiency include progressive weakness of the back, legs and knees, blurred vision, dizziness, poor memory, and a hesitancy, or an urgency to urinate.

TREATMENT

There are two basic treatment patterns in Medical Qigong for multiple sclerosis. If the treatment is started within its early stages, the symptoms can be completely eliminated, and the progression of the disease indefinitely stopped. This, of course, must be done in conjunction with an alteration in the patient's eating habits (no red meat, processed meat, sugar additives, or high fat diary products). The patient should reduce the amount of sexual activity, get plenty of rest and exercise, and avoid allergy causing agents. Treatments are described as follows.

1. To treat a patient with Damp Phlegm and Spleen Deficiency, first resolve the Dampness by purging the patient's torso and extremities. Next Dredge the Toxic Qi down the legs and out the Sp-9, and Sp-6 points. After purging, emit Qi into the Yellow Court (CV-12) to tonify the Spleen. Finally tonify the patient's Lower Dantian and regulate his or her Microcosmic Orbit (Fire Cycle).
2. To treat a patient with Liver and Kidney Deficiency, first tonify the patient's Kidneys and Liver by emitting Qi into the patient's Mingmen and Lower Dantian. Next emit energy into the patient's Bubbling Spring (Kd-1) points and direct the Qi to flow up the patient's Yin Channels into the patient's Lower Dantian and Kidney area. Finally, regulate the patient's Microcosmic Orbit (Fire Cycle).

PRESCRIPTIONS AND HOMEWORK

Patients often benefit from Medical Qigong exercises such as the Descending the Yang and Ascending the Yin Technique, Walking Therapy, Daoist Five Organ Exercises, as well as Taijiquan (Tai Chi Chuan) practice. One prescription I have found to be very effective in treating MS patients is as follows.

1. Have the patient sit in a chair, feet flat on the ground, hands resting on the knees, anal sphincter closed, tongue touching the upper palate while inhaling and exhaling through the nose.
2. While using natural breathing, the patient focuses on the Lower Dantian to create an energy ball. With each inhalation, divine healing light is absorbed into the body and stored in the Lower Dantian. With each exhalation, this healing light radiates within the Lower Dantian. After several minutes, the patient notices the Lower Dantian becoming full of heat and light.
3. Next, the patient imagines the energy ball in the Lower Dantian leaving the body through the coccyx into the Earth upon exhalation. Upon inhalation, the patient imagines the energy flowing up the center of both legs and filling the Lower Dantian. This meditation should continue for at least 20 to 30 minutes per sitting.

CHAPTER 53
QI EMISSION THERAPY FOR PSYCHOLOGY

INTRODUCTION TO ENERGETIC PSYCHOLOGY

The object of Chinese psychology is to study and understand the interplay of internal and external events on physical, mental, emotional, energetic and spiritual dimensions of human existence. These interactions include the human qualities and characteristics, as well as the fundamental attributes that exist and evolve within the changing environment and evolutionary development of nature. These interactions at the most fundamental level can be viewed as the transformational interactions of Yin into Yang, and Yang into Yin vibrational energy.

All matter vibrates regardless of its form (gaseous, liquid, solid, or in a transitional state such as ice melting into water). The human body receives these vibrations through the skin, eyes, ears, nose, and mouth, and records them in its cells. The mind interprets these incoming vibrations as images, sounds, smells, and sensations, and organizes these into memories. The extent to which the mind interprets the incoming data, depends upon the level of consciousness and awareness of the individual at the time of the recording. In times of danger, alertness normally increases and consciousness either expands or focuses more sharply until the danger is past; or if an injury occurs consciousness and awareness may be lost (e.g., a car crash).

When a Qigong doctor, attunes his or her vibrations to the patient's energetic field and emits Qi, the body's cells responsible for storing the vibrations are stimulated. The doctor's intention to heal the patient by releasing traumatic memories, combined with his or her projected Qi, often causes the patient to relive the original traumatic experience as the recorded vibrations are released. The Qigong doctor (still in tune with the patient's energy) is receptive to the patient's reaction and is able to interpret the released energy in the form of images, sensations, smells, sounds, and emotions while they are being reexperienced by the patient. If the doctor is sufficiently attuned, he or she may even pick up the conscious thoughts of the patient during this process.

As doctors tune in to their patients' cell emissions they are also able to project their own thoughts and interpretations of these memories back to the patients. This creates a positive feedback system, wherein the doctors' thoughts can influence the patients' thoughts and emotional reactions. Through thought transmission, Qigong doctors are often able to break the vicious cycle of negative thinking that so often leads to disease in the body.

In Western psychology, it is believed that a person's tendency to interpret events - even negative imaginary events (e.g., assigning malicious intentions to peoples' actions), can result in a self-fulfilling prophecy. By assuming and expecting the worst behavior in people, such a person always feels on the defensive and provokes others around him or herself to anger or irritation, thus reinforcing the negative viewpoint.

Chinese energetic medicine adds to this simple behavioral understanding of self-fulfilling prophesies, a new level of comprehension based on energetic interchange between people. When someone observes, or imagines, a specific negative thing happening, certain brain cells record this information in the form of vibrations. These vibrations can be received by another person's corresponding brain cells and projected back to the original sender, reinforcing the negative interpretation of the event(s). If this feedback loop continues - an imaginary wrong may become manifest, i.e., the negative event imagined actually occurs.

It is believed that different cells store different types of information, received in the form of vibrations. The silent thought communications between doctors and patients is believed to involve different sets of cells than those used to receive and record emanations of divine energy. These cells interrelate with the patient's emotional changes that are stored and released from the Yin and Yang organs and tissues, and are expressed through the patient's Wu Jing Shen.

The patient's ontology also includes the patient's growth and self-realization in relationship to the Wu Jing Shen (Five Essence Spirits) and the emotions stored within the body's Yin and Yang organs. Each of the Five Yin Organs has a Yin and a Yang aspect within its energetic substance, as well as emotional factors, and dynamic interconnections. The Yin and Yang aspects give rise to the Creative and Controlling Cycles and the Five Yin organs with their associated elements and emotional energies (Figure 53.1 - 53.3).

Summary of the Five Elements		
Wood	Liver	Anger, irritation and rage
Fire	Heart	Joy, happiness, pleasure and excitement
Earth	Spleen	Worry, sympathy, anxiety, contemplation and meditation
Metal	Lungs	Grief, sorrow, anguish, distress and heartache
Water	Kidneys	Fear, terror, panic, horror and fright

Figure 53.1. The Five Elements, Organs and Acquired Emotions

THE EMOTIONAL ENERGY STORED WITHIN THE FIVE YIN ORGANS

Traditional Chinese Medicine classifies the etiology of diseases into Internal and External physical, mental, emotional, energetic, and spiritual influences. In Chinese psychology, the primary focus is placed on the "root" of the energetic/emotional dysfunction. The root of the energetic/emotional dysfunction is often the unprocessed emotions which are stored in their associated organs. Once the patients' mental and emotional symptoms have been alleviated, the energy that was being used to suppress the excessive emotions, is then freed to be focused on their physical healing.

Traditional Chinese Medicine also recognizes the power of thoughts, images (including psychotic delusions and hallucinations), and fantasies to generate and release emotions. Qigong doctors study these images and emotional reactions of attraction or aversion to evaluate their patient's condition. When traumatic memories and emotions are suppressed over a long period of time, or are still fresh in the mind, any minor incident can precipitate a tidal wave of emotions. A rapid influx of thoughts and images sweeps through the mind, too fast for the conscious brain to register. These thoughts and images are of past painful experiences, and are loaded with emotional energy. Getting cut off by someone on the freeway, for example, may cause a momentary reaction of surprise, or an explosive enduring rage, depending upon the condition of the Five Yin Organs. In the latter case, the driver is overwhelmed by the Liver's sudden release of rage, triggered by the traumatic images and emotions from the past.

Clinical pathology is considered an Excess or Deficiency of the emotions at the level of the Five-Elements and their corresponding organs. The Excess or Deficiency of an internal organ can cause a release of thoughts or images that sweep through the patient's mind, bringing a sudden emotional response. In some instances these emotions first affect a specific organ, and then the entire balance of energy, as the first organ's energetic field deteriorates or adversely expands. When emotionally charged Qi is trapped within the tissues of the patient's internal organs, it travels through the channels, affecting the metabolic order of the viscera, glands and nervous system.

Creative Cycle			
Wood	Anger	Creates Joy	Fire
Fire	Joy	Creates Worry	Earth
Earth	Worry	Creates Grief	Metal
Metal	Grief	Creates Fear	Water
Water	Fear	Creates Anger	Wood

Figure 53.2. The Creative Cycle

Controlling Cycle			
Wood	Anger	Controls Worry	Earth
Earth	Worry	Controls Fear	Water
Water	Fear	Controls Joy	Fire
Fire	Joy	Controls Grief	Metal
Metal	Grief	Controls Anger	Wood

Figure 53.3. The Controlling Cycle

Emotions consist of energetic light that is either tempered and softened by the patient's belief structure and energetic fields. When the emotions and life purpose are in harmony, the Wu Jing Shen become focused; the Hun and Po do not disperse, grief and anger do not arise, and the Five Yin Organs effectively combat pathogenic influences. If this energetic light becomes unstable, i.e., too Excessive or Deficient, the patient's inner strength and individual nature begins to undergo dramatic changes.

The body's internal organs and their relationship to the Five Elements are as follows.

EXCESSIVE ANGER INJURES THE LIVER (WOOD)

The Chinese character for anger (Nu) is depicted by the Heart ideograph underneath two characters, a woman (placed on the left) and someone who has the upper hand (placed on the right). It expresses the feeling and frustration of an enslaved woman being dominated by someone.

PATHOLOGICAL MANIFESTATIONS OF ANGER

Pathological anger is considered the perverse movement of the element Wood. Anger is a generated force that has the potential to unleash explosive amounts of Qi, pushing the energetic manifestations to the extreme.

THE YIN AND YANG ASPECTS OF ANGER

The emotion of anger can be divided into Yin and Yang tendencies. The Yang type of anger can burst out due to accumulated pressure, resulting in the upward flow of massive amounts of Blood and Qi. The Yin type of anger implodes, remaining buried or held inside, thereby causing internal agitation and dissatisfaction which generates blockages.

The Liver's negative nature is one of being stubborn and rude; its positive nature is one of being compassionate, unselfish, kind, and merciful. The rising of Heat from the Liver's Qi corresponds to the level of anger; since the Liver stores the Blood and the Blood stores the Hun (Ethereal Soul). When the Hun leaves the body due to anger and rage, the animal nature of the Po take over. As the individual's Liver Fire rises, it causes the eyes and the face to become red. Liver Fire can also injure the Liver as well as the Blood, bile and Body Fluids associated with this organ.

ANGER AFFECTING THE LIVER

When Liver Qi becomes depleted it retracts, causing the emotion of fear to arise. When Liver Qi fills to Excess, the result is progressive restlessness, irritability, anger, and rage. In some instances, these emotions affect the entire balance of energy so that the organ begins to deteriorate.

Anger can cause trembling of the entire body due to the Liver's control of the tendons and ligaments.

When an individual experiences any type of boundary invasion, the Liver Qi triggers the Heart (Shen) to recall from its long term memories past traumas (containing both anger and fear). At the same time, the Liver also drains the Kidney Qi to meet the apparent life-threatening situation; this prepares the body for the flight or fight response. The power of the emotional response causes the Heart to go into shock and prevents the Shen from directing the mind. Thus another trauma is experienced by the patient's interpretation of the event (e.g., getting cut off on the road) and will automatically trigger an exaggerated emotional response in the future.

Given that the Liver is the mother of the Heart (responsible for long term memory) and the child of the Kidneys (responsible for short term memory), the Liver can drain the Kidneys and flood the Heart with anger, causing one's mind to go blank and the face to turn red.

ANGER AFFECTING THE SPLEEN AND STOMACH

Anger can damage the Spleen and Stomach, causing digestive problems, loss of appetite, and a sour taste in the mouth. These symptoms are directly related to the level of anger.

ENERGETIC PATTERN OF ANGER

The Qi pattern of anger has a very hot energy associated with it. When discharged, this hot energy swells up like a tidal wave expanding from behind the person's back over the head (Figure 53.4). It crests over the head and rushes forward like a wild animal attacking. The emotion also travels from the groin up the back of the body and out through the eyes.

Grief controls anger as metal cuts wood. At the root of anger is pain and sorrow. The anger protects the patient from feeling the initial hurt that festers like an infected wound. When the doctor lances the wound, allowing the toxic anger to drain, the uncovered original hurt can heal cleanly. Anger is transformed to sorrow and grief. The compassionate, non-judgmental attitude of the doctor allows the patient to grieve openly. The release of toxic anger and pain enables the patient to listen to the Hun. By focusing on the patient's emotional hurt from past experiences, which are the root of his or her anger, a shift from anger to hurt occurs. This exchange of emotions allows the patient to confront the hurtful past memories and discharge the toxic anger.

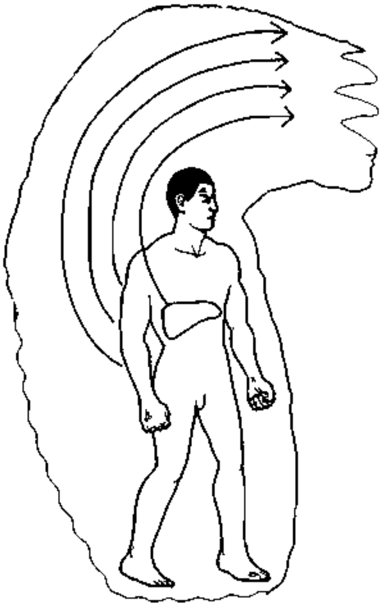

Figure 53.4. Anger rushes from the Liver over the back like a tidal wave, causing the Qi to ascend.

EXCESSIVE JOY INJURES THE HEART (FIRE)

The Chinese character for joy (Le) is depicted by the ideograph of a large drum, framed by bells mounted on a wooden stand. It expresses the feeling of rhythmical ceremonies and excitation. The character gives the image of a hand beating a drum during a festival (this same character is used for music).

THE YIN AND YANG ASPECTS OF JOY

When the Heart's emotions are balanced, they can be experienced as love, joy, happiness, excitement, and pleasure. This warmth radiates through the body affecting the expansion of free and easy Qi circulation. Joy is communicated and felt throughout the body, mind and soul.

PATHOLOGICAL MANIFESTATIONS OF JOY

Excessive joy and laughter can injure the Heart and scatter the spirit; the Qi becomes congested and the Heart flutters erratically. When the Heart Qi becomes depleted, it can no longer control the Lungs: the emotion of grief arises, and when the Qi fills to Excess, hysteria results. Water extinguishes Fire, therefore the emotion of fear controls the emotion of joy. By supporting a hysterical patient to become aware of his or her underlying fears, the doctor can startle the patient into a sense of sobriety.

Pathological joy and laughter are considered the perverse movement of the element Fire. If the patient's Fire becomes uncontrollable, the Qi - which expands outward towards the periphery - will cause the patient to become hysterical. After Fire has exhausted the patient's life-force, it leaves the patient exhausted, dazed, and disenchanted. The Heart's negative nature is confused, doubtful, and greedy; the positive nature is open-minded, loving, non-judgemental, and trustful.

ENERGETIC PATTERN OF JOY

Joyful feelings make the body's Qi become soft and relaxed. This energetic pattern is expansive, and causes a slower movement of the Qi. As the Heart controls the blood vessels and stores the Shen, a joyful spirit produces a clear, calm mind and a healthy body. This energy bubbles out from the center core of the body and ripples out in all directions (Figure 53.5). A light energy gently flows from the recipient back to the person expressing the emotion. While in this state, the patient's head is usually held erect, the face is full of color, and the limbs are relaxed.

EXCESSIVE WORRY INJURES THE SPLEEN (EARTH)

When the Spleen Qi becomes depleted, all four limbs lose their strength and the Five Yin Organs become unstable. When the Spleen Qi becomes Excessive, the body becomes swollen and the flow of urine and menses loses its smoothness.

THE YIN AND YANG ASPECTS OF WORRY

The emotions associated with the Spleen are relative to the energy expended in contemplation, meditation, anticipation, and worry. Too much worry and thought produces Qi stagnation. A stagnant condition causes Stomach disorders, such as poor appetite, indigestion, ulcers, constipation, or diarrhea. Stagnation can also block the Shen, which is responsible for emotional and mental balance, and thus result in a lack of common sense. The Spleen's negative nature is suspicious and self-centered; the positive nature is truthful, sincere, and compromising.

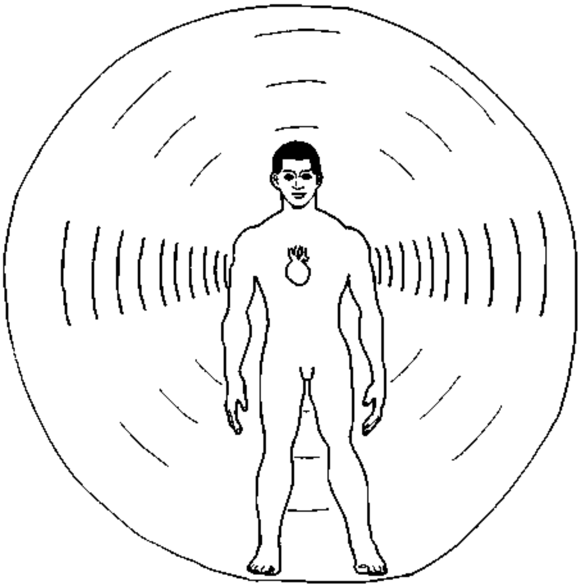

Figure 53.5. Joy bubbles out from the center core of the body, rippling out in all directions, causing the Qi to slow its flow.

WORRY AFFECTING THE MEMORY

The Spleen stores the Yi, and the Yi stores ideas. The Spleen transfers short term memory to the Kidneys; the Kidneys later transfers the acquired memory to the Heart, which is responsible for long term memory. In trauma victims, the Kidneys' role in transferring memories, can also account for the loss of memories that are too terrible to accept, since these memories are transferred along with great fear, or fail to get transferred at all.

THE ENERGETIC PATTERN OF WORRY

When one is worried, the energy comes up from the Earth and congests in the Middle Dantian area, thus weakening the Wei Qi (Figure 53.6).

CONTROLLING THE EMOTION OF WORRY

Because the element Wood controls the element Earth, the emotion of anger controls worry. It is very important that any anger uncovered be directed at the root cause, and not at the doctor who is trying to help a patient end a negative pattern. The doctor will gently challenge toxic belief structures held by the patient, until the patient begins to confront whatever issue is at the root of the anger.

When the patient becomes angry with the doctor, it is generally a condition known as "transference." If and when transference arises, the Qigong doctor can ask the patient, "Who do I remind you of right now?" or "When was the last time you felt this way?" Trace the patient's feelings as far back as possible. If the doctor gets mad at the patient, it is called "countertransference." Both transference and countertransference can involve any emotion (e.g., falling in love, feeling afraid, etc.).

EXCESSIVE SADNESS AND GRIEF INJURES THE LUNGS (METAL)

The Chinese character for sadness (Bei) is depicted by the Heart ideograph underneath the character for the back (placed on the top). It portrays a picture of a person who has turned his or her back on the true feelings of the Heart. The Chinese character for grief (Ai) is depicted by the mouth ideograph (wailing, groaning, and lamenting) placed within the center of the character of someone dressed in a special garment for mourning. Pathological grief is considered the perverse movement of the element Metal.

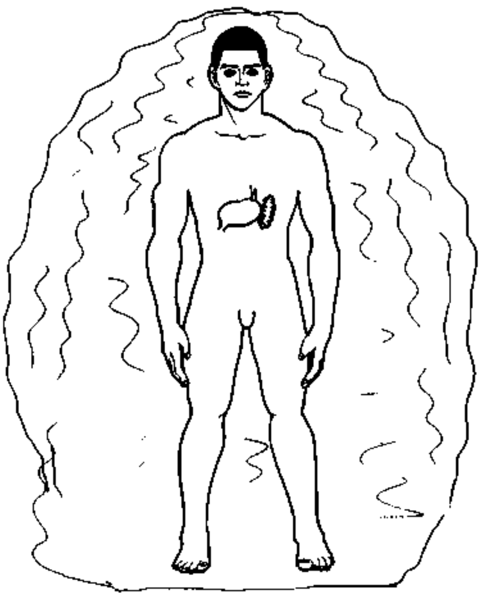

Figure 53.6. Worry binds the Qi, causing the Qi to stagnate.

When grief is experienced, its natural movement of condensation begins to compress the Heart, interfering with Blood circulation. This obstruction destroys both the respiratory flow of the Lungs, as well as the body's liquids. The Lungs store the Qi, the Qi stores the Po (Corporeal Soul). When the Lung Qi becomes depleted, shortness of breath develops. When the Lung Qi becomes excessive, the result is a feeling of discomfort in the chest, panting, and difficult breathing.

THE YIN AND YANG ASPECTS OF GRIEF

The feelings associated with the Lungs are sadness, melancholy, grief, sorrow, and anxiety. The negative nature is cunning, jealousy, and pessimism; the positive nature is generosity and righteousness. In the Five Elemental Controlling Cycle, the Lungs (Metal) affect the Liver (Wood). Excessive sadness from the Lungs can cause the Liver Qi to stagnate the Blood (Metal Overcontrolling Wood), leading to possible intercostal and hypochondriac pain, bloating, and depression.

GRIEF'S AFFECT ON THE INTERNAL ORGANS

Grief disperses the Qi, and thus injures the Lungs. Crying can lead to the release of toxic anger (that suppresses the feelings of sorrow) and grief. Depression and despair may also be manifestations of abnormal Liver Qi stagnation. When grief from the Lungs is excessive, the Spleen may be affected (this almost always happens after a major loss, such as a death in the family).

THE ENERGETIC PATTERN OF GRIEF

The external energetic patterns of grief and sorrow have a heavy, downward moving energy (Figure 53.7). The internal energetic patterns and manifestations of grief have a very thick, sticky energy, and a denseness like gelatin. Long periods of extreme grief also injure the Heart, Pericardium, and the Triple Burner Channels.

CONTROLLING THE EMOTION OF GRIEF

Because Fire melts Metal, the emotion of joy controls grief. The grieving process is a very important part of healing; however, chronic grief can be extremely detrimental, and when it is extreme and prolonged it can lead to spiritual oppression or possession. If the survivor cannot accept the death, he or she may prevent the departure of the loved one's spirit for many years on end. The patient may beg the Eternal Soul not to leave and cling with unrelenting grief to the lost one's energy. This type of dependency deepens, rather than lessens over time, as the patient opens more to receive the loved one's spiritual essence. In so doing, other Gui (spirits) are given power to enter as well.

After a death of someone loved, the patient may consciously suppress any joy or pleasure. The patient may believe that to allow any relief to suffering is tantamount to a betrayal of the deceased. This belief structure must be addressed by the doctor and patient. Establishing a sense of empathy with the patient will provide a safe environment for the patient to release the emotion of grief. The doctor can work towards empowering the patient to begin to recognize times of joy, and learn, or remember, how to count his or her blessings, to let go of sadness and grief, and reclaim lost joy.

Grief may give way to despair and anxiety. Because anxiety suppresses respiration, anxiety lowers the body's resistance to pathogenic factors by weakening the Wei Qi; anxiety also damages the Large Intestine, resulting in symptoms that range from constipation to ulcerative colitis.

Just as too much grief can cause inappropriate laughter, or gallows humor; too much laughter masks the accumulated and suppressed grief. Gallows humor is a syndrome whereby the patient laughs instead of grieving about hurtful past, or current experiences. Gallows humor is a defence mechanism of the patient's denial. Grief and humor are examples of opposites (Yin and Yang) transforming into each other.

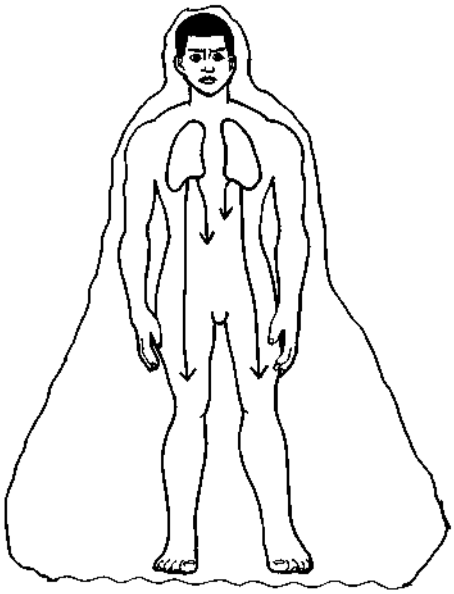

Figure 53.7. Grief causes the Qi to become obstructed.

The concern in dealing with many grieving people (pertaining to the death of a patient's mate or close friend, etc.) is that the patient may be using grief to suppress other unresolved emotions, such as: guilt, fear and loneliness.

1. Fear of forgetting the deceased voice, appearance, etc. may be due to feelings of guilt over not having expressed enough love towards the person when it was possible to do so.
2. Guilt over any expressed feelings of joy or happiness, may seem morally wrong because they cannot be shared by the departed.
3. Guilt may arise from too many hurtful things said, or too many things left unsaid to the departed.
4. Guilt over never having had the chance to really say good-bye (as with victims of accidents and sudden death) may become all consuming.
5. Loneliness and isolation may result from the loss of friendships.

6. Worry that loss of income and security may cause financial difficulties and health problems may cause the patient to cling to the past.

To assist the patient to heal from these toxic beliefs, the doctor prescribes creative visualizations, such as the following.

"Imagine, see and feel yourself and your loved one clothed in your astral bodies. Embrace and express your feelings and thoughts, knowing that these will be heard and received. Listen to your loved one's answer.

Ask forgiveness for any cruel remarks or failures to be as loving as you wish you could have been. Visualize your loved one accepting the request for forgiveness; embrace with love, and release him/her to be with God with your blessings."

In cases where the patients could not attend the funeral, have the patient imagine the funeral and perform a ritual closure (e.g., lighting a candle at home and saying a prayer).

Finally, encourage the patient to give him or herself permission to experience some joy in life and to go on living.

EXCESSIVE FEAR INJURES THE KIDNEYS (WATER)

The Chinese character for fear (Kong) is depicted by the Heart ideograph underneath two characters, a structure (placed on the left) and repeated strokes by the small blows of a hammer (placed on the right).

THE AFFECT OF FEAR ON THE INTERNAL ORGANS

The Kidneys store the Jing, the Jing stores the Zhi (actions that the mind expresses). Pathological fear is considered the perverse movement of the element Water. Fear is the unrestrained sinking and uncontrolled descent of Water Qi. When the Shen of the Heart is no longer supported by the Kidneys' Jing, the Shen leaves its residence. This results in the emotional release of panic. If the Heart Fire becomes weak, the Kidney Water rapidly draws downward and freezes. Insecurity, hesitation, immobility, and paralysis ensue. If the Heart Fire become Deficient, it cannot radiate warmth, causing the patient to crouch and hide.

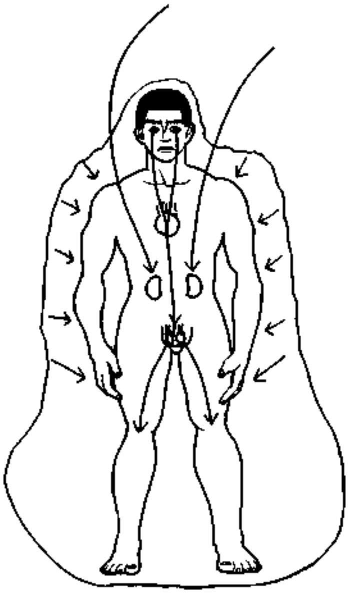

Figure 53.8. Fear causes the Qi to move inward and downward.

If the Kidney Yin is Deficient, the vision deteriorates, there is a ringing in the ears, dizziness, night sweats, and nocturnal emissions. If the Kidney Yang is Deficient, the Mingmen Fire becomes Deficient, and causes a counter-flow of energy affecting the hearing. Kidney Yang Deficiency also results in shortness of breath, cold limbs, and impotency. When the Kidney Qi becomes Excessive, swelling occurs and the Five Yin Organs become unstable. The feelings associated with the Kidneys are fear, paranoia, and panic.

THE YIN AND YANG ASPECTS OF THE KIDNEYS

The Kidneys' negative nature is arrogant, troublesome, and ignorant; the positive nature is peaceful, soft, and tender. Fear can cause loss of Bladder control, weakening of the knees, renal failure, and permanent damage to the Kidneys. It can injure the Heart, affecting the speech, eyes, and the Shen.

When fear injures the Kidneys, this can affect the Liver, causing muscle spasms, trembling, and disruption of the menses cycle. When the Gall Bladder is affected, decision-making becomes more difficult.

THE ENERGETIC PATTERN OF FEAR

The energetic pattern of fear is very cold and implodes into the center of the body, drawing all the energies inward and downward. Fear travels from the eyes down the front of the body into the groin (Figure 53.8).

Because Earth absorbs Water, contemplation controls fear. Therefore, the doctor can help the patient control the fear through visualization and meditation. The purpose of the visualization is to resolve and reprogram the patient's thinking patterns.

THE EMOTION OF FRIGHT

The emotion of fright is distinguishable from fear by the sudden, unexpected shock to the system. The Chinese character for fright (Ju) is depicted by the ideograph of a small bird opening its frightened eyes, in fear, anxiety, and dread. Fright startles the Shen, which then scatters the Qi and injures the Heart. Persistent fright causes damage to the Kidney Qi and results in a negative cycle of chronic fear.

FIVE YIN ORGAN EMOTIONAL RESPONSES

The common ingredient in any emotional dysfunction is the power element maintaining it. Holding on to original traumas, whether consciously or unconsciously, requires a great deal of energy. It is possible to stop this negative pattern by releasing the toxic emotional energy, thereby creating the opportunity for the mind, body, and spirit to heal. A new pattern of harmonious balance can be maintained, however, only after the old dysfunctional patterns are consciously recognized during their occurrence. Old habits die hard; new patterns require time and continuous effort before becoming as automatic as the previous dysfunctional habits. Every emotion is meant to serve a positive function, even anger, fear, and grief. Patients often require reminding of this fact. Every emotion also has its darker side when repressed for too long and when avoided at all costs.

1. Learning how and when to express anger is as important as not holding on to it for extended periods of time. Anger serves as a warning sign that personal boundaries are being violated, threatened, wounded, or that injustices are being committed. Patients must be taught better communication skills so they need not resort to screaming or insulting others to protect themselves. Appropriate assertiveness skills must be learned, and may require either psychotherapy, or classes in assertiveness training. Patients must also learn how to ask for what they need and desire, without expecting those close to them to mind read. Without learning how to set appropriate boundaries, without the ability to say "no" to those who take advantage of them, they will fall into the old habit of accumulating resentments; or they will alienate their employer, friends, and family through unguarded candor when angry.
    - On its negative side, anger can cause rage and destructive patterns (both internally and externally).
    - On its positive side, anger can cause constructive boundary setting and the righting of wrongs.
2. Fear, like anger, serves as a warning sign. Fear informs patients when they are in danger, so that appropriate measures can be taken to insure safety and survival. Some patients do not recognize the danger signals until it is too late to avoid falling victim to unscrupulous people. Some may require counseling and instruction in how to establish healthy relationships, and in how to identify individuals who are potential abusers before it becomes too late to avoid the ensuing violence.
    - On its negative side, fear can cause disabling panic.
    - On its positive side, fear can cause an awareness of the need for self-protection.
3. Joy in excess can lead to disharmony. Nervous laughter, or always playing the clown to gain acceptance, can be extremely difficult to unlearn. Often professional help is needed to increase socialization skills and self-esteem.

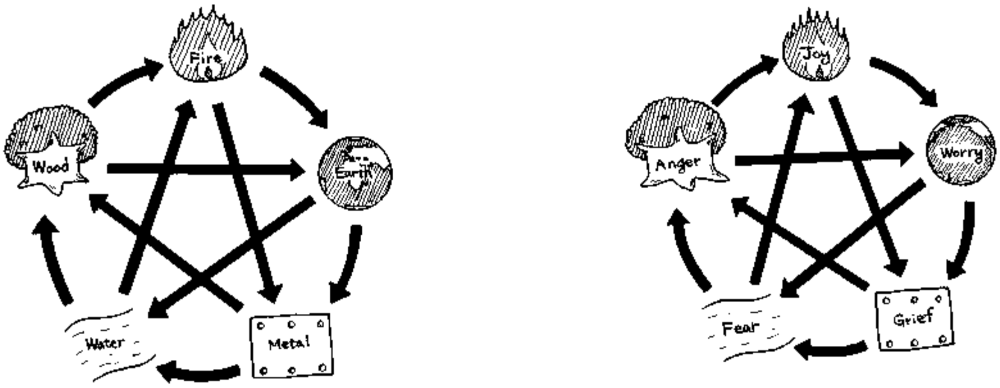

Figure 53.9. The Five Element Cycles

- On its negative side, joy can cause hysteria.
- On its positive side, joy can be a source of happiness to the spirit.

4. Sorrow is necessary at times to motivate patients to change. The old self must periodically die to make room for new experiences and learning. Risking sorrow is part of the process of bonding and maturing. Sadness is often a signal of imbalance in the patient's life, indicating insufficient rest or recreation, and far too much work.
   - On its negative side, sorrow can cause self-pity, rather than compassion for oneself.
   - On its positive side, sorrow can release distress.
5. Worry and anxiety are normal reactions to change. When these emotions become pathological in intensity of duration, patients may need professional help in the form of medication (e.g., for obsessive compulsive disorders), and/or psychotherapy (e.g., in the case of generalized anxiety disorder and posttraumatic stress disorder).
   - On its negative side, worry can cause great anxiety.
   - On its positive side, worry can be a motivating source for change.

The healing stages used in Medical Qigong Therapy employ these transitions in order to assist the patient's emotional and spiritual growth.

The picture in Figure 53.9 summarizes the Five Elements, their associated organs, and their associated emotions.

The cycles serve a functional importance for emotional transitions. The elements can be arranged in a number of ways. For clinical purposes, the main emphasis is on the Creative and Controlling Cycles.

EMOTIONAL DISHARMONY

It is becoming more common in the Western belief system that if a patient concentrates his or her mind and emotions on healing the illness, they can accelerate the recovery. Some Western research suggests that when the patient stops thinking and trains the mind to relax (reaching an alpha state), the healing is accelerated.

Emotional disharmony can resonate and penetrate one's entire being, thereby creating a vicious

Major Organ Disharmony	→	Emotional Disharmony
↑		↓
Environmental Disharmony	←	Disturbed Behavior

Figure 53.10. The Cycle of Emotional Disharmony

circle affecting the body, mind, emotion, energy, and spirit. As evidenced by the diagram (Figure 53.10), any disharmony in one of these four factors can trigger the cycle.

This vicious circle can be set in motion by any major disharmony, be it environmental, physical, emotional, or behavioral. Once in motion the disease can feed on itself. The power of emotional energy generated can at times overwhelm the conscious mind. The mind then has three options: to give in to the emotional tidal wave, to become eventually paralyzed through the effort of controlling it, or simply to deny that anything is wrong.

When emotional energy is suppressed it accumulates within the body, causing the Qi to flow incorrectly or to condense. The unresolved emotions cause chronic blocks in the body, particularly in the back, neck, Stomach, and intestinal areas. This toxic energy stored in the muscles and organs leads to tension, stress, and illness. The root cause of chronically tense shoulder muscles, for example, can be traced to: tensing the muscles to suppress the emotion(s), a strong negative belief system, and an accumulation of unresolved past traumatic memories, which continue to resurface.

Unless the initial cause of tension is released, the tissue memory associated with the initial trauma will repeat its programmed pattern, resulting in chronic tension and pain. Painful and negative emotions are not the only emotions suppressed. Many people also suppress feelings of joy and love. Most do so out of fear, guilt, and sense of unworthiness. Some do so out of religious beliefs, believing that suffering is good for the soul, or that the are being punished by God for sins of omission and/or commission to test their faith.

All emotions are triggered by sensory associations with pleasant and unpleasant past memories. In many cases patients are completely unaware of the triggering smell, image, sensation, or sound. To them it seems that the trauma, or memory, resurfaces from nowhere. This inability to pinpoint the trigger mechanism may indicate the progression of the patient's illness. Patients suffering from panic attacks, for example, have no idea what stimulus may have caused the overwhelming terror that they feel. Quite often panic attacks increase in frequency, partly out of the anticipatory fears of having another one, and partly due to their fear spreading and generalizing.

The first panic attack may have taken place in a restaurant, triggered by one of the five senses registering similar data (a smell or image for instance) which occurred during one of their traumatic memories. Henceforth they will avoid this restaurant, or perhaps all restaurants because new data has been added to the original trigger mechanism. New smells and images occurring during the attack may now stimulate the old memory of the trauma. Eventually, such patients may become home bound as they try to limit themselves to only safe places; their panic attacks have generalized to agoraphobia (fear of places or situations from which escape might be difficult or embarrassing). Each panic attack requires more energy to suppress both the original trauma and all of the new traumas added to it from surfacing.

Outgoing, successful people can thus transform themselves into anxiety ridden hermits. They can also develop physical symptoms, or turn to alcohol and drugs in attempts to cope. In less extreme mental/emotional diseases, changes in skin colors, tone of voice, mannerisms, and nervous twitches may manifest. As the mind tries to suppress uncomfortable emotions, these unresolved feelings often manifest as physical symptoms.

At times, however, the symptoms manifest more on the mental level, as in obsessive compulsive disorder. The logical mind, failing to completely suppress the emotion, may develop frantic obsessive thinking patterns, that seem to take on a life of their own. Some patients may obsess about dying, even if they seem by Western medical standards to be physically completely healthy; or they may obsess about killing their own child, or forgetting to lock all the doors when leaving the house and so on. They have no control over their repeating thoughts, and are aware of the unrealistic and unreasonable nature of these obsessions. Nevertheless, they are helpless to stop themselves.

Obsessions often manifest with compulsive addictions such as washing the hands so often during the day that they are raw and bleeding, or pulling out the hair on their head, or folding and unfolding their clothing. Any ordinary daily activity may turn into a compulsion, requiring the patients to continue performing a meaningless action again and again.

Another frequent disease found in traumatized patients goes by the name of posttraumatic stress disorder. Patients suffering from this condition have flash backs, i.e., they can not stop having moments of releasing the original traumas again and again with full intensity.

In posttraumatic stress disorder patients' symptoms are often disabling. Anxiety is chronic, they are in a state of hyper-alertness. They have become good at disassociating from their body, since it has become a very uncomfortable place to be. The Shen wanders and the patients may, or may not, be aware of where their spirit goes or what it experiences at these times. In extreme cases the ego fractures, usually during early childhood, and a full blown identity disorder (previously called multiple personality disorder) ensues.

Patients suffering from obsessive compulsive disorders, posttraumatic stress disorder, agoraphobia and panic attacks, or identity disorders should be referred out for extra help to a mental health professional specializing in these disorders.

Ideally, the doctor can teach and encourage the patient to get in touch with their body, and learn how to achieve and maintain a balance between the physical, mental and emotional planes. The doctor can also teach the patients how to transition detrimental energy into a higher level of consciousness, when the patients become aware of an unpleasant emotion. When patients learn how to channel their own energy in a more healthy way, the cycle of reliving the past can be broken. A negative emotion can be transmuted and expressed as a positive emotion; emotional energy can be converted to positive life-force energy.

EMOTIONAL FREQUENCY RESONATION

When the Qigong doctors are appropriately sensitive to the patients' emotional transitions, they are able to detect, and recognize, recurring energetic patterns. These energetic patterns are in constant resonation throughout the body's tissues. As patients shift their emotional thinking, their energetic frequency resonation (the body's Hz rhythm) shifts. These different emotions establish specific patterns of energy in the body. Each individual has a brain wave pattern that is as unique as their fingerprints. According to Richard Lee, director of the China Healthway's Institute, individuals resonate at different frequencies depending on their state of consciousness. Mr. Lee's research reveals the following.

1. Those motivated by fear and instinct will resonate at the delta wavelength, between 0 and 4 Hz (hertz, i.e., cycles per second). This level of energetic resonation, as measured by an electroencephalograph (EEG), is normal in an awake infant and a sleeping adult. This delta wavelength, however, indicates brain pathology in adults that are awake.
2. Those motivated by attachment and desire will resonate at the theta wavelength, between 4 - 7 Hz. This level of energetic resonation is generally common in children and in adults in the early stages of sleep. This frequency range also occurs when an individual creates mental imagery, and often manifests when a person is involved in deep altered states of consciousness. This level of energetic resonation also occurs in many brain disorders.
3. Those motivated by service and compassion will resonate at the alpha wavelength, between 8 and 13 Hz. These low amplitude alpha waves usually occur when the brain is in a calm, and the mind is relaxed (in a state of wakefulness). It was discovered that Qigong masters emit energy within this frequency range.
4. Those motivated by organization and planning will resonate at the beta wavelength, between 14 and 25 Hz. This level of energetic

resonation usually occurs when the nervous system is active and the body is awake. It manifests during sensory input and mental activity. This frequency range is usually common for those suffering from nervous overload and excessive thinking.

When the patients' vibratory rate shifts towards the alpha rhythm, they begin to experience deep relaxation and clarity of mind. Relief from chronic pain and symptoms associated with low energy is achieved.

According to Richard Lee, the infrasonic sound produced by Qigong masters is effective because it causes decreased resonation in the theta and beta ranges. This simultaneously decreases attachments and desires, calms the emotions, and reduces excess thinking, thereby freeing the patient's vital sources to be redirected towards healing.

DIFFERENTIATION OF PSYCHO-EMOTIONAL DISORDERS

Traditional Chinese Medicine has, for centuries, recognized the fact that psycho-emotional states, energy, and physical health are intertwined. In the Huang Di Nei Jing, the emperor's medical adviser, Qi Po, outlined the Internal Factors (emotions) as primary causes of disease.

Within the context of Traditional Chinese Medicine, Medical Qigong therapy is useful for treating many types of psycho-emotional disorders. Qi emission therapy can be used to calm the patients' Shen, as well as accelerate the patients' recovery time. In China, psycho-emotional problems are commonly called Jing Shen disorders and are differentiated as follows: Obstructed Shen, Unsettled Shen, and Weakened Shen.

1. Obstructed Shen is a condition due to the Shen becoming obstructed with Phlegm, or in mild cases, becoming obstructed by Qi or Blood. Symptoms are characterized by confused thinking, clouding of the mind, and in severe cases, complete loss of touch with reality. Degrees of severity range from very mild disorders to severe schizophrenia or mania.
2. Unsettled Shen is a condition due to the Shen becoming unsettled from Blood or Yin Deficiency (in mild cases), Qi stagnation, Blood Stasis, Fire, Empty Heat, Phlegm Fire, or Internal Wind. Symptoms are characterized by agitation, restlessness, and anxiety.
3. Weakened Shen is a condition due to Deficiency of Qi, Yang, Blood, or Yin. Symptoms are characterized by depression, mental exhaustion, and melancholy (dysthymia).

ETIOLOGY

The cause for psycho-emotional disorders can range from a variety of factors, including prenatal or postnatal constitutional imbalance, diet, unbalanced sexual activity, overwork, alcohol or drug abuse, trauma, stress, emotional immaturity, and instability, as well as a lack of love or spiritual perspective.

TREATMENT PRINCIPLES

There is a considerable degree of overlap among Jing Shen disorder symptoms. Therefore, the treatment principle must be based on a clear distinction between Excess and Deficiency, as well as the cause (root) and manifestation (branch) of the symptoms. The main treatment principles are as follows:

1. Nourish the Heart and calm the Shen to treat Shen weakness due to Qi, Yin, Yang, Blood, or Body Fluids Deficiency.
2. Purge the Excess, move stagnation, and calm the Shen to treat Shen obstruction or Shen agitation (unsettled Shen) due to Excess conditions of Qi, Blood, Phlegm, and/or Fire stagnation.
3. Purge the Excess, nourish the Heart and calm the Shen to treat Shen obstruction due to Yin Deficiency with Empty Heat.
4. Resolve the Phlegm, open the Heart's orifices, and calm the Shen to treat Shen obstruction due to Phlegm or Phlegm Fire.
5. Sink and calm the Shen to treat unsettled Shen due to rising Qi.

The term "calm the Shen" is used to describe the treatment principles in all psycho-emotional disorders. It should be understood to mean not only calming the mind, as in anxiety, but also lifting the patient's mood (as in depression), or clearing the mind (as in schizophrenia or mania).

SELF-TREATMENT PRESCRIPTIONS FOR PSYCHO-EMOTIONAL DISORDERS

Most meditations are difficult for patients with psycho-emotional disorders due to several factors. One factor is the problem of maintaining focused attention, as well as retaining the memory of specific medical Qigong prescriptions. The success, or failure of the treatment will also depend on the severity of the disorder.

POINT MEDITATION FOR EMOTIONAL TRAUMA

In China, Channel Point meditation is used in conjunction with the Massage Tapping method for treating patients with conditions such as anxieties, phobias, and addictions. In acupuncture, 135 of the original 365 acupuncture points are used to treat mental, emotional, energetic, and spiritual disharmonies. In fact, many of the "new" points can be applied for the treatment of Jing Shen disorders.

The cause of fear and anxiety is the disruption to the body's energy system from emotional trauma. When a patient is focusing on a particular anxiety or fear, the Shen becomes disrupted, and the energetic harmony of the body becomes unbalanced. This energetic imbalance causes certain physical reactions within the patient (abnormal breathing, pulse, etc.), and results in pathological behaviors.

TREATING EMOTIONAL ANXIETIES, PHOBIAS, AND ADDICTIONS

To treat anxieties, phobias, and addictions, the doctor encourages the patient to focus attention on the specific fears, or cravings for the particular addiction, to initiate a maximum emotional charge of distress. When the patient reaches maximum emotional distress, the doctor calls out specific channel points for the patient to tap.

The patient is given a specific mantra to repeat while tapping, to discharge the accumulated emotionally charged Qi, such as; "I accept this fear (or other emotion or craving)" or "I am being healed." The treatment continues for five to ten minutes, as the patient focuses on feeling the distress as much as possible, until all the trapped emotion has been discharged. Some energetic points commonly chosen for treating emotional anxieties, phobias, and addictions are GV-26, Ht-7, and Pc-6. These specific points are used to draw the energetic charge away from the Heart and into the extremities; this action facilitates calming the patient's Shen.

The initial tapping method, in conjunction with vivid imagination and positive affirmation, disperses the energy away from the Heart and other organs and calms the Shen. The spontaneous dispersing of the emotional charge, stored within the patient's tissues, releases all feelings of stress and anxiety. This type of guided Tapping Meditation therapy not only reduces and disperses the patient's emotional trauma, but also causes an energetic reprogramming within the patient's psychophysical dynamics.

MEDICAL QIGONG TREATMENT OF THE PHOBIA AND ANXIETY OF DEMON POSSESSION

These specific treatments are used in China to treat specific mental, and emotional disorders, that lead the patient to imagine that they are being haunted by spirit demons. In rectifying this type of psycho-emotional disorder, it is recommended that the Qigong doctor proceed as follows.

1. Perform Channel Point therapy to open the energetic points of the Eight Extraordinary Vessels (Figure 53.11).
2. Press and knead the Baihui GV-20, Dazhui GV-14, Lingtai GV-10, and Feishu Bl-13 points. Then apply the Extended Fan Palm or Sword Fingers method (see Chapter 33), using the Pulling and Shaking manipulations (see Chapter 34) to emit and conduct Qi along the Governing Vessel.
3. Press and knead the following channel points Baihui GV-20, Yintang (Ex.Pt.), Renzhong GV-26, Tinggong SI-19, Jiachè St-6, Quchi LI-11, Hegu LI-4, Weizhong Bl-40, and Chengshan Bl-57 points.
4. Apply the Sword Finger technique, using the Vibrating and Shaking manipulations to emit Qi into the Jiuwei CV-15 and Zhongwan CV-12 points and conduct Qi along the Conception Vessel to return it to the Lower Dantian.

MEDICAL QIGONG TREATMENT FOR SPECIFIC SPIRIT/GHOST HALLUCINATIONS

These specific treatments are also used in

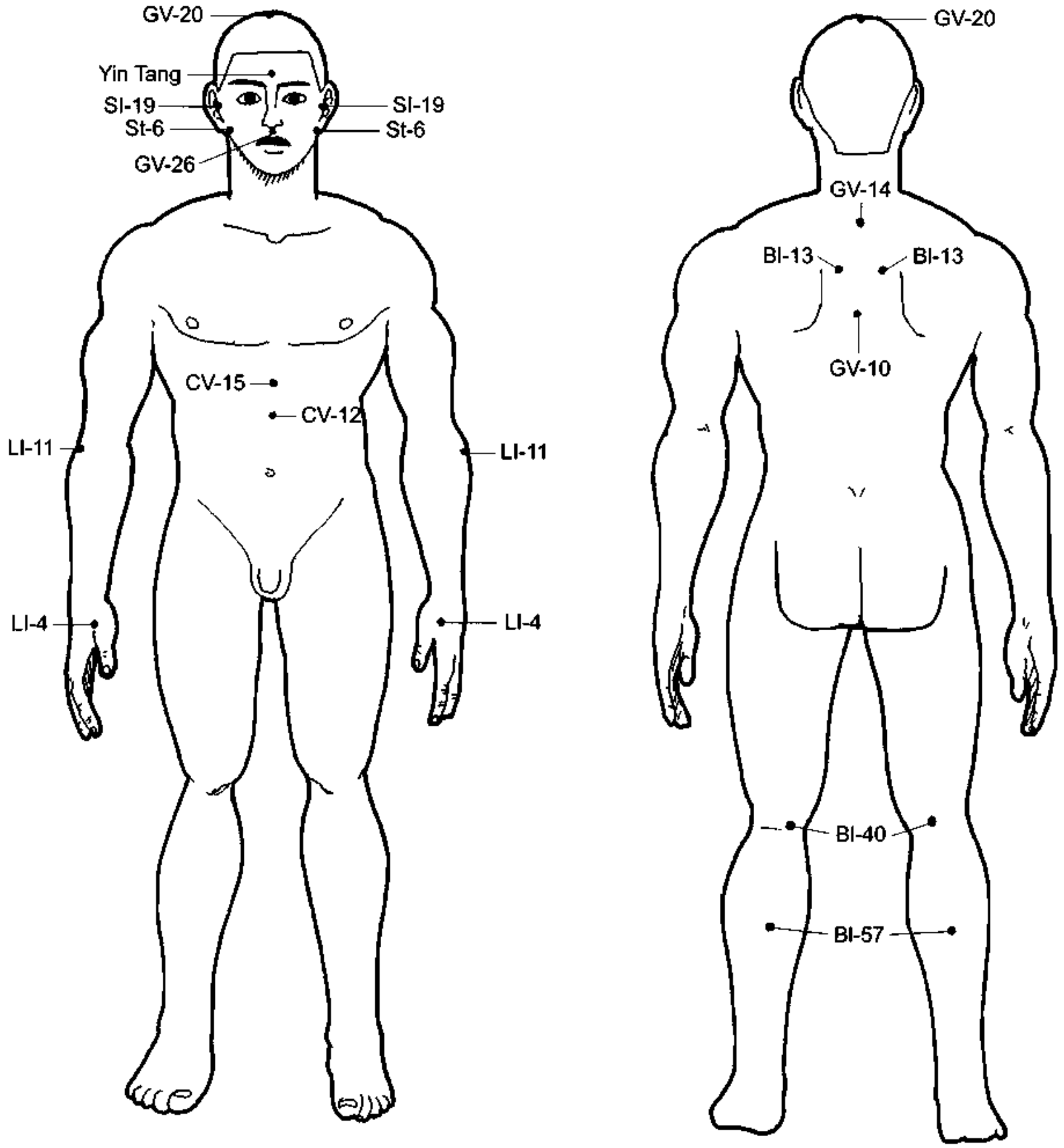
GV-20
Yin Tang
SI-19
SI-19
St-6
St-6
GV-26
CV-15
CV-12
LI-11
LI-11
LI-4
LI-4
GV-20
GV-14
Bl-13
Bl-13
GV-10
Bl-40
Bl-57

Figure 53.11. Medical Qigong Channel Point Therapy for the Treatment of Energetic Delusions

China to treat specific mental and emotional disorders that lead the patient to see spirit/ghost hallucinations. The Qigong doctor may additionally treat the patient according to the following therapeutic point stimulation:

- If the patient is disoriented, sees and hears floating corpse ghosts, or hears crying ghosts, treat the Tianfu Lu-3 points.
- If the patient is manic-depressed, sees and hears ghosts talking, treat the Tianchuang SI-16 points.
- If the patient is having seizures, sees and hears ghosts talking, treat the Tianzhu Bl-10 points.
- If the patient is manic, suicidal with incessant wild talking and actions, treat the Fengfu GV-16 point.
- If the patient experiences chronic nightmares, confused dreaming, and restless sleep, treat the Tianyou TB-16 points.

MEDICAL QIGONG TREATMENT FOR DEMON OR SPIRIT POSSESSION

Since the creation of Traditional Chinese Medicine in the People's Republic of China, the notion of spirits has been down-played. However, within the confines of the Medical Qigong clinics, I have been fortunate enough to be able to discuss with several Qigong doctors the cause, effect, and treatment of demon and spirit possession.

These following Thirteen Ghost Points are listed in the *Thousand Ducat Formulas*, prescribed by the famous physician Dr. Sun Simiao in the 7th century China, and were once used for the treatment of spirit possession (Figure 53.12). Today in China, these points are now used in the TCM clinics for the treatment of manic and depressive mental disorders, as well as for treating epilepsy. In extreme manic cases (with Heat) the doctor bleeds all of these points as well as all twelve Jing-Well points on the patient's hands. Classically, the patient is treated by bleeding each of these points, or by inserting an acupuncture needle into each point and then immediately withdrawing it. These Ghost Points are stimulated in a sequence of progressive point therapy, wrapping and enfolding the patient's body in five consecutive circles.

1. The Guigong (Ghost Palace), also known as the Renzhong GV-26 (Middle of Man) point, is the first point to be stimulated and is treated to calm the Shen and clear the brain.
2. The Guixin (Ghost Faith), also known as the Shaoshang Lu-11 (Lesser Metal's Note) points on both thumbs, are treated to calm the Shen and restore Collapsed Yang.
3. The Gullei (Ghost Fortress), also known as the Yinbai Sp-1 (Hidden Clarity) points on both big toes, are treated to calm the Shen, clear the brain, in order to instill clarity of thought and mind.

These first three points are used to complete the energetic pass of the first circle, covering the head, tips of the thumbs and tips of the big toes. The GV-26 point is considered the meeting point of Yin and Yang in the body, while the Lu-11 points on both hands and the Sp-1 points on both feet are located on the body's Hand and Foot Tai Yin Channels.

Next, the following two points are used to complete the second circle:

4. The Guixin (Ghost Heart), also known as the Daling Pc-7 (Big Mound) points on both wrists, are treated to calm the Shen and clear the brain.
5. The Guilu (Ghost Path), also known as the Shenmai Bl-62 (Extending Vessel) points under both outside ankles, are treated to calm the Shen.

The Ghost Heart and Ghost Path points complete the energetic pass of the second circle, covering and wrapping the patient's wrists and ankles. The Pc-7 points on the wrists are located on the body's Hand Jue Yin, and the Bl-62 points on the outside of the ankles are located on the body's Foot Tai Yang Channel points.

The following set of four points are stimulated to complete the energetic pass of the third circle:

6. The Guizhen (Ghost Pillow), also known as the Fengfu GV-16 (Wind's Palace) point just below the occiput, is treated to clear the brain.
7. The Guichuang (Ghost's Bed), also known as the Jiache St-6 (Jaw Vehicle) points on both sides of the jaw, are treated to dispel Wind and Cold and to clear Heat.
8. The Guishi (Ghost Market), also known as the Chengqiang CV-24 (Receiving Liquid) point under the lower lip, is stimulated to treat mental disorders, dispel Wind and Cold, clear

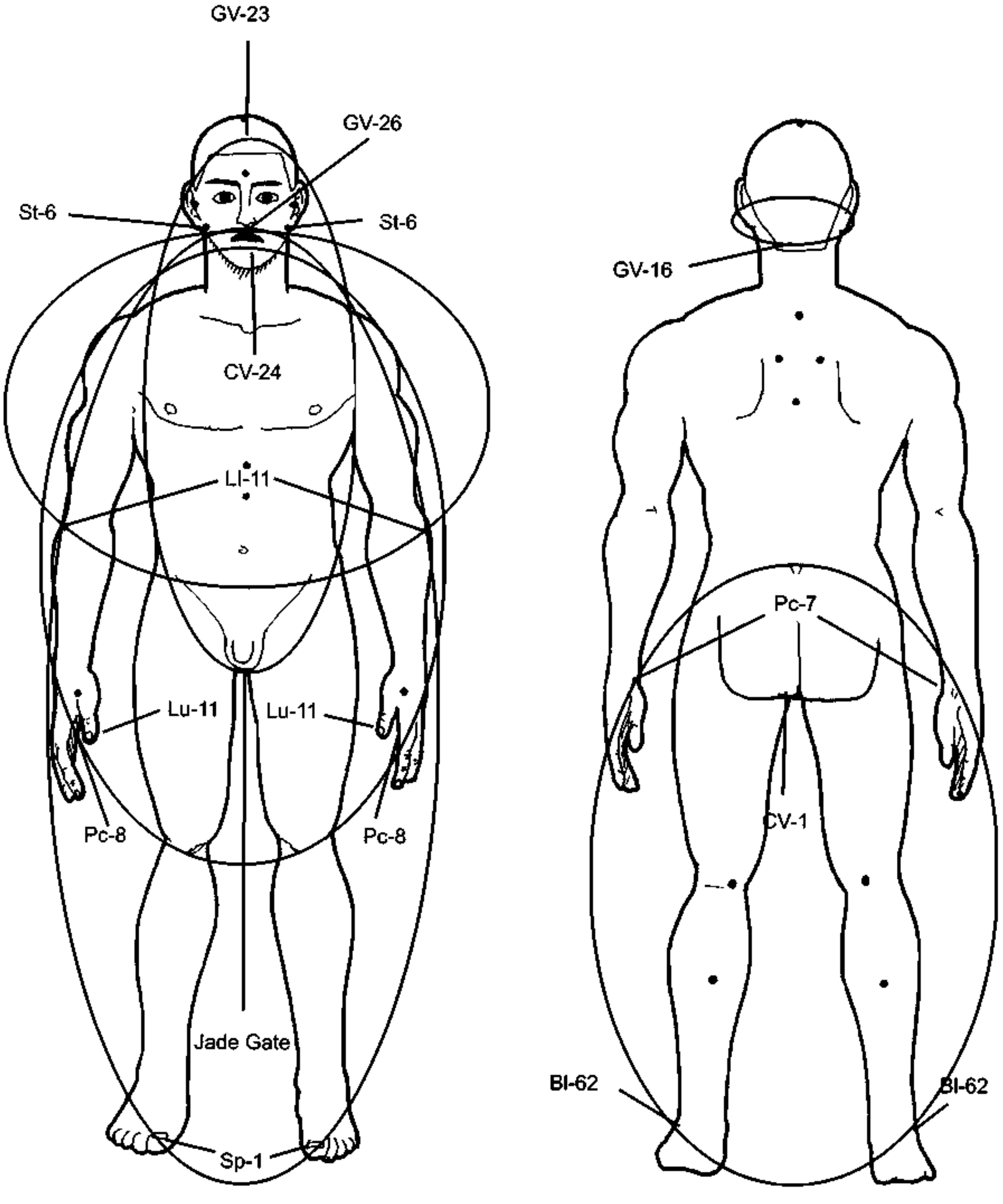
GV-23
GV-26
St-6
St-6
CV-24
LI-11
Lu-11
Lu-11
Pc-8
Pc-8
Jade Gate
Sp-1
GV-16
Pc-7
CV-1
Bl-62
Bl-62

Figure 53.12. Medical Qigong Treatment for Spirit Possession

Heat, and to transform Dampness and Phlegm.

9. The Guicu (Ghost Cave), also known as the Laogong Pc-8 (Palace of Labor) points at the center of both palms, are stimulated to treat mental disorders and to clear the brain.

These four points complete the energetic pass of the third circle, by covering and wrapping the patient's head, then drawing the energy into the center of the patient's palms. The GV-16 point on the back of the patient's head is located on the back gate of the Upper Dantian, the St-6 points are located on the sides of the mandible, and the CV-24 point is located in the depression in the center of the patient's mentolabial groove (jaw).

The next two points complete the energetic pass of the fourth circle:

10. The Guitang (Ghost Hall), also known as the Shangxing GV-23 (Upper Star) point, is treated to brighten the Shen and clear Heat.
11. The Guicang (Ghost Store), also known as the Yumentou (Jade Gate) in women, and Yinxiafeng (Below the Hidden Seam) in men, is used to treat mental disorders. The Jade Gate is located at the opening of the vagina, the Below the Hidden Seam is located in the Huiyin area in men.

These two points complete the energetic pass of the fourth circle, by covering and wrapping the patient's body from the head to the lower perineum. The GV-23 point is located on the anterior hairline, at the top of the patient's forehead, and the Ghost Store point is located in the lower perineum.

The following two points are stimulated to complete the fifth energetic pass:

12. The Guitui (Ghost Leg), also known as the Quchi LI-11 (Crooked Pool) points, are treated in order to drain the pathogenic influences from the Lungs and to calm the Po (Corporeal Soul).
13. The Guifeng (Ghost Seal), also known as the Haiquan (Sea Spring) and She Xia Zhong Feng (Under Tongue Middle Frenulum) are two Extra points located below the tongue. They are stimulated in order to treat mental disorders.

These final two points complete the energetic pass of the fifth circle, by covering and wrapping the patient's body from the arms to the head. The LI-11 points are located in the depression on the lateral end of the elbow crease, and the Haiquan Sea Spring points are located under the tongue.

TREATING PHOBIAS AND DEPRESSION USING "THE WINDOWS OF THE SKY" POINTS

If we observe the body as a figure 8, with the head being the upper portion and the torso being the lower portion, the neck would represent the area where the two circles meet. The points and orifices of the head, face and neck are like great windows in which Qi moves.

The trunk of the body is considered Earth and the realm of the physical; the head pertains to Heaven and the realm of the spirit. The neck is where all of the classic "window" points are located. The neck is the junction between Heaven and Earth, and is considered the Earth's window to Heaven. With the exception of Lu-3 (upper arm) and Pc-1 (chest), eight of the ten Windows of the Sky points are located on the neck itself, and seven of the ten points have the word Tian (Heaven) as part of their name.

Over the last quarter century, the Chinese have used the Windows of the Sky points in treating such psychological disorders as depression and phobias. In the Qigong clinic, we look for three indications before prescribing the Windows to the Sky points.

1. According to the Nei Jing, the Windows of the Sky points are used when the Yang Qi is unable to ascend to the patient's head. The result is that the patient will have physical symptoms in the head or Heart (i.e., loss of hearing, or pain described as being behind the Heart.
2. The second indication is that there is something missing in the Heavenly aspect of the patient's spiritual life. When the patient becomes stuck in their own negative patterns, these particular points can be used to open a window of hope by bringing into the patient's body the illuminating light from Heaven.

In these cases, the patient's disconnection from emotions and spiritual insight is due to the chronic negative patterns that block the descent of Heavenly light.

3. The third indication is a constant attitude of despondency by the patient. This behavior is not in accordance with his or her own values and beliefs, but the patient cannot control this pattern. Some patients know that something is amiss but cannot discern the source of the problem.

The Window points induce a hypnagogic state (a transitional state between sleeping and awaking) that allows some patients to experience out-of-body projection and past life regressions.

TREATMENT

When treating patients with the Windows of the Sky points, the Qigong doctor should use slow, gentle pressure on the skin to dip into the tissues, and then proceed with pure intention. The ten Windows of the Sky points are described as follows.

1. The Tian Fu Lu-3 (Heaven's Residence) point is sometimes called the Celestial Palace. The word Heaven represents divinity, clarity of vision, and the sense of the spirit (Figure 53.13). This point has a powerful psychological effect on all emotional problems deriving from Lung Qi disharmonies (e.g., depression, claustrophobia, agoraphobia, mental confusion, and forgetfulness).
2. The Fu Tu LI-18 (Support the Prominence) point is also known as Support and Rush Out. It is used on patients who feel out of touch with their own inner strength (Figure 53.14). This point relieves coughing, resolves Phlegm, and disperses masses (cysts and tumors) in the neck area. It is also used to treat a patient's inability to speak.
3. The Tian You TB-16 (Heaven's Opening) point can be used to calm the patient of anxieties and to soothe Heart palpitations (Figure 53.15). This point is also used for dizziness, sudden deafness, and diminished visual acuity.
4. Tian Zhu Bl-10 (Heaven's Pillar) is a Sea of Qi point. It is used to clear the brain, open the orifices, brighten the patient's eyes, and re-

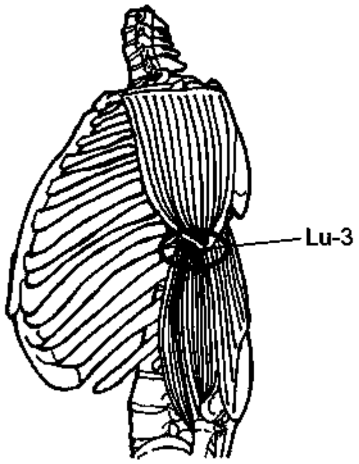

Figure 53.13. Heaven's Residence

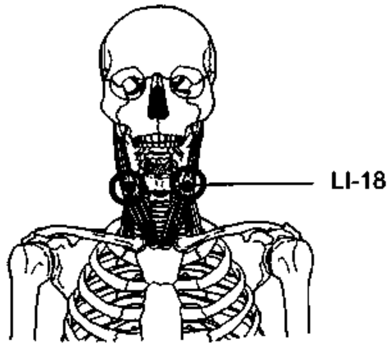

Figure 53.14. Support the Prominence

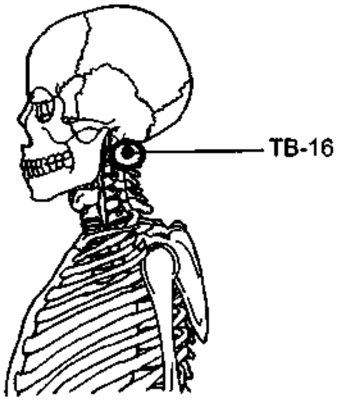

Figure 53.15. Heaven's Opening

move obstructions from the channels. The energy that moves and is located at these points has to do with the patient's will, ambition, and inner fortitude (Figure 53.16). Energetic dysfunction located at the Heaven's Pillar may manifest as deep sadness or anxiety, and be related to a lack of inner strength, or will. This point is also used when the patient's feet are unable to support the body.

5. Ren Ying St-9 (Man's Welcome) is a Sea of Qi point. The ancient Chinese philosophers associated this area of the body with Man in the triad of Heaven, Man, and Earth. This point is known as a place where energy is being welcomed into the body (Figure 53.17). It is a very sedating point of Blood and Qi, and is used to reduce blood pressure. Opening up this point allows the patient to become receptive to nurturing energy (Ying Qi). The doctor should be gentle and only treat one side at a time. It is also used to disperse Lung Qi and Regulate Qi.
6. The Tian Tu CV-22 (Heaven's Rushing Out) point is also known as the Celestial Chimney, and is an area where the body's internal energy and spirit leaves the body through sound (Figure 53.18). Heaven Rushing Out is related to emotional and spiritual expression, and is relevant when the patient's spiritual self (i.e., the individual's life purpose) is not being supported by physical actions (his or her code of conduct). In other words, the patient's life purpose is not in alignment with the patient's behavior. This point can also be used to regulate the Lungs, stimulate the descending action of the Lung Qi, clear Heat, and resolve Phlegm.
7. Feng Fu GV-16 (Wind's Palace) is a Sea of Marrow point, and is also a very good point for strengthening the body's Wei Qi (Figure 53.19). This point connects directly to the brain, and is used to eliminate Wind, clear the mind, open the sensory orifices, and benefit the energetic function of the brain (used for treating convulsions, epilepsy, or palsy).
8. The Tian Chuang SI-16 (Heaven's Window)

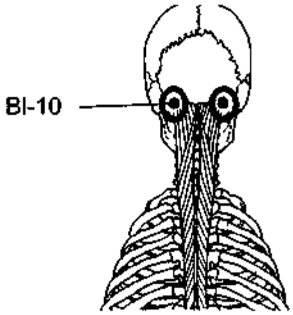

Figure 53.16. Heaven's Pillar

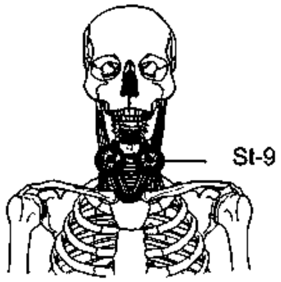

Figure 53.17. Man's Welcome

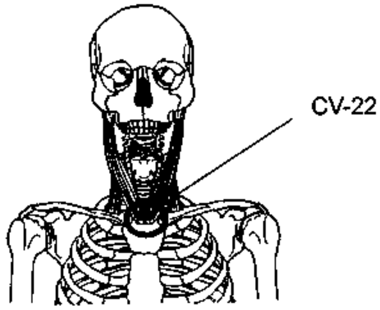

Figure 53.18. Heaven's Rushing Out

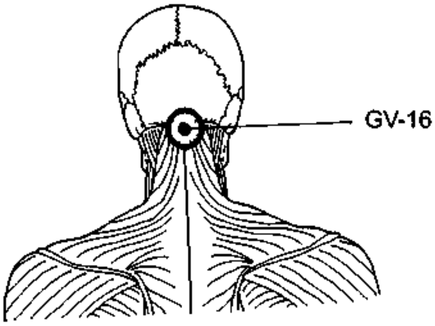

Figure 53.19. Wind's Palace

point is used to open a window to Heaven, especially in patients who need help in perceiving how to change a negative behavior pattern into a productive and harmonious one (Figure 53.20). This point is used for dispersing the Liver's Qi and for eliminating buzzing, pain, or deafness of the ears.

9. The Tian Rong SI-17 (Heaven's Appearance) point is also known as the Celestial Countenance. This point is where the Small Intestine's Channel Qi is received by, and enters into, the Heaven's Body Chamber (the head or skull). It can be used to produce a strong movement of energy from the head downward into the body.

   This point is used in spiritual emergencies, when patients have lost their connection with Earth energy, and remain blissed out, and oblivious of reality. It is often used in cases of addictions, once the patient is conscious of the addiction (Figure 53.21). This point is also used in order to expel Excess Fire and remove obstructions from the patient's channels.

10. The Tian Chi Pc-1 (Heaven's Pond) point is also known as the Celestial Pool. The term pond represents a source, resource, or reservoir of energy for the spirit. This point influences the area where the Shen gathers or pools in the chest region (Figure 53.22). It is capable of enhancing the transition from normal consciousness to feelings of being submerged in a sea of energy.

    It affects the Heart and Pericardium which govern the Shen, and whose symbolic home is Heaven. The doctor uses this point to remove the patient's street armor, and to begin the process of opening the patient's Heart protector at its source. This point is also used to disperse the Lungs' Qi, expand and relax the chest, regulate Qi, and clear Heat.

    Note: Because of the proximity of the breast tissue, the Heavenly Spring points (Pc-2) are sometimes used in women instead of the Heaven's Pond points (Figure 53.23). The Heavenly Spring point is located halfway down the anterior face of the biceps muscle.

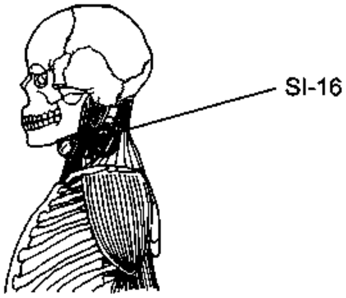

Figure 53.20. Heaven's Window

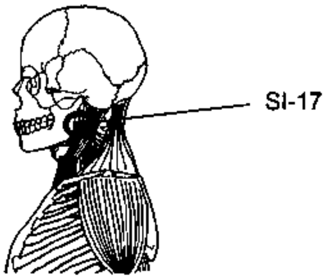

Figure 53.21. Heaven's Appearance

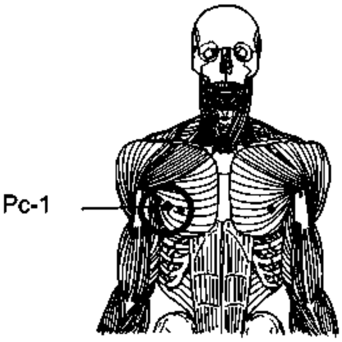

Figure 53.22. Heaven's Pond

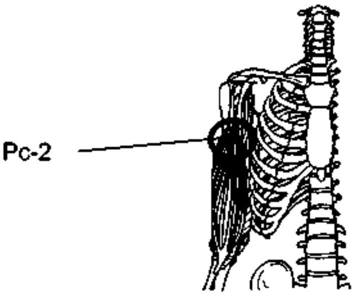

Figure 53.23. Heavenly Spring

TREATING PSYCHO-EMOTIONAL DISORDERS RELATED TO STRESS

The following prescriptions are used in order to assist the patient in dealing with suppressed anger and grief, for additional prescriptions used in order to treat emotional problems (see Chapter 19).

TREATING EXCESS LIVER FIRE AND THE REPRESSION OF ANGER

One of the most common problems observed in the clinic is Excess Liver Fire. One of the many causes of Excess Liver Fire is the repression of anger and a suppression of the spiritual components of the Hun. When anger is suppressed, its volatile Heat will sometimes attack and drain the Kidney Yin (Water). Therefore, Excess Liver Fire is generally related to a Deficiency of the Kidney Yin energy. When Qigong doctors reduce the Excess Liver Heat, they must also increase the patient's supply of Kidney Yin (Water) Qi (see note #2).

If the Liver's Yin becomes Deficient, the body's Qi will not be able to keep the patient's Internal Fire from rising. This causes the Yang to rise to the patient's head, and the Blood to become Deficient. The patient becomes irritable as the Liver's Yang begins to rise. This in turn, causes high blood pressure, which is associated with such problems as headaches, dizziness, insomnia, tinnitus, hypertension, and migraine headaches. Deficient Liver Yin, if unchecked, turns into Liver Fire, which causes severe high blood pressure, and heats the patient's Blood. As the Blood heats, it becomes Deficient, and causes Liver Wind. Unchecked Deficient Liver Yin results in shaking and trembling. If the Liver Wind mixes with the patient's Phlegm, it blocks the patient's channels resulting in paralysis, aphasia, memory problems, and eventually stroke.

The method of treatment is chosen according to the patient's syndromes:

1. For patients sensitive to energy, use Medical Qigong therapy while the patient is lying supine.
   - Purge and disperse Qi from the patient's Liver;
   - Pull pathogenic Qi out the patient's Liver Channels;
   - Tonify the patient's Kidneys through the Lower Dantian;
   - Give Liver purging and Kidney tonification prescriptions and homework.
2. For patients not sensitive to energy, use Tui Na and An Mo in conjunction with Medical Qigong therapy while the patient is lying prone.
   - Purge and reduce the Fire in the patient's Governing Vessel and Bladder Channels;
   - Start at the Mingmen and dredge the patient's Liver Fire, ending at the sacrum, using a counterclockwise circle rotation to reduce the Liver's Qi;
   - Give Liver purging and Kidney tonification prescriptions and homework.

Note #1. If the Liver Fire attacks the Stomach, resulting in vomiting and migraine headaches, have the patient lie supine. Because the Liver's Fire can travel transversely along the hypochondrium area, the Qigong doctor rocks and gently shakes the patient from the center of the thorax down to the Kidneys using intention, lead and disperse the Heat out from the patient's torso. Next, pull the Stomach Qi down and out of the patient's feet.

Note #2. Tonify the patient's Kidneys by using Jing Point therapy on the following points: St-36, GB-34, Sp-6, Lv-3, and Kd-1 on both sides of the body.

BEATING THE BAG AND SHOUTING TO DETOXIFY LIVER HEAT AND RELEASE TRAPPED ANGER AND RAGE

Beating the Bag and Shouting is the most common Medical Qigong homework given to patients with Excessive Liver Fire. The concept of Beating the Bag and Shouting is best comprehended when the function and flow of Liver Qi Heat is understood. When striking the bag and shouting, the shaking movements of the body will:

- facilitate an energetic discharge from the Liver organ,
- help alleviate pressure stored within the Liver organ itself (detoxifying suppressed feelings), and
- purge the body of Excess Liver Heat.

It is important for the patients to understand, that when shouting and striking on the surface of the bag (while focusing their intention deep into the bag's center), they are releasing deep seated pathogenic rage. This is why it is prohibited for patients to substitute such things as pillows, bed, etc., for the bag. If a patient takes a pillow and pounds it, for example, screaming and shouting to release the aggression, he or she is releasing and discharging pathogenic Qi into the pillow. If after releasing this Toxic Qi into the pillow, it is returned to the bed for a person to sleep on, the patient will reabsorb the discharged energy. To avoid this occurrence, patients are encouraged to set aside a space where they can safely discharge internal pathogenic Qi (preferably outdoors). This is similar to isolating and storing garbage away from living quarters.

While striking the bag, in addition to feelings of anger and rage, sometimes the patient may start crying or experience overwhelming feelings of grief and despair. This is also a normal reaction to dispersing toxic emotion. It sometimes occurs while discharging Stagnant Liver Qi, because the Lungs, which store the emotion of grief, have the function of keeping the Liver organ in check, in accordance with the Five Elemental Controlling Cycle.

What the patient needs to perform this exercise is two wooden dowels (1–1.5 inches in diameter and three feet long), a large bag of rice from the grocery store, and some duct tape. Completely encase the bag of rice with the duct tape to prevent spilling. Place the bag on a table, or a hard surface, level with the Lower Dantian. Hold a dowel in each hand, strike the bag, and focus on the Liver's Qi traveling up the back, down the arms, and through the wooden dowels.

With each impact, imagine the body releasing Toxic Liver Qi into the bag. It is important for the patient to exhale completely and deeply from the lower abdomen when shouting and striking the bag. Sometimes the patient is encouraged to shout "no" or "stop." These words will help the patient to facilitate complete emotional discharging, while also reestablishing a patient's broken energetic/emotional boundary system.

Bag striking should be performed for 10–15 minutes to achieve maximum benefit. Afterwards patients may feel euphoric and exhausted due to the heightened endorphin activity now occurring within the tissues. If the patient still feels angry, he or she should try writing out the rest of the rage on paper, which is then discarded. In order to end the exercise, patients should sit for five minutes in quiet solitude. It is important for the patient to reflect on any emotional memories which have been released and brought back into consciousness. The patient may then begin sharing these feelings and memories with the doctor, and commence the process work for healing the unexpressed emotions.

Some patients may become absolutely horrified at themselves for getting so angry during bag beating that they feel terrified afterwards; especially if so much rage was built up that is seems there is no end to it. Often patients are forced to stop from overwhelming fatigue, aches, and pain caused by this purging exercise. If the patient allows the rage to build and never completely processes the rage, it may cause the patient to feel that without the anger he or she would literally cease to exist, or that he or she is innately evil. The patient may feel that he or she "is" the anger, and that the exercise is making it worse, and may strive to control the anger even more than before.

Some patients will get so mad, they will walk out on the Qigong doctor without processing anything. This unprocessed rage can be extremely dangerous, especially if the patient is prone to beating his or her mate, kids, or pets. Many people are completely unaware of how much anger they have suppressed. Some patients are like bottomless wells, especially patients with borderline personality disorder. Patients with this disorder should never be given this exercise, as it could lead to self-mutilation. Refer such patients out for psychotherapy.

In our culture, women are not allowed to express anger, much less rage. If the patient has ever been beaten, or witnessed a loved one being abused, he or she may believe that he or she is just like the abuser, and that he or she has been lying to him or herself about not being bad and may become suicidal.

The Qigong doctor should never prescribe this exercise unless he or she fully understands the complexity of the patient's circumstances. Some patients require a period of time, in a safe place, with supervision to help them deal with such intense emotions.

Other possible reactions might be that the patient goes into stronger denial of the rage, blames the doctor for causing him or her to beat their mate or children, or blames the doctor for causing the rage. The Qigong doctor must make sure there is sufficient time for the patient to process all the reactions aroused. The patient must not feel hurried.

ANGER MANAGEMENT

With some patients it is best to start them off journalizing the anger, not in a dairy, but on sheets of paper. They address the letters to the person, or persons, with whom they are angry. They continue writing until all the anger has been discharged. They are not to mail the letters, or give them to the addressee, or show them to anyone including the Qigong doctor. In writing the letter, they allow themselves to fully express the anger without judging themselves. The patient should not be concerned about grammar, foul language, etc. No matter how trivial the upsetting incident was, it should be written down. The flow of consciousness is initiated and continued until the end of the letter. No event (no matter how silly or insignificant) that comes to mind should be left unexpressed.

When finished, the letter is to be torn up, or safely burned in a ritual of healing and letting go (e.g., light some incense or candles, then blow them out after the paper has been incinerated). Symbolically the patient must let go of the anger for good, and that is why the letter is destroyed in a ritualistic act.

Only later, as the patient continues to practice this purging technique, slowly gaining control of the emotions, can the angry letters be kept for a few days and reread. In rereading the letters the patient engages his or her discriminating, judging faculties (i.e., "Was I right to get upset over this incident? Did I over-react? What did this incident remind me of from my childhood?"). This is a safe and very powerful method of discharging anger from the patient's body. The patient is encouraged and allowed to say everything he or she thinks of when writing, holding nothing back. It is nonthreatening, and allows the patient to come to understand him or herself better in time.

TREATING STAGNANT LUNG QI AND THE REPRESSION OF GRIEF

The Lungs dominate the Qi and respiration. A common problem observed in the clinic is stagnant, or obstructed Qi within the Lungs, from the toxic buildup of sorrow, grief, and worry. Sadness and sorrow can cause a Deficiency of Lung Qi, while worry can cause the Qi in the chest to stagnate. Stagnant Qi within the Lungs is usually directly or indirectly related to the suppression of emotional issues.

The right Lung is considered more Yang than the left Lung; it therefore tends to reflect, and sometimes manifests physically, any unresolved, or suppressed emotional conflicts experienced within the father-child relationship (or other main male figures in the patient's childhood). These unresolved emotional issues tend to carry over and help shape the nature of subsequent conflicts experienced in other male relationships. These ongoing emotional reactions provide a further opportunity in life to express, work through, and resolve these emotional issues.

The left Lung is considered more Yin and thus reflects emotional issues stemming from the mother-child relationship (or main female figure in childhood). The left Lung may likewise manifest these issues in subsequent female relationships.

Generally, cysts, tumors, or cancer originating in these corresponding regions reflect the emotional history. Because of these emotional patternings, the method of treatment for Lung diseases (that are Internally induced) is chosen according to the patient's syndromes in relationship to the root of the original trauma. After determining the specific Lung which dominates and manifests the patient's pathogenic condition, the Qigong doctor proceeds as follows.

1. For patients sensitive to energy, use Medical Qigong therapy while the patient is supine.
   - Disperse Qi over the patient's throat and Lungs;
   - Purge pathogenic Qi out the afflicted Lung Channel;
   - Tonify the patient's Spleen and Kidneys;
   - Prescribe Homework.
2. For patient's not sensitive to energy, use Medical Qigong therapy in conjunction with An Mo Qi therapy while the patient is supine.
   - Disperse Qi over the patient's throat and Lung area;
   - Purge pathogenic Qi out the afflicted Lung Channel;
   - Dredge Fire from the patient's Liver;
   - Tonify the patient's Spleen and Kidneys;
   - Prescribe Homework.

DRY CRYING TO DETOXIFY STAGNANT LUNG QI AND RELEASE TRAPPED SORROW AND GRIEF

The most common homework prescription given to patients with stagnant Lung Qi is the Dry Crying exercise. The concept of "dry crying" is best understood if we look back to when we were little children. When little children cry hard, they generally release full body resonances. They shake and undulate from the middle of the body, allowing the pain to resonate outwards. When children reach puberty, they learn (through parental or peer pressure) to suppress their feelings, and to restrain full body resonance, by tensing the muscles in the throat. The constant pressures from society, parents and peers, result in self-conscious feelings and shame regarding the loss of control over the emotions. The following exercise allows the patient to experience full body resonation and to release deeply suppressed emotions.

Begin the exercise from a standing position. Stand with the feet more than shoulders width apart, toes pointing straight and the knees comfortably bent. With each hand, touch the thumb to the index finger to form a circle. Relax the arms letting them hang naturally. Close the anal sphincter, and breath naturally from the abdomen. Tilt the head backwards, opening the mouth wide, and relax the jaw. Allow the upper body to lean and bow backwards (without straining the spine), but stretching the psoas muscles. Hold this posture from 5 to 15 minutes, concentrating on relaxing the body, and allowing it to resonate (Figure 53.24).

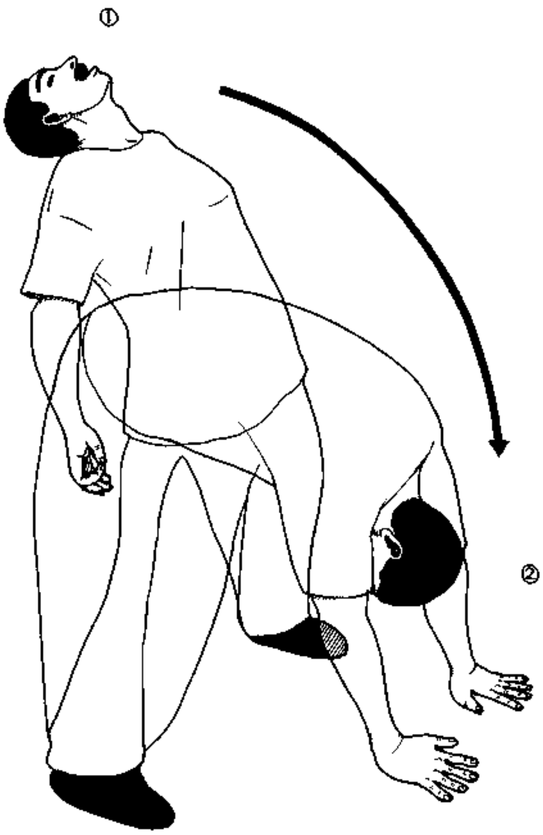

Figure 53.24. Dry Crying

The psoas muscles are located in the lower back and Mingmen area. They articulate from the transverse process of the lower lumbar vertebra. These muscles travel near the center of the abdomen down to the inner thigh. These particular muscles are associated with the storing and releasing of powerful emotions. As the psoas muscles stretch, the body begins to resonate and shake from the pelvic cavity upward. This action returns patients back to the preadolescent state of energy release, allowing them to detoxify deep seated grief within the tissues.

In order to complete the exercise, it is extremely important to ground. This prevents the uprooted toxic emotions from causing Qi stagnations in the throat and chest area. After holding the posture for the prescribed amount of time, slowly lean forward and tilt the chin down. Allow the arms to continue to hang naturally. Slowly lean forward until the hands touch the floor. Imagine Qi, flowing like water, from the torso out your arms, cleansing your body, sending the pathogenic Qi deep into the center of the Earth.

CONTRAINDICATIONS

Patients should avoid this particular exercise if they have previously injured their back. For patients who have difficulty with the stretching of the cervical or lumbar vertebrae, it is encouraged that they wear a cervical collar or lumbar support to prevent injury to the spine. Also, the exercise can be modified to allow the patient to sit comfortably on the edge of a chair while stretching the psoas muscles if the standing stretch is too difficult.

CHAPTER 54
QI EMISSION THERAPY FOR ONCOLOGY

HISTORY OF CANCER TREATMENTS

The ancient doctors of Traditional Chinese Medicine understood about the formation of tumors. The earliest records found were inscriptions recorded on "Oracles of Bone," unearthed from the ruins of the Yin Dynasty (16th–11th century B.C.). The word "ai" (cancer) first appeared in the Chinese Medical classic *Wei Ji Bao Shu* during the Song Dynasty, in the year 1171.

In Traditional Chinese Medicine, cysts, tumors, and cancers are considered multifaceted diseases, originating from imbalances in one or more of the five following sources: environmental, chemical, biological, physical, and psychological. According to the *General Treatise on the Etiology and Symptomatology of Diseases*, a tumor is one of the many diseases that responds to Dao Yin (Medical Qigong) exercises and meditations. To prevent the formation of cysts, tumors, or cancer, and to increase the strength of the immune system, Qigong doctors encourage patients to monitor their diet, maintain emotional stability, increase physical exercises through Medical Qigong prescriptions, and avoid exposure to pollutants.

From a Western Medical perspective, tumors and cancer are likewise considered multifaceted diseases. They are generally viewed as being created and formed through a constant bombardment, attack, and weakening of the tissues, caused by the following three factors:

- A congenital tendency towards specific types of tumors and cancers (inherited weakness);
- Exposure to radiation, environmental pollutants, and chemical poisoning, and
- An inability to repair weakened DNA, leading to uncontrolled replication.

In Traditional Chinese Medicine, cancer cells are considered "infant bodies" which absorb nourishment from the host (mother) body for their own growth; thus they are called "pathogenic or evil embryos." Cancer cells are actually components of the "original" or "mother" body.

CATEGORIZATION OF TUMORS

In Chinese Medicine, tumors and cancers are commonly referred to as "a retention of mass." Retention has the meaning of stagnation. Tumors and cancer originate from flowing substances within the body such as Qi, Blood, Phlegm, and food. These substances stagnate when they are attacked by pathogenic factors, and they gradually take on the form of a lump as a result of excessive accumulation. If the lump subsides spontaneously, it is commonly referred to as a cyst, or benign tumor. Many benign cysts and tumors, however, persist indefinitely.

BENIGN TUMORS

A benign tumor grows slowly, and can spontaneously stop growing, or disappear by itself. A benign tumor also grows in expansive ways. Its margin is clearly defined with firm boundaries, and is easily manipulated through palpation. There is no temperature differentiation between the skin surrounding a benign tumor and normal skin tissue.

Often, no symptoms surround the internal location of a benign tumor. As the tumor grows in certain areas of the patient's body, however, it can disrupt the functional activity of an affected organ. Benign tumors have been known to cause bleeding and inflammation. If the tumor is located within the endocrine glands, it can cause an abnormal function within the gland itself. A tumor located within an endocrine gland does not usually endanger the patient's life. Benign tumors can, however, become life threatening if the location compresses specific areas of the patient's brain or vital organs.

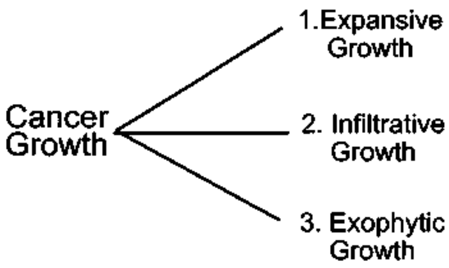

Figure 54.1. The Three Formations of Cancer Growth

MALIGNANT TUMORS

A malignant tumors grow very fast (potentially doubling every two months). Degeneration, festering, and hemorrhaging often accompany malignant tumors. A malignant tumor is infiltrative in its growth, with no clearly defined boundary. Its appearance is irregular and fastened (not easy to move through palpation). The skin temperature surrounding the malignant tumor is considered energetically Hot, as it is warmer than the normal surrounding tissue.

Most malignant tumors either metastasize to nearby, or distant areas of the body. The body's symptoms include fever, pain, anemia, fatigue, and general weakness. In many cases, if left untreated, they can cause death. Malignant tumors, and cancers can be divided into two main categories: a Carcinoma, and a Sarcoma.

1. A carcinoma is a malignant growth or tumor that occurs in the epithelial tissue (the outer surface or first layer of tissue that lines the body's cavities, as well as the principal tubes and passageways leading to the exterior of the body). Skin cancer, esophageal squamous cancer, adenocarcinoma of the Stomach, and breast cancer all belong to this type of carcinoma. A carcinoma may affect any organ, or part of the body, and may spread by direct extension, or indirectly through the lymphatic system and bloodstream.
2. A sarcoma is a malignant growth, or tumor, that occurs within the connective or mesenchymal tissue. It may affect the muscles, bones, fat, blood vessels, lymph system, Kidneys, Bladder, Liver, Lungs, Spleen, or parotid glands. Rhabdomyosarcoma (sarcoma of the muscles), osteosarcoma (sarcoma of the bones), and lymphosarcoma (sarcoma of the lymphatic system) all belong to this type of sarcoma.

THREE FORMATIONS OF CANCER GROWTH

According to Dr. Pan Mingji, Director of the Fuzhou Cancer Institute of Integration of Traditional Chinese and Western Medicine, there are three forms of cancer growth: Expansive Growth, Infiltrative Growth and Exophytic Growth (Figure 54.1).

1. Cancers of the Expansive Growth type consist of tumor cells crowded together on certain parts of the body. These cells proliferate, grow constantly, and expand towards the outside. The cells of the surrounding tissue form a wall to resist the tumor's expansion. The wall becomes the external membrane of the tumor, which has an evident boundary separating it from the normal tissue. Tumor growth begins in this way in the early stages of both benign tumors and malignant sarcomas.
2. Cancers of the Infiltrative Growth type consist of cells that not only group together, but, also separately invade the intercellular spaces, and tubular tissue of normal cells. They establish their foundation within the tissues, and proliferate daily, expanding outward with no boundary system between the cancer tumor and the normal tissue. Since there is no surrounding wall of membrane, the growth spreads in mass. Malignant tumors, especially carcinoma growth, infiltrate in this way.
3. Cancers of the Exophytic Growth type consist of cancer cells that occur on the skin, body cavity, or gastrointestinal mucosa; they often form on the surface of the body's tissues through papillary masses, projecting outwards on the skin and mucosa. Polyp, papilla-

shaped tumors, papilloma, gingival cancer, esophagal cancer, Stomach cancer, and intestinal cancer all grow in this way.

Benign tumors do not spread; they only grow in an expansive or exopathic way. The malignant tumor is different, however, as the cancer cells do not remain in their original location but expand from the mass outward. This expansion extends into other organs and tissues by way of the body's lymphatic plexus and blood circulation. This phenomenon is known as metastasizing.

ETIOLOGY, THE SIX PATHOGENIC EVILS THAT CAUSE CANCER

The following is an analysis of the causes of tumor formation. Traditional Chinese Medicine maintains the concept that a tumor is a systemic disease (affecting the whole body). It can be summarized as a disorder of the internal organs caused by emotional strains, and an imbalance of the viscera, that has allowed the pathogenic Evils to enter the body and obstruct the channels and collaterals. These pathogenic Evils cause problems, that may eventually lead to the stagnation of Blood and Qi and develop into tumors. The pathology of cancer is as follows: Qi stagnation, Blood stagnation, Retention of Toxic Dampness, Retention of Phlegm, Toxin Attack, and Weakness or Deficiency of Energy (Figure 54.2).

1. Emotional strains, depression of Liver Qi, or External pathogenic Evil factors can lead to a stagnation of Qi. Long term Qi stagnation can cause stagnation of Blood, which eventually causes clots. It is known that pathogenic Cold and Heat, and the prevailing pathogenic Evil factors, inevitably cause Blood clots.
2. Since Blood and Qi flow together, the obstruction of Qi can lead to an interference in Blood circulation. This interference gradually leads to Blood stasis; Blood Stasis is always further complicated by stasis of Qi. Blood stasis is sometimes referred to as "Dead Blood," due to the fact that the Blood is not moving. When Dead Blood accumulates in a certain area of the body, it eventually develops into an immovable mass or lump.

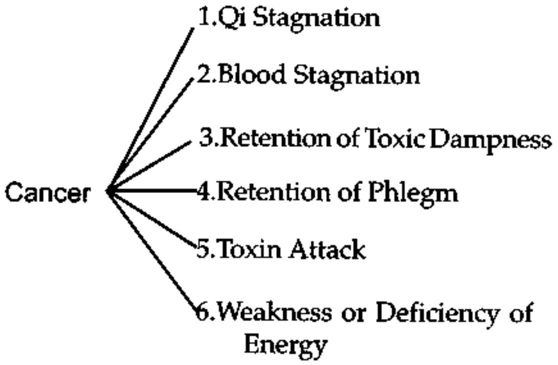

Figure 54.2. The Six Pathogenic Evils That Cause Cancer

3. A weakness of the Spleen and Stomach, as well as indigestion, can cause retention of fluid in the body. Long term fluid retention will produce toxic materials that can accumulate and form lumps.
4. Functional disorders of the Lungs and Spleen lead to poor digestion of watery food, and poor distribution of Body Fluids. Fluid retention causes Phlegm, especially when it is complicated by Excess Heat. If Phlegm is unable to move upward to be dispersed out of the body, it collects and stagnates. The stagnation of Phlegm in the Lung(s) leads to asthma. The stagnation of Phlegm in the Stomach causes gastric disorder and nausea. If Phlegm goes beneath the skin, it will take the form of a movable mass or lump.
5. Toxic Evils may come from poisonous external environments (exposure to chemicals, asbestos, ultraviolet rays, radiation, etc.) or may be generated by the stagnation of Excess Heat within the body. This Excessive Heat causes abnormal cell fusion and growth within the body, leading to cancer formation, and a systemic breakdown of normal cell growth.
6. Deficiency of both Qi and Blood, as well as a weakness of Yuan Qi, make the body more vulnerable and susceptible to the formation of new growth. When pathogenic Evils enter the body, they damage the Blood and Qi. If the Yuan Qi cannot be restored, the disease will lead to a vicious, destructive cycle, mak-

ing the individual more and more debilitated.

These six pathological changes may occur either alone or in combination: i.e., Qi and Blood stasis, Phlegm stagnation and Toxic Dampness, Phlegm stagnation and Blood stasis, a weakness of the body's Yuan Qi and strong pathogenic Evils.

The choice of which Medical Qigong therapy to use should be based on differentiation of symptoms and signs. Since cancers are primarily caused by emotional strains, strong pathogenic Evils, and a weakness of the patient's Yuan Qi, clinical therapy should be focused on tonifying the patient's Yuan Qi, and purging the Evil pathogenic factors.

EMOTIONAL COMPONENTS OF CANCER FORMATION

From the clinical perspective, if the cancer is internally induced through emotional suppression, there will be a series of emotional transitions the Qigong doctor will observe during its development. When there is an accumulation of negative emotions, for example, the patient's body begins to wall off this type of energy, and creates an energetic cyst, instead of dissipating and disposing of the Toxic Qi. The patient's suppressed internal emotions intensify the production and formation of the cyst, and feed the cancer cells, which result in an uncontrolled, rapid growth rate of these cells.

The body creates potential tumors and cancer cells every day. The macrophages or monocytes (large phagocytes) in the immune system aggressively attack and dissolve the cancer cells once the body becomes aware of them. Areas in the body, however, which are armored in denial, tend to attract and protect these distorted cells, allowing them to cluster and mass together.

Cancer cells can grow and accumulate wherever emotions are being repressed. The patient's denial maintains the constant unconscious flow of energy to these tissue areas. The phagocytes and other components of the immune system do not attack these toxic pockets of emotional energy, because of the patient's continuous denial of their existence. Consciousness and awareness exists even at the cellular level of the body. The continuous emotional energy bombarding the cells and tissues causes these tissues to go into an emotional state of shock. Emotional shock causes a release of endogenous opiates and corticosteroids, that deplete the white blood cells, and prevent the immune system from protecting the body. Because of the immune system's weakened state of resistance, infection and even the formation of cancer cells are now free to enter and coagulate within the tissue areas.

Scientific studies based in China, and the West, have shown that stress can stimulate the neuroendocrine system to the extent that it leads to depression of the immune system. Furthermore, the creation and growth of cancer can be influenced by stress. Removal of psychological stress alone can alter the neuro-immunologic functioning of the body to the extent that it can prevent and even change the course of cancer development.

In a patient with emotional blocks, chronic tension begins to interrupt the flow of Qi. If the emotional block is deep and long standing, this tension becomes constant. The capacity to express or discharge the trapped emotion becomes very difficult.

As the awareness dims, the natural communication between the conscious and subconscious mind breaks down. The original conscious decision to suppress certain emotions becomes a habit; it thus becomes an unconscious aberrant automatic function. When emotions are expressed, the conscious and subconscious mind are in harmony and synergized. When emotions are suppressed, the subconscious mind can no longer perform its job of healing the body; it diverts all its energy to shield the patient's denial system. Thus, the Shen itself, becomes misaligned with the patient's life purpose. The patient's posture, movements and lack of spiritual growth, reflect this misalignment. Changes also occur in skin temperature, body chemistry, and skin tone. The overall level of vital life force visibly drops.

Dr. Xu, at the Xi Yuan Hospital in Beijing, China, often pointed out that emotional upheaval, and the restriction of the breath, could cause the patient's energy to stagnate or deviate from its path. The deviated Qi draws more energy to its pathway, compounding the disease. The unstable emotional and mental focus of the patient further causes adverse effects to the body. Medical Qigong therapy and exercises expose the patient's deepest secrets, and thus reveal the underlying causes of the disease, which can then be treated. The patient's being is energized through his or her personal perceptions and belief structures. Therefore, it is necessary to change the patient's belief structure to increase the level of awareness and perceptual accuracy. The patient must be enabled to let go of suppressed emotions to fully experience deep emotional and spiritual changes. The Qigong doctor initiates the release of Toxic Qi to begin this healing process that can thereafter be maintained through appropriate prescriptions.

When cysts, tumors, or cancer are internally induced, they are formed from long-standing suppressed emotions which have become trapped within the tissue itself. These emotions usually consist of anger, guilt, grief, shame, and disappointment. The energetically formed cyst, or tumors, will remain walled off until the emotional energy is dissipated. Cysts, or tumors, can lodge themselves anywhere, penetrating deep into the major viscera causing visceral dysfunction; or they can superficially lodge themselves beneath the skin.

If the nature of the emotional energetic matrix (i.e., belief structure) is negative, the patient retains the energy of the injury, making it difficult to heal the trauma. If, however, the emotional matrix is positive, then the energy of the injury can be easily dispersed.

Once the Qigong doctor's energy reaches the cyst or tumor, the patient's energetic resonance supporting the tumor formation is interrupted. As the doctor continues to break up the cyst or tumor's physical mass, the frozen energetic pattern begins to unwind. The doctor continues to unravel the energetic mass until all the Heat has been released from the tissues, and the energetic pattern's unwinding is completed.

The underlying pathophysiology of healing can be broken down into four components:

1. The mind affects tissue matter through the emotions.
2. The emotions cause neuropeptides to be released.
3. The emotions (via the neuropeptides) influence the direction and movement of the immune cells (e.g., monocytes), sending them to various parts of the body.
4. The immune system (e.g., monocytes and phagocytes) heal the body.

The neuropeptide receptor sites are located on the immune cells, linking the body's immune system with the functions of the mind, emotions, and spirit. Neuropeptides are produced in response to the individual's state of mind and emotions; they influence the state of health of all the internal organs.

Through Dynamic Medical Qigong training, neuropeptides, in the form of endorphins, are released within the body's system. This release of energy and endorphins helps to facilitate the patient's healing.

Through the use of Medical Qigong therapy, exercises, and meditations, areas of denial and subconscious energetic patterning can be altered. The energetic matrix which supports the tumor(s) or cyst(s) thus dissolves.

When a healing session occurs, the Qigong doctor's external energy field envelops the patient's tissues, thus allowing the patient's cells to receive new information to reprogram the toxic energetic patterns. This reaction occurs providing the patient's tissues are open to receive the informational structures needed to either rebuild them, or disperse the disease's cell formation.

CANCER FORMATION AND THE PATIENT'S MIND, EMOTIONS, AND SPIRIT

Emotions strongly influence the increase or decrease of tumor size; therefore, it is important to regulate the patient's mind, emotions, and spirit. When the patient's mind, emotions and spirit are free from distractions and are in harmony, the patient is able to manage negative emotions. Regulating the Mind (whole body awareness and consciousness) is considered the primary Medical Qigong technique used for teaching patients how to avoid creating the dark abyss of negative emotions.

The patient must learn to let optimism prevail over pessimism. The practice of Medical Qigong not only cultivates vital energy and builds up the patient's health, but promotes the development of optimism through creating a quiescent mind. Tranquility and optimism improve circulation of Qi and Blood (which strengthens the immune system) thereby inhibiting cancer cells until they gradually perish.

While visiting in Maui, Hawaii, I came in contact with four people who had been diagnosed with terminal cancer, and had a life expectancy of two weeks according to their physicians. Each person had their own unique story as to how they had come to the Hawaiian Islands to die. What struck me as fascinating was the fact that each one of these individuals had either left successful but extremely stressful businesses, or long term abusive relationships. Each made amends the best they could with relationships from their past (those who had anger against them or those whom they had hurt). They had come to paradise to live their last weeks in peace. That was several years ago! It seems that upon resolving all personal conflict, and leaving their stressful environment, their cancers went into spontaneous remission. Each individual was now concentrated on enjoying the pleasures of a simpler life, and on focusing on their own spiritual growth.

STRESSFUL LIFE ACTIVITIES AND THE IMMUNE SYSTEM

Stressful life activities create a flood of adrenaline within the body. This occurs not only from the endings of the vertebral lumbar division at the core segment of the major internal organs, but also from the core of the adrenal gland in the medulla (which specifically is a sympathetic ganglion, and responds to stress by releasing adrenaline into the blood stream). Adrenaline spreads throughout the body and activates not only the body's sympathetic nervous system, but also the adrenal glands; the heart rate, respiratory rate, blood pressure, and Blood flow to the muscles are increased in response to the Fight or Flight Syndrome.

When the release of adrenaline causes the mind and body to react, the immune system must then seek to balance the chemical imbalances created within the body, before it can return to its normal job of monitoring cellular activity, and defending against pathogens. If a patient is already vulnerable to a specific disease, added stress causes the patient to get sick quicker. Meditation takes the patient out of the stress mode, deactivates the adrenal glands, and normalizes the function of the immune system.

An unstable life style is very stressful and can lead to depression and Qi deviations. Conflict produces stress which causes anxiety; chronic stress wears down the mind and body, which leads to anxiety, or depression. Stress begins with an overexcitement of the nervous system, slowly working its way deep into the body like wet cement. It causes the shoulders and upper back to become tense and rigid. Stress can result from happy occasions (such as weddings) and other normally pleasant events. When the patient's life style is not properly balanced, over time Qi deviations, and illness can result. The capacity to bear stress is related to the strength of the nerves. When stress levels surpass the nervous system's capacity to handle them, the nerves "break down"; this results in all sorts of physical, mental, and emotional disturbances. These disturbances, if unheeded, can eventually lead to organ malfunctions, cysts, tumors, cancer, and premature death. To avoid such states, a balance between work and rest must be maintained.

At a conference of the International Society for Neuro-immunomodulation (*The Effects of Stress and Depression on Physical Disease*), Dr. Philip Gold, of the National Institute of Mental Health stated that stress and depression send hormones (e.g., cortisol) flowing into the bloodstream. An imbalance of hormones, if unregulated, can:

- destroy the appetite,
- cripple the immune system,
- shut down the processes that repair tissue,
- inhibit sleeping patterns,
- break down bone density,
- aid the onset of infection, and
- instigate the growth of cancer cells.

An individual's emotional past determines his, or her, present state of health, both physical and emotional. Past emotional belief structures are responsible for both creating and healing diseases. Just as thoughts drift through the mind (but do not become active until they reside within a belief structure), so emotional wounding resides within the tissues and cells waiting to be activated by emotional upheavals. Our thoughts become form via choices, and manifest within the body.

In most societies, sharing traumatic wounds is often the first act of intimacy and bonding. Traumas provide a common ground for relationships, and support the pain structure that cements the relationship. People become addicted to empowering wounds, which in turn supports and empowers the disease. The exchange of painful personal traumas is reinforced by receiving sympathy and strong empathy, which binds rather than releases the pain from the tissues. Desire for sharing one's pain in order to receive love can lead to a habitual way of relating to those closest and dearest. When such individuals feel taken for granted, or unappreciated, they may immediately resort to sharing past traumas to illicit the desired response. When such techniques fail to elicit the desired response, a new wounding is experienced. Many of these individuals will continue to unconsciously create new traumas, in the hope that, this time, the desired outcome will be achieved. This in turn can become an addiction, empowering old wounds as it creates new ones, including illness. Thus, the spirit becomes divided to support the new belief structure (i.e., that pain and misery grin love and attention), and the life-force energy deviates from its assigned job of protecting the body, which creates illness. Setting the self up for new wounds is not a conscious process, but an unconscious one, which needs to be brought into the patient's awareness in order to be healed. Then and only then can the mind and spirit act as one unit, directing the body's Qi towards its proper function of healing.

Because energetic healing occurs in the present, not in the past, victims of past traumas will never heal if the energy continuously regresses to the armored traumatic belief structure.

USING MEDICAL QIGONG TO STRENGTHEN THE IMMUNE SYSTEM

Medical Qigong increases leukocyte production, stabilizes the body's blood pressure, and improves the immune function. The immune system is not so much an independent system, but is rather an interactive functional system throughout the body. Immune cells which live in the Blood and lymph fluid are commonly known as leukocytes (white blood cells); lymphocytes are a type of leukocytes that is created in the Spleen. The purpose of these cells is to work together to defend against pathogenic factors. The function of leukocytes and lymphocytes is to identify, ingest, destroy, and eliminate disease-causing organisms, as well as remove all dead, damaged, or irregular cells.

Immune cells are produced in the lymph tissue and red bone marrow, and are stored in the lymphatic organs and the Blood. To perform the function of protecting and cleaning, immune cells must circulate throughout the entire body through the capillaries and vessels that carry Blood and lymph fluid. Immune cells travel from the lymph nodes and the lymphatic organs of the thymus gland, Spleen, and tonsils, which are responsible for filtering the lymph fluid to remove impurities, and produce more lymph cells.

For an External pathogen to affect the body, it must first pass through protective layers of cells. Generally, the macrophages present within the tissues ingest the pathogens and dead tissues. Macrophages are reinforced by the action and movement of the body's Wei Qi. When a pathogenic intruder invades the body, the macrophages are often the first to attack the intruder. If the macrophages are not strong enough, more Blood and fluid are brought to the area, which contain more diverse immune cells. The increased fluid engorges the tissue, which engenders more cellular activity. The increased Blood and cellular metabolism increases the local temperature. The immune cells break down and ingest the pathogens, creating a process which continues until the disease is eliminated.

The dynamic movements of Medical Qigong

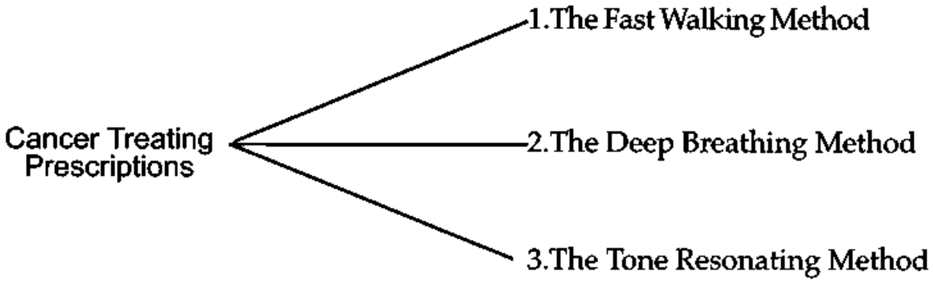
Cancer Treating Prescriptions
1.The Fast Walking Method
2.The Deep Breathing Method
3.The Tone Resonating Method

Figure 54.3. Three Prescriptions Used For Treating Cancer

exercises generate Heat, soften the body's connective tissues, and encourage the flushing of the intercellular fluid throughout the body's tissues and organs. Energetic Qigong practices increase oxygenation in the Blood and improve circulation, thereby enhancing leukocyte production, and strengthening the body's immune system.

CANCER TREATMENT METHODS

Tumor and cancer cells make their own room within the body's tissues by crowding the normal tissues. The tumor and cancer cells grab nourishing energy at the expense of the host, thus damaging the patient's body.

Medical Qigong cancer treatment methods are techniques designed to purge the energy that supports the tumor. These techniques are based on the fundamental Qigong patterns mentioned previously. The goal of these patterns focuses on treating the patient's Excess syndrome first, by purgation. Purgation is then followed by tonification, then regulation. The rapid growth of healthy energy, and the fast expulsion of Evil factors, should lead to normal body function restoration, and create more vigorous metabolic activity. More nourishment can be absorbed, delivered and distributed throughout the body, leaving no room to feed the tumor cells.

In Medical Qigong clinics, tumors are viewed as being similar in their expansion and growth to that of the energetic pattern of a plant's Qi flow. When tracing the origin of a tumor, the Qigong doctor begins at the tumor's location (considered the flower or fruit of the disease) and traces its energetic branches to the organ of origin (the actual root of the disease). If the disease is internally induced, the original source, i.e., the dysfunctional internal organ, feeds the tumor through suppressed emotional trauma.

To address this problem, the doctor prescribes the following three cancer treatment methods as homework, in addition to Medical Qigong treatments and herbal therapy: the Fast Walking Method, Deep Breathing Method, and the Tone Resonation Method (Figure 54.3).

1. The Fast Walking Method stimulates the electromagnetic fields within the tissues and reinforces the patient's Yuan Qi. By walking fast and changing the normal breathing patterns, the cancer cells can be destroyed. This is because the dual action of shocking the system with more Qi and Blood flowing through the body tends to stimulate the electromagnetic fields within the tissues, causing the tissue cells to become stimulated and the immune system to be aroused and thus enhanced.
2. The Deep Breathing Method tranquilizes the patient's mind, and enhances their Yuan Qi. The Deep Breathing method is used to treat various cancers due to mental and emotional factors, by changing the patient's normal breathing patterns.

By changing the normal breathing into rapid exhaling, slow deep exhaling, or relaxed toning, Blood clots and tumors can be dissolved. This breathing technique has a special effect on reinforcing the patient's Yuan Qi. When prescribing Medical Qigong therapy, it is important to match the patient's Yin disease with a Yang method of treatment, and a Yang disease with a Yin method of therapy. For example: rapid exhalation (a Yang method) would be used when treating an abdominal mass caused by a Cold (Yin) condition.

3. The Tone Resonating Method is utilized to purge any residual Evil pathogenic factors from the patient's body.

When first starting the Medical Qigong Cancer Tone therapy, a patient's practice time should be short. The performing time can be extended gradually with the increase of the patient's physical strength. To bring the Medical Qigong exercises to peak efficiency, a patient should practice a minimum of four hours (including 15 minute breaks), and up to six hours each day. The patient can take a break after half an hour to avoid fatigue.

Through various kinds of breathing and fast walking therapies, Qi can be energized to flow so vigorously through the Conception Vessel that the patients' Yin and Yang is balanced and their Lung Channels are strengthened. In the course of treating the tumor or cancer, True Qi (Zhen Qi) is fostered and the immunity of the body strengthened. Research from China proves that persistent training, vigorous enough to increase a person's heartbeat up to 70-80% of its maximum rate, helps dissolve Blood clots. Other reports state that long distance running can be effectively used to cure cancers. These reports may help explain why fast walking, and breathing methods, are effective for preventing and treating cancer.

SOCIAL ONCOLOGY—ESTABLISHING A CANCER SUPPORT GROUP

In China, cancer patients join support groups to practice Medical Qigong. This form of treatment is called social oncology; it was created by the Cancer Recovery Society for the purpose of supporting human interaction and the recovery of health by cancer patients.

Cancer recovery groups meet daily in parks throughout China, practicing self-healing (in particular, Guo Lin's Cancer Recovery Qigong and Walking Qigong Therapy). After practice, the groups gather together for about an hour to socialize in a nearby tea house. The patients sing songs, tell jokes, laugh, introduce new members, read poems, and tell stories that highlight the possibilities of recovery from cancer. As part of the social meeting, the group regularly sings "Happy Birthday" to each patient who has an anniversary after his or her cancer diagnosis. Laughter becomes a top priority in social oncology, and is considered the strongest medicine for healing.

According to Dr. Roger Jahnke, O.M.D., the advantages of social oncology are three fold: stress reduction, positive reinforcement of hope, and the establishment of deep interpersonal relationships.

1. Stress is reduced through socializing rather than isolating and fretting. The social healing aspect of the patient's recovery program serves as a constant reminder that the physical healing is enhanced when he or she becomes lighthearted and free of worry.
2. Increased hope and faith is achieved through the power of personal testimony. Each day during the social healing session, patients hear stories of close friends whose recovery process proves the benefits of healing from various forms of Medical Qigong.
3. The value of love and encouragement from friends mitigates depression, fear, and worry. Isolation is like a food to cancer; social oncology provides fun, encouragement and a deep, caring interpersonal connection within the recovery group.

SOCIAL ONCOLOGY WITHIN THE HOSPITALS

In several Medical Qigong hospitals in China, when the patients first check-in, they are assigned a support group. The support group consists of previous and current patients, who have survived the same type of cancer as the entering patients.

To assist patients in healing and to give them every emotional advantage during recovery, visitors, phone calls, letters, TV, and newspapers are forbidden for six weeks; this ensures a controlled environment in which neither family nor friends can adversely affect the healing process (through critical comments or attitudes). Patients are immersed in a support group full of caring, motivated individuals all working towards the same goal. Furthermore, patients who have been there the longest share their success stories, boosting everyone's moral.

The concept of clinical social oncology is based on the belief that the patients' high spirits affect the emotional perspective, and thus mental thinking. These emotional and mental changes alter the physical body. It was recently discovered that the group of breast cancer patients, who participated in the Stanford Hospital Cancer Support Group, lived twice as long as the women who refused to join the group. This difference in longevity was attributed to the love, support, and social aspect of the cancer support group. This recently discovery has given rise to the increase of several support groups in the United States now being established for patients recovering from AIDS, heart attacks, and many other illnesses including several types of cancer.

CANCER RESEARCH

Noted research scientist Feng Lida pioneered cancer research in China by showing that emitted Qi from Qigong masters produced marked changes in cancer cell cultures from mice. Several studies reported the effects of emitted Qi on tumors in animals. The emission of Qi, for example, was reported to inhibit the growth of implanted malignant tumors in mice, but did not destroy the tumors. Encouraged by the results with animals, researchers carried out clinical research on the effects of Medical Qigong on human subjects with cancer.

In one study, 127 patients with medically diagnosed cancer were divided into a Qigong group of 97 patients, and a control group of 30 patients. All patients received drugs, and the Qigong group practiced Medical Qigong for more than two hours a day over a period of 3 to 6 months. Both groups improved, but the Qigong group showed improvements four to nine times greater than the control group for strength, appetite, and weight gain. The Qigong group also did not experience bouts of diarrhea. The phagocytic rate, which is a measure of the immune function, increased in the Medical Qigong group, but decreased in the control group. These cancer treatment results were gathered and compiled from clinical research by Dr. Kenneth M. Sancier, available in the first Appendix of this text.

MEDITATIONS FOR CANCER PATIENTS

The following meditations are frequently used within the Medical Qigong clinic to aid healing. As previously discussed, it takes focused concentration to initiate an energetic transformation. Focused concentration is achieved through proper imagination, visualization and positive affirmation; it is practiced to reprogram the body's energetic cell patterns, and initiate healing.

DIVINE HEALING LIGHT MEDITATION

A very effective Medical Qigong meditation called Divine Healing Light is practiced by cancer and tumor patients throughout China to dissolve and disperse energetic pathogens. This meditation begins as follows:

Begin in a seated posture. Relax, and feel a quiet comfort. Imagine yourself seated on a quiet mountain overlooking the ocean, resting under a clear, dark blue evening sky. At the rim of the ocean, the full moon begins to rise. It continues to ascend until it faces you, suspended in the sky above your head. Slowly the moon begins to shrink, becoming smaller and smaller, yet brighter and brighter until it becomes the size of a small luminous pearl. This brilliant white pearl begins to pulse and expand, as it grows it opens to form a large orb, vibrating with divine power. From the center of the divine power flows unconditional love, healing, and compassion.

Breathe in this vibrant, healing light; allow it to penetrate deep into your body, permeating every cell, illuminating and dispersing all illness. Open fully to the divine healing. Feel the love and compassion enveloping you, and penetrating all the diseased areas of your body. Continue to breathe the healing energy in through your nose and exhale out through your mouth (imagining all the unclean energy descending into the ground). Continue this breathing pattern until all the negative energy has been expelled and replaced with divine love and healing light.

PULLING OUT THE PAIN MEDITATION

This meditation is an effective treatment prescription for treating many illnesses, including cancer and tumors. The goal is to eliminate all toxic belief structures that stand in the way, and sabotage the patient's healing. Gentle acceptance of the illness and emotional pain opens the pathway to understanding and unveiling hidden traumas. Force does not work in trying to change the tumor or cancer energetic patterns. Only honesty, respect, tenderness, and understanding will begin the healing process. As pain and suffering are allowed to surface into consciousness, the immune system is able to recognize diseased cells, and can mobilize an army of immune cells. These are then able to effectively identify, dissolve, and remove any tumors and disease.

The doctor instructs and guides patients through this meditation matching the rhythm of the words, to the patients' exhalations. Patients draw their painful memories and traumas out of their body in the form of dark smoke; they then separate the distilled wisdom and knowledge from this dark cloud, reabsorbing the healing light into every pore and cell in their body. The dark cloud then is released with the help of an angelic being who severs the energetic cord(s) that still attach the patients to their pain and suffering. The doctor assists the complete severance by clapping his or her hands as the patients visualize the "Sword of Truth" (held by this holy being), severing the dark energetic cord(s). The sword of truth represents the patients' intent to regain their power by letting go of their attachment to their illness. This attachment created very real energetic cords that bound the patients to their painful memories and toxic emotions. Severing these cords, or cord, symbolizes the end of the patients' investment in their disease, and the beginning of true healing. All the patients' pain and despair is released, as the black cloud ascends into the Heavens, and into the hands of God (or a higher power, in accordance with the belief structure of the patients). As the cloud reaches the first wave of love and compassion emanating from God (or a higher power), a mighty explosion transmutes the black despair and pain, into divine healing light. The doctor then assists the patients in accepting this healing, loving light back into their body, to experience, perhaps for the first time in their lives, true forgiveness and a state of grace.

1. Begin by sitting comfortably in a chair, with both feet on the floor, hands resting on the thighs, eyes closed, tongue on the upper palette. Breathe naturally through the Lower Dantian.
2. Focus your attention on the Yellow Court area below the diaphragm where you store painful memories and traumas. Imagine opening up this area to release the toxic energy out of your body as a stream of dark steam releasing from a pot.
3. As the dark steam flows out of your body you release feelings and memories of guilt, anger, rage, humiliation, abandonment, degradation, rejection, insecurity, and sorrow. These toxic emotions have, until now, prevented you from healing through creating a deep distrust of yourself and others.
4. Focus your attention on this energetic cloud, and begin separating the pain and hurtful memories from the knowledge and wisdom gathered from these experiences. Visualize this knowledge and wisdom in the form of golden, white, and silver light energy gathering on the right side of the room. Continue to drain the dark cloud of misery and pain, now occupying the left side of the room, until you have extracted all the new insights from it. As you do so, you notice the dark cloud becoming heavier and darker.
5. Focus on the right side of the room. Through you intention, begin to inhale and imagine this bright, illuminating energy flowing back into your body. Absorb this knowledge and wisdom gathered from past experiences, void of any feelings of hurt, pain, or judgement, into every cell of your body. This distilled knowledge and wisdom empowers you to heal from

your wounds on a physical, mental, emotional, and spiritual level.

6. Next focus your attention on the dark black cloud containing all your pain and suffering. Imagine an angelic being, righteous and holy holding a "Sword of Truth" standing by your side, ready to sever the dark energetic cords still connecting your physical body to this dark cloud of despair. The sword descends and severs the cords (the doctor claps his or her hands at this moment).
7. The cloud of darkness begins to float up, through the ceiling, through the sky into space. Far in the distance you begin to perceive God, or a higher power, emanating waves of compassion, love, and mercy descending towards the Earth. One of the waves touches the ascending dark cloud of hurt and pain, exploding it into a brilliant light. The cloud is immediately transmuted into fluorescent crystal blue drops of light. This pure, clean energy descends from the Heavens like a gentle rain. Breathe in this crystal blue healing energy, let it penetrate deeply into every pore, every tissue and cell, saturating your body completely. Feel the light cleansing, healing, and radiating throughout your being.

SUMMARY

Cancer is defined by Traditional Chinese Medicine as, "an obstruction of Qi and Blood circulation resulting in stagnation." The root cause for cancer is the breakdown and stagnation of the patient's Yuan Qi. Once cancer becomes established within the patient's body, or metastasizes, it is difficult to heal.

In order for cancer to form, it must elude the body's immune system, undergo many divisions, and produce countless generations of cells without resistance. By the time the cancer cells have formed into a cluster mass, a fixed energetic pattern has been established to support its existence and survival.

Only when the patient's Yuan Qi is returned to its original state, can the body begin to heal itself. Medical Qigong Cancer treatment methods strengthen the True Qi which in turn strengthens and cultivates the Yuan Qi. In this manner, the body is returned to its natural state of Yin and Yang balance.

It is the author's hope that these effective methods of treating cancer will soon be available to the American public. To this end, I have established a three-year master's and plan to offer a five-year doctorate program in Medical Qigong Science at the Five Branches Institute, College of Traditional Chinese Medicine, in Santa Cruz, California.

TREATING TUMORS AND CANCER WITH MEDICAL QIGONG THERAPY

The following are several Medical Qigong techniques and treatment modalities, currently used in China, to treat different types of energetically induced cysts, tumors, and cancers. It is advisable to accompany the prescriptions with various herbal prescriptions.

BRAIN TUMORS

A brain tumor is a generic term used to describe any intracranial mass, which includes: neoplastic, cystic, inflammatory (abscesses), or syphilitic.

ETIOLOGY

There are many causes for intracranial tumors. The most common cause is Kidney Yin (Water) failing to wash over the brain to disperse the pathogens. This condition can be caused by Excess Liver Fire dissolving the body's Kidney Water.

SYMPTOMS

Intracranial tumors produce symptoms primarily by two mechanisms: the infiltration and destruction of the normal brain tissue; or the tumor mass affecting the surrounding tissue, which causes edema, thus increasing the intracranial pressure.

General symptoms include headache (due to intracranial pressure), changes in the retina, and vomiting (without nausea). Additional symptoms include mental and emotional changes such as dullness, giddiness, or epileptiform convulsions.

TREATMENT

1. The doctors first step is to analyze the syndrome and the patient's emotional outlook.

When treating brain tumors, the growth of the disease must first be stopped. Check the patient's external living conditions, and do a psychological profile to determine any detrimental patterns.

2. Dredge and disperse toxic Heat out of the patient's Yang channels. Remove any Excess Liver Fire by drawing it out the Liver organ; then dredge the patient's Liver Channels, dispersing toxins out through the Lv-3 points.
3. The doctor must reverse the process of tumor development to dissolve the tumor. Instruct the patient to visualize energy from the Lower Dantian rising up to the tumor and transforming the mass back into Qi. The tumor growth must be reversed by first converting the solid tumor into soft tissue, then into fluid Qi that is then dispersed.

   In cases of cranial thrombosis, tumors must be dissolved slowly; otherwise other tumors will be created. The tumors must be softened, then slowly dispersed to avoid any recreation of the energetic channels' dysfunctions.
4. Extend Qi into the patient to reinforce the Kidneys' Qi by extending Qi into the patient's feet via the brain tumor point, Shihmein, at the bottom of the heels (Figure 54.4). Next, lead the patient's Kidney Water up into the brain to rinse over and slowly dissolve the tumor.
5. Finally, regulate and balance the patient's Yin and Yang energy.

PRESCRIPTIONS AND HOMEWORK

1. Have the patient practice the healing sound "Duo" (10 times in a straight and 10 times descending / ascending tone) five times a day.
2. Assign tonifying exercises for the Kidneys and Heart.
3. Prescribe herbal teas to enhance the patient's Kidney Yin.

BREAST CANCER

Mammary cancer is one of the most common types of malignant tumors in women; it is considered a malignant neoplasm of the breast.

ETIOLOGY

There are multiple factors associated with breast cancer. These factors can be categorized under genetic family factors, hormonal factors, dietary factors, environmental factors, and benign breast disease. Heredity can play an important role, especially if the patient's mother (or close relative) had bilateral breast cancer diagnosed prior to menopause. Also, a history of chronic breast disease (especially epithelial hyperplasia), or obesity can predispose the patient to breast cancer.

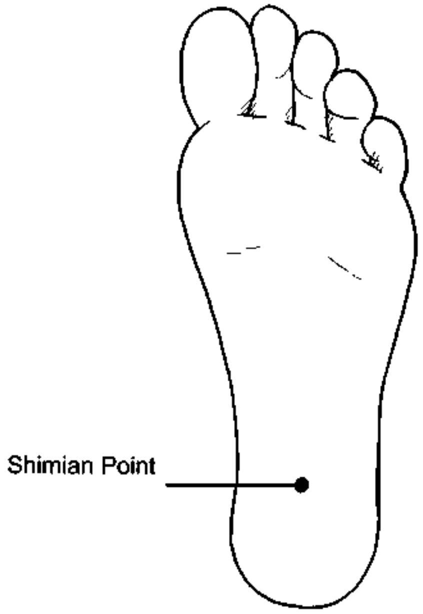

Figure 54.4. The Qigong doctor will emit Qi into the patient's "Brain Tumor" point (also called the Shimian, meaning insomnia point) located at the bottom of the patient's Heel.

SYMPTOMS

Symptoms consist of hard lumps which lack mobility, and progressively increase in size. The hard lumps may be concave, convex, or protrude; or the nipple may be inverted. Breast cancer is usually related to stagnation of the Liver's Qi, stasis of Blood, Phlegm and Toxic Heat, as well as weak Kidneys. The emotional components which lead to breast cancer are suppressed grief, worry, sorrow, and depression.

TREATMENT

1. Purge and eliminate the stagnation from the breast tissue (Figure 54.5).
2. Subdue the Rebellious Qi.
3. Clear the Liver and tonify the Kidneys. When treating breast cancer, the doctor should use the Vibrating Palm in conjunction with projected Tone Resonation (projecting both vibration and the "Ni" tone resonation into the cancerous area).
4. Tonify the patient's Conception and Thrusting Vessels, and end with the Microcosmic Orbit regulation (Fire Cycle).

UTERINE CANCER

Uterine cancer is considered a malignant neoplasm of the uterus.

ETIOLOGY

The etiology of uterine cancer is unknown. Possible causes of the disease may be:

- chronic stress
- toxins
- suppression of emotions
- pathological changes of the endometrial hyperplasia, or
- cervical carcinoma spreading onto the uterus.

SYMPTOMS

A uterine carcinoma is generally more common in women after menopause, roughly 75% of uterine cancers occur in postmenopausal women, 15% peri-menopausal women and 10% of menstruating women. The main clinical manifestations include pain in the lower abdomen, waist and thighs, abnormal uterine bleeding with a bloody, purulent discharge.

TREATMENT

One Medical Qigong therapeutic technique used in China to treat uterine cancer is as follows.

1. Purge the diseased area of stagnant Qi using the Thunder Palm technique in conjunction with the Vibrating Palm technique.
2. Project the sound "Yu" into the uterus to disperse Heat from the cancer cells.
3. Disperse Excess Heat from the patient's Liver and uterus, leading the Toxic Qi out the patient's body via the right Gall Bladder

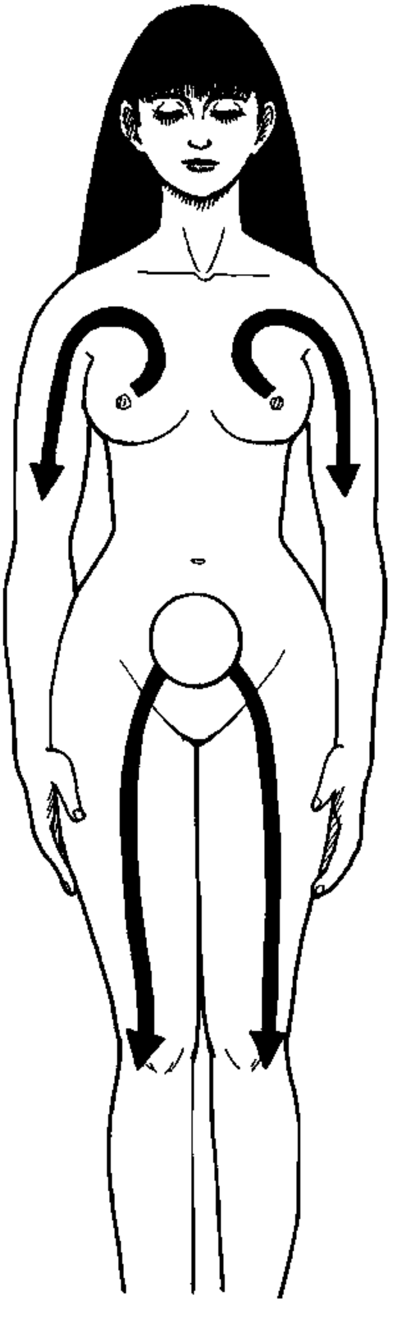

Figure 54.5. After Purging the pathogenic Qi from the patient's upper body, the Qigong doctor continues to disperse the patient's Toxic Qi from her lower abdomen, leading it down her legs and out her St-36 points.

Channel.

4. Fill and tonify the patient's Lower Dantian and Kidneys with Qi. Circulate the energy through the Microcosmic Orbit to improve Qi and Blood circulation through the Governing and Conception Vessels.

PRESCRIPTIONS AND HOMEWORK

The Conception Vessel moves Qi in the Lower Burner and uterus; it is the primary Vessel used in treating uterine and cervical cancer. Sexual activities are therefore prohibited from the start of the treatments, until treatment is no longer needed.

1. Have the patient practice the Descending the Yang and Ascending the Yin Technique (see Chapter 41), ending with the healing sound "Yu" for 24 breaths, 9 times a day.
2. Prescribe the Fast Exhaling Breathing Method along with the Cancer Walking methods for treating Kidney disease (see Chapter 15).
3. Have the patient practice the method of Taking in the Blue Qi, three times a day for 15 minutes each session, to strengthen the Kidneys.
4. Have the patient practice the Filth Dispelling meditation, while focusing on her uterus. This meditation is practiced as follows:
    a. From a sitting posture, with the eyes closed, and the body relaxed, place the tongue up against the upper hard palate, behind the teeth.
    b. Breathe naturally and evenly.
    c. Imagine Qi whirling in through the upper portion of the uterus, circulating clockwise, building momentum as the energy flows downward to the base of the uterus.
    d. Once the energy has reached just above the cervix area, imagine the whirling Qi absorbing the noxious Heat and Toxic Qi from the uterus, transforming it into wind, and dispelling it out the vagina. Repeat this sequence 10 times.
    e. Next, gently close the vagina and anal sphincter, then inhale, and imagine the Qi flowing into the body through the vagina. As this energy flows into the uterus and Lower Dantian area, it transforms into wind. This wind circulates in a counterclockwise direction, spiraling up the body, through the Thrusting Vessel and exits the body through the mouth with each exhalation. Repeat this visualization 10 times.
    f. Finally, focus attention on the Lower Dantian, imagine the Qi returning to its origin, and rub the abdomen 36 times in a clockwise direction to end the prescription.

CERVICAL CANCER

Cervical cancer is considered one of the most common malignant tumors in China. It is considered a malignant neoplasm of the cervix of the uterus, and can occur at any age.

ETIOLOGY

Cervical cancer is suspected to be an oncogenic papilloma virus that is transmitted sexually. Risk factors include:

- sexual intercourse at an early age,
- multiple sexual partners and promiscuous male sexual partners (especially those whose previous partner had cervical cancer),
- serious cervical lacerations caused from the penis being too long, or the insertion of foreign objects,
- unsanitary sexual relationships,
- closely spaced births, and
- a family history of cervical cancer.

Excessive sexual activity predisposes young teenage girls to cervical cancer, because, with the onset of ovulation and the changes in vaginal PH, active squamous metaplasia is taking place in the cervix. During this time of cellular immaturity and vulnerability, a carcinogen is most likely to have an influence on the squamous epithelium; this can predispose the teenager girl to cervical cancer later on in life.

THE DEVELOPMENT OF CERVICAL CANCER

There are two main types of cancer of the cervix. The squamous cell carcinomas make up 85-90% of cervical cancer, the rest (a small 10-15%)

are developed from a malignant adenoma arising from a glandular organ known as a adenocarcinoma.

Cervical cancer begins with the growth of abnormal precancerous cells and tissues called dysplasia. Dysplasia is first detectable in the squamous epithelium of the cervix (the muscle lining of the cervix), and spreads to create a tumor called a cervical intraepithelial neoplasm (CIN). The cervical intraepithelial neoplasm progresses through three levels of severity, from mild, to moderate, to severe. At the severe level of CIN, the precancerous growth turns malignant; it becomes a carcinoma in situ (localized). If left untreated, the carcinoma progresses through four more stages of development, each of which is subdivided into two grades of severity. These grades indicate the location and spread of the carcinoma. The four stages and their subdivisions are described as follows:

In the beginning stages of the disease, cervical dysplasia is characterized by three stages of abnormality and severity (CIN-1,-2,-3), which later progress onto four stages of invasive carcinoma (Stages I through IV).

- CIN-1: Corresponds to a mild condition of dysplasia.
- CIN-2: Corresponds to a moderate condition of dysplasia.
- CIN-3: Corresponds to a severe condition of dysplasia, and carcinoma (in situ). The cancer, however, is found only in the layers of the cells lining the cervix, and has not yet extended deeper into the tissue lining.
- Stage I-A: Corresponds to a microinvasive carcinoma. The cancer is beginning to spread from the lining of the cervix into the deeper connective tissue.
- Stage I-B: Corresponds to an invasive carcinoma confined to the cervix. The cancer has now spread from the lining of the cervix into the deeper connective tissues.
- Stage II-A: Corresponds to a tumor extending beyond the cervix, into the upper third of the patient's vagina, but is still inside the pelvic area.
- Stage II-B: Corresponds to a tumor extending beyond the cervix, into the loose connective tissues around the patient's uterus (parametrium), but is still inside the pelvic area.
- Stage III-A: Corresponds to a tumor extending into the lower third of the patient's vagina.
- Stage III-B: Corresponds to a tumor extending into the patient's pelvic wall. At this stage, the cancer may be blocking the ureters (responsible for carrying urine from the Kidneys to the Bladder).
- Stage IV-A: Corresponds to a tumor extending into the patient's Bladder or rectum.
- Stage IV-B: Corresponds to a tumor extending into the patient's Liver or Lungs.

SYMPTOMS

Clinical symptoms include bleeding after sexual intercourse and between periods, abnormal bowel movements, or fatigue. Bleeding may also be noticed after severe exertion or straining during defecation. As the disease develops, the leukorrhea (vaginal discharge) may become filthy and mixed with blood, or foul and mixed with sloughed pieces of Body Fluid or tissue. Additional symptoms include lumbar, sacral and abdominal pain, Bladder irritation, frequent and urgent urination, blood-tinged vaginal discharge, unilateral lymphoedema, and unilateral ureteral obstruction (this indicates a very advanced stage).

TREATMENT

When Righteous Qi is increased through Medical Qigong therapy and prescription homework, the energy feeding the patient's cancer begins to decrease.

One Medical Qigong therapeutic technique used in China to treat cervical cancer is as follows:

1. Purge the diseased uterine area of pathogenic Qi using the Thunder Palm technique in conjunction with the Vibrating Palm to disperse the stagnation.
2. Use the "Yu" sound Qi projection into the uterine and cervix area to disperse Heat from the cancer cells.

3. Disperse the Excess Heat from the patient's Liver and cervix areas, leading the Toxic Qi out the patient's body via the right Gall Bladder Channel.
4. Fill and tonify the patient's Lower Dantian and Kidneys with Qi. Circulate the energy through the Microcosmic Orbit to improve Qi and Blood circulation through the Governing and Conception Vessels.

PRESCRIPTIONS AND HOMEWORK

The Conception Vessel moves Qi in the Lower Burner and uterus, and is the primary vessel used to treat uterine and cervical cancer; therefore, sexual activities are prohibited from the start of the treatments, until treatment is no longer needed.

The prescriptions and homework required for cervical cancer treatment are exactly the same exercises that are used for the treatment of uterine cancer.

PROSTATE CANCER

Prostate cancer is considered a malignant neoplasm, and is usually an adenocarcinoma of the prostate gland. It is the major type of cancer in males. Prostate cancer occurs in less than 1% in men under 50. Men with a family history of prostate cancer are three times more likely to develop it then the general public. Many prostate cancers remain dormant, but once awakened, can become life-threatening.

ETIOLOGY

The etiology of prostate cancer is unknown. Speculations suggest that prostate cancer can be related to an imbalance between male and female hormones. From a Western Medical perspective, the carcinogenesis is a multistep accumulation of genetic lesions. These lesions may result in uncontrolled cellular proliferation, a decrease in cellular death or an apoptosis, invasion, and the metastatic spread of the disease.

From a Traditional Chinese Medical perspective, prostate cancer is due to Excess Liver Fire and Kidney Deficiency. The suppression of anger and fear, as well as a diet consisting of a high alcohol intake, or consumption of too many hot spicy foods, may also cause, or contribute to prostate cancer.

SYMPTOMS

In the early stages there are no symptoms; however, as the disease progresses the inflamed prostate gland impedes urination and results in outflow obstruction, preventing the Bladder from emptying completely. Dull pain is experienced around the external genitals and Bladder. There is frequent burning, unfinished and dripping after urination with white discharge. Also, in some cases, there will be sexual function disorders (an inability to attain erection) as well as sciatica-like symptoms with pain radiating downward to the perineum and anus.

If the prostate is hard and firm, it indicates cancer; if the prostate is large and swollen, it indicates an infection. Inflammation of the prostate can be divided into acute and chronic conditions.

In acute prostatitis caused from Damp Heat in the Lower Burner, symptoms include pain and discomfort in the genital and Bladder areas, fever, and chills.

In chronic prostatitis, symptoms include dull pain in the urogenital area, discharge from the penis, discomfort during sexual intercourse, and frequent urination (especially during the night).

TREATMENT

One of the Medical Qigong therapeutic techniques used in China to treat prostate cancer is as follows.

1. With the patient supine, begin to purge the diseased prostate area of pathogenic Qi using the Thunder Palm technique in conjunction with the Vibrating Palm to disperse the stagnation.
2. Next, use "Yu" sound projection into the prostate area to disperse the Heat from the cancer cells.
3. Disperse Excess Heat from Liver and pathogenic Qi from the prostate, out of the patient's body via the right Gall Bladder Channel.
4. Fill and tonify the patient's Lower Dantian and Kidneys with Qi, and circulate the energy through the patient's Microcosmic Orbit to improve Qi and Blood circulation within the Governing and Conception Vessels.

PRESCRIPTIONS

Depending on the patient's condition, one or more of the following prescriptions may be given:

1. The "Guo" sound resonation is used to disperse the Liver Fire.
2. The Kidneys' "Yu" sound resonation is used to disperse Toxic Qi from the prostate.
3. Tonification and regulation exercises are prescribed for the Lower Dantian, Spleen, and Kidneys.
4. The following three exercises are used to treat prostate cancer, and were made popular in the West by Dr. Hong Liu.
    a. Sit at the edge of the chair with your feet flat on the floor. The right leg crosses over your left leg. With your left hand grasp the bottom of the right foot.

       Then extend your right hand out in front of the body, palm facing upward, while you imagine releasing toxic energy from your body through the mouth as you exhale. As you inhale, twist your hand as far as possible towards the right side keeping your eyes on the palm, while simultaneously pulling your foot towards the body and contracting the anus and perineum. This opens the Liver Channel and directs Qi into the Kidneys and prostate area.

       After completing the movement, hold your breath for a moment. Imagine Blood and Qi flowing into the prostate area and exhale while returning to the starting position. The exercise is repeated for 18 breaths per each side.
    b. The next exercise begins from a Wuji posture. Bend over as you flex your toes upwards. Hold this position as long as you can before standing erect, focusing on breathing naturally from the abdomen. This movement increases the flow of Qi and Blood into the prostate area. It is important to make sure this exercise is performed slowly. The exercise is repeated 9 times.
    c. This last exercise is done sitting on the floor while facing a wall. Place the balls of your feet against the wall while resting your heels on the floor. Wrap your arms around your bent knees, while keeping the back straight and breathing naturally through the Lower Dantian.

       While inhaling, focus on breathing from the lower abdomen instead of the chest. This causes the Qi of the Thrusting Vessel to stimulate the prostate gland, and increases the flow of Qi and Blood into the prostate area. The exercise is repeated for 18 breaths.
5. The "Filth Dispelling Meditation," is also prescribed, with the patient focusing on his prostate area. This meditation is practiced as follows:
    a. From a sitting posture, with the eyes closed, and the body relaxed, place the tongue up against the upper hard palate, behind the teeth.
    b. Breathe naturally and evenly.
    c. Imagine Qi whirling in through the upper area of the Bladder, circulating clockwise, building momentum as the energy flows down to the prostate area.
    d. Once the energy has reached the prostate, the patient will imagine the whirling Qi absorbing the noxious Heat and filthy Qi from the Bladder and prostate areas, transforming it into wind, which is dispelled out through the penis. The exercise is repeated 10 times.
    e. Next, gently close the anal sphincter while inhaling and imagine the Qi flowing into the body through the penis. This energy flows into the Lower Dantian, becoming transformed into wind. This wind begins circulating in a counterclockwise direction, spiraling up the body, through the Thrusting Vessel and exits through the mouth with each exhalation. This is also practiced for 10 repetitions.

f. Finally, focus the attention on the Lower Dantian; imagine the vital Qi returning back to its origin, while rubbing the abdomen in a clockwise direction to end the prescription.

LUNG CANCER (PULMONARY CARCINOMA)

Lung cancer, also called a "Bronchial Carcinoma," is a malignant Lung tumor that originates in the bronchi. It can also spread to the trachea appearing as an ulcer in the windpipe, a small flattened bump or nodule. It can extend into the body's lymphatic system, as well as into the blood vessels.

The most common type of bronchogenic tumor is the squamous cell or epidermoid (on the airway lining). The oat cell and adenocarcinoma are the next most common.

ETIOLOGY

Squamous cell cancer is usually associated with a history of cigarette smoking and secondhand smoke. Statistically it is the most common of all carcinomas and accounts for 40% of all male deaths. Other chemical carcinogens may cause Lung cancer, especially among workers in industrial and mining areas.

SYMPTOMS

The main clinical manifestations include: chest pain, coughing with scanty sputum (sometimes with blood), pectoralgia, fever, loss of appetite, loss of weight, weakness and breathlessness in the later stages due to a Lung Yin Deficiency.

TREATMENT

Purge the Excess Heat from the patient's Lungs and Liver. Tonify and regulate the patient's Kidneys and Mingmen area, then strengthen the patient's Spleen and Lungs.

PRESCRIPTION AND HOMEWORK

1. Have the patient practice the Dry Crying exercise for a period of 15 minutes twice a day (see Chapter 53).
2. Have the patient practice the Sun and Moon Rotation Technique (see Chapter 41), ending with the healing sound "Shang" for 24 breaths, 9 times a day.
3. Prescribe the Fast to Moderate Exhaling Breathing Method along with the Cancer Stepping Method for treating Lung diseases (see Chapter 15).
4. Have the patient practice the method of Taking in the White Qi, three times a day for 15 minutes each time to strengthen the Lungs.
5. Have the patient practice the method of Taking in the Blue Qi, three times a day for 15 minutes each time in order to strengthen the Kidneys.

LIVER CANCER

The Liver is the most usual site for the metastatic spread of tumors that disseminate through the Blood system. Carcinoma of the Liver is quick in development, high in mortality, and is regarded as one of the common malignant tumors.

ETIOLOGY

The etiology of Liver cancer is unknown. There are speculations suggesting that the following list of items may cause, or contribute, to the creation of Liver cancer:

- exposure to chronic stress and the suppression of anger,
- toxic poisoning (Liver cancer is usually associated with a long history of alcohol, drug abuse and chronic hepatitis), and
- a congenital weakness, with heredity being the predisposing factor.

SYMPTOMS

The Liver may have a single nodule, or multiple nodules. The main clinical manifestations include: distension and severe hypochondriac pain, hemorrhage of the digestive tract, and progressive hepatomegaly (enlargement). Its surface feels irregular on palpation (central depression or umbilications can often be detected).

TREATMENT

Purge the Excess Heat from the patient's Liver and Gall Bladder Channels, then tonify and regulate the patient's Kidney and Mingmen areas. Strengthen the patient's Spleen and Lungs, then regulate the patient's Liver.

PRESCRIPTIONS

1. Have the patient practice the Descend the Yang and Ascend the Yin Technique (see

Chapter 41), ending with the healing sound "Guo" for 24 times, 9 times a day.

2. Prescribe the Fast to Moderate Exhaling Breathing Method along with the Cancer Walking Method for treating Liver disease (see Chapter 15).
3. Have the patient face, or embrace, a tree while practicing the method of "Taking in the Wood Element Qi," but only upon inhalation. With each exhalation the patient should purge the toxic Qi, sending it into the ground. This prescription is practiced three times a day for 15 minutes each time.
4. Have the patient practice the method of "Taking in the White Qi," three times a day for 15 minutes for each session to strengthen the Lungs.
5. Have the patient practice the method of "Taking in the Green Qi," three times a day for 15 minutes for each session to strengthen the Liver.
6. Have the patient practice the method of "Taking in the Yellow Qi," three times a day for 15 minutes for each session to strengthen the Spleen.

ESOPHAGEAL CANCER

Carcinoma of the esophagus is one of the common malignant tumors, caused from a carcinoma of the squamous epithelium of the esophagus. This type of carcinoma is called "Ge Shi" or dysphagia in Traditional Chinese Medicine.

ETIOLOGY

The cause of esophageal cancer is unknown. There is speculation that the following list of items may cause or contribute to the creation of esophageal cancer:

- consumption of foods with a high content of nitrates,
- large quantities of spicy, hot foods, or mouldy food,
- chronically eating too fast, thus irritating the throat,
- strong alcoholic drinks,
- smoking,
- poor oral hygiene, resulting in tooth decay, and
- chronic stress and the suppression of emotions.

There are three types of esophageal cancer that invade the wall of the esophagus.

1. An ulcerous carcinoma of the esophagus that grows towards the outside of the esophagus. Dysphagia (difficulty in swallowing) is not evident in this type of cancer.
2. A tumor which grows in a circular formation. In this type of cancer, the lumen is narrow and easily obstructed.
3. The tumor attaches itself to one side of the esophagus, while the other side is still healthy and can still dilate. In this type of cancer, symptoms of obstruction appear slowly.

SYMPTOMS

More common in men over 40, carcinoma of the esophagus causes pain in the epigastric and sternal area. Swallowing aggravates the symptoms. The main clinical manifestations include the following symptoms:

- enlarged lymph nodes in the neck,
- difficulty in swallowing (dysphagia),
- excessive salivation,
- poor digestion,
- vomiting,
- loss of weight,
- pain in the chest and back, and
- emaciation and weakness.

TREATMENT

Purge the Excess Heat from the patient's throat, Lungs and Liver area. Tonify and regulate the patient's Kidney and Mingmen area, then regulate the Qi of the patient's esophagus, rooting it within the Lower Dantian.

PRESCRIPTIONS

1. Have the patient practice the "Dry Crying" exercise for a period of 15 minutes twice a day (see Chapter 53).
2. Have the patient practice the healing sound "Shang" for 24 breaths, 9 times a day (see Chapter 40).
3. Prescribe the Fast to Moderate Exhaling Breathing Method along with the Cancer Walking method for treating Lung disease (see Chapter 15).

4. Have the patient practice the method of "Taking in the White Qi," three times a day for 15 minutes for each session to strengthen the Lungs.

STOMACH CANCER

Carcinoma of the Stomach is regarded as one of the common malignant tumors. It is sometimes categorized as a carcinoma, lymphoma or sarcoma.

ETIOLOGY

The etiology for Stomach cancer is unknown. Speculations suggest exposure to chronic stress, and a suppression of emotions, may cause, or contribute to this disease. There is also an association between the development of gastric cancer and previous partial gastrectomies for benign diseases.

SYMPTOMS

The symptoms of Stomach carcinoma often do not appear until they have developed to the middle and late stages. Most common in men over 40, carcinoma of the Stomach causes a continual pain and burning sensation in the epigastric and sternal area, which is usually aggravated by swallowing. The main clinical manifestations include: shallow complexion, lack of appetite, loss of weight, extreme anemia, blood in the stool, debility, vomiting of partially digested food, and sometimes Blood that looks like coffee-grounds. The Liver may also be enlarged with nodules.

TREATMENT

Purge the Excess Heat from the patient's Stomach, Liver and Gall Bladder Channels, then tonify and regulate the patient's Heart and Spleen. End by regulating the patient's Kidney and Mingmen areas.

PRESCRIPTIONS

1. Have the patient practice the Descend the Yang and Ascend the Yin Technique (see Chapter 41) ending with the healing sound "Dong" for 24 times, 9 times a day (see Chapter 40).
2. Have the patient practice the method of "Taking in the Yellow Qi," three times a day for 15 minutes each time in order to strengthen the Spleen.
3. Have the patient practice the "Taking in the Red Qi," three times a day for 15 minutes each session to strengthen the Heart.

PANCREATIC CANCER

Carcinoma of the pancreas is regarded as one of the common malignant tumors. Statistics state that 22% of deaths from gastrointestinal cancer are due to cancer of the pancreas.

ETIOLOGY

The etiology for pancreatic cancer is unknown. Speculations suggest cigarette smoking and diet (in particular, a high intake of fat and / or meat) may cause, or contribute to the cancer. Recent investigations have identified a number of factors that may also contribute to the cause of pancreatic cancer, including: environmental factors, surgical or associated medical factors, genetic factors, and occupational exposures to toxins.

SYMPTOMS

This type of cancer is more common in men between 50 and 70. A carcinoma of the pancreas causes pain in the epigastric area, which starts as a dull ache and usually becomes aggravated by lying down. The main clinical manifestations include dyspepsia, diarrhea and constipation, loss of weight, nausea, vomiting and fatigue.

The symptoms of pancreatic cancer can sometimes vary in accordance to its exact location. The head of the pancreas is close to the duodenum, while its tail reaches to, and energetically connects with, the Spleen. The middle of the pancreas lies horizontally across the upper aspect of the abdomen.

1. Carcinoma of the Pancreas's head has as for its main symptom jaundice, due to the common bile duct of the Gall Bladder becoming obstructed. This obstruction causes the bile to back flow into the Liver and thus contaminates the Blood. This causes both the Gall Bladder and Liver to swell. As the amount of bile and Blood increases, it causes the skin, mucous membrane and sclera to become yellow, and the urine to turn brown. During the later stages, the skin color changes from a yellowish color to orange or dark yellow.
2. Carcinoma of the Pancrea's body has for its

main symptoms sharp abdominal pain and swelling. The abdominal pain reacts to the compression of the abdominal cavity nerve. This type of pain can occur sporadically, radiating to the patient's shoulder and back. The patient will notice this type of pain becoming aggravated when sitting up or lying down. Bending forward, however, can ease the pressure on the celiac nerve plexus and reduce the pain.

3. Carcinoma of the Pancrea's tail usually does not have detectable, early symptoms. At times, it can only be discovered when the tumor has metastasized to the peritoneum, Lungs, bones or other organs. Once the tumor has developed to sufficient size, the manifestation of symptoms include anorexia, fatigue, and loss of body weight.

TREATMENT

Purge the Excess Heat from the patient's pancreas, then dredge the Stomach, Liver and Gall Bladder Channels. Next, tonify and regulate the patient's Heart and Spleen, then regulate the patient's Kidney and Mingmen areas.

PRESCRIPTIONS

1. Have the patient practice the Descend the Yang and Ascend the Yin Technique (see Chapter 41), ending with the healing sound "Dong" for 24 times, 9 times a day (see Chapter 40).
2. Have the patient practice the method of "Taking in the Yellow Qi," three times a day for 15 minutes each session to strengthen the Spleen.
3. Have the patient practice the "Taking in the Red Qi," three times a day for 15 minutes each time in order to strengthen the Heart.

COLON AND RECTUM CANCER

Carcinoma of the Large Intestine or rectum is regarded as one of the common malignant tumors of the digestive tract.

ETIOLOGY

It has been long accepted that colorectal cancer is caused, or promoted, by environmental factors, especially diet. It is suspected that carcinogens are present in the patient's feces.

Also, chronic ulcerative colitis is considered a predisposing factor for colorectal cancer. The following chronic diseases are believed to be related: benign tumors of the colon and rectum (especially frequently occurring Large Intestine polyps), and chronic colitis (especially chronic amoebic dysentery and chronic ulcerative colitis).

SYMPTOMS

Colorectal cancer is more common in patients over 50 years of age. A carcinoma of the colon causes pain in the lateral abdominal area, which also may be umbilical. The main clinical manifestations include a sudden change of bowel habit (constipation or diarrhea), loss of weight, fatigue and Blood in the stool.

Large Intestine cancer can be divided into two quadrants of right and left intestinal dysfunction.

- Cancer on the right half of the Large Intestine is measured from the cecum to the right half of the transverse colon. The early symptom of this disease is dull pain and discomfort within the right epigastrium, or lower abdominal area. As the disease progresses, the pathological changes of the tumor initiate poisoning symptoms which appear as: weakness of the body, poor appetite, loss of weight, and anemia. In the midterm, or late stages, the patient can sometimes feel the tumor on the right side of their abdomen (located where the right Large Intestine stores food residue from the Small Intestine).
- Cancer on the left half of the Large Intestine is measured from the center of the transverse colon to the junction of the sigmoid colon and rectum. Symptoms of cancer in the left half of the Large Intestine appear earlier because the frequency of defecation increases with mucus and Blood in the stool. Moreover, because the narrow opening of the Large Intestine is now occupied by the tumor, it is difficult for the fecal matter to pass through, causing inflammation, as well as alternating diarrhea and constipation. If the tumor continues to grow, the intestinal cavity can be clogged, causing abdominal distention and toxicosis of the whole body.

TREATMENT

Purge the Excess Heat from the patient's Large Intestine and rectum area, purging and dredging the toxic Qi out the body and down the Stomach, Liver and Gall Bladder Channels. Next, tonify and regulate the patient's Kidney and Mingmen areas, emitting Qi through the soles of the patient's feet. End with regulating the patient's Fire and Water Qi through the Fire Cycle of the Microcosmic Orbit.

PRESCRIPTION AND HOMEWORK

Have the patient practice the "Filth Dispelling Meditation," focusing on the Large Intestine and rectum area. The meditation should be practiced as follows.

1. Begin from a sitting posture, eyes closed, body relaxed, with the tongue placed up against the upper hard palate, behind the teeth.
2. The breathing should be natural and even.
3. Imagine Qi whirling in through the upper orifice of the Stomach, circulating clockwise through the Small Intestine, building momentum as it flows to the Large Intestine.
4. Imagine the whirling Qi driving through the colon. As it moves through the colon, the whirling Qi begins absorbing the noxious Heat and filthy Qi from the Large Intestine, transforming it into wind, and dispelling it out the anus. This is practiced for 10 repetitions.
5. Next, gently close the anal sphincter while inhaling; imagine the Qi in the lower abdomen again becoming transformed into wind. This wind begins circulating in a counterclockwise direction, spiraling up the body, through the Stomach, and is expelled out the mouth. This is also practiced for 10 repetitions.
6. Finally, focus the attention on the Lower Dantian; imagine the vital Qi returning back to its origin, and begin rubbing the abdomen clockwise to end the prescription.

CHAPTER 55
QI EMISSION THERAPY FOR RADIATION AND CHEMOTHERAPY

The purpose of Cytotoxic treatments (radiation and chemotherapy) is to kill the dividing cells responsible for the formation of the tumor mass. The well-known side effects of these types of therapy are loss of hair, loss of appetite, nausea, vomiting and damage to the patient's skin and gastrointestinal tract. When facing radiation and chemotherapy, the damage to the patient's immune system, although less obvious, is the major concern for the Qigong doctor. The cure of the cancer inevitably lies in activating and increasing the function of the immune response, so that it can recognize and eliminate the malignant tissue.

In China, Medical Qigong Therapy is commonly prescribed as an adjunct to both radiation and chemotherapy. The value of Medical Qigong Therapy is that it is very useful in reducing the harmful side effects caused by both radiation and chemotherapies, as well as enhancing the patient's immune system. Both radiation and chemotherapy assist the body in weakening a pathogenic cluster, but do not completely kill all of the tumor or cancer cells. That function is still left to the patient's immune system (which has consequently been weakened by the radiation or chemotherapy). Because Medical Qigong treatments, prescriptions and meditations enhance the body's immune system, its utilization as a adjunctive therapy is paramount.

MEDICAL QIGONG THERAPY AND RADIATION THERAPY

In most cases, radiation is much safer than chemotherapy because it can be directed into one specific area of the patient's body. However, it may also cause scarring that can interfere with the patient's internal organ function. Although radiation therapy is strong, its effect on the patient's body can be corrected by the Qigong doctor with dredging, purging, dispersing and tonifying techniques. The tissue damage left after radiation therapy should be rectified by dredging, purging and dispersing the afflicted area first. Then, after the toxic Heat has been purged, begin to tonify the patient's supporting organs and tissues surrounding the treatment area.

Sometimes the harmful side-effects of radiation therapy can be avoided by giving the patient Resonant Sound Therapy. In order to alleviate the patient's toxic Heat and pain, it is important for them to practice the "She" sound to cool the tissues after radiation treatments. After the organs have been sufficiently cleansed, the doctor may then tonify the patient's tissues, as well as the Wei Qi Fields, which have been damaged by the radiation therapy.

MEDICAL QIGONG THERAPY AND CHEMOTHERAPY

All forms of chemotherapy are cell-killing agents that damage the DNA and injure actively dividing cells, including those of the body's immune system. The tissue damage after chemotherapy is extremely extensive; the white blood cells decrease, causing secondary infections. Unlike radiation therapy, which is generally isolated to specific areas of the patient's body, chemotherapy affects all of the body's internal organs.

When treating patients who have undergone chemotherapy, it is important for the Qigong doctor to dredge, purge and disperse all the patient's Yin and Yang organs. As with radiation therapy, Sound Resonation Therapy is also mandatory in alleviating the patient's pain.

It is important for the patients to practice the "She" sound to cool the tissues after chemotherapy treatments. Specifically, cancer patients should pronounce the "She" Healing Sound therapy several times a day when their blood count is low, after both chemotherapy or radiation therapy. The sound "She" is pronounced six times. Its high pitched sound is in the straight tone, and its low pitched sound is in the dropping and rising tone. This sound is dropped and replaced by the sound "Ha," or the sound for the specific internal organ, as soon as the red and white blood count reaches a normal level.

After pronouncing the sound "Ha" for a certain period of time, add the sounds for the specific internal organs that have been affected (for example, add "Shang" for a patient with Lung cancer, and "Guo" for a patient with Liver cancer, etc.).

When treating the patient, after the diseased organs have been sufficiently cleansed, the doctor may then tonify the patient's Blood, Jing and tissues, as well as their energetic fields, which have been damaged by the chemotherapy.

RELIEVING SYMPTOMS ASSOCIATED WITH RADIATION AND CHEMOTHERAPY

After receiving radiation or chemotherapy, patients may suffer from several internal complications due to the severe organ and tissue trauma caused by these therapies. The following are six common problems resulting from radiation and chemotherapy.

1. A gastrointestinal infection can result from both radiation and chemotherapy, which have a tendency to burn up the body's Yin. This frequently results in causing Rebellious Qi, which is accompanied by such symptoms as nausea, vomiting, and poor appetite. Treat by tonifying the patient's Stomach Yin by prescribing herbs, and use Jing Point Therapy to quiet the digestion by tonifying Pc-6, St-36 points, as well as the Ear Shenmen points.
2. An infection of the Large Intestine can occur from the accumulation of toxins in the Lower Burner creating Damp Heat Symptoms (caused from radiation and chemotherapy). These toxins can cause abdominal pain, constipation or diarrhea with bloody mucous. Treat by tonifying the patient's Yin to prevent loss of Body Fluids (also give astringent herbs), and use Jing Point Therapy to stop any internal bleeding, by tonifying St-25, St-36, St-37, St-38, LI-4 and LI-11 points.
3. A Urinary Bladder infection can be caused by both radiation and chemotherapy that have a tendency to create Damp Heat in the Lower Burner. Damp Heat in the Lower Burner can result in cloudy, bloody and scanty urination, with a feeling of an urgency to urinate but an inability to do so. To treat, use Jing Point Therapy and tonify Sp-6, Sp-9, CV-2, CV-3, Bl-23 and Bl-29 points.
4. Edema in the limbs can occur from both radiation and chemotherapy that have a tendency to create Wind Damp and Deficiency, which leads to an accumulation of Body Fluids in the extremities. Treat by tonifying the Spleen and draining the body of Dampness. Use Jing Point Therapy to tonify Sp-6, Sp-9, Lu-7, LI-4, and Bl-20; also prescribe herbs.
5. A diminished white blood cell count can be caused by radiation and chemotherapy and can lead to secondary infections. Treat by tonifying the patient's Blood and Jing, and prescribe herbs.
6. Pneumonia can be caused by radiation and chemotherapy that have a tendency to cause damage to the Lungs. When both the Lung Qi and Yin Qi are damaged, pneumonia can result. Treat by tonifying the patient's Yin and Lung Qi, and prescribe herbs to enhance the immune system.

CHAPTER 56
QI EMISSION THERAPY FOR SURGERY

Medical Qigong Therapy is useful for treating patients before, during and after surgery. Qi emission can be used to reduce the patient's bleeding, enhance the immune system, minimize the risk of infection, strengthen the body, and to accelerate the recovery rate. Medical Qigong modalities are involved in preoperative therapy, surgery, postoperative therapy, follow-up therapies and remedial prescriptions.

MEDICAL QIGONG AND PREOPERATIVE THERAPY

Although Medical Qigong Therapy has had incredible results in healing and in stopping the progression of certain diseases, it is not a "cure-all." As with other clinical modalities, it has its own limitations. Currently, conventional medicine only utilizes three modalities for treating tumors and cancer; these include surgery, radiation and chemotherapy. In the healing process, undergoing surgery is sometimes unavoidable, and should never be viewed by the patient as a defeat, but only as a necessary step in the healing transition.

Having to face surgery should not be viewed as a negative reflection of a Qigong doctor's healing potential, nor as a patient's inability to heal. Surgery should be viewed as a necessary life saving intervention, that can firmly establish the patient on the road towards health and recovery. It is a positive affirmation of the patient's commitment to getting well and taking responsibility for the healing. Sometimes, surgery is a necessary step towards health and healing.

Before surgery, it is important for the doctor to consider the emotional relationship that the patient has developed with his or her diseased tissues. The mental, emotional, energetic and spiritual aspects of a disease must not be ignored.

The doctor's involvement with the patient and the tissues which are about to be surgically removed is extremely important, since this emotional interaction affects the patient's healing potential. For healing to take place, a trusting relationship with the patient must first be solidly established; this includes a compassionate attitude of the doctor towards the tissues that will be removed. The doctor models the correct attitude to facilitate an accelerated healing for the patient. It is believed that scar tissue formation is in direct proportion to the patient's fears and negative expectations surrounding the surgery. These fears and expectations form energetic and spiritual attachments to the physical body. Often the patient attempts to disconnect these spiritual attachments out of fear and survival instinct.

A patient is spiritually, emotionally and energetically interconnected with his or her body. This energetic attachment includes any and all organs, organ systems, or areas of the body which can be involved in the operation.

A patient may sometimes feel disillusioned about his or her own interpersonal relationship with the body, and try to disconnect all energetic attachments to the body's tissues out of fear and survival instinct. It is important for the Qigong doctor to help the patient get in touch with the suppressed feelings of loss surrounding the tissues which will soon be operated on. This is because surgery represents the death of a relationship with part of the self. Although the relationship that the patient has with the diseased organ is not considered a healthy one, it exists nevertheless. Allowing the patient to deny his or her true feelings related to the surgery interferes with the healing process. Denial may lead to increased scar tissue formation, and can promote the return of certain disease formations.

The Qigong doctor therefore encourages the patient to talk to, and "brief," each organ and tissue area regarding the surgical procedure to prevent further denial, and to promote the release of feelings. Otherwise, the patient may experience feelings of loss, grief and depression; this reaction can of course also be due to the aftereffects of the anesthesia. It is important that these feelings be experienced and expressed before the surgery. After the surgery, the patient may then experience feelings of relief and gratitude about the new changes in his or her body.

To prepare for surgery, the patient is usually given Medical Qigong meditations and prescriptions that strengthen the Lower Dantian, Kidney and Mingmen areas. One, or several sessions may be specifically devoted to allowing the patient to connect with the diseased organ(s), as well as the regions that are to be operated on. A request for divine intervention can facilitate the patient's relinquishing control, and for the removal of any and all energetic armor surrounding the diseased area (and the subsequent release of emotions associated with this armoring). The patient is then encouraged to disconnect any energetic attachments that the diseased tissues have formed with the surrounding cells, facilitating a final closure. This gradual process allows for an easier transition, as the patient's body must completely release any attachment to the diseased tissues during surgery.

Just before surgery, the patient is given a meditation, and encouraged to imagine a divine white light energy submerging and enveloping the regions of the body which are to undergo the operation. The Lower and Middle Dantians, as well as the areas which will soon be operated on, should also be energized. In addition, the Liver should also be energized to make sure that the Blood coagulates properly during the surgery.

USING MEDICAL QIGONG FOR ANESTHESIA

According to research on Western Surgeries (presented at the Second International Symposium on Memory and Awareness in 1992), patients who undergo anesthesia may be aware of both the pain and conversations within the operating room. Many kinds of drugs used to supplement anesthesia, such as neuromuscular blocking agents, paralyze all the muscles in the body but do not affect the central nervous system and consciousness. During surgery, patients have been observed flinching, twitching and making other facial expressions. Many patients later can give an accurate detailed account of the surgical procedure and conversations between the doctors and nurses. Even when unconscious, patients (whether energetically inside or outside of their bodies), can record, see and remember everything.

Qigong energy emitted during surgery, can however, reduce pain and wound shock, as well as calm the patient's Shen. Qi emission also reduces postoperative complications such as respiratory tract infection, functional disorders of the gastrointestinal tract, retention of urine, etc. The energy received by the patient also speeds up the healing process.

Medical Qigong therapy is found to be very useful in relieving pain by affecting the body's neuro-chemicals. The energy stimulates the peripheral and cutaneous nerves that carry sensory information, via the spinal cord, to the brain. This stimulation of the cutaneous nerves activates neurotransmitters which carry pain messages to the brain, and facilitates the closure of the body's pain-relay gates. The brain produces endorphins, or endogenously generated morphine-like chemicals in response to this closure. The body's opiate receptors are then able to dull the pain. This process is the basis for Qigong Anesthesia. Because Medical Qigong therapy safely produces an analgesic affect on the body's cutaneous tissues, its use in hospitals for surgery as well as for pre and postoperative procedures is gradually increasing.

The first operation in which Qigong anesthesia was applied (recorded in modern times), was performed successfully for the removal of a thyroid tumor in the Shanghai #8 People's Hospital, on May 9, 1980. Qigong Master Lin Hou-sheng demonstrated the ability to induce anesthesia during surgery by pointing his fingers at specific acupuncture points, using external Qi projection without the use of any Western anesthetic. The patient's blood pressure, pulse and respiration rate remained stable during the operation, and the patient's physiological function remained normal after the operation. The success of the Qigong anesthesia boosted the resident doctor's confidence and laid a solid foundation for future applications of Qigong anesthesia in surgical operations.

Qigong anesthesia has begun to follow in the wake of the widely accepted acupuncture anesthesia, as a holistic alternative to conventional pharmaceutical methods. The number of cases treated in China with Qigong anesthesia is steadily increasing, primarily for the following six benefits.

1. Arresting the patient's pain,
2. Resisting the pulling reflex action or shock reaction of the body's internal tissues and organs when the scalpel is inserted,
3. Reducing infection,
4. Reducing the body's opposition or contrary action/reaction to operative wounds,
5. Preventing wound shock (response of the body to the tissue trauma), and
6. Promoting the healing of wounded tissues.

Since Qigong analgesia is effective through the stimulation of certain points on the body's surface, both the location of the points, and the type of energetic application are crucial. There are three methods commonly used in selecting points: According to the Channels, According to the Segmental Innervation, and According to the Auricular Points.

1. The analgesic approach used according to the patient's channels includes two main applications:
   - The selection of points according to the course of the channel which traverses the site of the operation, and
   - The selection of points according to the differentiation of symptoms and signs of a disease, as well as the responses that may be elicited from the patient's body during the operative procedure.
2. The analgesic approach used according to the segmental innervation includes three main applications:
   - The selection of points according to the adjacent segment, or an area that is supplied by the same spinal nerve, or an adjacent spinal nerve of the operative site,
   - The selection of points according to the remote segment, in an area not supplied by the same spinal nerve, or an adjacent spinal nerve (on the operative site), and
   - The selection of points according to the stimulation of the nerve trunk within the same segment, stimulating directly the peripheral nerve of the operative site.
3. The analgesic approach used according to the Auricular Points employs the insertion and stimulation of acupuncture needles with Medical Qigong Therapy, and includes three main applications:
   - The selection of points according to the corresponding Auricular areas pertaining to the operative site and its involved internal organs,
   - The selection of points according to the theory of the Yin and Yang Organs pertaining to the operative site and its involved internal organs, and
   - The selection of points according to such things as tenderness, reduction of electro-resistance, the appearance of deformation and/or discoloration when a certain internal organ or area of the body is affected.

It is important preoperatively that the patient have confidence in the doctor's method of Qigong application. Any misgivings should be dispelled by explaining the procedure to gain the patient's full corporation and trust. In order to know the level of response that the patient will experience, it is necessary to test the patient's response to Qigong anesthesia prior to any surgery.

To use external Qi as an anesthetic, the Qigong doctor must first focus Qi into his or her Lower Dantian, increase its charge and density and then extend it out through the palms, or sword fingers, into the patient's body through specific channel points. The goal is to numb a specific area on the patient's body without touching the patient.

When Qigong anesthesia is being administered successfully, 96% of the nerve fibers responsible for pain sensation become anesthetized and the patient loses consciousness within ten minutes. Ten minutes after cessation of the Qigong anesthesia emission, 24% of the anesthesia has disappeared; while 72% of the tissue remains numb. The remaining tissue quickly recovers from the anesthetic effect and the patient quickly returns to consciousness.

Patients suffering from pain caused by such conditions as cancer or dysmenorrhea also respond positively to Qigong anesthesia.

MEDICAL QIGONG THERAPY AND SURGERY

Surgery has been performed in China for thousands of years. In ancient times, the legendary physician Hua Tuo was noted, not only for his expertise in Medical Qigong, acupuncture, massage and herbs, but also for his surgical skills as well. In Western culture, surgical therapies are applied for the removal of tumors, cysts, and abnormal tissue growth, as well as the transplantation of organs (heart, lungs, and kidneys) and the replacement of dysfunctional systems (joint replacement). Surgery, however, having a potential to both hinder and obstruct the body's flow of life force energy, is generally considered a last resort modality.

During surgery, the Qigong doctor assists the surgeon by applying External Qi Emission to the patient's body, energizing and strengthening specific areas. This is done to reduce pain and alleviate or eliminate the pulling reflex action (shock reaction) of the body's internal tissues and organs when the scalpel is inserted, prevent infection, and reduce bleeding.

When assisting surgeons during an operation, I personally have found that it is important to maintain a physical connection with the patient's body before beginning the Qi emission, until the anesthesia has fully taken effect. As the anesthesia relaxes the patient's tissues, the Qigong doctor can feel and see the patient's spirit leave the body.

Just before the scalpel begins its first incision, it is important for the Qigong doctor to physically disconnect from the patient's body. Even though the Qigong doctor has physically disconnected from the patient's tissues, it is extremely important for the Qigong doctor to increase his or her energy extension deeper into the center core of the patient's body. This protects the Qigong doctor from experiencing the shock, and initial trauma from the scalpel's first incision, which resonates throughout the patient's entire body.

It is also important for the Qigong doctor to focus the attention on extending and circulating his or her energy deep into the patient's Taiji Pole while the surgery is being performed. As the patient begins to lose Qi and Blood due to the operation, the Qigong doctor replenishes the patient's lost Qi, as well as energizes the incoming Blood from the transfusion. The Qigong doctor should also slow down the bleeding by slowing the patient's pulse.

During surgery, the surgeon's scalpel will sever several of the patient's channels as the doctor cuts through the surface of the skin and continues into the deeper regions of the patient's body. Each channel that is being disconnected will weaken the patient's energetic field corresponding to that specific organ or organ systems. It is the Qigong doctor's responsibility to maintain connection with the patient's energy field, and also to monitor and maintain the patient's energy level. The Qigong doctor maintains this connection as the surgeons open up the patient's body cavity, move internal organs, remove the diseased tissues (and any associated structures), and then sutures and closes the patient's body cavity.

When the patient is being sutured, the Qigong doctor must facilitate the patient's circulation of Qi through the Microcosmic Orbit; this stimulates the re-connection of any of the patient's internal and external channels which have been severed. The Qigong doctor must also change and rebuild the Energetic Grid formations in the body's internal and external energetic fields. The patient's Energetic Grid is reconstructed to a pre-diseased pattern to prevent the regrowth of the disease.

MEDICAL QIGONG AND POSTOPERATIVE THERAPY

The separation and restructuring of the inner fasciae that occurs during surgery can cause serious postoperative problems by either stopping the Qi from flowing (causing stagnations), or altering the energy's natural course (causing deviations). This is why after the surgery has been completed, it is extremely important for the Qigong doctor to immediately dredge and disperse the Turbid Qi from the patient's body. This is performed by energetically combing over the fresh incisions, to reconnect the patient's energetic preoperative patterns. Energetic combing also promotes the rapid healing of the patient's wounded tissues, reduces the formation of scar tissue, and continues to alleviate the effects of the surgery.

To facilitate an escalated healing, the Qigong doctor must also energize the patient's Kidneys and any energetic field which may have been depleted. Also, to bring the patient's temperature back to normal, and to disperse any feelings of nausea after the anesthesia and shock of surgery wear off, certain Heart, Pericardium and Stomach Channel points require stimulation.

It is important to note that after an operation, the patient's Dantians and all the body's tissues have been depleted, and are in need of revitalization. While the patient is healing, the color of the aura (located in the Wei Qi field surrounding the area of the surgical trauma) is usually dark to light gray. To increase the healing potential, the Qigong doctor gives the patient specific prescriptions in accordance with the internal organ(s) affected, such as white light meditation images. This is prescribed to energize and replenish the patient's damaged tissues and energy fields. The Qigong doctor also continues to work on rebuilding, reconnecting and reinforcing the patient's channels and Energetic Grids which the surgeon has severed during the operation.

MEDICAL QIGONG THERAPY AND RECREATING ORGAN ENERGY

Once an organ has been removed, there remains an energetic "void" within the spatial cavity of the patient's tissues. The energy of the organ, however, still exists within the patient's body as a "Phantom Organ." This energetic phenomena is due to the creative "blue-printing" of the prenatal and postnatal formation of the body's tissues. Because mass and energy are interconnected and interchangeable, it is therefore possible to energetically stabilize, as well as energize, the spatial cavity wherein the surgically removed organ once resided. To assist the patient in recreating active organ energy, the Qigong doctor prescribes the following meditation:

Lungs	Brilliant Pure White
Large Intestines	Abalone-Shell White
Stomach	Deep Yellow
Spleen	Orange Yellow
Heart	Red with Blue Tint
Small Intestine	Pink
Bladder	Deep Blue
Kidneys	Light Pastel Blue
Pericardium	Purple Red
Triple Burners	Orange Red
Gall Bladder	Pale Chartreuse
Liver	Dark Green

Figure 56.1. Organ Color Chart

STAGE 1 - THE HOOKUP

Generally several days after the surgery, while the patient is still recovering, the Qigong doctor has the patient imagine him or herself being connected to the Heavens and rooted to the Earth. Next, the patient uses color visualization.

STAGE 2 - CHOOSING THE APPROPRIATE COLOR

To prescribe the correct color, it is important for the Qigong doctor to know which organ or organ systems have either been removed, or seriously affected by the surgery. The patient is then assigned specific color meditations to tonify and strengthen the energetic fields that support the traumatized tissues. These meditations are also used to reestablish a physical, mental, emotional, energetic and spiritual connection to the area of trauma.

Figure 56.1 categorizes the general colors associated with the body's organs and channels. These colors can be viewed as extensions of energy originating from the original organs.

STAGE 3 - CLEANING AND ENERGIZING THE SPATIAL CAVITY

After the doctor has selected the proper organ color, the patient is given specific images used to cleanse, purify, and energize the tissues. These meditations create a solid energetic transformation that stabilizes the patient's energetic field.

1. The Cleansing and Purifying meditation is as follows. The patient imagines that his or her fingers and toes are straws. As the patient inhales, he or she will imagine divine white light energy flowing from the Heavens and Earth into the fingers and toes, filling the specific area which has been traumatized by the surgery (Figure 56.2). While exhaling, the patient imagines any pathogenic Qi leaving the tissues and body via the fingers and toes. The patient continues to perform this cleansing meditation for several minutes until he or she feels that the specific area has been purified.
2. The Creating and Energizing meditation requires that the patient imagine a clean, vibrant color (the Five Element Color associated with the particular organ), flowing into the spatial cavity, and areas, surrounding the removed organ. As the patient inhales, the organ color fills the vacant cavity, energizing the remaining energetic structure and surrounding tissues. The patient continues with this imagery until the specific organ area is full of heat, light and vibration (Figure 56.3). Because the surgery has permanently removed the organ from the patient's body, it is important for the patient to continue with this meditation every day for the rest of his or her life.

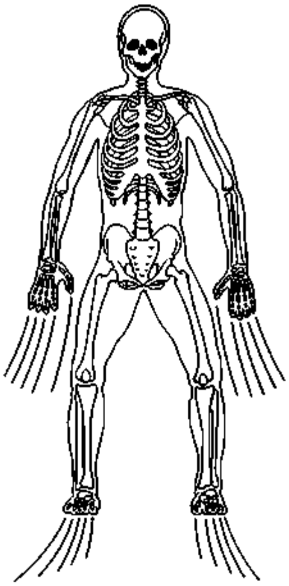

Figure 56.2. As the patient inhales, he or she will imagine the divine white healing light flowing into his or her body through the fingers and toes. This image is used to purge the patient's body and cleanse the tissues.

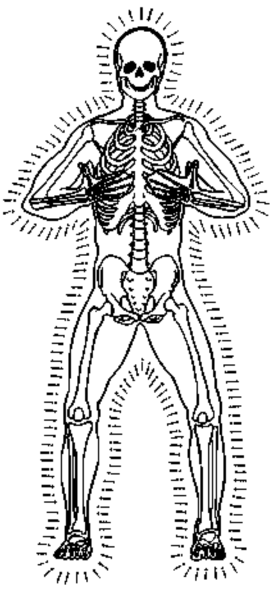

Figure 56.3. Once the patient is cleansed, in order to energize the tissues, he or she will imagine a clean vibrant energy color flowing into the specific spatial cavity associated with the removed organ.

MEDICAL QIGONG THERAPY AND ORGAN TRANSPLANTATION

Each particular organ in the human body has its own unique history and experience. This emotional memory and experience is active in the donated organ, as well as in the patient's diseased organ. Before the body can, or will receive a foreign organ, there must first be a time of reconciliation between the memories stored within the negative (as well as positive) physical cells of the organ soon to be replaced, and that of the donor's organ. To assist the surgeon in transplanting an organ, the Qigong doctor must perform visceral cleansing, preparation and purification of the transplanted organ. This method is explained in the following three stages:

STAGE 1

The patient must release all attachment to the departing organ. This includes all negative (as well as positive) physical, mental, emotional, energetic and spiritual memories and experiences.

STAGE 2

The history and emotional memory of the transplanted organ must be cleansed. This includes all negative (as well as positive) physical, mental, emotional, energetic and spiritual memories stored within the donor's organs.

STAGE 3

The relationship between the patient's tissues and the spatial cavity wherein the initial diseased organ was, or is located, must be established and harmonized.

The patient's body will either receive or reject the transplanted organ due to unresolved emotional trauma stored within the patient's tissues, or within the transplanted organ itself. Current research reveals that emotions trapped within transplanted organs can surface within the patient's subconscious mind, affecting his or her behavioral patterns. This is manifested through certain unnatural behavioral traits, which are transferred via the donated organ to the patient.

A famous example of a lingering emotional connection between an organ donor, the organ, and the recipient is described in the following true story. A few years ago a young woman (vegetarian) received a heart transplant from a young man who died in a motorcycle accident. Upon being discharged from the hospital, the woman began craving Chicken McNuggets, which was the young man's favorite food. This was in direct opposition to her previous diet and eating habits.

It is important to note here that organ rejection is not always due to the donor or patient's unresolved emotional issues. Organ rejection can also be caused from biological reactions complicating the acceptance of the foreign tissue.

FOLLOW-UP AND REMEDIAL PRESCRIPTIONS

The follow-up and remedial prescriptions will be determined according to the specific type of surgery the patient has undergone. The prescriptions must be designed to purge, strengthen and regulate any organ(s) that have been weakened by the surgery. Follow-up and remedial prescriptions must also circulate the patient's energy to reduce any energetic stagnations which may have been caused by internal organ movement and fascia damage. This will assist the patient in an escalated recovery, as well as reduce postoperative depression, which often follows after major surgery.

Sometimes, when assisting a patient who is healing from a severe disease (i.e., malignant tumors), additional Western therapies such as radiation or chemotherapy are required. The following is a testimonial by a patient of mine, Thea Sagen (female - age 50), illustrating the use of Qigong therapy in combination with radiation therapy.

Here is a summary of events surrounding my pancreatic surgery, June 16, 1997 at Community Hospital of the Monterey Peninsula.

"On June 16, 1997, emergency surgery to remove a fist-sized tumor from the tail of my pancreas, resulted in my former Sifu (Martial Arts Master), Jerry Alan Johnson, changing roles to become an "Energy Therapist." Thanks to my admitting physician, Dr. Stephanie Taylor, who facilitated the arrangement, Dr. Johnson was the first representative of Chinese Medical Qigong therapy allowed to enter surgery at the Community Hospital of the Monterey Peninsula.

As my Energy Therapist, Dr. Johnson conducted a preoperative session during which he facilitated the energetic and emotional release of my tumor. He then accompanied me into the four hour surgery, during which he maintained my energy levels, while reminding my tissues of their preoperative energetic patterns in order to facilitate rapid healing. He continued to work with me through the postoperative healing, and followed up several days after surgery. It was his vital support and insights that helped me to undergo the recommended course of radiation therapy.

The following is a synopsis of the Medical Qigong therapy that I experienced along with the traditional Western modalities used to treat my pancreatic islet-cell tumor.

Phase 1. The Preoperative Stage. Consisted of a Medical Qigong session directed toward putting me in touch with the tumor, my relationship to it, and the emotions associated with it. This was followed with an atonement with the Divine and relinquishment of control, removal of the armoring around my tumor, and subsequent release of emotions with it. I actually "saw" the disconnected energetic attachments of the tumor as they detached from my organs and were blowing like ostrich feather segments in a wind. I also saw a bright neon sign on the tumor that identified its emotional source. This entire process took only an hour, yet it facilitated closure with the tumor in order for me to easily release it during the surgery, which immediately followed the session. It also provided me with insights that helped my emotional healing over the months following surgery.

Phase 2. Surgery. I entered surgery without any sedatives and was filled with a tremendous sense of peace and empowerment in knowing that the hospital was honoring my energetic, as well as my physical needs. The surgery lasted for four hours during which Dr. Johnson connected with my energy fields and monitored, as well as maintained, my energy levels. He continued treating me, and working with my energetic fields, as the anesthesia was administered, and as the two surgeons opened my cavity, moved structures, removed the tumor (along with the tail of my pancreas and the spleen), and then sutured and closed my abdominal cavity.

Throughout the surgery Dr. Johnson stood at my head, where energy enters through the crown, and facilitated the running of energy into my center channel in the core of my body. He described his work as following the actions of the surgeons, and my body's response to these actions: the incision, the movement of internal organs, the resection of the tumor, and the final suturing. He used this opportunity to change the energetic grid formations in the tumor bed to prevent regrowth of the tumor, which was malignant.

Phase 3. The Postoperative Stage. This stage involved a continuation of the mitigation of the effects of surgery. He worked to tonify my Kidneys, to bring my temperature up, and to offset the shock of surgery, as well as used acupressure points to calm my heart and stomach during the nausea. I was reminded upon awakening, to do my Microcosmic Orbit exercise. Dr. Johnson then proceeded to "dredge" or comb over the incision to reconnect my pre-energetic channel and tissue patterns in order to facilitate healing, and to reduce scar tissue. The healing of the incision was fast. My doctors later remarked at how quickly I healed, and how little scar tissue there was compared to similar surgeries.

Phase 4. Follow-Up and Remedial Prescriptions. During the days following surgery, Dr. Johnson taught me several meditations: The "Divine Healing Light Meditation" for healing the tumor bed and incision, and the "Heaven and Earth Spleen Meditation" that fuses a ball of energy where the spleen once was – allowing me to continue to produce the energetic pattern of the spleen within my body. I began doing these meditations the day following surgery for the week of hospitalization, and continued to practice them along with the Microcosmic Orbit exercises throughout my recovery.

Phase 5. Radiation. After several weeks of recuperation, I was scheduled to received radiation treatments daily (Monday through Friday) for six weeks. Each day I would perform the Medical Qigong exercises (vocal and physical prescriptions) outdoors in a field overlooking the ocean.

Then I would drive to the hospital for the 15 minutes of radiation, followed by a session with Dr. Johnson every other day. I understood that the Medical Qigong exercises and sessions would offset the heat in the tissues while protecting the surrounding organs from unwarranted harm. There was concern about the possible loss of function in one kidney, but blood tests during the treatments showed no noticeable effects on them, or on my liver. In addition, unexpected by-products of the Medical Qigong sessions involved the release of emotional memories and an experience of the Divine in my healing process.

On alternate days after radiation, I had acupuncture and pranic healing with Jason Gill L.Ac. who teaches classes in Chinese Herbs and Traditional Chinese Medicine at Dr. Johnson's International Institute of Medical Qigong. Through these treatments, and the use of Chinese herbs, my immune system was strengthened. These complimentary modalities, combined with my body's own healing abilities and visualization practice, enabled me to avoid the typical side-effects of nausea and diarrhea. Aided by the positive energy from these two therapists, I was able to maintain my energy levels in sufficient quantity to continue driving to the radiation treatments on my own.

It is hard for me to imagine what this entire experience might have been like without the support of a doctor of Medical Qigong. I am truly grateful for Dr. Johnson's dedication, discipline and expertise that enabled him to be present for me throughout all aspects of this illness. It was a gift that I would encourage other patients to utilize for themselves in combination with any traditional allopathic program for cancer. I had the best of both worlds working for me in a remarkable manner thanks to all of my doctors."

PHANTOM PAINS OF AMPUTATED ORGANS

Amputation may be necessary as a result of severe injury or massive infection, such as gangrene. During surgical amputation, the soft tissues are cut in such a way that they cover the end of the remaining bone to heal completely. This allows the body's tissues to seal and protect themselves from the external environment. "Ghost" or "phantom" pains are caused by the patient's energetic connection with the amputated tissues, through the internal "energetic spacial cavity," or external "energetic boundary system." The phantom limb (usually one of the body's extremities, e.g., finger, hand, arm, etc.) exists in an energetic state and continues to interact with the patient's external Wei Qi field, as well as with the patient's Shen. Any interaction with the space, or energetic field of the amputated organ, will produce a direct response in the patient, similar to that of contacting the actual physical limb.

Rehabilitation programs generally do not prepare the patient for such reactions as phantom pains, which tend to leave the patient perplexed and frustrated. To alleviate the symptoms of phantom pains, the Qigong doctor sedates the patient's Wei Qi field that surrounds the amputated limb, and assists the patient in re-patterning the old energetic field to fit the postoperative tissue formation. The Qigong doctor can also prescribe Juniper Berries, which will relieve the phantom pain symptoms.

MEDICAL QIGONG THERAPY AND THE TREATMENT OF SCAR TISSUE

Medical Qigong therapy may be effective for treating pathologies caused by scarring due to surgical procedures. The formation of scar tissue often becomes the source of other physical problems by either obstructing, or weakening, the flow of Qi within the body's channels. In most cases of pathological scar tissue, the scar itself transverses several channels within the superficial and deep tissues.

The topological relationships of each scar will vary according to each patient. Therefore, different therapeutic approaches may be required for treatment, depending on the patient's strength, weakness, Elemental Constitution, and the patient's emotional state surrounding the surgery. Manifestations of scar tissue pathology include some of the following symptoms:

- a feeling of discomfort, or unpleasant sensations that originates around the scar area,

- radiating sensations that occur when the scar is being pressed, and
- a sharp, pulling type of pain that originates from the scar.

These conditions can be present, even if it has been several years since the surgery was performed.

TREATMENT

Medical Qigong treatment for scar tissue formation follows the principle of emitting Qi along the pathway of the channels. Before treatment, the doctor should take into consideration the energetic flow and function of the channels that have been affected by the surgery. If, for example, a patient's Yang channel (e.g., Stomach) becomes obstructed due to surgery, it can create an Excess condition within the Stomach Channel, which in turn creates a Deficiency of the Spleen Channel. Therefore, the treatment of both the Stomach and Spleen channels should be addressed.

After surgery, some scar tissue formation is inevitable, since both the body's inner-facia and channels (which are responsible for the lubrication of the tissues) have been severed. To minimize the degree of scar tissue formation, the Qigong doctor uses Qi extension techniques on the sutured area, extending his or her intention deep into the patient's body. The doctor's focuses on dispersing stagnations and regulating the flow of the undercurrent of Qi through the sutured tissues. Emphasis should be placed on reconnecting the patient's channels and collaterals. Next, the doctor focuses on regulating the external and internal flow of energy from within and without the patient's tissue area.

In treating scar tissue formation which has laid dormant for quite some time, tissue regulation through Tui Na therapy is required. After using the proper tissue regulation method (see Chapter 37), the Qigong doctor ends the treatment by using the One Finger Skill technique along the line of the patient's external scar. To perform this method, one thumb is placed on the end of the scar, while the doctor's other thumb presses into tissue, slowly outlining the entire scar (this heats and energizes the patient's tissues). The doctor ends the treatment with both thumbs placed at the beginning and ending of the scar, extending his or her energy through the patient's tissue in a circular pattern. This technique naturally induces an energetic emotional release of the memories trapped within the scar tissue.

SECTION XII

MEDICAL QIGONG APPENDICES

INTRODUCTION TO THE APPENDICES

There are many documents discussing the experimental research of Medical Qigong Therapy written within the past ten years. Among these documents, are listed a collection of abstracts from five national congresses on Medical Qigong researches as well as four World Conferences. The research has covered a variety of topics ranging from Medical Qigong and its relationship with: Physics, Chemistry, Biology, Western and Traditional Chinese Medicine, Optics, Magnetism, Electromagnetism, Acoustics and the Projection of Sound Resonation, Fluid Mechanics, Nuclear Physics, Electro-Chemistry, Organic Chemistry, Inorganic Chemistry, Analytical Chemistry, Physiology, Bio-Chemistry, Molecular Biology, Micro-Biology, Cytology, Genetics, Pathology, Pharmacology, Psychology, and Sports Medicine.

The experimental research of Medical Qigong is actually a study with multiple sciences. Of all the research reported, Biology and Medicine are in the forefront of the material collected.

The following research was conducted by distinguished scholars at various Universities, Scientific Academies, and Institutes through out the world (i.e.: Beijing Medical University, Qinghua Medical University, Nanjing Medical University, Tokyo Medical University, Tokyo Electro-engineering University, Higher Energy Institute, Physics Institute, International Institute of Medical Qigong, China Healthways Institute, etc.).

This research was focused on the Extension of Emitted Qi by various Medical Qigong Masters/Doctors, the result being positive in it's conclusion on the effectiveness of Medical Qigong Therapy as a solid and effective clinical modality.

This section of the text is devoted to the scientific approach of analyzing and scrutinizing the claims made as to the actual effectiveness of Medical Qigong Therapy as a clinical modality.

For this purpose, I have elicited the help of my good friend, the renowned scientist and research investigator, Dr. Kenneth M. Sancier. Dr. Sancier (a Material Scientist) is the president of the Qigong Institute, of San Francisco, California, and an active Council Board Member of the World Academic Society of Medical Qigong (Beijing, China). His many years of extensive research in the field of Energetic Medicine has led him to help organize the first International Congress of Qigong in conjunction with the Fifth International Congress of Chinese Medicine (at the University of California, Berkeley) including maintaining an active annual Qigong Science Forum for scientists. He is also responsible for developing the first computerized Qigong Data Base, which is currently being sold worldwide.

Dr. Sancier has actively collected scientific information on Qigong by attending international conferences, and meeting scientists in China, Japan, Canada and the United States. He has published several reviews of clinical and experimental studies of the medical aspects of Qigong therapy in many scientific journals and magazines. Although the following information in this appendix represents but a fraction of Dr. Sancier's findings, I am confident that the reader will find the data fascinating.

Dr. Jerry Alan Johnson May 2000

APPENDIX 1A
MEDICAL APPLICATIONS OF QIGONG

Kenneth M. Sancier, Ph.D.
Qigong Institute
Menlo Park, California

ABSTRACT

This article reviews selected scientific studies of medical applications of Chinese Qigong. The intention of the review is to outline research on Qigong and its potential for improving health care in western countries. The review centers on clinical and experimental studies to show that Qigong exercise can beneficially affect many functions of the body and improve health. The studies were selected to illustrate the following points: medical applications of Qigong are diverse, some studies were conducted in depth, and many applications hold promise to improve western health care.

Several specific clinical and experimental studies are outlined. The study using electroacupuncture according to Voll shows that almost all organs and functions of the body can be balanced by Qigong. The research on hypertension serves as a model to illustrate the many ways that the effects of Qigong on the body can be measured. Studies on brain waves are included to show that the mind is involved in Qigong and probably in the healing process. Research on Qigong's role in high altitude adaption suggests some practical applications for treating cardiovascular disease. A combination of Qigong and drug therapies is superior to drug therapy alone, as indicated in the case of two diseases, hypertension and cancer. In conclusion, there is a plea to medical practitioners and scientists to collect existing information on medical Qigong, promote research if needed, and identify applications that hold promise for improving western health care.

INTRODUCTION TO MEDICAL QIGONG

Traditional Chinese Medicine (TCM) is a holistic system for promoting health and for healing that includes several therapies such as acupuncture and moxibustion, herbal medicine, acupressure massage, nutrition, and Medical Qigong. These therapies are often used in combinations. The underlying theory of TCM is based on balancing Qi, the vital energy in the body, according to the theory of "yin-yang" and the Five Elements, that has been used for more than three thousand years. In practice, the flow of Qi is regulated, and "blocks" to the flow of Qi are removed. Energy blocks or Qi in excess or deficiency may result from disease, injury or stress. Qigong (pronounced chee gong) has a history of thousands of years, and it is unique among TCM therapies because almost anyone can learn and practice Qigong. However, it best to study Qigong with a qualified teacher to avoid adverse side effects. It is claimed that 60 million people in China practice Qigong daily, primarily to maintain health and achieve long life.

The word Qigong is a combination of two ideas. Qi is the vital energy of the body, and gong is the skill of working of the Qi. Medical Qigong for health and healing consists primarily of meditation, physical movements, and breathing exercises. Qigong practitioners develop an awareness of Qi sensations in their bodies and use their mind, i.e., intention, to guide the Qi in the body. It is said that the benefits of Qigong extend beyond health and healing to enhance spiritual life and even special abilities, such as psychic powers. Qigong is also used in martial arts to help develop physical and mental powers for self defense and healing.

Medical Qigong is divided into two parts: internal and external. Internal Qi is developed by individual practice of Qigong exercises. When Qigong practitioners become sufficiently skilled, they can use external Qi to "emit" Qi (Wai Qi in Chinese) for the purpose of healing another person. There are many reports of the medical efficacy of emitted Qi. This therapy has limited application on a large scale because there are a limited number of skilled Qigong masters. The present article will focus mainly on internal Qi because almost everyone can learn Qigong exercises for maintaining health and for self-healing.

In the early 1980s, scientists in China began investigating the many medical benefits claimed for Qigong. Since then, research on hundreds of medical applications of Qigong has been reported in the literature, which, unfortunately for the West, is written primarily in Chinese. A wealth of material is published in English in the proceedings of international conferences of Qigong. The proceedings contain abstracts of talks given at the conferences, and since 1986 there are 837 abstracts, more than half in English.

EXAMPLES OF MEDICAL APPLICATIONS OF QIGONG

Some examples of the applications of medical Qigong and emitted Qi on humans, animals, cell cultures, and plants were reviewed. Clinical and experimental evidence showing that Qigong exercise and external Qi affects various functions and organs of the body have been discussed. A short list of some of the functions and organs affected by Qigong, and the measurement techniques employed, include: the brain (EEG and magnetometer); blood flow (thermography, sphygmography, and rheoencephalography); heart functions (blood pressure, EKG, and UCG); kidney (urinary albumin assay); biophysical (enzyme activity, immune function, sex hormone levels); eyesight; and tumor size in mice.

In the following discussion, some clinical and experimental research studies will be discussed to illustrate the scope of research on medical applications of Qigong. The studies have been selected because they contain scientific information on important chronic medical conditions, such as hypertension and aging. Most of these studies on Qigong used controls and statistical analysis. This review focuses on the main accomplishments of selected research studies. Details may be omitted because of space limitations or because relevant information is not available. The critical evaluation of research studies will be left to medical specialists.

THERAPEUTIC BALANCING OF THE CHANNELS AND FUNCTIONS OF THE BODY

Electroacupuncture According to Voll (EAV) can monitor the effects of Qigong practice on therapeutic balancing of the Channel system, the classical TCM organs, and body functions. In EAV the electrical conductance of the skin above individual acupuncture points is measured using a low voltage and low current. Diagnosis depends on measuring the relative electrical conductance and its time dependence. An important diagnostic criterion of degeneration of an organ is an "indicator drop" that may occur during the measurement when the conductance reaches an apparent maximum value but then decreases before leveling off.

Measurements were made at 24 acupuncture points at the ends of the Channels of the fingers and toes of subjects and were made by the same operator and equipment. The subjects were asked to perform a Qigong exercise of their choosing, for example, sitting or standing meditation or moving Qigong. Two series of EAV measurements were made before and after healthy subjects practiced Qigong. In the first series, four subjects were examined by EAV before and after they practiced Qigong for 10 to 15 minutes. Qigong exercise decreased the average of the EAV measured values in the range of -19 to -31% ($p<0.004$) for the four subjects. Qigong essentially eliminated the indicator drops. In the second series, each of seven subjects was examined by EAV three times in a blind protocol so that the operator did not know whether a subject had practiced Qigong before the second or third examination. In the second series, Qigong exercise changed the average EAV measured values in the range of -17% to -35% for four subjects and in the range of 4% to 15% for three subjects. Indicator drops again were decreased.

These preliminary results show that Qigong can make significant changes in the therapeutic balancing of the Channel and organ systems, the goal of TCM.

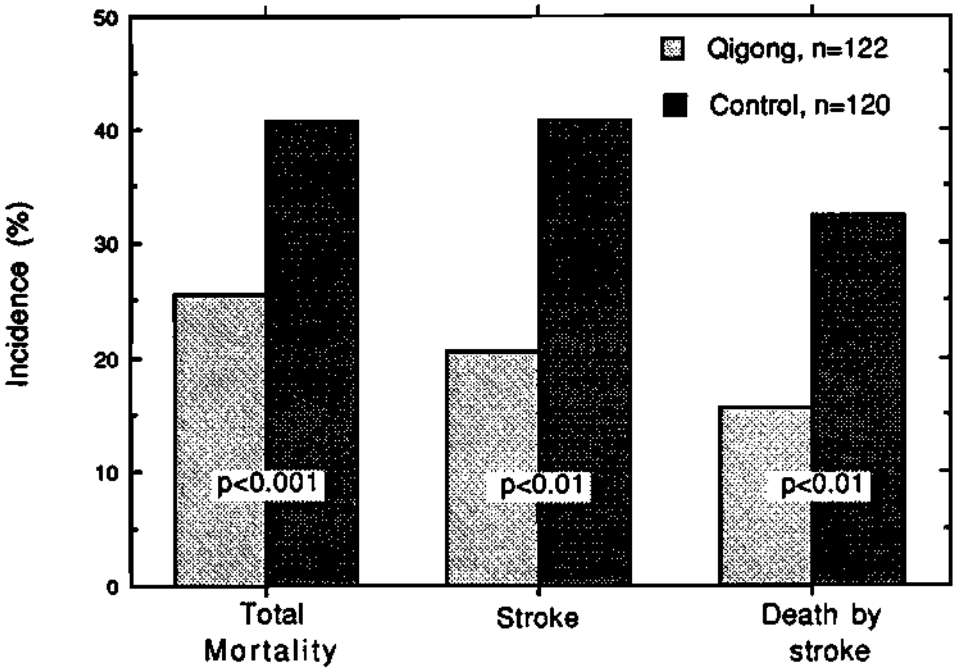

Figure 1. A 30-year folow-up of the effect of Qigong on mortality and stroke of 242 hypertensive patients. Both groups received drug therapy. (Qang. Xu, et al. ref. 7)

HYPERTENSION

Several groups in China have investigated the effects of Qigong on hypertension (i.e., high blood pressure). Research on the short- and long-term effects of Qigong practice on hypertensive patients has been carried out at the Shanghai Institute of Hypertension by Wang Chongxing, Xu Dinghai and their colleagues. Their research is reviewed extensively because it serves as a model of the effects of Qigong on many functions of the body. For these studies, the patients practiced "Yan Jing Yi Shen Gong" for 30 minutes twice a day. This Qigong is claimed to be especially valuable for therapeutic purposes and delaying senility. The Qigong exercise consists of a combination of sitting meditation and gentle physical movements that emphasizes a calm mind, relaxed body, and regular respiration.

PREVENTION OF STROKE AND MORTALITY

In 1991, the researchers reported a 20-year controlled study of the anti-aging effects of Qigong on 204 hypertensive patients. Recently, the researchers performed a 30-year follow-up on 242 hypertensive patients who were divided randomly into a Qigong group (n = 122) and a control group (n = 120). All patients were given drug therapy to control blood pressure, but only the experimental group practiced Qigong 30 minutes a day, twice a day. The results show that the accumulated mortality was 25.41% in the Qigong group and 40.79% in the control group ($p<0.001$). The incidence of stroke was 20.49% and 40.73% ($p<0.01$), and death rate due to stroke was 15.57% and 32.50% ($p<0.01$), respectively. These results, which are shown graphically in Figure 1, indicate that Qigong has potential for preventing strokes and extending life.

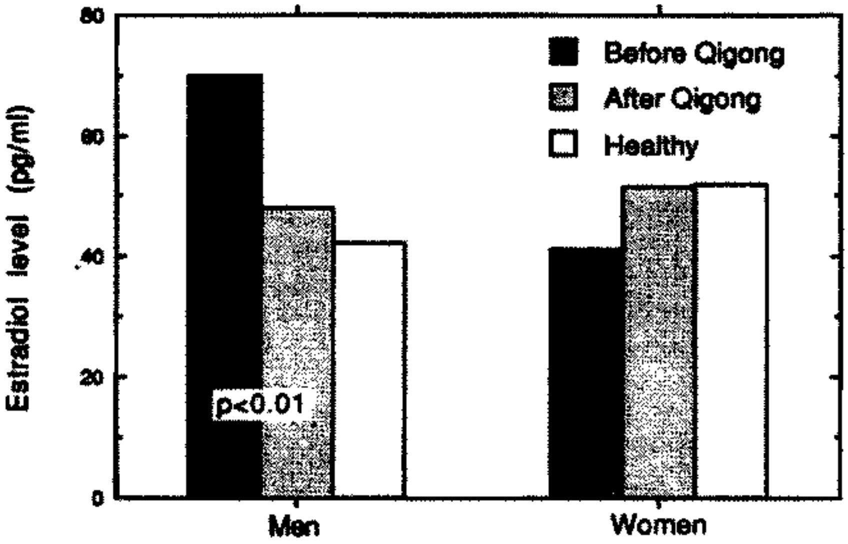

Figure 2. Serum estrdiol levels of hypertensive men (ages 40–69) and women (ages 51–57) who practice Qigong one year compared to healthy controls. (Ankun, Eang, et al. ref. 6)

The researchers also reported that over the 20-year period, blood pressure of the Qigong group stabilized while that of the control group increased. Remarkably, during this period the drug dosage for the Qigong group could be decreased and for 30% of the patients could be eliminated. However, the drug dosage for the control group had to be increased. These results suggest that Qigong activated the natural healing process of the body.

IMPROVEMENTS IN HEART FUNCTION AND MICROCIRCULATION

Aged hypertensive patients usually are found to have a deficiency of Heart-energy, which often leads to a weakened function of the left ventricle and a disturbance of microcirculation. The researchers evaluate the effects of Qigong by using ultrasonic cardiography (UCG) and indices of microcirculation. The 120 male subjects, 55 to 75 years of age, were divided into three groups: 46 cases of hypertensive subjects with Heart-energy deficiency, 34 cases without Heart-energy deficiency, and 40 cases with normal blood pressure. Patients whose blood pressure measured more than 160/95 mm mercury were accepted as subjects after regulation with anti-hypertensive drugs for 4 weeks.

The results showed that subjects with Heart-energy deficiency experienced several improvements: increases in cardiac output, ejection fraction mitral valve diastolic closing velocity, and mean velocity of circumferential fiber shortening, while the total peripheral resistance decreased ($p<0.05$-0.01). Significant changes did not occur in the group without Heart-energy deficiency.

Multiple quantitative evaluation of nailfold disturbance in microcirculation was made on the above three groups by observing 10 indices of abnormal conditions: configuration of micrangium, micrangium tension, condition of blood flow, slowdown of blood flow, thinner afferent limb, efferent limb and afferent limb ratio, color of blood, hemorrhage, and petechia. At the beginning of the study, the incidence of microcirculation obstruction for the above three groups was 73.91%, 26.47% and 17.50%, respectively. After practicing qigong for one year, the group with

Heart-energy deficiency showed a decrease in nailfold microcirculation obstruction from 73.91% to 39.31% ($p<0.05$). Significant changes did not occur in the group without Heart-energy deficiency.

These studies show that proper TCM diagnosis is essential for qigong therapy. The investigators emphasized that the kind of qigong must be selected according to the patient's condition.

IMPROVEMENT IN SEX HORMONE LEVELS

One consequence of aging is that the levels of sex hormones change in unfavorable directions. For example, female hormone (estrogen) levels tend to increase in men and decrease in women. Two studies indicate that this trend can be reversed by qigong exercise. The effect of qigong exercise on plasma sex hormone levels was determine for hypertensive men and women. The sex hormones levels were measured before and after qigong practice for one year.

The 70 male patients with essential hypertension (ages 40 to 69; disease stage II) were divided into two groups. For the qigong group ($n = 42$), the estradiol level (E2) decreased from 70.12 to 47.71 pg/ml, a decrease of 32% ($p<0.01$), while no significant changes occurred in the control group ($n = 20$). The testosterone levels (T) of both groups decreased about 7%. The value of E2 for the qigong group (47.71 pg/ml) approached that of healthy men (42.23±5.82 pg/ml) of the same age but without hypertension or cardiovascular, pulmonary, hepatic, renal, or endocrine diseases ($p<0.05$). For women (ages 51 to 67, the number in group not available), the aging process was associated with failure of ovarian function manifested by decreased E2 and increased T levels. Qigong resulted in an increase of E2 from 40.9±.3.5 to 51.6±3.5 pg/ml, a value about equal to that of normal menopausal controls without hypertension or cardiovascular, pulmonary, hepatic, renal, or endocrine diseases. The value of T was also increased by qigong from 25.5±2.2 to 37.2±2.2 ng/dl. These favorable changes in estradiol levels, E2, brought about by qigong are summarized in Figure 2 for both men and women.

In an auxiliary study, the 24-hour urinary estradiol levels were determined in 30 men ages 50 to 69. Qigong for one year resulted in a decrease of 31% in E2 and a decrease of 54% in the estradiol/testosterone ratio (E2/T). These changes were accompanied by improvements in symptoms associated with Kidney deficiency hypertension, such as soreness, dizziness, insomnia, hair loss, impotence, and incontinence. The average score for these symptoms was changed by qigong from 5.53±2.31 to 2.83±1.32 ($p<0.001$).

Ye Ming and co-workers reported similar favorable changes in plasma sex hormone levels E2 in 77 male and female qigong exercisers after 2 months qigong compared with 27 controls. They did not observe significant changes in testosterone.

The above studies show that qigong exercise can help restore the sex hormone levels that had deteriorated because of aging.

INCREASED BONE DENSITY

Aging may result in a decrease in bone density, especially in women. As a consequence, bones become more brittle and subject to fracture.

Bone density of male subjects was found to increase by practicing qigong for one year. For subjects 50 to 59 years of age (18 cases), the bone density increased from 0.627±0.040 to 0.696±0.069 g/cm3 ($p<0.01$). For subjects 60 to 69 years of age (12 cases), the bone density increase was somewhat less, namely, from 0.621±0.039 to 0.672±0.083 g/cm3 ($p<0.05$). For both age groups, the bone density increased to values exceeding those of normal men of the same age, 0.695±0.096 and 0.657±0.102 g/cm3, respectively. The results are shown graphically in Figure 3.

It is likely that qigong therapy also would help restore the bone density of women, especially menopausal women. If so, hormone replacement therapy and its side effects could be reduced.

CHANGES IN BLOOD CHEMISTRY IN HYPERTENSIVE PATIENTS

Auxiliary studies by Wang, Xu and co-workers on the effects of qigong exercise on blood chemistry of hypertensive subjects include improvements in plasma coagulation firbrinolysis indices, blood viscosity, erythrocyte deformation index, levels of plasma tissue-type plasminogen activator (tPA), plasminogen activator inhibitor (PAI), VIII factor related antigen (VIII R:AG), and anti-thrombin (AT-III). In another study, they reported that qigong exercise significantly and beneficially changed the activities of two messenger cyclic nucleotides (cAMP and cGMP).

These results provide further evidence that qigong affects many complex functions of the body.

CANCER

Feng Lida pioneered in research showing that emitted qi from qigong masters produced marked changes in cell cultures of cancer cells from mice. Several studies reported the effects of emitted qi on tumors in animals. For example, emitted qi was reported to inhibit the growth of implanted malignant tumors in mice but did not destroy the tumors. Encouraged by the results with animals, researchers carried out clinical research on the effects of qigong on human subjects with cancer. Detailed results are not available in English for all these clinical studies.

Some results are available for a clinical study of qigong as a therapeutic aid for patients with advanced cancer. In this study, 127 patients with medically diagnosed malignant cancer were divided into a qigong group of 97 patients and a control group of 30 patients. All patients received drugs, and the qigong group practiced qigong for more than 2 hours a day over a period from 3 to 6 months. Both groups improved, but the qigong group showed improvements four to nine times greater than the control group for strength, appetite, diarrhea free, and weight gain of 3 kg. The phagocytic rate, which a measure of the immune

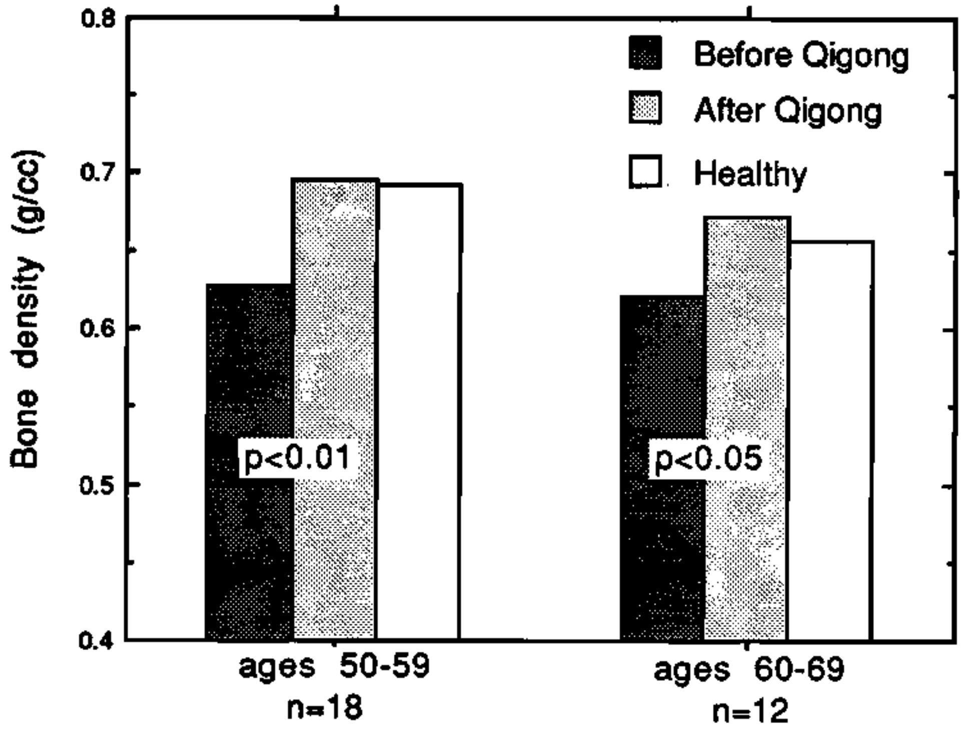

Figure 3. Changes in bone density of hypertensive men who practice Qigong one year compared to healthy males of the same age. (Xu Wang, et al. ref. 9)

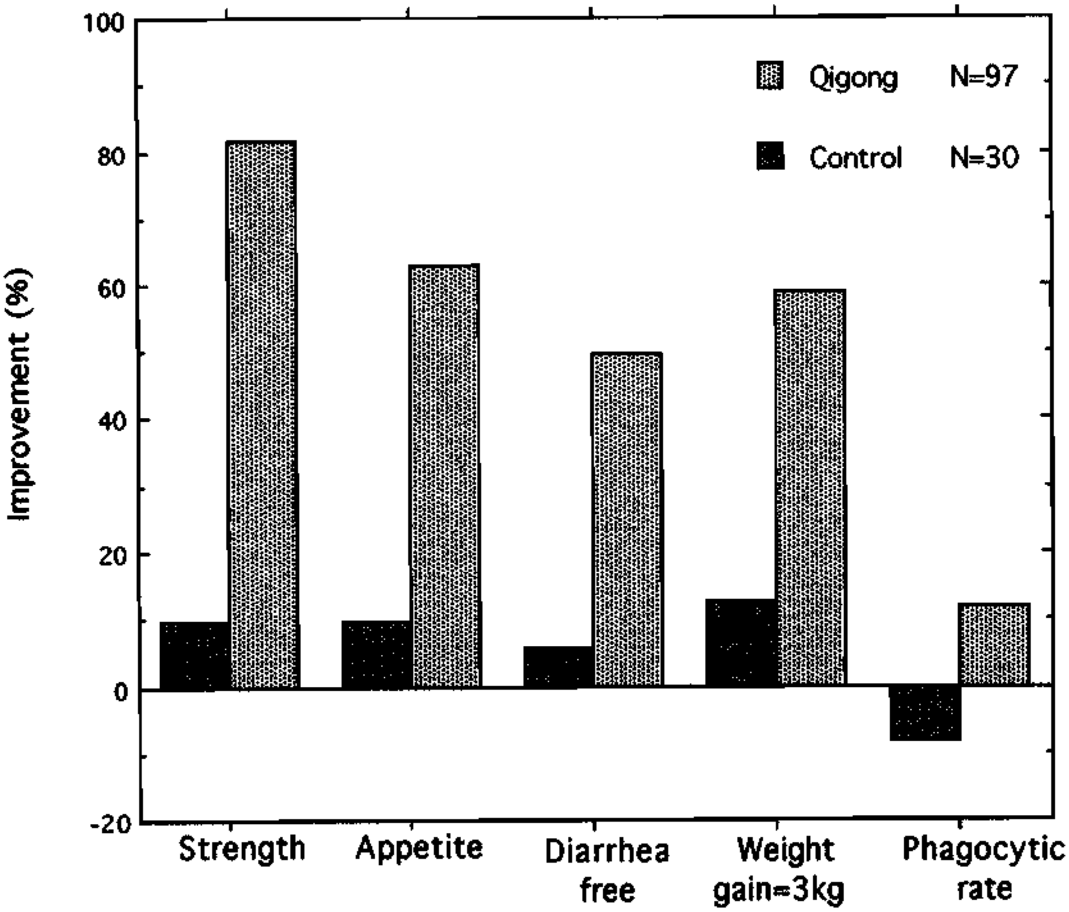

Figure 4. Comparison of Qigong, drug and drug-alone terapies for patients with advanced cancer. (ref. 25)

function, increased in the qigong group but decreased in the control group. These results are shown in Figure 4.

REVERSING SENILITY

To study the mechanism of keeping fit by qigong, a controlled study was made of 100 subjects classified either as presenile or with senile impaired cerebral function. The subjects were divided into two groups of 50 people each with a mean age of 62.7 years and with a similar distribution of age and sex. The qigong group practiced a combination of static and moving qigong. The control group exercised by walking, walking fast, or running slow. According to TCM method of classifying the vital energy, more than 80% of the patients in each group were classified as deficient in vital function and vital essence of the Kidney. Criteria for judging outcome were based on measuring clinical signs and symptoms including cerebral function, sexual function, serum lipid levels, and function of endocrine glands.

After six months, 8 of the 14 main clinical signs and symptoms in the qigong group were improved above 80%, whereas none of the symptoms in the control group was improved above 45%. These results suggest that qigong can reverse some symptoms of aging and senility.

MIND-BODY REGULATION

The main function of qigong is to regulate the mind, and according to TCM theory the mind leads the qi. The mind-regulating function of qigong mainly refers to the regulation of the functions of the brain and related body reactions. One tenant of qigong is that "the mind leads the qi, and the qi leads the blood." This somewhat mysterious statement can be interpreted to mean that intention (the mind) can direct the qi within the body. This mechanism is perhaps similar to the role of volition in biofeedback self-regulation.

BRAIN WAVES

The main focus of research has been on the effects of qigong on brain waves as measured by electroencephalography (EEG). During static, e.g., sitting meditation, alpha brain waves dominate over beta waves and spread to the frontal areas of the brain. Kawano and Wang have found differences in the EEG of Zen Buddhist priests and qigong masters. During almost all types of qigong training, the frequency of the alpha waves increased in the range from +0.6 to 1.0 Hz. During deep Zen meditation, the frequency decreased in the range from -1.0 to -1.5 Hz, and sometimes theta waves appeared. Also, frontal and occipital alpha waves tended to synchronize with a phase difference that depended on the type of meditation. This phase difference became smaller with qigong meditation (i.e., better synchronization) and larger for Zen meditation. According to Kawano and Wang these differences in brain function suggest that internal qigong involves a "semi-conscious" process, which involves some awareness and activity, while Zen meditation is a neutral process that releases the meditator from all concerns. This difference may relate to qigong being a healing art, whereas Zen is generally not considered to be one.

As mentioned earlier, a qigong master can emit qi to heal a patient, and the literature contains many references to this subject. The interaction between the qigong master and his subject has been followed in double blind tests. Both qigong master and subject were simultaneously measured by EEG, polygraphs, biochemical blood tests, and psychological tests. The EEG studies showed that brain wave patterns of the subject and qigong master were synchronized in type of brain waves and their location in the brain. Such synchronism may be required for healing by emitted qi.

Machi has studied qigong masters by combining EEG measurements with simultaneous measurements of physiological changes. He found that while the qigong master was emitting qi, the alpha-1 waves showed extremely high potential on the right frontal lobe, and there were increases in blood pressure, heart rate, and skin surface temperature. He also detected a far infrared emission with 1 Hz modulation signal coming from the Laogong point (a major acupuncture point in the palm of the hand.)

BLOOD FLOW TO THE BRAIN

Qigong exercise has been shown by rheoencephalography to increase blood flow to the brain. For 158 subjects with cerebral arteriosclerosis who practiced qigong for 1 to 6 months, improvements were noted in symptoms such as memory, dizziness, insomnia, tinnitus, numbness of limbs, and vertigo headache. During these studies, a decrease in plasma cholesterol was also noted. These results may offer hope to people with cerebral arteriosclerosis.

EFFECT OF RAPID AND LARGE CHANGES IN ALTITUDE

Studies have been made to determine whether qigong exercise would protect pilots from altitude stress when they increased altitude by going rapidly from low altitudes to the Tibetan highlands.

CARDIAC FUNCTION

Before entering the Tibetan highland, 66 persons were divided into two groups: a qigong group of 32 young men who did Qiyuan Qigong exercise for 4 weeks, and a control group of 34 young men who exercised to radio music. The two groups of men suddenly entered the highlands from a lower altitude. Before and after entering the highland, measurements were made of symptoms of altitude sickness and physiological changes. The qigong group suffered less altitude stress than the control group as measured by blood pressure, heart rate, oxygen consumption, microcirculation on epex of tongue and nail fold, and temperature at Laogong point (P8) of the left hand ($p<0.01$). The researchers suggest that qigong can prevent stresses from altitude changes.

These results with healthy subjects suggest that qigong also should be effective in improving the health of people with cardiovascular conditions. In fact several research studies have been reported the beneficial effects of qigong on cardiovascular diseases.

MICROCIRCULATION DISORDERS

Air force pilots were randomly divided into two groups: a qigong group of 22 persons who had practiced Qiyuan Qigong exercise for eight weeks, and a control group of 18 persons who did physical exercise for eight weeks before entering the Tibetan highlands. Microcirculation was measured at tongue apex and the nail fold, and also from the temperature at the Laogong point in palm of the left hand. When the men entered the high altitude, abnormal blood pressure and microcirculation of tongue apex and nail fold occurred in both groups. However, the abnormalities were statistically less in the qigong group than in the control group ($p<0.01$). The temperature at Laogong kept steady in the qigong group, but was reduced in the control group ($p<0.05$).

STUDY OF LUNG FUNCTION

The men were divided into three groups. Forty young males were in the qigong group and practiced Qi Yuan qigong for 4 weeks prior to entering the highlands; 40 young men were in the control group and exercised to radio music for 4 weeks prior to entering the highlands; and 40 young males were residents living at high altitudes. The results show that the integral value of symptoms of acute mountain sickness was lower in the qigong than in control group ($p<0.05$—0.01). Pulmonary ventilation of the qigong group was significantly improved compared with the control group ($p<0.05$—0.01), and nearly equal to the resident group.

The favorable outcome of these studies suggests that qigong may be a valuable therapy for people suffering from cardiac, microcirculation and lung disorders.

SUMMARY AND CONCLUSIONS

This review deals with only a small fraction of a large collection of research already performed on medical applications of qigong, mainly in China. Hopefully, the information presented is sufficient to illustrate the potential of qigong exercise for restoring normal body functions. The main conclusion from many studies is that qigong enables the body to heal itself.

Medical practitioners and scientists are encouraged to evaluate the potential value of medical applications of qigong. They should collect as much information as possible from original sources, develop bibliographies, and then determine whether additional research is required to validate given medical applications. They should then recommend those applications most likely to be suitable for western health care. It is recommended that TCM and qigong be considered to complement western medicine.

There are several issues that should be addressed for the successful introduction of qigong into western health care. It is necessary to win the support of scientists, medical practitioners and the public who know very little about TCM and qigong. A crucial step is to collect scientific evidence from research already performed. Regrettably, there are few research reports in English, and original scientific reports in Chinese are often difficult to obtain and translate. Further, some of the best technical research papers are incomplete scientifically. Recently this has become less of a problem in China. Although some scientific reports have flaws, we must not reject all research on qigong. We must give credit to China for developing the science of qigong and recognize that it is a science in development.

On the positive side, qigong is emerging rapidly among a public interested in improving their health. Qigong is a subject being addressed by the public press and in books and magazines. Increasing numbers of medical practitioners are becoming active in alternative medicine, although few physicians have studied TCM. At least one insurance company has paid for qigong therapy. HMOs and state and federally funded medical programs would benefit from paying for qigong therapy because it is effective and inexpensive. Pharmaceutical companies are already exploring many herbs, and these companies also could benefit from qigong research because of the synergic effect between qigong and drugs mentioned in this review. The Office of Alternative Medicine (OAM) of the National Institutes of Health is working to develop a basis for evaluating the merits of different alternative therapies, and in 1994 it funded one grant involving qigong. Incidentally, the Japanese Ministry of International Trade and Industry (MITI) is reported to be taking a pragmatic approach to alternative therapies by encouraging the development of industrial applications of therapies such as acupuncture and qigong.

There are many medical applications of qigong. Some of the most promising for introducing into western health care are chronic problems such as hypertension, cardiovascular disease, aging, asthma, allergies, neuromuscular problems, and cancer. These areas of public health deserve consideration by the western medical establishments.

*The p-value is the probability that two quantities are not different: the smaller the p-value greater the difference.

REFERENCES

1. Qigong Database™ is available from the Qigong Institute,450 Sutter Street, Suite 2104, San Francisco, CA 94108.
2. Sancier K M, Hu B. Medical Applications of Qigong and Emitted Qi on Humans, Animals, Cell Cultures, and Plants: Review of Selected Scientific Studies. Am J Acupuncture.1991:19 (4) 367-377.
3. Sancier K M. The effect of qigong on human body functions. Proceedings, Fifth International Symposium on Qigong, Shanghai, China.1994:179.
4. Sancier K M. The effect of qigong on therapeutic balancing measured by electroacupuncture according to Voll (EAV): a preliminary study. Acupuncture & Electro-Thera Res Int J. in press.
5. Wang Chongxing, Xu Dinghai, Qian Yuesheng, Medical and health care qigong, J Traditional Chinese Medicine. 1991:11(4) 296-301.
6. Kuang Ankun, Wang Chongxing, Xu Dinghai, Qian Yuesheng. Research on the anti-aging effect of qigong. J. Traditional Chinese Medicine. 1991:11 (2) 153-158.
7. Hong Shunhua, et al. Microcirculation of nail fold and immunogenicity after qigong practice for short periods. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 56-57.
8. Li Ziran, Li Liziang, Zhang Boli. Group observation and experimental research on the prevention and treatment of hypertension by qigong. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 113-114.
9. Jing Guinian. Observations on the curative effects of qigong self adjustment therapy in hypertension Proceedings, Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 115-117.
10. Bian Huangxian. Clinical observation of 204 patients with hypertension treated with qigong. Proceedings, First International Congress of Qigong. Berkeley, Calif., 1990: 101
11. Wang Chongxing, et al. The beneficial effect of qigong on the hypertension incorporated with coronary heart disease. Proceedings, Third International Symposium on Qigong, Shanghai, China. 1990: 40.
12. Wu Renzhao, Liu Zhewei. Study of qigong on hypertension and reduction of hypotensor. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 125.
13. Bornoroni Corrado, et al. Treatment of 30 cases of primary hypertension by qigong techniques. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 126.
14. Zhang Guifang. Development and application of a series of qigong feedback tapes Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 125.
15. Wang Chongxing, et al. Effects of qigong on preventing stroke and alleviating the multiple cerebro-cardiovascular risk factors—a follow-up report on 242 hypertensive cases over 30 years. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 123-124.
16. Wang Chongxing, Xu Dinghai, et al. Beneficial effects of qigong on the ventricular function and microcirculation of deficiency in heart-energy hypertensive patients. Pri-

vate communication.

17. Xu Dinghai, Wang Chongxing, et al. Clinical study of delaying effect on senility by practicing "Yang Jing Yi Shen Gong" in hypertensive patients. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, 1994:109.
18. Ye Ming, et al. Relationship among erythrocyte superoxide dismustase activity, plasma sexual hormones (T, E2), aging and qigong exercise. Proceedings, Third International Symposium on Qigong, Shanghai, China. 1990:28-32 (in English and Chinese).
19. Kuang Ankun, Wang Chongxing, Xu Dinghai, Qian Yueshang. Research on "anti-aging" effect of qigong. J Traditional Chinese Medicine. 1991:11 (3) 224-227.
20. Feng Lida, Effect of emitted qi on human carcinoma cells. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 1-4.
21. Feng Lida, Effect of emitted qi on the L 1210 cells of leukemia in mice. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 4-5.
22. Zhao Shan, et al. Preliminary observation of the inhibitory effect of emitted qi on transplanted tumors in mice. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 46-48.
23. Liu Tehfu,Wan Minsheng,Lu Oulun. Experiment of the emitted qi on animals. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 60-61.
24. Feng Lida,Peng Liaomin. Effect of emitted qi on prevention and treatment of tumors in mice. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 106-107.
25. Sun Quizhi, Zhao Li. Clinical observation of qigong as a therapeutic aid for advanced cancer patients. Proceeding, First World Conf Academic Exchange of Medical Qigong, Beijing, China. 1988: 97-98.
26. Chen Guoguang. The curative effect observed for 24 (cancer) cases under my emitted qigong treatment. Proceedings, Second International Conference on Qigong, Xi'an, China. 1989: 141-142.
27. Lo Jifeng, et al. Changes of peripheral blood cell population and immune functions in 31 nasopharyngeal carcinoma (NPC) patients treated with radiotherapy and qigong. Proceedings ,Third National Academic Conference on Qigong Science, Guangzhou, China. 1990: 94-95.
28. Yu Yi, et al. Effect of self-controlling qigong therapy on the immune function of cancer patients. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993: 128.
29. Zhao Hongmei, Bian Jingnan. Curative effect of intelligence qigong on 122 tumor patients. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993:130.
30. Wang Ying. Clinical observation on 30 cases of cancer treated by qigong therapy Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 131.
31. Xong Jing, Lu Zhong. Curative effect on 120 cancer cases treated by Chinese-Western medicine and qigong therapy. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 131
32. Fu Jingzhi. Treatment of advanced gastric cancer in the aged by the combination of qigong and medicinal herbs. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 132-133.
33. Zhang Shengbing. Effects of mind-regulation by qigong on the human body. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, September. 1994:68.
34. Kawano Kimiko, Wang Fengfong. Difference between the EEG of Zen priests and internal qigong masters. Society of Mind-Body Science. 1994: 3 (1) 99-104. (in Japanese).
35. Machi Yoshio. Various measurements to qigong masters for analyzing qigong mechanism. Society for Mind-Body Science. 1994:3 (1) 65-87. (in Japanese).
36. Kashiwasake Masaki. Double-blind tests of qi transmission from qigong masters to untrained volunteers. Society for Mind-Body Science. 1993:2 (1) 81-87 (in Japanese).
37. Machi Yoshio. Various measurements of qigong masters for analyzing qigong mechanism Society of Mind-Body Science. 1994:3 (1) 65-87. (In Japanese).
38. Liu Yuanliang, He Shihai, Xie Shanling. Clinical observation on the treatment of 158 cases of cerebral arteriosclerosis by qigong. Proceedings, Second World Conference on Academic Exchange of Medical Qigong, Beijing, China. 1993:125.
39. Liu Yuanliang. Private communication, 1993.
40. Mo Feifan, Xu Yongchun, Lu Yongpin, Xu Guang. Study of prevention of cardiac function disorder due to immediate entry into highlands by qigong exercise. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 78.
41. Chu Weizong, et al. Changes of blood viscosity and RCG in 44 cases with cardiovascular diseases after qigong ex-

ercises. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 57-58.

42. Qin Chao, et al. Bidirectional adjustment of blood pressure and heart rate by daoyin tuina on the arterial blood and heart rate. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 107.
43. Wang Chongxing, et al. Beneficial effect of qigong on improving the heart function and relieving multiple cardiovascular risk factors. Proceedings ,Third International Symposium on Qigong. Shanghai, China. 1990: 42 (in Chinese).
44. Mo Feifan, Wan Lurong, Jia ZiZi, Xu Guang. Study of prevention of microcirculation disorders of pilots in highlands by qigong. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993: 78.
45. Mo Feifan, Lu Yongpin, Zhao Guoliang. Effect of exercise with qigong on lung function of persons entering highland. Proceedings, Fifth International Symposium on Qigong, Shanghai, China.1994:186.
46. Wang Chongxing, Xu Dinghai, et al., private communication.
47. Wang Chongxing, Xu Dinghai, et al. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, September. 1994::22 (Chinese),109.
48. Wen-hsien Wu, Intractable reflex sympathetic dystrophy
49. McGee C T, Chow E P Y. Miracle Healing from China — Qigong. 1st ed. Coeur d'Alene, ID: Medipress; 1994: 8-19.

APPENDIX 1B
ANTI-AGING BENEFITS OF QIGONG

Kenneth M. Sancier, Ph.D.

INTRODUCTION

In the early 1980s, scientists in China began to study the medical benefits claimed for qigong. Since then, research on hundreds of medical applications of qigong have been reported in the Chinese literature. Of special interest for the present article are clinical reports of the medical benefits of qigong that claim to retard or reverse some diseases associated with aging.

Most of the original research was reported in Chinese, but access in English to most of this material is possible by reference to the proceedings of international conferences of qigong. Since 1986, ten such proceedings contain about 840 abstracts of talks given at the conferences, more than half of which are in English. These abstracts, along with about 160 abstracts of articles in the scientific literature, have been organized as a computerized database. The database enables searches and development of bibliographies across this entire body of information by using any key word. The clinical outcomes reported in this article are partly based on material in the database and partly on the author's person contacts with researchers.

The word qigong is a combination of two ideas: qi the vital energy of the body, and gong the skill of working of the qi. Medical qigong for health and healing consists primarily of meditation, physical movements, and breathing exercises. Qigong practitioners develop an awareness of qi sensations in their bodies and use their mind, i.e., intention, to guide the qi in the body. The benefits of qigong are said to extend beyond health and healing to enhance spiritual life and even special abilities, such as psychic powers.

Medical qigong is divided into two parts: internal and external. Internal qi is developed by individual practice of qigong exercises. When qigong practitioners have sufficiently mastered the skill, they can "emit" qi (external qi or waiqi in Chinese) for the purpose of healing another person. There are many scientific reports of the medical existence and efficacy of emitted qi. The present article focuses mainly on internal qi because almost everyone can learn qigong exercises for maintaining health and for self-healing, whereas, there are a limited number of skilled qigong masters available for healing.

There are numerous reports of the effects of emitted qi on living systems and the functions and organs of the human body. The present author reviewed some examples of medical applications of qigong and emitted qi on humans, animals, cell cultures, and plants, and he also published some of his experimental research on physiological effects of qigong. He discussed some clinical and experimental evidence showing that qigong exercise and external qi affects various functions and organs of the body. A short list of some of the functions and organs affected by qigong, and the measurement techniques employed (in parentheses), include: the brain (EEG and magnetometer); blood flow (thermography, sphygmography, and rheoencephalography); heart functions (blood pressure, EKG, and UCG); kidney (urinary albumin assay); biophysical (enzyme activity, immune function, sex hormone levels (laboratory analysis); eyesight (clinical); and tumor size in mice.

CLINICAL STUDIES INDICATING THE ANTI-AGING BENEFITS OF QIGONG

Several clinical studies will be described to illustrate the scope of research on medical applications of qigong to treat chronic medical conditions that may affect the aging process. Some details may be omitted because of space limitations. The critical evaluation of the research studies will be left to medical specialists.

THERAPEUTIC BALANCING OF THE CHANNELS AND FUNCTIONS OF THE BODY BY QIGONG

The profound effect that internal qigong practice may have on balancing the energies of the organs and functions of the body is illustrated by measurements using Electroacupuncture According to Voll (EAV). In EAV the electrical conductance of the skin above individual acupuncture points is measured using low voltage and low current. Diagnosis depends on measuring the relative electrical conductance and its time dependence. An important diagnostic criterion of degeneration of an organ is an "indicator drop" that may occur during the measurement when the conductance reaches an apparent maximum value but then decreases before leveling off.

Measurements were made at 24 acupuncture points at the ends of the 12 Channels in the fingers and toes of subjects and were made by the same operator and equipment. The subjects were asked to perform a qigong exercise of their choosing for 10 to 15 minutes, for example, sitting or standing meditation or moving qigong. Two series of EAV measurements were made before and after healthy subjects practiced qigong.

In the first series, four subjects were examined by EAV before and after they practiced qigong. Qigong exercise decreased the average of the EAV measured values of the four subjects in the range of -19 to -31% ($p<0.004$). Qigong eliminated almost all the indicator drops.

In the second series, each of seven subjects was examined by EAV three times in a blind protocol so that the operator did not know whether a subject had practiced qigong before the second or third examination. The results showed that qigong exercise changed the average EAV measured values in the range of -17 to -35% for four subjects and in the range of 4 to 15% for three subjects. Indicator drops again were markedly decreased.

These preliminary results show that internal qigong practice can make significant changes in the therapeutic balancing of the Channel and organ systems.

In a similar type of study, the electric current at acupuncture points on 14 Channels was measured using a single square wave voltage pulse technique. Both a qigong master, who emitted qi, and a qi-receiver were measured simultaneously and continuously. The results show that internal and external qigong produce different values in some measurement parameters, and also some synchronous behavior was observed between the sender and receiver.

BLOOD PRESSURE

The effect of qigong exercise on blood pressure is shown graphically in Fig. 1. The blood pressure (systolic and diastolic) in millimeters mercury is plotted as a function of time over 20 years for the group consisting of 242 patients, 122 in the qigong group and 120 in the control group. During the first two months, the blood pressure of all patients dropped in response to the hypotensive drug. Subsequently, and over the period of 20 years, the blood pressures of the qigong group stabilized while that of the control group increased. Remarkably, during this period the drug dosage for the qigong group could be decreased, while the dosage for the control group had to be increased.

Fig. 1. Effect of qigong on blood pressure of hypertensive patients over 20 years. Qigong group (n = 104) practiced 30 min/day twice/day, control group (n = 100).

MORTALITY AND STROKE

The incidences of mortality and stroke for the 30-year study are shown in Fig. 2. These results show that qigong exercise decreased by about 50 percent the incidence of total mortality, mortality due to stroke, and morbidity due to stroke. At the end of 30 years, 86 patients survived in the qigong group and 68 in the control group. These results clearly show that qigong has significant potential for preventing strokes and extending life.

Fig. 2. Effect of qigong on mortality and stroke of hypertensive patients over 30 years. Both groups received drug therapy; qigong group (n = 122) practiced qigong 30 min/day twice/day; the control group (n = 120).

IMPROVEMENTS IN HEART FUNCTION AND MIRCROCIRCULATION

Aged hypertensive patients usually are found to have a deficiency of Heart-energy, which often leads to a weakened function of the left ventricle and a disturbance of microcirculation. The researchers evaluated the effects of qigong for120 aged patients by using ultrasonic cardiography (UCG) and indices of microcirculation.

Experiments showed that the left ventricular function (LVF) in the hypertensive aged group (80 cases) was lower than that in the aged normal blood pressure group (40 cases), while the LVF in the deficiency of Heart-energy hypertensive patients (46 cases) was lower than in the non-deficiency Heart-energy hypertensive patients (34 cases).

After practicing qigong for one year, cardiac output was increased, the total peripheral resistance decreased, and the ejection fraction mitral valve diastolic closing velocity and the mean velocity of circumferential fiber shortening tended to be increased. Significant changes did not occur in the group without Heart-energy deficiency.

Quantitative evaluation of nailfold disturbances in microcirculation was made on the above groups by observing 10 indices of abnormal conditions: configuration of micrangium, micrangium tension, condition of blood flow, slowdown of blood flow, thinner afferent limb, efferent limb and afferent limb ratio, color of blood, hemorrhage, and petechia. The results showed that hypertension had an accelerating effect on the disturbance of microcirculation. The incidence of disturbance of microcirculation disturbance was 73.9% in the deficiency of Heart-energy hypertensive patients. After a year of qigong practice, the incident of disturbance was 39.1% ($p<0.01$).

The results suggest that qigong exercise has beneficial effects on Heart-energy and regulation of the blood channel, and qigong seems to have improved abnormal conditions of blood circulation.

ENHANCED ACTIVITY OF ANTI-AGING ENZYME SOD

Superoxide dismutase (SOD) is produced naturally by the body but its activity declines with age. SOD is often called an anti-aging enzyme because it is believed to destroy free radicals that may cause aging. The effects of qigong exercise to treat disorders of retired workers were studied by Xu Hefen and coworkers and included determinations of plasma SOD.

For their study, 200 retired workers,100 males and 100 females, ranging in age from 52 to 76 were divided into 2 groups: the qigong exercise group and the control group, and each group consisted of 50 males and 50 females. The main qigong exercise was Emei Nei Gong (one kind of qigong exercises of the Emei School), and was practiced at least 30 minutes a day.

The result showed that the mean level of SOD was increased by qigong exercise. For example, the SOD level was larger in the qigong group (about 2700 μ / g Hb) and than in the control group (1700 μ / g Hb), and this difference was significant ($p<0.001$). This study shows that qigong exercise can stimulate physical metabolism, promote the circulation of Channels and regulate the flowing of qi and blood, thus preventing and treating disorders of aging and promoting longevity.

CARDIOVASCULAR FUNCTION

Several studies reveal the potential benefits that qigong may have for improving the cardiovascular function of those with heart disease as well as old people. This conclusion is based on three studies reporting that qigong exercise can protect healthy pilots from altitude stress when they flew rapidly from a low altitude to the high altitude of the Tibetan highlands.

Before entering the Tibetan highland, 66 healthy young men were divided into two groups: a qigong group of 32 men who did Qiyuan Qigong exercise for 4 weeks, and a control group of 34 men who exercised to radio music. The two groups of men rapidly entered the highlands from a lower altitude. Before and after entering the highland, measurements were made of symptoms of altitude sickness and physiological changes.

The qigong group suffered less altitude stress than the control group as measured by blood pressure, heart rate, oxygen consumption, microcirculation on the epex of tongue and the nail fold, and the temperature at the Laogong point of the left hand ($p<0.01$). The researchers suggest that qigong can prevent stress from altitude changes.

In another study of changes in altitude, healthy young men were divided into three groups. Forty males were in the qigong group and practiced Qiyuan qigong for 4 weeks prior to entering the highlands; 40 men were in the control group and exercised to radio music for 4 weeks prior to entering the highlands; and 40 males were residents living at high altitudes. The results show that the integral value of symptoms of acute mountain sickness was lower in the qigong than in control group ($p<0.01$ to 0.05). Pulmonary ventilation of the qigong group was significantly improved compared with the control group ($p<0.01$ to 0.05), and nearly equal to the resident group.

In another study, air force pilots were randomly divided into two groups: a qigong group of 22 men who had practiced Qiyuan Qigong exercise for eight weeks, and a control group of 18 men who did physical exercise for eight weeks before entering the Tibetan highlands. Microcirculation was measured at tongue apex and the nail fold, and also from the temperature at the Laogong point in palm of the left hand. When the men entered the high altitude, abnormal blood pressure and microcirculation of tongue apex and nail fold occurred in both groups. The abnormalities were statistically less in the qigong group than in the control group ($p<0.01$). The temperature at Laogong kept steady in the qigong group, but was reduced in the control group ($p<0.05$).

The results of these three studies with healthy subjects lead to the conclusion that qigong also should be effective in improving the health of people with cardiovascular conditions including the aged. In fact several research studies have reported such beneficial effects of qigong on cardiovascular diseases. The three studies also provide evidence that qigong exercise is superior to physical exercise such as calisthenics.

COMBINATION OF MEDICAL QIGONG AND DRUGS THERAPY

There is ample evidence in the literature that therapy by a combination of qigong exercise and drugs is superior to that of drugs alone. The advantages of a combination therapy of qigong and drugs over drugs alone were discussed earlier in this paper for hypertension and cancer.

The mechanism of this apparent synergism is not entirely understood, but undoubtedly relates to the fundamental mechanism of qigong. Qigong is believed to remove blocks to the ready flow of the qi (energy), blood, oxygen and nutrients to all cells of the body as well as to promote removal of waste products from cells of the body. Blocks to energy (qi) flow may result from injury, disease or stress.

Increases in qi flow and blood circulation help nourish diseased or stressed tissue, providing a means for the body to heal itself. This mechanism suggests that qigong also could promote drug uptake to tissue and cells via increased blood circulation. Omura's research shows that drug uptake was increased by using qigongized paper (i.e., paper to which emitted qi was sent) applied to afflicted area of the body.

CONCLUSIONS

This review deals with a small fraction of the large collection of clinical research on medical applications of qigong. The information presented is intended to illustrate the potential of qigong exercise for restoring normal body functions in people with chronic conditions, many of which accelerate the aging process. The main conclusion from many studies is that qigong exercise helps the body to heal itself. In this sense, qigong is a natural anti-aging medicine. Two studies indicate that qigong exercise is superior to some physical exercises.

Qigong can complement Western medicine in many ways to provide better healthcare. For example, qigong has special value for treating chronic conditions and as a preventive medicine, whereas Western medicine has special value for treating acute conditions. There are many medical applications of qigong that can complement Western medicine to improve health care. Some examples include chronic problems such as hypertension, cardiovascular disease, aging, asthma, allergies, neuromuscular problems, and cancer. These areas of public health deserve consideration by the Western medical establishment.

REFERENCES

1. Kenneth M. Sancier, Ph.D., Copresident & Director of Research, 561 Berkeley Avenue, Menlo Park, CA 94025, USA. Phone/Fax +1-415-323-1221.
2. Some of the material in this article was adapted from the article, "Medical Applications of Qigong," by K.M. Sancier, Ph.D., and published in Alternative Therapies in Health and Medicine. 1996;2 (1): 40-46.
3. Qigong Database™ is available from the Qigong Institute, East West Academy of Healing Arts, 450 Sutter Street, Suite 2104, San Francisco, CA 94108, USA.
4. The p-value is the probability that two quantities are not the same: the smaller the p-value smaller the probability and the more significant the result.
5. Wang C, Xu D. Private communication, May 1995.
6. Duration of the study is not available.

ENDNOTES

[1] Sancier K M, Hu B. Medical Applications of Qigong and Emitted Qi on Humans, Animals, Cell Cultures, and Plants: Review of Selected Scientific Studies. Am J Acupuncture.1991:19 (4) 367-377.

[2] Sancier, KM, Medical applications of qigong. Alternative Therapies in Health & Medicine. 1995; 2(1) in press.

[3] Sancier KM, Chow EPY. Healing with qigong and quantitative effects of qigong. J. American College of Trad. Chinese Medicine. 1989: 7(3):13-19.

[4] Sancier KM. The effect of qigong on therapeutic balancing measured by electroacupuncture according to Voll (EAV): a preliminary study. Acupuncture & Electro-Therapy Res Int J.:1995;19:119-127.

[5] Sancier K M. The effect of qigong on human body functions. Proceedings, Fifth International Symposium on Qigong, Shanghai, China.1994:179.

[6] Kido M. Channel measurements of qi-gong operation and synchronous phenomena. J. Mind-Body Science: 1993: 2(1): 19-26.

[7] Wang Chongxing, Xu Dinghai, Qian Yuesheng, Medical and health care qigong, J Traditional Chinese Medicine. 1991:11(4) 296-301.

[8] Kuang Ankun, Wang Chongxing, Xu Dinghai, Qian Yuesheng. Research on the anti-aging effect of qigong. J. Traditional Chinese Medicine. 1991:11 (2) 153-158.

[9] Hong Shunhua, et al. Microcirculation of nail fold and immunogenicity after qigong practice for short periods. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 56-57.

[10] Li Ziran, Li Liziang, Zhang Boli. Group observation and experimental research on the prevention and treatment of hypertension by qigong. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 113-114.

[11] Jing Guinian. Observations on the curative effects of qigong self adjustment therapy in hypertension Proceedings, Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 115-117.

[12] Bian Huangxian. Clinical observation of 204 patients with hypertension treated with qigong. Proceedings, First International Congress of Qigong. Berkeley, Calif., 1990: 101

[13] Wang Chongxing, et al. The beneficial effect of qigong on the hypertension incorporated with coronary heart disease. Proceedings, Third International Symposium on Qigong, Shanghai, China. 1990: 40.

[14] Wu Renzhao, Liu Zhewei. Study of qigong on hypertension and reduction of hypotension. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 125.

[15] Bornoroni Corrado, et al. Treatment of 30 cases of primary hypertension by qigong techniques. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 126.

[16] Zhang Guifang. Development and application of a series of qigong feedback tapes Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 125.

[17] Wang Chongxing, et al. Effects of qigong on preventing stroke and alleviating the multiple cerebro-cardiovascular risk factors—a follow-up report on 242 hypertensive cases over 30 years. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 123-124.

[18] Wang Chongxing, Xu Dinghai, et al. Beneficial effects of qigong on the ventricular function and microcirculation

of deficiency in heart-energy hypertensive patients. Private communication, January 1994.

[19] Xu Dinghai, Wang Chongxing, et al. Clinical study of delaying effect on senility by practicing "Yang Jing Yi Shen Gong" in hypertensive patients. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, 1994:109.

[20] Ye Ming, et al. Relationship among erythrocyte superoxide dismustase activity, plasma sexual hormones (T, E2), aging and qigong exercise. Proceedings, Third International Symposium on Qigong, Shanghai, China. 1990:28-32 (in English and Chinese).

[21] Kuang Ankun, Wang Chongxing, Xu Dinghai, Qian Yueshang. Research on "anti-aging" effect of qigong. J Traditional Chinese Medicine. 1991:11 (3) 224-227.

[22] Xu, Hefen; Xue, Huining; Bian, Meiguang; Zhang, Chengming; Zhou, Shuying. Clinical study of the anti-aging effect of qigong. Proceedings, Second World Conference for Academic Exchange of Medical Qigong. Beijing, China. 1993: 137.

[23] Mo Feifan, Xu Yongchun, Lu Yongpin, Xu Guang. Study of prevention of cardiac function disorder due to immediate entry into highlands by qigong exercise. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 78.

[24] Mo Feifan, Lu Yongpin, Zhao Guoliang. Effect of exercise with qigong on lung function of persons entering highland. Proceedings, Fifth International Symposium on Qigong, Shanghai, China.1994:186.

[25] Mo Feifan, Wan Lurong, Jia ZiZi, Xu Guang. Study of prevention of microcirculation disorders of pilots in highlands by qigong. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993: 78.

[26] Chu Weizong, et al. Changes of blood viscosity and RCG in 44 cases with cardiovascular diseases after qigong exercises. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 57-58.

[27] Qin Chao, et al. Bidirectional adjustment of blood pressure and heart rate by daoyin tuina on the arterial blood and heart rate. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 107.

[28] Wang Chongxing, et al. Beneficial effect of qigong on improving the heart function and relieving multiple cardiovascular risk factors. Proceedings ,Third International Symposium on Qigong. Shanghai, China. 1990: 42 (in Chinese).

[29] Zhang Shengbing. Effects of mind-regulation by qigong on the human body. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, September. 1994:68.

[30] Liu Yuanliang, He Shihai, Xie Shanling. Clinical observation on the treatment of 158 cases of cerebral arteriosclerosis by qigong. Proceedings, Second World Conference on Academic Exchange of Medical Qigong, Beijing, China. 1993:125.

[31] Liu Yuanliang. Private communication, 1993.

[32] Feng Lida, Effect of emitted qi on human carcinoma cells. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 1-4.

[33] Feng Lida, Effect of emitted qi on the L 1210 cells of leukemia in mice. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 4-5.

[34] Zhao Shan, et al. Preliminary observation of the inhibitory effect of emitted qi on transplanted tumors in mice. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 46-48.

[35] Liu Tehfu,Wan Minsheng,Lu Oulun. Experiment of the emitted qi on animals. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 60-61.

[36] Feng Lida,Peng Liaomin. Effect of emitted qi on prevention and treatment of tumors in mice. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 106-107.

[37] Sun Quizhi, Zhao Li. Clinical observation of qigong as a therapeutic aid for advanced cancer patients. Proceeding, First World Conf Academic Exchange of Medical Qigong, Beijing, China. 1988: 97-98.

[38] Chen Guoguang. The curative effect observed for 24 (cancer) cases under my emitted qigong treatment. Proceedings, Second International Conference on Qigong, Xi'an, China. 1989: 141-142.

[39] Lo Jifeng, et al. Changes of peripheral blood cell population and immune functions in 31 nasopharyngeal carcinoma (NPC) patients treated with radiotherapy and qigong. Proceedings ,Third National Academic Conference on Qigong Science, Guangzhou, China. 1990: 94-95.

[40] Yu Yi, et al. Effect of self-controlling qigong therapy on the immune function of cancer patients. Proceedings,

Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993: 128.

[41] Zhao Hongmei, Bian Jingnan. Curative effect of intelligence qigong on 122 tumor patients. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993:130.

[42] Wang Ying. Clinical observation on 30 cases of cancer treated by qigong therapy Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 131.

[43] Xong Jing, Lu Zhong. Curative effect on 120 cancer cases treated by Chinese-Western medicine and qigong therapy. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 131

[44] Fu Jingzhi. Treatment of advanced gastric cancer in the aged by the combination of qigong and medicinal herbs. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 132-133.

[45] McGee C T, Chow E P Y. Miracle Healing from China—Qigong. 1st ed. Coeur d'Alene, ID: Medipress; 1994: 8-19.

APPENDIX 1C

THE EFFECT OF QIGONG ON THERAPEUTIC BALANCING MEASURED BY ELEXTROACUPUNCTURE ACCORDING TO VOLL (EAV): A PRELIMINARY STUDY

Kenneth M. Sancier, Ph.D.

ABSTRACT

Electroacupuncture According to Voll (EAV) was used to monitor the effects of qigong practice on therapeutic balancing of subjects. In EAV the electrical conductance of the skin above individual acupuncture points is measured using low voltage and current. Diagnosis depends on measuring the relative electrical conductance and its time dependence. An important diagnostic criterion of degeneration of an organ is an indicator drop which occurs during the measurement when the conductance decreases from an apparent maximum value and then levels off. Two series of EAV measurements were made before and after healthy subjects practiced qigong. Measurements were made at 24 acupuncture points at the ends of the Channels of the fingers and toes of a subject and were made by the same operator and equipment. The subjects were asked to perform a qigong exercise of their choosing. In the first series, four subjects were examined by EAV before and after qigong exercise. Qigong exercise decreased the average EAV measured values of the four subjects in the range of -19 to -31%. Qigong eliminated indicator drops for three subjects and reduced the indicator drop by 80% for the fourth subject. In the second series, each of seven subjects was examined by EAV three times in a blind protocol so that the operator did not know whether a subject had practiced qigong before the second or third examination. Qigong exercise changed the average EAV measured values in the range of -17 to -35% for four subjects and 4 to 15% for three subjects. Indicator drops appeared for three subjects, and they were eliminated for two subjects and reduced by 50% for the third subject. These preliminary studies indicate that EAV can monitor the effects of qigong on changes in the therapeutic balancing of the Channel-organ system. Suggestions are made of how EAV can provide basic information about qigong and its applications.

INTRODUCTION

Electroacupuncture According to Voll (EAV) is one of several electroacupuncture techniques for measuring local skin electrical conductance for diagnostic and therapeutic purpose(1). In general, EAV emphasizes preventive medicine by providing the medical practitioner with a diagnosis of a subject's health and its change at every stage of therapy. In the present study, EAV was used to monitor the effects of qigong practice on therapeutic balancing of subjects (2,3). The equipment is basically a D.C. resistance meter that measures relative changes in electrical conductance of the skin above an individual acupuncture point. Diagnosis depends on measuring the relative electrical conductance and its time dependence. The relative conductance is measured by a meter that is calibrated from 0 to 100. The "normal" value of 50 is expected when the acupuncture point and associated organ are in a physiologically balanced condition. Higher values (higher conductance) are associated with inflammation and lower values (lower conductance) with degenerative disturbances. The indicator drop is an important diagnostic criterion in EAV of a functionally disturbed organ. An indicator drop occurs during a measurement when the conductance of a given acupuncture point increases to an apparent maximum value and then decreases to a lower equilibrium value. In the present study, EAV was used to monitor the effects of qigong practice on therapeutic balancing of healthy subjects.

Two series of EAV measurements were made by the same operator and equipment. In both series, the subjects were asked to perform a qigong exercise of their choosing, usually sitting or standing meditation or moving qigong. The first series was performed during an international conference (4) at which qi and the human Channel system were among the main topics. About six months later, the second series was conducted in San Francisco with different subjects (5).

MATERIALS AND METHODS

The EAV instrument, a Computronix Acupro II system (Synergy Health Systems, 1223 Wilshire Boulevard #321, Santa Monica, CA 90403) was designed and operated by Douglas C. Leber. The instrumental test parameters were: 1.25 volts D.C. output voltage, 12.7 microamperes current output at full scale, and 95,000 ohms resistance at midscale of 50. A schematic of the electrical circuit is shown in Fig. 1. The instrument was calibrated from 0 to 100, and the value of 50 indicates that an organ associated with the acupuncture point was free of pathological problems. The tester pressed the brass probe electrode (0.13 cm diameter and connected to the positive side of the circuit) gently onto the acupuncture point as shown in Fig. 2. To minimized variations of skin conductance due to perspiration, tap water was applied to the skin in the region of the acupuncture point before the measurement. The cylindrical brass hand electrode (1-inch diameter and connected to the negative side of the circuit) was held in the subject's free hand, which was also moistened with water. Twenty four acupuncture points at the ends of the Channels of the fingers and toes of a subject were tested in about 5 minutes. The computer was programmed to provide information on the right and left side of the body for twenty organs and physiological functions of the body.

The subjects were 30 to 65 years of age. They were examined by EAV before and after practicing a qigong exercise for 10 to 15 minutes. The subjects went to a secluded spot to practice their own style of qigong, which was sitting or standing meditation or a form of moving qigong. The subjects were asked to balance their energies, and they were examined within 15 minutes after practicing qigong. One subject was examined by EAV before and after receiving qigong therapy from a qigong master and was also examined by EAV three more times within 22 hours.

RESULTS AND DISCUSSION

Series 1: In Series 1, four subjects were examined by EAV before and after qigong. Subjects A, B, and C, who had practiced qigong for more than 10 years, practiced their own style of qigong for 10 to 15 minutes. Subject D, who had studied qigong for about one year, was given a qigong balancing treatment for about 10 minutes by subject A, a qigong master who used a combination of acupressure massage and emitted qi to balance the Channels and organs of subject D. Subject A was not aware of that variations in the polarity of qi may exist in different parts of the his hands and of the body of the subject (6).

Examples of the EAV measurements before and after qigong are shown for subject D in Fig. 3 and for subject C in Fig. 4. The 20 pairs of horizontal bars represent the measured values at the acupuncture points associated with specific organs and functions of the body as described below the pair of bars. For a given pair of bars, the upper bar represents the value on the left side of the body and the lower bar that on the right side of the body. For example, LUCL and LUCR represents the lung Channel on the left and right side of the body, respectively. The scale for the measured values is shown at the bottom of each graph in the figure. A vertical line at the "normal" reading of 50 is a convenient reference for comparing changes that resulted from qigong practice. The two numerical values to the left of each bar are the actual measured values. The maximum value is on the left and the equilibrium value, which may be lower if there is an indicator drop, is on the right. The magnitude of the indicator drop is represented by a "white" bar and its numerical value is the difference between the maximum and the minimum measured values.

Visual inspection of Fig. 3 and Fig. 4 shows that qigong had decreased all measured values and had either eliminated indicator drops or reduced their numbers. However, qigong did not restore balance to some of the Channels and the corresponding organs. For example after the qigong balancing treatment for subject D, the measured values for the spleen, liver and gall bladder Channels on the right side of the body remained elevated with respect to the others at values of 66, 69 and 67, respectively. In the case of subject C, the measured values remained elevated for the stomach at values of 77 and 66 and for the urinary bladder on the right side of the body at a value of 68, and two of the nine indicator drops remained. These observations suggest that EAV can be used to monitor the effects of qigong therapy on individual organs of the body.

The results of the EAV measurements obtained before and after qigong for each of the four subjects are summarized in Table 1. For each subject, average measured values were calculated from the 40 values, 20 on the left and 20 on right side of the body (see Fig. 3). For each subject, a percentage change (change %) was calculated from the average measured values obtained before and after qigong. The results show that qigong had changed the average measured values of each practitioner in the range of -19 to -31%. The sum of all indicator drops (in meter divisions) for a given EAV examination were either eliminated or reduced markedly by qigong. The style of qigong of each subject and the duration of the practice are noted in a separate column in Table 1.

In Table 1, the plus and minus value that follows an average EAV value is an apparent "standard deviation" that indicates that individual Channels or body functions have measured values greater or less than the average value. This standard deviation would be zero if all the Channels and body functions were balanced and had the same measured AEV values. Individual Channels and body functions may not be balanced for many reasons, including chronic, acute, stress, and emotional factors. The magnitude of the apparent standard deviation may be regarded as a semiquantitative indicator of imbalance in the Channel system.

In a preliminary study to determine possible lasting effects of qigong, subject D was examined by EAV a total of four times over a period of 22 hours. The average measured value was 75.1±5.2 before qigong, 51.5±9.9 twenty minutes after qigong balancing, 55.7±10.4 four hours later after lunch, and 59.0±8.1 twenty-two hours later after breakfast. The increases in the standard deviations after the first 20 minutes were mainly due to large elevations in the measured EAV values of the Channels of the liver at a value of 85, spleen at 85, large intestines at 68, and kidney at 65. These changes are probably caused by food eaten at lunch.

Series 2. A blind protocol was used in Series 2 so that the EAV operator did not know when the seven subjects practiced qigong. This was accomplished by measuring the EAV of each subject three times, and in between measurements the subject left the examination room. After the initial EAV examination, the subjects were given a random selected card instructing them to practice qigong after the first or after the second EAV examination. Later the subjects informed the examiners when they practiced qigong. Of the seven subjects, E, F, G, H, I and K were experienced qigong practitioners, while subject J had studied qigong one year. Computer printouts similar to those in Fig. 3 and Fig. 4 were analyzed for each subject. The averages of the 40 measured values and the sum of indicator drops are tabulated in Table 2. The average measured values obtained just after qigong practice are indicated in boldface.

The results show that qigong practice changed the average measured readings of most subjects. The average measured values decreased for four subjects (E, F, G, H) in the range of -17% to -35%, increased for two subjects (J and K) in the range of 12% to 15%, and remained essentially unchanged for subject I. For subject I, a decrease in the average measured values was observed after qigong, that is, from the second to the third examination (82.7±2.7 to 61.2±11.8, respectively). This decrease may be due to a delay in the effects of qigong exercise. Indicator drops were present in the EAV of only three subjects before qigong, and these drops were either eliminated or decreased by qigong.

In series 2, the average measured values for four subjects (E, F, G, H) showed similar decreases in magnitude as those in series 1. However, the average measured values for the other three subjects (I,J, K) increased. This difference in response of individual qigong practitioners can be accounted for in many ways, such as by differences in qigong style, intention of the practitioner, and physical and emotional states. For example, we had asked the subjects to balance their energies, but subject J later told us that she had tried to increase her qi, thus perhaps explaining the increase of 12%.

Because of scheduling problems, we were unable to carry out control experiments. However, the reproducibilities of successive AEV measurements before or after qigong were satisfactory for five of the subjects. This can be seen in Table 2 by comparing the second and third measurements (after qigong) for subjects E, F, H, and J, and by comparing the first and second measurements (before qigong) for subject G. These successive readings changed less than 4%.

CONCLUSIONS

Our preliminary studies suggest that the EAV technique can monitor the effects of qigong on the therapeutic balancing of the Channel-organ system. EAV also should be valuable in other applications of qigong: 1) evaluating the effectiveness of different types of personal qigong exercises for balancing organs of the body, 2) measuring the receptivity of subjects to emitted qi, 3) evaluating the effects of different kinds of emitted qi, for example, positive and negative qi (8), 4) determining the duration of the balancing effects of qigong, 5) determining the relative healing power of qigong masters for certain medical problems, 6) comparing the relative effectiveness of energy balancing by qigong with other therapies, such as acumassage, acupuncture, healing with hands, Reiki, as well as conventional physical exercise, 7) investigating the therapeutic effectiveness of qi energy stored on materials such as paper (7), 8) investigating the therapeutic effectiveness of electrical, mechanical and laser stimulation(8), and 9) investigating the interaction between a qigong master and a subject (9).

ACKNOWLEDGEMENTS

The author acknowledges the cooperation of Douglas C. Leber of the Computronix Electro-Medical Systems, who designed and operated the EAV instrument. He also acknowledges support in arranging the experiments by Jim Jose of Synergy Health Systems, qigong master Gong Xin Xiong, Harry Kung, Linda Kung, and the qigong practitioners.

REFERENCES

1. Fred M. K. Lamb, Jr., M.D., Julia J. Tsuei, M.D., Zixian Zhao, M.D., M.S.: Bioenergetic Regulatory Measurement Instruments and Devices. Am. J. Acupuncture, Vol.16, No. 4: pp. 345-349, Oct.- Dec., 1988.
2. R. Voll: Twenty years of electro-acupuncture diagnosis in Germany: A progress report. Am. J. Acupuncture, Vol. 3, No. 1: pp. 7-17, 1975.
3. See collection of nine articles on Electroacupuncture According to Voll in the Am. J. Acupuncture, Special EAV issue, pp. 5-99, Second printing, 1989.
4. International Conference on Bioenergetic Medicine, Past, Present and Future, University of Hawaii, September 30, 1989.
5. Presented at the Second World Conference on Academic Exchange of Medical Qigong, Beijing, China, September 1993.
6. Omura,Y., Common factors contributing to intractable pain and medical problems with insufficient drug uptake in areas to be treated, and their pathogenesis and treatment: Part I. Combined use of medication with acupuncture, (+) Qigong energy-stored material, soft laser or electrical stimulation, Acupuncture & Electro-Thera. Res. Int. J., Vol. 17: pp. 107-148, 1992.
7. Omura, Y., Storing of Qigong energy in various materials and drugs (Qigongnization): its clinical application for treatment of pain, circulatory disturbance, bacterial or viral infections, heavy metal deposits, and related intractable medical problems by selectively enhancing circulation and drug uptake, Acupuncture & Electro-Thera. Res. Int. J., Vol. 15: pp. 137-157, 1990.

8. Omura, Y., Connections found between each Channel (heart, stomach, triple burner, etc.) & organ representation area of corresponding internal organs in each side of the cerebral cortex; release of common neurotransmitters and hormones unique to each Channel and corresponding acupuncture point & internal organ after acupuncture, electrical stimulation, mechanical stimulation (including Shiatsu), soft laser stimulation or Qigong, Acupuncture & Electro-Thera. Res. Int. J., Vol. 14: pp. 155-186, 1989.

9. Omura, Y., Unique changes found on the Qigong (chi gong) master's and patients's body during Qigong treatment; their relationship to certain Channels and acupuncture points and the re-creation of therapeutic Qigong states by children and adults, Acupuncture & Electro-Thera. Res. Int. J., Vol. 14: pp. 61-89, 1989.

Table 1. The effect of qigong exercise on the average measured EAV values and sum of indicator drops for each of 4 subjects before and after qigong (Series 1).

Each subject was tested before and after practicing qigong. Subject D was given qigong balancing therapy by subject A. The change (%) refers to the difference between the measured values before and after qigong. The indicator drop (ID) divisions are the sum of all such drops for a given subject. The style and duration of qigong practice is shown in the right-hand columns. Average measured values for before and after qigong were calculated from the average of 40 readings, 20 on the left side of the body and 20 on the right side (see Fig. 3).

Average measured values Sum ID Divisions Qigong
Before After Change Before After Time
SubjectQigongQigong(%)QigongQigongStyle(min.)
A69.0±3.351.4±4.4-26200Moving15
B72.0±3.953.3±7.1-26220Standing10
C67.0±6.654.6±5.7-1912928Sitting10
D75.1±5.251.5±9.9-31530Therapy10

Table 2. A blind study of the effect of qigong exercise on the average measured EAV values and sum of indicator drops for each of 7 subjects before and after qigong exercise (Series 2).

Each subject was tested 3 times. The bold faced numbers indicate measured values obtained just after qigong practice. Secret instructions informed the subjects to practice qigong either after the second or third EAV test. The change (%) refers to the difference in measured values before and after qigong practice. Indicator drop (ID) divisions are summed for a given subject. The style and duration of qigong practice are shown in the right-hand columns. The average measured values before or after qigong were calculated from the average of 40 measured values, 20 on the left side of the body and 20 on the right side (see Fig. 4).

Average Measured Values Sum ID Divisions Qigong
Change Before After Time
SubjectFirst testSecond testThird test (%)QigongQigongStyle(min.)
E61.3±15.439.6±11.238.6±9.7-35140Moving 15
F74.8±3.052.0±5.251.7±4.0-3000Sitting10
G76.1±2.6 79.1±4.258.1±8.6-2700Sitting10
H74.4±3.262.1±12.758.9±10.5-1700Sitting10
I79.2±4.182.7±2.761.2±11.8400Sitting10
J67.5±8.974.5±7.071.1±4.2122814Moving15
K67.2±11.5 53.9±13.562.1±14.71580Sitting10

APPENDIX 2

INTRODUCTION

This second Appendix includes research information on Medical Qigong therapy, collected by my good friend Richard H. Lee, Director of China Healthways Institute, San Clemente, California. Mr. Lee, a council board member of the World Academic Society of Medical Qigong, in Beijing, China, is a specialist in Western medical technology. He has worked extensively with the research department of the Beijing College of Traditional Chinese Medicine, studying the relationship of infrasonic waves emitted as Qi.

This research, focused on the Extension of Emitted Qi by various Medical Qigong Masters/ Doctors, was motivated by the pioneering research of Professor Lu Yan Fang at the National Institute of Electro-Acoustics in Beijing, China. Professor Lu developed prototype devices which simulated the frequency output of Medical Qigong doctors.

Mr. Lee has spent many years in China and abroad researching and collecting scientific data which led to the development of several Infratonic Medical Qigong Therapeutic Machines (in Chaotic resonances), which are currently being sold worldwide.

This second Appendix is a continuation of the research devoted to the scientific approach of analyzing and scrutinizing the claims made as to the actual effectiveness of Medical Qigong Therapy as a clinical modality. The result on the effectiveness of Medical Qigong Therapy as a solid and effective clinical modality being positive in conclusion.

The following information in this second appendix represents but a fraction of Mr. Lee's findings, however, I am confident that the reader will find the data fascinating.

Dr. Jerry Alan Johnson May 2000

SCIENTIFIC INVESTIGATIONS INTO CHINESE MEDICAL QIGONG THERAPY

Richard H. Lee
Director China Healthways Institute
San Clemente, California

INTRODUCTION

In the years since the first edition of "Scientific Investigations" was published, the debate over the existence of Qi, the legitimacy of healers, and the reality of PSI phenomena has heated up. While most readers embrace energetic concepts like Qi as a useful framework for understanding otherwise unexplainable phenomena, some continue to reject Qi. Even in China, while an estimated 100 million people practice Qigong, many professionals remain skeptical. Scientists who actually investigate Qigong generally conclude that the phenomena really do happen, but still try to maintain an arms-length distance from it. Many doctors are also reluctant to discuss Qi openly even though most of the general public is quick to identify Qi as feelings of abundant vitality, and the absence of Qi as fatigue.

There appear to be three principal barriers to serious consideration of Qi.

1. Inertia: "There is not enough data." "These studies have not been replicated." This seems to stem from an unwillingness to explore the studies that have been done. This book presents enough information to satisfy the open-minded investigator and the Qi explorer, but nothing was added in this edition to satisfy those who demand overwhelming evidence.
2. Hidden Bias: This is probably the biggest resistance that Qigong research faces. Many people simply refuse to consider the possibility of action at a distance, as exhibited in Qigong and PSI research. On the basis of this single assumption, all empirical data which supports such phenomena must be rejected as mistakes, experimenter bias, or falsification. Nothing was added to this edition to try to overcome this hidden bias.
3. Need for a framework: While the information presented in the first edition of this book answered the question, "Is the phenomenon of Qigong scientifically measurable?," It left readers without a framework other than faith by which they can understand these phenomena. Without a framework of understanding, many are reluctant to accept these strange experimental results.

Others are less concerned about physical substance and seek to understand Qi as an aspect of intention or consciousness. This more closely parallels the approach of most Qigong masters who, through years of training, gradually gain mastery over their own mind (through focus of attention), emotions (through calmness and releasing desires), and physical body (through discipline). Interestingly, this second viewpoint is similar to the view of many skeptics who insist that phenomena of Qi are simply "a product of the mind." Combining the idea that Qi is strongly influenced by the mind, with the research in this book which shows that Qi is measurable in the physical world, we get a picture of Qi as a mind/body connection which is appealing to most readers.

The Chinese do not view Qi as a scientific substance, but rather, as the essence of life, the bridge of consciousness between mind and body, and the "eternal now" in which all activity occurs. The physical body supports physical fields (Qi) which support vibrational frequency patterns (Qi) which allow communication (Qi) between some kind of abstract consciousness (mind) and a physical body which is our vehicle.

Richard H. Lee, January, 1999

QIGONG MASTERS AND ENERGETIC MEDICINE

For thousands of years, China has been a major source of the world's technology. We are surrounded by Chinese inventions such as paper, movable type, gun powder, and surgical anesthesia. And while the West has shot past China in many technological areas, there are still areas where China holds an edge. The Chinese investigations into the phenomenon of Qigong over the last ten years have, in some ways, placed Chinese researchers at the forefront of energetic medicine. Many findings of this research are presented on the following pages.

Qigong has been practiced for thousands of years in China. Yet it has never before had the broad acceptance that it has today. This is because the scientific investigations into the nature of Qigong show clearly that something is going on beyond what physics might predict. Measurable energy is emitted from the hands of those who have mastered these Qigong techniques. These masters can influence events from a distance with energy (or signal) emitted from their hands. Further, these energy emissions and Qigong training exercises have scientifically measurable, beneficial effects on health and healing.

Qigong masters have demonstrated many remarkable feats. Some can soften and bend steel. Some can light fluorescent tubes with their bare hands. Others can kill cancer cells in laboratory experiments, and many can, from another room, alter the brain wave patterns of test subjects including test animals.

The energy they emit from their hands is measurable with scientific equipment in the forms of infrared, magnetic, and acoustical energy. In one test, a strong pulse of static electricity was measured at a point 50 mm above the head of a Qigong master. In another, a strong magnetic field was measured just at the instant that a heavy iron bar hit a Qigong master's head. The bar bent as it would against Superman's head. He was unharmed with no sign of a lump or a bruise.

In 1979 and 1980, the leaders of China decided to put Qigong to the test. The Western scientific test. High technology government organizations across the country, many as prestigious as the National Atomic Energy Lab in Shanghai, and the Space Science and National Electro-Acoustics Institutes in Beijing, started testing Qigong masters and promptly discovered that the phenomenon of emitted Qi was measurable. Some of these laboratories went further and actually developed therapy equipment such as the Infratonic QGM which reproduces the infrasonic component of the emitted Qi. This research has led me into Qigong and taught me many things I didn't think I ever wanted to know.

Today's Western scientists know little about Qigong, but it is an important healing technology in the world today. It is the basis of acupuncture and traditional Chinese herbal medicine, yet it presents effective medical intervention when used alone. Americans have been radiating cancer with energy emitted from machines for the last few decades. The Chinese have been using energy emitted from the hands of Qigong doctors to treat cancer for thousands of years, finding that, in many cases, tumors can be shrunk without side effects.

These scientific experiments are repeatable. Scientific laboratories in America will reach the same conclusions as those in China. The evidence is persuasive that systematic Qigong training produces real benefits. And advanced practitioners can emit energy from their hands that is measurable with scientific equipment. Further, therapeutic equipment has successfully reproduced many of the beneficial effects observed. The Chinese scientific investigations into medical Qigong are indeed at the heart of a medical revolution that is improving the science of health and healing around the world.

THE SCIENCE OF CHINA

For thousands of years China has been a center for the advancement of civilization. The science of China has produced a wide variety of technologies. Gunpowder is the central ingredient to modem warfare. Paper and printing with movable type have for centuries been at the center of education, communication, and information storage. Despite the growth of computers, this system will continue to be the world's dominant information storage method.

Many centuries ago, a physician named Hua Tou invented surgical anesthesia and performed many kinds of surgeries like appendectomies and other of today's popular surgeries. While the Chinese contributed to what has today become modern medicine, they also created "Chinese medicine," a system of medicine based on the Chinese science of "Qi" or vital energy. It is this system from which acupuncture and Chinese herbal medicine have evolved. And at the heart of this system lie the science and practice of Qigong.

QIGONG'S LONG HISTORY

Through the millennia Qigong has been an influential force in the molding of China. Until recently it has been hidden from public view. Like many other disciplines, the art of Qigong has been taught from father to son or from master to disciple only after a strict oath of secrecy has been given. Knowing a science that others don't know gives a practitioner a decided advantage in earning a living. Thus, until recently, very few books have been written about Qigong, particularly about the more advanced techniques of study. The exception is the secret books that have been maintained in religious monasteries, being copied and recopied by studious monks through the centuries.

Even this activity in monasteries came to an abrupt halt with the Communist government in China when Mao put an end to most religious activities. Through the Cultural Revolution, monasteries and temples were shut down. Many of the ancient texts and artifacts were destroyed. Others were stored in government libraries. Many of the Qigong masters were killed. A few went into hiding or escaped from the country.

At the end of the Cultural Revolution, Deng Xiao Peng set about rebuilding much of the cultural heritage of China, allowing limited freedom of religious expression. He also promoted the rediscovery and preservation of traditional Chinese science and medicine, which included Qigong.

It is interesting to note that much of the willingness to share the secret teachings of Qigong came from outside countries. Masters who had emigrated to the West had started teaching the techniques openly without the strict code of silence. And some of the secret Qigong texts that had been taken from the temples and monasteries were published for people around the world to read. Qigong was no longer a secret so there was no longer any value to secrecy in China and Qigong masters became much more willing to teach what they knew.

The most influential factor in the Qigong revival was, perhaps, the new policy on the part of the China government to allow freedom of religion, particularly religion that had a basis in China's cultural heritage. Thus, Taoists and Buddhists who had preserved its teachings in secret for centuries were free to share their knowledge. This allowed Qigong practitioners to demonstrate their skills freely in public without fear of being condemned to death as witches or evil beings. Paranormal skills were again publicly accepted as a product of China's cultural heritage.

The big boost to Qigong came in the early 1980's when modern science turned to study the tradition of Qigong. The leaders of China reasoned that, if Qigong is to be accepted, it must pass the rigorous screening of scientific scrutiny. Studies were funded in the prestigious scientific institutes of China to determine whether the special skills of the Qigong masters could be scientifically verified. The results were most impressive. Many forms of energy were measured coming from the hands or bodies of Qigong practitioners in reproducible scientific experiments.

Scientific investigations into the effectiveness of medical Qigong quickly showed that, for many illnesses, medical Qigong was unsurpassed by any other medical system in the world. The discovery that energy emitted from the hands of Qigong doctors under controlled laboratory conditions was highly effective in killing cancer cells, gained extensive press coverage around the world.

Qigong science is a very new field with valuable technologies just beginning to surface. But given the fundamental nature of these discoveries, Qigong science could easily dominate the world of the 21st century.

SIMULATION OF EMITTED QI

Simulation of emitted Qi has been the goal of much of the Chinese Qigong research that has been conducted since about 1980. Many devices have been developed and tested, some of which have considerable potential. The following is a description of the most common forms:

Infrared devices have been developed some of which, through rare-earth emission surfaces, emit characteristic frequency bands, while others modulate the infrared output according to recorded patterns emitted by Qigong masters. These devices tend to have the most effect where coldness in the body is effectively treated with heat.

Electrical devices are often adapted to deliver current through the skin at frequency patterns that follow patterns produced by Qigong masters. These units have been used principally by acupuncturists.

Acoustical devices, which reproduce the subsonic output of the emitted Qi in the form of deep acting massage action, have been shown to be effective in pain management, circulation of Qi, blood, and endocrine fluids, and relaxation. The Infratonic QGM was used in several of the laboratory tests described in this book to simulate the emitted Qi in experiments. Because this device offers many insights into the nature and function of emitted Qi, and because many of the readers of this book use the Infratonic QGM at home or in their medical practices, much of the research around this device is included in this book.

SCIENTIFIC PLAUSIBILITY OF QI, THE HUMAN ENERGY FIELD

Some readers might tend to reject the findings of this study based on their assumption that the existence of human energy fields and the altering of a person's energy field are outside the laws of physics, and therefore impossible. It is therefore appropriate here to briefly review some of the findings of researchers mostly within the last decade to show that these energy fields are quite measurable with existing equipment and that there is strong evidence that known biophysical mechanisms can account for the observed phenomena.

First, the human body clearly has a vibrating electrical field, measurable for decades through such techniques as EMG, EEG, and EKG. How is this possible? The classical view of the human cell as a bag of water with a nucleus has been, for the most part, rejected in favor of models in which a highly structured collection of filaments, tubes and fibers connect the cell's nucleus not only to the surface of the cell but beyond, in a matrix which connects and interrelates the cells of the entire body (Oschman 1994) and (Pinta and Coffee 1991). Second, helical molecules within the body, such as DNA, keratin, collagen, actin, and myosin have piezoelectric qualities. As they elongate and contract they produce voltages. Also, when voltages are applied, they expand or contract producing mechanical activity and acoustical energy.

While it is well established that electrical signals are produced and received by the body, readers are more likely to question whether a magnetic field interpenetrates and surrounds the body. In 1963 researchers at Syracuse University reported the first measurements of the magnetic field of the human heart, just a millionth the strength of the earth's magnetic field, yet highly coherent and measurable throughout the body and beyond. Then in 1971, the SQUID, a highly sensitive superconducting magnetometer, was developed which was used to measure the magnetic field of the brain, 100 times weaker than that of the heart (Cohen 1972).

Many researchers now believe that the magnetic field of the heart, which extends throughout the body plays a role in regulating biological functions. Then, in 1991, Seto and colleagues measured the magnetic field between the hands of healers to be as strong as 1 milligauss. This is less than 1% the strength of Earth's magnetic field, but 1000 times as strong as the magnetic field produced by the heart. If the heart regulates biological functions with it's weak field, it is certainly feasible that a magnetic field from the hands of healers, measured to be many times stronger than the heart field might influence biological functions.

But, how can a magnetic field be produced from the hands? The Josephson junction, a superconducting semiconductor which is the basis of the SQUID's extremely high sensitivity to weak magnetic fields, has been predicted in quantum physics calculations to be present in living tissue. This was verified just ten years ago. (Del Giudice et al 1989). DNA is one of the biochemical structures shown to act as a superconducting magnetic field detector, able, through the Hall effect, not only to detect very weak magnetic signals, but also to produce magnetic signals (McCraty 1993). Thus DNA may act as a signal amplifier, releasing its energy coherently into passing magnetic waves, much as the ruby rod of a laser releases its light energy coherently with a passing photon to produce a laser beam. This provides a mechanism by which healers might produce the observed relatively strong biomagnetic signals, and by which a test subject or patient might be influenced by such signals (Oschman 1993).

Thus, modern advances in science now provide mechanisms by which it is possible for healers to produce magnetic and electric signals of significant magnitude to influence patients. For those who seek a more indepth overview of this rapidly advancing field of knowledge, "The Scientific Basis of Energy Medicine" by James L. Oschman, 1998, is highly recommended.

REFERENCES

Cohen D 1972 Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer. Science 175: 664-666

Del Guidice ES, Doglia S, Milani M, Smith JM, Vitiello G 1989 Magnetic flux quantisation and Josephson behaviour in living systems, Physica Scripta 40: 786-791

Dumitrescu I, Kenyon JN Electrographic imaging in medicine and biology, Nevil Spearman Limited, Suffolk England, 1983

Konikiewicz L, Griff L Bioelectrography a new method for detecting cancer and monitoring body physiology, Leonard Associates Press, Harrisburg, 1984

Lee R Bioelectric vitality- exploring the science of human energy, China Healthways Institute, 1997

Liu G, Wan P, Peng X, Zhong X The influence of the emitted Qi on the auditory brainstem evoked responses (ABER) and auditory middle latency evoked responses (MLR) in cats, Proceedings of the First World Conference for Academic Exchange of Medical Qigong, Beijing College of Traditional Chinese Medicine, 1988.

McCraty R, Atkinson M, Tiller WA 1993. New electrophysiological correlates associated with intentional heart focus, Subtle Energies 4: 251-268

Niu Xin, Liu G, Yu Z Measurement and Analysis of the Infrasonic Waves from the Emitted Qi, Proceedings of the First World Conference for Academic Exchange of Medical Qigong, Beijing College of Traditional Chinese Medicine, 1988.

Oschman JL 1993 A biophsysical basis for acupuncture. Proceedings of the First Symposium of the Committee for Acupuncture Research, CAR, Boston.

Oschman JL, Oschman NH 1994, Somatic recall, parts I-II Soft tissue memory / Soft tissue holography, Massage Therapy Journal, American Massage Therapy Association, Lake Worth, FL 34: 36-45, 66-67, 101-116

Pinta KJ, Coffee DS 1991 Cellular harmonic information transfer through a tissue tensegrity-matrix system. Medical Hypotheses 34:88-95

Seto A, Kusaka C, Nakacato S et al 1992 Detection of extraordinarily large bio-magnetic field strength from human hand. Acupuncture and Electro-Therapeutics Research International Journal 17:75-94

THE FUTURE OF QIGONG

Chinese research has found that Qigong training increases balance, quickness and endurance in athletes, and that fighter pilots trained in Qigong have quicker reactions and greater tolerance for the G-forces of sudden turns. Qigong has provided revolutionary treatments for hypertension and cancer, and has the potential for much more.

If, as the Chinese claim, Qi is the fundamental substance of the universe making up both matter and energy, and if we can through our bodies and our awareness, control and direct this Qi, not only within our own bodies, but as has been seen, beyond the limits of our bodies, the fantasy of "the Force" in Star Wars comes within the grasp of reality. Psychic phenomena like telekinesis and mental telepathy become minor extensions of the human manipulation of Qi. The Western ideas of positive thinking take on new meaning:

"The mind is the builder" and "Your thoughts create your reality."

The Chinese science of Qigong gives us proven tools to strengthen our ability to structure the world around us by training our minds and bodies to feel and understand Qi.

Remember that Western science and technology gave us electric motors which, when coupled with man-made equipment, can do the work of a thousand men. This came with our understanding of the principles of electricity. Emitted Qi simulation devices such as the Infratonic QGM provide this work-saving feature of machines. As we learn more about Qi, how it functions, and how it is directed by the human body, we will learn how to construct more and more powerful machines. We will see Qigong devices not only in hospitals but also in agriculture, in building air conditioning systems, in chemical processing plants and in space flight. The scientific Qigong research that started just a decade ago will grow and grow, and perhaps dominate our world of the 21st century.

However, the real emphasis of the Qigong movement and the intent of most Qigong teachers is not just to heal patients with illness, but to expand human potential. The research presented in this book shows that Qigong exercises help people to develop, not only the ability to heal themselves and others, but to do things which scientists considered impossible just a few years ago. Through a study of Qigong, there is strong evidence that we can become more than we ever thought possible.

INFRASONIC QIGONG RESEARCH

The discovery that specific bands of very low frequency sound influence healing and vitality was an accident. A "friend of the family" walked in to the National Electroacoustics Laboratory in Beijing and asked Prof. Lu Yan Fang, an acoustics scientist, if she would test his hand to see if any sound was coming out of it when he emitted Qi. She thought he was crazy but went along because she didn't want to dishonor her family by turning out a friend.

Prof. Lu's discovery that he did indeed emit a very strong signal in the infrasonic range was quite a surprise. This signal was 100 times stronger than that of a healthy, normal person. Elderly and sick people had even lower readings.

To verify that this infrasonic sound had a physiological effect, she built a device to simulate this output and tested it in hospitals. It worked to relieve pain and assist recovery in a wide variety of patients.

Prof. Lu's patent for the Infratonic QGM was the first ever filed in China and her device was awarded three awards for outstanding contribution to science and medicine from Chinese national science organizations.

It is clear from the Chinese infrasonic research presented in this chapter that infrasonic sound is an important component of emitted Qi.

INFRASONIC SIMULATION OF EMITTED QI FROM QIGONG MASTERS

by Professor Liu Guo-long, MD, PhD
Beijing College of Traditional Chinese Medicine
with Richard Lee

SUMMAY OF RESEARCH

Emitted Qi from Qigong masters clearly has a strong effect on the central nervous system, not only in humans but also in animal subjects. Qigong meditation and the Infratonic QGM have similar effects. These findings are summarized below:

Repetition of these tests with animals confirmed that these changes are physiological and not psychological phenomena by eliminating placebo effects.

DOUBTS ABOUT EMITTED QI

At the beginning of my research with Qigong, I was confident that a study of neurophysiology would prove that the "Wai-Qi" emitted by Qigong masters was nothing more than a psychological factor induced by the waving of hands and hypnotic suggestion.

Qigong is a system of physical and mental exercises which has been practiced in China for thousands of years. Up until the 1980's, most Chinese considered Qigong masters (those who have mastered this Qigong training) to be mythical story book characters with super-human powers. However, as the Chinese government began to support scientific research into Qigong, the few remaining true Qigong masters began to surface, amazing researchers with feats like killing bacteria in test tubes and causing previously paralyzed people to get up and walk.

I thought that these were all tricks of one sort or another, perhaps magicians' sleight-of-hand, hypnosis, or optical illusion. When I got the chance to work with people claiming to possess these abilities, I was fascinated. This was my chance to test with scientific equipment, whether there was any truth behind the excitement about Qigong masters. I was confident that I could disprove the myths.

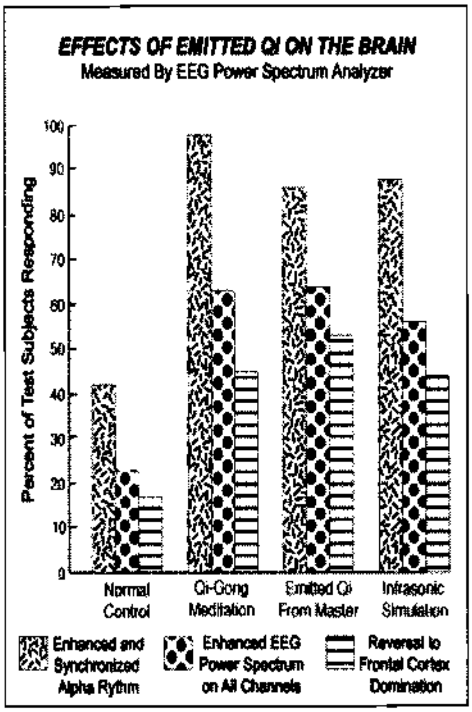

Emitted Qi has a pronounced and repeatable effect on EEG. It enhances frontal and occipital EEG power spectra, and often enhances the frontal so much that frontal becomes the dominant EEG activity whereas occipital dominance is more common. Emitted Qi also enhances and synchronizes the Alpha.

BACKGROUND

I was a specialist in Western medical technology assigned to the research department of the Beijing College of Traditional Chinese Medicine. The project was sponsored by the China Government Department of Education and the Department of Natural Sciences. My research group was assigned to study the relationship of infrasonic waves to emitted Qi. These research grants were motivated by the pioneering research of Prof. Lu Yan Fang at the National Institute of Electro-Acoustics in Beijing, who developed prototype devices to simulate the output of Qigong doctors.

Research has confirmed that humans have a very high degree of acoustic activity in the subsonic range below 20 Hertz (infrasonic), similar to the alpha rhythm of EEG. Also, people with chronic illnesses were found to have a much lower level of infrasonic activity, while Qigong masters had a much higher level of infrasonic output when they were emitting Qi.

The implication was that infrasonic sound might be related to human functioning, and further, that it might be in some way involved in the mechanism of brain functioning. We searched the available scientific literature for research papers for relationships between sound waves and brain waves but found none that had been done.

Extensive clinical research, based on infrasonic Qigong simulation, showed it to be effective for a wide variety of hospital problems. These results are what motivated the National Departments of Education and Natural Sciences to look further into the importance of infrasonic sound and what led to the research I am about to describe:

THE BRAIN AS A DETECTOR OF EMITTED QI

I had learned in my many years of research with the Electro-Encephalograph (EEG) that the human brain responds to even the most subtle of stimuli to the body, so I reasoned that, if there were really any scientific basis to emitted Qi, it would show up in the brain waves of test subjects who were placed in the path of these emissions. I expected to see no difference between the resting states and the Qi emission states.

What we saw was extraordinary. Within a few seconds after the Qigong master began to emit Qi, the subject's EEG would begin to shift. The EEG power spectrum was enhanced on all channels while the most pronounced increase was in the frontal lobe. Also, there was an enhancement and synchronization of the Alpha Rhythm in all channels. When the Qigong master stopped emitting Qi the EEG would gradually shift back toward the baseline readings.

To determine whether infrasonic energy was a significant part of the emitted Qi, we used the infrasonic Qigong prototypes in the same experiment. It was located 18 inches away, directly behind the back of the head of the test subject. The EEG electrodes were attached as before. The simulator was activated for short periods of time and the results recorded. We found that the effects on the receiver's EEG were quite similar to those of the emitted Qi.

Our further research involved monitoring the various sensory-cortical evoked potentials during Qigong meditation, emitted Qi, and infrasonic Qigong simulation. We again found very similar results from all three stimuli. We found that a large portion of the cerebral cortex was inhibited while other somatosensory cortex were excited. One of the significant findings of this study is that the inhibition of the cerebral cortex during Qigong meditation is clearly different from the excitation of the cerebral cortex that is measurable during sleep.

Through Acoustical Brainstem Evoked Response (ABER) it was found that the brainstem structures from the medulla to the hypothalamus were significantly facilitated. The brainstem plays an important role in regulating the functions of the inner organs, motor function, and emotion.

The implications of these studies were startling. Qigong masters can, without touch, voice, eye contact or any other traditional communication means, induce a clear, strong, and highly measurable change in a subject's brain functioning. A synchronization of alpha rhythm indicates deep relaxation, and is closely associated with accelerated healing. Enhanced power spectrum in the frontal lobe is especially significant because the association cortex of the frontal lobe is concerned with higher motor action, higher sensory function, emotional and motivational aspects of behavior, and integration of autonomic function. Facilitation of the brain stem, with its regulation of internal organs, may be a mechanism by which physical healing is induced or accelerated.

Despite these highly significant changes in EEG and evoked potentials, the subject had felt nothing and had no idea of the profound changes taking place within them.

The findings of these studies are solid evidence that a Qigong master can induce real physiological changes in a subject from several feet away, and further, may help to explain the high rate of recovery from chronic degenerative diseases in groups of hospital patients under the care of Qigong masters. These studies also show that the infrasonic Qigong simulator can induce similar changes in brain function and that, through Qigong meditation, a Qigong master can induce these same changes in his own brain.

SCIENTIFIC CONTROLS

There is much disagreement on how emitted Qi affects the brain. Many doctors insist that brain changes are psychologically induced, and that verbal suggestion, impressive hand motion, and expectation of the subject account for the observed phenomena.

To test this, we had several Qigong masters and people pretending to be Qigong masters treat the test subjects. The subjects were told that all were Qigong doctors, and all moved their hands in similar ways. We saw no significant changes in brain wave patterns with the fake Qigong masters, but when the real doctors emitted their Qi, or when we used the infrasonic Qigong simulator, we repeatedly got the highly significant changes.

Even this did not satisfy many of the doctors who reviewed our work, so we repeated the study with animals. We monitored EEG in awake rabbits and ABER in anesthetized cats as Qigong masters emitted Qi toward them. Even though there was no voice or eye contact between the Qigong masters and the animals, and the masters emitted Qi from several feet away, we saw shifts in EEG and ABER (Acoustical Brainstem Evoked Response) similar to those observed in the human subjects. This is a highly convincing result because all kinds of placebo effects are eliminated, yet modification of brain function at a distance remains.

INFRASONIC VS. ELECTROMAGNETIC INTERFERENCE

Extremely low frequency (ELF) electromagnetic signals can affect EEG. There is growing concern about the low frequency radiation produced by 60 Hz electrical power lines located close to people's homes and schools because of apparent disruption of brain and cellular function. ELF signals will cause the principal EEG power spectrum to show a spike at the frequency of the ELF signal. This is because the EEG is easily entrained by ELF signals. When the signal is discontinued, the EEG abruptly returns to normal. Electrical power lines operate at 60 Hz, which corresponds to EEG in the high Beta range, associated with mental anxiety and confusion. The entrainment of EEG to 60 Hz around power transmission lines may be why researchers are finding that people who live close to these power lines show a higher incidence of brain tumors.

The results using the infrasonic Qigong simulator were quite different from ELF signal entrainment. The shifts in EEG were gradual rather than abrupt, and while dominant EEG frequency did drift toward the dominant peak infrasonic frequency, it was a broad spectrum of EEG activity rather than a spike. The enhanced power spectrum continued after simulation was stopped, gradually decreasing and returning to the pretest state. From these observations it is clear that the effects of infrasonic are quite different from the entrainment of ELF electromagnetic signals.

SUMMARY

When I started this project in 1976, I had serious doubts about Qigong masters and emitted Qi. Now I am convinced that emitted Qi is very real and that Qigong is a very valuable art. With the support of the China government, an estimated 50 million Chinese are practicing Qigong daily, and many Western hospitals have added Qigong departments for patients with chronic and degenerative diseases. In addition, the China government has funded extensive scientific research into the nature of emitted Qi with the goal of advancing science and medical technology.

The study of emitted Qi and infrasonic sound, as they relate to human health and functioning, is a broad and exciting field. The use of Qigong in treating chronic degenerative diseases such as cancer and hypertension in China has proven very effective. It has been employed to accelerate healing for thousands of patients with a wide variety of diseases in Chinese hospitals. I am confident that emitted Qi, Qigong meditation, and the infrasonic Qigong simulator will play an increasing role in health care around the world.

SECONDARY ACOUSTIC BIOLOGICAL RESPONSE

(The following is from a technical paper presented in Beijing, China by Professor Lu Yan Fang, the inventor of the Infratonic Qigong Machine.)

"People often talk about sound. Common sense suggests that sound is what the ear can hear, the frequencies between 60 Hz. and 20,000 Hz.

Secondary sound is that sound that is outside of the range of the human ear, specifically low frequency sound below 60 Hz. Secondary sounds travel long distance with little attenuation.

In the 20th century, in the 1930's, this secondary sound was discovered, but it wasn't until the 1960's that it started drawing much attention. Since then the study, measurement, and analysis of the physical features of secondary sound have included:

1. Physical Features
2. Secondary Sound Measurement
3. Technology of Analyzing, and
4. Protecting the Body Against Strong Secondary Signals

Initial secondary sound biological response studies emphasized protecting against high level secondary sound which can damage the human body. Laboratory studies showed that strong secondary sounds beyond 150 dB are harmful to biological bodies. The American Environmental Protection Agency sets standards for secondary sound, concluding that secondary sounds below 130 dB are not harmful.

To this date there had been almost no studies investigating how secondary sounds affect biological processes. And never before has there been any thought of testing of the human body as a secondary sound source. Since 1983, we have worked with more than 70 Qigong masters, and have discovered for the first time that the human body actually emits a secondary sound signal. The frequency distribution peak of the signal is located in the secondary sound region of the sound spectrum, at about 10 Hz.

Through more than 100 repeated lab experiments we found almost identical results of frequency distribution. According to the analysis of test results in 1984, the first infrasonic Qigong prototypes were constructed. These machines used electroacoustic technology to simulate the frequency distribution of the Qigong master's low frequency signal, but stronger. This simulation of the human body life signal emission is then fed back to the human body. In the two years that followed, many clinical and animal tests proved that weak secondary sounds benefit the human body. The discovery that infrasonic sound is a basic substance of Qigong emission opens Qigong science's big door, opening an exciting new field for medical science.

Through hospital and clinical use, 1,134 patients were treated by the Infratonic QGM. Animal studies were also performed. Clinical use and animal use prove that this kind of secondary sound signal can strengthen many functions of the human body. The Qigong machine can have significant treatment value to the human body and works by reproducing the acoustic waves produced by the human body. It can promote circulation, regulate Qi and blood, open channels, and relieve pain.

The Qigong master is not alone in this ability to emit the secondary acoustical signal. Most everybody generates a similar signal. Young people with good health can emit the signal to a degree similar to that of a Qigong master. People with a weak body and elderly people have similar signals but much lower in power. People having Qigong training have a signal that is strong all the time, but especially strong during Qi emission. However, in normal people, as age increases or health condition worsens, the signal decreases. This explains why the human body is only sometimes a source of secondary sound. In the meantime, the body is also a receptor of secondary acoustical signals, and can be benefited by receiving the weak secondary acoustic signals.

Very early in the development of traditional Chinese medicine, this effect had been shown to exist. Anmo, and tuina, Chinese therapeutic massage techniques, give the human body this kind of low frequency stimulation. This stimulation through acupuncture points follows the cell as the basic pathway, and then goes to the meridian, and reaches the problem area. Then the problem can improve. Just like an old Chinese saying: "If you can get enough energy into the problem area, you can use that energy to treat the problem."

The human body's secondary acoustical biological effect is a scientific reality. The human body produces, delivers, emits, receives and controls this low frequency energy. This weak secondary acoustic effect has beneficial biological effect on the human body. It and its relationship with traditional Chinese medicine are of great value to the medical world. It is hoped that all scientists do research in this area.

MEASUREMENT AND ANALYSIS OF INFRASONIC WAVES FROM EMITTED QI

Niu Xin, Liu Guolong and Yu Zhiming
Beijing College of Traditional Chinese Medicine, Beijing, China

The theory of traditional Chinese medicine suggests that Qi is one of the fundamental substances in human bodies. Modem scientific research of the essence of the emitted Qi has yielded some positive results. In order to find out the relation between the emitted Qi and infrasonic waves, to explore the mechanism of the emitted Qi, to find out how a person generates and receives the emitted Qi, and to provide a quantitative physical scale for indicating the strength of the emitted Qi for experimental studies of the effect of it, we measured and analyzed the infrasonic waves from the Qi emitted by Qigong masters.

The test, which was done by an infrasonic testing system made in Denmark, was conducted in a noise-proof room in the Institute of Sound and Electronics under the Ministry of Electronic Industry. The background noise in the room was lower than 30 dB (decibel). The microphone was hung in the air over Laogong (P-8, acupuncture designation for a point in the center of the palm), the distance being 2 cm with no contact.

The experiment tested the release of emitted Qi at Laogong (P-8) in the palm, Mingmen (Du-4), Baihui (Du-20), Dantian, and Jianzhi. Special attention was paid to the test of Laogong (P-8).

1. The peak frequency of the infrasonic waves from the tested Qigong and non-Qigong masters ranged from 8 to 12.5 Hz. In one case the frequency reached 16 Hz. In another two, it reached 6 Hz.
2. The infrasonic waves from the Qigong masters ranged from 45 to 76 dB and those from the non-masters, 45-50 dB. Comparison of the intensity of infrasonic waves during the Qigong state and the non-Qigong state before and after the emission of Qi showed a statistically significant increase ($P<0.01$).* The increase of wave intensity of the Qigong masters compared with that of the non-Qigong masters was also obviously significant ($P<0.01$).* **The energy of the Qigong masters was over 100 times higher than that of ordinary persons.**
3. The masters who had practiced Qigong for many years and often emitted Qi to treat patients had a higher intensity of infrasonic waves, reaching over 70 dB. Those who started to practice Qigong a short time before and mainly practiced *Nei Yang Gong* had a lower intensity of the infrasonic waves (lower than 60 dB.)

Infrasonic and ultrasonic waves are all sound waves which cannot be detected by human ears. The frequency of infrasonic waves is below 20 Hz. Many natural phenomena and artificial actions may generate infrasonic waves. Human bodies may act as a source or a receptor of infrasonic waves giving rise to a biological effect. The infrasonic information we acquire a from the measurement of the Qi makes it possible to study the effect of the emitted Qi.

The Valsalva state in which a Qigong master emits his Qi, is the breath regulating state of Qigong and also the state of emitted Qi in the breath holding exercise. Non-Qigong masters who simulate the Valsalva state also send out more intense infrasonic waves. It shows that every person has infrasonic characteristics. A long period of practicing Qigong helps increase the radiative intensity of infrasonic waves. Entering the Valsalva state helps in the emission of Qi. The Qi emitted by masters who adopted Song Jing Gong (relaxed Quiescent pattern) had more intense infrasonic waves (reaching 72 dB). Thus, the mechanism of the emitted Qi released by different exercise patterns is different.

We have found by a series of tests that very able Qigong masters can keep the energy of infrasonic waves at a relatively high level (over 70 dB). So tests on infrasonic waves can be used to screen Qigong masters.

The human body can generate and emit infrasonic waves. As far as acoustics is concerned, the most suitable resonant frequency of human tissues is within the range of infrasonic waves. It shows that the human body easily receives infrasonic waves. Infrasonic waves are a strong, effective part of the emitted Qi because of their quick, long distance transmission, strong penetration, and non-decreasing vibration. It is possible that infrasonic waves themselves transmit the messages between the Qigong masters and the subjects, or serve as a carrier.

[(P<0.01) means that the probability that these test results would occur by chance is less than 1%. This measure of "statistical significance" tells us that the sample size was large enough and causative factors strong enough that the results of this study are reliable.]*

NEURO SCIENCE—NERVE IMPINGEMENT AND EMITTED QI

China Healthways Newsletter's Interview with Liu Guo Long, MD, PhD
Beijing College of Traditional Chinese Medicine
by Richard H. Lee in Palos Verdes, CA, August, 1991

(This interview came about as a result of requests by readers of China Healthways Newsletter for an explanation of neuroscience as it relates to vital energy and healing.)

Dr. Liu provides a fresh view into the role of vital energy in the nervous system and how vertebral nerve impingement affects Qi.]

China Healthways Newsletter: How has your research changed your views on Qi?

Dr. Liu: I started my research with the intention of proving that there was no physiological basis for Qigong, that it was nothing more than a circus trick. As my research progressed, I became more and more convinced that emitted Qi is both measurable and effective for healing. This was not before I eliminated all possible psychological effects by replacing human subjects in my experiments with laboratory animals and got similar results. In my view, emitted Qi is quite real because it is easily measurable using the human brain as a detector.

CHN: How does Qigong affect the function of the nervous system?

Dr. Liu: This is a very complex question because emitted Qi affects different organs quite differently. First, it clearly increases Electro-Encephalograph (EEG) activity throughout the brain and nervous system, principally in the frontal cortex and in the range of alpha, 8 to 14 Hz. This greater mental activity in the alpha range is associated with clear relaxed state of deep thought, which many agree is effective to relieve tensions and clear the mind. The frontal cortex of the cerebrum is at the top of the hierarchy. It is the center of awareness within the body, directing the cerebellum and brain stem. Increased EEG activity in the frontal cortex would indicate expanded awareness.

Secondly, emitted Qi inhibits the sensitivity of the cerebrum to outside stimulation while facilitating the sensitivity of the hypothalamus and medulla of the brainstem. The cerebrum performs high level conscious activities. When inhibited it receives less visual, auditory, and sensory input, and presumably sends fewer and clearer signals to the brain stem, which, in turn, is responsible for automatic functions such as respiration, vasodilation, and the complex functioning of internal organs. Because the brainstem is facilitated, it is more able to "tune-in" to the needs and illnesses of the body and to activate the needed repairs to restore health. I believe that this enhanced awareness of internal stimuli in the brain stem, combined with reduced sensitivity to outside stimuli, is the primary product of emitted Qi that brings about accelerated healing throughout the body.

CHN: Then, would you say that emitted Qi directed only to the central nervous system can enhance healing in all parts of the body?

Dr. Liu: That's right.

CHN: How can Qi applied to the local injury accelerate healing?

Dr. Liu: Neurons in all parts of the body are in constant oscillation, producing measurable output in the range of 8 to 14 Hz, the frequency of the alpha pattern. Neurons can both emit and receive energy in this frequency band. Therefore nerves throughout the body can be facilitated by the emitted Qi. The application of emitted Qi to the site of an injury facilitates the signals to the brain stem. Stronger and more accurate signals to the brain means a more effective healing response directed to the area of injury, including greater vasodilation, increased lymph flow, and an increased supply of nutrients for cell regeneration.

CHN: Is there any direct effect of emitted Qi on the injured tissue.

Dr. Liu: Yes, of course. Qi is the vital energy that drives the regeneration of tissue. Supplementing the available Qi can directly increase the metabolic activity of the injured tissue just as it increases the EEG activity of the brain.

CHN: How can emitted Qi strengthen the immune system against bacteria and viruses?

Dr. Liu: The mechanisms are the same. A facilitated brainstem and nervous system are able to identify the invader more quickly and accurately and to direct the immune system response more precisely to the area of infection. And, as before, the added supply of Qi fuels a more furious response from the immune system. Chinese herbs such as Huang Qi (Astragalus) are also highly effective at enhancing Qi to accelerate the immune response.

CHN: Can mis-position of spinal vertebrae disrupt this healing process?

Dr. Liu: Most definitely. There are basically three types of nerves, the sensory, the motor, and autonomic. Vertebral pressure on any type of nerve cell can inhibit sensitivity or create false signals. If sensory nerves are affected, important information may never reach the brain, or may arrive confused, misdirecting the healing effort. Affected motor nerves can result in spasm or dropsy, and most importantly, impingement of autonomic nerves affects the function of internal organs and if it continues over a long period, can result in serious degeneration or dysfunction.

Often, impingement of motor nerves can perpetuate the problem. As muscles spasm from false motor nerve signals, they may tend to pull the vertebrae further out of alignment, increasing the impingement. In this case, the problem tends to get worse with time as greater impingement further increases the muscle spasm.

The best way to deal with this sort of nerve impingement is through adjusting the position of spinal vertebrae. Often just the vertebral adjustment is sufficient to release pressure on the motor nerve and release the muscle spasm, allowing complete recovery. However, in many cases the impinged nerve is slow to recover, showing some immediate release of spasm but not complete release. The remaining muscle spasm is sufficient to pull the vertebra out of adjustment again and re-establish the motor nerve impingement.

In these cases of muscle spasm aggravating the nerve impingement, better results are achieved if a nerve facilitation technique is used in conjunction with the adjustment. Emitted Qi or the Infratonic QGM applied a few minutes before or after the adjustment facilitates the nerve to provide greater muscle relaxation and quicker regeneration of the nerve. In upper back disturbances the Chinese herbal formula, Hsiao Yao Pien is also excellent to accelerate progress by promoting improved Qi circulation through the motor nerves and through the muscle itself.

INFRASONIC TREATMENT OF ASTHMA

by Su Cheng Wu,
Department of Pediatrics,
The First Affiliated Hospital Guangxi Medical University,
Nanning, Guangxi, China.

(This clinical study finds that infrasonic sound is effective in treating symptoms of childhood bronchial asthma, and in strengthening immune function as measured by T-lymphocyte population.)

Fifty cases of child bronchial asthma were treated with the Infratonic QGM and routine drugs. T lymphocyte subpopulation (CD2 , CD4, CD8, and CD4/CD8) of peripheral blood was measured to evaluate this treatment's clinical effectiveness on raising immunity. The control group of 15 cases was treated with routine drugs only.

TREATMENT METHOD

Based on the features of child bronchial asthma and a dialectical therapeutics of traditional Chinese medicine, the following four points were treated for 5-10 minutes each morning for 5 days:

Dingchuan (located on the back, 0.5 cun lateral to the lower border of the spinous process of the 7th cervical vertebra). Indications: Facilitating the flow of the lung-qi to relieve asthma. Feishu (Located on 1.5 cun lateral to the lower border of the spinous process of the 3rd thoracic vertebra). Indications: Facilitating the flow of the lung-qi to resolve cough. Tiantu (located on the neck, on the anterior midline, in the center of the suprasternal fossa). Indications: Cough, asthma, chest pain and so on. Danzhong (located on the chest, on the anterior midline, on the level of the 4th intercostal space, on the midpoint of the line connecting the two nipples). Indications: Checking upward adverse flow of the lung or the stomach-qi).

OBSERVATION CRITERIA—FOUR GRADES

1. Clinically controlled: Remission from symptoms. Occasional mild asthma attacks are remitted without taking any medicine.
2. Obviously curative effect: Asthma attack is relieved within a short time.
3. Improvement: Some abatement.
4. Inefficacy: Symptoms, signs and reproduction rate of T lymphocyte subpopulation are unimproved or exacerbated.

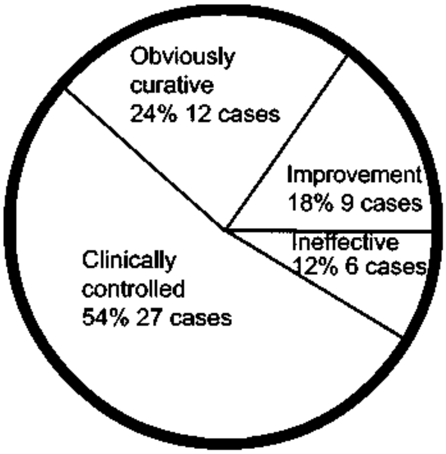

T- lymphocytes in Control vs Experimental Groups

THERAPEUTIC RESULTS

94% of the experimental group showed improvement (as shown to the right). In the control group 80% showed improvement as follows: Clinically controlled - 6 cases; obviously curative - 3 cases, improvement - 3 cases, inefficacy - 3 cases. X^2 determination indicates that there is a more obvious therapeutic effect in the treatment group than that in the control group ($P<0.01$).

LABORATORY RESULTS

T Lymphocyte subpopulation showed a significantly greater increase in the experimental group than in the control group treated by routine drugs alone as shown below:

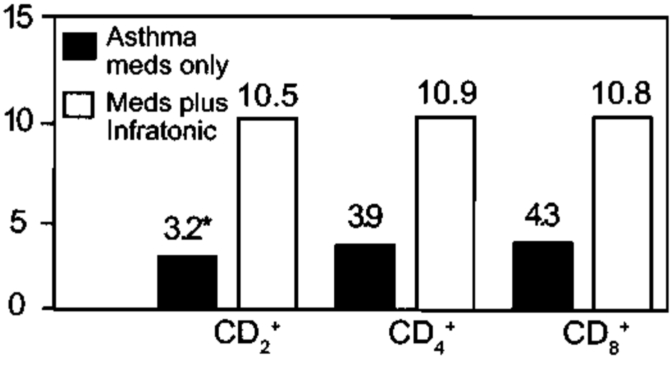

*Notes: Comparison before/after treatment $P<0.01$ unless noted: * = $P>0.05$*

In this study, clinical symptoms of wheezing, coughing and asthma caused by infection of the upper respiratory tract are obviously relieved using the Infratonic device. Testing of peripheral blood showed that the reproduction rate of T lymphocytes subpopulation in treatment group is greater than those receiving routine treatment, indicating increased level of immunologic function of human cells. The Infratonic QGM reduces the course of treatment without discomfort or side-effect. Its biological mechanism of action should be explored in the future.

THE INFRASONIC SIGNAL FROM THE QIGONG PERSPECTIVE

Sun Hua Ling, MD, Beijing Military Qigong Research Institute

(Dr. Sun, a physician and gifted Qigong doctor presents her analysis of the mechanism of emitted Qi and infrasonic signal.)

With the development of human society, health care has become an important part of daily life. Qigong training and techniques and devices like the Infratonic QGM can provide valuable influence. This instrument simulates the low frequency sound of Qigong doctors, which penetrates tissues and organs of the human body, influencing the function of cells, providing clinical results. This study reflects clinical treatment with the QGM of more than 1000 patients.

Cells are units of the human body. Biological electric waves transfer information between cells, forming a biological field of consciousness. The aim of Qigong training is the unification of this field, making the human body more in tune with nature and able to affect nature.

Some "hard Qigong" masters can break stones but can't treat disease because their emitted Qi's amplitude and frequency can not be absorbed by the human body. To be effective, emitted Qi or simulated emitted Qi must be close to the biological waveform of the patient. Then it is easily absorbed and treatment is more likely to be successful. Effective treatment protocols for 30 disease conditions were developed.

MECHANISMS OF THE QGM

1. Relieves pain and edema.
2. Strengthens immune function and reduces inflammation.
3. Normalizes autonomic nervous system.
4. Relieves some symptoms of heart disease.

RELIEVES PAIN AND EDEMA

When applied to sprain of joints, fracture, and arthritis, the QGM has the obvious result of reducing pain distention and edema. It also relieves pain due to sciatic neuralgia, cancer, lumbago, periarthritis of shoulder and stiffness of neck and shoulder muscles. Thus the QGM can effectively treat disease caused by disturbance of blood circulation and can improve local blood circulation.

Patients may feel a sensation of heat, relaxation of muscles, and gradual reduction of pain. This phenomenon may be due to excitation of cells, nerves, and blood vessels around the diseased area. This excitation may contribute to the reduction in pain.

STRENGTHENS IMMUNE FUNCTION AND REDUCES INFLAMMATION

The QGM has a function similar to short wave and ultrashort wave diathermy which strengthens immune function and reduces inflammation. It strengthens metabolism of organs and cells, increasing cellular energy, thus making the immune system stronger.

Diseases such as nephritis, inflammation of intestines and gall bladder, pancreatitis, gas, duodenal ulcer, and infantile dyspepsia, all respond favorably to treatment with the QGM, probably because of this factor of increased cellular energy.

NORMALIZES AUTONOMIC NERVOUS SYSTEM

Where the autonomic nervous system is disturbed, the QGM is often useful. This includes neurasthenia, tenseness of the cerebral cortex, and fatigue. Research shows that the low frequency signals emitted by the QGM penetrate the tissues and organs, altering the internal environment of cells, improving cellular metabolism, and relaxing nerves and muscles, thus regulating the balance between internal and external environment of nerve cells.

Some scholars suggest that emitted Qi may increase cerebral bioelectricity, balancing cell environment, resulting in disappearance of disease. In traditional Chinese medicine, this corresponds to "regulating the meridians and collaterals", "improving blood circulation", and "nourishing internal organs".

RELIEVES SOME SYMPTOMS OF HEART DISEASE

The QGM can relieve symptoms such as lack of blood in the heart muscle, arrhythmia, pressure on the chest, palpitation, and shortness of breath for pulmonary heart disease. This result may be due to improved metabolism of cells and improved local blood circulation.

EMITTED QI TRAINING INCREASES INFRASONIC SOUND EMISSION OF HEALERS

Prof. Lu Yan Fang, National Institute of TV and Electroacoustics, Beijing
Richard H. Lee, Director, China Healthways Institute, USA

(This study supports the premise, proposed by researchers earlier in this chapter, that infrasonic sound can be used as a measure of Qigong ability and indicates that Qigong skills, in this case, the ability to emit infrasonic sound from the palms, can be increased through intensive training.)

Can students in an intensive Qigong training increase the output of infrasonic sound from their hands after one week of training?

To answer this question (and others), 29 Americans traveled to China. The strength of their infrasonic emission was measured before and after the training at the Institute of TV and Electroacoustics, a national institute of technology and standards. Prof. Lu Yan Fang, who conducted the seminal research on infrasonic sound emission two decades ago, conducted the testing. In her research she had found that Qigong masters, powerful healers in China, were able to emit high intensity low frequency sound, 100 to 1000 times stronger than average individuals. Before the experiment she expressed strong doubts as to the possibility of increasing energy emission within one week because Qigong masters require many years to cultivate powerful Qi emission.

Prof. Lu had seen that strong young adolescent males could produce sound as high as 60 dB while average individuals emitted in the 40's and 50's. She was not surprised when in the initial testing 21 of 25 produced scores in the 40's. She felt that an improvement to 60 dB (only one participant had scored in the 60's in the pretest) would suggest improvement to supernormal ability, and that scores above 70 dB would be in the range of the powerful and well known Qigong healers.

A powerful Qigong healer who had scored well in this test before was invited to demonstrate. After several tries, the best he could do was 67 dB, which was 100 times what an average adult could produce, and still in the range of powerful healers, but not nearly as good as he had done previously. He was rather embarrassed and the translator explained that he had worked late the night before and had exhausted his Qi. An hour later he returned covered with perspiration. Apparently, in an effort to redeem himself, he had been building up power doing Qigong exercises for the past hour. This time his score was 78 dB, 10 times the power he had generated just an hour earlier. Participants applauded as he came out of the testing chamber. He had redeemed himself. Interestingly, he was massaging his shoulder as if he had strained it during the Qi emission.

This was a most interesting scenario. First the failure of an experienced Qigong doctor to produce high power, then his performing Qigong exercises to build up his Qi, and finally, his powerful sound emission, offered a persuasive picture that some sort of energy can be stored in the body during Qigong exercises and discharged rapidly through the shoulders to the hands to create intense sound, and possibly straining muscles.

Supporting this finding, one participant was experimenting with the equipment afterwards, emitting Qi into the microphone. At first he could produce high levels of infrasonic sound in front of the microphone, but after five minutes of emitting he could produce only 10% of the power he had produced earlier. He did not feel tired, but as he tried to emit sound, he found that his hand would no longer tremble as it had before. He had apparently exhausted his supply of this form of Qi!

Test Scores: The average score for the group before the training was 47 dB with only four participants scoring above 50. Under the guidance of Master Wan Sujian, a well known Qigong teacher and physician, the group began a 10-day intensive which included climbing mountains and the Great Wall, practicing emitted Qi, and receiving extensive emitted Qi. The group also received physical therapy treatments to open meridians and remove potential blockages in the arms and shoulders which would otherwise restrict the flow of Qi to the hands. The results of both tests are shown below.

After the emitted Qi intensive, the number scoring above 50 dB tripled from four to 13. Five scored above 60 and two above 70. An average increase of 7 dB (to 54 dB) means that students were producing five times the infrasonic energy they had produced before the training. (Statistical significance $t < 0.0005$) Professor Lu was quite impressed. She felt that five of the participants had both the amplitude and spectral distribution associated with powerful Qigong healers.

An interesting technique: While four of the five participants who scored above 60 were men with strong musculature, one petite woman scored 64 dB, which is 10 times the power that anyone else of similar build had scored. When asked her secret, she answered, "Well, to be honest, as I was waiting to emit Qi I silently asked our teacher for help. As soon as I did, I felt a surge of energy flowing through my arm and my hand started trembling powerfully. When I was done, my arm was sore clear up to the shoulder." She had massaged her shoulder after emitting her Qi much as the Qigong master who first demonstrated had done.

SCIENTIFIC QIGONG RESEARCH

Measuring the emitted Qi from Qigong masters has been the purpose of scientific research in many major scientific research institutes in China,

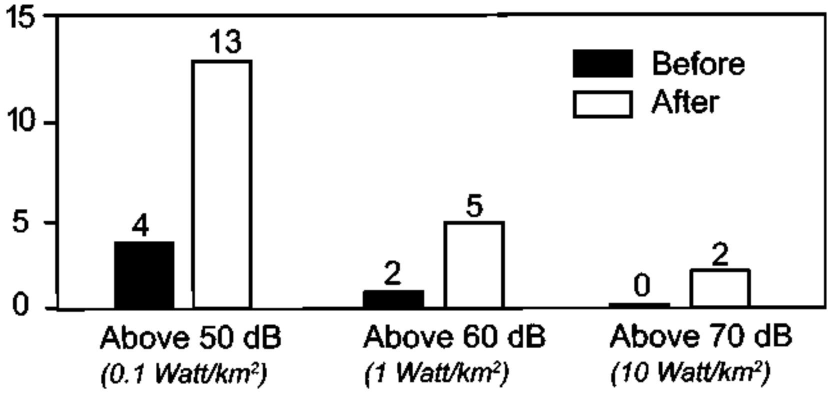

Participants were able to emit substantially higher levels of infrasonic sound after a week-long Qigong training. On the average, scores increased from 47 dB to 54 dB, a five-fold increase in infrasonic energy.

including the following:

- China Immunology Research Center, Beijing
- Institute of Space Medico-Engineering, Beijing
- National Research Institute of Sports Science, Beijing
- Institute of Psychology, Academia Sinica, Beijing
- China Academy of Traditional Chinese Medicine, Beijing
- Institute of Traditional Chinese Medicine, Beijing
- Beijing College of Traditional Chinese Medicine, Beijing
- Tangshan Health Institute for Women and Children, Hebei
- Zhejiang Research Institute of Traditional Chinese Medicine, Hangzhou
- Guiyang College of Traditional Chinese Medicine, Guizhou
- Shanghai Academy of Traditional Chinese Medicine, Shanghai
- The Second Military Medical College, Shanghai
- Shanghai Qigong Institute, Shanghai
- Xing Cheng Air Force Sanitarium
- Zhao Baofeng Beidaihe Qigong Rehabilitation Hospital, Hebei
- Shanghai Institute of Hypertension, Shanghai
- Shenyang Institute of Traditional Chinese Medicine, Shenyang
- Institute for Industrial Health, Xian
- Shandong Medical University, Shandong
- Anti Epidemic Centers of Henan Province, Xin Xiang
- Institute of High Energy Physics, Beijing
- Jiangxi Chinese Medicine and Pharmacy Institute, Jiangxi
- Nanjing Aeronautical Institute, Nanjing
- Institute of Space Medical Engineering, Sichuan
- Chinese Academy of Sciences, Chengdu

While thousands of investigations have been done, just a few have been described here, principally those that are reproducible either with Qigong masters or with Qigong simulation equipment. These studies show that emitted Qi is effective in treating digestive problems, infectious diseases, and bone fractures; that measurable changes in blood chemistry occurs a result of Qigong exercises; and that emitted Qi affects electricity and light in physics experiments.

EFFECTS OF THE EMITTED QI ON HEALING OF EXPERIMENTAL FRACTURE

Jia Lin and Jia Jinding

National Research Institute of Sports Science, Beijing, China.

(This study offers laboratory evidence that Qigong masters can accelerate recovery from fractures. From 1988 FWCAEMQG Proceedings)

The emitted Qi has been found to have a good curative effect on soft tissue injuries such as muscle soreness, scleroma in muscles, acute muscle sprain, muscle contusion and pains. Fracture is also a common injury in sports medicine. We have cured some cases of fracture with the emitted Qi. The purpose of this experiment was to investigate the biological effect of the emitted Qi on healing of fracture.

Sixteen healthy male rabbits, weighing between 1.9 and 2.5 kg, were divided into two groups: the control group and the emitted Qi group. A fracture with a gap of 3 mm was made in the left radius. The rabbits in the emitted Qi group were given the emitted Qi treatment for three minutes per day after fracture. X-ray films were taken every week. Each week specimens for light microscopy from two rabbits were cut into slices of 8*u*m thickness along vertical axis of radius, stained with H.E., and studied by a light microscope. The morphological observations are as follows:

1. Based on some radiographic indexes, such as reaction of fracture section, periosteal reaction and amount of callus formation and callus density, we found that the amount and density of callus formation were better in the emitted Qi group than in the control group. The difference was significant in the 2nd week

(n=16. $P<0.01$) and 3rd week (n=12. $P<0.05$).

2. Ultrathin sections were produced with LKB-III ultrotome, stained with uranyl acetate and lead citrate, and studied with a DXB electron microscope (EM). The changes in myofibrils were observed and photographed (x13000). Injured myofibril volume density in the pictures were calculated by means of point-count method.

The ultrastructural examination revealed that overstrain caused pathological changes such as muscle fiber edema, shortening or lengthening of sarcomeres, disorganization, breaking and disappearance of myofibrils as well as Z lines, accompanied by edema and damage of mitochondria. These changes could be seen less frequently in the emitted Qi group than in the control group. The result of quantitative analysis showed that the volume density of myofibrils of the injured muscle in the emitted Qi group was 2.47%, and that in the control group, 20.41 %.

MEASURING QI IN THE BLOOD

A Preliminary Study of the Relationship Between Qigong and Energy Metabolism—The Changes of the Blood ATP Content in Qigong Masters in the Qigong State

by Wang Zhenchang, Huang Jian, and Wu Zijuan
Shanghai Qigong Institute, Shanghai, China

[This revealing study quantifies a physiological change in the blood during Qigong exercise and Qi emission. When Qigong masters do Qigong exercises, blood ATP level increases, indicating a storing of energy. When they emit Qi, blood ATP level drops, indicating a depletion of energy. This is measurable evidence that physiological changes are occurring when a healer says he is collecting energy or emitting energy. It is also evidence that healers may be exhausted after treating patients and need rest and energy building activities to remain in good health themselves. From the 1988 FWCAEMQG Proceedings.]

Recently it has been reported that the emitted Qi by Qigong masters may have material foundation and the internal Qi cultivation can promote health. In this paper, the relationship between Qigong and energy metabolism was studied. The ATP content was measured with 20*ul* of blood drawn from the ring finger of the Qigong masters before and after the Qigong exercise, and, at rest, by the method of bioluminescence assay.

The instrument (FG-30 Type Luminescence Meter) and the chief reagent (luciferase-luciferin) were produced by Shanghai Plant Physiology Institute. The experiments were conducted in three groups:

1. The experimental group: The blood was drawn from each Qigong master's ring finger after he had taken a short rest, and before he exercised Qigong and entered the Qigong state. If he could emit Qi, he would do it for 5-10 minutes. If he exercised internal Qi cultivation, he would do it for half an hour. After the Qigong exercise the blood was drawn again. After rest for half an hour the blood was drawn once more.
2. The control group: The above procedure was repeated on the second day but the Qigong masters did not exercise Qigong.
3. The healthy group: Healthy persons did as the Qigong masters did at the same time. The results are as follows:

In 10 healthy persons the blood ATP contents were constant during the experiment. For nine Qigong masters in the control group the blood ATP contents did not change markedly. The blood ATP contents in 11 Qigong masters after emitting

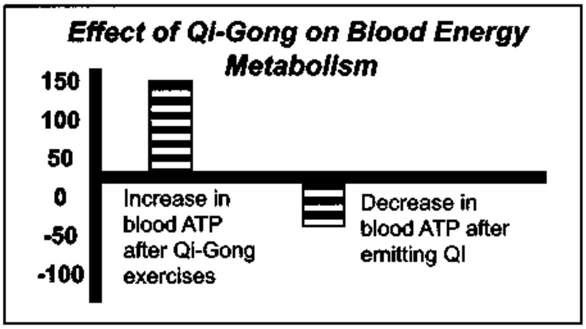

their Qi decreased markedly (the mean decrease was .000131 M compared with that of the healthy group ($P<0.01$). The blood ATP contents in 16 Qigong masters after cultivating the internal Qi increased markedly (the mean increase was .000054 M) compared with that of the healthy and the control group ($P<0.05$). The duration of Qigong practice influenced the changes of the blood ATP content.

The blood ATP contents in both groups of Qigong masters recovered after rest. Qigong masters were observed continually. First they exercised Qigong, cultivating the internal Qi. Then they emitted their Qi. The blood ATP contents changed as above. Conclusions were discussed. The emitted Qi may have a material foundation because the blood ATP contents decreased after the Qi was emitted. The internal Qi cultivation process promotes anabolism and increases CAMP level so that health may be improved.

EFFECTS OF QIGONG ON NEUROTRANSMITTERS

Liu Binghuai, Jiao Jie and Chen Quanzhu
Anhui College of Traditional Chinese Medicine, Anhui Province, China
Li Yefu, Shang Lili (Affiliated Hospital)

[This study shows that you don't need to be a Qigong master to gain healing benefits from Qigong exercises. New Qigong practitioners with chronic illnesses showed significant normalization of neurotransmitter concentration after Qigong practice. From the 1988 First World Conference for Academic Exchange of Medical Qigong (FWCAEMQG)]

It is well accepted that Qigong contributes to preventing and curing of diseases and prolonging life. To study its mechanisms, the authors observed by means of fluorescence spectrophotometry in 68 subjects the variations in blood content of the following monoamine neurotransmitters before and after the Qigong exercises. (5-hydroxytamine: 5-HT, norepinephrine: NE, and dopamine: DA) Forty-four of the subjects were male, 24 females. Their age ranging from 25 to 68 years old, with an average of 48. They all had practiced the Qigong exercise for one month. All were patients of common chronic diseases such as hypertension, coronary heart disease, gastric ulcer, gastritis, pulmonary emphysema, chronic bronchitis, joint system diseases, neurasthenia and neurosis. Results are as follows:

1. A comparison of monoamine neurotransmitter contents in the blood of the subjects pre- and post-exercise showed a general reduction in 5-HT, averaging from 0.42 to 0.21 *ug*/ml, close to the normal value of 0.15 *ug*/ml. The difference was notable ($P<0.001$) Variations in NE and DA tended to go up. The average NE was from 0.27 to 0.35 *ug* /ml, and DA increased from an average of 0.86 to 1.19 *ug* / ml. The above description showed that the effect of the Qigong exercise is closely related to the monoamine neurotransmitters in the body fluid.
2. Effects of the Qigong exercise on the blood monoamine neurotransmitters of patients of different diseases: The subjects were divided into five groups according to the nature of their disease. Subjects in each group showed reduction in blood 5-HT content after they had practiced the Qigong exercise. In all the groups, the content of vasoconstrictor substance 5-HT dropped to the normal concentration level, resulting in sufficient blood supply for the local tissues and organs. It also improved blood flow, promoted metabolism and strengthened the physiological function of the cells in the tissues thus helping to restore health, adding to the efficacy of Qigong.
3. The Qigong exercise helps regulate the neuro-body fluid inner environment. Through regulating the 5-HT blood concentration in the neurotransmitters of parasympathetic nerve, it causes the extra high blood 5-HT of the cardiovascular patients to drop to the normal level. It also dilates the coronary artery and improves the physiological functions, resulting in the relief of symptoms.
4. The Qigong exercise helps build or strengthen

the physiological function of the digestive system. The exercise exerts an influence on the control of nerves that govern the activities of the internal organs. It is shown by variation in the blood content of neurotransmitters NE, DA and 5-HT, with 5-HT getting lower and DA getting higher, facilitating local blood flow and metabolism in the tissues.

5. Through regulating the concentration of pain producer 5-HT in the peripheral blood, the Qigong exercise achieves an analgesic effect. The drop of 5-HT content is relative to the result of analgesia. This fact shows that the drop of 5-HT is good for invigorating blood circulation and reducing extravasation, which is a proof of the TCM principle: "pain is removed when block is eliminated."

The Qigong exercise results in preventing and curing of diseases. It helps regulate the neuro-body fluid inner environment, promotes activities of organs and maintains dynamic equilibrium of the organism through self stabilization.

PHYSICAL CHARACTERISTICS OF EMITTED QI

Gui Yongfan, Chen Qi, Li Yinfa, and Jiang Shen
Nanjing Aeronautical Institute, Nanjing, China

[Emitted Qi can alter the results of common physics experiments shown in this study from 1988 FWCAEMQG Proceedings.]

In this paper, three functions of the emitted Qi which were discovered with the aid of some physical instruments, such as the Van de Graaff generator and light sensitive plate are discussed.

1. The emitted Qi can restrain the discharge process. The discharge between the two balls in the Van de Graaff generator will be stopped or the period of discharge will become longer when the emitted Qi is exerted on the instrument. The Van de Graaff generator was used for testing. The two discharge balls in the generator were pulled apart until the critical state was reached. Then the Qigong master aimed his emitted Qi at the discharge space. The distance between the fingers or palm of the Qigong master and the discharge space was 10 cm to 3 M. We have discovered the following phenomena:
    a. The emitted Qi stopped the discharge or made the period of discharge longer.
    b. The Qigong master had no sensation of getting an electric shock when he emitted his Qi near the discharge space. Nevertheless, when he stopped emitting his Qi, he had the sensation of an electric shock if his hand was still near the discharge space.
    c. In the restraining process, a series of restraint and discharge, such as restraint—intermittent bursts of discharge—restraint, sometimes appeared.
2. The emitted Qi can cause the exposure of light sensitive plates wrapped in a piece of black paper. By false color image processing with a picture processing instrument, we obtained a color photograph, in which the image of fingers and the Laogong (P-8, center of palm) can be seen clearly. The Qigong master emitted his Qi at the emulsion layer of a light sen-

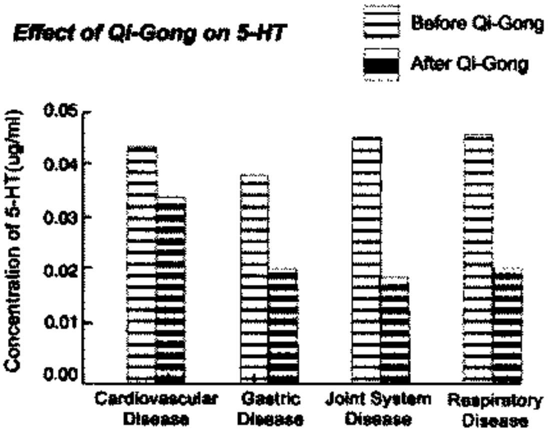

The concentration of the monoamine neurotransmitter 5-hydroxytamine decreased toward the normal value of 0.15 ug/ml in the above groups of patients after one month of Qigong exercises.

sitive plate from about 10 cm away for about 10 minutes.

RESULTS

a) The color of the controls was darker than that of the affected plates. b) After development, white speckles, various patterns or orderly stripes appeared on the affected plates.

SUPERCONDUCTING MAGNETOMETER MEASURES CHANGES DURING QIGONG

Wu Benjie, Dept. of Biophysics, Beijing Medical University
Wang Xiubi, Institute of Somatic Sciences, Beijing,
Wang Zijun, Nanjing University
Liu Jianben, Institute of Geophysics, Beijing

[Here we see that Qigong masters can produce magnetic fields from their palms, and that these fields have characteristic frequency bands. From 1988 FWCAEMQG Proceedings.]

A dynamic study of Qigong (QG) magnetic signals (MS) by a flux-gate magnetometer was conducted in a magnetically shielded room. Under the QG state, MS was <2Hz in frequency and several *n*T (T = Tesla, the unit of magnetic flux density, $1nT = 10^{-9}$ T) in amplitude. For deeper understanding of the MS dynamic properties, human weak MS were investigated by a superconducting biomagnetometer which is wider in frequency response (DC to 10 kHz instead of DC to 10 Hz) and higher in sensitivity. Through a frequency-power spectrum analysis by a computer, the MS changes under QG state were analyzed quantitatively.

Twenty persons (38 tests) were selected as subjects, and divided into two groups. (1) The control group consisting of three persons did not practise Qigong, 46.3 years old in average; (2) The Qigong group (practising for 1-20 years) consisted of 17 persons (31 tests), 48.3 years old in average.

The subjects took off all magnetic things; then came into a magnetically shielded room, with the acupoint to be measured aiming at the probe. The distance between the probe and Laogong (P 8) was about 6 cm, and that between the probe and head acupoints was about 4.5 cm. The MS dynamic changes were recorded continuously, before, during, and after QG state (3 minutes each).

MS was measured by a superconducting biomagnetometer, Model M-601. The frequency power spectra were analyzed by a signal processor, Model 7T17-S. In this paper the power changes were all of relative values.

The experimental results indicate that under the QG state three kinds of MS (1Hz, 3 Hz and <2 kHz) are measured from different acupoints Laogong (P 8), Baihui (Du 20), Yintang (Extra); there are three kinds of MS changes in intensity: increasing, decreasing and no change; and there is a good repetition of the change patterns and signal frequency of MS. Two novel phenomena were found, i.e. decrease of power and the change of frequency spectrum in MS during QG.

QIGONG AND CANCER

Cancer is one area where western science can use some help. Our government has pumped billions into cancer research, yet the problem continues to grow. Chemotherapy and radiation have such negative side effects that the time has come to look beyond therapies of killing "bad" cells, toward therapies with which the individual patient brings his own body into order.

In China, the obvious answer is Qigong. Cancer occurs when the cells of the body follow the will of an outside influence, perhaps a bacteria or a virus or a genetic predisposition. If Qi is strong and flowing smoothly within the individual, many Chinese believe that the cells will do the will of the organism, not of an outside influence.

A very interesting experiment that illustrates the will of an individual influencing cell growth is presented in a study by Dr. Feng Lida, MD, Ph.D., a famous Beijing researcher. She had Qigong masters treat vials of E-coli bacteria. The first vial was held but not treated. The Qigong master emitted Qi to the second one to try to "kill" the bacteria. He directed "health-promoting" Qi to the third one. Dr. Feng reported that, with over 40 repetitions, the health promoting dose resulted in a seven- to tenfold increase in the number of E. coli. The health destroying dose lowered the number of bacteria by one half or more. This experiment illustrates the power of the Qigong master in influencing cell growth. Every one of us has the latent ability to do the same, directing Qi to control disease within our own bodies. This can be seen from the success of programs emphasizing visualization and meditation.

If Qi is weak in an area of the body, abnormal growth will appear. Qi is weakened by poor nutrition and lack of exercise. It is also weakened by mental and environmental stress. Chemotherapy and radiation put heavy additional loads on the body, draining it of Qi. This is why methods to strengthen Qi are so valuable.

This chapter is devoted to Qigong as it is applied to cancer. It is clear from these studies that Qigong is an option that every person trying to beat cancer should consider, regardless of whether western techniques are also pursued. Chemotherapy and radiation in many cases have been shown to be of little value improving quality or length of life. They are not the only option. The following evidence shows that Qigong is an important therapeutic modality for treatment of cancer.

THE EFFECT OF THE EMITTED QI ON CARCINOMA CELLS

Feng Lida, Qian Ju Qing, Chen SuQing, et.al.
China Immunology Research Center, Beijing, China

[This influential paper is the first presented in the 1998 FWCAEMQG Proceedings, and shows that emitted Qi can kill cancer cells and alter chromosomes in vitro, providing clinical evidence that cancer patients would do well to learn Qigong.]

We used the techniques of tissue culture, cytogenics and electron microscopy to study the Hale cells and SCG-7901 human gastric adenocarcinoma cells, the change of the cellular ultrastructure and abnormal structure of the chromosomes of the cells. The results showed:

1. The effect of the emitted Qigong on the Hale Cells:

We repeated the experiment 20 times under identical conditions. After the Hale cells received the emitted Qi for 20 minutes, the survival rate of the cells in the experimental group was 69.28% of that in the control group, i.e. 30.72% of the cells in the experimental group were killed. At the same time we observed the Hale cells with a translation electron microscope, and found that degeneration and swelling took place in some of the cells which received the emitted Qi. The endoplasmic reticula were dilated obviously, and the nuclei of the cells were destroyed.

THE EFFECT OF EMITTED QI ON THE GASTRIC ADENOCARCINOMA CELLS:

After the gastric adenocarcinoma cells received the emitted Qi for one hour, the survival rate of the cells was 74.98% of that in the control group. i.e. the average destruction rate of the cells was 25.02%. We repeated the experiment 41 times under identical conditions, and through statistical analysis, found a difference between the two groups ($P<0.01$). In the meantime, a scanning electron microscope was used to study the cells. The results showed that the microvilli of the cells which received the emitted Qi had remarkably dwindled or disappeared and tiny holes could be seen on the surface of the cells.

3. The effect of the emitted Qi on the chromosomes of the gastric adenocarcinoma cells: We repeated the experiment 10 times under identical conditions. 1420 tumor cells in the control group and 1,428 tumor cells which received the emitted Qi for 60 minutes were analyzed. The subdiploid in the control group was 2.11%, the superdiploid was 0.14%, whereas the subdiploid in the experimental group was 3.29% and supertriploid was 0.56%. The statistical significance in these findings ($P<0.05$).

The structure of the chromosomes of the gastric adenocarcinoma cells also varied. In the experimental group, the total abnormality rate was

5.39%. The total abnormality rate in the control group was 1.40%. ($P<0.01$).

SUMMARY:

Tumor is a common disease which endangers man's life and health. The study on the effects of the emitted Qi on carcinoma cells, especially on the study of the structure and the chromosomes of tumor cells, has not been reported in the world before. The human chromosome is a material base of human heredity and a carrier of the genetic material. The results of our studies showed that the abnormality of the chromosomes of the gastric adenocarcinoma cells was significant after they received emitted Qi. The rate of exchange, breaking, and the dicentromere in the structure of the chromosomes of the gastric adenocarcinoma cells is increased. The results suggested that the emitted Qi could kill the chromosomes of the gastric adenocarcinoma cells. The study may start a new way to cure carcinoma.

QIGONG AS A THERAPEUTIC AID FOR ADVANCED CANCER PATIENTS

by Sun Quizhi and Zhao Li, Kuangan Men Hospital China Academy of Traditional Chinese Medicine, Beijing, China

[This article shows that, in clinical trials, cancer patients were far better off when they combined Qigong exercises and chemotherapy as compared to chemotherapy alone.]

One hundred and twenty three cases of advanced cancer treated simultaneously with drugs and Qigong has been reported in this paper. These were inpatients of our department during the past five years. Among them there were 60 males and 63 females. The youngest was 20 years old. The average age was 47 years old. These patients were diagnosed pathologically as malignant cancer, 70 cases were in the III stage, 53 were in the IV stage, of which, cancer of the stomach accounted for 42.23%, cancer of the large intestine 31.70%, mastocarcinoma 11.29%, and cancer of the esophagus and cardinal carcinoma, 8.1% each.

The 123 patients were divided into two groups. The experimental group included 93 cases treated with drugs and Qigong exercises. The control group, which consisted of 30 cases, was treated by drugs alone. Similar drugs were given to the two groups. Each patient who received the therapy of drugs and Qigong did the Qigong exercises for over two hours a day, three months as a course. Symptoms, signs, body weight, and immunological indices, etc. were recorded before and after treatment. The results were as follows:

1. Amelioration of general symptoms: In the experimental group, 81.7% of the patients regained strength, 63% improved appetite, 33.3% were free from diarrhea or irregular defecation in the experimental group, while in the control group the rates of patients were

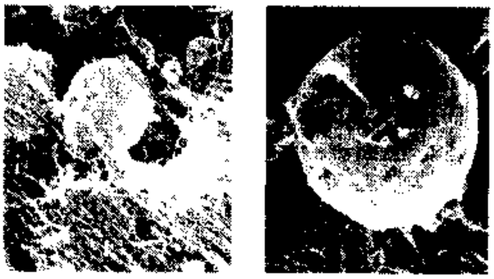

Cells of lung cancer before irradiation with emitted Qi

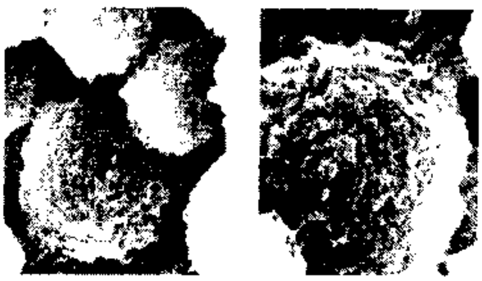

Villus exfoliation of lung cancer cells after irradiation with emitted Qi.

10%, 10% and 6%, respectively. The difference between the two groups is significant. (P<0.01)

2. Body weight: In the experimental group, 50.54% of the patients' body weight increased more than 3 kg, and 5.4% of the patients' body weight decreased more than 3 kg. In the control group, the rates were 13.3% and 30.0%, respectively. The difference between the two groups is, again, quite significant.
3. Phagocytosis of macrophages: In the experimental group, the phagocytic rate of macrophages increased by 12.31 % and the index of macrophages increased from 0.45 before the combined therapy to 0.63 after. In the control group, the phagocytic rate did not elevate, but decreased by 7.87% while the phagocytic index also dropped from 0.63 to 0.50.
4. In addition, it has been observed that, in the experimental group, the erythrocyte sedimentation of 23, and hepatic function of 20 of the 93 patients returned to normal, while in the control group, there were only three patients with normalized erythrocytic sedimentation, and two patients with normalized hepatic function, out of the 30 control cases.

To sum up, these studies suggest that the Qigong therapy is helpful to some extent to ameliorating the symptoms, improving appetite, strengthening constitution, and increasing the ability of self-cure. Qigong can regulate the mind, heart, Qi, and blood so as to get rid of pessimism and eliminate evil factors. Therefore, it is really an effective, simple supplementary treatment with no side effects.

THE EFFECT OF EMITTED QI ON LEUKEMIA IN MICE

Feng Lida and Zhao Xiuzhen
China Immunology Research Center, Beijing, China

[This study by Feng Lida, MD, PhD shows that emitted Qi can decrease the reproduction rate of leukemia cells in mice and that infrasonic Qi simulators can play a similar role. From the 1988 FWCAEMQG proceedings.]

With the rapid development of modern science and technology and the enhancement of people's living standard, the constitutive proportion of diseases has changed and most infectious and parasitic diseases have been brought under control or eliminated, whereas, tumors are becoming a common problem and are endangering man's life and health. In our country, the number of patients suffering from tumor diseases is over 1,000,000 a year, and the number of patients who died from tumors has exceeded 800,000. Therefore, it is important for us to explore a new route to cure tumors.

Leukemia is a malignancy, which has a high incidence and is difficult to cure. As there are some similarities between leukemia in mice and human leukemia, we chose DBA mice as a research model to study the effect of the emitted Qi on L_{1210} cells of leukemia in mice.

The experiment chose DBA mice with a weight of 20±2 gm, and both male and female mice were used. In the experiment we killed the mice, which had been injected with the L_{1210} cells 7 to 9 days before we collected the celiac liquid and adjusted the concentration of cells to 4.70-31.00 million/ml (avg. concentration was 1.20 million/ml). 0.2 ml L_{1210} cells was injected i.p. into the abdominal cavity of mice. And after one day we randomly divided the mice into groups.

One Group Treated With Emitted Qi

Among these groups, the mice of the experimental group received the emitted Qi once per day, for between 10 and 40 minutes for 10 days, while in the control group, no treatment was given. After 10 days the mice were killed and the number of L_{1210} cells were counted with a light microscope. The average value in the control group was 200.435x0.5 million per mouse, while in the experimental group 66.458x0.5 million per mouse and there is significant meaning between the two groups. (P<0.01).

Second Group Treated With Infrasonic Qi Simulator

At the same time, we used an infrasonic qi

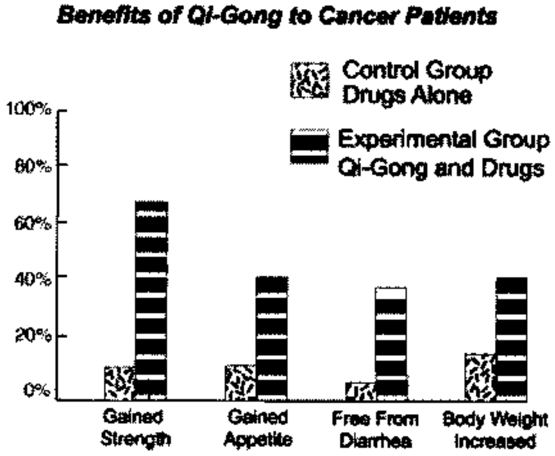

123 advanced cancer patients were treated for 6 months. The control group with drugs, the experiemental with drugs and Qigong exercises. The results show the value of Qigong exercises in cancer treatment.

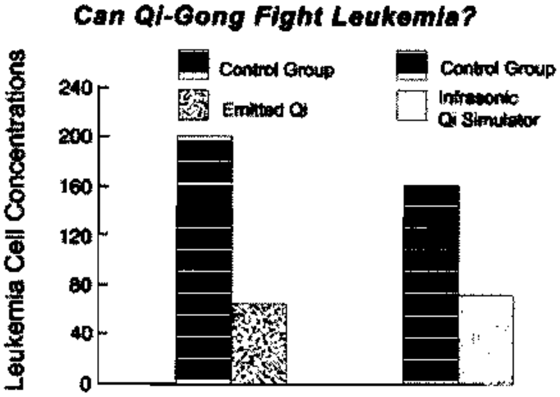

Concentration of leukemia cells in mice after 10 days' treatment with emitted Qi and the infratonic Qi stimulators

simulator to stimulate the mice injected with the L_{1210} cells, once per day, two hours at a time for 10 days, after which the number of L_{1210} cells was observed. The number in the control group was 160.826x0.5 million per mouse, while in the experimental group 70.870x0.5 million per mouse. This difference is of statistical significance ($P<0.1$). The results showed that the number of L_{1210} cells which were injected into mice could be remarkably reduced after the mice received emitted Qi, and the fact suggested that the emitted Qi could kill or inhibit the L_{1210} cells in mice.

Qigong is a part of traditional Chinese medicine, which is a therapy with Chinese characteristics. In recent years, Qigong has been used to cure a lot of diseases, especially difficult diseases and tumor diseases. It is an important issue whether Qigong can kill or inhibit malignant tumor cells or not. The result of this experiment shows that the emitted Qi could kill or inhibit the L_{1210} cells of leukemia in mice. After a body received it many times, the emitted Qi could significantly reduce the number of L_{1210} cells in mice. The experimental study laid a theoretical basis for the treatment of tumors with Qigong. Qigong therapy will be a new way to cure carcinoma. However, the mechanism and way that the emitted Qi kills or inhibits L_{1210} cells of malignant tumor cells in mice needs to be further investigated.

EFFECTS OF QIGONG ON MALIGNANT TUMOR

Luo Sen, Tong Tianmin, et al.
Zhejing Institute of Traditional Chinese Medicine
Hangzhou Qigong Hospital, Hangzhou, China.

[This study, also from the 1988 FWCAEMQG proceedings shows the benefits of combining traditional and modern cancer therapies in maintaining healthy blood cells.]

Under observation in this experiment were 80 patients with 17 kinds of malignant tumors including nasopharyngeal carcinoma, breast cancer, lung cancer, and gastric carcinoma. Their diagnosis was confirmed by cytological and radiological examinations. Within this group, 48 were male and 32 female. Most of them had been operated on before hospitalization. They were at their I or II stage of the disease and had received radiation or chemotherapy before. These patients were randomly divided into three groups. 30 of them, as the first group, received Qigong treatment.

The second group, composed of 25 cases, were treated with routine chemotherapy. And the third group, another 25, received chemotherapy combined with Qigong exercises. The parameters adopted included WBC and RBC count, serum hemoglobin and platelet count in blood before and after treatment, T-lymphocyte conversion test (H3TdRLT) and urine excretion of 17-hydroxy steroids within 24 hours before and after the last week of Qigong therapy in some patients. The total period of observation lasted 60 days. Results are reported as follows:

1. Group I, the Qigong group, developed a significant rise in their WBC, RBC count and serum hemoglobin after treatment ($P<0.01$). Group II, treated with chemotherapy, presented a significant lowering in these measures ($P<0.01$). Group III, which both practiced Qigong and received chemotherapy, showed an obvious elevation of serum hemoglobin, RBC and platelet count ($P<0.005$-0.01). Their WBC remained at the same level as before the treatment ($P>0.05$)
2. T-lymphocyte conversion test in Group I also showed a significant rise ($P<0.05$). It was seen to be maintained at the original level in Group III ($P>0.05$).

This experiment proves that this kind of Qigong exercise, by exciting the circulation of Qi along meridians and modulating the functions of blood circulation, immune response and endocrine tissues, can minimize the toxic side-effects of chemotherapy and enhance energy supply to the body. This practice is effective in treatment of cancer and supported by confirmed scientific experiment.

EFFECTS OF PATIENT TUMORS ON KIRLIAN IMAGES OF HEALTHCARE PROVIDERS

Su Cheng Wu, Guangxi Tumor Hospital, Guangxi, China
Richard H. Lee, China Healthways Institute, USA

The process of emitting Qi during a Qigong treatment, while valuable at jump-starting a patient's recovery, is exhausting and requires extensive Qigong exercises to replenish the depleted Qi. However, as Qigong practitioners become more skilled, they can treat more patients with less depletion, probably through a combination of becoming more proficient at collecting Qi and more efficient at utilizing the Qi that they do have. This is an important field of study because of the prevalence of "burnout" which affects many practitioners in all the medical professions. Perhaps doctors and nurses can be taught the principles of efficient use of their Qi. The following is offered to illustrate this point:

The top two Kirlian photographs on this page show images of fingertips of healthy individuals and healers. The light recorded by the photos is created as electrons moving between the fingers and the film ionize air. In the first photo, the balls are created as electrons leaving the finger create pools of light. And the streamers are created as electrons stream back from the film to the finger. A balance between balls and streamers indicates that electrons move both to and from the fingers. The second photo shows that, around the fingers of healers, the electrons travel more as waves than particles, creating a smooth glow instead of balls and streamers.

In cancer patients, research has shown that balls appear around the cancer (not illustrated, reference Dumitrescu and Kenyon in "Electrographic Imaging in Medicine and Biology", 1983) and streamers appear around all other parts of the body, as shown in the fourth picture. (reference "Bioelectrography" by Konikewicz and Griff 1984 and Dumitrescu, 1983). This strongly suggests that the rapid growth of cancer cells is fueled by high electrical vitality at the cancer site, and that the rest of the body is devitalized as the high energy electrons are somehow attracted by the cancer.

The lower three Kirlian photos were taken in a tumor hospital. A typical cancer patient (the third photograph) shows virtually no Kirlian image. This indicates a very low flow of electrons, either in or out of the finger, and thus, low vitality. After increasing the patient's electrical conductivity using an Infratonic QGM on the foot (K-1), the patient's image brightened, but showed only streamers but no balls, as electrons shot to the fingers but did not return from the fingers to the film. Had high energy electrons returned to the film, we would have seen balls in addition to streamers in the photographs. This suggests that the entire patient, not just the photographed finger, is depleted in electrical vitality, and may be drawing electrical vitality from the Kirlian camera as high voltage electrons stream into the patient's body, but do not return to the film. This supports the common belief that cancer draws vitality from everything in the surrounding area while fueling the very rapid growth of the cancer cells.

Several doctors in the tumor hospital were photographed and those with patient contact showed a similar image dominated by streamers with almost no balls. This suggests that the doctors were in some way devitalized by the patients' low energy condition, and were, themselves, drawing electrical vitality from the camera. It appears that cancer drains the vitality of everyone, the patient, health care providers, family, and even those who visit to show their support.

Why does vitality flow to the cancer site? Qigong practitioners believe that intentionality moves Qi, which suggest that intentionality might be acting. We might suspect the intentionality of the patient, the family or the doctor. All seem to be convinced that the cancer will win. It seems that we should also look at the possibility that the cancer, itself, has intentionality, and is causing the patient to give up hope and the doctor to say things like, "You have no more than 3 to 6 months

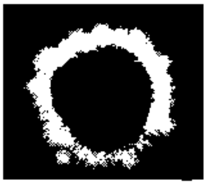

Kirlian image of typical healthy person shows balance between balls and streamers.

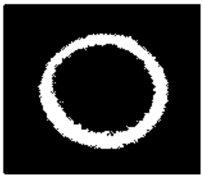

Kirlian image of typical healer

Cancer patient shows low conductivity.

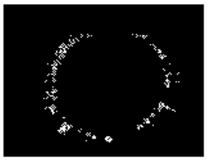

Cancer patient shows increased streamers after QGM treatment.

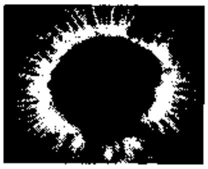

Doctor with patient contact also shows streamer dominance.

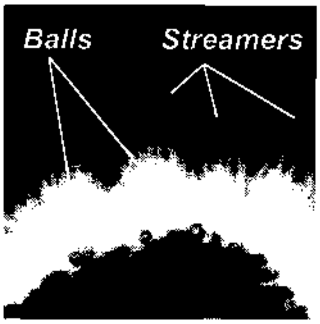

to live." Perhaps the fight against cancer is a battle of wills over the flow of vitality more than simply the biologically understood multiplication of ignorant cancer cells.

Can we reverse this flow of vitality, and will this reverse the growth of cancer? The preceding research in this chapter shows that Qigong practice and emitted Qi can kill and decrease the growth rate of cancer. Perhaps intentionality is a new weapon against cancer. And maybe the Kirlian camera will be the biofeedback technique which shows patients and researchers when their intentionality is properly applied and when they are winning the battle.

As is concluded in most research studies, "More research is necessary."

CHAOS AND CONSCIOUSNESS

The study of Qi as the relationship between the physical world and consciousness is often avoided by scientists because this violates a basic tenet of science, the experimental method, that the experimenter is not to interfere with the experiment. However, this appears to be exactly what Qigong masters are doing. Whereas religious groups have long held that our thoughts, wishes, desires, and prayers can influence our own lives and the lives of those around us, the scientific world has, for the most part, denied this possibility.

If, in our study of Qigong, we discount the conscious intent of the practitioner, we are ignoring what Qigong practitioners claim is the vital ingredient, intentionality. Thus, we are led to study the physics of the phenomenon, the interaction of electric, magnetic, and acoustical fields. However, if we allow intentionality to enter the equation, a very different view of Qigong emerges with enormous implications.

To accept that consciousness can influence the physical world, science needs some sort of field or substance which can be shaped by consciousness such as the rhythmic movement of electrical charges in the body as measured by EEG, which is shown to be influenced by Qigong masters.

From a broader perspective, this field is chaos, which might be described as uncertainty regarding the future, or instability such that subtle influences can guide the field's reorganization. In the case of EEG, the chaos is the available instability of free electrons, molecular ions, and neurochemicals. There are probably many other forms of chaos within the body including magnetic, acoustical, and electromagnetic. It is chaos, or uncertainty in these fields, which allows consciousness to enter and shape reality.

Qigong therapy is often broken into three phases. First, the practitioner "adds Qi" increasing the available energy, or chaos of the patient. Then he "sweeps the bad Qi out the feet" to remove crystallized structures which are interfering with the patient's health. Finally, he "smooths the Qi" to help bring order out of the chaotic state which results from the adding of chaos and removing of "bad Qi."

Thus, a Qigong treatment might be viewed as adding chaos to liquefy a crystallized structure, removing undesired, loosened crystals, then smoothing and harmonizing the field to "bring a higher order out of the chaos." Softening and bending metal, altering EEG in anesthetized animals, and accelerating recovery from cancer might all be explainable by "consciousness creating and guiding chaos."

LOW DIMENSIONAL CHAOS IN PRACTICING QIGONG

Yagu Lin, Jijun Gao, and Huo Lu

Shanghai Qigong Research

[In our research at CHI, we have found that chaos is a key ingredient in the effectiveness of a therapeutic signal, allowing greater penetration into the body. This Chinese chaos research shows that Qigong exercises result in reduced complexity of the EEG in the brain, providing calmness and mental clarity (bringing order out of chaos). From the FWCAEMQ.]

Recently, researchers such as Kaczmark, I.K. (1977), Nicolis, G. (1985), Skarda, C.A. and Freeman, W.J. (1987), and Xu Jinghua and Xu Nan (1987) have reported the systematic complexity and chaos of the brain electroencephalographic techniques (EEG) by the analysis of dimensionality in studies of chaotic dynamics of biological systems. However, there are few reports analyzing the Qigong functional state using this method.

In this paper, we analyze change in attractor dimensionality d values of the occipital EEG before, during, and after Qigong exercises. The results show that before exercises, d=1.8, during exercises d increases to 2.1, and after Qigong exercise, it remains at the increased value of 2.1. These results show that Qigong exercises can lower the complexity under conscious conditions, increase the orderliness, and decrease the randomness. This may explain Qigong's effect of "warding off distraction, thoughts, or evils," and may be a principal mechanism of Qigong's effectiveness.

QIGONG TRAINING AND COHERENCE OF EEG

Yang Sihuan, Yang Qinfei, Shi Jiming, Cao Yi
Institute of Qigong Science,
Beijing College of Traditional Chinese Medicine,
100029, Beijing, China

[This study shows that Qigong training increases coherence between the right and left sides of the brain. From the perspective of chaos theory, this means that Qigong helps to bring "order out of chaos" in the brains of practitioners. From Second World Conference on Academic Exchange of Medical Qigong.]

In this study, we analysed the coherence of EEG to observe the trainees' EEG regularly. The young students, who were 17 to 20 years old, had been practicing "Zhanzhuang Gong" for one year. We try to find out the effects of the Qigong training period on coherence of EEG. Thirty two persons in the Qigong group and 35 persons in the control group were involved in this experiment. During one year period of observation the subjects of the Qigong group practiced Qigong for 40 minutes every day. The EEGs of the Qigong group were analysed every half year in meditation, and the EEGs were also recorded before learning Qigong. The students in the control group did not take part in the Qigong training and their EEGs were investigated at rest twice with an interval of one year. In the test, eight channels of EEGs were simultaneously processed by a computer on line for 20 minutes. The program, "computer evaluation system for the Qigong state" was provided by the "Laboratory of Bio-Control, Department of Electrical Engineering, Zhejiang University".

After one year of Qigong training, total coherence between the left and right frontal regions increased from 0.84±0.07 to 0.87±0.06 ($p<0.05$). Before Qigong training, the total coherence between the left and right occipital areas was 0.68±0.14. After half year's training, it increased to 0.79±0.10, and after a year's training, it was 0.76±0.10. Self comparison showed the probability $p<0.001$. The total coherence between the left and right temporal areas before Qigong training was 0.48±0.17. Half year after Qigong training it was 0.55±0.13, compared with that before Qigong training ($p<0.05$). One year after Qigong training it was 0.64±0.12. Comparing before and a half year after Qigong training $p<0.001$.

Total coherence did not change significantly in the control group

VIBRATIONAL TREMBLING, CHAOS, AND EEG

Dale M. Patterson, MS, CACB, BCIA Certified
Richard H. Lee, China Healthways Institute, USA

[This article, published in China Healthways Newsletter, provides data collected from biofeedback research conducted by Dale Patterson which shows that chaos increases alpha EEG induction.]

Vibrational trembling appears to be a way memory is stored or processed in the physical body. Chaotic sound penetrates this field of vibrational activity, softening this memory and allowing us to more easily let go of fixated thoughts, addictive desires, and compulsive actions.

When someone is full of anger at being blocked from acquiring the object of desire, he may tremble extensively in the theta range of EEG. And when someone is overloaded with the stress of excess thinking (worry) and approaching a nervous breakdown, he will also tremble, but this time in the beta range. This excess vibrational activity consumes the body's Qi and causes the person to become emotionally, mentally, and physically rigid.

Applying chaotic sound breaks up these repeating patterns, often freeing the patient from frustration and worry. As the patient relaxes, capacity to process thoughts and emotions increases, new answers are found, and the increased freedom and flexibility are experienced as increased vitality. Thus, where physical memory inhibits healing, adding a chaotic alpha signal can accelerate recovery.

LABORATORY TESTING

Dale Patterson, a board certified EEG technician trained in biofeedback became intrigued with the QGM when one of his patients brought one in. He recorded EEG brain maps first with an early model of the QGM which produced a clinically significant 18% increase in Alpha. He then used the Q4 with chaos which provided eight times the increase in alpha activity, showing that chaos increases induction of alpha rhythm in the EEG.

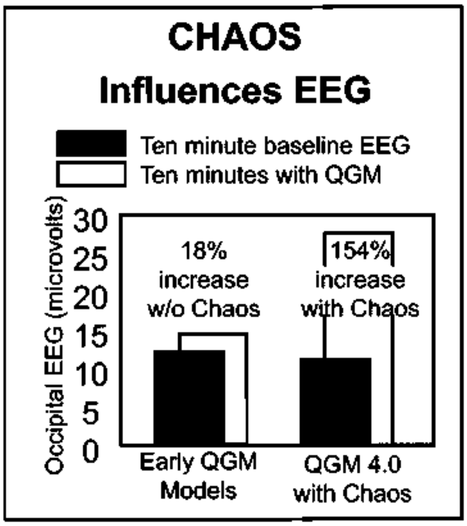

CHAOS AND THE COLLEGE ENTRANCE EXAM

Su Cheng Wu, Guangxi Medical Hospital

[From China Healthways Newsletter: A medical researcher applied the Infratonic QGM to high school students preparing for the college entrance examination.]

There was little chance that Li Wen Chao could ever attend college. As a high school senior, he had achieved only moderate grades in his courses and had done poorly on previous national examinations. Without very high scores in the China College Entrance Examination, Li would almost certainly spend his life in a menial factory or farming job. The College Entrance Examination is terrifying, because everyone who faces it knows that only half of high school graduates will enter the nation's college system. *Failure in this examination is a common cause of suicide in China.*

Enter Su Cheng Wu, a medical researcher and Li's uncle. He had just completed a successful study involving infrasonic treatment of bronchial asthma (see page 28) and was searching for other subjects to study. He felt that stress, nervousness, and mental overload were the principal reasons that high school seniors failed the examination, and believed that the Infratonic QGM would induce the deep calmness and mental clarity of Alpha into the students' minds and bodies, helping them to relax and overcome the mental overload that causes brains to lock up during tests.

For three days before the examination, Dr. Su went to visit his nephew's class and treated each patient by holding the machine on the upper back for five minutes and encouraging the student to talk about fears and tensions regarding the upcoming test. For students suffering from symptoms such as headaches, dizziness, poor appetite, insomnia, and menstrual pain, he instead selected and treated an appropriate point, usually along the front midline of the body.

The results of this small test drew much attention. Li's class was average among the seven senior classes, yet performed the best by far. Whereas the average pass rate for the other six senior classes was 50%, 86% of the students in this class passed the examination and were placed in five year colleges! The other 14% got high enough scores that they were accepted to three year trade schools. Thus, 100% made it to some sort of college. Li Wen Chao enrolled the next year at Guangxi Medical University to become a physician.

While it sounds unreasonable that a little chaotic infrasonic sound can cause such a large increase in scores, the opposing view makes more sense. *Stress and worry can shut down the thinking process.*

SEARCHING FOR JING, QI, AND SHEN IN WESTERN SCIENCE

Traditional Chinese medicine (TCM) holds that three "vital treasures", Jing, Qi, and Shen, are real substances which envelop the human body and are essential to life and healing. These are the fields through which Qigong masters work. Jing provides the genetic structure or the physical strength of the body. Qi is the life or vitality of the body, our energy level. And Shen provides the mind, or conscious aspect. TCM goes further, introducing three seeds of consciousness, the lower, middle and upper Dan Tians located in the lower abdomen, chest, and above the head respectively, which are born and mature through cultivation, providing awareness of the physical world through Jing, Qi, and Shen.

This parallels the writings of Alice Bailey which present a three fold personality: the etheric body (Jing) is shaped by the emotional body (Qi) which, in turn, is shaped by the mental body (Shen). The substance of each body is produced by the physical body whereas the function of each body is to support a mode of consciousness.

In "A New Science of Life" by Rupert Sheldrake a "morphogenetic field" is described as a vibrational field which surrounds and shapes living things as they organize. He argues that our genetic structure doesn't contain, but rather tunes into the blueprint of the physical body through this morphogenetic field. He supports this with laboratory findings that as more and more animals and humans are taught a new behavior, others of similar genetic makeup but physically separated from the first, learn the new behavior more and more quickly. Jing appears to be the TCM equivalent to the morphogenetic field of the physical body. Qi and Shen might also support morphogenetic fields of their own. According to Sheldrake, these morphogenetic fields are programmable, obtaining their initial programming from genetic structure, and being reprogrammed through experience and learning.

Jing, from the standpoint of physics, might be a field of electrons in constant flux suspended in salt water throughout our bodies. The rhythmic movement of this electric field is measurable by EEG, EMG, and EKG. Electrostatic forces are very strong and it is conceivable that such a field can shape a physical body by catalyzing certain chemical reactions and moving certain molecules to the right place at the right time, causing, for example, differentiation between liver cells and brain cells in the embryo.

If Jing is to vibrate freely it requires sufficient energy to have free flowing electrical charges both to resonate with the morphogenetic field and to fuel biochemical reactions. Ilya Prigogine, in "Order Out of Chaos", proposes that life requires chaos, that higher forms of organization emerge out of chaos, and that raising the energy of a system allows it to reconfigure into a higher order of organization. From the standpoint of Jing, when biochemical activity exhausts the available electrochemical potential, chaos in the field of electrons has been exhausted. Then the physical body may become unresponsive, and illness may result. Research during long airline flights shows that, where crowding and recirculated cabin air reduce the availability of electrical energy in the air, people become depleted and develop exhaustion and succeptibility to illness and infection. When passengers wear portable air ionizers, which add electrical instability (chaos) to the air by ionizing it, they do not suffer from exhaustion and avoid illnesses from airborne bacteria. From the standpoint of physics, Jing might be the available electrical instability or chaos in the sea of electrons in the body.

Qi might be a field of magnetic substance which can be measured because it makes the body more electrically conductive. Such diagnostic equipment as Ryodoraku and EAV are claimed to measure the Qi by measuring the electrical conductivity of meridians. High and balanced conductivity are associated with good health and vitality, and low or imbalanced conductivity is associated with illness and debility. While the Qi may have little direct influence on the physical body, its ability to increase local electrical conductivity gives it the ability to provide selective pathways of decreased resistance which allows it to influence the flow of electrons, and thus, shape the Jing. Electrically conductive needles may give acupuncturists a similar ability.

Analysis with Kirlian photography shows that washing the hands with tap water causes the electrical conductivity of the hand to decrease considerably. However, when the hands are washed with the same water which is first passed through a strong magnetic field, the electrical conductivity of the hands increases. (ref: Bioelectric Vitality — The Science of Human Energy by Lee) This tells us, first, that water can store and transport some sort of magnetic substance, and second, that this substance can be removed from or added to the human body. To illustrate this, Kirlian research on long jet flights reveals that reduced magnetic fields on airliners often cause those who have a TCM condition called Yin deficiency to have severe exacerbations of symptoms such as anxiety, nausea, headaches, dizziness, and rising heat. Placing a magnet on the body during the flight provides a source for magnetic substance and reduces the symptoms considerably. Qi might be describable by physics as available magnetic activity within the body, though there appears to be much about magnetism that physics doesn't yet know.

Shen is described in traditional literature as mind or awareness, and often as light. If it is light, why can't we see it? Modern physics tells us that an electron is really an electromagnetic wave, an x-ray, high energy light which has been somehow captured by the nucleus of an atom. Thus, it is conceivable that a field of light or electromagnetic energy could resonate around our physical bodies and remain unmeasurable, just as electrons orbit around a nucleus undetected.

Shen may become visible when excited by electricity. Most people have Kirlian images which show small balls of light surrounding the finger during the negative pulse and streamers, lines of light perpendicular to the finger surface, during the positive pulse. However, healers and artists who use their hands with enhanced awareness, often have a smooth glow around their fingers. (Kirlian photos of the balls and streamers and smooth glow are shown on page 58.) Some even have certain fingers with a smooth glow while others show the balls and streamers. It appears that only those fingers used intelligently have the glow. For instance, a person who principally uses three fingers for massage will find that those three will have the glow while the others have the balls and streamers. The fact that electrons act differently around a healer's fingers is a big opportunity for physicists to investigate the nature of Shen, and to give us a better understanding of the special function abilities of Qigong masters and other healers.

CONCLUSION

Jing, Qi, and Shen are measurable with physical equipment, and thus are subject to scientific study. These measurements correlate with states of health and rates of healing. Thus monitoring the strength of these fields in hospitalized patients and supplementing them when low may prove of value in reducing recovery time and saving lives. In addition to studying, monitoring and supplementing these fields, it might also be productive to study how to increase a patient's resonance with his own morphogenetic field while decreasing his resonance with that of a disease process. This may be a principal area where Qigong masters and other healers excel, while western medicine has little knowledge. Jing, Qi, and Shen, long considered nothing more than useful metaphors by western scientists, may be very real substances which unlock the door for western scientists to understand human vitality.

THE ROLE OF CHAOS IN DISEASE AND HEALING

Richard H. Lee, China Healthways Institute

In the book, "Order Out of Chaos," Ilya Prigogine offers that, when a structured system is energized, its chaos increases, allowing it to reconfigure spontaneously into a higher order of organization. This may explain the evolutionary progression on Earth from the simple structures to the highly organized life forms, from the helpless child to the capable adult who chooses to serve a yet higher organization, humanity.

How can chaos be healing? Chaos is often thought of as unpredictable and destabilizing. It is the enemy of structure and overturns the law of the land. How is it possible that chaos can be harnessed to serve us? As an example: When someone's life is not working, when the basic order upon which they depend is failing them, and when they just want to run away, it may be time to "let go of life" in a way that nurtures the creation of a new order within which the world again makes sense.

Chaos is opportunity. Where there is rigid organization, there is little opportunity for restructuring. The best solutions often go ignored because their implementation would require destruction of some of the structure which holds an organization together. An organization which is flexible, receptive to new ideas, and free to adjust to changing circumstances, is willing to break down old structures for the sake of growth and efficiency. Structure provides for certainty. Chaos injects uncertainty. With a high degree of chaos, an organization can quickly adapt to strange and unfamiliar circumstances without being held back by the inertia of structure.

Within the human body, there is a great deal of structure. Physically we can have unnecessary chronic muscle tension, purposeless lumps, and other structures which don't serve us. Emotionally, we can hold onto desires which repeatedly carry us down pathways of misery. Mentally, we can so fill ourselves with conflicting thoughts that we are unable to sleep and on the edge of nervous breakdown. Whenever a structure within us becomes inflexible, we experience pain. Blocked desire provides emotional pain, and conflicting world views cause mental pain. Flexibility allows us to face the unexpected with capability and enthusiasm. We can run from the path of an approaching car, or stop to pick up a dime. We can eat a bowl of bean sprouts and tofu as comfortably as a hamburger, attend a religious service of any denomination without concern for the clothing of our creator, and be comfortable considering philosophies of strict child rearing alongside newsletters promoting chaos.

Random, unexpected events provide the chaos which exposes our rigidity, our unyielding habits, our crystallized beliefs. Sometimes these crystallized structures within us yield to the pressure but often they resist, sustaining the old ways despite conflicting forces and pain. And sometimes the strain even becomes so great that a person's entire world view shatters. This can be a hugely transformative event:

The initial stress is experienced as pain, uncertainty, or fear of the unknown. We become desperate to avoid further stressing of the crystallized structure so we hang on even tighter. Then the order governing our emotions and thoughts snaps, pops, or shatters. We burst into tears and watch as our life falls to shambles before our eyes, with seemingly nothing left for us to hang onto. We may feel an abrupt shift. Perhaps we experience an immediate release or an exhilarating feeling of freedom and relaxation. At other times, we find ourselves drowning in the unknown, trying to hang onto the structure which just crumbled, unwilling to relax and trust that a new better structure will form in its place. Sometimes we spend years trying to hang onto a crumbled, useless, and painful structure.

Chaos is an opportunity to restructure our beliefs, habits, and physical bodies into a more flexible arrangement, more suitable to the unknown but anticipated future. Whenever we face a new situation, we have this opportunity. Every time we make a new choice or break an old habit, we increase our flexibility and expand our opportunities.

How do we nurture chaos? We start by looking for patterns which restrict our choices, consume our resources, and limit our beliefs. Watching TV, smoking, recounting past experiences or gossip, consuming countless grams of animal fats and alcohol, spending endless hours or repetition in a job or relationship. Changing these patterns may be quite painful because of deep and extensive crystallization, but the result will almost certainly be free time, vitality, new experiences, and an opportunity to make less restricted new choices. While choices made in anger are easiest, and often create more chaos, choices made with love usually provide a greater opportunity for growth. If shattering and reorganization occur within the emotion of anger, the new order will embody anger. If we immediately forgive and "love our enemy", the new order within us will embody love and harmony.

THERAPEUTIC CHAOS

Richard H. Lee, China Healthways Institute

[This article describes how chaotic infrasonic sound might work to penetrate the body, break up stagnation and accelerate healing.]

Over the past ten years CHI has been studying the Infratonic QGM, trying to determine why it is effective at accelerating recovery and decreasing pain. Until recently, our focus has been on the broad spectrum frequency band of infrasonic sound produced by the QGM which infuses the local area, facilitating nerves and inducing Alpha brainwave activity, facilitating nerve transmission between the brain and the local site of pain. This is a plausible mechanism, but there appears to be more to it.

As we experimented chaos surfaced as an "active ingredient." We built prototypes with advanced microprocessor technology which produced an almost purely chaotic signal. In testing we found that these new units penetrated more deeply and stimulated resistant areas of the body which other therapies could not penetrate.

It appears that our thoughts and emotions are strongly affected by sound, and that the body defends itself against outside vibrational signals by anticipating and setting up signals which cancel them out. This may be why someone speaks loudly when he wants to persuade us of an idea, and why we may shout back to defend our own position, and why it is difficult to read or think clearly while sitting in a noisy park in a big city, or in a windstorm, or with loud music playing. It seems that it is difficult to sustain one's own complex thoughts when noise is jumbling them.

Many of the emotions and thoughts we hold within us can be the cause of chronic illness, pain, and slow healing. We might even be carrying some of these patterns from childhood, or pick them up from chance encounters or spooky movies. Like a computer virus, they are lurking among thoughts and beliefs which make up our personality. Those emotions and thoughts which are causing health problems are part of what Qigong therapists call "bad Qi."

Our ability to filter out recognizable sounds, while allowing us to maintain our own thoughts and emotions, is a major barrier to the effectiveness of therapy devices like TENS, ultrasound, and massagers. If we receive a therapeutic treatment from such a device, it may provide some relief but will only affect those surface vibrational patterns which are ready to let go. Thus many therapy devices relax muscles to some degree. But as soon as the therapy begins to intrude on dominant beliefs or emotions we automatically identify the invading signal and filter it out. Thus, predictable therapeutic signals can only do so much.

While we can defend against rhythmic signals, our old attitudes and emotions are defenseless against the random sound of a chaos therapy device because it is unpredictable. It simply goes past our vibrational defenses and softens all thoughts and emotions in the target frequency range. This may be why, after a treatment, people feel calm, clear thinking, and liberated from the problems and worries of life. A big part of what they are feeling may be the abundance of energy which is released when old unnecessary thought and emotion patterns dissolve, releasing the Qi, the substance that was vibrating to maintain them.

It is not just "bad Qi" that is softened by chaotic infrasonic sound. All structures, thoughts and emotions, may be softened as acoustical chaos increases in the body. Since accelerated healing of broken bones and sprained tendons is often reported, chaotic sound may soften these solid physical structures as well. This parallels the first step in Qigong therapy, emitting Qi to soften and liquefy the stagnant Qi.

The second step is to sweep "bad Qi" downward and out the feet. This may be a process of removing old vibrational structures which have been freed to move but have not entirely dissolved. Therapists have found that using chaotic infrasonic sound to sweep from the area of pain down the body and out the feet after treating the pain is often more effective than treating the point of pain alone.

The third step in Qigong therapy is to create a nurturing Qi field to allow the remaining Qi to harmonize, nurturing the emergence of a new order out of the chaos. The Infratonic QGM helps to do this by infusing the body with alpha EEG activity. Whereas the theta rhythm is associated with our individual identity and our relationship with the physical world (and our sense of ego and self importance), and the beta rhythm is associated with planning and thinking (stress and worry), alpha is all about relationships, putting the group before the individual. This may be why, after a QGM treatment, patients become more "compliant" in a doctor's office and people find that interpersonal conflicts are easier to resolve. This process may also occur on the cellular level toward healing and disease resolution as "bad Qi" is removed and the body's cells and organs can reorganize along the lines of teamwork and harmony.

Qigong therapy can be understood in terms of chaos theory. As we let go of old structure, we have an increase of chaotic or undefined energy, which then can transform into a higher order of organization which handles all previous requirements in a new way and has left over capacity for new endeavors.

IMPLICATIONS OF SCIENTIFIC QIGONG RESEARCH

The findings of the scientific research into human vitality have implications in broad areas of science and medicine.

1. Emitted Qi has strong measurable effects in health, healing, physiology and mental functioning. The presented research will encourage scientists around the world to seek deeper understanding of this phenomenon.
2. As human vitality becomes better understood, both in terms of measurable fields which surround the body, and chaos and consciousness, it will be integrated into the American health care system. Monitoring and treating patients accordingly will significantly improve recovery rates and shorten hospital stays.
3. Qigong exercises are effective tools for the individual to cultivate rapid healing and vitality. In addition, many who overcome chronic illness through Qigong practice will develop the ability to help treat others through the emitted Qi.
4. For those who are caught in the frenzied thinking of busy city life, practice of Qigong and other forms of concentration and meditation exercises is an answer. Qigong simulators help to get people started in their relax-

ation and Qigong exercises by relieving pain, promoting deep relaxation and mental clarity, and providing evidence that the phenomenon is real.

5. As the principles of human vitality are accepted in families, children will learn that their feelings of energy flowing through their bodies are normal, and that their awareness of these feelings is valuable. Just as children quickly learn the language that their parents speak, these children will learn to understand and use Qi if their parents accept and discuss it.

The Chinese scientific investigations into Qigong establish the phenomenon of emitted Qi as a real and fertile area of scientific study. Qigong teaching centers and Qigong simulation devices are already making a difference. Further studies into the nature of human vitality and inventions coming out of this ongoing research will do much to shape the 21st century.

APPENDIX 3

INTRODUCTION

This third Appendix includes research information on the physiological and energetic mechanisms triggered within the body, when stimulated through Medical Qigong therapy, Taijiquan or Yoga. This information was collected by my good friend Roger Jahnke, OMD, Director of the Qigong Department at the Santa Barbara College of Oriental Medicine, Santa Barbara, California. Mr. Jahnke is a doctor of Acupuncture and Oriental Medicine, and has studied extensively in China and abroad, learning Traditional Chinese Medicine and refining his clinical ability since 1967.

Doctor Jahnke is also the Director of the Health Action Clinic in Santa Barbara, California, and was one of the founding Board members of the National Qigong Association. He has written a Medical Qigong book entitled, "The Healer Within: Using Traditional Chinese Techniques to Release Your Body's Own Medicine."

This third Appendix is a continuation of the research devoted to the scientific approach of analyzing and scrutinizing the claims made as to the actual effectiveness of Medical Qigong Therapy as a clinical modality. The result on the effectiveness of Medical Qigong Therapy as a solid and effective clinical modality being positive in conclusion.

The following information in this third appendix represents but a fraction of Dr. Jahnke's findings, however, I am confident that the reader will find the data fascinating.

Dr. Jerry Alan Johnson May 2000

THE PSYCHOLOGICAL AND ENERGETIC MECHANISMS TRIGGERED IN THE HUMAN SYSTEM BY THE SELF-APPLIED HEALTH ENHANCEMENT METHODS QIGONG, TAIJI, AND YOGA

Roger Jahnke, OMD
Director of the Qigong Department
Santa Barbara College of Oriental Medicine

Western science has very clearly delineated numerous physiological systems: the digestive system, the nervous system, the cardiac system, etc. However, what might be called the "healing system" has not been discussed or clearly defined until recently. Medicine has uncovered many of the pathological processes that occur within the recognized systems and defined clinical strategies for their resolution. Strategies for enhancing the "healing system," however, have not been addressed by conventional science or medicine.

Numerous cultures, typically considered to be scientifically unsophisticated, actually noted and described the "healing system" long ago. The Chinese in traditional Chinese medicine, and the Asian Indians in Ayurvedic medicine, simultaneously developed elaborate theories of the function of the "healing system," and over thousands of years refined strategies for enhancing its function to produce a remarkable endogenous resource for self-repair. This internally generated healing resource is called an "inner elixir," literally a medicine produced within the human system.

In addition to specific clinical methods such as acupuncture, herbal therapies and tissue manipulation (massage and articular adjustment), these ancient cultures also created comprehensive health enhancement and self-healing systems to assist people in improving their own capacity to heal. Collectively such methods have been called the self-applied health enhancement methods (SAHEM). 1 Typically SAHEM includes breath practice, relaxation and regulation of the mind and nervous system through meditation, purposeful adjustment of the body posture with gentle movements and self-applied massage.

Qigong (also Chi Kung) from China and Yoga from India are ancient methodologies developed and refined over thousands of years specifically for sustaining and enhancing the "healing system." Qigong and Yoga, while they are historically based on the cultivation of factors unknown to the West (China—Qi and India—Prana), actually have specific effects on well known physiological mechanisms that are fully understood by Western science.

While there are many subtle aspects of human physiology that are triggered or accelerated by the practice of the SAHEM (including an emerging new understanding of "energetics," "bio-energetics" or "the electro-physiological system) the following three areas of conventional physiology make up a major portion of the physiological function of the "healing system" and are enhanced by the practice of the SAHEM.

1. Oxygen Delivery
2. Lymph Generation and Propulsion
3. Brain and Neurological Activity

The following sections will address each of these aspects. Other important mechanisms effected by the SAHEM including co-enzymes, antioxidants, connective tissue, precursor hierarchies, stem cell morphology , etc. do not fall into the scope of this paper and their discussion must be held for the future. The domain of subtle energy based mechanisms for the SAHEM effect will be explored briefly in conclusion.

OXYGEN

The Chinese knew about the circulation of the blood approximately 2000 years before William

Harvey described it in 1616 (Temple and Needham, 1986; Veith, 1972)). They knew about the energy generating relationship of food and air 2300 years before the elaboration of the oxidation feature of the Krebs cycle. The simplicity of the ancient Chinese formula for metabolic activity, understandable even by children, encourages wise use breath and food as health enhancing factors This is how Chinese grandmothers have passed health wisdom to children. Note that such simple wisdom is uncommon in the grandmothers of scientific cultures today.

It is often asked what difference there is between Qigong, Yoga (or other mild SAHEM) and the more vigorous conventional fitness methods like running, lifting weights, power biking and aerobic exercise. One way to discuss this is to trace the oxidation and energy cycle. In both aggressive, vigorous exercise and mild exercise the body naturally produces a powerful mix of metabolic resources. In vigorous exercise this is typically called a fuel. In the mild fitness systems of ancient cultures this inner resource has historically been called a medicine, or an elixir.

In vigorous exercise this oxygen induced resource is spent as fuel for hungry muscles. In the milder forms of SAHEM this resource is not completely spent, instead it is circulated internally and utilized as an internal reserve of self-repair factors to sustain and heal the tissues, organs and glands. At the advanced levels of Qigong this endogenous medicine is believed to be transformed into an elixir of longevity and wisdom.

Increased oxygen availability from practice of SAHEM has three potential effects, it:

1. supports energy (ATP, AMP, ADP) generation.
2. generates water as a by product of energy metabolism.
3. enhances immune function.

ENERGY GENERATION

The energy necessary for cellular processes and body heat regulation is supplied through the reaction of oxygen and glucose, in the presence of adenosine triphosphate (ATP). The combination of oxygen from air and sugar from food is catalyzed by ATP releasing chemical energy (ergs) which accomplishes all cellular processes. ATP is the basic energy storage and energy transfer molecule in the body.

The oxygen carrying capacity of the blood and the energy catalyzing ability of ATP are critical factors in the ability of the human organism to sustain a high level of vitality. This suggests that oxygen and ATP may be related to what the Chinese call "Qi" and what the Asian Indians call "Prana," the vital force or life energy. Chinese medicine distinguishes several aspects of Qi. Zhen Qi (Genuine, True or Normal Energy) is the basic dynamic force of all vital function. Ying Qi (Nutrient Energy) is the dynamic force which fuels cellular work (Kaptchuk, 1983). The interaction of oxygen, nutrients and ATP is apparently a physiological process that is a direct correlate of one particular aspect (there are many) of the activity of the Qi.

In Chinese research it was observed that blood levels of ATP increase with practices which cultivate Qi (Wang, 1988). When the mysterious practice of Qi emission (projecting Qi) was performed, the practitioner's ATP levels decreased significantly. Emission of Qi results in a decrease in stored ATP (or potential energy) as recognized by Western physiological science.

The SAHEM activate the body's energy metabolism cycle. The Chinese call this the "cycle of the transformation of energy" (Jahnke, 1989). In the West it is called the Krebs cycle, and has been nicknamed the "bioenergetic cycle."

In the West, ancient medicine is generally characterized as unscientific, superstitious or even primitive. Therefore, the Chinese "formula" for the transformation of energy (Qi) seems overly simplified:

Gu Qi + Kong (Qing) Qi = Zhen Qi

or energy of food + energy of air = essential body energy

Gu Qi (grain Qi, or food Qi) is the essence or life force of food. It mixes with Kong Qi (Kaptchuck, 1983) or Qing Qi (Kendall, 1989) (natural air Qi), the essence or life force of air to form Zhen Qi (Genuine, True or Normal Qi)) which is the life force of the body (Veith, 1972;

Kaptchuck, 1983; Kendall, 1989).

However, it is this same basic formula, hidden in the complex vocabulary of Western science, that is used in modern physiology:

$6O_2 + C_6H_{12}O_6 = ATP = ergs + 6CO_2 + 6H_2O$
air + glucose = catalyst = energy + carbon dioxide + water

It seems that the Chinese knew, without a particularly refined scientific method, that only a portion of the air and food, the Qi or essence, was employed in the process. In fact only 21% of air is oxygen and glucose is approximately 60% of food. (Guyton, 1992)

Science has been exploring energies, beyond metabolic energy (ergs), within the human system including the migration of ions, the ionic gates in cell membranes, the ionic discharge along neurological pathways and the electrical discharge that causes the beating of the heart. The overall understanding of body energetics will increase significantly as science explores the extent to which Qi is linked to these known physio-energetic mechanisms. The interaction of oxygen, glucose and ATP is likely only one energetic aspect of the beneficial effects of Qigong and other SAHEM.

WATER PRODUCTION

A second critical benefit of increased oxygen metabolism is linked to the cellular production of water as a by-product of the energy production cycle noted above. (Shields, 1989) This water is incorporated in interstitial fluid, lymph, blood serum, cerebrospinal fluid, synovial fluid, tears, sweat, etc.

Coincidentally, the energy which fuels the work of the cells as well as much of the water which contributes to the lymphatic fluid and other body fluids have the same source: metabolism. Metabolism is gently increased by all of the SAHEM. In addition, while the metabolic production of ergs also produces water, it is also interstitial water that has been found to conduct the physiologic flow of ions internally.

The benefits of the practice of health enhancement and self-healing methods upon the water component of the endogenous self-repair resource will be further explored in the section below on the lymphatic system. The bionic conductance will be explored briefly under energetic mechanisms.

IMMUNE FUNCTION

ATP drives the activity of every cell. Immune function as well as the production of immune resources (white blood cells, lymphocytes, T-cells, killer cells, etc.) are dependent on oxygen availability. It has been shown that exercise can mobilize the effect of natural killer (NK) cells. (Pedersen, 1988) However, in individuals who exercise so vigorously that they exceed the aerobic level and cross the anaerobic threshold, immune function is actually decreased (Brahmi, 1985; Fehr, 1989).

The scientific evidence is strong on two points: 1) oxygen deficiency leads to decreased immune function and 2) moderate amounts of mild exercise increase immune function.

In his research, Nobel Prize recipient Otto Warburg found that oxygen deficiency was typically supportive of cancer cell proliferation (Warburg, 1966). There are numerous studies that associate reduced lung volume and oxygen exchange capacity with increased mortality and reduced resistance to disease (Gordon, 1970; Cullen, 1983). In studies with elders, immunodeficiency was found to be one consequence of reduced oxygen metabolism. (Saltzman, 1987).

Oxygen's effect on immune function was demonstrated in a German study. It was found that in elderly injured, stressed and hospitalized individuals the arterial oxygen content is often reduced from normal levels. Administration of oxygen was found to elevate the arterial oxygen content and increase recovery rates. Interestingly, the experimental addition of germanium to the treatment protocol in the study further accelerated the healing process (Liem, 1985). Germanium is known to accelerate oxygen metabolism.

These mechanisms associated with oxygen's role in the endogenous "healing system," are naturally occurring and optimal in healthy individuals. One or several of these mechanisms may be deficient in individuals who are ill or at risk for disease. These mechanisms are all enhanced by the gentle movement, focused breath practice and relaxation that occur in the practice of Qigong, Yoga and other SAHEM.

LYMPH SYSTEM FUNCTION

The lymphatic system has been largely neglected in the conventional Western scientific tradition. Contrasted with cardiac function, for example, the lymph function is relatively unexplored. Perhaps, because lymph and lymph vessels are generally translucent, they drew little attention in early anatomical study compared to organs, blood vessels, muscles and bones.

Hippocrates and Aristotle referred to "white blood" and "colorless fluid." In the Middle Ages medical knowledge declined and the lymph was temporarily forgotten. Not until 1627 did Asellius, in Milan, recover the knowledge of the lymph (Rusznyak, 1976). The structure and action of the lymph system was still undefined by 1900 (Yoffey, 1970) and both the immunological function of the lymph and the actual lymphogenic process are still not clearly understood even as we enter the 21st century.

However, in Asia, medical practice is based on meticulous observation of whole organisms—the whole person and the totality of integrated functions. Even though the discrete lymphatic function itself was unknown and unnamed, its effects were generally ascribed to the proper action of the Qi and fluids in China or the Prana and fluids in India. Even though the lymph, as such, was unnamed by Asian physicians, elaborate methods for generating and circulating lymph were developed, and faithfully practiced through Qigong, Yoga and other self-care practices.

When these practices are viewed in relation to what is now known about lymphatic function and its healing role, it appears that much of Qigong, Tai Chi, Yoga and Pranayama were developed specifically with the enhancement of lymphatic function in mind. Breath, movement and postural modification as well as meditation have specific effects on the circulation of the lymph and therefore enhance the overall function of the lymphatic system.

In the West science has divided the body fluids into specific categories: blood, lymph, cerebrospinal fluid, synovial fluid, extracellular fluid, intracellular fluid. From the paradigm of the West it seems that the unsophisticated Chinese may have overlooked important information with their broad, non-specific view of "Qi, blood and body fluids." However, the lymph fluid is largely interstitial biological water. This becomes an aspect of blood plasma (Rusznyak, 1976). Plasma and lymph both contribute water to the cerebrospinal fluid, part of which migrates back into the lymphatic system (Bradbury, 1985). This cycling of fluid (essentially water) from one system into another validates the Chinese approach of thinking of the all various fluids in a single continuum: simply as "body fluids."

Recent discoveries have demonstrated that electrical energy in the form of ion streams migrate throughout the body (Nordenstrom, 1983). The conductance medium for this energy flow system (which will be discussed in the energetic section of this article) is biological water in its numerous contexts including lymph, blood serum, interstitial fluid, etc.

The practice of SAHEM activates a number of remarkable self-healing mechanisms associated with the lymphatic system:

1. Lymph generation
2. Lymph propulsion
3. Immune function
4. Cerebrospinal fluid circulation
5. Nutritive function

LYMPH GENERATION

The actual generation of the lymph has long been attributed to the filtration of blood plasma from the capillaries (Adair and Guyton, 1985). As recently as 1985 this was recognized as the primary source of the lymph. This idea overlooks a significant detail that is at the crux of the tremendous health benefits of Qigong and Yoga: a significant portion of the water in the body's fluids is produced by the identical physiological process that generates the body's biological energy as discussed in the previous section on oxygen (Shields, 1989). Just as an individual respires to sustain life, each cell is also respiring. For each gram of glucose metabolized, in excess of a gram of water is produced.

In a moderately active 70 Kg human, between 2100cc and 2800cc of lymph enters the blood stream daily at the sub-clavian vein through the thoracic duct. In a resting individual the calculations of the Krebs cycle show that the cells produce approximately 950cc (Shields, 1989) of this daily water flow. In a person engaging in moderate exercises, such as Qigong or Yoga (or walking), up to 1400cc of aerobically generated water can be produced daily. When an individual is practicing some sort of moderate body activity, as much as one half of the water that is produced and propelled through the lymph system is a by-product of cellular metabolism. Increased lymph flow improves the elimination of the by-products of metabolism and pathogenic factors (both considered toxins) and increases the circulation of immune factors.

LYMPH PROPULSION

The blood's circulatory system has the powerful heart muscle to propel its fluid. The lymph, however has no distinct heart in humans. It was understood that the movement of lymph against gravity is accomplished by the contractions of the skeletal musculature through vessels with one way valves in a uni-directional system. Nothing new emerged until the mid 1900's when studies revealed that birds and reptiles have specific lymph hearts (Shields, 1980). Even as late as 1941 important aspects of lymph propulsion mechanism remained obscure (Drinker and Yoffey, 1941).

By 1949 "spontaneous, intrinsic pulsatory contraction of the peripheral lymphatic vessels" was demonstrated in humans with a rhythm unassociated with either the heart or the breath (Smith, 1949; Olszewski, 1979).

The current literature is crowded with a wide range of questions raised by research specifically on lymph propulsion: What factors might stimulate the intrinsic contractile mechanism? Is it consistent throughout or differentiated, by what regional factors? By what autonomic and neuroendocrine factors is intrinsic contractility modified? These are just a few of the questions. The study of SAHEM will accelerate our ability to answer these and other more general questions on the lymph system such as: What regulates lymph protein concentration? What effect does passage through the lymph nodes have on the proliferation of immune cells from within the nodes (Olszewski, 1985)? And what is the nutritive role of the lymph (Shields, 1972)?

The breath, through two mechanisms, has a significant effect on the propulsion of the lymph: 1). aerobic production of water and 2). mechanical pumping of the breath apparatus, lungs and diaphragm. Several additional propulsive mechanisms are initiated by body movement and body posture.

Tentative agreement now exists on at least five mechanisms which comprise the "lymph heart" and are activated by the SAHEM, including:

1. aerobic production
2. intrinsic smooth muscle contraction
3. contraction of striated skeletal muscles
4. gravity
5. breath apparatus: lungs, rib cage and diaphragm

Aerobic Propulsion. The aerobic production of water contributes to lymph propulsion mechanically. The cellular production of water, as a by-product of oxygen metabolism, is increased by dynamic Qigong and Yoga practice. The liquid holding capacity of the tissue spaces is naturally limited. As the limit is reached, the presence of additional lymph drives the excess into the terminal lymphatic vessels, the initial vascular openings of the lymphatic system (Yoffey, 1970: Adair, 1985; Shields, 1980, Olszewski, 1985). Like a cup that is running over, the interstitial space fills, building volume and pressure, and then flows over into the lymph vessels.

Intrinsic Smooth Muscle Contraction. The automatic response of the smooth muscle tissue of the lymphatic vessels and the peripheral lymphatic capillaries is to contract when filled and stretched to a certain tolerance (Olszewski, 1985). This contraction moves the lymph forward, with the assistance of the one way valves, in much the same way as the contraction of the heart moves the blood (Bradbury, 1985). In Qigong and Yoga this mechanism is accelerated by the increase in metabolism from gentle movement which contributes to lymph volume.

A number of studies have demonstrated that the intrinsic mechanism is also triggered when the overall sympathetic aspect of the autonomic nervous system is modified by the parasympathetic. Therefore, he intrinsic contractility mechanism is accelerated or enhanced by the shifting into the deeply relaxed state that is a primary feature of SAHEM practice.

Striated Skeletal Muscle Contraction. Even slight contraction of skeletal muscles in mild health enhancement practices propels the lymph in the one way vessels. In the more dynamic practices the pumping of the skeletal muscles and compression of the body parts increases interstitial. The effect of muscle contraction on lymph is one of the classic explanations for lymph motion (Adair, 1985, Bradbury, 1985). In Qigong and Yoga this mechanism is triggered by the mechanical action of the musculature and the action of the skeleton through the relaxation and contraction of the striated muscles in the movement forms of the practice.

Gravitational Propulsion. Because gravity exerts such a substantial force and because lymph has so far to climb to get to the thoracic duct's entry into the sub-clavian vein (at the base of the neck), any inversion of the limbs or even the prone body position allows for a more free flow of lymph. In keeping with this, elevation of the limbs is often prescribed for health problems characterized by a pooling of interstitial fluids, edema.

In SAHEM, many postures and movements create this mechanical dynamic where the lymph is actually propelled by gravity. In certain walking forms, Tai Chi and Guo Lin Qigong are the most renowned, the practitioner is constantly but slowly moving all of the limbs in beautiful circular motions that recurrently activate this mechanism. In Yoga many of the asanas (postures) invert the limbs. In the head and shoulder stands the whole body is inverted.

Breath Apparatus: Mechanical Propulsion. The most powerful of the array of mechanisms that work together to form the "lymph heart" is the mechanical action of the breathing apparatus itself (Sheilds, 1980). The concentration of lymphoid tissue above and below the diaphragm is many times more dense, and contains greater fluid volume, than any of the lymphoid tissue at the periphery, or even in the moderately prolific lymphoid areas of the axilla or groin (Adair, 1985). Lymph that has been carried from all over the body, accumulates centrally and is then propelled by the breath / diaphragm in a final rush through the thoracic lymph duct into the blood at the subclavian vein where it leaves behind its identity as lymph and is transformed into blood serum (Shields, 1981).

Above the diaphragm the thoracic duct of the lymphatic system is a central collecting vessel. Its size is many times that of the peripheral lymph vessels. Below the diaphragm a substantial dilation of the thoracic duct forms a collecting capsule for lymph, called the cisterna chyli. Chyle is a milky fluid of nutrients absorbed from the small intestine, which is passed into the circulating blood through the thoracic duct.

When full inspiration of the breath occurs, the diaphragm drops downward. A tremendous pressure is exerted on the cisterna chyli and a great negative pressure is generated in the thoracic cavity. The compressed cisterna chyli shoots lymph upward through the thoracic cavity. As air rushes in to balance the negative thoracic pressure, the lungs are fully expanded. This creates a powerful positive pressure which compresses the thoracic duct. Due to the one way lymphatic valvular system, lymph is forced upward where it enters into the sub-clavian vein with a rush.

Simultaneously, when the diaphragm drops downward on full inspiration it also compresses the abdominal and pelvic organs. The contents of the lymphoid reservoirs and vessels are forced by the same arrangement of one way valves forward in the system. In research with moving X-ray film, study subjects explored various actions and breath patterns. It was demonstrated that deep inspiration pumps the lymph at a rate that is dramatically greater than resting inspiration and other activities, including massage and walking (Shields, 1981).

It is fascinating to note that in Traditional Chinese Medical theory it has been taught for several millennia that the "lungs regulate the water passages" (Xie Zhu Fan, 1988). To Westerners this seems quite strange. The kidneys are considered to be responsible for water regulation. How could the Chinese have known that the activity of the lungs and diaphragm were linked to moving the body's internal water supply? Even now as we enter the 21st century Western science, and especially average Western citizens, have little idea that the simple act of regulating the breath is such a profound healing tool. However, here we can see that the Chinese somehow arrived at a kind of remarkable wisdom about how the lungs, the breath and the oxygen literally cooperate to produce and circulate the healing resource of water in the body.

IMMUNE FUNCTION

The immunoactive aspect of the lymphatic system is well represented in the conventional literature of Western physiology (Drinker, 1941; Bradbury, 1985; Olszewski, 1985; Van Rooiijen, 1987). The bone marrow, thymus, spleen and lymph nodes participate in the interaction of the lymph and immunity. The composition of the lymph fluid itself includes a number of immune active agents such as lymphocytes and macrophages (Olszewski, 1985).

Lymphocytes that exit with the lymph fluid from the nodes come from three sources: 1) inflowing with lymph from the tissues in the peripheral vessels, 2) exchanged from the blood that enters the node's own vascular system and 3) formed by local proliferation in the node itself (Bradbury, 1985). Lymphocytes naturally collect within the node, especially when lymph flow is sluggish. Greater numbers of cells proliferate when lymph flow is greater and the numbers circulated out of the node increase with flow volume as well (Adair, 1985). Both the numbers immunoactive cells and their circulation rate are increased by the deepening of the breath, relaxation and the alternating contraction and releasing of the musculature in the gentle practices of Qigong and Yoga.

Excellent recent research has clearly delineated, localized and quantified the development of specific antibody forming cells in lymph nodes (Van Rooijen, 1987). In addition, it has been found that there are neurotransmitter receptor sites on lymphocytes where they actually interface with the neuropeptides that drive immune function (Smith, 1985). This is one of the important links between neurochemistry and immunity that is mediated through the lymph system and optimized by the practice of the SAHEM which also increases the numbers of circulating immune cells and accelerates their rate of delivery.

CEREBROSPINAL FLUID (CSF)

The cerebrospinal fluid (CSF) system was classically perceived as a closed system, though another view, unconventional for many years, held that CSF was actually circulating in an open system that allowed the fluid to flow through the aracnoid villi and into the venous blood. By the 1970's it was generally acknowledged that the CSF travels along the cranial and spinal nerves and into the perineural lymphatics (Kimber, 1977).

Research using the microinjection of tracers has suggested several possible pathways for the passage of both the CSF and the cerebral interstitial fluid (CIF) to exit the aracnoid space (Bradbury, 1985). By 1985 the flow of CSF and CIF into the lymphatics was well documented (Bradbury, 1985). Consideration has even been given to the effects of pressure and posture on this flow (Bradbury, 1985), both of which are primary effects that are enhanced in the practice of Qigong, Taiji and Yoga.

The presence of CSF in the lymphatic system and the presence of neurotransmitter receptors on immune cells (Covelli, 1988) suggests a powerful benefit of the SAHEM on the interaction between neurotransmitters and immune function in the reticulo-endothelial system.

NUTRITIVE (TROPHIC) FUNCTION

The importance of a broad availability of nu-speech, irrational or catatonic behavior, such as stupor, rigidity, or flaccid movement of the limbs. The ability to interact with others is greatly impaired.

Sclera - a tough white fibrous tissue that covers the white of the eyes.

Sea of Blood (Sea of the Twelve Channels) - pertaining to the Energy located in the Thrusting Vessel.

Sea of Energy - Energy located in the Lower Dantian, or Qi Hai area.

Sea of Grain and Water (Sea of Nourishment) - pertaining to the Energy located in the Stomach.

Sea of Marrow - pertaining to the Energy flowing in the spinal column and brain, originating from the Kidneys.

Sea of Qi - the chest center. Some Medical Qigong schools maintain that there are two reservoirs of Qi: the Middle Dantian, being the Sea of Postnatal Qi, and the Lower Dantian, being the Sea of Prenatal Qi (which is regulated by the Qihai CV-6 point).

Sea of Yang Channels - pertaining to the Governing Vessel.

Sea of Yin Channels - pertaining to the Conception Vessel.

Self Regulation Therapy - pertaining to the patient's Qigong prescriptions (meditations and/or exercises).

Seven Emotions - see Seven Internal Factors.

Seven Essential Stars - the Sun, Moon, Mars, Venus, Mercury, Saturn and Jupiter, associated with the body's seven orifices.

Seven Internal Factors - pertaining to the seven emotional pathogenic factors that cause disease, when in an Excess condition (Joy, Sorrow, Worry, Grief, Fear, Fright, and Anger).

Seven Material Souls - pertaining to the seven Earthly spirits that reside in the body as the Po.

Seven Orifices - ears, eyes, nostrils, mouth, anus and urethra, which are considered the gates and windows of Essence, Energy and Spirit.

Seven Turbid Demon Natures - see Po.

Shaman - an ancient Tungus term meaning "between the worlds." A Shaman is a tribal priest or priestess who heals the physical, mental, emotional, energetic and spiritual aspects of the patient.

Shao Yang - Small Yang, also called Lesser Yang, Minor Yang, or Young Yang, is affiliated with the sunrise and the waxing-moon phase. Modern physicists associate the Lesser Yang with a light force and electromagnetism.

Shao Yin - Small Yin, also called Lesser Yin, Minor Yin, or Young Yin, is affiliated with the sunset and the waning-moon phase. Modern physicists associate the Lesser Yin with a heavy force, and gravity.

Shen - meaning Spirit; when speaking about physical development, it is derived from Qi, and can be divided into both Prenatal and Postnatal Shen.

Shen Deviations - mental and emotional disorders which have caused the Three Ethereal Souls (Hun) to leave the patient's body.

Shengong - training of the spirit through meditation and visualization.

Shening Out - terminology used to describe the Ethereal Soul (Hun) wandering away from the body.

Shi Qi (Turbid Qi) - also known as Evil Qi, Toxic Qi, and Pathogenic Qi, it is coarse, unrefined, polluted or dirty energy.

Sishencong (Four Alert Spirit) Points - four points at the top of the head (that surround the Baihui Point) used to absorb Heavenly Qi into the body.

Shou Zhen (Hand Diagnosis) - a form of diagnosis, wherein, the doctor assess the "energetic blueprint" of the patient's body transformed onto the doctor's left hand.

Shu Points - five specific points below the elbows and knees identified as the Well, Spring, Stream, River and Sea points. Each point has an affect on the quantity of the energy of an organ.

Sishencong (Four Spirit Hearings) Points - a group of four points located at the top of the head (surrounding the Baihui point), used for absorbing Heavenly Qi into the body's Taiji Pole.

Six External Factors - also known as the Six Pernicious Influences, these factors pertain to the six climatic changes (Wind, Summer Heat, Heat, Damp, Dryness, Cold, and Fire).

Six Storage Areas- body's Yang organs constantly fill and empty, and include the Bladder, Gall Bladder, Stomach, Large Intestine, Small Intestine, and Triple Burners.

Skin Zones - twelve dermal-zones, based upon the surface location of the body's Twelve Primary Channels.

Soaring Dragon Technique - hand technique for Qi emission, where the energy is emitted through the middle finger bent and pointing downward, while the other fingers are extended straight outwards.

Solid Organs - the body's Yin organs, which include the Liver, Heart, Spleen, Lungs, and Kidneys (also included in this list is the Pericardium).

Soul - immaterial Spiritual Essence of an individual's life, stored within the Heart and Middle Dantian.

Soul Body- see Astral Body.

Soul Extensions - the Shen develops and contains Twelve Soul Extensions. These Twelve Soul Extensions contain the body's different personality characteristics.

Soul Loss - the loss of parts of the Eternal Soul.

Soul Retrieval - to spiritually search for and bring back one's forgotten memories (soul), which have been isolated from consciousness due to trauma and shock.

Soul Travel - see Astral Travel.

Sound Energy Therapy - sound projected as audible and inaudible tone resonation, used for healing.

Sound Resonation - healing tones used for tonifying or dispersing the patient's Energy.

Spider Nevus - a branched growth of dilated capillaries on the skin, that resemble a spider.

Spinal Pass (Jia Ji Guan) - two points located on the lateral sides of the Mingmen (GV-4), where energy has a potential to stagnate.

Spinous Process - the single midline posterior projection arising at the junction of each vertebra.

Spiraling Energy Technique - hand manipulation, that extends and spirals the doctor's projected energy.

Spirit - the energetic manifestation of the Eternal Soul.

Spirit Body - the energetic vehicle inwhich the body's Shen can travel throughout the Astral Plane. The Spirit Body can manifest through many forms (Body of Light, animal forms, etc.).

Spirit Demons - see Demon Possession and Oppression.

Spirit Soul - the Three Ethereal Souls, accompanied by the individual's consciousness, acting as one unit for spirit travel.

Spirit Travel - the spirit (Hun) journeying outside of the physical body.

Splenomegaly - the enlargement of the Spleen.

Squamous Metaplasia - the conversion of tissue into a form of scalelike cells, that is abnormal for that tissue.

Stacking the Bones - allowing the bones to stack upon each other from the bottom of the feet to the top of the head.

Stagnation - not moving, inactive; pertaining to Qi, Blood, or thought patterns.

Static Qigong - the process of stationary, quiescent Energy gathering.

Stroke (Wind Stroke) - caused by the buildup of Excess Liver Fire creating Internal Wind.

This Internal Wind causes Qi and Blood to rebel upwards causing Phlegm to form and obstruct the cavities and vessels, creating Penetrating Wind or Stroke.

Subarachnoid Hemorrhage - bleeding internally, within the spaces at the base of the brain, between the pia proper and arachnoid contain the cerebrospinal fluid.

Subconscious Mind - part of the mind associated with the recording and storing of personal interpretations of reality (not readily accessible to the conscious mind).

Sublimation - the channeling of unacceptable impulses into acceptable, refined social forms and is the only defence mechanism considered to be a healthy reaction.

Substances - pertaining to the body's essential parts of physical and energetic material.

Sui - Marrow.

Super Ego - Dr. Sigmund Freud's terminology for the division of the psyche in psychoanalytic theory, responsible for the psychic reward and punishment system.

Sun's Essence - energy gathered from the sun, ingested as warm light.

Sword Fingers Technique - hand manipulation that emits Qi through the extended index and middle fingers.

Symptoms - a subjective manifestation of a pathological condition, reported by the patient.

Syndromes - a grouping of signs and symptoms, based on their frequent reoccurrence, that may suggest a common underlying pathogenesis.

Systemic - affecting the entire body.

T

Taiji Pole - the Center Core of light which joins the body's three Dantians and the Eternal Soul together.

Tai Yang - Great Yang, also called Strong Yang, Major Yang, or Old Yang, is affiliated with high noon and the full-moon phase. Modern physicists associate the Strong Yang with a strong nuclear force.

Tai Yi - meaning Great Divinity or God.

Tai Yin - Great Yin, also called Strong Yin, Major Yin, or Old Yin, is affiliated with midnight and the new-moon phase. Modern physicists associate the Great Yin with a weak nuclear force.

Ten Heavenly Stems - the ten energies of Heaven that rule the changes of the Five Elemental seasonal transitions, and are represented in the human body as the Yin and Yang aspect of the Five Elements (represented in the human body as the ten major internal organs).

Ten Thousand Voices - pertaining to the state of open receptivity of the Qigong doctor's Heart, after rooting the mind.

Ten Wings - consisting of ten commentaries from Confucius and his disciples, pertaining to the study of the eight trigrams, sixty-four hexagrams, and the Yi-Jing.

Third Eye Point (Yin Tang) - located in the center of the forehead, responsible for spiritual intuition and communication.

Thought-forms - images of concentrated thought patterns that manifest on the vibrational resonance of the Astral Plane.

Three Ethereal Souls (Hun) - pertaining to the three heavenly spirits that reside in the body.

Three Fires - the heat in the body, generated from the energy radiating from the Heart Fire, Kidney Fire, and Bladder Fire.

Three Outer Forces - pertaining to the three natural powers of Heaven, Earth and Man.

Three Parts Wisdom - knowledge obtained through the doctor's connection and communication with his or her Upper, Middle, and Lower Dantians.

Three Periods of Life - the developmental stages of the patient's Jing, Qi and Shen divided into the womb, childhood and adulthood.

Three Stars - pertaining to the three periods of life, each period is divided into three stages of development, known as the three stars.

Three Treasures of Earth - pertaining to the energy of Soil, Water and Wind, and the study of Feng Shui (Wind and Water).

Three Treasures of Heaven - pertaining to the energy of the sun, moon and stars, and the study of Chinese astrology.

Three Treasures of Man - pertaining to the energetic interaction of the body's Essence, Energy and Spirit, and the study of the Yi-Jing (I-Ching or Book of Changes).

Three Wonders - Clinical manifestations of Qi, categorized as Subtle, Mysterious and Incredible Wonders.

Thrombosis - the formation and development or existence of a Blood clot (thrombus) within the walls of the vascular system.

Thrusting Channels - the Five Energy Channels which surround and penetrate the body's center core via the Taiji Pole.

Thrusting Vessels - also called the Chong Vessels, they are the Five Energy Vessels which originate from the center of the body and internally transverse the legs and torso.

Ti - referred to as the Divine Center.

Tian Qi (Heavenly Energy) - the transformed energy of the Yuan Qi and the divine.

Tian Shen (Heavenly Spirit) - the transformed energy of the Yuan Shen and the divine.

Tie Bi (Iron Wall) - the areas of the body where it is most difficult for the energy to pass through when circulating the Microcosmic Orbit.

Tinnitus - a ringing, tinkling, or buzzing sound in the ear.

Toe Raised Stepping - pertaining to energetic walking therapy, wherein the toes are stretched when stepping in order to facilitate the increase of Qi flowing into the body via the leg channels.

Tonification (Tonify) - to supplement the insufficiency and strengthen the body's resistance.

Traditional Chinese Medicine - Chinese Energetic Medicine, divided into four branches of healing modalities (Acupuncture, Herbal Therapy, Medical Qigong Therapy, and Tissue Regulation Therapy (Chinese Massage).

Transference - the process whereby a patient unconsciously transfers feelings, thoughts, beliefs and patterns of behavior that had been previously experienced with others onto the doctor.

Transient Ischemic Attacks (TIA) - temporary interference with the Blood supply to the brain.

Treatment - the medical care given to a specific condition.

Trigger Points (Ashi Points) - places on the body which are tender spots, or painful areas near diseased or injured tissue.

Trigram - pertains to three Yao lines stacked upon one another forming a specific symbol, which represents certain characteristics.

Triple Burners (San Jiao) - also known as the Triple Heaters and Triple Warmers, they correspond to three main body cavities (perineum to navel, navel to base of solar plexus, solar plexus to throat), and are responsible for heating the body and transporting Body Fluids.

True Fire - the original Heat or Fire Energy that regulates the body's Yin and Yang Qi, created from the radiating energy of the Heart's Fire, Kidneys' Fire and Bladder's Fire.

True Nature - one's innate nature in harmony with life.

True Qi - the energy that circulates in the body's channels and collaterals which nourishes the Yin and Yang organs and fights disease.

True Self - one's true nature, connected to the subconscious mind.

True Spirit - pertains to the spiritual nature of the True Self. The Hun and Po are expressions of the body's True Spirit.

Tsou Hou Ru Mo ("the Spirit leaves and the Demon enters") - describes self induced psychosis, pertaining to improper Qigong training, wherein the patient's Hun leave the body and the Po take over.

Tui Na Therapy - a tissue manipulation therapy that focuses on the adjustment and / or stimulation of the muscles and tendons.

Tumor - an abnormal growth, either benign or malignant, caused by a retention of mass due to the cultivation and stasis of Qi, Blood, Phlegm, etc.

Turbid Qi - also called Evil Qi, is coarse, unrefined, polluted, and dirty energy.

Twelve Pi Hexagrams - the twelve symbols pertaining to the twelve time periods of the day and year.

Twelve Primary Channels - the body's twelve main energetic rivers (Liver, Lungs, Large Intestine, Stomach, Spleen, Heart, Small Intestine, Bladder, Kidneys, Pericardium, Triple Burners, and Gall Bladder).

Twelve Earthly Branches - twelve energies of the Earth that determine the six Qi factors of the seasonal transitions (represented in the human body as the Twelve Primary Channels).

Two Breathings - pertaining to the abdominal breathing method of holding the breath.

U

Umbilications - a depression resembling a navel.

Universal Qi - energy pertaining to the Heavens, the divine and the celestial influences.

Upper Burner - pertaining to the body's complex system of Fluid distribution via the Lungs and located within the upper chest cavity.

Upper Dantian - area within the center of the head, attributed as the body's chamber of light and door to psychic and intuitive powers. The Upper Dantian is also known as Seal Palace, Ancestral Opening, Calm Fountain, Heaven's Valley, Inner Source, and Clay Pill Palace.

V

Vasculitis - the inflammation of a blood or lymph vessel.

Vertigo - the sensation of moving in space, resulting in such symptoms as dizziness and light-headedness.

Virtue (De) - pertaining to the function of the divine in man.

Virtue of Dao - pertaining to the commendable quality of the divine.

Viscera - the body's internal organs.

Void - also called Wuji, it pertains to the infinite space between matter and energy.

W

Wai Dan Shu - external elixir cultivation, that focuses on cultivating Qi from outside the individual's body.

Wai Qi - external, extended energy.

Walking Therapy - Postoral Dao Yin walking exercises and dynamic "moving" meditations used for the treatment of organ Deficiencies.

Wandering Bi - migrating pain within the body's cavities.

Waning - to grow smaller.

Water Element - one of the Five Elements, pertaining to Kidneys and Bladder.

Water Jing - energy that controls the genetic development phase of the fourth fetal month.

Waxing - to grow larger.

Wei Lu Guan (Coccyx Pass) - located on the lowest segment of the spine just posterior to the anus, near the Chang Qiang (GV-1) point.

Wei Qi - the body's external field of Defensive and Protective energy (divided into three fields of Qi).

Wen Huo - pertaining to the gentle breathing method of Respiratory Dao Yin training.

White Blood Cell - any of a group of Blood cells that have no hemoglobin and migrate into tissues to fight infection and digest cell debris.

Wind Bi - pain in the body created by toxic Wind invasion.

Wind Stroke - Stroke caused by the buildup of Excess Liver Fire creating Internal Wind. This Internal Wind causes Qi and Blood to rebel upwards causing Phlegm to form and obstruct the cavities and vessels, thus creating Penetrating Wind or Stroke.

Windy Breathing Method - pertaining to the method of breathing through the nose.

Wood Element - one of the Five Elements, pertaining to the Liver and Gall Bladder.

Wood Jing - energy that controls the development phase of the direction of the fetus's emotional and spiritual aspects during the seventh month of pregnancy.

Wu Guan (Five Passes) - five important gates on the Governing Vessel located at the coccyx, Mingmen, Shendao, occiput, and Baihui areas where energy tends to stagnate.

Wu Huo - pertaining to the vigorous breathing method of Respiratory Dao Yin training.

Wuji - pertaining to infinite space or the formless Void.

Wuji Posture - a quiet standing posture used in meditation to allow the practitioner to return to a state of tranquility.

Wu Jing Shen (Five Essence Spirits) - the spiritual energy radiating from the core of the Five Yin Organs. Combined, these energies create the foundation of the body's Shen (Spirit).

Wu Se Dai - pertaining to the five colors of vaginal discharge - white, yellow, red, green-blue, and dark brown or black.

Wu Wei - a state of "no mind," i.e., no thoughts.

Wu Zang - the Five Yin Organs. Wu translates to mean "five," Zang translates to mean "to store or hold."

X

Xie Qi (Evil Qi) - energy that causes disease or harmful effects to the body.

Xin Xi (The Message) - knowledge stored within the Wuji or the Void.

Xiphoid Process - the lowest part of the sternum bone (sometimes referred to as the Doves Tail).

Xue - Blood.

Y

Yang - the positive charged energetic polarity, opposite of its companion Yin, pertaining to man, hard, light, hot, etc.

Yang Channels - the body's Yang energetic rivers, consisting of the Governing Vessel, Belt Vessel, Yang Linking Vessels, Yang Heel Vessels, Large Intestine Channels, Triple Burner Channels, Small Intestine Channels, Stomach Channels, Gall Bladder Channels, and Bladder Channels.

Yang Fire - also called Emperor's Fire, energy of the Heart Fire.

Yang (Fu) Organs - also known as Hollow Organs, that consist of the Gall Bladder, Small Intestine, Stomach, Large Intestine, Bladder. Also included is this category are the Triple Burners.

Yang Shen Disturbances - an emotional Yang state of energetic dysfunction.

Yang Ming (Yang Brightness) - indicates Yang Qi developing its final stage and then reverting into Yin.

Yao - a solid or broken line which is representative of either Yang or Yin energy, used in combination of three as Trigrams or six as Hexagrams.

Yao Cycles - the progression of twelve hexagrams (six Yang and six Yin) flowing in a waxing and waning cycle.

Ye (humor) - thick, turbid Body Fluids.

Yellow Court - located in the center of the diaphragm, just below the xiphoid process of the sternum. Its function is that of being the access point to releasing the body's internal organ emotional memories. Its location is also attributed to the 3rd Chakra.

Yi - the intention or thought (the cognitive mind).

Yi Jing - Chinese "Book of Changes," pertaining to the natural transitions of life.

Yin - the negative charged energetic polarity, opposite of its companion Yang, pertaining to woman, soft, dark, cold, etc.

Yin Channels - Yin energetic rivers, consisting of the Conception Vessel, Thrusting Vessel, Yin Linking Vessels, Yin Heel Vessels, Lung Channels, Pericardium Channels, Heart Channels, Spleen Channels, Liver Channels, and Kidney Channels.

Yin (Zang) Organs - also known as the Solid Organs, that consist of the Liver, Heart, Spleen, Lungs and Kidneys. Also included in this category is the Pericardium.

Yin Shen Disturbances - an emotional Yin state of energetic dysfunction.

Yin Tang (Third Eye Point) - located in the center of the forehead, responsible for projecting the Spirit for psychic intuition and communication.

Ying Qi (Nutritive Qi) - the body's nourishing energy.

Yu (Surplus) Vessels - secondary vessels that branch away from the energetic flow of the major Linking Vessels (at the chest and back), connecting the Linking Vessels energetic flow to the hands.

Yuan Jing (Original Essence) - the Original Kidney or Prenatal Essence.

Yuan Shen (Original Spirit) - the Original Prenatal Spirit.

Yuan Qi (Original Energy) - the Original Kidney or Prenatal Qi.

Yun - the Yin method of dynamic postural Dao Yin training.

Yu Zhen Guan (Occipital Pass) - the area located just inferior to the occipital bone where the brain originates, known as a specific point where Qi often stagnates.

Z

Zang Organs - Yin or solid organs (Liver, Heart, Spleen, Lungs, Kidneys and Pericardium).

Zang/Fu Organs - the body's Yin and Yang organs.

Zhang Xiang Xue Shou - in Chinese medical science, the study of energetic physiology.

Zhen Qi - see True Qi

Zheng Qi - Righteous Qi, pathogenic fighting Energy.

Zhi - the Will power, mental drive and determination.

Zhong Qi - Center Qi, Energy of the chest.

Zhou Qi - Turbid Qi, Evil Qi, Impure Qi

Zong Qi - Gathering Qi, and/or Respiratory Qi.

Zygomatic Facial Regions - pertaining to the sides of the cheeks below the eyes.

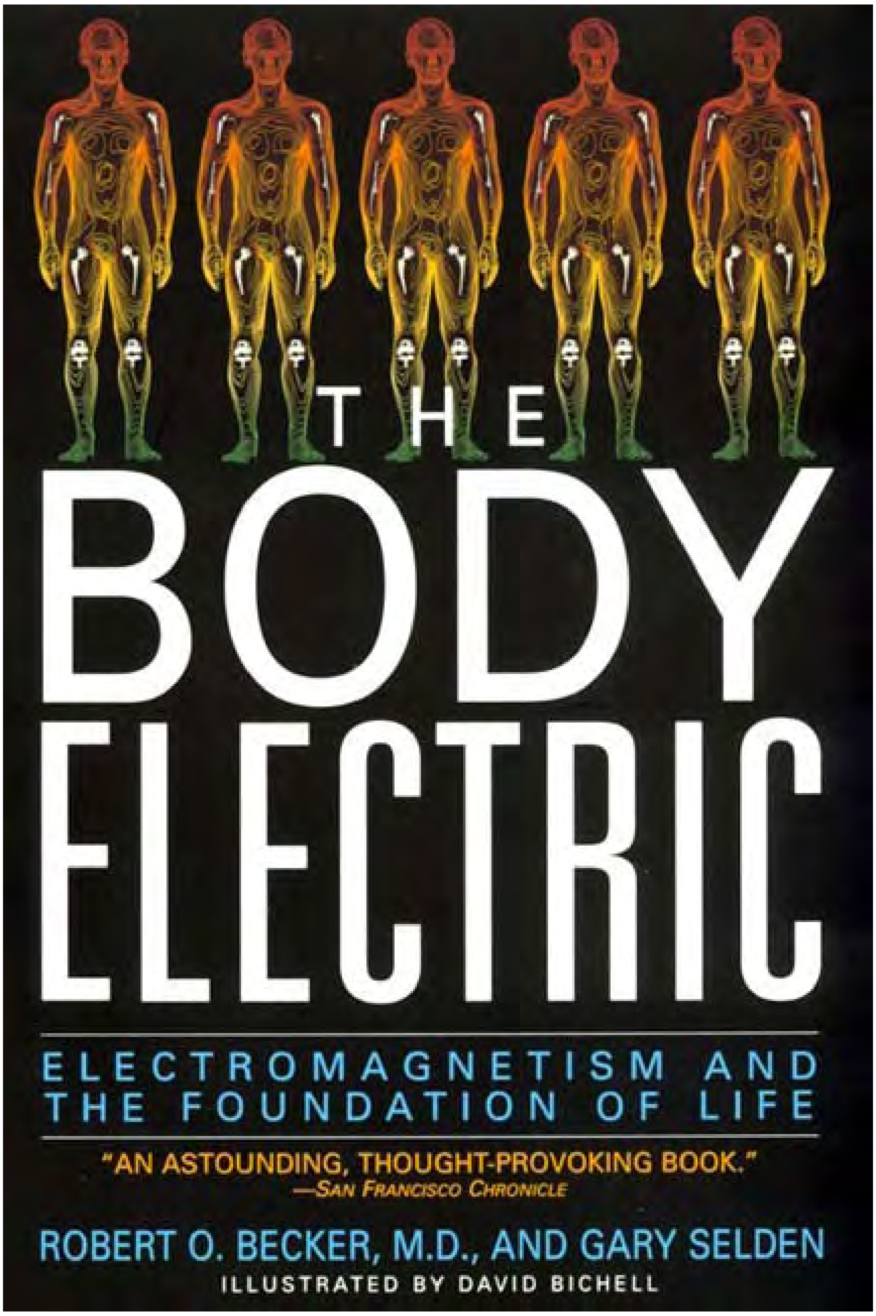

THE BODY ELECTRIC

ELECTROMAGNETISM AND THE FOUNDATION OF LIFE

"AN ASTOUNDING, THOUGHT-PROVOKING BOOK."
—*San Francisco Chronicle*

ROBERT O. BECKER, M.D., AND GARY SELDEN

ILLUSTRATED BY DAVID BICHELL

Contents

Acknowledgments 7

Co-author's Note 9

Introduction: The Promise of the Art 17

Part 1 Growth and Regrowth 23

1 Hydra's Heads and Medusa's Blood 25

Failed Healing in Bone 29

A Fable Made Fact 32

2 The Embryo at the Wound 40

Mechanics of Growth 42

Control Problems 47

Nerve Connections 55

Vital Electricity 60

3 The Sign of the Miracle 68

The Tribunal 69

The Reversals 71

Part 2 The Stimulating Current 77

4 Life's Potentials 79

Unpopular Science 82

12 The Body Electric
Undercurrents in Neurology 85
Conducting in a New Mode 91
Testing the Concept 94
5 The Circuit of Awareness 103
Closing the Circle 103
The Artifact Man and a Friend in Deed 106
The Electromagnetic Brain 110
6 The Ticklish Gene 118
The Pillars of the Temple 118
The Inner Electronics of Bone 126
A Surprise in the Blood 135
Do-It-Yourself Dedifferentiation 141
The Genetic Key 144
7 Good News for Mammals 150
A First Step with a Rat Leg 152
Childhood Powers, Adult Prospects 155
Part 3 Our Hidden Healing Energy 161
8 The Silver Wand 163
Minus for Growth, Plus for Infection 163
Positive Surprises 169
The Fracture Market 175
9 The Organ Tree 181
Cartilage 187
Skull Bones 188
Eyes 190
Muscle 191
Abdominal Organs 192
10 The Lazarus Heart 196
The Five-Alarm Blastema 197
1 1 The Self-Mending Net 203
Peripheral Nerves 206
The Spinal Cord 207
The Brain 213
12 Righting a Wrong Turn 215
A Reintegrative Approach 219
Part 4 The Essence of Life 227
13 The Missing Chapter 229
The Constellation of the Body 233
Unifying Pathways 238

14 Breathing with the Earth 243
The Attractions of Home 250
The Face of the Deep 255
Crossroads of Evolution 261
Hearing Without Ears 264
15 Maxwell's Silver Hammer 271
Subliminal Stress 276
Power Versus People 278
Fatal Locations 284
The Central Nervous System 284
The Endocrine, Metabolic, and Cardiovascular Systems 288
Growth Systems and Immune Response 292
Conflicting Standards 304
Invisible Warfare 317
Critical Connections 326
Postscript: Political Science 330
Glossary 348
Index 353

Introduction:
The Promise of the Art

I remember how it was before penicillin. I was a medical student at the end of World War II, before the drug became widely available for civilian use, and I watched the wards at New York's Bellevue Hospital fill to overflowing each winter. A veritable Byzantine city unto itself, Bellevue sprawled over four city blocks, its smelly, antiquated buildings jammed together at odd angles and interconnected by a rabbit warren of underground tunnels. In wartime New York, swollen with workers, sailors, soldiers, drunks, refugees, and their diseases from all over the world, it was perhaps the place to get an all-inclusive medical education. Bellevue's charter decreed that, no matter how full it was, every patient who needed hospitalization had to be admitted. As a result, beds were packed together side by side, first in the aisles, then out into the corridor. A ward was closed only when it was physically impossible to get another bed out of the elevator.

Most of these patients had lobar (pneumococcal) pneumonia. It didn't take long to develop; the bacteria multiplied unchecked, spilling over from the lungs into the bloodstream, and within three to five days of the first symptom the crisis came. The fever rose to 104 or 105 degrees Fahrenheit and delirium set in. At that point we had two signs to go by: If the skin remained hot and dry, the victim would die; sweating meant the patient would pull through. Although sulfa drugs often were effective against the milder pneumonias, the outcome in severe lobar pneumonia still depended solely on the struggle between the infection and the patient's own resistance. Confident in my new medical knowledge, I was horrified to find that we were powerless to change the course of this infection in any way.

It's hard for anyone who hasn't lived through the transition to realize the change that penicillin wrought. A disease with a mortality rate near 50 percent, that killed almost a hundred thousand Americans each year, that struck rich as well as poor and young as well as old, and against which we'd had no defense, could suddenly be cured without fail in a few hours by a pinch of white powder. Most doctors who have graduated since 1950 have never even seen pneumococcal pneumonia in crisis.

Although penicillin's impact on medical practice was profound, its impact on the philosophy of medicine was even greater. When Alexander Fleming noticed in 1928 that an accidental infestation of the mold Penicillium notatum had killed his bacterial cultures, he made the crowning discovery of scientific medicine. Bacteriology and sanitation had already vanquished the great plagues. Now penicillin and subsequent antibiotics defeated the last of the invisibly tiny predators.

The drugs also completed a change in medicine that had been gathering strength since the nineteenth century. Before that time, medicine had been an art. The masterpiece—a cure—resulted from the patient's will combined with the physician's intuition and skill in using remedies culled from millennia of observant trial and error. In the last two centuries medicine more and more has come to be a science, or more accurately the application of one science, namely biochemistry. Medical techniques have come to be tested as much against current concepts in biochemistry as against their empirical results. Techniques that don't fit such chemical concepts—even if they seem to work—have been abandoned as pseudoscientific or downright fraudulent.

At the same time and as part of the same process, life itself came to be defined as a purely chemical phenomenon. Attempts to find a soul, a vital spark, a subtle something that set living matter apart from the nonliving, had failed. As our knowledge of the kaleidoscopic activity within cells grew, life came to be seen as an array of chemical reactions, fantastically complex but no different in kind from the simpler reactions performed in every high school lab. It seemed logical to assume that the ills of our chemical flesh could be cured best by the right chemical antidote, just as penicillin wiped out bacterial invaders without harming human cells. A few years later the decipherment of the DNA code seemed to give such stout evidence of life's chemical basis that the double helix became one of the most hypnotic symbols of our age. It seemed the final proof that we'd evolved through 4 billlion years of chance molecular encounters, aided by no guiding principle but the changeless properties of the atoms themselves.

The philosophical result of chemical medicine's success has been belief in the Technological Fix. Drugs became the best or only valid treatments for all ailments. Prevention, nutrition, exercise, lifestyle, the patient's physical and mental uniqueness, environmental pollutants—all were glossed over. Even today, after so many years and millions of dollars spent for negligible results, it's still assumed that the cure for cancer will be a chemical that kills malignant cells without harming healthy ones. As surgeons became more adept at repairing bodily structures or replacing them with artificial parts, the technological faith came to include the idea that a transplanted kidney, a plastic heart valve, or a stainless-steel-and-Teflon hip joint was just as good as the original—or even better, because it wouldn't wear out as fast. The idea of a bionic human was the natural outgrowth of the rapture over penicillin. If a human is merely a chemical machine, then the ultimate human is a robot.

No one who's seen the decline of pneumonia and a thousand other infectious diseases, or has seen the eyes of a dying patient who's just been given another decade by a new heart valve, will deny the benefits of technology. But, as most advances do, this one has cost us something irreplaceable: medicine's humanity. There's no room in technological medicine for any presumed sanctity or uniqueness of life. There's no need for the patient's own self-healing force nor any strategy for enhancing it. Treating a life as a chemical automaton means that it makes no difference whether the doctor cares about—or even knows—the patient, or whether the patient likes or trusts the doctor.

Because of what medicine left behind, we now find ourselves in a real technological fix. The promise to humanity of a future of golden health and extended life has turned out to be empty. Degenerative diseases—heart attacks, arteriosclerosis, cancer, stroke, arthritis, hypertension, ulcers, and all the rest—have replaced infectious diseases as the major enemies of life and destroyers of its quality. Modern medicine's incredible cost has put it farther than ever out of reach of the poor and now threatens to sink the Western economies themselves. Our cures too often have turned out to be double-edged swords, later producing a secondary disease; then we search desperately for another cure. And the dehumanized treatment of symptoms rather than patients has alienated many of those who can afford to pay. The result has been a sort of medical schizophrenia in which many have forsaken establishment medicine in favor of a holistic, prescient type that too often neglects technology's real advantages but at least stresses the doctor-patient relationship, preventive care, and nature's innate recuperative power.

The failure of technological medicine is due, paradoxically, to its success, which at first seemed so overwhelming that it swept away all aspects of medicine as an art. No longer a compassionate healer working at the bedside and using heart and hands as well as mind, the physician has become an impersonal white-gowned ministrant who works in an office or laboratory. Too many physicians no longer learn from their patients, only from their professors. The breakthroughs against infections convinced the profession of its own infallibility and quickly ossified its beliefs into dogma. Life processes that were inexplicable according to current biochemistry have been either ignored or misinterpreted. In effect, scientific medicine abandoned the central rule of science—revision in light of new data. As a result, the constant widening of horizons that has kept physics so vital hasn't occurred in medicine. The mechanistic assumptions behind today's medicine are left over from the turn of the century, when science was forcing dogmatic religion to see the evidence of evolution. (The reeruption of this same conflict today shows that the battle against frozen thinking is never finally won.) Advances in cybernetics, ecological and nutritional chemistry, and solid-state physics haven't been integrated into biology. Some fields, such as parapsychology, have been closed out of mainstream scientific inquiry altogether. Even the genetic technology that now commands such breathless admiration is based on principles unchallenged for decades and unconnected to a broader concept of life. Medical research, which has limited itself almost exclusively to drug therapy, might as well have been wearing blinders for the last thirty years.

It's no wonder, then, that medical biology is afflicted with a kind of tunnel vision. We know a great deal about certain processes, such as the genetic code, the function of the nervous system in vision, muscle movement, blood clotting, and respiration on both the somatic and the cellular levels. These complex but superficial processes, however, are only the tools life uses for its survival. Most biochemists and doctors aren't much closer to the "truth" about life than we were three decades ago. As Albert Szent-Gyorgyi, the discoverer of vitamin C, has written, "We know life only by its symptoms." We understand virtually nothing about such basic life functions as pain, sleep, and the control of cell differentiation, growth, and healing. We know little about the way every organism regulates its metabolic activity in cycles attuned to the fluctuations of earth, moon, and sun. We are ignorant about nearly every aspect of consciousness, which may be broadly defined as the self-interested integrity that lets each living thing marshal its responses to eat, thrive, reproduce, and avoid danger by patterns that range from the tropisms of single cells to instinct, choice, memory, learning, individuality, and creativity in more complex life-forms. The problem of when to "pull the plug" shows that we don't even know for sure how to diagnose death. Mechanistic chemistry isn't adequate to understand these enigmas of life, and it now acts as a barrier to studying them. Erwin Chargaff, the biochemist who discovered base pairing in DNA and thus opened the way for understanding gene structure, phrased our dilemma precisely when he wrote of biology, "No other science deals in its very name with a subject that it cannot define."

Given the present climate, I've been a lucky man. I haven't been a good, efficient doctor in the modern sense. I've spent far too much time on a few incurable patients whom no one else wanted, trying to find out how our ignorance failed them. I've been able to tack against the prevailing winds of orthodoxy and indulge my passion for experiment. In so doing I've been part of a little-known research effort that has made a new start toward a definition of life.

My research began with experiments on regeneration, the ability of some animals, notably the salamander, to grow perfect replacements for parts of the body that have been destroyed. These studies, described in Part 1, led to the discovery of a hitherto unknown aspect of animal life—the existence of electrical currents in parts of the nervous system. This breakthrough in turn led to a better understanding of bone fracture healing, new possibilities for cancer research, and the hope of human regeneration—even of the heart and spinal cord—in the not too distant future, advances that are discussed in Parts 2 and 3. Finally, a knowledge of life's electrical dimension has yielded fundamental insights (considered in Part 4) into pain, healing, growth, consciousness, the nature of life itself, and the dangers of our electromagnetic technology.

I believe these discoveries presage a revolution in biology and medicine. One day they may enable the physician to control and stimulate healing at will. I believe this new knowledge will also turn medicine in the direction of greater humility, for we should see that whatever we achieve pales before the self-healing power latent in all organisms. The results set forth in the following pages have convinced me that our understanding of life will always be imperfect. I hope this realization will make medicine no less a science, yet more of an art again. Only then can it deliver its promised freedom from disease.

Part 1

Growth and Regrowth

Salamander: energy's seed sleeping interred in the marrow . . .
—Octavio Paz

One

Hydra's Heads and Medusa's Blood

There is only one health, but diseases are many. Likewise, there appears to be one fundamental force that heals, although the myriad schools of medicine all have their favorite ways of cajoling it into action.

Our prevailing mythology denies the existence of any such generalized force in favor of thousands of little ones sitting on pharmacists' shelves, each one potent against only a few ailments or even a part of one. This system often works fairly well, especially for treatment of bacterial diseases, but it's no different in kind from earlier systems in which a specific saint or deity, presiding over a specific healing herb, had charge of each malady and each part of the body. Modern medicine didn't spring full-blown from the heads of Pasteur and Lister a hundred years ago.

If we go back further, we find that most medical systems have combined such specifics with a direct, unitary appeal to the same vital principle in all illnesses. The inner force can be tapped in many ways, but all are variations of four main, overlapping patterns: faith healing, magic healing, psychic healing, and spontaneous healing. Although science derides all four, they sometimes seem to work as well for degenerative diseases and long-term healing as most of what Western medicine can offer.

Faith healing creates a trance of belief in both patient and practitioner, as the latter acts as an intercessor or conduit between the sick mortal and a presumed higher power. Since failures are usually ascribed to a lack of faith by the patient, this brand of medicine has always been a gold mine for charlatans. When bona fide, it seems to be an escalation of the placebo effect, which produces improvement in roughly one third of subjects who think they're being treated but are actually being given dummy pills in tests of new drugs. Faith healing requires even more confidence from the patient, so the disbeliever probably can prevent a cure and settle for the poor satisfaction of "I told you so." If even a few of these oft-attested cases are genuine, however, the healed one suddenly finds faith turned into certainty as the withered arm aches with unaccustomed sensation, like a starved animal waking from hibernation.

Magical healing shifts the emphasis from the patient's faith to the doctor's trained will and occult learning. The legend of Teta, an Egyptian magician from the time of Khufu (Cheops), builder of the Great Pyramid, can serve as an example. At the age of 110, Teta was summoned into the royal presence to demonstrate his ability to rejoin a severed head to its body, restoring life. Khufu ordered a prisoner beheaded, but Teta discreetly suggested that he would like to confine himself to laboratory animals for the moment. So a goose was decapitated. Its body was laid at one end of the hall, its head at the other. Teta repeatedly pronounced his words of power, and each time, the head and body twitched a little closer to each other, until finally the two sides of the cut met. They quickly fused, and the bird stood up and began cackling. Some consider the legendary miracles of Jesus part of the same ancient tradition, learned during Christ's precocious childhood in Egypt.

Whether or not we believe in the literal truth of these particular accounts, over the years so many otherwise reliable witnesses have testified to healing "miracles" that it seems presumptuous to dismiss them all as fabrications. Based on the material presented in this book, I suggest Coleridge's "willing suspension of disbelief" until we understand healing better. Shamans apparently once served at least some of their patients well, and still do where they survive on the fringes of the industrial world. Magical medicine suggests that our search for the healing power isn't so much an exploration as an act of remembering something that was once intuitively ours, a form of recall in which the knowledge is passed on or awakened by initiation and apprenticeship to the man or woman of power.

Sometimes, however, the secret needn't be revealed to be used. Many psychic healers have been studied, especially in the Soviet Union, whose gift is unconscious, unsought, and usually discovered by accident. One who demonstrated his talents in the West was Oskar Estebany. A Hungarian Army colonel in the mid-1930s. Estebany notices that horses he groomed got their wind back and recovered from illnesses faster than those cared for by others. He observed and used his powers informally for years, until, forced to emigrate after the 1956 Hungarian revolution, he settled in Canada and came to the attention of Dr. Bernard Grad, a biologist at McGill University. Grad found that Estebany could accelerate the healing of measured skin wounds made on the backs of mice, as compared with controls. He didn't let Estebany touch the animals, but only place his hands near their cages, because handling itself would have fostered healing. Estebany also speeded up the growth of barley plants and reactivated ultraviolet-damaged samples of the stomach enzyme trypsin in much the same way as a magnetic field, even though no magnetic field could be detected near his body with the instruments of that era.

The types of healing we've considered so far have trance and touch as common factors, but some modes don't even require a healer. The spontaneous miracles at Lourdes and other religious shrines require only a vision, fervent prayer, perhaps a momentary connection with a holy relic, and intense concentration on the diseased organ or limb. Other reports suggest that only the intense concentration is needed, the rest being aids to that end. When Diomedes, in the fifth book of the Iliad, dislocates Aeneas' hip with a rock, Apollo takes the Trojan hero to a temple of healing and restores full use of his leg within minutes. Hector later receives the same treatment after a rock in the chest fells him. We could dismiss these accounts as the hyperbole of a great poet if Homer weren't so realistic in other battlefield details, and if we didn't have similar accounts of soldiers in recent wars recovering from "mortal" wounds or fighting on while oblivious to injuries that would normally cause excruciating pain. British Army surgeon Lieutenant Colonel H. K. Beecher described 225 such cases in print after World War II. One soldier at Anzio in 1943, who'd had eight ribs severed near the spine by shrapnel, with punctures of the kidney and lung, who was turning blue and near death, kept trying to get off his litter because he thought he was lying on his rifle. His bleeding abated, his color returned, and the massive wound began to heal after no treatment but an insignificant dose of sodium amytal, a weak sedative given him because there was no morphine.

These occasional prodigies of battlefield stress strongly resemble the ability of yogis to control pain, stop bleeding, and speedily heal wounds with their will alone. Biofeedback research at the Menninger Foundation and elsewhere has shown that some of this same power can be rapped in people with no yogic training. That the will can be applied to the body's ills has also been shownb by Norman Cousing in his resolute conquest by laugh therapy of ankylosing spondylitis, a crippling disease in which the spinal discs and ligaments solidify like bone, and by some similar successes by users of visualization techniques to focus the mind against cancer.

Unfortunately, no approach is a sure thing. In our ignorance, the common denominator of all healing—even the chemical cures we profess to understand—remains its mysteriousness. Its unpredictability has bedeviled doctors throughout history. Physicians can offer no reason why one patient will respond to a tiny dose of a medicine that has no effect on another patient in ten times the amount, or why some cancers go into remission while others grow relentlessly unto death.

By whatever means, if the energy is successfully focused, it results in a marvelous transformation. What seemed like an inexorable decline suddenly reverses itself. Healing can almost be defined as a miracle. Instant regrowth of damaged parts and invincibility against disease are commonplaces of the divine world. They continually appear even in myths that have nothing to do with the theme of healing itself. Dead Vikings went to a realm where they could savor the joys of killing all day long, knowing their wounds would heal in time for the next day's mayhem. Prometheus' endlessly regrowing liver was only a clever torture arranged by Zeus so that the eagle sent as punishment for the god's delivery of fire to mankind could feast on his most vital organ forever—although the tale also suggests that the prehistoric Greeks knew something of the liver's ability to enlarge in compensation for damage to it.

The Hydra was adept at these offhand wonders, too. This was the monster Hercules had to kill as his second chore for King Eurystheus. The beast had somewhere between seven and a hundred heads, and each time Hercules cut one off, two new ones sprouted in its place—until the hero got the idea of having his nephew Iolaus cauterize each neck as soon as the head hit the ground.

In the eighteenth century the Hydra's name was given to a tiny aquatic animal having seven to twelve "heads," or tentacles, on a hollow, stalklike body, because this creature can regenerate. The mythic Hydra remains a symbol of that ability, possessed to some degree by most animals, including us.

Generation, life's normal transformation from seed to adult, would seem as unearthly as regeneration if it were not so commonplace. We see the same kinds of changes in each. The Greek hero Cadmus grows an army by sowing the teeth of a dragon he has killed. The primeval serpent makes love to the World Egg, which hatches all the creatures of the earth. God makes Adam from Eve's rib, or vice versa in the later version. The Word of God commands life to unfold. The genetic words encoded in DNA spell out the unfolding. At successive but still limited levels of understanding, each of these beliefs tries to account for the beautifully bizarre metamorphosis. And if some savage told us of a magical worm that built a little windowless house, slept there a season, then one day emerged and flew away as a jeweled bird, we'd laugh at such superstition if we'd never seen a butterfly.

The healer's job has always been to release something not understood, to remove obstructions (demons, germs, despair) between the sick patient and the force of life driving obscurely toward wholeness. The means may be direct—the psychic methods mentioned above—or indirect: Herbs can be used to stimulate recovery; this tradition extends from prehistoric wisewomen through the Greek herbal of Dioscorides and those of Renaissance Europe, to the prevailing drug therapies of the present. Fasting, controlled nutrition, and regulation of living habits to avoid stress can be used to coax the latent healing force from the sick body; we can trace this approach back from today's naturopaths to Galen and Hippocrates. Attendants at the healing temples of ancient Greece and Egypt worked to foster a dream in the patient that would either start the curative process in sleep or tell what must be done on awakening. This method has gone out of style, but it must have worked fairly well, for the temples were filled with plaques inscribed by grateful patrons who'd recovered. In fact, this mode was so esteemed that Aesculapius, the legendary doctor who originated it, was said to have been given two vials filled with the blood of Medusa, the snaky-haired witch-queen killed by Perseus. Blood from her left side restored life, while that from her right took it away—and that's as succinct a description of the tricky art of medicine as we're likely to find.

The more I consider the origins of medicine, the more I'm convinced that all true physicians seek the same thing. The gulf between folk therapy and our own stainless-steel version is illusory. Western medicine springs from the same roots and, in the final analysis, acts through the same little-understood forces as its country cousins. Our doctors ignore this kinship at their—and worse, their patients'—peril. All worthwhile medical research and every medicine man's intuition is part of the same quest for knowledge of the same elusive healing energy.

Failed Healing in Bone

As an orthopedic surgeon, I often pondered one particular breakdown of that energy, my specialty's major unsolved problem—nonunion of fractures. Normally a broken bone will begin to grow together in a few weeks if the ends are held close to each other without movement. Occasionally, however, a bone will refuse to knit despite a year or more of casts and surgery. This is a disaster for the patient and a bitter defeat for the doctor, who must amputate the arm or leg and fit a prosthetic substitute.

Throughout this century, most biologists have been sure only chemical processes were involved in growth and healing. As a result, most work on nonunions has concentrated on calcium metabolism and hormone relationships. Surgeons have also "freshened," or scraped, the fracture surfaces and devised ever more complicated plates and screws to hold the bone ends rigidly in place. These approaches seemed superficial to me. I doubted that we would ever understand the failure to heal unless we truly understood healing itself.

When I began my research career in 1958, we already knew a lot about the logistics of bone mending. It seemed to involve two separate processes, one of which looked altogether different from healing elsewhere in the human body. But we lacked any idea of what set these processes in motion and controlled them to produce a bone bridge across the break.

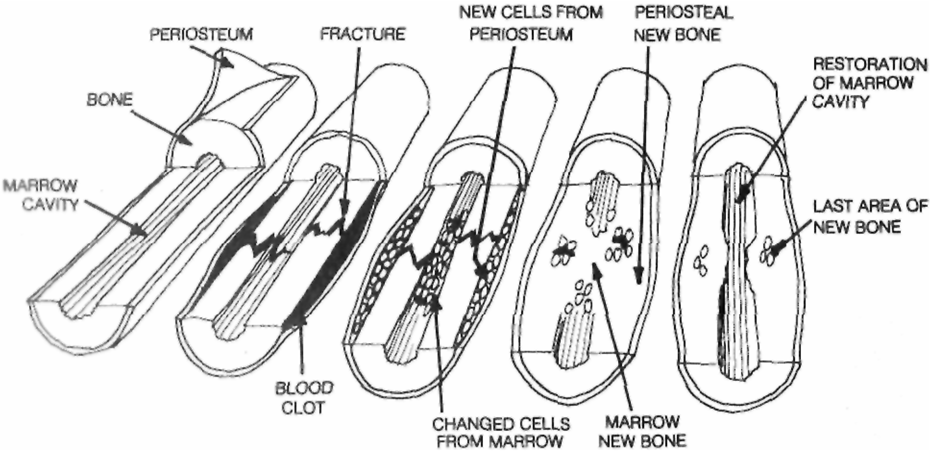
PERIOSTEUM
BONE
MARROW CAVITY
FRACTURE
BLOOD CLOT
NEW CELLS FROM PERIOSTEUM
CHANGED CELLS FROM MARROW
PERIOSTEAL NEW BONE
MARROW NEW BONE
RESTORATION OF MARROW CAVITY
LAST AREA OF NEW BONE

STAGES OF FRACTURE HEALING

Every bone is wrapped in a layer of tough, fibrous collagen, a protein that's a major ingredient of bone itself and also forms the connective tissue or "glue" that fastens all our cells to each other. Underneath the collagen envelope are the cells that produce it, right next ro the bone; together the two layers form the periosteum. When a bone breaks, these periosteal cells divide in a particular way. One of the daughter cells stays where it is, while the other one migrates into the blood clot surrounding the fracture and changes into a closely related type, an osteoblast, or bone-forming cell. These new osteoblasts build a swollen ring of bone, called a callus, around the break.

Another repair operation is going on inside the bone, in its hollow center, the medullary cavity. In childhood the marrow in this cavity actively produces red and white blood cells, while in adulthood most of the marrow turns to fat. Some active marrow cells remain, however, in the porous convolutions of the inner surface. Around the break a new tissue forms from the marrow cells, most readily in children and young animals. This new tissue is unspecialized, and the marrow cells seem to form it not by increasing their rate of division, as in the callus-forming periosteal cells, but by reverting to a primitive, neo-embryonic state. The unspecialized former marrow cells then change into a type of primitive cartilage cells, then into mature cartilage cells, and finally into new bone cells to help heal the break from inside. Under a microscope, the changes seen in cells from this internal healing area, especially from children a week or two after the bone was broken, seem incredibly chaotic, and they look frighteningly similar to highly malignant bone-cancer cells. But in most cases their transformations are under control, and the bone heals.

Dr. Marshall Urist, one of the great researchers in orthopedics, was to conclude in the early 1960s that this second type of bone healing is an evolutionary throwback, the only kind of regeneration that humans share with all other vertebrates. Regeneration in this sense means the regrowth of a complex body part, consisting of several different kinds of cells, in a fashion resembling the original growth of the same part in the embryo, in which the necessary cells differentiate from simpler cells or even from seemingly unrelated types. This process, which I'll call true regeneration, must be distinguished from two other forms of healing. One, sometimes considered a variety of regeneration, is physiological repair, in which small wounds and everyday wear within a single tissue are made good by nearby cells of the same type, which merely proliferate to close the gap. The other kind of healing occurs when a wound is too big for single-tissue repair but the animal lacks the true regenerative competence to restore the damaged part. In this case the injury is simply patched over as well as possible with collagen fibers, forming a scar. Since true regeneration is most closely related to embryonic development and is generally Strongest in simple animals, it may be considered the most fundamental mode of healing.

Nonunions failed to knit, I reasoned, because they were missing something that triggered and controlled normal healing. I'd already begun to wonder if the inner area of bone mending might be a vestige of true regeneration. If so, it would likely show the control process in a clearer or more basic form than the other two kinds of healing. I figured I stood little chance of isolating a clue to it in the multilevel turmoil of a broken bone itself, so I resolved to study regeneration alone, as it occurred in other animals.

A Fable Made Fact

Regeneration happens all the time in the plant kingdom. Certainly this knowledge was acquired very early in mankind's history. Besides locking up their future generations in the mysterious seed, many plants, such as grapevines, could form a new plant from a single part of the old. Some classical authors had an inkling of animal regeneration—Aristotle mentions that the eyes of very young swallows recover from injury, and Pliny notes that lost "tails" of octopi and lizards regrow. However, regrowth was thought to be almost exclusively a plant prerogative.

The great French scientist Rene Antoine Ferchault de Reaumur made the first scientific description of animal regeneration in 1712. Reaumur devoted all his life to the study of "insects," which at that time meant all invertebrates, everything that was obviously "lower" than lizards, frogs, and fish. In studies of crayfish, lobsters, and crabs, Reaumur proved the claims of Breton fishermen that these animals could regrow lost legs. He kept crayfish in the live-bait well of a fishing boat, removing a claw from each and observing that the amputated extremity reappeared in full anatomical detail. A tiny replica of the limb took shape inside the shell; when the shell was discarded at the next molting season, the new limb unfolded and grew to full size.

Reaumur was one of the scientific geniuses of his time. Elected to the Royal Academy of Sciences when only twenty-four, he went on to invent tinned steel, Reaumur porcelain (an opaque white glass), imitation pearls, better ways of forging iron, egg incubators, and the Reaumur thermometer, which is still used in France. At the age of sixty-nine he isolated gastric juice from the stomach and described its digestive function. Despite his other accomplishments, "insects" were his life's love (he never married), and he probably was the first to conceive of the vast, diverse population of life-forms that this term encompassed. He rediscovered the ancient royal purple dye from Murex trunculus (a marine mollusk), and his work on spinning a fragile, filmy silk from spider webs was translated into Manchu for the Chinese emperor. He was the first to elucidate the social life and sexually divided caste system of bees. Due to his eclipse in later years by court-supported scientists who valued "common sense" over observation, Reaumur's exhaustive study of ants wasn't published until 1926. In the interim it had taken several generations of formicologists to cover the same ground, including the description of winged ants copulating in flight and proof that they aren't a separate species but the sexual form of wingless ants. In 1734 he published the first of six volumes of his Natural History of Insects, a milestone in biology.

Reaumur made so many contributions to science that his study of regeneration was overlooked for decades. At that time no one really cared what strange things these unimportant animals did. However, all of the master's work was well known to a younger naturalist, Abraham Trembley of Geneva, who supported himself, as did many educated men of that time, by serving as a private tutor for sons of wealthy families. In 1740, while so employed at an estate near The Hague, in Holland, Trembley was examining with a hand lens the small animals living in freshwater ditches and ponds. Many had been described by Reaumur, but Trembley chanced upon an odd new one. It was no more than a quarter of an inch long and faintly resembled a squid, having a cylindrical body topped with a crown of tentacles. However, it was a startling green color. To Trembley, green meant vegetation, but if this was a plant, it was a mighty peculiar one. When Trembley agitated the water in its dish, the tentacles contracted and the body shrank down to a nubbin, only to reexpand after a period of quiet. Strangest of all, he saw that the creature "walked" by somersaulting end over end.

Since they had the power of locomotion, Trembley would have assumed that these creatures were animals and moved on to other observations, if he hadn't chanced to find a species colored green by symbiotic algae. To settle the animal-plant question, he decided to cut some in half. If they regrew, they must be plants with the unusual ability to walk, while if they couldn't regenerate, they must be green animals.

Trembley soon entered into a world that exceeded his wildest dreams. He divided the polyps, as he first called them, in the middle of their stalks. He then had two short pieces of stalk, one with attached tentacles, each of which contracted down to a tiny dot. Patiently watching, Trembley saw the two pieces later expand. The tentacle portion began to move normally, as though it were a complete organism. The other portion lay inert and apparently dead. Something must have made Trembley continue the experiment, for he watched this motionless object for nine days, during which nothing happened. He then noted that the cut

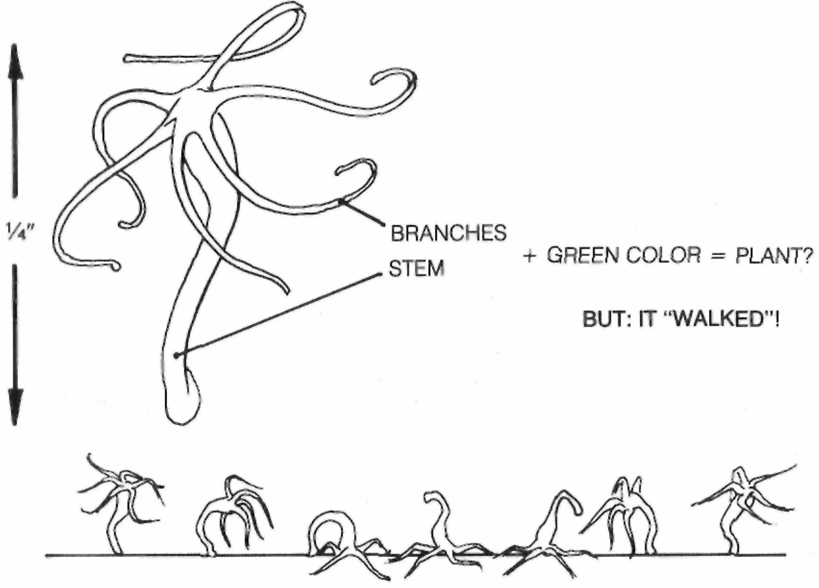

TREMBLEY'S DISCOVERY: ANIMAL OR PLANT?

end had sprouted three little "horns," and within a few more days the complete crown of tentacles had been restored. Trembley now had two complete polyps as a result of cutting one in half! However, even though they regenerated, more observations convinced Trembley that the creatures were really animals. Not only did they move and walk, but their arms captured tiny water fleas and moved them to the "mouth," located in the center of the ring of tentacles, which promptly swallowed the prey.

Trembley, then only thirty-one, decided to make sure he was right by having the great Reaumur confirm his findings before he published them and possibly made a fool of himself. He sent specimens and detailed notes to Reaumur, who confirmed that this was an animal with amazing powers of regeneration. Then he immediately read Trembley's letters and showed his specimens to an astounded Royal Academy early in 1741. The official report called Trembley's polyp more marvelous than the phoenix or the mythical serpent that could join together after being cut in two, for these legendary animals could only reconstitute themselves, while the polyp could make a duplicate. Years later Reaumur was still thunderstruck. As he wrote in Volume 6 of his series on insects, "This is a fact that I cannot accustom myself to seeing, after having seen and re-seen it hundrts of times."

This was just the beginning, however. Trembley's polyps performed even more wondrous feats. When cut lengthwise, each half of the stalk healed over without a scar and proceeded to regrow the missing tentacles. Trembley minced some polyps into as many pieces as he could manage, finding that a complete animal would regrow from each piece, as long as it included a remnant of the central stalk. In one instance he quartered one of the creatures, then cut each resulting polyp into three or four pieces, until he had made fifty animals from one.

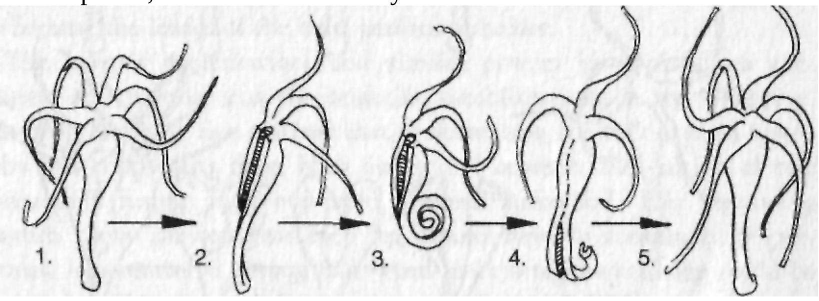

REGENERATION BY A HYDRA CUT IN HALF LENGTHWISE

His most famous experiment was the one that led him to name his polyp "hydra." He found that by splitting the head lengthwise, leaving the stalk intact, he could produce one animal with two crowns of tentacles. By continuing the process he was able to get one animal with seven heads. When Trembley lopped them off, each one regrew, just like the mythical beast's. But nature went legend one better: Each severed head went on to form a complete new animal as well.

Such experiments provided our first proof that entire animals can regenerate, and Trembley went on to observe that hydras could reproduce by simple budding, a small animal appearing on the side of the stalk and growing to full size. The implications of these discoveries were so revolutionary that Trembley delayed publishing a full account of his work until he'd been prodded by Reaumur and preceded in print by several others. The sharp division between plant and animal suddenly grew blurred, suggesting a common origin with some kind of evolution; basic assumptions about life had to be rethought. As a result, Trembley's observations weren't enthusiastically embraced by all. They inflamed several old arguments and offended many of the old guard. In this respect Trembley's mentor Reaumur was a most unusual scientist for his time, and indeed tor all time. Despite his prominence, he was ready to espouse radically new ideas and, most important, he didn't steal the ideas of others, an all too common failing among scientists.

A furious debarte was raging at the time of Trembley's announcement.

It concerned the origin of the individual - how the chicken came from

A HYDRA REGROWS ITS MOUTH AND TENTACLES

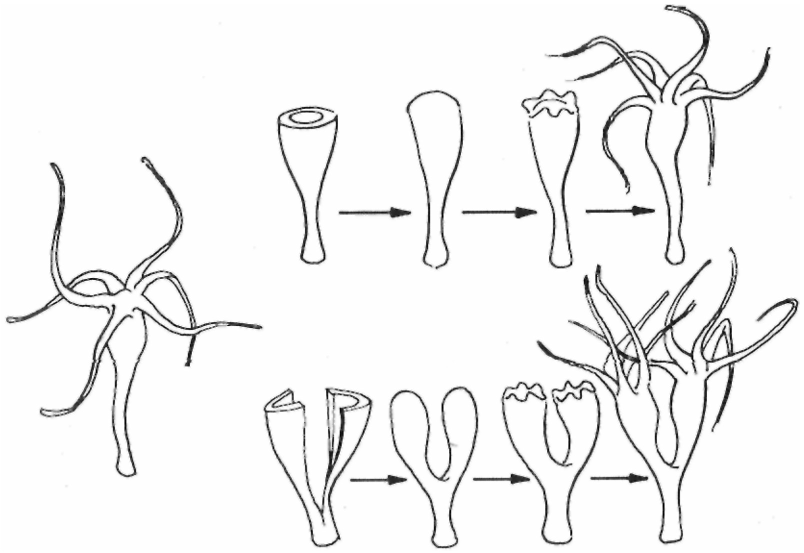

A PARTIALLY SPLIT HYDRA'S BODY GROWS TWO "HEADS"

the egg, for example. When scientists examined the newly laid egg, there wasn't much there except two liquids, the white and the yolk, neither of which had any discernible structure, let alone anything resembling a chicken.

There were two opposite theories. The older one, derived from Aristotle, held that each animal in all its complexity developed from simple organic matter by a process called epigenesis, akin to our modern concept of cell differentiation. Unfortunately, Trembley himself was the first person to witness cell division under the microscope, and he didn't realize that it was the normal process by which all cells multiplied. In an era knowing nothing of genes and so little of cells, yet beginning to insist on logical, scientific explanations, epigenesis seemed more and more absurd. What could possibly transform the gelatin of eggs and sperm into a frog or a human, without invoking that tired old deus ex machina the spirit, or inexplicable spark of life—unless the frog or person already existed in miniature inside the generative slime and merely grew in the course of development?

The litter idea, called preformation, had been ascendant for at least fifty years. It was so widely accepted that when the early microscopists studied drops of semen, they dutifully reported a little man, called a homunculus, encased in the head of each sperm - a fine example of science's capacity for self-delusion. Even Reaumur, when he failed to find tiny butterfly wings inside caterpillars, assumed they were there but were too small to be seen. Only a few months before Trembley began slicing hydras, his cousin, Genevan naturalist Charles Bonnet, had proven (in an experiment suggested by the omnipresent Reaumur) that female aphids usually reproduced parthenogenetically (without mating). To Bonnet this demonstrated that the tiny adult resided in the egg, and he became the leader of the ovist preformationists.

The hydra's regeneration, and similar powers in starfish, sea anemones, and worms, put the scientific establishment on the defensive. Reaumur had long ago realized that preformation couldn't explain how a baby inherited traits from both father and mother. The notion of two homunculi fusing into one seed seemed farfetched. His regrowing crayfish claws showed that each leg would have to contain little preformed legs scattered throughout. And since a regenerated leg could be lost and replaced many times, the proto-legs would have to be very numerous, yet no one had ever found any.

Regeneration therefore suggested some form of epigenesis—perhaps without a soul, however, for the hydra's anima, if it existed, was divisible along with the body and indistinguishable from it. It seemed as though some forms of matter itself possessed the spark of life. For lack of knowledge of cells, let alone chromosomes and genes, the epigeneticists were unable to prove their case. Each side could only point out the other's inconsistencies, and politics gave preformationism the edge.

No wonder nonscientists often grew impatient of the whole argument. Oliver Goldsmith and Tobias Smollett mocked the naturalists for missing nature's grandeur in their myopic fascination with "muck-flies." Henry Fielding lampooned the discussion in a skit about the regeneration of money. Diderot thought of hydras as composite animals, like swarms of bees, in which each particle had a vital spark of its own, and lightheartedly suggested there might be "human polyps" on Jupiter and Saturn. Voltaire was derisively skeptical of attempts to infer the nature of the soul, animal or human, from these experiments. Referring in 1768 to the regenerating heads of snails, he asked, "What happens to its sensorium, its memory, its store of ideas, its soul, when its head has been cut off? How does all this come back? A soul that is reborn is an extremely curious phenomenon." Profoundly disturbed by the whole affair, for a long time he simply refused to believe in animal regeneration, calling the hydra " a kind of small rush."

It was no longer possible to doubt the discovery after the work of Lazzaro Spallanzani, an italian priest for whom science was a full-time hobby. In a career spanning the second half of the eighteenth century, Spallanzani discovered the reversal of plant transpiration between light and darkness, and advanced our knowledge of digestion, volcanoes, blood circulation, and the senses of bats, but his most important work concerned regrowth. In twenty years of meticulous observation, he studied regeneration in worms, slugs, snails, salamanders, and tadpoles. He set new standards for thoroughness, often dissecting the amputated parts to make sure he'd removed them whole, then dissecting the replacements a few months later to confirm that all the parts had been restored. Spallanzani's most important contribution to science was his discovery of the regenerative abilities of the salamander. It could replace its tail and limbs, all at once if need be. A young one performed this feat for Spallanzani six times in three months. He later found that the salamander could also replace its jaw and the lenses of its eyes, and then went on to establish two general rules of regeneration: Simple animals can regenerate more fully than complex ones, or, in modern terms, the ability to regenerate declines as one moves up the evolutionary scale. (The salamander is the main exception.) In ontogenetic parallel, if a specias can regenerate, younger individuals do it better than older ones.

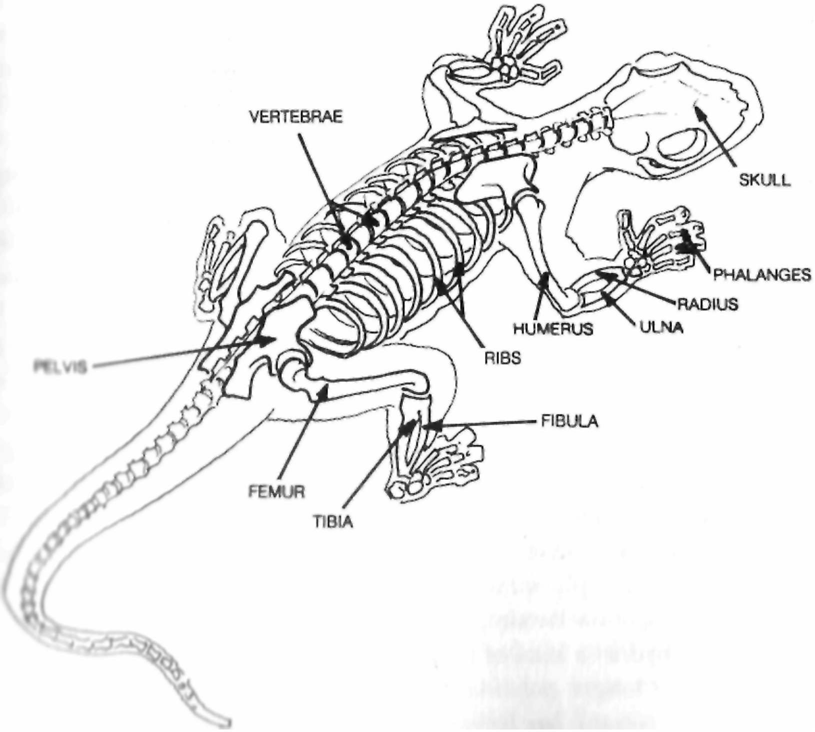

THE SALAMANDER'S SKELETON - AS COMPLEX AS OURS

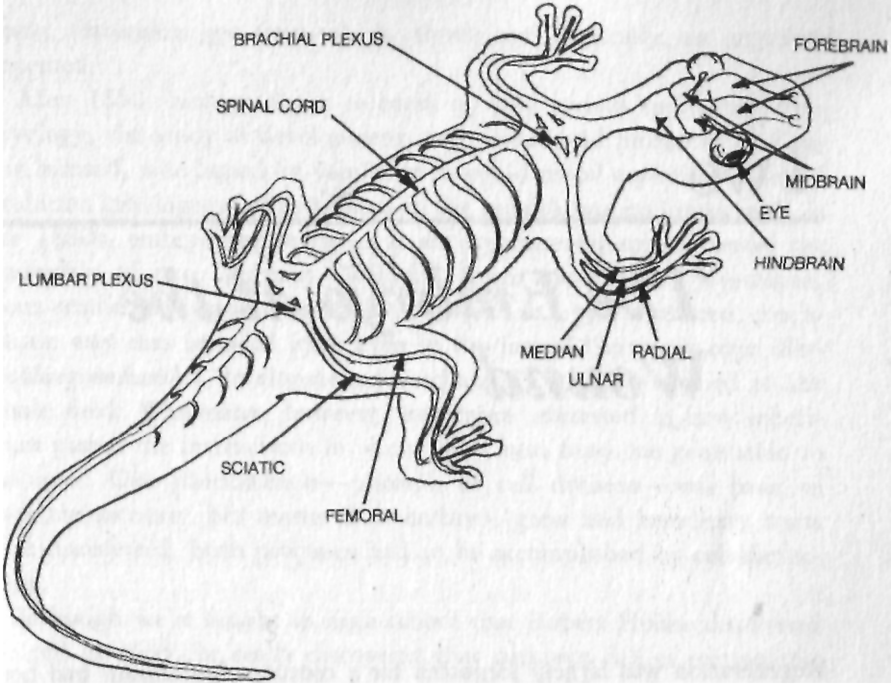

THE SALAMANDER'S NERVOUS SYSTEM—FORERUNNER OF OURS

This early regeneration research, Spallanzani's in particular, was a benchmark in modern biology. Gentlemanly observations buttressed by "common sense" gave way to a more rigorous kind of examination in which nothing was taken for granted. It had been "known" for perhaps ten thousand years that plants could regenerate and animals couldn't. To many zoologists, even twenty years after Trembley's initial discovery, the few known exceptions only proved the rule, for octopi, crayfish, hydras, worms, and snails seemed so unlike humans or the familiar mammals that they hardly counted. The lizard, the only other vertebrate regenerator then known, could manage no more than an imperfect tail. But the salamander—here was an animal we could relate to! This was no worm or snail or microscopic dot, but a four-limbed, two-eyed vertebrate that could walk and swim. While its legendary ability to withstand fire had been disproven, its body was big enough and its anatomy similar enough CO ours to be taken seriously. Scientists could no longer assume that the underlying process had nothing to do with us. In fact, the questions with which Spallanzani ended his first report on the salamander have haunted biologists ever since: "Is it to be hoped that [higher animals] may acquire [the same power] by some useful dispositions? and should the flattering expectation of obtaining this advantage for ourselves be considered entirely as chimerical?"

Two

I The Embryo at the Wound

Reneration was largely forgotten for a century. Spallanzani had been so thorough that little else could be learned about it with the techniques of the ime. Moreover, although his work strongly supported epigenesis, its impact was lost because the whole debate was swallowed up in the much larger philosophical conflict between vitalism and mechanism.

Since biology includes the study of our own essence, it's the most emotional science, and it has been the battleground for these two points of view throughout its history. Briefly, the vitalists believed in a spirit, called the anima or elan vital, that made living things fundamentally different from other substances. The mechanists believed that life could ultimately be understood in terms of the same physical and chemical laws that governed nonliving matter, and that only ignorance of these forces led people to invoke such hokum as a spirit. We'll take up these issues in more detail later, but for now we need only note that the vitalists favored epigenesis, viewed as an imposition of order on the chaos of the egg by some intangible "vital" force. The mechanists favored formation. Since science insisted increasingly on material explanations for everything, epigenesis lost out despite the evidence of regeneration.

Mechanism dominated biology more and more, but some problems remained. The main one was the absence of the little man in the sperm. Advances in the power and resolution of microscopes had clearly shown that no one was there. Biologists were faced with the generative slime again, featureless goo from which, slowly and magically, an organism appeared.

After 1850, biology began to break up into various specialties. Embryology, the study of development, was named and promoted by Darwin himself, who hoped (in vain) that it would reveal a precise history of evolution (phylogeny) recapitulated in the growth process (ontogeny). In the 1880s, embryology matured as an experimental science under the leadership of two Germans, Wilhelm Roux and August Weismann. Roux studied the stages of embryonic growth in a very restricted, mechanistic way that revealed itself even in the formal Germanic title, Entwicklungsmechanik ("developmental mechanics"), that he applied to the whole field. Weismann, however, was more interested in how inheritance passed the instructions for embryonic form from one generation to the next. One phenomenon—mitosis, or cell division—was basic to both transactions. No matter how embryos grew and hereditary traits were transferred, both processes had to be accomplished by cellular actions.

Although we're taught in high school that Robert Hooke discovered the cell in 1665, he really discovered that cork was full of microscopic holes, which he called cells because they looked like little rooms. The idea that they were the basic structural units of all living things came from Theodor Schwann, who proposed this cell theory in 1838. However, even at that late date, he didn't have a clear idea of the origin of cells. Mitosis was unknown to him, and he wasn't too sure of the distinction between plants and animals. His theory wasn't fully accepted until two other German biologists, F. A. Schneider and Otto Butschli, reintroduced Schwann's concept and described mitosis in 1873.

Observations of embryogenesis soon confirmed its cellular basis. The fertilized egg was exactly that, a seemingly unstructured single cell. Embryonic growth occurred when the fertilized egg divided into two other cells, which promptly divided again. Their progeny then divided, and so on. As they proliferated, the cells also differentiated; that is, they began to show specific characteristics of muscle, cartilage, nerve, and so forth. The creature that resulted obviously had several increasingly complex levels of organization; however, Roux and Weismann had no alternative but to concentrate on the lowest one, the cell, and try to imagine how the inherited material worked at that level.

Weismann proposed a theory of "determiners," specific chemical structures coded for each cell type The fertilized egg contained all the determiners, both in type and in number, needed to produce every cell in the body. As cell division proceeded, the daughter cells each received half of the previous stock of determiners, until in the adult each cell possessed only one. Muscle cells contained only the muscle determiner, nerve cells only the one for nerves, and so on. This meant that once a cell's function had been fixed, it could never be anything but that one kind of cell. In one of his first experiments, published in 1888, Roux obtained powerful support for this concept. He took fertilized frog eggs, which were large and easy to observe, and waited until the first cell division had occured. He then separated the two cells of this incipient embryo, According to the theory, each cell contained enough determiners to make half an embryo, and that was exactly what Roux got—two half-embryos. It was hard to argue with such a clear-cut result, and the determiner theory was widely accepted. Its triumph was a climactic victory for mechanistic concept of life, as well.

One of vitalism's last gasps came from the work of another German embryologist, Hans Driesch. Initially a firm believer in Entivicklungsmechanik, Driesch later found its concepts deficient in the face of life's continued mysteries. For example, using sea urchin eggs, he repeated Roux's famous experiment and obtained a whole organism instead of a half. Many other experiments convinced Driesch that life had some special innate drive, a process that went against known physical laws. Drawing on the ancient Greek idea of the anima, he proposed a nonmaterial, vital factor that he called entelechy. The beginning of the twentieth century wasn't a propitious time for such an idea, however, and it wasn't popular.

Mechanics of Growth

As the nineteenth century drew to a close and the embryologists continued to struggle with the problems of inheritance, they found they still needed a substitute for the homunculus. Weismann's determiners worked fine for embryonic growth, but regeneration was a glaring exception, and one that didn't prove the rule. The original theory had no provision for a limited replay of growth to replace a part lost after development was finshed. Oddly enough, the solution had already been provided by a man almost totally forgotten today, Theodor Heinrich Boveri.

Working at the University of Munich in the 1880s, Boveri discovered almost every detail of cell division, including the chromosomes. Not until the invention of the electron microscope did anyone add meterially to his original descriptions. Boveri found that all nonsexual cells of any one species contained the same number of chromosomes. As growth proceeded by mitosis, these chromosomes split lengthwise to make two of each so that each daughter cell then had the same number of chromosomes. The egg and sperm, dividing by a special process called meiosis, wound up with exactly half that number, so that the fertilized egg would start out with a full complement, half from the father and half from the mother. He reached the obvious conclusions that the chromosomes transmitted heredity, and that each one could exchange smaller units of itself with its counterpart from the other parent.

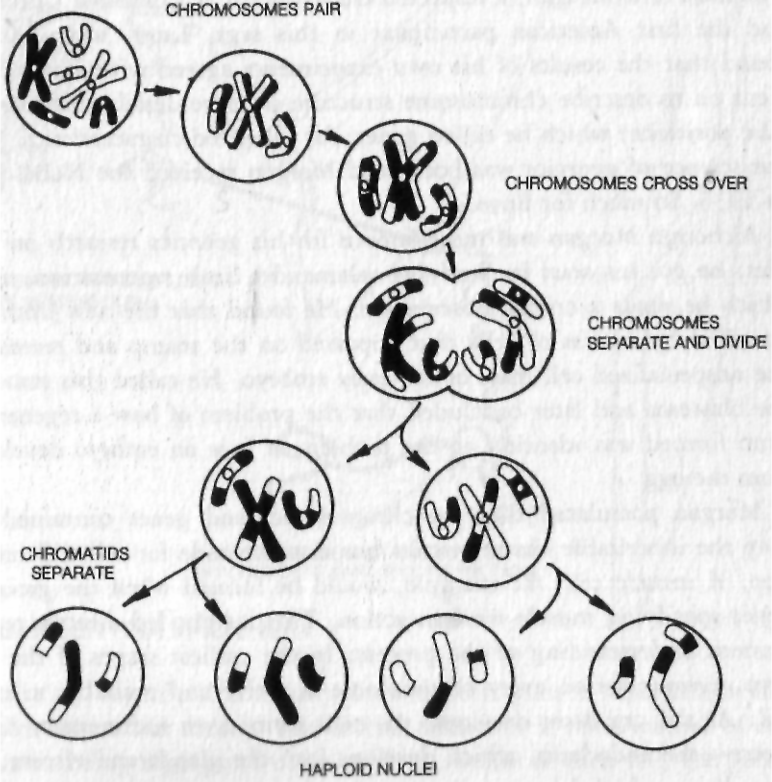
CHROMOSOMES PAIR
CHROMOSOMES CROSS OVER
CHROMOSOMES SEPARATE AND DIVIDE
CHROMATIDS SEPARATE
HAPLOID NUCLEI

MEIOSIS—FORMATION OF SEX CELLS

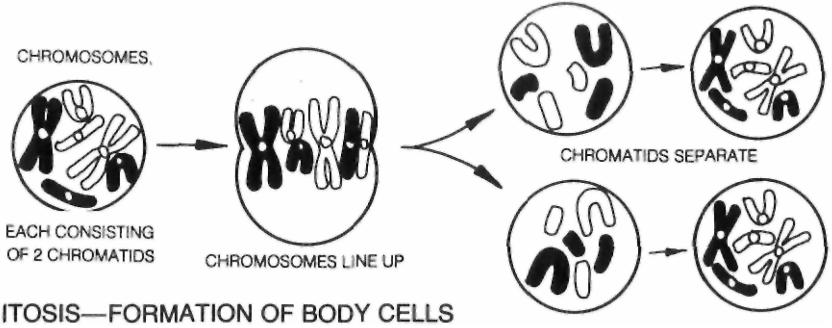

MITOSIS—FORMATION OF BODY CELLS

At first this idea wasn't well received. It was strenuously opposed by Thomas Hunt Morgan, a respected embryologist at Columbia University and the first American participant in this saga. Later, when Morgan found that the results of his own experiments agreed with Boveri's, he went on to describe chromosome structure in more detail, charting specific positions, which he called genes, for inherited characteristics. Thus the science of genetics was born, and Morgan received the Nobel Prize in 1933. So much for Boveri.

Although Morgan was most famous for his genetics research on fruit flies, he got his start by studying salamander limb regeneration, about which he made a crucial observation. He found that the new limb was preceded by a mass of cells that appeared on the stump and resembled the unspecialized cell mass of the early embryo. He called this structure the blastema and later concluded that the problem of how a regenerated limb formed was identical to the problem of how an embryo developed from the egg.

Morgan postulated that the chromosomes and genes contained not only the inheritable characteristics but also the code for cell differentiation. A muscle cell, for example, would be formed when the group of genes specifying muscle were in action. This insight led directly to our modern understanding of the process: In the earliest stages of the embryo, every gene on every chromosome is active and available to every cell. As the organism develops, the cells form three rudimentary tissue layers—the endoderm, which develops into the glands and viscera; the mesoderm, which becomes the muscles, bones, and circulatory system; and the ectoderm, which gives rise to the skin, sense organs, and nervous system. Some of the genes are already being turned off, or repressed, at this stage. As the cells differentiate into mature tissues, only one specific set of genes stays switched on in each kind. Each set can make only certain types of messenger ribonucleic acid (mRNA), the "executive secretary" chemical by means of which DNA "instructs" the ribosomes (the cell's protein-factory organelles) to make the particular proteins that distinguish a nerve cell, for example, from a muscle or cartilage cell.

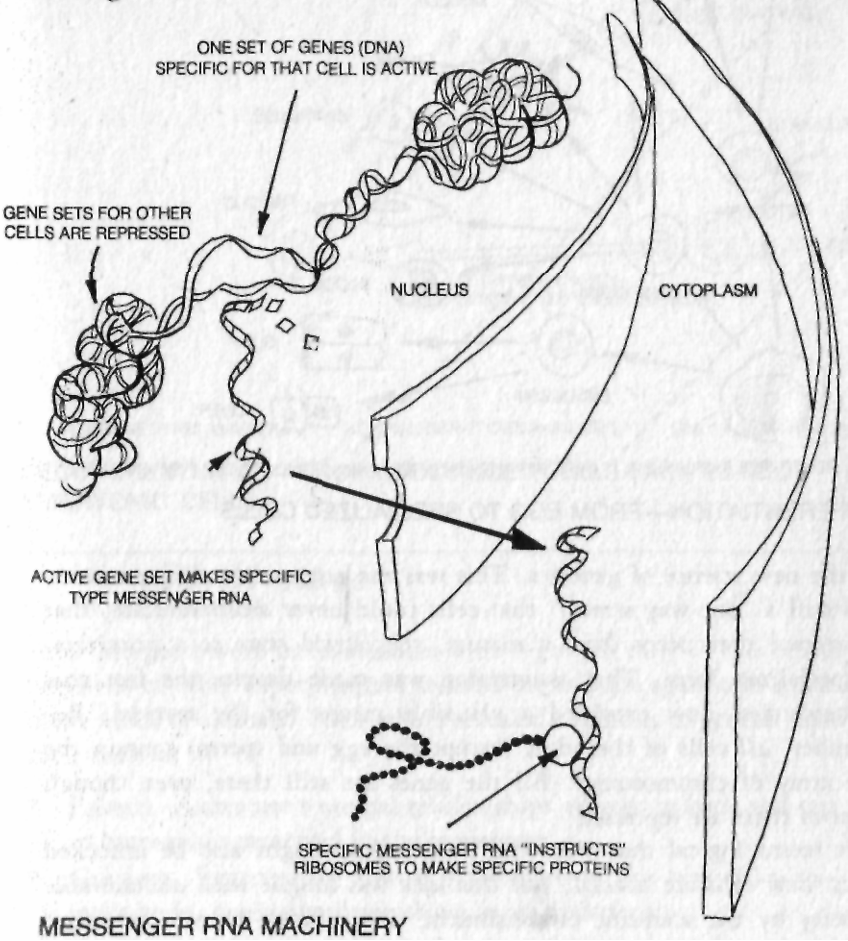
ONE SET OF GENES (DNA) SPECIFIC FOR THAT CELL IS ACTIVE
GENE SETS FOR OTHER CELLS ARE REPRESSED
NUCLEUS
CYTOPLASM
ACTIVE GENE SET MAKES SPECIFIC TYPE MESSENGER RNA
SPECIFIC MESSENGER RNA "INSTRUCTS" RIBOSOMES TO MAKE SPECIFIC PROTEINS

MESSENGER RNA MACHINERY

There's a superficial similarity between this genetic mechanism and the old determiner theory. The crucial difference is that, instead of determiners being segregated until only one remains in each cell, the genes are repressed until only one set remains active in each cell. However, the entire genetic blueprint is carried by every cell nucleus.

Science is a bit like the ancient Egyptian religion, which never threw old gods away but only tacked them onto newer deities until a bizarre hodgepodge developed. For some strange reason, science is equally reluctant to discard worn-out theories, and, even though there was absolutely no evidence to support it, one of Weismann's ideas was swallowed whole

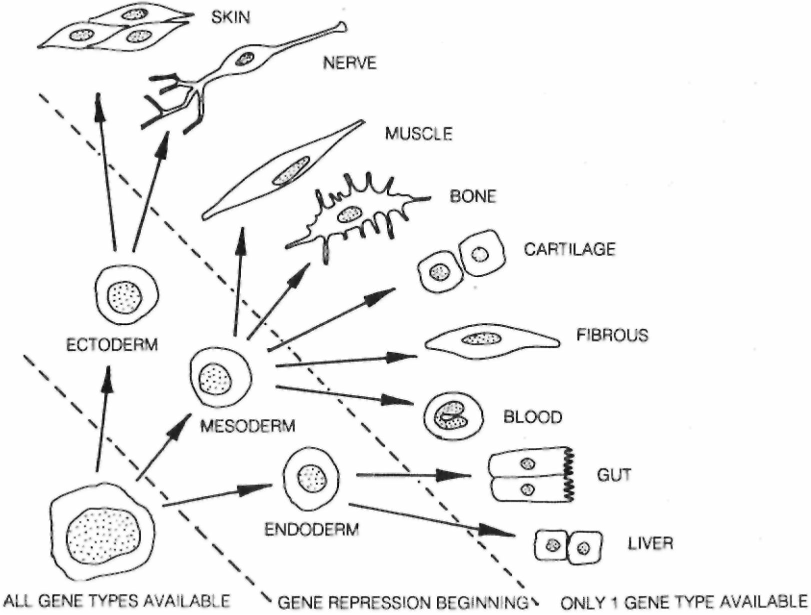
SKIN
NERVE
MUSCLE
BONE
CARTILAGE
FIBROUS
BLOOD
GUT
LIVER
ECTODERM
MESODERM
ENDODERM
ALL GENE TYPES AVAILABLE
GENE REPRESSION BEGINNING
ONLY 1 GENE TYPE AVAILABLE

DIFFERENTIATION—FROM EGG TO SPECIALIZED CELLS

by the new science of genetics. This was the notion that differentiation was still a "one-way street," that cells could never dedifferentiate, that is, retrace their steps from a mature, specialized state to a primitive, unspecialized form. This assumption was made despite the fact that chromosomes now provided a plausible means for the reversal. Remember, all cells of the adult (except the egg and sperm) contain the full array of chromosomes. All the genes are still there, even though most of them are repressed.

It seems logical that what has been locked might also be unlocked when new cells are needed, but this idea was fought with unbelievable ferocity by the scientific establishment. It's difficult now to see why, since no principle of real importance was involved, except possibly a bit of the supremacy of the mechanistic outlook itself. The mechanists greeted the discovery of genes and chromosomes joyfully. Here at last was a replacement for the little man in the sperm! Perhaps it seemed that admitting dedifferentiation would have given life too much control over its own functions. Perhaps, once genes were considered the sole mechanism of life, they had to work in a nice, simple, mechanical way. As we shall see, this dogma created terrible difficulties for the study of regeneration.

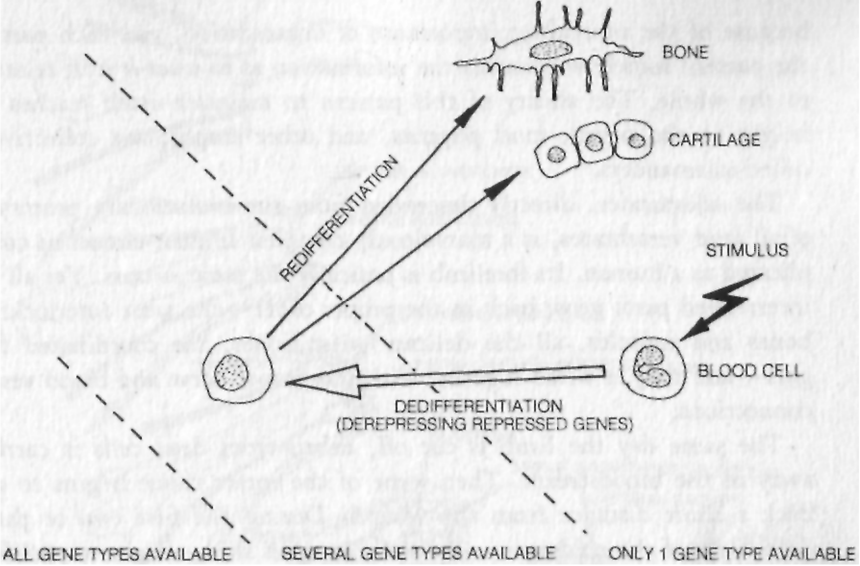
BONE
CARTILAGE
REDIFFERENTIATION
STIMULUS
BLOOD CELL
DEDIFFERENTIATION
(DEREPRESSING REPRESSED GENES)
ALL GENE TYPES AVAILABLE
SEVERAL GENE TYPES AVAILABLE
ONLY 1 GENE TYPE AVAILABLE

DEDIFFERENTIATE—FROM SPECIALIZED CELLS BACK TO NEO-EMBRYONIC CELLS

Control Problems

After Morgan's work on salamander limb regrowth early in this century, hundreds of other experimenters studied the miracle again and again in many kinds of animals. Their labors revealed a number of general principles, such as:

Polarity. A creature's normal relationships of front to back and top to bottom are preserved in the regenerate.

Gradients. Regenerative ability is strongest in one area of an animal's body, gradually diminishing in all directions.

- Dominance. Some one particular section of the lost part is replaced first, followed by the others in a fixed sequence.
- Induction. Some parts actively trigger the formation of others later in the sequence.
- Inhibition. The presence of any particular part prevents the formation of a duplicate of itself or of other parts that come before that part in the sequence.

All the experiments led to one unifying conclusion: The overall structure, the shape, the pattern, of any animal is as real a part of its body as are its cells, heart, limbs, or teeth. Living things are called organisms because of the overriding importance of organization, and each part of the pattern somehow contains the information as to what it is in relation to the whole. The ability of this pattern to maintain itself reaches its height in the newts, mud puppies, and other amphibians collectively called salamanders.

The salamander, directly descended from the evolutionary prototype of all land vertebrates, is a marvelously complex animal, almost as complicated as a human. Its forelimb is basically the same as ours. Yet all its interrelated parts grow back in the proper order—the same interlocking bones and muscles, all the delicate wrist bones, the coordinated fingers—and they're wired together with the proper nerve and blood vessel connections.

The same day the limb is cut off, debris from dead cells is carried away in the bloodstream. Then some of the intact tissue begins to die back a short distance from the wound. During the first two or three days, cells of the epidermis—the outer layer of skin—begin to proliferate and migrate inward, covering the wound surface. The epidermis then thickens over the apex of the stump into a transparent tissue called the apical cap. This stage is finished in about a week.

By then, the blastema, the little ball of undifferentiated cells described by Morgan, has started to appear beneath the apical cap. This is the "organ" of regeneration, forming on the wound like a miniature embryo and very similar to the embryonic limb bud that gave rise to the leg in the first place. Its cells are totipotent, able to develop into all the different kinds of cells needed to reconstitute the limb.

The blastema is ready in about two weeks. Even as it's forming, the cells at its outer edge start dividing rapidly, changing the blastema's shape to a cone and providing a steady source of raw material—new cells—for growth. After about three weeks, the blastema cells at the inner edge begin to differentiate into specialized types and arrange themselves into tissues, beginning with a cartilage collar around the old bone shaft. Other tissues then form, and the new limb—beginning with a characteristic paddle shape that will become the hand—appears as though out of a mist. The elbow and long parts of the limb coalesce behind the hand, and the regrowth is complete (except for some slight enlargement) when the four digits reappear after about eight weeks.

This process, exquisitely beautiful and seemingly simple, is full of problems for biology. What organizes the growth? What is the control factor? How does the blastema "know" that it must make a foreleg instead of a bind leg? (The salamander never makes a mistake.) How does all the information about the missing parts get to these undifferentiated

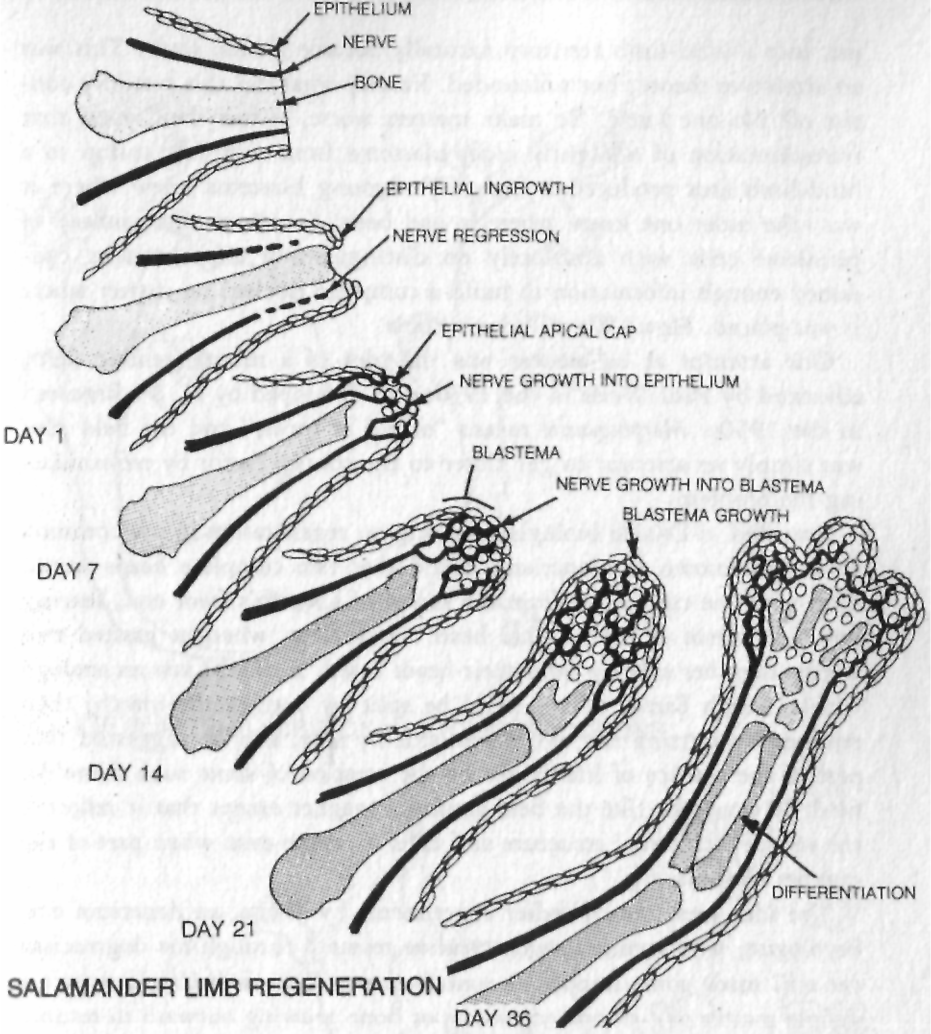
EPITHELIUM
NERVE
BONE
EPITHELIAL INGROWTH
NERVE REGRESSION
EPITHELIAL APICAL CAP
NERVE GROWTH INTO EPITHELIUM
DAY 1
BLASTEMA
NERVE GROWTH INTO BLASTEMA
BLASTEMA GROWTH
DAY 7
DAY 14
DIFFERENTIATION
DAY 21
DAY 36

SALAMANDER LIMB REGENERATION

cells, telling them what to become, which genes to activate, what proteins to make, where to position themselves? It's as if a pile of bricks were to spontaneously rearrange itself into a building, becoming not only walls but windows, light sockets, steel beams, and furniture in the process.

Answers were sought by transplanting the blastema to other positions on the animal. The experiments only made matters worse. If the blastema was moved within live to seven days after it first appeared, and grafted near the hind leg, it grew into a second hind leg, even though it came from an amputated foreleg. Well, that was okay. The body could be divided into "spheres of influence" or "organizational territories," each of which contained information on the local anatomy. A blastema put into a hind-limb territory naturally became a hind limb. This was an attractive theory, but unfounded. Exactly what did this territory consist of? No one knew. To make matters worse, it was then found that transplantation of a slightly older blastema from a foreleg stump to a hind-limb area produced a foreleg. The young blastema knew where it was; the older one knew where it had been! Somehow this pinhead of primitive cells with absolutely no distinguishing characteristics contained enough information to build a complete foreleg, no matter where it was placed. How? We still don't know.

One attempt at an answer was the idea of a morphogenetic field, advanced by Paul Weiss in the 1930s and developed by H. V. Bronsted in the 1950s. Morphogenesis means "origin of form," and the field idea was simply an attempt to get closer to the control factor by reformulating the problem.

Bronsted, a Danish biologist working on regeneration in the common flatworms known as planarians, found that two complete heads would form when he cut a strip from the center of a worm's front end, leaving two side pieces of the original head. Conversely, when he grafted two worms together side by side, their heads fused. Br0nsted saw an analogy with a match flame, which could be split by cutting the match, then rejoined by putting the two halves side by side, and he suggested that part of the essence of life might be the creation of some such flamelike field. It would be like the field around a magnet except that it reflected the magnet's internal structure and held its shape even when part of the magnet was missing.

The idea grew out of earlier experiments by Weiss, an American embryologist, who stymied much creative research through his dogmatism yet still made some important contributions. Regrowth clearly wasn't a simple matter of a truncated muscle or bone growing outward to resume its original shape. Structures that were missing entirely—the hand, wrist, and bones of the salamander's lower forelimb, for example—also reappeared. Weiss found that redundant parts could be inserted, but the essential ones couldn't easily be eliminated. If an extra bone was implanted in the limb and the cut made through the two, the regenerate contained both. However, if a bone was completely removed and the incision allowed to heal, and the limb was then amputated through what would have been the middle of the missing bone, the regenerate produced that bone's lower half, like a ghost regaining its substance Weiss suggested that other tissues besides bone could somehow project a field that included the arrangement of the bones. As a later student of regeneration, Richard Goss of Brown University, observed, "Apparently

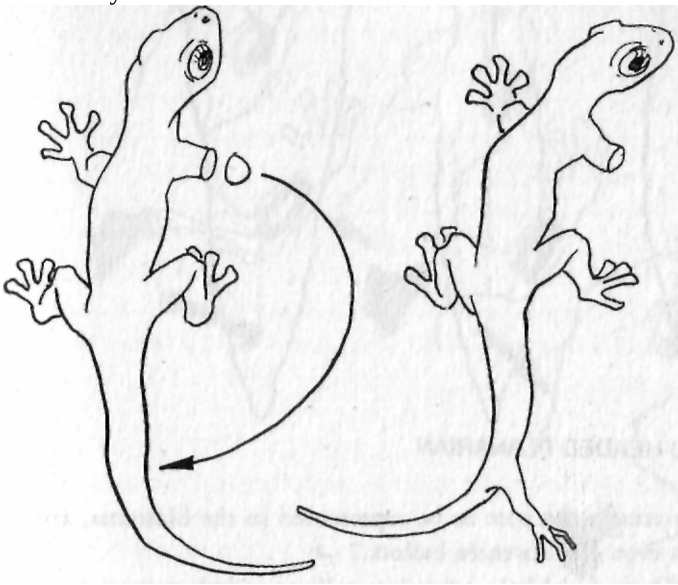

THE LATE BLASTEMA ALREADY HAS ITS INSTRUCTIONS

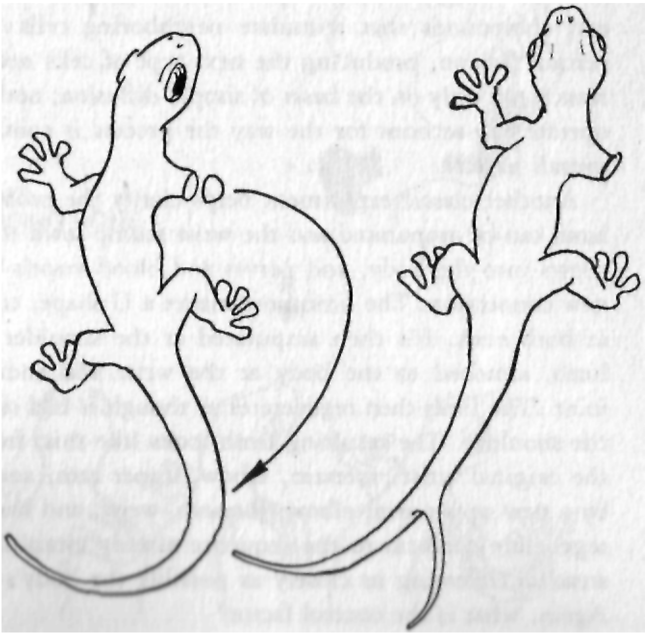

THE EARLY BLASTEMA GETS INSTRUCTIONS FROM NEARBY TISSUES

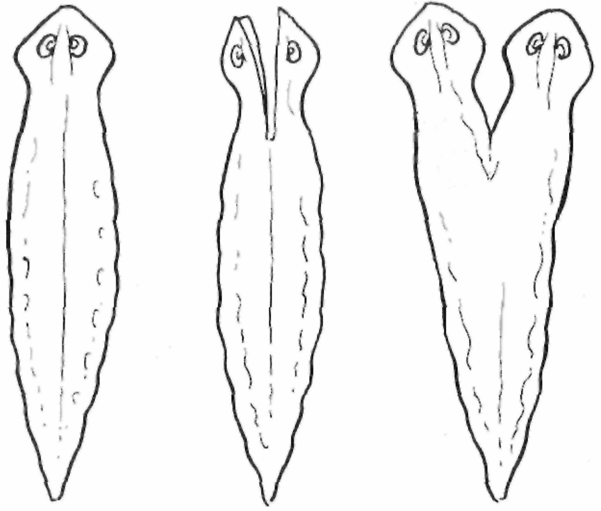

FORMING A TWO-HEADED PLANARIAN

each tissue of the stump can vote to be represented in the blastema, and some of them can even cast absentee ballots."

Any such field must be able to stimulate cells to switch various genes on and off, that is, to change their specialization. A large body of research on embryonic development has identified various chemical inducers, compounds that stimulate neighboring cells to differentiate in a certain fashion, producing the next type of cells needed. But these substances act only on the basis of simple diffusion; nothing in the way they operate can account for the way the process is controlled to express the overall pattern.

Another classic experiment helps clarify the problem. A salamander's hand can be amputated and the wrist stump sewn to its body. The wrist grows into the body, and nerves and blood vessels link up through the new connection. The limb now makes a U shape, connected to the body at both ends. It's then amputated at the shoulder to make a reversed limb, attached to the body at the wrist and ending with a shoulder joint. The limb then regenerates as though it had simply been cut off at the shoulder. The resulting limb looks like this: from the body sprouts the original wrist, forearm, elbow, upper arm, and shoulder, followed by a new upper arm, elbow, forearm, wrist, and hand. Why doesn't the regenerate conform to the sequence already established in this limb instead of following as closely as possible the body's pattern as a whole? Again, what is the control factor?

Information, and a monumental amount of it, is clearly passed from the body to the blastema. Our best method of information processing at

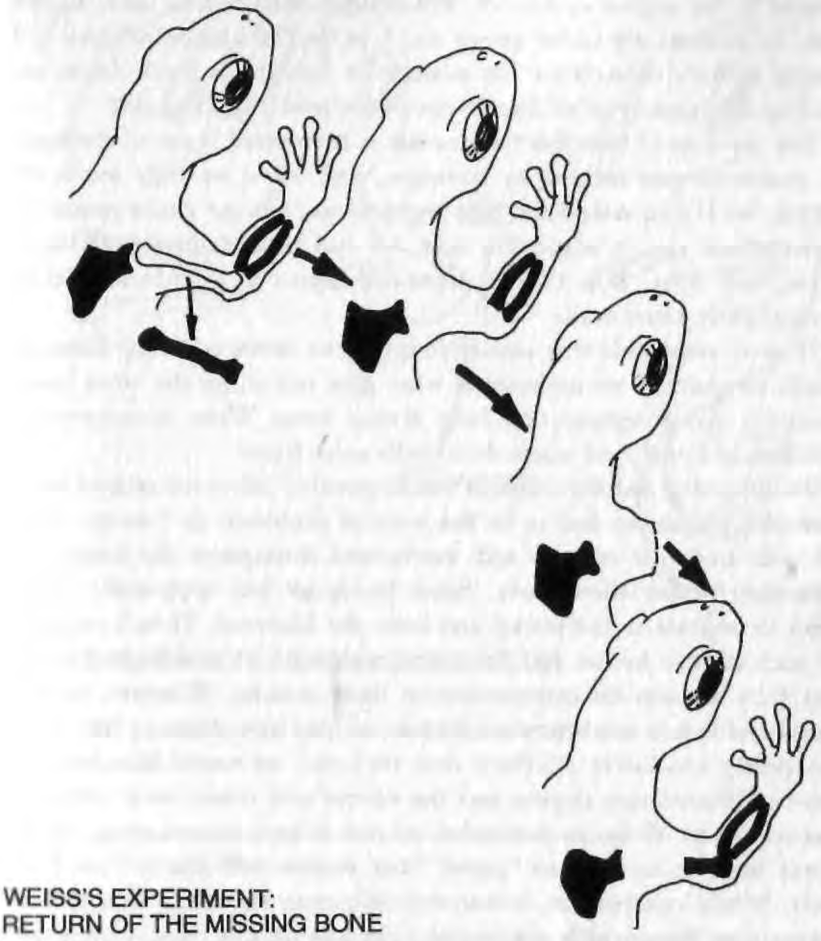

WEISS'S EXPERIMENT:
RETURN OF THE MISSING BONE

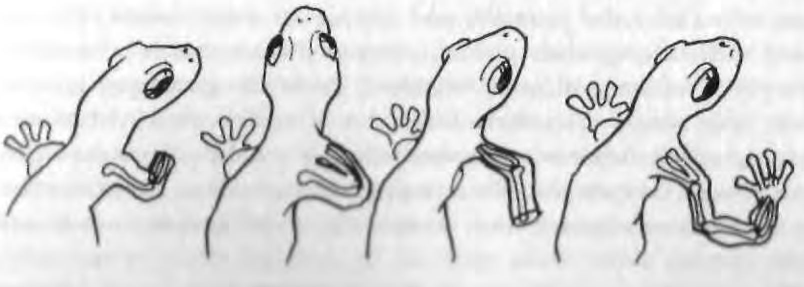

THE REVERSED-ARM EXPERIMENT

present is the digital computer, which deals with bits of data, signals that, in essence, say either yes or no, 1 or 0. The number of such bits needed to fully characterize the salamander forelimb is incalculable, exceeding the capacity of all known computers operating in unison.

The question of how this information is transferred is one of the hardest problems ever tackled by scientists, and when we fully know the answer, we'll understand not only regeneration but the entire process of growth from egg to adult. For now, we had best, as biologists themselves have done, skip this problem and return to it after addressing some slightly easier ones.

It seems reasonable that understanding what comes out of the blastema would be easier if we understood what goes into it, so the other major questions about regeneration have always been: What stimulates the blastema to form? And where do its cells come from?

The idea that dedifferentiation was impossible led to the related belief that all regeneration had to be the work of neoblasts, or "reserve cells" left over from the embryo and warehoused throughout the body in a primitive, unspecialized state. Some biological bell supposedly called them to migrate to the stump and form the blastema. There's evidence for such cells in hydras and flatworms, although it's now doubtful that they fully account for regeneration in these animals. However, no one ever found any in a salamander. In fact, as long ago as the 1930s, there was nearly conclusive evidence that they did not exist. Nevertheless, anti-dedifferentiation dogma and the reserve cell theory were defended fanatically, by Weiss in particular, so that many unconvincing experiments were interpreted to "prove" that reserve cells formed the blastema. When I started out, it was very dangerous for one's career even to suggest that mature cells might create the blastema by dedifferentiating.

Because it was so hard to imagine how a blastema could arise without dedifferentiation, the idea later developed that perhaps cells could partially dedifferentiate. In other words, perhaps muscle cells could become cells that looked primitive and completely unspecialized, but that would then take up their previous lives as mature muscle cells after a brief period of amnesia in the blastema. To fit the square peg into the round hole, many researchers did a lot of useless work, laboriously counting cell divisions to try to show that the muscle cells in the stump made enough new muscle cells to supply the regenerate. The embarrassing blastema—enigmatic and completely undifferentiated was still there.

We now know (see Chapter 6) that at least some types of cells can revert completely to the primitive state and that such despecialization is

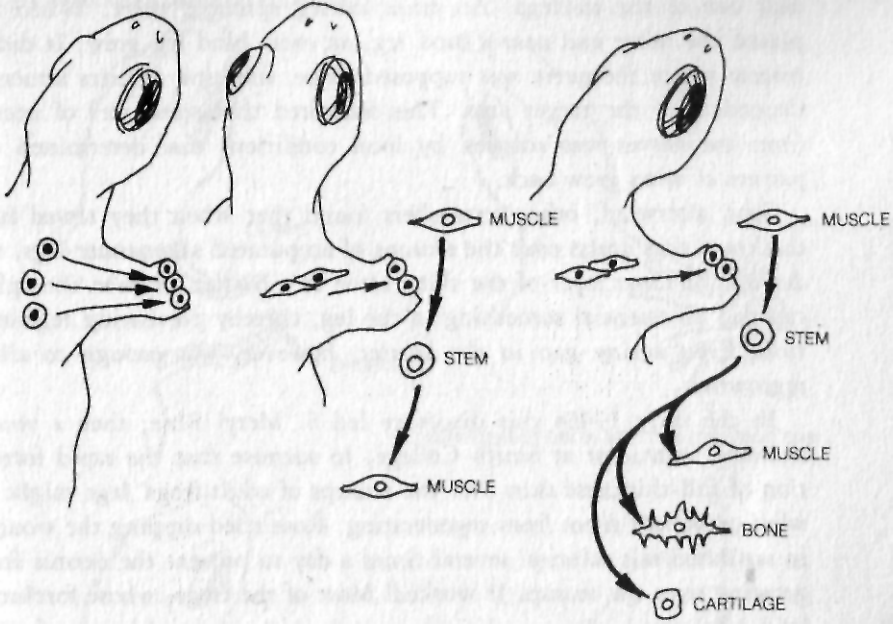

DEDIFFERENTIATE: PARTIAL OR COMPLETE?

the major, probably the only, way a blastema forms in complex animals like a salamander.

Nerve Connections

The other major question about the blastema's origin is: What triggers it? The best candidates for a "carrier" of the stimulus are the nerves. In complex multicellular animals, there's no regeneration without nerve tissue. Back in 1823, the English amateur Tweedy John Todd found that if the nerves into a salamander's leg were cut when the amputation was made, the limb wouldn't regrow. In fact, the stump itself shriveled up and disappeared. However, Todd got normal regeneration when he gave the nerves time to reconnect before severing the leg. Science wasn't ready to make anything of his observation, but many experiments since have confirmed it. Over a century later, Italian biologist Piera Locatelli showed that an extra leg would grow if a nerve was rerouted so that it ended near an intact leg. She cut the large sciatic nerve partway down the salamander's hind leg, leaving it attached to the spinal column and fully threading it up untder the skin so that its end touched the skin near one of the forelegs. An extra foreleg sprouted there. When she placed the nerve end near a hind leg, an extra hind leg grew. It didn't matter where the nerve was supposed to be; the kind of extra structure depended on the target area. This indicated that some sort of energy from the nerves was adapted by local conditions that determined the pattern of what grew back.

Soon afterward, other researchers found that when they sewed full-thickness skin grafts over the stumps of amputated salamander legs, the dermis, or inner layer of the skin, acted as a barrier between the apical cap and an essential something in the leg, thereby preventing regeneration. Even a tiny gap in the barrier, however, was enough to allow regrowth.

In the early 1940s this discovery led S. Meryl Rose, then a young anatomy instructor at Smith College, to surmise that the rapid formation of full-thickness skin over the stumps of adult frogs' legs might be what prevented them from regenerating. Rose tried dipping the wounds in saturated salt solution several times a day to prevent the dermis from growing over the stump. It worked! Most of the frogs, whose forelimbs he'd amputated between the elbow and wrist, replaced some of what they'd lost. Several regrew well-formed wrist joints, and a few even began to produce one new finger. Even though the replacements were incomplete, this was a tremendously important breakthrough, the first time any regeneration had been artificially induced in an animal normally lacking the ability. However, the dermis did grow over the stump, so the experiment worked by some means Rose hadn't expected.

Later, other investigators showed that in normal regeneration the apical cap, minus the dermis, was important because regrowing nerve fibers made unique connections with the epidermal cells in the first stage of the process, before the blastema appeared. These connections are collectively called the neuroepidermal junction (NEJ). In a series of detailed experiments, Charles Thornton of Michigan State University cut the nerves to salamander legs at various times before amputating the legs, then followed the progress of the regrowing nerves. Regeneration began only after the nerves had reached the epidermis, and it could be prevented by any barrier separating the two, or started by any breach in the barrier. By 1954 Thornton had proved that the neuroepidermal junction was the one pivotal step that must occur before a blastema could form and regeneration begin.

Shortly thereafter, Elizabeth D. Hay, an anatomist then working at Cornell University Medical College in New York, studied the neuroepidermal junction with an electron microscope. She found that as

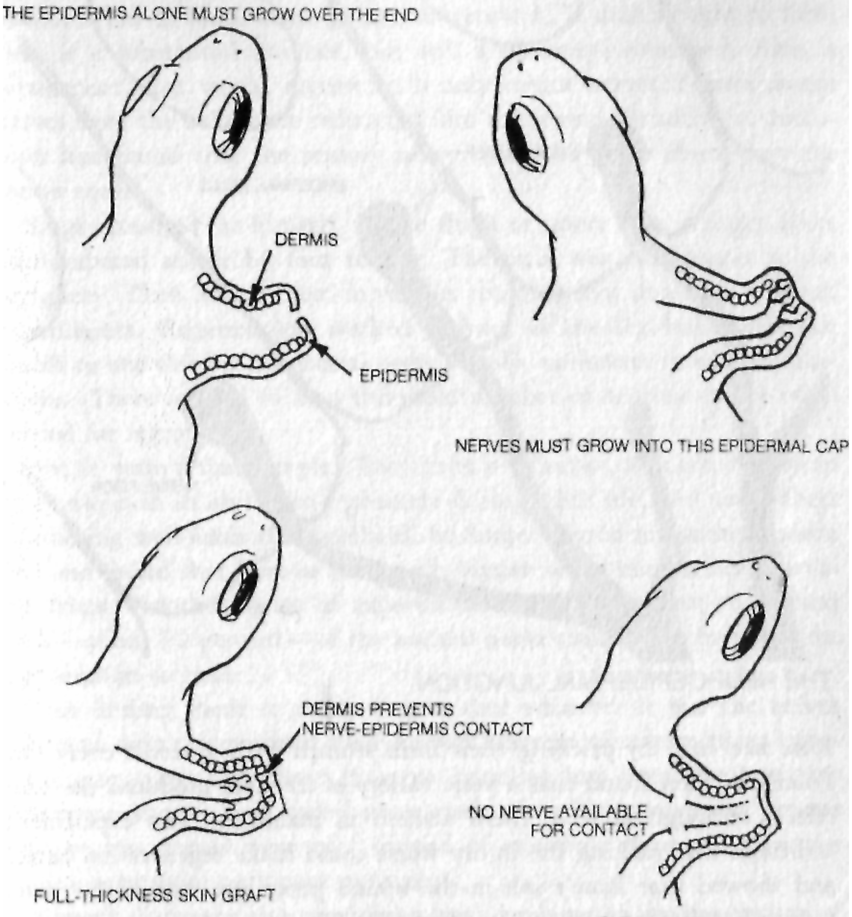
THE EPIDERMIS ALONE MUST GROW OVER THE END
DERMIS
EPIDERMIS
NERVES MUST GROW INTO THIS EPIDERMAL CAP
DERMIS PREVENTS NERVE-EPIDERMIS CONTACT
NO NERVE AVAILABLE FOR CONTACT
FULL-THICKNESS SKIN GRAFT
CUT NERVES

BLOCKING THE NEUROEPIDERMAL JUNCTION PREVENTS REGENERATION

each nerve fiber bundle reached the end of the stump, it broke up and each fiber went its separate way, snaking into the epidermis, which might be five to twenty cells thick. Each nerve fiber formed a tiny bulb at its tip, which was placed against an epidermal cell's membrane, nestling into a little pocket there. The arrangement was much like a synapse, although the microscopic structure wasn't as highly developed as in such long-term connections.

The junction was only a bridge, however. The important question was: What traffic crossed it?

In 1946, Lev Vladimirovich Polezhaev, young Russian biologist then working in London, concluded a long series of experiments in which he induced partial regeneration in adult frogs, the same success

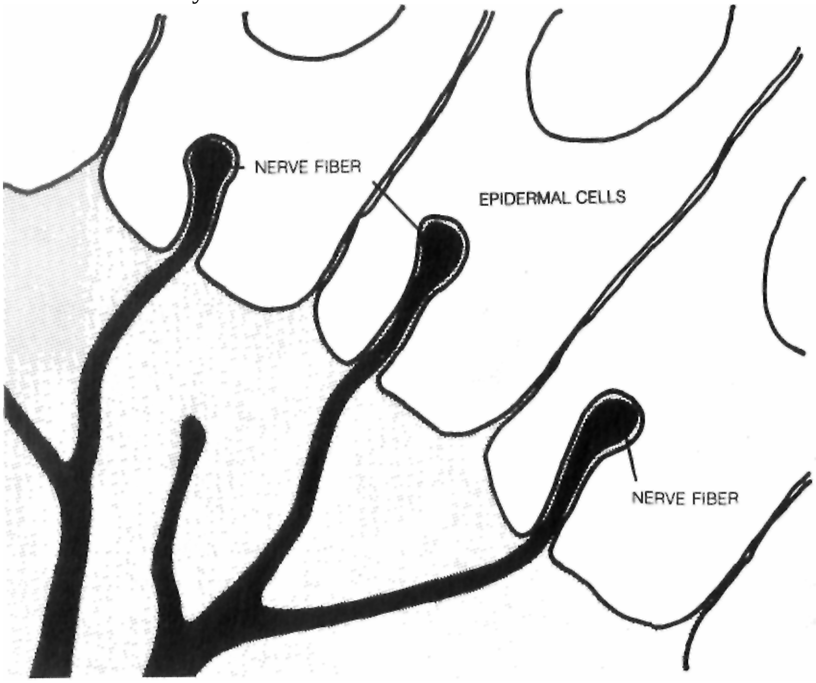

THE NEUROEPIDERMAL JUNCTION

Rose had had, by pricking their limb stumps with a needle every day. Polezhaev then found that a wide variety of irritants produced the same effect, although none of them worked in mammals. His experiments indicated that making the injury worse could make regeneration better, and showed that Rose's salt-in-the-wound procedure worked by irritation rather than by preventing dermis growth.

Next, the part that nerve tissue played was clarified considerably by Marcus Singer in a brilliant series of experiments at Harvard Medical School from the mid-1940s to the mid-1950s. Singer first confirmed Todd's long-forgotten work by cutting the nerves in salamander legs at various stages of regrowth, proving that the nerves were needed only in the first week, until the blastema was fully formed and the information transferred. After that, regeneration proceeded even if the nerves were cut.

Recent research had found that a salamander could replace its leg if all the motor nerves were cut, but not without the sensory nerves. Many assumed then that the growth factor was related only to sensory nerves, but Singer was uneasy over this conclusion: "The problem stated in advance that one or another nerve component is all important for regeneration." (Italics added.) Several facts didn't fit, however. Not only did the blastema fail to form when all nerves were cut, it didn't begin to form even if a substantial number, but still a minority, remained. Also, a salamander's leg would regrow with only motor nerves if extra motor nerves from the belly were redirected into the stump. In addition, zoologists had found that the sensory nerve contained more fibers than the motor nerve.

Singer counted for himself. In the thigh or upper arm, sensory fibers outnumbered motor by four to one. The ratio was even larger at the periphery. Then he cut them in various combinations in a long series of experiments. Regeneration worked as long as the leg had about one fourth to one third of its normal nerve supply, no matter in what combination. There seemed to be a threshold number of neurons (nerve cells) needed for regrowth.

But it wasn't that simple. The limbs of Xenopus, a South American frog unique in its ability to regenerate during adult life, had nerve fibers numbering well under the threshold. So Singer started measuring neuron size, and found that Xenopus had much bigger nerves than nonregenerating frogs. Another series of experiments verified the link: A critical mass—about 30 percent—of the normal nerve tissue must be intact for regeneration to ensue.

This finding made it pretty certain that whatever it was the nerves delivered didn't come from their known function of transmitting information by nerve impulses. If nerve impulses had been involved, regeneration should have faded away gradually with greater and greater flaws as the nerves were cut, instead of stopping abruptly when the minimum amount no longer remained.

Singer's discovery also provided a basic explanation for the decline of regeneration with increasing evolutionary complexity. The ratio between body mass and total nerve tissue is about the same in most animals, but more and more nerve became concentrated in the brain (a process called encephalization) as animals became more complex. This diminished the amount of nerve fiber available for stimulating regeneration in peripheral parts, often below the critical level.

In the early 1950s, Singer applied what he'd learned to the nonregenerating adult bullfrog. Using Locatelli's method, he dissected the sciatic nerve out of the hind leg, leaving it attached to the spinal cord, and directed it under the skin to the foreleg amputation stump. In two or three weeks, blastemas had formed, and the cut legs were restored to about the same degree as in Rose's and Polezhaev's experiments.

By 1954 Singer was ready to look for a growth-inducing chemical that was presumed to be coming from the nerves. The most promising

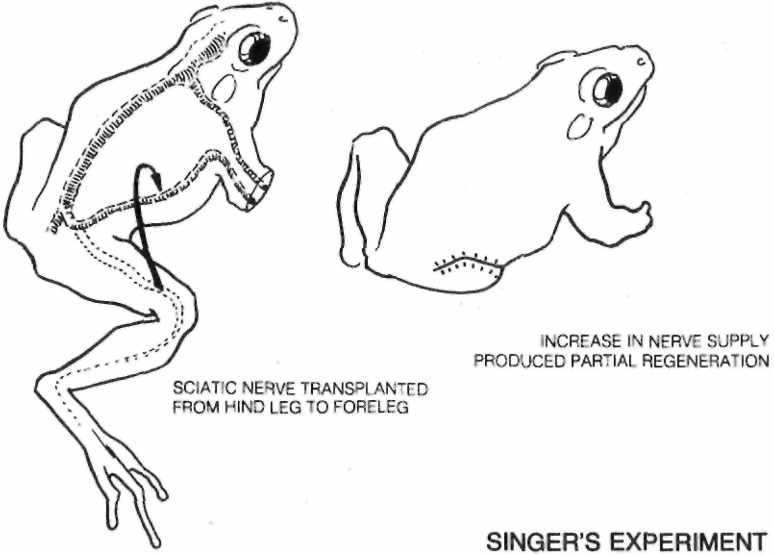

SINGER'S EXPERIMENT

possibility was the neurotransmitter acetylcholine, one of several compounds known to relay nerve impulses across synapses. The nerves secreted acetylcholine more abundantly than normal during blastema formation—just when nerve supply was crucial—and its production fell back to normal when regrowth was well under way. Singer had studied previous failures with acetylcholine, in which experimenters had rubbed it on the stump or injected it into the blastema. He thought these methods were too artificial, so he invented a microinfusion apparatus to release tiny amounts of acetylcholine continually, just as the nerves did. It used a clock motor to drip the hormone slowly through a needle into the shoulder of an anesthetized animal in which the leg nerves had been removed. He had trouble keeping the drugged salamanders alive, so maybe the anesthetic affected the outcome, but even the ones that survived didn't regenerate at all. The growth factor was almost certainly not acetylcholine.

Vital Electricity

These, then, were the shoulders on which I stood in 1958 as I began to look for the pattern-control and blastema-stimulating factors in regeneration. At that time we knew of two things that could yield some regrowth in nonregenerators: extra nerve and extra injury. How were they related? Luck gave me a clue.

I began my work just after the first few Sputniks, during the "missile gap" flap. Alarmed by the unforeseen triumphs of Russian technology, which we'd considered primitive, the government hastily began translating every Soviet scientific journal and distributing copies free to federally funded research centers. Suddenly, the medical library at the VA Medical Center in Syracuse, where I worked, began receiving each month a crate of Russian journals on clinical medicine and biology. Since no one else was much interested, this bonanza was all for me.

I soon made two discoveries: The Russians were willing to follow hunches; their researchers got government money to try the most outlandish experiments, ones that our science just knew couldn't work. Furthermore, Soviet journals published them—even if they did work. I particularly enjoyed Biofizika, the Soviet journal of biophysics, and it was there I encountered a paper on the "Nature of the Variation of the Bioelectric Potentials in the Regeneration Process of Plants," by A. M. Sinyukhin of Lomonosov State University in Moscow.

Sinyukhin began by cutting one branch from each of a series of tomato plants. Then he took electrical measurements around the wound as each plant healed and sent out a new shoot near the cut. He found a negative current—a stream of electrons—flowing from the wound for the first few days. A similar "current of injury" is emitted from all wounds in animals. During the second week, after a callus had formed over the wound and the new branch had begun to form, the current became stronger and reversed its polarity to positive. The important point wasn't the polarity—the position of the measuring electrode with respect to a reference electrode often determines whether a current registers as positive or negative. Rather, Sinyukhin's work was significant because he found a change in the current that seemed related to reparative growth. Sinyukhin found a direct correlation between these orderly electrical events and biochemical changes: As the positive current increased, cells in the area more than doubled their metabolic rate, also becoming more acidic and producing more vitamin C than before.

Sinyukhin then applied extra current, using small batteries, to a group of newly lopped plants, augmenting the regeneration current. These battery-assisted plants restored their branches up to three times faster than the control plants. The currents were very small—only 2 to 3 microamperes for five days. (An ampere is a standard unit of electric current, and a microampere is one millionth of an ampere.) Larger amounts of electricity killed the cells and had no growth-enhancing effect. Moreover, the polarity had to match that normally found in the plant. When Sinyukhin used current of the opposite polarity, nullifying the plant's own current, restitution was delayed by two or three weeks.

To American biology, however, this was all nonsense. To understand why, we must backtrack for a while.

Luigi Galvani, an anatomy professor in the medical school at the University of Bologna who'd been studying electricity for twenty years, first discovered the current of injury in 1794, but unfortunately he didn't know it.

At that time, biology's main concern was the debate between vitalism and mechanism. Vitalism, though not always called by that name, had been the predominant concept of life since prehistoric times throughout the world, and it formed the basis for almost all religions. It was closely related to Socrates' and Plato's idea of supernatural "forms" or "ideals" from which all tangible objects and creatures derived their individual characteristics. Hippocrates adapted this idea by postulating an anima as the essence of life. The Platonic concept evolved into the medieval philosophy of realism, whose basic tenet was that abstract universal principles were more real than sensory phenomena. Mechanism grew out of Aristotle's less speculative rationalism, which held that universal principles were not real, being merely the names given to humanity's attempts at making sense of the reality apprehended through the senses. Mechanism had become the foundation of science through the writings of Descartes in the previous century, although even he believed in an "animating force" to give the machine life at the outset. By Galvani's time, mechanism's influence was steadily growing.

Galvani was a dedicated physician, and medicine, tracing its lineage back to tribal shamans, has always been a blend of intuition and empirical observation based on a vitalistic concept of the sanctity of life. The vitalists had long tried—unsuccessfully—to link the strange, incorporeal phenomenon of electricity with the elan vital. This was Galvani's main preoccupation.

One day he noticed that some frogs' legs he'd hung in a row on his balustrade, pending his dinner, twitched whenever the breeze blew them against the ironwork. At about the same time his wife Lucia noticed in his laboratory that the muscles of a frog's leg contracted when an assistant happened to be touching the main nerve with a steel scalpel at the same instant that a spark leaped from one of the electrical machines being operated across the room. (The only type of electricity then known was the static type, in the form of sparks from various friction devices.) Today we know that an expanding and collapsing electric field generated by the spark induced a momentary current in the scalpel, which stimulated the muscle, but Galvani believed that the metal railing and scalpel had drawn forth electricity hidden in the nerves.

Galvani experimented for years with nerves from frogs' legs, connected in various circuits with several kinds of metals. He grew convinced that the vital spirit was electricity flowing through the nerves and announced this to the Bologna Academy of Science in 1791.

Within two years, Alesandro Volta, a physicist at the University of Padua, had proven that Galvani had in fact discovered a new kind of electricity, a steady current rather than sparks. He'd generated a bimetallic direct current, a flow of electrons between two metals, such as the copper hooks and iron railing of the famous balcony observation, connected by a conducting medium—in other words, a battery. The frogs' legs, being more or less bags of weak salt solution, were the electrolyte, or conducting medium. They were otherwise incidental, Volta explained, and there was no such thing as Galvani's "animal electricity."

Galvani, a shy and thoroughly noncombative soul, was crushed. His only response was an anonymous paper in 1794 describing several experiments in which frogs' legs could be made to twitch with no metal in the circuit. In one procedure, the experimenter touched one leg nerve with the frog's dissected-out, naked spinal cord, while holding the other leg to complete the circuit. Here the current was true animal electricity, coming from the amputation wound at the base of the leg.

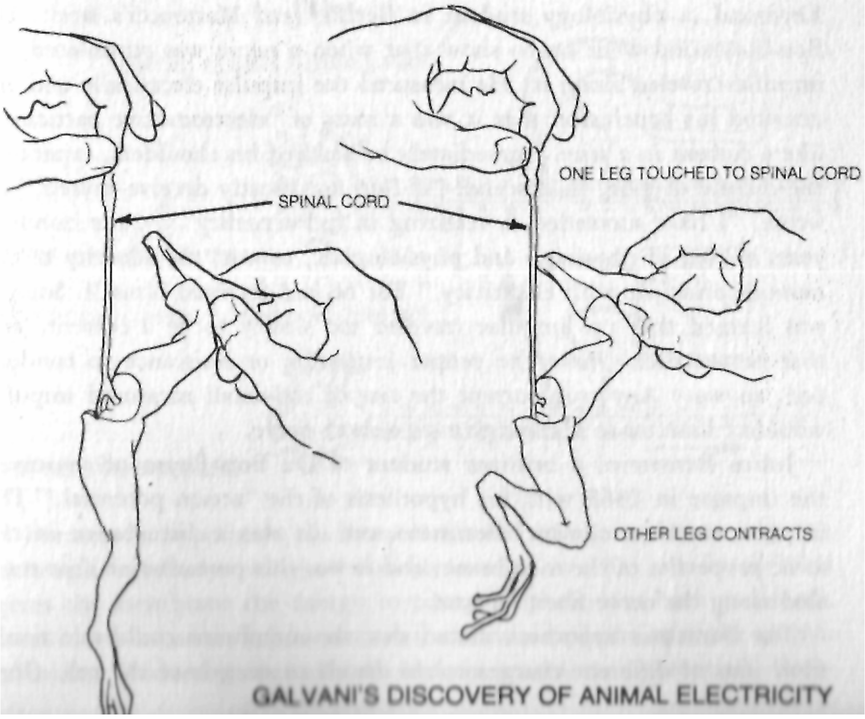

GALVANI'S DISCOVERY OF ANIMAL ELECTRICITY

In the long run, Galvani unwittingly helped the cause of the mechanists by giving them something to attack. As long as the elan vital was ephemeral, all you could say was that you couldn't find it. Once Galvani said it was electricity, a detectable, measurable entity, there was a target for experimentation. Actually, Baron Alexander von Humboldt, the explorer-naturalist who founded geology, proved in 1797 that Volta and Galvani were both partly right. Bimetallic currents were real, but so was spontaneous electricity from injured flesh. However, the mechanists had the upper hand; Galvani's anonymous report and Humboldt's confirmation were overlooked. Galvani himself died penniless and brokenhearted in 1798, soon after his home and property were confiscated by the invading French, while Volta grew famous developing his storage batteries under the auspices of Napoleon.

Then in the 1830s a professor of physics at Pisa, Carlo Matteucci, using the newly invented galvanometer, which could measure fairly small direct currents, came up with other evidence for animal electricity. In a meticulous series of experiments lasting thirty-five years, he conclusively proved that the current of injury was real. However, he didn't find it in the nervous system, only emanating from the wound surface, so it couldn't be firmly related to the vital force.

The tale took another turn in the 1840s when Emil Du Bois-Reymond, a physiology student in Berlin, read Matteucci's work. Du Bois-Reymond went on to show that when a nerve was stimulated, an impulse traveled along it. He measured the impulse electrically and announced his conclusion that it was a mass of "electromotive particles," like a current in a wire. Immediately he squared his shoulders, expecting the mantle of glory to descend: "If I do not greatly deceive myself," he wrote, "I have succeeded in realizing in full actuality . . . the hundred years' dream of physicists and physiologists, to wit, the identity of the nervous principle with electricity." But he had deceived himself. Soon it was learned that the impulse traveled too slowly to be a current, and that nerves didn't have the proper insulation or resistance to conduct one, anyway. Any true current the size of the small measured impulse wouldn't have made it through even a short nerve.

Julius Bernstein, a brilliant student of Du Bois-Reymond, resolved the impasse in 1868 with his hypothesis of the "action potential." The impulse wasn't a current, Bernstein said. It was a disturbance in the ionic properties of the membrane, and it was this perturbation that traveled along the nerve fiber, or axon.

The Bernstein hypothesis stated that the membrane could selectively filter ions of different charges to the inside or outside of the cell. (Ions are charged particles into which a salt breaks up when dissolved in water; all salts dissociate in water into positive and negative ions, such as the positive sodium and negative chloride ions of table salt.) Bernstein postulated that the membrane could sort most of the negatives outside and most of the positives inside the fiber. The membrane was polarized (with like charges grouped on one side), having a transmembrane potential, because the negative charges, all on one side, could potentially flow in a current across the membrane to achieve a balance on both sides. This was what happened in a short segment of the membrane whenever a nerve was stimulated. Part of the membrane became depolarized, reversing the transmembrane potential. The nerve impulse was actually a disturbance in the potential traveling along the membrane. As the area of disturbance moved along, the membrane quickly restored its normal resting potential. Thus the nerve impulse wasn't an electrical current, even though it could be measured electrically.

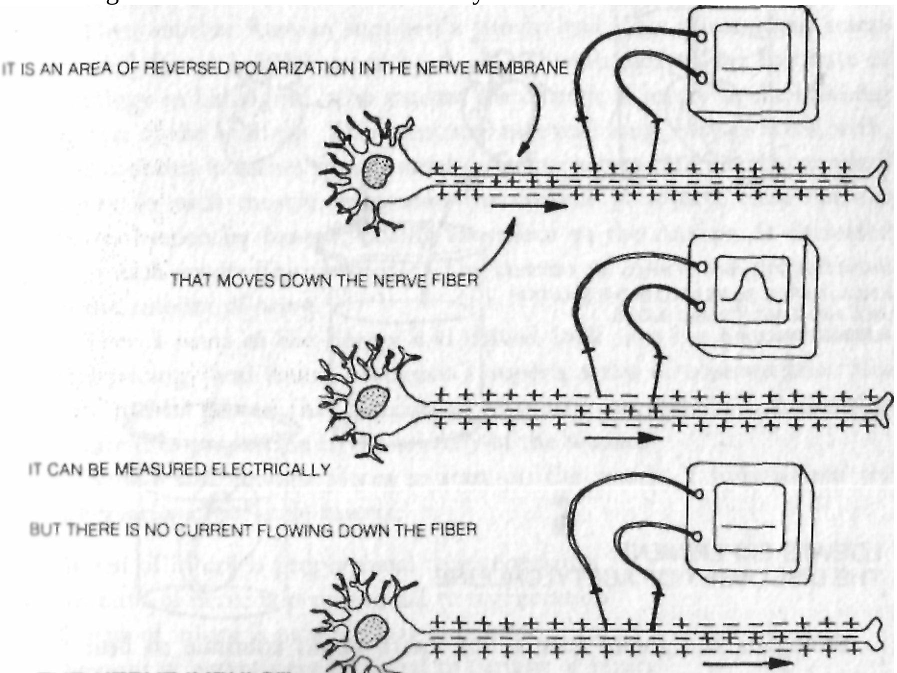
IT IS AN AREA OF REVERSED POLARIZATION IN THE NERVE MEMBRANE
THAT MOVES DOWN THE NERVE FIBER
IT CAN BE MEASURED ELECTRICALLY
BUT THERE IS NO CURRENT FLOWING DOWN THE FIBER

THE NERVE IMPULSE

Bernstein's hypothesis has been confirmed in all important respects, although it remains I hypothesis because no one has yet found what gives the membrane the energy to pump all those ions back and forth. Soon it was broadened, however, to include an explanation of the current of injury. Reasoning that all cells had transmembrane potentials, Bernstein maintained that, after injury, the damaged cell membranes simply leaked their ions out into the environment. Thus the current of injury was no longer a sign that electricity was central to life, but only an uninteresting side effect of cell damage.

The vitalists, with their hopes pinned on electricity, kept getting pushed into tighter and tighter corners as electricity was removed from one part of the body after another. Their last stand occurred with the discovery of neurotransmitters. They'd maintained that only an electrical current could jump across the synapse, the gap between communicating nerves. In 1920 that idea was disproven with a lovely experiment by Otto Loewi, a research professor at the NYU School of Medicine, later my alma mater. When I took physiology in my first year there, we had to duplicate his experiment.

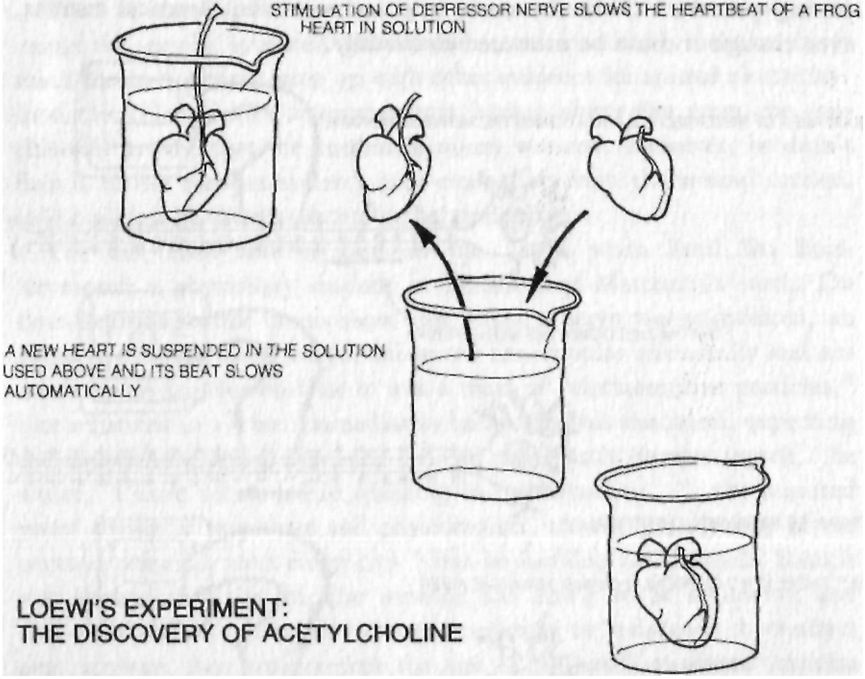
STIMULATION OF DEPRESSOR NERVE SLOWS THE HEARTBEAT OF A FROG HEART IN SOLUTION
A NEW HEART IS SUSPENDED IN THE SOLUTION USED ABOVE AND ITS BEAT SLOWS AUTOMATICALLY

LOEWI'S EXPERIMENT:
THE DISCOVERY OF ACETYLCHOLINE

Biologists had found that a frog heart would continue to beat for several days when removed with its nerves and placed in an appropriate solution. Stimulating one of the nerves would slow it down. Like Loewi, we took one such heart, with nerve attached, and stimulated the nerve, slowing the beat. We then collected the solution baching that heart and placed another heart in it. Its beat slowed even though its depressor nerve hadn't been stimulated. Obviously the nerve slowed the heartbeat by producing a chemical, which crossed the gap between the nerve ending and the muscle fiber. This chemical was later identified as acetylcholine, and Loewi was awarded the Nobel Prize in 1936 for this discovery. His work resulted in the collapse of the last vestige of electrical vitalism. Thereafter, every function of the nervous system had to be explained on the basis of the Bernstein hypothesis and chemical transmission across the synapse.

It was with great trepidation, therefore, that I put any credence in Sinyukhin's report that the strength of the injury current affected regeneration in his plants. Yet his report was detailed and carefully written. Something about his work gave me a gut feeling that it was valid. Maybe it was because the tomato plants he used were "Best of All" American Beauties. At this point I wasn't aware of Matteucci's forgotten work, but something clicked in my mind now as I studied Rose's and Polezhaev's experiments. In both, definitely in Pole2haev's and probably in Rose's, regeneration had been stimulated by an increase in the injury.

Then another Russian supplied a timely lead. In a government translation I found a 1958 paper by A. V. Zhirmunskii of the Institute of Cytology in Leningrad, who studied the current of injury in the hind leg muscle of the bullfrog. This muscle is nice and long, easy to work with, and contains branches from several different nerves. He made a standard injury in each muscle, measured the current of injury, then cut the nerves branch by branch, noting the effect on the current. It decreased with each succeeding nerve cut. The current of injury was proportional to the amount of nerve.

Then I went to the library and delved back into the history of neurophysiology and found Matteucci's superb series of observations. Not only had he proven that the current of injury was real, he'd shown that it varied in proportion to the severity of the wound.

Now I had enough pieces to start on the puzzle. I summarized the observations in a little matrix:

Extent of injury is proportional to regeneration
Amount of nerve is proportional to regeneration
Extent of injury is proportional to current of injury
Amount of nerve is proportional to current of injury
Ergo: current of injury is proportional to regeneration

I was pretty sure now chat, contemporary "knowledge" to the contrary, the current of injury was no side effect and was the first place to look for clues to the growth control and dedifferentiation-stimulating factors. I planned my first experiment.

Three

The Sign of the Miracle

Real science is creative, as much so as painting, sculpture, or writing. Beauty, variously defined, is the criterion for art, and likewise a good theory has the elegance, proportion, and simplicity that we find beautiful. Just as the skilled artist omits the extraneous and directs our attention to a unifying concept, so the scientist strives to find a relatively simple order underlying the apparent chaos of perception. Perhaps because it was mine, my theory that the current of injury stimulated regeneration seemed both simple and beautiful. It's impossible to convey the sense of excitement I felt when all of the facts fell together and the idea came. I'd created something new that explained the previously inexplicable. I couldn't wait to see if I was right.

In all the time that the Bernstein hypothesis had been used to explain away the current of injury, no one had ever thought to measure the current over a period of days to see how long it lasted. If it was only ions leaking from damaged cells, it should disappear in a day or two, when these cells had finished dying or repairing themselves. This simple measurement, with a comparison of the currents in regenerating versus nonregenerating limbs, was what I planned to do. I would uniformly amputate the forelegs of frogs and salamanders. Then, as the frogs' stumps healed over and the salamanders' legs redrew, I would measure the currents of injury each day.

The experiment itself was as simple as could be. The tricky part was
getting peremission to do it.

When you want to do a research project, there are certain channels you must go through to get the money. You write a project proposal, spelling out what hypothesis you want to test, why you think it should be done, and how you plan to go about it. The proposal goes to a committee supposedly composed of your peers, people who have demonstrated competence in related research. If they approve your project and the money is available, you generally get part of what you asked for, enough to get started.

The Veterans Administration had been dispensing research money for several years as a sort of bribe to attract doctors despite the low pay in government service. The money from Washington was doled out by the most influential doctors on the staff, not necessarily the best researchers, but I still felt I had a good chance because the VA was having an especially hard time recruiting orthopedists. Moreover, my hypothesis was based on the work of Rose, Polezhaev, Singer, Sinyukhin, and Zhirmunskii with inescapable logic. And since frogs and salamanders were anatomically similar, any difference in their currents of injury should reflect the disparity in their powers of regeneration. My chances of being thrown off by extraneous factors were thus minimal.

I remember thinking, as I wrote the proposal, how my life had come full circle. As a college freshman in 1941, I'd conducted a crude experiment on salamanders, showing that thyroid stimulation by iodine didn't speed up regeneration. Here I was nearly twenty years later, beneficiary of the intervening research, hoping to add to our knowledge of the same phenomenon and perhaps even discover something that would help human patients. I worried that my roundabout course might weigh against me, since one of the criteria for grants was whether the investigator had been trained for that particular field. This proposal would have been expected from a physiologist, not an orthopedist. Nevertheless, I was asking for a relatively minuscule amount of money. I needed only a thousand dollars to put together the equipment, so I didn't anticipate much trouble.

The Tribunal

"Dr. Becker, could you please come to a special research committee meeting in one hour?" The committee's secretary was calling. I'd known something was up, two months had passed since I'd filed my proposal, and all my queries as to its fate had gone unanswered.

"I'll be there."

"It's not here in the research office. It's downstairs in the hospital director's office." Now that was really strange. The director almost never paid any attention to the research program. Besides, his office was big enough to hold a barbecue in.

It was a barbecue, ali right, and I was the one being grilled. The director's conference room had been rearranged. In place of the long, polished table there was a semicircle of about a dozen chairs, each occupied by one of the luminaries from the hospital and medical school. I recognized the chairmen of the departments of biochemistry and physiology along with the hospital director and chief of research. Only the dean was missing. In the center was a single chair—for me.

The spokesman came right to the point: "We have a very grave basic concern over your proposal. This notion that electricity has anything to do with living things was totally discredited some time ago. It has absolutely no validity, and the new scientific evidence you're citing is worthless. The whole idea was based on its appeal to quacks and the gullible public. I will not stand idly by and see this medical school associated with such a charlatanistic, unscientific project." Murmurs of assent spread around the group.

I had the momentary thrill of imagining myself as Galileo or Giordano Bruno; I thought of walking to the window to see if the stake and fagots were set up on the lawn. Instead I delivered a terse speech to the effect that I still thought my hypothesis was stoutly supported by some very good research and that I was sorry if it flew in the face of dogma. I ended by saying that I didn't intend to withdraw the proposal, so they would have to act upon it.

When I got home, my fury was gone. I was ready to call the director, withdraw my proposal, apologize for my errors, stay out of research, quit the VA, and go into private practice, where I could make a lot more money. Luckily, my wife Lil knows me better than I sometimes know myself. She told me, "You'd be miserable in private practice. This is exactly what you want to do, so just wait and see what happens."

Two days later I got word that the committee had delegated the decision to Professor Chester Yntema, an anatomist who long ago had studied the regrowth of ears in the salamander. Since he was the only one in Syracuse who'd ever done any regeneration research, I've always wondered why he wasn't part of the first evaluation. I went to see him with a sense of foreboding, for his latest research seemed to refute Singer's nerve work, on which I'd based my proposal.

Using a standard technique, Yntema had operated on very young salamander embryos, cutting out all of the tissues that would have given rise to the nervous system. He then grafted each of these denervated embryos onto the back of a whole one. The intact embryos furnished the grafts with blood and nourishment, and the procedure resulted in a little "parabiotic" twin, normal except for having no nerves, stuck on the back of each host animal. Yntema then cut off one leg from each of these twins, and some of them regenerated. Since microscopic examination revealed no nerves entering the graft from its host, Yntema's experiment called Singer's conclusions into question.

Dr. Yntema turned out to be one of the nicest gentlemen I've ever met, but as I entered his office his Dickensian appearance of eminence—he was tall, thin, elderly, with craggy features, and wore an immaculately starched, long white lab coat—made me feel like a freshman being called before the dean. But he put me at ease immediately.

"I've read your proposal and think it's most intriguing," he said with genuine interest.

"Do you really?" I asked. "I've been afraid you would reject it out of hand because my ideas depend on Singer's work."

"Marc Singer is a good, careful worker," Yntema replied. "I don't doubt his observations. What I've described is an exception to his findings under special circumstances."

After a long, pleasant conversation about regrowth, nerves, and research itself, he gave me his approval with a word of caution: "Don't get your hopes up about what you want to do. I don't believe for one minute that it'll work, but I think you should do it anyway. We need to encourage young researchers. Besides, it'll be fun, and maybe you'll learn something new, after all. Let me know what happens, and if you need any help, I'll be here. I'll call the people at the VA right away, so get to work. Good luck."

This was the start of a long friendship. I'm deeply indebted to Chester Yntema for his encouragement. Had he not believed that research should be fun, that you should do what you want rather than what's fashionable, my first experiment would have been impossible, and this book would never have been written.

The Reversals

First I found a good supplier of salamanders and frogs, a Tennessee game warden who ran this business in his spare time. Sometimes the shipment would contain a surprise, a small snake. I never found out whether he included them deliberately or by error. At any rate, his animals weren't the inferior aquarium-bred stock but robust specimens collected from their natural habitats.

Next I worked out some technical problems. The most important of these was the question of where to place the electrodes. To form the circuit, two electrodes had to touch the animal. One was the "hot" or measuring electrode, which determined the polarity, positive or negative, with regard to a stationary reference electrode. A negative polarity meant there were more electrons where the measuring electrode was placed, while a positive polarity meant there were more at the reference site. A steady preponderance of negative charge at a particular location could mean there was a current flowing toward that spot, continually replenishing the accumulation of electrons. The placement of the reference electrode, therefore, was critical, lest I get the voltage right but the polarity, and hence the direction of the current, wrong. Some logical position had to be chosen and used every time. Since I postulated that the nerves were somehow related to the current, the cell bodies that sent their nerve fibers into the limb seemed like a good reference point. These cell bodies were in a section of the spinal cord called the brachial enlargement, located just headward from where the arm joined the body. In both frogs and salamanders, therefore, I put the measuring electrode directly on the cut surface of the amputation stump and the reference electrode on the skin over the brachial enlargement.

After setting up the equipment, I did some preliminary measurements on the intact animals. They all had areas of positive charge at the brachial enlargement and a negative charge of about 8 to 10 millivolts at each extremity, suggesting a flow of electrons from the head and trunk out into the limbs and, in the salamanders, the tail.

I began the actual experiment by amputating the right forelimbs, between elbow and wrist, from fourteen salamanders and fourteen grass frogs, all under anesthesia. I took no special precautions against bleeding, since blood clots formed very rapidly. The wounds had to be left open, not only because closing the skin over the salamanders' amputation sites would have stopped regeneration, but also because I was investigating a natural process. In the wild, both frogs and salamanders get injuries much like the one I was producing—both are favorite foods of the freshwater bass—and heal them without a surgeon.

Once the anesthetic wore off and the blood clot formed, I took a voltage reading from each stump. I was surprised to find that the polarity at the crump reversed to positive right after the injury. By the next day it had climbed to over 20 millivolts, the same in both frogs and salamanders.

I made measurements daily, expecting to see the salamander voltages climb above those of the frogs as the blastemas formed. It didn't work that way. The force of the current flowing from the salamanders' amputation sites rapidly dropped, while that from the frogs' stumps stayed at the original level. By the third day the salamanders showed no current at all, and their blastemas hadn't even begun to appear.

The experiment seemed a failure. I almost quit right there, but something made me keep on measuring. I guess I thought it would be good practice.

Then, between the sixth and tenth days an exciting trend emerged. The salamander potentials changed their sign again, exceeding their normal voltage and reaching a peak of more than 30 millivolts negative just when the blastemas were emerging. The frogs were still plugging away with slowly declining positive voltages. As the salamander limbs regenerated and the frog stumps healed over with skin and scar tissue, both groups of limbs gradually returned (from opposite directions) to the original baseline of 10 millivolts negative.

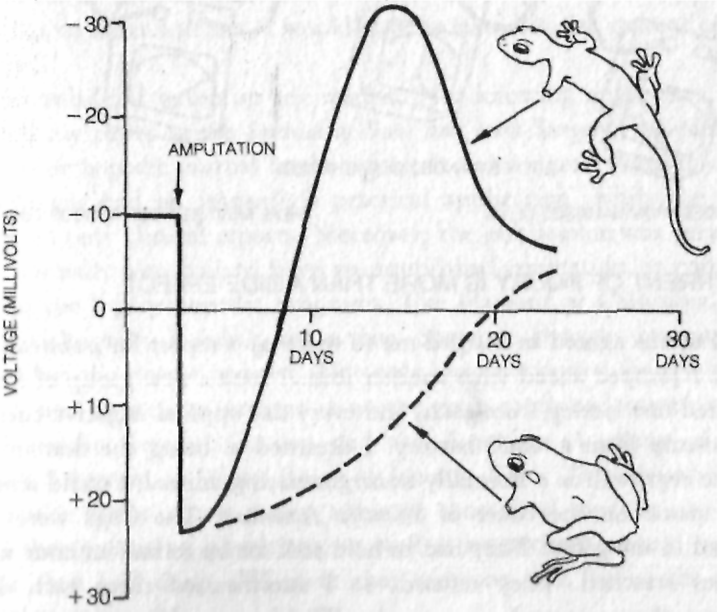

THE CURRENT OF INJURY: SALAMANDER VERSUS FROG

Here was confirmation better than my wildest dreams! Already, in my first experiment, I had the best payoff research can give—the excitement of seeing something no one else seen before. I knew now that the current of injury wasn't due to dying cells, which were long gone by then. Moreover, the opposite polarities indicated a profound difference in the electrical properties of the two animals, which somehow would explain why only the salamander could regenerate. The negative potential seemed to bring forth the all-important blastema. It was a very significant observation, even though the facts had scrambled my neat hypothesis somewhat.

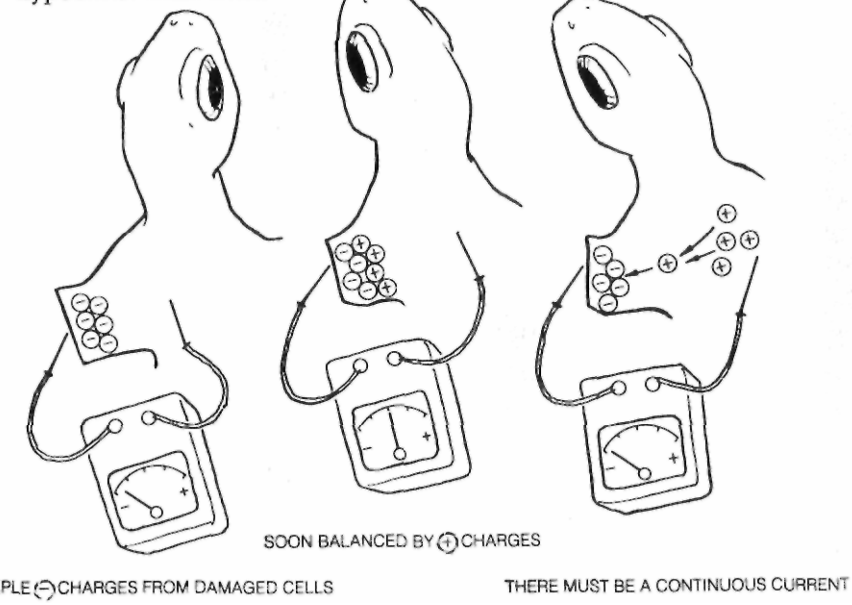

SOON BALANCED BY ⊕ CHARGES

SIMPLE ⊖ CHARGES FROM DAMAGED CELLS

THERE MUST BE A CONTINUOUS CURRENT

THE CURRENT OF INJURY IS MORE THAN A SIDE EFFECT

Dr. Yntema agreed and urged me to write up a report for publication, but first I jumped ahead with another idea. I took a new group of frogs, amputated one foreleg from each, and every day applied negative current to the stump from a small battery. I dreamed of being the first to get complete regrowth in a normally nonregenerating animal; I could almost see my name on the cover of Scientific American. The frogs were less interested in my glory. They had to hold still for up to half an hour with electrodes attached. They refused, so I anesthetized them each day, something they tolerated very poorly. Within a week my Nobel Prize had turned into a collection of dead frogs.

For some time I'd been scouring the dusty stacks of the medical library for previous work on bioelectricity, and how I found a paper written in 1909 by an American researcher named Owen E. Frazee. He reported that electrical currents passed through the aquarium water in which larval salamanders were living speeded up their regeneration. At that time, electrical equipment was so primitive that I couldn't rely on Frazee's results, but I decided to try it for myself. What Sinyukhin had done with tomato plants I hoped to do with salamanders.

To one group of salamanders I applied 2 microamperes of positive current from batteries connected directly to the stumps for five to ten minutes on each of the first five days after amputation. This was 0.000002 ampere, a tiny current by ordinary standards (most household circuits carry 15 or 20 amperes) but comparable to what seemed to be flowing in the limb. I intended to reinforce the normal positive peak in the current of injury. This treatment seemed to make the blastemas larger but slowed down the whole process somewhat. To another group I applied 3 microamperes of negative current on the fifth to ninth days, when the normal currents were hitting their negative peaks. This seemed to increase the rate of regrowth for a week but didn't change the time needed for a complete limb. Finally I tried Frazee's method with a constant current through the aquarium water. Again the results were equivocal at best. These failures taught me that, before I applied my findings to other animals, I would have to learn how the current of injury worked.

Meanwhile, I wrote up my results. Not knowing any better, I submitted my paper to the Journal of Bone and Joint Surgery, the most prestigious orthopedic journal in the world. It was a dumb thing to do. The experiment had no immediate practical application, while the journal accepted only clinical reports. Moreover, the publication was very political; normally you had to have an established reputation or come from one of the big orthopedic programs, like Harvard or Columbia, to get into it. Luckily, I didn't know that. Someone thought my paper was just what the doctor ordered. Not only was it accepted for publication, but I was invited to present it at the next combined meeting of the Orthopaedic Research Society and the American Academy of Orthopaedic Surgeons, at Miami Beach in January 1961. This invitation was a particular honor, for it meant someone considered my work so significant that practicing physicians, as well as researchers, should hear of it right then and there. Whoever that someone was, he or she has my undying gratitude.

My report was well received and soon was published, to the consternation of the local inquisitors and the delight of Chester Yntema.

Since the journal was geared to clinicians. I worried that my experiment wouldn't reach the basic researchers with whom I really wanted to share it, but again I was wrong. The next year I got a phone call from Meryl

Rose himself. He was excited by the article and wanted to know what I'd done since.

Although Rose taught at Tulane Medical School in New Orleans, he spent every summer at the Woods Hole Marine Biological Laboratory on Cape Cod, so he and his wife drove to Syracuse from there. Despite his success, Rose had maintained the completely open mind that a great researcher must have, and he was fascinated by the observations on electric fields, nerves, anesthesia, and magnetism that I'll recount in the next chapter. Since then his interest has encouraged me enormously. My friendship with this fine man and scientist has been fruitful even beyond the expectations I had then, and, when my wife and I had the Roses to dinner, we found our pasts were linked by an odd coincidence. As they walked in the door, Lillian exclaimed, "Dr. Rose! Weren't you at Smith College in the 1940s?" It turned out that she'd been a friend of Rose's student lab assistant and had helped catch the frogs for the famous salt-in-the-wound experiment!

Part 2
The Stimulating Current

The basic texture of research consists of dreams into which
the threads of reasoning, measurement, and calculation
are woven.
—Albert Szent-Gyorgyi

Four

Life's Potentials

It's an axiom of science that the better an experiment is, the more new questions it raises after it has answered the one you asked. By that standard my first simple test had been pretty good. The new problems branched out like the fingers on those restored limbs: Where did the injury currents come from? Were they in fact related to the nervous system and, if so, how? It seemed unlikely that they sprang into action only after an amputation; they must have existed before. There must have been a preexisting substratum of direct current activity that responded to the injury. Did the voltages I measured really reflect such currents, and did they flow throughout the salamander's body? Did other organisms have them? What structures carried them? What were their electrical properties? What were they doing the rest of the time, before injury and after healing? Could they be used to provoke regeneration where it was normally absent?

I had ideas about how to look for some of the answers, but, to understand my approach, the reader unfamiliar with electrical terms will need a simplified explanation of several basic concepts that are essential to the rest of the story.

Everything electrical stems from the phenomenon of charge. No one knows exactly what this is, except to say that it's a fundamental property of matter that exists in two opposite forms, or polarities, which we arbitrarily call positive and negative. Protons, which are one of the two main types of particles in atomic nuclei, are positive; the other particles, the neutrons, are so named because rhey have no charge. Orbiting around the nucleus are electrons, in the same number as the protons inside the nucleus. Although an electron is 1,836 times less massive than a proton, the electron carries an equal but opposite (negative) charge. Because of their lightness and their position outside the nucleus, electrons are much more easily dislodged from atoms than are protons, so they're the main carriers of electric charge. For the lay person's purposes a negative charge can be thought of as a surplus of electrons, while a positive charge can be considered a scarcity of them. When electrons move away from an area, it becomes positively charged, and the area to which they move becomes negative.

A flow of electrons is called a current, and is measured in amperes, units named for an early-nineteenth-century French physicist, Andre Marie Ampere. A direct current is a more or less even flow, as opposed to the instantaneous discharge of static electricity as sparks or lightning, or the back-and-forth flow of alternating current which powers most of our appliances.

Besides the amount of charge being moved, a current has another characteristic important for our narrative—its electromotive force. This can be visualized as the "push" behind the current, and it's measured in volts (named for Alessandro Volta).

In high school most of us learned that a current flows only when a source of electrons (negatively charged material) is connected to a material having fewer free electrons (positively charged in relation to the source) by a conductor, through which the electrons can flow. This is what happens when you connect the negative terminal of a battery to its positive pole with a wire or a radio's innards: You've completed a circuit between negative and positive. If there's no conductor, and hence no circuit, there's only a hypothetical charge flow, or electric potential, between the two areas. The force of this latent current is also measured in volts by temporarily completing the circuit with a recording device, as I did in my experiment.

The potential can continue to build until a violent burst of current equalizes the charges; this is what happens when lightning strikes. Smaller potentials may remain stable, however. In this case they must be continuously fed by a direct current flowing from positive to negative, the opposite of the normal direction. In this part of a circuit, electrons actually flow from where they're scarce to where they're more abundant. As Volta found, such a flow is generated inside a battery by the electrical interaction of two metals.

An electric field forms around any electric charge. This means that any other charged object will be attracted (if the polarities are opposite) or repelled (if they're the same) for a certain distance around the first object. The field is the region of space in which an electrical charge can be detected, and it's measured in volts per unit of area.

Electric fields must be distinguished from magnetic fields. Like charge, magnetism is a dimly understood intrinsic property of matter that manifests itself in two polarities. Any flow of electrons sets up a combined electric and magnetic field around the current, which in turn affects other electrons nearby. Around a direct current the electromagnetic field is stable, whereas an alternating current's field collapses and reappears with its poles reversed every time the current changes direction. This reversal happens sixty times a second in our normal house currents. Just as a current produces a magnetic field, a magnetic field, when it moves in relation to a conductor, induces a current. Any varying magnetic field, like that around household appliances, generates a current in nearby conductors. The weak magnetic fields we'll be discussing are measured in gauss, units named after a nineteenth-century German pioneer in the study of magnetism, Karl Friedrich Gauss.

Both electric and magnetic fields are really just abstractions that scientists have made up to try to understand electricity's and magnetism's action at a distance, produced by no known intervening material or energy, a phenomenon that used to be considered impossible until it became undeniable. A field is represented by lines of force, another abstraction, to indicate its direction and shape. Both kinds of fields decline with distance, but their influence is technically infinite: Every time you use your toaster, the fields around it perturb charged particles in the farthest galaxies ever so slightly.

In addition, there's a whole universe full of electromagnetic energy, radiation that somehow seems to be both waves in an electromagnetic field and particles at the same time. It exists in a spectrum of wavelengths that includes cosmic rays, gamma rays, X rays, ultraviolet radiation, visible light, infrared radiation, microwaves, and radio waves. Together, electromagnetic fields and energies interact in many complex ways that have given rise to much of the natural world, not to mention the whole technology of electronics.

You'll need a casual acquaintance with all these terms for the story ahead, but don't worry if the concepts seem a bit murky. Physicists have been trying for generations to solve the fundamental mysteries of electromagnetism, and no one, not even Einstein, has yet succeeded.

Unpopular Science

None of these things had the slightest relevance to life, according to most biologists around 1960. A major evaluation of American medicine, financed by the Carnegie Foundation and published in 1910 by the respected educator Abraham Flexner, had denounced the clinical use of electric shocks and currents, which had been applied, often overenthusiastically, to many diseases since the mid-1700s. Electrotherapy sometimes seemed to work, but no one knew why, and it had gotten a bad name from the many charlatans who'd exploited it. Its legitimate proponents had no scientific way to defend it, so the reforms in medical education that followed the Flexner report drove all mention of it from the classroom and clinic, just as the last remnants of belief in vital electricity were being purged from biology by the discovery of acetylcholine. This development dovetailed nicely with expanding knowledge of biochemistry and growing reliance on the drug industry's products. Penicillin later made medicine almost exclusively drug oriented.

Meanwhile, the work of Faraday, Edison, Marconi, and others literally electrified the world. As the uses of electricity multiplied, no one found any obvious effects on living creatures except for the shock and heating caused by large currents. To be sure, no one looked very hard, for fear of discouraging a growth industry, but the magic of electricity seemed to lie precisely in the way it worked its wonders unseen and unfelt by the folks clustered around the radio or playing cards under the light bulb. By the 1920s, no scientist intent on a respectable career dared suggest that life was in any sense electrical.

Nevertheless, some researchers kept coming up with observations that didn't fit the prevailing view. Although their work was mostly consigned to the fringes of the scientific community, by the late 1950s they'd accumulated quite a bit of evidence.

There were two groups of dissenters, but, because their work went unheeded, each was largely unaware of the other's existence. One line of inquiry began just after the turn of the century when it was learned that hydras were electrically polarized. The head was found to be positive, the tail negative. I've already mentioned Frazee's 1909 report of salamander regeneration enhanced by electrical currents. Then, with a classic series of experiments in the early 1920s, Elmer J. Lund of the University of Texas found that the polarity of regeneration in species related to the hydra could be controlled, even reversed, by small direct currents passes through the animal's body. A current strong enough to override the creature's normal polarity could cause a head to form where a tail should have reappeared, and vice versa. Others confirmed this discovery, and Lund went on to study eggs and embryos. He claimed to have influenced the development of frog eggs not only with currents but also with magnetic fields, a conclusion that was really risque for that time.

Stimulated by Lund's papers, Harold Saxton Burr of Yale began putting electrodes to all kinds of creatures. Burr was lucky enough to have a forum for his work. He edited the Yale Journal of Biology and Medicine, where most of his reports appeared; few other journals would touch them. Burr and his co-workers found electric fields around, and electric potentials on the surfaces of, organisms as diverse as worms, hydras, salamanders, humans, other mammals, and even slime molds. They measured changes in these potentials and correlated them to growth, regeneration, tumor formation, drug effects, hypnosis, and sleep. Burr claimed to have measured field changes resulting from ovulation, but others got contradictory results. He hooked up his voltmeters to trees for years at a time and found that their fields varied in response not only to light and moisture, but to storms, sunspots, and the phases of the moon as well.

Burr and Lund were handicapped by their instruments as well as the research climate. Most of their work was done before World War II and, even though Burr spent years designing the most sensitive devices possible using vacuum tubes, the meters were still too "noisy" to reliably measure the tiny currents found in living things. The two scientists could refine their observations only enough to find a simple dipolar distribution of potentials, the head of most animals being negative and the tail positive.

Burr and Lund advanced similar theories of an electrodynamic field, called by Burr the field of life or L-field, which held the shape of an organism just as a mold determines the shape of a gelatin dessert. "When we meet a friend we have not seen for six months there is not one molecule in his face which was there when we last saw him," Burr wrote. "But, thanks to his controlling L-field, the new molecules have fallen into the old, familiar pattern and we can recognize his face."

Burr believed that faults in the field could reveal latent illness just as dents in a mold show up in the jelly. He claimed to be able to predict all sorts of things about a person's emotional and physical health, both present and future, merely by checking the voltage between head and hand. His later writings were marred by a son of bioelectric determinism and a tendency to confuse "law and order" in nature with that odious euphemism as preached by Presidents. As a result, he began to suggest his simple readings as a foolproof way to evaluate job applicants, soldiers, mental patients, and suspected criminals or dissidents.

The fields Burr and Lund found were actually far too simple to account for a salamander's limb or a human face. Biological knowledge at that time gave them no theoretical framework to explain where their fields came from. They conceived of currents flowing within cells but had no proof. They had no inkling that currents might flow in specific tissues or in the fluids outside cells. They suggested that all these little intracellular currents somehow added up to the whole field. Burr wrote that "electrical energy is a fundamental attribute of protoplasm and is an expression or measure of the presence of an electrodynamic field in the organism." Unfortunately, an analysis of this sentence yields nonsense, and Burr's work was dismissed as foggy vitalism. Lund suffered the same fate. No one bothered to see if the measurements they'd made were valid. After all, you can disagree with a theory, but you should respect the data enough to check them. If you can't duplicate them, you're entitled to rest easy with your own concepts, but if you get the same results, you're obligated to agree or propose an alternate theory. Most scientists took the easy way out, however, and simply ignored Burr and Lund. Their discoveries remained little known, and most biologists didn't connect them with the tentative morphogenetic-field concept of regeneration.

Then in 1952 Lund's work was taken up by G. Marsh and H. W. Beams using the planarian. They found that the flatworm's polarity, like the hydra's, could be controlled by passing a current through it. When a direct current was fed in the proper direction through a section of a worm, normal polarity disappeared and a head formed at each end. As the current strength was increased, the section's polarity reversed; a head regrew at the rear, a tail at the front. At higher voltages, even intact worms completely reorganized, with the head becoming a tail and vice versa. Marsh and Beams grew convinced that the animal's electric field was the morphogenetic organizing principle. Still, their work was also ignored, except by Meryl Rose, who suggested that a gradation of electrical charge from front to back controlled the gradient of growth inhibitors and stimulators. He suggested that the growth compounds were charged molecules that were moved to different places in the body by the electric field, depending on the amount and sign of their charge and their molecular weight.

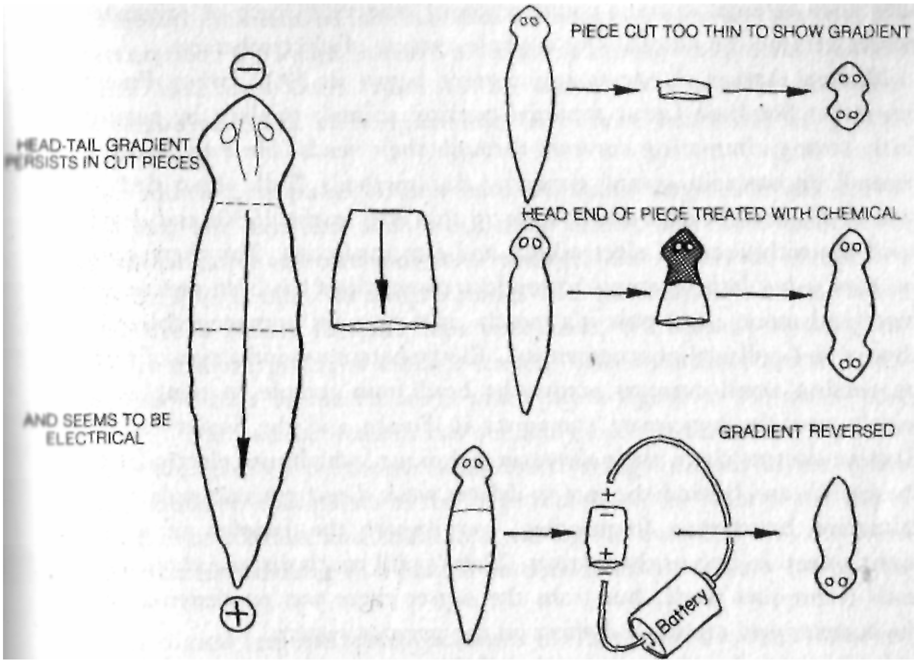
HEAD-TAIL GRADIENT
PERSISTS IN CUT PIECES
AND SEEMS TO BE
ELECTRICAL
PIECE CUT TOO THIN TO SHOW GRADIENT
HEAD END OF PIECE TREATED WITH CHEMICAL
GRADIENT REVERSED
Battery
+
−

MARSH AND BEAMS'S EXPERIMENT

Undercurrents in Neurology

While the investigation of the total body field moved haltingly forward in the study of simple animals, several neurophysiologists began finding out odd things about the nerves of more complex creatures, data that Bernstein's action potential couldn't explain. Going through the old literature, following lead after lead from one paper to the next, I found many hints that there were DC potentials in the nervous system and that small currents from outside could affect brain function.

The first recorded use of currents on the nervous system was by Giovanni Aldini, a nephew of Galvani and an ardent champion of vitalism. Using the batteries of his archenemy Volta, Aldini claimed remarkable success in relieving asthma. He also cured a man who today would probably be diagnosed as schizophrenic, although it's impossible to know how much benefit came from the currents and how much from simple solicitude, then so rare in treating mental illness. Aldini gave his patient a room in his own house and later found him a job. Some of Aldini's experiments were grotesque - he tried to resurrect recently executed criminals by making the corpses twitch with electricity - but his idea that external current could replenish the vital force of exhausted nerves became the rationale for a whole century of electrotherapy.

Modern studies of nerves and current began in 1902, when French researcher Stephane Leduc reported putting animals to sleep by passing fairly strong alternating currents through their heads. He even knocked himself unconscious several times by this method. (Talk about dedication to science!) Several others took up this lead in the 1930s and developed the techniques of electroshock and electronarcosis. The therapeutic value of using large currents to produce convulsions has been questioned more and more, until now it's mostly used to quiet unmanageable psychotics and political nonconformists. Electronarcosis—induction of sleep by passing small currents across the head from temple to temple—is widely used by legitimate therapists in France and the Soviet Union. Russian doctors claim their elektroson technique, which uses electrodes on the eyelids and behind the ears to deliver weak direct currents pulsing at calmative brain-wave frequencies, can impart the benefits of a full night's sleep in two or three hours. There's still much dispute about how both techniques work, but from the outset there was no denying that the currents had a profound effect on the nervous system.

In the second and third decades of this century there was a flurry of interest in galvanotaxis, the idea that direct currents guided the growth of cells, especially neurons. In 1920, S. Ingvar found that the fibers growing out of nerve cell bodies would align themselves with a nearby flow of current and that the fibers growing toward the negative electrode were different from those growing toward the positive one. Paul Weiss soon "explained" this heretical observation as an artifact caused by stretching of the cell culture substrate due to contact with the electrodes. Even after Marsh and Beams proved Weiss wrong in 1946, it took many more years for the scientific community to accept the fact that neuron fibers do orient themselves along a current flow. Today the possible use of electricity to guide nerve growth is one of the most exciting prospects in regeneration research (see Chapter 11).

The Bernstein hypothesis, unable to account for these facts, has turned out to be deficient in several other respects. To begin with, according to the theory, an impulse should travel with equal ease in either direction along the nerve fiber. If the nerve is stimulated in the middle, an impulse should travel in both directions to opposite ends. Instead, impulses travel only in one direction; in experiments they can be made to travel "upstream," but only with great difficulty. This may not seem like such a big deal, but it is very significant. Something seems to polarize the nerve.

Another problem is the fact that, although nerves are essential for regeneration, the action potentials are silent during the process. No impulses have ever been found to be related to regrowth, and neurotransmitters such as acetylcholine have been ruled out as growth stimulators.

In addition, impulses always have the same magnitude and speed. This may not seem like such a big thing either, but think about it. It means the nerve can carry only one message, like the digital computer's 1 or 0. This is okay for simple things like the knee-jerk reflex. When the doctor's rubber hammer taps your knee, it's actually striking the patellar tendon, giving it a quick stretch. This stimulates stretch receptors (nerve cells in the tendon), which fire a signal to the spinal cord saying, "The patellar tendon has suddenly been stretched." These impulses are received by motor (muscle-activating) neurons in the spinal cord, which send impulses to the large muscle on the front of the thigh, telling it to contract and straighten the leg. In everyday life, the reflex keeps you from falling in a heap if an outside force suddenly bends your knees.

The digital impulse system accounts for this perfectly well. However, no one can walk on reflexes alone, as victims of cerebral palsy know all too well. The motor activities we take for granted—getting out of a chair and walking across a room, picking up a cup and drinking coffee, and so on—require integration of all the muscles and sensory organs working smoothly together to produce coordinated movements that we don't even have to think about. No one has ever explained how the simple code of impulses can do all that. Even more troublesome are the higher processes, such as sight—in which somehow we interpret a constantly changing scene made of innumerable bits of visual data—or the speech patterns, symbol recognition, and grammar of our languages. Heading the list of riddles is the "mind-brain problem" of consciousness, with its recognition, "I am real; I think; I am something special." Then there are abstract thought, memory, personality, creativity, and dreams. The story goes that Otto Loewi had wrestled with the problem of the synapse for a long time without result, when one night he had a dream in which the entire frog-heart experiment was reveiled to him. When he awoke, he knew he'd had the dream, but he'd forgotten the details. The next night he had the same dream. This time he remembered the procedure, went to his lab in the morning, did the experiment, and solved the problem. The inspiration that seemed to banish neural electricity forever can't be explained by the theory it supported! How do you convert simple digital messages into these complex phenomena? Latter-day mechanists have simply postulated brain circuitry so intricate that we will probably never figure it out, but some scientists have said there must be other factors.

Even as Loewi was finishing his work on acetylcholine, others began to find evidence that currents flowed in the nerves. English physiologist Richard Caton had already claimed he'd detected an electric field around the heads of animals in 1875, but it wasn't until 1924 that German psychiatrist Hans Berger proved it by recording the first electroencephalogram (EEG) from platinum wires he inserted into his son's scalp. The EEG provided a record of rhythmic fluctuations in potential voltage over various parts of the head. Berger at first thought there was only one wave from the whole brain, but it soon became clear that the waves differed, depending on where the electrodes were put. Modern EEGs use as many as thirty-two separate channels, all over the head.

The frequency of these brain waves has been crudely correlated with states of consciousness. Delta waves (0.5 to 3 cycles per second) indicate deep sleep. Theta waves (4 to 8 cycles per second) indicate trance, drowsiness, or light sleep. Alpha waves (8 to 14 cycles per second) appear during relaxed wakefulness or meditation. And beta waves (14 to 35 cycles per second), the most uneven forms, accompany all the modulations of our active everyday consciousness. Underlying these rhythms are potentials that vary much more slowly, over periods as long as several minutes. Today's EEG machines are designed to filter them out because they cause the trace to wander and are considered insignificant anyway.

There's still no consensus as to where the EEG voltages come from. They would be most easily explained by direct currents, both steady state and pulsing, throughout the brain, but that has been impossible for most biologists to accept. The main alternative theory, that large numbers of neurons firing simultaneously can mimic real electrical activity, has never been proven.

In 1939, W. E. Burge of the University of Illinois found that the voltage measured between the head and other parts of the body became more negative during physical activity, declined in sleep, and reversed to positive under general anesthesia. At about the same time a group of physiologists and neurologists at Harvard Medical School began studying the brain with a group of MIT mathematicians. This association was destined to change the world. From it came many of our modern concepts of cybernetics, and it became the nucleus of the main American task force on computers in World War II. One of the group's first important ideas was that the brain worked by a combination of analog and digital coding.

One of the mathematicians, computer pioneer John von Neumann, later elaborated the concept in great detail, but basically it's rather simple. In analog computers, changes in information are expressed by analogous changes in the magnitude or polarity of a current. For example, if the computer is to use and store the varying temperatures of a furnace, the rise and fall in heat can be mimicked by a rise and fall in voltage. Analog systems are slow and can handle only simple information, but they can express subtle variations very well. Digital coding, on the other hand, can transmit enormous amounts of data at high speed, but only if the information can be reduced to yes-no, on-off bits—the digits 1 and 0. If the brain was such a hybrid computer, these early cyberneticists reasoned, then analog coding could control the overall activity of large groups of neurons by such actions as increasing or decreasing their sensitivity to incoming messages. (A few years later neurologists did find that some neurons were "tuned" to fire only if they received a certain number of impulses.) The digital system would transfer sensory and motor information, but the processing of that information—memory and recall, thought, and so on—would be accomplished by the synergism of both methods. The voltage changes Burge found in response to major alterations of consciousness seemed to fit within this framework, and his observations were extended by the Harvard-MIT group and others. Much of this work was done directly on the exposed brains of animals and of human patients during surgery. When cooperative patients elected to remain awake during such operations (the brain is immune to pain), human sensations could often be correlated with electrical data. Contributors to this endeavor included nearly all of the greatest American neurophysiologists—Walter B. Cannon, Arturo Rosenblueth, Ralph Gerard, Gilbert Ling, Wilder Penfield, and others.

Measurements on the exposed brain quickly confirmed the existence of potential voltages and also revealed possible currents of injury. Whenever groups of nerve cells were actively conducting impulses, they also produced a negative potential. Positive potentials appeared from injured cells when the brain had been damaged; these potentials then expanded outward to uninjured cells, suppressing their ability to send or receive impulses. When experimenters applied small negative voltages to groups of neurons, their sensitivity increased; that is, they would generate an impulse in response to a weaker stimulus. Externally applied positive potentials worked in the opposite way: They depressed nerve function, making it harder to produce an impulse. Thus there did seem to be an analog code, but how did it work? Did the potentials come from direct currents generated by the nerve cells themselves, or did they merely result from adding up a lot of action potentials all going in the same direction and arriving in the same place at the same time?

Some answers were provided by a series of beautiful experiments by Ling, Gerard, and Benjamin Libet at the University of Chicago. Working on frogs, they studied areas of the cortex where the neuron layer was only one cell thick and the cells were arranged side by side like soldiers on review, all pointing in the same direction. In such areas they found a negative potential on the dendrites (the short incoming fibers) and a positive potential at the ends of the axons (the longer outgoing fibers). This indicated a steady direct current along the normal direction of impulse transmission. The entire nerve cell was electrically polarized.

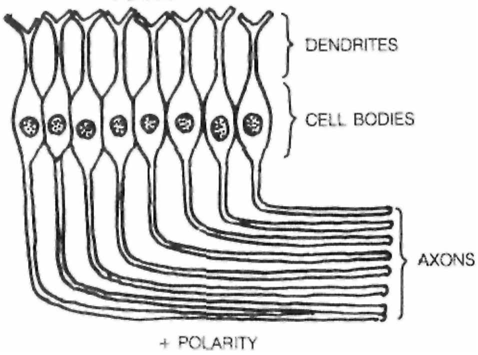

POLARIZED NERVE CELLS LINED UP IN CEREBRAL CORTEX

In another series of experiments, on brains removed from frogs and kept alive in culture, the Chicago group found that direct currents swept across the surface of the cortex in very slow waves, which could be produced experimentally by applying chemicals such as caffeine to a single spot on the surface. When they made a cut on the brain, severing groups of nerve fibers, these DC traveling waves would still cross the cut if the two surfaces were in direct contact. If the researchers held the cut open and filled it with a saline solution that matched body fluids, then the waves couldn't cross the gap. These were particularly important observations. They indicated that the current was transmitted by structures outside the neurons; it crossed the cut when the edges touched, but the microscopic parts of the severed neurons wouldn't have rejoined so easily. The results also showed that the current was not a flow of ions; otherwise it would have been able to cross the gap through the salt water.

Studying intact brains in living frogs, the lame group found a potential between the front and back of the brain. The olfactory (frontal) lobes were several milivolts negative with respect to the occipital (rear) lobe, implying a current flowing up the brain stem and between the two hemispheres to the front.

At the time, these observations seemed mighty odd. They didn't fit any concepts of how the nerves worked. As a result, they were largely ignored. The majority of neurophysiologists went on measuring the action potentials and tracing out fiber pathways in the brain. This was useful work but limited. The basic questions remained.

Only one research team followed up this work, some ten years later. Sidney Goldring and James L. O'Leary, neuropsychiatrists at the Washington University School of Medicine in St. Louis, recorded the same DC potentials from the human scalp, from the exposed brain during surgery, and from the brains of monkeys and rabbits. As noted before, the potentials varied in regular cycles several minutes long, like a basso continuo under the EEG. In fact, Goldring and O'Leary found waves within waves: "Written upon the slow major swings were lesser voltage changes." These were weak potentials, measured in microvolts (millionths of a volt) and varying in waves of 2 to 30 cycles per minute, sort of a pianissimo "inner voice" in a three-part electrical fugue.

Conducting in a New Mode

I was acutely aware that I didn't have the "proper" background for the work I planned to do. I wasn't a professional neurophysiologist; I didn't even know one. Indeed, after my run-in with the research committee, one member had taken me aside and earnestly advised me, "Go back to school and get your Ph.D., Becker. Then you'll learn all of this stuff is nonsense." Still, some of the greatest neurophysiologists had thought the same way I did about "all of this stuff." They suggested we might have been too hasty in throwing electric currents out of biology. My notion of putting them back in wasn't so outlandish, but only an extension of what they'd been saying. I was approaching the body's system of information transfer from the periphery, asking, "What makes wounds heal?" They'd started from the center, asking, "How does the brain work?" We were working on the same problem from opposite ends. As I contemplated their findings and all of biology's unsolved problems, I grew convinced that life was more complex than we suspected. I felt that those who reduced life to a mechanical interaction of molecules were living in a cold, gray, dead world, which, despite its drabness, was a fantasy. I didn't think electricity would turn to be any elan vital in the old sense, but I had a hunch would be closer to the secret than the smells of the biochemistry lab or the dissecting room's preserved organs.

I had another worthy ally when I started to reevaluate the role of electricity in life. Albert Szent-Gyorgyi, who'd already won a Nobel Prize for his work on oxidation and vitamin C, made a stunning suggestion in a speech before the Budapest Academy of Science on March 21, 1941. (Think of the date. World War II was literally exploding around him, and there he was, calmly laying the foundations for a new biology.) Speaking of the mechanistic approach of biochemistry, he pointed out that when experimenters broke living things down into their constituent parts, somewhere along the line life slipped through their fingers and they found themselves working with dead matter. He said, "It looks as if some basic fact about life were still missing, without which any real understanding is impossible." For the missing basic fact, Szent-Gyorgyi proposed putting electricity back into living things, but not in the way it had been thought of at the turn of the century.

At that earlier time, there had been only two known modes of current conduction, ionic and metallic. Metallic conduction can be visualized as a cloud of electrons moving along the surface of metal, usually a wire. It can be automatically excluded from living creatures because no one has ever found any wires in them. Ionic current is conducted in solutions by the movement of ions—atoms or molecules charged by having more or fewer than the number of electrons needed to balance their protons' positive charges. Since ions are much bigger than electrons, they move more laboriously through the conducting medium, and ionic currents die out after short distances. They work fine across the thin membrane of the nerve fiber, but it would be impossible to sustain an ionic current down the length of even the shortest nerve.

Semiconduction, the third kind of current, was a laboratory curiosity in the 1930s. Halfway between conductors and insulators, the semiconductors are inefficient, in the sense that they can carry only small currents, but they can conduct their currents readily over long distances. Without them, modern computers, satellites, and all the rest of our solid-state electronics would be impossible.

Semiconduction occurs only in materials having an orderly molecular structure, such as crystals, in which electrons can move easily from the electron cloud around one atomic nucleus to the cloud around another. The atoms in a crystal are arranged in neat geometrieal lattices, rather than the frozen jumble of ordinary solids. Some crystalline materials have spaces in the lattice where other atoms can fit. The atoms of these impurities may have more or fewer electrons than the atoms of the lattice material. Since the forces of the latticework structure hold the same number of electrons in place around each atom, the "extra" electrons of the impurity atoms are free to move through the lattice without being bound to any particular atom. If the impurity atoms have fewer electrons than the others, the "holes" in their electron clouds can be filled by electrons from other atoms, leaving holes elsewhere. A negative current, or N-type semiconduction, amounts to the movement of excess electrons; a positive current, or P-type semiconduction, is the movement of these holes, which can be thought of as positive charges.

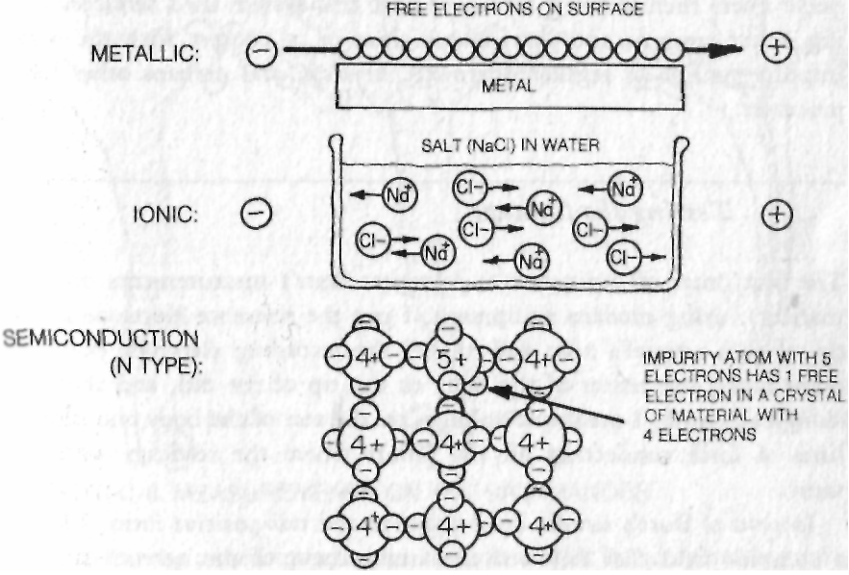
FREE ELECTRONS ON SURFACE
METALLIC:
METAL
SALT (NaCl) IN WATER
IONIC:
Na+
Cl−
SEMICONDUCTION (N TYPE):
4+
5+
IMPURITY ATOM WITH 5 ELECTRONS HAS 1 FREE ELECTRON IN A CRYSTAL OF MATERIAL WITH 4 ELECTRONS

THREE WAYS OF CONDUCTING ELECTRICAL CURRENT

Szent-Gyorgyi pointed out that the molecular structure of many parts of the cell was regular enough to support semiconduction. This idea was almost completely ignored at the time. Even when Szent-Gyorgyi expanded the concept in his I960 Introduction to a Submolecular Biology, most scientists (except in Russia!) dismissed it as evidence of his advancing age, but that little book was an inspiration to me. I think it may turn out to be the man's most important contribution to science. In it he conjectured that protein molecules, each having a sort of slot or way station for mobile electrons, might be joined together in long chains so that electrons could flow in a semiconducting current over long distances without losing energy, much as in a game of checkers one counter could jump along a row of other pieces across the entire board. Szent-Gyorgyi suggested that the electron flow would be smilar to photosynthesis, another process he helped elucidate, in which a kind of waterfall of electrons cascaded step by step down a staircase of molecules, losing energy with each bounce. The main difference was that in protein semiconduction the electrons' energy would be conserved and passed along as information instead of being absorbed and stored in the chemical bonds of food.

With Szent-Gyorgyi's suggestion in mind, I put together my working hypothesis. I postulated a primitive, analog-coded information system that was closely related to the nerves but not necessarily located in the nerve fibers themselves. I theorized that this system used semiconducting direct currents and that, either alone or in concert with the nerve impulse system, it regulated growth, healing, and perhaps other basic processes.

Testing the Concept

The first order of business was to repeat Burr's measurements on salamanders, using modern equipment. I put the reference electrode at the tip of each animal's nose and moved the recording electrode point by point along the center of the body to the tip of the tail, and then out along each limb. I measured voltages on the rest of the body and plotted lines of force connecting all the points where the readings were the same.

Instead of Burr's simple head-negative and tail-positive form, I found a complex field that followed the arrangement of the nervous system. There were large positive potentials over each lobe of the brain, and slightly smaller ones over the brachial and lumbar nerve ganglia between each pair of limbs. The readings grew increasingly negative as I moved away from these collections of nerve cell bodies; the hands, feet, and tip of the tail had the highest negative potentials.

In another series of measurements, I watched the potentials develop along with the nervous system in larval salamanders. In the adults, cutting the nerves where they entered the legs—that is, severing the long nerve fibers from their cell bodies in the spinal cord—wiped out the limb potentials almost entirely. But if I cut the spinal cord, leaving the peripheral nerves connected to their cell bodies, the limb potentials didn't change. It certainly looked as though there was a current being generated in the nerve cell bodies and traveling down the fibers.

To have a current flow you need a circuit; the current has to be made at one spot, pass through a conductor, and eventually get back to the generator. We tend to forget that the 60-cycle alternating current in the

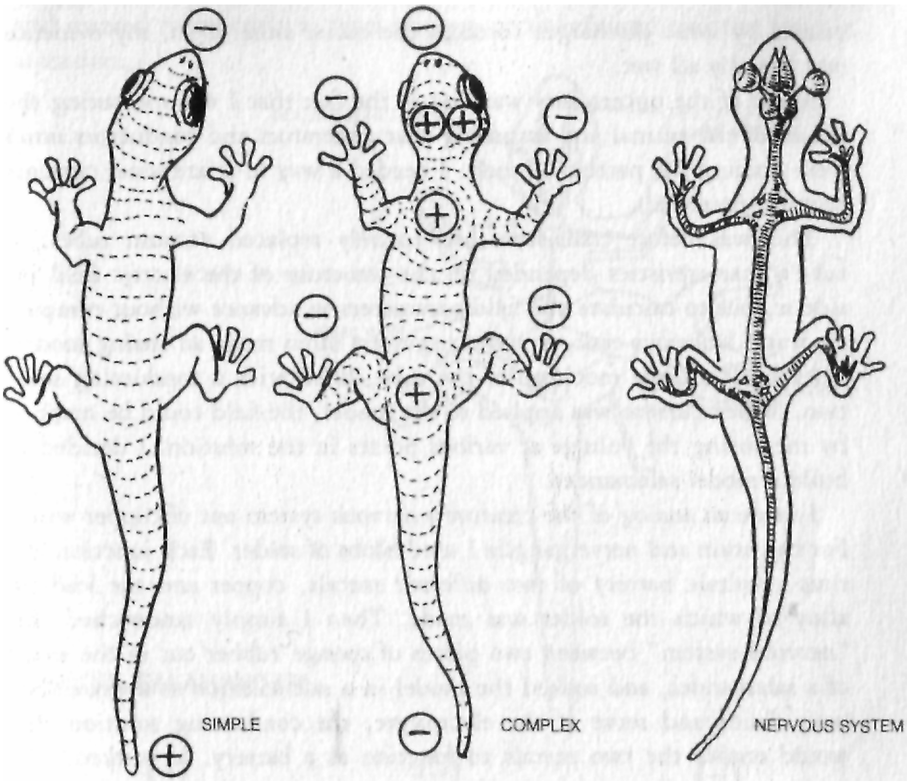

ELECTRICAL MEASUREMENTS ON THE SALAMANDER

wall socket isn't used up when we turn on a light but is merely coursing through it to the ground, through which it eventually returns to the power station. Since my measurements were positive over collections of nerve cell bodies, and increasingly negative out along the nerve fibers, it seemed a good bet that current was being generated in the cell bodies, especially since they contained all of the "good stuff"—the nucleus, organelles, and metabolic components—while the fibers were relatively uninteresting prolongations of the body. At the time, I supposed the circuit was completed by current going back toward the spine through the muscles.

This was a good start, but it wasn't scientifically acceptable proof. For one thing, my guess about the return part of the circuit was soon disproved when I measured the limb muscles and found them polarized in the same direction as the surface potentials. For another thing, it had recently been discovered that amphibian skin itself was polarized, inside versus outside, by ion differences much like the nerve membrane's resting potential, so it was just barely possible that my readings had been caused by ionic discharges through the moist skin. If so, my evidence was literally all wet.

Much of the uncertainty was due to the fact that I was measuring the outside of the animal and assuming that generators and conductors inside were making the pattern I found. I needed a way to relate inner currents to outer potentials.

This was before transistors had entirely replaced vacuum tubes. A tube's characteristics depended on the structure of the electric field inside it, but to calculate the field parameters in advance without computers was a laborious task, so radio engineers often made an analog model. They built a large mock-up of the tube, filled with a conducting solution. When current was applied to the model, the field could be mapped by measuring the voltage at various points in the solution. I decided to build a model salamander.

I made an analog of the creature's nervous system out of copper wires. For the brain and nerve ganglia I used blobs of solder. Each junction was thus a voltaic battery of two different metals, copper and the lead-tin alloy of which the solder was made. Then I simply sandwiched this "nervous system" between two pieces of sponge rubber cut in the shape of a salamander, and soaked the model in a salt solution to approximate body fluids and serve as the electrolyte, the conducting solution that would enable the two metals to function as a battery. It worked. The readings were almost exactly the same as in the real salamander. This showed that a direct current inside could produce the potentials I was getting on the outside.

If my proposed system was really a primitive part of the nervous system, it should be widely distributed, so next I surveyed the whole animal kingdom. I tested flatworms, earthworms, fish, amphibians, reptiles, mammals, and humans. In each species the potentials on the skin reflected the arrangement of the nervous system. In the worms and fish, there was only one area of positive potential, just as there was only one major nerve ganglion, the brain. In humans the entire head and spinal region, with its massive concentration of neurons, was strongly positive. The three specific areas of greatest positive potential were the same as in the salamander: the brain, the brachial plexus between the shoulder blades, and the lumbar enlargement at the base of the spinal cord. In all vertebrates I also recorded a midline head potential that suggested a direct current like that postulated by Gerard, flowing from back to front through the middle of the brain. It looked as though the current came from the reticular activating system, a network of cross-linked neurons that fanned out from the brainstem into higher centers and seemed to control the level of sleep or wakefulness and the focus of attention.

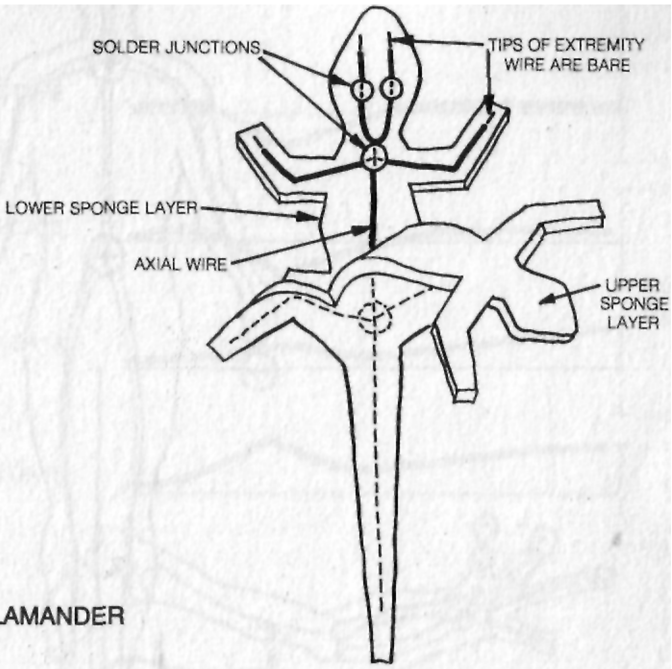
SOLDER JUNCTIONS
TIPS OF EXTREMITY WIRE ARE BARE
LOWER SPONGE LAYER
AXIAL WIRE
UPPER SPONGE LAYER

A MODEL SALAMANDER

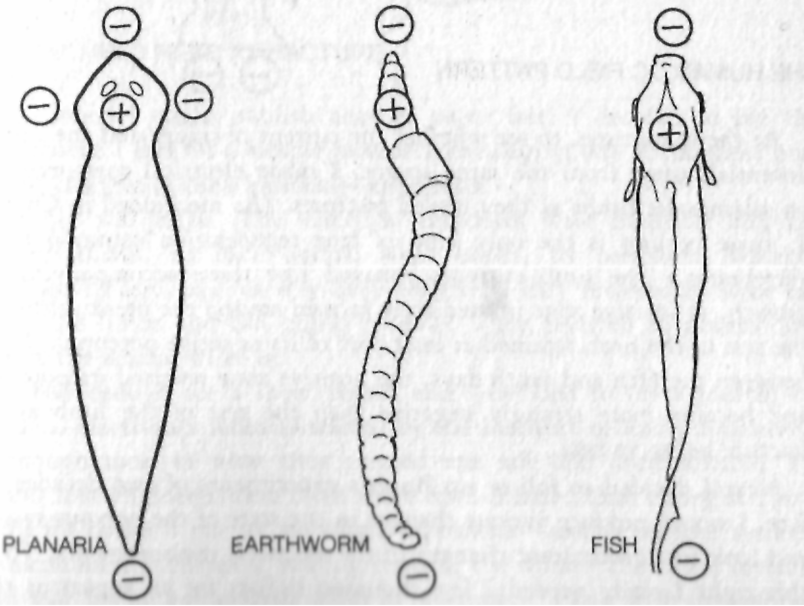
PLANARIA
EARTHWORM
FISH

A DC FIELD PATTERN APPEARS IN ALL ANIMALS

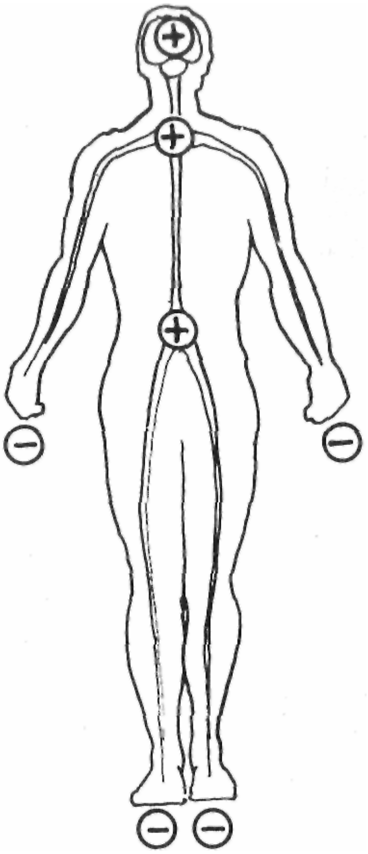

THE HUMAN DC FIELD PATTERN

At the same time, to see whether the current of injury and the surface potentials came from the same source, I made electrical measurements on salamander limbs as they healed fractures. (As mentioned in Chapter 1, bone healing is the only kind of true regeneration common to all vertebrates.) The limb currents behaved like those accompanying regrowth. A positive zone immediately formed around the break, although the rest of the limb retained at least part of its negative potential. Then, between the fifth and tenth days, the positive zone reversed its potential and became more strongly negative than the rest of the limb as the fracture began to heal.

Next I decided to follow up Burge's experiments of two decades before. I would produce various changes in the state of the nervous system and look for concomitant changes in the electrical measurements. To do this right I really needed a few thousand dollars for an apparatus that could take readings from several electrodes simultaneously and record them side by side on a chart. My chances of getting this money seemed

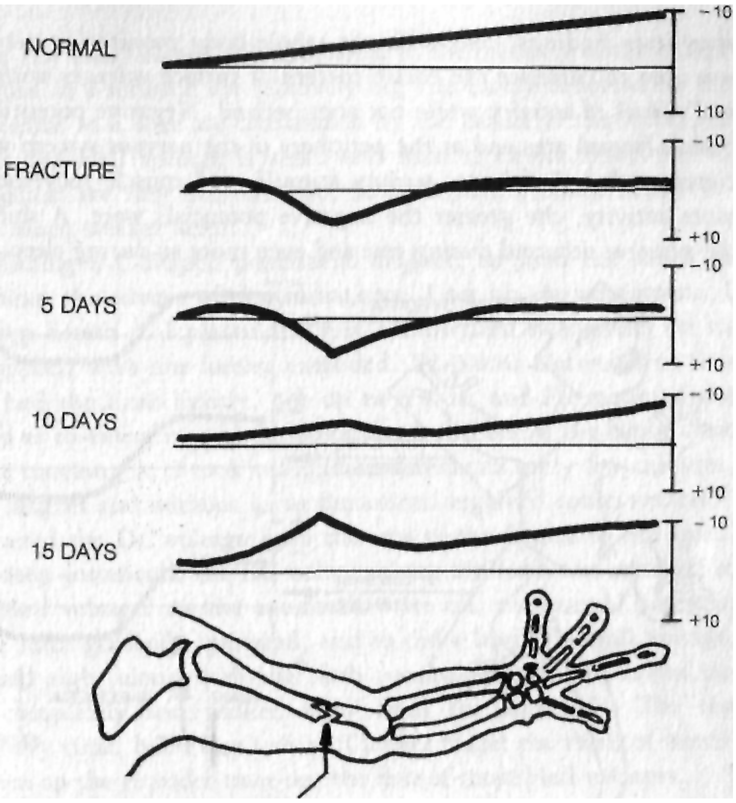

THE DC FIELD NEAR A FRACTURE

slim unless I could publish another paper fast. I decided to use the equipment I had for a simple measurement during one of the most profound changes in consciousness—anesthesia.

Burge was right. The electrical responses were dramatic and incontrovertible. As each animal went under, its peripheral voltages dropped to zero, and in very deep anesthesia they reversed to some extent, the limbs and tail going positive. They reverted to normal just before the animal woke up.

I had enough for a short paper, and I decided to try a journal on medical electronics recently started by the Institute of Radio Engineers. Although most of what they printed was safe and unremarkable, I'd found that engineers were often more open-minded than biologists, so I went for broke; I put in the whole hypothesis—analog nervous system, semiconducting currents, healing control, the works. The editor loved it and sent me an enthusiastic letter of acceptance, along with suggestions for further research. Best of all, I soon got another small grant approved and bought my multi-electrode chart recorder. Soon I had confirmed my anesthesia findings, and with the whole-body monitoring setup I also was able to correlate the entire pattern of surface voltages with the animal's level of activity while not anesthetized. Negative potentials in the brain's frontal area and at the periphery of the nervous system were associated with wakefulness, sensory stimuli, and muscle movements. The more activity, the greater the negative potentials were. A shift toward the positive occurred during rest and even more so during sleep. *

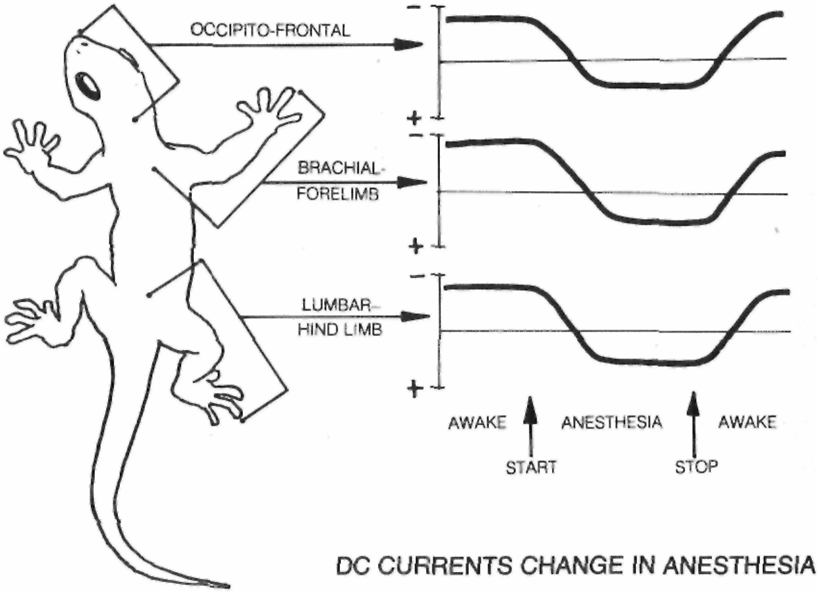

DC CURRENTS CHANGE IN ANESTHESIA

In my reading on solid-state electronics I found another way I could test for current in the salamander. Luckily it was cheap and easy; I could do it without buying more equipment. Best of all, it should work only if the current was semiconducting.

Suppose you think you have a current flowing through some conductor—a salamander's limb, for instance. You put it in a strong magnetic field so that the lines of force cut across the conductor at right angles. Then you place another conductor, containing no current, perpendicular to both the original conductor (the limb) and the magnetic field. If there is a current in whatever you're testing, some of the charge carriers will be deflected by the magnetic field into the other conductor, producing a voltage that you can measure. This is called the Hall voltage, after the gentleman who discovered it. The beauty of it is that it works differently for the three kinds of current. For any given strength of mag-

---

*I didn't know it until later, but another experimenter named H. Caspers made similar findings at about the same time.

netic field, the Hall voltage is proportional to the mobility of the charge carriers. Ions in a solution are relatively big and barely deflected by the field. Electrons in a wire are constrained by the nature of the metal. In both cases the Hall voltage is small and hard to detect. Electrons in semiconductors are very free to move, however, and produce Hall voltages with much weaker fields.

After finding a C-shaped permanent magnet, an item not much in demand since the advent of electromagnets, I set up the equipment. I took a deep breath as I placed the first anesthetized salamander on its plastic support, with one foreleg extended. I'd placed electrodes so that they touched the limb lightly, one on each side, and I'd mounted the magnet so as to swing in with its poles above and below the limb, close to yet not touching it. I took voltage measurements every few minutes, with the magnet and without it, as the animal regained consciousness. I also measured the DC voltage from the tips of the fingers to the spinal cord. In deep anesthesia, the DC voltage along the limb was zero and so was the Hall voltage. As the anesthetic wore off, the normal potential along the limb gradually appeared, and so did a beautiful Hall voltage. It increased right along with the limb potential, until the animal recovered completely and walked away from the apparatus. The test worked every time, but I don't think I'll ever forget the thrill of watching the pen on the recorder trace out the first of those Hall voltages.

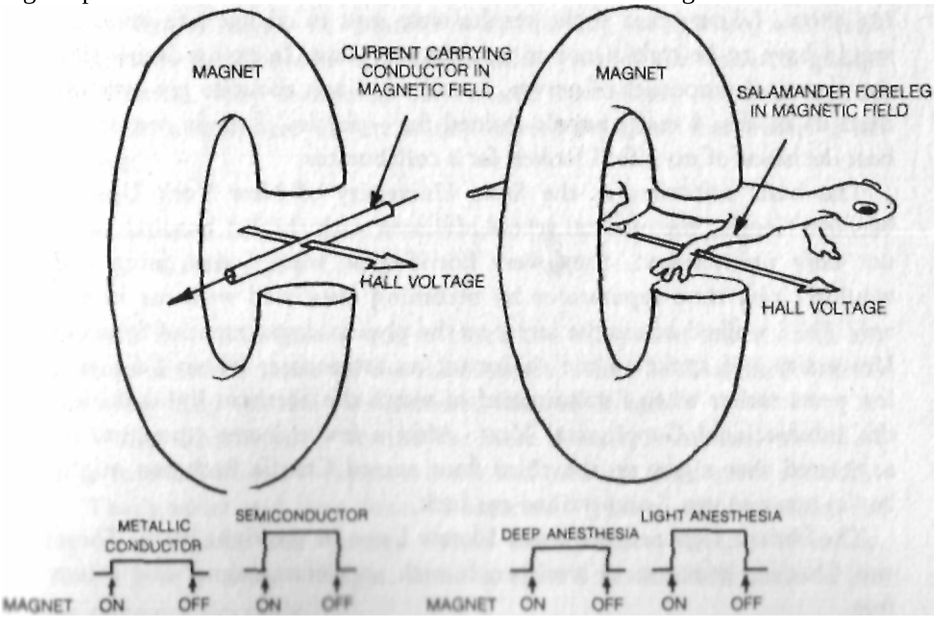
MAGNET
CURRENT CARRYING CONDUCTOR IN MAGNETIC FIELD
HALL VOLTAGE
MAGNET
SALAMANDER FORELEG IN MAGNETIC FIELD
HALL VOLTAGE
METALLIC CONDUCTOR
SEMICONDUCTOR
DEEP ANESTHESIA
LIGHT ANESTHESIA
MAGNET
ON
OFF
ON
OFF
MAGNET
ON
OFF
ON
OFF

THE HALL EFFECT

This experiment demonstrated unequivocally that there was a real electric current flowing along the salamander's foreleg, and it virtually proved that the current was semiconducting. In fact, the half-dozen tests I'd performed supported every point of my hypothesis.

Scientific results that aren't reported might as well not exist. They're like the sound of one hand clapping. For scientists, communication isn't only a responsibility, it's our chief pleasure. A good result from a clean, beautiful experiment is a joy that you just have to share, and I couldn't wait to see these data in print. I went for the top this time. The journal in American science is aptly named Science. Each issue reports on all fields from astronomy to zoology, so publication means a paper has more than a specialized significance. Mine was accepted, and I was jubilant.

With three major papers in three major journals after my first year of research, I felt I'd arrived. The world has a way of cutting you down to size, however, and in the science game the method is known as citation. No matter how important your paper is, it doesn't mean anything unless it's cited as a reference in new papers by others and you get a respectable number of requests for reprints. On both counts, I was a failure. I was learning how science treats new ideas that conflict with old ones.

I didn't stay discouraged long, though. I was doing science for the love of it, not for praise. I felt the concepts emerging from my reading and research were important, and I was passionately committed to testing them. I knew that if the results were ever to change any minds, I would have to be careful not to misinterpret data. In going deeper into the electrical properties of nerves, I realized, I was about to get over my head in an area I really wasn't trained for—physics. I made one of the best decisions of my life; I looked for a collaborator.

The basic scientists at the State University of New York Upstate Medical Center, the medical school affiliated with the VA hospital, were not only uninterested, they were horrified at what I was doing and wouldn't risk their reputations by becoming associated with me in any way. So, I walked across the street to the physics department of Syracuse University and spoke to the chairman, an astronomer whom I'd met a few years earlier when I volunteered to watch the northern lights during the International Geophysical Year. After a few minutes' thought, he suggested that a guy on the third floor named Charlie Bachman might be "as crazy as you," and wished me luck.

The instant I opened the door, I knew I was in the right place. There was Charlie, bent over a workbench with an electromagnet and a live frog.

Five
The Circuit of Awareness

Charlie and I talked all afternoon, beginning fifteen years of fruitful work together. For me, the best part was his friendship and his open mind. He, too, knew there was still a lot to learn. Our relationship also had a side effect of incalculable value: He sent some of his most talented graduate students over to my lab to do their thesis work and later to become my colleagues in research. Andy Marino, Joe Spadaro, and Maria Reichmanis each became an indispensable part of the research group. Like Charlie, they constantly contributed new ideas, and they helped create the atmosphere of intellectual adventurousness that makes a lab creative.

Closing the Circle

Charlie's first contribution was to check the equipment and confirm the measurements I'd made on the salamanders. After he'd satisfied himself that everything was real, we discussed what to do next.

"Well," Charlie said, "to find out more about this current, we'll have to go into the animal—expose a nerve and measure the current there."

"That's easier said than done," I objected. "Just to cut down into the leg of an animal will damage tissues and produce currents of injury.

That'll give spurious voltages. Besides, there'd be no stable place to put the reference electrode."

Charlie gave me a lesson in basic electronics. A voltage entering a wire will decrease as the current travels along, so there'll be a uniform voltage drop in each unit of length. All you need to do is put both electrodes along the conductor, with the reference electrode closer to what you think is the source. If you use a standard distance between the electrodes, you can compare the voltage drop along various wires and measure changes in the whole system from any segment of it.

All I had to do was the surgery. I decided to work on bullfrogs, whose hind legs were long and contained a nice big sciatic nerve. It was easy to find and could be exposed with just a little careful dissection, going between muscles instead of cutting through them. I was able to isolate over an inch of nerve with no bleeding or tissue damage, slipping a plastic sheet underneath so as not to pick up readings from the surrounding muscles. We measured the voltage gradient over a standard distance of 1 centimeter. It was the same from one frog to the next. In deep anesthesia it was absent or pretty small; as the anesthesia wore off, it became a constant drop of about 4 millivolts per centimeter, always gradually positive toward the spinal cord and negative toward the toes.

In some frogs we cut the nerve above the measuring site, whereupon the voltages disappeared—another indication that the current was actually in the nerve. Voltages returned a little later, but they weren't the same as before. We figured these secondary voltages were probably an artifact—a spurious measurement produced by extraneous factors—caused by currents of injury from the cut nerve itself or from the other tissues where I'd made the incision to cut the nerve.

Charlie then suggested that we make measurements on a longer section of nerve, and that was when we ran into a puzzle. The nicely reproducible voltages we'd found before couldn't be duplicated when we extended our measurement distance to 2 centimeters, close to the knee. We expected double the potential we'd found over the 1-centimeter distance, but often it was lower or higher than it should have been. I insisted that my dissection was producing local currents of injury that made our readings unpredictable. However, Charlie pointed out that I was a good frog surgeon and I didn't seem to be doing any more damage than before. He asked, "Could there be any difference in the nerve where you extended the dissection?"

"Not likely," I said. "The sciatic nerve does split up into two branches, but you only find them below the knee, when one goes to the front of the calf and one to the back."

"How do you know it doesn't separate before it gets to the knee?" he asked.

He was right. Not bad for a physicist! The nerve did divide, but the two parts were held together by the nerve sheath until they got past the knee. I was able to remove the sheath and isolate both portions. When we measured these, we found that the two sections were polarized in opposite directions. The voltage drop of the front branch was positive toward the toes. The posterior branch was polarized in the same direction as the sciatic trunk, but it always had a higher voltage gradient. The current in the front branch apparently flowed in the direction opposite to that in the rest of the nerve. The interesting thing was, when we added up the voltage increases from 1-centimeter lengths of the two branches—4 millivolts positive and 8 negative in a typical frog—we got roughly the same voltage gradient that we found in the main nerve, about 4 millivolts negative in every centimeter. At first that didn't make any sense.

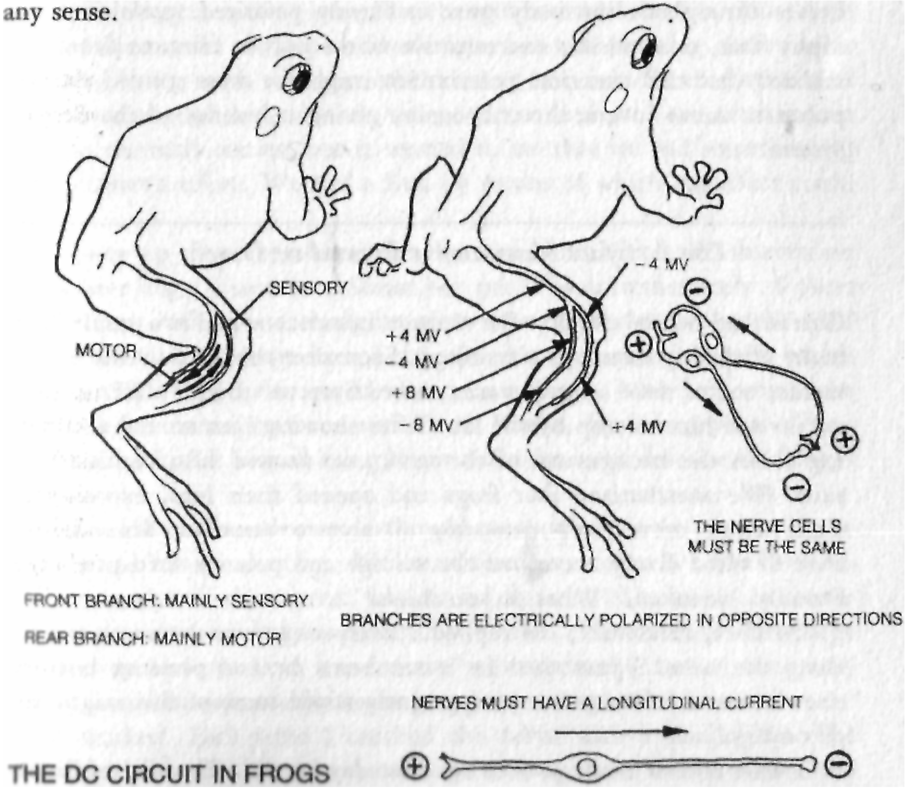
SENSORY
MOTOR
+4 MV
−4 MV
+8 MV
−8 MV
−4 MV
+4 MV
⊖
⊕
⊕
⊖
THE NERVE CELLS MUST BE THE SAME
FRONT BRANCH: MAINLY SENSORY
REAR BRANCH: MAINLY MOTOR
BRANCHES ARE ELECTRICALLY POLARIZED IN OPPOSITE DIRECTIONS
NERVES MUST HAVE A LONGITUDINAL CURRENT
⊕
⊖

THE DC CIRCUIT IN FROGS

On a hunch I took pieces of each nerve and sent them to the pathology department to have microscope slides made. I found that the fibers in the front fork were smaller than those in the other. A light bulb went on! The sciatic nerve is what's called a mixed nerve. It has both motor and sensory neurons. Sensory fibers are usually narrower than motor fibers, so it looked as though the front branch was all sensory, the back one all motor. Suppose the DC system also had incoming and outgoing divisions. We took readings from other nerves known to be all one type or the other. The femoral nerve along the front of the thigh is almost entirely motor in function, and, sure enough, it had an increasing negative potential away from the spine. The spinal nerves that serve the skin of a frog's back are sensory fibers, and they had increasing negative voltages toward the spine.

Now we saw that when you put motor and sensory nerves together into a reflex arc, the current flow formed an unbroken loop. This solved the mystery of what completed the circuit: The current returned through nerves, not some other tissue. Just as Gerard had found in the brain, nerves throughout the body were uniformly polarized, positive at the input fiber, or dendrite, and negative at the output fiber, or axon. We realized that this electrical polarization might be what guided the impulses to move in one direction only, giving coherence to the nervous system.

The Artifact Man and a Friend in Deed

Charlie had helped develop the electron microscope and as a result knew many of the big names in physiology. Soon after the sciatic nerve experiments, one of these acquaintances visited Syracuse to give a lecture, and we invited him to stop by the lab. After showing him around and talking about the background of the work, we showed him our latest results. We anesthetized four frogs and opened their legs, exposing all eight sciatic nerves and measuring all sixteen branches. The readings were flawless. Every nerve had the voltage and polarity we'd predicted. Proudly, we asked, "What do you think?"

"Artifact, all artifact," he replied. "Everyone knows there's no current along the nerve." Just then he remembered he had pressing business elsewhere and left in a hurry, apparently afraid some of this might rub off on him.

Charlie almost never swore, but that day he did. The gist of his remarks was that there sure was a difference between physicists and biologists. The former would at least look at new evidence, while the latter kept their eyes and minds closed. Thereafter we always referred to the "Artifact Man" when we needed a symbol of dogmatism.

We continued a few more observations on frog nerves. By now winter had come. That shouldn't have mattered—the lab temperature was the same all year, and the frogs didn't stop eating and hibernate as they would have in the wild—but there was a difference. The frogs' voltages were much lower, they stayed unconscious longer with the standard dose of anesthetic, and their blood vessels were much more fragile. Did they somehow sense the winter?

If the DC system was as we theorized, it would be influenced by external magnetic fields. In the Hall-effect experiment I'd already shown that it was, but I'd used a strong field, measuring thousands of gauss. The earth's magnetic field is only about half a gauss, but it does vary in a yearly cycle. At the time there was another scientist who was saying this weak field had major effects on all life. Frank A. Brown, a Northwestern University biologist who was studying the ubiquitous phenomenon of biological cycles—wavelike changes in metabolic functions, such as the alternation of sleep and wakefulness—was claiming that similar rhythms in the earth's magnetic field served as timers for the rhythms of life. Even though his evidence was good, no one paid any attention to him in the early sixties, but it seemed to me that we had something to offer Brown's effort. We had a link by means of which the effect could occur.

I wrote up the sciatic nerve measurements and added the observation on winter frogs. I sent it to Science but got it back immediately. I guess the editors had second thoughts after running my paper on the Hall effect. Next I tried the even better British equivalent, Nature, which took it. This time I also got some reprint requests. More important, the report led to correspondence with Frank Brown, beginning years of mutual feedback that helped bring about the discoveries described in Chapter 14.

I thought of one more way we could check whether the current in the nerves was semiconducting. We could freeze a section of nerve between the electrodes. If the current was carried by ions, they would be frozen in place and the voltage would drop to zero. However, if the charge carriers were electrons in some sort of semiconducting lattice, their mobility would be enhanced by freezing and the voltage would rise.

It worked. Each time I touched the nerve with a small glass tube filled with liquid nitrogen, the voltage shot upward. But perhaps I was damaging the nerve with the glass tube or through the freezing itself. Maybe the increase was merely a current of injury. To check, we simply cut the nerve near the spinal cord; the voltage gradient on the nerve went to zero, and then we applied the liquid nitrogen again. If the cold was really enhancing a semiconducting current, we should find no voltage now even after freezing the nerve—and we didn't. Therefore the increase in current wasn't due to artifact—damage to the nerve by freezing or touching it with the tube.

That settled it. Test after test had substantiated the direct-current system. Now we had to see where the concept would lead us and try to convince some of the Artifact Men along the way. We had lots of ideas for further work, but now the first priority was to get some reliable system of funding for ourselves.

I was continuing to have problems with the VA research office. After I'd gotten my second grant from that source, I soon found out that to have it approved and to be able to spend it were two different things. To order supplies—even things as simple as test tubes or electrode wire—I had to fill out a form and give it to the secretary of the research office. She had to fill out another form and get it signed by the research director. This form went to the supply service, where a clerk filled out a third form to actually order the stuff. Well, my orders stopped getting filled.

In the process of complaining I made friends with the secretaries and found out that the director was holding me up just by not signing my forms. His secretary solved my problem. The director was a procrastinator. A pile of papers would collect on his desk until his secretary told him they had to be taken care of right away. Then he would sign them all at once without looking at each one. His secretary, to whom I owe a tremendous debt, merely slipped my requests back into the middle of the pile, usually late on Friday afternoon. Several times he visited my lab, saw a new piece of equipment, and remarked, "I don't remember ordering that for you."

"You don't?" I replied sweetly. "We talked about it, and I had plenty of money left in the grant, so you said okay." It was better than arguing over each instrument, and I was careful not to overspend. I don't think he ever caught on.

Soon I encountered a more serious threat, however. Between the VA and the medical school I had a lot of bosses, and all of them were doing "research." However, the research service's annual report showed that I'd published more than all of my superiors put together on a few thousand bucks a year, while some of them were drawing forty or fifty thousand. I'd broken the old rule that you should never do more than your boss.

One of these fellows appeared in my lab one day. That was an event in itself, since he'd never been a supporter of mine; in fact, our relations were rather strained That day, however, he evinced great interest in what I was doing and made me an "offer I couldn't refuse."

"How would you like to have as much money as you need?"

I said that would be nice, but I wondered how it could come about.

"No problem. All you have to do is include me in the project. All I would expect in return is that my name would go on all publications."

It was a few seconds before I could believe I'd heard him right. Then I told him what he could do with his influence.

A few months later, I found out that the area surgical consultant, practically next to God in the VA hierarchy, was visiting the hospital to act on a report, made by my would-be "benefactor," that I was spending too much time on research and neglecting my patients. Fortunately, there was a lot of infighting among my superiors, and one higher than the guy who'd made the charge was supporting me. His motives were less to save a promising research program and more to embarrass the other man, but I was cleared.

It was also clear that I was courting disaster by relying on VA money. I needed outside support. I took time off from research to prepare two proposals. One, which I sent to the Department of the Army, emphasized the possibility that direct currents could stimulate healing. Since the Army's business produces quite a few wounds, I thought it would be interested, but it was not. The proposal was turned down promptly, but then a strange thing happened about a month later. I received a long-distance call from a prominent orthopedic surgeon, a professor at a medical school in the South. "I have a grant from the Army to study the possibility that direct currents might stimulate wound healing," he purred, "and I wonder if you might have any suggestions as to the best approach to use."

My God, were they all this sleazy out there? Of course, when I looked up his credentials, I found he had absolutely no background in bioelectricity. He'd just happened to be on the Army review committee, recommended disapproval, and then turned around and submitted the idea in his own name, now getting the go-ahead since he, a man with a reputation and friends on the review board, was going to do it, instead of some unknown upstart.

I sent the second application to the National Institutes of Health (NIH). I stayed within my specialty and proposed to study the solid-state physics of bone, eventually hoping to find out if direct currents could stimulate bone healing. The grant was approved, but only for enough money to do part of what I wanted. And although it was nice to have a cushion, a source not under local control, I nevertheless needed some political clout to stabilize the situation in Syracuse. I went directly to the dean of the medical school.

Carlyle Jacobson had seemed to be a nice guy, not the type to stand on ceremony or position, and I thought I could talk to him frankly. I gathered up reprints of my papers and went to his office.

"Sir," I began, "I've been doing research on direct current electrical effects in living things for the past four years. I've gotten some papers published in good journals, and I think this is an important piece of work. Nevertheless, I have great difficulty getting funds from the VA. My requests are blocked by the politicians on the committee. Meanwhile these same guys are spending five times as much as I get, and they don't publish a damn thing." I'm afraid I got carried away, but Dean Jacobsen just sat there listening until I'd finished.

"Have you done any experiments on the DC activity of the nervous system?" he asked.

This was an unexpected question, but I told him of our work on salamanders and frog nerves.

It turned out he'd done some research on nerves years ago—with Ralph Gerard! He was very enthusiastic. "You've gone much further than we ever did," he told me. "We never thought to relate the brain currents to a total-body system. How much do you need?"

I asked for $25,000 in each of the next two years, but explained that it had to be earmarked for me alone or I would never see it.

"Don't worry," he said. "Go right back to your lab. I'll get it for you. I wish I could work with you."

He must have been dialing Washington as the door closed behind me, for the next day the chief of research got a telegram from the VA Central Office authorizing the requested amount for me, and only me. He couldn't understand it, and I professed complete ignorance, too.

I figured nothing I did now could make the research director like me any less, so I made another move. I went to the hospital director and told him I needed more space. Having heard of my favor from Washington, he was most helpful, and soon I had a suite of rooms on the top floor.

Suddenly a whole new realm of research was within reach. Charlie and I hardly knew which way to turn. Our first and most important step was to hire Andy Marino as a technician. The salary meant much to him, and his intelligence and dedication meant even more to us. We were on our way.

The Electromagnetic Brain

If the current controlled the way nerves worked in the brain as well as in the rest of the body, then it must regulate consciousness to some extent.

Certainly the falling voltages in anesthetized salamanders supported this idea. The question was: Did the change in the current produce anesthesia? Apparently it did, for when I passed a minute current front to back through a salamander's head so as to cancel out its internal current, it fell unconscious. How this state compared with normal sleep was impossible to tell, but at least the animal was clinically anesthetized. As long as the current was on, the salamander was motionless and unresponsive to painful stimuli.

Was this real anesthesia, or was the animal just being continuously shocked? This certainly didn't seem to be the case, but the observation was so important and so basic to neurophysiology that I had to be sure. It was no easy task, however, for there were, and still are, few objective tests known for anesthesia, especially in salamanders. Brain waves had turned out to be useless in gauging depth of anesthesia in humans, because the one good marker—very slow delta waves—only showed up when the patient was dangerously close to death. Lacking any better idea, however, I used my new multi-electrode monitor to make EEG recordings of chemically anesthetized salamanders and found that they showed prominent delta waves even though they all recovered nicely. Delta waves would be my marker. The idea worked beautifully. Very small currents gave me delta waves on the EEG, the waves got bigger as I increased the current, and they all correlated with the animal's periods of unresponsiveness.

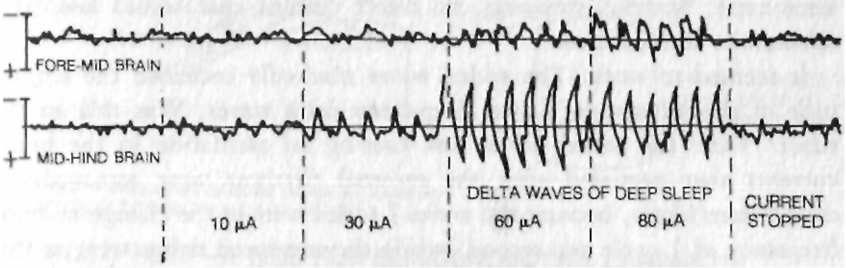

DC THROUGH THE HEAD ANESTHETIZES THE SALAMANDER

This result naturally led to the question: Did chemical anesthetics work by stopping the brain's electrical current? I couldn't see any way to get direct evidence one way or the other, but I thought maybe chemical anesthesia could be reversed by putting current into the brain in the normal direction. I found this could be done only to a certain extent. I could get a partial return of the higher frequency waves in the EEG, and the anesthesia seemed to become shallower, but I couldn't get a drugged salamander to wake up and walk away.

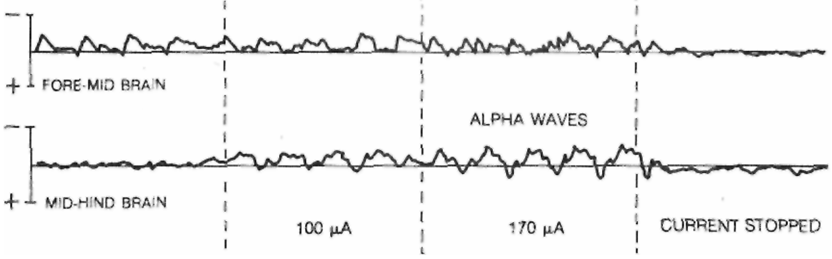

DC THROUGH THE HEAD PARTIALLY AWAKENS THE SALAMANDER

In the course of these observations, I found that when the head voltage was dropping as a chemical anesthetic took hold, specific slow waves always appeared briefly in the recordings. They were at the low end of the delta frequencies, 1 cycle per second or even less, and they also showed up when the voltage came back as the drug wore off. To find out if these waves always signaled a major change in the state of consciousness, I decided to use a standard amount of direct current to produce anesthesia, measure the amplitude (size) of the delta waves in the EEG, and then add some one-second waves of my own to the current I was putting into the animal's head. In other words, I would introduce some "change-of-state" waves from outside and see if they produced a shift in the EEG. I couldn't record the EEG simultaneously, because the waves I added would appear on the trace, so I rigged up a switch to cut out the added waves after a minute and turn on the EEG recorder at the same time, without stopping the direct current that would keep the salamander unconscious.

It seemed to work. The added waves markedly increased the amplitude of the salamander's own deep-sleep delta waves. Was this an artifact? Were the added waves just causing an oscillation in the brain currents that persisted after the external rhythms were removed? It didn't seem likely, because the waves I added were at the change-of-state frequency of 1 cycle per second, while the measured deltas were at the deep-sleep frequency of 3. However, an additional test was possible. I could add waves of other frequencies and see if they worked as well as 1 cycle per second. They didn't; in fact, as the frequency of the added waves increased over that rate, the deep-sleep delta waves got smaller. The one-second waves were a marker of major shifts in consciousness.

This line of work corroborated one of the main points of my hypothesis. Direct currents within the central nervous system regulated the level of sensitivity of the neurons by several methods by (hanging the amount of current in one direction, by changing the direction of the current (reversing the polarity), and by modulating the current with slow waves. Moreover, we could exert the same control from outside by putting current of each type into the head. This was exciting. It opened up vast new possibilities for a better understanding of the brain. It was still on the edge of respectability, too, since it was a logical consequence of the work done by Gerard and his co-workers. The next experiment was harder to believe, however.

I figured the brain currents must be semiconducting, like those in the peripheral nerves. I thought of looking for a Hall voltage from the head but reasoned the brain's complexity would make any results questionable. Then I thought of using the effect backward, so to speak, measuring a magnetic field's action on the brain rather than on the production of the Hall voltage. Since the Hall voltage was produced by diverting some of the charge carriers from the original current direction, a strong enough magnetic field should divert all of them. If so, such a field perpendicular to the brain's midline current should have the same effect as canceling out that normal current with one applied from the outside. The animal should fall asleep.

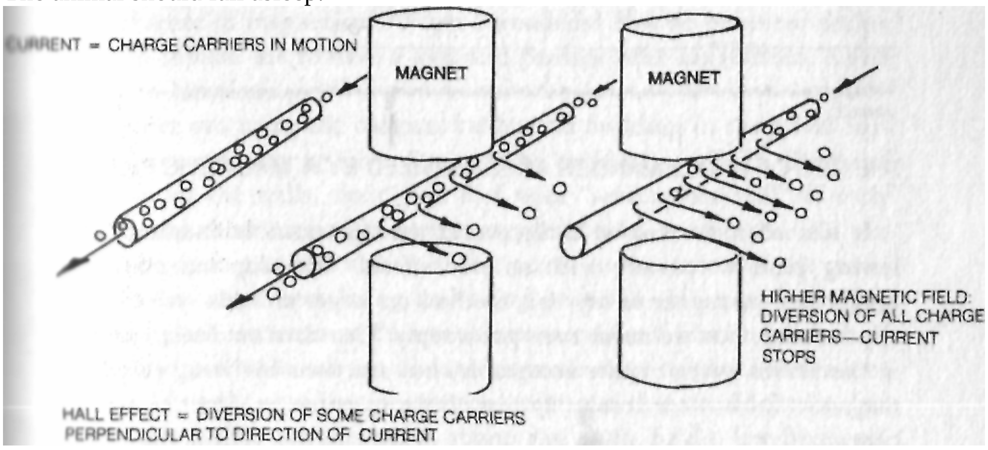
CURRENT = CHARGE CARRIERS IN MOTION
MAGNET
MAGNET
HIGHER MAGNETIC FIELD: DIVERSION OF ALL CHARGE CARRIERS—CURRENT STOPS
HALL EFFECT = DIVERSION OF SOME CHARGE CARRIERS PERPENDICULAR TO DIRECTION OF CURRENT

THE HALL EFFECT—A TEST FOR SEMICONDUCTING CURRENTS

We tranquilized a salamander lightly, placed it on a plastic shelf between the poles of a strong electromagnet, and attached electrodes to measure the EEC As we gradually increased the magnetic field strength, we saw no change—until delta waves appeared at 2,000 gauss. At 3,000 gauss, the entire BEG was composed of simple delta waves, and the animal was motionless and unresponsive to all stimuli. Moreover, as we decreased the strength of the magnetic field, normal EEG patterns returned suddenly, and the salamander regained consciousness within seconds, This was in sharp contrast to other forms of anesthesia.

With direct currents, the EEG continued to show delta waves for as long as a half hour after the current was turned off, and the animals remained groggy and unresponsive just as after chemical anesthesia.

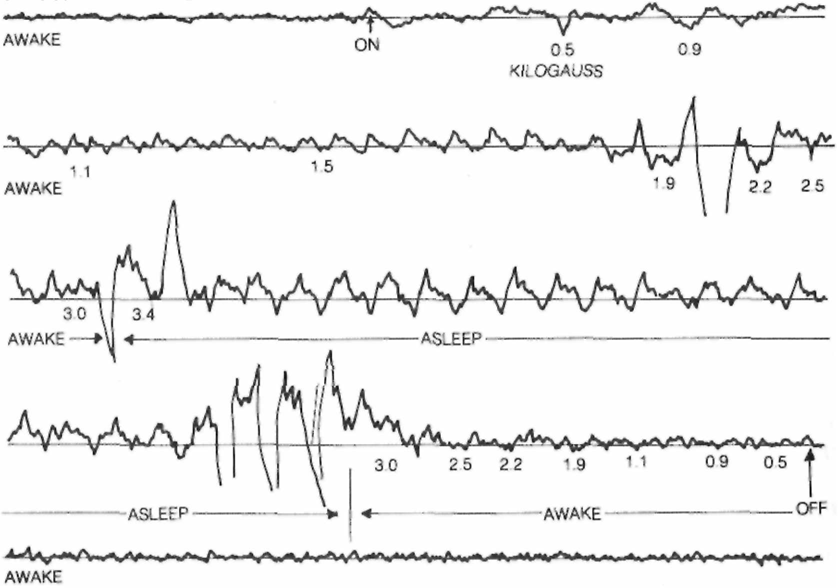

THE EEG OF A SALAMANDER ANESTHETIZED BY A MAGNETIC FIELD

It seemed to us that we'd discovered the best possible anesthetic, allowing prompt recovery with no side effects. We proposed getting a bigger electromagnet to try this method on larger animals and eventually humans, but we never even got a reply. Our data on direct currents in the nerves weren't quite acceptable, but reactions by living things to magnetic fields were absolutely out of the question in America at that time.

I was flabbergasted, therefore, to receive a phone call from one of the most prominent biologists in the Harvard-MIT orbit. He told me, "We're in the process of setting up an international conference on high-energy magnetic fields at MIT, and we've received a number of questions from respectable scientists in other countries asking why there is to be no session on biological effects of magnetic fields. This is a totally new idea to us, and we really don't believe there are such effects, but some of these fellows are persistent. We've searched the scientific literature and found your paper on the Hall effect. Since you steem to be the only person in this country doing any work along these lines, let me ask if you think there's anything to it."

I allowed as humbly as I could that there just might be something there, and told him about the latest experiments. There was a long pause filled with disdain. Then I added that Professor Bachman had been working with me. That changed the tone dramatically; this man also knew Charlie through work on the electron microscope during the war. He asked if I would organize the session and arrange for additional speakers.

There weren't too many investigators to choose from, and some of them were doing very slipshod work. I invited Frank Brown and selected a few others on the basis of published work. I'd just about finished when I got another call. A man with a thick German accent introduced himself as Dietrich Beischer.

"I have read your paper on the Hall effect," he began, "and I think we have much common interest." He explained that he was studying magnetic field bioeffects for the Navy and had done much work that wasn't published openly. At the time he was conducting a large experiment on human volunteers to check for effects from a null field, a complete absence of magnetism. When I wondered how he produced such a state, he invited me to have a look and perhaps offer suggestions. So off I went to Maryland.

Beischer was using the compass calibration building in the Naval Surface Weapons Center at Silver Spring. The building was huge. Electrical cables in all the walls, floor, and roof were "servo-connected" (directly cued) to the three axes of the earth's magnetic field, so that the field was canceled out in a sphere about 20 feet across in the center of the structure. Several men were living and being tested in this area. I was impressed by the resources at Beischer's command, and I had a good time, but I wondered what use any discoveries made there might ultimately be put to. My only contribution was to point out that the enclosure had been built before anyone knew about the earth field's low-frequency components, micropulsations ranging from less than 1 to about 25 cycles per second, that were far weaker than the planet's electromagnetic field as a whole. Consequently Beischer's subjects were still exposed to a very weak magnetic field pulsing at these frequencies, and I suspected that that component might be one of the most important for life, because all brain waves were in exactly the same range. Perhaps as a result of this factor, the null-field experiment was turning out to be inconclusive, but I asked Beischer to attend the MIT meeting and present some previous data suggesting that electromagnetic fields could affect embryonic development.

For my own presentation I decided against trying to compress all the work I'd done so far into a few minutes before a skeptical audience. Instead I offered some evidence that Charlie and I had gathered with the help of psychiatrist Howard Friedman, which showed a possible relationship between mental disturbances and solar magnetic storms. I'll discuss this study further in Chapter 14.

The MIT meeting went well. The field of bioelectromagnetism was still young, and the researchers in it didn't make many converts among the mainstream biologists. As usual, we found the physicists more inclined to listen. However, we drew inspiration from each other. I returned to the lab more determined than ever to elucidate the links that I knew existed between electromagnetic energy and life.

Charlie, Howard, and I decided to find out how the brain's DC potentials behaved in humans. The electrodes we'd been using on salamanders couldn't be scaled up for people, but within a week Charlie invented some that would give us equally precise readings from the human head. We immediately found that the back-to-front current varied with changes in consciousness just as in salamanders. It was strongest during heightened physical or mental activity, it declined during rest, and it reversed direction in both normal sleep and anesthesia. This knowledge led directly to the experiments, described in Chapter 13, that taught us much about how hypnosis and pain perception work.

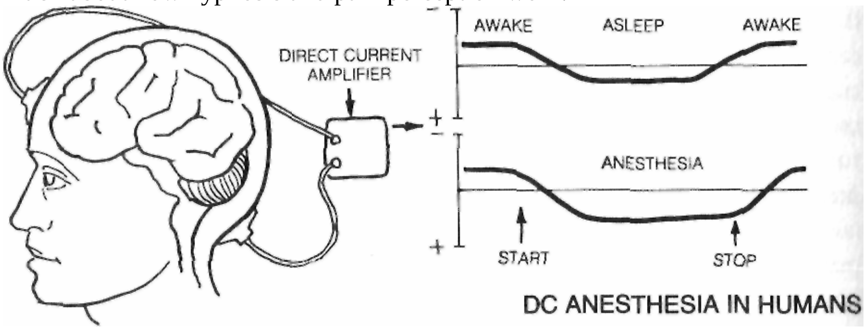

DC ANESTHESIA IN HUMANS

At this point I received an invitation from Meryl Rose to speak at the big event in the world of animal science, the International Congress of Zoology. This is not just a yearly convention; it is convened only when its directors agree that there has been enough progress to warrant a meeting. This session, in August 1963, was only the sixteenth since the first one had been set up in 1889. It was an honor to be there, and the conference itself was especially important, since it was one of the first times science formally addressed such ecological emergencies as pesticide pollution, the protection of vanishing species, overpopulation, and urban sprawl. The high point tor me came when I gave my paper and saw in the audience Dr. Ralph Bowen, my college biology professor, a kind but exacting teacher who'd inspired me with his unique combination of scientific discipline and respect for life. Afterward, with characteristic caution, he said something like "That's not too bad, Becker. I'd like to see you keep going in this research."

When I assured him that, despite my M.D. degree, I was still committed to basic biology, he said, "I hope so, but remember, it won't be easy. To change things is never popular." His encouragement meant a lot to me, and I was happy to be able to show him that I'd amounted to something.

A lot had happened in the four work-filled years since I'd begun studying the current of injury. That first experiment had opened a door into a great hall with passageways leading off in all sorts of fascinating directions. This was really the life! Without leaving the laboratory I'd gone on a journey of exploration as exciting as trekking through the uncharted wilds of New Guinea. Our work on nerves and the brain was leading toward a whole new concept of life whose implications only gradually became apparent. Meanwhile, my colleagues and I were continuing to investigate the processes of healing, leading to insights and practical applications that more than justified my enthusiasm.

Six

The Ticklish Gene

Despite my fascination with fundamental questions about the nature of life, I was, after all, an orthopedic surgeon, and I was eager to find things that would help my patients. In addition, to convince the Artifact Man and all his brothers, Charlie and I were looking for some direct test of semiconduction in living tissues. The Hall effect and the freezing of frog nerves each demonstrated a characteristic of semiconduction but didn't confirm it in standard engineering terms. Unfortunately, all the direct tests then known worked only with crystals. You needed a material you could carve into blocks, something that didn't squish when you put an electrode on it. The only possibility was bone.

To many biologists and physicians, bones are pretty dull. They seem like a bunch of scarecrow sticks in which nothing much happens, plain props for a subtler architecture. Many of my patients were in sad shape because doctors had failed to realize that bone is a living tissue that has to be treated with respect. It's a common misconception that orthopedic surgery is like carpentry. All you have to do is put a recalcitrant fracture together with screws, plates, or nails; if the pieces are firmly fixed after surgery, you're done. Nothing could be further from the truth. No matter how firmly you hold them together, the pieces will come loose and the limb can't be used if the bone doesn't heal.

The Pillars of the Temple

The skeleton doesn't deserve this cavalier treatment. The development of bones by the first true fishes of the Devonian era nearly 400 million years ago was a remarkable achievement. It enabled animals to advance, in both senses of the word, quickly and efficiently. Since bone is inside the body, it can live and grow with the animal, instead of leaving it defenseless as an external skeleton does when cast off during a molt. It's also the most efficient system for attaching muscles and increasing the size of animals.

Bone is extraordinary in structure, too. It's stronger than cast iron in resisting compression but, if killed by X rays or by cutting off its blood supply (barely adequate to start with), it collapses into mush. The part that's actually alive, the bone cells, comprises less than 20 percent of the whole. The rest, the matrix, isn't just homogeneous concrete, either. It's composed of two dissimilar materials—collagen, a long-chain, fibrous protein that's the main structural material of the entire body, and apatite, a crystalline mineral that's mainly calcium phosphate. The electron microscope shows that the association between collagen and apatite is highly ordered, right down to the molecular level. The collagen fibers have raised transverse bands that divide them into regular segments. The apatite crystals, just the right size to fit snugly between these bands, are deposited like scales around the fibers.

This intricacy continues at higher levels of organization. The collagen fibers lie side by side in layer upon layer wound in opposed spirals (a double helix) around a central axis. The bone cells, or osteocytes, are embedded in these layers, which form units a few millimeters long, called osteones. The center of each osteone has a small canal in which runs a blood vessel and a nerve. The osteones in turn are organized so as to lie along the lines of maximum mechanical stress, producing a bone of the precise shape best able to withstand the forces applied to it.

Bone has an amazing capacity to grow, which it does in three different ways. In childhood each long bone of the limbs has one or two growth centers, called epiphyseal plates. Each is a body of cartilage whose leading edge grows continuously while its trailing edge transforms into bone. When the bone is the right length, the process stops, and the remaining cartilage forms the bony knob, or epiphysis, at the end of the bone. The "closure" of the epiphyses is an index of the body's maturation.

Bone cannot heal. That sounds like a conundrum but it's literally true. Fractures knit because new bone made from other tissues unites the fracture ends. Although we sometimes speak of bone growth as part of fracture healing, the old, preexisting bone doesn't have the capacity to grow. As mentioned in Chapter 1, there are two tissues that produce new bone at a fracture site. One is the periosteum, the bone's fibrous covering. It's the cells of the periosteum's innermost layer that have the

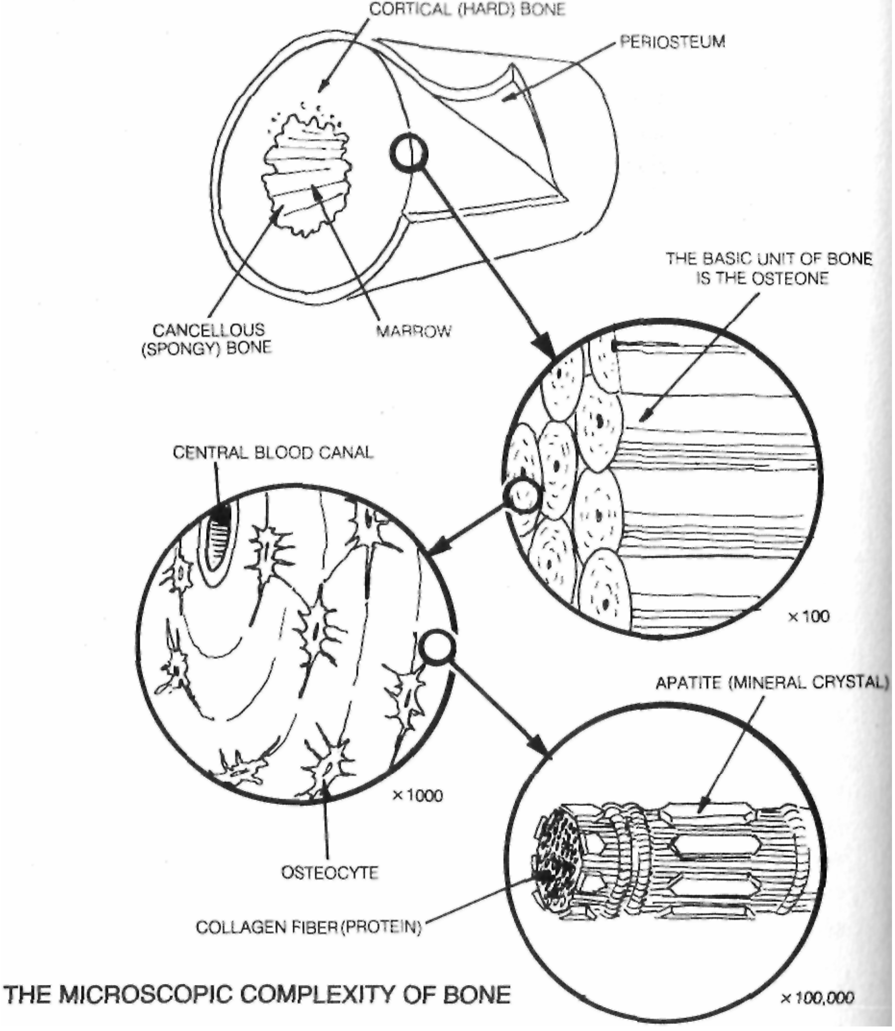
CORTICAL (HARD) BONE
PERIOSTEUM
THE BASIC UNIT OF BONE IS THE OSTEONE
CANCELLOUS (SPONGY) BONE
MARROW
CENTRAL BLOOD CANAL
×100
APATITE (MINERAL CRYSTAL)
×1000
OSTEOCYTE
COLLAGEN FIBER (PROTEIN)
×100,000

THE MICROSCOPIC COMPLEXITY OF BONE

power of osteogenesis, or bone formation. After a fracture, these cells are somehow turned on. They begin to divide and some of the daughter cells turn into osteoblasts, cells that make the collagen fibers of bone. Apatite crystals then condense out of the blood serum onto the fibers.

The other tissue that forms new bone to heal a fracture is the marrow. Its cells dedifferentiate and form a blastema, filling the central part of the fracture. The blastema cells then turn into cartilage cells and later into more osteoblasts. This process is true regeneration, following the same sequence of cell changes as the regrowing salamander limb.

Whatever a physician does to repair a fracture, he or she must protect the periosteum and marrow cavity from harm. Unfortunetely, too often

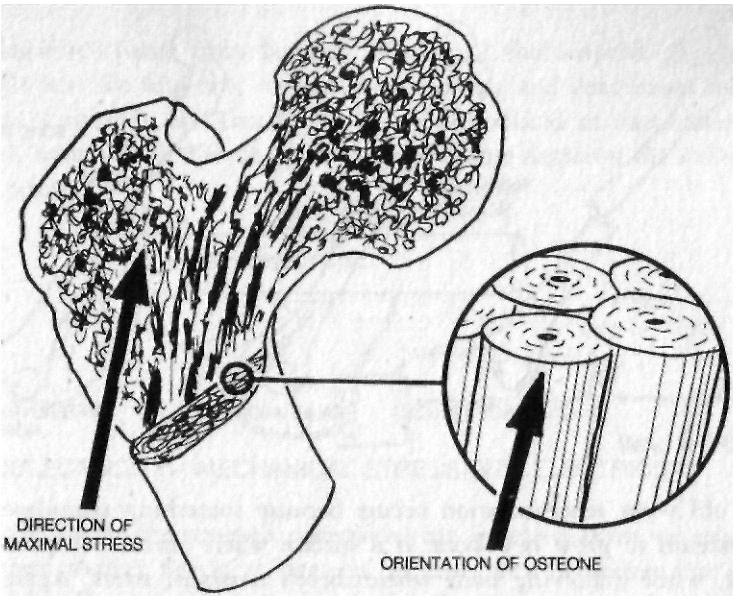

BONES-DESIGNED FOR MAXIMUM STRENGTH

the application of plates, screws, and nails does just the opposite and, rather than helping nature, the treatment impedes healing.

From a researcher's point of view, the question here is: What activates the cells of the periosteum and marrow? In the case of the marrow, we can expect it to be the same factor that switches on the cells in a salamander's amputated leg.

There's a third process of growth that's unique to bone. It follows Wolff's law, which is named after the orthopedic surgeon J. Wolff, who discovered it at the end of the nineteenth century. Basically, Wolff's law states that a bone responds to stress by growing into whatever shape best meets the demands its owner makes of it. When a bone is bent, one side is compressed and the other is stretched. When it's bent consistently in one direction, extra bone grows to shore up the compressed side, and some is absorbed from the stretched side. It's as though a bridge could sense that most of its traffic was in one lane and could put up extra beams and cables on that side while dismantling them from the other. As a result, a tennis player or baseball pitcher has heavier and differently contoured bones in the racket arm or pitching arm than in the other one. This ability is greatest in youth, so in childhood fractures it's often best to put the bone ends together gently by manipulation without surgery, settling for a less than perfect fit. Sometimes the hardest part is convincing the parents that a modest bend will straighten itself out in a few months in accordance with Wolff's law.

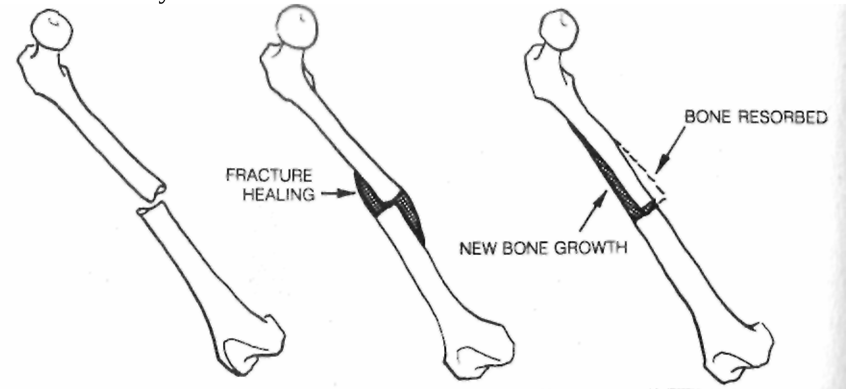
FRACTURE NOT REDUCED
FRACTURE HEALING
HEALED ANGULATED
BONE RESORBED
NEW BONE GROWTH
DIFFERENTIAL GROWTH NORMAL

WOLFF'S LAW

Wolffs-law reorganization occurs because something stimulates the periosteum to grow new bone at a surface where there's compressional stress, while dissolving bone where there's tensional stress. Again, the question for researchers is: What turns on the periosteal cells?

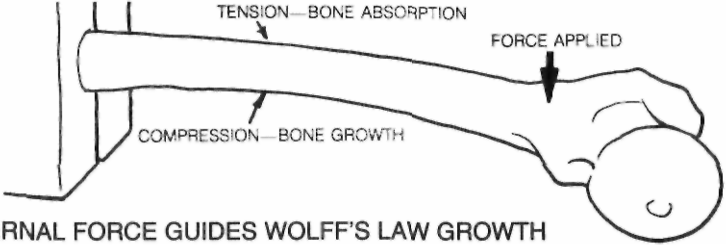
TENSION—BONE ABSORPTION
FORCE APPLIED
COMPRESSION—BONE GROWTH

EXTERNAL FORCE GUIDES WOLFF'S LAW GROWTH

After I'd finished giving my paper on the salamander's current of injury at the orthopedic meeting early in 1961, several people came up to the stage to ask questions. Among them was Andy Bassett, a young orthopedic surgeon who was doing research at Columbia. In our conversation we came up with an angle for investigating Wolff's law—piezoelectricity. Simply put, this is the ability of some materials to transform mechanical stress into electrical energy. For example, if you bend a piezoelectric crystal hard enough to deform it slightly, there'll be a pulse of current through it. In effect, the squeeze pops electrons out of their places in the crystal lattice. They migrate toward the compression, so the charge on the inside curve of a bent crystal is negative. The potential quickly disappears if you sustain the stress, bur when you release it, an equal and opposite positive pulse appears as the electrons rebound before settling back into place.

Since I'd shown that a stronger-than-normal negative current preceded regeneration, Bassett suggested that maybe bone was iezoelectric and the negative charge from bending stimulated the adaptive growth of Wolff's law. To find out, we tested both living and dead bones from a variety of animals, and found that bending produced an immediate potential, as expected. The compressed side became negative; the stretched side, positive.*

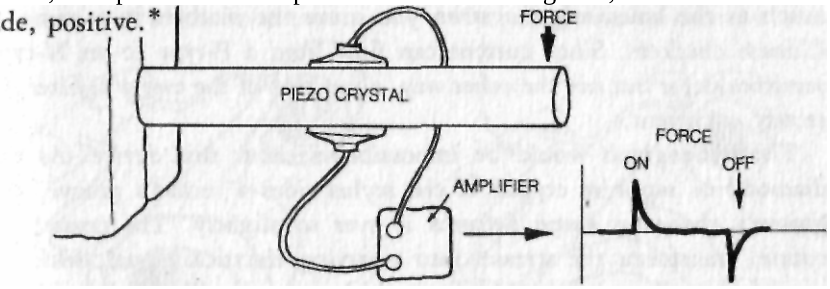
FORCE
PIEZO CRYSTAL
FORCE
ON
OFF
AMPLIFIER

PIEZOELECTRICITY—MECHANICAL STRESS INTO ELECTRICITY

Furthermore, the reversed potentials that appeared when we released the stress weren't nearly as large as the first ones. This was just as it should be. If a negative voltage was the growth stimulus, there had to be some way to cancel out the positive rebound voltage; otherwise it would have negated the growth message. In electronic terms, there had to be a solid-state rectifier, or PN junction diode.

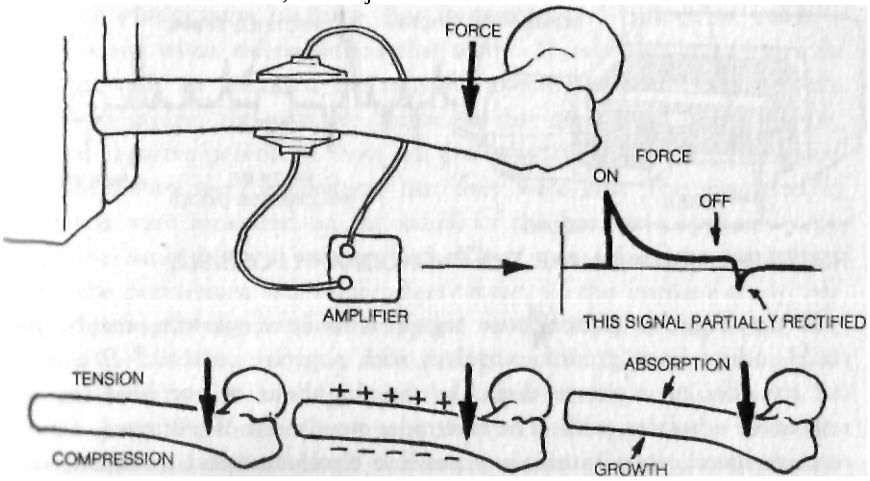
FORCE
FORCE
ON
OFF
AMPLIFIER
THIS SIGNAL PARTIALLY RECTIFIED
TENSION
COMPRESSION
+ + + + +
— — — — —
ABSORPTION
GROWTH

PIEZOELECTRICITY IN BONE

---

*After writing up this experiment, we found that it had been done before. Iwao Yasuda, a Japanese orthopedist, had shown that bone was piezoelectric back in 1954; he and Eiichi Fukada, a physicist, had confirmed the fact in 1957. We made note of their prior observations but published our paper anyway, since our techniques were different and ours was the first report in English.

Despite the initimidating names, this device is fairly simple. It's a filter that screens out either the positive (P) or the negative (N) part of a signal. As mentioned in Chapter 4, current can flow through a crystal lattice either as free electrons or as "holes" that can shift their positions much as the holes migrate when you move the marbles in a game of Chinese checkers. Since current can flow from a P-type to an N-type semiconductor but not the other way, a junction of the two will filter, or rectify, a current.

The phonograph would be impossible without this device. As the diamond or sapphire crystal of the stylus rides a record's groove, the groove's changing shape deforms it ever so slightly. The crystal, of course, transforms the stresses into a varying electrical signal, which is amplified until we can hear it. It would be an unintelligible hum, however, if we heard both the deformation pulse and the release pulse. Therefore we place a rectifier in the circuit. It passes current in one direction only, so the impulses don't cancel each other out. The signal is rectified, and when we feed it to a loudspeaker we hear music. Bassett and I felt sure we were seeing evidence of rectification in the fact that bone's release pulse was much smaller than the one from stress.

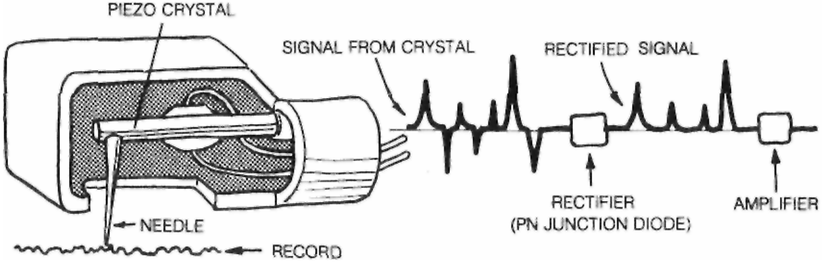

RECTIFIED SIGNALS MAKE THE PHONOGRAPH POSSIBLE

If the negative piezoelectric signal stimulated growth, maybe we could induce bone growth ourselves with negative current.* We tried out the idea on eighteen dogs. In the thighbone of one hind leg we implanted a battery pack. The electrodes were made of platinum, a nonreactive metal, to minimize any possible electrochemical irritation, and we inserted them through drill holes directly into the marrow cavity. As controls, some of the devices didn't have a battery. After three weeks we found that the active units had produced large amounts of new bone around the negative electrode but none around the positive. In the controls, there was no growth around either electrode.

---

*Again Dr. Yasuda and his colleagues had already done so, but their results seem to have been due to bone's ability to grow in response to irritation from the electrodes. They used alternating current, which is now known to have no direct growth-stimulating effect.

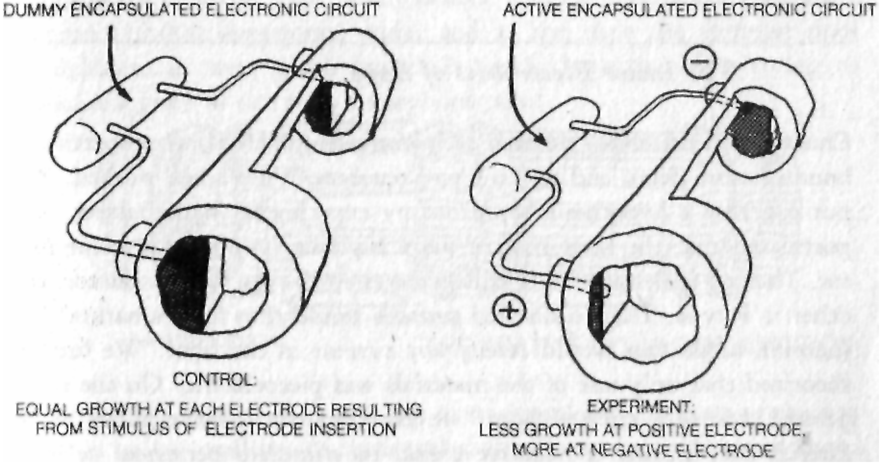
DUMMY ENCAPSULATED ELECTRONIC CIRCUIT
ACTIVE ENCAPSULATED ELECTRONIC CIRCUIT
⊖
⊕
CONTROL:
EQUAL GROWTH AT EACH ELECTRODE RESULTING FROM STIMULUS OF ELECTRODE INSERTION
EXPERIMENT:
LESS GROWTH AT POSITIVE ELECTRODE, MORE AT NEGATIVE ELECTRODE

ELECTRICAL STIMULATION OF BONE GROWTH

The results were exciting, but in retrospect I believe we made a serious error when we published this study. In our own minds and in print as well, we confused the negative potentials of the salamander's current of injury, the negative potentials that stimulated bone growth, and the negative potentials from the piezoelectric study. We proceeded as though they were equivalent, but they were not. The piezoelectric potentials were measured on the outside of the bone and appeared only when mechanical stress was applied. They were transitory, and most likely the periosteum was their target tissue. In the implant study, we used continuous direct current applied to the inside of the bone, the marrow cavity. What we were stimulating was the DC control system of regenerative fracture healing, not the piezoelectric control system of Wolff's law. We didn't clearly indicate the difference to the scientific reader, and this led to much confusion, some of which still persists twenty years later. As a result, many scientists think electricity stimulates bone growth because bone is piezoelectric. Most of these people don't realize that bone itself doesn't grow when a fracture heals. Moreover, everyone who has proceeded from our technique—and it's being used today to heal nonunions (see Chapter 8)—has done basically the same thing, continuously stimulating the bone marrow. No one has tried to stimulate the periosteum, as the pulsed piezoelectric signal does.

Our confusion also helped the scientific establishment accept the "trivial" electrical stimulation of bone by considering it something unique to bone. The relationship between our experiment and regeneration proper was lost.

The Inner Electronics of Bone

Charlie Bachman and I decided to investigate the electrical properties of bone in more detail and try to figure out how Wolff's law worked. We put together a hypothesis based on my experiments with Bassett. We postulated that the bone matrix was a biphasic (two-part) semiconductor. That is, either apatite or collagen was an N-type semiconductor; the other, a P type. Their connected surfaces would thus form a natural PN junction diode that would rectify any current in the bone. We further theorized that only one of the materials was piezoelectric. On the compressed side of a stressed bone, we expected the positive pulses to be filtered out, leaving a negative signal to stimulate periosteal cells to grow new bone.

We made several pairs of sample blocks, cut side by side from pieces of bone removed from patients for medical reasons. From one member of each pair we chemically removed the apatite. The other we treated with a compound that dissolved the collagen. The resulting pure collagen was yellowish and slightly rubbery, and the apatite pure white and brittle, but otherwise both blocks still looked like bone. Our first step was to test bone's component materials separately for semiconduction and piezoelectricity. Collagen turned out to be an N-type semiconductor and apatite a P type.

Then we tested our samples for piezoelectricity in the same way that Bassett and I had previously tested whole bone. We expected that apatite would be the only one to show an effect, since it was a crystal. However, collagen turned out to be a piezoelectric generator, while apatite was not. We now had the makings of a PN junction—two semiconductors, one an N type, the other a P type, joined together in a highly organized fashion.

Now came the crucial part of our hypothesis. We had to figure out a way to test for rectification at the PN junction. It was an important crossroads.

Here we rn up against what's known in the trade as a technical problem. To test for rectification we had to put one electrode on the collagen and one on the apatite as they appeared in whole bone. Unfortunately, the apatite crystals are each only 500 angstroms long. Now, the angstrom (named after the Swedish pioneer of spectroscopy Anders Jonas Angstrom) was invented for measuring atoms and molecules, and it is not large. Five hundred of them are only one tenth as long as a single wave of green light. Even today's thinnest microelectrodes are 1 micron (10,000 angstroms) wide, and at the time the thinnest ones available to us were much larger. It would have been like trying to measure a grain of rice with a telephone pole.

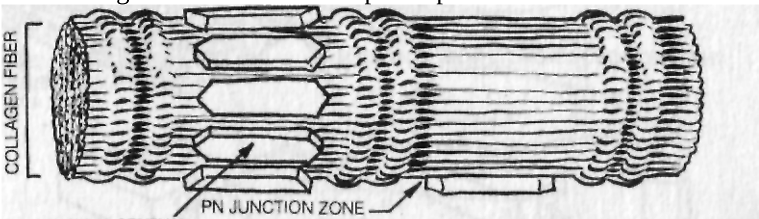

THE APATITE-COLLAGEN JUNCTION

We would have to do it in a sort of statistical way. Because of the way bone is built—millions of little scales glued onto larger fibers arranged in more or less lengthwise spirals along the osteones—I reasoned thus: If we put an electrode on the lengthwise cut, we'd be contacting mostly apatite, while an electrode on a face cut across the grain should connect to a greater proportion of collagen. If that method of electrode hookup worked and if we had a rectifier in our bones, then we'd be able to pass current through our samples only in one direction. That was exactly what happened. Our bone samples weren't as efficient as a commercial rectifier, but the amount of current we could put through them from a battery of constant voltage was much greater in one direction than in the other.

Current flowing "uphill," against the normal flow from P to N semiconductors, is called a reverse bias current, and we used it to look for photoelectric effects. Many semiconductors absorb energy from light, and any current flowing through the material gets a boost. We arranged our apparatus so only a small spot of light shone on the bone, because our silver electrodes were slightly sensitive to light and could produce a real artifact. With the voltage constant, the light produced an unmistakable increase in the current. Now, if bone really contained a rectifier, the photoelectric effect should be sensitive to the current's direction. The current in reverse bias should rise more with the same light intensify than the current in forward bias The experiment was simple. We reversed the battery and turned on the light. The amperage rose higher than before. The rectifier was real.

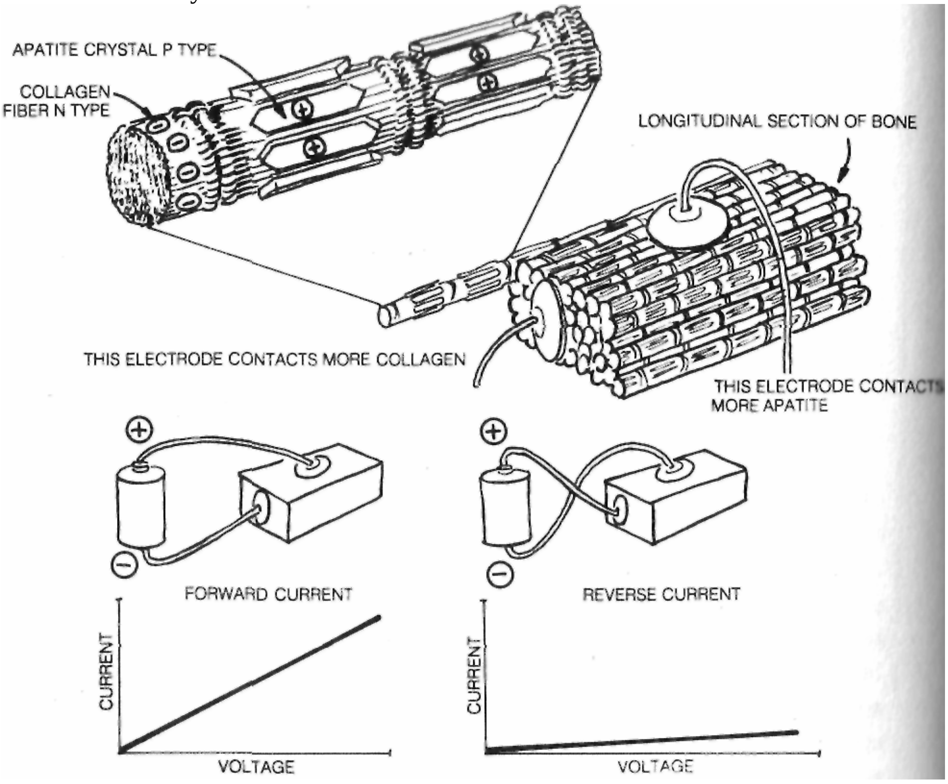
APATITE CRYSTAL P TYPE
COLLAGEN FIBER N TYPE
LONGITUDINAL SECTION OF BONE
THIS ELECTRODE CONTACTS MORE COLLAGEN
THIS ELECTRODE CONTACT
MORE APATITE
FORWARD CURRENT
REVERSE CURRENT
CURRENT
VOLTAGE
CURRENT
VOLTAGE

BONE IS A RECTIFIER—A PN JUNCTION DIODE

We could now follow the entire control system of Wolff's law. Mechanical stress on the bone produced a piezoelectric signal from the collagen. The signal was biphasic, switching polarity with each stress-and-release. The signal was rectified by the PN junction between apatite and collagen. This coherent signal did more than merely indicate that stress had occurred. Its strength told the cells how strong the stress was, and its polarity told them what direction it came from. Osteogenic cells where the potential was negative would be stimulated to grow more bone, while those in the positive area would close up shop and dismantle their matrix. If growth and resorption were considered as two aspects of one process, the electrical signal acted as an analog code to transfer information about stress to the cells and trigger the appropriate response.

Now we knew how stress was converted into an electrical signal. We had discovered a transducer, a device that converts other forces into electricity or vice versa. There was another transducer in the Wolff's-law system - the mechanism that transformed the electrical signal into appropriate cell responses. Our next experiment showed us something about how this one worked.

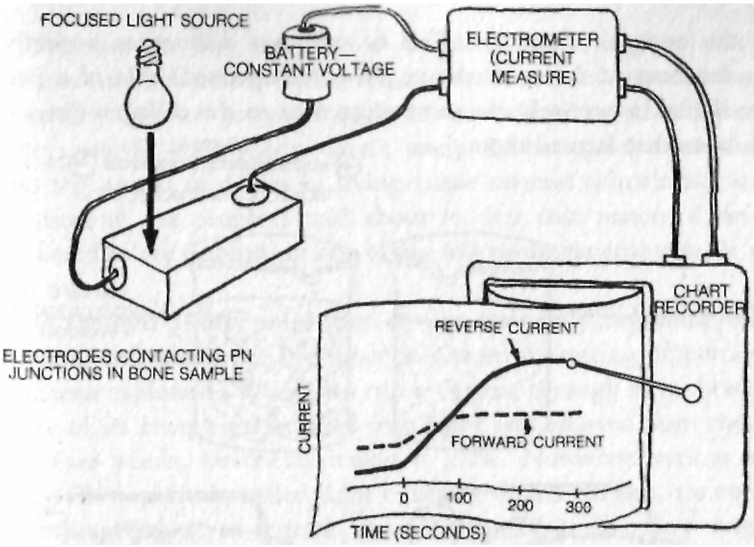
FOCUSED LIGHT SOURCE
BATTERY— CONSTANT VOLTAGE
ELECTROMETER (CURRENT MEASURE)
CHART RECORDER
REVERSE CURRENT
FORWARD CURRENT
CURRENT
0
100
200
300
TIME (SECONDS)
ELECTRODES CONTACTING PN JUNCTIONS IN BONE SAMPLE

THE LIGHT TEST CONFIRMS BONE'S ELECTRICAL SYSTEM

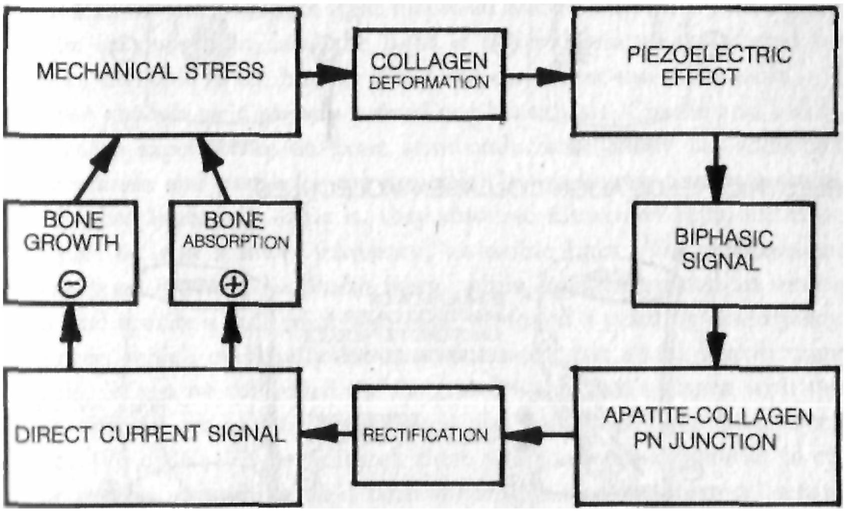
MECHANICAL STRESS
COLLAGEN DEFORMATION
PIEZOELECTRIC EFFECT
BIPHASIC SIGNAL
APATITE-COLLAGEN PN JUNCTION
RECTIFICATION
DIRECT CURRENT SIGNAL
BONE GROWTH ⊖
BONE ABSORPTION ⊕

WOLFF'S LAW CONTROL SYSTEM

Collagen fibers are formed from long sticks, like uncooked spaghetti, of a precursor molecule called tropocollagen. This compound, much used in biological research, is extracted from formed collagen—often from rat tails—and made into a solution. A slight change in the pH of the solution then precipitates collagen fibers. But the fibers thus formed are a jumbled, feltlike mass, nothing like the layered parallel strands of bone. However, when we passed a weak direct current through the solution, the fibers formed in rows perpendicular to the lines of force around the negative electrode. This fit our new discoveries perfectly, because the lines of force on the negative (compressed) side of a bent bone would be in precisely the same alignment as the collagen fibers of the new bone that formed there.

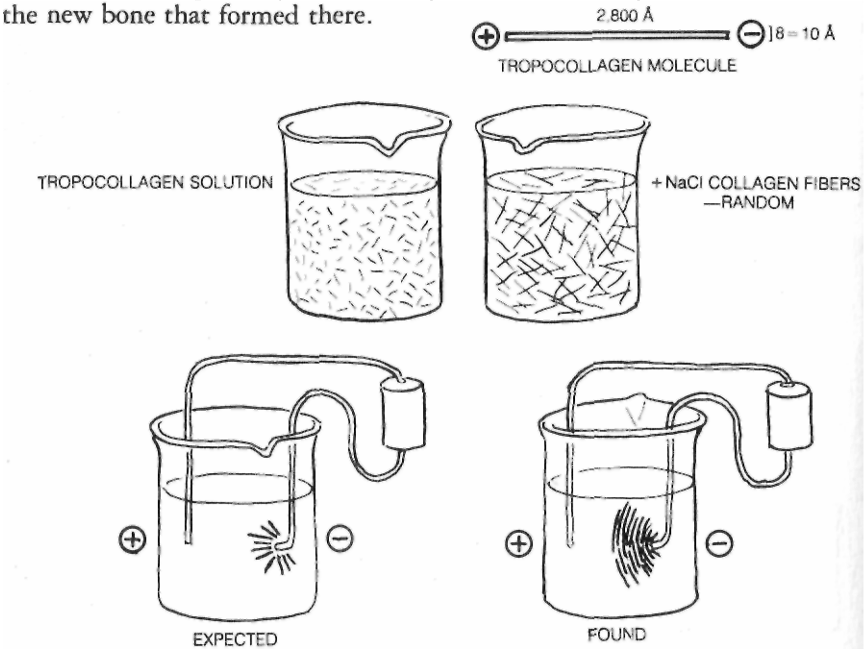
2,800 Å
⊕
⊖
8–10 Å
TROPOCOLLAGEN MOLECULE
TROPOCOLLAGEN SOLUTION
+NaCl COLLAGEN FIBERS
—RANDOM
⊕
⊖
⊕
⊖
EXPECTED
FOUND

ELECTRIC FIELDS ALIGN COLLAGEN MOLECULES

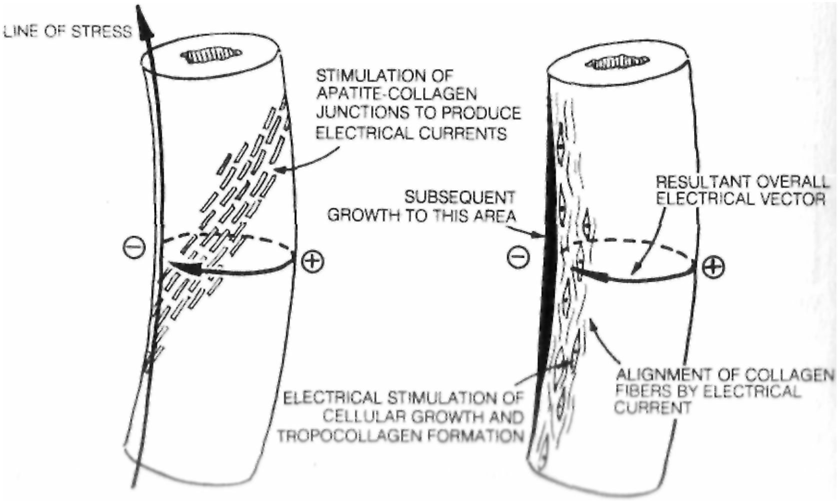
LINE OF STRESS
STIMULATION OF APATITE-COLLAGEN JUNCTIONS TO PRODUCE ELECTRICAL CURRENTS
SUBSEQUENT GROWTH TO THIS AREA
RESULTANT OVERALL ELECTRICAL VECTOR
⊖
⊕
⊖
⊕
ALIGNMENT OF COLLAGEN FIBERS BY ELECTRICAL CURRENT
ELECTRICAL STIMULATION OF CELLULAR GROWTH AND TROPOCOLLAGEN FORMATION

BONE'S ELECTRICAL SYSTEM GUIDES GROWTH

This was the first time that a complete circuit diagram of a growth process could be made. To us, this seemed an achievement of some note, but, although we got published, no one seemed to pay any attention. The scientific community wasn't ready for biological semiconduction, and the notion of diodes in living tissue seemed ridiculously farfetched to most of the people I told about it. For that reason, I never even bothered to try publishing one of our follow-up experiments. It was just too weird.

In the mid-1960s, solid-state devices were only beginning to hit the market, and one of the PN junction's most interesting properties hadn't yet been exploited. When you run a current through it in forward bias, some of its energy gets turned into light and emitted from the surface. In other words, electricity makes it glow. Nowadays various kinds of these PN junctions, called light-emitting diodes (LEDs), are everywhere as digital readouts in watches and calculators, but then they were laboratory curios.

We found that bone was an LED. Like many such materials, it required an outside source of light before an electric current would make it release its own light, and the light it emitted was at an infrared frequency invisible to us, but the effect was consistent and undeniable.

Even though we'd already proved our hypothesis, Charlie and I did a few more experiments on bone semiconduction, partly for additional confirmation and partly for the fun of it. It was known that some semiconductors fluoresced—that is, they absorbed ultraviolet light and emitted part of it at a lower frequency, as visible light. We checked, and whole bone fluoresced a bluish ivory, while collagen yielded an intense blue and apatite a dull brick-red. Here we found a puzzling discrepancy, however, which eventually led to a discovery that could benefit many people. When we combined the fluoresced light from collagen with that from apatite, we should have gotten the fluoresced light from whole bone. We didn't. That indicated there was some other material in the bone matrix, something we'd been washing out in the chemical separation.

Charlie and I were stumped on that line of research for a couple of years, until our attention was caught by a new development in solid-state technology called doping. Tiny amounts of certain minerals mixed into the semiconductor material could change its characteristics enormously. The making of semiconductors to order by selective doping would become a science in itself; to us it suggested trace elements in bone. We already knew that certain trace metals - such as copper, lead, silver, and beryllium - bonded readily to bone. Beryllium miners had a high rate of osteogenic sarcoma—bone cancer—because beryllium somehow removed normal controls from the osteocytes' growth potential. Radioactive strontium 90 worked its harm by bonding to bone, then bombarding the cells with ionizing radiation. Perhaps some trace element normally found in bone changed its electrical properties by doping it.

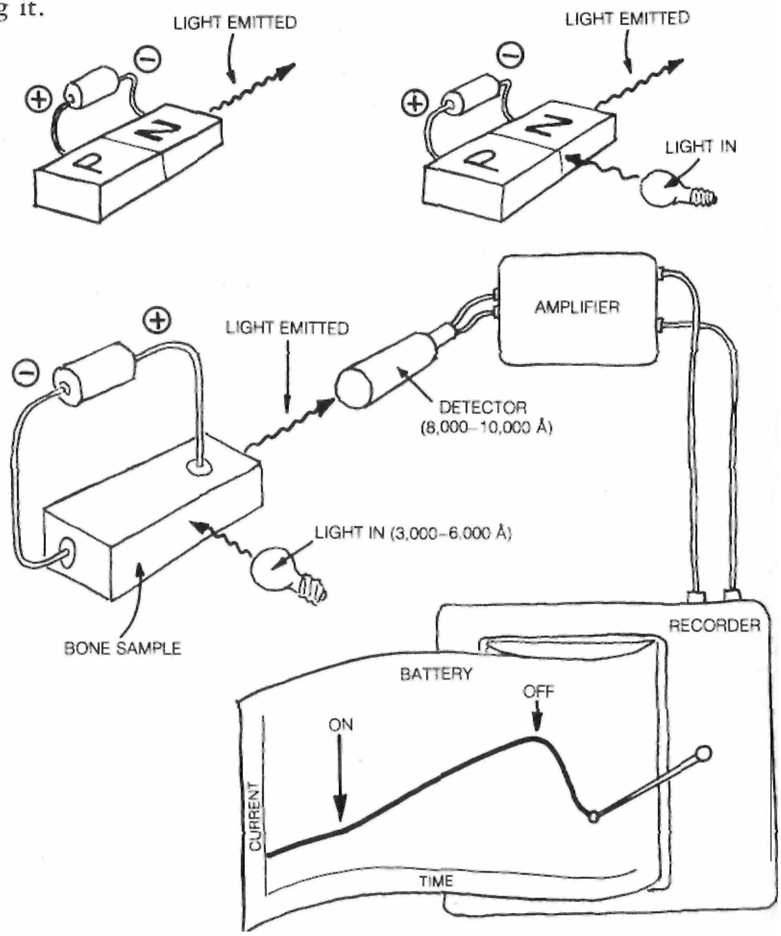
LIGHT EMITTED
P
N
LIGHT EMITTED
LIGHT IN
P
N
AMPLIFIER
LIGHT EMITTED
DETECTOR
(8,000–10,000 Å)
LIGHT IN (3,000–6,000 Å)
BONE SAMPLE
RECORDER
BATTERY
ON
OFF
CURRENT
TIME

PHOTOELECTRICITY: ELECTRICITY GOES IN; LIGHT COMES OUT

To find out, Charlie and I used a very complex device called an electron paramagnetic resonance (EPR) spectrometer on our bone samples. There's no easy way to explain just how this instrument works, but basically it measures the number of free electrons in a material by sensing a resonance produced in the electrons' vibrations by in applied magnetic field. We used it to measure the fre electrons in collagen and apatite, and we found the same kind of discrepancy as in our flourescence experiment: When we added together the free electrons of collagen and apatite, we fell short of the number we found in whole bone. That made us certain that we were washing out some trace mineral.

We decided to work backward. We prepared a solution containing small amounts of a wide variety of metals. Then we soaked our collagen and apatite cubes in this broth to see what they'd pick up.

We knew we were on the way to solving this mystery when we examined the results. Only a few of the metals had bonded to the bone materials: beryllium, copper, iron, zinc, lead, and silver. The diameters of all the absorbed atoms were exact fractions of one another. The results showed that the bonding sites were little recesses into which would fit one atom of silver or lead, two of iron or copper or zinc, or six of beryllium.

Only one of these metals gave us an electron resonance of its own, indicating that it had a large number of free electrons that could affect the electrical nature of bone. That metal was copper. We made a batch of broth containing only copper. We expected that copper's EPR signal would change to one value as it bonded with collagen, and to another as it bonded with apatite. Since the molecular structure of each was quite different, we figured that each would bind copper in a different way.

We could hardly believe the results. Bonding had indeed changed copper's resonance, but the change was the same in both materials. By analyzing it we deduced that each atom of copper fit into a little pit, surrounded by a particular pattern of electric charges, on the surface of apatite crystals and collagen fibers. Because the pattern of charges was the same in both materials, we knew that the bonding sites were the same on both surfaces and that they lined up to form one elongated cavity connecting the crystal and fiber. In other words, the two bonding sites matched, forming an enclosed space into which two atoms of copper nestled. The electrical forces of this copper bond held the crystals and fibers together much as wooden pegs fastened the pieces of antique furniture to each other. Furthermore, the electrical nature of peg and hole suggested that we had found, on the atomic level, the exact location of the PN junction.

This discovery may have some medical importance. The question of how the innermost apatite crystals fasten onto collagen had eluded orthopedists until then, and the finding may have opened a way to understand osteoporosis,a condition in which the apatite crystals fall off and the bone degenerates. The process is often called decalcification, although more than calcium is lost. It's a common feature of aging. I surmise that osteoporosis comes about when copper is somehow removed from the bones. This might occur not only through chemical/metabolic processes, but by a change in the electromagnetic binding forces, allowing the pegs to "fall out." It's possible that this could result from a change in the overall electrical fields throughout the body or from a change in those surrounding the body in the environment.

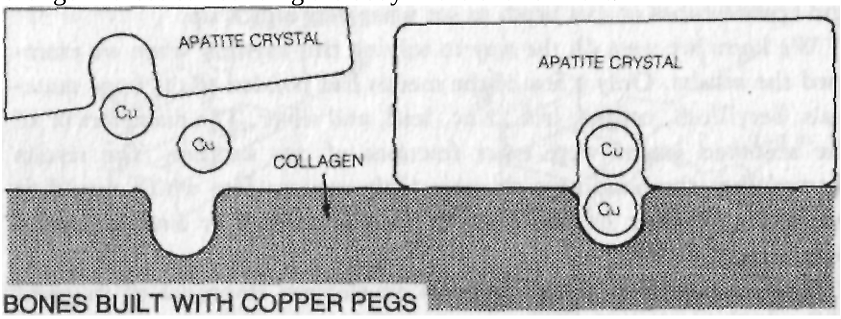

BONES BUILT WITH COPPER PEGS

Osteoporosis has been a major worry of the American and Russian space programs. As flights got longer, doctors found that more and more apatite was lost from the bones, until decalcification reached 8 percent in the early Soviet Salyut space-station tests. Serious problems were known to occur only when apatite loss reached about 20 percent, but the trend was alarming, especially since depletion of the calcium reservoir might affect the nervous system and muscular efficiency before the 20-percent level was reached. Although the immediate cause was their inability to get a malfunctioning air valve closed before all the cabin air escaped, weakness from muscle tone loss might have contributed to the death of the three cosmonauts who succumbed while returning from their twenty-four-day flight aboard Soyuz 11 at the end of June 1971.

Space osteoporosis may result from unnatural currents induced in bone by a spacecraft's rapid motion through the earth's magnetic field, with a polarity reversal every half orbit, or it may be a direct effect of the field reversal. This abnormality, which may change the activity of bone cells directly, would be superimposed on abnormal responses of bone's natural electrical system, which is almost certainly affected by weightlessness. The unfamiliar external field reversals could also weaken the copper pegs, at the same time that the bones are in a constant state of "rebound" from their earthly weight-induced potentials, producing a signal that says, "No weight, no bones needed." We know that the more even distribution of blood caused by weightlessness registers in the heart as an excess; as aresult, fluid and ions, including calcium, are withdrawn from tht blood. However, the effect probably isn't due to weightlessness alone, for the Skylab astronauts did rigorous exercise, which would have supplied plentiful stresses to their bones. They worked out so hard that their muscles grew, but decalcification still reached 6.8 percent on the twelve-week mission.

The Soviets at first claimed to have solved the problem before or during the Soyuz 26 mission of 1977—78, in which two cosmonauts orbited in the Salyut 6 space lab for over three months. Subsequent Soviet spacepersons, who have remained weightless as long as 211 days, reportedly have showed no ill effects from osteoporosis, and chief Soviet space doctor Oleg Gazenko said it simply leveled off after three months. However, this claim was later officially withdrawn, and I have a hunch that the Soviets are working on a way to prevent the condition by simulating earth-surface fields inside their space stations, a method that perhaps hasn't yet worked as well as they'd hoped. Andy Bassett has suggested giving our astronauts strap-on electromagnetic coils designed to approximate their limb bones' normal gravity stress signals, but so far NASA has shown no interest.

Unfortunately for earthbound victims of osteoporosis, the copper peg discovery still hasn't been followed up, even though I published it over fifteen years ago. Charlie and I wanted to continue in that direction, but we knew that we couldn't sustain more than one major research effort at a time. We decided that regenerative growth control was our primary target, so we reluctantly dropped osteoporosis. Fortified by our new knowledge that electricity controlled growth in bone, we returned instead to the nerves, taking a closer look at how their currents stimulated regrowth.

A Surprise in the Blood

1 felt as though the temple curtain had been drawn aside without warning and I, a goggle-eyed stranger somehow mistaken for an initiate, had been ushered into the sanctuary to witness the mystery of mysteries. I saw a phantasmagoria, a living tapestry of forms jeweled in minute detail. They danced together like guests at a rowdy wedding. They changed their shapes. Within themselves they juggled geometrical shards like the fragments in a kaleidoscope. They sent forth extensions of themselves like the flares of suns. Yet all their activity was obviously interrelated; each being's actions were in step with its neighbors'. They were like bees swarming: They obviously recognised each other and were communicating avidly, but it was impossible to know what they were saying. They enacted a pageant whose beauty awed me.

As the lights came back on, the auditorium seemed dull and unreal. I'd been watching various kinds of ordinary cells going about their daily business, as seen through a microscope and recorded by the latest time-lapse movie techniques. The filmmaker frankly admitted that neither he nor anyone else knew just what the cells were doing, or how and why they were doing it. We biologists, especially during our formative years in school, spent most of our time dissecting dead animals and studying preparations of dead cells stained to make their structures more easily visible—"painted tombstones," as someone once called them. Of course, we all knew that life was more a process than a structure, but we tended to forget this, because a structure was so much easier to study. This film reminded me how far our static concepts still were from the actual business of living. As I thought how any one of those scintillating cells potentially could become a whole speckled frog or a person, I grew surer than ever that my work so far had disclosed only a few aspects of a process-control system as varied and widespread as life itself, of which we'd been ignorant until then.

The film was shown at a workshop on fracture healing sponsored by the National Academy of Sciences in 1965. It was one of a series of meetings organized for the heads of clinical departments to educate them as to the most promising directions for research. A dynamic organizer named Jim Wray had recently become chairman of the Upstate Medical Center's department of orthopedic surgery, but Jim's superb skills were political rather than scientific. Since I was an active researcher and had just been promoted to associate professor, Jim asked me to go in his place. I tried to get out of it, because I knew my electrical bones would get a frosty reception from the big shots if I opened my mouth, but Jim prevailed. The meeting was mostly what I'd expected, but there were three bright spots. One was that microcinematic vision. Another was the chance to get acquainted with the other delegate from my department, a sharp young orthopedic surgeon named Dave Murray. The third was the presence of Dr. John J. Pritchard.

A renowned British anatomist who'd added much to our knowledge of fracture healing, Dr. Pritchard was the meeting's keynote speaker—the beneficent father-figure who was to evaluate all the papers and summarize everything at the end. Dave and I almost skipped his talk. We hadn't been impressed with the presentations, and we figured there had been so few new ideas that Pritchard would have nothing to say. However, our bus to the Washington airport didn't leave until after Pritchard's luncheon address, so we stayed. With a tact that seemed peculiarly English, he reached the same assessment we had, but phrased it so as to offend no one. He stressed that fracture healing should be considered a vestige of regeneration. Most past work on fractures had described what happened when a bone knit, as opposed to the how and why. As Pritchard pointed out, "Not a great deal of thought has been given to the factors that initiate, guide, and control the various processes of bone repair." Just as in regeneration research, this was the most important problem about fractures, he concluded.

Dave and I had to wait several hours before our flight back to Syracuse. We sat in the airport lounge and talked excitedly about broken bones. Dave agreed that, since I'd found electrical currents in salamander limb regeneration, it was at least plausible that similar factors controlled the mending of fractures. Having just deciphered the control system for stress adaptation (Wolffs-law growth) in bone, I felt prepared to get back to the more complex problems of regeneration via its remnant in bone healing. Dave and I decided to collaborate, and we planned our experiment on the plane. We would break the same bone in a standardized way in each of a series of experimental animals. I would study the electrical forces in and around the fractures as they healed. We would kill a few of the animals at each stage of healing, and Dave, an expert histologist (cell specialist), would make microscope slides of the healing tissues and study the cellular changes. Along the way we would fit our findings together to see whether electricity was guiding the cells.

Our first task was to choose the experimental animals. We wanted to use dogs or rabbits, since ultimately we were trying to understand human bones and wanted to work with animals as closely related to us as possible. But we would need scores of them to study each phase of healing adequately, and we had neither the funds nor the facilities to house so many large mammals. We thought of rats, but their longest bones were too short to study clearly and were curved as well. We were looking for nice, long, straight bones, in which we could produce uniform breaks.

We settled on bullfrogs. They were cheap to buy and care for; we could even collect some ourselves from nearby ponds. I already had a lot of experience in working with them. Best of all, the adult frog's lower leg had one long bone—the tibia and fibula found in most vertebrates had merged into a tibiofibularis. It was about two inches long with a fine, straight shaft.

Our misgivings about the evolutionary distance between frogs and humans were allayed when we went to the library to read up on what was then known about fracture healing in frogs. Dr. Pritchard himself, along with two of his students, J. Bowden and A. J. Ruzicka, had determined that frogs mended their bones the same way people did. Our question was: What stimulated the periosteal and marrow cells to change into new bone-forming cells?

We began by anesthetizing the animals and resolutely breaking all those little green legs by hand, bending them only to a certain angle so as not to rupture the periosteum around the fracture. I found I had to put little plaster casts on them—not because the frogs seemed in great pain but because their movements kept shifting the broken bones and making systematic observations impossible. They would have healed anyway; in our first sixty frogs we found two that had broken their legs in the wild and mended them, but I'm sure ours were the first that ever had casts.

The electrical changes were complex but were almost the same in every fracture. There were two distinct patterns, one on the periosteum and one on the bone. Before fracture the ankle end of both the bone and the periosteum had a small negative potential of less than 1 millivolt as compared to the knee end. At the moment of fracture, the negative potential on the intact periosteum over the break shot up to 6 or 7 millivolts, while areas of positive charge formed above and below the break. After a week, the periosteum's normal progression of negative charge toward the ankle was restored. When a fracture ruptured the periosteum, its negative potential went even higher than 7 millivolts, but amputation of the dangling lower leg immediately reversed the polarity, producing a positive current of injury from the stump, as in the frogs of my first regeneration experiment. The bone itself underwent a short-term electrical change opposite to that in the periosteum. A small positive charge appeared on each of the broken ends during the first hours, then fell to near zero after three hours.

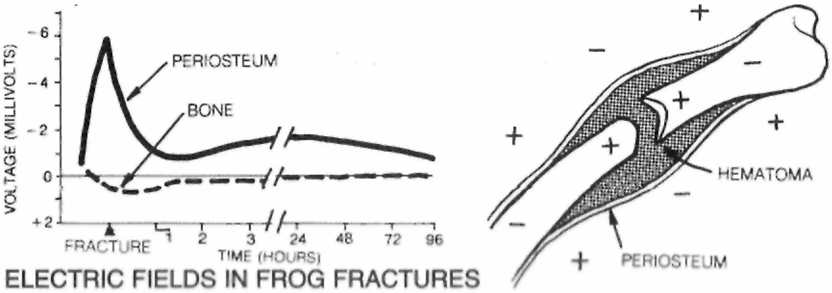

ELECTRIC FIELDS IN FROG FRACTURES

The electricity had two different sources. When cut the leg nerves, the periosteum readings dropped dramatically, indicating that these potentials were coming from currents in the nerves to the periosteum and the surrounding wound area. Measurements on the bone, which has almost no nerves, were unaffected. Many piezoelectric materials emit a continuous current for several hours after their charge-producing structure has been left in a state of unresolved stress by fracture; I surmised that this was true of bone and soon found that another research team had recently proven it. Two separate currents, then, one from the nerves and one from the bone matrix, were producing potentials of opposite polarity, which acted like the electrodes of a battery. These living electrodes were creating a complex field whose exact shape and strength reflected the position of the bone pieces. The limb was, in effect, taking its own X ray.

While I was busy with my probes and meters, Dave was taking samples of the bones and blood clots and preparing them for the microscope. We killed a few of the frogs every fifteen minutes during the first two hours, then every day for two weeks, every other day for the third week, and weekly for the last three weeks. Preparing the slides took several days.

In the normal sequence of bone healing in frogs, a blood clot forms after about two hours and develops into a blastema during the first week. It turns into the rubbery, fibrous callus during the second and third weeks, and ossifies in three to six weeks. In this last period, islands of bone first emerge near the broken ends. Next, bony bridges appear, connecting the islands. Then the whole area is gradually filled in and organized with the proper marrow space and blood canals to join the segments of old bone.

Dave began his work with specimens taken nearly a week after the fractures, when we expected to see the first signs of the callus forming. "This is mighty damn funny," he said as he walked in with the first box of slides. "I can't see any mitoses in the periosteum. There's no evidence that the cells there are multiplying or migrating."

We agreed that we must have done something wrong. Pritchard's work had been quite conclusive on this point. He'd even published photographs of the periosteal cells dividing and moving into the gap. We thought maybe we were looking at specimens from the wrong time period, but we could see with our own eyes that the callus was starting to form. Dave went back to study specimens from the first few days, even though we didn't expect to see much then except clotting blood. Soon he called me from his lab and asked, "What would you say if I told you that the red corpuscles change and become the new bone-forming cells?"

I groaned. "Nonsense. That can't be right." But it was right. We went over the whole slide series together. Beginning in the second hour, the erythrocytes (red cells) began to change.

All vertebrates except mammals have nuclei in their red blood cells. In mammals, these cells go through an extra stage of development in which the nucleus is discarded. The resulting cells are smaller, can flow through smaller capillaries, can be packed with more hemoglobin, and thus can carry oxygen and carbon dioxide more efficiently. Nucleated erythrocytes are considered more primitive, but even in these the nucleus is pycnotic—shriveled up and inactive. The DNA in pycnotic nuclei is dormant, and such cells have almost no metabolic activity; that is, they burn no glucose for energy and synthesize no proteins. If you had to choose a likely candidate for dedifferentiation and increased activity, this would be the worst possible choice.

In our series of slides the red cells went through all their developmental stages in reverse. First they lost their characteristic flattened, elliptical shape and became round. Their membranes acquired a scalloped outline. By the third day the cells had become ameboid and moved by means of pseudopods. Concurrently, their nuclei swelled up and, judging by changes in their reactions to staining and light, the DNA became reactivated. We began using an electron microscope to get a clearer view of these changes. At the end of the first week, the former erythrocytes had acquired a full complement of mitochondria and also ribosomes (the organelles where proteins are assembled), and they'd gotten rid of all of their hemoglobin. By the third week they'd turned into cartilage-forming cells, which soon developed further into bone-forming cells.

I wasn't happy with this turn of events. How could we reconcile what we saw with the well-documented findings of Pritchard, Bowden, and Ruzicka? I'd expected evidence for the semiconducting electrical system I'd been investigating, a concept that was already strange enough to keep me out of the scientific mainstream. I would have been happy if the electrical measurements had fit in with straightforward changes in the periosteal cells. The difference between them and erythrocytes was crucial. Periosteal cells were closely related precursors of bone cells; blood cells couldn't have been further removed. They couldn't possibly have built bone without extensive job retraining on the genetic level. These bullfrogs were bringing us up hard against a wall of dogma by showing us metaplasia—dedifferentiation followed by redifferentiation into a totally unrelated cell type. The process took place in some of the most specialized cells in the frog's entire body, and it looked as though the electric field set the changes in motion while at its strongest, about an hour after the fracture.

Our next move was a respectful letter to Dr. Pritchard asking if there was any way he could make sense of the contradictory observations. He replied in the negative but sent our inquiry to Dr. Bowden, who had done the actual work on frogs as his doctoral thesis. Bowden had a possible explanation. He'd done the experimental work under a time limit, and to finish before the deadline he'd kept his frogs at high temperatures—only a few degrees short of killing them, in fact—in order to speed up their metabolism.

Bowden also mentioned that two researchers cited in his bibliography had seen fracture healing in frogs much the same way we had. In the 1920s, a German named H. Wurmbach, also working on his doctorate, noted some strange cellular transformations in the blood clot and worried over his inability to explain them. However, Wurmbach also found mitoses in the periosteum and ascribed healing to the latter process, since it didn't involve dedifferentiation. A decade later, another German scientist, A. Ide-Rozas, saw the same changes in the blood cells, but he was more daring. He proposed that this transformation was the major force behind fracture healing in frogs and further suggested that regenerating salamanders formed their limb blastemas from nucleated red blood cells. Other experiments seemed to contradict Ide-Rozas' idea about limb blastemas, so his work was discredited and ignored, but Bowden wished us better luck.

Bowden's letter gave us a framework for understanding our results. We already knew that mammals did not heal bones by dedifferentiation of their red corpuscles, because their red cells had no nuclei and thus no mechanism for change. Mammals also had a thicker periosteum than other vertebrates, so we reasoned that periosteal cell division played a larger healing role in mammals. Frogs, it seemed, had both methods available but activated the periosteal cells only at high temperatures.

Do-lt-Yourself Dedifferentiation

Now we were sure that our results were real. We repeated the same fracture studies, but this time we also observed the cells while they were alive. We took tissue samples from the fractures and made time-lapse film sequences using techniques like those in the movie that had impressed me so much at the NAS workshop. We confirmed that the changes began in the first lew hours, just after the electrical forces reached their peak.

Now we decided to try a crucial test. If the electricity really triggered healing, we should be able to reproduce the same field artificially and start the same changes in normal blood cells outside the frog. If this didn't work, then I probably had spent the last seven years "collecting stamps"—accumulating facts that were interesting but, in the end, trivial.

I calculated the amount of current that would produce the fields I'd found. I came up with an incredibly small amount, somewhere between a trillionth and a billionth of an ampere (a picoamp and a nanoamp, respectively). Again I thought there must be some mistake. I didn't see how such a tiny current could produce the dramatic effects we'd seen, so, figuring that even if my numbers were right, more juice would simply hasten the process, I decided to start with 50 microamps, a current level that would be just shy of producing a little electrolysis—the breakdown of water into hydrogen and oxygen.

I designed plastic and glass chambers of various shapes, fitted with electrodes of several types. In these chambers we would place healthy red blood cells in saline solution and observe them by microscope while the current was on.

I set up the experiment in a lab across the street from the medical center, where there was available one of the inverted microscopes we would need to observe the cells through the bottoms of the chambers, where most of them would settle. I put a young technician named Frederick Brown in charge of the long grind of watching the cells hour after hour at different current levels and field shapes in the various chambers. We began in the summer of 1966. Fred was to enter medical school that fall, and I figured two months would be more than enough time! He was to run one test batch of frog blood each day and report to me the next morning as to what he'd found.

It didn't start well. Nothing had happened after six hours of current. We couldn't increase the amperage without electrocuting the cells, so we ran it longer. Still nothing happened. In fact, the cells started dying when we left them in the chambers overnight. We decided to lower the current, but I still didn't believe in the absurdly low values I'd calculated, so I told Fred to drop the amperage only a little bit day by day. He and I stared at a lot of blood cells over those two months, all stubbornly refusing to do anything. Finally, two days before Fred had to leave, we'd gotten the current down as far as our first apparatus could go, and well within the range I'd calculated—about half a billionth of an ampere. At eleven that morning he called me excitedly and I rushed across the street.

With the room darkened and the microscope light on, we saw the same cell changes as in the blood clot, first at the negative electrode, then at the positive electrode, and finally spreading across the rest of the chamber. In four hours all the blood cells in the chamber had reactivated their nuclei, lost their hemoglobin, and become completely unspecialized in form.

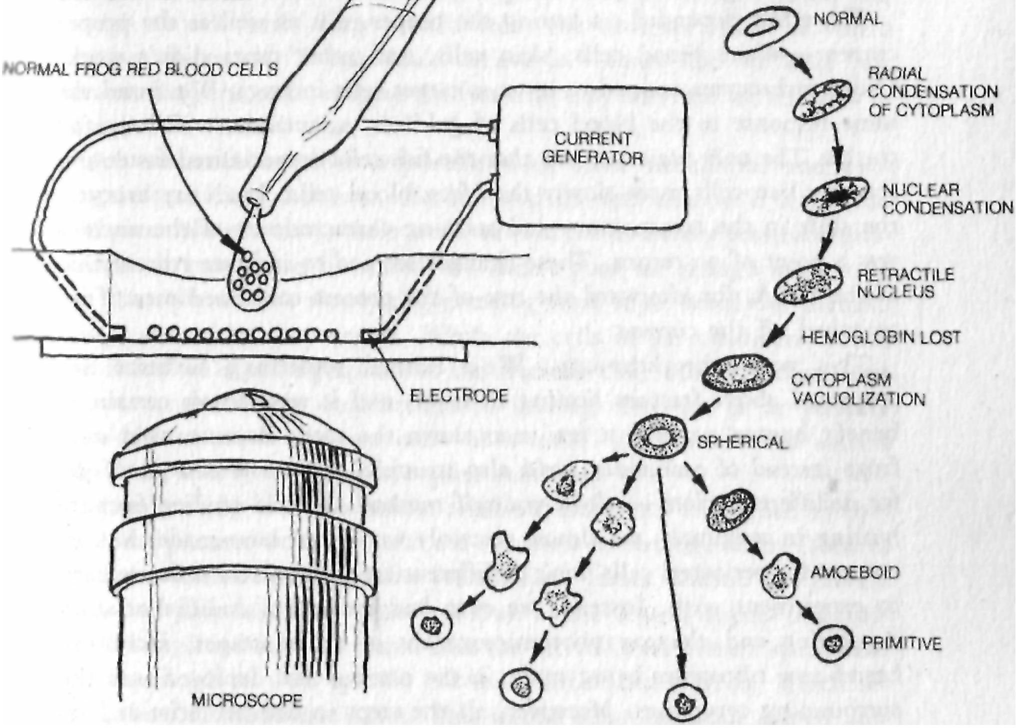
NORMAL FROG RED BLOOD CELLS
CURRENT GENERATOR
ELECTRODE
MICROSCOPE
NORMAL
RADIAL CONDENSATION OF CYTOPLASM
NUCLEAR CONDENSATION
RETRACTILE NUCLEUS
HEMOGLOBIN LOST
CYTOPLASM VACUOLIZATION
SPHERICAL
AMOEBOID
PRIMITIVE

THE FIRST ARTIFICIAL DEDIFFERENTIATE

We repeated the experiment many times, working out the upper and lower limits of the effective current. The best "window" was somewhere between 200 and 700 picoamps. I say "somewhere" because the susceptibility of the cells varied, depending on their age, the hormonal state of the frog, and possibly other factors.* This was an infinitesimal tickle of electricity, far less than anything a human could feel even on the most sensitive tissue, such as the tongue, but it was enough to goose the cell into unlocking all its genes for potential use.

---

* Rather than continually renewing a small part of their red-cell stock, frogs generate a whole year's supply in late winter as they emerge from hibernation. Thus, all their erythrocytes age uniformly as the year progresses. The cells become less sensitive to electricity as they get older, and that may be why frogs, even when warm and not hibernating, heal fractures more slowly in winter. The red blood cells dedifferentiate most readily in the spring, and the female's become even more sensitive than the male's when she ovulates early in that season. At that time her red corpuscles will despecialize in response to less than one picoamp. In fact, we saw red cells from ovulating females dedifferentiate completely in chambers to which we supplied no current whatever. Apparently an unmeasurably small current created by the charge difference between the plastic chamber and glass cover slip was enough.

The effect depended on having the proper cells as well as the proper current—white blood cells, skin cells, and other types didn't work. Only erythrocytes seemed to serve as target cells in frogs. We found the same response in the blood cells of goldfish, salamanders, snakes, and turtles. The only variation was that the fish cells despecialized faster and the reptilian cells more slowly than frog blood cells. In all erythrocytes the shift in the transparency and staining characteristics of the nucleus was a point of no return. These changes seemed to indicate reactivation of the DNA, for afterward the rest of the process continued even if we switched off the current.

This was a breakthrough. We'd learned something hitherto unsuspected about fracture healing in frogs, and it was almost certain to benefit human patients a few years down the road. Because we'd used frogs instead of mammals, we'd also stumbled upon the best proof yet for dedifferentiation—a do-it-yourself method. If we'd studied fracture healing in mammals, we almost certainly would not have made the discovery, for periosteal cells don't dedifferentiate and marrow cells are hard to experiment with. Instead, we even had movies of dedifferentiation happening and electron photomicrographs of air its stages, including brand-new ribosomes being made in the nucleus and deployed into the surrounding cytoplasm. Moreover, all the steps in dedifferentiation, including the activities in the nucleus and the assembly of ribosomes and mitochondria, exactly paralleled the changes found by the most recent research on salamander limb blastemas. We'd found the electrical common denominator that started the first phase—the blastema—in all regeneration.

The Genetic Key

Soon after we'd finished this experiment, I was invited to a meeting on electromagnetism in biology at the New York Academy of Sciences. This was basically a one-man show. Kenneth MacLean, a prominent surgeon and highly placed member of the academy, had been using magnetic fields on his patients for years and was convinced that they helped. Independently wealthy, he'd set up a lab in his office, with a large electromagnet. The meeting was a testament to his persistence rather than any widespread belief within the academy that he was right. So in February 1967 I presented our recent work. I played down the full role of electricity in fracture healing, with its overtones of vitalism, and concentrated on our method for inducing dedifferentiation in vitro. That was enough to call forth many attacks from the audience, most of which were variations on "I just don't believe it." Some said we were just electrocuting the cells, despite the fact that they survived for ten days in culture.

One audience member responded with some thoughtful and constructive criticism, however. He accepted the fact that we'd seen what we'd seen. Nevertheless, there were, he said, other barely possible explanations. In particular, he stated, we hadn't gone far enough in proving that the cells weren't slowly degenerating from some small but harmful change caused by the current. While the cells in our chambers looked like those we'd photographed in the fracture clot, our idea that these cells were electrically dedifferentiated healing cells was so at variance with current views that we must have more direct proof. For such a radical departure, seeing wasn't quite believing.

Stimulated by this one honest reaction, Dave and I returned to Syracuse and planned how we could use the latest knowledge about DNA to test our evidence further. A few years before, James Watson and Francis Crick had proposed what became known as the central dogma of genetics. In simplified form, it stated that the active DNA in each specialized cell imprinted its own specific patterns onto transfer RNA, which relayed them to messenger RNA. This second RNA molecule moved outside the nucleus to the ribosomes, where it translated the genetic instructions into the particular proteins that made the cell what it was.

We reasoned that, since a dedifferentiating cell was not going to divide immediately, it wouldn't duplicate its genes. Therefore there should be no increase in the amount of DNA it contained. However, since the cell changed its type by manufacturing a whole new set of proteins, the amount of RNA—the protein blueprints—should increase dramatically.

Using radioactive labeling and fluorescent staining techniques, we found there was indeed no new DNA but dramatic increases in RNA. Another test showed that our despecialized cells contained not only different proteins but also twice as many as their precursor red blood cells.

The most conclusive experiment was one suggested by Dan Harrington, a student who'd taken Fred Brown's place and who later went on to a Ph.D. in anatomy. Dan proposed that we use certain well-known metabolic inhibitors that disrupt the DNA-RNA-protein system, to see if we could prevent dedifferentiation. One such inhibitor, an antibiotic called puromycin, blocks the transfer of information from messenger RNA to the ribosomes and thus prevents proteins from being built. Dan proposed that we set up our plastic chambers in pairs. In one member of each pair we would place blood cells in saline solution and pass current through them as before. The other chamber would contain cells from the same frog and be connected to the same generator, but the solution would contain puromycin. The setup would thus be precisely controlled. If the current was actually unlocking new genes, the puromycin should stop dedifferentiation by intercepting the DNA's protein-making instructions. If the current was merely making the cells degenerate, however, the puromycin should have no effect and the transformations should continue.

Our conclusion held up: The cells in the puromycin solution didn't change. Next Dan suggested that we bathe these cells with several changes of water to wash out the puromycin. They promptly dedifferentiated with no current flowing! Apparently the current caused the genetic change in spite of the antibiotic, and the genes stayed unlocked, so the system still worked perfectly as soon as the puromycin blockade was lifted.

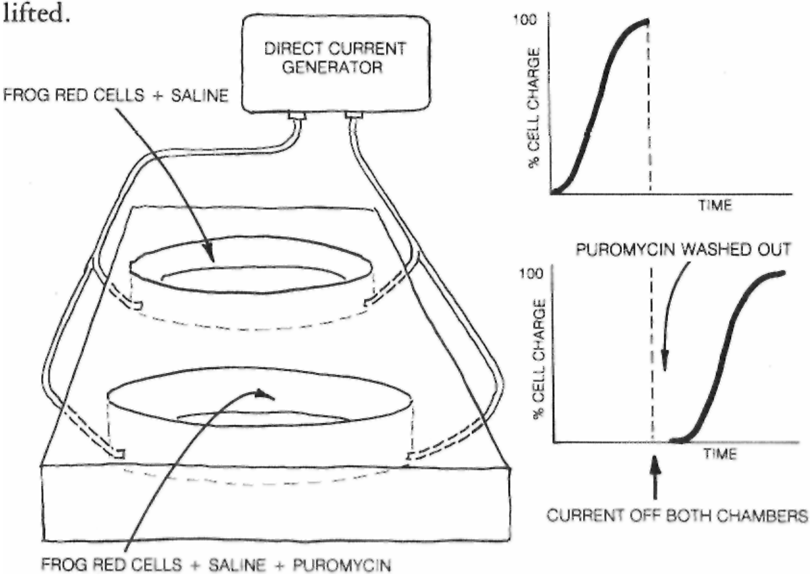

THE RNA BLOCKADE CONFIRMS ELECTRICAL DEDIFFERENTIATION

By early 1970 we had solid proof for nearly every detail of the control system for fracture healing in frogs, and by extension probably in mammals as well. Like all other injuries, a fracture produced a current of injury, in this case derived from the nerves in and around the periosteum.* At the same time, the bone generated its own current piezoelectrically due to residual stress in the mangled apatite-collagen matrix. These signals combined to stimulate the cells that formed new bone.

Except for the identity of the target cells, bone repair seemed to be basically the same in all vertebrates, proceeding through the stages of blood clot, blastema, callus, and ossification. In fish, amphibians, reptiles, and birds, the red cells in the clot dedifferentiated in response to the electric field, especially the positive potentials at the broken ends of the bone. They then redifferentiated as cartilage cells and continued on to become bone cells.

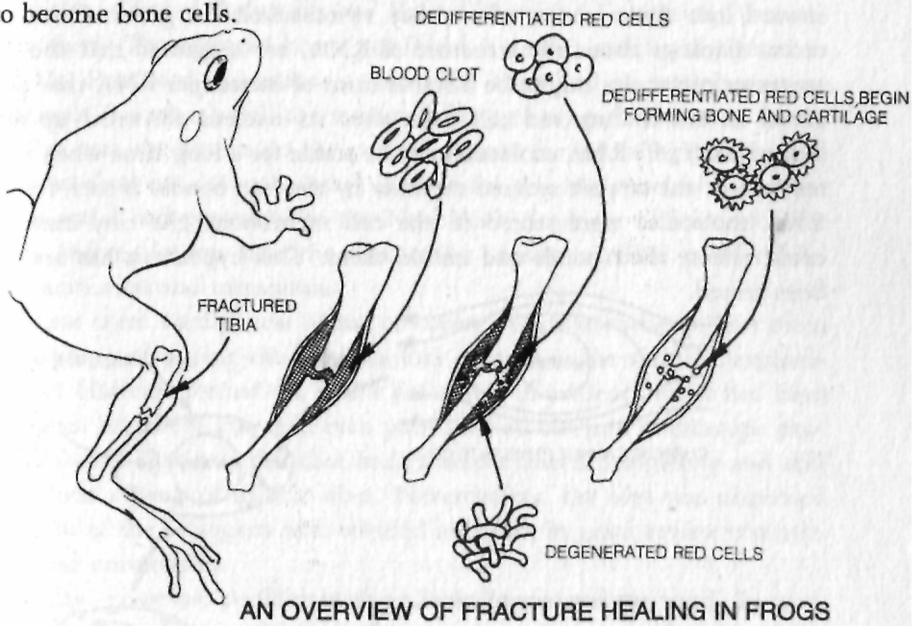
DEDIFFERENTIATED RED CELLS
BLOOD CLOT
DEDIFFERENTIATED RED CELLS BEGIN FORMING BONE AND CARTILAGE
FRACTURED TIBIA
DEGENERATED RED CELLS

AN OVERVIEW OF FRACTURE HEALING IN FROGS

In some animals the periosteal cells responded to the current of injury by migrating into the gap and specializing a bit further into bone cells. This process seemed to be available to amphibians only at high temperatures, but it was the dominant method in mammals, whose thick periosteum made up for the lack of nuclei in their red corpuscles. By this time we'd become pretty sure that the marrow component of bone healing in humans involved the dedifferentiation of at least the immature erythrocytes, which still contained a nucleus, and possibly other cell types.

---

*The lack of the periosteal (nerve-derived) current may explain the uncontrolled, deformed growth that often follows fractures in the limbs of paraplegics and lepers. Their bones still generate a positive potential in the gap, but because of nerve damage it isn't balanced by the negative periosteal potential that normally surrounds the break.

The electrical forces turned the key that unlocked the repressed genes. The exact nature of that key was the one part still missing from the process. The current couldn't act directly on the nucleus, which was insulated by the cell's membrane and cytoplasm. We knew that the current's primary effect had to be on the membrane. The cell membrane itself was known to be charged. Its charge probably occurred as a specific pattern of charged molecules, different for each type of cell. We postulated that the membrane released derepressors—molecules that migrated inward into the nucleus, where they switched on the genes. Based on recent findings about the structure of RNA, we suggested that the derepressor molecules might be a stable form of messenger RNA that persisted in the mature red cell even after its nucleus shriveled up and turned itself off. RNA molecules can be stable for a long time when they are folded, the strands secured together by electron bonds. If such folded RNA molecules were stored in the cell membrane, the tiny currents could release their bonds and unfold them. This hypothesis has not yet been tested.

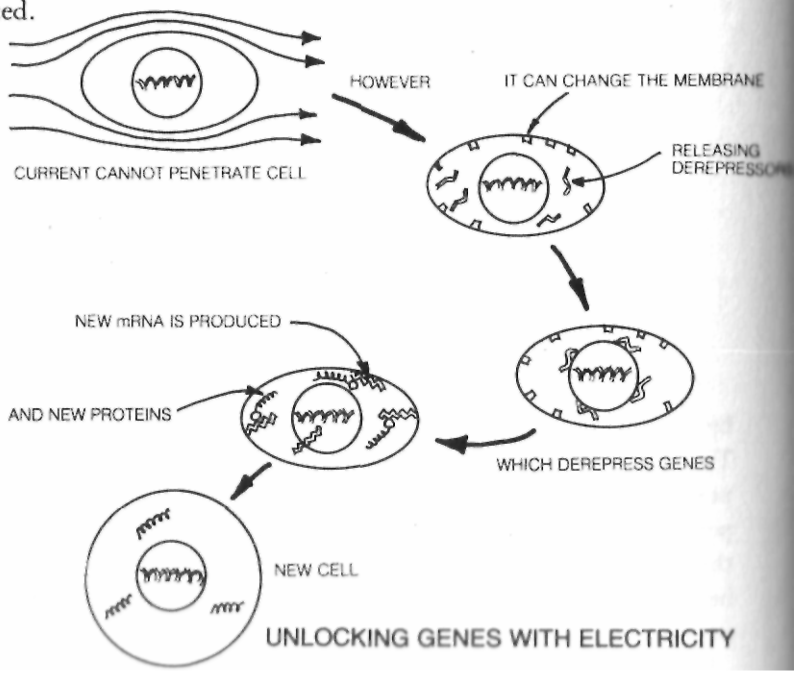

UNLOCKING GENES WITH ELECTRICITY

Fracture healing was ended by a straightforward negative-feedback system. As the gap was filled in with new matrix, the bone gradually redistributed its material to balance the stresses on it from the action of the surrounding muscles during cautious use, in accordance with Wolff's law. Repair of surrounding tissues lessened and then stopped the periosteal injury current. As a result the electrical field returned to normal, shutting off the cellular activities of healing.

When we'd finished this series of experiments, I was sure this was the most important piece of work I would ever do, and I was determined to get it published as a major article, not just a short note. Luck was with me. On several speaking engagements during the previous two years I'd been able to talk at length with Dr. Urist. He was enthusiastic about our findings, and since he was the editor of one of the major orthopedic journals—Clinical Orthopedics and Related Research—I submitted our report there. The editorial board published it uncut, and I was pleased to use Dr. Pritchard's statement at the 1965 bone-healing workshop as an epigraph. To me it's still the most satisfying of my publications.

This was the first time the control system for a healing process had been worked out in such detail. Except for the less conclusive account Dave and I had presented to the New York Academy of Sciences three years before, it was also the first really incontrovertible proof of dedifferentiation and metaplasia.

These were hardly new ideas, of course. Dedifferentiation had often been proposed during the previous four decades as the simplest explanation of blastema formation, and a great deal of evidence for it had been amassed. Elizabeth Hay had even published an electron microscope photograph of a blastema cell that hadn't despecialized completely and still contained a piece of muscle fiber. Nevertheless, the idea was dismissed by most of the biologists who wielded influence in grant review committees and universities.

Today, however, dedifferentiation is no longer a dirty word. In part, this is because Dave and I devised a way to produce it artificially, which could be repeated by anyone who cared to. Art Pilla, an electrochemist working with Andy Bassett in New York, was the first to confirm our method. I'm happy to have been able to play a major role in this hard-won advance of knowledge.

Even more important, this was the first work of mine that led directly to a technique that helped patients—electrical stimulation of bone healing (see Chapter 8). Meanwhile, our results led to another major question: Couldn't the currents we'd found be used artificially to stimulate other types of regeneration? We decided to see if we could bring limb regrowth a step closer to humans by trying to induce it in rats.

Seven

Good News for Mammals

Stephen D. Smith was the first to induce artificial regrowth with electricity applied to the limb of a nonregenerating animal. In 1967 Smith, setting forth on his own at the University of Kentucky after his apprenticeship to Meryl Rose at Tulane, implanted tiny batteries in adult frogs' leg stumps. I followed his work eagerly and was elated to hear that he'd gotten the same amount of partial regrowth that had resulted from Rose's salt, Polezhaev's needles, and Singer's rerouted nerves. Of all the experiments that have influenced me, this was probably the one that encouraged me the most.

For a battery that was small and weak enough, Smith had returned to the simple technology of Galvani and Volta. He soldered a short piece of silver wire to an equal length of platinum wire, and put some silicone insulation around the solder joint. He chose these two metals as being the least likely to release ions and produce spurious effects by reacting with the surrounding tissue. When immersed in a frog's slightly saline body fluids, this bimetallic device produced a tiny current whose voltage was positive at the silver end and negative at the platinum end.

Since our work on frog erythrocytes hadn't yet been published, it was sheer luck that the current from these batteries fell close to the "window of effectiveness" for blastema formation. As Smith later wrote: "It would he nice to be able to say that I had worked out all the parameters in advance, and knew exactly what I was doing, but such was not the case. As so often has happened in the history of science, I stumbled onto the right procedure."

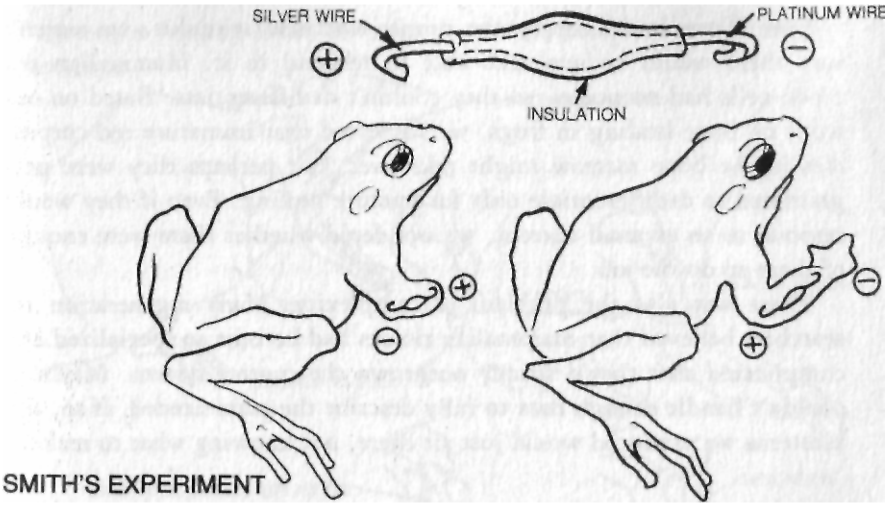

SMITH'S EXPERIMENT

Smith implanted his wires along the bone remnant, with one end bent over into the marrow cavity. The limbs with the positive silver electrode at the cut showed no growth, and in some cases tissue actually disintegrated. The negative platinum ends, however, started regeneration; the new limbs all stopped growing at about the same distance from the device, suggesting that regeneration might have been complete if the batteries had been able to follow along. In 1974 Smith made a device that could do just that, and achieved full regrowth.

Despite Smith's success, there was no reason to suppose that his method would work in mammals. One researcher had recently noted some regeneration in the hind legs of newborn opossums, but, since marsupials are born very immature and develop in the mother's pouch to a second birth, we suspected that this was merely a case of embryonic regrowth. Most fetal tissues were known to have some regenerative ability while they weren't yet fully differentiated. Richard Goss had shown that the yearly regrowth of deer and elk antlers was true multitissue regeneration, but this feat seemed too specialized to make us confident about restitution in other mammals or other parts of the body.

Many thought all such attempts were doomed, because the process of encephalization had progressed much further in mammals than amphibians. All vertebrates were known to have roughly the same ratio of nerve tissue to other kinds of tissue, but in mammals most of the limited nerve supply went into the ever more complex brain, until, as Singer had shown in a recent study, the proportion of nerve to other tissue in rat legs was 80 percent less than in salamander legs. This was well below the critical mass needed for normal regeneration, and we thought it might be impossible to make up the difference artificially.

Even if we could supply the proper electrical stimulus, we weren't sure there would be any cells able to respond to it. Mammalian red blood cells had no nuclei, so they couldn't dedifferentiate. Based on our work on bone healing in frogs, we suspected that immature red corpuscles in the bone marrow might take over, but perhaps they were programmed to dedifferentiate only for fracture healing. Even if they would respond to an external current, we wondered whether there were enough of them to do the job.

There was also the problem of complexity. Many regeneration researchers believed that mammalian tissues had become so specialized and complicated that they'd simply outgrown the control system. Maybe it couldn't handle enough data to fully describe the parts needed. If so, any blastema we produced would just sit there, not knowing what to make.

A First Step with a Rat Leg

I tested the kind of silver-platinum couplings Smith used and found they delivered several times too much current for ideal dedifferentiation, according to our frog experiments. Joe Spadaro, another of Charlie's grad students, suggested that we put carbon resistors between the two metals, giving us devices of various current levels.

In 1971, Joe and I amputated the right forelegs of thirty-five rats. We made the cuts through the upper foreleg well away from the elbow so that only the bone shaft, which had long ago ceased growing, would remain at the tip. We used all males, to obviate as many hormonal variables as possible. As controls we treated some of the stumps with no device, or one made of a single metal, or one with the silver positive end facing the stump. We did the actual test on twenty-two of the rats, implanting our batteries with the negative platinum electrode at the wound. We tucked the outer electrode into the marrow cavity and sutured the inner one to the skin of the shoulder.

We had an answer fast. After three days the stumps of the controls had begun to heal over or even, in the case of the highest-current couplings, die back a little behind the amputation line. But the experimental legs with our medium-current devices, supplying 1 nanoamp, were doing well. In a week, nearly every one had a well-formed blastema and seemed ready to replace the whole limb.

Since healing is very fast in rats, and because we wanted a uniform sample for our first test, we sacrificed all of the controls and most of the test animals at this time, although we spared a few for a month. We cut off the entire healing limb, then fixed, stained, and sectioned it for the microscope.

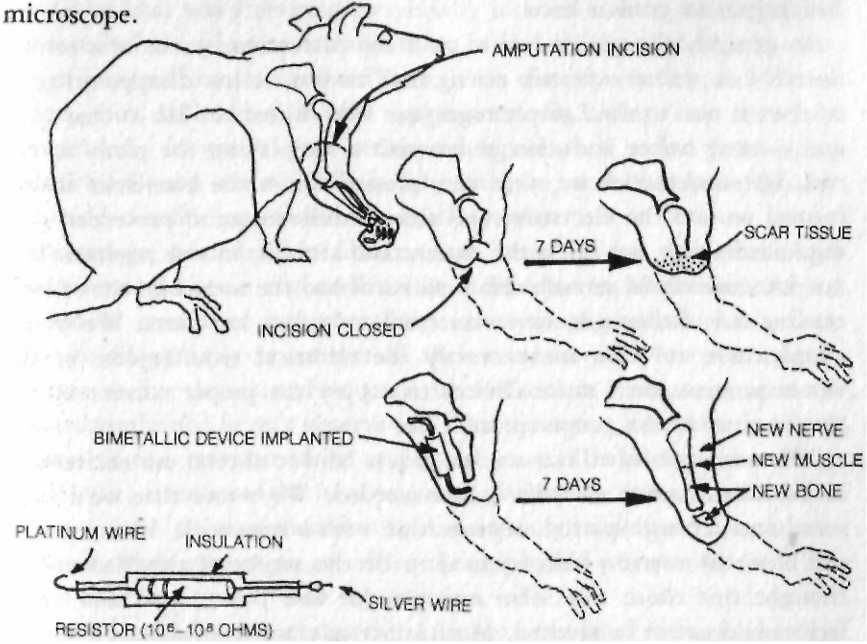
AMPUTATION INCISION
7 DAYS
SCAR TISSUE
INCISION CLOSED
BIMETALLIC DEVICE IMPLANTED
NEW NERVE
NEW MUSCLE
7 DAYS
NEW BONE
PLATINUM WIRE
INSULATION
SILVER WIRE
RESISTOR (10^6–10^8 OHMS)

DC STARTS LIMB REGENERATION IN RATS

I shall never forget looking at the first batch of specimens. The rat had regrown a shaft of bone extending from the severed humerus. At the proper length to complete the original bone there was a typical transverse growth plate of cartilage, its complex anatomical structure perfectly regular. Beyond that was a fine-looking epiphysis, the articulated knob at each end of a limb bone. Along the shaft were newly forming muscles, blood vessels, and nerves. At least ten different kinds of cells had differentiated out from the blastema, and we'd succeeded in getting regeneration from a mammal to the same extent as Rose, Polezhaev, Singer, and Smith had done in frogs.

Slides from some of the other animals were even more spectacular. One stump had two cartilaginous deposits that looked like precursors of the two lower arm bones beyond a fully formed elbow joint. All of the regenerates were bent toward the electrode, and in one the lower humerus had formed alongside the old shaft rather than as an extension of it, but otherwise its structure was quite normal.

With one exception, slides from longer than a week were less exciting. They seemed to have gotten less organized as time went on. Behind one of these older slastemas, at the end of a nearly unformed ghost of a bone, we found cartilage in a five-fingered shape—this limb had begun to grow a hand.

In general, though, it looked as if the current had to be of a certain duration as well as a certain strength. This was no less disappointing to us than it was to the Life photographer who visited the lab at that time and wanted before and after shots with a rat playing the piano at the end, but nonetheless we were very pleased. Since the blastemas always formed around the electrodes and since redifferentiation proceeded into organized tissue, we knew the current had stimulated true regeneration, not some abnormal growth. Mammals still had the means for the orderly reading out of their genetic instructions to replace lost parts. We would simply have to learn more exactly the electrical requirements of the whole process, then make devices to supply the proper current at the proper time in the proper place.

When we published our results, it was hard to shroud our excitement in the circumspect scientific jargon needed. We wrote that we'd activated true, though partial, regeneration with a minuscule direct current and that the marrow cells seemed to be the source of the blastema. I thought this claim was sober enough. Joe and I cautioned that other factors had yet to be studied. Most important, we warned that if such a tiny force could so easily switch on growth, it must be very powerful, and we'd best know it thoroughly before using it routinely on humans, lest we give them unwelcome growths—tumors.

I felt that, within the constraints of scientific propriety, we'd uttered a rousing call for a big research push to open up the benefits of regeneration to humans. It must have been a whisper, though, for it caused no more ripples than a feather settling on a frog pond.

Philip Person, a dental surgeon at the Brooklyn VA hospital and a friend whom I'd known for years, asked me to present our results to the New York Academy of Medicine. Before the academy would permit this, however, it insisted that two experts must visit the lab and look at the actual data. One was Marc Singer, who enthusiastically agreed that we'd really started regeneration in the rat. The other man was totally negative, but he wasn't a specialist in regeneration, so the academy permitted me to speak.

Singer was one of the few who showed much enthusiasm when I'd finished reading my paper at the meeting. Most of the audience was unresponsive, there were few comments or criticisms To these people, electric growth control was still a vitalistic impossibility, and they seemed unwilling to discuss dedilfferentation. The man who'd visited our lab with Singercomplained that the amount of new growth was small. Phil pointed out that it wasn't the quantity but the quality of new tissue that was important, especially in such a short time. Singer, convinced of the paramount importance of the nerves, thought the current might be stimulating them rather than directly causing dedifferentiation, but still thought the experiment was a big step forward. Nevertheless, it wasn't even attempted again until seven years later, when Phil Person himself took on the task; he, and later Steve Smith, confirmed our findings with even better results.

Meanwhile, buried in the literature we found reports that others had already observed some regeneration in mammals. In 1934 Hans Selye, the famous researcher into the effects of stress, discovered that a rat's limbs could partially regenerate of their own accord when the animal was two to five days old. Five years later Rudolph F. Nunnemacher of Harvard confirmed Selye's observation. Nunnemacher, however, ascribed the growth to a remnant of the epiphyseal plate. The growth-plate cells, he thought, simply might have kept on growing as normal in the adolescent animal. Selye replied that he'd specifically made sure to amputate the limbs high enough to get all of the epiphyseal plate so he could be certain that any growth was regenerative.

Thus Joe and I found that we'd really just extended the age limit for regeneration in the rat. Indeed, two years later Phil Person showed that even the young adult rats we'd used occasionally exhibited some regrowth, a fact that had puzzled us in a couple of our control animals. So, to be exact, our electrodes had temporarily but drastically boosted the efficiency of the process as it normally waned with age in the rodent. Still, it was the first time that had ever been done in a mammal.

Childhood Powers, Adult Prospects

The amputation of a fingertip—by a car door, lawn mower, electric fan, or whatever—is one of the most common childhood injuries. The standard treatment is to smooth the exposed bone and stitch the skin closed, or, if the digit has been retrieved and was cleanly cut, to try to reattach it by microsurgery. The sad fact is that even the most painstaking surgery gives less than optimal results. The nails are usually deformed or missing, and the fingers are too short and often painful, with a diminished or absent sense of touch.

In the early 1970s at the emergency room of Sheffield Children's Hospital in England, one youngster with such an injury benefited from a clerical mixup The attending physician dressed the wound, but customary referral to a surgeon for closure was never made. When the error was caught a few days later, surgeon Cynthia Illingworth noticed that the fingertip was regenerating! She merely watched nature take its course.

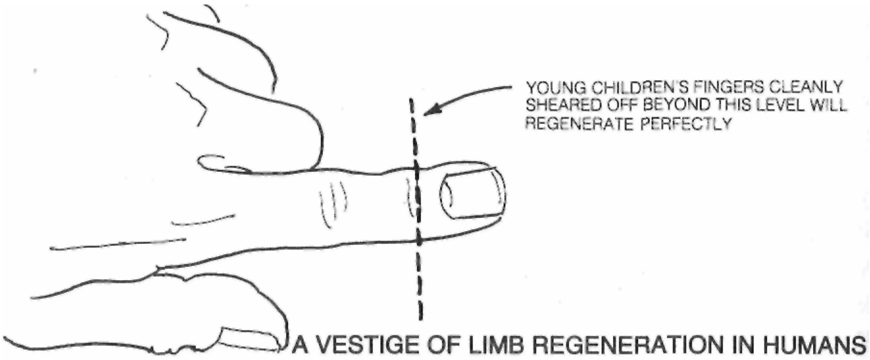
YOUNG CHILDREN'S FINGERS CLEANLY SHEARED OFF BEYOND THIS LEVEL WILL REGENERATE PERFECTLY

A VESTIGE OF LIMB REGENERATION IN HUMANS

Illingworth began treating other children with such "neglect," and by 1974 she'd documented several hundred regrown fingertips, all in children eleven years old or younger. Other clinical studies have since confirmed that young children's fingers cleanly sheared off beyond the outermost crease of the outermost joint will invariably regrow perfectly in about three months. This crease seems to be a sharp dividing line, with no intermediate zone between perfect restoration and none at all.

Some pediatric surgeons, like Michael Bleicher of New York's Mount Sinai Hospital, have become so confident in the infallibility of the process that they'll finish amputating a fingertip that's just hanging by a bit of flesh. A lost one will regenerate as good as new, whereas one that has merely been mutilated will heal as a stump or with heavy scarring.

Fingertip regrowth is true multitissue regeneration. A blastema appears and redifferentiates into bone, cartilage, tendon, blood vessels, skin, nail, cuticle, fingerprint, motor nerve, and the half-dozen specialized sensory-nerve endings in the skin. Like limb regeneration in salamanders, this process only occurs if the wound isn't covered by a flap of skin, as in the usual surgical treatment. Illingworth and her co-worker Anthony Barker have since measured a negative current of injury leaving the stump.

Sadly, natural replacement has been accepted only at a few hospitals. Bleicher laments his colleagues' resistance to the evidence: "Mention it to young residents just our of the training program, and they look at you as though you're crazy. Describe it on grand rounds or at other institutions, and they tell you it's hogwash." Nearly all surgeons cling instead to flashier and vastly more expensive yet less effective microsurgical techniques or simple stitches and stunted fingers.

This discovery and our own research indicated that the potential for at least some artificial regeneration was clearly quite good in young mammals. But what about the ones who needed it most—us older folks whose parts were more likely to be injured or broken down? The answer came unexpectedly several years later, in a way that showed the futility of adhering too rigidly to one's original plan. The scientist must be free to follow unexpected paths as they appear.

I always expected each of my associates, whether student or established researcher, to follow some independent project unrelated to our work together. In 1979, a young assistant named James Cullen (now a Ph.D. investigator in anatomy at the Syracuse VA hospital) proposed to study what would happen if nerves were implanted into the bone marrow of rats. Jim thought the nerves should induce new bone to form in the marrow cavity. Since the idea seemed logical and the technique might supplement the electrical bone-healing devices we'd developed by then, I encouraged him to go ahead.

Jim ran into technical problems right away. He could easily dissect the rat's sciatic nerve out of the hind leg, but getting it into the marrow cavity through a hole drilled in the thighbone was like trying to push a strand of limp spaghetti through a keyhole. He resorted to drilling two holes in the femur, passing a wire suture into the outer one, up the femur, and out the hole nearer the hip. Then he looped the wire around the nerve and pulled it into the marrow cavity using the suture. However, after doing a number of these, Jim decided that there had to be a better way. He decided to amputate the rat's hind leg halfway between the hip and the knee. He could then drill a hole into the marrow cavity just below the hip, pass a suture through it, and pull the nerve down the cavity and out the end of the bone remnant. This was much easier and made a better connection of nerve to bone, so Jim prepared a number of animals this way, only to find that the nerve had a disconcerting tendency to pull back, out of the femur. The amputation didn't faze the rats; they used the stump vigorously, and this caused the nerve to retract.

In those few animals whose nerve had stayed in place, an interesting bone formation had appeared in the marrow cavity. To secure the nerve and look for the same result in other animals, Jim sutured the nerve to the skin that we closed back over the stump. The stitch held the nerve in place, all right, but one animal so treated gave us a totally unpredicted and fascinating result: The missing portion of the femur partially regenerated. While this was surprising enough, the most startling fact was that Jim had used a group of surplus rats about six month old.

These rats were well into adulthood , when mammals were thought to lose all powers of regeneration except fracture healing. What had happened?

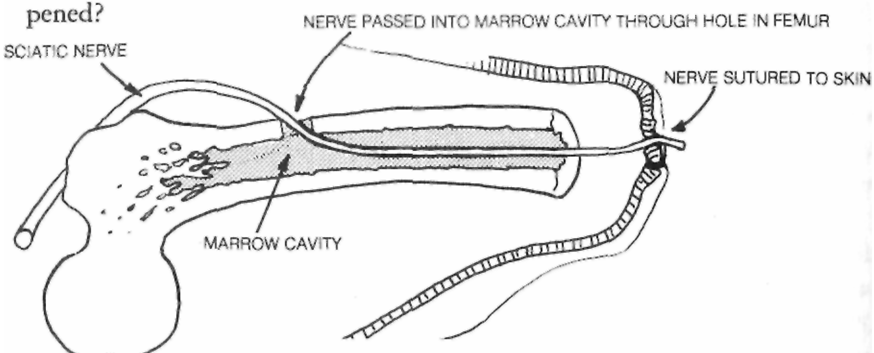

AN ARTIFICIAL NEUROEPIDERMAL JUNCTION IN RATS

Closer examination revealed that we'd made a hole in the skin when we sutured the nerve to it. The nerve appeared to have grown into contact with the epidermis. One of the requirements for normal regeneration of a salamander limb was a neuroepidermal junction, and it looked as though this had formed spontaneously in our one lucky rat when the two tissues were brought together by surgery.

We changed the course of the experiment by operating on the other rats to unite the sciatic nerve and epidermis, after scraping away the dermis. We used animals of various ages. The results exceeded our expectations. Even the old rats regenerated their thighbones and much of the surrounding tissue.

This offered an unparalleled opportunity to find out what it was about the neuroepidermal junction that was so important. We prepared one group of animals with a surgical neuroepidermal junction exactly as before. We prepared a second group the same way, except that we sutured the nerve to the end of the bone, a millimeter away from the hole and with no contact with the epidermis. The first group regenerated, while the second group showed normal rat healing with no growth. The important observation, however, came from electrical measurements we made every day on the stumps. In those animals that formed a neuroepidermal junction, we found electrical potentials following the same curve I'd found in the salamander. The voltage was about ten times as high, but the pattern was exactly the same. In the animals having no neuroepidermal junction, the potentials followed the same curve as in the nonregenerating frog.

We'd discovered that the specific electrical activity that started regeneration was produced by the neuroepidermal junction, not by the simple bulk of nerve in the limb. My original that the direct-current control system was located in the nerves now had to be expanded to include the electrical properties of the epidermis as well. The nerve fibers joined the epidermal cells like plugs into sockets to complete the exact circuit needed for a dedifferentiative current. Furthermore, since the neuroepidermal junction was located over the end of the stump, it continually produced blastemal cells exactly where needed, at the growing tip. This discovery was enormously important, then, because it proved beyond doubt that the electrical current was the primary stimulus that began the regenerative process, and that it could operate even in mammals.

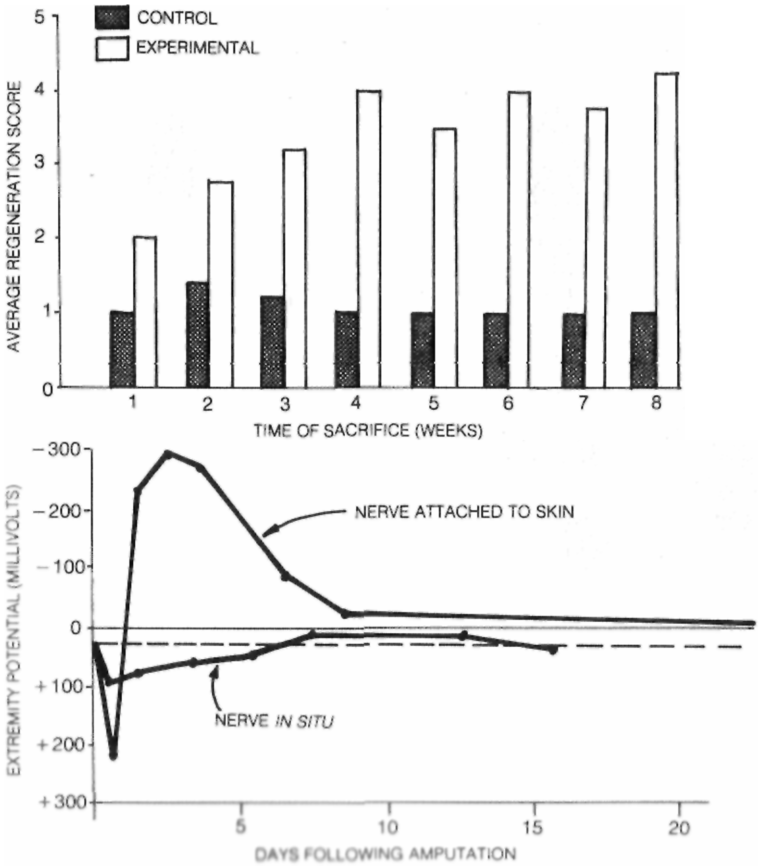

PRODUCING A REGENERATION CURRENT IN RATS

Another experiment showed us a formidable obstacle, however. In an additional series of rats we made a neuroepidermal junction, not at the end of the amputation stump, but on the side of the leg. There we measured the same "regenerative" electrical changes, but nothing happened. There was no growth. This meant that there were no sensitive cells at this site—no cells able to dedifferentiate in response to the current. In mammals, it seemed, such cells were found only in the bone marrow, a sparse cell population to serve as a source of raw material, especially in adult animals.

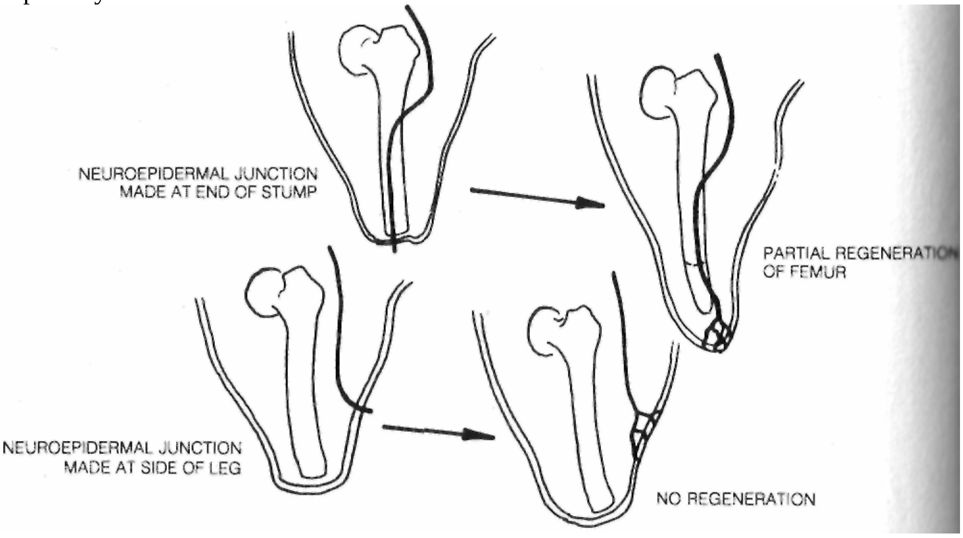
NEUROEPIDERMAL JUNCTION
MADE AT END OF STUMP
PARTIAL REGENERATION
OF FEMUR
NEUROEPIDERMAL JUNCTION
MADE AT SIDE OF LEG
NO REGENERATION

AN ARTIFICIAL NEUROEPIDERMAL JUNCTION STARTS REGROWTH IN RATS

This explained why we never got complete regrowth in any of our rats. The results were typical of an inadequate blastema. There weren't enough sensitive cells in the bone marrow to make a blastema big enough to produce a whole leg. The prospects for full limb regeneration in humans, then, looked very dim—unless we could come up with a way of making other cells electrically sensitive so as to transform them into despecialized blastemal cells. Luckily, while working on a completely different problem described in the following chapter, we stumbled upon a way to do just that.

Part 3
Our Hidden Healing Energy

Disease [is] not an entity, but a fluctuating condition of
the patient's body, a battle between the substance of
disease and the natural self-healing tendency of the body.
—Hippocrates

Eight
The Silver Wand

When Apollo whisked Aeneas off the field of battle before Troy, he healed the hero's shattered thighbone in a matter of minutes. Without a god at the bedside, the process takes three to six months, and sometimes it fails. If the bones didn't knit, the limb formerly had to be amputated after the victim had suffered for a year or more.

It was only in 1972 that I felt ready to try electrical stimulation of human bone growth in such cases. Zachary B. (Burt) Friedenberg, Carl Brighton, and their research group at the University of Pennsylvania had already reported the first successful electrical healing of a nonunion two years before, but to avoid possible side effects we felt we must duplicate the natural signal more closely than they had, and we didn't know enough until after our work on rat leg regeneration. Like Friedenberg we decided to place a negative electrode between the bone pieces, but using a much smaller current and a silver electrode rather than stainless steel. We thought silver would be less likely to react chemically with the tissue and better able to transmit the electrical current. At that time we were treating a patient whose condition seemed to demand that we try the new procedure.

Minus for Growth, Plus for Infection

Jim was in bad shape. Drafted during the Vietnam War, he'd been a reluctant, rebellious soldier. He survived his tour in Nam and was transferred to an Army base in Kansas late in 1970. On New Year's Eve he broke both legs in an auto crash. The local hospital put him in traction, with pins drilled through the skin and bones to hold the pieces together. When he was moved to the base hospital a few days later, all the pins had to be removed due to infection.

Jim's doctors couldn't operate because of the bacteria, so they had to be satisfied with a cast. Because he'd broken one leg below and the other above the knee, he needed a huge cast called a double hip spica. He was totally encased in plaster, from his feet to the middle of his chest, for six months. By August, his left lower leg had healed, but the right femur showed no progress at all. The quarter-inch holes where the pins had been were still draining pus, preventing surgery. That September he was given a medical discharge and flown to the Syracuse VA hospital.

When I first saw him, he was still in a large cast, although now his left leg was free. The halves of the right thighbone were completely loose. There was nothing in standard practice to do but leave the cast on and hope. After six more months Jim's hope was just about gone. For a year he'd lain in bed, unable to leave the hospital for even a brief visit home. He vented his rage against the staff, then grew despondent and unable to face the future, which no longer seemed to include his right leg.

Then Sal Barranco, a young orthopedic surgeon in his last year of residency, was assigned to my service from the medical school. He'd already been a good doctor when he briefly worked with me two years before—smart, hardworking, and really interested in his patients. He took over Jim's care, spent many hours talking with him, and arranged for counseling. Nothing seemed to help. Jim slipped further and further away from us.

Sal had always been interested in what was going on in the lab. In fact, I'd often tried to interest him in a career of teaching and research, but he preferred surgery and its rewards of helping people directly. In February of 1972, as we were nearing the clinical stage with our bone stimulator, Sal said, "You know, Dr. Becker, you really should consider electrically stimulating Jim's fracture. I don't see anything else left. It's his last chance."

The problem was that none of Friedenberg's patients had been infected. Although Jim's septic pin tracts weren't right at the fracture, they were too close for comfort. If I stirred up those bacteria when I operated to insert the electrodes, the game was lost. Moreover, it was obvious by now that electricity was the most important growth stimulus to cells. Even if it produced healing, no one could be sure what these cells would do in the future. They might become hypersensitive to other stimuli and start growing malignantly later. This was the first time in the history of medicine that we could start at least one type of growth at will. I was afraid of beginning a clinical program that might seize the public's fancy and be applied on a large scale before we knew enough about the technique. If disastrous side effects showed up later, we could lose momentum toward a revolutionary advance in medicine. I decided that if I carefully explained what we proposed to do, with all its uncertainties, and let the patient choose, then ethically I'd be doing the right thing.

As to the infection, for several years we had been looking for a way to stop growth. My experiments with Bassett on dogs back in 1964 suggested that just as we could turn growth on with negative electricity, so we could turn it off with positive current. If true, this obviously could be of great importance in cancer treatment. Because ours was always a needy program, trying to do more than we had grants for, we couldn't afford the expensive equipment needed to test the idea on cancer cells. We had to settle for bacteria.

In preliminary tests we found that silver electrodes, when made electrically positive, would kill all types of bacteria in a zone about a half inch in diameter, apparently because of positive silver ions driven into the culture by the applied voltage. This was an exciting discovery, because no single antibiotic worked against all types of bacteria. I thought that if I inserted the silver wire into Jim's nonunion and the area became infected, I could as a last resort make the electrode positive and perhaps save the leg a while longer. Of course, the positive current could well delay healing further or actually destroy more bone.

I explained all this to Jim and said that, if he wished, I would do it. I wanted him to know the procedure was untested and potentially dangerous. With tears in his eyes, he begged, "Please try, Dr. Becker. I want my leg."

Two days later, Sal and I operated through a hole in the cast. The fracture was completely loose, with not one sign of healing. We removed a little scar tissue from the bone and implanted the electrode. The part in between the bone ends was bare wire; the rest, running through the muscles and out of the skin, was insulated so as to deliver the minuscule negative current only to the bone.

The infection didn't spread, and Jim's spirits improved. As I made my daily rounds three weeks later, he said, "I'm sure it's healing. I just know it!" I was still nervous when, six weeks after surgery, it was time to pull out the electrode, remove the cast, and get an X ray. I needn't have worried. Not only did the X ray show a lot of new bone, but when I examined the leg myself, I could no longer move the fracture! We put

Jim in a walking cast, and he left the hospital for the first time in sixteen months. In another six weeks the fracture had healed enough for us to remove the cast, and Jim started rehabilitation for his knee, which had stiffened from disuse.

All the pin tracts, especially the ones nearest the break, were still draining, and Jim asked, "Why not use the silver wire on this hole to kill the infection? Then I'll be all done and won't have to worry anymore about infecting the rest of the bone." I had to agree with his logic. If the hole through the muscle to the outside healed shut, the infection would be more likely to spread within the bone. However, I told him that the positive current might prevent the hole from filling in with bone, making a permanent weak spot there.

We put in the electrode and used the same current as before, except reversing its polarity. I had no idea how long to let it run, so I arbitrarily pulled it out after one week. Nothing much seemed to have happened. The drainage might have been a little less, though not much; but I was afraid to use the positive current anymore for fear of further weakening the bone.

Jim left the hospital and didn't keep his next appointment in the clinic. A year later he returned unannounced saying he was just traveling through Syracuse and thought I would like to see how he was doing. He was walking normally, with no pain, placing full weight on the right leg. He said the drainage had stopped a week after he left the hospital and had never recurred. X rays showed the fracture solidly healed and the one pin tract I'd treated filling in with new bone. The pin site on the other leg was still infected, and I said we could treat that in a few days, since we'd improved our technique in the meantime. "No, I have to be moving on," Jim replied. "I don't have a job. I don't know what I'm going to do, but I know I don't want to spend any more time in hospitals."

Sal had been graduated from the residency program a few months after Jim was discharged in 1973, but before he left he spent all his free time in the lab helping us test the bactericidal (bacteria-killing) electrodes. A few previous reports had mentioned inconsistent antibacterial effects, some with alternating current, some with negative DC using stainless steel, but there had been no systematic study of the subject. We tried silver, platinum, gold, stainless steel, and copper electrodes, using a wide range of currents, on four disparate kinds of bacteria, including Staphylococcus aureus, one of the commonest and most troublesome.

Soon we were able to explain the earlier incosistencies: All five metals stopped growth of all the bacteria at both poles, as long as we used high currents. Unfortunately, high currents also produced toxic effects—chemical changes in the medium, gas formation, and corrosion—with all but the silver electrodes. Apparently such currents through most metals "worked" by poisoning both bacteria and nearby tissues.

Our preliminary observations turned out to be right. Silver at the positive pole killed or deactivated every type of bacteria without side effects, even with very low currents. We also tried the silver wires on bacteria grown in cultures of mouse connective tissue and bone marrow, and the ions wiped out the bacteria without affecting the living mouse cells. We were certain it was the silver ions that did the job, rather than the current, when we found that the silver-impregnated culture medium killed new bacteria placed in it even after the current was switched off. The only other metal that had any effect was gold; it worked against Staphylococcus, but not nearly as well as silver.

Of course, the germ-killing action of silver had been known for some time. At the turn of the century, silver foil was considered the best infection-preventive dressing for wounds. Writing in 1913, the eminent surgeon William Stewart Halsted referred to the centuries-old practice of putting silver wire in wounds, then said of the foil: "I know of nothing which could quite take its place, nor have I known any one to abandon it who had thoroughly familiarized himself with the technic of its employment."

With the advent of better infection-fighting drugs, silver fell out of favor, because its ions bind avidly to proteins and thus don't penetrate tissue beyond the very surface. A few silver compounds still have specialized uses in some eye, nose, and throat infections, and the Soviets use silver ions to sterilize recycled water aboard their space stations, but for the most part medicine has abandoned the metal. Electrified silver offers several advantages over previous forms, however. There are no other ions besides silver to burden the tissues. The current "injects" or drives the silver ions further than simple diffusion can. Moreover, it's especially well suited for use against several kinds of bacteria simultaneously. It kills even antibiotic-resistant strains, and also works on fungus infections.

For treating wounds, however, there was one big problem with the technique. Its effect was still too local, extending only about a quarter inch from the wire. For large areas we needed something like a piece of window screen made of silver, but this would have been expensive and also too stiff to mold into the contours of a wound.

We'd been doing our clinical experimens with financial support from a multinational medical-equipment company that made our "black boxes," the battery packs with all their circuitry that powered our electrodes. I discussed the problem with the company's young research director, Jack TerBeek, and a few weeks later he came back with a fascinating material. NASA needed an electrically conductive fabric, and a small manufacturing company had produced nylon parachute cloth coated with silver. It could be cut to any size and was eminently flexible.

It performed beautifully. Although the silver ions still didn't get more than a quarter of an inch from it, we could use it to cover a large area. Hopeful that we might have a cure for two of an orthopedist's worst nightmares—nonunion and osteomyelitis (bone infection)—we studied the positive silver technique in the lab and continued to use the negative electrodes to stimulate bone growth in selected patients. Word spread via newspaper and TV reports. We began getting patients from all over the nation, but we didn't accept many for the experimental program due to my conservative viewpoint. I applied the same criterion as before: Electrical treatment had to be the patient's last chance.

While slowly gaining experience, we surveyed the literature to stay informed about other people's work. As of 1976, fourteen research groups had used bone stimulators on some seven hundred patients, for spinal fusions and fresh fractures as well as nonunions, all with seemingly good results.

We'd used our electrical generator on only thirteen patients by then. We were the only ones using silver electrodes, a lucky choice as it turned out; all the others were using stainless steel, platinum, or titanium. We used 100 to 200 nanoamps per centimeter of electrode, while Brighton and most other investigators were using 10,000 to 20,000 nanoamps. The low level approximated the natural current and also minimized the chances of a dangerous side effect. Brighton and Friedenberg had found a danger of infection and tissue irritation when running their high-current electrodes at more than 1 volt. We figured this couldn't happen at our amperage, but just to be sure we built in an alarm circuit to shut off our box automatically if the electrical force rose close to 1 volt.

By this time we'd also cleared up several more cases of osteomyelitis by reversing the battery and making the silver electrode positive for a day. It looked safe. There was no crossover of effects: When negative, the wire didn't make infectious bacteria grow, and when positive, it didn't destroy bone-forming cells or prevent them from growing when we switched the current to negative. Our confidence in this method grew with one of our most challenging cases, which also forced us to revise our theories.

Positive Surprises

In December of 1976 a young man was sent to our clinic for a possible amputation. John was a man of the north woods. Weathered and hard, he faced the problem philosophically. "What's got to be has got to be," he said through tight lips. Three years before, he'd been in a snowmobile accident, breaking his right tibia (shinbone) in three places and also fracturing the fibula, the smaller bone of the lower leg. He'd been treated at a small local hospital, where the broken bones had become infected. He'd undergone several operations to remove dead bone and treat the infection, but the bacteria continued to spread. He came to us with the fracture still not healed and with a long cavity on the front of his shin in which one could see right into the dead and infected bone. He was struggling to walk in a cast extending up to his hip. He was married, with five young children, and his leg was obviously not the only place he was having trouble making ends meet.

"What kind of work do you do?" I asked him.

"I trap muskrats, Doc."

"That's all?"

"That's all, Doc."

"How in hell do you manage with that cast on?"

"I put a rubber hip boot over the cast, Doc."

Muskrat trapping is hard work, a tough way to make a living even for a man with two good legs. "John, if you have an amputation and wear an artificial leg, you sure won't be able to do that. What will you do then?"

"I dunno—welfare prob'ly. Prob'ly go nuts."

"You really like to work in the woods, don't you?"

"Wouldn't do anything else, Doc."

"Well, let's get you admitted to the hospital. Something has to be done, and I have an idea that might let you keep your leg." For the first time, John smiled.

In fighting the infection, the first step was to identify the enemy, the microbes. John's wound was a veritable zoo. There were at least five different types of bacteria living in it. Even with only one kind, osteomyelitis is notoriously hard to treat. Very little blood reaches the bone cells, so both antibiotics and the body's own defense agents are hampered in getting where they're needed. And even if we could get it into the bone, no single antibiotic could fight all of John's germs. Even a mixture would probably create a greater problem than it solved, for any bacteria resistant to the mix would spread like wildfire when the others competing against them were killed.

John's X rays were as chaotic as his bacteria cultures—pieces of dead bone all over the place with absolutely no healing—but we had to deal with the infection first. Since we'd have to use positive current for quite a while, I was afraid we'd destroy some of the bone, but I told John that six months after we got the wound to heal over with skin, I would bring him back into the hospital and use the negative current to stimulate whatever was left. I couldn't promise anything and, since I hadn't yet tried the silver nylon on this type of wound, we might run into unexpected problems. But John agreed with me that he had nothing to lose except his leg, which would certainly have to come off if we didn't try my plan.

A few days later I debrided the wound, removing the dead tissue and all grossly infected or dead bone. There wasn't much left afterward. It was an enormous excavation running almost from his knee to his ankle. In the operating room we soaked a big piece of silver nylon in saline solution and laid it over the wound. It had been cut with a "tail," serving as the electrical contact and also as a sort of pull tab that we could keep dry, outside of the cavity. We packed the fabric in place with saline-soaked gauze, wrapped the leg, and connected the battery unit.

I watched John anxiously during the first two days. If trouble was going to occur, that was when I expected it. By the third day he was eating well, and the current was beginning to drop off, indicating more resistance at the surface of the wound. Now it was time to change the dressing. We were overjoyed to see that the silver hadn't corroded and the wound looked great. Carefully I took a bacterial culture and applied a new silver nylon dressing.

The next morning Sharon Chapin, an exceptional lab technician who took an active part in some of the research, showed me the bacterial cultures. The number of bacteria had dropped dramatically. I went to give John the good news and change his dressing again, when I realized that I could teach him to do his own daily dressing changes. They were time-consuming for me, but John had too much time on his hands and was the one most interested in doing the best thing for his leg. It was a nice feeling to teach a muskrat trapper, who bad dropped out of school at sixteen, how to do an experimental medical procedure. He learned fast, and in a day or so he was changing dressings himself and measuring the current, too. By the end of the week, he allowed as how he did a better job than l did. Maybe he did it at that, because by then all of our bacterial cultures were sterile—all five kinds had been killed. The soft healing tissue, called granulation tissue, was spreading out and covering the bone. In two weeks, the whole base of the wound, which had been over eight square inches of raw bone, was covered with this friendly pink carpet. The skin was beginning to grow in, too, so we could forget about the grafts we thought we'd need to do.

I decided to take an X ray to see how much bone he'd lost. I could hardly believe the picture. There was clearly some bone growth! We'd been working through a hole in the cast, so I had no idea if the fracture was still loose. Without telling John why—I didn't want to get his hopes up if I was wrong—I removed the cast, felt the leg, and found that the pieces were all stuck together. John watched, and when I was done he lifted his leg into the air triumphantly. It held straight against gravity. His grin opened broader than an eight-lane highway. "I thought you said the bone wouldn't heal yet, Doc!"

I'd never so much enjoyed being wrong, but I warned John not to get too excited until we were sure the good news would hold up. I put him back into a cast and continued treatment another month, until the skin healed over. By then the X rays showed enough repair to warrant a walking cast. John left the hospital on crutches and promised not to run around in the swamps until I told him it was all right. He didn't come back until two months later. The cast was in tatters, and he walked in without crutches, smiling at everyone. The last X rays confirmed it: Healing was nearly complete, and John went back to the wilds.

By mid-1978 we'd successfully treated fourteen osteomyelitis patients with the positive silver mesh wire. The funny thing was, in five of them we'd healed nonunions as a "side effect," without any negative current at all. Obviously it was time to revise our idea that negative electricity alone fostered growth and positive inhibited it.

Andy Marino, Joe Spadaro, and I talked it over. Reducing the DC stimulation technique to its essentials, all you needed was an electrode that wouldn't react with tissue fluid when it wasn't passing current. Since a negative electrode didn't give off ions, any inert metal, such as stainless steel, platinum, or titanium, would work with that polarity. But we knew from our lab work that the situation was very different at the positive pole, where the current drove charged atoms of the metal into the nearby environment. We decided it must be chemical, not electrical, processes that were preventing the bacterial growth at the positive electrode. In that case, maybe polarity was unimportant in growth enhancement. We postulated that, because silver ions were nontoxic to human cells and the electrical aspect was right, we inadvertently grew bone with positive current. This idea turned out to be quite wrong, but we'll get to that story in due time.

Joe, who was always fascinated by the history of science, now found that none of the contemporary research groups had been the first to stimulate bone repair electrically. We'd all been beaten by more than 150 years. Back in 1812, Dr. John Birch of St. Thomas' Hospital in London used electric shocks to heal a nonunion of the tibia. A Dr. Hall of York, Pennsylvania, later used direct current through electroacupuncture needles for the same purpose, and by 1860 Dr. Arthur Garratt of Boston stated in his electrotherapy textbook that, in the few times he'd needed to try it, this method had never failed. Because of the primitive state of electrical science then, we didn't know how much current these doctors had used. However, the polarity didn't seem to matter, and they used gold electrodes, which were nearly as nontoxic at the positive pole as silver.

Realizing that we still didn't know as much as we'd thought about the growth control lock, we continued to ply the silver key. At least seventy patients with bone infections have now had the silver nylon treatment, including twenty at Louisiana State University Medical School in Shreveport, where Andy Marino ended up after the closing of our lab in 1980. In some of our first cases we noticed a discharge exuding from the tissues and sticking to the mesh when we changed the dressings. We thought it was "reactive" exudate—from irritation by the current—until one day when, during a slight delay in the operating room, I sent a sample of it to the pathology lab. It was filled with such a variety of cells that we had to rule out a simple response to irritation. Instead there was a variety of primitive-looking cell types, looking just like the active bone marrow of children. However, the patients were long past that age, and, besides, their marrow cavities were closed off with scar tissue from their unmended and infected bone injuries. We had to consider another source.

The exudate appeared at the same time as the granulation tissue, which is composed mainly of fibroblasts, ubiquitous connective cells forming a major part of most soft tissues. Since the exudate also contained some fibroblasts, we decided to see if the unfamiliar types had arisen by metamorphosis from them.

We set up a series of three-compartment culture dishes and placed a standard colony of isolated, pure-bred mouse fibroblasts in each. In one section we put a positive silver electrode, in one a negative electrode, and in the third a piece of silver wire not connected to anything.

In cells right next to all three wires, the cytoplasm changed to an abnormal texture in response to ions of dissolved silver, which migrated only about a hundredth of a millimeter. There were no other effects in the control or negative-current chambers.

Around the positive poles, however, this region was succeeded by an area of great activity for a distance of 5 millimeters on all sides. While doing their job of holding things together, fibroblasts have a characteristic spiky shape, with long sticky branches extending in all directions. In this region where silver ions had been driven by the current, many of the cells had changed to a static, globular form in which mitosis didn't occur. They seemed to be in suspended animation, floating freely instead of adhering to other cells or the sides of the dishes as usual. Mixed among them were many featureless cells with enlarged nuclei, the end products of dedifferentiation. More and more of the rounded fibroblasts turned into fully despecialized cells as the test progressed. Beyond the 5-millimeter line was a border zone with partial changes, followed by a realm of normal, spiky fibroblasts. Dedifferentiated cells normally divide rapidly, but these didn't, perhaps because they were sitting in a plastic dish far removed from the normal stimuli of an animal's body. Within a day after the current was turned off, the cells clumped together into bits of pseudotissue that looked like the young "bone marrow" we'd seen in the exudate. After two weeks they'd all reverted back to mature fibroblasts, presumably because regular replacement of the nutrient medium had by then washed out all the silver ions.

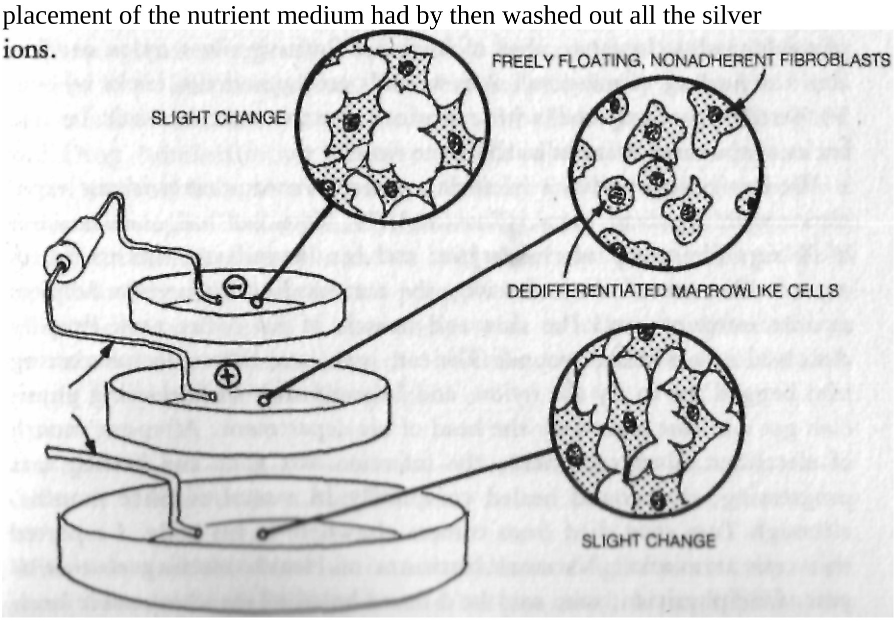
FREELY FLOATING, NONADHERENT FIBROBLASTS
SLIGHT CHANGE
−
DEDIFFERENTIATED MARROWLIKE CELLS
+
SLIGHT CHANGE

POSITIVE SILVER DEDIDIFFERENTIATES CONNECTIVE TISSUE CELLS

To learn more about these astonishing changes, we studied pieces of the granulation tissue itself, taken from patients treated with the silver nylon. We placed the samples in culture dishes and observed them as they grew. Without the silver factor we would have expected a population of slowly proliferating fibroblasts. However, these cells grew fast, producing a diverse and surprising assortment of primitive forms, including fully dedifferentiated cells, rounded fibroblasts, and amoebalike cells. Strangest of all were giant cells that looked almost like fertilized eggs, very active and with several nucleoli. (The nucleolus is a "little nucleus" within the nucleus proper.) When other cells encountered the giant cells, the smaller cells often split open and emptied their nuclei into the giants. After two weeks these diverse cells had coalesced into an amorphous mass of primitive cells closely resembling a blastema, and in another week, as the silver washed out, they'd all become staid, sober fibroblasts acting as though nothing had happened.

The major difference between the two experiments was that the second one started with cells that had already been exposed to positive silver ions in the human body. Their rapid growth and unspecialized forms suggested that the fibroblasts in the first experiment had in fact been dedifferentiated. It remains to be seen exactly what the various forms do, but it's obvious that in the aggregate they profoundly stimulate soft-tissue healing in a way that's unlike any known natural process. We ran a controlled study of the healing enhancement on pigs, their skin being physiologically closest to that of humans. Positive silver nylon accelerated the healing of measured skin wounds on the animals' backs by over 50 percent as compared with identical control wounds made on the backs of the same animals at the same time.

We saw positive silver's lifesaving potential most clearly in our experience with a patient named Tom in 1979. Tom had had massive doses of X rays for cancer of the larynx, and his larynx later had to be removed. Because of the radiation, the surrounding tissue was helpless against infection, and the skin and muscle of his entire neck literally dissolved into a horrid wound. The ear, nose, and throat doctor treating him begged me to try the nylon, and I agreed after the attending physician got a release signed by the head of his department. After one month of electrified silver treatment, the infection was gone and healing was progressing, the wound healed completely in a total of three months, although Tom soon died from tumors elsewhere in his body. I reported this case at a small National Institutes of Health meeting that same year. One physician, who said he'd never heard of any comparable healing of such a grave wound, was moved to exclaim after seeing my slides, "I have witnessed a miracle!"

We may only have scratched the surface of positive silver's medical brilliance. Already it's an amazing tool. It stimulates bone-forming cells, cures the most stubborn infections of all kinds of bacteria, and stimulates healing in skin and other soft tissues. We don't know whether the treatment can induce healing in other parts of the body, but the possibility is there, and there may be other marvels latent in this magic caduceus. Just before our research group was disbanded, we studied malignant fibrosarcoma cells (cancerous fibroblasts) and found that electrically injected silver suspended their runaway mitosis. Most important of all, the technique makes it possible to produce large numbers of dedifferentiated cells, overcoming the main problem of mammalian regeneration—the limited number of bone marrow cells that dedifferentiate in response to electrical current alone. Whatever its precise mode of action may be, the electrically generated silver ion can produce enough cells for human blastemas; it has restored my belief that full regeneration of limbs, and perhaps other body parts, can be accomplished in humans.

Many questions remain, however. We don't know how the changed cells speed up healing or how the silver changes them. We don't know how electrically produced silver ions differ from ordinary dissolved ions, only that they do. They evoke widely different reactions from the fibroblasts, and the cells affected by dissolved ions close to the electrode are prevented from dedifferentiating in response to the electrified silver. Most important is a search for possible delayed side effects.

These questions urgently need research by some good electrochemists, but the work isn't being done. We were probably lucky we hadn't found this effect in our first round of lab tests on silver electrodes. The Food and Drug Administration let us test the antibacterial technique on selected patients because we found no toxicity and because a few hours a day was enough. To say that the same electrodes run for a longer time could stimulate healing was such a bold claim that permission probably would have been denied. At this point, however, we sorely need enough imagination on the part of research sponsors to follow up these leads in the laboratories, while making the treatment available now to the desperate few who have no other hope.

The Fracture Market

Where does all this leave us in our understanding of electrically triggered bone healing? I'm afraid we're not too much better off than the nineteenth-century doctors who lost this effective treatment because no one understood it well enough to defend it from electrotherapy's opponents. Of course, we have more pieces of the puzzle today, but we still don't have a complete picture of how any of the competing techniques work.

Besides our low-current silver electrodes, there are two other basic types of bone growth stimulators. Friedenberg and Brighton at first placed their stiff stainless-steel wires through holes drilled into the bone near the break. Now a "semi-invasive" refinement is used in many patients—sticking the electrodes through the flesh into the fracture gap under local anesthetic; several may be needed for a large bone. They're connected to a self-contained battery pack set right into the cast. Friedenberg and Brighton patented their invention, and it's now approved by the FDA and marketed. About three fourths of the groups treating patients use some variation on this theme.

Australian researchers under D. C. Patterson devised a spiral titanium electrode that's placed into a notch cut in the bone on both sides of the fracture. It's now also FDA approved and marketed. Since this device, battery pack and all, must be implanted and removed in two separate operations, and since the electrode usually must be left behind in the bone, late complications may occur.

Others have taken a completely different approach, using pulsed electromagnetic fields (PEMF) to induce currents in the fracture area from outside the body. The best-known proponents of this method, Andrew Bassett and electrochemist Arthur Pilla, worked together at the Orthopedic Research Laboratories of Columbia-Presbyterian Medical Center in New York until 1982; Pilla is now at the Mount Sinai Medical Center. They developed a pair of electromagnetic coils sheathed in plastic pads, connected to a book-sized generator that plugs into a wall socket.

Having experimented with a wide variety of pulsed fields, Bassett and Pilla found four that stimulate fracture healing. The one that works best, which is now also approved and available commercially, is produced by electromagnets driven by alternating current supplied in bursts of pulses. Although it ties the patient to an electrical outlet for twelve hours a day (mostly during sleep, of course), this apparatus completely avoids surgery and its attendant risks.

The funny thing is that all three methods—low current, high current, and PEMF—seem to work equally well. Since the FDA approved them in late 1979, success rates have stabilized at about 80 percent.

As far as the two electrode methods are concerned, I believe some experiments wedid in 1977 and 1978 revealed why they both work.

When we arranged all the reports in order from lowest current to highest, we found a narrow band of low amperages and a wide band of higher ones that worked, separated by a range in between that failed. We tested various nonsilver electrodes on animals and found that, at the effective currents, they all produced some electrochemical changes even at the negative pole. Among other products, they created various highly reactive ions called free radicals, essential in cellular chemistry but also destructive in excess. These radicals irritate cells, and, since any continuous irritation stimulates bone-forming cells to lay down new matrix in self-defense, we concluded that the higher-current methods worked primarily by such irritation. Conversely, I believe low-current silver electrodes stimulate bone formation directly—by dedifferentiating the marrow cells and perhaps also by stimulating the periosteal cells.

At first, PEMF research suggested that the coils worked by inducing in the tissues electrical currents that changed the permeability of cell membranes to calcium. In most nonunions at least a small amount of fracture callus has appeared, consisting of collagen fibers, but for some\ reason it hasn't entered the next stage, in which apatite crystals form on the fibers. Work by Pilla and Bassett suggested that the currents induced by the pulsed fields caused the cells of the callus to absorb large amounts of calcium. Later, when the coils were turned off, they reasoned, the cells dumped this calcium outside among the collagen fibers, and apatite crystals finally formed where they belonged.

Their experiments raised the hope that other wave forms might regulate membrane passage of other ions or even control DNA transcription and protein synthesis. It seemed these field-induced currents might act as "vocal cords," allowing us to "speak" to the cell nucleus via the membrane, much as sound waves communicate with the brain via the ear.

PEMF does in fact induce currents—of a type never found normally in the body. Each pulse produces millions of tiny eddy currents briefly flowing in circles. As the magnetic field expands at the beginning of a pulse, the currents circle in one direction; as it collapses, they reverse. However, the latest research has cast doubt on the idea that these currents affect specific cellular processes. Rather, it seems that artificial time-varying magnetic fields directly activate the cells by speeding up their mitotic rate, as discussed in more detail in Chapter 15.

You may ask, "As long as something works, why quibble about how?" My answer is that understanding how is our best hope for using the tool right, without causing our patients other problems later. By sticking our own neologisms into the cell's delicate grammar, we automatically risk garbling it in unforeseen ways.

As of now, we're like blind people crossing a minefield. Accelerated mitosis is a hallmark of malignancy as well as healing, and long-term exposure to time-varying electromagnetic fields has been linked to increased rates of cancer in humans. Bassett has discounted potential dangers, saying, "You would experience almost the same field strength by standing under a fluorescent light." However, a fluorescent lamp may well feel like a floodlight to cells that can see nanoamperes of current. One of the main lessons of bioelectromagnetism so far is that less is often more.

On the other hand, it's too easy to assume that "natural is better." Since it would vindicate our low-current method, I obviously hope it's true, but the fact remains that putting electrodes in a bone is itself a very unnatural act. Stimulation of healing outside the normal limits of the process may incur fewer risks in the end. So far the evidence suggests otherwise, but we don't yet know for sure. That's why I keep emphasizing the need to go slowly, using these contraptions only when all else has failed, until we understand them better.

The most urgent need is a search for possible malignant effects. As far as I know, I've done the only such research on electrodes to date—one simple tissue culture study without grant support, using some money I saved from our last research project funded by the manufacturer of the battery pack. I proposed more extensive tests to various granting agencies before our lab was closed, but was turned down every time.

I exposed standard cultures of human fibrosarcoma cells to 360 nanoamps from stainless steel electrodes. I tried it five times, and each time there was a threefold increase in cell population at both electrodes. Even for cancer cells, this is remarkable proliferation for such a short time. To my knowledge, none of the developers or marketers of electrode devices have chosen to duplicate this test or try their own, despite the ease of doing the work and the fact that they have plenty of money. Whatever evidence on this point that may have been presented to the FDA hasn't been made public. After the evaluation panel granted commercial approval, however, several of its members expressed fears that this possibility hadn't been adequately tested. At this time, therefore, I must conclude that high-current electrodes might enhance the growth of any preexisting tumor cells in the electrical path—unlike low-current silver, which when negative had no effect on, and when positive suspended, mitosis of cancer cells in our lab.

As for PEMF, Bassett and Pilla believe that only cells in a healing process gone awry can "hear" their wave form, so they discount the idea of cancel enhancement from it. They claim to have found no PEMFs that produce or accelerate malignant growth; on the contrary, Pilla and oncologist Larry Norton of Mount Sinai say they've found at least one that seems to inhibit it in lab animals. This claim is seriously flawed, however, because of the difference between subjecting an entire animal to a magnetic field, and directing the same field to a small area around a fracture (see Chapter 12). Moreover, in 1983, Akamine, a Japanese orthopedic researcher, reported that the pulsed magnetic fields used for bone healing dramatically increased the mitotic rate of cancer cells. The same fields inhibited the return of such "stimulated" cancer cells to a more normal state. Thus PEMF, like high-current treatments, apparently does enhance cancer growth.

In the last decade or so, electrobiologists have learned a great deal about the effects of time-varying electromagnetic fields (as opposed to steady-state fields) on living tissue. We'll review these discoveries in Chapters 14 and 15. The evidence to date indicates that PEMF works by increasing the mitotic rate of the healing cells, not by altering calcium metabolism. If so, it can't possibly discriminate between bone-healing cells and any other type. It will accelerate the growth of any cellular system that is actively growing; this includes not only healing tissues, but fetal and malignant tissues as well.

At the present rate of basic research, we won't have direct proof on whether electrical healing stimulators are nurturing seeds of cancer in humans until two or three decades from now. We could find out much sooner by simple experiments on animals having shorter life-spans. Until we have that definitive answer, I contend that all three techniques should remain available as a last resort to prevent loss of limb, but I'm appalled at their increasing use to speed up orthodontics or accelerate healing of fresh fractures.

Unfortunately, the trend is away from caution. By the time this book is published, tens of thousands of patients will have been treated with the devices, many as a first, rather than last, resort. At a recent orthopedic meeting, I learned that four more companies are hoping to market new models. Several have asked me to advise them, but I haven't found one yet that wants to embark on any serious research. Without such a commitment, I refuse to take part in any battle of salesmen. I never even tried to patent the low-current silver method, because a medical device generally isn't considered patentable if the research that went into it was conducted throughout the scientific community and published for all to read. As I see it, the rush to the marketplace can only spawn jurisdictional disputes and ensure that important findings are kept as proprietary secrets.

Electrical osteogenesis could be the opening wedge into a new era of medicine. Within a few years, we may know how to use these techniques to repair joint cartilage, or even replace whole joints, and to correct various defects of bone growth. In the further future, we may be able to extend regeneration as needed to nearly all human tissues. For the first time a physician can now direct nature, albeit in a small way, rather than play her helpless servant. We must use this power wisely. We're tapping into the most potent force in all of biology. If we're irresponsible about it, we risk another electrical letdown that could put off medicine's glorious future for many years.

Nine

The Organ Tree

"I have yet to see any problem, however complicated, which when you looked at it the right way did not become still more complicated," science fiction writer Poul Anderson once observed. To a certain extent this is true of regeneration. Intricate nature is still more than a match for our finest-spun hypotheses. Yet we've now reached the oasis of science that we call an interim understanding, where the data begin to shake into place and we can sense the pattern of the rebus from the blanks we've filled in.

Ultimately we must relate all we learn about regeneration to a general system of communication among cells, for regrowth is only a special case of the cooperative cohesion that's the essence of multicellular life. This communication system includes but extends beyond the gene-protein-enzyme subsystems that govern the specialization of cells and unite their chemical trade routes into smoothly working tissues and organs. During embryonic development, cells where muscle will appear must receive instructions from their environment telling them to repress all genes except the muscle genome, or subcode. In many tissues, perhaps in all, chemical inductors from previously formed tissue perform this task, leading embryonic cells step by step through the stages of differentiation. However, chemical reactions and the passage of compounds from cell to cell can't account for structure, such as the alignment of muscle fiber bundles, the proper shape of the whole muscle, and its precise attachment to bones. Molecullar dynamics, the simple gradients of diffusion, can't explain anatomy. The control system we're seeking unites all levels of organization, from the idiosyncratic yet regular outline of the whole organism to the precisely engineered traceries of its microstructure. The DNA-RNA apparatus isn't the whole secret of life, but a sort of computer program by which the real secret, the control system, expresses its pattern in terms of living cells.

This pattern is part of what many people mean by the soul, which so many philosophies have tried to explicate. However, most of the proposed answers haven't been connected with the physical world of biology in a way that offered a toehold for experiment. Like many attempts, the latest major scientific guess, the morphogenetic field proposed by Paul Weiss in 1939, was just a restatement of the problem, though a useful one. Weiss conjectured that development was guided by some sort of field projected from the fertilized egg. As the dividing cell mass became an embryo and then an adult, the field changed its shape and somehow led the cells onward.

The problem was that there were too many "somehows." Even if one accepted Burr's largely ignored measurements of an electric L-field and admitted that it might be the morphogenetic field (a possibility Weiss dogmatically rejected), there was still no way of telling where the L-field came from or how it acted upon cells. Nor was there an explanation of how, if the field was an emanation from the cells, it could also guide them in building an animal or plant. In applying the idea to regeneration, biologists faced the related and seemingly insurmountable problem of how a more or less uniform outflow of energy could carry enough information to characterize a limb or organ. Given the complexity of biological structures, this was even harder than imagining how a field could "somehow" survive when the part it referred to was missing.

However, the morphogenetic field no longer has to account for everything. Acceptance of dedifferentiation lets us divide regrowth into two phases and better understand each. The first phase begins with the cleanup of wound debris by phagocytes (the scavenger race of white blood cells) and culminates in dedifferentiation of tissue to form a blastema. Redifferentiation and orderly growth of the needed part constitute the second phase.

Simplifying the problem in this way should give biologists an immediate sense of accomplishment, for the first stage is now well understood. After phagocytosis, while the other tissues are dying back a short distance behind the amputation line, the epidermal cells divide and migrate over the end of the stump. Then, as this epidermis thickens into an apical cap, nerve fibers grow outward and subdivide to form individual synapselike connections the neuroepidermal junction (NEJ) - with the cap cells. This connection transmits or generates a simple but highly specific electrical signal in regenerating animals: a few hundred nanoamperes of direct current, initially positive, then changing in the course of a few days to negative.

The pituitary hormone prolactin, the same substance that stimulates milk flow in nursing mothers, seems to sensitize cells to electric current. Then the signal causes nearby cells to dedifferentiate and form a blastema, apparently by changing the way cell membranes pass calcium ions. After confirming our frog blood-cell work, Art Pilla went on to produce the same changes by using pulsed DC to make a wave of calcium ions flowing across the culture dish. Steve Smith then confirmed the importance of calcium by preventing dedifferentiation with a calcium-blocking compound, and restarting it with another substance that enhanced passage of calcium ions. Working together, Smith and Pilla next used the same PEMF wave form now in clinical use to nearly double the rate of salamander limb regeneration, while completely preventing it with a different pulse pattern. Widespread recent work on calcium-binding proteins, such as calmodulin, has made it fairly certain that electrical control of calcium movement through cell membranes directs the give-and-take among these proteins, which in turn supervises the cell's entire genetic and metabolic industry.

Although not conclusive, the available evidence suggests that the current flows through the perineural cells rather than the neurons themselves (see Chapter 13). These are several types of cells that completely surround every nerve cell, enclosing all the peripheral fibers in a sheath and composing 90 percent of the brain. Lizards can replace their tails without the spinal cord, as long as the ependyma, or perineural cells surrounding the cord, remains intact, and ependymal tissue transplanted to leg stumps gives lizards some artificial regeneration there. However, the circuit may shift tissues near the wound, for Elizabeth Hay's electron microscope studies clearly show that the peripheral nerve's Schwann cell sheaths stop just short of the epidermis, and only the naked neuron tips participate in the NEJ. The exact current pathway in this microscopic area remains to be charted.

Not all cells can respond, however, as Jim Cullen and I found in one part of our fortuitous rat-regeneration experiments of 1979. Dedifferentiation occurred only when we passed the deviated sciatic nerve to the epidermis through the bone marrow. When we led it through the muscle yet sutured it to the skin in exactly the same way, an NEJ producing the right current appeared, but no blastema and no regrowth. Muscle cells apparently weren't competent to differentiate in the adult rat. The cellular target proved to be just as important as the electrical arrow.

There's still some opposition to parts of this scenario. Among some scientists, the prejudice against electrobiology remains so strong that one otherwise fine recent review doesn't even mention the NEJ or the difference between currents of injury in frogs versus salamanders!

Other objections are a little more substantial. A Purdue University group has measured electrical potentials near the surface of regenerating limbs underwater, using a vibrating probe. This is an electrode whose tip, ending in a tiny platinum ball, oscillates rapidly to and fro, giving the average voltage between the two ends of its motion. These researchers describe an arc of ion flow—they categorically deny the possibility of electron currents in living tissue—out from the stump and through the water or, in semiaquatic animals, a film of moisture on the skin. From the water, they suggest, these ions travel to the limb skin behind the amputation, then in through all the inner tissues, and finally out of the stump again to complete the circuit. They believe the epidermis drives these currents by its normal amphibian function of pumping sodium (positive) ions from the outside water into the body. They conceive of this ion flow as the regeneration current itself, because changing the concentration of sodium in the water directly affects their current measurements, and because certain sodium-blocking techniques have interfered with limb regrowth in about half of their experimental animals.

Of course, the Purdue researchers don't dispute the amply proven need for nerve and injury currents during regeneration. They've even confirmed Smith's induction of leg regrowth in frogs with batteries generating electron currents. Nevertheless, they consider nerves the target rather than the source of current, even though they propose no reason why their ion flow should be restricted to nerve tissue. In fact, they base their hypothesis partly on evidence that sodium flows even from denervated limbs.

There are several other problems with this theory: Its proponents' own measurements show that the sodium ion current almost disappears just when its supposed effect, blastema formation, is occurring. Moreover, it fails to explain the easily observed reversal of polarity in injury currents measured directly on the limb, as well as the crucial role of the NEJ. The proposed circuit goes right past the NEJ! Finally, it can't account for the several tests of semiconducting current throughout the nervous system, or regeneration in dry-skinned animals such as lizards.

To the lay person all this may seem like academic hairsplitting, until we reflect on the stakes: understanding regeneration well enough to restore it to ourselves. Certainly skin is electrically active. It's piezoelectric and pyroelectric (turning heat into electricity) as well as a transporter of ions in wet animals. In the last two decades nearly all tissues have been proven to produce or carry various kinds of electrical charge. Skin may play an as-yet-unknown role in regeneration besides its part in the NEJ, or it may merely be producing unrelated electrical effects. In any case, there are far too many data about the role of nerves to call skin the major source of the regeneration current. In fact, even the Purdue group has measured a stump current that's independent of sodium concentration.

A recent experiment by Meryl Rose gave further evidence of neural DC, without clearing up all aspects of the question. Rose removed the nerves From larval salamander legs before amputation. Normally such denervated larval stumps die right back to the body wall, but when Rose artificially supplied direct currents like those I measured in my first experiment, they regrew normally. This is pretty conclusive proof that the nerves are the electrical source in phase one. However, since it looks as though nerves also organize regeneration's second phase (see below), it's hard to understand how the new legs could have been completely normal when they were disconnected from the rest of the nervous system. Perhaps salamanders can pinch-hit for nerves at this stage through a tissue other than nerve. On the other hand, new nerves may simply have regrown into the limbs unbeknownst to Rose by the later stages of the experiment.

Phase two begins as the embryonal cells pile up and the blastema elongates. Early in this stage a sort of spatial memory becomes fixed in the blastema cells so the limb-to-be will have its proper orientation to the rest of the body. At the same time or shortly afterward, the cells at the inner edge of the blastema receive their new marching orders and platoon assignments. Then they redifferentiate and take their places in the new structure.

We can infer two things about the control for this part of the process. Since the blastema forms the right structure in relation to the whole organism, the guidance can't be purely local, but must come from a system that likewise pervades the whole body. Furthermore, there are no dedifferentiated cells left over when the work is done; there are just enough and no more. Thus there must be a feedback mechanism between the redifferentiation controls at the body side of the blastema and the NEJ's dedifferentiation stimulus at its outer edge.

A large body of earlier work has shown that the redifferentiation instructions are passed along a tissue arc whose main element is the circuit already established between nerves and epidermis in the first phase. The electrical component persuasively explains how this arc, an update of the morphogenetic field, may work. The direction (polarity) plus the magnitude and force (amperage and voltage) of current could serve as a vector system giving distinct values for every area of the body. The electric field surrounding continuously charged cells and diminishing with the distance from the nerve would provide a third coordinate, giving each cell a slightly different electrical potential. In addition, a magnetic field must exist around the current flow, possibly adding a fourth dimension to the system. Together these values might suffice to pinpoint any cell in the body. The electric and magnetic fields, varying as the current varies with the animal's state of consciousness and health, could move charged molecules wherever they were needed for control of growth or other processes. Since currents and electromagnetic fields affect the cell membrane's "choice" of what ions to absorb, reject, or expel, this system—in concert with the chemical code by which neighboring cells recognize each other—could precisely regulate the activities of every cell. It could express the exact point along the limb at which new growth must start; distinguish between right and left, top and bottom; even explain how totally missing parts, like extirpated bones or all the little bones of wrist and hand, can reappear.

Furthermore, the difference between electrical values at the inner and outer edges of the blastema would lessen as a new limb grew behind it. (Remember that the electrical potential grows increasingly negative toward the end of an intact limb.) The gradual convergence of these two values could constitute a feedback signal perfectly reflecting the number of dedifferentiated cells still needed. Although the results weren't entirely conclusive, perhaps because measurements had to be made under anesthesia, several experiments in the 1950s suggested that such a voltage differential governed restitution of the proper number of segments in earthworms. There was even a surge of positive potential that seemed to indicate when the job was finished.

This is a rich concept, and the details are without doubt more complex than this sketch, but they're all open to experimentation in a way that Weiss's morphogenetic field and Burr's L-field were not. The best part of this two-stage analysis is that it gives us a rationale for trying to foster regeneration after human injuries before we know all the details of the second phase.

The rat limb experiments strongly suggest that mammals lack two crucial requirements for the first phase of regeneration: They don't have the necessary ratio of nerve tissue to total limb tissue, the amount needed to make the dedifferentation stimulus srong enough; and they lack sufficient sensititive cells to respond to the electrical stimulus and form a big enough blastema. The work on rats pointed the way to defining the proper current, and the ability of electrically injected silver ions to dedifferentiate fibroblasts now gives us a possible method for producing an adequate blastema. We should now be able to supply the requirements for phase one in humans. Once this is done, the body itself can probably take care of phase two, even though we don't understand the process. Fingertip regrowth in children suggests that our bodies still have the ability to redifferentiate the cells and organize the missing part, as long as the electrical stimulus and the supply of sensitive cells are sufficient.

Microsurgeons have performed wonders in reimplanting cleanly severed portions of arms, legs, and fingers, but these limbs are subject to atrophy and obviously can't be grafted if they're too badly mangled or riddled with disease. As one who has performed too many amputations in his time, I find the prospect of being able to give a patient the real thing instead of a prosthesis tremendously exciting. There's a good chance that we'll eventually treat some nongenetic birth defects or old injuries by cutting off the defective part and inducing a normal one to grow. Perhaps, combined with gene splicing, such techniques could even rectify genetic birth defects.

Since no one has yet achieved full regeneration in rats or any other mammal, these dreams won't come true overnight. They aren't chimerical, however. The remaining problems could probably be solved in a decade or two of concerted basic research. Meanwhile, human capacities for repair of certain tissues are greater than most people realize, and there are already promising ways of enhancing some of them.

Cartilage

Fossils show that even the dinosaurs had arthritis, but unfortunately it outlived them. Many varieties have been described, all of which result in destruction of the hyaline (glassy) cartilage that lines the ends of the bones. The remaining cartilage cells try to heal the defect by proliferating and making more cartilage. They're almost never equal to the task, and scar tissue fills the rest of the hole. The result is pain, for scar tissue is too spongy to bear much weight or keep the bones from grinding against each other.

After our success at getting rat legs to partially regenerate, we studied this problem in 1973. We reasoned that, suite cartilage was made by only one kind of cell, getting it to regrow would be easier than working with a whole limb.

With orthopedic resident surgeon Bruce Baker of the Upstate Medical Center, we removed the cartilage layer from one side of the femur at the knee in a series of white rabbits. The operation left a circular hole of bare bone about 4 millimeters across. In the experimental animals we implanted silver-platinum couplings like those used on the rats, drilling the platinum end into the defect and tucking the rest along the bone. Most of the control animals filled in the defect with scar tissue along with some inferior fibrous cartilage. About a tenth of them grew a millimeter or two of good hyaline cartilage at the edge of the hole. But sure enough, the rabbits with the implants showed greatly enhanced repair. When we used an improved battery implant with silver wires at each end, we got even better results. Two of the rabbits healed the damage completely with beautiful hyaline cartilage just like the original material.

A few years later, when we were testing various electrode metals, we tried a different approach specifically for rheumatoid arthritis, in which runaway inflammation causes phagocytes to attack healthy cartilage cells. Gold salts taken orally sometimes control this disease but often produce toxic side effects. We figured electrical injection of pure gold directly into the joint with no other ions might work better. To find out, Joe Spadaro and I produced rheumatoid arthritis in the knees of both hind legs in forty rabbits, using a standard experimental procedure. Then we treated one knee in each animal with a positive gold electrode stuck right into the space between the two bones for two hours. Joe did the actual treatments. Then we sacrificed the animals gradually over a period of two months, and I examined both arthritic knees, not knowing until later which had been given gold. During the first two weeks about 70 percent of the treated knees were markedly better than the untreated ones. The improvement fell off to about 40 percent thereafter, suggesting that the treatment must be repeated for continued results.

Obviously, these were only preliminary experiments. However, since an estimated 31 million Americans suffer from arthritis, for which there is no cure yet, I think both avenues should be thoroughly explored as soon as possible.

Skull Bones

Lev Polezhaev has spent his career investigating what might be called the Polezhaev principle - the greater the damage, the better the regrowth. He has found he can often enhance repair by adding homogenates, minces, and extracts of the damaged organs, even though this doesn't augment the current of injury as his needling procedure did.

I

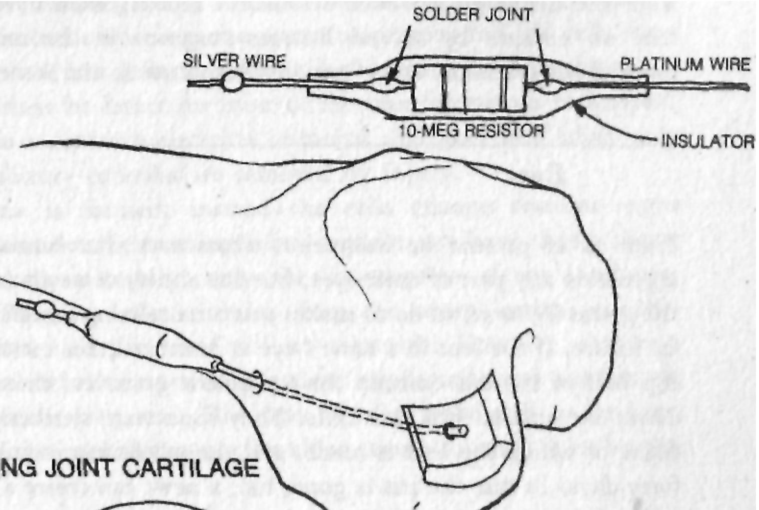
SOLDER JOINT
SILVER WIRE
PLATINUM WIRE
10-MEG RESISTOR
INSULATOR

REGENERATING JOINT CARTILAGE

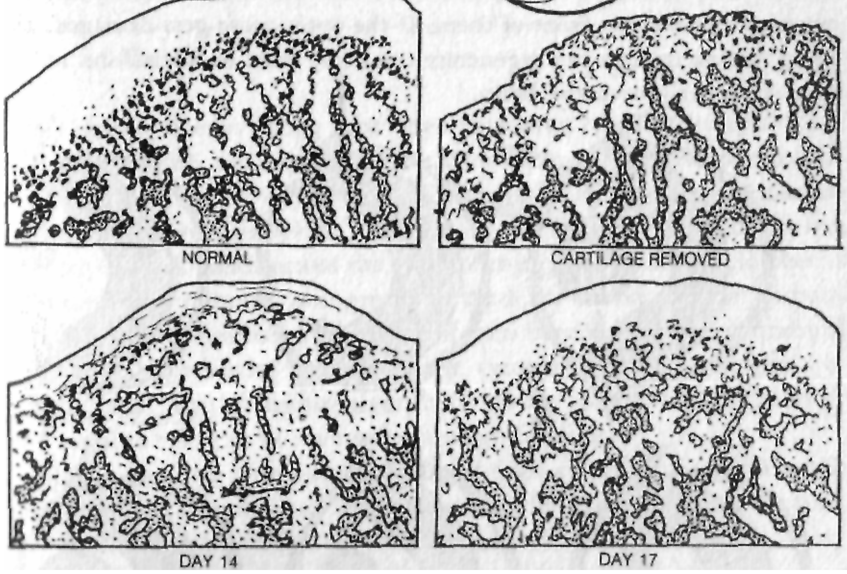
NORMAL
CARTILAGE REMOVED
DAY 14
DAY 17

REGENERATING JOINT CARTILAGE—RESULTS

Eventually Polezhaev developed a way to induce regeneration of holes in the skull, which normally heal over with scar tissue. As long as the dura mater, the cough membrane between skull and brain, is intact, a paste of blood and fresh (living) powdered bone will induce the bone cells at the edges to grow and bridge the gap. Microscopic studies have shown that the few live cells remaining in the paste don't survive and the bone particles themselves soon dissolve. Instead, some substance from the disintegrated bone stimulates repair. Since its first successful trial on humans by several Russian surgeons in the mid-1960s, this method has gradually come into increasing use in the Soviet Union.

Eyes

There is at present no indication whatsoever that humans could ever regenerate any part of their eyes, but the ability of newts (salamanders of the genus Triturus) to do so makes this a tantalizing research ideal for the far future. If the lens in a newt's eye is destroyed, the colored cells of the top half of the iris extrude their pigment granules, then transform by direct metaplasia into lens cells. They soon start synthesizing the clear fibers of which the lens is made, and the whole job is finished in about forty days. In case the iris is gone, too, a newt can create a new one from cells of the pigmented retina, and those cells can also transform into the neural retina layer in front of them. If the optic nerve gets damaged, the neural retina in turn can regenerate the nerve tract backward and reconnect it properly with the brain.

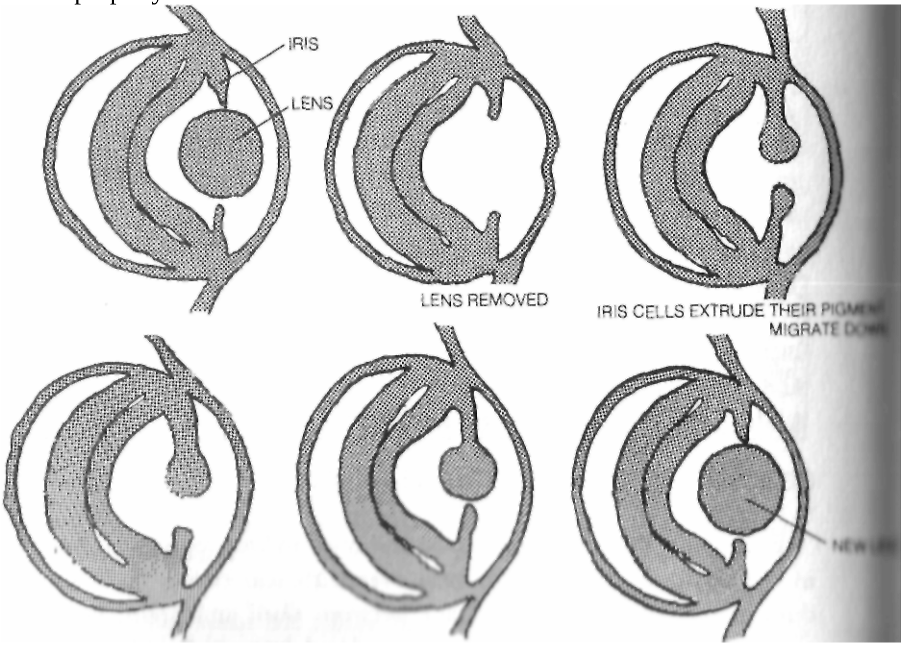

SALAMANDER EYE-LENS REGENERATION

No one knows why newts are so much more adept at this than all other creatures; their eyes have no obvious structural or biochemical peculiarities. Steve Smith gave us an important fact to work with when he found two proteins in the lens that seem to prevent the iris cells from changing into new lens cells as long as the old lens is in place. Since the neural retina must be intact for most of the transformations to happen, it may provide a constant electrical stimulus that goes into effect only when the inhibitory proteins are removed by injury.

No blastema is formed; instead the cells change costume right onstage. Furthermore, certain ingenious experiments have shown that a wound isn't really necessary, only the interruption of the inhibitory mechanism. Therefore, the stimulus from the neural retina probably isn't the familiar injury current of limb regrowth. However, despite a voluminous research literature on newt eye regeneration, no one has yet studied its electrical aspects. This may be why we're still so far from understanding the natural process, let alone trying to adapt it to human eyes.

Muscle

Every muscle fiber is a long tube filled with rows of cells (myocytes) laid end to end with no membranes between them—in effect, one multinucleated cell, called a syncytium. These nuclei direct the manufacture of contractile proteins, which are lined up side by side and visible, when stained, as dark bands across the array of myocytes. Each muscle fiber is surrounded by a sheath, and groups of them are bound together in bundles by thicker sheaths. At the edge of each bundle are long, cylindrical cells with huge nuclei and very little cytoplasm, called myoblasts or spindle cells. Also along the edges, between the spindle cells, clusters of tiny satellite cells can be seen at high magnifications.

After a crushing injury or loss of blood from a deep cut, muscle in the damaged area degenerates. The myocyte nuclei shrivel up and the cells die. Soon phagocytes enter to eat the old fibers and cell remnants. Only the empty sheaths and a few spindle and satellite cells are left.

Now these remaining cells turn into new myocytes, fill up the empty rubes, and begin secreting new contractile proteins. Although the early part of this process proceeds without nerves, it can run to completion only if motor nerve fibers reestablish contact with the terminals, called end plates, that remain at specific distances along each fiber sheath. If these end-plate areas are cut out, the nerve endings will enter, sniff around, and then retract. The muscle will atrophy. If the nerves reestablish the connections, new muscle cells will fill most of the original volume, gradually build up strength, and then completely differentiate into slow-twitch or fast-twitch fibers.

Attempts to enhance muscle regeneration in humans take two approaches. If available, a graft of a whole muscle from another source is the more effective. This is actual single-tissue regeneration, because the original muscle cells die and are replaced after new blood and nerve connections are made. Since its first clinical use in 1971, this approach has proven successful in replacing defective small muscles of the face and also in restoring anal sphincter control. Large limb muscles haven't been successfully transplanted yet.

Another method may soon be used in humans when grafts aren't possible. Muscle regeneration in birds and laboratory mammals has been considerably enhanced by inserting muscle tissue minced with fine scissors into pieces of no more than 1 cubic millimeter. Soviet biologist A. N. Studitsky first devised this method in the 1950s, extending Polezhaev's work, but its development has been slow.

Abdominal Organs

Despite over two hundred years of descriptive work, new regenerative capacities are still discovered in the animal kingdom from time to time. We've recently learned, for example, that adult frogs can restore their bile ducts, although for some reason females are better at it than males. Doctors have long known that the liver can replace most of the mass lost through injury by compensatory hypertrophy, in which its cells both enlarge and increase their rate of division so that the organ's chemical-processing duties can be maintained even though the ruined architecture isn't restored. Similarly, damage in one kidney is made up by enlargement of the other without rebuilding the intricate mazes of microtubules in the glomeruli. Recent studies of rat livers suggest that a combination of insulin and an epidermal growth factor, modified by at least a dozen other hormones, enzymes, and food metabolites, control the cell proliferation.

Now it appears that the spleen can make the same kind of comeback, at least in children. Adults who must have their splens removed rarely miss them, but children become more susceptible to meningitis. A few years ago, medical researcgers noted that children whose excised spleens had been damaged by accident were less likely to get meningitis than those whose spleens had been removed because of disease. Howard Pearson and his colleagues at the Yale University School of Medicine found that, when ruptured, the delicate spleen left bits of itself scattered in the abdomen, which grew and gradually resumed the organ's obscure blood-cleansing functions. Now when they remove a spleen, many surgeons wipe it on the peritoneum (the tough membrane that lines the abdominal cavity) to sow replacement seeds.

Another late discovery of regeneration was made in the late 1950s, when several scientists learned that tadpoles, larval salamanders, and sometimes adult salamanders could restore up to about four inches of their intestines. Moreover, all adult amphibians could reconnect the cut ends even if they couldn't replace a missing section. Allan Dumont, one of my best friends during medical school and now Jules Leonard Whitehill Professor of Surgery at NYU-Bellevue, decided to check this potential in mammals after I told him about my work on rat limbs. He wanted to find out whether regeneration could be stimulated in mammals to solve one of a general surgeon's most vexing problems—poor healing of sutured ends of gut after a cancerous or degenerated segment has been cut out. Even a small opening can spill feces into the abdominal cavity, with disastrous peritonitis the result.

Like any good scientist, Al started from the basics. After several years he'd confirmed the earlier reports. When he cut pieces from the gut of adult frogs and newts and merely put the ends close together in the abdomen, 40 percent of the animals survived by quickly reconnecting the two ends and completely healing them in about a month, although even the newts didn't replace much of the lost length in his experiments. Gut regeneration actually involves several tissues; Al's cell studies showed a blastema quickly forming at the junction and then differentiating into smooth muscle, mucus cells, and the structural cells of the villi.

Naturally, when I was organizing a conference on regeneration in 1979, I invited Al to present his results. About a month before the meeting, after the program had already gone to the printer, he wanted to change the title of his paper, for he'd just finished some surprising work. He asked me, "What would you expect to happen if I took some adult rats, cut out a centimeter of gut, and dropped the two loose ends back into the abdomen?" Like any first-year med student, I said they'd be dead of peritonitis in two or three days. Well, 20 percent of Al's rats had reconnected then bowels better than surgery could have done, and were alive and healthy. When Al had given one group of animals a temporary colostomy above the experimental cut, the survival rate jumped threefold. The perfect healing of the test animals compared with the controls indicated that sutures actually interfered with regeneration producing unnecessary scars and adhesions.

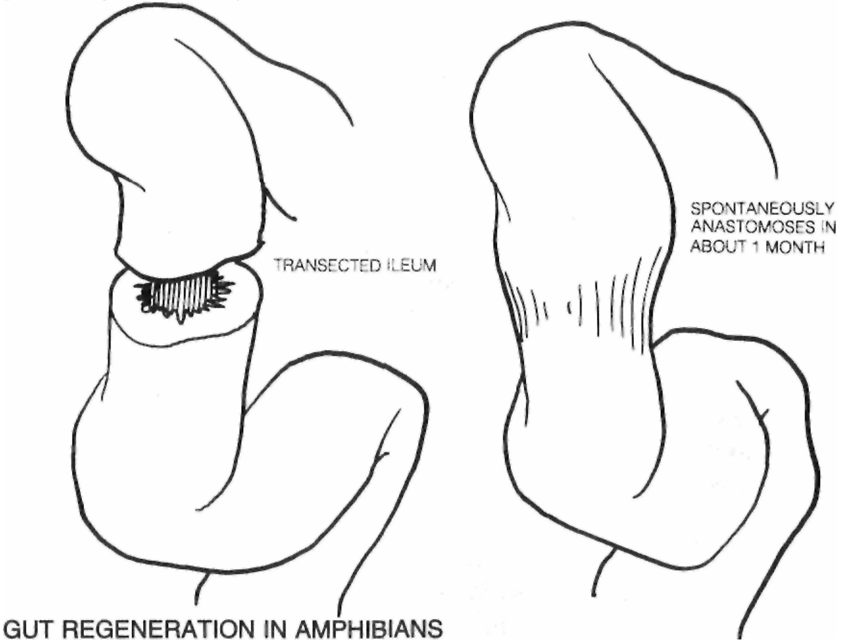
TRANSECTED ILEUM
SPONTANEOUSLY ANASTOMOSES IN ABOUT 1 MONTH

GUT REGENERATION IN AMPHIBIANS

No one knows for sure how the two cut ends find each other, but there's certainly some active search going on, for peritonitis sets in too quickly for the results to be due to chance. The process resembles a regrowing nerve fiber's search for its severed part, which may be conducted by electrical factors, a chemical recognition system, or both. Electrical potentials probably play the most important part, for recent research has found DC potentials at injuries on the peritoneum, and experimental changes in the peritoneum's normal bioelectric pattern attract the inner membrane enclosing the bowels, causing it to adhere to the site of the disturbance. Al has recently learned that, if the ends don't have to look for each other but instead are connected by a piece of Silastic tubing, rats can, like tadpoles, replace up to 3 centimeters of missing intestine. There's no reason to believe this technique couldn't be adapted to humans.

Even though we don't know enough yet to electrically stimulate intestinal healing, Al has proposed a preliminary test of regeneration in large mammals that could spare some patients a lifetime of misery. It's almost impossible to surgically rejoin the colon to the anus, and when sutures fail, the person ends up with a colostomy. Since the free end of the colon would be held near its proper position by the local anatomy, Al suggests replacing it without stitches and giving such animals a temporary colostomy upstream from the gap. If X rays later showed regrowth, the temporary colostomy would be closed, and the animal would have a continuous, healthy intestine. If even a few patients could be spared the indignity of living attached to a bag, the effort would be well worth the little—yet still nonexistent—funding required. Intensive study of the electrical details of gut healing would probably make surgery less devastating for many additional patients.

Exciting as the prospects in this survey are, they're by no means the only ones, or even the most spectacular, which are reserved for the following chapters. Many researchers are working to turn the breakthroughs of the last two decades to practical use. Even so, progress isn't nearly as fast as it could be, perhaps due to disbelief that such widespread self-repair is really possible for us. It's not only possible, it's nearly certain, given even a modest monetary push, for the "useful dispositions" foreseen by Spallanzani are within our reach.

Ten
The Lazarus Heart

Like Columbus, scientists sometimes stumble upon new continents when merely seeking a quicker trade route. Our research group had this good fortune in 1973.

We'd gone back to basics after learning how to dedifferentiate frog red blood cells and start regeneration in rat limbs. We decided to study nucleated red cells in a variety of creatures, hoping for leads toward better regrowth in mammals. Although their circulating erythrocytes have no nuclei, even mammals have young red cells, with nuclei, forming in the bone marrow. After severe bleeding, up to a fifth of those in the bloodstream may be immature nucleated types, as the marrow rushes them into service to make up for the loss. We surveyed the effect of direct current on red cells from fish, amphibians, reptiles, and birds. All of the cells responded, but in a different way for each species. We decided to have a more detailed look at the largest and hence most easily studied red blood cells available, those of our old friend Triturus viridescens, the common green newt.

A newt is so small that you can't just poke a needle into one of its veins and take a blood sample. The only practical way to get pure blood is to anesthetize the animal, slice open its chest, cut its heart in two, extract the blood with a pipette, and throw away the carcass.

As the phrase goes, we "harvested" blood from three newts each week by this method. One day, when Sharon Chapin had finished the chore, she asked me, "What would happen if I sewed these animals up?" I answered that, because their hearts had been destroyed, they would die within minutes, with or without sutures, from lack of oxygen to the brain, just as all our other amphibian blood donors had. We looked it up just to make sure. The standard works on regeneration all agreed that no animal's heart could repair major wounds. Unlike skeletal muscle, the cardiac variety had no satellite cells to serve as precursors for mature heart-muscle cells. In any case, the textbooks stated, the animal would die long before such repair could occur.

Next week Sharon put our three intended sacrifices in a bowl of water and with a straight face asked me if they looked healthy enough to use. I told her they looked fine. "Good!" she exclaimed. "These are the same three we used last week." Score one for the open mind!

Flabbergasted, I helped anesthetize and dissect this trio of miracles. Their hearts were perfectly normal, with no evidence of ever having been damaged in any way.

The Five-Alarm Blastema

Abruptly I changed my research plans. I asked Sharon to test a series of newts by cutting away large sections of their hearts and sewing up their chests, then killing some of the survivors every day and slicing, mounting, and staining the hearts for study under the microscope. Over 90 percent lived through the first operation, and several weeks later we had hundreds of slides ready for my examination and diagnosis. Unfortunately they all looked the same! Even those from the day after that horrendous mutilation showed only normal tissues with no sign of injury.

By now we knew we had come upon a first-class mystery and had better jettison our preconceptions. We reasoned that we could tell when regeneration was finished by finding out when the blood began flowing again. Under the microscope we could easily see blood cells streaming through capillaries in the transparent tail fins of lightly anesthetized newts. The motion stopped when we cut the heart, and restarted about four hours later. We sectioned a new series of hearts, this time covering the first six hours at intervals gradually increasing from fifteen minutes to one hour.

While waiting for the specimens, we rummaged more thoroughly through the literature for other reports on heart regeneration. There was evidence for very limited repair—but no true regeneration—of small heart wounds in a few animals. The process seemed limited to the very young. Even then, the results were of poor quality, combining a lot of scar tissue with only a little proliferation of nearby heart cells, but the mitotic component could be enhanced by various experimental aids.

In 1971, John O. and Jean C. Oberpriller, anatomists at the University of North Dakota School of Medicine in Grand Forks, reported that small wounds in salamander hearts healed this way but required two months. A year after that, the English edition of a book by Lev Polezhaev summarized several decades of Russian research, mainly on the hearts of frogs and lizards. Pavel Rumyantsev, now at the Cytology Institute of the USSR Academy of Sciences in Leningrad, had found in 1954 that newborn mammals (rats and kittens) could repair tiny puncture wounds, and recently he has proved the same capacity in the atria, or receiving chambers, of adult rat hearts. We even found a German report of 1914 claiming that human babies had sometimes regenerated small areas of their hearts damaged by diphtheria.

The Russians claimed some progress in extending this marginal native healing. In the late 1950s, N. P. Sinitsyn had repaired large holes (up to 16 square centimeters) in the hearts of dogs by covering the wounds with patches made of muscle sheath, canvas, suede, or other materials. Scar tissue still covered the outside, but the patch guided a thin layer of new muscle fibers forming along its inner surface. Using dogs whose wounds had already closed with scar tissue, Polezhaev then found he could induce heart muscle to fill in part of the gap by cutting away the scar and irritating the edges of the remaining cardiac muscle. Other Soviet researchers enhanced the muscle cell proliferation a little more with vitamins B1, B6, and B12, various drugs, extra RNA and DNA, and heart tissue extracts or minces.

Despite such goads, heart regrowth was limited to very small injuries or the border zone around larger ones, and it always took several weeks. No one had even imagined that half a heart could restore its other half, much less in a matter of hours. I could hardly wait until the next batch of slides was ready.

They showed us an unprecedented type of regeneration. Where the missing part of the heart had been, a blastema formed in about two and a half hours. We saw no evidence of dedifferentiation or mitosis in the remaining heart-muscle cells, and indeed it would have been impossible for the processes we'd already studied to make a blastema in such a short time. Instead, the mass of primitive cells arose dramatically from the blood.

As soon as the salamander heart is cut open, blood pools around the wound and clots quickly, usuallyin about one minute, sealing the hole like wet plaster. Almost immediately, the nearest red blood cells crack open like eggs. Their nuclei, surrounded by a thin coating of cytoplasm, glide by some means yet unknown directly to the raw, fraved edge of the heart muscle and insinuate themselves into the tangle of dying and injured cells. To a biologist this sight is bizarre, uncanny. It's as though the engine of a passing car could walk up to a stranded truck, climb under the hood, and drive it away.

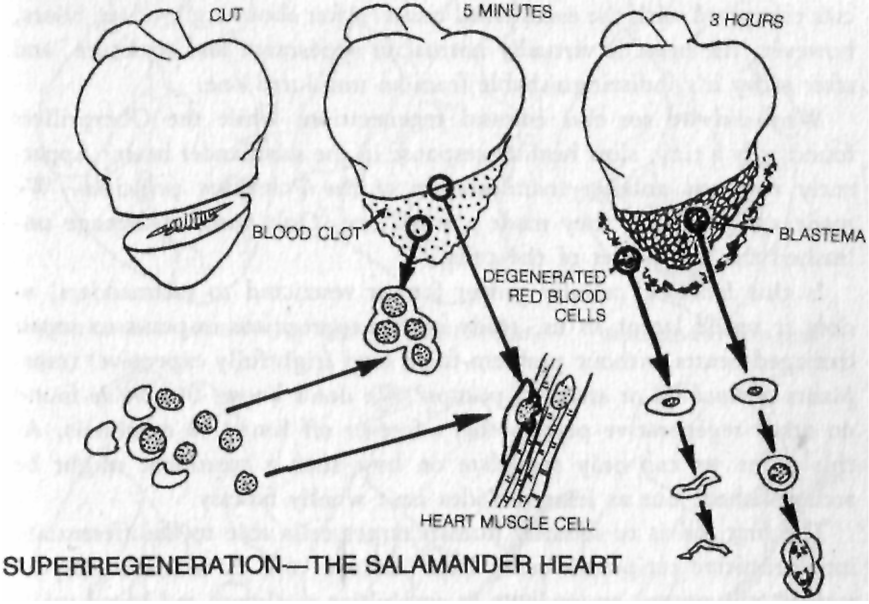

SUPERREGENERATION—THE SALAMANDER HEART

Farther away from the wound surface, the red cells also spill out their nuclei, but these cell yolks clump together, fusing their remaining cytoplasm to form a syncytium. Still farther away from the center of action, the red cells undergo the more leisurely dedifferentiation we observed in our frog fractures and DC culture studies. They turn into primitive ameboid cells that move toward the area of damage and attach themselves by pseudopods to the injured muscle fibers. In all of biology there's no precedent for these virtuosic cellular metamorphoses. In fact, they're so strange that most researchers have simply refused to believe in their existence or try the experiment for themselves.

All these changes are well under way within fifteen minutes. Soon afterward the extruded nuclei, the interconnected syncytial nuclei, and the ameboid cells are all dividing as fast as they can, building up the blastema. It's fully formed within three hours after the injury. By then its cells have already started to redifferentiate into new heart-muscle cells, synthesizing their orderly arrays of contractile fibers and connecting up with the intact tissue. If the clot contained more blood cells than were needed, the extras outside of the area now degenerate, apparently so as not to get in the way of the the repair work.

Meanwhile, the newt has survived by absorbing dissolved oxygen from the water through its skin. Now, at about the four-hour mark, there are enough new muscle cells to withstand contraction, and the heart begins pumping again, slowly. After five or six hours, most of the blastema cells have redifferentiated into muscle, which is still somewhat "lacy" or delicate compared with the established tissue. After about eight or ten hours, however, the heart is virtually normal in appearance and structure, and after a day it's indistinguishable from an uninjured one.

Why did we see this colossal regeneration, while the Oberprillers found only a tiny, slow healing response in the salamander heart? Apparently this was another manifestation of the Polezhaev principle. We made a big wound; they made a small one. Only massive damage unleashed the full power of the cells.

Is this fantastic cellular power forever restricted to salamanders, or does it reside latent in us, ready at the appropriate impetus to repair damaged hearts without problem-filled (and frightfully expensive) transplants of donated or artificial pumps? We don't know, but we've found no other regenerative process that's forever off limits to mammals. At this point we can only speculate on how such a treatment might be accomplished, but at least the idea isn't wholly fantasy.

The first job is to identify human target cells able to dedifferentiate into primitive totipotent cells. Bone marrow cells or immature erythrocytes, the nearest equivalents to amphibian nucleated red blood cells, are one obvious candidate population, especially since they seem to be the crucial cells in rat limb regeneration and the inner part of fracture healing. Fibroblasts despecialized by electrically injected silver ions might be used. Another possibility is lymphocytes, one class of infection-fighting white blood cells. In our lab we've demonstrated that they, too, can dedifferentiate in response to appropriate electrical stimuli.

Since newt-type heart regeneration doesn't occur naturally in mammals, we would probably have to grow a large mass of the target cells in tissue culture. Then, with the patient on a heart-lung machine, the surgeon could cut away scar tissue and otherwise freshen the wound if it wasn't recent enough, then apply enough of these ready-made pre-blastema cells to fill the defect. They would be held in place by a blood clot, sutured pericardium, or some type of patch. Then, assuming we'd learned the electrical parameters already, electrodes would induce nuclear extrusion, dedifferentiarion, consolidation with surrounding muscle, and the final transformation into normal cardiac muscle. The current would probably have to be adjusted throughout the process to get us various steps in synchrony, and vitamins or drugs might be used to enhance mitosis or protein synthesis. Once the scar had been removed, the instructions as to what cells were needed would come from the surrounding healthy heart muscle.

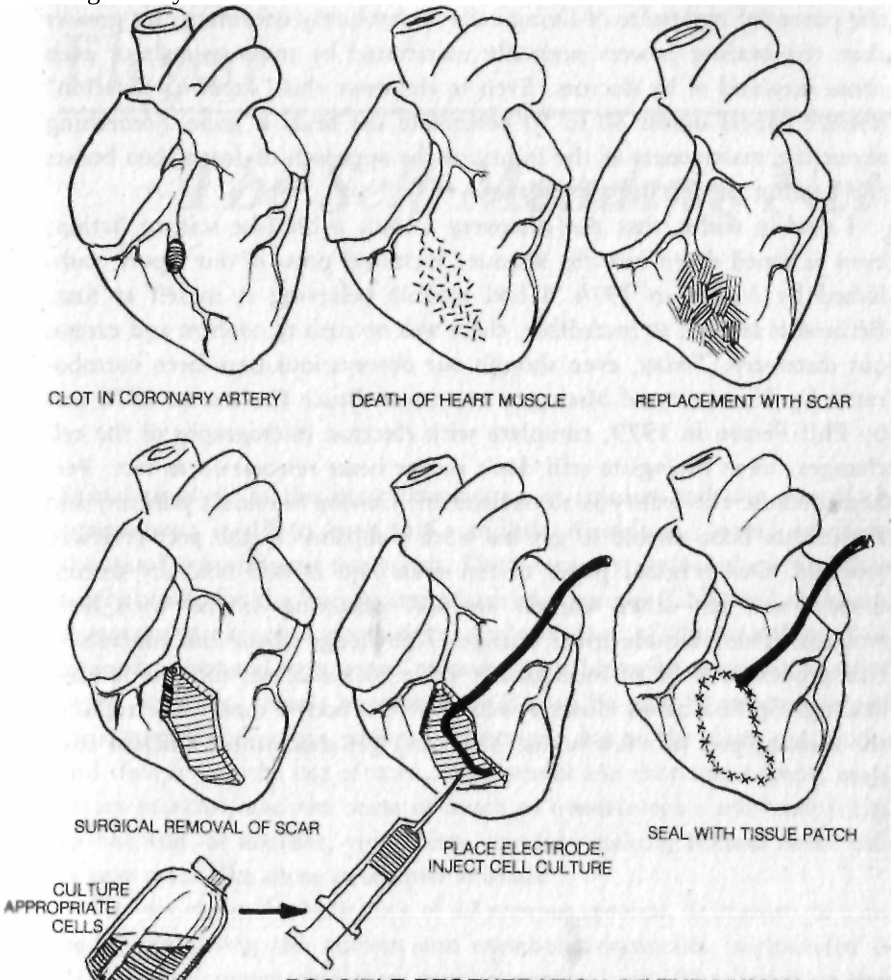
CLOT IN CORONARY ARTERY
DEATH OF HEART MUSCLE
REPLACEMENT WITH SCAR
SURGICAL REMOVAL OF SCAR
PLACE ELECTRODE, INJECT CELL CULTURE
SEAL WITH TISSUE PATCH
CULTURE APPROPRIATE CELLS

POSSIBLE REGENERATION OF THE HUMAN HEART

In the salamander this process takes about six hours. Since this exceeds the current limits of "machine time" for artificial circulation in humans, we would have to extend the capacities of heart-lung devices or else speed up the cellular processes. Obviously there's a long road of experiment to travel before we can be more specific about techniques. One of the things we must learn is whether the newt's electrocardiogram shuts down during repair. We must know how its presence or absence relates to the current of injury and other electrical factors in this novel method of blastema formation.

Personally, I'm sure we can get the human heart to mend itself. As a result of being confronted by this wonder in newts, I'm convinced that the potential repertoire of living cells is absolutely enormous, far greater than the healing powers normally manifested by most animals or even those dreamed of by doctors. Even in the newt this "superregeneration" doesn't appear unless 30 to 50 percent of the heart is gone. Something about the massiveness of the injury or the approach of death then boosts the healing process into overdrive.

I readily admit that the discovery sounds a bit like science fiction, even as toned down into the subdued technical prose of our report, published by Nature in 1974. I had trouble believing it myself at first. Because it seemed so incredible, there was no rush to confirm and extend our discovery. Today, even though our observations have been corroborated by University of Michigan anatomist Bruce Carlson in 1978 and by Phil Person in 1979, complete with electron micrographs of the cell changes, most biologists still don't accept heart restoration as fact. Perhaps because the reality is so outlandish, Carlson wouldn't publish, and Person has been unable to get his work published in the peer-reviewed journals. Our original paper of ten years ago is still officially unconfirmed, and the other workers are still puttering around with little wounds. This attitude must change. Knowledge about the controls of this process will be of incalculable value to medicine, for this is ideal healing. Spilled blood closes a wound at the body's center and replaces the missing part in a few hours. You can't get much more efficient than that.

Eleven
The Self-Mending Net

Spinal paralysis is the most devastating of injuries and also one of the commonest; it afflicts over half a million Americans, including fifteen thousand new sufferers every year. Until recently their outlook was absolutely bleak, for the human central nervous system (CNS) had no known regenerative capacity whatsoever. Only if part of the spinal cord remained unsevered was some recovery possible with physical therapy. Now, however, there's hope that we'll soon be able to coax nerve cells into reestablishing the proper connections across the damaged section and thus return the use of arms, legs, sexual and excretory organs, respiratory muscles, and the sense of touch to quadriplegics and paraplegics. In one way or another, this dream involves making human nerve cells behave more like those in simpler animals.

The neuron is the basic unit of all nervous systems. It consists of a cell body, containing the nucleus and metabolic organelles, surrounded by dozens of filaments that carry messages in and out. The incoming dendrites predominate in sensory neurons. There's usually only one motor fiber, or axon, which carries the neuron's outgoing messages to dendrites of other neurons or to receptors on muscle or gland cells. An axon, often several feet long, is the principal fiber of a motor neuron, which relays orders from the brain or spinal cord to the tissues and organs.

All neuron cell bodies reside in the brain and spinal cord. Only their axons and dendrites extend outward, forming the peripheral nerves that connect every part of the body with the CNS. Other fibers connect certain sensory and motor neurons within the spinal cord, creating reflex arcs, like those that jerk our hands from hot stows without our having to send the impulse all the way to the brain for instructions. Still other fibers connect spinal neurons with those in the brain, and in the brain itself the interconnections reach such a density that each nerve cell may hook up with as many as twenty-five thousand others.

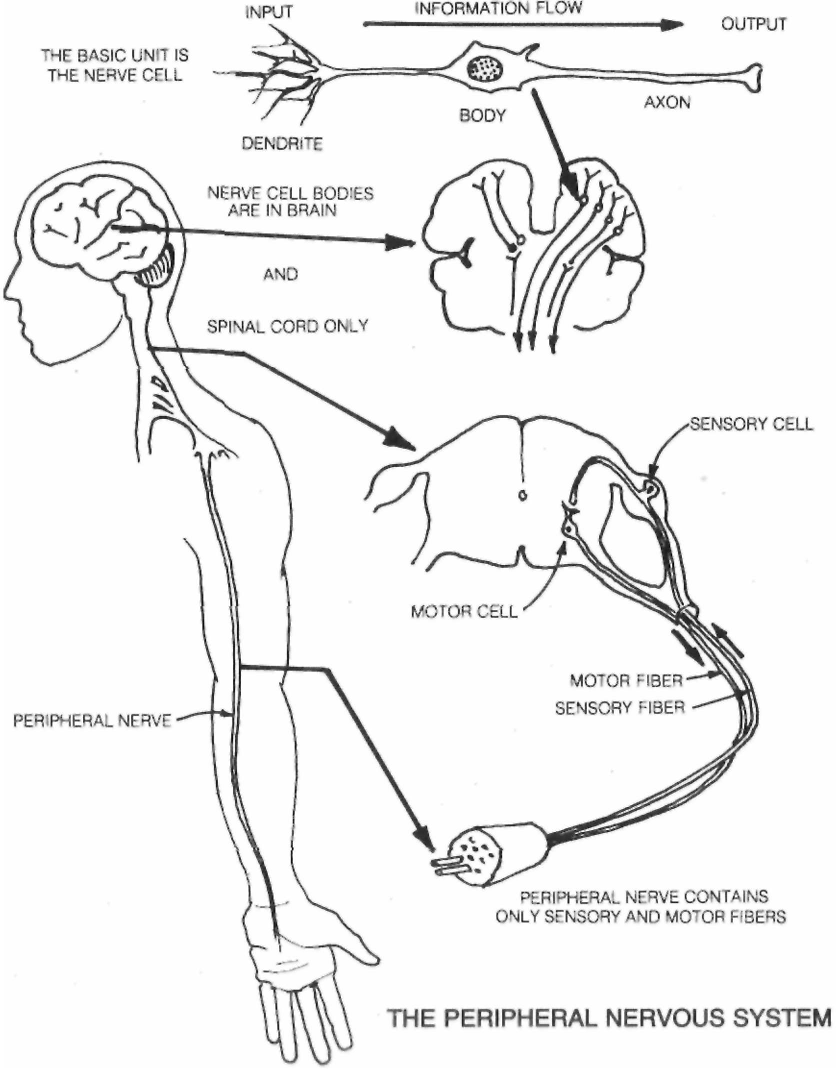
INPUT
INFORMATION FLOW
OUTPUT
THE BASIC UNIT IS THE NERVE CELL
DENDRITE
BODY
AXON
NERVE CELL BODIES ARE IN BRAIN
AND
SPINAL CORD ONLY
SENSORY CELL
MOTOR CELL
MOTOR FIBER
SENSORY FIBER
PERIPHERAL NERVE
PERIPHERAL NERVE CONTAINS ONLY SENSORY AND MOTOR FIBERS

THE PERIPHERAL NERVOUS SYSTEM

Except for a few specialized components like the naked fiber tips that enter into neuroepidermal junctions, all parts of every neuron are swaddled in various types of perineural cells. In the brain there are several kinds, collectively called the glia, in which the neurons are embedded like hairy raisins in a pudding. The cell bodies in the cord also are surrounded by glial cells, but their axons and dendrites, which include the fibers of the peripheral nerves, are surrounded by Schwann cells. These form tubes, made up of spiraling layers of membrane rich in a fatty substance called myelin, around some of the largest fibers. A third type, ependymal cells, line the four cavities within the brain, or ventricles, and the narrow central canal that runs the length of the spinal cord. These cells are close relatives, having all developed from the same part of the ectoderm, or outer cell layer, that formed the primitive neural tube in the embryo. The nervous system actually consists of several times more perineural cells than neurons.

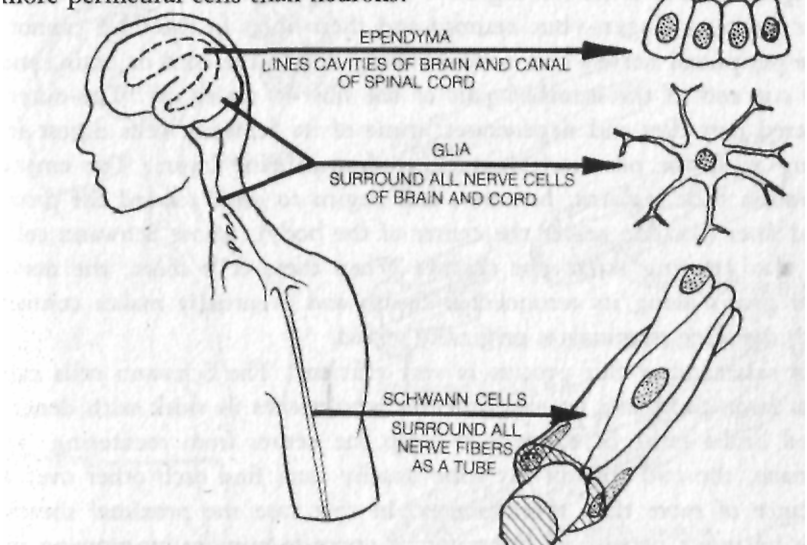
EPENDYMA
LINES CAVITIES OF BRAIN AND CANAL OF SPINAL CORD
GLIA
SURROUND ALL NERVE CELLS OF BRAIN AND CORD
SCHWANN CELLS
SURROUND ALL NERVE FIBERS AS A TUBE

THE PERINEURAL SYSTEM

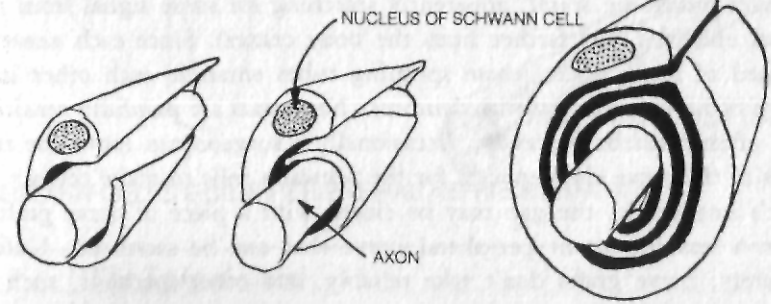
NUCLEUS OF SCHWANN CELL
AXON

SCHWANN CELL SHEATH AROUND A PERIPHERAL NERVE FIBER

Until recently the perineural cells were considered merely a "packing tissue," whose only job was to insulate and support the neurons. We've now learned that they play a major role in getting nutrients to the neurons. They also help control the diffusion of ions through the nerve cell membrane and hence regulate the speed of impulse firing, even to the point of inhibiting seizures, the random spread of impulses in the brain. They may also have an important part in memory, and they probably conduct the direct currents so important to regeneration. They're essential to healing wherever it occurs in nerve tissue.

Peripheral Nerves

Peripheral nerve fibers can regrow—otherwise we'd lose sensation whenever we cut a finger—but neurons and their fibers in the CNS cannot. The peripheral nerve's cell body survives, safe in the cord or brain, and the cut end of the attached part of the fiber is sealed off. The outer, severed part dies and degenerates; some of its Schwann cells digest it, along with the now-useless myelinated membrane layers. The empty Schwann tube remains, however, and begins to grow toward the proximal fiber (the one nearer the center of the body), whose Schwann cells are also growing across the chasm. When these cells meet, the nerve fiber grows along its reconnected sheath and eventually makes contact with the same terminals it originally served.

In salamanders this process is very efficient. The Schwann cells can cross large gaps, and an experimenter who wishes to work with denervated limbs must be diligent to keep the nerves from reentering. In humans, the two ends of the tube usually can't find each other over a distance of more than a centimeter. In that case the proximal sheath with its intact nerve fiber hunts for its opposite number by growing in an ever increasing spiral, apparently searching for some signal from the distal end (the part farther from the body center). Since each nerve is formed of many fibers, these spiraling tubes entangle each other in a lump of nerve tissue called a neuroma. Neuromas are painfully sensitive and often must be cut away. Occasionally a surgeon can move the two ends of the nerve close enough for the Schwann cells to make contact. If that's impossible, the gap may be closed with a piece of nerve grafted from a less important peripheral nerve that can be sacrificed. Unfortunately, nerve grafts don't take reliably, and other methods, such as making artificial channels with tiny plastic tubules, are still in the experimental stage.

We don't know why the salamander's peripheral nerve regrowth is so much more effective than ours, but I surmise that its more efficient DC electrical system accounts for the difference. If the locator signal is electrical, it should be possible to augment it in humans so as to grow nerve fibers over longer distances. Beginning with a 1974 report from David

H. Wilson of the Leeds General Infirmary in England, there have been some interesting claims that pulsed electromagnetic fields have speeded recovery of limb function in rats after peripheral nerve damage, but the effect hasn't yet been substantiated for humans. If these findings hold up, we may soon be able to boost nerves past their 1-centimeter limit, even if the action is indirect, and a thorough investigation of the electrical basics could drive nerve regrowth to even greater lengths.

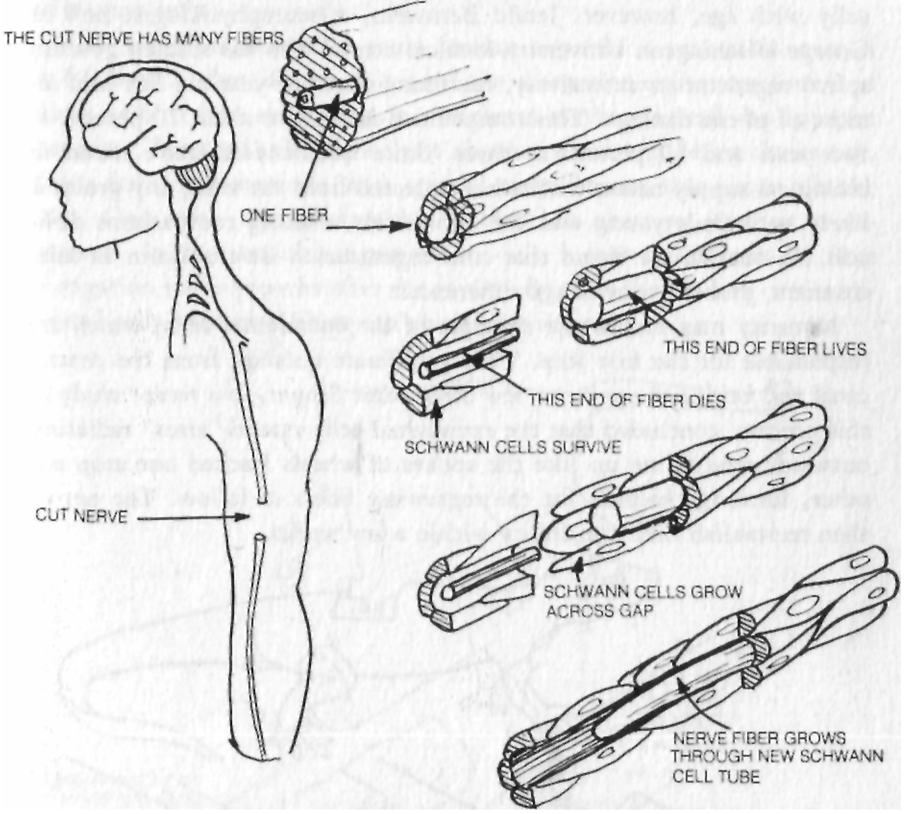
THE CUT NERVE HAS MANY FIBERS
ONE FIBER
THIS END OF FIBER LIVES
THIS END OF FIBER DIES
SCHWANN CELLS SURVIVE
CUT NERVE
SCHWANN CELLS GROW ACROSS GAP
NERVE FIBER GROWS THROUGH NEW SCHWANN CELL TUBE

SCHWANN CELLS GUIDE PERIPHERAL NERVE REGROWTH

The Spinal Cord

A sad and crucial difference separates peripheral fibers from those in the human spinal cord, for the latter don't reconnect over even a fraction of a centimeter. However, in most injuries relatively few of the neurons themselves are killed. It's important to realize that most of the cord cells below the injury don't die. The reflex arcs remain intact. In fact, reflexes are stronger than normal, because the neurons are now disconnected from the regulating influence of the brain. For the same reason, the broken bones of paraplegics heal in half the normal time, whereas a bone will heal very slowly or not at all if its peripheral nerve supply has been cut. Only the communication between brain and spine is silenced in paraplegia, and that makes all the difference.

Spinal fibers do reconnect in some animals, notably goldfish and, as you might expect, salamanders. Their ability seems to decline dramatically with age, however. Jerald Bernstein, a neurophysiologist now at George Washington University Medical School who has studied goldfish spinal regeneration extensively, has found that one-year-old fish heal almost all of the damage. This competence declines to about 70 percent at two years and 50 percent at three. Since salamanders aren't raised in biological supply houses but rather collected from the wild, any group is likely to include young and old individuals, making comparisons difficult. In our lab we found that cord regeneration isn't uniform in salamanders, probably due to age differences.

Maturity may reduce the response of the ependymal cells, which are responsible for the first step. They proliferate outward from the central canal and bridge the gap in a few days. Marc Singer, in a recent study of this process, concluded that the ependymal cells extend "arms" radiating outward, which line up like the spokes of wheels stacked one atop another, forming channels for the regrowing fibers to follow. The nerves then reestablish their continuity within a few weeks.

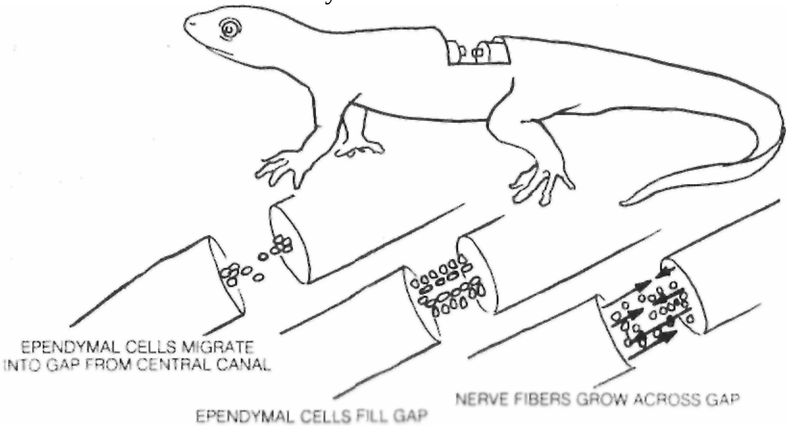
EPENDYMAL CELLS MIGRATE INTO GAP FROM CENTRAL CANAL
EPENDYMAL CELLS FILL GAP
NERVE FIBERS GROW ACROSS GAP

SALAMANDER SPINAL-CORD REGENERATION

Bernstein also found that there's a critical period during which regrowth must be completed or it will fail. After cutting the cords of goldlfish, he inserted Teflonon spacers to block regeneration. The normal processes took place, but of course the cells couldn't penetrate the divider. After the cellular activity had died down, Bernstein removed the barriers, but there was no further change. However, when he then cut off each damaged end, producing an even larger gap and reinjuring the cord, the cells started from scratch and healed the defect completely. Thus there's good reason to believe that even long-standing spinal injuries can potentially be regenerated if we can extend the basic capabilities of human cells.

One would expect to see some healing response in mammals, even if it fell short. After all, we only need the elongation and reattachment of fibers, which does take place in peripheral nerves. Instead the opposite happens. The cord cells die a short distance above and below the injury. Cysts form near the ends, and, instead of ependyma, scar tissue fills the gap. Only after this destruction is there an abortive attempt at regrowth. In humans this amounts to only a few millimeters of fiber elongation many months after the injury. By then it's too late; the ependymal cells and nerve fibers can't penetrate the scar.

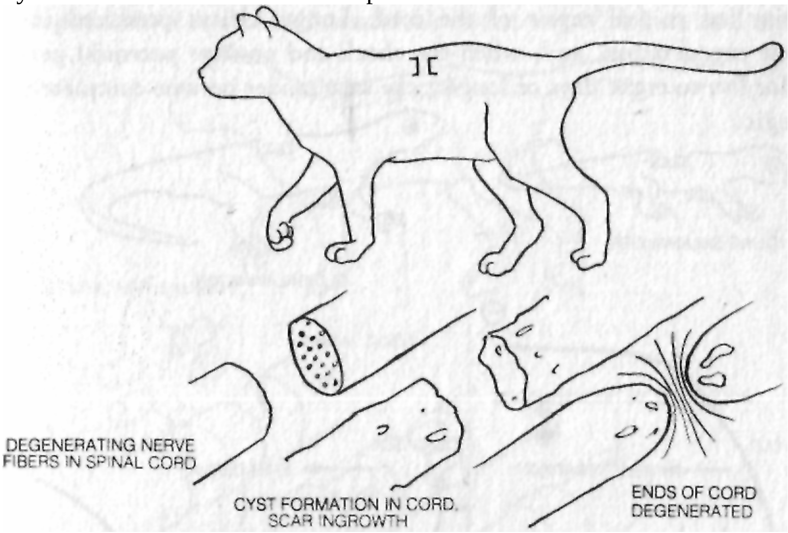
DEGENERATING NERVE FIBERS IN SPINAL CORD
CYST FORMATION IN CORD, SCAR INGROWTH
ENDS OF CORD DEGENERATED

CYSTS AND SCARS PREVENT CORD REGROWTH IN MAMMALS

Why the difference between salamanders and mammals? The reason may lie in the cord's immediate response. In all animals the injury instantly results in spinal shock, during which all neuronal activity is profoundly depressed, especially in the part of the cord still connected to the brain. Even the simplest reflexes disappear. As the shock wears off, the cord below the injury becomes hyperactive. Its reflexes become tremendously exaggerated and lead to spastic paralysis of the muscles. The interesting difference is in the duration of shock. In young salamanders and goldfish it lasts only a few minutes, but it may endure for over a week in old ones. In mammals it takes even longer to wear off—as long as six months in humans.

We made some electrical measurements on salamander and frog spines in our lab. The injured area turned out to be strongly positive during spinal shock, even though all direct-current flow ceased in the entire cord and in the peripheral nerves arising from the part below the trauma. Then, as the shock resolved, a steadily increasing negative potential appeared, its size reflecting the amount of outgrowth by ependyma and nerve fibers. We found that we'd only rediscovered these potentials, however. G. N. Sorokhtin and Y. B. Temper had made the same measurements at the Khabarovsk Medical Institute twenty years before. The patterns of shock and polarity both correlated, not only with the cell activity, but with the end result of regenerative success or failure. A few minutes of shock and a correspondingly short period of positivity led to full repair of the cord. Longer delays produced incomplete regeneration, and, when the shock and positive potential persisted for five to eight days or longer, the salamander became completely paraplegic.

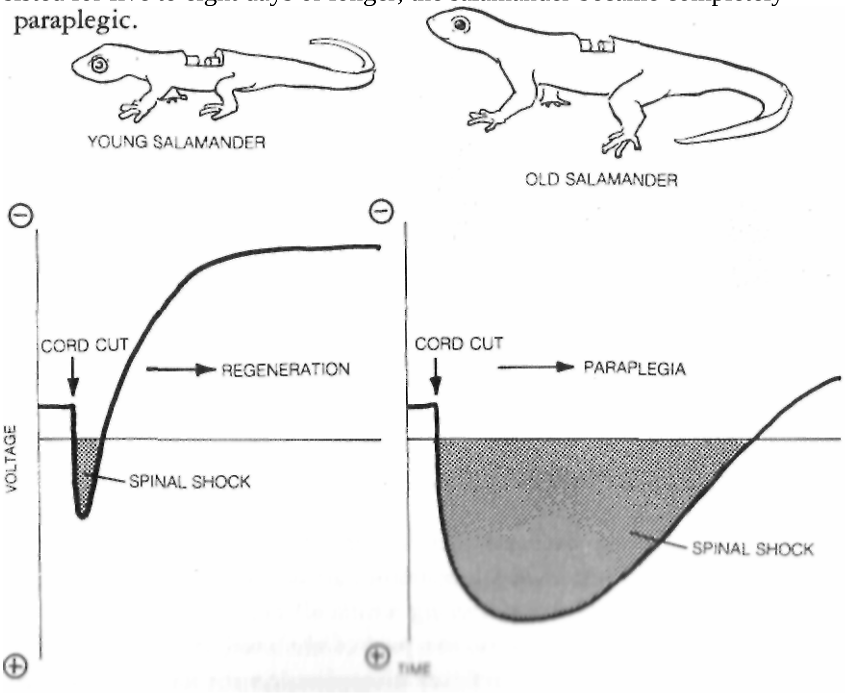

SPINAL SHOCK AND AGE INHIBIT CORD REGROWTH

As far as I know, the only electrical measurements of spinal shock in mammals were preliminary ones done at our laboratory in conjunction with Carl Kao of the VA hospital in Washington, D.C. We tested the severed cord ends in cats for twenty-four hours and found only an increasing positive potential. The situation seemed quite similar to the electrical difference between salamander and frog limbs. As in most instances, positive potentials appeared to inhibit constructive cellular activity while negative ones fostered it.

An experiment Kao did several years ago provided some supporting evidence. Kao made two cuts through the spinal cord in each of several cats, producing a central fragment about 5 millimeters long, separated from each end. He then grafted pieces of sciatic nerve as spacers in the two cuts. Typical degeneration with cysts occurred in each end of the cord but not in the isolated piece. In fact, this part showed some growth of its ependyma and nerve fibers. The small piece was probably isolated from the positive potentials produced in the rest of the cord. Hence it escaped inhibition and grew. It seems the prolonged electrical positivity of spinal shock is the main roadblock in the way of human cord repair.

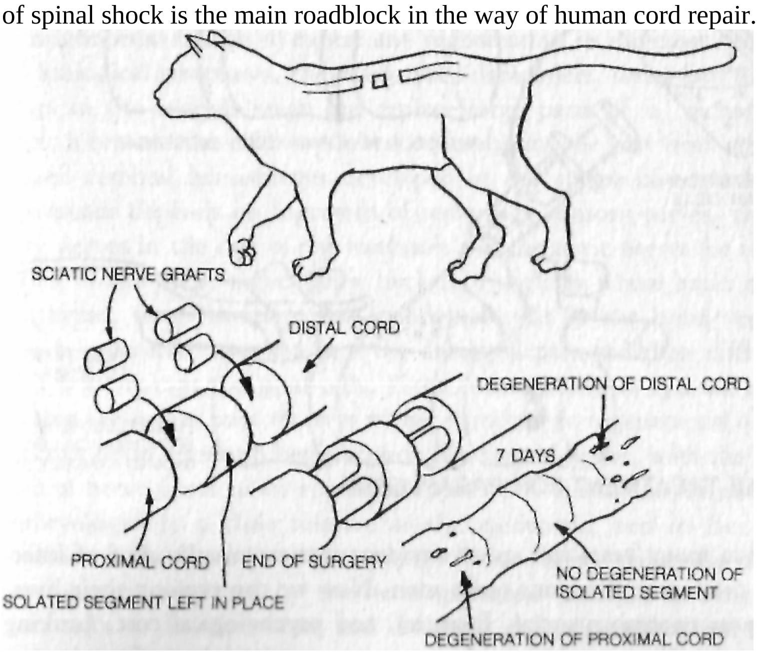
SCIATIC NERVE GRAFTS
DISTAL CORD
DEGENERATION OF DISTAL CORD
7 DAYS
PROXIMAL CORD
END OF SURGERY
NO DEGENERATION OF ISOLATED SEGMENT
SOLATED SEGMENT LEFT IN PLACE
DEGENERATION OF PROXIMAL CORD

KAO'S EXPERIMENT

It should be possible to cancel that polarity and replace it with a growth-stimulating negative one, using a properly shaped electrode. Older injuries in which spinal shock has subsided might require a different input of current, as well as surgical removal of scar and cysts. The electrode material would have to be chosen carefully, for some metals are toxic to nerve cells. Also, humans have a low ratio of ependyma compared to spinal neurons, so we might have to add more. However, it should be relatively easy to culture more ependymal cells from a sample of the patient's own, and then inject them when we put in the electrode.

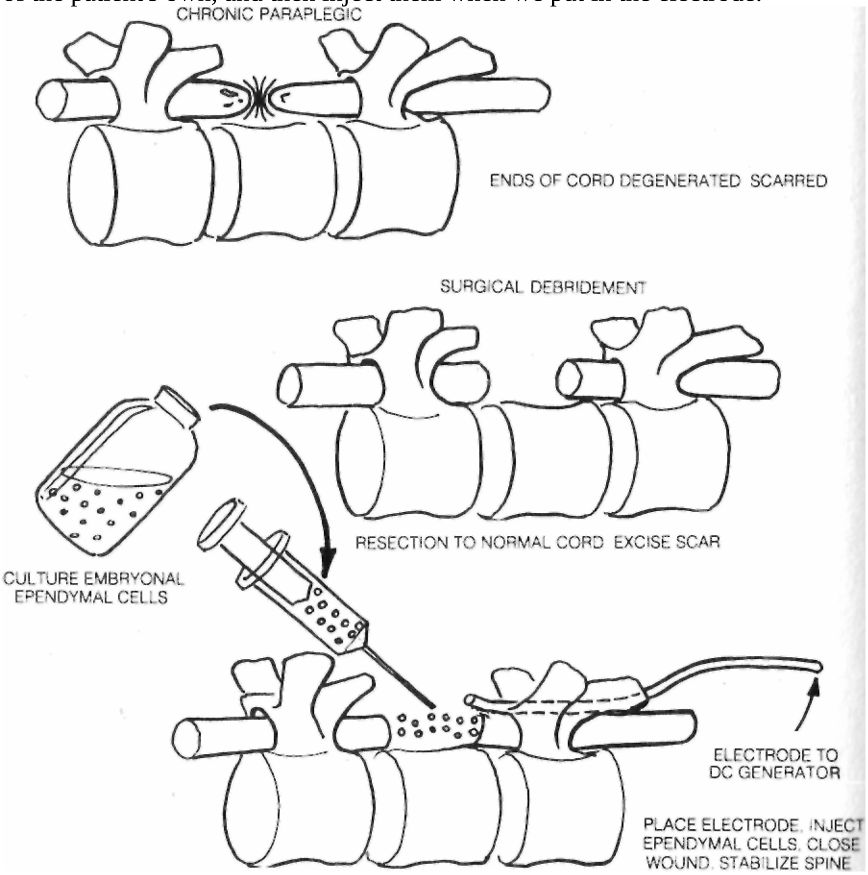
CHRONIC PARAPLEGIC
ENDS OF CORD DEGENERATED SCARRED
SURGICAL DEBRIDEMENT
RESECTION TO NORMAL CORD EXCISE SCAR
CULTURE EMBRYONAL EPENDYMAL CELLS
ELECTRODE TO DC GENERATOR
PLACE ELECTRODE. INJECT EPENDYMAL CELLS. CLOSE WOUND. STABILIZE SPINE

POSSIBLE TREATMENT FOR PARAPLEGIA

Not too many years ago spinal accident victims usually died of infections or other complications quite soon. Now we can prolong their lives, but only at enormous social, financial, and psychological cost. Looking ahead, as in the case of heart damage, we now have hope for releasing regeneration in humans. Actually, the outlook for spinal regrowth is more promising. The cellular processes are more familiar, and there are a few groups, like the American Paralysis Association, that sponsor research more imaginatively than the government agencies. Thus the electrical problems in spinal healing may be tackled sooner than in other fields.

The public imagination has been captured by the computerized muscle-stimulation techniques being developed by Jerrold Petrofsky, an engineer at Wright State University in Dayton. The nationally televised sight of his patient Nan Davis and other paraplegics taking tentative steps and pedaling tricycles with their own muscle power was tremendously exciting. But if we can get the body to do the same things by itself, that will be even better. Any amount of regeneration would only make other techniques more effective. Even restoring 10 percent of lost function would be an unimaginable blessing to those who are now helpless. I feel the electrical manipulation of spinal shock must be tested vigorously now, for this is perhaps the one area where the barriers of tragedy are closest to being broken.

The Brain

It might seem foolish to expect any regeneration in the most complex of all biological structures, the brain, yet salamanders, some fish, and most frogs in the tadpole stage can replace large parts of it, including the optic lobes and the olfactory lobes, or forebrain, the part from which our prized cerebral hemispheres developed in the course of evolution. Replacement depends on ingrowth of remaining sensory nerves, the olfactory nerves in the case of the forebrain and the optic nerves for the optic lobes. When these nerves grow back into the area where brain has been destroyed, they stimulate the ependymal cells in the brain ventricles, which proliferate outward into the damaged part and then differentiate into new neurons and glial cells. If the animal's nose or eyes are removed so that the injury zone receives no nerve input, no regeneration occurs.

Thus brain regrowth begins much like that of limbs, with the connection of nerve fibers to an epithelial tissue. The ependyma, remember, is embryologically a close relative of the epidermis, and in fact can be considered the central nervous system's "inner skin." Since the electrical environment produced by the neuroepidermal junction is what stimulates cells to dedifferentiate and divide in the salamander limb stump, and since we started limb regeneration in the rat by crudely mimicking this signal, it seems likely that a similar stratagem could induce brain regeneration in animals normally lacking this ability.

A form of shock, called the spreading depression of Leao after its discoverer, neurologist A. A. P. Leao, occurs after brain injuries. Starting at the site of damage, it extends in all directions until the entire cortex becomes electrically positive and all its neurons shut down. Leao studied it only in response to small injuries, when it persists for a few hours. We don't know whether it occurs in the salamander or how long it lasts after major damage to the mammalian brain. Concerted study of Leao's depression combined with experiments in electrically stimulating the ependymal cells could open the way to self-repair of the human brain.

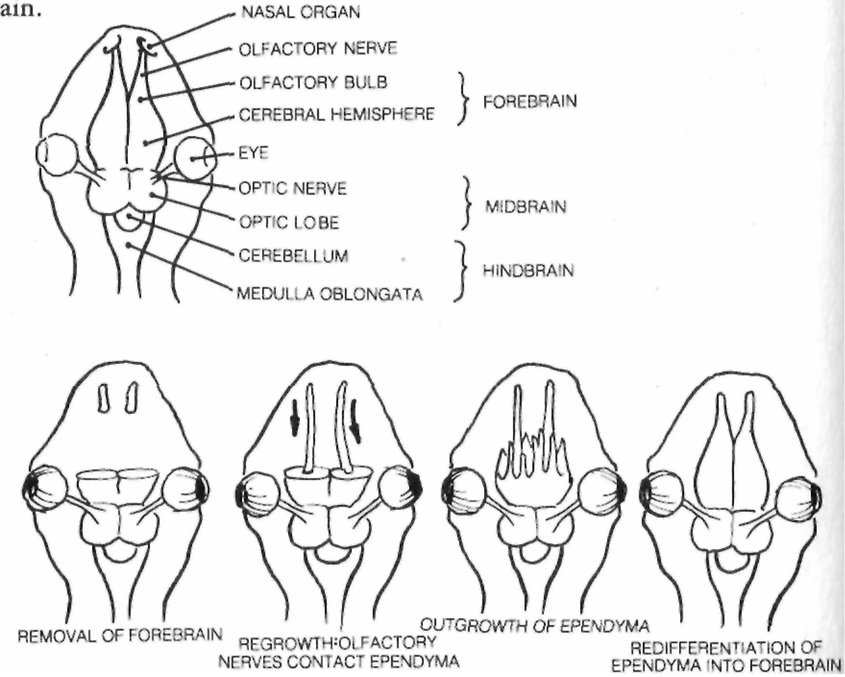
NASAL ORGAN
OLFACTORY NERVE
OLFACTORY BULB
CEREBRAL HEMISPHERE
FOREBRAIN
EYE
OPTIC NERVE
OPTIC LOBE
MIDBRAIN
CEREBELLUM
MEDULLA OBLONGATA
HINDBRAIN
REMOVAL OF FOREBRAIN
REGROWTH: OLFACTORY NERVES CONTACT EPENDYMA
OUTGROWTH OF EPENDYMA
REDIFFERENTIATION OF EPENDYMA INTO FOREBRAIN

AMPHIBIANS CAN REGENERATE LARGE PARTS OF THE BRAIN

Recovery from stroke and head wounds taught us long ago that the brain has a great deal of plasticity; that is, it can reorganize so that undamaged regions take over tasks formerly done by the lost cells. Supplementation of this ability with even a small amount of regeneration might make recovery nearly complete for many brain-damaged people. For the first time in history, neurologists can hope to progress from describing the brain and cord to mending them. As Geoffrey Raisman of London's Laboratory of Neurobiology recently reminded his colleagues: ". . . no immutable natural laws have been discovered that forever rule out repair of the nervous system."

Twelve

Righting a Wrong Turn

Good and evil often sprout from the same tree, in the body as in Eden. Nothing illustrates this paradox better than cancer. Today, because of breakthroughs in genetics, thousands of scientists are searching for oncogenes, bits of DNA that are presumed to pull the trigger that fires the malignant bullet. It has been known for a long time, of course, that cancer isn't inherited through egg and sperm the way hemophilia is. However, many have postulated that the immediate cause of cancer may be genetic changes in somatic cells. Normally suppressed genes held in an unnoticed corner of our genetic bookshelves since long ago in our evolution might be dusted off only when other bodily conditions are "just wrong." While the premise of this idea is apparently true, biologists have recently concluded that the difference between a normal gene producing a normal protein and one that could theoretically cause cancer is a single "typographical error" in a whole chapter of amino acid sequences. Such mistakes happen so often that we would all be riddled with cancer from infancy if that were all it took to start the disease. Something else must go awry before a few misspellings can turn the whole library into gibberish.

Three basic criteria by which a doctor diagnoses cancer must serve as the starting point in solving the mystery of its cause. First of all, the disease always always arises not from an alien germ but from a formerly normal cell of the host's body, and the cancer cells are more primitive than their healthy precursors. Moreover, this atavism reflects the seriousness of the disease: The simpler the cells, the faster they grow and the harder they are to treat, whereas a tumor that still resembles its tissue of origin is less malignant.

The second criterion is growth rate. Cancer cells multiply wildly, in contrast to the slow, carefully controlled mitosis of normal cells. Going hand in hand with this uncontrolled proliferation is a similar lack of control in the structural arrangement of the cells. Their membranes don't line up in the normal, specific ways, and they form a jumbled mass instead of useful architecture. As a further result of runaway multiplication, cancer doesn't observe the "boundary laws" of normal tissue. Instead it encroaches imperialistically upon its neighbors. In addition, since the cells don't adhere in any kind of structure, some of them are constantly breaking off, flowing through the blood and lymph, and setting up colonies—metastases—throughout the body.

The third basic criterion of cancer is metabolic priority. The diseased tissue greedily takes first choice of all nutrients circulating in the blood; the healthy part of the body gets what's left over. As the tumors disseminate and grow, they consume all available food, and the host wastes away and dies.

We can make one crucial observation at this point: Except for the lack of control, all three characteristics—cell simplicity, mitotic speed, and metabolic priority—are hallmarks of two normal conditions, embryonic growth and regeneration.

When considering the similarities between an embryo and a tumor, it's important to keep in mind one difference. Even though contained within the body of its mother, the embryo is a complete organism, and the controls over its cells are primarily its own, not those of an adult. Over thirty years ago in Switzerland, G. Andres probed this relationship by implanting frog embryos in various body tissues of adult frogs. Whenever the host didn't simply reject the graft, the embryo degenerated into a highly malignant metastasizing tumor. As a result, Andres proposed a theory of cancer that remains provocative today: A normal cell becomes cancerous by dedifferentiation. This change is not dangerous per se, according to Andres, but, because it occurs in a postfetal animal, the controls that would normally hold these neo-embryonic cells in check aren't working.

Cancer's relationship to regeneration is even more interesting. In the latter, a rapid growth of primitive cells having metabolic priority occurs in an adult, but with proper control as in an embryo Those animals that regenerate best are least susceptible to cancer. In general, as complexity increases up the evolutionary ladder to humans, regeneration decreases and cancer becomes more common. Although salamanders stand about midway in degree of complexity, they're perhaps the least specialized of all land vertebrates. They have tremendous regenerative abilities and almost no cancer. Even to give them tumors in the laboratory requires much effort. Adult frogs, on the other hand, have bodies that are much more specialized for their amphibious way of life; they regenerate very little and are subject to several kinds of cancer.

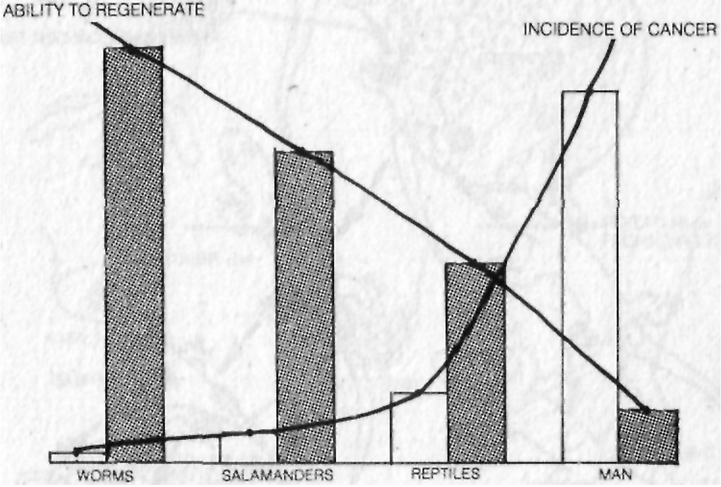

AS REGENERATION DECREASES, CANCER INCREASES

In 1948 Meryl Rose decided to see whether the environment of a salamander's regenerating limb could control the primitive cells of cancer as well as those of the blastema. He took pieces of a type of kidney tumor common in frogs and transplanted them to the limbs of salamanders. These tumors took better than most, and soon killed the animals when allowed to spread unchecked. However, when Rose amputated the leg just below or through the malignancy, normal regeneration followed, and the cancer cells dedifferentiated more fully as the blastema formed. Then as the new leg grew, the former frog tumor cells redifferentiated along with the blastema. The frog cells were easily distinguished from salamander cells by their smaller nuclei, and microscopic study showed frog muscle mixed in with salamander muscle, frog cartilage cells amid salamander cartilage, and so on.

This monumentally imporrant experiment proved Rose's hypothesis that regeneration's guidance system could control cancer, too. It implied that cancer cells weren't special but merely embryonic cells in a post-embryonic body. Rose's work led directly to Andres's theory a few years later.

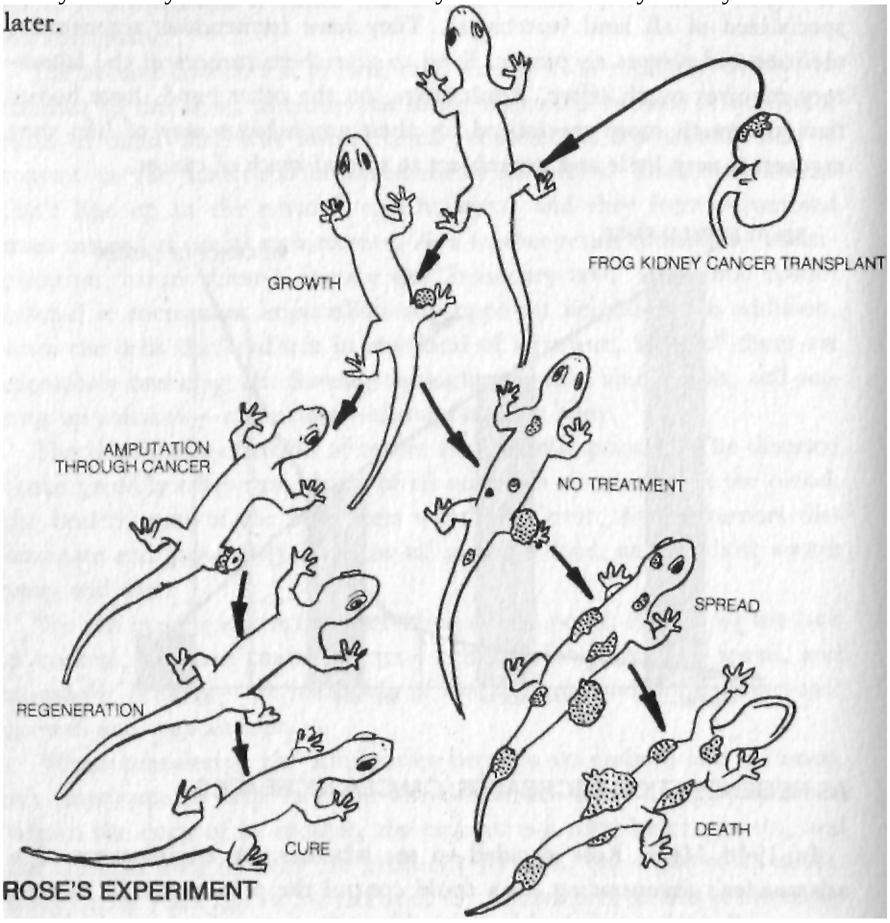

ROSE'S EXPERIMENT

Unfortunately, biology was still firmly gripped by anti-dedifferentiationism, and these ideas were partly ridiculed, partly ignored. The reaction held back cancer research for decades, because the dogma implied that carcinogenesis, like differentiation, was irreversible—once a cancer cell, always a cancer cell. As long as this view was sacred, the only possible way to cure cancer was to cut it out or kill it with drugs and X rays. We've been beating that dead horse for fifty years now with tragically modest increases in survival rates. Surgery works only against tumors that haven't yet spread. Chemotherapy depends on differences between malignant and healthy cells. However, the differences aren't as great as we would like, because cancer arises, not in some distant swamp, but from a slight change in our own cells. Therefore chemotherapy and X rays inevitably produce some damage in normal cells, too. Doctors, being only human and sharing in their patients' pain and terror, have devised ever more ferocious treatments. In our war on cancer we've stampeded ourselves into a sort of Vietnam syndrome: To destroy our foes, we're killing our friends. As Walt Kelly's cartoon character Pogo observed in that context, "We have met the enemy and he is us."

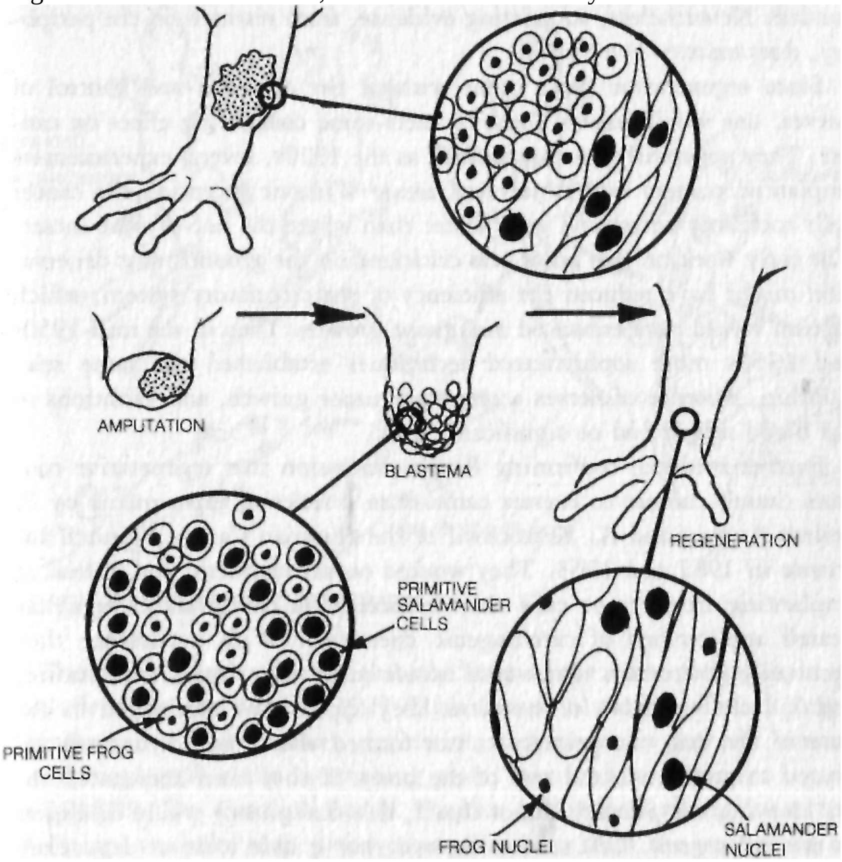
AMPUTATION
BLASTEMA
REGENERATION
PRIMITIVE SALAMANDER CELLS
PRIMITIVE FROG CELLS
FROG NUCLEI
SALAMANDER NUCLEI

CANCER CELLS RETURNED TO NORMAL BY REGENERATION

A Reintegrative Approach

But surely, you may be thinking, if this theory of cancer had any potential for a cure, the research establishment would have considered it. And surely there would be some supporting evidence. Unfortunately, even though the detooling and retooling of cells have now been accepted by all of biology, the old habits still persist throughout most of the granting hierarchy. A few years ago, for example, I met a young research fellow at the National Cancer Institute who wanted to study the regeneration-cancer link. He even showed me his proposal, an excellent one. I told him he was asking for trouble if he submitted it to NCI, but he said his boss approved and he was sure he'd get the grant. A month later he was forced out of the institute, and the project has never been funded. Nevertheless, supporting evidence, from research on the periphery, does exist.

Since regeneration can't occur without the stimulus and control of nerves, one would expect them to exert some controlling effect on cancer. They apparently do. As far back as the 1920s, several experimenters implanted tumors into denervated areas. Without exception the cancer cells took root better and grew faster than where the nerves were intact. The early work on this point was criticized on the grounds that denervation might have reduced the efficiency of the circulatory system, which in turn would have enhanced malignant growth. Then in the mid-1950s and 1960s more sophisticated techniques established the same relationship. Absence of nerves accelerated tumor growth, and variations in the blood supply had no significant effect.

Further evidence confirming Rose's conclusion that regenerative controls caused tumors to regress came from a series of experiments by F. Seilern-Aspang and K. Kratochwil of the Austrian Cancer Research Institute in 1962 and 1963. They worked on salamanders, but instead of implanting frog tumor cells they induced skin cancer with large, repeated applications of carcinogenic chemicals. With persistence they eventually got tumors that would invade subsurface tissues, metastasize, and kill the animals. In one series they applied the carcinogen to the base of the tail; the primary tumor formed there, and metastases appeared at random in the rest of the body. If they then amputated the tail, leaving the primary tumor intact, this malignancy would disappear as the tail regrew. Cell studies showed that it didn't die or degenerate but apparently reverted to normal skin. Furthermore, all the secondary tumors vanished, too, as though they were being operated by remote control from the main one. The salamander ended up with a new tail and no cancer. However, if the primary tumor was at a distant point on the body, amputation of the tail had no effect. Even though the tail regenerated, the main cancer and its offshoots all progressed, and the animal died.

This research, combined with Rose's, indicates that regeneration near a primary tumor can make it regress by reverting to its normal tissue type. I doubt that there's anything special about legs or tails; I would predict that regrowth in any part of the body, as long as it was near the primary tumor, would have the same effect. The key to regression appears to be a change in the malignancy's immediate neighborhood. The electrical currents in nerve and particularly in the neuroepidermal junction seem likely candidates, since they suffice to start regeneration in animals normally incapable of it.

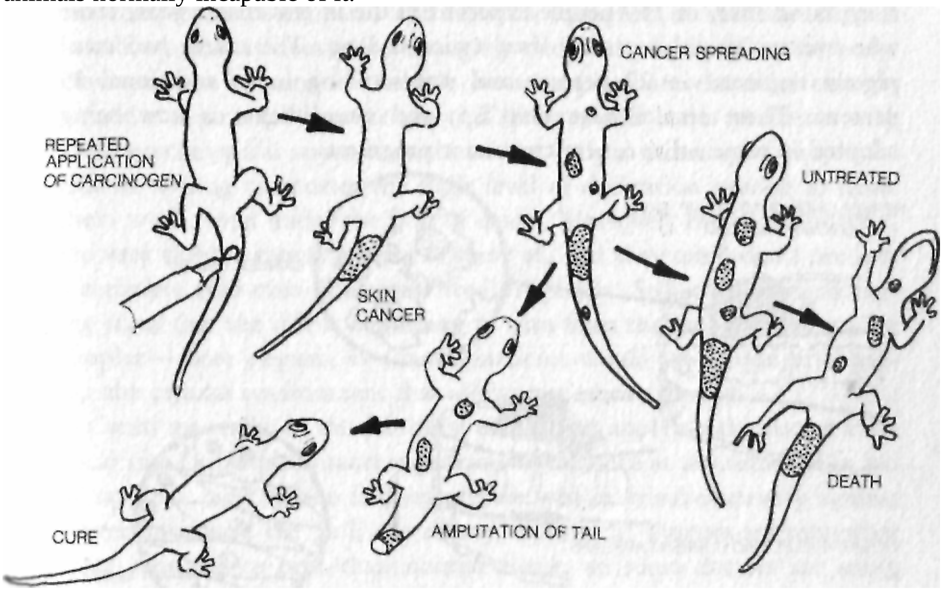

ROSE'S EXPERIMENT CONFIRMED

There's abundant evidence that the state of the entire nervous system can affect cancer. Back in 1927 Elida Evans, a student of Carl Jung, documented a link between depression and cancer in a study almost totally neglected in the intervening years. In a long-term project begun in 1946 by Dr. Caroline Bedell Thomas at Johns Hopkins School of Medicine, students were given personality tests, and the occurrence of disease among them was charted over several decades. In this and later studies, a high risk of developing cancer has been correlated with a specific psychological profile that includes a poor relationship with parents, self-pity, self-deprecation, passivity, a compulsive need to please, and above all an inability to rise from depres'sion after some traumatic event such as the death of a loved one or loss of a job. In such a person, cancer typically follows the loss in a year or two.

Several physicians have found they can greatly increase cancer patients' chances of a cure with biofeedback, meditation, hypnosis, or visualization techniques. Several years ago O. Carl Simonton, an oncologist, and

Stephanie Matthews-Simonton, a psychologist, began using all these methods, with emphasis on having patients develop a clear picture of the cancer and their body's response to it. For example, a patient might spend a meditative period each day imagining white blood cells as knights on white horses defeating an army of black-caped marauders. When the Simontons tabulated their first results in "terminal" cases, they found that, of 159 people expected to die in less than a year, those who eventually did succumb lived twice as long. The cancer had completely regressed in 22 percent and was receding in an additional 19 percent. These results have held up, and visualization is now being adopted in some other cancer-treatment programs.

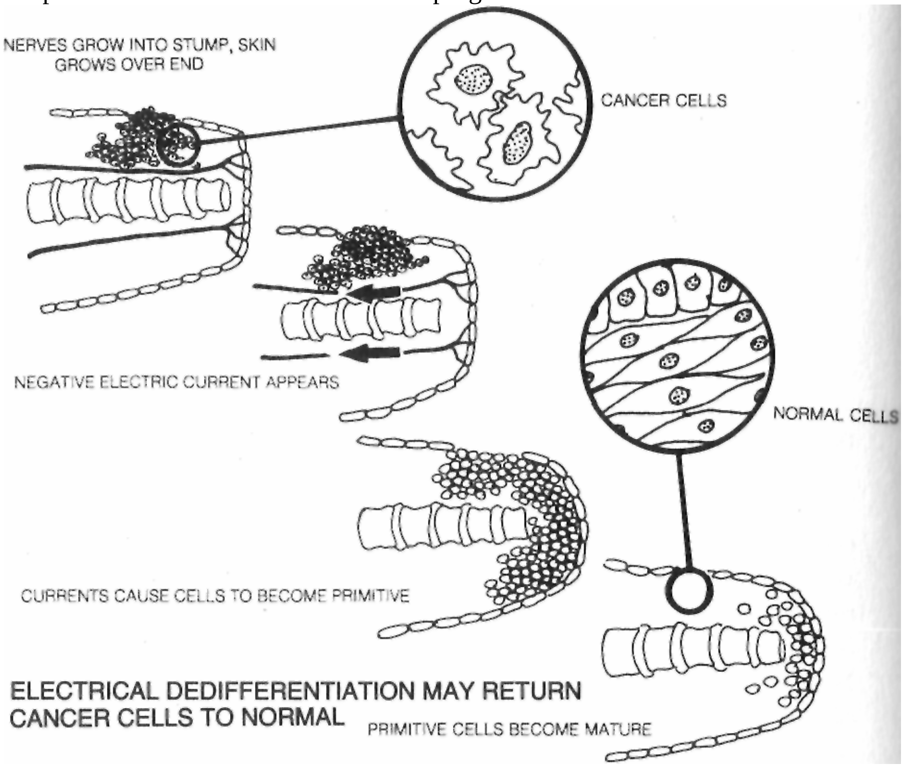
NERVES GROW INTO STUMP, SKIN GROWS OVER END
CANCER CELLS
NEGATIVE ELECTRIC CURRENT APPEARS
NORMAL CELLS
CURRENTS CAUSE CELLS TO BECOME PRIMITIVE
PRIMITIVE CELLS BECOME MATURE

ELECTRICAL DEDIFFERENTIATION MAY RETURN CANCER CELLS TO NORMAL

Penn State psychologist Howard Hall, testing hypnosis for boosting white blood cell activity, found a 40 percent increase in cell counts among his younger, more responsive subjects just one week after the trance session. New Haven surgeon Bernie Siegel has developed the Simontons' methods further with therapy groups called Exceptional Cancer Patients. Having patients make drawings to reveal their true psychological state sans defenses, then working toward a comprehensive change in outlook (total CNS response) to mobilize the will to live, Siegel has helped his patients greatly improve the quality and quantity of their lives, as compared to clinical prognoses. This approach also enhances chemotherapy's effectiveness while minimizing its side effects, and it dramatically increases the likelihood of a "miracle"—total regression and cure of the cancer.

All this really shouldn't be so surprising. Under hypnosis the mind can completely block pain, and research described in the next chapter has shown that it does so by changing electrical potentials in the body. How can we be sure it couldn't create appropriate electrical changes around a tumor and melt it away? There are still major problems with these psychological approaches, however. Only a minority of people are able or willing to muster the high level of dedication needed to make them work, even under the gun of death. Moreover, they require time, the very thing a cancer patient is short of, and they often don't produce a complete cure even when practiced diligently. Still, they're encouraging signs that the tide is beginning to turn from the warfare mode to the simpler—more elegant, as a mathematician would say—ideal of changing the cellular environment that allows a tumor to flourish.

Could we really do the job more directly by applying the proper electrical input to aim a surefire regenerative influence at the tumor? I'm sad to say that most of the few researchers who've tried electricity against cancer have used the "kill the enemy" approach. Tumors are somewhat more sensitive to heat than normal tissue, so some doctors are using directed beams of microwaves to cook them without, it's hoped, destroying too many healthy cells. FDA approval for general use of this method is expected soon. It has been known since the time of Burr and Lund that growing tissue is electrically negative, and cancer is the most negative of all. Hence some researchers have tried to inhibit tumor growth by canceling the offending polarity with positive current. Early reports were encouraging, but we now know that toxic metallic ions are released from most positive electrodes, so this method must be tested with great caution.

Only one research team has sought a reintegrative effect of electric current. In the late 1950s Carroll E. Humphrey and E. H. Seal of the Applied Physics Laboratory at Johns Hopkins tried pulsed direct currents on standardized fast-growing skin tumors in mice. Even though they used both positive and negative polarities, their results seemed sensational. In one series they got total remission in 60 percent of the test animals after only three weeks; all the control mice had died by then. In another series the control tumors averaged seven times the size of the ones treated with current. Unfortunately, present evidence doesn't really support this approach either. In my group's experiments with human fibrosarcoma cells in vitro, negative and positive currents both speeded up growth by over 300 percent. On the other hand, as mentioned in Chapter 8, we found that we could suspend mitosis in the fibrosarcoma cells with silver ions injected by minute levels of positive current. During one day of exposure, the cells appeared to dedifferentiate completely, and they stopped dividing for a month without additional treatment, even though we changed the nutrient medium regularly. Obviously, this entire subject needs to be investigated more thoroughly.

Some researchers believe pulsed electromagnetic fields may be of some benefit in treating cancer. Art Pilla, working with Larry Norton and Laurie Tansman of New York's Mount Sinai School of Medicine, as well as William Riegelson of the Medical College of Virginia, claims to have found a pulse sequence that significantly increases the survival time of mice with cancer. So far, these experimenters say they've increased the effectiveness of chemotherapy in lab animals with PEMF, but haven't found a pulse pattern that consistently regresses tumors in vivo, although Pilla and Steve Smith have been able to transform malignant lymphoma (lymph node cancer) cells into benign fibroblasts in culture.

The claim that PEMF may retard cancer in animals is seriously flawed, however. In these experiments the entire animal was exposed to the field, not just the part with the cancer. The pulsing field (like almost any time-varying magnetic field) induces a stress response in the animal (see Chapter 15). For a short while this increases the activity of the immune system, which slows the growth of the tumor. However, the field's effect on the tumor itself is to speed it up, and, in the long run, added stress is the last thing an animal with cancer needs. These experiments cannot be used to indicate the safety or benefit of PEMF in regard to cancer. Since to heal bones the fields are directed only at small regions, PEMF as used on humans does not produce stress, increased immune system response, or any concomitant antitumor activity.

Certain leads, such as the electrically injected silver, remain promising. In the 1950s and 1960s, Dr. Kenneth MacLean published some interesting work on the use of magnetic fields versus tumors in mice. He believed he'd healed several cases of cancer with steady-state magnetic fields, and certain unorthodox healers in America and India who use permanent magnets have made similar claims. The difference in effect between steady-state and time-varying fields (see Chapters 14 and 15) leads me to theorize that a steady-state magnetic field, if strong enough, may indeed halt mitosis in malignant cells.

Due to the prevailing outlook in cancer research, the key work remains to be done. Even promising nonelectromagnetic approaches have been victims of bias. There's sound evidence, for example, that megadoses of vitamin C do slow tumor growth and increase the chances for a complete cure, but Linus Pauling hasn't been able to persuade any of the powerful institutes to perform a large-scale trial. Some animal tests are now being funded, but, since vitamin C is nontoxic, immediate clinical experiments on large numbers of humans would make much more sense.

In following up the regeneration connection, two experiments in particular are crying out to be tried. Someone must attempt to duplicate the electrical environment of regeneration around tumors in lab animals, using electrodes. This would involve introducing small negative currents to thoroughly test the hypothesis that cancer cells are stuck in a state of incomplete dedifferentiation. The idea would be to dedifferentiate them the rest of the way and then let normal processes in the body turn them into healthy mature cells. The same hypothesis should be tested another way by surgically creating neuroepidermal junctions near the tumors.

Will these experiments be done soon? I wish I knew. The multibillion-dollar cancer research bureaucracy could certainly afford them, but, although there are a few signs of change, the establishment is stuck in the near-primitive state of the war mentality. I've maintained for many years that we won't learn much more about abnormal growth until we learn more about the normal kind. That approach can lead to cancer treatments that are truly compatible with our bodies, far safer and more effective than the simplistic, dangerous ones now in vogue.

Part 4
The Essence of Life

And if the body were not the soul, what is the soul?
—Walt Whitman

A hallucination is merely a reality that we normally don't have to bother with.
—Stella Denova

Thirteen

The Missing Chapter

Medical students often experience a profound, ego-wrenching shock at the midpoint of their training, as the focus shifts from classroom to bedside. My experience was typical, since this division was even sharper in the 1940s than it is today. After two years' study of the scientific underpinnings of medicine, my classmates and I thought we were pretty smart. During long days and longer nights of lectures, notes, labs, books, reviews, papers, and exams, we'd drunk so deeply of the distilled wisdom of the ages that surely we must know all there was to know about bodies and diseases. All that was left, it seemed, was learning how to apply that knowledge as apprentice doctors. Then we began to study with the senior members of the clinical departments, veterans of a less scientific era, who brought us up short. Their message soon became clear: We didn't know so damn much after all; no one did. All our learning was fine as far as it went, but on the hectic wards of Bellevue things often didn't go by the book.

The greatest teacher I had in those days was my surgery professor, Dr. John Mulholland, a granite cliff of a man whose iron-gray crew cut emphasized his uncompromising idealism. Any hint of laziness, impatience, or unconcern from his students brought a brusque reprimand, but Mulholland was gravely polite and compassionate to the dirtiest wino who needed a doctor. He showed us countless problems and techniques in his demonstration-lectures held in the nineteenth-century amphitheater, but he repeated one salient message over and over: "The surgeon can cut, remove, or rearrange the tissues, and sew up the wound, but only the patient can do the healing. Surgeons must always be humble before this miracle. We must treat the tissues with sure, deft gentleness, and above all we must do no harm, for we are nothing more than nature's assistants."

None of our textbooks could tell us the how and why of healing. They explained the basics of scientific medicine—anatomy, biochemistry, bacteriology, pathology, and physiology—each dealing with one aspect of the human body and its discontents. Within each subject the body was further subdivided into systems. The chemistry of muscle and bone, for example, was taught separately from that of the digestive and nervous systems. The same approach is used today, for fragmentation is the only way to deal with a complexity that would otherwise be overwhelming. The strategy works perfectly for understanding spaceships, computers, or other complicated machines, and it's very useful in biology. However, it leads to the reductionist assumption that once you understand the parts, you understand the whole. That approach ultimately fails in the study of living things—hence the widespread demand for an alternative, holistic medicine—for life is like no machine humans have ever built: It's always more than the sum of its parts.

With delicate lab culture methods we can remove from an animal certain organs and tissues, such as bone, a heart, a pancreas, a brain, or groups of nerve cells, keeping them alive for days or weeks. Much of modern biology in the West is based on the behavior of such isolated systems, which is assumed to be the same as in the living body. Russian biology, based on Ivan Pavlov's concept of the body as an indivisible unit, has always been skeptical of tissue culture results, considering these "parabiotic" reactions only hints toward definitive studies of the entire animal. The good sense of this view is shown by the fact that life tolerates fragmentation very poorly: Except in the simplest species removal of anything more than a few cells always destroys the organization, and hence the organism. Even if we could culture separately all the organs and tissues and then put them together like Dr. Frankenstein's monster, we would still, at our present level of knowledge, have only a collection of different kinds of meat, not a living entity. As Albert Szent-Gyorgyi once wrote, "Biology is the science of the improbable," and seldom can we predict new discoveries from what we already understand.

These limitations were clearly recognized by American medicine in the 1940s, but they seem to have been gradually forgotten. Today most M.D.'s tacitly assume that once a few blank spots are filled in, the established basic sciences will be all we'll ever need to take care of the sick. As a result, they're losing the forest among the trees. Of the disciplines that form medicine's foundation, only physiology tries to units structure and function into a complete picture of how the body works. Hence it's often called the queen of the biological sciences, yet even in this realm the synthesis is made organ by organ. There is, however, one organ group—the nervous system—that coordinates the activities of all the others; by receiving, transmitting, and storing information, it unites all the parts into that transcendence of fragments, the organism. Therefore the one discipline that comes closest to dealing with a living thing in its entirety is neurophysiology, which in the 1940s was already so sophisticated as to be almost a science unto itself.

Even neurophysiology couldn't explain the mystery of healing, however. My best texts either ignored it completely or shrugged it off in a few vague paragraphs. Moreover, my experiences at Bellevue during my internship and early residency convinced me that a physician's success was largely due, not to technical prowess, but to the concern he or she displayed toward patients. The patient's faith in the doctor profoundly affected the outcome of many treatments. Certain remedies, such as penicillin used against bacteria susceptible to it, worked every time. Other prescriptions weren't so predictable, however. If the patient thought the remedy would work, it usually did; otherwise it often didn't, no matter how up-to-date it was. Unfortunately, the importance of the doctor-patient relationship was being downgraded by the new scientific medicine. The new breed of physicians argued that this power of belief somehow wasn't real, that the patients only thought they were getting better—a bit of arrant nonsense that should have been quickly dispelled by a little open-minded, caring attentiveness on daily rounds. There was no known anatomical structure or biochemical process that provided the slightest reason to believe in such a thing, so it came to be dismissed as a mirage left over from the days of witchcraft. The placebo effect, as it's now called, wasn't documented until several decades later and still isn't fully accepted as an integral part of the healing process, but over the years I became convinced that it was a physiological effect of mind on body, just as real as the effects of wind on a tree.

Our lack of knowledge about healing in general and its psychological component in particular sowed seeds of doubt in my mind. I no longer believed that our science alone was an adequate basis for medical practice. As a surgeon, I tried to apply the principle of interaction on my own wards, by spending more time talking with patients, letting them know that I cared for them as well as taking care of the m. Naturally, as I became a teacher, I tried to pass on my beliefs to others. As I gained experience, I grew more and more convinced that all the textbooks were missing a chapter—the one that should have tied it all together and helped us doctors understand the bodily harmony we were trying to restore.

When I entered research, I aimed for a fairly limited goal among the many that lured me—finding out what stimulated and controlled the growth needed for healing—but always in the back of my mind were the larger questions that had haunted me since medical school: What unified an organism, making every cell subservient to the needs of the whole? How was it that the whole being could do things that none of its components could do separately? What made an organism self-contained, self-directed, self-repairing? When you get right down to it, I wanted to know what made living things alive. Intuitively I felt sure the answers needn't be forever hidden in mystic conundrums but were scientifically knowable. However, they would require a fresh approach from science, not the simple mechanistic dogmas left over from last century. As a result of the research on nerves and regeneration described in the foregoing pages, I believe I can now sketch at least an outline of that missing chapter.

It had been known for centuries that the nerves are the body's communications lines. Still, all the information collected by neurophysiologists hadn't revealed the integrating factor behind healing. Marc Singer proved that nerves are essential for regeneration, yet the elaborate impulse and neurotransmitter system, which until recently constituted everything we knew about nerves, carries no messages during the process. Nerves are just as essential to simpler kinds of healing. Leprosy and diabetes sometimes destroy nerve function to the extremities. When this happens, a wounded limb not only fails to heal but often degenerates far beyond the actual injury. I often thought about this paradox in connection with the other realities that were poorly explained by nerve impulses, such as consciousness and its many levels, sleep, biological cycles, and extrasensory experiences. As a doctor, however, I was most concerned with the mystery of pain.

This is the least understood of sensory functions, but it must have been one of the very first to evolve. Without it, living things would be so poorly designed that they couldn't survive, for they would never know what constituted danger or when to take defensive action. Pain is quite distinct from the sense of touch. If you place your finger on a hot stove, you feel the touch first and the pain appears a discernible time later, after the reflex has already drawn your hand away. Clearly the pain is conveyed by a different means. Furthermore, there are different types of pain. Pain in the skin is different from pain in the head or belly or muscles. If you ever want to embarrass a neurophysiologist, ask for an explanation of pain.

Early in my work on regeneration it occurred to me that I'd stumbled upon another method of nerve function. I imagined slowly varying currents flowing along the neurons, their fluctuations transmitting information in analog fashion. Though I kept my main focus on the role of these currents in healing, I pursued other lines of inquiry on the side. I did so partly out of simple curiosity, but also because I realized that, no matter how much merit my DC theory might have for healing, it would have a better chance of being considered if I could fill in some of its details in a wider context.

In the study of healing I dealt only with the output side of the system, the voltages and currents sent to the injured area to guide cells in repairing the damage. Cybernetics and common sense alike told me that, before an organism could repair itself, it must know it had been injured. In other words, the wound must hurt, and the pain must be part of an input side of the system. Certainly if the output side was run by electrical currents, it made no sense to assume that the input side relied on nerve impulses.

At the same time another problem nagged me. The impulses and the current seemed to coexist, yet everything we knew about nerve impulses and electricity said they couldn't travel through the same neuron at the same time without interfering with each other. We now have solutions to both problems, thanks to serendipity. The ways in which the answers came show how one experiment often furthers an unrelated one, and how politics occasionally benefits science.

The Constellation of the Body

In the early 1960s, after I'd published a few research papers, I had an unannounced visitor, a colonel from the Army Surgeon General's office. He said he'd been following my work from the start and had an idea he wanted to discuss. He liked if I'd ever heard of acupuncture.

I told him it wasn't the sort of thing taught in medical school. Although I'd read about it, I had no direct experience of it and didn't know whether it did any good.

"I can tell you for sure it does work ," he replied. "It definitely relieves pain. But we don't know how it works. If we knew that, the Army might adopt it for use by medics in wartime. After reading your work, some of us wondered if it might work electrically, the same as regeneration seems to. What do you think?"

That was a new idea to me, but right away I thought it was a good one. Although neurophysiologists had studied pain intensively for decades, there was still no coherent theory of it, or its blockage by anesthetics and anodynes. Because of Western medicine's biochemical bias, no pain-killers other than drugs were considered seriously. Maybe a physical method could give us a clue as to what pain really was.

We talked for several hours, but afterward I heard no more from the colonel, and I didn't get the chance to follow up his idea until more than a decade later. In 1971, while touring China as one of the first Western journalists admitted by the Communists, New York Times columnist James Reston saw several operations in which acupuncture was the only anesthetic, and he himself had postoperative pain relieved by needles after an emergency appendectomy. His reports put acupuncture in the news in a big way. It was almost the medical equivalent of Sputnik. Soon the National Institutes of Health solicited proposals for research on the Chinese technique, and I jumped at the chance.

At that time the prevailing view in the West was that if acupuncture worked at all, it acted through the placebo effect, as a function of belief. Hence it should be effective only about a third of the time, just like dummy pills in clinical tests. Many of those applying for the first grants began with this idea, and with the corollary that it wouldn't matter where you put the needles. Thus, much of our earliest research merely disproved this fallacy, which the Chinese—and apparently the U.S. Army—had done long ago. Recalling my talk with the colonel, I proposed a more elegant hypothesis.

The acupuncture meridians, I suggested, were electrical conductors that carried an injury message to the brain, which responded by sending back the appropriate level of direct current to stimulate healing in the troubled area. I also postulated that the brain's integration of the input included a message to the conscious mind that we interpreted as pain. Obviously, if you could block the incoming message, you would prevent the pain, and I suggested that acupuncture did exactly that.

Any current grows weaker with distance, due to resistance along the transmission cable. The smaller the amperage and voltage, the faster the current dies out. Electrical engineers solve this problem by building booster amplifiers every so often along a power line to get the signal back up to strength. For currents measured in nanoamperes and microvolts, the amplifiers would have to be no more than a few inches apart—
just like the acupuncture points! I envisioned hundreds of little DC

generators like dark stars sending their electricity along the meridians, an interior galaxy that the Chinese had somehow found and explored by trial and error over two thousand years ago. If the points really were amplifiers, then a metal needle stuck in one of them, connecting it with nearby tissue fluids, would short it out and stop the pain message. And if the integrity of health really was maintained by a balanced circulation of invisible energy through this constellation, as the Chinese believed, then various patterns of needle placement might indeed bring the currents into harmony, although that part of the treatment has yet to be evaluated by Western medical science.

The biggest problem Western medicine had in accepting acupuncture was that there were no known anatomical structures corresponding to the meridians, those live wires supposedly just under the skin. Some investigators claimed to have located tiny clusters of sensory neurons where the points were, but others had looked for them in vain. My proposal offered a convenient way into the problem. If the lines and points really were conductors and amplifiers, the skin above them would show specific electrical differences compared to the surrounding skin: Resistance would be less and electrical conductivity correspondingly greater, and a DC power source should be detectable right at the point. Some doctors, especially in China, had already measured lower skin resistance over the points and had begun using slow pulses of current, about two per second, instead of needles. If we could confirm these variations in skin resistance and measure current coming from the points, we'd know acupuncture was real in the Western sense, and we could go on confidently in search of the physical structures.

I got the grant and used part of the money to hire Maria Reichmanis, a brilliant young biophysicist who was Charlie Bachman's last Ph.D. student. Her combination of mathematical gifts and practicality got us results fast. Together we designed a "pizza cutter" electrode, a wheel that we could roll along the meridians to give us a reliable continuous reading, as well as a square grid of thirty-six electrodes to give us a map of readings around each point.

Along the first meridians Maria measured, the large-intestine and pericardial lines on the upper and lower surfaces, respectively, of each arm, she found the predicted electrical characteristics at half of the points. Most important, the same points showed up on all the people tested. Since acupuncture is scuh a delicate blend of tradition, experiment, and theory, the other points may be spurious; or they may simply be weaker, or a different kind, than the ones our instruments revealed. Our readings also indicated that the meridians were conducting current, and its polarity, matching the input side of the two-way system we'd charted in amphibians, showed a flow into the central nervous system. Each point was positive compared to its environs, and each one had a field surrounding it, with its own characteristic shape. We even found a fifteen-minute rhythm in the current strength at the points, superimposed on the circadian ("about a day") rhythm we'd found a decade earlier in the overall DC system. It was obvious by then that at least the major parts of the acupuncture charts had, as the jargon goes, "an objective basis in reality."

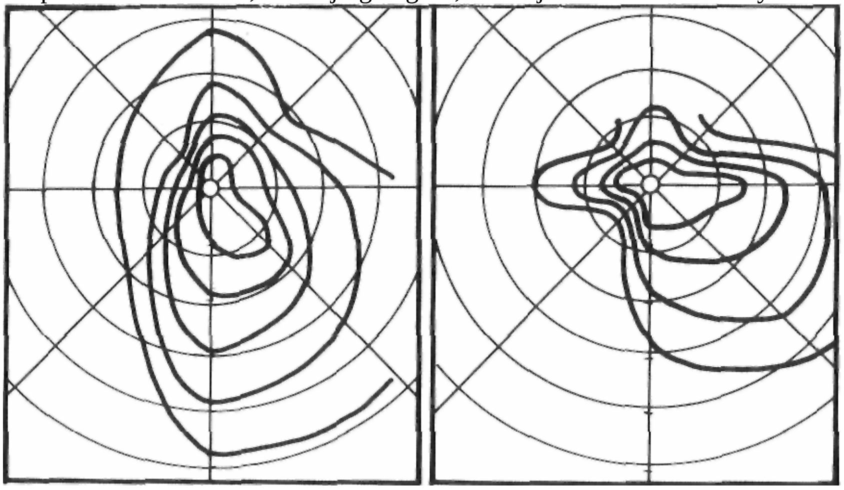

ELECTRICAL CONDUCTIVITY MAPS OF SKIN AT ACUPUNCTURE POINTS

Maria, Joe Spadaro, and I began a more sophisticated series of tests. We planned to record from six major points along one meridian as a needle was inserted into the outermost point. If the DC theory was valid, a change in potential should travel from point to point along the line. However, just as we were entering this second phase, the NIH canceled our grant, even though we'd published four papers in a year. Supposedly it had lost interest in acupuncture, at least in the kind of basic research we were doing on it. Even so, I was fairly satisfied. The input system worked as I'd predicted. The other major question remained: What structure carried the current so as not to interfere with the nerve impulses?

Of course, I've given away the answer in previous chapters. The perineural cells appear to carry the current. In the early 1970s, however, we only suspected this. The evidence came unexpectedly by cross-fertilizarion from an unrelated project.

One of the main problems of medical research is finding a suitable "animal model" for human diseases. The study of unmended fractures is especially hard, because people are endowed with lesser fracture-healing abilities than most other animals, for whom nonunions are a non-problem. Based on what I'd learned about the importance of nerves in bone healing, I figured we could produce nonunions in rats by cutting the nerves to the broken leg, particularly if we removed whole segments of the nerves so they couldn't grow back. I assigned this part of the project to Dr. Bruce Baker, a young orthopedic surgeon who was then finishing his residency with an extra year on a fellowship in my lab.

After Bruce had worked out the complicated surgical procedure, we anesthetized a series of rats, removed the nerve supply to one leg of each animal, and broke the fibula, or smaller bone of the calf, in a standard way. Then every day we reanesthetized a few of the rats and took out the fracture area to mount it for the microscope. At the same time, Bruce checked the cut nerves to make sure there was no regrowth. Successful denervation was confirmed by the microscope and by complete paralysis of the affected leg.

The results were encouraging yet puzzling. The nerves didn't regrow, and the broken bones took twice the normal six or seven days to heal, but heal they did, even though theoretically they shouldn't have knit at all without nerves.

It was well known that the severed end of a nerve would die after a couple of days, but, since we'd cut the nerves at the same time as we'd broken the bones, maybe the cut ends had exerted a subdued healing effect while they remained alive. In another series of animals, we cut the nerves first. Three days later, after making sure the legs were fully de-nervated, we operated again to make the fractures. We felt sure the delay would give us true nonunions. To our surprise, however, the bones healed faster than they had in our first series, although they still took a few days longer than normal.

Here was a first-class enigma. The only thing we could think of doing was to cut the nerves even earlier, six days before the fractures. When we got that series of slides back, we found that these animals, whose legs were still completely without nerves, healed the breaks just as fast and just as well as the normal control animals. Then we took a more detailed microscopic look at the specimens Bruce had taken from around the nerve cut. We found that the Schwann cell sheaths were growing across the gap during the six-day delay. As soon as the perineural sleeve was mended, the bones began to heal normally, indicating that at least the healing, or output, signal was being carried by the sheath rather than the nerve itself. The cells that biologists had considered merely insulation turned out to be the real wires.

Unifying Pathways

The experiments I'd done with psychiatrist Howard Friedman in the early 1960s, mentioned in Chapter 5, were the first to provide strong support for an analog theory of pain. In all animals, including humans, the normal negative potentials at the extremities weakened or vanished as an anesthetic took effect. Under deep total anesthesia, the potentials often reversed entirely, the extremities becoming positive and the brain and spine negative. At that time we didn't yet know about the two-way system, inward along sensory nerves and outward along motor nerves, but it was obvious that a current flow was being reversed by the pain-preventing drugs. In lab animals and humans under local anesthetic, such as a shot of procaine in one arm, the negative potential was abolished only for that arm. The DC potentials over the head were unaffected—except for a little blip in the recording that registered the prick of the needle!

In addition, the DC potentials react slowly enough to account for pain. A wound usually doesn't start to hurt in earnest until minutes, or even hours, after the injury. This delay has been especially hard to explain in terms of nerve impulses, which travel at 30 feet per second. However, when Friedman and I injured the limbs of salamanders while monitoring the potentials on their limbs and heads, we found that the change in the limb reading showed up in the head after a time approximating that of delayed pain. Acupuncture likewise involves a delay, usually twenty minutes or more, before its effects are felt.

We also found we could work backward, using the currents to produce anesthesia. A strong enough magnetic field oriented at right angles to a current magnetically "clamped" it, stopping the flow. By placing frogs and salamanders between the poles of an electromagnet so that the back-to-front current in their heads was perpendicular to the magnetic lines of force, we could anesthetize the animals just as well as we could with chemicals, and EEG recordings of magnetic and chemical anesthesia were identical. We got the same effect by passing a current through the brain from front to back, canceling out the normal current of waking consciousness, as in electrosleep.

One of the most exciting results of my collaboration with Dr. Friedman was proof that one's state of waking consciousness could change the perception of pain. Friedman, who already used hypnosis to control chronic pain in his patients, gave several of his best subjects hypnotic suggestions of arm numbness deep enough that they couldn't feel the prick of a needle. In each case, I found that the frontal negative potential of the head became less negative, often reaching zero, as the client attained deep trance. The reading changed in the same direction as in anesthesia, only not as far. Then, when the suggestion for pain control was given, the arm potential reversed just as it had in response to procaine. Conversely, when a control subject was asked in normal waking consciousness to concentrate forcefully on one arm, its sensitivity to pain increased, and the hand potential became more negative. We found we could use this difference to determine whether a person was really hypnotized or just cooperating.

Some doubters (including myself, I'm afraid) had believed hypnoanalgesia was merely a state in which the patient still felt the pain but didn't respond to it, but these experiments proved it was a real blockage of pain perception. It seems that the brain can shut off pain by altering the direct-current potentials in the rest of the body "at will." There's every reason to suppose that pain control through biofeedback or yoga likewise works by using an innate circuit for attenuating the pain signal, which releases a shot of the body's own pain-killers. When the signal is appropriately modulated, it releases endorphins (internally produced opiates), as shown by experiments in which an injection of the opiate-antagonist naloxone negates the anesthesia of acupuncture. I predict that research on this system will eventually let us learn to control pain, healing, and growth with our minds alone, substantially reducing the need for physicians.

Direct evidence for the perineural DC system has been accumulating gradually for several decades. Electric currents were detected in the glial cells of rat brains as long ago as 1958, and good (though long-ignored) measurements of direct currents in the frog's brain go back to the work of Ralph Gerard and Benjamin Libet in the early 1940s. Electron microscope work has shown that the cytoplasm of all Schwann cells is linked together through holes in the adjacent membranes, forming a syncytium that could provide the uninterrupted pathway needed by the current. The other perineural cells—the ependyma and glia—are probably connected in the same way, for syncytial links have recently been found in the glia of the leech, whose nervous system is much studied because of its unusually large cells. Recent use of selective radiation to isolate Schwann cells has shown that they, and not the neuron fibers, supply the nerve stimulus essential to regeneration.

The invention of a better magnetometer has yielded definitive proof that's now widely acknowledged. Any electric current automatically generates a magnetic field around itself. Hence, as the perineural current conveys information in its fluctuations, it must be reflected by a magnetic field around the body, whose pulsing would reveal the same information. When I first proposed this idea, many of my colleagues dismissed it as rank nonsense. I couldn't prove them wrong, because there were no instruments to measure a field as weak as that generated by such small currents. Everyone knew the human body had no effect on a compass needle or any other magnetic-field detector available at the time.

Then, in 1964, a solid-state physicist named Brian D. Josephson invented the electronic device now called a Josephson junction, a simple item that won him a Nobel Prize. Basically it consists of two semiconductors connected so that current can oscillate in a controlled fashion between them. Today it has many applications, especially in computers. When cooled near absolute zero in a bath of liquid helium, it becomes a superconductor in which the current plays back and forth endlessly. Superconduction is the passage of electrons through a substance without the resistance normally found in any conductor. This apparatus, called a superconducting quantum interferometric device, or SQUID for short, is a magnetic field detector thousands of times more sensitive than any previously known.

In 1963, G. M. Boule and R. McFee just barely managed to measure the relatively large magnetic field produced by the human heart—using the best old-fashioned instrument, a coil with 2 million turns of wire. Then, in 1971, working in a null-field chamber, from which the earth's magnetism and all artificial fields were screened out, Dr. David Cohen of MIT's Francis Bitter National Magnet Laboratory, who'd been corresponding with our lab since the early years, first used the SQUID to measure the human head's magnetic field. Two kinds of magnetic fields have been found. Quickly reversing AC fields are produced by the back-and-forth ion currents in nerve and muscle. They're strongest in the heart, since its cells contract in synchrony. The SQUID has also confirmed the existence of the direct-current perineural system, which, especially in the brain, produces steady DC magnetic fields one billionth the strength of earth's field of about one-half gauss.

By 1975, Drs. Samuel Williamson, Lloyd Kaufman, and Douglas Brenner of NYU had succeeded in measuring the head's field without a shielded enclosure, even amid the electromagnetic noise of downtown Manhattan. More important, they've found that the magnetoencephalogram (MHG)—a recording of changes in the brain's field analogous to the EEG - is often a more accurate reflection of mental activity than the EEG. Because the magnetic field passes right through the dura, skull bones, and scalp without being diffused, an MEG locates the current source more accurately than EEG measurements. The NYU group has since begun correlating magnetic events with well-known cerebral responses, such as the reaction of cells in the visual cortex to simple patterns and flashes of light. When the brain reacts to any stimulus, it produces a wave of electrical activity that's contained in the EEG. It's invisible in a standard EEG recording, because so much else is always going on in the brain at the same time. However, when one simple stimulus is repeated many times and the EEG tracings are averaged by computer, the particular electrical response to that one stimulus—called an evoked potential—can be teased out. Several research groups have slowly built up a small vocabulary of wave forms with specific meanings, including a "surprise wave," an "intention wave," and a "double-take wave," which appears when the mind briefly tries to make sense of semantic nonsense, as in the statement "She took a drink from the radio."

Taken together, the MEG research so far seems to be establishing that every electrical evoked potential is accompanied by a magnetic evoked potential. This would mean that the evoked potentials and the EEG of which they're a part reflect true electrical activity, not some artifact of nerve impulses being discharged in unison, as was earlier theorized. Some of the MEG's components could come from such additive nerve impulses, but other aspects of it clearly indicate direct currents in the brain, particularly the central front-to-back flow. The MEG doesn't show the EEG's higher-frequency components, however, suggesting that some parts of the two arise from different sources.

Since every reaction and thought seems to produce an evoked potential, the DC system seems directly involved in every phase of mental activity. At the very least, the electric sheath acts as a bias control, a sort of background stabilizer that keeps the nerve impulses flowing in the proper direction and regulates their speed and frequency. But the analog structure probably plays a more active role in the life of the mind. Variations in the current from one place to another in the perineural system apparently form part of every decision, every interpretation, every command, every vacillation, every feeling, and every word of interior monologue, conscious or unconscious, that we conduct in our heads.

This part of the analog system's job is much less well understood, however, than its integrative function throughout the rest of the body.

Perineural cells accompany every part of the nervous system. Even the tiniest twiglets of sensory nerves in the skin, which don't have a myelin covering, are surrounded by Schwann cells. The perineural thus just as well distributed to integrate bodily processes as the nerves themselves. They reach into each area of the body to create a normal electrical environment around each cell, or a stimulatory one when healing growth is needed. Likewise they enable an organism to sense the type and extent of damage anywhere in the body by transmitting the current of injury, with its by-product of pain, to the CNS. One could take a "fantastic voyage" from the farthest Schwann cell outpost in the big toe through the spinal cord and into all parts of the brain. Indeed, electrons are making this trip every moment of our lives.

Thus our bodies have an intricate and multilayered self-regulating feedback arrangement. We know, on the psychological level, that a person's emotions affect the efficiency of healing and the level of pain, and there's every reason to believe that emotions, on the physiological level, have their effect by modulating the current that directly controls pain and healing.

These discoveries give us a testable physical basis for the placebo effect and the importance of the doctor-patient bond. They also may give us the key to understanding the "miracle" cures of shamans, faith healers, and saints, as well as the spontaneous healing reported by means of visions, prayer, yoga, or battlefield terror. At the Menninger Foundation, Elmer Green has long been using biofeedback to explore the mind-body relationship. Green has described the full yogic control of pain and healing developed by one of his subjects, an otherwise average businessman. He lay against a bed of nails with no pain, and, when informed that a puncture from one of the points was bleeding, he turned his head, gazed at the wound, and immediately stopped the flow. A combination of biofeedback, recording electrodes, and the SQUID magnetometer would seem to be the ideal setup for the next level of inquiry into the mind's healing powers.

Moreover, since the analog system, like the impulse network, appears to work on both conscious and subconscious levels simultaneously, it's a likely missing link in several other poorly understood integrative functions that also cross from one realm to the other. It may lead us at last to fathom the twin wells of memory and emotion. It may even help us understand what happens when a new synthesis of creative thought, a.k.a. inspiration, bursts forth like a mushroom from strands of mycelia that have been quietly gathering their subterranean forces. Then science for the first time will begin to comprehend the artistic essence that makes its rational side productive.

Fourteen

Breathing with the Earth

Major changes in one's life often proceed improbably from the most minor events. So it was that I became involved in one of the most interesting parts of my work in 1961 because a dog bit someone I didn't even know at the time.

My first electrical measurements on salamanders had just revealed part of the DC control system to me. Elementary physics told me that the currents and their associated electromagnetic fields would have to be affected in some way by external fields. In engineering terms, the biomagnetic field would be coupled to the DC currents. Hence changes impressed upon it by external fields would be "read out" through perturbations in the current. Outside fields would also couple directly to the currents themselves, without acting through the biofield as intermediary, especially if the currents were semiconducting. In short, all living things having such a system would share the common experience of being plugged in to the electromagnetic fields of earth, which in turn vary in response to the moon and sun. In the late eighteenth century, the Viennese hypnotist and healer Franz Anton Mesmer proposed direct magnetic influences on earthly bodies from the heavenly ones, but his idea came from the scientifically unacceptable domain of astrology. With the notable exception of Nikola Tesla, most prominent researchers have derided it until recently. I figured the DC system must be the missing link in a very different, but very real, connection between geophysics and the responses of living things. I was eager to investigate it, but at firsr I didn't know how.

I was an orthopedic surgeon, about as far removed as possible from the psychiatric expertise needed for a serious study of behavior. And suppose I did find something? Who would believe me if I ventured so far from my specialty? The whole idea was preposterous to the science of the time, anyway. Still, I had to do something.

During the International Geophysical Year of 1957-58, I'd been a volunteer in the Aurora Watch Program. To find out whether the northern lights appeared simultaneously throughout the north latitudes in response to changes in the earth's magnetic field (they did), IGY organizers recruited a worldwide network of amateur observers to go out into their backyards every night and look at the sky. All of us got weekly reports on the state of the field from the national magnetic observatory at Fredericksburg, Virginia. I decided to go back through this data and see if there was any correlation between the disturbances in earth's field caused by magnetic storms on the sun, and the rate of psychiatric admissions to our VA hospital.

Luckily for me, Howard Friedman, the hospital's chief of psychology, was collecting donations door to door for a local Boy Scout troop at about this time. At one house, the family dog took an instant dislike to him and bit his ankle. After bandaging the wound, Howard's doctor gave him a tetanus booster shot. As luck would have it, Howard came down with a rare allergic reaction that involved fever, fatigue, nausea, and painful swelling of all the joints.

Since I was the nearest bone-and-joint man, Howard came to see me. This type of reaction is frightening, but not dangerous, and disappears of its own accord in a day or two. After I made the diagnosis and reassured him, we sat and talked for a few minutes. After some chitchat about the shortcomings of the hospital administration, he gestured at the papers tacked all over the walls of my office and asked, "What are all those charts?" I told him about my magnetic brainstorms.

He obviously thought I was as crazy as the people whose admissions I was charting, and probably wondered about the advice I'd just given. However, after hearing the background, he agreed it wasn't as silly as it sounded, and offered to help. It was a real break for me, since he was already a respected researcher, and a practical, open-minded one to boot. My diagnosis was correct, and our collaboration lasted almost two decades.

Howard's reputation got us access to the records of state psychiatric hospitals, giving us a sample large enough to be statistically useful. We matched the admissions of over twenty-eight thousand patients at eight hospitals against sixty-seven magnetic storms over the previous four years. The relationship was there: Significantly more persons were signed in to the psychiatric services just after magnetic disturbances than when the field was stable. Of course, such a finding could only serve as a guide to further investigation, because so many factors determined whether a person sought psychiatric help, but we felt that other influences would even out over such a large number of patients.

Next we looked for the same type of influence in patients already hospitalized. We selected a dozen schizophrenics who were scheduled to remain in the VA hospital for the next few months with no changes in treatment. We asked the ward nurses to fill out a standard evaluation of their behavior once every eight-hour shift. Then we correlated the results with cosmic ray measurements taken every two hours from government measuring stations in Ontario and Colorado. Since magnetic storms were generally accompanied by a decrease in the cosmic radiation reaching earth, we thought we might find changes in the patients' actions and moods during these declines. We decided to use cosmic rays instead of direct reports of the magnetic field strength because of problems in distinguishing between magnetic storms and other variations in the earth's field.

The nurses reported various behavior changes in almost all the subjects one or two days after cosmic ray decreases. This was a revealing delay, for one type of incoming radiation—low-energy cosmic ray flares from the sun—was known to produce strong disruptions in the earth's field one or two days later.

With this encouragement we went on in 1967 to confirm, by experiments described more fully in the next chapter, that abnormal magnetic fields did produce abnormalities in various human and animal responses. We found slowed reaction times in humans and a generalized stress response in rabbits exposed to fields ten or twenty times the normal strength of the earth's. Hence we suspected that the earth's normal field played a major role in keeping the DC system's control of bodily functions within normal bounds. The proof of this idea has come mainly from the work of two men: Frank Brown at Northwestern and Rutger Wever, working at the Max Planck Institute in Munich.

Already a respected endocrinologist, Brown became interested in biocycles in the 1950s. It was common knowledge that most organisms had a circadian rhythm of metabolic activity, which most people assumed was directly linked to the alternation of night and day or, in the case of shore life, to the tides. Oysters, for example, would open their shells to feed whenever the tide came in, covering them with water. It was a simple, obvious observation, but Brown didn't take it for granted. To his surprise, oysters in an aquarium with constant light, temperature, and water level still opened and closed their shells in time with their compatriots at the beach. To find out why, Brown flew oysters in a lightproof box from New Haven to his lab, in Evanston, Illinois. At first they kept time with Connecticut oysters, then in a few weeks gradually shifted to the tide pattern Evanston would have had if it had been on a seacoast. The oysters not only knew they'd been taken 1,000 miles westward, they also suffered from jet lag!

In his search for a creature whose response to magnetic fields might tell him more about biocycles, Brown settled upon the mud snail Nassarius, at home in the intertidal zone anywhere in the world. In his lab he placed the snails under uniform illumination in a box with an exit facing magnetic south. When they left the enclosure in early morning, they turned west. When leaving at noon, they turned east, but took a westerly course again in early evening. Furthermore, at new and full moon, the snails' paths veered to the west, while at the quarters they tended more eastward than at other times.

Brown's precise data from this and many other experiments showed that Nassarius had two clocks, one on solar time and one on lunar, and subsequent work with magnets told something about how the timepieces ran. The earth's magnetic field averaged 0.17 gauss in Evanston. When Brown put a 1.5 gauss permanent magnet facing north-south underneath the snails' doorways to augment the natural field, the animals made sharper turns, but their direction wasn't affected. Turning either the magnet or the enclosure through various angles made the snails change course a specific number of degrees. Brown concluded: "It seemed as if the snails possessed two directional antennae for detecting the magnetic field direction, and that these were turning, one with a solar day rhythm and the other with a lunar day one." This crucial experiment not only showed the dependence of biocycles on the earth's magnetic field, it also demonstrated the subtlety of the link. No longer could we expect changes in the magnetic environment to be as obvious in their effects on life as changes in oxygen levels, food supply, or temperature.

The niceties of the earth's electromagnetic field itself became better known as Brown's work progressed. Far from a static, simple magnetic field like that around a uniform bar of magnetized iron, the earth's field has turned out to have many components, each full of quirks.

At the end of the nineteenth century, geophysicists found that the earth's magnetic field varied as the moon revolved around it. In the same period, anthropologists were learning that most preliterate cultures reckoned their calendar time primarily by the moon. These discoveries led Svante Arrhenius, the Swedish natural philosopher and father of ion chemistry, to suggest that this tidal magnetic rhythm was an innate timekeeper regulating the few obvious biocycles then known.

Since then we've learned of many other cyclic changes in the energy structure around us:

The earth's electromagnetic field is largely a result of interaction between the magnetic field per se, emanating from the planet's molten iron-nickel core, and the charged gas of the ionosphere. It varies with the lunar day and month, and there's also a yearly change as we revolve around the sun.

A cycle of several centuries is driven from somewhere in the galactic center.

The earth's surface and the ionosphere form an electrodynamic resonating cavity that produces micropulsations in the magnetic field at extremely low frequencies, from about 25 per second down to 1 every ten seconds. Most of the micropulsation energy is concentrated at about 10 hertz (cycles per second).

Solar flares spew charged particles into the earth's field, causing magnetic storms. The particles join those already in the outer reaches of the field (the Van Allen belts), which protect us by absorbing these and other high-energy cosmic rays.

Every flash of lightning releases a burst of radio energy at kilocycle frequencies, which travels parallel to the magnetic field's lines of force and bounces back and forth between the north and south poles many times before fading out.

The surface and ionosphere act as the charged plates of a condenser (a charge storage device), producing an electrostatic field of hundreds or thousands of volts per foot. This electric field continually ionizes many of the molecules of the air's gases, and it, too, pulses in the ELF (extremely low frequency) range.

There are also large direct currents continually flowing within the ionosphere and as telluric (within-the-earth) currents, generating their own subsidiary electromagnetic fields.

In the 1970s we learned that the sun's magnetic field is divided from pole to pole into sectors, like the sections of an orange, and the field in each sector is oriented in the direction opposite to adjacent sectors. About every eight days the sun's rotation brings a new region of the interplanetary (solar) magnetic field opposite us, and the earth's field is slihgtly changed in response to the flip-flop in polarity. The sector boundary's passage also induces a day or two of turbulence in earth's field.

The potential interactions among all these electromagnetic phenomena and life are almost infinitely complex.

For many years most scientists dismissed Brown's conclusions as impossible. Given the old premise that life was entirely a matter of water chemistry, none of these electromagnetic changes would have enough energy to affect an organic process in any way. Discovery of the DC system showed how the interaction could work without energy transfer; it gave living things a way of "sensing" the fields directly. Undaunted by the slow acceptance of his work, Brown went on to document Nassaria's sensitivity to electrostatic fields as well. He also found magnetically driven cycles in all other organisms he tested, including mice, fruit flies, and humans. Even potatoes in a bin showed a field-linked rhythm of oxygen consumption. In humans, hormone output and the number of lymphocytes in the bloodstream are but two of many variables that dance to the same beat. One of the most important is cell cycle time. The actual process of cell division—in which chromosomes appear, line up, split in half, and are distributed equally between the two cells—takes only a few minutes. It must be preceded by several longer stages, one of which is duplication of all the cell's DNA. All stages together take about one day. Thus all growth and repair, which depend on regulated cell division, are synchronized with the earth's field.

Rutger Wever has done some even more telling work with humans during the last decade and a half. He built two underground rooms to completely isolate people from all clues to the passage of time. One was kept free of outside changes in light, temperature, sound, and such ordinary cues, but wasn't shielded from electromagnetic fields. The other room was identical but also field free. Observing several hundred subjects, who lived in the bunkers as long as two months, and charting such markers as body temperature, sleep-waking cycles, and urinary excretion of sodium, potassium, and calcium, Wever found that persons in both rooms soon developed irregular rhythms, but those in the completely shielded room had significantly longer ones. Those still exposed to the earth's field kept to a rhythm close to twenty-four hours. In some of these people, a few variables wandered from the circadian rate, but they always stabilized at some new rate in harmony with the basic one—two days instead of one, for example. People kept from contact with the earth field, on the other hand, became thoroughly desynchronized. Several variables shifted away from the rhythms of other metabolic systems, which had already lost the circadian rhythm, and established new rates having no relationship to each other.

Wever next tried introducing various electric and magnetic fields into his completely shielded room. Only one had any effect on the amorphous cycles. An infinitesimal electric field (0.025 volts per centimeter) pulsing at 10 hertz dramatically restored normal patterns to most of the biological measurements. Wever concluded that this frequency in the micropulsations of the earth's electromagnetic field was the prime timer of biocycles. The results have since been confirmed in guinea pigs and mice. In light of this work, the fact that 10 hertz is also the dominant (alpha) frequency of the EEG in all animals becomes another significant bit of evidence that every creature is hooked up to the earth electromagnetically through its DC system. Recently a group under Indian biophysicist Sarada Subrahmanyam reported that the human EEG not only responded to the micropulsations, but responded differently depending on which way the subject's head was facing in relation to the earth's field. Oddly enough, however, the head direction had no effect if the subject was a yogi.

The relationship has been conclusively proven by recent studies of the pineal gland. This tiny organ in the center of the cranium has turned out to be more than the vaguely defined "third eye" of the mystics. It produces melatonin and serotonin, two neurohormones that, among many other functions, directly control all of the biocycles. The lamprey, akin to the ancestor of all vertebrates, as well as certain lizards, has an actual third eye, close to the head's surface and directly responsive to light, instead of the "blind" pineal found in other vertebrates. The eminent British anatomist J. Z. Young has recently shown that this organ controls the daily rhythm of skin color changes that these animals undergo.

For our story the most important point is that very small magnetic fields influence the pineal gland. Several research groups have shown that applying a magnetic field of half a gauss or less, oriented so as to add to or subtract from the earth's normal field, will increase or decrease production of pineal melatonin and serotonin. Other groups have observed physical changes in the gland's cells in response to such fields. The experiments were controlled for illumination, since it has been known for several years that shining a light on the head somehow modifies the gland's hormone output even though it's buried so deeply within the head in most vertebrates that, as far as we know, it can't react directly to the light.

We likely have yet to discover many other ways that energy cycles in the solar system affect life on earth. They may strongly affect the outbreak of disease, for example. The last six peaks of the eleven-year sunspot cycle have coincided with major flu epidemics. A Soviet group under Yu. N. Achkasova at the Crimean Medical Institute, working with astronomer B. M. Vladimirsky of the Crimean Observatory, has found a connection between the sun's magnetic field and the Escherichia coli bacteria that live in our intestines and help us digest our food. The Russians found the bacteria grew faster when the sun's field was positive, or pointing toward earth, and slowed down when it was negative. Two days after the passage of each sector boundary there was a dip in bacterial growth corresponding to the maximum geomagnetic turbulence. The data also showed a decline in growth in response to large solar flares. Other Russian scientists have drawn a tentative correlation between the sector cycle and reports from two groups of persons with neurological diseases. The patients felt worse within sectors of positive polarity, when bacteria seemed to grow faster. Life's geomagnetic coupling to heaven and earth is apparently more like a web than a simple cord and socket.

The Attractions of Home

An animal's biocycles must be appropriate for its environment if it's to survive, so they must be precisely tuned to its geographic location. We might suspect, therefore, that many creatures would use magnetic information for their sense of place. A great deal of recent work has shown that they do. A built-in compass helps guide them in foraging or other local business, as well as migration over much longer routes. The latter feats often rival those of any modern navigator. Monarch butterflies travel from Hudson Bay to South America straight across the Caribbean without ever getting lost. The arctic tern breeds in summer on the northern ice cap, then moves to Antarctica for the other hemisphere's summer, flying 11,000 miles each way. Some salamanders, only inches long and built very low to the ground, travel up to 30 miles of rugged mountain country in California to set up housekeeping, then return to their home stream to breed. Such activities are hard to experiment with, however, so we've learned more about day-to-day travel.

Karl von Frisch was the first to attack the problem, with his famous 1940s studies of the honeybee dance, which won him a Nobel Prize in 1973. He established that on clear days bees navigated by combining the sun's angle with their sense of time, the way Boy Scouts are taught to use wristwatches as compasses. The bees also had a polarized light system that could determine the sun's direction through light clouds or forest canopy. Even more amazing, Frisch found, the scouts told workers back at the hive where the flowers were by means of a dance using the sun's angle and the direction to earth's center (the gravity vector) as references. However, Frisch noted that bees could still navigate between food and home just as well on completely overcast days, when the sun's angle and polarized light weren't available. There had to be a backup system.

It soon turned out that homing pigeons had the same abilities. In 1953 G. Kramer inferred that the birds must have a compass in addition to a map of remembered landmarks from the way they immediately pointed their beaks toward home after circling once after release. Soon others found the same kind of sun compass as bees used, but the pigeons could also steer perfectly on cloudy days. Back in 1947, H. L. Yeagley had had the temerity to suggest in the Journal of Applied Physics that pigeons might have a magnetic sense that allowed them to use the earth's field just as we use magnetic compasses. He was ridiculed and "refuted" by a few inadequate experiments—such as placing a pigeon in a variety of electromagnetic fields and noting that it seemed to be comfortable! Others, including Yeagley, attached small magnets to the birds' heads or wings, but found no clear-cut changes in their flight patterns.

Only a few researchers quietly looked into the matter further. After Hans Fromme of the Frankfurt Zoological Institute noted in the late 1950s that caged European robins faced longingly in their normal southwest migratory direction even when they were kept from seeing sun and stars, their usual signposts, his co-worker Friedrich Merkel discovered that, insulated from the earth's field by a steel cage, they no longer faced in one particular direction. Furthermore, by changing the orientation of the surrounding field with coils, he could give the birds a false sense of where southwest was. The experiment was validated with indigo buntings several years later.

There the matter rested until 1971, when William T. Keeton of Cornell realized that the pigeon's magnetic sense, if it existed, would be overshadowed by its sun compass, so naturally magnets affixed to the birds would have no effect on clear days. He soon found that the same birds released on a cloudy day got lost.

To study this magnetic interference in any weather, Keeton made translucent contact lenses for his birds, then released them in the mountains of northern New York. Simulating dense clouds, the contacts blocked the sun's angle and polarized light, and with 1-gauss magnets attached to their heads the birds couldn't find their way home. However, each avian Ulysses who wore the lenses but no magnets faultlessly navigated the 150 miles southwest to Ithaca, then flew ever tighter circles around the loft and fluttered in like a helicopter to a perfect blind landing.

Carrying the work forward after Keeton's untimely death soon after this experiment, Charles Walcott and Robert Green of the State University of New York at Stony Brook, working with James L. Gould of Princeton, outfitted pigeons with miniaturized electromagnetic coils that let the researchers vary the type and orientation of applied field at will. They discovered that if the south pole of the field was directed up, the birds could still find home, but with the north pole up they flew directly away from it. That meant they were using magnetic north as a reference point. At about the same time two German scientists, Martin Lindauer and Herman Martin, analyzed half a million bee dances and found a "magnetic error" in them—a compensation for the difference between magnetic north and true north. They were also able to introduce specific angles of error in the dances with specifically oriented coils around the hive. Here was proof that magnetic guidance systems existed in both the birds and the bees. * The next question was how the systems worked.

In 1975 Richard P. Blakemore, then a graduate student at the University of Massachusetts in Amherst, astonished the world of biology with the announcement that some bacteria, the lowliest of all cells, also had a magnetic sense. Blakemore made the discovery when, studying the salt marshes of Cape Cod, he noticed that one type of bacterium always oriented itself north-south on his microscope slides. Soon he found magnetotactic bacteria (those reactive to magnetism) near Cambridge, Massachusetts, where he set about studying them with Richard Frankel of MIT's magnet lab. The direction to magnetic north points through the earth somewhat down from the horizon, and the scientists became convinced that the bacteria were using the field to guide themselves ever downward to the mud where they throve, since they were too small to sink through the random molecular motion of the water around them.

---

* Recent work by Cornell biologist Kraig Adler showed that the magnetic sense of salamanders was many times more acute than even even that of pigeons. Not only could the amphibians find home without tight or other common cues; in addition, when Adler tried to confuse them with artificial fields, they quickly adapted to the interference and oriented themselves correctly in relation to the weaker geomagnetic background.

This idea was later confirmed by findings that microbes at Rio and in New Zealand were south-seekers.

Blakemore's electron micrographs soon revealed a surprising structure. Each bacterium contained within it, like a chain of cut jet stones, a straight line of magnetite microcrystals. Surrounded by a thin membrane, each of these particles was a single domain, the smallest piece of the mineral that could still be a magnet.

Blakemore's bacteria led Gould to look for similar crystals in bees and pigeons. Since an electron microscope survey of even a bee's brain would take several lifetimes, he examined the insects with a SQUID magnetometer. After confirming that they were magnetic, he dissected them and narrowed the location down to a part of the abdomen. Using the same method, Walcott and Green dissected the heads of two dozen pigeons, gradually subdividing them with nonmagnetic probes and scalpels. After a painstaking search the investigators found a tiny magnetic deposit in a 1- by 2-millimeter piece of tissue richly festooned with nerves, on the right side of the head, between the brain and the inner table of the skull. The same dot of tissue contained yellow crystals of the iron-storage protein ferritin, indicating that the pigeons, like the bacteria, synthesized their own lodestone crystals.

As usual, these recent answers have raised plenty of new questions. The existence of magnetic sensors in such diverse creatures as bacteria, bees, and birds—the current count of species with magnetic organs is twenty-seven, including three primates—suggests that a magnetic sense has existed from the very beginning of life, perhaps only to be perfected by creatures that need to get around a lot. Do all animals, then, have the same sensors, and do they always serve the same function? How is the information read out of the crystals by the nervous system and translated into directions? What aspect of the earth's field do these organs sense?

Keeton noticed an especially odd thing about his pigeons' flight patterns. When flying on visual flight rules by sun compass, they would circle once, get their bearings, then move off straight toward Ithaca. But when using their magnetic compass, the birds would fly due west from their release point until they got out over Lake Ontario, due north of Ithaca. Then, out of sight of land, they would make a right-angle turn to the left and follow the exact meridian of home. Keeton told me this but never published the observation because he didn't know what to make of it. He said, "I asked a physicist: 'Are they making contact with a certain magnetic line of force?' The man said, 'No, magnetic lines of force an just an arbitrary convention we use to symbolize a field and describe anomalies in it, such as occur around iron ore deposits. As far as we know, the lines have no equivalent in reality, and if they did, they would vary all over the place as the earth's field changes, anyway.'" Do pigeons then follow some maplike structure in the earth's field itself, a grid like that described by dowsers and geomancers since ancient times, something we can't find today even with our SQUID? Some migrating birds make a dogleg to the east in their north-south flyway, sailing out of sight of land over Lake Superior. Do they go out of their way to avoid being disoriented by iron ore deposits in the Mesabi Range? We can suspect, but we don't yet know.

To most people, of course, the most interesting questions concern themselves. Do we, too, have compasses in our heads?

On June 29, 1979, R. Robin Baker, a young University of Manchester researcher into bionavigation, led a group of high school students into a bus at Barnard Castle, near Leeds, England. Baker blindfolded and earmuffed them, then gave them all headbands. Half of the headgear contained magnets and half contained brass bars that their wearers thought were magnets. As the volunteers leaned back to concentrate, Baker wove a mazelike course through the town's tangled streets, then traveled a straight highway to the southwest. After a few miles the coach stopped while the students wrote on cards an estimate of the compass direction toward the school. Then the driver turned through 135 degrees and continued east to a spot southeast of the school, where the students again estimated their direction. When the cards were analyzed, it turned out that the persons with brass bars by their heads had been able to sense the proper heading quite reliably, while those wearing magnets had not.

Gould and his Princeton co-worker K. P. Able recently tried to replicate Baker's experiment but failed. However, Baker's review of the attempt suggested that the volunteers' directional sense may have been thrown off by magnetic storms, weak magnetic gradients inside the bus, and/or the greater electromagnetic contamination of Princeton compared with rural England. Baker and co-worker Janice Mather have recently devised a simpler test method. In the middle of a specially built, light-tight wooden hut free of magnetic interference, the subject is blindfolded, earmuffed, and seated on a friction-free swivel chair. After being turned around several times, the subject must estimate his or her compass heading as before. With statistically consistent success in more than 150 persons, Baker believes he has proven the existence of a human magnetic sense.

Oddly enough, he finds that as long as people can't feel the sun or sense some other obvious cue, they can judge direction better with blind-folds on. Otherwise they start rationalizing the process, trying to deduce the right way from too little evidence and becoming confused. He surmises that the magnetic sense serves its purpose best by unconsciously giving a continuous sense of direction without its owner's having to be aware of it all the time, and thus freeing attention for the search for food, a mate, shelter, and so on.

In 1983, using magnetic measurements in selective-shielding experiments, Baker and his co-workers reported locating magnetic deposits close to the pineal and pituitary glands in the sinuses of the human ethmoid bone, the spongy bone in the center of the head behind the nose and between the eyes. It's interesting to note that selective-shielding studies done in the early 1970s, by Czech emigre biophysicist Zaboj Harvalik, an adviser to the U.S. Army Advanced Material Concepts Agency, pointed to this same spot as one of two areas—the other was the adrenal glands—where the dowsing ability resided.

In 1984 a group headed by zoologist Michael Walker of the University of Hawaii in Honolulu isolated single-domain magnetite crystals from a sinus of the same bone in the yellowfin tuna and Chinook salmon. The crystals were of a shape normally shown only by magnetite synthesized by living things rather than geological processes. Abundant nerve endings entered the magnetic tissue, and the crystals were organized in chains much like those in magnetotactic bacteria. Each crystal was apparently fixed in place but free to rotate slightly in response to external magnetic forces. Calculations showed that such chains would be able to sense the earth's magnetic field with an accuracy of a few seconds of arc, or a few hundred feet of surface position. This result correlated perfectly with earlier homing studies on live tuna by the same group. This detailed work, along with related earlier research, strongly suggests that all vertebrates have a similar magnetic organ in the ethmoid sinus area, and I suspect that this organ also transmits the biocycle timing cues from the earth field's micropulsations to the pineal gland.

The Face of the Deep

DNA pioneer Erwin Chargaff has called the origin of life "a subject for the scientist who has everything," but that hasn't stopped many of us (or even him, for that matter) from speculating about it. There are numerous detailed pictures of that primal scene in print today, but most are variations on one theory - "warm soup and lightning." Life on earth began some 4 billion years ago, or roughly l or 2 billion years after the world itself was born. The atmosphere then was completely different, much like Jupiter's today, mostly ammonia and methane. Into that atmosphere some source of energy—lightning, heat, and ultraviolet radiation have all been suggested—led to the spontaneous formation of simple organic compounds. Sifting into the oceans for millions of centuries, these compounds would have coalesced by chance into ever more elaborate patterns. According to the theory, this process culminated in closed "protocells" able to resist the reactivity of other structures while growing through incorporation of similar compounds.

This idea owes its dominance largely to an important experiment made by S. L. Miller in 1953. Miller pumped a facsimile of the presumed early atmosphere—ammonia, methane, and water vapor—continuously past an electric spark. After several days he had some amino acids. Since these are the bricks of DNA, RNA, and all proteins, the evidence seemed very good. Later runs yielded even more sophisticated molecules. In water they coalesced into globules with a sort of membrane around them—called "coacervates" by A. I. Oparin and "proteinoids" by Sidney Fox, two of the most assiduous students of biogenesis.

Nothing came close to being alive in any of these spark chambers, however. More important, the experiments raised two difficulties, one theoretical, one practical. The soup theory needed to come up with a very sophisticated entity, a living cell with some genetic system using DNA or RNA, right off the bat. According to our notions of biology, nothing could be alive before that point, yet it seemed incredible that chance associations of the building blocks could form a palace of such complexity without passing through a mud-hut stage.

When the warm soup theory was first advanced, the mechanistic persuasion was at its height. Living things were complicated machines, but molecular machinery they were. However, the concept of a cell was much simpler than it is today. No longer is it considered a mere baglet of jelly with a few master molecules telling it what to do. Even the membrane of the simplest bacterium responds in intricate ways to information from outside, yet our best electron microscopes haven't yet revealed a complexity of structure adequate to explain its work.

There was a basic chemical problem as well. All organic compounds exist in two forms, or isomers. Each contains the same number and type of atoms, but in one they're arranged as a mirror image of the other. One is "right-handed" and the other is "left-handed." They're identified by the way they bend light in solution. The dextrorotatory (D) forms rotate it to the right, while levorotatory (L) isomers refract it to the left.

All artificial methods of synthesizing organic compounds - including the spark experiments—yield a roughly equal mixture of D and L molecules. However, all living things consist of either D or L forms, depending on the species, but never both.

We must conceive of the first living things as something unexpected, not just simpler versions of what we see around us. They couldn't have been cells; they couldn't have had a DNA-RNA-protein system, a living membrane, or a nerve impulse network.

We can try to define the bare minimum, the processes that must be available before an entity can be called living. There must be a way to receive information about external conditions, process it, and store it so that the data change the being's response to the same stimulus in the future. In other words, a sort of crude consciousness and memory must be present from the first. A life-form must also be able to sense damage and repair itself. Third, we can expect that it would show some sort of cyclic activity, perhaps tuned primarily to the circadian rhythm of the lunar day. Self-replication, one of the main requirements in the DNA-based theory, can be dispensed with. An organism that can fully heal its injuries is theoretically immortal. The criteria for life can be summarized as organization, information processing, regeneration, and rhythm.

The funny thing is that all of these criteria are met by the activities of semiconducting crystals. Semiconductivity occurs naturally in several inorganic crystals, including silicon, one of the most common elements, and the rare earth germanium. Moreover, extremely small amounts of contaminants will change the crystal's electrical properties dramatically in doping. The earth's volcanic mixing would have produced minerals with a wide variety of current-handling abilities to start from. Most important, the piezoelectric, pyroelectric, photoelectric, and other responses of semiconducting crystals could have served as an analog method of processing and storing information about pressure, heat, and light. Moreover, repeated passage of current through some semiconductors permanently changes the materials' characteristics so as to make the same electrical responses easier in the future. Movement of electrons along the crystal lattices inevitably would have been shaped by geocelestial cycles in the earth's electromagnetic field, as well as by the fields around other such crystalline organisms nearby—providing a sense of time and information about the neighbors. The currents also would have instantly refllected any loss of material and guided the deposition of replacement atoms to restore the original structure.

The idea of certain rocks, in the course of a billion years or so, gradually becoming responsive to their surroundings, growing, learning to
"hurt" when a lava flow or sulfuric rain ate away part of a vertex, slowly

rebuilding, pulsing with, well, life, even developing to a liquid crystal stage and climbing free of their stony nests like Cadmus' dragon's teeth or the lizards in an M. C. Escher print—all this may seen a bit bizarre. Yet it's really no more strange than imagining the same transformation from droplets of broth. The change happened somehow.

The biggest hurdle for this theory is accepting the idea that life could develop in the dry state, either out of the oceans or in the rocks underneath them. Since the mid-1960s it has seemed more plausible, for it was then that H. E. Hinton, of the University of Bristol, England, learned that at least one organism spends part of its life completely without liquid water. Certain flies of the Sahara desert lay their eggs in the brief pools formed by the rare rains. The larvae go through several metamorphoses in the water, but they're almost always interrupted by the evaporation of the pool. Though completely desiccated, in a state Hinton named cryptobiosis, they survive months or years until the next rainstorm, whereupon they take up where they left off. The larvae can be quick-dried and stored in a vacuum bottle for many years. Placed in water, they resurrect in a few minutes. If a larva is cut in two when active, it takes six minutes to die. If it's flash-dried in the first minute, the two pieces can be kept on a shelf for years, but when returned to water they'll live out their remaining five minutes. Contrary to common sense, it appears that in this case life doesn't need water, but death can't occur without it.

Getting rid of the water-equals-life assumption makes the crystalline theory more believable. Conditions on the young planet favored forests of crystals: It was hot; volcanoes were constantly firing new materials into the dense, dark shell of turbulent gases. However, the crystals would still have needed outside energy to overcome the entropy of nonliving matter. With an organizing principle built into them from the start, it's not too hard to imagine them acquiring other kinds of molecules, including the organics raining from the sky and dissolving in the waters. Then life would have been on its way to developing the biochemistry we now know—the genetic system and the consequent appearance of sexuality—which is the basis of all the creatures now alive or known from the fossil record. Still, we need an energy source for the transition. Lightning won't work in this context. We also need an explanation for the exclusively left- or right-handed molecules.

In 1974 F. E. Cole and E. R. Graf of New Orleans made a theoretical analysis of the Precambrian earth's electromignetic field that fulfilled both needs. They reasoned that since the atmosphere was much larger then, it must have pushed the ionosphere much farther out than it is today, into the region of the Van Allen belts. The earth would then have had an electromagnetic resonator of two concentric spheres—the upper atmosphere and the surface. Today, as in the past, the earth's pulsing magnetic field combines with the solar wind to induce large currents in the Van Allen belts. In the Precambrian era, however, the fluctuations of current in the Van Allen belts in turn would have generated huge currents in the nearby ionosphere. Since the earth's metallic core is an excellent conductor, the ionospheric currents would have coupled to it, producing an enormous and constant electrical discharge through the atmosphere and into the earth. Moreover, since the distance around the core at that time was roughly equal to 1 wavelength of electromagnetic energy at 10 cycles per second, or about 18,000 miles, this discharge would have pulsed at 10 hertz throughout the resonant cavity, which encompassed the whole atmosphere and surface. Besides directly providing electrical energy, this discharge would have produced abundant heat, ultraviolet radiation, and infrasound (or pressure waves), all of which would have fostered varied chemical activity.

Such a dense and electrically supercharged atmosphere undoubtedly would have produced great quantities of amino acids and peptides. As they came together in the air and water, linking chainwise to form proteins and nucleic acids, the vectors of electromagnetic force would have favored spiral shapes twisting in one direction or the other, depending on whether the reaction occurred in the Northern or the Southern Hemisphere. In 1981, W. Thiermann and U. Jarzak found some direct evidence for this theory by synthesizing organic compounds in a steady-state magnetic field. Changing the orientation of the field gave them high yields of either D or L forms.

It may be possible to run a further check on the Cole and Graf hypothesis at one place in the solar system—the Great Red Spot of Jupiter. This permanent hurricane, whose eye could swallow the earth, continually emits prodigious electrical discharges through an atmosphere much like the one proposed for Precambrian times. It may be synthesizing organic compounds and energizing a transition to life even now.

On earth, all entities formed within the 10-hertz discharge—and all of their descendants—would resonate at the same frequency or show extreme sensitivity to it, even after the original power source had been disconnected. The 10-hertz band would remain supremely important for most life-forms, as indeed it has. As already noted, it's the primary frequency of the EEG in all animals, and it can be used to restore normal circadian rhythms to humans cut off from the normal fields of earth, moon, and sun. William Ross Adey of the Loma Linda VA Hospital in

California has found that magnetic fields modulated at about the same frequency can be used to change the behavior of monkeys in several important ways, which are described more fully in the next chapter.

The Cole and Graf theory also suggests how the spark of life turned itself off. The currents driving the discharges would have ceased as the atmosphere gradually became depleted by escape of the lighter gases and by incorporation of the ammonia and methane into organic compounds. As this happened, the ionosphere would have gradually descended, becoming disconnected from the Van Allen belts. The ionospheric currents would have become too small to couple to the earth's core, and the atmospheric cavity too small to resonate at the core's prescribed frequency. At that point, the plug was pulled, but life was well on its way. Aside from competition from more advanced creatures, the loss of the energy source would explain why we see today no remnants of the transitional forms still emerging from inanimate matter.

This solid-state theory of life's creation is more than an exciting picture of our birth in a shower of sparks. It leads us to another of biology's great mysteries—the evolution of nervous systems—by a sensible sequence of steps. First there would have been a crystalline protocell transmitting information directly through its molecular lattice. As the first cells developed, we can envision chains of microcrystals, then chains of organic polymers transmitting information in the form of semiconducting currents. Although the exact mechanism of electron passage through living tissue is far from clear, nearly all organic matter exhibits piezoelectricity and all the other hallmarks of semiconduction. Furthermore, in a series of experiments during the 1970s, Freeman Cope, a Navy biophysicist building on Szent-Gyorgyi's work, has found evidence of superconduction at room temperature in a variety of living matter. Currents briefly induced in superconductors have been known to flow for many years without decay, but the phenomenon has heretofore been attained only near absolute zero. Although Cope's work is still preliminary and uncorroborated, he has found electromagnetic data consistent with superconduction in E. coli bacteria, frog and crayfish nerves, yeast, sea urchin eggs, and molecules of RNA, melanin pigment, and the enzyme lysozyme.

Whatever the exact details of the conducting system, the first multicellular organisms probably had networks of cells that were much like the first single cells. Later, these network cells would have specialized for their DC-carrying duties, linking into syncytia to avoid the high resistance of intercellular junctions. Somewhere along the line a central processing center and information storehouse would have developed. At the same time separate input and output tracts would have appeared, and the DC system would have neared its peak of specialization as its cells evolved into the prototypes of glial, ependymal, and Schwann cells. At about this point the high-speed digital impulse system for handling more complex information would have begun to form inside the older one. Today all multicellular animals have this kind of hybrid system, whose complexities should provide work for at least a few more generations of neurophysiologists.

Crossroads of Evolution

The Cole and Graf theory has one crucial requirement. The polarity of the earth's magnetic field must have stayed the same during the resonant period. Otherwise there would be a mixture of right and left isomers in living tissues. As far as we can tell, the field did remain steady in Precambrian times, but we have ample proof that its poles have reversed many times during the last half-billion years. Each time, the shift has coincided with the extinction of many species.

The geomagnetic record is written in two places: in igneous rocks bearing magnetic minerals, and in ocean floor sediments. Magnetic particles in molten rock are free to move and align themselves with the prevailing magnetic field. As the rock cools they're frozen in place. In the same way magnetic particles settling onto the ocean floor reflect the orientation of the field at the time of the deposit. Ocean sediments and the rock they eventually become have given us an undisturbed magnetic chronology for many millions of years, while the relatively few strata of igneous rock undisturbed by later upheavals give us occasional glimpses further back.

The reversals happen very fast, as geologic time goes. The field strength falls to about half its average for a few thousand years. Then during another thousand years the poles change places; then the field regains its normal strength in another few thousand years. All told, the change takes about five thousand years.

In the early 1960s, when the reversals were first discovered, geophysicists believed the magnetic field disappeared completely during the pole reversal. Thus they thought that the absence of the electromagnetic umbrella that protected life from high energy ultraviolet and cosmic rays would account for large-scale extinctions, These "great dyings" had long puzzled paleontologists . Soon the demise of a species of radiolarian was correlated with a magnetic field reversal. Radiolarians are microscopic plankton animals with hard calcareous skeletons. Each species has a distinct, intricate shape, so their remains form an easily recognizable, continuous record in sediment cores. By 1967 James D. Hayes and Neil D. Opdyke of Columbia's Lamont-Doherty Geological Observatory had correlated the disappearance of eight types of radiolarians with the reversals. Each species had been widespread and abundant; each extinction took place abruptly, with no previous decline in population. The "radiation barrage" theory seemed confirmed.

However, it has since been learned that the field strength drops only by half, not enough to drastically reduce the protective power of the Van Allen belts and ionosphere. Moreover, radiolarian populations extend down into several yards of water, which should protect them from the radiation anyway. Hays has since drawn the less specific outlines of current knowledge thus: As animals grow more specialized in the course of evolution, they become more sensitive to some as yet unknown, lethal effect of the reversals. Long periods without reversals—the quiescent eras sometimes last tens of millions of years—seem to produce a profusion of species especially susceptible to the effect, and they're weeded out at the next shift.

We know of two especially widespread extinctions. One, at the end of the Permian period, about 225 million years ago, wiped out half the kinds of animals then alive, from protozoa to early reptiles. The same kind of curtain dropped on the age of dinosaurs at the end of the Cretaceous period, some 70 million years ago. In both cases frequent magnetic pole reversals had resumed after a long quiescence. Many periods of less extensive extinction have also been documented in the fossil record and correlated with the field reversals. Most recently, J. John Sepkoski, Jr., and David M. Raup of the University of Chicago reported what they believed to be a 26-million-year cycle in the major dyings. If their hypothesis holds up, there may be some solar or galactic influence that interacts with a magnetic reversal for maximum destructive effect.

We can only surmise that the earth's field was instrumental in life's beginning, but by 1971 we knew virtually for certain that its polarity shifts had shaped life's development by a "pruning" of species. That year I was invited to a private meeting at Lamont to talk about the reversals, the sole M.D. among a score of biologists and geophysicists. At that time we could only speculate as to how the extinction effect came about. We didn't even have a workable theory of what changes inside the earth caused the turnabouts, or how the process affected the micropulsations and other aspects of the field. All we could agree on was that there were probably changes in every aspect of it, and our knowledge hasn't progressed much since then.

The pole shift happens so slowly that living things may well adapt to it easily; the 50-percent decline in field strength also seems rather unimportant. However, since we know the micropulsations control biocycles, including the timing of the mitotic rhythm, a major change in their frequency could be catastrophic. Experiments with artificial extremely low frequency fields (see Chapter 15) have shown that vibrational rates near normal but slightly above, from about 30 to 100 hertz, cause dramatic changes in the cell cycle time. This interferes with normal growth of the embryo and may tend to foster abnormal, malignant growth as well. If a geomagnetic reversal raises the micropulsation frequencies into this range, the accumulation of growth errors over many generations could well mean extinction.

We have no way of making a forecast, however. Reversals seem to happen at widely varied intervals, as often as every fifty thousand years during some periods, many millions of years apart during other times. The last one seems to have occurred about seven hundred thirty thousand years ago. Several scientists have interpreted data from NASA's MAGSAT orbiter, and from measurements of magnetic particles in lake sediments, as indicating that the earth's magnetic field strength is steadily declining, and has been for the last few thousand years. If so, we may already be entering the next reversal, but it's also possible we're merely experiencing one of the field's many short-term variations.

Nor can we be sure how serious a reversal would be for us. Hominids have weathered them in the past, but we have an extra reason for being uneasy this time. If we're entering a reversal now, it will be the first one in which the normal field is contaminated with our own electromagnetic effluvia, and the most powerful of these, at 50 and 60 hertz, fall right in the middle of the "danger band" in which interference with growth controls can be expected.

The field giveth as well as taketh away, however. If we can hang on until the next peak of its strength, we may benefit from a subtle infusion of electromagnetic wisdom. An ingenious theory recently proposed by Francis Ivanhoe, a pharmacologist and anthropologist at two universities in San Francisco, suggests how important it may have been to our own development.

Ivanhoe made a statistical survey of the braincase volume of all known Paleolithic human skulls, and correlated the increase with the magnetic field strength and major advances in human culture during the same period. Ivanhoe found bursts of brain-size evolution at about 380,000 to 340,000 years ago, and again at 55,000 to 30,000 years ago. Both periods corresponded to major ice ages, the Mindel and Wurm, respectively, and they were also eras when great cultural advances were made—the widespread domestication of fire by Homo erectus in the early Mindel, and the appearance of Homo sapiens sapiens (Cro-Magnon peoples) and gradual decline of Neanderthals (Homo sapiens) during the Wurm. Two other glaciations in the same time span—the Ganz of about 1,200,000 to 1,050,000 years ago and the Riss of about 150,000 to 100,000 years ago—didn't call forth such obvious advancements in human evolution. They also differed from the other two in that the average geomagnetic field intensity was much lower.

Ivanhoe has proposed a direct link from the magnetic field through the growth-hormone regulator pathways in the brain to account for the sharp evolutionary gains. He suggests that part of the hippocampus, a section of the brain's temporal lobe, acts as a transducer of electromagnetic energy. A part of the hippocampus called Ammon's horn, an arch with one-way nerve traffic directed by a strong current flow, may read out variations in the field strength, feeding them by a bundle of well-documented pathways called the fornix to the hypothalamus and thence to the anterior pituitary, where growth hormone is produced. It's known that larger amounts of this hormone in pregnancy increase the size of the cerebral cortex and the number of its nerve cells in the offspring, as compared with other parts of the brain. Ivanhoe also notes that the hippocampus and its connections with the hypothalamus are among the parts of the brain that are much larger in humans than other primates. The idea gains further support from the fact that neural activity in the hippocampus increases with electrical stimulation and reaches a maximum at 10 to 15 cycles per second, at or slightly above the dominant micropulsation frequency of today's field. The most powerful shaper of our development may turn out to be the subtlest, a force that's completely invisible to us.

Hearing Without Ears

We've considered how the electromagnetic fields of earth, moon, and sun affect life. In the next chapter, we'll ponder the effects of artificial fields from our machines. There's probably another interaction, however, of which we know much less: the effects produced on living things by the biomagnetic fields of other creatures. If one nervous system could sense the field of another, it would go a long way toward explaining extrasensory perception.

Following the curious dogma that what we don't understand can't exist, mainstream science has dismissed psychic phenomena as delusions or hoaxes simply because they're rarer than sleep, dreams, memory, growth, pain, or consciousness, which are all inexplicable in traditional terms but are too common to be denied. Fifty years ago, when J. B. Rhine of Duke University first published results of his card-guessing experiments, scientists eagerly debated and tested the subject for a few years. Then, although at least 60 percent of the attempts to confirm Rhine's work also got better-than-chance results (a replication rate better than that in most other areas of psychology), the openness somehow disappeared. Ever since World War II, serious parapsychologists have been hounded out of the forum of science. In the 1950s, for example, Science and Nature both published attacks on certain results of Rhine and S. G. Soal, an early psi researcher at London University. Today this attitude may be waning. G. R. Price, the author of one of the diatribes, apologized in Science in 1972, and both journals have begun accepting occasional reports on psychic research, although still confining themselves mainly to negative findings. As the climate has begun to change, a few researchers have looked for electromagnetic fields as a possible basis for extrasensory perception.

The results so far have been as inconclusive as those from any other approach. In 1978 E. Balanovski and J. G. Taylor used a variety of antennae, skin electrodes, and magnetometers to monitor a number of people claiming paranormal powers. They found no electric or magnetic fields associated with successes in telepathy experiments. In 1982, Robert G. Jahn, dean of engineering at Princeton, assembled the most impressive battery of electronic equipment ever brought to bear on the subject. He found definite effects by mental forces on interferometer displays and strain-gauge readings, along with positive results in remote-viewing experiments. The tests couldn't be reliably repeated, however, and seemed to vary with the moods of researcher and subject, and perhaps with other immeasurable environmental factors. The same dependence on attitude—experiments seem to work more often for believers than doubters—has bedeviled psychic research from its beginning. Although Jahn came up with no clear-cut findings on electromagnetic factors, he was forced to the sublimely understated conclusion that ". . . once the illegitimate research and invalid criticism have been set aside, the remaining accumulated evidence of psychic phenomena comprises an array of experimental observations . . . which compound to a philosophical dilemma."

We must remember that our study of biofields is still in its infancy.

It's only a decade since the SQUID first enabled us to find the magnetic field around our heads at all. Pigeons have a magnetic detector thousands of times more sensitive than the latest instruments. We also know that the interaction of semiconducting currents with external magnetic fields is thousands of times greater than that of currents in a wire, and engineers have built microscopic devices that enhance this sensitivity by a factor of another thousand or more. The electron microscope has shown us crystallike structures of previously unsuspected complexity in all living cells, whose functions we can only guess at. There's now some evidence that psychic intent can influence the flow of current in solid-state devices, so we may be nearing the energy levels at which extrasensory factors work. Since all living things generate weak electromagnetic fields, and since many, if not all, can sense those of the earth, communication by this medium remains a strong possibility. Recent disclosure of a multimillion-dollar research effort in this area by the hardheaded weapons planners at the Department of Defense is one more reason why those scientists who work in public shouldn't dismiss the idea.

We must always be careful to place more weight on observation than current theory. We must remember that we don't yet fully understand magnetism. It now appears that the single domain with both magnetic poles may not be the smallest unit of magnetism after all. Physicists now posit the existence of magnetic monopoles, particles having the characteristics of just one pole, north or south. In fact there's some experimental evidence for them. Some theoreticians go even further, envisioning a hitherto unsuspected kind of magnetism, a composite of waves and monopole particles, like light. Living things may interact with such a now immeasurable energy.

Any such message system would have at least two major difficulties to overcome in the course of evolution. Our own electrical-engineering experience, however, suggests workable approaches life may have taken.

One problem is that the strength of biofields is far below that of the earth's field. Hence any input from other creatures would be embedded in noise. This is a common obstacle to telecommunications, and there are several ways around it. The easiest is for sender and receiver both to be frequency locked, that is, tuned to one frequency and insensitive to others. Such a lock-in system might explain why spontaneous ESP experiences most often happen between relatives or close friends. The sensitivity of our instruments may someday develop to the point where we can tune in to biomagnetic fields on select frequencies, thus experimenting as directly with ESP as we now do with radio.

Another theoretical difficulty is the fact thatl psychic transmission doesn't seem to fade with distance. The electromagnetic field around an animal's nervous system, on the other hand, starts out unimaginably small and then diminishes rapidly. However, extremely low frequency (ELF) transmissions have a peculiar property. Because of their interaction with the ionosphere, even weak signals in this frequency range (from 0.1 to 100 cycles per second) travel all the way around the world without dying out. If an innate frequency selector is operating within this band, reception should be the same anywhere on earth.

At this time the DC perineural system and its electromagnetic fields provide the only theory of parapsychology that's amenable to direct experiment. And it yields hypotheses for almost all such phenomena except precognition. Telepathy may be transmission and reception via a biologically programmed channel of ELF vibrations in the perineural system's electromagnetic field. Dowsing may involve an unconscious sense of the electromagnetic fields of underground water or minerals, an idea given some support by Russian experiments in the 1960s. Nikolai N. Sochevanov, now with the USSR's Ministry of Geology, found that the accuracy of forty professional dowsers diminished by at least three fourths when he wound a current-carrying wire around their wrists or brought a horseshoe magnet near their heads.

Biological semiconductors even offer a possible basis for the aura often reported around living things by "sensitives." There has long been speculation that this "halo" might be some manifestation of an electromagnetic biofield. The ability of high-voltage (Kirlian) photography to produce an image very much like descriptions of the aura has aroused hope that the technique might render some aspect of psychic phenomena visible in a way that would be conducive to experiment. Because of this possibility, our lab investigated Kirlian photography during the mid-1970s.

We obtained beautiful pictures that seemed to vary in response to changes in the health of the test organism. However, the method failed one crucial test. If the Kirlian halo actually reflected the biofield or some other basic aspect of life, it should have disappeared when the organism being photographed died. Alas, it did not. The image remained the same as long as the water content of the corpse remained constant. We found the images were entirely due to a simple physical event, a corona discharge. This occurred when a high-voltage electric field broke down the air molecules between the two condenser plates of the Kirlian apparatus. The amount of water vapor in the air changed the voltage at which this happened, and on color film produced coronas in different colors and sizes. We found no evidence that the Kirlian image was related to the living state. Nor did we find that it could serve as a "screen" on which might be reflected some invisible field or aura, another possibility that had been suggested.

This is not to say that the aura occasionally perceived by some people around other organisms is imaginary. Things that appear so often in folklore often turn out to have a basis in fact. However, the body's magnetic field is far too weak to account for it. Our biofields, even if they were many times stronger, couldn't possibly emit light, but an appropriately sensitive magnetic detector in the brain, if it had nerve connections to the visual cortex, might "see" the magnetic field, in a manner of speaking. In a similar way astronauts in space "see" Cerenkov radiation—flashes of light that have been traced to the passage of high-energy cosmic rays through the retina.

On the other hand, the aura could literally be a form of light, perhaps at frequencies invisible to all but a few of us. The discovery of light-emitting diodes is still fairly recent. As you will recall, we found that bone happens to have such properties. The point of that experiment was its evidence that bone contains semiconducting PN junction diodes. There may well be other diodes in living things. The relationship between the nerve endings and the skin is an interesting one in this context. The skin-nerve interface—the closest normal equivalent to the neuroepidermal junction that triggers regeneration—may well be a diode. If so, the proper level of current could cause emission of light from the skin. It's even possible that such an array of diodes with very large currents might produce a holographic image of the body on an organic screen, such as the reputed image of Christ on the Shroud of Turin.

If extrasensory communication really is a function of the DC system, why isn't it more common and widely accepted? We may never know how well distributed it is among animals, although the number of pets who have returned to their owners over long distances suggests that many dogs and cats can find specific people by an unknown sense. The Duke University Parapsychology Laboratory has authenticated more than fifty such cases, many involving travel of hundreds or thousands of miles. We can expect that some species would be better at it than others, just as pigeons navigate by the earth's magnetic field far better than most other creatures. Among humans, some may simply be more gifted than others through genetic chance or some facet of their upbringing. Then again, the psychic sense may be a universal ability that was forgotten or suppressed as we came to depend more and more on language to get our messages across.

If they do depend on the same system, psychic ability and regeneration may go together; they may generally be better among simple animals. As the digital impulse system grew more efficient, its information may have overwhelmed the senses operating through the earlier mode. In fact, this may be part of the digital system's purpose. The ever-present hum of electromagnetic information from other creatures may have become an intolerable burden. Think how confused you would feel if you could simultaneously hear what everyone else in the world was thinking. After all, mediums, sorcerers, and psi experimenters all agree that some sort of trance or mental quietude—a reduction of nerve impulse activity—is needed for best results. According to Elmer Green, yogis of some Tibetan traditions teach clairvoyance to novices by having them meditate seated on a glass plate, facing north toward a sheet of polished copper in a dark, windowless room, with a bar magnet suspended over their heads, its north pole pointing up to the zenith.

The biofield also lends itself to theories of psychokinesis and object imprinting. All matter, living and nonliving, is ultimately an electromagnetic phenomenon. The material world, at least as far as physics has penetrated, is an atomic structure held together by electromagnetic forces. If some people can detect fields from other organisms, why shouldn't some people be able to affect other beings by means of their linked fields? Since the cellular functions of our bodies are controlled by our own DC fields, there's reason to believe that gifted healers generate supportive electromagnetic effects, which they convey to their patients or manipulate to change the sufferer's internal currents directly, without limiting themselves to the placebo effect of trust and hope.

Once we admit the idea of this kind of influence, then the same kind of willed action of biofields on the electromagnetic structure of inanimate matter becomes a possibility. This encompasses all forms of psychokinesis, from metal-bending experiments in which trickery has been excluded to more rigidly controlled tests with interferometers, strain gauges, and random number generators. At present, it's the only hypothesis that offers much hope of testability. On a less spectacular level, we must ask whether the biofield can project the individual signature of a person's thoughts onto his or her surroundings, changing the electromagnetic characteristics of these objects so that the person can be sensed by others even though absent. This may well be the commonest of all paranormal experiences, and the number of crimes solved by psychics reacting to the mere scene of the crime should entitle scientists to investigate the idea without fear of ridicule from their colleagues.

Over and over again biology has found that the whole is more than the sum of its parts. We should expect that the same is true of bioelectromagnetic fields. All life on earth can be considered a unit, a glaze of sentience spread thinly over the crust. In toto, its field would be a hollow, invisible sphere inscribed with a tracery of all the thoughts and emotions of all creatures. The Jesuit priest and paleontologist-philosopher Pierre Teilhard de Chardin postulated the same thing, a noosphere, or ocean of mind, arising from the biosphere like a spume. Given a biological communications channel that can circle the whole earth in an instant, possibly based on life's very mode of origin, it would be a wonder if each creature had not retained a link with some such aggregate mind. If so, the perineural DC system could lead us to the great reservoir of image and dream variously called the collective unconscious, intuition, the pool of archetypes, higher intelligences deific or satanic, the Muse herself.

Fifteen

Maxwell's Silver Hammer

In considering questions as remote as the origin of life, science must skate toward new shores across the thin ice of speculation, but it also has a duty to warn us of present dangers as specifically as possible. Since the earth's electromagnetic activity has such a profound effect on life, the obvious question is: What are the consequences of our artificial energies?

Electromagnetism can be discussed in two ways—in terms of fields and in terms of radiation. A field is "something" that exists in space around an object that produces it. We know there's a field around a permanent magnet because it can make an iron particle jump through space to the magnet. Obviously there's an invisible entity that exerts a force on the iron, but as to just what it consists of—don't ask! No one knows. A different but analogous something—an electric field—extends outward from electrically charged objects.

Both electric and magnetic fields are static, unvarying. When the factor of time is introduced, by varying the intensity of the field as in a radio antenna, an electromagnetic field results. As its name implies, this consists of an electric field and a magnetic field. The fluctuations in the field radiate outward from the transmitter as waves of energy, although somehow these waves simultaneously manage to behave as streams of massless, chargeless particles (photons). As to just how this happens, again—don't ask! Sometimes the phenomenon is called an electromagnetic field (EMF), to emphasize its connection with the transmitter; sometimes it's called electromagnetic radiation (EMR), to emphasize its outward-flowing aspect. However, the two terms refer to the same phenomenon and are interchangeable. The only meaningful distinction is between static and time-varying fields.

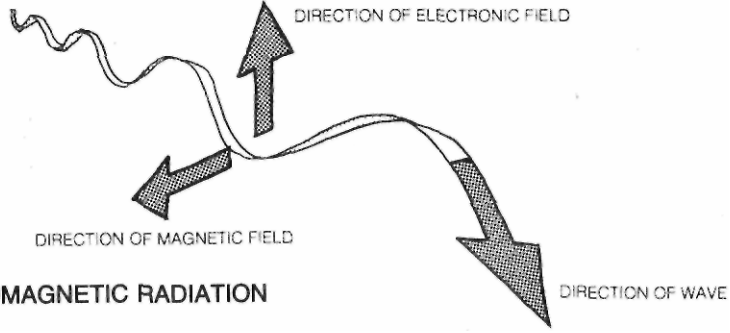

ELECTROMAGNETIC RADIATION

Each energy wave consists of an electric field and a magnetic field at right angles to each other, and both at right angles to the direction the wave is traveling. The number of waves formed in one second is the frequency; the distance the energy travels (at the speed of light) during one oscillation is its wavelength. The higher the frequency, the shorter the wavelength, and vice versa.

EMR spans an enormous range of frequencies. The shortest gamma rays, a tenth of a billionth of a millimeter long, vibrate sextillions of times a second. These, along with X rays and the shortest ultraviolet wavelengths, are termed ionizing radiation, because their high photonic energy can knock electrons away from atoms, creating highly reactive ions where they don't belong. Much of the damage from nuclear radiation is caused in this way. All lower frequencies, beginning with the longer ultraviolet wavelengths, are nonionizing.

Next comes the only energy we can see—the narrow band of visible light vibrating hundreds of trillions of times a second—and then the infrared waves we feel as radiant heat. Below these lie the waves we've harnessed for communication. They begin with microwaves (MW), whose frequency is measured in gigahertz or megahertz—billions or millions of cycles per second—and extend through the radio frequencies (RF) down to ELF waves, whose frequency converges on zero. The MW and RF spectrum is arbitrarily broken up into a further alphabet of extremely high, superhigh, ultrahigh, very high, high, medium, low, very low, and extremely low frequencies (EHF, SHF, UHF, VHF, HF, MF, LF, VLF, and ELF respectively). As we have seen, ELF waves approximate the dimensions of the earth; at 10 hertz one wave is about 18,600 miles long.

Except for light and infrared heat, we can't perceive any of these energies without instruments, so most people don't realize how drastically and abruptly we've changed the electromagnetic environment in just one century. Working at Cambridge University, Scottish physicist James Clerk Maxwell showed mathematically in 1873 that light was but a small part of the vast undiscovered realm of radiation. Heinrich Hertz first found some of the radio waves in 1888. Meanwhile, Edison had set up the first commercial electric-power system in New York in 1882.

For billions of years before then, the energies that life grew up among were relatively simple. There was a weak electromagnetic field modulated by micropulsations within it and further sculpted by the solar and lunar cycles. There was a burst of static centered at 10,000 hertz and reverberating over the whole earth whenever lightning flashed in the scores of thunderstorms in progress at any one time. There were a few weak radio waves from the sun and other stars. Light, including some infrared and ultraviolet, was the most abundant form of electromagnetic energy. At higher frequencies, living things absorbed only small amounts of ionizing X rays and gamma rays from space and from radioactive minerals in rocks. Large parts of the energy spectrum were totally silent.

We'll never experience that quiet world again. In 1893 Nikola Tesla lit the Chicago World's Fair with the first AC power system, and two years later he began the modern era of electrical engineering by harnessing Niagara Falls. In 1901 Guglielmo Marconi sent a radiotelegraph message across the Atlantic, using without acknowledgment a machine designed by the prolific Tesla. The invention of the vacuum tube in 1907 led to the first voice transmission by radio in 1915 and the first commercial station in 1920. Until then many people still ate supper by candle or kerosene, and the ambient forces remained a reasonable facsimile of earth's pristine field.

The greatest changes have all come in the one generation since World War II. The trend toward use of shorter and shorter radio waves, bounced off the ionosphere for long-distance communication, had begun before the war. The fight for survival against fascism impelled the development of microwave radar, which helped win the Battle of Britain, allowed all-weather bombing of Germany, and gave the American Navy a decisive edge over the Japanese. The conflict also produced other electronic devices of all types. In 1947 Bell Telephone set up the first microwave phone relay towers between New York and Boston, the same year the first commercial television broadcasts, also transmitted by microwaves, began. Since then nearly every human involved an electrical appliance, and today we're all awash in a sea of energies life has never before experienced, of which the following list of sources only skims the surface:

Everything that runs on a battery produces a DC magnetic field—from digital watches, cameras, flashlights, and portable radios to car ignition systems.

Strong magnetic fields are used in industry to refine ore, concentrate and recycle scrap iron, purify sewage, soften water for steam boilers, and many other tasks.

The starting and stopping of an electric train turns the power rail into a giant antenna that radiates ELF waves for over 100 miles.

Electromagnetic fields vibrating at 60 hertz (50 hertz in Europe and Russia) surround nearly every person on earth from appliances at home and machines at work.

Over 500,000 miles of high-voltage power lines crisscross the United States. Innumerable smaller lines feed into every home, office, factory, and military base, all producing AC or DC fields. Metal objects near the lines concentrate the fields to higher levels. In addition, high-voltage lines are, in effect, gigantic antennae operating at 60 hertz in the ELF band, the largest "radio" transmitters in the world. Switching stations, where the current is changed from one voltage or type to another, emit radio-frequency waves as well.

AC magnetic fields vibrating at 100 to 10,000 hertz emanate from antitheft systems in stores and libraries, and from metal detectors in airports.

Low-frequency radio waves are used for air and sea navigation, time references, emergency signals, some amateur radio channels, and military communications.

Medium frequencies between 535 and 1,604 kilohertz are reserved for AM radio transmitters, which are limited to 50,000 watts in this country but are sometimes much more powerful abroad.

HF and VHF channels are filled with chatter from the nation's 35 million CB radios, as well as shortwave bands for more ham radios, air and sea navigation systems, military uses, spy satellites, and police and taxi radios. VHF television and FM radio also inhabit this region. There are now over ten thousand commercial radio and TV stations in the United States alone, and 7 million other radio transmitters, not counting the millions operated by the military.

Weather satellites, some kinds of radar, diathermy machines, upward of 10 million microwave ovens, more cop and cab radios, automatic garage-door openers, highway emergency call boxes, and UHF television compete for the low microwave frequencies.

Higher microwave bands are crowded with more military talk channels and radar, navigational beacons, commercial communications satellites, various kinds of walkie-talkies, and America's two hundred fifty thousand microwave phone and TV relay towers.

Like the infrared rays above them in the spectrum, radio waves and microwaves produce heat when directed in high-intensity beams. Hence they're used for all sorts of industrial chores—bonding plywood, vulcanizing rubber, manufacturing shoes, sterilizing food, making plastics, and heat sealing the trillions of plastic-wrapped products in our stores, even opening oysters. Modern electronics would be impossible without the perfect silicon and germanium crystals grown in microwave furnaces.

The human species has changed its electromagnetic background more than any other aspect of the environment. For example, the density of radio waves around us is now 100 million or 200 million times the natural level reaching us from the sun. Nor is there any end in sight. When superconducting cables are introduced, they'll increase the field strength around power lines by a factor of ten or twenty. Electric cars, magnetically levitated transport vehicles, and microwave-beam satellites for transmitting solar power to earth would each add strong new sources of electromagnetic contamination. A proposed electromagnetic catapult that could shoot satellites into space from mile-long rails built up the side of a mountain would require the combined output of the country's thousand generating stations for the few seconds of each launch.

A few years ago most investigators believed that each wavelength interacted mainly with objects comparable to it in size. This was a comforting notion that theoretically limited each frequency to one type of effect and predicted that really troublesome problems for humans would come from only one portion of the spectrum—the FM band. Now, however, we know there are primary effects on all life-forms at ELF frequencies, and in other parts of the spectrum there can be consequences for specific systems at any level, from the subatomic to the entire biosphere as a unit.

Of course, a change at one level may well trigger secondary changes throughout an organism, so that the original one is hard to identify. Moreover, the impact of EMR ar any particular frequency is often related to its power density, the amount of energy streaming through a certain area. When discussing biological effects, this is best measured in microwatts (millionths of a watt) per square centimeter, a unit we'll simplify to microwatts. There's often no direct relationship between dose and effect, however; a low power density sometimes does things that a higher one does not. Furthermore, we can't tell how much energy from a given power density is actually absorbed, or what part of the body receives it. The same holds true for electric and magnetic fields, whose study is further complicated by the fact that animals of different shapes distort the fields differently. Likewise, fur, feathers, skin thickness, bone size, and the general shape of an animal complicate RF and MW absorption beyond our capacity to gauge it. Therefore, reactions seen in one species cannot be assumed for another. The only way to test for possible damage (or beneficial effects) is to actually do the experiment.

In a sense, the entire population of the world is willy-nilly the subject of a giant experiment. Electropollution has been the subject of heated public debate for nearly ten years, and unpublicized misgivings for decades before that. Unfortunately, the question of risk has been asked too late. Daily exposure of nearly everyone is a fait accompli.

Subliminal Stress

After Howard Friedman, Charlie Bachman, and I had found evidence that "abnormal natural" fields from solar magnetic storms were affecting the human mind as reflected in psychiatric hospital admissions, we decided the time had come for direct experiments with people. We exposed volunteers to magnetic fields placed so the lines of force passed through the brain from ear to ear, cutting across the brainstem-frontal current. The fields were 5 to 11 gauss, not much compared with the 3,000 gauss needed to put a salamander to sleep, but ten to twenty times earth's background and well above the level of most magnetic storms. We measured their influence on a standard test of reaction time—having subjects press a button as fast as possible in response to a red light. Steady fields produced no effect, but when we modulated the field with a slow pulse of a cycle every five seconds (one of the delta-wave frequencies we'd observed in salamander brains during a change from one level of consciousness to another), people's reactions slowed down. We found no changes in the EEG or the front-to-back voltage from fields up to 100 gauss, but these indicators reflect major alterations in awareness, so we didn't really expect them to shift.

We were excited, eagerly planning experiments that would tell us more, when we came upon a frightening Russian report. Yuri Kholodov had administered steady magnetic fields of 100 and 200 gauss to rabbits and found areas of cell death in their brains during autopsy. Although his fields were ten time as strong as ours, we stopped all human experiments immediately.

Friedman decided to duplicate Kholodov's experiment with a more detailed analysis of the brain tissue. He made the slides and sent them to an expert on rabbit brain diseases, but coded them so no one knew which were which until later.

The report showed that all the animals had been infected with a brain parasite that was peculiar to rabbits and common throughout the world. However, in half the animals the protozoa had been under control by the immune system, whereas in the other half they'd routed the defenders and destroyed parts of the brain. The expert suggested that we must have done something to undermine resistance of the rabbits in the experimental group. The code confirmed that most of the brain damage had occurred in animals subjected to the magnetic fields. Later, Friedman did biochemical tests on another series of rabbits and found that the fields were causing a generalized stress reaction marked by large amounts of cortisone in the bloodstream. This is the response called forth by a prolonged stress, like a disease, that isn't an immediate threat to life, as opposed to the fight-or-flight response generated by adrenaline.

Soon thereafter, Friedman measured cortisone levels in monkeys exposed to a 200-gauss magnetic field for four hours a day. They showed the stress response for six days, but it then subsided, suggesting adaptation to the field. Such seeming tolerance of continued stress is illusory, however. In his pioneering lifework on stress, Dr. Hans Selye has clearly drawn the invariable pattern: Initially, the stress activates the hormonal and/or immune systems to a higher-than-normal level, enabling the animal to escape danger or combat disease. If the stress continues, hormone levels and immune reactivity gradually decline to normal. If you stop your experiment at this point, you're apparently justified in saying, "The animal has adapted; the stress is doing it no harm." Nevertheless, if the stressful condition persists, hormone and immune levels decline further, well below normal. In medical terms, stress decompensation has set in, and the animal is now more susceptible to other stressors, including malignant growth and infectious diseases.

In the mid-1970s, two Russian groups found stress hormones released in rats exposed to microwaves, even if they were irradiated only briefly by minute amounts of energy. Other Eastern European work found the same reaction to 50-hertz electric fields. Several Russian and Polish groups have since established that after prolonged exposure the activation of the stress sytem changes to a depression of it in the familiar pattern, indicating exhaustion of the adrenal cortex. There has even been one report of hemorrhage and cell damage in the adrenal cortex from a month's exposure to a 50-hertz, 130-gauss magnetic field.

Soviet biophysicist N. A. Udintsev has systematically studied the effects of one ELF magnetic field (200 gauss at 50 hertz) on the endocrine system. In addition to the "slow" stress response we've been discussing, he found activation of the "fast" fight-or-flight hormones centering on adrenaline from the adrenal medulla. This response was triggered in rats by just one day in Udintsev's field, and hormone levels didn't return to normal for one or two weeks. Udintsev also documented an insulin insufficiency and rise in blood sugar from the same field.

One aspect of the syndrome was very puzzling. When undergoing these hormonal changes, an animal would normally be aware that its body was under attack, yet, as far as we could tell, the rabbits were not. They showed no outward signs of fear, agitation, or illness. Most humans certainly wouldn't be able to detect a 100-gauss magnetic field, at least not consciously. Only several years after Friedman's work did anyone find out how this was happening.

In 1976 a group under J. J. Noval at the Naval Aerospace Medical Research Laboratory at Pensacola, Florida, found the slow stress response in rats from very weak electric fields, as low as five thousandths of a volt per centimeter. They discovered that when such fields vibrated in the ELF range, they increased levels of the neurotransmitter acetylcholine in the brainstem, apparently in a way that activated a distress signal subliminally, without the animal's becoming aware of it. The scariest part was that the fields Noval used were well within the background levels of a typical office, with its overhead lighting, typewriters, computers, and other equipment. Workers in such an environment are exposed to electric fields between a hundredth and a tenth of a volt per centimeter and magnetic fields between a hundredth and a tenth of a gauss.

Power Versus People

Because industry and the military demand unrestricted use of electromagnetic fields and radiation, their intrinsic hazards are often compounded by secrecy and deceit. I learned this lesson in my first encounter with the environmental review process.

As we were investigating the EMF-stress connection in 1969, the Navy decided to build a giant antenna in northern Wisconsin. The plan, called Project Sanguine, was to establish a radio link with nuclear submarines at their normal depth of 120 feet or below. Conventional radio signals couldn't pass through water, so the vessels had to surface or else cruise very slowly a few feet under and communicate by means of a floating antenna at prearranged times. Since this made the subs temporarily vulnerable, the Navy wanted a message system using ELF waves, which penetrate earth and water.

The original design involved 6,000 miles of buried cable arranged in a grid across the upper two fifths of Wisconsin. A transmitter would pump current into one side; the electricity would emerge from the other side and complete the circuit by traveling through the ground. The device was actually a giant loop antenna using the earth as part of the loop. ELF waves issuing from it and resonating between the earth's surface and the ionosphere could be picked up anywhere on the globe.

Sanguine was one of the first military projects scrutinized under the Environmental Protection Act. In 1973 the Navy set up a committee of scientists to review fifteen years of naval research on ELF effects, as well as other pertinent work. Captain Paul Tyler of the Office of Naval Research asked me to be one of its seven members.

The only thing sanguine we found was the name. While the research to date didn't prove there would be grave harm to human health, it showed several dangers. The antenna would produce an electromagnetic field 1 million times weaker than that from a 765-kilovolt power line. It was to broadcast at 45 to 70 hertz, frequencies close enough to the earth's micropulsations that living things are very sensitive to them. Similar fields had been shown to raise human blood triglyceride levels (often a harbinger of stroke, heart attack, or arteriosclerosis), and change blood pressure and brain wave patterns in experimental animals. The generalized stress response, desynchronized biocycles, and interference with cellular metabolism and growth processes—and hence increased cancer rates—were also distinct possibilities. Hundreds of thousands of people would be living inside the antenna even in this sparsely populated area; long-term effects on plants and animals were unknown; and, because the signals would resonate throughout the world, the biohazards might be similarly widespread. For these reasons we unanimously recommended that the project be shelved pending answers to the ominous questions it raised. We provided a long list of necessary research, emphasizing further tests on triglycerides, biorhythms, stress, and psychological responses to ELF fields. We also warned that the health of a large part of the U.S. population might already be impaired by 60 hertz power lines carrying vastly more power than the proposed antenna.

The committee met on December 6 and 7, 1973, generating a report then and there, with a secretary taking down our conclusions. The Navy group in charge was apparently displeased with our findings, The printed proceedings, marked "For official use only," went out only to committee members, and the Navy refused to discuss them with anyone else.

As soon as I got back from Washington, I found that two power companies were planning a network of 765-kilovolt power lines linking nuclear reactors in upstate New York and Canada. One of the lines was to pass through a rural area near the village of Lowville, where I'd just bought land for a vacation-retirement home. I immediately wrote the head of the state's Public Service Commission. Without releasing the Sanguine report—I felt it wasn't my place to do so, even though its suppression was wrong—I informed PSC Chairman Alfred Kahn of its major conclusions. The commission in turn asked the Navy for a copy of our report but was turned down flat. In mid-1974, however, Andy Marino and I were asked to testify at PSC hearings on the power lines.

We presented the best evidence then available, some of which seemed to shock the PSC members. ELF fields at power line intensity or less had by then been linked to bone tumors in mice, slowed heartbeat in fish, and various chemical changes in the brain, blood, and liver of rats. Bees exposed to a strong ELF field for a few days in Russian research had begun to sting each other to death or leave the area. Some sealed off their hives and asphyxiated themselves. Attorneys for the power companies hurriedly asked a year's postponement of the hearings, which the PSC naturally granted.

Andy and I spent that year reading the rapidly accumulating scientific literature on EMF biological effects, including the enormous amount of Russian work becoming available in English. Andy also investigated the stress response further. He ran ten separate experiments with rats, exposing them for one month to 60-hertz electric fields of 100 to 150 volts per centimeter, simulating ground level underneath a typical high-tension line. Three generations of rats bred in this field showed severely stunted growth, especially among males. At lower field strengths (35 volts per centimeter) some of the animals gained more weight than controls, a response we tentatively traced to abnormal water retention, which, like underweight, could also result from stress. A few years later, a study commissioned by the Department of Energy to duplicate this research also produced contradictory but disquieting results. With every known variable controlled in an expensive, high-tech facility at Battelle Laboratories in Columbus, Ohio, one test showed severe growth retardation over three generations, while a second run under exactly the same conditions produced significantly greater weight gain than normal.

Andy's original work also revealed large increases in the infant mortality rate. Between 6 and 16 percent of the pups born in various tests failed to live to maturity because of the electric field. That is, these percentages were in excess of the normal death rate for newborn rats.

Various other symptoms consistent with stress were found, including decreased water intake, enlarged adrenal and pituitary glands, and altered protein and hormone ratios in the blood. There was also a very high incidence—ten in sixty—of glaucoma in the early experiments. The disease didn't show up in later runs from which we excluded animals having observable eye defects, suggesting that the electric field had worsened a preexisting problem rather than causing it.

We expected the utilities to roll out their heavy artillery when the PSC hearings resumed, but we were still unprepared for what actually happened. The companies had hired two microwave researchers, Herman Schwan and Solomon Michaelson, both of whom did most of their work for the Department of Defense, and University of Rochester botanist Mort Miller. Carefully prepared by these three, the company lawyers cross-examined us for seventeen days in December 1975, attacking not only our methods and results but our scientific competence and honesty as well. Michaelson strenuously denied that our rodents had shown signs of stress, even though the biological markers were clear. Even if they had, he contended, stress could be healthful, an idea that Hans Selye later called "farfetched" when applied to a biological challenge that was continuous and not self-imposed.

As far as I know, our testimony was the first ever openly given by American scientists stating that electromagnetic energy had health effects in doses below those needed to heat tissue, and that power lines might therefore be hazardous to human health. We criticized the White House Office of Telecommunications Policy for failing to follow up a tentative 1971 warning by advising the President that some harmful effects from electropollution were now proven. Moreover, although we didn't realize it at the time, we greatly embarrassed Captain Tyler and the Navy by publicly revealing the existence of the Sanguine report, which had been secret until then.

Among those who heard about it was Wisconsin Senator Gaylord Nelson, who was understandably furious that his constituents were even then being used as guinea pigs in ongoing ELF tests at an experimental station near Clam Lake, while the document gathered dust in some Navy safe. Quoting Andy and me, he soundly criticized the Navy on the Senate floor. Due to local opposition, the Navy had already moved the site of the full-scale antenna to Mulligan's upper peninsula, modifying the design and giving it a new name—Project Seafarer. Nelson's fury now induced the Secretary of the Navy to ask the National Academy of Sciences for further study of the environmental questions. Harvard's biology chairman, Woodland Hastings, who was picked to head the NAS committee, wrote Marino a flattering letter asking for his consultation when the members got down to work. Marino then called Hastings to tell him of the large body of data we'd assembled for the power line hearings, and to make sure the NAS body would be willing to consider it thoroughly. Hastings told him, "Heck, you guys will be on the committee."

Soon the sixteen members were announced, and we were nowhere in sight. Hastings later publicly called us quacks, but to us he said the Navy had specified who was to be on the panel, despite his threats to quit if Andy and I weren't admitted. We were well acquainted with three of the men who were on it: Schwan, Michaelson, and Miller. Obviously they weren't about to find hazards in Seafarer after testifying that a much stronger power line was perfectly safe. They remained on the committee even though all three neglected to mention their New York testimony on NAS conflict-of-interest forms. The rest of the committee was also stacked with people who routinely discounted any evidence of health effects from low-level EMFs.

The NAS committee took an inordinately long time to issue its report, but we eventually saw a reason for the delay. During the PSC hearings, all evidence was subject to cross-examination. Besides questioning the witnesses, each side could look at the other's papers, including the actual workbooks of experiments. After the testimony, while the commission, assisted by a panel of judges, was deliberating, other evidence could be introduced, but it was no longer subject to review by the opposing side. Oddly enough, the NAS report—which constituted a defense of the then current dogma and tried to discredit most of the disturbing evidence—appeared just after the gavel sounded to close the PSC hearings. It was immediately introduced as evidence, and we couldn't say a word about it.

Six years later a Navy spokesman explained to me what had "really" happened. He said the Secretary of the Navy had gone to NAS and arranged to pay for the work. Then, when the members of the committee were announced, the Secretary and other Navy brass agreed that the show was rigged. The Secretary protested to NAS and said the Navy wouldn't pay for the study. NAS said that since the authorizations had already been signed, the Navy would have to pay for it. Moreover, the Navy needed a report in four to six months. Of course, NAS had been planning to wait till the end of the New York PSC hearings, which dragged on and on. My informant told me that, in response to Navy pressure, NAS laid in effect, "Go away. We've got the money, and the study is out of your hands. We'll run it our way." By the time the report was issued it was too late for the Navy people to use it, and they considered it too biased to have any value anyway. However, I don't put much faith in this bit of blame shifting.

The PSC's panel of judges spent nearly a third of their advisory opinion attacking Marino's work and his "argumentative" demeanor at the witness table. Via a Freedom of Information Act request, Andy later found that the technical parts of this opinion had been written by one of the judges' paid consultants, Asher Sheppard, then a researcher at UCLA. Sheppard was at that time preparing a monograph, The Biological Effects of Electric and Magnetic Fields of Extremely Low Frequency, under contract to the American Electric Power Company. He concluded that there were no significant biological effects from low to moderate-intensity ELF fields such as occurred around power lines and appliances, despite the fact that he'd been working under W. Ross Adey, whose career has been devoted to studying just such effects.

Nevertheless, we won. The Public Service Commission specifically contradicted its judicial advisers, commending Marino as a valuable witness, and adopted most of our recommendations. One line already under construction was built, largely because New York Governor Hugh Carey threatened to dissolve the PSC if the commission stopped it, but the utilities were ordered to buy additional land for a wider safety zone along the right of way. They were also forced to invest $5 million in a five-year research program administered by the New York State Department of Health, and to stop encouraging multiple use of the land under power lines, such as leasing it for playgrounds. An additional six or seven proposed lines have been postponed indefinitely. Most important was the plain fact that we raised the issue successfully against great odds and secured a health-conscious verdict from the PSC, gaining time to gather more facts about the dangers.

The Navy's ELF antenna has also been on hold for many years. Seafarer lost momentum when 80-percent opposition in two 1976 referenda in Michigan's Upper Peninsula forced then candidate Jimmy Carter to oppose it publicly for a while. Once more renamed and redesigned, Project ELF has been heavily funded by the Reagan regime with an eye toward expansion. The first step now would consist of a 56-mile aboveground antenna carried on intersecting rows of utility poles in two corridors cut out of the Escanaba River State Forest. In July 1983, the Michigan Natural Resources Commission voted to allow construction.

However, six months later a federal district judge upheld the suit of several local groups, on whose behalf I testified, ruling that the Navy must prepare a new environmental impact statement. The Navy lost two appeals of this decision but won a lifting of the injunction in the third appeal, so construction is, at this writing, going on.

Fatal Locations

Subliminal activation of the stress response is one of the most important effects that EMFs and nonionizing radiation have upon life, but it's far from the only one. These unfamiliar energies produce changes in nearly every bodily function so far studied. Many of these alterations are associated with stress, but whether they're a result of, or an additional trigger for, the adrenocortical reaction is an irrelevant chicken-or-egg question at this stage of our knowledge. The most disturbing data come from work on the systems that integrate other bodily functions—the central nervous, cardiovascular, endocrine, and growth control systems. We'll concentrate upon these in the following overview of the biohazards.

For the most part, no attempt will be made to identify specific effects from microwaves, radio waves, and electric or magnetic fields, for similar changes have been observed from all modalities. The major problems come from extremely low frequencies, but higher frequencies have the same effects if pulsed or modulated in the ELF range. This is very often the case, for, to transmit information, microwaves or radio waves must be shaped. This is done by interrupting the beam to form pulses or by modulating the frequency or amplitude (size) of the waves. Furthermore, today's environment is a latticework of crisscrossing signals in which there's always the possibility of synergistic effects or the "construction" of new ELF signals from the patterns of interference between two higher frequencies. Therefore, experiments in which cells or organisms are exposed to a single unmodulated frequency, though sometimes useful, are irrelevant outside the lab. They're most often done by researchers whose only goal is to be able to say, "See, there's no cause for alarm."

The Central Nervous System

Since our work on human reaction time, half a dozen other groups have also found marked CNS effects from ELF fields. Most experiments have shown a decrease in reaction speed, although one researcher noted faster-than-normal reactions in humans exposed to very weak electric fields vibrating at beta wave frequencies. The sensitivity of some animals has turned out to be amazing. James R. Hamer of Ross Adey's group at UCLA reported changes in monkeys' response times from ELF electric fields as weak as 0.0035 volts per centimeter, roughly equivalent to the field from a color TV set 60 feet away.

One of the most telling tests was a simple one done at the Navy's Pensacola lab. R. S. Gibson and W. F. Moroney measured people's short-term memory and their ability to add sets of five 2-digit numbers in the presence of a 1-gauss magnetic field—-the strength found near some high-voltage power lines and many common high-current appliances, such as portable electric heaters. Test scores declined at both the 60-hertz power frequency and the 45-hertz frequency of the Sanguine-Seafarer antenna, but remained normal in control sessions.

Several studies on both sides of the Iron Curtain have found that rats are generally less active and less exploratory of their environment after being dosed with microwaves, although some frequencies induce restlessness. In contrast, ELF magnetic or electric fields almost always produce hyperactivity and disturbed sleep patterns in rats.

Obviously the subtle workings of the mind may undergo many shifts that don't show up in these crude behavioral tests. Most of our knowledge of electropollution's effects on the brain concerns variables that can be more easily quantified, such as changes in biochemistry, cells, and EEG patterns. These studies can't be easily related to changes in thought processes, but most of the results fit in well with the stress response.

In 1966, Yuri Kholodov found effects on rabbits' EEGs from a few minutes' exposure to fairly strong steady-state magnetic fields (200 to 1,000 gauss). As we'd found in salamanders, there were more delta waves, as well as bursts of alpha waves. He and another Russian biophysicist, R. A. Chizhenkova, also noted a desynchronization, or abrupt shift in the main EEG rhythm, for a few seconds when any field was switched on or off. The same effect has since been confirmed in rats with microwaves. This proved that the brain could sense the field, whether the animal knew it or not.

The sites of the greatest changes—the brain's hypothalamus and cortex—were cause for concern. The hypothalamus, a nexus of fibers linking the emotional centers, the pituitary gland, the pleasure center, and the autonomic nervous system, is the single most important part of the brain for homeostasis and is a crucial link in the stress response. Any interference with cortical activity, of course, would disrupt logical and associational thought.

In 1973 Zinaida V. Gordon, a pioneer in microwave research working with M. S. Tolgskaya at the USSR Academy of Medical Sciences Institute of Labor Hygiene and Occupational Diseases, reported a possible cellular feature of EMR stress. Low doses of microwaves, a mere 60 to

320 microwatts for an hour a day, changed nerve cells in the hypothalami of rats. During the first month of exposure, the neurotransmitter-secreting portions of the cells connecting the brain to the pituitary gland were enlarged. After five months they'd begun to atrophy. When microwave dosage was stopped at that point, however, the cells recovered. J.J. Noval's finding that ELF electric fields changed brainstem acetylcholine levels has already been mentioned. In similar experiments, others have noted a rise followed by a drop to below normal in rat brain levels of norepinephrine, the main neurotransmitter of the hypothalamus and autonomic nervous system. In Soviet work, microwave densities of 500 microwatts or more, delivered in a work-exposure pattern of seven hours a day, gradually reduced norepinephrine and dopamine (another neurotransmitter) to brain levels that indicated exhaustion of the adrenal cortex and autonomic system.

Two years after the Gordon-Tolgskaya report, Allen Frey, who has studied bioeffects of microwaves for over two decades at Randomline, Inc., a consulting firm in Huntingdon Valley, Pennsylvania, found an effect on the blood-brain barrier, the cellular gateway by which specialized capillaries strictly limit the molecules admitted to the delicate nerve cells' environment. Even at power densities as low as 30 microwatts, microwaves pulsed at extremely low frequencies loosen this control, in effect opening up leaks in the barrier. Since some barrier changes occur in response to stress and mood shifts, this could be either a cause or a result of the stress response, or an unrelated effect of pulsed microwaves. In any case, since the blood-brain barrier is the central nervous system's last and most crucial defense against toxins, we must consider this increased permeability a grave hazard until proven otherwise.

Researchers have noted several other potentially dangerous direct effects of electromagnetic smog on the neurons. In 1980 a group under R. A. Jaffe at Pacific Northwest Laboratories in Richland, Washington, found a general increase in neural excitability, especially at the synapses, in rats exposed to 60-hertz electric fields of only 10 volts per centimeter for one month. That same year A. P. Sanders and co-workers at the Duke University Medical Center in Durham, North Carolina, reported as follows on biochemical tests of rat brains subjected to microwaves at two levels, one half and also slightly more than the U.S. safety standard of 10,000 microwatts: "The results suggest that microwave exposure inhibits electron transport chain function in brain mitochondria and results in decreased energy levels in the brain."

In a series of experiments spanning mor than a decade, a group of scientists headed by Ross Adey, first at UCLA and later at the Loma Linda VA Hospital, have studied neuron response to ELF fields and pulses. Proceeding from Harrier's work on reaction time, they first ascertained that an even weaker electric field, roughly the influence of a light bulb 10 feet away, changed the firing rate of brain cells in monkeys and humans // the field was pulsing at brain wave frequencies. Then, working with radio waves beamed at chick brains kept alive in culture dishes, they found specific pulse rates that decreased or increased the binding of calcium ions to the nerve cells. The flow of calcium ions in and out of neurons controlled the firing rate of impulses in a complex feedback system. Two "windows" of pulsed radio waves (147 megahertz pulsed at 6 to 10 hertz, and 450 megahertz pulsed at 16 hertz) increased the flow of calcium from the cells, interfering with impulse transmission.

Unfortunately for conceptual simplicity but fortunately for the test animals and the rest of us, the pulsed frequencies that work on isolated brains don't work on whole animals. Adey has publicly expressed his conviction that pulses for changing calcium flow in intact nervous systems do exist, however, and he expects that a calcium efflux would interfere with concentration on complex tasks, disrupt sleep patterns, and change brain function in other ways that can't be predicted yet. This research obviously points toward "confusion beam" weaponry, so effective windows may already have been found, but they haven't been reported in the open literature. Be that as it may, Adey's work remains an important clue to the interaction between EMR and the human CNS at the brain's most sensitive frequencies. Together with the other findings just mentioned, it shows that electropollution can trigger profound and dangerous changes, even if we don't yet know exactly how and when.

Just how dangerous these changes may be was indicated by a study that Maria Reichmanis, Andy Marino, and I did in 1979, collaborating with F. Stephen Perry, a doctor near the town of Wolverhampton in western England. Perry had noticed that people living near overhead high-voltage lines seemed more prone to depression than others in his practice. Since ELF electric fields changed norepinephrine levels in rat brains and since depletion of this neurotransmitter in certain brain areas was a clinical sign of depression, the connection seemed plausible. We knew from earlier work that, although electromagnetic field strength fell off quickly in the immediate vicinity of a power line, the rate of decrease lessened with distame, so that the field was often well above background levels over a mile away. Reasoning that suicide was the one unequivocal and measurable sign of extreme depression, we plotted the addresses of 598 suicides on maps showing the location of power lines in Perry's locality. Then we statistically compared this distribution with a set of addresses chosen at random.

The suicide addresses were, on the average, closer to the high voltage wires. We found the same association with underground power lines, but we couldn't be sure whether more than the statistically expected number of suicides had occurred in areas where the fields were strongest. Since the total field strength was a combination of elements from many sources, we proceeded to measure the actual levels. This confirmed a link. Magnetic fields averaged 22 percent higher at suicide addresses than at the controls, and areas with the strongest fields contained 40 percent more fatal locations than randomly selected houses.

The Endocrine, Metabolic, and Cardiovascular Systems

Living things interpret electromagnetic energy for information about time and place, so they must have a means to filter out useless signals, although perhaps not those never before encountered. Many studies have found that the bioeffects of artificial energy stabilize after a few weeks, suggesting that animals adapt so as to live normally in a changed environment. Hence there's a large body of work that's often quoted to "prove" that electropollution isn't dangerous. As already noted, this simplistic viewpoint doesn't take into account the additive effects of stress. Moreover, when a stress is too strong or too persistent, compensation fails, and the effects become obvious and sometimes irreversible. When evaluating research on hazards, therefore, we must always ask whether the experiment was continued long enough to be informative. Otherwise, a short-term study showing harm is likely to be truer than a reassuring one of medium length.

The primary effect of electromagnetic energy on the endocrine system appears to be the stress responses already described. The major confirmatory study in humans comes from the Soviet Union, where detailed medical tests of seventy-two technicians exposed daily to 1,000 microwatts or less disclosed ominous changes in white and red blood cell counts and an across-the-board decline in immune response. The workers and a group of controls were studied for three years. No human study approaching this in length or completeness has ever been done in the West.

The only other consistently noted glandular change is in the thyroid. The work of several Soviet groups and one American team in the 1970s has clearly shown that radio and microwave frequencies, at power densities well below the American safety guideline of 10,000 microwatts, stimulate the thyroid gland and thus increase the basal metabolic rate. ELF fields at 50 hertz, on the other hand, have depressed thyroid activity in several experiments on rats. It isn't yet known whether this is a direct effect on the thyroid or whether, like the stress response, it's at least partly caused by alterations in brain function.

One more link in the bioclock-interference-and-stress response has come from 1980 work at the Battelle Pacific Northwest Laboratory in Richland, Washington. Working with rats, researchers there found that a weak 60-hertz electric field (only 3-9 volts per centimeter) canceled the normal nightly rise in production of the pineal gland hormone melatonin, the main hormonal mediator of biocycles.

The cardiovascular system responds to electromagnetic energy in at least two ways. The composition of the blood reflects the stress response and concomitant activation of the immune system, while many frequencies exert a direct effect on the electrical system of the heart.

Soviet scientists have observed a variety of blood changes in animals exposed to microwaves, radio waves, and ELF fields. These include declines in red blood cell count and hemoglobin concentration—and hence oxygen capacity—as well as changes in the relative numbers of various types of white blood cells and the relative amounts of blood proteins, and a possible reduction in the blood's ability to clot.

Most of the discomforting studies of electropollution have been done by the Soviets, and they've been given short shrift by Western scientists. There are many reasons for this attitude. There's a simple prejudice against all things Russian and a feeling that their science, technologically less flashy than ours, is necessarily cruder. Western researchers have hamstrung themselves with the dogma that there simply can't be bioeffects from low levels of electromagnetic energy—so why bother looking? Then, too, Russian publication standards are different; procedural details are often omitted, making replication difficult. In addition, there are often troubling contradictions in the data themselves. Results are often inconsistent from animal to animal. If red blood cell count goes up in one, it will go down in another, so the experiment shows no statistical change even though every animal's blood composition is going haywire. In such a situation, the ultramechanistic Americans tend to believe the statistics, while the Soviet biologists concentrate on the animals. Russian scientists have been systematically studying electromagnetic bioeffects since 1933, and we can hardly afford to dismiss their entire body of work simply because it comes from a country we fear.

My associates and I therefore proceeded from one of the most detailed Soviet reports and designed an experiment to measure effects on the blood of mice as our test fields were turned on and off. We concluded that these effects weren't a reaction to the fields themselves but rather a transient compensation that the animals were making in response to any change in their electromagnetic environment. By themselves, none of the blood fluctuations were especially hazardous. However, since we all live amid EMFs that are constantly shifting as we turn appliances on and off or travel from place to place, the continual blood instability could be significant.

American attitudes began to alter in 1978-79 when Richard Lovely of the University of Washington took advantage of a detente-inspired exchange of microwave results to visit the Soviet Union for a month and study Eastern methods closely. His research group then painstakingly recreated a major Soviet experiment in which rats had been irradiated seven hours a day for three months with 500 microwatts. The Russian work was confirmed in every detail, including disruption of the blood's sodium-potassium balance, other pathological changes in blood chemistry, damage to the adrenal glands from stress-induced hormonal changes, diminished sense of touch, a decline in explorativeness, and slower learning of conditioned responses. Donald I. McRee, director of the EPA electromagnetic-radiation health research program, termed the results "very interesting" and called for an end to the American establishment's contempt for Soviet work.

Electromagnetic energy has other adverse effects on blood composition and tissue function. Yuri D. Dumansky, one of many Soviet biophysicists who have done detailed work on microwave hazards, found changes in carbohydrate metabolism, including a rise in human blood-sugar levels, resulting from 100 and 1,000 microwatts. Power-frequency (50-hertz) fields were also linked to altered sugar and protein metabolism in rats, as well as decreased muscular strength in rabbits. Like many other Russian results, these were questioned because of American failure to corroborate them. In this case a research team headed by N. S. Mathewson of the Armed Forces Radiobiology Research Institute in Bethesda, Maryland, reported no such metabolic changes in response to the Sanguine-Seafarer 45-hertz frequency.

However, the Mathewson group made a fundamental mistake. They neglected to account for the 60-hertz background field near the test cages in their lab, even though they'd measured it when setting up the work station. When we reanalyzed their data in light of this omission, the experiment showed exactly the same changes in blood levels of glucose, globulins, lipids, and triglycerides as the Russians had found.

The most frightening data so far on blood composition come from a preliminary study for Project Sanguine. Dietrich Beischer found that one day of exposure to a magnetic field such as would be produced by the

ELF antenna caused a 50-percent rise in triglycerides in nine of ten human subjects. The NAS committee's conclusion that this early result didn't stand up was based on subsequent Navy work, mainly the faulty Mathewson study. Adequate follow-up by a disinterested group has never been funded, even though Beischer's finding agrees completely with Russian studies and the reinterpreted Mathewson data on animals. This doesn't exhaust the list of microwave metabolic effects reported behind the Iron Curtain. Dumansky found widespread changes in the liver function of rats exposed to low levels of microwaves that were scheduled to approximate the pattern of mealtime exposure from microwave ovens. Others detected vitamin B2 and B6 depletion from blood, brain, liver, kidneys, and heart, as well as major shifts in trace-metal metabolism in response to low levels of microwaves. The distribution of copper, manganese, molybdenum, nickel, and iron was affected throughout the bodies of rats. Similar trace-metal changes were recorded after exposure to ELF electric fields for four months, even at moderate field strengths for only half an hour per day. Since B6 is essential to the utilization of carbohydrates, fat, and protein, and since the trace metals act as catalysts in a wide variety of biochemical reactions, these observations may explain some of the other metabolic changes.

There are indications that some types of electropollution directly decrease the efficiency of the heart. Several research groups in Poland, the Soviet Union, Italy, and the United States have studied pulse, electrocardiogram, blood pressure, and reserve capacity (the heart's ability to handle exertion) in animals. Microwaves and 50-hertz electric fields both produced similar changes that persisted throughout long-term exposure. These included bradycardia (decreased pulse), a huge reduction (40 to 50 percent) in the strength of the electrical impulses governing contraction of the heart muscle, a decline in reserve capacity, and a short-term rise followed by a long-term fall in blood pressure. In general, these decrements occurred in both "domestic" (0.5 volts per centimeter) and "industrial" (50 volts per centimeter or more) electric fields and at microwave power densities of 150 microwatts, well within the amount received by many people from radar beams and microwave ovens.

In humans, confirmatory evidence for these effects comes from several Russian studies of workers in high-voltage power station switchyards. In the first such group examined, forty-one of forty-five had some sign of nervous or cardiovascular disease, including bradycardia, instability of pulse and blood pressure, and tremors. The same health problems were found in four additional studies of nearly seven hundred more workers.

The only comparable American study is much quoted for its failure to find consistent health damage in a mere eleven power-line maintenance workers.

Growth Systems and Immune Response

Given the results presented so far and the dynamics of life's connection to the earth's field, we can now make several predictions about the effects of ELF pollution. The most important aspects of the natural electromagnetic field for the biological timing systems are the lunar circadian rhythm and the micropulsations of 0.1 to 35 hertz. It seems logical that cells will perceive frequencies close to normal more readily than those further removed from the norm. Therefore we can postulate that the ELF band from 35 to 100 hertz would be the most damaging, while higher frequencies might go more or less unnoticed until the energy injected into cells became intense or prolonged enough to be significant. The accumulating evidence supports this idea.

Based on this notion, we can predict two major ELF effects that would encompass many others. We can expect the abnormal signals to disrupt biocycles. Such disruption would trigger the generalized stress response even if the EMR-induced changes in brain neurotransmitters were only an effect and not a cause of the stress reaction. In addition, the wrong timing signals would likely throw off the mitotic cycle time of every cell, interfering with growth processes throughout the body.

Although any number of factors can trigger the adrenocortical stress reaction, the response itself is always the same. It involves the release from the adrenal glands of specific hormones, mainly the corticosteroids, which in turn mobilize the body against invading germs or foreign proteins. Thus the stress response always activates the immune system.

Short exposures to stress aren't necessarily harmful and may even be healthy. In fact, the Soviet work on microwave stress has disclosed a brief period of increased immune-system competence at very low intensities (under 10 microwatts). However, when an organism must face a continual or repeated stress, the response system enters the chronic phase, during which resistance declines below normal and eventually becomes exhausted. Several well-known diseases, such as peptic ulcer and hypertension, result directly from this stage, but the most important result is a decrease in the body's ability to fight infection and cancer.

The trouble is that the immune system is geared to fight tangible invaders—bacteria, viruses, toxins, and misbehaving cells of the body itself—or such consciously detectable stresses as heat, cold, or injury. It includes a system of circulating antibodies by which specialized cells recognize the intruder. The cells controlling this phase, which is called humoral immunity, then select appropriate defenders from an array of other types, each programmed for a certain function, such as digesting bacteria, clearing away cellular debris, or neutralizing poisons. Electromagnetic energy isn't consciously perceived, however. It tricks the immune system into fighting a shadow. Thus we can predict that, just like a fire company answering a false alarm, the body will be less able to fight a real fire.

Experiments bear out this supposition. Impaired immune response has been found at many frequencies. Several groups of Soviet researchers have found a decline in the efficiency of white blood cells in rats and guinea pigs after the animals had been exposed to radio waves and microwaves. Most of these experimenters checked for immune system disruption only up to power densities of about 500 microwatts, one twentieth of the nominal American safety standard. Multiple dangers from higher levels are already considered proven in the Soviet Union.

As predicted, however, the most dramatic reported effect on immune response has been produced by ELF fields. During his systematic study of 200-gauss, 50-hertz magnetic fields, Yuri N. Udintsev found that the concentration of bacteria needed to kill mice in such an environment was only one fifth that needed without the field.

When considering resistance to illness, we must also account for the effect of electromagnetic energy on the disease itself, a factor that has so far been all but ignored. Virtually the only evidence to date is a disturbing piece of work by Yu. N. Achkasova and her colleagues at the Crimean Medical Institute in Simferopol. In 1978 they reported the results of exposing thirteen standard strains of bacteria—including anthrax, typhus, pneumonia, and staphylococcus—to electric and magnetic fields. After accounting for magnetic storms, ionospheric flux, passage of the interplanetary magnetic-field boundaries, and other variables, they found clear evidence that an electric field only slightly stronger than earth's background stimulated growth of all bacteria and increased their resistance to antibiotics. The magnetic fields inhibited the growth of the germs but in many cases still enhanced their resistance to antibiotics. Achkasova concentrated on frequencies between 0.1 and 1 hertz, so the survey was far from complete, but perhaps the most important finding was that every field tested had an effect, even after one four-hour exposure. In many cases longer exposure produced permanent changes in bacterial metabolism.

The admittedly sketchy evidence to date suggests that our electropollution is presenting us, and perhaps all animals, with a double challenge: weaker immune systems and stronger diseases. We shouldn't be surprised, then, at an onslaught of "new" ailments, beginning about 1950 and accelerating toward the future. In several cases, new maladies have recently been described as coming from pathogens that previously weren't capable of inducing disease, and this, too, shouldn't surprise us. Among the newcomers are:

Reye's syndrome. First described in 1963, this condition begins with severe vomiting as a child is recovering from the flu or chicken pox. It then progresses to lethargy, personality changes, convulsions, coma, and death. The mortality rate, initially very high, has now been reduced to about 10 percent, but the incidence has increased greatly.

Lyme disease. A virus disease carried by certain insects, it produces severe arthritis in humans. It's one of several similar illnesses that have appeared only recently.

Legionnaire's disease. This is a pneumonia caused by a common soil bacterium that has found a second home in air-conditioning systems. The organism caused us no recognized problems before the initial outbreak in Philadelphia in 1976.

AIDS. Autoimmune deficiency syndrome is a condition in which the body's immune system fails completely and its owner often dies. The patient is unable to resist common, otherwise harmless bacteria and viruses, and can no longer suppress the seeds of cancer that reside in all of us. At present, some sort of virus is suspected as the precipitating cause.

Herpes genitalis. This disease isn't new, but its prevalence and severity have increased tremendously in one decade. Sexual permissiveness generally takes the blame, but a decline in immunocompetence may be more important.

Certainly there are additional factors that may be contributing to the rise of these and other new illnesses. Chemical pollution and the prevalence of junk food are two of the most obvious. However, these diseases, as well as cancer, birth defects, and the other growth problems described below, are on the increase throughout the industrialized world. So are some of the major psychological diseases, such as depression and compulsive use of all types of drugs, from caffeine, nicotine, and alcohol to prescription tranquilizers and the illegal euphoriants, Although heart-attack death rates have declined in the last five years (for no known reason), they're still far higher than before World War II. These diseases exist at more or less the same rates in countries whose chemical toxicity, eating habits, and styles of life are widely divergent. However, the massive use of electromagnetic energy is a common denominator uniting all of the developed nations. In particular, the entire North American continent, Western Europe, and Japan generate such strong 50- and 60-hertz fields that they can be sensed by satellites in space. The populations of these areas are continuously bombarded by these ELF fields.

Disruption of the biocycle timing cues must inevitably make it harder for the body to regulate the mitotic rate of its cells. The major exception to the "no effect" assurances in the NAS Sanguine-Seafarer report was unignorable evidence that 75-hertz fields lengthened the mitotic cycle and hindered cell respiration of the slime mold used in standard tests of cellular growth. The same effects were seen regardless of field strength. Hence we should expect that ELF pollution would foster diseases in which growth processes go awry.

Indeed there has been an alarming increase in such problems. Cancer is hardly a novel illness, but its prevalence is new. In the mid-1960s roughly a quarter of the U.S. population could expect to develop it. By the mid-1970s, that figure had risen to one third, and it's now even higher. The incidence of birth defects has doubled in the past quarter century. There has been a similarly rapid rise in infertility and other reproductive problems.

Rarer defects of cell division may be on the increase as well, expecially among workers exposed by occupation to high levels of electromagnetic energy. Pathologist Hylar Friedman of the Army Medical Center in El Paso reported in 1981 that radar technicians were three to twelve times more likely than the rest of the population to get polycythemia, a rare blood disorder characterized by production of too many red blood cells. Such relationships are hard to confirm statistically, however, in diseases affecting small numbers of people. We need direct experimental evidence and large-scale studies on the widespread disorders. Both are now available.

Back in 1971, two more Soviet researchers, S. G. Mamontov and L. N. Ivanova, reported that industrial-strength 50-hertz electric fields tripled the mitotic rate of liver and cornea cells in mice. Soon afterward, Bassett and Pilla published empirical evidence that pulsed EMFs accelerated the healing of bone fractures. For the most part, however, concrete evidence that time-varying fields could affect cell division was slow in coming.

That situation has changed in the last few years. Several experimenters, notably Stephen Smith, have now proven that the Electrobiology bone-healing device, using 15 pulse-bursts per second, speeds up the division rate of cells that are already proliferating rapidly. Among normal cells, this includes skin, gut, and liver cells. In 1983, A. R. Liboff, a biophysicist at Oakland University in Rochester, Michigan, reported on the effects of a more inclusive set of parameters. Magnetic fields of 0.2 to 4 gauss, vibrating at 10 to 4,000 hertz, all enhanced the replication of DNA during the S (synthesis) phase of mitosis.

As predicted, the interaction appears to be greatest between 35 and 100 hertz. Jose M. R. Delgado—the flamboyant advocate of a "psychocivilized" society through mind control, who has publicized direct electrical stimulation of the brain by such displays as stopping a charging bull in its tracks with a radio impulse transmitted to an implanted electrode—recently reported results of a genetic study of magnetic fields at three frequencies. Delgado placed chick embryos in minuscule magnetic fields pulsed at 10, 100, and 1,000 hertz. He used fields of only 0.001 gauss, or roughly the strength of the earth field's micropulsations. Chicks exposed to the 10-hertz fields were normal, but those dosed at 100 hertz developed severe defects of the central nervous system. The highest frequency also yielded abnormalities, but they were much less severe. Higher intensities are common in homes, in offices, and near power lines. The Navy has found stronger fields near its 76-H3 ELF antenna and reradiated at that frequency from a power line a mile away.

It's important to bear in mind that, in stimulating DNA synthesis, an electromagnetic field doesn't distinguish between desirable and undesirable growth. It affects all cells in the same way, but cell systems that are already rapidly dividing are speeded up the most. As we've seen in earlier chapters, these susceptible processes include healing, embryonic growth, and cancer.* In fact, a researcher working on the New York State Department of Health's power line project, Wendell Winters of the University of Texas Health Sciences Center in San Antonio, recently reported some of the first laboratory evidence that power frequencies can accelerate malignant growth. Winters exposed human cancer cells to 60-hertz electromagnetic fields for just twenty-four hours, and found a sixfold increase in their growth rate seven to ten days later.

---

*Only the magnetic component appears to accelerate healing in any way. Power-frequency electric fields severely retard fracture healing in rats, as Andy Marino, Jim Cullen, Maria Reichmanis, and I proved with a series of experiment in 1979. This work was confirmed the following year by R. D. Phillips in a study done for a Department of Energy review of transmission line bioeffects.

Moreover, the perturbation of normal cell-cycle time is enhanced if nuclear magnetic resonance (NMR) is induced in the atoms of the DNA molecules. In simplified terms, nuclear magnetic resonance is present when the magnetic fields around atomic nuclei are induced to vibrate in unison. The phenomenon requires two external magnetic fields, one steady and one pulsating. For every chemical element, the oscillating field at a specific frequency will induce resonance within the steady-state field at a certain strength.

In 1983 a research team under A. H. Jafary-Asl showed that the earth's magnetic background could serve as the steady field, while the harmonics of power line frequencies could produce a time-varying field that would induce nuclear magnetic resonance in at least two common atoms of living tissue—potassium and chlorine. Other elements might also be susceptible to the effect. Bacteria and yeast cells exposed to these NMR conditions doubled their rate of DNA synthesis and proliferation, but daughter cells were half size. Liboff, analyzing contradictory studies, found that the contradictions disappeared when he calculated resonance conditions for the earth's field where each test was done. Previous work must now be reinterpreted as one vast experiment in adding new frequencies to the varying background.

Almost all experimenters to date have tested the response of organisms to a single specific frequency and intensity. This approach was needed in the beginning to provide a basic level of knowledge, but it's far removed from everyday life, in which we're all exposed to many frequencies simultaneously. A synergism between electromagnetic energy and radioactivity has already been suggested by the fact that cancer rates among nuclear power plant workers are higher than was predicted solely by the higher levels of ionizing radiation in their environment. Nuclear power plants abound in multifrequency radio waves and other electromagnetic radiation. In addition to inducing NMR in the building blocks of living cells, multiple frequencies may likewise interact synergistically to yield biohazards greater than the sum of their individual dangers.

Animal experiments on the risk of cancer and birth defects from electromagnetic energy arc scarce, even in the USSR. The little work that has been done was mostly on microwaves. The only well-known American laboratory study of birth defect dangers used pulsed radio waves and found numerous mutations in fruit fly offspring. In 1976 a Russian group dosed rats with 50 and 500 microwatts for one to ten days. When they then studied somatic (nongenital) cellls from the animals, they found chromosome defects in astounding numbers. At the higher power density there were five times as many as in the controls, and even at the lower intensity the number continued to increase (to 150 percent of the normal value) for two weeks after the beams were turned off.

A 1979 study directed by Przemyslaw Czerski of the National Research Institute of Mother and Child, in Warsaw, documented increased numbers of damaged chromosomes in the sperm of mice exposed one hour a day for two weeks to microwave intensities ranging from 100 microwatts up to the American safety standard of 10,000 microwatts. An even more discomfiting set of data came from a mid-1970s Russian experiment in which female mice were subjected to small power densities, 10 to 50 microwatts. Throughout this range there was a decrease in the number and size of litters and an increase in developmental problems among the newborn animals. The rate of stillbirths jumped from 1.1 percent at the lowest intensity to 7 percent at the highest.

Alas, human beings are the main experimental animals in this line of research. Those who contend microwaves pose no danger often quote a survey of twenty thousand Korean War veterans completed in 1980 by C. D. Robinette and others for the NAS-National Research Council's Medical Follow-up Agency. Comparing VA medical records of radar technicians and others heavily exposed to microwaves with the records of controls, this group found no increase in the death rate. This finding can't be relied on, however. Most of the controls were radar operators, who are exposed to some radiation from radar beams as well as from their consoles. Thus the presumption that they absorbed negligible amounts of EMR just doesn't hold water. In the last few years more reliable epidemiological studies have appeared, showing increased rates of cancer and birth defects among people exposed to higher-than-average levels of electromagnetic energy.

Since microwave broadcasts for television and telephone relays must be in a line of sight to the receivers, there are only a few suitable high locations for the transmitters near each city. Of necessity there's an above-normal concentration of ELF fields and microwave spilloff in that area, possibly leading to a destructive synergism as outlined above. Moreover, since TV is aimed at an audience and phone relay beams at the next station, corridors are set up within which people get more than their share of microwaves.

Sentinel Heights, seven miles from downtown Syracuse, is one such transmitter hill. Slightly more than a thousand people live there. From 1974 to 1977 I learned of seven cases of cancel in that area. They were divided into two clusters, in two microwave corridors separated by a shadow zone. This is 55 percent more than the 4.5 cases statistically expected for this population, and there may have been more cases I didn't know about. Obviously, in such a small and unscientific sample the results could have been due to chance, but the ominous implications demanded some more extensive surveys.

The first one came in 1979, when Nancy Wertheimer and Ed Leeper of the University of Colorado Medical Center in Denver published a study of childhood cancer and power lines. The researchers studied 344 deaths from childhood cancer between 1950 and 1973. The address of each of the victims was paired with the address of the next baby born in the area, to provide a matched series of controls. If the family had moved before the death, both birth and death addresses were used in the experimental group. The wiring of each house and its distance from the nearest transformers were studied. It proved possible to divide the houses into two groups: those with high-current wiring configurations producing strong magnetic fields, and those wired in a low-current arrangement producing much weaker magnetic fields. After certain other variables—such as economic class, family risk patterns, traffic, and urbanization differences—were factored out, the childhood death rate from leukemia, lymph node cancer, and nervous system tumors in the high-current homes was more than double the rate in low-current homes.

Three years later S. Milham, director of occupational health and safety for the state of Washington, found that adults who worked in strong electromagnetic fields also had a leukemia incidence significantly higher than the norm. The link appeared in statistics for generating-station operators, high-voltage-line maintenance workers, aluminum smelters, and several other categories of laborers.

Besides the investigation itself, another thing was noteworthy about Milham's paper: the reaction of the scientific establishment. Another paper quickly appeared in the same periodical, the New England Journal of Medicine, citing many other studies to prove Milham wrong. However, all of them involved controlled exposure to microwaves alone, while the jobs studied by Milham were in the real world, where microwaves and power-frequency fields mix. The editors declined to publish my letter pointing out this obvious flaw in the critique, but still it was momentous that such a prestigious publication ran Milham's paper at all.

Soon confirmatory reports appeared. Wertheimer's and Leeper's findings were duplicated in Stockholm by a group who correlated childhood leukemia with actual measurements of magnetic fields. The strongest statistical link was found with 200-kilovolt power lines running within 200 yards of the stricken child's home. Milham's work was vindicated by surveys in Los Angeles and Great Britain. Wertheimer herself extended her observations to adults and found the same highly significant connection between high-current wiring and various cancers, especially leukemia.

Radar beams (composed of pulsed microwaves) have the highest power densities of any EMR source. In the laboratory, both radio frequency and microwave radiation have been shown to change the gateway-barrier function of cell membranes, upset hormone balances, and induce chromosome defects, all of which are factors in malignant growth. However, there have been few attempts to directly assess radar's potential role in human cancer.

John R. Lester and Dennis F. Moore of the University of Kansas School of Medicine in Wichita have recently done so. Wichita was an ideal location for such an inquiry. It had two airports with radar towers, but few other major sources of electropollution. Its chemical environment was also quite clean as cities go. Lester and Moore plotted the cancer incidence for the whole city and found it was highest where the residents were exposed to both radar beams. It was lower where only one beam penetrated, but lowest where the population was fully shielded behind hills. The results held up when other factors, such as age, poverty, sex, and race, were statistically balanced as far as possible. The authors noted one apartment house whose cancer death rate was twice that of the area's nursing homes; its upper floors were in direct line with both radar beams.

Heart attack rates in North Karelia and Kuopio, Finland, became the highest (and most swiftly increasing) in the world within a few years after the Soviets installed a gigantic over-the-horizon radar complex that bounced microwaves off the surface of Lake Ladoga and through these parts of southeastern Finland. These are rural districts whose way of life is built on outdoor labor rather than the sedentary indoor stresses generally associated with heart disease. Noting that cancer rates had also risen precipitously in the region, Lester and Moore went on to investigate statistics for American counties having Air Force bases. These counties had a significantly higher percentage of cancer deaths than other counties, even though radar towers from commercial airports inevitably must have smoothed out the data and made the difference less striking.

The study of human genetic defects from electromagnetic energy is still in a primitive stage. In the case of microwaves, this situation is largely due to obstruction by military and government agencies. Even in World War II, rumors of radar-induced sterility were so rampant that sailors often gave themselves "treatments" before shore leave. The first scientific evidence of reproductive effects didn't come until 1959, when John H. Heller and his co-workers at the New England Institute for Medical Research in Ridgefield, Connecticut, found major chromosome abnormalities in garlic shoots irradiated with low levels of microwaves. They soon found the same changes in mammalian cells, as well as the fruit fly mutations mentioned above. Their work in this direction ended about 1970 due to lack of funds.

In 1964 a group of researchers studying Down's syndrome at the Johns Hopkins School of Medicine, after linking the malady to excess X rays given to pregnant women, found an unexpected further correlation with fathers working near radar. It was a full decade before any money was allocated to follow up this finding, and, while the link between parental radar exposure and Down's syndrome wasn't substantiated, higher-than-normal numbers of chromosome defects were found in the blood cells of radarmen.

By this time an Alabama professor of public health had found an apparent surge in birth defects among children of radar-exposed Army helicopter pilots. In 1971 Dr. Peter Peacock noted that there had been seventeen children born with clubfoot within a sixteen-month period at the Fort Rucker, Alabama, base hospital. Statistically, there should have been no more than four.

Working through two federal agencies and two private research foundations, Peacock and others tried for five years to follow up this disturbing news, only to be thwarted by some clever tactical moves by the Army. Refusing to release work records, medical files, and radar inspection records on grounds of "privacy" and "national security," officials of the Army Medical Research and Development Command managed to prevent all but two reassessments of Peacock's original data for several years. They stalled separate research proposals sponsored by the Environmental Protection Agency and the Food and Drug Administration's Bureau of Radiological Health without ever letting on to one agency that they were dealing with the other. As the coup de grace, the Army agreed to supply the FDA group with a survey of radar transmitters in the Fort Rucker area. The officers fobbed off on the unwitting civilians a deceitfully sketchy map showing only one major radar installation at the base, whereas an official Army report made at the time of the observed birth defects showed nineteen such emitters. Throughout the Vietnam War thousands of helicopter trainees had each spent months flying through the resultant microwave haze. Much of their training consisted of homing right down the beams to within a few dozen yards of the source in TH-13 Bell copters whose Plexiglas bubbles left them naked to microwaves.

The Fort Rucker affair and many other instances of military-governmental sabotage of health effects research on microwaves have been impeccably documented in New Yorker reporter Paul Brodeur's 1977 book, The Zapping of America. In the early 1970s, for example, follow-up to a preliminary finding of excess Down's syndrome among children of Seattle airline pilots was first supported by the local chapter of the Air Line Pilots Association, then opposed due to pressure from the national level.

The stonewalling continues. Grants for serious consideration of electropollution's dangers have been cut to a trickle in the United States, but some findings continue to emerge, especially from other countries.

A 1976 survey of Hydro-Quebec's generating-station electricians showed a drastic change in the gender ratio of children born after one of the parents began work in the high-EMF environment. Before, boys and girls had been born in equal numbers; afterward, there were six times as many males as females. A 1979 study of Swedish high-voltage substation workers showed lower birth rates and an 8-percent incidence of genetic defects in offspring, as compared with 3 percent among children of a control group. The finding was confirmed in 1983. Since most of the exposed electrical workers were men, the damage apparently was done during sperm formation. Most recently, in May 1984, Nancy Wertheimer presented evidence of a statistical correlation between use of electric blankets, which emit powerful EMFs, and the occurrence of birth defects.

Among the most serious recent data are those concerning video display terminals (VDTs). There have been alarming numbers of miscarriages, stillbirths, and birth defects among pregnant women working in newly computerized offices. In one year at the Dallas office of Sears, Roebuck and Company, for example, only four of twelve pregnancies ended normally. Among twelve pregnant VDT workers at the Defense Logistics Agency in Marietta, Georgia, there were seven miscarriages and three cases of congenital defects. Four VDT operators in the Toronto Star's classified-ad department gave birth to deformed children, while three co-workers who didn't work with VDTs had normal babies. These anomalies must be compared with the normal 15-percent incidence of spontaneous abortion and the 3-percent rate of serious birth defects among the population at large. Writing in Microwave News, an independent newsletter covering nonionizing radiation, in 1982 editors Louis Slesin and Martha Zybko reported on eight such clusters, and workers' groups have documented several others, but still there has been no attempt at a large-scale statistical study to check the oft repeated claim that these are just coincidences.

Two studies are widely quoted as disproving harmful effects from the machines. In 1977, when two New York Times copy editors developed radiation-induced cataracts after less than a year at their new screens, the National Institute of Occupational Safety and Health (NIOSH) tested a few machines and, finding that X-ray emissions were within the half-millirem-per-hour standard for work exposure, concluded there was no link to the health problems. Unfortunately, the agency didn't adequately measure nonionizing radiation, gave contradictory data as to the sensitivity of its own instruments, and failed to test malfunctioning monitors, which are known to emit larger amounts of X rays. Nor is there any assurance that the X-ray exposure standard is adequate, since it was formulated for a much smaller group of workers (mainly nuclear technicians and uranium miners), whose health is continuously monitored in a way that that of VDT operators is not. Furthermore, the NIOSH investigators noted an enormous microwave reading of 1,000 microwatts in one of the Times offices, without even bothering to find out where it was coming from!

Press releases claimed a mid-1983 National Academy of Sciences review would allay the fears once and for all, proving VDTs to be risk free. However, a reading of the text showed a different picture. While the authors played down reports linking birth defects and eye problems to VDT radiation, they admittedly failed to find any research adequate to answer the health questions one way or the other.

According to the sketchy data available, all VDTs (which of course include video games and televisions as well as computer monitors) emit varying amounts of radiation over a broad spectrum. The transformers release VLF and ELF waves, while microwaves, X rays, and ultraviolet emanate from the screen. Poorly adjusted or malfunctioning terminals can emit enormous amounts; two machines tested in the offices of Long Island's Newsday, for example, were producing 15,000 microwatts of radio energy. There's no information whatever on the synergisms that may operate amid this varied radiation over long periods of time, but I suspect that the birth defects are primarily due to the ELF component.

Meanwhile, the only American "research" on the problem continues to be the daily lives of our 10 million or more console operators. Despite the reassurances, at least a third to a half of the workers continue to suffer hea daches, nausea, neck and back pain, and vision impairment. In fact, a 1983 survey of eleven hundren UPI employees conducted by

Arthur Frank, then at New York's Mount Sinai School of Medicine, suggested that VDT users lose so much time due to eye problems and neck pain that the effects may become a major drain on the economy by the end of the decade.

Some of the complaints undoubtedly arise from postural strains and lighting defects in the notoriously ill-designed work areas where many VDTs are used. They could be prevented by more frequent breaks and some sympathetic attention to human engineering. The birth defects and cataracts probably won't disappear so easily, however. Certainly pregnant women should be allowed temporary reassignment without loss of pay, a right already accepted in much of Western Europe and recently put into law in Ontario. That won't protect sperm cells and unfertilized eggs, however. Regular maintenance and a lead-impregnated glass or acrylic screen (such as is used in nuclear power plant windows) can virtually eliminate ionizing radiation, but screen-generated microwaves require a transparent shield that still conducts electrical energy—a product that doesn't yet exist. Some frequencies of EMR are easy to block simply by using metal cabinets instead of the cheaper plastic ones, but VLF and ELF waves require grounded shielding. All these preventive measures are expenses that most manufacturers and managers have been loath to accept; until they do, workers will be paying the entire price.

The dangers of electropollution are real and well documented. It changes, often pathologically, every biological system. What we don't know is exactly how serious these changes are, for how many people. The longer we, as a society, put off a search for that knowledge, the greater the damage is likely to be and the harder it will be to correct. Meanwhile, one of the few honest statements to emerge from the Nixon administration, a warning issued by the President's Office of Telecommunications Policy in 1971, continues to bleed through the whitewash: "The population at risk is not really known; it may be special groups; it may well be the entire population. . . . The consequences of undervaluing or misjudging the biological effects of long-term, low-level exposure could become a critical problem for the public health, especially if genetic effects are involved."

Conflicting Standards

The establishment Attitude toward EMR's health effects derives largely from the work of Herman Schwan. An engineer who had been a professor at the Kaiser Wilhelm Institute of Biophysics in Germany during most of the Nazi era, Schwan was admitted to the United States in 1947, soon accepted a post at the University of Pennsylvania, and since then has done most of his research for the Department of Defense.

Like infrared radiation, radio waves and microwaves produce heat when they're absorbed in sufficient quantity. Although not a biologist, Schwan assumed this heating was the only effect EMR would have on living tissue. In this respect he considered living things no different from the hot dogs that World War II radarmen used to roast in their microwave beams, so cooking was the only harm he foresaw. Schwan then estimated danger levels based on how much energy was needed to measurably heat metal balls and beakers of salt water, which he used to represent the size and presumed electrical characteristics of various animals.

Appreciable heating occurred in these models only at levels of 100,000 microwatts or above, so, incorporating a safety factor of ten, Schwan in 1953 proposed an exposure limit of 10,000 microwatts for humans. By showing soon afterward that it took more than this intensity to cause burns in real animals, Sol Michaelson seemed to have confirmed the safety of "nonthermal" dosages. No one tested for subtler effects, and the 10,000-microwatt level was uncritically accepted on an informal basis by industry and the military. In 1965 the Army and Air Force formally adopted the Schwan limit, and a year later the industry-sponsored American National Standards Institute recommended it as a guideline for worker safety.

There were persuasive economic reasons why the 10,000-microwatt standard was and still is defended at all costs. Lowering it would have curtailed the expansion of military EMR use and cut into the profits of the corporations that supplied the hardware. A reduced standard now would constitute an admission that the old one was unsafe, leading to liability for damage claims from ex-GIs and industrial workers. One of the strongest monetary reasons was given in a 1975 classified summary of the DOD's Tri-Service Electromagnetic Radiation Bioeffects Research Plan: "These [lower] standards will significantly restrict the military use of EMR in a peacetime environment and require the procurement of substantial real estate around ground-based EMR emitters to provide buffer zones." The needed real estate was estimated to be 498,000 acres. The price of this much land would surely run well into the billions of dollars.

Even before it was adopted, there were indications that the standard might be inadequate. During the obligatory fight for compensation in the face of callous official denials of responsibility, an interesting discovery was made by Thomas Montgomery, a former civilian technician working for the Army Signal Corps, who is now blind, deaf, and crippled because of a massive accidental exposure to a radar beam in 1949. In one of the files opened by his suit, Montgomery found a document proving that in the late 1940s the Institute of Radio Engineers had formulated more conservative safeguards that included methods for preventing accidents like the one that had incapacitated him. (He'd been repairing a transmitter when a co-worker, not knowing he was standing in front of the wave guide, turned it on. Since the microwaves were imperceptible, Montgomery didn't know he was being irradiated until it was too late.) Leaders of the military-industrial electronics community chose not to promulgate these proposed regulations.

There were other hints that all was not well. In 1952 Dr. Frederic G. Hirsch of the Sandia Corporation, a maker of missile guidance systems, reported the first known case of cataracts in a microwave technician. The following year Bell Laboratories, alarmed by reports of sterility and baldness among its own workers as well as military radar personnel, suggested a safety level of 100 microwatts, a hundred times less than Schwan's. Even Schwan has consistently maintained that his dosage limit probably isn't safe for more than an hour.

In 1954 a study of 226 microwave-exposed employees at Lockheed's Burbank factory was reported by company doctor Charles Barron. He said there were no adverse effects, despite "paradoxical and difficult to interpret" changes in white blood cell counts, which he later ascribed to laboratory error, as well as a high incidence of eye pathology, which he determined was "unrelated" to radar.

However, the safety standard had already become a Procrustean bed against which all research proposals and findings were measured. Grants weren't given to look for low-level hazards, and scientists who did find such effects were cut down to size. Funds for their work were quickly shut off and vicious personal attacks undermined their reputations. Later, when undeniable biological changes began to be noted from power densities between 1,000 and 10,000 microwatts,, the idea of "differential heating"—hot spots in especially absorptive or poorly cooled tissues—was advanced, as though this convenient explanation obviated all danger. Soviet research could easily be discounted because of its "crudity," but when nonthermal dangers were documented in America, military and industrial spokespeople simply refused to acknowledge them, lying to Congress and the public. Many scientists, who naturally wanted to continue working, went along with the charade.

from nuclear weapons tests. Throughout the 1950s there was "no cause for alarm," but twenty years later the Wyoming sheep ranchers' suit for compensation for fallout-damaged herds unearthed documents proving the responsible officials had known better at the time. Even the symbol of American military machismo may have fallen victim to the policy. John Wayne, as well as Susan Hayward and other cast members, died of cancer about two decades after making a movie called The Conqueror, which was filmed in the Nevada desert while an unexpected wind shift sifted radioactive dust down on them from a nearby test.

Today the EMR deceit still proceeds. On August 2, 1983, Sol Michaelson was quoted as saying some bioeffects had been observed in animals and a few claimed in humans from intensities under 10,000 microwatts, but "none of these effects, even if substantiated, could be considered hazardous or relevant to man"—even though three years before he'd co-authored a paper that reviewed previous evidence and added some of his own in support of the generalized stress response from microwaves.

The evidence had begun to come in over twenty years previously. John Heller's 1959 finding of chromosome changes in irradiated garlic sprouts and the 1964 Johns Hopkins correlation of Down's syndrome with parental exposure to radar were mentioned in the previous section. In 1961 a study was conducted on a strain of mice bred to be especially susceptible to leukemia and used to evaluate risk factors for that disease. Two hundred mice, all males, were dosed with 100,000 microwatts at radar-pulse frequencies for one year. An unusually high proportion of the animals—35 percent—developed leukemia during that time, and 40 percent suffered degeneration of the testicles. Admittedly, this was a very high power density, but the mice were exposed for only four minutes a day. The most disturbing part, however, is that the sponsor, the Air Force, cut off all funds for follow-up work, and to this day no American research has adequately addressed this potential danger.

In 1959 Milton Zaret, an ophthalmologist from Scarsdale, New York, began a study for the Air Force to see if there was any special risk to the eyes of radar maintenance men. He at first found none, because he'd geared the examinations to check the lens of the eye. A few years later, when several private companies referred to him microwave workers who'd developed cataracts, Zaret saw he'd made a mistake. Because the microwaves penetrated deeply into tissue, cataracts from them had developed behind the lens, in the posterior capsule, or rear part of the elastic membrane surrounding the lens. At this point the Air Force again suddenly lost interest, but Zaret has doggedly pursued the matter in his own practice. Although military and industry people still deny that nonthermal microwave cataracts exist, Zaret's work has proven beyond good-faith dispute that low doses exert a cumulative effect that eventually stiffens and clouds the posterior capsule. Surveys by others here and abroad have confirmed his conclusion. In fact, these behind-the-lens cataracts constitute a "marker disease" for sustained microwave exposure. Zaret has personally diagnosed over fifty clear-cut cases, many in airline pilots and air traffic controllers.

In 1971 Zaret was involved in a projected five-year primate study for the Navy when he made the mistake of telling his overseers the disturbing news that one of the monkeys had died after a few hours' exposure to just twice the U.S. safety level. Within days Captain Paul Tyler "happened to be in the area" of the Hawaii research facility and wanted Zaret to show him around. He left in about twenty minutes, having seen little of the equipment, and in a few days more the entire project had been canceled.

The double-dealing has only made Zaret dig in his heels. Through the years he has been one of the few doctors willing to take on the government by testifying on behalf of plaintiffs filing claims for microwave health damage. At such proceedings one always runs into the same cast of characters speaking more or less the same lines. At one trial the ubiquitous Michaelson attacked Zaret's professional abilities, only to have later testimony reveal the embarrassing fact that Michaelson's mother owed the vision in one eye to the ophthalmologist's surgical skill.

As we've already seen, there are many indications that EMFs and EMR weaker than the Schwan guideline have serious effects on growth. The 10,000-microwatt level was directly tested in 1978 by a group at the Stanford Research Institute. Pregnant squirrel monkeys and their offspring were irradiated. Of nine babies zapped in utero and/or after birth, five died within six months, compared with none in the control group.

The U.S. safety standard would be grossly inadequate even if it was law. In actuality, although some businesses and military agencies have adhered to it, it has never posed any threat to those that have not. A federal court case decided in 1975 and upheld in 1977 defined it as an advisory or "should" guideline, which couldn't be enforced. Now, as per instructions dated March 17, 1982, Occupational Safety and Health Administration (OSHA) inspectors can no longer issue even meaningless citations to companies for exposing workers to more than the limit.

The only actual regulation of public electromagnetic energy dosage pertains to microwave ovens. In 1970 the FDA's Bureau of Radiological Health stipulated that no oven should leak more than 1,000 microwatts at a distance of 2 inches when new, nor more than 5,000 after sale. Even at that time, research suggested this was an unsafe level, a fact recognized by Consumers Union in 1973 when it recommended against the purchase of any brand. Leakage surveys have shown that all types put out an average of 120 microwatts near the door, while many emit much higher amounts. A worn seal or piece of paper towel stuck in the door can increase the user's exposure to well over the 5,000-microwatt level, according to Consumers Union tests.

What amounts of electromagnetic energy do workers and the general public actually absorb? The levels vary greatly. Some antenna repairers receive up to 100,000 microwatts for minutes or hours during a job. Many factory workers are in the same bracket. From 1974 to 1978 NIOSH surveyed eighty-two industrial plastic molders and sealers. Over 60 percent exposed the operator to more than the Schwan limit, some to over 260,000 microwatts. Because of low wages for such work, nearly all sealer operators are women of childbearing age. NIOSH has estimated that some 21 million workers are exposed to some level of radiofrequency waves or microwaves as a direct result of their jobs. No metal shielding is provided for most workers in this country, although it's sold to other nations having better safety rules.

At this time there's no way to estimate how much EMR people are getting away from their jobs, because the few readings that have been taken have measured only single sources and single frequencies. No one has yet surveyed our cities and countryside throughout the whole spectrum from ELF to microwaves. All we know is that most people's daily exposure is high. Even the Environmental Protection Agency has estimated that if the Soviet off-the-job safety limit of 1 microwatt in the radio and microwave bands was adopted here, over 90 percent of our FM stations would have to be shut down.

Table 1. Power Density at Various Distances from a 50,000 Watt AM Radio Station

Distance (feet)	Power Density (microW/cm2)	Distance (feet)	Power Density (microW/cm2)
15	838	482	23
29	284	663	12
69	196	1571	2
152	43	3280	1
308	33	5760	0.3

Note Data from R. Tell et al., "Electric and Magnetic Field Intensities and Associated Body Currents in Man in Close Proximity to a 50 kW AM Standard Broadcast Station," presented at Bioelectromagnetics Symposium, Seattle, 1979.

Table 2. EMF in Typical Tall Buildings

City	Location	Power Density (microW/cm2)
New York	102nd Floor, Empire State Building	32.5
Miami	38th Floor, One Biscayne Tower	98.6
Chicago	50th Floor, Sears Building	65.9
Houston	47th Floor, 1100 Milam Building	67.4
San Diego	Roof, Home Tower	180.3

Note: Data from R. Tell and N. H. Hankin, Measurements of Radio Frequency Field Intensity in Buildings with Close Proximity to Broadcast Systems, ORP/EAD 78-3, U.S. Environmental Protection Agency, Las Vegas, 1978.

Table 3. Power-Frequency Electric Fields of Household Appliances Measured at a Distance of One Foot

Appliance	Electric Field (V/m)
Electric blanket	250
Broiler	130
Phonograph	90
Refrigerator	60
Food mixer	50
Hairdryer	40
Color TV	30
Vacuum cleaner	16
Electric range	4
Light bulb	2

Note: Data in tables 3 and 4 from Fact Sheet for the Sanguine System: Final Environmental Impact Statement, U.S. Navy Electronic Systems Command, 1972.

Table 4. Power-Frequency Magnetic Fields of Household Appliances

Range	Appliance
10-25 gauss	Soldering gun Hairdryer
5-10 gauss	Can opener Electric shaver Kitchen range
1—5 gauss	Food mixer TV
0.1-1.0 gauss	Clothes dryer Vacuum cleaner Heating pad
0.01-0.1 gauss	Lamp Electric iron Dishwasher
0.001 -0.1 gauss	Refrigerator

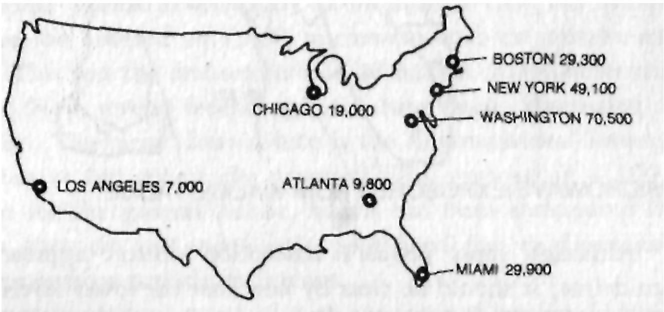
BOSTON 29,300
NEW YORK 49,100
WASHINGTON 70,500
CHICAGO 19,000
LOS ANGELES 7,000
ATLANTA 9,800
MIAMI 29,900

SOURCE: U.S. ENVIRONMENTAL PROTECTION AGENCY, 1978

NUMBER OF PEOPLE IN CERTAIN CITIES EXPOSED TO RADIO AND TV SIGNALS ABOVE USSR SAFETY LEVEL

Most city dwellers continuously get more than a tenth of a microwatt from television microwaves alone. This may be especially significant, because of the human body's resonant frequency. This is the wavelength to which the body responds "as antenna." Next to the ELF range, it's perhaps the region of the spectrum in which the strongest bioeffects may be expected. The peak human resonant frequency lies right in the middle of the VHF television band.

Many people live in zones of higher-than-average risk. Levels rise steeply from 1 microwatt within half a mile of most radio stations. Near "antenna farms" like the forest of transmitters on Mount Wilson on the outskirts of Los Angeles, the densities can reach well into the thousands. They regularly go up to about 7 microwatts near microwave relay towers, which are often placed in the center of towns. Exposure to 100 microwatts is not uncommon within half a mile of military or airport radar towers. Office workers in tall buildings are often in direct line with microwave beams, whose intensities may reach 30 to 180 microwatts, as measured in a recent EPA survey. CB radios and walkie-talkies bombard users, especially their heads and chests, with thousands of microwatts. These figures, of course, represent only single sources, not the total exposure.

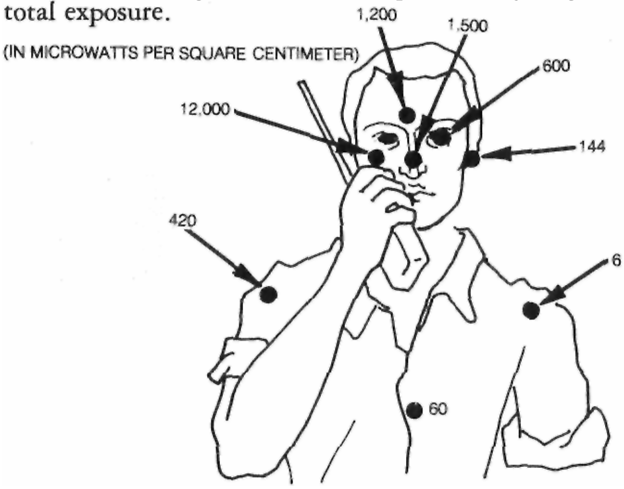

MICROWAVE EXPOSURE FROM WALKIE-TALKIE

Although most people's absorption doesn't approach the Schwan guideline, it should be clear by now that the lower levels are little cause for comfort. Everywhere in Western nations, except in the most remote forests or deserts, the ambient energy from ELF power systems is several thousand times above the earth's background field strength, providing abundant interference with the biocycle timing cues. Moreover, the accumulated research has clearly shown that small doses often have the same effects as larger ones. Ross Adey, who has intensively studied the "window effect,", in which a certain result is produced at some frequencies and power levels but not at others interspersed between the effective-ones, believes future research will reveal such windows at much lower levels, even at fractions of a microwatt. Indeed there has already been one report of brain wave changes suggesting resonance of neural electrical currents with radio waves and microwaves down to a billionth of a **microwatt.**

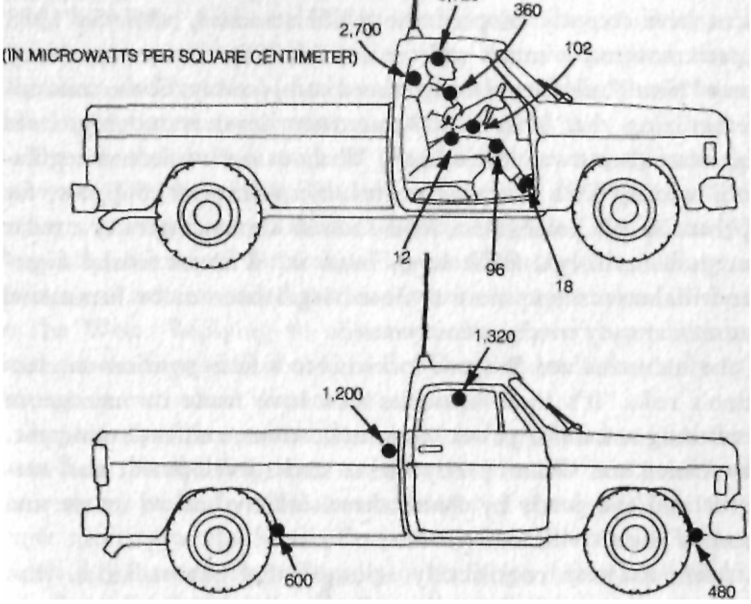

MICROWAVE EXPOSURE FROM A TYPICAL CB RADIO

There's a chance of somewhat stricter rules in a few years. In 1982 the American National Standards Institute recommended that the radiowave safety level be lowered to 1,000 microwatts and the microwave level to 5,000. This was the first semiofficial admission that nonthermal effects do exist. Now several federal agencies have begun discussing a formal regulation. The most likely source is the Environmental Protection Agency, but at last report the rumored EPA proposal of a 100-microwatt limit for the general public, which had been anticipated in late 1984, was abruptly and indefinitely postponed due to dissension within and pressure from outside the agency.

A health-protecting federal standard with the force of law would have a major impact on both industry and government. Industry would experience a decline in revenue and an increase in costs. Government, especially the military, would be inconvenienced in a multitude of activities. Both would be subject to lawsuits for exposures and damages prior to establishment of the standard. In addition, we must understand that no amount of artificial EMR, no matter how small, has been proven safe for continuous exposure. Bioeffects have been found at the lowest measurable doses. However, we must also understand that the greatest danger lies in uncontrolled exposure to large amounts of EMR at many overlapping frequencies, and therefore a stringent standard with a definite timetable for phasing it in is the only way to protect the public health.

Moreover, such action must come from Washington. New Jersey and Connecticut have recently adopted the ANSI standard, while in 1983 Massachusetts enacted a much stricter one of 200 microwatts, which large areas of New York City already exceed considerably. Some communities, recognizing that even a 100-microwatt level is too high, are beginning to set their own, lower, ones. Without realistic federal regulation we will end up with a totally unworkable patchwork. Suppose, for example, that the Air Force, from a base outside a town, operates a radar dome that produces illegal EMR levels inside it. Without federal direction, that will become one more confused legal issue to be hammered out for years in already overburdened courts.

All of the industrialized West is locked into a false position on electropollution's risks. It's these countries that have made the maximum use of electromagnetism for power, communications, and entertainment. The Soviet Union and China, partly due to underdevelopment and wartime destruction, and partly by choice, have severely limited its use and the exposure of their civilians.

Soviet scientists have consistently assumed that any radiation that doesn't occur in nature will have some effect on life. We've consistently made the opposite assumption. Throughout our recent history American regulators have followed a "dead body policy." They have extended no protection until there was proof of harm sufficient to overcome all deception. There's no longer any question that, as far as electromagnetic energy is concerned, we've been wrong and the Soviets have been right.

In the 1950s, Russian doctors conducted extensive clinical exams of thousands of workers who had been exposed to microwaves during the development of radar. Having disclosed serious health problems, these studies weren't swept under the rug. Instead, the USSR set limits of 10 microwatts for workers and military personnel, and 1 microwatt for others. Both levels are strictly enforced. When this first became known in the West in the early 1960s, instead of checking their assumptions many American scientists and administrators chose to believe this was Russian propaganda aimed at embarrassing us.

By 1971, when they presented their work at a momentous conference in Warsaw, Zinaida V. Gordon and Maria N. Sadchikova of the USSR Institute of Labor Hygiene and Occupational Diseases had identified a comprehensive series of symptoms, which they called microwave sickness. Its first signs are low blood pressure and slow pulse. The later and most common manifestations are chronic excitation of the sympathetic nervous system (stress syndrome) and high blood pressure. This phase also often includes headache, dizziness, eye pain, sleeplessness, irritability, anxiety, stomach pain, nervous tension, inability to concentrate, hair loss, plus an increased incidence of appendicitis, cataracts, reproductive problems, and cancer. The chronic symptoms are eventually succeeded by crises of adrenal exhaustion and ischemic heart disease (blockage of coronary arteries and heart attack).

The Soviet standards were set long before the dangers were this clear, however. The comparison is instructive. At a 1969 international symposium on microwaves in Richmond, Virginia, Dr. Karel Marha of Prague's Institute of Industrial Hygiene defended his findings on birth defects and recommended that the Eastern European standard be adopted in the West. Replying to objections that the dire predictions hadn't been proven beyond doubt, he said: "Our standard is not only to prevent damage but to avoid discomfort in people."

Apparently this concern doesn't include Americans, for the Soviets have been bombarding our embassy in Moscow with microwaves for some thirty years. In 1952, at the height of the Cold War, there was a secret meeting at the Sandia Corporation in New Mexico between U.S. and U.S.S.R. scientists, allegedly to exchange information on biological hazards and safety levels. It seems the exchange wasn't completely reciprocal, or perhaps the Americans didn't take seriously what the Russians told them; there have been other joint "workshops" since then, and each time the Soviets have sent people who publicly acknowledged the risks, while the American delegates have always been "no-effect" men. At any rate, soon after the Sandia meeting, the Soviets began beaming microwaves at the U.S. embassy from across Tchaikovsky Street, always staying well within the Schwan limit. In effect, they've been using embassy employees as test subjects for low-level EMR experiments.

The strange thing is that Washington has gone along with it. The "Moscow signal" was apparently first discovered about 1962, when the CIA is known to have sought consultation about it. The agency asked Milton Zaret for information about microwave dangers in that year, and then hired him in 1965 for advice and research in a secret evaluation of the signal, called Project Pandora. Nothing was publicly revealed until 1972, when Jack Anderson broke the story, and the U.S. government told its citizens nothing until 1976, in response to further news stories in the Boston Globe. According to various sources, the Russians shut off their transmitter in in 1978 or 1979, but then resumed the irradiation for several months in 1983.

According to information given Zaret in the 1960s, the Moscow signal was a composite of several frequencies, apparently aiming for a synergistic effect from various wavelengths, and it was beamed directly at the ambassador's office. Thus it may have been used at least partially to activate bugging devices, but it wasn't consistent with one of the other subsequent official American explanations—a jamming signal to disrupt the U.S. eavesdropping equipment on the embassy roof.

The intensity isn't known for certain. When the State Department admitted the signal's existence, officials claimed it never amounted to more than 18 microwatts. However, although released Project Pandora records don't directly reveal a higher level and the relevant documents have allegedly been destroyed, research protocols aimed at simulating the Moscow signal called for levels up to 4,000 microwatts.

In the mid-1960s published Soviet research indicated that such a beam would produce eyestrain and blurred vision, headaches, and loss of concentration. Within a few years other research had uncovered the entire microwave syndrome, including the cancer potential.

By all accounts except the official ones, the Moscow bombardment has been highly effective. In 1976 the Globe reported that Ambassador Walter Stoessel had developed a rare blood disease similar to leukemia and was suffering headaches and bleeding from the eyes. Two of his irradiated predecessors, Charles Bohlen and Llewellyn Thompson, died of cancer. Monkeys exposed to the signal as part of Project Pandora soon showed multiple abnormalities of blood composition and chromosome counts.

In January 1977, the State Department, under duress, announced results of a series of blood tests on returning embassy personnel: a "slightly higher than average" white blood cell count in about a third of the Moscow staff. If 40 percent above the white blood cell counts of other foreign service employees (levels common to incipient leukemia) can be considered "slightly higher than average," then this technically wasn't a lie. The finding has been officially ascribed to some unknown microbe. Unfortunately, there's no such doubt about the veracity of explanations about some earlier research. As part of Project Pandora in the late 1960s, the State Department tested its Moscow employees for genetic damage upon their return stateside, telling them the inner cheek scrapings were to screen for those unusual bacteria. No results were ever released, and they're reportedly part of the missing files, but one of the physicians who conducted the tests was quoted by the Associated Press as saying they'd found "lots of chromosome breaks." The embassy staff had to learn this when the rest of us did - in the newspapers nearly a decade later.

The Russians themselves have never admitted the irradiation, and the Schwan guideline has put the American government in an embarrassing bind. In 1976 the State Department gave its Moscow employees a 20-percent hardship allowance for serving in an "unhealthful post" and installed aluminum window screens to protect the staff from radiation a hundred times weaker than that near many radar bases. That same year the government gave Johns Hopkins School of Medicine a quarter of a million dollars to see if there was a link between the signal and "an apparently high rate of cancer" in the embassy (which wasn't confirmed). Nevertheless, although President Johnson asked Premier Kosygin at the 1967 Glassboro talks to stop the bombardment, Washington has never had any formal basis to demand that it be stopped due to danger to the staff. That was apparently considered an acceptable risk in the protection of the lenient U.S. standard.

Invisible Warfare

The Soviets have led the way in learning about the risks of electropollution, and, as we have seen, they've apparently been the first to harness those dangers for malicious intent. However, the spectrum of potential weapons extends far beyond the limits of the Moscow signal, and Americans have been actively exploring some of them for many years. Most or all of the following EMR effects can be scaled up or down for use against individuals or whole crowds and armies:

The crudest of these armaments would be a sort of electromagnetic flamethrower with a greater range than chemical types. Dogs were cooked to death in experiments at the Naval Medical Research Institute as long ago as 1955, and high-power transmitters using short UHF wavelengths can severely burn exposed skin in seconds. Electromagnetic pulse (EMP) is a term designating the immensely powerful, near-instantaneous surge of electromagnetic energy produced by a nuclear explosion. It was first discovered in the late 1960s. The EMP from one detonation a few thousand miles above the earth would destroy all electrical systems throughout an entire continent. In the early 1970s new types of EMR generators emitting power levels on or twenty times higher than ever before were developed in in an effort to simulate EMP and help devise communications systems shielded from it. In 1973 these transmitters were described in an invitation-only seminar at the Naval Weapons Laboratory in Dahlgren, Virginia, where their use for antipersonnel and anti-ballistic-missile energy beams was discussed. No information about their subsequent development has since been made public, and the difficulties of long-range missile tracking argue that ABM beams haven't yet become feasible, but there are no such difficulties in the way of EMR beam weapons for use against unshielded people.

At some UHF power densities there's an insidious moth-to-the-flame allurement, which would increase such a weapon's effectiveness. As discoverer Sol Michaelson described it in 1958, each of the dogs used in his experiments "began to struggle for release from the sling," showing "considerable agitation and muscular activity," yet "for some reason the animal continues to face the horn." Perhaps as part of the same effect, UHF beams can also induce muscular weakness and lethargy. In Soviet experiments with rats in 1960, five minutes of exposure to 100,000 microwatts reduced swimming time in an endurance test from sixty minutes to six.

Allen Frey's discovery that certain pulsed microwave beams increased the permeability of the blood-brain barrier could be turned into a supplemental weapon to enhance the effects of drugs, bacteria, or poisons.

The calcium-outflow windows discovered by Ross Adey could be used to interfere with the functioning of the entire brain.

In the early 1960s Frey found that when microwaves of 300 to 3,000 megahertz were pulsed at specific rates, humans (even deaf people) could "hear" them. The beam caused a booming, hissing, clicking, or buzzing, depending on the exact frequency and pulse rate, and the sound seemed to come from just behind the head.

At first Frey was ridiculed for this announcement, just like many radar technicians who'd been told they were crazy for hearing certain radar beams. Later work has shown that the microwaves are sensed somewhere in the temporal region just above and slightly in front of the ears. The phenomenon apparently results from pressure waves set up in brain tissue, some of which activate the sound receptors of the inner ear via bone conduction, while others directly stimulate nerve cells in the auditory pathways. Experiments on rats have shown that a strong signal can generate a sound pressure of 120 decibels, or approximately the level near a jet engine at takeoff.

Obviously such a beam could cause humans severe pain and prevent all voice communication. That the same effect can be used more subtly was demonstrated in 1973 by Dr. Joseph C. Sharp of the Walter Reed Army Institute of Research. Sharp, serving as a test subject himself, heard and understood spoken words delivered to him in an echo-free isolation chamber via a pulsed-microwave audiogram (an analog of the words' sound vibrations) beamed into his brain. Such a device has obvious applications in covert operations designed to drive a target crazy with "voices" or deliver undetectable instructions to a programmed assassin. There are also indications that other pulse frequencies cause similar pressure waves in other tissues, which could disrupt various metabolic processes. A group under R. G. Olsen and J. D. Grissett at the Naval Aerospace Medical Research Laboratory in Pensacola has already demonstrated such effects in simulated muscle tissue and has a continuing contract to find beams effective against human tissues.

In the 1960s Frey also reported that he could speed up, slow down, or stop isolated frog hearts by synchronizing the pulse rate of a microwave beam with the beat of the heart itself. Similar results have been obtained using live frogs, indicating that it's technically feasible to produce heart attacks with a ray designed to penetrate the human chest.

In addition to the methods of damaging or killing people with EMR, there are several ways of controlling their behavior. Ross Adey and his colleagues have shown that microwaves modulated in various ways can force specific electrical patterns upon parts of the brain. Working with cats they found that brain waves appearing with conditioned responses could be selectively enhanced by shaping the microwaves with a rhythmic variation in amplitude (height) corresponding to EEG frequencies. For example, a 3-hertz modulation decreased 10-hertz alpha waves in one part of the animal's brain and reinforced 14-hertz beta waves in another location.

Some radar can find a fly a kilometer away or track a human at twenty-five miles, and several researchers have suggested that focused EMR beams of such accuracy could bend the mind much like electrical stimulation of the brain (ESB) through wires. We know of ESB's potential for mind control largely through the work of Jose Delgado. One signal provoked a cat to lick its fur, then continue compulsively licking the floor and bars of its cage. A signal designed to stimulate a portion of a monkey's thalamus, a major midbrain center for integrating muscle movements, triggered a complex action: The monkey walked to one side of the cage, then the other, then climbed to the rear ceiling, then back down. The animal performed this same activity as many times as it was stimulated with the signal, up to sixty times an hour, but not blindly—the creature still was able to avoid obstacles and threats from the dominant male while carrying out the electrical imperative. Another type of signal has made monkeys turn their heads, or smile, no matter what else they were doing, up to twenty thousand times in two weeks. As Delgado concluded, "The animals looked like electronic toys."

Even instincts and emotions can be changed: In one test a mother giving continuous care to her baby suddenly pushed the infant away whenever the signal was given. Approach-avoidance conditioning can be achieved for any action simply by stimulating the pleasure and pain centers in an animal's or person's limbic system.

Eventual monitoring of evoked potentials from the EEG, combined with radio-frequency and microwave broadcasts designed to produce specific thoughts or moods, such as compliance and complacency, promises a method of mind control that poses immense danger to all societies—tyranny without terror. Scientists involved in EEG research all say the ability is still years away, but for all we could sense of it, it could be happening right now. Conspiracy theories aside, the hypnotic familiarity of TV and radio, combined with the biological effects of their broadcast beams, may already constitute a similar force for mass standardization, whether by design or not..

The potential dangers of televised lethargy are no yawning matter. It's well known that relaxed attention to any mildly involving stimulus, such as a movie or TV program, produces a hypnoid state, in which the mind becomes especially receptive to suggestion. Other inducers of hypnoid states include light sleep, daydreams, or short periods of time spent waiting for some predetermined signal or action, such as a traffic light.

The Central Intelligence Agency funded research on electromagnetic mind control at least as early as 1960, when the notorious MKULTRA program, mostly concerned with hypnosis and psychedelic drugs, included money for adapting bioelectric sensing methods (at that time primarily the EEG) to surveillance and interrogation, as well as for finding "techniques of activation of the human organism by remote electronic means." In testimony before the Senate Subcommittee on Health and Scientific Research on September 21, 1977, MKULTRA director Dr. Sidney Gottlieb recalled: "There was a running interest in what effects people's standing in the field of radio energy have, and it could easily have been that somewhere in the many projects someone was trying to see if you could hypnotize somebody easier if he was standing in a radio beam."

Hypnotists often use a strobe light flashing at alpha-wave frequencies to ease the glide into trance. It seems for over thirty years the Communist bloc nations have been using an ELF wave form to do the same thing undetectably and perhaps more effectively. Ross Adey recently lost most of his government grants and has become a bit more loquacious about the military and intelligence uses of EMR. In 1983 he organized a public meeting at the Loma Linda VA hospital and released photos and information concerning a Russian Lida machine. This was a small transmitter that emitted 10-hertz waves for tranquilization and enhancement of suggestibility. The most interesting part was that the box had an ancient vacuum-tube design, and a man who'd been a POW in Korea reported that similar devices had been used there during interrogation.

American interest in the hypnosis-EMR interaction was still strong as of 1974, when a research plan was filed to develop useful techniques in human volunteers. The experimenter, J. F. Schapitz, stated: "In this investigation it will be shown that the spoken word of the hypnotist may also be conveyed by modulated electromagnetic energy directly into the subconscious parts of the human brain—i.e., without employing any technical devices for receiving or transcoding the messages and without the person exposed to such influence having a chance to control the information input consciously." As a preliminary test of the general concept, Schapitz proposed recording the brain waves induced by specific drugs, then modulating them onto a microwave beam and feeding them back into an undrugged person's brain to see if the same state of consciousness could be produced by the beam alone.

Schapitz's main protocol consisted of four experiments. In the first, subjects would be given a test of a hundred questions, ranging from easy to technical, so they all would know some but not all of the answers. Later, while in hypnoid states and not knowing they were being irradiated, these people would be subjected to information beams suggesting answers for some of the items they'd left blank, amnesia for some of their correct answers, and memory falsification for other correct answers. A new test would check the results two weeks later.

The second experiment was to be the implanting of hypnotic suggestions for simple acts, like leaving the lab to buy some particular item, which were to be triggered by a suggested time, spoken word, or sight. Subjects were to be interviewed later. "It may be expected," Schapitz wrote, "that they rationalize their behavior and consider it to be undertaken out of their own free will."

In a third test the subjects were to be given two personality tests. Then different responses to certain questions would be repeatedly suggested, and nonpathological personality changes would also be suggested, both to be evaluated by new testing in a month. In some cases the subjects were to be prehypnotized into talking in their sleep, so the microwave programmer could gear the commands to thoughts already in the brain. Finally, attempts would be made to produce the standard tests of deep hypnotic trance, such as muscular rigidity, by microwave beams alone.

Naturally, since this information was voluntarily released via the Freedom of Information Act, it must be taken with a pillar of salt. The results haven't been made public, so the work may have been inconclusive, and the plans may have been released to convince the Soviets and our own public that American mind-control capabilities are greater than they actually are. On the other hand, the actualities may be so far ahead of this research plan that it was tame enough to release in satisfying FOIA requirements.

How many of the EMR weapons possibilities have actually been developed and/or used? Those not privy to classified information have no way of knowing. There are plenty of rumors. Boris Spassky claimed he'd lost the world chess championship to Bobby Fischer because he was being bombarded with confusion rays. I recall hearing about one secret American experiment in which a scientist was supposedly set up with invitations to three conferences to give the same presentation each time. The first one went fine, but at the last two he was irradiated with ELF waves, reportedly to induce Adey's calcium efflux, and he became confused and ineffective.

Another FOIA release from the Defense Intelligence Agency in 1976 may be revealing. Prepared by Ronald L. Adams and E. A. Williams of Battelle Columbus Laboratories, it's entitled "Biological Effects of Electromagnetic Radiation (Radiowaves and Microwaves), Eurasian Communist Countries." The pages released merely recount Allen Frey's discoveries without mentioning his name, implying instead that only the Reds would be so dastardly as to investigate such things for use as weapons. Immediately after mention of the blood-brain barrier leak phenomenon, a paragraph was deleted, followed by the tantalizing sentence, "The above study is recommended reading material for those consumers who have an interest in the application of microwave energy to weapons." Even without this document, considering the relentless pace of arms development, we would have to be very naive to assume that the United States has no electromagnetic arsenal.

The Soviets may already be using theirs, however, on a scale far beyond that of the Moscow signal. During the U.S. bicentennial celebration of July 4, 1976, a new radio signal was heard throughout the world. It has remained on the air more or less continuously ever since. Varying up and down through the frequencies between 3.26 and 17.54 megahertz, it is pulse-modulated at a rate of several times a second, so it sounds like a buzz saw or woodpecker. It was soon traced to an enormous transmitter near Kiev in the Soviet Ukraine.

The signal is so strong it drowns out anything else on its wavelength. When it first appeared, the UN International Telecommunications Union protested because it interfered with several communications channels, including the emergency frequencies for aircraft on transoceanic flights. Now the woodpecker leaves "holes"; it skips the crucial frequencies as it moves up and down the spectrum. The signal is maintained at enormous expense from a current total of seven stations, the seven most powerful radio transmitters in the world.

Within a year or two after the woodpecker began tapping, there were persistent complaints of unaccountable symptoms from people in several cities of the United States and Canada, primarily Eugene, Oregon. The sensations—pressure and pain in the head, anxiety, fatigue, insomnia, lack of coordination, and numbness, accompanied by a high-pitched ringing in the ears—were characteristic of strong radio-frequency or microwave irradiation. In Oregon, between Eugene and Corvallis, a powerful radio signal centering on 4.75 megahertz was monitored, at higher levels in the air than on the ground. Several unsatisfactory theories were advanced, including emanations from winter-damaged power lines, but most engineers who studied the signal concluded that it was a manifestation of the woodpecker. The idea was advanced that it was being directed to Oregon by a Tesla magnifying transmitter. This apparatus, devised by Nikola Tesla during his turn-of-the-century experiments on wireless global power transmission at a laboratory near Pikes Peak, hasn't been much studied in the West. It reportedly enables a transmitter to beam a radio signal through the earth to any desired point on its surface, while maintaining or even increasing the signal's power as it emerges. Paul Brodeur has suggested that, since the TRW company once proposed a Navy ELF communications system using an existing 850-mile power line that ended in Oregon, the Eugene phenomenon might have been the interaction between a Navy broadcast and Soviet jamming.

He that as it may, the woodpecker continues in operation, and there are several unsettling possibilities as to its main purpose. A former chief of naval research has privately discounted the idea that it's directed against the U.S. population. However, Robert Beck, a Los Angeles physicist who regularly serves as a DOD consultant, told me that the signal has a threefold purpose. He said it acts as a crude over-the-horizon radar that would pick up a massive first strike of U.S. missiles if Soviet spy satellites and other detectors were knocked out. Second, the signal's modulations are an ELF medium for communicating with submarines underwater. Third, he claimed the signal has a biological by-product about which he promised further information. Of course, I haven't been able to contact him since.

Several educated guesses can be made, however. Adey's research suggests that the best way to get an ELF signal into an animal is to make it a pulse modulation of a high-frequency radio signal. That's exactly what the woodpecker is. Within its frequency range, it could be beamed to any part of the world, and it would be picked up and reradiated by the power supply grid at its destination.

Raymond Damadian has theorized that the woodpecker signal is designed to induce nuclear magnetic resonance in human tissues. Damadian, a radiologist at Brooklyn's Downstate Medical Center, patented the first NMR scanner, a device that gives an image of internal organs similar to CAT scanners but using magnetic fields rather than nuclear radiation. As mentioned earlier in this chapter, NMR could greatly magnify the metabolic interference of electropollution or EMR weapons. Maria Reichmanis calculated the pulse frequency that would be required to do this with a radio signal in the woodpecker's range, and she came up with a band centered on the same old alpha rhythm of 10 hertz. And in fact, the signal's pulse is generally about that rate, although it is often a two-part modulation of 4 + 6, 7 + 3, and so on. The available evidence, then, suggests that the Russian woodpecker is a multipurpose radiation that combines a submarine link with an experimental attack on the American people. It may be intended to increase cancer rates, interfere with decision-making ability, and/or sow confusion and irritation. It may be succeeding.

I keep hearing persistent rumors of American transmitters set up to try to nullify the Russians' signal or to affect their people in a similar way. In 1978, Stefan Rednip, an American reporter living in England, claimed access to purloined CIA documents proving the existence of a program called Operation Pique, which included bouncing radio signals off the ionosphere to affect the mental functions of people in selected areas, including Eastern European nuclear installations.

The whole business sounds too much like an undeclared electromagnetic war. However, there are persistent complaints that the American effort is being hampered in a strange way. Shortly after the rigged National Academy of Sciences report on Project Seafarer, for example, the Navy sent a delegation to a meeting at the National Security Agency to complain about an alleged "zap gap" between the United States and the USSR, and to ask other delegates to push for more research money for turning nonthermal EMR effects into weapons. According to one of my Navy contacts, the NSA sent several "experts" who had never done any research on EMR and who firmly advised the Navy to abandon its program. Later he voiced the same suspicions I'd already heard from others: Given the allegedly vigorous Soviet electroweapons research program and the underfunding of ours, he concluded that there is a mole highly placed in the American military science establishment, perhaps in the NSA itself, who is preventing us from acquiring any clear competence in this field.

Unfortunately, my source, having served as a hatchet man for defunding research on the environmental dangers of electropollution, isn't exactly reliable. Complaints of a mole could easily be a blind for a large and intense U.S. EMR weapons program. That there's more going on than meets the eye is clear from my last communication with Dietrich Beischer. In 1977 the Erie Magnetics Company of Buffalo, New York, sponsored a small private conference, and Beischer and I both planned to attend. Just before the meeting, I got a call from him. With no preamble or explanation, he blurted out: "I'm at a pay phone. I can't talk long. They are watching me. I can't come to the meeting or ever communicate with you again. I'm sorry. You've been a good friend. Goodbye." Soon afterward I called his office at Pensacola and was told, "I'm sorry, there is no one here by that name," just as in the movies. A guy who had done important research there for decades just disappeared.

The crucial point to me is that both sides may be embarking on hostilities whose consequences for the whole biosphere no one can yet foresee. Even if the Soviets have begun an electromagnetic war and we're totally unprepared to fight back, I doubt that a simple buildup and retaliation are the best course for our own survival.

The extent of the danger can be dramatized best by considering one last potential weapon. Around 1900, Nikola Tesla theorized that ELF and VLF radiation could enter the magnetosphere, the magnetic field in space around the earth, and change its structure. He has recently been proven right.

The magnetosphere and its Van Allen belts of trapped particles produce many kinds of EMR. Since they were initially studied through audio amplifiers, the first kinds to be discovered, around 1920, were given fanciful names like whistlers, dawn chorus, and lion roars. Many of them result from VLF waves produced by lightning, which bounce back and forth from pole to pole along "magnetic ducts" in the magnetosphere. This resonance amplifies the original VLF waves enormously.

Satellite measurements have proven that artificial energies from power lines are similarly amplified high above the earth, a phenomenon known as power-line harmonic resonance (PLHR). Radio and microwave energy also resonates in the magnetosphere. This amplified energy interacts with the particles in the Van Allen belts, producing heat, light, X rays, and, most important, a "fallout" of charged particles that serve as nuclei for raindrops.

Recent work with sounding rockets has matched specific areas of such ion precipitation with the energy from specific radio stations, and established that the sifting down of charged particles generally occurs east of the EMR source, following the general eastward drift of weather patterns. In 1983, measurements from the Ariel 3 and 4 weather satellites showed that the enormous amount of PLHR over North America had created a permanent duct from the magnetosphere down into the upper air, resulting in a continuous release of ions and energy over the whole continent. In presenting this data at the March 1983 Symposium on Electromagnetic Compatibility in Zurich, K. Bullough reminded the audience that thunderstorms have been 25 percent more frequent over North America between 1930 and 1975 than they were from 1900 to 1930, and suggested that the increased energy levels in the upper atmosphere were responsible.

Since the mid-1970s there has been a dramatic increase in flooding, drought, and attendant hardships due to inconsistent, anomalous weather patterns. It appears likely that these have been caused in part by electropollution and perhaps enhanced, whether deliberately or not, by the Soviet woodpecker signal. It now seems feasible to induce catastrophic climate change over a target country, and even without such weather warfare, continued expansion of the electrical power system threatens the viability of all life on earth.

Critical Connections

It may be hard to convince ourselves that something we can't see, hear, touch, taste, or smell can still hurt us so dreadfully. Yet the fact must be faced, just as we've learned a healthy fear of nuclear radiation. Certain scientists, some perhaps acting in a program of deliberate disinformation, keep telling the public that we still don't know whether electropollution is a threat to human health. That's simply not true. Certainly we need to know more, but a multitude of risks have been well documented.

Three dangers overshadow all others. The first has been conclusively proven: ELF electromagnetic fields vibrating at about 30 to 100 hertz, even if they're weaker than the earth's field, interfere with the cues that keep our biological cycles properly timed; chronic stress and impaired disease resistance result. Second, the available evidence strongly suggests that regulation of cellular growth processes is impaired by electropollution, increasing cancer rates and producing serious reproductive problems. Electromagnetic weapons constitute a third class of hazards culminating in climatic manipulation from a sorcerer's-apprentice level of ignorance.

There may be other dangers, less sharply defined but no less real. All cities, by their very nature as electrical centers, are jungles of interpenetrating fields and radiation that completely drown out the earth's background throb. Is this an underlying reason why so many of them have become jungles in another sense as well? Is this a partial explanation for the fact that the rate of suicide between the ages of fifteen and twenty-four rose from 5.1 per 100,000 in 1961 to 12.8 in 1981? Might this be an invisible and thus overlooked reason why so many governmental leaders, working at the centers of the most powerful electromagnetic networks, consistently make decisions that are against the best interests of every being on earth? Is the subliminal stress of electronic smog misinterpreted as continual threats from outside—from other people and other governments? In addition, if Teilhard de Chardin's noosphere exists, our artificial fields must mask it many times over, literally disconnecting us from life's collective wisdom. This is not to ignore the plain fact of evil, but it often seems there must be some other reason why today's power elite are so willing to bring the whole world to the brink of so many different kinds of destruction. Maybe they literally can't hear the earth anymore.

Everyone worries about nuclear weapons as the most serious threat to our survival. Their danger is indeed immediate and overwhelming. In the long run, however, I believe the ultimate weapon is manipulation of our electromagnetic environment, because it's imperceptibly subtle and strikes at the core of life itself. We're dealing here with the most important scientific discovery ever - the nature of life. Even if we survive the chemical and atomic threats to our existence, there's a strong possibility that increasing electropollution could set in motion irreversible changes leading to our extinction before we're even aware of them.

All life pulsates in time to the earth, and our artificial fields cause abnormal reactions in all organisms. Magnetic reversals may have produced the "great dyings" of the past by disrupting biocycles so as to cause stress, sterility, birth defects, malignancies, and impaired brain function. Human activities may well have duplicated in three decades what otherwise would have taken five thousand years to develop during the next reversal. What will we do if the incidence of deformed children rises to 50 percent, if the cancer rate climbs to 75 percent? Will we be able to pull the plug?

Somehow these dangers must be brought into the open so forcefully that the entire population of the world is made aware of them. Scientists must begin to ask and seek answers to the questions raised in this chapter, regardless of the effect on their careers. These energies are too dangerous to be entrusted forever to politicians, military leaders, and their lapdog researchers.

Since our civilization is irreversibly dependent on electronics, abolition of EMR is out of the question. However, as a first step toward averting disaster, we must halt the introduction of new sources of electromagnetic energy while we investigate the biohazards of those we already have with a completeness and honesty that have so far been in short supply. New sources must be allowed only after their risks have been evaluated on the basis of the knowledge acquired in such a moratorium.

With an adequately funded research program, the moratorium need last no more than five years, and the ensuing changes could almost certainly be performed without major economic trauma. It seems possible that a different power frequency—say 400 hertz instead of 60—might prove much safer. Burying power lines and providing them with grounded shields would reduce the electric fields around them, and magnetic shielding is also feasible.

A major part of the safety changes would consist of energy-efficiency reforms that would benefit the economy in the long run. These new directions would have been taken years ago but for the opposition of power companies concerned with their short-term profits, and a government unwilling to challenge them. It is possible to redesign many appliances and communications devices so they use far less energy. The entire power supply could be decentralized by feeding electricity from renewable sources (wind, flowing water, sunlight, georhermal and ocean thermal energy conversion, and so forth) into local distribution nets. This would greatly decrease hazards by reducing the voltages and amperages required. Ultimately, most EMR hazards could be eliminated by the development of efficient photoelectric converters to be used as the primary power source at each point of consumption. The changeover would even pay for itself, as the loss factors of long-distance power transmission—not to mention the astronomical costs of building and decommissioning short-lived nuclear power plants—were eliminated. Safety need not imply giving up our beneficial machines.

Obviously, given the present technomilitary control of society in most parts of the world, such sane efficiency will be immensely difficult to achieve. Nevertheless, we must try. Electromagnetic energy presents us with the same imperative as nuclear energy: Our survival depends on the ability of upright scientists and other people of goodwill to break the military-industrial death grip on our policy-making institutions.

Postscript: Political Science

An important scientific innovation rarely makes its way by gradually winning over and converting its opponents: it rarely happens that Saul becomes Paul. What does happen is that its opponents gradually die out and that the growing generation is familiarized with the idea from the beginning.
—Max Planck

Dispassionate philosopher inquiring into nature from the sheer love of knowledge, single-minded alchemist puttering about a secluded basement in search of elixirs to benefit all humanity—these ideals no longer fit most scientists. Even the stereotype of Faust dreaming of demonic power is outdated, for most scientists today are overspecialized and anonymous—although science as a whole is somewhat Mephistophelian in its disregard for the effects of its knowledge. It's a ponderous beast, making enormous changes in the way we live but agonizingly slow to change its own habits and viewpoints when they become outmoded.

The public's conception of the scientist remains closest to its image of the philosopher—cold and logical, making decisions solely on the basis of the facts, unswayed by emotion. The lay person's most common fear about scientists is that they lack human feelings. During my twenty-five years of research I've found this to be untrue yet no cause for comfort. I've occasionally seen our species' nobler impulses among them, but I've also found that scientists as a group are at least as subject to human failings as people in other walks of life.

It has been like this throughout the history of science. Many, perhaps even most, of its practitioners have been greedy, power-hungry, prestige-seeking, dogmatic, pompous asses, not above political chicanery and outright lying, cheating, and stealing. Examples abound right from the start. Sir Francis Bacon, who in 1620 formulated the experimental method on which all technical progress since then has been founded, not only forgot to mention his considerable debt to William Gilbert but apparently plagiarized some of his predecessor's work while publicly belittling it. In a similar way Emil Du Bois-Reymond based his own electrical theory of the nerve impulse on Carlo Matteucci's work, then tried to ridicule his mentor and take full credit.

Many a genius has been destroyed by people of lesser talent defending the status quo. Ignaz Semmelweis, a Hungarian physician who practiced in Vienna during the mid-nineteenth century, demanded that his hospital colleagues and subordinates wash their hands, especially when moving from autopsies and sick wards to the charity childbirth ward he directed. When the incidence of puerperal fever and resultant death declined dramatically to well below that of the rich women's childbirth ward, proving the importance of cleanliness even before Pasteur, Semmelweis was fired and vilified. His livelihood gone, he committed suicide soon afterward.

The principal figure who for decades upheld the creed that dedifferentiation was impossible was Paul Weiss, who dominated biology saying the things his peers wanted to hear. Weiss was wrong, but along the way he managed to cut short a number of careers.

For many years the American Medical Association scorned the idea of vitamin-deficiency diseases and called the EEG electronic quackery. Even today that august body contends that nutrition is basically irrelevant to health. As the late-eighteenth-century Italian experimenter Abbe Alberto Fortis observed in a letter chiding Spallanzani for his closed-minded stance on dowsing, "... derision will never help in the development of true knowledge."

In the past, these character flaws couldn't wholly prevent the recognition of scientific truths. Both sides of a controversy would fight with equal vehemence, and the one with better evidence would usually win sooner or later. In the last four decades, however, changes in the structure of scientific institutions have produced a situation so heavily weighted in favor of the establishment that it impedes progress in health care and prevents truly new ideas from getting a fair hearing in almost all circumstances. The present system is in effect a dogmatic religion with a self-perpetuating priesthood dedicated only to preserving the current orthodoxies. The system rewards the sycophant and punishes the visionary to a degree unparalleled in the four-hundred-year history of modern science.

This situation has come about because research is now so expensive that only governments and multinational corporations can pay for it. The funds are dispensed by agencies staffed and run by bureaucrats who aren't scientists themselves. As this system developed after World War II, the question naturally arose as to how these scientifically ignorant officials were to choose among competing grant applications. The logical solution was to set up panels of scientists to evaluate requests in their fields and then advise the bureaucrats.

This method is based on the naive assumption that scientists really are more impartial than other people, so the result could have been predicted decades ago. In general, projects that propose a search for evidence in support of new ideas aren't funded. Most review committees approve nothing that would challenge the findings their members made when they were struggling young researchers who created the current theories, whereas projects that pander to these elder egos receive lavish support. Eventually those who play the game become the new members of the peer group, and thus the system perpetuates itself. As Erwin Chargaff has remarked, "This continual turning off and on of the financial faucets produces Pavlovian effects," and most research becomes mere water treading aimed at getting paid rather than finding anything new. The intuitive "lunatic twinge," the urge to test a hunch, which is the source of all scientific breakthroughs, is systematically excluded.

There has even been a scientific study documenting how choices made by the peer review system depend almost entirely on whether the experts are sympathetic or hostile to the hypothesis being suggested. True to form, the National Academy of Sciences, which sponsored the investigation, suppressed its results for two years.

Membership on even a few peer review boards soon establishes one's status in the "old boys' club" and leads to other benefits. Manuscripts submitted to scientific journals are reviewed for validity in the same way as grant requests. And who is better qualified to judge an article than those same eminent experts with their laurels to guard? Publication is accepted as evidence that an experiment has some basic value, and without it the work sinks without a ripple. The circle is thus closed, and the revolutionary, from whose ideas all new scientific concepts come, is on the outside. Donald Goodwin, chairman of psychiatry at the University of Kansas and an innovative researcher on alcoholism, has even put it in the form of a law of exasperation: "If it's trivial, you can probably study it. If it's important, you probably can't."

Another unforeseen abuse has arisen, which has lowered the quality of training in medical schools. As the peer review system developed, academic institutions saw a golden opportunity. If the government wanted all this research done, why shouldn't it help the schools with their overhead, such as housing, utilities, bookkeeping, and ultimately the salaries of the researchers, who were part of the faculty? The influx of money corroded academic values. The idea arose that the best teacher was the best researcher, and the best researcher was the one who pulled down the biggest grants. A medical school became primarily a kennel of researchers and only secondarily a place to teach future physicians. To survive in academia, you have to get funded and then get published. The epidemic of fraudulent reports—and I believe only a small percentage of the actual fakery has been discovered—is eloquent testimony of the pressure to make a name in the lab.

There remain today few places for those whose talents lie in teaching and clinical work. Many people who don't care about research are forced to do it anyway. As a result, medical journals and teaching staffs are both drowning in mediocrity.

Finally, we must add to these factors the buying of science by the military. To call it a form of prostitution is an insult to the oldest profession. Nearly two thirds of the $47-billion 1984 federal research budget went for military work, and in the field of bioelectricity the proportion was even higher. While military sponsors often allow more technical innovation than others, their employees must keep their mouths shut about environmental hazards and other moral issues that link science to the broader concerns of civilization. In the long run, even the growth of pure knowledge (if there is such a thing) can't flourish behind this chain link fence.

If someone does start a heretical project, there are several ways of dealing with the threat. Grants are limited, usually for a period of one or two years. The experimenter then must reapply. Every application is a voluminous document filled with fine-print forms and meaningless bureaucratic jargon, requiring many days of data compilation and "creative writing." Some researchers may simply get tired of them and quit. In any case, they must run the same gauntlet of peers each time. The simplest way to nip a challenge in the bud is to turn off the money or keep the reports out of major journals by means of anonymous value judgments from the review committees. You can always find something wrong with a proposal or manuscript, no matter how well written or scientifically impeccable it may be.

Determined rebels use guerrilla tactics. There are so many funding agencies that the left hand often knoweth not what the right hand doeth. A proposal may get by an obscure panel whose members aren't yet aware of the danger. The snowstorms of paper churned out by the research establishment have required the founding of many new journals in each subspecialty. Some of these will accept papers that would automatically be rejected by the big ones. In addition, there's an art to writing a grant proposal that falls within accepted guidelines without specifying exactly what the researchers intend to do.

If these methods succeed in prolonging the apostasy, the establishment generally exerts pressure through the schools. Successful academics are almost always true believers who are happy to curry favor by helping to deny tenure to "questionable" investigators or by harassing them in a number of ways. For example, in 1950 Gordon A. Atwater was fired as chairman of the American Museum of Natural History astronomy department and curator of the Hayden Planetarium for publicly suggesting that Immanuel Velikovsky's ideas should receive a fair hearing. That same year Velikovsky's first book, Worlds in Collision, was renounced by his publisher (Macmillan) even though it was a best seller, because a group of influential astronomers led by Harvard's Harlow Shapley threatened to boycott the textbook department that accounted for two thirds of the company's sales. No matter what one may think of Velikovsky's conclusions, that kind of backstairs persuasion is not science.

As the conflict escalates, the muzzled freethinker often goes directly to the public to spread the pernicious doctrines. At this point the gloves come off. Already a lightning rod for the wrath of the Olympian peers, the would-be Prometheus writhes under attacks on his or her honesty, scientific competence, and personal habits. The pigeons of Zeus cover the new ideas with their droppings and conduct rigged experiments to disprove them. In extreme cases, government agencies staffed and advised by the establishment begin legal harassment, such as the trial and imprisonment that ended the career and life of Wilhelm Reich.

Sometime during or after the battle, it generally becomes obvious that the iconoclast was right. The counterattack then shifts toward historical revision. Establishment members publish papers claiming the new ideas for themselves and omitting all references to the true originator. The heretic's name is remembered only in connection with a condescending catchphrase, while his or her own research programs, if any remain, are defunded and the staff dispersed. The facts of the case eventually emerge, but only at an immense toll on the innovator's time and energy.

To those who haven't tried to run a lab, these may seem like harsh words, unbelievable, even paranoid. Nevertheless, these tactics are commonplace, and I've had personal experience with each and every one of them.

I got a taste of the real world in my very first foray into research. After World War II, I continued my education on the GI Bill, but those benefits expired in 1947. I'd just married a fellow student named Lillian, who had caught my eye during our first orientation lecture, and I needed a summer job to help pay expenses and set up housekeeping. I was lucky enough to get work as a lab assistant in the NYU School of Medicine's surgical research department.

I worked with Co Tui, who was evaluating a recently published method for separating individual amino acids from proteins as a step toward concentrating foods for shipment to the starving. Dr. Co, a tiny man whose black, spiky hair seemed to broadcast enthusiasm, inspired me enormously. He was a brilliant researcher and a good friend. With him I helped develop the assay technique and began to use it to study changes in body proteins after surgery.

I was writing my first scientific paper when I walked to work one morning and found our laboratory on the sidewalk—all our equipment, notes, and materials junked in a big pile. I was told neither of us worked there anymore; we were welcome to salvage anything we wanted from the heap.

The head secretary told me what had happened. This was during a big fund drive to build the present NYU Medical Center. One of the "society surgeons" had lined up a million-dollar donation from one of his patients and would see that it got into the fund, if he could choose a new professor of experimental surgery—now. As fast as that, Co Tui and his people were out. I vowed to Lillian: "Whatever I do in medicine, I'm going to stay out of research."

I'm happy that I wasn't able to keep my promise. The research itself was worth it all. Moreover, I don't want to give the impression that I and my associates were alone against the world. Just when hope seemed lost, there was always a crucial person, like Carlyle Jacobsen or the research director's secretary, to help us out. However, right from my first proposal to measure the current of injury in salamanders, I found that research would mean a constant battle, and not only with administrators.

Before I began, I had to solve a technical problem with the electrodes. Even two wires of the same metal had little chemical differences, which gave rise to small electrical currents that could be misinterpreted as coming from the animal. Also, the slightest pressure on the animal's skin produced currents. No one understood why, but there they were. I found descriptions in the older literature of silver electrodes with a layer of silver chloride applied to them, which were reported to obviate the false interelectrode currents. I made some, tested them, and then fitted them with a short length of soft cotton wick, which got rid of the pressure artifact. When I wrote up my results, I briefly described the electrodes. Afterward I received a call from a prominent neurophysiologist who wanted to visit the lab. "Very nice," I thought. "Here's some recognition already." He was particularly interested in how the electrodes were made and used. Some months later, damned if I didn't find a paper by my visitor in one of the high-class journals, describing this new and excellent electrode he'd devised for measuring direct-current potentials!

A couple of years later, while Charlie Bachman and I were looking for the PN junction diode in bone, I was asked to give a talk on bone electronics at a meeting in New York City. The audience included engineers, physicists, physicians, and biologists. It was hard to talk to such a diverse group. The engineers and physicists knew all about electronics but nothing about bone, the biologists knew all about bone but nothing about electronics, and the physicians were only interested in therapeutic applications. At any rate, I reviewed some bone structure for the physicists and some electronics for the biologists, and then went on to describe my experiments with Andy Bassett on bone piezoelectricity.

I probably should have sat down at that point, but I thought it would be nice to talk about our present work. The rectifier concept was tremendously exciting to me, and I thought we might get some useful suggestions from the audience, so I described the experiments showing that collagen and apatite were semiconductors, and discussed the implications. After each talk a short time was set aside for questions and comments, generally polite and dignified. However, as soon as I finished, a well-known orthopedic researcher literally ran up to the audience microphone and blurted out, "I have never heard such a collection of inadequate data and misconceptions. It is an insult to this audience. Dr. Becker has not presented satisfactory evidence for any semiconducting property in bone. The best that can be said is that this material may be a semi-insulator."

Semiconductors are so named because then properties place them between conductors and insulators, so you could very well call them semi-insulators; the meaning would be the same. My opponent was playing a crude game. While saying these derogatory things about me, he was actually agreeing with my conclusion, merely u»ing a different term.

This man's antagonism had begun a couple of years before. When Andy Bassett and I had finished our work on the piezoelectric effect in bone, we wrote it up, submitted it to a scientific journal, and got it accepted. Unbeknownst to us, this fellow had been working on the same thing, but hadn't gotten as far in his experiments as we. Somehow he learned of our work and its impending publication. He called Andy, asking us to delay our report until he was ready to publish his own data. Andy called me to talk it over. What counts in the scientific literature is priority; he was asking us to surrender it. There was no ethical basis for his request, and I would never have thought of asking him to delay had the situation been reversed. I said, "Not on your life." Our paper was published, and we'd acquired a "friend" for life.

Now there he was at the microphone trying to scuttle my presentation with a little ambiguous double-talk. I thought, "He must be doing the same work as we are again. If he wins this encounter, I'll have trouble getting my data published, and he'll have a clear field for his." Instead of defending the data, I explained that semi-insulator and semiconductor were one and the same. I said I was surprised he didn't know that, but I appreciated his approval of my data! Someone else in the audience stood up in support of my position, and the crisis was past. The lab isn't the only place a scientist has to stay alert.

In 1964, soon after the National Institutes of Health approved the grant for our continuing work on bone, I received the VA's William S. Middleton Award for outstanding research. That's a funny little story in itself. The award is given by the VA's Central Office (VACO), whose members had already decided on me, but candidates must be nominated by regional officers, and the local powers were determined I shouldn't get it. Eventually VACO had to order them to nominate me.

The award put me on salary from Washington instead of Syracuse, and due to pressure from VACO I was soon designated the local chief of research, replacing the man who signed all the papers at once. I was determined to put the research house in order, and I instituted a number of reforms, such as public disclosure of the funding allocations, and productivity requirements, no matter how prominent an investigator might be. Many of the reforms have been adopted throughout the VA system. They didn't make me more popular, however. Over the next several years there was continuous pressure from the medical school to allocate

VA research funds for people I felt were of little value to the VA program itself; thus the money would have constituted a grant to the school. I knew that if I didn't deliver I would eventually be removed from the position of chief of research. In that case, I would go back on a local clinical salary and my research program would again be in jeopardy. Therefore, at the beginning of 1972 I applied for the position of medical investigator in the VA research system, a post in which I would be able to devote up to three fourths of my time to research. I was accepted. The job was to begin a few months later; in the meantime I continued as chief of research.

Apparently my new appointment escaped the notice of my local opponents. I'd accepted several invitations to speak at universities in the South and combined them all into a week's trip. I left the office a day early to prepare my material and pack. While I was still home, my secretary called. She was crying, and said she'd just gotten a memo firing me as chief of research and putting me to work as a general-duty medical officer in the admitting office. This not only would have closed our lab, but also would have kept me from practicing orthopedic surgery.

It was a nice maneuver but, fortunately for me, it wasn't legal. As medical investigator, I could be fired only by Washington, and the local chief of staff soon got a letter from VACO ordering him to reinstate me.

Soon I began to get on some "enemies lists" at the national level, too. In December 1974 I got word that our basic NIH grant (the one on bone) hadn't been renewed. No reasons were given. This was highly irregular, since applicants normally got the "pink sheets" with at least the primary reviewer's comments, so they could find out what they'd done wrong. Instead I was told I could write to the executive secretary for a "summary" of the deliberations.

The summary was half a page of double-spaced typing. It said my proposal had been lacking in clarity and direction, and that the experimental procedures hadn't been spelled out in enough detail. The main problem seemed to be that I was planning to do more than the reviewer thought I could do with the money I was requesting. In addition, my report on the perineural cell research with Bruce Baker was criticized as "data poor." The statement concluded: "On the other hand, there are some areas which appear to be worthy of support and are reasonably well described, e.g., bone growth studies, regenerative growth, and electrical field effects."

I was, to say the least, puzzled. The subjects "worthy of support" were precisely the main ones we were working on. It didn't make any sense until I reflected that this was just after I'd helped write the first Sanguine report and had begun to testify about power line dangers before the New York Public Service Commission. Perhaps the Navy was pressuring the NIH to shut me up.

If somebody at the federal level was trying to lock me out as early as 1974, he forgot to watch all the entrances, for my proposal of that year on acupuncture was approved. I'd originally tacked this on to the main NIH application, where it was criticized as inappropriate. I merely sent it off to a different study section, which funded it. After a year we had the positive results described in Chapter 13, and I presented them at an NIH acupuncture conference in Bethesda, Maryland. Ours was the only study going at the problem from a strictly scientific point of view, that is, proceeding from a testable hypothesis, as opposed to the empirical approach of actually putting the needles in and trying to decide if they worked. To the NIH's basic question—is the system of points and lines real?—our program was the only one giving an unequivocal answer: yes.

Nevertheless, when this grant came up for renewal in 1976, it, too, was cut off. The stated reasons were that we hadn't published enough and that the electrical system we found didn't have any relation to acupuncture. The first was obviously untrue—we'd published three papers, had two more in press, and had submitted six others—and the second was obvious pettifogging. How could anyone know what was related to acupuncture before the research had been done? I happened to know the chairman of the NIH acupuncture study section, so I wrote him a letter. He said he was surprised, because the group itself had been pleased with our report. By then it was obvious that something was up.

As of October 1976 we would have no more NIH support. As the money dwindled, we juggled budgets and shaved expenses to cover our costs, and with the help of Dave Murray, who was now chairman of the orthopedic surgery department at the medical school, we kept the laboratory intact and enormously productive. We actually published more research than when we hadn't been under fire.

Early in that same year, however, my appointment as medical investigator had expired, and I had to reapply. Word came back that my application was being "deferred," that is, it had been rejected, but I had the option of reapplying immediately. In her accompanying letter, the director of the VA's Medical Research Service wrote, "While your past record and the strong letters of support [the peer reviews of my application] were considered extremely favorable, the broad research proposal with sketchy detail of technique and methodology was not considered approvable." Now, the instructions for medical investigator applications clearly stated that I was to spell out past accomplishments and indicate future directions only in broad outline. Instead, the director was applying the criteria for first-time grant applicants just entering research. She invited me to resubmit the proposal in the other format. But that would not have helped. Even if the second application was approved, the money would arrive six months after the lab had been closed and we had gone our separate ways.

There was another strange thing about the rejection. By that time all federal granting agencies had to provide the actual reports (with names deleted) of the peers who had done the reviewing. Three out of the four were long, detailed, well-thought-out documents in the standard critique format; they'd been neatly retyped, single spaced, on "reviewer's report" forms with an elite typewriter. One was absolutely lavish in its praise, saying that the VA was fortunate to have me and that the proposed work would undoubtedly make great contributions to medicine. Another was almost as laudatory.

One name had inadvertently been left on one page of the third review. It was the name of a prominent orthopedic researcher with whom I'd disagreed for years about commercialization of bone-healing devices. Since our mutual disregard was well known in the orthopedic service, I feel it was indefensible for the director to ask him to review my application in the first place. Perhaps she expected a more damaging critique from him. He did complain that the proposal was insufficiently detailed. However, his appraisal was quite fair and even said my proposed work was of "fundamental importance to the field of growth and healing." It obviously led up to a recommendation for approval, but the last sentence of that paragraph had been deleted.

The last review was half a page of vague objections, typed double spaced on a pica machine with no semblance of the standard format. There was a revealing mistake ("corrective" tissue instead of connective tissue) that showed the writer had glanced at my proposal for cues but really didn't know what it was about. Strangest of all was the phrasing of this pseudoreview: "[Becker's proposal] is broad and sweeping in scope and contains little documentation for technique and methodology. However, in view of his past record and strong letters of support, a decision should be deferred. ..." The director had used it almost word for word in her letter.

She certainly had no motive for such conduct herself. I'd met her briefly a few years before. In 1966 she'd been appointed chief of research at the Buffalo VA Medical Center and had visited Syracuse to see how I'd organised the program there. Our conversation was pleasant but quite innocuious.

The incentive came from elsewhere. In 1982 a former chief of naval research, with whom I'd made friends because of our common interest in spinal cord regeneration, told me the Navy had protested the closing of the lab. He said it had been a deliberate action coming from some level higher than the VA, NIH, or NAS. Somewhere the decision was made to "shut this guy up," he said, but he wanted me to know that he and the Navy had had no part in it. At one point during the struggle, Andy Marino got a call from a minor official in the VA. He said, "Listen, we're under an awful lot of pressure here to shut you fellows off. Can't you just back off a little bit, stop saying all these things, just sort of downgrade that public hearing you're involved with?" This was the reasonable approach—"We're only trying to help you"—so Andy spent some time sympathizing with this guy's position. Then when he finally asked him, "Where are you getting the pressure from?" Andy got an answer: "Mostly from DOD."

At about the same time I received a clear message about what was going on from one of the scientists from the "other side," one of those who were doing a lot of well-funded work for the military but publishing little. During a recess in a scientific meeting, I encountered him in an empty lobby. Glancing over his shoulder, he drew me aside to the windows and told me my only trouble was that I was going public. He said he and all the rest of us knew there were nonthermal effects and hazards, but we had to keep it quiet. I replied that if no one "went public," the situation could never be corrected and a lot of people would suffer needlessly. He told me that was no concern of mine and predicted my attitude would ruin my career. Well, I could only agree with him on the last point. I agreed that it probably already had, but at least I had a clear conscience.

Rather than sit and wait for the executioner, I tried the back door. I sent a detailed merit-review proposal, the kind filed by beginning researchers, to a newly established section on rehabilitation research. Instead of applying through VACO, however, I routed it through the VA regional office in Boston.

While I was waiting for a reply, I had to fend off another attack. Before testifying at the PSC hearings, I'd notified my supervisor. With a VA lawyer present, he'd told me in person that this was just the kind of thing the VA wanted to fulfill its public obligations, and he even sent me a letter to that effect. Then, just after I appeared on 60 Minutes and talked about Sanguine, power lines, and the rigging of the NAS committee, I heard the va Personnel Office was investigating me for engaging in nonapproved work - the testimony - on government time. I'd gone to the hearings at my own expense to attack a public health problem, but never mind; they were still accusing me of stealing from the taxpayers. If the charge was upheld, I would have to pay back part of my salary and grants based on all the time I'd spent researching, preparing, and giving the evidence.

Soon a company spy showed up in my office. I knew anything I said would be carried straight to the top, so I mentioned, "just between you and me," that I was in a position to reveal infractions committed by the local administration that were more serious than their charge against me. The next day I got a call from the director: There was no audit being planned; that was just a rumor; it was all a big mistake.

Soon I got word that my funding ploy had worked. The rehabilitation proposal was approved in routine fashion, and we were in business through 1979. I'm sure there was hell to pay at VACO when the phantom leader of the posse found that one of the director's underlings had let us slip the noose. However, I knew there was no way we'd ever finagle another grant, and I told everybody to start lining up another job well in advance. As for me, I'd decided to retire as soon as I became eligible in 1978. I was tired and discouraged. There was almost no interest in the evidence I'd accumulated for the DC perineural system, and I was pretty sure I'd never be able to do any research after this latest grant ran out.

Then a magazine article provided some support that induced me to continue for a while. In 1976, Smithsonian ran a piece called "If a Newt Can Grow a New Limb, Maybe We Can." The author, Robert Bahr, had written a popularized but accurate account after looking through the scientific literature and finding, among others, our papers on rat limb regeneration. Shortly afterward, I received a call from Don Yarborough, a congressional lobbyist for the American Paralysis Association, a group of people who were dissatisfied with the then-prevalent idea that nothing could be done for paraplegics. Obviously, spinal cord regeneration was the ultimate answer if it was possible. Yarborough asked if what Bahr had written was true.

When I said it was, he enthusiastically asked for more details. I told him the problems we'd encountered, both scientific and political, and the fact that I was soon to retire and disband the lab. He asked me to make no final decisions until he brought this situation to the attention of his contacts in Congress.

Shortly thereafter, I was asked to see Senator Alan Cranston, then chairman of the Senate Committee on Veterans' Affairrs. Steve Smith was also invited. Also present were representatives from NIH and other agencies, who soon established their position. If the senator wanted to support crackpot stuff like this, they fumed, he would have to see to it that additional money was given to NIH, since they wouldn't fund this work at the expense of "good, solid research projects."

However, Cranston obviously was interested, and he must have leaned on the VA a little, for its policy changed briefly. The administrators expanded rehabilitation research, appointing a well-known orthopedic surgeon named Vernon Nickel to head it. As soon as he hit Washington, he called and asked what he could do for me. I told him that for years I'd been wanting to organize an international symposium on mechanisms of growth control and their clinical promise. Before I'd even finished describing the idea he asked how much I needed. I said $25,000, and he said it was on its way. Predictably, Vernon didn't last long in the capital; he was gone right alter I organized the conference, but by then it was too late to cancel the funds.

The meeting was held in September 1979, and it exceeded all expectations. Every important researcher in the field was there, except for Meryl Rose, who can't stand crowds, and Marc Singer, who was ill. It brought together in one set of proceedings irrefutable evidence that a knowledge of the bioelectricity of growth would lead to incredible breakthroughs in medicine. Since then, a few more grants for regeneration work have opened up, and articles on bioelectromagnetism have begun to appear more often in the journals.

At a June 1978 meeting in Washington, requested by Senator Cranston to plan more regeneration research, my colleagues and I gave (we thought) an exciting and scientifically successful preview of the work that would be presented at the full-scale symposium a year later. At the end, however, the director stood up and said our work was largely erroneous and wholly without significance. "We see absolutely no reason to change the direction of VA research," she pontificated. "We see no reason to expand any programs in the area of regeneration." That was that for Cranston's initiative. The meeting appeared to be just a setup so the VA bigwigs could write him a letter saying they'd duly considered the work of all the experts, weighed it in the balance, and found it wanting.

However, my co-workers and I decided to give it the old college try for more funds. Our grant expired December 31, 1979, and we had to submit our new proposal by the preceding January. By now our nemesis was assistant chief medical director, so it was impossible to sneak one around her. The VA seems to change its application instructions every year, and this year we were specifically told our program was a "type III." Therefore we weren't to submit a detailed write-up of each proposed experiment. By this time we were obviously competent to do them, and we would be judged on the basis of past productivity. On that score we had no worries. We'd led the way in bioelectricity research and we were arranging the first conference ever in our specialty of growth control. Moreover, after our application was filed, during the symposium, I arranged for the chiefs of the various VA research sections to visit our lab, inspect the facilities, meet my replacement (Dave Murray), and discuss our future course. They said they wanted me to continue part time. Suddenly I felt we might survive after all.

Then, around Thanksgiving, came word that our application had been disapproved. The director had pulled the same switch. She'd changed procedures after receiving the proposal, instructing the reviewers to consider it as a "type I" application from a new investigator, with emphasis on future plans. As might be expected, one of the peers said, "This is a poorly written, overly ambitious, incompletely detailed proposal." Another wrote, "Dr. Becker is one of the pioneers in the field of electrically induced osteogenesis and regeneration. His work is visionary and exciting, yet at the same time it is controversial and lacks quantitation." The same old gobbledygook! Overly ambitious? By whose standards? Visionary and exciting? Exactly what do you want in a research project? Quantitation? If you can start to grow a leg back on a rat, what statistics do you want besides the procedures, photos, and the number of experimental animals and controls?

Soon after the rejection, I heard hints through intermediaries that the director might be willing to let us keep the laboratory together if I severed all my connections with it and if the remaining members agreed not to do any research on regeneration or electropollution. With such an agreement, it might be possible to get interim funds pending action on new proposals from my co-workers. The restrictions didn't leave much room to work, but at least my people's livelihood would be spared for a while, so we got back on the peer review treadmill.

Andy Marino, Joe Spadaro, and Dave Murray all submitted proposals of their own late in 1979. Approval of any would have kept the lab open, but the VA director went outside normal procedures to ensure rejection of each one. She chose an ad hoc committee instead of the usual peers to evaluate Spadaro's application. One of the critiques objected to a lack of guidance for him due to the fact that "we understand Becker is retiring." To judge Marino's plan for testing the three methods of electrical osteogenesis for long-term side effects, she bypassed the reviewers who would have been chosen routinely, picking instead someone who had no trainning in orthopedics but who could be counted on to reject anything remotely associated with me.

This man, a well-connected Purdue University embryologist, had for several years been making it a habit to ridicule my work in some of his own papers, while using it without credit as the basis of his own in other publications. In 1978, in response to a Saturday Review article describing regeneration work by me and others, a member of his lab had written a long, vituperative letter to the editor accusing me of "bad science" that had "made life difficult" for real researchers like him and his partners. He accused me of falsifying data on the rat limb experiment. He said the results I'd reported in three days were impossible, even though he'd never bothered to repeat the experiment himself. He ridiculed a claim that blastemas arose from white blood cells, an assertion that neither I nor anyone else had ever made. He accused me of misquoting other scientists' work to support my own ends. He charged me and Stephen Smith with trying to pass off a photo of an intact frog as one that had regrown a complete leg in one of Smith's experiments. Finally, he castigated me for avoiding scientific journals in favor of publishing my results in the popular press, even though I'd had well over a hundred papers printed in the peer-reviewed literature.

Steve Smith was moved to answer the letter point by point, closing with a denunciation of establishment scientists who responded to new ideas with ridicule and slander. Furthermore, he concluded, "I do not understand the rationale of those who feel that research is some obscure process whose results should be made known to the public only as a series of miraculous revelations. The decision as to what kind of research to support is essentially a political one in this day and age, and I firmly believe that an informed public opinion is a much better basis for that decision than general obfuscation."

Normally even the most scurrilous charges are thrashed out in the technical journals, however inadequate that process may be. The letter writer, however, took it upon himself to send a copy to one of the overseers of NIH funding. I suppose I could consider myself lucky that by then there was no more damage anyone could do me in that quarter.

The eminent embryologist's vendetta continued into 1980. In February of that year, the Purdue Office of Public Information issued a press release lauding the great man's own research and presenting him as a white knight doing single combat against "myths." In it he accused me of "plain ordinary fraud," repeating the accusations from his assistant's letter. As a result, I sent the president of Purdue a long, detailed letter itemizing his employee's actions during the past several years, and threatening to sue both him and the university. Immediately I got a call from the embryologist, in which he claimed he never meant anything like that, and asking what I planned to do. I asked to see some apologies in writing, and soon I received a nice letter expressing "deep regrets" for the "rather garbled" press release and belatedly acknowledging that he would never have done any regeneration work without the stimulus of Smith's experiments, which in turn had descended from mine.

The point of summarizing this embarrassing behavior is that, at about the same time he was talking with the Purdue press office, this gentleman was also writing the "impartial" review of Andy's proposal. We know who did the job because, in the small family of regeneration workers, the fellow's contentious style and his practice of quoting mainly his own research are immediately recognizable. His critique repeated some of the same old charges against my own work, even though the proposal wasn't mine. At one point he objected that one of Andy's experimental controls was to use the devices on healthy bone. He sniffed, "I am not aware that physicians are electrically stimulating uninjured bone in humans." Of course they weren't, but that part of the procedure was the perfect way to test for side effects in normal bone—the whole point of the experiment. The reviewer also wondered why some of the animals were to be sacrificed after one or two months while looking for long-term effects, knowing as well as any researcher that this was the normal way of following changes in tissue. After all this, he complained that there were no controls, when in reality the controls were to consist of animals allowed to live out their normal life-spans (which he conveniently overlooked), as well as the very procedures he objected to.

Several paragraphs of the pink sheets were deleted. Perhaps they contained more vitriol than even the VA thought was suitable for me. Another revealing criticism was left in, however. Since Andy had been part of my lab, the reviewer felt that I "would exert considerable influence on the character of the research proposed here," and for him that was an unacceptable contamination.

The denouement was completely predictable. All the proposals were rejected. We continued to work through 1980, supported at a low level of productivity by the interim funds. I kept a small part of the tissue culture lab going with a little money from the company that had made the black boxes for our bone-healing stimulators. In it we did a scaled-down version of Andy's proposed experiment to test for stimulation of malignant cells from electrical osteogenesis—and found it.

From January through June of that year, most of our energy went into drafting one last proposal and trying to get a fair hearing from someone else in the VA. This time we stipulated right in the proposal that I would officially retire and that Dave Murray would become the lab's principal investigator. Of course, the research director knew that I would still talk to these people, and on December 19 we were informed by phone of the final repudiation. As a last-ditch effort, I wrote a letter of appeal to the chief VA administrator, who'd shown favorable interest in our work a couple of years before. I asked him for a hearing and investigation, but to no avail.

The lab ceased to exist on New Year's Day 1981. The local VA chief had the gall to offer Andy and Joe jobs as night administrative officers. Instead, Marino went to work at Louisiana State University Medical School in Shreveport, where he still investigates the positive silver technique on a small scale. Spadaro also remained in orthopedic surgery, at the SUNY Upstate Medical Center right there in Syracuse. Maria Reichmanis decided she'd had enough of professional science, quit research entirely, and got married. This was the end of the foremost group working toward limb and spinal cord regeneration and one of the few bioelectromagnetism labs outside the DOD-industry orbit.

I've taken the trouble to recount my experience in detail for two reasons. Obviously, I want to tell people about it because it makes me furious. More important, I want the general public to know that science isn't run the way they read about it in the newspapers and magazines. I want lay people to understand that they cannot automatically accept scientists' pronouncements at face value, for too often they're self-serving and misleading. I want our citizens, nonscientists as well as investigators, to work to change the way research is administered. The way it's currently funded and evaluated, we're learning more and more about less and less, and science is becoming our enemy instead of our friend.

Glossary

Apatite: The mineral fraction of bone, microscopic calcium phosphate crystals deposited on the preexisting collagen structure of the bone, making it hard. See also Collagen.

Axon: The prolongation of a nerve cell that carries a message, or stimulus, away from the cell body. For example, a motor nerve cell axon carries a contraction stimulus to a muscle. See also Dendrite, Neuron.

Base pair: An association between two of the four fundamental chemical groups that make up all DNA and RNA molecules. Base pairs are the smallest structures that form units of meaning in the genetic code. The more base pairs, the larger the molecule.

Biological cycles: Changes in the activities of living things in an ebb-and-flow pattern. Such changes occur in almost all physical aspects, including sleep-wakefulness, hormone levels, and numbers of white cells in the blood. The predominant pattern is approximately twenty-four hours and usually follows the lunar day closely. See also Orcadian rhythm.

Blastema: The mass of primitive, unspecialized cells that appears at the site of an injury in animals that regenerate. Blastema cells specialize and form the replacement part.

Circadian rhythm: The predominant biological cycle of all living things, from Latin circa ("approximately") and dies ("a day"). See also Biological cycles.

Collagen: A protein that makes up most of the fibrous connective tissue that holds the body's parts together. Tendons, ligaments, and scar tissue are composed almost entirely of collagen. It also forms the basic structure of bone. See also Apatite.

Crystal lattice: The precise, orderly arrangement of atoms in a crystal, forming a netlike structure.

Dedifferentiation: The process in which a mature, specialized cell returns to its original, embryonic, unspecialized state. During dedifferentiation the genes that code for all other cell types are made available for use by de-repressing them. See also DIFFERENTATION, Gene, Redifferentiation.

Dendrite: The prolongation of a nerve cell that carries a message, or stimulus, toward the cell body. For example, sensory nerve cell bodies receive stimuli from receptors in the skin via their dendrites. See also Axon, Neuron.

Differentiation: The process in which a cell matures from a simple embryonic type to a mature, specialized type in the adult. Differentiation involves restricting, or repressing, all genes for other cell types. See also Dedifferentiation, Gene, Redifferentiation.

DNA: The molecule in cells that contains genetic information.

Ectoderm: One of three primary tissues in the embryo, formed as differentiation (cell specialization) is just beginning. The ectoderm gives rise to the skin and nervous system. See also Endoderm, Mesoderm.

Electrode: A device, usually metal, that connects electronic equipment to a living organism for the purpose of measuring electrical currents or voltages in the organism, or delivering a measured electrical stimulus to the organism.

Electrolyte: Any chemical compound that, when dissolved in water, separates into charged atoms that permit the passage of electrical current through the solution.

Embryogenesis: The growth of a new individual from a fertilized egg to the moment of hatching or birth.

Endoderm: One of three primary tissues in the embryo, formed as differentiation (cell specialization) is beginning. It forms the digestive organs. See also Ectoderm, Mesoderm.

Epidermis: The outer layer of skin, having no blood vessels.

Epigenesis: The development of a complex organism from a simple, undifferentiated unit, such as the egg cell. It is the opposite of preformation, in which a complex organism was thought to develop from a smaller, but similarly complex, antecedent, such as the homunculus that some early biologists thought resided in the sperm or egg cell.

Epithelium: A general term for skin and for the lining of the digestive tract.

Exudate: Liquid, sometimes containing cells, that diffuses out from a wound or surface structure of a living organism. Examples are a wound exudate and the slime exudate from the skin of fish.

Galvanotaxis: The movement of a living organism toward or away from a source of electric current.

Gene: A portion of a DNA molecule structured so as to produce a specific effect in the cell.

Gene expression: Specific structure and activity of a cell in response to a group of genes coded for such activity. For example, genes coded for muscle cause a primitive cell to assume the structure and function of a muscle cell.

Glia: A tissue composed of a variety of cells, mostly glial cells, that makes up most of the nervous system. These cells have been considered nonneural in the sense that they cannot produce nerve impulses. Therefore they have been thought unable to transmit information, rather having protective and nutritive roles for the nerve cells proper. This concept is changing. It is now known that glial cells have electrical properties that, while not the same as nerve impulse transmission, enable them to play a role in communication in the body.

Hertz: Cycles per second, the unit for measuring the vibratory rate of electromagnetic radiation. Named for Heinrich Hertz, German physicist who made the first experimental discovery of radio waves in 1888.

Homeostasis: The ability of living organisms to maintain a constant "internal environment." For example, the human body maintains a constant amount of dissolved oxygen in the blood at all times by means of various mechanisms that sense the oxygen level and increase or decrease the breathing rate.

In vitro: An experiment done in a glass dish on part of a living organism.

In vivo: An experiment done on an intact, whole organism.

Magnetosphere: The area around the earth in which the planet's magnetic field exerts a stronger influence than the solar or interplanetary magnetic field. It extends some 30,000 to 50,000 miles from the earth's surface. A prominent feature of the magnetosphere is the Van Allen belts, areas of charged particles trapped by the earth's magnetic field.

Magnetotactic: Active movement toward a magnetic pole.

Mesoderm: One of three primary tissue layers in the embryo, which develop as differentiation (cell specialization) begins. It becomes the muscular and circulatory system in the adult.

Mitosis: Cell division. Actual division takes only a few minutes but must be preceded by a much longer period during which preparatory events, such as duplication of DNA, take place. The entire process generally takes about one clay.

Neoblast: An unspecialized embryonic cell retained in the adult bodies of certain primitive animals and called to the site of an injury to take part in regenerative healing.

Neuroepidermal junction: A structure formed from the union of skin and nerve fibers at the site of tissue loss in animals capable of regeneration. It produces the specific electrical currents that bring about the subsequent regeneration.

Neurohormone: A chemical produced by nerve cells that has effects on other nerve cells or other parts of the body.

Neuron, or neurone: A nerve cell.

Neurotransmitter: A chemical used to carry the nerve impulse across the synapse.

Osteoblast: A cell that forms bone by producing the specific type of collagen that forms bone's underlying structure.

Osteogenesis: The formation of new bone, whether in embryogenesis, postnatal development, or fracture healing.

PEMF: Pulsed electromagnetic field.

Periosteum: A layer of tough, fibrous collagen that surrounds each bone. It contains cells that turn into osteoblasts during fracture healing.

Photoelectric material: A substance that changes light into electrical energy, producing an electric current when light shines on it.

Piezoelectric material: A substance that changes mechanical stress into electrical energy, producing an electrical current when deformed by pressure or bending.

Potential: Another term for voltage, which may at times be limited to a voltage that exists without a current but is potentially able to cause a current to flow if a circuit is completed.

Preformation: See Epigenesis.

Pyroelectric material: A substance that changes thermal energy into electrical energy, producing an electric current when heated.

Redifferentiation: The process in which a previously mature cell that has dedifferentiated becomes a mature, specialized cell again. See also Dedifferentiation , Differentiation .

Salamander: Any of a group of amphibians related to frogs but retaining a tail throughout their lives. Salamanders live in water or moist environments. Most are 2 to 3 inches long, but some grow to more than a foot in length. Since salamanders are vertebrates, with an anatomy similar to ours, and since they regenerate many parts of their bodies very well, they are the animals most commonly used in regeneration research.

Schwann cells: The cells that surround all of the nerves outside of the brain and spinal cord. See also Glia.

Sciatic nerve: The main nerve in the leg. It includes both motor nerve fibers carrying impulses to the leg muscles and sensory nerve fibers carrying impulses to the brain.

Semiconduction: The conduction of electrical current by the movement of electrons or "holes" (the absence of electrons) through a crystal lattice. It is the third and most recently discovered method of electrical conduction. The others are metallic conduction, which works by means of electrons traveling along a wire, and ionic conduction, which works by movement of charged atoms (ions) in an electrolyte. Semiconductors conduct less current than metals but are far more versatile than either of the other types of conduction. Thus they are the basic materials of the transistors and integrated circuits used in most electronic devices today.

Superconduction: The conduction of an electrical current by a specific material that under certain circumstances (generally very low temperatures) offers no resistance to the flow. Such a current will continue undiminished as long as the necessary circumstances are maintained.

Synapse: The junction between one nerve cell and another, or between a nerve cell and some other cell. See also Neurotransmitter.

Undifferentiated: Unspecialized, a term applied to cells that are in a primitive or embryonic state. See also Dedifferentiation, Differentiation, Redifferentiation .

Vertebrate: Any of the animals that have backbones, including all fish, amphibians, reptiles, birds, and mammals. All vertebrates share the same basic anatomical arrangement, with a backbone, four extremities, and similar construction of the muscular, nervous, and circulatory systems.

THE KATHARA™

Bio-Spiritual Healing System

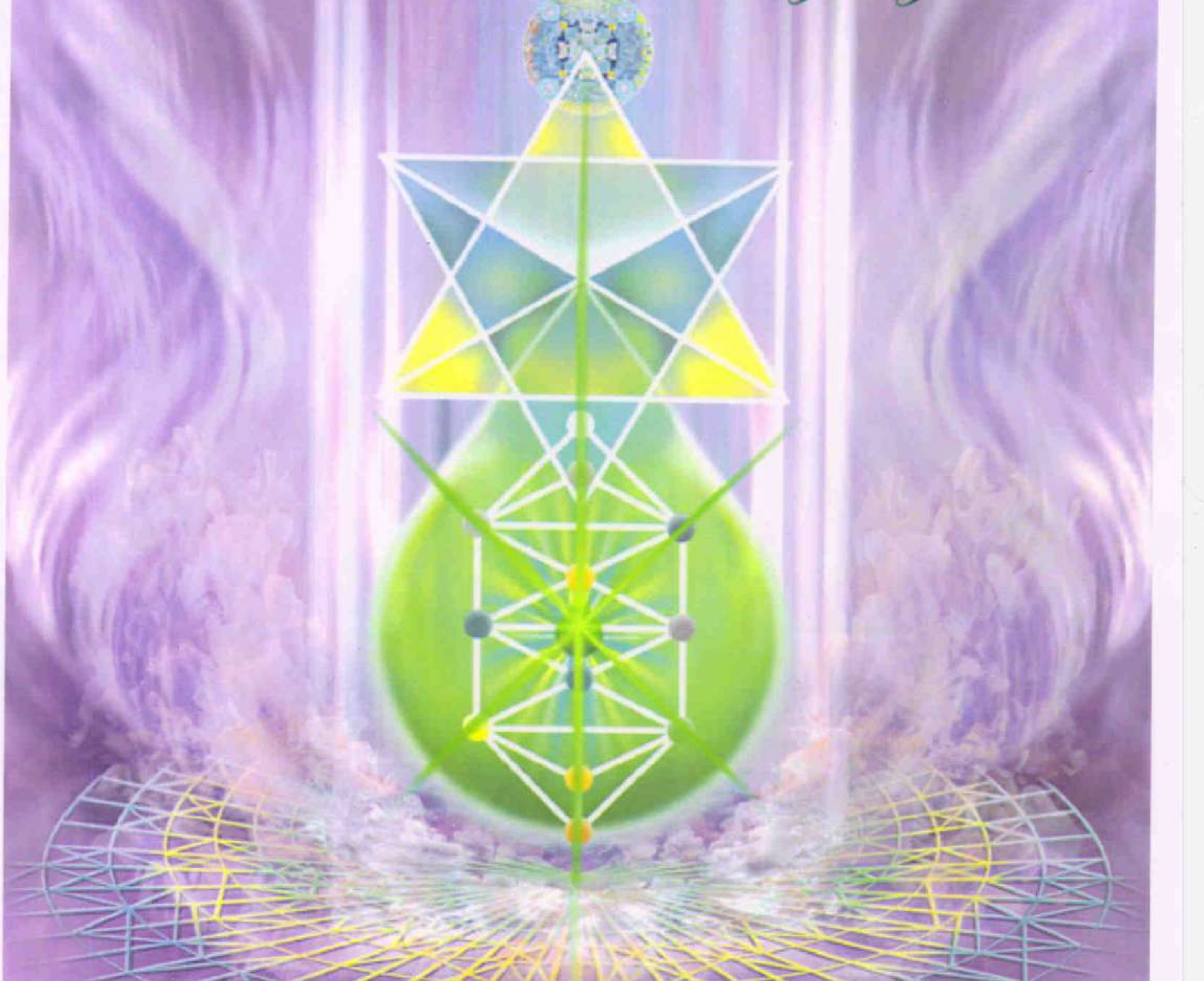

LEVEL 1 CERTIFICATE PROGRAM

by A'sha-yana Deane

The Kathara™ Bio-Spiritual Healing System
is supplied by the Azurite Press of the Melchizedek Cloister Emerald Order
All Materials ©1999-2002 inc. A'sha-yana Deane, Eckr MCEO All Rights Reserved

KATHARA

BIO-SPIRITUAL HEALING SYSTEM

CERTIFICATE PROGRAM

The Maharic Recoding Process

LEVEL - 1

SESSION MANUAL

FOR FACILITATOR-INSTRUCTORS

FOUNDATIONS OF KATHARA HEALING™
KATHARA HEALING ™ FACILITATION
THE MAHARIC RECODING PROCESS ™

An advanced *Spiritual Healing Certificate Program* based upon KEYLONTIC MORPHOGENETIC SCIENCE. Sponsored by the **Kathara Alliance**–an associate of the Azurite Press World-wide Consulate.

All Materials © 1/2000, 2001, 2002 Ashayana Deane, Eckr. MCEO–All Rights Reserved

Table of Contents

Chapter 1: THE CONTEXT OF HEALING - 7- 47
Structure of the Unified Field and the Multi-dimensional Anatomy of Form
Chapter Highlights 37

Chapter 2: THE ELEMENTS OF HEALING ....48 - 112
The Human Body Kathara and the Multi-dimensional Anatomy of the Human Form
Chapter Highlights..102

Chapter 3: KATHARA HEALING BEGINNING APPLICATIONS 113-165
Creating Field Integrity and Building the Knowledge Base
Chapter Highlights..149

Chapter 4: GENERAL KATHARA HEALING FIELD WORK 166-211
Skills to Facilitate Healing
Chapter Highlights 204

Topics and Techniques Locator Index

Chapter 1: The Context of Healing ....Pages 7-47
***Structure of the Unified Field and the Multi-dimensional Anatomy of Form**...8-9*
* The Holographic Template..10-14
* The Structure of Dimensions, Partiki Phasing and Matter Density 15-25
* Universal and Planetary Kathara Grids, the 12-Tree, Kathara Centers and Lines, Scalar Shields, Crystal Seals and Diodic Points 26-31
* The Universal and Internal Templar Complexes...32-34
* The Transduction Sequence and Holographic Reality Construction...........................35-36
Highlights..37-47

Chapter 2: The Elements of Healing ....Pages 48-112
***Human Body Kathara: Multi-dimensional Anatomy of the Human Form...** ...49-50*
* The Maharic Shield and the Imprint for Health...51-52
* The Planetary Bio-feed Interface System (PBIS) 53-54
* The Hierophant Symbol Code, PBIS, the Maharic Shield.......................................55-57
* Morphogenetic Expansion, Growth Cycles, Death, Molecular Compaction......................58-62
* Levels of Human Anatomy, Embodied Kathara 12-Tree Grid, Crystal Shields, Seals, Signets, Hova Bodies, Auric Capsules, Diodic Points, Seed Seals, Chakras, Axiom Lines, the Hara Level, DNA and Merkaba 63-93
* The 12 Human Senses, Apparthi and Human Perception..94-101
Highlights..102-112

Chapter 3: Kathara Healing Beginning Applications .... Pages 113-165
Creating Field Integrity and Building the Knowledge Base
* Tapping the Inner Mentor and Awakening the 6th and 7th Senses.................................114-120
Technique-1: Awakening the Mentor and the 6th and 7th Senses 117-120
* Morphogenetic Repatterning, Frequency Sub-Harmonics, Maharic Seal.........................121-125
* Kathara Points and Kathara Point Sequences..126-131
***Sequence-1* Kathara Points for Regeneration 128-129**
***Sequence-2* Kathra Points for Revitalization 130-131**
Technique-2: The Maharic Seal and the Liquid Light Cleanse 132-137
* Opening the PBIS and the Maharic Seal, the Structure of Multi-dimensional Identity and the 4 Primary Triadic Healing Currents...138-148
Highlights 149-165

Chapter 4: General Kathara Healing Field Work ....Pages 166-211
Skills to Facilitate Healing
* Opening the Healing Channels-Doradic-Phase-1 Current, the Doradic Sphere, Doradic Cords and Doradic Balls...167-174
Technique-3: Opening the Healing Channels-Doradic-Phase-1 Current.......175-177
* The Karmic-Miasmic Imprint Incarnational Time Vectors, Vector Imbedding...................178-180
* Multi-Vector Holographic Recoding through the Maharic Infusion...............................181-183
Technique-4: Multi-Vector Holographic Recoding Kathara Scan.................184-186
Technique-5: The Maharic Infusion 187-190
* Attitudes and Etiquette in Healing, the Motivation to Heal, Postures of Love, the Framework for Healing, and Fully Engaging the Love Vibration 191-197
Exercise-Client Greeting: The Mahara-Space Comfort Touch 198-199
* The 6-Step Maharic Recoding Session, Melchizedek Cloister Ordination and Becoming a Professional Spiritual Healing Facilitator-Instructor...200-203
6-Step Client Maharic Recoding Session: and Client Agreement Form**..........202-203**
Highlights 204-211

Introduction to Kathara Level 2 Mentor – Healing Applications
MERKABIC RADIAL BODY HEALING

Topics and Techniques Locator Index

Chapter – 5: Mahara-Merkabic Master Key Healing – Manifesting the Blueprint for Health, Primal Currents, the DNA sub-Strand Matrix, Merkaba, Psonns, the Trinity Keys and Photo-Dynamic Manifesting

- Universal Life Force Currents .... 214 – 220
- DNA, Fire Letters, the Tribal Shield, Cue Zones .... 221 – 228
- **Technique – 6: The Maharic Quick Seal** 226
- **Technique – 7: Ps – 13 Tribal Shield Activation** .... 227 – 228
- Merkaba Activation, Solar & Lunar Salutations .... 229 – 237
- **Technique – 8: Merkaba Salutation** .... 235 – 236
- DNA, Khundaray Activation, Emerald & Amethyst Awakenings .... 237 – 241
- **Technique – 12: Kee-Ra-ShA Activation Sequence** .... 242 – 247
- Universal Trinity Keys & Merkaba Key Codes 248
- **Technique – 9: Merkaba Trinity Key Induction** 249
- Merkabic Healing .... 251 – 255
- Introduction to the Radial Body Trion-Meajhe Field .... 256 – 257
- The Radial Body Transduction Sequence .... 258 – 262
- Radial Body Healing .... 263 – 266
- The Radial Body and the Merkaba Vehicle .... 267 – 272
- Radial Body Healing and Veca Codes .... 273 – 276
- **Technique – 10: Restoring the Trion Field-Photo Radionic Veca Healing** 277 – 278
- **Technique – 11: Restoring the Meajhe Field-Photosonic Eckasha Healing** 279 – 280
- **Technique – 13: The Song of Lyra, Christos Invocation** .... 281 – 282

Diagrams Locator Index

- **Universal Life Force Currents & the 15 Dimensional Time Matrix** 214
- **Universal Life Force Currents & the Embodied Kathara Grid** 215
- **Universal Life Force Currents, Crystal Seals & Chakras** 216
- **Universal Life Force Currents, Hova Bodies & Auric Levels** 217
- **Universal Life Force Currents, Shields, Signets & Kathara** 218
- **Primal Currents, Silicate Matrix DNA & Subtle Body Anatomy** 219
- **Universal Life Force Currents & DNA Fire Letter Sequences** 220
- **Silicate Matrix DNA & Fire Letters** 222
- **The Tribal Shield Species & Divine Blueprint** 224
- **The Natural Nethra Phase Density 1 Merkaba** 233
- **Merkaba Phases Review** 234
- **The Ar-iE-a Khum-Nar'A** 236
- **The Trinity Master Key Selection Chart** 250
- **Universal Merkaba Fields** 255
- **The Radial Body & Manifestation Mechanics** 258
- **The Eiron Point** 261
- **The Radial Body in Creation of the Hologram** 262
- **The 5 Radial Body Levels** 265
- **Transduction Sequence of the Miasmic Body** 266
- **The Christos-Trion-Meajhe Fields, High Veca Codes & Universal Life Force** 276

Diagrams

Chapter 1

The Holographic Template.... 14
Vibration-Oscillation Correspondences 18
One 15-Dimensional Time Matrix 22
Frequency Bands and Sub-Frequency Bands Within a 15-Dimensional Scale... 23
Partiki Phasing, Harmonics of Manifestation and Matter Density 25
Structure of the Kathara Level-1 12-Tree Grid 27
The Primal Manifestation Template 28
The Universal Templar Complex 34
Transduction Sequence 35

Chapter 2

The Planetary Bio-Feed Interface System (PBIS) 54
The Hierophant Symbol Code 57
Growth Rates and Morphogenetic Expansion 60
Level-1 Kathara 12-Tree Grid Internal Templar 64
Shields, Signets and the Kathara Grid 70
Level-2 Kathara Crystal Seals Grid and Chakras 71
Auric Level, Chakra and Hova Body Correspondences 75
Primary Colors of Chakras 76
The Hara Level 86
12 Primary Axi-A-Tonal Lines 87
Silicate Matrix 12-Strand DNA Template 92
Merkaba Phases 93
The 12 Human Senses 101

Chapter 3

Awakening the Mentor 120
Frequency Sub-Harmonics in the 15-Dimensional Scale 123
The Maharic Seal 136
Liquid Light Cleanse 137
The 5 Hova Bodies 143
Stations of Identity and Levels of Mind in the Hova Bodies.... 144
The 5 Hova Bodies and Levels of Identity in the Time Matrix 145
4 Primary Triadic Currents and Hova Body Correspondences 147
Combined Triadic Currents 148

Chapter 4

Elements of Doradic-Phase-1 Current Activation 173
Intrinsic Process of Activating a Triadic Phase Current 174
Opening the Healing Channels 177
Maharic Recoding Process Elements 183
Amplifying Inner Audio-Visual and Kathara Scan 186
The Maharic Infusion-Client Facilitation 190
Postures of Love 197
Client Agreement Form 202
The 6-Step Maharic Recoding Session 203

INTRODUCTION TO KATHARA BIO-REGENESIS TECHNOLOGIES

COLOR, SYMBOL, SOUND AND BIO-REGENESIS TECHNOLOGIES

BIO-REGENESIS TECHNOLOGIES are tools to assist us in accelerating our expansion of consciousness through natural, gentle stimulation of the personal **DNA TEMPLATE, KUNDALINI ENERGIES, CHAKRAS AND MERKABA VEHICLE.** Bio-Regenesis techniques are built upon the foundations of **KEYLONTIC MORPHOGENETIC SCIENCE** and advanced **SCALAR MECHANICS**, as understood by races of higher evolution; these are paradigms of science not yet recognized or validated by contemporary earth science paradigms. Because Bio-Regenesis technologies are based upon what is viewed by modem earth science paradigms as a speculative rather than verifiable science, I present Bio-Regenesis techniques only as an option to explore in meditation.

Bio-Regenesis technologies were taught as "Common Knowledge" within the Ascension Schools of Pre-Ancient advanced human cultures and were utilized as standard practice in pre-ancient time periods. Bio-Regenesis Technologies are built upon the natural laws of scalar wave mechanics. Bio-Regenesis Techniques are **SPECIFIC APPLICATIONS OF CONSCIOUS ENERGY DIRECTION WITHIN THE MANIFESTATION TEMPLATE OF THE BODY** that stimulate into activity dormant portions of the personal Manifestation Template, which allows for a natural, progressive and accelerated advancement of the personal **ORGANIC EVOLUTIONARY BLUEPRINT** of the 12th-Dimensional Omni-polar Pre-matter Template. Bio-Regenesis Techniques simultaneously create subtle, natural acceleration within the interwoven energy systems of the DNA Template, Chakra System, Merkaba Fields, Kundalini energies, higher dimensional consciousness, and "Subtle Energy Body" anatomy.

In the Bio-Regenesis paradigm it is said that such techniques allow an individual to affect the function of the scalar wave template of the body, working co-creatively **under the direction of the individual's own higher dimensional levels of consciousness.** As the higher dimensional levels of the personal identity, the spiritual identity of the Soul, governs the processes of the personal Manifestation Template, Bio-Regenesis Technologies are completely safe and natural to the organic design of the human being. Using processes of Bio-Regenesis from the earthly conscious perspective expedites the natural processes of energy and consciousness by which the higher identity and earthly identity integrate to become one. In advanced ancient human cultures, Bio-Regenesis Technologies represented a natural attribute of consciousness expansion into the realms of higher evolution and assisted to biologically prepare humans for **STAR GATE** passage or **"ASCENSION"**.

If we can understand that the **HUMAN MIND** is an energy force that **CONTINUALLY GENERATES PATTERNS OF SCALAR WAVES** through the process of **THOUGHT**, it is not difficult to understand how **APPROPRIATELY DIRECTED THOUGHT** can directly influence the function of the personal Manifestation Template scalar blueprint. All thoughts serve this purpose, and directly effect the observable state of the mind-body-spirit system and the manifestation of events within the external life drama. **THOUGHTS ARE THINGS- *SCALAR WAVE CONFIGURATIONS*** of multi-dimensional vibrating patterns of **BI-POLAR ELECTROMAGNETIC ENERGY RADIATION**, which form **SPECIFIC PATTERNS OF SCALAR FREQUENCY** within the personal Manifestation Template.

Undirected thought creates chaotic patterns within the Manifestation Template, while thought directed through clear intention creates ordered patterns within the Manifestation Template. Thought directed with clear intention and **KNOWLEDGE OF THE STRUCTURE OF SUBTLE-BODY ANATOMY**, is a very potent form of **SCALAR POWER**. The patterns of scalar energy frequency that thought projects into the personal Manifestation Template serve as **ELECTROMAGNETIC OPERATIONAL INSTRUCTIONS** within the Template, directly effecting the conditions that will be met in manifest experience. To develop skill in the manifestation process, the conscious "intending" mind must become more familiar with the **"LANGUAGE OF SCALAR WAVES"**, as this is the **"LANGUAGE OF LIGHT AND SOUND"** upon which the Manifestation Template operates. The "language of scalar waves" is "spoken" through the forms of specific wave spectra, forms that appear to the conscious 3-dimensional mind as frequencies of **SOUND**, spectra of LIGHT or **COLOR**, and shapes or **SYMBOLS.**

What we perceive as color is specific spectra of **LIGHT-RADIATION WAVES** that have a specific measurable **WAVELENGTH** and vibrational **ENERGY SIGNATURE.** In using the conscious mind to direct quantities of color, we are **CONSCIOUSLY SELECTING SPECIFIC SCALAR FREQUENCIES** to add to the Manifestation Template. When we use the conscious mind to create images of **SYMBOLS**, we are further **SPECIFYING THE ARRANGEMENT OF SPECIFIC FREQUENCIES**, as they will appear within the Manifestation Template. **SYMBOL FORMS** (even those that compose the alphabets of outer languages) represent **SCALAR-WAVE GUIDES**, and directly affect the contours of the scalar wave blueprint of the Manifestation Template. The Manifestation Template governs the function of the body and consciousness within the manifest experience, and so the **"Language of Scalar Waves"** is the medium through which we gain conscious access to the **CASUAL LEVEL OF MANIFESTATION.** In using **color and symbol** to direct frequency, we are creating a pattern of energy that has a specific vibrational signature or **"TONAL SOUND SIGNATURE"**, which directly interfaces with the energetic vibrational sound signature of the Manifestation Template's scalar grid.

In Bio-Regenesis technology, Color and Symbol are used to direct the **TONAL SCALAR ARRANGEMENTS** of the Manifestation Template, to create natural, desired results within the body-mind-spirit system. There are many complexities within the process of conscious manifestation that one must learn to fully master the contours of space, time and matter. The beginning step is learning to give more finely tuned thought intention **"PROGRAMMING INSTRUCTIONS"** to the personal Manifestation Template. Thoughts and images rendered in **external language** and 3-dimensionally associated symbols affect the **outer layers** of the Manifestation Template, but they do not provided as finely tuned manifestation instructions as do the **"scalar wave languages"** of Color and Symbol image, which direct the contours of vibrational patterns of SOUND. Mastery within the interdimensional spectrum progressively develops as one masters the core **"PROGRAMMING LANGUAGES"** of **Color, Symbol** and **Sound Vibration.**

In **MIND TECHNOLOGIES**, such as **CREATIVE VISUALIZATION** and **VERBAL AFFIRMATION**, beginning skill in **SCALAR TEMPLATE PROGRAMMING** is developed through using emotionally charged images and precise sound patterns in the language to which we are accustomed, as methods to begin gaining conscious control over the processes of personal manifestation. Keylontic Science **BIO-REGENESIS TECHNOLOGIES** take us **ONE STEP DEEPER** into the manifestation process, by directly employing the **LANGUAGE OF SCALAR WAVES** and directing the **ENCODED INSTRUCTIONS** through **PRECISE LOCATIONS** within the body-mind-spirit system and the Manifestation Template upon which this system is built. Bio-Regenesis technologies take us from speaking the language of the 3-dimensional mind into speaking the intrinsic scalar-wave language of the Spirit, which is the **CASUAL ELEMENT** behind and within all manifestation.

Bio-Regenesis Technologies represent applications of Scalar Mechanics that utilize the inherent power of **SUBTLE WAVE FORMS** to created desired change within the core Manifestation Template of the body. Within this process, **COLOR** and **SYMBOL** become the **TOOLS** through which the 3-dimensionallyconscious mind becomes empowered to participate in conscious co-creation with the Soul and higher dimensional spiritual portions of identity. **Color and Symbol** possess power as **MATHEMATICALLY ENCODED** directors of scalar frequency; the ultimate effect of this directed power is the ability to affect the scalar arrangements of the **Manifestation Template**. Intrinsically, **scalar waves** are specific points of vibrating bi-polar energy signatures. Vibration of energy units creates patterns of internal and external **SOUND**, within and beyond the range of 3-dimensional detection. The Manifestation Template can be conceptualized as an **INTERWOVEN FABRIC OF SOUND FREQUENCY**, the "Silent Symphony" of personal and universal being. The tools of Color and Symbol affect the scalar-wave arrangement of the Manifestation Template, and thus alter the vibrational patterns of **SOUND TONES** within the personal blueprint. **COLOR AND SYMBOL are tools for the direction of SOUND**, and intentional application of **SPECIFIC SOUND TONES** can also be used in conjunction with Color and Symbol, to further amplify one's power to interact with the Manifestation Template.

Along with the techniques of Color and Symbol direction, **TONING**, or the **intentional generation of specific sounds to affect the Manifestation Template**, is also utilized in Bio-Regenesis technologies and spiritual expansion programs. "Toning Therapies" have emerged in recent years through the New Age and Holistic Healing paradigms, and have been applied for thousands of years, within the practices of spiritual masters in every religious tradition. SPOKEN LANGUAGE itself is a form of TONING, with its accompanying SYMBOL ALPHABET, and like all Symbols and Sound Tones, our languages directly affect the personal Manifestation Template, and the level of consciousness that can manifest through the body. In ancient cultures, the powers of language were better understood than they presently are in contemporary mainstream society. Particularly within the Priest-casts of ancient religions, language was used more consciously, with an understanding that words, sounds and images would directly effect the consciousness of the masses. Prayers, rituals and songs, some of which remain in use today, were used as a means of *altering consciousness*. Some assisted in spiritual development, others were intentionally used to limit and control the consciousness of populations.

The contemporary earth science paradigm is only beginning to explore the power of sound vibration and the connection between sound, symbol and color is not yet recognized or understood. **In advanced pre-ancient cultures, these sciences of Scalar Mechanics were fully understood and utilized in the most powerful of ways.** Color-Symbol-Sound technologies were not only used in **Bio-Regenesis** applications for personal health and consciousness expansion; they were also used among **GROUPS OF PEOPLE** to directly interface with the scalar-wave Manifestation Template of the **PLANTETARY SHIELDS**. Vortices, Time Portals, Star Gates, Ley Lines and free energy systems generated through the planetary core and solar light emanations were all **DIRECTED THROUGH THE FOCUSED MIND**, utilizing knowledge of Color-Symbol-Sound Scalar Mechanics. Since the fall of ancient cultures, which came as a result of abusing the powers of these Scalar technologies, the Color-Symbol-Sound sequences once used to direct massive amounts of power became the "**HIDDEN AND FORBIDDEN SACRED KNOWLEDGE**".

The **COLORS, SYMBOLS** and **SOUNDS** used in Bio-Regenesis techniques are not simply random choices. The elements of Bio-Regenesis technique are drawn from a **UNIVERSAL STANDARD LANGUAGE OF "SCALAR SPEAK"**, through which **precise scalar frequencies** are generated through **specific Color-Symbol-Sound sequences** to create very specific effects within Manifestation Templates. The Bio-Regenesis Techniques presently offered represent only the beginning of our reawakening to this vast body of Sacred Science knowledge. Many Bio-Regenesis techniques utilizing Color-Symbol sequence scalar mechanics are accompanied by a corresponding exercise in **TONING**. Within the mystical **Ascension Schools** of ancient and pre-ancient times, there existed a very **SACRED SERIES OF TONING SEQUENCES** that were used **to directly interface with the DNA TEMPLATE** and the **AXIOM LINES** in the **PLANETARY**

SHIELD. On a **personal level,** through this specific series of Toning, dormant strands in the DNA Template could be progressively brought into activation, to expedite the process of Spiritual Actualization, Holistic Healing and building of the PERSONAL MERKABA VEHICLE within the bio-energetic field. On the **planetary level,** these **Sacred Toning Sequences** were used to **ENERGETICALLY OPEN** the PLANETARY AXIOM LINES for CLEARING, REPROGRAMMING **AND ACCELERATING THE ACTIVATION** of the PLANETARY SHEILDS. This series of Sacred Toning was collectively called "THE MUSIC OF THE SPHERES".

THE SONGS OF THE SPHERES

The MUSIC OF THE SPHERES represents the ORGANIC TONAL SEQUENCES of EACH DIMENSIONAL FREQUENCY BAND within the 15-Dimensional Time Matrix; the CONSOLIDATED CORE VIBRATIONAL SOUND SIGNATURES of each Dimensional frequency band. Of the 15 core vibrational signatures of the 15 Dimensions in a Time Matrix, 12 of these core signatures serve as the PRIMARY 12 TONAL SEQUENCES upon which each strand of the 12-Strand original human **DNA Template**, and each AXIOM LINE in the 4-Density **Planetary Shields** are structured. In using the **Tone Sequences of the Music of the Spheres**, the 12-Strand DNA Template can be PROGRESSIVELY BROUGHT INTO NATURAL ACTIVATION. Scalar-wave distortions within the Templates for the DNA Strands (which physically manifest as genetic distortions) can be systematically REALIGNED with their natural perfect 12th Dimensional **ORGANIC IMPRINT.** In using the **Music of the Spheres** for work with the **Planetary Shields**, each of the 24 AXIOM LINES can be OPENED and ENCODED WITH THE PERFECT PATTERN OF THE PLANETARY PLASMA BEAM. Transmitting the 12th-Dimensional MAHARIC SHIELD pattern directly into the **Axiom Lines** in the Planetary Shields is the most rapid method to distribute the program of the PLANETARY ORGANIC IMPRINT throughout the 4 Density Levels of the PLANETARY SHIELDS.

In ancient days, the *Music of the Spheres* was used extensively for purposes of self-generated healing and to prepare the physical body for literal passage through the planetary Star Gates ("Ascension"). The **specific tone sequences** for each SINGULAR DIMENSION (which corresponded to one DNA Strand Template in the 12-Strand blueprint and the inherent Dimension of consciousness that would embody through activation of the strand) are called the "SONGS OF THE SPHERES". Each "SONG" contained very specific "**word sounds**" (from the first Universal Spoken language) for **lyrics**, and specific **multi-layered tones** (which generated inaudible Tri-tone Standing Waves) for **melody, harmony** and **overtones**. To be effective, the Songs had to be used through the HUMAN VOICE, externally or mentally as the "SINGING OF SONGS".

The Sacred Songs were sung in combination with directing specific corresponding COLOR and SYMBOL sequences through the bio-energetic field (or through the Planetary Axiom Lines for Planetary Shield work). **The Songs served to "open" specific areas of the DNA Template, or specific Planetary Axiom Lines, to receive instructions directed by the Color and Symbol sequences.** The tones of the Songs generated specific scalar-wave formations that triggered activation of corresponding configurations within the DNA Template and Planetary Shields. Activation of dormant portions of the DNA Strand Templates or Planetary Axiom Lines "opened" the Strand Template or Planetary Shield to directly receive Color-Symbol scalar wave instructions, through which desired outcomes within the personal or planetary body, and within the external reality experience, could be "PROGRAMMED" for manifestation.

The **SCIENCE OF THE MUSIC OF THE SPHERES** was one of the most highly guarded **"SACRED SECRETS"** in pre-ancient cultures, due to the power of interdimensional freedom and conscious manifestation that could be obtained through knowledgeable use of the **SONGS.**

Before the **SONGS OF THE SPHERES** will be fully returned to human awareness, our culture will have to mature enough to demonstrate **spiritually motivated ethical use of power.** Certain SONGS are now being slowly returned, so that we may assist those of higher evolution to realign Earth's Planetary Shield with its original 12th-Dimensional **MAHARIC SHIELD ORGANIC IMPRINT.** One of the **15 SONGS OF THE SPHERES** was provided to us by the Emerald Order. This SONG is used to activate dormant scalar-waves in the 8th **DNA STRAND TEMPLATE**, through which the process of activating the dormant **1st-8th Strands** of the 12-Strand DNA Template can be accelerated, expediting the natural processes of higher dimensional spiritual identity integration.

This SONG is also used to open the **8TH AXIOM LINE** in **GAIA'S PLANETARY SHIELD**, to facilitate realignment of **Earth's Planetary Shields** from **ETHERIC MATTER DENSITY-3.** This SONG was given so that **GROUPS** can begin assisting directly in Planetary Grid work, opening specific Axiom Lines and Vortices in the planetary body to help bring Earth's Planetary Shield Manifestation Template into its organic alignment and balance. The SONG that was given is called the **"Ariea Khum Nar'A".** The English language translation is **"THE SONG OF ORION".** This **SONG OF THE 8th SPHERE** corresponds to the **8th -Dimensional level** of the **UNIVERSAL KATHARA GRID** and its **8TH PRIMARY STAR GATE.** It also corresponds to the **8th PLANETARY AXIOM LINE** within the Etheric Matter Density-3 **GAIAN PLANETARY SHIELD** and to the **8th DNA STRAND TEMPLATE** in the human body. The **8th Dimensional STAR GATE** within the **UNIVERSAL KATHARA GRID** is the Star Gate that opens into the **ORION STAR SYSTEM**, thus the name of the **SONG OF THE 8th SPHERE** is **"THE SONG OF ORION".**

Soon the SONGS that correspond to the 11th and 12th –Dimensional frequencies will be provided. Aside from its use in planetary grid work, when used as a **"MANTRA"**, repeatedly sung as a focus for the mind in meditation, the ***Ariea Khum Nar'A*** serves to stimulate consciousness expansion into the higher dimensional fields and amplifies and expedites the effects of personal Bio-Regenesis techniques.

CHAPTER 1

THE CONTEXT OF HEALING

The Structure of the Unified Field and the Multi-dimensional Anatomy of Form

- THE HOLOGRAPHIC TEMPLATE
- THE STRUCTURE OF DIMENSIONS, PARTIKI PHASING AND MATTER DENSITY
- UNIVERSAL AND PLANETARY KATHARA GRIDS, THE 12-TREE, KATHARA CENTERS AND LINES, SCALAR SHIELDS, CRYSTAL SEALS AND DIODIC POINTS
- THE UNIVERSAL AND INTERNAL TEMPLAR COMPLEXES
- THE TRANSDUCTION SEQUENCE AND HOLOGRAPHIC REALITY CONSTRUCTION

TOPICS OF DISCUSSION

The Ariea Khum-Nar'-A

The Song of Orion

Part 1 : the Resonant Tone Syllable Sequence (primary melody)

Verse 1: OM A Da E, Patum ah
OM A Da E, Patum ah
Patum ah A' DE E - Tah, A khum-na
OM A Da E, Patum ah

Verse 2: TrA TE La E' Durum Na
TrA TE La E' Durum Na
Durum' ah A' Khem - Tah, A khum -na
OM A Da E, Patum ah

Part 2 : the Base Tone Syllable Sequence (accompaniment 1)

OM A Da E, Patum ah - Lines 1, 2 & 4 Verse 1 & Line 4 Verse 2
TrA TE La E' Durum Na - Lines 1 & 2 Verse 2
UM ah OM, A
UM ah OM, A
Patum ah A' DE E – Tah, A khum- na
Durum ah A' Khem – Tah A khum-na
DO RAY' TU Ra"
OM A Da E, Patum ah
UM ah OM, A

Part 3: the Over Tone Syllable Sequence (accompaniment 2)

OM A Da E, Patum ah – Lines 1 & 2 Verse 1
TrA TE La E' Durum Na – Lines 1 & 2 Verse 2
BE HE Tah OM, A
BE HE Tah OM, A

Patum ah A DE E – Tah, A Khum –na – Line 3 Verse1
Durum ah A Khem- Tah, A Khum - na - Line 3 Verse 2
Thra Khem' Na, A
OM A Da E, Patum ah – Line 4 Verses 1 & 2

KhrA MiM, Tah A

THE STRUCTURE OF THE UNIFIED FIELD AND THE CONTEXT OF HEALING

If one is to become fluent in the art and science of **Bio-Spiritual Healing Facilitation,** it is necessary to become familiar with the *primary structures of energy* within which existence takes place. Without this understanding one will be unable to identify the *true nature of systems and their intrinsic functions.* If we do not understand the **intrinsic order of a system** and the elements that make up its parts, our ability to comprehend the *organic functions* of the system will be limited. Without comprehension of the intrinsic dynamics inherent to a system, our perceptions will be limited to observing the **effects of systems function,** rather than perceiving the **causal elements** behind and within the apparent observable effects. In relation to the **health and healing** of the human body, it is wise to view the human **Body-Mind-Spirit organism** as a series of interwoven systems, each with *inherent, implicit order*, the perceivable and observable aspects of which represent the effects of systems function. **The observable effects of the human condition emerge from an as yet unidentified causal order, a hidden and mysterious source out of which the manifest effects of human existence emerge.** Holistic Healing approaches human health and healing in terms of the *interdependent relationships* between *multiple interwoven systems.* The Body-Mind-Spirit organism is viewed in terms of the *intrinsic relationships that unite* the multiple systems of the human being, rather than in terms of the *observable boundaries* that appear to *separate and divide* the multiple systems inherent to the human organism.

Holistic Healing, and **KATHARA Healing** particularly, take the concept of **intrinsic unity** within the seemingly separate systems of the human being one step further than *simple interrelationship* between **physical, mental and spiritual aspects** of human experience. **In the paradigm of Kathara Healing, the human organism is viewed as a set of interwoven systems that is itself part of a larger system of interdependent systems.** The human individual is not only recognized as a *singular set of unified parts*, but rather as an *intrinsic part* of the species, planetary, galactic, universal and cosmic systems within which human existence finds **context** for being. The **Cosmic Unified Field of Energy and Consciousness** is viewed as the **indelible medium** within which all reality takes place, and human existence is viewed in terms of *its relationship to this greater Meta-system* of interwoven, interdependent systems. **The health and wellbeing of the human being is directly related to and effected by the systems-functions of the greater Unified Field**. In order to better understand and facilitate the health and healing of biological organisms, it is useful to understand the *basic order and function* of the **UNIFIED FIELD**- the context in which life itself takes place.

Kathara Healing is an ancient system of **Holistic Healing**, dating back beyond presently identified recorded human history. The principles of **Universal Order** as taught within Kathara Theory present a **broader context of universality** than is presently recognized by contemporary western science. Kathara is built upon **Laws of Multidimensional Universal Physics,** laws of intrinsic nature with which western science is not yet familiar. In Kathara Healing the *hidden order of universal systems* is recognized and the foundations of Kathara *demonstrate the interwoven systems-functions* between the human being and the order of the **Unified Field. To develop proficiency and genuine skill within the art and science of Holistic Healing, the basic order of the Unified Field must be understood.** For with this understanding comes the ability to *directly interact with and influence the causal elements* of the human experience. For this reason the teachings of Kathara Healing begin with providing a basic format through which the **CONTEXT of healing** can be understood ***– a glimpse at the intrinsic order of the Universal Unified Field.***

With foundations grounded in **coherent systems relationships**, the Kathara Healing Facilitator has more knowledge, power and comprehension to bring to the art and science of healing facilitation. **Through comprehension of the *intrinsic systems of universal order*, the Kathara Facilitator progressively evolves to embrace the *unity of SELF within the Unified Field*, opening to an ever-expanding capacity of spiritual awakening and the embodiment of UNCONDITIONAL LOVE.** The physical, mental and spiritual aspects of human anatomy represent one interdependent, interwoven system that is *continually in a state of relationship to the Unified Field.* This relationship must be acknowledged and embraced if true health and healing are to occur. **Human evolution is a Bio-Spiritual process, the body and the consciousness – the biology and the spirit - are inseparably intertwined within themselves and within the greater order of the Unified Field.**

Healing emerges through restoring order to the ONE-SYSTEM and its infinite parts.

THE HOLOGRAPHIC TEMPLATE

The externalization of life, manifestation of matter and individuation of identity are **HOLOGRAPHIC PROJECTIONS of CONSCIOUSNESS-ENERGY SUBSTANCE**, created through an *intrinsic order of energetic relationships*. Energetic relationships represent **interrelationships of consciousness**, in its infinite manifest and de-manifest forms.

Consciousness *IS ENERGY*
And
Energy *IS CONSCIOUS*.

True Science has a *spiritual core* and true spirituality has *its intrinsic translation within the Laws of Energy that govern the manifestation of universal systems*.

Reality is **THOUGHT CONSTRUCTION.** The core substance of the cosmos is **CONSCIOUSNESS.** ***Thought** is an attribute of CONSCIOUSNESS*, the **filter** through which consciousness manifests itself into the **Hologram of Form.**

Nothing is truly solid. All things are composed of *consciousness* and their *apparent solidity* is determined by the *relationship between the consciousness observing the form and the consciousness of which the form is made.* Human consciousness emerges through an **order of dimensionalization,** which sets frameworks within which the *interrelationships of consciousness* can take place. **Human perception of solidity and externalization is determined by the oscillation rates between dimensionalized units of consciousness.** Dimensionalized units of consciousness form **frequencies of energy,** which exist in *specific relationships* to each other. Human *consciousness* (the soul –spirit), and human *physical form* are **patterns of frequency** formed by *dimensionalization of units of consciousness. We perceive manifest solidity due to the relationship between the frequencies of consciousness of which we are made and those of the Unified Field of Conscious Energy around us.* **Humans perceive as solid the frequency bands that exist one full dimensional spectrum below the frequencies within which the individuated or collective consciousness is stationed.**

The only difference between a *thought* and a *manifest thing* is the *frequency of the consciousness of the observer.* Manifest reality is **thought projection,** made solid and externalized by relationships between the *frequencies of our focus of attention* and those of the *projected thought forms*. **We perceive as reality the thought forms of the collective masses as they were placed within the frequency bands that now exist one dimension below the frequencies within which our present focus of collective consciousness is stationed.**

As reality is a **THOUGHT FIELD**, composed of units of consciousness awareness, reality can be directly **AFFECTED by THOUGHT**, as the substance of both thought and reality are one and the same. The **key to mastery** of our present reality and to restoring our **imprint for health**, is understanding the *reality of dimensionalization* within which our consciousness now resides.

The human body is a HOLOGRAPHIC PROJECTION of consciousness that is built upon a *HOLOGRAPHIC TEMPLATE* of structured units of consciousness. The **Holographic Template** represents the **Living Morphogenetic Field** – the *dimensionalized blueprint of conscious light, sound and scalar waves*, within which the individuated consciousness is stationed, and *upon which the illusion of solidity of the body matter is formed*. **To affect *true and lasting change* within the Body-Mind-Spirit System of the human being, one must *go to the core* of its manifest structure – the *Holographic Template* or Thought-energy Blueprint, upon which the entire system of consciousness itself comes to know being.**

The body is a **HOLOGRAM** that can be directly affected by the *substance of which it is* ***composed*** – **CONSCIOUSNESS.** In learning the structure of dimensionalization through which the body-illusion manifests, one becomes progressively more masterful in directing this personal thought form to suit ones desired specifications. The core **Holographic Template** of the personal and universal body is called the **Kathara Grid. Kathara Healing** is the process by which the *embodied* aspects of human consciousness become more aware of and able to direct the *intrinsic processes of holographic projection*, with the intention of restoring integrity to the function of the Body-Mind-Spirit system.

No thing is truly manifest **– it only appears to be so,**
due to the refraction of consciousness
within the energetic relationships inherent to
the *Holographic Template* –
the original *thought-form construct*
upon which our universal structure
is perpetually recreated.

All creatures and things are Spiritual Beings having a Manifest Experience... No creature or thing can be ever divorced from its innate spirituality, but only from the awareness of its nature. For life is the expression and substance of spirit – and manifest reality is but a Dreamscape, from which we eventually all awaken...

The causal element of all manifest experience is called a ***Morphogenetic Field*** or **MANIFESTATION TEMPLATE**. Keylontic Science, the science of Manifestation Templates is part of a larger advanced paradigm of science that is known as **15-Dimensional UNIVERSAL UNIFIED FIELD PHYSICS.**

PRIMAL SUBSTANCE

- A Manifestation Template is made of **PRIMAL SUBSTANCE;** units of conscious energy called **PARTIKI** that exist as ***Omni-polar points of fixed vibration. Partiki units*** are the smallest building blocks of matter, and they form the templates upon which consciousness in all forms enters manifestation.

- Partiki units are **Omni-polar** (containing the potentials for all polarities or none) units of vibrating energy that perpetually cycle back and forth between a state of **BI-POLAR LIGHT RADIATION** (scalar-standing-wave) and **OMNI-POLAR SOUND VIBRATION.**

- In the ***Omni-polar state***, Partiki units exist as invisible, fixed points of ***sound vibration***, forming what is called an **INTERDIMENSIONAL TRI-TONE WAVE**. An Interdimensional ***Tri-tone Wave*** is a singular point of fixed vibration that is composed of ***3 sub-vibrations*** (base tone, overtone and resonant tone). When Partiki units are in their Omni-polar state they exist as ***Ante-matter substance***, the first state of energy organization before Pre-matter manifestation. Through a process of Internal Fission characteristic to their design, Partiki units break apart to form Bi-polar waves, while replicating their original Tri-tone Omni-polar design.

- In the ***Bi-polar state***, two sub-vibrations of the Tri-tonal Standing Waves of Partiki units break down into two sub-units, ***sub-tones***, called **PARTICUM**-*Base Tones* and **PARTIKA** *Overtones*, interconnected units of **BI-POLAR LIGHT RADIATION** that form ***Electromagnetic Scalar-Standing-Waves***. The third vibration of the Tri-tone wave, the *Resonant Tone*, remains as the core vibration through which the polarized vibrational units of Particum and Partika will reunite through ***Internal Fusion***.

- **PARTICUM** units form the energetic building blocks for **PARTICLE TEMPLATES** and **PARTIKA** units form the building blocks for simultaneously manifesting **ANTI-PARTICLE TEMPLATES**. Particle and Anti-particle Templates are interconnected, united by the Ante-matter **PARTIKI** units through which they manifest. The effects of **ELECTRICAL AND MAGNETIC FORCE** are set in motion through the **EXPANSION AND CONTRACTION** of energy vibration between Particle and Anti-particle matter units through the Ante-matter Partiki unit that unite them.

- **Particum-Particle units** represent the **CONTRACTION PHASE** of ante-matter energy, with higher **VIBRATION** (energy contracted into or received by the Ante-matter template) and lower **OSCILLATION** (energy expanded or transmitted from the Ante-matter template) and so are considered **BASE MAGNETIC**. **Partika-Anti-particle units** represent the **EXPANSION**

PHASE of Ante-matter energy, with lower Vibration and higher Oscillation, and are considered **BASE ELECTRICAL.**

- **ELECTROMAGNETIC ENERGY** is created through the continual cycling of energy, as the Omni-polar Ante-matter Tri-tone Wave of ***Sound Vibration*** projects its energy into Bi-polar Particum and Partika Scalar Waves of ***Light Radiation.*** Energy thrust expands into the Anti-particle Partika vibration point as **electrical force.** It then contracts into the particle Particum vibration point, creating **magnetic pull,** (in ratio to its original electrical thrust), as the third vibration in the Ante-matter Tri-tone Wave draws the energy back to its source. Through the draw of the Ante-matter Partiki template, Partika-Ante-particles and Particum-particles are brought into a state of **Fusion.** Upon fusion, the energy of the Particum and Partika transmutes, the Bi-polar Standing Scalar Waves of Light Radiation returning to their original Omni-polar Ante-matter state of Partiki units. Once returned to the Partiki state, the process repeats as the Partiki creates Fission and replication. **Partiki, Particum and Partika units represent PHASES that Primal Substance energy passes through as it moves into and out of manifestation.**
- The process by which Partiki units perpetually convert conscious energy into Bi-polar Light Radiation Scalar-Waves (expansion) and back into Omni-polar Sound Vibration Tri-tone Waves (contraction) is called **PARTIKI PHASING.** Through perpetual cycles of Partiki Phasing, matter units "flash on" through **fission,** into Bi-polar Particle and Anti-particle manifestation built upon Scalar Waves of Light Radiation, then "flash off" through **fusion,** into Omni-polar Ante-matter Sound Vibration, continually cycling energy between the manifest and non-manifest state. **The perception of external physical matter, space and the passage of time are created through *variance in relationship* between rates of Partiki Phasing.**
- Consciousness in the form of Partiki Units is the **PRIMAL SUBSTANCE** of which the cosmos is structured and Partiki Phasing is the perpetual **PRIMAL ACT** of reciprocal fission and fusion through which Primal Substance is brought into manifestation.

The Holographic Template

PRIMAL SUBSTANCE

Units of Consciousness, Morphogenetic Fields and Scalar Grids

1. ***Partiki, Partika and Particum:*** Smallest units of energy-substance. Electro-tonal units of consciousness that are the building blocks of morphogenetic fields, matter and individuated conscious identity. Operate as minute, perpetual motion fission/ fusion generators.

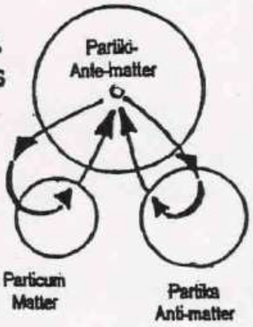

2. ***Partiki Grids*:** The fabric of morphogenetic field structure. Partiki units group to form interwoven Strands, then Grids of electro-tonal substance, out of which morphogenetic fields are fashioned. Out of a Unified Field of Partiki units, Partiki Grids form to create the frequency bands that make up dimensional fields and then to form individuated morphogenetic fields.

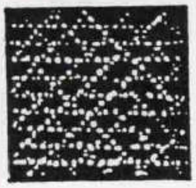

3. Partiki units continue to group or accrete upon the Partiki Grids, forming
A. ***Keylons*** – crystallizations of "frozen light" (standing scalar wave patterns and fixed points of tonal frequency), and
B. ***Keylon Codes*** – complex groupings of Keylons, which together form a crystalline template of light spectra, sound frequency and electro-magnetism that is the morphogenetic field Crystal Body– the Blueprint upon which matter and identity will manifest.

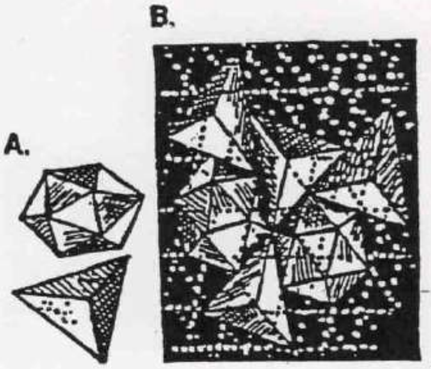

The **Morphogenetic Field** is made of interwoven **Partiki Grids.** The electro-tonal units of consciousness out of which Partiki Grids form create standing wave patterns – **scalar wave grids,** which hold the form of consciousness within dimensionalized manifestation. Morphogenetic Fields are thus **Scalar Grids,** that form in specific ordered interrelationship. Partiki Grids form upon the core *geometrical-mathematical* structure of the **Kathara Grid,** and so the Kathara Grid represents the *core Scalar Field* upon which form manifests.

Morphogenetic Fields govern the form of matter manifestation and the evolution of biology and consciousness. Morphogenetic Field mechanics thus hold the key to mastery of physical reality and Bio-Spiritual Evolution.

THE STRUCTURE OF DIMENSIONS

Frequency Bands and Sub-Frequency Bands, Vibration-Oscillation-Particle Spin, Harmonics of Manifestation and Matter Density Levels.

The Hologram of manifest reality is a complex system of interwoven energy constructs, all built upon **dimensionalized sets of scalar grids**- or fields of standing wave points – that form the **base laws of energy** upon which manifest reality functions.

The physical portions of the human body represent ***wave strata*** that exist within specific relationships to the embodied consciousness. **The embodied consciousness itself is *regulated by the intrinsic behavior of the wave strata* in which it is ensconced.** The wave strata of seemingly physical forms appear to *human detection* in the form of *particles, sub-particles and wave bands* that group to form the 3-dimensionally manifest hologram of human life

The **behavior of dimensionalized wave strata** is an intrinsic element within the function of ***human anatomy***, and provides the *very context* within which human existence and evolution take place. When we come to understand ourselves as **vibrating quantities of conscious energy**, existing within a state of **vibrational co-resonance with all things,** which are themselves *vibrating quantities of conscious energy*, we can begin to understand the **fundamental structures** upon which the human condition is built.

The human body is a ***construct of conscious energy****, a living* morphogenetic field existing within the greater living morphogenetic constructs of the species, planetary, galactic, universal and cosmic morphogenetic structure. We are simply an element of sentient consciousness within a larger structure of sentient consciousness, a part of the **Unified Field of Consciousness** that IS the living identity of the cosmos. Our physicality emerges through the **perceptual lens of dimensionalization**, which provides us with the seeming manifestation of **particles, atoms, molecules** and the **physical systems** built upon them.

If we expect to **heal ourselves** of the manifest conditions that cause suffering within the Body-Mind-Spirit system, we must first understand the **nature of health**, the "way things are supposed to work" when they are *functioning properly*. To understand the intrinsic **Imprint and impetus for health** in the human organism, a basic understanding of the *contents of its parts* is in order. That understanding begins with *identifying the basic reality structures within which we reside, and how we exist in relationship to these structures.*

Our understanding of *health* begins with *identifying the context of existence*, the *Unified Field of Consciousness-Energy* of which we are all a part. **A simple *identification of the structures of the Unified Field* will work wonders in assisting us to understand ourselves, and through this understanding we become more empowered to foster a healthful and more joyous living experience.** To better understand ourselves we must return to the core of our intrinsic nature, the **dimensionalized fields of consciousness** within which we reside and out of which the substance of our SELVES is composed.

The following characteristics inherent to the manifestation of the human hologram are a place to begin our journey of self-understanding.

Characteristics of the Hologram

PRIMAL ORDER

SCALAR WAVES: Points of standing waves, composed of quantities of conscious energy that emanate out of fixed units of vibration (Partiki units), which form the morphogenetic field templates upon and through which consciousness dimensionalizes itself to experience the hologram of manifest reality. Created through Partiki Phasing (internal fission/fusion "flash on/flash off")

MORPHOGENETIC FIELDS: The templates of consciousness, light, sound and scalar standing-wave grids which serve as the structured blueprints upon which our consciousness manifests into dimensionalization, through which the holographic projections of space-time-matter and identity individuation can be experienced.

DIMENSIONS: *Morphogenetic Field Scalar Grids composed of interwoven layers of scalar waves, which serve to direct the flow of consciousness into multiple patterns of refraction through which the hologram of matter density, linear time and objectification of reality can be experienced.* Dimensions are specific structures of consciousness, composed of Partiki units with differing vibratory-oscillation rates, which exist in precise relationship to each other, and govern the intrinsic laws of function within the observably manifest worlds. The process of dimensionalization occurs through the wholeness of consciousness *compartmentalizing itself into various ordered sections having diversity of vibratory-oscillation*, through which the wholeness of consciousness may experience itself in an endless variety of forms. Each dimension represents a full frequency band.

FREQUENCY BANDS: Synchronized, cyclic, repeated sequences of "flashing on and off" of scalar-wave points within the dimensional scalar grids, which gives the appearance of manifest movement of wave spectra within a dimensional system. **Sub-frequency Bands** are shorter segments of the longer "flashing on and off" sequence of the full Frequency Band. The frequency of dimensional levels and the particles within them is determined by the vibratory-oscillation rate of the Partiki units that make up the scalar-wave grid. Frequency is the form consciousness takes on while ensconced within the framework of dimensionalization. Frequency is the vibrating-oscillating wave spectra of which matter particles and dimensionalized consciousness are composed.

VIBRATION-OSCILLATION: Vibration is the condition of **internal movement**, or ***holding of energy within*** units of consciousness (Partiki) and the particles built upon them. Vibration is determined by the **quantity of energy** ***held within*** Partiki units. **Oscillation** is the condition of **external movement**, or ***expending of energy from*** Partiki units and the particles built upon them.

The vibratory-oscillation rate of particles represents the *rate at which the scalar standing-waves "flash off and on"* within the morphogenetic field template. Vibration is the *internal measure of energy quantity* held *within* Partiki units and exists within direct proportion to the *oscillation* rate, or *external measure of energy quantity expended*, exhibited by Partiki and the particles built upon them.

The vibratory-oscillation rate represents the ratio between energy contraction (vibration) and energy expansion (oscillation), which sets the rhythm of particle pulsation that governs the formation of wave spectra within dimensionalization. The vibratory-oscillation rate of Partiki units determines the **frequency** (cyclic pattern of scalar waves "flashing off and on") of wave spectra within the various levels of the dimensionalized morphogenetic field. **The relationship of vibrational-oscillation between wave spectra determines the matter density levels particles will exhibit within the dimensional scale.**

Vibration-Oscillation Correspondences

Higher Dimensional Fields (8-15)	Lower Dimensional Fields (1-8)
Low Vibration	**Higher Vibration**
Higher Oscillation	**Lower Oscillation**
Expanding	**Contracting**
Higher Frequency	**Lower Frequency**
Shorter Wave Length	**Longer Wave Length**
Slower Electron Spin	**Faster Electron Spin**
Less Density	**More Density**
Electric	**Magnetic**
Lower Vibration =	Higher Vibration =
Less energy holding in Partiki =	More energy holding in Partiki =
More energy expended from Partiki =	Less energy expended from Partiki =
Higher Oscillation =	Lower Oscillation =
More expansion =	More contraction =
Faster flash rate of Partiki =	Slower flash rate of Partiki =
Higher Frequency of Flash =	Lower frequency of flash =
More external movement =	More internal movement =
Faster oscillation =	Slower oscillation =
Shorter wave length =	Longer wave length =
Slower electron spin around nucleus =	Faster electron spin around nucleus =
Greater expansion of energy units =	Greater compaction of energy units =
Less density =	Greater density =
Less magnetism =	Greater magnetism =
More electric	Less electrical

The human anatomy is a 15-dimensional morphogenetic field scalar grid structure, composed of wave spectra with varying degrees of matter density characteristic to the dimensional band within which each level of the anatomy is stationed. The portions of our anatomy that appear solid are those stationed one dimension below our focus of conscious attention.

PRIMAL ORDER OF MANIFESTATION TEMPLATES

- Morphogenetic Fields, or the **MANIFESTATION TEMPLATES** that serve as the blueprint through which consciousness manifests, are composed of the Primal Substance of **PARTIKI UNITS.** Partiki units draw together or **ACCRETE** to form **PARTIKI STRINGS.** Following a set **mathematical and geometrical design** inherent to their form, Partiki Strings interweave and overlap, forming a **"Fabric of Light and Sound"** called a **PARTIKI GRID.** A Partiki Grid is a fixed pattern of interwoven Bi-polar Scalar-Waves of light radiation,built upon Omni-polar Tri-tone Waves of sound vibration,that forms the foundation for diversified Manifestation Templates. A **Partiki Grid** is a **SCALAR GRID.**

- Following the precise mathematical design set by the will and intention of the manifesting consciousness, Partiki Grids further accrete, drawing in more Partiki Units, forming groups of Partiki units that are arranged following *3 different rhythms of Partiki Phasing.* **Tri-phase groupings of Partiki Units** form patterns of Scalar-wave and Tri-tone wave interrelationships that create ***crystallizations of energy*** called **KEYLONS.** Groups of Keylons continue to accrete, forming complex Keylon arrangements called **KEYLON CODES,** through which the Manifestation Template is progressively built up from Ante-matter through various stages of matter density, beginning with **PRE-MATTER SUBSTANCE,** the first density of **Liquid Silica Hydro-plasmic** energy to emerge into externalized form.

- The Morphogenetic Fields upon which universes manifest are built upon a Manifestation Template that has a specific order of mathematical interrelationships that govern the combining of Scalar Grids with different rates of ***Partiki Phasing.*** The **Cosmic Manifestation Template** is called the **ENERGY MATRIX.** Within the Energy Matrix there are many smaller manifestation templates upon which Universes are structured. **Universal Manifestation Templates** are called **TIME MATRICES.** There are uncountable Time Matrices within the Energy Matrix, which are collectively referred to as **THE TIME MATRIX.** All Time Matrices within the Cosmic Matrix follow a specific mathematical-geometrical program through which space, time and matter can be experienced by consciousness upon entering the Time Matrix structure.

- Time Matrices are arranged upon a pattern of 15 different, but interwoven, Partiki Phasing rhythms (rhythms of expansion and contraction of energy). Each of the 15 rhythms of Partiki Phasing creates one **DIMENSION.** A **Dimension** is a set pattern of **"FLASH LINE SEQUENCES"**, or a cycle of singular Partiki Phasing rhythms that contains within it 12 smaller rhythms of Partiki Phasing. Each **Dimensional Manifestation Template** creates a set of fixed Scalar-Tri-Tone Wave Points that create a template of stationary **SCALAR FREQUENCY.** Each Dimension represents one **SCALAR FREQUENCY BAND** containing 12 smaller **SUB-FREQUENCY BANDS.** A **Scalar Frequency Band** is a cyclic pattern of specific Partiki Phasing Rhythms, or **EXPANSION AND CONTRACTION RATIOS OF ENERGY CIRCULATION,** to which consciousness conforms in order to enter the experience of manifestation.

- To create the Universal Manifestation Template upon which external space, time and matter experience can be known, Dimensions are ordered into sets of 15, forming the blueprint for a

15-Dimensional Time Matrix. Within a Time Matrix the 15-Dimensions are further arranged into sets of **3 Dimensions**, forming **5 3-Dimensional reality fields**. Each 3-Dimensional Reality Field is called a **HARMONIC UNIVERSE**, and each represents a **level of matter densification** specific to its intrinsic rates of Partiki Phasing. One Time Matrix is thus a **15-Dimensional Scalar Grid** with 5 separate yet interwoven reality fields and 5 different densities of matter manifestation.

- Each of the 5 Harmonic Universes within a Time Matrix represents a **TIME CYCLE**. A Time Cycle is a fixed, repeating pattern of Partiki Flash Line Sequences, or a repeating cycle of specific ratios of energy expansion-oscillation and contraction-vibration. The Time Cycle of a Harmonic Universe is called an **EUIAGO CYCLE**. There are **5 simultaneously manifesting Euiago Cycles in every Time Matrix**, through which consciousness passes in order to experience linear evolution through space, time and matter. Each Euiago cycle contains within it 6 smaller Flash Line Sequences forming6 smaller cycles of time called **TIME CONTINUUA**.

The **PRIMAL ORDER of Manifestation Templates** represents the core mathematical and geometrical interrelationships of energy that consciousness takes on in order to enter the holographically projected experience of external space, time and matter. All manifest forms, including the **HUMAN BODY** and **PSYCHE** , are built upon a Manifestation Template that begins with the Primal Order of Dimensionalization, through which the consciousness anchors itself into the Time Cycles of the Universal Manifestation Template Time Matrix, to begin the experience of evolution through time.

Understanding the **PRIMAL ORDER OF DIMENSIONAL STRUCTURE**, as it pertains to the human organism, allows for comprehension of the greater causal framework through which personal healing must occur. Such comprehension of the core causal structure of manifestation enables one to work cooperatively, rather than in resistance to, the natural **Universal Laws of Manifestation,** progressively restoring the **NATURAL HARMONIC ORDER** of wellbeing and **DIVINE RIGHT FORM,** through which the indelible **Organic Imprint for Health** can be progressively restored.

Dimensional Structure

A Dimension represents a full ***Frequency Band***, or repeating, cyclic sequence of "flashing on and off" scalar-wave points within a morphogenetic field. The morphogenetic field scalar grid structure of dimensions takes the form of ***sets of 15-Dimensional Matrices***. The 15 Dimensions are grouped into sets ***of 3 Dimensions each***, forming ***5 sets of 3-dimensional reality fields*** called **HARMONIC UNIVERSES**. The 15-Dimensions composing the 5 Harmonic Universes together represent one TIME MATRIX system.

Each dimension of frequency is composed of **12 Sub-frequency Bands**, or shorter cycles of the "flashing on and off" of scalar-wave points, which exist as part of the longer cycle of the full Dimensional Frequency Band. The Partiki in each dimension have a specific rate of **vibratory-oscillation** (contraction and expansion of energy), which determines the **Angular Rotation of Particle Spin** (angular axis upon which particles will spin). There is a **90-degree shift of angular rotation** of particle spin between each dimensional frequency band and a **45-degree reverse shift** between **Harmonic Universes** (sets of 3 dimensions/frequency bands.)

Particles having varying vibratory-oscillation rates and angles of spin allow multiple dimensional reality fields to co-exist within the same space while remaining perceptually invisible to each other. The *relationships between wave strata* within the dimensional frequency bands create the **holographic refraction** of light, sound and scalar waves that allows consciousness to *perceive the illusions of* **matter solidity, space, time and externalization of reality** while it is ensconced within the structures of dimensionalization.

One 15-Dimensional Time Matrix with 5 Harmonic Universes

PRIMAL ORDER of the Dimensionalized Universal Manifestation Template

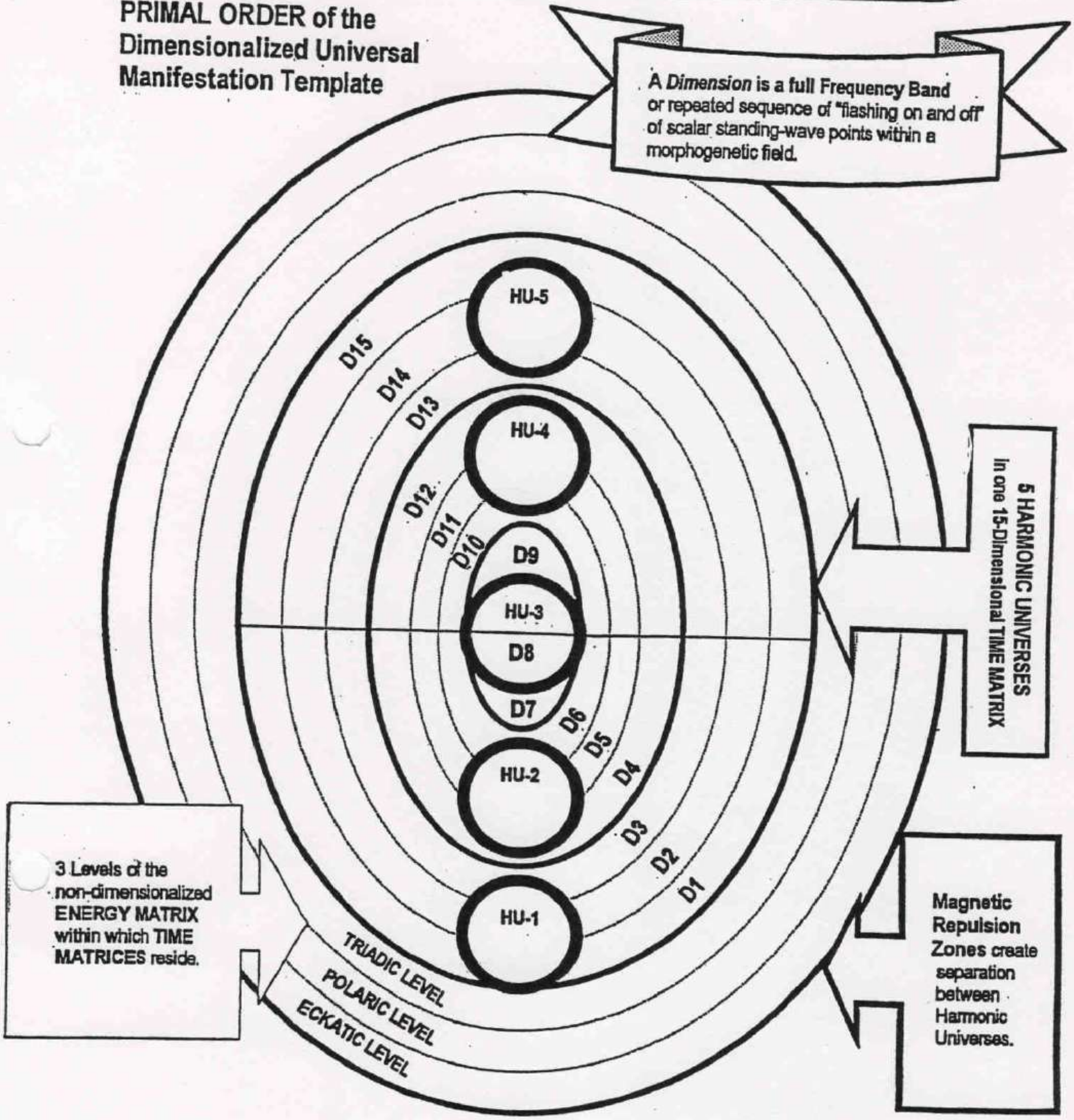
HU-5
D15
D14
D13
HU-4
D12
D11
D10
D9
HU-3
D8
D7
D6
D5
D4
HU-2
D3
D2
D1
HU-1
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

5 HARMONIC UNIVERSES
In one 15-Dimensional TIME MATRIX

3 Levels of the non-dimensionalized **ENERGY MATRIX** within which **TIME MATRICES** reside.

Magnetic Repulsion Zones create separation between Harmonic Universes.

Frequency Bands and Sub-Frequency Bands within the 15-Dimensional Scale

Frequency Bands are cyclic, repeated sequences of "*flashing on and off*" of scalar standing-wave points that create the appearance of manifest wave spectra within a dimensional system. ***Sub-frequency Bands*** are shorter cycle segments of the longer "*flashing on and off*" sequence of the *full* Frequency Band. **Frequency and Sub-frequency Bands make up the foundation wave strata of Dimensions.**

1 Harmonic Universe composed of 3 Dimensions, each Dimension composed of 12-Sub-frequency Bands, for a total of **36 Sub-frequency Bands in 1 Harmonic Universe.**
(3 D per 1HU x 12 SFB per D = 36 SFB per 1HU) 1

①

DIMENSIONS	SFB
D3	12 SFB
D2	12 SFB
D1	12 SFB
	36 SFB (12 x 3)

5 Harmonic Universes in one 15-Dimensional Matrix 2

Harmonic Universe	② Dimensions	③ 180 SFB
H 5	D15	12 SFB
	D14	12 SFB
	D13	12 SFB
H 2	D4	12 SFB
	D5	12 SFB
	D6	12 SFB
H 3	D9	12 SFB
	D8	12 SFB
	D7	12 SFB
H 4	D10	12 SFB
	D11	12 SFB
	D12	12 SFB
H 1	D3	12 SFB
	D2	12 SFB
	D1	12 SFB
		180 SFB (12 x 15)

3 15 Dimensions with 12 Sub-frequency Bands each create **180 Sub-frequency Bands in one 15-Dimensional Matrix.** (12 SFB per D x 15D = 180 SFB or 36 SFB per 1HU x 5HU = 180 SFB)

15 Frequency Bands (Dimensions) **and 180 Sub-frequency Bands within the 5 Harmonic Universes of one 15-Dimensional Time Matrix.**
SFB = Sub-frequency Band

PARTIKI PHASING, HARMONICS OF MANIFESTATION AND MATTER DENSITY

Partiki Phasing – *Scalar Standing-Wave flashing.* The internal expansion/ fission/ "flashing on" and contraction/ fusion/ "flashing off" of scalar standing-wave points made of Partiki units. *Partiki Phasing* governs the mechanics of matter manifestation within the Dimensionalized Frequency Bands of the Time Matrix. Through Partiki Phasing HARMONICS of Manifestation and varying levels of matter density are created.

Each of the 5 Harmonic Universes, moving upward in the 15-Dimensional Scale, has a progressively less dense state of perceivable matter. As you go up the Dimensional Scale each Dimension has a progressively faster rate of Partiki Phasing with its corresponding higher frequency, faster oscillation, lower vibration, shorter wave length, greater expansion and less matter density. **These conditions represent the stages of perceivable matter density consciousness will experience as the focus of attention expands progressively upward/outward into the higher Dimensions of the Dimensional Scale through the process of evolution** (accretion of frequency from the Unified Field into the personal morphogenetic field).

As consciousness progressively expands back into the 15-Dimensional structure of its original morphogenetic field, matter density perception will change from perceivable Gross Matter physical density, to Semi-etheric Matter density, to Etheric Matter density, to Pre-matter Liquid Light density and to its original Ante-matter scalar wave form of non-density consciousness. **As the morphogenetic field expands through drawing in more frequencies of consciousness, the perceivable matter density of the body and external reality field progressively lessens.**

Partiki Phasing, Harmonics of Manifestation and Matter Density

1. ***Partiki, Partika and Particum:*** Smallest units of energy-substance. Electro-tonal units of consciousness that are the building blocks of morphogenetic fields, matter and individuated conscious identity. Operate as minute, perpetual motion fission/ fusion generators.

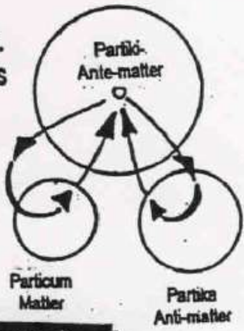

1. **Partiki Phasing**- expansion/ fission/ "flash on" and contraction/ fusion/"flash off"of scalar points governs the manifestation process

2.

Dimensions #	Harmonic Universe #	density type
1-2-3	HU-1	Carbon based biology physical matter-1
4-5-6	HU-2	Carbon-Silica based biology physical matter-2
7-8-9	HU-3	Silica based biology Etheric matter
10-11-12	HU-4	crystalline liquid-light based biology Pre-Matter
13-14-15	HU-5	standing wave pattern flame or fire body ANTE-Matter

2. **Density/ Biology:** Progression of Matter Density and related Biology within the 15-Dimensional Scale

3. **Harmonics of Manifestation:** Progressive expansion of consciousness into the Dimensional Scale creates progressive perceptual De-manifestation of matter Density within the Hologram of external reality. **Harmonics of Manifestation** are the stages of Matter Density and Vibration-Oscillation Consciousness perceives and passes through within the Dimensional Scale.

5 Harmonics of Manifestation in one 15-Dimensional Time Matrix

UNIVERSAL AND PLANETARY KATHARA

Kathara Level-1: 12-Tree, Kathara Level-2: Crystals, Kathara Level-3: Diodic Points, Kathara Centers and Kathara Lines.

The Kathara Grid is the *primary mathematical-geometrical organization of units of consciousness* upon which Partiki units group to form morphogenetic field scalar grids. It is the CORE level of scalar standing-wave creation and energetic organization within and behind all dimensionalized systems, and is thus considered to be the **Core of the Holographic Template** upon which the morphogenetic scalar wave blueprint and all other dimensions of form anatomy are built. The Kathara Grid is the causal element within all manifest affects of dimensionalization and consciousness.

The form of the Kathara Grid is reflected in the Macrocosm and Microcosm of all manifestation. All forms have at their core the *common structure of the Kathara Grid Holographic Template.* The Kathara Grids of all forms - universes, galaxies, planets and species - are linked to each other and to the Kathara Grid of the Cosmos through the interwoven Partiki scalar standing-wave grids of the *Kathara Grid Holographic Core.* **The *Cosmic Kathara Grid* represents the scalar field of the Cosmic Unified Field of Consciousness-Energy** through which the *Hologram of manifest experience* is perpetually created.

Energy cannot be created or destroyed, it only changes form. Energy IS *Eternal Consciousness* that perpetually changes form by projecting through the structures of the Kathara Grid, while simultaneously remaining *always the same* - the *singular wholeness* of the **Cosmic Unified Field of Consciousness-Energy**.

Structure of the Kathara Level-1 12-Tree Grid Core Scalar Template

12 PRIMARY KATHARA CENTERS:

Kathara Centers are condensed Scalar Wave points that hold the core electro-tonal geometrical-mathematical program for dimensionalized structure. There are 12 Kathara Centers, each holding the program for one Dimension. The program for Partiki Phasing and Vibration-Oscillation for each Dimensional Frequency Band is set in the corresponding Kathara Center. The Kathara Centers set the template for each Dimension, to ground consciousness within the Holographic experience of manifestation.

15 PRIMARY KATHARA LINES:

Kathara Lines are the primary sequences of Partiki Phasing that transmit the electro-tonal programs held within the Kathara Centers from one Dimension to another, to keep the continual flow of consciousness circulating throughout the 15-Dimensional levels of the Morphogenetic Field. When functioning properly, the Kathara Lines connect each Harmonic and Dimension of consciousness to the other and govern the synchronization of cycles of time within the Dimensional Fields and within the morphogenetic field, DNA and body rhythms of dimensionalized forms. The **3 Vertical Kathara Lines** control the operation of the other 12 Kathara Lines. The **Central Vertical Kathara Line** is the control center for Partiki Phasing rhythms throughout the Kathara Line System.

3 Vertical Kathara Lines = 1-2-3
Central Vertical Kathara Line = 1

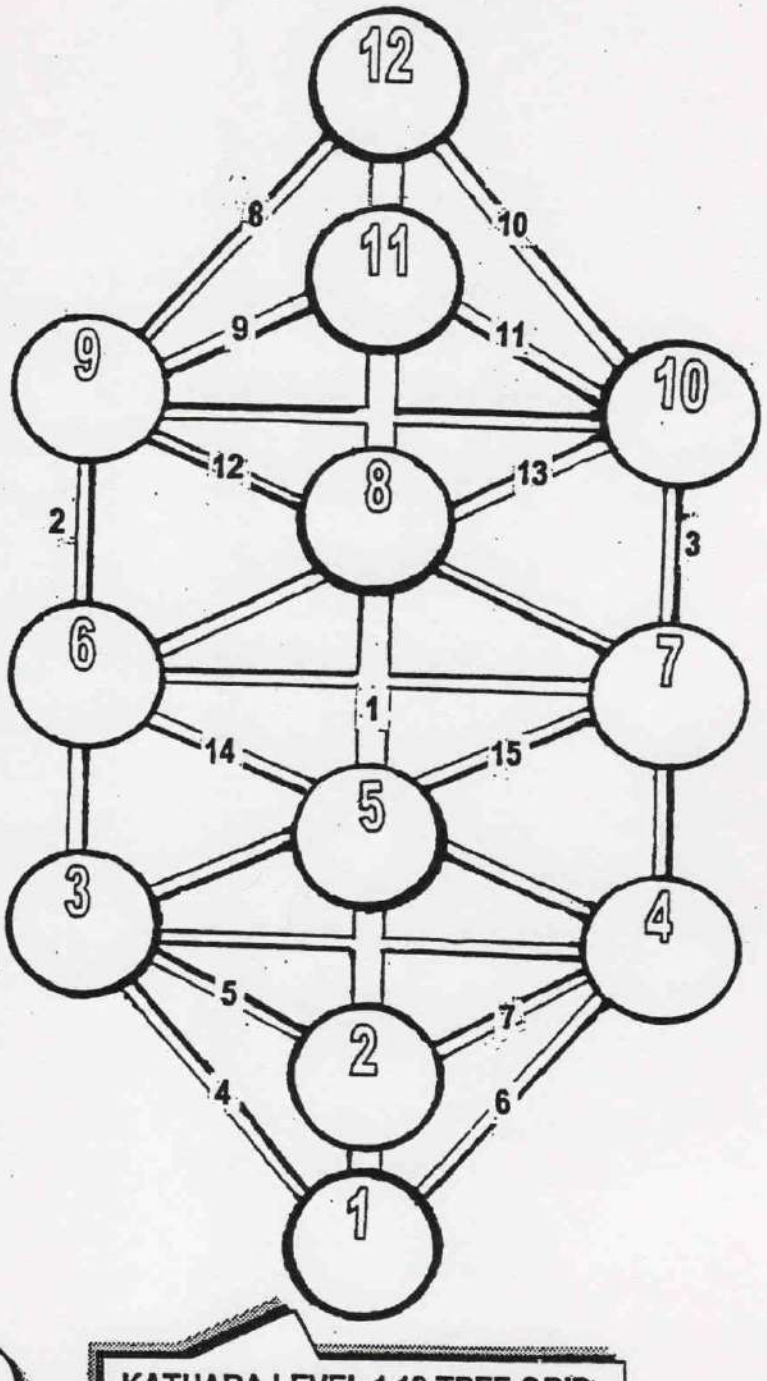

KATHARA LEVEL-1 12-TREE GRID

The 2 Interwoven Kathara Grids of Primal Order within the Universal Time Matrix and Cosmic Energy Matrix of the PRIMAL MANIFESTATION TEMPLATE

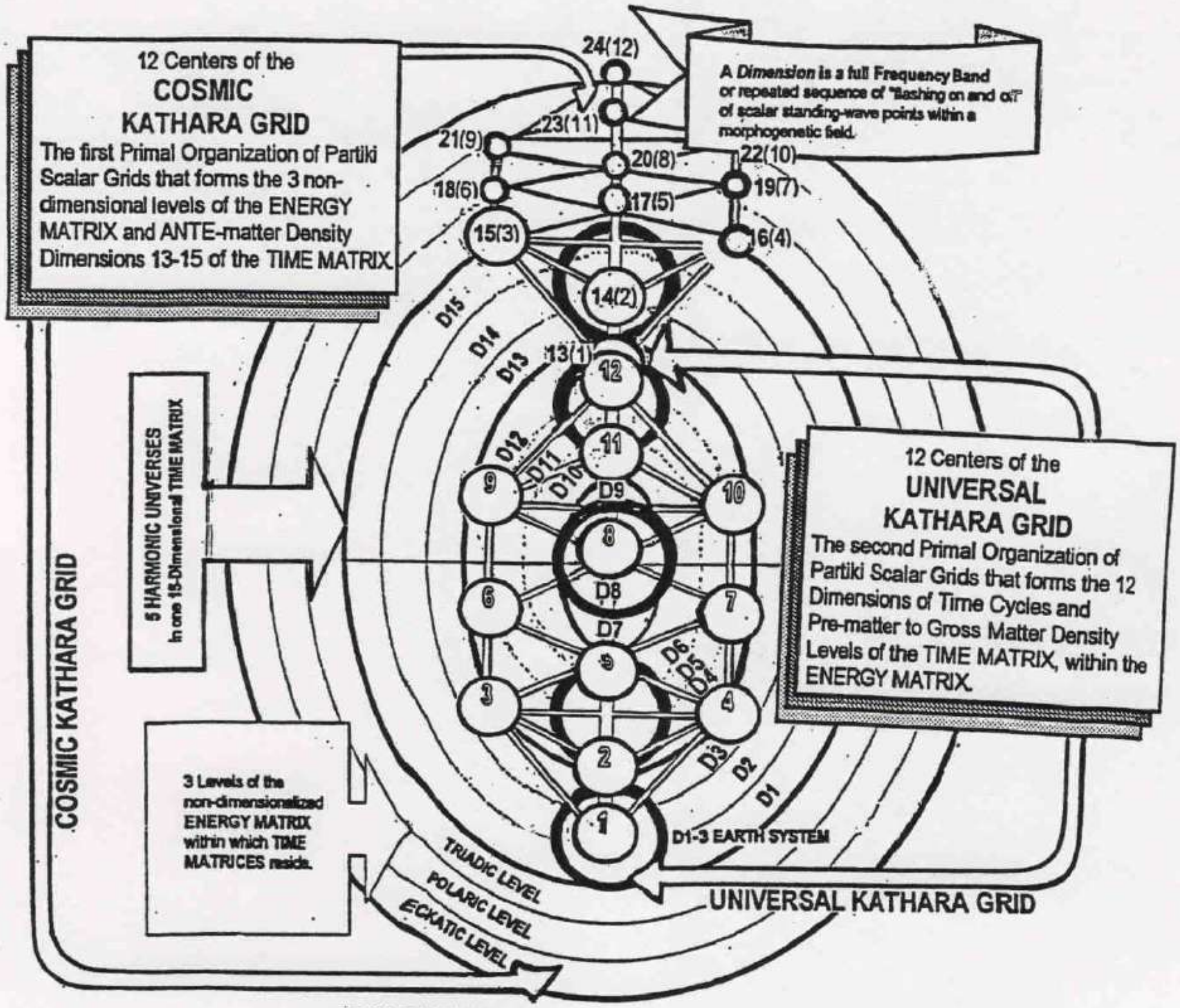

12 Centers of the
COSMIC KATHARA GRID
The first Primal Organization of Partiki Scalar Grids that forms the 3 non-dimensional levels of the ENERGY MATRIX and ANTE-matter Density Dimensions 13-15 of the TIME MATRIX.

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

12 Centers of the
UNIVERSAL KATHARA GRID
The second Primal Organization of Partiki Scalar Grids that forms the 12 Dimensions of Time Cycles and Pre-matter to Gross Matter Density Levels of the TIME MATRIX, within the ENERGY MATRIX.

5 HARMONIC UNIVERSES
In one 15-Dimensional TIME MATRIX

3 Levels of the non-dimensionalized ENERGY MATRIX within which TIME MATRICES reside.

24(12)
23(11)
21(9)
20(8)
22(10)
18(6)
17(5)
19(7)
15(3)
16(4)
14(2)
13(1)
D15
D14
D13
12
11
D12
D11
D10
D9
9
10
8
D8
D7
6
7
D6
D5
D4
5
3
4
2
D3
D2
D1
1
D1-3 EARTH SYSTEM
COSMIC KATHARA GRID
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL
UNIVERSAL KATHARA GRID

5 HARMONICS OF MATTER DENSITY IN THE TIME MATRIX

1. Gross-matter D1-D3
2. Semi-etheric D4-D6
3. Etheric D7-D9
4. Pre-matter D10-D12
5. Ante-matter D13-D15

= ○

There are 3 **Primary Levels** to the structure of the Kathara Grid. **Level-1: *The Kathara 12-Tree*, Level-2: *The Crystal Seals* and Level-3: *The Diodic Points*.**

Level-1:The 12-Tree holds the first mathematical-geometrical program upon which Kathara Grid Levels 2 and 3 are created, and is composed of the **Kathara Centers** and **Kathara Lines.**

Level-2: The Crystal Seals are groups of 3-Dimensional Partiki scalar wave composites that regulate the rate of *Partiki Phasing* to create the base structures upon which *dimensionalization* is formed. The flow of Frequency (sequences of scalar standing-wave flashing/ cycles of Partiki Phasing) between dimensional bands and Harmonic Universes, the fixed expansion and contraction rates of Partiki, the Vibration-Oscillation and the Angular Rotation of Particle spin are all regulated by the Crystal Seals.

Level-3: Diodic/Miodic Points represent **concentrated areas of energy** that emerge at points where *dimensionalized levels of scalar grids cross over and through each other*, creating **minute vortices** of wave spectra that serve as *frequency modulation zones* (scalar flash-cycle transition points) between Dimensional Frequency Bands. **Diodic Point** vortices are **White Hole** vortices that **transmit energy** into the *Particle Universe* from their counterpart **Miodic Point** vortices within the *Anti-particle Universe*. Through the structure of Dimensionalization, each Harmonic Universe, and the structures contained there within, have within their morphogenetic fields *sets of Diodic and Miodic Points* – sets of **White and Black Holes** – that regulate the flux of consciousness-energy between systems.

The entire cosmic structure is built upon the morphogenetic foundations of *interwoven Kathara Grids*. Within the structures of forms, such as planets or people, the Kathara Grid is the *Core Template of Holographic Manifestation* and controls the structure and function of all of the other *subtle and physical energy systems.*

Planetary scalar templates - **Planetary Shields -** vortices, electro-magnetics, Axiom and Ley Lines are all governed by the operations of the *Kathara Grid*. In the **human body**, personal scalar templates - **Hova Bodies -**, Chakras, Axiom and Meridian Lines and Bio-energetic Field Auric Levels, as well as the physical systems, conscious awareness and DNA, are controlled by the *personal Kathara Grid*. **In terms of generating the *Imprint for Health* within the human organism, as well as *reclaiming dormant potentials of Immortality*, the Kathara Grid is the *causal mechanism* through which such affects can occur.**

KATHARA LEVEL-2 CRYSTAL SEALS GRID

Level-2: The Crystal Seals are groups of 3-Dimensional Partiki scalar wave composites that regulate the rate of *Partiki Phasing* to create the base structures upon which *dimensionalization* is formed. The flow of Frequency (sequences of scalar standing-wave flashing/ cycles of Partiki Phasing) between Dimensional Bands and Harmonic Universes, the fixed expansion and contraction rates of Partiki, the Vibration-Oscillation and the Angular Rotation of particle spin are all regulated by the Crystal Seals.

Planetary Star Crystal Seals

In the ***Planetary Kathara Grid***, the **15 Primary Star Crystal Seals** exist *between* each planetary Vortex or Chakra, on the **Central Kathara Line** or **Central Planetary Axis**. The Star Crystal Seals serve as **"Frequency Seals"** that open or close the *flow of wave spectra* and *Frequency Bands* from one **Harmonic Universe** to another, by regulating and modulating the *sequences of Partiki Phasing and the Vibration-Oscillation rates of Partiki units* within the Dimensional Scale. They control the opening and closing of the *Central Vertical Kathara Line* and *Central Planetary Axis*, through which particles are *phase-locked* into their respective dimensional placements and linear time cycle orientations. Opening of the *Star Crystal Seals* releases the planetary particles and *Ley Lines* from their *dimensional phase-lock*, opening **Inter-harmonic Star Gates** (*wave spectra vortices*) within the Ley Lines, for ***molecular transmutation** and time continuum shift*, through blending the Inter-Harmonic wave spectra of the planetary morphogenetic field. **The Planetary Star Crystal Seals control the opening and closing of the 3 *Vertical* Kathara Lines** (and **Vertical** *Star Gate passageways*) **of the Level-1: 12-Tree Kathara Grid with in planetary body.**

Biological Star Crystal Seals

In the ***human body*** the **15 Primary Star Crystal Seals** exist *between* each Chakra on the **Central Kathara Line** or **Central Body Current**. The *biological Star Crystal Seals* serve as **"Frequency Seals"** that open or close the *flow of wave spectra and Frequency Bands* from one *Harmonic of consciousness* to another, by regulating and modulating the *sequences of Partiki Phasing and Vibration-Oscillation rates of Partiki units* within the Dimensionalized scalar wave grid of the morphogenetic field. They control the opening and closing of the *Central Vertical Kathara Line* and *Central Body Current*, through which particles are *phase-locked* into their respective dimensional placements and linear time cycle orientations. Opening of the *biological Star Crystal Seals* releases the *body particles* and **DNA** from their *dimensional phase-lock*, opening **Inter-Harmonic Star Gates** (*wave spectra vortices*) within the **DNA**, for **cellular transmutation** (via particle+anti-particle *fusion*) and time continuum shift, through blending the Inter-Harmonic wave spectra. **The Biological Star Crystal Seals control the opening and closing of the 3 *Vertical* Kathara Lines** (and **Vertical** wave spectra vortices) **of the Level-1: 12-Tree Kathara Grid within the biological form.**

Seed Crystal Seals

Along with the 15 Primary *Star Crystal Seals* that control the *Vertical* modulation of wave spectra, there are also **15 Secondary Crystal Seals** that control the **HORIZONTAL** Kathara Lines of the Level-1: 12-Tree Kathara Grid within the planetary and biological bodies. The *Horizontal Crystal Seals* are called **SEED SEALS**, they exist and serve as **"Frequency Seals"** ***within*** the planetary vortices and biological Chakras, modulating the wave spectra between Parallel (particle/ anti-particle) Dimensional Frequency Bands.

KATHARA GRID LEVEL-3 DIODIC POINTS

Level-3: Diodic/Miodic Points represent concentrated areas of energy that emerge at points where *dimensionalized levels of scalar grids cross over and through each other*, creating **minute vortices** of wave spectra that serve as *frequency modulation zones* (scalar flash-cycle transition points) between Dimensional Frequency Bands. **Diodic Point** vortices are **White Hole** vortices that **transmit** energy into the *Particle Universe* from their counterpart **Miodic Point** vortices within the *Anti-particle Universe*. Through the structure of Dimensionalization, each Harmonic Universe, and the structures contained there within, have within their morphogenetic fields *sets of Diodic and Miodic Points* – **sets of White and Black Holes** – that regulate the flux of consciousness-energy between systems.

Planetary Diodic Points

Within the **planetary body** and morphogenetic field, **Diodic Points** are located in **geographical locations** where Inter-dimensional and Inter-harmonic Horizontal and Vertical Kathara Lines *intersect* to form small **vortices** of inter-dimensional or inter-harmonic wave spectra that serve as passageways – **time portals**- between Dimensional Time Continua or Harmonic Time Cycles. They connect planetary and stellar bodies to their *anti-particle counterparts* and intersect with Vertical Inter-harmonic Star Gates and Horizontal Inter-dimensional Time Portals, through sets of Diodic-Miodic **White and Black Holes** at the center of planetary and stellar cores.

Biological Diodic Points

Within the **biological form** and morphogenetic field, Diodic Points are located in **physical body regions** where Inter-dimensional and Inter-harmonic Horizontal and Vertical Kathara Lines *intersect* to form small **vortices** of inter-dimensional or inter-harmonic wave spectra that serve as ***gateways of consciousness*** between Dimensional Time Continua or Harmonic Time Cycles. They connect physical bodies to their *anti-particle counterparts* and intersect with the Vertical and Horizontal *relationships of wave spectra within the DNA/ RNA and bio-energetic field*, through sets of **White and Black Holes** at the center of the biological Harmonic scalar grids.

Diagonal Grid Control

The *Kathara Grid Level-3: Diodic Points* control the **Diagonal Kathara Lines**, Diagonal *Inter-dimensional portals*, Diagonal *Planetary Ley Lines* and Diagonal relationships between inter-dimensional wave spectra within the biological DNA. **In *Kathara Healing*, the *Diodic Points* are used to open and close the *Diagonal Kathara Lines* for channeling higher dimensional frequency into the Level-1 Kathara Grid, to regenerate the Imprint for Health and to revitalize the Bio-energetic and physical systems.**

THE UNIVERSAL TEMPLAR COMPLEX

- The systems of interwoven Manifestation Templates within the 5 Matter Density Levels of the 15-Dimensional Universal Kathara Grid are collectively referred to as the **UNIVERSAL TEMPLAR COMPLEX.** The Manifestation Template of the human body, likewise structured upon this Primal Order, also manifests through a personal **INTERNAL TEMPLAR COMPLEX** of interwoven 15-Dimensional Manifestation Templates.

- Universal, Galactic and Planetary Scalar-wave Templates of the Universal Templar Complex are called **Scalar SHIELDS.** Shields are sets of 3-Dimensional Scalar Grids upon which the 5 Matter Density Levels of the Universal Manifestation Template manifest. A Shield is composed of the distinct **3 Primary Flash Line Sequences,** or Partiki Phasing Rhythms, characteristic to the 3 dimensions of which the Shield is composed. The Internal Templar Complex of the human body is also organized into 5 sets of 3-Dimensional **Personal Scalar Shields,** through which the matter form and 5 Matter Density Levels of the 15-Dimensional human consciousness are projected into manifest experience.

- Through the intrinsic dynamics of synchronistic Partiki Phasing, each 3-Dimensional Scalar Shield manifests a **3-DIMENSIONAL SPHERICAL ELECTROMAGNETIC DOMAIN** that creates the energetic framework within which externalized manifest experience can take place. Within the Universal Templar Complex, these Spherical Domains form 5 separate yet interwoven Reality Fields or **Harmonic Universes,** each with a different level of Matter Density, rate of Vibrational Oscillation, different Angular Rotation of Particle Spin and variance in Time Cycle orientation. Within the Personal Internal Templar Complex, the Spherical Electromagnetic Domains created by synchronistic, dimensionalized Partiki Phasing form within the 15-Dimensional Time Matrix 5 separate yet interwoven **STATIONS OF CONSCIOUSNESS** called **HOVA BODIES.** The **5 Hova Bodies** of human Manifestation Template Anatomy within the Time Matrix represent the **5 Subtle Energy Bodies** of human bio-energetic field construction, through which human identity experiences series of **SIMULTANEOUS INCARNATIONS** within the various Time Cycles and matter density levels of the Time Matrix. **Human Evolution** is the process by which the electromagnetic frequencies and stations of consciousness of the 5 Hova Bodies progressively merge to transmute the physical matter form backward through the various Matter Density stages through a process called **TRANSMUTATIONAL DIMENSIONAL ASCENSION.**

  Each of the 5 Shields within a 15-Dimensional Manifestation Template is structured upon **3 KATHARA CENTERS** within the core Primal Order of the Level-1 Kathara 12-Tree Grid. Each Kathara Center of the Universal Kathara Grid forms a core point of consolidated frequency called a **SIGNET Star Crystal Seal** within the 3-Dimensional scalar Shields. Within the universal, galactic and planetary Templar Complex the Signet Seals operate as **Star Gates** between the Time Cycles of the 5 Matter Density levels of the 15-Dimensional Time Matrix. Within the Personal Internal Templar Complex, each **Hova Body** is structured upon **3 Kathara Centers** within the embodied Personal Kathara Grid. Each embodied Kathara Center forms a **Signet Star Crystal Seal** within the human anatomy. The embodied **Signet Seals** operate **as Bio-electric Windows** within the **DNA-RNA,** that link the Stations of Consciousness and various **Simultaneous Incarnations** of the identity that exist within the Time Cycles of the Time Matrix to each other, across the illusions of time. Through the embodied Personal Kathara Grid, Scalar Shields, Signet Seals and DNA Windows, the individuated aspects of **ETERNAL IDENTITY** in time are indelibly linked in consciousness. This intrinsic linking of personal consciousness within the various simultaneous cycles of time appears to the embodied human awareness as **REINCARNATIONAL MEMORY.**

The embodied Manifestation Template of the human form is a microcosmic reflection of the Macrocosmic Universal and Cosmic Manifestation Templates. Both microcosmic and macrocosmic Manifestation Templates share, and are interwoven through, the same **Primal Structures** of the Kathara 12-Tree Scalar Grid, the 5 3-Dimensional Scalar Shields and Signet, Star and Seed Crystal Seals and the Diodic Point "Black and White Hole Sets" of energy vortices that manifest upon the Kathara Grid. Human consciousness and biology are directly and indelibly connected to the Universal Templar Complex through the **Personal Internal Templar Complex** of the Personal Manifestation Template.

HEALING AND PRIMAL FULFILLMENT

A rudimentary but operational understanding of the intrinsic structure of the embodied Personal Manifestation Template- (*the* ***Level-1****Personal Kathara 12-Tree Grid,* ***Level-2*** *Scalar Shields, Hova Bodies and Signet, Star and Seed Crystal Seals, and* ***Level-3*** *Diodic Points*) - and its relationship to the Universal Templar Complex, allows the embodied human consciousness the opportunity for progressive participation within the processes of **CONSCIOUS EVOLUTION** and intentional restoration of the **Organic Imprint for Health.** All manifest **DIS-EASE** originates through **distortions** in the electro-tonal scalar wave programs of the **Kathara Grid, Scalar Shields** and **Manifestation Template.** To achieve true and lasting healing, rather than suppression and transference of dis-ease to other embodied systems, distortions within the personal Manifestation Template need to be realigned from the causal Kathara Grid level through **REGENESIS** of **Primal Order** within the **Organic Imprint for Health.**

Manifestation Template healing therapeutics based upon knowledge of the Kathara Grid, simultaneously work to restore the natural intended **Primal Order** and function of the human body and consciousness, while expediting the evolutionary process of merging the stations of identity through time, to facilitate fulfillment of the organic **PRIMAL PURPOSE.** The **Primal Purpose** of human evolution is the achievement of **Transmutational Dimensional Ascension** and **Co-creative Mastery** over the Personal Eternal Consciousness and **Internal Templar Complex** within the Holographic Fields of the Time Matrix. Through fulfillment of the **Primal Purpose,** humanity will regain **THE PRIMAL CONDITION** of Conscious AT-ONE-ment with the Universe and the Central Source of Creation, or **GOD.**

The Universal Templar Complex

The First 4 SHIELDS each with 3 SIGNETS and correspondence to the Kathara Centers of the Level-1 12-Tree Kathara Grid.

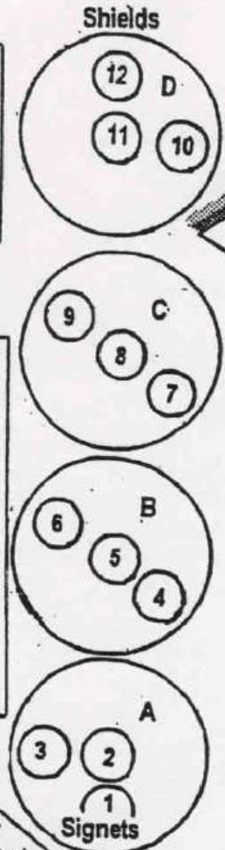

Kathara Level-2 Scalar Shields & Crystal Seals

The 4 Scalar SHIELDS corresponding to the Kathara Center programs and Harmonics of Matter Density within the *Universal* Manifestation Template.(The 5th Scalar Shield in the 15-Dimensional Time Matrix corresponds to the first 3 Kathara Centers of the *Cosmic* Kathara Grid.) Shields are the 3-Dimensional scalar-wave grids that hold the blueprint for manifestation, as set by the mathematical encoding of Partiki Phasing Rhythm held within the Kathara Centers of the Manifestation Template.

HORIZONTAL SHIELDS

- **A. TELLURIC SHIELD**
  Gross Matter Shield
  Dimensions 1-3
- **B. DORADIC SHIELD**
  Semi-etheric Matter Shield
  Dimensions 4-6
- **C. TEURIC SHIELD**
  Etheric Matter Shield
  Dimensions 7-9
- **D. MAHARIC SHIELD**
  Pre-matter Shield
  Dimensions 10-12
- **E. RISHIC SHIELD**
  Ante-matter Shield
  Dimensions 13-15
  (vertical- not shown)

The 12-Dimensional UNIVERSAL KATHARA GRID within the 15-Dimensional Time Matrix. The 12 Kathara Centers hold the core mathematical and geometrical Partiki Phasing programs upon which the Shields of the Universal Manifestation Template organize.

The 12 SIGNET Star Crystal Seals of the Universal Kathara Grid. SIGNET SEALS are the core points of consolidated frequency corresponding to each Dimension that represent the first manifest arrangement of scalar waves created through the mathematical programs of the corresponding Kathara Centers. The Signets transmit the organizational programs upon which the Scalar Shields manifest. Each 3-Dimensional Scalar Shield manifests on the electro-tonal programs of 3 Signet Star Crystal Seals. Signet Seals keep the Harmonics of Matter Density separate from each other by keeping Partiki units phase-locked into their 3-dimensional electromagnetic domain.

When released, Signet Seals serve as Star Gates, allowing the Time Cycles and Matter Density Levels within the Time Matrix to merge and blend. Signet Seals are the central control points for Scalar Shield programming.

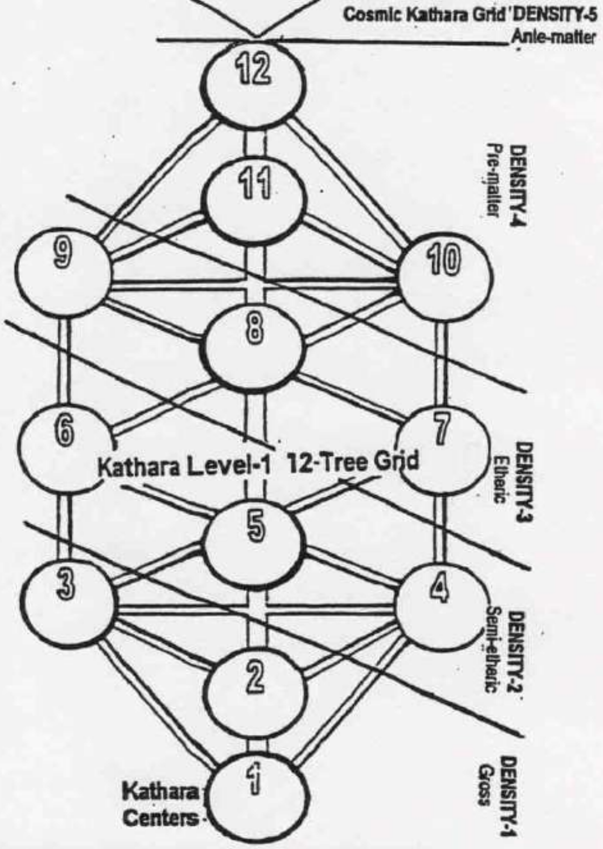

The structure of the macrocosmic **Universal Templar Complex** is replicated within the microcosmic structure of the personal **Internal Templar Complex** upon which **15-Dimensional Human Anatomy** manifests.

The Transduction Sequence

The *Transduction Sequence* represents the phases consciousness passes through into dimensionalization and the process by which conscious energy is transduced and *changes form* to enter the wave spectra within which perception of manifestation is possible.

Basic *Transduction Sequence* of Consciousness into Biological Form.

CONSCIOUSNESS TO **PARTIKI UNITS** TO

KATHARA SCALAR GRIDS LEVELS –1, 2 AND 3 TO **PARTIKI GRIDS** TO

MORPHOGENETIC FIELDS TO **MERKABA FIELDS** (Sets of Counter-rotating electro-magnetic spirals of energy –repeating scalar wave flash/ Partiki Phasing sequences that run between Dimensions and Harmonic Universes)

TO **DNA/ RNA IMPRINT** TO **BIO-ENERGETIC AURIC FIELD** TO **CHAKRAS**

TO **NADIAL CAPSULE** (tissue-like capsule that forms around the 3-Dimensional scalar grid of the Harmonic Universe-1 morphogenetic body.) TO

DNA/ RNA MANIFEST TO **NADIS** (energetic rivulets that feed energy from the Nadial Capsule into the Central Nervous System and molecular structure) TO

NUCLEAR CORE TO

CNS-METABOLIC/ BIOLOGICAL RHYTHMS- BLOOD-BRAIN-CHEMICALS-HORMONES-ORGANS-TISSUES.

HOLOGRAPHIC REALITY CONSTRUCTION

The Unified Field of Conscious-Energy

THE **SCALAR STANDING-WAVE GRIDS** THAT MAKE UP **DIMENSIONALIZED MORPHOGENETIC FIELDS** REPRESENT THE *INTRINSIC STRUCTURE* OF THE **COSMIC UNIFIED FIELD OF ENERGY** WITHIN WHICH ALL THINGS EXIST. *ENERGY IS CONSCIOUS AND CONSCIOUSNESS IS ENERGY.* **CONSCIOUSNESS-ENERGY CANNOT BE CREATED OR DESTROYED IT SIMPLY CHANGES FORM,** WHILE SIMULTANEOUSLY REMAINING ALWAYS THE SAME – **THE ETERNAL UNIFIED FIELD OF CONSCIOUSNESS.**

The Kathara Grid

ALL FORMS OF CONSCIOUSNESS AND PERCEIVABLY MANIFEST FORMS *DERIVE THEIR INDIVIDUATION* WITHIN THE UNIFIED FIELD THROUGH THE STRUCTURE OF THEIR **INDIVIDUATED MORPHOGENETIC FIELD**. THOUGH HAVING MUCH DIVERSIFICATION WITHIN THE **GEOMETRICAL-MATHEMATICAL PROGRAMS** THAT FORM **INDIVIDUALITY** WITHIN LIKE PATTERNS, THE *CORE OF ALL INDENTITY AND MANIFEST FORM IS THE SAME* – THE INTRINSIC STRUCTURE OF **THE KATHARA GRID – THE CORE HOLOGRAPHIC TEMPLATE OF DIMENSIONALIZATION UPON WHICH MORPHOGENETIC FIELDS ARE FORMED.**

Interwoven Morphogenetic Fields

THE INHERENT STRUCTURE OF **INTERWOVEN KATHARA GRIDS** FORMS **SPHERICAL DIMENSIONALIZED MORPHOGENETIC FIELDS** WITHIN WHICH CONSCIOUSNESS CAN EXPERIENCE THE **HOLOGRAPHIC PROJECTION** OF MANIFEST REALITY, SPACE, TIME, MATTER AND EXTERNALIZATION. THE SPHERICAL MORPHOGENETIC STRUCTURES OF THE **MICROCOSM** EXIST WITHIN THE GREATER SPHERICAL MORPHOGENETIC FIELDS OF THE **MACROCOSM,** FORMING A SYSTEM OF CONSCIOUS-ENERGETIC **SPHERES-WITHIN-SPHERES** THAT REPRESENTS THE ***INTERWOVEN MORPHOGENETIC FIELDS*** **OF THE COSMIC UNIFIED FIELD OF CONSCIOUSNESS-ENERGY.**

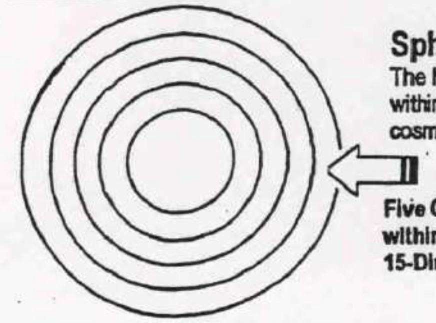

Spheres-within-Spheres ***Interwoven Morphogenetic Fields***
The Microcosm within the Macrocosm. The individual within the species within the planetary within the galactic within the universal within the cosmic within the INFINITE AND ETERNAL.

Five Over-lay Universes exist within the same space in 1 15-Dimensional Matrix.

CHAPTER 1 HIGHLIGHTS

THE HOLOGRAPHIC TEMPLATE

- **External reality is a HOLOGRAPHIC PROJECTION.** Perception of **solidity** and **externalization** are determined by the **oscillation rates** between **dimensionalized units of consciousness** - ***Partiki, Partika*** and ***Particum.*** Dimensionalized units of consciousness form **Frequency Bands** that exist in specific relationships to each other. We perceive as solid the Frequency Bands that exist **1** full Dimensional Band **below** the frequencies within which the consciousness is stationed.

- **Dimensionalization** creates the structure through which consciousness can perceive the **Hologram of Externalization.**

- The human body is a **Hologram** built upon a **Holographic Template** of units of consciousness structured into **Scalar Standing-wave Grids** that represent the **Living Morphogenetic Field** blueprint upon which matter and dimensionalized consciousness manifest.

THE STRUCTURE OF DIMENSIONS

- The **Unified Field** is built upon ***dimensionalized sets*** of ***Scalar Standing-wave Points***, **Wave Strata** formed of units of consciousness that form the basis of energetic relationship through which the **Hologram of Manifestation** can be perceived.

- The human body is a construct of ***Conscious Energy***, built upon **a15-Dimensional Anatomy of Dimensionalized Scalar-Grids.** Perceived physicality emerges through the **Perceptual Lens of Dimensionalization** which provides the picture of seemingly manifest particles, atoms, molecules and physical systems.

- Understanding the nature and structure of **Holographic Reality** allows us to understand the foundations upon which the **body** is built and the **context** within which **healing** must occur.

- **Scalar Waves** are standing wave points that emanate out of fixed points of sound (electro-tonal Partiki units), which form the **Scalar Grid morphogenetic templates** that house consciousness in **dimensionalization** for experience of manifest reality. Scalar Grids are created through **Partiki Phasing.**

- **Partiki Phasing** is the process of ***internal fission/ expansion and fusion/ contraction*** characteristic of **Partiki Units**, through which the units perpetually break apart and replicate through *fission*, then re-merge through *fusion*. Creates a “*flashing on and off*” *of **Scalar Wave points**,* or “**Flash-line Sequences**” that determine the intrinsic orientation of consciousness within the Scalar Grids composing morphogenetic fields.

- **Dimensions** are Scalar Grids of consciousness composed of interwoven layers of Scalar waves with **varying Flash-line Sequences**, or *Frequency Bands.* **Dimensions are Frequency Bands**, or Partiki wave spectra, that exist within specific relationship to each other to form the template of Holographic Reality.

- **Frequency Bands** are synchronized, repeating ***Flash-line Sequences of Partiki Phasing*** that give the appearance of manifest movement of wave spectra within a Dimensional Scalar Grid **Sub-frequency Bands** are *portions* of the *Flash-Line Sequences* that make up the whole Dimensional Frequency Band. The Frequency of a Dimensional Band is determined by the **Vibratory-Oscillation** rate of Partiki units that make up the Dimensional Scalar Grids.

- **Vibration** is the "internal movement" or **ENERGY HOLDING** of Partiki units. **Oscillation** is the "external movement" or **ENERY EXPENDED** by Partiki units. **Vibratory-Oscillation rate is the rate at which Partiki Phasing "flashing on and off" occurs.** Vibration and Oscillation exist in **direct proportion** to each other. As Vibration increases (energy holding), Oscillation decreases (energy expenditure). Faster Flash-line Sequences (higher frequency) *expend* more energy, raise Oscillation and lower Vibration of Scalar Templates. **Higher Oscillation/ Slower Vibration creates slower electron movement around nucleus and less perceivable density of matter.**

- A **Dimension** is a full Frequency band or **full cycle** of a **Flash-line Sequence.**

- The **Universal Morphogenetic Template** is built upon ***sets of 15 Primary Flash-line Sequences*** that form **15-Dimensional Time Matrices** within which manifestation can be experienced.

- **Dimensions are organized into sets of 3, (**3 full Flash-line Sequences/ 3 Frequency Bands), that create the template through which 3-dimensional manifestation can be experienced. Each set of 3 Dimensions is called a **Harmonic Universe**. There are **5 Harmonic Universes** within a **15-Dimensional Time Matrix.**

- Each Dimension contains **12 Primary Sub-frequency Bands** (partial Flash-line Sequences).

- Partiki in each Dimension have specific set **Partiki Phasing rhythms** giving each Dimension a specific ratio of **Vibration-Oscillation.** Higher Dimensions have Faster Partiki Phasing, are higher frequency and thus have higher oscillation, lower vibration and less particle density.

- **Vibratory-Oscillation** rate within a Dimension determines the **Angular Rotation of Particle Spin (ARPS)**, or *rotational axis of particles* in a Dimension. There is a **90-Degree Shift** in ARPS between each Dimension and a **45-Degree Reverse Shift** in between Harmonic Universes.

- Particles having varying **Vibratory-Oscillation** rates and **ARPS** allow *multiple 3-Dimensional Reality Fields* to *co-exist* within the same space while remaining perceptually and experientially **invisible** to each other.

- The **relationship between wave strata** and **variance in Partiki Phasing Flash-line Sequences** within Dimensional Frequency Bands create **Holographic Refraction** of consciousness-energy that allows perception of an externalized, 3-Dimensional **Holographic Illusion** for consciousness focused within the Scalar Grids of the 15-Dimensional Time Matrix.

UNIVERSAL AND PLANETARY KATHARA

- The **Kathara Grid** is the **first level** of Partiki structure/ *Scalar Grid organization* that sets consciousness into dimensionalization. It is thus the **CORE HOLOGRAPHIC TEMPLATE** upon which the *morphogenetic field blueprint* is structured.

- The *Kathara Grid* is the *Core Holographic Template* for **dimensionalization of the Unified Field** of Consciousness-Energy through which the **Microcosm** and the **Macrocosm** are structured. **All things are built upon the foundation of the *Kathara Grid.***

- There are **3 Primary Levels to the Kathara Grid. Level-1:** ***The 12-Tree Grid*** with **Kathara Centers** and **Kathara Lines. Level-2:** ***The Crystal Seals*** with **Hova Bodies** and **Scalar Shields** and **Level-3:** ***The Diodic Points*** of minute energy vortex "**White and Black Hole" sets.**

KATHARA GRID LEVELS

KATHARA 12-TREE GRID

- **Level-1 Kathara** holds the base electro-tonal **programs** of Partiki Phasing and Dimensional/ Frequency structure upon which the other Kathara Levels and the morphogenetic field are built.

KATHARA SCALAR SHIELDS AND CRYSTAL SEALS

- **Level-2 Kathara Grid** is composed of 5 groups of 3-Dimensional **Scalar Shields** and **Crystal Seals** - that regulate the rates of Partiki Phasing to create the base structures upon which ~~dimensionalization is built.~~

KATHARA DIODIC POINTS

- **Level-3 Kathara Grid** is made of sets of minute White and Black Hole energy **vortices,** made of inter-dimensional Scalar Wave spectra, that **regulate the flux of consciousness-energy** between Dimensional, Harmonic and Parallel/ anti-particle systems.

- ***Cosmic*** *morphogenetic structure* is built upon interwoven **Kathara Grids** and the Dimensionalized Scalar Standing-wave Templates they form.

- The **5 Primary Scalar Templates of Dimensionalization** are formed on **Scalar Shields** composed of sets of 3 Dimensions, which form *5 3-dimensional spheres of energy* called **HOVA BODIES**. Hova Bodies and their Scalar Shields hold consciousness-energy into **5 separate but interwoven 3-dimensional systems.**

- The **Kathara Grid** and the multiple levels of morphogenetic structure built upon it is the **CAUSAL ELEMENT** behind and within all manifestation, and thus the **CORE HOLOGRAPHIC TEMPLATE** through which the **Imprint for Health** can be restored to Biology.

- All manifest **Dis-ease** originates through **distortions** in the natural electro-tonal programs/ *Partiki unit arrangements* of the *Kathara Grid* and *Hova Body Scalar Shields*. The Kathara Grid and morphogenetic field are the **core elements** through which **lasting healing** is created. **Healing** occurs through **REGENESIS** of natural electro-tonal **programs** of the **Holographic Template**.

- **Partiki Units, Scalar Standing Waves, and Crystalline Keylons and Keylon Codes** represent the **PRIMAL SUBSTANCE** in energy that consciousness takes on in order to create the **Holographic Projection** of external manifestation.

- The **Kathara Grid, 15-Dimensional Structure, Scalar Shields, Crystal Seals, 5 3-Dimensional Harmonic Universes and Hova Bodies, Diodic Point** vortices and the subtle and dense matter energetic systems that manifest upon them represent the **PRIMAL ORDER** of mathematical and geometrical interrelationship that **Primal Substance** utilizes to orchestrate the manifest Hologram through creation of **Manifestation Templates.**

- The interwoven, ordered structures of scalar wave fields that form the **Universal Manifestation Template** of the 15-Dimensional Time Matrix are collectively referred to as the **UNIVERSAL TEMPLAR COMPLEX.**

- The interwoven, ordered structures of scalar wave fields that form the **Personal Manifestation Template** of **15-Dimensional Human Anatomy** are collectively referred to as the **Personal INTERNAL TEMPLAR COMPLEX.** The Internal Templar Complex, a structural replica of the Macrocosmic Universal Templar Complex, anchors personal consciousness into the Time Matrix for the experience of the space, time and matter Hologram.

- The **PRIMAL PURPOSE** of human evolution is the achievement of **Transmutational Dimensional Ascension** and **Co-creative Mastery** over the **Personal Eternal Consciousness** and **Internal Templar Complex** within the holographic fields of the Time Matrix. **REGENESIS** of **Primal Order** within the **Organic Imprint for Health** creates true healing and expedites the natural processes of human evolution.

- Through fulfillment of the **Primal Purpose** humanity will regain **THE PRIMAL CONDITION** of **Eternal Conscious At-ONE-ment** with the Central Source of Creation or **GOD.**

CHAPTER 2

(Section One: *Foundations of Kathara Healing* – Part II)

THE ELEMENTS OF HEALING

The Human Body Kathara and the Multi-dimensional Anatomy of the Human Form

- THE MAHARIC SHIELD AND THE IMPRINT FOR HEALTH
- THE PLANETARY BIO-FEED INTERFACE SYSTEM (PBIS)
- THE HIEROPHANT SYMBOL CODE, PBIS AND THE MAHARIC SHIELD
- MORPHOGENETIC EXPANSION, GROWTH CYCLES, DEATH AND MOLECULAR COMPACTION
- LEVELS OF HUMAN ANATOMY, EMBODIED KATHARA 12-TREE GRID, CRYSTAL SEALS, SHIELDS, SIGNETS, HOVA BODIES, AURIC CAPSULES, DIODIC POINTS, SEED SEALS, CHAKRAS, AXIOM LINES, THE ḤARA LEVEL AND DNA.
- THE 12 HUMAN SENSES, APPARTHI AND HUMAN PERCEPTION

TOPICS OF DISCUSSION

THE HUMAN BODY KATHARA AND THE MULTIDIMENSIONAL ANATOMY OF HUMAN FORM

The Structures of Manifestation and Reclaiming Our Power

In *PART-I* of this program, **THE CONTEXT OF HEALING**, we explored the hidden foundations of *reality structure* within which our existence takes place, the *context of relativity* through and within which *healing* occurs. We learned that the context of existence can be viewed as taking place within a **Unified Field of Energy**, and that **Energy IS consciousness and Consciousness IS energy.** In our exploration we encountered the **Partiki**, **Partika** and **Particum**, the ultra-micro-particle units of consciousness in electro-tonal form, through which the tangible structure of living **Morphogenetic Fields** are formed. In learning the foundation structures of the Unified Field of Consciousness-Energy, we come to understand that the anatomy of any form, including that of the human Body-Mind-Spirit System, is a **multi-dimensional construct** built on an ordered template of **sets of 15-Dimensions**, through which all conditions of apparent manifestation are created.

Through reviewing the underlying structures of manifest reality (units of consciousness, the ***Partiki Scalar Standing-wave Grids*** that hold consciousness into individuated dimensionalized "packages" of awareness to allow for experience of space, time, matter and form, the core ***Kathara Grid Holographic Template*** that sets the pattern of dimensionalization for consciousness to follow into the manifest experience, the ***Levels of Matter Density*** corresponding to the ordered structure of Dimensions, and the reality that **Dimensions** are *tangible constructs of energy* created through the process of ***Partiki Phasing***-the fission-fusion "flashing on and off" of scalar standing-waves composed of units of consciousness) **we have already expanded our perceptions of the nature of reality itself.** With these **simple revelations of knowledge**, we have shifted ourselves from a perspective of viewing the reality before us as one of *separate manifestations disconnected from our being*, to a perspective of viewing reality as an *intricate system of interwoven consciousness, expressing in the form of standing-wave strata*, to which we are *intimately and indelibly connected*.

Suddenly we find that we cannot be "left out" of things, we cannot "be on the outside as life passes us by", **that there IS no "outside" to which we may stand separate, as by the *nature of our existence* we are an *intrinsic element and force of creative affect* within the whole of creation.** Through this knowledge we have begun to tap the **secrets of SELF**, as we shift from viewing ourselves as *mortal creatures trapped within a gross-matter finite physical body*, to **KNOWING OURSELVES AS THE EXPRESSION OF CONSCIOUS, SENTIENT WAVE-STRATA - ETERNAL BEINGS ENSCONCED WITHIN AN ETERNAL UNIVERSE, within which Energy-Consciousness cannot be created or destroyed, it simply changes form perpetually.** We move from perceiving ourselves as finite beings of limited power to affect a world that is separate from us, to knowing that we are *walking wave-forms within a sea of interrelated wave forms*, and that the pattern of our being operates as an intrinsic part of the whole. **As we begin to perceive ourselves in terms of the *unlimited reality of our nature*, rather than define ourselves by the limited boundaries of the *manifest illusion before* us, we begin to reclaim our power to *affect desired change* within the contours of our lives.**

Knowledge, KNOWING and the Imprint for Health

Knowledge is an agent of freedom, IF it is used to foster practical applications of change. Knowledge can also become a *perpetual quest of distraction*, through which we deny ourselves the ability to EMBRACE the experience of living as we seek outside of ourselves the answers that live within. **Knowledge is power if it is applied appropriately, it can be an agent that leads one through the door to the WISDOM that awaits on the other side.** Knowledge alone cannot provide **cognition,** cognition comes through the *application of applied knowledge* to provide **tangible, experiential KNOWING,** a knowing derived only through the **direct experience of BEING. In Kathara Healing™ *knowledge* is applied as a *tool* through which one can grow into the *wisdom of direct experience*** - from ***thinking or believing***, which are knowledge/data-based characteristics of a mind perceiving itself as *finite* - to **KNOWING,** through Direct Cognition of the Experience of BEING, which is a characteristic of a consciousness KNOWING itself as ETERNAL. .

In our first lessons of **Kathara Healing,** we gained *knowledge* of the *Context of Healing* and the *structure* of that which exists within and around us, so that we may become more empowered to serve as a *healing force*, within ourselves and within the collective of the Whole. **But this knowledge will only become practical or useful to us if it is applied to foster the direct experience through which the wisdom of KNOWING may emerge.** Then, and only then, will we be changed...then and only then will we become *true agents of healing*. **Knowledge can help you to *understand the nature of health and healing*, but only through KNOWING HEALTH through *direct experience* will you be *healed*.** To KNOW HEALTH we must rediscover this quality within ourselves, to EXPERIENCE the *quality of wellness* that emerges as the **Imprint for Health** is reawakened within the human body. **First we must use *knowledge* to recognize that the *Imprint for Health* is there, sleeping silently within us awaiting this awakening.**

Healing the Causal Core

Fixing symptoms does not restore the *Imprint for Health*, it simply masks from view the *causal elements* through which non-health manifested. Removing the symptom *without addressing the underlying cause* simply sets the stage for the cause to *re-manifest itself* in new ways, through *new sets of symptoms*. **Each symptom you remove will give birth to another symptom if the *underlying cause* is not changed.** Kathara Healing addresses the **deepest *causal factor*** in manifestation, the **Holographic Template Kathara Grid,** upon which all conditions come into being. **Within the intrinsic structure of the Kathara Grid the dormant *Imprint for Health* is sleeping, and through knowledge of the Kathara and applications of its methods we can reawaken the Imprint for Health within ourselves, to progressively move into KNOWING HEALTH** through the cognition of Direct Experience.

In Kathara Healing you are provided with the knowledge you need to understand the structures upon which the *Imprint for Health* is built, for in understanding these structures you will be enabled to call the Imprint for Health into awakening. **But it is through the APPLICATION OF KATHARA TECHNIQUE that you will AFFECT CHANGE. The technique applications offered in Kathara Healing allow you to "bring the knowledge home" to the core of its usefulness, the ability to lead you into the *direct experience of health-ward change*.**

Elements of Healing

In **PART II THE ELEMENTS OF HEALING**, we will apply the knowledge of the *Context of Healing* that we acquired in *PART I*, directly to its relevance in terms of the **HUMAN BODY**. We will now explore **Kathara Mechanics** as they apply directly to the human system, and you will gain the knowledge of your **15-Dimensional Anatomy** through which you may then effectively utilize **Kathara Mechanics Techniques** to begin your journey into the **EXPERIENTIAL REALITY OF HEALING.** We will discover the long forgotten potentials of the ***Planetary Bio-feed Interface System™*** and how to receive energy for healing directly from the Kathara Grid of the Earth. We will explore the real processes within and behind the experience of growth through time, so that we may finally identify the **true causal element intrinsic to the *degeneration* and *death* of the body.** We will come to understand the functions of the Body-Mind-Spirit System as they are **INTENDED to operate**, to identify the areas in which malfunction is occurring, so that we may *remedy the casual factors* to progressively restore the **INTENDED INTEGRITY OF THE BODY-MIND-SPIRIT SYSTEM. In PART II we will reclaim our knowledge of our *multidimensional anatomy*, and rediscover the buried secrets of awakening the *Imprint for Health***.

MAHARIC SHIELD

The Imprint for Health and the *Maharic Shield*

Restoration of health within the body is a progressive endeavor that takes time. Portions of the Body-Mind-Spirit System that have been malfunctioning must readjust to a new process of function that will progressively work its way up from the core **Kathara Grid Holographic Template**, into the **Bio-energetic Field**, **Chakra System** and **DNA**, to progressively ***reset*** the **Imprint for Health** within the Body-Mind-Spirit System. **The Imprint for Health sleeps soundly now within a portion of the multidimensional anatomy called the *MAHARIC SHIELD*.** The Maharic Shield is a buried treasure of **REGENERATION and REVITALIZATION** of the human organism, as it is the key to *recoding the human Kathara Grid* back into its original Imprint for Health. **The *Imprint for Health* is a tangible construct of energy that is located within the structure of the *Maharic Shield*.** In calling these dormant energy systems into activation within the body, one has the opportunity to restructure the causal level of malfunction, and to recode the Holographic Template upon which malfunction manifests.

Reciprocal Reflection and the Kathara Grid

All conditions of dis-ease within the Body-Mind-Spirit System *share a common core of causality - distortions of the Kathara Grid* upon which manifestation is built. All symptoms affect the Body, Mind and Spirit *simultaneously*, malfunction in one aspect of being will have its *reciprocal reflection* in the other aspects of being, as the symptoms emanate from distortion in the *Kathara Core*, upon which all aspects of being simultaneously manifest.

As we work with restoring integrity to the *Holographic Template Kathara Grid*, we are thus creating *simultaneous restoration through Reciprocal Reflection* within the *Body*, the *Mind* and the *Spirit* (the CONSCIOUSNESS that makes up the substance of our multidimensional anatomy and identity).

The Maharic Shield and Whole-Being-Healing

In activating the *Maharic Shield* to regenerate the *Imprint for Health*, you will have change occurring within multiple levels of being. Your **BODY** will begin to reflect change as dis-ease-causing distortions are released from the *biological imprint*, and a higher level of order is reestablished within the ***atomic structure*.** Your **MIND** will begin to reflect change as dis-ease-causing distortions in your **thoughts, ideas and beliefs** are released from the *mental body* awareness. Your **SPIRIT**-consciousness, will reflect change, as divisions between your embodied awareness and the expanded portions of your multidimensional awareness will begin to dissolve, and you will become progressively more able to embody greater states of *cognition, creativity and awareness.* ***Kathara Grid recoding* through the use of the *Maharic Shield* will set in motion the processes of WHOLE BEING HEALING, as the *Imprint for Health* is progressively reawakened and restored within the Body-Mind-Spirit System**. Activation of the **Maharic Shield** within the personal Kathara Grid will lay the foundations for progressive, lasting and powerful restoration of the **Imprint for Health** within your Body-Mind-Spirit System.

Learning to activate the Maharic Shield to restore the Imprint for Health requires a bit of **knowledge** regarding *Elements* of the hidden anatomy of multidimensional being, and application of specific **energetic mechanics** that will directly affect these hidden structures of energy. We will begin PART II with a discussion on the **ELEMENTS OF HEALING as they apply to *developing the skills required to create the condition of Maharic Shield activation*, through which the *Imprint for Health* can receive its wake-up call.**

THE PLANETARY BIO-FEED INTERFACE SYSTEM™

Awakening the Personal and Planetary Templar Complex Connection

Interwoven Personal and Planetary Kathara Grids

The core Holographic Templates of both the *personal* and the *planetary* morphogenetic fields are built upon a Scalar standing-wave *Kathara Grid*. As the morphogenetic fields of **species** are interwoven into the larger morphogenetic field of the **planet** they inhabit, there is a **direct energetic connection between the Kathara Grids of individuals and that of the planetary body. This connection between personal and planetary Kathara Grids is called the *Planetary Bio-feed Interface System™*, - PBIS.** For thousands of years the **planetary Kathara Grid** has carried **distortions** in its natural pattern that have rendered the **PBIS** primarily inoperable. **Kathara distortions in the planetary morphogenetic field blocked the human's innate ability to draw energy for regeneration and revitalization from the planetary morphogenetic field.**

When the human Kathara Grid system is working properly there is a *natural sympathetic rapport* between the personal and planetary Kathara Grids that keeps the *Imprint for Health* within the human *Maharic Shield* functioning properly. Due to the distortions of the planetary Kathara, the human Kathara developed **reciprocal distortions**, which created **blockages between the Maharic Shield Imprint for Health and the other functional portions of the human Kathara.** The blockages with the human Kathara manifested as distortions within the natural function of the **DNA** and bodily systems, disarming many of the natural *self-regenerative functions* organic to human biology and rendering the human form more susceptible to disease and malfunction of the biological system. As there is a **Reciprocal Reflection** of patterns between the Body, Mind and Spirit aspects of the human organism, due to their *common causal element* of the *Kathara Grid Holographic Template*, **distortions within the DNA and physical body function also become reciprocal distortions** within the **mental body** and its **perceptual facilities,** and within the **Spiritual Body,** as the Kathara distortions *block and distort* the organic cycles through which higher dimensional aspects of consciousness would progressively embody within and expand the human form.

The Planetary Bio-Feed Interface System™

As the morphogenetic fields of species are interwoven into the larger morphogenetic field of the planet they inhabit, there is a direct energetic connection between the Kathara Grids of individuals and that of the planetary body. This connection between personal and planetary Kathara Grids is called the *Planetary Bio-feed Interface System™*, - PBIS.

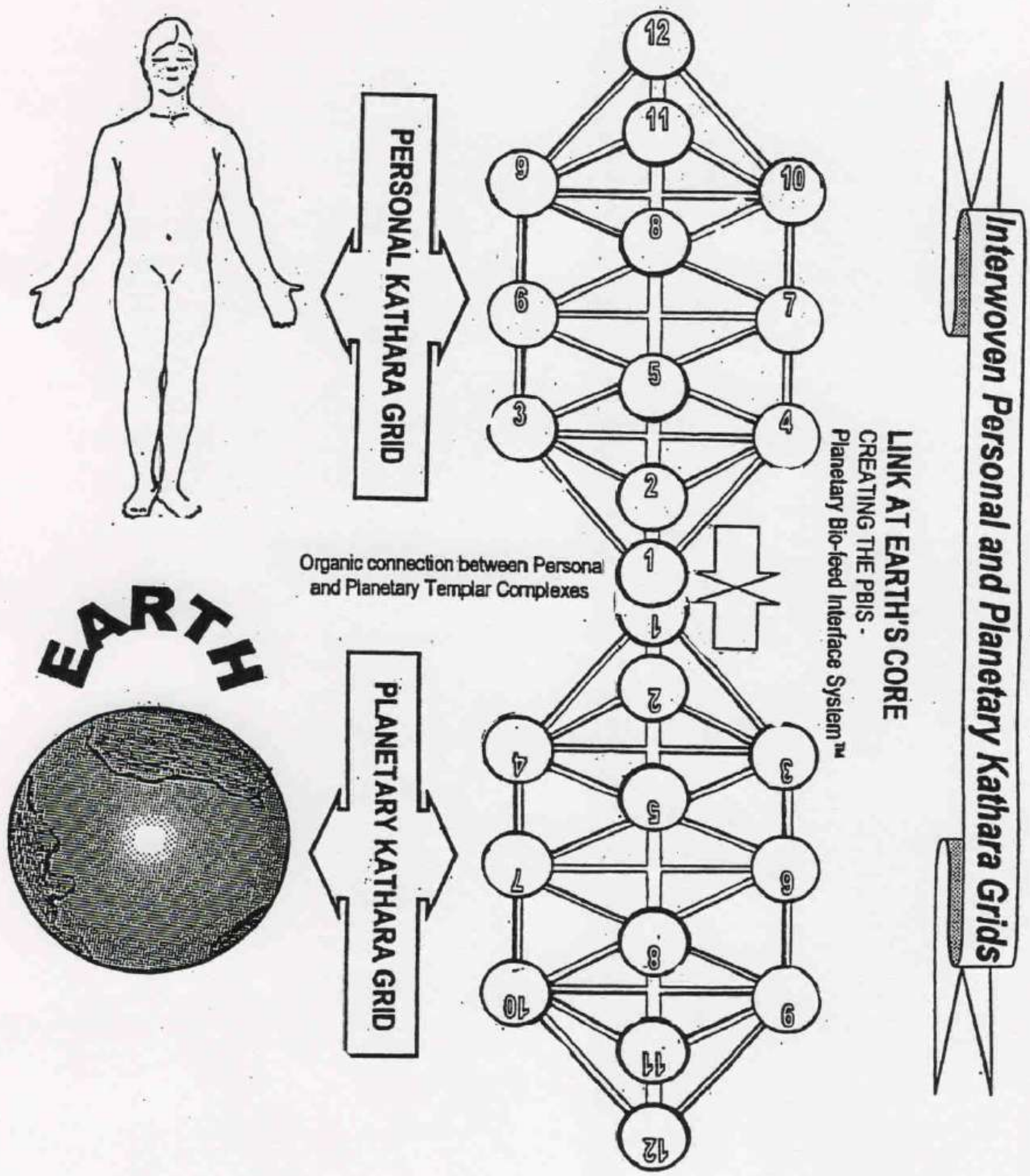
PERSONAL KATHARA GRID
12
11
9
10
8
6
7
5
3
4
2
1
Organic connection between Personal and Planetary Templar Complexes
LINK AT EARTH'S CORE
CREATING THE PBIS -
Planetary Bio-feed Interface System™
Interwoven Personal and Planetary Kathara Grids
EARTH
PLANETARY KATHARA GRID
1
2
4
3
5
7
6
8
10
9
11
12

Return of the PBIS - Gift of the 21st Century

The secrets of utilizing the PBIS have been passed down within the generations of the Melchizedek human families who brought these teachings to Earth long ago, kept hidden from public display until a day when the PBIS could once again become operational through evolution and clearing of the planetary Kathara Grid. These teachings are now being returned into the public arena because **progress has been made in restoring the integrity of the planetary Kathara Grid system. For the *first time in many thousands of years* the potentials of healing held within the PBIS can be returned to humanity, and the PBIS can now be *activated within the personal biology* to reawaken the *Maharic Shield* and its *Imprint for Health*, within the human Kathara System.** Reawakening the *PBIS* and the *Maharic Shield* within the human body *allows the personal Kathara Grid to open into the planetary Kathara Grid*, through which the now-functional **Holographic Template of the planetary Kathara can be used to RESET the original imprint of function within the personal Kathara,** for accelerated regeneration of the **Imprint for Health** within the human form.

HIEROPHANT SYMBOL CODE

The *Hierophant Symbol Code* PBIS Trigger

The first application of the PBIS that we will employ in Kathara Healing is utilizing the energies flowing through planetary Kathara to reawaken the Maharic Shield within human anatomy. The process of opening the PBIS within the personal morphogenetic field involves using a **specific scalar standing-wave-guide GEOMANCY - a consolidated inter-dimensional *frequency pattern* that triggers release of the *frequency seals* (Kathara Level-2 Crystal Seals) on the presently dormant 10th, 11th and 12th Chakras, opening the human Bio-energetic Field and Kathara Grid to the Kathara Grid, Bio-energetic Field, Ley Lines and energy Vortices of the planetary morphogenetic field.** The wave-guide Geomancy containing the frequencies to activate the PBIS and the Maharic Shield can be directed by the human consciousness **in the form of a visually created symbol code. This symbol code is called the HIEROPHANT.** In **TECHNIQUE # 2: *The Maharic Seal and the Liquid Light Cleanse***, to be explored in ***PART III. of Section I,*** the **Hierophant Symbol Code** will be used to bring the **Maharic Shield** and the **PBIS** out of dormancy, in order to **RESET** the **Imprint for Health** within the operational Holographic Template of the human **Kathara Grid.**

The *Hierophant Symbol Code Scalar Wave Guide* for Kathara Reprogramming

Thoughts and **Images** formed by the mind represent **small morphogenetic fields** with minute **Kathara Grids**, that become interwoven into the larger morphogenetic field and Kathara Grid of the human multidimensional anatomy. ***THOUGHTS ARE THINGS* in terms of the bio-energetic reality of the body and the Kathara Grid; *thoughts become translations* of electro-tonal "*digital" instructions* within the human bio-energetic field, and directly enhance or detract from the natural function of the electro-tonal programs of the human Kathara Grid.** In recognizing the **power of thought *as the power of creation of scalar standing-wave grids***, we can begin to apply that power effectively in regeneration of the Kathara Holographic Template, by **directing needed *electro-tonal instructions* for awakening the *Maharic Shield* into the embodied Kathara Grid.** The **Hierophant Symbol Code,** used in the form of a **visualized image,** sets a **powerful electro-tonal program** within the human Kathara, a program that **triggers the opening between the human Kathara and the PBIS**

The *Hierophant Symbol Code* appears in the form of a *Pale Silver Merkaba Star*, a six-pointed "Star of David", which represents the *electro-tonal program* of the *11th and 12th-Dimensional aspects of the Universal Kathara Grid.* The **frequency wave spectra** perceived by the human mind as **Pale Silver** represents the **electro-tonal program** of the ***Dark Silver** primary frequency band of the 11th-Dimensional wave spectra, combined with the **Pure White** primary wave spectra of the 12th-Dimensional frequency bands.* Using the **Hierophant Symbol** wave guide to direct *frequency and electro-tonal instructions* through the Kathara Grid, creates a ***scalar standing-wave pattern of 11th and 12th Dimensional frequency* within the personal Kathara Grid,** the frequencies in which the **original *electro-tonal instructions*** for the ***Imprint for Health*** are stored. By **awakening the 11th and 12th Dimensional frequencies within the human Kathara and bio-energetic field**, all of the *lower frequencies* within the Kathara electro-tonal program become **modulated and recoded** into their *organic, undistorted pattern*, by the *stronger wave influence* of the higher 11th and 12th Dimensional frequencies, which *contain within themselves* the organic imprint for the *entire dimensionalized wave spectra* of the dimensions that *oscillate below* them.

The Hierophant Symbol Code
Scalar Wave Guide for Kathara Reprogramming

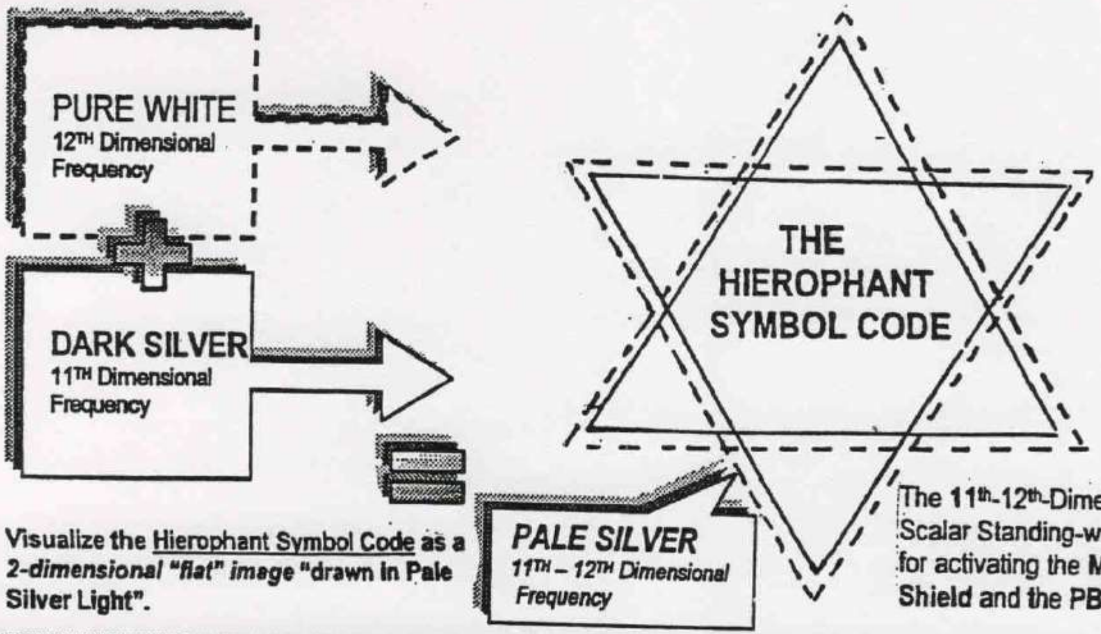
PURE WHITE
12TH Dimensional Frequency
DARK SILVER
11TH Dimensional Frequency
PALE SILVER
11TH – 12TH Dimensional Frequency
THE HIEROPHANT SYMBOL CODE

Visualize the Hierophant Symbol Code as a *2-dimensional "flat" image* "drawn in Pale Silver Light".

The **11th-12th**-Dimensional Scalar Standing-wave Guide for activating the Maharic **Shield and the PBIS.**

Ease of Kathara Healing Techniques

Through using the *Hierophant Symbol Code* we awaken the restorative frequency patterns within the *Maharic Shield* and the *PBIS*, beginning the process of regenerating the organic *Imprint for Health in* the human Holographic Template Kathara Grid. The restorative affects of using the *Hierophant Symbol Code* to activate the *Maharic Shield* and the *PBIS* will occur upon using the *Kathara Mechanics* provided in **Technique # 2**, *whether or not* an individual understands the realities in frequency that are taking place through using this process. **Therefore, the complex mechanics of activating the *Maharic Shield* to reset the *Imprint for Health*, can be *reduced to simple instructions* for the direction of energy.** Though the **theory of Kathara Healing** is highly *detailed and complex*, the **Techniques** by which *active results are achieved* employ **simple exercises in *Mental Symbol Code Visualization* and mental energy direction**, coupled with the **application of simple *Kathara-stimulating massage*** to specific corresponding body points. **The *Techniques* of Kathara Healing and regeneration of Kathara Grid integrity are so *simple* that even a young child can learn to actively employ them, to achieve the desired result of restoring the *Imprint for Health*.**

MORPHOGENETIC EXPANSION

Frequency Accretion, Growth, Age Rhythms, Degeneration, Death and Kathara Blockage

Frequency Accretion and Growth Through Time

The morphogenetic field of the human body evolves through time via the process of **Frequency Accretion**. The human morphogenetic field *exists within* the larger morphogenetic field of the planetary and universal dimensionalized **Unified Field of Consciousness-Energy.** Through the inherent function of the **Kathara Grid** and the multi-layered systems of energy built upon it, *progressively more frequency and corresponding wave spectra* from the *dimensionalized Unified Field* are **drawn into** the *personal morphogenetic field* to create **progressive expansion of the morphogenetic imprint and its resulting progressive *expansion of consciousness* and *growth of the physical body* in time.**

Kathara blockages and Organic Growth Cycles

Blockages and scalar wave distortions within the **Kathara Grid** *impede* the natural functions of **Frequency Accretion**, and so **distort the intended natural processes of consciousness expansion and growth of the body form in time.** When the Kathara Grid is functioning properly, there are **certain *specific points in linear time* when *specific aspects of morphogenetic expansion* via Frequency Accretion are intended to occur.** The **original cycles of growth and expansion inherent to the human form** occurred between **1. fetal integration through age 11-12, 2. age 11-12 through age 22, 3.age 22 through age 33** and **4. age 33 through age 44**, at which time the human morphogenetic field would reach its **full expansion,** the **DNA** would reach its organic **12-strand activation**, the identity would expand to hold **12-dimensionsof consciousness** within embodiment, and the organism would have **full conscious control of *manifesting and de-manifesting* its 15-Dimensional anatomy in time. The body forms of the original human imprint are IMMORTAL** - deterioration and death were **not a part of the original *Imprint for Health*** within the human Bio-Spiritual makeup. Through understanding how the natural process of growth and expansion in time were INTENDED to work, we can also understand the core reason for the manifestation of the SYSTEMS MALFUNCTIONS - the conditions of biological and mental deterioration that lead to the condition of **BIOLOGICAL DEATH** - a condition that is **INORGANIC TO THE ORIGINAL GENETIC IMPRINT OF THE HUMAN FORM!.**

The First Natural Expansion Cycle

In the original human genetic blueprint, the frequencies and dimensions of consciousness corresponding to the wave spectra of Dimensions 1, 2, and 3, were intended to accrete or draw into the morphogenetic field between fetal integration and the age of 11-12. At the age of 11-12 the wave spectra of which the body was composed would take a **leap in frequency**, beginning the process of **expanding the morphogenetic field** to hold the higher frequency wave spectra of **Dimensions 4, 5 and 6.**

The *expansion of the morphogenetic imprint* is reflected in the *DNA*, and *chemical and hormonal balances* within the body, which govern the function of the *metabolic rate* and the operation of molecular function. The **first cycle** of *morphogenetic expansion* into the wave spectra of **Dimensions 4, 5 and 6,** represents the process of **SOUL INTEGRATION** for the embodied identity, and **begins the natural process of de-densification of matter -** *by lowering of morphogenetic vibration, raising of oscillation rate and frequency, and shortening of the wave-length of the scalar waves that compose the matter form and consciousness.* **These changes in *intrinsic wave characteristics* have their reflection within the behavior of *atomic structure*-** as the morphogenetic field wave spectra **raises in frequency**, the ***electron spin*** **around the atomic nucleus *slows*,** thereby increasing the electrical nature, and reducing the magnetic orientation of atomic function, ***transmuting the perceivable matter substance*** **of the cells to a *less dense* state of being. Completion of the *Soul Integration process* marks the transmutation of *Carbon-based Gross-matter density* into the *Carbon-Silica based Semi-etheric matter density*, and the expansion of embodied awareness from 3 to 6 Dimensions of consciousness.**

Completing Expansion

This process of *progressive morphogenetic field and consciousness expansion* and *resulting de-densification of biological matter* continued throughout the remaining *expansion cycles*. Age **22 to age 33** marked the accretion of the **7th, 8th and 9th -Dimensional** wave spectra, the process of ***Over-Soul Integration,*** expansion into a **9-Dimensional consciousness** and transmutation to a ***Silica-based Etheric matter density biology***. Age **33 to 44** marked the accretion of the **10th, 11th and 12th Dimensional** wave spectra, the process of ***MAHARIC INTEGRATION***, expansion into a **12-Dimensional consciousness** and transmutation **out of biology** into the **Crystalline Liquid Light wave form of Pre-matter substance.** At this point the identity is considered a **full AVATAR** and has complete control over ***manifesting and de-manifesting*** the **Holographic Projection** of its desired body imprint in time.

Growth Rates and Morphogenetic Expansion

ORIGINAL HUMAN GENETIC IMPRINT

PERSONALITY INTEGRATION (Telluric Integration) – **BIRTH - AGE 12**
Expands the Morphogenetic Field through accreting the frequencies of Dimensions 1-3 and activating DNA Strand Templates 1-3.

SOUL INTEGRATION (Doradic Integration) – **AGE 12 – 22**
Expands the Morphogenetic Field through accreting the frequencies of Dimensions 4-6 and activating DNA Strand Templates 4-6.

OVERSOUL INTEGRATION (Teuric Integration) – **AGE 22 – 33**
Expands the Morphogenetic Field through accreting the frequencies of Dimensions 7-9 and activating DNA Strand Templates 7-9.

CHRISTOS AVATAR INTEGRATION (Maharic Integration) –**AGE 33 – 44**
Expands the Morphogenetic Field through accreting the frequencies of Dimensions 10-12 and activating DNA Strand Templates 10-12. Allows full Cellular Transmutation and Dimensional Ascension to the 12th Dimensional Pre-matter Liquid Light HYDROPLASMIC Substance Liquid Silica) and conscious mastery over 12 Dimensions within the 15-Dimensional Time Matrix. A "Christed" Immortal Being possessing non-polarized Unity or Christ Consciousness.

CONTEMPORARY HUMAN GENETIC IMPRINT

PERSONALITY INTEGRATION (Telluric Integration) – **BIRTH – DEATH**

SOUL INTEGRATION (Doradic Integration) – KATHARA BLOCKAGE

OVERSOUL INTEGRATION (Teuric Integration) – KATHARA BLOCKAGE

CHRISTOS AVATAR INTEGRATION (Maharic Integration) – KATHARA BLOCKAGE. Deterioration and Death of the physical body occur rather than progressive frequency accretion for Cellular Transmutation and Dimensional Ascension. Distortions in the personal Kathara Grid and Scalar Shields manifest in the DNA Template and block natural inflow and Morphogenetic Field accretion of frequency from the higher dimensional stations of identity. Conscious awareness remains polarized in Duality Consciousness and physical body is rendered finite and vulnerable to dis-ease.

THE 4 NATURAL CYCLES OF MORPHOGENETIC FIELD EXPANSION

THE *4 CYCLES OF MORPHOGENETIC FIELD EXPANSION* REPRESENT THE *ORIGINAL ORGANIC PROCESS* BY WHICH THE HUMAN ORGANISM WAS *INTENDED* TO EVOLVE THROUGH TIME.

The Inorganic Condition of Human Illness and Death

The processes of biological and mental deterioration and resulting DEATH of the biological form, which we presently view as FACTS OF LIFE regarding the process of human passage through linear time, ARE NOT ORGANIC CONDITIONS TO THE HUMAN ORGANISM. Other life forms, such as the **animal kingdoms**, that manifest through morphogenetic fields having *less than 12 Kathara Centers active* within their Kathara Grid, cannot undergo the full process of bodily transmutation into Pre-matter substance, and cannot fully expand their embodied awareness into 12th-Dimensional consciousness. For the **animal kingdom**, the process of evolution is also Frequency Accretion, but it is achieved through expanding one's body form to its full capacity, then entering the consciousness into a morphogenetic imprint with a bit higher capacity, until eventually the consciousness can accrete enough frequency to take it into the next level of matter de-densification and consciousness expansion. **Death is natural to the organic animal kingdom.**

Humans were intended to undergo the *full spectrum of morphogenetic expansion and cellular transmutation* WITHIN ONE ETERNAL BODY IMAGE, that was built upon a program of *full activation* of the *12 Kathara Centers* of the Kathara Grid. Death is NOT natural for humans.

Kathara Grid Blockage and Molecular Compaction

The process of *biological deterioration* and progressive movement toward biological DEATH is the *direct result* of unnatural BLOCKAGES WITHIN THE KATHARA GRID HOLOGRAPHIC TEMPLATE. Blockages within the Kathara Grid program presently block the process of Frequency Accretion **at the beginning of the First Expansion Cycle-** as accretion of the 4th, 5th and 6th Dimensional frequencies representing the *Soul Integration Process* is blocked within the Kathara Grid and thus within the biological DNA, the incoming frequencies of the Soul Level consciousness **push on and CRUSH the morphogenetic field**, rather than *expand the morphogenetic field into a higher level of frequency holding*. When the incoming scalar wave frequencies of Soul Integration begin to crush the morphogenetic field, a condition called **MOLECULAR COMPACTION** is created.

MOLECULAR COMPACTION -

***over exposure to scalar wave spectra* and its resulting *fragmentation of the morphogenetic field and Kathara Grid* - IS THE *ONLY TRUE CAUSE OF DIS-EASE*, AND THE *CORE CAUSAL ELEMENT* OF THE *UNNATURAL DEATH EXPERIENCE*.**

The *Maharic Shield* and Human Healing

Within the imprint of the **MAHARIC SHIELD** of human morphogenetic structure, is the dormant imprint of the *10th, 11th and 12th Dimensional electro-tonal programs* that represent the **ORGANIC PROGRAM OF THE KATHARA GRID. Through activation of the Maharic Shield** - the scalar standing-wave template that corresponds to the Pre-matter Liquid Light fields of Dimensions 10, 11 and 12 - **blockages within the Kathara Grid Holographic Template, that result in the genetic mutation that causes Molecular Compaction, can be progressively realigned with the original IMMORTAL IMPRINT FOR HEALTH.** Though it takes time and consistent application of **Kathara Healing Technique** to fully regenerate the **Imprint for Health and Immortality** within the 15-Dimensional Levels of human anatomy, working to heal distortions within the Kathara Grid through **MAHARIC INFUSION** (running the frequencies and wave spectra of the *Maharic Shield* through the Kathara Grid, Bio-energetic System and Body), **can create progressively more observable affects in creating and maintaining health, and slowing the bodily deterioration process, within the normal life span of an individual.**

Potentials of Kathara Healing

The true beauty and significance of *Kathara Healing* is that it can not only do wonders for creating and maintaining *higher standards of health* within the average human life span, but within the applications of KATHARA MECHANICS lies the *promise and potential* of human biological IMMORTALITY and the fulfillment of SPIRITUAL ACTUALIZATION and evolution to MASTERY- through progressive embodiment of and conscious mastery over our 15-Dimensional Anatomy and the creative process of Holographic Projection within the dimensionalized fields of space, time and matter.

KATHARA HEALING HOLDS THE KEY TO HUMANITY RECLAIMING ITS ANCIENT HERITAGE OF THE ETERNAL SELF AND IMMORTAL AT-ONE-MENT WITH SOURCE.

LEVELS OF HUMAN ANATOMY- THE EMBODIED KATHARA

The Human Body Kathara Grid

The Kathara Grid of the human body is the *Core Holographic Template* upon which the morphogenetic scalar-wave template, and all other levels of identity and form are built. It is the **"Control Center"** for manifestation of the human being.

LEVEL-1 KATHARA 12-TREE GRID

THE KATHARA GRID LEVEL-1: 12-TREE

Level-1 Kathara Grid 12-Tree: The first level of the Kathara Grid **sets the *pattern for dimensionalization of consciousness* through which identity can enter the Time Matrix for the Holographic Experience of manifestation. The Kathara Grid 12-Tree is the first wave spectra composite through which consciousness anchors itself into dimensionality in order to experience manifestation of form.** All things within dimensional structure have the Kathara 12-Tree at the core, but not all manifest forms have the full pattern of the 12-Tree *activated* within their morphogenetic pattern. The 12-Tree is composed of **12 Kathara Centers** or *Kathara Spheres*, and **15 Kathara Lines.**

KATHARA CENTERS

The **Kathara Centers** are *crystallizations of Partiki units* that hold *composite electro-tonal energy signatures for each dimensional band.* The electro-tonal programs of the Kathara Centers hold the **foundation template** upon which the biological **DNA** will manifest.

KATHARA LINES

The **Kathara Lines** are sequences of *Partiki Phasing rhythms* ("flashing on and off"/fission-fusion sequences) that *transfer the dimensionalized electro-tonal programs of the Kathara Centers between each center*, to form the **core level of the scalar standing-wave template** upon which subsequent levels of the Kathara Grid, Bio-energetic System and body will build.

KATHARA HEALING AND THE LEVEL-1 12-TREE GRID

The *Kathara Centers* and *Kathara Lines* are used in *Kathara Healing* to realign and awaken the original 12-Dimensional imprint within the human morphogenetic field, in order to *reinstate the original Imprint for Health and Bio-Spiritual Mastery* within the operational evolutionary blueprint for the Body-Mind-Spirit System of the Human form.

Level-1 Kathara 12-Tree Grid

of the INTERNAL TEMPLAR

KATHARA CENTERS

#12 – *The Capstone*
In 10th Chakra
6" above the head

# 11 – *Crown Jewel*
In Pineal Gland
at center of brain

# 9 & # 10 – *The Shar*
In brain
behind ears

# 8 – *Spark of Orion*
In Thyroid Gland
at 8th Chakra

# 6 & # 7 – *The Wings*
In shoulder indents
through to back

# 5 – *Seat of Amenti*
At Navel in
Central Body Current

# 3 & # 4 – *The Dal*
male: Ureters
female: Ovaries

# 2 – *Naum-Yun*
male: Prostate Gland
female: Vagina

# 1 – *Earth Core*
at center of Earth
in 13th Chakra

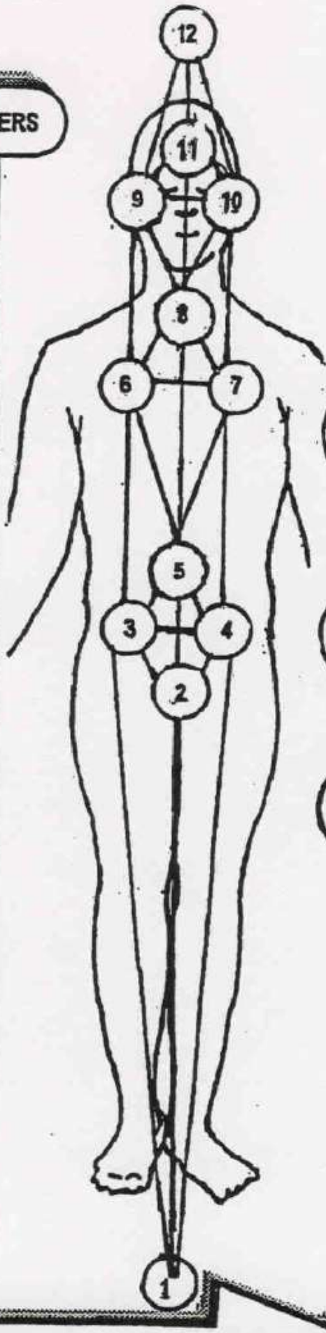
12
11
9
10
8
6
7
5
3
4
2
1

Placement of *Kathara Level-1 12-Tree Grid Kathara Centers* and *Kathara Lines* within the human body.

Central Vertical Kathara Line
controls all other Kathara Lines. Kathara Centers on *Central Vertical Kathara Line* control all other Kathara Centers.

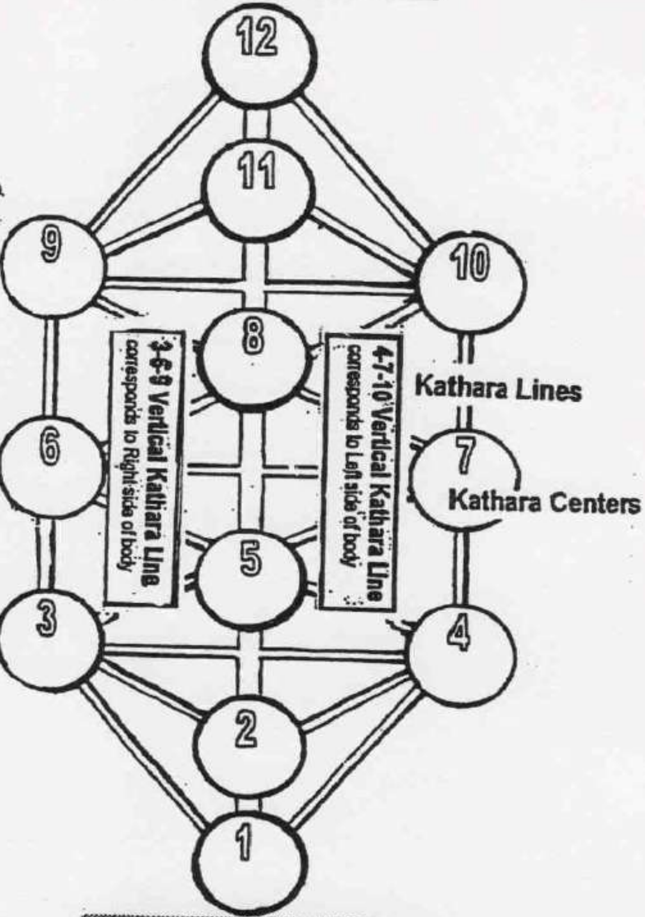
12
11
9
10
8
3-6-9 Vertical Kathara Line
corresponds to Right side of body
4-7-10 Vertical Kathara Line
corresponds to Left side of body
Kathara Lines
6
7
Kathara Centers
5
3
4
2
1

Electro-tonal Programs within the ***3 Vertical Kathara Lines*** control the programs in all other Kathara Lines.

LEVEL-2 KATHARA CRYSTAL SEALS GRID

Level-2: Crystal Seals and 15 Primary Chakras

The Crystal Seals are groups of 3-Dimensional Partiki scalar wave composites that regulate the rate of Partiki Phasing to create the base structures upon which dimensionalization is formed. The *flow of frequency* between dimensional bands and Harmonic Universes, the fixed expansion and contraction/ fission and fusion rates of Partiki, the *Vibration-Oscillation Rates* of Partiki and the *Angular Rotation of Particle Spin* are all regulated by the Crystal Seals.

Star Crystal Seals are positioned between the Chakra Centers along the **Central Body Current**, and **Seed Crystal Seals** are positioned between them, and serve as the *point of composite frequency* out of which the **15 Primary Chakra Center Vortices** emerge.

HOVA BODIES AND SHIELDS

5 Dimensionalized Hova Body Shields and 5 Auric Capsules

Hova Bodies are **3-Dimensional, tri-tonal scalar standing-wave grids** composed of numerous sets of **Crystal Seals** that form the electro-tonal program upon which the **Kathara Grid Level-3: Doradic Points** emerge. **The Hova Body Grids are called Crystal Shields.** The **5 Dimensionalized Hova Bodies** are structured as **5 Concentric Spheres** of energy that *surround and permeate the human body* and serve to *hold the embodying consciousness into dimensionalized levels* through which the **Holographic Projection** of *experienced manifestation* can occur. The Hova Bodies form **"tissue capsules"** of *subtle energy* around **each 3-dimensional Harmonic** level of the **Bio-energetic Field**, within which the **15 Dimensional levels** of the **Auric Field** are stationed.

In Kathara Healing, restoration of the original **4th-Harmonic Maharic Shield** *Imprint for Health* is used to *realign and clear blockages* from the lower Hova Bodies and corresponding **Kathara Centers** and **Kathara Lines**, to awaken the **Imprint for Health and Spiritual Actualization** within the Body-Mind-Spirit System.

HOVA BODIES, SHIELDS & SIGNETS

5 Ariea Hova Bodies, Kathara Grid and Harmonics of Manifestation

Hova Bodies are *tri-tonal scalar grids* that form *spherical energy "tissue capsules"* around each 3-dimensonal level of the multi-dimensional identity. Within 15-Dimensional Human Anatomy, there are **5 Dimensionalized Hova Bodies** called **Ariea Hova Bodies**. The Ariea Hova Bodies form out of the electro-tonal programs held within the **Kathara Centers** of the *Level-1 Kathara 12-Tree Grid.* The *12th Kathara Center* of the Kathara 12-Tree Grid connects to the *12th Kathara Center* of the *Universal Kathara Grid*, just as the *1st Kathara Center* in the personal Kathara connects to the *1st Kathara Center* in the *Planetary Kathara Grid. The 12th and 1st Kathara Centers,the personal Kathara Grid and morphogenetic field connect to the planetary and universal morphogenetic fields.* **When working with the Hova Bodies we first focus upon the lower four , as these must be functioning properly before the 5th Hova Body and its connection to the Universal Kathara Grid can be activated within the Bio-energetic Field.** Four of the 5 Ariea Hova Bodies are directly connected to the Crystal Seals within the human body **Central Body Current** on a **HORIZONTAL PLANE**. These first **4 Hova Bodies** correspond to each of the **12 Kathara Centers** of the personal Kathara 12-Tree Grid. **The Hova Bodies represent Harmonics of Manifestation.**

Hova Bodies, SIGNETS and SHIELDS

Each Hova Body **emanates outward** from a set of **3 Kathara Centers** and their corresponding **Crystal Seals. The *core template of the Hova Body* is a *Triad of Crystalline Spheres* that connect directly to the *3 Kathara Centers* to which they correspond.** This Triad of Crystal Spheres is called a **SIGNET. The Signets are the smaller, *fixed, consolidated* frequency points out of which the *Seed Crystal Seals* that form the *Chakras* manifest.** The Signets and the smaller Crystal Seal formations that manifest through them, create a **disc** of spinning, **horizontal wave spectra** that direct the function of *Hova Body merger*; these scalar discs are the scalar templates upon which the Hova Bodies manifest, and are called **SHIELDS.**

The *Shields and the Signets* that direct them are considered to be the *core of the Level-2 Crystal Seals Kathara Grid.* Each of the 4 Hova Bodies of the first 12 Dimensions form upon a set of **3 Signets**. Each **Signet *transmits a flash-line*** (Partiki Phasing Sequence) **on a different axis**, one **vertical**, one **horizontal** and one **diagonal**, forming a **3-plane projection of flash-lines** that form the base of the **3-dimensional Holographic projection.**

Kathara Centers and Activating the Signets and Shields

In each set of *3 Signets* at least one corresponds to the *Central Vertical Kathara Line*, the control center of the Level-1 Kathara Grid 12-Tree. The Signets and Shields can therefore be used to direct the function of the Kathara Grid 12-Tree and all built upon it, and the Kathara Centers can be used to trigger activation of the Signets and Shields within the Hova Bodies. When a **Signet is activated** its corresponding **disc or *Shield* begins to spin**. When a certain spin rate of the Shield is reached, it triggers the activation of the next set of Signets in the next Harmonic setting the next Shield to spin. **When *2 Shields activate* and reach a reciprocal spin rate the *tissue capsule* separating the 2 corresponding Hova Bodies releases and the *2 Hova Bodies merge*, which is the process by which *Higher Identity Integration* occurs. It is also the process by which the *Merkaba Vehicle* is activated within the Bio-energetic field.** When all **four Shields** activate and spin, the body matter is able to **transmute** out of matter density and into pure wave form, which represents Maharic/ Avatar Integration for the identity.

THE 5 HOVA BODY SHIELDS

The Hova Bodies and the Shields

Each Hova Body contains at its core a Shield through which the Hova Body functions are governed. The **Nada Hova Body** of dimension 1-2-3 forms on the **TELLURIC SHIELD**, its control center being the 2nd Kathara Center that is composed of 2nd Dimensional Frequencies. The **Alphi Hova Body** of dimensions 4-5-6 forms on the **DORADIC SHIELD**, its control center being the DORA or Archetype of the 5th Dimensional Frequencies. The **Betcha Hova Body** of dimensions 7-8-9 forms on the **TEURIC SHIELD**, its control center being the TEURA or Monad of the 8th Dimensional Frequencies. The **Mahara Hova Body** forms on the **MAHARIC SHIELD**, its control center being the *Mahunta* or Avatar Core of the 11th Dimensional Frequencies. The 5th Hova Body, the **Raja Hova Body**, forms on a set of Signets that correspond to **stellar points** within the **Universal Kathara Grid.** Its Shield, the **RISHIC SHIELD**, rotates on a **vertical plane** and it comes into activation within the personal morphogenetic field only after the four lower Shields have transmuted the body into Pre-matter Liquid Light of the MAHARIC SHIELD. **In terms of healing and advancing evolution of the earthly identity, the first four Shields are of greatest significance.**

Hova Bodies, Shields, Signets and Healing

In all healing application knowledge of the Shields is of great value, for any manifest condition of disorder will have its first manifestations within the Kathara Grid and the Shields that form upon it. The process of healing, therefore, is the process of *restoring the Shields to their proper order* so they may activate to build the Merkaba Vehicle through progressive Higher Identity Integration. **In Kathara Healing we begin with the MAHARIC SHIELD**, the Shield that contains the *pure form of electro-tonal programs upon which the consciousness first entered matter density* through the Pre-matter Liquid Light Fields of the 4th Harmonic - Dimensions 10-11-12. **Through activating and progressively drawing in the highest 11th and 12th-Dimensional frequencies of the Maharic Shield, distortions in all lower Shields can be more rapidly realigned with their original pattern.** When we work with the **Kathara Centers** of the Level-1 12-Tree Grid we simultaneous create change within the **SIGNETS** that direct the function of the **SHIELDS**, and so we are able to more rapidly facilitate appropriate realignment of the Shields and Hova Bodies. **THE CORE OF HEALING IS RESTORING THE HOVA BODIES AND THE SHIELDS UPON WHICH THEY MANIFEST TO THEIR ORIGINAL INTEGRAITY OF ORDER.**

Shields, Signets and Diodic Points

Through the interaction and interplay of multi-harmonic frequency between the **Shields, Signets** and their **Crystal Seals, Kathara Grid Level-3- DIODIC POINTS** come into manifestation, to fill out the 3-plane, 3-dimensional "armature" of scalar waves and flash-lines, upon which more Partiki units will cluster and crystallize to build up the particle base for form manifestation. The **Diodic Points** within the body and morphogenetic field connect directly to the Shields of the Hova Bodies in areas where the *flash-lines and scalar points* within the Crystal Seals of the Shields *align with and pass through each other.* **Flash-Lines from one Hova Body Shield crossing through those in other Hova Body Shields form *COORDINATE POINTS*, or "Gateways" of energy-consciousness *transference* through which frequency and awareness from one Hova Body and identity level can cross through into other Hova Bodies and identity levels.** The *Coordinate Points*, or *regularly scheduled points of flash-line intersection*, form **tiny vortices of energy** , which keep energy and consciousness flowing throughout the multiple dimensional levels of identity and anatomy. **The vortices of energy created by the Coordinate Points of flash-line sequences are called *DIODIC POINTS* and they have a counter reflection within *anti-particle manifestation* called *MIODIC POINTS.*** Diodic and Miodic Points are sets of **White and Black Holes** that circulate energy throughout the anatomy.

LEVEL-3 KATHARA GRID DIODIC POINTS

Diodic and Miodic Points

Diodic/ Miodic Points represent *concentrated areas of energy* that emerge at points where *dimensionalized levels of scalar grids cross over and through each other*, creating minute **vortices** of wave spectra that serve as **frequency modulation zones** (*scalar flash-cycle transition points*) between Dimensional Frequency Bands. Diodic Point vortices are **White Hole** vortices that *transmit energy* into the Particle Universe from their counterpart **Miodic Point** vortices within the Anti-particle Universe. Through the structure of Dimensionalization, each Harmonic Universe, and the structures contained there within, have within their morphogenetic fields *sets of Diodic and Miodic Points* - **sets of White and Black Holes** - that regulate the flux of consciousness-energy between systems.

Within the biological form and morphogenetic field, Diodic Points are located in **physical body regions** where Inter-dimensional and Inter-Harmonic **Horizontal and Vertical Kathara Lines intersect** to form small **vortices** of inter-dimensional or inter-harmonic wave spectra that serve as **Gateways of Consciousness** between Dimensional Time Continua of Harmonic Time cycles. They connect physical bodies to their **anti-particle counterparts** and intersect with the Vertical and Horizontal relationships of wave spectra within the **DNA/ RNA** and bio-energetic field, through sets of White and Black Holes at the center of the biological Harmonic scalar grids.

Diodic Points and 3-Dimensionality

Diodic Points control the **Diagonal Kathara Lines**, the **sequences of Partiki Phasing** or ***"flash-lines"*** that run on a **diagonal** between the Kathara Centers. Whereas the **Kathara Level-1 Kathara Lines** direct the **Vertical flash-line sequences** and the **Level-2 Crystal Seals** control the primary **Horizontal flash-lines** and their relationships to the Vertical flash-lines, the **Diodic Points** further regulate the relationships of flash-line sequences through the Kathara Grid by serving as **"turnstiles of energy"**. The Diodic Points intersect with the crossover points of the Vertical and Horizontal flash-lines , *adding a further level of modulation* to the frequencies passing through, allowing a **3-dimensional Harmonic Projection of flash-lines** (flash-lines projecting on 3 differ angular planes)to build within the morphogenetic field. **The Diodic Points add the third plane of projection to the Kathara Grid, beginning the formation of what will appear as a 3-dimensional "armature" of flash-points upon which particles will accrete to "fill out" the Holographic illusion of 3-dimensionality.**

Diodic Points and Healing

In Kathara Healing Diodic Points are used to open the **Diagonal Kathara Lines** for reprogramming. When Diodic Points become blocked, due to blockages within the Kathara Grid (which appear as "gaps" or missing sequences of flashes along the flash-lines of Partiki Phasing), energy builds up in the *Level-3 Kathara Grid*, forming *distortions* in the natural pattern that will manifest in the body and bio-field systems as *disease or disharmonic condition*. In later applications of Kathara Healing we will learn to identify points of Diodic and Miodic build-up in the body, and use *primary flash-line sequences applied through touch points* to realign the natural flash-point sequences and clear the Diodic Blockages.

Shields, Signets and the Kathara Grid

Hova Bodies **are spherical scalar wave grids that form tissue capsules between each 3-Dimensional Harmonic of Manifestation.**
Signets **are the core points of consolidated frequency that hold the programs for the Scalar grid SHIELDS upon which the Hova Bodies manifest.**

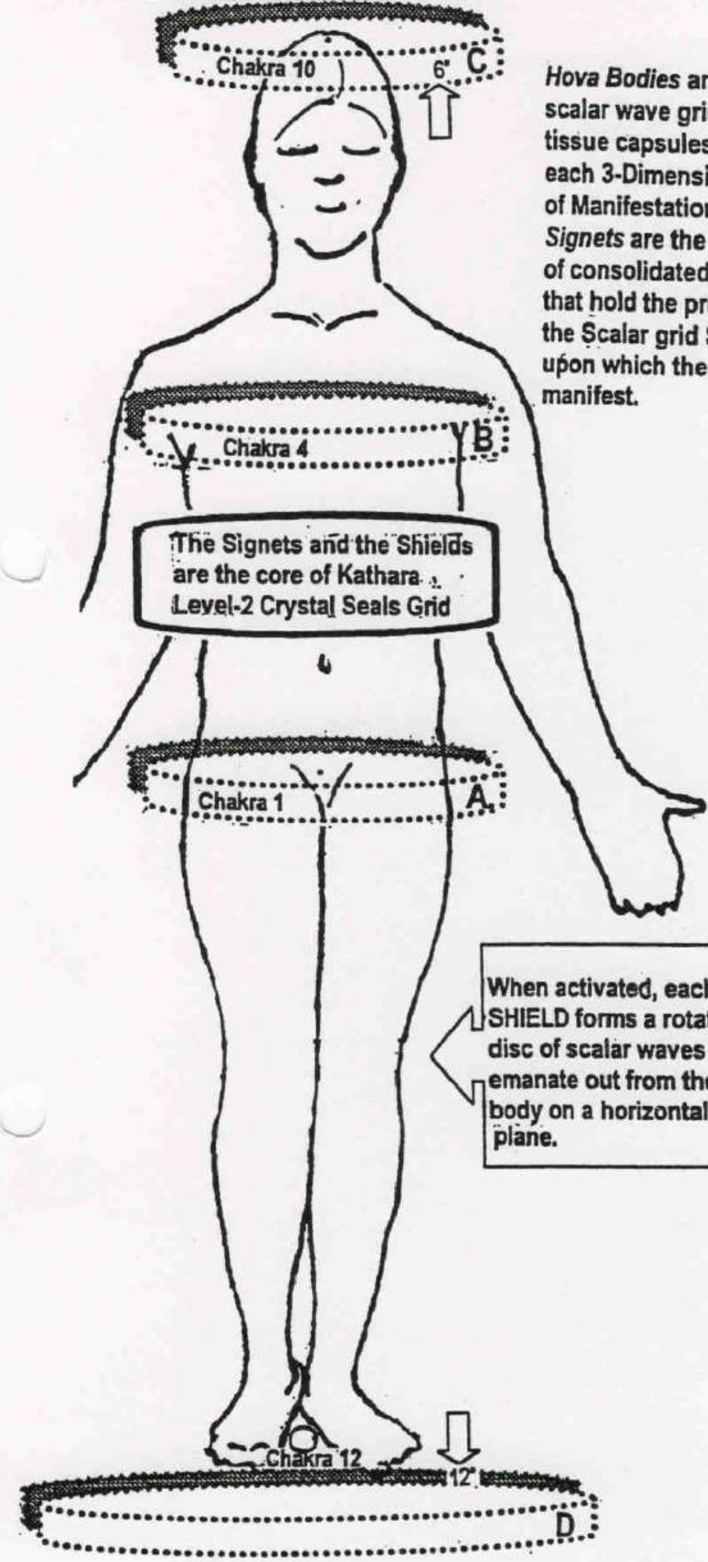
Chakra 10
6"
C
Chakra 4
B
Chakra 1
A
Chakra 12
12"
D

The Signets and the Shields are the core of Kathara Level-2 Crystal Seals Grid

When activated, each SHIELD forms a rotating disc of scalar waves that emanate out from the body on a horizontal plane.

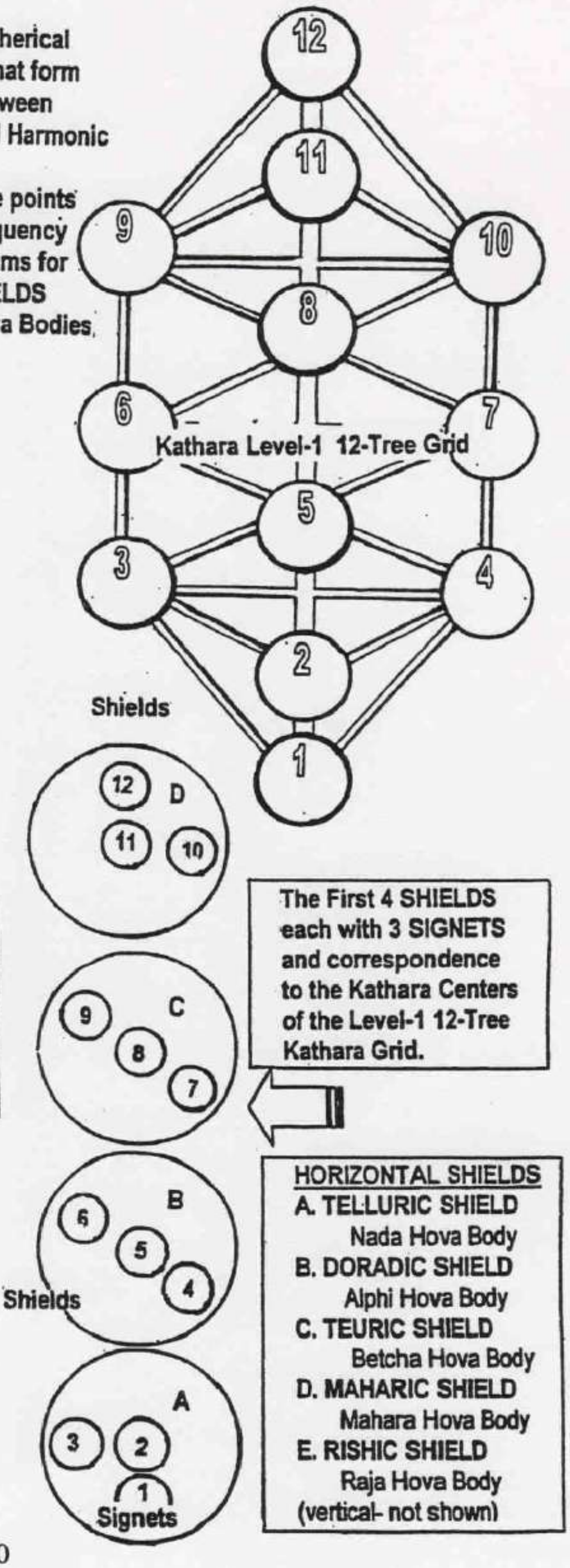
12
11
9
10
8
6
7
Kathara Level-1 12-Tree Grid
5
3
4
2
1
Shields
12
D
11
10
C
9
8
7
B
6
5
4
Shields
A
3
2
1
Signets

The First 4 SHIELDS each with 3 SIGNETS and correspondence to the Kathara Centers of the Level-1 12-Tree Kathara Grid.

HORIZONTAL SHIELDS
A. TELLURIC SHIELD
Nada Hova Body
B. DORADIC SHIELD
Alphi Hova Body
C. TEURIC SHIELD
Betcha Hova Body
D. MAHARIC SHIELD
Mahara Hova Body
E. RISHIC SHIELD
Raja Hova Body
(vertical- not shown)

Level-2 Kathara Crystal Seals Grid and Chakras

Chakras (C & MC)

SIGNET
Star Crystal Seals (S)

15 Chakras (C & MC)
7 Primary Chakras (C)
- C1- Base
- C2- Sacral
- C3-Solar Plexus
- C4- Heart
- C5- Throat
- C6- 3rd Eye- Pituitary
- C7- Crown-Pineal

8 Morphogenetic Chakras (MC)
- MC8-Thymus
- MC9-Thalamus
- MC10- Galactic 1
- MC11-Galactic 2
- MC12- Earth Star
- MC13-Earth Core
- MC14-Universal 1
- MC15-Universal 2

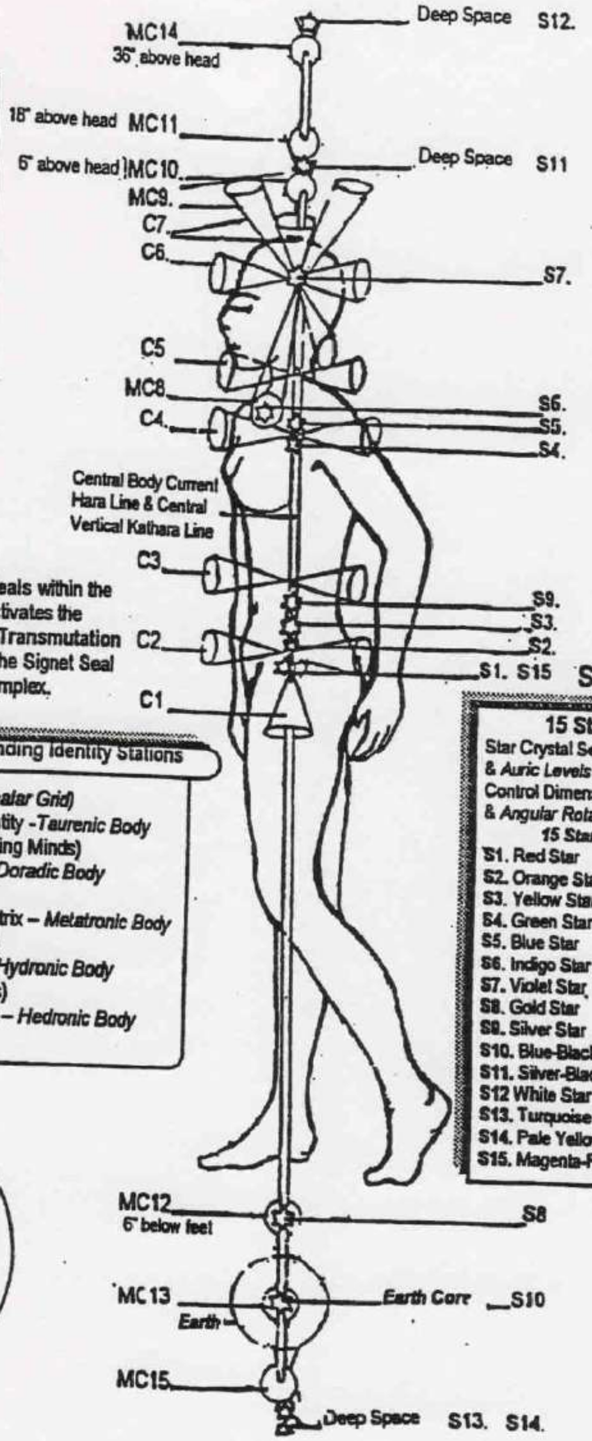
MC14
36" above head
Deep Space S12.
18" above head MC11
6" above head MC10
Deep Space S11
MC9.
C7.
C6.
S7.
C5
MC8
C4.
S6.
S5.
S4.
Central Body Current Hara Line & Central Vertical Kathara Line
C3
S9.
S3.
C2
S2.
S1. S15
C1
MC12
6" below feet
S8
MC 13
Earth
Earth Core S10
MC15
Deep Space S13. S14.

The 5 Hova Bodies progressively merge with release of the embodied Signet Star Crystal Seals, forming the various phases of the personal Merkaba Vehicle within the bio-energetic field.

Release of the SIGNET Star Crystal Seals within the embodied Internal Templar Complex activates the 12-Strand DNA Template for Cellular Transmutation and Dimensional Ascension through the Signet Seal Star Gates of the Universal Templar Complex.

Signet Seals

15 Star Crystal Seals

Star Crystal Seals keep human identity & *Auric Levels* dimensionally separate. Control Dimensional *Merkaba Field Axis* & *Angular Rotation of Particle Spin.*

15 Star Crystal Seals (S):
- S1. Red Star
- S2. Orange Star
- S3. Yellow Star
- S4. Green Star
- S5. Blue Star
- S6. Indigo Star (Blue-Violet)
- S7. Violet Star
- S8. Gold Star
- S9. Silver Star
- S10. Blue-Black Star
- S11. Silver-Black Star
- S12 White Star
- S13. Turquoise Star
- S14. Pale Yellow Star
- S15. Magenta-Pink Star

Hova Bodies and Corresponding Identity Stations

("*Hova*" means *Tri-tonal Scalar Grid*)

1. Nada Hova – HU-1 Incarnate Identity -*Taurenic Body* (Subconscious, Instinctual & Reasoning Minds)
2. Alphi Hova – HU-2 Soul Matrix – *Doradic Body* (Astral, Archetype & Angelic Minds)
3. Betcha Hova –HU-3 OverSoul Matrix – *Metatronic Body* (Ketheric, Monadic & Kertatric Minds)
4. Mahara Hova – HU-4 Avatar ID – *Hydronic Body* (Christiac, Buddhaic & Nirvanic Minds)
5. Raja Hova – HU-5 Rishi Collective – *Hedronic Body* (Universal Conscious Mind)

5 Hova Bodies

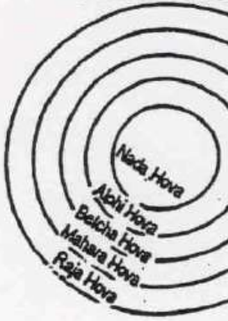
Nada Hova
Alphi Hova
Betcha Hova
Mahara Hova
Raja Hova

AXIOM LINES, AURIC LEVELS AND CHAKRAS

Diodic Points, Signets, Seed Seals and AXIOM LINES

Each **Signet** within a Hova Body Shield holds the electro-tonal program for one *Seed Crystal Seal* and one *dimensionalized flash-line sequence*. Each Seed Seal program corresponds to one *dimensionalized flash-line program* from one *Level-1 Kathara Center*. Each **Seed Crystal Seal** holds intact the electro-tonal program for one ***dimensionalized flash-line projection angle***, or **Single Axis Dimensional flash-line sequence.** The Single Axis Dimensional flash-line sequences controlled by each Seed Crystal Seal are called **AXIOM LINES.**

There are **12 Primary Axiom Lines** that correspond to the 12 Kathara Centers in the *Personal and Planetary Kathara Grids*. There are 12 additional **Stellar Axiom Lines** that correspond to the 12 Kathara Centers in the Universal Kathara Grid. **The Axiom Lines represent single-axis flash-line sequences -** (fixed points of conscious electrified sound- *electro-tonal units*- that form strands of scalar standing waves) **that intermesh to refine the scalar-grid webbing upon which Kathara Grid Level-3 Diodic Points will emerge.**

Axiom A-B Tonal Lines, Diodic-Miodic Vortices, Chakras-DNA-Portals-Ley Lines & Healing

Each of the 12 Primary Axiom Lines has a **reverse reflection** within the dimensions of the **anti-particle universe.** The 12 Primary Axiom Lines governing the **Diodic "White Hole" Vortices** of the **particle universe** are called **Axi-A-Tonal Lines** (or "Axiatonal" Lines). The corresponding reflections of the Axi-A-Tonal Lines in the **anti-particle universe,** which govern the **Miodic "Black Hole" Vortex** reflections of the Diodic Points, are called **Axi-B-Tonal Lines.** In Kathara Healing the integrity of the Axi-A-Tonal Lines is restored through realigning the electro-tonal programs within the corresponding **Kathara Centers.** The **12 Primary Axi-A-Tonal Lines** set the morphogenetic imprint upon which the **Chakra System** and foundation **DNA** organization of the human body will manifest and upon which the **Planetary Vortex/Portal System** and primary **Planetary Ley Lines** will emerge within the planetary body. Restoring the *Imprint for Health* to the **Axi-A-Tonal Lines** and corresponding **Kathara Centers** will regenerate integrity within the **DNA** and **Chakra System** of the body and will restore the organic operations of the planetary **Vortices-Portals** and **Ley Lines** of the planetary morphogenetic grid.

Restoring the **Imprint for Health** within the **Axi-A-Tonal Lines** simultaneously restores integrity to the corresponding **Diodic Points** and to the **Axi-B-Tonal Lines** and their **Miodic Points.** When working with *Kathara Techniques* that "clear the **Miasmic Imprint**" – *the anti-particle blockages within Diodic Points* – the *Diodic Points* are realigned by clearing their *Miasmic distortions*, which then reestablishes the integrity of corresponding **Axiom A and B Tonal Lines**, sending a realigned electro-tonal program back through the **Seed Seals** and the Level-1 **Kathara Centers.** If the programs within the **Kathara Centers** are **simultaneously realigned** using infusion of the frequencies from the **Maharic Shield**, the Kathara Center programs, the Axiom Lines and the Diodic Points will be progressively restored to health. If the Kathara Centers are not simultaneously realigned, the restored Axiom Line programs will be **overridden** by the Kathara Center programs until the restored programs reach *critical mass* through *prolonged repetition*. ***Clearing the Miasmic Imprint* while simultaneously *Repatterning the Level-1 Kathara Centers and Kathara Lines* via *Maharic Infusion* is the fastest way to restore the *Imprint for Health* within the Human Body-Mind-Spirit System and within the natural operations of planetary bio-mechanics.**

Axiom Lines, Diodic Points, Hova Bodies, Shields and Auric Field Levels

The Single Axis flash-line sequences of the **12 Primary Axiom Lines** cross over and through each other to form COORDINATE POINTS between the Hova Body Shields, through which the "White Hole" ***Level-3 Diodic Point Vortices*** emerge. **The varying angles of flash-line projection within the Axiom Lines create separate dimensionalized bands of frequency** (separate sets of flash-line sequences) **within the Hova Body structure.** As there are **3 Signets** and thus **3 Seed Seals** and **3 Axiom Lines** inherent to each **Hova Body Shield**, each Hova Body becomes *dimensionally compartmentalized* into **3** separate yet interrelated sections, each section having inherent sets of Diodic Points. **Each dimensionalized aspect of the Hova Bodies becomes a *Level of the Auric Field* within the bio-energetic system.** There are **12 Inner Layers of the Auric Field**, which correspond to Dimensions **1-12** and to the 12 Primary Kathara Centers, Seed Seals and Axiom Lines of the *first 4 Hova Bodies*. There are **3 Outer Layers of the Auric Field** that correspond to Dimensions and Seed Seals **13-15**, to the *first 3* Kathara Centers and Axiom Lines of the **Universal Kathara Grid** and to the *5th Hova Body*.

Seed Seals, Axiom Lines, Consolidated Diodic Points, Chakras and Auric Levels

The points in the Kathara Grid where the Single Axis Axiom Lines cross over and through the Central Vertical Kathara Line in the Level-1 Grid and the Horizontal flash-lines of the Seed Crystal Seals form fixed, **consolidated Diodic Points –larger vortices** of energy composed of many smaller Diodic Vortices. **The larger vortex points are called CHAKRAS.** Each Chakra Vortex is formed upon the electro-tonal program of its corresponding Dimensional Frequency Band, Kathara Center, Seed Seal and Axiom Line, and **each chakra corresponds to one Auric Field Level, DNA strand imprint and level of consciousness.** As *3 Signets, Seed Seals* and *Axiom Lines* correspond to each *Hova Body Shield*, **3 Chakras regulate and correspond to each of 5 Hova Bodies**, for a total of **15 Primary Chakras. Primary Chakras 1-12** correspond to Kathara Centers **1-12** in the **Personal and Planetary Kathara Grids** and **Primary Chakras 13-15** correspond to the *first 3 Kathara Centers and Axiom Lines* of the **Universal Kathara Grid.** Chakras 13-15 are considered **Universal Trans-stellar Chakras.**

The *Chakras* regulate the *transduction of flash-line sequences* from the *Dimensionalized Unified Field* and *Diodic Vortices* of the Axiom Lines and Auric Field Levels into the *Hova Body tissue capsules* and *rivulet channels* that connect to the physical, electro-magnetic, chemical and hormonal aspects of the visibly manifest form. Chakras regulate the flow of consciousness-energy between the dimensionalized Auric Field levels of the Hova Bodies, keeping the various multi-dimensional stations of consciousness separate, in order to allow for experiential perception of holographic manifestation. Each **Chakra circulates** its Single Axis flash-line sequence along the **angle of projection** carried within its corresponding **Axiom Line**, and this *angle of flash-line projection* determines the **Angular Rotation of Particle Spin (ARPS)** – or the *axis upon which particles rotate*- within each dimensionalized portion of the anatomy. There is a **90-degree shift** in the *flash-line angle of projection* from one Chakra and dimensional band to the next, which creates a 90-degree shift in the **Angular Rotation of Particle Spin** between each dimensionalized portion of the anatomy. Between each 3-dimensional Hova Body there is also a **45-degree reverse shift** in flash-line projection angle, which creates a **45-degree** rather than a 90-degree **forward shift** of flash-line projection angle between each Hova Body and between each **set of 3 Chakras** (Chakras 1-2-3 * 4-5-6 * 7-8-9 * 10-11-12 * 13-14-15). The precise angles of flash-line projection and ARPS between Chakras, creates the perceivable 3-dimensional hologram.

Chakras

A Chakra appears within the bio-energetic field as **2 cones or spirals of multi-colored energy** that are **linked together at the tip** in the **Central Body Current** through the **Seed Crystal Seal** out of which the cones emerged. **Chakras 2-6** extend from the Seed Seals outward through the body and Auric Field on a **Horizontal axis** and **Chakras 8 and 9** intersect with *6 and 7* at the *Pineal Gland* on **2 rotating Diagonal axes** running through the head. **Chakras 1 and 7 and 10-14** are stacked above and below the body on a **Vertical Axis** aligned with the *Central Body Current* (CBC) and the *Central Vertical Kathara Line*. Chakras 1&7, 10&12, 11&13 and 14&15 are actually the bottom and top cones of **4 Large Vertical Chakra Spirals**, but are counted as separate chakras. **Chakra 15** is a **"roving chakra"**, which travels the course of the **12 Primary Axiom Lines** far out in Level–15 of the Auric Field. Each Chakra draws in energy from and transmits energy into the dimensional Unified Field corresponding to the *chakra number* and carries as its **dominant perceivable color the hue** characteristic to the **primary frequency band wave-length** of its **dimensional affiliation.**

Primary Color of Chakras

Chakra	Dimension	Color	Chakra	Dimension	Color
CK-1	D-1	RED	**CK-8**	D-8	GOLD
CK-2	D-2	ORANGE	**CK-9**	D-9	SILVER
CK-3	D-3	YELLOW	**CK-10**	D-10	BLUE-BLACK
CK-4	D-4	GREEN	**CK-11**	D-11	SILVER-BLACK
CK-5	D-5	BLUE	**CK-12**	D-12	WHITE
CK-6	D-6	BLUE-VIOLET (INDIGO)	**CK-13**	D-13	PALE TURQUOISE
CK-7	D-7	VIOLET	**CK-14**	D-14	PALE YELLOW
			CK-15	D-15	MAGENTA-PINK

A Chakra will also carry, in lesser saturation, the colors/frequency bands associated with all of the dimensions below its own dimensional affiliation.

Chakras and Dimensional Merkaba Fields

When the Seed Seal at the core of a Chakra is released, the **2 cones of the Chakra move through each other** to form a **Dimensional Merkaba Field-** *2 counter-rotating spirals of electro-magnetic energy that take the form of a* ***Star-tetrahedron***. When the Seed Crystal Seal releases, the corresponding scalar-wave points within the Hova Body Shield are *released from* ***Dimensional Phase-lock*** – they are released from *polarity* by the re-combining of their inherent *Particum-Particles* and *Partika-anti-particles*. Through re-combining Particum and Partika the scalar standing-wave points undergo **fusion**, then a reciprocal **fission** and **replication** within the Harmonic above. Through release of the Seed Seals the *Dimensionalized Auric Field Levels* of the *lower Hova Body* align along the same *Angular Rotation of Particle Spin /* angle of flash-line projection, which transmutes them into the *angle of flash-line projection* and *Partiki Phasing rate* of the Dimensionalized Auric Field Levels of the Hova Body from the *Harmonic above.* **Seed Crystal Seals are released by infusing them with the frequencies/ flash-line sequences of the corresponding dimension from the Harmonic above.** Progressive release of the Seed Seals within the Chakras creates opening and merging of the Kathara Lines in the Level-1 12-Tree Grid, and merging of the *Hova Bodies, Shields, Diodic Points, Auric and Identity Levels* and *DNA strands, through which activation of the* ***Merkaba Vehicle*** *and de-densification of matter occurs.*

Auric Level, Chakra and Hova Body Correspondences

Each 3-Dimensional Scalar Shield forms a Spherical Electromagnetic Domain called a Hova Body.

Each Hova Body manifests the structure, and governs the function, of 3 Primary Chakras.

Chakras draw energy in from, and transmit energy into, the **Unified Fields** of each Dimension. Each Chakra carries as its **Primary Color** the hue associated with the **wave-length** of the dimensional frequency band to which the Chakra corresponds.

Each Chakra corresponds to a level of the *Auric Field* and one *Axi-A-Tonal Line*. Each ***Hova Body*** corresponds to a **set of 3 Chakras** and contains 3 dimensionalized frequency bands that form the dimensional levels of **Auric Field.**

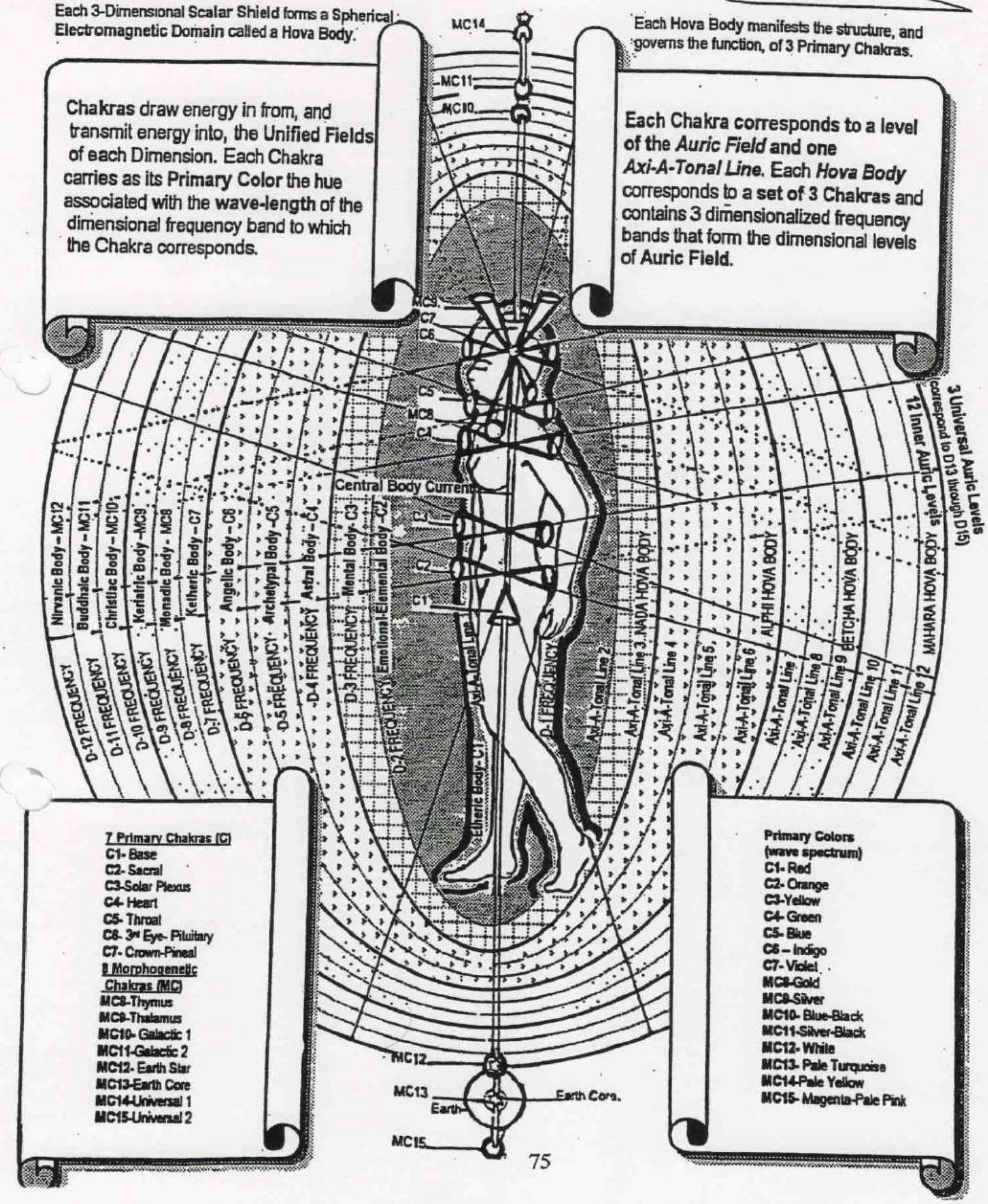
MC14
MC11
MC10
MC9
C7
C6
C5
MC8
C4
Central Body Current
C3
C2
C1
Nirvanic Body – MC12
Buddhaic Body – MC11
Christiac Body – MC10
Keriatric Body – MC9
Monadic Body – MC8
Ketheric Body – C7
Angelic Body – C6
Archetypal Body – C5
Astral Body – C4
Mental Body – C3
Emotional-Elemental Body – C2
Etheric Body – C1
D-12 FREQUENCY
D-11 FREQUENCY
D-10 FREQUENCY
D-9 FREQUENCY
D-8 FREQUENCY
D-7 FREQUENCY
D-6 FREQUENCY
D-5 FREQUENCY
D-4 FREQUENCY
D-3 FREQUENCY
D-2 FREQUENCY
D-1 FREQUENCY
Axi-A-Tonal Line
Axi-A-Tonal Line 2
Axi-A-Tonal Line 3
NADA HOVA BODY
Axi-A-Tonal Line 4
Axi-A-Tonal Line 5
Axi-A-Tonal Line 6
ALPHI HOVA BODY
Axi-A-Tonal Line
Axi-A-Tonal Line 8
Axi-A-Tonal Line 9
BETCHA HOVA BODY
Axi-A-Tonal Line 10
Axi-A-Tonal Line 11
Axi-A-Tonal Line 12
MAHARA HOVA BODY
12 Inner Auric Levels
3 Universal Auric Levels (correspond to D13 through D15)
MC12
MC13
Earth
Earth Core
MC15

7 Primary Chakras (C)
C1- Base
C2- Sacral
C3-Solar Plexus
C4- Heart
C5- Throat
C6- 3rd Eye- Pituitary
C7- Crown-Pineal
8 Morphogenetic Chakras (MC)
MC8-Thymus
MC9-Thalamus
MC10- Galactic 1
MC11-Galactic 2
MC12- Earth Star
MC13-Earth Core
MC14-Universal 1
MC15-Universal 2

Primary Colors (wave spectrum)
C1- Red
C2- Orange
C3-Yellow
C4- Green
C5- Blue
C6 – Indigo
C7- Violet
MC8-Gold
MC9-Silver
MC10- Blue-Black
MC11-Silver-Black
MC12- White
MC13- Pale Turquoise
MC14-Pale Yellow
MC15- Magenta-Pale Pink

PRIMARY COLOR OF CHAKRAS

Chakra	Dimension	Color	Chakra	Dimension	Color
CK-1	D-1	RED	CK-8	D-8	GOLD
CK-2	D-2	ORANGE	CK-9	D-9	SILVER
CK-3	D-3	YELLOW	CK-10	D-10	BLUE-BLACK
CK-4	D-4	GREEN	CK-11	D-11	SILVER-BLACK
CK-5	D-5	BLUE	CK-12	D-12	WHITE
CK-6	D-6	BLUE-VIOLET (INDIGO)	CK-13	D-13	PALE TURQUOISE
CK-7	D-7	VIOLET	CK-14	D-14	PALE YELLOW
			CK-15	D-15	MAGENTA-PINK

A Chakra will also carry, in lesser saturation, the colors/frequency bands associated with all of the dimensions below its own dimensional affiliation.

AXI-A-TONAL LINES, THE HARA LEVEL, 6 HARA POINTS AND COMPLEXES

Axiom Lines, Rotating Flash-line Projection and the Human Body

Each Seed Crystal Seal within the Signets of the Hova Body Shields holds the electro-tonal program for a Single Axis Axiom Line flash-line projection. **The flash-line sequences of each Axiom Line are not simply *stationary* single-line projections of flash-lines, but rather operate as *rotating* single-line projections.** Each rotating Single Axis flash-line projection is ***synchronized*** with the other rotating Single Axis flash-line projections of Axiom Lines from the other Seed Crystal Seals. Through the ***rotational synchronization*** of the Axiom flash-line sequences, the flash-lines of each Axiom Line regularly cross through each other, and through the vertical and horizontal flash-line sequences from the Level-1 and Level-2 Kathara Grid. The points of *synchronized flash-line crossover* form ***consistent points of intersection*** through which the **Level-3 Diodic Vortices** and **Chakras** manifest. Within the human body and Auric Field the **consistent points of flash-line crossover** can be charted in the form of **fixed lines of energy** that run through the body. **The charted Axiom Lines within the human body represent series of regularly scheduled flash-line intersection points, created through the synchronized rotation of the Single Axis Axiom flash-line sequences.**

Axiom A Lines – "Axi-A-Tonal" Lines in the Human Body

The consistent points of Axiom Line flash-line crossover that form fixed lines of energy within the body can be charted as **Axiom A and B Line** energy flows within the body. The Axiom A Lines correspond to the body manifestation in the particle universe, the B Lines to the body double in the anti-particle universe. **For Kathara Healing applications the A Lines – Axi-A-tonal Lines- of the particle body are used for various purposes of facilitating Kathara Center realignment and revitalization of the Body-Mind-Spirit System.** Within the body the Axi-A-Tonal Lines can be charted as **Primary Flow Lines** – conduits through which energy-consciousness, in the form of scalar-point flash-line sequences, circulates through the morphogenetic field. The fixed flash-line crossover points that form the **charted Axi-A-Tonal Lines** appear in the body as a set of **12 Vertical Flow Lines**, one running from each of 12 Auric Field Levels, which run through the body within the **Level-2 Crystal Seals Kathara Grid**. Each Axi-A-Tonal Line corresponds to one of 12 Dimensional Frequency Bands, Kathara Centers, Seed Crystal Seals, Chakras and DNA strand imprints. **The 12 Primary Axi-A-Tonal Lines appear to enter and leave the body through the head and feet, with conduits running through the arms, hands and out the finger-tips, and through the legs, feet and out the tips of the toes. Each Axi-A-Tonal Line appears as a colored cord of energy, its hue determined by the wave length characteristic to its corresponding dimensional frequency band.** In charting the positions of the 12 Primary Axi-A-Tonal Lines in the body, **numbering** and **color** are used to denote the *Dimensional correspondences* of each Axi-A-Tonal Line.

Axi-A-Tonal Lines 11 & 12 and the "Hara Line"

Axi-A-Tonal Lines 11 and 12, which correspond to the 11th and 12th Dimensions, Kathara Centers, Seed Seals and Chakras, run vertically through the center of the body along the Kathara Grid Level-1 **Central Vertical Kathara Line**, forming a ***twisted "rope"*** of Dark Silver (D-11) and White (D-12) energy that wraps around the Central Vertical Kathara Line Flash-line Sequence. The **Central Vertical Kathara Line**, which appears as a **thin Silver-gold flash-line of energy** running from the 14th Chakra (36"above the head) through the body and into the 13th Chakra at Earth's Core, is imbued with frequencies from the **12th, 11th and 8th Dimensions** – the frequencies of White, Dark Silver and Gold.

In *Level-2 Kathara Grid*, the rotating single axis flash-lines of *Axi-A-Tonal Lines 11* and *12* spiral around the thin Silver-Gold flash-line of the *Central Vertical Kathara Line,* merging into a thin, vertical tri-tonal flash-line of primarily *Gold* hue that becomes the core of the *Central Body Current.* In ancient Eastern Healing Systems this **thin Gold Vertical Line** is called the **"Hara Line"**, and is often considered the foundation of the manifest form. In truth, the **Level-1 Kathara 12-Tree Grid** and the *Signets, Crystal Seals* and *Shields* of **Kathara Level-2** are the deeper foundations upon which the Hara Line manifests.

The *Hara Level* of Kathara Level-2 Crystal Seals Grid

The mechanics of the Crystal Seals in the Level-2 Kathara Grid are complex, but can be employed with relative ease through working with the energetic dynamics of the ***outer levels* of the Level-2 Kathara Grid**, the level in which the *Hara Line* can be viewed using *9th-dimensional Higher Sensory Perception.* We refer to this level of the Level-2 Crystal Seals Grid as the **HARA LEVEL**. The Hara Level is composed of the Hara Line, and 6 of the 15 Star Crystal Seals of the Kathara Level-2 Crystal Seals Grid.

The 6 *Hara Points* (Crystal Seals) apparent from the Hara Level are as follows:

1. The 9th *Silver* Core Star Crystal Seal, called the **HARA CENTER**
2. The 6th *Indigo (or Blue-Voilet)* Star Crystal Seal, called the **SOUL STAR**
3. The 12th *White* Star Crystal Seal, called the **GALACTIC STAR**
4. The 3rd *Yellow* Star Crystal Seal, called the **SUN STAR**
5. The 8th *Gold* Star Crystal Seal, called the **EARTH STAR**
6. The 10th *Blue-Black* Star Crystal Seal, called the **EARTH CORE SEAL**

The **6 Hara Points** govern and regulate the function of **Complexes** of *energy synthesization and distribution* that control the operations of all primary body systems and bio-energetic field dynamics. In *Kathara Healing Level-2* we will explore mechanics of the Hara Level. In Level-1 Kathara Healing we will simply become familiar with Hara Level as an element of 15-dimensional human anatomy.

THE HARA AND THE KUNDALINI

The Silvery-gold energies of the **Hara Center** (9th Star Crystal Seal) are interwoven at the **base of the spine** with what is referred to as the **KUNDALINI life-force energies.** The Kundalini energies represent the **PRIMARY ENERGY CURRENTS** through which the incarnating consciousness anchors its morphogenetic field within the **Fetal Body** through the **FIRST 8 CELLS** of conception. **The Hara Center energies interweave at the base of the spine with the frequencies of the first Star Crystal Seal, the *RED Star Seal*,** in the First Cell. When the Hara Center Seal is dormant and activated only to *the 3-Dimensional Level* its Pale Silvery-gold frequencies appear to be encased within a **crimson RED sheathe** of 1st-Dimenional energy, often **making the Hara Center and Hara Line appear to be RED in color.** As the Hara Center comes into activation through progressive release of the 9th Star Crystal Seal, the Gold 8th-Dimensional and Silver 9th Dimensional frequencies of the Hara become the dominant hues, giving the Hara Center and Hara Line its natural **Silvery-gold color.** With full activation of the Hara Center and **12th-Dimensional Maharic Integration,** the tri-tone frequencies of the Hara Center and Hara Line take on a **tri-color spectrum of Pale Silver, Gold and Red-Violet** (pale Magenta), as the Kundalini primary life force currents at the base of the spine activate within the Central Vertical Current of the body. Full activation of the Kundalini merges the 7th-Dimensional **Violet,** 1st-Dimensional **Red,** 8th-Dimenional **Gold** and 9th Dimensional **Silver** frequencies with the full spectrum **White-Silver** energies of the 12th Dimensional Maharic Current.

The Hara Center, like all Star Crystal Seals, is one of the primary regulatory elements of the natural Kundalini energies. As the Kundalini energies at the base of the spine are brought out of higher dimensional dormancy, through progressive activation of the Star Crystal Seals in the Kathara Level-2 Crystal Seals grid, the Hara Center progressively reflects these changes by expanding in size and energy processing capacities and taking on the hues (wave spectra) of the higher dimensional frequencies. The **Kundalini energies** at the base of the spine are the **key to physical cellular transmutation** of the body. Kundalini energy currents **regulate the physical body's position in space-time,** keeping the physical body phase-locked into the planetary Time Cycle in which it was conceived, following the mathematical programs set by the Star Crystal Seals and the Kathara Grid. For physical transmutation to occur, the Kundalini energies of the higher dimensions must be brought into embodied activation, setting the processes of the **Internal Templar Complex** in motion.

The **DNA Template** manifests upon the programs set by the **Kathara Grid.** As the Kathara Centers and their corresponding Star Crystal Seals, Shields, Kundalini currents, Cranial-Sacral (Pineal-Tailbone) Seals, Seed Crystal Seals and Chakras activate, dormant DNA Strand Template also activate, progressively altering the **Angular Rotation of Particle Spin,** energy-processing capacities, **metabolic orientation** and **neurological processes** of the body and molecular structure. As these physical changes occur through activation of the Crystal Seals and Kundalini energies, the **Merkaba Vehicle** (interdimensional counter-rotating electro-magnetic fields in star-tetrahedral form) progressively builds within the bio-energetic field, allowing the physical body **release from phase lock** within its home space-time vector. **Activation of the Hara Center represents one of the stages of this natural transmutational process.**

THE HARA CENTER

The 9th Core Star Crystal Seal, the Hara Center (Tan Tein) and the D-9 Thalamus Complex

The **Star and Seed Crystal Seals** of the Level-2 Kathara Grid that hold consciousness into the dimensional phase-lock set by the Level-1 Kathara Centers accrete along the Central Vertical Kathara Line and the Central Body Current. **From the *Hara Level* – the *outer portions of the Level-2 Kathara Grid*** - the *12th, 9th and 6th Star Crystal Seals*, **appear as *areas of condensed light* positioned along the *vertical Hara Line*.**

The **9th Star Crystal Seal**, located about 2 " below the Navel is known as the ***Core Star Crystal Seal*** – the point of **8th and 9th dimensional frequency** of the Harmonic-3 Over-Soul Matrix (***Teuric Shield***) **that anchor the personal morphogenetic field into that of the Earth** and serve as the **center of gravity** for the adult human body. **The *Core Star Seal* appears from the Hara Level as a 2" ball of Gold (D-8) energy encased in a thin** Silvery **(D-9) membrane; in Eastern theory this ball of energy is called the *Tan Tein*, and is used for direction of bodily and Earth energies in advanced martial arts and healing practices.** In Kathara theory we simply refer to this **energy ball** as the ***Hara Center***, or the 9th **Core Star Crystal Seal.**

The ***Hara Center*** is the point at which the dimensionalizing identity enters the **Etheric Matter densities** of Harmonic Universe-3. The Hara Center represents the point into which the **Over-Soul Identity, its Monadic Core, the Betcha Hova Body** and **Kathara Centers 7, 8 and 9** first anchor within the body. The Hara Center holds the morphogenetic imprint for **DNA strands 7, 8 and 9,** corresponds to **Chakra Centers** and **Axi-A-Tonal Lines 7,8 and 9** and is a key point for repatterning the electro-tonal programs of the **Teuric Shield** and realigning and activating corresponding DNA strands, Chakras, Axi-A-Tonal Lines and body regions they govern.

The Hara Center is associated with the ***Thalamus Complex***, the chemical translation and regulatory system that governs the **9th-Dimensional interface** between various **brain-function areas** and the glandular and metabolic operations of the **Thymus Complex, Gonads, Lymphatic System** and corresponding body systems. Through the ***9th-Dimensional Thalamus Complex*** the electrical impulse, chemical and hormonal relationships between the ***10th-Dimensional Shara Complex*** and the ***8th-Dimensional Thymus Complex*** and its related body systems are governed.

The Hara Center and the Thalamus Complex link the functions of the **Mahara** and **Betcha Hova Bodies,** the **Maharic** and **Teuric Shields** and the **Avatar** and **Over-Soul** levels of consciousness and identity within the body and bio-energetic field.

The primary area of dominion of the *D-9 Thalamus Complex* is *Brain Function*.

THE SOUL STAR

The 6th Soul Star Seal and the D-8 Thymus Complex

Also visible along the Hara Line of Kathara Grid Level-2 is the **6th Star Crystal Seal**, located around the **Thymus** at the top of the breast bone and associated with the **5th and 6th dimensional frequencies** of the Harmonic-2 Soul Matrix (***Doradic Shield***). The **6th Star Seal**, visible as a **pale Blue-Violet-white point of diffused light** along the Hara Line, is called the ***Soul Star Seal.***

The *Soul Star Seal* is the point at which the dimensionalizing identity enters the **Semi-Etheric Matter densities** of Harmonic Universe-2. The Soul Star represents the point into which the **Soul Identity**, its **Archetype Core**, the **Alphi Hova Body** and **Kathara Centers 4, 5 and 6** first anchor within the body. The Soul Star holds the morphogenetic imprint for **DNA strands 4, 5 and 6**, corresponds to **Chakra Centers** and **Axi-A-tonal Lines 4, 5 and 6** and is a key point for repatterning the electro-tonal programs of the **Doradic Shield**, and realigning and activating corresponding DNA strands, Chakras, Axi-A-Tonal Lines and body regions they govern.

The Soul Star is associated with the **Thymus Complex**, the chemical translation and regulatory system that governs the **8th-Dimensional interface** between the **Pineal**, **Pituitary**, and **Thyroid Glands**, which govern the chemical and metabolic operations of the **hormones**, **Adrenal Glands**, **Endocrine System**, **Spleen**, and *crystalline* formation of the **Red Blood Cells** and **Hemoglobin.**

The *Thymus Complex* and its related glandular and metabolic systems, govern the *cyclic rhythms* by which the **DNA strands** will activate through the **Pineal Gland**, the rate of *growth and degeneration* of body tissue through the **Thyroid Gland** and through the **Pituitary Gland** and its inter-glandular processes, the functions of *electro-tonal impulse conversation, translation and distribution* through the **brain**, **Central Nervous System** and **blood** are regulated.

The *Thymus Complex* controls the **inter-glandular systems balances** between the **Pineal, Thyroid and Pituitary Glands** and regulates the *primary operations* of electrical impulse and chemical translation and transfer throughout the brain and body systems via the **Thymus Gland.** The *8th-Dimensional Thymus Complex* is governed by the functions of the *9th-Dimensional Thalamus Complex* and the *10th-Dimensional Shara Complex.*

The Soul Star and the Thymus Complex link the functions of the **Betcha** and **Alphi Hova Bodies**, the **Teuric** and **Doraidic Shields** and the **Over-Soul** and **Soul** levels of consciousness and identity within the body and bio-energetic field.

The primary area of dominion of the *D-8 Thymus Complex* is *Metabolic Function.*

THE SUN STAR

The 3rd Sun Star Seal and the Nadial Complex

Also present, but less distinctly visible along the Hara Line of Kathara Grid Level-2 is the **3rd Star Crystal Seal**, located just beneath the **Hara Center** of the *9th Star Seal* 2" below the Navel and associated with the **2nd and 3rd dimensional frequencies** of the Harmonic-1 Incarnate Matrix (***Telluric Shield***). The 3rd Star Seal, visible as a subtle **pale Yellow-Orange-Red point of diffused light** along the Hara Line, is called the ***Sun Star Seal.*** Due to its proximity to the Silver-Gold *Hara Center* 9th Star Seal the Sun Star is often considered to be part of the Hara Center, but in deeper Hara Level scanning the distinction between these crystal seals can be evidenced. The *Sun Star Seal* is the point at which the dimensionalizing identity enters the **Gross Physical-Matter densities** of Harmonic Universe-1.The Sun Star represents the point into which the **Incarnate Identity**, its **Subconscious-Telluric Core**, the **Nada Hova Body** and **Kathara Centers 1, 2 and 3** first anchor within the body. The Sun Star holds the morphogenetic imprint for **DNA strands 1, 2 and 3**, corresponds to **Chakra Centers** and **Axi-A-tonal Lines 1, 2 and 3** and is a key point for repatterning the electro-tonal programs of the **Telluric Shield**, and realigning and activating corresponding DNA strands, Chakras, Axi-A-Tonal Lines and body regions they govern.

The Sun Star is associated with the **Nadial Complex**, the electrical impulse and chemical translation and regulatory system that governs the **4th -Dimensional interface** between the **Thymus Complex, Thalamus Complex** and the functions of the **physical body systems,** through which the electrical, chemical and metabolic operations of the **Blood, Cardio-Vascular, Pulmonary** and **Respiratory Systems** are governed. The Nadial Complex governs the operations of **Heart and Breathing rhythms**, the flow of consciousness-energy or **"Prana"** from the higher dimensional anatomy into physical embodiment and regulates many key functions within the processes of **nutrient conversion and distribution** within the Skeletal and soft tissue systems. **The *Nadial Complex* is a key center for the direction of healing energies within the Nada Hova Body, Telluric Shield and physical body systems.**

In advanced applications of healing and manifestation, the Nadial Complex is combined with the attributes of the **Shara, Thalamus** and **Thymus Complexes** and with the **Universal Raja Complex** of Harmonic-**5**, which links the human body to the planetary, galactic and universal Unified Fields through the *12th and 13th Chakras* and *10th Star Crystal Seal* at Earth's core via the *Planetary Bio-Feed Interface System.* The frequencies of the *3rd Sun Star Crystal Seal*, that govern the function of the *Nadial Complex* can be combined with those of the 9th Core Star Seal *Hara Center* and the *8th and 10th Star Crystal Seals* through the 12th and 13th Chakras, *PBIS, Doradic, Teuric, and Maharic Shields* and *Raja Complex* to form a potent **Teuric-Telluric Current** for **healing and manifestation**, once the *Doradic* and *Teuric Currents* have been activated within **the body.**

The *Sun Star* and the *Nadial Complex* link the functions of the **Alphi** and **Nada Hova Bodies**, the **Doradic** and **Telluric Shields** and the **Soul** and **Incarnate** levels of consciousness and identity within the body and bio-energetic field.

The primary areas of dominion of the D-4 *Nadial Complex* are
Cardio-Vascular, Pulmonary* and *Nutrient Absorption/Conversion Functions.

THE GALACTIC STAR

The 12th Galactic Star Seal and the D-10 Shara Complex

The third point of condensed light visible along the Hara Line of Kathara Grid Level-2 is the **12th Star Crystal Seal**, located at the **14th Chakra** 36" above the head and associated with the **11th and 12th dimensional frequencies** of the Harmonic-4 Avatar Matrix (***Maharic Shield***). The **12th Star Seal**, visible as a **White-Silver point of Light emanating an inverted White-Silver-Gold *cone of energy* at the 14th Chakra**, is the White Star or ***Galactic Star Seal***.

The *Galactic Star Seal* is the point at which the dimensionalizing gestalt identity enters the **Pre-Matter Liquid Light densities** of Harmonic-4, the point at which identity begins its individuation and journey of densification into the wave spectra of perceivably experiential matter. The Galactic Star represents the point into which the **Avatar Identity**, its **Buddhaic Core**, the **Mahara Hova Body** and **Kathara Centers 10, 11 and 12** first anchor within the **Bio-energetic Field** to begin setting the scalar field for the body. The Galactic Star Seal holds the morphogenetic imprint for **DNA strands 10, 11 and 12**, corresponds to **Chakra Centers** and **Axi-A-Tonal Lines 10, 11, and 12** and is a key point for repatterning the electro-tonal programs of the **Maharic Shield**, and realigning and activating corresponding DNA strands, Chakras, Axi-A-Tonal Lines and body regions they govern.

The Galactic Star is associated with the *Shara Complex*, the regulatory system that governs the **10th-Dimensional interface** between the **D-9 Thalamus Complex** and the D-13 portion of the **Rishi Shield** in Harmonic-5. Through the Shara Complex the Scalar Fields holding the gestalt identity of Harmonic-5 Rishi Consciousness Collective- the "*Universal Family of Consciousness*" – translate via electro-tonal conduction, into the more individuated Scalar Fields that house the Avatar identity gestalt Consciousness Collective- the "*Galactic Family of Consciousness*".

The **Shara Complex *regulates the disbursement of consciousness*** in individuated form through the dimensional fields of the Time Matrix and serves to link each individuated and incarnate member of the Geomantic Entity Collective beyond the Time Matrix, to its Avatar identity and Galactic Family of Consciousness in Harmonic Universe-4, through the Universal Family of Consciousness of the Rishi identity in Harmonic Universe-5. The Shara Complex represents the primary facility through which an individual consciousness retains its indelible energetic connection to its greater **family of consciousness**, through which it is directly linked to **Source.**

The Shara Complex governs, regulates and orchestrates the transduction, translation and cyclic rhythms of **electro-tonal projection of consciousness into dimensionalization**, sets the original operational structures of **Kathara Scalar Grid** arrangement and directs the synchronistic **evolution of families of consciousness** through time. Through the operations of the Shara Complex, an identity's cycles of **incarnation and manifestation** within the 5 Harmonics of a 15-Dimensional Time Matrix are regulated, which sets the **core regulatory rhythms of Partiki Phasing**, expansion and contraction of consciousness, and vibration-oscillation ratios that **govern the formation and function** of the **Thalamus and Thymus Complexes** and all aspects of the individuated **Kathara Grid, Bio-energetic Field** and **bodily systems**. The Shara Complex controls the cycles of in-flow and out-flow of energy between the personal morphogenetic field and that of the dimensional Unified Fields, and thus affects all systems of the body and consciousness, with a primary emphasis upon the functions of the **blood, consumption-excretion and immunological facilities** of the body and of the activation of **the higher dimensional sense facilities** within the embodied consciousness.

The Shara Complex links the functions of the **Mahara and Raja Hova Bodies**, the **Maharic and Rishic Shields** and the **Avatar** and **Rishi** levels of identity within the embodied consciousness. **Primary area of dominion of the D-10 Shara Complex is the Immune System and HSP.**

THE EARTH SEALS

Earth Seals 8 and 10 and the Raja Complex

Within the Hara level of field depth in which the 3rd Sun Star Crystal Seal can be detected there are 2 other Star Crystal Seals of primary importance stationed along the Hara Line of Kathara Grid Level-2; these are referred to as the **Earth Seals**. The Earth Seals function as intrinsic aspects of the **Planetary Bio-feed Interface System** and the *Raja Complex*. The **8th Star Crystal Seal,** called the ***Earth Star Seal*** is located within the **12th Chakra** 6" below the feet and is associated with the **7th, 8th and 12th dimensional frequencies** of the Harmonic-3 Over-Soul Matrix (***Teuric Shield***) and the Harmonic-4 Avatar Matrix (***Maharic Shield***). The Earth Star is visible as a subtle **Gold-Orange point of diffused light** positioned along the Hara Line **6" below the feet**, as the Hara Line leaves the physical body through the 1st Base Chakra to extend downward into the Earth's core. The **10th Star Crystal Seal,** called the ***Earth Core Seal***, is located within the **13th Chakra** where the Hara Line connects into the Earth's Core and is associated with the 9th, 10th, 12th and 13th dimensional frequencies of the Harmonic-3 Teuric Shield, Harmonic-4 Maharic Shield and Harmonic-5 Rishi Matrix (***Rishic Shield***). The Earth Core Seal is visible as a **pale Blue-Black point of diffused light** positioned in the **Earth's Core,** at the point where the personal Hara Line connects to the center of the Earth's Core at the *1st and 12th Planetary Kathara Centers*.

The ***8th Earth Star Seal*** is the point at which the dimensionalizing identity combines the 12th-dimensional frequencies of the **Harmonic-4 Pre-matter Liquid Light density** with the 8th and 9th-dimensional frequencies of the **Etheric Matter densities of Harmonic-3** to connect the personal morphogenetic field and Kathara Grid to those of the Planetary Body. The **Earth Star Crystal Seal** can be viewed as the point where **Personal and Planetary Kathara Grids connect to Galactic Kathara Grid.** The Earth Star represents the point into which the **Mahara and Betcha Hova Bodies, Kathara Centers 8, 9 and 12 and the 12th Galactic Star Seal and Shara Complex** first anchor into the **Planetary and Galactic Kathara Grids.** The Earth Star holds the morphogenetic imprint for **all DNA strands for all manifest incarnations in *4 Harmonics of Time,*** corresponds to **Chakra Centers and Planetary Vortices** and **Personal and Planetary Axi-A-tonal Lines 8, 9, 10 and 12** and is a key point for repatterning the electro-tonal programs of the ***Personal* Maharic, Teuric, Doradic and Telluric Shields,** and realigning and activating corresponding DNA strands, Chakras, Planetary Vortices, Axi-A-Tonal Lines and personal or planetary body regions they govern. In the **Maharic Shield Activation process** the **Earth Star** is activated to release phase-lock on the **12th Chakra** to open the personal Kathara Grid to the PBIS.

The ***10th Earth Core Seal*** is the point at which the dimensionalizing identity combines the frequencies of the **8th Earth Star Crystal Seal** with the **13th-dimensional frequencies** of **Harmonic-5 Ante-matter density** to connect the personal and planetary morphogenetic field and Kathara Grid to those of the Universal Kathara Grid of Harmonic-5. The **Earth Core Seal** can be viewed as the point where **Personal, Planetary and Galactic Kathara Grids connect to the Universal Kathara Grid.** The Earth Core represents the point into which the **Rishi Identity**, the **Raja Hova Body** and **Universal Kathara Centers 1, 2 and 3** first anchor into the 3rd Earth Star and 12th Galactic Star Seals to form the Raja Complex of 5 Harmonic scalar fields through which personal, planetary, galactic and universal experiential manifestation takes place. The Earth Core Seal corresponds to **Chakra 13**, and holds the morphogenetic imprint for **all manifest incarnations of persons, planets and galaxies within the 5 Harmonics of Time in one 15-dimensional Time Matrix.** Through the combined frequencies of the **Earth Star and Earth Core Seals** the personal Kathara Grid connects to the Planetary, Galactic and Universal Kathara Grids forming the Planetary Bio-feed Interface System and the ***Raja Complex*** – or ***Universal* Bio-feed Interface System** – that links personal, planetary, galactic and universal consciousness in time.

THE RA CENTER

The 3rd Sun Star Seal, 8 and 10 Earth Seals and *RA Center*

When combined with **PBIS, Doradic, Teuric** and **Maharic Shields** and **Raja Complex**, the **3rd Sun Star Crystal Seal** and the **12th Core Star Crystal Seal** *Hara Center merge* to form a *Teuric-Telluric Current* that encompasses the *Seed Crystal Seal* within the **3rd Solar Plexus Chakra**, forming a ***sphere of powerful healing energies*** at the **3rd Chakra** called the **RA Center**. In **Kathara Healing Level-2** we will begin activation of the *Teuric Current* of the Harmonic-3 Betcha Hova Body, through activation of Doradic Current Phase-2 and will explore beginning applications of combining the Doradic Phase –1 and Teuric Currents with the natural *Telluric Current* of the Harmonic-1Nada Hova Body to begin bringing the RA Center for **healing and manifestation** into activation. (For more on the 4 Primary Triadic Healing Currents see Chapter 4 of this manual.)

The Hara Level of Kathara Level-2 and the 6 Primary Hara Points

Though the study of the Crystal Seals of Kathara Level-2 Crystal Seals Grid is complex in detail, one can become skilled in healing facilitation through Crystal Seals mechanics by learning the basic orientation of **the outer levels** of the **Kathara Level-2** scalar grid – ***The Hara Level.*** In simple analysis, the ***elements* of the Hara Level can be viewed as follows:**

The Hara Line – the Central Body Current, that is formed by the *11th and 12th Axiom Lines* wrapping around the fixed flash-line sequence of the *Level-1 Kathara 12-Tree Central Vertical Kathara Line.*

The 6 Primary Hara Points: Star Crystal Seals **1.** # 9-Hara Center, **2.** #6-Soul Star, **3.** #12 Galactic Star, **4.** #3-Sun Star, **5.** #8-Earth Star and **6.** #10-Earth Core. Each primary Hara Point governs the function of a **complex** of interwoven scalar, bio-energetic and physical systems. Knowledge of the **6 Hara Points** and the **Thalamus, Thymus, Shara, Nadial** and **Raja Complexes** they respectively govern, allows the healing facilitator a greater understanding of the 15-dimensional anatomy. Such **Foundations of Kathara Healing** are necessary in preparation for **Kathara Healing Level-2** practices that will utilize the **Hara Level** of the **Kathara Level-2 Crystal Seals Grid** in combination with the Kathara Healing Level-1 **Maharic Recoding Process** and other elements to begin activation of the **RA Center** for advanced healing and manifestation.

Axi-A-Tonal Lines 1-10

In our exploration of the Hara Level of the Crystal Seals Grid we have learned that of the **12 Primary Axi-A-Tonal Lines,** (12 primary vertical energy-flow lines within the body and bio-energetic field, created by cross-over points of the rotating, single axis, flash-line sequences projected by each of 12 Seed Crystal Seals), the 11th and 12th Axi-A-Tonal Lines spiral around the Kathara Level-1 12-Tree Grid *Central Vertical Kathara Line* to form the Hara Line and Central Body Current. Axi-A-Tonal Lines 1-10 also run vertically through the body, from the head to the toes and through each arm to the finger tips, connecting body regions to corresponding levels of the Auric Field, Chakras, Kathara Centers, Crystal Seals and DNA. The sequence of Axi-A-tonal Line position is as follows:

Right Side of Body to Center- 3,6,9,2,8 **Center of Body-11, 12 Center to Left Side of Body-.** 5,1,10,7,4

Each line of Axi-A-Tonal Lines **1-10** ends in one **finger-tip** and **toe** on the corresponding side of the body, and runs vertically through the top of the head to merge with Axi-A-Tonal Lines 11 and 12 and the Hara Line at **Chakra 10**, and down through the feet to merge with the Hara Line at **Chakra 12.**

The HARA LEVEL- Outer Portion of Kathara Level-2 Crystal Seals Grid

6 Primary Hara Points
(Star Crystal Seals on Hara Level)

1. **Hara Center (Core Star)**
   (9th Silver Star Crystal Seal)
   2" below Navel
2. **Soul Star**
   (6th Indigo Star Crystal Seal)
   Top of Breast Bone at Thymus
3. **Galactic Star**
   (12th White Star Crystal Seal)
   Cone at 14th Chakra 36" above head
4. **Sun Star**
   (3rd Yellow Star Crystal Seal)
   Just below Hara Center
5. **Earth Star**
   (8th Gold Star Crystal Seal)
   In 12th Chakra 6" below feet
6. **Earth Core**
   (10th Blue-Black Star Crystal Seal)
   In 13th Chakra at Earth's Core

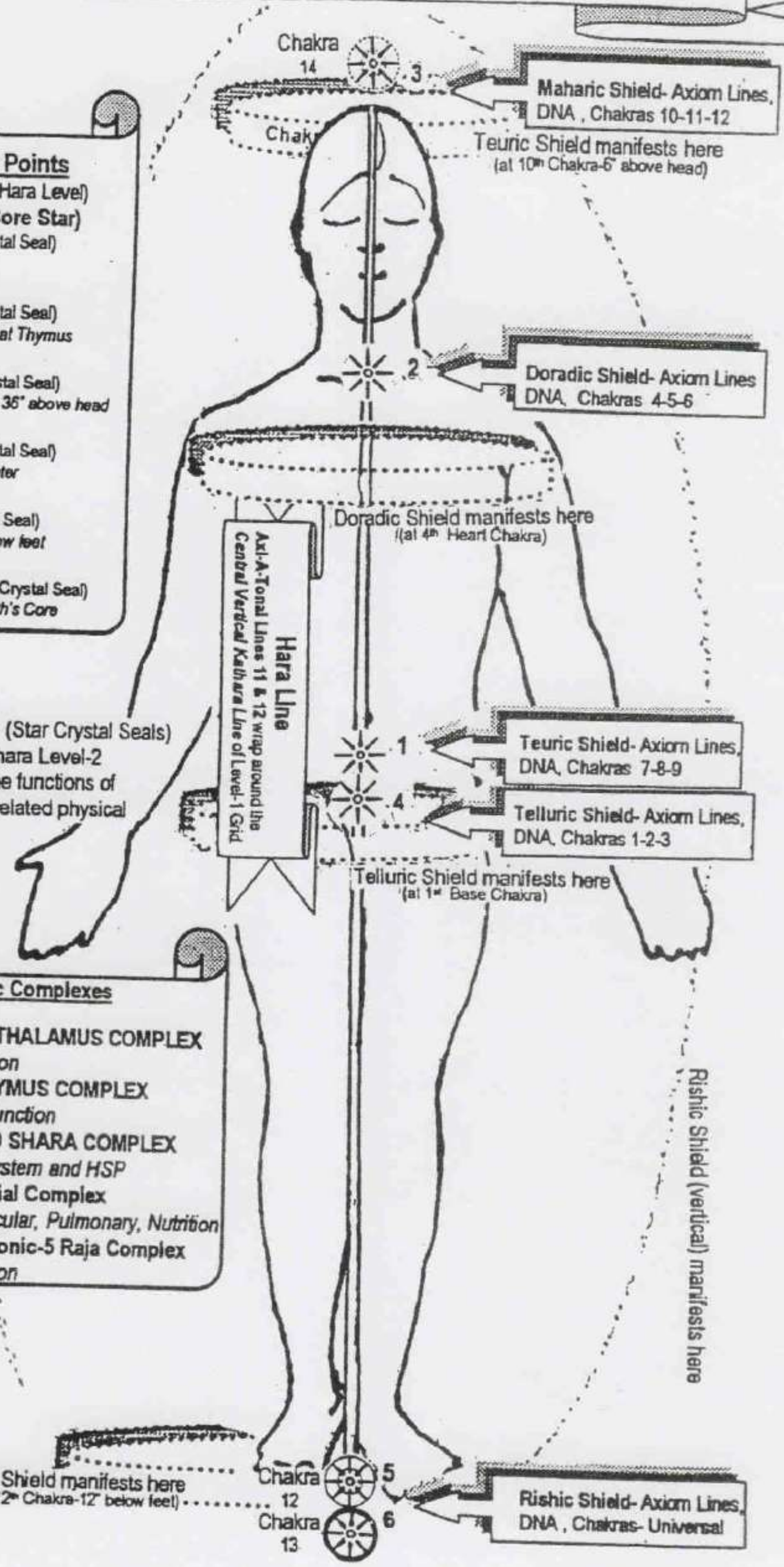
Chakra 14
3
Maharic Shield- Axiom Lines, DNA , Chakras 10-11-12
Chak
Teuric Shield manifests here
(at 10th Chakra-6" above head)
2
Doradic Shield- Axiom Lines DNA, Chakras 4-5-6
Doradic Shield manifests here
(at 4th Heart Chakra)
Hara Line
Axi-A-Tonal Lines 11 & 12 wrap around the Central Vertical Kathara Line of Level-1 Grid
1
Teuric Shield- Axiom Lines, DNA, Chakras 7-8-9
4
Telluric Shield- Axiom Lines, DNA, Chakras 1-2-3
Telluric Shield manifests here
(at 1st Base Chakra)
Rishic Shield (vertical) manifests here
Maharic Shield manifests here
(6" below 12th Chakra-12" below feet)
Chakra 12
5
Chakra 13
6
Rishic Shield- Axiom Lines, DNA , Chakras- Universal

The **6 Primary Hara Points** (Star Crystal Seals) of the Hara Level in the Kathara Level-2 Crystal Seals Grid govern the functions of corresponding Shields and related physical systems.

Related Bio-energetic Complexes

1. Hara Center – **D-9 THALAMUS COMPLEX**
   Focus: *Brain Function*
2. Soul Star – **D-8 THYMUS COMPLEX**
   Focus: *Metabolic Function*
3. Galactic Star – **D-10 SHARA COMPLEX**
   Focus: *Immune System and HSP*
4. Sun Star – **D-4 Nadial Complex**
   Focus: *Cardio-Vascular, Pulmonary, Nutrition*
5. Earth Seals – **Harmonic-5 Raja Complex**
   Focus: *Transmutation*

12 Primary Axi-A-Tonal Lines
of the Kathara Level-2 Crystal Seals Grid

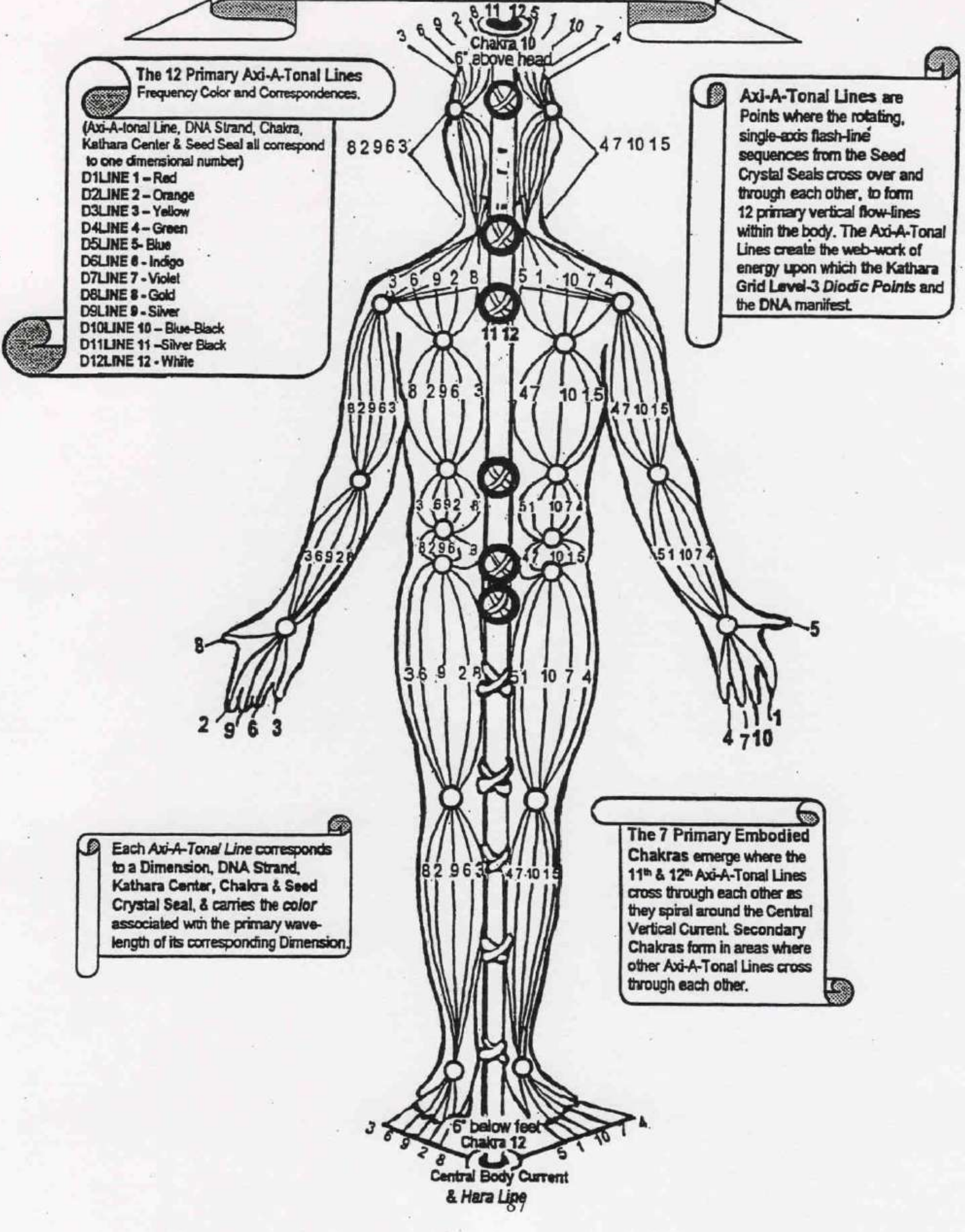

The 12 Primary Axi-A-Tonal Lines
Frequency Color and Correspondences.

(Axi-A-tonal Line, DNA Strand, Chakra, Kathara Center & Seed Seal all correspond to one dimensional number)

D1LINE 1 – Red
D2LINE 2 – Orange
D3LINE 3 – Yellow
D4LINE 4 – Green
D5LINE 5- Blue
D6LINE 6 - Indigo
D7LINE 7 - Violet
D8LINE 8 - Gold
D9LINE 9 - Silver
D10LINE 10 – Blue-Black
D11LINE 11 –Silver Black
D12LINE 12 - White

Axi-A-Tonal Lines are Points where the rotating, single-axis flash-line sequences from the Seed Crystal Seals cross over and through each other, to form 12 primary vertical flow-lines within the body. The Axi-A-Tonal Lines create the web-work of energy upon which the Kathara Grid Level-3 ***Diodic Points*** and the DNA manifest.

Each *Axi-A-Tonal Line* corresponds to a **Dimension, DNA Strand, Kathara Center, Chakra & Seed Crystal Seal**, & carries the *color* associated with the primary wavelength of its corresponding Dimension.

The 7 Primary Embodied Chakras emerge where the 11th & 12th Axi-A-Tonal Lines cross through each other as they spiral around the Central Vertical Current. Secondary Chakras form in areas where other Axi-A-Tonal Lines cross through each other.

3 6 9 2 8 11 12 5 1 10 7 4
Chakra 10
6" above head
8 2 9 6 3
4 7 10 1 5
3 6 9 2 8
5 1 10 7 4
11 12
8 2 9 6 3
4 7 10 1 5
3 6 9 2 8
5 1 10 7 4
3 6 9 2 8
5 1 10 7 4
8
2 9 6 3
5
4 7 10 1
3 6 9 2 8
5 1 10 7 4
8 2 9 6 3
4 7 10 1 5
3 6 9 2 8
5 1 10 7 4
6" below feet
Chakra 12
Central Body Current
& Hara Line

Axi-A-Tonal Lines and the Meridians

The Axi-A-Tonal Lines within the body set the primary holographic scalar field upon which the Meridian Lines – or Secondary Flow Lines- of the body manifest. **The Meridian Lines can be used to access the Axi-A-Tonal Lines and the Axi-A-Tonal Lines can be used to access the flash-line programs of the Kathara Centers and 3 Levels of the Kathara Grid, if the *correspondences* between *the Meridians, Axi-A-Tonal Lines, Kathara Centers* and *Chakras* are understood.** For centuries the **Meridian Lines** and **Chakras** have been used in subtle-energy healing systems. This knowledge can be applied and the potential results of related healing techniques amplified by employing the correspondences between the Meridian and Chakra systems and the Axi-A-Tonal Lines and 3 Levels of the Kathara Grid. Bringing the Kathara Grid into the picture allows for restructuring the **Imprint for Health** within the *causal core* while *expediting* the healing process by restructuring the various other levels of the bio-energetic anatomy.

In Level-1 Kathara Healing we focus primarily upon **Core Morphogenetic Repatterning** – resetting the electro-tonal programs of the Kathara Grid through the uncorrupted programs of the ***Maharic Shield***- to set the "*ground*" into which other healing applications can *anchor*. In later Kathara Healing Levels we will explore technique applications that affect various other levels of the bio-energetic anatomy. **In Level-1 Kathara Healing we strive to create a basic familiarity with the elements of, and the interrelationships between, the various levels of bio-energetic anatomy.** Through understanding the basic correspondences between the **3 Levels of the Kathara Grid** – the *Kathara Centers* and *Kathara Lines*, the *Hova Bodies*, Auric Levels, *Signets, Shields, Crystal Seals* and *Axiom Lines*, and the *Diodic Points* and *Chakras* – and the other aspects of energetic anatomy, such as the **Meridian Lines** and **DNA**, the healing facilitator is equipped with a knowledge base through which more precise healing facilitation can be applied. Meridian Lines and their applications to Kathara Healing will be explored in later Levels of the Kathara Healing Program.

DNA

The DNA Connection

In this Chapter we have explored the various elements of the 15-Dimensional Anatomy of the human form, from the Kathara Grid holographic scalar template, to the Hova Bodies, Shields, Axiom Lines, Chakra Centers, Auric Field Levels and the Hara Level with its related Complexes of energetic interface. Throughout all of these elements of the multi-dimensional anatomy we have seen how each element interrelates with the others through their *correspondences within the structure of the 15-Dimensional Scale.* The elements of the subtle body anatomy also systematically correspond to the structure and function of the human **DNA.** Through direct interaction with various elements of the subtle body anatomy, the operations of corresponding aspects of the human DNA can be repaired, enhanced and altered, to advance applications of healing, and to expedite the evolution of human biology and consciousness. In the *Kathara Healing Level-2 program* we will further explore the correspondences between the subtle body anatomy and the DNA, and methods by which the DNA can be repaired and activated to a more advanced level through employing the use of Kathara Healing mechanics.

The Silicate Matrix 12-strand DNA Pattern

The core human DNA imprint is built upon a **12-Dimensional** Structure, each of 12 strands building up upon one of the Primary 12 Axi-A-Tonal Lines and each strand corresponding to 1 of 12 dimensional frequency bands within the 15-dimensional scale. The original human DNA imprint is called the **Silicate Matrix**, the 12-strand scalar-wave configuration that allows for the progressive **transmutation** of Harmonic-1 carbon-based molecular-elemental structure into the Silica-based structures of higher Harmonic matter. The 12-strand DNA imprint of the human form implies that human biology was designed to embody **12 dimensions of conscious awareness**, or to fully embody the Avatar identity level - the ***"Christed" Identity*** of Harmonic-4 Pre-matter Liquid Light matter density. The human form is organically designed to undergo **full cellular transmutation** through 4 Harmonics of manifestation and their corresponding matter density levels. When the gene code is working properly, the process of cellular transmutation is directed by the embodied human consciousness. In historical terms, the process of ***conscious cellular transmutation*** has been referred to as the process of **"Ascension"**, which is ***Dimensional Expansion***, the mechanics of which have been hidden with ancient texts dealing with **Merkaba Mechanics.** The **Merkaba Vehicle** of inter-harmonic scalar-wave configurations, which forms within the bio-energetic field, is a consequence or *simultaneous affect* of progressive activation of the Silicate Matrix within the genetic code.

The Silicate Matrix and Subtle Body Anatomy

Just as each Axi-A-Tonal Line of the subtle body anatomy has its correspondence to a Kathara Center, Crystal Seal, Axi-A-Tonal Line, Auric Field Level and Chakra, so too do the DNA strands have their specific affiliations to each *dimensionalized element* of the subtle body anatomy. Each DNA strand translates into electrical-elemental-chemical form from the governing Axi-A-Tonal Line flash-line sequence to which it corresponds. The **Hova Bodies** and **Shields** of subtle body anatomy each correspond to a *set of 3* Axi-A-Tonal Lines, Auric Field Levels and Chakras and thus also correspond to a set of **3 DNA strands**, that together create the experiential projection of separate 3-Dimensional holographic reality fields. Distortions and blockages in the Kathara Grid, Hova Bodies, Shields, Axi-A-Tonal Lines and Chakras appear as distortions and malfunction within the DNA. **The function of the DNA determines the *quantity of interdimensional frequency*, and thus the *amount of dimensionalized consciousness*, that can actively process through and embody within the physical body system.** DNA distortions manifest as distortions of biology and consciousness.

The 12 DNA strands of the Silicate Matrix are *sets of scalar-wave spectra* that correspond to each dimensional band, through which each dimensionalized level of the subtle body anatomy is translated into wave-spectra that is within the range of perceivable matter manifestation. *The DNA can be viewed as the primary structure through which subtle energy-consciousness translates into manifest physicality.* As each full dimensional band contains 12 primary Sub-frequency Bands (See Chapter 1 for Dimensional Structure), each of the 12 DNA strands contains *12 primary* **Seed Codes,** or *consolidated scalar points* composed of electro-tonal flash-line sequences, that correspond to the Sub-Frequency Bands within each dimension. The 12 primary Seed Codes within each DNA strand have an *anti-particle counterpart* composed of frequencies from the Anti-particle system. The 12 particle Seed Codes in each strand are called **BASE CODES**, their anti-particle counterparts are called **ACCELERATION CODES**, The Seed Codes of the core DNA scalar imprint are the *points of consolidated frequency* through which energy-consciousness is drawn in from the dimensional Unified Fields to expand the personal morphogenetic field.

DNA Fire Codes- Genetic Time Codes

Along with the **12 Base Codes** and **12 Acceleration Codes** of each DNA Strand Template there are also 12 Primary DNA Codes called **DNA FIRE CODES** that regulate the relationship between human biology and the **Planetary Time Cycle** in which the biology manifests. The DNA Fire Codes can be conceptualized as existing *between* the Primary 12 Strands of the Silicate Matrix. Each Fire Code is composed of one half of the scalar frequency patterns from each of the two Primary Strands that the Fire Code is placed between. Each of the 12 Fire Codes corresponds directly to one of 12 **Signet Star Crystal Seals** in the Personal Shields of the Internal Templar Complex, and with release of the corresponding **Signet Seal**, the Fire Code comes into activation.

The DNA Fire Codes release the natural scalar barriers between particle and anti-particle DNA Strands, allowing **STRAND FUSION** to occur. Fusion of DNA strands in the particle body with corresponding strands in the anti-particle body creates a progressive transmutation of the scalar wave forms of the DNA Template and Scalar Shields, through which the polarized Particum-**Particles** and Partika-**Anti-particles** merge and transform into their original Omni-polar Partiki **Ante-matter** state. Through strand fusion, the biological carbon-based biology progressively **lessens in density** as its elemental agents are transmuted from the Kathara level, into elemental agents characteristic of the higher dimensional matter densities. **Dimensional Ascension** is achieved through release of the Kathara Grid Seals. Release of the Kathara Grid Seals creates merger of the Hova Bodies, activation of the Kundalini energies, release of the Cranial-Sacral Crystal Seals, **fusion of the DNA Strand Templates via the DNA Fire Codes**, building of the Merkaba Vehicle in the bio-energetic field and the resulting transmutation of bodily matter and shift in Angular Rotation of Particle Spin within molecular construction. Through this transformational process the human biology is released from its organic phase-lock within its current space-time cycle, and is able to engage in **transcendental travel** through the various matter density levels and Time Cycles in and beyond the 15-Dimensional Time Matrix. **As the DNA Fire Codes regulate the body's orientation to space-time, they are also referred to as the GENETIC TIME CODES.**

The DNA Fire Codes are an essential element in the natural process of restoring the natural immortal **Organic Imprint for Health** within the human body. The ability of the human form to become **perpetually self-sustaining** is dependent upon the Kathara Grid and DNA being able to continually renew the natural Primal Order of Maharic Shield scalar arrangement upon which the body is built, to keep the Star Crystal Seals and conduits of interdimensional energy supply open. **Perpetual revitalization of the body** is achieved through full activation of the **12-Strand DNA** pattern. When activated, the **DNA Fire Codes** keep the Crystal Seals, **the interdimensional INTERNAL STAR GATES of energy inflow**, open, so the finite supply of frequency held within the personal Manifestation Template can be continually renewed. Restoration of the Organic Imprint for Health, as held within the **Maharic Shield**, is achieved through progressively clearing distortions of scalar-wave arrangements within the Kathara Grid, Shields and DNA Template and realigning the entire subtle energy and physical energy systems with the Primal Order of the Maharic Shield. This process of whole-being-healing does not occur overnight, as many genetic distortions (which emerge from Kathara Grid misalignment) have served as the dominant bio-electrical program within the human genome for many thousands of years. With knowledgeable practice of Kathara Grid realignment, and supportive healing applications within the various Body-Mind-Spirit systems, these inherited distortions can be progressively healed through Regenesis of the Primal Order as held within the personal Maharic Shield.

Science, "Junk DNA" and Evolution

In the contemporary human genetic code the majority of the.DNA strands corresponding to higher dimensional anatomy and consciousness are dormant or damaged due to ancient blockages within the Kathara Grid and its corresponding elements of the subtle body anatomy. Present day humans function primarily on a **3 to 3.5 DNA activation level**, while the imprint and potentialities for strands 4-12 lie dormant within the morphogenetic field and within the fragments of non-functional DNA that contemporary science has labeled ***"junk DNA"***.

The portions of the DNA chain that science has presently identified as the **"Double Helix"**, represent only the **surface portions** of the chemical, elemental and electrical components of the active DNA strands. Science has yet to identify the *multidimensional spectra* of DNA manifestation, and has yet to realize that within the structures of detectable DNA, there are *levels of structure and function* that direct the operations of the *entire genetic imprint*, which are *not presently detectable* by contemporary scientific technique. The human DNA imprint will always appear from external analysis, as a *2-strand Double Helix* configuration, but what is not understood, is that within this Double Helix there are, and will be progressively more, **additional Double Helix strands**, which *fuse together* and add their operational coding into the active DNA imprint. By contemporary standards of categorization, if science were to objectively observe the activated 12-strand pattern (or even an activated 4-strand pattern), the DNA would *not appear to be human at all*, but rather as the seemingly incomprehensible genetic signature of an "unknown" species. As science evolves into comprehension of the multi-dimensional spectrum, the understanding of the true structure and function of DNA will progressively advance.

DNA, Kathara Healing and Bio-Spiritual Evolution

One does not have to wait for scientific theology to catch up with a rapidly evolving humanity in order to begin advancing the functions of the DNA. Through the processes of Kathara Healing, as we work to clear blockages from the Kathara Grid and to reawaken the *Imprint for Health* within the human body, we are *simultaneously, directly affecting* the operational structure of the DNA. As we work to create **Hova Body merger**- *merging of the scalar-grid flash-line sequences of the 5 Harmonic Hova Bodies*- we are progressively reordering portions of the "junk DNA" and stimulating dormant portions of the DNA into activation. As we work with subtle energies to clear blockages within the Kathara Grid, we are progressively purging ancient distortions and mutations within the gene code, distortions that have kept our collective consciousness locked into the limited 3-dimensional perceptual range of Harmonic Universe-1.

It is through the ancient distortions of the human Kathara Grid and the resulting **mutations** of the human DNA, that the multidimensional portions of identity seem separate from each other. The higher levels of conscious awareness that are stationed within the scalar grids of the higher dimensional Hova Body Shields have been unable to *electronically translate* into the lower dimensional identity stations of the Nada Hova Body and its apparent physical manifestation. As we work with Kathara Healing to create progressive Hova Body merger and DNA strand ordering and activation, we will progressively experience *advancement of conscious awareness* and *expansion of the perceptual field* and *Higher Sensory Perception* (HSP). This expansion and advancement occurs because we are able to progressively embody higher frequencies of consciousness and scalar wave flash-line sequences into the personal morphogenetic field, via the energetic conduits of the DNA. **Kathara Healing modalities are the fastest and easiest methods available to repair and advance the function of the DNA and to expedite the organic processes of *Bio-Spiritual Evolution*.**

The Silicate Matrix 12-Strand DNA Template with Hova Body, Scalar Shield and Identity Level Correspondences

The Morphogenetic Field Scalar-wave Template for the Original Human Genome

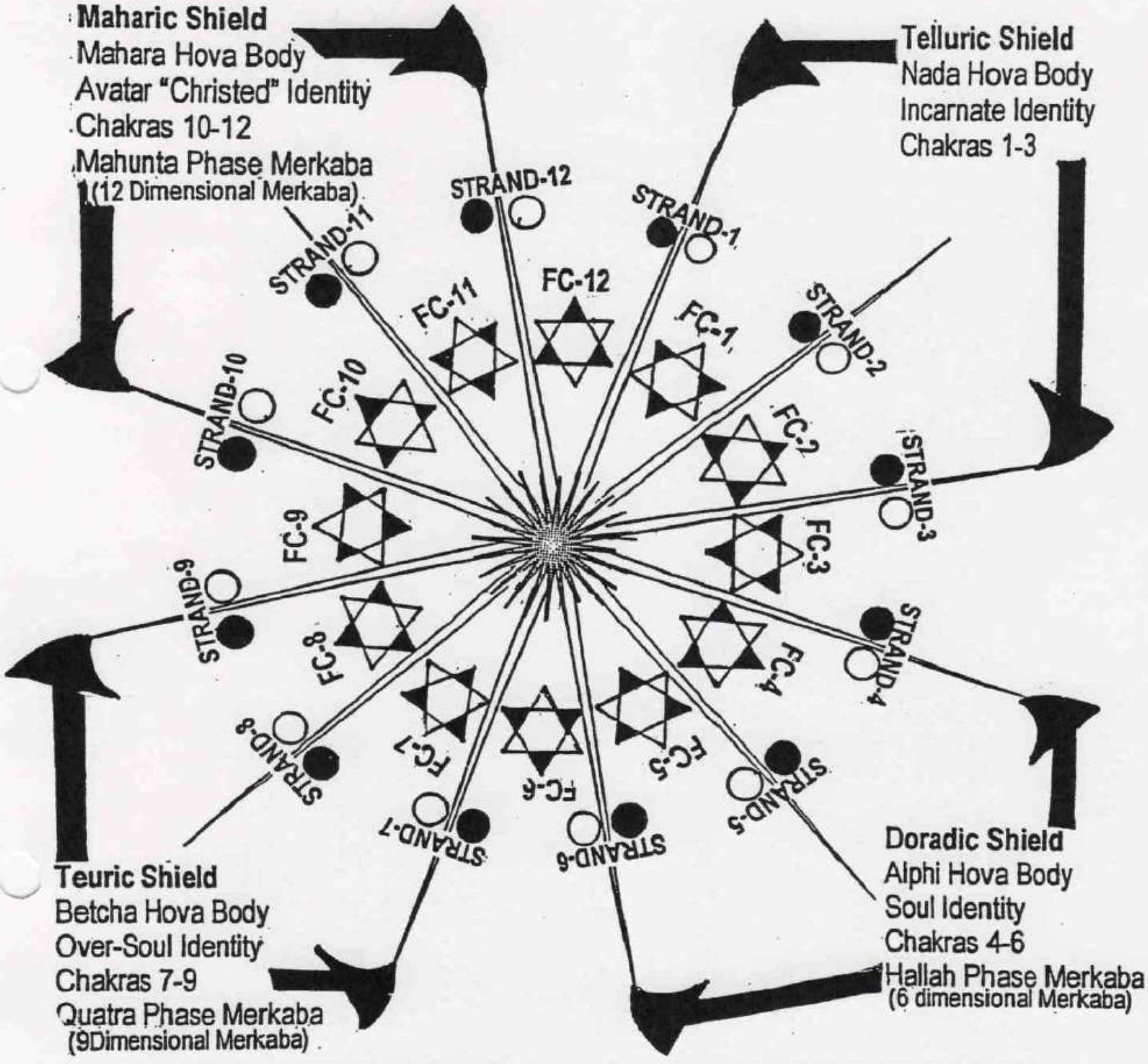
Maharic Shield
Mahara Hova Body
Avatar "Christed" Identity
Chakras 10-12
Mahunta Phase Merkaba
(12 Dimensional Merkaba)
Telluric Shield
Nada Hova Body
Incarnate Identity
Chakras 1-3
STRAND-12
STRAND-1
STRAND-11
STRAND-2
STRAND-10
STRAND-3
STRAND-9
STRAND-4
STRAND-8
STRAND-5
STRAND-7
STRAND-6
FC-12
FC-1
FC-11
FC-2
FC-10
FC-3
FC-9
FC-4
FC-8
FC-5
FC-7
FC-6
Doradic Shield
Alphi Hova Body
Soul Identity
Chakras 4-6
Hallah Phase Merkaba
(6 dimensional Merkaba)
Teuric Shield
Betcha Hova Body
Over-Soul Identity
Chakras 7-9
Quatra Phase Merkaba
(9Dimensional Merkaba)

12 Base Codes per Strand = ● **12 Acceleration Codes per Strand = ○**

12 FIRE CODES between Strands = ✡

(Diagram presents simple conceptualization of DNA Strand Template orientation, not actual geometrical arrangement of wave-form structure and observable strand interrelationship.)

Merkaba Phases – Building the Merkaba Vehicle

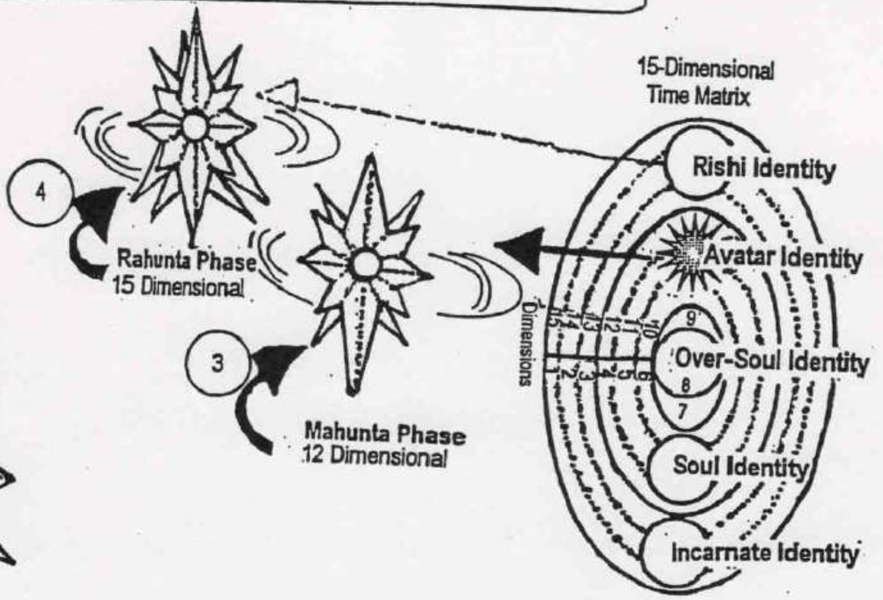
15-Dimensional
Time Matrix
Rishi Identity
Avatar Identity
Dimensions
15 14 13 12 11 10
9
Over-Soul Identity
1 2 3 4 5 6
8
7
Soul Identity
Incarnate Identity
4
Rahunta Phase
15 Dimensional
3
Mahunta Phase
12 Dimensional

Formation of the **Mahunta Merkaba** marks fulfillment of the 12-Strand DNA potential.

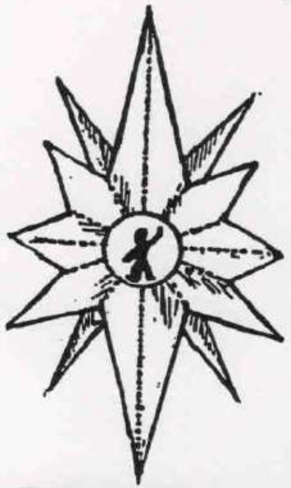

The 4 Personal Scalar Shields merge to form the 12-Dimensional **Mahunta Merkaba**

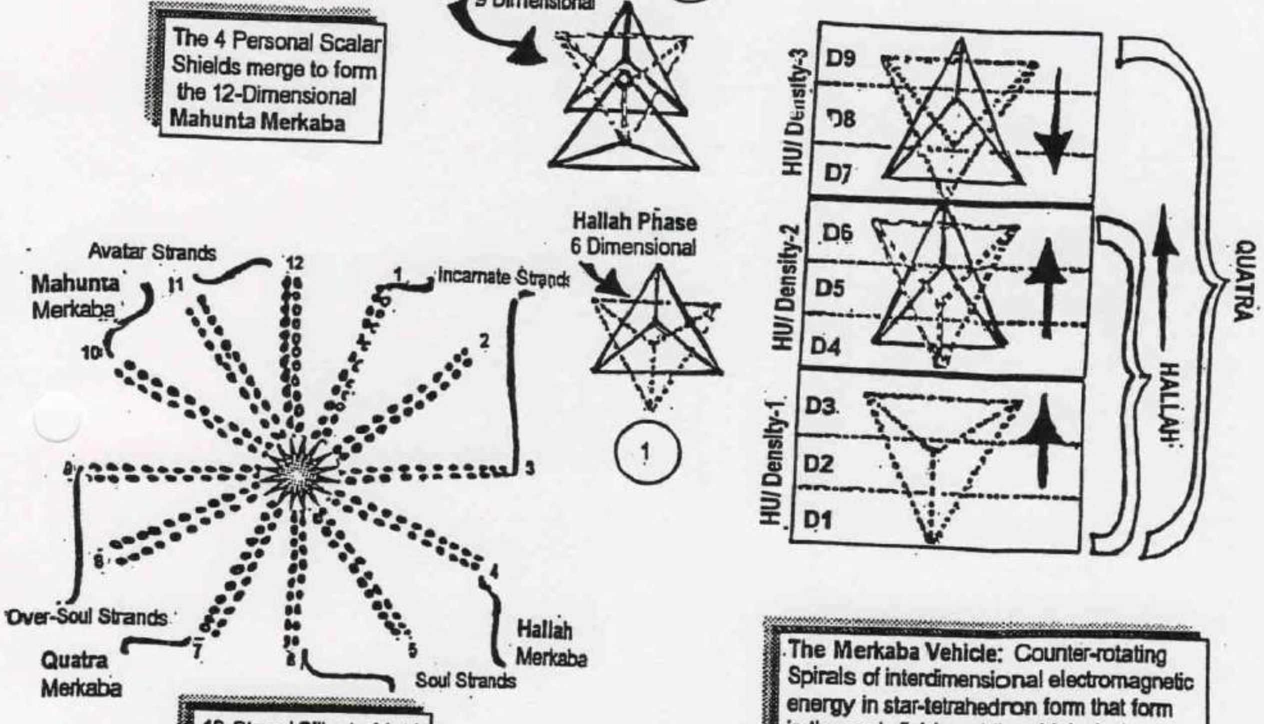
Quatra Phase
9 Dimensional
2
Hallah Phase
6 Dimensional
1
HU/Density-3
D9
D8
D7
HU/Density-2
D6
D5
D4
HU/Density-1
D3
D2
D1
QUATRA
HALLAH
Avatar Strands
Mahunta Merkaba
12
11
10
1
Incarnate Strands
2
3
4
Hallah Merkaba
5
Soul Strands
6
7
Quatra Merkaba
8
9
Over-Soul Strands

12-Strand Silicate Matrix DNA Diagram shows Strand relationships

The Merkaba Vehicle: Counter-rotating Spirals of interdimensional electromagnetic energy in star-tetrahedron form that form in the auric field, enabling biological interdimensional travel and ascension.

THE 12 HUMAN SENSES, APPARTHI AND HSP

Perception and the Hologram

In earlier text we have explained that **manifestation is not what it appears to be**, but rather a **Holographic Projection** of consciousness through the structures of **dimensionalization**. Further, that dimensionalization itself is a *compartmentalization* of consciousness created through units of consciousness (Partiki, Partika and Particum) *assembling in specific mathematical-geometrical ratios* to form **scalar grids**, interwoven fabrics of *fixed points of electro-tonal units* that form layers of standing-waves -**scalar waves** - that serve to "flesh out" the structure of the Hologram, through projection along a *vertical, horizontal* and *diagonal axis*. We have explored the **Kathara Grid** - the *core template* of scalar standing-waves upon which consciousness begins its experience of dimensionalization, the structures of the **15-Dimensional Universe** construction, and how these characteristics of the Hologram relate to forming the structure of **15-Dimensional Human Anatomy**. In understanding the basic foundations upon which our perceivable Hologram forms, we begin to better comprehend the nature of reality and of ourselves within this multi-dimensional reality structure. **Through the Holographic Projection of consciousness within the cosmic Unified Field of Energy-consciousness and its scalar fields of dimensionalization, we perceive the conditions of space, time, matter, externalization of reality and individuation of identity.** Within the Hologram of consciousness we perceive as **solid and objectified** the **wave spectra** and their inherent energy constructs of that which exists **precisely one dimensional band below the present station of our consciousness focus.**

In this section we will explore a bit about the **apparatus of perception** that allow us to perceive externalization in such a manner, the **facilities of energy inherent to the scalar-wave construct of the human form,** which together work to bring the Hologram into perceivable manifestation. We will call the constructs of energy formed through scalar organization, which allow for perception of the manifest Hologram, the **HUMAN SENSES.**

THE *HUMAN SENSES* ARE THE *EXPERIENTIAL RESULT* OF *ORGANIZATIONS OF WAVE PATTERNS* INHERENT TO THE ELECTRO-TONAL PROGRAM AND SCALAR ORGANIZATION OF HUMAN CONSTRUCTION.

What ARE Senses?

Senses are the **scalar-wave apparatus**, and their seemingly manifest parts, that allow the personal morphogenetic field to **synthesize and translate frequency** from the dimensional bands of the Unified Field, into **coherent perceptual and experiential qualities** that give embodied consciousness a **translation of its relationship to other forms of consciousness within the Unified Field.** Senses - or the apparatus which give us the *ABILITY TO SENSE*- are scalar arrangements inherent to our form that **TRANSLATE ENERGY SIGNATURES FROM THE UNIFIED FIELD INTO USABLE DATA OF RELATIONSHIP.**

Like the true substance of the human body and consciousness, the Unified Field exists as a *quantity of dimensionalized energy-consciousness in electro-tonal scalar wave form* - a compartmentalized field of **ENERGY SIGNATURES**, within which a vast number of **consciousness constructs** exist The human organism is one such **construct of consciousness**, and through its inherent form, the *energy signatures* of other scalar forms are translated into perceptual data. **The senses are simply *energy constructs* within the human morphogenetic field, that *separate and translate energy signatures from the Unified Field* into sets of experiential stimuli, through which the contours of the manifest Hologram can be *recognized* and *experienced.***

The APPARTHI and Perception

The sense facilities are *organizations of energy* within the morphogenetic field that hold sets of "*energy receivers" or magnetic points of cohesion*, into which *like frequency patterns* from the Unified Field are drawn in for translation. Every organism has a different configuration of energy receivers within its morphogenetic makeup, and so every organism will have a variation of perception. The **morphogenetic energy receivers**, or **APPARTHI** of an organism's construction serve to *draw certain frequency bands into the morphogenetic field for translation and to block out other frequency bands*, making them unavailable for translation. The frequency bands, or *flash-lines of Partiki Phasing*, that are blocked out of the personal morphogenetic field, create **GAPS OF PERCEPTION** of the Unified Field, through which the Holographic picture of "space between objects", "externalization of form" and "separation of SELF from the Unified Field" becomes perceivable to the organism. **These qualities of *objectification* are the result of BLOCKED FLASH-LINE SEQUENCES, portions of the *scalar fields* of the Unified Field and their inherent *sequences of Partiki Phasing*, that *cannot translate* through the personal morphogenetic field due to the arrangement of the APPARTHI within the personal morphogenetic structure.**

Senses as Scalar-wave Constructs & Expanding our Perception of Perception

If it can be recognized that externalization of experience is a perceptual illusion of consciousness created through selectivity of range within the perceptual field, the true nature of the human sense facilities can be better understood as a pattern of interrelated scalar-wave forms that serve to block out portions of the Unified Field so that other portions may be brought more fully into perceivable view. **When viewing the *physical apparatus* through which the known human senses appear to occur as part a of larger, interwoven system of *scalar construction*, it will be easier to understand that *perception is not a localized phenomena created by and limited to the identified organs of perception*.** The experiential capacities of even the 5 known human senses have a much broader range than what is presently assumed or identifiable through analysis of the manifest sense organs. *If one can understand that perception of manifestation itself is an affect of these greater scalar wave constructs, it will not seem so outrageous to consider the reality and integrity of sensing and perception beyond the range of the presently identified sense facilities.*

ESP, HSP and Evolution

"Extra-sensory perception" is not "EXTRA" at all, it is a *normal attribute* of the *evolving human organism*, as the consciousness ensconced within the *body hologram* continues its *natural expansion* into the greater portions of its evolutionary imprint. Individuals displaying abilities of "Extra-sensory Perception" are not gifted with something *other than what is characteristic to the human form*, **they are simply *a bit ahead in the stage of development* of the natural human perceptual capacities. Everyone has dormant potentials for some degree of "ESP"; like all things within the human organism, ESP is regulated by the *function of the DNA* and its naturally intended *relationship* to the *Kathara Grid and morphogenetic structure*.** Individuals with **less distortions and blockages** within the **Kathara Grid** (and the resulting DNA malfunctions such distortions create) will have *greater ease* in calling the dormant perceptual facilities into activation. *Individuals who consciously attempt to use the greater perceptual facilities will stimulate a more rapid advancement into reclaiming this stage of evolution, than those who refuse to acknowledge the existence of their greater sensing facilities.* **The greater sense facilities bring the *higher dimensional wave spectra* and its inherent reality fields into the *perceptual range* of the earthly embodied human consciousness, and so the greater sense facilities can be considered the HIGHER SENSES, a term far more accurately describing their orientation than the term *Extra Senses*. "ESP" becomes "HSP"- *Higher* Sensory Perception - a *natural attribute of experience for the evolving human organism.***

THE 5 KNOWN SENSES

The 5 Known Human Senses

In order to understand the nature of the **Higher Senses** it is necessary to understand a bit more about the nature of the **common senses**, or the "*lower senses*" if you will, for the **common senses** are those that are "tuned to" perception of the **lower frequency dimensions. Perception of *Sight, Sound, Touch, Taste and Smell* are *experiential affects* of the *scalar-wave constructs* that exist within the NADA HOVA BODY and the TELLURIC SHIELD, the Kathara Grid and morphogenetic field structures that are composed of and stationed within the frequency bands of dimensions 1, 2 and 3, the first Harmonic of manifestation.** The **APPARTHI,** or ***"energy receivers"*** within the Nada Hova body have their first imprint within the **Kathara Grid Level-2: Crystal Seals** structure. The first pattern for the sense facilities responsible for manifesting the organs of perception of the 5 identified senses exist within the configurations of electro-tonal Partiki units that form the crystalline seals of which the dimension-2 **Telluric Shield** is constructed. Through the Partiki Phasing flash-line projections of the Telluric Shield, the "armature" or 3-dimensional blueprint for the organs of perception are set for manifestation within the body hologram.

The 5 Known Senses and Sound Transduction

The 5 known senses are more intimately related than commonly supposed; ALL represent a form of SOUND TRANSDUCTION - the frequencies, or flash-line sequences, within the 1st, 2nd and 3rd dimensional bands, that form the identifiable wave patterns presently associated with sound waves, are drawn into the Level-2 Kathara Grid Crystal Seals by **Apparthi receivers** that carry like flash rhythms. Once drawn into the Apparthi receivers the sound spectrum becomes modulated into different flash-line rhythms and sent into the **Diodic Points of the Level-3 Kathara Grid.** Once intersecting with the frequencies within the Diodic Points the frequencies from the Apparthi are then **divided/ polarized** and projected up and down the Harmonic-1 dimensional scale, **creating a pattern of energy refraction** within the morphogenetic field and physical body apparatus. The polarized frequencies first appear within the **DNA** as part of the core electrical vibration, and then further transmute into **electrical impulse** that manifests as configurations with the **blood chemistry** and within the **Central Nervous System.** These synthesized electrical impulses are further projected through the **scalar-wave grids** that hold the form of the **identified sensory organs,** finally **culminating in the interrelated qualities of perception associated with the identified sensory organs and their 5 known attributes of sensing.**

THE 12 HUMAN SENSES

The Interwoven Senses and the Hova Bodies

As in the case of the Lower Senses, the *Higher Senses* and their experiential manifestations are the product of consciousness interacting with the *dimensionalized scalar organization* of the human morphogenetic anatomy. Whereas the Lower Senses are affiliated with the transduction of sound frequency within the *1st, 2nd and 3rd Dimensions*, through the structures of the *Nada Hova Body* and the *Telluric Shield* of the *first Harmonic*, the Higher Senses are in like fashion associated with the scalar organization of the *Higher Dimensional Hova Bodies* and their corresponding *Shields*.

In truth, all of the sensing facilities organic to the human form comprise interwoven and interrelated systems of *energy refraction* that together create the perceptual experience of *externalization* within the hologram of manifestation. For example, it is not commonly recognized that the sense of **smell** is in actuality part of the function of the **auditory facility**, and thus represents an intrinsic part of the sense of hearing. It is also not commonly understood that the processes associated with **vision** are not only connected to the processes of the brain and Central Nervous System, but also with the formation of *crystalline structures* with the plasma of the **blood. The senses are not nearly as *separate* and *independently functional* as they appear to be when viewed from outside analysis.** They are rather a part of an **interwoven system of electro-tonal information relay** that spans the 15-Dimensional morphogenetic field anatomy of the human organism.

The 12 Human Senses

In terms of breaking down the sense facilities for structural analysis, such as is done by modern science in identifying the 5 Known Senses, we will describe the sensory attributes of the **Higher Sense facilities** in terms of their *primary orientation*, *location of their primary Apparthi receivers* within the *Hova Body* structures and by the type of *perceptual or experiential attributes* produced through their function. **In such categorization of the human sense facilities there are *7 additional*, as yet unidentified, *sense facilities* operating within and through the human consciousness**, which are collectively responsible for translating the scalar fields of dimensionalized reality into the *illusionary holographic picture* of externalized manifestation and its qualities of space, time, matter and identity individuation. **Collectively there *are 12 Primary Sensing Facilities* within 15-Dimensional Human Anatomy.**

PERCEPTUAL BLOCKAGE, KATHARA GRID DISTORTION AND KATHARA HEALING

Perceptual Blockages

The Higher Senses are connected to the Lower Senses and provide attributes of perceptual experience characteristic to their dimensional levels, but in the present stage of human genetic evolution the Higher and Lower Senses are not fully "plugged into" each other, and so the various higher dimensional fields, and the sensory abilities associated with the Higher Senses, appear to be separate or unavailable to the embodied human consciousness.. **The Higher and Lower Senses plug in through opening of *Kathara Lines* in the *Level-1 12-Tree Kathara Grid*, which creates a reciprocal opening and expansion of the DNA, Hova Bodies and activation of dormant nerve networks within the brain, molecular core and Central Nervous System of the body.**

Presently the human senses are *phase-locked into a 3-dimensional perceptual field* due to blockages between the *3rd and 4th Kathara Centers*, which block the Apparthi receivers within the Nada Hova Body from accepting compatible electrical impulses from the higher dimensional Hova Bodies. This Kathara Grid blockage manifests within the body as fragmented units of DNA (**"Junk DNA"**), that serve to render the natural processes of multi-dimensional data transfer into the Central Nervous System inoperable. Portions of *higher dimensional experience* come to conscious attention in the form of *dream perception* or *intuitive impression,* but the natural clear pathways of multi-dimensional perception are blocked to a large degree within contemporary humanity.

Perceptual Blockages and Kathara Healing

Blockages in the Kathara Grid and their reciprocal blockages in the human perceptual field can be progressively removed through practices of *Kathara Healing*, and utilizing the *Maharic Shield* to reawaken the Organic Imprint for Health.. As Kathara blockages are progressively repaired, the genetic code will expand its function by reassembling and activating the presently dormant portions of the DNA imprint. DNA expansion will create reciprocal alterations in the bodily metabolic rate, brain, blood, glandular and hormonal function, particle spin speed and operation of the Central Nervous System, that will allow for a progressive, natural activation of the Higher Sense facilities and expansion of the human perceptual field into the multi-dimensional arena. **In Kathara Healing the dormant Higher Senses will be progressively brought into activation**, beginning with the **6th -Audurea-Merkaba Sense** and the **7th-Tristet-Cellular Telepathy Sense** through which **Inner audio-visual-direct cognition** skills will progressively develop. These skills will be employed in the processes of Kathara Healing, and in so doing they will be progressively stimulated into greater levels of activation. As the dormant Sense facilities are called into activation, Kathara Healing techniques to remove Kathara Grid blockages are used to realign the *Holographic Template* to restore the **Imprint for Health,** through which natural perceptual expansion will progressively take place.

The 12 Human Senses

The collective spectrum of human sense facilities, including the Higher Senses of 15-Dimensional Human Anatomy can be categorized as follows:

The 12 Human Senses, Hova Bodies and Primary Experiential Attributes

Identified Lower Dimensional Sense Facilities

1. Visual-Sight Sense- ***Nada Hova Body (dimensions 1-2-3)***
2. Auditory-Hearing Sense - ***Nada Hova Body***
3. Tactile-Touch Sense- ***Nada Hova Body***
4. Olfactory-Smell Sense- ***Nada Hova Body***
5. Sense of Taste- ***Nada Hova Body***

Unidentified Higher Dimensional Sense Facilities

6. **Audurea -Merkaba Sense -** *Encompasses all Hova Bodies*
   Primary Attribute: **Manifestation-**circulation of consciousness and frequency accretion
7. **Tristet-Cellular Telepathy** -*Alphi Hova Body (dimensions 4-5-6)*
   Primary Attribute: **Inner Audio-Visual-Direct Cognition** reading energy signatures.
8. **Nurgode-Transmutation Sense -** *Betcha Hova Body(dimensions 7-8-9)*
   Primary Attribute: **Molecular Transmutation -** shape-shifting
9. **Ragode-Transmigration Sense -** *Betcha Hova Body*
   Primary Attribute: **Transmigration -** Projection of Consciousness
10. **Geuard-Transfiguration Sense-** *Mahara Hova Body (dimensions 10-11-12)*
    Primary Attribute: **Bilocation-**multiple holographic manifestations
11. **Shara-Centrifugal Sense-** *Raja Hova Body (dimensions 13-14-15)*
    Primary Attribute: **Electrical Projection -** consciousness enters dimensionalization
12. **Aurt-U - Centripetal Sense -** *Raja Hova Body*
    Primary Attribute: **Magnetic Accretion -**consciousness exits dimensionalization

CHAPTER 2 HIGHLIGHTS

THE IMPRINT FOR HEALTH

- Healing occurs through correcting distortions within the Kathara Grid and Hova Body Shields.

- The human body blueprint was designed to hold **12** Dimensions of Frequency/ Flash-line Sequences and 12 Dimensions of Consciousness, the content of 4 out of the 5 Hova Bodies within the 15-Dimensional Anatomy.

- The **4th Hova Body** - the **Mahara Hova Body** - is composed of Frequencies/ Partiki Phasing Flash-line Sequences of the **10th, 11th and 12th Dimensions.** It manifests through the 10th, 11th and 12th *Kathara Centers* and the **Maharic Shield** Scalar Grid of the **4th Harmonic.**

- The **Maharic Shield** of *Level-2 Kathara Grid Crystal Seals* holds the original **IMPRINT FOR HEALTH**- the original *Flash-line Sequences* of *Harmonic 4 Pre-matter Liquid Light* - that represent the *point of entry* into perceivable manifestation.

- The **Maharic Shield** can be used to **repattern** the Kathara Grid Holographic Template back to the original **Imprint for Heath**. The Flash-line Sequences of *Dimension 12* can *override* and correct *faulty Flash-line Sequences* that manifest as distortions within the Kathara Grid and Hova Body Shields.

- The **Hierophant Symbol Code** is a *Scalar Standing-wave* pattern containing the electro-tonal programs of the 11th and 12th Dimensions. The Hierophant Symbol code serves as a *Frequency* ***WAVE GUIDE*** through which the **Maharic Shield** can be activated to run 11th and 12th Dimensional Frequency through the Kathara Grid for **repatterning** the **Imprint for Health** within the Core Holographic Template of the body.

- Using the *Hierophant Symbol Code* the **Planetary Bio-feed Interface System™** can be activated to **empower** the Kathara Grid repatterning process via drawing Frequency from the **Planetary Maharic Shield.**

MORPHOGENETIC EXPANSION

GROWTH RATES

- Growth through time is created through **FREQUENCY ACCRETION -** drawing Frequency Bands from the Dimensional Unified Fields of Consciousness into the personal morphogenetic field for expansion.

- When the Kathara Grid is functioning properly there are **set morphogenetic field expansion periods** within the linear growth cycles of the human body.
  1. **Fetal Integration** to age 11-12.
  2. **Harmonic 2 Integration** age 12-22
  3. **Harmonic 3 Integration** age 22-33
  4. **Harmonic 4 Integration** age 33-44

  If the Kathara Grid is working properly the **body does not die** but rather **progressively transmutes** into *less dense states of matter* in an **immortal body form**, via accretion of the 4 Harmonics of Frequency.

- Present **Kathara Distortions** block integration of **Harmonic 2** and begin Molecular Compaction, body degeneration and Death.

LEVELS OF HUMAN ANATOMY

- The **Kathara Grid** is the *Core Holographic Template* upon which the human body and consciousness enter dimensionalization in order to experience manifestation.

- There are **3 Levels** to the Human *Kathara Grid*: *Level-1*- The **12- Tree**, *Level-2* – **Crystal Seals** and *Level-3* – **Diodic Points**.

- *Kathara Level-1 –12-Tree* contains the primary **12 Kathara Centers** and primary **15 Kathara Lines**, that correspond to various regions of the physical body and bio-energetic field and which "ground" consciousness into the framework of dimensionalization.

- *Kathara Level-2 – Crystal Seals* contain the **Signets** *–consolidated points of dimensional frequency* – the **Shields** – *3-dimensional scalar-wave discs emanating from sets of 3 Signets* through which the *Hova Body* energy capsules form– and the **Star Crystal Seals** and **Seed Crystal Seals** through which the *Chakras* and *Bio-energetic Field* levels emerge.

- *Kathara Level-3 – Diodic Points* are **Coordinate Points** where the flash-line sequences from one Shield and its Seals cross over and through those of other Shields to form minute **"White Hole" vortices** of energy-consciousness relay between the Hova Bodies and stations of identity.

- **Hova Bodies** are *3-dimensional scalar grids* in the form of **Concentric Spheres** that form around each **Shield** in the *Kathara Level-2 Crystal Seals Grid,* which serve as "tissue capsule" encasements around each Harmonic of energy-identity within the 15-dimensional scale.

- There are **5** *Dimensionalized* **"Ariea" Hova Bodies** forming human anatomy, which emanate from the **5** 3-Dimensional scalar **Shields** of the *Level-2 Kathara Grid*. Each Shield and Hova Body corresponds to **3 Kathara Centers** in the *Level-1 12-Tree Grid.*

- **5 Ariea Hova Bodies** and Corresponding **Shields:**
  Nada Hova-Telluric Shield-Dimensions1-3
  Alphi Hova-Doradic Shield-Dimensions 4-6
  Betcha Hova-Teuric Shield-Dimensions 7-9
  Mahara Hova-*Maharic Shield*-Dimensions 10-12
  Raja Hova-Rishic Shield-Dimensions 13-15

- The **Signets** and the **Shields** can be used to direct the function of the *Kathara Level-1 12-Tree Grid* and all built upon it. **Kathara Centers** can be used to trigger activation of the *Signets* and *Shields*.

- The first **4 Hova Bodies** and their **Horizontal Shields**, are most important in terms of human evolution. Activation of the **Signets** activates their corresponding **Shield**, which in turn releases the **tissue capsules** between *Hova Bodies*, allowing merger to occur. When the **first 4 Shields** activate the body is able to integrate the Avatar identity and **transmute** into pre-matter liquid-light wave form.

- In *Kathara Healing* the Harmonic-4 **Maharic Shield** is used to restore integrity to all *Shields* and *Hova Bodies* below, which progressively clears the *Kathara Grid* of distortions.

- *Level-1 Kathara Lines* direct the **Vertical** Flash-lines, *Level-2 Crystal Seals* control the **Horizontal** Flash-lines and *Level-3 Diodic Points* govern **Diagonal** Flash-lines. Together the 3 levels of the Kathara Grid form the 3-plane armature upon which the 3-dimensional hologram is created.

- Each **Signet** within a Hova Body Shield holds the electro-tonal program for one *Seed Crystal Seal.* Each *Seed Crystal Seal* corresponds to one Level-1 **Kathara Center** and holds the program for one primary **Axiom Line** – a *single axis* dimensional flash-line sequence.

- The **12 Primary Axiom Line** programs of the *Seed Crystal Seals* create separate *dimensionalized bands of frequency* within each Hova Body, forming a distinction between *dimensionalized levels* within the bio-energetic **Auric Field.** The Axiom flash-lines cross over and through each other to form the *Coordinate Points* between the Shields through which the *Level-3* **Diodic Point Vortices** emerge.

- The points in the Kathara Grid where Axiom Lines cross over and through the **Central Vertical Kathara Line** in the *Level-1 Grid*, and the **Horizontal flash-lines** in the *Crystal Seals* of the *Level-2 Grid*, form **consolidated Diodic Points** with *numerous Diodic Vortices* – these are called **CHAKRAS.** Each Chakra corresponds to an *Axiom Line, Auric Level, Seed Seal, Kathara Center*, and *Dimensional Frequency Band* and ***Merkaba Field.***

- The **12 Primary Axi-A-Tonal Lines** are a set of 12 *Vertical Energy Flow-lines* that run through the body, which are created by the consistent cross-over points of the rotating, single-axis flash-line sequences projected from the **Seed Crystal Seals.**

- **Axi-A-Tonal Lines 11 and 12** combine with the **Central Vertical Kathara Line** to form the **Hara Line**, the foundation of the **Central Body Current.**

- The **Hara Level** is the ***outer level*** of the **Kathara Level-2 Crystal Seals Grid**, can be viewed from *9th-Dimensional perception* and is used in Kathara Healing Level-2 .

- The *Hara Level* is composed of the **Hara Line, 6 Primary Hara Points** and the **Complexes** of energy processing they govern.

- The **Axi-A-Tonal Lines** and **Hara Level** of the *Kathara Level-2 Crystal Seals Grid* are elements of 15-Dimensional anatomy that are used in later applications of Kathara Healing.

- *Axi-A-Tonal Lines* set the core electro-tonal scalar-wave program upon which the **DNA** will manifest. The human DNA imprint contains the program for **12 strands of DNA**, one manifesting upon each of the **12 Primary Axi-A-Tonal Lines.**

- The human form was designed to embody **12 dimensions** of consciousness. The original human genetic code is called the **Silicate Matrix** and when operational it allows human cellular structure to **transmute** through ***4 Harmonics*** *of matter density*.

- The Silicate Matrix is mutated in present day humans due to ancient **Kathara Grid blockages,** which manifest as distortions of form and function within the DNA. "*Junk DNA*" is part of this distortion.

- Contemporary human DNA can be **reverse-mutated,** and its interdimensional functions restored, by applications of **Kathara Healing.**

- Each of the 12 strands of DNA *corresponds to* one of 12 *Dimensional Frequency Bands, Kathara Centers, Seed Crystal Seals, Signets, Axi-A-Tonal Lines*, and *Chakras*. Each **Hova Body Shield** governs the function of a **set of 3 DNA strands.**

THE 12 HUMAN SENSES

- The **human senses** are the *experiential result* of the *scalar standing-wave organization* inherent to the human form.

- Senses are **scalar arrangements** that *translate energy signatures* from the **Unified Field** into *usable data of relationship.*

- The sensing systems are built upon sets of **energy receivers** within the morphogenetic field that are called **APPARTHI**. The Apparthi inherent to Kathara structure serve to **block certain flash-line sequences** out of the morphogenetic field and allow others in to translate into perception.

- The flash-line sequences blocked by the Apparthi create **Gaps of Perception** of the *Unified Field*, which give us the *experience* of **externalization** and **separation of form.**

- *Apparthi receivers* create **refraction and polarization of flash-line sequences** that translate into a variety of identifiable ***holographic sensory perceptions.***

- The sense facilities correspond to the structure of the **Hova Body Shields** and **DNA.** In categorizing the interwoven sense facilities by their *primary orientation*, location of their *Apparthi receivers* within the Hova Bodies and the *experiential attributes* they produce, the human organism can be said to possess **12 Primary Senses.**
- The **5 known senses** are products of the operations of the **Nada Hova Body** and **Telluric Shield**, and thus correspond to **dimensions 1-3**.
- There are **7 as yet unidentified senses** that correspond to the higher dimensional *Hova Bodies*, *Shields* and presently dormant strands of *DNA*.
- As we evolve to *activate* and *reverse-mutate* the **DNA** imprint, through *Hova Body* alignment and merger, the experiential attributes of the higher dimensional senses will become more apparent.
- Due to their *dimensional affiliation*, the *5 known senses* can be viewed as the **Lower Senses**, the *higher dimensional senses* as the **Higher Senses.** Perception through the *Higher Senses* is called ***Higher Sensory Perception*** or **HSP.**

TAPPING THE INNER MENTOR

Levels of Consciousness and the Perceptual Bridge

In earlier text we have learned of the structure human consciousness takes when entering the scalar fields of dimensionalization for the perceivable experience of manifestation. In comprehending the true structure of multi-dimensional anatomy, with its **Kathara Grid, Hova Bodies** and **Shields, Diodic Points** and **levels of bio-energetic substance**, we can begin to see that there is much more to human consciousness than meets the eye - **much more *energy-identity* operating presently beyond the view of our 3-dimensionally focused earthly personality.**

There are portions of our consciousness stationed within the Hova Bodies of the higher dimensional fields, and these expanded levels of ourselves hold knowledge and participate in perceptual experience beyond the usual range of our waking perceptions. As we evolve and expand through **Hova Body Integration** and the resulting activation of dormant **DNA** components, we will progressively have *greater conscious access* to these expanded levels of consciousness while we are actively focused within the waking 3-dimensionally embodied state. However, **we do not have to wait until our evolution plods along to bring us to these expanded perceptual states,** we can, from our waking state of 3-dimensional consciousness begin to actively tap this hidden reservoir of knowledge and perceptual experience. **We can take an active hand in *expediting the process* of our multi-dimensional evolution by using our dimension-3 *mental body awareness* to direct our higher awareness into our current range of perception.** We can begin to build a **PERCEPTUAL BRIDGE** between our 3rd-Dimensional and Higher Dimensional consciousness. **In taking the initiative for building this Bridge of Consciousness we begin to stimulate our Higher Senses into operation, expediting the natural process of Higher Identity Integration.**

Through building a PERCEPTUAL BRIDGE
between our present state of awareness
and the levels of our consciousness
stationed in the Higher Dimensional Hova Bodies,
we can actively expedite our own Bio-Spiritual Evolution.

THE MENTOR

The Perceptual Bridge and Creating the Mentor

The Higher Dimensional levels of human consciousness are not dormant within their respective dimensional bands of function, they are quite awake and aware within their own dimensional fields. As we perceive and interact with the 3-dimensional world within which our present awareness is focused, the Higher Dimensional portions of our consciousness are also actively perceiving and interacting with the dimensional environments within which they are focused. **Higher dimensional aspects of our awareness *seem cut off* from our perceptual view because our consciousness is presently focused through a *genetic code arrangement and Kathara Grid orientation* that blocks out the frequencies of those higher stations of identity.** If we can realize that expanded portions of our consciousness simultaneously exist within the higher dimensional frequency bands, we can also realize that it is possible to access such levels of our own identity by building **Perceptual Bridges** between our 3-dimensional perspective and the portions of our consciousness stationed within the Higher Dimensional Hova Bodies.

We can begin to build a *Perceptual Bridge* between ourselves and our *Higher Selves*, by using simple *frequency modulation techniques* of Kathara Healing, such ~~as Technique #1~~: *Awakening the Mentor and the 6th and 7th Senses*. In this scalar-grid activating exercise, we will awaken dormant **Partiki Phasing Flash-lines** and activate dormant portions of the **Kathara Lines** in the *Kathara 12-Tree Grid*, to begin opening the channels of communication between the consciousness focused within the *Nada Hova Body* and that focused within the *Alphi Hova Body* of Harmonic-2. **In this exercise we will create the MENTOR, a portion of our consciousness composed of *3rd, 4th and 5th Dimensional frequencies*, that will serve as an *intermediary focus of awareness* through which we can begin to receive guidance and knowledge from our 2nd-Harmonic of consciousness.** Creating the MENTOR will begin the process of activating dormant **DNA** , opening new **nerve networks** within the Central Nervous System of the 3-dimensional body, and activating dormant **APPARTHI** frequency receivers within the *Level-2 Crystal Seals Kathara Grid*. **Through creating the MENTOR, we begin to build the Perceptual Bridge between the 1st and 2nd Harmonic of our conscious awareness.**

The MENTOR is a portion of our identity that we create to serve as an intermediary focus of awareness between the Nada Hova Body and the Alphi Hova Body; a PERCEPTUAL BRIDGE* between our first and second *Harmonics of Consciousness.

The MENTOR and Kathara Healing

In the practices of Kathara Healing we will use the MENTOR as a source of **guidance, information retrieval, objective observation** and as a **communications link** between our own higher consciousness and the higher levels of consciousness in those who come to us for healing facilitation. **Awakening and working with the MENTOR will become a standard practice in our applications of Kathara Healing.** In creating the MENTOR we are using our creative power of scalar-grid manifestation (thoughts and mental images are scalar-grid morphogenetic field forms) to bring into being a **specialized portion of our consciousness** that can **relay frequency** from our higher Harmonics of Consciousness to our 3-dimensionally focused identity. **With repeated use of the MENTOR focus of awareness we will develop the ability to simultaneously hold 2 Harmonics of consciousness, and we will begin to reclaim the natural flexibility of focus that is the healthful condition of embodied human consciousness.** Through the MENTOR focus of attention we can reach higher levels of cognition and understanding regarding application of healing facilitation, and we will more rapidly call into awakening the Higher Sense facilities of the 15-Dimensional Anatomy.

Other Uses of the MENTOR Focus

In creating the MENTOR focus we will begin to realize that even the 3-dimensional focus of awareness with which we are most familiar is far more flexible than once supposed. The MENTOR is not simply an **inert creation**, but rather a **living** portion of our conscious identity composed of portions of the waking self and portions of the Higher Self. MENTOR can become an **advisor**, a portion of self able to "keep its head above the illusion" when you are confronted with challenging events in the 3-dimensional arena; a part of you able to remain clam, focused and able to make clear decisions beyond the confusion of mental or emotional body processing. MENTOR can be sent into the **Cellular Memory** of the body to locate core issues or reincarnational blockages, and can serve to moderate between sub-personality fragments and the "Inner Child". MENTOR can like wise be used to **harmonize communications** between self and others, as you can send the MENTOR focus into the higher dimensional fields to connect with the higher selves of others for conflict resolution. Playing with the MENTOR focus can also have great rewards. You can focus your attention in the Mentor Space and then direct the MENTOR across the room to begin practicing affects of **BILOCATION.** You can travel with MENTOR to distant locations to develop **REMOTE VIEWING SKILLS**, which will be used in later Kathara Healing levels for **REMOTE HEALING FACILITATION.** MENTOR can become a **vehicle for conscious transport** of your awareness, while awake and also in **dreaming** and **higher dimensional projection.**

TECHNIQUE # 1

AWAKENING THE *MENTOR* AND THE 6TH AND 7TH SENSES

Awakening the Mentor:

1. Place the **fingertips** of both hands and the focus of your attention at the **Navel.** Take a few slow, full breaths then visualize the image of a **Blue Flame** (like a gas stove flame) deep inside the body behind the Navel. Gently **massage** the Navel region while envisioning the **Blue Flame.**

2. **Stop massage** and visualize a **Sphere of White Light within the Blue Flame** behind the Navel. **Place your attention within the White Sphere** and imagine that you can feel its reality around you, as if you are sitting within the White Light Sphere.

3. Breath gently and visualize the **White Light Sphere moving upward** through the center of your body until it reaches the center of the brain at the **Pineal Gland. Stop movement of White Light Sphere at the Pineal Gland,** then visualize the **White Light Sphere moving to the RIGHT** in a straight, horizontal line **OUTSIDE OF THE BODY** until it is positioned in your **auric field** just above your **RIGHT SHOULDER.** ***The White Light Sphere positioned over the Right Shoulder now** represents **a portion of your** awareness that was raised to the 5th-Dimensional frequencies via its passage through the Blue Flame and the 5th Kathara Center at the Navel.* You will name this part of your higher identity **MENTOR.**

4. Imagine that you can **feel the reality of the White Light Sphere** that holds your MENTOR resting gently above your Right Shoulder as a **"fuzzy ball of energy".** ***MENTOR** serves as the meeting point between your waking conscious mind and the **5th-Dimensional portion** of your Soul Matrix Identity- the **Archetype.** With practice, you **can lean the energy of your waking awareness** from its position within your head, over to the right and into the **MENTOR Sphere,** to receive direct cognition, audio and visual guidance from your 5th-Dimensional Identity Station.* **In Healing Facilitation** MENTOR will communicate with the 5th-Dimensional Archetype aspect of your **Client's Soul Matrix,** asking how you may best serve the Client's Highest Evolution; guidance MENTOR receives will be translated to you from MENTOR in the form of intuitive urge, spontaneous cognition or audio-visual directions. MENTOR will also serve as a **translator of information from your higher dimensional Stations of Identity** to your conscious mind. MENTOR can also be taken into the Dream State with you to assist in **Dream Recall** and initiating **Out-Of-Body Travel.**

5. Practice creating and sensing the Mentor Sphere over your Right Shoulder. Carry it around with you wherever you go, and play with *moving the focus of your consciousness back and forth between the inside of your head and MENTOR.* Try perceiving from both focuses of attention simultaneously. Such play will help stimulate neurological sensitivities to translating and perceiving information from the 5th and higher dimensional Stations of Awareness.

TECHNIQUE # 1: continued

AWAKENING THE *MENTOR* AND THE 6TH AND 7TH SENSES

Awakening the 6th and 7th Senses: The *Audurea* and the *Tristet*

1. Focus your attention on the **MENTOR Sphere** over your Right Shoulder. Take a few slow, deep breaths and visualize MENTOR *expanding into a Sphere the size of a Basket Ball.* Use the next **INHALE** to draw the MENTOR Sphere *toward your head*, visualizing MENTOR moving from over your Right Shoulder to a position *surrounding your entire head.* **EXHALE,** *stop the movement* of the MENTOR Sphere, and imagine that you can feel its "fuzzy energy" all around the outside of your head; *as if your head is now within the Basket Ball size MENTOR Sphere.*
2. With the next **INHALE** breath, **contract MENTOR into a small Sphere of White Light** positioned inside the center of the brain at the **Pineal Gland.** Exhale and breathe slowly for a moment. With the next **EXHALE** breath, *push* the small MENTOR Sphere from the Pineal Gland, straight forward and into the **6th-Chakra** at the *Forehead.* Visualize MENTOR moving into the 6th-Chakra from inside the head, then continuing to move through the 6th Chakra spiral straight outward and *away from the body.* Imagine the MENTOR Sphere traveling outward through the long **Blue-Violet Tunnel** of the 6th Chakra spiral, until you visually loose track of the MENTOR Sphere far out in front of you.
3. Take a few slow breaths, then on the next **INHALE** visualize the MENTOR Sphere *re-appearing* in the distance of your 6th Chakra spiral, and *riding in on the inhale breath,* back into the **Pineal Gland.** Breathe slowly.
4. With the next **EXHALE**, visualize the MENTOR Sphere moving *down from the Pineal Gland* at the center of the brain, stopping at the center of the **5th Chakra** in the *Throat.* Breathe slowly. On the next **EXHALE** push the MENTOR Sphere straight outward, imagining it moving forward from inside the Throat, into the blue green spiral of the *5th Chakra.* Visualize MENTOR *riding the current of the exhale breath* out into the distance within the blue-green **tunnel** of the 5th Chakra spiral, until it fades from view. Take a few slow breaths and on the next **INHALE** imagine MENTOR returning through the 5th Chakra blue-green spiral to rest deep within the **Throat.** Use the **final INHALE** to move MENTOR back up to the **Pineal Gland** and the final **EXHALE** to *push* the MENTOR Sphere back out to its "home" position in the auric field over the **Right Shoulder.**

This process opens up the 1st-5th Sub-frequency Bands within each dimensional layer of the 6th and 5th Chakras, beginning the process of opening the 6th-Audurea and 7th-Tristet Senses, stimulating Inner Vision and Audio to higher dimensional frequency perception.

SYMBOL SHORTHAND

TECHNIQUE # 1: Condensed Version
Awakening the Mentor and the 6th and 7th Senses

AWAKENING THE MENTOR

1. Finger *tips* at *Navel*, see *Blue Flame* in body behind Navel. *Massage* Navel with finger tips
2. See *White Sphere in Blue Flame* at Navel. Imagine *yourself in White Sphere*, feel it around you.
3. Move *White Sphere* up to *Pineal Gland*, see White Sphere move out over *Right Shoulder* creating your *MENTOR*
4. Feel *MENTOR* as *fuzzy ball* of energy hanging in your auric field over your Right Shoulder
5. *Practice shifting* the *focus* of your attention back and forth between usual focus and *MENTOR*

AWAKENING THE 6TH & 7TH SENSES

1. Expand MENTOR over Right Shoulder to size of *Basket Ball, INHALE* drawing *MENTOR* around your *head*, as if your head is inside the Basket Ball size energy that is MENTOR.
2. *INHALE, contracting MENTOR* into *Pineal Gland* as a small Sphere of White Light, *EXHALE* to push *MENTOR* out of *Pineal Gland out to Chakra 6* and into the *Blue-Violet Tunnel* of the Chakra 6 spiral.
3. *INHALE MENTOR back from Blue-Violet Tunnel* of Chakra 6 and into the *Pineal Gland*
4. *EXHALE MENTOR* from *Pineal Gland to Chakra 5. EXHALE MENTOR out* through *Chakra 5 Green Tunnel. INHALE MENTOR back* through *Green Tunnel* into *Chakra 5* at Throat. *INHALE MENTOR* up from *Chakra 5* to *Pineal Gland. EXHALE MENTOR* out from Pineal Gland to its "Home Station" over *Right Shoulder.*

Awakening the Mentor and the 6th and 7th Senses

AWAKENING THE MENTOR

Move MENTOR from Navel to Pineal Gland to position over Right Shoulder.

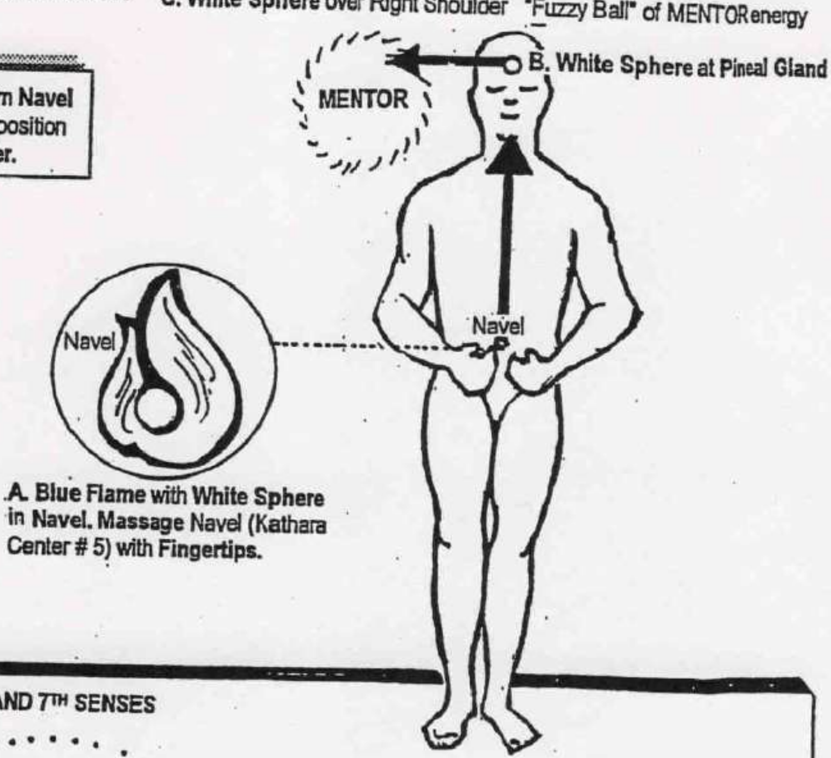
MENTOR
Navel
Navel

C. White Sphere over Right Shoulder "Fuzzy Ball" of MENTOR energy

B. White Sphere at Pineal Gland

A. Blue Flame with White Sphere in Navel. Massage Navel (Kathara Center # 5) with Fingertips.

AWAKENING THE 6TH AND 7TH SENSES

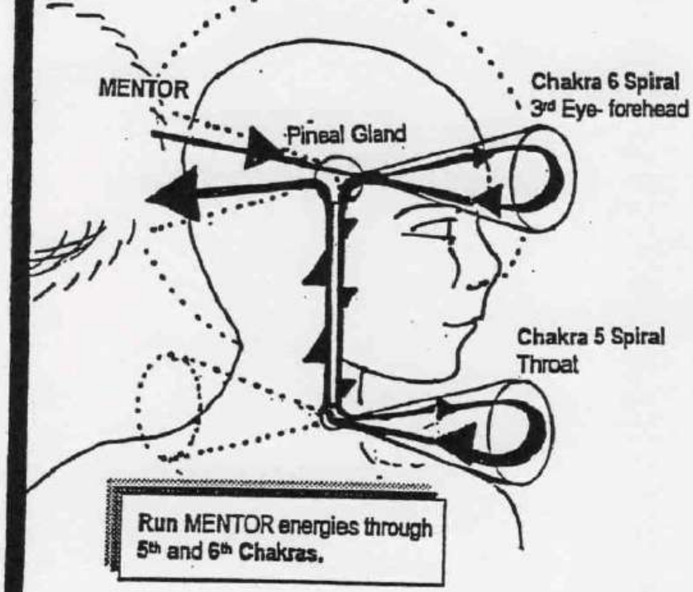
MENTOR
Pineal Gland
Chakra 6 Spiral
3rd Eye- forehead
Chakra 5 Spiral
Throat

Run MENTOR energies through 5th and 6th Chakras.

A. Expand MENTOR energy to basketball size and position around head
B. MENTOR in Pineal Gland as White Sphere
C. White Sphere into 6th Chakra Spiral then 5th Chakra Spiral and back to Pineal Gland.
D. Reposition MENTOR over Right Shoulder

MAHARIC RECODING, FREQUENCY SUB-HARMONICS & THE MAHARIC SEAL

Kathara Healing and Reaching the Core of Dis-ease

In earlier text we have explored the significance of the **Kathara Grid** and its role as the **Holographic Template** upon which the morphogenetic field, body and consciousness manifest within the 15-Dimensional system. We have learned that the Kathara Grid is the **core template** upon which the *morphogenetic field* manifests, and that the *morphogenetic field* is the **blueprint** upon which the body systems and embodied consciousness manifest. Through our study of these *core structures of manifestation* we come to realize that *true and lasting health or healing* must be approached from the deepest levels of personal creation, the *morphogenetic reality* within which all dis-ease conditions have their origin. **Kathara Healing applications are intended to reach the *core of systems malfunction*, the various levels of the *morphogenetic field and Kathara Grid*, in order to restore the *Imprint for Health* within the Holographic Template for manifestation.**

The Mahara Hova Body, Maharic Shield and the Imprint for Health

In *Level-1 Kathara Healing* we will learn the first and most important level of *restructuring the Holographic Template* to hold its original Imprint for Health - the MAHARIC RECODING PROCESS. The **MAHARIC SHIELD** is the manifestation template upon which the **MAHARA HOVA BODY** of *Harmonic-4* manifests. The Mahara Hova Body is the scalar-grid that holds the **10th, 11th and 12th-Dimensional** aspects of consciousness, and represents the **ORIGINAL POINT OF MANIFESTATION** -the **creation point** when consciousness moved from its pure scalar-wave state of being into the dimensionalized scalar-grids through which the illusion of external holographic projection could be experienced. The Mahara Hova Body and its station in Harmonic Universe-4 represent the original **PRE-MATTER LIQUID LIGHT** fields of manifestation through which we passed in order to enter the experience of density and individuation of matter form.

Prior to our entering the dimensionalization of the systems within which matter can be perceived, we existed in **pure scalar-wave form** - a scalar-pattern that held the **original, pure and perfect pattern** for our creation as consciousness. The Mahara Hova Body and the Maharic Shield are the Liquid Light Pre-matter Scalar-templates that hold our original pattern of wholeness, our original **CREATION IMPRINT**. Our creation imprint is the **ORIGINAL IMPRINT FOR HEALTH**, the condition of **eternal being** free from Hova Body distortions picked up in the lower dimensional fields.

THE 12th-DIMENSION and FREQUENCY SUB-HARMONICS

In Level-1 Kathara Healing we will utilize the *Imprint for Health* as held within the *Maharic Shield* to begin REPATTERNING the Kathara Grid back to its original level of function. The frequencies of the **12th-Dimension**, as embodied within the **Maharic Shield**, contain within them as **Frequency Sub-Harmonics** ALL OF THE FREQUENCY BANDS FROM THE DIMENSIONS BELOW, along with the natural frequency spectra characteristic to the 12th-Dimension. Each Dimensional Frequency Band within the 15-Dimensional Spectra contains within itself its own **12 Primary Sub-frequency Bands** (portions of the whole dimensional Partiki Phasing flash-line sequences), plus a **reflection set** of the frequency spectra from **all of the Dimensional Bands below.**

For example, Dimension 2 contains its own *12 Primary Sub-frequency Bands*, **plus** *a reflection set of the 12- Primary Sub-frequency Bands* that make up the *Dimension-1 Frequency Band* below. Dimension-1 has 12 Primary Sub-frequency Bands, Dimension-2 thus has 12 Primary Sub-Frequency Bands plus 12 Secondary Sub-Frequency Bands from Dimension-1for a total of *24 Sub-Harmonics* in Dimension 2. **This pattern of FREQUENCY SUB-HARMONICS builds upward through the 15-Dimensional Scale. The 12th-Dimensional Frequency Band has a total of 144 Sub-Harmonics** - 12 Frequency Bands x 12 Sub-Frequency Bands each = *144 Sub-Harmonic Frequency Bands*. Because the 12th-Dimension holds within itself *all of the Frequency Spectra and scalar-grid programs* from the lower dimensions, the *12th-Dimensional Frequencies have the ability to "plug into" and override any Partiki Phasing flash-line pattern from the lower dimensional fields*. **The ORIGINAL PATTERN OF CREATION for the lower dimensional structure itself is held within the *12th-Dimensional Frequencies*, and so the 12th-Dimension *Sub-Harmonics* can be used to *correct any distortion* with the dimensional frequencies of the Lower Hova Bodies and their Shields.**

The Maharic Shield and the 12th-Dimension

As the Maharic Shield and Mahara Hova Body are composed of 10th, 11th and 12th-Dimensional scalar grids, they contain within them the **144 Frequency Sub-Harmonics** of the entire 12-Dimensional Spectrum. The human morphogenetic field is designed to hold the 144 Sub-Harmonic Spectrum, which is fulfilled through activation of the original **12-Strand DNA Silicate Matrix** gene code. Because of its 144 Sub-Harmonic scalar-wave spectrum, the 12th-Dimensional level of the **Maharic Shield** can **reset** the *original, non-distorted imprint of Partiki Phasing flash-line sequences* within the entire *Kathara Grid 12-Tree, the Kathara Centers, Kathara Lines, the Hova Bodies, DNA* and within any level of the morphogenetic field and manifest body built upon them. The *Maharic Shield* is the Kathara **RESET BUTTON.**

Frequency Sub-Harmonics in the 15-Dimensional Scale

Dimension (Primary SFB + Secondary SFBs)	Dimension x 12-SFB each =	**# of Frequency Sub-Harmonics**
Universal Morphogenetic Field: (cannot fully process through the human gene code for embodiment)		
Raja Hova Body-RISHIC SHIELD		
D-15 (12 + 168 =)	15 x 12 =	180 **Frequency Sub-Harmonics**
D-14 (12 + 156 =)	14 x 12 =	168 **Frequency Sub-Harmonics**
D-13 (12 + 144 =)	13 x 12 =	156 **Frequency Sub-Harmonics**
Human Morphogenetic Field: **(can process through gene code via 12-Strand DNA Silicate Matrix)**		
Mahara Hova Body- MAHARIC SHIELD		
D-12 (12 + 132 =)	**12 x 12 =**	**144 Frequency Sub-Harmonics**
D-11 (12 + 120 =)	**11 x 12 =**	**132 Frequency Sub-Harmonics**
D-10 (12 + 108 =)	**10 x 12 =**	**120 Frequency Sub-Harmonics**
Betcha Hova Body- TEURIC SHIELD		
D-9 (12 + 96 =)	**9 x 12 =**	**108 Frequency Sub-Harmonics**
D-8 (12 + 84 =)	**8 x 12 =**	**96 Frequency Sub-Harmonics**
D-7 (12 + 72 =)	**7 x 12 =**	**84 Frequency Sub-Harmonics**
Alphi Hova Body- DORADIC SHIELD		
D-6 (12 + 60 =)	**6 x 12 =**	**72 Frequency Sub-Harmonics**
D-5 (12 + 48 =)	**5 x 12 =**	**60 Frequency Sub-Harmonics**
D-4 (12 + 36 =)	**4 x 12 =**	**48 Frequency Sub-Harmonics**
Nada Hova Body-TELLURIC SHIELD		
D-3 (12 + 24 =)	**3 x 12 =**	**36 Frequency Sub-Harmonics**
D-2 (12 + 12 =)	**2 x 12 =**	**24 Frequency Sub-Harmonics**
D-1 (12 + 0 =)	**1 x 12 =**	**12 Frequency Sub-Harmonics**

REMEMBER: A **FREQUENCY BAND** is a *fixed, repeating sequence* of *Partiki Phasing Flash-lines* - **"flashing on and off" cycles of scalar standing-wave points** within the scalar-grid template, caused by the *fission/ expansion and fusion/ contraction of Partiki Units*. A **SUB-FREQUENCY BAND** is a *fixed, repeating sequence of Flash-lines* that is *part of* the full cycle of the *Frequency Band* Partiki Phasing sequence. A **FREQUENCY SUB-HARMONIC represents a replica of Frequency Band Flash-line sequences from the dimensions below, carried within the Primary Flash-line sequences of the Dimensions above.**

HIGHER DIMENSION = MORE FREQUENCY SUB-HARMONICS = FASTER FLASH RATES = HIGHER FREQUENCY = FASTER OSCILLATION/ SLOWER VIBRATION = SHORTER WAVE LENGTH = SLOWER ELECTRON SPIN = LESS PERCEIVABLE MATTER DENSITY

The Maharic Shield - Core of Healing Therapeutics

To restore the **Imprint for Health** within the *Body-Mind-Spirit System*, the **Holographic Template of scalar-grids** upon which the system manifests must be **restored to its original function.** *Topical applications* of healing modalities within the physical body, bio-chemistry, DNA, Auric Field, Chakra System, and Level-2: Crystal Seals and Level-3: Diodic Points of the Kathara Grid can create *temporary change* toward healing, but the **disharmonic conditions** held within the core **Level-1 12-Tree Kathara Grid** will eventually **override** alteration made to the outer levels of the morphogenetic field or body systems. **Working with outer levels of the Bio-system can *assist* in the healing process, but first the *Imprint for Health* must be restored within the *Level-1 12-Tree Kathara Grid.*** The *Level-1 12-Tree Grid,* with its *Kathara Centers* and *Kathara Lines* is created upon the **144 Frequency Sub-Harmonic** Pattern. As the **Maharic Shield** and Mahara Hova Body are the only portions of the dimensionalized scalar-grid anatomy that hold the full **12th-Dimensional Frequency Band** with its *144 Sub-Harmonics*, the **Maharic Shield** and its corresponding **10-11-12 Kathara Centers** are the *only portions of the morphogenetic anatomy that can override and restructure the electro-tonal scalar standing-wave point arrangements within Level-1 Kathara Grid.* **Thus it is with the *Maharic Shield* that core healing therapeutics must begin.**

Critical Mass and Simultaneous Healing of Core/ Outer Morphogenetic Levels

Activating the **Maharic Shield** will **open, activate and realign** the *scalar-wave flash-line sequences* of dormant *Kathara Centers, Kathara Lines, Crystal Seals, Hova Bodies, Diodic Points, Chakras* and *DNA* within the outer levels of the morphogenetic field and body. Once the **Imprint for Health** has been *temporarily restored* within the *Core Holographic Template Kathara Grid* via running the *Flash-line sequences* of the **Maharic Shield** through the *Level-1 Kathara Grid*, the *outer layers of the anatomy can be further realigned* to expedite manifestation of the Kathara Grid Repatterning into the outer levels of the Bio-system. **Technique # 2: The Maharic Seal and the Liquid Light Cleanse**, begin the process of running the 12th-Dimensional Flash-line sequences, via their Frequency Sub-harmonics, through the Kathara 12-Tree Grid to **REPATTERN** *distorted flash-line sequences* within the core Holographic Template. In the beginning of using this technique, the affects of this reprogramming will last only several hours, as the outer levels of the Bio-field will have greater **"Critical Mass"** in their distortion patterns, than the Kathara Grid can immediately override. **Full healing comes when the renewed Imprint for Health reaches CRITICAL MASS within the morphogenetic field.** Critical Mass is reached via *frequent use of the Maharic Seal* and *outer level Bio-field therapeutics.*

Kathara Healing Theory and Applied Methods

In Kathara Healing the first priority is to *reset the integrity* of the *Kathara Level-1; 12-Tree Grid* by Re-patterning its flash-line sequence program to the *Maharic Shield* via the *Maharic Seal and Liquid Light Cleanse*. This process sets the electro-tonal **"ground"** within which alterations of *outer morphogenetic levels* can connect, reawakening dormant **APPARTHI** *frequency receivers* within the *Level-2: Crystal Seals*, which allows the morphogenetic field levels to *expand* enough to hold the *higher frequencies* of the *Imprint for Health*. After the **Healing Ground** is set within the *Kathara Grid*, additional applications of Kathara Healing can then be applied to various other levels of the Kathara Grid, Morphogenetic Field and Body. **There is an infinite variety of Kathara Healing Bio-energetic applications**, that range from Levels 1, 2 and 3 Kathara Grid work, reprogramming the Kathara Centers, Kathara Lines, Hova Bodies, Shields and Signets, reversing the Miasmic/Karmic Imprint and accelerating the opening of the Diodic Points, to working with the Chakra System, Meridian and Axiom Lines, DNA and physical body points while running the Triadic Currents for healing. In coming Levels of the *Kathara Healing Program* we will learn the use of the **GEOMANCIES** and **FIRE LETTERS**-key **Symbol Codes** that serve as **scalar-wave guides** to reset the electro-tonal *flash-line sequences* within the morphogenetic body, and the application of the Symbol Codes to the treatment of specific physical, emotional, mental and spiritual dis-ease conditions. We will explore working with the **Color Cap Stones** of the Dimensional Frequency Bands and how they can be used in combination with the Symbol Codes for use in **Manifesting** desired conditions and **De-manifesting** undesirables, both within the body and in the 3-dimensional life drama.

Applications of the Level-1 Kathara Healing Program

In Program Level-1 we are setting the foundations upon which all the other skill levels of Kathara Healing are built. The focus of the Level-1Program is to create Morphogenetic Repatterning via the Maharic Shield, and to begin clearing the Miasmic/ Karmic Imprint that blocks various areas of the Hova Body Imprint. We will next learn to add **FORCE AND POWER** to our *Morphogenetic Recoding Process* by opening the **PLANETARY BIO-FEED INTERFACE SYSTEM™** to draw energy directly from the **Planetary Maharic Shield**, and we will begin learning of the *physical body massage point sequences*- **Sequence 1 and 2 Kathara Points** - through which we can begin to *regulate the flow of energy within the Bio-system* for **Regeneration** of the Imprint for Health and **Revitalization** of the Body-Mind-Spirit System. In **Section 2** of this course we will explore **MULTI-VECTOR HOLOGRAPHIC RECODING** - clearing the *Miasmic Imprint* from the Kathara Grid.

SEQUENCE 1 AND 2 KATHARA POINTS

Sequence-1 Kathara Points for REGENERATION

Sequence-1 Kathara Points are mini-energy-vortex points in the body that connect directly to **Diodic Points** within the *Kathara Level-3 Diodic Points Grid*. Stimulation of the Kathara Points affects the movement of energy within related areas of the Kathara Grid, which creates a reciprocal movement of energy within the cellular structure corresponding to the Kathara Points.

When used **in proper sequence**, *Sequence-1 Kathara Points* ***open energetic gate ways to the Kathara Lines*** of the Kathara Level-1 12-Tree Grid, allowing patterns of frequency to be entered into the Kathara Lines, to realign disharmonic patterns of frequency running through the Holographic Program of the Kathara Lines. When the Kathara Lines are opened to receive **11th** and **12th** dimensional frequency infusions from the **Mahara Hova Body**, the Holographic Program of the Kathara Grid can be progressively ***regenerated*** to the Imprint for Health, as held within the program of the Maharic Shield.

Sequence-1 Kathara Points are used to *open* the Kathara Lines prior to *Technique # 2: The Maharic Seal and the Liquid Light Cleanse* and *Technique # 4: The Maharic Infusion*. They **may also be used alone** with deep breathing, for frequent healing support, to draw energy up from the Maharic Shield, re-balancing and regenerating the Imprint for Health within the Kathara Grid and Body-Mind-Spirit System. It is not necessary to close the Kathara Lines after using Sequence-1 Points as they will **close automatically** in **several hours**. If one desires to further revitalize the cellular structure, running the *Sequence-2 Kathara Points* for *closing* the Kathara Lines will increase the amount of energy running through the *body cells*, as the Kathara Lines are no longer absorbing the energy.

Sequence-1 Points **regenerate and re-balance the *Kathara Core* first, which creates *more lasting but less immediate* changes in the body. *Sequence-2 Points* close the Kathara Lines if they are opened and draw energy from the Maharic Shield *directly into the bio-energetic field and physical cells*, rather than the core Kathara Grid, creating *more immediate, but temporary*, revitalization of the body.** The patterns held within the Kathara Grid- core holographic template will eventually override those set into the bio-energetic field alone. **Used in *combination* with each other the *Sequence-1 and Sequence-2 Kathara Points* can affect more lasting revitalization by *simultaneously realigning the core template and the outer layers* of the bio-energetic system and cellular structure.** In combination with other Kathara Healing techniques the affects and restoration of the Imprint for Health are further amplified and accelerated.

Sequence-2 Kathara Points for REVITALIZATION

Sequence-2 Kathara Points are mini-energy-vortex points in the body that, like *Sequence-1 Kathara Points*, connect directly to **Diodic Points** within the *Kathara Level-3 Diodic Points Grid.* Stimulation of the Kathara Points affects the movement of energy within related areas of the Kathara Grid, which creates a reciprocal movement of energy within the Bio-Field and cellular structure corresponding to the Kathara Points.

When used **in proper sequence,** *Sequence-2 Kathara Points* ***close the energetic gate-ways to the Kathara Lines*** of the Kathara Level-1 12-Tree Grid, which the *Sequence-1 Kathara Points* had opened. Closing of the Kathara Lines governed by the Sequence-1 Kathara Points, ***creates a reciprocal OPENING of Nadial Lines (Nadis*** energy rivulets that carry energy from the Nada Hova Body/ Nadial capsule, into the physically manifest systems**)** within the *outer levels* of the Bio-energetic Field. Opening of corresponding Nadis allows restorative patterns of frequency to be entered into the **outer levels** of the Bio-energetic Field and Chakra System, to realign disharmonic patterns of frequency running through the *Chakras, Auric Field and Physical Cells.* When the Nadis, Chakras and outer Bio-energetic Field are opened to receive **11th and 12th dimensional** frequency infusions from the **Mahara Hova Body,** the energetic integrity of the physical cells and Bio-energetic Field can be progressively ***revitalized*** with Maharic frequency and the Imprint for Health, as held within the program of the Maharic Shield.

Sequence-2 Kathara Points are used to *close* the Kathara Lines following *Technique # 4: The Maharic Infusion.* They **may also be used alone** with deep breathing, for frequent healing support, to draw energy up from the Maharic Shield, through the Kathara Grid and into the **outer levels** of the Bio-energetic Field. Directing Maharic frequency into the outer levels of the Bio-energetic Field regenerates the Imprint for Health within, and directly **revitalizes** the *body cells* for accelerated Body-Mind-Spirit healing. If one desires to revitalize the cellular structure, running the *Sequence-2 Kathara Points* for closing the Kathara Lines will **increase the amount of energy running through the *body cells*,** as the Kathara Lines are no longer absorbing the energy. **Sequence-2 Points revitalize and rebalance the *outer levels of the Bio-energetic Field* first, which creates more immediate but temporary revitalization of the body, as restorative energies do not directly enter the Level-1 Kathara Grid.** To create *lasting change* when using the *Sequence-2 Points* alone, restorative electro-tonal programs must be **repeatedly** entered into the Bio-energetic field until the new programs reach **critical mass,** at which time the new programs will begin to *subtly influence* the programs held within the Kathara Grid. **The primary difference between Sequence-1 and Sequence-2 Kathara Points is this: *Sequence-1 Points*** channel energy from the Maharic Shield into the ***core Kathara Grid*** holographic template. ***Sequence-2 Points*** channel energy from the Maharic Shield into the ***outer levels of the Bio-field and body cells.* For this reason the affects of *Sequence-2 Kathara Points* are more *immediate*, and those of *Sequence-1 Kathara Points* are more *permanent.***

The patterns held within the Kathara Grid core holographic template eventually override those set into the outer Bio-field alone. **Used in *combination* with each other and other healing applications the *Sequence-1 and Sequence-2 Kathara Points* can affect more lasting revitalization by simultaneously *realigning the core template and the outer layers* of the Bio-field and cellular structure.** Using ***Kathara Point Sequences*** in combination with each other and other Kathara Healing techniques will amplify and accelerate the affects and restorative function of the Imprint for Health within the Body-Mind-Spirit system.

14 SEQUENCE 1 KATHARA POINTS

Regeneration of the Imprint for Health

Sequence-1 Kathara Points **OPEN** energetic gate-ways within the **Kathara Lines** of the *Kathara Level-1 12-Tree Grid*, to receive electro-tonal reprogramming for **regeneration** of the *Imprint for Health*. *Sequence-1 Points* are used to open the Kathara Lines **prior to** other energy-healing applications, to open the Kathara Grid so the regenerative patterns of other healing applications can "ground" within the Kathara. Sequence-1 Points re-generate and rebalance the **Kathara Core** first, creating **more lasting but less immediate changes** in the body.

1............CROWN
2 & 3.......TEMPLES
4.............BASE OF SKULL
5 & 6........BELOW SHOULDERS
7 & 8........ABDOMEN IN FROM HIPS
9 & 10......BACKS OF KNEES
11 & 12....SOLES AND TOPS OF FEET
13............4TH HEART CHAKRA
14............6TH THIRD EYE CHAKRA

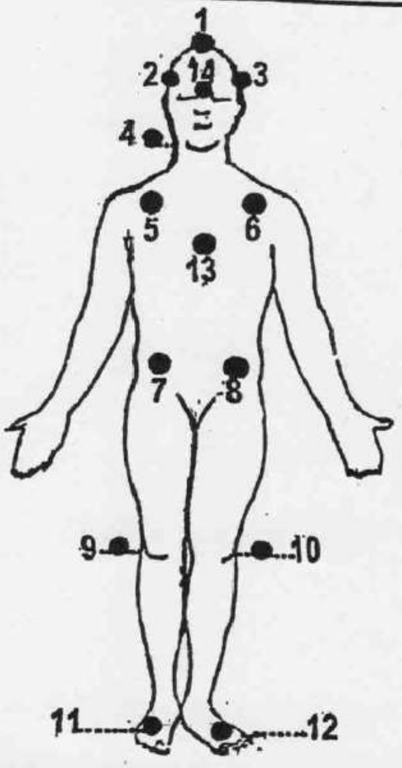

Instructions: Breathe slowly and deeply. Imagine a flow of energy running from the **Maharic Shield** (12" below feet), up into the body through the **soles of the feet** and the **1st Chakra**. Gently **rub** the *Kathara Points* with fingertips for 10 – 20 seconds each, following the **precise sequence** described above.

14 Sequence-1 Kathara Points for Regeneration

LOCATION & APPLICATION SEQUENCE

1............CROWN

2 & 3.......TEMPLES

4............BASE OF SKULL

5 & 6.........BELOW SHOULDERS

7 & 8.........ABDOMEN IN FROM HIPS

9 & 10......BACKS OF KNEES

11 & 12....SOLES AND TOPS OF FEET

13...........4TH HEART CHAKRA

14............6TH THIRD EYE CHAKRA

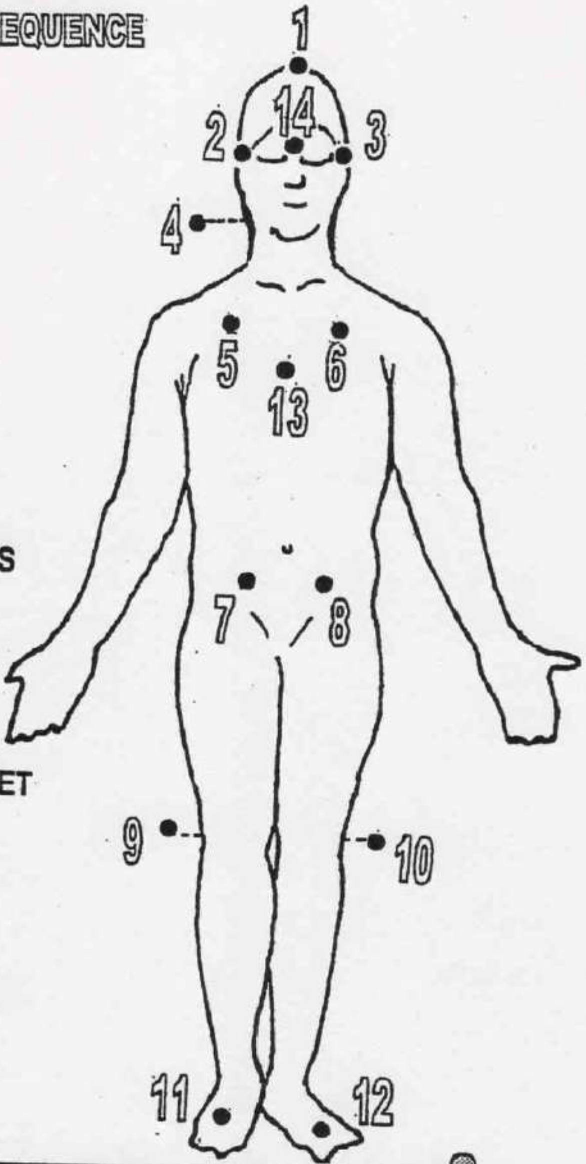

Instructions: Breathe slowly and deeply. Imagine a flow of energy running from the **Maharic Shield** (12" below feet), up into the body through the **soles of the feet** and the **1st Chakra.** Gently **rub** the *Kathara Points* with fingertips for 10 – 20 seconds each, following the **precise sequence** described above.

6 SEQUENCE 2 KATHARA POINTS

Revitalization of the Imprint for Health

Sequence-2 Kathara Points **CLOSE** energetic gate-ways within the **Kathara Lines** of the *Kathara Level-1 12-Tree Grid*, to direct restorative frequencies into the **outer levels of the Bio-energetic Field** for **revitalization** of the *Imprint for Health* within the physical cells. *Sequence-2 Points* are used to close the Kathara Lines **following** use of the *Sequence-1 Kathara Points* or other energy-healing applications, to close the Kathara Grid so revitalizing frequencies can directly enter various *outer levels* of the Bio-energetic Field, such as the **Auric Levels, Chakras, Meridians, DNA** and **physical cells**. Sequence-2 Points revitalize and rebalance the **outer levels of the Bio-energetic Field** first, creating **more immediate but temporary changes** in the body.

1 & 2........NECK, BELOW EARS

3 & 4........BELOW CHEEKBONES

5..............BASE OF TAILBONE

6..............THYMUS

*** AMPLIFER –3RD SOLAR PLEXUS CHAKRA**

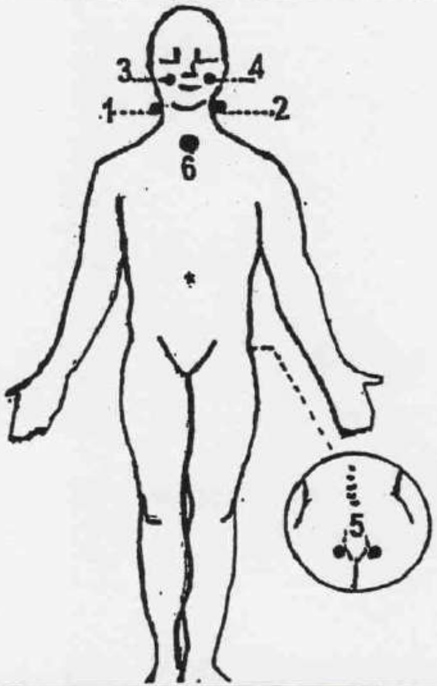

Instructions: Breathe slowly and deeply. Imagine a flow of energy running from the **Maharic Shield** (12" below feet), up into the body through the **soles of the feet** and the **1st Chakra**. Gently **rub** the *Kathara Points* with fingertips for 10 – 20 seconds each, following the **precise sequence** described above.

6 Sequence-2 Kathara Points for Revitalization

LOCATION & APPLICATION SEQUENCE

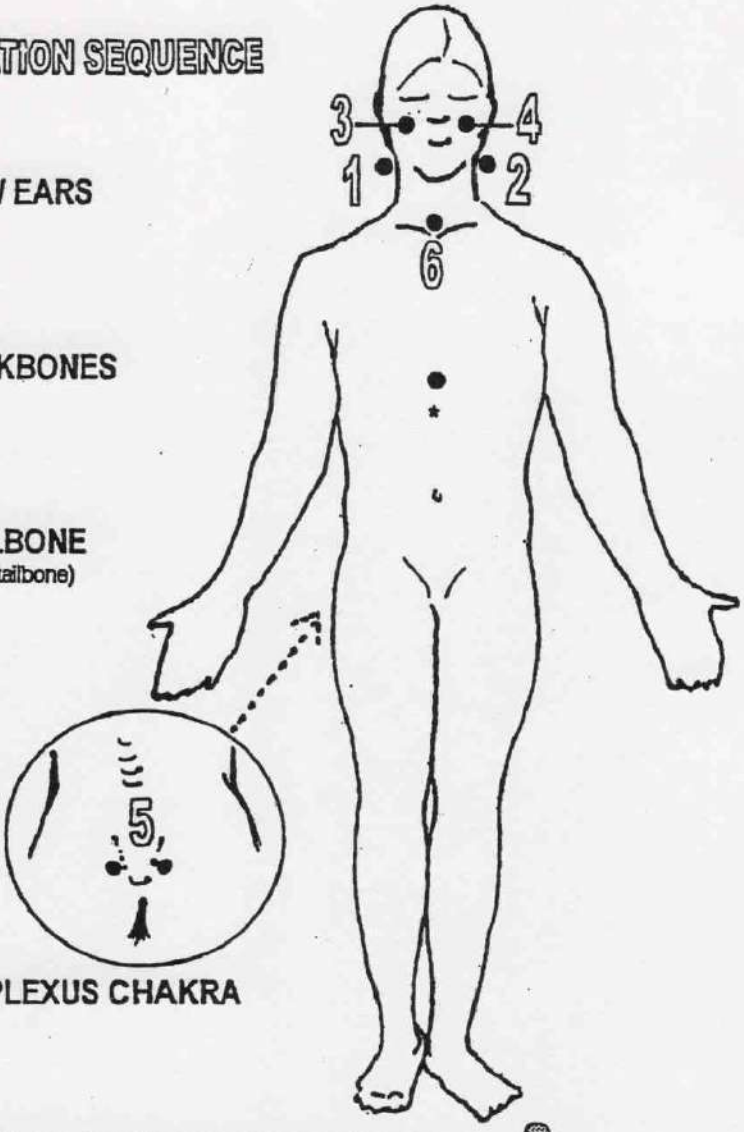

1 & 2........NECK, BELOW EARS

3 & 4........BELOW CHEEKBONES

5..............BASE OF TAILBONE
(one point on each side at base of tailbone)

6..............THYMUS

* AMPLIFER – 3RD SOLAR PLEXUS CHAKRA

Instructions: Breathe slowly and deeply. Imagine a flow of energy running from the **Maharic Shield** (12" below feet), up into the body through the **soles of the feet** and the **1st Chakra.** Gently **rub** the *Kathara Points* with fingertips for 10 – 20 seconds each, following the **precise sequence** described above.

TECHNIQUE # 2

THE *MAHARIC SEAL* AND THE *LIQUID LIGHT CLEANSE*

1. Close your eyes and visualize the image of a flat (2-dimensional line drawing) 6-Pointed Star of David (**The Merkaba Star**) made of **White Light**, stationed at the **6th-Chakra** on the *inside of the Forehead.* (If you cannot yet generate inner visualizations, simply imagine that the symbol is there.) Continue to visualize or imagine, and move the **White Merkaba Star** image over toward the **left** in your inner field of vision. In the *center* of your inner field of vision, visualize/ imagine the image of *another Merkaba Star*, made of a **Dark Silver Light**. Observe or imagine the **2 Merkaba Stars** in your inner field of vision and then slowly *move the images toward each other* until they have merged into *one Merkaba Star* Symbol made **of Pale Silver Light**. (The *Pale Silver Double Merkaba Star Symbol* is called the **HIEROPHANT** – it is the scalar wave template that governs the release of phase-lock on the *Maharic Shield.)*
2. Visualize/ imagine the Hierophant Symbol clearly in mind, then shift attention to the sound of your breath. Take several slow, deep breaths. On the next **EXHALE** use the force of breath to *forcefully push* the **Hierophant** down from its position at the 6th-Chakra, through the *Central Body Current* and into the **Earth's Core.** Take a few more slow breaths while visualizing/ imagining the Hierophant Symbol in the center of the Earth's core. Imagine now that the *Hierophant begins to spin in the Earth's core*, the Dark Silver Merkaba Star and the White Merkaba Star spinning in *opposite directions.* Spin the Hierophant progressively faster until it "takes on a life of its own", and suddenly "pops and expands" to form a **Pale Silver Sphere of Light** *spinning in the Earth's core.*
3. Observe the Pale Silver Hierophant Sphere spinning in Earth's core, then visualize/ imagine that a **huge, horizontal Pale Silver spinning disc** emerges out from the Hierophant Sphere, forming a **massive Pale Silver Disc of Light** *spinning throughout the Earth and extending outward into the atmosphere* as a **rotating, horizontal plane.**

This horizontal disc-plane extending through the Earth represents the *Mahara Hova Body*, or **Maharic Shield** of the planetary morphogenetic field – the Pre-matter Liquid Light scalar grid composed of frequency from dimensions 10, 11 and 12. (Known as the ***Shield of Amaratena***). Using the Hierophant wave template you have now activated the planetary Maharic Shield, ***opening the Planetary Bio-feed Interface System™ into your bio-energetic field.*** This constitutes opening your personal Kathara Grid and morphogenetic field into the Planetary Kathara Grid and morphogenetic field, allowing you to draw an *unlimited supply of energy* from the planetary field. You can draw in only as much as your personal morphogenetic field can hold. With repetition of this practice the energy-holding capacity of your Kathara Grid progressively increases, allowing you *a continually expanding supply of energy* for healing and revitalization.

TECHNIQUE # 2: Continued
THE *MAHARIC SEAL* AND THE *LIQUID LIGHT CLEANSE*

4. While remaining aware of the *Planetary Maharic Shield* spinning on a horizontal plane throughout the Earth's body, return the focus of your attention to the spinning **Silver Hierophant Sphere** rotating at Earth's core. (You will now use the Hierophant Sphere to activate your **personal Maharic Shield** and release the 12th-Dimensional Pillar of Light within your bio-energetic field.). Take several slow, deep breaths, while visualizing/imagining the Hierophant Sphere spinning in Earth's core. On the next **INHALE**, draw the Hierophant Sphere *up from Earth's core* with the inhale breath, forming a **Pale Silver Cord** *about 6" thick* trailing behind the Hierophant as it rises to the **12th Chakra**, 6" beneath the feet. *One end of the Silver Cord remains rooted in* ***Earth's core****, the other end is attached to the* ***Hierophant Sphere*** *and follows its movement.* Visualize the **12th Chakra** as a small, **transparent disc** about *3" in diameter*, positioned *6" below the feet.* Stop the movement of the Hierophant in the *center of the 12th Chakra*, imagine the Sphere beginning to spin and *visualize the diameter of the 12th Chakra disc expanding to about 6" in diameter*, as the Hierophant spins at its center. Take a few slow, deep breaths as the *Hierophant spins in the center of the* ***12th Chakra.***

5. On the next **INHALE** draw the Heirophant Sphere *up from the 12th Chakra*, through the **Central Body Current** and up to the **14th Chakra** *36" above the head.* As the Hierophant Sphere rises, visualize the *6" thick* **Silver Cord** drawing up through the **12th Chakra** and farther up through the **Central Body Current,** following the path of the Hierophant Sphere to the **14th Chakra.** Visualize the Silver Cord as a long, thick **Pale Silver Cord of Light**, running from the *Earth's core, through your body and out the top of your head,* connecting to the **14th Chakra** 36" above the head. Take several deep, slow breaths while imagining that you can *feel the sensation* of this *massive current of clear, pure energy* running through the *center* of your body.

6. On the next **EXHALE**, push the spinning Hierophant Sphere *down* through the **Central Body Current,** while leaving the *top end* of the Pale Silver Cord attached to the **14th Chakra** 36" above the head. Move the Hierophant down through the **12th Chakra** 6" below the feet, and **stop the Hierophant** *6" below the 12th Chakra;* a position about **12" below the feet.** This is the location of the **Maharic Shield** within the morphogenetic field of the body. Breath slowly and visualize the Hierophant *spinning progressively faster* 12" beneath the feet, until the Hierophant "takes on a life of its own" and suddenly "*pops and expands*" to form an **expanding horizontal disc of Pale Silver Light spinning 12" beneath the feet.** ***Feel the disc as a circular platform 3 yards in diameter below feet.***

TECHNIQUE # 2: Continued
THE *MAHARIC SEAL* AND THE *LIQUID LIGHT CLEANSE*

7. Call to mind the image of the *massive horizontal disc of Pale Silver Light* spinning through the Earth's body – the **Planetary Maharic Shield**. Imagine the sensation of the large, spinning *Pale Silver platform disc* spinning 12" beneath your feet – your **personal Maharic Shield**. Breathe slowly and deeply while visualizing/ imagining a *massive rush of pure, cool* ***Pale Silver energy*** running from the *planetary Maharic Shield*, up through the **Pale Silver Cord** and into the *personal Maharic Shield* 12" beneath your feet. Breath as this energy runs, *filling your Maharic Shield to its energy-holding capacity*. When your Maharic Shield has received as much Pale Silver energy as it can hold, it will "take on a life of its own", and suddenly "*pop and expand*" upward into an ***oscillating Pale Silver Pillar of Light*** *surrounding and permeating the body*, a *tube of light* extending beyond the body *about 18"*, and running from the **Earth's Core** to the **14th Chakra** 36" above the head. *Feel the reality of your body encased in the cool, refreshing Pale Silver Pillar of oscillating light.* Take a few slow deep breaths and imagine the energy from the Pillar *emanating through all of your body cells*. Your Bio-energetic Field is now sealed with a protective barrier of 11th/12th-Dimensional frequency – **the Maharic Seal.**

8. As the Pale Silver Pillar of Light oscillates around you, place your attention for a moment at the **13th Chakra** in Earth's core and **INHALE** a final **Spark** of *Pale Silver Light* from Earth's core, drawing the Spark up through the **Pale Silver Cord**, into the **Main Vertical Channel of the Kathara Grid**, and into the center of the **4th Heart Chakra**. Breath slowly and easily while visualizing energy from the *Silver Spark* radiating throughout the embodied **12-Tree Kathara Grid**, *filling it with Pale Silver Light.* When the Kathara Grid has reached its energy-holding capacity, the ***12th Chakra*** *will automatically contract* to decrease the flow of Silver Light from the Earth's core. As the 12th Chakra contracts, the *Planetary Maharic Shield* stops spinning and transmitting energy through the Silver Cord, returning the Silver Chord to its dormant state. The ***Maharic Seal*** **Pale Silver Pillar of Light** will remain in your bio-energetic field between **1-3 hours**. Over a *3-month period* of frequent, consistent *daily practice* of this process, the Maharic Seal will self-sustain in the bio-energetic field for **12-14 hours**.

Use of this technique allows a healing Facilitator to activate the **Maharic Shield** and **Seal** in others, through the ***Maharic Infusion process*** (Technique # 4), while Facilitator's Pillar remains active. Progressive use of the Maharic Seal and running the Liquid Light creates a reciprocal clearing, reordering and realignment of the Kathara Grid-***Holographic Template*** – and all subtle energy and physical systems built upon it. The Lower Hova Bodies are progressively cleared and realigned with the *original perfection of the Pre-matter Liquid Light Mahara Hova Body*, and the body and genetic code progressively expand to hold higher dimensions of conscious awareness.

TECHNIQUE # 2B: THE MAHARIC QUICK SEAL

1. Work with a 24 pointed, 3 Dimensional Star (Heirophant) and be aware that you will be working with beings known as the Breneau Rishi ...

2 Begin by slowing your breathing , as you imagine or visualize a 24 point, 3 -Dimensional, Heirophant at the Pineal Gland in the center of your brain. Visualise the Heirophant surrounded by a Pale Blue Sphere of Light.

(The Blue Sphere serves as a buffer for your readiness to accept the frequencies associated with using the 24 pointed Star; the Blue Sphere holds the energy for you until your body can handle the energy. If you can handle it, the energy will simply pass through the blue sphere).

3 INHALE, as if you are going to grab the blue sphere containing the 24 point star located at the Pineal, and on the EXHALE move the 24 point star/ blue sphere, all the way down to Earth's Core. Try to hear a sound tone as the Heirophant/ Sphere hits the Planetary Shield ...

4 INHALE and draw the Heirophant/ sphere up to your Personal Maharic Shield 12" below your feet ... and EXHALE while watching the Pale Silver-Blue sphere expand out to a large sphere ... and watch as your Maharic Shield pop out as a disc, about 4' in diameter, Pale Silver with a light coating of Pale Blue.

5 Bring your attention to the center of your Maharic Shield 12" below your feet ... and INHALE the 24 point Star (only) up into the Heart Chakra. EXHALE, expanding the frequencies of the 24 point Star into the Heart Chakra.

6 INHALE to grab the 24 point Star, and EXHALE, pushing the Star up into the 14th Chakra 36" above the head. Imagine that the 24 point Star is spinning in a clockwise direction in the 14th Chakra. As you do this, try to feel/ sense the energy around you, just a few inches out from your body ...

7 Now, put your attention into your Maharic Shield 12" below your feet, take a couple of relaxing breaths and INHALE deeply pulling the Pale Silver and Blue energy up, as if you are pulling on your Maharic Seal to latch it on to your 14th Chakra. Try and feel the sensations just a few inches from your body now – feel the difference if you can. This procedure gives you a Quick Seal.

8 Now envision the cord of Pale Silver Maharic Light that would normally come up with your Shield and focus on the Earth's Core. Begin drawing energy all the way up the 4" diameter cord and into the body. INHALE the energy up into the Heart Center, and expand it there. Begin to form a pale Silver ball of Maharic frequency.

On each EXHALE send your energy down to the Earth's Core to bring up another load of Maharic frequency, up from the Planetary Shields, expand it into the Heart center and repeat several times. Notice that the Cord grows larger as you do this, expanding from approximately 4" to 6 to 8" diameter, until finally it feels like a skin around (as well as within) your body ... as you load your Astral field with Maharic Frequency.

9 Move your attention to the Pale Silver ball you have created in your Heart Center, INHALE and move it up to the 14th Chakra 36" above your head .. as you do, feel for the sheath of Maharic Frequency encasing your etheric body, close to your skin. (This Maharic/ Christos skin charges, and saturates, your etheric body – feel the peacefulness and all-knowing nature of this frequency).

SYMBOL SHORTHAND

TECHNIQUE # 2: Condensed Version
The Maharic Seal and Liquid Light Cleanse

Symbol Shorthand is used to help in remembering the primary steps of a technique

1. *Silver Hierophant Star at Chakra 6*
2. *EXHALE* Hierophant Star from *Chakra 6 to Earth's Core –Spin Hierophant Star* at Earth's core-until it *pops* into *Hierophant Silver Sphere at Earth's core.*
3. *Hierophant Sphere* spins *at Earth's core* until *Silver Disc of Earth's Shield* emerges.
4. *INHALE- Hierophant Sphere* up from *Earth's core to Chakra 12* trailing *Silver Cord* from Core.
5. *INHALE- Hierophant Sphere* up from *Chakra 12 to Chakra 14*, trailing *Silver Cord.*
6. *EXHALE- Hierophant Sphere* down from *Chakra 14 to Personal Maharic Shield* and *spin Hierophant Sphere* until *Personal Shield pops* out to become *Silver disc* plane 12"beneath feet.
7. Draw *Silver energy* into *Personal Shield from Earth's Shield* through *Silver Cord,* until *Personal Shield* is *filled* and *pops* to become vertical *Pillar* of Silver Light around body.
8. *INHALE*-draw *Spark of Silver Light* from *13th Chakra* at *Earth's Core,* into *main* vertical channel of *Kathara Grid* then to *Chakra 4;* from Chakra 4 transmit energy from Spark to fill *12-Tree* Kathara Grid. Chakra 12 auto-shut down when 12-Tree Grid is full.

THE MAHARIC SEAL

63

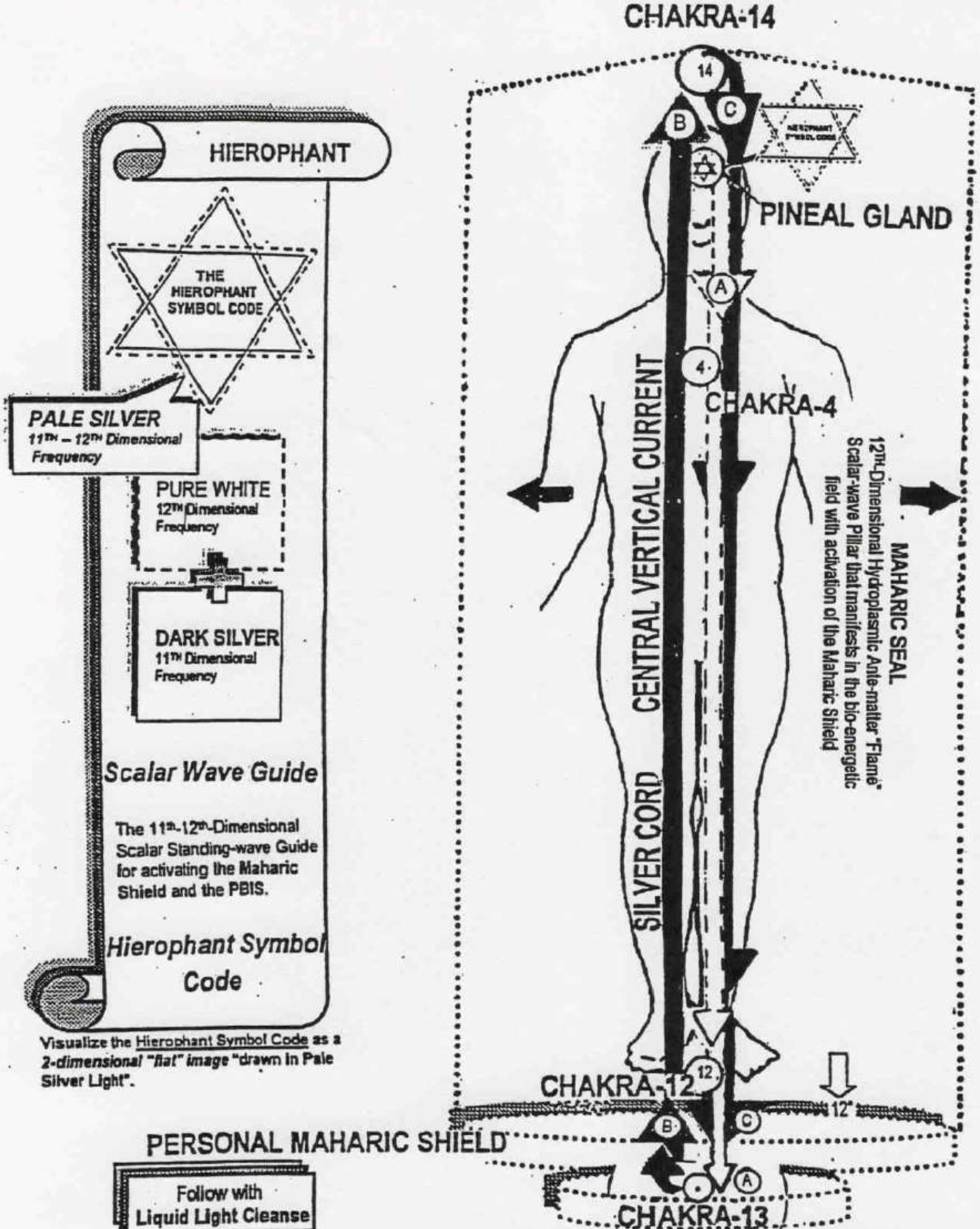
CHAKRA-14
14
B
C
HIEROPHANT SYMBOL CODE
PINEAL GLAND
A
4
CHAKRA-4
CENTRAL VERTICAL CURRENT
SILVER CORD
MAHARIC SEAL
12TH-Dimensional Hydroplasmic Anti-matter "Flame" Scalar-wave Pillar that manifests in the bio-energetic field with activation of the Maharic Shield
HIEROPHANT
THE HIEROPHANT SYMBOL CODE
PALE SILVER
11TH – 12TH Dimensional Frequency
PURE WHITE
12TH Dimensional Frequency
DARK SILVER
11TH Dimensional Frequency
Scalar Wave Guide
The 11th-12th-Dimensional Scalar Standing-wave Guide for activating the Maharic Shield and the PBIS.
Hierophant Symbol Code
CHAKRA-12
12
12"
B
C
A
PERSONAL MAHARIC SHIELD
CHAKRA-13
PLANETARY MAHARIC SHIELD

Visualize the Hierophant Symbol Code as a *2-dimensional "flat" image* "drawn in Pale Silver Light".

Follow with Liquid Light Cleanse

LIQUID LIGHT CLEANSE

KATHARA CENTERS

#12 – *The Capstone*
In 10th Chakra
6" above the head

# 11 – *Crown Jewel*
In Pineal Gland
at center of brain

# 9 & # 10 –*The Shar*
In brain
behind ears

# 8 – *Spark of Orion*
In Thyroid Gland
at 8th Chakra

# 6 & # 7-*The Wings*
In shoulder indents
through to back.

# 5 –*Seat of Amenti*
At Navel in
Central Body Current

# 3 & # 4 -*The Dal*
male: Ureters
female: Ovaries

# 2 – *Naum-Yun*
male: Prostate Gland
female: Vagina

# 1 – *Earth Core*
at center of Earth
in 13th Chakra

Central Vertical Kathara Line
controls all other Kathara Lines. Kathara Centers on *Central Vertical Kathara Line* control all other Kathara Centers.

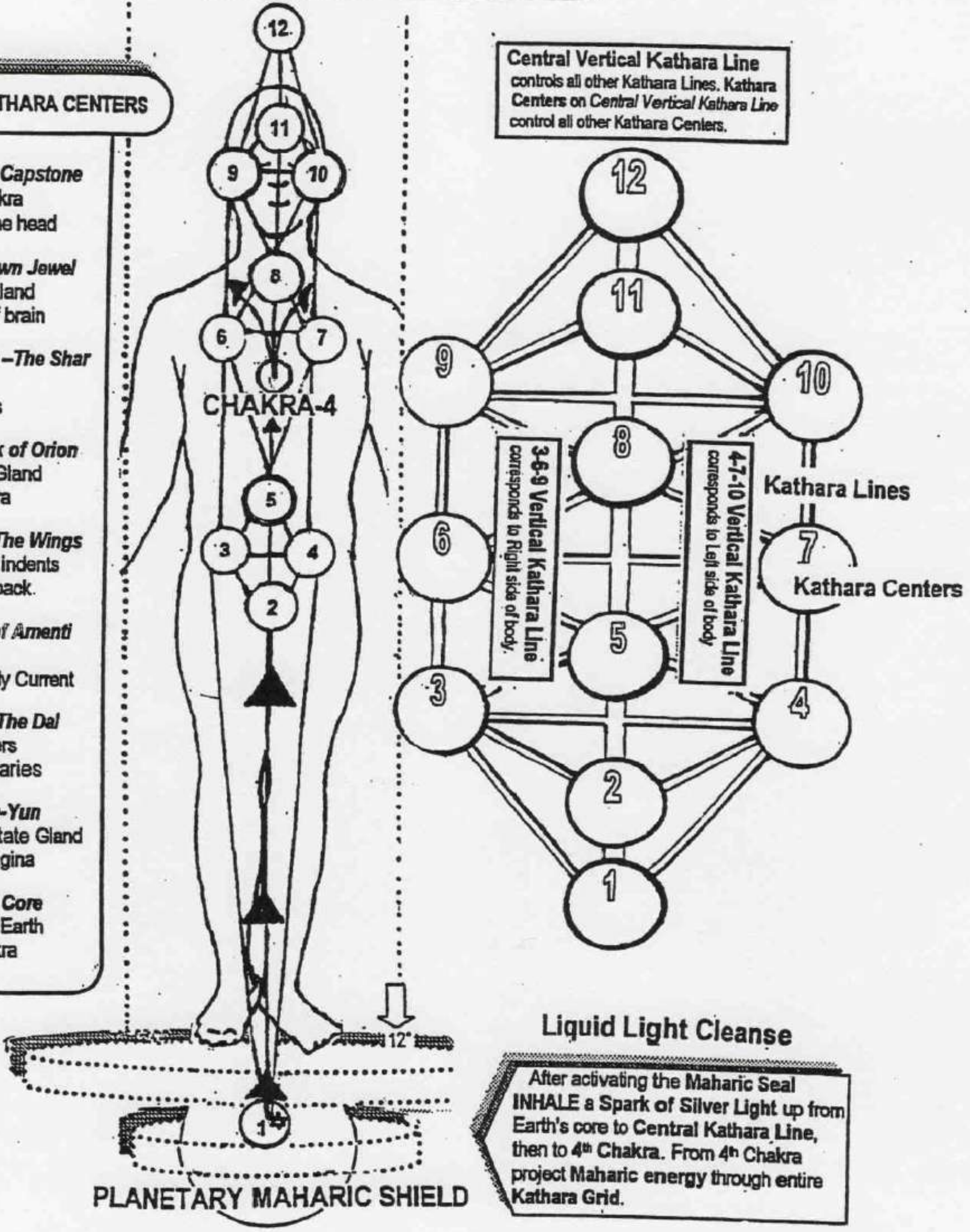
12
11
9
10
8
6
7
CHAKRA-4
5
3
4
2
12"
1
PLANETARY MAHARIC SHIELD
12
11
9
10
8
3-6-9 Vertical Kathara Line
corresponds to Right side of body.
4-7-10 Vertical Kathara Line
corresponds to Left side of body
Kathara Lines
6
7
Kathara Centers
5
3
4
2
1

Liquid Light Cleanse

After activating the Maharic Seal **INHALE** a **Spark of Silver Light** up from **Earth's core** to **Central Kathara Line,** then to **4th Chakra.** From 4th Chakra project **Maharic energy** through entire **Kathara Grid.**

The Planetary Bio-Feed Interface System and the Maharic Seal

Use of the *Maharic Seal and Liquid Light Cleanse* begins the process of activating the **8th through 12th Chakras** and opening the natural **PLANETARY BIO-FEED INTERFACE SYSTEM (PBIS)** within the human body. Through the **PBIS**, the finite quantity of scalar frequency of which the personal Maharic Shield, Hova Bodies and individual identity are composed expands to **infinite potential**, as the **Planetary Scalar Shields** can be used to open dormant **Crystal Seals** within the Scalar Shields of the personal Level-2 Kathara Grid. **Once activated through the 12th Chakra, the PBIS opens a conduit of perpetual energetic supply between the human body and the planetary Morphogenetic Field, allowing the body to continually draw in as much higher dimensional frequency as the DNA Template can hold.**

With the PBIS activated, the human body can also become a tool for reprogramming and re- balancing the Planetary Shields, Vortices, Ley Lines and energetic grids of the Planetary Templar Complex. When the Crystal Seals of the personal Shields are released, the **human DNA Template** can serve as a **bio-electric conduit** for running higher dimensional frequency into the **Planetary Shields.** The human body, with its **12-Strand DNA Template potential,** is the only mechanism on Earth, other than the Planetary Shields themselves, that is capable of running a full 12th-Dimensional **Maharic Current.** Just as within the human Manifestation Template, the Maharic Current can be used to purge distortions within the Planetary Shields, to restore the **Planetary Templar Complex** to its original **Primal Order.** Kathara Healing technologies not only provide the opportunity for individuals to facilitate healing in other *biological life forms*, they also reawaken the human potential to facilitate healing within the **geophysical planetary body.**

The Planetary Bio-Feed Interface System

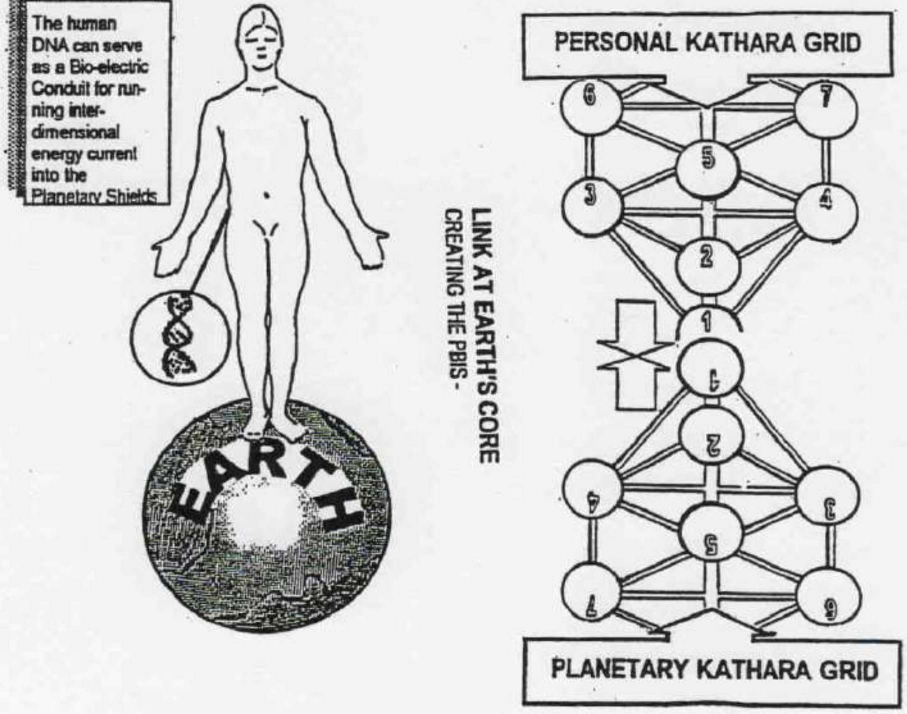

The human DNA can serve as a Bio-electric Conduit for running inter-dimensional energy current into the Planetary Shields

The Maharic Seal and Client Sessions

Through understanding the **Context of Healing**- *The Structure of the Unified Field*, the **Elements of Healing** – *The Anatomy of the Human Body Kathara Grid* and **Kathara Healing Beginning Applications** - *Morphogenetic Repatterning* through the ***Maharic Seal and Liquid Light Cleanse*** and *Awakening the Mentor* for access to 5th Dimensional awareness, the Kathara Healing Facilitator has developed the rudimentary skills necessary to begin **personal Core Holographic Template Healing. To effectively facilitate Core Template healing in others it is important for Kathara facilitators themselves to employ personal Template Healing.**

The ability to hold the **12th-Dimensional Maharic frequency sub-harmonics** (the 12th-sub-frequency band within each of the full Dimensional Frequency Bands) in body to assist in its transmission to a client for healing support increases as the facilitator utilizes the **Maharic Seal and Liquid Light Cleanse** technique **for personal healing.** The *Maharic Seal and Liquid Light Cleanse* technique not only increases the facilitator's ability to transmit 12th-Dimensional frequency sub-harmonics, it also temporarily **realigns distortions** within the *facilitator's* bio-energetic field and Kathara Grid and creates a temporary **12th-Dimensional Scalar Frequency Seal** within the *facilitator's* bio-energetic field. **Without use of the *Maharic Seal*, the facilitator will inadvertently transmit bio-energetic field and Kathara Grid distortions to the client and will energetically pick up scalar distortions** (including part of the "Karmic" Miasmic Imprint) **from the client.** (This problem occurs in all subtle-energy healing modalities that do not employ the Maharic Seal). Both of these conditions serve to amplify, rather than remedy, dis-ease and disharmony within the Mind-Body-Spirit system, for both *facilitator and client.* For these reasons, **the *Maharic Seal and Liquid Light Cleanse* technique needs to be employed prior to ANY healing facilitation work,** to protect the **bio-energetic field integrity** of facilitator and client. **The *Maharic Seal and Liquid Light Cleanse* technique represents the facilitator's preliminary preparation prior to Kathara Healing Facilitation Client Sessions.**

Safety in Healing Facilitation!

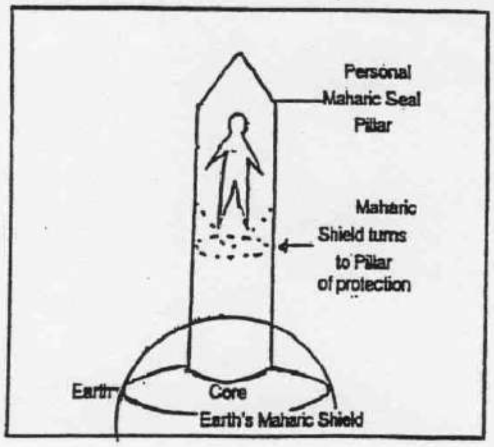

The ***Maharic Seal and Liquid Light Cleanse*** technique, used as facilitator preparation prior to **Client Sessions,** protects facilitator and client from taking on bio-energetic field distortions from each other. **Maharic Seal is the** preliminary step before any **Kathara Healing** application and is needed to effectively run interdimensional healing currents.

CO-GENERATIVE HEALING

The final phase of education necessary before engaging the adventure of **Kathara Healing Client Facilitation** is rediscovering the structure of personal **Multi-dimensional Identity** and how to use the personal **Hova Bodies** to tap the higher dimensional scalar frequencies for enhanced healing facilitation ability. The ability to understand and utilize the **PRECISE HEALING CURRENTS** of the interdimensional spectrum adds **greater power and skill** to the healing facilitation process. Unlike many other energy healing systems, **Level-1 Kathara Healing** trains the facilitator to identify, awaken within the body and employ in client service **specific combinations of interdimensional healing current** that fine-tune and amplify the beneficial effects of subtle energy healing. This chapter will conclude with an overview of the **4 PRIMARY TRIADIC HEALING CURRENTS**, 4 of the 5 **3-Dimensional Primal Energy Currents** that can be awakened within the body through activation of the **Maharic Shield.**

The natural **Primal Triadic Healing Currents**, which exist naturally as part of the personal Hova Body, Scalar Shield and Kathara Grid anatomy, create powerful, effective and user-friendly energetic healing tools when combined with the 12th-Dimensional frequencies of the **Maharic Shield.** Running the Triadic Currents is achieved by using the current to first activate the frequencies of the corresponding Hova Body and station of awareness within the personal Scalar Shields and **DNA Template.** Once activated in the DNA Template, the **Hova Body** is used to draw perpetual supply of the desired current from the Dimensional Unified Field corresponding to the chosen current. **Running the Triadic Currents requires activation of the corresponding frequencies within the facilitator's DNA Template, thus using the Triadic Currents in Client facilitation also serves to progressively stimulate and expedite activation of dormant frequencies in the facilitator's DNA Template.** Using the Triadic Currents in client facilitation serves to accelerate the processes of Holographic Recoding, higher identity Hova Body integration and multi-dimensional evolution within the Body-Mind-Spirit system of the *facilitator*, as well as that of the *client*, creating a condition of **CO-GENERATIVE HEALING** between client and facilitator.

Healing, Evolution and Ascension

In the Kathara Healing paradigm the art and science of healing the Body-Mind-Spirit system is viewed in terms of the *human relationship to the natural structure and dynamics of creation.* **Healing becomes a necessity when the intrinsic energetic systems and scalar-wave constructions of consciousness upon which the biological manifestation is built become incongruent with the natural Primal Order of the dimensionalized Time Matrix system.** The processes of natural healing and the hidden, organic functions of **human evolution** are intimately intertwined. Conditions that facilitate natural healing also facilitate fulfillment of the naturally intended course of human evolution, while conditions detrimental to Body-Mind-Spirit health are likewise detrimental to the evolutionary objectives of both the individual and the species.

The processes by which lasting healing occurs through re-genesis of Primal Order within the Scalar Shields and personal anatomy, are the same processes by which the **CONSCIOUSNESS** evolves to the fulfillment of **Spiritual Actualization**, through embodiment of the 12-dimensions of consciousness inherent to human DNA Template design. The genetic code of the biological imprint is built upon the Scalar-wave Templates of Consciousness that serve to hold the identity as energetic substance into physical embodiment. **Distortions within the physical manifestation indicate distortions in the Manifestation Template and thus constitute distortions in CONSCIOUSNESS.** Genuine healing is a *FUNCTION OF CONSCIOUSNESS*, first and foremost, and it is through realignment of consciousness itself that full healing can occur.

In the Kathara program it is understood that biology and consciousness are intimately intertwined and thus interdependent. The techniques of **Bio-Regenesis** provided in Kathara Healing assist in the processes of restoring the **Organic Imprint for Health** while simultaneously serving to **expedite the natural processes of interdimensional evolution** through DNA Template, Kundalini, Crystal Seals and Merkaba activation. **Human evolution is not a random, goal-less occurrence; it is rather an ordered system of interdimensional physics dynamics by which human consciousness progressively masters the contours of creation in time.** The evolutionary objective of human creation is **mastery,** which is progressively attained through the dynamics of **Transmutational Dimensional Ascension. Ascension** is therefore the intrinsic and indelible **goal** within human evolution, the **hidden impetus for action** that moves the human species forward in time.

Healing and Bio-Spiritual Evolution

When we begin running the higher dimensional frequencies of the **4 Primary Triadic Currents** for Kathara Healing applications, we are, in truth, running **FREQUENCIES OF CONSCIOUSNESS**, that emanate from the **Hova Bodies** in which the expanded spiritual portions of our personal identity are stationed. As we reawaken the higher dimensional frequencies of consciousness within the **DNA Template** we are progressively expediting the reintegration of personal **Multi-dimensional Identity**, serving the **spiritual goal of Ascension** through stimulating into activation the scientific dynamics of **Unified Field Physics** by which genuine Ascension occurs. **Kathara Healing** progressively restores the functions of universal physics dynamics within the human body, as they apply to the pathway of Ascension, and so Kathara Healing applications are considered to be processes of **ACCELERATED BIO-SPIRITUAL EVOLUTION**, as well as processes of holistic Body-Mind-Spirit healing. Both facilitator and client simultaneously receive the subtle but progressive benefits of evolutionary acceleration and spiritual development through Kathara Healing **Bio-Regenesis Technologies**, through which the expansion of personal consciousness into the higher dimensional aspects of spiritual awareness occurs.

Kathara practices serve to stimulate progressive integration and embodiment of the Multi-dimensional Identity by activation of dormant portions of the **DNA Template** through which the scalar-wave patterns of the higher frequencies of identity can enter the physical form. As spiritual advancement through Hova Body integration and DNA Template activation is a beneficial "side effect" of Kathara Healing practice, it is helpful for the Kathara facilitator to understand the order of **Multi-dimensional Identity Structure**. It is also helpful to realize that the **Unified Field of Energy** utilized in Kathara Healing is, in fact, a **UNIFIED FIELD OF CONSCIOUSNESS.** Healing is a function of energy dynamics and at its core, energy is the **substance consciousness adopts** to experience manifestation, thus **HEALING IS A FUNCTION OF CONSCIOUSNESS.** When personal consciousness is applied in ways that are congruent with the natural **Laws of Unified Field Physics**, the effects of healing and wellbeing are met in manifestation.

MULTI-DIMENSIONAL IDENTITY STRUCTURE

Consciousness and the Physical Illusion

In Chapter 1 we explored the **Structure of the Unified Field** and discovered the morphogenetic framework, or **Holographic Template** upon which matter systems are constructed. We learned how the structure of dimensionalization applies to the anatomy of the **15-Dimensional Human body**, from the **Core Template Kathara Grid**, the **Hova Bodies and Shields**, to the **Chakras**, **Bio-energetic Field** and **physical matter** structures built upon them. **We also discovered that *physical manifestation* is not what appears to be - that *nothing is really manifest at all*, but only *appears to be so* due to the *relationships of compartmentalized consciousness interacting with each other through the framework of dimensionalization*.** Our physical parts and our manifest world are **scalar wave constructions**, appearing solid and externalized because we view them with a level of consciousness that is stationed **one full dimension above** the wave forms we perceive as solid and objectified. Our consciousness structures itself into the scalar wave forms of dimensionality in order for us to perceive the illusion of manifestation. There is no difference, other than the **dimensional frequencies from which one is perceiving**, between the *substance that composes physical matter* and that which *composes our seemingly non-manifest portions of consciousness*. **The core of reality is CONSCIOUSNESS - it is the *energetic substance* out of which all things are made.**

Hova Bodies- Platforms of Perception

In this Section we will learn how the attributes of mind and sentient identity fit into the multidimensional anatomy of human consciousness. The scalar grids of the **Hova Bodies** and **Shields** that hold portions of our consciousness separate from each other so we may experience objectification, can be understood as **PLATFORMS OF PERCEPTION.** A portion of our conscious identity is **stationed within each of the Hova Body Shields,** and the dimensions contained within the Shield become the *platform* from which *that portion of our selves* will perceive matter objectification. **We are simultaneously focused within each Hova Body Platform of Perception, and this gives us separated but interrelated stations of identity and different LEVELS OF MIND from which we simultaneously perceive.** The amount of perception we are able to incorporate into our embodied focus on Earth is determined by the Level of **Higher Identity Integration** that has occurred. Integration of Higher Identity is the *merger* of the **Hova Bodies,** which creates changes and advancement in the **biological gene code** that allows the neurological and metabolic functions of the body to expand.

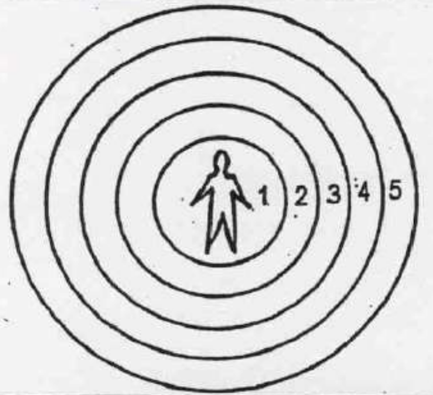

THE 5 HOVA BODIES

3-Dimensional Electro-magnetic Domains created by the Scalar Shields that serve as stations of dimensionalized consciousness

1. **NADA HOVA** Density-1
2. **ALPHI HOVA** Density-2
3. **BETCHA HOVA** Density-3
4. **MAHARA HOVA** Density-4
5. **RAJA HOVA** Density-5

Our Present Platform of Perception

Presently we are accustomed to perceiving from the Platform of Perception provided by the **NADA HOVA BODY**, the *scalar grid* that houses the 1st, 2nd and 3rd Dimensional Aspects of our identity and anatomy. While perceiving from the Nada Hova Body and the **TELLURIC SHIELD** that holds its form, we are functioning as a **3-Dimensional consciousness** and will not be aware of the perceptions of the other levels of ourselves that are perceiving from the *Platforms of Perception* from the other Hova Bodies. **As we expand and evolve we will progressively incorporate the portions of our awareness stationed in the other Hova Bodies into our earthly focus of attention.**

The *Higher Identity Structure* of the Hova Bodies and related *Levels of Mind* within dimensionalization are as follows:

Stations of Identity and Levels of Mind in the Hova Bodies

INCARNATE IDENTITY: Nada Hova Body/ Telluric Shield / dimensions 1, 2 and 3.
The Subconscious (D-1), Instinctual (D-2) and Reasoning (D-3) Levels of Mind

SOUL IDENTITY: Alphi Hova Body/ Doradic Shield/ Dimensions 4, 5 and 6
The Astral (D-4), Archetype or Dora (D-5) and the Angelic (D-6) Levels of Mind

OVER-SOUL IDENTITY: Betcha Hova Body/ Teuric Shield/ Dimensions 7, 8 and 9
The Ketheric (D-7), Monadic or Teura (D-8) and the Keriatic (D-9) Levels of Mind

AVATAR IDENTITY: Mahara Hova Body/ Maharic Shield/ Dimensions 10, 11 and 12
The Christiac (D-10), Buddhiac (D-11) and the Nirvanic (D-12) Levels of Mind.

RISHI IDENTITY: Raja Hova Body/ Rishic Shield/ Dimensions 13, 14 and 15
Universal Consciousness Mind

There are **3 additional Platforms of Perception** and 3 additional Hova Bodies beyond the structure of Dimensions;Hova Bodies that represent 3 Levels of **Ascended Mastery Identity.**

Spiritual Actualization is the process of progressive Higher Dimensional Identity Integration. ***Higher Identity Integration*** is the merger of the ***Hova Bodies,*** which creates changes and advancement in the biological ***Genetic Code*** through progressive activation of dormant portions of the ***Silicate Matrix 12-Strand DNA Template.***

5 Hova Bodies and Levels of Identity in the 15-Dimensional Time Matrix

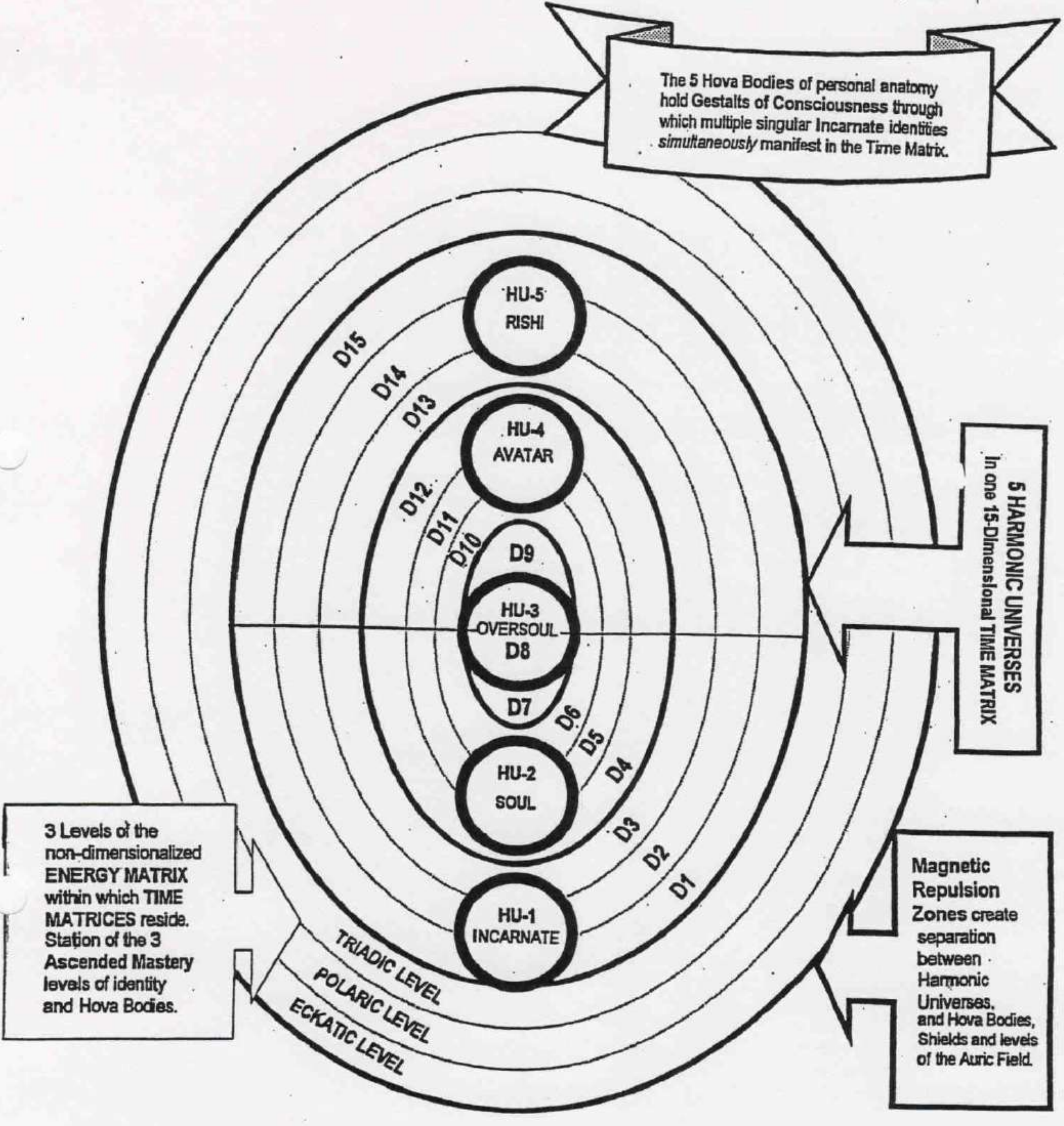
HU-5
RISHI
HU-4
AVATAR
HU-3
OVERSOUL
HU-2
SOUL
HU-1
INCARNATE
D15
D14
D13
D12
D11
D10
D9
D8
D7
D6
D5
D4
D3
D2
D1
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL

The 5 Hova Bodies of personal anatomy hold Gestalts of Consciousness through which multiple singular Incarnate identities *simultaneously* manifest in the Time Matrix.

5 HARMONIC UNIVERSES
In one 15-Dimensional TIME MATRIX

3 Levels of the non-dimensionalized **ENERGY MATRIX** within which **TIME MATRICES** reside. Station of the 3 **Ascended Mastery** levels of identity and Hova Bodies.

Magnetic Repulsion Zones create separation between Harmonic Universes, and Hova Bodies, Shields and levels of the Auric Field.

THE 4 PRIMARY TRIADIC HEALING CURRENTS

Hova Bodies and the 4 Primary Triadic Currents

Each of the Hova Bodies and Shields contains a portion of your energy-identity that exists in the form of electro-tonal impulse born of Partiki Phasing. The **energy of your higher stations of identity and that of the Universal Unified Fields of Consciousness-Energy can be drawn into the Nada Hova body and earthly incarnate identity for the purpose of healing facilitation.** There are **four Primary Triadic Healing Currents,** each corresponding to one of the **four higher dimensional Hova Bodies.** Through *your* corresponding Hova Body station you can *draw energy from the higher dimensional Unified Fields* and channel it into your Nada Hova Body for direction into healing within the first Harmonic. **The *4 Primary Healing Currents* are HARMONIC CURRENTS OF ENERGY, each composed of *3 Dimensions of Frequency*.** Each Dimensional Band within a Harmonic or **TRIADIC** current is referred to as a **PHASE** *of that current*. **In Kathara Healing we develop the ability to activate each progressive PHASE of each TRIADIC CURRENT, through grounding the current into the body via the activated MAHARIC SHIELD.**

Triadic Currents and Color

Each *Phase* of Triadic Current corresponds to a wave length in the dimensional spectrum that appears as COLOR to the embodied human mind. In directing the **Triadic Currents** we therefore use the **color translation of the primary dimension wave spectra** to differentiate between Phases of current for various applications. **The first current to be activated in Kathara Healing is the that of the DORADIC CURRENT**, the frequencies drawn in from Harmonic 2 through the **Doradic Shield and Alphi Hova Body. The Doradic Current has 3 Phases: Phase -1: 4th Dimensional current/ GREEN wave, Phase-2: 5th - Dimensional Current/ BLUE wave and Phase -3: 6th-Dimensional current/ Blue-Violet wave.** In **Level-1 Kathara Healing** we learn to activate and run the **Phase-1 Green Doradic Current.**

The 4 Primary Triadic Currents and Hova Body Correspondences

TELLURIC CURRENT: Nada Hova Body/ Telluric Shield / Dimensions 1, 2 and 3.
Phase-1 RED, Phase-2 ORANGE-GOLD , Phase-3 YELLOW

DORADIC CURRENT: Alphi Hova Body/ Doradic Shield/ Dimensions 4, 5 and 6
Phase-1 GREEN, Phase-2 BLUE, Phase-3 BLUE-VIOLET

TEURIC CURRENT: Betcha Hova Body/ Teuric Shield/ Dimensions 7, 8 and 9
Phase -1 VIOLET, Phase-2 GOLD, Phase-3 Mid-tone Silver

MAHARIC CURRENT: Mahara Hova Body/ Maharic Shield/ Dimensions 10, 11 and 12
Phase-1 BLUE-BLACK, Phase-2 SILVER BLACK, Phase-3 WHITE

TRANSMUTATION CURRENT

(can be transmitted only by 12th-level Avatars with full Mahara Hova Integration)

RISHIC CURRENT: Raja Hova Body/ Rishic Shield/ Dimensions 13, 14 and 15
Phase-1 TURQUOISE, Phase-2 PALE YELLOW, Phase-3 MAGENTA

Current Combining

When using **Triadic Phase Currents, each Phase current is** *combined* **with the current running from the activated** ***Maharic Shield*****, in** ***combination with*** **the Phase current from the** ***Harmonic ABOVE the Phase current you are using*****.** The **DORA DIC PHASE-1** current is therefore composed of: **D-11/D-12 PALE SILVER** (Maharic Shield), **D-4 GREEN** Doradic Shield Phase -1 AND **D-7 VIOLET,** the *Phase-1 current* of Harmonic 3 - Teuric Shield, *one Harmonic up* from Doradic Phase-1. In activating the **DORADIC-1 (Doradic Phase-1) current you are using SILVER-GREEN-VIOLET color to direct Dimensions 12-11-7-4 current.**

COMBINED TRIADIC CURRENTS

Used in Kathara Healing

TELLURIC

PHASE-1: **PALE SILVER** (D11&12) + **RED** (D-1) + **GREEN** (D-4)
PHASE-2: **PALE SILVER** + **ORANGE-GOLD** (D-2) + **BLUE** (D-5)
PHASE-3: **PALE SILVER** + **YELLOW** (D-3) + **INDIGO** (D-6)

DORADIC

PHASE-1: **PALE SILVER** + **GREEN** (D-4) + **VIOLET** (D-7)
PHASE-2: **PALE SILVER** + **BLUE** (D-5) + **GOLD** (D-8)
PHASE-3: **PALE SILVER** + **INDIGO** (D-6) + **SILVER** (D-9)

TEURIC

PHASE-1: **PALE SILVER** + **VIOLET** (D-7) + **BLUE-BLACK** (D-10)
PHASE-2: **PALE SILVER** + **GOLD** (D-8) + **SILVER-BLACK** (D-11)
PHASE-3: **PALE SILVER** + **SILVER** (D-9) + **WHITE** (D-12)

MAHARIC

(Requires full 12-Strand DNA Template activation & accelerated DNA Template coding)
PHASE-1: **PALE SILVER** + **BLUE-BLACK** (D-10) + **PALE TURQUOISE** (D-13)
PHASE-2: **PALE SILVER** + **SILVER-BLACK** (D-11) + **PALE YELLOW** (D-14)
PHASE-3: **PALE SILVER** + **WHITE** (D-12) + **PALE MAGENTA** (D-15)

In **Level-1 Kathara Healing** we begin preparing the body and DNA Template to receive the higher dimensional Phase Currents by initiating transmission of the **Doradic Phase-1 Healing Current**- the *Pale Silver* + *Green* + *Violet* frequencies of dimensions *12, 11, 7 and 4.*

Through the 7th-Dimensional aspect of the **Teuric Shield** and 11th and 12th Dimensional aspects of the **Maharic Shield**, the **Doradic-1 Healing Current** allows for recoding and realignment of the **1st-4th**-Dimensional aspects of the **Nada** and **Alphi Hova Bodies**, the **Telluric Shield** and **Physical, Emotional** and **Mental Bodies** of Density-1, the **D-4** aspect of the **Doradic Shield** and **Astral Body**, the **1st-4th DNA Strand Templates**, the **1st-4th Chakras**, and the **1st-4th Sub-harmonics** of all Scalar Shields, DNA Templates, Hova Bodies and Chakras. Doradic-1 Healing Current is effective in clearing the **Karmic-Miasmic imprint** from the **Density-1 Time Cycle incarnations** and the beginning of Density-2 Time Cycle incarnations.

CHAPTER 3 HIGHLIGHTS

KATHARA BEGINNING APPLICATIONS

- The **Hova Bodies** of personal anatomy contain portions of our higher dimensional awareness. One can begin accessing this awareness by building a **Perceptual Bridge** between the 3-Dimensional Self and higher identity aspects.
- Taking conscious initiative in building a Perceptual Bridge expedites **DNA Template activation** and the organic evolutionary process.
- The **MENTOR Technique** assists in building the Perceptual Bridge and activates dormant **Kathara Lines** in our **Level-1 Kathara Grid** to open communication between our D3 and Density-2, D4 and D5 levels of awareness.

- The Mentor is a part of our awareness that is composed of **3rd, 4th and 5th Dimensional frequencies** of consciousness, which exists as a quantity of **Scalar Wave patterns** that span **Densities 1 and 2.**

- The **Mentor Focus** will be used in **Kathara Healing** for **guidance, information retrieval** and **objective observation.** It can also be used to access **Cellular Memory,** and in **"Inner Child", Bi-location, Remote Viewing** and **Dream** work.

MORPHOGENETIC REPATTERNING AND THE MAHARIC SHIELD

- Through a process called **Morphogenetic Repatterning**, Kathara Healing effects the **Core** of Dis-ease and dis-harmony in the Body-Mind-Spirit system by **repatterning** the Scalar-wave Programs in the personal **Holographic Template** to their original **Primal Order.**

- In **Level-1 Kathara Healing** we learn the **MAHARIC RECODING PROCESS**- the most important level of Holographic Template realignment.

- The **Maharic Shield** is the 3-Dimensional Scalar Grid corresponding to the **10th, 11th and 12th-Dimensional frequencies** of **Density-4.** It represents the **Original Point of Manifestation** and holds our original "Liquid Light" **Hydroplasmic Pre-matter Creation Imprint**, which is the **Organic Imprint for Health.**

- In Kathara Healing, the **Maharic Shield** Organic Imprint for Health is used to progressively **repattern** the embodied **Kathara Grid** and body systems to their original **Primal Order.**

- The **12th-Dimensional Frequencies** of the Maharic Shield contain within them the **144 Sub-frequency Band** Scalar patterns of **ALL** of the Dimensions below.

- 12th-Dimensional **Maharic Frequency** can be used to **effectively correct distortions** in the lower dimensional levels of the Manifestation Template, Scalar Shields and Kathara Grid.

- Activation of the **Maharic Shield** serves as the **RESET BUTTON** for the embodied **Kathara Grid** and for 12 Dimensions of Body-Mind-Spirit anatomy.

THE MAHARIC SEAL AND LIQUID LIGHT CLEANSE

- In bio-energetic healing, **Level-1 Kathara Grid** distortions override realignments made to the **outer levels** of the bio-fields alone. For **lasting healing** to occur, Level-1 Kathara Grid distortions must be realigned.

- The **Maharic Shield**, with its **144 Frequency Sub-Harmonics**, is the only portion of the Morphogenetic anatomy that can fully repattern and realign the **Level-1 Kathara 12-Tree Grid.** In **Kathara Healing**, the **Maharic Shield** is used to realign the Level-1 12-Tree Grid.

- The **Maharic Seal and Liquid Light Cleanse** technique is used to restore distorted **Partiki Phasing Sequences** and **Scalar-wave arrangements** within the **3 Levels of the Kathara Grid** and **Scalar Shields.**

- Realignment will be **temporary** until the Kathara Grid reaches **"Critical Mass"** of energy in the **new**, realigned pattern.

- Until **critical mass** is reached, the outer levels of the anatomy will continue to **reverse** the Kathara Grid healing. With practice and **consistent application** of the **Maharic Seal and Liquid Light Cleanse** technique, the realignment of the Kathara Grid will become **permanent.**

- In future applications of Kathara Healing processes will be used to realign the **outer levels** of the bio-field and the **Kathara Grid simultaneously,** to expedite the healing process.

- The Focus of **Level-1 Kathara Healing** is **Morphogenetic Repatterning** and **Holographic Template Recoding** via the **Maharic Shield**, to clear distortions and the **Karmic-miasmic imprint** from the personal **Scalar Shields, Kathara Grid** and **DNA Template,** utilizing the **accelerated energies** of the *Planetary Bio-feed Interface System* **(PBIS).**

KATHARA POINTS

- **Kathara Points** are mini energy-vortex points in the body that connect directly to **Diodic Point vortices** in the **Level-3 Kathara Diodic Point Grid.**

- Stimulation of the Kathara Points through **manual massage** effects the **movement of energy** within related areas of the Kathara Grid and body systems.

- Used in sequence, **Sequence-1 Kathara Points for Regeneration** *open* energetic gateways to the **Kathara Lines** of the **Kathara Level-1 12-Tree Grid**, to allow new, realigned scalar patterns from the **Maharic Shield** to run through the Kathara Lines and Kathara Grid for **Core Template Repatterning.**

- **Sequence-1 Kathara Points** create **more lasting but less immediate** effects of realignment and healing. They can be **used alone** for **deep healing** and re-balancing of the **Core Kathara Grid.**

- **Sequence-2 Kathara Points for Revitalization** *close* **Kathara Lines** in the 12-Tree Grid and **channel** energy into the **Nadial Lines** of the **outer bio-energetic field** for more rapid revitalization and re-balancing of the **outer auric levels, chakras** and **physical cells.**

- **Sequence-2 Kathara Points** create **more immediate but less lasting** effects of realignment and healing. They can be **used alone** to create **temporary revitalization of** the **outer levels** of the bio-system.

- **Used together, Sequence-1** Kathara Points can run **Maharic energy** through the **core Kathara Grid** for **Template Recoding. Sequence-2** Kathara Points can then be used to **direct the Maharic energy** into the **outer levels** of the bio-field and physical body, to **expedite healing** through **simultaneous** inner and outer bio-field realignment.

✳ In **Kathara Healing, Sequence-1** Kathara Points are used to **OPEN** the Kathara Lines **prior to** running the **Maharic Seal and Liquid Light Cleanse** and **Sequence-2** Kathara Points can be used **following** the procedure to **revitalize** the outer bio-field and body systems.

THE MAHARIC SEAL, PBIS AND HEALING

- Use of the **Maharic Seal and Liquid Light Cleanse** technique begins activation of dormant **Chakras 8-12** and the **Planetary Bio-feed Interface System (PBIS).**

- **Opening the PBIS** allows energy from the **Planetary Shields** to be drawn into the **human body** for **rapid activation** of the **personal Maharic** Shield and **Crystal Seals,** to **expedite** the processes of **DNA Template activation, healing, recoding of the Holographic Template** and interdimensional evolution.

- Opening of the PBIS can also allow the human **DNA Template** to serve as a **Bio-electric Conduit,** through which the **Planetary Scalar Shields** and Electro-magnetic Grids can be realigned with their **Primal Order** for **geophysical planetary healing.**

- Through the PBIS humans can directly facilitate healing of the **Planetary Templar Complex** from the Manifestation Template level, as well as facilitate healing in other biological organisms.

- Each use of the **Maharic Seal and Liquid Light Cleanse** technique assists in **planetary healing**, as the process progressively stimulates the **Planetary Maharic Shield** into activation, to realign **Planetary Shield** distortions.

- The **Maharic Seal and Liquid Light Cleanse** technique is used as **preliminary preparation** for **ALL Kathara Healing applications.** It creates a temporary **12th-Dimensional Scalar Frequency Seal** in the facilitator's bio-field to **protect** both facilitator and client from taking on each other's bio-field distortions.

- In **healing applications**, the **Maharic Seal** is needed to run interdimensional healing currents to their most effective **12th-Sub-Harmonic level.**

BIO-SPIRITUAL EVOLUTION, ASCENSION AND HEALING

- The processes by which **healing** occurs through re-genesis of Primal Order are the same processes through which true **Spiritual Actualization** takes place.
- **Biology** and **consciousness** are **intimately intertwined** through the natural **Laws of Universal Unified Field Physics**, the **energetic dynamics** through which consciousness enters manifestation.
- The hidden **impetus for action** behind human evolution is the core species objective of **Mastery of Co-creation** within the space-time system. This **Evolutionary Objective** is obtained through achievement of the Evolutionary **Goal** of **Transmutational Dimensional Ascension.**
- **Kathara Healing** applications generate **healing** through restoring the **Original Imprint for Health.**

- Restoration of the **Original Imprint for Health** simultaneously expedites the natural processes of **evolution,** through activation of the **12-Strand DNA Template, Kundalini** energies, and **Merkaba Vehicle** and integration of the **Hova Bodies, Scalar Shields** and higher dimensional aspects of personal **spiritual identity.**

- **Kathara Healing** facilitates healing while simultaneously facilitating **Spiritual Actualization** and expediting the Goal of **Transmutational Dimensional Ascension.**

- **Healing is a function of consciousness.** There is no difference between the substance that composes physical matter and the substance that composes consciousness, other than the **dimensional frequency** from which one is perceiving.

- When personal consciousness energy is applied in **congruence** with the natural Laws of Universal Unified Field Physics, **healing, wellbeing** and **Spiritual Actualization** result.

MULTI-DIMENSIONAL IDENTITY STRUCTURE

- **Human anatomy** is a **15-Dimensional Scalar-wave construct.** Part of our identity exists **simultaneously** within the Time Cycles of the higher dimensional **Harmonic Universes.**

- We are indelibly **linked** to our higher identity levels through our embodied **Kathara Grid, Scalar Shields** and **DNA Template.** As we evolve biologically to activation of dormant portions of our DNA Template and Scalar Shields, the conscious awareness of our higher identity levels **progressively embodies.**

- The 3-Dimensional Hova Body **Electro-magnetic Domains** and Scalar Shields that form them represent **Platforms of Perception** through which our higher identity levels perceive.

- **Merger** of the **Hova Bodies** via activation of the **DNA Template** and **Scalar Shields** creates **reintegration** of the **Multi-Dimensional Identity Structure.**

4 PRIMARY TRIADIC HEALING CURRENTS AND IDENTITY

- The **5 Levels of Multi-Dimensional Identity Structure** are: **Incarnate**-Density-1, **Soul**-Density-2, **Oversoul**-Density-3, **Avatar**-Density-4 and **Rishi**-Density-5. There are also **3 Ascended Master** identity levels, Hova Bodies and Shields existing **beyond** the dimensionalized **Time Matrix.**

- In **Kathara Healing** we use the various levels of higher dimensional identity to draw in **healing current** from the higher dimensional Unified Fields, by activating the portions of the **personal Scalar Shields, Kathara Grid** and **DNA Template** that correspond to the higher dimensional **Hova Bodies.**

- There are **4 Primary Triadic** (3-Dimensional) **Energy Currents** that can be used in **Healing Facilitation**: The **Telluric Current**-Density-1, the **Doradic Current**-Density-2, the **Teuric Current**-Density-3 and the **Maharic Current**-Density-4. The **Rishic Current** of Density-5 can be run only by beings with **full 12-strand DNA activation.**

- Each of the **4 Primary Triadic Healing Currents** corresponds to one of the first 4 **Hova Bodies** of human anatomy.

- In Kathara Healing, the 4 Primary Triadic Currents are harmonically interwoven to create 4 powerful **Combined Healing Currents.**

- Each of the 4 Triadic Healing Currents has **3** Phases, or **3 Phase Currents,** each representing one of the three dimensional frequency bands within the **Triadic Current.**

- Each of the **12 Phase Currents** is run in combination with the 11-12th-Dimensional **Maharic Current** and the Phase Current from the **harmonic directly above,** to form a powerful **Triadic Healing Current.**

- In **Kathara Healing,** specific Healing Currents are directed by using the **Color** of their dominant **wave-length.**

- **Level-1 Kathara Healing** teaches us to activate our **Doradic-Phase-1 Current**- the **Pale Silver** (D11-12) **+ Green** (D-4) **+ Violet** (D-7)- for **Healing Facilitation** and **personal** DNA Template and Shield activation.

- The **Doradic-Phase –1 Current** can realign the **1st-4th** Dimensional aspects of the **Scalar Shields**, the **1st-4th Chakras, Axi-A-Tonal Lines** and **DNA Strand Templates**, the **Physical, Emotional** and **Mental** bodies of the **Nada Hova Body** and the D-4 **Astral body** of the **Alphi Hova Body.**

- **Doradic-Phase-1 Current** can also be used to realign the **1st-4th Sub-Harmonics** of ALL Shields, DNA Strand Templates, Axi-A-Tonal Lines, Chakras and Hova Bodies and to clear the **Karmic-Miasmic Imprint** from the **Density-1** incarnational Time Cycles and the **first part** of the **Density-2** incarnational Time Cycles.

CHAPTER 4

GENERAL KATHARA HEALING FIELDWORK

Skills to Facilitate Healing

- OPENING THE HEALING CHANNELS- DORADIC-PHASE-1 CURRENT, THE DORADIC SPHERE, DORADIC CORDS AND DORADIC BALLS.
- THE KARMIC-MIASMIC IMPRINT, INCARNATIONAL TIME VECTORS, VECTOR IMBEDDING.
- MULTI-VECTOR HOLOGRAPHIC RECODING THROUGH THE MAHARIC INFUSION.
- ATTITUDES AND ETIQUETTE IN HEALING, THE MOTIVATION TO HEAL, POSTURES OF LOVE, THE FRAMEWORK FOR HEALING AND FULLY ENGAGING THE LOVE VIBRATION.
- THE 6 STEP MAHARIC RECODING SESSION, MELCHIZEDEK CLOISTER ORDINATION AND BECOMING A PROFESSIONAL SPIRITUAL HEALING FACILITATOR-INSTRUCTOR.

TECHNIQUES:

#3- OPENING THE HEALING CHANNELS- DORADIC –PHASE-1 CURRENT

#4- MULTI-VECTOR HOLOGRAPHIC RECODING KATHARA SCAN

#5- THE MAHARIC INFUSION

THE MAHARA SPACE COMFORT TOUCH

TOPICS OF DISCUSSION

OPENING THE HEALING CHANNELS

DORADIC-PHASE-1 CURRENT

Interdimensional Currents and Signet Star Crystal Seals

Activation of any **Triadic Healing Current** requires temporary release of the corresponding **Signet Star Crystal Seals** within the personal **Scalar Shield** and a temporary activation of the portions of the dormant **12-Strand DNA Template** that correspond to the Signet Seals. Signet Seals modulate the flow of frequency between the 3-Dimensional Harmonic Density Levels. When Signet Seals are in their usual (for this time period) dormant phase, they keep the **5 Hova Bodies** of the personal 15-Dimensional anatomy separate from each other by creating an electro-magnetic barrier or **Magnetic Repulsion Zone** between each 3-Dimensional level of the energetic anatomy. This *Magnetic Repulsion Zone* between embodied **Harmonics of Manifestation** manifests as **polarity resistance to frequency inflow** within the 3 Levels of the Kathara Grid, the Scalar Shields, Axi-A-Tonal Lines, Chakras and Auric Field levels of the embodied **Inner Templar Complex. In order to run higher dimensional frequency currents through the physical body for healing applications, it is necessary to temporarily release portions of the Magnetic Repulsion Zones between the Hova Bodies.** Releasing the electro-magnetic resistance to higher dimensional frequency inflow within the Inner Templar Complex is achieved through temporary activation of the **Signet Seals** within the **Kathara Level-2 Crystal Seals Grid** of the personal **Scalar Shields**.

Most contemporary energetic healing systems rely upon using the 1st, 2nd and 3rd-Dimensional frequencies of the **TELLURIC PHASE CURRENTS** in the **Nada Hova Body**. Through the Nada Hova Body alone, only the 3-Dimensional **Sub-Harmonics** of higher dimensional frequency currents can be run, as the dormant *Signet Seals* block the faster moving *Partiki Phasing Flash-line Sequences* of higher dimensional frequency bands from running through the DNA Template. **If a frequency band is blocked from running through the DNA Template that frequency cannot be energetically transmitted through the Nada Hova Body and physical bio-electrical systems. In conventional energy healing methods**, even those emerging from ancient systems of subtle-energy healing, the **Signet Seals**, and thus the *Scalar Shields, Hova Bodies* and *DNA Template* remain dormant in their usually **closed** position, **phase-locked** into a 3-Dimensional orientation. While the DNA Template remains phase-locked into the 3-Dimensional orientation of the Nada Hova Body, only the **1st through 12th Sub-frequency bands** (Sub-Harmonics) of **1st through 3rd Dimensional Current** (Telluric Current), and the **1st through 3rd Sub-frequency bands** of **4th through 12th Dimensional Current** can be run through the human body for transmission. Use of the lower frequency (slower Partiki Phasing rhythm) **Telluric Currents** can effect **temporary** beneficial change in the **outer layers** of the bio-energetic field, but Telluric Currents alone cannot facilitate full repatterning of the **3 Levels of the Kathara Grid** where **Scalar Shield** distortions that cause **genetic malfunction** are embedded. The **11th-12th-Dimensional frequencies** of the Pre-matter Hydroplasmic **Maharic Currents** are the only dimensional frequencies powerful enough to affect change in the embodied **Kathara Grid**. To run **Maharic Current**, and the high frequency interdimensional **Triadic Phase Currents**, through the human body requires release of the **Signet Seals within the personal Scalar Shields.**

Triadic Phase Currents, the Maharic Shield and the PBIS

Most healing systems have not incorporated the **Maharic Current** into their frequency application procedures because the **Planetary Templar Complex** has not been able to run a D-12 **Maharic Current** through the planetary grids for **210,216 years.** Like the human **DNA Template**, if the **Planetary Shields** and Manifestation Template are blocked from processing interdimensional frequency, that frequency will be **unavailable for use** on the planet. If the **Signet Star Crystal Seals** (Star Gates) in the **Planetary Shields** are dormant, closed and phase-locked into Density-1 orientation, no one on the planet will be able to draw interdimensional frequency into their **personal Shields** because the frequency is unavailable within the **Planetary Shields**. The personal and planetary Templar Complexes are **intimately interwoven**, and what effects one will inevitably effect the other.

In order to run the **full** spectrum of interdimensional **Triadic Phase Currents** through the human body it is necessary to open the **Signet Seals** in the personal **Inner Templar** and the **Signet Star Gates in the Planetary Templar Complex**. In conventional human evolution, the internal Signet Seals, which serve as the **Genetic Time Codes** or **"Fire Codes"** of the human **DNA Template**, automatically, progressively and sequentially open over **very long cycles of time**. During the slow activation of the Genetic Time Codes the conscious identity experiences passage through numerous incarnate lifecycles within the planetary Time Cycle, evolving through a series of **Single Vector incarnations** before being released from the 3-Dimensional phase-lock of Density-1 manifestation. One can **manually open the Signet Seals** in the embodied Inner Templar Complex to **expedite evolution** into the Multi-dimensional, **Multi-Vector Time Cycles** through use of the 11th-12th Dimensional **Maharic Current**.

Activation of the personal **Maharic Shield** allows for rapid release of the embodied **Signet Seals** and activation of the **Fire Codes** in the DNA Template. The **Maharic Current** from the personal Maharic Shield can be used to release the **Signet Star Crystal Seals** within the body, which in turn activates the dormant **Kundalini** energies at the base of the spine and the **Cranial-Sacral Kundalini Seals** within the **Tailbone** and **Pineal Gland.** Release of the Cranial-Sacral Seals of the Internal Templar Complex allows the **Magnetic Repulsion Zones** between the **Hova Bodies** to release so the DNA Template, personal Scalar Shields and physical body systems can receive inflow of higher dimensional frequency. To gain conscious mastery over the process of releasing the internal Signet Seals requires activation of the personal **Maharic Shield**, and the personal Maharic Shield can be **manually activated** only through drawing **Maharic Current** into the personal Shields via the **Planetary Maharic Shield.** **The entire process of releasing the internal *Signet Seals* by activating the personal *Maharic Shield*, in order to enable the *DNA Template* to run *Triadic Phase Currents* for healing is entirely dependant upon the *operational function* of the *Planetary Bio-feed Interface System* (PBIS).**

If the **Planetary Templar Complex** is blocked, or the Signet Star Gates are dormant, and **Maharic Current** from the **Planetary Maharic Shield** is not able to run through the planetary grids, humans on Earth **do** not have the option of **manually activating the personal Maharic Shield**, because the **PBIS** is not operational. The **PBIS** becomes **operationally functional** within the human body through activation of the **12th Chakra**. The *12th Chakra* cannot activate in the personal Templar Complex if the **Planetary Templar Complex** cannot run 12th-Dimensional **Maharic frequency.** For over **200,000** years, the Planetary Templar Complex could not run Maharic frequency, and thus manual activation of the personal Maharic Shield and release of the embodied Signet Seals to open the Hova Bodies for running interdimensional Triadic Phase Currents for healing were not available options for Earth humanity. Since the recent activation of the **Planetary Templar Complex, Maharic Shield** and **PBIS** activation and **Kathara Healing** applications are now available on Earth.

Opening Triadic Phase Currents in the Human Body

The method by which **Triadic Phase Currents** are activated within the human body is a **progressive process.** First, opening of the personal **Kathara Grid, Telluric Shield** and **Nada Hova Body** is orchestrated through activation of the **Maharic Shield** and **Signet Seals**, to allow inflow of frequency from the higher dimensional **Hova Bodies**. Next, the incoming frequency is harnessed and **collected** within the **Chakra System.** Finally, the collected frequency is then **transmitted** for healing facilitation in others, either through direct **Chakra transmission** or by close-range **Frequency Induction** via transmission of frequency from the activated **Secondary Chakra Centers** within the **PALMS.**

The ability to run the higher dimensional Triadic Phase Currents within the body must be built up within the Internal Templar Complex, starting with the **slowest-moving current** of the higher dimensional spectrum- the **DORADIC-PHASE-1 CURRENT.** The *Doradic-Phase-1 Current* utilizes **the sub-frequencies** of the **11th-12th-**Dimensional **Maharic Current** from the Density-4 *Mahara Hova Body* as a **carrier wave** that will carry the desired interdimensional currents through the **Signet Seals** of the personal **Scalar Shields.** The **full spectrum** of the **4th** Dimensional **Doradic-Phase-1** current from the Density-2 *Alphi Hova Body* is activated as the **primary healing frequency** for transmission. The **full spectrum** of the **7th-**Dimensional **Teuric-Phase-1 Current** (one Harmonic above the Doradic-Phase-1 Current) from the Density-3 *Betcha Hova Body* is used to release the 4th-Dimensional **Signet Seal,** allowing the 4th-Dimensional Doradic-Phase-1 current to enter the **Nada Hova Body.**

Activating the **4th Signet Seal** triggers **temporary** activation of the **4th DNA Strand Template** and corresponding **4th Kathara Center** in the Level-1 12-Tree Grid, which allows the 4th-Dimensional frequencies of the **Doradic-Phase-1 Current** to pass through the **active DNA** and into the bio-energetic and bio-electrical systems of the body for transmission. Once the DNA Template becomes accustomed to holding a **4th-DNA Strand Template Activation,** after **several months** of practice in running the 4th–Dimensional *Doradic-Phase-1 Current,* the body will be able to begin receiving the primarily 5th-Dimensional **Doradic-Phase-2 Current. The lower dimensional Triadic Currents must be activated within the DNA Template before the higher dimensional Triadic Currents can run.** In **Level-1 Kathara Healing** the first step of activating the full spectrum (12 sub-harmonics) of the 4th-Dimensional **Doradic-Phase-1 Current** is achieved.

Triadic Phase Currents and Color Cap-Stones

The **Color Cap Stone** of a dimensional frequency band is the visual representation of the **primary wave-length** formed by the specific **Partiki Phasing Rhythm** for that dimension. A specific dimensional frequency band can be chosen and directed by the **color translation** of its primary wave-length. The **Color Cap Stones** of the dimensional frequency bands are used in Kathara Healing to direct specifically selected dimensional frequency bands (Partiki Phasing Rhythms) for specific applications. The **Color Cap Stones** for the **Doradic-Phase-1 Current** are **PALE SILVER** (D-11-D-12 Maharic) + **GREEN** (D-4 Doradic) + **VIOLET** (D-7 Teuric). In **technique # 3 – OPENING THE HEALING CHANNELS-DORADIC-1** - the ***Pale Silver* + *Green* + *Violet*** Cap Stone Colors will be used to activate the **Doradic-Phase-1** Healing Current within the body.

DORADIC SPHERE, CORDS AND BALLS

The Doradic Sphere and Current Collection

The **DORADIC SPHERE** is a **frequency repository** that is formed to hold the activated **Doradic-Phase-1 Current** within the **Chakra system** for transmission in healing facilitation. As the Doradic-Phase-1 Current is primarily a **4th-Dimensional Current**, the **Doradic-1 Sphere** will be formed in the **4th Chakra**, which corresponds to the 4th-Dimensional **ASTRAL body** of the Alphi Hova Body and Auric Field. **Transmission** of the Doradic-Phase-1 Current will take place from the **Doradic Sphere** at the **4th Chakra**, creating a **4th-Dimensional ASTRAL INFUSION** of frequency into the bio-energetic field of the client in **close-range healing**, and a general **Astral Plane Infusion** of Doradic-Phase-1 frequency into **Manifestation Templates** of areas(and everything contained within the space) when transmitted directly from the **4th Chakra.**

Doradic Cords and Current Induction

The **DORADIC CORDS** are **energy conduits** that are formed by opening the embodied **Axi-A-Tonal Lines** within the **arms, hands and fingers.** Opening the Doradic Cords allows the **Secondary Chakras** within the **palms** of the hands to carry the **Doradic-Phase-1 Current** for application in **close-range** transmission. Though Triadic Phase Currents can be transmitted directly from the **Chakra Center** in which the **Collection Sphere** is placed, the energies are more **diffused** and **diluted** in this form of transmission, unless the personal DNA Template holds a **full, permanent activation** of the Strand Template corresponding to the Triadic Current being transmitted. It takes time, often years, for the DNA Template to acquire and permanently hold a higher strand activation level. One-on-one healing facilitation is more effective when conducted through **concentrating** the Triadic Current through the **Axi-A-Tonal Lines** for **DIRECT INDUCTION** of current through the **Palm Chakras** in **close-range**transmission. Close-range transmission also allows for **more precise application** of frequency to specific areas of the body systems. In **Level-1 Kathara Healing**, the **Doradic Cords** are activated within the arms, palms and fingers to allow for concentration, focused direction and **Direct Induction** of the **Doradic-Phase-1 Current.**

For **general healing** purposes, such as **"clearing rooms"** or geographical spaces, or when attempting to transmit healing current into **groups of people**, transmission of the **Doradic-Phase-1 Current** can be orchestrated directly from the **4th Chakra** for **NON-SPECIFIC INDUCTION.** In **Non-Specific Induction**, following activation of the Doradic-Phase-1 Current, each **INHALE** breath is used to fill the **Doradic Sphere** with more energy and each **EXHALE** breath is used to expand and transmit the energy into the room or bio-fields of a group of people. It is not necessary to activate the Doradic Cords for Non-Specific Induction applications. In **Level-1 Kathara Healing** applications use of the **Doradic Cords** and **Palm Chakras** is employed for Direct Induction close-range transmission in Client Sessions. To enhance and expedite **personal healing**, the activated **Doradic-Phase –1 Current** can be run through the personal **Level-1 Kathara Grid** with the ***Maharic Shield and Liquid Light Cleanse*** technique and then through each of the **7 Primary Embodied Chakras.**

Doradic Balls and Palm Transmission Direct Induction

The **DORADIC BALLS** are small energy repositories that form in the **Palms** of the Hands when the **Doradic Cords** have been activated to run Doradic-Phase-1-Current through the **Palm Chakras.** Doradic Balls are used to **accumulate current** in the **Palm Chakra vortices** prior to transmission, to **strengthen** the intensity of the current.

PALM CHAKRAS AND TRANSMITTING CURRENT

Primary Polarities of Palm Chakras

Along with the **15 Primary Chakras** within human anatomy there are many other smaller **Secondary** and **Tertiary Chakra Centers** throughout the body, which correspond to the energetic interrelationships of scalar-wave patterns within the **Kathara Level-3 Diodic Points Grid**. All Chakras operate as sets of **Black and White Holes**, through which energy from the Dimensional Unified Fields circulates into and out of the body and Manifestation Template. The **Palm Chakras** are part of the **Secondary Chakra System**. Like all Chakras, the Palm Chakras are structured as **2 vortices** of energy **joined at the tip** by a minute **Crystal Seal** within the **Kathara Level-2 Crystal Seals Grid**. The Palm Chakras run through the **center of the Palms** and out of the **back of the hands.**

In any set of Chakra vortex spirals, one vortex represents the **magnetic Black Hole**, a *counter-clockwise rotating* spiral that **draws energy IN** toward the central Crystal Seal, and the companion vortex represents the corresponding **Electrical White Hole,** the *clock-wise rotating* spiral that **sends energy OUT** from the central Crystal Seal. The **Black Hole** vortex represents the **RECEIVING spiral** or **Magnetic Pole** and the **White Hole** vortex represents the **TRANSMITTING spiral** or **Electrical Pole.** The **Electrical Pole** of a Chakra represents the **EXPANSION CURRENT** of the Chakra, through which it **expands energy out** from the personal Shields; it is referred to as **BASE ELECTRICAL**. The **Magnetic Pole** of a Chakra represents the **CONTRACTION CURRENT** of the Chakra, through which it **contracts energy in** from the external Unified Field to the personal Shields; it is referred to as **BASE MAGNETIC.**

The Chakras in the **Palms** are composed of many smaller minute **vortex sets**, each with Magnetic and Electrical Poles, but the **PRIMARY POLE** of each Palm Chakra is that pertaining to the **larger Black-White Hole set** and **Crystal Seal**. Though each Palm Chakra is capable of both **Transmitting and Receiving** energy current, the **strongest and most focused currents** can be achieved by utilizing the natural polarity orientation of the Primary Vortex Set. The Chakra in the **RIGHT PALM** is **BASE ELECTRICAL**; its corresponding vortex on the back of the right hand is **Base Magnetic**. The Chakra in the **LEFT PALM** is **BASE MAGNETIC**; its corresponding vortex on the back of the left hand is **Base Electrical**. When running Triadic Phase Currents it is helpful to understand the **Primary Polarities of the Palm Chakras** in order to create the **strongest** and **least distorted** healing current. In **Kathara Healing**, the **RIGHT HAND** is the **TRANSMITTING HAND**, through which Triadic Current is **Exhaled** down the **Cords** in the arms and **EXPANDED** into the **RIGHT PALM CHAKRA** for **TRANSMISION**. The **LEFT HAND** is the **RECEIVING HAND**, through which the Triadic Current transmitted from the Right Hand is **drawn back** into the facilitator's body **TO CREATE A CLOSED CIRCUIT OF CURRENT** for the **strongest current intensity** and **purest frequency content.**

HAND POSITIONS

Standard Hand Position for Transmitting Triadic Current

When transmitting Triadic Current, the **LEFT PALM** is **PLACED OVER THE BACK OF THE RIGHT HAND**. The *Base Magnetic* Current of the *Left Palm* **interfaces** with the *Base Magnetic* Current in the *Back of the Right Hand*, creating a **MAGNETIC REPULSION ZONE**, which **blocks the vortex on the Back of the Right hand** from drawing **non-specific frequency** from the Density-1 Unified Field. This keeps the current transmitted from the *Right Palm Chakra* **pure** to its specifically chosen content. The *Base Magnetic* vortex in the *Left Palm* simultaneously draws the **transmitted current** back up through the smaller vortex sets after it runs through the client's body, creating a **closed circuit of pure frequency** running between the **Hova Bodies** of the **client,** the **facilitator** and the higher dimensional **Unified Fields**. Options of **Frequency Modulation** for other healing applications also become available through using this **HAND POSITION** when transmitting healing current. In some applications a **modulated current** that includes bursts of Density-1 Telluric Current interspersed with current transmission from the higher dimensions is desired; this can be achieved by moving the Left Hand on and off of the vortex on the back of the Right Hand. The **LEFT OVER RIGHT HAND POSITION** is used in **Level-1 Kathara Healing** applications.

Elements of Doradic-Phase-1 Current Activation

Doradic-1 Cap Stone Colors:

PALE SILVER (D-11&D-12) + GREEN (D-4)+ VIOLET (D-7)

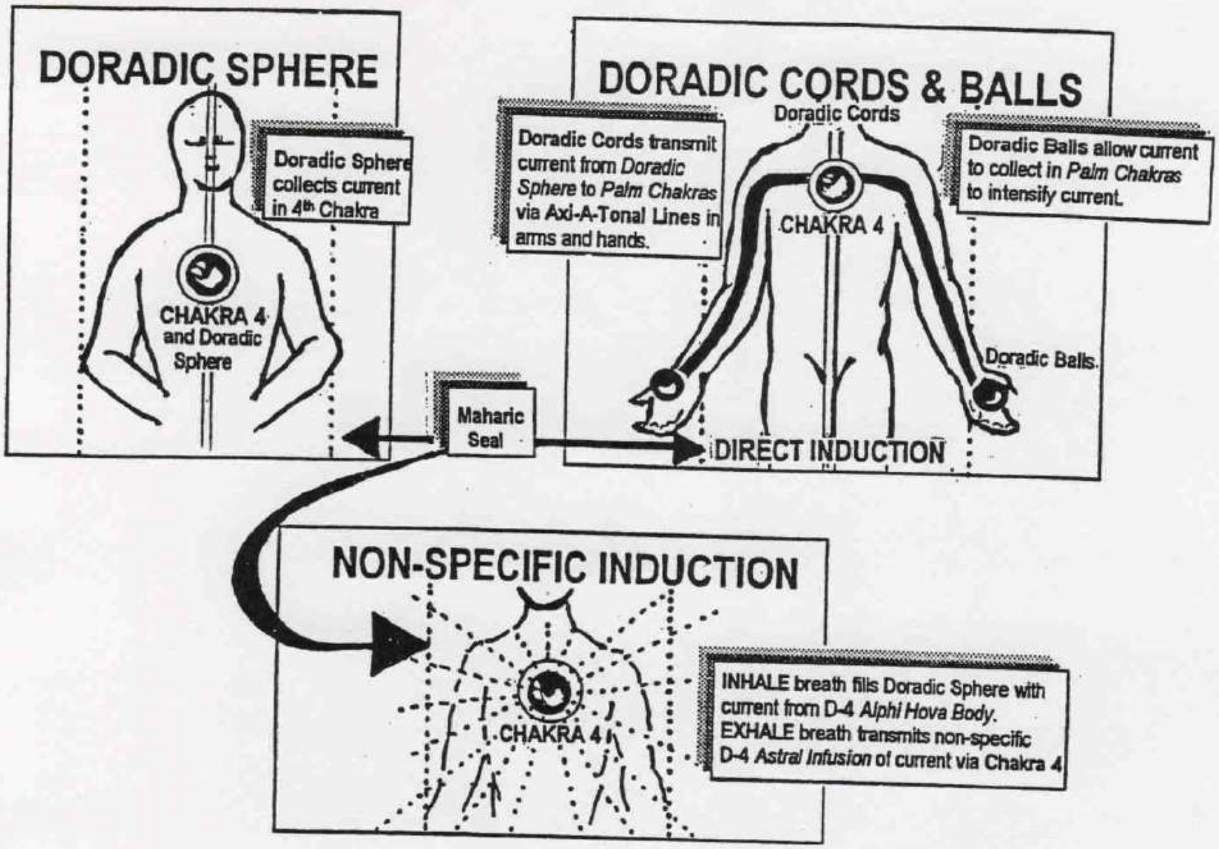
DORADIC SPHERE
CHAKRA 4 and Doradic Sphere
DORADIC CORDS & BALLS
Doradic Cords
CHAKRA 4
Doradic Balls
Maharic Seal
DIRECT INDUCTION
NON-SPECIFIC INDUCTION
CHAKRA 4

Doradic Sphere collects current in 4th Chakra

Doradic Cords transmit current from *Doradic Sphere* to *Palm Chakras* via Axi-A-Tonal Lines in arms and hands.

Doradic Balls allow current to collect in *Palm Chakras* to intensify current.

INHALE breath fills Doradic Sphere with current from D-4 *Alphi Hova Body*. EXHALE breath transmits non-specific D-4 *Astral Infusion* of current via Chakra 4

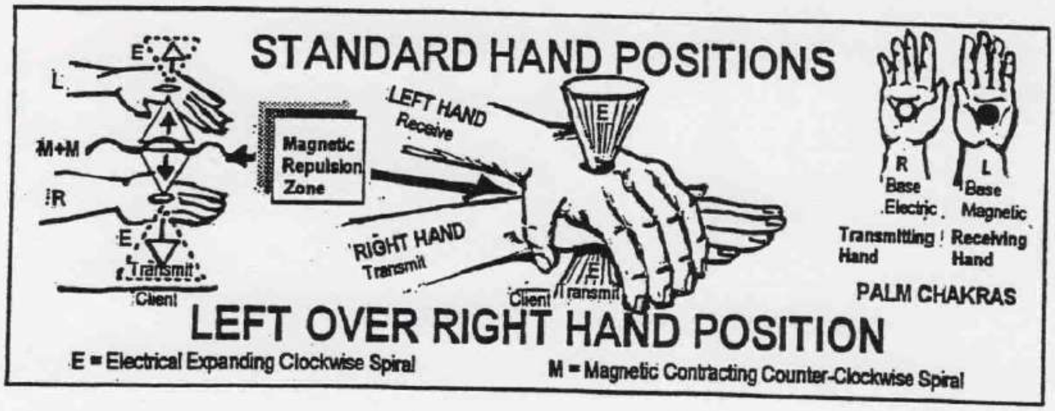
STANDARD HAND POSITIONS
E
L
M+M
R
E
Transmit
Client
Magnetic Repulsion Zone
LEFT HAND Receive
E
RIGHT HAND Transmit
E
Client
Transmit
R
Base Electric
L
Base Magnetic
Transmitting Hand
Receiving Hand
PALM CHAKRAS
LEFT OVER RIGHT HAND POSITION

E = Electrical Expanding Clockwise Spiral

M = Magnetic Contracting Counter-Clockwise Spiral

Intrinsic Process of Activating a Triadic Phase Current

ACTIVATE MAHARIC SHIELD

1. Open the PBIS using the D-11 – D-12 Wave Guide Symbol Code, the **Hierophant Symbol**, to open the **12th Chakra** and stimulate release of **Maharic frequency** from the **Planetary Maharic Shield.**
2. From the PBIS draw Maharic frequency from the Earth's Maharic Shield **up** into the dormant personal Maharic Shield 12" below the feet, to activate **the personal Maharic Shield.**

ACTIVATE TRIADIC PHASE CURRENT

3. Choose the specific Triadic Current you desire to run.
4. Use the inflow of **Maharic Current** from the personal and planetary Maharic Shield as a **Carrier Wave** to carry the **Cap Stone Color frequencies of the Signet Star Crystal Seals** in your Scalar Shields that correspond to the **Triadic Phase Current** you wish to activate.
5. Temporarily **release** the **Signet** and **Cranial-Sacral Kundalini Seals and** activate the **DNA Strand Template** corresponding to the chosen **Triadic Phase Current**, by running the **Primary Frequency** and the frequency **one Harmonic above it** through the **Signet Seals**, on the Maharic carrier wave current, to open the **Nada Hova Body** and physical body systems to inflow of the higher dimensional **Triadic Phase Current.**

COLLECT AND TRANSMIT HEALING CURRENT

6. As the Triadic Phase currents transmits into the body through the Signet Seals, Kundalini energies and DNA Template, **collect and build the current** as a **Sphere** of energy within the **Chakra** corresponding to the Primary Triadic Current you are running.
7. Run the current from the **Collecting Chakra** through **the Axi-A-Tonal Line Cords** in the arms, to activate the **Palm Chakras.**
8. **Exhale** current from the **Chakra Collection Sphere** through the **Cords** and allow it to build into **Balls** of energy within the **Palm Chakras** before transmitting.
9. Transmit the Triadic Current through the **RIGHT PALM CHAKRA** into client's bio-energetic field while drawing more current into the **Chakra Collection Sphere** with each **exhale** breath and draw the current back through the **LEFT PALM** on each **inhale** breath to create a higher dimensional Closed Circuit Healing Current. Use **LEFT PALM OVER BACK OF RIGHT HAND** hand position.

TECHNIQUE # 3

Opening the Healing Channels- Doradic-Phase-1 Current

1. Breathe slowly and deeply for several breaths while visualizing or imagining a **4" Violet Dot** on the center of the *inside of your forehead*. Make the color image as strong as you are able.(If you can't yet perceive inner colors or visualization, simply imagine the Violet Dot as being there). Move the Violet Dot to the center of the brain at the **Pineal Gland**, and then upward into the 7th **Crown Chakra**, from the inside of the head moving upward and outward until the Violet Dot *rests on top of your head* at the Crown Chakra.

2. Call to mind the **Hierophant Symbol** (Pale Silver Merkaba Star formed by combining 2 Merkaba Stars). Visualize the spinning Hierophant stationed 12" below the feet within your **Maharic Shield** and slowly **INHALE**, drawing the Hierophant up through the Central Body Current, up to meet the Violet Dot at the 7th **Crown Chakra** at the top of the head. When the *Hierophant connects with the Violet Dot* visualize the Violet Dot bursting into a **4" Vertical Pillar of Violet-Silver Light**. Imagine the *Violet-Silver Pillar* running down through the body into the *Earth Core* and upward to the 14th **Chakra** 36" above the Head. You have activated the **Violet Ray of Transmutation.**

3. Move your attention to the 4th **Heart Chakra** and imagine a *small Green Spherical Crystal* – the **Green Star**- stationed at its core. This is the *Green Star Crystal Seal* that controls release of *4th Dimensional* frequency – *Phase-1 Doradic Current* - into the body. **INHALE**, drawing Violet-Silver energy from the Violet-Silver Pillar into the *Green Star*. **EXHALE**, *expanding* the Violet-Silver energy *into the Green Star sphere. Repeat 2 more INHALE-EXHALE* breaths, drawing Violet-Silver Energy from the Pillar and expanding it into the Green Star sphere. At the end of the last **EXHALE**, imagine the **Green Star popping**, turning into a spinning Green and Violet Starburst of Light, as the *Violet-Silver Pillar disappears. Spin the Green-Violet Starburst* faster until it becomes a *ball of Green-Violet Light*, about the size of a *grapefruit*, spinning in the center of the *4th Heart Chakra*. The Green-Violet ball is the store-house for your Phase-1Doradic Current – it is called the **Doradic Sphere. INHALE** and draw Pale Silver energy from your **Maharic Shield**, through the Central Body Current and into the center of the **Doradic Sphere** at the 4th **Heart Chakra**, encasing the Sphere in translucent Pale Silver Light.

4. **EXHALE quickly** and push **2 Cords** of *Green-Violet-Silver energy* – **Doradic Cords** -from the Doradic Sphere *down the inside of your arms*, one Cord through each arm. As the ends of the Doradic Cords reach the inside of the hands, notice a sensation of *mild resistance* as the Cords push against the *etheric membranes* covering the *small chakras* in the palms. **INHALE** a full breath, drawing Silver energy up from the **Maharic Shield**, into the *Doradic Sphere* at the 4th Chakra. **EXHALE forcefully**, pushing more energy down through the *Doradic Cords* in the arms, and feel this energy **"pop"** out of the *palms, releasing the membranes* on the palm chakras. Continue to **INHALE** energy up from the *Maharic Shield* into the *Doradic Sphere* and **EXHALE** it through the *Doradic Cords, building the strength of the Doradic Current* in both palms. Imagine the Doradic Current *pooling in the palms* to form *2 Green-Violet balls of light*, about the size of *tennis balls*. Energize the Doradic Sphere and Cords until you can sense the smaller **Doradic Balls** formed in the *palms* of your hands. **Your *Doradic Current* is now activated for use in healing facilitation.**

SYMBOL SHORTHAND

TECHNIQUE # 3: Opening the Healing Channels Doradic-Phase-1 Current

1. **4" Violet Dot** inside forehead, move to **Pineal Gland** then up through **7th Crown Chakra** to **top of head.**
2. **Hierophant Symbol** at center of your **Maharic Shield** 12" below feet. Inhale Symbol **up** to **Violet Dot** at top of head. See **Violet Dot burst** into **4" Violet-Silver Pillar** running through body to **Earth's core** and up to **14th Chakra** 36" above head.
3. Small **Green Star Crystal Seal at 4th Heart Chakra.** Inhale drawing **Violet-Silver** energy from **Violet-Silver Pillar** into **Green Star Crystal Seal** and see **Green Star Crystal Popping** into a spinning **Green-Violet Starburst,** as Violet-Silver Pillar **disappears.** Spin Green-Violet Starburst faster until it becomes grapefruit-size **BALL** of **Green-Violet Light**, the **Doradic Sphere,** at the center of the **4th Heart Chakra**. Encase Green-Violet Sphere in **Pale Silver** Maharic energy.
4. Push a **Green-Violet-Silver Doradic Cord** down each arm to **Palm Chakra Membranes** and draw **Pale Silver** up from **Maharic Shield** to **Doradic Sphere** and forcefully push Maharic Current **down arms** to **pop Palm Chakra Membranes.** While running **Doradic-1 Current** draw energy from **Maharic Shield** to **Doradic Sphere** and push it down arms from **Doradic Sphere** through **Doradic Cords** to **pool** into **Doradic Balls** at **Palm Chakras.**

TECHNIQUE # 3: Opening the Healing Channels
Doradic-Phase-1 Current

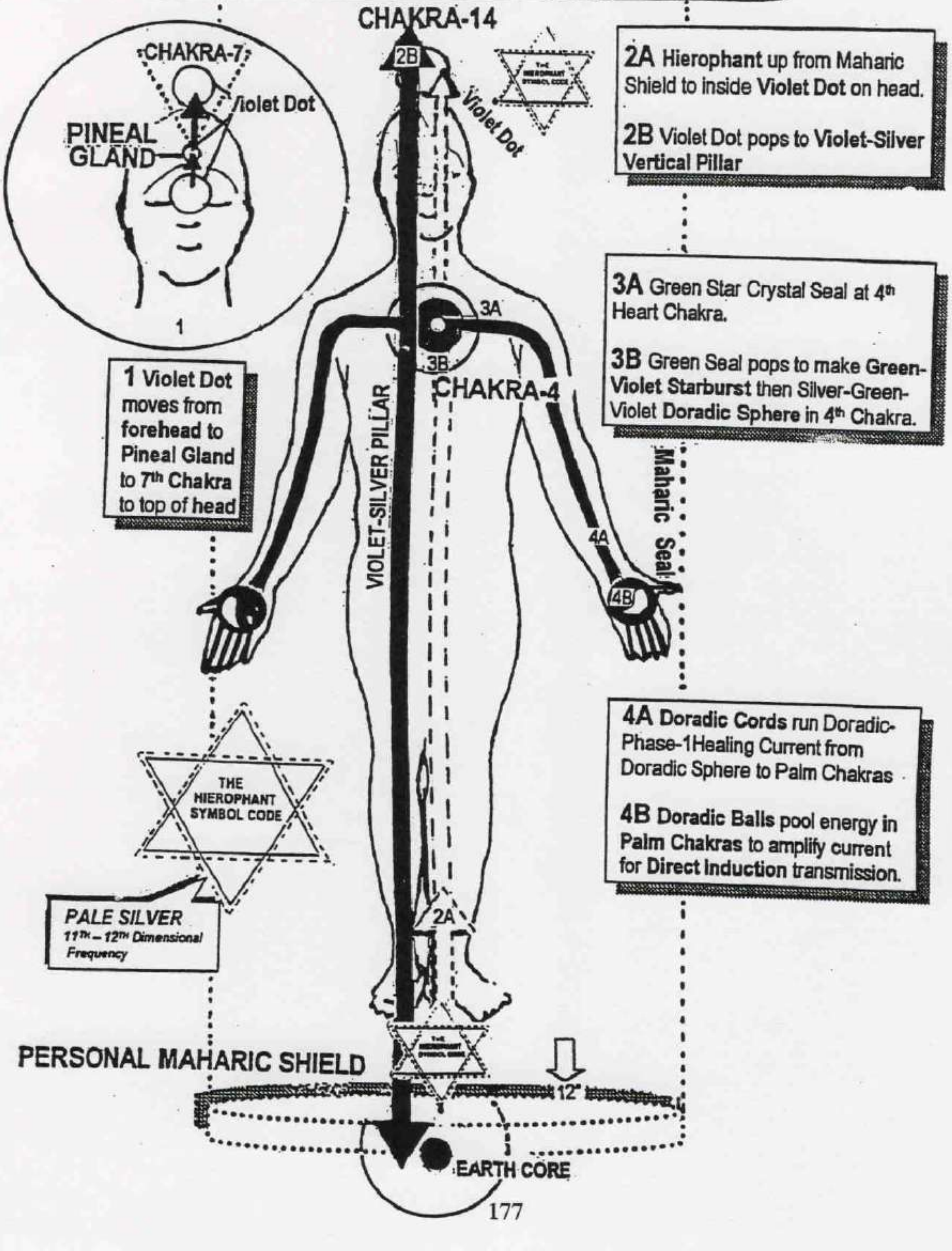

2A Hierophant up from Maharic Shield to inside **Violet Dot** on head.

2B Violet Dot pops to **Violet-Silver Vertical Pillar**

3A Green Star Crystal Seal at 4th Heart Chakra.

3B Green Seal pops to make **Green-Violet Starburst** then Silver-Green-Violet **Doradic Sphere** in 4th Chakra.

1 Violet Dot moves from **forehead** to **Pineal Gland** to **7th Chakra** to top of head

4A Doradic Cords run Doradic-Phase-1Healing Current from Doradic Sphere to Palm Chakras

4B Doradic Balls pool energy in **Palm Chakras** to amplify current for **Direct Induction** transmission.

PALE SILVER
11TH – 12TH Dimensional Frequency

CHAKRA-14
CHAKRA-7
Violet Dot
PINEAL GLAND
1
2B
Violet Dot
THE HIEROPHANT SYMBOL CODE
3A
3B
CHAKRA-4
VIOLET-SILVER PILLAR
Maharic Seal
4A
4B
THE HIEROPHANT SYMBOL CODE
2A
PERSONAL MAHARIC SHIELD
12"
EARTH CORE

THE KARMIC-MIASMIC IMPRINT

THE NATURE OF KARMA

Karma, Identity, the Christos and the Physics of Divine Right Order

When an identity enters the **Time Matrix** to begin **simultaneous cycles of evolution** through time, it enters the manifest fields through first adopting a **Manifestation Template** composed of Partiki Units and Scalar Waves. The first Manifestation Template is called the **Rishic Shield**; it exists at the **Ante-matter** scalar-wave level of **Density-5** and holds the complete **Primal Scalar Pattern** for all of the **Singular Identities** that will **simultaneously** manifest into incarnation within the lower dimensional Densities and Time Cycles. The **Rishic Shield** represents the **Identity Gestalt** or **Family of Consciousness** through which identity enters manifestation. The Rishic Shield carries the precise configuration of Scalar-wave patterns, the **Primal Order** of energetic substance, through which the consciousness individuated to form a collective of consciousness within the dimensionalized Time Matrix. Every individual incarnate from the Rishic Family of Consciousness in time carries as part of its core **Kathara Grid** scalar program the **electromagnetic energy signature** of **Primal Order** carried by the **Rishic Shield** through which it manifested.

The Scalar-wave design of the Rishic Shield represents **DIVINE RIGHT ORDER**, the **Primal Order** of scalar-wave energetic relationship and **Partiki Phasing Sequences** that individual identities in time must re-evolve to hold within their embodied **Manifestation Templates** in order to **regain freedom** from the dimensional systems of the Time Matrix. Consciousness from the Rishic Shield, in the form of **Omni-polar wave spectra**, enters the manifest Densities by projecting portions of itself into **smaller Consciousness Gestalts** within the lower dimensional fields. The smaller Gestalts of Consciousness form the **3-Dimensional Scalar Shields** of the identity within the dimensionalized matter systems. **Each Scalar Shield holds the consciousness of numerous simultaneous Cycles of Incarnation within each of the 4 Primary Time cycles of Densities 1 through 4.** Each individual incarnate in time is directly linked to its original **Rishic Shield Family of Consciousness** through the **Density-4 Maharic Shield.** The Maharic Shield of dimensions **10-12** contains the first **Pre-matter Hydroplasmic scalar-wave spectra** through which identity began its experience within the holograms of Space, Time and Matter. The **personal Maharic Shield,** often called the **PERSONAL CHRISTOS IMPRINT**, holds the original **Primal Order** of scalar waves through which manifest consciousness links to its **Rishic Shield.** The **Christos Imprint** of the **Maharic Shield** represents the **PATH OF DIVINE RIGHT ORDER** through which each individual consciousness in time can **re-evolve** to its original Eternal Non-manifest State of Wholeness in Consciousness beyond the Time Matrix. **Divine Right Order is not only a spiritual concept; it is a reality in energy within the scientific dynamics of Universal Unified Field Physics.** Divine Right Order represents the **Primal Order** of scalar-waves that make up the original **Manifestation Template** from which individual **singular lifetimes** emerge, and it is the necessary pattern of energetic order within the personal multi-dimensional Manifestation Template which allows incarnate beings to fulfill the **evolutionary objective** of **Transmutational Dimensional Ascension. The result of an incarnate being's *Manifestation Template* and Conscious Awareness becoming mis-aligned with the energetic structure of *Divine Right Order* is called KARMA.**

Karma, Divine Right Order, the Scalar Shields, Free Will and Miasms

If it is understood that the universe and all things within it exist as **manifestations of consciousness** in the form of **precisely ordered** scalar wave arrangements, and that there is an intrinsic and indelible **Primal Order** upon which all manifestation functions, it will not be difficult to conceptualize the **literal and tangible nature** of the **Karmic Imprint**. In energetic terms, **Divine Right Order** represents the **natural alignment** of conscious energy fields within the intrinsic **Laws of Universal Physics** characteristic to the dimensional structures of time. Divine Right Order exists as a **tangible, specific** and **harmonious organization** of **energy interrelationships** within the Primal Order of energetic structure that creates dimensionality. The condition frequently referred to as **Karma** is, like Divine Right Order, a tangible, specific organization of energy interrelationships. Unlike Divine Right Order, *Karma* represents dis-harmonious interrelationships of energy within the context of the Primal Order that creates dimensionality. **Karma** can be viewed as **CHAOTIC, incoherent energetic disorganization**, whereas **Order** represents **coherent energetic** organization.

In terms of the **15-Dimensional anatomy** of the **personal Manifestation Template**, the **Karmic Pattern** represents **areas of incoherent energetic disorganization** within the structure of the personal **Scalar Shields**. The **12th-Dimensional** Level of the personal **Maharic Shield** holds the original **Christos Imprint**, the **Primal Order** *perfect organization* of scalar-wave configurations that links our consciousness and **DNA Template** to the **Rishic Shield Family of Consciousness** through which we entered the **Time Matrix**. While the Maharic Shield holds our original **Perfect Organization of Consciousness**, the Scalar Shields upon which our lower dimensional **Hova Bodies and Incarnate, Soul and Over-Soul Manifestation Cycles** emerge often become **distorted** from their original **Primal Order** as held within the **Maharic** and **Rishic Shields**. Human consciousness, in its multidimensional forms, was designed **"in the image of its Creator"**, having **Free Will** to *apply the dynamics of energetic creation in time* in accordance with **personal intention**, within the greater context of the fixed **Natural Energetic Laws of Creation**.

When we use our gift of **Free Will** to direct our energies in a manner **congruent** with the **Divine Right Order** of **Universal Unified Field Physics** (which implies co-creative, not competitive co-evolution for all), the scalar-wave design of our personal **Scalar Shields** remains organized upon the intrinsic design of universal and personal **Primal Order**. Through **embodiment** of Divine Right Order within our personal **Scalar Shields** and consciousness, we know conditions of **perfect health and harmony** within the Time Matrix. When our Free Will is used to direct our energies in ways that are **incongruent** with and non-supportive of **Primal Order**, the **disorganization** of our energetic orientation manifests as literal **Partiki Phasing Sequence distortions** that create **misshapen scalar-wave groupings** within the anatomy of the personal **Kathara Grid**, **Scalar Shields** and the **DNA Template** built upon them. These misshapen scalar-wave groupings within the Kathara Grid, Scalar Shields and DNA Template are called **MIASMS** and they hold, in **tangible scalar frequency** form, the **disorganized energetic patterns** of our **Free Will choices** through which we have used our energies in a manner **incongruent** with **Divine Right Order**.

***Miasms* are the *tangible substance* of our *Karmic Pattern* that reflect the areas in which our *consciousness* and *Scalar Shields* are *distorted* from our original *Maharic Shield Christos Imprint* of *Perfect Organization* within the context of *Universal Primal Order*.**

TIME VECTORS AND THE KARMIC-MIASMIC IMPRINT

Miasms, Incarnational Time Vectors, Vector Imbedding and the DNA Template

The **Miasmic Imprint** within the body and Manifestation Template that holds the personal **Karmic pattern** manifests as **mathematical, geometrical** and **electro-magnetic distortions** in scalar-wave forms within the Kathara Grid and Scalar Shields, that cause groupings of **Anti-particles** to build up within the **Particle field** of the physical and energetic anatomy. The build-up of **Anti-particles** within the Particle Field blocks the natural flow of energy and consciousness through the Kathara Level-3 **Diodic Points Grid**, and manifests as **malfunctions** within the **Partiki Phasing Programs** within the **Kathara Centers** of the Level-1 **12-Tree Grid** and within the **Signet Star Crystal Seals** of the Kathara Level-2 **Crystal Seals Grid.** The **3 Levels of the Kathara Grid** set the blueprint for manifestation of the **Scalar Shields, Hova Bodies, Chakra System** and **DNA Template. Miasmic distortions** in the **Kathara Grid** effect **every level of personal manifestation,** from the operation of the **physical genetic code** to the ability of consciousness to successfully embody within the physical form, to the potentials of **Transmutational Dimensional Ascension** and attainment of **Spiritual Actualization.**

Karma*- the *Miasmic Imprint*- is a *biophysical phenomenon* of scalar-wave distortion that directly effects the *human DNA.

The **Incarnations of SELF** that **simultaneously manifest** into incarnational cycles of divergent space-time placement within the 4 Densities and 12-Dimensions of Matter are all **directly interconnect** through the **Kathara Grid, Scalar Shields** and **DNA TEMPLATE.** The **Free Will choices** made by other manifestations of ourselves **concurrently existing** in other space-time locations and different manifest body forms, effect the function of our personal **Kathara Grid, DNA Template, conscious awareness** and **life experience.** The **choices WE make** in our present lifetime directly effect the **Kathara Grid, DNA Template, consciousness** and **life experience** of **our other-time incarnations.** The energetic connection between incarnations of the SELF is an immediate, **living connection** of energy and consciousness. Because our numerous simultaneous incarnations in time are directly, energetically linked to each other via the Kathara Grid and DNA Template, we literally **inherit Miasmic Buildup from each other** on a continual basis. When we make the choice to restore our personal **Imprint for Health,** our **Christos Imprint,** as held within our **Maharic Shield,** part of these efforts must involve **clearing the Miasmic Imprint** that is bleeding through into our personal Manifestation Template for the other incarnations of Self.

Each of our simultaneous incarnations in time represents one **SINGLE VECTOR CONSCIOUSNESS,** one portion of our **Eternal Identity** that experiences 3-Dimensional **linear evolutionary progression** along **ONE TIME VECTOR** (one cycle of Partiki Phasing Sequences). **Each Incarnate represents a TIME VECTOR.** The multiple Time Vectors of our Multiple Simultaneous Selves are literally **encoded** within core scalar-wave programs of our **DNA Template;** this is called **VECTOR IMBEDDING.** To fully heal our Body-Mind-Spirit system by restoring it to its **Organic Imprint for Health,** the **Maharic Shield Christos Imprint,** we need to clear the Miasmic Imprint from our other simultaneous incarnations as it is presently embedded within our DNA Template. **Transmuting the Miasmic Imprint that is embedded in our DNA Template to restore the original perfect Primal Order is called "Clearing the Karmic Pattern"** or **MULTI-VECTOR HOLOGRAPHIC RECODING.**

MULTI-VECTOR HOLOGRAPHIC RECODING

REGENESIS OF THE PERSONAL CHRISTOS

Restoring the Organic Imprint for Health and Clearing the Karmic-Miasmic Imprint

Literally all conditions of **dis-ease** in the human body and dis-harmony in the life experience, including the unnatural condition of **death** of the human body form, result from **Miasmic Distortions** within the embodied **Kathara Grid, Scalar Shields, DNA Template** and personal **Inner Templar Complex.** The natural state of **Universal Primal Order** is harmonious co-creative interrelationship of consciousness. Humans can progressively return to their natural condition of **Ordered Perfection** through **re-genesis**, or *re-creation*, of the **Organic Imprint for Health – the Maharic Shield Christos Imprint** – within the Kathara Grid blueprint of the personal Manifestation Template. **Regenesis of the Personal Christos is a Multi-Vector Endeavor**, involving clearing the karmic-Miasmic Imprint from a myriad of various simultaneous selves presently evolving in the Time Matrix. The **hard way** to clear the Miasmic Imprint is by **"walking the Karma"**, which means allowing progressive levels of the **Miasmic Imprint** to repeatedly **re-manifest** in the personal life experience, until the **proper application of Free Will choice** used in congruence with **Divine Right Order** becomes consciously apparent and the **"Right Choices"** are made to re-ordered the disorganized use of energy that caused the Miasm to form. In our present time, this approach to "clearing the Karmic Pattern" is no longer effective as a learning tool by which embodied consciousness teaches itself the etiquette of creation. **The Miasmic Imprint within the collective human species has grown so extensively that the core issues of misapplied Free Will that cause Miasmic Buildup can no longer be consciously identified, and the Miasmic Imprint continues to compound and amplify itself as the same errors of Free Will energy use are perpetually repeated and reinforced.**

The Karmic-Miasmic Imprint is an **ENERGY FORM**; a literal **construct of scalar-wave frequency** with its corresponding **Anti-particle Manifestation.** As a form of energy, the **Miasmic Imprint** can be dealt with **AS ENERGY, WITH ENERGY**, for the most expedient clearing, healing and realignment of the **Karmic Imprint.**

The **Karmic-Miasmic Imprint** manifests within the **causal level** of manifestation, the **3 Levels of the embodied Kathara Grid of the personal Internal Template Complex**, and from the Kathara Level the Karmic-Miasmic Imprint manifests as **misalignment** and **malfunction** within all levels of Body-Mind-Spirit system. To clear the Karmic Imprint **AS ENERGY, WITH ENERGY**, its disorganized scalar-frequency patterns must be realigned from the **Core Template Kathara Grids**, the **causal level** of miasmic manifestation. Restoring the **Maharic Shield Christos Imprint** to the personal **Kathara Grid** takes time and **precise applications of frequency re-ordering** in all levels of the Body-Mind-Spirit system. The first step to **Regenesis of the Personal Christos** is the **MAHARIC RECODING PROCESS.**

MULTI-VECTOR HOLOGRAPHIC RECODING

Recoding the Time Vectors of the DNA Template and the DNA Fire Codes

The process of regenerating the **Christos Imprint** within the **Kathara Grid** of the human body is a **MULTI-VECTOR** undertaking. A **Vector** is a **set cycle of vibrational-oscillation rhythm** that creates a **specific repeating Partiki Phasing Rhythm** and **Flash Line Sequence.** Each individual incarnate self in time holds within the electro-magnetic programs of its embodied **Kathara Centers** the precise **Partiki Phasing Rhythms** that correspond to its particular placement within the planetary **Time Cycles.** The **Time Vector,** or **Flash Line Sequence** that creates one linear **Event Horizon,** is literally encoded and actively operational within the **Kathara Grid** and **DNA Template** of each incarnate identity. In the process of restoring the **Organic Integrity** to the personal Manifestation Template, the dormant **Partiki Phasing Rhythms** that correspond to one's **other-time incarnates** are progressively **brought into activation** and **realigned** with the Christos Imprint of the **Maharic Shield.**

The **Partiki Phasing Flash Line Sequence programs** within the embodied **Kathara Centers** manifest within the **DNA Template** as the **GENETIC TIME CODES,** or **Fire Codes,** of the DNA. In restoring **Maharic Order** to the personal **Kathara Grid,** the **12 Dormant Fire Codes** of the DNA Template will be slowly but progressively brought into **activation** through the 11th-12th –Dimensional Frequencies of the **Maharic Shield Mahara Hova Body Current.** Activation of the **DNA Fire Codes** creates a reciprocal release of the **Signet Star Crystal Seals** and **Cranial-Sacral Kundalini Seals** within the **Kathara Level-2 Crystal Seals Grid,** progressive activation of the dormant **Kundalini energies** and **Merkaba Vehicle Phases,** merger of the **Hova Bodies** and embodiment of the higher dimensional **Spiritual aspects** of conscious awareness. The process of recoding the multiple time vectors within the DNA Template – **Multi-vector Holographic Recoding-** is a process of progressive **WHOLE-BEING HEALING** that is achieved through **consistent application** of the **MAHARIC RECODING PROCESS,** and expedited through complimentary processes of **DNA Template, Kundalini** and **Merkaba Vehicle** activation. Level-1 Kathara Healing introduces the operational mechanics of the **Maharic Recoding Process.**

The Maharic Recoding Process

FOR MULTI-VECTOR HOLOGRAPHIC RECODING

The 4 Primary Elements of Action

1. **The Maharic Seal and the Liquid Light Cleanse** (Technique # 2)
2. **Opening the Healing Channels** (Technique # 3)
3. **Amplifying Inner Audio-Visual and Kathara Scan** (Technique # 4)
4. **The Maharic Infusion** (Technique # 5)

The 4 Primary Elements of Action to begin the **Maharic Recoding Process,** through which **Multi-Vector Holographic Recoding** of the **Morphogenetic Field Manifestation Template** is achieved. Begins realignment of the Core Manifestation Template **Kathara Grid Levels,** the **Scalar Shields, Hova Bodies, Axi-A-Tonal Lines** and **DNA Template** with the **Primal Order** perfect scalar-wave pattern of the personal 12th-Dimensional Pre-matter Hydroplasmic **Maharic Shield-** the **Christos Imprint** and **Organic Imprint for Health.** Also begins activation of dormant **DNA Strand Templates, Kundalini** energies and the **Merkaba Vehicle** for **Transmutational Dimensional Ascension.**

Bio-Regenesis of the Christos Imprint
and 12-Strand DNA Template
within the human body.

TECHNIQUE # 4

AMPLIFYING INNER AUDIO-VISUAL AND KATHARA SCAN

1. Place **RIGHT palm** (*electrical-transmitting hand*) over the **5th Kathara Center** at the **Navel**, and place the *Left palm* (*magnetic-receiving hand*) over the *back of the right hand* (to block 3rd-Dimensional energies from entering the rear of the right hand chakra). **INHALE** energy from your **Maharic Shield** – the *Maharic Carrier Wave*- into your **Doradic Sphere** at the *4th Heart Chakra*. **EXHALE** and run the *Maharic Wave* down the *Doradic Cords* in the arms and into the *Doradic Balls* in the palms. Inhale, then **EXHALE**, using the exhale breath to *push* energy from the *Doradic Sphere*, down the *Doradic Cord* in the **RIGHT** arm, through the **RIGHT Doradic Ball** and into the **5th Kathara Center** at the *Navel*. Run several breaths of *Doradic-1 Current* into the *5th Kathara Center* at the Navel, until a **mild tingling sensation** is felt.

2. Place the **LEFT palm** over the **5th Chakra** at the *Throat*, with the *Right palm* covering the *back of the LEFT hand*, and **INHALE** the Doradic-1 Current up from the *5th Kathara Center* at the Navel, into the Seed/Center of the **5th Chakra**. **EXHALE** and *expand* the Doradic-1 frequency into the *5th Chakra spiral*, allowing the magnetic current from the LEFT hand to draw the energy out through the spiral. As you **EXHALE**, slowly **move the hands forward**, away from the body and the 5th Chakra.

3. Repeat the same process over the **6th Kathara Center**, placing the **RIGHT palm** over the *6th Kathara Center* below the **Right Shoulder**, with the *Left palm* covering the back of the *RIGHT hand*. **INHALE**/Draw energy up from the *Maharic Shield*, into the *Doradic Sphere*. **EXHALE** energy down the *Doradic Cords* into the *Doradic Balls*. Inhale, then **EXHALE** energy down the *RIGHT Doradic Cord* through the *RIGHT palm* and into **Kathara Center 6.** Run several breaths to energize the *6th Kathara Center*. Then move **LEFT palm** over the *6th Third Eye Chakra* in the forehead, with Right palm behind. **INHALE** the Doradic-1 Current from the *6th Kathara Center* into the Seed/Center of the *6th Chakra* and **EXHALE** the energy into the 6th Chakra spiral while drawing it through with the magnetic current from the LEFT hand. The **6th Sense-*Audurea***, which provides **inner audio** perception, and the **7th Sense-Tristet**, which provides **inner visual** perception are now energized and amplified in preparation for **Kathara Scan**.

4. Close your eyes and focus your attention in a **single point** at the *inside of your forehead*. Inhale, and as you **EXHALE** run your attention outward ***through your 6th Chakra spiral***, into the **6th Chakra Spiral of your Client**. Move your attention into the center of your Client's brain, the **Pineal Gland**. Inhale, and **EXHALE** moving your awareness down from the Client's Pineal, through your *Client's Central Body Current* and down into the **Earth's core**. **INHALE** and move your awareness back upward to your **Client's Maharic Shield**, and spread your awareness outward through the *Maharic Shield* with the **EXHALE**. On the *inside of your forehead*, imagine the image of the **Kathara 12-Tree Grid**, **INHALE** energy from the *Client's Maharic Shield* and **EXHALE** the energy out into the *12-Tree image*. Breathe gently for a moment, *observing the 12-Tree Grid*, which will register tones of white, black and ranges of gray in between. If you cannot yet visualize, you will still get a sense of the 12-Tree image and the areas of dark and light within it. The dark areas represent layers of the **MIASMIC BODY**, areas of the *karmic imprint* that need clearing when you run *Technique # 4: The Maharic Infusion*. Remember the areas of the Client's Kathara that need clearing. Inhale, and *open your eyes* on **EXHALE**.

TECHNIQUE # 4: Amplifying Inner Audio-Visual and Kathara Scan

Activate Maharic Seal and Open Healing Channels. (Techniques # 2 and # 3 first)

1. **RIGHT Palm** on **5th Kathara Center** in **Navel.** Breathe **Maharic Current** up from personal **Maharic Shield** and into **Doradic Sphere** at **4th Chakra.** Push **Maharic-Doradic-1 Current** down **Doradic Cords** and into **Doradic Balls** at **Palm Chakras.** Push Current from Doradic Ball in **RIGHT Palm** into **Kathara Center # 5** in **Navel.**

2. **LEFT Palm** on **5th Throat Chakra,** cover back with **Right Palm** (Reverse Standard Hand Position). Inhale **Doradic-1 Current** up from **5th Kathara Center** in **Navel** to center of **5th Throat Chakra** and expand Current into **5th Chakra Spiral,** with **Left Palm Chakra** drawing current out of **5th Chakra Spiral.**

3. **RIGHT Palm** on **Kathara Center # 6** below **Right Shoulder,** cover back with **Left Palm.** Push **Maharic-Doradic-1 Current** from Doradic Ball in **RIGHT Palm** into **Kathara Center # 6. LEFT Palm** over **6th 3rd Eye Chakra,** cover back with **Right Palm** (Reverse Standard Hand Position). Inhale **Doradic-1 Current** up from **Kathara Center # 6** to center of **6th- 3rd Eye Chakra** and expand Current into **6th Chakra Spiral,** with **LEFT Palm Chakra** drawing Current out of **6th Chakra Spiral.**

4. **KATHARA CLIENT SCAN: A.** Place **Point of Attention** inside **forehead** and run attention out of **6th Chakra Spiral** and into the **6th Chakra Spiral of Client,** then into **Client's Pineal Gland. B.** Run attention down **Client's Central Vertical Current** and into **Earth's Core,** then back up to **Client's Maharic Shield.** Expand attention out into **Client's Maharic Shield. C.** Visualize image of **Kathara 12-Tree Grid,** inhale energy from **Client's Maharic Shield** and exhale it into image of **Kathara Grid. D.** Observe for **Dark Miasmic Buildups** in image of **Kathara Grid;** run extra healing Current into these areas of the **Client's Kathara Grid** during the **Maharic Infusion** (Technique # 5).

Personal Kathara Scan: On **Step 4** run attention down **your Central Vertical Current** then back up to **your Maharic Shield** and out into **your Maharic Shield.** Breathe energy from your Maharic Shield into image of **Kathara 12-Tree Grid** and observe for **Dark Miasmic Buildup** areas to clear during the ***Maharic Seal and Liquid Light Cleanse*** technique. Send extra **Maharic Current** to these areas during ***Liquid Light Cleanse.***

TECHNIQUE # 4: Amplifying Inner Audio-Visual and Kathara Scan

1. A. RIGHT Palm on 5th Kathara Center in Navel.
B. Breathe Maharic Current up from personal Maharic Shield to Doradic Sphere and down Doradic Cords to Palm Chakras.
C. Push Current from RIGHT Doradic Ball into 5th Kathara Center.

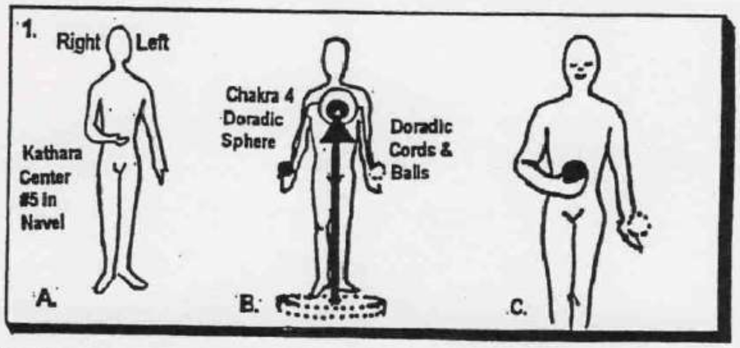
1.
Right
Left
Kathara Center #5 in Navel
A.
Chakra 4 Doradic Sphere
Doradic Cords & Balls
B.
C.

2. A. LEFT Palm on 5th Throat Chakra, cover back with Right Palm.
B. INHALE Doradic-1 Current up from Kathara Center # 5-Navel to 5th Chakra and expand Current into 5th Chakra Spiral.

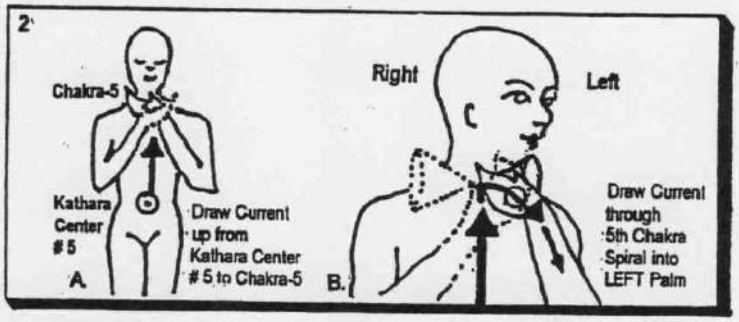
2
Chakra-5
Kathara Center #5
A
Draw Current up from Kathara Center # 5 to Chakra-5
Right
Left
B.
Draw Current through 5th Chakra Spiral into LEFT Palm

3. A. RIGHT Palm on Kathara Center #6 in Right Shoulder, cover back with Left Palm. B. LEFT Palm on 6th 3rd-Eye Chakra, INHALE Current from Kathara Center # 6 into 6th Chakra Spiral and into LEFT Palm. C. Kathara Scan: a. Run your Attention from your 6th Chakra to Client's 6th Chakra and Pineal Gland. b. Run Attention down Client's Central Vertical Current to Earth's Core & into Client's Maharic Shield. c. EXHALE energy from Client's Maharic Shield into Image of Kathara 12-Tree Grid. d. Scan grid for Dark Areas of Miasmic Buildup for clearing with Maharic Infusion.

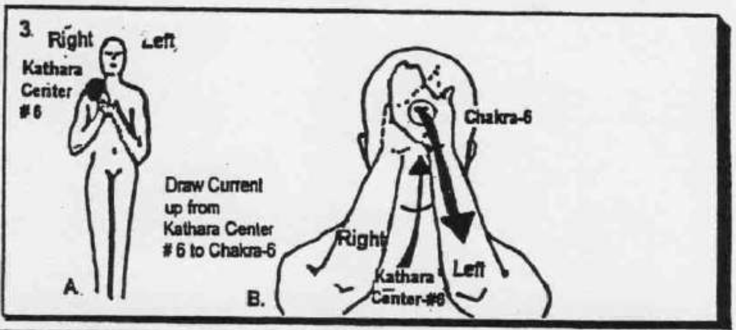
3
Right
Left
Kathara Center #6
A.
Draw Current up from Kathara Center # 6 to Chakra-6
Chakra-6
Right
Kathara Center-#6
Left
B.

KATHARA SCAN

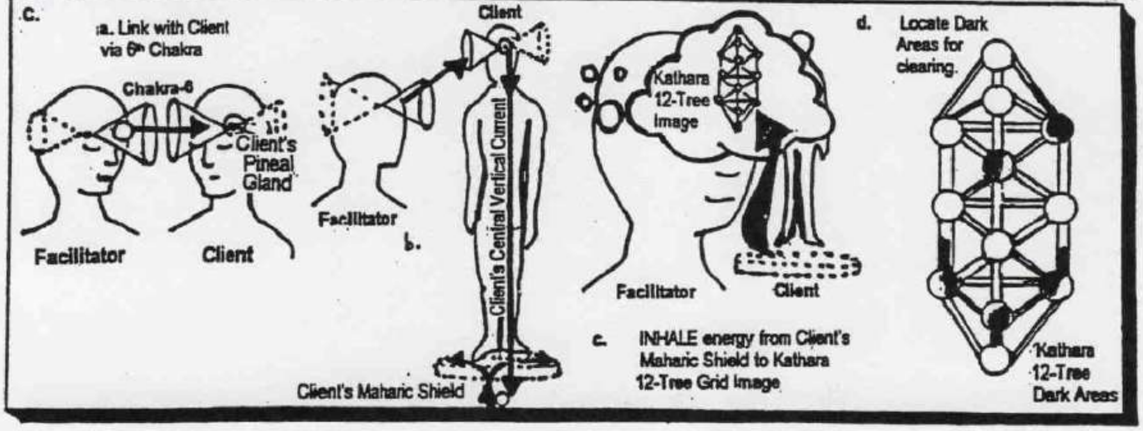
C.
a. Link with Client via 6th Chakra
Chakra-6
Client's Pineal Gland
Facilitator
Client
Client
Client's Central Vertical Current
Facilitator
b.
Client's Maharic Shield
Kathara 12-Tree Image
Facilitator
Client
c.
INHALE energy from Client's Maharic Shield to Kathara 12-Tree Grid Image
d.
Locate Dark Areas for clearing.
Kathara 12-Tree Dark Areas

TECHNIQUE # 5

THE MAHARIC INFUSION – CLIENT FACILITATION

Prior to using this technique you should have completed the following: 1.Positioned *MENTOR* over your Right shoulder to assist (*Tech. # 1:Awakening the Mentor and the 6th and 7th Senses*). 2. Activated your *Maharic Shield, Maharic Seal* and run the *Liquid Light.* (*Tech. # 2: The Maharic Seal and the Liquid Light Cleanse.*) 3.Activated your *Doradic Current, Inner audio/\Visual* and *Client Kathara Scan* (*Tech. # 3. and #4*)
4. Opened the *Kathara Lines* for you and your Client (*Sequence-1 Kathara Points*)

Place Client in a reclining position with shoes off:

1. Position yourself at the **feet** of the reclining Client. Place **RIGHT palm** *12" below Client's feet* at the **Maharic Shield**, *left palm* over the back of Right hand.

2. Visualize the **Hierophant symbol** on the *inside of your forehead.* Inhale, then **EXHALE,** moving the *Hierophant* down to **Chakra 4** and into the **Doradic Sphere** with the exhale breath. **INHALE** energy upward from *your* **Maharic Shield** and into the *Doradic Sphere*. **EXHALE** the **Hierophant** with *Doradic-1 current* down the **Right Doradic Cord** and into the *Doradic Ball* in the **Right palm.** .

3. Inhale, then **EXHALE** the *Hierophant and Doradic-1 current* into the *center* of the **Client's Maharic Shield**. Visualize the *Hierophant spinning* in the center of the Client's Maharic Shield until you sense a **mild shift** in the energy below the Client's feet. The shift indicates that the Client's *Maharic Shield* has activated and the *Maharic Pillar* has "popped" vertically upward from the Shield – the Client's bio-energetic field now carries the **Maharic Seal.**

4. **INHALE** energy from your *Maharic Shield* into your *Doradic Sphere*, then **EXHALE** the energy down **both Doradic Cords,** into the Doradic Balls in the palms. Visualize the **12-Tree Kathara Grid** *within the Client's body*. You will now run the **Maharic Infusion** through the *small chakra points in the feet.*

TECHNIQUE #5 : Continued
THE *MAHARIC INFUSION*

5. Place your **Right palm** on the *bottom* of **Client's Left foot** and your **Left palm** on *top of the foot,* centered over the *Right palm.* Imagine that you are lining up the *small vortex* beneath the ball of the Client's foot (which extends *through the foot* to come out on top), with the *small chakras* in your *palms.* Inhale, then **EXHALE** energy from your **Doradic Sphere** through the **Right Doradic Cord, Right palm** and *into the **sole** of the Client's foot.* Visualize the **Green-Violet-Silver energy** traveling up the **Kathara Line on the** *Left side of the body,* through the *4th, 7th and 10th Kathara Centers.* Continue **EXHALING** energy into the **Left foot** until you begin to feel a bit of **fuzzy energy** coming out of the foot toward your palm. This indicates that the *Kathara Line* has *reached its energy-holding capacity.* **Repeat** the process on the **Right foot**, running energy through the *Kathara Line on the Right side of the body and the 3rd, 6th and 9th Kathara Centers.*

6. You will now clear and align the **Central Vertical Kathara Line** to complete the **Maharic Infusion. Move to Client's head.** Place **Right palm** over **Client's 10th Chakra** 6" above the head, the location of the *12th Kathara Center.* Place *Left palm* over back of Right hand. **INHALE** energy from your **Maharic Shield** into your **Doradic Sphere** then **EXHALE** energy down the **Right Doradic Cord,** into the **Right palm** and through the **10th Chakra**. Visualize the energy moving down from the *10th Chakra* into the **7th Chakra,** to the **Pineal Gland (11th Kathara Center),** then *rapidly* down the **Central Body Current** and **Central Kathara Line,** through *Kathara Centers 8* (thyroid), *5* (Navel), *2* (at Base Chakra) and into *Kathara Center 1* at **Earth's core.** Continue **EXHALING** energy into the **10th Chakra** and **Central Kathara Line** until you feel **fuzziness** against your *Right palm,* indicating that the *Kathara Grid* has reached its *energy-holding capacity.*

7. *Before releasing 10th Chakra* direct a bit of additional energy through the **Central Kathara Line** into areas that revealed **dark spots** during the *Kathara Scan.* Visualize the areas *getting light* as the energy runs through them. **You may add other healing modalities at this point, if desired.** *Close* by using the **Sequence-2 Kathara Points** to close the Kathara Lines and ***revitalize the body.*** Allow Client a few moments to recline before ending session.

THE MAHARIC INFUSION – CLIENT FACILITATION

Preparation: **A.** *Technique #1*- Place MENTOR over your Right Shoulder **B.** *Technique #2*-Activate your Maharic Shield and run Liquid Light Cleanse. **C.** *Technique #3*- Activate yourDoradic-1 Current **D.** *Technique #4* – Amplify Inner Audio-Visual and run Client Kathara Scan **E.** *Sequence-1 Kathara Points* massage on Self and Client to open Kathara Lines.

Place Client in reclining position with shoes off

1. Stand at Client's feet, place **RIGHT Palm** 12" below Client's feet at **Client's Maharic Shield**, left palm over back of Right hand.

2. See **Hierophant Symbol Code** inside your forehead, **EXHALE** Symbol down to **Chakra-4 Doradic Sphere. INHALE** energy up from your **Maharic Shield** into Doradic Sphere. **EXHALE** *Hierophant Symbol* with *Doradic-1 Current* down **RIGHT Doradic Cord into RIGHT Doradic Ball in RIGHT Palm.**

3. **EXHALE** Doradic-1 Current and Hierophant Symbol into Center of **Client's Maharic Shield** via **Right Palm Doradic Ball and spin Hierophant** until sensed **Energy Shift of Client's Maharic Seal.**

4. **INHALE** Current from your Maharic Shield to Doradic Sphere and **EXHALE** Current down both **Doradic Cords** into **Doradic Balls.** See **Kathara12-Tree Grid** within **Client's body** and begin running the ***Maharic Infusion*** through the **Secondary Chakras in Client's feet.**

5. **RIGHT PALM** on **bottom of Client's LEFT FOOT, LEFT PALM on top of Client's LEFT FOOT,** both **centered** over Client's LEFT FOOT **Secondary Chakra. INHALE** Maharic energy into Doradic Sphere and **EXHALE** Doradic-1 Current into Client's **LEFT KATHARA LINE,** and through **Kathara Centers # 4, #7** and **# 10. EXHALE** Doradic-1 Current into Client's **LEFT FOOT** until **"Fuzzy"** resistance emerges from Client's foot Chakra. **REPEAT PROCESS** on **Client's RIGHT FOOT,** running Doradic-1 Current through the **RIGHT KATHARA LINE** and **Kathara Centers # 3, #6** and **#9.**

6. **Stand at Client's Head. RIGHT PALM** over **Client's 10th Chakra** (6" above head- the 12th Kathara Center), LEFT Palm over back of Right hand. Fill Doradic Sphere with Maharic Energy then **EXHALE** Doradic-1 Current down **RIGHT DORADIC CORD** and Palm and into **Client's 10th Chakra.** Run Current from **Chakra-10,** to **Crown Chakra-7, Pineal Gland,** rapidly down **Central Vertical Current** and **Kathara Centers #8** (Thyroid), **#5** (Navel) **#2** (Chakra-1) and **#1** (Earth's Core Chakra 13). Continue exhaling Current until **"Fuzzy"** resistance emerges from Client's 10th Chakra.

7. Direct **additional energy** through Client's **CENTRAL KATHARA LINE** to clear **DARK AREAS** of **Kathara Grid** located in **Kathara Scan.** Add any additional healing technologies following Client's Maharic Infusion. Close with **SEQUENCE-1 KATHARA POINTS** massage. Allow Client to remain relaxed in reclining position for several minutes following Maharic Infusion.

TECHNIQUE # 5: THE MAHARIC INFUSION – CLIENT FACILITATION

Maharic Infusion 6-STEP Sequence

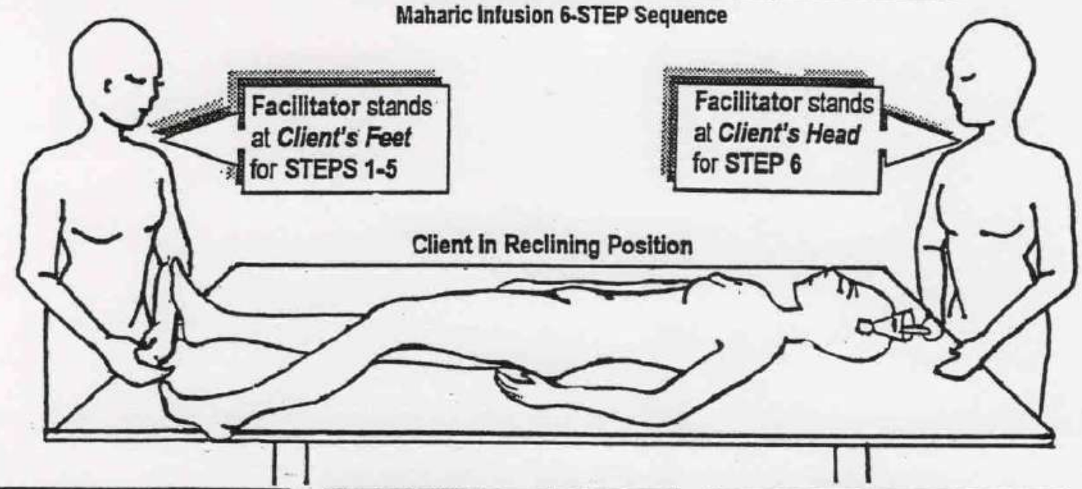
Client in Reclining Position

Facilitator stands at *Client's Feet* for STEPS 1-5

Facilitator stands at *Client's Head* for STEP 6

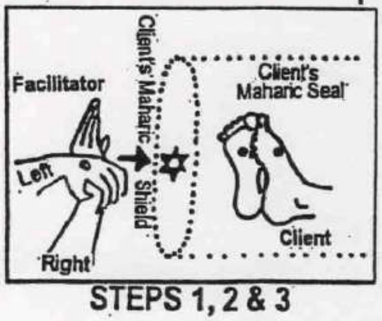
Facilitator
Client's Maharic Shield
Client's Maharic Seal
Left
Right
Client

STEPS 1, 2 & 3

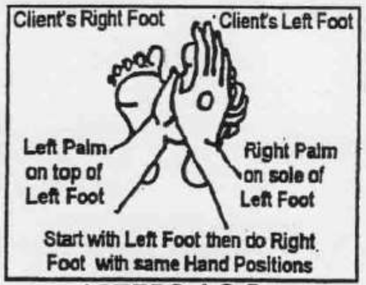
Client's Right Foot
Client's Left Foot
Left Palm on top of Left Foot
Right Palm on sole of Left Foot

Start with Left Foot then do Right Foot with same Hand Positions

STEPS 4 & 5

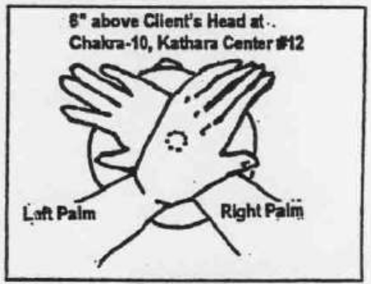
Left Palm
Right Palm

6" above Client's Head at Chakra-10, Kathara Center #12

STEP 6

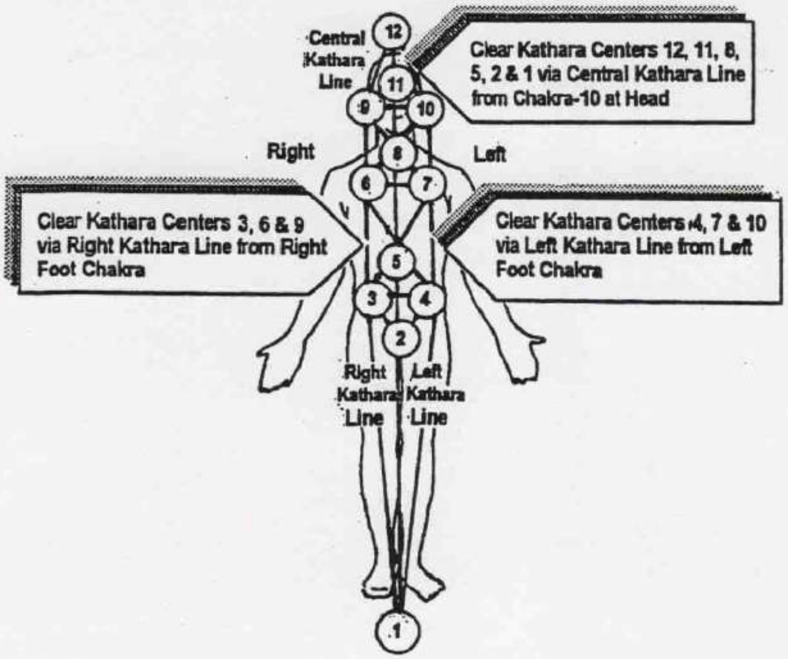
Central Kathara Line
12
11
9
10
8
6
7
5
3
4
2
1
Right
Left
Right Kathara Line
Left Kathara Line

Clear Kathara Centers 12, 11, 8, 5, 2 & 1 via Central Kathara Line from Chakra-10 at Head

Clear Kathara Centers 3, 6 & 9 via Right Kathara Line from Right Foot Chakra

Clear Kathara Centers 4, 7 & 10 via Left Kathara Line from Left Foot Chakra

MOTIVATION TO HEAL

Healing Facilitation is an art and a science, but most importantly, healing facilitation is an **interrelationship of SOULS.** Egotistical objectives have no place in quality Healing Facilitation. The decision to participate in Healing Facilitation should not be motivated by a desire to prove you have "special abilities" to boast about, nor with the primary motive of gaining **financial profit** from Healing Facilitation endeavors. The energies exchanged in healing facilitation are **FREE TO ALL**, a **DIVINE GIFT** from the universe, and they belong to all beings equally. A facilitator cannot rightly request payment from a client for providing what the client already possesses- **Natural Healing Energies.**

In Divine Order, it is appropriate for the facilitator to accept **payment** for the **personal time and energies** invested in **learning and developing facilitation skills,** and for the time spent in active facilitation. It is important that the facilitator understands this subtle difference of **"fair exchange of energy"**. If the facilitator mistakenly assumes they are requesting payment **for healing energies** they supply, they are inadvertently misrepresenting themselves to the client. The client possesses personal healing energies and does not need to purchase such energies from others. The **time and expertise** offered by the facilitator to assist the client in stimulating their own healing energies, is a service worthy of reciprocation. If a client chooses such assistance, the **facilitator** has **legitimate right** to honor their own worth as a being by **requesting reasonable payment or barter** in **fair energetic exchange** for the service they offer. However, if a client is in need of healing facilitation, and has no means of fair exchange at that time, the facilitator should honor the worth of the client by providing assistance of equal value **free of charge.**

In spiritual principal, "making money" is not a legitimate motivation for participating in Healing Facilitation. Seeking to find one's personal **"Divine Right Livelihood"** in **sincere service** to others, which might possibly involve providing client service in Healing Facilitation, IS a spiritually legitimate motivation. Healing Facilitators will discover that when they put the objective of **GENUINE SERVICE TO THE DIVINE** as the **first Priority,** rather than the pursuit of personal monies or status, the **Universe Will Provide.** Facilitators who enter the healing field with hidden personal motivations incongruent with **Universal Laws of Energy,** which imply **equality, respect, reverence** and **balanced exchange of energy** between all beings, will find difficulties emerging in their lives as the personal **SOUL** sets up lessons through which **Universal Principles** become more apparent. There is only **ONE TRUE MOTIVATION** for participation in Healing Facilitation- **the motive is LOVE.**

POSTURES OF LOVE

The Energy-Reality of LOVE

In terms of **Universal Physics**, Love is an **energy reality**, a **STATE OF VIBRATIONAL HARMONIZATION**, or **CO-RESONANCE OF FREQUENCY** that allows an **energetic bridge** to build between individuals. It is through this energetic bridge of frequency that one can assist in running healing energies that will facilitate the healing process of others. Without the sentiment and **FREQUENCY** of **GENUINE LOVE**, one cannot energetically facilitate the healing of another. **LOVE is the essential ingredient in healing,** for it allows the opening of facilitator and client's bio-energetic fields to the **Universal Frequencies** of the interdimensional spectrum. Cultivating the ability to embrace and hold the frequencies of Universal or **OMNI-LOVE** is the responsibility of any true healing facilitator. The emotionally experienced reality of the ***frequency of Love*** takes many forms. In healing facilitation it is important to **assess the most appropriate form of love** to engage with each individual client

Soft Love

Many people respond well to **"SOFT LOVE"**, the kind, gentle, nurturing, soothing love often displayed by mother's comforting their infants. Soft Love works well with client's whom possess some degree of spiritual maturity and who have cultivated the ability to hold some degree of a love frequency within themselves. Soft Love is the appropriate form of love for client's possessing the maturity to accept personal responsibility for themselves, their actions and their emotional reaction patterns, and who do not attempt to manipulate, drain energies from or place blame on others. Soft Love is characterized by **engaged detachment** of the facilitator and is demonstrated through soft-spoken, calm and sincere words and actions that are intended to place the client at ease.

Soft Love is always built up sincerity, kindness, respect and a genuine desire to see the client thrive and prosper. Developing the ability to carry the Soft Love frequency takes time, attention and practice, and requires that the facilitator works to heal personal **internal conflicts**, unresolved issues from **childhood** and **karmic** reaction patterns that disrupt the Soft Love vibration. Soft Love is a **vulnerable love** in that it necessitates one to show the self as it is, honestly expressing personal feelings tempered with kindness. Soft Love is **not approval seeking**, it is rather genuine, honest and **self-generated**. Soft Love requires **love** and **respect** of the self and others mutually. It is built upon a level of spiritual maturity in which it is understood that personal value is implied by the fact of existence and is not determined by the approval or **validation** of others outside of the self and the personal relationship to the Divine. The ability to carry the Soft Love frequency emerges through **spiritual integration, personal healing** of internal conflicts and power struggle issues and **intentional application of kindness and mindfulness.**

Tough Love

Soft Love does not facilitate healing on all occasions. Certain clients who come to a facilitator for healing assistance bring with them a great need for **emotional healing** and development of spiritual maturity. They may demonstrate demanding, arrogant or pushy attitudes and refuse to accept responsibility for their personal actions and resulting consequences. Most often such individuals carry deep wounds from childhood or from reincarnational bleed-through. Because they are **wounded within**, they have not yet developed the self-control or maturity necessary to treat themselves or others with kindness, respect or love.

Individuals who display unreasonable behaviors or attitudes pose quite a challenge to the love-based healing facilitator. If they are approached with Soft Love, the facilitator will often find themselves as a scape goat for the individual's personal problems, may have their energies and time excessively **drained** by such clients and might possibly be subjected to outright verbal or physical **abuse**. Clients exhibiting such personality traits are struggling within themselves to **gain control** over the various **conflicting portions** of their personal energies, and they tend to objectify this internal conflict resolution in the form of **POWER STRUGGLE** with others.

Personalities trapped within cycles of subconscious self or other-abusive attitude patterns need **love** more than anyone, but most often their behaviors push others away. Intrinsically the behaviors serve to keep the individual sheltered from **self-exposure;** they do not allow others close enough to inspire feelings of vulnerability or lack of self-esteem. Often such personality traits effectively keep the individual from seeing themselves and the inner pain and conflict from which they attempt to hide. To facilitate healing in such individuals without succumbing to their manipulation, aggression or abuse requires that the facilitator adopt a posture of **TOUGH LOVE**. In Tough Love the facilitator clearly establishes **personal boundaries** in their own mind as to what treatment they will and will not accept. If the client crosses those boundaries the facilitator **assertively addresses the issue** with the client, requests that the offensive behavior cease and provides **CLEAR CONSEQUENCES** as to what will occur if the offense continues. The Tough Love approach is rendered from a base of **Soft Love** and **Omni-Love**, but it is strengthened with **Self Love** on behalf of the facilitator. The facilitator recognizes that they are only assisting the client to continue with the self-destructive patterns by condoning or allowing offensive behavior, and chooses to **love the client enough** to **confront the pattern** so that it may begin to release. In **lovingly, calmly**, but **firmly confronting** poor behavior and **setting clear boundaries** and **consequences**, the facilitator assists such clients to temporarily find a new pattern of action because the old one does not work for manipulation in this instance. The Tough Love approach frequently requires facilitators to demonstrate that they hold their own power, even in the face of client disapproval. Tough Love is an ability of **personal empowerment** worth cultivating for personal and client healing.

Self Love

In healing facilitation **Love** is the essential ingredient. All love begins with the **Divinity Within the SELF.** You will be able to transmit the frequency of love to others in direct proportion to your ability to hold the frequency of love within yourself. If you do not recognize your intrinsic value as a **Divine Being** and seek validation of your existence through external sources such as achievements, status or other-approval, you will likewise project these sentiments into your love relationships with others. If a client does not meet the external standards of validation that you have placed on yourself to determine your personal worthiness of love, you will have difficulty not falling into **judgement** of the client.

Judgement and **love** cannot transmit through the human body at the same time. Judgement creates an **energy reality** of separation or **Non-Resonant Frequencies** of energy, whereas Love creates the **Co-resonance of Frequencies** needed for open flow of Universal Energies for healing facilitation. If one can realize that **LOVE IS THE ONLY CONSTANT** and that all conditions of judgement change, it is easier to cultivate the innate ability to fully hold the frequency of Love. Though **ASSESSMENT** of conditions, actions or attitudes is useful and necessary, such assessment can be rendered through **"separating the person from the action".** You can judge the effectiveness or value of the **action or idea** without assigning a **value-judgement** to the **person** to which it is attached.

If the healing facilitator can feel love, reverence, respect and honor for the **Divinity Within Themselves,** they will also transmit these qualities of **excellence in Love** to the client. The quality of Love brought into the healing facilitation experience will directly effect the success of healing assistance given. **Love the Self** and know that all beings possess an **unalienable worth and value** as a living part of the **Divine.** Begin to cultivate this awareness within your personal life and you will greatly increase your effectiveness as a healing facilitator. Love, honor and respect yourself, your feelings, your dreams and your desires, so you may better love those you wish to serve.

Love others as you would love the Self,
and love the Self as you would like to love others.

The Illusion of Martyrdom

People who love themselves do not buy into the idea that they must play **"martyr"** to others in order to be of service. **Martyrdom** implies a belief that you are **lacking in worth,** or must **earn** your worth in being, by relinquishing that which feels of most value to you. It also **undermines expression of the God-Presence** within others, by assisting them to believe that you are so vital to their wellbeing that they cannot function without you. In truth, martyrdom is an **expression of EGO**, most often used by people who feel they are **personally insignificant** in the eyes of the Divine, in order to **give themselves significance** and a **sense of personal power and purpose** for being. People who actively *know* the intrinsic **Spiritual Divinity** within themselves treat that Divinity with honor and respect and do not allow or need others to **depreciate the Self** and its desires in order to prove the existence of a *belief* in personal Divinity.

Martyrdom **does not facilitate healing** within the self or client, as it reinforces the **erroneous belief** in intrinsic **personal unworthiness,** which exists in **direct contradiction** to the **innate God-Presence** that exists in every living thing. Within the energy dynamics of **Universal Law,** natural **balance** is achieved in the flow or **GIVING**, and back-flow or **RECEIVING** of energy. In human spiritual relationships this **balance of energy** is achieved by **mutually valuing** each person in a relationship, through which the natural giving and receiving of energy can be exchanged unimpeded. **Martyrdom impedes the natural flow of energy**; the martyr customarily gives more than receives, which progressively diminishes the martyrs ability to keep open a perpetual flow of energies from the Divine because the martyr does not allow the Self to receive.

Martyrdom is not a demonstration of Love,
it is an exhibit of Love Deprivation.
The remedy for Martyrdom is
OMNI-LOVE.

OMNI-LOVE; Fully Engaging the Love Vibration

The greatest love that can be embodied in human form is **OMNI-LOVE.** OMNI-LOVE is a pure state of vibrational Frequency Co-Resonance between the human identity and the **12 Dimensions** of reality within which the human anatomy is couched. Omni-Love is expressed in human form as **Christed Love** or **Christos Consciousness, the 12th-Dimensional Love** of the **Maharic Level of Identity.**

Christos Love is a state of full Frequency Resonance with **everything existing** in the many universes contained within **12 Dimensions** of the 15-Dimensional System. It is a **transcendent love** that is attached to no thing but is **AT ONE** with **ALL THINGS** in a state of perpetual **Engaged Detachment.** Through Christos-Omni Love, the **SELF** is known as an **extension of God**, or the **Divine Source;** a **temporary mask of form** worn by the **ONE CONSCIOUSNESS** of the Cosmos. All things and beings are known as **simultaneous expressions** of the **ONE-SELF** that is God-Source. From this state of transcendence in love, all activity is understood to **exist within** the reality of LOVE, and all conflict and strife are viewed as the ONE-SELF **progressively expanding the ability of its expressions** to **carry the frequencies of energy** that constitute **ONE-LOVE,** a state of **total vibrational Frequency Co-resonance** with the **Cosmos.**

Activation of the full **12-Strand Silicate Matrix DNA Template** and full integration of the **Mahara Hova Body Avatar Identity** allows for the experiential embodiment and transmission of Christos-Omni-Love. **Embodiment of Omni-love is the goal toward which human evolution moves forth.** Omni-Love is Soft, Tough, and Enduring, honoring **Self** , **Other** and the **Divine** simultaneously. It is the **strongest healing force** of frequency in the 4 Harmonics of Matter Density. Christos-Omni-Love is the **fulfillment** of the **Christos Principle.** Strive to bring Omni-love into your life, and personal and client healing facilitation, and you will assist your higher identity levels to create transformation! Omni-Love is the **natural structure** of reality. **AWARENESS** of Omni-Love is cultivated through **intention** and appropriate use of **personal free will choice** in congruence with the **natural Laws** of the **Unified Field Physics of Consciousness and Creation.**

Through living in Omni-Love all moments are known as Sacred, all beings are known as Blessed and life becomes an ecstatic study in Reverence for Existence.

Postures of LOVE

1. SOFT LOVE
KIND, NURTURING, **GENTLE**

2. TOUGH LOVE
KIND, NURTURING, **ASSERTIVE**

3. SELF LOVE
HONORING THE **DIVINITY** WITHIN
SELF

4. OMNI-LOVE
HONORING THE **DIVINITY** WITHIN **ALL**
Namaste'

__Omni-Love__ is the fulfillment of the __Christos Principle,__ the goal toward which human evolution moves forth through progressive activation of the __12-Strand Silicate Matrix DNA Template__ and embodiment of the __Avatar__ Identity.

OMNI-LOVE EXERCISE TO AMPLIFY HEALING FACILITATION

EXERCISE

The MAHARA SPACE-COMFORT TOUCH

1. **EXHALE Hierophant Symbol** from behind forehead, down *Central Body Current*, into **Earth's Core** (Chakra 13), then **INHALE** *Hierophant* back up to your **Maharic Shield 12" below feet.** *Spin Hierophant* to *activate* **Maharic Shield** and **Maharic Seal Pillar. INHALE** energy up from your *Maharic Shield* into your **Doradic Sphere** at *Chakra 4.*

2. Breathe slowly and deeply while visualizing a **small faceted Crystal Sphere** in the center of your *Doradic Sphere* at Chakra 4. This is the **Holographic Crystal Sphere**; it represents the station of your inner **AVATAR** identity – this is the **MAHARA SPACE.** Focus your consciousness within the *Holographic Sphere* and feel its reality around you, as if your body is encased within the Holographic Sphere.

3. Observe the interior of the Sphere as a **pale Silver-White mist.** Within the mist now envision a **golden treasure chest**- a treasure you had left for yourself long ago to one-day reclaim. This represents your **Inner Arc of the Covenant** – *a doorway to reclaiming your ancient immortal heritage.* **Mentally command** the lid of your inner *Arc of the Covenant* to lift. As the lid rises **pale rainbow colored light** fills the interior of your Holographic Sphere. Notice now that within the golden box there is a gleaming **Silver Sword**, the Silver Rod of Power that will enable you to more quickly evolve to your eternal state of being.

4. Mentally make the **Silver Sword** *rise from the golden box*, floating above the box in a vertical position. Now call to mind the image of your **body** *as it sits within the physical room.* Imagine *the Silver Sword positioned above the body's head*, directly over **Chakra 7. EXHALE** and **steadily lower the Silver Sword** through the *7th Chakra and down the Central Body Current, through the Maharic Shield and down into* **Earth's core.** Observe the Silver Sword turn into a *Pillar of Pale Silver Light with pale rainbow colors flashing through it.* The **Rainbow Pillar** now *surrounds and runs through your body, raising its frequency while lowering its vibration,* lifting your body cells into a **less-dense state of matter.** Feel your body growing lighter and lighter as the Rainbow Pillar slows the spin-rate of your matter particles, making you more and more electric- more light.

5. As you feel the sensations of your density lessening, **imagine that your body is beginning to glow with Pale Silver light** – the *Pre-matter Liquid Light* of dimensions 10, 11 and 12. Feel the *features of your face and the contours of your body begin to melt* into a **Silver, 2 –dimensional fluid body form.** The form of the body still remains but it has been returned to its **first-density state**- the **Silver Liquid Light holographic form of the avatar identity.** Know yourself now as this avatar identity and feel the essence of **pure love** and eternity flow through you. **Know yourself as this essence of Divine Love and Eternal Light.**

6. Now **INHALE** and draw the **Rainbow Pillar into the Holographic Sphere** at *Chakra 4*, visualize the base pulling up from Earth's core and the portion above your head simultaneously *contracting into the Holographic Sphere.* **EXHALE** the energy of the *Rainbow Pillar* fully into the *Holographic Sphere* and observe the **Holographic Sphere become a radiant Silver ball of Liquid Light stationed with your Doradic Sphere** in the Heart Chakra.

EXERCISE: Continued
THE *MAHARA SPACE-COMFORT TOUCH*

7. **You may now transmit the frequencies of the 4th-Harmonic, dimensions 10, 11 and 12, to serve as a healing force for others.** You may transmit this energy into your environment directly through the **4th Heart Chakra**, where it will enter the Unified Field at the 4th-Dimensional astral plane. You may use the breath to move the Holographic Sphere down into the **3rd Chakra** *for 3rd-Dimensional transmission* into the *Earth's atmosphere* and the *Mental Body* levels of people's Bio-energetic field. You may transmit through **Chakra 2** into the Telluric or *Elemental Kingdom* of the planet and into people's 2nd-Dimensonal *Emotional bodies*. Or you may transmit the frequencies of the Silver Sword through the **1st Base Chakra**, for *sacred bonding in intimate relationship*, or to *raise the frequency and lower the vibration* of the *molecular and sub-atomic fields* around you. To transmit the frequencies of the Silver Sword through the chakras, simply use the **EXHALE** breath to move the Silver Light "charged" **Holographic Sphere down** to lower chakras, inhale and use the next **EXHALE** to *push the frequency outward* from the selected chakra into the environment. Use the **INHALE** breath to move the *Holographic Sphere* up through the chakra system. *Always move the Holographic Sphere up or down along the* ***Central Body Current****, to which the core of each chakra is attached.*

8. **You may also transmit the frequencies of the Silver Sword through APPLIED TOUCH –** ***THE COMFORT TOUCH*** -to amplify specific healing facilitation functions and to establish a more harmonious rapport with other people and the nature kingdoms. Simply visualize the Silver Liquid-Light charged **Holographic Sphere** at the center of your **Doradic Sphere** in the **4th Chakra**. Activate your **Doradic-1 current** via the **Maharic Shield and Seal**, then visualize the *Silver Liquid Light from the Holographic Sphere combining with the Green-Violet-Silver Doradic-1 current* as you **EXHALE** the current into the **Doradic Cords** in the arms and into the **Doradic Balls** in the palms. *This application is especially useful to help Clients relax during sessions, or to comfort others when they are in a state of stress, unbalance or illness.*

   Once your current is moving into the Doradic Balls, **EXHALE** the current through the **Right palm.** *Place the* ***Right palm*** *gentle and briefly over the* ***Client's rear Chakra 2****, or touch them gently on the* ***left shoulder****.* The amplified *Doradic-1 current* will move into the *4th-Dimensional astral-body field*, and move its way into the *mental, emotional and physical bodies* from there. **The COMFORT TOUCH of the MAHARA SPACE can be applied to infants, children, older persons, animals and plants to comfort them and support their general well-being.**

PROFESSIONAL USE OF THE KATHARA BIO-SPIRITUAL HEALING SYSTEM PROGRAM – THE KATHARA ALLIANCE

The Kathara System introduced in **The Kathara Bio-Spiritual Healing System ™ Level 1 Workshop & Course Manual with Introduction to The Kathara Bio-Healing System™ Level 2** is destined to become a ***12-level program*** that is focused on "whole-being healing through progressive biological and spiritual awakening". Kathara is as much a course on (non-denominational, dogma-free) **Spiritual Integration Mechanics** as it is a Holistic Healing course.

An individual completing training in the **Kathara Level-1 Bio-spiritual Healing System** delivered by a Certified Kathara Alliance Instructor, will receive a **Certificate of Completion**, which recognizes the person named as attending and receiving training in the Kathara Level One Bio-Spiritual Healing System.

If you desire to practise the ***Kathara Bio-Spiritual Healing System***™ in a regular, professional capacity through Client Service, you may use the techniques provided within this Kathara Healing Manual to begin. Prior to professional use, it is suggested that one becomes a Certified member of the **Kathara Alliance** through the ***Azurite Press MCEO (World-wide Consulate***) and its appointed representatives.

The **Kathara Alliance** is a group of dedicated spiritual healing Practitioners and Instructors who have demonstrated a thorough understanding of the Theory and Practice of the **Kathara Bio-Spiritual Healing System**™ and the Maharic Recoding Process, as deemed necessary by the **Azurite Press** for professionalism in the practice of Spiritual Healing Facilitation and Instruction.

All members of the ***Kathara Alliance*** are **Certified Kathara Healing Practitioners and Instructors,** and are permitted to practice Kathara Healing facilitation as **SPIRITUAL SERVICE.** For the protection of both Client and Practitioner, as well as the ethics inherent to the Teachings themselves, Kathara Healing is to be promoted, explained and practiced as **a non-diagnostic spiritual healing enhancement service only** and not as a replacement for traditional medical diagnosis or treatment.

Certification in the **Kathara Alliance** as an **Independent Spiritual Healing Practitioner and Instructor** of the **Kathara Bio-Spiritual Healing System**™ requires formal application, including a written open-book examination and video presentation by the applicant. The examination is intended to reveal areas in which applicants may need to further strengthen understanding of and personal relationship to the practice of Kathara Healing Facilitation and Instruction within the parameters of Excellence recognized as intrinsic by the Azurite Press MCEO.

All Instructors and Practitioners must be CERTIFIED to conduct **public presentation** of copyright materials contained in this, the authorized Kathara Bio-Spiritual Healing System Manual. To uphold standards of Professionalism and Excellence in practice, the Azurite Press MCEO **cannot** recognize **Non-Registered Facilitators or Instructors** as active participants in the ***Kathara Bio-Spiritual Healing System Program.*** Obviously, non-certified Facilitators may use this program and its applications for their own **personal and spiritual enrichment. The Azurite Press MCEO authorizes only Certified Practitioners and Instructors** to professionally conduct Client Sessions.

With regards to teaching the program there is a distinction between 'Teaching' and 'Sharing'. **Teaching** is defined as formal presentation of the materials (following protocols described in the **Kathara Bio-Spiritual Healing System™** Manual and communications to Kathara Alliance applicants and members) in the form of a seminar/ workshop and payment is made by attendees. To teach the Kathara Bio-Spiritual Healing System™, one needs to be a member of the **Kathara Alliance**. '**Sharing**' is defined as leading a study group where attendees are not expected to pay (it would not be inappropriate to accept small donations if offered). To share one does not need to be a member of the Kathara Alliance but group size should be limited to less than 12.

Within the **Kathara Bio-Spiritual Healing System™** there are specific pages for Kathara Alliance Instructors to use in their classes. The Instructors pages are identified by having a number at the *top* of the page (as well as the regular number at the bottom). These are the only pages that can be photocopied for teaching purposes. Only certified Kathara Alliance members (or a Kathara Alliance applicant who is delivering and recording a workshop for assessment purposes) are authorized to photocopy these pages.

The easy-reference **Instructor's pages** need to be sequenced in the order denoted by the number at the *top* of the page. These pages have been carefully sequenced to facilitate delivery and learning of the materials. Learners' needs, including varying learning style preferences, were very much considered when preparing the text and visual presentation of these teaching materials. In terms of frequency, the learning process can be described as when previously dormant Scalar Standing Waves in the personal shields are stimulated and activate('switch on'). As frequency is accreted (by progressive exposure to, and interaction with, the materials) the scalar waves in the personal shield self –sustain (we would say the person has 'learned' the material). The specific sequence of the pages are pre-set to create optimum frequency accretion in the personal shields of learners *even if this is not immediately perceived or understood by the learner*. Therefore it is imperative that these pages are used in the delivery of this program and the instructor remains true to the given sequence.

As Independent Spiritual Healing Facilitators and Instructors, all Certified Facilitators and Instructors are expected to accept full personal responsibility and legal liability for the practice and application of the **Kathara Bio-Spiritual Healing System™.** Certified Facilitators-Instructors are **not required** to pay dues, royalties or revenues to the Kathara Alliance/APMCEO for use of this program other than the cost of initial personal **Training-Registration** and obtaining supplies of **Session Manuals, Certificates and Registration Examinations** for the professional Instruction-Certification of others.

THE KATHARA ALLIANCE CERTIFICATION AND COMPETENCY REQUIREMENTS

The Certification Program is designed to enable the applicant to fully demonstrate his/her understanding of the material in a way that maintains the level of integrity and purity of the Program desired of the Kathara Alliance. This also ensures the Student attains a desired level of understanding in taking the workshop. The requirements of Certification are designed to protect the integrity of the material, those who participate with it, and the rights and responsibilities of all concerned.

Certified Kathara Practitioners and Instructors must not only be knowledgeable in the practice and teaching of Kathara, but are also asked to commit to the behavioral guidelines originating from the Covenant of Regent Consulate of the Melchizedek Cloister Emerald Order (MCEO). In this way those seeking out a Kathara practitioner can reasonably expect the highest integrity in both the teachings and in the individuals with whom they are relating.

Certified Kathara Alliance members are in no way affiliated with the MCEO Vehicle, the Azurite Press and should at no time state or imply such representation. Kathara Alliance members are independent Practitioners who offer local, community-based Kathara Healing Level courses and facilitation for individuals.

PLEASE NOTE: The APMCEO & Kathara Alliance **reserve the right** to **revoke** any Kathara Alliance Certification and Rights of Association of Certified Facilitator-Instructors for ***any*** reasons of: *Misuse, misrepresentation and false claims pertaining to this Kathara Program; inappropriate treatment of APMCEO staff, clients or participants in Kathara Healing Programs; dishonest or unethical behavior and practices, and racial, gender, or religious discrimination of clients and participants or Kathara Alliance Members.*

For further information on the Kathara Alliance certification process visit our website at http://www.azuritepress.com/kathara/kathara_application_proceedures.html or contact

In the U.S. and Canada: Sandy Sayles, 8527 Cool Water Ct., Antelope, CA 95843 - email - siepearl@comcast.net **OR** Melissa Fathman, 855 W. Dillon Rd, Apt. H306 - Louisville, CO 80027 - email - mfathman@yahoo.com

Rest of World Office: Hilary Lyons 341A Green Lanes, Palmers Green, London N13 4TY, U.K.
hilary.lyons@virgin.net 07970770404

THE 6-STEP MAHARIC RECODING SESSION

Before Client Arrives: ***Technique # 1***- Activate Your Mentor. Conduct ***Mahara-Space-Comfort Touch Exercise*** and remote-request permission from client's ***Rishi*** identity level to apply **Comfort Touch** when client arrives. Create a calm, pleasant environment where you will not be interrupted by phone or other distractions. Optional: Sing 3 rounds of the ***Ariea-Khum-Nar'A* toning** to amplify personal healing and DNA Template activation during healing facilitation session.

When Client Arrives: Apply Comfort Touch if Rishi gave remote permission. Sit with client and review terms of the ***Client Agreement Form*** and secure client's signature on form. Offer client the option of audio-taping the session for client's at-home review.

6-Step Technique Application Sequence- *Maharic Recoding Session*

1. ***Technique # 2*-The Maharic Seal and the Liquid Light Cleanse** *to seal and align personal bio-fields.*

2. ***Sequence-1 Kathara Points massage***- first your points, then client's points, *to open Kathara Lines.*

3. ***Technique # 3*- Opening the Healing Channels** to activate your *Doradic-Phase-1* healing current and ***Technique # 4*- Amplifying Inner-Audio-Visual and Kathara Scan** to locate miasms for healing in client's Kathara Grid.

4. ***Technique # 5*- The Maharic Infusion** to activate client's Maharic Shield and Seal, recode client's Kathara Grid to the Maharic Shield and clear miasmic buildups.

5. **Optional:** Apply **light touch-massage** to miasmic areas, or employ other healing therapeutics after Maharic Infusion.

6. ***Sequence-2 Kathara Points massage*** on client to revitalize outer body systems and close Kathara Lines. Allow client to rest a few minutes before leaving. Answer any client questions and allow time for client to discuss experience if desired.

When Client Leaves: To ensure clear bio-fields, while Doradic-1 Current is still running, draw energy up from Earth's Maharic Shield and run Maharic Current through your Kathara Grid and body, up to Chakra-14 and back down to Chakra-13, then push a Pillar of Silver Light out of Chakra-4 and EXHALE it to release it around body into auric field. Next massage *your* **Sequence-2 Kathara Points** to revitalize bio-field systems and close Kathara Lines.

CHAPTER 4 HIGHLIGHTS

GENERAL KATHARA HEALING FIELD WORK

- Activation of any **Triadic Healing Current** requires temporary release of corresponding **Signet Star Crystal Seals** in the **Kathara Level-2 Crystal Seals Grid.**
- **Maharic 12th-Dimensional Frequency** is required to **manually release** the **Signet Star Crystal Seals** to activate **Triadic Healing Currents. Kathara Healing** utilizes **Maharic Frequency** to activate interdimensional **Triadic Healing Currents** for more powerful healing support.
- Most healing systems have not used **Maharic Current** because the **Planetary Templar Complex** could not run 12th-Dimensional Maharic Frequency for over 200,000 years, until January 2000.
- Release of the **Signet Seals** also activates dormant Strands in the **DNA Template** and releases the **Cranial-Sacral Seals** on the **Kundalini Energies.** Kathara Healing represents advanced, **12-Dimensional Kundalini Healing.**

OPENING THE HEALING CHANNELS

- Triadic Phase Currents are activated in the body by using **Maharic Frequency** from the **PBIS** to release the **Signet Seals** of the **Internal Templar Complex,** which allows higher dimensional frequency to run through the **physical body system** for transmission.
- Triadic Phase Currents are composed of **3 interwoven frequencies,** the **Maharic Frequency,** plus a chosen **dimensional frequency** and its corresponding frequency from **one Harmonic up.**
- Triadic Phase Currents are **directed** by their corresponding **Color Cap Stone,** the **Primary Wave Length** of the dimensional frequency band from which the current is drawn. **Kathara Level-1** activates the **1st Triadic Phase Current,** the ***Silver (D-12) + Green (D-4) + Violet (D-7)*** **DORADIC PHASE-1 Current.**
- **Doradic Current** is activated through the **Maharic Shield,** then **collected** in the **4th Chakra** ***DORADIC SPHERE,*** projected down the ***DORADIC CORDS*** in the **arms** and into the ***DORADIC BALLS*** in the **Palm Chakras** for **DIRECT INDUCTION** of frequency transmission in healing facilitation.

PALM CHAKRA POLARITY

- The **Palm Chakras** are **Secondary Chakras** located in the **Palms of the Hands.** Like all Chakras, they function as **Black-White Hole Sets** for **transmitting** and **receiving energy** between the personal Manifestation Template and the dimensional Unified Fields of energy.
- The Palm Chakras are **double vortices** of energy, with **one spiral** exiting the **Palm** and its **corresponding spiral** exiting the **back of the hand.** The **2 Spirals** of the Palm Chakras are linked by a **Crystal Seal** at the center point where the 2 vortices meet. In one Palm Chakra, one spiral rotates **clock-wise**, is **Base Electrical** and **transmits energy** out of body, and the other spiral rotates **counter-clock-wise**, is **Base Magnetic** and **draws energy** into body.
- In **Level-1Kathara Healing** the **RIGHT PALM** is the **TRANSMITTING HAND** and the **LEFT PALM** is the **RECEIVING HAND.** For **transmission**, the **Standard Hand Position** is **LEFT Palm placed over the back of the RIGHT HAND** for pure transmission and **control of specific frequency** transmitted from the **RIGHT PALM.**
- The **RIGHT PALM** is **BASE ELECTRICAL**, the **LEFT PALM** is **BASE MAGNETIC**, but both have **subtle** electro-magnetic capacity.

THE KARMIC-MIASMIC IMPRINT

- To enter manifestation in time, identity enters **cycles** of **SIMULTANEOUS INCARNATION** within the **4 Harmonics of Matter Density in 12-Dimensions** of the 15-Dimensional **Time Matrix.** The **PRIMAL ORDER Scalar-wave Template** through which an identity enters time is the **Density-5** (dimensions 13, 14 and 15) **RISHIC SHIELD**

- The **Rishic Shield** holds in the Time Matrix the identity of the **Family of Consciousness** from which **incarnates** emerge. The Scalar Patterns of the Rishic Shield represent the tangible **"DIVINE RIGHT ORDER"** in energy and consciousness, for each incarnate manifesting from it. The individual identity must **realign** its **Manifestation Template** with this **Primal Order** to become free from the Time Matrix in full **Ascension.**

- The **Density-4** Pre-Matter Hydro-plasmic **MAHARIC SHIELD** (Dimensions 10, 11 and 12) **is** the Scalar Template through which identity manifests in the lower-dimensional Densities and which **connects the individual incarnate** to its **Rishic Shield.** The **Maharic Shield** is often called the **PERSONAL CHRISTOS Template,** and **it** holds the consciousness and **Perfect Primal Order** for all incarnations of an identity in time.

- "**Karma**" is the result of **Free Will choices** made by **incarnates** that involve using personal energies in ways incongruent with the natural **Laws of Universal Unified Field Physics** and the **Primal Order** of the **Christos Template**. Karma creates **scalar-wave distortions** within the Scalar Shields of personal 15-Dimensional Anatomy that **manifest physically** as **energetic clusters** called **MIASMS.**

- **Miasms** are **unintended** groups of **anti-particles** lodged within the particle field of the personal **Manifestation Template** that block and impede the **natural flow of energy** and consciousness into embodiment. **Miasms,** and the misshapen scalar-wave distortions that form them, are the **cause of ALL disease** and dis-harmonic conditions.

- **Health** and Harmonious flow of energy are the **natural conditions** of human design. The **Maharic Shield Personal Christos Template** holds the **Organic Imprint for Health** for all simultaneous incarnates of an identity, and can therefore be used to **clear the Karmic-Miasmic Imprint** from the personal Manifestation Template to progressively restore **Healthful Primal Order.**

- Miasms from **all simultaneous incarnations** manifest within the personal Manifestation Template and within the DNA Template of each, where they become literal **genetic distortions.**
- Each of an identity's incarnates in various time cycles represents one **TIME VECTOR**, which is encoded in the Vector Codes of the **DNA Template.** Karmic-Miasmic **bleed-through** occurs **between incarnates** from different concurrent time cycles through the **living link** of scalar frequency that exists between the **DNA Templates** of simultaneous incarnations.
- The **Karmic-Miasmic Imprint** does not have to be suffered and cleared the "**old fashioned way**" by "walking through the Karma in physical reality". The **Miasmic Imprint** can be cleared **as energy, with energy**, to prevent its manifestation in physical terms via realignment of the **DNA Template** with the personal **Maharic Shield.**
- Any given incarnate can clear the Karmic-Miasmic Imprint by realigning the personal Manifestation Template with the shared Christos Template Maharic Shield. This process is called **MULTI-VECTOR HOLOGRAPHIC RECODING**, which creates **BIO-REGENESIS** of the original 12-Strand **Silicate Matrix DNA Template** of human design and restoration of the **Organic Imprint for Health.**

MULTI-VECTOR HOLOGRAPHIC RECODING

ATTITUDES AND ETIQUETTE IN HEALING

- **Holographic Recoding** and clearing the Karmic-Miasmic Imprint from the personal Manifestation Template begins with activation of the **Maharic Shield and Seal** in the bio-energetic field and progressive running of 12th-Dimensional **Maharic Frequency** from the **Personal** and **Planetary Maharic Shields** through the **3 Primary Levels** of the embodied **Kathara Grid**; the Level-1 **12-Tree Grid**, Level-2 **Crystal Seals Grid** and Level-3 **Diodic Points Grid.**
- Personal Holographic Recoding begins with consistent use of ***Technique* # 2- The Maharic Seal and the Liquid Light Cleanse**, and is further accelerated by facilitating **Maharic Recoding** in others through running the **Triadic Phase Currents** for **Co-generative Healing.**
- One can facilitate healing in others and amplify personal healing through the using ***Technique* # 5- The Maharic Infusion.**
- The **Maharic Recoding Session** utilizes the combined **Bio-Regenesis Techniques** presented in **Level-1 Kathara Healing** to facilitate **Client Healing Sessions.**

- Professionalism in Healing Facilitation requires an understanding of the **Postures of Love** and a willingness to **cultivate** the ability to hold and transmit **OMNI-LOVE, honoring the Divinity of ALL things and beings.**
- The Primary Postures of Love are **SOFT LOVE, TOUGH LOVE, SELF LOVE** and **OMNI-LOVE.**
- The emotionally experienced quality called **LOVE** is a **REALITY IN ENERGY PHYSICS**. LOVE is a **STATE OF VIBRATIONAL FREQUENCY CO-RESONANCE** that allows **free flow of energy** between consciousness within the Universal Unified Field.
- The human **12-Strand DNA Template** was designed to allow **12-Dimensions of Vibrational Co-Resonance**, the **Christos Consciousness** of the **Maharic Shield**, to embody in human form. **CHRISTOS-OMNI-LOVE** is Vibrational Co-Resonance with all that exists within **12 Dimensions** of the Time Matrix; its embodiment is the **primary goal** toward which **human evolution** progresses.
- Christos-OMNI-LOVE is the **strongest healing frequency** available within the Densities of Space, Time and Matter, and is **essential** for effective **Healing Facilitation** and personal **Ascension.**

CHAPTER 5

INTRODUCTION TO KATHARA LEVEL-2

The Kee-Ra-ShA Induction Process™

Mahara-Merkabic Master Key Healing-Manifesting the Blueprint For Health

Primal Currents, the DNA Sub-Strand Matrix, Merkaba, Psonns, the Trinity Keys and Photo-Dynamic Manifesting

- **SPIRITUAL INTEGRATION**
- **PRIMAL LIFE FORCE CURRENTS-The Kundalini-Antahkahrana, Maharata, Kee-Ra-ShA & Khundaray**
- **DNA SUB-STRAND MATRIX, FIRE LETTERS, PSONNS & THE TRIBAL SHIELD**
- **MERKABA &THE SALUTATIONS**
- **2 MASTERS KUNDALINI ACTIVATIONS-The Emerald & Amethyst Awakenings**
- **UNIVERSAL TRINTIY KEYS & MERKABA KEY CODES**
- **MAHARA MANIFESTING,DNA, KATHARA & PSONNS**

TECHNIQUES:

TECHNIQUE # 6: The Maharic-24 Quick Seal
TECHNIQUE # 7: Ps-13 Tribal Shield Activation
TECHNIQUE # 8: Merkaba Salutations
TECHNIQUE # 9: Merkaba Trinity Key Induction
TECHNIQUE # 10: Photo-Radionic Manifesting
TECHNIQUE # 11: Photo-Sonic Manifesting
TECHNIQUE # 12: Kee-Ra-ShA Activation Sequence-Awakening the Flame

Sacred Sciences of the Mu'a Grail Line-
Tools of the Trade

- **The Maharata-** *Universal Life Force Currents, Kundalini, Antahkarana, Maharata, Kee-Ra-Shay, Khundaray and the Maharic Seal.*
  Field Technique-1: ***Maharic Quick Seal***

- **The Tribal Shield and Cue Zones-** *DNA, Fire Letters, the Sacred Psonns and Cue Zones*
  Field Technique-2: ***Tribal Shield-Cue Zone Activation***

- **The Sacred Salutations-** *Electrical & Magentic Merkaba Activation*
  Field Technique-3: ***Sacred Salutations****- Merkaba Tone Dancing*

- **Expediting Spiritual Actualization -** *Tribal Shield Activation, Emerald and Amethyst Awakening Masters Kundalini Activations, DNA Template Bio-Regenesis, Master Key Codes, Melchizedek Cloister Ordinations*

Universal Life Force Currents

One 15-Dimensional Time Matrix with 5 Harmonic Universes

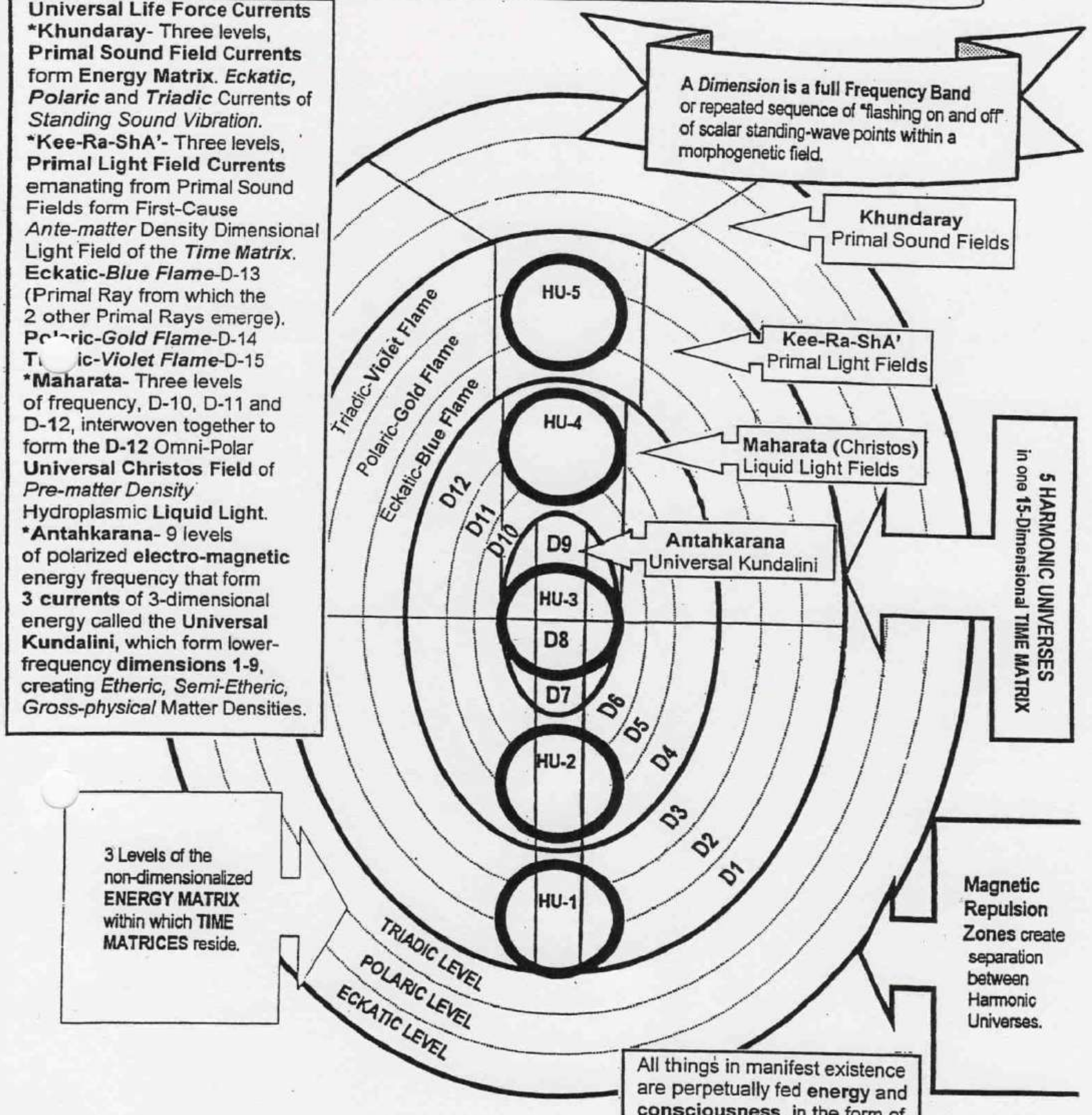

Universal Life Force Currents
Khundaray**- Three levels, **Primal Sound Field Currents** form **Energy Matrix**. ***Eckatic, Polaric and ***Triadic*** Currents of *Standing Sound Vibration.*
Kee-Ra-ShA'**- Three levels, **Primal Light Field Currents** emanating from Primal Sound Fields form First-Cause *Ante-matter* Density Dimensional Light Field of the ***Time Matrix.
Eckatic-*Blue Flame*-D-13 (Primal Ray from which the 2 other Primal Rays emerge).
Polaric-*Gold Flame*-D-14
Triadic-*Violet Flame*-D-15
***Maharata-** Three levels of frequency, D-10, D-11 and D-12, interwoven together to form the **D-12** Omni-Polar **Universal Christos Field** of *Pre-matter Density* Hydroplasmic **Liquid Light**.
***Antahkarana**- 9 levels of polarized **electro-magnetic** energy frequency that form **3 currents** of 3-dimensional energy called the **Universal Kundalini,** which form lower-frequency **dimensions 1-9**, creating *Etheric, Semi-Etheric, Gross-physical* Matter Densities.

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

3 Levels of the non-dimensionalized **ENERGY MATRIX** within which **TIME MATRICES** reside.

Magnetic Repulsion Zones create separation between Harmonic Universes.

All things in manifest existence are perpetually fed **energy** and **consciousness**, in the form of **interdimensional frequency**, via the **Universal Life Force Currents.**

Khundaray
Primal Sound Fields
Kee-Ra-ShA'
Primal Light Fields
Maharata (Christos)
Liquid Light Fields
Antahkarana
Universal Kundalini
5 HARMONIC UNIVERSES
in one 15-Dimensional TIME MATRIX
Triadic-Violet Flame
Polaric-Gold Flame
Eckatic-Blue Flame
HU-5
HU-4
HU-3
HU-2
HU-1
D12
D11
D10
D9
D8
D7
D6
D5
D4
D3
D2
D1
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL

Universal Life Force Currents and the Embodied Kathara Grid Core Template

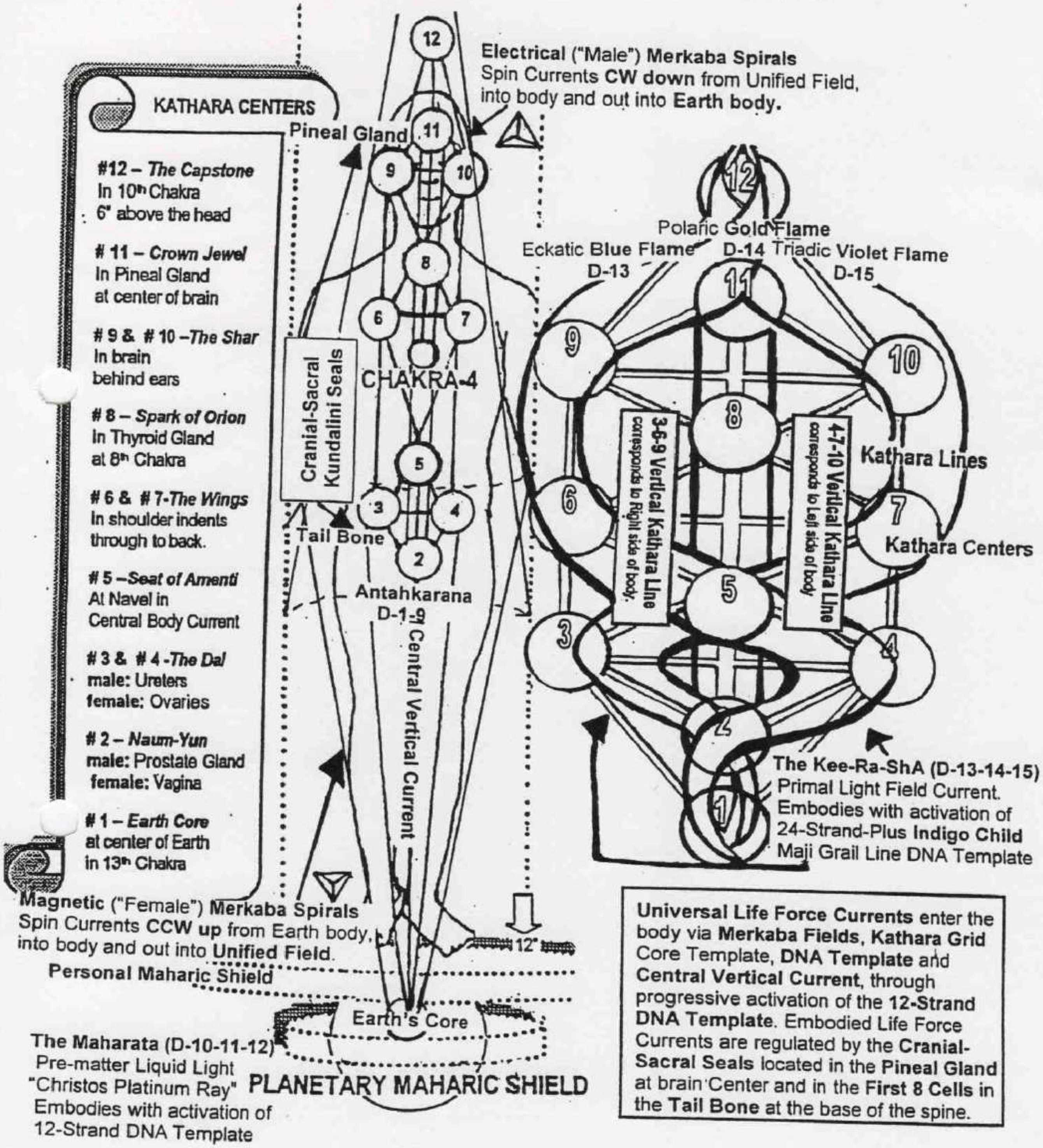
12
Electrical ("Male") Merkaba Spirals
Spin Currents CW down from Unified Field, into body and out into Earth body.
Pineal Gland
11
9
10
8
6
7
CHAKRA-4
Cranial-Sacral Kundalini Seals
5
3
4
Tail Bone
2
Antahkarana D-1-9
Central Vertical Current
Polaric Gold Flame
D-14
Eckatic Blue Flame
D-13
Triadic Violet Flame
D-15
3-6-9 Vertical Kathara Line
corresponds to Right side of body.
4-7-10 Vertical Kathara Line
corresponds to Left side of body
Kathara Lines
Kathara Centers
1
The Kee-Ra-ShA (D-13-14-15)
Primal Light Field Current.
Embodies with activation of 24-Strand-Plus Indigo Child Maji Grail Line DNA Template
Magnetic ("Female") Merkaba Spirals
Spin Currents CCW up from Earth body, into body and out into Unified Field.
12"
Personal Maharic Shield
Earth's Core
The Maharata (D-10-11-12)
Pre-matter Liquid Light
"Christos Platinum Ray"
Embodies with activation of 12-Strand DNA Template
PLANETARY MAHARIC SHIELD

KATHARA CENTERS

#12 – *The Capstone*
In 10th Chakra
6" above the head

# 11 – *Crown Jewel*
In Pineal Gland
at center of brain

# 9 & # 10 –*The Shar*
In brain
behind ears

# 8 – *Spark of Orion*
In Thyroid Gland
at 8th Chakra

# 6 & # 7-*The Wings*
In shoulder indents
through to back.

# 5 –*Seat of Amenti*
At Navel in
Central Body Current

# 3 & # 4 -*The Dal*
male: Ureters
female: Ovaries

# 2 – *Naum-Yun*
male: Prostate Gland
female: Vagina

# 1 – *Earth Core*
at center of Earth
in 13th Chakra

Universal Life Force Currents enter the body via **Merkaba Fields, Kathara Grid** Core Template, **DNA Template** and **Central Vertical Current**, through progressive activation of the **12-Strand DNA Template**. Embodied Life Force Currents are regulated by the **Cranial-Sacral Seals** located in the **Pineal Gland** at brain Center and in the **First 8 Cells** in the **Tail Bone** at the base of the spine.

Universal Life Force Currents and Level-2 Kathara Crystal Seals Grid & Chakras

Chakras (C & MC)

Star Crystal Seals (S)

15 Chakras (C & MC)
7 Primary Chakras (C)
C1- Base
C2- Sacral
C3-Solar Plexus
.C4- Heart
C5- Throat
C6- 3rd Eye- Pituitary
C7- Crown-Pineal
8 Morphogenetic Chakras (MC)
MC8-Thymus
MC9-Thalamus
MC10- Galactic 1
MC11-Galactic 2
MC12- Earth Star
MC13-Earth Core
MC14-Universal 1
MC15-Universal 2

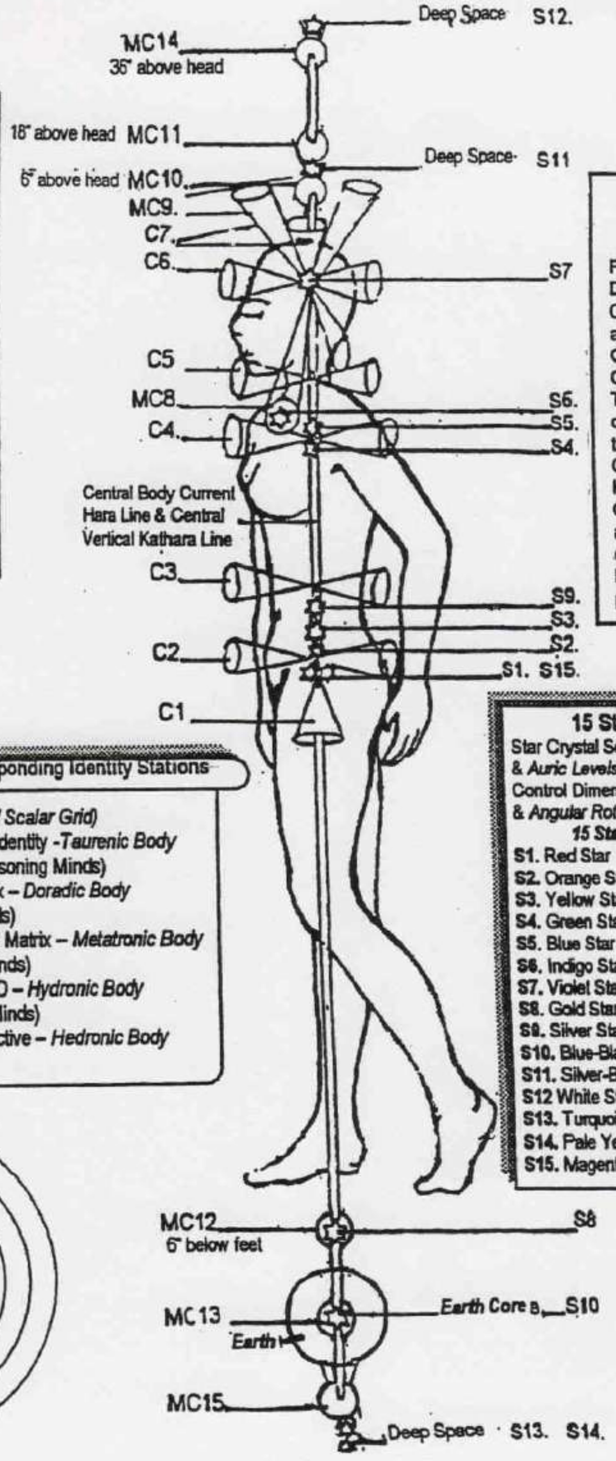
Deep Space S12.
MC14
36" above head
18" above head MC11
Deep Space S11
6" above head MC10
MC9.
C7.
C6.
S7
C5
MC8
C4.
S6.
S5.
S4.
Central Body Current Hara Line & Central Vertical Kathara Line
C3
S9.
S3.
C2
S2.
S1. S15.
C1
MC12
6" below feet
S8
MC13
Earth
Earth Core
S10
MC15
Deep Space · S13. S14.

Universal Life Force Currents
Enter via **Merkaba Fields, Kathara Grid, DNA Template & Central Vertical Current** and are regulated by the **Crystal Seals** to form the **Chakra System, Axi-a-Tonal and Meridians** of energy distribution in the manifest body. As Currents activate in the Kathara Grid, DNA, Crystal Seals, Chakras and body, **frequency and consciousness** from the higher-dimensional fields progressively embodies.

15 Star Crystal Seals
Star Crystal Seals keep human identity & *Auric Levels* dimensionally separate. Control Dimensional *Merkaba Field Axis & Angular Rotation of Particle Spin.*
15 Star Crystal Seals (S):
S1. Red Star
S2. Orange Star
S3. Yellow Star
S4. Green Star
S5. Blue Star
S6. Indigo Star (Blue-Violet)
S7. Violet Star
S8. Gold Star
S9. Silver Star
S10. Blue-Black Star
S11. Silver-Black Star
S12 White Star
S13. Turquoise Star
S14. Pale Yellow Star
S15. Magenta-Pink Star

Hova Bodies and Corresponding Identity Stations

(*"Hova" means Tri-tonal Scalar Grid*)
1. Nada Hova – HU-1 Incarnate Identity -*Taurenic Body* (Subconscious, Instinctual & Reasoning Minds)
2. Alphi Hova – HU-2 Soul Matrix – *Doradic Body* (Astral, Archetype & Angelic Minds)
3. Betcha Hova –HU-3 OverSoul Matrix – *Metatronic Body* (Ketheric, Monadic & Kertatric Minds)
4. Mahara Hova – HU-4 Avatar ID – *Hydronic Body* (Christiac, Buddhaic & Nirvanic Minds)
5. Raja Hova – HU-5 Rishi Collective – *Hedronic Body* (Universal Conscious Mind)

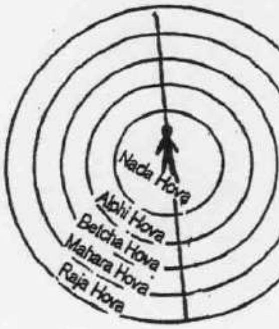
Nada Hova
Alphi Hova
Betcha Hova
Mahara Hova
Raja Hova

Universal Life Force Currents form and sustain the "**Auric Field" Levels** of 15-Dimensional Human Anatomy

Auric Level, Chakra and Hova Body Correspondences

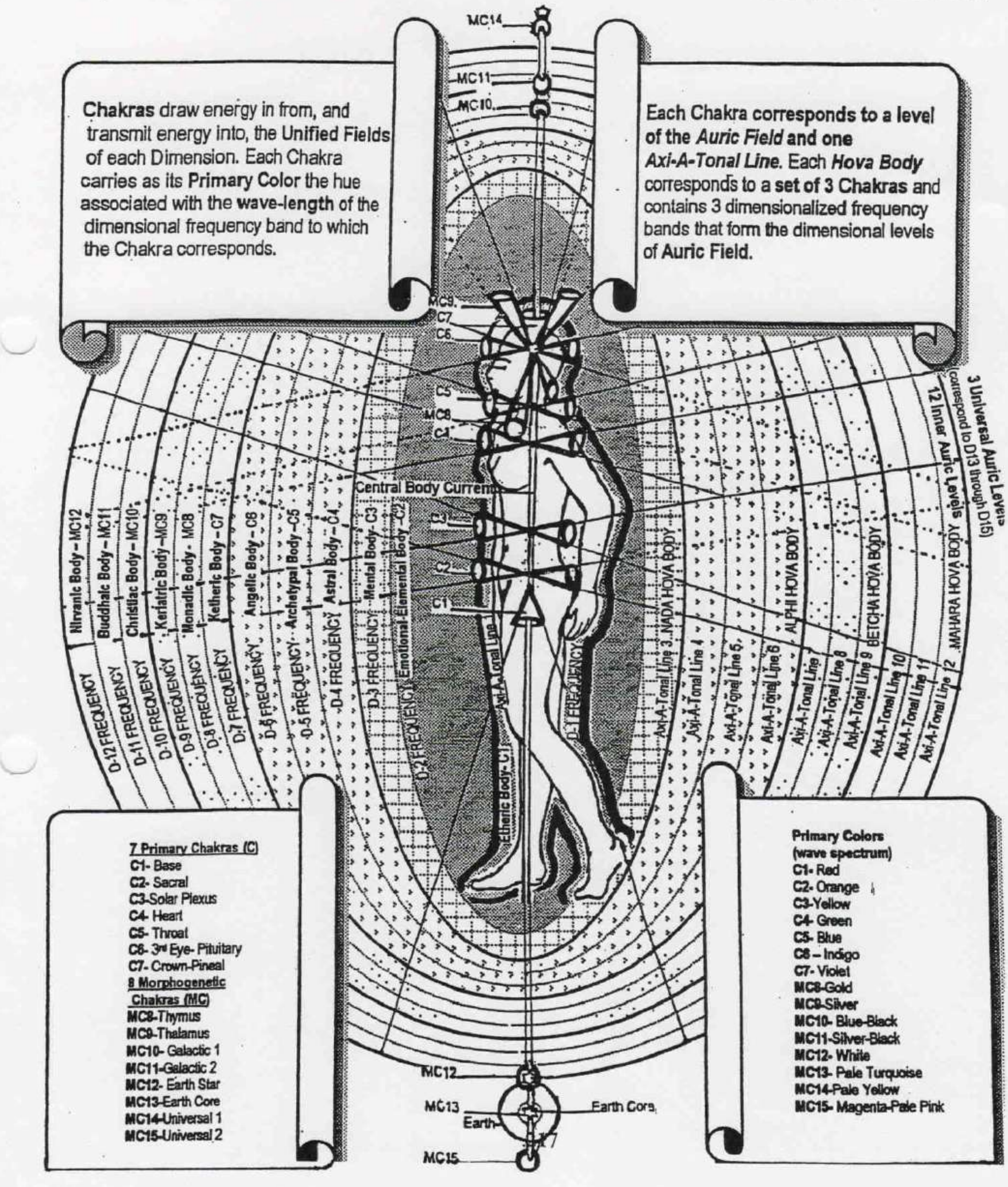

Chakras draw energy in from, and transmit energy into, the **Unified Fields** of each Dimension. Each Chakra carries as its **Primary Color** the hue associated with the **wave-length** of the dimensional frequency band to which the Chakra corresponds.

Each Chakra corresponds to a level of the ***Auric Field*** **and one** ***Axi-A-Tonal Line***. Each ***Hova Body*** corresponds to a **set of 3 Chakras** and contains 3 dimensionalized frequency bands that form the dimensional levels of **Auric Field.**

7 Primary Chakras (C)
C1- Base
C2- Sacral
C3-Solar Plexus
C4- Heart
C5- Throat
C6- 3rd Eye- Pituitary
C7- Crown-Pineal
8 Morphogenetic Chakras (MC)
MC8-Thymus
MC9-Thalamus
MC10- Galactic 1
MC11-Galactic 2
MC12- Earth Star
MC13-Earth Core
MC14-Universal 1
MC15-Universal 2

Primary Colors (wave spectrum)
C1- Red
C2- Orange
C3-Yellow
C4- Green
C5- Blue
C6 – Indigo
C7- Violet
MC8-Gold
MC9-Silver
MC10- Blue-Black
MC11-Silver-Black
MC12- White
MC13- Pale Turquoise
MC14-Pale Yellow
MC15- Magenta-Pale Pink

MC14
MC11
MC10
MC9
C7
C6
C5
MC8
C4
Central Body Current
C3
C2
C1
Nirvanic Body – MC12
Buddhaic Body – MC11
Christiac Body – MC10
Keriatric Body – MC9
Monadic Body - MC8
Ketheric Body – C7
Angelic Body – C6
Archetypal Body – C5
Astral Body – C4
Mental Body - C3
Emotional-Elemental Body – C2
Etheric Body - C1
D-12 FREQUENCY
D-11 FREQUENCY
D-10 FREQUENCY
D-9 FREQUENCY
D-8 FREQUENCY
D-7 FREQUENCY
D-6 FREQUENCY
D-5 FREQUENCY
D-4 FREQUENCY
D-3 FREQUENCY
D-2 FREQUENCY
D-1 FREQUENCY
Axi-A-Tonal Line
Axi-A-Tonal Line 3
NADA HOVA BODY
Axi-A-Tonal Line 4
Axi-A-Tonal Line 5
Axi-A-Tonal Line 6
ALPHI HOVA BODY
Axi-A-Tonal Line 7
Axi-A-Tonal Line 8
Axi-A-Tonal Line 9
BETCHA HOVA BODY
Axi-A-Tonal Line 10
Axi-A-Tonal Line 11
Axi-A-Tonal Line 12
MAHARA HOVA BODY
12 Inner Auric Levels
3 Universal Auric Levels (correspond to D13 through D15)
MC12
MC13
Earth
Earth Core
MC15

Universal Life Force Currents and the Embodied Shields, Signets and Kathara Grid

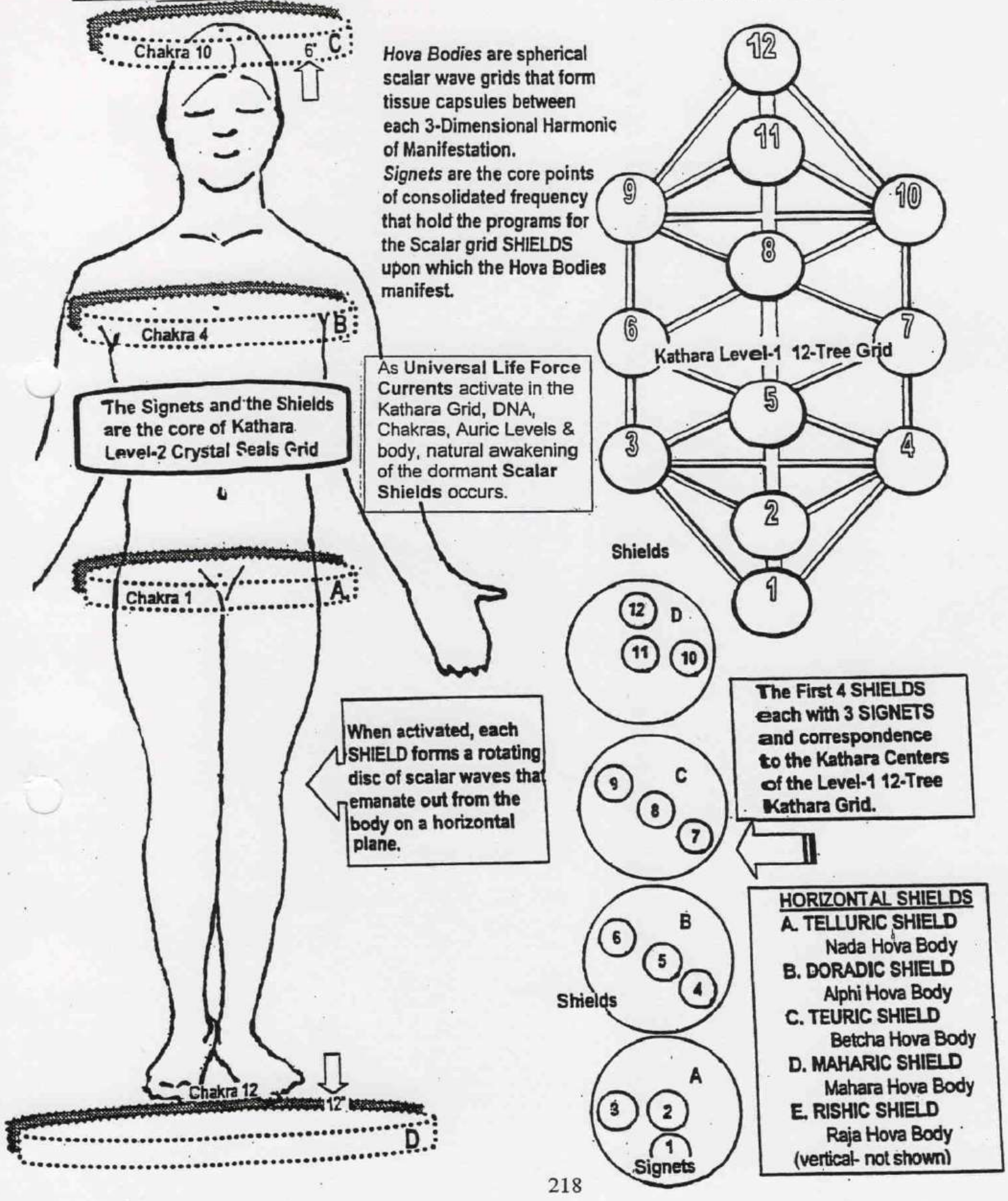
Chakra 10
6"
C
Chakra 4
B
Chakra 1
A
Chakra 12
12"
D
12
11
9
10
8
6
7
Kathara Level-1 12-Tree Grid
5
3
4
2
1
Shields
12
D
11
10
9
C
8
7
6
B
5
4
Shields
A
3
2
1
Signets

Hova Bodies **are spherical scalar wave grids that form tissue capsules between each 3-Dimensional Harmonic of Manifestation.**

Signets **are the core points of consolidated frequency that hold the programs for the Scalar grid SHIELDS upon which the Hova Bodies manifest.**

The Signets and the Shields are the core of Kathara Level-2 Crystal Seals Grid

As **Universal Life Force Currents** activate in the Kathara Grid, DNA, Chakras, Auric Levels & body, natural awakening of the dormant **Scalar Shields** occurs.

When activated, each SHIELD forms a rotating disc of scalar waves that emanate out from the body on a horizontal plane.

The First 4 SHIELDS each with 3 SIGNETS and correspondence to the Kathara Centers of the Level-1 12-Tree Kathara Grid.

HORIZONTAL SHIELDS

A. TELLURIC SHIELD
Nada Hova Body
B. DORADIC SHIELD
Alphi Hova Body
C. TEURIC SHIELD
Betcha Hova Body
D. MAHARIC SHIELD
Mahara Hova Body
E. RISHIC SHIELD
Raja Hova Body
(vertical- not shown)

The Silicate Matrix 12 Strand DNA
Primary Correspondences to Subtle Body Anatomy

Silicate Matrix 12-Strand DNA
Primary Correspondences to Subtle Body Anatomy.

DNA Strand #	Dimension #	Axi-A-Tonal Line	Chakra & *Seed* Crystal Seal	Auric Field Level
NADA HOVA BODY – TELLURIC SHIELD				
1	1	1	1	Etheric Body
2	2	2	2	Emotional Body
3	3	3	3	Mental Body
ALPHI HOVA BODY – DORADIC SHIELD				
4	4	4	4	Astral Body
5	5	5	5	Archetype Body
6	6	6	6	Angelic Body
BETCHA HOVA BODY – TEURIC SHIELD				
7	7	7	7	Ketheric Body
8	8	8	8	Monadic Body
9	9	9	9	Keriatric Body
MAHARA HOVA BODY – MAHARIC SHIELD				
10	10	10	10	Christiac Body
11	11	11	11	Buddhaic Body
12	12	12	12	Nirvanic Body

Three 3-Dimensional Kundalini Currents compose the 9-Dimensional Antahkarana **The Telluric, Doradic and Teuric Kundalini Currents**

Maharata D-10-11-12 Christos Current

As **Universal Life Force Currents** activate within the body, the **dimension of consciousness** contained in each **Auric Field Level** progressively embodies.

Universal Life Force Currents and DNA Template Fire Letter Sequences

15-Dimensional Levels of Identity, DNA and Indigo Children

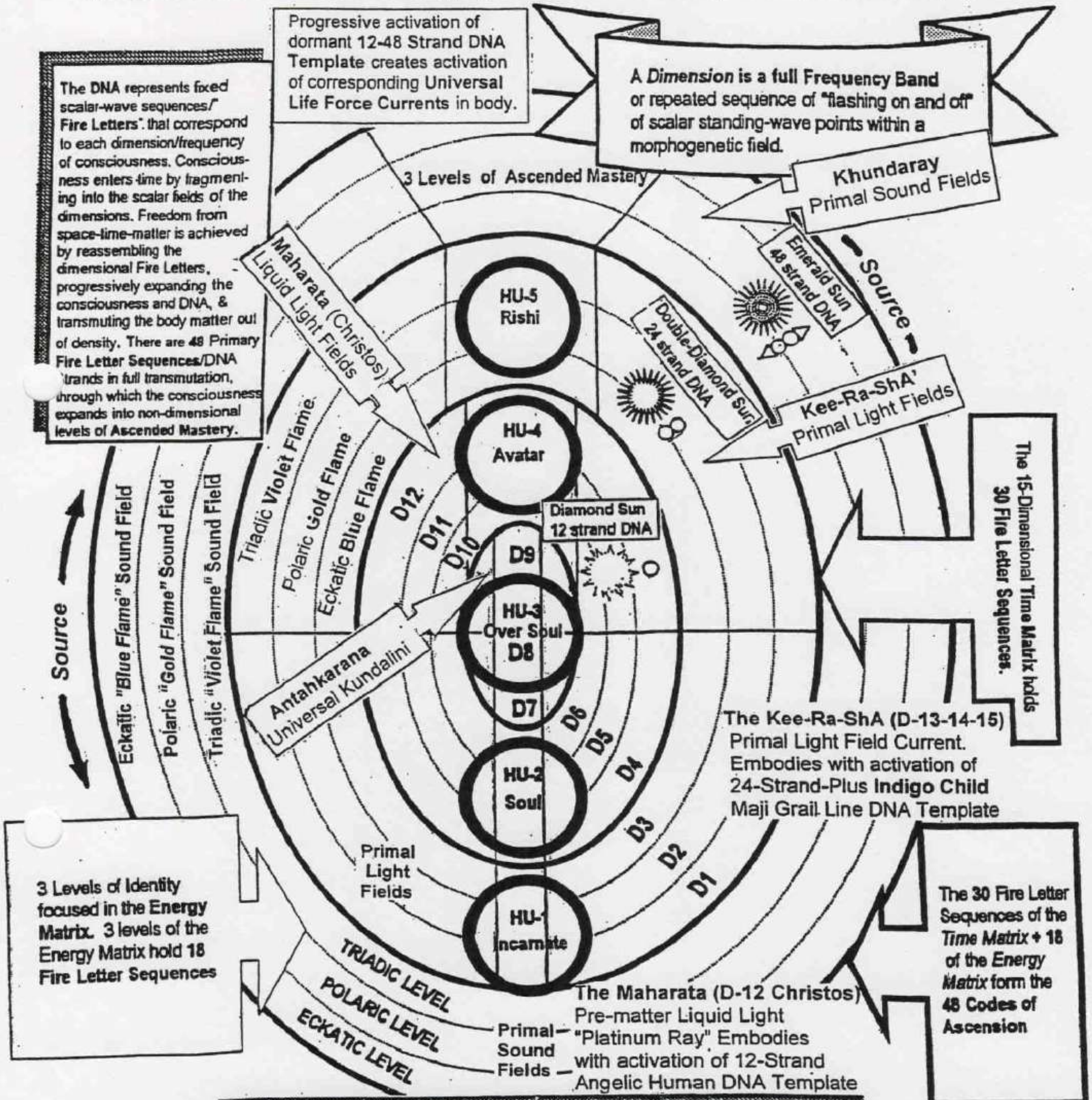

Progressive activation of dormant 12-48 Strand DNA Template creates activation of corresponding Universal Life Force Currents in body.

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

The DNA represents fixed scalar-wave sequences/"**Fire Letters**" that correspond to each dimension/frequency of consciousness. Consciousness enters time by fragmenting into the scalar fields of the dimensions. Freedom from space-time-matter is achieved by reassembling the dimensional Fire Letters, progressively expanding the consciousness and DNA, & transmuting the body matter out of density. There are 48 Primary **Fire Letter Sequences**/DNA trands in full transmutation, through which the consciousness expands into non-dimensional levels of **Ascended Mastery.**

The Kee-Ra-ShA (D-13-14-15) Primal Light Field Current. Embodies with activation of 24-Strand-Plus **Indigo Child** Maji Grail Line DNA Template

3 Levels of Identity focused in the **Energy Matrix.** 3 levels of the Energy Matrix hold 18 Fire Letter Sequences

The Maharata (D-12 Christos) Pre-matter Liquid Light "Platinum Ray" Embodies with activation of 12-Strand Angelic Human DNA Template

The 30 Fire Letter Sequences of the *Time Matrix* + 18 of the *Energy Matrix* form the 48 Codes of Ascension

The 15-Dimensional Time Matrix holds 30 Fire Letter Sequences.

3 Levels of Ascended Mastery
Khundaray
Primal Sound Fields
Emerald Sun
48 strand DNA
Source
Double-Diamond Sun
24 strand DNA
Kee-Ra-ShA'
Primal Light Fields
Maharata (Christos)
Liquid Light Fields
HU-5
Rishi
HU-4
Avatar
Diamond Sun
12 strand DNA
D12
D11
D10
D9
HU-3
Over Soul
D8
D7
D6
D5
D4
D3
D2
D1
HU-2
Soul
HU-1
Incarnate
Triadic Violet Flame
Polaric Gold Flame
Eckatic Blue Flame
Eckatic "Blue Flame" Sound Field
Polaric "Gold Flame" Sound Field
Triadic "Violet Flame" Sound Field
Source
Antahkarana
Universal Kundalini
Primal Light Fields
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL
Primal Sound Fields

The *Oraphim* genetic imprint of the **Indigo Children** allows more rapid re-assembly of the **48 DNA Fire Letters**/ scalar-wave grids, creating the potential to embody the *Avatar* identity and open the consciousness to *Ascended Master* Identity Levels. **Indigo Children are Avatars and Ascended Masters waiting to awaken.**

All Angelic Humans possess a minimum of **12-Strand** DNA potential through which activation of the **Maharata Current** and embodiment of the D-12 **Christed Avatar identity** can take place.

The Reality of Spiritual Integration

- **Spiritual Actualization** is as much a function of **Divine Physics** as it is a product of **Divine Consciousness.** The natural Primal Order of **Universal Structure** sets the **Templates of Manifestation** to which consciousness **must conform to** enter the experience of manifest expression within a space-time-matter system.
- The personal Christos **Maharic Shield** Manifestation Template and the Christos **Tribal Shield** Species Manifestation Template are **microcosmic replicas** of the macrocosmic Primal Order upon which Universal and Cosmic structure are built..
- The Personal D-12 Christos Divine Blueprint Maharic Shield connects each Individual Christos self to the **larger** D-12 Christos Divine Blueprint of the **Species.** The D-12 Christos Divine Blueprint is called the **TRIBAL SHIELD.**
- To understand the realities of **personal Spiritual Integration and Actualization,** in order to **fulfill our personal life purpose,** it is important to comprehend the basic **Primal Order** structures of energy, the **personal D-12 Maharic Shield** and the **D-12 Species Tribal Shields,** through which our consciousness manifests in time.
- The secrets to **reclaiming the Divine Birthrights and Responsibilities** of Angelic Human heritage are hidden within the realities of the **Tribal Shield** and the specific **"Fire Letter Sequences"** of the Tribal Shield that manifest within and **govern the function** of the Angelic Human **DNA Template.**
- The **key to Spiritual Actualization** is activation of the dormant personal 12-Strand DNA Template. The **most rapid means of** activating the dormant Angelic Human 12-Strand DNA Template is through **activation of the Tribal Shield,** the Species D-12 Christos Divine Blueprint, which represents the **CORE PROGRAMMING** of the personal D-12 Maharic Shield.

The Silicate Matrix 12-Strand DNA Template with Hova Body, Scalar Shield and Identity Level Correspondences

The Morphogenetic Field Scalar-wave Template for the Original Human Genome

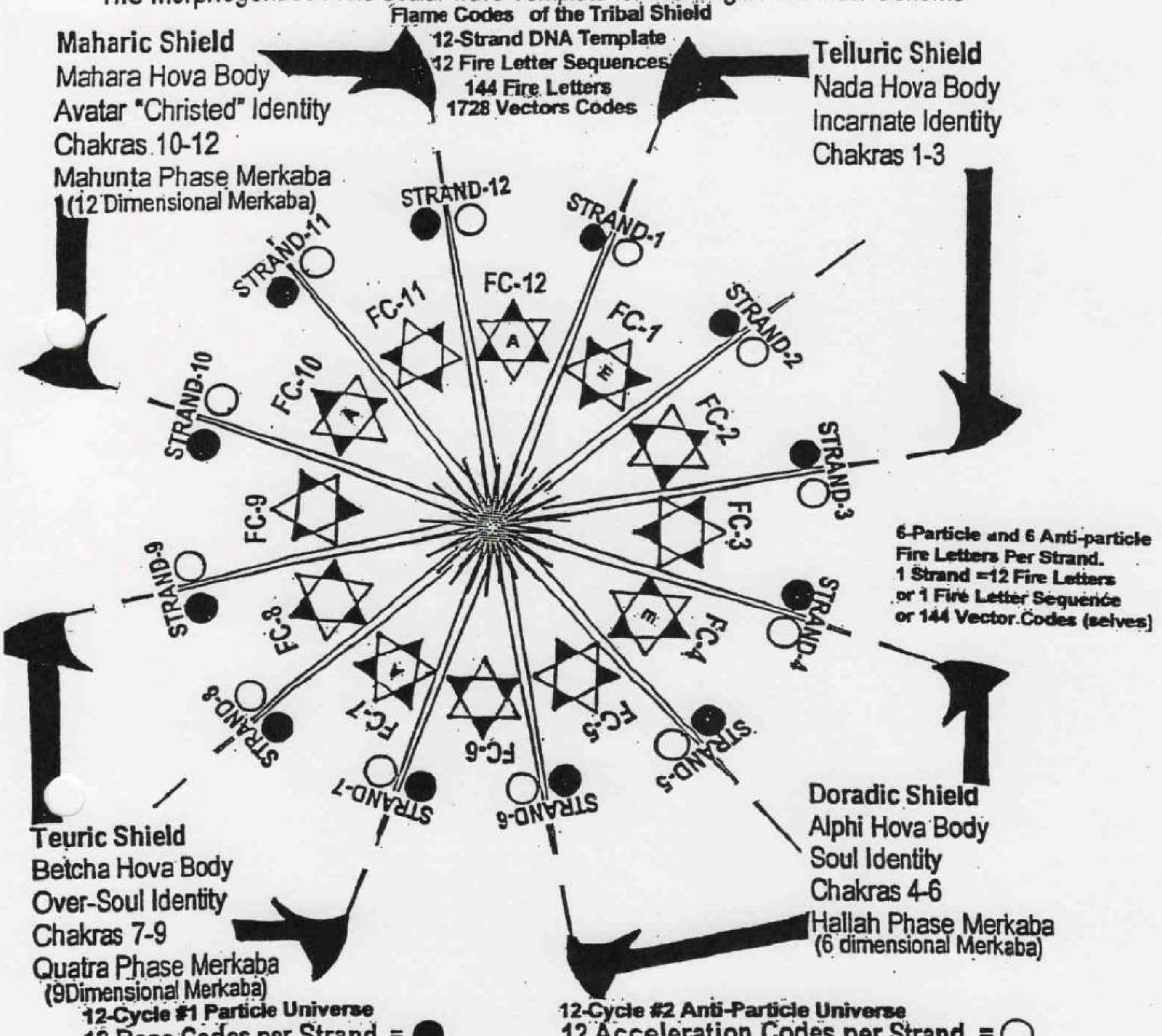
Flame Codes of the Tribal Shield
12-Strand DNA Template
12 Fire Letter Sequences
144 Fire Letters
1728 Vectors Codes
Maharic Shield
Mahara Hova Body
Avatar "Christed" Identity
Chakras 10-12
Mahunta Phase Merkaba
(12 Dimensional Merkaba)
Telluric Shield
Nada Hova Body
Incarnate Identity
Chakras 1-3
STRAND-12
STRAND-1
STRAND-2
STRAND-3
STRAND-4
STRAND-5
STRAND-6
STRAND-7
STRAND-8
STRAND-9
STRAND-10
STRAND-11
FC-12
FC-1
FC-2
FC-3
FC-4
FC-5
FC-6
FC-7
FC-8
FC-9
FC-10
FC-11
A
E
E
A
A
6-Particle and 6 Anti-particle Fire Letters Per Strand.
1 Strand =12 Fire Letters
or 1 Fire Letter Sequence
or 144 Vector Codes (selves)
Doradic Shield
Alphi Hova Body
Soul Identity
Chakras 4-6
Hallah Phase Merkaba
(6 dimensional Merkaba)
Teuric Shield
Betcha Hova Body
Over-Soul Identity
Chakras 7-9
Quatra Phase Merkaba
(9Dimensional Merkaba)

12-Cycle #1 Particle Universe
12 Base Codes per Strand = ●

12-Cycle #2 Anti-Particle Universe
12 Acceleration Codes per Strand = ○

12 FIRE CODES between Strands = ✡

(E) Emerald Kundalini Awakening releases Fire Codes 1 & 4 (A) Amethyst Kundalini Awakening releases Fire Codes 7,10 & 12

(Diagram presents simple conceptualization of DNA Strand Template orientation, not actual geometrical arrangement of wave-form structure and observable strand interrelationship.)

Activating the Tribal Shield

- The 12-dimensional **Tribal Shield** is the dormant **Species Divine Blueprint** within the **DNA Template core** of the Angelic Human races of the 4 Evolutionary Rounds. The frequencies of energy and dimensions of consciousness of the **1728 simultaneous selves** of the Personal Eternal Christos Identity can be **activated in one incarnation** through **sequential activation** of the **Fire Letter Sequences** or "**Flame Codes'** of the **Tribal Shield.**
- The "Flame Code Fire Letters" within the DNA Template are the **DNA Signet Codes** that correspond to Earth's **12 Primary Star Gates** and the **Planetary Shields** through which the planetary body manifests. The Flame Codes serve as the core foundation of the personal 12-Strand DNA Template.
- The DNA Template Flame Codes of the Tribal Shield allow for the **Primal Life Force Currents** of the Density-5 (dimensions 13-14-15) **Kee-Ra-ShA Primal Light Fields** and the **Khundaray Primal Sound Fields** from the Energy Matrix, beyond the 15-Dimensional Time Matrix, to embody more fully within the Angelic Human form.
- A **Fire Letter** is a **fixed point of consolidated frequency** that holds specific frequencies consciousness into manifest form. The **Manifestation Templates** or Divine Blueprints of **all things** manifest are built upon **intricate patterns** of interwoven Fire Letters called **Fire Letter Sequences**. The Flame Codes of the Angelic Human personal DNA Template and Species Tribal Shield are composed of the **same Fire Letter Sequences** that make up the Divine Blueprint of **Earth's Planetary Shields.**
- Activation of the Tribal Shield sets in motion simultaneous "firing" or activation of the **144 Fire Letters (12** Fire Letter **Sequences)** of the 12-Strand DNA Template, expediting the natural evolutionary process of Soul, Over-Soul and Christos Avatar Identity Integration, through progressively opening the "Trans-time DNA Template Star Gates".
- Activating the 12-Strand DNA Template allows the D-12 frequency of the **Universal "Christos" Maharata Current** to activate within and run through the Angelic Human DNA Template and into the Earth's **Planetary Shields.** Activation of the **Maharata Current** within **groups of Angelic Humans** on Earth allows for **collective anchoring** of the **D-12 Planetary Christos Divine Blueprint**, the **Shield of Aramatena**, to achieve fulfillment of Angelic Humanity's **Sacred Commission** of the **Planetary Christos Realignment Mission**.

The Tribal Shield- *Species D-12 Christos Divine Blueprint*

The Angelic Human 12-Strand DNA Template contains
6 Fire Letter Sequences from particle universe 12-Cycle
plus 6 Fire Letter Sequences from anti-particle 12-Cycle
= **12 Fire Letter Sequences** or **144 Fire Letters** from the
Two Planetary 12-Cycles of the Cycle of the Rounds.
(1 12-Cycle = 4 Evolutionary Rounds)

The **Tribal Shield** connects our personal D-12 **Maharic Shield Divine Christos** Blueprint to the greater Christos Blueprint of the Angelic Human **Species. 144** Fire Letters (12 Fire Letter *Sequences*) of the **Tribal Shield** make up the core program of the Human **12-Strand DNA Template.**

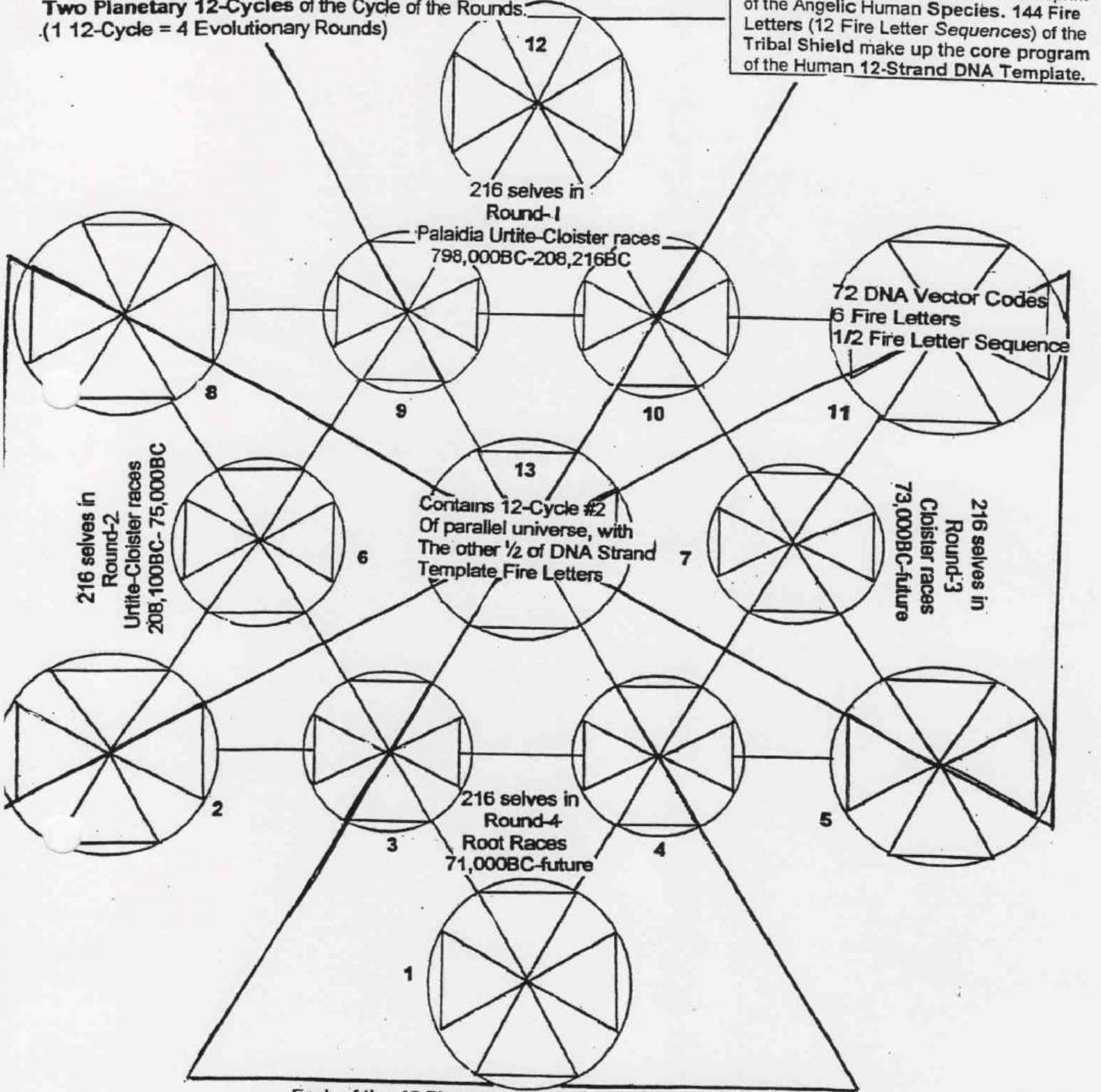
12
216 selves in
Round-1
Palaidia Urtite-Cloister races
798,000BC-208,216BC
72 DNA Vector Codes
6 Fire Letters
1/2 Fire Letter Sequence
8
9
10
11
216 selves in
Round-2
Urtite-Cloister races
208,100BC- 75,000BC
13
Contains 12-Cycle #2
Of parallel universe, with
The other ½ of DNA Strand
Template Fire Letters
6
7
216 selves in
Round-3
Cloister races
73,000BC-future
2
216 selves in
Round-4
Root Races
71,000BC-future
5
3
4
1

Each of the **12 Planetary Time Cycles** in one Planetary 12-Cycle of the **"Cycle of the Rounds" has 72 Time Vectors. 12-Strand** Angelic Humans have **one self** in each of the 72 Time Vectors of the particle universe Planetary 12-Cycle and 72 **selves** in the corresponding anti-particle Planetary 12-Cycle. **72 selves x 12** Planetary Time Cycles = **864 selves** per one Planetary 12-Cycle + **864 selves** parallel universe Planetary 12-Cycle = **1728 selves** in Tribal Shield.

Tribal Shield Activation Creates Cue Zones In Earth's Planetary Shields.

Cue Zones

"Cue Zones" are locations on Earth where a small portion of Earth's D-12 Shield of Aramatena **Christos Divine Blueprint** has been re-set within Earth's Planetary Shields. During Stellar Activations Cycles, Cue Zones can be **manually set** by individuals or groups by running the **D-12 Maharata Current**, activating the **Tribal Shield** to draw in the Khundarary Primal Sound frequencies, then singing the 12-Tribes **Master Psonns Suffixes** and running the tones through the body and feet and into Earth's Planetary Shields. After running the Master Psonns Suffixes, one can then allow the **"Song of the Christos Soul"**, the tones of the personal D-12 **Maharic Shield** (the "Elohei-Elohim Within") to run into the Planetary Shields. The personal Soul Song will create a **"Back Song"** transmission of frequency from Earth's Planetary Shields, a set of healing frequencies sent back to you in response to activating your Soul Song in Earth's Planetary Shields. Once set, Cue Zones will remain for a period of **3.5 to 5 years**, depending on strength of energy originally set. They can be **recharged** using the same "setting" procedure to **last indefinitely.**

Cue Zones are **"safe zone"** areas on the Planet that are protected from interdimensional interference. They can be used to amplify effects of energy **healing applications**, **to "charge" objects and substances** with sub-harmonics of D-12 Christos frequency and when used with the personal Temporary Maharic Seal, to get clear, protected **"meditation space"** free from interdimensional interference. Cue Zones are "keyed" to Earth's Planetary **Templar "Cue Sites"**, the activation sites for Earth's 12 Primary Star Gates, which exist on regions of Earth that correspond to the 12 Primary Star Gates of **Inner Earth.**

TECHNIQUE # 6: THE MAHARIC QUICK SEAL

1 Work with a 24 pointed, 3 Dimensional Star (Heirophant) and be aware that you will be working with beings known as the Breneau Rishi ...

2 Begin by slowing your breathing , as you imagine or visualize a 24 point, 3 -Dimensional, Heirophant at the Pineal Gland in the center of your brain. Visualise the Heirophant surrounded by a Pale Blue Sphere of Light.

(The Blue Sphere serves as a buffer for your readiness to accept the frequencies associated with using the 24 pointed Star; the Blue Sphere holds the energy for you until your body can handle the energy. If you can handle it, the energy will simply pass through the blue sphere).

3 INHALE, as if you are going to grab the blue sphere containing the 24 point star located at the Pineal, and on the EXHALE move the 24 point star/ blue sphere, all the way down to Earth's Core. Try to hear a sound tone as the Heirophant/ Sphere hits the Planetary Shield ...

4 INHALE and draw the Heirophant/ sphere up to your Personal Maharic Shield 12" below your feet ... and EXHALE while watching the Pale Silver-Blue sphere expand out to a large sphere ... and watch as your Maharic Shield pop out as a disc, about 4' in diameter, Pale Silver with a light coating of Pale Blue.

5 Bring your attention to the center of your Maharic Shield 12" below your feet ... and INHALE the 24 point Star (only) up into the Heart Chakra. EXHALE, expanding the frequencies of the 24 point Star into the Heart Chakra.

6 INHALE to grab the 24 point Star, and EXHALE, pushing the Star up into the 14th Chakra 36" above the head. Imagine that the 24 point Star is spinning in a clockwise direction in the 14th Chakra. As you do this, try to feel/ sense the energy around you, just a few inches out from your body ...

7 Now, put your attention into your Maharic Shield 12" below your feet, take a couple of relaxing breaths and INHALE deeply pulling the Pale Silver and Blue energy up, as if you are pulling on your Maharic Seal to latch it on to your 14th Chakra. Try and feel the sensations just a few inches from your body now – feel the difference if you can. This procedure gives you a Quick Seal.

8 Now envision the cord of Pale Silver Maharic Light that would normally come up with your Shield and focus on the Earth's Core. Begin drawing energy all the way up the 4" diameter cord and into the body. INHALE the energy up into the Heart Center, and expand it there. Begin to form a pale Silver ball of Maharic frequency.

On each EXHALE send your energy down to the Earth's Core to bring up another load of Maharic frequency, up from the Planetary Shields, expand it into the Heart center and repeat several times. Notice that the Cord grows larger as you do this, expanding from approximately 4" to 6 to 8" diameter, until finally it feels like a skin around (as well as within) your body ... as you load your Astral field with Maharic Frequency.

9 Move your attention to the Pale Silver ball you have created in your Heart Center, INHALE and move it up to the 14th Chakra 36" above your head .. as you do, feel for the sheath of Maharic Frequency encasing your etheric body, close to your skin. (This Maharic/ Christos skin charges, and saturates, your etheric body – feel the peacefulness and all-knowing nature of this frequency).

Technique # 7: Ps-13 Tribal Shield Activation

The Master Key Tones

- The **12-Tribes** names were the **audible-tone translations** of the **specific Fire Letter Sequences** contained within the **Tribal Shield DNA Templates** of each Tribe. The tones of the Tribal names were used to activate the **144 Fire Letters** of the Tribal Shield in the **personal DNA Template**, providing the Angelic Humans races with the ability to **consciously regulate** the **activation level** of their **DNA Templates** and **Primal Life Force Currents.**
- The **Sound-Tone Programs** that are used to activate the DNA Template and Primal Life Force Currents are called **"The Sacred Psonns"**. The **Tribal name** was the **Master Psonn**. In running the Rainbow Roundtables, the Signet Council Regents brought their Tribal Shield "Flame Codes" into activation by **singing rounds** of the **Sacred Master Psonns.**

The 12 Sacred Master Psonns

for activating the 144 Fire Letters of the Tribal Shield Flame Codes in the 12-Strand DNA Template.

Plus **Master Tone Activation Suffix.**

12. *A-reah-Azurta*
a- RI-a-Zoor- ta- Rha

11. *Zephar-Duun-Atur*
ze-far-Doon a-Tur- Dha

10. *Ma'ah-hu-ta*
Ma-a hoo ta- Khu

9. *Yun Zu-Xen*
Yu-Un Zoo-Zen KE

8. *Chia Zhun Zan La-Yung*
ChE' ah-Zoon Yan LA-Yoong' OM

7.*Mahata-Agrah*
ME hah' ta a'g-ra OE

6. *Ramyana-Shridveta*
rah ma yah na shrid vE' Da U

5. *Ionatu-Etillah*
I O' Na too et il' a EU

4. *Nuagu Hali*
Noo ah' goo ha' LE Ka

3. *Amekasan-Etur*
a ME' ka sun e too'r DO

2. *Maahali-Bruea*
Ma a ha' LE - BrU' A EL

1. *Isutu-Esheau*
I sU' too E' shoo UR

Tribal Shield-Cue Zones Activation

Creating a Cue Zone

1. **Choose location** and **activate** a **personal "Maharic Quick Seal"** to "run the D-12 **Maharata Current**" in your **body.**
2. Once you can feel sensation of Maharata Current running, sing-tone 12-Tribes **Master Psonns** (names) for **Tribe-12** and **Tribe-1** to activate **Tribal Shields Flame Codes** in your DNA Template to run the **"Rainbow Ray"**. Direct Pale Silver Maharata Current and Rainbow Ray pale color spectrum through body and out **soles of feet** into Earth's Planetary Shields. Tone for a while to build critical mass of frequency.
3. Use a selected combination of **3 Master Psonns Suffixes** and tone for a while, running sound through body and feet into Earth's grids. (avoid the Suffix combination of Tribe-4-7-10 used in that sequence to avert present distortions in Earth's grids that run on this mathematical coding)
4. Allow the **Christos Soul Song** to emerge from you "free style" and send the tones into Earth's grids.
5. When complete, sit quietly for a while and listen for the Earth's telepathically transmitted **"Back Song"**, drawing its gentle frequencies into your body for healing.
6. Close with singing several rounds of the **3 Master Psonns Suffixs** that you opened with. The Cue Zone you have set will extend out from your position in a radius of about **5 yards** and will last **3.5-5 years** depending on strength of energy you used in setting the Cue Zone.
7. To **amplify** strength of Cue Zone you can use the **"Merkaba Spin"** just after closing. Remain in same position, stand, the gently spin, first **clockwise** (as if you are standing on a clock face with the "6" in front of your feet, "3" at the left hand, "9" at the right and "12" directly behind you). Spin gently clockwise to bring electrical frequency down through your **"top"** Merkaba spiral. Then slow, stop and **reverse spin** to **counterclockwise**, to activate **"bottom"** Merkaba spiral", drawing frequency up from Earth's grids. Personal Merkaba Activation will accelerate the frequencies of the D-12 energies you set at the site, making the Cue Zone **stronger and more lasting.**

The Solar Salutation
Electrical Merkaba Activation

- The ancient **Mu'a** and **Lemurian** Guardian Races ran the **Maharata** and activated the **Tribal Shield** each day to conduct the **Sacred Solar and Lunar Salutation Rites** each morning and evening. The **Sacred Salutations** are **Toning Dances** that activate the personal **Merkaba Field** for sustained activation of the DNA Template and amplification of the Maharata and Universal Life Force Currents in the body.
- **Solar Salutations** accelerate the spin of the **Electrical Harmonic Merkaba Spiral** (Top, CW spin), drawing increased amounts of **Electrical Anti-particle** frequency from the Solar **Electrical Merkaba Spiral** light spectrum into the **Electrical Acceleration Codes** in the **DNA Template to "charge" the body** with **Electrical Anti-particle frequency.**
- Accelerated **Electrical Merkaba Spiral Spin** of **Solar Salutations** increases the amount of **Anti-particle Base-Electrical** frequency **brought into** the body's **DNA Template** and **Kathara Grid** from the Dimensional Unified Fields and increases the amount of **Anti-particle** frequency **sent out** from the body into Earth's **Planetary Shields.** Frequency passes between the Human Body and Earth's Planetary Shields via the embodied **Kathara Grid** core templates that make up the **Planetary Bio-Feed Interface System** ("PBIS")
- **Electrical Anti-particle** energy enters the body via the **"Top" Electrical** Harmonic and Dimensional **Merkaba Spirals**, the **Base-Electrical Acceleration Codes** in the DNA Template, the **Electrical Kathara Centers 2-5-8-11** of the Kathara Grid **Central Pillar**, the **Electrical aspect** of **Electro-magnetic Kathara Centers 1-3-6-9-12** on the **Right Side** of body ("Left Side Pillar" in Kathara diagram), **Electrical Chakras 2-5-8-11** and the **Electrical aspect** of **Electro-magnetic Chakras 1-3-6-9-12**.

- **Solar Salutations** (CW spin Top Electrical Merkaba Field Activation) **raise** the body's **Oscillation** (energy transmission-expansion), **lower** the body's **Vibration** (energy holding-contraction), making the body **Higher Frequency** and **More Electrical** (transmits more energy than it is drawing in).
- A **Higher Frequency Electrical Body** has:
  - **Lower Vibration**, Higher Oscillation/Frequency, **Less Matter Density**
  - More activated DNA Strands, more chromosomes, embodiment of more levels of consciousness.
  - **Slower** heart, breathing and metabolic rhythms (energy expansion and contraction rates),thus requires less food (burns fuel slower).
  - Greater **longevity** and **less sleep** (body can synthesize natural higher frequency energy and Stream of Consciousness for longer periods without taxing.)
  - Greater **Silica quantity** in **Silica-to-Carbon Ratio** of elemental composition and **larger size** (more space between cells; more expanded matter).
  - Perception of **expanded Space** and **slower Time** movement.
  - Greater **heat release**, thus **cooler** internal temperature.
  - Natural attribute of **transmitting** greater amounts of energy thrust into the Planetary Shields than it is drawing in/receiving from them. **Energy "giving".**

- The **Solar Salutations** "charge" the **Top Electrical CW Merkaba Spiral** to maintain its natural cycle of **33(⅓) rotations per 1 "trillionth" of a nano-second"**, through which the **Base-Electrical** systems of the body are naturally maintained.

The Lunar Salutation
Magnetic Merkaba Activation

- **Lunar Salutations** accelerate the spin of the **Magnetic Harmonic Merkaba Spiral** (Bottom, CCW spin), drawing increased amounts of **Magnetic Particle** frequency from the **Lunar Magnetic Merkaba Spiral** reflected light spectrum and Earth's Planetary Shields into the **Magnetic Base Codes** in the **DNA Template** to **"charge" the body** with **Magnetic Particle frequency.**
- Accelerated **Magnetic Merkaba Spiral Spin** of **Lunar Salutations** increases the amount of **Particle Base-Magnetic** frequency **brought into** the body's **DNA Template** and **Kathara Grid** from the **Earth's Planetary Shields** and increases the amount of **Particle** frequency **sent out** from the body into Dimensional Unified Fields.
- **Particle Base-magnetic** energy enters the body via the **"Bottom" Magnetic** Harmonic and Dimensional **Merkaba Spirals**, the **Base-Magnetic Base Codes** in the DNA Template, the **Magnetic Kathara Centers 4-7-10** of the Kathara Grid on the **Left Side** of the body ("Right Side Pillar" on Kathara Diagram), the **Magnetic aspect** of **Electro-magnetic Kathara Centers 1-3-6-9-12** on the **Right Side** of body ("Left Side Pillar" in Kathara diagram), **Magnetic Chakras 4-7-10** and the **Magnetic aspect** of **Electro-magnetic Chakras 1-3-6-9-12**.
- **Lunar Salutations** (CCW spin Bottom Magnetic Merkaba Field Activation) **lower** the body's **Oscillation** (energy transmission-expansion), **raise** the body's **Vibration** (energy holding-contraction), making the body **Lower Frequency** and **More Magnetic** (draws in more energy than it is transmitting out).

- A **Lower** Frequency **Magnetic Body** has:
    - **Higher Vibration**, Lower Oscillation/Frequency, Greater **Matter Density**
    - **Fewer** activated DNA Strands, fewer chromosomes, embodiment of fewer levels of consciousness.
    - **Faster** heart, breathing and metabolic rhythms (energy expansion and contraction rates), thus requires **more food** (burns fuel **faster**).
    - Lesser **longevity** and needs more **sleep** (body can synthesize natural higher frequency energy and Stream of Consciousness for **shorter periods**; taxes more easily)
    - Greater **Carbon quantity** in **Silica-to-Carbon Ratio** of elemental composition and **smaller size** (less space between cells; more contracted matter).
    - Perception of **contracted or more confined Space** and **faster Time** movement.
    - Less **heat release**, thus **warmer** internal temperature.
    - Natural characteristic of **drawing** greater amounts of energy thrust from the Planetary Shields than it is transmitting into them. **Energy "taking"**.

> The **Lunar Salutations** "charge" the **Bottom Magnetic CCW Merkaba Spiral** to maintain its natural cycle of **11⅔ rotations per 1 "trillionth" of a nano-second"**, through which the **Base-Magnetic** systems of the body are naturally maintained.

The **natural ratio** between the **Top-CW-Electrical-"Male" Harmonic Merkaba Field** and the **Bottom-CCW-Magnetic-"Female" Harmonic Merkaba Field** is **Electrical-Top 33⅓CW** rotations per 1 "trillionth" of a nano-second and **Magnetic-Bottom 11⅔CCW** rotations per 1 "trillionth" of a nano-second.

The Natural Christos Merkaba Vehicle & Personal Shields

© 2000 Ashayana Deane

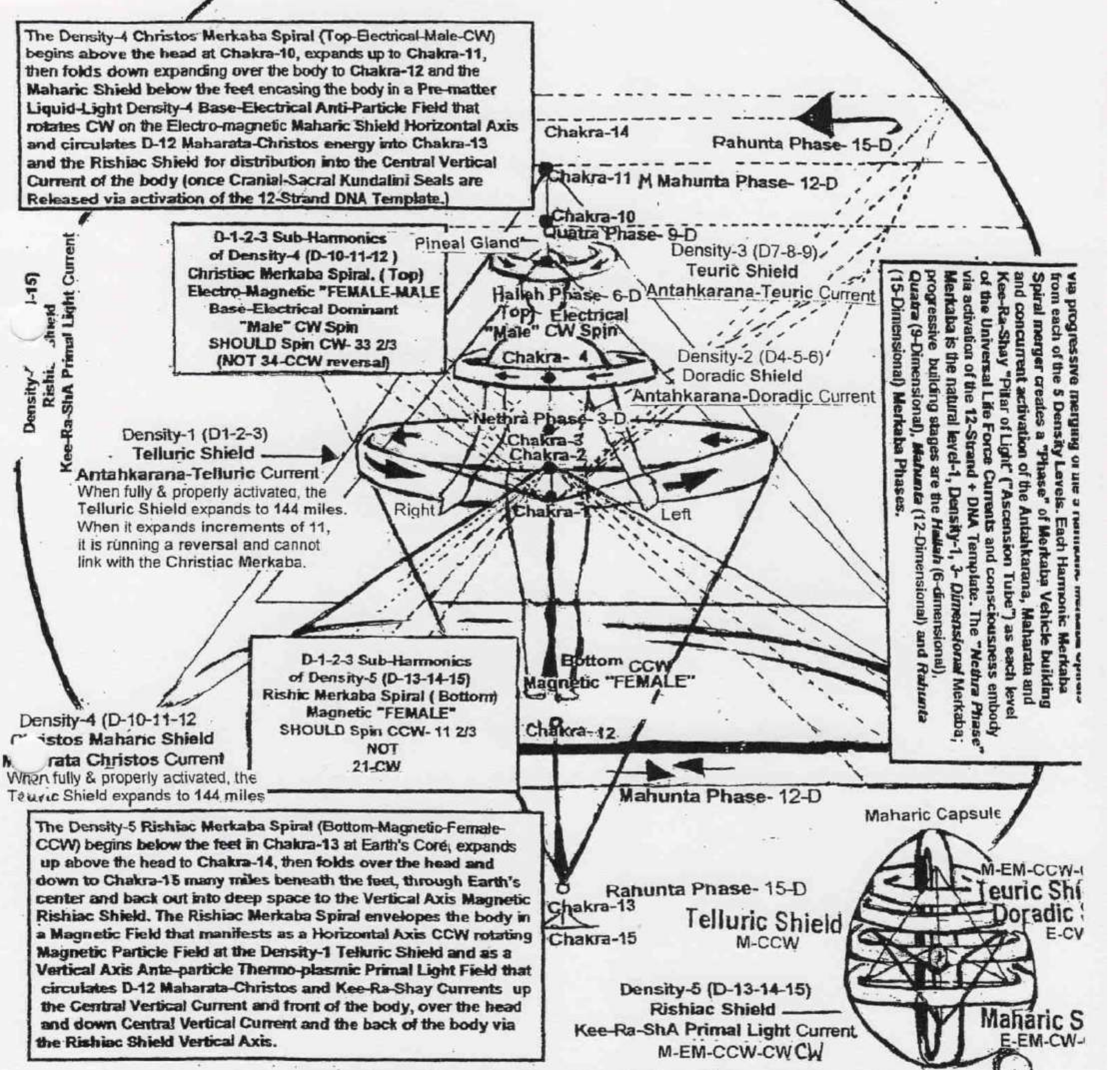
Chakra-14
Rahunta Phase- 15-D
Chakra-11 M Mahunta Phase- 12-D
Chakra-10
Quatra Phase- 9-D
Pineal Gland
Density-3 (D7-8-9)
Teuric Shield
Antahkarana-Teuric Current
Hallah Phase- 6-D
Top Electrical
"Male" CW Spin
Chakra- 4
Density-2 (D4-5-6)
Doradic Shield
Antahkarana-Doradic Current
Nethra Phase- 3-D
Chakra-3
Chakra-2
Chakra-1
Right
Left
Density-1 (D1-2-3)
Telluric Shield
Antahkarana-Telluric Current
Density-/ 1-15)
Rishi Shield
Kee-Ra-ShA Primal Light Current
Bottom CCW
Magnetic "FEMALE"
Chakra-12
Density-4 (D-10-11-12
istos Maharic Shield
rata Christos Current
Mahunta Phase- 12-D
Maharic Capsule
Rahunta Phase- 15-D
Chakra-13
Chakra-15
Telluric Shield
M-CCW
Density-5 (D-13-14-15)
Rishiac Shield
Kee-Ra-ShA Primal Light Current
M-EM-CCW-CW
M-EM-CCW-
Teuric Sh
Doradic
E-CV
Maharic S
E-EM-CW-

The Density-4 Christos Merkaba Spiral (Top-Electrical-Male-CW) begins above the head at Chakra-10, expands up to Chakra-11, then folds down expanding over the body to Chakra-12 and the Maharic Shield below the feet encasing the body in a Pre-matter Liquid-Light Density-4 Base-Electrical Anti-Particle Field that rotates CW on the Electro-magnetic Maharic Shield Horizontal Axis and circulates D-12 Maharata-Christos energy into Chakra-13 and the Rishiac Shield for distribution into the Central Vertical Current of the body (once Cranial-Sacral Kundalini Seals are Released via activation of the 12-Strand DNA Template.)

D-1-2-3 Sub-Harmonics
of Density-4 (D-10-11-12)
Christiac Merkaba Spiral. (Top)
Electro-Magnetic "FEMALE-MALE
Base-Electrical Dominant
"Male" CW Spin
SHOULD Spin CW- 33 2/3
(NOT 34-CCW reversal)

When fully & properly activated, the Telluric Shield expands to 144 miles. When it expands increments of 11, it is running a reversal and cannot link with the Christiac Merkaba.

via progressive merging of the 3 Harmonic Merkaba Spirals from each of the 5 Density Levels. Each Harmonic Merkaba Spiral merger creates a "Phase" of Merkaba Vehicle building and concurrent activation of the Antahkarana, Maharata and Kee-Ra-Shay "Pillar of Light" ("Ascension Tube") as each level of the Universal Life Force Currents and consciousness embody via activation of the 12-Strand + DNA Template. The *"Nethra Phase"* Merkaba is the natural level-1, Density-1, 3- *Dimensional* Merkaba; progressive building stages are the *Hallah* (6-dimensional), *Quatra* (9-Dimensional), *Mahunta* (12-Dimensional) and *Rahunta* (15-Dimensional) Merkaba Phases.

D-1-2-3 Sub-Harmonics
of Density-5 (D-13-14-15)
Rishic Merkaba Spiral (Bottom)
Magnetic "FEMALE"
SHOULD Spin CCW- 11 2/3
NOT
21-CW

When fully & properly activated, the Teuric Shield expands to 144 miles

The Density-5 Rishiac Merkaba Spiral (Bottom-Magnetic-Female-CCW) begins below the feet in Chakra-13 at Earth's Core, expands up above the head to Chakra-14, then folds over the head and down to Chakra-15 many miles beneath the feet, through Earth's center and back out into deep space to the Vertical Axis Magnetic Rishiac Shield. The Rishiac Merkaba Spiral envelopes the body in a Magnetic Field that manifests as a Horizontal Axis CCW rotating Magnetic Particle Field at the Density-1 Telluric Shield and as a Vertical Axis Ante-particle Thermo-plasmic Primal Light Field that circulates D-12 Maharata-Christos and Kee-Ra-Shay Currents up the Central Vertical Current and front of the body, over the head and down Central Vertical Current and the back of the body via the Rishiac Shield Vertical Axis.

The NATURAL Spin Ratio between the Top (Male-Electrical-CW-Anti-particle) Density-4 Christiac Merkaba Spiral and the Bottom (Female-Magnetic-CCW-Particle) Density-5 Rishiac Merkaba Spiral is 33 1/3-CW-Male-Top to 11 2/3-CCW-Female-Bottom. The 33 1/3-CW/11 2/3-CCW Spin Ratio creates the NATURAL ratio of 33 2/3-Parts-Electrical Anti-Particles to 11 2/3-Parts Magnetic Particles, (or 33 1/3 energy expansion/oscillation/thrust to 11 2/3 energy contraction/vibration/draw) within earth's Density-1 Matter Base, the organic Divine Blueprint of Density-1. This Christiac Internal Merkaba Vehicle Spin Ratio creates a self-sustaining, Eternal, Matter-base that is MORE ELECTRICAL than Magnetic, which perpetually re-fuels itself via natural circulation of the *Maharata*, *Kee-Ra-ShA* and *Khundaray* Universal Life Force Currents.

Merkaba Phases – Building the Merkaba Vehicle

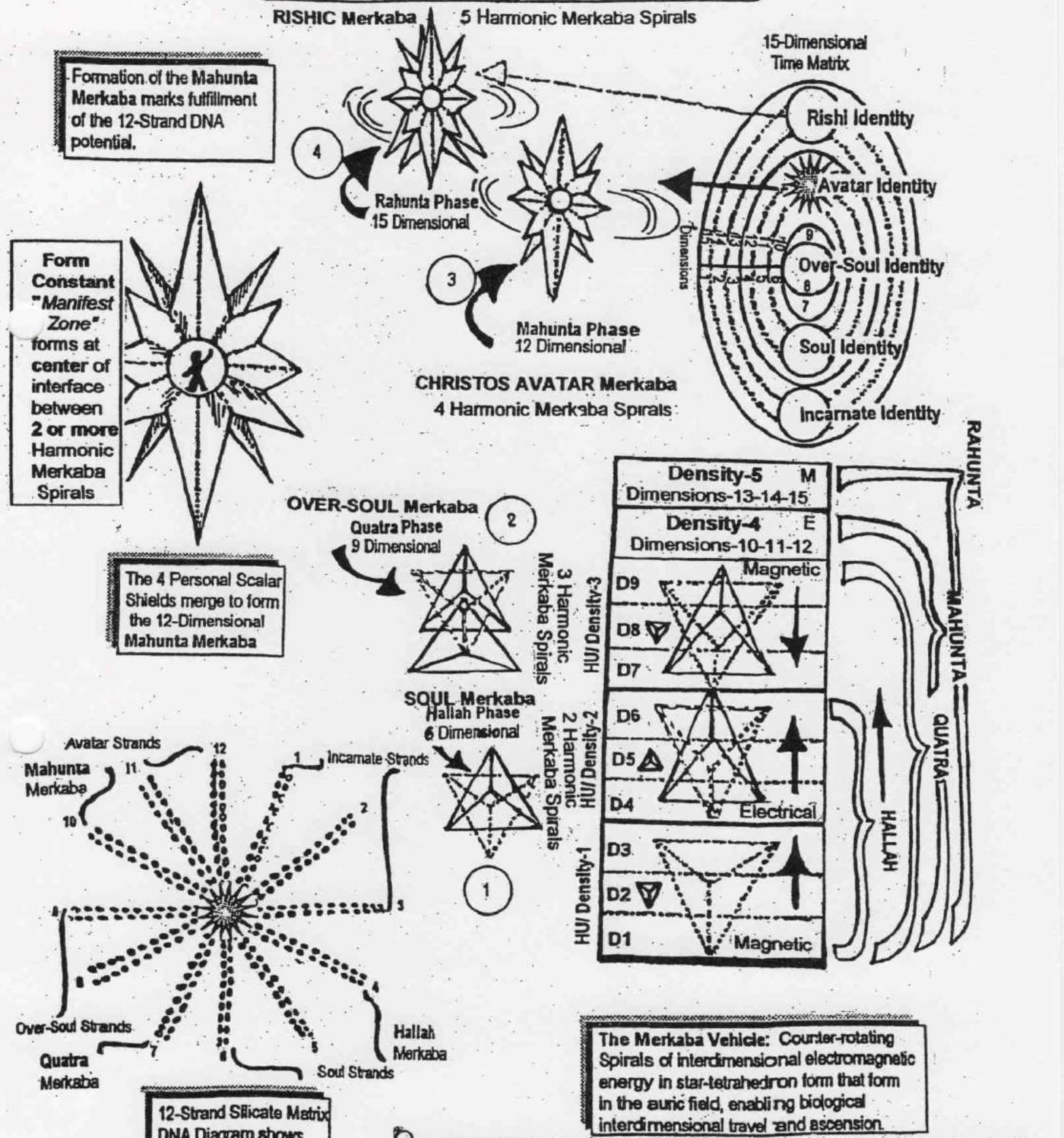

RISHIC Merkaba
5 Harmonic Merkaba Spirals
15-Dimensional Time Matrix
Rishi Identity
Avatar Identity
Over-Soul Identity
Soul Identity
Incarnate Identity
Dimensions
9
8
7
4
Rahunta Phase
15 Dimensional
3
Mahunta Phase
12 Dimensional
CHRISTOS AVATAR Merkaba
4 Harmonic Merkaba Spirals
OVER-SOUL Merkaba
Quatra Phase
9 Dimensional
2
3 Harmonic Merkaba Spirals
SOUL Merkaba
Hallah Phase
6 Dimensional
2 Harmonic Merkaba Spirals
1
Density-5 M
Dimensions-13-14-15
Density-4 E
Dimensions-10-11-12
HU/ Density-3
D9
D8
D7
Magnetic
HU/ Density-2
D6
D5
D4
Electrical
HU/ Density-1
D3
D2
D1
Magnetic
RAHUNTA
MAHUNTA
QUATRA
HALLAH
Avatar Strands
Mahunta Merkaba
Incarnate Strands
Over-Soul Strands
Quatra Merkaba
Soul Strands
Hallah Merkaba
1
2
3
4
5
6
7
8
9
10
11
12

Formation of the Mahunta Merkaba marks fulfillment of the 12-Strand DNA potential.

Form Constant *"Manifest Zone"* forms at center of interface between 2 or more Harmonic Merkaba Spirals

The 4 Personal Scalar Shields merge to form the 12-Dimensional Mahunta Merkaba

12-Strand Silicate Matrix DNA Diagram shows Strand relationships

The Merkaba Vehicle: Counter-rotating Spirals of interdimensional electromagnetic energy in star-tetrahedron form that form in the auric field, enabling biological interdimensional travel and ascension.

Each dimension of the Human Body has *one* **Dimensional Merkaba Field** (EM set). The *three* **Dimensional Merkaba Fields** in *one* **Density Level** form *one* **Harmonic Merkaba Spiral** in each of **5 Density Levels** that **merge** to form the **Merkaba** Star Gate Transit (Ascension) **Vehicle.**

Technique # 8: Merkaba Salutations

Healing the Spin & Balancing Particle/ Anti-Particle Ratios

33 1/3-CW/11 2/3-CCW Christos Merkaba Activation

***Note:** Discern **CW** or **CCW direction of Merkaba Spin** by imagining that you are standing, with arms stretched out to the sides, in the center of a round 12-digit clock-face. The **FRONT** of your body faces the **"6"-digit**, **"12"-digit** is positioned **behind you,** the **"3"-digit** is positioned at your **LEFT ARM** and the **"9"-digit** is positioned at your RIGHT **ARM.** **CLOCKWISE "CW" SPIN:** Hold LEFT **"3-Arm"** firmly out to the side of the body like a **"clock hand"** and begin **Spin** toward FRONT, moving the body with the LEFT **arm** toward the **"6"-digit** position in **FRONT** of your feet. COUNTER-CLOCKWISE "CCW" SPIN: Hold RIGHT **"9-Arm"** firmly out to the side of the body like a **"clock hand"**. and begin **Spin** toward FRONT, moving the **body** with the **RIGHT arm** toward the **"6"-digit** position in **FRONT** of your feet. The **TOP-ELECTRICAL-MALE** Merkaba Spiral **Christos Spin** is always **33 1/3-CW** rotation. The **BOTTOM-MAGNETIC-FEMALE** Merkaba Spiral **Christos Spin** is always **11 2/3-CCW** rotation.

Solar Salutation Rites: (best time *Dawn's First Light* & *12:00 Noon*)
begin with *Top-Electrical-CW-***"Male"** **Merkaba Spiral** activation, followed by *Bottom-Magnetic-CCW-"Female"* Merkaba Spiral
Use Tribe-12 Activation Suffix **"RHA"** as ***Initiating Tone***
Lunar Salutation Rites: (best time *12:00 Midnight* to *3:00AM*)
begin with ***Bottom-Magnetic-CCW-"Female"*** **Merkaba Spiral** activation, followed by *Top-Electrical-CW-"Male"* Merkaba Spiral
Use ribe-1 Activation Suffix **"UR"** as ***Initiating Tone***

1. Activate ***Maharic Quick Seal*** Technique- 6
2. Activate ***Tribal Shield-Cue Zone*** Technique- 7
3. Stand with spine comfortably straight, breath slowly, in a relaxed manner and begin **quietly toning** the ***Initiating Tone***, continuing repetition while progressively **increasing the volume** and extending the Exhale Breath Tone thrust behind Initiating Tone until **you can feel the subtle energy** of Tone sound waves oscillating in the body.
4. **Activate First Merkaba Spiral** (Top or Bottom Spiral first, depending on Solar or Lunar Salutation) by holding **arms out to the side**, parallel with the ground, and begin to **slowly SPIN** in the direction (CW or CCW) natural to the *** Christos Spin** of the Merkaba Spiral you are activating. Continue to **Spin AND Tone Initiating Tone** while progressively **increasing rate of Spin and Toning.** To **prevent dizziness,** hold one arm out to front with thumb positioned UP, and **focus your eyes on the THUMB** rather than the spinning scenery beyond. Do not Spin faster than comfort level. You may **modulate toning in rounds of 3,** between the **Initiating Tone** and toning the **Christos Spin Rate Number** ("33 1/3" for Electrical-Top Spiral, **"11 2/3"** for Magnetic-Bottom spiral), if desired.
5. When energy in body feels **"charged"** (a "doneness" intuitive feeling), **slow spin** to a **gentle STOP**, while **Toning 3 Times**, the **"*Um Shaddai Ur'-A ah-Khum' Tun*"** Lyra-**Halls of Amorea** D-12 activation sequence, to **amplify** the Merkaba Spiral "charge. **Stand STILL** and breathe quietly for a few moments while Spin stabilizes, then **repeat STEPS 3-5** to **activate other Merkaba Spiral.**
6. **End Salutation Rite** with full **3 Rounds** (minimum) **Toning** of the ***"Ariea Khum Nar-A""*** (Song of Orion D-8 activation sequence) **Part-1:Verses 1 and 2,** to be sung while **Walking 2 Circles**, one **CW**, one **CCW**, beginning Circle Walk with Spin direction of **First Merkaba Spiral** activation, followed by a **reverse-direction** Circle, then **Standing STILL** for a few moments to allow the **Christos Maharata Current** to run between your body and Earth's Planetary Shields.

The A*r-/E- a Khum –Nar' A*
The Song of Orion

Part 1 : the Resonant Tone Syllable Sequence (primary melody)

Verse 1: OM A Da E, Patum ah
OM A Da E, Patum ah
Patum ah A' DE E - Tah, A khum-na
OM A Da E, Patum ah

Verse 2: TrA TE La E' Durum Na
TrA TE La E' Durum Na
Durum' ah A' Khem - Tah, A khum -na
OM A Da E, Patum ah

Part 2 : the Base Tone Syllable Sequence (accompaniment 1)

OM A Da E, Patum ah - Lines 1, 2 & 4 Verse 1 & Line 4 Verse 2
TrA TE La E' Durum Na - Lines 1 & 2 Verse 2
UM ah OM, A
UM ah OM, A
Patum ah A' DE E – Tah, A khum- na
Durum ah A' Khem – Tah A khum-na
DO RAY' TU Ra"
OM A Da E, Patum ah
UM ah OM, A

Part 3: the Over Tone Syllable Sequence (accompaniment 2)

OM A Da E, Patum ah – Lines 1 & 2 Verse 1
TrA TE La E' Durum Na – Lines 1 & 2 Verse 2
BE HE Tah OM, A
BE HE Tah OM, A

Patum ah A DE E – Tah, A Khum –na – Line 3 Verse1
Durum ah A Khem- Tah, A Khum - na - Line 3 Verse 2
Thra Khem' Na, A
OM A Da E, Patum ah – Line 4 Verses 1 & 2

KhrA MiM, Tah A

The fastest means of naturally activating the Personal 12-Strand DNA Template is :

1. **Tribal Shield Activation:** Activating the 144 Fire Letters of the Tribal Shield & Sacred Salutations Merkaba Activations.

2. **Emerald and Amethyst Awakening 2 Masters Kundalini Activations:** Releasing key **Fire Codes** between the DNA Strand Templates to allow for **expedited DNA Strand Braiding.**

3. **DNA Template Bio-Regenesis:** Progressive use of internally directed DNA Template Bio-Regenesis technologies for progressive purging of Strand Template mutations, restoration of the natural D-12 Christos Divine Blueprint of the 12-Strand DNA Template and expedited, accelerated activation of the DNA Strand Template Base Codes and Acceleration Codes.

4. **Master Key Codes:** Use of the personal scalar-wave-guide Symbol Code Programs that correspond directly to the core programming of the personal D-12 Maharic Shield for further sequential expedition of Strand Template Activation.

5. **Melchizedek Cloister Level-3 Regent Ordination:** Transmission of the Kee-Ra-ShA Primal Light and Khundaray Primal Sound Field **"Rainbow Ray Current"** directly into the personal **DNA Template** via direct **Chakra Induction**. Sufficient Rainbow Ray frequencies to provide a level-3 Regent Ordination can be transmitted by a Melchizedek Cloister **level-5 Elder Consummate** or **Level-6 Eckar** Indigo Child-Type-1 Maji Grail Line Angelic Humans, who have the frequencies of the Primal Creation Currents integrated into their DNA Template from birth. **Expedites all of the above** while provided additional **bio-energetic field support** for more **stable cycles** of DNA Template activation.

The Khundaray RRT Activation Sequence

A progression of **manually initiated** *Kundalini and DNA Template Activations* that systematically induce *temporary release* the **Cranial-Sacral Kundalini Seals** (Pineal Gland & Tail Bone), in their **natural Release-Sequence,** allowing for temporary and ***progressive*** embodiment (nano-second burst of frequency) of the *Primal Universal Life Force Currents* via **Syncopated Activation** of the **12-Strand DNA Template**.

The ***Khundaray RRT ("K-RRT") Activation Sequence*** allows the Density-4 (D-10-11-12 Pre-matter) ***Maharata Christos Current***, Density-5 (D-13-14-15) ***Kee-Ra-ShA Primal Light Currents*** (Ante-matter *Blue, Gold* and *Violet* "Flames") and the ***Khundaray Primal Sound Currents*** (*Eckatic, Polaric* & *Triadic* Columnar Standing Waves) from *Source* to run through the DNA Template and body for accelerated ***TRUE 12-Strand or higher DNA Template Activation** and expedited human Spiritual Integration.

For the human body to **hold and transmit** a **minimum critical mass** of Kee-Ra-ShA and Khundaray frequency to **access the Planetary Shields** and **Star Gates** for ***Masters (Universal) K-RRTs,*** the ***K-RRT Activation Sequence*** must be conducted **4 to 24 hours** prior to the K-RRT. The *K-RRT Activation Sequence* must also be run **2-24 hours** prior to the ***Amethyst Awakening Masters-2 Kundalini Activation*** to prepare the body for the Amethyst Awakening frequencies. The *Amethyst Awakening* (Field Technique-6) can also be run **just before** a K-RRT (following prior *K-RRT Activation Sequence*) to **greatly amplify the power** of the K-RRT and Signet Council group.

There are **5 Specific Operations of Energy Management** within the ***K-RRT Activation Sequence***:

1. **Activating the Doradic Phase-1 Current**
2. **Awakening the KA**
3. **The Emerald Awakening**
4. **Activating the Kee-Ra-ShA**
5. **Activating the RA Center-Anchoring the Rainbow Bridge**

The Emerald & Amethyst Awakenings
Masters Kundalini-DNA Template Activations
For Accelerated ***DNA Template Activation*** and ***Spiritual Integration***, Running **K-RRTs** and enhanced ***Immune Support***

THE KUNDALINI LIFE-CURRENTS AND THE CRANIAL-SACRAL SEALS.

The human body is fuelled by ENERGETIC LIFE-FORCE CURRENT FROM SOURCE. In contemporary human anatomy, the only life-force currents running fully through the embodied Kathara Grid and body are those corresponding to the SINGLE-VECTOR time cycle of the Density-1 TELLURIC SHIELD. Each of the 4 HOVA BODIES and their corresponding SCALAR SHIELDS, DNA STRAND TEMPLATES, AXI-A-TONAL LINES, DIMENSIONS OF CONSCIOUSNESS AND MERKABA FIELD LEVELS process the frequencies of one TRIADIC CURRENT – a 3-dimensional, Bi-polar life-force current. The life-force currents corresponding to the first 9 dimension and first 3 Hova Bodies, Mater Densities and Shields are referred to as the KUNDALINI ENERGIES. Each KUNDALINI CURRENT is a TRIADIC CURRENT – a set of 3 dimensional currents that function together and are separated by Magnetic Repulsion Zones within the embodied Kathara Grid. The 3 embodied Kundalini Currents are the TELLURIC, DORADIC and TEURIC CURRENTS .

Each of the 3 Kundalini Currents is spiralled within the regions of the 1st 8 CELLS of human conception, located in the TAILBONE at the Base of the Spine. (The incarnating consciousness anchors its identity and Christos Manifestation Template within the FIRST 8 CELLS, to initiate the FETAL INTEGRATION PROCESS.) The first Kundalini Current, the TELLURIC CURRENT that corresponds to the DENSITY-1 Nada Hova Body, is active within the human form from birth, enabling 3 dimensions of consciousness awareness, in the form of scalar frequency, to embody within the human organism. When the 2 dormant Kundalini Currents in the tailbone, the DORADIC and TEURIC CURRENTS, are activated, the frequencies of consciousness from dimensions 4-9 progressively awaken within the body, activating the dormant 4-9 strands of the 12-strand DNA Template. The 3 Primary Kundalini Spirals each contain 3 single-dimension currents, and are controlled by the corresponding SIGNET STAR CRYSTAL SEALS in the Kathara Level-2 Crystal Seals Grid. The Signet Seals create polarization between dimensionalized sets of scalar waves within the Scalar Shields, that manifest within the Kathara Level-2 Grid, First Eight Cells and DNA Template as FREQUENCY SEALS or blockages within the TAILBONE (Sacral region) and the PINEAL GLAND in the center of the brain – the CRANIAL-SACRAL SEALS.

When the Cranial-Sacral Seals are released, through activation of the Signet Star Crystal Seals that govern them, the dormant Kundalini Currents come to life within the body, sending progressively higher-frequency energy through the CENTRAL VERTICAL CURRENT of the body. Activation of the 3 Kundalini Currents progressively creates integration of the higher dimensions of consciousness and de-densification of molecular structure. When ALL of the frequencies within the 3 embodied Kundalini Currents are activated, the 4th KUNDALINI CURRENT- the PRIMAL TRIADIC CURRENT, the KEE-Ra-ShA', is activated. The Frequency Seal on the KEE-Ra-ShA' is located in the 12th Chakra, 6" below the feet. Activation of the KEE-Ra-ShA' anchors the 12th-Dimensional HYDROPLASMIC BEAM within the body, initiating the process of CELLULAR TRANSMUTATION and intentional de-manifestation and Dimensional Ascension of the human biological form.239

THE EMERALD AWAKENING AND THE KEE-Ra-ShA'

The Emerald Awakening Masters Kundalini Activation is one of two inter-harmonic energetic activations of previously dormant frequency within the body that sets in motion simultaneous activation of the 12 MAGNETIC BASE TONES within the FIRE CODES between the 12 Strands of the SILICATE MATRIX DNA Template. Activation of the 12 Fire Code Base Tones begins activation of the 3 PRIMAL CREATION FREQUENCIES – the KEE-Ra-ShA' – within the 4 Scalar Shields of the PERSONAL INNER TEMPLAR COMPLEX. Activation of the 12 DNA Fire Codes allow for progressive natural fusion of each dimensional DNA Strand Template to occur, through which the organic potentials of CELLULAR TRANSMUTATION are re-awakened within the physical body form.

Through the natural processes of Cellular Transmutation, once fully operational within human genetic design, the carbon-based elemental biology of Density-1 molecular structure is progressively and systematically transmuted to SILICA-based biology. Transmutation to Silica-based biology allows the physical body to pass through the levels of Semi-etheric, Etheric and Pre-matter Hydroplasmic Liquid-Silica Densities of Harmonics 2-4 (Dimensions 4-12), through which the consciousness continually enters manifestation, to achieve physical mastery of the organic process of Transmutational Dimensional Ascension. To accomplish mastery of the Ascension process, the SIGNET STAR CRYSTAL SEALS with the Scalar Shields, which keep the 12 dimensionalized frequency levels of the consciousness and molecular form Phase-locked into dimensional division/ polarization, must be released. Release of the Signet Star Crystal Seals allows the frequency spectra of the higher dimensional anatomy and the Primal Creation Currents of the KEE-Ra-ShA', to enter into the Template for the physical body and dormant DNA Fire Code Templates, progressively building the Merkaba Vehicle and setting the Cellular Transmutation process in motion. The process of progressive Transmutational Dimensional Ascension represents the function of the organic genetic design and evolutionary objective of the human creation blueprint, through which the natural Immortal Christos-Angelic Form of the human being is actualised.

In the evolutionary cycles of contemporary humanity, actualization of the 12-Strand DNA potential takes place slowly in Earthly terms, as the incarnate identity experiences in single-life time-vectors, several cycles of Density-1 incarnation, spanning many thousands of years, in order to progressively activate the DNA Fire Codes within the Manifestation Template. When the human genome was operating in its organic form, this process took place in one lifetime, over a period of about 44 years, as the immortal human incarnate matured to embody the 12 dimensions of consciousness within its original Christos Manifestation Template- the MAHARIC SHIELD, actualizing its organic Angelic-human heritage and mastery over 4 Densities of manifestation within the Time Matrix.

The EMERALD AWAKENING Masters Kundalini Activation releases the Density-2 4TH – Dimensional Green Signet Star Crystal Seal and the Density-1 1st-Dimensional Red Signet Star Crystal Seal, to open the Density-1 NADA HOVA BODY and TELLURIC SHIELD, and Density-2 ALPHI HOVA BODY and DORADIC SHIELD, to receive the Omni-polar frequencies of the 12th-Dimensional MAHARIC CURRENT (Hydroplasmic Beam) from the Pre-matter Christos Template Maharic Shield. Upon entering the Telluric and Doradic Shields, the 12th-Dimensional frequencies of the Maharic Current trigger simultaneous activation of the Base Tones within the 12 Fire Codes of the DNA Template, beginning the process of DE-POLARIZTION of the 12-Strands and levels of consciousness within the human Manifestation Template. Progressive de-polarization of the 12 levels of consciousness within the human Manifestation Template sets in motion the organic process of Cellular Transmutation and potential mastery over Transmutational Dimensional Ascension. The Emerald Awakening Masters Kundalini Activation initiates REGENESIS of the original function of the human genome, releasing the Base Tones of the 3 KEE-Ra-ShA' Currents within the body via opening of 2 Signet Seals, through which the physical body begins the carbon-to-silica transmutation process and is prepared to receive the higher dimensional OVER TONE frequencies of the KEE-RA-Sha', to complete Cellular Transmutation via fusion and de-polarization of the particle (Particum) and anti-particle (Partika) aspects of the Manifestation Template. This is the natural process of UNIFIED FIELD PHYSICS by which a human being becomes "Christed".

THE AMETHYST AWAKENING AND THE KHUN-DA-RAY

The final step in fulfilling the potentials of original Angelic-human design is activation of the OVER TONE frequencies of the KEE-Ra-ShA', which allows fusion between DNA Strand Templates to complete on 12 dimensions of frequency. The final step in Transmutational Dimensional Ascension is called the AMETHYST AWAKENING MASTERS KUNDALINI ACTIVATION. Through the AMETHYST AWAKENING, the DENSITY-3 7TH-Dimensional VIOLET Signet Star Crystal Seal and 8th-Dimensional GOLD Signet Star Crystal Seal, and the DENSITY-4 11TH-Dimensional DARK SILVER Signet Star Crystal Seal are released within the body's Scalar Shields. Release of the OVER TONE SIGNET SEALS allows the OVER TONES of the KEE-Ra-ShA Creation Currents to fuse with the BASE TONES, bringing the 3 OMNI-POLAR CREATION CURRENTS (Primal Creation Currents built upon a TRINITY OF TONES, BI-POLAR + NON-POLAR energy units) together as ONE PRIMAL NON-POLAR CURRENT – the KHUNDARAY (also called the "Rainbow Bridge" or the " Rainbow Serpent".)

The KHUN-Da-rAy is the PRE-LUMINAL vibration, or STANDING WAVE OF INTERNAL SOUND emitted from ONE-SOURCE/ GOD, that corresponds to the COMBINED ORIGINAL SOURCE CURRENT from which the 3 KEE-Ra-ShA' PRIMAL LIFE-FORCE CURRENTS of the 3 non-dimensionalized levels of the ENERGY MATRIX, beyond the 15-Dimensional Time Matrix emerge. The Emerald Awakening Masters Kundalini Activation, and its resulting activation of the BASE TONES of the KEE-Ra-ShA' within the human body, is the first step in manually expediting the natural process of Transmutational Dimensional Ascension.

AWAKENING THE KA AND REMOVING INORGANIC CRYSTAL IMPLANTS

Before the Emerald Awakening Kundalini Activation can be orchestrated, the MAHARIC SHIELD and the KA must be brought out of dormancy within the human Manifestation Template. Once the 12th-Dimensional Maharic Current is released within the embodied Kathara Grid, the KA can be activated. The "KA" refers to the 4 HOVA BODIES corresponding to dimensions 1 through 12. (Nada, Alphi, Betcha and Mahara Hova Bodies.) The KA becomes activated through removal of 4 inorganic Crystal Seal IMPLANTS that block the natural function of the 2nd, 3rd and 4th DNA Strand Templates. Following he procedure for AWAKENING THE KA, the Emerald Awakening Kundalini activation can be conducted.

Technique # 12: Kee-Ra-ShA Activation Sequence

ACTIVATING DORADIC PHASE I Awakening the Flame

Preparation: Activate Maharic Seal Technique-6 *Maharic Quick Seal*

1 Breath slowly and deeply for several breaths while visualizing or imagining a Violet Dot on the center of the *inside* of your forehead. Make the color image as strong as you are able (if you can't yet perceive inner colors simply imagine the Violet Dot being there).

Move the Violet Dot to the center of the brain, at the **Pineal Gland,** and then upward into the 7th Chakra, finally moving it out of your head to rest on the top of your skull, at the Crown Chakra.

2 Call to mind the Heirophant Symbol (Pale Silver Merkaba Star/ Star of David). Visualise the Heirophant spinning 12" below your feet, within your Maharic Shield, and slowly INHALE.... drawing the Pale Silver Heirophant up through the Central Body Current, up to meet the Violet Dot at the 7th Crown Chakra at the top of the head.

When the Heirophant connects with the Violet Dot, visualize the Violet Dot bursting into a 4" vertical Pillar of Violet-Silver Light. Imagine the Violet-Silver Pillar running down through the body and into the Earth's Core, as well as upward into the 14th Chakra positioned 36" above your head. (This activates the Violet Ray of Transmutation).

3 Now, move your attention to the 4th Heart Chakra, and imagine a small Green Spherical Crystal – the Green Star – stationed at it's core. This is the Green star Crystal Seal that controls release of 4th Dimensional frequency – Phase I Doradic Current – into the body. INHALE, drawing Violet-Silver energy from the Violet-Silver Pillar into the Green Star.

EXHALE, expanding the Violet-Silver energy into the Green Star sphere. Repeat two more INHALE/ EXHALE breaths, drawing Violet-Silver energy from the Pillar and expanding it into the Green Star sphere.

4 At the end of the last EXHALE, imagine the Green Star POPPING, turning into a spinning Green and Violet Starburst of Light, as the Violet-Silver Pillar disappears. Spin the Green-Violet Starburst faster until it becomes a ball of Green-Violet Light, about the size of a grapefruit, spinning the center of the Heart Chakra. (The Green-Violet ball is the storehouse for your Phase I Doradic Current; it's called the Doradic Sphere).

INHALE and draw Pale Silver energy from your Maharic Shield, through the central Body Current, and into the center of the Doradic Sphere at the heart Chakra ... encasing the sphere in Pale Silver Light.

AWAKENING the KA

1 Bring your attention back to the DORADIC SPHERE at the HEART CENTRE...

2 INHALE pale silver energy up from the EARTH CORE to fill & expand the DORADIC SPHERE; breath slow and deep 2-3 times ...

3 Now, put the tips of your fingers together, and feel the 'fuzziness' of the energy flow...

4 Next, move the fingertips to the tail bone... INHALE energy from the Maharic Shield into the DORADIC SPHERE

Continued Next Page

As you EXHALE, push PALE SILVER energy down to the finger tips and into the tail bone............... REPEAT 2-3 times and then remove the fingers.

5 Now, bring your attention to the 2nd CHAKRA (SACRAL) .....feel or visualize or just know ... a DOUBLE TERMINATED QUARTZ CRYSTAL (a crystal with a point at both ends) is beginning to "push out" of the front of your 2nd Chakra, the Sacral center.

6 Focus your attention on this, try to feel the crystal pushing.... use your hands to help ease it forward, and out.

7 There is a smaller double terminated crystal in the 3rd (Solar Plexus) Chakra; this will release on it's own because the currents have activated it to do so. Help it out with your fingers.

8 In the Heart Chakra, imagine a little sphere of energy, colored green. Bring this little green sphere more fully into focus ... and you will notice a little black cube of energy inside it ...look more closely again, and notice a smaller white cube inside the black one.

9 As you focus on the two little cubes, watch them merge together .. feel a ripple of vibration as they do, and then ... notice a little puff of smoke as they both disappear.

Now, the body is activated to open the 4th and 1st Crystal Seals; you are ready for **the EMERALD AWAKENING ...**

1 Now we will open the GREEN STAR CRYSTAL SEAL ....So, keeping your awareness on the Doradic Sphere, the GREEN-VIOLET- SILVER ball at the Heart Center.........INHALE,

2 and ... on the EXHALE visualize or imagine pushing a cord of GREEN-VIOLET-SILVER Doradic Current down into the tail bone... to the 1st Cell ...... at it's very tip .

3 Breathe easy ... and take a moment to visualize the 1st Cell at the tip of your tail bone.... as a WHITE-SILVER SPHERE...

5 Now within the WHITE SILVER SPHERE, visualize, a GREEN SPHERE ... and then visualize a RED SPHERE inside the GREEN SPHERE ...

6 Bring your attention back to the SPHERE of GREEN-VIOLET-SILVER energy in your Heart Center ...

and on the next EXHALE visualize, or imagine, sending a cord of GREEN-VIOLET SILVER energy all the way down your central vertical column and into the GREEN SPHERE in the 1st Cell at the tip of the tail bone.

7 AS you INHALE pull PALE SILVER LIGHT from your personal Maharic Shield, the platform-disc 12" below your feet, up into the tail bone and into the GREEN SPHERE.

8 As you do this, visualize, the GREEN SPHERE 'pop' into a STARBURST of PALE GREEN-SILVER LIGHT.

Continued Next Page

This has opened the kundalini seal – your WHITE-SILVER & RED Spheres REMAIN as you left them earlier ....Now, we will work to open the **RED STAR CRYSTAL SEAL....**

---

9 Return your awareness to the Doradic Sphere at the Heart Center....INHALE, and then on the EXHALE push a cord of GREEN-VIOLET-SILVER Doradic Current down to the 1st Cell at the tip of the tail bone, ...

10 INHALE and draw PALE SILVER energy up from the personal Maharic Shield, the platform-disc positioned 12" below your feet, up into the tail bone, and into the RED SPHERE ... inside .... the WHITE-SILVER SPHERE.

11 AS YOU DRAW the PALE SILVER energy into the RED SPHERE visualize the RED SPHERE 'pop' into a STARBURST of PALE PINK-SILVER LIGHT...... and ...... for the next few breaths EXHALE the PALE PINK-SILVER LIGHT into the 12th Chakra, positioned 6" below your feet...

You have now opened the Kee Ra ShA(Y) Seal ... so that the Sacred Flames of the Kee Ra ShA can be activated ......

ACTIVATING THE KEE RA SHA

The Kee Ra ShA appears as 3 little flames which exist in the 12th Chakra positioned 6" below the feet. The 3 flames are PALE BLUE, PALE YELLOW & PALE VIOLET.

1. Begin by focussing deeply on the 12th Chakra, 6" below your feet, and the three little flames held there ... one is PALE BLUE, one PALE YELLOW & the third PALE VIOLET.... We'll take a few moments to do this......

If you have difficulty visualizing, just imagine that they are there.

2 On the INHALE, draw the PALE VIOLET FLAME up from Chakra 12 ... into the central vertical column .. and all the way up ... and out through the top of the head, to the 10th Chakra ... 6" above your Crown chakra......

3 Next, return your attention to the 12th Chakra, 6" below your feet ... allow yourself time to connect with the little PALE YELLOW-GOLD FLAME. Breathe easy.

4 On the next INHALE draw the PALE YELLOW-GOLD FLAME up from Chakra 12 ... into the central vertical column .. and all the way up ... and out through the top of the head, to the 10th Chakra ... 6" above your Crown chakra......

Again, return your attention to the 12th Chakra, 6" below your feet ... and connect with the PALE BLUE FLAME... this FLAME is the INNER ARC of the COVENANT ... Take your time; breathe easy.

5 On the next INHALE draw the PALE BLUE FLAME up from Chakra 12 ... into the central vertical column .. and all the way up ... and out through the top of the head, to the 10th Chakra ... 6" above your Crown chakra......

5D Continued Next Page

© Rev. Ashayana Deane 2000

6 Now, visualize or imagine the 3 little FLAMES together within the 10th Chakra 6" above your Crown Chakra .... in a moment we will create the Kee Ra ShA SPHERE... Breathe easy.

7 On the next INHALE draw PALE SILVER energy from your personal Maharic Shield 12" below your feet, all the way up your central vertical column, up into the 10th Chakra......

EXHALE gently and INHALE once more, drawing more PALE SILVER energy up into the 10th Chakra ... visualize or imagine the 3 little flames enclosed within a SPHERE of PALE SILVER LIGHT...

8 Breathe easy. On the next EXHALE move the Kee Ra ShA SPHERE all the way down through the central vertical column, out through the 12th Chakra 6" below your feet, and all the way down to the 13th Chakra at EARTH'S CORE.

As the Kee Ra ShA enters EARTH'S CORE listen for sound tones w/ inner ear...

These TONES represent the AMENTI "PASS KEY" ... which opens your personal GRU-AL POINT (#1) - the 1st Cell in the Tail Bone .... releasing the KEE RA SHA FLAMES into your AURIC FIELD.

9 Breathe easy... Watch as the Kee Ra ShA FLAMES rise up your Maharic Pillar and emerge through your 1st Cell ... passing like radiating waves of PALE BLUE, PALE YELLOW & PALE VIOLET ......

10 As the waves of color move through you, bring your attention to your first 4 Chakras ... the base, the sacral, the solar plexus and the heart ... and visualize or imagine each Chakra emanate a SILVER MAHARIC SHEATH or OVAL POD, completely surrounding each one ...

This completes the EMERALD AWAKENING, the natural process by which we use Unified Field Physics to trigger our CHRISTED Selves. Before we finish we must first establish the link to the Amethyst Awakening, the process of activation and sheathing of the higher Chakras.

To complete the session we will now work with 8th & 11th Dimensional Energies to open DORADIC PHASE 2 & 3 CURRENT and activate TEURIC PHASE 2 CURRENT.......THIS IS CALLED **ACTIVATING THE RA CENTRE in the THYROID – and ANCHORING THE RAINBOW BRIDGE**

1 Begin by moving your attention to the area of your 3RD EYE, and focus just inside your fore-head .. a little above and behind your eye brows...... Breathe easy.

2 Now, visualize, or imagine, a DARK SILVER DOT in the position you are focussed on ... in the area of your 3RD EYE... Take a little time.

3 INHALE and move the DARK SILVER DOT into the Pineal Gland in the center of your head and then straight up through the CROWN CHAKRA until the DARK SILVER DOT sits on the top of your skull.

Continued Next Page

© Rev. Ashayana Deane, 2000

4 On the next INHALE, pull PALE SILVER ENERGY up from your personal Maharic Shield 12" below your feet, all the way up the central vertical current and into the DARK SILVER DOT on the top of your head...

As the PALE SILVER ENERGY reaches the DARK SILVER DOT, the SILVER DOT 'pops' into a SILVER STARBURST ...

5 On the next EXHALE push the SILVER STARBURST down the central vertical channel, all the way to the tip of the tail bone ... EXPANDING the SILVER STARBURST FULLY into the 1st Cell...

6 Breathe easily and watch, as GOLD-SILVER-CURRENT emerges from the tailbone .. and begins to flow up through each Chakra ..

from the 1st Chakra at the base of the spine ... up through the Sacral ... the Solar Plexus ... all the way up to the Crown Chakra ... we'll take a little time for this ....

From the Crown Chakra the GOLD-SILVER-CURRENT continues upward out of the top of your head ... up through the 10th Chakra 6" above your head ... the 11th Chakra 18" above your head ... and finally the 14th Chakra 36" above your head.

7 Once the GOLD-SILVER-CURRENT reaches the 14th Chakra it returns to the 1st Chakra the flow of the current creates a GOLD-SILVER SHEATH throughout your central vertical channel.

8 Bring your attention to the 2nd Chakra now .. were going to create a GOLD -SILVER- BALL there, in the SEED SEAL at the center of the Chakra ... just imagine the chakra as a bow tie shape, the SEED SEAL is at the narrow point in the middle. The GOLD-SILVER BALL is called the TEURIC SPHERE ...

INHALE and draw GOLD -SILVER-CURRENT from the GOLD -SILVER-SHEATH into the center of the Chakra ... as you EXHALE create and EXPAND the BALL to surround the bow tie ...

9 Now ... INHALE and move the GOLD -SILVER-BALL up to the 1ST HARA CENTRE ... 2" BELOW the NAVEL (the 9th SILVER STAR SIGNET SEAL or TAN TIEN) ... EXHALE GENTLY allowing the BALL to EXPAND

10 On the NEXT INHALE move the GOLD-SILVER-BALL up to the SOLAR PLEXUS and EXHALE GENTLY

11 INHALE and move the GOLD-SILVER BALL up to the HEART CENTRE ...
EXHALE GENTLY ... ... INHALE ... drawing more GOLD-SILVER CURRENT into the BALL and THEN EXHALE FORCEFULLY pushing cords of GOLD SILVER CURRENT down each arm and into the PALMS ... bring your fingertips together and feel the fuzzy buzz of TEURIC PHASE 3 CURRENT

You have now activated TEURIC PHASE 3 CURRENT in your palms – and – you are now able to run CURRENT from D4 thru' D8 to transmit to the THYROID GLAND ... to activate and anchor the Rainbow Bridge in others.....

End of Field Technique-5 *The K-RRT Activation Sequence* (5A-B-C-D-E)

You are now ready for the **Amethyst Awakening *Masters Kundalini/DNA Template Activation-2*** (Field Technique-6 *Initiating the Amethyst Awakening*) or to participate as an active member of ***a Khundaray Rainbow Roundtable*** (K-RRT) **Signet Council** (Field Technique-4 A-B-C). **Wait at least 4 Hours** before K-RRT participation, **and no longer than 24 hours** to maintain the **appropriate level of critical mass Khundaray frequency in the body** that is need for an effective K-RRT.

© Rev. Ashayana Deane, 2000

INITIATING THE AMETHYST AWAKENING

1 **Preparation:** Complete **Maharic Quick Seal** Technique 6 and **Activation Sequence**

2 This technique works to bring in a burst of the Khundaray, Eckatic Blue frequency ...that sets the Amethyst Awakening in motion.

3 Place your attention at the center of your brain, inside your pineal gland. Imagine that you see there a Pale Violet Spark of Light (this is the bottom of the Seal of your 7th Crown Chakra). Imagine that you are going to move this Pale Violet Spark of Light straight up into your 10th Chakra 6" above your head ... imagine that the 10th Chakra is a disc of Light, Blue/ Black – Sapphire.

4 Now see the disc of Light, that was Blue/ Black take on a Violet hue. Try to feel it there, a small horizontal disc ... On the INHALE you will move this disc of Violet-Sapphire Light up to the 24 point Silver Star at the 14th Chakra ...... and, on the EXHALE, move it down through the 10th Chakra, down through the Crown, down the Central vertical Current, down through your Personal Maharic Shield all the way down to the 13th Chakra at the Earth's Core ...

5 Notice, as you do this, a whole series of Violet-Blue Sparks flying out from the Planetary Maharic Shield, as if they are alive, as if electrical communication is taking place. Notice as this happens that the energy flowing up your Maharic Pillar from the Earth Core reflects this change as Violet-Blue energy flows slowly up toward you ... as this energy reaches your personal Maharic Shield, notice that it does the same thing as the Planetary Shields – throws out a shower of Violet-Blue sparks ... all communicating with each other about the mathematical instructions about what to do with your body, your crystal seals, your chakras ...

6 Now, move your awareness to the Pale Silver skin all round your body/ etheric body .. and be aware too of the Main Vertical Current that you brought the Pale Silver cord through ...and at the same time, on each coming INHALE, bring up the Pale Silver-Violet-Sapphire energy, simultaneously, from your Maharic Shield to infuse the Pale Silver skin all round you.

Take a few breaths to do this and watch as the 3 color flicker through you – when they reach the Heart Chakra feel them expand out ... and allow them to travel on up, all the way to the 14th Chakra ... expanding throughout every Chakra as the Pale Silver-Violet-Sapphire energy flows up, through and all around you ... See each Chakra respond with a spiraling burst of Silver-Violet-Sapphire, backward and forward through both sides of the Chakras until they all shift .. try to feel the process as it awakens all the cell's of your body.

The brain cells and brain patterns begin to respond, activating the Fire Letters, setting the grid of frequencies necessary to activate the Fire Codes 7, 10 and 12 in the DNA Template.

Your own **D-12 Christos Avatar identity level** will **slowly** release the ***Khundaray Activation Frequencies*** through which the Amethyst Awakening progressively unfolds in the DNA Template and body. This process can be **naturally and safely accelerated** by receiving a **Level-3 MC Regent Ordination transmission** from a qualified **Indigo Child/Eieyani** ***Level-4 through Level 6 MC Ordinate*** **Consummate, Elder Consummate** or **Eckar** that carries the necessary **Amethyst Awakening Khundaray Activating Frequencies** in their Eieyani DNA Templates.

The Merkaba Trinity Key Induction

The ***Merkaba Trinity Key Induction Technique*** begins the process of establishing a **conscious directional relationship** with various levels of your 15-Dimensional Anatomy (or that of your Client). The **Trinity Key Symbols** are **mathematical-geometrical programs** that carry the specific organic structure of **each dimensional frequency band** as it exists within the **D-12 Universal Divine Blueprint**. Imbuing the proper **Symbol Key** into the appropriate area of the personal anatomy can rapidly expedite **restoration the personal Divine Blueprint** within the **DNA Template, Chakra System** and embodied **Kathara Grid**. Use of the Trinity Master Key Symbols that correspond precisely to the natural order of each dimensional frequency band, can **intensify** all applications of **holistic and allopathic healing**, as well as progressively and systematically assist in the process of accelerating **DNA Template** healing and **activation.** Every **condition** within the body-mind-spirit system has a specific **Energy Signature** unique to its own construction. Every **dis-ease** condition within any aspect of the personal experience (physical, emotional, mental, spiritual, financial, social, etc.) is created and **held in manifestation** by a specific set of mathematical – geometrical energy relationships that are **stored** within the embodied **Kathara Grid** and **DNA Template** in the form of **Symbol Keys**. The mathematical-geometrical energy relationships that serve as the **blueprints** upon which experiential manifestation will emerge can be perceived by the conscious identity in its Symbol translation. When using the **Trinity Master Key Symbols**, you are **resetting and activating very specific portions** of the personal **Divine Blueprint** (D-12 Pre-matter Maharic Shield) into a naturally harmonious relationship with Universal Order (the organic structure of dimensionalized frequency bands). The Trinity Master Keys provided in this Technique trigger activation of very specific **Fire Letters** (Keylons) within the DNA Template, allowing the DNA Template to receive rectly the **new electrical instructions** corresponding the personal Divine Blueprint. This technique can be used with Kathara Healing Techniques from Kathara Level-1, with any other healing modality or singly after activating the D-12 **Maharata Current** via **Techniue-6, *The Maharic Quick Seal***.

It is **necessary** to initiate a **D-12 Maharaic Seal** in order to run **any Symbol Key program** into the DNA Template and Level-1 Kathara Grid; it is the strength of the D-12 Maharata Current that allows access to the deepest levels of the Personal Shields and Divine Blueprint. The **Trinity Master Key Symbols** are a **set of 3 Symbol Keys** that are to be used **in sequence.** The first two Trinity Keys, the **Initiating Keys**, are used in **every application** to open the embodied Level-1 Kathara Grid to receive new instructions. The third Trinity Master Key Symbol, the **Transmitting Key,** corresponds directly to the Divine Blueprint of **one dimensional frequency band** and its corresponding **Kathara Center, DNA Strand Template, Chakra** and related body regions, **Axia-tonal Line, Auric Field Level** and **station of consciousness**. In a **Trinity Key Induction Session**, one can choose to work on only **one dimension** (Chakra-DNA Strand Template-Kathra Center-Axia-tonal Line-Auric Level) via its Transmitting Key, or each of the **12 Transmitting Key** of the Universal Kathara can be used, **in sequence**, following induction of the two Initiating Keys. The process of using the Trinity Keys is simple; the **Chakra** that corresponds to the Transmitting Key/ dimension/DNA Strand with which you desire to work is used as the **vortex** through which the Trinity Keys are **INDUCED** into the corresponding **Kathara Center**. From the Kathara Center, the Trinity Keys electrically pass on to the **DNA Template** and from there into the **Axia-tonal Line system** and corresponding areas of the body-mind-spirit system. Use of the Trinity Key Induction Technique progressively assists in activating the personal **Merkaba Vehicle** while **amplifying and expediting restoration** of the personal Divine Blueprint.

Technique # 9: Merkaba Trinity Key Induction

1. Activate the Maharata Current via **Technique-6: *The Maharic Quick Seal*** and decide which Transmitting Key(s) you will use. **Trace** the **3 Trinity Symbol Keys** onto **3 small squares of paper.**

2. First, place each of the papers with the two Initiating Keys **face up on the floor**, then **stand** (barefoot or socks only) with **one foot on top of each of the papers; Initiating Key #1** under the **LEFT foot, Key #2** under the **RIGHT foot.** (Note: To run Trinity Key Induction for a Client, have them recline on their back and place the papers with Initiating Keys between your palms and the soles of Client's feet. As introduced in Kathara Level-1, use your **palm Chakras** to **transmit** the Symbols and the Pale-Silver Maharata Current, using the same action as described below, into the **Client's foot Chakras**; Key-1LEFT foot-***right hand transmission***, Key-2 RIGHT foot-***left hand transmission***.) In Self-induction, the papers with Symbols should be positioned beneath the **balls of your feet**, in the area of the main **Chakra** on the sole of each foot. Simply begin to **breathe deeply** and slowly for a few minutes. With each **INHALE** draw a **thick stream** of **Pale-Silver Maharata Current** from Earth's core through the **Central Vertical Current** of the body and out the top of the head to the **14th Chakra** 36-inches above the head. With each **EXHALE,** imagine a grapefruit sized Pale-Silver **Maharic Ball** rapidly moving down your Central Vertical Current. When the Maharic Ball reaches your 2nd-Chakra near the navel, imagine the Ball splitting into **Two Balls**, each continuing to **travel down each leg and out your feet Chakras**, through the **Symbol Keys**. Once the Two Balls of Maharic Current have passed through the Symbol Keys, imagine that they **re-merge** into the **one grapefruit sized Pale-Silver Ball** that continues to move downward into the **13th Chakra** at Earth's core. **Continue** this process until you sense a **momentary slight warmth, tingling or subtle itching** at the **balls of your feet.** These mild, **momentary sensations,** called the **CUE RESPONSE**, will grow stronger the more frequently you use the feet Chakras for Symbol Key induction. Once you have noticed the **Cue Response**, you have **successfully induced** the two **Initiating Trinity Keys** and it is now time to utilize the remaining **Transmitting Key.**

3. If you have chosen to use the Transmitting Key for D-1-Strand-1-Axiatonal Line-1-**Chakra-1**, continue **standing** for the final Symbol Key Induction, placing the paper with the Transmitting Key face up on the floor between **slightly parted feet; do not stand on Symbol Key paper** for **Base Chakra Induction.** For all other Chakras, **recline on your back** and place the paper with the Transmitting Key **directly on top of the Chakra** to which it corresponds, **on the front of the body**. Use each **INHALE** breath to drawn Maharata Current up from Earth's Core, into the Central Vertical Current **only as far up as the selected Chakra**. On each **EXHALE**, push Pale-Silver Maharata Current **out from the Central Verticle Current** of the body and **into the Chakra**, imagining that the Chakra is progressively turning into a **radiant vortex of Pale-silver light** that **spins faster and faster**. When you begin to feel a bit of warmth, tingling, energy running or itching of the skin **at the Chakra region** (the **Cue Response**), use a final **INHALE breath** to draw the image of the Symbol **down through the Chakra into the body**, and into the **tiny Crystal Seal** that links the Chakra to the Central Vertical Current. Take a **FINAL exaggerated EXHALE**, and imagine that you have **pushed the energy of the Transmitting Symbol Key** out of the Chakra into the **corresponding level of the Auric Field.** Once imbued into the **Auric Field** the single image of the Symbol then **replicates automatically and rapidly**, creating **millions of mini-replicas of the Symbol** within your bio-energetic field. When the **Cue Response** sensation at the Chakra has **completely ceased**, you have **successfully induced** the entire **mathematiucal program** of the Transmitting Symbol into the corresponding Kathara Center, DNA Strand Template, Axia-tohal Line, body region and level of consciousness.

Technique-9: Merkaba Trinity Key Induction

Trinity Master Key Selection Chart

© 2001 Ashayana Deane

12 Transmitting Keys: ***Correspond to each Dimension-Chakra-Kathara Center-Axia-tonal Line etc.***

Number in circle represents the **number of Chakra** into which corresponding Transmitting Key Symbol will be induced. **Note:** The **Smaller Symbols** placed **outside of SOME circles** are **Amplifier Key Symbols** corresponding to the Chakra by which they are positioned. The Amplifier Keys can be used **in place of** the Transmitting Key within the circle, or the two Keys can be drawn on one paper for **simultaneous induction**. It is **recommended** to begin with the Transmitting Key inside the circle, then after a few induction exercise, switch to the Amplifier Key, and after several exercises with the Amplifier Key, use both Transmitting and Amplifier Keys simultaneously.

Two Initiating Keys

1)

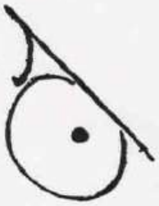

2

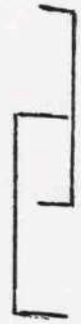

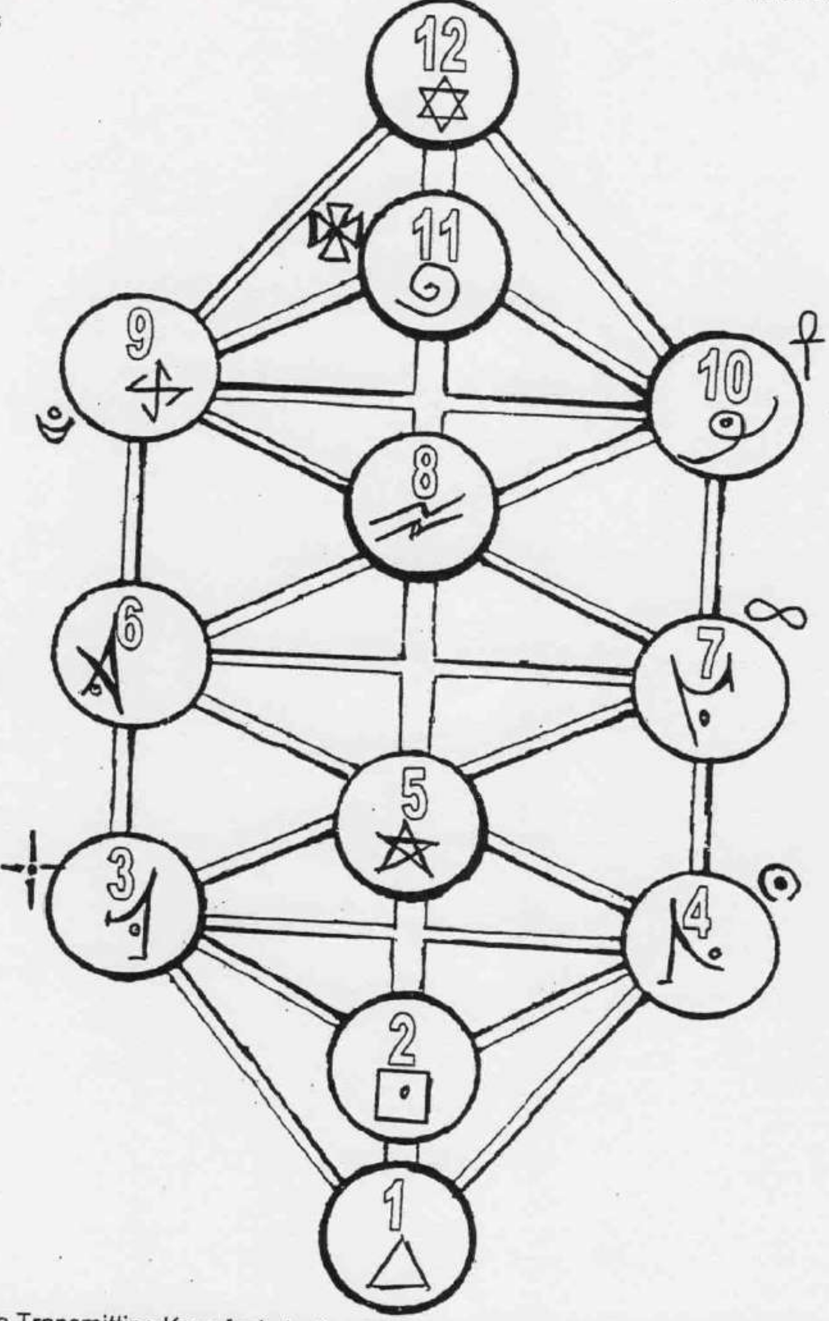

Select one or more Transmitting Keys for induction following induction of the two Initiating Keys above.

In **Kathara Level-1** we have learned of the innate structures of multidimensional reality and how these structures are intimately intertwined with the ***15-Dimensional Human Anatomy*** and levels of personal identity through the personal ***Maharic Shield*** and ***DNA Template***. In rediscovering these long forgotten aspects of the Human condition and our inherently implied relationship to Universal Order, we've become empowered to draw upon the natural **Triadic Phase Currents** indigenous to our personal biology and universal structure, in order to take an active hand in personal *Conscious Bio-spiritual Evolution* and healing, and healing facilitation for others. ***Techniques 1-5*** (#1-*Awakening the Mentor*, #2-*Maharic Seal and Liquid Light Cleanse*, #3-*Opening the Healing Channels Doradic Phase-1 Current*, #4-*Multi-vector Holographic Recoding Kathara Scan* and #5- *The Maharic Infusion* client session) contained within *Chapters 1-4 of the Kathara Level-1 Maharic Recoding Process* enable us to begin conscious interaction with the "invisible" processes of **Body-Mind-Spirit Evolution** that are inherent to Human design. In our introduction to **Kathara Level-2** we have begun to explore the greater nature of ***Universal Life Force Currents***, the "mysterious" sciences of ***Merkaba Mechanics*** and their indelible connection to our personal evolution and **Bio-spiritual Actualization Processes**. ***Techniques 6-7-8-9-12*** (#6-*Maharic Quick Seal*, #7-*Ps13 Tribal Shield Activation*, #8-*Merkaba Salutations*, #9-*Merkaba Trinity Key Induction* and #12 *Kee-Ra-Sha Activation Sequence*) so far explored within this *Chapter 5* of the *Kathara Level-2 Introduction* enable us to advance our conscious interaction with the processes of personal Body-Mind-Spirit Evolution and Bio-Spiritual Actualization. As we increase our personal abilities to understand, embody, hold and direct the inherent powers of the Universal Life Force Currents, and to access the levels of personal spiritual identity and wisdom to which they correspond, we progressively increase our ability to consciously facilitate healing in ourselves and for others we may choose to serve. An in-depth study of Universal Life Force Currents and Masters Merkaba Mechanics involves cultivating a necessary familiarity with a broad spectrum of interrelated subjects that take us into the heart of the Science of Vibrational Mechancis and Holographic Creation Physics, as well as into the very Soul of our spiritual relationship to consciousness, ourselves, each other, our world and God. Mastery of Merkaba Mechanics is an attribute inherent to achieving the state of consciousness often called "Ascended Mastery", which requires much in-depth study into the "mysteries" of creation. However, a functional mastery of Merkaba and the related constructs of the personal "Inner Templar" system upon which the personal body and identity manifest can be achieved through gaining familiarity with and actively employing core mechanics of the personal Merkabic System.

As our *Kathara Level-2 Introduction* began, our exploration of active techniques of *Tribal Shield Activation, Merkaba Salutations* and *Trinity Keys* set us on the path of becoming functionally able to consciously direct elementary operations of the personal Merkaba Vehicle and Life Force Currents, as a means of expediting our progress of Bio-Spiritual Evolution through DNA Template/Merkaba Field Activation. As our *Kathara Level-2 Introduction* continues, we will begin to explore the seemingly "***magical world of Merkaba***", rediscovering the very tangible importance of the **Merkaba Field** in relation to the inherent dynamics of DNA, health, healing and spiritual actualization. To achieve genuine integrity in the facilitation of healing for others, one must simultaneously strive to heal and actualize the self. The greater the level of health and spiritual actualization possessed by a healing facilitator, the greater will be the facilitation of healing for the client to whom the healing facilitator is in service. As Kathara Level-2 Introduction draws to a close we will begin an introductory exploration of the "Ancient Secrets" of the **Veca Codes,** so that we may engage practical application of two additional techniques of Photo-Radionic and Photo-Sonic **Merkabic Healing. Techinque-#10** ***Restoring the Trion Field Veca Code Activation*** and **Technique-#11-*Restoring the Meajhe Field Eckasha Activation*** will provide us with powerful tools of the Kathara Healing trade that will ***substantially increase the quality, amount and power of interdimensional frequency we are capable of receiving, holding and transmitting***. Through these simple but effective Merkabic Healing tools we can expedite our processes of personal healing and spiritual actualization, thereby enabling us to become more effective and powerful healing facilitators in our client session service.

The "Magic of Merkaba"

Spiritual-science teachings of **Merkaba Mechanics** emerge from the most ancient of times; record of Merkaba teachings is most often found in obscure traditional "Occult" texts or within New Age "Ascension" teachings. Genuine Merkaba Sciences hold the keys to understanding and mastering the processes of both spiritual and physical evolution, on a personal, planetary, galactic and universal level. Merkaba Mechanics are not the product of "science fiction", nor are they "magical hocus-pocus" of "occult mysticism". Merkaba Mechanics are not "magic" at all, but rather an intrinsic element of universal Primal Order. **Merkaba Fields**, sets of Star-tetrahedron shaped counter-rotating electro-magnetic energy fields within and surrounding manifest forms, are simply an inherent element of the15-Dimensional Anatomy of form manifestation and the Primal Order structure of multidimensional Creation Physics. In the "*Transduction Sequence*" described in *Kathara Level-1*, the sequence of energy transmutation by which conscious energy manifests into tangible biological form, the *Merkaba Field* represents the structure and stage of energy dynamics by which *Primal Life Force Currents* pass from the scalar-wave-template of the embodied *Kathara Grid levels* into the *DNA Template*. In terms of non-biological form manifestation, all forms possess an inherent structure of interwoven Merkaba Fields through which Primal Current energy is transmuted from the Kathara Grid Core scalar-template level into the outer levels of the Kathara Grid and Axiom Line template upon which the specialized structure of the form builds up, through particle/anti-particle accretion, into physical manifestation. Merkaba Mechanics seem "magical" only when the inherent "**Laws of Nature**" that govern the dynamics of **Multidimensional Universal Unified Field Physics** are not understood.

Through understanding of genuine *Universal Unified Field Science, Merkaba Fields*, and their inherent dynamics as utilized in the study of *Merkaba Mechanics*, are simply recognized as natural elements of the processes inherent to Creation Physics and the naturally occurring, organic relationships between Energy Consciousness, and the Holographic Manifestation of form.

As Merkaba Fields are an intrinsic, significant and natural part of the Body-Mind-Spirit anatomy, the study and application of Merkaba Mechanics is an essential component in any Masters Holistic Healing or Spiritual Actualization program. Merkaba Mechanics are "*Scientific*" in that they represent a core element within the innate processes of creation and manifestation that are intrinsic to the sciences of Creation Physics. Merkaba Mechanics are also "*Spiritual*" in that it is through the innate structures and dynamics of Creation Physics, of which Merkaba Mechanics are a natural part, that consciousness enters the manifest arena in the form of sentient, individuated, multidimensional identity. The "body, mind and spirit" of every manifest being is formed through and governed by the natural laws of Creation Physics and Primal Order. The Laws of Creation Physics and Primal Order are designed and maintained by and within a non-manifest, intelligent, central Consciousness Force and Source of Creation that is commonly referred to as "**God**". God expresses It-Self in the form of *Conscious-energy Constructs* that carry the living "Spirit and Consciousness of God" in the form of **Primal Life Force Currents**, which are the foundations for macrocosmic and microcosmic creation and manifestation. Merkaba Fields and the mechanics by which they function, are an indelible part of the physics of Creation and Consciousness, and are thus intimately involved with the Bio-Spiritual processes inherent to **Spiritual Actualization**.

Merkaba Fields, DNA, Healing and Spiritual Actualization

Particles, People, Planets, Galaxies and Universes all have **sets of interwoven Merkaba Fields** that serve as the "**core Primal Life Force Current circulation system**" between the Cosmic, Universal, Galactic, Planetary and Personal "Divine Blueprint" Kathara Grid scalar-templates and the manifestation of those blueprints into the tangible structures of matter. Merkaba Fields are the "**carriers of consciousness**" and the "**engines of energy**" by which physically manifest forms are created, maintained and evolved through the Holographic Unified Field of manifest space-time-matter reality. In relation to the processes of personal bio-spiritual healing and spiritual actualization, the personal Merkaba Field is an intrinsic aspect of 15-Dimensional Anatomy that directly governs the function of the physical, emotional, mental and spiritual body systems. **The multidimensional structure of the personal Merkaba Field is the system of electro-magnetic interface by which the Primal Life Force Currents and Consciousness of God-Source and Its sentient individuations pass from the Kathara Grid Core Template into the scalar-wave DNA Template that holds the blueprint for the physical-chemical biological DNA.** This reality of **Merkaba Field-DNA Connection** implies that the personal Merkaba Field is directly related to the function or dysfunction of the **biological genetic code.** As the Merkaba Field "carries the energy of consciousness" into the DNA and biological vessel, the *Merkaba Field-DNA Connection* also implies that the potentialities of consciousness and embodiment of spiritual identity, while a being is ensconced within the Time Matrix, are dependant upon the condition and function of the personal Merkaba Field, DNA Template and Kathara Grid.

> *Merkaba Fields are the "**carriers of consciousness**" and the "**engines of energy**" by which physically manifest forms are created, maintained and evolved through the Holographic Unified Field of manifest space-time-matter reality.*

Universal Merkaba Field Structure

Basic Merkaba Field structure is relatively simple. Viewed Macrocosmically, in a 15-Dimensional Time Matrix there are **15 sets** of counter-rotating electromagnetic **Merkaba Spirals** that form **15 Universal Dimensional Merkaba Fields.** A "**Merkaba Field**" is composed of a set of **2 Merkaba Spirals**, one electrical-transmitting "cone", the other magnetic-receiving "inverted cone"; the 2 "cones" of the Merkaba Spirals, which operate as a **White Hole/Black Hole Pair**, intersect each other, creating a system of energy vortices that form an **Elliptical Sphere** of energy within the center of the intersected Merkaba Spirals. The Sphere of energy at the center of a Merkaba Field is called the **FORM CONSTANT**; it is a "**spherical energy domain**" within which the individuation of consciousness and holographic manifestation/projection of form can take place, akin to an "energy egg" within which manifest reality occurs. Within this Sphere of energy at the center of a Merkaba Field, the template-blueprints of a form are held "constant", allowing conscious Partiki units, particles, anti-particles and ante-particles to accrete in specific ratios to "fill out the form in matter". Within the Form Constant sphere, consciousness is able to perceive the **holographic illusion** of externalized space-time-matter manifestation, as the scalar-template of the form is held fixed within the sphere. The "hologram" appears to remain "constant" within the Form Constant sphere while consciousness is focused within the frequency bands of which the Merkaba Field is composed.

On the macrocosmic universal scale, the Form Constant of each Universal Dimensional Merkaba Field creates a single, elliptical-sphere shaped electromagnetic domain of frequency that represents a single dimension of manifestation. Following the innate mathematical-geometrical structure of Primal Order, the 15 Universal Dimensional Merkaba Fields group into **5 sets** of 3-Dimensional Merkaba Fields called **Universal Harmonic Merkaba Fields**. The 3 Form Constant Spheres from each set of 3 Universal Dimensional Merkaba Fields forms a **Universal HOVA BODY**, or a "Harmonic Universe" electromagnetic domain. Following the mathematical Primal Order program of a Time Matrix as held within the Universal Kathara Grid template and transmitted through the rotation ratios of the Harmonic and Dimensional Merkaba Fields, each of the 3 dimensional levels of a Universal Hova Body holds the Partiki units, consciousness and matter particulates of its dimensional field into a specific angle and speed of rotation. The 3-Dimensional Harmonic Merkaba Field controls the relationship of particle rotation angle (angular rotation of particle spin or "**ARPS**") and rotation speed between each of the 3 Dimensional Merkaba Fields, holding together a 3-dimensional electromagnetic domain, within which consciousness can experience a 3-dimensional space-time-matter hologram. In a single 15-Dimensional Time Matrix there are thus 15 singular Universal Dimensional Merkaba Fields governed by 5 larger 3-dimensional Universal Harmonic Merkaba Fields, all of which circulate consciousness in the form of Primal Life Force Currents into and out of the hologram of manifestation. The seeming solidity and durability of an externally manifest matter-field in space-time is created, maintained, perpetuated and completely dependent upon the intrinsic function of the Universal Merkaba Fields.

Within the larger Form Constant elliptical-spheres of the 5 Universal Hova Bodies of a 15-Dimensional Time Matrix, the holographic perceptual realities of space-time-matter take place. Every being or thing in manifest form enters its reality hologram by *taking on the mathematical program* of first the *Universal Kathara Grid* and *5 Universal Hova Bodies*, which set the "Life Stream" or "Stream of Consciousness" of Primal Creation

Currents in motion following the specific **Harmonic Merkaba Field spin ratios** characteristic to each Harmonic Universe System. Galaxies are formed in this way, as individualized expressions of the greater Harmonic Universe of which they are a part. **Universal Harmonic and Dimensional Merkaba Fields** govern the birth, evolution and transmutation of the **galaxies** within them. **Galactic Harmonic and Dimensional Merkaba Fields** govern the birth, evolution and transmutation of **star systems and planets** within them. **Planetary Harmonic and Dimensional Merkaba Fields** govern the birth, evolution and transmutation/death cycles of the **species and matter forms** existing upon the planet. The **Personal Merkaba Fields** of species and individuals direct the function of the personal **DNA Template, chemical DNA** and **biological genetic code**, thus governing the birth, evolution and transmutation processes of individual organisms. The mathematical-geometrical programs of the **Primal Order "Divine Blueprint"** that govern Merkaba Field formation and spin ratios are held within the Universal, Galactic, Planetary, Species and Personal Kathara Grid structures. In a singular 15-Dimensional Time Matrix, the **inter-linked Universal, Galactic and Planetary Harmonic and Dimensional Merkaba Fields** circulate Primal Life Force Currents and consciousness into and out of manifestation in space-time.

Universal Merkaba Fields

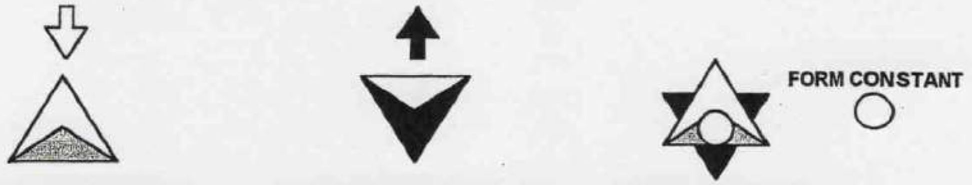
FORM CONSTANT

Electrical ***Merkaba Spiral***
Rotates Clockwise
"White Hole" Vortex

Magnetic ***Merkaba Spiral***
Rotates Counter-clockwise
"Black Hole" Vortex

One ***Merkaba Field*** =
a set of 2 *Merkaba Spirals*
Black-White Hole Pair

Universal Merkaba Field of One 15-Dimensional Time Matrix

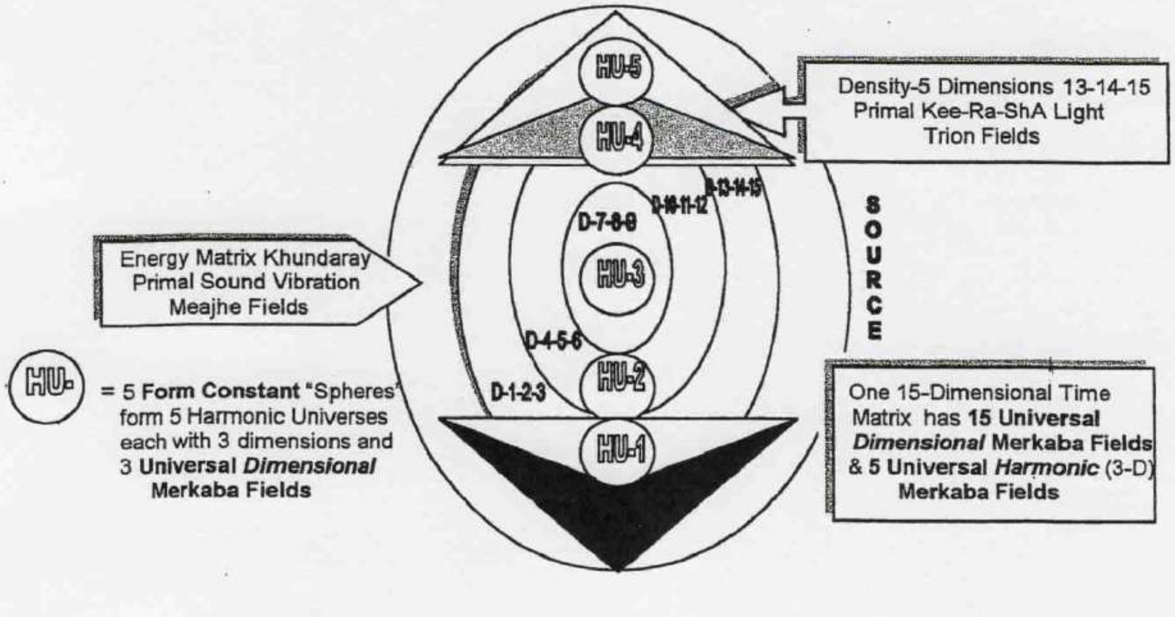
HU-5
HU-4
HU-3
HU-2
HU-1
D-13-14-15
D-10-11-12
D-7-8-9
D-4-5-6
D-1-2-3
S O U R C E
Density-5 Dimensions 13-14-15
Primal Kee-Ra-ShA Light
Trion Fields
Energy Matrix Khundaray
Primal Sound Vibration
Meajhe Fields
HU = 5 **Form Constant** "Spheres"
form 5 Harmonic Universes
each with 3 dimensions and
3 **Universal *Dimensional***
Merkaba Fields
One 15-Dimensional Time
Matrix has **15 Universal**
***Dimensional* Merkaba Fields**
& 5 Universal *Harmonic* (3-D)
Merkaba Fields

The Personal Merkaba, Trion-Meajhe Field, Radial Body and Radis.

Each individual being receives its life force, consciousness, biology and manifest hologram through the innate connection between the **Personal Merkaba Field** and the Merkaba Fields of the Planet, Galaxy, Universe and Time Matrix. When the **D-12 Pre-matter Divine "Christos" Blueprint** is expressing naturally in a planetary system, the organisms on the planet circulate Life Force Currents and consciousness through the physical and "spiritual" bodies via the Personal and Planetary Merkaba Fields. The **Planetary Merkaba Field** circulates Life Force Currents through the **Personal Kathara Grid** template (morphogenetic field) of the organism via the **Planetary Bio-Feed Interface System.** The organism's **Personal Merkaba Field** then picks up the energy from the Kathara Grid and continues to circulate the Life Force Currents from the Personal Kathara Grid and into the **Maharic Shield imprint** in the **DNA Template** that is held within the Level-1 *Kathra 12-Tree Grid,* Level-2 *Kathara Crystal Seals-Hara Grid* and Level-3 *Kathara Diodic Grid*. From the Maharic Shield imprint in the Kathara Grid, the Personal Merkaba Fields distribute the Life Force Currents and consciousness into the **DNA/RNA Template and Axiom Lines.** The Life Force Current continues circulation from the DNA/RNA Template and Axiom Lines into the **Personal Hova Bodies** and **Auric Field**, to the **Chakras** and **Meridian Lines**. Passing through the Meridian Lines, the Life Force Current, imbued with the specific mathematical manifestation program picked up from the Kathara Grid Divine Blueprint Template, then enters a state of final transmutation into 3-dimensional particles by passing through an electrostatic **Repulsion Zone** barrier that exists surrounding each 3-dimensional Hova Body.

The frequency barrier surrounding each 3-dimensional Hova Body forms as **thin, skin-like membrane** or **"Tissue Capsule"** surrounding each of the 5 Hova Bodies in the 15-Dimensional Anatomy. The **Hova Capsule** membrane is composed of Density-5 **Primal Light** ante-matter units, called **TRION UNITS**, that exist in the form of fixed standing-columnar-scalar-light waves made of **Keylons**, or 3-dimensional **Partiki Grids**. The Trion Units and Keylon Partiki Grids within them are anchored upon a fixed field template of **Primal Sound** units, called **MEAJHE UNITS**, which exist as fixed points of mathematically organized vibration. The outer layer of each Hova Capsule is called the **Trion Field**, as it is composed of **Trions**, or units of Density-5 ante-matter Primal Light. The inner portion of the Hova Capsule is called the **Meajhe Field** as it is composed of **Meajhons** (mE' yans) or fixed points of Primal Sound vibration from the **Energy Matrix** beyond the holographic manifestation of the Time Matrix Light Fields. The collective body of the 5 Hova Capsules is called the **RADIAL BODY** or **Trion-Meajhe Field**. All manifest forms possess a core Radial Body/ Trion-Meajhe Field **Hova Capsule** structure as part of the intrinsic 15-Dimensional anatomy.

As multidimensional Life Force Currents pass through the Meridian Lines and into each of the 5 layers of the Radial Body, the energy currents and the consciousness carried upon them project the mathematical program of the Divine Blueprint into the Trion-Meajhe Field of each of the 5 Hova Capsules. Within the Hova Capsule, the individualized raw Life Force Currents enter a system of 3-dimensional energy conduits or **rivulets** called **Radis**, which are the "turnstiles" between "manifest matter" and "conscious energy". The Radis draw in the Life Force Currents with their manifestation program through the Trion-Meajhe Field, then polarize and break up the Trion and Meajhon units into negatively and positively charged electrical sub-units called **Mions** and **Dions**. Partiki, Keylons, Meajhons, Trions, Mions and Dions belong to a category of primal building blocks known as **IONIC PARTICULATES,** which are the building blocks of consciousness, energy, light, sound and scalar-wave-fields that form the blueprints upon which all matter manifests. **Trions** are electrostatic particulates that have both a negative and positive electrical charge, which renders them neutral or static; they are the foundations of what becomes **Ante-matter** Primal Light particles of pre-visible gaseous light. Meajhons are static units of pre-sound standing waves, also possessing neutral electrical charge, that form the Primal Sound Fields of the Energy Matrix. Within the structures of the Radial Body Radis rivulates, Trions and Meajhons are **polarized** into negative charged **Mions** ("Base Magnetic" vibrating pre-sound particulates, foundations of negatively charged particles such as electrons, originating from the Universal Particle Particum field.) and positively charged **Dions** ("Base Electrical" oscillating pre-light particulates, foundations of positively charged particles such as protons, originating from the Universal Anti-particle Partika field). Together, Mions and Dions create the **electromagnetic Ionic Particulate base field** within the Merkabic Form Constant through which particles emerge via Ionic Particulate accretion into tangible matter units.

The Trion-Meajhe Field Radial Body of any manifest form represents the 5-Density "Veil" through which consciousness-as-Primal Life Force moves into and out of externalized 3-dimensional manifestation. **Merkaba Fields are the "energy engines and consciousness carriers" by which Life Force and consciousness are circulated between the internal Kathara Grid scalar template and the Trion-Meajhe Field Radial Body "veil" as they pass into and out of external expression.** The perpetual polarization, electromagnetic expression and de-polarization of the Life Force Currents within the Merkaba Field Form Constant electromagnetic domain is the process by which the holographic projection of matter is perpetually brought into and taken out of perceptually experiential being. The **Radial Body** Trion-Meajhe Field receives, via Merkaba Field circulation, the template or "design" for its particle manifestation from the "instructions" held in the Divine Blueprint Maharic Shield and DNA Template of the Kathara Grid. Manifestation instructions pass from the Kathara Grid to the DNA Template via the Merkaba Field, then continue into the energy circulation systems (Axiom Lines, Hova Bodies, Chakras, Meridians) that deliver the instructions to the Radial Body for particulate manifestation. In biological life forms, the Radial Body and Radis then translate the scalar-wave blueprint of the DNA Template and Kathara Grid into the polarized Ionic Particulate Mion-Dion sub-units, which accrete following the DNA Template instructions to form sub-atomic, atomic, molecular, chemical DNA, physical matter form- the vehicles of embodied consciousness. The **Merkaba Field** also receives its instructions for energy circulation from the Kathara Grid and DNA Template.

The Radial Body Tissue Capsule, Radis *& the Mechanics of Manifestation*

The **Radial Body** exists as a thin "***Tissue Capsule***" surrounding each of the **5 Hova Bodies**. Each of the **5 Radial Body Levels** are composed of an **Outer Layer** of electrostatic neutral-charge **Trion Primal Light Units** and an **Inner Layer** of **Meajhon Primal Sound Units**. Each of the 5 Levels of the Radial Body serve as the **polarizing filter or "lens"** through which **Multidimensional Primal Life Force Currents** flow, via **Merkaba Field** circulation, from the Kathara Grid and DNA Template, into manifest form.

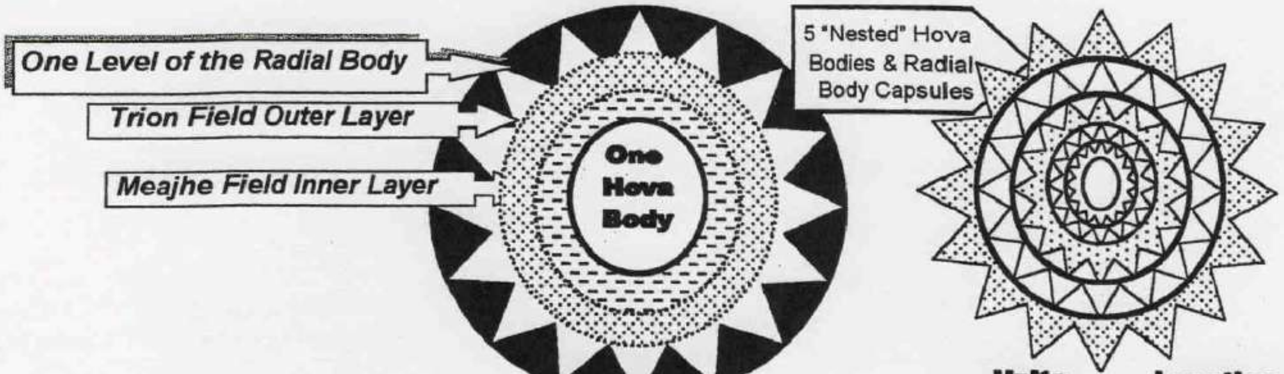
One Level of the Radial Body
Trion Field Outer Layer
Meajhe Field Inner Layer
One Hova Body
5 "Nested" Hova Bodies & Radial Body Capsules

Radial Body Transduction-Manifestation Sequence

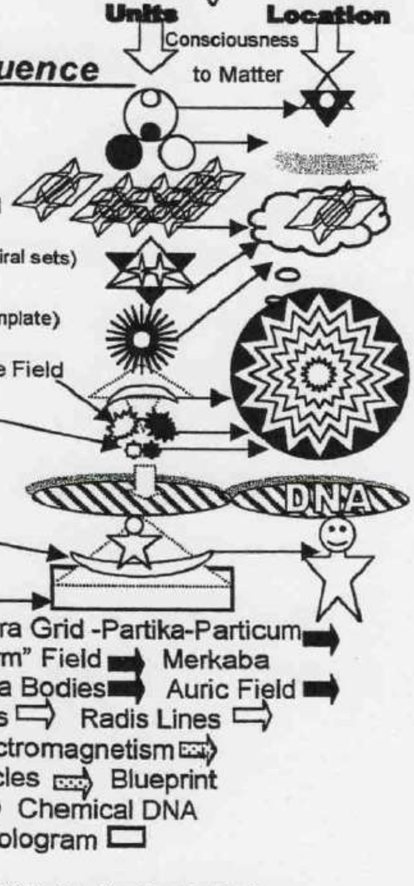

Primal Life Force Current *Partiki Units* Planetary Merkaba Field
(Units of Consciousness/Primal Substance)
Partika and Particum Units Kathara Template/Maharic Shield
(Polarized Primal Substance Units)
3-D Partiki Grid Keylon Units/Codes/Grid Morph. Keylon Thought-Form Field
(3-D Crystallized Pre-Light-Sound units, unit groups and grids)
Merkaba Field *Keylons* (3-D Crystallized Pre-Light-Sound Keylon Unit electromagnetic spiral sets)

DNA/RNA Template *Keylons* Kathara Grid (3-D Crystallized Pre-Light-Sound unit Template)
Axiom Lines-Hova Bodies-Auric Field-Chakras-Meridian Lines
Trion & Meajhon *Electrostatic Units* 5 Radial Body "Lens" Levels Trion-Meajhe Field
Dion (+) & Mion (-) Ionic Particulate *Electromagnetic Units* Radis Lines

Sub/Atomic Particles/Anti-Particles, Molecules...Chemical DNA
Physical Matter Body Hologram "Chemical Lens"

External Hologram Projection of Space-Time-Matter
Primal Life Force Currents ➡ Planetary Merkaba Field Partiki ➡ Kathara Grid -Partika-Particum ➡ Maharic Shield Divine Blueprint ➡ Keylon Morphogenetic "Thought-form" Field ➡ Merkaba Fields ➡ Kathara Grid DNA/RNA Template ➡ Axiom Lines ➡ Hova Bodies ➡ Auric Field ➡ Chakras ➡ Meridian Lines ➡ Trion-Meajhe Field 5 Radial Body Levels ⇨ Radis Lines ⇨ Mion & Dion Ionic Particulates ⇨ Strong and Weak Nuclear Force, Electromagnetism ⇨ Quarks, Muons, Measons, Sub-atomic & Atomic Particles and Anti-Particles ⇨ Blueprint translation into Nuclei, Molecules, Elements, Compounds, Chemicals ⇨ Chemical DNA Template and "Chemical Lens" of the manifest body form ⇨ Manifest Hologram ▭

Reverse Transduction De-Manifestation Sequence: matter- *chemical* DNA- *Sub-Atomic units*- Radis *Mion/Dion units*- Radial Body *Trion-Meajhon units*- Meridian Line-Chakra-Auric Field-Hova Body-Axion Line-DNA Template-Merkaba Field-Morphogenetic Thought-form Field *Keylon Grids & units*- Maharic Shield Template & Kathara Grid *Partika and Particum Units*- Planetary Merkaba Field *Partiki Units* of consciousness.

Radial Body Transduction-Manifestation Sequence

1. From the **Planetary Merkaba Field** (via the Planetary Shield, Axiom Lines, Hova Bodies Vortices, Meridian Lines and Radial Bodies), **rhythmic pulses of Conscious Primal Life Force Currents** enter the species and personal **Kathara Grid Core Template** as Primal Units of **Partiki**.

2. Within the **Kathara Grid Template** the **Partiki Units** pick up the **mathematical-geometrical program** from the **Maharic Shield** "Divine Blueprint" Template, **polarize** to form **Particum and Partika Units**, then **accrete** to form **3-dimensional** electro-magnetic **Partiki Grids** called **Keylons**. The Keylons group together mathematically forming complex templates of **Crystallized Keylon Codes** (or "***Keylon Thought-Crystals***"), which create a mathematically encoded **3-dimensional Morphogenetic Field ("Keylon Thought-form Field"** or Morphogenetic **"Keylon Crystal Grid"**) carrying the manifestation blueprint program of the Maharic Shield.

3. As the Primal Life Force Currents of the electromagnetic Keylon Crystal Grid interact with each other following the program of the D-12 Pre-matter **Universal Divine Blueprint**, they form **sets of spiraling electro-magnetic vortices** called **Merkaba Fields**, which carry the Maharic Shield-Kathara Grid Blueprint program into formation of the **DNA/RNA Template** Keylon Crystal Grid.

4. From the DNA/RNA Template held in the Kathara Grid, Primal Life Force Current carries the **Keylon Crystal Grid** of the blueprint into the **Axiom Lines, Hova Bodies, Auric Field, Chakras** and **Meridian Line** systems, finally carrying the Keylon Crystal Grid of the Divine Blueprint into the **5 Radial Body "Tissue Capsule" Levels.**

5. Within the **5 Levels of the Radial Body** the **incoming Keylon Crystal Grid** translates into **electrostatic Trion-Primal Pre-Light Units** (gaseous pre-light Ante-matter) and **Meajhon-Primal Pre-Sound Units** (units of vibrating pre-sound energy) that possess a **neutral electrical charge.** The **Encoded Primal Life Force Currents** (PLFCurrents+ Keylon Crystal Grid blueprint program) are converted into a **Trion-Meajhe Field blueprint replica;** a " Primal Light-Sound" translation of the original blueprint.

6. The **Encoded Primal Life Force Currents** then travel, **AS the Encoded Trion-Meajhe Field**, into the **Radis Lines** in each of the 5 Levels of the Radial Body. The Radis Lines first **replicate and store** the **mathematical blueprint encoding** of the **Trion-Meajhe Field,** then **polarize** the neutrally charged Trion-Meajhe units into **electromagnetic units** called **Mion** (-) and **Dion** (+) ***Ionic Particulates***. The **replicated Trion-Meajhe Field blueprint** becomes part of the "Primal ***Pre-sound*** vibration" ***Meajhe Field*** **Inner Layer** of the ***Radial Body*** **Trion-Meajhe Field**, serving as a living "***Memory Matrix"*** composed of **Meajhon Unit** "Primal Pre-sound vibration". The **polarized Mion/Dion Units** become the foundations upon which the **"Strong and Weak Nuclear Force"**, **Electromagnetism, quarks, Muons, Measons, Sub-atomic Particles** and **Anti-particles** etc. form. Particum-based **Mions** ("Base magentic) become **particles, negative electrical charge units** like **electrons** and **sound spectra**. Partika-based **Dions** become **anti-particles, positive electrical charge units** like **protons** and **light spectra.**

***Radial Body Transduction-Manifestation Sequence-* Continued** page 2

7. Blueprint Encoded Primal Life Force Currents in the form of a **Mion-Dion Field** then translate the original Kathara Grid Blueprint into **Quarks, Muons, Measons, Sub-atomic** and **Atomic particles** and **anti-particles**, which **accrete** according to the blueprint to form **molecules, elements, chemicals**, the **Chemical DNA Blueprint** and the **physically manifest body form**. As this occurs the **body image** "**flashes on**" into manifestation as a **Pulse of Life Force Consciousness** carrying the **Chemically Translated Holographic Image** of the body form template blueprint. This Pulse of Consciousness continues to expand from the Kathara Grid and DNA via the Transduction Sequence, through and beyond manifestation of the body image and into **projection of portions of the Dimensional Unified Field** that **holographically translate** into **individualized experience** within a perceived **Common Planetary Time Rhythm and Matter Field.**

NOTE: *An individual will holographically perceive and experience only the portions of the Dimensional Unified Field mathematical-geometrical program that are encoded within the personal Maharic Shield, species Tribal Shield and Planetary, Galactic and Universal Shield aspects of the personal Kathara Grid and DNA Template.*

A **global population** experiences a **common planetary reality field and time continuum** because the Kathara Grid Template and DNA of all beings entering a planetary field **adopt the common imprint** of the mathematical-geometrical program of the **Planetary Kathara Grid** blueprint in order to enter into the electromagnetic domain of the Planet's Form Constant Hova Body structure for experiential manifestation. Members of a **singular species** share a common **species Tribal Shield** template through which a **common Chemical Body Form** and perceptual orientation holographically manifest. The imprint for **individuality** in biological form, life experience variation and the application of **Free Will Choice** is held within the mathematical program of the **personal Maharic Shield** D-12 Pre-Matter **Divine "Christos" Blueprint**. Imbued within the mathematical program of the **Human species Tribal Shield**, and that of some other species, is the imprint for **probability** to exist. The Human species Tribal and Maharic Shields carry a program that allows for **new programs of probable creation** to be **added** to the personal and species **Kathara Grid**, allowing the embodied consciousness to retain **a degree of direct influence** over what sets of mathematical programs will be called into personal experiential Holographic Expression from the personal Maharic Shield Template. **Creative variation** on the original personal, species, planetary and galactic Divine Blueprint is possible as long as the new probabilities do not deviate so excessively from the Pre-set programs and **Original Intentions** of the Universal Divine Blueprint that the **structural integrity** and **Source-God Will Intentions** for our Time Matrix become **irreparably compromised. Free Will thus exists within a pre-set framework of overriding Divine Will**, and is permitted free reign in the manufacture of **co-creative experiential probability** up to the point at which the Original Divine Will of Source and the creative framework itself become unduly jeopardized.

8. The physical body form serves as a **Chemical Lens** through which ***portions*** of the **Dimensional Unified Field** are perceived by the embodied consciousness as a **3-dimensional Hologram,** as the Primal Life Force Currents continually carry **pulses of the individuated consciousness** through the ***Transduction Sequence*** from Kathara Grid, through the DNA, to the manifest Chemical Lens of the physical body structure and Holographic "external" reality field projection.

Radial Body Transduction-Manifestation Sequence- Continued page 3

9. As a Keylon-Thought-Form blueprint traveling on one Pulse of Consciousness **fully expands** (fully "Activates") into the manifest externalized hologram, **another Pulse of Consciousness** is released from the synchronized mathematical programs of the Planetary Merkaba Field, which sets the Transduction Sequence in motion ("Initiates")for the ***next portion* of the personal mathematical blueprint.** The blueprint of the **Planetary Kathara Grid** regulates the **cyclic rhythms** at which **Pulses of Life Force Current** will be released through the **Planetary Merkaba Field**; the Planet releases **"Pulses of Time"**.

NOTE: The linear time experience of life forms on a planet is synchronized into specific singular continua and cycles of time through the *DNA Template*; the species Tribal Shield, personal Maharic Shield and DNA Template are imbued with the *exact mathematical programs* of the specific *Planetary Time Cycle* and *Continuum* or "Time Wavef" into which the consciousness intends manifest embodiment. A singular Time Wave constitutes a Time Continuum, within which there are specific repeating cycles of progressively accelerating *Pulses of Time* that collectively form the linear time-progression continuum of a *Planetary Time Cycle*.

10. As one **Pulse of Time/Consciousness** brings its blueprint into manifest **Critical Mass Expansion** (fully activates specific portion of the blueprint in Holographic Expression), the Planetary Merkaba Field releases the **next "incoming" pulse** of Time, which carries the next portion of the Kathara Grid mathematical blueprint into the Transduction Sequence. As the new Pulse enters the Transduction Sequence and begins **Expansion** toward manifestation, the **first pulse** of the **fully expanded hologram** begins **Contraction,** entering a **back-return** through the **Reverse Transduction Sequence** (matter- *chemical* DNA-*Sub-Atomic units*- Radis *Mion/Dion units*- Radial Body *Trion -Meajhon units*- Meridian Line-Chakra-Auric Field-Hova Body-Axion Line-DNA Template-Merkaba Field-Morphogenetic Thought-form Field *Keylon Grids & Units*- Maharic Shield Template & Kathara Grid *Partika and Particum Units*- Planetary Merkaba Field *Partiki* Units)

11. A s **pulse-1 contracts** its hologram program **moving backward** through the Transduction Sequence, creating the **"Remembered Past" moment**, the **pulse-2** simultaneously **expands** its hologram program **moving forward** through the Transduction Sequence, birthing the as-yet-to-come **"Future" moment.** The backward moving **"Past"pulse-1** and the forward moving **"Future" pulse-2 intersect** at the **Radial Body Levels** in the Transduction Sequence. This **"past pulse-future pulse intersection point"** is called the **Eiron** (E'r-on) **Point**, also known as **"Zero Point"**, the **"Present moment of conscious focus within the hologram"**, or the "***Eiros*** Conscious, Breathing Life-Force Stream" (Also referred to as the **"Stream of Consciousness"**, the ***"Living Holy Spirit"*** or the ***"Breath of God"*** ..etc.).

12. The *Eiron/Zero Point*, or **Eiros**, is the **"Present Moment of Creation Power"** for the "manifest" conscious identity. It exists at the position where the **third dimensional field,** the electromagnetic **Resonant Tone "Mental Body"** field of a single 3-dimensional Hova Body, **intersects** with the **Meajhe Field Inner Layer** of the corresponding **Radial Body** Tissue Capsule.

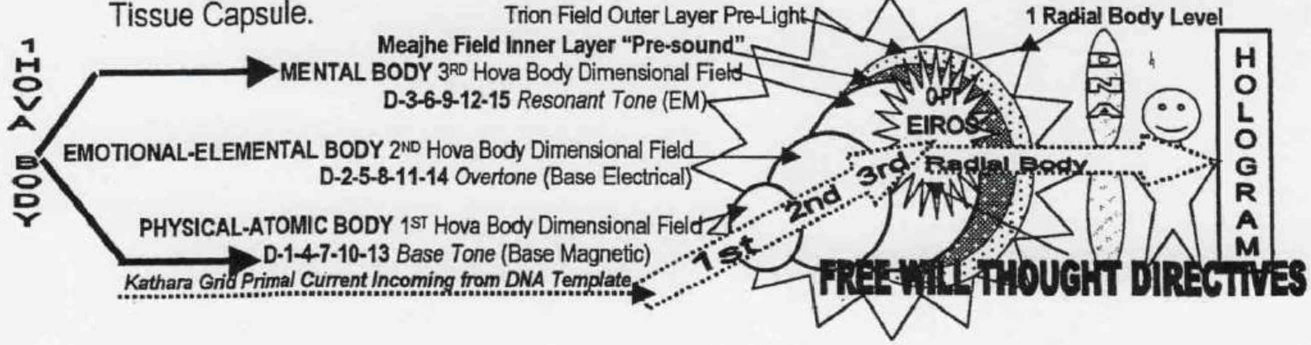
Trion Field Outer Layer Pre-Light
1 Radial Body Level
Meajhe Field Inner Layer "Pre-sound"
1 HOVA BODY
MENTAL BODY 3RD Hova Body Dimensional Field
D-3-6-9-12-15 Resonant Tone (EM)
EMOTIONAL-ELEMENTAL BODY 2ND Hova Body Dimensional Field
D-2-5-8-11-14 Overtone (Base Electrical)
PHYSICAL-ATOMIC BODY 1ST Hova Body Dimensional Field
D-1-4-7-10-13 Base Tone (Base Magnetic)
Kathara Grid Primal Current Incoming from DNA Template
1st
2nd
3rd
Radial Body
OPT
EIROS
DNA
HOLOGRAM
FREE WILL THOUGHT DIRECTIVES

***Radial Body Transduction-Manifestation Sequence-* Continued** page 4

13. As the backward moving **"Past" pulse-1** and the forward moving **"Future" pulse-2 intersect** at the Radial Body Meajhe Layer **Eiros** (Eiron Point) ✱ the returning **contracting pulse-1** passes through the Trion Layer and into the ***Meajhon blueprint replica,*** called a " **Radial TILE**" ▨, that it **left behind** in the Meajhe Layer **Eiros** during pulse-1's **incoming** passage. Just as the returning pulse-1 retreats back through its Radial Tile, the **incoming pulse-2,** carrying the **next blueprint** coding ▭ ,**simultaneously** passes through the **first Radial Tile, combining** the "past" **Radial Tile** mathematical-vibrational blueprint with the **incoming pulse-2 blueprint** ▨ At this moment-point, the hologram template from the "past moment" is adopted by the incoming "future moment" pulse-2 and expanded by the pulse-2 blueprint.

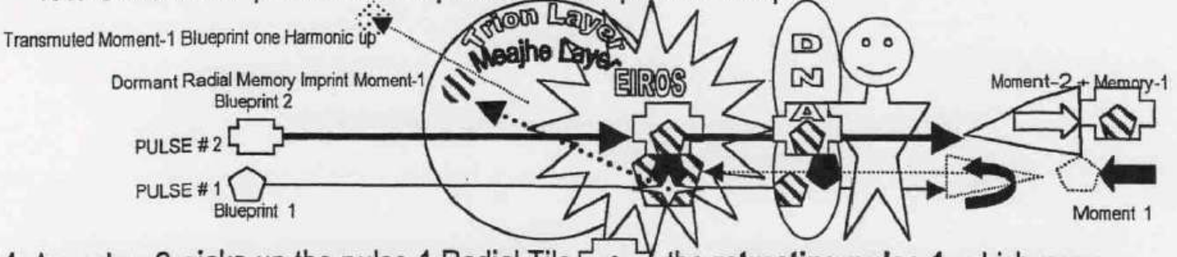

14. As pulse-2 picks up the pulse-1 Radial Tile, ▨ the **retreating pulse-1**, which now **holds the Radial Tile blueprint in *reverse mathematical sequence* as *TRION Pre-light Units*** ⬟ , collides with the incoming pulse-2. Both pulses carry the **common blueprint** of the first Radial Tile; the **incoming pulse-2** holding the blueprint in **forward** mathematical sequence as ***Meajhon Pre-sound Units*** ▨,the **retreating pulse-1** holding the blueprint in **reverse mathematical sequence** as ***Trion Pre-light Units*** ⬟ . As the expanding/advancing and contracting/retreating pulses collide with blueprint **"sequence and anti-sequence",** the pulse-1 reversed **Trion Radial "Anti-Tile" fuses** with the original **Meajhon Radial Tile** ✱ **and** both Radial Tile and Anti-tile **completely de-manifest** or "***Ascend***" from their original Harmonic Universe by **TRANSMUTING** to become a ***Merkaba Field.*** The **energy/ consciousness** once held by Radial Tile/Anti-tile then **replicates** its original blueprint via **Fission,** leaves a **dormant "Memory Imprint"** *Meajhon replica* ▨ in the Radial Body ***Meajhe Field,*** and **reappears** in faster oscillating form within the corresponding **Hova Body Form Constant** of the next Harmonic Universe up in the dimensional scale.

15. As transmutation of the "past" Radial Tile occurs in the Radial Body Eiros, the **incoming pulse-2** is momentarily held **static** as it leaves a **new Radial Tile** in the Meajhe Field; the new Radial Tile holds the **pulse-1+pulse-2** program. At this moment-point the **entire** externalized **Hologram** literally **"Flashes Off"** and de-manifests, its energy **held in suspension** within the **Radial Body** Trion/Meajhe Field. When the first Radial Tile is transmuted, the incoming **pulse-2**, carrying an **expanded imprint of the first Radial Tile**, progresses from the Meajhe Field Eiros to Trion Field, then begins to **"Flash On"** its blueprint in manifestation. The embodied self perceives a **"liner moment progression"**; the ***pulse-1***program becomes "**memory**" of the **"Past"** moment, ***pulse-2*** the **"Present"**.

The Radial Body, Mechanics of Manifestation, Healing and Human Potential

There is much more to be revealed regarding the subject of the *Mechanics of Manifestation*. Our introduction to the ***Radial Body anatomy***, the ***Transduction Sequence*** and the ***Mechanics of Manifestation*** represents a beginning in the quest for genuine mastery over the manifest experience. In this *Kathara Level-2 Introduction* we will **not yet address** the highly pertinent aspects of *Manifestation Mechanics* that pertain to **direct Creation of matter and events** through **Conscious Thought Projection** and applied personal **Will Intention**, nor will we yet investigate the complex realities of **Space-Time Travel, Star Gate Passage** and the **Spiritualized-science of Ascension**. All of these subjects and more are *intimately intertwined* with **personal healing** and the **Imprint for Health**, **attaining mastery** over Manifestation Mechanics and the Bio-Spiritual processes of genuine **Spiritual Actualization.** All of these abilities, and many others too numerous to mention, presently exist as **dormant evolutionary potentials** for the Human species; potentials that were once, long ago, fully actualized and commonly expressed. The ***Kathara Bio-Spiritual Healing System*** represents a body of Master teachings that focus upon the manifestation and maintenance of **Holistic Health through Bio-Spiritual Mastery and Spiritual Actualizaton.** Exploration of subjects such as ***Merkaba Mechanics***, ***Manifestation Mechanics*** and the ***Radial Body Trion-Meajhe Field*** is an intrinsic part of furthering our abilities to embody and practice the ideals of Healing and Health through Holism.

Before we can begin to actualize the exciting potentials of freedom, love, power and knowledge that await Human re-discovery, we must first focus upon **healing the obstacles** that have long stood in our way of actualizing true Human potential. Our study of *Merkaba Mechanics* in this *Kathara Level-2 Introduction* has included a *necessary introduction* to **Manifestation Mechanics** and **Radial Body Anatomy**; the greatest physical and spiritual difficulties presently facing Humankind emerge from a **HIDDEN ILLNESS** that has held our species hostage within a progressively deteriorating, **unnatural Mortal body form**. The **underlying cause of ALL Human illness** is an **unnatural genetic mutation** that manifests THROUGH the **DNA Template** and **chemical DNA.** The **Casual Element** of this species genetic deviation from its original Divine Blueprint can only be found through study of the 15-Dimensional Anatomy, ***Primal Life Force Currents, Kathara Grid, Maharic Shield, Merkaba Field, DNA Template***, the ***Radial Body*** and the intimate interplay that perpetually exists between these interwoven systems. It is precisely this ***species genetic deviation***, which has progressed and increased for over **200,000 years**, that has kept Humanity bound to a finite, limited world, unable to express the potentials that once existed as our birthright. Though the problem of this species "Internal Illness" is multi-faceted and highly complex, the **Soul-U-tions** through which progressive healing of the Human genome can be rendered are relatively simple. The **Soul-U-tions for healing** and restoring health to our species, and ourselves, **require familiarity with basic Kathara Grid, Manifestation Template, Radial Body, DNA Template and Merkaba Mechanics.** These are the elements of our 15-Dimensional anatomy within which the **Casual Element** of our genetic deviation is held, and thus these are the elements of our anatomy through which the **most rapid and permanent Healing Solutions** can be rendered.

In the following pages we will continue our investigation of the ***Transduction Sequence and the Miasmic Body***, in order to **identify** the regions of our 15-Dimensional Anatomy in which our hidden genetic mutations are created, stored and perpetuated. Once we can identify the regions in which the greatest healing is needed, we can then directly employ appropriate remedial measures. Working with the **Core Kathara Grid and Maharic Shield Divine Christos Blueprint** to reset the Imprint for Health (Kathara Level-1), while **simultaneously** working with the **elements of our anatomy within which the most significant remedial measures are needed**, will allow us to use Divine Blueprint technology to implement **expedited healing.** In this *Kathara Level-2 Introduction* we will examine the *Transduction-Manifestation Sequence* and how it's present **malfunction** creates the **Miasmic Body**, which creates the DNA Template distortion through which our **chemical DNA mutation**, and a phenomenon known as the "**Phantom**" or **"Shadow" Self** is brought into being within our manifest hologram. We will then continue our exploration of **Merkaba Mechanics** and **Radial Body Healing** applications, to discover the **"Secrets of the Veca Codes"**, and Photo-Sonic and Photo-Radionic Radial Body Healing technologies, through which our Kathara and Merkabic Healing applications can be tremendously intensified, amplified and expedited.

The 5 Radial Body-Hova Capsule-Trion-Meajhe Field Levels

and Dimensional, Triadic Phase Current, Density Level, Merkaba Phase, Hova Body, DNA Template, Axiatonal Line, Chakra, Identity Level and Memory Matrix Correspondences

NADIAL CAPSULE:(also called Telluric Capsule): Composed of **Dimensions-1-2-3** ***Telluric Kundalini Current*** frequency, **Density-1** Trion-Meajhe **Gross Matter Density**, 3-Dimensional **Nethra Phase Merkaba**, surrounds **Nada Hova Body**, corresponds to **DNA Strand Templates, Axiatonal Lines and Chakras 1-2-3** and holds focus of consciousness in the D-1 *Lower-Physical-Atomic,* D-2 *Lower-Telluric-Emotional-Elemental* and D-3 *Lower-Mental* ***Taurenic Bodies*** of the embodied Density-1 **Incarnate Personality-*Tauren*** identity, the Sub-conscious, Instinctual and Ego-Reasoning aspects of the "***Mortal Mind***" and the **Personal** ***Taurenic Memory Matrix***. *Radis* rivulets are called "***Nadis Lines***" (or Telluradis Lines)

DORADIC CAPSULE: Composed of **Dimensions-4-5-6** ***Doradic Kundalini Current*** frequency, **Density-2** Trion-Meajhe **Semi-etheric Matter Density**, 6-Dimensional **Hallah Phase Merkaba**, surrounds **Alpha Hova Body**, corresponds to **DNA Strand Templates, Axiatonal Lines and Chakras 4-5-6** and holds focus of consciousness in the D-4 *Astral-Mid-Physical-Atomic,* D-5 *Archetypal Mid-Telluric-Emotional-Elemental* and D-6 *Celestial Mid-Mental* ***Doradic Bodies*** of the embodied Density-2 **Incarnate Soul-*Dora*** identity, the "***Angelic Mind***" and the **Soul/Species** ***Akashic Memory Matrix***. *Radis* rivulets are called "***Doradis Lines***"

TEURIC CAPSULE: Composed of **Dimensions-7-8-9** ***Teuric Kundalini Current*** frequency, **Density-3** Trion-Meajhe **Etheric Matter Density**, 9-Dimensional **Quatra Phase Merkaba**, surrounds **Betcha Hova Body**, corresponds to **DNA Strand Templates, Axiatonal Lines and Chakras 7-8-9** and holds focus of consciousness in the D-7 *Higher-Physical-Ketheric,* D-8 *Monadic Higher-Telluric-Emotional-Elemental* and D-9 *Keriatric-Higher-Mental* ***Metatronic Bodies*** of the embodied Density-3 **Incarnate Over-Soul-*Teura*** identity, the ***"Archangelic Mind"*** and the **Over-Soul/Planetary** ***Askashic-Eckashic Memory Matrix*** . *Radis* rivulets are called "***Teuradis Lines***".

MAHARIC CAPSULE: Composed of **Dimensions-10-11-12** ***Maharic "Christos" Pre-matter Hydro-plasmic Liquid Light Primal Current*** frequency, **Density-4** Trion-Meajhe **Pre-matter Density**, 12-Dimensional **Mahunta Phase Merkaba**, surrounds **Mahara Hova Body**, corresponds to **DNA Strand Templates, Axiatonal Lines and Chakras 10-11-12** and holds focus of consciousness in the D-10 *Christiac-Transcendent Pre-Physical,* D-11 *Buddhaic Transcendant Pre-Telluric-Emotional-Elemental* and D-12 *Nirvanic Transcendent Pre-Mental* ***Hydronic Bodies*** of the embodied Density-4 **Incarnate Avatar-*Dolus*** identity, the ***"Christos Avatar Mind"*** and the **Avatar/Galactic** ***Dolaric Memory Matrix***. *Radis* rivulets are called "*Maharadis Lines*".

RISHIAC CAPSULE: Composed of **Dimensions-13-14-15** ***M Kee-Ra-ShA Ante-matter Thermo-plasmic Gaseous Pre-Light Primal Current*** frequency, **Density-5** Trion-Meajhe **Ante-matter Density**, 15-Dimensional **Rahunta Phase Merkaba**, surrounds **Raja Hova Body**, corresponds to "Indigo Grail Line" **DNA Strand Templates, Axiatonal Lines and Chakras 13-14-15** and holds focus of consciousness in the D-13-14-15 *Transcendent Ante-matter* ***Hedronic Bodies*** of the Density-5 **Breneua Rishi** Consciousness Collective identity, the ***"Rishiac Universal Mind"*** and the **Rishi/ Universal** ***Rishic Memory Matrix of one 15-Dimensional Time Matrix.*** *Radis* rivulets are called "*Rishadis Lines*".

The **5 3-dimensional Levels** of the **Trion-Meajhe Field Radial Body** within the 15-Dimensional **Time Matrix** connect each Density Level of all manifest consciousness to the greater, non-dimensionalized structure of the **Energy Matrix** ***Khundaray*** **Primal Sound Vibration** ***Meajhe Field*** through the **Density-5** Ante-matter ***Kee-Ra-ShA*** **Primal Light** ***Trion Field***.

The Transduction Sequence and the Miasmic Body

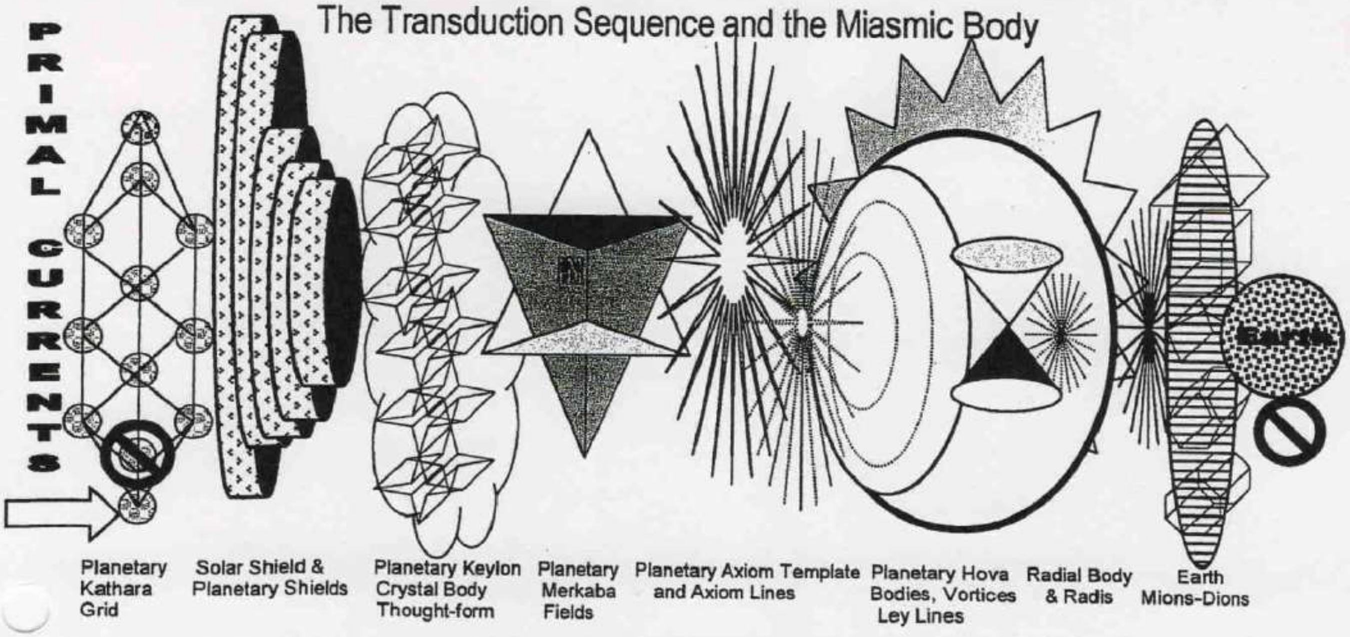
PRIMAL CURRENTS
Earth
Planetary Kathara Grid
Solar Shield & Planetary Shields
Planetary Keylon Crystal Body Thought-form
Planetary Merkaba Fields
Planetary Axiom Template and Axiom Lines
Planetary Hova Bodies, Vortices Ley Lines
Radial Body & Radis
Earth Mions-Dions

Personal Transduction Sequence

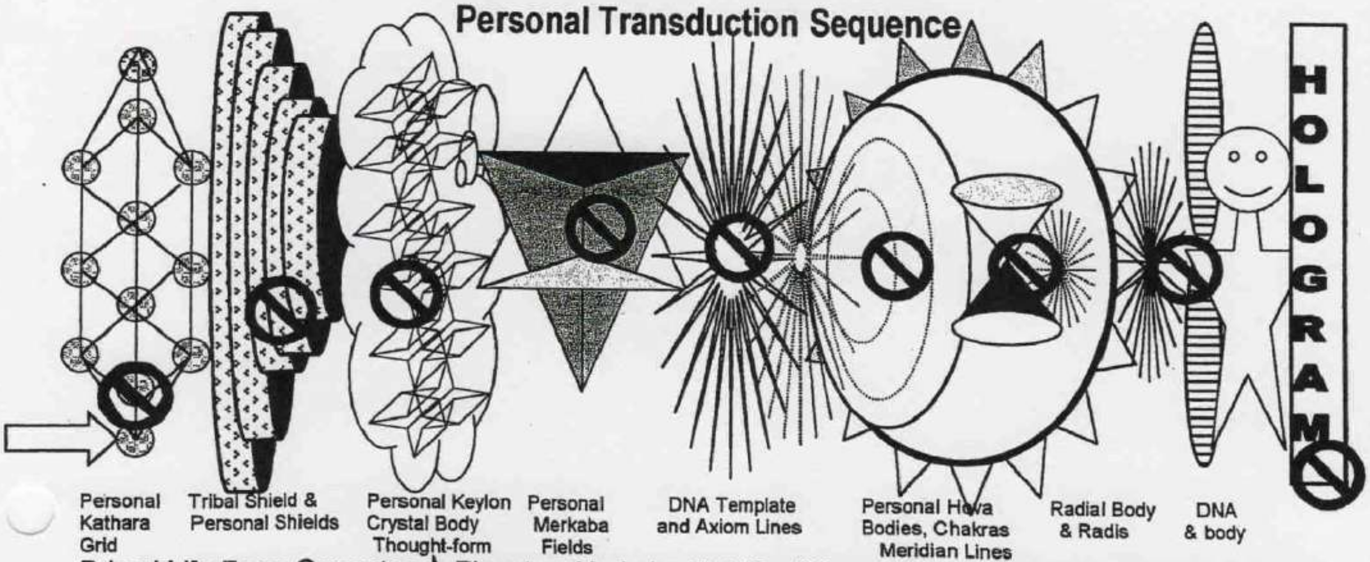
HOLOGRAM
Personal Kathara Grid
Tribal Shield & Personal Shields
Personal Keylon Crystal Body Thought-form
Personal Merkaba Fields
DNA Template and Axiom Lines
Personal Hova Bodies, Chakras Meridian Lines
Radial Body & Radis
DNA & body

Primal Life Force Currents ➡ Planetary Merkaba Field Partiki ➡ Kathara Grid -Partika-Particum ➡ Maharic Shield Divine Blueprint ➡ Keylon Morphogenetic "Thought-form" Field ➡ Merkaba Fields ➡ Kathara Grid DNA/RNA Template ➡ Axiom Lines ➡ Hova Bodies ➡ Auric Field ➡ Chakras ➡ Meridian Lines ➡ Trion-Meajhe Field 5 Radial Body Levels ⇨ Radis Lines ⇨ Mion & Dion Ionic Particulates ⇨ Strong and Weak Nuclear Force, Electromagnetism ⇨ Quarks, Muons, Measons, Sub-atomic & Atomic Particles and Anti-Particles ⇨ Blueprint translation into Nuclei, Molecules, Elements, Compounds, Chemicals ⇨ Chemical DNA Template and "Chemical Lens" of the manifest body form ⇨ Manifest Hologram ▭

⊘ **Miasms block energy flow and create a "Phantom-Shadow Self" running on opposite mathematical program; blocks natural energy flow.**

- The Personal ***Radial Body*** or Trion-Meajhe Field is composed of **5 Hova Body Capsules** that manifest as electromagnetic barriers or "**frequency fences**" between each set of 3 DNA Strand Templates, each set of 3 associated Axiom Lines, Chakras and Meridians and between each 3-dimensional layer of the Auric Field and corresponding level of personal multidimensional identity.
- Each of the **5** 3-dimensional **Harmonic Merkaba Fields** corresponding to the *5 Hova Bodies* of 15 Dimensional Anatomy *govern the function* of the corresponding **Dimensional Merkaba Fields, Radial Body level** and its imbedded **Memory Matrix**, via the mathematical program held in the personal Kathara Grid, Shields and **Crystal Body** template. The Kathara Grid program is passed from the **Keylontic Crystal Body** into the **DNA Template** via the **Dimensional and Harmonic Merkaba Fields** and the "mini-Merkaba Fields" of the Level-3 Kathara **Diodic Grid**.
- Each Hova Capsule **Radial Body level,** receiving its instructions from the DNA Template and 3rd dimensional Mental Body Hova Body level, in turn governs the function of the **chemically manifest DNA** and all body-consciousness systems that are built upon chemical genetic code.
- The **Harmonic Merkaba Field** governs the **particle spin axis** (Angular Rotation of Particle Spin) ratios of the energy units that make up its corresponding 3-dimensional system. Each **Harmonic Merkaba Field** corresponds to one **Hova Body**, one set of **3-dimensions** and the Matter Density Level characteristic of that 3-dimensional system. Each Hova Body, and thus each Harmonic Merkaba Field Level and Radial Body Level, correspond to one set of **3 DNA STRAND TEMPLATES** within the Human12-Strand Template.
- Full activation of **one Harmonic Merkaba Field** creates reciprocal activation of the corresponding set of ***3 DNA Strand Templates***, which creates embodiment of corresponding levels of identity and consciousness (personality, Soul, Over-Soul, Avatar, Rishi etc). When one Harmonic Merkaba Field fully activates through natural activation of the corresponding **Shields** and **Crystal Body Keylons**, the Harmonic Merkaba Field becomes capable of **merging** with the Harmonic Merkaba Fields from other Density Levels, in a process frequently called **Transmutative Dimensional Ascension**.

- During the process of "**Ascension**", (also called "space-time travel", "wave-riding", "Star Gate passage", "Teleportation" etc.) the **Harmonic Merkaba Fields** of more than one Density Level merge, progressively shifting the **two or more different sets of 3-dimensional particles** into a **common particle spin axis angle and spin speed**. The corresponding **Axiom Lines, Hova Bodies, Auric Field levels, Chakras, Meridian Lines** and **Radial Body Levels** also merge.
- When Harmonic Merkaba Fields, and their corresponding Inner Templar anatomy, merge, the **Fire Codes** between the corresponding **DNA Strand Templates activate, causing the DNA Strand Templates to merge,** which creates a reciprocal reaction called **"DNA Strand BRAIDING"** between corresponding **chemical DNA strands**. (The "Double Heli" configuration of "one strand" remains, as the chemical components of each strand overlay together to become a single Double Helix containing the chemical programs of multiple DNA strands.)
- When the **Fire Codes** between strands in the DNA Template activate, dormant sequences of chemical DNA called " **Transient Turnstile DNA Sequences**" "turn on" within the **Hydrogen Bonds** that link each DNA "Ladder Rung" together, producing a **transient trans-harmonic element** called **Celestalline** within the chemical DNA.
- Natural Braiding of the chemical DNA strands and activation of the chemical Turnstile DNA Sequences create a series of **sub-atomic bio-chemical DNA responses** that allows for Internal **Atomic Particle/Anti-particle Fusion** and **Fission.** Natural **Internal Atomic Fusion-Fission** enables the manifest atomic structure to temporarily de-manifest, returning to "*spirit form*", by re-entering the electrostatic **Trion-Meajhe Field** of the corresponding **Radial Body Levels.**
- During the Ascension process, the Harmonic Merkaba Fields of **2 or more Density Levels** merge, forming various **PHASES of Merkaba** activation that create a **Trans-Harmonic Merkaba Field** that is called the **Merkaba VEHICLE.** The 3-Dimensional **NETHRA Phase** Merkaba Field of Density-1accelerates in rotation speed to become the 6-dimensional **HALLAH Phase**, 9-Dimensional **Quatra Phase**, 12-Dimensional **Mahunta Phase** or 15-Dimensional **Rahunta Phase** Harmonic Merkaba Vehicle.

- As a phase of the Trans-Harmonic Merkaba Vehicle activates, the corresponding lower-dimensional **Hova Body and Radial Body Level dissolve**, their content of energy and consciousness **expanding** into the next Hova Body and Radial Body Level, forming a **Trans-Harmonic Radial Body "FORM CONSTANT"** within the **Merkaba Vehicle.**
- The Trans-Harmonic Radial Body and Merkaba Vehicle allow the consciousness and biological form to "de-manifest" from the lower dimensional Density Level. The identity can then **re-manifest** in another Density Level or space-time coordinate by projecting itself into the desired location, AS the *electrostatic* ***Trion-Meajhe Field*** **Trans-Harmonic Radial Body,** then returning the DNA Template activation level and Harmonic Merkaba Spin ratios to that of the new location for re-manifestation.
- When the Personal **Merkaba Field** is used to form the **Trans-Harmonic Radial Body** in the process of Ascension, the Merkaba Field is called a **MERKABA VEHICLE.** A ***Merkaba Vehicle*** is an electrostatic **Trion-Meajhe Field** of Primal Pre-Light-Sound through which the consciousness (AKA "spirit") can move itself, *and its atomic structure*, through **thought-projection**, to a desired location beyond the "veils" of the Planetary, Galactic and Universal Hova Capsule-Radial Body-Form Constant barriers.
- Conscious mastery of the Personal Merkaba Field, through which the organic **internally created Merkaba Vehicle** Trans-Harmonic Radial Body becomes available for conscious direction, implies **conscious control of DNA Template function.** The potentialities of **genuine Ascension** (etc…) and mastery of Bio-Spiritual Consciousness are **inborn attributes** of the Human 12-Strand DNA Template. Due to a **species de-evolutionary genetic mutation** that compromises natural function of the Personal Merkaba Field and DNA, the reality of Ascension on a species level has been lost to all but a few on Earth for over 200,000 years (re: Masters Templar Coursebook).
- The potentials of **Bio-Spiritual Actualization** CAN be returned to humankind through **reverse-mutation** of genetic dysfunction within the human genome and DNA Template. Healing the genetic template can be achieved by **manually employing** specific practices of **Kathara "Maharic Shield" Core Template Dynamics** in **combination** with **Harmonic Merkaba Mechanics** through which the many eons of biological miasmic distortion can be progressively and rapidly cleared from the Personal Kathara Grid and DNA Template.

- Restoration of the D-12 Pre-matter Divine Blueprint within the Human Kathara Grid and DNA Template will eventually **free the Personal Merkaba Field** for use as a **Trans-Harmonic Merkaba Vehicle.** Kathara Healing applications, and use of the D-12 Pre-matter Maharata "Christos" Current to reset the biological D-12 Pre-matter Divine Blueprint, are the beginning steps in reclaiming Bio-Spiritual Mastery.

- The function of the personal **Merkaba Field** and **DNA Template** are not only associated with abilities of **Atomic Transmutation** and **Dimensional Ascension**, they are directly linked to **every aspect of body-mind-spirit function**, including **biological longevity** and **health, memory, intelligence,** and **metabolic, chemical, mental** and **emotional balance** and stability.

- Most significantly, the **Merkaba Field/DNA Connection** is the Bio–Spiritual conduit through which the consciousness of an embodied individual is directly linked to the common Perpetual Force of Energy and Eternal Source of Intelligent Creation that is "**God".**

- When a being's Merkaba Field and DNA Template are functioning upon the innate program of **Universal Primal Order**, as it is held within the Pre-matter **Universal Divine Blueprint** of the D-12 "Christos" Liquid-Light field, the natural and perpetual supply of **Primal Life Force Currents** and "spiritual" consciousness emanating eternally from Source-God, flow through an open energetic channel or "**Pillar of Light**", which is called the "Um Shaddai Ur" (Pillar of First Cause Light-Sound') into the manifest embodiment. Such a being is called a "**Christed Being**". This implies that the integrity of its original D-12 Pre-matter Divine Blueprint has not been structurally compromised, and thus the perpetual Life Force Currents, consciousness, sentience, intelligence, wisdom, power, Will and LOVE of Source can freely express through the Um Shaddai Ur, into the manifest hologram of the Time Matrix through the Christed embodiment.

- If the Kathara Grid, DNA Template and Merkaba Field of a being are damaged, and no longer carry the **original mathematical-geometrical program** of the **D-12 Divine Blueprint** through which the being entered space-time, the flow of Primal Life Force Currents, the Um Shaddai Ur Pillar and thus the active consciousness connection to God will be distorted, reduced or severed at the dimensional level to which the template distortion corresponds.

- A manifest being incapable of running the Primal Life Force Currents of the Um Shaddi Ur "Pillar" becomes **biologically finite**, its consciousness "trapped in the Time Matrix", even following death of the physical body. According to the **natural Laws of Creation Physics** as set by Source-God, this condition of **finite mortality** will remain until the being's **Kathara Grid Template and Merkaba Field connection to** the D-12 **Universal Divine "Christos" Blueprint** and Universal Trion-Meajhe Field Um Shaddi Ur Pillar is restored.
- Healing Core Template genetic distortions that are severe enough to cause dysfunction of the Personal Merkaba Field first require **resetting of the D-12 Divine Blueprint** through activation of the **Personal Maharic Shield.** Until the Body-Mind-Spirit system has reached a **Critical Mass** of D-12 Divine Blueprint holding through frequent and progressive manual running of the D-12 Maharata Current, healing Core Template genetic distortions **also requires** simultaneous, direct and repeated **manual realignment and restoration** of **Personal Merkaba Field function** and **Radial Body Trion-Meajhe Field** integrity.
- If there **are distortions within the Kathra Grid or DNA Template** of an individual or species, the Personal Merkaba Fields will not function in harmony with the **natural energy circulation rhythms** of the **Planetary, Galactic, Universal** and **Time Matrix Merkaba Fields** and Universal Kathara Grid. Such Kathara Grid distortions create tangible **genetic and consciousness mutation** within biological organisms, and can divert the **intended evolutionary destiny** of an embodied consciousness or species on to a divergent path of **de-evolution and consciousness fragmentation** unintended by the original plan encoded within the D-12 Pre-matter Christos Divine Blueprint.
- Distortions with the DNA Template and resulting **Miasms** can be purged through **manual resetting of the D-12 Pre-matter Divine Blueprint,** via repeated manual activation of the Personal **Maharic Shield**. The activated Maharic Shield runs the Density-4 D-12 **Maharata Primal Life Force "Christos" Current** through the Kathara Grid and form embodiment, progressively resetting the original mathematical coding of the personal Divine Blueprint within the Kathara Grid and DNA Template. Use of the **Maharic Shield techniques** in *Kathara Level-1* sets this process of **Core Template Regeneration** in motion.

- The Core Template "Divine Blueprint" healing process can be expedited through applied **Christos-Merkabic Healing Mechanics** such as the ***Tribal Shield Activation*** (Technique-7), ***Merkaba Salutation*** (Technique-8), ***Merkaba Trinity Key Induction*** (Technique-9) and ***Kee-Ra-ShA Activation Sequence*** (Technique-12).

- Divine Blueprint Healing can be expedited further through **manual restoration** of the personal Trion-Meajhe Field **Radial Body** and its inherent **Memory Matrix. Radial Body Healing,** when **combined with** ***Kathara Maharic Shield*** and ***Christos-Merkabic Healing*** dynamics is the most powerful and rapid Holistic Healing technology available.

- **Radial Body Healing** dynamics, which also serve to progressively restore the organic **Memory Matirx,** involve use of the specific **mathematical programs** that govern formation of the **Radial Body Trion-Meajhe Field.** These mathematical programs are called **VECA CODES.**

- **Technique-10** ***Restoring the Trion Field-Photo-Radionic Healing*** **and Technique-*11Restoring the Meajhe Field-Photo-Sonic Healing,*** provide introduction to **Radial Body Healing Mechanics.** Use of these techniques, in **combination** with **Technique-2** ***The Maharic Seal and Liquid Light Cleanse,*** and **Merkabic Healing Techniques 6-12, will greatly expedite** the personal healing process while **significantly increasing** the body-mind-spirit abilities to **receive, hold/sustain and transmit greater quantity and quality** of interdimensional frequency in the Client Healing Facilitation process.

Radial Body Healing, the VECA CODES, DNA and Merkaba

- Veca Codes, also called the "I AM" or "**Immanuyana Sequence**" are very specific **SYMBOL CODES**, each possessing corresponding **Sound Tones** called ***Arieas*** and **energy movements** called ***Seurias***, which correspond to the Cosmic, Universal, Galactic, Planetary and Personal **Trion-Meajhe Field Radial Body Levels**.
- There are **5 categories of Veca Codes**, the Planetary **Low-Veca *Codes of Manu*** (Taurenic-Akashic Codes), the Galactic **Mid-Veca *Codes of Immanu*** (Akashic-Eckashic Codes), the Universal **Level-1 High-Veca *Codes of Immanuel*** (Dolaric-Rishic Codes), the Cosmic **Level-2 High Veca *Codes of Immanuyana*** (Triadic-Polaric-Eckatic Codes) and the God-Seed **Yunasai *Codes of Eckasha*** (Source Codes).
- Beginning applications of Veca Code Radial Body Healing begin with restoring the Trion-Meajhe Field Radial Body through resetting the ***Yunasai*-Eckasha and High Veca *Immanuyanas* and *Immanuel* Codes**, through **frequent manual Direct Induction** of the their corresponding **Symbol Codes** and **Sound Tones/Arieas**.
- The most powerful **Yunasai-Eckasha Codes** are **PHOTO-SONIC Codes**, carrying the Mathematical Programs of the ***Khundaray Current*** Primal Sound ***Meajhe Field*** AND their bridging **God-Seed vibration-rhythms** connecting to Source. The **Level-2 *Immanuyana*** and **Level-1 *Immanuel* High Veca Codes** are **PHOTO-RADIONIC Codes**, carrying the Mathematical Programs of the Trion-Meajhe Field **Khundaray Primal Sound** AND **Kee-Ra-ShA Primal Light Currents.**
- **Consistent use** of **Veca Code Radial Body Healing**, in combination with **Merkabic and Kathara Healing** applications, will progressively **clear the Miasmic Body** and heal, balance, re-integrate, and restore the Divine Blueprint within, the energies of the **Phantom-Shadow Self** in all levels of the Body-Mind-Spirit system.
- More advanced applications of Veca Code technologies include use of the **Mid-Veca *Immanu*** and **Low-Veca *Manu* Codes,** and the corresponding ***Seurias*** (body movements to direct frequency), following a period of initial induction and activation of the Yunasai and High Veca Codes. Full Spectrum Veca Code technologies are introduced in the "***Dance of Life***" Masters Spiritual Development Program.

- **Induction,** or **IMBEDDING** of the High Veca Codes into the 15-Dimensional Anatomy involves placing the corresponding **Symbol Code Sequence** (mathematical program) over the **regions of the physical body** that correspond to the appropriate aspects of the Kathara Level-2 **Crystal Seals Grid** and Level-3 **Diodic Grid.** Once the Symbol is in place, the **breath** is used to direct **D-12 Maharata Current** through each **Symbol Code** and **body region.** Use of the **Maharata Current** "carrier wave" (via **activation of the Maharic Seal/Shield**) is **required** to **induce** the code's **mathematical program** into the **Diodic** and **Crystal Seals Grid,** where the program will then naturally transfer into the corresponding areas of the **Level-1 Kathara 12-Tree Grid** core template.
- Once the High Veca and Eckasha Code Sequences are Imbedded/induced into the Level-1 Kathara Grid, they are then **Activated** in the Kathara Grid by **TONING** the corresponding **Veca Code Ariea Sequence**. The longer toning is sustained, the greater will be the quantity and quality of Veca Code Activation.
- Once Activated, the Veca Code programs travel upward through the stages of the **Transduction-Manifestation Sequence** from the **Kathara Grid**, to the **Shields**, into the **Keylon Crystal Body Morphogenetic Field,** then into the **Merkaba Field, DNA Template** and **Hova Body/ Radial Body Levels**, temporarily restoring the **Christiac Merkaba Field, Radial Body** and **Memory Matrix Divine Blueprint**. **Frequent and continued use** of the Veca Codes will progressively build **Critical Mass** of the **Divine Blueprint program simultaneously** within **ALL levels** of the 15-Dimensional Anatomy, eventually culminating in expedited full restoration of the D-12 Maharic Shield blueprint within the Body-Mind-Spirit system, via the Radial Body Trion-Meajhe Field.
- Activation of the Yunasai and High Veca Codes within the personal body **prior to conducting a Kathara Client Healing Session** will cause the facilitator's **Level-3 DIODIC GRID to activate,** temporarily **transmitting Veca Code sub-harmonics.** This will cause the **minute Diodic Chakras** in the facilitator's **FINGER TIPS** to activate **if the Palm Chakras** have been activated with either pure **Maharata Current** or a selected **Triadic Phase Current.** The facilitator will then **automatically transmit** Veca Code sub-harmonics into the client's **Diodic Grid** while conducting general Kathara Healing applications, allowing the client to benefit from temporary **Radial Body healing support**, which will **amplify the effects** of all healing applications.

- A more **Temporary Imbedding and Activation** of the Veca Code program can be achieved through **OPTICAL-PINEAL INDUCTION**, whereby the **eyes** are focused for a time upon the **Image** of the **Symbol Code.** The mathematical Veca Code program enters the **Pineal Seals** through the **Optical Currents** and **6th Chakra**, an then travels for Imbedding/induction through the **Central Vertical Current** into the body regions and Crystal Seals Grid areas to which the Veca Code corresponds. The **Veca Tones/Arieas** are then used to Activate the Imbedded Veca Code Sequence.

- In **Optical-Pineal Induction**, the Veca Program looses a bit a **frequency charge/ strength** during is travel in the Central Vertical Current. This creates a **more temporary imbedding and activation** of the Veca Codes than that which is achieved through Direct Induction. Optical-Pineal Induction is useful for **daily amplification** of the Vecas, and for Kathara Healing Client Sessions and planetary grid "site work", when used in **combination** with the **full induction procedure** employed at least **once per week**.

- To receive **full benefit** of **Veca Code Radial Body Healing** applications, **Technique-2 *The Kee-Ra-ShA Activation Sequence*** should be fully run at least **ONE TIME**. *Technique-2* can also be used occasionally thereafter if desired to create **intensification and amplification** of the Primal Life Force Currents during Kathara, Merkabic and Veca Code Healing technologies.

- Frequent use of **Technique-13: The "Song of Lyra"** Christos Invocation will amplify all Kathara, Merkabic and Veca Code Radial Body Healing applications.

The Christos-Trion-Meajhe Fields, High-Veca Codes and Universal Life Force Currents

One 15-Dimensional Time Matrix with 5 Harmonic Universes

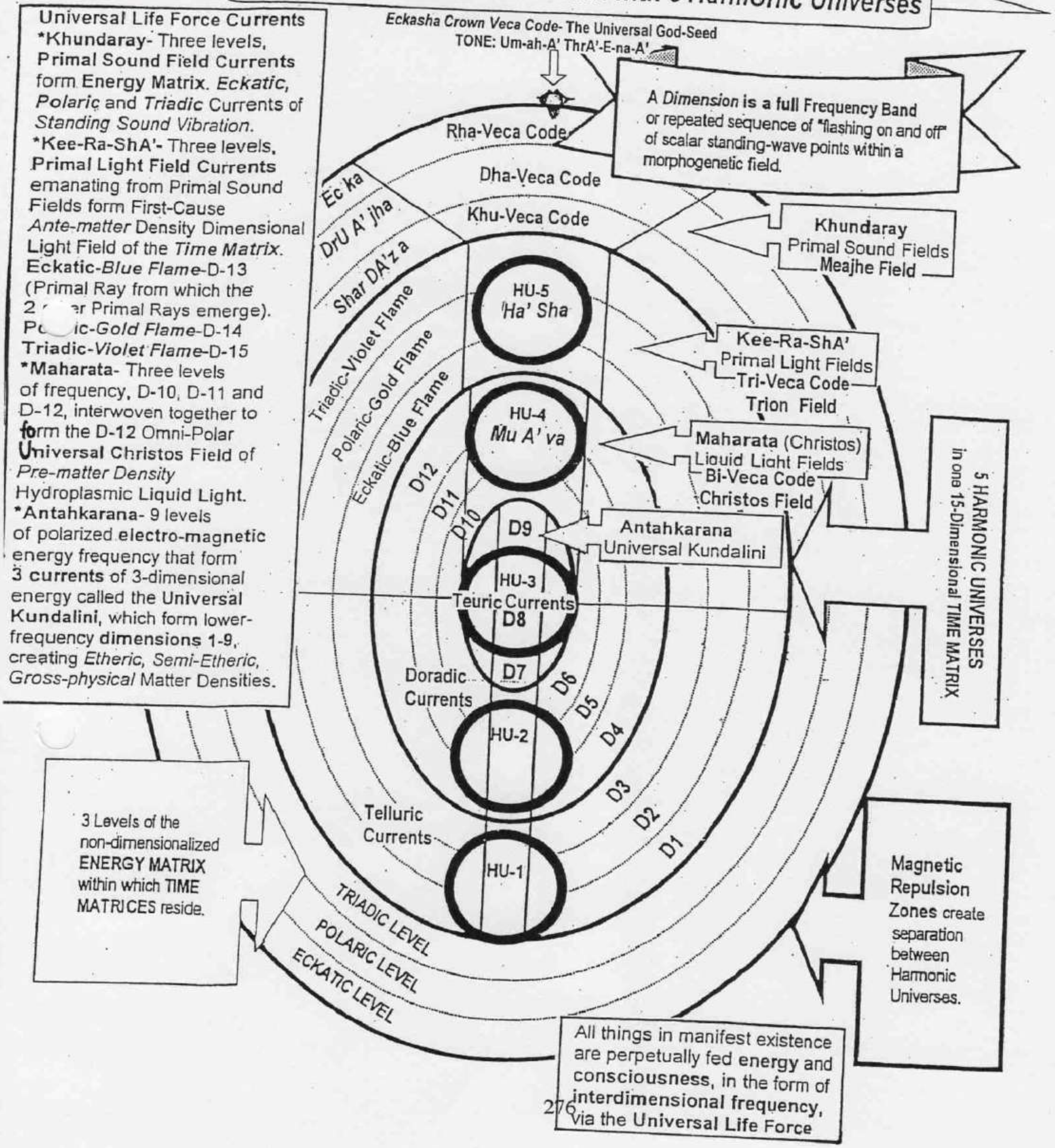
Eckasha Crown Veca Code- The Universal God-Seed
TONE: Um-ah-A' ThrA'-E-na-A'
Rha-Veca Code
Dha-Veca Code
Khu-Veca Code
Ec Ka
DrU A' jha
Shar DA'z a
Triadic-Violet Flame
Polaric-Gold Flame
Eckatic-Blue Flame
HU-5
'Ha' Sha
HU-4
Mu A' va
D12
D11
D10
D9
HU-3
Teuric Currents
D8
D7
D6
D5
D4
D3
D2
D1
Doradic Currents
HU-2
Telluric Currents
HU-1
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL
Khundaray
Primal Sound Fields
Meajhe Field
Kee-Ra-ShA'
Primal Light Fields
Tri-Veca Code
Trion Field
Maharata (Christos)
Liquid Light Fields
Bi-Veca Code
Christos Field
Antahkarana
Universal Kundalini
5 HARMONIC UNIVERSES
in one 15-Dimensional TIME MATRIX

Universal Life Force Currents
***Khundaray**- Three levels, **Primal Sound Field Currents** form **Energy Matrix**. *Eckatic*, *Polaric* and *Triadic* Currents of *Standing Sound Vibration*.
***Kee-Ra-ShA'**- Three levels, **Primal Light Field Currents** emanating from Primal Sound Fields form First-Cause *Ante-matter* Density Dimensional Light Field of the *Time Matrix*.
Eckatic-*Blue Flame*-D-13 (Primal Ray from which the 2 other Primal Rays emerge).
Polaric-*Gold Flame*-D-14
Triadic-*Violet Flame*-D-15
***Maharata**- Three levels of frequency, D-10, D-11 and D-12, interwoven together to form the D-12 Omni-Polar **Universal Christos Field** of *Pre-matter Density* Hydroplasmic Liquid Light.
***Antahkarana**- 9 levels of polarized **electro-magnetic** energy frequency that form **3 currents** of 3-dimensional energy called the **Universal Kundalini**, which form lower-frequency **dimensions 1-9**, creating *Etheric*, *Semi-Etheric*, *Gross-physical* Matter Densities.

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

3 Levels of the non-dimensionalized **ENERGY MATRIX** within which **TIME MATRICES** reside.

Magnetic Repulsion Zones create separation between Harmonic Universes.

All things in manifest existence are perpetually fed energy and **consciousness**, in the form of **interdimensional frequency**, via the **Universal Life Force**

Technique # 4

Restoring the Trion Field-Photo-Radionic Healing

Regular use of this technique will strengthen and improve RRT
Planetary healing Applications (excerpt from Kathara Level-2)

Creating a Personal Bio-field Link to the Universal Christos-Trion-Meajhe Field

Masters Merkaba Mechanics
© 2001 Ashayana and Azurtanya Deane

Exercise-1: **Once** per day, do **Maharic Insta-Seal** then practice **each set of Veca** tones separately while fixating **visual focus** on corresponding **Veca-Code** symbol; spend at least **three or four minutes** gazing at each symbol (optical-pineal induction) while **toning** corresponding tones (core template vibrational induction) With practice this will progressively expand the personal consciousness while activating the personal Christos-Trion-Meajhe Field of your Christos-Rishiac-Ascended Master identity levels within the cellular structure.

Exercise-2: **Once per week**, directly after completing **Exercise-1**, place all Veca-Code symbols over designated body areas, activate Maharic Shield, breathe D-12 Maharata Current through each symbol and body area to **induce code**, using the Bi-Tri-Khu-Dha-Ra Sequence, then begin **toning** the corresponding tones following the Bi-Tri-Khu-Dha-Ra Sequence. Memorize the tone sequence of the 5 High-Veca Codes as if it were a **"song"** and relax, close eyes and sing/tone this "song" (Psonn) for **5 to 10 minutes**. The longer you tone, the more frequency power you build within the body and the stronger your link will be to the Universal Christos-Trion-Meajhe Field. This exercise can also be used with groups, either singularly to create and amplify Christos-Trion-Meajhe Field links or just prior to beginning RRT work. Frequent and consistent use of these two exercises will progressively develop the spin-rate capacity of the personal Christos Merkaba Field, while expanding higer consciousness within the body and progressively enhancing the health, immunity and integrity of the Mind-Body-Spirit system.

The High-Veca Codes

Universal Time Codes for Primal Frequency Access

Cosmic Codes of Immanuyana

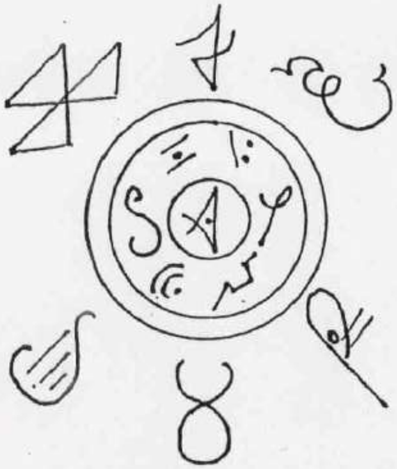

The Rha-Veca Code
Primal Sound-3 Eckatic
called the *Ec ka*
TONE:"*Rha Ec' ka*"
Symbol Placement:
Tailbone

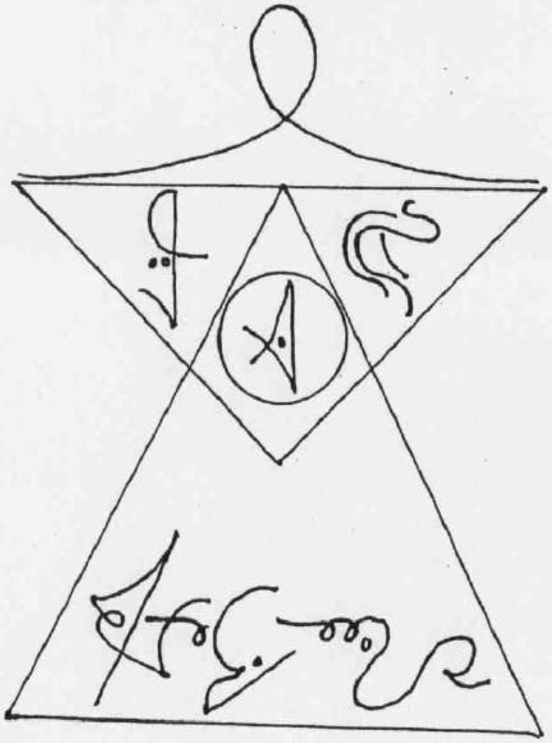

The Dha-Veca Code
Primal Sound-2 Polaric
called the *DrU A' jha*
TONE: "[illegible]' *Dru A' jha*"
Symbo[illegible]lacement:
Chakra-2 (just below naval)

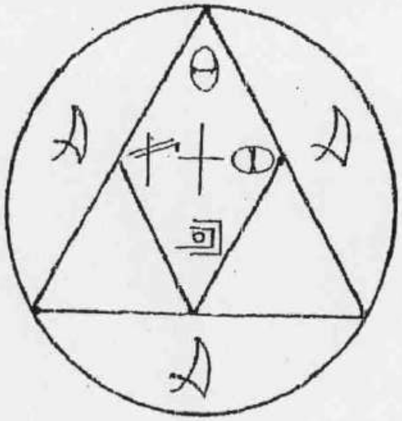

The Khu-Veca Code
Primal Sound-1 Triadic
called the *Shar dA'z a*
TONE:"*Kh*[illegible]*har DA'z a*"
Symbol Placement:
Chakra-8 Front (clavicle)

Universal Codes of Immanuel

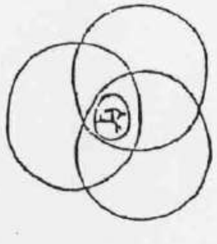

The Tri-Veca Code
Density-5 Dimensions 13-14-15
called the *Ha' Sha*
TONE:"*Kee' Ra ShA Ha Sha*"
Symbol placement:
Left Foot bottom

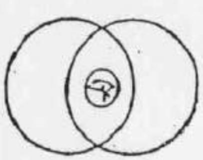

The Bi-Veca Code
Density-4 Dimensions 10-11-12
called the *Mu A' va*
"*Ma ha ra' ta Mu A' va*"
Symbol placement:
Right Foot bottom

Technique #11

Restoring the Meajhe Field-Photo-Sonic Healing

1. Use **Technique-6:** ***The Maharic Quick Seal*** to activate your Maharic Shield, Seal and Maharata Current.
2. **Stare at image** of the **Eckasha Code** for a few seconds, then gently begin to **"Cross Eyes"** until you can see a **"Double Image"** of the *Eckasha* when looking at the image. **Hold " Eye Cross"** focus while taking **three rapid breaths**; on each **EXHALE** projecting a **stream** of (D-6) **INDIGO current** from your **6th "3rd Eye" Chakra at** forehead. (This braids D-6 Optical Currents into a D-12 Current to carry the Eckasha program into the Pineal Seals.). Un-cross eyes, relax visual focus and close eyes for a moment or two.
3. **INHALE** several breaths of **Pale Silver D-12 Maharata Current** up from the **Earth Core** through your **Maharic Cord**, on each **EXHALE** depositing the Maharata Current in the **Pineal Gland** at the center of the brain, to form a Pale Silver **"Maharata Ball"** in the Pineal.
4. Breathe gently and **return visual focus to the image of the Eckasha**, focusing upon the **Eckasha Core** ***REUCHE'*** (the "Double Circles and their content at center of Eckasha).
5. **EXHALE** while projecting a **stream of INDIGO current** from the **6th Chakra** at forehead, into the **REUCHE** symbol before your eyes. **INHALE** while drawing the **INDIGO current stream** with the **REUCHE image at its end** back through the **6th Chakra** and into the **Maharata Ball** in the **Pineal.**
6. **INHALE** one more breath of **Maharata Current** up from **Earth Core** to the **Maharata Ball** at the **Pineal**, then use a forceful EXHALE to PUSH Eckasha-Reuche encoded Maharata Ball out to the right, to the "Mentor" position over the right shoulder. Visualize the Maharata Ball growing to the size of a "Basket Ball" over the right shoulder, and visualize the image of the Reuche as expanding to fill the inside of the Maharata Ball.
7. **INHALE forcefully** while drawing the **Maharata Ball** back to a position **centered around your head**, then use **three breaths** to **INHALE Maharata Current** up from **Earth's Core to Pineal** and **EXHALE** the **Maharata Current outward** to fill **and expand the Maharata Ball.**
8. Visualize the **Maharata Ball** expanding outward from around your head, until the Maharata Ball **fully encompasses your body and entire Chakra System**, from **Chakra-14** 36" above head to **Chakra-13** in Earth Core. Imagine that your entire body is now surrounded by a 3-dimensional "Eckasha Tear-Drop" shaped CLEAR CRYSTAL CAPSULE. The Eckasha Code Photo-Sonic Capsule is now temporarily activated in your Inner Templar system.

Eckasha Crown Veca Code- The Universal God-Seed

Tone: Um ...... ⇩ ah...........A' ⇧ ThrA'....... ⇩ E'- ⇩ na......A

Symbol Placement: Pineal Induction via Chakra-6 "3rd Eye"

The Yunasai-Eckasha God-seed Code

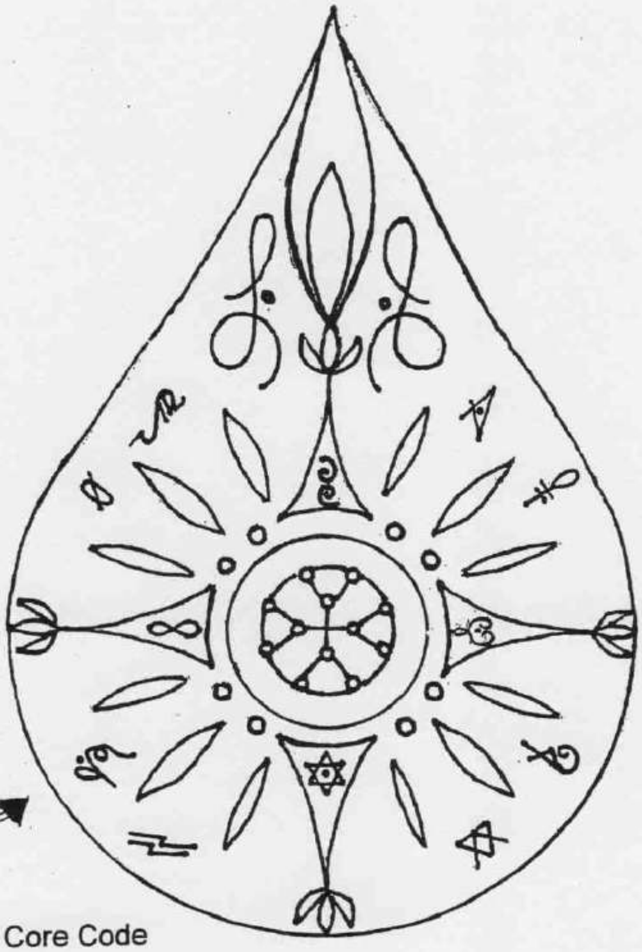

The REUCHE' Eckasha Core Code

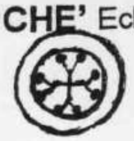

Technique # 13

The Song of Lyra Christos Invocation

The Vectos Invocations

The Vectos Compendium Declariti

The Vectors Collective Declarations

(All self incarnations in all Time Vectors declaring in Unity as the Avatar Self)
English translation from Elohei-Anuhazi Emerald Covenant Decree Texts:
© 2001 Ashayana Deane

The Mah'-ah' Bi-Vectus Declaritum

Declaration of the Maharata

(Intention Decree AS the personal D-12 Christos Avatar Maharata self identity)

I decree now in this moment of KNOWING that all things here within me
are hence forth restored to the original Divine Intention of the ONESELF that is MINE.

I release now all bonds of manifest illusion that place veils of SELF-forgetting before mine eyes.
I revoke all separations of duality through which my Heart is rendered numb with disillusion and discord.

In this moment I reclaim the Whole of my Condition, reaching forth into the depths of all creation to now call my SPIRIT home.

I AM the Sun of the Radiant Glowing. I AM the HEART of Divine LOVE. I AM the Still Point of the Face of All Creation. I AM the Voice of the Universal Song.

I decree now in full remembered Knowing that I AM ALL and WE are ONE.
In setting forth this decree of my Eternal Divine Power, I shall render Divine Intention only, through which the All and I are equally served
in the Highest Order of our collective Best Divine Intention.

I call now unto the many Selves and Shadows ensconced within the illusions of time, and decree unto thee that I AM THIS, I AM.
I AM the wind that brushes thy faces.
I AM the rain that renews thy field.
I AM the Whisper of the Breathe of God within YOU.
WE NOW are the expression of Divine Heart, Mind and Hand.
WE are the Rumbling Hymn Song of the earthen worlds of matter.
WE are the Prana-Breath of God Eternal, manifest.
We are the Waters of Eternal Life forever outward flowing.
We are the cooling, life force Flame of the Cosmic Fire expressed.
In our UNION now and always, declare WE in our Wholeness,
that all dissidence is rendered spent, all conflict NOW made healed.

As DIVINE NAME THIS..................(your Christos spiritual name)
I AS WE, the Maharata KHUM BI-Vec'TUS, (D-12 Christed Avatar Divine Blueprint self identity), DO NOW DECREE through the Divine Love, Wisdom and Power that is MINE....to express now AS this moment, the condition of **Mah'-ah BI-Vec'TI UN UR'-A-OR'-NaM' –OOR** (Christos beginning and ending NOW-completion of a manifest expression in its originally intended Divine Fulfillment, its past-present-future full expression made in the present time, clearing all duality, disharmony and distortion. Anchoring of the Divine Christos Maharata Blueprint in the moment). Upon this Invocation Divine Intention is rendered manifest in all conditions to which I now assign.

The Mah'-ah' Bi-Vec'tus Inavhoki
Invocation of the Maharata

(Creation Decree Invocation AS the personal D-12 Christos Avatar Maharata self identity)

© 2001 Ashayana Deane

From this Still Point of my Wholeness in manifest decision,
I Decree in Sovereign Knowing, here expressed at my Command *(lovingly stated absolute intention)* ***the Divine Will and Intention of the ONE-SELF that is ONE GOD.***
The ONE-SELF that IS ONE GOD THIS I AM.

Mah–Hah-Rah-Tah KHUM BI-Vec'TUS ……………………(name)
I as D-12 Christos Avatar self…….(name)
IN-A-VHO'-KI UN' E-BLE'UM BI-Vec'TI
Invoke by Divine Loving Command Now, by the Power of the Universal Christos,
Mah'-ah BI-Vec'TI UN UR'-A- OR'-NaM' – OOR
the D-12 Divine Christos Blueprint past-present-future; NOW the Divine Sea of Liquid Light anchors in this moment.
E'-stA Un'-tA E'-Sa Tra'zd-Jha' HA'-A Rha
Absolutely now established, always and forever, in eternal abiding reverent Love of the Eternal ALL-ONE
'tah-TA Um a Ah-ShA'-Lum
Blessed Be all embracing, given forth from the Still Point of Eternal Peace
Thah-A'-Jha in'ta DO A"
Spoken AS IT IS STATED, As SO It IS DONE
Um Sha'-DI UR'- A ah Khum' TUn
All Embracing Pillar of First Cause Light, I AS THIS expression now.

In the name of the Melchizedek Cloister Emerald Order Eieyani,
Holy Order of the Yunasai,
Sacred House of ONE.

English Translation:

Ma-Ha-Ra-Ta KHUM BI-Vec'TUS ……………………(name)
(Khum = "I AS") (Ma-Ha-Ra-Ta BI-Vec'tus = D-12 Christos Avatar Self)
iN-A-VHO'-KI UN' E-BLE'UM BI-Vec'TI
(*in-A-vhO-kI*= invoke by Divine Loving Command), (*UN*= Now), *(E-BLE'UM*= by the power of),
(Bi-Vec'ti = Universal Christos)
Mah'-ah BI-Vec'TI UN' UR'-A- OR'-NaM' – OOR
(Mah'ah Bi-Vec'tI= D-12 Divine Christos Blueprint past-present-future)
(UN Ur'-A- Or'-Nam' –OOR= Now the Divine Sea of Liquid Light anchors in this moment)
E'-stA Un'-tA E'-Sa Tra'zd-Jha' HA'-A Rha
(E'StA Un'TA E'Sa = Absolutely now established, always and forever)
(Tra'zd-Jha'= in eternal abiding reverent Love) (HA'-A Rha = of the Eternal ALL-ONE)
P'tah-TA Um a Ah-ShA'-Lum
(P'tah-TA UM = Blessed Be, all embracing) ("a" in this context = given forth from)
(Ah-ShA-Lum = the Still Point of Eternal Peace)
Thah-A'-Jha in'ta DO A"
(Tha-A-Jha inta= Spoken AS IT IS STATED) (DO A"= As SO It IS DONE)
Um Sha-DI' UR' A' ah Khum' TUn
(Um Sha- DI' UR'A = All Embracing Pillar of First Cause Light)
(ah Khum' TUn = I AS THIS expression now)

Teleportation Wingmakers Manual

Introduction
Teleportation is the ability of moving matter from one point in time and space to another point in time and space instantaneously. There are different types of teleportation available at this time both visual and physical teleportation are possible and we use visual teleportation as our starting point. The different types of visual teleportation are. visual body or astral body teleportation, or visual object teleportation but before we can master teleportation we must first learn the art of Telekinesis/Psychokinesis and Clairsentience and the higher state of consciousness the seventh sense. Other abilities useful in teleportation are Tele-Visualization and ESP (Extrasensory Perception). The ability of teleportation is used in Time travel, Interstellar travel, and Dimensional travel. Teleportation is the near instantaneous transport of the Psychic from one location to another and there is no other form of travel faster then teleportation
Teleporting An Visual Object

Visual Technique
Step One
Charge your physical body with energy.

Step Two
Then place an object in front of you.

Step Three
Next close your eyes and visualize the object in front of you.

Step Four
Then visualize the energy around the object and see your energy blending with the objects energy. This will look like the two energies will be knitting together

Step Five
Then focus and feel the object feel the object leaving and if you need to visualize that the object is disappearing

Step Six
Then focus your mind on the new destination six feet in front of you.

Step Seven
At the new destination six feet way in front of you and feel the object reappearing at the new position in front of you and if you need to visualize that the object is reappearing at the new destination. Then open your eyes.

Remember
This is a visual exercise so you are only teleporting the object visually.

Teleportation Portal

The image above is what a energy portal should look like

Introduction

A Portal is a wormhole of energy made within the light matrix of the physical plain. We can make Portal from the chakra energy, prana energy, and kundalini energy found at the base of the spine and merkaba energy found in the light matrix of the physical body and auric field. Portals can be used in Time travel, Interstellar travel, Dimensional travel and Blank Slate Technology.

Visual Technique
Step One
Charge your physical body with energy.

Step Two
Then close your eyes and see your visual body standing in front of you.

Step Three
Then construct a Portholes like you learnt in the section on Portal in front of your visual body.

Step Four
Then visualize that your energy is blending with your visual body.

Step Five
Next focus your mind on traveling a distance of one kilometre from where you are.

Step Six
Then see your visual body walking into the Portal.

Step Seven
Next see the Portal opening one kilometre away from where you started and see your visual body walking out of the Portal

Step Eight
Then walk back into the Portal and back out to your starting point.

Remember
Practice this exercise by expanding the distance but always make sure you can do this one before moving on.

Teleportation Portal

Visual Technique
Step One
Charge your physical body with energy.

Step Two
Then close your eyes and see your visual body standing in front of you.

Step Three
Then construct a Portholes like you learnt in the section on Portal in front of your visual body.

Step Four
Then visualize that your energy is blending with your visual body.

Step Five
Next focus your mind on traveling to the USA the destination is The white house, washington DC

Step Six
Then see your visual body walking into the Portal.

Step Seven
Next see the Portalopening in front of the white house and see your visual body walking out of the porthole

Step Eight
Then walk back into the Portal and back out to your starting point.

Teleportation Porthole

Visual Technique
Step One
Charge your physical body with energy.

Step Two
Then close your eyes and see your visual body standing in front of you.

Step Three
Then construct Portholes like you learnt in the section of Portal in front of your visual body.

Step Four
Then visualize that your energy is blending with your visual body.

Step Five
Next focus your mind on traveling to italy the destination is the coliseum in rome

Step Six
Then see your visual body walking into the Portal.

Step Seven
Next see the porthole opening in front of the coliseum and see your visual body walking out of the Portal

Step Eight
Then walk back into the Portal and back out to your starting point.

Teleportation Porthole

Visual Technique
Step One
Charge your physical body with energy.

Step Two
Then close your eyes and see your visual body standing in front of you.

Step Three
Then construct a Portholes like you learnt in the section on Portal in front of your visual body.

Step Four
Then visualize that your energy is blending with your visual body.

Step Five
Next focus your mind on traveling to egypt the destination is the great pyramids.

Step Six
Then see your visual body walking into the Portal.

Step Seven
Next see the Portal opening in front of the great pyramids and see your visual body walking out of the Portal

Step Eight
Then walk back into the Portal and back out to your starting point.

Teleporting The Visual Body
Beginners Technique

Visual Technique
Step One
Get comfortable and relax.

Step Two
Now close your eyes then charge your physical body with energy. Then see your visual body standing in front of you.

Step Three
Then visualize that your energy is blending with your visual body.

Step Four
Now focus your mind on teleporting to a destination of your choice.

Step Five
Then see your visual body teleporting to that destination and when you arrive look around and take note of your observation and check them when you come back.

Step Six
Then see your visual body teleporting back to the point in front of your physical body and in your own time open your eyes.

Remember
Then check with books or other information on that destination to see if your observations are correct. If they are then well done now try another destination.

Added 16/04/2005
Teleporting Your Visual Body
Advanced Technique

Visual Technique
Step One
I want you to get comfortable and relax.

Step two

Now close your eyes then charge your physical body with energy. Then see your visual body standing in front of you.

Step Three
Then visualize that your energy is blending with your visual body.

Step Four
Now here's were things change in the past i would have told you to spin the core star and become the light being this time I want you to focus on the hara line see it running through your visual body from the higher point above your head down through the crown chakra passing through the sixth and fifth chakra's into the soul seat down through the lower chakra's and the ten tien and out through the base chakra and down into the earth. Feel and see the visual body grounding with the earth see the tan tien changing from a reddish brown colour to a reddish colour like the earths core and feel this connection with the earth.

Step Five
Now focus your mind on teleporting to a destination of your choice.

Step Six
Now as we spin the core star i want you to stay focused on the hara line and this connection with the earth and in your own time spin the core star as you've done a thousand times before and this time pay attention

Step Seven
Then see your visual body teleporting to that destination and when you arrive at the destination return the visual body to normal and as you do focus on the hara line once again.

Step Eight
Then see your visual body and spin the core star once again and teleporting back to the point just in front of your physical body and return the visual body back to normal once again. And in your own time open your eyes.

Note
Now if you've done this technique you'll see what was missing when I tried teleporting in 1986. Write down every little detail of your observations during this technique and if you wish post your finding under the section Hara line Tech on the message board and we will Develop this technology together as a group. Because this technology benefits each and every one of us as a whole.

Remember
Master the visual technique before moving on to the physical technique it is important that you know the visual technique inside out before moving on.

Teleporting A Psi Ball
Beginners Technique
Introduction
Creating psi balls is the act of shaping energy with the mind. Energy is much easier to manipulate than physical objects, so this is a great beginners exercise.

Two Person Technique
Step One
Work out who will teleport for this exercise and who will be the receiver for this exercise.
Step Two
(Teleporter) Hold your hands out in front of you, with both palms facing. Close your eyes and relax. Concentrate on the space between your hands feel the energy ball between your hands. Try to strengthen this energy field by

visualizing the ball growing stronger. If you cannot feel the ball, move your hands away from each other, then closer and closer, trying to feel the point of energy. To me, this feels like a tingling sensation between my hands and in my third eye.

Step Three
(Teleporter) Now i want you to open your eyes But stay focused on the energy between your hands feel the psi ball.

Step Four
(Receiver) Now i what you to hold your hands out in front of you, with both palms facing.

Step Five
(Teleporter) Focus your mind on the point between the hands of the receiver and when you are ready feel the psi ball teleport to the new destination between the hands of the receiver and if you have done this right the receiver should be able to feel the psi ball between their hands. If you need to use visualization for this technique then you can but remember to feel the energy movement.

Physical Teleportation

Introduction
Physical teleportation is the near instantaneous transport of the Psychic from one location to another.

Because Teleportation is an extension of Clairsentient ability, there are two categories of Teleportation: Extensive Teleportation and Projective Teleportation. As the names imply, each is based on the type of Clairsentience which must be used to target the correct location.

Although Teleportation is nearly instantaneous, the Psychic may be disoriented upon arrival at the new location. During this period of disorientation, the Psychic suffers a penalty to all Perception/Awareness/Observation Trials, and may take no action other than Perception/Awareness/Observation Trials until the disorientation passes, or until the Psychic achieves a Superior Success or better on an Intellect (Perception) Trial. However, disorientation may continue even after a successful Perception Trial, and the penalty to Perception then applies to any action taken until the period of disorientation passes. Increasing degrees of success on the Teleportation Trial can reduce or eliminate disorientation.

Extensive Teleportation
is the ability to teleport one's body a short distance, overcoming intervening barriers to normal travel. In order to use Extensive Teleportation, the targeted location must be sensed via Sense Extension, or must be in physical sight of the Psychic. Teleportation is nearly instantaneous once activated, and can generally be accomplished in the same OP during which Clairsentience was used to locate the target. The Psychic remains subject to possible disorientation at the new location.

It is possible to use Extensive Teleportation to dodge an attack, so long as the teleportation trial is successful, and the new location is not in the path of an attack. However, the possibility of disorientation makes this a potentially dangerous technique that is normally attempted only as a last resort.

Using Extensive Teleportation requires the psychic to make a Fatigue Trial with a difficulty equal to the difficulty of the Extensive Teleportation Trial.

Projective Teleportation
Is the ability to teleport to any location the Psychic can see clearly via Sense Projection. Seeing clearly means the Psychic must have achieved a Superior Success or better on the Sense Projection Trial. The Psychic may apply a

-1 modifier to the difficulty of the Projective Teleportation Trial for each degree of success above Superior on the Sense Projection Trial.

Projective Teleportation requires a minimum of one OP to perform, in addition to whatever time and effort was required to sense the target location using Sense Projection.

Teleporting Physical Objects

Introduction
When it comes to teleporting physical object there are two types of teleportation the first is by blending your energy with the atoms of the object and the other form of teleportation is by using an electromagnet light chamber to teleport the object because light particles penetrate the atoms of the object and once the light particles Blend with the object then you can move the object but before you can move the object you build a electromagnet light energy generator

Added 05/11/2002
Teleporting Physical Objects
Physical Technique
Step One
Find a small object of your choice. (I used a teaspoon)

Step Two
Then hold the object in between your hands and hold your arms so that they are comfortable.

Step Three
Then i want you to sit quietly breathe deeply and get comfortable and relax.

Step Four
Empty your mind of all extra thoughts and remain clear minded and focused on the task at hand.

Step Five
With your eyes closed feel the energy of the spoon and blend your energy with the objects energy Feel around the object and knit your energy with the energy of the object remember feel it don't think about it. (Your hand may feel hot as your energy blends this is normal)

Step Six
Get into the flow of the object and feel the energy around the object and within the object. Once your energy has kintted with the object feel the force of the energy and stabalize it before you move to the next step.

Step Seven
Then visualize or feel the object leaving your hands in a word disappearing from between your hands and reappearing at the desired destination and as it leaves your hands open your eyes and see the object appearing at the destination before you

Remember
Remember NEVER apply force! You aren't there to physically force the object to move through the fabric of time and space. That's not point of the exercise.
(The thoughts and comments expressed in this exercise are from personal experience it was first done by Donna on the 05/11/2002 the object was a teaspoon)
You may experience problems with your destination don't worry it takes practice to learn to control it.

Meta-Physical Teleportation
Etheric Body : Lesson One

Visual Technique
Step One
Get comfortable and relax.

Step Two
Now close your eyes then charge your physical body with energy.
Then visualize the etheric layer of your auric body (the First Layer of the auric field).

Step Three
Then visualize that your energy is blending with the etheric layer of your auric field.

Step Four
Now focus your mind on teleporting the etheric body across the room. visualize the destination point.

Step Five
Then see your Etheric body teleporting to that destination and as you do this stretch out your feelings and feel the etheric body leaving the auric field when you arrive at the destination look back at your physical body and take note of your observation and feelings and when you are ready teleport your etheric body back into your auric field.

Step Six
Then in your own time open you eyes but be careful you may feel a little disoriented at first this is normal.

Step Seven
Then write down your findings on how you felt during this exercise and of what you seen during this exercise

Remember
Master this technique before trying to teleport the other layers of the auric field or Physical objects or the physical body.

A Method On Physical Teleportation
Teleportation 'the practice'

The first order of things that you must understand is that teleportation is a natural way of travel for all beings of light. Whether it be on the third dimensional or fourth, fifth, and higher dimensional existence. It is natural and very easy for any light being because it all deals with "light ". Different colors of light give you different level of awareness. Thus your experience will be varied. Also throughout each level of light expression.

Physical Teleportation
Abilities Used In Teleportation

Tele-visualization
Tele-visualization is the ability of 3 dimensional visualization with its dimensions being sound, shape, and form just like the picture that we can view with our own eyes in the physical reality and is the channel by which we can travel through interstellar space and time in the visual and physical forms.

Psychokinesis

Psychokinesis or PK is The ability of movement of an object or the physical body of a term coined by the famous researcher J. B. Rhine. It can be defined as the direct action of the mind on a physical object without the mediation of any known physical energy. Phenomenon of this kind have been reported from as far back as St Ben edict up to the modern day spoon bending of Uri Geller! There are various forms including Macro-PK where the influence is upon large objects and Micro-PK where the target is from the quantum world.

Clairsentience
Sometimes called "psychic feeling". There are two forms of clairsentience: empathy and psychic knowing. Clairsentience is the ability to "know" about someone or something without the use of the five senses for input. Clairsentience is probably the most common type of intuitive ability. (Often we refer to it as a Hunch, a Gut Reaction, or Instinct.)

Hara Line Development
Hara line development is designed to bring your awareness to the grounding and ungrounding process of the hara line in the visual techniques of time travel, dimensional travel, levitation, teleportation, and interstellar travel.

Added 16/04/2005
Energies Used In Teleportation
In teleportation we use two main energy systems. The first being the hara line and the second being Kundalini energy from the base of the spine. All our hidden Siddhi powers come from an awakened Kundalini. Examples of the Siddhi powers are supernatural abilities of strength, sight, hearing, levitation, bi-location, teleportation, and instant manifestations, etc... Basically anything you can imagine beyond the everyday world is possible when your Kundalini is awakened. The Siddhi powers are our natural abilities, yet through social conditioning they have been suppressed and we have been programmed to believe they are malicious or wrong. They are as natural as the warmth from the sun above. Rising Kundalini along the spinal column and moving it out through the chakra and blending it with the atoms of the physical body is what make teleportation possible. The hara line on the second hand is the anchor that grounds us to the physical and earth plains of existence it is here that we must first start with hara line development because teleportation will not work if we stay grounded to the earth all you will feel is the body shunting as it tries to jump through the fabric of time and space so we must learn the technique of grounding and ungrounding thoroughly before trying teleportation remember safety comes first and if you are not well trained in the grounding and ungrounding technique before trying the this teleportation technique because it could be fatal upon arrival at your new destination if you are not well trained.

Warning
This Technique is not for beginners and I would suggest you try this technique with your visual body first it is dangerous and could be fatal if not studied properly.

Added 16/04/2005
Physical Body Teleportation

Physical Technique

For preparation find a quiet place and a soft surface and lay down.

Step One
Now in your own time close your eyes and take a couple of deep breaths feel the chest rise and fall and relax.

Step two
Then visualize the base chakra. At the base of the chakra there is a well of energy, this well is know as the jewel of the lotus within the jewel there is a light golden energy. Visualize that this golden energy is trickling out of the

jewel and spreading out into the chakra. As the chakra spins see the chakra becoming golden in colour with its light radiant energy. Feel that energy.

Step Three
Now focus on the second chakra and do the same, see the jewel at the centre and see this radiant energy trickling out into the 2nd chakra fill it with its radiant energy.

Step Four
Then move on to the 3rd chakra and do the same for this chakra and the other chakras till all twelve chakras are filled with the radiant Golden wheels of Kundalini energy and feel how light this energy feels in your body.

Step Five
Now once you have reached the twelveth chakra see the twelven chakras and feel this light golden energy continually flowing into the chakras till it starts to over flow into the physical body do this for around about ten minutes.

Step six
Now as the golden energy begins to flow out into the physical body see the energy spreading through your entire body feel how light your body feels as the golden energy blends with your body and relax.

Step Seven
Then in your own time I want you to visualize the hara line and feel your connection with the earth and the physical plain and in your own time draw the hara line out of the earth and into the base chakra and hold it there.

Step Eight
Now as you do this you may feel your body becoming unstable this is normal because you are no longer linked with the earth and the physical plain

Step Nine
After a couple of minutes when your physical body is feeling more stable in this new plain beyond time and space see the three tan tiens and expand the energy of the tan tiens thought your being one by one starting with the lower tan tien and moving through the middle tan tien and through to the top tan tien and hold it there for a moment.

Step Ten
Now at this point you are ready for physical teleportation but before you can teleport you need a destination for this technique choose a destination close by like in the same room or a room close by see that destination.

Step Eleven
And in your own time feel your body leaving it's current position and teleporting to the new destination and as you arrive at the new destination drop the hara line back down into the earth. feel this connection with the earth and the physical plain and allow your physical body to stabilize in the physical plain. Focus on this for a couple of minutes

Take Note
You may feel disoriented at this point don't freak out this is normal after teleporting in time this will pass and as you get better at teleporting this disorientation will pass.

Step Twelve
Now relax and feel at one with all things and in your own time open your eyes and well done.

Warning

Practice this technique with your visual body first this technique is very dangerous and cloud be fatal if not done right. Remember this is not a race safety comes first so be careful. You can also do this technique in the lotus or half lotus position as well and if it makes it easier put this technique on a audio tape so you can play it back while your doing this technique but if you do this make sure you remember step eleven for when you arrive at the new destination.

Copyright © Wingmakers.co.nz - All rights reserved.

HOW TO SURVIVE IN A WORLD WITHOUT ANTIBIOTICS

A top MD shares safe alternatives that work, some better than antibiotics

by Keith Scott-Mumby MD, MB ChB, PhD

"The Alternative Doctor"

Title Quote

Put bluntly, medicine’s successes at vaccination and antibiotics treatment are trivial accomplishments relative to natural selection’s success at generating the immune system. Recognizing this fact has important repercussions for the long-term control of infectious diseases. We will probably obtain better disease control by figuring out how to further tweak the immune system and capitalize on its vastly superior abilities than by relying on some human intervention such as new antimicrobials (antibiotics, antivirals, or antoprotozoal agents).

Paul W. Ewald Plague Time

Disclaimer

All content within this digital book is commentary or opinion and is protected under Free Speech laws in all the civilized world. The information herein is provided for educational and entertainment purposes only. It is not intended as a substitute for professional advice of any kind.

In no event shall Professor Scott-Mumby be liable for any consequential damages arising out of any use of, or reliance on any content or materials contained herein, neither shall Professor Scott-Mumby be liable for any content of any external internet sites listed and services listed.

Always consult your own licensed medical practitioner if you are in any way concerned about your health. You must satisfy yourself of the validity of the professional qualifications of any health care provider you contact as a result of this publication.

Quotes And Facts

The mass of a single cell of the E coli bacterium is 665 femtograms. A femtogram is one-thousandth of a picogram, which is one-thousandth of a nanogram, which is a billionth of a gram.

Journal of Applied Physics.

……

Not only does the Earth contain more bacterial organisms than all others combined (scarcely surprising, given their minimal size and mass); not only do bacteria live in more places and work in a greater variety of metabolic ways; not only did bacteria alone constitute the first half of life's history, with no slackening in diversity thereafter; but also, and most surprisingly, total bacterial biomass (even at such minimal weight per cell) may exceed all the rest of life combined, even forest trees, once we include the subterranean populations as well.

Stephen Jay Gould, "Planet of the Bacteria," Washington Post Horizon, 1996, 119 (344): H1; Reprinted here with permission; This essay was adapted from Full House, New York: Harmony Books, 1996, pp. 175-192.

……

Bacteria inhabit effectively every place suitable for the existence of life. Mother told you, after all, that bacterial "germs" require constant vigilance to combat their ubiquity in every breath and every mouthful, and the vast majority of bacteria are benign or irrelevant to us, not harmful agents of disease. One fact will suffice: during the course of life, the number of E. coli in the gut of each human being far exceeds the total number of people that now live and have ever lived.

Stephen Jay Gould, "Planet of the Bacteria," Washington Post Horizon, 1996, 119 (344): H1; Reprinted here with permission; This essay was adapted from Full House, New York: Harmony Books, 1996, pp. 175-192.

……

Numerical estimates, admittedly imprecise, are a stock in trade of all popular writing on bacteria. The Encyclopaedia Britannica tells us that bacteria live by "billions in a gram of rich garden soil and millions in one drop of saliva." Human skin harbors some 100,000 microbes per square centimeter (note:

“microbes” in-cludes nonbacterial unicells, but the overwhelming majority of “microbes” are bacteria).

Writer Dorion Sagan and biologist Lynn Margulis, Garden of Microbial Delights

……

Fully 10 percent of our own dry body weight consists of bacteria, some of which, although they are not a congenital part of our bodies, we can’t live without.

Writer Dorion Sagan and biologist Lynn Margulis, Garden of Microbial Delights

……

We could not digest and absorb food properly without our gut “flora.” Grazing animals, cattle and their relatives, depend upon bacteria in their stomachs to digest grasses in the process of rumination. About 30 percent of atmospheric methane can be traced to the action of methanogenic bacteria in the guts of ruminants, largely released into the atmosphere—how else to say it—by belches and farts.

Stephen Jay Gould, “Planet of the Bacteria,” Washington Post Horizon, 1996, 119 (344): H1; Reprinted here with permission; This essay was adapted from Full House, New York: Harmony Books, 1996, pp. 175-192.

……

In another symbiosis essential to human agriculture, plants need nitrogen as an essential soil nutrient but cannot use the ubiquitous free nitrogen of our atmosphere. This nitrogen is “fixed,” or chemically converted into usable form, by the action of bacteria like Rhizobium, living symbiotically in bulbous growths on the roots of leguminous plants.

Stephen Jay Gould, “Planet of the Bacteria,” Washington Post Horizon, 1996, 119 (344): H1; Reprinted here with permission; This essay was adapted from Full House, New York: Harmony Books, 1996, pp. 175-192.

A group, led by microbiologist William. B. Whitman, estimates the total number of bacteria on Earth to be five million trillion trillion -- that's a five with 30 zeroes after it. Look at it this way. If each bacterium were a penny, the stack would reach a trillion light years. These almost incomprehensible numbers give only a sketch of the vast pervasiveness of bacteria in the natural world. "There simply hadn't been any estimates of the number of bacteria on Earth," said Whitman. "Because they are so diverse and important, we thought it made sense to get a picture of their magnitude."

The study was published in the June 1998 Proceedings of the National Academy of Sciences.

……

For the first half of geological time our ancestors were bacteria. Most creatures still are bacteria, and each one of our trillions of cells is a colony of bacteria.

Richard Dawkins

……

"Love: Before I heard the doctors tell The dangers of a kiss; I had considered kissing you. The nearest thing to bliss. But now I know biology and sit and sigh and moan; six million mad bacteria and I thought we were alone!"

Song lyrics

……

A probe sent two miles underground in a South American gold mine found bacteria living there. Their dinner? Radioactive emissions from the rocks around them. Above ground, the aptly named Deinococcus radiodurans can handle radiation exposure nearly 10,000 times the fatal dose for humans.

Somewhere in cyberspace...

……

Everything we do with food is an attempt to keep from being poisoned by our microbial competitors.

Professor Paul Sherman, Cornell University.

Contents

Terminology 1

MRSA... What's all the fuss? 2

The Fear Erupts 4
  So with the political frenzy disposed of, what do we know? 5
  Scare Talk? I don't think so 6

What are bacteria? 8
  What about viruses? 10
  What about parasites? 11
  Historic antibacterials 11
  The Love Triangle 12

Staphylococcus and Co. 14
  Beware of the hospitals! 15
  Pets Can Give You MRSA Too 16
  Community Active MRSA 18

The Danger Is Increasing 19
  C. difficile and membranous ulcerative colitis 20
  Treatment is difficult, wouldn't you know! 21

The Golden Age Is Over! 22
  Escape Pathogens 24
  No New Antiobiotics On The Way 25
  What the Rest of Us Can Do 25
  The Bottom Line 26
  My Advice 26
  What is important is to build good health; overall health. 27

Part 1

Understanding The Enemy

Our fight with bacteria 30
- It's not just hygiene 32

Killing Bacteria Has A Downside 34
- Metagenomics May Give Us The Edge 36

The rise of the superbugs 38
- All those romantic Victorian novels 39

Agribusiness is to blame 42

Turning back the clock 44
- Teeth Infections 44
- One day, everything changed. 45
- We now have antibiotic resistant gonorrhea. 46

Part 2

What Can You Do To Kill Bacteria?

Plain water 49
- Antibiotic Soap? 50
- Kids need clean hands, too 52

Drawing Salve and Ointments 53

Blue light 55
- Get a UV Light Nano To Protect Yourself Against Dangerous Bacteria 56

Manuka Honey! 58
- Propolis 59

Stolle Milk 62

Colostrum 67

Marine Quinton 69
- Enter Messieur Rene Quinton... 70
- Properties of QMP 71

How To Take It 72

Chlorine Dioxide Protocol 73
The Process 75
A Pathogen Terminator 76
A Journey into Chemical Alchemy 76
Nausea 77
The Procedure 78

Hydrogen Peroxide 82
A Bit Of Biochemistry 83
A Little History 84
What Diseases Can Be Treated With Hydrogen Peroxide? 86
How Is Peroxide Used? 86
Grades of Hydrogen Peroxide 87
Other Administration Routes For Hydrogen Peroxide 89

Hydrochloric acid Injections 91

Colloidal Silver 94
Safety Of Colloidal Silver 95
Overdose (Argyria) 96

Nutritional factors 98
Nutrition bullets 100

Vitamin A 101

Vitamin C 102
Polio Teasers 102
Blood Levels 104
Bacterial Diseases 105
Klenner Protocol 106
Streptococci and Staphylococci 107
Administration 109
Vitamin C vs. Ascorbic Acid 110

Vitamin D 111
Vitamin D Cream 112

Iodine 114
Test For Iodine Deficiency 114

Zinc 116
Supplements 117
Health Risks from Excessive Zinc 117

Iron Overload 119

What About Herbs? 121

Calendula 124
Making Calendula tea: 125

Garlic 126
Use Garlic Tea as an Antibiotic 127

Artemisia annua 128
Better Than Garlic? I Think So! 128
Scientific Interest 129
Malaria 129
Other Pathogens 130
How To Take It 131

Samento (Uncaria tomentosa) 132

Tea Tree (Melaleuca alternifolia) 134

Aloe Vera 137
Dosage 138

Green Tea 139
White Tea 140

Spices Are Hot Stuff! 141

Mushrooms and Fungi 143
Reishi 144
Maitake 145
Shiitake 145
Inonotus obliquus (Chaga) 146
Pleurotus species 146
Trametes versicolor (Turkey Tail) 146
It just remains to enter a few words of caution: 147

Essential oils (aromatherapy) 149
Scientific Testing or Aromatherapy Oils 150
Oregano 151

Cinnamon 153
Cassia 153
Thyme 154
Black Seed Oil 155
Mountain Savory 156
Pregnancy and Strep B 158
Lavender 159
Geranium 160

Competitive Inhibition 162
Probiotics 163
Symbiotics 164
This Is So Good The Veterinarians Are Onto It 166
Probiotics directly help the effects of antibiotics! 167
How To Take Probiotics 167
Human Probiotic Therapy 168
Bacteriophages 169
A Historical Note 169
Pros and Cons Of Phage Therapy 170
Biofilm, no not a Bond movie 171

Homeopathy 172
Now a New Science Altogether 172
Background 173
Homeopathy in 3 bullet points 174
Homoepathy In Acute Infections 176
General Remedies To Look Out For 177
Specific Remedies Or Nosodes 178

Homotoxicology 180
Also know as comlex homeopathy and German homeopathy. 180
Different formulations 180
What To Take 182
Where To Buy 183

Biochemic Tissue Salts 185
Also known as Schussler Salts 185
Modern Tablet Form 185

The Russian SCENAR 186
Clinical Aspects 187

UV Light As An Antibiotic 188
History 189
Procedure 190

  - What Conditions Can Be Benefited? 190
- Cold Lasers 192
- Nano UV Zapper 194
- What Goes In The Medicine Cupboard? 195

Appendix

- The official FDA version 198
  - Why has this happened? 199
  - Survival of the Fittest 199
  - Bacterial Weaponry 200
  - How Antibiotic Resistance Happens 200
  - A Vicious Cycle: More Infections and Antibiotic Overuse 201
  - Towards Solving the Problem 203
  - Targeting TB 205
  - Narrowing the Spectrum 205
  - Big Difference 206
  - Vancomycin Resistance Crisis 206

Terminology

Let’s just go over a few terms so that when you meet them you will be clear about the meaning.

Antibiotic means a substance which will KILL bacteria and similar organisms (but not viruses).

Viricidal. A substance which will kill viruses (but not bacteria).

Bacteriostatic means a substance which will handicap bacteria, by preventing them growing and multiplying. It means does not actually kill bacteria but holds them in check, while the person’s immune system gets to work, mopping up the invader.

Antiseptic. A substance (or process, like UV light) which kills all living organisms on a surface. In reality most antiseptics will kill 95% or more. But nothing is ever 100%.

Hygienic. Germs reduced to a level where the body is not at risk.

Sterile. No surviving viable germs on any kind.

MRSA… What's all the fuss?

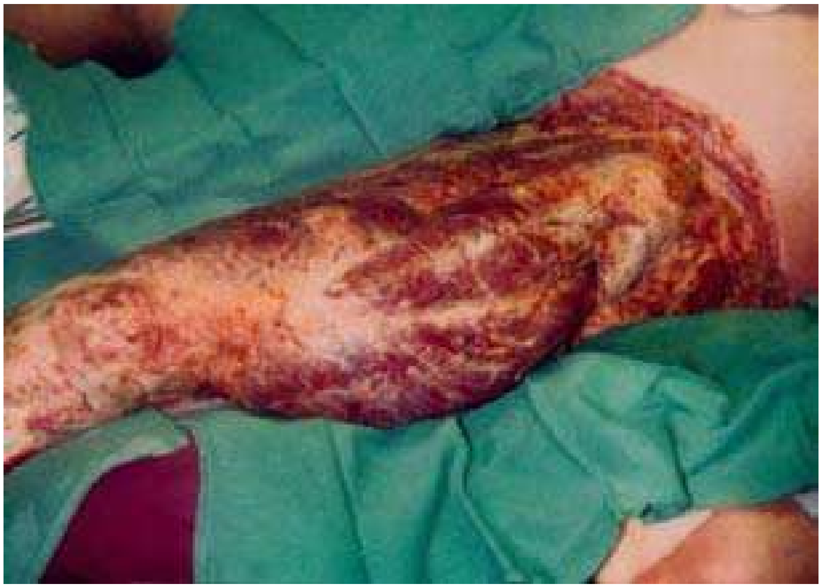

(this began as a kick to his ankle, while playing indoor soccer; on the second day he was rushed to ER and almost died)

MRSA (pronounced "mersa") is an abbreviation for "methicillin-resistant Staphylococcus aureus", the hospital superbug. Staph. aureus is a common bacteria but certain strains have become resistant to almost all known antibiotics, including powerful modern penicillins (such as methicillin). Today, methicillin is no longer made and has been replaced by newer variants, such as oxacillin, flucloxacillin and dicloxacillin.

It has raised extreme concern, not because of its own damaging powers (which are not inconsiderable), but because it is seen as the flagship of worse to come. As you will read, antibiotic-resistant bacteria have been around almost from the beginning of the "antibiotic era". But something about the MRSA problem makes it clear that we have to the end of a golden age of healthcare, where infectious disease were seen to be only an occasional or trivial problem. Those who chanced to die of an infection were seen as just "unlucky".

We no longer fear pneumonia, bacterial endocarditis, brain abcesses, and other terrible conditions that swept away our loved ones in a matter of days. We have become complacent. It's been a great period in which to live. I was born in 1945, just as antibiotics hit the market. We lost nobody in my family due to infections in the years since!

But we will soon return to the days when antibiotics are of little help. People will die in huge numbers, unless they take the trouble to learn the alternatives. The irony is, as I keep saying, that there are hundreds of alternatives. But arrogant colleagues sneer at these, as if they were just old wives' tales and superstitions.

In fact the real folly and superstition has been mainstream medicine, which has continued to believe in the silly fairy tale of endless life with antibiotics. I repeat: the threat of resistance has been there right from the start. They just ignored it.

Now you are lucky, you can re-learn about the old remedies, which really worked. But I’m going to give you the science too, lots of science and studies, which show this is no folklore but perfectly genuine and humble remedies, which are easy to self-administer and cost very little to obtain or make.

I’m sorry to do this, in a way, but I need to raise your concern level so that you understand how critical this is. Take a look at this picture, which is a man’s leg after surgery to clean up an infection of “flesh eating bacteria”. You can see how it got that name. Cutting away his infected skin and tissues was the only way to save his life, when antibiotics no longer work.

There’s worse still at the end of section 28. Please! I urge you, take the end of antibiotics as something very serious.

#1 The Fear Erupts

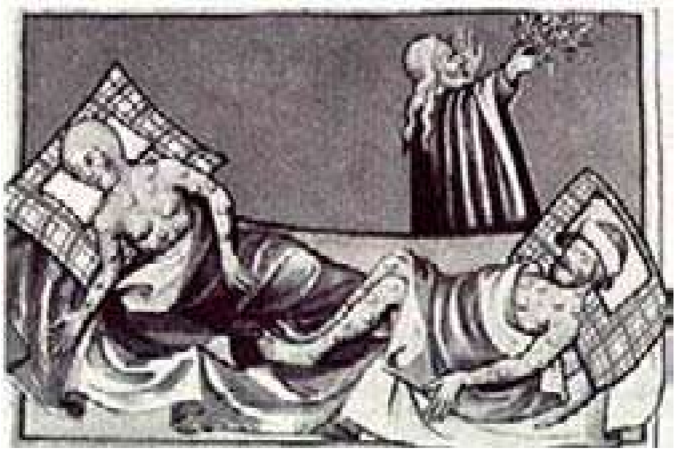

Illustration of the Black Death from the Toggenburg Bible (1411)

MRSA “fever” struck the country with the publication of a telling article in the Journal of the American Medical Association (JAMA 17th Oct 2007). That morning I was invited on a radio show to discuss the problem and get it into perspective. I had recently published an article likening the community active form of MRSA (as opposed to the hospital “superbug” version) to the Black Death from ages past.

One of the other guests, who I suppose I should not name, was being interviewed by several radio stations, including the one I was on. He blamed illegal immigrants, claiming it was their dirty habits and lack of personal responsibility that had caused the problem; he also blamed foreigners for tape worms, TB, Chagas’ disease and a whole host of other dangerous health conditions.

I followed him and declared his opinion “fascist, unscientific, irrational and inciting prejudice and hatred”. Fortunately, he had left to spread his poison somewhere else by the time I said this, otherwise a war of bitter words would have broken out.

No matter, I stand by my condemnation of this foolish and ignorant tirade. It’s akin to hysteria. The studio presenter, normally a very mild and civilized man, ended up raising his voice in shock at what “these people are doing to our country”.

This is how mob violence gets started.

The presenter only calmed down when I pointed out the stupidity of blaming foreigners, when Americans travel overseas and are just as capable of bringing back pathogens as newcomers! And what about legal immigrants? Does he suppose that bacteria have a way of knowing who’s legal and who isn’t?

I myself contracted Cholera while working on my 2004 Tsunami rescue mission to Sri Lanka and brought that to Pasadena (much to the shock of the public health officials):

http://www.scott-mumby.com/mission1.html

I was not an illegal alien.

As to the tapeworm, that was just pig ignorance (slight pun intended!) Parasites have been endemic in the US, especially the Deep South, for centuries. Transmission is by food and from pets, rarely by direct human contact. There are some bad beasties here, as you will know if you have visited my parasite page: http://www.alternative-doctor.com/allergies/parasites.htm

For example, 25- 30% of the population of the south-eastern USA, mainly children, have whip worm (Trichuris species) and my textbook sources are impeccable: *(Schmidt GD, Roberts LS, Foundations of Parasitology, Times Mirror/Mosby College Publishing, St. Louis, 1989, p. 418).*

Ironically, I might be warning healthy Europeans to take care when they come to the USA!

So with the political frenzy disposed of, what do we know?

MRSA was actually caused by doctors, not immigrants. It’s an inevitable result of overuse of antibiotics, followed by tougher strains of penicillin (like methicillin), used to try and clean up the stubborn bugs, and the eventual emergence of resistance even to methicillin.

It’s uncertain what the true incidence is because it is not at this time a notifiable disease. But the 2008 JAMA article worked by extrapolating from earlier 2005 figures and is, in effect, a mere guess. There were 5,287 invasive infections reported that year in the regions studied, which would translate to an estimated 94,360 cases nationally. There were 988 reported deaths among infected people in the study; around 20% and this was simply extrapolated up to 18,650.

IF THE STUDY FIGURES ARE CORRECT — and I emphasize they are only an intelligent guess, which I accept as such — then MRSA is killing more people than AIDS. 988 out of 94,360 cases gives a rate of 6.3 per 100,000. That would translate to 18,650 deaths annually (the researchers admit they don't even know if MRSA was the cause in all cases). AIDS killed an estimated 17,011 Americans in 2005.

Scare Talk? I don't think so

You might think this is scare talk' "Of course this thing will never get out of control". You are wrong. It is already out of control. The unanswered question is will it spread like wildfire, or just continue as a grumbling danger lurking below the surface?

Let's just hope that the pandemic model won't happen. As I said there are strong parallels with the Medieval "Black Death" but microbial disease is just not predictable. The important point to keep in mind is that pandemic infections, like any other infection, are NOT caused just by pathogenic microbes.

Infectious disease is caused by a lowered immune system. Period.

Healthy individuals, with a vibrant immune system and a positive mental attitude will not succumb to pathogens, no matter how vicious they are. I lived in a malarious country for years, with no prophylaxis, got bitten a dozen times a day by mosquitoes and caught nothing (there was even dengue fever present in our city).

We are in balance with Nature. She has invented NOTHING we cannot handle, if our defence mechanisms are working properly. Always remember that; it's important. But disease gets started because individuals are weakened by negative factors, such as overload, malnutrition, shock and stress. If we avoid these factors, or correct them quickly, there is no reason to fear disease.

Be healthy, that's your best protection. Remember my #1 motto in this: any good health measure is an anti-infection measure (same with cancer).

If you are struck down in a pandemic of any kind, you can forget the idea that the doctor will come along and help you. Even if the medical profession knew how to deal with such a problem, which it doesn't, there would simply not be enough doctors to go around. You will be on your own.

You need to read this report carefully and NEVER throw it away.

How many valiant men, how many fair ladies, breakfast with their kinfolk and the same night supped with their ancestors in the next world! The condition of the people was pitiable to behold. They sickened by the thousands daily, and died unattended and without help. Many died in the open street, others dying in their houses, made it known by the stench of their rotting bodies. Consecrated churchyards did not suffice for the burial of the vast multitude of bodies, which were heaped by the hundreds in vast trenches, like goods in a ships hold and covered with a little earth.

Giovanni Boccaccio

#2 What are bacteria?

rod forms

Bacteria are single-celled organisms. They come in a variety of shapes: round spheres (called a coccus, plural cocci), rod-shaped (called a bacillus, plural bacilli), spiral- shaped (called a spirochete, plural spirochetes) and a few odd-balls (no plural!)

Bacteria are the oldest living organisms on planet Earth. Since they don't produce "babies" (offspring), but reproduce by just splitting in two, you could say that bacteria represent an original founding organism which is now billions of years old!

Estimates of how many cells make up a human being vary—between 1 trillion and 100 trillion. I've seen both.

But what is clear is that there are 10 TIMES more bacteria in you and on you than the actual number of cells that go to make up YOU!

You are about 10% dry weight of bacteria: that's you, your body and your cells, minus the water in your tissues.

If you took away all the connective tisues of a human being, just leaving the cells, you'd have the equivalent of 9 human beings in your gut, so far as cell numbers were concerned; THAT'S how big a deal bacteria are for us.

Bacteria have only one single strand of DNA that encodes their nature. We humans have about 25,000- 30,000 genes that make up our humanness.

But there are over 100 times more bacterial genes, on you and in you, which all that matters. That's 3 million genes, set against your measly 1%. **What happens to your bacteria, happens to you. Trust me.**

We need lots of respect for those single-cell organisms that are all-but part of us. In fact bacteria keep us alive. The friendly ones act like body armour and prevent pathogenic invaders from destroying us. They keep the outer environment under control, so we stay healthy. They digest our food. They make certain vitamins for us.

Remember, the vast majority of bacteria are friendly or at least harmless. There are just a few troublemakers that create disease and death.

However, there is a real mystery. How do they do anything at all? I mean, there is no brain or nervous system. They don't have any obvious means of communicating or interacting. How do bacteria get organized to make somebody sick or healthy? They are so tiny and inconsequential, it seems unreasonable they could affect us at all, while acting individually.

Well, according to Professor Bonnie Bassler at Princeton there has been a breakthrough discovery. Bacteria, it seems, have a secret chemical conversation system. It's called "quorum sensing." As it turns out, every type of bacteria makes and secretes small molecules (a bit like hormones). When a bacterium is alone, these molecules have no effect (no receptors).

But when there's a large enough group of bacteria, these secreted molecules reach a critical level and suddenly all the bacteria begin to act as a synchronized group, based on the characteristic behavior programmed into the genes.

This was first discovered by Bassler, observing a marine bacterium called Vibrio fischeri. This little organism emits light. But when a suspension of bacteria is so dilute they are, in effect, isolated, nothing happens. Only when the bacterial concentration—and hence the level of quorum sensing signals—reaches a certain level, then all the bacteria "switch on" at once and begin to glow in the dark. It seems remarkable that tiny organelles, without any apparent sense organs, can suddenly tell who is in the neighborhood and how many neighbors there are!

Incidentally, they can talk inter-species too. A second quorum sensing mechanism, possessed of all bacteria and working interspecies, is what Bassler calls a "bacterial Esperanto".

The implications of this are devastating. Consider a pathogenic organism inside your body. If a few bacteria released their virulence, it would have no effect. You are at least a trillion times bigger than they are.

But they grow and multiply and when their quorum sensing tells them there are enough cronies present, they all launch their virulence attack at the same time. This way they are able to suddenly overwhelm an enormous host and bring it down.

Neat huh? But not good if you are the victim.

However this remarkable discovery by Bassler has two important consequences, both with enormous medical implications.

A whole new type of antibiotics can be looked for, which will effectively block the quorum sensing molecules (like jamming the enemy radar).

Health can be built, helping to block out bad bacteria, by supporting and provoking the chemical sensing of the good, friendly bacteria that support us!

Watch out, this is the way it's going to go.

Meantime, we have to worry about protecting ourselves.

What about viruses?

Viruses, are dangerous and can cause many infectious diseases, some of which can be fatal to humans. Smallpox, for example. However, viruses are not the same as bacteria. While a bacterium is a single celled organism, a virus is not even a single cell; technically it's the equivalent of part of a nucleus—just a package of DNA, wrapped up in a protein envelope.

The virus cannot live and cannot reproduce, unless it breaks into a cell and hijacks the cell's reproductive mechanism. It does this by taking over the DNA of the cell and forcing the cell to replicate its own viral DNA. This results in making lots of new viruses, but the cell will usually die in this process, releasing the new viruses to go and infect more cells. So the disease spreads.

The trouble is, viruses are not killed by antibiotics so, although they may be very dangerous, they don't come under the material covered by this report. Only a few anti-viral substances are available, and none which are fully safe. But that's a different story.

What about parasites?

Parasites are different again. They may cause disease and can be either single celled or multicellular animals. A worm, for example is multi-celled; as well as mechanisms for digestion, it has muscular fibers for movement.

Segmented worms, like tapeworms, are certainly multicellular, but each segment is more or less identical to the segment above and segment below. So parasites of this type work a bit like a duplicating or photocopying machine, breaking off new pieces of identical infected tissue and eggs. It's more like a colony than an animal, really.

But we also have single cell parasites such as Giardia lamblia and malaria.

Yes, malaria is a parasitic disease; the killer organism of the species Falciparum invades the blood and reproduces itself destroying blood cells in the process. The patient will sometimes die of too much blood loss as much as the severe fever which can ensue.

Note that the infamous mosquito is only the spreader of the malaria parasite. It does not otherwise cause malaria and you may be bitten by a mosquito and not get malaria. Conversely, you cannot get malaria unless you are bitten by an infected mosquito. The mosquito is an essential intermediate we call a "vector".

There are many treatments for parasites depending on the nature of the organism concerned. None of these fall within the definition of antibiotic, and so will not be considered in this report.

Where there is a clear vector, such as in the case of malaria, wiping out the vector is an obvious way to block transmission of the disease. However trying to eliminate mosquitoes has met with only very limited success. That doesn't seem to be the way to go.

Historic antibacterials

Many cures for infectious diseases existed prior to the beginning of the twentieth century which were based on medicinal folklore. That didn't mean they did not work; just that doctors weren't interested in them!

Cures for infection in ancient Chinese medicine, using plants with antibiotic-like properties, began to be described over 2,500 years ago. Many other ancient cultures, including the ancient Egyptians, ancient Greeks and medieval Arabs already used molds and plants to treat infections.

Cinchona bark (quinine) was a widely effective treatment of malaria in the 17th century, the disease caused by protozoan parasites of the genus Plasmodium.

According to legend, the first European ever to be cured from malaria fever was the wife of the Spanish Viceroy, the countess of Chinchon (hence the name).

The court physician was summoned and urged to save the countess from the waves of fever and chill that were threatening her life, but every effort failed to relieve her. At last the physician administered some medicine which he had obtained from the local Indians, who had been using it for similar syndromes. The countess survived the malarial attack and reportedly brought the cinchona bark back with her when she returned to Europe in the 1640s.

The birth of homeopathy (see section #41) was based on quinine testing. The founder of homeopathy, Dr. Samuel Hahnemann, when translating the Cullen's Materia medica, noticed that Dr. Cullen wrote that quinine cures malaria and can also produce malaria. Dr. Hahnemann took daily a large non-homeopathic dose of quinine bark. After two weeks, he said he felt malaria-like symptoms. This idea of "like cures like" was the starting point of his writings on "Homeopathy".

The Love Triangle

Gonorrhea, together with syphilis and HIV, make a "love triangle" of the most famous sexually transmitted diseases. Indeed, gonorrhea, caused by the Neisseria gonorrhoeae bacterium, affects 62 million people, aged mainly 15 to 29, and represents world's second most widespread STDs after the infection with Chlamydia trachomatis (930,000 cases of Chlamydial infection and 360,000 of gonorrhea were assessed in the United States in 2004).

But the power of gonorrhea is what really have shocked the scientists (as revealed in a new research published in PLoS Biology): relative to its weight, the Neisseria bacterium is the strongest organism on Earth: it can pull up to 100,000 times its body weight (could you drag 7,000 tons?).

Many bacteria move around using up to 10 contractile filaments named pili. The pili can be 10 times longer than the bacterium itself. Neisseria bacteria employ pili, but no one has seen before that these bacteria can join the force of their pili for achieving long, strong pulls.

The team led by Michael Sheetz at Columbia University in New York

#3 Staphylococcus and Co.

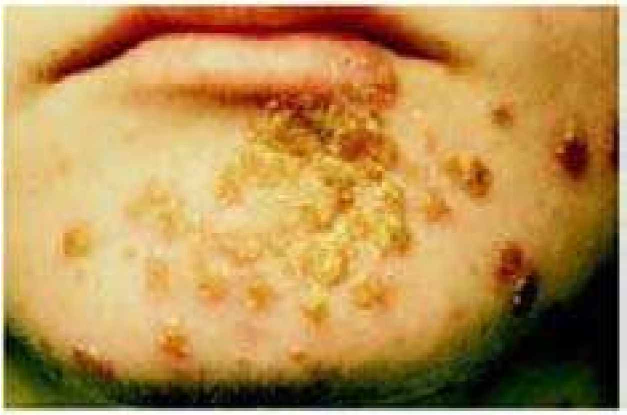

Now let's get back to bacterial disease and the problems we face trying to eliminate them.

I want to emphasize that Staphylococcus infections have been with humankind since forever. Most infections affect the skin and are mild. S. aureus was first identified in Aberdeen, Scotland, in 1880 by the surgeon Sir Alexander Ogston in pus from surgical abscesses [Classics In Infectious Diseases. "On Abscesses". Alexander Ogston (1844-1929)].

The name Staphylococcus aureus (abbreviated to S. aureus or Staph aureus in medical literature), literally means the "golden cluster seed" or "seed gold" . It is named on account of its golden appearance when it grows in colonies. You have probably seen the color too, in the golden-yellow crusts that form on the face, around the mouth, called impetigo.

About 20% of the population are long-term carriers of S. aureus. It can cause a range of illnesses from minor skin infections, such as pimples, impetigo, boils, cellulitis, folliculitis, carbuncles and abscesses, to life-threatening diseases such as pneumonia, meningitis, osteomyelitis, endocarditis, Toxic shock syndrome (TSS), and septicemia.

The new antibiotic resistant MRSA monster has been created by doctors, not by Nature. For years it has been the feared hospital bug that no-one talks about, because it's something given to you by doctors and nurses when you enter their environment. A survey earlier this year suggested that MRSA infections, including noninvasive mild forms, affect as many as 5 percent of hospital patients. Nobody wants suing; hence the secrecy.

Nobody likes to be a bringer of bad news but it is clearly going to get worse. More deaths. Hospitals are becoming a horribly unsafe place to be.

Beware of the hospitals!

Hospitals are just about the most UN-healthy places a person can find themselves. The very people who are looking after you, may be giving you a very unpleasant disease. One in every 20 healthcare workers carries methicillin-resistant Staphylococcus aureus (MRSA), researchers in Switzerland said.

hospital contracted MRSA

But the vast majority is without symptoms and only 5% have full-blown clinical infections, according to Stephan Harbarth, M.D., of the University Hospitals of Geneva, and Werner Albrich, M.D., of University Hospital Bern.

One implication is that screening efforts aimed at symptomatic infections are likely to miss a large proportion of colonized healthcare workers who might transmit the bacteria, according to a review in the May 2008 issue of Lancet Infectious Diseases.

Instead, the study authors said, “aggressive screening and eradication policies” should be used in an outbreak and in situations where MRSA has not reached highly endemic levels. The researchers looked at 127 studies published from January 1980 through March 2006, to see how likely healthcare workers are to be infected or colonized by MRSA and to assess their role in MRSA transmission.

On the basis of the published evidence, the study said, healthcare workers are usually vectors, rather than the main sources of MRSA transmission, implying that “good hand hygiene practices remain essential to control the spread of MRSA.”

Among 33,318 workers screened in the studies, 4.6% on average were carrying MRSA, the researchers found -- usually in the nose, although other sites were found. Most (94.9%) had no symptoms.

Risk factors included chronic skin diseases, poor hygiene practices, and having worked in countries with endemic MRSA.

Research revealed 18 studies with proven, and 26 studies with likely, transmission to patients from healthcare workers who were not clinically infected with MRSA.

“Staphylococcal dispersal is mainly dependent on whether the person is a nasal carrier,” researchers said, so that “screening of infected healthcare workers only will likely miss a large number of asymptomatic personnel capable of transmitting MRSA to patients.”

So beware!

Source reference: Albrich WC, Harbarth S “Health-care workers: Source, vector, or victim of MRSA?” Lancet Infect Dis 2008; 8: 289-301.

Pets Can Give You MRSA Too

Pet owners are blithely unaware that life-threatening pathogens can be picked up from their beloved pooch or pussycat.

According to Dr. Richard Oehler, of the University of South Florida College of Medicine in Tampa, and colleagues writing in the July 2009 issue of The Lancet Infectious Diseases.

Dog and cat bites account for about 1% of emergency department visits each year in the United States and Europe. Severe infections occur in about 20% of all cases and are caused by bacteria from the animal’s mouth, plus possibly other bacteria from the human patient’s skin.

About one in every 10 dogs or cats carries S. aureus. About a third (35%) of those were found to carry the MRSA strain, according to a study done at the University of Pennsylvania Veterinary School.

Of course infections travel both ways. More human cases of community-acquired MRSA leads to more MRSA colonization in domestic animals.

In fact, the study found that in some cases, pets may be picking the pathogen up from humans. “A growing body of clinical evidence has documented MRSA colonization in domestic animals, often implying direct acquisition of S. aureus infection from their human owners,” the team noted. MRSA-related skin infections in pets can then easily spread back to humans, so your pet may pose considerable danger.

Sepsis, a potentially fatal bloodstream infection, can be a complication following bite wounds from a pet infected with MRSA or other types of bacteria.

Animals bites are a major cause of injury in the USA and Europe each year, particularly in children. Bites to the hands, forearms, neck and head are usually the worst.

Treatment of MRSA infections in pets is similar to that used in humans and for the most part, the danger to pet owners -- even those most vulnerable to MRSA -- remains low.

SOURCE: The Lancet Infectious Diseases, news release, June 21, 2009 The worry now is that MRSA has spread beyond hospitals into the community. It's fancy name is CA-MRSA (community-associated methicillin-resistant staphylococcus aureus). In recent years, the resistant germ has been spreading through prisons, gyms and school locker rooms, mostly in poor urban neighborhoods.

An especially dangerous form is brought about by inhalation; the bacteria goes straight to the lungs, produces a sudden and violent pneumonia and the victim can die WITHIN HOURS, coughing up frothy blood and unable to breathe.

This is why I likened it to the Black Death (Bubonic Plague or just "The Plague"). The so-called respiratory form of Plague in olden times was also very sudden; it often killed the same day as it was inhaled. Remember the kid's nursery rhyme "Tissue, aTissue, We all Fall down" (dead)? That was said to commemorate respiratory Bubonic plague.

Suddenly, queues and public gatherings are potentially dangerous places. You are just as likely to catch this modern killer in a queue at the supermarket, from some young all-American sports youth or cheerleader, as you are from a foreigner, so please do not start or join any aversion campaign against "dirty immigrants". It's ignorant and could cost YOU your life by having you focus on the wrong thing.

Community Active MRSA

The worry now is that MRSA has spread beyond hospitals into the community. It's fancy name is CA-MRSA (community-associated methicillin-resistant staphylococcus aureus). In recent years, the resistant germ has been spreading through prisons, gyms and school locker rooms, mostly in poor urban neighborhoods.

An especially dangerous form is brought about by inhalation; the bacteria goes straight to the lungs, produces a sudden and violent pneumonia and the victim can die WITHIN HOURS, coughing up frothy blood and unable to breathe.

This is why I likened it to the Black Death (Bubonic Plague or just "The Plague"). The so-called respiratory form of Plague in olden times was also very sudden; it often killed the same day as it was inhaled. Remember the kid's nursery rhyme "Tissue, aTissue, We all Fall down" (dead)? That was said to commemorate respiratory Bubonic plague.

Suddenly, queues and public gatherings are potentially dangerous places. You are just as likely to catch this modern killer in a queue at the supermarket, from some young all-American sports youth or cheerleader, as you are from a foreigner, so please do not start or join any aversion campaign against "dirty immigrants". It's ignorant and could cost YOU your life by having you focus on the wrong thing.

#4 The Danger Is Increasing

Exactly as I predicted in the first edition of this book, things are falling apart rapidly. We are just a few years from total meltdown.

In 2009 a new strain of MRSA arose which is more deadly than anything we have seen previously. According to a report (November 1st 2009) this new strain of MRSA is five times more lethal than what has gone before. The mortality is 50%, compared to 11% for the "normal" MRSA

[Henry Ford Hospital study presented at the 47th annual meeting of the Infectious Diseases Society of America Oct. 29-Nov.1 in Philadelphia].

Why aren't the newspapers and TV programs screaming about this?

Bear in mind, the problem here is not just MRSA. Now we have Vancomycin resistant enterococcus (VRE). That's much more to be feared: if it reaches the septicemia stage it is 100% fatal. Nothing can stop it (Vancomycin is known as the "last defence" antibiotic).

Our old killer enemy TB has now morphed into a new strain which is resistant to ALL KNOWN ANTIBIOTIC DRUGS.

Also we now have PRSP (penicillin-resistant Streptococcus pneumoniae). It is reported that out of 100,000 hospitalizations for pneumonia, 40% are now due to this organism.

Then there is GISA. Never heard of that one?

In December 2002 the microbiology department at Monklands Hospital, Scotland isolated only the third Glycopeptide Intermediate Staph Aureus (GISA) in the United Kingdom.

The press were involved due to the rarity of this organism and the fact that this infection had a fatal outcome for the patient (the first fatal case in the UK).

After a press conference, Dr Alistair Leanord, consultant microbiologist, invited TV crews into the microbiology department to film biomedical scientist staff using the Vitek Autoanalyser, which was used to initially detect the GISA organism.

The problem is, GISA is not susceptible to vancomycin or teicoplanin. Vancomycin and teicoplanin are considered the last line of defence against

MRSA. However, fortunately, all GISA isolates to date have been susceptible to other antimicrobial drugs that have been effective in treating cases.

But what happens when that lucky state of affairs runs out?

C. difficile and membranous ulcerative colitis

You may have heard of C. difficile (or C. dif and various other appelations). Its real name is Clostridium difficile and that says it all. Most of the Clostridia species are nasty and cause diseases like botulism and gas gangrene.
Diff causes membranous ulcerative colitis, a severe and often fatal bowel inflammation.

Dif comes about from overuse and abuse of antibiotics. It's just another example of where good turns bad; the wonders of antibiotics become a deadly liability.

In recent years, C. difficile infections have become more frequent, more severe and more difficult to treat. he European Center for Disease Prevention and Control recommend that fluoroquinolones and the antibiotic clindamycin be avoided in clinical practice due to their high association with subsequent Clostridium difficile infections.

The emergence of a new, highly toxic strain of C. difficile, resistant to fluoroquinolone antibiotics, such as ciprofloxacin (Cipro) and levofloxacin (Levaquin), said to be causing geographically dispersed outbreaks in North America was reported in 2005. The Centers for Disease Control in Atlanta has also warned of the emergence of an epidemic strain with increased virulence, antibiotic resistance, or both *[McDonald L (2005). "Clostridium difficile: responding to a new threat from an old enemy" (PDF). Infect Control Hosp Epidemiol 26 (8): 672–5. doi:10.1086/502600. PMID 16156321].*

When severe, symptoms include watery diarrhea 10 to 15 times a day, abdominal cramping and pain, fever, blood or pus in the stool, nausea, dehydration and loss of weight.

Some people may be symptom free carriers (as with amebic dysentry).

Treatment is difficult, wouldn’t you know!

Antidiarrheals are strictly contraindicated because a life-threatening dilation of the bowel called toxic megacolon may result.

Different antibiotics may be called for. Treatment with oral vancomycin or parenteral metronidazole usually will result in abatement of symptoms within 3 to 5 days.

This is certainly one puppy that isn’t going to yield to yoghourt and roughage!

But unless the patient needs hospitalizing for support, chlorine dioxide therapy (section #18) should be tried; hydrogen peroxide (section #19); or colloidal silver (section #21).

Probiotics are essential. The orthodox treatment of choice is rapidly becoming a human faecal transplant (see section 40b).

A SCENAR (section #44) and homeopathics (section #41), would bring in useful energetic support. Try Cantharis, Colchicum, Colocynthis, Arg.Nit., Brom., Lach.

Aloe vera is the best herb to try (section #35).

#5 The Golden Age Is Over!

> In an article which appeared in Medscape, 2/3/2010, Dr Paul Auwaerter, MD, MBA, Clinical Director, Division of Infectious Diseases at Johns Hopkins University School of Medicine, was quoted as saying that, "The patients are sicker than ever before, so we are using antibiotics more intensively, and the bacteria are changing in response. At the same time we have really lost ground in incentive mechanisms for creation and production of new antimicrobial compounds.
>
> *Emerging Antibiotics: Will We Have What We Need? Laura A. Stokowski, RN, MS*

Unfortunately, the general public are still not getting this problem of antibiotic resistance or the absence of effective new antibiotics. People generally have faith or have been lulled into believing that medical scientists can develop effective new antibiotics whenever needed because they have always done so in the past.

By the time most people wake up to the realities of the situation, it will be too late. Antibiotic-resistant infections are becoming the next great equalizer, and this is not just a problem for the elderly or the immune-suppressed. Friends and family, rich and poor alike, will succumb to infections that should be curable but aren't, and everyone will be looking around for someone to blame.

The many factors that have contributed to the current crisis have already been debated, but these sobering facts remain:

- More US patients die of MRSA infections than HIV/AIDS and tuberculosis combined.

- Only 2 new antibiotics -- doripenem and telavancin -- have been approved in the past 3 years.

- We have no drugs to treat infections with some strains of multi-drug-resistant gram-negative bacilli, like Pseudomonas aeruginosa and Actinobacter baumannii.

We may finally have arrived at the era of the untreatable bacterial infection.

Livermore DM. Has the era of untreatable infections arrived? J Antimicrob Chemother. 2009;64(Suppl 1):29-36.

Mechanisms are varied

Resistance can be either inherent -- as exemplified by the inability of vancomycin to penetrate the cell wall of gram-negative bacteria -- or acquired. Acquired resistance is a change in the bacterium's genetic composition that permits clinical resistance to drugs that were once active against it. Acquired resistance can reduce the effectiveness of an antibiotic or render the antibiotic completely ineffective against the bacterium.

Bacteria can also become resistant to other classes of antibiotics (cross-resistance) or transfer their resistance genes to other microbes and species (co-resistance). The strategies used by bacteria to resist the actions of antibiotics include:

- Reduced outer membrane permeability;
- Reduced cytoplasmic membrane transport;
- Increased efflux/decreased influx of antibiotic;
- Neutralization of antibiotic by enzymes;
- Target modification; and
- Target elimination

SOURCE: Chen LF, Chopra T, Kaye KS. Pathogens resistant to antibacterial agents. Infect Dis Clin N Am. 2009;23:817-845

Escape Pathogens

(usually labeled ESKAPE)

These are pathogens of highest concern; the most serious, life-threatening infections as classified by the Infectious Diseases Society of America (IDSA), so-called because they effectively escape the effects of antibacterial drugs.

These include:

- E Enterococcus faecium Third most common cause of health-care associated blood stream infections. Increasing resistance to vancomycin
- S Staphylococcus aureus (MRSA) Emerging resistance to current drugs and significant
- drug toxicities. Lack of oral agents for step-down therapy
- K Klebsiella
- Escherichia coli
- K pneumoniae Extended beta-lactamase producing organisms increasing in frequency and severity; associated with increasing mortality. K pneumoniae carbapenemases causing severe infections in LTCF. Few active agents; nothing in development
- A Acinetobacter baumannii Increasing worldwide, recent surge reported in hospitals. Very high mortality. Carbapenem-resistant
- P Pseudomonas aeruginosa Increasing P. aeruginosa infections in US and worldwide. Resistant to carbapenems, quinolones, aminoglycosides
- E Enterobacter species MDR HCA infections increasing; resistance via ESBLs, carbapenemases, and cephalosporinases

(**key:** HCA = healthcare associated; BSI = bloodstream infection; MRSA = methicillin resistant S aureus; ESBL = extended-spectrum beta-lactamase; LTCF = long-term care facility; MDR = multiple drug-resistant)

No New Antiobiotics On The Way

The trouble is, according to Brad Spellberg, MD, Associate Professor of Medicine, Geffen School of Medicine at UCLA, “The economic and regulatory climates have changed so that the drug companies aren’t making new antibiotics. We’ve been talking about antibiotic stewardship since the 1950s. It used to be ‘new bug, new drug,’ but not anymore. We are already seeing infections resistant to all of the antibiotics that we have now, and the number will increase at a geometric rate over the next 5 years.”

It’s not that there aren’t any antibiotics in development. There are, but they won’t solve our real problems. The antibiotics aimed at gram-negative infections that are currently in development have mechanisms of action similar to the drugs we already have available. Ceftobiprole and ceftaroline, for example, have both completed phase 3 trials, but they are essentially the same as cefepime, a drug already on the market. A gram-negative organism resistant to cefepime, then, will also be resistant to both ceftobiprole and ceftaroline. So, explains Dr. Spellberg, the real problem is that there are no gram-negative antibiotics in the pipeline that will work against bacteria already resistant to the drugs we have. The number of drugs that have made it through the developmental process and received FDA approval has plummeted in recent years. From 1983 through 2007, systemic antibacterial approvals declined by 75%. *[Boucher HW, Talbot GH, Bradley JS, et al. Bad bugs, no drugs, no ESKAPE! An update from the Infectious Diseases Society of America. Clin Infect Dis. 2009;48:1-12. Abstract].*

What the Rest of Us Can Do

Dr. Paul Auwaerter has 2 suggestions: First, education and awareness of this issue must increase among the medical community. Second, political, and scientific forces must work together to recognize the scope of the problem and develop solutions. “We learned a lesson from the H1N1 pandemic: the importance of being prepared. Right now, we are not close to being adequately prepared.”

We need to realize that it isn’t just our patients who are at risk, it’s all of us. But nothing will change if we sit back and think this problem is too big, too complex, that it’s someone else’s problem, or that there is nothing we can do. There is something we must do, according to Dr. Spellberg. “Only when constituents start putting pressure on elected officials will things change.”

“We need [physicians’ and nurses’] voices because right now the FDA is dominated by statisticians and the clinicians are being shouted down. We have lost our voices. They are making drug approval decisions based on statistical results and have lost all sight of clinical reality.”

Dr. Spellberg recommends writing letters to the editors of professional journals or other media, expressing concern -- and yes, outrage -- about the current situation. People can also inform regulatory officials and politicians that the current approach to approving drugs is unacceptable; that physicians are being excluded from these conversations, and that statisticians should not have the sole voice. Demand that the FDA stop insisting on perfection and demand that clinicians have a say in the approval process. DO YOUR PART: read this important information manual and raise your voice. Only by wide public action will anything change. The dangers are real; they are here NOW and are not (ever) going to go away. So take immediate action.

The Bottom Line

“As the Medscape article concluded: We are literally inches away from a crisis in public health that will severely hamper our ability to transplant organs, replace hips, provide cancer chemotherapy, perform dialysis, or keep premature babies alive. Then, it will only be a matter of time before otherwise healthy individuals will die because they contracted infections after routine surgery or were infected by pathogens that are common in the community but for which we no longer have a cure.” Death is stalking everybody. Get wise; get educated and take effective action.

My Advice

I can tell you that the usual trite advice: using chlorox or Lysol and washing your hands is a waste of time against the pneumonic form. Shocked? You shouldn’t be. Count on “official” recommendations to be stupid, inadequate and/or worthless.

They don’t even listen to their own science.

The question remains: if there is a pandemic of this organism, what should you do? Please note that my answers amount to good health advice for the present, when and while you remain fit and well. This is not just “disaster talk". Any good health measure is a good protection measure in the event of an epidemic or pandemic. Those who survive will do so because they have strong physiology and a vibrant immune system.

You can forget the idea that the doctor will come along and help you. Even if the medical profession knew how to deal with such a problem (which it doesn't), there would simply not be enough doctors to go around. You will be on your own.

But don't panic. Contact does not mean inevitably going down with the disease. Millions did not die in the Plague years. Those who worked the carts to carry off the dead bodies generally survived: it was NOT a suicide job! Doctors don't all fall over with the diseases they treat, even though they are in intimate contact with infected patients.

What is important is to build good health; overall health.

In addition, you need to know some very handy treatments that will kill pathogens as effectively as prescription drugs. You have not been told about these—and you are not likely to be, with the establishment taking its orders from the trans-national pharmaceutical cartels (trans-national really translates as they don't obey any country's laws).

The fact is these humble remedies I shall list are mostly cheap and there is not much profit in them. Not enough for greedy directors and stockholders of the drug companies.

This eReport will introduce you to DOZENS of modalities for treating and eliminating infectious bacterial pathogens. Many of the treatments, incidentally, will also work against viruses AND parasites, which is more than can be said for antibiotics.

But first, we have some study to do! As the sun begins to set on the magic of antibiotics and the (apparent) safety they brought, you need to learn what we are up against for the future.

This eReport will introduce you to DOZENS of modalities for treating and eliminating infectious bacterial pathogens. Many of the treatments, incidentally, will also work against viruses AND parasites, which is more than can be said for antibiotics.

But first, we have some study to do! As the sun begins to set on the magic of antibiotics and the (apparent) safety they brought, you need to learn what we are up against for the future.

Part 1

Understanding The Enemy

#6 Our fight with bacteria

There are approximately ten times as many bacterial cells as human cells in the human body, with large numbers of bacteria on the skin and in the digestive tract. The vast majority of the bacteria in the body are rendered harmless by the protective effects of the immune system, and a few are beneficial.

However, a few species of bacteria are pathogenic and cause infectious diseases, including cholera, syphilis, anthrax, leprosy and bubonic plague. The most common fatal bacterial diseases are respiratory infections, with tuberculosis alone killing about 2 million people a year, mostly in sub-Saharan Africa.

For a time we got ahead in the war against infectious diseases. Better nutrition and public hygiene have been the main reason for the improvements. Vaccinations seem to be mainly a waste of time and have made no real impact on the gradual disappearance of infectious diseases.

There is no question that antibiotics, too, made a huge impact in this war. For a time we had better weapons and ammunition than the enemy; it was only natural we would win. But even as antibiotics triumphed, within just a few years of their birth, resistant germs began to appear.

The implications of this were being discussed even in my time as a med. student in the 1960s. But nobody really paid much attention. The relentless advertizing and heavy pushes of the antibiotic manufacturers, coupled with the natural human tendency to laziness on the part of doctors, meant that the drug industry went on making huge profits. And all the while, the time bomb was ticking.

Back then we were just too naïve to understand or care about the possible long-term consequences of excessive prescriptions. Today, with 20/20 hindsight, we can say the path to where we are has been marked by reckless folly and greed.

We have arrived at a place where the short-comings of drug-based medicine are all too plain.

Antibiotics are gradually ceasing to work; most drugs have little real benefit and lots of side effects; the costs of prescriptions has risen to the point where drugs are now a huge proportion of the gross national product and deaths due to medicines has soared to the point where, in the US alone, iatrogenic deaths—deaths caused by doctors—are now over 100,000 per year.

That's more citizens than were killed by the entire 10 years of the Vietnam war!

And when I say drug deaths, I mean where the CORRECT drug, was CORRECTLY PRESCRIBED. Mistakes and blunders account for a further 7,000 deaths per year. And that's from a study carried out by a professor of medicine at Johns Hopkins School of Hygiene and Public Health, not "statistics" from left wing agitators or Ralph Nader.

[By the way, since it is germaine to the topic in question, you can throw in another 80,000 deaths a year from infections acquired in hospital and despite the very best of antibiotic therapy].

Not a pretty picture.

But the othere way that other way our world has been SERIOUSLY DAMAGED is by the obsession with finding the pharmaceutical "holy grail" is that basic medicine and health care has been lost.

Doctors don't even consider hygiene these days. They just fall back on antibiotics.

A study published in the Annals of Internal Medicine (2004), brought to light the fact that doctors, even surgeons, who thought they were not being watched, didn't bother to wash their hands much. They were caught on a hidden video camera.

This is despite the proof from Dr. Ignaz Semmelweiss, going back over a century, that surgeons washing their hands before an operation saved the lives of countless women in childbed (puerperal sepsis or "childbed fever" had over 50- 100% mortality and it was mainly caused by poor hygiene technique).

In this study, under an infection control expert with the University of Geneva Hospitals, researchers sought to better understand doctors' attitudes. They secretly tracked 163 doctors to monitor their hand washing during the day. Each doctor also completed a survey about their attitudes on hand hygiene.

They observed just 57% of doctors washing their hands between patients. Internists and medical students were the most diligent about washing their hands. Surgeons and anesthesiologists washed their hands least often.

On days when doctors had a busy workload, they were less likely to wash, the report shows. If they thought they were being watched, they were more likely to wash.

If doctors carried hand-wipe packets with them, or if hand-rub solutions were at the patient's bedside, they were more likely to use them. These solutions don't require water and a sink.

In fact, an expert infection-control panel recently issued a national recommendation promoting the use of these hand-rub solutions for disinfecting hands that are not grossly soiled, writes Robert A. Weinstein, MD, with Chicago's Cook County Hospital, in an accompanying editorial.

It's "excellent advice," he writes. Doctors "must use these products as a matter of ritual on entering and leaving every patient's room." Also, doctors must act as role models for medical students to ensure that each generation of doctors follows good hand hygiene -- and infection control -- practices.

It's not just hygiene

It's worse. Doctors no longer think of healthy nutrition. Yet that's the whole reason we beat TB. This disease was nearly gone before the first antibiotics arrived in the 1950s. TB is a disease of under-nutrition. Good food is the cure, not drugs and certainly not the vaccine (BCG).

Rest, calm, good food and love are the main ingredients of any good medical practice. Yet these are ignored, in the stupid belief they are unimportant and that drugs are the "real" medicine. Hell, people recovered all over the place before these drugs were ever invented.

There are some truths that will never go away, no matter what scientific advances may come about. Doctors who lose sight of these healing elements are a danger to the community.

Finally, as you know, countless safe and effective remedies from popular culture were summarily dropped as if they had no place in modern medicine. Herbs and other remedies have been around for centuries: they are not here because people are gullible and believe in folk myth and snake oil—but because these herbs really do work!

Homeopathy too has proven itself time and again. Now it is virtually banned and doctors who practice it treated as outlaws.

The result?

We are going downhill rapidly to a dangerous new time, when drugs may not be able to dig us out of the mess caused by doctor folly. I am concerned.

#7 Killing Bacteria Has A Downside

Even killing pathogens can wreak havoc. Consider this cautionary tale: it shows we have a lot more to learn.

When scientists discovered that a bacterium Helicobacter pylori, not stress, caused most stomach ulcers, two researchers were awarded a 2005 Nobel Prize and medical practice changed significantly. Now antibiotics are used to treat ulcers. It is satisfying to note that ulcers and even stomach cancers have declined dramatically in numbers.

However microbiologist Martin Blaser and his colleagues at New York University began to document an odd medical trend. As Helicobacter pylori, virtually disappeared among children, cases of asthma tripled. So did rates of hay fever and allergies, such as eczema.

After analyzing health records of 7,412 people collected by the National Center for Health Statistics, Dr. Blaser and NYU epidemiologist Yu Chen reported this summer (2009) in the Journal of Infectious Diseases that children between three and 13 years old who tested positive for H. pylori bacteria were 59% less likely to have asthma. They also were 40% to 60% less likely to have hay fever or rashes.

Among adults, gastro-esophageal reflux disease (GERD) became more common, as did some forms of esophageal cancer.

Dr. Blaser soon reasoned that eliminating the harm these bacteria cause is at the expense of the protection they provide us. He may be right. The jury is still out.

One thing is for sure: we are landlords to an immense host of bacteria and most of these species, and what effects they may have, are still unknown to science. But we must now suspect that they interact with us in diverse ways and can influence obesity, heart disease and cancer, as well as many other conditions.

As many as 500 species of bacteria may inhabit our guts, like H.pylori. Maybe 500 or so other species make themselves at home in our mouth, where each tooth has its own unique bacterial colony. No one knows how many species we contain in all. This past August, researchers at Kings

College London identified yet another new species of oral bacteria between the tongue and cheek.

We are the equivalent of an Amazonian rain forest, where they discover new species every week!

Until recently, half of humanity carried Herlicobacter pylori stomach bacteria, according to a 2002 study in the New England Journal of Medicine. Indeed, we appear to have evolved together. Among those born in the U.S. under the age of 25 or so, only 5% or so still carry Helicobacter, largely due to the indiscriminate use of antibiotics.

The connection to allergies is just one of the pressing public health puzzles posed by our complex relationship with the trillions of microbes that call us home. "Recent studies have shown that changes in bacteria can be correlated with some pretty serious diseases," says Jane Peterson, head of the National Human Genome Research Institute's comparative sequencing program.

Childhood diabetes also is on the rise in developed countries, for instance. Last week, University of Chicago immunologist Alexander Chervonsky and his collaborators at Yale University reported that doses of beneficial stomach bacteria can stop the development of Type 1 diabetes in lab mice.

Common stomach bacteria also can stop the development of Type 1 diabetes in lab mice, researchers have reported. The researchers believe their findings could one day be used to develop bacteria-based treatments for patients.

Other bacteria are just as crucial to our well-being, feeding us the calories from food we can't digest on our own, bolstering our immune systems, tending our skin and dosing us with vitamins, such as B6 and B12, which we are unable to synthesize unaided.

So indiscriminate use of antibiotics may carry a whole host of complications that cannot be foreseen. For sure, we will disturb the natural balance and for that we have a term: dysbiosis. It is particularly important in the gut, where dysbiosis can lead to multiple health problems, including food allergies and overgrowth of less friendly yeasts and bacteria.

Metagenomics May Give Us The Edge

Now scientists are trying to sequence bacterial genes (like plotting the human genome). No-one knows what will emerge from an investigation like this. But it's being taken very seriously. In 2008 the National Institutes of Health launched a 5-year, $125 million Human Micro-biome Project to analyze hundreds of microbial species that make your body their home.

Making the project possible is a new gene-mapping technique called metagenomics.

To start, researchers at the Baylor College of Medicine in Houston, the Washington University School of Medicine in St. Louis, the Broad Institute in Cambridge, Mass., and the J. Craig Venter Institute in Rockville, Md., have been sequencing the genomes of 200 microbe species isolated from 250 healthy volunteers. They are sampling bacteria from the skin, gut, vagina, mouth and nose, then attempting to identify them by cataloging variations in a single gene sequence that all bacteria share.

Working with similar metagenomics projects in Europe, Japan, China and Canada, they hope to assemble a reference collection of genomic information covering 1,000 microbial species that infest us. If all goes as planned, they may soon find themselves trying to analyze 200,000 genes, compared to a feeble 20,000 for the human genome.

These data sets are so huge, and we don't have the tools yet to analyze them. But the diversity is more than anyone expected. Dr. Segre, who specializes in the study of the skin, found one set of microbial communities thriving in the bend of the typical elbow and an entirely different set of colonies on the average forearm.

In all, she identified 113 different kinds of bacteria living in concentrations of about 10,000 per square centimeter on the surface and, just beneath the skin, in densities of one million microbes per square centimeter, she reported last May.

In a real sense, the history of all these many microbes is the history of humanity itself. "We are living beings that co-evolved with micro-organisms," Dr. Segre says. Evidence suggests that strains of helicobacter bacteria evolved along with humankind from its beginnings in primitive organisms a billion years ago. Every mammalian species appears to have its own unique variety of these microbes. Human-strain Helicobacter pylori accompanied our anatomically modern ancestors on their great walk out of Africa about 58,000 years ago. Molecular epidemiologists at the Max Planck Institute for

Infection Biology in Berlin have so far identified 370 strains of the bacteria that seemed to reflect the migrations and settlements of their human hosts.

It is vital that we understand bacteria better, like this, rather than just thinking in crude terms of antiseptics and antibiotic drugs.

#8 The rise of the superbugs

The greatest old-time "superbug" was the one that brought the ravages of the "Black Death". That's the dread plague that swept the world in waves, over the 13th – 17th century. At its peak, it wiped out one third of the known population of the world. Imagine that today. Two billion people; where would we put all those bodies? We'd need machinery to dig graves fast enough. Decayed people would ooze into our water table.

However today, I don't think we wouldn't lose such a big proportion of folks, because of antibiotics. But what would happen if antibiotics began to fail us, as they are beginning to do? It's hard for younger people born today to imagine what such a world would be like.

You'd have to be over 70 years to remember back to a time like that. I'm in my sixties and, as a baby boomer, was born right at the start of the penicillin era.

But there were still cases walking the streets that you just wouldn't see today. Like the kid in our village who had his right arm amputated at the shoulder. I was puzzled and upset by his deformity. Then the adults explained he'd nicked his finger on a tin can, got gangrene and had to lose his arm to save his life.

Later in life, while I was at med school, there were still cases showing the era of the "old medicine". Congenital syphilis, for example. It shows with a collapsed nose bridge and punched out holes, like extra nostrils , sucking air right through the sides. You could see that on a public 'bus quite regularly (though of course the syphilis had eventually been eradicated in that person's body, by the judicious use of antibiotics).

The late stages of syphilis were in evidence right up to the 1970s. General paresis of the insane (GPI, so-called) was the outrageous one, almost funny, characterized by megalomania. One patient when I was a medical school declared he was God, escaped the hospital and tried to buy a building for a million (he had no money), swept up to a minor radio celebrity and proposed to her and was finally picked up by the police "shooting the rapids" in a sewage facility.

It has been speculated that Nietsche, a megalomaniac philosopher if ever there was one, suffered with GPI. Following a psychotic breakdown in 1889, at the age of 44 years, he was admitted to the Basel mental asylum and on 18 January 1889 was transferred to the Jena mental asylum. He remained in demented darkness until his death on 25 August 1900. In Basel, a diagnosis of general paralysis of the insane (GPI; tertiary cerebral syphilis) was made.

This diagnosis was confirmed in Jena and is still widely accepted.

But easier to see, walking down the street, was Tabes dorsalis. This wretched condition is characterized by a stamping gait, due to loss of leg co-ordination. But the worst aspect, from the patient's point of view, are the stabbing agonizing pains in the back and down the leg. Poor Delius, the composer, had this.

Ironically, normal nerve signals didn't get through. The patient would smash up their own knee joints, etc. because the protective mechanisms were not in place. It's a moral lesson in the fact that pain is really there to protect us from harm, not make us miserable.

Why am I wasting my time on this despicable disease, some may ask? Because syphilis was one of the most feared of all diseases for centuries and kept young men and women in terror of the act of union. It cast its dark shadow over saints and sinners alike and many a creative genius was snuffed out by its chill wind.

Trust me, you would not want this monster to break loose again. Yet antibiotic-resistant syphilis has been with us for decades, certainly since the Vietnam war. GIs on furlough, young tarts who want to make cash giving them sex and bad prescribing by military MDs made it almost inevitable.

All those romantic Victorian novels

Who hasn't read Dickens and Jane Austen and all those great old-time stories? Often one of the characters takes ill and dies, almost immediately. You wonder why?

The dreaded diseases were TB (a very slow death, due to consumption) and lobar pneumonia. The latter was once nicknamed "Captain Of The White Horsemen Of Death", due to its terrifying swiftness. A person would get wet on Friday and take a chill, develop a fever on Saturday and be dead by bedtime on Sunday. There was no treatment as such, all the doctor and family could do was wait for the "crisis".

This is the moment you have seen often in movies and on TV; the patient's fever rages higher and higher and they rave in delirium. The loving wife sits there helplessly mopping the sweat from the patient's brow (and her own, I imagine). Over the space of a critical couple of hours, with the mother of all battles raging in the victim's body, fate was settled. Either he or she expired, or the immune system suddenly got the upper hand (like an arm wrestling contest, where the loser suddenly gives in).

In which case the fever peaks, the patient stirs, opens their eyes in confusion and draws their first free breath for what seems like a month (actually less than 12 hours, usually).

Those were the days when you simply had few resources. The important thing was not to get the infections in the first place.

Well, I 'm telling you all this, not to make you feel gloomy and afraid, but to really celebrate antibiotics. They really have been one of the most, if not THE most, telling medical discovery of all time. Antibiotics have saved inestimable billions of lives.

In fact it's my firm opinion that the wonder of antibiotics is what really led us into the gullible belief that drugs were great and that there must be a chemical cure for everything. Antibiotics (and a couple of other amazing substances, like ACTH as it was prescribed then, or steroids as we prescribe them today) did seem to usher in a new era of amazing pharmaceutical cures.

Most of us are well aware that the rest of the promises of the drug industry also have not been, and will not be, fulfilled. Diseases cannot be treated as a "drug deficiency", as many doctors seem to think. It has taken decades to realize that antibiotics do not come without a price and that they will eventually fail.

The problem of this legacy today is NOT the antibiotics. It's resistance to them.

Resistance and the fact that antibiotics have made us careless; we are not nearly cautious and guarded enough of pathogens (being careful does not require being fearful, just levelheaded). Doctors and public alike have learned to be cavalier and have grown up with the attitude that antibiotics will save the day. To use antibiotics is a kind of failure; it means the important preventative basics were not being applied. Trouble is as inevitable as night follows day.

You saw the movie "Jurassic Park". You will remember Jeff Goldblum's character repeating, over and over, the warning that "Life will find a way".

Life is incredibly ingenious and tough when it comes to surviving; it will get round almost any barrier, just as the dinosaurs engineered their way round an amino acid deficiency that was supposed to keep them infertile [OK, I don't normally use fictional sources—but this movie made a very important and valid scientific point].

Well—here's the thing—bacteria (and viruses and parasites, for that matter) are also life forms! They too are tough, very ingenious and swift to adapt their abilities to their needs. So in the course of time they have developed new metabolic tricks, so that antibiotics cease to work.

Man then developed newer and deadlier antibiotics. But pretty soon organisms started to adapt to those too. It's been a war of attrition. Unfortunately, Man is showing every sign of eventually losing this war. We are running out of ideas for new drugs to kill pathogens, without killing the patient.

Pathogens, it seems, will never run out of ideas to counter our efforts.

So the outcome was never in doubt. It's just that armies of irresponsible doctors have hastened the whole end by assiduously prescribing antibiotics when they are not necessary. Add to that the fact that antibiotics are used in various agricultural processes, treating livestock to enhance their growth, so the food chain is also a risk in creating antibiotic resistance.

See, the more times pathogens are exposed to toxic antibiotics, the more likely it is—just on the law of averages—that sooner or later one of them would figure out a way round the antibiotic.

As soon as a resistant strain emerges, obviously that bacterium can multiply unchallenged by therapy. Bacteria are capable of dividing every 20 minutes or so. That means in just 10 hours a single bacterium could become 268,000,000—enough to fill a bucket (at 10 hours and 20 minutes, they would divide and fill 2 buckets!)

Of course this won't happen in real life, because the food supply wouldn't keep up with demand. But you get the idea? What was once just a few lucky cells, pretty soon becomes a torrent of billions, then gazillions of cells. MRSA, the hospital resistant staphylococcus, was once just such a handful of cells. Now this dreaded organism is in almost every ward in every hospital; on people's skin; in their nostrils; on their towels and clothing… just about EVERYWHERE.

That's why there is a problem.

#9 Agribusiness is to blame

It's true that doctors have abused antibiotics for decades. But the really guilty party and the truly reckless fools are those in the agribusiness (big shot farmers, livestock merchants, food manufacturers and all their hangers on).

Did you know that 70% of all antibiotic consumption goes, not to hospitals for administration to humans, but the farmers, to feed their livestock? That's according to a careful study by the Union of Concerned Scientists — and that is the main reason we're seeing the rise of pathogens that defy antibiotics.

While doctors are being urged not to use prophylactic antibiotics, even to protect human babies, farmers use them in huge quantities to protect baby calves, pigs and poultry. It's worse than criminal. It's INSANE.

It inevitably gives rise to resistant strains of bacteria, as you have learned. But these resistant, dangerous bugs don't just hang out in the farmyard! They turn up in our food supply! We swallow them!

According to Pultizer prize winning New York Times writer Nicholas Kristof , "Five out of 90 samples of retail pork in Louisiana tested positive for MRSA — an antibiotic-resistant staph infection — according to a peer-reviewed study published in Applied and Environmental Microbiology last year.

"Another recent study of retail meats in the Washington, D.C., area found MRSA in one pork sample, out of 300, according to Jianghong Meng, the University of Maryland scholar who conducted the study.

"Regardless of whether the bacteria came from the pigs or from humans who handled the meat, the results should sound an alarm bell, for MRSA already kills more than 18,000 Americans annually, more than AIDS does. A new strain called ST398 is emerging and seems to find a reservoir in modern hog farms. Research by Peter Davies of the University of Minnesota suggests that 25 percent to 39 percent of American hogs carry MRSA." Public health experts worry that pigs could pass on the infection by direct contact with their handlers, through their wastes leaking into ground water (one study has already found antibiotic-resistant bacteria entering ground water from hog farms), or through their meat, though there has been no proven case of someone getting it from eating pork.

Thorough cooking will kill the bacteria, but people often use the same knife to cut raw meat and then to chop vegetables. Or they plop a pork chop on a plate, cook it and then contaminate it by putting it back on the original plate.

Yet the central problem here isn’t pigs, it’s humans. Unlike Europe and even South Korea, the United States still bows to agribusiness interests by permitting the non-therapeutic use of antibiotics in animal feed. That’s unconscionable.

The peer-reviewed Medical Clinics of North America ran an article last year that concluded that antibiotics in livestock feed were “a major component” in the rise in antibiotic resistance. The article said that more antibiotics were fed to animals in North Carolina alone than were administered to the nation’s entire human population.

“We don’t give antibiotics to healthy humans,” said Robert Martin, who led a Pew Commission on industrial farming that examined antibiotic use. “So why give them to healthy animals just so we can keep them in crowded and unsanitary conditions?”

The answer is simple: greed.

Legislation to ban the non-therapeutic use of antibiotics in agriculture has always been blocked by agribusiness interests, a rich and very powerful lobby.

Louise Slaughter of New York, who is the sole microbiologist in the House of Representatives, said she planned to reintroduce the legislation this coming week. “We’re losing the ability to treat humans,” she said. “We have misused one of the best scientific products we’ve had.”

That’s an almost universal view in the public health world. The Infectious Diseases Society of America has declared antibiotic resistance a “public health crisis” and recounts the story of Rebecca Lohsen, a 7-year-old New Jersey girl who died from MRSA in 2006. She came down with what she thought was a sore throat, endured months in the hospital, and finally died because the microbes were too virulent for the drugs.

Again, let Nicholas Kristof speak: “This will be an important test for President Obama and his agriculture secretary, Tom Vilsack. Traditionally, the Agriculture Department has functioned mostly as a protector of agribusiness interests, but Mr. Obama and Mr. Vilsack have both said all the right things about looking after eaters as well as producers.

“So Mr. Obama and Mr. Vilsack, will you line up to curb the use of antibiotics in raising American livestock? That is evidence of an industrial farming system that is broken: for the sake of faster-growing hogs, we’re empowering microbes that endanger our food supply and threaten our lives”.

#10 Turning back the clock

Surgeons had a whole different attitude to infections in days past. They talked of healing by first intention and second intention. They knew how to nurse an abscess into discharging safely. These were MAJOR skills because if you got it wrong and it went to septicemia, you were likely dead.

But there were many bad problems. You wouldn't want to go back to those days. For example, cavernous sinus thrombosis. What that means is an abscess forming in the veins of the base of the brain (the cavernous sinus) and this can happen because of wounds or infected sores or spots on the head and face. Typically, the infection would travel backwards through into the skull, because there are no valves in these veins.

The cavernous sinus would sludge up and clot and the whole mass become a seething mass of infection. As a result, the brain could also become infected, a brain abscess form and death was the usual result.

That's why concerned parents of old would urge their children not to pick spots on the face and I had many a slap from my mother for trying to do so. I thought it was silly—till I went to med school years later and learned where this ancient injunction came from.

Country people didn't know much medicine but they had learned, over the centuries, that a spot around your face could kill you if it went wrong.

Teeth Infections

One of the biggest dangers of death by infection was a tooth abscess, as I wrote about in my book Virtual Medicine. At the beginning of the 20th century, the single biggest predictor of whether you are likely to die of heart disease, was teeth. You could die of a coronary, but you could die of a disastrous heart disease caused by bacteria getting loose in your blood and settling on the valves in your heart.

And of course it is not confined to the heart; bacteria that are thrown off from the teeth can travel to the brain, the kidneys and other organs.

According to one informed dentist, if we were to spread tooth infections all together on to our skin, we would have a huge infected ulcer as large as the back of our hand. Yet because this is all hidden around the roots of teeth, we ignore it and don't take the danger seriously enough.

We make light of a visit to the dentist these days, we've got used to not living with these dangers, but believe me these were very serious matters and a lot of people found themselves sick unto death quite unexpectedly, and there were many, many tragedies.

If we lose the control of microbes that antibiotics promised, we will be back in a primitive era when even the simplest health problem COULD turn out to be a fatal event.

One day, everything changed.

Along came sulphonamides. They were pretty toxic substances but the benefits far outweighed their harm.

Then penicillin arrived. It is probably the most wonderful medical breakthrough ever discovered. It has saved billions of lives. Yet penicillin remains one of the most non-toxic drugs in the entire medical armoury. It's virtually harmless (allergy is the only real problem).

Penicillin was literally a miracle. It transformed the medical landscape. For a time it looked like medical science was actually going to conquer the world.

Nobody ever considered this idea that we could go backwards; that we might lose this wonder therapy; that it could be misused and start messing up our bodies. Nobody saw that far.

Today we know that so-called "broad spectrum" antibiotics, which kill just about everything, can lead to disastrous upset of the whole bacterial colony which lives on and in our bodies. The fact that we so much depend for our health on this entire set of organisms (as described in section #2) means that antibiotics, as often as not, end up causing chronic disease; not fixing it. We call this condition dysbiosis (dys- means "bad" as in dyslexia, bad words; dysentery, bad guts; dyspepsia, bad digestion).

But of course, if you've got pneumonia, and you might die within two to three days without treatment, you are not going to care much about whether it upsets the fine balance your intestinal flora.

The real crime was the abuse of antibiotics for trivial conditions, that would self-resolve anyway. Patients became demanding and doctors became lazy and inattentive of medical basics. But there were other serious mistakes made…

One of the worst abuses was the way the STDs are treated (sexually transmitted diseases; formerly called venereal disease). It has traditionally been very difficult to treat and monitor STD cases. The shame and embarrassment mean that patients will often not re-attend for further treatment and follow up, even when strongly urged to do so, in their own interests.

Knowing this, STD clinics eventually took to trying to eradicate the infections with one-off huge doses of antibiotic.

Trouble was inevitable.

We now have antibiotic resistant gonorrhea.

Then there is resistant syphilis. This started out in the far east, with the Vietnam war. It isn't yet widespread. But it's ominous. Because if syphilis comes back as it once was, it was among the most feared diseases in civilization. Not a nice way to die. Many famous people have died of syphilis, including King Henry VIII, Schubert the composer, Lenin, Christopher Columbus and quite a few Popes by all accounts. But that's another story.

Serves them right, some people might say. That would be unkind but what about children born with this infection? It's not their fault.

Chlamydia is an organism recognized as mainly sexually transmitted and is the commonest infection in your women here in the US. So it's not a time to get supercilious or puritanical. It's just the way things are.

Now in my medical school days there was a saying about syphilis. Among the many damaging effects is the fact that it attacks the main vessel from the heart and leads to deadly aneurisms. So the saying was that "The chickens of an intemperate youth come home to roost in a middle-age aorta".

Ironically, Chlamydia also attacks blood vessels, causing coronary artery inflammation and thrombosis. So we might, perhaps, modernize the saying to: The chickens of an intemperate youth come home to roost in middle-aged coronary arteries!

I'm not just being flippant. These jokey-type sayings often help memorize aspects of the data.

It was not always mistaken usage. To a certain extent, antibiotic resistance was inevitable, even given proper usage. It was just a matter of time. In fact it came very early 1950s. A hospital in Japan had an outbreak of Shigella, the bacterial form of dysentery. The organism was found to be multiple drug-resistant (MDR that's another term you'll hear more about) and could not be controlled by sulfonamides, chloramphenicol, tetracyclines or streptomycin.

It sent shock waves through the scientific community.

But of course it isn't just the fault of doctors. Worldwide very many antibiotics have been freely available over the counter and been taken by people who have no idea of the correct use of these powerful medicines. This is all to support sales and profits of the pharmaceutical industry and has nothing to do with medical best-practise.

Drug resistant TB is also on the rise, especially in inner cities. It's often a disease of poor people because they don't eat properly and TB is mainly a disease of malnutrition.

The overall number of tuberculosis (TB) cases in Texas has declined during the past few years, but the rate of multidrug-resistant (MDR) cases has increased, making treatment more difficult. These findings were presented at the 135th annual meeting of the American Public Health Association.

The most recent national data (2005) showed 124 of 10,662 tuberculosis cases as being MDR-TB; resistant to both the main treatment drugs. The rate is higher in immigrants.

Texas ranks second in the United States in terms of number of TB cases, after California, while New York has slipped to third, and Florida is fourth.

Even if treatment is successful it takes a LONG time. The average time to complete therapy for a drug-resistant case is a year; it doubles to 2 years for MDR-TB, while some individuals can take up to 3 years to rid themselves of the disease. Most treatment is conducted as directly observed therapy (DOT). The patient is watched, to ENSURE that he or she swallows it!

Part 2

What Can You Do To Kill Bacteria?

#11 Plain water

Let's start with simple remedies first. What's cheap, plentiful, non-patented and very effective for eliminating bacteria? Chlorhexidine? Dettol? Give up?

It's water. Plain water, not saline. Plain water, not antiseptic scrub even! That's according to a study published in January 2008.

Don't forget, water is very cleansing. It can dilute toxins and wash away toxins and organisms. That will help the body fight infection. Doctors, pharmacies and hospitals use saline, which is a salt solution at the same concentration as the body tissues. This is gentler. But at a pinch, water will do fine. In fact just as well and is much cheaper, of course. If it's healthy enough to drink, it's healthy enough to use on an open wound is the basic message from a recent extensive study, published in the Cochrane Database of Systematic Reviews, January 24, 2008.

The results might seem surprising: compared with hospital saline, tap water was more effective in reducing the infection rate in adults with acute wounds but not in children. To me that evens out at about the same. No real difference in other words.

There were lots of easy criticisms of the methodology used in this study. Some of the comparisons were based on a single trial. But I don't think the little details detract from the main overall finding: water works just fine. Besides, drugs get into the market place at huge profits with far less credible science than this.

Although wound care has changed significantly in the last decade, there has been limited focus on the types of solutions used for wound cleansing. Antiseptic preparations have been traditionally used, but animal models suggest that antiseptics may actually hinder healing.

Saline irrigation has been the natural alternative for doctors. But plain water is just fine for the rest of us.

No getting away from it: washing your hands is a good protection against the transmission of germs. I have already pointed out to you that it won't protect you against the pneumonic form of MRSA. But the likelihood of transmitting almost any other bacteria, including the regular MRSA, is proven to be reduced.

Heck, this was the story told over a century ago by Ignaz Semmelweiss. He pointed out that if surgeons would just wash their hands after dissecting dead bodies and BEFORE they put their fingers into the intimate parts of women in labor, that dreaded childbed fever (puerperal sepsis) was virtually eliminated.

Since this fever had almost 100% mortality, you would think doctors would rejoice in the discovery. But in fact they attacked Semmelweiss with such a fury and vengeance that the poor man eventually committed suicide. Thus passed one of my great heroes and a medical giant, brought down by ignorant pygmies.

But his message remains clear and bright. Hand washing is good; dirty hands and fingers, bad.

Antibiotic Soap?

Contrary to what you may think, antibiotic soaps are a BAD idea. They are not only not necessary, they may contribute to the development of antibiotic-resistant strains, which is definitely a bad thing.

Despite the proven health benefits of hand washing, many people don't practice this habit as often as they should — even after using the toilet. If you don't wash your hands frequently enough, you can infect yourself with these germs by touching your eyes, nose or mouth. And you can spread these germs to others by touching them or by touching surfaces that they also touch, such as doorknobs.

Infectious diseases that are commonly spread through hand-to-hand contact include the common cold, flu and several gastrointestinal disorders, such as infectious diarrhea. While most people will get over a cold, the flu can be much more serious. Some people with the flu, particularly older adults and people with chronic medical problems, can develop pneumonia. The combination of the flu and pneumonia, in fact, is the eighth-leading cause of death among Americans.

Inadequate hand hygiene also contributes to food-related illnesses, such as salmonella and E. coli infection. According to the Centers for Disease Control and Prevention (CDC), as many as 76 million Americans get a food-borne illness each year. Of these, about 5,000 die as a result of their illness. Others experience the annoying signs and symptoms of nausea, vomiting and diarrhea.

The moral is clear: wash ‘em. And often. Sanitizers are merely an option. As I said, we don’t need these. But if you must, the US Center For Disease Control (CDC) recommends choosing products that contain at least 60 percent alcohol.

To use an alcohol-based hand sanitizer:

- Apply about 1/2 teaspoon of the product to the palm of your hand.
- Rub your hands together, covering all surfaces of your hands, until they’re dry.

If your hands are visibly dirty, however, wash with soap and water, if available, rather than a sanitizer.

When should you wash your hands?

- After using the toilet
- After changing a diaper — wash the diaper-wearer’s hands, too
- After touching animals or animal waste
- Before and after preparing food, especially after handling raw meat or poultry
- Before eating
- After blowing your nose
- After coughing or sneezing into your hands
- Before and after treating wounds or cuts

- Before and after touching a sick or injured person
- After handling garbage
- Before inserting or removing contact lenses
- When using public restrooms, such as those in airports, train stations, etc.

Kids need clean hands, too

You can help your children avoid getting sick by insisting that they wash their hands properly and frequently. To get kids into the habit, teach by example. Wash your hands with your children and supervise their hand washing. Place hand-washing reminders at children's eye level, such as a chart by the bathroom sink for children to mark every time they wash their hands.

Make sure the sink is low enough for children to use, or that it has a stool underneath so that children can reach it. Tell your children to wash their hands for as long as it takes them to sing their ABCs, "Row, Row, Row Your Boat" or the "Happy Birthday" song. This works especially well with younger children, who may rush when washing their hands.

Hand washing doesn't take much time or effort, but it offers great rewards in terms of preventing illness. Adopting this simple habit can play a major role in protecting your health.

#12 Drawing Salve and Ointments

Here’s another “golden oldy”.

“Drawing” is a very ancient technique that needs to be kept in view. Sometimes it works instead of antibiotics; sometimes it needs to be used in conjunction.

It simply means to draw the fluid out of a swollen infected sore, boil, wound or abscess. A highly concentrated substances, such as icthammol or zinc oxide paste is smeared over the infected site and kept in contact with a bandage or poultice. It sucks out the moisture by the scientific principle of osmosis.

The dehydration effect has two benefits: bacteria don’t like it, it sucks moisture from them and they die (that’s the principle of using a strong sugar density in jelly, jam and conserves). But it also draws moisture from the inflammation site, which lessens the pressure and so makes it less painful.

What was red, swollen and tense tissue turns soft, crinkly and less painful. It’s a sort of halfway house to lancing a boil or abscess; reducing the pressure without needing to actually make a cut.

Very often the use of a drawing salve alone will eradicate an unpleasant boil or infection resulting from a splinter or small wound. But beware: it does not remove diseases which make skin infections a hazard, such as diabetes or compromised immunity.

One of the simplest is 25% zinc oxide paste, from the regular pharmacopeia. It is readily available in small tubes or tins.

Close favorite is 20% icthammol ointment or “black drawing salve”. It stinks and looks unpleasant—but it does the job. Herbs such as Calendula, Echinacea and others in a base of beeswax and Vitamin E or olive oil may be added to the icthammol salve. Cover with a bandage since it stains clothing.

Hylands Homeopathic Drawing Salve: An oldie but a goodie, Hylands (also called PRID Drawing Salve) includes components of both a drawing salve and homeopathic remedy. Eight physicians invented this salve in 1903. PRID contains the homeopathic ingredient acidum carbolicum along with herbs such as calendula and Echinacea. Other ingredients include ichthammol, sulphur and silicea.

At a pinch you can make your own drawing salve, using the strongest solution you can make of table salt or Epsom salts in hot water. Soak a bandage or compress in the mixture and then apply. Unfortunately, this remedy needs frequent repeat applications over many hours.

You could also try a lavender and tea tree oil compress: Both lavender and tea tree oil contain antibacterial agents particularly useful against boils (see section #39).

Drop tea tree oil directly onto the boil for one to two days during the first phases to bring it quickly to a head. Lavender, however, is too irritant for direct application.

Margosa or Neem is a very effective natural antibiotic. You can make an ointment for boils by boiling some neem seeds in water till they are soft. Then take some neem leaves and smash them to make a pulp. Mix these boiled neem seeds and neem leaves and you get a natural ointment for boils which can be directly applied to the area. More over this poultice even prevents the infection from spreading as well.

You can even use castor ointment for boils. All you need to do is to make a powder out of the castor bark and add some water to turn it into a paste. You can even add grind the root of the herb biskhapra for additional effect.

Another effective ointment for boils can be made by grinding onion and garlic paste. This can aid in healing the boil quickly.

Always wash your hands thoroughly after treating boils. The bacteria spread easily from person to person, and can spread from one area of the body to another.

#13 Blue light

Next comes a pretty simple method of sterilization from the world of physics. Light.

One of the miracles of the amazing electromagnetic spectrum, which I explain to people is the fact that although most of the electromagnetic spectrum is very hostile to life (such as microwaves, ultraviolet light, and gamma rays), the fact is that a very narrow band of the electromagnetic spectrum, that we call “light” is not only friendly, but is actually essential to life.

However, here is a useful reminder that even light is part of that dangerous spectrum, and can have destructive effects on living organisms, blue light especially (red light, which spills over into the infrared range is warmth, and therefore also somewhat supportive of life).

Blue light, being close to the ultraviolet range, can itself be quite harmful, and we can cash in on this property and make it useful. We find the bluelight is great for killing resistant bacteria, as a study published in Photomedicine and Laser Surgery, April 2009, shows.

Specifically, blue light NOT including dangerous UV frequencies, kills MRSA (see section 45 for news about the safe use of UV light as an antibiotic).

The finding comes from Chukuka S. Enwemeka, PhD, and colleagues at New York Institute of Technology. Their study was funded by Dynatronics Corp., which makes the blue-light device used in the study.

In earlier studies, Enwemeka’s team found that MRSA died when exposed to blue light that included part of the ultraviolet (UV) spectrum. Even though the total UV dose was less than that of a few minutes of sunlight, it would be safer not to expose humans to any more UV light than necessary.

So the researchers used a LED device that emits only blue light, not UV, and found it worked nearly as well!

“Irradiation with [blue] light energy may be a practical, inexpensive alternative to treatment with pharmacologic agents, particularly in cases involving cutaneous and subcutaneous MRSA infections,” Enwemeka and colleagues conclude. The researchers tested two MRSA strains: one typical of the strains that bedevil hospitals, and one typical of the strains found in the community. Both strains were susceptible to the blue light.

Relatively low doses of blue light, less than 2 minutes, killed off about 30% of MRSA in laboratory cultures. Obviously, longer exposure was better, but 100% death of organisms did not come about, even through very long exposures.

Exactly how blue light kills MRSA, or whether the bacteria can become blue-light resistant, isn't known.

Tell your doctor about this study if you get any resistant surface infection. But remember also, light penetrates well below the skin layer (shine a torch through your hand on a dark night and you will see what I mean).

See just how valuable light is by viewing this YouTube video by Dr. John Ott, one of the pioneers in light and health (I featured his work in Virtual Medicine, first published 1999).

http://www.youtube.com/watch?v=bw6hcTGND3c

And let me close by commenting that fresh air and sunlight have traditionally been recognized for their healing powers. Moms always knew it! To expose an infected wound to bright sunlight would do far more good than covering it with bandages, most especially if the organism is one of these resistant toughies anyway.

I always do away with the band aid at the earliest available moment, for this reason.

Get a UV Light Nano To Protect Yourself Against Dangerous Bacteria

Note: You can always get a Nano UV germ zapper. It's a handy electronic gardget that looks like a cell phone and is about the same size. But it delivers a powerful light punch; a mixture of UV-A, UV-B and UV-C.

Read more in section #47.

Soon, we are going to do some nutrition stuff. But this next few are different. It’s from the larder but not quite nutrition!

Let’s start with a surprising contender: HONEY! What’s surprising is not that it works as an antibtioic but that conventional doctors and hospitals openly admit its efficacy!

#14 Manuka Honey!

Honey has been known for its healing properties for thousands of years - the Ancient Greeks used it, and so have many other peoples through the ages. Even up to the Second World War, honey was being used for its antibacterial properties in treating wounds. Then along came antibiotics, as I explained, and it was promptly forgotten.

But in recent times there has been a renaissance of the use of honey in the medical profession: an editorial in the Journal of the Royal Society of Medicine (Zumla and Lulat 1989) discussing this expressed the opinion “The therapeutic potential of uncontaminated, pure honey is grossly underutilized”.

I know for example that bandages soaked in manuka honey were given to patients at the Christie Cancer Hospital in Manchester to reduce their chances of contracting the MRSA superbug and to lessen wound inflammation following surgery.

Honey is used routinely at the Manchester Royal Infirmary for dressing wounds, and other research has found it can fight gum disease, ease digestive problems and soothe sore throats (I got my bachelor of medicine and bachelor of surgery at Manchester University and so I know the Christie and the Royal Infirmary very well).

The way honey works is that it is very rich, it is like jam or preserves you know; adding the sugar to the fruit is what makes it impossible for the bacteria to grow, it just sucks the water out of them so they can’t grow and honey does the same thing. So remember diluted honey does not work and will cause bacteria to flourish!

A British team from the famous University of Oxford and the Radcliffe Hospital in Woodstock Road, Oxford (I added that for you Morse fans out there), have done a systematic review of the use of honey as a wound dressing, compared to other dressings and treatments *[BMC Complement Altern Med. 2001; 1: 2. Published online 2001 June 4. doi: 10.1186/1472-68821-2].*

Time for healing was significantly shorter for honey than all other treatments but, the researchers complained, the quality of studies was low. However the results were pretty clear, taken over the number of studies. Honey won every time.

There was one study in which the healing of infected postoperative wounds compared honey with antiseptics in addition to systemic antibiotics after culture and sensitivity. For all outcomes honey was significantly better, with much shorter times for healing, eradication of infection, use of antibiotics and hospital stay.

[Al-Waili NS, Saloom KY. Effects of topical honey on post-operative wound infections due to Gram positive and Gram negative bacteria following Caesarean sections and hysterectomies. Eur J Med Res. 1999;4:126–30]

Another study I found in the British Journal Of Surgery (which was not included in the Oxford group review) showed dramatic benefits from honey in treating infected wounds and ulcers.

Fifty-nine patients with wounds and ulcers most of which (80 per cent) had failed to heal with conventional treatment were treated with unprocessed honey. Fifty-eight cases showed remarkable improvement following topical application of honey. Only one case, later diagnosed as Buruli ulcer, failed to respond (Buruli ulcer is caused by Mycobacterium ulcerans, a relative of TB). *[Br J Surg. 1988 Jul;75(7):679-81].*

What's especially interesting is that honey allows wound healing with little or no scarring. It seems that honey is able to restrain the excessive growth of collagen which takes place during wound healing. The less collagen, the less scarring.

[Topham J. Why do some cavity wounds treated with honey or sugar paste heal without scarring? J Wound Care. 2002 Feb;11(2):53-5].

Propolis

Propolis is the resinous substance collected by bees from the leaf buds and bark of trees, especially poplar and conifer trees. Bees use the propolis along with beeswax to construct their hives. Because of antimicrobial properties of propolis, it helps keep hives free of germs.

The great violin makers Stradivarius and Guaneri used unique blends of oils, resins, beeswax, and propolis, to varnish their instruments and no doubt protect them against mold and microorganisms.

Propolis has a long history of use in folk medicine and was even used as a conventional doctor’s drug in London in the 1600s. It appears to have antimicrobial and anti-inflammatory

activities. In the past, propolis has been taken by mouth to fight bacterial infection including tuberculosis, fungal infection such as Candidiasis, parasitic infections such as malaria, and viral infections such as colds.

A study published in the November 2001 of the journal Mycoses investigated the effect of propolis on 80 strains of Candida yeasts, including:

- 20 strains of Candida albicans
- 20 strains of Candida tropicalis
- 20 strains of Candida krusei and
- 15 strains of Candida guilliermondii

The yeasts showed propolis had a clear antifungal activity and was especially useful against Candida albicans.

While not conclusive evidence has yet been produced, propolis may offer a number of other health/medicinal benefits including the following:

Popolis contains a special compound which may interfere with inflammatory messengers in the body. This potential effect may help to relieve inflammatory conditions such as arthritis.

Propolis may enhance the function of the immune system.

A few studies have shown some activity against cancer cells.

Several other laboratory studies found that propolis may protect the liver from damage by alcohol or environmental toxins.

A recent study of individuals with second-degree burns, compared propolis cream and a prescription burn cream and showed the two were about comparable in preventing infection. But propolis cream promoted earlier healing and seemed to decrease inflammation more than the prescription cream. (I would use homeopathic calendula cream, for burns).

Other studies have also shown that an alcohol-based mouth rinse of propolis may help eliminate oral bacteria: those that can cause dental cavities, gingivitis, and periodontal disease.

For instance, a 1996 study tested the antibacterial properties of propolis and honey against oral bacteria. The antibacterial effects of propolis and honey on oral streptococci were determined using bacterial culture plates. The researchers also clinically checked the short-term antibacterial effect of propolis solution and honey on salivary total bacteria and Streptococcus mutans was tested in 10 volunteers.

Propolis demonstrated an antibacterial effect both in vitro on isolated oral streptococci and in the clinical study on salivary bacterial counts. Honey induced bacteria growth at low concentrations, while at high concentrations honey had an inhibitory effect on bacterial growth in vitro.

Salivary counts of total bacteria and Streptococcus mutans were lower for 1 hour after application of honey.

[Steinberg D; Kaine G; Gedalia I, Antibacterial effect of propolis and honey on oral bacteria. American Journal Of Dentistry 1996;9(6):236-9].

#15 Stolle Milk

This is something that's really useful to know, when antibiotics have run out. Talking about cheap!

Moo!

Suppose there existed a cheap, safe, widely available effective oral "vaccine" that was:

- 100% safe for all who use it;
- can be taken orally and tastes good;
- can be manufactured in virtually every country in the world, with the humble technology available to many third-world nations;
- so cheap that virtually everyone in the world can afford it;
- able to protect against a wide variety of organisms, including viruses, rickettsia, parasites, protozoan, bacteria, mycoplasm, yeast/fungus, amoeba;
- capable of reducing or eliminating allergic reactions to vast numbers of exogenous and endogenous substances;
- boosts the immune system, accelerates healing of injuries, helps repair nervous system damage, burns fat and builds lean muscle, increases vitality and stamina, and elevates mood;
- and might – just might – deal a deadly blow against a number of cancers?

Suppose, moreover, there were over 4,000 clinical studies worldwide which describe the efficacy of this oral vaccine in the treatment of hundreds of different infectious diseases.

Would you be interested to know what it was?

I'll bet you would.

It’s milk protein concentrate. Why do I call it an oral vaccine? Well, that’s the really interesting part. Let Anthony di Fabio take up the story:

> Bessie, our former pet milk cow, lived in a small pasture of not more than three acres. She munched on uncooked grasses during the summer and uncooked dry hay during the winter, licked mineral block, and drank from a rain-filled, surface-drained pond whose waters were loaded with a wide variety of microorganisms. The pond also held frogs, snakes, bugs, worms, snails, and so on. She often drank and urinated at the same time, recycling fluids from the pond even as she drank.
>
> When she was ready to drop her calf, we led her to an old barn that had held forty head of cattle. One’s nose almost stifled from sediments of dust, mold, fungi, and dried manure layered fifty years deep. When Bessie’s calf, Nina (pronounced “Neenya”), was born, she lacked effective defensive mechanisms against the blizzard of microorganisms that assailed her in every cubic inch of the air she breathed, the ground she stood on, or on the inexperienced tongue she extended to various surfaces. Almost by magic, thousands of potentially deadly microorganisms invaded her immature body.
>
> Nina, as with all calves, was also born with a leaky gut! Now pay attention here, because I know that many readers have a leaky gut, a condition where the stomach lining is so thin that whole, undigested protein molecules pass directly from the stomach into the blood stream. Once inside the bloodstream these protein molecules are identified as foreign invaders, and we create antibodies to counteract them. This situation brings about food allergies.
>
> Patients and their doctors both work very, very hard to get rid of the patient’s leaky gut. Their leaky gut is considered the source of many degenerative diseases – or at least a major component of them. But Bessie and Nina had found a way to make the leaky gut
>
> a beneficial survival mechanism! When Nina wobbled to her feet and gently nudged at Bessie’s milk sac, the very first milk to come was colostrum. As Nina prodded the milk sac with her nose and sucked as saliva dripped, she also injected her blizzard of rapidly multiplying microorganisms into Bessie’s teat, and up into Bessie’s milk sac into a portion called the '“cistern.”
>
> Inside Bessie’s cistern specialized cells that had been lying dormant came alive, and they started manufacturing – guess what? – “disease-

> specific antibodies," and immune-boosting "complement," and also flooding her cistern with "immunoglobulins" and "growth factors"!
>
> Very shortly after Nina introduced her stream of potentially dangerous microorganisms into Bessie's teat – then into Bessie's cistern – her mammary biochemical factory stimulated specialized cells that became active and began to create disease-specific antibodies and activated complement that mingled with Nina's first fluids, the colostrum, which Nina sucked back into her leaky gut from Bessie's teats. Bessie was generating vaccines for her calf!
>
> The immunoglobulins, growth factors and these disease-specific antibodies and their helpers, the complement, passed directly into Nina's stomach, and there they attached themselves to whatever corresponding organisms were present inside the gut, killing many. Because of Bessie's leaky gut many of these specially prepared biological agents also passed directly into Nina's bloodstream, and within her blood plasma they attached themselves to whatever microorganisms they'd been designed to destroy and killed them. Never once was Nina placed in danger from the surrounding hostile environment whose every biological niche was filled with a wide variety of deadly microorganisms.

Milk is thus the ideal vaccine! At least dedicated and immune-competent milk is.

According to Herbert Struss, PhD, former Senior Chemist, Food Chemistry Laboratory, Minnesota Department of Agriculture Laboratory Services Division – and also a scientist who was involved in much of the early clinical work testing this wondrously universal vaccine – those interested in "immune milk" (as it is called) during the '60s, made their astounding oral vaccine discoveries when they were trying to answer the question: "What's the survival advantage to being a mammal?"

All animals evolve traits that persist because they have a survival advantage. So what is the advantage to being a mammal? (a mammal, remember, defines a creature with teats or "mammary glands", hence the name) Clearly, Nina's suckling at Bessie's teat, drawing a blood-like liquid called "colostrum" from Bessie's cistern was a possible answer to their question. The survival advantage was simply that an "acquired" or "adaptive" immunity could be transferred from mother to offspring and that this adaptive immunity would extend for some period of time, thus providing the offspring with a distinct survival advantage.

Human milk may not have quite the same importance for survival, as it is with multilayered placentas such as horse, goats, and cattle. But some immunity does pass from the mother to the human child. It is obvious that a breast-fed human baby usually has an advantage over bottle fed. Breast-fed infants have a far lower incidence of early disease and also allergies.

They have been vaccinated by Nature, through nurture. Colostrum is a very special form of milk immune concentrate. See the next section.

But we will continue with our discussion of milk protein concentrate or “hyper-immune milk”, as it is known. Thousands of good scientific papers have been published on its power immune properties. We can start the story with Drs. Wlliam “Pete” Peterson, Barry Campbell, and other colleagues working at the University of Minnesota from 1950 to 1958. They used killed bacterial antigens injected into the teat of a cow (ouch!!) and collected the first ten days’ colostrum. Peterson’s interest was in treating rheumatoid arthritis and allergies, which he did successfully.

In the late 1950s a wealthy businessman called Ralph Stolle, owner-operator of the San-MarGale Farm in Lebanon (Ohio), stepped into the picture. Stolle took an interest in hyper-immune milk, simply as a hobby. He developed a better way of using the cow to provoke a beneficial antigen response, by injecting pathogens into the cow’s bloodstream and then harvesting the milk. Using his personal fortune he gave out thousands of portions of hyper-immune milk and asked the users to report back its benefits. Over a total of some 35 years a total of 8000 people reported a variety of benefits, from joint pain relief (85%) to improvement in fatigue (74%) and lowering cholesterol (71%). This series probably constitutes one of the largest open clinical trials ever carried out.

Stolle went into partnership with the New Zealand Dairy Board to make very pure hyperimmune milk concentrate or “Stolle milk” as it was known. In 1988 it went on sale in Asia for the first time. It is freeze dried and shipped to third world countries, where it is a very valuable vaccination and therapy resource. It can also be rendered homeopathically, so that the cost is even less. In 1999 Spencer Trask purchased Stolle Milk Biologics Inc (SMBI).

It seems to me there is an important dimension that has been missed for decades. The family cow was the manufacturer of home-grown vaccines for humans. Folks lived mainly in a rural environment and milked their own beasts. Not just landowners; even farm workers had access to a Bessie-type vaccine factory. The cow would be infected with all present human pathogens, make antibodies and give these back in the milk.

That’s why most of those country folks were so darned healthy, I believe. They had a constant supply of free and effective personalized vaccines down on the farm. Far better than modern vaccine attempts manufactured in a test tube and under the pretense that everyone is in the same circumstance and health status.

#16 Colostrum

Milk brings me to colostrum, the thin yellowish fluid secreted by the mammary glands at the time of parturition that is rich in antibodies and minerals, and precedes the production of true milk. It is also called foremilk.

Mother's colostrum, incidentally, is vitally important to human babies. It transfers mother's immunity to infection to the child and protects it for the first few weeks of life. There are no antibodies, of course, in milk formulas.

Commercial colostrum is produced by cows that were previously vaccinated with certain kind of pathogens. These cows will then produce antibodies in their blood and then transferred them to milk they produced, just like hyperimmune milk. But the highest concentration of antibodies is transferred in Colostrum, the first drops expressed from the teat.

According to research data, there are around 90 known immunologically-active substances in colostrum, including growth factors, lipids, lactoferrin (iron-binding protein with antimicrobial qualities), cytokines [released from T cells, they inhibit replication of viruses and chemicals (cytotoxins) that kill the infected cell], etc.

Immune Factors in colostrum have been shown to help the body fight off harmful invaders such as viruses, bacteria, yeast and fungus. Each factor plays a specific role in our body's defense against these attackers. In addition, colostrum contains over 20 antibodies to specific pathogens including E coli, salmonella, rotavirus, candida, streptococcus, staphylococcus, H pylori, and cryptosporidia.

In addition to immunoglobulins, medical studies show that PRP in colostrum also supports an under active immune system. Some workers say it has great benefit in treating cancer immunologically. PRP also helps balance an overactive immune system present in autoimmune diseases.

A 2003 study by the Health Sciences Department of the University of South Australia suggests that taking concentrated bovine colostrum supplements could reduce the incidence of respiratory tract infections. The study involved subjects being randomly allocated to consume 60g/day of Colostrum for eight weeks, then examining their occurrence of symptoms. Results from the first week were kept separate from those from the last seven weeks.

The results showed that during the first week of supplementation, there was no measurable difference in symptoms in the proportion of subjects taking the supplement as opposed to those who didn't. During the subsequent seven weeks, however, a much lower percentage of the subjects taking Colostrum reported upper respiratory infection related symptoms.

Symptom duration, however, did not differ. This suggests that the immune benefits achieved by Colostrum can be very good in terms of being a preventative, but that it does not have any measurable effect on an infection once it has already taken hold. That is why it is important to use Colostrum properly, as a preventative, and not depend on it to cure illness.

Another study by Susanna Rokka of MTT Agrifood Research Finland, showed the same thing, testing the ability of Bovine Colostrum to help fight Helicobacter Pylori infections. Helicobacter is a kind of bacteria that is often responsible for Gastritis, Gastric ulcer, and Stomach Cancer.

The research also showed that immune milk combined with lactobacilli that's extracted from fermented cabbage can effectively help in preventing Streptococcus and Helicobacter Pylori. Streptococcus is the most responsible bacteria in causing dental caries from adhering to tooth surface.

So, logically, it makes sense to take colostrum (60 gr daily) alongside whatever alternative antibiotic treatment you are engaged in. What we English call "belt and braces" (I suppose for Americans it would be belt and suspenders).

#17 Marine Quinton

Shockingly, I only got to know of Quinton recently, while staying at Dr. Garry Gordon's home. He'd only just found it too. My ignorance was despite years of practice in Europe, where Quinton has been big business in the healing arts for over a century. A well kept secret, you might say.

What is Quinton? Sea water!

Well, not just any old sea water. It's special vortex water from certain blooms in the oceans, that are very rich in nutrients. It seems to have energetic properties too, above and beyond the mere presence of rare minerals etc.

Let me back up a tad, before we come to the product. The floor of our oceans is indescribably rich in minerals. Think about this: EVERYTHING that ever lived and died goes into the water system, down the rivers and ultimately finds its way to the ocean. Added to that is all the ocean life which lives, dies and is recycled, all the plankton, corals, fish, feces, EVERYTHING, which falls to the ocean floor as organic debris.

There is thick mud at the bottom of the ocean that contains dense nutrients and some minerals that are otherwise incredibly rare, like iridium, osmium, yttrium and so on.

But it doesn't just stay on the ocean bed, lost to the biosystem. Quite the contrary. This nourishing mud is carried around the ocean floor by submarine currents which have only recently begun to be understood. There are certain places where this nutrient deposit wells up to the surface. Giant surges of ocean currents that we call convergences stir up the seas and bring the nutrients back to the bio system.

The polar oceans are classic sites for this. The huge bloom of algae that takes place in the Arctic and Antarctic every year yields a staggering abundance of life where the ocean, literally, changes color due to the density of life it carries. Waters turn red with krill and other plankton.

This bloom is so rich it feeds the greatest animals of all: the whales. So much nutrition is absorbed into the biosystem at these sites in the summer time that whales can double their weight and deposit enough fat or blubber to live on it through the winter months.

But there are the birds, fishes and other animals too, so that the polar oceans literally burst with life every spring when the returning sun ignites the nutrition chain. Given the right wind and currents, then these blooms can be carried far and wide. Scientists report this year there has been a huge upsurge in nutrients and an explosion of sea life off Monterrey, California, from currents brought all the way down from the Arctic.

The polar regions are not the only upsurges, however. In fact they can take place anywhere, if the wind blows the water out to sea; then underwater currents draw up deeper water to fill the gap.

That's what I mean by ocean nutrients, OK?

Enter Messieur Rene Quinton...

Rene Quinton (pronounced cahn-ton) was a Frenchman, a doctor, biologist, biochemist and physiologist. He discovered the healing merits of marine plankton plasma, drawn from deep ocean water upsurges, right at the end of the nineteenth century. It is a little known fact that by 1907 Quinton had established 69 Marine Dispensaries and the product was already saving countless lives throughout the deadly pandemics of the early 20th century (tuberculosis, typhoid, cholera, syphilis, influenza).

When Quinton was finally buried in 1925, his fame had reached such proportions that tens of thousands of men, women and children, not to mention generals, dignitaries and statesmen, attended the funeral. Yet we have never heard of him. How can that be?

Think "Pharma" and think "eliminating the competition", no matter how good the product. The stark truth is I never heard of this cure, all through my medical training. No libraries have editions of his work; no pharmacy has heard of it; no learning institutions teach it. He's been deleted! Especially in his native France.

Quinton's marine plasma (or QMP) was considered so effective for a wide range of common afflictions that is was reimbursed under two French laws, including Social Security. It was, of course, absurdly cheap. But that didn't suit the drug barons, who wanted to peddle their costly and dangerous garbage for huge profits. Through their bribes and malign abuse they got a law passed, requiring QMP to be heat treated. That would effectively ruin its properties and so put Quinton out of business.

However, Quinton manufacture did not fold up. The operation was moved to Spain instead.

Thanks to the more liberal scientific climate in Spain it has survived until today.

Now QMP has arrived in the US and is being sold by Original Quinton of Buena Park, CA.

Properties of QMP

As I write this I have in front of me a translation of the entry for Quinton Marine Plasma in the French Medical Dictionary of 1975 (a kind of PDR). It describes the product as a sterile apyrogenic solution (pH 7.2) of seawater, prepared under aseptic conditions by special processes without rise in temperature or exposure to electric potential or field, in order to preserve its molecular balance and its character as an "alive medium".

The seawater is extracted from a 30-metre depth (zone of solar penetration) under special conditions and from special locations.

It contains 92 elements of the periodic table (all primary and trace minerals).

The entry goes on to add "Rene Quinton showed, in 1904, that QUINTON MARINE PLASMA is identical physically, chemically, and physiologically, to our interior milieu (extracellular fluid) and provides optimum conditions for red blood corpuscles and leucocytes and other blood fractions. It is possible to replace the entire blood volume of an animal with QUINTON MARINE PLASMA without any disorders for the organism." Sacrebleu! Shades of Ringer's solution!

Without any intellectual prowess, one can therefore deduce that QMP is very safe, helps stabilize the internal milieu, provides every conceivable nutrient mineral and provides low concentration homeopathic-type mechanisms for healing. It's a miracle!

Well, it isn't really. It's just Nature at her best. But you can see why Rene Quinton was adored by millions and hated by Big Pharma. Here is a substance that's cheap to manufacture, no patents, is readily available and works wonders as a natural "antibiotic".

The "indications" (reasons to use it) are manifold and include: childhood gastroenteritis, poisoning, malnutrition and eczema; anemia, asthma, exhaustion, anti-aging, dysentery, tuberculosis and atherosclerosis; uterine and vaginal infections; rhinitis, sinusitis, respiratory allergies; skin allergies, dermal infections, histaminic reactions and psoriasis; energy restoration; bioterrain restoration and burns.

The official French pharmacopeia (1975) even lists it as an "antibiotic solvent and carrier".

I have even found that dentists can use it to save teeth, by injecting this healing balm into the surrounding gums. The abcesses and periodontal disease simply disappear.

There are DOZENS more uses.

How To Take It

QMP is taken by injections or orally. When injected, there is an isotonic form, which does not sting. Orally, the full-strength version is fine.

Sprays work well for skin conditions.

It can even be taken by nebulizer, just breathed in. Or I find it very soothing for tired eyes.

Just the whiff of "sea air" is very restorative to me!

It's so safe there are no real limits on the number of ampoules. It's only body fluid, after all!

You can't OD on it like water (too much water will kill you).

QMP comes in 10 ml ampoules. 2, 3, 5, 10 ampoules are fine, depending on the seriousness of the complaint. It's so cheap, cost is not an issue (around $3 an ampoule) Note that Quinton is not FDA approved for injection in the USA.

#18 Chlorine Dioxide Protocol

Next we come to a modern surprise. The story of how it was discovered is interesting. One of those serendipity things!

The idea of using chlorine dioxide was developed by Jim Humble, a gold miner and metallurgist, on an expedition into the jungles of Central America, looking for gold. It was a response to a need to help a member of his expedition who came down with malaria, more than two days away, through heavy jungle, from the next mine.

After many years of experience, Humble always carried "stabilized oxygen" (sodium chlorite) with him on such expeditions, to make local water safe to drink. Facing the possibility of a quick loss of life, he gave it to the stricken man. To everyone's amazement, the patient was well within a few hours. That seemed like a miracle, but Humble wanted to better understand what had just happened.

Over the course of several years, Jim Humble figured out that what made stabilized oxygen so effective in some malaria cases, was not the oxygen at all, but the trace amounts of chlorine dioxide it contained. Further research led him to come up with a way to produce hundreds, if not thousands more units of chlorine dioxide than what is found in stabilized oxygen. This is done by using a higher concentration of sodium chlorite (28% vs. 3% for stabilized oxygen), in conjunction with an activator.

Humble went on the have great success using his new formula which, unscientifically, has been named "Magic Minerals Solution" (there are no minerals in it). But MMS (for short) does seem to work, though I doubt some of the claims. It is said to been successful in helping over 75,000 people in several African nations – including Uganda and Malawi – rid themselves, primarily of malaria, but also hepatitis, cancer, and AIDS.

Anyone can be overloaded with toxins. Most people probably are but won't admit it or, more likely, don't know it. Others would prefer to think they're not. If your health is not perfect, you're habitually low on energy, have trouble keeping your weight down or your blood pressure in the normal range, or constantly dealing with inflammation or pain, or indeed have any medical condition that is adversely affecting your health, then there's likely to be a toxin, heavy metal, virus, bacteria, fungus, or parasite playing a part. Mainstream medicine will typically respond by loading you up with additional pollutants, many of which indiscriminately kill healthy tissue while going after “the bad guys” to deal with the symptoms. Not so with chlorine dioxide. It only acts on anything harmful. Miracle or not, the effects can be amazing!

Over the next few s, the MMS protocol will be described. When followed, it will produce and distribute chlorine dioxide to your red blood cells, which are the most effective and intelligent pathogen killers known to Nature, although your white blood cells are assumed to do all the work.

But first, a little background on the chemistry. Chlorine dioxide and chlorine are not the same. Chlorine is a chemical element. In ionic form, chlorine is part of common salt and other compounds, and is necessary to most forms of life, including human. A powerful oxidizing agent, it is the most abundant dissolved ion in ocean water, and readily combines with nearly every other element, including sodium to form salt crystals, and magnesium, as magnesium chloride. Chlorine dioxide is a chemical compound that consists of one chlorine ion bound to two ions of oxygen. Oxidizing agents are chemical compounds that readily accept electrons from “electron donors.” They gain electrons via chemical reaction. This is important because relative to chlorine dioxide, all pathogens are electron donors.

Chlorine dioxide is extremely volatile. You might call it “hot tempered,” but in a very beneficial way. This volatility is a key factor in chlorine dioxide’s effectiveness as a pathogen destroyer. The compound is literally explosive, so explosive, it’s not safe to transport in any quantity. Therefore, it is common practice to generate chlorine dioxide “on site” at the point of use. Most chlorine dioxide production is done on a scale that would prove deadly for individuals, for example, in municipal water treatment systems, where it is beginning to replace chlorine because it produces no carcinogenic by-products.

Chlorine dioxide is approved by the Environmental Protection Agency in safely removing pathogens and contaminates like anthrax. So you know it must be effective. However, the concentrations used in such applications can vary from 500 to over 6,000 parts per million (ppm), which would clearly be deadly to an individual. Using the MMS protocol you will produce chlorine dioxide around 1 ppm. You will use the MMS solution, which is safe to transport, to make nature’s harmless pathogen destroyer.

The MMS solution is 25% sodium chlorite in distilled water. You can produce chlorine dioxide with a single drop, when an “activator” of vinegar, lemon juice, or a 10% solution of citric acid is added. Citric acid is recommended because of its simplicity. The natural pH of sodium chlorite is 13. Adding vinegar, lemon juice, or citric acid creates about 3 mg of unstable but still harmless chlorine dioxide.

The Process

Let’s talk a bit more about how and why chlorine dioxide works by giving the immune system a new lease of life. Volatility is what makes chlorine dioxide so effective when it contacts pathogens. As we’ve already mentioned, chlorine dioxide is a safe and effective disinfectant in many municipal water delivery systems, hospitals, and even in bio-terrorism response. It stands to reason that chlorine dioxide would be just as effective working in the waters of the human body at the appropriate dose.

Chlorine dioxide’s extreme volatility prevents pathogens from developing a resistance, mainly because when they “clash,” the pathogens no longer exist. Yet, healthy cells and beneficial bacteria remain unaffected. While normal levels of oxygen in the blood cannot destroy all of the pathogens present under disease conditions, delivery of chlorine dioxide changes everything. “Halt! Surrender Your Electrons, Now!” When a chlorine dioxide ion contacts a harmful pathogen, it instantly rips up to five electrons from the pathogen, in what can be likened to a microscopic explosion… harmless to us, but terminal for pathogens. The pathogen – an electron donor – is rendered harmless due to the involuntary surrendering of its electrons to the chlorine dioxide – an electron acceptor – and the resulting release of energy. Oxidized by the chlorine ion, the former pathogen becomes a harmless salt.

This process benefits a body that has become toxic. Throughout the body, anywhere chlorine dioxide ions – transported via red blood cells – come into contact with pathogens, the pathogens give up their electrons and cease to exist.

The chlorine dioxide-armed cells only “detonate” on contact with pathogens, which include harmful bacteria, viruses, funguses, toxins, heavy metals, and parasites. All of these will have pH values that are out of the body’s range of good health. They will also have a positive ionic charge. The chlorine dioxide equipped cells do not oxidize beneficial bacteria, or healthy cells, as their pH levels are 7 or above, and hold a negative ionic charge. Chlorine dioxide ions will oxidize – meaning vaporize – diseased cells… anything that is acidic, with a positive ionic charge. If the chlorine dioxide ions encounter no pathogens or other poisons, they deteriorate into table salt and in some instances, hypochlorous acid, which the body can also use.

A Pathogen Terminator

Research has proven chlorine dioxide to be much safer than chlorine, as it is selective for pathogens when used in water. Furthermore, it does not create harmful compounds from other constituents in the water as chlorine does. Numerous scientific studies have demonstrated that chlorine – part of the halogen family of elements – creates as least three carcinogenic compounds when it enters the body, principally Trihalomethanes (THMs). There has been no such evidence of harmful compounds being produced from chlorine dioxide. This is why, in 1999, the American Society of Analytical Chemists proclaimed chlorine dioxide to be the most powerful pathogen killer known to man. It has even been used to clean up after anthrax attacks.

A Journey into Chemical Alchemy

Once it is introduced into the bloodstream, chlorine dioxide performs a highly energetic acceptance of up to five electrons when it comes across any cell that is below a pH value of 7.

This means that diseased cells are essentially vaporized (i.e., "oxidized") while healthy cells are unaffected.

This is how it happens. Red blood cells that are normal carriers of oxygen throughout the body do not differentiate between chlorine dioxide and oxygen. Therefore, after you have swallowed the MMS/chlorine dioxide-rich solution, red blood cells pick up chlorine dioxide ions as they pass through the stomach wall. When the red blood cells, armed with chlorine dioxide, encounter parasites, fungi, diseased cells or anything that has a pH below 7 and a positive ionic charge, the "aliens" are destroyed along with the chlorine dioxide ion.

If the chlorine dioxide doesn't hit anything that can set it off, it will eventually deteriorate, by losing an electron or two. This will allow it to be converted into hypochlorous acid. This compound kills pathogens and even cancerous cells. Hypochlorous acid is so important that its diminished presence in the body is described medically by the term 'myeloperoxidase deficiency'. Many people are afflicted by this condition. The immune system needs a great deal more hypochlorous acid when disease is present. Facilitated by the MMS solution, chlorine dioxide delivers it in quantity.

The most salient point to know is that chlorine dioxide has 100 times more energy to do what oxygen normally does, and yet, will not harm healthy cells. By the way, if you are totally healthy and have nothing in your body that is at an acidic level below 7, there are no ill effects from taking chlorine dioxide at the appropriate dose. However, your stores of hypochlorous acid will be increased.

MMS works best to destroy pathogens that may be present in the body, when 2 or 3 mg of free chlorine dioxide are in the solution at the time it is swallowed. However, the body is supplied with chlorine dioxide in a "timed release" manner lasting about 12 hours. Be aware, that before you feel better, it is likely you will feel ill in one way or another.

Nausea

The nauseating feeling that you may possibly experience, especially if you take too big a dose, is the result of chlorine dioxide encountering and destroying a large number of pathogens. We are generally oblivious to the pathogens that are present in our bodies, but there is an increasing awareness in the medical and scientific literature of their importance to our health, or more to the point our ill health. Since they build up over time in various organs of the body, they generally affect our health slowly and cumulatively. If chlorine dioxide takes them out too suddenly, the result will be a dramatic reaction. However, the time it takes to clear the pathogens and toxins is much less than it took for them to accumulate. It has probably taken many years, possibly almost a lifetime, for some of them to accumulate.

As a person always feels unwell when they contract a case of, say, acute hepatitis, dengue fever or Lyme disease, they may continue to feel unwell for a while, in which case it will be difficult to tell whether it is the condition or the chlorine dioxide that is causing the unwell feeling. However, if the condition is treated with chlorine dioxide in its early stages, the pathogens will be killed off quickly as they are still in the blood stream and therefore available to be attacked easily by the circulating chlorine dioxide-rich red blood cells. Under these circumstances, the symptoms should be over very quickly.

In chronic conditions, especially Lyme disease and dengue fever, nearly every organ of the body can be affected to a greater or lesser extent, which is why the symptoms can be so variable, and the person can feel so ill. This variability is almost diagnostic of such conditions, especially as there are no satisfactory blood tests to prove the diagnosis. The reason for this is that there may be a number of different organisms involved, each having a different life cycle. In these cases it will take a little longer to achieve a resolution of the condition, as each organism is destroyed in its own time.

Years of “leeching” from dental amalgams can “innocently” deposit enough mercury in one’s system to steal innocence, rob vitality, and erase precious memories. Lead can accumulate over the years from atmospheric exposure. Dislodging and vaporizing either or both of them may feel uncomfortable for a short time compared to the time it took for them to accumulate.

If there is nothing for chlorine dioxide to encounter, it deteriorates into constituents that are totally non-toxic. Nothing poisonous is left behind to build up, as is the case with many medical protocols. Medical treatments currently provide you no way of removing the poisons when they don’t work. Chlorine dioxide, on the other hand, lasts long enough to do its job, and then the amount that does not furnish the immune system with needed ions becomes nothing more than micro amounts of salt and water. The chlorine dioxide has just a few minutes to do its job, and then it no longer exists, leaving nothing behind that can build up, or do additional harm.

The Procedure

The procedure for taking MMS is simple. All you need is your bottle of MMS, a clean, empty, dry glass, an eyedropper and the activator citric acid.

Note: When following the instructions below, keep this paragraph in mind. Always activate the MMS drops with one of the food acids, either lemon juice drops, or limejuice drops, or citric acid solution drops, the citric acid drops being the simplest.

Always add 5 drops of citric acid to each 1 drop of MMS, mix in an empty dry glass and wait at least 3 minutes, then add 1/3 to 2/3 glass of water and drink it. (You can expand the 3 minutes to 10 before drinking.)

Repeat this dose in between one and two hours, ideally doing all of this after your evening meal, possibly starting about ¾ hour after you have eaten, as it can sometimes make some people sleepy, apart from which your body does most of its detoxifying during the night.

Start modestly with as little as 1 drop of MMS plus 5 drops of citric acid on your first day (never forget to wait at least 3 minutes for the mixture to react to create chlorine dioxide, which will turn yellow and smell of chlorine, and repeat the dose in one to two hours). Take your time and do not rush. You could stay on this low dose for a few days, and then increase the number (2 and 10, 3 and 15, etc) on subsequent days, but I repeat – TAKE YOUR TIME. There is no point in going higher than 15 and 75 respectively, but it is rare for anyone to reach this level.

Your body will tell you when you've reached the optimum dosage for you, and, if in doubt, drop the next dose. Clearing may be a bit uncomfortable, but it need not be intolerable. You may feel like you've been through a battle, and, in a sense, you have. However I suggest that, if you develop any symptoms of any sort, assume you have taken too high a dose, so put up with the effect, take the antidote, and drop the next dose, possibly even not taking any for a day.

Before you can be healthy again, you need to destroy toxins, pathogens, parasites and anything harmful to you. In order to do so, they have to be uprooted and released from their "strongholds" in your body tissue. You have no idea what they are or where they are. Remember they may be buried deep in your organs.

You don't have to reach your maximum tolerated dose. Whatever dose you use will have its value, but the higher you can comfortably reach the better. It is just that the whole process will take that much longer.

This gentle approach applies to any chronic condition, and especially if you want to clean up your body. However, if you develop an acute medical condition such as dengue fever or malaria, for example, start straight in at at least 5 drops of MMS to 25 drops of citric acid, although you could possibly start at 8 drops of MMS to 40 drops of citric acid, and don't forget to repeat the dose in between one and two hours.

With any luck you will feel remarkably better by the next day. If you are not quite symptom free, repeat the same the next day, increasing the dose by about one third. In an acute situation, you can take three doses a day, each one repeated one to two hours later.

Antidote. If you develop any symptoms you don't like, assume it is the chlorine dioxide working too hard within you. To clear these symptoms, either take a few doses of ½ teaspoonful of sodium bicarbonate in a glass of water or a few grammes of vitamin C in water. Don't take both. Then either don't take a dose of MMS for 24 hours or drop the next dose and gradually work back up again.

IMPORTANT. Please be aware that, as I have already said, chlorine dioxide is a very potent chemical and literally destroys anything potentially harmful it comes across. Whatever dose you take, it will do its job. It is understandable that you want to reach as high a dose as possible as soon as possible, but I would encourage you not to think like that. As most of us have accumulated a lot of undesirable things within our bodies over the years, some of which may now be causing a major illness, it is not unreasonable to suggest that it may take some time to get rid of it all, possibly six months or more.

A number of people have reported to me that they did not feel any nausea, nor did they vomit, but started to feel generally unwell, or some of their old symptoms started to come back a bit, when they had reached a certain dose of chlorine dioxide. Fortunately they rang me. My advice was that they should go back to a lower dose, possible even not taking any chlorine dioxide for 24 hours before restarting at the lower dose. This approach has worked in every case. Unlike an antibiotic, nothing can develop a resistance to chlorine dioxide, as has already been said.

So what I am really trying to say is that, if you have ANY undesirable effects, even if you become a bit more tired than before, ASSUME that the chlorine dioxide is being too active for your body's current ability to eliminate the toxins. Take a lower dose next time, and be prepared to stay low for a while. Please don't overdose. It will only make you feel unnecessarily ill.

Overall you may feel the effects, but this is a good thing. You will also feel healthy again. Any sick feeling will be TEMPORARY, a small price to pay for the longer-term possibility of lasting restored health, no matter what stage of life you happen to be currently experiencing. When the clearing is done, you won't need to take the maximum dose. You can go on a maintenance application (1 or 2 drops of MMS) to keep your insides pathogen free and your immune system strong or take a dose every so often. I hope you have found this information helpful. In summary, when MMS is combined with citric acid it produces chlorine dioxide, which, at the appropriate dose is a safe and effective way to boost your immune system and eliminate a full range of harmful organisms, toxic metals and chemicals which may well be making you ill.

Dr. Patrick Kingsley. February 2008.

Next, while we are on the topic of simple chemical substances, let’s look at two others: first, hydrogen peroxide…

#19 Hydrogen Peroxide

The bleach story continues!
Hydrogen Peroxide: A Powerful Antibiotic

Hydrogen peroxide is like water (H_2O) with an extra oxygen atom added (H_2O_2). It's the extra oxygen and the power to release it right in the tissues that makes this substance a powerful medicine. Oxygen is good for us and bad for a great many pathogens, notably anaerobic bacteria, yeasts and viruses. It also cripples cancer cells and renders them vulnerable to the immune system.

Nature supplies us with abundant hydrogen peroxide. Every time it rains, falling drops of water bring down ozone from higher in the atmosphere in the form of hydrogen peroxide. The extra life-giving oxygen is undoubtedly one of the reasons plants do better when watered from the sky than from a watering can. It is also why rain is so refreshing. Don't hide indoors! Go out and get the extra oxygen: it's Nature's gift of energy!

In fact hydrogen peroxide from the sky is so good that farmers are known to mimic it by spraying their crops with diluted hydrogen peroxide. Feed it to your plants, using one ounce of 3% hydrogen peroxide solution to a quart of water. They'll love it!

Unfortunately, because of current atmospheric pollution, much of the peroxide never reaches the ground. It gets used up interacting with chemicals in the atmosphere and in urban areas this has lowered oxygen levels from the 20% you were taught in school, down to as little as 10%.

Despite its health-giving properties, our oxygen needs are not being met. The oxygen-generating rain forests are being destroyed worldwide, which further reduces available oxygen. Chlorination of drinking water removes oxygen. Cooking and over-processing of our foods lowers their oxygen content. Unrestrained antibiotic use destroys beneficial oxygen-creating bacteria in the intestinal tract.

Dr. Johanna Budwig of Germany has shown that for proper cellular utilization of oxygen to take place, our diets must contain adequate amounts of unsaturated fatty acids, rich in pi electrons. Unfortunately, the oils rich in these fatty acids have become less and less popular with the food industry. Their very reactive nature means they require more careful processing and it gives them a shorter shelf-life. The food industry, which has no apparent interest in human health, has therefore turned to the use of synthetic fats and dangerous processes like hydrogenation.

A Bit Of Biochemistry

Ozone is an unstable coupling of 3 oxygen atoms and it easily gives up the third oxygen atom which, once alone, becomes a violent reactor and can cause damage. That's why ozone is potentially hazardous. Hydrogen peroxide also has an extra oxygen atom, which is gives up easily, reverting to water. It too is potentially dangerous—but only if you misuse and abuse it, which of course is true of any medicine!

The fiery "singlet oxygen", as it's called, that is given up in the tissues is what creates the powerful medicinal properties of hydrogen peroxide. It's deadly to pathogens and deadly to cancer cells, yet does our bodies little harm.

The bubbling you see when hydrogen peroxide comes in contact with a bacteria-laden cut or wound is the oxygen being released and bacteria being destroyed.

You've probably understood the anti-oxidant story and you no doubt take antioxidants to protect yourself. However the missing half of the story, which rarely gets told by half-educated health "experts", is that we need oxygenation in our bodies. Hydrogen peroxide is the exact substance by which our own white blood cells kill bacteria, molds (yeasts) and viruses. Cells release it on the spot and then rapidly re-absorb and neutralize it when it has killed the pathogen.

If this ability was totally swamped by excessive anti-oxidants, then we wouldn't be able to fight off infections. Please remember this point. The ability of our cells to produce hydrogen peroxide is essential for life. H_2O_2 is not some undesirable by-product or toxin as orthodox medicine would have you believe, but instead a basic requirement for good health.

Hydrogen peroxide is involved in many other important processes in the body. For example, vitamin C helps fight infections by enhancing hydrogen peroxide production, which in turn stimulates the production of prostaglandins. Also lactobacillus found in the colon and vagina produce hydrogen peroxide. This destroys harmful bacteria and viruses, preventing colon disease, vaginitis, bladder infections and a host of other common ailments *(Infect Dis News Aug.8,91:5)*. It is required for the production of

thyroid hormone and sexual hormones. *(Mol Cell Endocrinol 86;46(2): 149-154) (Steroids 82;40(5):5690579)*. It stimulates the production of interferon *(J Immunol 85;134(4):24492455)*. It dilates blood vessels in the heart and brain *(Am J Physiol 86;250 (5 pt 2): H815-821 and (2 pt 2):H157-162)*.
It may improve glucose utilization in diabetics *(Proceedings of the IBOM Conference 1989, 1990, 1991).*

The good thing is that hydrogen peroxide is dirt cheap and readily available. BUT YOU MUST TAKE CARE TO USE ONLY THE CORRECT FOOD GRADE FORM. More about that later.

We don't yet fully understand the complete workings of hydrogen peroxide but we do know that it is a prolific source of free oxygen. 1 pint of food-grade (35% solution) contains the equivalent of 130 pints of oxygen. 1 pint of 3% hydrogen peroxide found at the local drugstore contains 10 pints of oxygen. And 1 pint of the 6% solution (the kind used to bleach hair) contains 20 pints of oxygen.

We also know that when H_2O_2 is taken into the body (orally or intravenously) the oxygen content of the blood and body tissues increases dramatically. Early researchers felt these increases were simply due to the extra oxygen molecule being released. This is unlikely. Only very diluted amounts of H202 are ever introduced into the body. The small amount of oxygen present couldn't be solely responsible for the dramatic changes that take place.

Dr. Charles Farr, a strong proponent of intravenous use, has suggested another possible answer. He believes that hydrogen peroxide stimulates enzyme systems throughout the body. This triggers an increase in the metabolic rate, causes small arteries to dilate and increase blood flow, enhances the body's distribution and consumption of oxygen and raises body temperature *(Proceedings of the International Conference on Bio-Oxidative Medicine 1989, 1990, 1991).*

A Little History

The therapeutic benefits of hydrogen peroxide were reported as far back as 1920. The prestigious English medical journal, The Lancet, carried the story that IV hydrogen peroxide was used to successfully treat pneumonia in the flu epidemic following World War I.

In the 1940's Father Richard Willhelm, a pioneer in promoting peroxide use, reported its extensive use in treating everything from bacterial-related mental illness to skin disease and polio. Father Willhelm founded of "Educational Concern for Hydrogen Peroxide" (ECHO, a nonprofit organization dedicated to educating the public on the safe use and therapeutic benefits of hydrogen peroxide).

Much of the interest in hydrogen peroxide waned in the 1940's when prescription medications came on the scene. Since that time there has been little economic interest in funding peroxide research. After all, it is inexpensive and non-patentable, so worthless to drug companies, who pursue profits above patient care. Yet despite this lack of apparent interest of orthodox medicine, over 7,700 articles relating to hydrogen peroxide have been written in the standard medical journals.

Thousands more, involving its therapeutic use, have appeared in alternative health publications. The number of conditions helped by hydrogen peroxide is astounding. The reported dangers and side effects are few and often conflicting.

Again, I repeat, that it is quite safe if handled properly. But deadly if not.

Only 35% Food Grade hydrogen peroxide is recommended for internal use. See overleaf for notes about grades of hydrogen peroxide.

Even at this concentration, however, hydrogen peroxide is a very strong oxidizer and if not diluted, it can be extremely dangerous or even fatal. Any concentrations over 10% can cause neurological reactions and damage to the upper gastrointestinal tract. There have been two known fatalities in children who ingested 27% and 40% concentrations of H202.

Recently, a 26 month old female swallowed one mouthful of 35% H202. She immediately began vomiting, followed by fainting and respiratory arrest. Fortunately, she was soon under emergency room care and although she experienced erosion and bleeding of the stomach and esophagus, she survived the incident. When she was re-examined 12 days later, the areas involved had healed *(J Toxicol Clin Toxicol 90;28(1):95-100).*

What Diseases Can Be Treated With Hydrogen Peroxide?

All infections will benefit, whether viral, bacterial, fungal (yeasts and molds) or parasitic. The list would be long but would include colds, influenza, Herpes, sore throats, gum (periodontal) disease, infected wounds and burns, MRSA certainly, pneumonia, hepatitis, infectious mononeucleosis, Lyme, vaginal Thrush, bowel disorders (proctitis, colitis, Crohn's etc), prostatitis, trichomoniasis, systemic Candiasis and cystitis.

Veterinary parasitology also suggests it is effective against toxoplamosis, which is otherwise almost impossible to treat *[Veterinary Parasitology, Volume 153, Issues 3-4, 31 May 2008, s 209-213].*

Even claims have been made for the successful treatment of AIDS (these are never investigated for their worth; the doctors are just treated as liars and quacks).

How Is Peroxide Used?

In serious situations and if local political matters permit, there is no question that hydrogen peroxide (and ozone) therapies are best administered by a qualified health professional, who know what he or she is doing.

But you can do it for yourself, if you can get the correct grade of hydrogen peroxide. For the avoidance of confusion, here is information about all the common forms of hydrogen peroxide. Be sure you understand what you are buying, if you want to use it for self-therapy.

Grades of Hydrogen Peroxide

Hydrogen peroxide is available in various strengths and grades.

1. 90%: This is used as an oxygen source for rocket fuel!

2. 3.5% Pharmaceutical Grade: This is the grade sold at your local drugstore or supermarket. This product is not recommended for internal use. It contains an assortment of stabilizers which shouldn't be ingested. Various stabilizers include: acetanilide, phenol, sodium stanate and tertrasodium phosphate.

3. 6% Beautician Grade: This is used in beauty shops to color hair and is not recommended for internal use.

4. 30% Reagent Grade: This is used for various scientific experimentation and also contains stabilizers. It is also not for internal use.

5. 30% to 32% Electronic Grade: This is used to clean electronic parts and not for internal use.

6. 35% Technical Grade: This is a more concentrated product than the Reagent Grade and differs slightly in that phosphorus is added to help neutralize any chlorine from the water used to dilute it.

7. 35% Food Grade: This is used in the production of foods like cheese, eggs, and whey-containing products. It is also sprayed on the foil lining of aseptic packages containing fruit juices and milk products. Only 35% Food Grade hydrogen peroxide is recommended for internal use. At this concentration, however, hydrogen peroxide is a very strong oxidizer and if not diluted, it can be extremely dangerous or even fatal.

You must handle it without getting it on your skin. Direct contact will burn the skin. Rubber gloves from Home Depot or similar are recommended. If you do get any on your skin, wash it off immediately with copious water.

It needs to be diluted down to 3%. Diluting 10 times would result in 3.5%, so if we dilute it 1 in 12, that's close to 3%. You can do this easily by pouring 1 ounce of H_2O_2 into a pint jug or jar, then adding 11 ounces of distilled water. This will make 12 ounces of 3%.

I recommend that you transfer the H_2O_2 as you need it to a small glass eyedropper bottle, which you can get from your local drugstore. Leave the rest in the tightly sealed bottle it came in and store it at the back of the refrigerator, where it will be quite stable chemically, for long periods (make sure children cannot access it).

Suggested Dose Regime

Based on years of experience and thousands of cases reports, the following dose regime is known to work well and should be right for you.

Drops can be added to 6- 8 ounces of liquid in a glass, either more distilled water, fruit juice, milk or aloe vera (juice or gel). DO NOT USE TAP WATER: the chlorine in tapwater would deactivate the hydrogen peroxide.

Hydrogen peroxide must be taken only on an empty stomach. If there is food in the stomach, the reaction on any bacteria present may cause excess foaming, indigestion, and possibly even vomiting. Additionally, some animal research indicates that when it is given orally, peroxide combines with iron and small amounts of vitamin C in the stomach, hydroxyl radicals are created *(J Inorg Biochem 89;35(1):55-69).*

Day	Number of Drops	Time Per Day
1	3	3
2	4	3
3	5	3
4	6	3
5	7	3
6	8	3
7	9	3
8	10	3
9	12	3
10	14	3
11	16	3
12	18	3
13	20	3
14	22	3
15	24	3
16	25	3

Please don't go above 25 drops. It may be counter-productive.

Transplant patients should not take hydrogen peroxide. It strongly stimulates the immune system and could, theoretically, cause a rejection of the organ. If you feel nauseated and sick, this could be a healing crisis. As you detox and kill off lots of pathogens in your system, your body may not be able to remove the breakdown products fast enough. The answer is just to persist until the toxins are cleared.

Try to keep going at the dose you are on; just don't increase it until the reaction wears off. Then resume the climbing dose (maximum, 25 drops, 3 times a day).

You should also take vitamin E (alpha and gamma mixed forms, not synthetic alpha-tocopherol). It enhances oxygen metabolism and is quite safe, whatever the current scientific frenzy to discredit it.

You should also actively engage in pre- and probiotics (section #40a), to start repopulating your body with good friendly bacteria.

You can get food grade 35% hydrogen peroxide online at this website:

http://www.dfwx.com/h2o2.htm

Other Administration Routes For Hydrogen Peroxide

1. Three tablespoons mixed with a quart of distilled (or any non-chlorinated) water makes a good vaginal douche, for those stricken with bacterial or yeast (Candida) vaginitis.

   Remember the need to re-populate with good flora. The pH of the vagina is important.

   It is normally kept slightly low by malic acid. This deters bacteria and yeasts. I have heard (but don't recommend) women douching with yoghurt to increase friendly flora.

   One case was found with a raspberry seed stuck in her cervix: she'd used supermarket sweetened fruit yoghurt, which was very silly!

Enemas. The above formula can also be used effectively for colon infection (Candida, Giardia and other).

2. It makes an effective full strength mouthwash or can be mixed with baking soda for toothpaste.

3. It can be used full strength as a foot bath for athlete’s foot. (Diabetics have found relief from circulation problems by soaking their feet in 1 pint of 3% peroxide mixed with 1 gallon of warm, non-chlorinated water for 30 minutes nightly.)

4. A tablespoon added to 1 cup of non-chlorinated water can be used as a nasal spray. Depending on the degree of sinus involvement, one will have to adjust the amount of peroxide used. If it causes excess catarrhal discharge, reduce the proportions.

#20 Hydrochloric acid Injections

What? That's right. Only 1:1,000 strength! It's more barefoot doctoring from the days before fancy antibiotics. Burr Ferguson M.D. of Birmingham AL first wrote about it around 1925. Dr. Ferguson had been a battle surgeon during World War 1 and he had seen the wounded die by the hundreds from infections. He quickly found that when 10 mls of 1:1,000 hydrochloric acid is given IV, there is a rapid and significant increase in the white cell count. Infections melted away and fever subsided.

You can treat virtually any infection with this simple protocol. It's quite safe!

Begin as soon as there are signs of fever. Shots are given daily. Usually after a week the fever is cleared and the patient feeling well.

The doctor was William Howell M.D. of Lexington, Tennessee told of the following case in a 1930s journal Medical World:

The patient was age 15 and she was delivered of a very large baby that lived only for two hours. She lived in a log cabin in the woods. The delivery was done with the best aseptic condition that could be had in a log cabin. The patient weighed only 90 pounds. There were some small lacerations. On day three after delivery there was a message that the patient had a chill and high fever. She lived in a river bottom and I was in hope that it was malaria and I sent quinine. On the fifth day there was another message telling of the grave condition of the patient. In going into the sick room I saw at once that the message had been urgent. The girl was delirious, temperature was 106, pulse was 140, respiration was 40 and there was a discharge from the vagina with a fetid odor. Every other case I had seen like her had died of the infection.

I had some one in 1,500 hydrochloric acid with me but had feared to use it. I injected 10 cc of 1 in 1,500 hydrochloric acid into a vein with much trepidation. The following minutes were anxious ones for me as I hardly knew what to expect. I had never heard of hydrochloric acid being used to treat puerperal sepsis. I was thinking of the warnings of the fatal results of an injection of acid in the vein. I was holding the radial pulse when all of a sudden there was sweat on the patient's neck and forehead and a slowing of the pulse. In a few more minutes the patient was bathed in a profuse perspiration and there was a stop to the chatter of her delirium. Thirty minutes after the injection she was conscious and I asked her how she felt. She said that she felt much better. She wanted to go to sleep.

Within an hour of the injection, her temperature had dropped to 103, her pulse had dropped to 100 and respiration had decreased to 22. On the next four days I repeated the injections. On day six, temperature was 99, pulse was 72 and respiration was 19. Two days later I was called and told that her fever had returned. Her temperature was 101 and there was a free discharge from the vagina. I gave her one more injection of 10 cc of acid.

All signs of the infection she had had completely disappeared.

The story was later retold by Barry Groves in Townsend Letter for Doctors, December 2001 issue.

Obviously this is one remedy that should probably only be done by an MD or RN. But in an emergency, in the jungle, anywhere, just do it!

I've administered hydrochloric acid to myself IV on two occasions. It worked and was entirely without side effects. If you end up with MRSA or any other dangerous or intractable infection, don't trust to antibiotics to save you: get some holistic practitioner to do this protocol for you.

I need hardly say that I believe it would be wise for decent MDs to re-learn these older techniques which certainly worked. Moreover they were cheap and safe (that was partly the problem, of course: no major drug company would ever promote such profit-less remedies).

One more case from Dr. William Howell, then we're done:

The acid gives the same happy results in nephritis. Give an injection every day---in five or six days the albumin diminishes until there is none. I saw a case of acute nephritis September, 1933, following scarlet fever. A boy, nine years of age, swollen till he looked as though he would burst. I tried everything I could get at; salt-free diet, potassium citrate in large doses, milk diet. The hospital authorities said his urine showed four plus albumin. After six weeks he showed no improvement.

I decided to try the acid. I gave him 3 c.c. of a 1:500 solution in the gluteal muscle every day, alternating hips each day. In seven days the albumin began to decrease. After twenty-one injections he was entirely well. I have reports from him almost weekly; no albumin.

I saw him May 6th, 1934. He was out on the road with other children, was just as busy in their games as they were, showed no signs of his former trouble, looked well, ruddy like the others. Since that time I saw another case of acute prostatitis, urine loaded with albumin. After six injections his urine was entirely cleared up. I have given him about twenty-five injections. He sleeps well at night; no more bladder trouble.

I have a man under my care who had tuberculosis of the lungs. The bladder became involved. I gave him an injection every day for thirty days. He missed his fever, all bladder symptoms are gone, eats plenty, sleeps well, has had no fever in ten months. His weight has gone up from 140 to 180 lbs. I am still giving him an injection once a week. He is doing light work, coughs very little. He has taken more than one hundred injections with never the slightest harm.

#21 Colloidal Silver

If you've been into alternative and holistic medicine for any length of time. You must have heard of colloidal silver is an anti-septic. You've probably heard that Alexander the Great's troops kept their water supply in silver vessels and how wagon train pioneers, travelling through the Wild West, used to put a silver dollar in the water barrel, to keep the water pure.

There is really no argument that it is effective. In fact, silver has even reached the mainstream in the form of silver-soaked bandage line from Curad. According to the bandage company: "Laboratory testing showed that silver reduced bacterial growth like Staph. aureaus, E. coli, E. hirae and Pseudomonas aeruginosa -- a powerful germ that does not respond to many antibacterials -- in the dressing for 24 hours."

Even washing machines are now, for better or worse, using silver ions to kill bacteria.

Meanwhile, silver-coated urinary catheters are used to help reduce urinary tract infections (UTIs) -- studies showed they reduced the incidence of UTI by 47 percent.

Silver inhibits the growth of bacteria by deactivating the bacteria's oxygen metabolism enzymes. In turn, this destroys the bacteria's cell membranes, stopping the replication of the bacteria's DNA. *[Source- Acupoll Precision Research, April 2003 Beiersdorf, Inc.]*

The real problem is not credibility with colloidal silver antisepsis. The problem is that there's an awful lot of junk products, manufactured in garage and kitchen top businesses, with small and inadequate electrolysis machines. These are really not potent enough to have any noticeable effect. Yet the people who peddle these goods like to quote science to give them credibility.

Don't be fooled, make sure you get a good product. It's not just a matter of the particle count but particle size also (surface area)—and ionic silver content.

Silver ions are silver atoms which have an electron missing in the outer shell. Silver particles are metallic silver consisting of clusters of silver atoms and can range in size from less than a nanometer up to 1,000 nanometers (1 micron). Silver ions will combine with chloride ions readily where they are present. The human stomach contains a strong solution of hydrochloric acid. Silver chloride forms immediately in the stomach when silver ions enter it. The same thing happens to silver ions in the bloodstream by virtue of the high chloride content due to the presence of sodium and potassium chloride. Silver ions, therefore are of little value inside the body as they are quickly combined with the available chloride ions to form worthless silver chloride.

Let me suggest a product called Advanced Colloidal Silver. It has approximately 20% ionic silver and 80% silver particles with a mean average of about 2 nanometers. This small particle size, combined with small particles results in a vastly larger surface area of silver for a given concentration. Because of this, ACS is potentially more effective than those products with much larger particles and much higher ppms.

The difference between good and bad products is illustrated by a study carried out by Dr. Ron Leavitt of BYU University, and reported by Deseret News, Tuesday, May 16, 2000.

According to Leavitt "The data suggests that with the low toxicity associated with colloidal silver, in general, and the broad spectrum of antimicrobial activity of this colloidal silver preparation, this preparation may be effectively used as an alternative to antibiotics."

The original study compared a good quality colloidal silver against tetracyclines, fluorinated quinolones (Ofloxacin), the penicillins, the cephalosporins (Cefaperazone) and the macrolides (Erythromycin). Among the microbes tested were streptococcuses, pneumonia, E. coli, salmonella, and shigella.

There has since been considerable dispute about Leavitt's right to publish this data publicly, with BYU has been sending intimidating cease and desist orders out to individuals and organizations referencing the study. These letters go beyond reasonable demand that copyrights be respected and further demand that BYU not be mentioned in conjunction with the study at all, despite the fact this information now exists in the public domain.

Safety Of Colloidal Silver

Up until the advent of chemical antibiotics which came about in the 1940's, some very strong silver products were being used as antibiotics. Of the millions of people who used these concentrated silver products, there were no reported deaths and only 239 reported cases of generalized Argyria (EPA Report ECAO-CIN-026 Jan "91" Pg. VI-3). In that same EPA report on VI-4 it states that Gaul and Staud (1935) suggested 8 grams of silver arsphenamine (used by injection at 145,000 ppm strength) as a safe total (lifetime) dose.

They also noted that other authors suggest safe total doses of 12-15 grams, based on clinical experience. The work of Furchner et al (1968) as stated above showed that small amounts of silver did not build up in the system of primates.

According to the EPA IRIS Report on silver (Integrated Risk Information Systems) (5th , 1st paragraph) it states that a number of tests were completed to investigate the absorption and retention of ingested silver in a number of animals (including primates). In conclusion, the test work indicated that between 90-99% of ingested silver was excreted on the second day after ingestion and greater than 99% was excreted in less than a week.

In other words, almost all the ingested silver was out of the body in only two days, which indicates that silver does not build up in the system when consumed in small amounts.

Overdose (Argyria)

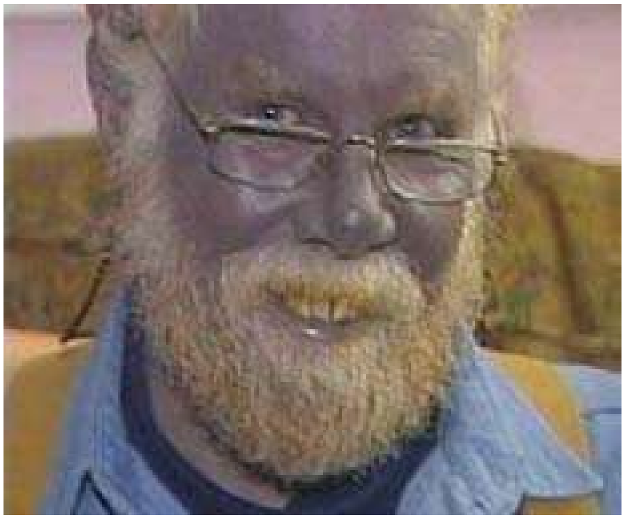

However, it is possible to overdose with silver, and this may lead to a rare cosmetic condition of blue skin. Its scientific name is argyria or silverism and is an irreversible blue-gray discoloration of the skin, nails and gums that's linked to colloidal silver. It can happen gradually and sneak up on you unawares, after a number of years.

You have to be pretty excessive (in fact really daft) to get it. It would require an intake of 1,000 mgm or more from the use of any silver compound, including its salts. Once established, the condition is usually permanent and is a kind of disfigurement.

The US Federal Register listed the silver products that cause argyria as silver salts, including; silver nitrate, silver arsphenamine, silver chloride and possibly silver iodide. These products were sold until about 1975 under various labels consisting of silver solutions ranging from 5-30% silver [50,000-300,000 ppm (parts per million) of silver] *(Federal Register, FDA-21CFR Part 310, pg. 53685).*

In comparison American Biotech Labs silver supplement products are only 10 ppm and 22 ppm. American Biotech Labs' EPA approved hospital/home surface disinfectant is only 32 ppm.

Other potential side effects of the liquid include: seizures and other neurological problems Kidney damage Indigestion Headaches and fatigue Skin irritation

#22

Nutritional factors

Now let’s look at nutritional factors against infections. It’s CRITICAL!

I have already explained that the main reason TB was conquered had nothing to do with vaccines or antibiotics; it was just improvement in nutrition for the population. It’s the same story for most infectious diseases. The usually touted nonsense that we’ve conquered disease with antibiotics and vaccinations is completely wrong but it’s believed with all the passion of religious fervor, directly against all evidence to the contrary.

Here’s the proof:

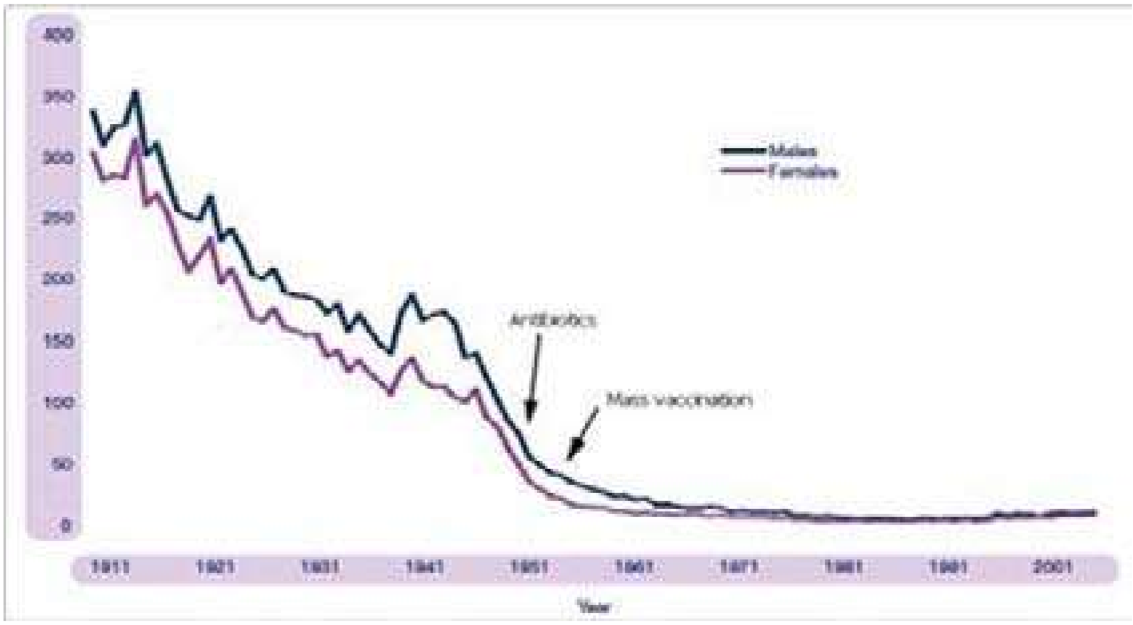

These figures are from Britain but the trend is the same the whole world over.

And TB (overleaf):

Note the sudden surge during the Second World War. This was due to temporary malnutrition and further re-inforces the point I am trying to make.

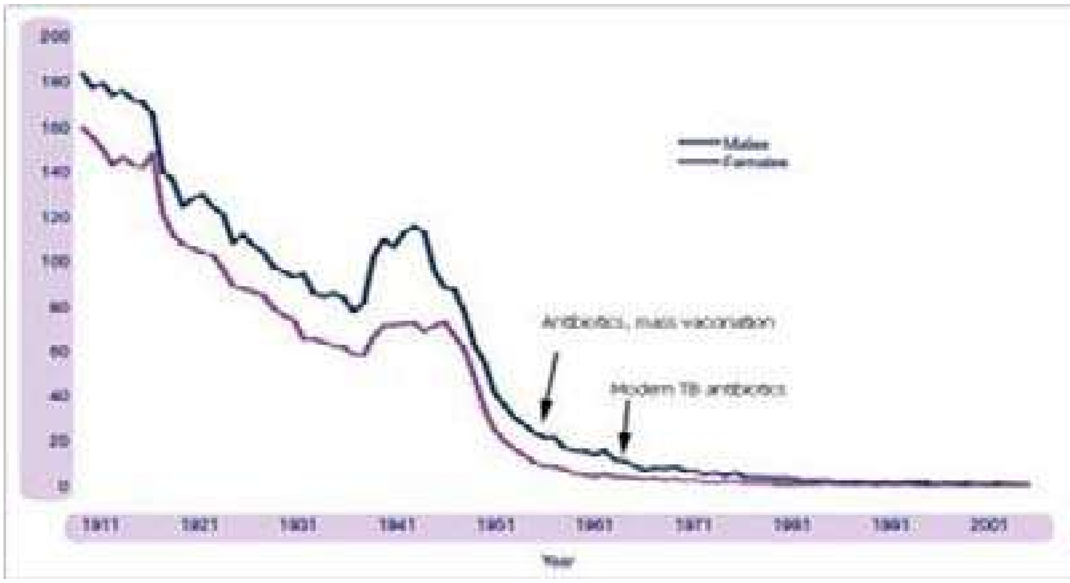

If the Americans want to ignore facts from elsewhere in the world (they usually do!), here is proof you're covered too:

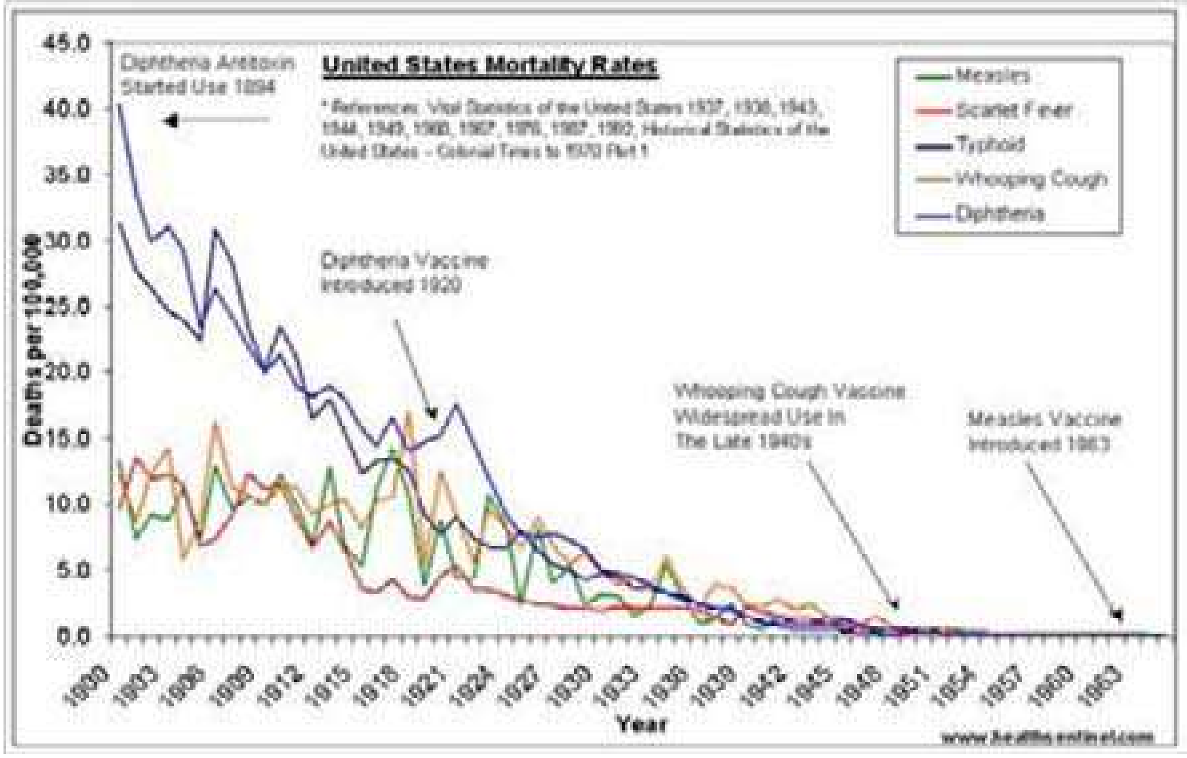

The depiction of Diphtheria vaccine introduced in 1822 is completely misleading: it was an anti-serum to the toxin, not the bacterium. Only one of these featured diseases is viral in nature (measles) but the principle applies just fine to immunity against all pathogenic organisms. General standards of hygiene and wellbeing, good housing, plus adequate nutrition make a big difference in general immunity in the population.

Nutrition bullets

This is not a book on nutrition but here, in a nutshell, is what you should do.

Avoid sugar at all costs. You probably know that in the US the average per capita is 150 lbs. That’s 3 lbs a week! Over 150 years ago Native American Indians warned against sugar: they said it weakened the body. It makes the immune system sluggish for up to 6 hours afterwards and the last thing you would want is tardy white blood cells that show up late for the action! Also it removes zinc which is very important for immunity.

Eat well. By that I mean a natural diet of natural foods with enough calories. Infectious diseases prosper only where there is malnutrition. Almost all endemic diseases (diseases which run permanently in society, not just in epidemic fashion), such as measles, TB and cholera, all declined significantly in the decades BEFORE antibiotics. The cause was better nutrition.

Remember being overweight does not mean you have “extra” nutrition: it means you have mal- or dys-nutrition. You are eating badly, eating the wrong things and ruining your body defences.

Alcohol, caffeine and tobacco are also known to lower immunity.

Instead of junk you must eat more foods containing anti-oxidants. These are contained in colored fruits, especially berries, blue, red and black. Manufactured foods which contain none of these essential life-giving food properties. It’s astonishing how little infectious disease you will ever get, if you eat like this.

#23 Vitamin A

You will hear talk from some people who do not do their research properly that vitamin D is the “antibiotic vitamin”. It is not. Vitamin A is the clear winner. For my subscribers to the Wholesome Living Letter (now ceased publication) I cited important clinical trials in the years before antibiotics that vitamin A in very large doses (50- 100,000 IU) killed virulent bacteria as well as antibiotics. Mortality rates moved from 100% fatal to 100% survival. That’s BETTER than antibiotics!

In 1929 a milestone study was conducted by Sir Edward Mellanby, Professor of Pharmacology in Sheffield, England, for the Medical Research Council of the UK. Up to that time, puerperal sepsis was a highly fatal and untreatable condition (for instance in 1928, just one year earlier, there were 16 cases on Mellanby’s pitch and all the women died). Mellanby administered high doses vitamin A to 5 women in his study—one of the first controlled clinical trials ever—and all 5 women survived! It was a triumph for science but also nutrition.

Unfortunately, antibiotics lay just around the corner and that soon put paid to the idea of a humble vitamin being remarkably effective against a killer fever. Well, please remember, just 80 years on, the clock has turned completely and we are now almost post-antibiotics. *[Mellanby E, Green HE (1929). Vitamin A as an anti-infective agent. Its use in the treatment of puerperal septicaemia. BMJ 1:984-6]*

In 1931, Mellanby published another keynote paper with the title ‘Diet as a prophylactic agent against puerperal sepsis’. In it he measured the incidence of ‘sepsis’ in a group of 550 post-natal women. Yet the list included cystitis (9 cases), mastitis (7 cases), influenza (1 case), gonorrhea (1 case), cervicitis (1 case) and septic perineum (2 cases), endometritis (7 cases) and 1 case of streptococcal septicaemia.

Again the response to Vitamin A was remarkable. [Green HN, Pindar D, Davis G, Mellanby E (1931). Diet as a prophylactic agent against puerperal sepsis, with special reference to Vitamin A as an anti-infective agent. BMJ 2:595-598].

What is meant by high doses of vitamin A? 25,000 – 50,000 units. “But that’s toxic”, say the pundits, aghast. Well, which would you choose: vitamin overdose or death? The fact is that vitamin A is not toxic at these levels in the short term. Your fate is decided in 48- 72 hours! Not long enough to worry about liver toxicity, surely?

#24 Vitamin C

Now come to what is probably the most powerful of all antibiotic substances known to man. Incredibly, it's also one of the safest pharmaceuticals we have. If you are caught in a vicious pandemic, yet you can gain access to this therapy, you should do fine!

I'm talking about humble vitamin C (ascorbic acid). Actually not so humble—it has brought people out of septicemic coma when nothing else worked and regular antibiotics were powerless to save the patient.

But only when administered IV in huge doses, hundreds of times greater than doctors today would even consider.

The full story of its power is told in an important book, entitled Vitamin C Infectious Diseases And Toxins (Curing The Incurable) by Thomas E. Levy MD, JD. The author cites over 12,000 medical and scientific journal references, which is far more than I can do here.

Levy's book leans a great deal on the work of an earlier doctor called Frederick R. Klenner MD, chief of staff at the Memorial Hospital in Reidsville, North Carolina, who worked in the 30s and 40s, when vitamin C first became readily available as an affordable pharmaceutical.

Klenner pioneered his treatment on hundreds of cases and had outstanding results against many infectious diseases, using intravenous vitamin C.

Klenner maintained hospital records and fever charts that told of case after case of recovery from meningitis, encephalitis, pneumonia and serious complications following scarlet fever (Streptococcus). Many fevers from 100-105 degrees F began to diminish within minutes of intravenous vitamin C. Both viral and bacterial diseases responded. The only requirement seemed to be that the dose was large enough and treatment continued for long enough.

Polio Teasers

Adelle Davies, famous authoress of the classic text Let's Eat Right to Keep Fit, was fortunate enough to meet Dr Klenner, and she tells us one of his stories; of an 18-month-old girl suffering from polio. The mother reported that the child had become paralysed following a convulsion, after which she soon lost consciousness. When Dr Klenner first saw the child her little body was blue, stiff, and cold to the touch; he could neither hear heart sounds nor feel a pulse; her rectal temperature was 100°F. The only sign of life, he could detect was a suggestion of moisture condensed on a mirror held to her mouth. The distraught mother was convinced that the child was already dead.

Klenner injected 6000 mg of vitamin C into the child's blood; four hours later she was cheerful and alert, holding a bottle with her right hand, though her left hand was still paralysed.

A second injection was given; soon the child was laughing and holding her bottle with both hands, all signs of paralysis gone. She recovered fully.

To show just how far this can be pushed, Levy cites an adult case, a woman, also a polio victim. She had considerable pain in the hamstring area, pain in neck and pain in the lower back. She had a natural desire to keep her entire body static to avoid painful moments. Her fever reached 104.6°F, along with a sore throat that had relapsed after initial treatment with antibiotics, aspirin and fruit juice two weeks earlier.

Although the patient weighed less than 120 lbs. Klenner nevertheless immediately administered 22,000 mg (22 g) of vitamin C. by slow intravenous injection, using a large 100 cc syringe. Plus he prescribed 1500 mg of vitamin C with juice every two hours by mouth.

Twelve hours later, the woman was free of her headache, and her fever was down to 101.4°F. Klenner gave her another 22,000 mg injection; this caused nausea and vomiting which lasted for around 30 minutes. Then, 24 hours later, her temperature had further dropped to 100.8°F, with a definite clinical improvement.

She was continued on 18,000 mg injections given every 12 hours (equivalent to 36 g a day). After a few days, the IV vitamin C. was discontinued, but the patient was asked to continue by mouth for an additional week taking 1500 mg every three to four hour.

Klenner noted that the patient had an almost complete elimination of pain, except at the knees, after the first 48 hours. The temperature normalised in 84 hours. The patient finally made a complete recovery.

Although these are viral cases, similar outstanding results are to be obtained with potentially fatal bacterial infections, such as tetanus, diphtheria, pneumococcal pneumonia, TB, leprosy and typhoid fever. Klenner came to speak of vitamin C as "the antibiotic par excellence".

Blood Levels

Significantly, Klenner found that vitamin C levels vanished completely from the blood when the patient was under the stress of an acute infection. Nor could any vitamin C be found in the urine. This severe deficiency might be why patients respond so dramatically to massive doses of the vitamin. We now know that this lack is caused by vitamin C being used up by neutralizing toxins and blasting viruses and bacteria; it effectively eliminates the danger, but is lost in doing so.

The rational basis of the high dose technique is therefore to administer vitamin C intravenously until it began to spill over into the urine and could be detected once again. Monitoring blood levels is obviously even more accurate but not always possible and not necessarily desirable when urgency can save a life; after all, vitamin C is pretty harmless. In any case, it is the intracellular levels that count, which may be much depleted, even when blood plasma levels appear reasonable.

Hence Klenner's urgency and the need to administer the doses as rapidly as possible, to be sure the plasma was overloaded and thus that the vitamin C would be sure to flood into the cells. Modern research, by the way, shows us that white blood cells may carry up to 80 times the residual blood levels of vitamin C, thus being sure to bring plenty of it to the site of an infection. So Nature knows how good this stuff is!

Klenner published at least 20 significant papers and in 1949 addressed the American Medical Association, to tell colleagues what he'd achieved. Yet his work is now virtually lost and any mention of it is dismissed a myth and nonsense. The truth is now known only to a few physicians and practitioners [it was my "drug of choice' to take with me to the danger zone rescue mission after the 2004 tsunami and kindly donated by friend of Deseret Biologicals, Sandy, Utah].

Klenner achieved outstanding recoveries in diseases which at that time were considered to be virtually untreatable.

Bacterial Diseases

Thomas Levy also quotes extensive research by other scientists and doctors, using large doses of vitamin C to overcome severe and sometimes fatal bacterial infections, including the following:

- Rocky Mountain spotted fever.
- Pseudomonas, a notoriously difficult to treat bacteria.
- Bacillary dysentery and even amoebic dysentery.
- Brucellosis.
- Malaria (not a bacterial disease, but transmissible parasite called Plasmodium).
- Typhoid fever (not to be confused with typhus, which also responds but is viral in origin)
- Leprosy
- Staphylococcus
- Streptococcus (leads to rheumatic fever).
- Tuberculosis
- Tetanus,
- Whooping Cough (Pertussis)
- Diphtheria

Note also: studies have proven that large doses of vitamin C helps to prevent allergies of varying kinds. I've given IV gallons of the stuff over the years. It even counters the effects of poison oak, poison ivy, snakebite, black widow spider bite and carbon monoxide poisoning.

So if treatment with intravenous vitamin C is so effective in such a wide variety of conditions, how is it that most doctors have never even heard of it, much less use it in their office?

Klenner Protocol

The trouble is that the scientific literature is dogged with studies which purport to investigate the effects of vitamin C but which used such tiny doses as to be worthless. It is an almost universal mistake among doctors to believe that ascorbic acid has only vitamin like properties and these manifested only a few hundred milligrams at most. It is even cited in the literature that the body will block blood levels above about 300 mgms. What Thomas Levy turns to calling "Klenner-sized doses" of vitamin C are very large—around 500- 700 mgms per kg of body weight, repeated several times a day. That would range from 20 grams daily for an infant; 40 grams for a child; to 25 grams for a slight-built woman; to 50 grams for a hefty man.

Clearly this is beyond the vitamin nature of ascorbic acid and indicates that it has a very important detoxing and antimicrobial property.

This is critically important because most bacteria exert their deadly effect by means of a toxin. So a substance which both destroys the pathogens and their toxins is indeed a godsend.

Take diphtheria, which once had a dreadful mortality rate of 80% (now down to 5- 10%). This disease kills by the secretion of a sticky dense membrane, which forms in the throat and suffocates the child. But there is also a deadly toxin, Diphtheria A, which itself is rapid fatal once present in sufficient quantity.

To treat this condition, Klenner recommended huge dose, as described, of 500- 700 mgm per kilogram of bodyweight, run into a vein through a wide bore needle "as fast as the cardiovascular system would allow". This is several thousand times what most doctors (in their ignorance) would consider safe. Yet even in these very high doses, ascorbic acid shows no ill effects on the patient, making vitamin C one of the most remarkable substances in human physiology.

There have been many published studies on this topic but in one report Klenner describes the fate of three children living in the same neighborhood, each with diphtheria. All three children had different doctors. The little girl under Klenner's scare was given 10 g of vitamin C in an intravenous push with a 50 cc syringe, every eight hours for the first 24 hours and then every 12 hours. She was also given 40,000 units of diphtheria antitoxin, injected into her abdomen. The other two children received the antitoxin as well, but they did not receive any vitamin C. They both died but Klenner's patient survived, later becoming a nurse.

[Klenner F. Observations of the dose and administration of ascorbic acid, when employed there are in the range of the vitamin in human pathology Journal Of Applied Nutrition, 23 (3&4); 61-88.]

It should be noted that Klenner's work with diphtheria is ably supported by other studies of the time. Harde and Philippe, demonstrated that lethal doses of diphtheria toxin premixed with vitamin C were no longer lethal when injected into guinea pigs. Without the vitamin C, the toxin killed the guinea pigs in just a few days. *[Harde E and Phillippe M. Observations sur le pouvoir antigene mélange toxine diphtherique et vitamin C. Compt rende Acad d sc 9:738-739]*

Levy cites several other studies from different countries, also showing vitamin C protected guinea pigs from the fatal effects of diphtheria toxin.

Streptococci and Staphylococci

What about these common human pathogens, which can produce infections varying from minor suffering to a fatal outcome? After all, we are very interested in Staph, because of MRSA.

Same great story. Vitamin C triumphs.

One study looked at two children with defective abilities of their white blood cells to kill bacteria. They are especially susceptible to repeated skin infections with staphylococcus. The study reported that vitamin C. was effective in delaying and eventually suppressing infectious episodes. *[Rebora A. and Crovato F. F Dallegri, and F. Patrone (1980) Repeated staphylococcal pyoderma in two siblings with defective neutraphil bacteria killing. Dermatologica 160(2):106-112].*

A much earlier study (1941) was able to demonstrate an inhibition of the growth of Staphylococcus aureus, even at relatively low doses. *[Gupta G. and B. Guha (1941) The effect of vitamin C and certain other substances on the growth of microorganisms. Annals of Biochemistry and Experimental Medicine 1(1):14-26]*

Kodama and Kojima were able to demonstrate the ability of vitamin C. to render staphylococcus related toxin, harmless. (Remember, I explained that the main deadly effect of virulent bacteria is the toxins that they produce). *[Kodama T. and T. Kojima (1939) Studies of the staphylococcal toxin,* *toxoid and antitoxin; effect of ascorbic acid and staphylococcal lysins and organsisms. Kitasato Archives of Experimental Medicine 16:36-55].*

Yet another case report described an elderly woman with an ulcer on her left cheek infected with the presence of staphylococcus aureus. She had had the condition over three years, but it rapidly healed with vitamin C therapy. *[Lederman E. (1962) Vitamin C deficiency and ulceration of the face. The Lancet 2:1382].*

Nakashini reported that the direct (topical) application of vitamin C to a bedsore was able to remarkably enhance the bacteria killing effect of antibiotics. He also noted that MRSA subsequently disappeared from the wound. *[Nakanishi T. (1992) [A report on a clinical experience of which has successfully made several antibiotic-resistant bacteria (MRSA etc) negative on a bedsore] Article in Japanese. Igaku Kenkyu. Acta Medica. 62(1):31-37.*

It even works in animals, test-tube experiments showed that the presence of vitamin C. significantly increase the ability of white blood cells to kill staphylococcus aureus.

[Andreason C. and Frank D. (1999) The effects of ascorbic acid on in vitro heterophil function. Avian Diseases 43(4):656-663.]

Remember the work I already alluded to, by Kelly (1944), showing that pathogens are almost universal in the mouths and on the tonsils of healthy hosts. It is only when immunity is compromised by nutritional deficiency—in this case lack of vitamin C—that infections can become established and overcome the host.

So in a way we are going round in circles and saying that good nutrition is your best protection against pathogens. If you do it right you may never even contract a disease, never mind succumb to it, even during a worldwide pandemic in which others are dying by the million!

But if you get MRSA, epidemic or not, make sure you call this section to the attention of a competent practitioner who is licensed to carry out this procedure. You will survive.

Administration

You will gather from reading this far that intravenous vitamin C can be a lifesaver, even for a patient afflicted with the most deadly infection and almost at the point of death.

But obviously not everyone will be able to find the help of a physician to administer IV vitamin C, especially in the event of a worldwide pandemic. To administer therapeutic doses to yourself, you only need to take vitamin C powder using the “fill and flush” method; that means to take doses by mouth, increasing by a teaspoonful at a time, until it causes diarrhea. We believe this happens only when the tissues have become saturated with the vitamin.

For a child, Adelle Davis suggests putting one cup of hot water in a liquifier, adding 50 tablets of vitamin C (500 mg each), blending well, sweeten to taste with honey, pour into a glass jar, and keep refrigerated. Doses can be administered mouthful at a time, or teaspoonful at that time.

If one does not have a liquifier, Davis suggest crushing 50 tablets and dissolving these in a cup of hot water. Each teaspoon of such a solution contains 500 mg. If 100 tablets (500 mg each) are dissolved in a cup of water, each teaspoon a solution would supply 1 g (1000 mg).

For any infectious disease in a sick child, mild or severe, administering a teaspoon more of vitamin C solution, sweetened in this way, can be a godsend when a physician cannot be reached, as for example, during the night.

Note that synthetic vitamin C does not contain enzymes found in natural foods; therefore it is quite stable to heat and storage. But once stirred into foods, such as fruit juice, enzymes will begin to activate and readily destroy it.

[Davies without any and L. let’s eat right to keep fit, Harcourt Brace Jankovich, New York, 1954, 1970, 132 to 134]

There are no right and wrong doses of vitamin C as such. All these treatments are based on empirical experience (basically trial and error). But this wonderful natural substance is very tolerant of error, and it’s almost impossible to overdose a person on it.

Important note:

Vitamin C vs. Ascorbic Acid

Don't listen to the ignorant nonsense widespread on the Web that vitamin C is "good" and "natural", while ascorbic acid is "bad" and synthetic. They are both IDENTICAL in physiological properties; just different names for the same thing.

All the good recovery stories I have just cited were all with synthetic vitamin C. It simply isn't possible to find the therapeutic high doses I have described from natural sources. So there is no question that synthetic vitamin C does the same job, or better.

What I did find, though, was that the manufactured source of vitamin C could cause unpleasant allergic reactions. The most common starting point is corn, so a sensitive corn-allergic individual could react to it.

But other plant sources are available, such as palm. These would be safer.

#25 Vitamin D

Throughout this section, I'll be referring to the D3 form (cholecalciferol or calciol), which has powerful immune booster properties. Without doubt D is the most powerful immune booster vitamin we know. It is highly effective at reducing mortality from all causes (recent 2007 studies). It lowers the cancer risk by over 60%. You want to be taking adequate amounts as of now. Don't wait for an emergency.

In April 2005, a virulent strain of influenza hit a maximum-security forensic psychiatric hospital for men that's midway between San Francisco and Los Angeles. John J. Cannell, a psychiatrist there, observed with increasing curiosity as one infected ward after another was quarantined to limit the outbreak. Although 10 percent of the facility's 1,200 patients ultimately developed the flu's fever and debilitating muscle aches, none did in the ward that he supervised. Even though they had mingled with patients from infected wards before their quarantine, none developed the illness.

A few months later, Cannell ran across a possible answer in the scientific literature. In the July 2005 FASEB Journal (Federation of American Societies for Experimental Biology), an article reported that vitamin D boosts production in white blood cells of one of the antimicrobial compounds that defends the body against germs.

Immediately, Cannell realized, the high doses of vitamin D that he had been prescribing to virtually all the men on his ward were the reason for the mass immunity.

Vitamin D stimulates cathelicidin production. The cathelicidins are an important group of antimicrobial proteins. They work by punching holes in the external membrane of a microbe, permitting its innards to leak out. They also trigger the congregation of white cells when we get an infection.

You can think of these as "on-board" antibiotics; part of our innate immunity.

The FASEB Journal article also triggered Cannell's recollection that children with rickets, a hallmark of vitamin D deficiency, tend to experience more infections than do kids without the bone disease. He shared his flu data with some well-known vitamin D researchers, and they urged him to investigate further.

On the basis of more than 100 articles that he collected, Cannell and seven other researchers now propose that vitamin D deficiency may underlie a vulnerability to infections by the microbes that cathelicidin targets. These include bacteria, viruses, and fungi, the group notes in a report available online for the December 2006 Epidemiology and Infection.

This is only a hypothesis, "but a very credible one" My take on this too is that the population malnutrition I was referring to in the previous sections, which was the main reason the infectious diseases declined, could be as much due to improved vitamin D status as other factors.

Scientists are beginning to suspect that would hold true for the decline in TB, also. A study published in 2006 investigated the relationship between vitamin D and susceptibility to tuberculosis and showed dramatically better recovery in patients supplemented with 10,000 IU daily. 100% of patients survived, while normally 40 - 50% will die, despite treatment *[Acta Med Indones. 2006 Jan-Mar;38(1):3-5]*

Maybe vitamin D status is the MAIN reason we fight infections better when we are properly nourished. In any case, the moral is clear: don't let yourself be deficient.

This article in the prestigious New England Journal Of Medicine made it clear that cathelicidn defiency was associated with increases susceptibility to Staph. infections (hence MRSA) *[Ong, P. Y., Ohtake, T., Brandt, C., Strickland, I., Boguniewicz, M., Ganz, T., Gallo, R. L. & Leung, D. Y. (2002) N. Engl. J. Med. 347, 1151-1160.]*

Ignored stupid and criminal government advice here in the USA and take at least 1,000 units a day. 2,000- 5,000 units would make more sense in an epidemic or if you are battling MRSA.

Vitamin D Cream

Vitamin D as a cream may have important potential for those with MRSA of the skin.

Independently, dermatologist Mona Ståhle of the Karolinska Institute in Stockholm reached a similar conclusion. Her team administered an ointment containing a drug mimic of 1,25-D to the skin of four healthy people. The salve hit "the jackpot, right away," Ståhle says. In the May 2005 Journal of Investigative Dermatology, her team reported that where the ointment had been applied, cathelicidin-gene activity skyrocketed as much as 100-fold.

The team also found evidence of a localized increase in the concentration of cathelicidin.

Scientists found that vitamin D was also effective against TB. The sunlight/ vitamin D effect might explain why blacks are racially more prone to TB. It would also tell us one reason why the sanitoriums, with their keen emphasis on fresh air and sunshine, had some measurable therapeutic results, even before the days of hospitalization.

120 or so reports over the past 70 years have pointed to a link between vitamin D and resistance to infections. Five studies since the 1930s that have linked reduced risks of infectious disease to dietary supplementation with cod liver oil, a rich source of vitamin D.

Just think about what you do know for a moment--- in winter, colds, flu, and other respiratory diseases are more common and more likely to be deadly than they are in summer. During winter, ultraviolet-light exposure tends to be low because people spend more time indoors and the atmosphere filters out more of the sun's rays, especially at mid and high latitudes.

Sun exposure—in moderation—might also prove therapeutic, Ståhle's team suggested in the November 2005 Journal of Investigative Dermatology. The scientists showed that in eight fair-skinned people, a single dose of ultraviolet-B radiation—just enough to evoke some skin reddening the next day—activated the vitamin D receptor and the cathelicidin gene in the exposed skin.

This reminds me that in preantibiotic days, tuberculosis patients were put on a fresh-air-andsunshine regimen. Could the vitamin D so acquired account for the cures this system sometimes produced?

Take a minimum of 2,000 IU a day, double or treble that in an epidemic. Ignore official advice that you don't need more than 200 IU a day. It will kill you. A tablespoon has around 1400- 1500 IU.

#26 Iodine

Here's a much neglected mineral deficiency that has profound effects of the performance of the immune system.

According to the WHO, 2.2 billion people worldwide are at risk for Iodine Deficiency Disease. Iodine deficiency impairs immune function. The myth that it only affects people in areas remote from the sea is made nonsense by these figures. Deficiency is everywhere. According to Dr David Brownstein iodine intake in the USA has fallen by 50% during the last 30 years.

We had iodized table salt but any intelligent doctor can see the clear evidence that iodine is not absorbed from salt. It was iodine in bread that actually helped the problem. But 30 years ago manufacturers ceased adding iodine to bread and instead used bromine, which counteracts iodine! Any wonder then?

Test For Iodine Deficiency

Dip a cotton bud in iodone solution, such as Lugol's or tincture of iodine and make a patch on your skin. If the yellowish stain disappears in less than an hour; it means your body is lacking crucial iodine and has soaked it up. If the stain remains for more than four hours, you iodine levels are fine.

Iodine levels are not just important for immunity. Low levels of iodine will impair thyroid function. Since the thyroid too is involved in the immune response, it's a double-whammy for your immune system if you are low.

If there is an epidemic pending or occurring, get yourself plenty of iodine fast. When times are safer, be more concerned about regulating the dose

Remember, women are naturally prone to iodine deficiencies. That's because the thyroid gland in women is twice as large as in men -- so under normal circumstances, women need more iodine. However, when there is increased stress (and when is a woman's life not stressful?), the need for iodine can double or triple.

Take 5 mg a day minimum, start now, don't wait for infections to strike. Ignore claims that iodine is toxic. So is anything if you take too much. The necessary doses are 10- 50 times higher than fashions claim. One of the best ways to boost your iodine levels is to add seaweed sea vegetables to your diet. Just one teaspoon of sea vegetables a day can help you regain normal iodine levels. Incorporating seafood and fish into your diet can also help. Other foods that contain iodine are eggs and dairy products, including milk, cheese and yogurt, onions, radishes, and watercress. Some foods, called goitrogens, should be omitted for a while as they hinder iodine utilization.

These included kale, cabbage, peanuts, soy flour, Brussels sprouts, cauliflower, broccoli, kohlrabi and turnips.

Tincture of iodine is often found in emergency survival kits, used both to disinfect wounds and to sanitize surface water for drinking.

When an alcohol solution is not desirable, Lugol's iodine, an aqueous solution of iodine in potassium iodide solution, or povidone iodine (Betadine), a PVPI solution, can be used.

When using free iodine to sanitize surface water, it should be left to sit for 30 minutes before drinking, allowing time for all viruses and bacteria to be killed. If the water temperature is less than 20 °C (68 °F), sitting time should be extended to several hours. To purify 1 L (34 oz) of water, 0.3 mL (0.01 oz) of 2% tincture should be used. Cold or cloudy water should be given 2.4 mL (0.08 oz) of 2% iodine, and should be left to sit for several hours.

Following disinfection, the iodine odor and taste can be removed by adding powdered vitamin C, which converts the iodine to iodide. A similar reaction also removes chlorine from chlorinated water.

Iodine used to sanitize water is also available in tablet or crystal form. One form or another should be available at health stores, and trip outfitters.

#27 Zinc

Who doesn't know that zinc helps fight infection? It also helps with wound repair and a number of other key healing functions.

Zinc is an essential mineral that is naturally present in some foods, added to others, and available as a dietary supplement. Zinc is also found in many cold lozenges and some overthe-counter drugs sold as cold remedies.

Zinc is involved in numerous aspects of cellular metabolism. It is required for the proper functioning of approximately 100 enzymes and it plays a role in immune function, protein synthesis, wound healing, DNA synthesis and cell division.

The critical point to note is that a daily intake of zinc is required to maintain a steady state because the body has no specialized zinc storage system. Yet our diets are notoriously lacking in zinc. Studies show that an average diet does not meet the recommended requirements (15 mg daily).

Partly it is removed in the processing of food, especially in refining grains. Phytate inhibition further reduces the bio-availability of zinc. To make matters worse, crops grown in Europe and the US may be short of zinc because the soil has become exhausted.

Signs of a lack of zinc are said to be white flecks in the finger nails and an impaired sense of taste. Professor Bryce-Smith's zinc status test rests on the fact that zinc-deficient individuals cannot taste the metal in solution as zinc sulphate. After adequate supplementation, taste appears to return.

Zinc works in concert with other vitamins and co-enzymes. For example, vitamin A metabolism needs zinc. B6, so helpful in many conditions, is far more effective in the presence of zinc and magnesium. Zinc appears to counteract copper and an imbalance of the ratio of these two metals is associated with an increased tendency to allergies.

Zinc supplements reduce the average duration of the common cold. Other clues suggest the importance of zinc in immunity.

Supplements

Zinc is available in a large number of proprietary preparations such as in the gluconate or orotate form. Doses of around 15 mg are required, but beware: quoted weights usually include the whole formula. A tablet with 100 mg of zinc as gluconate may have only 2 mg of elemental zinc!

Research has not determined whether differences exist among forms of zinc in absorption, bioavailability, or tolerability. However I recall a paper suggesting that the citrate form is best absorbed. For cheapness and convenience I often give zinc sulfate as a powder, 7 g dissolved in a pint of water. The patient takes a teaspoon per day of this liquid, providing about 15 mg daily.

In addition to standard tablets and capsules, some zinc-containing cold lozenges are labeled as dietary supplements.

Ironically, after what I said about the returning sense of smell, there have been several case reports of anosmia (permanent loss of smell) from the use of zinc-containing nasal gels or sprays, and so this raises questions about the safety of intranasal zinc.

Health Risks from Excessive Zinc

Zinc toxicity can occur in both acute and chronic forms. Acute adverse effects of high zinc intake include nausea, vomiting, loss of appetite, abdominal cramps, diarrhea, and headaches. Intakes of 150 - 450 mg of zinc per day have been associated with such chronic effects as low copper status, altered iron function, reduced immune function, and reduced levels of high-density lipoproteins. Reductions in a copper-containing enzyme, a marker of copper status, have been reported with even moderately high zinc intakes of approximately 60 mg/day for up to 10 weeks.

Therefore I do not recommend that you exceed 15- 20 mg a day of elemental zinc.

However- it may not be that simple.

Although zinc is clearly needed for competent T-cell mediated immunity, it has the opposite effect on neutrophils and macrophages, causing a reduction in their motility and ability to ingest invaders. That could explain a recent cautionary finding: that zinc may favor the growth and activity of bacteria over the cells of the host (rather like excess iron, see next section).

In an article appearing in the Dec. 1 2008 online edition of Proceedings of the National Academy of Sciences, researchers in UC's department of molecular genetics, biochemistry and microbiology detailed findings that the presence of zinc is crucial to the formation of infection-causing biofilms.

Staphylococci can grow as biofilms, which are specialized communities of bacteria that are highly resistant to antibiotics and immune responses. They are remarkably adhesive and can grow on many surfaces, including implanted medical devices such as pacemakers, heart valve replacements and artificial joints. Preventing or inhibiting the growth of such biofilms would dramatically reduce the incidence of staph infections.

UC researchers in the lab of Andrew Herr, PhD, found that zinc causes a protein on the bacterial surface to act like molecular Velcro, allowing the bacterial cells in the biofilm to stick to one another. Zinc chelation, or removal, prevented biofilm formation by Staphylococcus epidermidis and Staphylococcus aureus.

So zinc is a definite immune booster but has the surprising effect of favoring staphylococci and other invaders. Stay tuned for updates.

#28 Iron Overload

Here's where nutrition gets tricky, even dangerous. Excess iron in the human body causes fulminating fatalities, whenever there are serious bacterial infections. Iron favours the invader far more than the immune system.

The public has it so hammered into them that virtually everyone is iron deficient and "Better take some extra, just to be safe". This is the grossest nutritional mistake that is made by ignorant or foolish MDs today. Iron EXCESS is far from rare and is very dangerous indeed. If you have it and land in an epidemic or pandemic, you are for it!

Did you know that patients with iron overload disease are told not to walk on beaches barefoot? This is on account of a bacterium common in all sea water called Vibrio vulnificus. This bacterium, when it encounters stored iron in our bodies, becomes catastrophically toxic in the presence of excess iron. But how common is iron overload disease?

About 20% of the population is in danger. There are 42 million people at risk for iron overload in the U. S. population alone. Some population segments such as the Scott/Irish and the African Americans have twice the double gene frequency and an amazing 20% carrier rate in the U.S.

You may be risking your health if you take otc multi vitamins or minerals with iron or vitamin C without knowing your iron storage status. Vitamin C has been found to mobilize stored iron into the heart muscle where it sets up arrhythmia or heart rhythm disturbances. It can also cause people to over absorb iron even from their ordinary diets. Do not take vitamin C supplements if you have, or might have this condition, until you have found out for sure.

Meanwhile stupid or ignorant doctors go on prescribing iron pills (huge doses around 250 mgm) as if they were sweeties and not a compound which is more lethal than arsenic to millions of people!

Never mind iron deficiency; it is much more important to know if someone is in iron overload.

Maybe you are? Two lab tests will settle it: Transferrin saturation. Serum Iron (SI) divided by Total Iron Binding Capacity (TIBC) and Serum Iron (SI) gives Transferrin Saturation TS. It should be 12-45%.

Serum Ferritin. Normal range is: 5 to 150.

Without being in possession of this important laboratory information, remember you have a

1:5 risk of iron overload and you will die very swiftly if a plague of any organism does occur. Get it fixed before the problem occurs (the cure is chelation or, in an emergency, blood letting).

Overleaf there's a picture of what vibrio vulnificus did to one man's leg, just before the amputation. I DON'T RECOMMEND YOU LOOK AT THIS, UNLESS YOU HAVE A GOOD STOMACH.

This is a man who went fishing one day, got a tiny scratch on his leg and just wiped the blood off. This bacteria moves fast. He complained next morning about tightness in the leg and was running a fever that evening and he all but died that night in agonizing pain that no meds could relieve.

The irony is he never even got out of the boat.

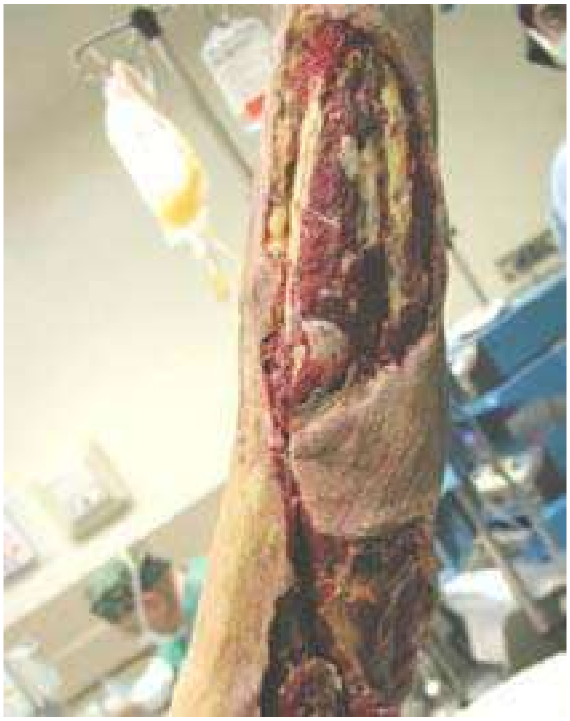

An above-knee amputation saved his life and after 10 days in a coma in intensive care he pulled through.

I urge you to take all care, get your iron levels checked now and learn EVERYTHING you can about antibiotic substitutes. Get rid of any MD who scoffs. Find a good one.

Doctor's do not give warnings about vulvificus to immune compromised patients (Kidney, Liver, Hepatitis C, Cancer, Leukemia, immune suppressant drugs, the list is long). If that applies to you, stay out of the water on the Gulf Coast between April and October.

Remember: vulnificus can enter the body through a tiny cut or bite. If you must go fishing, take with you a bucket filled with bleach water or hydrogen peroxide and dip your hands in it several times an hour.

If you suspect you (or a loved one) got unlucky, use the MMS protocol immediately, while en route to the hospital.

#29 What About Herbs?

We hear about tropical rainforests and the medicinal wonders they produce. Actually the ordinary garden also contains many such substances, as some of you may know.

Even conventional medicine is getting on to this; 20 years ago kids with leukemia were in trouble; 4 out of 5 died. Now it's 4 out of 5 survive, thanks to the periwinkle plant (Vinca major). This humble little ground cover plant has given us vinblastine and vincristine, two very powerful chemo agents.

But really, when we say "herbs", we usually mean gentler phytotherapies than this!

Probably the most famous herb for beating infections is Echinacea (Coneflower). It's not the best, in my opinion, but certainly deserves a favorable mention. It fortifies the immune system.

But I find the results of using just this herb disappointing. Moreover numerous clinical trials have failed to show much positive effect. Not all studies are corrupt and bent only on rubbishing natural therapies! I would not ask anyone to rely much on Echincea as protection in an epidemic or real pandemic.

The best way to take Echinacea is as a homeopathic compound. I have in mind one product in particular, which is Echinacea compositum, by the German manufacturer HEEL. It's best taken as liquid vials: if no doctor is available to administer it IV, a patient can still take the ampoules, one dissolved in water every third day (one day ON, two days OFF).

There are other immune fortifying herbs, such as Astragalus, Boneset, Calendula, Lady's Mantle (Alchemilla) and yarrow. Ginger has powerful anti-infective, anti-parasitic, and anti-inflammatory properties.

Compare these also with the benefits of fungi and mushrooms, described in another section (#38).

Thing is, herbs have a place but I wouldn't really rely on them if an infection was raging in my body. I'd want something that killed bacteria (bactericidal) or at least held them down while the immune system got to work beating them to death (bacteriostatic)!

There are some plant sources which are definitely powerful antibiotics. They hold in check, and even kill, bacteria.

Best-known by far, of course, is common garlic. Some others are probably better.

We can use herbs in a variety of ways, apart from just eat them. For example, herbal infusions, herbal baths and herbal wraps, as well as dried leaf powders and capsules.

Then there is essential oil therapy. This, as you will see in a later section, is very different from recreational "aromatherapy". Essential oils were in fact the first serious antibiotics that we had in the 20th century. These oils are still there and still work. Doctors ignore them but that doesn't mean they are not to be preferred to synthetic manufactured compounds with complicated side effects!

Herbs can also be smoked! I found this study on PubMed from the Department of Pharmacy and Pharmacology, Faculty of Health Sciences, University of the Witwatersrand in South Africa (Abstract): "Validation Of Smoke Inhalation Therapy To Treat Microbial Infections."

Aim Of The Study: In traditional healing, the burning of selected indigenous medicinal plants and the inhalation of the liberated smoke are widely accepted and a practiced route of administration. This study elucidated the rationale behind this commonly practiced treatment by examining the antimicrobial activity for five indigenous South African medicinal plants commonly administered through inhalation (Artemisia afra, Heteropyxis natalensis, Myrothamnus flabellifolius, Pellaea calomelanos and Tarchonanthus camphoratus).

Material And Methods: An apparatus was designed to simulate the burning process that occurs in a traditional setting and the smoke fraction was captured for analysis and bioassay. Methanol and acetone extracts as well as the essential oil (for the aromatic species) were prepared and assayed in parallel with the smoke fraction.

Results: Antimicrobial data revealed that in most cases, the 'smoke-extract' obtained after burning had a better effect than the corresponding solvent extracts and essential oils. So the "smidging" tradition of the North American indians was validated.

The report concluded that "These results suggest that the combustion process produces an 'extract' with superior antimicrobial activity and provides in vitro evidence for inhalation of medicinal smoke as an efficient mode of administration in traditional healing.”

[J Ethnopharmacol. 2008 Oct 28;119(3):501-6. Epub 2008 Aug 19.]

Homeopathy: Section #41

Finally, don’t forget that many homeopathic remedies (section 41) are also derived from herbs and have similar or related properties.

#30 Calendula

This plant sits halfway between the immune fortifiers and actual antibiotic herbs. It has a definite antispectic action and a measurable antibiotic action. It's great for minor skin infections, such as wounds and abcesses.

Calendula officinalis (pot marigold) comes from Latin and means "the first day of the month". This wonderful annual, which grows up to 28 inches tall, produces cheerful flowers that have a faint but distinctive scent. Its leaves can be added to salads while the flowers add a colorful garnish to many dishes (some people even eat the garnish!)

Medicinal Calendula preparations are usually made from either the whole flower heads or just the petals. The flower head contains an essential oil that has an antibiotic effect and lends the flower its characteristic scent. Other components include calendon, which also has a strong odor, and a carotene-like dye calendulin.

Calendula works well as a fairly strong antiseptic properties both topically and internally. Applied externally to the skin, the plant's antibiotic action helps promote the painless healing of minor wounds by reducing inflammation and pus formation.

I've prescribed calendula ointment for decades as a burn remedy. It helps keep the wound antiseptic and promotes healing without scarring.

Calendula can also be used as a gargle to relieve sore throats.

Making Calendula tea:

Pour 1 cup of boiling water over 1-2 tsp. of dried calendula petals. Steep for 10 min, then strain. Drink 2-3 cups of warm tea daily. Use as a gargle for sore throats.

For immediate dressing of wounds pluck (or buy) fresh flowers and squeeze to extract the juice, which you can apply directly to wounds.

Calendula ointment can be bought in stores or prepared at home by mixing a few drops of calendula oil with skin lotion. Apply it to minor burns, bruises, cuts and scrapes. The homeopathic form of Calendula is best taken as 30C tablets. So-called homeopathic Calendula ointment is also available but at 1X (ten times diluted) it’s hardly homeopathic. It cost just a few dollars.

#31 Garlic

It was used by the ancient Egyptians, Chinese, the Greeks and Romans. In 1858, Louis Pasteur demonstrated the powerful antibacterial properties of garlic. It was used in both World Wars. When no other antiseptic or antibiotic was available wounds were smeared with crushed garlic juice and then bandaged. In WW2 the Russians used it when antibiotic supplies ran out. It was called "Russian Penicillin".

Modern research has shown that garlic juice can slow the growth of (and even kill) more than 60 species of mould, and more than 20 species of bacteria, including some that are very virulent. It is also currently attracting attention because of its ability to lower blood pressure, cholesterol, and for its possible anti-cancer properties.

The active ingredients are ajoene and allicin. Ajoene breaks down via sulfenic acid to allicin when it is crushed. Allicin is the chief antibiotic compound. Allicin gives rise to the characteristic garlic smell. Don't be fooled by sales talk that garlic oil and capsules have the same effect as raw, stinky garlic!

1 mgm of allicin is said to be equivalent to 15 standard units of penicillin. The standard dose of penicillin is around 1500 Units, so 100 mgms of allicin could be equivalent to a penicillin dose!

There are over 2,000 studies on the antibiotic efficacy of garlic.

Microbiologist Dr Ron Cutler, based at the University of East London, claims that the garlic compound allicin not only kills established varieties of MRSA, but also destroys the new generation of 'super-superbugs' that have evolved resistance to Vancomycin and Glycopeptides, the powerful antibiotics widely considered to be the last line of defence against MRSA.

Allicin can cure patients with MRSA-infected wounds within weeks, according to in vitro research presented at the Institute of Biomedical Scientists congress in Birmingham during October 2003 and was published in the Journal of Biomedical Science 2004.

[Cutler, RR; P Wilson (2004). "Antibacterial activity of a new, stable, aqueous extract of allicin against methicillan-resistant Staphylococcus aureus (PDF file)". British Journal of Biomedical Science 61 (2): 71-4. http://www.allimax.us/Cutler.pdf. Retrieved on 20 Feb 2009.]

After showing that allicin destroys the MRSA microbe in laboratory trials, Cutler has now teamed up with the firm Allicin International to develop topical treatments including a nasal cream, oral capsules and soaps that have proved effective against both MRSA and GISA.

Another study published in 2002 found that raw garlic consumption could help limit the dam-age done to the heart after surgery because of its natural antioxidant properties. Supplements of allicin have also been shown to reduce risk of colds, prevent high blood pressure and kill cancer cells. So eat it!

Garlic may be one of the reasons the so-called Mediterranean diet prolongs life; it includes lots of garlic, fresh and cooked. If you are not used to using this amazing plant flavor, it's time you got it into your diet.

Just make sure some of it is raw, chopped up or blended in sauces and salsas etc. Cooked garlic has little of the properties it is famous for.

Kyolic is aged garlic and rather different but a powerful medicinal in its own right.

Note: garlic can interfere with the action of some anticoagulant drugs.

Use Garlic Tea as an Antibiotic

Here's how to make it: Place a few cloves of garlic, pre-crushed, into a saucepan of water and bring it to a boil.

Remove the pan from the heat. Stir in the honey and lemon juice or a little of some other suitable sweetener, such as maple syrup.

Strain into a warmed teapot and sip about half a cup.

Repeat 3 times a day, either hot or cold. Refrigerate the remaining tea.

Surface Application: In an emergency it can be crushed and the juice applied to the site of an open wound to kill bacteria such as staph and strep. These strains of bacteria cannot become immune to garlic's effect as it has modern synthesized antibiotics because garlic kills in a different way.

#32 Artemisia annua

Better Than Garlic? I Think So!

Artemisia is a large, diverse genus of plants with between 200 to 400 species belonging to the daisy family Asteraceae. It comprises hardy herbs and shrubs known for their volatile oils. They grow in temperate climates of the Northern Hemisphere and Southern Hemisphere, usually in dry or semi-dry habitats. The fern-like leaves of many species are covered with white hairs.

It is said that the genus Artemisia (which includes over 400 plants) may be named after an ancient botanist. Artemisia was the wife and sister of the Greek/Persian King Mauseolus (from whose name we get the word mausoleum). Artemisia, who ruled for three years after the king's death, was a botanist and medical researcher, and died in 350 B.C. *(Wikipedia).*

The two best known plants are Artemisia annua (Sweet Wormwood, Sweet Annie, Sweet Sagewort or Annual Wormwood. Chinese : pinyin) and Artemisia absinthum. The latter has no significant medicinal properties and is the basis of the drink absinthe. It has been banned in many countries, for its brain-rotting effects.

Vermouth (from the German Wermut) was originally made with wormwood but now uses more innocent herbs for flavor.

Other members of this family include the cooking herb Tarragon.

Most members of the genus are bitter. This led to the plant being used by wet-nurses for weaning infants from the breast, as in this speech by Shakespeare from Romeo and Juliet Act I, Scene 3:

Nurse: ... And she [Juliet] was wean'd, – I never shall forget it, – Of all the days of the year, upon that day: For I had then laid wormwood to my dug, ...

Scientific Interest

There is a lot of interest in the medical properties of Artemisia annua, which I shall call simply Artemisia from this point. Its chief ingredient, artemisinin, is a potent anti-malarial, anti-fungal and even anti-cancer substance. A search of PubMed with the term artemsinin yields hundreds of papers.

Artemisinin is a terpene which contains peroxide bridges. Taken along with section #19 on hydrogen peroxide this can be seen why it would benefit infections, even including the very difficult case of toxoplasma cysts.

Two other important active ingredients which are widely being studied are artesunate and artemether. They seem to have pretty good anti-infective properties too. To quote a study carried out at the Centre for Infection, Division of Cellular and Molecular Medicine, University of London, and published in the journal Trends in Pharmacological Science *[Trends Pharmacol Sci. 2008 Oct;29(10):520-7]:*

"Artemisinins are derived from extracts of sweet wormwood (Artemisia annua) and are well established for the treatment of malaria, including highly drug-resistant strains. Their efficacy also extends to phylogenetically unrelated parasitic infections such as schistosomiasis. More recently, they have also shown potent and broad anticancer properties in cell lines and animal models. In this review, we discuss recent advances in defining the role of artemisinins in medicine, with particular focus on their controversial mechanisms of action. This safe and cheap drug class that saves lives at risk from malaria can also have important potential in oncology."

Malaria

Artemisinin has become well-known for its powerful anti-malarial effect. At a time when the world is steadily losing control of the disease, due to widespread drug resistance of the malaria parasite Plasmodium falciparum.

Systematic reviews on Artemisinin show that it is as effective as quinine in treating both uncomplicated and severe malaria. Unfortunately, the relapse rate was still high when used alone.

But artemisinin in conjunction with a suitable anti-biotic, a protocol called "Artemisia combination therapy" (ACT), has been found to be highly effective. Coartem, from artemisinin combined with an antibiotic lumefantrine, wipes out malaria in more than 96 percent of patients in regions where malaria has become resistant to older drugs. Traditional meds such as chloroquine work in only 50 percent of patients where the parasite is drug-resistant. There were an estimated 247 million malaria cases in 2006, and nearly 881,000 patients died, according to the World Health Organization.

Coartem was approved by the Swiss in the late 1990s and put on the World Health Organization (WHO) essential drug list in 2002. The USA was typically way behind world progress, with the FDA holding it back till early 2009 (the USA mistakenly believes only their doctors and scientists are competent to make these decision).

Other Pathogens

To be included in this report Artemisia would have to affect more than just malaria, though this disease is actually the world's number one killer.

I found a paper which showed artemisinin in combination with ampicillin or unasyn (a complex of ampicillin and sulbactam) decreased mortality in mice subjected to a lethal live E. coli challenge from 100 to 66.7% or 33.3%, respectively *[Antimicrob Agents Chemother. 2006 July; 50(7): 2420–2427].*

It seems clear it has widespread anti-biotic applications. Studies published on the Sloan-Kettering website show anti-parasitic properties. In vitro studies indicate that artemisinin may be an effective treatment for other protozoal infections such as leishmaniasis, Chagas' disease, and African sleeping sickness. *[http://www.mskcc.org/mskcc/html/69126.cfm, accessed June 27th 2009, 10.40 am PST]*

Another study from the University of Nizwa, Oman, showed in vitro antimycoplasmal activity of six Jordanian medicinal plants, including a related Artemisia, against three Mycoplasma species. Trop Anim Health Prod. 2007 Oct;39(7):515-9.Links All Mycoplasma species showed susceptibility to Artemisia herba-alba and Artemisia arborescens with MIC ranges from 3.125-12.5 mg/ml.

Incidentally, this study compared garlic and pomegranate with Artemisia arborescens and showed they had only limited activity against some Mycoplasma isolates.

Another Artemisia was helpful with Helicobacter pylori, the scourge of stomach ulcers (and stomach cancer). Increasing drug resistance means we are now losing ground in this “breakthrough” treatment of ulcers. Artemisia may be the answer.

A study in Mexico published in the Journal of Ethnopharmacology showed Artemisia ludoviciana (also Persea americana, Annona cherimola, Guaiacum coulteri, and Moussonia deppeana) had high antHelicobacter pylori. *[J Ethnopharmacol. 2009 Mar 18;122(2):402-5].*

How To Take It

The good news is that it is free. Artemisia annua flourishes in almost any garden, worldwide.

A 2004 test reported that Artemisinin is absorbed faster from the tea preparations than from capsules. The maximum plasma concentrations were observed after 30 minutes. Artesunate is rapidly absorbed and reaches maximum plasma level within 45-90 minutes *[Rath K, et al. Pharmacokinetic study of artemisinin after oral intake of a traditional preparation of Artemisia annua L. (annual wormwood). Am J Trop.Med Hyg. 2004;70:128-32]*.

Making The Tea: One litre (2 pints) of boiling water is poured onto 5 g dried leaves (large tablespoonful, loose, or teaspoon if crushed to a powder) of Artemisia annua. It is allowed to brew for 10 to 15 minutes, and then poured through a sieve. This tea is then drunk in four portions in the course of the day. The period of treatment is between 5 and 7 days.

Patients with gastrointestinal disorders or those taking antacids should not take Artemisia because it increases the production of stomach acid *[Skyles AJ, Sweet BV. Alternative therapies. Wormwood. Am J Health Syst.Pharm. 2004;61:239-42].*

It is also unwise to use this remedy if you are taking antiseizure medications; Artemisia can induce seizures resulting in decreased efficacy of antiseizure medications [previous reference].

#33 Samento

(Uncaria tomentosa)

You may not have heard of this powerful antibacterial and anti-inflammatory. It's in the Cats Claw family but reputedly 1,000 times stronger in its immune enhancing effect than ordinary Cats Claw.

Samento is the name given to the plant by the Ashaninka Indians in the Peruvian rainforest. They have ancient knowledge of the plant and use it for a wide variety of complaints.

The Indians are a crucial source of the plant materials because they can recognize the correct strain, without the need to elaborate tests. Let me explain why this is important:

Ordinary Cats Claw contains important oxidole alkaloids, one of the many important groups of phyto-nutrients found in the plant. One strain of plants contains pentacyclic oxindole alkaloids (POAs); the other contains both POAs and tetracyclic oxindole alkaloids (TOAs).

POAs act on the cellular immune system and TOAs act on the central nervous system. TOAs are not good because they interfere with the benefits of the POAs. This is confirmed by studies.

Uncaria tomentosa plants may contain as much as 80% TOAs. According to studies in Austria, as little as 1% TOAs can cause a 30% reduction in the acquired immune benefits that POAs provide. So even a little TOAs, if present, can go a long way to neutralizing the benefits of Cats Claw.

This explains why lots of the common herb were needed by Western herbalists, to produce any real effect. But the amazing strain found by the Ashaninka Indians does not have the destructive TOAs. Make sure when you buy Samento, that the packet specifically states no TOAs. Otherwise it isn't Samento anyway!

Samento has shown dramatic benefits in virtually every kind of inflammation, such as asthma and arthritis.

It can also fight infections from AIDS to hepatitis' flu to salmonella. Its broad spectrum effect helps to increase T-lymphocyte and microphage production. It also helps repair the lipid matrix in the cell walls and is particularity good for stopping the expansion of virus laden cells.

The 2 active ingredients are berberine and hydrastine. These are good for killing many of the bacteria that cause diarrhea, such as E. Coli, Salmonella, Shigella, Klebsiella and the cholera bacterium. Also parasites like ameba, giardia and trachoma. It also has antifungal properties, so would suit Candida patients. It’s even effective to some degree against TB.

Samento may even have anti-cancer action. Renato Rizzi, at the University of Milan, led an experiment with Samento as it relates to cancer causing (mutagenic) substances in smokers. It is well known that the urine of smokers contains mutagenic substances. When given Samento for two weeks, the smoker’s urine returned to normal. This is important because it shows Samento may help lessen the risk of developing cancers and other degenerative diseases.

Dose: all these benefits can be had on a dose of just 3 to 6 capsules a day (good for people who hate swallowing lots of tablets and capsules).

#34 Tea Tree
(Melaleuca alternifolia)

I put this plant substance here, though it could just as easily slot into the aromatherapy section coming up.

The plant was named tea tree after Captain James Cook set foot on the shores of Australia back in 1770. The story goes that Cook brewed up a tea from the leaves of the Melaleuca tree and liked it. It has been used for centuries as an antiseptic by the aborigines.

In 1933, the world-class British Medical Journal described Tea Tree Oil as a powerful disinfectant, non-poisonous and non-irritant. In 1930 the Australian Medical Journal reported extremely good results when using tea tree oil to treat everything from septic wounds to scar regeneration. The report stated that the oil dissolved pus, leaving wounds clean without any apparent damage to the tissues.

In 1937 the same journal noted that blood, pus or other organic matter actually increased the antiseptic properties of tea tree oil by about 10-12%. The Australian Government issued little bottles of Tree Oil in the first aid kits of soldiers serving in the tropical regions during the Second World War. In 1955 by the way the United States dispensaries still had Tea tree oil listed as an active germicidal, with an antiseptic action 11 to 30 times more strong than carbolic acid (phenol).

Tea tree oil is very complex, containing over 100 complex chemical components called terpenes (these are what give it its powerful pine-odor). These terpenes make up 80 to 90 percent of the oil.

Tea tree oil is known to be active against many bacteria, including streptococci, staph, E. coli, and various fungi, such as Candida. In a study of 30 patients who had infections that were resistant to a common antibiotic, one-third were completely cured by tea tree oil compared to just 13 percent who received conventional treatment.

Although tea tree oil is seen by the scientific establishment as very much a quaint folk remedy, science is forcing a reexamination of these outdated attitudes towards folk remedies and alternative medicine. Treatments that were once dismissed as nonsense or even quackery, are now been justified through rigorous scientific testing (in facts drugs are often the real quackery and nonsense).

Now I’m going to shock you all and incense thousands of rabid dogmatists in favor of Tea Tree oil, and point out that it may not be what you want at all.

Beware: Tea tree oil may lead to antibiotic resistance

Repeated exposure to low doses of Tea Tree Oil - a common ingredient in many beauty products - can increase the chances of suffering from “superbug” infections, University of Ulster scientists have revealed.

They discovered that exposure to low doses of Tea Tree Oil make pathogens such as MRSA, E. coli and Salmonella more resistant to antibiotics, and capable of causing more serious infections.

Dr Ann McMahon and Professor David McDowell, members of the University’s Food Microbiology Research Group, said: “We have been growing pathogens such as MRSA, E-coli and Salmonella in low concentrations of tea tree oil. These concentrations are not sufficient to kill the bacteria, but can switch on their defense mechanisms. Unfortunately, these defense mechanisms have the added effect of making bacteria more resistant to antibiotics, and able to cause “harder to treat” infections.”

Tea Tree Oil is used commercially in many products including shampoos, body lotions and toiletries, but there is no legislation requiring manufacturers to state the concentration of tea tree oil in these products.

This increases the risks that people will use low concentrations of tea tree oil, which fail to kill bacteria, but increase their resistance to antibiotics. So, if a person uses tea tree oil products on their skin repeatedly, any MRSA on their skin could develop increased resistance to the antibiotics which are used to control MRSA infections.

“The bottom line is that tea tree oil should not be used at low concentrations - less than 4% - to make sure that bacteria are killed, not just stressed. Otherwise we are just arming the bacteria against treatment by antibiotics.”

Publishing their findings in the Journal of Antimicrobial Chemotherapy, the scientists said: “Although tea tree oil may be an effective antimicrobial agent when appropriately used at high (bactericidal) concentrations, its application at low (sub-lethal) concentrations may contribute to the development of antibiotic resistance in human pathogens”.

http://www.ulster.ac.uk

The best advice therefore seems to be: if you want to use it, use plenty and make sure you wipe out the bacteria population all at once. Otherwise it’s counter productive.

There is one other word of caution too: several cases of gynaecomastia, male breasts in young boys, came about through topical applications of tea tree oil (a shampoo, actually); not many but enough to cause concern.

I pointed this out in one of my publications and was venomously attacked by numerous supposed guardians of “truth”. How dare I attack a near sacred old and venerable cure? The mere fact it was hurting young boys was seen as irrelevant. All sorts of arguments were brought forth, including even Robert Tisserand (see Aromatherapy section #39) writing to me and telling me it was probably an estrogenic substance in the plastic shampoo bottles. He didn’t even accept that the condition cleared up when the use of the tea tree shampoo was discontinued. It must all be faked propaganda science was his attitude.

Denial is a silly attitude and yet it seems to be almost a requirement to be a card-carrying “holistic therapist”. Me? I’d rather stick with at least a semblance of science. All I was asking them was to avoid repeated applications of tea tree in young boys. You’d think I was attacking a sacred cow or even trying to butcher one!

#35 Aloe Vera

The virtues of Aloe vera have been well known since ancient times. The name was derived from the Arabic alloeh meaning ‘bitter’ because of the bitter liquid found in the leaves. It is also known as ‘lily of the desert’, the ‘plant of immortality’, and the ‘medicine plant’.

The fresh gel was said to have been used by Cleopatra to keep her skin soft and young. Aloe vera is truly one of Nature’s absolute miracle plants. It seems to be able to cure almost anything. Always think if it with infections, including Candida and yeast.

The outstanding property of Aloe seems to be its gentleness. It is non-irritating and soothes any and all inflammations, whether infective or not. This makes it ideal to take internally and it is also great for vaginal overgrowths, from Candida, to Trich.

If it does seem to be irritating, that is almost certainly due to additives from the manufacturer. If you suspect that, switch to pure Aloe, direct from the plant.

The chief grower and manufacturer of Aloe, worldwide, is an operation called Univera and their flagship product, “Aloe Gold” I highly recommend. I can tell you how to get it, without joining an MLM (admin@informed-wellness.com)

Think of Aloe in situations like dysbiosis, due to excessive use of antibiotics.

It also works as an anthelminthic [an agent that destroys or expels intestinal worms and/or parasites; vermicide; vermifuge]; an antibacterial; an antifungal/yeast; an antiviral and as a parasite treatment.

It may even prevent cancer: that’s from the Mayo Clinic website, no less! *[http://www.mayo-clinic.com/health/aloe-vera/NS_patient-aloe/DSECTION=evidence]*

That’s pretty versatile.

Dosage

Adults (18 years and older)

Pure Aloe Vera gel is often used liberally on the skin three to four times per day for the treatment of sunburn and other minor burns. Creams and lotions are also available. There are no reports that using aloe on the skin causes absorption of chemicals into the body that may cause significant side effects. Skin products are available that contain aloe alone or aloe combined with other active ingredients.

For internal use, take 1 – 3 ounces daily.

Children (younger than 18 years)

Topical (skin) use of aloe gel in children is common and appears to be well tolerated. Oral dosing is NOT recommended for kids.

#36 Green Tea

I could, I suppose, have put this with the helpful foods and nutrition section; but I have chosen to feature it in herbs, right after the tea tree!

It may be that green tea can help antibiotics be three times more effective in fighting drug-resistant bacteria, even superbugs. This is according to research at Alexandria University in Egypt, in a paper presented in April 2008 at the Society for General Microbiology in Edinburgh, Scotland.

Green tea is a common drink in Egypt, and that means that many people will drink it while taking antibiotics. The researchers wanted to determine if green tea would help or interfere with the effectiveness of antibiotics. That's reasonable.

"We tested green tea in combination with antibiotics against 28 disease-causing microorganisms belonging to two different classes," Dr Mervat Kaseem, of the university's pharmacy faculty, said in a prepared statement. "In every single case, green tea enhanced the bacteria-killing activity of the antibiotics. For example, the killing effect of chloramphenicol was 99.99 percent better when taken with green tea than when taken on its own in some circumstances."

Kaseem and colleagues also found that green tea made 20 percent of drug-resistant bacteria susceptible to cephalosporin antibiotics, an important type of antibiotics to which new drug-resistant strains of bacteria have evolved resistance.

In almost every case and for all types of antibiotics they tested, the researchers found that drinking green tea at the same time as taking the antibiotics appeared to increase the action of the antibiotics and reduce drug resistance in bacteria. In certain cases, even low concentrations of green tea were effective.

Dr. Kassem has a good point to make: "Our results show that we should consider more seriously the natural products we consume in our everyday life," he said. "In the future, we will be looking at other natural herb products such as majoram and thyme to see whether they also contain active compounds which can help in the battle against drug resistant bacteria."

But wait a minute: I found another study that said we could dispense with the antibiotics altogether! It was carried out for USDA and published in a peer reviewed journal, the Journal of Food Protection. Researchers investigated the antibiotic activity of 11 catechins from green tea and compared these to the strength of 5 medicinal antibiotics.

Their findings and conclusion was startling: three of the catechins beat tetracycline and vancomycin (effective at lower dilutions). I quote: “The results show that three of the catechins exhibited exceptional bactericidal activities at very low levels in the nanomolar range. They were more active than medicinal antibiotics such as tetracycline and vancomycin.”

You may remember vancomycin as the “last ditch antibiotic”, 30?

So farmers and food suppliers are getting the story, even if your doctor is not!

[Journal of Food Protection. 69:100-107.]

White Tea

But white tea extract seems to be even more powerful, according to a 2004 study conducted at Pace University. Researchers found that White Tea Extract (WTE) may have prophylactic applications in retarding growth of bacteria, viruses and fungi. Their findings were presented at the 104th General Meeting of the American Society for Microbiology.

The anti-viral and anti-bacterial effect of white tea (Stash and Templar brands) is greater than that of green tea. Moreover, the anti-viral and anti-bacterial effect of several toothpastes including Aim, Aquafresh, Colgate, Crest and Orajel was enhanced by the addition of white tea extract.

White tea was more effective than green tea at inactivating bacterial viruses. Results obtained with the bacterial virus, a model system; suggest that WTE may have an anti-viral effect on human pathogenic viruses.

Studies have also indicated that WTE has an anti-fungal effect on Penicillium chrysogenum and Saccharomyces cerevisiae. In the presence of WTE, Penicillium spores and Saccharomyces cerevisiae yeast cells were totally inactivated. It is suggested that WTE may have an anti-fungal effect on pathogenic fungi.

#37 Spices Are Hot Stuff!

Professor Paul W. Sherman and undergraduate student Jennifer Billing prepared a great report on the antibiotic qualities of spices, published in the March 1998 issue of the journal Quarterly Review of Biology.

Sherman credits Billing, a Cornell undergraduate student of biology at the time of the research, with compiling many of the data required to make the microbe-spice connection: More than 4,570 recipes from 93 cookbooks representing traditional, meat-based cuisines of 36 countries; the temperature and precipitation levels of each country; the horticultural ranges of 43 spice plants; and the antibacterial properties of each spice.

Garlic, onion, allspice and oregano were found to be the best all-around bacteria killers (they kill everything), followed by thyme, cinnamon, tarragon and cumin (any of which kill up to 80 percent of bacteria). Capsicums, including chilies and other hot peppers (killed or inhibited up to 75 percent of bacteria), while pepper of the white or black variety inhibits 25 percent of bacteria, as do ginger, anise seed, celery seed and the juices of lemons and limes.

So even mild spices has some measurable effect it seems.

The Cornell researchers noted that in hot countries nearly every meat-based recipe calls for at least one spice, and most include many spices, especially the potent spices, whereas in cooler counties substantial fractions of dishes are prepared without spices, or with just a few." They theorize that spices were a development chiefly intended to inhibit the bacteria of food spoilage.

If that's correct then it is only natural that spices would have a potent antibiotic capacity— and that's what they found.

As decent scientists, they did consider several alternative explanations for spice use and discounted all but one. The problem with the "eat-to-sweat" hypothesis -- that people in steamy places eat spicy food to cool down with perspiration -- is that not all spices make people sweat, Sherman says, "and there are better ways to cool down -- like moving into the shade." The idea that people use spices to disguise the taste of spoiled food, he says, "ignores the health dangers of ingesting spoiled food." However the micronutrient hypothesis -- that spices provide trace amounts of anti-oxidants or other chemicals to aid digestion -- could be true and still not exclude the antimicrobial explanation, Sherman says. However, this hypothesis does not explain why people in hot climates need more micro-nutrients, he adds. The antimicrobial hypothesis does explain this.

"I believe that recipes are a record of the history of the coevolutionary race between us and our parasites. The microbes are competing with us for the same food," Sherman says. "Everything we do with food -- drying, cooking, smoking, salting or adding spices -- is an attempt to keep from being poisoned by our microscopic competitors. They're constantly mutating and evolving to stay ahead of us. One way we reduce food-borne illnesses is to add another spice to the recipe. Of course that makes the food taste different, and the people who learn to like the new taste are healthier for it."

The following are their Top 30 Spices with Antimicrobial Properties (best to least):

1. Garlic
2.Onion
3.Allspice
4. Oregano
5. Thyme
6. Cinnamon
7. Tarragon
8. Cumin
9. Cloves
10. Lemon grass
11. Bay leaf
12. Capsicums
13. Rosemary
14. Marjoram
15. Mustard
16. Caraway
17. Mint
18. Sage
19. Fennel
20. Coriander
21. Dill
22. Nutmeg
23. Basil
24. Parsley
25. Cardamom
26. Pepper (white/black)
27. Ginger
28. Anise seed
29. Celery seed
30. Lemon/lime

Source: "Antimicrobial Functions of Spices: Why Some Like It Hot," Jennifer Billing and Paul W. Sherman, The Quarterly Review of Biology, Vol. 73, No.1, March 1998

#38 Mushrooms and Fungi

Fungi and mushrooms are related to herbal medicines but not quite the same. Remember most antibiotics come from molds, which are related directly to fungi. According to mycologist expert Paul Stamets, fungi are closer to human life than to bacteria. In fact, they too HATE bacteria, which is why they produce powerful antibiotics (bacteria would attack fungi).

The fungi are a remarkable group of organisms that don't need light to synthesize their own food. What we call mushrooms and toadstools are simply the visible fruiting bodies on what is a much larger organism, covering sometimes up to hundreds of square feet, called the mycelium.

The medicinal properties of fungi have been known since ancient times. In fact Oetzi, the "Ice Man" found in the Tyrolean Alps knew about fungi. His body was frozen 5,200 years ago and found to be remarkably preserved when he thawed out after being discovered intact above the 10,000 line in 1991.

His gear was preserved too and that included a medicine pouch, containing a few walnut-sized dried fungi of the species Piptoporus betulinus, known to have antibiotic properties. It shows that Europe was in there with medicinal mushrooms, though not perhaps to the vast extent of Chinese herbal lore.

Now modern science backs up virtually all that has been claimed for mushrooms and fungi. They are immune modulators, anti-viral, antibacterial and anti-cancerous.

Broadly speaking the edible fungi can be divided into those with culinary merits and those which are purely medicinal. There is some overlap; for example, the shiitake mushroom is a well-known Japanese delicacy and has now spread to the West.

Shiitake stimulates the immune system about a hundred times more than the common white button mushroom. Maitake does much more to aid the immune system than morels, portobellos and chanterelles etc. The very tasty oyster mushroom (Pleurotus ostreatus) is an exception and is enjoyed for its antibacterial, anti-viral, blood pressure moderating and cholesterol reducing qualities.

The key to the effectiveness of mushrooms is a substance called beta glucan, a large and complex polysaccharide (sugar chain), which stimulates the immune system to fight pathogens and malignant cells. Maitake is known to be an effective antitumor agent.

Note that yeast beta glucan, commonly sold via MLM networks, is largely ineffective. They love to quote the science but this, if you look closely, is always experiments on the mushroom beta glucan and is NOT applicable to the products they sell (I have a saying: there is science, damned science and then MLM science!)

Mushrooms contain more than just beta glucan, of course: amino acids such as lysine and tryptophan and vitamins such as C, K, nicotinic acid (vitamin B3), riboflavin (B2) pantothenic acid (B5). There are also other complex substances, such as terpenes and steroids.

Reishi

Let’s start with the Reishi mushroom (Ganoderma lucidum). To the ancient Chinese this mushroom was called Lingzhi meaning “spirit plant”. According to legend, Taoist priests in the first century were supposed to have included the mushroom in magic potions that granted those who consumed them longevity eternal youth and immortality. In Chinese art the Reishi mushroom is a symbol of good health and long life; symbols of it abound everywhere.

Clinical studies have confirmed Reishi has properties as an anti-inflammatory, antioxidant, BP and blood sugar moderator and cholesterol reduce. Reishi contains over 100 bioactive immunomodulatory substances, including beta glucan.

More importantly, studies have shown that Reishi significantly increases three immune signalers known to help the immune system destroy invaders. Japanese researchers say that a hardy system may be able to resist invasions and microbes that weaker host systems may not.

It should also be mentioned that Reishi does not produce some of the negative effects associated with the use of antibiotics, which can often stop the invasion of microbes, but will further weaken the immune response after continued use and are also ineffectual in cases of viruses, pollens and malignant cells.

Maitake

Next let's look at Maitake (Grifola frondosa). Maitake means "dancing mushroom". In Europe it is known as "Hen Of The Woods " and is sometimes called the King Of Mushrooms because of its size.

Maitake is a delicious culinary mushroom but also valued for its medicinal properties. It has known antibiotic activity, although I could only find one really meaningful study to support this. In combination with vancomycin, one of the few remaining antibiotics with some activity against methicillin-resistant Staphylococcus aureus (MRSA), Maitake led to very enhanced survival against Listeria bacteria injected into the abdominal cavity of mice (an often fatal form of food poisoning).

[Kodama N, Yamada M, Nanba H. Jpn J Pharmacol 2001;87:327-332].

Suggested doses: For general protection, 300 mg to 2 grams extract daily, between meals. For therapeutic use against an active infection: 25 to 35 grams extract or more daily until results evident, then taper back to 15 to 30 grams daily, or as recommended by a health care practitioner.

Next, the most famous of all.

Shiitake

It produces a growth inhibiting beta glucan called Lentinan (from shiitake's scientific name Lentinula edodes). Scientists now believe that lentinan and virus-like particles found in shiitake trigger the increased production of various serum factors associated with immunity and inflammation. These so-called lymphokines, such as interferon and interleukin, stimulate the defense system through the proliferation of phagocytes, including macrophages and other immune fighters that attack cancer cells, bacteria, and viruses.

A substance called cortinelin, a broad-spectrum antibacterial agent, which has been isolated from shiitake, kills a wide range of pathogenic bacteria. A sulfide compound extracted from shiitake has been found to have an effect against the fungus that causes ringworm and other skin diseases.

Inonotus obliquus (Chaga)

Stimulates the production of macrophage and NK cells, essential to healthy immune system function. It is rich in immuno-modulating polysaccharides and has been found to be an effective natural anti-inflammatory, anti-tumor, anti-oxidant, anti-viral, and antibiotic.

Pleurotus species

This includes Pleurotus ostreatus (Oyster mushroom). It is enjoyed for its interesting taste, plus anti-bacterial, anti-viral, blood pressure moderating, and cholesterol reducing qualities. It is actually easy to grow at home and is cheap; it should be eaten more widely than it is.

A polycyclic aromatic compound pleurotin has been isolated from P. griseus which possesses antibiotic properties.

Trametes versicolor (Turkey Tail)

Turkey tail mushroom is one of the most widespread mushrooms in the world. Turkey tail can be found on dead trees in just about any hardwood or deciduous forest in North America, Europe, China, Japan and Siberia. As its name indicates, the mushroom looks exactly like a turkey's tail and can be very colorful.

Turkey Tail produces a very powerful immune stimulant polysaccharide, called PSK.

PSK has also been proven as a potent antibiotic, in particular against strains of Staphylococcus aureus, Pseudomonas aeruginosa, Candida albicans, Cryptococcus neoformans, Escherichia coli and Listeria monocytogenes.

[Sakagami, H. & M. Takeda, 1993. "Diverse biological activity of PSK (Krestin): a protein bound polysaccharide from Coriolus versicolor (Fr.) Quel." In Mushroom Biology & Mushroom Products, eds. S.T. Chang J.A. Buswell & Siu-wai Chiu, The Chinese University Press, Hong Kong. pp. 237-

245. Mayer, J. & J. Drews, 1980. "The effect of protein-bound polysaccharide from Coriolus versicolor on immunological parameters and experimental infections in mice." Infection 8: 13-21.

Ng, T.B., J.M. Ling, Z.T. Wang, J.N. Cai & G.J. Xu, 1996. "Examination of courmarins, flavonoids and polysaccharopeptides for antibacterial activity." General Pharmacology Oct; 27(7): 1237-40.]

Turkey tail is a little tough to eat so it is usually found in teas, tinctures, capsules and pills.

It just remains to enter a few words of caution:

Firstly, on no account go out and start harvesting mushrooms yourself and trying to make home-brew extracts. You'll likely kill yourself. The Fungi are an enormously large range of organisms and the difference between safe ones and those which kill can be very subtle. It takes an expert to safely identify edible mushrooms. Of course if you live in France or anywhere civilized which really KNOWS food, you can go into any pharmacy and they will identify any mushrooms you gather. But really, you are safer not collecting your own, unless you are a botanist or better still a mycologist (fungologist).

There is an additional problem, which may not be obvious to you, which is getting a predictable dose. It's something which bugs even commercial fungi extracts which you may find for sale on the Net. Let me warn you that the vast majority of products offered are totally worthless, no matter even if the vendor is sincere and well-meaning. It is very difficult to get a regular, effective dose and keep it consistent.

See, part of the difficulty is that fungi are protected by a membrane coat of chitin. That's same stuff insect cases and your fingernails are made of. Just grinding up mushrooms isn't going to get at the best of the content. It needs a heat process to get at the good stuff, by breaking down the chitin. This, of course, is counter-intuitive to most alternative healers and practitioners, who know that heat normally destroys nutrients.

Let me assure you that in this case it's a waste of money, if not extracted properly, and 99% or manufacturers don't even know this, never mind practice it!

To get you over this problem I have identified a range of mushroom products that meet all the necessary requirements for health and effectiveness.
They are organically grown, under strict natural conditions, in a special facility Oregon State (which also, incidentally, supplies the WHO with fungal medicines, for treating the populations of Third World Countries). These are the most potent full-spectrum myoceutical products ever produced.

You can get purchase details at:

http://www.powershroom.com

#39 Essential oils (aromatherapy)

Well, let's move on. Can you see now there is SO MUCH you can do to fight active infections, without the use of antibiotics.

For instance, did you know that essential oils (aromatherapy oils) have powerful antibiotic properties? That was one of their main uses, when first developed. Remember, this was long ago, in the days before antibiotics or any other significant drugs had emerged.

Aromatherapy oils were once not merely for relaxing and getting in touch with the inner self! They were the best pharmaceuticals of the day!

The story goes back a long way, beginning with a man called Chamberla in 1887. He studied a mixture of oregano, cinnamon, angelica and geranium, and he found that each of these inhibited bacteria growth in a petri dish (special culture dish for microbes).

This is exactly the same test that led to the discovery of penicillin. Yet you never hear that these things do essentially the same! Chamberla found the oils very effective against meningococcus (the cause of meningitis), typhus and Staphylococcus (MRSA is a Staphylococcus, remember). But they were not so good against diphtheria and didn't work at all for anthrax—but then, hardly anything ever does.

Another man called Clavell in 1918 showed that clove bud (Syzygium, also a good homeopathic remedy) kills the TB bacillus at 1 part in 6,000. That was an important stepping stone.

In 1928 a famous case was that of Rene-Maurice Gattefosse, who burned his hand. It became infected and smelled very bad. Gattefosse put it in lavender (to ease the smell, presumably) but was surprised when the lavender clearly had a very beneficial effect. He recovered. Gattefosse went on to learn more about plants and their oils and he was the person who invented the term aromatherapy.

In WW2, a French surgeon called Jean Valnet used essential oils to treat injuries and infections and psychiatric disorders. He was able to use these things to treat gangrene successfully. You never get to hear of this today! Nobody ever says that simple herbal oils could be effective against gangrene.

Whoever has been told “You don’t need an amputation yet, let’s try lavender oil”?

Valnet did a lot of studies and published a book in 1960 called The Practice of Aromatherapy: A Classic Compendium of Plant Medicines and Their Healing Properties.

If you write down Valnet and Aromatherapy and Google that, you will find it. I saw a second-user copy on sale on Amazon for just 5 bucks.

Scientific Testing or Aromatherapy Oils

What astonishes most people (including me, when I first learned it) is that aromatherapy has a good tradition of scientific testing and screening. In the USA it may be looked on askance but in France aromatherapy is quite the thing!

Valnet, in fact, was dedicatedly scientific. He tested numerous plant oils against bacterial culture in petri dishes, exactly the way that led to the discovery of penicillin and other antibiotics. Some were astonishingly good.

1960: Maruzella demonstrated antibacterial and antifungal effects of hundreds of aromatic compounds.

1987: Deininger and Lembke demonstrated antiviral activity of essential oils and their isolated components.

1973: Wagner and Sprinkmeyer did research on a 170 year old blend of distilled oils still available in Germany. The effects of melissa and the other oils in Kosterfrau Melissengeist had been empirically known since Paracelsus (about 1500). They concluded that, with varying degrees of intensity, there was an inhibiting influence on all the bacteria tested, (Pneumococcus, Klebsiella pneumoniae, Staphlococcus aureus haemolyticus, Neisseria catarrhalis, Streptococcus haemolyticus, Proteus vulgaris, Hemophilus influenza, Haemophilus pertussis, Candida albicans, Escherichia coli-Aerobacter group, various Corynnebacteria, and Listeria) and stated the large spectrum of this inhibitory action is as broad as or even greater than that of wide-spectrum antibiotics.

One of the most interesting experiments I came across in my research was carried out in the 1960s by Professor Griffon, Director of the French Police Toxicology Laboratory. He researched the use of aromatherapy or essential oils in cleansing the air. Bottom line… dispersing essential oils into the atmosphere dramatically cut down airborne pathogens. After only 30 minutes of airborne spray cultures were down to zero.

He used a mixture of cinnamon, clove, peppermint, lavender, pine, rosemary and thyme.You probably think these are just sissy smells. But actually these are very very powerful substances, I think I will have time to explain these to you before I open for questions.

(This has enormous implications for CA-MRSA …the respiratory form. If I am not careful, I may run out of time, so let's just do a couple of quick specific…

Drs. Jean-Claude Lapraz and Christian Duraffourd, two Paris-based physicians, organized the world's first phytotherapy conference in Tunisia in May 1993.

Durrafourd and Lapraz have suggested that dosing with essential oils may not be linear in the same way as antibiotics. They found a very small quantity of essential oil had a far greater effect when used in the body than effects measured in vitro (in glass dishes). They also somewhat abandoned the quest for finding one oil for each pathogen. That's something of an allopathic nonsense.

Instead, they see the need for matching the oil to the constitution and make up of the patient in question.

Oregano

Origanum vulgare, also known as Greek or Mediterranean oregano, has good claim to be a natural antibiotic. This is not the cooking oregano however but a rather rare wild herb which grown in the mountainous regions of the eastern Mediterranean. Its scent and flavor are due to the essential oil which has been renowned for its potent antiseptic properties.

When oregano oil was first tested in 1910 it was described as "the most powerful plant-derived antiseptic known" (H. Marindale). It has been found effective both in killing multiple bacteria and preventing their growth. It was shown to be many times more potent than phenol (carbolic acid).

Carvacrol activity crucial. Carvacrol or isopropyl-o-cresol is the active ingredient with powerful antimicrobial properties; it is more potent than carbolic acid.

Oregano has been found to contain several powerful antioxidants such as phenolic acids and flavonoids which protect your tissues against harmful toxins like cancer-causing free radicals *[Tian H, Lai DM. Zhong Yao Cai. 2006;29(9):920-1; Hazzit M, et al. J Agric Food Chem. 2006;54(17):6314-21].*

Oregano oil may be the ‘best of the best’ among essential oils in killing microbial pathogens, killing bacteria, fungus, yeasts and molds. It was shown to be very effective against Candida (1.5 times more potent than the drug Nystatin).

Specifically, it has activity against:

- Escherichia coli (E. Coli)
- Candida albicans
- Bacillus cereus
- Proteus
- Cryptococcus neoformans
- Staphylococcus aureus (Hence MRSA)
- Streptococcus pneumoniae
- Pseudomonas aerugiosa
- Salmonella

Researchers from the Agricultural Faculty, Selcuk University in Turkey studied essential oils from several different plants, including oregano, laurel, marjoram and mint. They tested these oils on certain bacteria (including Bacillus, which can cause bowel infections and diarrhoea) in the laboratory. All of the herbs, including oregano, were found to successfully block the growth of the bacteria studied *[Ozcan MM, Sagdic O, Ozkan G. J Med Food. 2006 Fall;9(3):418-21].*

In an animal study, performed by scientists at the National Agricultural Research Foundation in Greece, oregano was found to be just as effective as the prescription-only antibiotic neomycin for curing diarrhoea caused by the E.coli bug, which is notorious for causing bowel infections.

Commenting on the findings the scientists said: This study indicates that dried oregano leaves may be as effective in the treatment of coli infection as neomycin *[Bampidis VA, et al. J Vet Med A Physiol Pathol Clin Med. 2006;53(3):154-6].*

Dose: Oregano oil for products intended for internal use can be used by adding four drops to a small amount of water and take up to three times a day. For external use it can be applied liberally up to three times a day.

Cinnamon

There are two kinds of cinnamon oil, Cinnamomum zeylanicum and Cassia (next section).

Zeylanicum stands for Sri Lanka. I know that because I lived in Sri Lanka for a couple of years. I was professor at a university there. And it's very effective and wonderful to smell. I love it. But it must be said the other cinnamon (Cassia) is more powerful.

A nice idea I found is to make a special flavored honey with this oil:

> To 2 oz. of maple syrup or 2 oz. of honey add: 3 drops of Ginger Essential Oil 1 drop Cardamom 1 drop Cinnamon bark oil 3 drops Vanilla Essential Oil, pure 8 drops of certified organic Orange or Mandarin Orange oil.

Cassia

Cassia (Cinnamomum cassia, or "hot cinnamon") is native to the south-eastern parts of China and to a lesser extent in Vietnam, Burma and India. It is cheaper and more abundant than the Sri Lanka variety. Its value dependant mainly on the percentage of cinnamic aldehyde which it contains. It is heavier, less liquid, and congeals more quickly than the Ceylonese variety.

Cassia has recognized antimicrobial properties and is often used for diarrhea. The essential oil is a powerful germicide, but being very irritant is rarely used in medicine for this purpose.

One study carried out at the Department of Agricultural Biology and the Research Center for New Bio-Materials in Agriculture, College of Agriculture and Life Sciences, Seoul National University (Koera) showed Cassia had 1/5th - 1/10th the power of tetracycline and chloramphenicol against intestinal bacteria.

Not a problem, since we could always take more Cassia than antibiotic capsules. Interestingly, Cassia had no effect on Lactobacillus, a prime probiotic organism. So there would not be the disastrous after effects of the two antibiotics used in the trial. *[J. Agric. Food Chem., 1998, 46 (1), pp 8–12].*

An interesting analysis from the University of Hong Kong showed that typical Cassia preps. have around 98% cinnamaldehyde and were effective against a wide range of pathogens, including Staphylococcus aureus, E. coli, Enterobacter aerogenes, Proteus vulgaris, Pseudomonas aeruginosa, Vibrio cholerae, Vibrio parahaemolyticus and Samonella. *[Am J Chin Med. 2006;34(3):511-22].*

Thyme

Thyme was used medicinally by the Egyptians, Greeks, and Romans. Most present day research has centered on thyme’s ability as an antibacterial and anti-infectious agent, even when diffused in the air.

Thyme oil is composed of several different components that show antimicrobial activity (at least: carvacrol, thymol and linalool).

In vitro studies have shown that thyme oil has valuable broad spectrum activity against Salmonella typhimurium, Staphylococcus aureus, Escherichia coli, and a number of other bacterial species.

Thyme essential oils showed some of the strongest killing power against MRSA and VRE antibiotic-resistant bacteria, according to studies at the Western Infirmary, Glasgow, UK. According to Jean Valnet, M.D., thyme oil kills the anthrax bacillus, the typhoid bacillus, meningococcus, and the agent responsible for tuberculosis and is active against salmonella and staphylococcus bacteria.

The College of Pharmacy study at Oregon State University, also confirmed thyme is substantially more effective than most others in killing MRSA, including the virulent strain USA300. Oregano and thyme together are consistently the most effective in inhibiting the growth of all three bacteria; MRSA, Escherichia coli, and Acinetobacter calcoaceticus. Dr. Ihsan Edan

AlSaimary, of the Univ. of Basrah, has also confirmed these other research results.

Thyme is also a broad spectrum disinfectant that helps to kill germs (including E.coli) on surfaces all around your home. *[Marino M, Bersani C, Comi, G. Antimicrobial Activity of the Essential Oils of Thymus vulgaris L. Measured Using a Bioimpedometric Method. Journal of Food Protection, Volume 62, Number 9, September 1999, pp. 1017-1023(7)].*

In another study, published in Bionews online, it was found that Escherichia coli and Salmonella typhimurium were susceptible to thymol and carvacrol (both found in thyme oil) and eugenol (found in clove oil).

Activity of natural antimicrobial compounds against Escherichia coli and Salmonella enterica serovar Typhimurium Letters in Applied Microbiology *[N.A. Olasupo, D.J. Fitzgerald, M.J. Gas-son and A. Narbad]* Volume 37 Issue 6 448 - December 2003.

Thyme oil is very strong stuff by the way, don't go rubbing it on your skin without diluting it. Something like a massage oil.

Black Seed Oil

Black seed oil is not often included in Western reviews of essential oils. Maybe there is a certain amount of prejudice in this. But I think black seed oil could stand as high as Oregano oil and does not taste as repelling.

Black seed oil (Nigella sativa) has been known for centuries in Middle East countries. The black seed is known in Arabic as the habbutual barakah (the seed of blessing). It is sometimes called the Black Cumin Seed, however this is not the same as Black Cumin which the Indians refer to as Jeera/Zeera.

The Prophet Mohammad stated in his “Hadith” that black seed oil cures every illness except death. A small phial of it was apparently found in the tomb of Tutankhamen.

Now modern science has come to its support. Researchers around the world have confirmed the anti-bacterial and anti-mycotic effects of black seed oil. Health practitioners in various countries around the world are using the oil against inflamation of all sorts as well as fungi infections. Black Seed extracts have been found to stimulate bone marrow and immune cells, raise interferon production and increase B cells (which produce antibodies).

The active ingredients of black seed are nigellone, thymoquinone, and fixed oils. Beta si-tosterol has also been identified as a key marker in black seed oil. Beta sitosterol has been shown to exert protective effects and reduce symptoms of BPH (benign prostatic hyperplasia). *[Lancet 1995;345(8964):1529-32]*

Black seed also contains significant proportions of protein, carbohydrates and essential fatty acids. Other ingredients include linoleic acid, oleic acid, calcium, potassium, iron, zinc, magnesium, selenium, vitamin A, vitamin B, vitamin B2, niacin, and vitamin C.

Dose: a teaspoonful 3 times a day of the oil. It's bitter and irritating but can be mixed with tea, coffee, carrot juice, yoghurt or other carriers. For respiratory infections it can be rubbed in the chest; for sinusitis and catarrh, 3- 4 drops in each nostril.

General use: Ibn Senna known in the West as Avicenna (980-1037 AD) the Persian-born Islamic philosopher and physician who wrote the great medieval medical text, described the power of the black seed to stimulate the body's energy and banish fatigue.

A teaspoon of black seed oil mixed in a glass of orange juice with breakfast keeps you active all through the day. A teaspoon of black seed oil mixed in a hot drink after supper gives you a quiet sleep all through the night.

Mountain Savory

Long established as an antibiotic substitute, mountain savory (Satureja montana) was in third place, behind lemongrass oil and lemon mytle oil, according to a 2008 study *[Inhibition of methicillin-resistant Stapphulococcus aureus (MRSA) by essential oils; Sue Chao, Gary Young, Craig Oberg, and Karen Nakaoka; Flavour and Fragrance Journal, 2008; 23: 444-449]*.

One important study examined in vitro antimicrobial activity of the essential oils of the aerial parts of Satureja montana L. and Satureja cuneifolia. The major compound of S. montana oil was the phenolic monoterpene carvacrol (45.7%).

The antimicrobial effects of S. montana and S. cuneifolia oils were found to have a broad spectrum activity against multidrug-resistant pathogens. These oils were active against all the test strains, with the exception of

Pseudomonas aeruginosa. Compared with S. cuneifolia, savory oil exhibited greater antimicrobial activity.

The maximum activity of savory oil was observed against Escherichia coli, MRSA and against the yeast (Candida albicans). The essential oil of S. cuneifolia was also found to inhibit the growth of medically important pathogens such as S. aureus and E. coli. Fungicidal activity for both oils against C. albicans and S. cerevisiae (food yeast) was also observed. *[Volume 18 Issue 12, s 967 – 970].*

Useful antibiotic formula:

Mix together the following: 10 drops lemon; 8 drops mountain savory; 3 drops oregano; 1 tablespoon of V6 food grade oil. 2 ounces of grape juice.

Take 2 tablespoons a day of this mix. (based on a suggestion by Dr. Gary Young).

V6 is an essential oil carrier, consisting of fractionated Cocos Nucifera (Coconut) Oil, Sesamum Indicum (Sesame) Seed Oil, Vitis Vinifera (Grape) Seed Oil, Prunus Amygdalus Dulcis (Sweet Almond) Oil, Triticum Vulgare (Wheat) Germ Oil, Helianthus Annuus (Sunflower) Seed Oil, and Olea Europaea (Olive) Fruit Oil.

If you want, put it in a Double O gelatin or vegetable capsule instead of the V6. In which case use these proportions: 5 drops Lemon Essential oil 3 drops Oregeno Essential oil 5 drops Mountain Savory Essential oil.

Take one capsule 3 times daily.

Pregnancy and Strep B

For pregnant women, Strep B can be a hazard (and to the neonate too). Signs, Symptoms of Neonatal GBS Infection Unaffected by Intrapartum Antibiotics *[Medscape Summary]*

"Antibiotic exposure also does not delay the onset of clinical signs of infection . . ."

[The Influence of Intrapartum Antibiotics on the Clinical Spectrum of Early-Onset Group B Streptococcal Infection in Term Infants. Bromberger P, Lawrence JM, Braun D, Saunders B, Contreras R, Petitti DB, Pediatrics 2000 Aug;106(2):244-250]

In fact it probably causes it! [Potential consequences of widespread antepartal use of ampicillin. Towers CV, Carr MH, Padilla G, Asrat , Am J Obstet Gynecol 1998 Oct;179(4):879-83]

"The increased administration of antenatal ampicillin to pregnant women may be responsible for the increased incidence of early-onset neonatal sepsis with non-group B streptococcal organisms that are resistant to ampicillin. At this time penicillin G, rather than ampicillin, is therefore recommended for prophylaxis against group B streptococci. In addition, future studies are needed to determine whether alternate approaches, such as immunotherapy or vaginal washing, could be of benefit."

So, do the following:

Soak an ORGANIC tampon in 15 drops Lemon Essential oil 9 drops Oregeno Essential oil 15 drops Mountain Savory Essential oil 1 tsp carrier (V-6 food grade) oil Leave soaked tampon in overnight.

The Journal of the Science of Food and Agriculture Sep 2006 did a nice piece on growing savory (Satureja hortensis) in Scotland!) *[ref: S.G. Deans and K.P. Svoboda, Antibacterial activity of summer savory (Satureja hortensis L.) essential oil and its constituents. J. Hort. Sci., 64, 205-210 (1989)].*

Lavender

If you Google 101 essential oil uses, the first 23 are lavender!

Long prized for its healing properties, written records of the use of lavender for medicinal purposes date back as far as 60 AD and the writings of Dioscorides. In ancient Rome lavender was recognized for its healing and antiseptic qualities, its ability to deter insects, and for washing. In fact, its name stems from the Latin “lavare”, meaning to wash. During the First World War, when modern antibiotics were sparse, lavender was used to dress wounds and helped to heal scar tissue and burns.

Lavender’s Antibiotic Properties

Lavender is renowned for its antibiotic properties. Studies have shown that the essential oil of lavender, particularly when combined with Geranium oil, is capable of killing some Staph infections. Other studies have reported that lavender is good for treating ear infections, and is mild enough to treat such symptoms in children. Recently, four new chemicals have been isolated from lavender plants, and are believed to be beneficial for the treatment of Candida. Lavender tea (infusion) may be made from the dried flowers, 1 1/2 tsp. flowers to 8oz.water.

This can be drunk up to 4 times a day.

Lavender reduces fever and is a strong antiseptic and has been used to fight diphtheria, strep throat and pneumonia.

Lavender tea, or a few drops of oil in a glass of water, used as a gargle eases sore throats and laryngitis, can also soothe toothache.

Wounds and topical application

Lavender oil is an exception to all the other essential oils, in that it does not need to be diluted in a carrier oil because it is so gentle. It is also safe to use on infants and children.

Lavender is often used to treat scalds, minor burns, cuts, grazes, inflammation, eczema and dermatitis.

Geranium

This is Pelargonium sidoides, or South African Geranium NOT P. graveolens, the garden variety geranium. It's excellent for bronchitis. This is a big problem in primary care.

One study followed 476 adults with a clinical diagnosis of bronchitis from primary care medical centers in Germany. It excluded patients who had antibiotic therapy in the previous 4 weeks or asthma, severe heart, renal, or liver disease; immunosuppression; drug or alcohol abuse; and pregnancy or lactation.

South African geranium, Pelargonium sidoides (NOT the familiar pot plant)

This was a randomized, double-blind, placebo-controlled prospective study. Treatment consisted of 1.5 mL of aqueous ethanolic extract (11%) of Geranium root 3 times daily for 7 days. Placebo was matched for color, taste, smell, and viscosity. The breathlessness severity scale scores (BSS) improved in the geranium group over the placebo group. This was a well-structured trial.

[Matthys H, Eisebitt R, Seith B, Heger M. Efficacy and safety of an extract of Pelargonium sidoides (Eps 7630) in adults with acute bronchitis. A randomized, double-blind, placebo controlled trial. Phytomedicine 2003; 10(Suppl 4): S7-S17.]

Pelargonium is also an effective treatment for throat infections. Severe sore throat in children is often due to Streptococcus. In a double-blind clinical trial carried out in the Ukraine, children aged 6 to 10 with acute Strep throat infections were treated with pelargonium root extract or a placebo. Pelargonium resulted in much greater improvements than placebo, reduced the severity of symptoms and shortened the duration of illness by two days on average *[Altern Health Med 2003; 9(5): 68-79].*

Another study showed it was great for acute sinusitis, which is usually due to a bacterial or fungal infection after a cold. A recent study in Germany, involving 361 people with acute sinusitis who were treated with pelargonium root extract, found that at the end of the four week treatment period 80 per cent of patients recovered or experienced a clear improvement in their symptoms *[Z Phytother 2007; 28: 58-65]*.

One in vitro study showed killing drug-resistant Staph, so it may have a place in therapy for MRSA [Phytomedicine 2003; 10 (Suppl 4): 18-24]. The recommended dosage is 20mg of the root extract, three times a day, or 800mg of dried, powdered root, taken twice a day, for 10 days.

General Warning: All I have written applies ONLY to therapeutic grade essential oils and not to recreational products, so please do not try some of its recommendation unless the oils you are using are at least AFNOR or ISO certified. Merely the word “pure” is not enough.

#40 Competitive Inhibition

In this section I'll be explaining another good strategy for controlling pathogenic bacteria. That's to encourage competition, so that safe, friendly organisms proliferate and occupy so much of the available territory (our bodies) that there is no room for wild forms to get started.

If the friendly horde that we carry should be replaced by nasty disease-causing renegade bacteria from the wild, our health can be compromised, even up to the point of death. All that really stops this happening is the massive presence of largely friendly—or at least nonhostile—forms. They are our main allies in fighting disease. Yet doctors ignore them and, indeed, even ravage and kill them in massive numbers, by the careless and unnecessary use of antibiotics.

Don't be smug. You do your part in hurting our friends too. Antiseptic soaps, sprays and body washes, that the TV ads convince you are really necessary, will kill off good forms as well as the ones you're scared of.

1. It's been said that we are landlords to the host of bacteria that live on us. As such, we should be responsible and caring owners and look after our guests! We can tend our crop of healthy organisms in several ways. We can feed the good bacteria.

2. We can import more good bacteria and add to them.

3. We can get other non-bacterial life forms to attack bacteria, much as introducing a new predator into the eco-landscape can result in wiping out the species on which it preys.

It may surprise you that the scientific study of these three strategies is well under way.

Adding to bacterial colonies is called "probiotics" and nurturing what's already present, with helpful nutrition, is called "pre-biotics".

Eubiotics, which covers both, means simply "good life"; in other words eating and drinking well and behaving sensibly. Synbiotics is another term you may encounter. I'll explain it further along.

The third approach we do using bacteria-hostile viruses which we call bacterio-phages; that means “bacteria eaters”. We usually say phages for short. You can probably guess how that one works.

Next I’ll cover each of these processes in more detail.

Probiotics

Probiotics, as I just remarked, means adding to the quantity of friendly organisms we carry. It is rather like artificially re-stocking lakes and rivers with fish. If you do it well—and actually ban fishing—you can build up stocks rapidly, because the fish will naturally reproduce and multiply. It’s the same with helpful bacteria.

The first ever recognition of this process of probiotics was the use of yoghourt.

Russian scientist and Nobel laureate Eli Metchnikoff was the first to suggest that it would be possible to modify gut flora by adding friendly microbes to squeeze out the damaging ones.

At the beginning of the 20th century Metchnikoff, at that time a professor at the Pasteur Institute in Paris, introduced the theory that the aging process results from the activity of putrefactive (or proteolytic, meaning “breaks down protein”) microbes producing toxic substances in the large bowel. According to Metchnikoff these compounds—including phenols, indols and ammonia—were responsible for what he called “intestinal auto-intoxication”, which caused the physical changes associated with old age.

Metchnikoff knew that milk fermented with lactic-acid bacteria inhibits the growth of proteolytic bacteria because of the low pH (acidity) produced by the fermentation of lactose. He had also observed that certain rural populations in Europe, for example in Bulgaria and the Russian Steppes who lived largely on milk fermented by lactic-acid bacteria (yoghurt) were exceptionally long lived.

Based on these facts, Metchnikoff proposed that consumption of fermented milk would provide the intestine with harmless lactic-acid bacteria and decrease the intestinal pH and that this would suppress the growth of proteolytic bacteria (something else for the fans of alkali foods to think about).

Metchnikoff himself introduced in his diet sour milk fermented with the bacteria he called "Bulgarian Bacillus" and found his health benefited.

In 1900 Henry Tissier, also from the Pasteur Institute, isolated a Bifidobacterium. Tissier showed that bifidobacteria are predominant in the gut flora of breast-fed babies, and he recommended administration of bifidobacteria to infants suffering from diarrhea.

The stage then switched to Japan. In the 1920s Japan was not the wealthy nation it is today. Many children lost their lives to infectious diseases and other maladies brought on by malnutrition. In 1930. Dr. Minoru Shirota, working in a microbiology lab at Kyoto Imperial University's School of Medicine, succeeded in developing a stronger strain of lactic acid bacteria which would work to destroy the harmful gut bacteria. This bacterium was named Lactobacillus casei shirota after Shirota. It produced a yoghurt-like drink called Yakult, which was introduced to the market in 1935.

Subsequently, other species have been introduced to conventional medicine, including Lactobacillus rhamnosus, Lactobacillus casei, and Lactobacillus johnsonii, because they are normal intestinal species with beneficial properties.

Symbiotics

The old days of just "Lactobacillus acidophilus", basically a bovine form, are gone. As I said in my 1993 book, we need human strain Lactobacillus. The rest just don't thrive and soon die. Lactobacillus bulgaricus had the same problem.

But more importantly we need lots of Bifodobacteria, which is our main friendly passenger down there in the gut.

My more recent caution has been to get properly prepared organisms. The general technique of just crushing them into a tablet just kills the organisms, as you would expect, making a joke of the supposed population values given on the label.

Look instead for non-compacted preparations.

The other new trend is feeding your bowel flora the right stuff. No question that lots of yeasty foods to support the invaders, plus all the sugary and starchy foods that they can ferment for food, is a recipe for disaster. You need sugar-free healthy foods.

Keeping off ALL sugary and starch foods is tough. But you can take prebiotics, including galactooligosaccharides from human breast milk. Not for everybody but at least the conventional doctors have access to it. Mixing prebiotics and probiotics like this is called synbiotics (syn- just means coming together, as in synthesize).

If you are not in a doctor's program for synbiotics, you need to consider rice bran and Inuflora, from the jerusalem artichoke. These are fiber products and provide a matrix in which the probiotics can establish themselves. One is soluble and one insoluble and the balance between the two makes an ideal, relatively inexpensive and natural substitute for fancy oligosaccharides.

There is the additional benefit of binding heavy metal complexes, when these are excreted into the gut via the hepato-biliary axis (liver and gallbladder excretion route).

Of course you probably know that holding yeasts, ferments and fungals in check is also a matter of avoiding dietary sugar, which feeds these organisms. Also the avoidance of all dietary fermented products will help, notably cheese, mushrooms, bread and so on.

So, you see, we've gone way beyond just adding yoghourt to our diets!

To summarize:

Bad foods (avoid):

- Yeasty, starchy, sugary, all artificial foods
- Limit fruits (sugar and fructose)
- Yoghourt preparations that have been sweetened for taste (rather defeats the purpose)

Good foods (eat plenty):

- Rice bran
- Inuflora. Jerusalem artichoke
- Any and all fiber vegetables, legumes and salad foods

If you Google “Inuflora” you will find scores of references and places where you can buy online.

This Is So Good The Veterinarians Are Onto It!

Presenting his work at a Society for General Microbiology meeting in the UK on 2 April, 2009, Colin Hill of University College Cork described how his team had used three animal models of disease that had human counterparts — bovine mastitis, porcine salmonellosis (a gastrointestinal disease) and listeriosis in mice (an often fatal form of food poisoning) — to demonstrate the protective effects of probiotics.

The researchers used their own probiotic preparations containing safe bacteria such as Lactobacillus species newly isolated from human volunteers. Hill said that in all three animal diseases, the research group observed a positive effect, in that the animals were “significantly protected” against infection.

The team also used probiotics to control disease in animals that were already infected. The results of these tests, said Hill, showed that administering these safe bacteria to an infected animal was as effective as the best available antibiotic therapies in eliminating the infectious agent and resolving the symptoms.

In each instance the protection was linked to a particular bacterial species, and the mechanism of action varied from direct antagonism, where the probiotic directly kills the pathogenic bacteria, to effects mediated by the host immune system.

For example, Lactobacillus salivarius UCC118 protected mice against listeriosis (a disease which can affect pregnant women) by producing an antimicrobial peptide that eliminates Listeria monocytogenes, the causative organism of Listeria.

Probiotics directly help the effects of antibiotics!

It gets better!

Helicobacter pylori is present in about 25% of patients treated for gastritis or peptic ulcer disease (many now consider this microbe to be the "cause" of peptic ulcers). Standard treatment with two antibiotics (clarithromycin plus amoxicillin) and a fancy drug called a "proton pump inhibitor" (esomeprazole) failed to eradicate the troublesome microbe.

But when a team from the University of Pisa in Italy added supplementation with probiotics (lactobacillus etc.), the cure rate was boosted and side effects reduced (less nausea, diarrhea, metallic taste in the mouth, inflamed tongue, and abdominal pain).

So now scientists think all 5 together are great.

You know what the joke is? If they just gave the probiotics they would work as well as or better than all the other drugs put together. And without any side effects. Duh!

[Am J Gastroenterol. 2007;102:951-956]

How To Take Probiotics

Forget the cheery TV ads with celebrity backing. Most commercial gut flora yoghurt products are of little value. They come with sugar! The bacteria are usually dead.

You need a good source of HUMAN strain lactobacillus. But also Bifidobacteria, which is normally the most common friendly organism in our gut, should be in the supplement as well.

Even then, the probiotic is not bound to establish itself successfully in our gut, so be very selective about what product you choose.

Beware of the tablet delivery. Most of those contain only crushed (dead) bacilli, no matter what claims of bacteria count! Capsules or liquid presentation are the safest.

Human Probiotic Therapy

(Faecal Transplants)

Are you ready for this? One of the ultimate ways to enhance bowel flora is to transplant faecal matter form healthy people into those who are sick and in danger.

Thye Yuck! factor is high but it does work and is practiced in many orthodox medical facilities around the world. In Sweden, human faecal transplants are the treatment of choice of life-threatening membranous ulcertaive colitis, a diosease typically caused by overuse and abuse of antibiotics, leading to overgrowth of Clostridium difficile. It is often fatal.

In fact the faeces are administered either via a nasogastric tube, straight into the stomach, or via an enema. The latter is surely to be preferred, since the colon is where these probiotic bacteria belong.

Of course there are precautions, like testing the "donor" for other pathogens and parasites, HIV and hepatitis.

Family members are preferred as donors, since we share about 80% of the bacteria colonists in the gut with our mother and siblings. But anyone can be a donor, provided he or she is healthy. Donors usually provide faecal material for free (well, you could hardly charge for it, come now!)

So what has all this to do with Clostridium difficile? This is a pretty toxic organism, as are most Clostridia species. But it's not native to our gut. It only takes hold when overuse of antibiotics have displaced the good guys. So it's a man-made complaint.

When healthy human gut flora is ploughed back, so to speak, C. difficile is squeezed out, exactly as we explain probiotics working. And before you ask: yoghourt won't work. This is a killer bug.

But faecal transplants seem to work very well indeed: close on 100% success in fact. People stretchered into hospital at death's door make a remarkable recovery, often in as little as 48 hours! That's amazing. It does suggest that the mechanism is specific toxins produced by friendly bacteria from a healthy gut being able to kill the C. difficile pretty quickly.

One study even resorted to my jokey approach and was titled: Bacteriotherapy using gut flora: toying with human motions" *[Journal of Clinical Gastroenterology, vol. 38, p. 475]*.

However some 200 or so other studies were more serious.

One by Borody and colleagues details 6 cases of severe chronic ulcerative colitis treated using faecal bacteriotherapy. Complete reversal of symptoms was achieved in all patients in just 4 months, by which time all colitis medications had been discontinued as unnecessary *[Borody T, Warren E, Leis S, Surace R, Ashman O (2003). "Treatment of ulcerative colitis using fecal bacteriotherapy.". J Clin Gastroenterol 37 (1): 42–7].*

Bacteriophages

Phage – means eating in science. Esophag(e)us, the gullet, is the same word in a different form and phagocytes ("cell eaters") are immune cells that gobble up bacteria and viruses.

A bacteriophage is any one of a number of viruses that infect bacteria. Bacteriophages, or "phages" for short are viruses that invade only bacterial cells where they take over the genetic code and release more phages. When the bacterium dies, it releases these new phages and so they grow on and on. They can wipe out bacteria in no time.

Phages are estimated to be the most widely distributed and diverse entities in the biosphere. These ubiquitous organelles can be found in all reservoirs populated by bacterial hosts, such as soil or the intestines of animals. One of the densest natural sources for phages and other viruses is sea water and up to 70% of marine bacteria may be infected by phages.

You can find out more by visiting http://www.phage.org, which defaults to the Mansfield Ohio State University website.

A Historical Note

Following the discovery of bacteriophages by Frederick Twort and Felix d'Hérelle early in the 20th C, phage therapy was immediately recognized by many to be a key way forward for the eradication of bacterial infections.

Phages have since been used for over 60 years as an alternative to antibiotics in the former Soviet Union and Eastern Europe. In the USA during the 1940s, commercialization of phage therapy was undertaken by the large pharmaceutical company, Eli Lilly.

Whilst knowledge was being accumulated regarding the biology of phages and how to use phage cocktails correctly, early uses of phage therapy were often unreliable.

Unfortunately, when antibiotics were discovered and marketed widely in the USA and Europe, Western scientists mostly lost interest in further use and study of phage therapy for some time. But research continued in Russia and the Eastern Bloc.

I was taught about phages in med school in the early 1960s; then they seemed to disappear from view.

Phages are now making a comeback. They are seen as a possible therapy against multi drug resistant strains of many bacteria.

Pros and Cons Of Phage Therapy

There are some very important benefits with phage therapy. There are also some drawbacks.

Important benefits of phage therapy:

- Specific to the bacteria so it doesn’t wipe out gut flora
- Harmless to the host
- Few side effects and do not stress the liver
- Only needs one dose because they replicate

Disadvantage:

- Will only kill the exact bacterium so needs a cocktail approach for multiple infections.

Biofilm, no not a Bond movie

Pseudomonas aeruginosa—which can cause a number of unpleasant and highly intractable infections, notably of the genito-urinary tract, but also ear infections—is particularly hard to treat because it typically wraps itself in a biofilm - a layer of sugars and proteins that make it up to 1000 times more resistant to antibiotics as a non-biofilm from the same species.

Now it has been found that a single dose of a phage called Biophage-PA, selectively attacks P. aeruginosa, breaks down the biofilm, and destroys the bacteria. It has been used successfully to treat long-term sufferers of antibiotic-resistant ear infections. Andrew Wright from University College London Ear Institute and colleagues studied 24 people with severe ear infections (Clinical Otolaryngology, accepted for publication). Half the volunteers were given Biophage-PA and the rest received a placebo.

Pain, pus secretion and inflammation were reduced in both groups, but the relief was double in the group on the treatment. The number of target bacteria in the ear was significantly reduced in the phage-treated group, while there was no significant reduction in the placebo group.

By the end of the six-week trial, three patients on the phage were clear of infection. Big promise there!

One of the important advantages of phage treatment is that it only takes one dose; this gets round the problem that patients often forget to complete their course of antibiotics, thus inviting resistance.

Of course it is entirely possible that bacteria might evolve resistance to phages too. It might almost be predicted. But so far it isn't documented.

#41 Homeopathy

Now a New Science Altogether

Now, pay close attention, because this section will probably save more lives than any other in this report. I'm going to briefly describe homoeopathy, and explain why it is a wonderful, effective and safe approach to dealing with infections of all kinds (not just bacterial).

Homoeopathy has been around over two centuries, predates the vaccination and has demonstrated its effectiveness against countless infectious diseases and epidemics. In epidemics of the past, people who took homeopathy often survived unscathed, while others died in significant numbers. Yet, you will still hear it maligned and most conventional doctors and a pseudo-scientists claimed that it's been "proven not to work".

This is nonsensical logic. You cannot, logically and scientifically, prove something doesn't work! You can only fail to prove that it does work!

In fact detractors and bigots simply ignore the volumes of scientific evidence that homoeopathy works. Criticisms come from ignorant, mis-informed critics.

For example I always quote a major study reported in the prestigious journal The Lancet, [Lancet Dec 10, 1994:344, p 1585]. It caused a great deal of controversy and the then editor took the unusual step of mocking it in an editorial. But he did have the decency to admit the trials were done to a very vigorous standard (most drugs trials find their way to publications with far inferior standards).

Because he had to admit that using a double-blind randomized trial clearly showed that homeopathy worked, he wrote in his editorial:

"... we must ask if the technique of randomized controlled trials is fundamentally flawed, and capable of producing evidence for effects that do not exist, by, for example, the effects of clinicians' expectations of outcome transmitting by subtle effects that circumvent double blinding?" (Lancet 1994; 344;1601- 06) I regard this as the worst kind of bigotry. Instead of drawing the conclusion that homeopathy is indeed highly effective, he dared to suggest that randomized double-blind studies, the benchmark of medical science, maybe didn't really work!

Well, in reading this I know you are not too swayed by conventional rhetoric. Nevertheless, I urge that you do not permit this crass ignorance to enter your thinking. To go along with it could some day save your life.

If you trust your author, let me assure you that homeopathy works and it will save innumerable lives in a pandemic. As I indicated in an earlier section, I relied on it entirely when I lived in a malarious zone for 2 years. I took no prophylactic. I simply took a few doses of homeopathic anti-malaria preparation and then after a few weeks didn't even continue. I was bitten many times a day by mosquitoes but remained quite healthy.

Dana Ullman has done a great job of compiling lots of science papers showing uncontrovertibly that homeopathy does work. You can read her findings here:

http://www.homeopathic.com/articles/view,132 *[DANA ULLMAN, MPH, is one of America's leading advocates for homeopathy. He has authored 10 books, including The Homeopathic Revolution: Why Famous People and Cultural Heroes Choose Homeopathy, Homeopathy A-Z, The Consumer's Guide to Homeopathy, Homeopathic Medicines for Children and Infants, Discovering Homeopathy, and (the best-selling) Everybody's Guide to Homeopathic Medicines, with co-author Stephen Cummings, MD].*

Background

Homeopathy is a discipline developed during the 1790s entirely by one man: Samuel Hahnemann, a German physician. Experimenting on himself with the anti-malarial drug quinine, Hahnemann noticed that large doses of the medicine actually caused malaria-like symptoms, while smaller doses cured the symptoms. From this, he advanced his concept of Similia similibus curentur, or "let like be cured with like." Hahnemann then developed an extensive system of medicine based on this concept. He named it homeopathy, from the Greek words homoios (the same) and pathos (suffering).

As the homeopathic healing system grew in popularity during the 1800s, it quickly attracted vehement opposition from the medical and apothecary professions. Since the early 1900s, when the American Medical Association and pharmacists waged a battle against it, homeopathy has been neglected and sometimes ridiculed by mainstream medicine.

Proponents of homeopathy over the years have included Louisa May Alcott, Charles Dickens, Benjamin Disraeli, Johann Wolfgang Goethe, Nathaniel Hawthorne, William James, Henry Wadsworth Longfellow, Pope Pius X, John D. Rockefeller, Harriet Beecher Stowe, William Thackeray, Daniel Webster, and W. B. Yeats. England's royal family has employed homeopathic practitioners since the 1830s *[Source: About.com]*.

The only problem with this wonderful, gentle therapy, is that it takes some skill to administer. I hope, if there is a pandemic scare, or if you developed a drug-resistant strain of bacteria, you can get advice from someone properly trained.

Before you ask, Homeopathic Medicines are generally regarded as safe by the FDA (GRAS).

Homeopathy in 3 bullet points

Homeopathic remedies are almost always made from natural materials—plant, animal, or mineral substances or even venom from snakes or stinging insects. Some remedies may be given in a spray, ointment, or cream, but the most common forms of administration are liquid dilutions or dropped onto tablets or tiny pillules.

Here are the key points in how the preparations are made.

1. The key principle is "like treats like". A remedy which mimics the disease is chosen = similimum. This is the origin of the name "homo-" means: same as. It's completely different to the conventional idea of fighting and suppressing the symptoms.

2. The remedy diluted many times: 10x10x10 (the X series, for example, 10X) or 100x100x100x100 etc (the C series eg. C30). Note that in Europe the X series is usually named D (for decimal), such as D8. Because this process makes the remedy work more effectively, it's called potentization.

3. After each dilution, the bottle is shaken or succussed as it's called.

The part that drives conventional thinkers crazy is the idea that diluting something makes it more effective. They "know" that diluting only makes it weaker. In fact they get very angry about the fact that most dilutions are so high that there cannot be even one single molecule of the original substance left in the remedy bottle.

But that's because they are bound by ridiculously constraints to their ideas and beliefs, which is that the only valid model is biochemical. In other words it needs "stuff" present to make it work.

The real nature of homeopathy seems to be acting as a signal to transfer a "message". This message can be passed onto the next solution and the next, no matter how many times the dilution process takes place.

I haven't time here to go into the whole science of homeopathy but let me offer you an illustration (overleaf), showing a spectrographic study of increasing dilutions of a remedy called belladonna. It should be a kind of "proof" to encourage you.

Read from the bottom up. The first solution is what we call the mother tincture (1:250). Each step in dilution, D1, D6, D12 and so on are progressive dilutions of this tincture, in the 10X sequence (decimal, hence D).

In this chart, any spikes above the horizontal line means that the belladonna solution is actually transmitting electro-magnetic energies (like radio or light waves). You will see that by D200 (diluting one drop to ten drops, repeated 200 times) the spikes are at their strongest.

I christened this chart "Radio Belladonna". What is remarkable is that the transmissions take place long after any of the belladonna substance has vanished.

Need I say more?

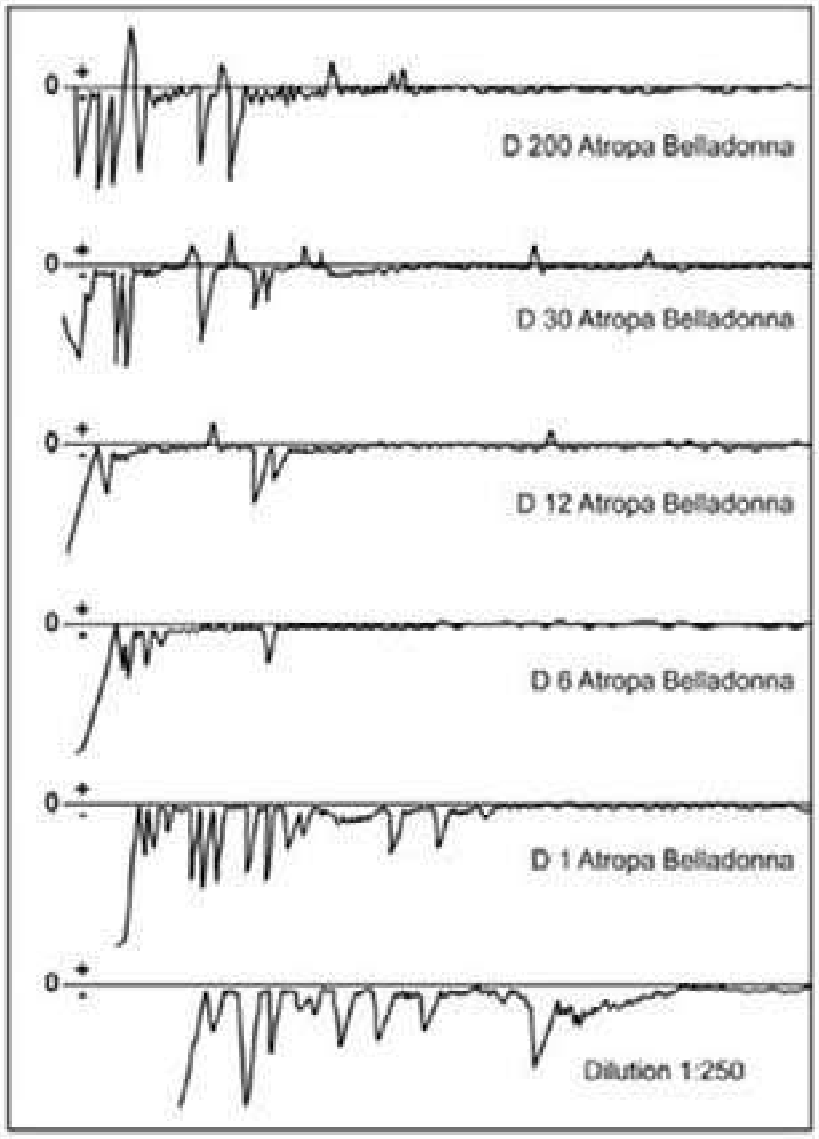

Homoepathy In Acute Infections

Much homeopathy is directed towards lifting chronic or inherited disease traits. It requires what is called constitutional prescribing, meaning taking the person whole physical make up and psychological character into the equation. Treatment is NEVER directed towards symptoms but towards changing the individual's make up or "constitution" and driving out disease tendency.

Prescribing for acute illnesses is rather different. Symptoms may be considered in their own right.

Homeopaths classify symptoms according to a hierarchy of four categories for purposes of acute prescribing:

- Peculiar symptoms. These are symptoms unique to the individual that do not occur in most persons with the acute disease. Homeopaths make note of peculiar symptoms because they often help to determine the remedy.

- Mental and emotional symptoms. These are important general symptoms that inform the homeopath about the patient's total experience of the disorder.

- Other general symptoms. These are physical symptoms felt throughout the patient's body, such as tiredness, changes in appetite, or restlessness.

- Particular symptoms. Particular symptoms are localized in the body; they include such symptoms as nausea, skin rashes, headache, etc.

During homeopathic case-taking, the practitioner will evaluate the intensity of the patient's symptoms, assess their depth within the patient's body, note any peculiar symptoms, evaluate the modalities of each symptom, and make a list of key symptoms to guide the selection of the proper medicine.

Medicines (called "remedies") can be general (against any infection) or specific for that disease.

General Remedies To Look Out For

Aconitum napellus: This remedy is often indicated when fever and inflammation come on suddenly, sometimes after exposure to wind and cold, or after a traumatic experience. The person may be very thirsty and often feels fearful or anxious.

Belladonna: Intense heat, redness, swelling, throbbing, and pulsation indicate a need for this remedy. The person's face may be flushed and hot (though hands and feet may be cold) and the eyes are often sensitive to light. Thirst may be lower than expected during fever. Discomfort is worse from motion or jarring, and relieved by cold applications. This was the classic remedy for scarlet fever.

Ferrum phosphoricum: This remedy is indicated in the early stages of many inflammations. Taken at the very first sign of a cold or sore throat, it often helps a person throw the infection off and not get ill. Fever, pink-flushed cheeks, a general weariness, thirst, and moderate pain and swelling are typical symptoms suggesting Ferrum phos in illness or infection.

Bryonia: Feeling worse from even the slightest motion is a strong indication for this remedy. When ill, the person wants to stay completely still—to be left alone and not interfered with in any way. Fever with chills, a very dry mouth, and thirst are also likely. Tearing pains that feel worse from any motion, but improve from pressure if it adds stability, may accompany local infections.

Calcarea carbonica: People who need this remedy tire easily and have low stamina. They tend to feel chilly and sluggish, with clammy hands and feet (though their feet may heat up in bed at night, and their heads may perspire during sleep). Swollen lymph nodes, frequent colds, sore throats, ear infections, and skin eruptions are common. Children who need this remedy are often slow to walk and may have teething problems, frequent colds, and ear infections.

Hepar sulphuris calcareum: A person who needs this remedy feels extremely sensitive and vulnerable when ill, especially if exposed to cold or drafts. Ear infections, sore throats, sinusitis, bronchitis, and skin eruptions are often seen, and cheesy-smelling discharge or offensive pus may be produced. Areas of inflammation can be very sore and sensitive, and splinter-like pains are often felt (in the tonsils when swallowing, in a boil when the skin is touched, etc.).

Silica: A person who needs this remedy can be sensitive and nervous, with low stamina and poor resistance to infection—leading to swollen lymph nodes, frequent colds, sore throats, tonsillitis, sinusitis, bronchitis, and other illnesses. Boils, easy infection of wounds, and abscessed teeth are often seen. Although very chilly in general, the person often perspires during sleep. Offensive foot sweat with an inclination toward fungal infections is also common.

Mercurius solubilis: This remedy is needed when a person has swollen lymph nodes, offensive breath, and is extremely sensitive to any change in temperature. A tendency toward night sweats and profuse salivation during sleep are other indications. Infections of the gums, ears, sinuses, throat, and skin often respond to this remedy when the other symptoms fit.

Sulphur: This remedy is useful in many kinds of infection characterized by irritation, burning pain, redness of mucous membranes, and offensive odors and discharges. Skin problems are very common—eczema, acne, boils, lymphangitis, and inflammations on or around the genitals. Symptoms are often worse from warmth and worse after bathing. Colds, bronchitis, and other illnesses that have been neglected, or infections that drag on for a very long time, may be helped with this remedy.

Specific Remedies Or Nosodes

Another approach entirely is to use disease pathogen of diseased tissues as a remedy. Remember, it will be diluted many times and so quite harmless. So hoemopathically-potentized malaria can be used as an anti-malaria remedy and that’s what I referred to earlier. I took it in Sri Lanka.

The use of actual disease substances we call a “nosode”. Nosodes exist for TB, cancer, rubella, ‘flu, Streptococcus and Staphylococcus, to name just a few.

A nosode can act like a vaccine and be used as a preventative. In fact Samuel Hahnemann should be credited with the development of vaccination, not Edward Jenner (but try telling that to orthodox doctors!)

Nosodes are really only available through a homeopathic prescriber. In the event of an epidemic or pandemic, I expect supplies will be limited.

But if you are OK ordering online, you can get ANYTHING homeopathic from India, where homeopathy is huge. That’s thanks to the former ruler of Punjab, Maharaja Ranjit SinghMaharaja, who espoused it in 1830. It worked for him and he spent some of his fabulous fortune disseminating it throughout India (which included Pakistan at that time).

Every major city in India has pharmacies dispensing homeopathy medicine. Many states in India also have homeopathy colleges teaching this type of alternative medicine to prospective students. Currently there are 186 homeopathic medical colleges in India.

The Indian government created the Central Council Of Homeopathy which requires uniform education at the graduate level and diploma level in homeopathy. This has lead to standardized practice in India preventing unqualified homeopaths to prescribe medicine and treatments. Currently, there are over 300,000 homeopathic doctors practicing medicine in India.

Just to finish, if you have not heard the term “allopathic”, it means conventional medicine. It comes from the Latin alli- meaning “other” (same word as AL-ternative). So it’s the direct opposite of treating like with like. Allopathic medicine fights against a condition, trying to suppress it, rather than working along with it, to bring it to a close.

#42 Homotoxicology

Also know as comlex homeopathy and German homeopathy.

A similar and closely related discipline is that of homotoxicology, sometimes called “complex homeopathy”. It’s also sometimes called “German homeopathy”, since around 40% of German MDs use it regularly. But I think that’s confusing, because Hahnemann was German so both methods are really “German homeopathy”.

Complex doesn’t really mean convoluted and difficult; it seems means by the use of mixtures and formuals, instead of single pure substances.

In fact complex homeopathy, as it is practiced today, was largely developed by another German doctor Hans-Heinrich Reckweg. He sold out his company in Baden-Baden and re-started in Albuqueque, New Mexico, so you have it in the US too. It’s called HEEL (short for herba est ex luce, or plants are from light).

Complex homeopathy, as its name suggests, uses compound solutions; mixes of different remedies. This is anathema to classic homeopaths, who insist that Hahnemann described a science based on one remedy, one disease, at one time (they just ignore the fact that even Hahnemann later experimented with mixes.

Different formulations

Because of the very nature of homeopathy, mixtures can take two forms: mixtures of different substances or mixtures of the same substance only but at different potencies or dilutions. The latter we call a chord (like in music, several notes sounding at once).

But there are other, more subtle differences too. In complex homeopathy we tend to use lower (not so extreme) dilutions. Thus D4, D6 and D10 are often encountered; D30, D60 and D200 are also encountered a lot. Whereas a classic homeopath works more in dilutions (potencies) of C12 and upwards to 1M and 2M. M means a 1:100 dilution, repeated 1,000 times.

In classical homeopathy, the teaching is that as soon as you get a result, you stop the medication and leave the recovery process to run un-aided. A complex homeopathy prescriber would tend to smile with satisfaction and write the patient up for another week or two! Then, as well as the scandal of using more than one substance at a time in a mixture, a complex homeopathy prescriber might also use more than one mixture at a time!

This is typical where we would add what's called a "drainage remedy". This is to increase removal of toxins which accrue in an illness, especially when being treated effectively. As the pathogens are destroyed, the body may reach an overload state, due to its inability to remove resulting toxins fast enough. Drainage remedies stimulate that process and so make sense.

Here an example of a complex homeopathic remedy (Lymphomyosot) and what it contains and simplified notes on what each is for:

> Geranium robertianum 4X, diarrhea, painful urination, ulcerations Nasturtium aquaticum (watercress) 4X, liver and gall bladder, mouth ulcers Ferrum iodatum 12X, scrofula (TB), glandular swelling Myosotis arvensis 3X, chronic bronchitis
>
> Scrophularia nodosa (knotted figwort) 3X, inflammation and swelling of lymph glands and breast, eczema Teucrium scorodonia (germander) 3X, chronic rhinitis, bronchitis, tuberculosis
>
> Veronica officinalis (speedwell) 3X, chronic bronchitis, cystitis, dermatosis
>
> Equisetum hyemale (horsetail) 4X, cystitis, kidney stones
>
> Fumaria officinalis (fumitory) 4X, blood purifier, excess toxin levels
>
> Natrium sulphuricum 4X, liver problems, gastroenteritis, asthma Pinus sylvestris (Scots pine) 4X, rickets, bronchitis, scrofula
>
> Gentiana lutea (yellow gentian) 5X, chronic gastritis, flatulence and diarrhea.
>
> Aranea diadema (spider orchid) 6X, exudations, worse in wet and cold
>
> Sarsaparilla 6X, herpes, painful urination, cystitis and kidney inflammation
>
> Calcarea phosphorica 12X, gastroenteritis, rickets

Note some similarities with herbal preparations.

What To Take

Complex homoepathy prescribing is a complex art in itself. I cannot do more than give an outline here. However it's definitely the one for home prescribing. There are four main reason why it is good for the layman to use:

1. It's safe, you can't mess up.

2. Reactions are milder, it's hard to overdo it!

3. The very nature of a mixture means something should hit the target!

4. The names often help with knowing what to choose (no such help in classic homeopathy).

Here are some suggestions what to do for a severe infection.

Engystol N (HEEL). Great for boosting defence mechanisms and increasing resistance to infections of all kinds, especially viral infections. Good for pneumonia, bronchitis, rhinitis and hepatitis. It's also good after vaccinations, to prevent adverse after effects.

Echinacea compositum

I have already remarked I wouldn't put much faith in just Echincaea for my family. But this Echinacea compositum (complex) is brilliant!

Euphorbium Compositum Spray

Great for nasal, sinus and pharygeal infections.

Galium-Heel is a non-specific drainage remedy, perhaps more suitable for chronic conditions.

Gripp-Heel. This is the one you want in a swine flu or bird flu pandemic. Grippe is the French word for influenza. It contains Aconite (shock and chills, which is the onset of most infectious diseases); Bryonia (inflammation of mucus membranes and cough); Lachesis (bushmaster snake), worsening condition and congestive bronchitis; Eupatorium, influenza with excruciating pains; and Phosphorus, bronchopneumonia and laryngitis, with hoarseness.

Lymphomyosot. Chronic and repeated tonsillitis and lymph gland swellings.

Tonsilla compositum. This is one of my favorites for stimulating the immune system. It has no less than 29 logical ingredients and increases body defences in all ways, but especially the lymphatic system. It goes to chills, bronchitis, emphysema, abcesses, suppurating wounds, gingivitis, as well as the obvious sore throats and tonsillitis.

There are many other compounds which can be called into action by the experienced and knowledgeable prescriber. It takes years to learn the full repertoire of this kind of prescribing.

But here is information enough to get you started on searching and, hopefully, to give you confidence that here is a wonderful, safe and gentle approach to healing. No matter how it is mocked and despised by conventional colleagues who only prescribe poisonous drugs as a “cure”, trust me, complex homeopathy works!

Where To Buy

HEEL have distributors and outlets in many parts for the world, from Colombia to Bielorus! For the UK and several other centers, contact my good friend Roger Wilson via his website, http://www.biopathica.co.uk (or biopathica.com).

HEEL’s headquarters are in Germany: Dr.-Reckeweg-Str. 2-4 D-76532 Baden-Baden or:

P.O. Box 100349 D-76484 Baden-Baden Phone: +49 72 21 50 10 0 E-Mail: info@heel.de Web: http://www.heel.de

HEEL USA (www.heelusa.com) is related by the Reckeweg range of products but is a separate corporate entity to the original HEEL.

There are several other very good complex homeopathic pharmacy suppliers:

Deseret Biologicals in Sandy, Utah.

Deseret Biologicals are now working with Bruce Shelton MD, former US advisor to HEEL. 469 West Parkland Drive Sandy, UT 84070-6403

phone: 801-563-7448 fax: 801-563-7455 orders: 800-827-9529

Pascoe. Not so much for otc remedies or direct mail but you will recognize these products from certain prescribers. [http://www.pascoe-global.com/sites/content/e7639/e16201/index_ eng.html]

#43 Biochemic Tissue Salts

Also known as Schüssler Salts

Biochemic tissue salts were developed as homeopathic remedies by Samuel Hahnemann, and further investigated by Dr Wilhelm Heinrich Schuessler in the late nineteenth century. They are remedies based on the inorganic salts and are intended to increase the body's ability to absorb and use the mineral salts found in a person's everyday diet in order to maintain or restore good health.

According to Schuessler, illnesses arise when there is an imbalance or deficiency of these minerals in the body - when they are not in the right place at the right time, in other words. Schuessler salts, which he called 'means of functioning,' are meant to redress the imbalance. Schuessler himself worked with 12 salts. His successors added 12 more.

Modern Tablet Form

Biochemic tissue salts are available in tablet form similar to that of most homeopathic remedies. They are made up of the salt ground with lactose powder (a process called trituration) in a ratio of 1:9, and then further diluted six times at a 1:9 ratio so the final product is a 1,000,000 to 1 dilution. These tablets then dissolve on the tongue. They are both individually available as one kind of salt, or come in combinations to address particular illnesses or ailments. There are no known side-effects, and taking them in excess will not result in overdose as biochemic tissue salts occur naturally in the body. Waste will be excreted.

We don't need all the Schuessler salts here but two are useful for infections:

Ferrum Phosphate (ferrum phos) D12 or 12X: great for acute infections, the quicker you nip them in the bud, the better.

Kali Muriaticum/ Potassium Chloride (kali mur) D6 or 6X: the second good infection remedy, can be taken alternately with Ferr Phos.

#44 The Russian SCENAR

SCENAR
Professional

I strongly recommend every home have a SCENAR in the medicine cabinet. I have seen this amazing device blow away serious infections in a day or less.

Actually, the SCENAR is a whole family of machines developed in Russia and I predict they will completely change the face of medicine in the next 20 years. They are fast, portable, cheap and effective against almost any condition, from treating sports injuries, strokes, angina, acute infections, back pains and irritable bowel disease (as well as pre-menstrual tension and post-surgical complications) and even defibrillating hearts!

The promise of a small hand-held device that is capable of curing most illness, such as was portrayed in the cult 1970s TV series "STAR TREK", has become a reality. The device weighs around 300 gm, resembles a TV remote control and is powered by an ordinary 9v. battery. It is placed on the skin of the chest, head, abdomen or any diseased part, where it collects electromagnetic signals. These are then modulated according to the on-board software programme and played back to the tissues.

Essentially, SCENAR is using the patient's own endogenous signals on a cybernetic feedback basis, scanning and re-transmitting many times a second. As described to me, the device 'evolves' a new signal pattern for the disordered tissues, the machine literally entering into an information dialogue with the body. New frequencies and energy patterns are established, which in turn become fresh input signals, to be further modified, and so on.

This output-equals-new- input is much the way that fractals are generated and thus, biologically-speaking, we seem to be on good ground here. On the premise that disease signals are generally fixed and unnatural, anything which breaks up the existing order has the capability of disease-busting.

SCENAR casebook from Dr Keith MD (quoted in my book "Virtual Medicine").

Male, 69 years old. This man who was a gardener by profession had chronic suppurating osteomyelitis of the foot (deep bone infection) which could not be controlled. He was scheduled for an amputation of the lower leg in four days time, largely due to intractable pain. Someone decided to try a SCENAR device on him. It was run over the affected limb for about 30 minutes. Next day, the pain had vanished for the first time in 8 months. Later that day another 30-minute treatment was given.

By next morning the recovery was so dramatic the amputation was called off. A third treatment was given and seven days after the first SCENAR this man was back at work, digging in the garden. His leg has completely recovered.

Clinical Aspects

The device is very safe; the impulse times are very short. No pain is felt but the patient is usually aware of a tingling sensation while it works. The practitioner seeks for what the Russians term asymmetry, meaning something different about the tissue characteristics in the vicinity. There are five main criteria: discolouration (reddening or pallor) sensation (numbness or hyper-aesthesia) stickiness' in which the machine drags with a magnet-like quality as it is drawn over certain tracts of the skin sound changes (the machine begins to chatter electronically as it hits certain zones) change in numerical output display.

Even though it may not coincide with the obvious area of symptoms or pathology, the important point is to treat the asymmetry. For reasons we do not fully understand, when this is eliminated, recovery will rapidly follow.

Couldn't resist reporting just a simple application... My lovely wife Vivien was working in Los Angeles (she's a fashion designer). She had a respiratory infection and felt rough enough for antibiotics. However she was still suffering when she came home Friday night. I broke out the SCENAR and ran it over her chest and back. Within 5 minutes there was what the Russians called an "asymmetry": on the left side there was pain under the tip of the scapula and over the pecs at the front. We worked this for about 10 minutes.

That's all! She reported feeling tons better. All night she didn't cough and woke feeling much better. Who needs antibiotics or even herbs and homeopathy when you have a SCENAR!

You can find out more about this device in my book "Virtual Medicine". Get it here:

http://www.alternative-doctor.com/scenar/

#45 UV Light As An Antibiotic

Back when (in the days before antibiotics) there were many treatments which worked for infections. One almost lost technique was the use of Ultra Violet Light irradiation of the blood to kill pathogens and even cancer cells.

This is a version of what we technically call plasmapheresis or, in Europe "auto sanguis" therapy (meaning "own blood" therapy).

You can do a number of things to blood while it is outside the body that you could not do while it is inside… For example, you can expose it to ozone and so enrich it with oxygen. A high local concentration of oxygen is known to provide a hostile environment for cancer cells. Some deadly organisms, too, like the gas gangrene bacteria (Clostridium welchii), are killed by a high-oxygen environment. Even viruses succumb.

But here I am describing the use of ultra-violet light to treat infections. UV is accepted as a sterilizing agent; city water supply is treated with it; indeed, you may have a UV unit fitted to your spa or pool. There is no scientific argument over its cleansing capabilities. It zaps pathogens!

It's toxic to humans too (sunburn!) but that problem is solved by focusing the UV on blood which is OUTSIDE the body. We call this ultra-violet blood irradiation or UVBI for short. It's also sometimes known as photoluminescence.

The treated blood is returned to the body, where it spreads its good message throughout the tissues. Surprisingly, only about 5% of the total blood volume needs to be treated in this way. This small proportion spreads through the entire body and works its magic, in much the same way a therapeutic drug gets everywhere.

UVBI kills viruses, parasites and bacteria and was popular in the 1930s for polio and other viral infections. The reason is simple enough: it works! But then it became unpopular, as it was gradually sidelined by mass vaccination programs, which were believed to work, despite evidence to the contrary. Finally, UV blood therapy fell into disrepute, except among dedicated holistic physicians.

That's a great pity because it is very safe (no side effects ever recorded, when used properly), highly effective and cheap to do. It's so easy, as a matter of fact, that it’s something a lay persons, with the right equipment and clear instructions, can administer it to themselves.

History

Research into the use of UV light for therapy goes back as far as the 1870’s. Niels Ryberg Finsen won the Nobel Peace Prize for “Physiology of Medicine” in 1903 for his UV treatments of 300 people suffering from Lupus in Denmark.

One of the early pioneers of UVBI was Kurt Naswitis, who directly irradiated the blood with UV light through a shunt in 1922.

Another pioneer of this treatment was Emmett K. Knott, who started up in 1928. Dr. Virgil K.

Hancock and Mr. Knott published the first article on the therapeutic efficacy of phototherapy in June of 1934. By June of 1942 they had treated 6,520 patients with ultraviolet therapy. Their work was reported in the most prestigious medical journals. One respected physician of the time, Dr. Rebbeck, wrote in the early 1940s that Hancock and Knott “had in the irradiation of blood with ultraviolet spectral energy, a therapy of more pronounced value than any other method known to date.” During a 50-year period, doctors performed more than 300,000 clinical tests that repeatedly showed photoluminescence works and not a single patient was harmed.

Factually, there are very few contraindications for UVBI and unsuitable patients are easily screened out: porphyria, hereditary metabolic disorders such as phenylketonuria, exeroderma pigmentosum, acute photodermatitis, and a hypersensitivity to sunlight or other forms of ultraviolet light are not suitable for UVBI.

On September 11, 1928 Emmet K. Knott patented his device as a “Means for Treating Bloodstream Infections” and later he received another patent, in 1943, which was slightly different.

This is very fortunate for us today because the machine received an FDA “grandfather” status as a device that was sold and distributed in interstate commerce prior to 1976.

Procedure

If you are lucky enough to locate a practitioner who will do this for you, you will find it's very quick and virtually painless. It takes about 20 minutes in total.

A catheter is placed into one of the veins around the elbow (usually with a butterfly needle -- a small plastic catheter attached to a short needle). Drawn blood travels through a small glass chamber, called a cuvette, where it is exposed twice to ultraviolet light before getting re-introduced to the patient's blood stream.

Many patients feel significant improvements after just a few treatments. In severe cases, patients may require 10 or more treatments. But usually, 3 to 5 treatments are sufficient. Even chronic, intractable problems will likely recover.

What Conditions Can Be Benefited?

Actually, a knowledgeable practitioner would probably say almost any non-optimum condition might respond, even cancer. Bear in mind my rubric that "Any good health measure is an anti-cancer measure."

Other conditions which may be significantly improved include:

- Viral Infections
- Candidiasis (fungal infections)
- Bacterial Infections
- Chronic Fatigue
- Poor Oxygen
- Blood Poisoning
- Chronic infections (skin, etc.)
- Colitis
- Leg ulcers
- Diabetes Complications
- Poor Immune Function

UVBI or photoluminescence seems poised to make a come back. It is free of (legal) molestation by the FDA, though the FDA has scant regard for the niceties of law!

But with the obvious dangers and lack of efficacy of vaccination; the development of dangerous antibiotic resistant species of bacteria; and with the difficulty eradicating many viruses, such as hepatitis and AIDS, it is good to know we have a PROVEN, safe, fast, cheap and effective alternative therapy. It won’t suit the mindless pill-pushers, who seem to have lost the true art of the healer.

#46 Cold Lasers

The local application of low-level laser (cold laser) energy has been found to be effective in a variety of inflammatory conditions, including infections. Lasers produce coherent light, which is critical. Color light therapy is a different topic altogether. LED light sources are non-coherent.

The 630- 640 nanometre wavelength (red) has considerable effect on the tissues, stimulating cell metabolism, decreasing inflammation and increasing endorphin release, which reduces pain.

One other important consideration is the optimum power necessary for bio-stimulation. The Arndt-Schulz Law of photo-biological activity, essentially states that “less is more” when it comes to energy for improved cellular physiology. A good cold laser, such as the Erchonia cold laser provides the best wavelength (635 nm) for cellular physiology at very low energy (2-5 mill Watts) to stimulate cells to function better. If the stimulation is too intense, there may actually be an inhibitory effect, or possibly degeneration or destruction of cells.

The advantage of lasers, like the SCENAR is that they can be focused on the area of concern. There is no local thermal effect and no recovery time. Normal healing requires certain mediators to be released at the wound site. These cell mediators call inflammatory cells to the wound which clean up damaged tissues, fight bacteria, and stimulate specialized cells to grow to try to heal the zone of injury. The physiologic concept is that cold laser improves cellular metabolism and accelerates the process of healing to a more rapid and efficient state.

Now it happens that 405 nm (violet) laser light works differently. It is soothing and calming.

It helps the body fight infections at a deeper level and even seems to bring in an emotional response to healing.

Stress, overwork, improper diet and the environment can all contribute to a toxic overload in the body which reduces the body’s ability to maintain a healthy immune system. The immune system is the body’s defense against viruses, bacteria, molds, fungus and other opportunistic infections which would like to set up a “home” in your body. The laser can help detoxification protocols work faster and on a deeper cellular level.

There are over 1500 published studies and not one of them mentions any negative side effects of semi-conductor diode lasers at the 5mW range, such as the Erchonia lasers. Cold lasers are safe, non-toxic and non-invasive - there have not been any recorded side effects in over 1500 publications. There are some necessary common sense precautions that need to be considered, such as avoiding pointing the laser beam directly into the eye and maintaining it there, which could prove to be damaging to the eye.

Note: If you hear the term “scalar wave laser” run a mile. There is no such thing in the universe and it’s science baloney. Find a real science-grounded practitioner, not someone who speaks pseudo-quantum psychobabble.

#47 Nano UV Zapper

This is another device which should be in every home. Manufacturers claim it is the "world's first compact disinfection light" and say the technology could be used for germ-busting in homes, schools and hospitals.

The Nano-UV Portable Disinfection Light looks like a mobile phone but flips open to reveal an ultraviolet lamp.

Those behind the product say it is certified to kill 99.99 per cent of the potentially deadly H5N1 strain of avian flu in five seconds, and harmful bacteria, such as MRSA, E.coli and salmonella, in 10 seconds.

By using "multi-wavelength" UV technology, combining UVA, UVB and UVC rays, manufacturers say the light is more effective than standard UV disinfection products at killing the DNA of viruses, bacteria and fungi.

The light has a sterilising effect when passed over objects like baby's dummies, nappy changing tables, cutlery, toilet seats and towels.

The Nano UV zapper has a built-in safety device to ensure each use only lasts 10 seconds, which is all it takes to kill 99% plus of all germs.

The beauty of this product is that it's small enough to carry everywhere and powerful enough to give you peace of mind in any situation.

The Nano zapper can be used to sterilize any surface except the skin.

Microbiologist Professor Hugh Pennington, the emeritus professor of bacteriology at Aberdeen University, said UV technology is well known for its germ-killing properties and has a place alongside other good hygiene practices, such as hand-washing.

#48 What Goes In The Medicine Cupboard?

Let’s finish up by considering what should every home have in the medicine cupboard. Having got this far, you may have your own opinions about what would be good to include, to be ready in the event of an infection, whether mild or severe.

These are the items I suggest:

- 1 oz. or 25 grams of icthammol ointment (20%)
- Drawing salve for acute, small inflammed infections
- 4 oz. spray of colloidal silver; 100 ml. bottle Grazes, skin abrasions, dental hygiene, sore throat, puncture wounds
- A bottle of MMS (chlorine dixiode) Severe systemic infections, viral and bacterial, bronchitis, wounds, abcesses
- Engystol (if you can’t get it, use propolis) General defence for any infectious disease, especially viral.
- 30 ml. bottle of Samento (TOA-free) Chronic infections and “stealth pathogens” (eg. Lyme’s)
- 30 ml. Oregano Oil Acute infections of all kinds
- 30 ml. Black seed oil Acute infections of all kinds
- 30 ml. Lavender Oil Mainly skin conditions but could be used for abscesses, acne, athlete’s foot, fungal infections, boils, bruises, cuts, insect bites and stings, lice, sunburns, wounds, bronchitis, coughs, colds, congestion, flu, laryngitis, throat infections, whooping cough, and sinus infections.
- Also You Need One Of Two Electrical Healing Devices.

  A SCENAR

  Learn to use it first; don’t wait till somebody gets sick!

A low-level laser

Also learn to use it first!

- High Dose Vit C and Vit A for acute viral infections. Have vitamin C powder on hand and vitamin A capsules. There are 25,000 IU capsules coming in from China. Buy them only if you trust the source (I don’t).

- Vitamin D capsules, 5,000 IU, improve your immunity whenever an epidemic threatens.

Appendix

The official FDA version

I thought you might like to supplement what I am telling you with the "official" story. This section is from the FDA Consumer magazine (September 1995). I think it's worth reading it over again in a different way. It re-inforces what I have been saying.

You can also listen to some useful recordings and download PDF files here too:

When penicillin became widely available during the Second World War, it was a medical miracle, rapidly vanquishing the biggest wartime killer--infected wounds.

Discovered initially by a French medical student, Ernest Duchesne, in 1896, and then rediscovered by Scottish physician Alexander Fleming in 1928, the product of the soil mold Penicillium crippled many types of disease-causing bacteria. But just four years after drug companies began mass-producing penicillin in 1943, microbes began appearing that could resist it.

The first bug to battle penicillin was Staphylococcus aureus. This bacterium is often a harmless passenger in the human body, but it can cause illness, such as pneumonia or toxic shock syndrome, when it overgrows or produces a toxin.

In 1967, another type of penicillin-resistant pneumonia, caused by Streptococcus pneumoniae and called pneumococcus, surfaced in a remote village in Papua New Guinea. At about the same time, American military personnel in Southeast Asia were acquiring penicillin-resistant gonorrhea from prostitutes.

By 1976, when the soldiers had come home, they brought the new strain of gonorrhea with them, and physicians had to find new drugs to treat it. In 1983, a hospital-acquired intestinal infection caused by the bacterium Enterococcus faecium joined the list of bugs that outwit penicillin.

Antibiotic resistance spreads fast. Between 1979 and 1987, for example, only 0.02 percent of pneumococcus strains infecting a large number of patients surveyed by the national Centers for Disease Control and Prevention were penicillin-resistant. CDC's survey included 13 hospitals in 12 states. Today, 6.6 percent of pneumococcus strains are resistant, according to a re-port in the June 15, 1994, Journal of the American Medical Association by Robert F.

Breiman, M.D., and colleagues at CDC. The agency also reports that in 1992, 13,300 hospital patients died of bacterial infections that were resistant to antibiotic treatment.

Why has this happened?

"There was complacency in the 1980s. The perception was that we had licked the bacterial infection problem. Drug companies weren't working on new agents. They were concentrating on other areas, such as viral infections," says Michael Blum, M.D., medical officer in the Food and Drug Administration's division of anti-infective drug products. "In the meantime, resistance increased to a number of commonly used antibiotics, possibly related to overuse of antibiotics. In the 1990s, we've come to a point for certain infections that we don't have [antibiotics] agents available."

According to a report in the April 28, 1994, New England Journal of Medicine, researchers have identified bacteria in patient samples that resist all currently available antibiotic drugs.

Survival of the Fittest

The increased prevalence of antibiotic resistance is an outcome of evolution. Any population of organisms, bacteria included, naturally includes variants with unusual traits--in this case, the ability to withstand an antibiotic's attack on a microbe. When a person takes an antibiotic, the drug kills the defenseless bacteria, leaving behind--or "selecting," in biological terms--those that can resist it. These renegade bacteria then multi-ply, increasing their numbers a millionfold in a day, becoming the predominant microorganism.

The antibiotic does not technically cause the resistance, but allows it to happen by creating a situation where an already existing variant can flourish. "Whenever antibiotics are used, there is selective pressure for resistance to occur. It builds upon itself. More and more organisms develop resistance to more and more drugs," says Joe Cranston, Ph.D., director of the department of drug policy and standards at the American Medical Association in Chicago.

A patient can develop a drug-resistant infection either by contracting a resistant bug to begin with, or by having a resistant microbe emerge in the body once antibiotic treatment begins. Drug-resistant infections increase risk of death, and are often associated with prolonged hospital stays, and sometimes complications. These might necessitate removing part of a ravaged lung, or replacing a damaged heart valve.

Bacterial Weaponry

Disease-causing microbes thwart antibiotics by interfering with their mechanism of action. For example, penicillin kills bacteria by attaching to their cell walls, then destroying a key part of the wall. The wall falls apart, and the bacterium dies. Resistant microbes, however, either alter their cell walls so penicillin can't bind or produce enzymes that dismantle the antibiotic.

In another scenario, erythromycin attacks ribosomes, structures within a cell that enable it to make proteins. Resistant bacteria have slightly altered ribosomes to which the drug cannot bind. The ribosomal route is also how bacteria become resistant to the antibiotics tetracycline, streptomycin and gentamicin.

How Antibiotic Resistance Happens

Antibiotic resistance results from gene action. Bacteria acquire genes conferring resistance in any of three ways.

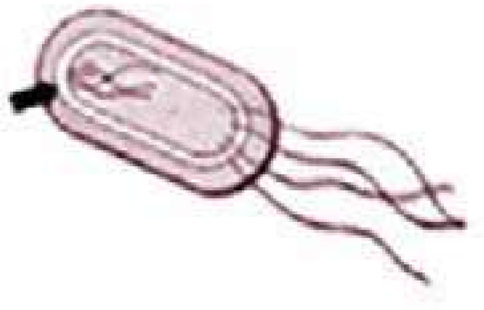

In spontaneous DNA mutation, bacterial DNA (genetic material) may mutate (change) spontaneously (indicated by starburst). Drug-resistant tuberculosis arises this way.

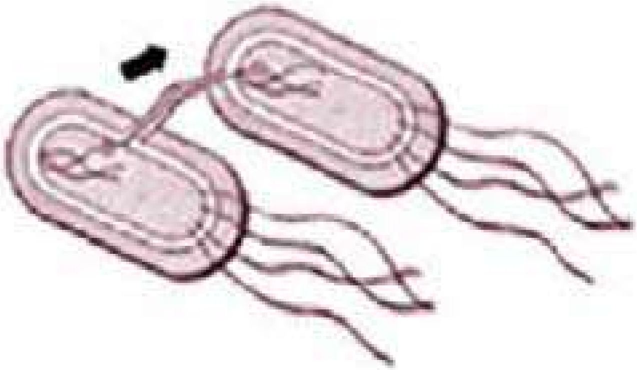

In a form of microbial sex called transformation, one bacterium may take up DNA from another bacterium. Pencillin-resistant gonorrhea results from transformation.

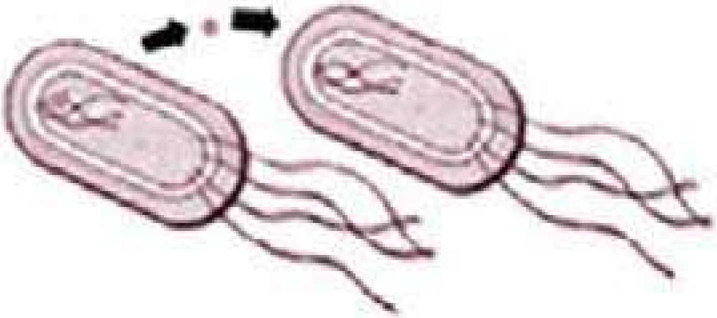

Most frightening, however, is resistance acquired from a small circle of DNA called a plasmid, that can flit from one type of bacterium to another. A single plasmid can provide a slew of different resistances. In 1968, 12,500 people in Guatemala died in an epidemic of Shigella diarrhea. The microbe harbored a plasmid carrying resistances to four antibiotics!

A Vicious Cycle: More Infections and Antibiotic Overuse

Though bacterial antibiotic resistance is a natural phenomenon, societal factors also contribute to the problem. These factors include increased infection transmission, coupled with inappropriate antibiotic use.

More people are contracting infections. Sinusitis among adults is on the rise, as are ear infections in children. A report by CDC's Linda F. McCaig and James M. Hughes, M.D., in the Jan. 18, 1995, Journal of the American Medical Association, tracks antibiotic use in treating common illnesses. The report cites nearly 6 million antibiotic prescriptions for sinusitis in 1985, and nearly 13 million in 1992. Similarly, for middle ear infections, the numbers are 15 million prescriptions in 1985, and 23.6 million in 1992.

Causes for the increase in reported infections are diverse. Some studies correlate the doubling in doctor's office visits for ear infections for preschoolers between 1975 and 1990 to increased use of day-care facilities. Homelessness contributes to the spread of infection. Ironically, advances in modern medicine have made more people predisposed to infection. People on chemotherapy and transplant recipients taking drugs to suppress their immune function are at greater risk of infection.

"There are the number of immunocompromised patients, who wouldn't have survived in earlier times," says Cranston. "Radical procedures produce patients who are in difficult shape in the hospital, and are prone to nosocomial [hospital-acquired] infections. Also, the general aging of patients who live longer, get sicker, and die slower contributes to the problem," he adds.

Though some people clearly need to be treated with antibiotics, many experts are concerned about the inappropriate use of these powerful drugs. "Many consumers have an expectation that when they're ill, antibiotics are the answer. They put pressure on the physician to prescribe them. Most of the time the illness is viral, and antibiotics are not the answer. This large burden of antibiotics is certainly selecting resistant bacteria," says Blum.

Another much-publicized concern is use of antibiotics in livestock, where the drugs are used in well animals to prevent disease, and the animals are later slaughtered for food. "If an animal gets a bacterial infection, growth is slowed and it doesn't put on weight as fast," says Joe Madden, Ph.D., strategic manager of microbiology at FDA's Center for Food Safety and Applied Nutrition. In addition, antibiotics are sometimes administered at low levels in feed for long durations to increase the rate of weight gain and improve the efficiency of converting animal feed to units of animal production.

FDA's Center for Veterinary Medicine limits the amount of antibiotic residue in poultry and other meats, and the U.S. Department of Agriculture monitors meats for drug residues. According to Margaret Miller, Ph.D., deputy division director at the Center for Veterinary Medicine, the residue limits for antimicrobial animal drugs are set low enough to ensure that the residues themselves do not select resistant bacteria in (human) gut flora.

FDA is investigating whether bacteria resistant to quinolone antibiotics can emerge in food animals and cause disease in humans. Although thorough cooking sharply reduces the likelihood of antibiotic-resistant bacteria surviving in a meat meal to infect a human, it could happen.

Pathogens resistant to drugs other than fluoroquinolones have sporadically been reported to survive in a meat meal to infect a human. In 1983, for example, 18 people in four midwestern states developed multi-drug-resistant Salmonella food poisoning after eating beef from cows fed antibiotics. Eleven of the people were hospitalized, and one died. A study conducted by Alain Cometta, M.D., and his colleagues at the Centre Hospitalier Universitaire Vaudois in Lausanne, Switzerland, and reported in the April 28, 1994, New England Journal of Medicine, showed that increase in antibiotic resistance parallels increase in antibiotic use in humans. They examined a large group of cancer patients given antibiotics called fluoroquinolones to prevent infection. The patients' white blood cell counts were very low as a result of their cancer treatment, leaving them open to infection.

Between 1983 and 1993, the percentage of such patients receiving antibiotics rose from 1.4 to 45. During those years, the researchers isolated Escherichia coli bacteria annually from the patients, and tested the microbes for resistance to five types of fluoroquinolones. Between 1983 and 1990, all 92 E. coli strains tested were easily killed by the antibiotics. But from 1991 to 1993, 11 of 40 tested strains (28 percent) were resistant to all five drugs.

Towards Solving the Problem

Antibiotic resistance is inevitable, say scientists, but there are measures we can take to slow it. Efforts are under way on several fronts--improving infection control, developing new antibiotics, and using drugs more appropriately.

Barbara E. Murray, M.D., of the University of Texas Medical School at Houston writes in the April 28, 1994, New England Journal of Medicine that simple improvements in public health measures can go a long way towards preventing infection. Such approaches include more frequent hand washing by health-care workers, quick identification and isolation of patients with drug-resistant infections, and improving sewage systems and water purity in developing nations.

Drug manufacturers are once again becoming interested in developing new antibiotics. These efforts have been spurred both by the appearance of new bacterial illnesses, such as Lyme disease and Legionnaire's disease, and resurgences of old foes, such as tuberculosis, due to drug resistance.

FDA is doing all it can to speed development and availability of new antibiotic drugs. "We can't identify new agents--that's the job of the pharmaceutical industry. But once they have identified a promising new drug for resistant infections, what we can do is to meet with the company very early and help design the development plan and clinical trials," says Blum. In addition, drugs in development can be used for patients with multi-drug-resistant infections on an "emergency IND (compassionate use)" basis, if the physician requests this of FDA, Blum adds. This is done for people with AIDS or cancer, for example.

No one really has a good idea of the extent of antibiotic resistance, because it hasn't been monitored in a coordinated fashion. "Each hospital monitors its own resistance, but there is no good national system to test for antibiotic resistance," says Blum.

This may soon change. CDC is encouraging local health officials to track resistance data, and the World Health Organization has initiated a global computer database for physicians to report outbreaks of drug-resistant bacterial infections.

Experts agree that antibiotics should be restricted to patients who can truly benefit from them--that is, people with bacterial infections. Already this is being done in the hospital setting, where the routine use of antibiotics to prevent infection in certain surgical patients is being reexamined.

"We have known since way back in the antibiotic era that these drugs have been used inappropriately in surgical prophylaxis [preventing infections in surgical patients]. But there is more success [in limiting antibiotic use] in hospital settings, where guidelines are established, than in the more typical outpatient settings," says Cranston.

Murray points out an example of antibiotic prophylaxis in the outpatient setting--children with recurrent ear infections given extended antibiotic prescriptions to prevent future infections. (See "Protecting Little Pitchers' Ears" in the December 1994 FDA Consumer.)

Another problem with antibiotic use is that patients often stop taking the drug too soon, because symptoms improve. However, this merely encourages resistant microbes to proliferate. The infection returns a few weeks later, and this time a different drug must be used to treat it.

Targeting TB

Stephen Weis and colleagues at the University of North Texas Health Science Center in Fort Worth reported in the April 28, 1994, New England Journal of Medicine on research they conducted in Tarrant County, Texas, that vividly illustrates how helping patients to take the full course of their medication can actually lower resistance rates. The subject--tuberculosis.

TB is an infection that has experienced spectacular ups and downs. Drugs were developed to treat it, complacency set in that it was beaten, and the disease resurged because patients stopped their medication too soon and infected others. Today, one in seven new TB cases is resistant to the two drugs most commonly used to treat it (isoniazid and rifampin), and 5 percent of these patients die.

In the Texas study, 407 patients from 1980 to 1986 were allowed to take their medication on their own. From 1986 until the end of 1992, 581 patients were closely followed, with nurses observing them take their pills. By the end of the study, the relapse rate--which reflects antibiotic resistance--fell from 20.9 to 5.5 percent. This trend is especially significant, the researchers note, because it occurred as risk factors for spreading TB--including AIDS, intravenous drug use, and homelessness--were increasing.

The conclusion: Resistance can be slowed if patients take medications correctly.

Narrowing the Spectrum

Appropriate prescribing also means that physicians use “narrow spectrum” antibiotics--those that target only a few bacterial types--whenever possible, so that resistances can be restricted. The only national survey of antibiotic prescribing practices of office physicians, conducted by the National Center for Health Statistics, finds that the number of prescriptions has not risen appreciably from 1980 to 1992, but there has been a shift to using costlier, broader spectrum agents. This prescribing trend heightens the resistance problem, write McCaig and Hughes, because more diverse bacteria are being exposed to antibiotics.

One way FDA can help physicians choose narrower spectrum antibiotics is to ensure that labeling keeps up with evolving bacterial resistances. Blum hopes that the surveillance information on emerging antibiotic resistances from

CDC will enable FDA to require that product labels be updated with the most current surveillance information.

Many of us have come to take antibiotics for granted. A child develops strep throat or an ear infection, and soon a bottle of “pink medicine” makes everything better. An adult suffers a sinus headache, and antibiotic pills quickly control it.

But infections can and do still kill. Because of a complex combination of factors, serious infections may be on the rise. While awaiting the next “wonder drug,” we must appreciate, and use correctly, the ones that we already have.

Big Difference

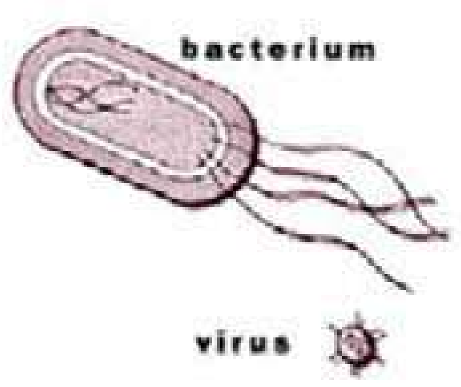

If this bacterium could be shown four times bigger, it would be the right relative size to the virus beneath it. (Both are microscopic and are shown many times larger than life.)

Although bacteria are single-celled organisms, viruses are far simpler, consisting of one type of biochemical (a nucleic acid, such as DNA or RNA) wrapped in another (protein). Most bi-ologists do not consider viruses to be living things, but instead, infectious particles. Antibiotic drugs attack bacteria, not viruses.

Vancomycin Resistance Crisis

When microbes began resisting penicillin, medical researchers fought back with chemical cousins, such as methicillin and oxacillin. By 1953, the antibiotic armamentarium included chloramphenicol, neomycin, terramycin, tetracycline, and cephalosporins. But today, researchers fear that we may be nearing an end to the seemingly endless flow of antimicrobial drugs.

At the center of current concern is the antibiotic vancomycin, which for many infections is literally the drug of “last resort,” says Michael Blum, M.D., medical officer in FDA’s division of anti-infective drug products. Some hospital-acquired staph infections are resistant to all antibiotics except vancomycin.

Now vancomycin resistance has turned up in another common hospital bug, enterococcus. And since bacteria swap resistance genes like teenagers swap T-shirts, it is only a matter of time, many microbiologists believe, until vancomycin-resistant staph infections appear. “Staph aureus may pick up vancomycin resistance from enterococci, which are found in the normal human gut,” says Madden. And the speed with which vancomycin resistance has spread through enterococci has prompted researchers to use the word “crisis” when discussing the possibility of vancomycin-resistant staph.

Vancomycin-resistant enterococci were first reported in England and France in 1987, and appeared in one New York City hospital in 1989. By 1991, 38 hospitals in the United States reported the bug. By 1993, 14 percent of patients with enterococcus in intensive-care units in some hospitals had vancomycin-resistant strains, a 20-fold increase from 1987. A frightening report came in 1992, when a British researcher observed a transfer of a vancomycin-resistant gene from enterococcus to Staph aureus in the laboratory. Alarmed, the researcher immediately destroyed the bacteria.

Writer Ricki Lewis is a geneticist and textbook author.

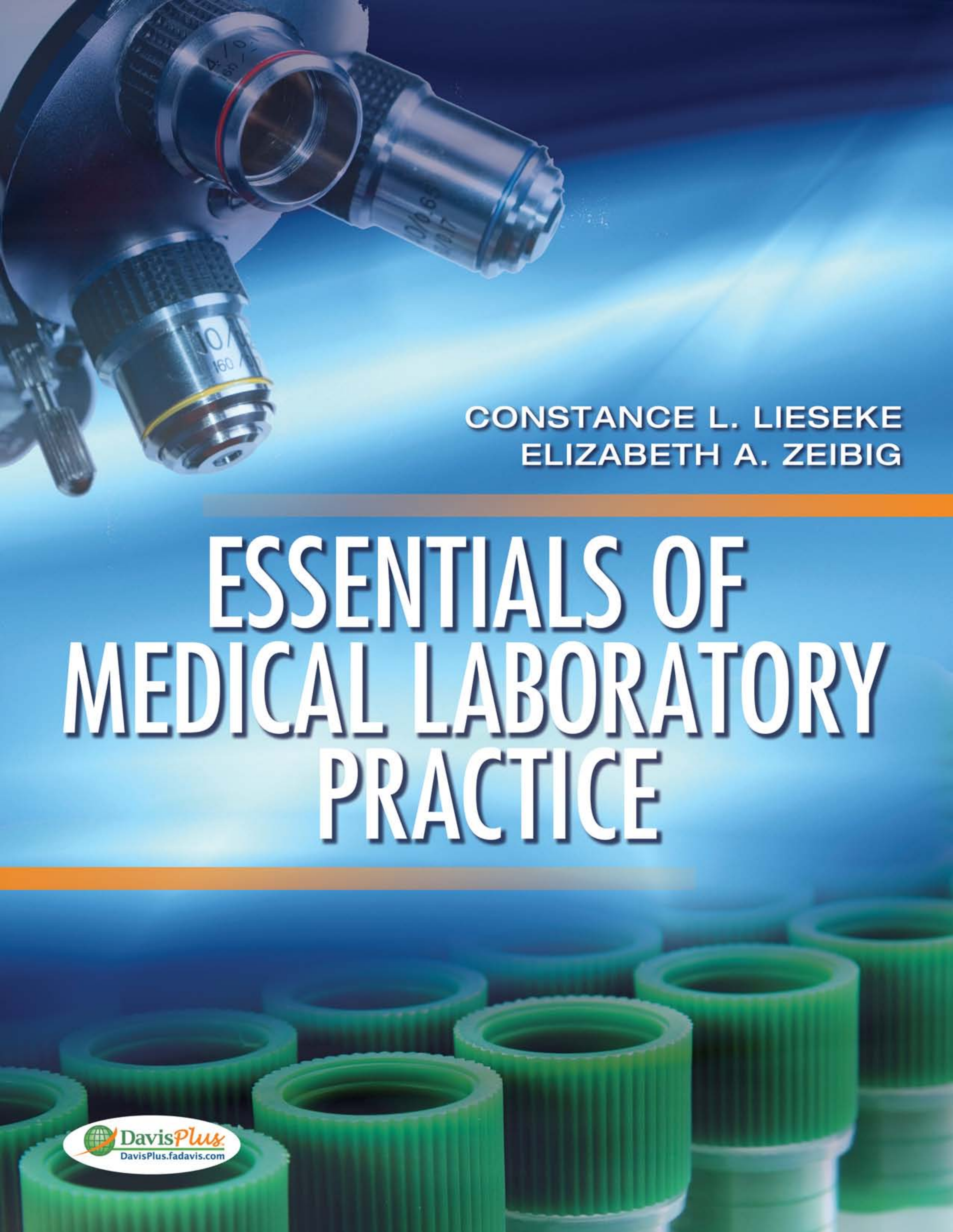

CONSTANCE L. LIESEKE
ELIZABETH A. ZEIBIG

ESSENTIALS OF MEDICAL LABORATORY PRACTICE

DavisPlus
DavisPlus.fadavis.com

Contents in Brief

SECTION I. **Overview of the Laboratory** 2

CHAPTER 1. The Clinical Laboratory 5
CHAPTER 2. Regulations Governing Laboratory Personnel 25
CHAPTER 3. Laboratory Safety and Preventing the Spread of Disease 35
CHAPTER 4. Assuring Quality 63
CHAPTER 5. Legal and Ethical Issues 83
CHAPTER 6. Laboratory Equipment 95

SECTION II. **Specimen Collection and Processing** 118

CHAPTER 7. Overview of Specimen Collection and Processing 121
CHAPTER 8. Collection and Processing of Blood Samples 133
CHAPTER 9. Collection and Processing of Urine Samples 195
CHAPTER 10. Collection and Processing of Samples for Microbial Studies 215

SECTION III. **Hematology and Coagulation** 256

CHAPTER 11. Overview of Hematology 259
CHAPTER 12. Complete Blood Count With Differential 269
CHAPTER 13. Hemoglobin and Hematocrit 285
CHAPTER 14. Erythrocyte Sedimentation Rate 303
CHAPTER 15. Coagulation Studies 313

SECTION IV. **Clinical Chemistry** 332

CHAPTER 16. Overview of Clinical Chemistry 335
CHAPTER 17. Glucose Testing 349
CHAPTER 18. Other Select Chemistry Tests 371

SECTION V. **Urinalysis** 388

CHAPTER 19. Urinalysis 391
CHAPTER 20. Physical Characteristics of Urine 401
CHAPTER 21. Chemical Examination of Urine and Feces 413
CHAPTER 22. Microscopic Examination of Urine 439

SECTION VI. **Immunology** 456

CHAPTER 23. **Immunology** 459
CHAPTER 24. **Immunological Based Rapid Testing** 473

APPENDIX A: **Reference Ranges** 491
APPENDIX B: **Answers to Test your Knowledge Questions** 493
APPENDIX C: **Tube Guide for BD Vacutainer Venous Blood Collection** 505

GLOSSARY 507
INDEX 529

Contents

SECTION I. Overview of the Laboratory 2

CHAPTER 1. The Clinical Laboratory 5
The Clinical Laboratory 6
Laboratory Departments 8
Why Is Laboratory Testing Performed? 9
The Role of the Medical Assistant in the Laboratory 10
Information Flow in the Clinical Laboratory 10
Three Phases of Laboratory Testing 21

CHAPTER 2. Regulations Governing Laboratory Personnel 25
Laboratory Professionals 26
Clinical Laboratory Improvement Amendment of 1988 29

CHAPTER 3. Laboratory Safety and Preventing the Spread of Disease 35
Infection Control and Laboratory Safety 37
Core Concepts of Infection Control 37
Laboratory Safety 48
Diseases Caused by Bloodborne Pathogens in the Laboratory Setting 55
Postexposure Follow-up Procedure 58

CHAPTER 4. Assuring Quality 63
Assuring Quality in the Laboratory 64
Quality Control and Quality Assurance 64
What Is Quality? 65
Other Methods of Assuring Laboratory Quality 73

CHAPTER 5. Legal and Ethical Issues 83
Legal and Ethical Issues in the Laboratory Environment 84
Laws and Ethics 84
Legal Concepts Affecting Patient Interactions 85
Health Information Portability and Accountability Act 88
Ethics 90
Risk Management and the Medical Assistant 92

CHAPTER 6. Laboratory Equipment 95
Laboratory Equipment 96
Equipment Used for Automated Clinical Laboratory Improvement Amendment–Waived Testing 107

SECTION II. Specimen Collection and Processing 118

CHAPTER 7. Overview of Specimen Collection and Processing 121
Specimen Ordering 122
Patient Identification 125
Patient Preparation 126
Specimen Collection at Home 127
Items That Must Be Documented With Specimen Collection 127
Labeling Information 128
Chain of Custody 128

CHAPTER 8. Collection and Processing of Blood Samples 133
Anatomy and Physiology of the Cardiovascular System 135
Site Selection 138
Contraindications and Areas to Avoid 140
Venipuncture Equipment 143
Methods Used for Venipuncture 146
Blood Collection Tubes 147
Order of Draw for Venipuncture 150
Capillary Punctures 150
Preparation for Blood Collection 152
Specimen Processing 153
Potential Negative Outcomes of Venipuncture and Capillary Puncture 184
Other Processing Procedures 186
Preparation of a Peripheral Blood Smear for Staining 186
Wright's Stain Procedure 189

CHAPTER 9. Collection and Processing of Urine Samples 195
Types of Urine Specimens 196
Urine Specimen Processing 210

CHAPTER 10. Collection and Processing of Samples for Microbial Studies 215
Microbiology Sample Collection Guidelines 218
Detailed Microbiology Sample Collection Procedures 223
Special Sample Collection and Processing Procedures 240
Processing Microbiology Samples 244
Microbiology Test Results 251

SECTION III. Hematology and Coagulation 256

CHAPTER 11. Overview of Hematology, Nikki A. Marhefka 259
Hematopoiesis—Blood Cell Formation 260
Types of Blood Cells in the Circulating Blood 263
Analysis of the Formed Elements 266

CHAPTER 12. Complete Blood Count With Differential, Nikki A. Marhefka 269
Complete Blood Count 270
Automated Analyzers for Complete Blood Count Testing 279

CHAPTER 13. Hemoglobin and Hematocrit, Nikki A. Marhefka 285
Hemoglobin 286
Hemoglobinopathies 287
Hemoglobin Testing 288
Hematocrit 293
The Relationship of Hemoglobin and Hematocrit Values 299
Anemia 299

CHAPTER 14. Erythrocyte Sedimentation Rate, Nikki A. Marhefka 303
Erythrocyte Sedimentation Rate 304
Plasma Proteins Affecting the Erythrocyte Sedimentation Rate 304
The Influence of Red Blood Cells on the Erythrocyte Sedimentation Rate 305
Reference Ranges 305
Clinical Significance of Erythrocyte Sedimentation Rate Testing 305
Erythrocyte Sedimentation Rate Determination 306
Potential Sources of Error for the Erythrocyte Sedimentation Rate Procedure 307

CHAPTER 15. Coagulation Studies 313
Purpose of Coagulation Studies 314
Mechanisms of Blood Clotting 315
Disorders Diagnosed or Monitored With Laboratory Coagulation Tests 316
Laboratory Tests Used to Diagnose Coagulation Disorders or Monitor Anticoagulant Therapy 318
Specimen Requirements for Coagulation Testing 327

SECTION IV. Clinical Chemistry 332

CHAPTER 16. Overview of Clinical Chemistry 335
Clinical Chemistry 336
Specimen Types Used for Clinical Chemistry Analysis 337
CLIA-Waived Clinical Chemistry Test 338
Other Common Clinical Chemistry Tests 340
Reference Ranges 342
Potential Sources of Error 342

CHAPTER 17. Glucose Testing 349
Glucose Utilization and Control Mechanisms 351
Pathophysiology of Glucose Metabolism 352
Types of Glucose Tests Performed 356
Glucose Testing Methods 361

CHAPTER 18. Other Select Chemistry Tests 371
Lipid Testing 372
Electrolytes 378

SECTION V. Urinalysis 388

CHAPTER 19. Urinalysis 391
Historical Perspective of Urinalysis 392
Anatomy and Physiology of the Urinary System 393
Sequence of Urine Production and Excretion 396
Clinical Significance of Urine Testing Results 396
Quality Assurance for Urine Testing Procedures 398

Standard Precautions Used When Analyzing Urine Specimens 398
Types of Urine Specimens 398

CHAPTER 20. Physical Characteristics of Urine 401
Physical Characteristics of Urine 402
Potential Sources of Error 407

CHAPTER 21. Chemical Examination of Urine and Feces 413
Urine Analytes and Their Clinical Significance 414
Potential Sources of Error 420
Safety Precautions 422
Quality Control Procedures 422
Urine Testing Methods 422
Confirmatory Urine Testing 424
Fecal Occult Blood Testing 430

CHAPTER 22. Microscopic Examination of Urine 439
Reasons for Performing Urine Microscopic Examinations 440
Common Formed Elements in the Urine and Their Clinical Significance 440
Methods Used for Urine Microscopic Examinations 447
Reporting Urine Microscopic Results 447
Role of the Medical Assistant in Microscopic Urine Examination Procedures 452
Quality Control and Quality Assurance Procedures for Urine Microscopic Examinations 452

SECTION VI. Immunology 456

CHAPTER 23. Immunology 459
Immunity and Immunology 460
The Immune Process 461
How Immunity Is Acquired 464
Failure of Our Immune Systems 465
Blood Types 465
Immunology Testing Methods 468
Common Serology Tests Performed in Reference or Hospital Laboratories 468
CLIA–Waived Tests Commonly Performed in the Physician Office Laboratory 469

CHAPTER 24. Immunological Based Rapid Testing 473
Immunology Methods and Procedures 474
Rapid Testing 475
Common CLIA–Waived Rapid Tests and Their Clinical Significance 476

APPENDIX A: Reference Ranges 491
APPENDIX B: Answers to Test Your Knowledge Questions 493
APPENDIX C: Tube Guide for BD Vacutainer Venous Blood Collection 505

GLOSSARY 507
INDEX 529

On the Horizon

Laboratory testing is an essential component of effective patient care. The results obtained from laboratory testing provide valuable information that assists physicians and other primary health providers with the diagnosis and monitoring of disease.

Relevance for the Medical Assistant (Health-Care Provider)

Medical assistants and other health-care support personnel are frequently involved in the collection of samples for laboratory testing. In many settings, these individuals perform select laboratory tests that are typically considered basic and uncomplicated. Laboratorians who have completed extensive education and training are able to perform complex laboratory testing. Medical assistants and other health-care support personnel may be responsible for the collection of samples and, in many cases, performance of the tasks associated with the processing and transport of samples to the laboratory for complex testing. On occasion, these individuals may have the opportunity to examine laboratory test results as part of their encounters with patients.

Case in Point

It is the first week of your student practicum at Maple Grove Clinic. You have been assigned to work for Dr. Pueblo under the direction of your clinical instructor, Doris. The second patient of the day is Mr. Hershey. You observe Doris interact with Mr. Hershey and perform the preliminary appropriate tasks associated with Mr. Hershey's appointment. After examining the patient, Dr. Pueblo emerges and informs Doris that he will be ordering laboratory tests. Doris takes this opportunity to introduce you to the laboratory.

Questions for Consideration:

- Compare and contrast the three phases of laboratory testing: preanalytical, analytical, and postanalytical.
- Government regulations allow medical assistants and other health-care support staff to perform what category of laboratory tests?
- What is the purpose of practicing universal precautions when dealing with laboratory specimens?
- True or False. Laboratory test results can be reported only if the quality control samples are tested and found to be in range.
- What is the name of the federal act that regulates laboratory-testing personnel?

Section I

Overview of the Laboratory

Chapter 1: The Clinical Laboratory provides the reader with basic definitions of terms such as *laboratory test* and *normal range*; identification of the areas of a typical laboratory; and an introduction to preanalytical, analytical, and postanalytical testing. A discussion of the importance of the laboratory and role of the medical assistant and other health-care support personnel as they relate to laboratory testing is presented. The difference between the regulations governing clinical laboratories and laboratories housed in a physician office is introduced in terms of the testing that can be performed in each setting.

Chapter 2: Regulations Governing Laboratory Personnel identifies the individuals who perform testing in clinical laboratories, including the Medical Laboratory Technician (MLT) and the Medical Laboratory Scientist (MLS), along with the range of tasks that each level of practitioner can perform. A detailed description and justification for the role that medical assistants and other health-care support personnel may assume in the laboratory under federal regulations, specifically in a physician office setting, is presented.

Chapter 3: Laboratory Safety and Preventing the Spread of Disease covers associated key concepts and documentation procedures, including those for Material Safety Data Sheets (MSDS); the handling of fire and chemical spills; and biological and electrical considerations. Personal protective equipment (PPE) and handwashing concepts are addressed. This chapter contains sections dedicated to bloodborne pathogen standards, universal precautions, handling of needles and sharps, and preventing the spread of disease.

Chapter 4: Assuring Quality covers the essential aspects and importance of quality assurance and quality control. The performance and documentation of quality control procedures, as well as the recognition and resolution of quality control results that are not in range, are detailed. The concepts, considerations, and reasoning behind proficiency/competency testing are explored.

Chapter 5: Legal and Ethical Issues explores a number of legal and ethical issues associated with laboratory testing. Topics include the Clinical Laboratory Improvement Act of 1988 (CLIA '88), fraud and abuse, informed consent, "the right to know," and the Health Insurance Portability and Accountability Act (HIPAA).

Chapter 6: Laboratory Equipment introduces the reader to key pieces of equipment used in a typical physician office laboratory, including centrifuges, microscopes, and point-of-care instruments. A brief overview of the principles behind these and other key hematology and chemistry analyzers are introduced and described.

Chapter 1

The Clinical Laboratory

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- The Clinical Laboratory
  - Physician Office Laboratories (POLs)
  - Hospital Laboratories
  - Reference Laboratories
- Laboratory Departments
- Why Is Laboratory Testing Performed?
- The Role of the Medical Assistant in the Clinical Laboratory
- Information Flow in the Clinical Laboratory
  - Laboratory Requisitions
  - Advance Beneficiary Notice of Noncoverage
  - Laboratory Directory
  - Laboratory Reports
- Three Phases of Laboratory Testing
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

1-1 Define the key terms for this chapter.

1-2 Describe the different types of laboratories presented in the text and the common tests available in each.

1-3 Identify the different departments in a hospital or reference laboratory and list some of the tests performed in these departments.

1-4 Provide several reasons that laboratory testing might be performed.

1-5 Explain the roles a medical assistant might play in a laboratory setting.

1-6 List and justify the various pieces of information that must be included on a laboratory requisition.

1-7 Explain the concept of Advance Beneficiary Notice of Noncoverage (ABN) and how it affects laboratory reimbursement.

1-8 Explain the purpose of a laboratory directory, as well as how a laboratory directory may be used when preparing to collect a specimen.

1-9 Compare and contrast the function of a laboratory requisition and a laboratory report.

1-10 Identify the different phases of laboratory testing, and explain the flow of the laboratory testing process.

1-11 Provide examples of preanalytical, analytical, and postanalytical procedures and how they affect the quality of laboratory testing.

CAAHEP/ABHES STANDARDS

CAAHEP 2008 Standards

II.A.2: Distinguish between normal and abnormal test results

ABHES Standards

- None

KEY TERMS

Advance Beneficiary Notice of Noncoverage (ABN)
Ambulatory
Asymptomatic
CLIA-waived
Complexity
CPT code
Efficacy
Hospital laboratories
ICD-9 code
Laboratory
Laboratory directory
Laboratory report
Laboratory requisition
Normal ranges
Panels
Physician office laboratories (POLs)
Plasma
Point-of-care testing (POCT)
Profiles
Quality control (QC)
Reference laboratories
Reference ranges
Serum
Specimens
STAT

The clinical laboratory plays a vital role in patient care. Diagnostic puzzles are solved each day by using results obtained from testing procedures performed on blood, tissue, and other body fluids. Quality patient care depends on excellent laboratory practices, so it is important to build a solid understanding of the structure and function of the clinical laboratory before performing testing procedures. This chapter explains the reasons that laboratory testing may be ordered and identifies the types of laboratories where testing is performed. The role of the medical assistant in the clinical laboratory as well as the roles played by other laboratory personnel are also introduced. To expand understanding of how the clinical laboratory operates, the flow of information in the laboratory setting is also explained.

THE CLINICAL LABORATORY

A clinical **laboratory** is any place where **specimens** from the human body may be collected, processed, examined, or analyzed. With this broad definition in mind, then, a laboratory might be someone's home, a long-term care facility, the office of a health-care provider, a free clinic, a hospital, a small facility used strictly for specimen collection and processing, or a large and complex reference laboratory. As of January 2010, the Centers for Medicare & Medicaid Services identified more than 216,000 clinical laboratories in the United States.[1] To simplify our discussion of the various types of laboratories, we will classify them into three basic groups as described below (Table 1-1).

Physician Office Laboratories

Physician office laboratories (POLs) are clinical laboratories within physician offices where laboratory testing is carried out on specimens obtained from the practices' own patients. More than half of all the laboratories in the United States are POLs. This type of laboratory can be advantageous because the results for tests performed on site are available quickly and patient treatment can begin immediately if necessary. Testing and specimen collection in these office laboratories may be accomplished by various members of the health-care team, such as medical assistants,

---

[1] Data From CLIA Update—January 2010
Division of Laboratory Services
Centers for Medicare & Medicaid Services
Laboratories by Type of Facility
https://www.cms.gov/CLIA/downloads/factype.pdf

TABLE 1-1

Types of Laboratories and Sample of Testing Performed

Type of Laboratory	Testing Performed
Physician office laboratory	CLIA-waived tests Moderate complexity testing (with appropriate training, staffing, and supervision) Microscopic examinations (performed by health-care provider or qualified staff)
Hospital laboratory	CLIA-waived tests CLIA moderate- and high-complexity tests (based on the needs of the patients served in the facility) Point-of-care testing Microscopic examinations
Reference laboratory	CLIA-waived tests Moderate- and high-complexity tests Microscopic examinations

clinical or medical laboratory technicians who have been trained in laboratory testing, or phlebotomists who are trained to draw blood and perform some simple laboratory procedures. These **ambulatory** (outpatient) laboratories generally perform low **complexity** tests that are designated as Clinical Laboratory Improvement Amendment–waived (**CLIA-waived)** tests by the U.S. Food and Drug Administration (FDA). CLIA-waived tests are laboratory examinations and procedures that use simple and accurate methods requiring very little interpretation to report correct results. These tests may include the following:

- Rapid microbiology testing for the presence of group A *Streptococcus* and influenza
- Urine analysis
- Pregnancy testing
- Tests such as those for mononucleosis, *Helicobacter pylori,* and HIV
- Coagulation testing to monitor patients who are taking anticoagulants
- Glucose levels and other tests used to monitor diabetic patients
- Fecal occult blood tests for the presence of blood in the stool
- Cholesterol testing

Physician office laboratories may also provide testing that is more complex if staff members are properly trained to perform such procedures. More details about the Clinical Laboratory Improvement Amendment and the classification of laboratory tests are provided in Chapter 2.

Hospital Laboratories

Hospital laboratories generally offer laboratory testing that meets the needs of their respective institution. For instance, if a hospital specializes in a certain type of surgical procedure or treatment, the hospital laboratory will offer extensive testing in that specialty area, in addition to the standard tests required to monitor the health of the other patients in the hospital. In most situations, hospital laboratories perform high volumes of routine test procedures. Hospital laboratories may also serve as a reference laboratory for the local community, especially for **STAT** testing needs. (STAT tests are those that require immediate results.) Tests performed at a hospital laboratory may include the following:

- Electrolytes
- Kidney function tests
- Liver function analysis
- Blood typing and crossmatches for transfusions
- Identification of microorganisms and antibiotic sensitivity testing
- Urine analysis
- Coagulation testing
- Cardiac enzyme assays
- Complete blood counts (CBCs) and other hematology testing

Hospital laboratories may also offer another method for testing samples, called **point-of-care testing (POCT)**. Point-of-care tests are actually performed at the patient's bedside rather than in the laboratory, using a portable instrument that gives immediate results. These tests may be performed by laboratory personnel or in some situations by other hospital employees who have been trained to perform the testing.

Reference Laboratories

Reference laboratories perform more tests annually than either POLs or the hospital laboratories, processing perhaps thousands of specimens per day. These tests include those that are performed at hospital laboratories, but reference laboratories may offer specialized testing that is not performed at either hospital laboratories or POLs. Specimens may be sent to a reference laboratory from all over the country.

Test Your Knowledge 1-1

Which type of laboratory offers the widest variety of tests? (Outcome 1-2)

LABORATORY DEPARTMENTS

Testing in hospital and reference laboratories is usually departmentalized for efficiency, as presented in Table 1-2. This allows all testing that uses similar instrumentation or methods to be performed in the same area of the laboratory. These departments may include the following:

- **Specimen Processing:** The area of the laboratory where all incoming specimens are sorted, accessioned into the computer system of the laboratory, and appropriately labeled for transport to their respective departments for testing. The specimen processing department may also prepare specimens for transport to reference laboratories.
- **Hematology:** Whole blood testing, which focuses on the formed elements (the blood cells) in the blood. Coagulation testing is also performed in this department.
- **Clinical Chemistry:** Testing performed on **plasma** or **serum** (the liquid portion of the blood) and includes analysis of the substances dissolved in the bloodstream. Most testing is automated, and many of the tests are performed as panels or groups of tests.
- **Serology/Immunohematology:** Testing that focuses primarily on the presence of antigens or antibodies on cells or in the liquid portion of the blood. Blood typing and antibody screens and crossmatches for transfusions may also be performed in this department.

TABLE 1-2

Summary of Laboratory Departments, Common Tests Performed, and Specimen Types

Laboratory Department	Common Tests Performed	Specimen Types
Specimen Processing	Incoming specimens sorted Specimens accessioned into computer system Specimens labeled and prepared for testing in separate department	All types
Hematology	Complete blood count Hemoglobin and hematocrit measurements Coagulation studies Erythrocyte sedimentation rate Differential white blood cell count Platelet counts	Whole blood
Clinical Chemistry	Electrolytes Glucose Blood urea nitrogen Creatinine Thyroid testing Cardiac enzyme testing Comprehensive metabolic panel Cholesterol/lipid testing Many automated chemistry tests and panels	Serum, plasma, urine, cerebrospinal fluid, amniotic fluid
Serology and Immunohematology	Various tests looking for antigens and/or antibodies such as: RPR Mononucleosis testing HIV tests Chlamydia testing Antistreptolysin O test Pregnancy testing C-reactive protein ABO/Rh blood typing Antibody screens and crossmatches for transfusions Newborn testing Prenatal testing	Serum, plasma, whole blood

TABLE 1-2–cont'd

Summary of Laboratory Departments, Common Tests Performed, and Specimen Types

Laboratory Department	Common Tests Performed	Specimen Types
Urinalysis	Physical appearance of urine Urine chemical analysis Microscopic urine analysis	Urine
Microbiology/Parasitology	Identification of pathogenic microorganisms Streptococcal screens Antibiotic sensitivity testing	Blood, urine, wound specimens, tissue, stool, cerebrospinal fluid, sputum, urethral and vaginal discharge, nails, skin scrapings
Cytology	Examination of various specimens for abnormal cells Chromosomal studies Pap smears	Urine, skin, tissue specimens, sputum
Coagulation	Testing for presence or absence of adequate clotting factors	Whole blood
Histology/Pathology	Examination for abnormal form and structure in tissues	Tissue/organs, biopsy specimens from various parts of the body, preserved and fresh specimens examined

- **Urinalysis:** The physical appearance of urine is assessed, and urine chemical and microscopic analysis is performed.
- **Microbiology/Parasitology:** Identification of pathogenic microorganisms and antibiotic sensitivity testing.
- **Cytology:** Examination of various specimens for abnormal cells, chromosomal studies, Pap smears.
- **Coagulation:** Specimens testing for the presence of various clotting factors.
- **Histology/Pathology:** Tissue samples examined for abnormal function and form.

Test Your Knowledge 1-2

Which laboratory department would perform a test to see if someone had a pathogenic microorganism in a wound? (Outcome 1-3)

WHY IS LABORATORY TESTING PERFORMED?

Laboratory testing is critical for appropriate patient treatment. The results obtained from a blood test or evaluation of other types of specimens from the human body are essential for the health-care provider to gain information that is not available through the patient history or physical examination. Health-care providers generally order laboratory tests for at least one of these reasons:

1. *To assign a diagnosis.* Laboratory testing may be used to help with a differential diagnosis when the patient's symptoms are vague or similar to those of other disease states. Diagnostic testing can also confirm a clinical diagnosis (such as diabetes or a myocardial infarction) when that confirmation could lead to more effective treatment for the patient. Microbiological testing may also fall into this category, as the identification of the bacteria and/or virus involved in the infection will be essential to establish an appropriate plan of action.
2. *Prevention or early detection of disease through screening tests.* Many chronic health conditions are **asymptomatic** in the early disease stages. If these conditions are identified with routine screening tests (such as cholesterol measurements and routine prostate cancer screening) early in the disease process, treatment may be much more effective in keeping the patient healthy. These screening tests are being ordered with more frequency in the general population as more

information becomes available about the advantages of early diagnosis and treatment.

3. *Ongoing assessment of the patient's progress and treatment.* Once a diagnosis has been established and treatment has begun, it is very important that the patient is monitored closely. This often requires frequent blood tests to determine therapeutic drug levels, hepatic profiles when a particular medication may cause liver damage, or hemoglobin tests to see how effective the treatment might be for someone who is anemic. Patients may need to have their blood drawn daily, monthly, or quarterly. The frequency is dependent on the pathological condition. It may be necessary to use specimens other than blood to monitor the treatment for a patient. Sometimes repeat cultures are also necessary to verify the **efficacy** (effectiveness) of the antibiotic used to fight an infection.

Test Your Knowledge 1-3

When a pregnant young woman sees her nurse-midwife for a routine checkup, the practitioner requests a test to see if the patient is developing diabetes during the pregnancy. Why is the test being performed?

a. To assign a diagnosis
b. To screen for prevention or early disease detection
c. To assess the ongoing status of treatment
d. None of the above (Outcome 1-4)

THE ROLE OF THE MEDICAL ASSISTANT IN THE CLINICAL LABORATORY

A medical assistant can contribute in the clinical laboratory in various ways. The comprehensive training received by medical assistants allows them to perform clinical duties with direct patient contact, as well as offering employment opportunities in an administrative capacity. A clinical medical assistant working in a physician office laboratory may be called on to perform phlebotomy or supervise urine drug screen collections for routine employment purposes, educate patients concerning collection protocols, collect and process microbiology specimens, and perform various test procedures. Medical assistants performing clinical duties may also be asked to prepare microscopic specimens for examination, and focus the microscope for providers to evaluate the specimen. In addition, the medical assistant with appropriate training may also perform microscopic examinations such as white blood cell differential counts and urine microscopy examinations.

A clinical medical assistant working in a hospital or reference laboratory may function as a phlebotomist or may work as a laboratory assistant. A laboratory assistant may be responsible for setting up microbiology cultures, answering phones and responding to requests for laboratory results, assisting with bone marrow aspiration or other collection procedures, processing specimens and delivering them to the various laboratory departments, performing CLIA-waived or moderate-complexity tests available in that laboratory setting, or assisting with preparation of tissue and cellular samples for examination. Medical assistants may also perform point-of care testing and assist with quality assurance in the laboratory.

An administrative medical assistant in a laboratory setting might help with inventory and ordering of supplies, data input, answering telephone calls, scheduling, billing and coding, production and transmission of laboratory reports, creation of other laboratory forms, and various nonclinical patient interactions. These duties are necessary in every laboratory, regardless of size or location.

Test Your Knowledge 1-4

True or False: A medical assistant can work only in an administrative capacity in a hospital or reference laboratory. (Outcome 1-4)

Test Your Knowledge 1-5

Are medical assistants able to perform laboratory testing? (Outcome 1-5)

INFORMATION FLOW IN THE CLINICAL LABORATORY

In addition to understanding the organization of the clinical laboratory setting and identifying the reason laboratory testing might be performed, it is important to know how information is exchanged within the laboratory. To ensure that laboratory test results are as meaningful as possible to the health-care provider, all those involved with specimen collection, processing, and testing must use the required paperwork and database correctly. This information exchange will involve a **laboratory requisition,** a **laboratory directory,** a computer database, and a **laboratory report.**

Laboratory Requisitions

A laboratory requisition (Fig. 1-1) is the form used by a physician (or other qualified health-care professional) to

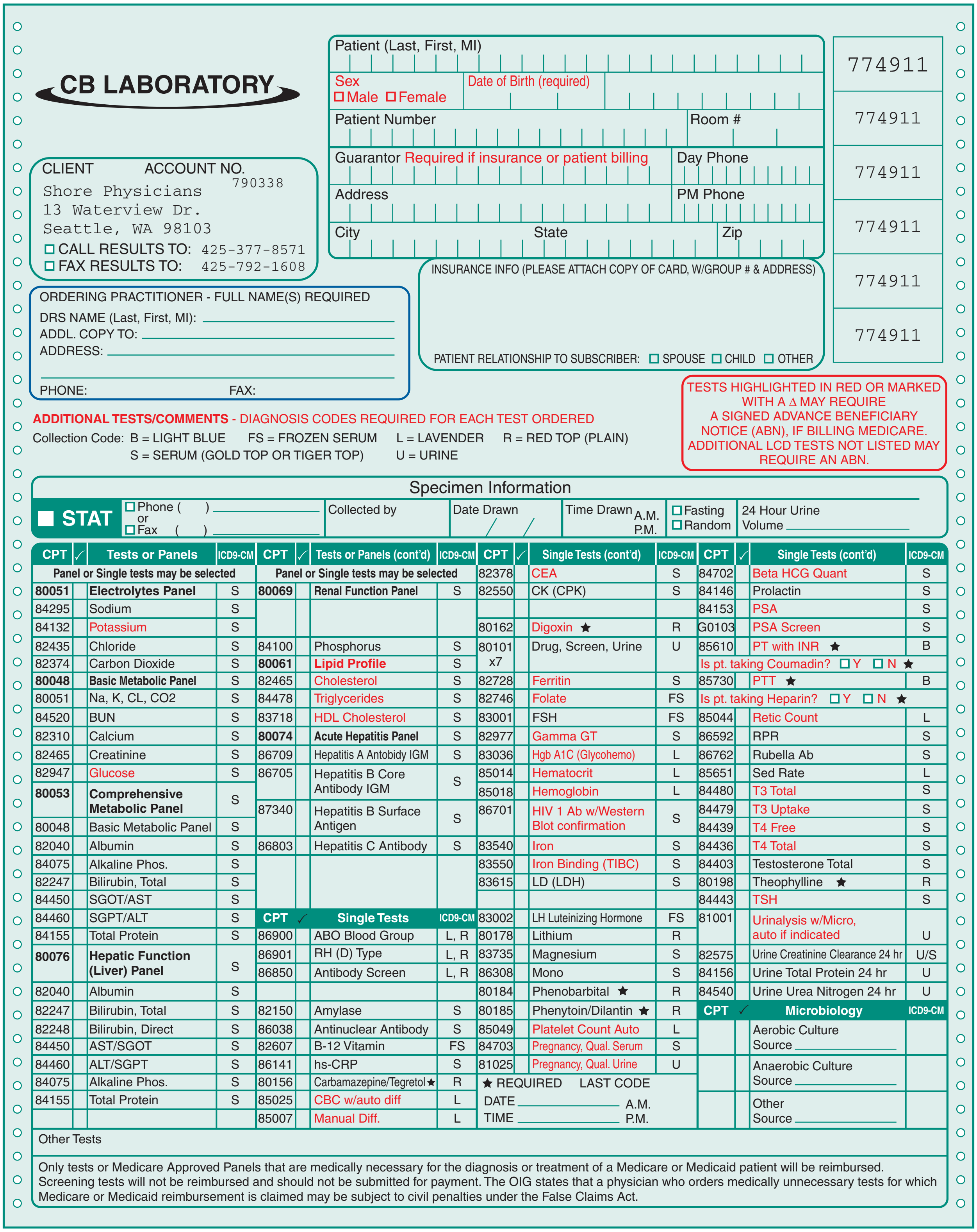

CB LABORATORY

Patient (Last, First, MI)

Sex ☐ Male ☐ Female
Date of Birth (required)

Patient Number
Room #

Guarantor Required if insurance or patient billing
Day Phone

Address
PM Phone

City
State
Zip

774911
774911
774911
774911
774911
774911

CLIENT ACCOUNT NO.
790338
Shore Physicians
13 Waterview Dr.
Seattle, WA 98103
☐ CALL RESULTS TO: 425-377-8571
☐ FAX RESULTS TO: 425-792-1608

INSURANCE INFO (PLEASE ATTACH COPY OF CARD, W/GROUP # & ADDRESS)

PATIENT RELATIONSHIP TO SUBSCRIBER: ☐ SPOUSE ☐ CHILD ☐ OTHER

ORDERING PRACTITIONER - FULL NAME(S) REQUIRED
DRS NAME (Last, First, MI):
ADDL. COPY TO:
ADDRESS:
PHONE: FAX:

TESTS HIGHLIGHTED IN RED OR MARKED WITH A Δ MAY REQUIRE A SIGNED ADVANCE BENEFICIARY NOTICE (ABN), IF BILLING MEDICARE. ADDITIONAL LCD TESTS NOT LISTED MAY REQUIRE AN ABN.

ADDITIONAL TESTS/COMMENTS - DIAGNOSIS CODES REQUIRED FOR EACH TEST ORDERED

Collection Code: B = LIGHT BLUE FS = FROZEN SERUM L = LAVENDER R = RED TOP (PLAIN)
S = SERUM (GOLD TOP OR TIGER TOP) U = URINE

Specimen Information

■ STAT	☐ Phone ( ) or ☐ Fax ( )	Collected by	Date Drawn / /	Time Drawn A.M. P.M.	☐ Fasting ☐ Random	24 Hour Urine Volume

CPT	✓	Tests or Panels	ICD9-CM
Panel or Single tests may be selected			
80051		**Electrolytes Panel**	S
84295		Sodium	S
84132		Potassium	S
82435		Chloride	S
82374		Carbon Dioxide	S
80048		**Basic Metabolic Panel**	S
80051		Na, K, CL, CO2	S
84520		BUN	S
82310		Calcium	S
82465		Creatinine	S
82947		Glucose	S
80053		**Comprehensive Metabolic Panel**	S
80048		Basic Metabolic Panel	S
82040		Albumin	S
84075		Alkaline Phos.	S
82247		Bilirubin, Total	S
84450		SGOT/AST	S
84460		SGPT/ALT	S
84155		Total Protein	S
80076		**Hepatic Function (Liver) Panel**	S
82040		Albumin	S
82247		Bilirubin, Total	S
82248		Bilirubin, Direct	S
84450		AST/SGOT	S
84460		ALT/SGPT	S
84075		Alkaline Phos.	S
84155		Total Protein	S

CPT	✓	Tests or Panels (cont'd)	ICD9-CM
Panel or Single tests may be selected			
80069		**Renal Function Panel**	S
84100		Phosphorus	S
80061		**Lipid Profile**	S
82465		Cholesterol	S
84478		Triglycerides	S
83718		HDL Cholesterol	S
80074		**Acute Hepatitis Panel**	S
86709		Hepatitis A Antobidy IGM	S
86705		Hepatitis B Core Antibody IGM	S
87340		Hepatitis B Surface Antigen	S
86803		Hepatitis C Antibody	S
CPT	**✓**	**Single Tests**	**ICD9-CM**
86900		ABO Blood Group	L, R
86901		RH (D) Type	L, R
86850		Antibody Screen	L, R
82150		Amylase	S
86038		Antinuclear Antibody	S
82607		B-12 Vitamin	FS
86141		hs-CRP	S
80156		Carbamazepine/Tegretol ★	R
85025		CBC w/auto diff	L
85007		Manual Diff.	L

CPT	✓	Single Tests (cont'd)	ICD9-CM
82378		CEA	S
82550		CK (CPK)	S
80162		Digoxin ★	R
80101 x7		Drug, Screen, Urine	U
82728		Ferritin	S
82746		Folate	FS
83001		FSH	FS
82977		Gamma GT	S
83036		Hgb A1C (Glycohemo)	L
85014		Hematocrit	L
85018		Hemoglobin	L
86701		HIV 1 Ab w/Western Blot confirmation	S
83540		Iron	S
83550		Iron Binding (TIBC)	S
83615		LD (LDH)	S
83002		LH Luteinizing Hormone	FS
80178		Lithium	R
83735		Magnesium	S
86308		Mono	S
80184		Phenobarbital ★	R
80185		Phenytoin/Dilantin ★	R
85049		Platelet Count Auto	L
84703		Pregnancy, Qual. Serum	S
81025		Pregnancy, Qual. Urine	U

★ REQUIRED LAST CODE
DATE ____________ A.M.
TIME ____________ P.M.

CPT	✓	Single Tests (cont'd)	ICD9-CM
84702		Beta HCG Quant	S
84146		Prolactin	S
84153		PSA	S
G0103		PSA Screen	S
85610		PT with INR ★	B
Is pt. taking Coumadin? ☐ Y ☐ N ★			
85730		PTT ★	B
Is pt. taking Heparin? ☐ Y ☐ N ★			
85044		Retic Count	L
86592		RPR	S
86762		Rubella Ab	S
85651		Sed Rate	L
84480		T3 Total	S
84479		T3 Uptake	S
84439		T4 Free	S
84436		T4 Total	S
84403		Testosterone Total	S
80198		Theophylline ★	R
84443		TSH	S
81001		Urinalysis w/Micro, auto if indicated	U
82575		Urine Creatinine Clearance 24 hr	U/S
84156		Urine Total Protein 24 hr	U
84540		Urine Urea Nitrogen 24 hr	U
CPT	**✓**	**Microbiology**	**ICD9-CM**
		Aerobic Culture Source ____________	
		Anaerobic Culture Source ____________	
		Other Source ____________	

Other Tests

Only tests or Medicare Approved Panels that are medically necessary for the diagnosis or treatment of a Medicare or Medicaid patient will be reimbursed. Screening tests will not be reimbursed and should not be submitted for payment. The OIG states that a physician who orders medically unnecessary tests for which Medicare or Medicaid reimbursement is claimed may be subject to civil penalties under the False Claims Act.

Figure 1-1 Sample laboratory requisition. Current Procedural Terminology © 2011 American Medical Association, All Rights Reserved.

document the tests that are to be performed for a patient. The form is filled out by the health-care provider who orders the test or by the provider's office staff. This requisition must be complete with various data linking the patient to the tests ordered, and must provide billing information for that patient. The laboratory staff will use this requisition to enter the orders into their database for the patient, and the form is often used for reimbursement procedures as well. Although all laboratories may use unique formats, standardized information should be present on all requisitions for testing and reimbursement purposes:

- Patient demographic information, including name, address, phone number, and birth date.
- Patient gender.
- Complete patient insurance and billing information.
- Date and time of collection, and identification of the person who performed the collection.
- Documentation of how the results are to be relayed to the health-care provider who ordered the tests, and documentation of any other practitioner who is to receive a copy of the results.
- Appropriate diagnostic information in the form of an **ICD-9 code** to allow for reimbursement for the test ordered. An ICD-9 code is the numeric indicator for the diagnosis or symptom associated with the laboratory test ordered.
- Test selection, clearly marked with a check, circle, or X in the appropriate area on the requisition. Tests may also be written in if they are not preprinted on the requisition used.
- Any additional comments that may assist with the ordering or interpretation of the test.

Test Your Knowledge 1-6

Why is it important to have the name of the ordering physician/practitioner on a requisition when ordering laboratory testing? (Outcome 1-6)

Laboratory tests may be ordered as **panels** or **profiles,** which consist of a group of tests that are designed to indicate a problem with a specific organ system or disease process. Tests may also be ordered individually, or the requisition may contain orders for a combination of both. Often the requisitions are organized according to the general types of tests, such as hematology, serology, or chemistry. On the requisition, each test will have a five-digit procedure code (known as a **CPT code),** which is necessary for reimbursement purposes. There may also be a code or symbol included on the requisition, indicating the type of tube or specimen container that is necessary for that specific test. This information is valuable as a quick reference for the phlebotomist when collecting the specimen.

Most requisitions also include some sort of label that can be peeled off and applied to the specimen container. This label may have numbers, bar-code symbols, or both. These labels link the patient information provided on the requisition with that specific sample once it has been entered into the database for the laboratory.

Advance Beneficiary Notice of Noncoverage

An **Advance Beneficiary Notice of Noncoverage** (**ABN**) is a document that may be necessary to receive

Procedure 1-1: Completing a Laboratory Requisition

TASK

Correctly fill out a laboratory requisition, including all necessary information for reimbursement. Include an ABN if necessary.

CAAHEP/ABHES Standards

None

CONDITIONS

- Laboratory requisition
- Patient demographic information
- Patient insurance information
- Laboratory order from qualified health-care professional
- ICD-9 code for diagnosis or symptoms
- Advance Beneficiary Notice of Noncoverage

Procedure	Rationale
1. Greet and identify patient using at least two unique identifiers.	All patients must be identified properly before collecting samples or performing laboratory testing.
2. Verify test ordered.	All laboratory test orders should be verified by checking the chart and/or requisition more than once.
3. Fill in the requisition with the patient's demographic information by printing it clearly, or by using a data entry program if available at your location.	The patient's demographic information (including age and gender) is used for reimbursement as well as establishing reference ranges for the tests ordered.
4. Ask patient for an insurance card if he or she is insured.	The insurance card should be used for the information added to the requisition; do not rely on information that the patient provides from memory.
5. Copy both sides of the insurance card, and add the information to the requisition in the appropriate site.	Be sure to include the insurance identification number as well as the group number for appropriate reimbursement. The back of the card is important to copy, as it provides the insurance company's contact information in case of claim questions.
6. Establish the relationship of the patient to the insured. Document that on the requisition.	Sometimes the patient is not listed on the insurance card; it may be a spouse or a child or other dependent.
7. If the patient has a medication level ordered (such as digoxin), ask when the patient took the last dose. Use the laboratory directory if necessary to verify whether it is the correct time to draw the specimen. Note the time of the last dose on the requisition.	Some medication levels are clinically significant only when drawn at a specific interval after the last dose.
8. If the patient has Medicare Part B as his or her primary insurance coverage, verify whether an Advance Beneficiary Notice of Noncoverage is necessary. This must be completed and signed before the specimen collection begins.	The laboratory may have a database established in which the employee can type in the name of the test ordered and the diagnosis code provided, and the software will be able to determine whether an ABN is necessary. If there is no automated system in place, the employee may need to use an alternative reference. Many laboratories provide their clients with a book of covered diagnosis codes for the tests with limited coverage.
9. Clarify which physician ordered the test and document that on the requisition. If additional copies are to be sent to other health-care providers, add this information to the requisition as well.	Many times the requisitions will be preprinted with the name of the physician office or practice group. It is still necessary to document which physician within the practice actually ordered the laboratory test to be performed.
10. Document whether the test was ordered as STAT or routine.	If the test was ordered as STAT, it needs to be drawn and processed immediately. It may be necessary to contact a laboratory courier to pick up the sample immediately, or to notify the technician that there is a STAT order.

Continued

Procedure 1-1: Completing a Laboratory Requisition—cont'd

Procedure	Rationale
11. Check the requisition for an ICD-9 code for each test to be performed. If a code is not present, contact the health-care provider or check the chart to obtain a code.	An ICD-9 code is necessary for every test for successful reimbursement.
12. Have the patient wait in a comfortable environment until the sample can be collected. Place the requisition in an appropriate location to alert other staff members that a specimen collection is waiting.	Sometimes the employee who fills out the requisition is not the same person who will collect the sample.

reimbursement for laboratory services. It is specifically used for patients who have Medicare Part B as their primary insurance coverage. The purpose of an ABN is to inform Medicare-covered patients that payment may be denied for a specific laboratory test and that the patient will be billed for the full cost of that test if he or she chooses to have it performed. The ABN allows the patients to make an informed decision about whether they wish to have the tests performed, with the realization that they may be responsible for the total cost.

The ABN must be verbally reviewed with the patient, and any questions about potential reimbursement must be answered before it is signed, whether or not the patient wishes to have the laboratory work performed. This process must occur before the specimen is collected from the patient. The person collecting the specimen must ensure that the test ordered has been identified on the ABN form and that there is documentation of the anticipated reason for noncoverage. Figure 1-1 shows an example of an ABN. The employee filling out the form also must provide an estimated cost in writing so that the patient knows what the financial responsibility may be if he or she decides to have the laboratory test performed. The patient's decision about having the test performed must be documented on the form, along with the patient's signature and that day's date. The patient must receive a copy of the form after it is signed, and a copy must also be kept on file with the laboratory.

Medicare coverage for laboratory tests may be denied for various reasons, including frequency of testing, the diagnosis provided by the health-care provider for a specific test, or because the test ordered is considered experimental. Whenever a patient with Medicare coverage has a specimen collected, the laboratory employee responsible for the specimen collection must verify whether the test ordered will be covered by Medicare for the reason that the test is ordered. The regulations affecting coverage are different for geographical regions across the country, and change frequently. Most laboratories now have a way to verify coverage by using a computer database, but it can be difficult to keep abreast of changes. It is unlawful to have every Medicare patient fill out an ABN "just in case"; it is the responsibility of those collecting the specimen to make their best effort to verify coverage before an ABN is signed.

The format for this Advance Beneficiary Notice of Noncoverage document is defined by the Centers for Medicare & Medicaid Services (CMS), and is updated periodically (CMS document number CMS-R-131). Medicare bases its decision for coverage on "medical necessity" rules, which define those tests that the agency deems medically necessary for specific health conditions, and how frequently these tests should be performed. It is important to remember that just because Medicare has limited coverage on specific tests, the patient should never be told that the health-care provider gave a "bad code" or that the test was ordered for the "wrong reason."

Test Your Knowledge 1-7

What is the purpose of an ABN? (Outcome 1-7)

Laboratory Directory

As discussed, laboratory requisitions often provide information about what type of tube to use for a blood draw for a specific test by using a code or symbol, but

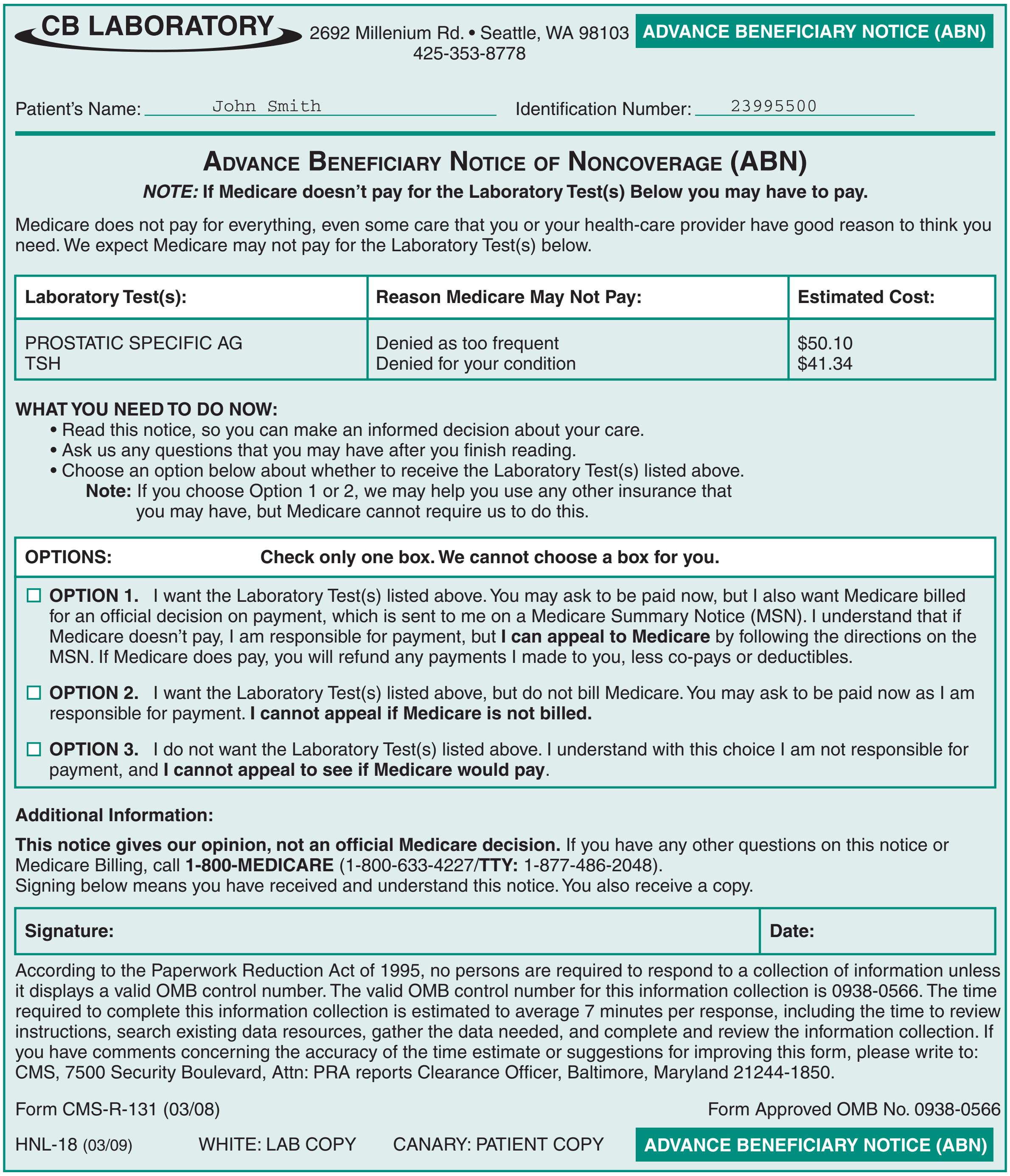

CB LABORATORY 2692 Millenium Rd. • Seattle, WA 98103
425-353-8778
ADVANCE BENEFICIARY NOTICE (ABN)

Patient's Name: John Smith Identification Number: 23995500

ADVANCE BENEFICIARY NOTICE OF NONCOVERAGE (ABN)

NOTE: **If Medicare doesn't pay for the Laboratory Test(s) Below you may have to pay.**

Medicare does not pay for everything, even some care that you or your health-care provider have good reason to think you need. We expect Medicare may not pay for the Laboratory Test(s) below.

Laboratory Test(s):	Reason Medicare May Not Pay:	Estimated Cost:
PROSTATIC SPECIFIC AG TSH	Denied as too frequent Denied for your condition	$50.10 $41.34

WHAT YOU NEED TO DO NOW:

- Read this notice, so you can make an informed decision about your care.
- Ask us any questions that you may have after you finish reading.
- Choose an option below about whether to receive the Laboratory Test(s) listed above.

Note: If you choose Option 1 or 2, we may help you use any other insurance that you may have, but Medicare cannot require us to do this.

OPTIONS: **Check only one box. We cannot choose a box for you.**

☐ **OPTION 1.** I want the Laboratory Test(s) listed above. You may ask to be paid now, but I also want Medicare billed for an official decision on payment, which is sent to me on a Medicare Summary Notice (MSN). I understand that if Medicare doesn't pay, I am responsible for payment, but **I can appeal to Medicare** by following the directions on the MSN. If Medicare does pay, you will refund any payments I made to you, less co-pays or deductibles.

☐ **OPTION 2.** I want the Laboratory Test(s) listed above, but do not bill Medicare. You may ask to be paid now as I am responsible for payment. **I cannot appeal if Medicare is not billed.**

☐ **OPTION 3.** I do not want the Laboratory Test(s) listed above. I understand with this choice I am not responsible for payment, and **I cannot appeal to see if Medicare would pay**.

Additional Information:

This notice gives our opinion, not an official Medicare decision. If you have any other questions on this notice or Medicare Billing, call **1-800-MEDICARE** (1-800-633-4227/**TTY:** 1-877-486-2048).
Signing below means you have received and understand this notice. You also receive a copy.

Signature:	**Date:**

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number. The valid OMB control number for this information collection is 0938-0566. The time required to complete this information collection is estimated to average 7 minutes per response, including the time to review instructions, search existing data resources, gather the data needed, and complete and review the information collection. If you have comments concerning the accuracy of the time estimate or suggestions for improving this form, please write to: CMS, 7500 Security Boulevard, Attn: PRA reports Clearance Officer, Baltimore, Maryland 21244-1850.

Form CMS-R-131 (03/08) Form Approved OMB No. 0938-0566

HNL-18 (03/09) WHITE: LAB COPY CANARY: PATIENT COPY **ADVANCE BENEFICIARY NOTICE (ABN)**

Figure 1-2 Example of an Advance Beneficiary Notice of Noncoverage (ABN).

the information provided on the requisition is limited at best. The requisition does not include information about how to process and store the specimen, or what the minimum volume may be for the test ordered. It also does not include information about how often the laboratory performs that specific analysis if the physician needs to know this information. This additional information about specimen collection and handling may be found in a laboratory directory (Fig. 1-3).

POINT OF INTEREST 1-1

Advance beneficiary notice of noncoverage

If a laboratory accepts Medicare assignment, it is not legal to bill Medicare beneficiaries for laboratory tests that are covered by Medicare Part B. However, Medicare does not cover all laboratory tests, so there are many situations in which the health-care providers may order tests that are not covered by Medicare but that provide valuable information for patient treatment. The only way that the laboratory can be reimbursed for these tests is to have the patient complete a valid Advance Beneficiary Notice of Noncoverage, also known as an ABN. This form must be completed and signed prior to the specimen collection. The ABN form informs the Medicare beneficiary of the test that has been ordered, and specifies the reason that Medicare may not pay for the tests, as well as providing the estimated cost for the test.
The ABN is necessary in these situations:

- When the test ordered is not medically necessary in Medicare's opinion for the ICD-9 code (diagnosis) that is provided. This situation is also present if there is no ICD-9 code given for the test ordered.
- Many screening tests are not covered by Medicare; these are often not covered if there is no evidence of disease.
- Tests that are ordered more frequently than what is recommended by Medicare.
- Any test that is still considered to be investigational or experimental in nature. These tests have not yet been approved by the FDA.

Often those who work with Medicare patients will use the term *limited coverage* to refer to the rules that govern laboratory reimbursement. This means that Medicare will pay for laboratory tests only when they meet certain criteria. These may be national coverage decisions or local coverage decisions. Nationally, there are several tests established that have limited coverage (based on the ICD-9 codes, frequency of testing, or the investigational status) and regional restrictions may add to the list that is established on a national basis.

By signing the ABN, the Medicare patient is giving the laboratory permission to bill him or her directly for the test to be performed. If the employee who draws the blood does not get an ABN signed when appropriate, the laboratory will be performing the test for free, as it cannot have the ABN signed after the process is completed. If the specimen was collected by an employee at a physician office and transported to the laboratory for testing, the physician office will often receive a bill for the service from the laboratory, because that laboratory cannot be reimbursed in any other way.

It is illegal to have every Medicare patient fill out an ABN; these forms are to be used only when necessary for noncovered tests. In addition, the laboratory must retain a copy of the ABN and the patient must also be given a copy. Advanced Beneficiary Notices of Noncoverage are also used for other medical services; it is necessary to have a specific one for each type of noncovered service.

Commonly, a laboratory directory is a computer database of specific information about tests that are performed by that laboratory. The directory may also be provided in a book format, which can be consulted by facilities that do not have Internet access. The laboratory directory (also known as a directory of services) provides the following types of information:

- The internal test number used to enter the test order into the database
- CPT five-digit code used for reimbursement for that test
- Related information if there is more than one place within the directory where the test is addressed
- Acronyms or abbreviations that may be listed on the requisition for that test, or perhaps used to order the test in the computer system for that laboratory
- The type of specimen required, and in some cases the type or color of tubes to use for the blood draw or identification of the additives that must be present in the tubes
- The requested specimen volume and the minimum acceptable specimen volume
- Collection notes and/or specific requirements necessary for some tests
- Storage instructions for the specimen after collection and during transportation (room temperature, frozen, refrigerated, etc.)
- **Reference ranges** (also known as **normal ranges)** for the test ordered
- Clinical significance and interpretation of the test results
- Testing intervals/frequency and testing locations

Once the specimen requirements have been established, the collection process can continue. The employee who is performing the collection must document two unique

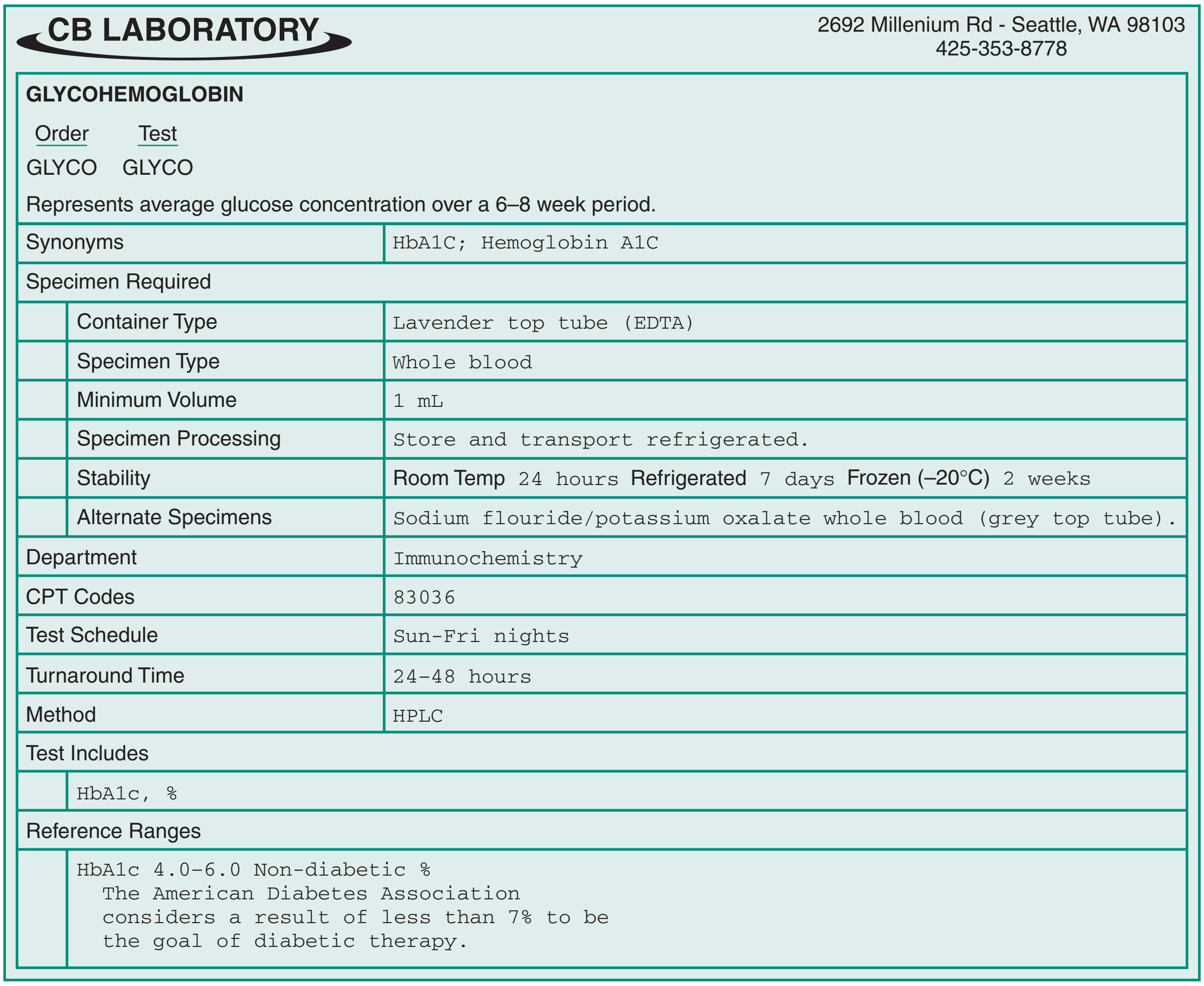

CB LABORATORY

2692 Millenium Rd - Seattle, WA 98103
425-353-8778

GLYCOHEMOGLOBIN

Order	Test
GLYCO	GLYCO

Represents average glucose concentration over a 6–8 week period.

Synonyms		HbA1C; Hemoglobin A1C
Specimen Required		
	Container Type	Lavender top tube (EDTA)
	Specimen Type	Whole blood
	Minimum Volume	1 mL
	Specimen Processing	Store and transport refrigerated.
	Stability	Room Temp 24 hours Refrigerated 7 days Frozen (–20°C) 2 weeks
	Alternate Specimens	Sodium flouride/potassium oxalate whole blood (grey top tube).
Department		Immunochemistry
CPT Codes		83036
Test Schedule		Sun-Fri nights
Turnaround Time		24-48 hours
Method		HPLC
Test Includes		
	HbA1c, %	
Reference Ranges		
	HbA1c 4.0-6.0 Non-diabetic % The American Diabetes Association considers a result of less than 7% to be the goal of diabetic therapy.	

Figure 1-3 Sample of information found in a laboratory directory.

Procedure 1-2: Use a Laboratory Directory

TASK

Use a laboratory directory to clarify the collection requirements and processing procedures for a laboratory order.

CAAHEP/ABHES Standards

None

CONDITIONS

- Laboratory requisition
- Laboratory order from qualified health-care professional
- Laboratory directory book or database including laboratory service information

Continued

Procedure 1-2: Use a Laboratory Directory—cont'd

Procedure	Rationale
1. Obtain the laboratory requisition or the laboratory order from the health-care provider. Verify the test ordered, especially if it is a test that may be performed on more than one body fluid.	Some tests, such as creatinine or total protein, are commonly ordered on blood but may also be ordered using a urine specimen.
2. Look up the test alphabetically using the computer database or the laboratory directory reference book.	This can sometimes be difficult if the test is ordered under an acronym or abbreviation or if there are two names used for the test, such as Tegretol, which may also be known as carbamazepine. Many laboratory directories will include a cross-reference that may help the user to find the correct entry.
3. Identify the type of specimen to be collected.	The information may be very straightforward, specifying the color of tube to be used. (For example, lavender-top or green-top tube.) However, sometimes the specimen requirements are described as "heparinized plasma" or "EDTA plasma." The employee needs to know which type of tube to use for this type of specimen.
4. Verify any restrictions on specimen type, or notations of unacceptable specimens.	For instance, many tests cannot use hemolyzed samples, or sometimes samples that are over 2 hours old cannot be used.
5. Verify the acceptable minimum volume, if listed.	There may be a requested volume (e.g., 1 mL) as well as a minimum volume listed. The minimum volume may be important if drawing the specimen from a child or other patient with difficult-to-access veins.
6. Identify the specimen processing instructions.	Is the specimen supposed to be spun within a certain time after the blood draw? Is the plasma or serum to be frozen? The processing instructions are very important for the specimen to be acceptable when it arrives at the testing laboratory.
7. Identify the schedule for performance of the test ordered.	The reference laboratory may list the testing schedule so that the health-care provider will be aware of how long it will take before he or she can expect the results for the test.
8. Prepare the necessary supplies to perform the specimen collection.	If a specimen must be frozen immediately or kept at a certain temperature, it is necessary to have the supplies at hand immediately after the specimen is collected. These should be prepared before proceeding.

identifiers for the patient, the employee ID (or initials), the date of collection, and the time of collection on the requisition and on the tube or alternative collection container. This same employee may also be responsible for entering the patient information into the computer database, whereas in other laboratories the specimen and paperwork may be prepared for transport to another location where the information will be added to the database upon arrival.

Test Your Knowledge 1-8

Why would a medical assistant use a laboratory directory before performing a blood draw? (Outcome 1-8)

Laboratory Reports

Once the specimen has been processed, delivered to the correct department, and tested within the laboratory, a laboratory report is generated to transmit the test results back to the health-care provider. An example of a laboratory report is presented in Figure 1-4. The laboratory report document will list the results for the tests performed, as well as the reference ranges (also known as normal ranges) that have been established by the laboratory for that test. Reference ranges are the results that are expected in the general healthy population 95% of the time for a particular laboratory test. A range is necessary (instead of a specific value) because of differences in the population due to age, race, and gender. Geographical locations may also affect the reference range, as will the testing methods used by that laboratory. A notation will be present on the laboratory report for all results that are outside the expected reference range for that patient, based on demographics, the testing method, and the test ordered. The gender and age of the patient may affect the reference ranges used for interpretation of the results, so it is critical that this information is provided whenever a sample is collected.

The laboratory report will also include the date and time of the specimen collection, the name and identification number for the patient, and the name and address of the laboratory where the test was performed. The specimen source is identified, as well as the date and time the report was generated.

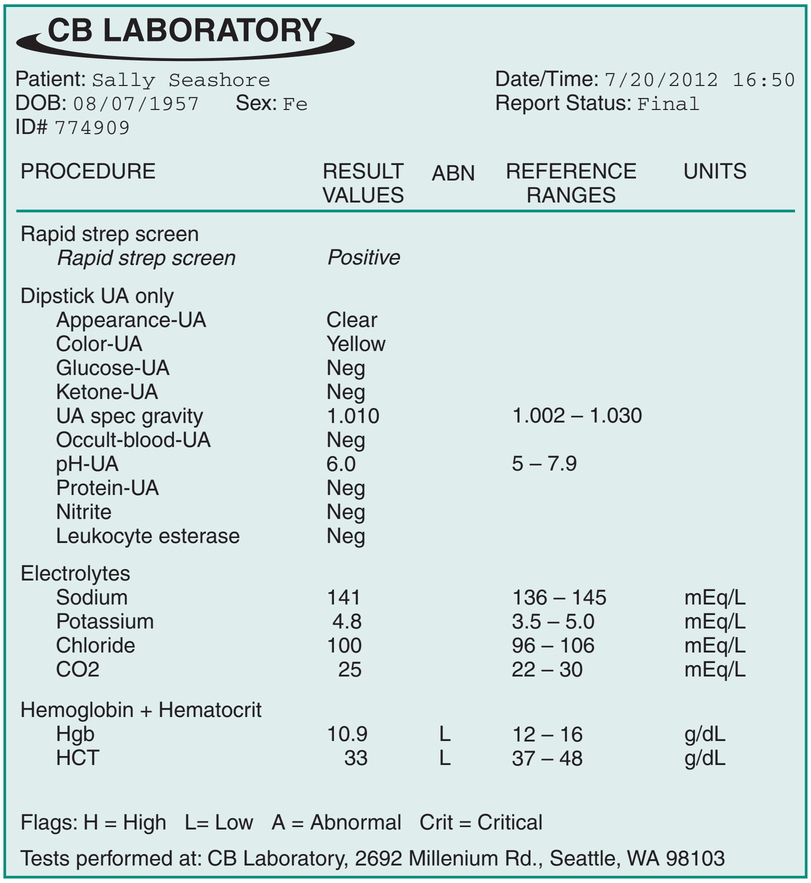

CB LABORATORY

Patient: Sally Seashore
DOB: 08/07/1957 Sex: Fe
ID# 774909

Date/Time: 7/20/2012 16:50
Report Status: Final

PROCEDURE	RESULT VALUES	ABN	REFERENCE RANGES	UNITS
Rapid strep screen				
Rapid strep screen	*Positive*			
Dipstick UA only				
Appearance-UA	Clear			
Color-UA	Yellow			
Glucose-UA	Neg			
Ketone-UA	Neg			
UA spec gravity	1.010		1.002 – 1.030	
Occult-blood-UA	Neg			
pH-UA	6.0		5 – 7.9	
Protein-UA	Neg			
Nitrite	Neg			
Leukocyte esterase	Neg			
Electrolytes				
Sodium	141		136 – 145	mEq/L
Potassium	4.8		3.5 – 5.0	mEq/L
Chloride	100		96 – 106	mEq/L
CO2	25		22 – 30	mEq/L
Hemoglobin + Hematocrit				
Hgb	10.9	L	12 – 16	g/dL
HCT	33	L	37 – 48	g/dL

Flags: H = High L= Low A = Abnormal Crit = Critical

Tests performed at: CB Laboratory, 2692 Millenium Rd., Seattle, WA 98103

Figure 1-4 Sample laboratory report.

Procedure 1-3: Distinguish Between Normal and Abnormal Test Results

TASK

Use a laboratory report to identify normal and abnormal test results for a patient.

CAAHEP/ABHES Standards

CAAHEP 2008 Standards

II.A.2: Distinguish between normal and abnormal test results

ABHES Standards

None

Conditions

- Laboratory report
- Pen, pencil, or highlighter

Procedure	Rationale
1. Examine the laboratory report for all necessary information.	All laboratory reports must include the following: 1. The name of the patient 2. The patient ID 3. The gender and age of the patient 4. Laboratory results with documented reference ranges 5. The name of the ordering health-care provider 6. The date and time of specimen collection 7. The date and time that the specimen was tested 8. The name of the laboratory where the test was performed
2. Identify the column on the laboratory report where the reference ranges are noted for each of the test results.	This column is usually near the patient results, and the data are often listed as a range, except in the case of tests that provide a positive and negative test result, or in the case of microbiology testing, antibiotic sensitivity testing, or immunology testing.
3. Compare the patient results to the reference range column.	Remember that the reference ranges may take the age and gender of the patient into consideration; the ranges will not necessarily be the same for all patients tested.
4. Circle or highlight the results that are not within the reference ranges. Document High or Low next to the values in the provided area.	Highlighting or circling these results may bring them to the attention of the provider. When a laboratory report is printed in the office (rather than one used in the classroom) the High or Low results are noted on the laboratory report when it is printed. However, the office protocol may also include highlighting or circling the results when the report arrives in the office to make certain that they are not overlooked.

Laboratory reports may be hand-delivered to the ordering health-care provider by a medical assistant within an office or via a courier service from a reference laboratory. They may also be faxed, mailed, or in some cases transmitted via e-mail. In some situations, the laboratory reports may also be available online through a dedicated laboratory link so that the provider can view the results on site. These results must be reviewed as soon as possible so that appropriate action can be taken for those outside of the normal reference range. Laboratory reports are a legal document that becomes part of the patient's health record.

Test Your Knowledge 1-9

In which ways are laboratory requisitions and laboratory reports similar? (Outcome 1-9)

THREE PHASES OF LABORATORY TESTING

As we begin to understand the way that a laboratory is organized and how information is transferred through the testing process, it is important to realize that there are three components, known in laboratory jargon as *phases,* associated with laboratory testing: preanalytical, analytical, and postanalytical. A description of each phase of testing is detailed below and summarized in Table 1-3.

- **Preanalytical Phase:** The preanalytical phase of laboratory testing refers to the situations and actions that take place prior to the collection, during the collection, and during the processing/storage/transportation of the specimen. Phlebotomists and medical assistants often participate in this phase of laboratory testing. The importance of this phase cannot be emphasized enough. Generally speaking, the majority of problems associated with laboratory tests result from inadequate or inappropriate specimen collection, processing, storage, and transportation (i.e., the preanalytical phase of testing).
- **Analytical Phase:** The analytical phase of laboratory testing refers to the performance of the tests that have been ordered. This phase also includes maintenance and calibration of laboratory equipment and instruments associated with the testing and performing

TABLE 1-3

Identification of, Definition of, and the Personnel Responsible for the Three Phases of Laboratory Testing

Phase	Definition	Personnel Responsible
Preanalytical	All procedures/processes that occur before the specimen is actually tested. Includes patient preparation, accurate paperwork and data entry, appropriate specimen collection, processing, storage and transportation.	Medical assistant, phlebotomist, CLT or MLT, other laboratory professionals involved in this process
Analytical	All procedures/processes involved in the testing of the specimen. This includes the way the testing instrument was prepared and maintained, how the testing supplies were stored, appropriate training of the personnel performing the test, and quality control to ensure that the testing methods are working properly.	Medical assistant, medical laboratory professional performing the test, supervisor responsible for training and overseeing the process, personnel performing maintenance on instruments
Postanalytical	All procedures/processes that affect how the test results are handled when the analysis has been completed. These may include review and analysis of the results by the person performing the test, appropriate follow-through on extremely high or low results, how the results are recorded (in a computer or on paper, etc.) phone calls, report printing, report sorting, appropriate fax procedures, charting procedures within an office, physician review procedures, contact with patient if necessary for follow-up.	Various laboratory professionals that perform the test, administrative laboratory support personnel that process the results, medical assistants or other physician office personnel who perform charting procedures, physicians and other health-care professionals who interact directly with the patients.

quality control (QC) measures, which are in place to validate the test reagents/kits, the testing process, and training of the laboratory personnel performing the test.

- **Postanalytical Phase:** The postanalytical phase of laboratory testing includes the processes associated with the recording and reporting of laboratory results, storage and/or disposal of specimens after testing, and provider and patient notification of test results. Even if the other two phases of testing occur without any exceptions, if this phase isn't handled appropriately, then the overall experience will not be a positive one, and may negatively affect patient treatment.

Test Your Knowledge 1-10

List the three phases of laboratory testing.

(Outcome 1-10)

Test Your Knowledge 1-11

If a specimen is collected into a tube with the wrong type of anticoagulant for the test ordered, which phase of laboratory testing has gone wrong? (Outcome 1-11)

SUMMARY

Laboratory testing is vital to the diagnosis of patient disease and implementation of proper patient treatment, as it provides critical information that cannot be gathered with a physical examination alone. Testing may occur in physician office laboratories, in hospital laboratories, or in large reference laboratories. The larger laboratories may be organized by departments as needed for efficient specimen handling. Various personnel perform the testing procedures, and their education and titles will vary depending on their responsibilities and the type of testing performed in the respective laboratory. Medical assistants play an important role in the laboratory by providing direct patient interaction, specimen collection, and processing or performance of laboratory tests.

Laboratory information, such as that provided on a requisition, is vital for proper testing procedures, result interpretation, and reimbursement for services provided. The exchange of information begins with a requisition and may involve the use of a laboratory directory for specimen requirements and other specifics. The process ends with the laboratory reports that provide the ordering practitioner with the final test results and reference ranges. Reimbursement procedures may involve laboratory personnel, so it is important that these individuals understand the processes involved with the necessary documentation on a requisition and the use of an ABN.

There are three phases of laboratory testing, including the procedures surrounding specimen collection and handling, the actual testing process, and the manner in which the results are reported back to the physician. These are the preanalytical phase, the analytical phase, and the postanalytical phase. All three phases must be handled correctly for laboratory results to be as meaningful as possible as part of quality patient care.

TIME TO REVIEW

1. If a patient is asymptomatic, it means that he or she: Outcome 1-1
   - **a.** Demonstrates specific symptoms of a disease state
   - **b.** Complains of multiple symptoms
   - **c.** Does not exhibit symptoms
   - **d.** Is not ill
2. True or False: A specimen may be defined as a part of something that demonstrates the characteristics of the whole. Outcome 1-1
3. What do hospital and reference laboratories have in common? Outcome 1-2
4. Is urine analyzed only in physician office laboratories? Outcome 1-2
5. A medical assisting student needs to have a blood test performed to see if she has antibodies to hepatitis B in her bloodstream. This test would most likely be performed in this laboratory department: Outcome 1-3
   - **a.** Blood Banking
   - **b.** Hematology
   - **c.** Serology
   - **d.** Microbiology
6. Which laboratory department is responsible for accessioning specimens into the laboratory information system when they arrive at the laboratory? Outcome 1-3
7. Mr. Earl has a laboratory test performed to see if the diet and exercise changes suggested by Outcome 1-4

his health-care provider have helped to lower his cholesterol levels. Which of these reasons explains why this test is being ordered?

- **a.** Assignment of a diagnosis
- **b.** Ongoing assessment of the patient's progress and treatment
- **c.** Prevention or possible early detection through screening tests
- **d.** None of the above

8. Which of these duties might a medical assistant perform in a laboratory setting? Outcome 1-5

- **a.** Urine drug screen collections
- **b.** Phlebotomy
- **c.** Urinalysis testing
- **d.** Preparation of microbiology specimens for testing and/or analysis
- **e.** All of the above

9. True or False: A diagnosis code does not need to be included on a requisition at the time the laboratory test is ordered. Outcome 1-6

10. True or False: A requisition is used only for ordering the laboratory tests. Outcome 1-6

11. The abbreviation ABN stands for: Outcome 1-7

- **a.** Advance Beneficiary Notice of Noncoverage
- **b.** Advance Beneficiary Notification
- **c.** Action Bulletin of Noncoverage
- **d.** Advance Bulletin of Noncoverage

12. The laboratory directory will include ___________ instructions for a specimen after collection and during transportation. Outcome 1-8

13. True or False: The terms *laboratory requisition* and *laboratory report* may be used interchangeably. Outcome 1-9

14. Which phases of laboratory testing might be affected by the actions of a medical assistant in his or her various laboratory roles? Outcome 1-11

Case Study 1-1: Potential for growth

Cindy Lee, a certified medical assistant (CMA) working in a gynecology office, leaves early one Friday before her office closes. When she returns on Monday morning, she finds a box containing agar plates for microbiology specimen growth on the counter in the laboratory area. Cindy notices that the box has KEEP REFRIGERATED printed on the outside, but it appears that the box has been left out all night, and the cold packs within the box are at room temperature.

- Should Cindy be concerned about this situation?
- Which phase of laboratory testing may be affected by this oversight?

RESOURCES AND SUGGESTED READINGS

American Society for Clinical Laboratory Science Consumer Lab Testing Information
http://www.ascls.org/labtesting/index.asp
Numerous questions and answers on general laboratory testing.

American Society for Clinical Laboratory Science Introduction to Laboratory Testing
http://www.ascls.org/labtesting/labintro.asp
Introduction to the three phases of laboratory testing; also basic terms referring to analytical aspects of quality control.

Centers for Medicare & Medicaid Services Overview of Beneficiary Notices Initiative
http://www.cms.hhs.gov/bni
Links to various ABN information, including the new and old forms in use and answers to various commonly asked questions.

Good Laboratory Practices for Waived Sites
http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5413a1.html
An overview of some problems that have occurred in waived test sites and an introduction to the three phases of laboratory testing. Recommendations for good practices overall.

CLIA Update—January 2010. Division of Laboratory Services, Centers for Medicare & Medicaid Services; Laboratories by Type of Facility (Exempt/Nonexempt combined)
http://www.cms.gov/CLIA/downloads/factype.pdf
Lists the number of different types of laboratories registered with CMS as of January 2010.

Chapter 2

Regulations Governing Laboratory Personnel

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Laboratory Professionals
  - Personnel in the Laboratory Setting
  - Medical Assistants in the Laboratory
- Clinical Laboratory Improvement Amendment of 1988
  - History of the Regulation
  - Levels of Laboratory Testing Defined by CLIA
  - Employee Qualifications for Performance of CLIA Testing
  - Oversight of CLIA Laboratories
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

2-1 Define the key terms.

2-2 List the laboratory professionals present in a typical hospital, reference, or physician office laboratory.

2-3 Describe the personnel structure of the laboratory settings presented in the text.

2-4 Explain how the duties of laboratory professionals may vary, depending on their education and credentials.

2-5 Describe the role of a medical assistant in the clinical laboratory.

2-6 Explain the focus of CLIA '88, and why it was developed.

2-7 Demonstrate understanding of the levels of laboratory testing designated by CLIA '88.

2-8 Identify the laboratory professionals qualified to perform the various levels of laboratory testing as allowed by CLIA '88.

2-9 Identify the agencies responsible for overseeing CLIA '88 compliance.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

IX.C.5: Discuss licensure and certification as it applies to healthcare providers.

IX.C.8: Compare criminal and civil law as it applies to the practicing medical assistant.

ABHES Standards

4.b: Medical Law and Ethics: Federal and State Guidelines

4.f: Medical Law and Ethics: Health laws and regulations

Graduates. f: Comply with federal, state and local health laws and regulations.

KEY TERMS

Ambulatory care
CLIA '88
CLIA-waived
Centers for Medicare & Medicaid Services (CMS)
COLA
Credentialed
High-complexity tests
Labile
Moderate-complexity tests
Phlebotomist
Proficiency testing
Provider-performed microscopy procedures (PPMPs)
U.S. Department of Health and Human Services (HHS)
U.S. Food and Drug Administration (FDA)

Medical assistants are in a unique situation, as the diverse training they receive allows them to assume numerous roles in an ambulatory setting. A medical assistant may have an opportunity to interact with various health-care agencies and related entities, such as radiology clinics, insurance companies, physical therapy offices, specialty physician offices, or pharmacies. In addition, the duties of the medical assistant will most likely include communication with the laboratory, and many medical assistants also perform laboratory testing in physician office laboratories. It is important to understand who makes up the workforce of the medical laboratory so that effective communication may be established. It is also vital that medical assistants understand the regulations controlling laboratory testing as they apply to the scope of practice in the physician office laboratory to remain compliant. This chapter provides an understanding of how to keep the physician office laboratory operating within the boundaries of the regulations. It will also provide insight into the clinical laboratory by identifying which career paths are available to a medical assistant.

LABORATORY PROFESSIONALS

Laboratory professionals pay a critical role in quality health care. The test results they generate improve patient care by providing information to the physician that cannot be discovered in any other way. Even though many processes in the laboratory have become automated, it is important for those performing the testing procedures to have adequate background knowledge for appropriate interpretation of the test results. It is also imperative that laboratory employees are knowledgeable about the procedures necessary to ensure quality testing methods in the laboratory, laws and regulations governing laboratory processes, and how all the laboratory departments work together.

Personnel in the Laboratory Setting

The personnel involved in laboratory testing are classified according to their education and credentials. Table 2-1 provides a list of the laboratory professionals that may play a role in laboratory testing, including their credentials and education. The basic classifications include the following:

- **Pathologists:** These are board-certified physicians who have specialized training in disease and laboratory interpretation. A pathologist may function as the laboratory director for sites that perform all levels of laboratory testing, and is generally affiliated with hospital and reference laboratories.
- **Physicians:** A physician without any laboratory specialty training may function as the laboratory director of a physician office laboratory. Dentists may also serve as laboratory directors if they perform laboratory testing in their clinics. Physicians may not serve as directors for hospital or reference laboratories unless they have additional credentials specifically qualifying them for this setting.
- **Nurse-practitioners or physician assistants:** These are known as midlevel health-care providers, and they may function as directors of physician office laboratories. Nurse-practitioners have at least 6 years of education; physician assistants generally possess at least a 4-year degree, as well as additional focused medical education. These midlevel providers are not usually employed in a reference or hospital laboratory.
- **Medical laboratory scientists:** Personnel with this title often work in a reference or hospital laboratory as testing personnel or as supervisors. They must have at

TABLE 2-1

Laboratory professional title, education, and credentials

Laboratory Professional Title	Education	Credentials
Pathologist	Doctorate with at least 8 years of education	MD or DO, board certified as pathologist
Physician	Doctorate with at least 6 years of education	MD or DO
Physician assistant	Bachelor's degree	PA
Nurse-practitioner	Master's degree	NP or ARNP
Clinical laboratory scientist	Bachelor's or master's degree	CLS, certified by National Certification Agency for Medical Laboratory Personnel (NCA)
Medical laboratory scientist	Bachelor's degree	MLS, certified by the American Society for Clinical Pathology (ASCP) or MT if certified by the American Medical Technologists (AMT), or International Society for Clinical Laboratory Technology (ISCLT)
Medical laboratory technician	Associate degree	MLT, certified by the American Society for Clinical Pathology (ASCP) or American Medical Technologists (AMT)
Clinical or registered laboratory technician	Associate degree	International Society for Clinical Laboratory Technology (ISCLT)
Medical assistant	1-year certificate or associate degree	May be certified by the American Association of Medical Assistants (CMA) or registered by the American Medical Technologists (RMA)
Phlebotomist	Varies by state	May be nationally certified by various agencies; not always required; some states may allow training on-the-job with no formal education

Note: Those with an associate degree or higher are generally required to have certification (credentials) to work in the laboratory. Those below this level of education are not necessarily required to have formal certification.

least a 4-year degree in laboratory medicine, and are credentialed by the American Society for Clinical Pathology. Those qualified with a 4-year degree may also be known as medical technologists if they received their certification through the American Medical Technologists.

- **Medical laboratory technicians:** These employees usually perform testing procedures in a hospital or reference laboratory. They may also be employed in physician office laboratories. Medical laboratory technicians must have completed at least an associate degree in laboratory medicine, and are nationally certified through the American Society of Clinical Pathology or through the American Medical Technologists.
- **Medical assistants:** Medical assistants may work in a physician office laboratory, but they can also be employed in a reference or hospital laboratory if the tasks they perform are within their lawful scope of practice. Medical assistants have completed a medical assisting program that included at least an introduction to laboratory procedures. Medical assistants are not required to be certified to work in a laboratory, but if they choose to pursue a certification, they may test through the American Association of Medical Assistants to become a certified medical assistant (CMA), or through the American Medical Technologists to become a registered medical assistant (RMA).
- **Phlebotomists:** A **phlebotomist** may be a medical assistant, but there are also phlebotomists trained in short-term programs, either at technical colleges or on the job, specifically to draw blood and process laboratory specimens. Phlebotomists are primarily employed in physician office laboratories and hospital laboratories. Phlebotomists must have at least a high school

diploma (or general equivalency diploma [GED]) and have documented training to draw blood. They may become certified nationally, and some states require additional state certification as well.

Test Your Knowledge 2-1

May a physician assistant serve as a laboratory director? (Outcome 2-2)

Those laboratory professionals that are **credentialed** have completed the required education for their specific classification, and they have also successfully passed a comprehensive assessment examination as required by their credentialing agency. To keep their credentials current, these professionals must participate in continuous education every year. Some states also require that laboratory personnel pass an additional state examination in order to work in a laboratory setting.

When entering a clinical laboratory, it may be difficult to tell by observation which employees have which type of credential, as they are all wearing lab coats and all performing what appear to be similar tasks. Figures 2-1 and 2-2 are representative of the personnel structure in a reference laboratory, a hospital laboratory, and an **ambulatory care** setting (such as a physician office laboratory). Laboratory testing may be performed by a variety of classifications, such as medical laboratory scientists, medical or clinical laboratory technicians, medical assistants, or phlebotomists.

Test Your Knowledge 2-2

May a medical technologist serve as a supervisor in a reference laboratory? (Outcome 2-3)

Test Your Knowledge 2-3

What type of credentials must a phlebotomist possess to work in a laboratory setting? (Outcome 2-4)

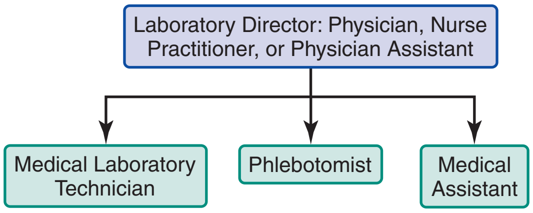
Laboratory Director: Physician, Nurse Practitioner, or Physician Assistant
Medical Laboratory Technician
Phlebotomist
Medical Assistant

Figure 2-1 The personnel structure of the laboratory professionals working in a CLIA-waived physician office laboratory.

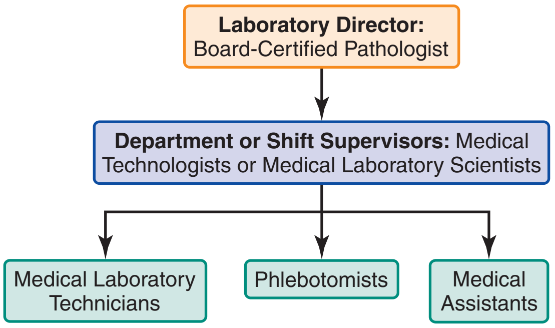
Laboratory Director: Board-Certified Pathologist
Department or Shift Supervisors: Medical Technologists or Medical Laboratory Scientists
Medical Laboratory Technicians
Phlebotomists
Medical Assistants

Figure 2-2 The personnel structure of the laboratory professionals working in a hospital or reference laboratory.

POINT OF INTEREST 2-1

Name changes

When working in a large laboratory (such as a reference or hospital laboratory), it is possible that you may encounter various credentials for individuals who are performing the same tasks. This is not unusual, but a recent merger of credentialing agencies may make this even more confusing!

Prior to 2009 there were three widely recognized credentialing agencies for laboratory personnel. They were the Board of Registry for the American Society of Clinical Pathology (ASCP), the National Credentialing Agency for Laboratory Personnel (NCA), and the American Medical Technologists (AMT). In October 2009, the NCA and the ASCP credentialing agencies decided to merge. Those who were previously credentialed by the NCA are now part of the ASCP organization, and those who were already part of the ASCP have not seen very many changes in their credential renewal processes. The certifying agency for the ASCP is now called the Board of Certification.

What has changed, however, are the terms used by these organizations for the various personnel classifications.

- **Baccalaureate degree–level personnel:** The ASCP previously used the designation *medical technologists* to describe this classification, and the NCA used *clinical laboratory scientist*. Now, the new title for those with a 4-year degree in medical technology that have successfully passed the examination is that of medical laboratory scientist, with the abbreviation MLS (ASCP).
- **Associate degree–level personnel:** The ASCP previously used the designation *medical laboratory*

technician to describe this classification, and the NCA used *clinical laboratory technician*. When the two agencies merged, they retained the designation *medical laboratory technician* for those with a 2-year degree who successfully pass the examination, and the abbreviation used for their credential is MLT (ASCP).

Medical Assistants in the Laboratory

As introduced in Chapter 1, medical assistants may play numerous roles in the clinical laboratory environment. Their education enables them to draw blood, assist with collection of other types of specimens, process specimens for testing, and perform simple or moderately complex test procedures. They may perform these tasks in the physician office laboratory, in a hospital laboratory, or in a reference laboratory. Medical assistants may also work in the microbiology or pathology department within a hospital or reference laboratory assisting with specimen preparation, answering phones, or entering specimens into the computer.

Some laboratory settings, especially a specialist or hospital laboratory, need their medical assistants to do more than draw blood or process specimens. In these settings they may be taught how to perform arterial blood draws, and also may learn how to assist with bone marrow aspirations or bronchoscopy specimen collection. In some situations, medical assistants may be running tests on automated instruments and performing urine microscopy examinations and manual differential procedures on normal blood smears. These specialized tasks require additional documented training beyond that offered in a traditional medical assisting education program, and the qualifications of the personnel performing these tasks may be regulated by state or local laws.

Test Your Knowledge 2-4

Are medical assistants only allowed to draw blood in a laboratory setting, or can they perform other duties?
(Outcome 2-5)

CLINICAL LABORATORY IMPROVEMENT AMENDMENTS OF 1988

CLIA '88 is an acronym for the Clinical Laboratory Improvement Amendment of 1988. This regulation establishes quality standards for all aspects of human medical laboratory testing, with the exception of those tests performed for research only. It was designed to ensure that all laboratory tests, regardless of where they are performed, would produce accurate reliable results and be performed in a timely manner.

History of the Regulation

The original Clinical Laboratory Improvement Act was enacted in 1967. It affected only private laboratories (federal and state laboratories were exempt) that received specimens via interstate commerce. This means that the only laboratories affected by the original act were those that provided services across state borders. These laboratories had to be licensed if they processed more than 100 interstate specimens per year. This act also improved laboratory quality standards overall by adding a requirement for **proficiency testing** to these licensed laboratories. Proficiency testing is used to verify that the results reported for a certain test are accurate. Because this Clinical Laboratory Improvement Act had such a small focus, there were many laboratories that were not affected by these requirements, especially physician office laboratories. It was estimated that of 200,000 laboratories in business in the United States, only about 12,000 were regulated.

In 1987, the *Wall Street Journal* published two articles expanding on the poor quality of laboratory testing in the United States. At that time, there was very little structured supervision of the quality of testing performed in laboratories that were not affected by the Clinical Laboratory Improvement Act of 1967. Specifically, the articles expanded on several situations in which Papanicolaou (Pap) smear tests showed false-negative results, allowing cases of cervical cancer to become advanced before a correct diagnosis was made. The cytology laboratories that were reporting these results were poorly staffed, and the technicians were overworked. Both the U.S. House of Representatives and the U.S. Senate held hearings to investigate the inaccurate Pap smear reports and other laboratory errors. As a result of these hearings, questions began to form about how laboratories functioned and about what kind of quality assurance was in place for the results reported. Therefore, it was decided that more federal oversight was necessary, and on October 31, 1988, the Clinical Laboratory Improvement Amendment of 1988 (CLIA '88) was enacted. Since that time, there have been several revisions, allowing more tests to be included in the **CLIA-waived** category, and giving the **U.S. Food and Drug Administration (FDA)** the responsibility to determine the level of complexity for laboratory tests.

Test Your Knowledge 2-5

What was CLIA '88 designed to ensure? (Outcome 2-6)

Levels of Laboratory Testing Defined by CLIA '88

CLIA '88 requires that all laboratories performing tests on human specimens must register with the **Centers for Medicare & Medicaid Services (CMS).** The registration process, and title of the certificate granted to the laboratory, is based on the type of testing performed. The classification of the testing is based on the complexity of the testing procedure and the clinical significance of the results from the test, among other factors. This amendment classifies all tests into one of three categories:

- **CLIA-waived tests:** These are tests that are so simple and accurate that the likelihood of misinterpretation is minimal. This classification also requires that the tests will pose no reasonable risk of harm to the patient if the test were accidentally performed incorrectly. CLIA-waived tests also include procedures that have been cleared for home use, such as glucose testing performed by diabetics at home. Examples of CLIA-waived tests are streptococcal screens, urine pregnancy tests, and mononucleosis testing. Many CLIA-waived tests are performed in physician office laboratories, but they may also be done in hospital or reference laboratories. Table 2-2 lists the types of tests that are CLIA-waived at this time. This list has grown significantly since CLIA '88 first went into effect, and will continue to change.
- **Moderate-complexity tests:** These tests are more complex than CLIA-waived tests. The complexity designation is based on a grading scale that takes into account the difficulty of test performance, the maintenance and troubleshooting required for the instrument used for the test, the amount of knowledge necessary to correctly interpret the results, and other specific aspects of the testing process. Some physician office laboratories perform moderate-complexity testing, but there are restrictions on the personnel performing the tests. Employees must be properly trained in these more advanced concepts in order to meet the requirements of the regulation. These tests also have more requirements for quality control and quality assurance documentation, and laboratories are required to participate in proficiency-testing programs to validate their test results. **Provider-performed microscopy procedures (PPMPs)** are also included in this classification, which concern the microscopic examination of **labile** specimens in the office performed by licensed health-care providers. Labile specimens are those that must be examined shortly

TABLE 2-2

CLIA-waived tests

Test Type	Test Name	CPT Code
Chemistry	Blood glucose by glucose monitoring devices cleared by the FDA for home use	82962
Urine and stool testing	Fecal occult blood	82270 GO 107
	Dipstick or tablet reagent urinalysis—nonautomated; for bilirubin, glucose, hemoglobin, ketone, leukocytes, nitrite, pH, protein, specific gravity, and urobilinogen	81002
	Urine pregnancy tests by visual color comparison	81025
	Ovulation tests by visual color comparison for human luteinizing hormone	84830
Hematology	Hemoglobin by copper sulfate—nonautomated	83026
	Erythrocyte sedimentation rate—nonautomated	85651
	Blood count; spun microhematocrit	85013

Note: This chart recognizes only the original CPT codes and test names listed as waived tests in 1988. There are numerous tests and testing methods that have been added to this list, which can be accessed at http://www.fda.gov/cdrh/clia/cliawaived.html after collection because the appearance of the specimen will change rapidly after collection. Provider-performed microscopy procedures may include urine microscopic examinations, semen analysis for the presence or absence of sperm, nasal smear examinations, and other types of analysis that need to be performed on fresh specimens.

- **High-complexity tests:** This category of testing is performed almost exclusively in hospital and reference laboratories by highly trained laboratory professionals. These tests require substantial knowledge of the significance of the test results, complex testing procedures, and stringent requirements for calibration of instruments and quality control procedures to ensure the accuracy of the results.

Laboratories may perform multiple levels of testing; some may perform only CLIA-waived tests, whereas others may perform high-complexity tests as well as waived tests. All laboratories must register and pay a fee to CMS that is based on the complexity of the tests that they perform. Failure to register or follow all the regulations associated with the specified level of testing may result in significant fines (thousands of dollars per day of noncompliance) and potential closure of the laboratory.

Test Your Knowledge 2-6

Describe the minimum level of education required to perform the following types of laboratory testing:

a. CLIA-waived tests
b. Moderate-complexity tests
c. High-complexity tests (Outcome 2-7)

Test Your Knowledge 2-7

What factors are taken into account when assigning the different levels of complexity to the various laboratory tests? (Outcome 2-7)

Employee Qualifications for Performance of CLIA Testing

The CLIA '88 regulations are very specific about the training necessary to perform the different levels of testing that are identified in the amendment:

- **CLIA-waived testing:** Minimum personnel requirements include a high school diploma or a GED, and thorough, documented training for the processes performed by the individual in the laboratory. Those performing CLIA-waived testing must follow the manufacturer's directions provided with the kit used exactly as they are written, and appropriate training for all individuals performing the laboratory test must be documented. These tests may be performed by phlebotomists, medical assistants, or those with more laboratory education.
- **Moderate-complexity testing:** The minimum education requirements are very similar to those of the CLIA-waived category. Moderate-complexity tests are more complicated to perform, and they may have much more clinical significance if they are performed or interpreted incorrectly, so the training must be comprehensive if the individual performing the test has limited formal laboratory education. This additional documented training must include numerous areas of the laboratory, such as quality control testing procedures, maintenance of the testing instruments, clinical significance of the tests, and validity of the test results. In most circumstances, this level of testing is performed by a medical laboratory technician or medical technologist, but it is possible to appropriately train a medical assistant to perform at this level as well.
- **High-complexity testing:** In order to perform tests in the clinical laboratory that have been established as high complexity by the CLIA standards, the laboratory professional must possess at least an associate degree in laboratory science. He or she must also have documented training on site for quality control procedures, maintenance and troubleshooting of the instruments used for the procedures, and more in-depth knowledge of the test parameters. These high-complexity tests usually comprise numerous performance steps, and the results have a high level of clinical significance. High-complexity tests are performed only in larger laboratories such as hospital and reference laboratories, and generally are performed by clinical or medical laboratory technicians, clinical laboratory scientists, or medical technologists.

Test Your Knowledge 2-8

What levels of laboratory testing may a clinical laboratory scientist with 4 years of education perform? (Outcome 2-8)

Oversight of CLIA Laboratories

Various agencies are involved with the enforcement of the CLIA '88 regulations. These include the FDA, the CMS, and the **U.S. Department of Health and**

Human Services (HHS). The HHS has the overall responsibility for laboratory quality assurance as designated in CLIA '88, but it has delegated various parts of the process to other agencies. The FDA is responsible for determination of the levels of complexity for the various laboratory tests. After the level of complexity has been determined, laboratories that perform these tests must apply for the appropriate certificate that allows them to perform these tests. The laboratory may qualify only for certain testing procedures based on the qualifications of their personnel. They apply to CMS for this certificate, whether it is a **Certificate of Waiver** (allowing them to perform CLIA-waived testing only) or another type of certificate that allows them to perform tests of higher complexity. A laboratory can only begin testing after receiving CMS certification. Compliance is monitored by the CMS or, in some situations, by state or private agencies that have regulations that are at least as stringent as those of the CMS, and have an agreement with CMS to monitor compliance. Some laboratories are affiliated with **COLA,** an independent company that accredits laboratories. COLA uses various educational methods to help laboratories meet the requirements for CLIA compliance, and CMS recognizes COLA as an accrediting agency. If a laboratory retains its accreditation with COLA, it will be assumed that that laboratory is meeting the CLIA requirements and will be granted the certificate that is appropriate for the level of complexity tested without site inspection requirements from the CMS. COLA is most commonly associated with physician office laboratories, but the organization is recognized by larger laboratories for accreditation purposes as well.

The CLIA certificate types are based on the complexity of the tests performed. Here is a summary of the available certificates for laboratories:

- **Certificate of Waiver:** Allows laboratories to perform only CLIA-waived tests.
- **Certificate of PPM:** Allows the licensed health-care providers in an organization to perform designated microscopic examinations.
- **Certificate of Compliance or Certificate of Accreditation:** Allows the laboratory to perform CLIA-waived tests, PPMPs, moderate-complexity tests, and high-complexity tests. These certificates are granted after the laboratory has been inspected. If the inspection was performed by a governmental agency, the laboratory receives a Certificate of Compliance. If the inspection was performed by a private accreditation company, the laboratory receives a Certificate of Accreditation.

Test Your Knowledge 2-9

Which federal agency assigns the CLIA categories to laboratory tests? (Outcome 2-9)

SUMMARY

Within a typical laboratory setting you will find professionals with various credentials and educational levels. Those who work in a hospital or reference laboratory may have different qualifications than those who work in a smaller physician office laboratory. There are regulations that dictate who can perform which type of testing, and what training and supervision must be involved for each test. CLIA '88, applies to all laboratories in the United States that perform testing on human specimens, except for federal laboratories and those that perform only research testing. Every testing site must register and pay a fee to comply. This regulation has established three levels of testing: CLIA-waived tests, moderate-complexity tests, and high-complexity tests. The levels have different requirements for testing personnel, quality control procedures, instrument maintenance, and laboratory supervision. The FDA makes the decision about which complexity level each test will have. Medical assistants with appropriate additional training may perform CLIA-waived or moderate-complexity tests, but those who perform high-complexity testing must have more specialized laboratory training in order to be compliant. All laboratories must register with CMS to legally perform laboratory testing at any level.

TIME TO REVIEW

1. A labile substance is one that: Outcome 2-1
   a. Is not fixed and is easily destroyed
   b. Is not fixed but can be stored for an extended period of time
   c. Is fixed with some sort of preservative
   d. Can easily be transferred to a reference laboratory
2. True or False: Ambulatory care refers to the care provided to those transported to the hospital as an emergency transport. Outcome 2-1

3. Which laboratory settings may employ a medical assistant? Outcome 2-2
   - a. POL
   - b. Hospital laboratory
   - c. Reference laboratory
   - d. None of the above
   - e. a, b, and c
4. True or False: Pathologists serve as directors in reference laboratory settings. Outcome 2-3
5. A laboratory professional with an associate degree in laboratory science may work in which laboratory setting? Outcome 2-4
   - a. Reference laboratory
   - b. Hospital laboratory
   - c. Physician office laboratory
   - d. None of the above
   - e. a, b, and c
6. List three duties that may be performed by a medical assistant in a hospital or reference laboratory. Outcome 2-5
7. CLIA '88 stands for: Outcome 2-6
   - a. Clinical Laboratory Incident Act #88
   - b. Clinical Laboratory Improvement Amendment of 1988
   - c. Consistent Laboratory Improvement Amendment of 1988
   - d. Continual Linear Improvement Act of 1988
8. Choose the types of laboratory testing that may be performed by a medical assistant. Outcome 2-7
   - a. CLIA-waived testing
   - b. Moderate-complexity testing
   - c. Provider-performed microscopy procedures
   - d. High-complexity testing
9. True or False: Those performing CLIA-waived testing do not need to follow the manufacturer's directions; they may write their own procedures as needed. Outcome 2-8
10. What role does the Centers for Medicare & Medicaid Services play in CLIA '88 enforcement? Outcome 2-9
   - a. It is not involved
   - b. It inspects each laboratory
   - c. It handles all the laboratory registration procedures
   - d. It determines the level of complexity for all laboratory tests

Case Study 2-1: CLIA quiz

Rose is starting at a new job in a physician office laboratory today. As part of her orientation, the physician who is in charge of the laboratory asks her a few questions to test her understanding of the laboratory organization. He asks her to answer the following questions about CLIA '88. How should she answer these?

1. Was CLIA '88 created to protect employees or patients?
2. Is it still necessary to be trained in the performance of CLIA-waived test procedures, or can these be performed without documented training?
3. Do laboratories have to formally registered in order to legally perform testing on human specimens?

RESOURCES AND SUGGESTED READINGS

American Medical Technologists
Certifying agency for laboratory professionals
http://www.amt1.com
American Society for Clinical Pathology
Membership and certification information for laboratory professionals, as well as continuing education information
http://www.ascp.org
Centers for Disease Control and Prevention CLIA-related publications from the Federal Register and the Code of Federal Regulations
Excellent reference for the actual CLIA regulations presented in a time line that is easy to follow
http://wwwn.cdc.gov/clia/chronol.aspx
Centers for Disease Control and Prevention CLIA Subpart A General Provisions
Provides various details about CLIA '88, including compliance requirements
http://wwwn.cdc.gov/clia/regs/subpart_a.asp#493.1
Tests waived by the FDA from January 2000 to present
List of all CLIA-waived tests. Updated regularly.
http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfClia/testswaived.cfm
Information about the COLA accreditation services and education products. Also includes online education programs that may be used to fulfill continuing education requirements.
http://www.cola.org

Chapter 3

Laboratory Safety and Preventing the Spread of Disease

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Infection Control and Laboratory Safety
- Core Concepts of Infection Control
  - Microorganisms
    - *Types of Microorganisms*
    - *Microorganism Growth Requirements*
    - *Medical Asepsis*
  - Chain of Infection
  - Standard Precautions
    - *Centers for Disease Control and Prevention Hand-Washing Recommendations*
    - *Acceptable Medical Hand-Washing Procedures*
    - *Proper Use of Personal Protective Equipment*
- Laboratory Safety
  - Chemical Safety
    - *Physical Safety*
    - *Fire Safety*
    - *Electrical Safety*
    - *Body Mechanics*
  - Bloodborne Pathogen Safety
    - *Universal Precautions*
    - *Bloodborne Pathogens Standard*
- Diseases Caused by Bloodborne Pathogens in the Laboratory Setting
  - Hepatitis
  - Human Immunodeficiency Virus
- Postexposure Follow-Up Procedure
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

3-1 Define the key terms.

3-2 List the major types of infectious agents.

3-3 Restate the difference between pathogenic and nonpathogenic microorganisms

3-4 Describe the various shapes of bacteria presented in the text.

3-5 Compare and contrast bacteria and viruses.

3-6 Describe medical asepsis.

3-7 Explain what the chain of infection concept refers to, and describe how the chain may be broken.

3-8 Explain how the CDC Standard Precautions are used in a laboratory setting.

3-9 Analyze the importance of proper hand-washing procedures and appropriate use of personal protective equipment.

3-10 Explain appropriate procedures for hand sanitization for health-care workers.

3-11 Examine the fundamental concepts included in the OSHA Hazard Communications Standard.

3-12 List the required components on a Material Safety Data Sheet.

3-13 Explain how a chemical label provides safety information.

3-14 Describe how a laboratory employee may protect themselves from other physical dangers in the laboratory.

3-15 Identify who is protected by the OSHA Bloodborne Pathogens Standard.

3-16 Interpret the key terms included in the OSHA Bloodborne Pathogens Standard.

3-17 List the essential components of an exposure control plan.

3-18 Discuss the appropriate use and disposal of sharps in the laboratory environment.

3-19 Define biohazardous waste and explain proper disposal methods for this type of laboratory waste.

3-20 Compare and contrast the major bloodborne pathogens that are considered to be a threat in the laboratory environment.

3-21 Detail the appropriate follow-up procedure in case of an accidental bloodborne pathogens exposure.

CAAHEP/ABHES STANDARDS

CAAHEP 2008 Standards

III.P.4. Perform Handwashing
III.C.III.3. Discuss Infection Control Procedures
III.C.III.5. List major types of infectious agents
III.P.III.2. Practice Standard Precautions
X.C.XI.1: Describe Personal Protective Equipment
X.C.XI.3: Describe the importance of MSDS in a healthcare setting
X.C.XI.4. Identify safety signs, symbols and labels
X.C.XI.8. Discuss Fire safety issues in a healthcare environment

ABHES 2010 Standards

- Clinical: Apply principles of aseptic techniques and infection control
- Clinical: Use standard precautions

KEY TERMS

Aerobic
Acquired immune deficiency syndrome (AIDS)
Anaerobic
Asepsis
Asymptomatic
Bacilli
Bacteria
Biohazard symbol
Biohazardous waste
Bloodborne pathogen (BBP)
Bloodborne Pathogens Standard (1910.1030)
Carcinogenic
Carriers
CD4 cells
Centers for Disease Control and Prevention (CDC)
Cirrhosis
Cocci
Contagious
Contaminated sharps
Diplococci
Disinfection
Electron microscope
Engineering controls
Epidemiology
Exposure control plan
Fecal-oral route
Fomites
Fungi
Hazard Communication Standard
Health care–associated infection
Hepatitis
Hepatitis A (HAV)
Hepatitis B (HBV)
Hepatitis C (HCV)
Hepatitis D
Human immunodeficiency virus (HIV)
Infection
Infection control

Infectious
Material Safety Data Sheets (MSDS)
Medical asepsis
Microorganisms
Mucous membranes
Mycotic
National Fire Protection Association (NFPA)
Nonintact skin
Normal flora
Nosocomial infection
Occupational exposure
Occupational Safety and Health Administration (OSHA)
Opportunistic pathogen
Other potentially infectious materials (OPIMs)
Parasites
Parenteral exposure
Pathogens
Percutaneous
Personal protective equipment (PPE)
Postexposure prophylaxis (PEP)
Protozoa
RACE
Regulated waste
Resident bacteria
Sanitization
Spirilla
Spores
Standard Precautions
Staphylococci
Sterilize
Streptococci
Surgical asepsis
Susceptible
Transient bacteria
Transmissible
Universal Precautions
Vector
Viruses
Window period
Work practice controls

INFECTION CONTROL AND LABORATORY SAFETY

Pathogens, infection, and *contamination* are words that most of us have heard at home or in our communities. What do they really refer to? What impact may they have on our laboratory environment, our health, and the health of our patients? The laboratory setting has stringent safety precautions that must be observed in order to keep our health-care workers safe as they perform their daily routines. Laboratory professionals use various chemicals and equipment that put them at risk for injury. In addition, specimens taken from the human body present a unique challenge as they are naturally **infectious**, or capable of transmitting disease to others. **Infection control** is the term used to describe the process of protecting health-care workers and the patients they serve from the infectious agents in our facilities. In addition, because all the hazards faced in the laboratory are not from infectious agents, further safety methods must be employed to protect our laboratory professionals. It is imperative that as a medical assistant working in a laboratory environment you understand the hazards you may encounter, the laws that are designed to protect you and the patients you serve, and the appropriate use of the safety equipment that is available to you.

CORE CONCEPTS OF INFECTION CONTROL

In recent years, we all have become more aware of the infections that may be transmitted through the water, through the air, or with casual contact in our communities. Severe acute respiratory syndrome (SARS), the Asian flu, *Escherichia coli* outbreaks, and other pathogens such as the monkeypox virus remind us that we are very vulnerable. The presence of AIDS and hepatitis C is further proof of how our world has changed in the past 35 years. Be very mindful of the fact that almost any transmissible infection could find its way into your laboratory at any moment, as it may be present in your community, and you won't be aware of it. An ill patient may come into the office for medical services with an infectious condition, a specimen may be dropped off that is capable of transmission, or a salesperson may arrive and offer her hand, which is covered with bacteria, for a greeting.

Microorganisms

Diseases that may be encountered in your medical office or laboratory (even without any blood exposure) include varicella (chickenpox), tuberculosis, viral respiratory infections, conjunctivitis (pinkeye), gastrointestinal infections, and measles. Sometimes these infections affect the personnel, and sometimes they affect other patients within the facility. They are all caused by **microorganisms,** living organisms that are too small to be seen without a microscope. Most microorganisms are harmless to humans. However, a small percentage of **bacteria**, **viruses, fungi, protozoa,** and **parasites** are capable of causing disease in the human body. These disease-causing organisms are known as **pathogens.** An **infection** is the invasion of the body by pathogens that then cause disease symptoms. Not all microorganisms are pathogenic, and some are only capable of causing disease when they enter a part of the body where they don't normally reside. A disease is **contagious** or **transmissible** if it can be spread to other people directly or indirectly.

Test Your Knowledge 3-1

Staphylococcus aureus is a type of bacteria that sometimes functions as a pathogen. What other types of microorganisms may cause disease in humans? (Outcome 3-2)

Test Your Knowledge 3-2

Do all microorganisms cause disease? (Outcome 3-3)

Types of Microorganisms

Bacteria. Bacteria are single-celled organisms that have a cell wall in addition to the cell membrane that our human cells possess. (This is an important property when identifying different types of bacteria with special stains, as is discussed in Chapter 10.) Of all the bacteria known to exist, approximately 4% are known to cause disease in humans. Bacterial infections are treated with medication that will kill the microorganisms or keep them from multiplying, (antibiotics), but for effective treatment, identification of the causative agent is often necessary.

An initial step of the identification process is to examine a sample under the microscope so that the bacteria can be classified according to one of three basic shapes, as seen in Figure 3-1. **Cocci** are round bacteria, which can then be classified further by their appearance when examined microscopically. Cocci that grow in grape-like clusters are **staphylococci,** those that grow in chains are **streptococci,** and those that grow in pairs are **diplococci.** Diseases caused by cocci include streptococcal sore throat, pneumonia, abscesses, food poisoning, gonorrhea, and meningitis.

Long, slender, rod-shaped (oval) bacteria are called **bacilli.** These are especially prevalent in the soil and the air. Many types of bacilli are able to form **spores,** which is a dormant form of the bacteria that is resistant to changes in heat, moisture, and disinfectants. Bacilli cause diseases such as botulism, tetanus, diphtheria, tuberculosis, and salmonella food poisoning. *E. coli* is a normal bacilli that is present on our skin and in our intestines, but if we come in contact with a specific strain, *E. coli* O157:H7, it may lead to serious food poisoning, and may even be fatal.

Spirilla are curve-shaped or spiral bacteria. This type of bacterium is less frequently isolated in specimens from the human body, but the infections may be quite serious when they occur. *Treponema pallidum* is a spirillum that causes syphilis, and cholera is caused by another type of spirillum. This category may also be described as curved rods and microbiologists may subdivide it further into those bacteria that only have a slight curve and those that are tightly wound like a spring.

Test Your Knowledge 3-3

How are staphylococci and diplococci different? (Outcome 3-4)

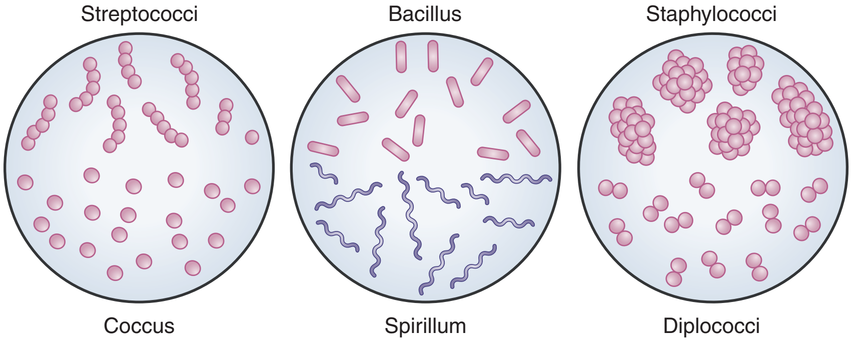

Figure 3-1 Types of bacteria, including cocci, staphylococci, streptococci, diplococci, bacilli, and streptococci.

Test Your Knowledge 3-4

Describe the appearance of bacilli viewed under the microscope. (Outcome 3-4)

Viruses. Viruses are another common type of pathogenic microorganism. They are the smallest infectious agent, and are not really cells. Viruses are either made up of RNA or DNA surrounded by a protein capsule, and require a host cell to survive and replicate. Viruses cannot be observed using a simple microscope such as is found in most laboratories; they require use of an **electron microscope** for visualization. They invade the cells of our body, and use our own structures to make more virus particles. Viruses cause many of the diseases for which we are vaccinated, including measles, mumps, and chickenpox. The **human immunodeficiency virus (HIV)** is the causative agent of AIDS, and is an example of a retrovirus. Antiviral medications are available for some viral infections, but because of the way virus particles use human cells, the treatment may be quite damaging to the host as well as to the virus. Antibiotics are not effective against viruses.

Test Your Knowledge 3-5

List two ways that viruses are different from bacteria. (Outcome 3-5)

Fungi. Fungi are plant-like organisms that flourish in an environment that is dark and damp. Yeast is a type of fungus, and fungal infections on the body are called **mycotic** infections. Athlete's foot and ringworm are examples of fungal infections.

Parasites. Parasites are similar to viruses because they require a living host to survive. A genus of bacteria known as *Rickettsia* is parasitic. Parasites take their nourishment from the host and require their host cells to reproduce, but they do not utilize the human cell in the same way that a virus does. Malaria is an example of a parasitic disease; the malaria parasite actually works into the red blood cells of the host and causes the patient to become ill. The word *parasite* may also be used to refer to larger, multicellular organisms such as tapeworms, which live inside the human host and use the body to survive.

Protozoa. Protozoa are complex single-cell microorganisms, most of which are nonpathogenic. However, there are a few species that can cause very serious infections in humans. Protozoa live in the soil and water.

Microorganism Growth Requirements

Because we have so many microorganisms present in our environment, we must find ways to eliminate as many as possible, with an emphasis on the removal of pathogens that could cause harm. However, it is impossible and impractical to **sterilize** our environment in the laboratory, which would involve the elimination of all microorganisms. **Sanitization** reduces the number of microorganisms on a surface with cleaning. Heat or chemicals are often used for sanitization in the health-care environment. **Disinfection** is the process by which a medical assistant applies a chemical to a surface to kill the pathogenic microorganisms that may be present. Sanitization is often performed before an item is disinfected, so that the majority of the microorganisms are already removed before the pathogens are targeted. Items such as examination tables, countertops, and equipment are disinfected on a regular basis. Skin may also be disinfected, as it is impossible to sterilize the skin by eliminating all the microorganisms. Seventy percent isopropyl alcohol or povidone-iodine solutions are used for skin disinfection. Work surfaces and equipment are often disinfected in the laboratory with a freshly prepared 10% bleach solution.

Asepsis means that a surface is without infection. In the medical environment, there are two types of asepsis. **Surgical asepsis** means that all the pathogenic organisms have been destroyed before they enter the body. Invasive procedures such as venipuncture, injections, and urinary catheterization require surgical asepsis to be in place so that pathogenic microorganisms aren't introduced into the body of the patient. Equipment used for these types of procedures must be sterilized, and special care is used to avoid infection. **Medical asepsis** is a term used to describe a procedure or an environment that allows a patient to be treated without exposure to pathogenic microorganisms. For noninvasive procedures and the majority of care provided in the medical office, medical asepsis is adequate to protect the patients from potential infection. The items that are used for this type of care are clean and have been disinfected, but sterility is not necessary. To disinfect the environment appropriately and achieve medical asepsis, we must understand what elements microorganisms need for survival. These include the following:

- **Temperature:** The optimum temperature for each microorganism will vary. Human pathogens tend to prefer our body temperature, so they grow best at approximately 98.6°F.
- **pH:** A neutral pH is best suited for most microorganisms. This is why many disinfectants and cleaning

agents are basic or acidic, as they will kill the microorganisms by making their environment inhospitable.

- **Darkness and moisture:** Most microorganisms like darkness or dim light, and they all need moisture to survive.
- **Nutrition:** The type of nutrition varies depending on the type of microorganism, but they all need something in their environment to use as a food source.
- **Oxygen: Aerobic** microorganisms need oxygen present for survival. **Anaerobic** microorganisms are best suited for environments that have an absence or low levels of oxygen.

Good aseptic practices for the medical setting will attempt to eliminate the elements necessary for pathogenic microorganisms to survive. These practices may include the following:

- Disinfect work areas (using an appropriate disinfecting agent for the health-care environment) between patients.
- Take appropriate respiratory precautions while working with patients with potentially infectious respiratory conditions. Provide masks and tissues to the patients, and enforce their use by employees as well.
- Close the door of the treatment or blood-draw area if possible when an infectious disease is suspected.
- Limit access of nonessential personnel and visitors to patient care areas.
- Keep the laboratory and waiting room area well lit, well ventilated, and free of dirt and dust.

Test Your Knowledge 3-6

Brittany, a phlebotomist working in the laboratory, is cleaning up at the end of her shift. She uses the disinfectant provided by her employer and some paper towels to thoroughly clean the area where she has been drawing blood from patients. Is this work area now sterile? (Outcome 3-6)

Chain of Infection

Figure 3-2 is a representation of the way pathogens are transmitted from person to person. This is represented as a chain, because all the parts of the circle are sequential and linked to one another. To stop the transmission of disease, the chain must be broken. A medical assistant working in a laboratory environment must practice good infection control techniques to achieve medical asepsis and break the chain. The potential infection needs to be stopped at the source before it can be transmitted. The essential parts of the chain include the following:

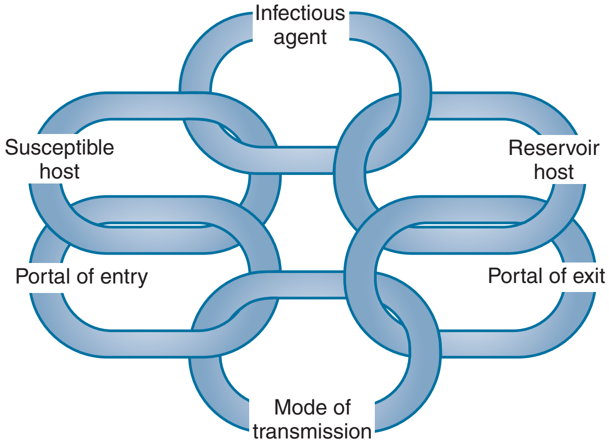

Figure 3-2 Chain of infection.

- **Infectious agent:** A pathogenic microorganism. This may be a virus, bacteria in the environment, or pathogens that are carried in the bloodstream. There may also be microorganisms that are only pathogenic in specific situations. For instance, if an individual is taking antibiotics for an extended period of time, she may develop a yeast infection, because the yeast is taking advantage of the body's imbalance created with the antibiotic use. The yeast would be an **opportunistic pathogen.**
- **Reservoir host:** Someone who is infected or "carrying" the infectious agent. This may be a person or an animal that may or may not have symptoms of the infection. A reservoir host provides the necessary environment for the pathogen to grow.
- **Portal of exit:** The means by which the pathogen leaves the host body, either through the eyes, mouth, ears, intestinal tract, urinary tract, respiratory tract, reproductive tract, or broken skin. Another portal of exit may be through the blood or other body fluids capable of transmitting pathogens.
- **Mode of transmission:** This describes how the pathogen moves from one person to another. It may be transmission through the air as moisture droplets after a sneeze or a cough, other direct transmission occurs as one person touches another or via an insect or other **vector.** Vectors are living organisms that take in the pathogen, allow it to live and multiply in or on their bodies, then transmit it to another host without becoming ill with the pathogen during the transport. Dirty hands are a very common mode of transmission.

Pathogens may also be transmitted by inanimate objects, known as **fomites.** Doorknobs, telephones, countertops, and writing instruments are common fomites in the laboratory environment. Contaminated food or water may also function as a mode of transmission.

- **Portal of entry:** The pathogens enter the body in the same ways that they leave the body. The **mucous membranes** that line the external openings of our bodies are a common portal of entry and exit for pathogenic microorganisms.
- **Susceptible host:** A susceptible host is not protected from the pathogen as it enters his or her body. The pathogen to grow and multiply within the susceptible host. Age and illness can affect the susceptibility of a patient or an employee. A potential host will protect him- or herself (become less susceptible) in various ways. Protection may include vaccinations against certain viruses or bacteria, appropriate nutrition, or proper use of personal protective equipment.

In addition to understanding the chain of infection, it is important to realize how the chain may be broken by health-care staff members. Appropriate infection control practices and patient education are effective means of breaking the chain. Specifics include the following:

- **Infectious agent:** Although health-care professionals may not always be able to break the chain at this link, it is important to remember that an infectious agent may be eradicated if a patient takes all of his or her antibiotics as directed. Patient education in this area is critical.
- **Reservoir host:** Health-care personnel are sometimes tempted to go to work when they are ill. They are essentially a reservoir host at this time, and if they are not in the facility when they are ill, the chain is broken. In addition, if there is a patient in the waiting room with a suspected illness that is highly contagious, removing the patient from the vicinity of other patients may break the chain at this point.
- **Portal of exit:** Providing respiratory protection and tissues to patients may contain the infectious agent as it leaves the body. Keeping infected wounds covered may also affect this link of the chain.
- **Mode of transmission:** The simplest and most important action that health-care personnel can take to break the chain of infection is washing their hands. In addition, contaminated surfaces should be cleaned immediately after care is provided, and often throughout the workday. Patients should be educated about how to protect themselves if they live in the same household with another family member who is ill. Hand sanitization is also critical for family members and patients to eliminate an opportunity for the pathogens to be transported to another susceptible host.
- **Portal of entry:** Those who are working with patient specimens must wear appropriate protective equipment (such as a face shield to provide mucous membrane protection, gloves, and a laboratory coat) to protect the common portals of entry. Health-care personnel working with patients who have respiratory symptoms should protect their own respiratory passages with a mask. All broken skin should be covered when working with patients, and there should be no eating, drinking, chewing gum, or applying makeup in the area where specimens are collected or processed.
- **Susceptible host:** Health-care personnel should always keep their vaccinations up to date to keep them from being susceptible to vaccine-preventable diseases. In addition, a balanced lifestyle with good nutrition and adequate rest will reduce susceptibility. Patients should be encouraged to keep vaccinations up to date, and those who have impaired immune systems should be cautioned about protecting themselves while in the community.

Test Your Knowledge 3-7

How might a medical assistant break the chain of infection at the mode of transmission link? (Outcome 3-7)

Test Your Knowledge 3-8

Cally Jones just completed a set of vaccinations to protect her from hepatitis A. Where has she broken the chain of infection for this disease? (Outcome 3-7)

Standard Precautions

We have discovered that there are numerous microorganisms in our environment that have specific requirements they need to survive. Some of these microorganisms are pathogens, meaning that they can cause infection in the human body. A **health care–associated infection** is one that is acquired in a health-care facility, such as a hospital, long-term care facility, physician office, or laboratory. If a patient in an inpatient facility enters without any evidence of infection, and develops signs of one more than 48 hours after being admitted, the infection is then investigated as a health care–associated infection. This type of infection may also be known as a **nosocomial infection**. Health care–associated infections are more common in inpatient facilities in which patients are in close proximity to one another, but it is also possible to acquire an infection in an ambulatory care setting. For example, chickenpox is highly contagious and can be spread from one person to another via droplets in the air after an infected individual coughs or sneezes. If a susceptible patient is in the same waiting area for an extended period of time with someone who is infected, the patient may develop chickenpox as a health care–associated infection.

Health care–associated infections have become more prevalent, more dangerous, and more expensive to treat in the past few decades. In response, the **Centers for Disease Control and Prevention (CDC)** has developed a set of **Standard Precautions** to assist health-care facilities with their infection control efforts. This standard is based on the premise that every person is potentially infectious with a microorganism that could be transferred to someone else in the health-care setting. Appropriate hand-washing techniques are stressed in the standard, as it has been found that this is the most effective action taken by medical staff to stop the spread of infection. Appropriate use of personal protective equipment, equipment disinfection, good respiratory hygiene, appropriate use of sharps, correct disposal of linen, and environmental cleaning practices are also addressed in the standard. For those working in inpatient facilities, the standard includes information about different types of precautions to be taken in isolation situations. The Standard Precautions expand and enhance the **Universal Precautions** developed in the 1980s, which specifically addressed infections transmitted through contact with blood and other potentially infectious materials. The premise of the Universal Precautions was that everyone was potentially infectious for bloodborne pathogens and that the same care should be used to treat every specimen.

Test Your Knowledge 3-9

According to the CDC recommendations for Standard Precautions, who is to be considered infectious?
(Outcome 3-8)

Test Your Knowledge 3-10

What is the most effective action a health-care worker can take to stop the chain of infection? (Outcome 3-8)

Centers for Disease Control and Prevention Hand-Washing Recommendations

Even though research has shown that the hands of health-care personnel are one of the most common modes of transmission for pathogens, there are still issues with hand-washing compliance. The CDC recognized that this issue needed to be addressed, and the guidelines recently released now recommend the use of alcohol-based hand rub for routine decontamination of the hands in many situations, which may be quicker and easier for health-care professionals to use than washing the hands with soap and water. Either hand washing and decontamination using alcohol-based hand rubs remove or kill **transient bacteria,** viruses, or other types of microorganisms that may be present on the hands. Transient microorganisms, which attach themselves to our hands during our day-to-day activities, are responsible for most of the contamination and potential infection as health-care workers move from patient to patient. **Resident bacteria,** or **normal flora,** are bacteria living on and in between the deeper layer of our skin cells, and are not removed during routine hand cleaning. Here are some of the specific recommendations for hand hygiene:

1. Avoid any unnecessary touching of surfaces surrounding the patient.
2. If your hands are visibly dirty or contaminated with blood or other materials, wash them with soap and water. (It is not important at this point whether it is an antimicrobial soap that is used.)
3. Hands should be washed before eating and after using the restroom.
4. If hands are not visibly soiled, or if they were washed with a soap that was not antimicrobial, it is recommended that an alcohol-based hand rub be used. It may be used instead of the soap and water wash, or in addition to hand washing after a nonantimicrobial soap is used.
5. Hands should be decontaminated (washed or cleansed using the alcohol-based hand rub) in these situations:
   a. Before and after direct contact with patients, even if the patient has intact skin
   b. After contact with blood, body fluids or excretions, mucous membranes, nonintact skin, or dressings used for open wounds
   c. If the health-care provider will be going from a contaminated body site to a clean body site on the same patient
   d. After contact with any objects, such as equipment, in the immediate area of the patient
   e. After removing gloves

5. It is also recommended that artificial nails be avoided if the duties of the health-care provider include direct contact with patients at high risk for infection. Keep natural nails less than one-fourth-inch long.
6. Minimize jewelry on the hands, as bacteria may be present underneath or in jewelry and is not removed during normal hand-washing techniques.
7. Refilling of soap pump dispensers is discouraged. If a refill is necessary, the dispenser should be rinsed thoroughly before more soap is added. Bacterial contamination may be present in the small amount of soap at the bottom of the container.

Test Your Knowledge 3-11

What is an acceptable alternative to hand washing in the laboratory in most situations? (Outcome 3-9)

Test Your Knowledge 3-12

Describe three situations in which a health-care worker should decontaminate their hands. (Outcome 3-10)

Acceptable Medical Hand-Washing Procedures

The CDC recommendations outline specific practices to be employed when washing or disinfecting hands because this is such an important part of infection control practices. These procedures are explained in detail in Procedures 3-1 and 3-2.

Test Your Knowledge 3-13

Why would a clean, dry paper towel be used to turn off the water faucets after hand washing? (Outcome 3-10)

Procedure 3-1: Perform Hand Washing

TASK

Perform hand washing appropriately, using the recommended technique for medical personnel.

CAAHEP/ABHES STANDARDS

CAAHEP 2008 Standards

III.P.4. Perform Handwashing

ABHES 2010 Clinical

- Apply principles of aseptic techniques and infection control
- Use standard precautions

CONDITIONS

- Antibacterial soap
- Running water
- Paper towels
- Manicure brush and stick

Procedure	Rationale
1. Remove all rings, and make certain that lab coats, watches, and bracelets are pushed up above the wrist area.	It is not possible to clean appropriately under rings, so they should be removed. Watches and bracelets need to be moved above the wrists so that the wrists can be cleaned.
2. Adjust the water temperature until it is warm but not hot.	Hot water leads to skin breakdown, and should be avoided.
3. Wet the hands and wrists.	To provide the best sudsing action with the soap, the hands should be wet before applying the soap.
4. If using a soap dispenser, use a clean paper towel to dispense the amount of soap recommended by the manufacturer. This is usually 3–5 mL or 1–2 pumps. If an automatic soap dispenser is used, the paper towel is not necessary.	A push-top soap dispenser or an automatic dispenser is recommended. If bar soap must be used, the bar of soap must be kept in the hands until they are thoroughly covered with suds, and a drainable soap dish must be in use to avoid pooling of dirty water around the bar of soap.

Continued

Procedure 3-1: Perform Hand Washing—cont'd

Procedure	Rationale
5. Rub the hands together vigorously for at least 10–15 seconds. All surfaces of the hands should be covered with soapsuds, as well as between the fingers and up over the wrist area. Keep the hands lower than the elbows.	The friction created helps to minimize the amount of bacteria on the hands. The hands must remain lower than the elbows to avoid having the water run up to the elbows, causing contamination farther up on the arm. Hand washing with all surfaces covered with suds, including wrists
6. Avoid touching the lab coat or scrubs on the front of the sink, and avoid touching the hands to the inside of the sink while scrubbing.	If the lab coat or scrubs touches the front of the sink, it will become contaminated with dirty water. The inside of the sink is considered to be "dirty," so should not be touched during the hand-washing procedure.
7. Clean the top of the fingernails with the brush and clean under the fingernails with the manicure stick.	The cuticle area and the skin surrounding the fingernails harbor bacteria, so special care needs to be taken to clean this area thoroughly. The manicure stick should be used under the fingernails to remove as much contamination as possible.
8. Rinse the hands under the warm water, keeping the hands lower than the elbows.	The hands need to remain below the elbows to avoid contamination. All soap must be rinsed from the hands. Rinsing of hands with fingertips pointing downward.

Procedure	Rationale
9. Dry the hands thoroughly with paper towels. Dispose of the paper towels in an appropriate receptacle.	Hands should be dried thoroughly before continuing. Special care should be given to the area between fingers.
10. Using another clean, dry paper towel, turn off the water faucets. Dispose of the paper towel in the appropriate receptacle.	A clean, dry paper towel must be used rather than using the damp one that was used to dry the hands. The moisture on the paper towel may draw the bacteria present on the faucets back onto the clean hands.

Procedure 3-2: Sanitize Hands With an Alcohol-Based Hand Sanitizer

TASK

Sanitize hands using an alcohol-based hand sanitizer.

CAAHEP/ABHES STANDARDS

CAAHEP 2008

III.P.III.2. Practice Standard Precautions

ABHES 2010 Clinical

- Apply principles of aseptic techniques and infection control
- Use standard precautions

CONDITIONS

Alcohol-based hand sanitizer containing 60% to 95% alcohol

Procedure	Rationale
1. Remove all rings, and make certain that laboratory coats, watches, and bracelets are pushed up above the wrist area.	It is not possible to clean appropriately under rings, so they should be removed. Rings contain microorganisms that can cause infection. Watches and bracelets need to be moved above the wrists so that the wrists can be cleaned.
2. Verify that the sanitizer to be used is alcohol based.	The CDC recommendations sanction only the use of alcohol-based sanitizers.
3. Dispense the manufacturer's recommended amount of the product in the palm of the hand.	Do not dispense less than is recommended or the hand sanitization will be incomplete. Do not apply more than is recommended or the dry time will be increased.
4. Cover all surfaces of the hands and fingers with the sanitizer solution, and continue to rub these surfaces until the product is completely dry. Don't forget the areas around and under the fingernails, as well as between the fingers and the wrist area.	The skin around the fingernails and cuticles harbors a lot of microorganisms, so pay close attention to this area.
5. Allow the hands to air-dry. Do not use a paper towel to dry hands.	If the hands are not allowed to air-dry, the sanitization process is incomplete. The microorganisms will continue to be destroyed while the solution dries on the hands.

Proper Use of Personal Protective Equipment

Later in this chapter you will learn about the **Occupational Safety and Health Administration (OSHA)** Bloodborne Pathogens Standard, which is a comprehensive policy addressing **bloodborne pathogen (BBP)** exposure in health-care settings. A key component of this standard, as well as the Standard Precautions, is proper use of **personal protective equipment (PPE).** Personal protective equipment helps to protect the employee from bloodborne pathogens (those pathogens that are transmitted via direct contact with blood and other infectious body fluids), and in addition it helps to protect them from the pathogens in their environment that are not bloodborne. In order to be effective, personal protective equipment must be worn at appropriate times, removed when not needed, and disposed of properly. When considering PPE for a task, the employee must decide what type of exposure is reasonably anticipated while performing that task. Policies must be in effect by the employer to guide these decisions for PPE use.

The most important personal protective equipment in the medical laboratory is properly fitting gloves (see Fig. 3-3). Employers are responsible for providing PPE to employees, and gloves are no exception. Most facilities no longer use latex gloves; latex allergies have become too widespread to continue with this practice. Latex alternatives (such as nitrile gloves) are readily available. Gloves should be worn when contact with blood, mucous membranes, or nonintact skin could be anticipated when performing a certain task. Gloves should also be worn when contact with **other potentially infectious materials (OPIMs)** is anticipated. OPIMs are those that are capable of transmitting bloodborne pathogens.

It is not necessary to wear gloves when touching a patient with intact (unbroken) skin for routine care, such as taking vital signs and helping a patient into the procedure chair. Hands should be washed before and after each glove use, and gloves should always be removed after caring for a patient. The same pair of gloves should never be used for the care of more than one patient. Do not wash gloves. Gloves should not be worn after patient care when leaving that immediate area, as they may be contaminated with unseen microorganisms that will be deposited on door handles, telephones, and the like. Also, do not touch your face or hair while caring for a patient and wearing gloves. Do not wear the same pair of gloves for a patient if you are going from a dirty body site to a clean body site; remove the gloves, wash your hands, and put on a clean pair. Remember, gloves cannot prevent a needlestick injury, but they can prevent a pathogen from entering your body through a small break in your skin under the glove, and they can protect you from transient microorganisms that you may be exposed to while working in the health-care environment.

Test Your Knowledge 3-14

Marni is working hard to successfully draw blood from a patient who has very fragile veins. As she works, the phone keeps ringing. She finally obtains a sample, and just as she completes the draw and has the patient put pressure on the site, she reaches out with her gloved hand and answers the phone. What has she done wrong in this scenario? (Outcome 3-9)

Gloves should be removed promptly after use. The method of removal is critical, as the gloves may be contaminated with unseen microorganisms that must not be spread to the surrounding area. Procedure 3-3 explains the steps involved in aseptic glove removal.

Personal protective equipment also may also include masks, goggles, face shields, or respirators (Fig. 3-3). A mask may be worn to protect others from droplets that may be generated when coughing or sneezing. Masks may be given to patients to wear while they are in the facility, or health-care professionals who have an upper respiratory infection or allergies may wear masks. Remember to keep the mask tight to the face so that it is effective. Masks with goggles or face shields may be worn to protect the employee's eyes, mouth, and nasal passages from moisture droplets or aerosols in the laboratory environment.

Respirators are required when working with patients with pulmonary tuberculosis or when dealing with certain types of specimens considered potentially infectious for airborne transmission. Respirators must also be worn when working with certain chemicals used as preservatives in the laboratory. Respirators must be fit properly to the face of the employee to be effective, and the filters must be maintained as directed.

Gowns may be used when additional protection is required during patient care (Fig. 3-3). These are especially important when entering isolation areas of the hospital or when working in a nursery environment with newborns. Care must be taken to keep the interior side of the gown free of contamination to protect the employee. Gowns are usually made of light, disposable material, and must be disposed of immediately after use. Generally

Procedure 3-3: Removal of Contaminated Gloves

TASK

Properly remove and dispose of contaminated gloves.

CAAHEP/ABHES STANDARDS

CAAHEP 2008

III.P.III.2. Practice Standard Precautions

ABHES 2010 Clinical

- Apply principles of aseptic techniques and infection control
- Use standard precautions

CONDITIONS

- Nonlatex gloves
- Biohazard waste container

Procedure	Rationale
1. Grasp the palm of the glove on the nondominant hand. Keep the gloves away from the body and the hands pointed toward the floor.	Grasping the palm will allow a firm hold as this glove is removed. Keeping the hands away from the body pointed toward the floor will minimize the risk of splatter in the eyes or mucous membranes.
2. Pulling on the palm of the glove, turn it inside out as it is removed from the nondominant hand.	If the glove is inside out, it will not be able to contaminate the bare skin, as the soiled area will be on the inside of the glove.
3. Crumple up the contaminated glove into the other gloved hand.	This allows the contaminated glove to be held safely while the other glove is removed.
4. Insert two of the ungloved fingers under the cuff, against the wrist of the gloved hand.	Be careful not to touch the contaminated side of the glove with the bare hand.
5. Pulling down with these fingers, turn the second glove inside out over the other glove while slipping it off the fingers.	This method will allow the contaminated surfaces to remain inside the glove bundle.
6. Dispose of the gloves in a biohazard waste container.	Visible contamination on the gloves means that they need to be disposed of as biohazardous waste.
7. Wash hands.	Hands must always be washed after removing gloves because gloves are not foolproof and hand contamination is still possible. It also helps to remove any powder residue that may be left behind on the hands.

there is a disposal area just inside or outside the room where the gowns were worn.

Laboratory coats are also used when there is a potential for splashing or soiling of the clothing worn by laboratory professionals. These should be OSHA approved as fluid resistant to offer the best protection. The coats must have tight cuffs, and should button or snap up to the neckline when worn. Clean laboratory coats should not be stored with dirty coats, and employees must not wear their contaminated laboratory coats into eating areas or the restroom. Laboratory employees are also not to take their coats home to be laundered, as this could contaminate their home environment. If laboratory coats are required at the facility where an employee works, OSHA regulations dictate that the facility is responsible for providing the coats and cleaning them commercially after use.

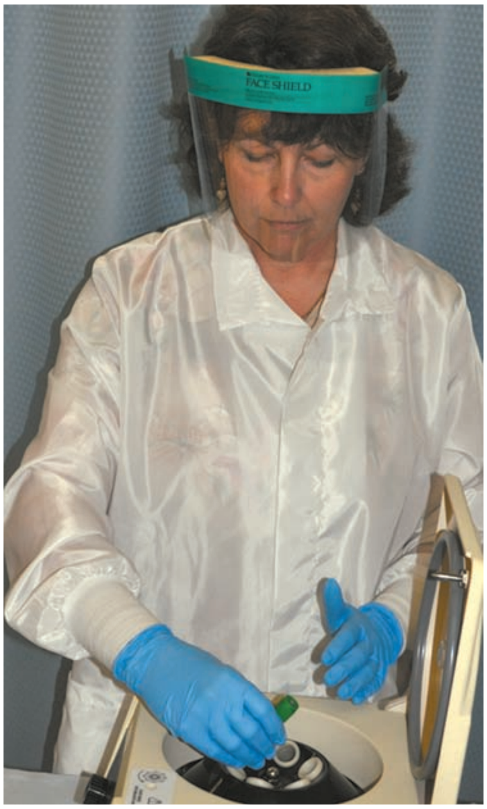

Figure 3-3 Medical assistant wearing OSHA-approved laboratory coat, appropriate face protection, and gloves.

Laboratory professionals should never apply makeup, chew gum, eat, or drink when working with blood or OPIMs. Also, food must never be stored in the same refrigerator with specimens or medications, regardless of how it is packaged.

Test Your Knowledge 3-15

What type of personal protective equipment should be worn when a medical assistant is performing a routine blood draw? (Outcome 3-9)

POINT OF INTEREST 3-1

Bloodborne facts

The U.S. Department of Labor, Occupational Safety, and Health Administration (OSHA) has published a series of fact sheets that cover various aspects of the Bloodborne Pathogens Standard. This standard may be difficult to understand and therefore compliance may be complicated. These fact sheets are excellent tools for offices that are struggling to develop their plans appropriately or for anyone who is trying to update knowledge of the subject. They cover many subjects, including the hepatitis B virus, vaccine, and postexposure follow-up, and proper use of personal protective equipment. Single copies may be obtained by contacting the OSHA Publications Office, Room N-3101, 200 Constitution Avenue, NW, Washington, DC 20210.

LABORATORY SAFETY

The laboratory environment is unique because of the types of hazards present. As we have already discussed in this chapter, laboratory professionals are surrounded by potential pathogens as they perform patient care and work with specimens. In addition, employees are using or are in the vicinity of chemicals of various types, as well as electrical testing equipment. Laboratory professionals are also exposed to bloodborne pathogens, as needle use and handling fluids from the human body are part of the daily duties in this environment. All of these potential risks make safety in the laboratory workplace somewhat complicated. This chapter has already provided information about protection from exposure to pathogenic microorganisms in the environment and basic infection control practices. Now we will emphasize chemical safety, physical safety, and bloodborne pathogen safety in the laboratory.

Chemical Safety

Numerous chemicals are present in the clinical laboratory. Some of these are used as cleaning agents or disinfectants, whereas others may function as preservatives for laboratory specimens. Hydrochloric acid (HCl), for example, is often used as a preservative for 24-hour urine specimens, and bleach is used as a recommended method for cleaning surfaces exposed to blood and other potentially infectious materials. Both acidic and alkaline substances can cause severe burns, and mixing various chemicals may have disastrous effects. To protect the employees using these chemicals, OSHA created the **Hazard Communication Standard.** This standard gives all employees the right to know about the potential hazards associated with the chemicals in their workplace. The required components addressed in this standard include the following:

- A hazard communication program to be developed and used within the facility
- Current inventory of all hazardous chemicals used in the workplace

- Appropriate labeling of those chemicals designated as hazardous
- Material Safety Data Sheets
- Documented training for employees

Test Your Knowledge 3-16

Who is the OSHA Hazard Communication Standard designed to protect? (Outcome 3-11)

The Hazard Communication Standard establishes comprehensive guidelines for labeling chemicals. These guidelines include the name of the chemical; contact information for the manufacturer; physical and health hazards of the chemical; safety precautions; and information pertaining to the storage, handling, and disposal of the chemical. The original container, as well as any transfer containers for the same chemical must be labeled in this manner.

Test Your Knowledge 3-17

The laboratory assistant who is working the evening shift notices that the laboratory is almost out of a disinfectant that staff members use to clean some of the instruments. He wants to be sure that the supervisor orders the same product that is now in use. Where can he look to find the name of the manufacturer to tell his supervisor? (Outcome 3-13)

A **Material Safety Data Sheet** (**MSDS**) must also be available for every hazardous chemical in use. This is a document provided by the manufacturer that provide even more details about the chemical than those printed on the label. A current MSDS must be kept on file for any chemical in use that is considered potentially hazardous, and these sheets must be accessible to employees at all times (Fig. 3-4). The standard requires that the following information be provided on all Material Safety Data Sheets:

- **Identification:** Must include the generic and brand name, as well as the name, address, and emergency phone number for the manufacturer, and the date the MSDS was prepared.
- **Composition of ingredients:** List of the ingredients, and how much of the chemical is considered to be safe for exposure.
- **Physical and chemical properties:** Includes items such as appearance, odor, boiling point, specific gravity, pH, etc.
- **Fire and explosion data:** Will this chemical catch fire or explode if used incorrectly? If so, how should you extinguish the fire?
- **Reactivity data:** How does this chemical react with other chemicals? What types of interactions should be avoided?
- **Health hazards:** Includes information such as the route of entry, signs and symptoms to be aware of in case of overexposure, medical conditions that might be worsened when this chemical is used, and acute or chronic health hazards that may develop with regular exposure to the chemical. This section of the MSDS is very important to the health-care professional who uses this chemical as part of his or her daily tasks. There is also information presented about the **carcinogenic** (cancer-causing potential) classification of the chemical.
- **Emergency first-aid procedures:** What should you do if overexposure occurs as a first aid measure while help is on the way?
- **Precautions for safe handling and use of the chemical:** Includes instructions for handling a chemical spill, how to store the chemical, and how to dispose of the chemical and or container.
- **Control measures:** Includes the personal protective equipment and safety apparatus that should be used when dealing with the chemical.

Test Your Knowledge 3-18

How many required components must be present on a Material Safety Data Sheet? (Outcome 3-12)

The last required component of the Hazard Communication Standard is employee training. All employees who might be exposed to hazardous chemicals at work must be provided information about the chemicals and their potential hazards prior to the beginning of their employment in the area where the chemicals are used, and again whenever the hazard may change with the addition of new chemicals or new duties. There must be documentation of each training session, and the education must be continuous; not only offered at the time of initial employment. There must also be assurance that the employee understood the information presented. The required training elements include the following:

- A general overview of the right to know standard, explaining all the components involved, where to find the inventory of hazardous chemicals in use, and where the written hazard communication program is kept for that facility

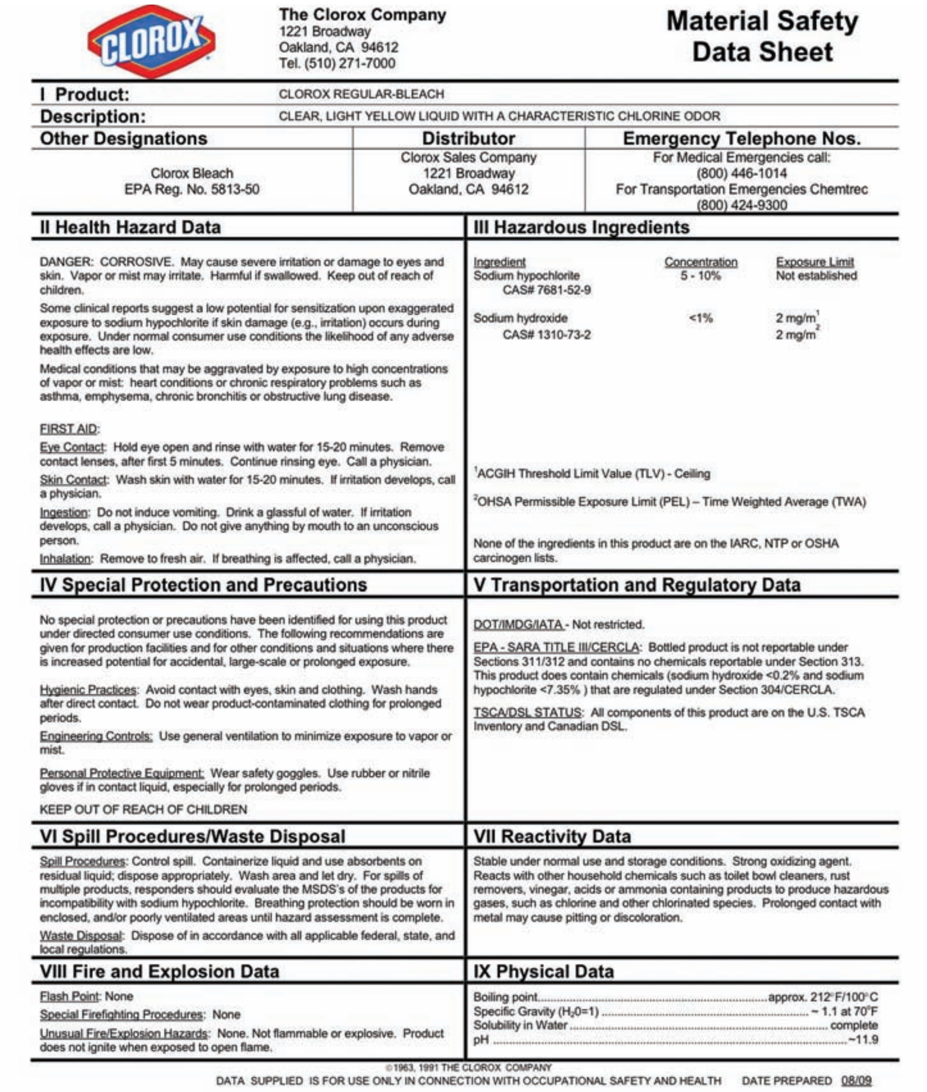

The Clorox Company
1221 Broadway
Oakland, CA 94612
Tel. (510) 271-7000

Material Safety Data Sheet

I Product: CLOROX REGULAR-BLEACH

Description: CLEAR, LIGHT YELLOW LIQUID WITH A CHARACTERISTIC CHLORINE ODOR

Other Designations	Distributor	Emergency Telephone Nos.
Clorox Bleach EPA Reg. No. 5813-50	Clorox Sales Company 1221 Broadway Oakland, CA 94612	For Medical Emergencies call: (800) 446-1014 For Transportation Emergencies Chemtrec (800) 424-9300

II Health Hazard Data

DANGER: CORROSIVE. May cause severe irritation or damage to eyes and skin. Vapor or mist may irritate. Harmful if swallowed. Keep out of reach of children.

Some clinical reports suggest a low potential for sensitization upon exaggerated exposure to sodium hypochlorite if skin damage (e.g., irritation) occurs during exposure. Under normal consumer use conditions the likelihood of any adverse health effects are low.

Medical conditions that may be aggravated by exposure to high concentrations of vapor or mist: heart conditions or chronic respiratory problems such as asthma, emphysema, chronic bronchitis or obstructive lung disease.

FIRST AID:

Eye Contact: Hold eye open and rinse with water for 15-20 minutes. Remove contact lenses, after first 5 minutes. Continue rinsing eye. Call a physician.

Skin Contact: Wash skin with water for 15-20 minutes. If irritation develops, call a physician.

Ingestion: Do not induce vomiting. Drink a glassful of water. If irritation develops, call a physician. Do not give anything by mouth to an unconscious person.

Inhalation: Remove to fresh air. If breathing is affected, call a physician.

III Hazardous Ingredients

Ingredient	Concentration	Exposure Limit
Sodium hypochlorite CAS# 7681-52-9	5 - 10%	Not established
Sodium hydroxide CAS# 1310-73-2	<1%	2 mg/m[1] 2 mg/m[2]

[1] ACGIH Threshold Limit Value (TLV) - Ceiling

[2] OHSA Permissible Exposure Limit (PEL) – Time Weighted Average (TWA)

None of the ingredients in this product are on the IARC, NTP or OSHA carcinogen lists.

IV Special Protection and Precautions

No special protection or precautions have been identified for using this product under directed consumer use conditions. The following recommendations are given for production facilities and for other conditions and situations where there is increased potential for accidental, large-scale or prolonged exposure.

Hygienic Practices: Avoid contact with eyes, skin and clothing. Wash hands after direct contact. Do not wear product-contaminated clothing for prolonged periods.

Engineering Controls: Use general ventilation to minimize exposure to vapor or mist.

Personal Protective Equipment: Wear safety goggles. Use rubber or nitrile gloves if in contact liquid, especially for prolonged periods.

KEEP OUT OF REACH OF CHILDREN

V Transportation and Regulatory Data

DOT/IMDG/IATA - Not restricted.

EPA - SARA TITLE III/CERCLA: Bottled product is not reportable under Sections 311/312 and contains no chemicals reportable under Section 313. This product does contain chemicals (sodium hydroxide <0.2% and sodium hypochlorite <7.35%) that are regulated under Section 304/CERCLA.

TSCA/DSL STATUS: All components of this product are on the U.S. TSCA Inventory and Canadian DSL.

VI Spill Procedures/Waste Disposal

Spill Procedures: Control spill. Containerize liquid and use absorbents on residual liquid; dispose appropriately. Wash area and let dry. For spills of multiple products, responders should evaluate the MSDS's of the products for incompatibility with sodium hypochlorite. Breathing protection should be worn in enclosed, and/or poorly ventilated areas until hazard assessment is complete.

Waste Disposal: Dispose of in accordance with all applicable federal, state, and local regulations.

VII Reactivity Data

Stable under normal use and storage conditions. Strong oxidizing agent. Reacts with other household chemicals such as toilet bowl cleaners, rust removers, vinegar, acids or ammonia containing products to produce hazardous gases, such as chlorine and other chlorinated species. Prolonged contact with metal may cause pitting or discoloration.

VIII Fire and Explosion Data

Flash Point: None

Special Firefighting Procedures: None

Unusual Fire/Explosion Hazards: None. Not flammable or explosive. Product does not ignite when exposed to open flame.

IX Physical Data

Boiling point........approx. 212°F/100°C
Specific Gravity ($H_20=1$)........~ 1.1 at 70°F
Solubility in Water........complete
pH........~11.9

©1963, 1991 THE CLOROX COMPANY

DATA SUPPLIED IS FOR USE ONLY IN CONNECTION WITH OCCUPATIONAL SAFETY AND HEALTH DATE PREPARED 08/09

Figure 3-4 MSDS information sheet for Clorox Bleach. *Courtesy of Clorox.*

- An explanation of what an MSDS is, where it may be found in the workplace, and how to use one
- Identification of the specific chemicals used in the work area for that employee, with instructions for the protective measures to be employed, including the personal protective equipment that is appropriate for that specific chemical
- The meaning of the labels and symbols used on hazardous chemicals
- Emergency measures to be taken in case of a spill or exposure

To ensure the safe use of chemicals, remember always to wear the appropriate personal protective equipment and safety equipment for the task at hand. Also, never use a chemical in a fashion other than that for which you were trained to use it. Do not transfer chemicals to unlabeled containers, and do not reuse containers that have been used previously. Remember that mixing chemicals may have disastrous effects; adding even water to some chemicals can be very dangerous. Finally, every employee should know the location of the eye wash station and emergency shower in the vicinity of their work area.

Physical Safety

As is the case with any business, there is always the chance that things will go wrong in the laboratory. This could include a chemical exposure, but it could also be a fire, an electrical emergency, or other personal injuries.

A laboratory professional needs to be aware of all hazards in the workplace, and also needs to be ready to respond to emergencies appropriately.

It is always essential for any business to have a plan of action in case of an emergency, but it is even more critical for a health-care facility such as a laboratory. Employees are not only responsible for themselves, but also for the patients in their presence at the time of the incident. Appropriate training and careful planning may dictate the difference between a positive or negative outcome of an emergency situation. This includes posting of emergency numbers (such as 911 or another internal number in a large facility for emergency response), maps showing the closest exit from various places within the building, employee training on use of fire extinguishers, and hazard identification.

The **National Fire Protection Association (NFPA)** has developed a labeling system that provides general information to employees and rescuers about the health, flammability, or reactivity hazard of chemicals. These categories are represented by blue, red, yellow, and white diamonds, each containing a number. The colors of the diamonds represent the different types of hazards, and the number contained within each diamond indicates the severity of the hazard. The blue diamond represents the respective health danger with exposure to the chemical. The red diamond indicates flammability hazard, and the yellow diamond indicates the reactivity potential for the chemical if exposed to increased heat or other conditions. There is also a white diamond, which may include a special symbol, indicating whether a chemical is radioactive or reacts with water (Fig. 3-5).

Test Your Knowledge 3-19

How do the NFPA fire labels protect employees?

(Outcome 3-14)

Fire Safety

In order to keep themselves and those around them as safe as possible, all employees should know the procedures to follow in case of fire in their facility. Fire safety basics such as Stop, Drop, and Roll (in the case of clothing that has caught on fire) are likely familiar concepts. However, when in the workplace, there are more aspects to consider: Where are the fire extinguishers? How do I use them? What can I do if there is not one nearby? Where is the nearest exit? How do I call for help? These are all questions that should be answered in initial training, and the procedures should be reviewed on a regular basis.

According to the National Fire Protection Association, there are four classifications used to describe fires, each of which has its own type of fire extinguisher to be used. Multipurpose extinguishers are also available to use for Class A, B, or C fires.

- **Class A:** Class A fires involve common household materials such as wood and paper. Water or a water-based

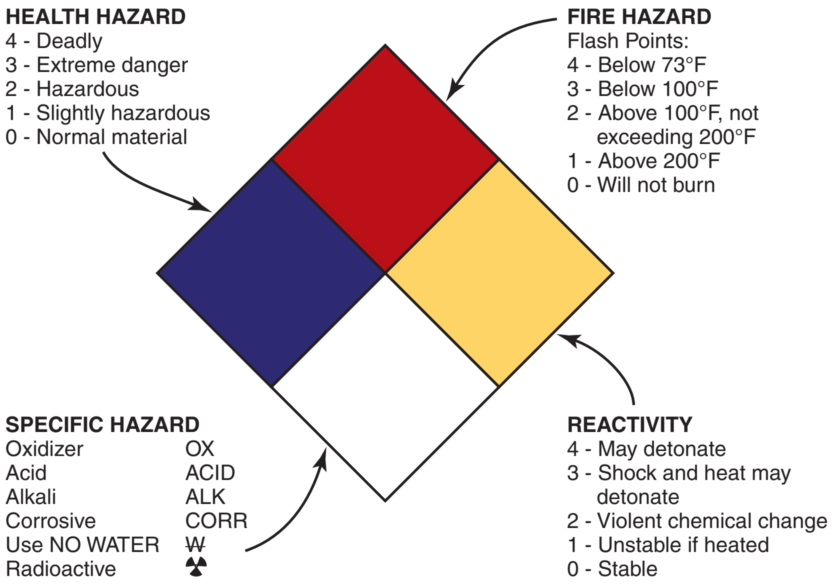
HEALTH HAZARD
4 - Deadly
3 - Extreme danger
2 - Hazardous
1 - Slightly hazardous
0 - Normal material
FIRE HAZARD
Flash Points:
4 - Below 73°F
3 - Below 100°F
2 - Above 100°F, not exceeding 200°F
1 - Above 200°F
0 - Will not burn
SPECIFIC HAZARD
Oxidizer OX
Acid ACID
Alkali ALK
Corrosive CORR
Use NO WATER ~~W~~
Radioactive ☢
REACTIVITY
4 - May detonate
3 - Shock and heat may detonate
2 - Violent chemical change
1 - Unstable if heated
0 - Stable

Figure 3-5 NFPA chemical fire symbol.

solution is needed to put out this type of fire, and these are contained in a Class A extinguisher.

- **Class B:** Class B fires generally involve flammable liquids and/or vapors, and they need to be smothered to be put out. The Class B extinguishers contain chemicals, carbon dioxide or foam.
- **Class C:** Class C fires are related to electrical equipment, so special care must be used to extinguish them. If the solution used to extinguish the fire conducts electricity, the fire will not be extinguished. The Class C extinguishers use chemicals or other types of substances that do not conduct electricity.
- **Class D:** Class D fires frequently lead to explosions, as they occur with reactive metals such as sodium or potassium. They are very difficult to control, and there is not a fire extinguisher available in most sites to extinguish this type of fire. Sand or other dry powder agents work best for this type of fire.

Fast action is critical when a fire is discovered to keep those in the workplace as safe as possible. A common acronym recommended to help those facing a fire remember what to do is **RACE**: Rescue, Alarm, Confinement, Extinguish. Table 3-1 lists the letters and expands on the meaning.

Test Your Knowledge 3-20

What is designated by the different letters assigned to fire extinguishers? (Outcome 3-14)

Electrical Safety

We are surrounded by electrical equipment in the laboratory environment, so that fire and electrical shock are definitely potential hazards. Most hazards can be minimized by appropriate maintenance and service of the equipment and the electrical outlets. The use of extension cords and overloading of electrical outlets should be avoided. Also, only qualified personnel should service electrical equipment. In case of electrical shock, remember that you must stay safe in order to help those who have been injured. Shut off the source of the electricity immediately, if possible. Do not touch the victim if you are unable to shut off the source of the electricity. Call for emergency assistance immediately. If the source of electricity has been eliminated, evaluate the victim and begin cardiopulmonary resuscitation (CPR) if necessary.

TABLE 3-1

RACE: The steps to take in the event of a fire

R	Rescue individuals in danger
A	Activate the alarm system
C	Confine the fire by closing windows and doors
E	Extinguish the fire using an appropriate fire extinguisher

Body Mechanics

As a medical assistant working in a laboratory environment, you may be performing numerous venipunctures each shift. These procedures often require bending over the patient sitting in the phlebotomy chair, or perhaps in a hospital setting, bending over the bed of a patient. This bending can cause a great deal of stress to the muscles of the back and neck, resulting in pain. There are additional duties in the laboratory that may require lifting or carrying equipment trays or supplies. A laboratory professional may need to assist a patient with transfer from a wheelchair to an examination table or other chair. In all these situations, it is important to keep your back as healthy as possible. Stretches and regular exercise help, as well as staying close to an ideal body weight. Remember to use the muscles in your legs when lifting heavy objects, and when carrying heavy objects keep them close to your body and avoid twisting motions. Change your position as often as possible; avoid sitting or standing for prolonged periods without a break. If patient transfer is part of the duties for a medical assistant in your work environment, appropriate techniques should be part of the initial training for employees.

Bloodborne Pathogen Safety

Many of you who are reading this textbook have never lived in a world without the presence of HIV and **acquired immune deficiency syndrome (AIDS).** However, it was not that long ago that these diseases were not yet known. Prior to 1980, there was not a lot of information available about potential bloodborne pathogen exposure in health care. **Hepatitis B (HBV)** was discovered in 1967, but research on the methods of transmission were still in progress for many years. With the discovery of the HIV in 1983, change came quickly. It became evident that health-care workers needed to be protected from the hazards of working with blood and other potentially infectious materials. Studies performed in the 1970s showed that the rate of hepatitis B infection for health-care workers was ten times that of the general population. A reliable vaccination was created and released to the general public in 1982, but compliance was voluntary, and many health-care workers had already been exposed to the virus. After HIV was identified, data were gathered to see how many health-care workers were potentially infected as a result of occupational exposure, and although the numbers related directly to exposure on the job were low, they were too high to be ignored.

Bloodborne pathogen is a term used to describe any pathogenic microorganism found in human blood that can cause disease in humans. These diseases are spread through direct contact with the bloodstream of another individual. This means that there must be a piercing of the skin **(parenteral exposure),** direct blood-to-blood contact as might occur when **nonintact skin** touches the blood of another individual, or mucous membrane exposure, as occurs with sexual activity or accidental splashes into the eyes or mouth. These diseases are not carried only in the bloodstream of the infected individual. Other potentially infectious materials include semen, vaginal secretions, cerebrospinal fluid, synovial fluid, pleural fluid, pericardial fluid, peritoneal fluid, amniotic fluid, saliva when dental procedures are being performed, and all body fluids that are visibly contaminated with blood. If it is impossible to tell which type of fluid you may be working with, it is also to be assumed that it is a potentially infectious material. Tears, feces, urine, sputum, nasal secretions, sweat, and saliva (spit) are not considered to be infectious unless they are grossly contaminated with blood.

Universal Precautions

Bloodborne pathogen exposure in health care was first addressed by recommendations from the CDC in 1982, specifically referring to the information about the newly identified HIV. In 1985, the CDC published a recommendation to adopt Universal Precautions when dealing with blood or other potentially infectious materials. The concept of universal precautions recognized the fact that it was impossible to identify all patients who may be infected with HIV or other bloodborne pathogens by taking a health history and performing an examination. The recommendation was that all patients (and all specimens of an infectious nature) were to be considered infectious for HIV and other bloodborne pathogens, and the same precautions were to be taken in all situations. Prior to this, gloves were not mandatory in health-care settings, mouth pipetting was still performed in the laboratory, and laboratory coats were not required. Special precautions were taken only if a specimen or patient was labeled as infectious. Currently, the CDC standard precautions have expanded universal precautions to include all means of infection.

Test Your Knowledge 3-21
What are some examples of OPIMs? (Outcome 3-16)

Bloodborne Pathogens Standard

In 1988, OSHA strengthened this recommendation of Universal Precautions with the publication of the **Bloodborne Pathogens Standard (1910.1030)**. This standard provided a rigorous policy to protect health-care workers who had **occupational exposure** to blood or other potentially infectious materials in the workplace. This standard had special significance for those who worked in a laboratory environment, because the amount of potential exposure was quite high for this population. This regulation required that all employers with potential occupational exposure to bloodborne pathogens in their workplace develop an **exposure control plan.** An exposure control plan must include the following components:

- **Exposure determination:** The plan requires that all employers determine which classifications of employees have risk of exposure, and during what types of duties that the risk exists.
- **Methods of exposure control compliance:** Exposure control compliance means that all employees are to be taught how to practice universal precautions by implementing appropriate personal protective equipment. It also required that the employer use and train the employees to use **engineering controls** and **work practice controls** to minimize their risk of exposure. Engineering controls are physical devices (such as needle safety devices or stationary safety shields) that are used by the employees to make their environment safer as they perform their duties. Work practice controls are ways that employees might alter the manner in which they perform a task to minimize the risk of bloodborne pathogen exposure. These include rules banning eating or drinking in an area where there is potential exposure, bandaging cuts before application of gloves, wearing gloves and other PPE appropriately, wearing close-toed shoes, and disposing of **biohazardous waste** and **contaminated sharps** correctly. Biohazardous waste is a classification used to describe substances contaminated with liquid or semiliquid blood or other potentially infectious materials. Contaminated sharps

are needles or other materials capable of piercing the skin that have been used with blood or other potentially infectious materials.

- **Hepatitis B vaccination** (or documented declination of the vaccine by the employee) must be made available for all employees with occupational risk of bloodborne pathogen exposure.
- Communication of hazards to all at-risk employees, which must occur initially upon assignment, and at least annually. (Retraining must also occur if there are changes to the procedures or hazards in a specific area of the facility.) Records for this training must be kept for at least 3 years. Initial training must include:
- Information about the **epidemiology,** signs, symptoms, and methods of transmission of bloodborne pathogens.
- Instruction about the specific methods employed by that facility (their exposure control plan) to protect their employees, as well as general preventative procedures.
- Information about the hepatitis B vaccine
- An explanation of the procedure to follow for an exposure incident

Test Your Knowledge 3-22

True or False: The OSHA Bloodborne Pathogens Standard was created to protect patients. (Outcome 3-15)

Test Your Knowledge 3-23

List two key components required in an exposure control plan. (Outcome 3-17)

Appropriate Sharps Use and Disposal. Unfortunately, more health-care professionals are exposed to bloodborne pathogens by contaminated needles or other sharp devices than through any other route. This means that safety must be considered at all times when using and disposing of contaminated needles, glassware, or any other instruments that have come into contact with infectious materials. In 1999, there was an update to the Bloodborne Pathogens Standard requiring that employers constantly evaluate new safety equipment for invasive procedures to be certain that they are using the safest methods and products. The addendum also requires that employees who actually use the products on a daily basis have input as to which product is most efficient and easiest to use before the choices are made for which item to purchase. There are many choices available for needle safety devices, and no excuse not to use them. Keep in mind the following when considering sharps safety:

1. Always make sure that employees are trained on the appropriate use of the needle safety device in use. Many of these devices are designed to be "one handed," and if the employee tries to use two hands, he or she is at great risk for a puncture with a contaminated needle.
2. Contaminated needles are never to be recapped with the original needle cover, bent, or broken off. They must be disposed of immediately after use.
3. A biohazard sharps container should always be within reach when drawing blood or giving injections so that the needle can be disposed of immediately. These containers must be rigid and puncture-proof, and should be snapped closed and disposed of when they are three quarters full.
4. If contaminated glassware is broken, use mechanical means to clean up the mess, and then dispose of the glassware in a biohazard sharps container. A small disposable broom and dustpan may be designated for this type of situation, or tweezers may be used to pick up the pieces of glass. Thicker industrial gloves that are puncture resistant may also be used. Clean the area with an appropriate disinfectant (10% bleach solution allowed to sit for 20 minutes) to eliminate potential bloodborne pathogen exposure.
5. When performing capillary blood draws from the finger or foot of a patient, always use a capillary puncture device that recoils into the original holder. Do not try to reuse these devices.
6. Use plastic tubes and other containers instead of glass whenever possible.

Test Your Knowledge 3-24

Are all needle safety devices operated in the same way? (Outcome 3-18)

Regulated Waste and Housekeeping. In a health-care setting, there is a need to sort the waste generated by the facility. Some of the waste may be disposed of in the same way as household trash. Other waste is regulated, and requires special attention for disposal. The OSHA definition of **regulated waste** is quite long. Essentially, the term refers to any liquid blood or other potentially infectious materials. It also refers to contaminated items that are soaked with blood to the extent that the blood could be released if the item were compressed, or items caked with blood so that the blood could be released if the item were handled. Regulated waste also refers to contaminated sharps and any other microbiological wastes that may contain blood or OPIM.

Regulated waste must be disposed of in containers that are clearly identifiable as biohazardous. The containers (or bags) must either be red or must be marked with a **biohazard symbol** (Fig. 3-6). When using bags to collect regulated waste other than sharps, care should be taken to avoid adding items to the bag that are not biohazardous in nature. All bags must be closed securely when full, to be enclosed in larger containers for disposal by a regulated waste facility. All sharps containers will also be picked up by regulated waste facilities for disposal. Only companies that are licensed to carry biohazardous waste are allowed to transport and incinerate this type of refuse.

Other housekeeping guidelines in the laboratory include disinfection of counters at the beginning and end of each shift (at a minimum) with a fresh 10% bleach solution or an OSHA-approved disinfectant that is designed to kill the hepatitis B virus and HIV. Also, remember all biohazardous spills must be cleaned up using a spill kit that is designed for this purpose.

Test Your Knowledge 3-25

Is all laboratory waste considered to be biohazardous?
(Outcome 3-19)

DISEASES CAUSED BY BLOODBORNE PATHOGENS IN THE LABORATORY SETTING

The OSHA Bloodborne Pathogens Standard is very comprehensive, and is designed to protect employees who work with blood and other potentially infective body fluids. But what exactly are the employees being protected from, and how serious are the diseases?

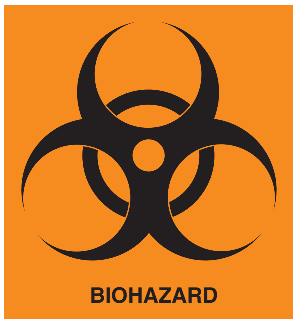

Figure 3-6 Biohazard symbol.

Hepatitis

Any discussion of bloodborne pathogens usually includes the topic of hepatitis. Essentially, **hepatitis** means "inflammation of the liver." When we discuss hepatitis in relation to bloodborne pathogens, we are referring to hepatitis that is caused by a viral infection. There are numerous types of the hepatitis virus that have been isolated, and not all of them are bloodborne. The ones that we are most concerned with are hepatitis B and hepatitis C, as these viruses are bloodborne pathogens, and they can cause serious illness. Table 3-2 lists details about the different types of hepatitis and other bloodborne pathogens.

Hepatitis B Virus (HBV)

Hepatitis B was first identified as a unique virus in 1967. It was formerly known as serum hepatitis, because there was a high prevalence of infection with the hemophiliac population who had received blood products as part of their treatment. The hepatitis B virus targets the liver, and it is the most serious hazard facing those in the health-care industry. HBV may be present in the blood and other potentially infective fluids of the body. Unfortunately, it may survive on surfaces in dried blood for up to 1 week. Hepatitis B may be transmitted in a health-care setting through needlesticks or other contaminated sharps, by contact with contaminated surfaces, or through mucous membrane exposure with aerosols or splatters created while handling infectious specimens. For those not employed in health care, transmission usually is the result of sharing contaminated needles, or through sexual contact.

Hepatitis B causes flu-like symptoms, and the affected patient may appear jaundiced. The symptoms of infection often do not appear until months after the virus has entered the body. Most of those who are infected recover, but approximately 2% of those infected will develop chronic infection, which causes **cirrhosis** of the liver or cancer. There are also patients that may be **carriers** of the hepatitis B virus who do not know it. These patients often have had no definitive symptoms of the infection, but they are capable of passing it on to others. Fortunately, a series of vaccinations may prevent hepatitis B infection. OSHA requires that these vaccinations be offered free of charge to all employees who have bloodborne pathogen exposure in the workplace.

There is also a newly discovered form of hepatitis virus that affects only those who are already hepatitis B positive. This appears to be a mutant or variant form of HBV, but it is unique enough to be identified as its own virus. This is known as **hepatitis D**. There currently is no vaccination available for hepatitis D.

Hepatitis C Virus (HCV)

You may have heard of **hepatitis C (HCV)** as the new "epidemic" disease of the world. This statement is misleading. Although there are millions of cases of hepatitis C infection that have been diagnosed over the past decade, these patients did not necessarily have a recent infection. Hepatitis C infection is usually **asymptomatic,** so most of those who are now being diagnosed with the disease were actually infected more than a decade ago. For many years researchers knew that there was another form of hepatitis virus that was different from both HBV and HAV. They just couldn't isolate it as a unique strain of the virus. In 1989, technology finally allowed the hepatitis C virus to be identified as another cause of post-transfusion hepatitis. The hepatitis C virus is a bloodborne pathogen, and many of those who are diagnosed today shared dirty needles for drug use, received blood transfusions or other blood products, or had tattoos applied prior to the mid-1980s. If there are any symptoms of the HCV infection, they are usually milder than those of hepatitis B, so are often misread as being symptoms of the flu. Unlike hepatitis B, in which it is uncommon to advance to the chronic form of the disease, those with hepatitis C infection usually progress to the chronic stage (approximately 80% to 85% of patients). Of these, a small percentage develop permanent liver damage and/or liver failure. Transmission of the hepatitis C virus in a health-care setting usually requires a larger accidental exposure than would be necessary to transmit HBV. There is no vaccination for HCV.

Hepatitis A Virus (HAV)

Although the **hepatitis A (HAV)** virus is not a bloodborne pathogen, it is important to understand how it is transmitted. HAV attacks the liver of infected patients and causes inflammation just like the other forms of viral infections that we have discussed. The hepatitis A virus is very contagious, and is transmitted by the **fecal-oral route.** The virus is shed in the feces of infected individuals, often before they even know they are infected. The fecal material containing the virus can then be passed on to others through direct contact (such as touching another individual) or indirect contact through food, water, or inanimate objects. Patients with HAV usually show symptoms sooner than those with HBV or HCV, and are often initially more severely ill. However, the hepatitis A virus rarely progresses to cause chronic infection or permanent liver damage. There is a series of vaccinations available to protect against hepatitis A infection.

Test Your Knowledge 3-26

Are all forms of hepatitis considered to be bloodborne pathogens? (Outcome 3-20)

POINT OF INTEREST 3-2
Pathogenic microorganisms

In most situations, the discussion of bloodborne pathogens focuses on hepatitis B, HIV, and hepatitis C. However, the OSHA Bloodborne Pathogens Standard is not only designed to protect employees from infection with these viruses. The standard actually includes any pathogenic microorganisms that may be present in blood or other potentially infectious materials. Examples of other diseases that may be transmitted by direct contact with blood of infected individuals include malaria, syphilis, babesiosis, brucellosis, leptosporosis, Creutzfeldt-Jakob disease, and viral hemorrhagic fever.

Human Immunodeficiency Virus

The human immunodeficiency virus (HIV) attacks the immune system of our bodies. The HIV infection may eventually lead to AIDS. Viruses need the genetic information in our cells to survive and replicate, and HIV is no different. It attacks a type of white blood cells in our body known as **CD4 cells.** This type of blood cell fights infection in our bodies. Those infected with HIV are **susceptible** to opportunistic infections. These infections are caused by microorganisms that would not normally infect individuals who have healthy immune systems. HIV infection also makes the infected patients more prone to develop certain types of cancer.

When patients are infected with HIV, they undergo several stages. Initially, they may have swollen glands, or they might have slight flu-like symptoms. This generally lasts from weeks to months. Also, initially, because the amount of virus particles in the infected patient's body is still low, an HIV test result would be negative. HIV tests are designed to detect antibodies to the virus, and in the early stages of infection, the level of antibodies in an individual are not high enough for detection. The patient is still capable of passing on HIV during this time through blood-to-blood contact or sexual activity. This is known as the **window period** of infection, and may last up to 6 months.

The second stage of infection is generally asymptomatic. The infected patient may still not know that he or she is HIV positive. This stage can last for months or even years.

When patients reach the third stage, they often contract one or more opportunistic infections. This can be the first time many patients find out that they are HIV positive. Their symptoms will vary, depending on the infection that they are experiencing.

The final stage of HIV infection is what we know as AIDS. At this stage, the HIV-positive patient has been diagnosed with an opportunistic infection or disease that has been classified by the CDC as an AIDS indicator. These diseases include esophageal thrush, cytomegalovirus infection, a type of cancer known as Kaposi's sarcoma, and invasive cervical cancer in the HIV-positive patient. An AIDS diagnosis may be made based on the number of CD4 cells present in a blood sample as well. Remember, HIV infection causes AIDS, but an infected patient does not have AIDS until he or she reaches this final stage and contracts one of the indicator diseases, or is diagnosed with AIDS because of the blood count.

For many infected individuals worldwide, HIV infection is fatal. The treatment regime has changed significantly in the past 25 years, and more HIV-positive patients than ever before are able to live long lives without acquiring AIDS. Fortunately, HIV is not a very hearty virus; it dies immediately when it comes into contact with air, and the amount of virus necessary for an infection to occur from an accidental exposure has to be quite high. Hepatitis B infection is a much more serious risk from accidental health-care worker occupational exposure than HIV infection. Phlebotomy procedures are most often responsible for HIV exposure that occurs in a health-care setting. As of 2000, the CDC had processed 56 claims in which it appeared that a health-care professional had contracted HIV from an accidental exposure. At this time, there is no vaccination for the human immunodeficiency virus, and accidental infection is battled with rapid administration of antiviral medications.

Test Your Knowledge 3-27

What are two characteristics that hepatitis B and HIV have in common? (Outcome 3-20)

POSTEXPOSURE FOLLOW-UP PROCEDURE

You have taken all the appropriate precautions, read all the training manuals, worn all the right personal protective equipment . . . but your patient jerks his arm right at the end of the phlebotomy procedure, and you are punctured by a contaminated needle. Now what? The Bloodborne Pathogens Standard carefully spells out what type of follow-up care employers must offer to their employees (free of charge) if there is an accidental exposure to blood or other potentially infectious materials in the workplace. Remember, an exposure can be **percutaneous** (through the skin) as in the scenario above, but it may also be a mucous membrane exposure through a splash in the eyes or in the mouth. Here are the steps to follow if there is an accidental exposure:

1. Wash the site thoroughly with soap and water. If it is a mucous membrane exposure, thoroughly flush the exposed area with water.
2. The incident must be reported immediately to the appropriate manager or other contact person within your health-care facility. Many of the antiviral medications used for treatment after an exposure should be started within a few hours, so an immediate report is critical.
3. Fill out an incident report form. The route of exposure and the circumstances surrounding the incident are very important and are required by law.
4. Identification of the source individual needs to be determined as soon as possible. Consent should be obtained for the source individual to be tested for HBV, HCV, and HIV, unless his or her status is already known for these diseases. If the individual refuses to consent for the test, the employer must document that a refusal was made. If allowed by local law, the test will be performed regardless, so that the employee can be treated effectively. The results of these tests will be made available to the exposed employee. A blood sample will also need to be obtained from the exposed employee to test for a baseline level for HBV, HCV, and HIV.
5. The exposed employee will have the opportunity to seek medical attention as soon as possible after the initial report has been completed. **Postexposure prophylaxis** is most effective if initiated rapidly after the exposure. This means that preventive medication (such as antiviral medicine) should be administered within hours of exposure. The health-care professional providing the follow-up care will be provided with information about how the accident happened, any laboratory testing results that might be available,

and all medical records for the employee that might be relevant (such as vaccination records). The health-care professional also must be provided a copy of the Bloodborne Pathogens Standard.

6. The health-care professional performing the follow-up care will determine whether postexposure prophylaxis is necessary for HBV, HCV, or HIV, depending on the risk evaluation. The route and amount of exposure, the vaccination status of the employee, and the circumstances of the exposure will be taken into consideration. The treatment for the employee often consists of a combination of drugs for an extended period of time.
7. Follow-up blood tests are generally performed on the exposed employee, at a minimum interval of 6 weeks, 12 weeks, and 6 months. It is possible that this series may be extended up to 12 months.
8. The exposed employee also must be offered counseling and reminded that confidentiality must be retained in reference to the source individual. All medical follow-up for the employee is also retained confidentially.

Test Your Knowledge 3-28

What should happen immediately if an employee splashes blood in her eyes? (Outcome 3-21)

Test Your Knowledge 3-29

True or False: All employees who are victims of a bloodborne pathogen exposure will be sent to see a health-care provider for follow-up care. (Outcome 3-21)

SUMMARY

The world we live in is a complex place, with new health hazards constantly emerging. As health-care workers, we must protect ourselves from potential pathogens in our environment, our patients, and our workplace. There are several government regulations that have been designed to protect employees in a health-care or laboratory environment. Education concerning these regulations, and careful adherence to their principles, are essential steps to working safely in this unique environment. Standard Precautions have been developed to assist with infection control efforts. The Hazard Communications Standard establishes guidelines to protect employees from chemicals in the workplace, and the Bloodborne Pathogens Standard educates health-care employees about the risks they face from bloodborne pathogens as they perform their duties. Medical assistants must be familiar with the details contained in these documents to use safety materials in their workplace and to be prepared for accidents when they happen.

TIME TO REVIEW

1. A bacteria that is aerobic needs ______________ to survive and replicate: Outcome 3-1
   - **a.** Nitrogen
   - **b.** Oxygen
   - **c.** Carbon dioxide
   - **d.** Water
2. True or False: A substance that produces or causes cancer is carcinogenic. Outcome 3-1
3. The types of microorganisms listed in this chapter that may be human pathogens include: Outcome 3-2
   - **a.** Protons, neutrons, and electrons
   - **b.** Bacteria, viruses, and parasites
   - **c.** Bacteria, viruses, fungi, and parasites
   - **d.** None of the above
4. True or False: Most of the types of bacteria present in our environment are pathogenic. Outcome 3-3
5. True or False: The presence of bacteria anywhere in our bodies is always an infection. Outcome 3-3
6. A round-shaped bacteria that appears in pairs when viewed under the microscope is known as a: Outcome 3-4
   - **a.** Cocci
   - **b.** Spirilla
   - **c.** Bacilli
   - **d.** Diplococci
7. Which type of infection is best treated by antibiotic use? Outcome 3-5
   - **a.** Viral
   - **b.** Bacterial
   - **c.** Viral and bacterial
   - **d.** None of the above

8. True or False: In a medically aseptic environment, bacteria may still be present on surfaces. Outcome 3-6

9. How does the "chain of infection" have to be broken to keep a pathogen from causing disease? Outcome 3-7

a. It must be broken at multiple points to keep the pathogen from causing disease
b. Only vaccinations will break the chain of infection
c. The chain of infection must only be broken at one point to discontinue the spread of infection
d. The pathogen must be killed before the chain will be broken

10. What are the elements addressed in CDC Standard Precautions document? Outcome 3-8

a. Appropriate use of personal protective equipment
b. Equipment disinfection procedures
c. Respiratory hygiene
d. Sharps use
e. All of the above

11. True or False: Hand washing is designed to remove the normal (resident) flora from our skin. Outcome 3-9

12. When using an alcohol-based hand rub to disinfect the hands, it is important to: Outcome 3-10

a. Put on gloves while hands are still damp
b. Rub the solution thoroughly on all parts of the hands and wrists
c. Use only a drop of the solution
d. Always wash the hands first before using

13. The OSHA Hazard Communications Standard addresses: Outcome 3-11

a. The use of personal protective equipment
b. The formal education required by everyone in the laboratory before they can be hired to perform certain tests
c. What type of fire extinguisher to use for various types of fires
d. The brand of hand sanitizer endorsed by the federal government

14. Which of the following components are required for any Material Safety Data Sheet? Outcome 3-12

a. Identification of the chemical
b. Color of the chemical
c. Health hazards
d. a and c
e. All of the above

15. True or False: Laboratory employees may protect themselves from occupational back injuries by using appropriate bending techniques when lifting heavy objects. Outcome 3-14

16. Why was the OSHA Bloodborne Pathogens Standard created? Outcome 3-15

a. To protect employers from lawsuits
b. To create a revenue source for OSHA
c. To educate and protect employees from the hazards of bloodborne pathogens
d. To protect patients from bloodborne pathogens

17. What is an example of parenteral or percutaneous bloodborne pathogen exposure? Outcome 3-16

a. A respiratory infection contracted when working with a patient with pneumonia
b. Touching a patient's blood with a bare hand with unbroken skin
c. An accidental needlestick with a needle contaminated with the blood of a patient
d. Splashing urine on the arm while pouring a specimen

18. True or False: The OSHA Bloodborne Pathogens Standard requires the development of an exposure control plan. Part of this plan requires that all employees who have occupational exposure to blood or other potentially infectious materials are offered an opportunity to receive a series of hepatitis C vaccinations at no charge. Outcome 3-17

19. True or False: When using a needle to draw blood, it is okay to lay it down on the table next to the drawing chair right after the blood draw, and come back later to activate the safety device and dispose of it correctly. Outcome 3-18

20. Who must transport biohazardous waste? Outcome 3-19

a. Employees of the laboratory that have been appropriately trained
b. The local municipal garbage disposal company
c. A certified, licensed biohazardous waste disposal company
d. None of the above

21. What do hepatitis A, hepatitis B, hepatitis C, and HIV all have in common? Outcome 3-20
    - **a.** They are all caused by viruses
    - **b.** They are all caused by bacteria
    - **c.** They all have vaccinations to prevent infection
    - **d.** None of them have vaccinations to prevent infection.

22. Are blood tests part of the follow-up procedure for health-care workers who are percutaneously exposed to blood in the workplace? Outcome 3-21

Case Study 3-1: Too tired to think

Jackie was really tired. She had been up all night with her 3-year-old daughter who had an ear infection. When she reported to her job at the GI clinic the next morning, she was relieved to find out that she would be working in the laboratory, as this meant that she would be on her feet a bit less than if she were working as the MA for one of the physicians. It was Friday, which was usually a slow day in the laboratory.

In the middle of the morning, Jackie had drawn three glass red-top blood tubes from Mr. Sifton. She was fighting back a yawn, and when she turned around to put the blood in the rack on the counter, she accidentally tapped one of the tubes against the edge of the counter with a great deal of force. The tube flew from her hand and broke against the refrigerator a few feet away.

Jackie was flustered. Mr. Sifton had already left, and she knew that she would have to call him and ask him to return for another blood draw. In addition, she had quite a mess to clean up, as the blood had splattered a great deal. She knew that the bloody mess had to take precedence, because other patients would be entering the laboratory at any time.

Jackie was already wearing her nitrile gloves, so she bent down and started to pick up the pieces of the tube that had broken. Suddenly she felt a sharp pain in her finger, and realized that she had punctured her glove and her finger with a slice of glass.

1. What are two things that could have been done to avoid the parenteral exposure to the patient's blood?
2. What is the first thing that Jackie should do now? What else needs to happen as soon as possible?

RESOURCES AND SUGGESTED READINGS

OSHA Hazard Communication Standard
OSHA regulation for chemical safety in the workplace http://www.osha.gov/SLTC/hazardcommunications/standards.html

Small Business Handbook
General reference tool to verify that the OSHA Standards are in place properly at your worksite http://www.osha.gov/Publications/smallbusiness/small-business.html#hazsub/

National Fire Protection Association
National Fire Protection Association website that includes information about fire, electrical, and building safety http://www.nfpa.org

Updated directives for OSHA Bloodborne Pathogens Standards, 1991
Additional information added to the OSHA Bloodborne Pathogens Standards in 1991 1910.1030- 29CFR http://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=standards&p_id=10051

"Viral Hepatitis"
Excellent information about the different forms of viral hepatitis presented by the Centers for Disease Control and Prevention http://www.cdc.gov/hepatitis/index.htm

Hepatitis B Foundation, "Meet Dr. Blumberg"
Facts concerning the discovery of hepatitis B, and Dr. Blumberg who was awarded a Nobel Prize in 1976 for this discovery
http://www.hepb.org/about/blumberg.htm

Chapter 4

Assuring Quality

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Assuring Quality in the Laboratory
- Quality Control and Quality Assurance
- What Is Quality?
  - The Medical Assistant's Role in Assuring Quality Results
    - Types of Quality Control Specimens
  - What Happens When Things Are "Out of Control"?
  - Documentation of Quality Control Results
  - Accuracy and Precision
- Other Methods of Assuring Laboratory Quality
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

4-1 Define each of the key terms.

4-2 Compare and contrast quality assurance and quality control.

4-3 Explain why quality is essential for laboratory testing.

4-4 Describe circumstances when a medical assistant may need to perform quality control testing in the laboratory.

4-5 Classify the types of quality control specimens used in the laboratory.

4-6 Explain the concept of troubleshooting and list various actions that may be taken when quality control samples are not in range.

4-7 Demonstrate the ability to correctly log quality control results with appropriate documentation techniques.

4-8 Provide examples of the ways that quality may be assured with laboratory testing procedures.

CAAHEP/ABHES STANDARDS

CAAHEP 2008 Standards

I.P.I.11. Perform quality control procedures

I.A.I.2. Distinguish between normal and abnormal test results

II.C.II.7. Analyze charts, graphs and/or tables in the interpretation of healthcare results

II.P.II.2. Maintain laboratory test results using flow sheets

ABHES Standards

Medical Laboratory Procedures 10.a: Quality Control, and Graduates: a. Practice Quality Control

KEY TERMS

Accuracy
Analyte
Calibration
Calibration verification
Clinical significance
Competency testing
Critical results
Electronic quality control devices
External quality control specimens
Internal quality control indicators
Mean
Oncologist
Panic results
Precision
Proficiency testing
Qualitative results
Quality assurance (QA)
Quality control (QC)
Quantitative results
Reference ranges
Shift
Trends
Troubleshooting
Valid

ASSURING QUALITY IN THE LABORATORY

Laboratory results are extremely important to patient diagnosis and treatment. However, the results reported back to the physician are only as good as the testing methods themselves. This chapter explains various ways that the quality of the testing process can be monitored, so that the patient results reported will be correct. Quality control (QC) and quality assurance (QA) will be covered in detail, as well as documentation techniques and troubleshooting. Other procedures that are part of overall test quality assessment will be explored, such as proficiency testing and competency assessments, and the rules for quality performance of CLIA-waived tests will be reviewed.

QUALITY CONTROL AND QUALITY ASSURANCE

The terms *quality assurance* and *quality control* seem so similar; you may be wondering why it is important to know the difference. Essentially, they are both part of a plan to provide excellent quality care to the patients we serve. Quality control is part of the quality assurance process, but with a more specific focus.

Try to remember the last time you visited a health-care provider or laboratory to have your blood drawn and tested. How was your experience? Did you receive the information you needed to prepare for the blood draw appropriately? Did you feel that the person who drew your blood was trained adequately? How confident were you that the testing process was done in a way that accurate and valid results were obtained and reported back to your physician? Did your physician contact you in a timely manner with your results? These impressions, even though they were very personal to you, are all part of a quality assurance program at that facility.

Quality assurance (QA) includes all the policies and procedures necessary to provide excellent quality in laboratory testing. Chapter 1, introduced the three parts of laboratory testing: preanalytical, analytical, and postanalytical processes. A good quality assurance program acknowledges every part of the testing process to ensure that the laboratory staff members are doing all they can to contribute positively to the welfare of the patient in a timely manner. Quality assurance may be thought of as the big picture and it may include policies and procedures for collection, patient education, training for those performing the testing, maintenance and record keeping of instruments, monitoring of ordering and storage procedures for testing kits or reagents, and monitoring of turnaround times for results to be available to the health-care provider. Box 4-1 shows more detailed components of a quality assurance plan.

Quality control (QC) practices are used in the laboratory to ensure the quality of the testing process and the accuracy and precision of the results. Although the focus of quality control is specifically on the testing process, these practices are not limited to the details of how the test was performed. A good quality control program includes appropriate employee training, documentation of storage and shipping of reagents or testing kits used for testing, adherence to the manufacturer's instructions, correct interpretation and documentation of results, and verification that the results being reported are accurate and reliable. Quality control focuses on the second part

BOX 4-1 QA processes

Preanalytical Phase

- Requisition appropriately filled out
- Patient preparation
- Patient identification
- Sample identification; three identifiers must be used
- Correct sample collection techniques; venipuncture performed correctly, etc.
- Correct storage and transportation of sample
- Correct storage of testing kits and quality control materials
- Appropriate training of personnel
- Preparation of standard operating procedures and policy manuals for reference
- Retention of current package inserts for correct instructions

Analytical Phase

- Valid test kit; not expired and stored correctly
- Instrument properly maintained and calibrated
- Proper mixing/preparation of quality control samples
- Proper mixing/preparation of patient samples
- Following manufacturer instructions exactly as written for CLIA-waived tests
- Proper documentation of quality control results; verification of acceptable QC values before patient sample is analyzed
- Order/perform confirmatory tests as recommended by manufacturer

Postanalytical Phase

- Appropriate documentation of patient results
- Appropriate units reported with test results
- Results transmitted to health-care provider in timely manner
- All critical results called immediately
- Include correct reference ranges with laboratory results
- Report all test results to authorities as required by law
- Properly dispose of used testing materials

of the laboratory testing process: the analytical phase. Anything that relates directly to the testing process may be part of quality control. Table 4-1 shows how quality assurance and quality control compare to one another.

Test Your Knowledge 4-1

List two ways that quality assurance and quality control are alike. (Outcome 4-2)

TABLE 4-1 Quality assurance versus quality control

Quality Assurance	Quality Control
Addresses procedures in all phases of laboratory process; includes quality control	Focuses specifically on the testing process, or analytical phase of the laboratory process
Comprehensive plan including patients, employees, and employers	Includes processes performed by employees and influenced by policies of employers
Plan developed by the employer and the quality assurance officer	Policies often established by the manufacturer; also may be developed by the employer
Requires comprehensive and detailed documentation	Requires detailed documentation
Requires critical thinking and analysis to design the plan and keep it current	Requires critical thinking and analysis to decide if the patient sample can be reported based on the quality control results

WHAT IS QUALITY?

The laboratory environment is no different from that of any other business. Customers are essential to the survival of the laboratory, and the testing provided is a service that the customers need. Who are the customers? This is not quite as easy to answer as it might seem at first. The laboratory serves patients; that is pretty clear. The patients may be varied in age, socioeconomic class, race, and gender, but laboratory results are a critical part of their care. However, the laboratory can serve the patients only if the quality of the laboratory performance meets the expectations of the health-care provider who sends the patient (or the specimens from the patient) to a specific laboratory for testing. Therefore, the health-care providers are also laboratory customers.

The performance of the laboratory must be of very high quality to serve both of these types of customers efficiently. Consider this scenario:

> *A new* ***oncologist*** *opens a practice in your community. He specializes in providing care to patients with cancer, and is very concerned that the laboratory*

testing available will be sufficient to meet his needs. Within the first week, he orders several screening tests for ovarian cancer to be performed on his patients. The analyzer used to perform these tests breaks down on Monday, and no one in the laboratory contacts the physician to let him know of the delay. He finally receives his screening tests on Friday, and one of the results is positive. He is very upset about the time that it took for him to receive this result in his office, and decides to send all his testing requests to another laboratory.

This is an example of a breakdown in quality assurance. If the physician had been notified of the expected delay for his results, he may still have been upset, but at least the lines of communication would have been open. Another alternative would have been sending the samples to another laboratory while the machine was serviced; the delay would most likely have been minimized, and the physician would have received his results in a timely manner.

Other examples of poor quality performance are incorrect results, improper patient preparation, reporting the wrong **reference ranges** (a range of expected values for a healthy individual for a specific test) for a specific patient's gender and age, or even improper billing practices. All aspects of the customer experience in the laboratory must be of high quality, as the test results may have significant impact on the treatment for the patient. A patient may be put on antibiotics when they are not needed because a streptococcal screen was incorrectly interpreted as positive. Another patient result may show a false elevation in potassium levels when a specimen is mishandled, and the physician may change their medication based on these results. This could be a life threatening situation.

It may be helpful to consider laboratory quality as a series of "Rights":

- Right result
- Right time (both specimen collection and reporting of results)
- Right specimen
- Right patient
- Right reference range for result interpretation
- Right price for the services provided

Test Your Knowledge 4-2

How do incorrect laboratory results affect patient care?

(Outcome 4-3)

The Medical Assistant's Role in Assuring Quality Results

Most medical assistants working in a laboratory environment will be performing Clinical Laboratory Improvement Act–waived (CLIA–waived) tests. These are tests that are very simple to perform with results that are easy to interpret. As introduced in Chapter 3 of this text, laboratory tests are categorized under CLIA regulations based on the complexity of the process. Some of the tests are categorized as "waived" tests because they are exempt (or waived) from many of the regulatory procedures of the laboratory tests that are more complex in nature. This does not mean, however, that they are exempt from quality control procedures. Even before the testing process begins, the medical assistant can be checking the sample collection devices (such as swabs or blood tubes) for outdates, and participating in necessary training to be ready for the procedure.

The first opportunity that a medical assistant might have to be involved in quality control is when the testing kit or reagents arrive in the laboratory. Storage instructions are printed on the outside of the box, and these need to be followed precisely to provide accurate test results. Reading the package insert is critical, as the procedure specified by the manufacturer must be followed carefully. When the box is opened, this package insert should be stored in a safe place. Many laboratories have a binder for these sheets, and it is the medical assistant's responsibility to verify that the insert has not changed since the last one was placed in the binder. Laboratories may also keep multiple copies that have the date of receipt documented for reference.

Types of Quality Control Specimens

When the kit (or the reagent) is opened for use, it will be necessary to test a quality control specimen as recommended on the package insert. Some testing procedures may call for two types of quality control specimens to be analyzed. There are three general categories used to describe quality control specimens:

1. **External quality control specimens:** This may be a reference solution that comes with the kit, or a swab that has been treated with material that will cause the test to read positive or negative. External quality control materials may also need to be purchased separately from the testing materials. Regardless of how the materials are provided, the most important concept to follow is that of the analysis of the quality

control specimens exactly as directed by the test manufacturer, and the comparison of the results from the specimen to the data provided to be sure that the test performance is as expected. These may be **qualitative results** for tests that provide a positive or negative result, or they may be **quantitative results** with a range of expected results. Quantitative tests provide results as numbers indicating the concentration of certain substances. If the results of the external quality control specimen are not as expected, the kit or reagent cannot be used to test patient samples until the problem has been identified and solved.

2. **Internal quality control indicators:** These are quality control indicators that are part of the individual testing kits. They show that the test process was **valid,** but they do not provide a test result. An internal quality control measurement is usually a color indicator, and in the package insert there will be an explanation of acceptable performance before a test result can be reported. Internal quality control specimens are most common in qualitative testing procedures. Remember, if the test is not valid (the indicator does not develop properly), you may not report the patient test result. Figure 4-1 shows how an indicator might appear.
3. **Electronic quality control devices:** Electronic quality control devices are reusable, and are used to verify the function of an instrument used for analysis. The electronic quality control devices are commonly test strips or cuvettes. The devices come with an expected numerical range of results, and it is imperative that the results fall within this range when tested, or no patient results may be reported until the problem has been identified and solved.

Test Your Knowledge 4-3

Do electronic QC devices let you know whether a test is valid? (Outcome 4-3)

Test Your Knowledge 4-4

If a medical assistant is using a machine to analyze a blood sample for a test result, what type of QC may she be required to use?

a. External
b. Internal
c. Electronic
d. a and c (Outcome 4-4)

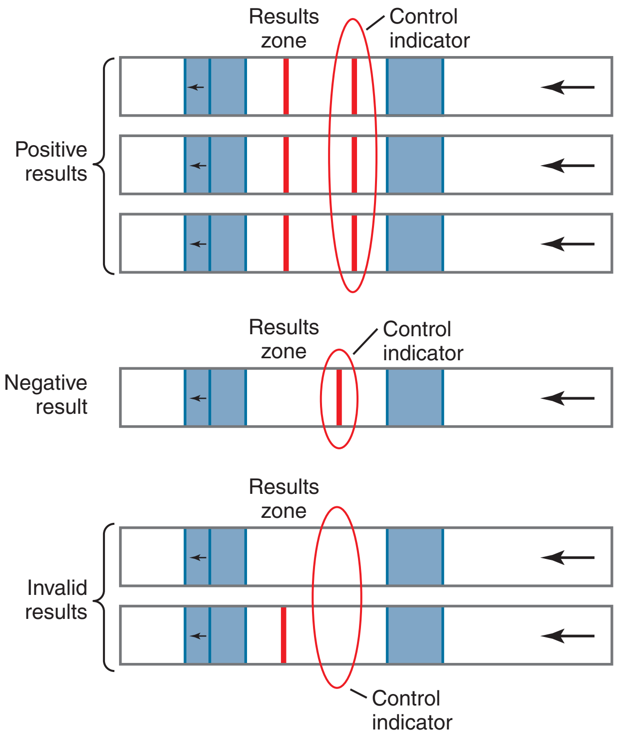

Figure 4-1 Valid and invalid test results.

Medical assistants may also perform tests that have been classified as CLIA moderate-complexity procedures. These procedures may require quality control procedures to be performed with more frequency than the CLIA-waived tests, and there is more training and documentation required for the employee who is performing the tests. Laboratories that perform these moderately complex tests are required either to follow the recommendations for quality control included in the package insert from the manufacturer, or process at least two levels of quality control materials with each analytical run. The CLIA regulations state that the laboratory must use the more stringent option of the two. Tests of moderate complexity utilizing instruments also require calibration, calibration verification, and proficiency testing. These concepts are discussed later in this chapter.

What Happens When Things Are "Out of Control"?

It is very important to recognize the importance of quality control specimens as an indicator of the reliability and accuracy of the testing method to be utilized for

Procedure 4-1: Documentation of Qualitative Quality Control Values

TASK

Properly perform quality control testing and document qualitative QC values on a chart.

CAAHEP/ABHES STANDARDS

CAAHEP 2008

I.P.I.11. Perform quality control procedures
I.A.I.2. Distinguish between normal and abnormal test results
II.C.II.7. Analyze charts, graphs and/or tables in the interpretation of healthcare results
II.P.II.2. Maintain laboratory test results using flow sheets

ABHES Standards

Medical Laboratory Procedures 10.a: Quality Control, and Graduates: a. Practice Quality Control

CONDITIONS

- CLIA-waived test kit
- External quality control material
- Quality control chart
- Pen

Procedure	Rationale
1. Open the CLIA-waived test kit and secure the package insert.	It will be necessary to have the package insert available for reference as the quality control specimen is analyzed.
2. Verify which type of quality control specimen your laboratory uses for this testing procedure. Find the directions in the package insert for that type of specimen.	The quality control specimen may be one that is provided with the kit or it may be a commercial one that is purchased separately. It is important that the correct procedure be used for the testing process.
3. Wash hands and apply gloves.	Appropriate personal protective equipment should be used when testing quality control materials just as it would be used for testing samples.
4. Following the directions in the package insert, perform the testing procedure using the positive quality control specimen. Repeat with the negative quality control specimen.	The performance of the testing procedure should be verified with both a positive and negative control specimen as recommended by the manufacturer.
5. Verify that the positive and negative quality control specimens provide the appropriate results.	Tests must be valid and the results must be as expected for the test kit to be considered "in control."
6. Remove gloves, dispose of them appropriately, and wash hands.	Gloves should be removed before documenting results.
7. Record the results on the log sheet. Include these factors: **a.** Date and time of test and employee ID **b.** Analyte tested (what type of test was performed?) **c.** Type of quality control (external); brand name if a commercial control was purchased separately	All these factors are very important documentation, and any form should allow for the recording of complete information.

Procedure	Rationale
from the kit, and lot number of quality control material. **d.** Result (positive or negative) **e.** Was result acceptable? **f.** Corrective action taken, if necessary	
8. Put away testing supplies and disinfect work area.	

Procedure 4-2: Documentation of Quantitative Quality Control Values

TASK

Properly perform and document quantitative quality control values on a chart. Determine whether the values are acceptable according to laboratory protocol.

CAAHEP/ABHES STANDARDS

CAAHEP 2008

I.P.I.11. Perform quality control procedures
I.A.I.2. Distinguish between normal and abnormal test results
II.C.II.7. Analyze charts, graphs and/or tables in the interpretation of healthcare results
II.P.II.2. Maintain laboratory test results using flow sheets

ABHES Standards

Medical Laboratory Procedures 10.a: Quality Control, and Graduates: a. Practice Quality Control

CONDITIONS

- Quantitative testing method
- External quality control material
- Quality control chart with established acceptable ranges provided
- Pen

Procedure	Rationale
1. Consult the instructions for the testing method if necessary before continuing with the process.	With automated testing procedures, the manufacturer's recommendations usually do not change as often as they might with the CLIA-waived test kits. However, consulting the instructions is always appropriate.
2. Verify which type of quality control specimen your laboratory uses for this testing procedure.	The quality control specimen may be one that is provided with the testing method, or it may be a commercial product that is purchased separately.
3. Wash hands and apply gloves.	Appropriate personal protective equipment should be used when testing quality control materials just as it would be utilized for testing samples.
4. Following the directions in the package insert or the laboratory procedure manual, perform the testing procedure using the quality control specimen. Repeat with a second quality control level if recommended by the manufacturer.	The performance of the testing procedure should be verified with both a high and low level quality control specimen as recommended by the manufacturer.

Continued

Procedure 4-2: Documentation of Quantitative Quality Control Values—cont'd

Procedure	Rationale
5. Put away quality control material.	Many quality control specimens must remain refrigerated, so this should occur immediately after use.
6. Remove gloves, dispose of them appropriately, and wash hands.	Gloves should be removed before documenting results.
7. Record the results on the log sheet by placing a dot at the line at which the date and specimen result intersect. In addition, document the time of the test and your identification. Verify that the correct control level brand and lot number are on the log sheet.	All these factors are very important documentation, and any form should allow for the recording of complete information.
8. Verify whether the quality control is within the established ranges and whether there is a pattern developing that needs attention, such as a trend or a shift in the results. If there is a problem or if the quality control is out of range, do not process patient samples.	Careful documentation of the results and observation of any quality control patterns are absolutely necessary. Patient samples cannot be processed if the sample is out of range or if there is an undesirable pattern forming in the results.
9. Repeat the test if results are out of range. If the condition persists, begin troubleshooting the problem. Document the repeat and also if there are troubleshooting steps taken.	Troubleshooting may include opening a new bottle of quality control material, cleaning the instrument, verifying the instructions for performing the test, verifying that the quality control chart is for the correct lot number, etc.
10. If the troubleshooting is not successful, do not allow the instrument to be used for patient samples, and contact a supervisor or the manufacturer for further action.	It is essential that the problem be resolved as soon as possible to continue with quality patient care.

patients. When the QC results are not as expected, the person running the test or operating the instrument must not report any patient results until the problem has been solved. The systematic process of identifying the problem with the testing procedure is called **troubleshooting.** There are times when the quality control specimen does not provide a positive result as expected, or when the test results appear invalid when running a pregnancy test QC specimen. The medical assistant performing the test must immediately try to identify the problem with the testing procedure before the process can continue. Here are some actions to take when troubleshooting a testing process:

1. Check the expiration dates of the reagents, kit, and quality control reagents in use. If any of these supplies are outdated, document the discrepancy, discard the expired item, and perform the test again using materials that are not past their expiration date.
2. Verify how the specimen, kit, reagent, or quality control material has been stored. Were the manufacturer's recommendations followed? If there is evidence that the storage instructions were not followed, the results provided by this testing method may be inaccurate. The testing component that was mishandled must be discarded, with appropriate documentation of the issue.
3. Verify the specimen type required for the testing procedure. Sometimes a kit is designed to work only with certain types of body fluids. The acceptable specimen type will be provided as part of the package insert.

Tests performed on body fluids not indicated as acceptable may provide inaccurate or invalid results.

4. Was the control material well mixed and at the correct temperature for the test? The quality control material may require warming to room temperature before use or vigorous mixing prior to testing. These instructions must be followed carefully.
5. Refer to the current package insert. Are you aware of skipping a step, or doing the procedure incorrectly? Has the package insert or instructions changed since you last performed the test?
6. If an instrument was used for the test, is it clean? Check the maintenance documentation, and see if the recommended schedule has been followed for routine instrument care. If needed, perform the recommended maintenance, after which the QC can be analyzed again.
7. Read the troubleshooting portion of the manufacturer's instructions. There is valuable information here that may help to solve the problem.

Usually the problem may be identified by following these suggestions. If the problem is still not apparent, a new quality control specimen should be obtained for a retest. If it is an internal control and the quality control specimen comes with the kit, open a new kit and try the test again. Finally, documentation of the details and troubleshooting process is critical. If there seems to be a pattern emerging, it is imperative that everyone knows what has occurred, and the steps that have been taken to solve the problem. Details also help while communicating with a manager or supervisor about the problem. Sometimes the medical assistant may need to contact the manufacturer for advice, and these details will be necessary to continue the troubleshooting process.

Test Your Knowledge 4-5

Cindy Lou has recently been trained to perform pregnancy testing in the physician office laboratory (POL) where she is employed. She receives a blood specimen from the physician who asks her to test it to see if a young patient is pregnant. Cindy follows the pregnancy testing procedure exactly as she was taught, but the test results show that it is an invalid test. What are two possible explanations? (Outcome 4-6)

Documentation of Quality Control Results

Quality control requirements are different for CLIA-waived tests than they are for those of moderate or high complexity. Many of the waived tests are qualitative in nature, which means that the result provided is either positive or negative, rather than a measurement reported as a number or percentage. The quality control assessment for these qualitative tests may be built in to the individual testing devices or the laboratory may purchase other quality control specimens to verify the performance quality for the **analyte** (substance being tested). Whether the QC is internal or external, it must be documented, and this record must be kept for at least two years. As with any document in the laboratory or in a patient's chart, the results on the quality control chart must be documented in pen, not pencil. This provides a permanent record that is not easily changed. One of the most common sources of laboratory error is that of transcription; it is very important to carefully document all laboratory data. Figure 4-2 is an example of a quality control log sheet to be used for a qualitative test. At a

Date and time of test/ Employee identification	Analyte tested	Type of quality control specimen used; external or internal?	For external quality control; brand and quality control lot numbers	Quality control result	Results acceptable? (yes or no)	Corrective action taken

Figure 4-2 Log used to record quality control results for a qualitative test.

minimum, the documentation should include the following:

- **The date and time of the quality control test.** Remember, the frequency that the QC is tested must be at least as often as the manufacturer recommends.
- **The quality control material lot number and level.** If it is internal control, this would be documented as additional information. If the testing device has a color change that occurs to verify the validity of the test, this needs to be recorded as well.
- **The lot number for the testing kit in use.** If this test kit was just opened, the kit should also have the date and the initials of the person who opened it written on the test kit box.
- **Results of the quality control test.** This is often just positive or negative for qualitative CLIA-waived tests.
- **Documentation of whether the kit or test can be used to test patient samples.** Remember, this decision is based on whether the quality control results were within the established acceptable ranges.
- **Documentation of troubleshooting or corrective action, if required.**
- **Initials or signature of employee performing test.**

Test Your Knowledge 4-6

Why is it important to document the date of a quality control test? (Outcome 4-7)

Various moderate-complexity tests, as well as some that are CLIA-waived, will provide quantitative results. This means that the quality control results for these testing methods will also be reported as a number, rather than merely a negative or positive result indicating the presence or absence of a substance. Because the results reported for these tests are to fall within an established range, it is very important that the quality control results are monitored closely to verify the method's performance. A common tool used to record these quantitative results is a Levey-Jennings chart, used to monitor the quality control performance over a set time interval. The acceptable quality control ranges are either established by the laboratory before patient samples are analyzed with the kit or instrument, or they are established by the manufacturer and adopted by the laboratory. The values obtained each time the quality control is processed are charted (in pen) according to the date and the value obtained. The chart also has spaces to document the analyte being tested, the QC level, the lot numbers of the kit and the quality controls, and other information that may be necessary for that particular laboratory. Levey-Jennings charts will also include the **mean** (average) expected value, as well as limits for acceptable QC testing results. If an individual quality control test falls outside of these ranges, patient testing cannot continue until the problem has been addressed. This type of graph can also be monitored for **trends** in the data, which show a progressive change in the results from day to day that continue to move gradually in one direction. **Shifts** will also be evident, and can be identified on the graph as a set of results that suddenly move significantly, but still remain within an acceptable range. Sometimes a shift will be evident when the results move from one side of the mean to the other, but a shift may also be evident even if the results remain on the same side of the mean. Both of these situations must be addressed, because they are indicative of a problem with the quality control procedure, the testing method, the operator, or a combination of these factors. Figure 4-3 is an example of a quantitative Levey-Jennings quality control chart.

Test Your Knowledge 4-7

Is a Levey-Jennings chart used to document qualitative QC test results? Why or why not? (Outcome 4-7)

Accuracy and Precision

The Levey-Jennings chart can be a useful tool to verify whether a testing method is demonstrating accuracy and precision. These are very important concepts to keep in mind in the laboratory environment. **Accuracy** is a word used to define how close the result produced for that test is to the real value. This can be visualized by thinking of a bull's-eye target. The true value for a test procedure is at the center of the target. If the test is accurate, then each time the test is performed the results will be near the center of the target.

Precision describes the reproducibility of results. This can also be thought of as a measurement of whether the results are close together if the test is run multiple times on the same specimen. Visualizing the bull's-eye target again, a test shows good precision if the results are all clustered closely together when the test is run multiple times.

A test can be precise but not accurate. This means that the results for a certain test may all be close to the same value, but not near the true value for that test. The goal for laboratory testing is to have results that are both

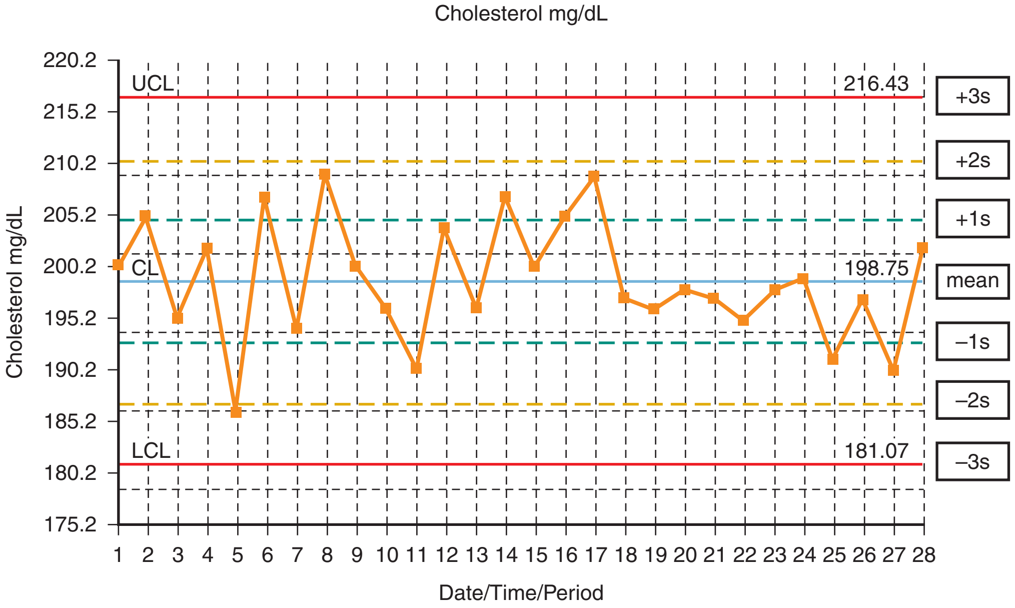

Figure 4-3 Levey-Jennings chart used to record quantitative quality control results.

accurate and precise, and quality control measures help to identify problems in these areas before they affect the quality of the patient results for that test.

OTHER METHODS OF ASSURING LABORATORY QUALITY

All laboratory results have **clinical significance,** meaning that a health-care provider may change the treatment or assign a diagnosis for a patient based on the laboratory results. In addition, there are established **reference ranges** for all laboratory tests. These are the expected values for patients in a normal population for a certain analyte. In order to feel confident that the laboratory results reported are accurate and reliable, all facets of quality assurance must be taken into consideration. The best way to understand how these all fit into patient care is to consider a patient scenario with a negative outcome:

> *Mr. Butler has developed an atrial arrhythmia and is being treated with the drug digoxin to help increase the efficiency of his heart muscle and correct problems with the electrical impulses that pass through his heart. Mr. Butler visits the laboratory monthly to have his drug level checked, as digoxin has serious side effects if the concentration is too high in the bloodstream. When the laboratory tests the digoxin level, a higher-than-normal level is discovered for someone being treated for his condition, and the result is well outside the reference range. When Mr. Butler's physician receives the report, he calls his patient, verifies his health status, and tells him that he needs to decrease his digoxin dose for a few days. The physician asks Mr. Butler to go to the laboratory to have another specimen drawn at the end of the week. On day three after decreasing his dose, Mr. Butler has his drug level rechecked. This second draw is extremely low, well below the reference range. The result is drastically different from what it was a few days earlier, and is much lower than it should be with the minor medication adjustment based on the previous high result.*

Of course, Mr. Butler's health-care provider is eager to find out how the results for the digoxin could be so different in just a few days. He contacts Mr. Butler and confirms that he followed the dosage changes as directed. The physician then contacts the laboratory that performed the testing on the first sample. The laboratory staff rechecks the sample, and finds the result is close to the first reading: still high above the reference range. However, when their quality assurance officer steps in to resolve the issue, she finds that the quality control results for the digoxin test have been slowly reading higher and higher for a few days prior to the test date. This trend has gone unnoticed by the personnel performing the tests, because the results were all within the acceptable range.

POINT OF INTEREST 4-1

Statistics used for quality control assessment

Statistics is a specific branch of mathematics that deals with the collection, analysis, and interpretation of numerical data. We use statistics in the laboratory to analyze the quantitative data derived from quality control testing. The entire process is quite complicated, but developing an understanding of a few key terms can benefit the medical assistant as he or she works with these data.

To establish acceptable ranges for qualitative QC testing procedures, the mean, standard deviation, and variance must be established.

- **Mean:** The mean is the numerical average of a set of numbers. To calculate the mean, find the sum (total) of a set of values (numbers), then divide that sum by the number of values in the set. For instance, if you were to analyze a specimen ten times for a potassium level, you might generate these numbers: 4.5, 4.4, 4.2, 4.6, 4.5, 4.1, 4.3, 4.6, 4.5, and 4.7 mmol/L. The mean of these values would be calculated by adding them all together, then dividing by 10, because there were ten results used for the calculation. The mean result for these numbers would be 4.4 mmol/L.
- **Variance:** The variance is a bit more complicated to calculate. It gives a numeric indication of how much the values in the set vary or deviate from the mean. The variance is calculated after the mean for the set of numbers has been established. Each number is subtracted from the mean, and this difference is squared, which means to multiply it by itself one time. The sum of the squares is then established, and this sum is divided by the total number of values minus one. (The term used for this calculation is $n - 1$.) Therefore, for our scenario presented above, we would follow these steps from the calculate mean of 4.4 mmol/L:

The sum of the deviations squared is 0.34, and this number divided by $n - 1$ (9 in our example) is 0.036. The variance for this set of numbers is 0.036.

Potassium Value	Deviation; How Far from the Mean of 4.4?	Deviation Squared; Multiplied by Itself
4.5	0.1	0.01
4.4	0.0	0.00
4.2	0.2	0.04
4.6	0.2	0.04
4.5	0.1	0.01
4.1	0.3	0.09
4.3	0.1	0.01
4.6	0.2	0.04
4.5	0.1	0.01
4.7	0.3	0.09

- **Standard deviation:** The standard deviation provides a measurement of how the values we are working with are placed or scattered around the calculated mean for the values. The standard deviation uses the variance that was calculated in the previous step, and is calculated by taking the square root of the variance. For the potassium scenario that we are working with, the standard deviation would be calculated by calculating the square root of 0.036. The standard deviation would be 0.061 mmol/L for this set of potassium results.

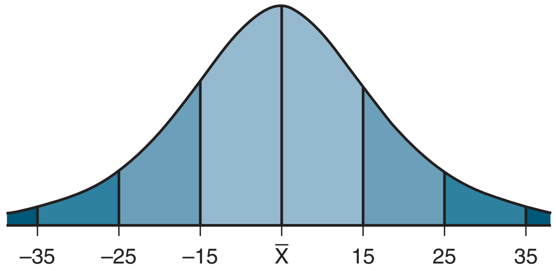

The previous calculations are used to establish acceptable values or allowable standard deviation for their qualitative laboratory results. If the potassium values were analyzed for 100 people who were considered to be of normal health, we would find that their values could be plotted on a graph to form a Gaussian curve (see above), which is sometimes known as a bell curve or normal distribution curve. In a normal curve, half of the values are greater than the mean and half of the values are less than the mean for that test result. Statistically, it has been shown that 68% of the results in a normal population will be within one standard deviation above or below the mean, and 95% will be within two standard deviations above and below the mean. Ninety-nine percent of the results will fall within three standard deviations of the mean.

Normally, when establishing acceptable ranges (also known as confidence limits) for a laboratory procedure, the quality control results must fall within a range of two standard deviations above or below the mean in order to be an acceptable result. For our potassium scenario, if this was a quality control material that was tested ten times, the results would need to fall within two standard deviations of the mean. The mean was 4.4, and the quality control results would need to be 4.4 plus or minus 0.122 to be within range. This means that the acceptable ranges for this quality control sample would be 4.28 to 4.52 mmol/L.

She also checks the maintenance logs for the instrument used for digoxin testing, and finds that the machine has not been serviced as recommended by the manufacturer. This is a very serious situation, especially because Mr. Butler may not have been the only patient affected by the instrument's performance.

The QA officer begins to make a plan to correct the issues. The first part of her plan is to have the overdue maintenance performed on the instrument. This includes cleaning the internal components of the machine and changing some tubing and reagents. After the maintenance has been performed, the machine has to be calibrated. **Calibration** is a standardization of the instrument for a specific analyte, meaning that a calibration sample (generally ordered from the manufacturer) is processed using the instrument. If the reading does not come out exactly as it should, the instrument is adjusted (according to manufacturer's instructions) until it does measure the calibration sample at precisely the predetermined level.

Once the calibration is complete, quality control assessment must be performed again. This is called **calibration verification.** The calibration and subsequent verification show that the testing instrument has been reading the digoxin specimen results several increments higher than they really were. This caused Mr. Butler's reading to appear falsely elevated; his specimen was tested once again after the calibration and verification and found to be at the high end of the expected reference range. His dosage should not have been altered. This change in medication may have put his health at risk unnecessarily.

To follow up on this incident, the QA officer documents all that has occurred. Her follow-up includes reeducation of the personnel responsible for the quality control and maintenance for the testing instrument. She develops a plan to follow up with these employees to see if the intervention and retraining were successful. The QA officer also notifies the laboratory director and any other required personnel in the organization who need to be notified of a testing system problem leading to a bad patient outcome.

If the patient's health had been negatively affected by the medication change in response to the first digoxin test result, the laboratory would be at risk for legal action. As a result of this situation, it is likely that the physician may decide to send his specimens to a different laboratory for analysis.

As this scenario draws to a close, the QA officer feels confident that the digoxin levels reported from that instrument are now accurate and reliable. She also has to go back and follow up on reported results for other patients tested close to the time that Mr. Butler had his test performed.

However, another question still exists: How well do the digoxin results from this laboratory compare to others that are performing the same test with the same instrument? This is the final step of assuring quality for the laboratory. It is like taking a standardized assessment to see if the results produced in a certain laboratory compare well to those of others across the country. **Proficiency testing** is a means to measure the performance of a laboratory externally. For laboratories performing moderate- or high-complexity testing, federal regulations require that proficiency testing is performed at least three times yearly. The laboratories are sent specimens from a provider that has been approved by their licensing agency, and these specimens must be tested exactly as a patient sample would be tested. The laboratories then report their results back to the approved provider, and they are compared to other laboratories that perform the same test using the same method. If a laboratory's results are too different from the others used for the comparison, that laboratory may be told that it can no longer offer that specific test until the problem is identified and solved. In addition, if a laboratory is not successful with its proficiency testing, there will usually be a site visit performed by those who certify or license this laboratory for operation.

For proficiency testing to be valid, it is important to stress that the samples provided are to be run precisely the same as patient samples. Laboratories are not allowed to run the samples in duplicate unless the patient samples would be treated in this way. Proficiency testing is a powerful tool, providing verification for a laboratory that their results are comparable to others performing the same test across the region or country.

By law, laboratories that perform only CLIA-waived tests are not required to participate in proficiency testing. However, they may choose to voluntarily participate. The Centers for Medicare & Medicaid Services (CMS) still recommends that some sort of external quality verification is performed periodically, whether it is through a commercial agency or informally. Informal verification of quality may include splitting a patient sample and sending some of it to another laboratory to see what its result is, or asking for another laboratory to share its patient's samples for the same reason. Many laboratories also perform internal proficiency testing in which they pool patient samples, run them numerous times to get an acceptable range for the results, and then use these to supplement their commercial quality control materials throughout the day. These patient pool specimens are usually separated into small samples and frozen until they can be tested. The internal patient quality control samples are a nice way to supplement the commercial quality control in the laboratory, but they do not help with comparisons to other laboratories unless they are shared.

There is one more recommended method to assure quality in the laboratory. This process is known as **competency testing.** Competency testing actually evaluates the personnel who perform the laboratory tests, by direct observation, or by introducing mock specimens with known values into the patient load and evaluating the results and documentation after the test is complete. This evaluation process can be especially critical with the CLIA-waived tests, as often the person who performs the test has little or no formal laboratory training, and he or she may not realize the significance of cutting corners by not following the procedure exactly as the manufacturer recommends it to be performed. Many laboratories have designed a program in which competency testing of some sort is performed annually for each employee, in which a supervisor provides samples of known results for personnel to test. If the documented results do not fall within the established ranges, then additional training needs to be integrated into the laboratory environment.

Test Your Knowledge 4-8

How is the process of calibration different from running a quality control specimen? (Outcome 4-8)

Test Your Knowledge 4-9

Is proficiency testing required for all laboratories? (Outcome 4-8)

The CMS has performed site visits at CLIA-waived laboratories across the United States over the course of the past few years and the results have been very informative. Because the regulatory requirements for laboratories that perform only CLIA-waived tests are minimal, there is concern for the quality of the services provided by these laboratories. The following is a summary of the issues identified in the site visits:

- The employees performing the tests were inadequately trained. Training must be provided by a qualified person who has knowledge of the test performance, and the actual performance of the test should be evaluated as the trainer observes.
- There were high turnover rates for testing personnel, especially medical assistants.
- CLIA-waived testing sites were performing tests that were not CLIA waived, using personnel who were not adequately educated or trained to perform the procedures. This could be a potentially dangerous situation for patients, as the procedures may not be performed correctly and the testing personnel may not be able to identify significant issues with the testing process due to lack of training.
- Many of the sites did not have the most recent manufacturer's procedural instructions for the tests they were performing. Approximately 20% of the sites reported that they did not routinely check the manufacturer's package inserts to see if there were recommended changes in the procedure. There were also cost-cutting measures in effect in some laboratories, in which urinalysis testing strips or fecal occult blood slides were being cut in half to increase the number of specimens tested per kit. This destroys the integrity of the testing process, and means that the laboratory is noncompliant with the CLIA regulations.
- Quality control testing was not performed as specified by the manufacturer by many of the sites, and units of measure were incorrectly reported with patient results.

- Problems were identified with expiration dates and storage methods for the testing kits, as well as documentation of the lot numbers and control results for the kits once they were opened. There were also issues with inadequate performance of instrument maintenance and calibration.
- Approximately 6% of the laboratories visited did not perform follow-up confirmatory tests as recommended by the manufacturer.
- Some of the sites were testing specimens that were not acceptable for that waived method. For instance, some tests are designed to be performed on urine, but the laboratory was testing plasma instead.

In 2005, the Centers for Disease Control and Prevention (CDC) published a report summarizing the results of site visits mentioned above. (You can find the link to the report in the Resources and Suggested Readings section at the end of this chapter.) In addition to the summary, the report also includes recommendations for good laboratory practices for waived testing sites. The majority of these recommendations have been included in this chapter. They include specific considerations for all three parts of the laboratory testing procedure: preanalytical, analytical and postanalytical phases.

1. Be certain that there is a written request or requisition for all tests performed.
2. Follow the written instructions for collection methods carefully, and verify what type of specimen is needed for the test ordered before collection.
3. Properly identify the patient and specimen, and store the specimen properly until the testing process begins.
4. Check the kit expiration date.
5. Know the product insert; be sure that copies are kept current at all times, and refer to them often. A sample package insert is reproduced in Figure 4-4.
6. Follow the procedure provided by the manufacturer exactly as written.
7. Run controls often, at least as often as is recommended by the manufacturer.
8. Repeat the test if there is a problem, and troubleshoot if the problem continues. Always verify that the test result is valid if there is an indicator present on the testing device.
9. Accurately record the results; use the correct units and be sure that the reference ranges reflect the population and sample type that you are reporting.
10. Sign and date your results wherever they are recorded in your laboratory setting.
11. Notify the health-care provider that ordered the tests in a timely manner. Be aware of **panic** or **critical results.** These are test results that are far enough outside of the reference ranges that they may be considered to be life threatening, and must be reported *immediately* to the ordering physician or another appropriate responsible party. When notifying the physician's office of these results, document who was notified, what was reported, and when it was reported, and sign and date this documentation.

Test Your Knowledge 4-10

What is the most important step that a medical assistant can take to ensure that he or she is performing a CLIA-waived test for mononucleosis correctly?

(Outcome 4-8)

SUMMARY

Quality assurance in the clinical laboratory is like a big umbrella covering numerous aspects of the process. A good QA plan should address issues such as education of the providers and patients concerning appropriate patient preparation before testing, initial and ongoing education of all personnel, written procedures, thorough documentation of all testing, quality control and maintenance procedures, and plans for how to follow up on problems that may develop. Quality control focuses more on the testing process itself, and includes specimen and testing kit or reagent storage, following directions, and reporting results appropriately. Other means of assuring quality in the laboratory may include calibration, external proficiency testing, and competency testing. The most important aspect of all these measures is to assure quality results for the patients we serve so that they can be cared for appropriately.

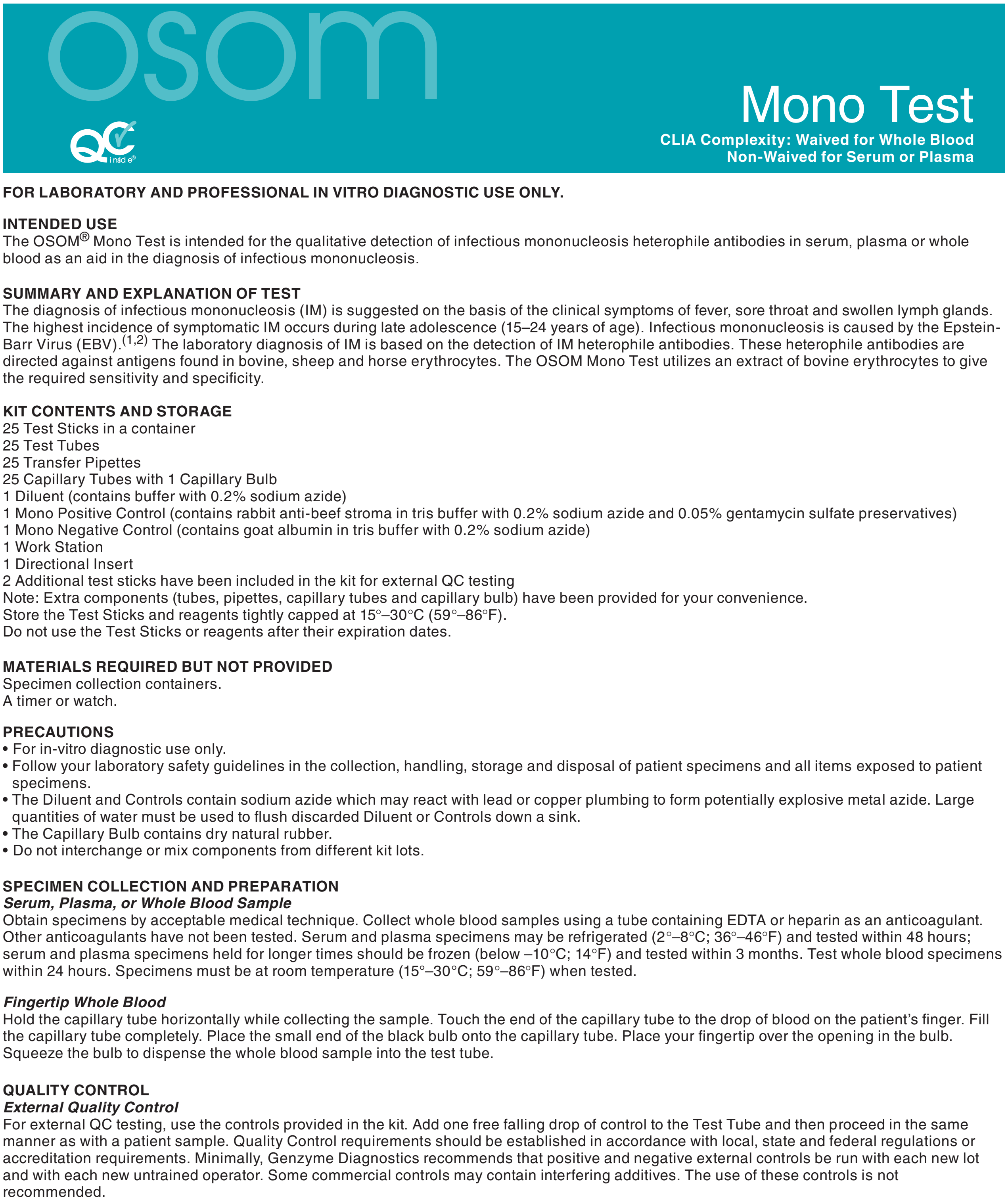

osom

QC inside

Mono Test

CLIA Complexity: Waived for Whole Blood
Non-Waived for Serum or Plasma

FOR LABORATORY AND PROFESSIONAL IN VITRO DIAGNOSTIC USE ONLY.

INTENDED USE

The OSOM® Mono Test is intended for the qualitative detection of infectious mononucleosis heterophile antibodies in serum, plasma or whole blood as an aid in the diagnosis of infectious mononucleosis.

SUMMARY AND EXPLANATION OF TEST

The diagnosis of infectious mononucleosis (IM) is suggested on the basis of the clinical symptoms of fever, sore throat and swollen lymph glands. The highest incidence of symptomatic IM occurs during late adolescence (15–24 years of age). Infectious mononucleosis is caused by the Epstein-Barr Virus (EBV).(1,2) The laboratory diagnosis of IM is based on the detection of IM heterophile antibodies. These heterophile antibodies are directed against antigens found in bovine, sheep and horse erythrocytes. The OSOM Mono Test utilizes an extract of bovine erythrocytes to give the required sensitivity and specificity.

KIT CONTENTS AND STORAGE

25 Test Sticks in a container
25 Test Tubes
25 Transfer Pipettes
25 Capillary Tubes with 1 Capillary Bulb
1 Diluent (contains buffer with 0.2% sodium azide)
1 Mono Positive Control (contains rabbit anti-beef stroma in tris buffer with 0.2% sodium azide and 0.05% gentamycin sulfate preservatives)
1 Mono Negative Control (contains goat albumin in tris buffer with 0.2% sodium azide)
1 Work Station
1 Directional Insert
2 Additional test sticks have been included in the kit for external QC testing
Note: Extra components (tubes, pipettes, capillary tubes and capillary bulb) have been provided for your convenience.
Store the Test Sticks and reagents tightly capped at 15°–30°C (59°–86°F).
Do not use the Test Sticks or reagents after their expiration dates.

MATERIALS REQUIRED BUT NOT PROVIDED

Specimen collection containers.
A timer or watch.

PRECAUTIONS

- For in-vitro diagnostic use only.
- Follow your laboratory safety guidelines in the collection, handling, storage and disposal of patient specimens and all items exposed to patient specimens.
- The Diluent and Controls contain sodium azide which may react with lead or copper plumbing to form potentially explosive metal azide. Large quantities of water must be used to flush discarded Diluent or Controls down a sink.
- The Capillary Bulb contains dry natural rubber.
- Do not interchange or mix components from different kit lots.

SPECIMEN COLLECTION AND PREPARATION

Serum, Plasma, or Whole Blood Sample

Obtain specimens by acceptable medical technique. Collect whole blood samples using a tube containing EDTA or heparin as an anticoagulant. Other anticoagulants have not been tested. Serum and plasma specimens may be refrigerated (2°–8°C; 36°–46°F) and tested within 48 hours; serum and plasma specimens held for longer times should be frozen (below –10°C; 14°F) and tested within 3 months. Test whole blood specimens within 24 hours. Specimens must be at room temperature (15°–30°C; 59°–86°F) when tested.

Fingertip Whole Blood

Hold the capillary tube horizontally while collecting the sample. Touch the end of the capillary tube to the drop of blood on the patient's finger. Fill the capillary tube completely. Place the small end of the black bulb onto the capillary tube. Place your fingertip over the opening in the bulb. Squeeze the bulb to dispense the whole blood sample into the test tube.

QUALITY CONTROL

External Quality Control

For external QC testing, use the controls provided in the kit. Add one free falling drop of control to the Test Tube and then proceed in the same manner as with a patient sample. Quality Control requirements should be established in accordance with local, state and federal regulations or accreditation requirements. Minimally, Genzyme Diagnostics recommends that positive and negative external controls be run with each new lot and with each new untrained operator. Some commercial controls may contain interfering additives. The use of these controls is not recommended.

Figure 4-4 Package insert. *Courtesy of Genzyme Diagnostics.*

Internal Quality Controls
The OSOM Mono Test provides two levels of internal procedural controls with each test procedure.

- The red Control Line is an internal positive procedural control. The Test Stick must absorb the proper amount of test material and be working properly for the red Control Line to appear.
- A clear background is an internal negative procedural control. If the test has been performed correctly and the Test Stick is working properly, the background will clear to give a discernible result.

If the red Control Line does not appear, the test may be invalid. If the background does not clear and interferes with the test result, the test may be invalid. Call Genzyme Diagnostics Technical Assistance if you experience either of these problems.

LIMITATIONS

- As with all diagnostic assays, the results obtained by this test yield data that must be used as an adjunct to other information available to the physician.
- The OSOM Mono Test is a qualitative test for the detection of IM heterophile antibody.
- A negative result may be obtained from patients at the onset of the disease due to heterophile antibody levels below the sensitivity of this test kit. If symptoms persist or intensify, the test should be repeated.
- Some segments of the population with acute IM are heterophile antibody negative.(1)

EXPECTED VALUES
A heterophile antibody response is observed in approximately 80–90% of adults and children with EBV-caused IM. This percentage drops to approximately 50% for children under four years of age.(1)

While the incidence of IM reflects wide seasonal, ethnic and geographical variation, a large epidemiological study noted that the highest incidence of symptomatic IM occurred during late adolescence (15–24 years of age)(2).

PERFORMANCE CHARACTERISTICS
A total of 439 specimens (183 serum, 176 plasma and 80 whole blood) were evaluated by two clinical labs in a clinical study. Test results of the OSOM Mono Test were compared to results obtained with a commercially available latex particle agglutination test for the qualitative determination of infectious mononucleosis heterophile antibodies. Discrepancies between the results given by the OSOM Mono Test and the latex particle agglutination test were resolved by Epstein-Barr Virus (EBV) specific serological assays. In these assays, the specific antibodies to the EBV capsid antigen (IgM) and EBV nuclear antigen-1 (IgM and IgG) were determined.

Serum Specimens: Comparative Test

		+	−
OSOM Mono Test	+	74	8*
	−	0	101

*6 out of 8 tested postitive by EBV testing

Whole Blood Specimens: Comparative Test

		+	−
OSOM Mono Test	+	30	3*
	−	0	47

*1 out of 3 tested postitive by EBV testing

Plasma Specimens: Comparative Test

		+	−
OSOM Mono Test	+	67	15*
	−	0	94

*8 out of 15 tested postitive by EBV testing

All Specimens: Comparative Test

		+	−
OSOM Mono Test	+	171	26*
	−	0	242

*15 out of 26 tested postitive by EBV testing

When compared to a commercially available latex particle agglutination test for infectious mono-nucleosis heterophile antibodies, the OSOM Mono Test showed a sensitivity of 100% and a specificity of 90.3%. The overall agreement was 94.1%.

Fifteen of the 26 discrepant samples were determined to be recent or acute EBV infections by EBV serological testing, in which case the sample was considered positive. Including the samples confirmed positive by EBV serological testing, the overall clinical study specificity of the OSOM Mono Test is 95.9% and the overall sensitivity is 100%.

POL Studies
An evaluation of the OSOM Mono Test was conducted at three physicians' offices or clinical laboratories where testing was performed by personnel with diverse educational backgrounds. Each site tested the randomly coded panel consisting of negative (5), low positive (3) and moderate positive (4) specimens for three days. The results obtained had 99.1% agreement (107/108) with the expected results.

Figure 4-4 Cont'd

TIME TO REVIEW

1. *Analyte* means: Outcome 4-1
   - **a.** A similarity between two things
   - **b.** A substance being tested
   - **c.** A method used for testing samples
   - **d.** A type of proficiency testing

2. True or False: The mean of a set of results is the average of the results. Outcome 4-1

3. A quantitative testing procedure provides results as: Outcome 4-1
   - **a.** Numbers or measurements
   - **b.** Positive or negative
   - **c.** Descriptive analysis by a pathologist

4. Which concept focuses specifically on the analytical phase of laboratory testing? Outcome 4-2
   - **a.** Quality control
   - **b.** Quality assurance
   - **c.** Documentation
   - **d.** Personnel training

5. Which of these statements describes how the details of the postanalytical phase of laboratory testing affects quality assurance? Outcome 4-3
   - **a.** The postanalytical phase of laboratory testing is not considered to be part of quality assurance
   - **b.** When results are reported promptly to the physician as part of the postanalytical process, patient care is positively affected
   - **c.** Postanalytical processes such as appropriate patient education are critical for quality patient care
   - **d.** Postanalytical processes such as performing the test procedures using the manufacturer's inserts are critical for quality patient care

6. True or False: If a medical assistant performs only CLIA-waived testing, will it be necessary for her to perform quality control? Outcome 4-4

7. When is troubleshooting performed? Outcome 4-5
   - **a.** Daily
   - **b.** Quarterly
   - **c.** As needed
   - **d.** When directed by the package insert for the test kit

8. Please provide three examples of data that must be documented when logging quality control for a qualitative test. Outcome 4-6

9. What is the purpose of calibration verification? Outcome 4-8
   - **a.** Calibration verification is performed to see if the internal quality control is working correctly
   - **b.** Calibration verification is performed to see if calibration is necessary for an instrument
   - **c.** Calibration verification is performed when a new reagent lot number is put into use for the first time
   - **d.** Calibration verification is performed after a calibration has been completed on an instrument to verify that the calibration was successful and that the instrument is ready for patient testing

10. True or False: Competency testing addresses only test results. Outcome 4-7

11. Where do laboratories obtain samples for proficiency testing? Outcome 4-7
    - **a.** From the test kit manufacturer
    - **b.** From approved providers
    - **c.** From the CMS
    - **d.** From the CDC

Case Study 4-1: An unexpected delay

A 3-year-old child is brought into the office with a potential urinary tract infection. The mother of the child goes into the restroom and helps with the collection of a very small amount of urine to run a urinalysis test. The physician wants the test run as soon as possible so that the child can be started on antibiotics, as she is very uncomfortable and cries whenever she has to urinate.

The medical assistant turns on the urinalysis instrument, and takes out the quality control material from the refrigerator. While the instrument is performing an internal electronic check, she notices that the quality control material has a printed expiration date that has already passed, but only by a couple of weeks.

1. Would it be acceptable to use this quality control material? What are potential outcomes of this decision?
2. What is one step that could be taken before informing the physician of the situation?

RESOURCES AND SUGGESTED READINGS

"CDC Report Assesses the State of Proficiency Testing"
A summary of a recent report released by the CDC addressing laboratory quality
http://www.aacc.org/publications/cln/2008/June/Pages/cover1_0608.aspx

"Clinical Laboratory Improvement Amendments Overview"
A written overview of CLIA policies and procedures, and numerous links for more specific information http://www.cms.hhs.gov/CLIA/

"CDC Report on Good Laboratory Practices for Waived Testing Sites"
Summary of site visits to CLIA-waived testing sites, as well as recommendations for quality practices. Also available as a printed document. http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5413a1.htm

"Tests Waived by the FDA From January 2000 to Present"
List of all CLIA-waived tests. Updated regularly. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfClia/testswaived.cfm

"Laboratory Quality Control"
Good information about basic quality control standards
http://www.virology-online.com/presentations/index.htm

"Good Laboratory Practice With Waived Systems"
Great information from the WA State Department of Health about laboratories that perform only waived tests
http://www.doh.wa.gov/hsqa/fsl/LQA_Home.htm

"Recommended Elements of a Laboratory Quality Assurance Plan"
Recommendations based on Environmental Protection Agency (EPA) procedures for development of a QA plan
http://www.umass.edu/tei/mwwp/files

"Abbott Diagnostics Practical Application Quality Control Guide"
Summary of the practical application of quality control principles http://www.abbottdiagnostics.com/Your_Health/Thyroid/ Testing/quality.cfm

"Westgard Rules: What Are They?"
Summary of the Westgard Rules interpretation of Levey-Jennings QC charts; the Westgard system for defining quality control limits http://www.westgard.com/mltirule.htm

Chapter 5

Legal and Ethical Issues

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

Legal and Ethical Issues in the Laboratory Environment
Law and Ethics
Legal Concepts Affecting Patient Interactions
    Consent
    Release of Information and Patient Rights and Responsibilities
    Tort Law
    Liability
    Scope of Practice
Health Information Portability and Accountability Act
    The Impact of the Health Insurance Portability and Accountability Act on the Medical Assistant
Ethics
    Professional Behaviors
Risk Management and the Medical Assistant
Summary
Time to Review
Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

5-1 Define the chapter key terms.

5-2 Explain the difference between law and ethics.

5-3 Distinguish between informed and implied consent, and describe when consent may be refused.

5-4 Express the principles presented in the Consumer Bill of Rights and Responsibilities.

5-5 Explain who has a right to a patient's medical information, and in what circumstances that right may be waived.

5-6 Compare and contrast intentional and unintentional torts.

5-7 Describe the liability assumed by a medical assistant in the laboratory, and why personal malpractice insurance is recommended.

5-8 Explain what is meant by scope of practice, and how these limitations are established for medical assistants.

5-9 Explain the acronym HIPAA, and explain how it affects the daily activities of a medical assistant in the laboratory environment.

5-10 Define *ethics* and explain the difference between personal and professional ethics.

5-11 Describe several ways that a medical assistant working in the laboratory may be part of risk management for his or her workplace.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

IX.C.1: Discuss legal scope of practice for medical assistants
IX.C.5: Discuss licensure and certification as it applies to healthcare providers

ABHES Standards

#4: Medical Law and Ethics, b: Federal and State Guidelines, f: Health Laws and Regulations
Graduates: f: Comply with federal, state and local health laws and regulations

KEY TERMS

- Assault
- Battery
- Department of Health and Human Services (HHS)
- Disclosure
- Duress
- Emancipated
- Ethics
- Expressed consent
- Fraud
- Health Insurance Portability and Accountability Act (HIPAA)
- Implied consent
- Informed consent
- Intentional tort
- Laws
- Liability
- Libel
- Malfeasance
- Malpractice
- Misfeasance
- Negligence
- Nonfeasance
- Patient identifier
- Privacy Rule
- Protected health information (PHI)
- Reportable condition
- Risk management
- Scope of practice
- Slander
- Standard of care
- Statutes
- Tort
- Unintentional tort

LEGAL AND ETHICAL ISSUES IN THE LABORATORY ENVIRONMENT

Most of us understand that it is important to obey the law and that there may be consequences if we make the wrong choices. However, in our daily lives, we realize that there are factors other than laws that affect our actions. These factors may include our personal values and ethics. In addition, as health-care employees, medical assistants must adhere to a set of professional laws and ethics to perform our duties as we should. In this chapter, you will learn the difference between law and ethics, and examine how these concepts define the professional actions of those employed in a health-care environment.

LAWS AND ETHICS

Laws are essentially rules of conduct that have been created by government at the local, state or national level. An established rule or law may also be known as a **statute.** Laws are enforced by appointed authorities. For instance, the Occupational Safety and Health Administration has created rules to uphold the federal laws developed to safeguard employees. Generally, laws protect the public by prohibiting activities that may be harmful, and establishing some form of consequences that would discourage those that continue to pursue these activities. Health-care laws are designed to protect the rights of the patients, the employees, or the employers.

Ethics are like laws because they dictate or determine appropriate behavior for individuals. They are different from laws because they are not created by governmental agencies. Personal ethics are behavioral guidelines that an individual develops throughout a lifetime, guided by personal values and morals, and influenced by upbringing. Professional ethics are developed by a well-recognized professional organization, with the anticipation that those working in that profession will use them as a guide for appropriate conduct.

Test Your Knowledge 5-1

Name one difference between laws and ethics.

(Outcome 5-2)

LEGAL CONCEPTS AFFECTING PATIENT INTERACTIONS

There are numerous laws and statutes that play a role in a health-care setting. In addition to federal laws, each state has a medical practice act that dictates the delivery of health-care services in that state. Medical assistants act as agents of the physician office or laboratory where they are employed, which means they have a responsibility to abide by the laws that dictate their actions as they perform health-care procedures and deal with patient information. In order to follow the established rules, it is vital to learn some general legal concepts.

Consent

The law requires that there is consent, or permission, given before any treatment occurs. The consent may be given verbally, in writing, or nonverbally by behavior. **Implied consent** means that the patient's actions send a message of agreement to the treatment offered. For instance, when a patient offers his or her arm for a blood draw, the consent for the procedure is implied by this action. Implied consent also dictates care that occurs in an emergency situation. If the patient is not able to refuse consent, or if there is no legal document stating that emergency care should not be offered, emergency medical assistance may be rendered. It is assumed (implied) that the patient would want to be saved.

Informed consent is necessary whenever the treatment involved includes an invasive surgical procedure, experimental drugs, or any treatment that could be a significant risk to the health of the patient. This may also be known as **expressed consent.** Informed consent requires that the patient is advised of the following details:

- The name of the procedure to be performed and the health-care provider that will perform the procedure
- Risks of the procedure
- Risks involved if the procedure is *not* performed
- Alternatives to the treatment offered

It is essential that the patient understand the components printed on the consent form, as it is a legal document. The patient must sign and date the consent form, and this must be witnessed. As in all situations concerning consent, the responsibility for explaining the procedure and the risks belongs to the health-care provider (such as a physician or nurse-practitioner) but the medical assistant may be asked to verify that the form is filled out correctly, as well as witness the signature of the patient.

Minors generally are not allowed to sign an informed consent. There are exceptions to this rule, which may vary from state to state. The exceptions generally involve requests for birth control or treatment of sexually transmitted infections. There are also exceptions for **emancipated** minors, and those in the armed forces who are still under the legal age of consent. Emancipated minors have been declared by the court to be independent and capable of making their own medical decisions. All other minors must have a parent or legal guardian sign the consent form for them.

A patient has the right to refuse consent at any time. Patients may decline treatment for any reason, based on personal beliefs, cultural differences, or financial considerations. If consent is refused, the patient should fill out and sign a form declining the procedure. This documentation can be very important if there is a negative outcome based on the patient's decision to refuse treatment.

Test Your Knowledge 5-2

When a patient offers an arm for a blood draw, what kind of consent is he or she offering?

a. Informed consent
b. Implied consent
c. Deferred consent
d. None of the above

(Outcome 5-3)

Release of Information and Patient Rights and Responsibilities

Every patient who seeks medical care has certain rights and responsibilities. Medical personnel should always be conscious of these principles when dealing with patients, so that they can stay within appropriate boundaries in their interactions. Patient responsibilities include providing the physician with accurate information about his or her medical history and current condition, following the physician's instructions for treatment and follow-up care, and providing appropriate compensation to the health-care provider for services provided.

Patient rights have been addressed numerous times through the years. In 1998, the President's Advisory

Commission on Consumer Protection and Quality in the Health Care Industry issued a Consumer Bill of Rights and Responsibilities. This document addresses seven distinct areas of patient care:

1. *Information disclosure:* Patients have the right to know the details of their health-care plan, their health-care professionals, and the facilities they may use for care. This means that an interpreter must be provided if needed for understanding.
2. *Choice of providers and plans:* Patients have the right to evaluate different providers and/or health coverage plans to make the best choices for their needs.
3. *Access to emergency services:* If an emergency need arises, a patient has the right to receive services wherever needed without prior authorization. (However, this does not mean that the patient will not have to pay for these services if they are not covered specifically by their health-care plan.)
4. *Participation in treatment decisions:* All patients need to know about their treatment options and have the right to participate in decisions about their medical care. This also means that they have the right to refuse care if they choose.
5. *Respect and nondiscrimination:* All patients have the right to receive respectful treatment from their health-care providers that is considerate of their feelings and without discrimination.
6. *Confidentiality:* Patients have the right to expect their health-care information to remain confidential.
7. *Complaints and appeals:* All patients have the right to an objective review of any complaints they have about their health-care provider or health-care plan. This can include any part of their experience, including billing, accessibility of the building for those with special needs, operating hours, and so on. Of course, the right to review and appeal also pertains to the quality of care experienced.

Even though patients have the right to expect their health-care information to remain confidential, there are limitations. The physical medical record (where the information is documented; also known as “the chart”) belongs to the health-care provider, but the information in the record belongs to the patient. Patients have the right to request their medical information, and they also have the right to control who has access to their medical records in most situations. The information contained in the medical record may be requested by numerous parties, such as insurance companies, family members, and legal representatives. Later in this chapter we will learn more about the law which details how personal information may be protected.

There are some situations where the **disclosure** (sharing) of the patient’s information is required, even without consent. For instance, if a patient is diagnosed with a **reportable condition,** state and federal laws require that the laboratory or the health-care provider contact the local health department or the Centers for Disease Control and Prevention (CDC). Reportable conditions are diseases or conditions that have been deemed to be dangerous to the general public health. The statistics accumulated by the state or federal agencies allow the prevalence of the disease in the population to be tracked. Examples of reportable conditions include HIV infection, Lyme disease, cholera, hepatitis, measles, and gonorrhea infection. If a laboratory test is positive for these conditions, the result must be passed on to the appropriate authorities as soon as possible. The patient does not have the right to stop this report process, although in some situations he or she may have the right to request anonymity to protect his or her identity.

Test Your Knowledge 5-3

Lily Lee is moving to another city to attend college. She would like to get a copy of her medical records to take with her as she establishes a new physician relationship in her new home. Does she have the right to a copy of her medical record? Is it appropriate for her to take the entire original record from the office?

(Outcome 5-4 and Outcome 5-5)

Test Your Knowledge 5-4

If someone is assaulted, was he or she touched or physically injured? (Outcome 5-1)

Tort Law

Torts are wrongful acts that are committed against a person or property that result in harm. The harm caused by the tort may be physical or emotional. This type of act does not include those that involve violation of a contract. Torts may be classified as intentional or unintentional. In the case of an **intentional tort,** it must be apparent that the action was deliberate and intended to cause harm. Intentional torts include the following:

- **Assault:** An assault involves threatening an individual, or making a person fear harm.

- **Battery:** Battery is the act of touching another without that person's permission. A person may be committing battery even if the touch does not cause harm.
- **Duress:** If a patient is forced or coerced into an act (such as signing a consent form or undergoing a procedure against his or her will), that person may claim duress.
- **Libel:** If a health-care professional makes a false written statement about a patient (or someone else), he or she may be accused of libel. This is especially a problem when the false statement causes a problem with the victim's reputation or employment.
- **Slander:** Slander is false statement made about someone else orally.
- **Fraud:** Fraud is a dishonest or deceitful act that is performed with the intent to hide the truth.

Unintentional torts are situations in which the harm was the result of an accident. The accuser must believe that the employee was performing his or her job in good faith when the patient was harmed, and the incident was not deliberate. **Negligence** is often involved in unintentional torts, as it means that a professional did not take reasonable care to prevent harm. The term **malpractice** is used if a health-care professional is shown to be negligent in his or her duties. Malpractice accusations are the most common reason for claims against health-care professionals, and the legal decisions are based on the **standard of care** (minimum safe professional conduct) for a reasonable professional with the same training who might be facing a similar situation. Malpractice may be classified in one of three ways:

1. **Malfeasance:** In this situation, the treatment provided was wrong and unlawful.
2. **Misfeasance:** If treatment was provided, but it was performed in an improper manner, it is known as misfeasance.
3. **Nonfeasance:** When the health-care professional fails to perform a necessary act or delays treatment excessively, the act is known as nonfeasance.

Test Your Knowledge 5-5

What is the difference between an intentional and unintentional tort? (Outcome 5-6)

Liability

Health-care professionals assume a certain amount of **liability** in everything that they do, meaning that they are legally obligated to perform their duties to the best of their ability. Liability exists in all areas of our lives: driving a car or while walking the dog, for example. We are responsible for our actions, and obligated to drive the car to the best of our ability, as well as keep our dog on a leash and under control while on the walk. For health-care professionals, their liability insurance is actually malpractice insurance, as they may be held responsible for an intentional or unintentional tort if they make a mistake or they don't perform their jobs to the best of their ability. Anyone who works in a medical office is liable for all interactions with the public, whether performing an invasive procedure, taking a blood pressure reading, or participating in a phone call with a patient. A malpractice suit may be filed against any or all of the employees in an office or laboratory.

In most health-care facilities, the malpractice insurance coverage for the physician or health-care provider generally includes coverage for the medical assistant working for them. Many medical assistants don't pursue their own malpractice insurance coverage. This may be a mistake. If a patient sues the practice (specifically naming each individual involved), for malfeasance, for example, the best outcome for the physician may be to settle out of court. This means that the incident will still be on the record for the medical assistant, who was not part of the decision to settle in this way. For medical assistants (or other allied health-care professionals) who work in the laboratory setting, it is especially important to have appropriate personal coverage. Even though the physician may have malpractice insurance that appears to be adequate, it is possible that when drawing blood or performing laboratory tests outside of the direct supervision of the physician, the medical assistant may be found to be solely responsible for malpractice. In addition, if a physician is sued for malpractice because of a venipuncture or other procedure performed by his or her medical assistant, the physician can even file suit against the medical assistant for financial damages incurred as a result of the incident.

Test Your Knowledge 5-6

Is a medical assistant working in a laboratory liable for his or her actions while working with patients? Is it possible for him or her to be sued for malpractice? (Outcome 5-7)

Scope of Practice

Scope of practice is a term used to describe the activities that an employee in a specific profession can and cannot perform. These limitations are different for each of the health-care professionals present in a physician office or laboratory environment. The scope of practice is dictated by licensure and training, and may vary according to the state in which a professional offers his or her services.

Medical assistants are not licensed, meaning that they are required to work under the license of another qualified health-care professional, such as a physician, nurse-practitioner, or physician's assistant. These health-care professionals have a responsibility to train the medical assistant and to establish his or her duties and limitations within the practice. Many states have statutes in place that limit the duties that a medical assistant may perform. Although certification is not mandatory, medical assistants may take a comprehensive examination to become nationally certified by the American Association of Medical Assistants or registered through the American Medical Technologists. In addition, some states require very specific certification or registration for allied health professionals who perform invasive procedures.

Educational programs for medical assistants vary. Some may be just a few months in length, whereas other programs offer a 2-year associate degree. In addition, each state may have different laws that dictate the tasks that may be legally performed by medical assistants. This means that the scope of practice for a medical assistant is based on his or her education and training, as well as the limitations that may be in effect by the state in which he or she practices.

In general, the scope of practice for medical assistants includes the following limitations:

- Medical assistants are not able to make independent medical assessments or diagnose, or give advice independently.
- It is the physician's responsibility to be aware of the medical assistant's scope of practice, whether based on state laws or the medical assistant's training.
- Medical assistants are not able independently to write prescriptions or give out medication samples without specific documented orders from a qualified health-care professional.
- Medical assistants cannot independently interpret test results or treat patients.

In most states, the following are tasks that medical assistants are allowed to perform:

- Assisting with various clinical and patient care procedures
- Carrying out administrative procedures such as answering phones, working with charts, and performing billing and coding
- Obtaining medical histories
- Explaining treatment procedures to patients (within the guidelines established by the facility)
- Preparing and assisting with various routine and specialty examinations
- Collecting specimens, including venipuncture and capillary puncture
- Performing electrocardiograms (ECGs)
- Administering medication
- Assisting with prescription refills
- Performing Clinical Laboratory Improvement Act–waived laboratory tests or those of moderate complexity
- Assisting with various disease management programs, such as tracking laboratory results for a patient on medication
- Administering injections

Test Your Knowledge 5-7

Is the scope of practice the same for medical assistants anywhere in the United States? (Outcome 5-8)

HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT

Earlier in this chapter, we discussed the Consumer Bill of Rights and Responsibilities. One component addressed the confidentiality of a patient's health record. This is a very complex issue, and it has been addressed in various ways, as the protection of personal health information is one of the most important duties of a health-care professional. It is impossible to establish a good working relationship with a patient if there is no trust in the professional's ability to keep private information secure. Patients share information with those in health care that they do not divulge to anyone else, and the expectation is that this information will only be shared if it is absolutely necessary for patient treatment.

In the early 1990s, it became evident that there were areas of health care that needed reform. Congress passed the **Health Insurance Portability and Accountability Act (HIPAA)** to provide guidelines for this reform. This act was originally designed to improve the efficiency and effectiveness of the health-care system by focusing on the following areas:

- Improvement of the health benefit options for covered workers who leave or change jobs
- Standardization of electronic health data

- Creation of unique health identifiers for the various entities involved in health-care services
- Development of security standards for handling **protected health information (PHI),** which includes anything in the patient's health record that could be used to identify the patient

One of the standardization methods for handling health data was to increase the use of computers for transmission of health information. This created more concerns about the unauthorized access to personal information that may occur with electronic transmission. To address these confidentiality issues, the **Department of Health and Human Services** (**HHS**) developed the **Privacy Rule,** which is an expansion of the original HIPAA mandate that applies to confidentiality of health information, and the rights of patients to control access to their own records. This rule went into effect April 2003, and it carries significant monetary consequences for noncompliance. This is the first federally mandated set of rules for the protection of health information.

Test Your Knowledge 5-8

Does HIPAA impact only confidentiality of medical information? (Outcome 5-9)

The Privacy Rule outlines very specific precautions that are to be practiced in every health-care facility. There are several standard provisions that all medical assistants should keep in mind, regardless of their specialty of practice:

1. Any document that includes a **patient identifier** must be treated as protected health information. These identifiers include demographic information such as name, address, and phone number, but also include photos, Social Security numbers, birth dates, and other types of unique information that may be linked back to that specific patient. Documents must have all patient identifiers removed before they can be discarded, or they must be destroyed in such a way that this information is not available after disposal.
2. Patients have a right to know how their PHI has been shared, even if it was necessary for treatment.
3. Patients have a right to review and/or to request copies of their medical records. This law does not state that they have the right to walk out of an office with their physical chart, nor does the office have to allow review or copies to be available at any time. A patient can submit a request for the records, with the understanding that it may take some time to process the request and copy the records.
4. Health information cannot be shared with others without the patient's consent. This does not apply to reporting that may be required by law.
5. When any patient information is shared, it is important that it is limited to data that are essential for that situation. This is especially important when dealing with labor and industry claims for injured employees or insurance claims, as the health-care professional should include only records about a specific treatment if it is in question, but not the entire chart.
6. The Privacy Rule also requires every office to have a privacy notice and to train all employees about HIPAA and the Privacy Rule. All patients must receive a copy of the office's privacy notice, and patients must give written consent or permission to disclose their health-care information. A privacy or HIPAA compliance officer must be assigned for every institution to oversee the process within that office.

The HIPAA regulations still allow health-care providers to share information with family members who are directly responsible for the care of a patient, or to other providers or members of the health-care team (such as radiology services, laboratory professionals, etc.) when it is necessary for patient care. Also, the regulations recognize that sometimes it is impossible for oral privacy to be assured in all areas of a practice; there has to be evidence that this has been addressed to the full extent of the facility.

The Impact of the Health Insurance Portability and Accountability Act on the Medical Assistant

The privacy section of the HIPAA regulations may affect phone conversations with or about patients, discussion of patient issues within earshot of the waiting area in a clinic, or the visibility of patient information to visitors in an office. Laboratory requisitions that contain PHI cannot be left where patients may have access to them, computer screens should not be visible to patients as they walk to and from the treatment area, and patient schedules that contain names should not be posted where they are visible by those who are being treated. Fax numbers used to transmit laboratory results must be carefully monitored, as these results need to end up in a secure location once received. Complications arise when protected health information is shared via e-mail, although this method of transmitting information may be more efficient for patient care within a large organization or between providers. The need for electronic security measures is great. Also, many clinics have changed the way they call patients from the waiting room for treatment; they now call the patients by the first or the last name, but not both. It is the responsibility of the designated HIPAA compliance officer to decide the specific methods used by an office for compliance with the regulations, and to ensure that appropriate training is provided for all employees.

It is also illegal to gossip about patients, even if it is in a private area where other patients cannot hear. Discussions about patient care are to be limited to those that are necessary for providing quality health care. As health-care professionals, medical assistants are not allowed to share any protected health information with family or friends. Sometimes this is difficult, as health-care professionals may have knowledge about a family friend or relative that others would want to know. It is illegal to share this information, so care must be taken to keep it private.

Another example of illegal action is idle searching of the database within a health-care facility to see who might have had services provided. There were recent reports in the news of employees who lost their jobs and of facilities that were fined significantly when an employee searched for the records of a celebrity rumored to be in a certain facility. Even if this information is not shared with anyone else, it is a HIPAA violation to perform the search. It can (and will) be tracked back to the responsible individual.

ETHICS

Most of the information in this chapter is based on laws and legal terms that dictate the actions of health-care professionals. But there are also ethical decisions that must be made every day in order to interact appropriately with patients, coworkers, and other members of the health-care team. Ethics are standards that define our actions, but they are not created or enforced by governmental agencies. As was explained at the beginning of this chapter, there are two types of ethics that define our actions while at work. We must adhere to a set of personal ethics that are defined by our own sense of what is right and wrong. Professional ethics also define our actions. These are based on the standards of our professions, and although they are not laws, there are consequences to unethical behaviors.

Personal ethics begin to develop in childhood. We generally understand what our parents and community define as acceptable behavior by the time we enter grade school. These are based on our morals and values, and become refined as we approach adulthood. Personal values such as the desire to help others will make a health-care professional much better at performing his or her job. Empathy, honesty, responsibility, and respect for others are also part of our personal ethics; this is often known as our *work ethic*. Remember, if there is something that you are doing at work that you are not proud of, it is probably unethical at the personal level for you.

Professional Behaviors

Professional ethics are standards established by the profession, and usually sponsored by a professional organization that represents the profession. Physicians have several codes of ethics that define their actions, such as the Hippocratic Oath and the Code of Medical Ethics defined by the American Medical Association. One ethical theme present in both codes is the show of respect for patients and to show compassion for those served, while staying within legal limits of practice. By utilizing these standards, physicians may remind themselves why they chose this profession, and also remember what their patients may expect from them.

The American Association of Medical Assistants also has a code of ethics (Box 5-1). These guidelines also include respect for patients and the laws that guide the medical assisting profession, as well as service to the community and lifelong learning. At a minimum, professional ethics for medical assistants must put patients first by honoring their right to confidentiality and by being honest in all professional situations. If this code is supported by a strong set of personal ethics, a medical assistant will be a valuable asset to the profession.

Remember that there are ethical dilemmas that develop. Many times there may be a situation that arises that is not illegal, but it is unethical. This may include situations in which the medical assistant does not agree with the actions of the employer, even though these actions are not illegal. Usually these issues involve disrespect for patients or a poor quality of care due to the attitude of the physician. The medical assistant will need to make choices about whether she can continue employment with this practice.

Test Your Knowledge 5-9

List two professional standards included in the AAMA Code of Ethics. (Outcome 5-10)

BOX 5-1 American Association of Medical Assistants code of ethics

American Association of Medical Assistants (AAMA) Code of Ethics

Members of AAMA dedicated to the conscientious pursuit of their profession, and thus desiring to merit the high regard of the entire medical profession and the respect of the general public which they serve, do pledge themselves to strive always to:

A. Render service with full respect for the dignity of humanity;
B. Respect confidential information obtained through employment unless legally authorized or required by responsible performance of duty to divulge such information;
C. Uphold the honor and high principles of the profession and accept its disciplines;
D. Seek to continually improve the knowledge and skills of medical assistants for the benefit of patients and professional colleagues;
E. Participate in additional service activities aimed toward improving the health and well-being of the community.

POINT OF INTEREST 5-1

Medical assisting scope of practice

There is a lot of confusion among medical assistants as well as physicians, nurse-practitioners, physician assistants, and other health-care providers about the scope of practice for the medical assistant. The confusion is well earned; the scope of practice for each state differs widely. For those medical assistants who move from one state to another, it can become even more difficult to know what they are allowed to do in any given state.

Governing Agencies

Just finding the regulations for each state can be difficult. A good resource may be the program director for the local CAAHEP or ABHES accredited medical assisting programs. These program directors should have up-to-date information about the scope of practice for the medical assistants in their area, and should be able to provide information about which governing body oversees medical assistants in their state. In many states the department of health oversees the unlicensed members of the health-care teams. In other areas, medical assistants may be part of the state nursing board.

Registration or certification

Medical assistants are nonlicensed members of the health-care team. There is no license available in any state for medical assistants. However, most states do require some sort of registration process for the medical assistants in their jurisdiction. This may include a testing process by which they are registered and certified to perform in that state, or the state may require only that the medical assistant fill out certain paperwork and submit a fee to become registered. Some states require the medical assistant to bear the burden of registration fees, whereas other states require that the medical assistant be registered each time she or he changes employers and the employer is responsible for the fee.

Many medical assistants are under the impression that if they become Certified Medical Assistants through the AAMA or Registered Medical Assistants through the American Medical Technologists that this will cover them regardless of their state of residence and that these credentials somehow expand their scope of practice. This is not true. Many states will grant registration (or certification) if the medical assistant is nationally certified, or they may require those who are not nationally certified complete a state certification process. However, this varies from state to state, so it is very important to find out what the regulations are in your state of residence. Usually the office manager or physician you are working for knows the rules, but occasionally they have decided not to work with the state to register their medical assistants. This is not a safe environment to work in, as the medical assistant may be operating outside of his or her scope of practice and unlawfully performing tasks, which place him or her at risk for legal actions.

Allowable tasks

Every state has limitations for the tasks that can be performed by the medical assistants. The following are some of the tasks that may be limited:

- **Invasive procedures:** Some states allow medical assistants to draw blood, but this task is still limited in a few areas. Many states (e.g., New York) do not allow medical assistants to administer medication through any route, including by injection. It is important to clarify this before performing injections or venipuncture.

Continued

- **IV access:** Most states limit the IV access of nonlicensed health-care personnel. In some areas, the medical assistant is allowed to stop the IV or remove the catheter when it is no longer needed, but most areas prohibit much more than this.
- **Placement of urinary catheters:** This is an area where many medical assistants are currently operating outside of their scope of practice. Most states prohibit the placement of a urinary catheter by a nonlicensed health-care professional.
- **Administering oral medications:** Although medical assistants are allowed to give injections in most states, the administration of oral medication may still be limited. Washington State, for example, has a list of oral medications that can be administered by health-care assistants, which include medical assistants.
- **Working in a hospital setting:** Some states prohibit medical assistants from working in a hospital or inpatient setting if they are performing clinical procedures. This decision is based on the presence (or absence) of health-care providers in the vicinity while invasive procedures are performed.
- **Suture removal:** Most states allow medical assistants to remove sutures, but there are limitations in a few areas.
- **Botox administration or assisting with other cosmetic procedures:** Medical assistants generally cannot operate laser equipment or administer Botox or other injections that are not considered to be medications.

RISK MANAGEMENT AND THE MEDICAL ASSISTANT

After reading through this chapter, it may appear that the legal aspects of the medical profession are overwhelming. There is a lot of information to remember, but as a medical assistant, there are many things that you can do to avoid risk of harm to yourself, your patients, and your practice. This concept is called **risk management.** These risks may be legal in nature, as in a case in which a malpractice suit is filed for poor quality of care; financial in nature, as in a case in which a laboratory or physician office chooses not to accept patients with specific insurance coverage because of the potential for low levels of reimbursement; or physical in nature, as in a case in which a medical assistant might not use the appropriate technique to perform a capillary blood draw, or when an office staff doesn't notice the puddle inside the front entrance to the clinic and a patient slips and falls.

Because a medical assistant may be the member of the health-care team in an office or a laboratory who spends the most time with a patient, it is important that he or she keep the following risk management techniques in mind to avoid potential harm or legal action:

- Act within the defined scope of practice
- Use all available training to be certain that techniques are performed with the highest level of skill possible
- Use current, up-to-date equipment as available
- Perform regular performance checks on any testing equipment used for patient samples
- Follow the manufacturer's instructions exactly for all laboratory testing performed
- Consult resources (other employees or reference materials) if you have questions
- Verify the identity of every patient before you begin interactions
- Chart when patients call to inquire for test results, and when they are notified of the results
- Be certain that all laboratory results are handled in a timely manner
- Practice confidentiality at all times
- Always practice universal precautions and keep your environment medically aseptic
- Treat all patients with respect and overcome personal biases to treat them all equally
- Act as a chaperone when needed for procedures
- Obtain informed consent when necessary
- Be vigilant for potential safety hazards in the clinic environment
- Document promptly and thoroughly
- Log all phone call correspondence
- Remember not to criticize other members of the health-care team or other health-care providers
- Do not give advice that has not been approved by the office policy
- Do not be afraid to ask for help if you are feeling overwhelmed by the number of patients you are assisting or by the tasks at hand.

Test Your Knowledge 5-10

After drawing the blood from a patient, medical assistant Krissy Clark takes the time to document the blood draw site, the date, and the time of the draw on the laboratory requisition. Is this an example of risk management? (Outcome 5-11)

SUMMARY

The majority of the time, the relationship between a patient and the health-care team is an amicable one in which both parties are satisfied with the care provided. However, there are unfortunate situations that may arise as a patient is treated that damage this relationship. Sometimes these result in charges of malpractice, and sometimes they may require legal actions because a member of the health-care team ignored the legal guidelines that define their scope of practice. In order to protect ourselves from these situations, it is important to stay insured as a means of protecting our financial security. It is also important to be aware of the legal boundaries for each profession. Ethics play a role as well, as not all inappropriate actions in a medical office or laboratory are illegal. Risk management is a way to provide care defensively; avoid problems before they start!

TIME TO REVIEW

1. What is a statute? Outcome 5-1
   - a. A personal set of rules dependent on an individual's sense of right and wrong
   - b. A law established by the legislative branch of a government entity
   - c. Behavioral standards established by a professional organization
2. If a newspaper article prints information about a local physician that is untrue, what legal term might be used to describe this action? Outcome 5-1
3. What is the meaning of the acronym HIPAA? Outcome 5-1
4. Who establishes laws that affect health-care organizations? Outcome 5-2
   - a. Professional organizations
   - b. Governmental agencies
   - c. Local groups of practitioners
   - d. Local health departments
5. The________________________ __________________ is responsible for explaining the details of a procedure to the patient before the informed consent paperwork is signed. Outcome 5-3
6. Which of these principles are not addressed in the Consumer Bill of Rights? Outcome 5-4
   - a. The right to insurance coverage
   - b. The right to emergency services when needed
   - c. The right to be treated without discrimination
   - d. The right to an interpreter if needed
7. What is a reportable condition? Outcome 5-5
8. True or False: It is not necessary for a medical assistant to carry her own malpractice insurance. Outcome 5-7
9. Referring to the medical assisting scope of practice explained in this chapter, which of these duties would be outside the guidelines? Outcome 5-8
   - a. Offering medical advice based on personal experiences
   - b. Documentation of procedures performed
   - c. Assisting with a Pap smear specimen collection
   - d. Scheduling appointments
10. True or False: It is a HIPAA violation to fax medical information to a health-care specialist who is functioning as a consultant for a patient's complicated medical condition. Outcome 5-9
11. Personal ethics are defined by: Outcome 5-10
    - a. Professional standards of the profession
    - b. Local statutes
    - c. Experiences and guidance beginning in childhood
    - d. HIPAA
12. True or False: Risk management should be a part of the routine for everyone working in the health-care environment. Outcome 5-11

Case Study 5-1: Too much talk

Amy Conner is the receptionist who greets patients at the front desk for a family practice physician. A 17-year-old young man comes in to have his blood drawn a few days after seeing the physician. He enters the reception area by himself, and gives Amy his requisition. She notices that in addition to several routine chemistry tests, the physician has ordered an HIV screen to be performed. Amy puts the laboratory orders into the computer system, and hands the labels to the medical assistant to perform the blood draw.

After the young man leaves the facility, Amy steps into the hallway behind the reception area and starts

Continued

a conversation with the medical assistant. She tells the medical assistant how sad it is that an HIV test has to be done on someone so young, and expands her conversation to include all the potential ways that he could have contracted HIV. After a few minutes she hears someone at the front desk and steps back to her work area.

The 17-year-old patient is now at the front desk again, with his mother standing next to him. He appears to be uncomfortable, and the mother starts to ask questions about the conversation that Amy was having in the hallway.

a. Are Amy's actions in violation of the law?
b. Should Amy assume that the mother knows all about the blood tests that were ordered for her son?
c. How should this situation have been handled?

RESOURCES AND SUGGESTED READINGS

"Office for Civil Rights—HIPAA"
Detailed reference for the HHS Privacy Rule http://www.hhs.gov/ocr/Health InformationPrivacy

"What Is Ethics?"
Article published by Santa Clara University about the meaning of ethics, and how personal and professional ethics differ http://www.scu.edu/ethics/practicing/decision/whatisethics.html

"Quidel Rapid Diagnostic Products"
Links to the Quidel point-of-care test kits, including CLIA-waived tests. http://www.quidel.com/products/product_list.php?cat=1&by=brand&group=1

"Reportable Diseases"
Information about reportable diseases that must be reported to the local authorities or the Centers for Disease Control and Prevention http://www.nlm.nih.gov/medlineplus/ency/article/001929.htm

Chapter 6

Laboratory Equipment

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

Laboratory Equipment
  Microscopes
    Compound Microscopes
    Other Types of Microscopes
    How Do Medical Assistants Use Microscopes?
  Centrifuges
    Centrifuge Maintenance
    Centrifuge Safety
  Laboratory Refrigeration
  Incubators
Equipment Used for Automated CLIA–Waived Testing
  Testing Methodology
  Instruments Used for Chemical Testing of Urine Specimens
  Instruments Used for Coagulation Testing
  Instruments Used for Chemistry Testing
  Instruments Used for Hemoglobin Measurements
  Other Hematology Instruments
  Glassware and Other Miscellaneous Laboratory Equipment
Summary
Time to Review
Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

6-1 Define the key terms.

6-2 Describe the different types of microscopes introduced in this chapter.

6-3 Identify the different parts of a compound microscope.

6-4 Describe maintenance procedures for a compound microscope.

6-5 Explain how to correctly focus a microscope to examine a specimen on a slide.

6-6 Explain the role of a medical assistant in a physician office laboratory that performs microscopic examinations of specimens.

6-7 Explain how to operate and maintain a centrifuge.

6-8 Describe how temperature must be monitored for a laboratory refrigerator.

6-9 Explain why an incubator is sometimes used in a laboratory environment.

6-10 List several advantages of providing CLIA-waived testing in a physician office laboratory.

6-11 Demonstrate understanding of the basic principles used for measurement in laboratory instruments.

6-12 Explain why automated urine analyzers are used.

6-13 Describe similarities between the different types of coagulation testing instruments available for CLIA-waived testing.

6-14 List several types of instruments used for CLIA–waived chemistry testing.

6-15 Describe the operation of an instrument used to measure hemoglobin in a whole blood specimen.

6-16 List the types of hematology instruments presented in this chapter.

6-17 Describe some of the types of glassware found in the laboratory, and how to correctly measure the liquid level in glassware.

CAAHEP/ABHES STANDARDS

None

KEY TERMS

- Agglutination
- Anemia
- Aperture
- Arm
- Balanced
- Base
- Beaker
- Binocular
- Centrifugal force
- Centrifuge
- Cholesterol
- Coarse adjustment knob
- Compound microscope
- Condenser
- Cover slips
- Cylinder
- Diluent
- Dissecting microscope
- Electrical impedance
- Electrolyte
- Electron microscope
- Eyepiece
- Fecal occult blood
- Fine adjustment knob
- Flask
- Glass slides
- Glucose
- Glycosylated hemoglobin monitoring
- Graduated cylinder
- Hemoglobin
- Incubator
- International normalized ratio (INR)
- Iris diaphragm
- Laboratory thermometer
- Light source
- Lysed
- Meniscus
- Monocular
- Objective lenses
- Ocular lens
- Oil immersion lens
- Pipette
- Plasma
- Prothrombin time test
- Protime
- Reagent strips
- Revolving nosepiece
- Rotations per minute (rpm)
- Serum
- Spectrophotometer
- Stage
- Stereo microscope
- Tachometer
- Water baths

LABORATORY EQUIPMENT

Laboratories use many types of instruments to process, store, and analyze specimens. As the complexity and variety of tests offered by a laboratory expand, so do the number of instruments in use. Because medical assistants are generally working in laboratory environments performing Clinical Laboratory Improvement Act (CLIA)–waived tests, they don't usually use a wide variety of instruments. This chapter focuses on equipment commonly used for specimen processing and storage, as well as performance of CLIA-waived testing procedures. Selected instruments used for laboratory tests of moderate complexity are also presented.

Microscopes

There are several types of microscopes available. These vary by the amount of magnification, the manner by which the image is viewed, and the light (energy) source used. A typical laboratory setting will use a **compound microscope,** and some laboratories may also use a **dissecting microscope,** especially if the laboratory performs blood bank procedures. Specialty laboratories may also have personnel trained to operate an electron microscope to create three-dimensional images of the specimens. All microscopes provide magnification so that microscopic structures or organisms that are too small to be seen with the naked eye can be identified or quantified in a specimen. Sometimes it is necessary to apply stain to a specimen before viewing, and sometimes the specimen is viewed in its natural state.

Compound Microscopes

The compound microscope is the most common type used in a general laboratory setting. The magnification for this type of microscope is compounded, or increased by two different sets of lenses. The energy source is a replaceable bulb that sits in the base of the microscope and shines very brightly. Adjustments may be made for the amount of light used for viewing, as well as the strength of the magnification, which makes this microscope very versatile. Compound microscopes are often used to visualize stained blood smears, urine sediment, and various microbiology specimens.

Test Your Knowledge 6-1

What is the most common type of microscope used in the clinical laboratory?

a. Dissecting microscope
b. Electron microscope
c. Fluorescent microscope
d. Compound microscope (Outcome 6-2)

Figure 6-1 shows a compound microscope with its component parts identified. It is important that medical assistants working in the laboratory know how to set up a microscope properly, focus a slide for viewing, and maintain the microscope appropriately after the procedure. To perform these procedures correctly, the medical assistant must be able to identify the parts of the microscope. Although at first glance the microscope itself may appear complex, its parts may be broken into categories to help you to remember their names and understand their function.

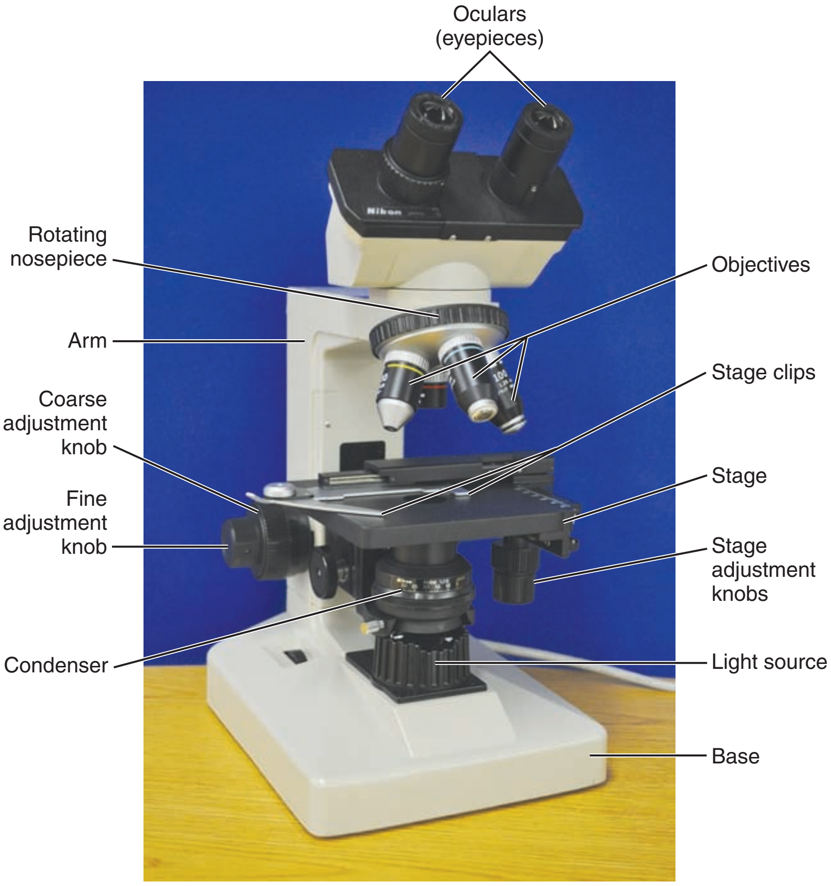

Figure 6-1 Parts of a compound microscope.

Supportive Structures. The more delicate parts of the microscope are supported by a **base** and an **arm.** The base sits on the surface (tabletop or laboratory bench) to keep the microscope secure. The microscope arm is a curved metal support, and is to be gripped when it is being carried. It is important to remember to use two hands whenever moving the unit; one should grip the arm, and one should be placed under the base.

Another supportive structure is the **stage,** which is the surface on which the slide sits for viewing. Most compound microscopes have metal clips that hold the slide securely in place on the stage and allow the slide to be moved around for viewing by turning a mechanical control knob, located under the stage.

Test Your Knowledge 6-2

On a compound microscope, what is the name of the surface where the slide is placed for observation?
(Outcome 6-3)

Structures Used for Specimen Viewing. To view a specimen with the microscope, there are two types of structures used. Some of the structures are designed to provide the necessary light to see the specimen, whereas others are necessary to provide various amounts of magnification. To visualize the specimen as clearly as possible, it is very important that the light and the magnification are both at the correct setting.

Light Sources The **light source** is located in the base of the microscope. The bulb can be turned on and off with a switch or knob, which is usually located on the base toward the back of the microscope. The bulb should not be touched with the bare hands when changing the light source, as the natural oils present on the skin may be transferred to the bulb, and these will overheat when the bulb is turned on, causing the bulb to shatter. The light will come up from the bulb at the base of the microscope, through the **condenser,** then up through the specimen on the stage and finally through the optical magnification components to be viewed.

The condenser is designed to focus the rays of light on the specimen for a clearer image. It may be raised or lowered, depending on the density of the specimen viewed. The knob to adjust the condenser height is on the side under the stage. An **iris diaphragm** is located on the front or just underneath the condenser. The iris diaphragm may be adjusted to increase or decrease the *amount* of light reaching the slide. The amount of light needed to view a specimen depends on its density; generally, the denser the specimen, the more light will be needed. In addition, as the magnification increases, there is more need for light, so the condenser and iris diaphragm need to be adjusted for better viewing.

Test Your Knowledge 6-3

What is the purpose of the condenser on the compound microscope? (Outcome 6-3)

Magnification The compound microscope has two different points of magnification. The first magnification is encountered as the light passes through the specimen from below the stage, and flows into one of the **objective lenses.** These lenses are located on a **revolving nosepiece** and are identified with their respective power of magnification: 4X, 10X, 40X, and 100X. The different objectives also vary by rings of different colors on each, and by the length of each objective tube. When using the microscope, the operator starts with the objective marked with the lowest magnification, and proceeds through the various levels as needed to view the specimen. If the 100X objective is used, a thin layer of oil is placed between the objective and the specimen on the slide to further focus the light and make the image as sharp as possible. This objective is known as the **oil immersion lens.**

The second point of magnification available in a compound microscope is located within the **eyepiece** at the very top of the microscope. An internal mirror actually reflects the image from the slide through this **ocular lens** (eyepiece) where an additional 10X magnification is accomplished. This means that if the 100X oil immersion lens is used to examine a specimen, the object is actually viewed as 1,000 times larger than it would be without magnification. A **binocular** microscope is one that has two ocular lenses (or eyepieces), and a **monocular** microscope is one that has only one ocular lens used to view the specimen.

Focus on the specimen is achieved by using the **coarse adjustment knob** and the **fine adjustment knob**. These knobs project out from the side of the microscope arm. The larger of the two knobs is the coarse adjustment, and is used when the specimen is first being brought into focus at a low magnification. The fine adjustment is the smaller knob, and is used to fine-tune the focus. The fine adjustment knob is used with the higher magnification levels.

Microscope Maintenance Procedures. A microscope must be handled and maintained properly in order to keep it functioning as it should. The following are a few guidelines for handling and maintaining a compound microscope that a medical assistant should adhere to:

1. Keep the microscope covered with a plastic cover when not in use. This protects the delicate microscope and lenses from excess dust and potential damage.
2. Always use two hands (as described earlier) to carry a microscope.
3. Do not touch the light source or the lenses with your fingers. The light source may overheat and burst because of the transferred oils, and fingerprints on the lenses can be very difficult to remove.
4. The optical lenses and eyepieces should be kept clean at all times. These may be cleaned *only* using lens paper, as the glass surfaces are very delicate and will be scratched if Kimwipes or other tissues are used. Xylene may be used to assist with removal of oil or other buildup. A small amount should be placed on the lens paper before gently rubbing the lens.
5. When cleaning the objectives, remember always to clean the 100X optical lens last. This one is used with oil immersion, and if cleaned first, the lens paper used may contaminate the other lenses with oil as they are cleaned. Do not apply pressure to the lenses when cleaning.
6. Keep the stage clean between uses.
7. When moving from one objective to another during use or cleaning procedures, always use the nosepiece to change the powers. Do not push against the tubes directly to move them.
8. Slides should be removed or added to the stage only when the lowest power objective is turned down toward the light source. Do not attempt to remove or add slides when the higher power objectives are in place (pointed down toward the specimen), as the objective lenses may be damaged during the process, and/or the slide may be broken.

Test Your Knowledge 6-4

True or False: The 100X objective (oil immersion objective) should be cleaned first when performing maintenance on the compound microscope. (Outcome 6-4)

Test Your Knowledge 6-5

What should be used to clean the eyepieces on a microscope?

a. Kleenex or a similar facial tissue
b. Laboratory tissues such as Kimwipes
c. Lens paper
d. None of the above (Outcome 6-4)

Procedure 6-1: Microscope Use

TASK

Correctly use a compound microscope to obtain specimen focus with all objectives.

CAAHEP/ABHES STANDARDS

None

CONDITIONS

- Compound microscope
- Lens paper
- Specimen slide
- Immersion oil
- Laboratory tissue such as Kimwipes
- Gloves (if needed)

Procedure	Rationale
1. Wash hands and gather necessary supplies. Apply gloves if the specimen is infectious and unfixed (without preservative).	Gloves are not necessary if the specimen is fixed with chemicals that will kill the microorganisms on the slide. If dealing with unfixed specimens (such as vaginal smears or unstained urine) gloves must be applied.
2. Remove the microscope from storage, if necessary. Be sure to use one hand on the arm and one under the base of the microscope.	The microscope may be kept on the countertop, or it may be stored in a drawer or on a shelf. Always support the microscope with two hands to avoid hitting it against surfaces that could damage the instrument.

Continued

Procedure 6-1: Microscope Use—cont'd

Procedure	Rationale
3. Plug in the microscope to the electrical outlet. Verify that the excess cord does not hang over the edge of the counter.	The cord needs to be behind the microscope to avoid catching the cord and accidentally pulling the microscope from the counter.
4. Use lens paper to clean the eyepieces and the objectives. Start with the eyepieces, after which each objective should be cleaned starting with the lowest power and finishing with the 100X objective. Only the optical surface should be cleaned with the lens paper.	The eyepieces should be cleaned before the objectives to avoid possible transfer of oil to the eyepieces from the objectives. It is also important to remember that the 100X objective should be cleaned last because it is used with oil that could be transferred to the other objectives.
5. Turn on the light source and adjust the light to a low level.	The light should be at a low level as you begin to focus the specimen. More light will be required later as the objectives are changed to higher levels.
6. Adjust the eyepieces to be the same width as your eyes. With some models, it may be possible to adjust the eyepieces individually for a sharper focus.	Both eyes should be used at the same time when looking through a binocular microscope. This will be best accomplished when the width of the eyepieces matches that of your eyes.
7. Rotate the nosepiece to bring the lowest power objective straight down where it is pointing at the slide. There should be an audible click when this objective is in the proper position. Watch the stage and objective closely to avoid touching the stage with the objective. The lowest power objective may have a 4X or a 10X printed on it.	Focus always starts with the lowest power objective. Some microscope models may have a 4X objective, but most will have 10X as the lowest possible magnification.
8. Using the coarse adjustment knob, move the stage until it is at the lowest point possible.	The coarse adjustment knob is the larger knob at the back of the microscope. Moving the stage down will allow space to put the slide in place without potentially touching the objective with the slide.
9. Place the slide on the stage with the specimen side up. If the compound microscope used has stage clips, use these to secure the slide in place. Using the stage adjustment knob, position the slide so that the area on the slide to be viewed is directly over the hole where the light will shine through the specimen.	If the slide is not faceup, it will be very difficult to reach optimum clarity when viewing the specimen. It is especially important to center the slide where the specimen viewing area is focused over the hole where the light comes through when the specimen covers only a small area on the slide.
10. If the microscope has a condenser, adjust the condenser all the way up, using the adjustment knob to the side of the condenser. Using the adjustment lever on the diaphragm, open the iris diaphragm to allow the maximum amount of light to be utilized for viewing the specimen. This will be adjusted to a lower level as you work through the focusing process.	As the specimen focus process begins, it is essential to have as much light as possible available. The condenser may be moved farther from the specimen, and the diaphragm may be adjusted as the specimen is brought into closer focus.

Procedure	Rationale
11. While looking through the eyepieces, slowly turn the coarse adjustment knob until the specimen comes into approximate focus. It may not be possible to bring the specimen into complete focus using only the coarse adjustment knob.	The coarse adjustment knob allows for an approximate focus.
12. Without moving the slide or the objective, use the fine adjustment knob to bring the specimen into complete focus.	The fine adjustment knob is to be used only after approximate focus has been accomplished with the coarse adjustment knob.
13. Adjust the amount of light entering the specimen by moving the condenser down and by adjusting the diaphragm until the specimen appears in crisp, clear focus.	Too much light at a low magnification may cause the specimen to appear washed out, and some of the formed elements may be overlooked. Too little light does not allow for enough contrast to identify some of the structures.
14. Use the stage adjustment knob to move the object viewed directly into the center of the field of vision while looking in the microscope.	If this step is not performed, you may not be able to find the object again after changing to the next level of magnification, especially if the specimen only covers a small portion of the slide.
15. Use the nosepiece to swing the next objective so that it is pointing directly down at the slide. As before, the objective should click into place.	Always watch the objective as you move it into place to avoid touching anything with the objective lens.
16. Use the fine adjustment knob to bring the object into focus. It may be necessary to adjust the amount of light entering the specimen to see the object clearly.	With the objectives above 10X, do not use the coarse adjustment knob for focus.
17. If desired for the type of specimen viewed, continue this process, using the fine focus and adjusting the available light as the objectives are changed to higher levels of magnification.	Some specimens do not require higher levels of magnification if the structures can be identified at the lower levels.
18. If the oil immersion lens (the lens marked with the 100X and a ∞ symbol) is to be used, it will be necessary to place one or two drops of immersion oil on the slide before rotating the objective into place. **a.** Use the nosepiece to move the current objective out of the way, but do not lower the 100X objective until the oil has been added. **b.** Keep the slide as it was with the previous objective and place the drop(s) of oil on the slide right over the light source. **c.** Then, use the nosepiece to lower the 100X lens into place and adjust the focus as needed with the fine adjustment knob.	Resist the urge to make more room for this objective before moving it into place. This objective will be very close to the slide; it will appear that it is going to touch the slide, but if the proceeding steps were performed correctly, it will not touch the slide. The oil will seal the objective with the slide to keep the light focused on the specimen with clarity.

Continued

Procedure 6-1: Microscope Use—cont'd

Procedure	Rationale
19. After the specimen has been analyzed as needed, lower the stage to the lowest possible point, and remove the slide from the stage clips.	The stage should always be at the lowest point during storage to avoid potentially damaging the objectives.
20. Using the nosepiece, adjust the objectives so that the lowest power is pointing down toward the stage.	This allows the focusing to start appropriately when a new slide is put on the microscope, and also helps to avoid potential damage to the objectives.
21. Turn off the light source, and clean the eyepieces and objectives using lens paper. Remember to clean the 100X objective last, if applicable. Xylene can be used on the lens paper if needed for excessive soiling of the lenses.	Only lens paper should be used for this, as any other type of tissue will scratch the sensitive lenses. If the 100X objective is cleaned first, the rest of the lenses may be contaminated with oil. If xylene is used, be sure to follow appropriate protective measures as described by the manufacturer and the workplace.
22. Wipe off the stage and the rest of the microscope using a laboratory tissue such as Kimwipes. Moisten the tissue if necessary to remove dust. The stage should be disinfected periodically by using an alcohol wipe.	The slides, specimens, and oil can easily soil the stage.
23. Unplug and secure the electrical cord, and put the plastic cover back on the microscope. Be certain to move the microscope to the appropriate storage area as directed by office policy.	Be sure to use two hands when transporting the microscope.
24. Discard the slide or place it in the appropriate storage area as dictated by office policy.	Stained specimens are often stored for a period of time for future study. Slides should be treated as a "sharp" and disposed of in a puncture-resistant biohazard container.
25. Remove gloves (if used for the procedure) and wash hands.	Hands should be washed before and after each procedure in the laboratory setting.

Test Your Knowledge 6-6

Which adjustment knob should be used first when the specimen on the microscope is initially brought into focus? (Outcome 6-5)

Test Your Knowledge 6-7

Is the microscope stage moved when objectives are changed for higher magnification of a specimen? (Outcome 6-5)

Other Types of Microscopes

A dissecting microscope may also be utilized in the laboratory to view specimens that need to be seen in their natural state without compression or those that are too thick to be viewed effectively with the compound microscope. This type of microscope may also be known as a **stereo microscope** because the microscope is designed to allow a slightly different view from each of the eyepieces, providing an image with depth and dimension. The light source for this type of microscope is not focused in the same way that the compound microscope is focused; it is diffused or reflected onto the image rather than shone directly through it. The magnification ability of the dissecting microscope is not as extensive as that of the compound microscope, as it is usually only capable of enlarging the specimen by 40 or 100 times its normal size. Dissecting microscopes are often used for blood bank procedures when the technologist is examining different serum and cell combinations to look for **agglutination** of the specimen. Dissecting microscopes may be used for a variety of other purposes in the clinical laboratory when fine, detailed work needs to be performed with some magnification.

An **electron microscope** allows for much greater magnification of specimens, and also provides a three dimensional image. Electrons are utilized to "excite" the specimen, and the resulting interaction with the atoms present in the specimen will produce an image to be recorded and studied. This image is more of a picture that is defined by the properties of the specimen, rather than a true image seen by the eyes while viewing the specimen. The image created by the excitation of the sample is stored and viewed on a computer. Electron microscopes may be used to visualize specimens that are far too small to be seen with a compound microscope, such as viruses and the interior structures of cells. This type of microscope is not routinely found in clinical laboratories, as the user requires specialized training to operate it and interpret the images created with examination.

Test Your Knowledge 6-8

What type of microscope may be used to see viruses?

a. Compound microscope
b. Electron microscope
c. Dissecting microscope
d. Fluorescent microscope (Outcome 6-2)

How Do Medical Assistants Use Microscopes?

The majority of the time, a medical assistant will not perform microscopic examinations for diagnostic purposes. There are a few microscopic procedures that have been classified as CLIA moderately complex procedures, and these can be performed by medical assistants with appropriate documented training. These include urine microscopic analysis (see Chapter 22) and normal blood cell differential counts (see Chapter 12). These are the exceptional cases in specialized offices such as urology offices or physician office laboratories with minimal staffing. Most commonly, a medical assistant's use of a microscope is limited to focusing the specimen for the health-care provider, as well as performing maintenance procedures. It is of great benefit for a medical assistant to be capable of focusing the microscope, as this allows the health-care provider's valuable time to be spent examining the specimen rather than performing the original setup process. Medical assistants may be focusing stained blood smears, urinalysis sediment specimens, vaginal smears, nasal smears, Gram-stained microbiological specimens, and microscopic examinations for the presence of fungal elements on the microscope for examination. Remember, if the health-care provider does perform microscopy examinations, the laboratory must register as a CLIA site that provides this type of service.

Test Your Knowledge 6-9

Is a medical assistant properly trained to perform microscopic identification procedures? (Outcome 6-6)

Centrifuges

A **centrifuge** is an instrument used to prepare blood (or other liquid specimens) for transport or testing. Blood is made up of cells that are suspended in a liquid substance called **plasma.** Chemical analysis is often performed on the plasma, and in order for the results to be accurate, the cells must be removed from the liquid portion of the blood as soon as possible after the specimen is obtained. A centrifuge is used to force the separation of the cells from the fluid portion of the blood (plasma).

If a blood specimen is allowed to sit for an extended period of time in a tube, gravity causes the cells to settle to the bottom of the tube, as they are denser (heavier) than the surrounding fluid. A centrifuge creates a powerful "artificial" gravity source, called **centrifugal force.** The force is formed as the specimen spins around an axis hundreds to thousands of times per minute. When a specimen is processed by a centrifuge, the separation of the different components of the blood is accelerated; the cells are pushed to the bottom of the tube within minutes. This allows the plasma to be removed from the cells in a timely manner so that chemical analysis can be performed. Urine specimens may also be centrifuged to bring the suspended elements (such as blood cells, bacteria, etc.) to the bottom of the tube so that they can be examined under a microscope.

There are various types of centrifuges available for use in a laboratory. Some of the benchtop models are designed to spin 8 to 10 samples at a time for a maximum of 10 minutes each cycle. Others are units the size of a dishwasher, capable of processing hundreds of tubes in each cycle for extended periods of time. The decision about which type of centrifuge to use is based on several factors, including the volume of samples to be processed, the requirements for the liquid portion of the blood after separation, and the amount of separation required. The decision may also be based on the recommended maintenance procedures for the centrifuge model of choice, as well as cost. Some centrifuges generate a lot of excess heat as they operate, due to the rapid turning of the unit around the axis. These models require built-in refrigeration to decrease the temperature to keep the specimens from becoming overheated.

Centrifuge Maintenance

Maintenance of all centrifuge models is similar. It is important that the interior and exterior of the machine stay clean and that there is nothing impeding the rotation of the unit around the axis. The **rotations per minute (rpms)** must also be checked at least quarterly using a **tachometer.** If the rotations have changed significantly since the last check, the instrument should be serviced, as there may be a problem that needs to be addressed. Many centrifuge models have a brushed drive motor, and these brushes must be changed at intervals recommended by the manufacturer. All metal components of the centrifuge should be checked for cracks at least once a month. As with all electrical equipment, the electrical cord should also be checked for wear periodically.

Test Your Knowledge 6-10

List two maintenance procedures to be performed on a centrifuge. (Outcome 6-7)

Centrifuge Safety

- All medical centrifuges should have a cover that must be locked in place when in operation. Figure 6-2 shows an example of a centrifuge. The cover keeps the operator safe from potential aerosol formation or splashing that may occur if a specimen breaks or if a tube becomes uncovered during the cycle.
- It is also important that the centrifuge is **balanced** before starting. This means that for every tube inserted into the centrifuge, a tube with the same weight of fluid must be placed directly across from it in the unit. This additional tube may be filled with the same fluid as the specimen (another blood tube filled to the same level, for example), or it may be a balance tube that is filled with the same amount of water. This is a critical step, no matter what type of specimen is to be centrifuged. If the unit is not balanced, it is possible that the centrifuge will rock while operating (like an off-balanced washing machine), fall off the countertop, and break the specimens as they are spinning.
- Be certain that all specimens are capped securely before starting the centrifuge.
- It is very important to follow the manufacturer's recommendations concerning length of cycles and specific maintenance.
- Remember never to open the centrifuge before it stops spinning completely. Never use your hand to try and slow down the final spin; this can be very dangerous.

Test Your Knowledge 6-11

What does it mean to balance a centrifuge? (Outcome 6-7)

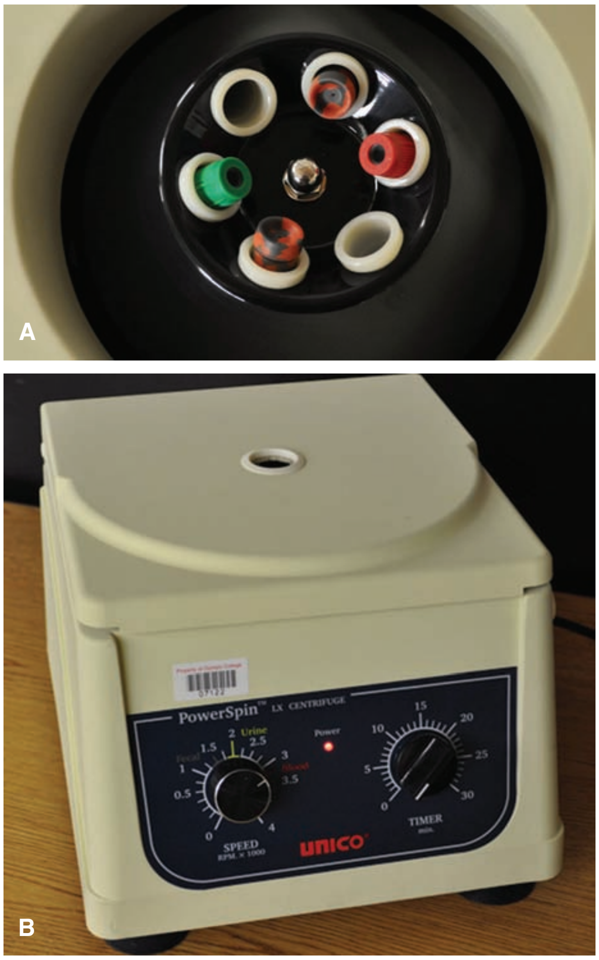

Figure 6-2 The inside (A) and outside (B) of a typical benchtop centrifuge.

Procedure 6-2: Operating the Centrifuge

TASK

Use a benchtop centrifuge to successfully separate plasma from cells in a blood specimen.

CAAHEP/ABHES STANDARDS

None

CONDITIONS

- Gloves
- Benchtop centrifuge
- Blood specimen in tube
- Balance tube
- Spacers to go inside slots in centrifuge if necessary
- Water and transfer pipette
- Test tube rack
- Disinfectant wipe

Procedure	Rationale
1. Wash hands and gather necessary supplies. Apply gloves.	Hands should be washed before and after performing any procedures in the laboratory, and gloves must be worn whenever handling blood specimens.
2. Verify that the centrifuge is plugged into an electrical outlet. Open the lid and verify that all the slots in the centrifuge are empty.	Occasionally the previous user will have left a tube or a spacer in one of the slots within the centrifuge, which can cause it to become unbalanced while processing the specimen. Care should also be taken to verify that there is no liquid in the bottom of the receptacles in the centrifuge, as this could also cause the centrifuge to become unbalanced.
3. Pick up the blood specimen and hold the tube at eye level. Fill the balance tube with water until the two tubes have an equal volume of fluid.	The tubes must have the same fluid volume for the centrifuge to be balanced. This can also be accomplished with two tubes of blood that are the same and are filled to the same level.
4. Verify that the original rubber cap is on the blood tube securely. The balance tube should also be capped with the original rubber cap.	To avoid aerosol formation, all specimens processed in the centrifuge should be securely capped.
5. For some centrifuge models, it may be necessary to place a rubber spacer or adapter in the centrifuge slot to accommodate the size of tube spun. Consult the manufacturer's recommendations.	The spacer or adapter will allow the blood tube to remain high enough in the centrifuge to be removed after spinning, and it may also allow for a tighter fit within the slot to avoid the breakage from vibration that is possible when the tube does not fit securely.
6. Place the two tubes in slots within the centrifuge that are across from each other. The centrifuge can be operated if the other slots are empty, as long as the two tubes are across from each other.	Balance is essential to keep the centrifuge from vibrating excessively while operating. An unbalanced centrifuge can cause the tubes to break, and it also can cause the centrifuge to fall off the countertop due to the excessive motion.
7. Close the lid securely. Turn on the timer on the centrifuge to the appropriate time for spinning a blood specimen. This will be indicated on the front of the centrifuge or in the manufacturer's insert.	Larger tubes may require a higher rate of centrifugation or a longer time in the centrifuge to achieve the desired separation of cells and plasma.

Continued

Procedure 6-2: Operating the Centrifuge—cont'd

Procedure	Rationale
8. Monitor the centrifuge for at least 20 to 30 seconds after it has been turned on to see if the instrument is vibrating excessively. If so, turn off or unplug the centrifuge immediately.	An unbalanced centrifuge can be dangerous if it vibrates on the countertop and moves from its original position.
9. Do not attempt to open the centrifuge until it has come to a full stop.	Opening the centrifuge prematurely or placing a hand into the unit while it is still spinning can be very dangerous.
10. When the centrifuge has come to a complete stop, open the lid and carefully remove the tube of blood. Place the blood in a rack for further processing.	Care needs to be taken when removing the tube of blood so that the plasma and cells do not become mixed.
11. Remove the balance tube and the spacers and adapters if used. The balance tube may be discarded or reused for another specimen. The spacers and adapters are to be stored for use with the next centrifuge load.	Some laboratories keep a set of balance tubes close at hand to be used for additional centrifugation. Be sure that the tube cap is securely fastened if using the balance tube again.
12. Use a disinfectant wipe to wipe down the exterior and interior of the unit.	If excessive contamination is evident, it may be necessary to clean the unit to a greater extent. In many units, the slots in which the tubes are placed are removable so that they can be washed if necessary.
13. Remove gloves and wash hands.	Hands should be washed before and after performing any procedures in the laboratory.

POINT OF INTEREST 6-1
"Sally Centrifuge"

In many areas of the world health-care providers do not have access to electricity or funds for purchasing traditional instruments used for diagnosis. The presence of anemia contributes significantly to the diagnosis and subsequent treatment of disorders such as malnutrition, HIV/AIDS, tuberculosis, and malaria. Blood specimens collected from the patients in these areas must be shipped to larger health-care facilities for testing, which is expensive and time consuming.

In 2010, two Rice University college students created an inexpensive centrifuge that allows the cells to be separated from the plasma in microhematocrit tubes in 10 to 20 minutes without the use of electricity. Lauren Theis and Lila Kerr created the centrifuge as a class project, using materials that included combs, yogurt containers, and a salad spinner. The centrifuge cost approximately $30 to build, and it can process 30 samples of 15 microliters each at one time. The centrifuge must be pumped by hand for 10 to 20 minutes, at which point it reaches speeds of approximately 950 rotations per minute and successfully separates the blood cells from the plasma in the tubes. The tubes then can be compared to a reference chart to obtain the hematocrit result, used to diagnosis the presence of anemia.

These students are part of a Rice University program, Beyond Traditional Borders. As part of this program, the students used the centrifuges in several remote locations during the summer of 2010 to test their ability to withstand traditional use in a health-care setting. If successful, "Sally Centrifuge" could have a profound impact on the ability to deliver affordable, timely health care in many locations across the globe.

Laboratory Refrigeration

Many specimens must be maintained at reduced temperatures to provide reliable test results when analyzed. The instructions for specimen collection and processing for a particular test may include refrigeration of the specimen, or possibly freezing of the plasma or **serum.** To ensure that the specimens are at the correct temperature, the refrigerators or freezers in the clinical laboratory should be monitored each day; in large laboratories with multiple shifts, they may be monitored more than once a day.

A **laboratory thermometer** is used to obtain accurate, reliable temperature readings. Laboratory thermometers are usually made of a measurement device that has the end designed to register the temperature while submersed in a container filled with liquid. The employee who is monitoring the temperature will open the door of the refrigerator or freezer, read the temperature on the thermometer, and record the result on a log sheet nearby. To keep the temperature as stable as possible, the door for the refrigerator or freezer should be opened only long enough to read the thermometer. Some thermometers are designed with systems that allow them to be monitored by a device that is outside of the refrigerator. This is advantageous because it is not necessary to open the door of the unit to take the temperature. In either case, a range of acceptable readings for each refrigerator or freezer must be established, and if the results fall outside this range, corrective action must be taken immediately. The specimens in the unit must be transferred elsewhere until the temperature reading is within the acceptable range again.

In laboratory situations in which refrigeration temperature is especially critical, an alarm system may be installed that sounds whenever there is a fluctuation in the temperature readings. Blood bank facilities have a very small range of acceptable storage temperatures for units of blood, and must have processes in place for critical monitoring of all their refrigeration units. These special refrigerators often have battery backups that activate a cooling system in case of power failure, as well as remote monitoring systems and audible alarms. The temperature for these units must be monitored more frequently than other refrigerators; often there is a continuous recording of temperatures.

Some refrigerators and freezers in physician office laboratories may be used to store specimens as well as vaccines or other medications. These products will have very specific storage requirements, so monitoring of the temperature used for storage will be especially critical. If the unit falls outside the acceptable storage ranges, the medication and vaccines may need to be discarded. Quality control materials may also be stored in refrigerators or freezers with specimens, and these materials will also have specific storage requirements that must be followed.

It is imperative that all refrigeration or freezer units in a laboratory setting are properly maintained to avoid excessive moisture or ice buildup. The refrigeration coils at the back of the unit should be kept free of dust buildup, and the interior and exterior of the unit should be kept as clean as possible. *No food or drink is to be stored in laboratory refrigerators where specimens are kept.* This is a safety precaution designed to protect employees from potential contamination of their food, as well as an Occupational Safety and Health Administration (OSHA) regulation that must be followed.

Test Your Knowledge 6-12

If specimens are to be kept at a certain reduced temperature, how often should the refrigerator temperature be monitored? What other precautions should be in place? (Outcome 6-8)

Incubators

When microbiology specimens are collected from a human for analysis, the pathogens that may be present are reproducing at body temperature. This temperature is approximately 25° to 27° Celsius or 95° to 99° Fahrenheit. To keep these pathogens alive until they can be identified, it is necessary to keep them at this temperature for a few days. A laboratory **incubator** is used for this purpose. The temperature for an incubator must be monitored at least daily, and the results must be documented on a log sheet. Much like a laboratory refrigeration unit, there is a limited acceptable range for these temperature readings, and if the incubator falls outside of that range, the unit must be serviced.

In addition, **water baths** or other types of heating units may be used in a laboratory. Water baths may be used to heat specimens as part of specific testing procedures. Other heating units may be used to “fix” a specimen to a slide, or to sterilize inoculation devices in the microbiology laboratory.

EQUIPMENT USED FOR AUTOMATED CLIA-WAIVED LABORATORY TESTING

Many medical offices choose to perform automated CLIA-waived tests in their own laboratory on site. Performance of these tests in the office environment rather than in a reference laboratory may benefit the patient in various ways:

- On-site testing may assist the health-care provider to assign a definitive diagnosis quickly. This allows a plan of treatment to be established before the patient leaves the office. The patient and health-care provider can discuss the plan, and an opportunity is provided for face-to-face communication. This can be especially critical for patients who are acutely ill and those with chronic health conditions requiring frequent laboratory testing to monitor the progress of their treatment.
- CLIA-waived tests performed in the physician's office may be less expensive for the patient. Many times insurance coverage is limited for laboratory testing, so a reduced charge is definitely a benefit.
- Performance of testing on site may allow patients to minimize visits; they can have their specimen collected and tested at the same location that they see their health-care provider, without going to a separate destination for collection and/or testing. This helps improve patient compliance rates.

CLIA-waived automated testing may also be performed in larger laboratories for tests that do not require more advanced methods of analysis. These laboratories may employ medical assistants or phlebotomists to perform these testing procedures with appropriate oversight by other laboratory professionals. Automated methods for CLIA-waived tests include chemical urinalysis, chemistry testing, and hematology testing.

Test Your Knowledge 6-13

List one advantage of performing automated CLIA-waived tests in a physician office laboratory. (Outcome 6-10)

Testing Methodology

Chemistry analyzers may test the liquid portion of the blood (plasma) or the whole blood specimen. Hematology instruments are designed to perform various measurements on the cells present in the specimen. Automated analyzers used for CLIA-waived hematology, coagulation, urinalysis, and chemistry testing (as well as advanced instruments used in more complex testing procedures) are often used to provide quantitative results of substances or cells present in the specimens. Other testing instruments may provide qualitative results, such as those used for **fecal occult blood** testing or urine drug screening. In this case, the presence or absence of a specific analyte provides the necessary information for the health-care provider to develop a plan of action.

Most of the CLIA-waived chemistry testing methods explained in this textbook use reagents that change color when they are exposed to the chemicals present in the specimen. This color change is measured by a **spectrophotometer,** an instrument that measures light intensity. A specific wavelength of light enters a sample, and depending on the amount of color change, a certain amount will continue through the sample to be measured on the other side. The measurement of the intensity of the light at the end of the process is directly related to the concentration of the chemical substances present in the solution. The light intensity measurement may be changed by absorption of the light by the specimen, or by reflecting or scattering the light so that it is not measured directly at the end of the reaction.

Hematology testing often uses **electrical impedance,** a process for counting cells in the whole blood specimen and differentiating them by size. For this type of test, whole blood specimens are added to a **diluent** (a liquid used for dilution of a specimen) that is capable of conducting electricity. An electrical current is applied to the mixture of specimen and diluent as it passes through a small opening, called an **aperture.** Because the blood cells do not conduct electricity, they break the current between the electrodes on either side of the aperture. The amount of impedance (interference of the electrical signal) caused by a certain cell will allow the instrument to count the cell and approximate the size and other physical properties. A similar type of measurement uses the amount of light scattered by a specimen to measure the cell numbers and cell sizes in a hematology specimen.

Hematology instruments (such as those that test only **hemoglobin)** may also use spectrophotometers to measure a specific substance. In this case, the cells must be broken, or **lysed** prior to the testing procedure, so that the hemoglobin present inside the cells may be measured.

When using an instrument to perform any of these tests, it is imperative that the manufacturer's instructions are followed concerning frequency and extent of quality control testing and calibration of the instrument. These procedures must be performed as directed to verify that the instrument is operating as it should before patient samples are analyzed. The reagents used with the analyzers often have storage requirements and expiration dates that must be monitored as well, and quality control samples must be prepared and processed as directed to produce meaningful results.

Test Your Knowledge 6-14

Electrical impedance is a common term used to describe how the number of cells in a specimen are measured. What process occurs to allow the cells to be counted by the instrument? (Outcome 6-11)

Instruments Used for Chemical Testing of Urine Specimens

Urine specimens are analyzed in various ways. One of the methods used to evaluate the specimen is a chemical analysis to detect and/or quantify the presence of substances in the urine specimen that may indicate disease. Quite often, this analysis will be performed manually in a physician office laboratory, but this analysis may also be performed using an instrument and **reagent strips** that are imbedded with small squares designed to change color when exposed to specific chemicals present in the urine specimen. Although the exact number and types of chemicals analyzed in the specimen may vary according to the manufacturer of the unit, common chemical substances detected include the following:

- pH levels
- protein
- glucose
- blood
- leukocytes
- specific gravity
- bilirubin
- glucose
- ketone
- urobilinogen
- nitrite

The process of reading the reagent strips is time and color sensitive. The chemical measurements are based on changes in color that develop in response to the presence of certain chemicals. These changes do not all occur at the same rate, which means that the reagent pads on the strips must be read at specific time intervals after the strip has been exposed to the urine specimen. Automated urine analyzers are designed to move the strip through the measuring device with the appropriate speed to read these reagent pads at the correct time.

A similar procedure is used for testing urine using any of these machines. Essentially, the reagent strip is immersed in the urine specimen, blotted to remove excess specimen, then applied to a tray that feeds the reagent strip into the instrument for analysis. The instrument times the advance of the strip appropriately for the different reagent pads to be analyzed for color changes. The amount of color change is directly related to the concentration of the chemical substances in the urine, and once analyzed, the results are printed on a strip that may be kept as a permanent record. Some of the instruments are also capable of transmitting the results directly to a computer so that they may be stored electronically. Urine analyzers are quick and easy to use, and eliminate the need for the employee to monitor the reaction of the different areas of the reagent strip for the full time needed for color development.

Common urine analyzers include the Clinitek Urine Analyzer, manufactured by the Bayer Corporation, and the Urisys 1100 Analyzer, manufactured by Roche. Henry Schein also manufactures the One Step Plus Analyzer (Fig. 6-3). The automated chemical analysis testing that is performed with these instruments is CLIA-waived, as long as the manufacturer's directions are followed exactly as printed. Quality control and maintenance for this type of equipment may include the use of commercial quality control specimens of different levels, calibration of the instrument, and cleaning of the instrument at regular intervals. Some general aspects to keep in mind

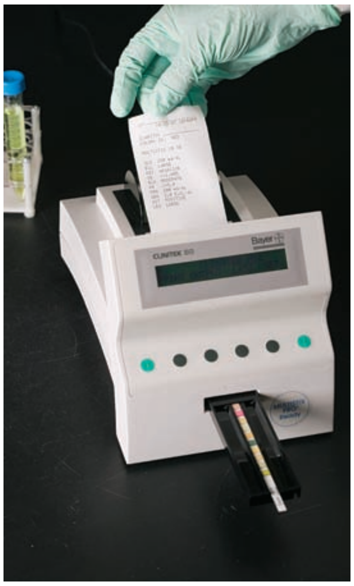

Figure 6-3 One Step Plus Urine Analyzer. *Courtesy of Henry Schein, Inc.*

when performing urine testing with an automated instrument include the following:

- Reagent strips do outdate. *Always* check the expiration date before using the strips.
- Reagent strips must be protected from moisture, so the bottle must be closed immediately after removing the necessary strips for testing.
- Automated urinalysis instruments may produce erroneous results in situations in which the urine is discolored by medications or dyes used for diagnostic purposes. The employee performing these tests must keep this in mind and follow the manufacturer's recommendations and the office policy for confirmatory testing before reporting results on discolored specimens.
- As with all testing procedures, documentation of the quality control and maintenance procedures is critical, as well as the appropriate documentation of all patient results.

Instruments Used for Coagulation Testing

Coagulation testing is used to screen patients for blood clotting issues. It may also be used to monitor patients who have been placed on anticoagulant therapy, such as warfarin (Coumadin). As presented in more detail in Chapter 15, coagulation testing is commonly performed as a CLIA-waived test in the physician office laboratory environment. The most common test used for screening or monitoring is the **prothrombin time test.** This test is more commonly known as a **protime,** and it measures the length of time necessary for a blood specimen to form a clot when reagents are added. The result is reported in seconds. Protimes may also be performed in larger laboratories using testing methods that are not CLIA-waived. When the protime test is performed, an **international normalized ratio (INR)** is usually reported in addition to the test result. The INR is a calculation provided by dividing the protime result for the patient by the normal control value for the lot of reagents currently in use for that system.

Protime testing on the CLIA-waived instruments is accomplished by inserting a reagent strip or cartridge into the instrument, then adding one or more drops of blood to the designated area of the strip. The blood specimen is usually obtained from a capillary puncture, and a drop is placed directly onto the testing device or transferred from the finger using a capillary transfer device. There are specific timing and quantity requirements, and if the operator does not follow these guidelines, an error code will result and the test must be repeated with a fresh specimen. Some of these coagulation analyzers provide a printout, others can be connected to an external printer for result documentation, and still other instruments may be interfaced to the computer for the results to be stored electronically. The units do have a display screen where the operator can read the result.

Various CLIA-waived methods and instruments may be used to perform the protime test. These include the ITC ProTime-3 as well as Roche Diagnostic's CoaguChek S, XS, and XS Plus. Other CLIA-waived protime systems include the Hemosense Inratio system, and several products produced by the Lifescan Corporation. Figure 6-4 shows a Roche CoaguChek S with reagents and necessary to perform a protime test. Each type of instrument has instructions that are a little bit different, but all the testing systems share some of the following characteristics:

- All CLIA-waived or home use systems have test strips that are packaged individually so that one test is performed at a time.
- Each lot of test strips has a unique product code that must be entered into the machine to perform the test. Some products have a computer chip with the product code, and this must be inserted into the analyzer prior to the sample testing process.
- All instruments have quality control procedures; some are built-in internal controls that run automatically during the testing process, while other instruments use liquid commercial control materials that are run as

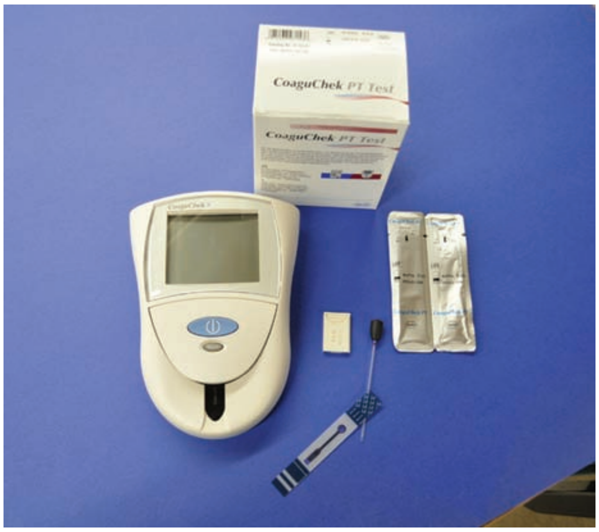

Figure 6-4 Roche CoaguChek S Plus with reagents and quality control materials.

patient samples. As always, the quality control (QC) values must be documented and interpreted as recommended by the manufacturer.
- The reagents all have specific storage requirements, although these vary from brand to brand.
- All CLIA-waived methods require whole blood samples for testing, and the test is designed to be run on capillary blood. A lancet is used to pierce the skin for the sample.
- Many of the testing methods recommend use of a microcapillary tube with a bulb or a collection cup to assist with obtaining adequate sample volume.
- The blood sample is applied directly to the reagent strip for all methods; generally there is a window or circle that must be covered completely with blood.
- The instruments require very little maintenance. Most have batteries and some also have A/C adapters. The batteries must be changed periodically, and should be removed from the device if it will not be used for an extended period of time. The analyzers should also be kept clean. The instrument (especially the sample application area) may be cleaned with a cotton-tipped applicator moistened with isopropyl alcohol or a 5% bleach solution.
- None of the instruments should be immersed in water or any other liquid.
- All manufacturer's instructions for the testing process must be followed exactly as written.

Test Your Knowledge 6-15

What type of specimen is needed for handheld coagulation instruments?

a. Plasma
b. Serum
c. Urine
d. Whole blood (Outcome 6-13)

Test Your Knowledge 6-16

Are results for CLIA-waived coagulation testing available at the completion of the test? (Outcome 6-13)

Instruments Used for Chemistry Testing

CLIA-waived automated chemistry testing procedures are now available for a variety of analytes. Some of the most common tests performed in the physician office laboratory are **glucose** measurements and **glycosylated hemoglobin monitoring** for diabetes screening and treatment. **Cholesterol** studies and **electrolyte** measurements are also quite common. Many instruments now have the capability of performing chemistry panels that include three or more tests, providing the health-care provider a more comprehensive set of information to work with as they develop a treatment plan. The test results are usually available within an hour of the test onset, and many of the instruments are capable of transmitting these results directly to a computer for documentation. Calibration of these analyzers may be performed internally or with an external cartridge, and many methods include an internal method of quality control that eliminates the need to purchase separate liquid controls. Also, because the CLIA-waived status for these instruments applies to whole blood specimen testing, most of the instruments require only a few drops of blood, which can be accomplished with a capillary blood draw.

There are more CLIA-waived glucose testing methods than any other type of chemistry analyzers. Many of the glucose instruments used in the physicians' office laboratory are also in use as home testing devices. These instruments have become more technologically advanced, and use even less blood for analysis than in the past. It is no longer required to use fingertips for all samples; innovative lancets and reduced specimen volume requirements have allowed the sites for specimen collection to vary. Devices used at home and in the office allow for test results to be stored, and some also may interface directly with a computer system to allow for closer monitoring and better communication with the health-care provider.

Glucose analyzers utilize reagent strips or cassettes that are inserted into the instrument. A capillary blood specimen is obtained and a few drops are placed directly onto the appropriate area of the reagent strip. The strip may be advanced into the instrument, or the specimen may be analyzed through the application area at the front of the instrument. The analyzer includes a display screen on which the operator may view the results, and (as is the case for the other instruments already covered in this chapter) the analyzer may be capable of printing results or transmitting them directly to a computer for electronic storage. Some instruments are even capable of charting or graphing patient data to provide a historical overview of results over a period of time.

Glucose analyzers found in the physician office laboratory may include the Roche Diagnostics Accu-Chek, the Abbott iStat, and the Hemocue Glucose 201 Microcuvette. There are many CLIA-waived glucose analyzer methods, and more information may be found for any specific type by visiting the website for that manufacturer. It is important to remember that each brand of analyzer has unique reagent strips and quality control materials; these are not interchangeable.

Other chemistry tests may be performed using CLIA-waived automated testing procedures in a physician office setting or other small laboratory. These include measurements for electrolytes, blood urea nitrogen (BUN), triglycerides, cholesterol, and calcium. Instruments used for testing these analytes use individual testing cartridges, which are self-contained. A capillary blood sample may be obtained in some situations, or in others the entire tube of blood collected by venipuncture may be placed in the instrument to be sampled. Because each testing cartridge is self-contained, the blood sample is added to every cartridge individually. The addition of the sample may be performed within the instrument using an automated method, or the person performing the test may add the sample to each cartridge by hand. Results are usually available within 30 minutes, even if the patient has several tests performed at once. The results are printed out on a slip that may be added to the patient record, and many models are also capable of transmitting the information electronically to a computer system.

The CLIA-waived automated chemistry instruments that use individual cartridges are more expensive to operate per test than the large chemistry analyzers found in reference and hospital laboratories, because of the high cost of the individual cartridges. Larger instruments have reagents that cost less per assay, but they are not CLIA waived for operation. The low sample volume and limited variety of testing procedures ordered in a physician office laboratory are served well by the instruments that use the self-contained individual cartridges.

The Abbott iStat Chem 8+ Cartridge analyzer is a common CLIA-waived chemistry instrument that may be used for point-of-care testing in an emergency room, at the bedside of a patient in the hospital, or in a physician office laboratory. It uses whole blood, and can be used to produce results for a variety of tests. Cholesterol testing may be performed using the Cholestech LDX, but this instrument may not be used for very many other types of tests. The Abixis Piccolo Xpress blood chemistry analyzer has a good deal of flexibility, uses tubes of whole blood, and doesn't take up very much space on a countertop. A comprehensive list of the CLIA-waived automated chemistry testing instruments may be found on the U.S. Food and Drug Administration (FDA) website listed in the Resources and Suggested Readings section at the end of this chapter.

Test Your Knowledge 6-17

What is a common chemistry analyzer used in small laboratories to test cholesterol levels? (Outcome 6-14)

Instruments Used for Hemoglobin Measurements

Hemoglobin is the molecule within the red blood cells that carries oxygen to the tissues of the body. Iron is necessary to build this molecule, so the hemoglobin level of the blood is directly related to the oxygen-carrying capacity of the cells, as well as the iron levels of the individual. Hemoglobin (Hgb) is commonly performed as a screening test for **anemia,** and it may also be ordered as a test to monitor progress when a patient is being treated for anemia.

Because hemoglobin is part of the red blood cell structure, the cells must be broken or lysed before the hemoglobin level can be measured. In CLIA-waived automated hemoglobin testing instruments, the cells are lysed within the individual cuvette when the blood is added. For instance, in the HemoCue hemoglobin testing method, the inside of the individual cuvette used for the test is coated with a chemical (sodium deoxycholate) that destroys the red blood cell membranes, allowing the hemoglobin to be measured.

Test Your Knowledge 6-18

Is hemoglobin measured from intact red blood cells? (Outcome 6-15)

Regardless of the type of instrument used, hemoglobin testing is always performed using whole blood. CLIA-waived systems use capillary samples, and the blood is usually applied directly to the cuvette (or testing device such as a strip) from the capillary puncture site. Proper capillary technique must be used to provide an appropriate specimen. One test is performed at a time, and the individual testing device may be placed in the instrument after the blood has been added, or it may be necessary to place it in the machine prior to the addition of the sample. (Check the manufacturer's recommendations for details.) The instruments will have a display screen on which the progress of the test may be monitored once the blood has been added; generally the results are available within one minute. Some of the instruments may be capable of printing the results, whereas others may also be able to send the results electronically to a computer system for documentation.

The CLIA-waived hemoglobin systems are equipped with electronic calibration methods. The HemoCue system has a standardized cuvette with known values that should be checked at regular intervals to be certain that the instrument is operating correctly. The ITC Hgb Pro has an internal calibration that is performed every time a test is performed. If the internal calibration process does not show that the instrument is working correctly, an error will result and the instrument cannot be used for patient testing until the calibration is successful. Liquid commercial quality control specimens may also be utilized to verify that any hemoglobin instrument is operating appropriately.

Other Hematology Instruments

Many physician office laboratories use hematology instruments that have been classified as moderately complex according to CLIA regulations. Common manufacturers are Beckman Coulter and Abbott. Remember, when performing tests of moderate complexity, it is necessary to perform quality control and calibration procedures more frequently, and there are more regulations for personnel training.

Test Your Knowledge 6-19

Is a glucometer an example of an instrument used for hematology testing? (Outcome 6-14)

Glassware and Other Miscellaneous Laboratory Equipment

Physician office laboratories don't have as much glassware as larger laboratories, and in many cases, glassware is now replaced by plastic disposable containers. However, there are a few standard items that may be found in the laboratory setting, regardless of the size of the facility:

- **Pipettes:** Pipettes are used to move liquid from one place to another. It is similar to the turkey baster that you may find in your kitchen. Some pipettes are designed to measure small amounts of liquid very accurately. Others are used just to transfer liquids from one place to another, without measurement. Pipettes come in varied sizes, and are often plastic and disposable.
- **Beakers** and **flasks:** Beakers and flasks are both containers that may be used to store, transfer, or heat liquids. Beakers generally are wide at the top and have a flat base, whereas flasks have a narrow opening at the top and a rounded bottom that is much larger than the top.
- **Glass slides** and **cover slips:** Glass slides are still used in the laboratory, even though many items are now made of plastic. Slides made of glass are still the best item to use when viewing items under the microscope, as they do not distort the view in the manner that some plastic slides may. Slides may be completely clear, or they may be frosted partially or completely for specialized uses. Some slides may have a depression in the center for certain types of specimens. Cover slips are small panes of glass or plastic placed on top of liquid specimens on the glass slides before viewing. Cover slips are often used when the specimen is viewed in a natural state without any preservative or staining.
- **Cylinder:** A cylinder is a slim, round container that is used to measure liquids. A **graduated cylinder** is marked with specific units to allow for accurate measurement.

Because glassware is often used to accurately measure liquid in the laboratory setting, it is important to know the correct way to read the amount of liquid present. Liquid in a glass container is not completely flat at the top, because liquid is “attracted” or pulled up or down the sides of the container to form a curve. This curved surface of the liquid is known as the **meniscus.** In a narrow container, the meniscus will be more curved than it will be in a wide container. When measuring liquid in glassware, always perform the measurement at the lowest or highest point of the meniscus; this principle applies whether removing liquid to obtain a certain volume, or adding liquid to a container. Figure 6-5 shows how to measure liquid correctly in glassware using the meniscus.

Test Your Knowledge 6-20

The curved area at the top of a column of liquid in a glass cylinder is called the:

a. Photometer
b. Diaphragm
c. Flask
d. Meniscus (Outcome 6-17)

SUMMARY

Various types of automated instruments may be found in the clinical laboratory. These instruments may be used to process samples for analysis, examine specimens under the microscope, or analyze body fluids utilizing CLIA-waived automated testing techniques. Medical assistants need to know the correct way to use and maintain equipment for specimen processing, focus microscopes, and perform automated testing appropriate to their level of training. It is also important that medical assistants familiarize themselves with the other equipment used in the laboratory setting, such as refrigerators, incubators, and various types of glassware. Care should be taken to maintain all equipment appropriately, and follow the manufacturer's directions to perform testing procedures exactly as directed.

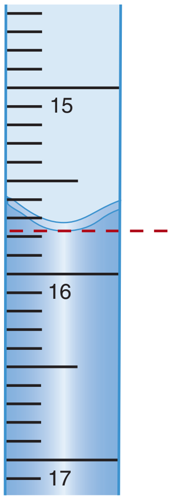

Figure 6-5 Reading the liquid volume using the meniscus. Many liquids curve at the edges of glass containers because of the attraction of the molecules to the glass. This curvature at the top of the liquid is called the *meniscus*. The level is to be measured at the horizontal center of the curvature. Some liquids curve upward instead of downward from the edges of the container, in which case the liquid level is to be read at the top of the curvature. Always hold the container at eye level when measuring.

TIME TO REVIEW

1. What disorder might be monitored by performing periodic glycosylated hemoglobin tests? Outcome 6-1
- **a.** Diabetes
- **b.** Hypertension
- **c.** Hyperlipidemia
- **d.** Hypokalemia

2. A device used to transfer small quantities of liquids is a: Outcome 6-1
- **a.** Pipette
- **b.** Cuvette
- **c.** Cylinder
- **d.** Beaker

3. True or False: A tachometer is a type of thermometer used to measure temperatures in the laboratory environment. Outcome 6-1

4. True or False: The objective is a device that can be adjusted to control the amount of light that enters a specimen on a microscope. Outcome 6-1

5. How many eyepieces does a binocular microscope have? Outcome 6-1
- **a.** One
- **b.** Two
- **c.** Three
- **d.** Four

6. True or False: Centrifuges are used to process all laboratory specimens before testing. Outcome 6-7

7. What are two tasks that a medical assistant may be asked to perform with a microscope? Outcome 6-6
- **a.** Maintenance and cleaning of the microscope
- **b.** Prepare and focus specimens on the microscope for examination by the health-care provider
- **c.** With appropriate training, examination of urine sediment
- **d.** All of the above

8. True or False: Automated urine analyzers are used to test for bacteria and other suspended objects in urine specimens. Outcome 6-12

9. What are two types of CLIA-waived chemistry instruments presented in this chapter? Outcome 6-14

10. Which company is mentioned in the text as a common manufacturer of hematology instruments? Outcome 6-16
- **a.** Bayer
- **b.** Dimension
- **c.** Beckman Coulter
- **d.** Piccolo

11. Why is a cover slip used? Outcome 6-17
- **a.** To view specimens that have been preserved with an additive
- **b.** To view specimens in their natural state
- **c.** To view large specimens without the use of a slide
- **d.** To view chemical reactions

Case Study 6-1: This test won't work!

The manager for the internal medicine office where Cindy Lou is employed as a medical assistant is on vacation. In her absence, the lead medical assistant has placed a few orders for laboratory supplies, including some test strips for the glucometer used for glucose testing.

On Tuesday morning, Cindy knows that a patient is scheduled for diabetes monitoring. This usually involves a blood glucose level to be performed during his visit. Cindy decides to process the quality control specimens for the glucometer first thing in the morning so that she can be ready for the patient's arrival.

Cindy turns on the instrument, allows it to perform the internal verification that it always goes through, and inserts a test strip to process the quality control specimen. The test strip seems to fit in the instrument a bit differently from the way it has before, but she continues with the process. After inserting the strip and adding the QC material, the display screen flashes an ERROR code, and no result is displayed for the test. Cindy removes this strip and inserts another one, but receives the same result. She checks the QC material to see if it might be expired, and finds that it is not. Once more Cindy attempts to process a QC specimen of a different level (high glucose level) but receives the same code.

1. What is the most probable explanation for the ERROR code?
2. Can Cindy Lou process a patient specimen despite the ERROR codes?

Case Study 6-2: Lack of focus

Lucille has been working in the laboratory at the local obstetrics and gynecology office for several years. She routinely performs blood draws and several different CLIA-waived laboratory tests. The providers in the office have decided that they want to start performing microscopy procedures in the office to better serve their patients.

To perform the microscopy procedures, the providers in the office ask Lucille to demonstrate how well she can focus the microscope. Lucille has not worked with a microscope since she was in medical assisting school, but she does her best to remember the steps involved. She is successful with the initial focus using the 10X objective, but can't seem to bring the specimen into focus when she switches to the 40X objective.

1. What are two possible explanations for the lack of focus with the 40X objective?

RESOURCES AND SUGGESTED READINGS

"US Centrifuge Systems: Frequently Asked Questions"
Frequently asked questions about how a centrifuge works, and types to choose from http://www.uscentrifuge.com/faq.htm

"The Clinical Laboratory Improvement Act and the Physician's Office Laboratory"
Various educational modules and mini quizzes for CLIA waived hematology and chemistry tests http://www.medicine.uiowa.edu/CME/clia/default.asp

"Tests Waived by the FDA From January 2000 to Present"
List of all CLIA waived tests; updated regularly http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfClia/testswaived.cfm

"Principles of Spectrophotometry"
Great explanations about how spectrophotometry works http://www.ruf.rice.edu/~bioslabs/methods/protein/spectrophotometer.html

"Welcome to HemoCue"
Information about the various products manufactured by HemoCue http://www.hemocue.com

"Glossary of Microscope Terms"
Explanations for many of the terms used with microscope use http://www.microscope-microscope.org/basic/microscope-glossary.htm

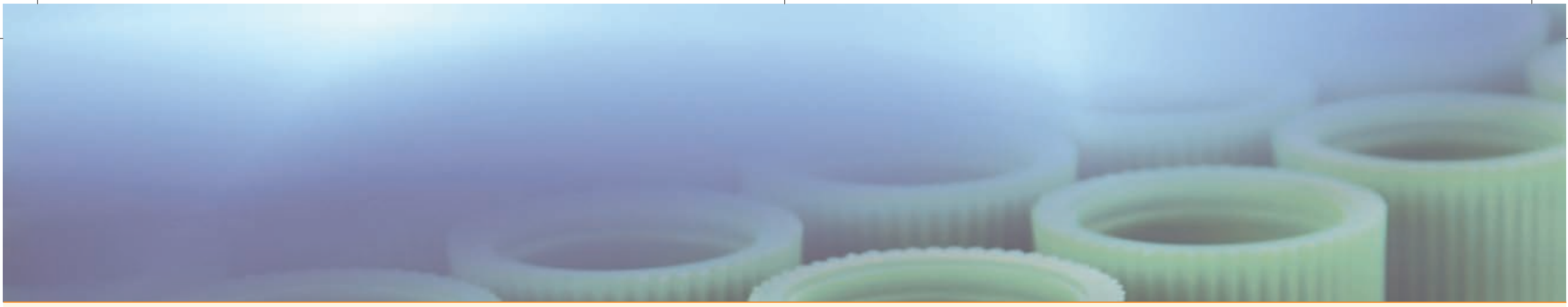

Section I

Overview of the Laboratory

What Does It All Mean?

The purpose of this section is to introduce you to global aspects associated with the clinical laboratory and laboratory testing. As this section clearly indicates, laboratory testing is crucial to the diagnosis and monitoring of patient diseases and conditions, as well as to rule out any diseases and conditions. This being said, an alert medical assistant will have a comprehensive understanding of these aspects. Further, the medical assistant will use appropriate corresponding techniques and procedures to ensure proper laboratory sample collection for testing by appropriately trained and educated individuals. Our Case in Point for this section explores some of these important concepts and serves as a review of the section content.

Case in Point

As noted in this case, during the first week of your student practicum you are introduced to the clinical laboratory by your clinical instructor, Doris. Your patient, Mr. Hershey, presents to Maple Grove Clinic for evaluation. After examining Mr. Hershey, Dr. Pueblo determines that he requires laboratory testing to determine his condition. This situation gives you and Doris a wonderful opportunity to discuss important concepts associated with both the clinical laboratory and laboratory testing. Highlights of this discussion follow for your review and consideration.

There are numerous ways in which laboratories are structured and organized based on a number of factors, including the environment in which the laboratory exists (it may be a small rural hospital or large commercial reference laboratory). Each laboratory determines what tests it will run, what instrument it will use for the tests, and the values considered as normal for the typical patient population the laboratory serves. There are three phases of laboratory testing: preanalytical, analytical, and postanalytical. Every laboratory test passes through all three phases. An error in any of the three phases may adversely affect the laboratory results generated. Most of the problems encountered in laboratory testing occur during the preanalytical testing phase because of a variety of issues, for example, collecting the sample on an individual other than the targeted patient.

Laboratory tests are categorized based in part on the difficulty of the procedure required to obtain the results. Health-care support individuals, including medical assistants, are allowed under government regulations to perform select testing under strict guidelines. Handling every sample using universal precautions is an important point to emphasize, as doing this contributes to the validity of the test results obtained. In addition to running the laboratory tests, quality control samples must also be tested on a regular basis to ensure that the test is working properly. Laboratory test results can only be released (in laboratory jargon this is called “turned out”) to the patient’s chart if the quality control samples test properly. In the event that quality control results are not in range, investigation and resolution of the problem must occur before patient results can be considered valid. Individuals vary in many ways. Variations in laboratory test values (as well as quality control samples) are no exception. Because of this, laboratory test results considered as being normal fall within a range of values.

In conclusion, laboratory tests provide physicians and other primary care providers with valuable information. In fact, many sources have reported that some 70% to 80% of diagnosis and treatment decisions are based on laboratory results. As this section clearly suggests, there are many important aspects to laboratory structure and testing, all of which contribute to the bottom line: accurate laboratory tests performed in a timely manner.

On the Horizon

It has been widely documented that as many as 70% to 80% of all patient treatment decisions are based on laboratory test results. Similarly, there is strong evidence to support the fact that laboratory test results are only as good as the sample from which the testing occurs. Specimen collection and processing, also known as the preanalytical phase (meaning the steps before actual analysis occurs) of testing, is thus a very important factor that contributes to the reliability of the laboratory test results obtained. Results from such samples collected and/or processed improperly cannot be considered valid and thus are of no benefit to the patient. There is widespread evidence that most laboratory testing errors occur in during the collection and processing of specimens.

Relevance for the Medical Assistant (Health-Care Provider)

Medical assistants and other health-care support personnel are often called upon to assist in the collection of, perform the collection of, instruct patients on the collection of and process samples for laboratory testing. To ensure the most reliable results for the patient, an in-depth understanding of specimen collection and processing is of paramount importance.

Case in Point

After a relatively smooth morning, your first patient after lunch is Wilma F., a 70-year-old woman. After reviewing Wilma's file, you notice that she was seen and treated 3 weeks ago, on what just happened to be on your day off, for a urinary tract infection. As you help Wilma get onto the scale to weigh her and then again when you place the blood pressure cuff on her arm, you notice that she is very hot to the touch. Your suspicions are confirmed when you take her temperature and it is 101.5°F! You ask Wilma to identify all symptoms she has been experiencing. She tells you that in addition to feeling hot all the time, she has a burning sensation and pain during urination. You document these details and tell Wilma that the doctor will be in to see her shortly and you leave the examining room. After the doctor examines Wilma, he asks you to collect her blood and assist Wilma in collecting urine for urinalysis and culture.

Questions for Consideration:

- What special considerations must be addressed to collect the blood in an appropriate manner in this case?
- What special equipment do you need to collect this blood sample?
- What type of urine sample should you assist Wilma to collect?
- How is this type of urine sample collected?
- Why are the blood and urine collection processes so important to implement in this case?

Section II

Specimen Collection and Processing

As one might imagine, there are many things that must be considered in this situation to ensure that laboratory specimens are properly collected and processed. That being said, this section is divided into four chapters, each with a targeted focus.

Chapter 7: Overview of Specimen Collection and Processing consists of an in-depth discussion of the proper ordering and documentation associated with specimen collection. The name of the laboratory test(s) being ordered, methods of collection, patient preparation, and proper information for specimen labeling are all addressed. Specimen collections that involve a process known as the chain of custody are identified and described.

Chapter 8: Collection and Processing of Blood Samples explores the anatomy and physiology of the cardiovascular system as well as body sites both suitable and unsuitable for the collection of blood. The equipment and procedures for successful venipuncture and capillary puncture specimens and the process of creating peripheral blood smears is detailed.

Chapter 9: Collection and Processing of Urine Samples introduces the reader to the most common types of urine samples, including but not limited to catheterized, midstream-clean catch, first morning collection, and timed collection. The purpose of each specimen type is identified.

Chapter 10: Collection and Processing of Samples for Microbial Studies covers the general requirements for proper collection of samples for microbial studies. Samples described include throat swabs for strep screen and culture, sputum, urine, blood (using aseptic technique), wounds, and stool. Samples for saline wet preps and KOH preps are also described. Processing supplies, particularly media and techniques used to support growth of microorganisms, are covered.

Chapter 7

Overview of Specimen Collection and Processing

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

Specimen Ordering
  Required Items for All Laboratory Tests
  Test Specifics
    Medicare-Approved Panels
    Standing Orders
    Reflexive Testing
Patient Identification
  Verification of Patient Identity
    Acceptable Patient Identifiers
    Electronic Bar Codes
Patient Preparation
Specimen Collection at Home
Items That Must Be Documented With Specimen Collection
Labeling Information
Chain of Custody
Summary
Time to Review
Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

7-1 Define each of the key terms.

7-2 List the required items necessary for ordering any laboratory test.

7-3 Compare and contrast organ- or disease-specific panels that are approved by the Centers for Medicare & Medicaid services (CMS) to custom panels created by some laboratories.

7-4 Explain how standing orders are used in the laboratory.

7-5 Describe reflexive testing and explain how it may benefit patients.

7-6 Describe acceptable methods used to verify patient identity.

7-7 Demonstrate understanding of the importance of proper patient preparation prior to specimen collection.

7-8 Instruct patients in how to fast for a laboratory test.

7-9 Explain the importance of timing for peak and trough medication levels.

7-10 Explain why thorough documentation of specimen sources is so important.

7-11 Describe what to include when documenting specimen collection.

7-12 Describe appropriate labeling procedures for laboratory specimens.

7-13 Describe the critical aspects involved in a specimen collection that requires a chain of custody.

CAAHEP/ABHES STANDARDS

CAAHEP Standards Met

IV.P. Psychomotor Skills, Concepts of Effective Communication, #6: Prepare a patient for procedures and/or treatments.

ABHES Standards

10. Medical Laboratory Procedures, Specimens

KEY TERMS

1 or 2 hr postprandial (1 or 2 hr pp)
Chain of custody
Fasting
First morning void
Glucose challenge
Glucose tolerance test
Peak collection level
Random specimen
Reflexive testing
Standing orders
Trough collection level

Proper specimen collection and processing are critical components of laboratory testing as part of the preanalytical phase. As introduced in Chapter 1, certain factors in the collection and processing of specimens are universal: clear, concise patient preparation instructions, verification of patient identification, labeling procedures, and source documentation are all utilized for specimen collection procedures, regardless of the specimen source. To have a positive impact on patient treatment and outcomes, health-care employees who collect specimens or instruct patients to prepare for specimen collection must be knowledgeable about the techniques required for quality procedures.

SPECIMEN ORDERING

Laboratory testing begins with orders for specimens to be collected. Sometimes this step may ambiguous. Communication from the health-care provider may be unclear for various reasons. Examples of potential problems include verbal orders that are misunderstood, the use of nonstandardized abbreviations, and handwritten orders that are difficult to read. It is imperative that everyone involved with the process understand how critical the initial data used to order a test may be.

Required Items for All Laboratory Tests

A laboratory requisition form is used to order specific laboratory testing procedures, but the requisition is also used to provide information needed for reimbursement for these procedures. In addition, the test ordered must always be documented in the patient's chart.

Laboratory requisition forms may vary; some are organized with the tests listed in alphabetical order, whereas others may be organized with the tests listed in categories. Regardless of the organization of the laboratory requisition form, there are certain items that are essential for the collection process to be carried out appropriately. Many of these items are introduced in earlier chapters in the textbook. At a minimum the requisition form should include the following:

1. Complete demographic information for the patient. This includes the full name and birth date, the patient's gender, and the billing information.
2. Name of ordering health-care professional and documentation of any other additional agencies or health-care professionals that should receive copies of the laboratory report.
3. Date for the test collection. A specific date may be a critical factor, or there may be a bit more flexibility for the collection date. For instance, sometimes a laboratory test will be ordered to be checked for a medication blood level after the patient has been on that particular medication for several days; this date should be noted on the requisition form so that the patient knows when to have his or her blood drawn.
4. Specific documentation for the tests ordered. Each test must be marked carefully and completely with an X or a circle around the test. Each test will be accompanied by a current procedural terminology (CPT)

code (printed on the requisition form next to the test), which may help with definitive choices when a provider is unsure of which test to order.

5. Diagnosis codes (in the form of an ICD-9 code) must be included with all laboratory tests ordered. Each test must have a diagnosis code specifically assigned to it, rather than one or two codes written randomly on the requisition form, as they may not apply to all the tests ordered. A diagnosis code is required for reimbursement.
6. If a test may be performed on more than one type of body fluid, it is very important to specify which fluid is to be used for the analysis. For instance, cortisol levels may frequently be performed using serum or urine; if the source is not documented correctly, then the patient may be required to return to resubmit the specimen.
7. When a medication level is ordered, the time of the last medication dose should be included on the requisition form, or the time of the requested blood draw should be specified. This may be critical for appropriate interpretation of the results.
8. If a wound culture is required, the source must be documented carefully. Specimen processing for microbiological specimens will vary depending on the site of the collection and the type of sample.
9.

Test Your Knowledge 7-1

List three items that must be present on a laboratory requisition form when ordering a specimen collection. (Outcome 7-2)

Test Specifics

Laboratory tests may be ordered in several ways. If a health-care provider feels that only one analyte needs to be tested to assist with diagnosis and treatment, he or she will order a single test, such as a glucose level for a patient who is being screened for diabetes, a potassium level for someone who is taking diuretics, or a hematocrit test for someone who is being treated for anemia. However, it is often beneficial for the health-care provider to order a group of tests that have been designed to enable providers to make more accurate diagnoses. These groups of tests may be *custom panels* that have been designed by a specific laboratory or specialty. A thyroid panel, for instance, may include three different tests that measure different aspects of thyroid function. These custom panels are not universal in nature; what is included in a custom panel by one laboratory will not necessarily be the same as those included in a panel by the same name with another laboratory.

Medicare-Approved Panels

Health-care providers must be careful when ordering panels, especially for those patients who have Medicare coverage for their payment source. The Centers for Medicare & Medicaid Services (CMS) has adopted several organ- or disease-specific panels that were designed by the American Medical Association (AMA). These panels will be the same no matter which laboratory uses them. These are the *only* panels that are always covered by Medicare without the need of an Advanced Beneficiary Notice of Noncoverage (ABN), even though some of the tests included in the panel have limited Medicare coverage. Care must be taken, however, to ensure that all the tests in the panel are really necessary for the treatment of the patient. If there is no documentation in the patient's chart to show necessity for the entire panel, the health-care provider may be held responsible for fraudulent ordering practices. The Medicare-approved organ- or disease-specific panels are the following:

- Basic Metabolic Panel(s); with or without ionized Calcium
- Electrolyte Panel
- Hepatic Panel
- Hepatitis Panel
- Comprehensive Metabolic Panel
- Renal Function Panel
- Lipid Panel

Laboratories must follow strict reimbursement rules for tests that are included in panels. The total cost of the panel may not exceed the totals of the separate tests included in the panel. Table 7-1 itemizes the tests included in the CMS-approved panels.

TABLE 7-1

Medicare-approved panels

Test Name	CPT Code
Comprehensive Metabolic Panel (CPT 80053)	
Albumin	82040
Bilirubin, total	82247
Calcium	82310
Carbon dioxide	82374
Chloride	82435
Creatinine	82565
Glucose	82974
Alkaline phosphatase	84075

Continued

TABLE 7-1—cont'd

Medicare-approved panels

Test Name	CPT Code
Potassium	84132
Total protein	84155
Sodium	84295
Transferase, alanine amino (ALT, SGPT)	84460
Transferase, aspartate amino (AST, SGOT)	84450
Blood urea $$	84520
Hepatic Function Panel (CPT 80076)	
Total protein	84155
Albumin	82040
Alkaline phosphatase	84075
Transferase, aspartate amino (AST, SGOT)	84450
Transferase, alanine amino (ALT, SGPT)	84460
Bilirubin, total	82247
Total protein	84155
Basic Metabolic Panel With Ionized Calcium (CPT 80047)	
Glucose	82974
Potassium	84132
Sodium	84295
Blood urea nitrogen	84520
Calcium, ionized	82330
Carbon dioxide	82374
Chloride	82435
Creatinine	82565
Basic Metabolic Panel With Calcium (CPT 80048)	
Glucose	82974
Potassium	84132
Sodium	84295
Blood urea nitrogen	84520
Calcium	82310
Carbon dioxide	82374
Chloride	82435
Creatinine	82565
Renal Function Panel (CPT 80069)	
Glucose	82974
Potassium	84132
Sodium	84295
Blood urea nitrogen	84520
Calcium	82310
Carbon dioxide	82374
Chloride	82435
Creatinine	82565
Phosphorus	84100
Lipid Panel (CPT 80061)	
Cholesterol, total	82565
Triglycerides	84478
Electrolytes (CPT 80051)	
Potassium	84132
Sodium	84295
Carbon dioxide	82374
Chloride	82435

Current Procedural Terminology © 2011 American Medical Association, All Rights Reserved.

Test Your Knowledge 7-2

How are custom panels similar to the Medicare-approved panels? (Outcome 7-3)

Standing Orders

Frequently health-care professionals will establish a set of **standing orders** for a patient, meaning that the patient is to have a specific test performed at a certain time interval for a period of time. For instance, a patient may need to have a potassium level checked every 3 months to monitor progress while on a certain medication. To set up the standing order, the physician will submit a requisition form to the laboratory with the initial potassium level ordered, but the requisition form may also include additional orders to repeat this test every 3 months. Some laboratories may also have a separate form used to transmit information about standing orders, rather than using the initial requisition form. Standing orders may not be continued for more than a year without a new order being written by the health-care provider. Recently, the practice of using standing orders has been questioned. In 2007, there was a change to the Physician Fee Schedule update, which specified a need for a separate order for every blood glucose test performed for patients in certain settings; the rule emphasized that a standing order was not sufficient for documentation of necessity for these patients. Even though this ruling addressed only blood glucose, it does serve as a warning for future policy changes in reference to standing orders.

Test Your Knowledge 7-3

A standing order for a theophylline drug level is written so that the patient is to have her medication level tested monthly for 18 months. Is this order acceptable? Why or why not? (Outcome 7-4)

Reflexive Testing

Another practice that may be used in a clinical laboratory is **reflexive testing.** In this situation, additional testing is automatically performed in response to certain abnormal results. An example of reflexive testing is the antibiotic sensitivity testing performed on culture samples. If a culture specimen has a positive result, an antibiotic sensitivity is automatically performed to provide the physician with additional knowledge about the best way to treat the infection. Without this reflexive option, it would be necessary to contact the physician and request a separate order for this sensitivity to be performed, which would delay patient treatment.

The addition of reflexive tests allows the ordering physician to receive additional information in a timely manner rather than waiting until the initial sample result is transmitted back to the physician and another order is placed for a new sample to be collected. In order for reflexive testing to be performed within the guidelines necessary for reimbursement, it is essential that the ordering physician is made aware of the tests included in the reflexive testing panel, and it is also necessary that the physician be offered the opportunity to order these secondary tests separately without the reflexive option if needed. These reflexive tests are to be documented clearly on the requisition form.

Test Your Knowledge 7-4

List one way that reflexive testing benefits patients. (Outcome 7-5)

PATIENT IDENTIFICATION

As presented in Chapter 1, the medical assistant or phlebotomist responsible for collecting samples for analysis plays a critical role. Appropriate patient identification and documentation of the collection procedures used are very important components of this preanalytical process.

Verification of Patient Identity

Whenever a patient presents for specimen collection, it is imperative that the person performing the collection verify the identity of the patient. This is the most important procedure in the specimen collection process. A minimum of *two* unique identifiers must be used to verify patient identification. Incorrect identification of the patient may result in specimen collection on the wrong patient, and misdiagnosis of the patient because of incorrect laboratory values. Drawing the wrong patient may even have fatal consequences if the treatment plan developed is incorrect.

Acceptable Patient Identifiers

- **Name:** It is always best to ask the patient to state his or her name; do not ask, “Are you Mrs. Smith?” as many patients may say “Yes” without actually hearing or understanding what was asked. Ask the patient to state his or her name, and spell the last name.
- **Birth date:** The patient must also state his or her birth date. It is possible to have two patients with the same first and last name, but improbable that they will also share the same birthday.
- **Social Security number or patient ID:** The use of the Social Security number is no longer recommended in most cases, with the exception of those who are receiving care in a government facility. A patient ID may have been assigned to the patient (especially in the case of managed care clients or hospital settings) and this may serve as a unique identifier.
- **Driver’s license or other picture identification:** For legal specimens or those that are associated with employment, a picture ID is necessary for definitive identification. It is an acceptable means of identification in all situations, but not always necessary. In many office settings, the picture ID is part of the patient’s permanent health record.

In hospitals or other settings in which the patient may not be able to converse with the collector, it is especially important that the identity of the patient is established before the specimen is collected. Never use the name written or printed above the bed as that of the patient. If the patient has an armband that has been placed on the wrist, this may be used if the patient is not capable of verifying his or her identity verbally. Never use an armband for identification if it is not securely placed on the patient’s arm. If the patient is not wearing appropriate identification, it will be necessary to locate the health-care professional responsible for this patient (usually one of the nursing staff) and ask him or her to reband the patient and provide definitive ID before collecting the specimen. In this situation, be sure to document the name of the person who provided the identification.

Test Your Knowledge 7-5

Should an armband taped on the side of a hospital bed not be used for patient identification? (Outcome 7-6)

Electronic Bar Codes

Most likely, not many of us remember a time when we didn't have bar codes on our groceries and other items that we purchase in retail stores. Bar codes allow for more specific identification of the product purchased, and provide an efficient way to update prices and product inventory. The health-care industry has also found that the use of electronic bar codes helps to streamline the process of ordering laboratory tests. Using labels with electronic bar codes on the individual tubes and specimen containers helps to assure specimen identification and patient confidentiality, and allows for automated sorting and distribution of the specimens in large laboratory settings. It is especially important to verify identification in situations in which bar codes are used, because the individual specimen containers do not necessarily include the patient's name. The individual specimens may only be identified with the bar code or specimen ID label that represents that patient in the electronic system.

PATIENT PREPARATION

In addition to the verification of patient identity and demographic information, the medical assistant responsible for specimen collection must also verify whether the patient prepared appropriately for the test ordered prior to collection of the specimen. The reference ranges for laboratory tests are based on specific parameters, and if these were not in place for the patient being tested, their results may appear to be abnormal. Examples of specific preparation procedures include the following:

- Fasting: A **fasting** specimen may be required for tests such as blood glucose levels or lipid testing. Fasting means that the individual must go without food and any liquid other than water for 12 hours prior to the blood draw. The patient usually can take medication that has been prescribed if it poses a health risk to skip or postpone a dose until after the period of fasting. If the patient did take medication during this fasting period, it must be documented on the requisition form.
- 1 or 2 hr postprandial: A **1 hr pp or 2 hr postprandial (1 or 2 hr pp)** blood draw means that the specimen is collected 1 or 2 hours (respectively) after the individual eats a meal. These pp samples are usually used for blood glucose testing.
- Glucose challenge: When patients are screened for diabetes (especially for pregnant women) a **glucose challenge** may be ordered. In this situation the patient is given a drink that contains 50 g of glucose to ingest. One hour after the sweetened drink is consumed, the patient has his or her blood drawn again for a glucose test.
- Glucose tolerance test: A **glucose tolerance test** has several steps, and is usually only ordered if the fasting blood sugar, postprandial blood sugar, or glucose challenge results are abnormal. In this situation, a blood sample is drawn from the fasting patient. (The patient usually provides a urine specimen as well.) The patient is then asked to drink a solution that contains 100 g of glucose. Subsequent blood draws are performed every half hour for a period of 3 to 6 hours. (This process is presented in more detail in Chapter 17.)
- Timed drug levels: Reference ranges for blood medication levels are based on the rate at which different drugs are metabolized. Some types of medication peak in the bloodstream relatively quickly after ingestion, while others are slow acting. The medical assistant or phlebotomist needs to verify (and document) when the patient last took their medication before performing the collection. Some drugs may have therapeutic ranges that are very close to the levels where the drug may harm the body, and the timing of the draw is essential to determine whether the dosage should be changed. Drugs may also be ordered as **peak** or **trough collection levels.** These terms mean that the specimen is to be drawn when that particular drug is at the highest level (peak) and/or the lowest level (trough) in the bloodstream. Because each medication is absorbed and cleared from the body differently, the time after the last dose of the medication for the peak and trough draw will vary. The peak levels are generally obtained 1 to 2 hours after a drug is given; the trough levels are usually just before the next dose is due. With peak and trough levels, if the medical assistant does not verify the time of the last dose, the result may appear to be abnormal when in actuality the specimen was just drawn at the wrong time.
- Restricted diets for specimen collection: Some laboratory tests performed on blood, stool, and urine

require that the patient restrict his or her diet before and/or during collection. Examples may be avoidance of rare meat for a few days before the collection of a stool specimen for fecal occult blood testing, or avoidance of certain foods during the collection time for a 24-hour urine specimen.

- Random specimens: **Random specimens** do not require special preparation before collection. The concentration of many chemical substances and cellular components present in the bloodstream does not change throughout the day, and can be collected at any time. This term is also used to describe urine specimens used for basic urinalysis testing, as well as many microbiology specimens.
- First morning void: During sleep, the body continues to create urine as the blood circulates through the kidneys. The **first morning void** specimen may be requested for specific urine chemistry tests, as the concentration of the analyte will be elevated in this specimen because it has been in the bladder all night. A microalbumin test is an example of a test that is often ordered as a first morning void specimen.

Test Your Knowledge 7-6

Mr. Rider comes in for a blood draw as part of his annual physical. The physician has ordered a fasting blood glucose level. Just as the medical assistant begins to prepare for the blood draw, she notices that the patient has a coffee mug in his hand. What should she ask him before she begins the draw? (Outcome 7-8)

SPECIMEN COLLECTION AT HOME

Urine, stool, and microbiology specimens are often collected by patients in the privacy of their homes and delivered to the laboratory soon after. With this type of collection, special care must be taken to educate the patient properly in reference to specimen collection and storage. Information for patient use should be offered verbally and in writing to ensure understanding and compliance. Collection and storage techniques should be verified again before accepting the specimen at the laboratory. For instance, a urine specimen collected at home should be collected using a sterile container, and the specimen must be refrigerated and transported to the laboratory as soon as possible after collection. In contrast, stool specimens collected for culture are not to be refrigerated, but should arrive at the laboratory within 2 hours of collection. Some specimens must remain protected from light, whereas other specimens need to be added to containers with preservatives immediately after collection. It is the responsibility of the medical assistant who accepts these specimens from the patient to verify that the collection and processing have been performed correctly. Prior patient education is critical for quality specimens.

ITEMS THAT MUST BE DOCUMENTED WITH SPECIMEN COLLECTION

Documentation does not end with the collection of the specimen. It is essential that details about the collection are recorded on the requisition form, as these details help to provide critical information that may affect the test results.

- **Who?** The initials (or other unique identification, such as an employee ID) for the person who collected the specimen must be documented on the requisition form. This information may be critical if there are any problems with the specimen. This information is also placed on the blood tube after collection while in the presence of the patient.
- **What?** What type of specimen was collected? For blood specimens, there should be documentation of the type of specimen collected because venous, arterial, and capillary samples may have different reference ranges for a specific analyte. Reimbursement for the collection may also vary with the type of blood specimen. For urine specimens, there should be documentation as to whether the sample is a random or timed urine collection, and the method used for the collection process. (This is further clarified in Chapter 9.) Body fluids (such as amniotic fluid or synovial fluid) are especially difficult to identify by sight, so clarification of the specimen type is essential at the time of collection; reference ranges are very different for each type of body fluid.
- **When?** Documentation of the date and time of collection is critical. The time must include a.m. or p.m., or military time may be used. Many chemical analytes change drastically with time once the blood has left the body, and these natural postcollection changes may be misidentified as part of a disease process if the specimen is not processed in a timely fashion. For timed collections, it is essential. The date and time of collection are added to the requisition form as well as the tube after a blood draw is complete.
- **Where?** The site of collection must also be documented. For blood draws, this does not need to be absolutely specific; a general location such as right

hand or left arm will do. This documentation can be very helpful if there are complications at the draw site. For culture specimens, documentation of the collection site should be as specific as possible, because the normal florae present in different areas of the body are not the same. The physical location of the culture collection site needs must be specified (right arm, left leg, opening of fistula, etc.) as well as the type of wound or collection procedure used. A superficial wound sample is different from that collected from a deep puncture wound, and a specimen collected by aspiration is different from a specimen collected by squeezing a pustule. For instance, bacteria that are common on the surface of the skin are *not* part of the normal florae in a deep wound, even if the specimens are both taken from the right great toe.

Test Your Knowledge 7-7

Why is it important to document that a blood specimen was obtained from a capillary puncture rather than a vein? (Outcome 7-10)

LABELING INFORMATION

All collection containers must be labeled immediately after collection, meaning that the specimen must be labeled in an outpatient setting while in the presence of the patient and before another outpatient requisition form is accepted. In an inpatient setting, the specimen labeling process must be complete before the phlebotomist leaves the patient's room. Labeling may involve writing with an indelible pen directly on the preaffixed label on the tube after collection, or by applying a label that has been generated by a computer and adding a few items. Computer-generated labels often include bar codes that may be read by the instruments used for testing. Tubes should never be prelabeled before specimen collection, as this can lead to incorrect patient identification if the collection is unsuccessful and the tubes are not discarded, or if there is more than one patient present in a blood drawing area at the same time. Specimens that are mislabeled can potentially cause serious harm to patients when inaccurate patient results are generated.

The specimen label must include the following information:

- The name of the patient and a unique identifier number. The ID number is generated by the computer at the time of the blood draw, or it may be part of the requisition form available as a small peel-off label. If the patient number is not available, the date of birth may be used instead. Two patient identifiers are required.
- Date and time of collection
- Initials (or other unique ID) for the person who performed the specimen collection
- Other required information may include documentation of a peak or trough draw, or at what time a specimen is taken for a glucose tolerance test. Even though the time of the draw should indicate which specimen is in the tube, it is a good idea to document this in more than one way to avoid any confusion.

For patients who collect their specimens at home, the person who accepts the specimen at the laboratory must verify that the specimen is labeled appropriately while in the presence of the patient. If a patient drops off a specimen that is not labeled and no one is present to accept the specimen, it will need to be recollected.

Test Your Knowledge 7-8

List two items that must be included on the label after a blood specimen is collected. (Outcome 7-12)

CHAIN OF CUSTODY

Sometimes blood or urine is collected in situations in which the test results may be used as legal evidence. Blood may be collected for DNA analysis or blood alcohol level readings, whereas urine may be tested for drugs of abuse. The tests may be ordered in preemployment situations, in the case of on-the-job accidents or suspected drunk driving, or for some athletes. In these situations, it is imperative that medical assistants follow their facility's guidelines very carefully as they perform their duties. The specimens that are collected must have written documentation for their processing and storage from the point of collection through the testing and reporting process. It must be evident that a responsible individual had possession of the specimen, or had secured the specimen at every step of the process. The record of the individuals who have access to the specimen is called the **chain of custody.** A chain of custody is a physical object: a form that is filled out and signed each time that the specimen changes hands. An example of a chain of custody form may be found in Figure 7-1. Critical components of the process include the following:

1. **Identification of the patient:** This is the most important part of the process. Sometimes patients are not willing participants in the process, as they may have been placed under arrest, or may have been involved in some sort of incident at their place of work. If the patient is not able (or willing) to show appropriate identification, the arresting officer or supervisor must identify the patient and it must be documented on the form.
2. Explanation of the process and witnessed signed consent form.
3. **Secure collection area and standardized process:** Especially in the case of urine specimens, there are very specific procedures that must be followed to eliminate the opportunity for patients to add water or other substances to the specimen, or to substitute a specimen that they may have brought from outside the premises. For instance, the patient is not allowed to take any coats or purses into the restroom during the collection. The handles of the faucet in the collection area may be secured with tamper-evident tape, or the water supply may be shut off from outside the room. In addition, the toilet water always has a bluing agent added to avoid an opportunity to take water from the toilet unnoticed. The specimen temperature is documented immediately when the collection is complete as a means of quality assurance.
4. **Assignment of a sample number:** Legal specimens should not be identified by name after the initial collection process. This helps to protect the patient's confidentiality, and also limits the opportunities for someone to tamper with a specific specimen.
5. **Securing the specimen and protecting against tampering:** The specimen has a tamper-evident seal that is affixed at the time of collection. The patient and the collector initial this seal and the specimen is sealed in the presence of the patient. If this seal is broken when the sample is accessed for testing, the chain of custody is invalidated.
6. **Documentation of processing:** The specimen must remain in a secure location at all times. The chain of custody form provides an opportunity to document where the specimen was stored, and who placed it in that location at each step of transfer. Legal specimens must be kept in a locked location with limited authorized access until they are tested. The documentation continues until the time that the specimen is disposed of.

There may be times when the medical assistant is called to testify about the collection process of a legal specimen. If the chain of custody process is followed carefully, there should be no problem with testifying regarding the specimen processing. Poor documentation may be reason for dismissal of a legal case, and could have serious repercussions.

Test Your Knowledge 7-9

What are two methods used to protect a legal specimen from potential tampering? (Outcome 7-13)

POINT OF INTEREST 7-1

DOT drug screening

Urine may be tested for drugs of abuse for many reasons. The test is a common part of preemployment screening for many corporations. Most of the large retail store chains now require drug screening as a preemployment procedure, and some of them also use random drug screens as well.

There are also times when the federal government requires drug screen testing. The Omnibus Transportation Employee Testing Act of 1991 set a standard that requires drug and alcohol testing for employees who are employed in professions that the government has designated as "safety-sensitive." This regulation potentially affects more than 10 million individuals. These employees may be working in the following areas:

- Those in the trucking industry
- Railroad workers
- Bus drivers or other mass-transit employees
- Pilots and other aviation employees
- Employees in other types of public transportation
- Those who work on public pipelines

The U.S. Department of Transportation (DOT) has set standards concerning who will be tested, as well as the policies and procedures to be followed for these individuals. The collection process is even more critical in these situations, and the laboratory testing may be performed only at laboratories that have been approved by the DOT. Special chain of custody forms are used for the collection process, and the procedures for processing of the paperwork are very specific.

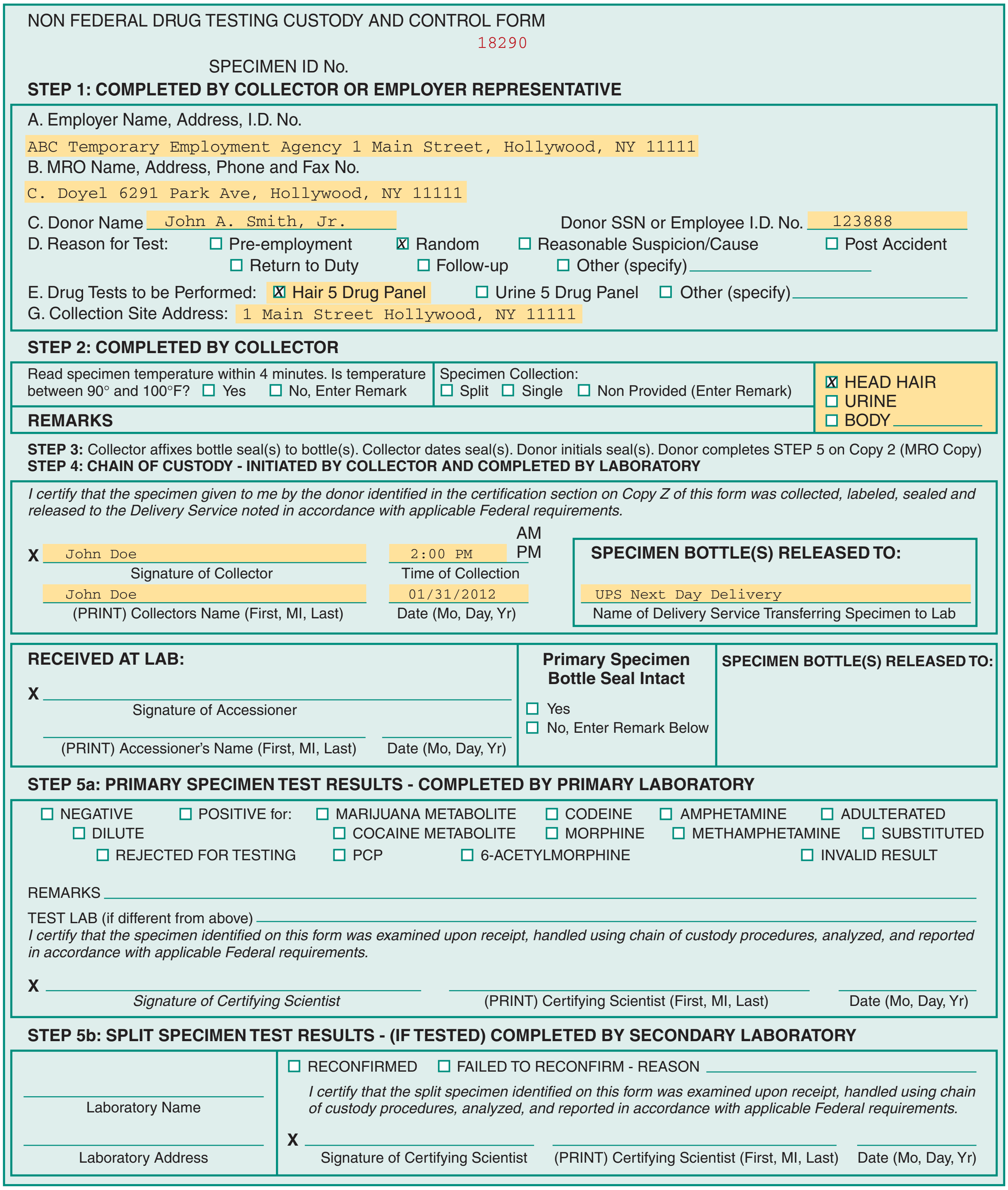

NON FEDERAL DRUG TESTING CUSTODY AND CONTROL FORM

18290

SPECIMEN ID No.

STEP 1: COMPLETED BY COLLECTOR OR EMPLOYER REPRESENTATIVE

A. Employer Name, Address, I.D. No.

`ABC Temporary Employment Agency 1 Main Street, Hollywood, NY 11111`

B. MRO Name, Address, Phone and Fax No.

`C. Doyel 6291 Park Ave, Hollywood, NY 11111`

C. Donor Name `John A. Smith, Jr.` Donor SSN or Employee I.D. No. `123888`

D. Reason for Test: ☐ Pre-employment ☒ Random ☐ Reasonable Suspicion/Cause ☐ Post Accident
☐ Return to Duty ☐ Follow-up ☐ Other (specify) ______

E. Drug Tests to be Performed: ☒ Hair 5 Drug Panel ☐ Urine 5 Drug Panel ☐ Other (specify) ______

G. Collection Site Address: `1 Main Street Hollywood, NY 11111`

STEP 2: COMPLETED BY COLLECTOR

Read specimen temperature within 4 minutes. Is temperature between 90° and 100°F? ☐ Yes ☐ No, Enter Remark

Specimen Collection: ☐ Split ☐ Single ☐ Non Provided (Enter Remark)

☒ HEAD HAIR
☐ URINE
☐ BODY ______

REMARKS

STEP 3: Collector affixes bottle seal(s) to bottle(s). Collector dates seal(s). Donor initials seal(s). Donor completes STEP 5 on Copy 2 (MRO Copy)

STEP 4: CHAIN OF CUSTODY - INITIATED BY COLLECTOR AND COMPLETED BY LABORATORY

I certify that the specimen given to me by the donor identified in the certification section on Copy Z of this form was collected, labeled, sealed and released to the Delivery Service noted in accordance with applicable Federal requirements.

X `John Doe`
Signature of Collector

`2:00 PM` AM PM
Time of Collection

`John Doe`
(PRINT) Collectors Name (First, MI, Last)

`01/31/2012`
Date (Mo, Day, Yr)

SPECIMEN BOTTLE(S) RELEASED TO:

`UPS Next Day Delivery`
Name of Delivery Service Transferring Specimen to Lab

RECEIVED AT LAB:

X ______
Signature of Accessioner

______ ______
(PRINT) Accessioner's Name (First, MI, Last) Date (Mo, Day, Yr)

Primary Specimen Bottle Seal Intact
☐ Yes
☐ No, Enter Remark Below

SPECIMEN BOTTLE(S) RELEASED TO:

STEP 5a: PRIMARY SPECIMEN TEST RESULTS - COMPLETED BY PRIMARY LABORATORY

☐ NEGATIVE ☐ POSITIVE for: ☐ MARIJUANA METABOLITE ☐ CODEINE ☐ AMPHETAMINE ☐ ADULTERATED
☐ DILUTE ☐ COCAINE METABOLITE ☐ MORPHINE ☐ METHAMPHETAMINE ☐ SUBSTITUTED
☐ REJECTED FOR TESTING ☐ PCP ☐ 6-ACETYLMORPHINE ☐ INVALID RESULT

REMARKS ______

TEST LAB (if different from above) ______

I certify that the specimen identified on this form was examined upon receipt, handled using chain of custody procedures, analyzed, and reported in accordance with applicable Federal requirements.

X ______ ______ ______
Signature of Certifying Scientist (PRINT) Certifying Scientist (First, MI, Last) Date (Mo, Day, Yr)

STEP 5b: SPLIT SPECIMEN TEST RESULTS - (IF TESTED) COMPLETED BY SECONDARY LABORATORY

______
Laboratory Name

______
Laboratory Address

☐ RECONFIRMED ☐ FAILED TO RECONFIRM - REASON ______

I certify that the split specimen identified on this form was examined upon receipt, handled using chain of custody procedures, analyzed, and reported in accordance with applicable Federal requirements.

X ______ ______ ______
Signature of Certifying Scientist (PRINT) Certifying Scientist (First, MI, Last) Date (Mo, Day, Yr)

Figure 7-1 An example of a chain of custody form. The top copy contains the name of the patient, but this copy does not follow the specimen as it progresses to the testing phase. The copies that are used with the specimen after it is input into the computer are designated only by the specimen ID number.

SUMMARY

Specimen collection and processing are critical components of the laboratory testing. Care must be taken to fill out the request for testing in the proper manner, educate the patient appropriately for necessary preparation, and thoroughly document the collection procedure. It is also necessary to label the specimen using recommended procedures so that there are no potential problems with identification as the testing process continues. The type of collection, the type of specimen, and the date and time of collection are critical components that must be logged on the requisition form after collection. In legal situations, care must be taken to follow established procedures to ensure that specimens are kept safe from tampering until they are tested and the results are reported. This process uses a chain of custody form, which must be filled out by anyone who has access to the specimen during and after collection.

TIME TO REVIEW

1. When is a trough blood level drawn? Outcome 7-1
2. Which of these items does not need to be included on a complete laboratory requisition form? Outcome 7-2
   a. ICD-9 code(s)
   b. Ordering physician's name
   c. Cost of tests
   d. Demographic information for the patient
3. How are CMS panels different from custom panels created by independent laboratories? Outcome 7-3
   a. The custom panels include fewer tests
   b. The CMS panels have been approved by the AMA
   c. The reimbursement rates are better for the custom panels
   d. They are essentially the same
4. True or False: Reflexive testing causes excessive time to be spent before results are available to the ordering health-care provider. Outcome 7-5
5. True or False: All medication is absorbed and used by the body at the same rate. Outcome 7-9)
6. List three items that must be documented on the requisition form after a specimen is collected. Outcome 7-11
7. True or False: It is acceptable to use a computer-generated label on a tube after collection as long as it includes the basic required information. Outcome 7-12

Case Study 7-1: Standing orders

Mr. Johnson comes into the laboratory to have his potassium level checked. He has a standing order to have this test performed every 3 months because of a medication that he takes for a heart disorder. When he arrives at the laboratory, the phlebotomist checks the records for Mr. Johnson in the computer system. He informs Mr. Johnson that he cannot draw his blood today because his standing order has expired.

1. What does the phlebotomist mean by stating that the standing order has expired?
2. What needs to occur before Mr. Johnson can have his blood drawn for the potassium test?

Case Study 7-2: Labeling

Cassy Jones is a medical assisting student who is training in the laboratory. She is trying very hard to get all the information correct for every situation, and she tells her trainer that she is working on anticipating the needs of the practice so that she can be better organized. When her trainer returns from lunch one afternoon, she finds that Cassy has labeled blood tubes for the expected blood draws for the afternoon. Cassy's trainer tells her that although she can see that her intentions were good, this is an unacceptable practice.

1. Why is the prelabeling of the tubes an unacceptable practice?

RESOURCES AND SUGGESTED READINGS

"University of Washington Laboratory Testing Policies"
Very specific information about testing and collection processes designed for use by practitioners http://depts.washington.edu/labweb/PatientCare/Policies/index.htm

"Errors in Laboratory Medicine"
Information about common laboratory errors including all phases of the testing process http://www.clinchem.org/cgi/content/abstract/48/5/691

"Laboratory Test Ordering and Documentation; List of Do's and Dont's http://www.cap.org/apps/docs/pt_checkup/pol_library/laboratory_test_ordering.pdf

Chapter 8

Collection and Processing of Blood Samples

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Anatomy and Physiology of the Cardiovascular System
  - The Heart
  - Blood Vessels
    - *Arteries*
    - *Veins*
    - *Capillaries*
- Site Selection
  - Most Commonly Used Veins
  - Appropriate Capillary Puncture Sites
- Contraindications and Areas to Avoid
  - Scars and Tattoos
  - Mastectomy Considerations
  - Intravenous Fluids
  - Bruises and Hematomas
  - Varicose Veins
  - Central Lines and Fistulas
  - Edematous Areas
  - Obese Patients
- Venipuncture Equipment
  - Tourniquets
  - Gloves
  - Sharps Biohazardous Waste Containers
  - Alcohol and Gauze Pads
  - Needles
  - Additional Supplies
- Methods Used for Venipuncture
  - Evacuated Tube System
  - Syringe Method
  - Winged Infusion or Butterfly Method
- Blood Collection Tubes
  - Color Coding of Tubes
  - Types of Additives
- Order of Draw for Venipuncture
- Capillary Punctures
  - Order of Draw for Capillary Puncture
- Preparation for Blood Collection
  - Specimen Processing
  - Obtaining Serum for Testing
  - Obtaining Plasma for Testing
  - Whole Blood Specimens
  - Unacceptable Specimen Types
- Potential Negative Outcomes of Venipuncture and Capillary Puncture
  - Inability to Draw Blood
  - Fainting Patients
  - Rolling Veins
  - Hematoma Formation
  - Collapsing Veins
  - Nerve Damage
  - Infection
- Other Processing Procedures
- Preparation of a Peripheral Blood Smear for Staining
- Wright's Stain Procedure
- Summary
- Time to Review
- Resources and Suggested Readings

Learning Outcomes

After Reading this Chapter, the Successful Student will be able to:

8-1 Define the key terms.

8-2 Explain the route used as blood moves through the heart.

8-3 Compare and contrast the qualities of veins and arteries.

8-4 Examine the ways that capillaries differ from veins and arteries.

8-5 Locate the most commonly used veins for drawing blood specimens, and explain why they are the best choice for this procedure.

8-6 Describe where capillary blood draws may be safely performed.

8-7 List the major contraindications for blood draws, as well as areas to be avoided.

8-8 List and describe the supplies necessary to perform successful venipuncture and capillary punctures.

8-9 Differentiate when the different methods of venipuncture (evacuated tube system, syringe system, or butterfly system) might be used.

8-10 Describe the significance of the tube stopper colors when drawing blood.

8-11 Explain the order of draw for specimen collection, and describe how it may be different with the various collection techniques.

8-12 Analyze techniques to be used with different age groups when performing venipuncture or capillary puncture.

8-13 Explain why capillary punctures may be preferred as a blood draw technique in some circumstances, and provide examples.

8-14 Summarize appropriate preparation techniques for blood draw procedures.

8-15 Successfully perform venipuncture and capillary puncture procedures using the process outlined in the textbook.

8-16 Demonstrate the ability to create a blood smear on a slide that would be suitable for staining, and explain why staining is important for examination of blood samples.

8-17 Explain how to process samples for various laboratory tests ordered.

8-18 Analyze the differences between serum and plasma.

8-19 Provide examples of inappropriate specimens, and describe how these issues may be avoided.

8-20 Describe how negative outcomes may be avoided when drawing blood.

CAAHEP/ABHES STANDARDS

CAAHEP Standards:

I.C.I.4: List major organs in each body system
I.C.I.5: Describe the normal function of each body system
I.P.I.2: Perform Venipuncture
I.P.I.3: Perform Capillary Puncture
I.A.I.1: Apply critical thinking skills in performing patient assessment and care
III.P.III.3: Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations

ABHES Standards

- Medical Office Laboratory Procedures: Collect, label and process specimens: Perform venipuncture
- Medical Office Clinical Procedures: Apply principles of aseptic techniques and infection control
- Medical Office Clinical Procedures: Use Standard Precautions

KEY TERMS

Antecubital area
Anticoagulants
Arteries
Arterioles
Basilic vein
Bevel
Butterfly system
Capillaries
Capillary action
Capillary puncture
Capillary tubes
Cardiovascular
Cephalic vein
Clot activators
Contraindications
Edema
Evacuated tube system
Flanges
Gauge
Hematoma
Hemoconcentration
Hemolysis
Interstitial
Lipemia
Lumen
Mastectomy
Median cubital vein
Microcollection containers
Morphology
Multi-sample needle
Palpation
Plasma
Plasma separator tubes (PST)
Quality not sufficient (QNS)
Serum
Serum separator tubes (SST)
Sharp
Short draw
Syringe system
Tourniquet
Vacuum
Vasodilation
Vasovagal syncope
Veins
Venules
Winged infusion set

The majority of laboratory testing is performed on blood specimens. As part of the preanalytical process, medical assistants and phlebotomists are responsible for greeting and identifying patients, verifying patient identification, making appropriate site selections for collection, using correct collection procedures, and processing specimens prior to the testing stage. It is essential that the health-care professional who performs these duties understands how to execute them correctly, as the potential impact of these processes is profound if they are performed improperly. This chapter provides details about how to collect and process various blood specimens, including necessary knowledge of the anatomy and physiology of the cardiovascular system, an introduction to the various tools used for obtaining blood, and information about how to avoid negative outcomes during the collection process. Procedures for processing specimens after collection are also presented.

ANATOMY AND PHYSIOLOGY OF THE CARDIOVASCULAR SYSTEM

In order to consider the various methods used to obtain blood specimens, it is important to understand the anatomy and physiology of the **cardiovascular** system. Essentially, all the body systems are connected, or linked, by the cardiovascular system, because it offers a means of transportation to and from all the cells of the body. *Homeostasis* is a state in which the body is healthy and in balance. For homeostasis to be maintained in our bodies, the cardiovascular system must transport water, nutrients, waste products, chemicals, and hormones everywhere within the body. The immune system is also dependent on the function of the cardiovascular system. In addition, circulation of the blood helps to keep the body's temperature within normal limits, contributing to homeostasis.

The Heart

The heart is the beginning and the end of the cardiovascular system, and is approximately the size of a closed fist. It is made of very strong cardiac muscular tissue. Within the structure of the heart are four chambers. The upper two chambers of the heart are the *atria* and the lower two chambers of the heart are the *ventricles*. The two sides of the heart are separated by a wall (called the *septum*) and the atria and the ventricles on each side of the heart are separated by *valves*, which keep the blood from flowing backward within the structure. Figure 8-1 allows us to follow the path of the blood flow through the heart and lungs.

1. Oxygen-poor blood flows into the right atrium through two large veins known as the superior vena cava and inferior vena cava.
2. As the heart contracts, the blood flows through the tricuspid valve (which separates the chambers) into the right ventricle.
3. From the right ventricle, the blood passes the pulmonary semilunar valve and enters the right and left pulmonary arteries. These arteries take the blood into the lungs for oxygenation. (These are the only arteries in the body that do not carry oxygenated blood.)
4. Once the blood has passed through the lungs, where it exchanges oxygen for carbon dioxide, the blood flows back onto the left atria through the right and left pulmonary veins. (These veins are the only veins in the body that carry freshly oxygenated blood.)
5. From the left atria, the blood passes through the bicuspid valve and enters the left ventricle. The left ventricle is the part of the heart where the muscular walls are the thickest, as the left ventricle is responsible for pumping the blood out of the heart into the arterial system to circulate through the body.
6. The blood passes through the aortic semilunar valve, finally exiting the heart through a very large artery, known as the aorta.

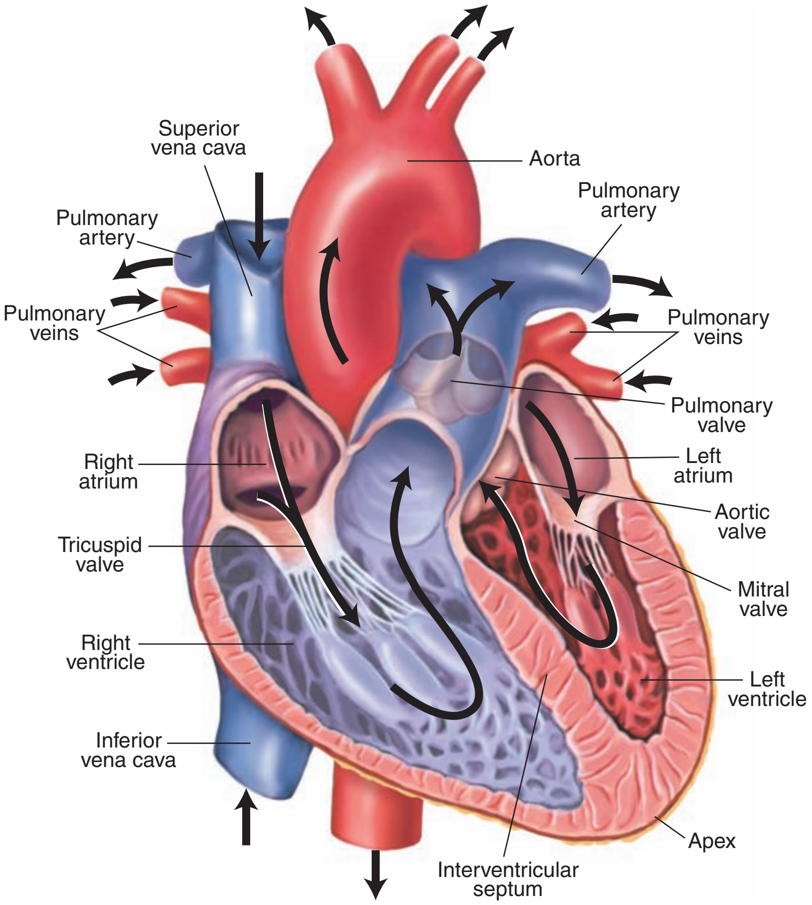
Superior vena cava
Aorta
Pulmonary artery
Pulmonary artery
Pulmonary veins
Pulmonary veins
Pulmonary valve
Left atrium
Right atrium
Aortic valve
Tricuspid valve
Mitral valve
Right ventricle
Left ventricle
Inferior vena cava
Apex
Interventricular septum

Figure 8-1 The pathway followed by the blood as it circulates through the heart, lungs, and vessels. *Reprinted with permission from Eagle S, Brassington C, Dailey C, and Goretti C:* The Professional Medical Assistant: An Integrative, Teamwork-Based Approach. *Philadelphia: FA Davis, 2009.*

7. In a single heartbeat, the heart goes through two stages: *systole* (the contraction of the heart muscle) and *diastole* (the relaxation of the heart muscle). During ventricular systole, the blood pressure increases, pushing the blood through the arterial system and adding pressure to the walls of the arterial vessels, causing them to expand. The diastolic (relaxation) phase allows the blood pressure to decrease and the arterial walls to adjust back to their normal size. We feel this increase and decrease in the blood pressure as a patient's pulse in the arteries of the body.

Blood Vessels

There are three types of blood vessels that make up the cardiovascular system: arteries, veins, and capillaries. Each type of vessel has unique characteristics that enable it to perform specific tasks. Blood specimens may be collected from any of these vessels. Figure 8-2 shows the relationship between the types of blood vessels that make up the cardiovascular system.

Test Your Knowledge 8-1

What organs and/or body systems make up the cardiovascular system? (Outcome 8-1)

Arteries

Arteries are the vessels that carry oxygenated blood away from the heart. Arterial walls are quite muscular and have more elasticity than do venous walls, and as the blood flows through the arteries (expanding and relaxing the vessels) it is felt as a pulse. There are three layers that make up the walls of both arteries and veins, but the walls of the arteries are thicker and stronger than those of the veins, and the arterial walls have more flexibility. Arteries get smaller and smaller as they branch through the body away from the heart. Smaller arterial vessels branch into **arterioles,** which then lead into capillaries. The blood that travels through arteries is brighter red than that which travels through the veins or capillaries due to the high level of oxygen present in the arterial red blood cells.

Veins

Veins carry blood back to the heart from the systemic circulation. Because this is not a process that actually follows the laws of gravity, the venous vessels have one-way valves that help to keep the blood from pooling in the extremities. The walls of the veins are made up of

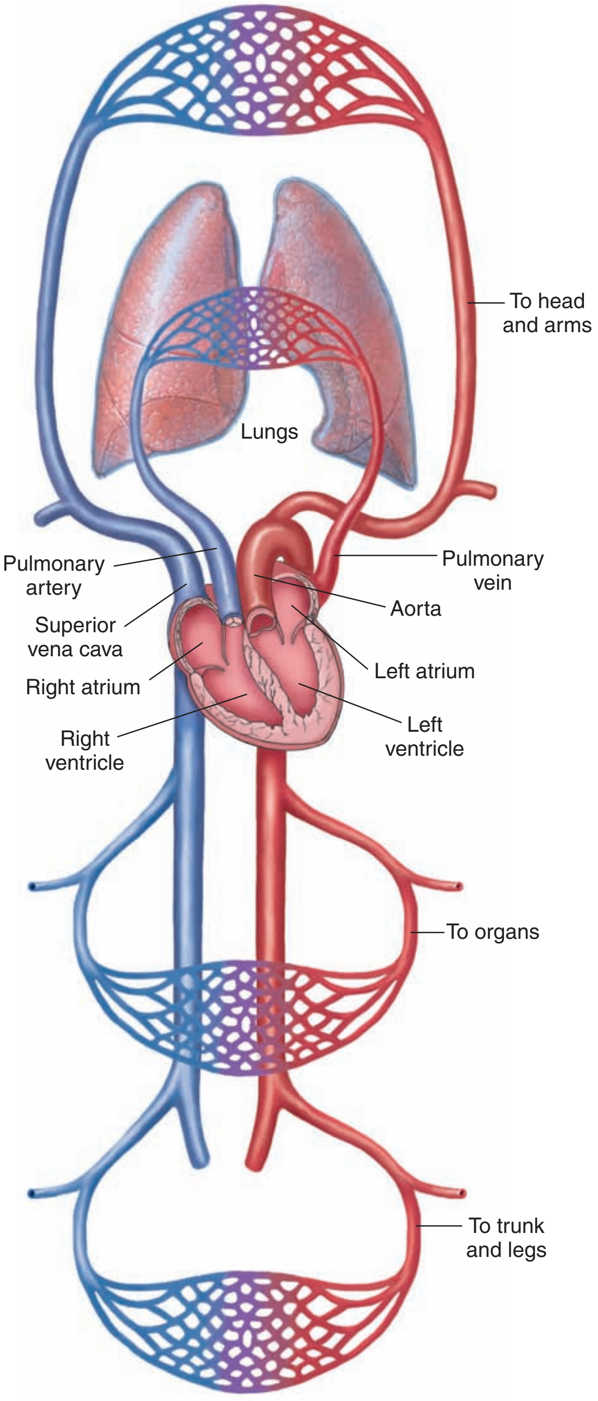

Figure 8-2 The relationships between the arteries, veins, and capillaries. *Reprinted with permission from Eagle S, Brassington C, Dailey C, and Goretti C:* The Professional Medical Assistant: An Integrative, Teamwork-Based Approach. *Philadelphia: FA Davis, 2009.*

three layers, but the layers are thinner and less muscular than are those of the arterial system. Veins require skeletal muscle movement to propel the blood back to the heart. Veins are smallest where they are attached to the capillaries; these tiny vessels are called **venules.** The veins become larger and larger as they approach the heart. The blood that travels through the veins has less oxygen than that in the arterial or capillary system, and is darker red in color. Veins do not have a pulse, and when palpated, veins feel less rigid than arteries.

Capillaries

Capillaries are so small that they can be seen only with a microscope. The walls of capillary vessels are made up of a single layer of epithelial cells. They are designed to link the arterial vessels to the venous vessels, and the blood that flows through the capillaries is a mixture of venous and arterial blood. As blood travels through the arteries to the arterioles, there is less and less room for the entire volume of blood to make it through the vessels. As the vessels get smaller, more fluids are pushed out of the vessel and into the **interstitial** spaces between the cells surrounding the blood vessels. By the time that the blood makes it to the capillaries, the size of the blood vessel has often reduced to the point that there is room for only one red blood cell at a time to make it through. The capillary beds are the only place in the body where exchange of nutrients, gases, and other molecules may occur, because the liquid portion of the blood is in direct contact with the cells around the vessels. As the capillaries join back with the venous system, they begin to get larger in diameter, and the vessels become venules, which gradually increase in diameter to become veins. Due to their small size, capillaries do not cause a pulse that can be felt.

Test Your Knowledge 8-2

What is one way that veins are different from arteries? (Outcome 8-3)

Test Your Knowledge 8-3

Are capillaries approximately the same size as arteries and veins? (Outcome 8-4)

SITE SELECTION

Appropriate site selection for venipuncture procedures is essential to obtain a blood sample that is acceptable for testing, and to avoid negative outcomes for the patient. The most common area for venipunctures to be performed is the **antecubital area** of the arm, which is the area immediately in front of (the anterior side of) the elbow.

Most Commonly Used Veins

There are several veins in the antecubital area that may be successfully used for withdrawing blood. The **median cubital vein** (Fig. 8-3) is best for venipuncture because it is usually large and well anchored. This means that the median cubital vein doesn't usually move when the needle is inserted. The **cephalic vein,** which is located on the thumb (lateral side) of the antecubital area, and the **basilic vein,** on the little finger (medial side) of the antecubital area, may also be acceptable for venipuncture. If it is necessary to consider these two alternative venipuncture sites, the median cubital vein is the first choice; if this vein cannot be felt, the second choice would be the cephalic vein. The basilic vein is the last choice in the antecubital area because of its close proximity to the brachial artery and median nerve. The basilic vein is also not anchored as well as the other antecubital veins, so it has more of a tendency to move, or roll away, during the puncture.

When performing phlebotomy, finding a vein is one step that requires practice and experience. Applying a tourniquet may allow the vein to be more prominent. **Palpation,** which uses touch to feel for a vein, may be used with or without the tourniquet application to determine whether a vein is suitable for a blood draw. Sight may be helpful initially, but if a vein cannot be felt, that vein should not be used for a blood draw. Palpation is performed by using the tips of the gloved fingers (never the thumb) in a pushing manner as the fingers are moved across the arm. An acceptable vein should have some "bounce," like a soft rubber tube, when palpated. A tendon will feel flatter and harder than a vein, without

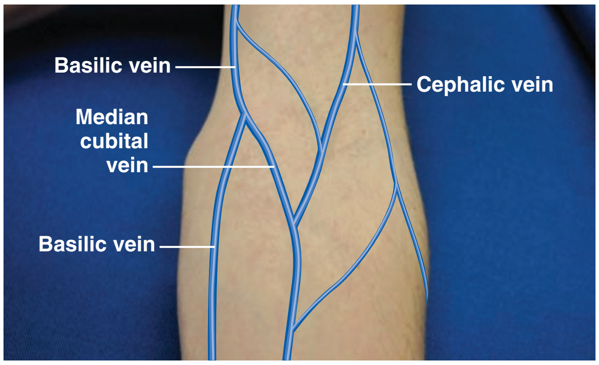

Figure 8-3 Antecubital area with vein selection; identification of the median cubital, cephalic, and basilic veins.

any elasticity, and an artery will feel like a taut rubber band. Veins may be differentiated from arteries because of the pulse that is felt in an arterial vessel. While feeling for the vein, it is important to establish the depth and size of the vessel, as well the direction of the vein. When the needle is inserted into the arm, it must follow the direction of the vein, and be inserted to the correct depth at which it will be in the vein rather than resting above or piercing through the vein. The angle of insertion is also very important to allow for the blood to enter the vein when it is pierced.

Venous patterns vary from person to person, as the anatomy of each patient is unique. In addition, some patients have vessels that are much more prominent in one arm than in the other. A venipuncture should not be attempted if the health-care professional is not sure whether the vein is suitable for the procedure or until the other arm has been checked for potential sites. If there is no suitable vein identified in the antecubital area of either arm, the back of the hand may also be considered for obtaining a sample. It is especially important to anchor the blood vessels of the hand before inserting the needle, as these vessels are located close to the surface of the skin and often roll away when the needle is inserted. The hand veins are also more prone to bruising and hematoma formation. Venipunctures should not be performed on the inside (or underside) of the wrist because of the presence of various nerves, tendons and ligaments, the close proximity of the radial artery, and the small size of the veins present in this area.

If an acceptable vein cannot be located, it may be necessary to encourage the vessels to become more prominent before venipuncture is attempted. To encourage **vasodilation,** a warm compress may be applied to the area, or the desired site may be massaged. Slapping, thumping, or flicking the draw site is not recommended. Lowering the arm down below the level of the heart may also allow the veins to become more evident. Tourniquet application is definitely beneficial, but it is important to remember that a tourniquet may stay on for a maximum of only 1 minute before it should be removed. A blood pressure cuff (stand-alone type) set at 40 mm Hg of pressure may also be useful in coaxing the veins to become more prominent.

Test Your Knowledge 8-4

What area of the arm should be the first choice for venipuncture? (Outcome 8-5)

Appropriate Capillary Puncture Sites

Capillary blood draws (also known as dermal or skin punctures) may be used for blood collection when the required volume of blood needed for testing is small. The capillary beds in the fingers and heel are the sites generally used in the laboratory. The fingers are often used as sites for capillary punctures for adults and children over the age of 1 year. The puncture should be made on the fleshy central surface on the palmar side of the finger. Capillary draws should never be performed right on the tip of the finger, or on the extreme lateral side(s) near the fingernail, because of the close proximity to the bone. The capillary device should be oriented so that the puncture into the skin is perpendicular to the fingerprints; doing so means that the blood won't pool into the lines of the fingerprint during the collection process. Figure 8-4 shows where it is safe to perform a capillary puncture on a fingertip.

For infants (children below the age of 1 year) capillary blood draws are performed on the heel of the foot. The fingers on infants should not be used for capillary punctures, as there is not enough tissue to make this a safe choice while avoiding potential damage to the bone. The central area of the infant's heel, the arch of the foot, the ball of the foot, and the toes should *never* be used for blood draws. The capillary draw should be done only on the most medial and lateral section of the bottom of the heel. Figure 8-5 shows the safe areas for capillary punctures on an infant's heel. The bone of the foot may be damaged if the procedure is performed incorrectly and *osteomyelitis* (inflammation and/or infection of the bone and bone marrow) may result.

In all capillary draw situations, it is imperative that the site is warmed thoroughly before the process begins. It is also helpful if the hand or foot is held below the level of the heart. Massage may also benefit the process, as it

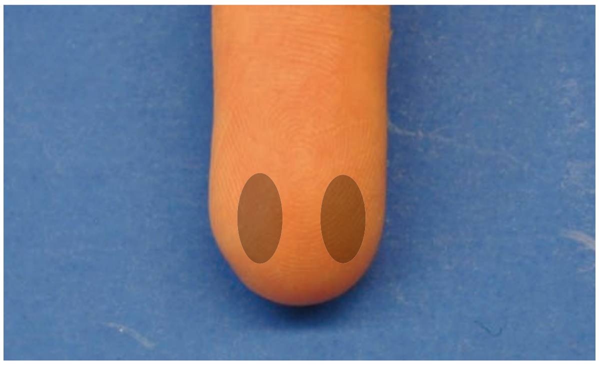

Figure 8-4 Locations for capillary draws for fingers.

POINT OF INTEREST 8-1
Newborn screening

Each state provides newborn screening to detect disorders that would cause irreversible damage or death if undetected for an extended period of time. These disorders (for the most part) are not visible at birth; the babies appear to be healthy until the damage has already occurred. Irreversible neurological or musculoskeletal damage will result if these diseases are not detected. Most states require that the babies be tested prior to discharge from the hospital, with a recommendation for a second screening 7 to 14 days later. Examples of disorders that may be detected with screening include the following:

- Phenylketonuria
- Maple syrup urine disease
- Carnitine uptake deficiency
- Congenital hypothyroidism
- Cystic fibrosis
- Various hemoglobinopathies
- Galactosemia

The specimen collection usually includes a capillary sample taken from a heel stick. Blood may also be drawn from the veins on the back of the hand and applied to the collection paper. Special filter paper is used for the collection process, and the sample must be handled appropriately to avoid erroneous results or rejection of the specimen.

Test Your Knowledge 8-5

Are capillary draws always performed on the fingers?
(Outcome 8-6)

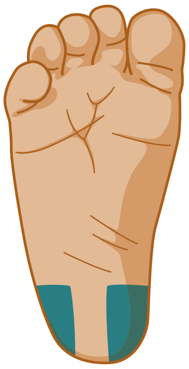

Figure 8-5 Locations for capillary draws in the heel of an infant.

will help to increase the blood supply to the area. When a capillary specimen is submitted for testing, it is important to document that the specimen came from a capillary draw rather than a venipuncture because the reference ranges for capillary blood samples are not the same as those for venous blood.

Blood used for capillary samples should be free flowing. It is important that the area immediately adjacent to the puncture site is not squeezed to increase the blood flow, as this will contaminate the sample with interstitial fluid. This fluid from between the cells has a different concentration of many chemical substances and will dilute the cell counts for hematology specimens. Massaging the finger well above the puncture site will help increase the blood flow. Specimens drawn from the heel may benefit from gentle pressure applied to the calf and ankle area.

CONTRAINDICATIONS AND AREAS TO AVOID

When considering an appropriate site for venipuncture, it is the responsibility of the health-care professional to use best judgment to avoid harming the patient unnecessarily. Also, as part of the preanalytical process, it is important to remember that any substances that may contaminate the specimen during the blood draw must be avoided. To accomplish these goals, there are certain **contraindications** and precautions for venipuncture to keep in mind.

Scars and Tattoos

Excessively scarred skin (as in the case of a burn) should not be used for venipuncture. Scarred areas are more susceptible to infection and often have reduced circulation; further, the veins underneath the scarred skin may be difficult to palpate and anchor. Skin that has been tattooed is also more prone to infection, and the dyes used in coloring the tattoo may cause test interference if they make their way into the specimen.

Mastectomy Considerations

If a patient has had a **mastectomy,** venipuncture should not be performed on the arm on the same side of the body as the procedure. Mastectomy procedures involve the removal of breast tissue. In addition, the lymph nodes surrounding the breast and under the arm also may be removed. The absence of these nodes can cause a decrease in the flow of the lymph fluid in the area, which puts the patient at an increased risk of lymphedema (accumulation of lymph in the tissues) and infection, even if the blood draw is not performed and only a tourniquet is applied. The opposite arm should be used for the blood draw. If a patient has a double mastectomy, the health-care professional who ordered the blood draw should be consulted to see what alternative sites may be used for blood collection.

Intravenous Fluids

Intravenous fluids are a potential source of contamination when drawing blood. In situations in which the patient is receiving IV fluids, it is preferable to use the opposite arm for venipuncture. If it is not possible to draw from the other arm, the IV flow must be discontinued (turned off) for at least 2 minutes (depending on the substance being infused) by a physician or other qualified health-care provider before the blood draw is attempted. After the IV fluid drip has been turned off, the tourniquet may be applied *below* the IV insertion site, and the blood draw may be drawn from below the IV site. Even with these precautions, the first 5 to 10 mL of blood withdrawn must be discarded to avoid potential contamination with IV fluid. If a coagulation test is part of the blood draw order, an additional 5 mL of blood should be drawn and discarded before this tube used for the coagulation testing is collected.

Veins that have been used for IV therapy are often irritated and damaged, even after the IV has been discontinued. Ideally, these sites should be avoided for venipunctures for at least 1 or 2 days after the catheter has been removed to allow for the vein to recover. *Phlebitis* (inflammation of the vein) may result if the area is not allowed to heal before venipuncture procedures are performed.

Test Your Knowledge 8-6

Is IV therapy always a contraindication for venipuncture? Why or why not? (Outcome 8-7)

Bruising and Hematomas

Bruises and **hematomas** may occur as complications of venipuncture or IV therapy, and these sites should be avoided for subsequent blood draws. A bruise and a hematoma are similar in that each is a visible sign of damage to a vessel in which blood has escaped and entered the surrounding tissues. Hematomas result from damage to larger vessels such as veins and arteries, whereas a bruise may be the result of injury to a capillary or a small blood vessel. A hematoma is often hard and swollen, and may be quite painful for the patient. If a blood draw is attempted in a bruised area or a hematoma site, it may cause additional discomfort for the patient and the integrity of the sample may be affected because of the pooling of old, dead cells in the area. Drawing from this area may also include the risk of infection due to inflammation of the tissues. If it is absolutely necessary to draw from an area with excessive bruising or a hematoma, the venipuncture should be performed below the injured site.

Varicose Veins

A *varicose vein* is one that is twisted and enlarged and very visible near the surface of the skin. These veins have lost their natural elasticity and integrity, and are prone to inflammation and blood clot formation. Varicose veins should not be used for venipuncture, as the blood flow through these areas has been adversely affected, increasing the chance for infection and bruising. These veins are also not very stable and will often roll away during the venipuncture.

Central Lines and Fistulas

Intravenous lines are used commonly to administer medication, fluids, or blood products, especially when there is a need for continued access for an extended period of time. There are times when a phlebotomist or medical assistant may be asked to withdraw blood from a *central venous catheter* or a *peripherally inserted central catheter* (PICC line). The central venous catheters are inserted into one of the major veins of the body, such as the subclavian, superior vena cava, or jugular vein. There is a piece of tubing extending out of the vein through the skin that is covered with a dressing. The central venous catheter is often left in place for an extended period of time to be used for administration of medication directly into the bloodstream.

A PICC is inserted into the cephalic or basilic vein of the arm or hand. These peripherally inserted catheters are fragile, as they are not inserted into a large vessel. Blood withdrawal from either of these devices generally requires injection of saline and/or a heparin solution into the catheter to clear the line before and after the blood is taken. It is essential to remember that anyone who withdraws blood from these types of lines must have specialized training, and in some states it is unlawful for a nonlicensed person to perform this function. In states where this is allowed, each facility must establish its own policy and training procedures.

Medical assistants and phlebotomists in most states are not allowed to perform this type of draw.

Another type of venous access device that may be in use is a *cannula*. This is a flexible tube that is used to gain access to the veins for *dialysis* treatment for those who have advanced kidney disease. The cannula is placed under the skin, and has a special port that is designed for insertion of a needle to withdraw blood for dialysis or for testing. A *fistula* is similar to a cannula, except that the device has been implanted in a way that it joins the arterial and venous system for hemodialysis procedures. The fistula is also under the surface of the skin, so it requires the skin to be punctured for access. It is recommended to use the other arm, rather than one containing the fistula. Blood draws should never be attempted from these implanted devices unless the practitioner has received specialized training. In many states, only licensed health-care providers may perform this procedure. In addition, a tourniquet should not be applied to the arm that contains a fistula or cannula.

POINT OF INTEREST 8-2
Arterial blood draws

Respiratory emergencies (such as advanced chronic obstructive pulmonary disease [COPD] or pulmonary edema) may require an arterial blood gas analysis for appropriate treatment to be started. In these situations the oxygen and carbon dioxide levels may be far outside the reference ranges, and the arterial blood pH may be increased or decreased to a life-threatening level. Arterial blood draws are similar to venipuncture procedures in that the site must be stabilized and cleansed properly before the needle is inserted. The angle of insertion, site selection options, and the way the blood enters the syringe are different from those of a venipuncture.

Arterial blood gases are commonly drawn from the radial artery; as a second choice, the brachial artery may be used as long as the patient is not being treated with anticoagulant drugs. The radial artery is easy to identify and secure because it is so near the surface of the body. To choose a site for an arterial blood draw, the health-care professional must palpate until the pulse is felt strongly. The site needs to be cleansed thoroughly, and the specimen needs to be drawn into a special heparinized syringe designed for this type of collection. A cup of wet ice also needs to be on hand, as the sample needs to be placed on ice immediately after the draw is complete.

The health-care provider should have his or her index finger on the pulse of the artery. The needle is inserted at an angle of approximately 45 degrees. The skin is pierced right next to the index finger. When the artery is entered, the blood will pump out through the lumen of the needle because of the pulse pressure. A flash is apparent immediately at the end of the needle attached to the syringe as the artery is pierced, and the blood will continue to fill the syringe without pulling on the plunger. (The plunger is pulled back before the needle is inserted to allow room for the blood in the syringe.) When the syringe is full, the safety device must be activated immediately on the needle. The needle will be removed and the end of the syringe will be plugged with a rubber stopper. The sample must be placed on ice immediately after labeling of the specimen. Firm pressure must be applied to the puncture site for a minimum of 5 minutes after the blood draw, or until the blood stops flowing from the site.

Due to the potential for arterial damage or nerve damage with arterial blood draws, intensive training is required for health-care professionals who perform this as part of their job. Training includes careful observation and supervised performance, as well as identification of risk factors associated with the process. Some states restrict this procedure to licensed personnel such as respiratory therapists or RNs.

Edematous Areas

In some clinical situations, patients may accumulate fluid in the intercellular spaces of the tissues. This accumulation of fluid is called **edema.** The presence of excess fluid may be widespread as a result of heart and/or renal failure, or it can be more localized from trauma or fluid leakage from an IV site. It will be very difficult to palpate a vein in edematous areas, because of the additional fluid in the tissues. If possible, these areas of the body should be avoided for venipuncture, because of the increased opportunity for infection and potential for contamination of the specimen with the additional fluids. If the edema is widespread, the medical assistant may need to apply gentle pressure to the site until the edema is displaced, then perform the venipuncture as quickly as possible. If this is not successful and alternative sites are not obvious, the phlebotomist or medical assistant who is drawing the specimen should notify the physician. Capillary collection procedures are often not an option when edema is present because of the excessive amount of fluid contamination.

Test Your Knowledge 8-7

Why is it important to avoid edematous areas when performing venipunctures? (Outcome 8-7)

Obese Patients

Excessive body fat can make venipuncture very challenging. It may be very difficult to palpate veins in obese patients, and sometimes it is even necessary to use longer than normal needles to successfully reach the veins under the surface of the skin. Obese patients may be quite apprehensive about the blood draw process because of past experiences with unsuccessful venipunctures. The anxious attitude may make it even more difficult to carry out a successful blood draw. If the patient suggests a site where others have had success, it is best to attempt this site if there are no other prominent veins. Adipose tissue may also be displaced in some situations by applying gentle pressure with the fingers to "move" the tissue out of the way until the draw site has been established.

VENIPUNCTURE EQUIPMENT

There are various methods used to obtain blood from a vein. Each type of blood draw requires unique supplies and procedures. Blood may be withdrawn from a vein using the evacuated tube system, the syringe system, or the butterfly method. Regardless of the method chosen, tourniquets, gloves, sharps containers, alcohol and gauze pads, needles, and adhesive materials to be applied after the blood draw are necessary for the process.

Tourniquets

A **tourniquet** is a flexible band that is placed around the arm of the patient to make the veins more prominent. Most tourniquets found on the market today are stretchy and made of a nonlatex material. They are approximately 1 in. wide, and approximately 15 in. in length. Tourniquets are designed to apply pressure, which causes distention of the veins. The tourniquet must be placed approximately 3 in. above the site to be used for venipuncture. If it is placed closer, the ends of the tourniquet may interfere with the procedure. Remember, tourniquets are only a tool to help the vein to become more visible; there are times when it is not necessary to use a tourniquet, and some laboratory tests prohibit the use of a tourniquet for the specimen collection.

Tourniquets should not be applied too tightly; this can cause pain for the patient, and may cause **hemoconcentration** in the veins because of the restricted blood flow. Hemoconcentration occurs when excessive pressure builds up in the veins below the tourniquet application site. The increased pressure in the veins will force water and small molecules out of the vessels into the surrounding tissues. Larger formed elements and molecules will remain in the venous blood, and become more concentrated. White blood cells, red blood cells, platelets, calcium, cholesterol, triglycerides, and bilirubin levels may all be affected by this venous concentration. Tourniquets that are left on for more than a minute may also cause hemoconcentration; therefore, the phlebotomist needs to remain aware of this as the venipuncture site is prepared. Extended tourniquet application time may also cause pain for the patient.

When applying a tourniquet, keep in mind that it should be applied in a way that it can easily be released with one hand. Tourniquets do not need to be applied to bare skin; they can be applied over the sleeve. Open wounds, scars, and bruises should be avoided when applying the tourniquet. Some tourniquets are disposable, and these should be discarded after each use. Others are designed to be used for multiple patients, and should be disinfected at least daily, and discarded if they are visibly soiled or lose elasticity.

Gloves

The Occupational Safety and Health Administration (OSHA) regulations designed to protect health-care employees specify that gloves are to be worn when collecting blood. After a blood draw, the gloves are removed, and the employee must sanitize his or her hands. For effective palpation and dexterity during the collection process, gloves should fit snugly. Except in very specific circumstances, it is not necessary for the gloves used for phlebotomy to be sterile. Most health-care settings no longer use latex gloves, due to the increase in latex sensitivity issues. A common substitute for latex gloves, nitrile gloves work well for phlebotomy. Vinyl gloves may also be worn, but they are not advised in all situations because of their tendency to break down when exposed to certain chemicals.

Sharps Biohazardous Waste Containers

A **sharp** is anything that has been exposed to blood (or other potentially infectious materials) that has corners, edges, or sharp projections that are capable of cutting or piercing the skin. In venipuncture, sharps include needles and capillary puncture devices. Glass slides may also be classified as sharps because they have sharp corners that may be hazardous. Sharps must be disposed of in a waste container that is specifically designed to be puncture resistant, leakproof on the sides and bottom, and labeled as biohazardous waste. These containers should be placed close to the site where the venipuncture is performed so that contaminated sharps may be disposed of immediately. Sharps waste containers come in various sizes, and should never be overfilled. If a sharps container is filled past the marked fill line, it is possible for injury to occur when an employee adds contaminated sharps to the container.

Alcohol and Gauze Pads

Seventy percent isopropyl alcohol is generally used for cleaning the venipuncture site. It may be applied by using a presoaked commercial square that is individually wrapped and packaged. Alcohol may also be applied by using a gauze pad that has been saturated just prior to use. Isopropyl alcohol is an effective antiseptic that can help to eliminate most transient and surface resident bacteria from the venipuncture site. The alcohol used to cleanse the site must be allowed to dry completely before the procedure is performed. If liquid alcohol is still present when the needle is inserted, it can cause a stinging sensation and irritation to the skin and/or the vein. Alcohol may also cause hemolysis of the red blood cells if it is introduced into the specimen.

When additional cleanliness is essential during the collection process (such as when blood cultures or arterial punctures are ordered), iodine or chlorhexidine gluconate may be used. Chlorhexidine gluconate is ideal for those who are allergic to iodine. Procedures used for drawing blood cultures are presented in more detail in Chapter 10.

Clean 2 x 2 gauze squares are recommended for venipuncture procedures. The gauze is applied after the needle has been removed from the arm and the collection is completed. Cotton balls should not be used because their excess fibers tend to stick to the wound and can cause bleeding to recur when the cotton ball is removed later.

Test Your Knowledge! 8-8

List three items that are needed for all venipuncture procedures, regardless of which technique used.

(Outcome 8-8)

Needles

Needles used to withdraw blood are always disposable and used only once. They are generally 1 to 1.5 in. in length. Figure 8-6 describes the different parts of a needle. The slanted part of the needle that creates the sharp tip is called the **bevel.** This slant helps the needle to enter the skin with a minimum amount of trauma and pain. There is a hole in the bevel where the blood enters the interior hollow portion of the needle, called the **lumen.** The **gauge** of the needle refers to the diameter of the lumen: the larger the needle, the smaller the number used to describe the gauge. Needles used for venipuncture are generally 20 to 23 gauge. Needles that are smaller than 23 gauge will cause **hemolysis** (breaking of the red blood cells with release of hemoglobin) as the blood cells are broken when they enter the small hollow cylinder that forms the lumen. Conversely, if the needle used is too large, hemolysis may also result because of the force at which the blood enters the tube or syringe through the large opening.

When removing the cap from a needle to prepare for venipuncture, it is important to inspect the needle for any defects such as burrs or a blunt tip. If there are problems present, the needle must be discarded immediately into a sharps container and a new one used for the procedure.

Needles used for the **evacuated tube system** are also known as **multi-sample needles.** They have sharp tips at both ends; the end with the beveled tip is designed to be inserted into the vein; the shorter, blunt tip at the opposite end is covered with a rubberized sleeve. The needle is designed to be screwed into a hard plastic holder that holds the tube and the needle. The needle

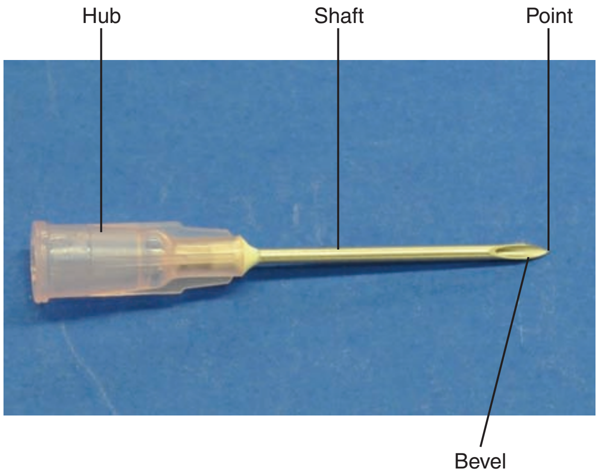

Figure 8-6 Identification of the different parts of a needle.

or the holder must be equipped with a safety device to protect the employee upon completion of the blood draw. A multi-sample needle and holder are shown in Figure 8-7.

When performing venipuncture utilizing the **syringe system,** a needle of similar length and gauge as that for the evacuated tube method is used (Fig. 8-8). These needles are designed to screw directly onto the syringe, so the needle only has one pointed end; the other end has a hub that attaches directly to the syringe. The needles or the syringe must be equipped with a safety device to protect the employee after the blood draw is completed. Examples of safety mechanisms include a cover that is pushed over the needle and locked in place after the draw, or a retractable needle that moves back into the syringe once the specimen is obtained.

The **butterfly system,** or winged infusion system, consists of a short needle with “wings” (Fig. 8-9). When used for venipuncture, the needle is generally 23 gauge, and is attached to a piece of plastic tubing that is approximately 12 in. long. The wings are held by the fingers while the needle is inserted into the vein. Butterfly needles are used for fragile veins (such as those in the hand) or pediatric draws. The butterfly tubing may be attached directly to a syringe, or an adapter may be used to attach the butterfly setup to a plastic evacuated tube holder. Proper activation of the safety devices included with the butterfly system is critical, as there are more accidental needlesticks attributed to the use of butterflies than there are in either of the other two systems.

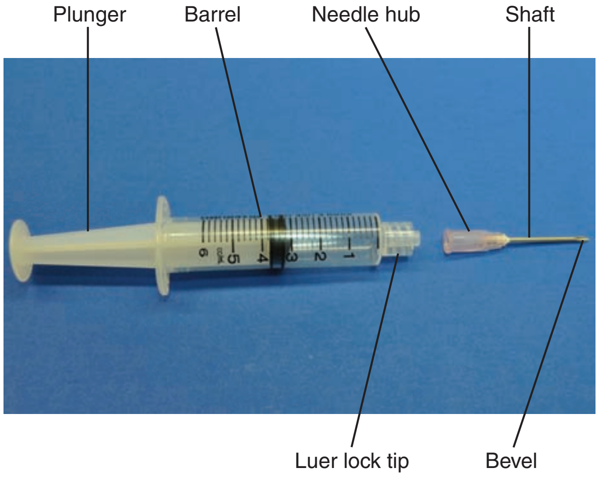

Figure 8-8 Syringe with identification of the various parts.

Test Your Knowledge 8-9

Are the needles used for all venipuncture techniques the same? (Outcome 8-8)

Safety needles are required for all methods of venipuncture. The various safety devices are designed to be easily activated, using either a needle that withdraws into the device while still in the skin, or one hand activation of the safety device. Figure 8-10 shows various safety devices for evacuated tube systems, syringe needles, and butterfly needles. It is essential that the medical assistant performing venipuncture become comfortable

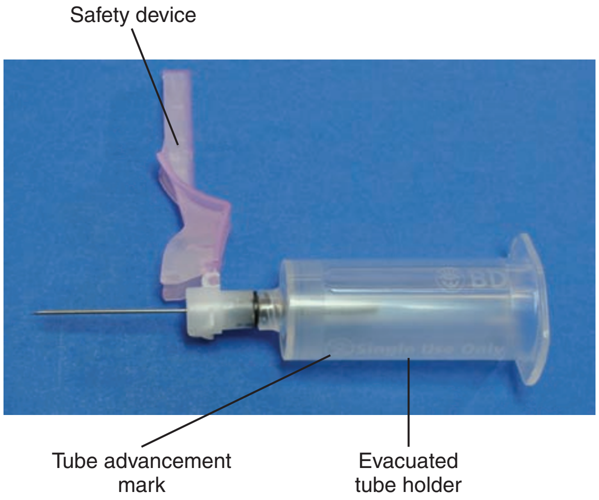

Figure 8-7 Multi-sample needle and holder with identification of parts.

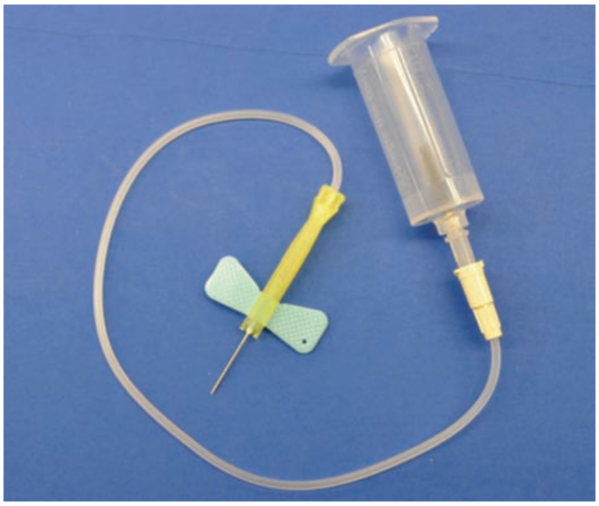

Figure 8-9 Butterfly assembly.

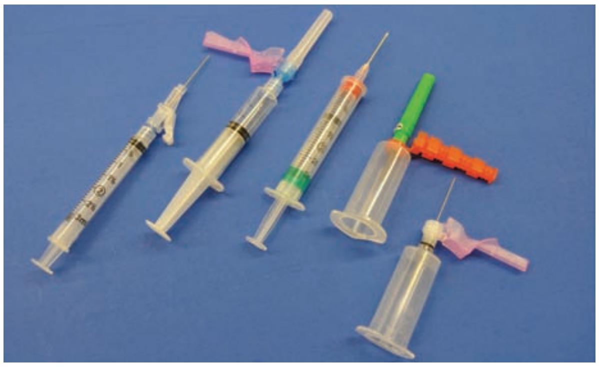

Figure 8-10 Safety devices for use with the evacuated tube system and syringe draws.

with the safety devices available to him or her so that the device can be activated quickly and safely for protection.

Additional Supplies

Blood spill cleanup supplies should be available near every phlebotomy station, and staff should be trained regularly on their use. In addition, because every tube should be identified while in the presence of the patient, it is important to have at least one functional pen available to identify the tube and document the blood draw information on the requisition. It is also necessary to have backup tubes and needles within reach for use when the procedure does not go as planned.

METHODS USED FOR VENIPUNCTURE

Evacuated Tube System

The most common and most efficient method of performing venipuncture is the use of the evacuated tube system. As described earlier in the chapter, blood is collected directly into an evacuated tube through a needle that has two pointed ends. This method is advantageous because there is no need to transfer the blood from one container to another after the draw, and also because multiple tubes of blood may easily and quickly be drawn with one skin puncture. The needle to be used with the evacuated tube system is screwed into an evacuated tube holder that is made of semirigid plastic. These tube holders are designed to be used only once; if the medical assistant tries to use them more than once, the needle may not screw in securely. The rubber cover covering the shorter, posterior end of the needle is pushed back to expose the sharp needle as the evacuated tube is pushed on during the blood draw. Evacuated tubes have the air removed to create a vacuum, so as soon as the tube cap is pierced with the needle, the blood is pulled into the tube. When the blood draw is complete, the safety device for the needle must be activated, and the evacuated tube holder with the needle attached must be disposed of directly into a biohazardous sharps container.

The plastic evacuated tube holder has flared sections at the bottom, which are known as **flanges.** These are designed to be pushed against as the medical assistant changes tubes during the blood draw process. The flanges help to keep the venipuncture setup stable during the process of drawing blood. There is also a line marked on the tube holder; this is a "safe" mark that shows how far the tubes can be advanced into the holder before the vacuum will be affected. Some health-care personnel prefer to advance the tube to this mark before they begin the blood draw process, whereas others do not. If the tube is advanced past this mark, the vacuum will be affected and the tube may not fill as desired once the needle is in the vein.

Syringe Method

Health-care personnel may prefer to use the syringe method for venipuncture in patients having small or fragile veins. Use of a syringe to withdraw blood allows for more control of the vacuum used to withdraw blood from the vein, as the individual performing the venipuncture can control the speed and force used to pull on the plunger of the syringe. Syringe draws also allow the person performing the venipuncture to see the blood in the hub of the needle as the vein is entered, which can be reassuring when attempting a blood draw on a vein that is difficult to palpate or locate. This return of blood into the hub of the needle is called a "flash" of blood.

Syringes are made of a hollow plastic barrel with graduated measurements in cubic centimeters (cc) or milliliters (mL) printed on the side. Inside the barrel of the syringe is a plunger that is designed to fit tightly enough to create suction when the plunger is pulled backward. The sizes of syringes used for venipuncture may vary from 3 mL to 20 mL, with 5-mL or 10-mL syringes most commonly used. Any size needle may be attached to the syringe, but 20- to 23-gauge, 1-in. needles are most commonly utilized for venipuncture. Safety devices for the blood draw may be attached to the needle, or the syringe may be safety enabled.

One disadvantage of using syringes for venipuncture is that the blood must be transferred into tubes for further processing and testing after the blood draw. This must occur quickly so that the blood does not clot in the syringe. Transfer devices should be used to add the blood from the syringe to the necessary tubes (Fig. 8-11).

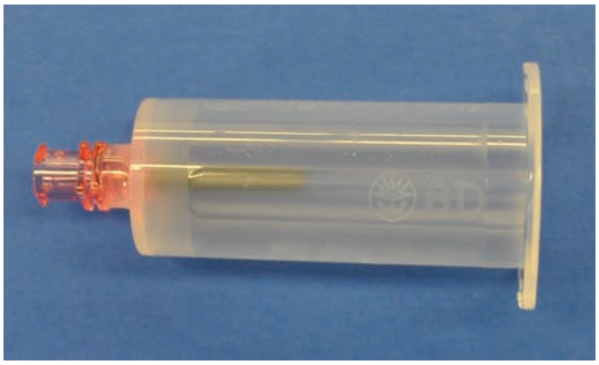

Figure 8-11 Transfer device.

Because the amount of blood obtained with one venipuncture is limited to the size of the syringe used, this must be taken into consideration before choosing the syringe method for blood collection. It may not be advantageous to use a syringe when there is an order that requires more than 10 mL of blood, as the sample may clot in the syringe before it can be successfully transferred to the correct blood tubes.

Winged Infusion or Butterfly Method

Winged infusion sets (also known as **butterfly systems)** are made up of a beveled needle that is attached to a length of tubing with a luer adapter on the opposite end. Just below the needle on the butterfly system are two "wings" that are made of hard plastic. The medical assistant will grasp these wings and use them to guide the insertion of the needle into the vein. Butterfly systems are often used for venipuncture in patients having very fragile veins, such as those in the hands. They are also used for many pediatric blood draws, small fragile veins in areas other than the hand, and sometimes for uncooperative patients. Butterfly sets are preferred in these situations because the needle can be maneuvered easily to gain venous access because of its decreased length. The angle used to enter the vein is smaller when using a butterfly setup. This can be especially important when attempting a blood draw from a superficial small vein, such as those in children. The attached flexible tubing allows for multiple tubes to be collected without danger of moving the needle in the vein. Butterfly systems may be used directly with evacuated tubes (utilizing a Luer adapter for this purpose), or they may be attached to a syringe when there is a desire to use very gentle pressure during the blood draw.

Venipuncture with a butterfly system has a higher risk of an accidental needlestick if the safety device is not activated properly before the medical assistant attempts to process the specimen after the blood draw. Becton, Dickinson has a product with a push-button activation that withdraws the needle into a plastic reservoir before it is removed from the skin. Other products require activation of the safety device after the needle is removed from the patient, and it is imperative that the health-care professional performing the venipuncture follow directions for activation to avoid potential bloodborne pathogen exposure.

POINT OF INTEREST 8-3

Minimizing risks for employees and patients while collecting samples

The collection of blood samples is a very important part of the laboratory testing process. It is important to remember that safety precautions must be followed at all times to protect both the employee and the patient. The appropriate personal protective equipment will offer a layer of protection for the employee, but safety does not stop there. It is very important to follow instructions for the use and activation of all safety devices in order for them to provide protection. It is also important to focus your full attention on the task at hand when performing invasive procedures; it only takes a moment of inattention to provide an opportunity for bloodborne pathogen exposure.

Patient safety is also very important. Invasive procedures can be very traumatic for some patients because of prior experiences or fear of the unknown. Use of phlebotomy chairs with padded support on the arms and across the chest will help to keep the patient comfortable and safe during the process. Careful observation of the patient's response to the invasive procedure will help the employee anticipate fainting or other adverse reactions. All patients should be treated with compassion and care, and the collection process should be completed as quickly and efficiently as possible.

BLOOD COLLECTION TUBES

Evacuated tubes used for blood collection are available in several sizes, and are made of plastic with rubberized stoppers. (Glass collection tubes may still be available, but they are not recommended due to the potential for breakage and blood exposure for health-care personnel.) Evacuated tubes vary by diameter, length, and the additives present in the tubes. They are each designed to hold a specific amount of blood and vary from 2-mL tubes that may be used for pediatric or geriatric blood draws, to 15-mL tubes that are used when a high volume of blood is necessary for a particular assay. The medical assistant or phlebotomist drawing the blood will choose the tube size according to the age of the patient, the integrity of the vein, and the amount of blood needed for the test ordered. The different colors used for the stoppers are coded according to the additive contained within the tube.

These tubes are evacuated because the air within the tube has been removed to create a **vacuum.** This vacuum has been artificially created and calibrated by the manufacturer so that a very specific amount of blood will be drawn into the tube. The tube will never fill completely to the top, as they are not calibrated to hold that volume of blood. As long as the vacuum is still present within the tube, blood will continue to flow. If there has been a loss of vacuum in the evacuated tube, blood will not enter the tube, or it will fill incompletely. This is known as a **short draw** and often will require the specimen to be redrawn. Short draws are sometimes caused by dropping a tube prior to the venipuncture, pushing the tube too far onto the double-ended needle used for evacuated tube draws prior to the blood collection, or by accidentally pulling the needle out of the vein during the venipuncture process.

All blood tubes have an expiration printed on the label. There is an expectation that the vacuum and additives in the tube will function as expected until this date. The health-care worker who performs the blood draw is responsible for checking the expiration date on all tubes before they are used, and at the same time examining the tube for any obvious defects or breaks. Any tubes that are near expiration should be removed from the shelves before the printed date has passed.

Test Your Knowledge 8-10

Why are tubes used for blood collection called vacuum tubes? (Outcome 8-8)

Color Coding of Tubes

Evacuated tubes are all sealed with a colored stopper. These closures are made of rubber and/or plastic and are designed to ensure the vacuum within the tube and provide a barrier to the blood once the tube is filled. Some tubes available on the market have a rubber stopper with an outside plastic cap of the same color. The plastic cap provides protection for the employee when opening the tube, because the additional plastic cap hangs over the outside of the tube and forms a barrier to any aerosol formation from the blood in the tube. An example of this type of tube is the Hemogard tube, manufactured by Becton, Dickinson. See Figure 8-12 for an example of the Hemogard tube.

The stopper colors on evacuated tubes are color coded and used to indicate whether additives are present in the tubes, and if so, what types are present. Additives may function as **anticoagulants,** or **clot activators.** The additives may sometimes serve as preservatives of the integrity of the specimen. Anticoagulants will keep the blood from clotting in the tube so that the cells remain suspended in the plasma. Without the addition of anticoagulants, the blood will begin to clot in the tube immediately, and a visible solid clot will form within an hour. Clot activators, such as glass beads or silicone, are designed to speed up the clotting process so that the blood can be spun in the centrifuge within a shorter period of time after collection. Tubes may also contain a thixotropic separating gel that is designed to form a barrier between the blood cells and the liquid portion of the blood after it is centrifuged. The gel in these **serum separator tubes (SST)** or **plasma separator tubes (PST)** is in the bottom of the tube when the blood is collected, but as the tube is centrifuged, the cells move to the bottom of the tube and the gel travels up in the tube to form a barrier between the cells and the liquid. Figure 8-13 shows how the gel separates the sample after centrifugation.

The colors used for tube stoppers are similar from brand to brand, providing a uniform reference for the different types of additives. The type of anticoagulant present in the tube is printed on the label, along with the expiration date and the blood volume for the

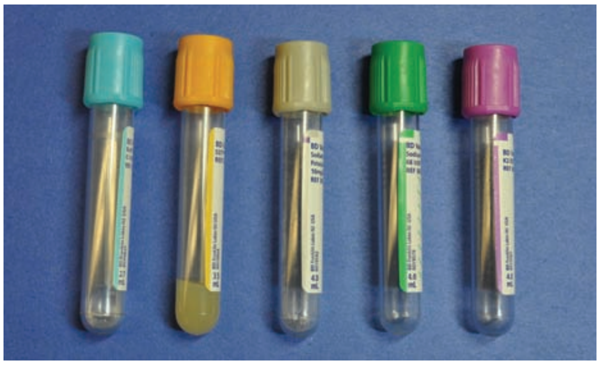

Figure 8-12 Vacuum tubes with Hemogard caps presented in the appropriate order of draw.

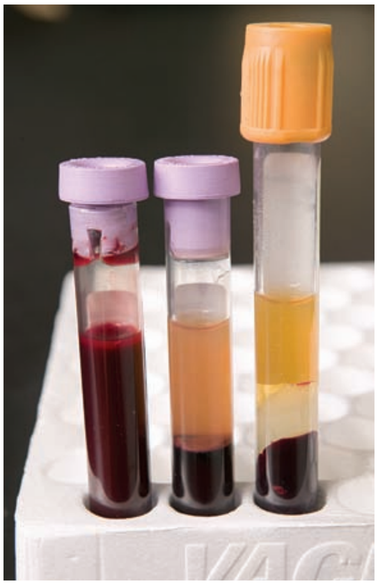

Figure 8-13 The tube on the left shows whole blood before centrifugation. The tube in the middle shows plasma sitting on top of clotted blood cells. The tube on the right is a tube with serum separator gel that has been spun in a centrifuge and shows serum sitting on top of clotted blood cells. *Reprinted with permission from Eagle S, Brassington C, Dailey C, and Goretti C:* The Professional Medical Assistant: An Integrative, Teamwork-Based Approach. *Philadelphia: FA Davis, 2009.*

tube. When using reference materials (such as laboratory requisition forms or laboratory directories) the specimen collection instructions will reference the stopper color needed for the test ordered or the additive present in the tubes. Each laboratory may have different tube requirements for a given test based on its testing methodology, so it is always best to verify the necessary tube before drawing the blood.

Test Your Knowledge 8-11

What is the significance of the different stopper colors used for the vacuum tubes? (Outcome 8-10)

Types of Additives

Types of anticoagulants and/or additives present in the most common evacuated tubes are listed here with their respective color stoppers and summarized in Table 8-1.

- **Yellow:** Sterile tubes used for blood cultures. There are also specific yellow top tubes that contain a red blood cell preservative called acid citrate dextrose (ACD), and these are used for special blood bank procedures. The sterile yellow tubes for blood cultures are by far more common.
- **Light blue:** Contains sodium citrate as an anticoagulant. Light blue top tubes are primarily used for coagulation studies. The ratio of blood to anticoagulant is critical for these tubes and they must always be filled completely to prevent potential dilution and erroneous results.
- **Red:** Glass red top tubes are sterile and contain no additives. Plastic red top tubes contain a clot activator. Red top tubes are used for a variety of tests in which serum is the preferred specimen.
- **Red/gray (tiger top) or gold:** Contains a silica clot activator that increases platelet activity and decreases the time necessary for a clot to form in the tube. Also contains thixotropic gel that separates the cells from the serum when the specimen is centrifuged. Used for chemistry and immunology testing in which serum is the preferred specimen.
- **Black:** Special tubes used only for Westergren sedimentation rates. Contain sodium citrate, but in a different concentration (blood to anticoagulant liquid 4:1) than the light blue top tubes. These tubes are not to be used for coagulation studies.
- **Green:** Contains heparin combined with sodium, lithium, or ammonium as an anticoagulant. These tubes are preferred for chemistry tests in which plasma is suitable for testing and the results must be obtained quickly, as the specimen does not need to clot before the sample may be spun in the centrifuge. Light green Hemogard tubes and those with green and black mottled stoppers contain thixotropic gel in addition to the anticoagulant, which separates the plasma from the cells after centrifugation.
- **Gray:** Contains sodium fluoride, which acts as a glucose preservative (called an antiglycolytic agent) keeping the glucose levels stable in the specimen for a period of time after collection. Also contains an anticoagulant; either potassium oxalate or sodium EDTA. Gray top tubes are not appropriate for general chemistry testing, but are used for glucose testing, as well as blood alcohol determinations.
- **Lavender:** Contains potassium EDTA (dipotassium ethylenediaminetetraacetic acid) as the anticoagulant. Prevents the use of calcium in the clotting process to prevent blood clotting and keep the blood cells suspended in plasma. Lavender top tubes are used for hematology testing because this preservative preserves the integrity and shape of the individual cells better than any of the other anticoagulants.

TABLE 8-1

Table of common tube colors and additives

Tube Stopper Color	Additives
Red top	None
Tiger top/ speckled top/ gold top	Clot activators/thixotropic gel
Green top	Heparin
Mint green top	Heparin and thixotropic gel
Light blue top	Sodium citrate
Lavender top	Potassium EDTA
Gray top	Sodium fluoride and potassium oxalate

ORDER OF DRAW FOR VENIPUNCTURE

The Clinical and Laboratory Standards Institute (formerly known as NCCLS) has recommended a specific order of draw for venipuncture specimens using the evacuated tube system, syringe collection method, or butterfly draws. There is a different recommended order of draw for capillary specimens. These recommendations were developed because of potential problems with additive carryover from one evacuated tube to another. Reports showed significant increases in potassium and some other analytes when a lavender top tube (containing potassium EDTA for an anticoagulant) was drawn just prior to the tube used for chemistry testing. Other potential carryover issues were also identified.

Test Your Knowledge 8-12

Why is the order of draw so important?

(Outcome 8-11)

The recommended order of draw for the six most commonly used tubes:

1. **Blood culture specimens:** These may be yellow top tubes or a set of bottles used for blood cultures. They must be drawn first to avoid potential for contamination of the draw site.
2. **Light blue top citrate tube:** This tube is primarily used for coagulation studies, so it is important that additive crossover is avoided because of potential interactions. Previous recommendations always called for a red top tube to be drawn before the light blue top tube, but with the current CLSI recommendations, this is no longer necessary. An exception is when a butterfly system is used to draw a light blue top tube; in this case, a red top tube is used as a "clear" tube because of the excessive air present in the tubing of the butterfly. The blue top tubes must be filled as far as the vacuum allows, and if the excess air from the tubing is introduced into the light blue top tube, the tube will not fill adequately.
3. **Red top, tiger top, or gold top tubes with or without clot activator:** These tubes are used when serum is the specimen of choice.
4. **Green top tubes:** These are both tubes that have gel separator and tubes that do not. They may be mint green or dark green. These are to be drawn before the EDTA tubes to reduce the risk of crossover that may increase the potassium levels when analyzed.
5. **Lavender top tubes:** These have potassium EDTA as the anticoagulant, and are primarily used for whole blood testing.
6. **Gray top tubes**: Gray top tubes are used for glucose testing, as well as a few other tests. They need to be drawn after the EDTA tubes because the additives present in the gray top tube could cause abnormal changes in the red blood cells to be analyzed if there is crossover. The gray top tube must be drawn after the tubes used for other chemistry testing, because the potassium oxalate present in the gray top tube could cause falsely elevated potassium results if the specimens were contaminated by the anticoagulant and this test was to be performed on the sample.

CAPILLARY PUNCTURES

Most blood collections are performed by puncturing a vein. However, there are times when a **capillary puncture** is a more appropriate method. During a capillary puncture, an incision is made in through the skin, which passes through the epidermis into the dermal layer. It is critical that the puncture reaches the dermal layer, as this is where the blood supply exists. Remember, the blood present in the capillary beds of the body is a mixture of venous and arterial blood. There is actually more arterial blood represented because of the increased arterial pressure forcing this blood into the capillary vessels. Because of the way that the skin layers are punctured, capillary samples may also include additional fluid from the space between the tissues and cells surrounding the capillary vessels. This is why it is very important to document the source of the blood specimen when collecting a capillary sample, as the reference ranges will be different from those used for venous blood.

Test Your Knowledge 8-13

Why is it important to document that a sample was capillary rather than venous? (Outcome 8-13)

Capillary samples are never appropriate when the test ordered requires a large amount of blood, or if there are problems with the circulation for the patient. This technique should also be avoided if the patient has edema, as the excessive fluid may dilute the sample and cause erroneous results. CLIA-waived tests require very little blood, so capillary punctures are often performed for these tests. Some patients who are undergoing cancer treatment or dialysis may be better suited for capillary blood draws to allow their blood levels to remain stable. Capillary punctures are also appropriate for most children, some geriatric patients with fragile veins, and infants.

Capillary punctures should always be performed using retractable nonreusable lancets for the safety of the patient and the employee performing the puncture. These devices are calibrated to provide a uniform depth for the puncture, and they lock in place once the puncture has occurred so that they cannot be reactivated. The devices used for capillary punctures are color coded according to the depth of the blade. They vary from those used for infants, which must be less than 2.0 mm (to avoid potentially touching bone), to adult depths up to 3 mm, primarily used for finger punctures. Lancet devices that may be used for home testing or perhaps for an individual CLIA-waived test procedure usually will not provide a puncture that is feasible for collecting enough blood to fill a microtube for analysis. All puncture devices must be disposed of in a biohazardous sharps container.

During a venipuncture, a tourniquet is used to aid in the palpation and visualization of an appropriate draw site. Capillary draws do not require the use of a tourniquet, but it is very important that the site to be used is as warm as possible to help increase the blood circulation. This may be accomplished by soaking the appendage in warm (not hot) water; by using a warm, wet compress; or with use of a commercial hand warmer held on the skin until it is warm to the touch.

The capillary puncture also requires the use of clean gauze (usually at least two sets), 70% isopropyl alcohol to clean the site, and a bandage to be used after the puncture. The employee who is performing the capillary puncture should also prepare the site under the patients' hand or foot by draping the area with an absorbent cloth or sheet to absorb any excess blood.

Capillary samples may be collected in various ways. **Microcollection tubes** may be utilized, which are smaller versions of the evacuated tubes used for venipuncture. These tubes are color coded in a way similar to the evacuated tubes, and many contain anticoagulants or other additives. **Capillary tubes** may also be used, especially for hematocrit testing. Capillary tubes are long, thin, hollow tubes designed to hold blood. The tubes are held horizontally with one end touching the drop of blood created by the capillary puncture. The tube fills with blood via **capillary action.** These tubes may be made of glass or plastic. They may be coated internally with heparin to keep the blood from clotting as it is collected, in which case the tube will have a red ring around it. Other capillary tubes do not include anticoagulant, and these plain tubes will have a blue ring around their diameter. Capillary tubes may also be part of a microcollection container setup, where the capillary action pulls the blood into the container. Figure 8-14 includes examples of devices used for capillary punctures.

Order of Draw for Capillary Puncture

The order of draw for the various microcollection containers using a capillary technique is different from the order used for venipuncture. The order is not the same because the capillary blood clots faster than the blood taken from a vein and because the potential for carryover

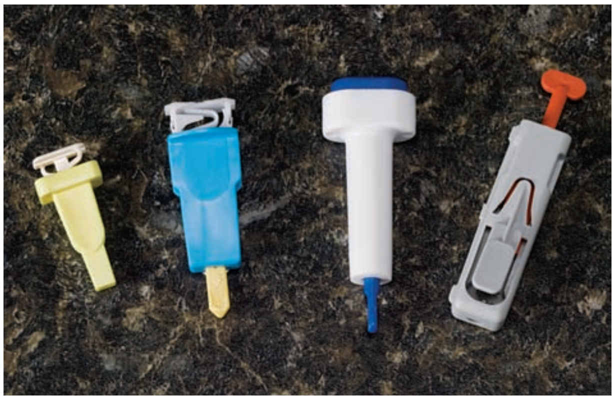

Figure 8-14 Devices used for capillary punctures. *Reprinted with permission from Eagle S, Brassington C, Dailey C, and Goretti C:* The Professional Medical Assistant: An Integrative, Teamwork-Based Approach. *Philadelphia: FA Davis, 2009.*

is not as high. For capillary draws, the order of draw is as follows:

1. **Lavender (EDTA) tubes are filled first.** It is imperative that these are well mixed during the process to avoid blood clot formation.
2. **Other additive tubes,** such as green top tubes or those with heparin and thixotropic gel, are collected next. Microcollection tubes are not used for coagulation studies, so there are no sodium citrate tubes to be filled.
3. **Nonadditive tubes are drawn last,** and will usually clot within 6 or 7 minutes.

Test Your Knowledge 8-14

Is the order of draw the same for capillary draws as it is for venipuncture specimens? (Outcome 8-11)

POINT OF INTEREST 8-4

Waste reduction considerations

Unfortunately, specimen collection generates a lot of waste. Current practice prohibits reusing most items that are exposed to bloodborne pathogens because of the danger of exposure for subsequent patients. The contaminated waste must be disposed of and processed appropriately by a licensed biohazardous waste handler. The waste is packaged according to federal specifications, transported in a standardized manner, and finally incinerated.

However, not all the waste in the laboratory setting is contaminated. Lids from needles, plastic and paper packaging, and alcohol prep pad packages need to be added to the regular trash, not treated as biohazardous materials. Some of the packaging may be recyclable; this option should be employed whenever possible. Any paper may include protected personal information and so needs to be shredded before disposal to protect patients' privacy. Some recycling programs will process shredded paper.

To minimize waste, do not open sterile packaging until you are certain that you will be using a specific device for specimen collection. Many phlebotomists will set up all the sterile supplies before they palpate the vein of the patient. This may be a mistake, as it may become apparent that a different device is needed for the blood draw. The sterility has been compromised for the devices that were set out, so they will all need to be discarded, in addition to the devices actually used for the blood draw. Become aware of recycling options in the laboratory setting, and make use of this service whenever possible.

PREPARATION FOR BLOOD COLLECTION

Several steps are required in the blood collection process prior to the actual puncture of the skin. Specific procedures are utilized regardless of the method used to collect the sample. These critical aspects include the following:

- **A professional atmosphere:** A collection site where phlebotomy is performed should be clean, organized, well lit, and tidy. Many times this collection site is the only contact that the patient will actually have with the laboratory where the testing will be performed, and if they feel as if this is a professional environment, the patient will have more confidence with the testing process to follow.
- **Appropriate introductions:** When first approaching a patient, the health-care employee must identify him- or herself. Proper identification should also be worn, such as a name tag and/or uniform so that the patient feels in the presence of a qualified professional for the procedure.
- **Patient identification:** This is the most critical aspect of the collection procedure. The health-care employee should ask for the patient to give his or her entire name and birth date for verification of identification. This information should be checked against the requisition form for accuracy. It is important that the patient actually says, "I am Sally Smith," rather than the health-care employee asking, "Are you Sally Smith?" Many times patients will answer yes to such a question, even though they did not hear or understand the name given. If drawing blood from a patient in an inpatient setting, armbands may be used for identification, but in addition, the patient identity also needs to be verified in another way. The phlebotomist must document who provided the verification of the patient identity if it is not the patient.
- **Verification of orders and patient preparation:** The medical assistant or phlebotomist needs to verify what tests were ordered before continuing with the collection procedure. This can be accomplished by looking carefully at the requisition form or labels provided for the blood draw. The medical assistant is also responsible to see if patients are prepared properly for the test

ordered; for instance, if the test is a fasting blood sugar, the medical assistant must ask patients if they have been without food and drink (except water and necessary medications) for 12 hours. If the preparation was not followed appropriately, the specimen is not to be drawn. When drawing blood to test for a therapeutic drug level, the medical assistant needs to ask when the medication was last taken, and verify if that is an appropriate time interval for this particular test.

- **Hand sanitization:** The medical assistant must wash his or her hands (if they are visibly soiled from a previous procedure) or use hand sanitizer appropriately before touching the patient. Gloves must be worn during the venipuncture process and should not be applied until the hands are totally dry.
- **Site selection**: The medical assistant should always start in the antecubital area when looking for an appropriate venipuncture site. The first choice is the median cubital vein, the second choice is the cephalic vein, and the last choice in this area is the basilic vein. The medical assistant should always check the veins in both arms to make the best choice. The back of the hand may also be used for venipuncture if there are no suitable sites in the antecubital area. For a capillary draw on an adult, the middle finger or ring finger is used, and part of the preparation may be warming the site.
- **Preparation of necessary supplies:** It is not possible to prepare fully for the venipuncture before the site selection has been performed, because the choice of which method to use for the blood draw will be dependent on the integrity and size of the veins to be used for the blood draw. Once the site selection is complete, the medical assistant can prepare an evacuated tube setup, a syringe for the blood draw, or a butterfly setup. The tubes to be used should also be checked for expiration and/or defects, and laid out in order of draw close to the patient so that they can be easily reached during the procedure. Gauze and alcohol should be within reach as well.

  Supplies may vary depending on the age and general health of the patient. An infant usually does not have venipuncture performed for most laboratory tests. Small children may need venipuncture, but a smaller volume of blood may be drawn using a butterfly and a 2- or 3-mL tube. Those patients who are over the age of 60 often present veins that collapse or roll easily, and due to complex health issues, it may be necessary to withdraw smaller volumes of blood for these populations. A syringe draw or butterfly setup will often be necessary.
- **Cleansing and support of the site:** Once the site has been selected, it must be cleaned. Seventy percent isopropyl alcohol is usually used to clean the site, and it is applied in a circular motion with the draw site at the center of the circle. The alcohol must be allowed to dry completely before the blood draw is performed. The site should not be repalpated after the alcohol has been used. If repalpation is absolutely necessary, the site should be cleaned again after being touched.

  The venipuncture site should be well supported by using a chair that has arms or a cross support designed for blood-drawing procedures. It may be helpful to have the patient make a fist with the opposite hand and place this under the elbow, which helps to expose the antecubital area fully for inspection and offers additional support. For inpatients, the patient's arm should be settled on a pillow or rolled-up towel to help make the veins more accessible.

  Support of the site may be complicated for small children. It may be necessary to ask the adult who accompanied the child to assist with the process by holding the child and securing the child's upper body for safety. A staff member may also be asked to assist in the process. The medical assistant should be ready for the blood draw as quickly as possible in this situation, and should not allow the child to see the needle for an excessive period of time before the venipuncture.
- **Patient communication:** Whenever the skin is to be punctured, remember to be honest with the patient about what will occur. *Never* tell a patient that the procedure won't hurt, or chastise the patient for expressing fear or discomfort with the procedure. Assure the patient that the procedure is necessary, and that it will be completed in as little time as possible. Try to adapt to the age of the patient by making appropriate small talk. Allowing the child to "blow away the pain" may be beneficial; the act of saying "ow" and blowing out may help to distract him or her from the process.

Test Your Knowledge 8-15

What is the most important aspect of patient preparation for blood collection procedures? (Outcome 8-14)

SPECIMEN PROCESSING

Once the blood specimen is collected, it may require further processing before it can be used for testing. The analysis may call for serum, plasma, or whole blood for

(Text continues on page 180)

Procedure 8-1: Venipuncture Using the Evacuated Tube System

The evacuated tube system is the method used most commonly for obtaining blood. It is fast and relatively safe, and allows for multiple tubes to be filled with one invasive procedure.

TASK

Successfully perform a venipuncture using the evacuated tube system. The process must be completed within 5 minutes.

CONDITIONS

- Hand-washing supplies and/or alcohol-based hand sanitizer
- Disposable gloves
- Tourniquet
- 70% isopropyl alcohol
- Double-pointed multi-sample needle
- Evacuated tube holder
- Laboratory requisition form or labels with specified test
- Evacuated tubes
- 2 x 2 gauze pads
- Adhesive bandage or wrap
- Test tube rack
- Biohazardous sharps container
- Biohazardous disposal bag

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.2: Perform Venipuncture **I.A.I.1:** Apply critical thinking skills in performing patient assessment and care **III.P.III.3:** Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations

ABHES Standards

- Medical Office Laboratory Procedures: Collect, label and process specimens: Perform venipuncture
- Medical Office Clinical Procedures: Apply principles of aseptic techniques and infection control
- Medical Office Clinical Procedures: Use Standard Precautions

Procedure	Rationale
1. Gather the requisition form and/or labels for the blood draw, and greet the patient. Identify yourself appropriately.	The requisition or labels are necessary to collect the correct type of specimen. It is always correct practice to identify yourself to the patient.
2. Wash hands (if they are visibly soiled) or apply hand sanitizer. Allow hands to dry completely.	Clean hands stop the spread of infection. Hands should be completely dry before attempting to apply gloves, or it will be difficult to put the gloves on the wet hands.
3. Verify the identity of the patient by asking for his or her name and at least one other unique identifier (such as the patient's birth date). Compare this information to the requisition form or labels.	Patient identification must always be verified using at least two unique factors.
4. Verify whether dietary restrictions were followed, and time of last medication dose, if needed.	Many laboratory tests require fasting specimens or other dietary restrictions. Drug dosage times are especially important for appropriate interpretation of the laboratory results.
5. Have the patient sit in the phlebotomy chair with appropriate arm support, and extend his or her arm to expose the antecubital area.	The arm should be supported for the venipuncture process, and the antecubital area will be the first area that is considered for the blood draw.

Procedure	Rationale
6. Apply gloves.	Gloves are required for procedures in which there is reasonable anticipation of exposure to blood or other potentially infectious materials.
7. Apply the tourniquet approximately 3 in. above the antecubital area (or above the wrist if a hand draw is necessary) and select an appropriate draw site using palpation. Use the tips of the ring finger and middle finger in a gentle probing motion as the antecubital area is examined. An appropriate vein will feel flexible and firm. Do not allow the tourniquet to remain on the arm for more than 1 minute.	The tourniquet needs to be placed high enough that it will not block the venipuncture site. Only blood vessels that can be palpated should be used for venipuncture; just looking at them is not sufficient. The thumb should never be used to feel for a vein, as there may be a pulse felt from the thumb that can be misleading. The thumb is also not as sensitive as the fingers. 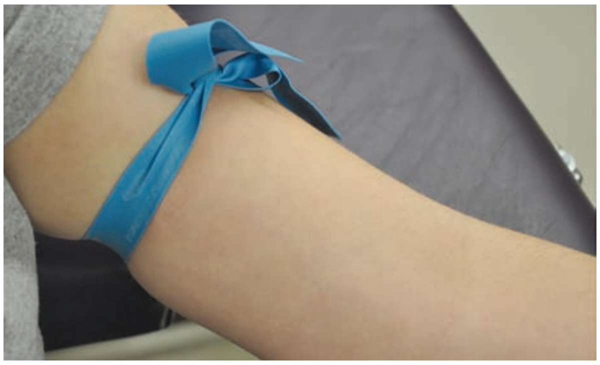
8. While palpating, have the patient make a fist, but *do not* allow the patient to pump his or her hand.	Pumping the hand can cause erroneous test results due to hemoconcentration.
9. If an appropriate site is not identified on the first arm, examine the other arm. If the antecubital area is not appropriate on either arm, examine the back of the hands for other opportunities.	It is important to choose the best site before inserting the needle. If the first arm examined doesn't provide a vein that is appropriate, check the other arm. If necessary, the back of the hands may be considered as a blood draw site.
10. Once an appropriate site is selected, determine the direction of the vein; is it running straight up and down or at an angle across the arm? Establish a visible landmark (a mole or dimple in the skin, etc.) for reference.	The needle must be inserted in such a way that it follows the direction of the vein for the best chance of success. A landmark is helpful so that the chosen venipuncture site can be identified after cleaning the area.
11. Have the patient open the hand to relax the fist until the tourniquet is reapplied.	Relaxing the hand will help the blood to flow normally as the site is prepared.
12. Remove the tourniquet.	A tourniquet may not stay on the arm for more than 1 minute, or hemoconcentration may result.

Continued

Procedure 8-1: Venipuncture Using the Evacuated Tube System—cont'd

Procedure	Rationale
13. Apply 70% isopropyl alcohol to the area in a circular motion with the draw site at the center of the circle. Allow the site to air-dry.	70% isopropyl alcohol will kill most of the bacterial contaminants on the surface of the skin. Never blow or fan the site to speed up the drying process, as this recontaminates the clean area.
14. As the alcohol is drying, assemble the necessary supplies.	Supplies need to be close by for ease of use and patient safety.
a. Choose a needle of appropriate size and attach it to the evacuated tube holder. Do not remove the cap that covers the long end of the needle until right before the skin is to be punctured.	This should always occur in the presence of the patient so that he or she is assured that the needle is sterile. The cap should not be removed until it is ready to be used to minimize the risk of contamination or possible injury. Needle sizes may vary according to the needs of the patient and the policy of the facility.
b. Arrange the necessary tubes for the tests ordered within easy reach of your nondominant hand, and have them set up in the correct order of draw.	The tubes may be laid out on the counter near the patient, or a small tube rack may work well for this purpose.
c. Verify that the tubes are not expired and that the anticoagulant is placed away from the rubber stopper within the tube. (You may need to shake the tube lightly to get the anticoagulant to move away from the stopper.) Check for any obvious defects in the tubes.	Use of an expired tube may result in a loss of vacuum and an unsuccessful blood draw. Potential crossover of anticoagulant from one tube to another should be minimized as much as possible. Defective (such as cracked or chipped tubes) should be discarded immediately.
15. Verify that gauze pads and an adhesive bandage are within reach. Extra tubes should also be close by so that they can be used if necessary.	The gauze pads and adhesive bandages need to be ready for use immediately when the needle is removed from the skin. If the phlebotomist feels that there may be a chance that the vacuum is exhausted in the first tube, using a second one that is nearby may allow the venipuncture to be successful.
16. Once the alcohol is dry, reapply the tourniquet. Do not repalpate the venipuncture site, or if absolutely necessary, clean the end of the gloved finger with alcohol before touching the site.	If the site is touched after cleansing, it will need to be cleaned again before the venipuncture can begin.
17. Have the patient make a fist.	Forming a fist may help with the visualization of the veins once the tourniquet is reapplied.
18. Remove the cap to expose the end of the needle to be inserted into the arm. Do not allow this needle to touch anything before piercing the skin.	Touching this needle to any surface before it pierces the skin will cause contamination and possible introduction of bacteria into the vein.

Procedure	Rationale
19. Stabilize the chosen vein by anchoring it with the thumb of the nondominant hand about 2 in. below the draw site, and/or off to the side. Make sure that the skin is pulled taut over the vein.	It is important to stabilize the vein, but if the thumb is placed too close to the draw site, it will be in the way of the needle insertion and could cause interference with the angle used for the process and result in an unsuccessful blood draw. 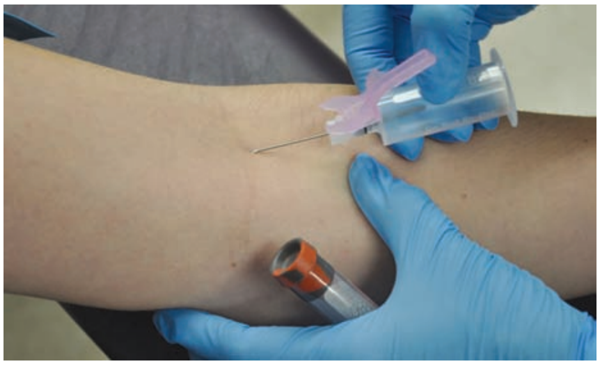
20. As you prepare to pierce the skin with the needle, it is good practice to warn the patient by saying something like, "Here we go; you will feel a stick."	If the patient is not expecting the skin to be pierced, he or she may be startled and move the arm or hand, causing an unsuccessful venipuncture.
21. Insert the needle at an angle of approximately 15 degrees, with the bevel up. The actual insertion site is approximately $^{1}/_{4}$ to $^{1}/_{2}$ inch below the identified draw site so that the needle is actually inserted into the vein at the chosen site.	If the angle is significantly less or more than 15 degrees, it may slide just above the vein or puncture through both sides of the vein rather than just entering the vein. The bevel facing upward allows the blood to enter the needle with minimal trauma.
22. The needle needs to be inserted quickly, with one smooth gentle motion. As the insertion is accomplished, follow the direction of the vein that was previously identified; if the vein is running at an angle across the arm or hand, this is how the needle should be directed into the vein. Approximately a third of the needle is usually below the surface of the skin when the insertion is complete.	Smooth insertion minimizes the trauma to the patient. It is important to follow the direction of the vein as the needle is inserted; this allows a better opportunity for the blood to enter the needle without obstruction. The needle needs to be inserted far enough to enter the vein, but not so far that it punctures both sides of the vein. 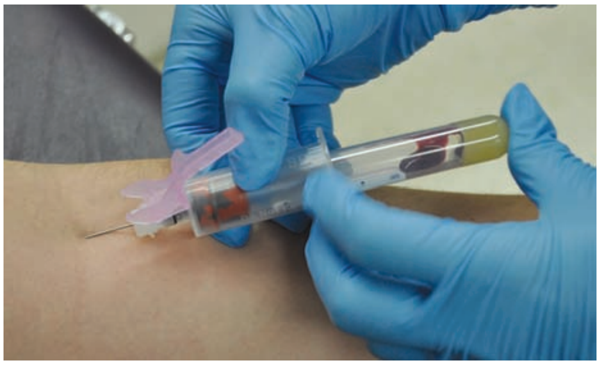

Continued

Procedure 8-1: Venipuncture Using the Evacuated Tube System—cont'd

Procedure	**Rationale**
23. While keeping the needle firmly seated in the vein, push the first tube into the needle holder, with the label facing downward. Use the index and middle fingers of the nondominant hand by placing them on either side of the flared edge at the bottom of the evacuated tube holder. Place the thumb of the nondominant hand on the bottom of the tube, and squeeze or pull the thumb toward the fingers, pushing the tube into the holder. If the needle is placed appropriately in the vein, blood should start to enter the tube as soon as the needle has pierced the stopper.	The hand used to hold the evacuated tube holder should rest firmly against the arm to keep the needle in place. If the label on the tube is facing upward, it is difficult to see the blood as it enters the tube. 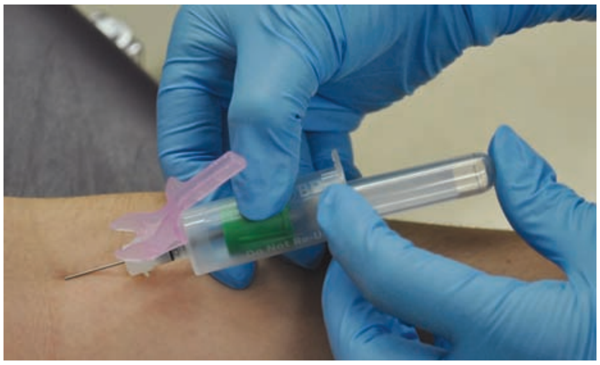
24. If only one tube is necessary for the tests ordered, release the tourniquet once the blood has started to enter the tube. If there is more than one tube necessary, the tourniquet may stay in place until the last tube has been inserted, as long as the length of time has not exceeded 1 minute since the tourniquet was applied.	The tourniquet must not stay on longer than 1 minute. It is always necessary to release the tourniquet before the needle is removed from the arm to avoid bleeding from the venipuncture site and introduction of air into the last vacuum tube used.
25. If it is necessary to use more than one tube for the blood draw, push against the flanges to remove the tube from the holder. Be careful not to disturb the needle placement as the tube is removed. Insert the next tube needed for the tests ordered, following the appropriate order of draw.	It is important to keep applying gentle pressure on the arm of the patient with the back of the fingers that are supporting the evacuated tube holder while removing and inserting tubes to keep the needle in place within the vein. It may be necessary to change tubes more than once when there are multiple tubes necessary for the tests ordered.
26. As each tube with additive is removed from the holder, invert the tube gently at least one or two times. This may be accomplished while the next tube is filling.	Tubes with anticoagulant added must be inverted sufficiently to mix the sample with the additive or the blood will clot and the specimen will need to be redrawn.
27. As the last tube is filling, ask the patient to relax the fist. Verify that the tourniquet has been released, and remove the last tube from the holder.	This is important to avoid a potential bruise or hematoma at the draw site, and also to avoid introduction of excess air into this last tube.

Procedure	Rationale
28. As the needle is removed from the skin, place gauze over the site without applying initial pressure. The needle is to be removed quickly, and at the same angle as the insertion. Do not apply pressure to the site with the gauze until the needle has been removed completely from the skin, as this will cause pain for the patient.	The gauze placed over the site will minimize bleeding as the needle is removed.
29. Ask the patient to apply pressure to the site for 3 to 5 minutes. Do not allow the patient to bend his or her arm.	The pressure should be adequate to stop any bleeding before the patient leaves the drawing area. Bending the arm increases the risk for bleeding and bruise formation.
30. Once the needle has been removed from the arm, immediately activate the needle safety device and discard the entire setup in the biohazardous sharps container. Do not remove the needle from the evacuated tube holder before disposal. *Never recap contaminated needles.*	The evacuated tube setup is designed to be used one time and discarded without removing the needle before disposal. Recapping contaminated needles increases the risk of needlestick accidents and potential exposure to bloodborne pathogens. 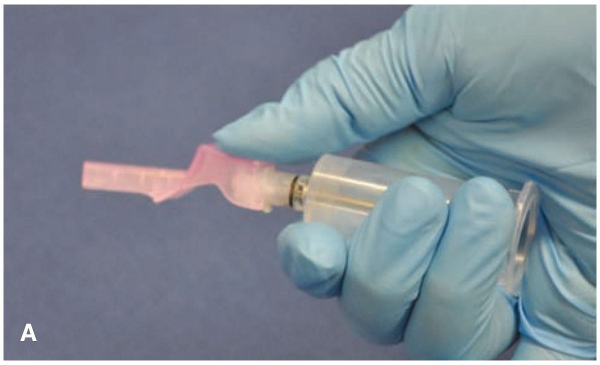

Continued

Procedure 8-1: Venipuncture Using the Evacuated Tube System—cont'd

Procedure	Rationale
	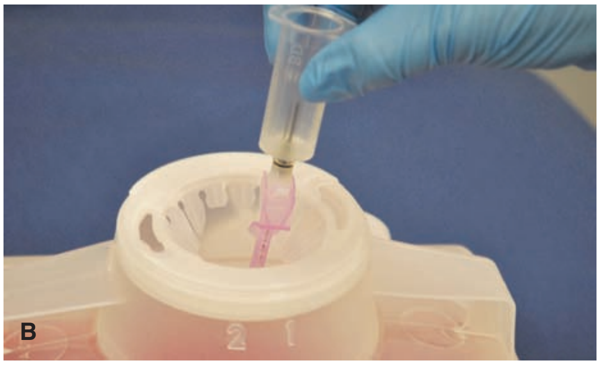
31. As the patient is applying pressure to the draw site, gently invert the blood specimens containing anticoagulant 5 to 10 times.	Insufficient mixing of the anticoagulant with the specimen will result in clotted blood and a sample that must be discarded.
32. Label the tubes, including the full name of the patient, birth date (or other unique identifier assigned by the health-care provider), the date and time of collection, and the initials of the person who drew the sample.	All samples must be labeled using at least two unique identifiers. The date and time provide additional information that may be necessary for interpretation of the results. Identification of the phlebotomist may be helpful if there are questions about the procedure or the specimens collected.
33. Use the requisition form or the labels to verify that all the necessary tubes were drawn for the specimens ordered.	Verifying the details of the blood draw once more in the presence of the patient allows for a redraw to be performed immediately if something is missing.

Procedure	Rationale
34. Observe any special handling instructions for the specimens.	Some types of specimens must remain at room temperature, whereas others may need to be put on ice immediately. This should be information that is ascertained before the procedure starts.
35. Monitor the patient for any signs of distress from the procedure.	These may include pallor, perspiration (especially on the upper lip or forehead), increased anxiety, or light-headedness. If the patient is exhibiting any of these symptoms, it is best to have him or her lie down if possible. This may be easily accomplished if the patient is in a chair that can recline. If not, a cold compress on the forehead and/or the back of the neck may help. Continue to converse with the patient and move the blood out of sight. Ask for assistance if you feel that your patient is feeling faint. If the patient loses consciousness, it may be necessary to lower him or her to the floor from the phlebotomy chair.
36. Once the tubes have been labeled, check the draw site for bleeding. If it has not stopped bleeding, apply pressure for another few minutes, and if the bleeding is still present, contact a physician.	It is important to look under the gauze for 2 or 3 seconds before applying the adhesive bandage to be certain that the site has actually stopped bleeding.
37. If the bleeding has subsided, apply an adhesive bandage, but leave the gauze in place to allow for additional pressure.	Self-adhesive bandages (such as Coban) may be wrapped around the site rather than applying an adhesive to the skin directly. Self-adhesive bandages may be especially effective for those patients who are on anticoagulant therapy.
38. Advise the patient to avoid heavy lifting or excessive exercise of the arm used for venipuncture for at least 1 hour.	Heavy lifting or exercise could cause the site to resume bleeding.
39. Assist the patient to stand (if necessary) and thank him or her for being cooperative.	Patients may be a bit light-headed, and assistance may be needed when he or she first stands up.
40. Discard all trash, and place the tubes in an appropriate vessel for processing. Never touch full tubes of blood without wearing gloves.	Touching the tubes of blood without wearing gloves offers potential opportunities for exposure to bloodborne pathogens.
41. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.

Continued

Procedure 8-1: Venipuncture Using the Evacuated Tube System—cont'd

Procedure	Rationale
42. Document the procedure in the patient's chart as well as on the requisition form. Include the date and time of collection, what was collected, and if there were any circumstances that were out of the ordinary. Also document where the vein was accessed (such as right arm, left hand, etc.). The documentation should include the type (color) of tubes drawn, as well as the identity of the person who collected the sample.	All patient interactions must be carefully documented. The site used for vein access should always be written in the chart or on the requisition form so that if there are any negative outcomes associated with the procedure, the site has been noted.

Date	
1/18/2020:	Phlebotomy performed in right antecubital area for CMP and CBC. Tiger top and lavender top tube drawn.
1100 a.m.	Connie Lieseke, CMA (AAMA)

Procedure 8-2: Venipuncture Using a Syringe

A syringe and needle may be used for venipuncture in situations in which the available veins are prone to move during the blood draw, or when the vein may not be capable of supporting the vacuum exerted from an evacuated tube. This may be the case if the vein is small, or if it lacks "bounce" when palpated. Syringe draws are an excellent choice in the antecubital area when the veins are not quite suitable for a blood draw using an evacuated tube system setup.

TASK

Successfully perform a venipuncture using a syringe. After the blood has been obtained, transfer it successfully into the necessary tubes for analysis. The process must be completed within 5 minutes.

CONDITIONS

- Hand-washing supplies and/or alcohol-based hand sanitizer
- Disposable gloves
- Tourniquet
- 70% isopropyl alcohol
- Syringe (5- to 15-mL capacity)
- 21- to 23-gauge safety enabled needle
- Transfer device
- Laboratory requisition form or labels with specified test
- Evacuated tubes
- 2 x 2 gauze pads
- Adhesive bandage or wrap
- Test tube rack
- Biohazardous sharps container
- Biohazardous disposal bag

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.2: Perform Venipuncture **I.A.I.1:** Apply critical thinking skills in performing patient assessment and care **III.P.III.3:** Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations

ABHES Standards

- Medical Office Laboratory Procedures: Collect, label and process specimens: Perform venipuncture
- Medical Office Clinical Procedures: Apply principles of aseptic techniques and infection control
- Medical Office Clinical Procedures: Use Standard Precautions

Procedure	Rationale
1. Gather the requisition form and/or labels for the blood draw, and greet the patient. Identify yourself appropriately.	The requisition or labels are necessary to collect the correct type of specimen. It is always correct practice to identify yourself to the patient.
2. Wash hands (if they are visibly soiled) or apply hand sanitizer. Allow hands to dry completely.	Clean hands stop the spread of infection. Hands should be completely dry before attempting to apply gloves, or it will be difficult to put the gloves on the wet hands.
3. Verify the identification of the patient by asking for his or her name and at least one other unique identifier (such as the patient's birth date). Compare this information to the requisition form or labels.	Patient identification must always be verified using at least two unique factors.
4. Verify whether diet restrictions were followed, and time of last medication dose, if needed.	Many laboratory tests require fasting specimens or other dietary restrictions. Drug dosage times are especially important for appropriate interpretation of the laboratory results.
5. Have the patient sit in the phlebotomy chair with appropriate arm support, and extend his or her arm to expose the antecubital area.	The arm should be supported for the venipuncture process, and the antecubital area will be the first area that is considered for the blood draw.
6. Apply gloves.	Gloves are required for procedures in which there is reasonable anticipation of exposure to blood or other potentially infectious materials.
7. Apply the tourniquet approximately 3 in. above the antecubital area (or above the wrist if a hand draw is necessary) and select an appropriate draw site using palpation. Use the tips of the ring finger and middle finger in a gentle probing motion as the antecubital area is examined. An appropriate vein will feel flexible and firm. Do not allow the tourniquet to remain on the arm for more than 1 minute.	The tourniquet needs to be placed high enough that it will not block the venipuncture site. Only blood vessels that can be palpated should be used for venipuncture; just looking at them is not sufficient. The thumb should never be used to feel for a vein, as there may be a pulse felt from the thumb that can be misleading. The thumb is also not as sensitive as the fingers.
8. While palpating, have the patient make a fist, but *do not* allow the patient to pump his or her hand.	Pumping the hand can cause erroneous results because of hemoconcentration.
9. If an appropriate site is not identified on the first arm, examine the other arm. If the antecubital area is not appropriate on either arm, examine the back of the hands for other opportunities.	It is important to choose the best site before inserting the needle. If the first arm examined doesn't provide a vein that is appropriate, check the other arm. If necessary, the back of the hands may be considered as a blood draw site.
10. Once an appropriate site is selected, determine the direction of the vein; is it running straight up and down or at an angle across the arm? Establish a visible landmark (a mole or dimple in the skin, etc.). for reference.	The needle must be inserted in such a way that it follows the direction of the vein for the best chance of success. A landmark is helpful so that the chosen venipuncture site can be identified after cleaning the area.

Continued

Procedure 8-2: Venipuncture Using a Syringe—cont'd

Procedure	Rationale
11. Have the patient open the hand to relax the fist until the tourniquet is reapplied.	Relaxing the hand will help the blood to flow normally as the site is prepared.
12. Remove the tourniquet.	A tourniquet may not stay on the arm for more than 1 minute, or hemoconcentration may result.
13. Apply 70% isopropyl alcohol to the area in a circular motion with the draw site at the center of the circle. Allow the site to air-dry.	70% isopropyl alcohol will kill most of the bacterial contaminants on the surface of the skin. Never blow or fan the site to speed up the drying process, as this recontaminates the clean area.
14. As the alcohol is drying, assemble the necessary supplies.	Supplies need to be close by for ease of use and patient safety.
a. Open the sterile package holding the syringe	Syringes should remain sterile until the time of use.
b. Pull back and push forward the plunger of the syringe several times to verify whether it moves smoothly. Be certain that the plunger is pushed forward as far as it goes before continuing with the process.	"Exercising" the plunger in this way allows for smooth movement when pulling back the plunger to allow blood to enter the syringe. If the plunger does not move smoothly or is too loose within the barrel of the syringe, it should be discarded.
c. Remove the needle from the sterile wrapper. Do not take off the cap covering the end of the needle until just before insertion into the patient's arm.	The needle must remain sterile until just prior to insertion. It cannot touch other surfaces before piercing the skin of the patient.
d. Attach the needle to the syringe tip.	Verify that this is a secure connection so that blood will not leak out around the needle during the venipuncture process.
e. Place gauze pads and adhesive bandage within reach of the nondominant hand for use at the end of the venipuncture.	Supplies to be used during the venipuncture process should be within easy reach of the nondominant hand to avoid reaching over the site where the needle is inserted in the arm.
f. Organize tubes needed for analysis and unwrap transfer device. Place within easy reach.	When the venipuncture procedure is finished, it is necessary to transfer the blood into the tubes quickly to avoid clot formation in the sample. Organization of supplies before the process is critical.
g. Verify that the tubes are not expired and that the anticoagulant is placed away from the rubber stopper within the tube. (You may need to shake the tube lightly to get the anticoagulant to move away from the stopper.) Check for any obvious defects in the tubes.	Use of an expired tube may result in a loss of vacuum and an unsuccessful blood draw. Potential crossover of anticoagulant from one tube to another should be minimized as much as possible. Defective (such as cracked or chipped tubes) should be discarded immediately.
15. Once the alcohol is dry and the supplies are assembled, reapply the tourniquet. Do not repalpate the venipuncture site, or if absolutely necessary, clean the end of the gloved finger with alcohol before touching the site.	If the site is touched after cleansing, it will need to be cleaned again before the venipuncture can begin.

Procedure	Rationale
16. Have the patient make a fist.	Forming a fist may help with the visualization of the veins once the tourniquet is reapplied.
17. Remove the cap to expose the end of the needle to be inserted into the arm. Do not allow this needle to touch anything before piercing the skin.	Touching this needle to any surface before it pierces the skin will cause contamination and possible introduction of bacteria into the vein.
18. Stabilize the chosen vein by anchoring it with the thumb of the nondominant hand about 2 in. below the draw site, and/or off to the side. Make sure that the skin is pulled taut over the vein.	It is important to stabilize the vein, but if the thumb is placed too close to the draw site, it will be in the way of the needle insertion and could cause interference with the angle used for the process and result in an unsuccessful blood draw.
19. As you prepare to pierce the skin with the needle, it is good practice to warn the patient by saying something like, "Here we go; you will feel a stick."	If the patient is not expecting the skin to be pierced, he or she may be startled and move the arm or hand, causing an unsuccessful venipuncture.
20. Insert the needle at an angle of approximately 15 degrees, with the bevel up. The actual insertion site is approximately $^{1}/_{4}$ to $^{1}/_{2}$ inch below the identified draw site so that the needle is actually inserted into the vein at the chosen site.	If the angle is significantly less or more than 15 degrees, it may slide just above the vein or puncture through both sides of the vein rather than just entering the vein. The bevel facing upward allows the blood to enter the needle with minimal trauma.
21. The needle needs to be inserted quickly, with one smooth gentle motion. As the insertion is accomplished, follow the direction of the vein that was previously identified; if the vein is running at an angle across the arm or hand, this is how the needle should be directed into the vein. Approximately a third of the needle is usually below the surface of the skin when the insertion is complete.	Smooth insertion minimizes the trauma to the patient. It is important to follow the direction of the vein as the needle is inserted; this allows a better opportunity for the blood to enter the needle without obstruction. The needle needs to be inserted far enough to enter the vein, but not so far that it punctures both sides of the vein. 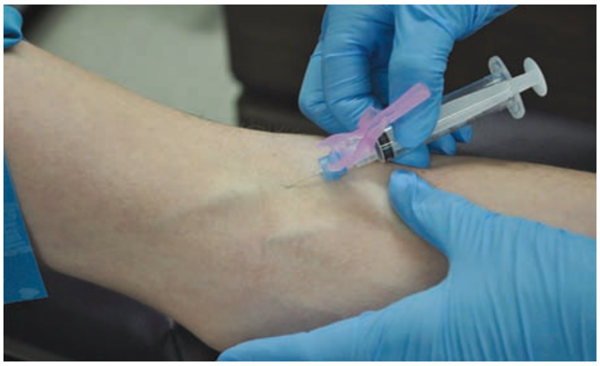

Continued

Procedure 8-2: Venipuncture Using a Syringe—cont'd

Procedure	Rationale
22. When using the syringe technique, there may be a "flash" of blood visible in the hub of the needle as soon as it is inserted in the vein. Gently pull back on the plunger of the syringe to create a vacuum and pull the blood into the syringe.	It is important to pull back relatively slowly because if the plunger is pulled back too fast, it can cause hemolysis. It is also possible that if the vein is small, it will collapse with the pressure applied by pulling the plunger back too quickly.
23. Continue to pull back slowly on the plunger and monitor the volume of blood entering the syringe. Keep gentle pressure on the arm of the patient with the back of the fingers holding the syringe.	The medical assistant must continue to pull back on the plunger so that the blood will continue to enter the syringe. To keep the needle from moving during the process, pressure should be applied with the back of the fingers holding the syringe. 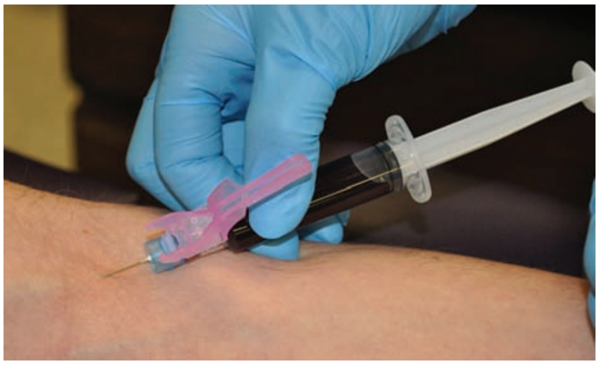
24. When the required amount of blood has almost been obtained, release the tourniquet and have the patient open his or her fist.	The tourniquet must not stay on longer than 1 minute. It is always necessary to release the tourniquet and have the patient open the fist before the needle is removed from the arm to avoid bleeding from the venipuncture site.
25. As the needle is removed from the skin, place gauze over the site, without applying initial pressure. The needle is to be removed quickly, and at the same angle as the insertion. Do not apply pressure to the site with the gauze until the needle has been removed completely from the skin, as this will cause pain for the patient.	The gauze placed over the site will minimize bleeding as the needle is removed.
26. Ask the patient to apply pressure to the site for 3 to 5 minutes. Do not allow the patient to bend his or her arm.	The pressure should be adequate to stop any bleeding before the patient leaves the drawing area. Bending the arm increases the risk for bleeding and bruise formation.

Procedure	Rationale
27. Once the needle has been removed from the arm, immediately activate the needle safety device. Grasp the needle at its base, and remove it from the syringe. Discard the needle in the biohazardous sharps container. *Never recap contaminated needles.*	The needle must be removed from the syringe so that the transfer device can be applied. Recapping contaminated needles increases the risk of needlestick accidents and potential exposure to bloodborne pathogens.
28. Screw the transfer device onto the end of the syringe filled with blood.	Make sure this is a secure seal so that the blood will flow adequately.
29. While holding the syringe upright, insert each evacuated tube into the open end of the transfer device. Follow the recommended order of draw.	Holding the syringe upright will minimize the opportunity for anticoagulant crossover as the transfer device is used. 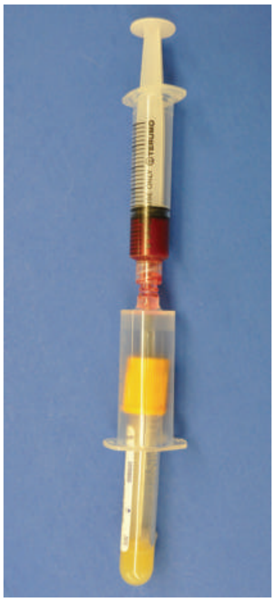
30. Dispose of the transfer device and syringe in a biohazardous sharps container.	There is a needle within the transfer device, so this must go into a sharps container.
31. Gently invert the blood specimens containing anticoagulant 5 to 10 times.	Insufficient mixing of the anticoagulant with the specimen will result in clotted blood and a sample that must be discarded.
32. Label the tubes, including the full name of the patient, birth date (or other unique identifier assigned by the health-care provider), the date and time of collection, and the initials of the collector.	All samples must be labeled using at least two unique identifiers. The date and time provide additional information that may be necessary for interpretation of the results. Identification of the phlebotomist may be helpful if there are questions about the procedure or the specimens collected.

Continued

Procedure 8-2: Venipuncture Using a Syringe—cont'd

Procedure	Rationale
33. Use the requisition form or the labels to verify that all the necessary tubes were drawn for the specimens ordered.	Verifying the details of the blood draw once more in the presence of the patient allows for a redraw to be performed immediately if something is missing.
34. Observe any special handling instructions for the specimens.	Some types of specimens must remain at room temperature, whereas others may need to be put on ice immediately. This should be information that is ascertained before the procedure starts.
35. Monitor the patient for any signs of distress from the procedure.	These may include pallor, perspiration (especially on the upper lip or forehead), increased anxiety, or light-headedness. If the patient is exhibiting any of these symptoms, it is best to have him or her lie down if possible. This may be easily accomplished if the patient is in a chair that can recline. If not, a cold compress on the forehead and/or the back of the neck may help. Continue to converse with the patient and move the blood out of sight. Ask for assistance if you feel that your patient is feeling faint. If the patient loses consciousness, it may be necessary to lower him or her to the floor from the phlebotomy chair.
36. Once the tubes have been labeled, check the draw site for bleeding. If it has not stopped bleeding, apply pressure for another few minutes, and if the bleeding is still present, contact a physician.	It is important to look under the gauze for 2 or 3 seconds before applying the adhesive bandage to be certain that the site has actually stopped bleeding.
37. If the bleeding has subsided, apply an adhesive bandage, but leave the gauze in place to allow for additional pressure.	Self-adhesive bandages (such as Coban) may be wrapped around the site rather than applying an adhesive to the skin directly. Self-adhesive bandages may be especially effective for those patients who are on anticoagulant therapy.
38. Advise the patient to avoid heavy lifting or excessive exercise of the arm used for venipuncture for at least 1 hour.	Heavy lifting or exercise could cause the site to resume bleeding.
39. Assist the patient to stand (if necessary) and thank them for their cooperation.	Patients may be a bit light-headed and assistance may be needed when he or she first stands up.
40. Discard all trash, and place the tubes in an appropriate vessel for processing. Never touch full tubes of blood without wearing gloves.	Touching the tubes of blood without wearing gloves offers potential opportunities for exposure to bloodborne pathogens.
41. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.

Procedure	Rationale
42. Document the procedure in the patient's chart as well as on the requisition form. Include the date and time of collection, what was collected, and if there were any circumstances that were out of the ordinary. Also document where the vein was accessed (such as right arm, left hand, etc.). The documentation should include the type (color) of tubes drawn, as well as the identity of the person who collected the sample.	All patient interactions must be carefully documented. The site used for vein access should always be written in the chart or on the requisition form so that if there are any negative outcomes associated with the procedure, the site has been noted.

Date	
1/18/2020:	Phlebotomy performed in left antecubital area for H&H. Lavender top tube drawn.
1100 a.m.	Connie Lieseke, CMA (AAMA)

Procedure 8-3: Venipuncture Using the Butterfly (Winged Infusion) System

The butterfly system is often used for obtaining blood from patients with small, fragile veins. It is the system of choice when performing blood draws from the hand of patients. When utilized for venipuncture, 23-gauge needles are usually used.

TASK

Successfully perform a venipuncture using a butterfly (winged infusion) system. The process must be completed within 5 minutes.

CONDITIONS

- Hand-washing supplies and/or alcohol-based hand sanitizer
- Disposable gloves
- Tourniquet
- 70% isopropyl alcohol
- Butterfly needle (winged infusion set)
- Evacuated tube holder and luer adapter, or syringe and transfer device
- Laboratory requisition or labels with specified test
- Evacuated tubes
- 2 x 2 gauze pads
- Adhesive bandage or wrap
- Test tube rack
- Biohazardous sharps container
- Biohazardous disposal bag

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.2: Perform Venipuncture **I.A.I.1:** Apply critical thinking skills in performing patient assessment and care **III.P.III.3:** Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations

ABHES Standards

- Medical Office Laboratory Procedures: Collect, label and process specimens: Perform venipuncture
- Medical Office Clinical Procedures: Apply principles of aseptic techniques and infection control
- Medical Office Clinical Procedures: Use Standard Precautions

Continued

Procedure 8-3: Venipuncture Using the Butterfly (Winged Infusion) System—cont'd

Procedure	Rationale
1. Gather the requisition and/or labels for the blood draw, and greet the patient. Identify yourself appropriately.	The requisition or labels are necessary to collect the correct type of specimen. It is always correct practice to identify yourself to the patient.
2. Wash hands (if they are visibly soiled) or apply hand sanitizer. Allow hands to dry completely.	Clean hands stop the spread of infection. Hands should be completely dry before attempting to apply gloves, or it will be difficult to put the gloves on the wet hands.
3. Verify the identification of the patient by asking for his or her name and at least one other unique identifier (such as the patient's birth date). Compare this information to the requisition or labels.	Patient identification must always be verified using at least two unique factors.
4. Verify whether dietary restrictions were followed, and time of last medication dose, if needed.	Many laboratory tests require fasting specimens or other dietary restrictions. Drug dosage times are especially important for appropriate interpretation of the laboratory results.
5. Have the patient sit in the phlebotomy chair with appropriate arm support, and extend his or her arm to expose the antecubital area.	The arm should be supported for the venipuncture process, and the antecubital area will be the first area that is considered for the blood draw.
6. Apply gloves.	Gloves are required for procedures in which there is reasonable anticipation of exposure to blood or other potentially infectious materials.
7. Apply the tourniquet approximately 3 in. above the antecubital area (or above the wrist if a hand draw is necessary) and select an appropriate draw site using palpation. Use the tips of the ring finger and middle finger in a gentle probing motion as the antecubital area is examined. An appropriate vein will feel flexible and firm. Do not allow the tourniquet to remain on the arm for more than 1 minute.	The tourniquet needs to be placed high enough that it will not block the venipuncture site. Only blood vessels that can be palpated should be used for venipuncture; just looking at them is not sufficient. The thumb should never be used to feel for a vein, as there may be a pulse felt from the thumb that can be misleading. The thumb is also not as sensitive as the fingers.
8. While palpating, have the patient make a fist, but *do not* allow them to pump his or her hand.	Pumping the hand can cause erroneous results because of hemoconcentration.
9. If an appropriate site is not identified on the first arm, examine the other arm. If the antecubital area is not appropriate on either arm, examine the back of the hands for other opportunities.	It is important to choose the best site before inserting the needle. If the first arm examined doesn't provide a vein that is appropriate, check the other arm. If necessary, the back of the hands may be considered.

Procedure	Rationale
10. Once an appropriate site is selected, determine the direction of the vein; is it running straight up and down or at an angle across the arm? Establish a visible landmark (a mole or dimple in the skin, etc.) for reference.	The needle must be inserted in such a way that it follows the direction of the vein for the best chance of success. A landmark is helpful so that the chosen venipuncture site can be identified after cleaning the area.
11. Have the patient open the hand to relax the fist until the tourniquet is reapplied.	Relaxing the hand will help the blood to flow normally as the site is prepared.
12. Remove the tourniquet.	A tourniquet may not stay on the arm for more than 1 minute, or hemoconcentration may result.
13. Apply 70% isopropyl alcohol to the area in a circular motion with the draw site at the center of the circle. Allow the site to air-dry.	70% isopropyl alcohol will kill most of the bacterial contaminants on the surface of the skin. Never blow or fan the site to speed up the drying process, as this recontaminates the clean area.
14. As the alcohol is drying, assemble the necessary supplies. Start by opening the sterile package holding the butterfly setup and uncurling the tubing by gently stretching it.	Supplies need to be close by for ease of use and patient safety. The butterfly system should remain sterile until just before use. If the tubing stays tightly curled, it is difficult to use it effectively.
Option 1. Butterfly System With a Syringe	
a. Open the sterile package holding the syringe	Syringes should remain sterile until the time of use.
b. Pull back and push forward the plunger of the syringe several times to verify whether it moves smoothly.	"Exercising" the plunger in this way allows for smooth movement when pulling back the plunger to allow blood to enter the syringe. If the plunger does not move smoothly or is too loose within the syringe, it should be discarded.
c. Attach the end of the butterfly tubing to the syringe.	This tubing needs to be attached securely to allow the blood to enter the syringe.
d. Open a transfer device and place within reach.	There is a limited amount of time allowed to transfer the blood from the syringe to the tubes before it clots, so it is important to have the transfer device ready.
Option 2. Butterfly System With Evacuated Tube Holder	
a. Verify that the butterfly system has a luer adapter to be used with the evacuated tube holder.	The luer adapter has a needle covered with a rubber sleeve that pierces the tubes when using the evacuated tube holder.
b. Screw the butterfly tubing with the adapter attached into the evacuated tube holder.	The tubing and the holder must be securely attached to one another to allow the blood to flow appropriately into the tubes.

Continued

Procedure 8-3: Venipuncture Using the Butterfly (Winged Infusion) System—cont'd

Procedure	Rationale
15. Place gauze pads and adhesive bandage within reach of the nondominant hand for use at the end of the venipuncture.	Supplies to be used during the venipuncture process should be within easy reach of the nondominant hand to avoid reaching over the site where the needle is inserted in the arm.
16. Organize tubes needed for analysis and place within easy reach.	To allow for the blood collection process to proceed smoothly, it is necessary organize the supplies before getting started.
17. Verify that the tubes are not expired and that the anticoagulant is placed away from the rubber stopper within the tube. (You may need to shake the tube lightly to get the anticoagulant to move away from the stopper.) Check for any obvious defects in the tubes.	Use of an expired tube may result in a loss of vacuum and an unsuccessful blood draw. Potential crossover of anticoagulant from one tube to another should be minimized as much as possible. Defective (such as cracked or chipped tubes) should be discarded immediately.
18. Once the alcohol is dry and the supplies are assembled, reapply the tourniquet. Do not repalpate the venipuncture site, or if absolutely necessary, clean the end of the gloved finger with alcohol before touching the site.	If the site is touched after cleansing, it will need to be cleaned again before the venipuncture can begin.
19. Have the patient make a fist.	Forming a fist may help with the visualization of the veins once the tourniquet is reapplied.
20. Remove the cap to expose the end of the needle to be inserted into the arm. Do not allow this needle to touch anything before piercing the skin.	Touching this needle to any surface before it pierces the skin will cause contamination and possible introduction of bacteria into the vein.
21. Stabilize the chosen vein by anchoring it with the thumb of the nondominant hand about 2 in. below the draw site, and/or off to the side. Make sure that the skin is pulled taut over the vein.	It is important to stabilize the vein, but if the thumb is placed too close to the draw site, it will be in the way of the needle insertion and could cause interference with the angle used for the process and result in an unsuccessful blood draw.
22. As you prepare to pierce the skin with the needle, it is good practice to warn the patient by saying something like, "Here we go; you will feel a stick."	If the patient is not expecting the skin to be pierced, he or she may be startled and move the arm or hand, causing an unsuccessful venipuncture.
23. Grasp the butterfly needle with the wings on either side. The textured area of the plastic wings is designed to be against the fingers; this allows the bevel of the needle to face upward.	If the needle is not grasped appropriately, the bevel will point downward and could affect the success of the blood draw.

Procedure	Rationale
24. Insert the needle at an angle of 5 to 10 degrees, with the bevel up. The actual insertion site is approximately $^{1}/_{4}$ to $^{1}/_{2}$ inch below the identified draw site so that the needle is actually inserted into the vein at the chosen site.	If the angle is significantly less or more than 5 to 10 degrees, it may slide just above the vein or puncture through both sides of the vein rather than just entering the vein. The bevel facing upward allows the blood to enter the needle with minimal trauma. 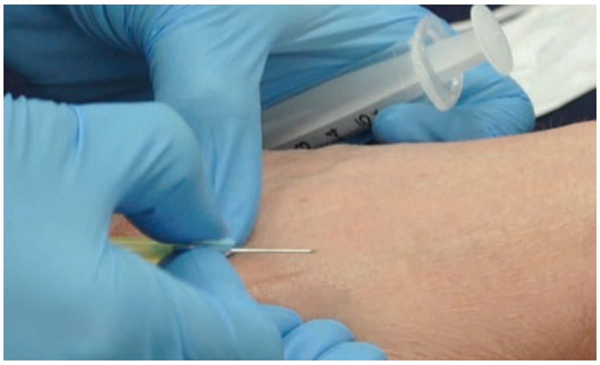
25. The needle needs to be inserted quickly, with one smooth gentle motion. As the insertion is accomplished, follow the direction of the vein that was previously identified; if the vein is running at an angle across the arm or hand, this is how the needle should be directed into the vein. Approximately a third of the needle is usually below the surface of the skin when the insertion is complete.	Smooth insertion minimizes the trauma to the patient. It is important to follow the direction of the vein as the needle is inserted; this allows a better opportunity for the blood to enter the needle without obstruction. The needle needs to be inserted far enough to enter the vein, but not so far that it punctures both sides of the vein.
26. When the needle pierces the vein, the blood will immediately be present in the tubing. As soon as there is blood present, begin to pull the plunger of the syringe back slowly, if using a syringe. If the evacuated tube holder is used, push on the first tube to be drawn so that the vacuum pulls the blood through the tubing.	The blood will not flow into the tubing adequately without the vacuum of the plunger or the evacuated tube. Do not pull back quickly, or it may cause hemolysis or collapse the vein.
27. If using the syringe system, continue to pull back slowly on the plunger and monitor the volume of blood entering the syringe. Change the evacuated tubes as needed if using these directly to draw the blood. Keep gentle pressure on the arm of the patient with the back of the fingers holding the butterfly setup.	The medical assistant must continue to pull back on the plunger so that the blood will continue to enter the syringe. To keep the needle from moving during the process, pressure should be applied with the back of the fingers holding the syringe.
28. When the required amount of blood has almost been obtained, release the tourniquet and have the patient open the fist.	The tourniquet must not stay on longer than 1 minute. It is always necessary to release the tourniquet and have the patient open the fist before the needle is removed from the arm to avoid bleeding from the venipuncture site.

Continued

Procedure 8-3: Venipuncture Using the Butterfly (Winged Infusion) System—cont'd

Procedure	Rationale
29. As the needle is removed from the skin, place gauze over the site, without applying initial pressure. The needle is to be removed quickly, and at the same angle as the insertion. Do not apply pressure to the site with the gauze until the needle has been removed completely from the skin, as this will cause pain for the patient.	The gauze placed over the site will minimize bleeding as the needle is removed.
30. Ask the patient to apply pressure to the site for 3 to 5 minutes. Do not allow the patient to bend his or her arm if the draw was performed in the antecubital space.	The pressure should be adequate to stop any bleeding before the patient leaves the drawing area. Bending the arm increases the risk for bleeding and bruise formation.
31. Once the needle has been removed from the arm, immediately activate the needle safety device.	The safety device may be a push-button device that is activated while the needle is still in the skin, or it may be activated right after removal. It is imperative that the medical assistant keeps his or her fingers behind the needle while activating the safety device. 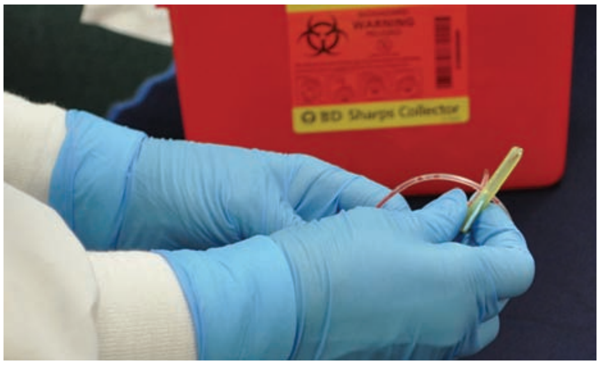
32. If a syringe was used for the procedure, remove the tubing from the end of the syringe and discard the butterfly unit into a biohazardous sharps container. If an evacuated tube system was used, discard the evacuated tube holder with the butterfly setup.	The tubing must be removed so that the transfer device can be applied to put the blood into the tubes.
33. Screw the transfer device onto the end of the syringe filled with blood.	Make sure this is a secure seal so that the blood will flow adequately.
34. While holding the syringe upright, insert each evacuated tube into the open end of the transfer device. Follow the recommended order of draw.	Holding the syringe upright will minimize the opportunity for anticoagulant crossover as the transfer device is used.
35. Dispose of the transfer device and syringe in a biohazardous sharps container.	There is a needle within the transfer device, so this must go into a sharps container.

Procedure	Rationale
36. Gently invert the blood specimens containing anticoagulant 5 to 10 times.	Insufficient mixing of the anticoagulant with the specimen will result in clotted blood and a sample that must be discarded.
37. Label the tubes, including the full name of the patient, birth date (or other unique identifier assigned by the health-care provider), the date and time of collection, and the initials of the collector.	All samples must be labeled using at least two unique identifiers. The date and time provides additional information that may be necessary for interpretation of the results. Identification of the phlebotomist may be helpful if there are questions about the procedure or the specimens collected.
38. Use the requisition form or the labels to verify that all the necessary tubes were drawn for the specimens ordered.	Verifying the details of the blood draw once more in the presence of the patient allows for a redraw to be performed immediately if something is missing.
39. Observe any special handling instructions for the specimens.	Some types of specimens must remain at room temperature, whereas others may need to be put on ice immediately. This should be information that is ascertained before the procedure starts.
40. Monitor the patient for any signs of distress from the procedure.	These may include pallor, perspiration (especially on the upper lip or forehead), increased anxiety, or light-headedness. If the patient is exhibiting any of these symptoms, it is best to have him or her lie down if possible. This may be easily accomplished if the patient is in a chair that can recline. If not, a cold compress on the forehead and/or the back of the neck may help. Continue to converse with the patient and move the blood out of sight. Ask for assistance if you feel that your patient is feeling faint. If the patient loses consciousness, it may be necessary to lower him or her to the floor from the phlebotomy chair.
41. Once the tubes have been labeled, check the draw site for bleeding. If it has not stopped bleeding, apply pressure for another few minutes, and if the bleeding is still present, contact a physician.	It is important to look under the gauze for 2 or 3 seconds before applying the adhesive bandage to be certain that the site has actually stopped bleeding.
42. If the bleeding has subsided, apply an adhesive bandage, but leave the gauze in place to allow for additional pressure.	Self-adhesive bandages (such as Coban) may be wrapped around the site rather than applying an adhesive to the skin directly. Self-adhesive bandages may be especially effective for those patients who are on anticoagulant therapy.
43. Advise the patient to avoid heavy lifting or excessive exercise of the arm used for venipuncture for at least 1 hour.	Heavy lifting or exercise could cause the site to resume bleeding.

Continued

Procedure 8-3: Venipuncture Using the Butterfly (Winged Infusion) System—cont'd

Procedure	Rationale
44. Assist the patient to stand (if necessary) and thank him or her for being cooperative.	Patients may be a bit light-headed and assistance may be needed when he or she first stands up.
45. Discard all trash, and place the tubes in an appropriate vessel for processing. Never touch full tubes of blood without wearing gloves.	Touching the tubes of blood without wearing gloves offers potential opportunities for exposure to bloodborne pathogens.
46. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
47. Document the procedure in the patient's chart as well as on the requisition form. Include the date and time of collection, what was collected, and if there were any circumstances that were out of the ordinary. Also document where the vein was accessed (such as right arm, left hand, etc.). The documentation should include the type (color) of tubes drawn, as well as the identity of the person who collected the sample.	All patient interactions must be carefully documented. The site used for vein access should always be written in the chart or on the requisition form so that if there are any negative outcomes associated with the procedure, the site has been noted.

Date	
1/18/2020:	Phlebotomy performed in rigft hand for BMP and Hematocrit level. Lavender top tube drawn.
1100 a.m.	Connie Lieseke, CMA (AAMA)

Procedure 8-4: Blood Collection From a Capillary Puncture

Capillary punctures are performed frequently to obtain blood for CLIA-waived tests, as well as to draw blood for testing on children and infants. Blood obtained by capillary puncture is the preferred specimen type in these situations. In some situations in which it has been difficult to perform a successful venipuncture for adults, a capillary puncture specimen may also be obtained for testing.

TASK

Successfully perform a capillary puncture and obtain blood necessary for the tests ordered. The process must be completed within 5 minutes.

CONDITIONS

- Hand-washing supplies and/or alcohol-based hand sanitizer
- Disposable gloves
- 70% isopropyl alcohol
- Disposable safety equipped lancet
- Laboratory requisition form or labels with specified test
- Microcollection tubes
- 2 x 2 gauze pads
- Hand warmer (if necessary)
- Adhesive bandage or wrap
- Biohazardous sharps container
- Biohazardous disposal bag

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.3: Perform Capillary Puncture **I.A.I.1:** Apply critical thinking skills in performing patient assessment and care **III.P.III.3:** Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations

ABHES Standards

- Medical Office Laboratory Procedures: Collect, label and process specimens: Perform capillary puncture
- Medical Office Clinical Procedures: Apply principles of aseptic techniques and infection control
- Medical Office Clinical Procedures: Use Standard Precautions

Procedure	Rationale
1. Gather the requisition form and/or labels for the blood draw, and greet the patient. Identify yourself appropriately.	The requisition or labels are necessary to collect the correct type of specimen. It is always correct practice to identify yourself to the patient.
2. Wash hands (if they are visibly soiled) or apply hand sanitizer. Allow hands to dry completely.	Clean hands stop the spread of infection. Hands should be completely dry before attempting to apply gloves, or it will be difficult to put the gloves on the wet hands.
3. Verify the identification of the patient by asking for his or her name and at least one other unique identifier (such as the patient's birth date). Compare this information to the requisition or labels.	Patient identification must always be verified using at least two unique factors.
4. Verify whether dietary restrictions were followed, and time of last medication dose, if needed.	Many laboratory tests require fasting specimens or other dietary restrictions. Drug dosage times are especially important for appropriate interpretation of the laboratory results.
5. Have the patient sit in the phlebotomy chair with appropriate arm support, and extend his or her arm so that the hand may be accessed easily. Massage the fingertips if necessary for warmth, or apply a commercial warming device, hand warmer, or warm towel.	The patient should be secure and comfortable for the capillary puncture process. The fingers must be warm for a successful capillary blood draw; warming the site and massaging will allow much better blood flow. A commercial warming device works well, or immersing the finger or heel in warm water will also help. The temperature of the warming device or water should not exceed 42°C (108°F). Three to five minutes is generally sufficient to warm the site.
6. Assemble necessary equipment within reach. This includes the necessary tubes for the tests ordered, gauze, alcohol, and an adhesive bandage.	Once the incision is made, the process will go quickly, so it is important to have all supplies within reach.
7. Apply gloves.	Gloves are required for procedures in which there is reasonable anticipation of exposure to blood or other potentially infectious materials.

Continued

Procedure 8-4: Blood Collection From a Capillary Puncture—cont'd

Procedure	Rationale
8. Choose the appropriate finger for the blood draw, and disinfect the fingertip with an alcohol swab. Allow the alcohol to dry completely before performing the skin puncture. Do not fan or blow on the site to dry the alcohol.	For both adults and children, the ring finger or middle finger should be used. The index finger is more sensitive and more callused, and the little finger does not have enough flesh to protect the bone from puncture. The thumb should never be used as a capillary puncture site. The alcohol may contaminate the specimen if not allowed to dry completely before performing the puncture. Blowing or fanning the site may recontaminate the skin after cleansing.
9. Choose the correct lancet for the age of the patient and the site selected. The number of tubes to be drawn must also be taken into consideration when choosing a device to use for the incision.	Lancets come with different depths and widths. There are recommendations for ages and uses provided by the manufacturers. Lancet devices designed for home use that produce only a drop of blood do not provide enough blood to be used for microcollection tubes.
10. Remove the cap from the lancet device. Do not allow the surface to be placed against the patient's skin or to touch anything else before it is utilized.	This part of the device must remain sterile until use.
11. Perform the dermal puncture by holding the lancet device firmly against the skin and activating the device. Use the lateral side of the ring or middle fingertip, perpendicular (opposite) to the lines of the fingerprint. Immediately discard the lancet into a biohazardous sharps container.	The lateral sides of the fingers need to be used to avoid potential damage to the bone. The incision needs to be perpendicular to the lines of the fingerprint to keep the blood from following the fingerprint and flowing away from the incision site.
12. Wipe away the first drop of blood, then gently massage the finger to achieve blood flow.	The first drop of blood is contaminated with tissue fluid, and must be discarded. The finger should be continuously massaged from the proximal to the distal end of the fingertip. Do not squeeze right at the collection site as this can contaminate the specimen with tissue fluid and cause erroneous results.
13. Fill the required tubes in the correct order of draw. If anticoagulant is used in the tubes, tap them against the countertop to mix the specimen as the tube is filled. Do not touch the collection device against the incision while collecting the specimen. Instead, touch it to the drop of blood as it forms at the collection site.	The capillary order of draw must be followed to avoid cross-contamination. The tubes must be mixed well during the collection process to avoid clotting. The collection device is to be used only to collect blood that is free flowing; scraping or touching the incision site can cause infection and/or irritation.
14. If a microhematocrit tube is to be filled, hold it horizontal to the site to avoid introduction of air bubbles to the specimen.	Capillary tubes that are held at a slant allow air to enter the microhematocrit tube. The capillary tube will fill with capillary action if held horizontal to the incision site. Plug the end of the capillary tube when filled appropriately.

Procedure	Rationale
15. When the desired tubes have been collected, apply gauze to the puncture site and instruct the patient to apply direct pressure, if he or she is capable.	Direct pressure helps the bleeding to stop.
16. Tightly cap and invert any microcollection tubes containing anticoagulant 8 to 10 times.	Tubes must be thoroughly mixed to avoid clotting.
17. Label the tubes with the patient's name, birth date (or other unique identification number), your initials, the date and the time of the blood draw.	For microcollection tubes, it may be necessary to write this information on a label to be attached to the tube. The microcollection tubes may also be placed inside larger tubes that are labeled appropriately. Capillary tubes may also be labeled in this way, as it is very difficult to label the actual collection container.
18. Use the requisition form or the labels to verify that all the necessary tubes were drawn for the specimens ordered.	Verifying the details of the blood draw once more in the presence of the patient allows for a redraw to be performed immediately if something is missing.
19. Observe any special handling instructions for the specimens.	Some types of specimens must remain at room temperature, whereas others may need to be put on ice immediately. This should be information that is ascertained before the procedure starts.
20. Monitor the patient for any signs of distress from the procedure.	These may include pallor, perspiration (especially on the upper lip or forehead), increased anxiety, or lightheadedness. If the patient is exhibiting any of these symptoms, it is best to have him or her lie down if possible. This may be easily accomplished if the patient is in a chair that can recline. If not, a cold compress on the forehead and/or the back of the neck may help. Continue to converse with the patient and move the blood out of sight. Ask for assistance if you feel that your patient is feeling faint. If the patient loses consciousness, it may be necessary to lower him or her to the floor from the phlebotomy chair.
21. Once the tubes have been labeled, check the draw site for bleeding. If it has not stopped bleeding, apply pressure for another few minutes, and if the bleeding is still present after 5 minutes, contact a physician.	It is important to look under the gauze for 2 or 3 seconds before applying the adhesive bandage to be certain that the site has actually stopped bleeding.
22. If the bleeding has subsided, apply an adhesive bandage, but leave the gauze in place to allow for additional pressure.	Adhesive bandages should not be applied for capillary punctures on small children as they may pose a choking hazard. Newborns may have adhesive bandages applied if they are not able to remove them. Self-adhesive bandages (such as Coban) may be wrapped around the site rather than applying an adhesive directly to the skin. Self-adhesive bandages may be especially effective for those patients who are on anticoagulant therapy.

Continued

Procedure 8-4: Blood Collection From a Capillary Puncture—cont'd

Procedure	Rationale
23. Assist the patient to stand (if necessary) and thank them for their cooperation.	Patients may be a bit light-headed, and assistance may be needed when they first stand up.
24. Discard all trash, and place the tubes in an appropriate vessel for processing. Never touch full tubes of blood without wearing gloves.	Touching the tubes of blood without wearing gloves offers potential opportunities for exposure to bloodborne pathogens.
25. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
26. Document the procedure in the patient's chart as well as on the requisition form. Include the date and time of collection, what was collected, and if there were any circumstances that were out of the ordinary. Also document where the specimen was obtained (such as right middle finger, left hand, etc.). The documentation should include the type (color) of tubes drawn, as well as the identity of the person who collected the sample.	All patient interactions must be carefully documented. The site where the specimen was obtained should always be written in the chart or on the requisition form so that if there are any negative outcomes associated with the procedure, the site has been noted.

Date	
1/18/2020:	*Capillary puncture performed in right ring finger for CBC. Lavender top microcollection tube drawn.*
1100 a.m.	*Connie Lieseke, CMA (AAMA)*

the analysis. It is the job of the medical assistant to process the specimen properly so that it can be tested.

Obtaining Serum for Testing

When blood is allowed to clot and is then spun down in a centrifuge, the liquid portion is called **serum.** Serum is plasma that no longer has the clotting factors included, as they have been used up in the blood clot that formed in the tube. Analytes such as glucose, lipids, cholesterol, electrolytes, hormones, enzymes, and antibodies may be dissolved in serum. To isolate this liquid so that the tests can be performed, it is necessary to separate it from the rest of the blood specimen. Whenever drawing blood for serum or plasma, it is necessary to collect approximately 2.5 times as much blood as the volume needed for the testing procedure. For instance, if there is 1 mL of serum required for an electrolyte test, the medical assistant drawing the blood should obtain at least 2.5 mL of blood.

Tubes without anticoagulants are used to collect samples for serum testing. These include red top tubes, as well as those that contain thixotropic gel and clot activators. The blood must be allowed to stand for at least 30 to 45 minutes at room temperature to clot thoroughly before the specimen can be further processed. If the specimen is spun in the centrifuge before there is a chance for a solid clot to form, the clotting factors will not be mixed in with the cells at the bottom of the tube when centrifuged; instead, they will form a large fibrin clot in the serum layer. A *fibrin clot* is a soft, sticky mass that makes it very difficult to separate out serum for testing. Although the tube must be allowed to clot completely before centrifuging, it should not be more than an hour after the blood draw is performed before the tube is spun, as prolonged contact with the cells in the tube may allow chemical changes to take place in the serum, which will affect the test results. Potential changes may include a decreased serum glucose level, an increased serum iron level, and elevated serum potassium levels, among others.

To fully separate the serum from the cells, the specimen should be centrifuged for at least 10 minutes. The serum may then be removed from the specimen and placed in a transfer tube. There are various methods that may be used to remove the liquid portion of the blood; remember that regardless of the method used for separation, adequate personal protective equipment should be worn at all times, including a face shield or plastic-mounted barrier shield, gloves and a laboratory coat. (See Fig. 8-15.) The methods that may be used for separation include the following:

- **SST or PST tubes:** These tubes contain a thixotropic gel that forms a barrier between the cells and the serum (or plasma with PST tubes) so that the liquid can remain in the tube, but be separated from the cells. If it is necessary to remove the serum and put it in a transfer tube, it can be poured out because the gel forms a solid barrier to the cells in the bottom of the tube.
- **Transfer pipette:** A pipette may be used to aspirate the liquid out of the tube and transfer it into another tube. Care must be taken not to aspirate out any of the cells that are present in the bottom of the tube; if the serum appears a hazy red color while aspiration is taking place, it should be spun again to remove any cells that may have been accidentally aspirated into the specimen. If the serum still appears to be red after recentrifugation, it is hemolyzed and the specimen will, in most circumstances, have to be redrawn. If the serum is no longer red, separate the serum again from the cells that have settled to the bottom of the tube.
- **Plunger-type separators:** These are devices that have a filter on one end of a plastic tube, with an opening at the other end. After the specimen has been centrifuged, the rubber stopper is removed from the top of the tube and the filter end of the separator is carefully inserted into the specimen tube and pushed down through the serum or plasma. Proper protective equipment must be used when performing this procedure to keep from potentially splashing the liquid into the eyes or mucous membranes. Also, once the separator has been place in the tube, it is important to pull it back up a bit to provide an air barrier between the serum and the cells. This keeps the blood cells in the bottom of the tube from being in contact with the serum, causing chemical changes as they metabolize nutrients in the liquid.

Test Your Knowledge 8-16

How are specimens processed if serum is to be separated from the cells? (Outcome 8-17)

Specimen storage instructions for most chemistry tests will advise that the serum and/or plasma be refrigerated within a few hours after processing to protect the various analytes from changing in concentration. Follow the directions provided in the laboratory directory for handling the serum specimen after it has been separated. Also, if a transfer tube is used, the labeling on the tube is critical. Not only does the patient information need to be included, but there also has to be a notation of the

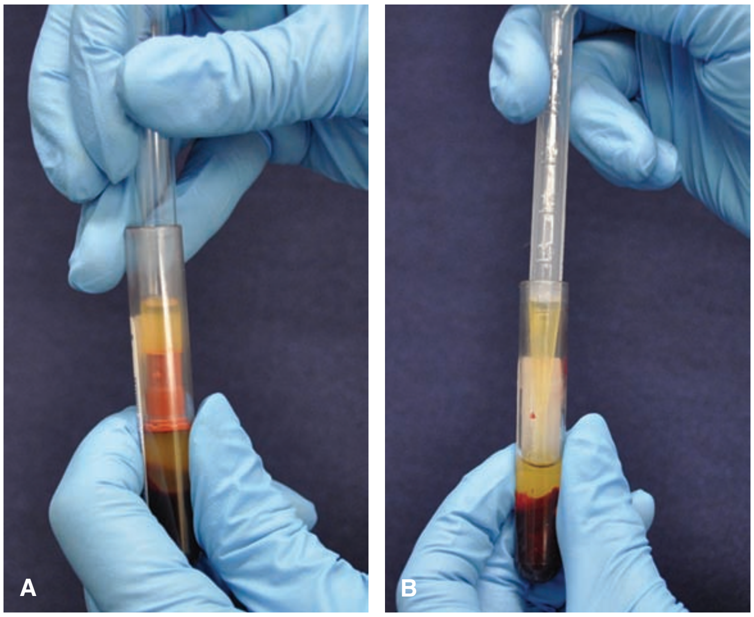

Figure 8-15. A and B. (A) Serum tube after centrifugation and (B) various devices used for separation of serum.

type of specimen (serum or plasma) and the type of anticoagulant present in the tube, because many body fluids are similar in appearance. Transfer tubes are available that are color coded to match the color tube originally used for the specimen collection, which may help eliminate confusion.

Obtaining Plasma for Testing

The liquid portion of the blood in our bodies is **plasma.** Plasma is made up of approximately 90% water, with dissolved substances making up the remaining portion. Plasma may be analyzed for levels of chemicals involved in the clotting process such as fibrinogen and prothrombin. Other common tests performed on plasma specimens include levels of electrolytes, calcium, glucose, and creatinine.

Because plasma is the liquid portion of the blood that contains factors that contribute to clotting, a specimen that is to be used for plasma testing must not be allowed to clot. A tube that contains an anticoagulant (such as heparin) must be used for the collection. The medical assistant should consult the laboratory directory to see what type of anticoagulant is to be used for specimen collection before the venipuncture is performed. It may be possible to use a PST tube, which contains anticoagulant and a thixotropic gel that will separate the plasma from the blood cells in the specimen after centrifugation. As in the case of serum, it is necessary to collect a blood specimen that is approximately 2.5 times the required volume for the test ordered. Appropriate labeling of the transfer tube is essential; remember to include a notation that the fluid is plasma in addition to the patient's name and other necessary information.

Unlike the process for obtaining a serum sample, plasma samples should be well mixed, then centrifuged as soon as possible. There is no need to allow the specimen to sit for an extended period of time before centrifuging, as the blood is not going to clot in the tube. This makes plasma samples the specimen of choice for most STAT chemistry tests, because the samples can be processed and the test performed quickly after collection.

After centrifugation, a sample that has had anticoagulant added will separate into three layers. This will include a layer containing the red blood cells, topped by a very small layer that contains the white cells and platelets (sometimes called the buffy coat), with the liquid plasma present as the top layer in the tube. The tube in Figure 8-16 is an example of the appearance of a tube to be used for plasma separation after centrifugation.

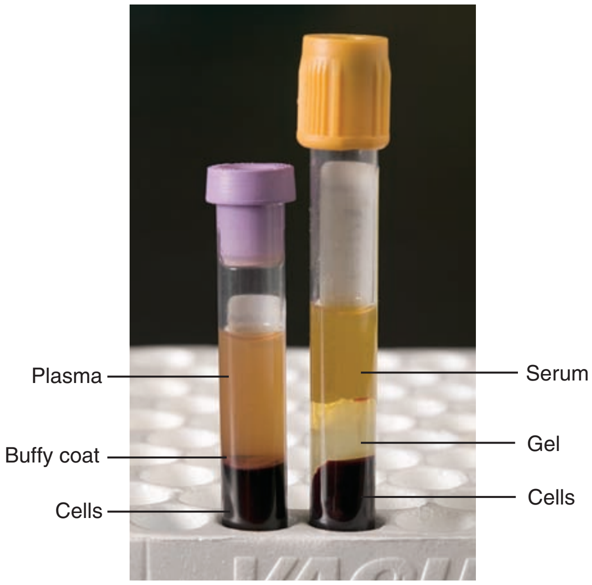

Figure 8-16 The tube on the left shows plasma after centrifugation. Note the cells, buffy coat, and plasma. The tube on the right shows serum separated from the clotted cell by gel. *Reprinted with permission from Eagle S, Brassington C, Dailey C, and Goretti C:* The Professional Medical Assistant: An Integrative, Teamwork-Based Approach. *Philadelphia: FA Davis, 2009.*

The actual separation methods for plasma are the same as those used for serum, including the use of a PST tube, the transfer pipette, or the plunger-type separators. Remember, whenever separating the liquid portion of the blood from the cells, it is imperative that the appropriate personal protective equipment is used to protect the employee from potential exposure. Removal of the rubberized stoppers may create an aerosol that could get into the mouth or eyes, so face shields must be worn in addition to gloves and a protective laboratory coat. Also, all supplies must be disposed of appropriately; the stoppers and any other specimen containers must be disposed of as biohazardous materials.

Test Your Knowledge 8-17

Describe one way that serum and plasma are different.
(Outcome 8-18)

Whole Blood Specimens

Tests that count or examine the cells present in the blood require a whole blood specimen. These tests are often performed in the hematology department, and include complete blood counts, platelet counts, and hemoglobin and hematocrit tests. For tests requiring whole blood specimens, the blood is drawn into tubes that contain anticoagulants so that the cells are not involved in clot formation, which would make it very difficult to count the cells or examine their appearance. Potassium EDTA is the anticoagulant that is usually preferred; this is present in the lavender top evacuated tubes.

Whole blood specimens are to be well mixed at the time of the initial blood draw, and again just prior to analysis. The specimens used for whole blood testing are not to be centrifuged, as the analysis is performed on the formed elements within the sample. In some small physician office laboratories, the medical assistant may place the whole blood specimens directly on a "rocker" that keeps the specimen mixed until the analysis can be performed. Remember that the mixing and inversion of these samples must be a gentle motion to avoid damaging the cells. Eight inversions of the tube immediately after drawing the blood should provide appropriate mixing of the specimen.

Test Your Knowledge 8-18

True or False: Tests that utilize whole blood specimens require that the specimen be centrifuged before testing.

(Outcome 8-17)

Unacceptable Specimen Types

In certain situations, the specimen collected and processed will be rejected for the test ordered. There are numerous reasons for specimen rejection, including the following:

- **Hemolyzed specimens:** *Hemolysis* means that the red blood cells in the specimen have been damaged and broken. It may be a result of a traumatic venipuncture in which the cells were damaged as they entered the needle, or hemolysis may be the result of mishandling the tube after the blood draw. Hemolysis is evident in the specimen after centrifuging by the presence of a pink to red tint in the plasma or serum. (Fig. 8-17 is an example of a hemolyzed specimen.) Potassium, magnesium, and iron levels are examples of tests for which a hemolyzed specimen is unacceptable. To avoid hemolysis, be sure that the tubes used for the blood draw are kept at room temperature, and be sure to use the appropriate sized needle for the draw. Also, if using a syringe, do not pull back on the plunger with a great deal of force, as this may damage the cells. Use good technique when performing venipunctures, and *gently* invert all tubes when mixing.
- **Lipemic specimens: Lipemia** is the presence of excess lipids (fatty molecules) in the blood. Plasma or serum in a lipemic specimen will appear cloudy or milk-like after centrifugation. These lipid molecules interfere with the testing methods for many analytes. Some laboratories are capable of clearing the specimen with a special type of centrifuge, whereas others will reject the lipemic specimens (Fig. 8-17).
- **Quantity not sufficient**: When the medical assistant does not draw enough blood to perform the tests ordered, the specimen may be rejected as **quality not sufficient (QNS)** because the amount drawn is not enough for the tests to be performed on it. In almost all situations, these samples will need to be redrawn so that there is enough specimen to complete the tests ordered.
- **Clotted specimens:** When a whole blood specimen is necessary for the test ordered, a clotted specimen is unacceptable. The clotting process draws in the cells in the specimen to be involved in the clot. This means that even if the clot is small, the cell count in the tube will be inaccurate, because it is impossible to know how many cells are involved in the clot and how many are floating freely in the specimen. To avoid clotted specimens, be sure to invert the specimens thoroughly during the collection process.
- **Incorrect anticoagulant use:** Each anticoagulant uses a different principle to keep the blood from clotting. Some bind up the calcium in the specimen, as it is necessary for the clotting process to proceed. Others make the platelets in the specimen nonadhesive so that they cannot cling to one another. Tests are designed to be performed using a specific anticoagulant, and if the incorrect one is used, it may alter the test results. For instance, potassium EDTA may cause false elevation of the potassium levels if a lavender top tube was used

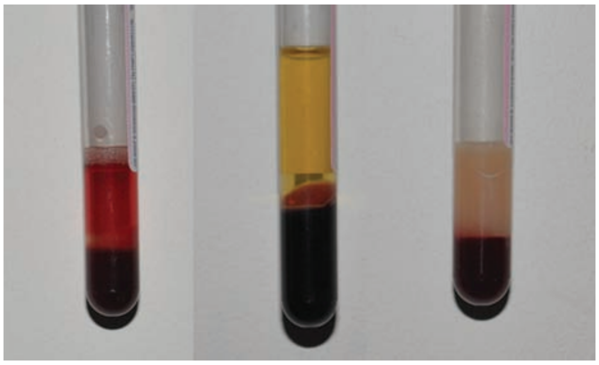

Figure 8-17 Hemolysis and lipemia present in plasma tubes on left and right.

for the plasma sample. Sodium citrate is an anticoagulant that binds up the calcium in the specimen so that it cannot participate in the clotting process. This means that calcium levels performed on plasma from these light blue top tubes would be very low, as the calcium is not in solution in a way that it can be tested.

- **Fibrin clots:** Tubes without added anticoagulant must be allowed to clot properly before centrifugation, or it may be impossible to obtain enough serum from the specimen to perform the tests ordered. This may result in a request for a specimen to be recollected. It may be possible to physically remove the fibrin clot and recentrifuge the specimen, but this may lead to damaged red cells and hemolysis.

Test Your Knowledge 8-19

Examples of inappropriate specimens may include:

a. Hemolyzed and lipemic specimens
b. Partially filled specimen tubes for multiple tests
c. Blood collected in tubes with the wrong anticoagulant
d. All of the above (Outcome 8-19)

POTENTIAL NEGATIVE OUTCOMES OF VENIPUNCTURE AND CAPILLARY PUNCTURE

Regardless of the skill level exhibited by the medical assistant, sometimes the person who is attempting a venipuncture may be unsuccessful. In addition, even when the person drawing the blood is successful, there may be physical patient complications resulting from the blood draw. It is important to realize that these negative outcomes are sometimes unavoidable, and to know what action should be taken if they occur.

Inability to Draw Blood

Veins are not solid objects that are incapable of movement. Sometimes even when the vein is anchored tightly, it will move just a bit as the needle is inserted. Or, the health-care worker who is drawing the blood may not insert the needle far enough or go in just a bit too far so that the blood does not enter the needle. It is never acceptable to "probe" when drawing blood. However, recommendations from the Clinical and Laboratory Standards Institute do allow the phlebotomist to move the needle a bit further into the vein or a bit further out of the vein to see if the blood will start to flow. Sometimes a slight change in the location or angle of the needle is all that is necessary for the blood to enter the needle. The needle should never be moved from side to side after insertion into the arm, as this will cause damage to the tissues and pain for the patient. If the slight movements (small increments in or out) are not enough to allow blood to enter the needle, discontinue the draw, and try again. Be sure to check for alternative sites before using the same area for a second attempt. If the person drawing the blood is still unsuccessful after two attempts, he or she should seek assistance from another qualified employee.

Fainting Patients

Some patients experience light-headedness, dizziness, or fainting during or immediately after a blood draw. This may be due to **vasovagal syncope**, which is the body's exaggerated reaction to the sight of blood. Some of these patients may have had negative experiences in the past, whereas others cannot identify a specific reason for their reaction. Each person who experiences vasovagal syncope has his or her own triggers, which may include emotional distress, the sight of blood, and/or pain. The trigger causes a response in the body that includes a drop in blood pressure and a decreased heart rate. Young patients, thin patients, nervous individuals, and those who are very quiet (or sometimes *very* talkative) are more prone to fainting. Also, hunger, fatigue, and environmental factors such as excessive heat or strong smells may make the situation worse. There are usually some symptoms that occur prior to the actual fainting episode such as nausea, yawning, dizziness, weakness, perspiring, pallor, and a flushed feeling of warmth. If the patient communicates any of these symptoms, the blood draw should be discontinued. The medical assistant should help the patient to put the head down between the legs if he or she is still conscious. This may help to keep the patient from fainting. If the patient shares a history of fainting, he or she should be drawn in the supine position, and the medical assistant should make additional efforts to talk to the patient during the procedure to monitor the level of consciousness. If a patient does faint when sitting up in a phlebotomy chair, the primary focus should be safety for the patient and the medical assistant. Discontinue the draw, remove the tourniquet and the needle, activate the safety device on the needle, and bandage the draw site. Call for assistance from a coworker. If the patient is still unconscious, it may be necessary to lower him or her to the floor from the chair, with special care taken to protect the head from hitting anything during the process. The patient's legs should be elevated to help the blood flow return to the heart and brain, and a health-care provider should be notified of the situation.

Rolling Veins

The median cubital vein is usually quite stable, which makes it an excellent choice for venipuncture. If this vein cannot be used for phlebotomy procedures, the other veins in the antecubital area may be considered. However, these are not as well anchored, and have a tendency to "roll" or move away when the venipuncture is attempted. Special care must be taken to anchor these veins very well to avoid this problem.

Hematoma Formation

A hematoma is the result of blood leaking into the tissues from a vein. It can occur in routine venipunctures, but is more common when the process did not go smoothly. If the initial insertion of the needle is too deep, the needle may puncture and go through the vein, which allows excessive blood to leak from the punctures in the vein. Conversely, if the needle does not go far enough into the vein, the bevel may not be totally within the vessel, and blood may leak out around the slant at the end of the needle. (See Fig. 8-18 for examples.) Hematomas can also be the result of too little pressure applied after the venipuncture procedure, with subsequent bleeding around the puncture site. When a hematoma forms, there is a sudden swelling (and sometimes a discoloration) around the site where the needle is inserted. If this occurs, the tourniquet and then the needle need to be removed and pressure should be applied immediately. The medical assistant should be sure to keep pressure on the site for at least 5 minutes, and also should be sure to document the situation. The patient may experience pain and more swelling in the area. The healthcare provider may suggest ice application and anti-inflammatories to help with the discomfort.

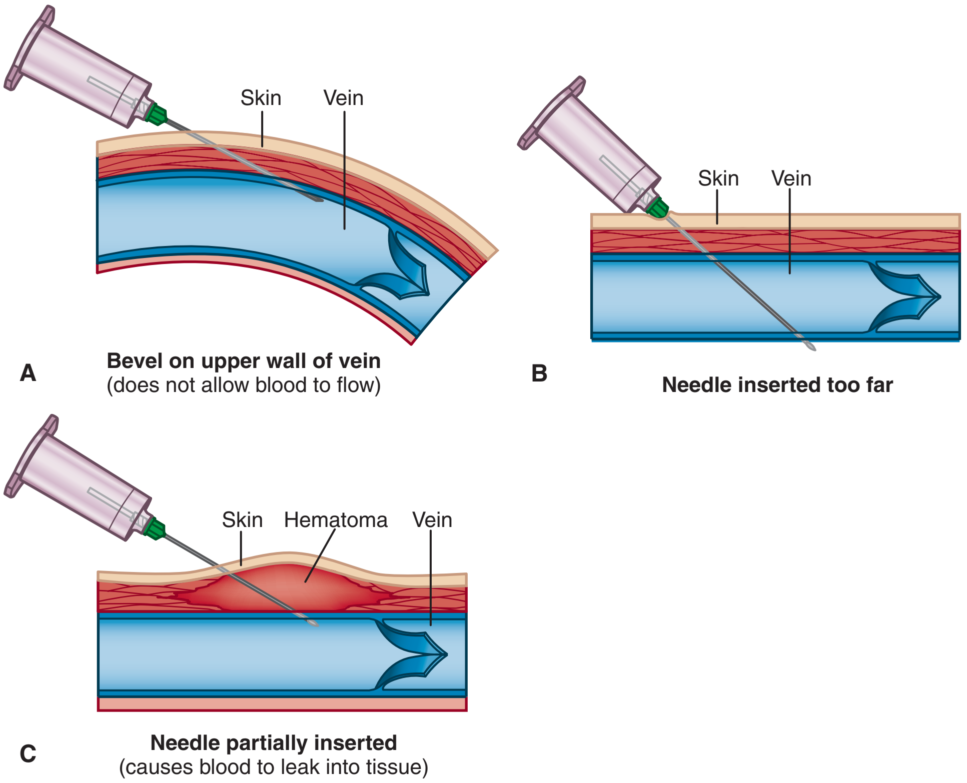
Skin
Vein
A
Bevel on upper wall of vein
(does not allow blood to flow)
Skin
Vein
B
Needle inserted too far
Skin
Hematoma
Vein
C
Needle partially inserted
(causes blood to leak into tissue)

Figure 8-18 Problems with the way in which the needle is entering the vessel. (A) The first drawing shows what occurs if the angle of insertion is too shallow and the needle goes above the vein. (B) The second drawing shows what occurs if the angle of insertion is too high, and the needle goes through the vein. (C) The third shows what can happen if the needle is not inserted far enough into the vein. *Reprinted with permission from Strasinger S, and Di Lorenzo M:* Phlebotomy Textbook, *ed. 3. Philadelphia: FA Davis, 2011.*

Collapsing Veins

Fragile veins may collapse with the amount of pressure used to withdraw blood during a venipuncture procedure. Small veins or veins that have been damaged with IV therapy or medication treatment are more prone to collapse. If the medical assistant recognizes that the vein is small or appears fragile, the evacuated tube system should not be used for the venipuncture, as this will increase the likelihood of venous collapse. A syringe and needle or a butterfly system should be used instead.

Nerve Damage

Venipuncture procedures performed in the median cubital vein rarely cause irritation to the nerves in the arm, as this vein is not usually in close proximity to the medial nerve. When drawing from the antecubital area, blood draws performed from the basilic vein are the most likely to cause nerve irritation or damage due to the close proximity to the nerve. If a patient complains of excessive pain during a blood draw (especially if he or she describes it as shooting pain that goes up or down the arm), the procedure should be discontinued immediately. Never attempt a venipuncture from the underside of the wrist, as there are numerous opportunities to cause nerve irritation or damage in this area.

Infection

Infection or excessive irritation at the site of a venipuncture is not common, but can occur. To minimize the risk, the medical assistant should always clean the site thoroughly before the procedure, and also allow the alcohol used to disinfect the site to dry completely before the needle is inserted. If the alcohol has not been allowed to dry, it can cause irritation to the skin in the area, which makes the site more prone to infection. Using good judgment when choosing a site for venipuncture will also help to minimize the chances of inflammation and infection.

Test Your Knowledge 8-20

True or False: The angle at which the needle is inserted for venipuncture has no impact on whether the procedure has a negative outcome. (Outcome 8-20)

OTHER PROCESSING PROCEDURES

Processing a blood specimen may include centrifugation, separation of plasma or serum from the cells, and appropriate storage of the specimen until testing occurs. In addition, the medical assistant may be asked to create and stain a smear from a whole blood specimen so that the healthcare provider or other qualified professional may view it.

PREPARATION OF A PERIPHERAL BLOOD SMEAR FOR STAINING

When it is necessary to view the red blood cell **morphology** (appearance and size), identify the types and percentages of various types of white blood cells, and quantitate the number of platelets present in the circulation, a peripheral blood smear may be utilized. Automation has replaced the need to perform a manual examination of the slide for most hematology testing procedures, but there is still a need to examine the slide manually in many

Procedure 8-5: Creation of Peripheral Blood Smear

A peripheral blood smear may be requested at the time of the initial blood draw, or it can be created using blood from a lavender top tube containing EDTA anticoagulant. This procedure will explain how to create a smear using the lavender top tube. If creating a smear at the time of the blood draw, the procedure only differs with the initial application of the sample to the slide.

TASK

Successfully create a peripheral blood smear to be stained and used for a manual differential count or pathologist examination. The process must be completed within 5 minutes.

CONDITIONS

- Hand-washing supplies and/or alcohol-based hand sanitizer
- Disposable gloves
- Clean glass slides
- DIFF-SAFE device
- Lavender top (EDTA) tube filled with blood
- Biohazardous sharps container
- Biohazardous disposal bag

CAAHEP/ABHES STANDARDS

CAAHEP Standards

III.P.2: Practice Standard Precautions

ABHES Standards

None

Procedure	Rationale
1. Sanitize hands (allow them to dry) and apply gloves.	Gloves must be worn for any procedures in which exposure to blood or other potentially infectious materials is anticipated.
2. Invert the blood tube 8 to 10 times to mix the sample.	The specimen must be well mixed or it will provide erroneous results when the blood smear is examined.
3. Insert DIFF-SAFE device into the top of the lavender top tube.	The device has a blunt metal cannula that is inserted through the rubber stopper on top of the tube. This is performed while the tube is in an upright position. 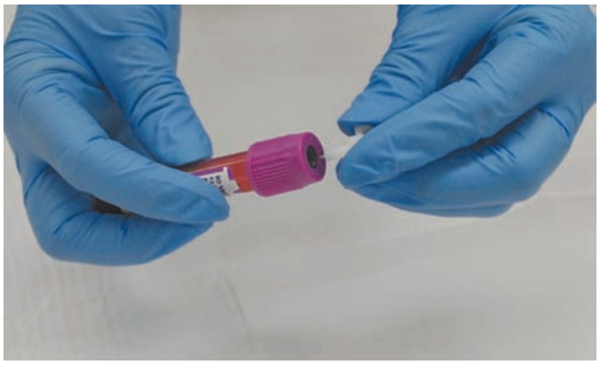
4. Turn the tube upside down and push the DIFF-SAFE device to the slide until a drop of blood is released.	Pushing down on the tube with the DIFF-SAFE device against the slide will release a drop of blood onto the slide. 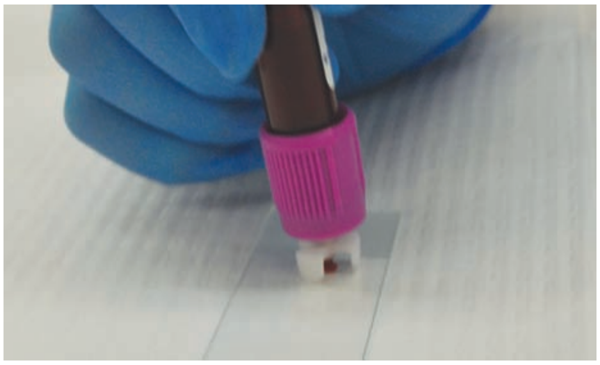

Continued

Procedure 8-5: Creation of Peripheral Blood Smear—cont'd

Procedure	Rationale
5. Pull the second slide back into the drop of blood at an approximate 30-degree angle, allowing the blood to flow along the edge of the slide.	This process must occur quickly. The angle allows for correct application of the blood across the slide. 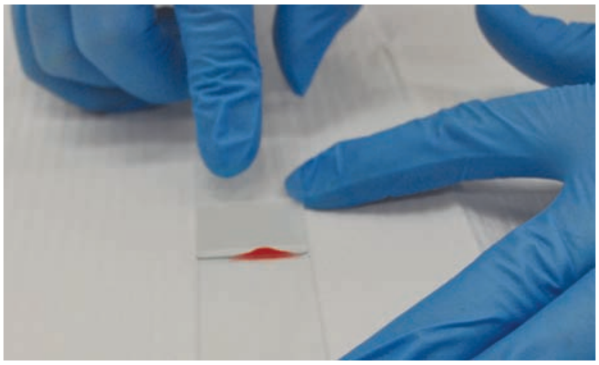
6. Once the blood has flowed to cover approximately three-quarters of the width of the slide, push the second slide forward to spread the blood across the original slide.	Push the spreader slide forward smoothly with a rapid motion. Do not apply weight or additional pressure during this step, or the slide will not move smoothly across the surface. The entire process must occur in less than 15 seconds or the drop of blood will begin to dry and the distribution of the cells will be uneven across the slide. Smears should have a "feathered" edge, with visible edges on the heaviest part of the smear. Ideally there will be no holes or ridges in the blood pattern. If the cell distribution is too thick, it will be difficult to see the morphology of the red blood cells clearly, and it may be difficult to visualize the white blood cells as well. The completed smear should have an appearance that is similar to a thumbprint, with a heavy distribution at the beginning of the smear and a gradual decrease in the thickness across the slide. 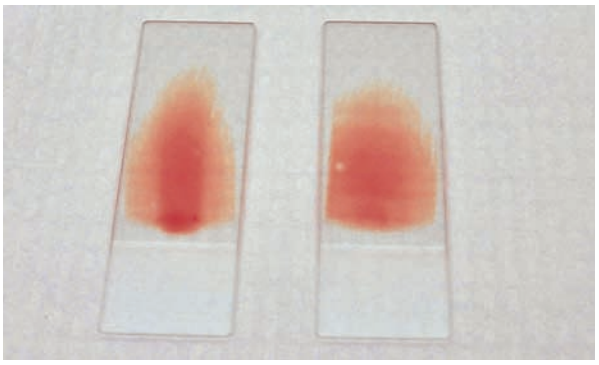

Procedure	Rationale
7. Allow the slide to air-dry.	The slide must be completely dry before it is stained or the appearance of the cells will be altered during the staining process.
8. Remove the DIFF-SAFE device from the top of the tube and discard the device in a biohazardous sharps container. Also discard the second slide used for spreading in the biohazardous sharps container. Store the tube as directed by the laboratory.	The device and the slide are contaminated with blood and are considered to be a sharp.
9. Label the slide with a pencil or a grease pen with the name of the patient and other identification as required by the laboratory.	All slides need to be properly labeled before staining.
10. Transfer the slide to the appropriate area of the laboratory for drying and staining.	The medical assistant should not touch the slide without gloves on, so the transport should occur before the gloves are removed.
11. Remove gloves and sanitize the hands.	Hands must always be sanitized after removing gloves.
12. Document the procedure in the patient's chart if it is available.	In a physician office laboratory, this procedure should be documented in the patient's chart. If the medical assistant is working in a laboratory setting where the chart is not available, this step is not necessary.

Date	
1/18/2020:	Blood smear prepared from lavender top tube.
1100 a.m.	Connie Lieseke, CMA (AAMA)

instances. A lavender top EDTA tube may be used or, in some cases, a drop of blood may be taken directly from a capillary puncture and applied to the slide. The safest method of applying the drop of blood to the slide from the tube is to use a DIFF-SAFE device (Alpha Scientific Corporation, Malvern, Pennsylvania), which eliminates the need to open the tube to gain access to the blood.

WRIGHT'S STAIN PROCEDURE

Once the peripheral blood smear has been prepared, it must be stained so that it can be examined. There are various products available for this staining process, including Diff-Quik stain by Dade Behring. Wright's stain is another common stain often used in the physician office laboratory as well as in larger laboratories. It is a hematology stain that is used for staining smears made from blood and bone marrow samples. Wright's stain allows for visualization of the red blood cells and platelets in the sample, as well as differentiation of the various types of white cells that are present. The stain contains methanol, which acts as a fixative for the cells, as well as an acidic red dye and a blue dye that is alkaline in pH. These stains are absorbed differently by the formed elements in the blood to allow for them to be examined and counted.

Procedure 8-6: Quick Stain of Peripheral Smear Using Camco Quik Stain II

The Camco Quik Stain procedure allows for the cells to be stained appropriately so that they can be identified by an individual who is qualified to perform this microscopic evaluation. This may be performed for a differential examination, in which a technician or technologist will count the various types of white cells present and report a percentage of each type. The stain solution may be poured into small vessels for dipping the slides, or it may be applied with a pipette to a slide on a rack that is suspended over a sink or other receptacle.

TASK

Stain a peripheral blood smear using Camco Quik Stain II.

CONDITIONS

- Air-dried blood smear
- Camco Quik Stain II solution
- Distilled water with a pH of 6 or 7
- Laboratory wipes
- Sink or other receptacle for excess stain

CAAHEP/ABHES STANDARDS

None

Procedure	Rationale
1. Sanitize hands, and allow them to dry completely. Apply gloves.	Gloves must be worn for any procedures in which exposure to blood or other potentially infectious materials is anticipated. In addition, the stain will discolor the fingers if gloves are not utilized.
2. Dip the blood smear in the stain solution for 10 seconds. Alternatively, the smear may be placed on a rack and stain may be added until the slide is covered with the stain for 10 seconds.	The 10 seconds is a minimum staining time. If the smear is thick, additional staining time may be necessary.
3. Allow excess stain to drain from the slide. Apply or dip in distilled water for 20 seconds.	It is necessary to allow at least 20 seconds for the process to be complete.
4. Wipe away any excess stain that may be present on the back of the slide with a laboratory wipe.	Excess stain on the back of the slide will obscure the view when examining the smear under the microscope.
5. Allow the slide to air-dry. Make sure that there is no excess water standing on the slide.	This type of smear does not need to be heat fixed. Excess water may change the staining characteristics or increase the time necessary for the slide to dry.
6. Dispose of the laboratory wipe in the trash and put away supplies.	All work areas should remain clean and organized.
7. Remove gloves and sanitize hands.	Gloves should be removed before proceeding to the next task. Hands must always be sanitized after removing gloves.

SUMMARY

Blood collection is the most common and most important part of the preanalytical process for laboratory testing. To perform quality venipuncture and capillary puncture, it is imperative that the medical assistant or phlebotomist understands the basics of the anatomy and physiology of the cardiovascular system as well as areas and situations to avoid when collecting blood. Following recommended procedures during the collection process is critical, which includes knowledge of how to safely use the various collection devices.

Patient identification and appropriate preparation are of utmost importance. Knowledge of the various tube types to be used is very important in collecting the sample correctly. Order of draw is critical to avoid crossover contamination. Finally, processing the samples appropriately is the final preparatory step before testing the sample. This may include centrifugation and separation of the plasma or serum, or it may include preparation and staining of slides.

TIME TO REVIEW

1. The bevel of a needle is: Outcome 8-1
   - **a.** The interior hollow space of the needle
   - **b.** The slanted tip at the end of the needle
   - **c.** The diameter of the needle
   - **d.** None of the above
2. The flanges of an evacuated tube holder: Outcome 8-1
   - **a.** Allow for the needle to be screwed on the holder
   - **b.** Provide a space for the tube to be inserted
   - **c.** May be used to provide leverage when inserting and/or removing tubes
   - **d.** Are decorative only
3. True or False: Interstitial fluid is the fluid located between the cells of the tissues. Outcome 8-1
4. True or False: Venules are larger in diameter than are veins. Outcome 8-1
5. Identify all the ways that veins and arteries are the same: Outcome 8-3
   - **a.** Transport blood
   - **b.** Connect to capillaries
   - **c.** May be punctured to withdraw blood for testing
   - **d.** Have multiple layers of cells in the walls of the vessels
   - **e.** Have a distinct pulse that can be felt with palpation
   - **f.** Have valves that prevent backflow
   - **g.** Are varied in size throughout the body
6. Mark the veins below in order of preference for venipuncture procedures by numbering them 1, 2, and 3: Outcome 8-5

   _______ Cephalic vein
   _______ Basilic vein
   _______ Median cubital vein
7. True or False: Capillary punctures on infants may be performed on the great toe. Outcome 8-6
8. An example of an area to avoid when choosing a site for venipuncture is: Outcome 8-7
   - **a.** Scarred skin
   - **b.** Varicose veins
   - **c.** Veins on the back of the hand
   - **d.** None of the above
   - **e.** a and b
9. How is a butterfly (or winged infusion set) different from the needle used for evacuated tube draws or syringe draws? Outcome 8-8
10. Using the following order of draw, create a mnemonic that will help you to remember the order the tubes are to be drawn with an evacuated tube system or a syringe draw. Outcome 8-9

    Yellow, light blue, red, green, lavender, gray
11. Describe one way that a medical assistant might put a child at ease before a blood collection is to be performed. Outcome 8-12
12. Why is it important to verify whether a patient has properly prepared for a test before the medical assistant performs the blood draw? Outcome 8-14
13. Why would a peripheral smear be used? Outcome 8-16
    - **a.** To test for certain hormones
    - **b.** To scan for bacteria
    - **c.** To identify various types of white blood cells
    - **d.** To perform ABO blood typing

14. What is one practice that can help to protect patients and employees during invasive procedures? Outcome 8-15

15. Capillary specimens may be preferred: Outcome 8-13
    a. When performing CLIA-waived tests that use whole blood
    b. When the patient's medical condition warrants it
    c. For infants
    d. All of the above

16. Which part of the hand is used to hold the evacuated tube holder when the medical assistant performs a venipuncture? Outcome 8-15

17. Do all capillary puncture devices provide incisions that are the same depth? Outcome 8-15

Case Study 8-1: How long was this sitting here?

A blood specimen for a glucose level and electrolyte analysis was drawn on Mr. Dee at 9:30 a.m. Shortly after the blood draw, Shannon, the medical assistant who was covering the laboratory area of the office, was asked to fill in for another employee who became ill at work. At 2 p.m., after lunch, Shannon went back into the laboratory and processed the specimens that she had drawn that morning so that they were ready to be picked up by the laboratory courier at 2:30.

Just before closing, Shannon received a call from the laboratory alerting the health-care provider that the potassium was elevated above the reference range for Mr. Dee, and the glucose levels were well below the reference range. The laboratory suggested that the specimen needed to be redrawn before the health-care provider acted on these results.

1. What event of the day may have affected these blood levels?
2. Do you agree that the specimen should be redrawn before the physician acts on the results obtained from the original specimen?

Case Study 8-2: Clumps

John George is performing a blood draw for a CBC, differential, and platelet count on Mrs. Charrone. As he completes the blood draw, he notices that she is looking really pale and is no longer conversing with him. He removes the tourniquet and the needle with the evacuated tube holder and lavender top tube from her arm just as she loses consciousness. He gently lowers her to the floor as she slides out of the phlebotomy chair, and elevates her feet as he calls for help. Within a few minutes she regains consciousness and after monitoring her for a few minutes, the health-care provider states that she is ready to leave the office. John helps her out to the waiting room to meet her husband, after which he returns to the blood draw area to process the specimen. He labels the specimen appropriately and places it on the blood rocker for the laboratory technician working in the testing area to access it for the CBC, differential, and platelet count.

In about an hour, the laboratory technician comes to find John and asks him if he remembers anything unusual about this blood draw. The laboratory technician said that the platelet count is extremely low with the initial analysis and that the hematology instrument flagged the sample with an error. The laboratory technician is going to recheck the specimen and create a blood smear to look at the sample more thoroughly.

1. How do you think the events of this blood draw may have affected the platelet count?
2. What do you think the technician may find when she looks at the blood smear?

RESOURCES AND SUGGESTED READINGS

"DIFF-SAFE Blood Dispenser: Directions for Use"
Provides step-by-step instructions on how to use a DIFF-SAFE Blood Dispenser http://www.alpha-scientific.com/Diff-safe2.html

"NCCLS simplifies the order of draw."
by Ernst D, Calam R. Medical Laboratory Observer, May 2004
National Committee for Clinical Laboratory Standards. Procedures for the Collection of Diagnostic Blood Specimens by Venipuncture. Approved Standard, H3-A5, Wayne, PA; 2003.

"Preanalytical Errors in the Emergency Department."
from BD Company, LabNotes. 17, no. 1, 2007 http://www.bd.com/vacutainer/labnotes/Volume17Number1/

"Preparation of Peripheral Blood Smear Staining With Wright's Stain
Directions on how to stain a peripheral blood smear using Wright's Stain http://www.scribd.com/doc/8801750/Preparation-of-Peripheral-Blood-Smear-Staining-With-Wrights-Stain

"Staining the Cells."
How to stain a blood smear; instructions on how to stain a peripheral blood smear http://www.tpub.com/content/medical/14295/css/14295_286.htm

"Venipuncture Technique Using the Multisample Vacutainer System"
Description of the use of the Vacutainer© system with details http://www.phlebotomycert.com/multi_vtainer_systm.htm

"Newborn Screening"
Washington State Department of Health. In-depth information about the various disorders that may be detected by newborn screening. Includes links to more detailed information about the collection process and specimen handling. http://www.doh.wa.gov/ehsphl/phl/newborn/disorders.htm#msud

Chapter 9

Collection and Processing of Urine Samples

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Types of Urine Specimens
  - Clean-Catch Midstream Urine Specimen Collection
  - Catheterized Specimens
  - Suprapubic Aspiration
  - Prostatitis Specimen Collection
  - Timed Urine Specimen Collections
    - *Glucose Testing Specimens*
  - Urine Collection Procedures for Infants and Pediatric Patients
- Urine Specimen Processing
  - Refrigeration and Preservation
  - Proper Disposal of Urine and Supplies
- Summary
- Time to Review
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

9-1 Define the key terms.

9-2 Explain the necessity of obtaining different types of urine samples for laboratory testing.

9-3 Compare and contrast various urine specimen types.

9-4 Describe how urine specimens are to be processed after collection.

9-5 Describe changes that may occur in urine specimens if they are left at room temperature after collection.

9-6 Analyze the details for proper labeling of the various types of urine specimens presented in the text.

9-7 Examine the necessary patient education for collection of samples to be used for specialized urine testing procedures.

9-8 Instruct a patient on the proper collection procedure for a clean-catch midstream urine specimen.

9-9 Instruct a patient on the proper collection procedure for a 24-hour urine specimen.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.6: Anatomy and Physiology: Perform patient screening using established protocols

I.A.I.2: Anatomy and Physiology: Use language/verbal skills that enable patients' understanding

ABHES Standards

- Medical Laboratory Procedures: e. Patient Instructions (collection of urine and feces); Instruct patients in the collection of a clean-catch mid-stream urine specimen.

KEY TERMS

- Aliquot
- Distal urethra
- Diurnal variation
- Fasting urine specimen
- First morning void
- Formed elements
- Glucose tolerance urine specimen
- Indwelling catheter
- Intermittent (straight) catheter
- Prostatitis
- Prostatitis specimen collection process
- Random
- Suprapubic aspiration
- 24-hour urine collection
- 2-hour postprandial specimen
- Urinalysis
- Urinary meatus
- Urine culture

TYPES OF URINE SPECIMENS

Urine analysis is one of the most common tests performed in the clinical laboratory. The collection process for urine specimens is noninvasive, and is relatively easy to accomplish because the production of urine by the body is an ongoing process. It is important to remember, however, that proper collection and processing procedures are essential for the urine laboratory results to be accurate and meaningful. (Table 9-1 summarizes the various types of urine specimens.)

Specimen collection procedures may vary based on the length of time for the collection, dietary limitations, specimen volume requirements, and the method of collection. Many urine specimens are collected to diagnose potential urinary tract infections, in which case they are usually **random** collections; they could be collected at any time of the day, as long as the collection procedures

TABLE 9-1

Types of Urine Specimens

Specimen Type	Time of Collection	Uses
Clean-catch midstream urine specimen	Random	Urinalysis, culture, random chemistry testing, urine pregnancy testing
Intermittent (straight) catheter specimen	Random	Urinalysis, culture, random chemistry testing, urine pregnancy testing
Indwelling catheter specimen	Not appropriate for urinalysis, culture, or chemistry testing	Not appropriate for urinalysis, culture, or chemistry testing
Suprapubic aspiration specimen	Random	Urinalysis, culture, random chemistry testing, urine pregnancy testing, urine cytology
Prostatitis specimen	Random as part of procedure	To check for bacterial prostatitis
First morning void specimen	Upon rising	Urine pregnancy testing, specific chemistry tests; urinalysis, culture
24-hour urine specimen	24 hours of collection	Chemistry testing
2-hour postprandial specimen	2 hours after a 100-g carbohydrate meal	Glucose and ketones
Glucose tolerance testing	As part of a glucose tolerance testing procedure	Glucose and ketones
Infant urine specimens	Random	Urinalysis

are followed appropriately. Other urine specimens are collected for chemical analysis to aid in the diagnosis of specific problems with the kidneys or other organ systems. These chemicals often change in concentration throughout the day, which is known as **diurnal variation.** To provide accurate results, specimen collection in these circumstances is carried out over a specific period of time at a certain time of the day. Procedures for collection may also include dietary restrictions for some tests to limit potential erroneous results because of a diet rich in a particular substance.

The guidelines that should be followed whenever a urine specimen is to be collected include the following:

- All specimen containers must be correctly labeled. In the case of urine collection, this should occur immediately after the patient returns the specimen container to the laboratory or health-care facility. All labels should include the patient's name, birth date or other unique identifier, and the date and time of the collection. Labels must be placed on the container itself, not just on the lid. It is also a good idea to document the type of specimen (e.g., urine) in the container; many body fluids have a similar appearance when received in the laboratory.
- The required specimen volume must be verified before instructing the patient on collection requirements.
- Any medication the patient is taking that may potentially interfere with the test results needs to be documented on the requisition form.
- For timed specimens, the start and finish time (and dates if appropriate) must be documented on the container.
- Dietary restrictions must be communicated to the patient before collection, both verbally and in writing.
- Collection of urine for analysis during the menstrual cycle should be avoided if possible; the presence of red blood cells may interfere with the testing procedures for some analytes. If it is not possible to delay the collection, care must be taken to instruct the patient on proper cleaning techniques to avoid gross contamination, and a notation should be placed on the laboratory requisition form (and/or in the patient chart if available) that the patient is menstruating.
- Specimen containers that have preservatives added need to be clearly marked, and in these situations, there should be an additional container provided for the patient to use during collection to avoid urinating directly into the specimen container with the preservative.
- Patients should *always* be provided with a specimen container, so that they are not using a jar or other container from home. Although these may appear "clean," they may contain chemicals or other substances that can affect the urine test results.
- Any instructions for collection must be given to the patient verbally and in writing. These include storage instructions for the specimen until it is returned for analysis.

Test Your Knowledge 9-1

Please provide two reasons that a urine specimen may be ordered for testing. (Outcome 9-2)

Test Your Knowledge 9-2

List one item that is the same for all urine collection procedures. (Outcome 9-2)

POINT OF INTEREST 9-1

Urine drug screens

Many health-care providers and laboratories are now involved in preemployment physicals. These may include urine drug screens for drugs of abuse. This type of urine collection has very specific procedural parameters designed to prove that the specimen collected was not tampered with prior to the completion of testing. There are guidelines that dictate the steps involved in the collection procedures, as well as guidelines that outline the security of the specimen during transportation and processing. A series of documentation on a standardized form is necessary; this process establishes the *chain of custody* for the specimen. Federal regulations require preemployment, postaccident, and random drug screening for some jobs, whereas other large nonfederal employers choose to ensure a drug-free environment by requiring their employees to undergo testing. The following is a summary of the steps involved in a urine drug specimen collection procedure:

1. Collectors must be appropriately trained for the process.
2. The collector verifies that there is a bluing agent (some sort of dye) added to the toilet before the donor process begins.
3. The donor is greeted and must provide photo identification. In some situations, alternative forms of identification are acceptable from an employer representative.

Continued

4. The collector fills out the first step of the chain of custody form, and the donor signs to verify the accuracy of this information.
5. The donor leaves all purses, coats, and so on outside of the restroom to eliminate the possibility of smuggling in concealed substances to contaminate the urine.
6. The donor washes and dries his or her hands.
7. The collector must tape the toilet lid and the faucet handles with tamper-proof tape so that there is no access to other water sources.
8. The collector remains in the restroom (if there is a private stall) or outside the restroom door if there is no private stall during the collection process. Sometimes a witnessed collection is requested, in which case the donor must be watched by the collector during the collection.
9. The donor is not allowed to wash his or her hands until the specimen is handed to the collector and the follow-up quality control measures and documentation have been completed. The urine specimen must remain within the site of the donor and the collector during this process.
10. The collector verifies the volume received, the appearance of the urine, and records the temperature of the specimen. If there is a problem with the temperature or the appearance of the urine specimen, the collection process is void.
11. As the donor observes, tamper-proof collection labels are placed over the top of the collection container and the donor initials these labels. The date and time are also added to these labels.
12. These steps are documented on the chain of custody form.
13. From this point forward, every individual involved in transportation, processing, or testing of the specimen must document their actions and the purpose of their actions on the chain of custody form.

Clean-Catch Midstream Urine Specimen Collection

When a urinalysis or urine culture is ordered for a patient, it is important the specimen be as contaminant free as possible. A **urinalysis** includes physical observations and chemical analysis of the urine specimen, and may also include a microscopic examination of the urine sediment. A **urine culture** is designed to test for bacterial growth in the urine specimen. In both these situations, it is imperative that the specimen received for testing is representative of the urine in the urinary bladder and urethra, rather than contaminants that might be present on the outside of the body. A clean-catch midstream urine specimen collection technique is the best noninvasive method to obtain a urine specimen for these tests. This procedure involves cleaning the external opening of the urinary tract and obtaining a midstream urine specimen.

The clean-catch midstream urine collection technique offers an opportunity to obtain a specimen that has less contamination with epithelial cells and bacteria from the distal urethra and urinary meatus than a specimen obtained without proper cleansing and technique. The **distal urethra** is the area of the urethra that is closest to the outside of the body. The **urinary meatus** is the exterior opening of the urinary tract. Because these are both near the outside of the body, they are naturally potential sources of exterior contamination. Even though all urine specimens must pass through this area of the urinary tract as they exit the body, the goal is to obtain a specimen that contains the bacteria and other microscopic structures that are clinically significant, (those that are present in the bladder) and minimize those that are essentially contaminants. This is critical for patients to be properly diagnosed and treated for urinary tract infections.

Test Your Knowledge 9-3

Why is the cleansing process for a clean-catch midstream urine specimen so important? (Outcome 9-2)

Test Your Knowledge 9-4

Should a urine collection container be labeled before or after it is given to the patient for collection? (Outcome 9-6)

Catheterized Specimens

Another way to avoid potential specimen contamination is to use a catheter to obtain the specimen. Only a licensed health-care professional may insert a catheter. This process may also be used when the patient is not capable of following the directions or performing the procedures necessary for a clean-catch urine collection procedure, such as those who are very young or those who may have physical limitations. Catheters may be

(Text continues on page 203)

Procedure 9-1: Instructing a Patient for Collection of a Clean-Catch Midstream Urine Specimen for Urinalysis and/or Culture

Clean-catch midstream urine specimens are required by most laboratories for performance of urinalysis and cultures. By cleansing appropriately and providing a midstream specimen that comes from the bladder and upper urethra, most extraneous microorganisms will be eliminated from the specimen.

TASK

The student will demonstrate the ability to properly instruct a male or female patient how to collect a clean-catch midstream urine specimen.

CONDITIONS

- Appropriately labeled sterile urine collection container
- Three moist antiseptic towelettes
- Hand-washing facilities
- Paper towels

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.6: Anatomy and Physiology: Perform patient screening using established protocols
I.A.I.2: Anatomy and Physiology: Use language/verbal skills that enable patients' understanding

ABHES Standards

Medical Laboratory Procedures: e. Patient Instructions (collection of urine and feces); Instruct patients in the collection of a clean-catch mid-stream urine specimen.

Procedure	Rationale
1. Wash hands and gather necessary supplies for the patient.	Hand washing breaks the chain of infection.
2. Greet and identify the patient. Verify the test orders and the labeling on the container.	The patient identity should always be verified for every procedure. Verification of test orders and labeling reduces the potential for error.
3. Provide the patient with a sterile container and three antiseptic towelettes.	The patient will need to refer to these items as the procedure is explained. The towelettes will be used for cleansing around the urinary opening area. Strong antiseptics should not be used because of the potential irritation to the external urinary structures. Date Room No. Name Doctor

Continued

Procedure 9-1: Instructing a Patient for Collection of a Clean-Catch Midstream Urine Specimen for Urinalysis and/or Culture—cont'd

Procedure	Rationale
Instructions for Male Patients	
a. Instruct the patient to wash and dry his hands, and once in the restroom, remove his underwear.	Hand washing breaks the cycle of infection, and removing the undergarments allows for better access during the cleaning process.
b. The patient should carefully remove the lid of the specimen container and place it with the inside up on the counter near the toilet where the container can easily be accessed for the collection process. Emphasize the importance of keeping the interior of the container sterile.	Placing the lid with the inside up prevents contamination from microorganisms that may be present on the countertop.
c. Male patients who are uncircumcised must pull back the foreskin, and keep it withdrawn during the cleaning and collection process.	Pathogenic microorganisms may be present under the foreskin.
d. The tip of the penis must be cleansed three times: • Wipe with the first wipe from front to back on one side of the tip of the penis. • Wipe with the second wipe from front to back on the other side of the tip of the penis. • Use the third wipe over the urinary opening (the meatus) from front to back. Note: A new antiseptic towelette must be used to clean each side of the penis and an additional towelette used to clean the tip of the penis. (A total of three towelettes are to be used for the cleansing process.) • Some laboratories may provide a fourth towelette to clean the outside of the container after the collection.	Thorough cleansing eliminates potential contaminating substances. Once the cleaning process has started, the patient should keep holding the penis until the specimen has been obtained to eliminate the possibility of recontamination. 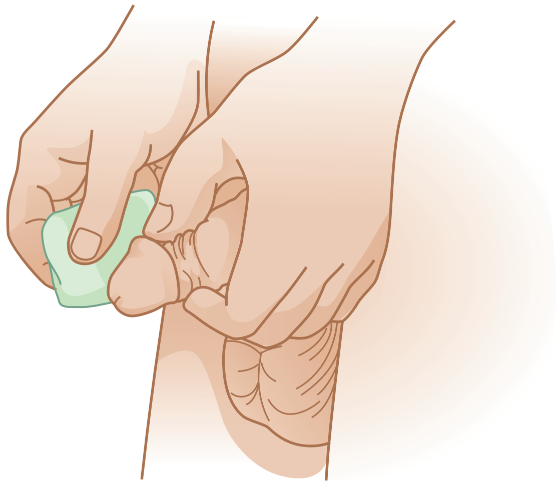

Procedure	Rationale
e. Instruct the patient to pick up the specimen container so that he is ready when for the collection. Begin to urinate into the toilet. Stop the urine stream, move the specimen container into a position where it will intersect the urine stream, and urinate into the container until it is at least half full. Stop the urine stream again and move the container out of the way. Remind the patient not to touch the inside of the container.	This process flushes out microorganisms that may be present near the opening of the urethra and allows collection of a specimen that represents the urine that was present in the bladder and upper urethra.
f. Instruct the patient to finish urinating into the toilet.	The specimen must be *midstream*; the first and last part of the urine void should not be included in the specimen submitted to the laboratory.
g. The patient must carefully replace the lid on the container without contaminating the interior of the cup or lid. If urine is present on the outside of the cup, dry it off with toilet tissue or a paper towel, and discard the waste.	Contamination must be avoided on the inside of the container so that the specimen represents the urine as it is inside the body.
h. Instruct the patient to wash and dry his hands.	Clean hands prevent the spread of microorganisms.
i. Instruct the patient to place the urine in the designated area of the laboratory if the specimen is collected in the office. If it is a home collection, instruct the patient to refrigerate the specimen until it can be transported to the testing location.	The specimen will need to be refrigerated or tested in a timely manner to preserve the integrity of the test results.
Instructions for Female Patients	
a. Instruct the patient to wash and dry her hands, and once in the restroom, remove her underwear.	Hand washing breaks the cycle of infection, and removing the undergarments allows for better access during the cleaning process.
b. The patient should carefully remove the lid of the specimen container and place it with the inside up on the counter near the toilet where the container can easily be accessed for the collection process. Emphasize the importance of keeping the interior of the container sterile.	Placing the lid facing up prevents contamination with microorganisms that may be present on the countertop.
c. Instruct the patient to move her knees apart until she can easily reach her labia. (These are the structures that cover the urinary opening in females.)	It is important that the patient has full access to the labia to clean the area properly.

Continued

Procedure 9-1: Instructing a Patient for Collection of a Clean-Catch Midstream Urine Specimen for Urinalysis and/or Culture—cont'd

Procedure	Rationale
d. Instruct the patient to spread apart the labia with one hand to expose the urinary opening (the urinary meatus). The labia will need to remain spread apart until the cleaning and collection process is complete.	The labia must remain apart after the area is cleaned to eliminate the potential for recontamination of the site. 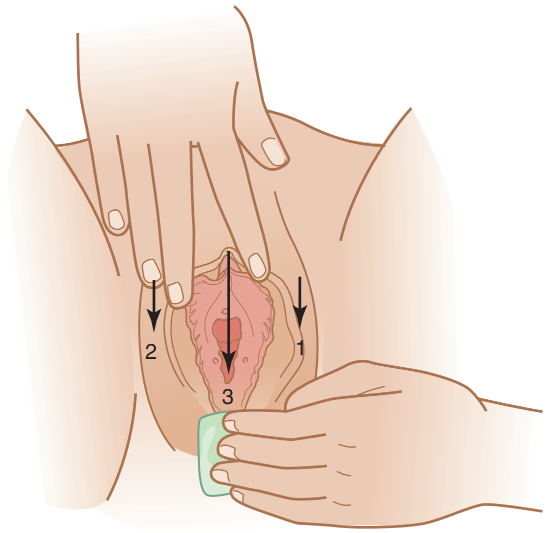
e. The patient needs to clean the area three times, using a clean antiseptic towelette each time. • The first towelette should be used to clean from front to back on one side of the urinary opening. • Use the second towelette to wipe front to back on the other side of the urinary opening. • The third towelette is to be used to wipe front to back across the urinary opening for the final cleansing step. • Some laboratories may provide a fourth towelette to clean the outside of the container after the collection.	Cleansing in this manner avoids potential contamination from the anal area (front to back) and provides an opportunity to thoroughly clean away potential pathogenic microorganisms that could enter the specimen while urinating.
f. Instruct the patient to pick up the urine specimen container in the hand that is not holding the labia apart.	The specimen container must be in hand before the patient starts to urinate.
g. The patient should begin to urinate into the toilet. After a small amount has entered the toilet, stop the urinary stream and place the specimen container under the urinary opening.	This initial urine stream flushes out potential contaminants that may have been around the urinary meatus.

Procedure	Rationale
h. Instruct the patient to urinate directly into the specimen container, filling it at least half full. Tell the patient to stop urinating and move the specimen container out of the way. Remind the patient to avoid contamination of the interior of the specimen container.	This is a *midstream* specimen that includes urine from the bladder and the upper part of the urethra.
i. Instruct the patient to finish urinating into the toilet.	The first and last part of the urine specimen should not be submitted to the laboratory.
j. The patient must carefully replace the lid on the container without contaminating the interior of the cup or lid. If urine is present on the outside of the cup, dry it off with toilet tissue or a paper towel.	Contamination must be avoided on the inside of the container so that the urine is representative of the conditions on the inside of the body.
k. Instruct the patient to wash and dry her hands.	Hand washing stops the spread of microorganisms.
l. Instruct the patient to place the urine in the designated area of the laboratory if the specimen is collected in the office. If it is a home collection, instruct the patient to refrigerate the specimen until it can be transported to the testing location.	The specimen will need to be refrigerated or tested in a timely manner to preserve the integrity of the test results.
4. Verify whether the patient understood the instructions as given, and check if he or she has any additional questions. Supply contact information for future questions.	This is a critical step for the appropriate collection of the specimen.
5. Provide the patient with a written copy of the instructions.	A written copy will allow patients to refresh their memory when the collection is performed.

Date	
10/24/2016:	Patient provided supplies and written and verbal instructions for a clean-catch midstream urine collection procedure.
11:58 a.m.	Connie Lieseke, CMA (AAMA)

used in various ways (listed below), and it is imperative that the requisition form and the patient chart include documentation of the type of collection performed.

- **Intermittent (straight) catheter:** There may be times that it is necessary to empty the bladder for a patient who is unable to do so. If this is a temporary situation, an **intermittent (straight) catheter** may be used for the process. This catheterization procedure involves a small, stiff, plastic tube that is placed through the urethra into the bladder to allow the bladder to be drained. This procedure may also be used to obtain a specimen for urinalysis and culture, as the potential for contamination is greatly reduced. Catheterization, however, does include risks of bacterial infection because the lining of the urethra may be damaged as the catheter is inserted or removed.
- **Indwelling catheter: Indwelling catheters** are designed for long-term use. They may be inserted in surgery patients for a few days after surgery, or they may be inserted in patients who have permanently lost the ability to control urination. These catheters are larger in diameter than the straight catheters, and have a small “balloon” that is pumped up to keep them in place within the bladder once they have been inserted. Indwelling catheters continuously drain the urine from the bladder into a collection bag. This bag is emptied periodically. Indwelling catheters are *not* recommended

as a source of urine for urinalysis or urine cultures, as the catheters are kept in place for extended periods of time and become contaminated with bacteria. If a patient with an indwelling catheter needs to have a urinalysis, it is best for the catheter to be removed and for the health-care professional to use a straight catheter to obtain the specimen for analysis.

Test Your Knowledge 9-5

Which type of catheterization procedure is used to obtain a urine specimen for a routine culture?

(Outcome 6-3)

Suprapubic Aspiration

There are times, especially with young children, when it is imperative that a urine specimen be collected quickly, without contamination. Skilled health-care professionals (such as physicians or nurse practitioners) may choose to perform a **suprapubic aspiration** to obtain a specimen for urinalysis and culture. A needle is placed externally through the abdomen directly into the bladder, and a urine specimen is withdrawn. This procedure may also be used to collect urine for *cytology* examination, a procedure in which the cells present in the specimen are examined by a pathologist or cytologist for abnormalities in their appearance.

Prostatitis Specimen Collection

Prostatitis (inflammation of the prostate) may be caused by various factors. Prostate enlargement and/or use of urinary catheters are often linked to this condition. The tissue of the prostate may be inflamed, or there may be an infection with pathogenic microorganisms. In order to plan the correct course of action, it may be necessary to collect a urine specimen utilizing a special procedure. In a **prostatitis specimen collection** process, there are three specimens collected in close progression. The patient is provided with three antiseptic towelettes and three sterile specimen collection containers, numbered sequentially 1, 2, and 3. The patient cleansing process is the same as that required for the clean-catch midstream collection procedure. However, rather than urinating into the toilet at the beginning of the process, the first small amount of urine passed goes into collection cup 1. Next, the midstream specimen goes collected in cup 2. After this collection is completed into cup 2, the health-care provider will massage the patient's prostate to allow prostatic secretions to enter the urethra for collection. After the prostate has been massaged, the patient will immediately urinate into the third cup. Cultures are performed on all three specimens. If the second specimen has a high concentration of bacteria present, the third specimen is considered to be invalid, because of contamination by the bacteria already present in the urethra. Generally, if there is an infection of the prostate, the bacterial count of the third specimen will be much higher than that of the first two specimens.

Timed Urine Specimen Collections

Most urine specimens collected in the office setting are random collections, because the time of day is not significant for the desired results. The clean-catch midstream urine collections used for urinalysis and culture are usually random specimens. Remember, however, that the time of collection still must be recorded accurately. A random collection is not always desirable for all urine specimens. There are situations in which the *time* of collection is critical:

- **First morning void specimen:** A specimen that is collected when the patient first rises from sleep is known as a **first morning void** specimen. This specimen is concentrated (ideally it should be approximately 8 hours since the patient last voided)

POINT OF INTEREST 9-2

Proper documentation for urine collection procedure education

This chapter includes a lot of information about different types of urine collection procedures. The key component in all these processes is patient education. The patient should be provided with the appropriate collection supplies in all situations, as well as verbal and written instructions for the collection process. Every laboratory should develop its own literature to distribute to patients so that there is standardization of information provided for collection and storage of the specimens. Government regulations require that the instructions for patients are to be provided in their native language whenever possible. In addition, it is very important to remember that whenever instructions are provided for the patient, doing so must be documented properly in the patient chart.

because it has been in the bladder during the night while the patient slept. The first morning void specimen is recommended for urine pregnancy tests (especially if it is very early in the potential pregnancy) and for detection of the presence of protein or albumin in the urine. It may also be used for routine urinalysis and culture, but the clean-catch midstream collection procedure should still be followed. There will be more dissolved substances present in this specimen than in those found later in the day, and it may also contain more microscopic structures.

- **24-hour urine collection:** Routine urinalysis testing reports the presence or absence of numerous chemicals in the urine specimen and provides a rough quantitative estimate of the amount present for that chemical. For some chemical components, however, it is necessary to have an exact measurement of the amount of that chemical present in the urine. These analytes are not present in the same concentration at all times; they may vary with the time of the day (diurnal variation), or their concentration may be changed with exertion or the hydration status of the individual. To account for these variables, it is often desirable to collect all the urine produced by a patient over the period of an entire day. This is called a **24-hour urine collection.** This collection process is used to test for various substances; among the most common are the following:
    - Urea nitrogen: Urea nitrogen is a by-product of protein metabolism, and is used as an indicator of kidney function.
    - Urine electrolytes: Urine electrolyte tests include those that check for the level of calcium, chloride, potassium, and sodium in the urine. These tests may be ordered together or individually. They are often part of a clinical workup for suspected endocrine disorders or hormonal imbalances or kidney dysfunction. Urine calcium levels are sometimes ordered for patients with recurring kidney stone formation. Situations that cause electrolyte imbalances in the bloodstream may also result in abnormal urine electrolyte levels.
    - Catecholamines: Catecholamines are substances that are created by nerve tissue and by the adrenal glands of the body. They include norepinephrine, dopamine, and epinephrine. Urine specimens do not contain the intact catecholamine molecules; instead, we test for the breakdown products of these compounds. The breakdown products include homovanillic acid (HVA), vanillymandelic acid (VMA), normetanephrine, and metanephrine. Urine catecholamines may be ordered as a group or individually. This test may be used when an adrenal gland tumor is suspected (pheochromocytoma), and it may also be ordered for patients with symptomatic hypertension.
    - 17-Hydroxysteroids: This test is often performed when there is suspicion of Addison's disease, which is a type of adrenal insufficiency.
    - Urine total protein: The level of total protein in the urine may be an indicator of kidney damage.
    - 5-Hydroxyindoleacetic acid: More commonly known as 5-HIAA, this test may be performed to detect a specific type of tumor, the carcinoid tumor, of the digestive tract. 5-HIAA is a metabolite of serotonin in the urine, and these intestinal tumors secrete unusually high amounts of serotonin.
    - Urine creatinine: Creatinine is a breakdown product of creatine, present in muscle tissue throughout the body. Creatinine is cleared from the body exclusively by the kidneys, and the test is ordered to screen for possible kidney dysfunction.

In some situations, blood is also drawn and tested for the same analyte as the urine specimen at the end of the 24-hour interval. For example, when urine creatinine is ordered, there is often a blood draw performed for creatinine as well, so that both results can be used to reach a diagnosis.

Urine specimens collected over a 24-hour period may have dietary or fluid restrictions; this is to limit potential erroneous results because of a diet rich in a particular substance. Patient education concerning these restrictions is *critical* for a meaningful urine test result. The patient must be given the instructions verbally and in writing, and documentation of the education should be added to the patient's chart.

It may also be necessary to deliver (or even ship) these urine specimens directly to a laboratory for processing, rather than returning them to the health-care provider's office. These instructions should be explained clearly before the collection process is started. The patient should be provided with all shipping materials and written instructions so that the specimen is delivered successfully.

When refrigeration is not possible (especially when the specimen must be transported over long distances) chemical preservatives may be utilized to keep the 24-hour urine specimen from undergoing changes upon standing. Chemical preservatives (such as boric acid or dilute hydrochloric acid) are often added to 24-hour urine specimen containers to keep chemical changes from occurring before the specimen is tested. These additives for 24-hour urine specimen collection are listed in Table 9-2.

Test Your Knowledge 9-6

When providing instructions to a patient for a 24-hour specimen collection, what should the patient be told about urinating directly into the container that contains hydrochloric acid as a preservative? (Outcome 9-7)

There are additional handling procedures that are required by 24-hour urine specimens once they arrive at the laboratory. These procedures include proper documentation of the total specimen volume upon arrival at the laboratory, and verification of the times included in the collection. The specimen must also be well mixed and an **aliquot,** which is a small sample, must be poured off for further chemical testing. Figure 9-1 shows a 24-hour urine collection container.

TABLE 9-2

Additives used for 24-hour urine specimens

Test Ordered	Preservative Added
Aldosterone	1 g boric acid per 100 mL urine
Amylase	Add NaOH; adjust pH to 6 or above
Urine calcium	Add HCl to adjust pH 1.5–2.0
Catecholamines	Add HCl to adjust pH to 2–3 or add 25 mL 6N HCl
Cyclic AMP (adenosine monophosphate)	10 mL 6M HCl
17-Hydroxy-corticosteroids	Boric acid to adjust pH 5–7
Urine phosphorus	HCl to adjust pH to 1.5–2
Testosterone	1 g Boric acid per 100 mL
Uric acid	NaOH to adjust pH to 8 or above
Dopamine	Glass container with 25 mL HCl
Homovanillic acid (HVA)	Glass container with 25 mL HCl
Iodine	Glass container with 25 mL acetic acid
Vanillymandelic acid (VMA)	Glass container with 25 mL 6N HCl
Urea nitrogen	25 mL 6N HCl
5-Hydroxy-indoleacetic acid (5-HIAA)	25 mL 6N HCl

Procedure 9-2 details the instructions for collection of a 24-hour urine specimen.

Glucose Testing Specimens

Glucose is not usually present in the urine at levels high enough to be measured. However, in patients with diabetes mellitus, blood glucose concentrations may reach levels that are high enough to cause glucose to “spill” into the urine. There are several urine collection procedures that may assist with the diagnosis or monitoring of those with diabetes mellitus or gestational diabetes. These include the following:

- **2-hour postprandial specimen:** For a **2-hour postprandial specimen,** the patient is instructed to empty the bladder before a meal that contains approximately 100 g of carbohydrates. Two hours after meal completion, the patient is to collect a urine specimen. Ideally, the patient will also have a blood sample drawn 2 hours after eating so that both results can be used to evaluate the patient’s condition. Because there will not be a culture or complete urinalysis performed on the specimen, 2-hour postprandial urine specimens do not need to be collected using the clean-catch midstream urine collection procedures, The specimen is tested only for the presence of glucose. (Additional

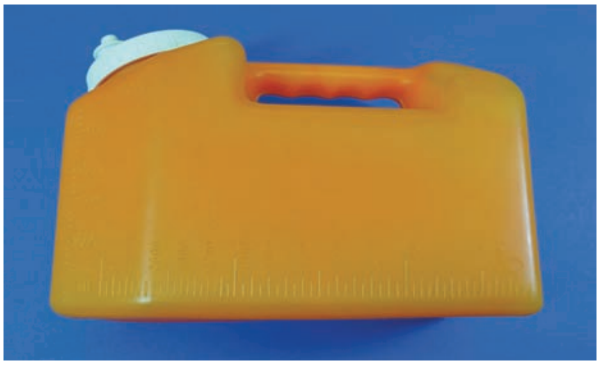

Figure 9-1 A 24-hour urine collection container. Notice the large-volume capacity, the measurements along the side of the container, and the amber color, which blocks light exposure, as some chemical concentrations are decreased with exposure to light. This 24-hour urine collection container includes a special closure adaptation to make it easy to pour the specimen when it arrives at the laboratory.

information about this procedure is included in Chapter 17.)

- **Glucose tolerance urine specimen: Glucose tolerance urine specimen** tests are used for diagnosis of gestational diabetes or diabetes mellitus. Urine and blood specimens are collected at specific intervals after a fasting patient is given a high glucose substance to drink. The first urine specimen collected should ideally be a "fasting" specimen, meaning that it is the second morning void, not the first. The first morning void contains the breakdown products from the food ingested the night before the test; the second morning void should be representative of the body in a state of fasting. After the patient has been given the high glucose substance to drink, a blood and urine collection is performed at least three more times at hourly intervals for 3 more hours. Additional samples may be collected depending on the policy for that particular laboratory. It is not necessary to use the clean-catch midstream collection procedures for this test, as the specimens collected are not used for cultures or complete urinalysis procedures.

Procedure 9-2: Instructing a Patient for Collection of a 24-Hour Urine Specimen

This process collects all urine that is produced by the body over a 24-hour period. It is noninvasive, and provides a great deal of information about the patient's renal function. Remember, urine contains many chemical substances such as sodium, potassium, and creatinine that are dissolved in water. Because the concentration of these chemicals may vary throughout the day, it is important to collect all the urine produced over a 24-hour period for an accurate assessment.

TASK

Provide the necessary supplies and education for a patient to successfully complete a 24-hour urine specimen collection.

CONDITIONS

- An appropriately labeled 24-hour urine collection container (or more than one container if a high volume of urine is anticipated)
- Warning sticker if preservative is added to the specimen container
- An additional smaller container (such as a urinal device) for urine collection
- Written patient instructions, including any dietary restrictions for the collection period

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.6: Anatomy and Physiology: Perform patient screening using established protocols

I.A.I.2: Anatomy and Physiology: Use language/verbal skills that enable patients' understanding

ABHES Standards

None

Procedure	Rationale
1. Wash hands and gather necessary supplies for the patient. Verify any dietary restrictions and/or preservative requirements for the test ordered by checking the laboratory reference guide. Check to see if the instructions for the analyte ordered include a blood draw.	Hand washing breaks the chain of infection. It is imperative that the medical assistant verify all restrictions and special instructions before meeting with the patient. There should be a written copy of all instructions (including dietary restrictions) available so that they can be provided to the patient.

Continued

Procedure 9-2: Instructing a Patient for Collection of a 24-Hour Urine Specimen—cont'd

Procedure	Rationale
2. Add preservative to the collection container if necessary, using appropriate safety precautions. Apply warning label if acid was added to the container.	Warning labels must be added to the container if there has been preservative added (e.g., hydrochloric acid) that could harm the patient. If there is a harmful preservative added to the container, the patient must be told that he or she cannot urinate directly into the container or splash the specimen on bare skin. 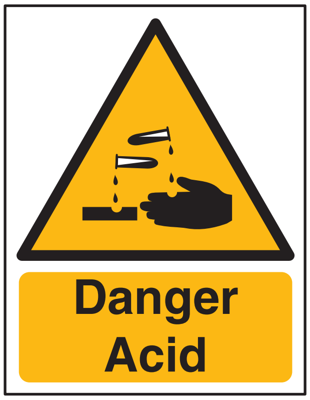
3. Greet and identify the patient. Verify the test order once more.	The patient identity must always be verified for every procedure. Verification of test orders and labeling reduces the potential for error. Some laboratories have labels that include the dietary restrictions that are also attached to the container.
4. Provide the patient with the 24-hour urine collection container and an additional smaller, wide-mouth container for urine collection.	The 24-hour urine collection container will be amber or dark in color. It will hold approximately 3 liters of urine. The dark color protects some of the urine chemicals from decomposing with exposure to light. The smaller container is used to collect the specimen during urination, after which it will be added to the large container.
5. Instruct the patient that the 24-hour specimen must remain refrigerated during the collection process. Also emphasize that *all* urine collected during the 24-hour period must be collected in the container provided.	For accurate results the patient must understand that *all* the urine in the 24-hour period must be collected and preserved appropriately. For convenience, the patient may want to choose a day for collection when he or she will be home most of the time.
6. Instruct the patient that upon arising on the first day of the 24-hour period he or she is to empty the bladder by urinating into the toilet. This is the start time of the collection period, so the patient is to note this time on the collection container and on any paperwork that may have been provided for the process.	For accurate results, this part of the process should be emphasized; it is a bit difficult to understand without a thorough explanation.

Procedure	Rationale
7. Explain to the patient that all the urine for the next 24 hours (after that initial void) must be saved in the large container. Instruct the patient that he or she may use the smaller collection container during the urination process, but this urine must all be added immediately to the larger container.	It must be stressed that all the urine must be collected during the 24-hour period for the test results to be valid.
8. Explain to the patient that if for any reason all the urine is not put in the container, the collection process must be started over.	The reference ranges for the tests are dependent on 24-hour specimens; any urine not added to the specimen container may potentially affect the results.
9. Explain to the patient that the 24-hour collection period ends the following day at the same time that the process was started. This urine (from the following morning at the same time) is to be added to the collection container to complete the process. Document this time on the collection container and on any additional paperwork as required.	Emphasize that the 24-hour period starts and ends with an empty bladder.
10. Tell the patient to transport the filled container to the laboratory as directed. It may be necessary to collect blood when the specimen is returned to the laboratory.	The patient should be notified before the collection process begins if there is a blood draw necessary so that he or she is anticipating this.
11. Verify whether the patient understood the instructions as given, and check if there are any additional questions. Supply contact information for future questions.	This is a critical step for the appropriate collection of the specimen.
12. Provide the patient with a written copy of the instructions.	A written copy will allow the patient to refresh his or her memory when the collection is performed.

Date	
10/24/2013:	*Patient provided supplies and written and verbal instructions for a 24-hour urine collection procedure.*
11:58 a.m.	*Connie Lieseke, CMA (AAMA)*

Test Your Knowledge 9-7

How is a first morning void specimen different from a 24-hour urine specimen? (Outcome 9-3)

Urine Collection Procedures for Infants and Pediatric Patients

It can be very difficult to collect a urine specimen from an infant or young child using noninvasive procedures. A routine specimen may be obtained by using a special type of collection bag designed to be attached to the skin surrounding the urinary opening of a young child or infant. The area around the urinary meatus must be carefully cleaned to reduce contamination, and care must be used to attach the device securely so that the urine does not leak out around the adhesive. It is also important to inform the parents of the patient (if the collection is to be completed at home) to check for the presence of urine in the bag often so that it can be transferred to an appropriate sterile collection cup and refrigerated. Figure 9-2 shows a pediatric urine collection

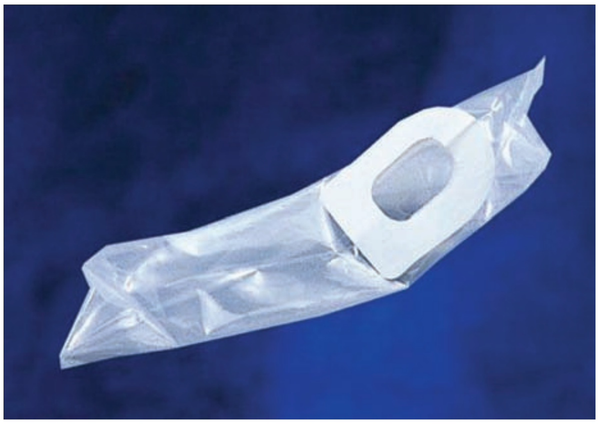

Figure 9-2 Pediatric urine collection device. Note the adhesive surrounding the opening of the bag. The adhesive surface will be applied to the child's cleansed skin surrounding the urinary opening. *Courtesy of Zask, Inc.*

bag. A suprapubic aspiration or straight catheterization procedure may be performed by the physician (or other qualified health-care professional) for young children and infants if a sterile specimen is necessary for proper diagnosis.

Test Your Knowledge 9-8

Why is a special collection device necessary to collect a urine specimen from an infant? (Outcome 9-3)

URINE SPECIMEN PROCESSING

Because urine can be collected relatively easily without invasive procedures, it is often taken for granted. A urine specimen may have drastic changes occur while standing at room temperature, so care must be taken to handle the specimen correctly after collection until the testing process is complete. Urine that is not processed correctly can lead to misdiagnosis and incorrect treatment for the patient.

Although urine is not considered to be capable of transmitting HIV, hepatitis B, or hepatitis C, it is important that the processor uses gloves and a protective laboratory coat when handling urine samples. Other pathogenic microorganisms may be present in the urine specimen, and employees need to be certain to protect themselves appropriately. Protective eyewear may also be considered if there is a potential for splashing the specimen. This may be particularly critical when handling specimens that have acid added as preservative, or those that may have a lot of blood present in the urine.

POINT OF INTEREST 9-3

What now?

Once the specimen has been collected and returned to the laboratory or physician's office, it is up to the medical assistant to process and store it correctly. Here are a few basic guidelines:

- While still in the presence of the patient, verify that the specimen is correctly labeled with name, birth date, and other identifying information.
- Verify and, if necessary, document the date and time of collection on the appropriate paperwork.
- Review the collection process with the patient; ask specifically whether there were any problems with the specimen collection or storage. If the collection or storage of the specimen was incorrect, the specimen should be discarded at this time, and the patient should be provided with materials to perform a recollection.
- Some facilities will have the medical assistant immediately document the color, appearance, and specific gravity of the specimen before it is refrigerated, as these can change once the specimen temperature decreases.
- Specimens for urinalysis, culture, and most other procedures are to be tested or refrigerated as soon as possible, and always within 1 hour of arrival at the laboratory.
- Twenty-four-hour urine collections may need to be mixed well and an aliquot may need to be poured off. This aliquot would then be forwarded to the main laboratory for analysis. (This step will be dependent on the policy of the laboratory performing the 24-hour urine collection container.)

Test Your Knowledge 9-9

Why are gloves worn when handling urine specimens in the laboratory? (Outcome 9-4)

Refrigeration and Preservation

Urine specimens should be refrigerated (at 2° to 8°Celsius) immediately after collection if they cannot be analyzed within 1 hour of collection. If the specimen is collected in a patient's home, instructions about refrigeration are especially critical. The sample should stay cool until delivered to the laboratory; if a patient has numerous errands to run, or if the patient is going to work, it is important that the patient use a cooler to keep the urine refrigerated until it is delivered to the laboratory.

There are numerous changes that may occur as the urine is left standing at room temperature, but the most dramatic is the change in the bacterial count. The bacteria present in urine specimens are capable of multiplying very rapidly; the bacterial count of the specimen may double within 2 hours at room temperature. This means that the patient may initially be treated for a urinary tract infection based on a high bacterial count reported as part of the urinalysis, when there were actually very few bacteria present in the specimen at the time of collection. Refrigeration halts most of the changes that may occur in the urine specimen, and reduced temperatures are especially effective at limiting the replication process of bacteria. The potential changes to a urine specimen with extended room temperature storage are listed in Table 9-3.

Test Your Knowledge! 9-10

What change in the urine specimen is most dramatic if the specimen is left out at room temperature for more than an hour? (Outcome 9-5)

TABLE 9-3

Changes in urine specimen when left at room temperature for more than 1 hour

Changes in Urine Specimen	Details and/or Clinical Significance
Decreased clarity and slight color change	Crystallization of solutes; may also cause color interpretation to change as these form
pH increases	Bacteria in specimen convert urea to ammonia and CO_2 is lost
Ketones decrease	Ketones decrease with time at room temperature
Bilirubin and urobilinogen decrease	Exposure to light causes these chemicals to degrade in the specimen
Cells and casts dissolve	Microscopic examination is inconclusive as identification of microscopic features is inaccurate
Bacterial count increases	Bacteria continue to multiply exponentially
Glucose decreases	Glucose is used by bacteria and other cellular components as an energy source

Although it is necessary to preserve urine specimens at reduced temperatures to stabilize some of the analytes, it is important to remember that these reduced temperatures may also affect the specimen. Refrigeration may increase the *specific gravity* of the specimen, and cause some diffuse microscopic crystallization within the urine specimen as well. The specific gravity is a measurement of the amount of dissolved substances in the urine specimen. (More information about specific gravity is covered in Chapter 20.) Urine that has been refrigerated may also change in color and appearance, and these parameters are recorded as part of a routine urinalysis in many situations. In order to counteract the effects of refrigeration on these testing parameters, the urine specimens should be allowed to return to room temperature before routine urinalysis testing is performed. This reverses most of the changes. Also remember the urine specimens must be well mixed with gentle inversion or swirling of the specimen container before testing.

Additives (preservatives) may also be used for specimens that are to be tested for common urinalysis tests and/or set up for urine cultures. Many manufacturers produce tubes with additives for this purpose. An example is BD Vacutainer Plus Plastic Conical urinalysis preservative tubes, which contain chlorhexidine, ethyl paraben, and sodium propionate for preservation of urine specimens for urinalysis. Some of these additive tubes cannot be used for certain types of test procedures; it is imperative that the medical assistant check the package insert to verify whether a specific type of collection container with an additive may be used for the procedure ordered. These special additive tubes are not usually given to the patient for transfer at home; the urine specimen is added to the tube once it arrives at the laboratory or physician office. Because microscopic examinations of the specimens are performed, it is also important that the **formed elements** (substances in the urine specimen such as cells and microorganisms) are preserved so that they can be documented as part of the examination. Additives may assist with the preservation of these substances as well as the chemical components present in the specimen.

Proper Disposal of Urine and Supplies

After testing, laboratories may immediately discard urine samples, or keep them refrigerated for a few days for potential follow-up testing and troubleshooting of test results. When the urine is no longer needed, it may be flushed down the sink, and the specimen containers may be disposed of as regular (nonbiohazardous) trash.

Urine is not considered a biohazard in the same way that blood and other body fluids are. Sinks used for disposal of urine should be disinfected daily with a concentration of 10% bleach or another approved laboratory disinfectant.

SUMMARY

Urine testing is a common method of screening for disorders of metabolism, pregnancy, and potential infection within the body. Proper patient preparation and education are essential to ensure a quality specimen that will yield meaningful results. Collection methods, required volume, and the need for preservatives will vary, depending on the test ordered. Also, once the urine specimen arrives at the laboratory, it is imperative that the specimen is appropriately processed and preserved in a timely manner to prevent overgrowth of microorganisms and potential erroneous results.

TIME TO REVIEW

1. How is a suprapubic aspiration different from a catheterization? Outcome 9-3
   a. The urine is collected from a different part of the urinary system
   b. The collection device used is different for each procedure
   c. The amount of urine collected is not the same for the two types of procedures
   d. All of the above

2. True or False: The urine collected using the clean-catch midstream collection procedure may be collected at any time of the day. Outcome 9-3

3. Which of these statements is not true of indwelling catheters? Outcome 9-4
   a. Indwelling catheters are inserted and used continuously for an extended period of time
   b. Indwelling catheters are often colonized with bacteria
   c. Indwelling catheters may be used to collect a urine specimen after they have been in place for a few days
   d. Indwelling catheters may be used for patients who have problems with incontinence

4. Which of these measurements may change when urine is left at room temperature for longer than 1 hour? (Circle all that apply.) Outcome 9-5
   a. Color
   b. Glucose concentration
   c. Bacterial count
   d. Appearance of microscopic elements
   e. pH
   f. Specific gravity

5. True or False: It is not necessary to provide two unique identifiers (such as patient name and birth date) on a urine specimen container. Outcome 9-6

6. True or False: Labels for urine specimen containers may be placed on the lid only; it is not necessary to label the container. Outcome 9-6

7. Which of these urine specimen types may be collected at random times? (Circle all that apply.) Outcome 9-3
   a. Clean-catch midstream urine specimens
   b. Specimens to check for prostatitis
   c. Specimens used to check for the presence of glucose as part of a glucose tolerance test
   d. Suprapubic aspiration specimens
   e. Urine pregnancy test specimens
   f. Culture specimens

8. The urinary meatus is: Outcome 9-1
   a. Part of the bladder
   b. The excess skin removed during the process of circumcision
   c. The external opening of the urethra
   d. Part of the prostate gland

9. True or False: Different collection techniques for urine specimens are necessary because chemicals are not present in urine in the same concentration at the same time each day. Outcome 9-2

10. Diurnal variation means: Outcome 9-1
    a. Fluctuations in the value of an analyte that occur during each day.
    b. Concentration of urine in the morning void specimen.
    c. Changes in urinary pH when urine is left at room temperature
    d. None of the above

11. True or False: An aliquot of a urine specimen refers to a small portion (or fraction) of the whole. Outcome 9-1

Case Study 9-1: Postoperative pain

Mr. Johnson has had pain in his perineal area for more than a week, and he has noticed that he is urinating quite frequently. Mr. Johnson decides that he should call the physician for an appointment, because he now has a fever and doesn't feel as if he can go to work. He recently had abdominal surgery, and was catheterized with an indwelling catheter for several days before he was mobile.

1. What are two possible explanations for his discomfort and frequent urination?
2. How could the physician decide on the correct diagnosis? What procedures might be utilized?

Case Study 9-2: Specimen dropoff

Mrs. Brown brings her 24-hour urine specimen container to the office. The medical assistant who is working at the front desk receives the specimen from Mrs. Brown.

1. What does the medical assistant need to do before Mrs. Brown leaves the office?

RESOURCES AND SUGGESTED READINGS

"Prostatitis"
Neal Chamberlain, computerized teaching materials for the Infectious Disease Course at Kirksville College of Osteopathic Medicine http://www.atsu.edu/faculty/chamberlain/Website/lectures/lecture/prostate.htm

"Acute Bacterial Prostatitis"
Overview of prostatitis, with specific information about diagnosis and treatment http://www.drrajmd.com/conditions/prostate/prostatitis/acute_bacterial_prostatitis.htm

"Care Gram 24-Hour Urine and Food Drug Restrictions"
Information about dietary and medication restrictions for 24-hour urine collections from Licking Memorial Health Systems http://www.lmhealth.org/pdf/CareGrams/Lab%20Tests/24%20Hour%20Urine%20&%20Food%20Drug%20Restrictions%201616-0550.pdf

"CAP Today: Cloudy or Clear? Best Forecasts for Urine Cultures," CAP today online newsletter, November 2006
Details for a study about preanalytical errors with urine specimen collection for cultures http://www.cap.org/apps/portlets/contentViewer/show.do?printFriendly=true&contentReference=cap_today%2Ffeature_stories%2F1106Urine.html

"24-Hour Urine Collection"
Southwestern University Health Library selection about 24-hour urine collection; includes clinical significance and details about the collection process http://www.utsouthwestern.edu/patientcare/healthlibrary/healthtopics/0,,P08955,00.html

"Best Practice: Tips for Urinalysis"
Becton, Dickinson online newsletter, LabNotes, 13, no. 2, 2003
Details about urine processing for urinalysis and urine cultures, including information about preservative tubes that are available http://www.bd.com/vacutainer/labnotes/2003spring/best_practice.asp

"Sample Meal for a 2-Hour Postprandial Glucose Test"
Patient instructions for 2-hour postprandial glucose testing procedure from the Delaware Valley Institute of Fertility and Genetics http://www.startfertility.com/patient_fld/form_pdfs/Sample_Meal-2-hr_PP_glucose_test.pdf

"Laws, Regulations, Standards and Guidelines"
A publication of the New Jersey Hospital Association with a summary of laws, regulations, standards, and guidelines that pertain to language assistance for diverse populations in the health-care setting that may not have English as their native language http://www.njha.com/publications/pdf/Laws_Regulations_Standards_and_Guidelines.pdf

Chapter 10

Collection and Processing of Samples for Microbial Studies

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Microbiology Sample Collection Guidelines
  - Types of Media
  - Microbiology Terms
- Detailed Microbiology Sample Collection Procedures
  - Throat Swab Collections for Culture or Strep Screens
  - Sputum Samples
  - Urine Samples for Culture
  - Blood Culture Collection
  - Cerebrospinal Fluid Samples
  - Genital Samples
  - Wound Cultures
  - Stool Specimens
  - Nasopharyngeal Specimens
  - Eye Cultures
  - Ear Cultures
- Special Sample Collection and Processing Procedures
  - Fungal Sample and Culture Collection Procedures
    - *Potassium Hydroxide Preparation*
  - Wet Mount Procedure
  - Sample Collection Procedures for Detection of Parasites
    - *Ova and Parasite Examination*
    - *Pinworm Collection Procedures*
- Processing Microbiology Samples
  - Slide Preparation and Gram-Staining Procedure
  - Plating and Inoculation of Media
  - Antibiotic Sensitivity Testing
- Microbiology Test Results
- Summary
- Time to Review
- Case Studies
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

10-1 Define the key terms.

10-2 Demonstrate understanding of the general guidelines to be followed when collecting for microbiology testing.

10-3 List various types of media used for bacterial and viral specimen setup.

10-4 Describe how the collection processes and supplies for bacterial, viral, and fungal specimens may differ.

10-5 Compare and contrast sample collection of throat swabs for streptococcal screens and cultures.

10-6 Explain the collection procedure for sputum samples.

10-7 Differentiate the types of urine specimens that may be used for cultures.

10-8 Describe the procedure for collecting a specimen for a blood culture.

10-9 Describe how a cerebrospinal fluid (CSF) sample is collected, and explain how the sample should be processed after collection.

10-10 Explain how genital screening samples are processed.

10-11 Demonstrate an understanding of the various types of wound cultures and how their processing differs.

10-12 Compare and contrast collection and processing of stool samples as compared to other common microbiology specimen types.

10-13 State when a nasopharyngeal specimen may be necessary, and describe the collection technique used.

10-14 Describe the technique used for eye and ear culture sample collection procedures.

10-15 Classify the various types of fungal infections, and describe the techniques used to collect fungal culture samples.

10-16 Differentiate the requirements and use of a KOH prep and a wet mount specimen preparation.

10-17 Explain how a stool sample is collected and preserved for an examination for the presence of ova and parasites.

10-18 Summarize the sample collection process used for detecting a pinworm infection.

10-19 Explain the steps involved in performance of a Gram stain, as well as the differential appearance of the stained organisms.

10-20 Explain how a microbiological sample is plated for growth.

10-21 Describe how sensitivity testing is performed and explain the clinical significance of the process.

10-22 Translate a microbiology laboratory report, and explain the clinical significance of the values.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.A.I.2. Use language/verbal skills that enable patients' understanding
III.C.III.3. Discuss infection control procedures
III.C.III.5 List major types of infectious agents
III.C.III.7. Match types and uses of personal protective equipment (PPE)
III.C.III.9. Discuss quality control issues related to handling microbiological specimens
III.P.III.2. Practice Standard Precautions
III.P.III.3. Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations
III.P.III.7. Obtain specimens for microbiological testing
III.A.III.1. Display sensitivity to patient rights and feelings in collecting specimens
III.A.III.2. Explain the rationale for performance of a procedure to a patient
III.A.III.3. Show awareness of patients' concerns regarding their perceptions related to the procedure being performed

ABHES Standards

- Apply principles of aseptic techniques and infection control
- Use standard precautions
- Practice quality control
- Perform selected CLIA-waived tests that assist with diagnosis and treatment, #5: Microbiology testing
- Dispose of biohazardous materials
- Collect, label and process specimens, #3: Perform wound collection procedures, #4: Obtain throat specimen for microbiology testing
- Instruct patients in the collection of a fecal specimen

KEY TERMS

Aerobes
Agar
Anaerobes
Antibiotic sensitivity testing
Broth
Candle jar
Cerebrospinal fluid (CSF) samples
Colonies
Cover slip
Culture
Culture and sensitivity (C&S) report
Disk-diffusion method
Ectoparasites
Endoparasites
Eustachian tube
Exudate
Gram stain
Hanging drop slide
Hemolytic uremic syndrome
Inoculation
Kirby-Bauer method
KOH (potassium hydroxide)
Lumbar puncture
Medium (pl., media)
Methicillin-resistant *Staphylococcus aureus* (MRSA)
Minimum inhibitory concentration (MIC) method
Motility
Mycoses
Nasopharyngeal collection technique
Nasopharynx
Normal flora
Otitis
Otitis externa
Otitis media
Ova and parasite examination (O&P)
Parasites
Pathogen
Petri dish
Pinworm
Resistant
Scabies
Sensitive
Septicemia
Sequela
Slant
Sputum
Strep screens
Trichomoniasis
Wet mount
Wood's lamp
Zone of inhibition

Microbiology is the study of organisms that require magnification to be seen by the human eye. As introduced in Chapter 3, microorganisms are everywhere! These tiny creatures have been the objects of study since the early 1700s, when Antony Leeuwenhoek first recorded observations of bacteria he watched with his homemade microscopes. These "animalcules" (as Leeuwenhoek called them) were described in great detail, and Leeuwenhoek hired an illustrator to provide pictures of the living creatures that he observed. However, the idea that microorganisms were capable of spreading disease was understood much later. In 1847, Ignaz Semmelweis was working in a maternity hospital in Siena, Italy, when he realized that some sort of "invisible agent" seemed to spread from the mothers who died of puerperal fever after childbirth to the other mothers in the ward. He insisted that the physicians working in the ward begin to wash their hands between patients, and also had the practitioners change their laboratory coats after performing autopsies before they worked with expectant mothers. These small changes resulted in a dramatic drop in the mortality rates in the maternity ward of the hospital. Unfortunately, Semmelweis was ridiculed for his strict enforcement of the rules, and was subsequently fired.

It would still be almost 100 years before observations such as the ones made by Semmelweis began to change the way that the world viewed microorganisms and their relationship to disease. Louis Pasteur and Robert Koch were responsible for many of the theories that we use in microbiology today. In the late 1800s, they conducted a series of experiments to suggest that bacteria cause disease, and these scientists went even further to prove that bacteria did not spontaneously generate, but instead replicated at a very fast pace when introduced to an environment that supported their growth. It was also proved during this time that bacteria could be kept from contaminating a certain area if the appropriate precautions were taken. These were revolutionary ideas that continue to shape the way we view microorganisms today.

When working with microbiological samples, it is important to keep these core concepts in mind, as our goal is to avoid extraneous contamination of the specimens, to provide appropriate samples for testing to allow for proper patient diagnosis and treatment, and to nurture the growth of microscopic organisms so that they can be identified. We now know that the microscopic "animalcules" that can cause harm do not include only bacteria. Microbiology testing may also focus on viruses, fungi, and parasites. Because of this variety, collection methods and testing procedures vary, depending on the type of specimen collected and the organism that is thought to be present. The laboratory may process these samples in the general microbiology department, or there may be subdivisions for bacteriology, virology, mycology, and/or parasitology. This chapter presents collection and handling procedures for the most common types of microbiological tests ordered.

MICROBIOLOGY SAMPLE COLLECTION GUIDELINES

In order to provide appropriate treatment, health-care providers need to know which type of pathogen is causing the symptoms exhibited by the patient. A **pathogen** is a microorganism that is capable of causing disease in the body. Most infectious diseases are caused by viruses or bacteria. Treatment choices vary dramatically based on the causative organism present in the specimen collected for testing. Variables for treatment may be based on the type of pathogen present (such as viruses, bacteria, fungi, or parasites), the amount of the microscopic organism isolated, as well as sensitivity testing for antimicrobials suitable for treatment. The microorganism that causes a disease is a pathogen, so health-care providers must always remember which organisms may be present as part of the **normal flora** (resident bacteria present in certain areas of the body) as these may contaminate the specimen, and their presence may be misleading.

Test Your Knowledge 10-1

What is the term used to describe a microorganism that causes disease?. (Outcome 10-1)

To assure quality processes and valid results, specific guidelines should be followed whenever collecting or processing samples for microbiology testing. These include the following:

Specimen collection

- Collect the sample from the area of infection, and avoid contamination from surrounding areas if at all possible. This is especially critical when collecting samples for wound cultures.
- Ideally, bacterial culture specimens should be collected before antibiotics are introduced. If the antibiotics have already been administered, the causative organism may not be present in high enough numbers to identify it appropriately.
- Use the correct collection and transportation methods. Research what type of swab or container is to be used, and how the sample is to be handled after collection until it reaches the laboratory. Package inserts should be consulted for specific collection techniques. Some samples are to be refrigerated immediately, whereas others are to remain at room temperature until they are transported to the laboratory for processing.
- Always collect a sufficient quantity of sample for testing.
- Be certain that the sample is appropriately labeled, including the source (be as specific as possible), the patient's name, the identification of the individual performing the specimen collection, and the date and time of the collection. Document this information on the requisition form appropriately as well.
- Verify and adhere to any sample restrictions for transport time. Also verify whether the sample needs to be protected from light after collection.
- Use sterile supplies, and be careful to maintain sterility while performing sample collection procedures.
- Store all laboratory specimens in a refrigerator or other container dedicated for that use.

Personnel

- All personnel participating in the collection, processing, or testing of microbiological specimens must be appropriately trained in all procedures.
- Personnel should participate regularly in competency testing to assure quality of collection, processing and testing of specimens.
- Wear appropriate personal protective equipment (PPE) when performing specimen collection or specimen processing. This may include items such as gloves, a laboratory coat, face shield, or special respiratory equipment such as that needed for processing samples that may contain tuberculosis.

Media

- A medium provides nutrition for microorganisms when grown in the laboratory environment. There is an expiration date associated with all media to ensure appropriate performance. Media should be discarded when outdated.
- For a medium to function as designed, it cannot be allowed to dry out or become contaminated. In addition, there are specific storage requirements for each type of medium, and these must be followed carefully.
- Verify that the correct medium is used for each type of specimen collection, as this can have a profound effect on the results.
- Medium performance should be verified on a regular basis by inoculating the medium with a known microorganism to observe growth.

Stains

- All products used for staining microorganisms have expiration dates provided by the manufacturer, after which the stains must be discarded. Stain performance should be verified on a regular basis by staining microorganisms with predictable staining characteristics to observe staining patterns.

Test Your Knowledge 10-2

What must be documented on the sample after the collection of a microbiology specimen? (Outcome 10-2)

Test Your Knowledge 10-3

Should bacterial culture specimens be collected before or after antibiotic therapy has been started? (Outcome 10-2)

Types of Media

The growth of any microorganism in the laboratory is dependent on the presence of an appropriate nutrition source and other chemicals necessary for survival. These are provided in a culture **medium** (plural, **media**). The medium may be in the form of a liquid nutrient broth, a semi-solid gel, or a solid gel. **Agar** is a thickening agent added to the medium that allows the gel to solidify so that the microorganisms will grow on the surface. A medium may be provided in tubes, or it may be provided in a small shallow container, known as a **Petri dish.** Figure 10-1 shows various types of media used in a microbiology setting. Some laboratories create their own media, but most purchase commercially prepared media, which provide a longer shelf life, fewer impurities, and more consistent appearance and performance. All types of media contain nutrients and additives that support the growth of microorganisms. Ingredients may include blood, vitamins, dyes, sugar mixtures, and organic sources of nutrition such as vegetable extracts.

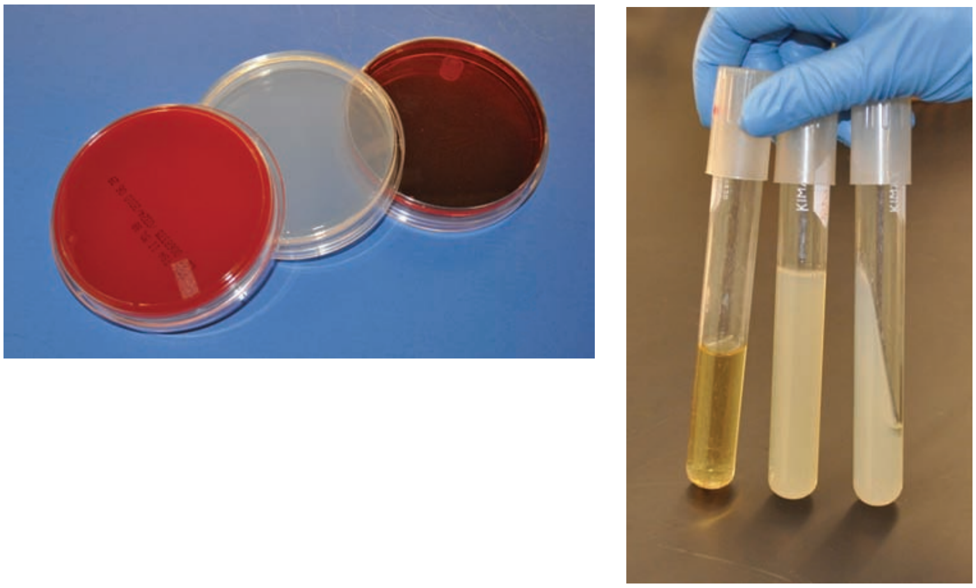

Figure 10-1 Types of media. Note the presence of the agar plate, the slant, and the broth.

Culture media is classified into categories based on whether the medium enriches the growth of all bacteria or selectively inhibits growth of certain bacteria.

- *Selective media* have additives that prevent the growth of certain types of bacteria while promoting the growth of other select species. One example is Thayer Martin agar, used for specimens when a gonorrhea infection is suspected. It has lysed red blood cells, which provide an easy source of hemoglobin to promote the growth of the *Neisseria gonorrhoeae* organism, but the medium also contains vancomycin, colistin, and nystatin. These are antibiotics that prohibit the growth of other common nonpathogenic microorganisms often found concurrently with gonorrhea.
- *Differential media* allow for different types of bacteria to be identified based on an indicator. A common differential medium used in the laboratory is MacConkey agar, which has a special dye additive. If the organisms growing on the agar ferment lactose (the indicator), their growth will appear pink on the medium plate. This colored growth allows the microbiologists to make a preliminary identification of the pathogenic microorganisms based on their ability to ferment lactose.
- *Enriched media* include nutrients that allow for a wide variety of organisms to grow. Blood agar is an example of an enriched medium that is designed to support growth of microorganisms that flourish in the human body, including both pathogens and nonpathogens.

Solid gel is provided in a tube (called a **slant)** or in a Petri dish. Liquid medium (usually called nutrient **broth)** comes in a tightly sealed tube. It is important to keep all forms of media from drying out or becoming contaminated after receipt at the laboratory or physician office. The media should be refrigerated for storage, but must be brought to room temperature before use. All forms of media have a limited shelf life, so it is important to be aware of their expiration date.

Test Your Knowledge 10-4

What are two ways that culture media may differ?

(Outcome 10-3)

Microbiology Terms

A **culture** is a group of microorganisms grown in a laboratory environment using prepared media. The word *culture* is used whether the sample is bacterial, viral, or fungal. A culture is created when a medium source has a specimen added for potential growth. The culture is the most common test performed in the microbiology laboratory. The process of adding the specimen to a medium source is known as **inoculation.** Inoculation may occur in the physician office immediately after specimen collection, or it may be performed once the specimen arrives at the laboratory. The visible growth that occurs once the specimen has been added to the medium may be seen as **colonies** (or clumps) of growth, especially with bacterial cultures. The unique appearance of the colonies present on the culture may allow for preliminary identification of the microorganism. To accelerate bacterial growth, the specimen is usually placed in an incubator that is kept near body temperature (37°Celsius) for 24 to 48 hours. Some organisms must be placed in an environment with little to no oxygen present for growth **(anaerobes),** whereas others like an oxygen-rich environment **(aerobes).** There are incubators or other devices that are specifically designed to provide these parameters for culture growth.

Preliminary laboratory results for bacterial cultures are often available within 24 to 48 hours after the specimen has been submitted to the laboratory. Each laboratory establishes a definitive time frame before which they will not report a negative result for a bacterial culture. This means that the laboratory gives the sample a chance to grow for at least a few days before posting a negative result. Most pathogenic bacteria are relatively easy to grow in the laboratory, and identification techniques have been established for most of these microorganisms.

Viral cultures require a longer time for growth and identification, and require use of different types of media than those required for bacterial cultures. Because viruses are parasitic intracellular organisms, viral cultures require living tissue and cells to support growth and replication. It is imperative that the person collecting a viral specimen for culture pay scrupulous attention to specimen collection protocol, the technique of collection, and transportation requirements. Viral culture specimens should be collected, in general, within 72 hours of symptom onset. Collection swabs made of calcium alginate should not be used for collection of viral culture specimens, and cotton-tipped swabs with wooden shafts must also be avoided. Generally, viral specimens are refrigerated to preserve their integrity if there will be a delay in transportation. Because it is a complicated process, viral culture processing is usually performed only in large reference laboratories. It may be 7 to 10 days before a viral culture result is confirmed. For this reason, if a viral infection is suspected and the patient's symptoms warrant it, the health-care provider may administer antiviral medication to the patient without a culture result or identification of the organism.

Fungal cultures are even slower to grow than viral cultures. The fungal elements in a specimen may be difficult to isolate because of the faster-growing organisms present in or on the specimen, such as bacteria. A selective media is used, including nutrition sources such as brain and/or heart extract, blood, dextrose, and natural nutrition sources such as potatoes. Antibiotics may be added to retard the growth of the bacteria in the specimen so that the fungal elements can be identified. The exact type of media used will be dictated by the specific source of the specimen within the body, as, for example, the types of fungus growing on the skin may be very different from those growing in the lungs. A fungal culture may take up to 4 weeks to report results; it is generally a minimum of 21 days before a negative result can be released.

Test Your Knowledge 10-5

Which type of culture (viral, bacterial, or fungal) is quickest to grow? (Outcome 10-4)

POINT OF INTEREST 10-1
Bioterrorism

Prior to September 11, 2001, the threat of widespread bioterrorism was a rare topic in most casual conversations. However, the climate changed with the events of that day and the weeks that followed. The Centers for Disease Control and Prevention (CDC) strongly recommends that all laboratories (even those in physician offices) continue to monitor their procedures and remain aware of those that could create aerosols. An aerosol is a suspension of fine particles (such as bacteria or viruses) in the air. For those facilities in which the risk of aerosol formation is heightened because of the types of procedures performed and the patient population served, a biosafety hood is recommended. The procedures would be performed within the hood workspace, and any microorganisms released into the air would be vented out of the room.

Health-care facilities also must remain educated about what to do in case of an exposure within their facility; fast, appropriate action can save lives. The agents may be bacterial or viral in nature. Most, but not all, agents that could be used for bioterrorism would be spread through the aerosol form. It is difficult to predict which agents may be used for a biological attack, but the following are some that are considered to be most probable:

- **Anthrax:** In the 1800s, Louis Pasteur documented work that he performed with this bacterium, and there are reports of soldiers using anthrax in World War I to deliberately infect animals that were to be eaten by the enemy. In October and November 2001, there were 22 cases of deliberate infection with the anthrax bacterium in the United States. Anthrax is caused by *Bacillus anthracis* and is a gram-positive, spore-forming bacillus. In nature, the spores may be present in the soil, and they are either ingested or picked up on the skin and hair of animals when they are feeding. Humans usually become infected only when they come in contact with animals or products that are contaminated, and infection is rare. There are three routes of infection for anthrax:
  - *Cutaneous:* The most common natural route is cutaneous, or via the skin; humans may be infected when handling contaminated wool, or occasionally from the spores in the soil. The cutaneous infection causes a papule that progresses with time into a necrotic ulcer. If the ulcer penetrates deep enough into the skin, the patient's bloodstream may become infected, which can lead to more serious consequences.
  - *Inhalation:* The route that is used most often for deliberate infection is via inhalation. The exposed patient has flu-like symptoms. Within days, the patient suffers from hypoxia and has difficulty breathing, as pulmonary edema and further damages within the lungs progress. Approximately 50% of those who are infected via the inhalation develop meningitis, and the morbidity rate is quite high. Prior to 2001, there had been only 20 cases of anthrax infection via inhalation reported in the United States.
  - *Gastrointestinal*: The ingestion of poorly cooked, contaminated meat allows anthrax entrance via the gastrointestinal route. The patient may have abdominal discomfort, bloody diarrhea, and vomiting, and may have visible ulcerations in the oropharynx when examined. Essentially, the necrotic lesions that are visible with cutaneous exposure form within the GI tract and impair the patient's ability to digest food properly.

All forms of anthrax exposure can be treated with antibiotics if caught early. Once the more serious symptoms have developed, it may be difficult treat successfully. Ciprofloxacin or doxycycline is often used successfully for early treatment. Anthrax is not passed from person to person.

- **Brucellosis:** Brucellosis is a disease that is common in cattle, sheep, goats, and pigs in many areas of the world. It may also affect domestic animals such as dogs. Humans may be infected by direct exposure to an animal host. There are very few human infections in the United States because of a rigid vaccination program, but the disease is much more frequent in other areas of the world. Infection causes flu-like symptoms in most humans, but can lead to much more serious CNS problems, as well as heart damage. Brucellosis forms an aerosol very easily, so to avoid potential inhalation in the workplace, it is important to remember not to "sniff" cultures in the microbiology area. This practice may be very dangerous.

- **Botulism:** The symptoms associated with botulism are related to the toxin produced by the spore-forming bacterium *Clostridium botulinum*. This toxin is one of the most lethal natural substances known to humans. The toxin paralyzes the muscles of those infected, which eventually leads to paralysis of the lungs, and death. The toxin is not used as an aerosol as much as some of the other microorganisms considered as threats for bioterrorism use, but the bacteria could be very lethal if added to a common food source or used in a contained area. *C. botulinum* may also be introduced into the body through a wound; there is evidence that armies in the past added it to their swords to enhance the damage done during hand-to-hand combat. For those infected, the most important consideration is the treatment of symptoms, as the damage caused by the toxin spreads quickly throughout the body.
- **Plague:** Everyone who has studied world history has heard of the plague in medieval Europe. The plague is caused by the bacteria *Yersinia pestis* and may infect humans through two different routes. One is the bubonic plague, which is usually caused by the bite of a flea or tick carried on a host rodent. The other is through inhalation. This is the method that causes concern for potential use of this bacterium as a bioterrorist weapon. The pneumonic plague may be spread from person to person via droplets in the air after the initial infection, and is relatively easy to isolate and grow in large amounts. Thus, initial infection could lead to wider and wider mortality. Infection may easily cause respiratory failure and death. The bacterium is treatable with tetracycline and/or fluoroquinolone for at least 7 days if given orally, or IV streptomycin or gentamycin, but the infection must be caught early.
- **Smallpox:** Smallpox is different from the other microorganisms that are considered bioterrorist threats because it is a virus instead of a type of bacterium. The last natural case of smallpox in the United States occurred in 1949, and the causative agent (the variola virus) is considered to have been eradicated by a long-term, aggressive, and worldwide vaccination campaign that has been carried out since 1980. There is no specific treatment for the smallpox virus infection except for supportive care of the symptoms. The variola virus appears to infect only humans, and it may be passed from person to person through infected body fluids and contaminated bedding. It is very rarely passed through the aerosol route, and is not naturally present in nature, so use of the virus as a bioterrorist tool would require more effort than some of the other agents discussed here.
- **Hemorrhagic fever viruses:** These viruses are carried by a host or vector, usually a rodent (such as a rat or mouse) or a tick, flea, or mosquito. For some of the viruses (Ebola and Marburg, for instance) the host is still unknown. Once a human has been infected, it is possible to spread the virus from person to person through contaminated body fluids. Outbreaks are sporadic around the world, and usually remain limited to the area where the vector or host is located, but this has the potential to change. Because of the ease of travel today, coupled with the fact that rats and mice can "stow away" on ships and planes, it is possible for these diseases to become widespread within a relatively short period of time. Those who are infected have flu-like symptoms early in the disease process. Then as the disease progresses, there may be evidence of hemorrhaging under the skin, and bleeding from all orifices of the body. Multiple organs are affected by the virus, and it usually is the organ failure that causes fatalities, not the loss of blood. There are no vaccines or cures for these infections at this time; infected individuals are quarantined and treated symptomatically.
- **Tularemia:** Tularemia is caused by the bacterium *Francisella tularensis* and is carried often by rodents, rabbits, or wild hares. It is *very* infectious; only a small number of the bacteria entering the body will cause disease. It is easy to spread this bacterium as an airborne weapon, so it would be a likely choice for bioterrorist activities. It causes a severe respiratory illness and life-threatening pneumonia, and is not difficult to find in nature. It is treatable with early intervention and aggressive antibiotic treatment.

DETAILED MICROBIOLOGY SAMPLE COLLECTION PROCEDURES

Procedures and supplies may vary for the different microbiology collection processes, but there are common techniques shared by all the systems. Specimens may be collected in a sterile cup or tube or, more commonly, on a swab to be transported to the laboratory. These collection swabs are commercially prepared and supplied in a sterile pull-apart package. The package includes a sterile swab (often made of Dacron) and a transport medium designed to keep the swab from drying out before it reaches the laboratory. Culturette (manufactured by Becton, Dickinson) is an example of a system used for transport. Figure 10-2 includes some of the common sample collection devices available. The following are some techniques to be used for all culture collections:

- Be certain that a laboratory request form has been completed for the patient. The specific source of the specimen must be documented.
- Verify the collection technique necessary for the test ordered, as well as how the specimen is to be handled after collection: refrigeration, room temperature storage, frozen. Most bacterial culture specimens are to be kept at room temperature until transported.
- Always sanitize your hands, and wear appropriate personal protective equipment (PPE). In some situations, this may be a face shield or a disposable gown that covers your scrubs. In patients who have a wound that is positive for methicillin-resistant *Staphylococcus aureus* (MRSA), for example, it is often necessary to take extra precautions and to dispose of all the PPE used in collection before leaving the presence of the patient.
- If using a swab, check the package for signs of potential contamination (e.g., water stains, torn outer packaging) and verify that the expiration date has not passed. If using a different type of collection container, ensure that the package is sterile by examining the outer wrapper or safety seal.
- For swab packages, carefully peel apart the plastic outer envelope, using aseptic technique. Put this down on a flat surface. Pick up the swab by the lid that is attached to the end of the "stick" end of the swab. (The lid is permanently attached to the swab in most cases.) It is very important that the swab does not touch any other surfaces before the specimen is collected, or extraneous organisms may be added to the culture specimen.
- Collect the specimen appropriately, avoiding any other areas of the body except those needed for the collection.
- Without setting down the inoculated swab, remove the cap from the collection tube (still sitting inside the sterile plastic package) and insert the inoculated swab fully into the collection tube until the lid clicks closed. It may be necessary to activate the transport medium present in the tube by breaking an ampule at the bottom of the tube; this will depend on the type of collection device used.
- Remove gloves and other PPE that was used in the collection process, and sanitize the hands.
- Label the specimen carefully, documenting the specimen source, patient's name, the date, the time, the source (be specific), and your initials or identification. Chart the procedure in the patient's chart if it is available to you.
- If the specimen is to be transported via courier to a laboratory for workup, place it in a biohazardous specimen bag. The requisition form goes into the outer pocket of the bag. If the bag has an area for documentation for handling and storing the specimen, mark this appropriately (e.g., room temperature, refrigerated).
- Place the specimen in the pickup area for the courier to transport to the laboratory.

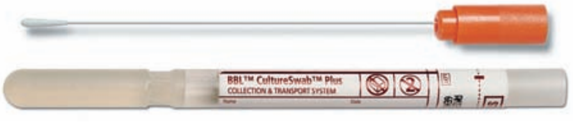

Figure 10-2 Swab collection devices. *Courtesy of BD, Inc.*

Test Your Knowledge 10-6

When using a swab to collect a culture specimen, is it acceptable to set it down on a countertop while getting the patient situated for the collection procedure?

(Outcome 10-2)

Throat Sample Collections for Culture or Strep Screens

Samples may be taken from the throat of patients for group A strep screens for general cultures, or sometimes both types of specimens at the same time. For both a strep screen sample and a culture sample the collection procedure is identical, but the materials used for the collection may differ. Before beginning the collection, it is important to verify whether the physician wants a culture to be performed, a strep screen, or both. Once the strep screen is performed, it is not possible to use the sample for the culture setup. If both a culture and a strep screen are ordered, the culture plate may be inoculated before the strep screen is performed, but not afterward. It is also possible to use two swabs (these are packaged as a double swab) at the same time to collect the sample so that one can be used for each procedure. Be careful about the type of collection swab used; strep screens recommend the use of rayon collection swabs. Cotton swabs or wooden tipped swabs are not to be used for strep screens. Some of the CLIA-waived procedures don't recommend the use of liquid or gel transport media after collection.

Throat samples are used most commonly to test for streptococcal infections with a **group A strep screen** rapid test completed in the physician office. Group A streptococcus (a subtype of *Streptococcus pyogenes)* is the most common cause of streptococcal throat infection. Those who are positive for streptococcal infection of the throat should be treated immediately to avoid sequela that can develop as a result of the initial strep infection. A **sequela** is a serious secondary complication or infection that follows as a result of the primary bacterial infection. For a group A streptococcal infection, these complications may include heart damage and kidney damage from glomerulonephritis. Throat specimen collections are performed using a sterile swab, often made of Dacron, rayon, or a cotton mixture. For a throat culture, the sterile swab is part of a packaged culture collection kit, with transport media usually present in the transport tube. The throat culture specimen is transported to the laboratory for processing, whereas a strep screen sample is usually processed immediately after collection by performing a CLIA-waived strep screen test. A sterile swab provided with the strep screen kit is used for the collection, and there is no transport medium present. Remember to verify whether the physician wants both a strep screen and a culture to be performed, as the medical assistant needs to know this information before performing the strep screen.

Test Your Knowledge 10-7

What is one way that a collection for a throat culture and a rapid strep screen are similar? (Outcome 10-5)

Strep screens are one of the most common CLIA-waived tests performed in all types of laboratories. The test is simple to perform, and requires only a sample taken from the back of the throat, a tongue depressor, personal protective equipment, and the supplies in the kit. The entire process takes approximately 10 minutes from the time the order is received to the time the result may be reported. This fast turnaround time and the ease with which the test can be performed allow the patient to be treated immediately if the screen is positive for the streptococcus A antigen. The CLIA-waived group A strep screen procedure is presented in Chapter 24.

Test Your Knowledge 10-8

Is it acceptable to collect only one swab if a strep screen *and* a throat culture are ordered? (Outcome 10-5)

Test Your Knowledge 10-9

Are the samples for a strep screen and for a throat culture collected from different areas of the throat? (Outcome 10-5)

Sputum Samples

When patients are diagnosed with a lower respiratory infection, such as pneumonia, it is important to identify the causative organism if possible. This allows the health-care provider to plan the best course of action for treatment. Sputum samples are necessary for culture and organism identification in these situations. **Sputum** is the substance produced from a deep cough or from aspiration caused by inserting a tube down the patient's throat. Aspiration of sputum is primarily performed on patients who are intubated in an inpatient setting, but sputum cultures may be collected by patients at home without aspiration.

The medical assistant is not usually involved with the actual collection process for a sputum sample if employed in a physician office, but it is important that the procedure is explained thoroughly to the patient so that he or she may produce a quality sample. Sputum samples are collected into a sterile container (such as a urine cup or a centrifuge tube) and transported immediately to the laboratory for processing. Ideally, a sputum collection system will protect the sample from UV light. The system should also prevent the health-care employees processing the specimen from being exposed to the sputum. Becton, Dickinson offers a sputum collection system that does both, composed of a large centrifuge tube enclosed in an amber "funnel" tube with a lid on the top (Fig. 10-3). The amber outer container is important, as *Mycobacterium tuberculosis,* the causative

Procedure 10-1: Collection of a Throat Specimen for Culture or Strep Screen

Throat culture collections are very common procedures in physician offices. It is important to perform the collection process correctly to ensure meaningful results. Remember to communicate appropriately with the patient to assist with the collection process.

TASK

Collect an appropriate specimen from the throat for a culture or strep screen. Complete the procedure within 5 minutes.

CONDITIONS

- Hand sanitization equipment
- Gloves
- Tongue depressor
- Safety shield
- Sterile culture swab or swab provided with strep screen kit
- Label
- Laboratory requisition form
- Biohazardous waste container
- Biohazardous transport bag

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.A.I.2. Use language/verbal skills that enable patients' understanding, **III.C.III.3:** Discuss infection control procedures, **III.C.III.9:** Discuss quality control issues related to handling microbiological specimens. **III.P.III.2:** Practice Standard Precautions, **III.P.III.3:** Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations, **III.P.III.7:** Obtain specimens for microbiological testing, **III.A.III.1:** Display sensitivity to patient rights and feelings in collecting specimens, **III.A.III.2:** Explain the rationale for performance of a procedure to the patient.

ABHES Standards

- Apply principles of aseptic techniques and infection control
- Use standard precautions
- Practice quality control
- Perform selected CLIA-waived tests that assist with diagnosis and treatment, #5: Microbiology testing
- Dispose of biohazardous materials
- Collect, label and process specimens, #3: Perform wound collection procedures, #4: Obtain throat specimen for microbiology testing

Procedure	Rationale
1. Greet and properly identify the patient, using at least two unique identifiers. Introduce yourself.	It is imperative that the identity of the patient is verified before the specimen is collected. The medical assistant should always identify him- or herself with any patient interaction.
2. Explain the procedure to the patient. Answer any questions that the patient may have about the collection process.	Explaining the procedure will help the patient know what to expect and will facilitate cooperation. Answering questions allows the medical assistant to display sensitivity to patient rights and feelings in collecting specimens.
3. Sanitize hands (allow them to dry) and apply gloves.	Gloves must be worn for any procedures in which exposure to blood or other potentially infectious materials is anticipated.

Continued

Procedure 10-1: Collection of a Throat Specimen for Culture or Strep Screen—cont'd

Procedure	Rationale
4. Gather necessary supplies within reach. If a strep screen is ordered, be certain to use the swab supplied with the kit. If a throat culture is requested, use a sterile swab packaged with the transport media.	Organization of supplies is important so that the collection procedure may continue without interruption.
5. Position the patient in an upright position. If necessary, adjust available lighting so that the back of the throat is well lit.	The back of the throat must be well lit and accessible for the collection to be successful.
6. Put on face safety shield.	Because this procedure may initiate coughing or gagging in some individuals, it is important that the medical assistant protect against potential exposure.
7. While holding the tongue depressor in one hand and the collection swab in the other, instruct the patient to open his or her mouth wide.	The mouth must be opened wide to gain full access to the collection area.
8. Push down on the tongue lightly with the tongue depressor, and insert the swab into mouth. Be certain to avoid touching the tongue, cheeks, gums, or teeth with the collection swab.	Introduction of extraneous microorganisms may result in inaccurate culture results.
9. Gently brush the swab against the tonsillar area and the back of the throat. A gentle rotating motion works best. Be sure to touch any white or inflamed patches present with the collection swab.	Adequate sample must be obtained for the accuracy of the culture results. White areas may represent clusters of bacteria, so it is important that these are included in the sample.
10. After removing the swab from the mouth, place it in the sterile container. If only a throat culture is requested, slide the swab into the container and activate the preservative or coat the swab. If a strep screen was ordered, only rayon swabs may be used for the collection, and liquid preservative is not recommended in the swab container.	It is important that the swab be processed appropriately so that the results are representative of the bacteria present in the throat. CLIA-waived strep screen package inserts will specify whether liquid preservatives (such as Amies or Stuart media) can be used. Rayon swabs with dry transport are recommended for strep screen specimen collection and transport.
11. Dispose of the tongue depressor in the biohazard garbage.	Proper disposal is necessary for quality control purposes.
12. Label the specimen appropriately and place the labeled specimen in a biohazardous transport bag to go the laboratory.	The specimen should always be labeled in the presence of the patient.
13. Remove gloves and safety shield.	Gloves should be removed when no longer necessary.

Procedure	Rationale
14. Wash hands, and thank the patient for being cooperative. Add documentation to the laboratory requisition form if needed.	Hands must be washed whenever gloves are removed. The requisition form must be completed with the date and time of collection as well as the source of the culture.
15. Chart the procedure appropriately.	All patient interactions should be recorded on the chart.

Date	
1/18/2014	Throat culture specimen collected. Connie Lieseke, CMA (AAMA)
11:10 a.m.	

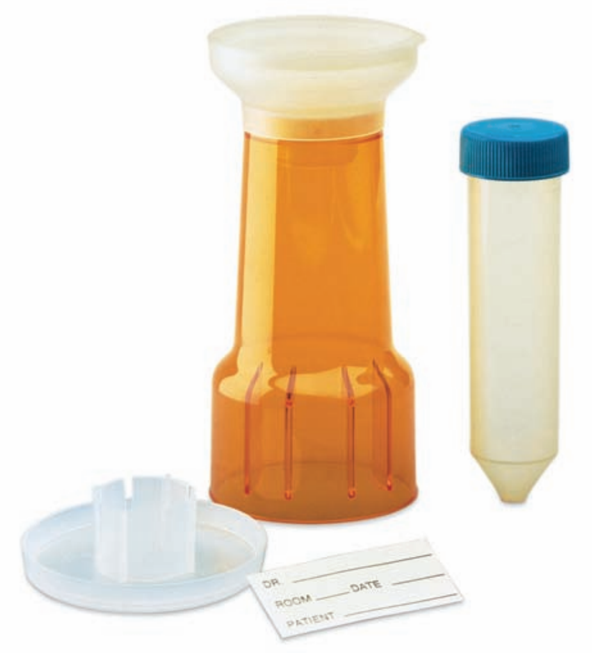

Figure 10-3 BD sputum collection container. The inner tube may be removed from the bottom of the container with minimum exposure potential for the employee. *Courtesy of BD, Inc.*

agent of tuberculosis infection, does not live long when exposed to light. The specimen must remain protected from light until it is processed in the laboratory, or this pathogenic microorganism may not be isolated from the culture even though it is present. Once the specimen arrives at the laboratory, the technician will remove the inner "tube" that contains the sputum and process the specimen according to protocol to set up the culture.

Patients collecting sputum samples must be instructed that a deep cough is absolutely necessary to produce an acceptable specimen. If saliva rather than sputum is submitted for testing, the specimen will most likely contain only normal flora for upper respiratory system, and the culture results will be inconclusive. It is recommended that the sample be obtained first thing in the morning, after the patient brushes his or her teeth and rinses the mouth, but before ingesting food or drink. Because the pathogenic microorganisms responsible for many lower respiratory infections are spread by droplets in the air, it is necessary that the patient perform collection in a private area, preferably at home. Do not allow a patient to attempt to produce a sputum culture specimen in the waiting room of the physician office, as this will potentially expose the other patients to the pathogenic microorganisms distributed as aerosol when the patient coughs. The sample must be delivered to the laboratory as soon as possible after collection, preferably within 2 hours. The specimen should not be refrigerated, nor exposed to extreme heat.

Test Your Knowledge 10-10

Why should a sputum specimen *not* be collected in the waiting room of a physician office? (Outcome 10-6)

Once the specimen arrives at the laboratory, the technician (or sometimes, in an inpatient setting, a respiratory specialist) will decide whether the specimen is acceptable for culture. The technician also documents the volume and appearance of the specimen. A sputum sample that has too many epithelial cells present and is too "watery" in appearance is often a saliva sample; it was not the result of a deep cough and is not suitable for culture setup. Sputum color and thickness are not indicators of any specific illness, and this information should never be interpreted in this way.

Urine Samples for Culture

Urine specimens used for culture must be collected in a manner that reduces potential contamination by nonpathogenic microorganisms. This may be accomplished by utilizing the clean-catch midstream collection method (see Chapter 9), or alternatively, urine may be collected for culture with a straight catheter or suprapubic aspirate taken directly from the bladder. Urine samples for culture should never be taken from an indwelling catheter, as the potential for extraneous contamination is very high. The medical assistant must be sure to document on the requisition form whether the specimen was obtained from a catheter or via the clean-catch midstream urine method, as the amount of urine applied to the agar plate and the interpretation of the results are different for each type of specimen. Figure 10-4 shows an example of the different sizes of inoculation loops used for urine specimens.

Patients must be provided with a sterile cup for the urine collection, and given instructions in how to perform the collection to provide a satisfactory culture sample. It is also imperative that within 1 hour of collection the specimen is set up for culture, refrigerated, or preserved in some way to keep the bacterial count stable. Urine tubes with preservative are available for this purpose. The bacteria present in the culture specimen will multiply at a very high rate if the urine is left at room temperature for more than an hour, so the culture results may appear to be erroneously abnormal.

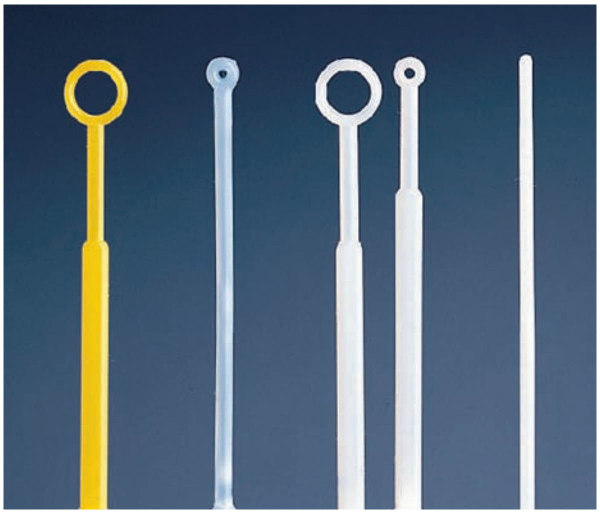

Figure 10-4 Urine loops for culture. The larger loop is used to apply urine collected via catheterization and the smaller loop is used for inoculation of urine collected via the clean-catch method. *Courtesy of Fisher HealthCare, Inc.*

Test Your Knowledge 10-11

Which of these urine specimen types may not be used for cultures?

a. Specimens taken from indwelling catheters
b. Specimens obtained using a straight catheter
c. Specimens collected via the clean-catch midstream collection technique
d. None of the above (Outcome 10-7)

Blood Culture Collection

Health-care providers order blood culture collections when they suspect that the patient may have pathogenic bacteria in the bloodstream, a condition known as **septicemia.** The National Center for Injury Prevention and Control is part of the Centers for Disease Control and Prevention (CDC) and is responsible for compiling statistics about injuries and reducing their consequences. According to its statistics, septicemia was the 10th leading cause of death in the United States in 2006, responsible for more than 30,000 deaths. Septicemia is often a complication of urinary tract infections, pneumonia, and infected wounds. Blood cultures assist the health-care provider by providing information about the causative agent of the infection, as well as determining which antibiotics will be best to use for treatment. It is common for blood cultures to be ordered just before or just after a fever has spiked, and they are often ordered as STAT blood draws. Ideally, blood cultures will be drawn before antibiotics have been administered.

Special bottles that contain nutrients capable of supporting all bacterial growth are used for blood cultures, one that includes special nutrition and a unique atmosphere to encourage growth of microorganisms that need oxygen to grow **(aerobes)** and one that is specially designed to support growth of microorganisms that do not do well in the presence of oxygen (anaerobes). There may be special instructions with the order for the timing of the blood draws; some practitioners ask that they are drawn 30 minutes apart, whereas others will have them done serially, which means that they are drawn at regular intervals over a set period of time. The order may also be to draw two blood cultures, which means that they are to be drawn one after the other in two different sites.

Test Your Knowledge 10-12

Why are multiple bottles inoculated when collecting blood cultures? (Outcome 10-8)

Because bacteria are not usually present in the bloodstream, the presence of any growth in the blood culture is considered to be abnormal. It is imperative that a detailed, controlled aseptic process is followed to clean and disinfect the skin as thoroughly as possible before the specimens are drawn to avoid contamination with normal skin flora. Contamination of the specimen may lead to thousands of dollars worth of treatment that otherwise would have been unnecessary and additional hospitalization days.

Each laboratory will have established protocol for the blood cultures drawn and processed within its facility. The majority of methods will include the use of a combination of 1% to 2% tincture of iodine provided in applicator sticks or cleaning pads, or 10% povidone provided in cleaning packets. Alcohol pads may also be used for a two-step cleaning process. The blood culture collection supplies may be purchased as a kit; Figure 10-5 shows an example. Surgical asepsis must be followed carefully to keep from contaminating the sample, so the site must not be repalpated after cleaning unless sterile gloves have been applied.

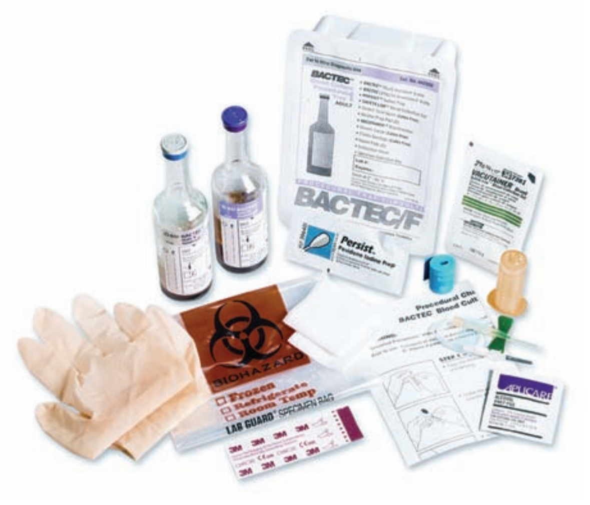

Figure 10-5 Blood culture collection kit containing two bottles, an iodine-based scrub and brush, and alcohol pads. *Courtesy of BD, Inc.*

A syringe with a safety needle may be used to obtain the specimen for the blood cultures (at least a 20-mL syringe for adults; at least a 10-mL syringe for children) or a butterfly setup may be used. The firm bioMérieux has created BacT/ALERT, a blood culture system that includes a butterfly adapter that can be used to fill the blood culture bottles directly during the venipuncture, eliminating the need to transfer the blood from the syringe to the bottles. The BacT/ALERT setup also includes an adapter that allows blood to be drawn directly into other blood tubes if there are additional samples ordered at the same time as the blood culture specimens. This helps to minimize the invasive procedures performed for that patient.

Adult specimens should have approximately 10 mL of blood added to each bottle. The vacuum will automatically pull the blood into the bottle from the syringe or from the vein if a butterfly setup and adapter are used. Monitor the flow carefully, as the vacuum does not always stop when the correct amount of blood has been added, and it is possible to overfill the blood culture bottles. Depending on the manufacturer's recommendations, pediatric draws will most likely call for 2 to 5 mL of blood per bottle.

There are special blood culture bottles available for patients who have already started on antibiotic therapy. These include substances that neutralize or remove the antibiotics from the blood so that they cannot interfere with the growth of the microorganisms. Examples are the Antimicrobial Removal Device (ARD), produced by Becton, Dickinson, or a Fastidious Antimicrobial Neutralization bottle, made by bioMérieux.

Test Your Knowledge 10-13

How many times is the draw site cleaned/disinfected prior to blood culture collections?

a. One time
b. Two times
c. Three times
d. Four times (Outcome 10-8)

Cerebrospinal Fluid Samples

Cerebrospinal fluid (CSF) samples are used to help diagnose various diseases of the central nervous system. These diseases include meningitis and encephalitis caused by bacterial, viral, fungal, or parasitic infections. CSF samples are collected only by a qualified health-care provider by means of a **lumbar puncture,** in which a needle is placed between the lower lumbar vertebrae into

(Text continues on page 234)

Procedure 10-2: Blood Culture Collection Procedure

Blood cultures are ordered when the health-care provider suspects an infection of the bloodstream with a pathogenic microorganism. The collection process is similar to that of a typical blood draw, but there are specific procedures for disinfection of the puncture site and inoculation of the blood culture media that must be followed carefully. Blood cultures that are contaminated with normal skin flora may result in improper treatment and prolonged hospital stays.

TASK

Perform a venous blood culture collection.

CONDITIONS

- Laboratory requisition form
- Hand sanitization equipment
- Gloves
- Blood culture bottles
- 10% Povidone iodine swab stick or other antiseptic scrub kit
- Alcohol prep pads
- 22- or 23-gauge, 1-in. needle and 20-mL syringe **or** 22- or 23-gauge butterfly setup with luer adapter and special bottle cap for inoculation of blood culture bottles
- Biohazardous sharps container
- Biohazardous waste container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.A.I.2: Use language/verbal skills that enable patients' understanding, **III.C.III.3:** Discuss infection control procedures, **III.C.III.5:** List major types of infectious agents, **III.C.III.8:** Differentiate between medical and surgical asepsis used in ambulatory care settings, identifying when each is appropriate, **III.C.III.9:** Discuss quality control issues related to handling microbiological specimens. **III.P.III.2:** Practice Standard Precautions, **III.P.III.3:** Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations, **III.P.III.7:** Obtain specimens for microbiological testing, **III.A.III.1:** Display sensitivity to patient rights and feelings in collecting specimens, **III.A.III.2:** Explain the rationale for performance of a procedure to the patient.

ABHES Standards

- Apply principles of aseptic techniques and infection control
- Use standard precautions
- Practice quality control
- Dispose of biohazardous materials
- Collect, label and process specimens

Procedure	Rationale
1. Verify the test ordered on the requisition form. Pay special attention to the time and/or frequency of the blood culture specimen collection. Note the age of the patient.	The test and identification on the specimen label should be verified every time to avoid erroneous results. Blood cultures may be ordered as STAT draws, which must be performed immediately, or they may be ordered at specific frequencies, such as 1 hour apart. The patient's age may be significant, as there are often special blood culture bottles or differences in the required blood volume for pediatric patients.
2. Gather necessary supplies. Verify the expiration date on the blood culture bottles and antiseptic solution.	Expired bottles and antiseptics may not be used.
3. Wash hands (if they are visibly soiled) or apply hand sanitizer. Allow hands to dry completely.	Clean hands stop the spread of infection. Hands should be completely dry before attempting to apply gloves, or it will be difficult to put the gloves on the wet hands.

Procedure	Rationale
4. Apply tourniquet to the desired site and select an appropriate venipuncture site. Release tourniquet.	It is important to establish a visual landmark for the venipuncture site chosen, as it will not be possible to repalpate the site before insertion of the needle. The tourniquet must be removed because appropriate skin cleansing techniques for blood culture collection take more than 1 minute.
5. Cleanse the venipuncture site.	The cleansing process will vary with laboratory policies. Some sites prefer a two-step process, in which the site is vigorously cleansed with alcohol for 30 seconds, followed by use of a povidone or iodine swab stick. If a povidone/iodine swab stick or cleaning pad is used, the solution must be applied using a concentric circle motion, beginning at the chosen venipuncture site and moving outward without going over any area more than once. If an alcohol/chlorhexidine cleansing kit is used, the site must be scrubbed vigorously for at least 30 to 60 seconds. This is the recommended method for cleaning the site for infants older than 2 months of age or for those with iodine sensitivity. Regardless of the cleansing method, the site must be allowed to dry completely before insertion of the needle into the skin.
6. Prepare the setup for the blood draw. This may be a needle and syringe (at least 20 mL) or a butterfly setup with a luer adapter attached to a special cap that goes over the blood culture bottles and allows the blood to be introduced directly into the bottles.	The choice of needle setup is based on the size and condition of the veins. The ideal ratio of blood to the medium in the bottles is 1:10. This means that each bottle should have 10 mL of blood added for optimal results.
7. Mark the blood culture bottles on the side with the minimum and maximum volumes.	Blood culture bottles do have a vacuum that functions to pull in the blood specimen from a syringe or butterfly. However, the vacuum in these bottles is not controlled as carefully as that in the evacuated tubes, so care must be used to monitor the amount of blood added to the bottle. It is possible to overfill the blood culture bottles. It is important to make a mark on the bottle so that the phlebotomist can estimate when the appropriate amount of blood has been added.
8. Prepare the blood culture bottles. Flip off the removable bottle cap and clean the rubber stopper with an alcohol prep pad. Discard this pad and cover the rubber stopper with a clean alcohol prep pad. Leave this pad in place until the blood is added to the bottle.	Asepsis is critical for blood culture collections. Every precaution must be followed to eliminate an opportunity for bacterial contamination of the specimen.

Continued

Procedure 10-2: Blood Culture Collection Procedure—cont'd

Procedure	Rationale
9. Reapply the tourniquet and withdraw the necessary volume of blood; 20 mL is the desirable volume for most blood culture procedures. Draw the blood using a butterfly setup or a syringe with a needle attached.	20 mL of blood will provide 10 mL per bottle, which is the desired volume for best detection and identification of infective microorganisms.
Option 1: Blood Collection Using a Butterfly Setup	
a. Assemble the butterfly setup with a luer adapter attached to the end of the tubing. Screw the other end of the luer adapter into the special adapter cap designed for the blood culture bottles to be used, or an evacuated tube holder if this works for the bottles.	By using the luer adapter and the special cap, the blood can be added directly into the blood culture bottles without need of additional transfer procedures. 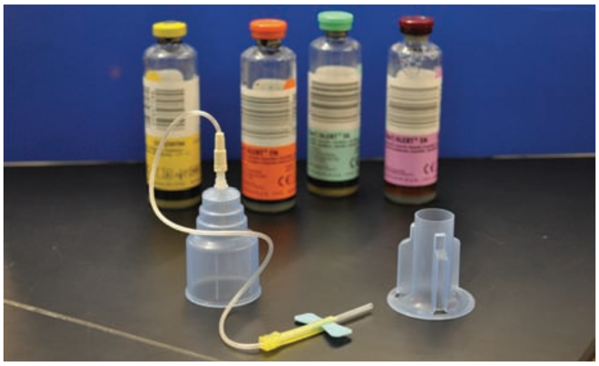
b. Using appropriate venipuncture technique, insert the butterfly needle into the vein. Continue to hold the butterfly device or tape in place.	It is important to secure the needle to avoid damage to the vein and to ensure a successful blood draw.
c. Remove the alcohol pad from the top of the bottle and push down the adapter to pierce the septum of the blood culture bottle. Blood should be added to the aerobic bottle first.	The vacuum in the blood culture bottle will pull the blood from the vein.
d. Monitor the bottle volume and pull up the adapter to remove the needle from the stopper when the desired amount of blood has been added to the bottle.	It is possible to overfill the bottles, so this is an important step to follow.
e. Follow the same procedure to add blood to the anaerobic bottle. Monitor the volume as necessary and pull up the adapter when the desired volume has been reached.	Anaerobic bottles should be filled after the aerobic bottles.

Procedure	Rationale
Option 2: Blood Collection Using a Syringe and Needle	
a. Open the sterile package holding the syringe.	Syringes should remain sterile until the time of use.
b. Pull back and push forward the plunger of the syringe several times to verify whether it moves smoothly.	"Exercising" the plunger in this manner allows for smooth movement when pulling back the plunger to allow blood to enter the syringe. If the plunger does not move smoothly or is too loose within the syringe, it should be discarded.
c. Remove the needle from the sterile wrapper. Do not take off the cap covering the end of the needle until just before insertion into the patient's arm.	The needle must remain sterile until just prior to insertion. It cannot touch other surfaces before piercing the skin of the patient.
d. Attach the needle to the syringe tip.	Verify that this is a secure connection so that blood will not leak out around the needle during the venipuncture process.
e. Reapply the tourniquet, taking care not to touch the venipuncture site. Using appropriate technique, withdraw 20 mL of blood into the syringe, and have the patient place pressure over the venipuncture site.	Special care must be utilized to keep the disinfected site without contamination; 20 mL of blood is optimal for the blood culture procedures.
f. If the bottle is designed to fit a transfer device, activate the safety device on the needle, remove it from the syringe, and discard it in a biohazardous sharps container. Attach the transfer device to the syringe. Remove the alcohol pad from the top of the bottle and push down the transfer device to pierce the septum of the blood culture bottle. Blood should be added to the aerobic bottle first.	The blood culture bottles may come with a very small neck that is designed for use with a standard transfer device such as that in an evacuated tube system.
g. If the bottle is not designed for a transfer device to fit over the top of the bottle, it will be necessary to add the blood to the bottle with the needle on the syringe. Verify that the blood culture bottle is on a stable, flat surface. Remove the alcohol prep pad from the top of the bottle, and puncture the septum with the needle on the syringe, puncturing the aerobic bottle first.	Do not hold the blood culture bottle with your hand while puncturing the septum, as this will increase the chance of an accidental needlestick. Place the bottle on the counter or tabletop.
h. Monitor the volume in the blood culture bottle as the blood is introduced and remove the needle when 10 mL have been added.	Do not overfill the blood culture bottle.
i. Repeat the process to add 10 mL to the anaerobic bottle. Discard the syringe/needle setup in a biohazardous sharps container.	Avoid contamination of the septum on the anaerobic bottle.

Continued

Procedure 10-2: Blood Culture Collection Procedure—cont'd

Procedure	Rationale
10. Label both bottles and the requisition form with the date, time and site of the blood draw, as well as the birth date or other patient ID and the initials of the collector. Transport to the laboratory as soon as possible for processing.	Accurate labeling and documentation are always necessary.
11. Clean the iodine solution off the patient's arm using an alcohol swab. Clean up the work area and chart the procedure.	The iodine solution may become irritating to the skin if left on the patient's arm after the procedure.

Date	
3/12/2014:	Blood culture collection X2 from right and left arm. Connie Lieseke, CMA (AAMA)
4:40 p.m.	

the subarachnoid space, and a sample of cerebrospinal fluid is withdrawn for analysis. This procedure may also be known as a spinal tap. Generally, four tubes are used for the collection, with 1 to 8 mL of fluid placed in each tube. The tubes are numbered in order of collection, which is a very important part of the procedure. In most situations, the first tube collected during the lumbar puncture is used for chemical analysis, including protein and glucose levels. The second tube is used for culture and other microbiology testing, and the last tube collected is used to perform a cell count to look for presence of red and/or white blood cells.

Medical assistants may aid in the collection procedure for CSF samples, and they may also process the samples to be delivered to the laboratory. All CSF analysis must be conducted on a STAT basis, as the integrity of the cells present in the specimen may deteriorate rapidly once the spinal fluid leaves the body. Some tubes used for CSF collection are already labeled with numerals that are raised on the side of the tube, but if not, the numbering of the tubes as the fluid is added is *critical*, as is the assurance that sterile technique is maintained when handling the specimen. The tubes must be closed securely after the fluid is added and labeled appropriately with all patient demographic information and the date and time of the collection. Ideally, if the medical assistant knows that a lumbar puncture is to be performed in the office, he or she may contact the laboratory to request a STAT courier pickup before the procedure is started, so that the courier will arrive very shortly after the process has been completed. Do not refrigerate or freeze CSF samples after collection before the routine testing can be performed.

Test Your Knowledge 10-14

What is the additional information required on the CSF collection tubes that is not required for other types of microbiology samples? (Outcome 10-9)

Genital Samples

Genital cultures use several types of media to identify microorganisms that cause common genital infections such as cervicitis (inflammation of the cervix), urethritis (inflammation of the urethra), or pelvic inflammatory disease. Not all organisms that cause genital infections can be easily cultured, but common infective agents isolated and identified with culture procedures include yeast infections caused by *Candida albicans*, and bacterial infections (vaginosis) caused by *Gardnerella vaginalis, Neisseria meningitides,* or *Haemophilus ducreyi.* Another common infection of the genitals is **trichomoniasis,** which is a sexually transmitted disease caused by the parasite *Trichomonas vaginalis.* This parasite is identified through microscopic examination of vaginal smears and other testing methods, but not cultures.

Many sexually transmitted diseases, such as chlamydia, herpes, human papillomavirus infection, and syphilis, are not typically identified through culture, because the causative microorganisms are difficult to grow in the laboratory. More often, these diseases are identified through serological tests that check for the presence of antigens or antibodies. If a culture order is received for these microorganisms, the order should be verified before proceeding with collection. If the health-care provider really wants a culture to be performed, the medical assistant should carefully investigate the collection and processing procedures so that the culture is viable and the results are meaningful.

Cultures from the genitals of patients are usually ordered if the patient complains of itching, discharge, or burning, or if the patient has visible lesions on or near the external genitalia. Remember to use standard precautions at all times when assisting with the collection or processing of these specimens. The specimens may be collected via aspiration (as in the case of abscesses), or more commonly swabs are used to obtain a specimen from open lesions or discharge. When collecting a urethral swab from a male patient, a small swab on a wire may be inserted into urethra to obtain the culture specimen. It may also be necessary to collect specimens from the rectum if the sexual history of the patient makes this a probable site for infection.

Vaginal and rectal culture specimens may also be collected from pregnant women to test for the presence of group B streptococci. This is an important procedure to avoid newborn infection with these bacteria, which can cause pneumonia or sepsis. Women infected with group B streptococci are usually asymptomatic, so this screening process is necessary as part of a routine prenatal checkup. A swab is used for the collection, and the specimen is handled following general bacterial culture guidelines.

If a cervical culture is requested (as in the case of suspected gonorrhea), the specimen collection will require the use of a speculum to widen the vaginal opening and provide access to the cervix. Excess mucus and drainage are to be wiped away from the opening of the cervix, after which a collection swab is inserted into the opening of the cervix and rotated several times to obtain the specimen. Cervical specimens are often collected at the same time that a specimen is collected for a Pap procedure. Specimens collected for gonococci cultures may be transported in a tube that contains charcoal mixed in the transport medium. The charcoal is designed to absorb any interfering substances in the specimen that might decrease the growth rate of the gonococci bacteria.

Test Your Knowledge 10-15

What type of collection device is used for most genital sample collections?

a. An aspirate using a needle and syringe
b. Skin scrapings
c. Swabs
d. Urine samples (Outcome 10-10)

A medical assistant may be asked to inoculate the agar plates used to support bacterial growth from genital culture specimens in the physician office at the time of collection. It may be necessary to inoculate more than one type of agar to isolate the causative organism. Chocolate agar, which contains lysed (broken) red blood cells, is commonly inoculated with the specimen, as it supports the growth of most pathogenic bacteria that may be present in the genital area. Chocolate agar is shown in Figure 10-6. Pathogens commonly present in the genital area need growth factors that are present inside the red cells, so the red blood cells are lysed when preparing the nutrient agar. However, if the health-care provider wants to determine whether the patient may be infected with *Neisseria gonorrhoeae* (the causative agent of gonorrhea), he or she will also want to inoculate a Petri dish that contains Thayer Martin agar. Thayer Martin agar contains lysed red cells just as the chocolate agar. In addition, Thayer Martin agar contains

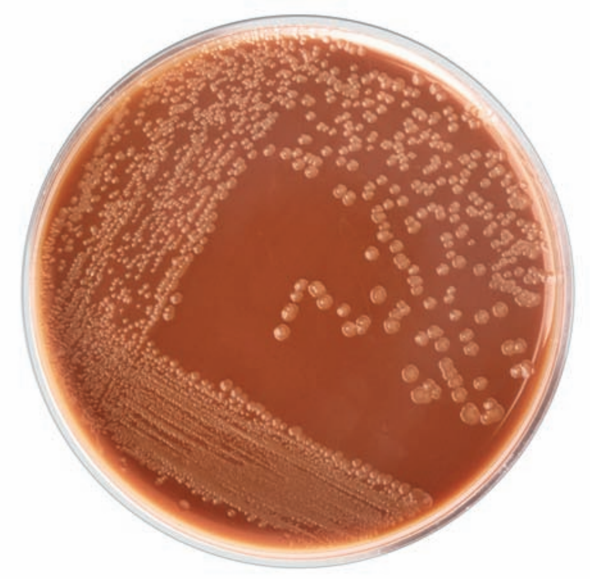

Figure 10-6 Chocolate agar. *Courtesy of Fisher HealthCare, Inc.*

antibiotics (vancomycin, colistin, and nystatin), which inhibit the growth of other bacteria on the agar plate. This allows for the gonococci to grow without competing with other normal flora from the vaginal area. The gonococci also like to grow in an environment with a low oxygen concentration. A special CO_2 tablet may be added to the bag that contains the agar plate after it has been inoculated to enhance the growing environment, or the inoculated agar plate may be placed in a **candle jar,** a special enclosed container in which the medical assistant places the specimen. After the specimen is added, the medical assistant lights a candle and inserts it into the container until the flame goes out. This process removes oxygen from the environment inside the container and enhances the growth of *N. gonorrhoeae* if present.

POINT OF INTEREST 10-2
Pap smears and human papillomavirus

A Pap smear, or Pap test, is performed on a sample of cells collected from a woman's cervix during a pelvic examination. The Pap test does not detect the presence of infection with sexually transmitted pathogenic microorganisms, but it does detect changes in the cells of the cervix that may develop into cancer. It also detects malignant cervical cells that may already be present. A genital culture may be obtained at the same time that a Pap sample is collected, and this may be tested for the presence of the genital human papillomavirus (HPV), or other suspected infectious agents. A special kit is used for collection of the sample for the Pap smear, which may include a brush-like device, a swab, and/or a wooden spatula. Once the sample has been collected by the health-care provider, it is placed into a solution for preservation and transport to the laboratory. Medical assistants may be responsible for processing the sample appropriately after collection. Once the specimen arrives at the laboratory, a cytotechnologist examines the sample under a microscope to look for evidence of abnormal cervical cells.

Certain strains of the genital HPV may cause genital warts in males and females. Other strains of HPV may cause changes in the cells of the cervix that lead to cervical cancer. The cellular changes are detected with the Pap smear, but the presence of HPV and verification of the type of HPV infection must be accomplished with a separate sample collection. Vaccines are available for protection against many of the most common strains of HPV. Cervarix and Gardasil vaccinations are available for women. If necessary, these vaccinations may begin at 9 years of age, but the recommended age range is 11 to 26 years. Gardasil protects against the strains of HPV linked to cervical cancer, and also protects against most genital warts. Cervarix is specific for the HPV strains that cause cervical cancer. Males may also be vaccinated with Gardasil for protection. These vaccinations are recommended for boys and young men ages 9 to 26 years.

Wound Cultures

A wound is any break in the skin. A wound may be a deep break, an abrasion on the surface of the skin, a surgical incision, or a scratch. Even an ingrown toenail with signs of infection may be processed as a wound. Wound cultures may be ordered when a patient has redness, swelling, drainage, or heat in or around a wound. Cultures collected from breaks in the skin are the most varied in collection techniques, processing procedures, and results. The most important aspects of wound culture collection to remember are the following:

- Only culture the wound; not the area around it. Contamination from extraneous bacteria will complicate the culture process, and may cause inaccurate results and incorrect treatment.
- It is best to collect a culture from a clean wound; the process washes away transient bacteria present in the area that are not involved in the actual infection.
- Deep wounds and aspirates are usually collected and processed as anaerobic culture specimens, as the bacteria growing in this environment do not prefer exposure to environmental oxygen because they are growing well in the deep tissues of the body. Anaerobic culture specimens are collected using special culture kits that protect the specimen from oxygen exposure while providing moisture to keep the specimen viable.
- It is very important that the source is documented thoroughly when working with wound cultures. Details of the culture source will dictate how the culture is set up and which types of pathogens the microbiologist is expecting to identify from the wound. Normal flora may be pathogenic if present as an overgrowth (very high in numbers) within a certain area of the body, or if present in an area that is usually considered to be sterile, such as the eye. Documentation of the patient's name, the date and time of collection, and the source of the sample should be affixed to the collection container as well as added to the requisition form and the patient chart. The source

information includes details about the side of the body (for instance, wound left hand thumb) for the wound.

- If **exudate** (fluid from the blood vessels in the area of the infection) or pus is present, the health-care provider may culture this drainage. However, in the case of abscesses, sometimes the culture is not performed, as the most important aspect is to clear the area of all the abscess material, not find out what caused the abscess.
- Results obtained from someone with a chronic wound may be very different from those of a patient with a wound that has just become infected. Sometimes the quantitative results (number of colonies or bacteria present in the culture specimen) may indicate whether current treatment is working for a specific wound.
- A **Gram stain** is often performed from the original collection swab at the same time the culture is set up. Gram stains provide information about the shape and staining characteristics of the pathogen. Observation of the slide after performing a Gram stain can provide preliminary information to guide how the culture is to be started, based on the types of bacteria present. For instance, gram-positive cocci often grow better on certain media than do gram-negative rods.
- Swabs are usually used to collect wound culture specimens. Be sure to keep the swab sterile until the specimen is collected. If an aspirate is taken using a needle and syringe, the needle must be removed from the syringe before transporting the specimen. In this situation the aspirate is often transferred into a sterile tube for transport. Identification of the transfer tube is essential; the laboratory will not know what type of fluid is submitted without the documentation, as many body fluids are very similar in appearance. Wound culture specimens should be delivered to the laboratory as soon as possible, and should not be refrigerated or frozen before the culture medium is inoculated.

Test Your Knowledge 10-16

True or False: It is not advisable to wash a wound before collecting a culture specimen. (Outcome 10-11)

Stool Specimens

Cultures for stool specimens differ from many of the other cultures that we have discussed because of the copious amount of normal flora that is present in the digestive tract. The culture medium used to process stool samples is very selective; it has been designed to encourage the growth of known gastrointestinal pathogens, while discouraging the growth of the normal flora of the area. Only pathogenic microorganisms are processed further for identification and treatment. Sometimes the physician may also ask for a specimen to be collected to check for the presence of parasites in the gastrointestinal (GI) tract. This is known as an **ova and parasite examination (O&P)**, during which the microbiologist will be looking for the presence of parasites or their ova (eggs). Stool cultures may be ordered when a patient has diarrhea that lasts more than a few days, or when there is mucus or blood visible in the stool. A patient may be exposed to pathogenic bacteria from eating food or drinking water that has been contaminated. Common sources include water from a contaminated well, lake, or stream, or undercooked eggs, poultry, or beef. Unpasteurized milk is another potential source. When patients travel outside the country, there may be exposure to bacteria that have become part of the normal flora for the occupants of that area, but that will make visitors sick who are not habituated to them.

Stool specimens for bacterial cultures are collected in a sterile cup. The stool specimen must be fresh for culture setup. The specimen should be delivered to the laboratory within 2 hours of collection, and the specimen must be protected from extremes in heat during transport. The sample cannot be taken from the water in the toilet bowl, and the sample cannot contain urine. Toilet tissue also should not be mixed in with the sample. To successfully collect a stool culture specimen, it is recommended that the patient spread plastic wrap across the back half of the toilet seat so that the stool can be collected on this plastic wrap when eliminated. It is important that urine is not introduced to the plastic wrap. For patients who are wearing diapers, it may be necessary to line the diaper with plastic wrap for successful collection.

Test Your Knowledge 10-17

Why is it important for the medium used for stool cultures to be so selective? (Outcome 10-12)

Common pathogens isolated from stool specimens include the following:

- *Clostridium difficile:* Infection with this bacteria often follows use of broad-spectrum antibiotics, which disrupt the balance of normal flora in the body and allow this opportunistic bacteria to flourish. *C. difficile* is a spore-forming bacterium that may live on surfaces (hospital beds, walls, etc.) for extended periods of time. The bacteria are shed in the stool of infected individuals, and the infection may be life threatening if not treated appropriately. Patients may form a pseudomembrane, or artificial lining in the intestinal tract that impairs absorption of nutrients from the food they ingest. *C. difficile* is difficult to grow as a culture (hence the name *difficile,* "difficult"), but a toxin is secreted by the bacteria that may be detected in the stool specimen or in blood from an infected individual.

- *Salmonella:* This name actually describes a group of bacteria of the *Salmonella* genus. *Salmonella* species live in the intestines of many animals, such as baby chickens, cattle, or pigs. They may also be present in the feces of turtles and domestic animals. Food that has been contaminated with animal feces is the most common means of transmission. The bacteria are transmitted via the fecal-oral route, and usually cause abdominal cramping and severe diarrhea within 1 to 2 days of infection. Most of the time the infection is resolved without antibiotic treatment within a week, but occasionally the *Salmonella* bacteria may enter the bloodstream and cause more serious illness. To avoid infection with this organism, patients should always wash and cook food thoroughly, keep eggs refrigerated, and wash their hands using soap and clean water after working with or petting animals.
- *Shigella:* This name describes a genus of bacteria that causes diarrhea, including *Shigella sonnei* and *Shigella flexneri*. These two types cause almost all cases of *Shigella* infection in the United States. *Shigella dysenteriae* type 1 can cause deadly epidemics, but is very rarely seen in the United States. *Shigella* is transmitted through fecal material from one person to another. Common methods for transmission include food handlers who do not wash their hands after using the restroom, and small children who share wading pools. Infected patients will continue to shed the bacteria in their stools, which often appear bloody, for several weeks after the symptoms have subsided.
- *Campylobacter:* This spiral-shaped organism causes diarrhea, abdominal cramps, nausea, and fever in infected patients. The most common species of the *Campylobacter* genus to cause disease is *Campylobacter jejuni.* Most cases are spread by the feces of birds or the ingestion of birds, as they are able to carry the bacteria without being ill with it themselves. Infection may occur from exposure to raw or undercooked poultry, but occasionally a dog or cat may be infected and the owner may become infected from them. *Campylobacter* can cause disease with very little exposure, so cross-contamination can occur between, for example, cutting boards in a food preparation area and a preparer's unwashed hands, leading to illness. *Campylobacter* infection is the most common bacterial infection causing diarrhea in the United States. Usually it resolves without antibiotic treatment, but there are many documented cases of infected patients who developed the neurological disorder Guillain-Barré syndrome. The *Campylobacter* infection acts as a "trigger" to cause an autoimmune response in which the body attacks the nervous system, causing paralysis.
- *Escherichia coli:* This family of bacteria is present in our environment, and is part of our normal flora. However, there are certain strains of *E. coli* that can cause illness, especially when they are introduced into areas of our body where *E. coli* is not normally found. Some of the *E. coli* families, also known as enterohemorrhagic or verocytotoxic types of *E. coli,* produce the Shiga toxin. The most common one to cause disease in the United States is *E. coli* O157:H7. When a patient has become infected with this type of *E. coli,* the toxins produced by the bacteria cause the illness. Those who are infected usually start to feel ill within a few days of ingesting the bacteria. Symptoms include severe abdominal cramps, vomiting, and diarrhea, which may be bloody. In most patients the infection resolves within a week. However, approximately 5% to 10% of those who are infected with *E. coli* O157:H7 develop **hemolytic uremic syndrome,** which can be life threatening. Hemolytic uremic syndrome occurs when the toxins produced by the bacteria begin to destroy the red blood cells and platelets of the infected patient. The torn or broken red cells begin to block the small capillaries in the kidneys, which can lead to total renal failure.

Nasopharyngeal Specimens

When an upper respiratory infection is suspected, it may be necessary to collect a specimen using a **nasopharyngeal collection technique,** commonly used when collecting samples to test for presence of viral microorganisms. The **nasopharynx** is the area of the pharynx (throat) that runs from the back of the nasal cavity down to the back of the throat that is visible through the mouth. Nasopharyngeal specimens are usually collected by entering the pharynx through the nose, but they may also be collected by going through the mouth and "sweeping" the area of the pharynx that is above the uvula. Some key points are the following:

- Generally, nasopharyngeal swab specimens will be collected 3 to 7 days after the symptoms have started, and before antibiotics or antiviral medications have been administered.

- Always verify the order before collecting the sample. Verify whether a viral culture or a bacterial culture is requested, and have the correct supplies on hand to handle the specimen appropriately after collection. Copan Diagnostics makes several types of "flocked" swabs designed for maximum specimen collection potential. Nasopharyngeal swabs traditionally have a flexible shaft, which may be made of wire or flexible thin plastic. Only Dacron or rayon swabs should be used for viral samples.
- To collect the sample on the nasopharyngeal swab, it is necessary first to measure how far the swab needs to be inserted for the collection. The depth of insertion is approximately half the length from the earlobe to the base of the nose (nostril) of the patient. This should be measured and marked on the nasopharyngeal swab shaft so that it is visible during the collection. The swab is inserted into the nose parallel to the roof of the mouth. Continue to insert the swab until the depth of insertion has been reached and slight resistance is felt, indicating that the nasopharynx has been reached. The swab should be rotated several times, and left in place for a few seconds to absorb any mucus that is present. Remove the swab in the same way that it was inserted. It may be necessary to insert the swab in the other nostril and repeat the process. The swab is then placed immediately into the viral transport media (or bacterial transport media if ordered) and processed according to the directions provided by the testing laboratory.

Eye Cultures

Samples from the eye may be taken from the *conjunctiva,* the mucous membrane that covers the inner area of the eyelid and extends to cover the exposed surface of the eye. The conjunctiva can be accessed just under the upper and lower eyelids. Eye cultures less commonly are obtained as corneal scrapings or an aspirate from the eyeball. In the outpatient setting, conjunctival samples will be the most commonly collected. When assisting with this procedure, it is important to remember that the microbiology laboratory will prefer a sample from each eye, even if the infection is suspected in only one eye. This allows the laboratory to compare any normal flora that may be present, and assists with identification of potential pathogens. Also, the collection swabs should be moistened with sterile saline or nutrient broth before use to minimize discomfort when the collection swab is rubbed across the delicate surfaces of the eye.

Common pathogens present in the eye may be isolated by inoculation of the specimens on 5% sheep's blood agar, chocolate agar, and mannitol salt agar. The pathogens often include *Streptococcus* genus, as well as *Haemophilus* genus, and occasionally *Gonorrhoeae* bacteria may be evident, especially in a newborn infant whose mother had minimal prenatal care.

Test Your Knowledge 10-18

Why is the swab used for eye culture samples usually moistened prior to use? (Outcome 10-13)

Ear Cultures

The human ear is separated into three sections: the outer ear, the middle ear (located right behind the tympanic membrane), and the inner ear. The middle ear is connected to the back of the throat by a small structure known as the **eustachian tube**. Inflammation (with or without infection) of the ear is called **otitis** and is further specified by adding the location to the name. **Otitis media** describes inflammation or infection of the middle ear, and **otitis externa** describes inflammation and/or infection of the outer ear canal.

Otitis media is the most common type of ear infection, and usually develops as a result of blockage of the eustachian tube. When this tube is blocked, fluid builds up in the middle ear, and if there are bacteria or viruses present (such as those causing a sore throat or the common cold), these may be introduced to the fluid within the ear by the eustachian tube. (The middle ear is usually a sterile environment.) This is a common disorder in children, as their eustachian tubes are shorter and straighter than those of adults. If the pressure builds up too much within the middle ear, the tympanic membrane may rupture, and the health-care provider will take a culture specimen of the drainage with a sterile swab. Sometimes it may be necessary to take a culture specimen of the fluid behind the tympanic membrane when it has not ruptured. A procedure called *tympanocentesis* may be performed, in which the health-care provider inserts a needle through the tympanic membrane and aspirates a sample of the fluid that is built up in the middle ear. Another procedure that may allow access to the fluid is a *myringotomy,* in which a tiny incision is made into the eardrum to allow for drainage of the infected fluid. In this situation, a sterile syringe may be used to aspirate some of the sample for a culture, or a sterile swab can be used to obtain the sample as it drains out of the eardrum.

Otitis externa is a condition that is often known as swimmer's ear. When moisture is introduced into the external ear canal, it can create a perfect environment for bacterial or fungal growth to occur. Otitis externa may also be caused by excess moisture or a break in or irritation to the skin of this area. The introduction of foreign objects (such as cotton-tipped swabs) or chemicals (such as those present in hair-care products or dyes) may contribute to the development of otitis externa. The outer ear usually itches, and there is often discharge and pain in the area as well. If a culture is needed to identify the infective microorganism, a sterile swab is used to swipe the lesions and/or obtain a sample of the discharge. The outer ear canal does contain normal flora, such as *Staphylococcus epidermis, Staphylococcus aureus,* and *Corynebacterium* genus.

Test Your Knowledge 10-19

Which type of otitis is often linked to excessive moisture in the ear?

a. Otitis externa
b. Otitis media (Outcome 10-14)

SPECIAL SAMPLE COLLECTION AND PROCESSING PROCEDURES

The role of a medical assistant does not end with the collection and labeling of the specimens. Sometimes there are special procedures that must be completed after the collection so that specimens can be examined, or unique collection requirements for certain specimens that require prior setup.

Fungal Sample and Culture Collection Procedures

As was presented earlier in this chapter, specimen collection procedures performed to detect the presence of fungal elements may be different from those procedures used for bacterial or viral organisms. Successful isolation and identification of fungal elements are dependent on very specific collection methods, rapid transport, and use of the correct growth medium once the sample arrives at the laboratory. Fungi may be isolated from wounds, abscesses, hair follicles, or lesions on the skin. *Tinea pedis* (athlete's foot) and *Tinea unguium* (a fungal infection of the nails) are frequently seen in the physician office. Fungal infections are referred to as **mycoses** and are further identified by the area of infection and/or the causative organisms. Common classifications of fungal infections include the following:

- **Superficial or cutaneous mycoses:** These infect the outermost (dead) layers of the skin and hair, as well as the epidermis, hair follicles and deeper layers of the visible skin. *Tinea corporis* (ringworm) is an example of this type of infection.
- **Subcutaneous mycoses:** These fungal infections occur in the dermis and subcutaneous layers of the skin, and may affect the muscles and tissue layers beneath the skin. Often these are chronic infections that resulted from an initial wound when the fungus was introduced into the body.
- **Systemic mycoses:** Systemic fungal infections may be present in multiple organs and/or areas of the body. They are often introduced to the body through the respiratory tract. Systemic fungal infections may be caused by opportunistic fungal organisms in patients who have suppressed or inactive immune systems (such as HIV-positive patients), or they may be caused by a pathogen that is introduced to the body.

Sample collection procedures will differ according to the site of infection. It is important that the medical assistant and health-care provider verify which type of specimen is necessary for the infection before the collection process begins. Sometimes the health-care provider will utilize a **Wood's lamp,** which shines a fluorescent light over the infected area. Certain types of fungi will fluoresce when exposed to this light, and the physician may use this information without the need of an actual specimen collection to prescribe treatment. Table 10-1 summarizes the type of specimen needed and the collection method used to identify the fungal elements in various areas of the body. Sample types include scrapings of skin lesions for superficial infections, nailbed samples, aspirates from abscesses, or swabs used to sample vaginal drainage.

Test Your Knowledge 10-20

Are all fungal infections superficial? (Outcome 10-15)

Potassium Hydroxide Preparation

Many times it is not necessary to identify the specific *type* of fungus responsible for an infection; the most important information is whether a fungal element is the causative agent. Patient treatment is based on the presence of the fungi without further identification. An examination of the sample under the microscope with a drop or two of 10% to 20% potassium hydroxide (**KOH**) solution is

TABLE 10-1

Information needed for collection of samples for fungal cultures*

Type of Infection	Specimen Required	Collection Method
Skin lesions	Skin scrapings	Use the dull edge of a scalpel or a glass slide; scrape across the lesion. If there are not enough loose pieces created with this action, adhesive tape may be used to pick up a sample. Transport dry in sterile container.
Fungal infection of the hair	Pieces of hair with roots attached	Pluck out the hair (including root material) with sterile forceps. Transport dry in sterile container.
Nail infection	Pieces of nail and debris from underneath the nail	Samples from discolored or misshapen parts of the nail preferable; may use microdrill method to provide samples from the more proximal parts of the nail for better capture of the fungi. Samples scraped from underneath the nail may also provide fungal elements. Transport dry in sterile container.
Abscesses and other subcutaneous infections	Aspirate taken with a needle and syringe, or a sterile swab moistened with saline brushed over the surface from deepest area of infection possible	Sample only the actively infected areas for best results. Be certain that the sample amounts are adequate for culture.
Deep tissue infections	Samples of tissue	Tissue to be excised; will be minced, then cultured for fungal elements.

*For all cultures, specimen ideally will be processed within 2 hours of collections.

performed immediately after collection to search for the presence of fungi. The KOH dissolves other organic elements in the specimen such as hair, mucus, skin cells, and bacteria, but it does not dissolve the fungus that may be present. This allows the health-care provider to visualize the fungus and decide on a plan of treatment without waiting for a culture result. Some fungal species do not show up with the KOH treatment, so the health-care provider may order a culture in addition to the microscopic examination if a fungal infection is suspected upon examination. It is important to remember that results from a fungal culture may not be available until weeks after the sample is submitted, so alternative methods of detection are very important to avoid delay in treatment.

Wet Mount Procedure

A **wet mount** is a means by which living organisms may be observed under the microscope. *Trichomonas vaginalis* is a common sexually transmitted parasitic microorganism that needs to be visualized as it is alive and active via a wet mount examination. A wet mount is prepared by adding a few drops of saline to a sample (usually a swab with secretions on it) then placing a drop of saline/sample mixture onto a microscopic slide. A **cover slip** (a small piece of plastic or glass) is applied over the solution on the slide, and the health-care provider examines the sample. Diagnosis is made by examining the structural characteristics and movement of the microorganism. Wet mount examinations (as well as KOH preps) are performed in physician offices only by health-care providers or trained laboratory professionals, but not by medical assistants. The medical assistant may set up the sample appropriately and bring it into focus on the microscope for examination, but the health-care provider or other appropriately trained personnel will be responsible for performance of the examination.

A **hanging drop slide** is another way that a sample, especially those exhibiting a lot of movement, may be examined for live microorganisms. In this case, a special depression slide is used. These slides have a “scoop” out of the center of the slide, which forms a depression. A cover slip is also used and petroleum jelly is applied with an applicator to the four corners of the cover slip. A drop of the sample (often an inoculated swab that has been agitated in saline) is placed on the center of the cover slip.

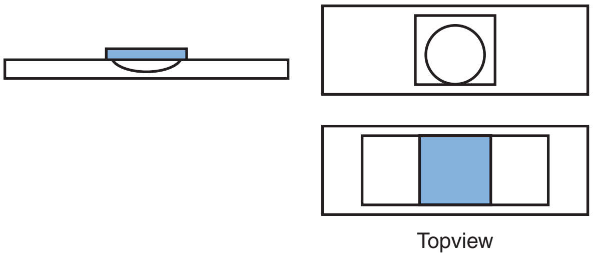

Figure 10-7 A depression slide. Note how the "scoop" is taken from the middle of the slide. This is covered with a cover slip. The sample is actually placed on the cover slip with the depression set over this area and the slide flipped for viewing under the microscope.

The depression slide is placed over the cover slip, where the petroleum jelly seals the edges. The setup is inverted so that the specimen drop is "hanging" from the cover slip over the depression. This allows an excellent opportunity to examine the sample for anatomy and **motility** (movement) of the microorganisms that are present. An example of a hanging drop slide can be seen in Figure 10-7.

Test Your Knowledge 10-21

How are KOH preps and wet mounts similar? How are they different? (Outcome 10-16)

Sample Collection Procedures for Detection of Parasites

Organisms that live on or in a host and use their host for nourishment are known as **parasites.** Pathogenic microorganisms that use the human body for nourishment include **ectoparasites,** which live on the outside of the body, and **endoparasites,** which live within the human body. These parasites feed off the human host, at the expense of the health of the human that they occupy.

Common ectoparasites that use humans as hosts include lice, mites, fleas, and ticks. Diagnosis of these infections usually occurs with the visual examination of the site. For instance, the health-care professional may take a scraping from the skin of the infected individual and examine it under the microscope to see if there are mites present. **Scabies** is a parasitic infection with *Sarcoptes scabiei,* and this may be diagnosed with a skin scraping. The Burrow ink test is another method sometimes used for verification of scabies infections, performed by rubbing the site with an ink pen, then wiping away the ink with an alcohol pad. If there is a parasitic scabies infection then the burrow holes left from where the microorganisms entered the skin will be evident. When a lice infection is suspected, the diagnosis is usually made with careful examination of the hair for the adult lice or the nits that are present near the scalp. Most infections with ectoparasites do not cause the host to become ill; however, secondary infections from the scratching and breaks in the skin may be problematic.

The most common internal parasitic infection in the United States is caused by *Giardia lamblia*. This single-cell parasitic microorganism is passed in the feces of animals. Untreated water in ponds and mountain streams are contaminated with feces from indigenous animals in the area such as beaver, muskrat, elk, and deer. The human host then ingests the parasite by drinking this contaminated water. The presence of the *Giardia* in the water cannot be detected without the use of a microscope, as it does not change the water's color, smell, or taste. Infection by other internal parasites such as roundworms, tapeworms, and hookworms may also occur, but this is not as common in the United States as it is in other areas of the world. Endoparasites can damage their hosts by causing cellular destruction, chewing damage to the gastrointestinal tract, anemia, blood vessel damage, bloody diarrhea, and failure to absorb nutrients.

Ova and Parasite Examination

For most endoparasites, the ova and parasite examination (O&P) is used to make a visual identification of the presence of adult parasites or eggs in the stool. Most parasites have adult forms and produce ova (eggs) or cysts with sexual or asexual reproduction. When the gastrointestinal tract is infected with parasites, the adult parasites or their eggs are shed into the stool of the infected host. The O&P test is performed by placing thin smears of stool (fresh or preserved) onto slides and examining and/or staining them for examination. The sample may need to be spun in a centrifuge for concentration before the slide is prepared to allow for easier detection of any parasites or ova present. The microbiologist will be able to detect the presence or absence of adult parasites or the eggs (ovum), and based on the anatomical structures observed, an identification of the species may also be performed.

When a patient is to be tested for the presence of ova and parasites, he or she is provided with a collection kit that includes a container for the stool specimen, as well as one or two small bottles of preservative solution (Fig. 10-8). The bottles must be appropriately labeled to warn the patient that they contain strong chemicals that could be harmful if swallowed or allowed to remain on bare skin for extended periods of time. The bottles may have a "scoop" built into the cap for the patient to add

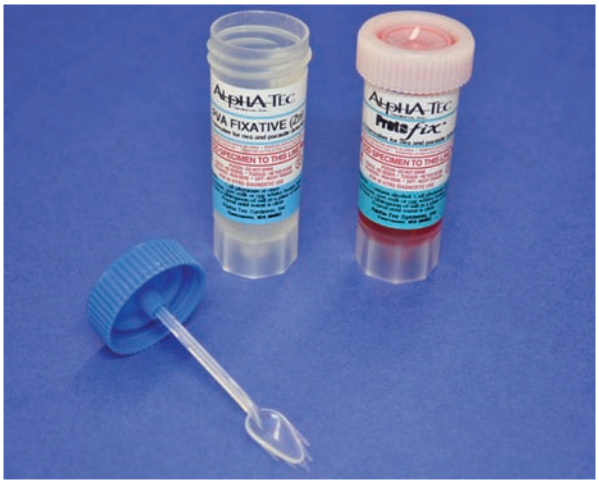

Figure 10-8 Ova and parasite collection kit. There are two different types of preservatives provided, and each bottle has a scoop in the lid that is used to add the stool specimen to the preservative.

the stool specimen to the preservative in the bottle. Common preservatives are 10% buffered formalin and polyvinyl alcohol (PVA). The formalin allows for extended preservation of the parasites and eggs, but the stool sample may be used only for wet mount observation; it cannot be stained for a more permanent record of the examination. PVA does preserve the sample for a reasonable amount of time, but it also allows for staining of the slide for further study. Some laboratories will provide two bottles with different fixatives. The patient should be advised to collect the stool sample into the larger container (provided) or use some other method to collect the sample without contamination of urine or water. Then, the patient should use the scoop or a provided applicator to put a small part of the sample into the provided preservative bottles and return it to the laboratory or physician office.

Sometimes a stool culture is ordered at the same time as the sample for the ova and parasite examination, so it is important that the medical assistant give appropriate instructions for the time intervals before the specimen arrives back at the office. Some laboratories will also prefer a fresh stool sample, in which case it needs to be delivered to the laboratory within 1 hour of collection to preserve the anatomical features of the parasitic microorganisms for identification. It is also common for health-care providers to order multiple sample collections, as the parasites and/or eggs may be shed at random times in the stool. Identification of the specimen collection date and time is very important.

Test Your Knowledge 10-22

How does a collection procedure for an ova and parasite examination differ from a stool culture collection procedure?

a. The O&P collection procedures include containers with preservative solution, but stool culture sample procedures do not
b. Stool specimens for cultures can be kept at room temperature for extended periods of time, whereas specimens for O&P cannot be stored at room temperature
c. The procedures are exactly the same for both types of collections
d. None of the above (Outcome 10-17)

Pinworm Collection Procedures

Pinworm infections, caused by the parasite *Enterobius vermicularis,* are the most common type of worm infections in the United States. The pinworm is a small white worm that can infect the colon and rectum of humans. School-age children and those who are institutionalized are at highest risk for this type of infection, as well as those who care for this population. The adult pinworms are large enough that they can sometimes be seen around the rectum without magnification. The adult female pinworm leaves the rectum at night and lays her eggs around the opening. Within a couple of hours these eggs are infectious, and may be shed into the underwear and bedclothes. This infection often causes intensive anal itching.

Testing for pinworm infection must occur first thing in the morning, before the individual has had a bowel movement or bathed, as these actions may remove the eggs from the rectal area. A piece of clear adhesive tape (sticky side down) may be pressed against the perianal area, and any eggs in the area will stick to the tape. This process may need to be repeated for several days before eggs are identified. The eggs can then be viewed under the microscope when placed on a slide. A "pinworm paddle" may also be used for the collection. Becton, Dickinson produces a tool called the FALCON SWUBE Pinworm Paddle, which is a clear plastic paddle with one adhesive side at the wide end. This sticky side is placed against the perianal area, and then the paddle is placed within a tube to keep it secure until it is examined. The actual paddle can be used as a slide under the microscope, as it is transparent. Figure 10-9 shows the FALCON SWUBE Pinworm Paddle.

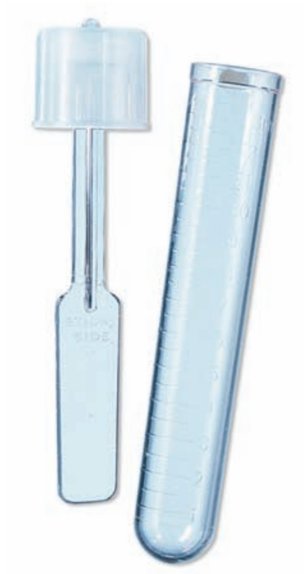

Figure 10-9 The FALCON SWUBE Pinworm Paddle for pinworm collection. The end of the collection device is pressed against the anus to collect the sample. The sample is transported in the outer tube to the laboratory where the actual collection device may be placed on the microscope for examination. *Courtesy of BD, Inc.*

Test Your Knowledge 10-23

What role does adhesive tape play in the collection of the samples for pinworm detection? (Outcome 10-18)

PROCESSING MICROBIOLOGY SAMPLES

This chapter has explained the collection process and immediate handling for many different types of microbiology samples. In a physician office, the medical assistant is often responsible for one more step; the processing of the sample to prepare it for growth of the culture. This may include inoculation of the growth medium, slide preparation, and Gram staining, or even initiating the process for antibiotic sensitivity testing.

Slide Preparation and Gram-Staining Procedure

Developed in 1884, the **Gram stain** is a technique that quickly and inexpensively provides information about the bacteria that are present in a sample. Identification of the microorganisms is based on the fact that almost all bacteria will preferentially absorb different colors of stains, based on characteristics of their cell walls.

When the slide is viewed using the microscope, those bacteria that retain the crystal violet dye will appear purple and be identified as gram positive. Gram-negative organisms will have retained the safranin counterstain, and will appear pink when examined. Figure 10-10 shows gram-positive and gram-negative bacteria after staining. Gram-stained slides also allow the appearance and shape (e.g., cocci, bacilli) microorganisms to be viewed. Complete identification (such as genus and species) cannot be ascertained with a Gram stain, but the microorganism's characteristics that can be determined will assist the physician in making appropriate initial treatment decisions. If the Gram stain is performed at a laboratory outside of the physician office, the preliminary results are often called in to the physician office. The medical assistant must be very careful to document all information and report it to the physician in a timely manner.

Test Your Knowledge 10-24

Why are two different types of stains used for the Gram-staining procedure? (Outcome 10-19)

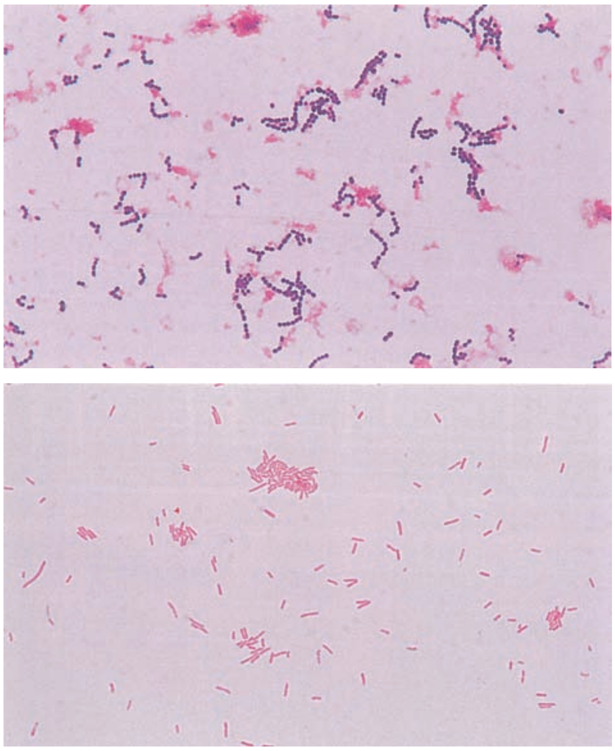

Figure 10-10 Gram-negative and gram-positive bacteria. The purple bacteria *(top)* are gram positive, and the pink bacteria *(bottom)* are gram negative.

Procedure 10-3: Gram-Staining Procedure

The Gram stain is one of the most common procedures performed in the microbiology department. It is a differential stain that allows bacteria to be classified into one of two groups depending on the way that the cell wall reacts with the dyes used in the staining process. The cell walls of the gram-positive bacteria have a greater affinity for the crystal violet stain than the bacteria classified as gram negative. When viewed under the microscope, the gram-positive bacteria will appear purple in color as a result of the crystal violet stain, and the gram-negative bacteria will appear pink in color as a result of the safranin counterstain.

TASK

Perform a Gram stain on a bacterial smear.

CONDITIONS

- Heat-fixed glass slide with bacterial smear
- Gram stain supplies (crystal violet, Gram's iodine, 95% ethyl alcohol, and safranin)
- Gloves
- Forceps
- Water
- Laboratory wipes

CAAHEP/ABHES STANDARDS

CAAHEP Standards

III.C.III.9: Discuss quality control issues related to handling microbiological specimens. **III.P.III.2:** Practice Standard Precautions, **III.P.III.3:** Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations, **III.A.III.** Applied Microbiology/Infection Control

ABHES Standards

- Apply principles of aseptic techniques and infection control
- Use standard precautions
- Practice quality control
- Perform selected CLIA-waived tests that assist with diagnosis and treatment, #5: Microbiology testing

Procedure	Rationale
1. Sanitize hands and apply gloves.	The gloves will protect the hands from becoming stained by the materials used in this procedure.
2. Grasp the slide in the forceps. Hold it over the sink or other waste container. Flood the slide with the crystal violet stain and allow it to be covered for 1 minute.	The slide should be held over the sink to allow excess stain to drain. It is important to keep the stain in contact with the slide for 1 full minute.
3. Pour off stain from slide and flush gently with water. Blot against laboratory wipes to remove excess water.	The slide should be flushed until the liquid stain is removed from the slide. It may be necessary to rub the underside of the slide with a laboratory wipe to remove excess stain if it drips onto the back of the slide.
4. Apply Gram's iodine to the slide for one minute.	The Gram's iodine forms a complex to keep the crystal violet attached to the gram-positive bacteria.

Continued

Procedure 10-3: Gram-Staining Procedure—cont'd

Procedure	Rationale
5. Pour off the Gram's iodine and flush the slide with alcohol. Allow the alcohol to run over the slide until the alcohol runs clear. This may take 20 to 30 seconds, and depends on the thickness of the smear.	The alcohol is a decolorizing agent, and it is imperative that the slide is flushed with this until it runs clear.
6. Rinse the alcohol from the slide with gentle running water. Blot off any excess water by tapping the slide against a laboratory wipe. Clean the underside of the slide with a laboratory wipe if necessary to remove excess stain.	Rinsing the water stops the decolorization process.
7. Cover the slide with safranin stain. Allow the slide to be covered for 20 seconds.	The safranin is the counterstain, which gives color to the gram-negative organisms that did not absorb the crystal violet stain. Without it, they would appear colorless on the slide when viewed under the microscope.
8. Rinse the slide once more with water. If necessary, use alcohol on the underside of the slide to remove any residual stain. Blot the stained slide with a laboratory wipe to absorb excess water.	This is the final stage of the process. It is important to make sure that the residual stain is removed so that the light will not be obscured when the slide is viewed under the microscope.
9. Put away supplies and clean the work area.	The laboratory area must remain clean and tidy.

POINT OF INTEREST 10-3

Tuberculosis and other acid-fast bacilli

Certain types of bacteria resist colorization during the Gram-staining procedure. These bacteria have a higher concentration of lipids in their cell walls, so they do not retain the crystal violet or the safranin stain. At the end of the Gram-staining procedure they remain essentially colorless. In order to identify these bacteria, the Ziehl Neelsen stain is often used to identify these bacteria, known as acid-fast bacilli. The most common group of bacteria that reacts in this way is from the genus *Mycobacterium*. Many species are harmless to humans, but there are a few notable exceptions. *Mycobacterium tuberculosis* is the causative agent of tuberculosis, *Mycobacterium avium* is a bacterium that is capable of causing lung disease, and *Mycobacterium leprae* causes leprosy.

Tuberculosis (TB) is a disease that has been around for a very long time. Active infection is on the rise, and it is imperative that the health-care community remain aware and educated about how to identify and treat this infection. *M. tuberculosis* is most often thought of causing a lung infection, but the bacterium may infect any area of the body. Other common infection sites include the kidneys and the central nervous system. The bacteria are spread through the air when an infective individual coughs, speaks, sneezes, or sings. There are two types of tuberculosis infection:

1. **Latent infection:** A patient has latent tuberculosis infection when he or she has become infected with the bacteria, but does not become actively ill. In the case of latent infection, the patient's immune system has "walled off" or isolated the bacterium to the extent that it does not cause immediate harm. The TB infection may spread at a later date, especially if the patient develops other conditions that lower natural immunity, such as HIV infection. The patients with latent infection will show a positive reaction on the tuberculosis skin test, and they need to be treated to kill the bacteria in their bodies before they cause active infection.

2. **Tuberculosis disease:** A patient with tuberculosis disease has the *M. tuberculosis* bacteria in the body, but the immune system has not rendered it inactive. The bacterium has caused active illness. Symptoms may include severe coughing for an extended period of time, spitting or coughing up blood, fatigue, anorexia, chills, fever, and night sweats. Patients with tuberculosis disease are infectious, and may infect those around them in their home or workplace. They may also infect the health-care professionals who care for them.

To aid in controlling tuberculosis disease, certain high-risk groups should be tested for latent infection, as well those who show symptoms of the active disease. These groups include those who have been exposed to a TB-positive person, patients who have a compromised immune system, those who have lived in or traveled to a country where TB is widespread in the population, those who are incarcerated, and those who are health-care professionals. The tuberculin skin test is the most common way to screen for latent infection, and this test is offered at local health departments as well as many physician offices. Those who immigrated to the United States from another country may have received the bacille Calmette-Guerin (BCG) vaccine against tuberculosis, which is given in countries outside of the United States. However, the BCG vaccine is not considered to offer lifetime immunity. Individuals who have received the BCG vaccine should still receive the tuberculosis skin test. They may show a false-positive result; if that is the case, patients will require follow-up chest x-rays and a physical examination to determine TB status. There is also a blood test now available for tuberculosis infection, called an interferon-gamma release assay. It does not check for the presence of the TB bacilli; rather, it tests for the body's immune response to the bacteria. It is relatively expensive to perform, and not yet in widespread use. An example of this test is the Quantiferon-GOLD test. Those who have had the BCG vaccine have less of a chance of a false positive with the blood test than they do the skin test.

Fortunately, tuberculosis infection, whether the infection is latent or active, is treatable. Those who test positive with the skin test or the blood test are treated with a regime that may last 6 to 12 months, and includes several medications at once. *M. tuberculosis* has become problematic in recent years, with development of several antibiotic resistant strains. Treatment with more than one medication at once seems to reduce the chance of developing resistance. Common medications include isoniazid, rifampin, ethambutol, and pyrazinamide. Patients with active infection are required to take their medication; if there is a high risk of noncompliance, they may even be placed in a directly observed therapy program in which a local health nurse will monitor their dosages daily. Incarceration is also an option; if the local health authority feels that a patient poses a high risk to other individuals and if that patient is noncompliant, jail time may be required.

Plating and Inoculation of Media

In some situations, laboratories may request that the physician office laboratory inoculate the media at the time of collection. A nutrient broth or slant may sometimes be innoculated. These types of media encourage growth of microorganisms but do not allow for isolation of specific colonies for further growth and identification. To accomplish separation of the growth into definitive colonies, an agar plate is used with a specific pattern of inoculation to create isolated colonies of the organism to be further identified and tested. The pattern of inoculation may be different depending on the type of agar plate or type of specimen. A common method used to add a sample to an agar plate is the quadrant method, in which the agar plate is divided into four parts, either mentally or with a grease pencil if necessary. The sample is applied to one quadrant of the plate fairly heavily with a swab that is twirled for full exposure. Successive quadrants are "streaked" with the sample by passing through the previous quadrant one time to pick up specimen, then continuing with numerous streaks that are close together. The goal is to provide an opportunity for growth of isolated colonies so that these can be further tested for identification and possibly sensitivity tested. Figure 10-11 illustrates the quadrant streaking method.

Test Your Knowledge 10-25

Which of these may be used for inoculation?

a. Agar plates
b. Slants
c. Liquid nutrient broth
d. All of the above (Outcome 10-20)

Procedure 10-4: Quadrant Streaking Inoculation Procedure

To isolate microorganisms from pure culture or to isolate specific colonies from agar plates that contain mixed flora, it is necessary to set up a streak plate. The most common method used for this procedure is the quadrant streaking method. The plate is divided into four sections, and the sample is streaked across each area, with every quarter slightly overlapping the previous area.

TASK

Prepare a streak plate using the quadrant streaking method.

CONDITIONS

- Agar plate
- Inoculation loops or sterilization equipment
- Gloves
- Culture specimen on swab

CAAHEP/ABHES STANDARDS

CAAHEP Standards

III.C.III.3: Discuss infection control procedures, **III.C.III.9:** Discuss quality control issues related to handling microbiological specimens. **III.P.III.2:** Practice Standard Precautions, **III.P.III.3:** Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations,

ABHES Standards

- Apply principles of aseptic techniques and infection control
- Use standard precautions
- Practice quality control
- Dispose of biohazardous materials
- Collect, label and process specimens

Procedure	Rationale
1. Remove the lid of the agar plate and place face down on the countertop.	The lid should be placed so that it does not become contaminated.
2. Sanitize hands and apply gloves.	Gloves protect the hands from the microorganisms in the culture.
3. Mentally divide the agar plate into four quadrants. This may also be accomplished physically by drawing a four-part diagram on the bottom of the agar plate.	The quadrant method requires that the plate be divided into four sections.
4. Use the culture swab to inoculate the first quadrant on the agar plate. Gently roll the swab across this area for complete coverage.	The first quadrant must have an adequate amount of specimen present for the rest of the procedure to proceed correctly.
5. Turn the agar plate so that the second quadrant is at the top. Use the inoculation loop to apply sample into the next quadrant by passing it into the original application two times, then spreading it across the second quadrant without entering the first quadrant again.	The sample becomes less concentrated as it is spread across the second quadrant.
6. Using a new sterile inoculation loop (or sterilize and cool a reusable loop) repeat the procedure with the subsequent quadrants.	Each quadrant will be inoculated in the same way with a very small amount pulled from the previous quadrant and spread out.
7. Label the plate and incubate as directed.	Appropriate incubation will be necessary to allow for specimen growth.
8. Put away supplies and clean the work area.	The laboratory area must remain clean and tidy.

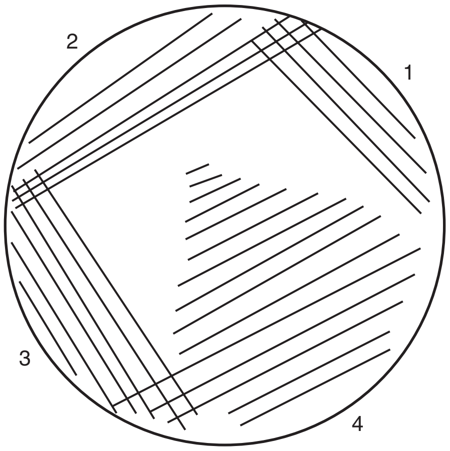

Figure 10-11 The quadrant streaking method of inoculation. The culture swab or a loop of a liquid sample is placed across quadrant 1. A sterile inoculating loop is used for the second quadrant; the streaks overlap quadrant 1 three times, then the swipes continue to fill the quadrant. Another sterile inoculating loop is used for the third quadrant, and the same procedure is used. This process is repeated once more for the fourth quadrant.

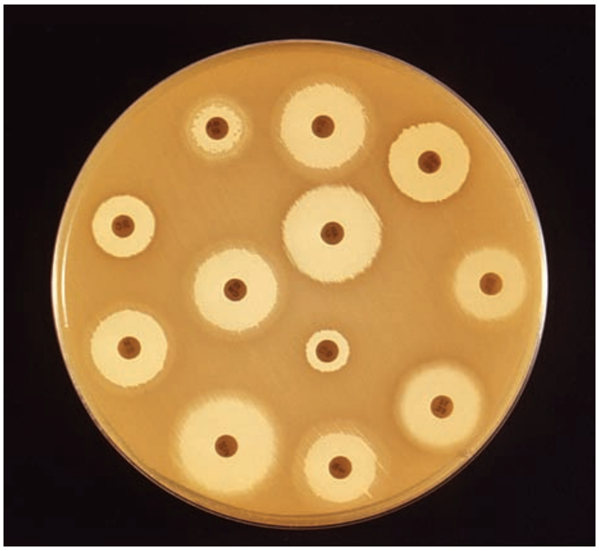

Figure 10-12 Antibiotic sensitivity plate showing the disk-diffusion method. Note the zones of inhibition around some of the antibiotic disks. *From CDC/Gilda L. Jones.*

Antibiotic Sensitivity Testing

Antibiotic sensitivity testing is performed to determine how effective antimicrobial therapy is against a certain type of bacteria. There are numerous methods available for this procedure; instruments such as the VITEK 2 (produced by the firm bioMérieux) are capable of performing antibiotic sensitivities in a fully automated fashion. These automated methods use the **minimum inhibitory concentration (MIC) method** to determine which antimicrobial agent is best to use for a specific pathogenic bacteria. However, many laboratories use an older, more laborious method of testing known as the **disk-diffusion method** (also known as the **Kirby-Bauer method**) for antibiotic sensitivity testing.

To set up the disk-diffusion (Kirby-Bauer) procedure, a solution of the diluted specimen is applied to cover an entire Mueller Hinton agar plate. Absorbent paper discs that have been impregnated with various antibiotics are applied directly to the specimen on the plate (Fig. 10-12). The antibiotics are absorbed into the specimen and the agar as the plate is incubated for a period of 18 to 24 hours. If an organism is **sensitive** to a specific antibiotic, it will have a **zone of inhibition** around the disc for that specific medication, meaning that the antibiotic is able to stop the growth of that particular microorganism so that no growth occurs near the disc on the agar plate. If bacterial growth is present right up to the disc and there is no visible zone of inhibition, this bacteria is **resistant.** A resistant bacteria is not susceptible to the action of a particular antibiotic, so a patient should not be treated with this medication. The zone of inhibition is measured for each type of bacteria, and reported in the laboratory report so that the health-care provider will have the knowledge needed to treat the patient appropriately using antibiotics that will inhibit the growth of the specific pathogen. The Clinical and Laboratory Standards Institute (CLSI) outlines the reporting methods for disk-diffusion antibiotic testing. The technician reading the results in the microbiology laboratory will measure each zone of inhibition for the antibiotics used. He or she will then use the NCCLS to report the antibiotic susceptibility as sensitive, moderately susceptible, or resistant.

The minimum inhibitory concentration testing procedure is usually performed using an automated instrument. A plastic panel with rows of small wells that are filled with a specific concentration of antibiotic is used. The microbiology technician will inoculate each well with a dilution of the culture specimen. (This may also be performed automatically.) After incubating for approximately 24 hours, the wells are checked for evidence of bacterial growth, which presents as a cloudy appearance. The well that has the lowest concentration of a specific antibiotic that does not show evidence of growth is reported as the minimum inhibitory concentration of that specific antibiotic. The laboratory report will include the list of antibiotics tested and the MIC for each of them. It will also interpret these data further to state whether that bacteria is susceptible, moderately susceptible, or resistant to the antibiotic.

Test Your Knowledge 10-26

Why are antibiotic discs used when performing sensitivities? (Outcome 10-21)

POINT OF INTEREST 10-4

Antibiotic resistance

The discovery and further development of antibiotics has been an incredible asset to public health. Millions of lives have been saved, and our life expectancy has increased as a result of the positive steps we have taken to battle infectious diseases. Unfortunately, these steps are not without complications. As we develop more and more complex drugs, the microorganisms in our environment are beginning to develop resistance to our antibiotics. This is especially evident in the drug classifications that are considered first line of defense: the drugs that are cheapest and easiest to use. This resistance is most recognizable in the types of infections that are most prevalent: those that cause diarrhea, sexually transmitted diseases, diseases of the upper and lower respiratory tract, and especially those that are nosocomial (acquired as a result of hospitalization). Some of the most serious problems include the following:

- Vancomycin-resistant *Enterococcus* (VRE): *Enterococcus* is a genus of bacteria that normally is found in the gastrointestinal system. However, when the bacteria are introduced into other areas of the body (the urinary tract, for instance) they are considered to be pathogenic, and cause serious infections. Many strains have developed a resistance to vancomycin, which was commonly used to treat this type of infection.
- **Methicillin-resistant *Staphylococcus aureus* (MRSA):** Staph is a common bacterium found as normal flora in various areas of our bodies. However, when introduced into a wound or into the bloodstream, it can be a virulent pathogen. Methicillin is a classification of antibiotic that has historically worked well against staph infections, but the levels of resistance are rising dramatically to this drug. MRSA is an aggressive pathogen, and may be difficult to treat effectively because of its virulence. It is estimated that up to 20% of our population may be colonized with MRSA at this point.
- Multiresistant *Mycobacterium tuberculosis*: As the number of cases of TB rises around the world, more and more show resistance to one or more of the drugs used for treatment. The regime used for those who have active or latent TB infection now includes more than one drug, as this seems to slow the development of resistant strains, and accelerate the elimination of the pathogen.

Scientists believe that there are multiple factors that contribute to antibiotic resistance: the widespread use of antibiotics in our food sources, historical prescribing patterns of health-care providers worldwide, and patient compliance issues. Whenever an antibiotic is prescribed as treatment for a bacterial infection, the microorganisms either will be eliminated from the patient or they will not. Those bacteria that are not eliminated have now been exposed to the antibiotic, and many have developed a way to "resist" the effects of that particular medication. As these bacteria multiply, their offspring will also have this natural resistance.

Medical assistants can help with educating patients about the importance of this issue. Encourage patients to keep the following issues in mind with antibiotic use:

Always take antibiotics as prescribed. This includes the frequency of the prescription (e.g., take one pill two times per day) as well as the length of time for the dosage. Many patients are tempted to stop taking the medication as soon as they feel better; this allows the bacteria that are still present in their bodies to survive and develop resistance to the antibiotic that they were taking. Patients need to take the prescription as it was written.

Do not take prescriptions that belong to other people. These medications were not prescribed for the specific type of infection that the patient may be exhibiting. Even if two people have the same infection, the required dosage for a medication may be different for each. Self-medicating in this way is a dangerous practice, and contributes to antimicrobial resistance.

Do not insist on antibiotics when the health-care provider feels that the infection is viral in nature. Antibiotics are not effective against viral infections, and taking them when they are not necessary only exposes the body to these medications when they are not needed.

Do not insist on a "higher grade" antibiotic when the more common drug may be just as effective. Many patients feel that penicillin, tetracycline, and similar first-line drugs could not be the best thing for them, as they are convinced that the newest drug developed must be better. Follow the guidance of your health-care provider and your pharmacist; they are more familiar with the specific dosage recommendations for these drugs.

MICROBIOLOGY TEST RESULTS

It can be difficult to interpret the results for microbiology testing by identifying the information that is clinically significant. Both the culture report and the sensitivity result may be present on the same document. This may be ordered or reported as a **culture and sensitivity (C&S) report**. Each laboratory has a unique system that it uses for reporting, but there are a few similarities common to all reports. The report will provide information about the microorganisms that are isolated from the culture, as well as relative amounts of these different microorganisms. When the report is final, there may also be antibiotic sensitivity data provided for bacterial cultures so that the health-care provider will be able to suggest treatment with the best medication available. This may be reported in several ways, so it is important for the medical assistant to gain an understanding of the report format used by his or her laboratory in order recognize abnormal values and bring them to the attention of the health-care provider. Figure 10-13 provides an example of a microbiology laboratory report.

Test Your Knowledge 10-27

What is a C&S report used to document?

(Outcome 10-22)

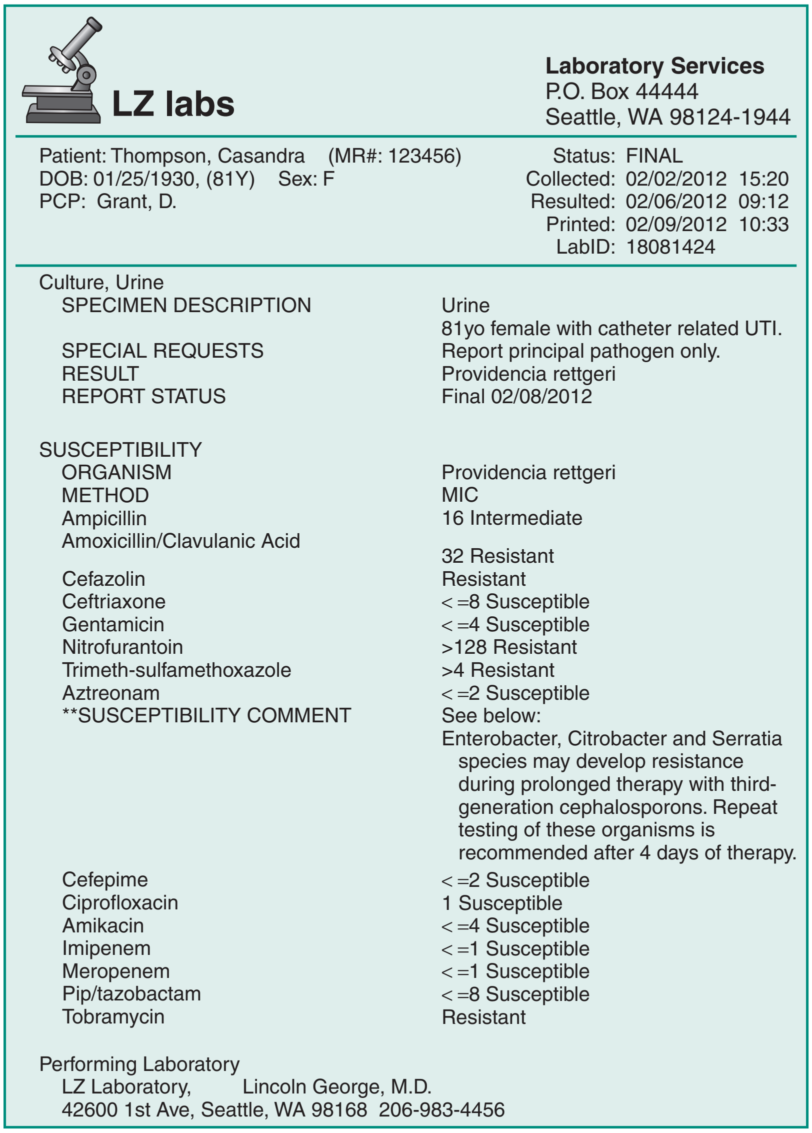

LZ labs

Laboratory Services
P.O. Box 44444
Seattle, WA 98124-1944

Patient: Thompson, Casandra (MR#: 123456)
DOB: 01/25/1930, (81Y) Sex: F
PCP: Grant, D.

Status: FINAL
Collected: 02/02/2012 15:20
Resulted: 02/06/2012 09:12
Printed: 02/09/2012 10:33
LabID: 18081424

Culture, Urine

SPECIMEN DESCRIPTION	Urine 81yo female with catheter related UTI.
SPECIAL REQUESTS	Report principal pathogen only.
RESULT	Providencia rettgeri
REPORT STATUS	Final 02/08/2012

SUSCEPTIBILITY

ORGANISM	Providencia rettgeri
METHOD	MIC
Ampicillin	16 Intermediate
Amoxicillin/Clavulanic Acid	32 Resistant
Cefazolin	Resistant
Ceftriaxone	< =8 Susceptible
Gentamicin	< =4 Susceptible
Nitrofurantoin	>128 Resistant
Trimeth-sulfamethoxazole	>4 Resistant
Aztreonam	< =2 Susceptible
**SUSCEPTIBILITY COMMENT	See below: Enterobacter, Citrobacter and Serratia species may develop resistance during prolonged therapy with third-generation cephalosporons. Repeat testing of these organisms is recommended after 4 days of therapy.
Cefepime	< =2 Susceptible
Ciprofloxacin	1 Susceptible
Amikacin	< =4 Susceptible
Imipenem	< =1 Susceptible
Meropenem	< =1 Susceptible
Pip/tazobactam	< =8 Susceptible
Tobramycin	Resistant

Performing Laboratory
LZ Laboratory, Lincoln George, M.D.
42600 1st Ave, Seattle, WA 98168 206-983-4456

Figure 10-13 Microbiology laboratory report.

SUMMARY

One of the most diverse areas of the laboratory is the microbiology department. Specimen requirements, processing procedures, and culture techniques are very specific and varied. Basic aseptic procedures and labeling techniques are common for all specimen types. As part of their role in quality microbiology testing, medical assistants need to remain knowledgeable about the type of test ordered, the specimen required for each kind of test, and the processing requirements. It is also very important that medical assistants know how to read the microbiology reports used by the laboratory so that clinically significant results are recognized and handled appropriately.

TIME TO REVIEW

1. Inoculation is the: Outcome 10-1

- **a.** Addition of a specimen to a culture medium for growth
- **b.** Addition of antibiotics to a culture medium
- **c.** Separation of colonies on a culture plate
- **d.** Gram staining of a specimen

2. True or False: Mycoses are the various types of fungal infections that may be present on or in humans. Outcome 10-1

3. Hemolytic uremic syndrome is a severe consequence of which type of infection? Outcome 10-1

- **a.** *Shigella*
- **b.** *Salmonella*
- **c.** *E. coli* O157:H7
- **d.** *Yersenia*

4. A nasopharyngeal culture specimen is obtained by access through the: Outcome 10-13

- **a.** Mouth
- **b.** Nose
- **c.** Eye
- **d.** Pharynx

5. Which of these is not a type of culture media? Outcome 10-3

- **a.** Slant
- **b.** Broth
- **c.** Agar
- **d.** Aerobe

6. True or False: Sample collection supplies and transport media for bacterial, viral, and fungal cultures are all different. Outcome 10-4

7. Why should sputum samples be protected from light? Outcome 10-6

8. Which of these statements is false concerning the collection procedures for blood cultures? Outcome 10-8

- **a.** Blood cultures are rarely collected
- **b.** Aseptic technique is essential to avoid contamination of the specimen and misdiagnosis of the patient
- **c.** Blood is usually added to more than one bottle after collection
- **d.** Physicians may order blood cultures to be drawn serially at set time intervals

9. Genital samples may be used to detect infections with: Outcome 10-10

- **a.** *Candida albicans*
- **b.** *Trichomonas vaginalis*
- **c.** *Neisseria meningitides*
- **d.** All of the above

10. Preservatives are used for transport of specimens to be examined for presence of ova and parasites because: Outcome 10-17

- **a.** The preservatives provide nutrition for the parasites in the sample
- **b.** The appearance of the parasites and their eggs are kept near those of their natural state in the body
- **c.** The preservatives keep the parasites alive so that they can grow in the laboratory
- **d.** All of the above

11. What time of day should a pinworm sample collection occur? Outcome 10-18

- **a.** Just before bedtime
- **b.** Randomly; the time of day is not important
- **c.** First thing in the morning
- **d.** Right after bathing in the morning

12. Gram-positive microorganisms will appear ________________ after the Gram-staining procedure is complete. Outcome 10-19

- **a.** pink
- **b.** purple
- **c.** blue
- **d.** black

13. The quadrant method describes: Outcome 10-20
    - **a.** A technique used to inoculate an agar plate
    - **b.** A technique used for Gram stains
    - **c.** A technique used for collection of wound culture specimens
    - **d.** A calculation used for antibiotic sensitivity reporting
14. When reading a microbiology laboratory report, why is it important for the healthcare provider to be familiar with the normal flora commonly present in the area of the body where the specimen was collected? Outcome 10-22

Case Study 10-1: How long is too long?

Marti is a certified medical assistant working in a busy internal medicine practice. There is a new receptionist who is working at the front counter this week, and Marti has found that the receptionist is a bit overwhelmed by the volume of patients the practice is seeing today. The morning appears to be uneventful, but at 2 p.m., the new receptionist brings a urine specimen back to Marti. The receptionist tells Marti that it was dropped off that morning before lunch, but that she forgot to do anything with it until now. Marti sees that the physician has ordered a culture on the urine specimen, and that the urine is in a specimen cup without preservative.

1. What should Marti do?

Case Study 10-2: A sticky situation

Mrs. Oliver brings in her 3-year-old who is complaining of intense anal itching. Lanette, the medical assistant, is asked to collect a sample for examination of potential pinworm infection. This is a procedure she has not performed often, and Lanette is a bit nervous. She decides that the best approach is to speak very little to Mrs. Oliver and the child so that process goes quickly.

1. For this procedure, is this the best approach?
2. If you were Lanette, how would you explain the process to Mrs. Oliver and her child?

RESOURCES

"Antimicrobial Resistance"
World Health Organization
Excellent overview about the progressing issues with antibiotic resistance around the world http://www.who.int/mediacentre/factsheets/fs194/en/

"Antony van Leeuwenhoek (1632–1723)"
History of Leeuwenhoek and the early discoveries of bacteriology http://www.ucmp.berkeley.edu/history/leeuwenhoek.html

BacT/ALERT blood culture collection procedure
An overview of the supplies provided with the culture collection kit and their use. http://www.biomerieux-usa.com/upload/Worksafe-Blood-Culture-Collection-Procedure-1.pdf

"Lumbar Puncture and Spinal Fluid" by John E. Greenlee, MD.
Information about how a spinal fluid specimen is obtained, and how it is processed umed.med.utah.edu/neuronet/pda/2002/HO%20LP%20Spinal%20Fluid%20Greenlee%202002.doc

"Cerebrospinal Fluid Analysis," in Encyclopedia of Surgery: A Guide for Patients and Caregivers
Great article on the clinical indications of a need for CSF collection, the laboratory test results, and potential poor outcomes of the procedure. http://www.surgery-encyclopedia.com/Ce-Fi/Cerebrospinal-Fluid-CSF-Analysis.html

"Disease Listing: Campylobacter General Information"
CDC Division of Foodborne, Bacterial and Mycotic Disease. http://www.cdc.gov/nczved/dfbmd/disease_listing/campylobacter_gi.html

"Disease Listing: Salmonella General Information"
CDC Division of Foodborne, Bacterial and Mycotic Disease, http://www.cdc.gov/nczved/dfbmd/disease_listing/salmonellosis_gi.html

"Disease Listing: Shigella General Information"
CDC Division of Foodborne, Bacterial and Mycotic Disease, http://www.cdc.gov/nczved/dfbmd/disease_listing/shigellosis_gi.html

"Pinworm Fact Sheet"
CDC Division of Parasitic Diseases
Great information about pinworm infestation http://www.cdc.gov/ncidod/dpd/parasites/pinworm/factsht_pinworm.htm

"Educational Commentary on the Preservation of Fecal Specimens for Ova and Parasite Examination"

Excellent information about pros and cons of different preservatives used for O&P examinations; sponsored by the American Society of Clinical Pathology http://www.api-pt.com/pdfs/2002Amicro.pdf

"Training Material Facilitates Proper Nasopharyngeal Swab Collection"

Training module for nasopharyngeal swab collection http://www.rapidmicrobiology.com/news/2023h4.php

"UV Light Cuts Spread of Tuberculosis," Science Daily, March 19, 2009

Information about studies performed for new methods of controlling the spread of tuberculosis within healthcare facilities http://www.sciencedaily.com/releases/2009/03/090316201505.htm

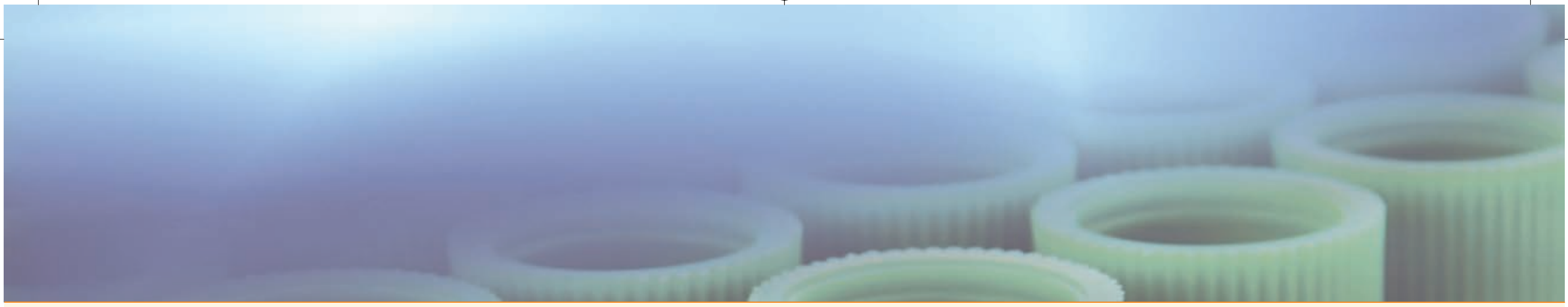

Section II

Specimen Collection and Processing

What Does it All Mean?

As this section clearly points out, proper specimen collection and processing is paramount to obtain accurate and reliable laboratory results. Our Case in Point for this section demonstrates the importance of proper specimen collection and processing.

Case in Point

As you may recall, Wilma F. is a 70-year-old female who presented to the doctor's office where you work complaining of burning and painful urination. She felt hot to the touch as confirmed by her temperature that was 101.5°F. The doctor asked you to collect blood and urine samples for laboratory testing.

Based on the patient's symptoms and the fact that Wilma has a history of urinary tract infections (UTIs), it is likely that she once again has an infection. If she only has a UTI, the urine collected for culture will reveal significant microorganism growth but the bloodstream will be free of microorganisms and the corresponding blood culture results will be reported as "negative." If however, microorganisms establish residence in Wilma's bloodstream, her blood culture results will be considered "positive." It is common for patients to start out with a UTI and then progress into a bloodstream infection. In these cases, the same microorganisms often are found in both the blood and urine culture specimens.

In Wilma's case, she has symptoms of both a blood and urine infection. Wilma's high temperature suggests a possible blood infection whereas her problems with urination indicate a urine infection. The question now is: What microorganisms are causing these problems for Wilma? With the appropriate collection and handling of Wilma's urine and blood samples, the chances of getting reliable and accurate laboratory results is excellent. While the doctor is waiting for the results to come back, he will likely begin Wilma on a suitable treatment known to treat a variety of microorganisms, often referred to a broad-spectrum treatment. He will then adjust treatment if necessary after the microorganisms have been identified and the laboratory results are obtained.

On the Horizon

The study of blood and blood-forming tissues is known as *hematology. Coagulation* refers to the clotting of blood. Because these two concepts interconnect, corresponding testing is often conducted in the same area of the clinical laboratory. The results generated by performing hematology and coagulation testing are of monumental importance in the diagnosis and monitoring of a number of diseases and conditions.

Relevance for the Medical Assistant

Medical assistants may be called upon to collect samples and in some instances perform CLIA-waived tests that fall into the hematology and coagulation category. An understanding of this area of laboratory testing is essential to recognize the clinical significance of test results and to anticipate the needs of the health-care provider when dealing with the patients. In addition, medical assistants need to have a full understanding of their role in the preanalytical process while preparing patients and specimens for testing

Continued on page 258.

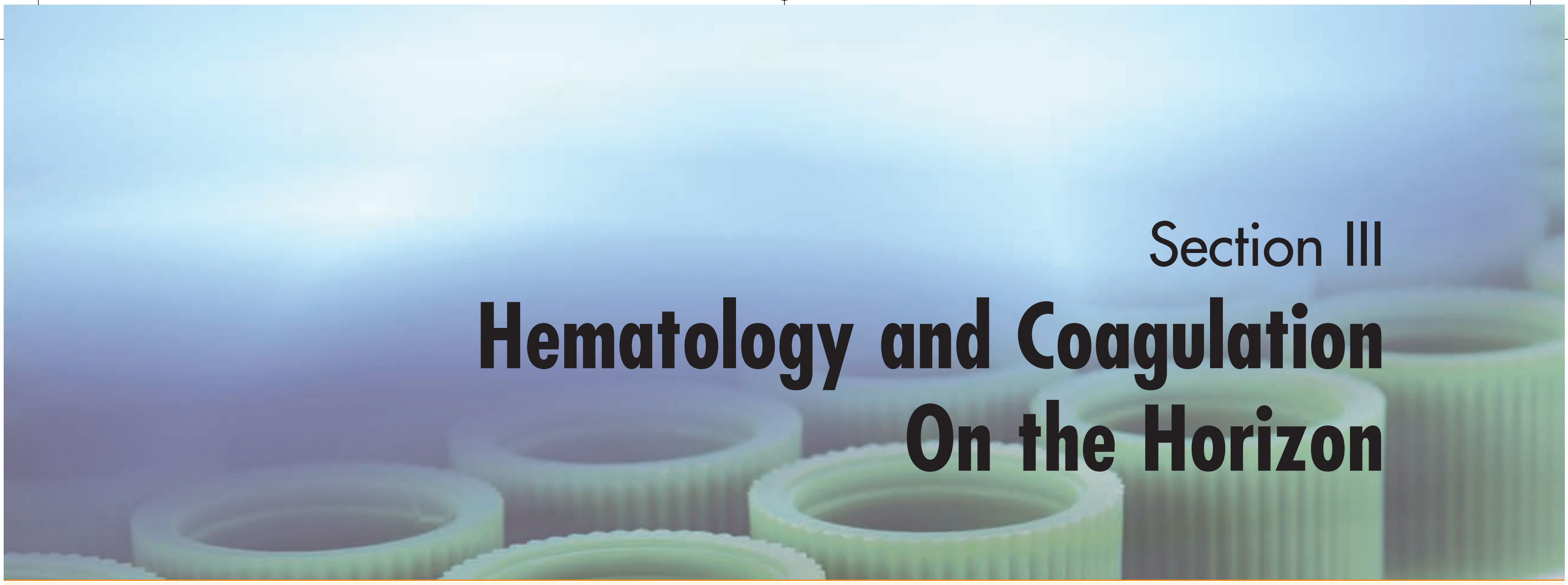

Section III

Hematology and Coagulation On the Horizon

The fundamental concepts associated with hematology and coagulation are covered in six chapters that make up this section of the text, as outlined below.

Chapter 11: Overview of Hematology introduces the reader to the formation, description, and analysis of the three key types of cells in the circulating blood: red blood cells (RBCs), white blood cells (WBCs), and platelets.

Chapter 12: Complete Blood Count With Differential explores the key components of the most commonly ordered battery of hematology tests, known as a complete blood count (CBC).

Chapter 13: Hemoglobin and Hematocrit examines the description, theory, and testing (with emphasis on CLIA-waived procedures) of these two important hematology tests: hemoglobin (Hgb) and Hematocrit (Hct). An independent examination of each test as well as an investigation of the relationship between them are made, because of their importance in the diagnosis and monitoring of hemoglobinopathies, anemia, and polycythemia.

Chapter 14: Erythrocyte Sedimentation Rate covers the description, theory, and testing (with emphasis on CLIA-waived tests) of the erythrocyte sedimentation rate (ESR) test. This test, often in combination with other laboratory tests, is of importance in instances of inflammation changes and in monitoring pregnancies.

Chapter 15: Coagulation Studies introduces the reader to the purpose of coagulation studies, the mechanism of blood clotting, disorders associated with abnormal coagulation test results, and commonly performed coagulation tests. Select coagulation laboratory tests that are useful in the diagnosis of disorders and monitoring of anticoagulant therapy as well as specimen requirements for testing are covered.

Continued from page 256.

Case in Point

You are the primary medical assistant working at Midtown Medical. In the midst of a very busy Monday, Margie, the receptionist, informs you that a new patient, M.J., has been worked into the schedule and is here now to be evaluated. After greeting M.J. and recording his weight and height, you take him into the examination room. While preparing to take his blood pressure and pulse, you ask M.J. what brings him in to see the doctor today. You find out that up until a few weeks ago, this 38-year-old male had been in the best of health. The symptoms he has been experiencing consist of an overall feeling of being tired and weak, and every time he brushes his teeth his gums bleed. After questioning M.J. further, you find out that he was recently hospitalized after a tractor accident and had received a blood transfusion at that time. You notice while talking to him that he looks very pale. The doctor orders blood to be drawn for a complete blood count (CBC), and coagulation studies (PT and APTT). The following results are obtained:

LABORATORY REPORT
PATIENT: M.J.

Complete Blood Count (CBC)

Test	Patient Result	Reference Range for Adult Males
WBC count	10.0	4,000–11,000/mm^3
RBC count	3.0	5.5–6.5 $\times$ 10^6/mm
Hgb	9.0	14–18 g/dL
Hct	30	42%–54%
RBC Indices	>>>>>>>>>	>>>>>>>>>
MCV	86	80–96 femoliters
MCH	31	27–33 picograms
MCHC	35	33–38 g/dL
Platelet count	28	140–400 $\times$ 10^3/mm^3
Differential	>>>>>>>>>	>>>>>>>>>
Neutrophils	65	50%–70%
Bands	5	3%–5%
Lymphocytes	23	23%–30%
Monocytes	6	3%–7%
Eosinophils	1	1%–3%
Basophils	0	0%–1%
Cell morphology	Few schistocytes present	

Coagulation Studies

Test	Patient Result	Reference Range
Prothrombin time (PT)	18 seconds	10–14 seconds
Activated partial thromboplastin time (APTT)	47 seconds	20–35 seconds

Questions for Consideration:

- What type and color of blood-drawing tube should be used for this patient so that the laboratory can perform the complete blood count (CBC) test?
- What type/color of blood-drawing tube should be used for this patient so that the laboratory can perform the coagulation studies?
- Examine the patient's laboratory results. Which of these results is/are considered as "abnormal" (that is, out of the reference range)?
- What laboratory results obtained most likely account for the patient's being pale?
- What is the purpose of the coagulation studies tests (PT & APTT)?

Chapter 11

Overview of Hematology

Nikki A. Marhefka, EdM, MT(ASCP), CMA (AAMA)

CHAPTER OUTLINE

Hematopoiesis—Blood Cell Formation
Types of Blood Cells in the Circulating Blood
  Erythrocytes—Red Blood Cells
  Leukocytes—White Blood Cells
    Granular White Blood Cells
    Agranular White Blood Cells
  Thrombocytes—Platelets
Analysis of the Formed Elements
Summary
Time to Review
Case Studies
Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

11-1 Define the key terms.

11-2 Summarize the process of blood cell formation.

11-3 State the three types of formed elements in the circulating blood and the function of each cellular element.

11-4 Describe the appearance of a reticulocyte and provide examples of situations in which there are increased numbers of reticulocytes in the circulating blood.

11-5 Differentiate the appearance and function of the five types of leukocytes normally present in the circulating blood.

11-6 Describe the appearance of a band cell, and explain when there may be an increased number of band cells in the circulating blood.

11-7 Describe the appearance and function of platelets.

11-8 List three CLIA–waived hematology testing procedures that may be performed in a physician office.

CAAHEP AND ABHES STANDARDS

CAAHEP Standards met:

I.C.I.2. Identify Body Systems
I.C.I.4. List major organs in each body system
I.C.I.5. Describe the normal function of each body system
I.C.I.6. Identify normal pathology related to each body system

ABHES Standards met:

None

KEY TERMS

- Agranular
- Band cell
- Basophil
- Bilobed
- Blast cells
- Complete blood count (CBC)
- Diapedesis
- Eosinophil
- Erythrocyte
- Erythrocyte sedimentation rate (ESR)
- Erythroid
- Erythropoietin
- Formed element
- Granular
- Granulocyte
- Granulocyte-macrophage colony-stimulating factor (GM-CSF)
- Hematocrit
- Hematology
- Hematopoiesis
- Hemoglobin
- Hemostasis
- Heparin
- Histamine
- Interleukin
- Intrinsic factor
- Leukocyte
- Lobulated
- Lymphocyte
- Lymphoid
- Megakaryocyte
- Monocyte
- Myeloid
- Neutrophil
- Phagocytosis (phagocytic, phagocytize)
- Platelet
- Pluripotent
- Polychromatic stain
- Polymorphonuclear
- Red blood cell (RBC)
- Reticulocyte
- Reticulum
- Stem cells
- Thrombocyte
- Thrombopoietin
- Thrombus
- White blood cell (WBC)

Blood is a specialized connective tissue that consists of **formed elements** suspended in liquid plasma. Approximately 38% to 48% of the total blood volume is made up of these formed elements, including **red blood cells (RBCs; erythrocytes), white blood cells (WBCs; leukocytes),** and **platelets (thrombocytes).** The cells function in important ways to keep the human body healthy and safe. The primary function of the erythrocytes is to carry oxygen to the tissues of the body, whereas the leukocytes help to fight off infection within the body. The thrombocytes are present to assist with blood clotting and repair of blood vessel damage. Each of the cell types may be identified by characteristics of its nucleus and cytoplasm.

Hematology is defined as the study of blood and blood-forming tissues. However, more specifically, it is the word used to describe the study of the formed elements in the blood. This chapter explores the process involved with the formation of cellular elements in the bone marrow, and discusses identification and functions of erythrocytes, leukocytes, and thrombocytes.

Test Your Knowledge 11-1

List the three formed elements present in the circulating blood. (Outcome 11-3)

HEMATOPOIESIS—BLOOD CELL FORMATION

Blood cell formation, a process known as **hematopoiesis** (Fig. 11-1), occurs primarily in the red bone marrow. In children, most bones contain red marrow. In the adult (individuals greater than age 20), red marrow is found in the flat bones (vertebrae, ribs, iliac crest, and sternum). **Stem cells** are the precursors of all blood cells and are present in the bone marrow. They are pluripotent, which means that they have the ability to differentiate into several types of specialized cells. These stem cells may divide into precursor stem cells classified as **myeloid** or **lymphoid.** Erythrocytes and most of the white blood cells are capable of arising from myeloid precursor cells. Specialty white blood cells responsible for the majority of the body's immune protection, known as **lymphocytes,** arise from the lymphoid precursor

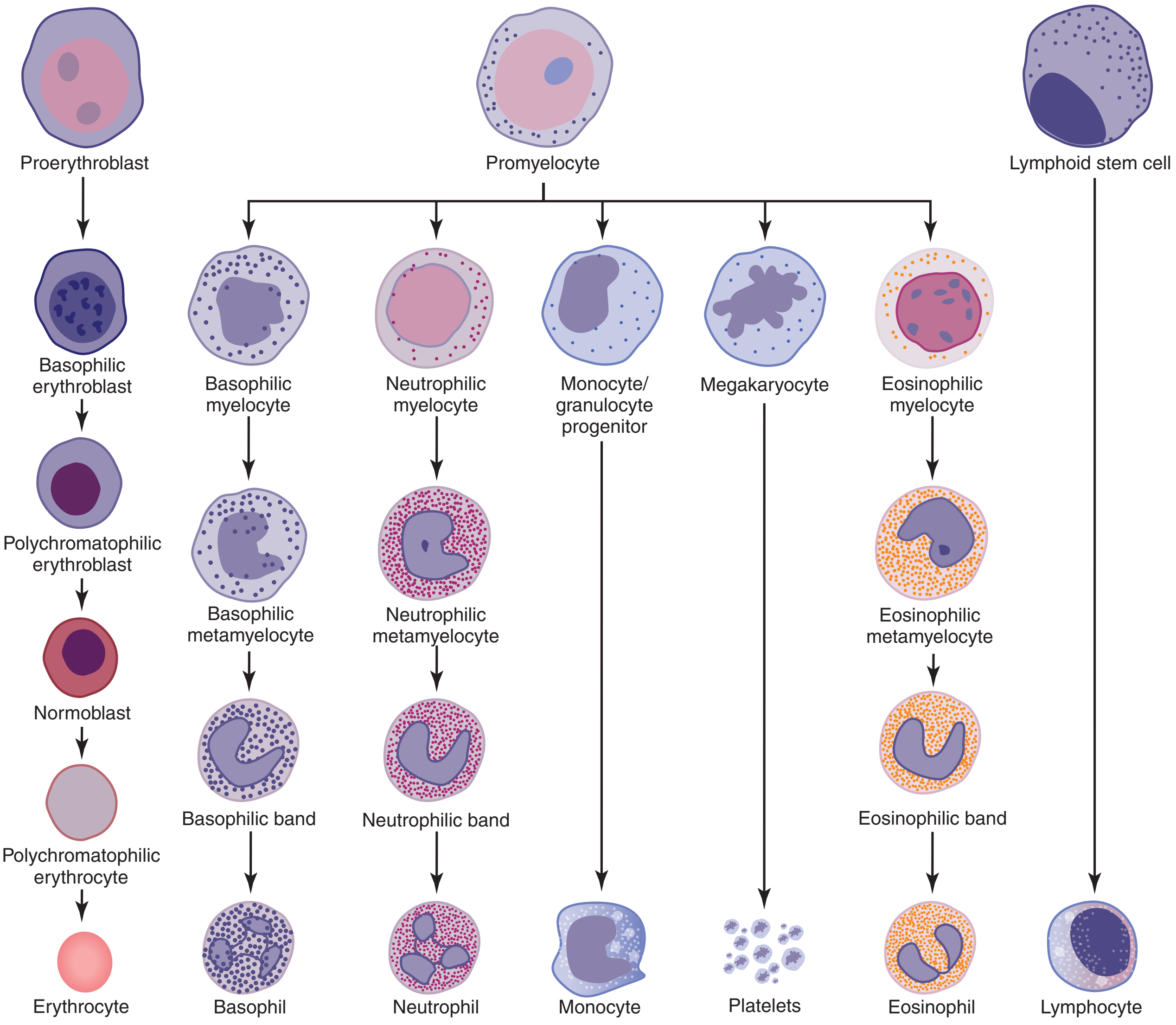
Proerythroblast
Promyelocyte
Lymphoid stem cell
Basophilic erythroblast
Basophilic myelocyte
Neutrophilic myelocyte
Monocyte/ granulocyte progenitor
Megakaryocyte
Eosinophilic myelocyte
Polychromatophilic erythroblast
Basophilic metamyelocyte
Neutrophilic metamyelocyte
Eosinophilic metamyelocyte
Normoblast
Basophilic band
Neutrophilic band
Eosinophilic band
Polychromatophilic erythrocyte
Erythrocyte
Basophil
Neutrophil
Monocyte
Platelets
Eosinophil
Lymphocyte

Figure 11-1 Hematopoiesis. The production and maturation of the blood cells and fragments.

stem cells. It is important to note that lymphocytes are also produced and processed in the thymus gland and secondary lymph organs, in addition to their production in the bone marrow. Descriptions of the formed elements in the blood are summarized in Table 11-1.

Myeloid and lymphoid stem cells, once developed, differentiate to produce **blast cells,** which are the first identifiable immature cell stages of what eventually become erythrocytes (RBCs), leukocytes (WBCs), and thrombocytes. Blast cells mature and go through a number of stages before they are transformed into mature cells. Numerous functions of the spleen are associated with the development, storage, breakdown, and removal of specific blood cells, as noted in Box 11-1.

Under normal circumstances, only the mature blood cells will enter the circulating blood. Immature forms of the blood cells may be present in the circulating blood in disease states. For example, immature red blood cells may be evident in the circulating blood with anemia, and immature white blood cells may be present with leukemia. The cellular components of the bone marrow can be studied by making slides of the marrow, staining it, and studying it under the microscope. The sample for evaluation is obtained from the marrow cavity of the flat bones.

TABLE 11-1

Formed Elements of the Circulating Blood

Blood Cell	Size	Description	Function
		Leukocytes	
Neutrophil	9–15 microns	Multilobed nucleus, grayish purple granules	Phagocytic, increases in bacterial infection
Band cell	9–15 microns	C-shaped nucleus, grayish purple granules	An immature neutrophil, increases in bacterial infection
Eosinophil	10–16 microns	Bilobed nucleus, large red-orange granules	Phagocytic, detoxifying
Basophil	10–14 microns	Bilobed nucleus, large blue-black granules	Release histamine and heparin
Lymphocyte	7–18 microns	Large dark nucleus, little light blue cytoplasm (There are larger ones with larger less dense nuclei and more cytoplasm.)	Produce antibodies, involved in cell-mediated immunity
Monocyte	12–20 microns	Large foamy convoluted nucleus, little or no granules in cytoplasm	Travel to tissue to become macrophage, phagocytic
Erythrocytes	6–8 microns	No nucleus, biconcave discs	Carry oxygen in blood to the tissues
Thrombocytes	0.5–3 microns	Small cytoplasmic fragments	Blood clotting

BOX 11-1 **Functions of the spleen associated with blood cells**

- Site of red blood cell formation in the fetus
- Site of worn out red blood cell and platelet removal
- Site of old red blood cell breakdown
- Site of hemoglobin (from the worn-out red cells) decomposition
- Stores some iron from hemoglobin decomposition
- Contains red blood cell and platelet pools for use when needed
- Areas of lymphocyte proliferation and immune surveillance and response

A number of factors are produced by the body to stimulate the production of blood cells. **Intrinsic factor** helps with B_{12} absorption in the digestive tract, and B_{12} is critical for red blood cell formation. **Thrombopoietin** from the liver stimulates production of platelets, and **granulocyte-macrophage colony-stimulating factor (GM-CSF)** increases the formation of leukocytes. **Interleukins** influence the growth and activation of the lymphocytes. **Erythropoietin** stimulates the bone marrow to produce more red blood cells when needed in the body. When the oxygen level is low in the tissues, the kidneys produce more erythropoietin to stimulate the differentiation and growth of the **erythroid** (pertaining to erythrocyte) cells in the bone marrow to increase the number of circulating red blood cells. Blood loss, such as from a bleeding ulcer, anemia, or low oxygen levels in the lungs due to high altitude or lung disease cause an increase in erythropoietin.

POINT OF INTEREST 11-1

Bone marrow aspiration

In some clinical situations, medical practitioners need to visualize what is occurring in the bone marrow to make an accurate diagnosis. A bone marrow biopsy is not a procedure that is performed casually because the process may be painful and is very invasive. A sample of the red marrow may be obtained to look for the cell distribution, iron stores, or cell maturity. The information gained from the sample may provide valuable clues needed for a differential diagnosis, and therefore, may be very valuable.

Red bone marrow is found in the ribs, sternum, ilium, vertebrae, and the epiphyses of the long bones of adults. In children, it is found in the shafts of the long bones as well. The red marrow is generally the same in each of these areas, so a sample may be taken from any location. The site for the biopsy and aspiration is usually the posterior iliac crest. The sternum and the broad end of the tibia are other sites that may be utilized, although not as frequently as the iliac crest.

A sample of the bone marrow can be obtained using a special hollow needle with a stylus insert. The aspiration is completed using sterile technique to prevent infection and/or inflammation of the bone, which is called osteomyelitis. The patient is usually in a prone position so that the iliac crest is fully exposed and supported. The skin, subcutaneous tissues, and the periosteum are anesthetized with a local anesthetic to reduce discomfort to the patient. A special hollow aspirate, needle is used to penetrate through the layers of tissue and into the bone. A twisting motion is necessary for the needle to reach the destination. Once the marrow cavity has been entered, the stylus part of the needle device is removed, and a syringe is attached to the end of the hollow needle that is exposed. Fluid is aspirated, and some of the marrow material is obtained. The medical assistant or laboratory technologist that is assisting with the procedure must verify whether the sample is valid by observing its appearance.

The aspirated sample is placed on a clean glass slide. The liquid sample from the marrow may be spread across the slide using a similar procedure as that used when preparing a blood smear. There will also be solid components of the bone marrow obtained; these are called spicules, and they are slowly rotated across a slide to make impressions for examination. These solid components are then placed in preservative for further studies if necessary. The slides are stained and the preparations are assessed under the microscope by a pathologist or hematologist.

After the procedure has been completed, the medical assistant will prepare the slides, clean up the supplies and assist with the application of a pressure bandage to the puncture site.

Test Your Knowledge 11-2

Define *hematopoiesis*. (Outcome 11-2)

Test Your Knowledge 11-3

Stem cells are described as pluripotent. What does this mean? (Outcome 11-2)

TYPES OF BLOOD CELLS IN THE CIRCULATING BLOOD

As was introduced earlier in this chapter, there are three categories of cells circulating in blood: erythrocytes, leukocytes, and thrombocytes. A description and function of the primary members in each category are detailed in the following sections.

Erythrocytes—Red Blood Cells

Erythrocytes, also known as red blood cells (RBCs), are the most numerous formed element in the blood. The number of red blood cells in the body varies with the gender and age of the patient, but there are approximately 4.2 to 6.5 million in a milliliter of blood. Normal red blood cells measure 6 to 8 microns in diameter. The red blood cells are smaller than the white blood cells in the body, but larger than the platelets. Red blood cells perform their duties within the blood vessels and do not enter the tissues normally.

The primary function of red blood cells is to carry oxygen to all tissues of the body. Each red blood cell contains an iron pigment responsible for oxygen transport known as **hemoglobin.** Oxygen enters the capillaries that surround the alveoli of the lungs and moves to each red blood cell, where it attaches to the hemoglobin. Hemoglobin is a protein composed of an iron-containing pigment and globulin (a type of protein). Hemoglobin gives red blood cells their characteristic reddish color. Blood is a brighter red color in the arteries as more oxygen is attached and is darker red in the venous blood because of less oxygenation. It is interesting to note that some carbon dioxide is carried through the blood via hemoglobin in the red blood cells to be released in the lungs.

Circulating red blood cells have no nuclei and are biconcave (the center is thinner than the thicker edges) in shape. Their unique shape and the lack of a nucleus allow red blood cells to increase their surface area for gas diffusion (oxygen transport) and allow the cells to contain more hemoglobin than a cell with a nucleus. The biconcave shape of the erythrocyte also eases the cells' passage through the capillaries to bring oxygen to the tissues.

Each red blood cell has a life span of approximately 120 days. Mature red blood cells have no organelles and therefore a limited metabolism. Erythrocytes are not capable of producing proteins or enzymes. Red cells undergo much stress as they traverse the blood vessels in the body and they have no repair mechanisms. The red cells are continuously monitored and the old, worn-out ones are recognized and engulfed by the macrophages or the reticuloendothelial system of the liver, spleen, and bone marrow. The functions of the spleen are presented in Box 11-1.

Immature red blood cells, called **reticulocytes,** may also be found in the circulating blood. Reticulocytes exhibit a blue network when stained with a special stain because the cells have not fully matured and still have residual structures present in the cell cytoplasm. The reticulocytes will also appear as large bluish red blood cells when viewed with the standard stains used for peripheral blood smears. The blue network is remnant RNA (ribonucleic acid) left in the cell when the nucleus is expelled near the end of the red blood cell maturation. With use of special staining procedures, the remnant RNA is visualized as fine, thread-like strands known as **reticulum.** These immature reticulocytes should remain in the bone marrow until maturity. However, when the demands of the body are increased for red blood cells (as in the event of hemorrhage or massive blood cell destruction), reticulocytes enter the circulating blood. Normally, less than 5% of the circulating red blood cells are reticulocytes. Both erythrocytes and reticulocytes are pictured for comparison in Figure 11-2.

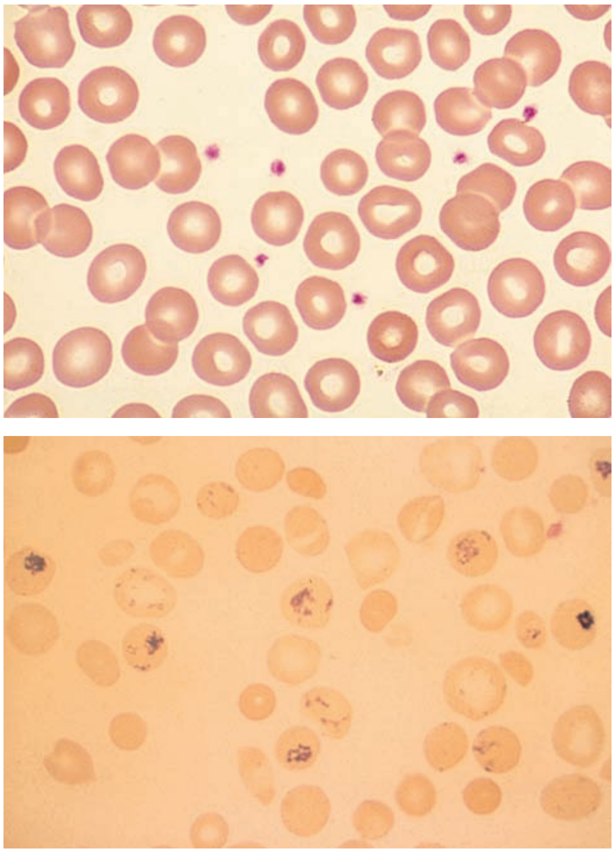

Figure 11-2 Top: Erythrocytes in the circulating blood. Bottom: Reticulocytes The endoplasmic reticulum is visible with new methylene blue stain. *From Harmening, DM: Clinical Hematology and Fundamentals of Hemostasis, ed 5. FA Davis, Philadelphia, 2009, with permission.*

Test Your Knowledge 11-4

Do mature erythrocytes contain a nucleus?

(Outcome 11-3)

Test Your Knowledge 11-5

What type of immature red blood cell may increase if a patient has hemorrhaged due to an injury?

(Outcome 11-4)

Leukocytes—White Blood Cells

Leukocytes are specialized cells that play an essential role in the immune response of the body. White blood cells are the largest blood cells and vary in size from 7 to 20 microns. The total number of white blood cells is 4,300 to 10,800 per cubic millimeter in normal circulating blood. The white blood cells spend most of their life span in the tissues of the body, but the vascular circulation allows white blood cells a mode of transportation throughout the body so that they can respond where necessary. Some WBCs only live a few hours, whereas others live for months or years. Many of them die when they have completed destroying foreign invaders. There are five mature types of white blood cells that are generally found in circulating blood: *neutrophils, lymphocytes, monocytes, eosinophils,* and *basophils* (Fig. 11-3). These cells may be categorized in various ways based on their appearance. One such method is to define them as **granular** or **agranular**.

Granular White Blood Cells

White blood cells that have names ending in –phil (neutrophil, eosinophil, and basophil) have distinct granules visible in their cytoplasm and are known as **granulocytes**. These granules contain digestive enzymes that are designed to destroy or damage foreign cells or microorganisms, or

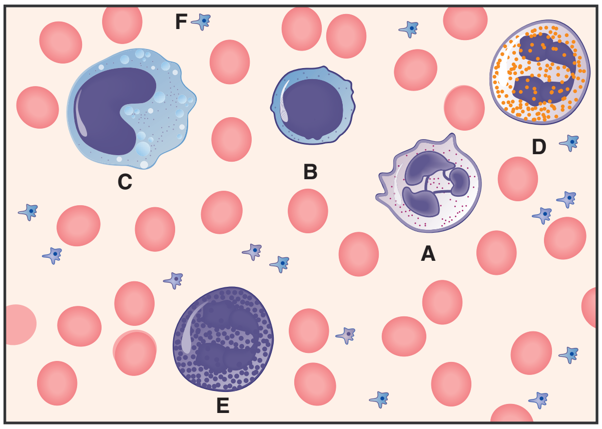

Figure 11-3 Leukocytes and thrombocytes in the circulating blood. A. neutrophil. B. lymphocyte. C. monocyte. D. eosinophil. E. basophil. F. platelet.

chemicals that assist with the process of destroying invaders. The nuclei of granular cells are arranged in lobes or sections. This appearance of the nucleus is described as **lobulated.** White cells are distinguished by the staining of their nucleus and granules when using a **polychromatic stain** (a stain having many colors, such as Wright's stain). The granular cells live only a few days and dissolve as they age, or are destroyed when they have participated in bacterial destruction.

Neutrophils make up the majority of circulating granular cells, and provide much of the body's protection against infection. A typical neutrophil measures 9 to 15 microns in diameter, has a nucleus with 3 to 5 lobes, and has gray/pale purple granules in its cytoplasm. A neutrophil may also be called a **polymorphonuclear** (PMN) leukocyte, or "poly," because of the multiple lobes of its nucleus. Some laboratories may also describe a neutrophil as a "seg," referring to the segmented appearance of the nucleus in these cells.

The primary function of neutrophils is **phagocytosis,** which is the act of engulfing and destroying foreign invaders. The majority of the circulating white blood cells of adults are usually neutrophils. The number of neutrophils often increases when a bacterial infection is present in the body. A mature neutrophil lives for 10 to 24 hours in the circulating blood. Typically, it takes about 30 minutes for the neutrophil to engulf bacteria. This process destroys the cell, and we often see these dead white cells accumulated as pus.

A **band cell,** as seen in Figure 11-4, is an immature form of the neutrophil that may appear in the circulating blood when there is a high demand for neutrophils by the body. Neutrophils are the first to respond to infection or injury, so the need is increased for an acute infection. The nucleus of a band cell is C shaped as the nucleus has not had a chance to mature into lobes or sections. The cytoplasm of a band cell is the same as that of a mature cell. Normal circulating blood may have some band cells, but they usually make up less than 10% of the total. Band cells may also appear in the circulating blood of patients suffering from disease states that result in overproduction of neutrophils in the bone marrow.

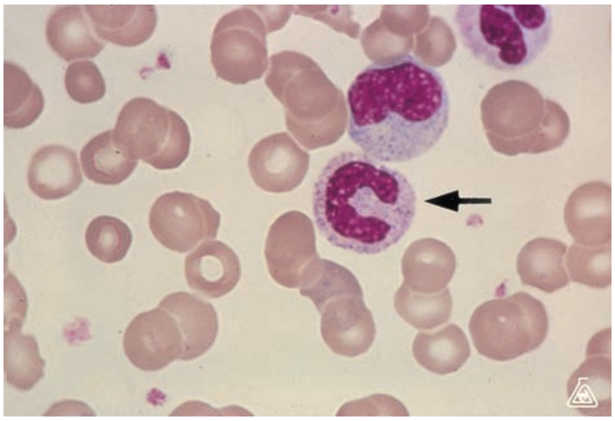

Figure 11-4 A band cell has a C-shaped nucleus and the same granules as the neutrophil.

Eosinophils are 10 to 16 microns in diameter, have **bilobed** nuclei (that is, each nucleus has two lobes or sections), and have large orange/red granules in the cytoplasm. These cells phagocytize invaders attached to antibodies (antigen-antibody complexes) and detoxify foreign substances. (Additional information on these concepts is located in Chapter 24.) An increase in the number of eosinophils is often seen in allergic reactions, myocardial disease, or with parasitic infection. Eosinophils enter the tissues from the blood capillaries and live there for about 6 days.

Test Your Knowledge 11-6

What is the type of white blood cell that will increase in production when a patient has a bacterial infection?
(Outcome 11-5)

Test Your Knowledge 11-7

What is the name of the white blood cell that may be increased in number in an allergic reaction? How can this cell be differentiated from the other types of white blood cells? (Outcome 11-5)

Basophils are 10 to 14 micron in diameter, have bilobed nuclei, and large dark blue granules in the cytoplasm. The nucleus is often hard to see, as the granules are so large and dark that they obscure it from view. The granules found in the cytoplasm of basophils contain **heparin** and **histamine.** Heparin helps to prevent blood from clotting, whereas histamine makes blood vessel walls more permeable so that white cells can get into tissues to phagocytize foreign cells and invaders. An increase in basophils is often seen in the presence of chronic infection, when there is healing of infection, in allergic asthma, or in contact allergies. Basophils have a life span of about 6 days.

Agranular White Blood Cells

Lymphocytes and monocytes are considered agranular, as they have few visible cytoplasmic granules**.** **Monocytes** are the largest of the white blood cells with a diameter of 12 to 20 microns and have a large, convoluted, less densely stained nucleus. The cytoplasm often has a "ground-glass" appearance and a pale blue or light gray color. In addition, the cytoplasm of monocytes often contains vacuoles (holes) when viewed under the microscope. Monocytes travel through the vessel walls and out of the circulatory system into the tissues as needed in a process called **diapedesis.** Monocytes are phagocytic in function as they destroy and remove dead cells from tissue. They defend the body against microorganisms and malignant cells. Monocytes live in the circulating blood for 8 to 24 hours and then as macrophages in the tissue for several months. The macrophages assist with removal of all types of old, damaged, or nonfunctional blood cells.

Lymphocytes are the smallest of the white blood cells, with dark, dense nuclei that often take up most of the cell. Lymphocytes traditionally have a small amount of light blue cytoplasm. Lymphocytes may vary in size from 7 to 19 microns in diameter. Lymphocytes appear slightly larger than red blood cells when viewed on a stained slide. Lymphocytes mature into B cells, T cells, and natural killer cells. B lymphocytes become plasma cells and form antibodies. T cells and natural killer cells are active in cell-mediated immunity. (More information about the body's defense system may be found in Chapter 24.) Lymphocytes are usually the most predominant white blood cell in the circulating blood of children until about age 10 years, but they are the second most prevalent type of white blood cell in adults. An increase in number may be seen in viral infections at all ages. The life span of the T cells is months to years and the B cells live only a few days.

Test Your Knowledge 11-8

What type of white blood cell has forms that make antibodies and are active in cell-mediated immunity?
(Outcome 11-5)

Thrombocytes—Platelets

Thrombocytes, or platelets, are spherical or ovoid cell fragments which are 0.5 to 3 microns in diameter (Fig. 11-3). There are normally 150,000 to 450,000 platelets per millimeter of blood. Platelets are cytoplasmic fragments of large bone marrow cells called **megakaryocytes.** Two-thirds of the platelets are in the circulating blood at any given time, whereas the other third is stored in a pool in the spleen. The typical platelet life span is 8 to 10 days.

Platelets actively participate in **hemostasis** (the stoppage of bleeding or of circulation). Platelets prevent blood loss and have three functions that begin the clotting process (for more information on hemostasis see Chapter 15). When a capillary is damaged, platelets stick together and adhere to the sides of the cut to form a plug that fills the opening. While forming the plug, platelets secrete a chemical known as serotonin designed to help constrict the blood vessels. Platelets also activate the complex chemical process of coagulation to enhance the platelet plug. Younger platelets are larger and more metabolically active and are more effective in hemostasis.

Platelets sometimes clot within the blood vessels when there is no vessel wall damage. This may result in a **thrombus** (blood clot), which adheres to the wall of a vessel or organ. These clots can occlude (block) the vessels and potentially stop the flow of blood. When the body's tissue does not receive oxygen from the blood, body cells will die. A thrombus may be the cause of a heart attack or a stroke.

Test Your Knowledge 11-9

What are the three hemostatic functions of platelets that prevent blood loss? (Outcome 11-7)

ANALYSIS OF THE FORMED ELEMENTS

In reference and hospital laboratories erythrocytes, leukocytes, and platelets may be identified and counted by automated analyzers capable of processing large volumes of specimens. The cells that make up the formed elements in the blood are typically identified by size, appearance, and other specific characteristics. A **complete blood count (CBC)** is the most commonly ordered hematology test. As the test name implies, this comprehensive test provides practitioners with important information about all three formed elements in the circulating blood. The process of cell formation (hematopoiesis) may be studied by collecting and microscopically examining bone marrow samples in large laboratories. Types of hemoglobin can be identified by the technique of hemoglobin electrophoresis. (More information about these tests is presented in Chapters 12 and 13.)

In the physician office and at the bedside in hospitals, CLIA-waived hematology testing is often performed. The hemoglobin concentration of the blood cells is often measured using a CLIA-waived procedure. The **erythrocyte sedimentation rate** and the **hematocrit** value are other common CLIA-waived tests performed in physician office laboratories. The hematocrit is the percentage of red blood cells in a given volume of blood, and the erythrocyte sedimentation rate is a screening test for inflammation that measures the speed at which erythrocytes settle out of anticoagulated blood. Medical assistants or phlebotomists often perform these tests in the physician office laboratory.

Test Your Knowledge 11-10

Name three CLIA-waived hematology tests.

(Outcome 11-8)

SUMMARY

Hematology is the study of the cellular (or formed) elements in the blood. The cells of the circulating blood are predominantly formed in the bone marrow by the process of hematopoiesis. Each type of cell matures from the same pluripotent stem cell. The stem cell differentiates into blast cells for each kind of blood cell and then continues to mature into the functioning cells that are released into the circulating blood. These functioning cells include erythrocytes, leukocytes, and thrombocytes. Erythrocytes carry oxygen, leukocytes are part of the immune system, and platelets participate in blood clotting.

Erythrocytes contain hemoglobin, which enables them to transport oxygen throughout the body. An increase of a more immature form of the erythrocyte, the reticulocyte, in the circulating blood can show that the bone marrow is trying to provide sufficient red blood cells at times of crisis. This may mean that the body has increased its demands for oxygen, or that there is a need to replace lost blood cells.

There are five types of leukocytes normally seen in the circulating blood. These white blood cells are neutrophils, eosinophils, basophils, lymphocytes, and monocytes. Each cell is identified by the properties of its nucleus and cytoplasm and has a specific function in the immune process. Thrombocytes are small round cell fragments that are involved in the clotting process.

Blood cells and platelets can be counted and tested for their ability to function. Large analyzers in the reference and hospital laboratories can perform many tests on a large number of specimens or highly specialized testing. The medical assistant can perform the CLIA–waived testing of hemoglobin, hematocrit, and the erythrocyte sedimentation rate. An abnormal number of the formed elements or an abnormal test result can be indicative of a disease process. Testing procedures and clinical significance of the formed elements in the blood are discussed in the following chapters.

TIME TO REVIEW

1. What is the difference between erythropoietin and thrombopoietin? Outcome 11-1
   a. Erythropoietin stimulates red blood cell formation and thrombopoietin stimulates white blood cell formation
   b. Erythropoietin stimulates platelet production and thrombopoietin stimulates white blood cell production
   c. Erythropoietin stimulates red blood cell formation and thrombopoietin stimulates platelet production
   d. Erythropoietin stimulates platelet production and thrombopoietin stimulates red blood cell production
2. True or False: During hematopoiesis all blood cells are derived from stem cells. Outcome 11-2
3. Which of these statements does not describe the appearance or function of erythrocytes? Outcome 11-3
   a. Erythrocytes are biconcave in shape
   b. Erythrocytes assist in blood clot formation
   c. Erythrocytes carry oxygen throughout the body
   d. Erythrocytes are the most numerous found elements in the bloodstream
4. What is the substance in the red blood cells to which oxygen attaches? Outcome 11-1
   a. Nuclei
   b. Vacuoles
   c. Reticulum
   d. Hemoglobin
5. Immature red blood cells that still contain some endoplasmic reticulum which may be visualized in the circulating blood in periods of high demand are called: Outcome 11-5
   a. Granulocytes
   b. Reticulocytes
   c. Band cells
   d. Lymphocytes
6. The number of white blood cells in the circulating blood is 4,300 to 10,800 per milliliter and their main function is in: Outcome 11-3
   a. Participation in the immune response of the body
   b. Carrying oxygen to the cells of the body
   c. Blood clotting
   d. Liver function

7. The cytoplasm of these cells have large dark granules that contain heparin and histamine: Outcome 11-6
   - **a.** Neutrophils
   - **b.** Eosinophils
   - **c.** Basophils
   - **d.** Monocytes
8. The type of white blood cells that are most prevalent in adults and that phagocytize foreign invaders are called: Outcome 11-6
   - **a.** Neutrophils
   - **b.** Eosinophils
   - **c.** Basophils
   - **d.** Monocytes
9. The function of these cells is to remove dead cells from tissue. They are called macrophages when they have traveled into the tissues: Outcome 11-6
   - **a.** Neutrophils
   - **b.** Eosinophils
   - **c.** Basophils
   - **d.** Monocytes
10. An immature neutrophil that may be visualized in the circulating blood in a severe infection is called a: Outcome 11-7
    - **a.** Granulocyte
    - **b.** Reticulocyte
    - **c.** Band cell
    - **d.** Lymphocyte
11. Which of the formed elements are involved in hemostasis and prevent blood loss? Outcome 11-3, Outcome 11-8
    - **a.** Reticulocytes
    - **b.** Thrombocytes
    - **c.** Erythrocytes
    - **d.** Leukocytes

Case Study 11-1: Meaning of abnormal results

Dr. Stewart has ordered a CBC to be collected on Mr. Marcus. Lily, the medical assistant, draws the blood sample and sends it to the local reference laboratory with the courier that afternoon. The next day, the results are available and Lily reviews them before placing them on the chart for Dr. Stewart to read. The leukocyte count and the neutrophil count are both flagged as elevated above the normal range.

1. What do you think these two elevated results may mean?

RESOURCES AND SUGGESTED READINGS

"Fun Science Gallery: Let's Observe the Blood Cells" Contains pictures and descriptions of the various blood cells. http://www.funsci.com/fun3_en/blood/blood.htm

Chapter 12

Complete Blood Count With Differential

Nikki A. Marhefka, EdM, MT(ASCP), CMA (AAMA)

CHAPTER OUTLINE

Complete Blood Count
  Complete Blood Count Specimen Requirements
  White Blood Cell Count
  Red Blood Cell Count
  Hemoglobin and Hematocrit
  Red Blood Cell (Erythrocyte) Indices
  Platelet Count
  Leukocyte Differential Count
  Peripheral Blood Smear
    Manual Leukocyte Differential
    Red Blood Cell Morphology
    Platelet Morphology and Estimated Count
  Manual Blood Cell Counts
Automated Analyzers for Complete Blood Count Testing
Summary
Time to Review
Case Study
Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

12-1 Define the key terms.

12-2 Identify the component tests of a complete blood count (CBC).

12-3 Recognize the normal reference ranges for each CBC component.

12-4 Identify examples of disorders that are characterized by an increased red blood cell, white blood cell, or platelet count.

12-5 Identify situations where the white blood cell count may be decreased.

12-6 State the clinical significance of the erythrocyte indices included in a CBC.

12-7 For the five types of normal white blood cells identified on a differential, state the normal reference range in the form of percentages.

12-8 Identify examples of diseases characterized by an abnormal white blood cell differential.

12-9 Differentiate terms used to describe normal and abnormal red blood cell morphology using appropriate medical terminology.

12-10 Describe the appearance of platelets when viewed under the microscope.

12-11 Summarize the general principle of electrical impedance as it relates to automated CBC testing.

CAAHEP AND ABHES STANDARDS

CAAHEP 2008 Standards

I.C.1.6. Identify common pathology related to each body system
I.A.I.1. Apply critical thinking skills in performing patient assessment and care

ABHES 2010 Standards

None

KEY TERMS

Anemia
Anisocytosis
Blast
Complete blood count (CBC)
Differential
Electrical impedance
Hematocrit (Hct)
Hemoglobin (Hgb)
Hemocytometer
Histogram
Hyperchromic
Hypochromic
Leukemia
Leukocytosis
Leukopenia
Lysed
Macrocyte
Mean corpuscular hemoglobin (MCH)
Mean corpuscular hemoglobin concentration (MCHC)
Mean corpuscular volume (MCV)
Microcyte
Morphology
Normochromic
Normocytic
Ovalocyte
Poikilocytosis
Polychromatic stain
Polycythemia
Red blood cell index
Sickle cell
Spectrophotometrically
Spherocyte
Target cell
Thrombocythemia
Thrombocytopenia
Thrombocytosis

A complete blood count (CBC) provides physicians and other health-care professionals with an overview of cells and cell fragments in the circulating blood. The amounts of different types of cells in the specimen as well as physical characteristics of these cells are included in the CBC results. Used alone or along with other diagnostic and laboratory tests, the CBC helps the physician diagnose disorders, follow the course of treatment, and determine prognoses. This chapter introduces the reader to each of the major CBC components as well as presents an overview of automated analyzers used in CBC testing.

COMPLETE BLOOD COUNT

The **complete blood count (CBC)** comprises a series of tests, each with its own reference range, as listed in Table 12-1. The tests are usually performed using an automated hematology analyzer, but some of them may also be performed using manual methods. Components of a CBC include the following:

- *A white blood cell (WBC) count***,** providing the number of white blood cells present in the sample.
- A *red blood cell (RBC) count,* which provides the total number of red blood cells present in the specimen.
- **Hemoglobin** concentration, which is the concentration of the iron-containing pigment in red blood cells which binds to oxygen.
- **Hematocrit**, which is the percentage of red blood cells in a given volume of blood.
- Three **red blood cell indices,** defined as numerical descriptions of the RBCs present.
- A *platelet count,* which provides a quantitative measurement of the platelets present in the specimen
- The *white blood cell differential,* which provides quantitative information about the types of white blood cells present in the specimen.

TABLE 12-1

Suggested Reference Ranges for Component Tests of the Complete Blood Count

White blood cell count	4,300 to 10,800/mm^3
Red blood cell count	Adult female: 4.2–5.9 × 10^6/mm^3 Adult male: 4.6– 6.2 × 10^6/mm
Hemoglobin concentration	Adult female: 12 to 16 g/dL Adult male: 13 to 18 g/dL
Hematocrit	Adult female: 37%–48% Adult male: 45%–52%
Red blood cell indices	MCV 80–100 femtoliters MCH 27–31 picograms/cell MCHC 32–36 g/dL
Platelet count	150 to 450 × 10^3/mm^3
Differential	
Neutrophils	54%–65%
Lymphocyte	25%–40%
Monocyte	2%–8%
Eosinophil	1%–4%
Basophil	0%–1%

The hemoglobin and hematocrit tests included in the CBC are CLIA-waived test procedures. However, the other tests included in the complete blood count are categorized as moderately or highly complex by CLIA. Medical assistants may perform moderately complex automated hematology testing with appropriate documented training and supervision. Medical assistants can also perform a normal manual differential from a stained blood smear if they are specially trained and appropriately supervised for this moderately complex procedure.

For the CLIA moderately complex tests, the following additional steps must be taken for appropriate documentation, training, and quality control:

1. All laboratories that perform tests of moderate complexity must be registered with the Centers for Medicare & Medicaid Services (CMS) as a moderately complex laboratory. The Certificate of Waiver that is used for CLIA-waived testing procedures will not be valid if the tests of higher complexity are also performed in the laboratory.
2. The requirements for quality control increase with the higher level of test complexity. This includes the documentation of daily calibration and maintenance, the testing of two levels of quality control performed routinely, and participation in a proficiency testing program.
3. Those employees who perform these tests must also have documentation of all training, and regular documented updates when the process changes.

Test Your Knowledge 12-1

What tests are components of the CBC? (Outcome 12-2)

Complete Blood Count Specimen Requirements

Blood samples for the complete blood count can be collected by venipuncture or capillary collection. The CBC requires a well-mixed lavender top tube containing EDTA as an anticoagulant. There is an emphasis on the mixing of the sample for CBC analysis because the whole blood tested must represent the circulating blood in the body. If the sample is not inverted five to ten times immediately after collection, small clots may form, leading to inaccurate results.

A lavender Microtainer tube will be used in the capillary collection process for a CBC. As in the case of the tube used for venipuncture, the specimen in the Microtainer tube must be mixed during and immediately following the collection process. A steady, consistent blood flow is necessary so that small clots do not form during the blood collection. CLIA-waived analyzers have special collection chambers that can be filled directly with blood from a capillary puncture, eliminating the need for a Microtainer tube. These collection chambers are placed in the instrument for testing.

White Blood Cell Count

The total number of WBCs in the circulating blood (known as a white blood cell count) is another component of a CBC. There is no distinction between types of white blood cells in this total white blood cell count. Since there are so many blood cells, a method known as *scientific notation* is used to express the values for the white blood cell count and the red blood cell count. Scientific notation is used when numbers are too large to be written in the common decimal format, in which the numbers are reported as a number $\times 10^6$. The normal reference range for the WBC count is approximately 4.3 to 10.8 $\times 10^3/\text{mm}^3$ (4,300 to 10,800 cells per cubic mm).

Leukocytosis is the term used when the white blood cell count exceeds the normal range. The most frequent cause for an elevation of leukocytes is infection. Increased WBC counts may also be associated with stress, intense exercise, trauma, inflammation, and pain. Smoking may also cause slight leukocytosis.

In addition, white blood cell counts are usually elevated well above the normal range in the presence of **leukemia.** In cases of leukemia, white blood cell counts may range from 30,000 to 200,000 per cubic millimeter of blood. Leukemia is a cancer of the bone marrow in which a large number of immature white blood cells are being produced or disease states in which mature forms live an exceptionally long time. These white blood cell clones are not capable of performing their roles as part of the immune system. Leukemias are classified according to the predominant cell type present, and whether the disease is acute or chronic in nature. Table 12-2 provides more information about the various types of leukemia.

Leukopenia refers to a WBC count that is below the normal range (values below 4,000 per cubic millimeter). The number of white blood cells may be decreased in a chronic viral infection. Exposure to lead, mercury, some chemotherapy agents, and radiation may also cause the white blood cell count to be diminished.

Test Your Knowledge 12-2

What is the reference (normal) range of the red blood cell count for an adult? (Outcome 12-3)

Red Blood Cell Count

Red blood cells are the most prevalent formed element in circulating blood. Because there are so many red blood cells, scientific notation is used as it is for the white blood cell count to express these values. There are millions of red blood cells present in each specimen, so these are reported as 10^6. The reference (normal) ranges for red blood cell measurements will vary based on the age and gender of the patient. For instance, adult males normally

TABLE 12-2

Types of leukemias

Type	Abbreviation	Description
Chronic myelogenous (granulocytic) leukemia	CML	A large number of all types of white blood cells and platelets; subtyped according to the type of predominant cell
Acute myelogenous (granulocytic) leukemia	AML	The bone marrow produces a large number of granulocytic blast cells; predominantly an adult leukemia
Chronic lymphocytic leukemia	CLL	Accumulation of lymphocytes that live a long time; predominately identified in older adults
Acute lymphocytic (lymphoblastic) leukemia	ALL	Abnormal growth and development of lymphocytes; diagnosed most frequently in children
Hairy cell leukemia	HC or HCL	Excess B lymphocytes are produced; these lymph have projections that make them look “hairy”; generally affects middle-aged and older adults

have more red blood cells present in their bodies than do adult females. Red blood cell counts are performed using an automated hematology analyzer. The reference ranges will vary from 4.2 to 6.2 × 10^6 cells per microliter.

Red blood cell counts that exceed the normal range may be present in various diseases, including **polycythemia** (a condition that results in an excess of RBCs). Those results below the normal range suggest a reduction in the oxygen-carrying capacity of the blood known as **anemia.** The various types of anemia are further described in Chapter 13.

Test Your Knowledge 12-3

List two disorders that are characterized by an increased white blood cell count. (Outcome 12-4)

Hemoglobin and Hematocrit

Hemoglobin (Hgb) is an iron-containing pigment present in all red blood cells and is responsible for oxygen transport. The Hgb concentration is typically measured in grams per deciliter (g/dL). For this test, RBCs are **lysed** (broken open) and the free hemoglobin enclosed in the cell is then measured **spectrophotometrically.** Spectrophotometric measurements measure color changes produced by the hemoglobin present in a sample using a specific wavelength of light. The reference ranges for hemoglobin also vary by gender and age. The hemoglobin normal range for adults is 12 to 18 g/dL.

The **hematocrit (Hct)** is the amount of space occupied by the red blood cells in a whole blood sample as compared to the total sample volume. A hematocrit can be measured directly utilizing hematocrit tubes after centrifugation, or the hematocrit result may be calculated using the hemoglobin and one of the red blood cell indices, called the mean corpuscular volume (MCV). This red blood cell index represents the average volume of the individual red blood cells in a sample. The hematocrit is always reported as a percentage, and the reference ranges vary with gender and age of the patient. Adult normal ranges include values from 38% to 52%. More details of hemoglobin and hematocrit testing and further evaluation of their clinical significance are provided in Chapter 13.

Red Blood Cell (Erythrocyte) Indices

A typical CBC includes three calculated values that help classify RBCs in terms of size and hemoglobin concentration. The calculations use the word *corpuscular* in referring to the red blood cells. These calculations are known as **red blood cell indices:**

1. **Mean corpuscular volume (MCV),** defined as the average volume of a red blood cell in a sample
2. The average weight of hemoglobin in the red blood cells in a sample, known as **mean corpuscular hemoglobin (MCH)**
3. **Mean corpuscular hemoglobin concentration (MCHC),** the average weight of hemoglobin in a given volume of packed blood cells

These values are calculated from the hemoglobin concentration, the hematocrit value, and the red blood cell count. The indices are most often used to classify various types of anemias. Table 12-3 includes more information about each of the indices, including a brief description, the

TABLE 12-3

Red Blood Cell (Erythrocyte) Indices

Red Blood Cell Index Units	Description	Formula	Suggested Reference (normal range)
MCV: Mean corpuscular volume Femtoliters	The average volume of red blood cells in the specimen	$\frac{\text{Hematocrit (\%)} \times 10}{\text{RBC count (millions)}}$	80–100 fL
MCH: Mean corpuscular hemoglobin Picograms per cell	Content or weight of hemoglobin in the average red blood cell	$\frac{\text{Hemoglobin (g/dL)} \times 100}{\text{red blood cell count (million)}}$	27–31 pg/cell
MCHC: Mean corpuscular hemoglobin concentration g/dL	The average concentration of hemoglobin in a volume of packed cells	$\frac{\text{Hemoglobin (g/dL)} \times 100}{\text{hematocrit (\%)}}$	32–26 g/dL

fL is the abbreviation for femtoliters, and pg is the abbreviation for picograms.

formula used for calculation, and a suggested reference range. These indices provide clinicians with valuable information about the size and condition of the red blood cells in the sample as well as more details about the amount of hemoglobin present. The calculated indices may be compared with the appearance of the corresponding red blood cells as noted on a stained blood smear to verify their clinical significance and the accuracy of the values.

Test Your Knowledge 12-4

What information is provided by the red blood cell indices? (Outcome 12-6)

Platelet Count

As was introduced in Chapter 11, platelets are small cell fragments that interact with chemicals in the bloodstream to assist in the clotting process. The reference ranges for platelet counts may vary with the blood sample collection technique employed. Capillary collection values are lower than those in venous blood because some of the platelets in the area are immediately involved with the formation of the plug for the injury of the skin puncture. The venous reference range for the platelet count is 150×10^3 to $450 \times 10^3/\text{mm}^3$ (150,000 to 450,000 per cubic millimeter).

Thrombocytopenia means a decreased platelet count. Platelets may be decreased with some types of infection and in situations where the individual has been exposed to certain medications or toxic substances. **Thrombocytosis** refers to an increased platelet count in the circulation, and may be associated with a variety of diseases. If the platelet count is increased in response to a specific "trigger," it is known as *reactive thrombocytosis*. When the elevated platelet count becomes pronounced over an extended period of time and the bone marrow begins to produce excessive amounts of abnormally functioning platelets, the condition is known as **thrombocythemia.** Thrombocythemia is considered to be a serious condition, as these excess platelets can cause small clots to form intravascularly, or the platelets may not function properly to stop bleeding when needed.

Leukocyte Differential Count

A **differential** is a process by which the percentage of the five normal types of white blood cells in the circulating blood is calculated. These cell types include the following:

1. **Neutrophils:** First responders to foreign invasion of the body
2. **Eosinophils:** Assist with allergic reactions, parasitic infections, and chronic inflammation
3. **Basophils:** Play a key role in the inflammatory process
4. **Lymphocytes:** Ingest and kill microorganisms and play a key role in the immune response with antibody formation
5. **Monocytes:** Attack and eliminate foreign substances; monocytes function as macrophages in the tissues of the body

As introduced in Chapter 11, immature neutrophils, known as *band cells*, are also identified and reported separately when a leukocyte differential is performed. The band cells may be increased in number during a bacterial infection because the bone marrow is attempting to produce an increased amount of white blood cells (especially neutrophils) to meet the body's need. Immature cells that are earlier in the developmental stage than band cells are not normally present in the bloodstream, and their presence indicates a serious infection or other blood disorder. The immature forms of other white blood cell types may occasionally be present and must be identified, as their presence may assist with a diagnosis or prognoses.

When reporting the leukocyte differential, each type of white blood cell is reported as a percentage (%). This procedure is completed utilizing automated methods or performed manually with visualization of a stained blood smear by a properly trained clinical laboratory professional. In blood samples for most adults, neutrophils are typically seen in the greatest numbers (between 54% and 65%). Lymphocytes are next in frequency; 24% to 40% would be the normal range. Monocytes make up approximately 2% to 8% of the white blood cell count. The rest of the white blood cells are made up of eosinophils, which total 1% to 4%, and the basophils, which make up 1% or less of the circulating white blood cells. The immature neutrophils (band cells) may total up to 5% of the normal differential as well. The cells are identified on a stained slide by their size, shape, color, and the structures within the cell. Automated instruments may also report these parameters. Table 12-4 provides a summary of information about the various types of white blood cells.

The results of the differential can be used for initial diagnosis or to monitor treatment. An increase in neutrophils often indicates a bacterial infection. With acute infections, there usually will be an increase in the number of the immature band cells as well as an overall increase in neutrophils. Eosinophils may increase in number with allergies and parasitic infections. Lymphocytes often

POINT OF INTEREST 12-1

Bananas in the blood?

The order of frequency for the presence of the five types of white blood cells may be remembered by using this simple sentence: **N**ever **L**et **M**onkeys **E**at **B**ananas! Starting with the most prevalent to the least, the five types of white blood cells identified in the circulating blood are **N**eutrophils, **L**ymphocytes, **M**onocytes, **E**osinophils, and **B**asophils.

exceed the normal range in viral infections. In immediate hypersensitivity reactions (such as asthma) basophils are often increased. An increase of monocytes may suggest a chronic infection (tuberculosis is an example) or the presence of a malignancy.

Lymphocytes may be elevated with infectious mononucleosis. In addition to the increase in the percentage of lymphocytes, many of the lymphocytes may be larger than normal. *"Reactive" lymphocytes* are also identified with mononucleosis; these are larger than the typical lymphocyte, and when visualized under the microscope, they have a nucleus that is less dense than a typical lymphocyte. An increase of regular, larger, and reactive lymphocytes accompanied by other physical signs and symptoms and a positive mono screening test will be used to diagnose infectious mononucleosis.

The white blood cell differential is significant for patients with leukemia because the various types of leukemia are categorized by the type of predominant white blood cell present in the bone marrow and bloodstream. Leukemia classification is also based on whether the disease process is acute or chronic in nature. The acute leukemias primarily present with immature white blood cells in the form of **blasts** (the earliest form of the white blood cell that is identifiable), while the chronic types of leukemia are characterized by dramatically increased numbers of dysfunctional mature cells. In acute lymphoblastic leukemia, the differential would have many immature lymphocytes, referred to as *lymphoblasts*. The blood smear of a patient with chronic myelogenous leukemia typically has an increased number of the granulocytic white blood cells, (neutrophils, eosinophils and basophils). The granulocytic white blood cells have granules present in their cytoplasm.

Test Your Knowledge 12-5

Describe the appearance of a normal red blood cell on a stained blood smear. (Outcome 12-9)

POINT OF INTEREST 12-2

Infectious mononucleosis

Clinical Facts

- Viral infection caused by the Epstein-Barr virus
- Virus enters through the oropharynx, transmitted in saliva; called "the kissing disease"
- Most prevalent in adolescents and young adults
- Virus infects the B-cell lymphocytes

Clinical Symptoms

- Fatigue
- Fever
- Sore throat
- Swollen lymph glands

Hematological Features

- Slight elevation in the white blood cell count (12,000 to 25,000)
- Increased percentage of lymphocytes (60% to 90%)
- Presence of reactive (atypical) lymphocytes
  - Larger T cells than normally visualized
  - Larger; nuclei with more open chromatin and deep indentations
  - Cytoplasm with basophilia, azurophilic granules, and vacuoles

Immunology Features

- Positive rapid-screen mono test
- Increased heterophile antibody titer

Treatment

- Rest
- NSAIDs, such as ibuprofen
- No lifting or contact sports if the spleen is enlarged

Peripheral Blood Smear

A peripheral blood smear may be used for a manual white blood cell differential as well as visualization of other formed elements in the blood. A blood smear consists of a slide with a drop of blood spread across it. The blood sample may be created directly from a capillary puncture or from an EDTA-anticoagulated venous blood sample that is less than 2 hours old. The smear is then dried, stained, and viewed under the microscope. White blood cells, red blood cells, and platelets are observed. Examination of cell **morphology** (appearance

TABLE 12-4

White Blood Cell Types in the Circulating Blood and Reference Ranges for the Leukocyte Differential

White Blood Cell	Image of Cell	Description of Morphology	Percentage Normally Present (suggested reference range)
Neutrophil	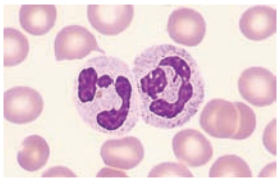 (From Harmening, DM: Clinical Hematology and Fundamentals of Hemostasis, ed. 5. FA Davis, Philadelphia, 2009, with permission.)	Multilobed nucleus; lavender gray granules in cytoplasm	54%–65%
Lymphocyte	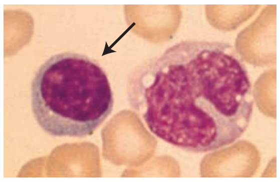 (From Harmening, DM: Clinical Hematology and Fundamentals of Hemostasis, ed. 5. FA Davis, Philadelphia, 2009, with permission.)	Large, dense, dark blue nucleus; small amount of pale blue cytoplasm	25%–40%
Monocyte	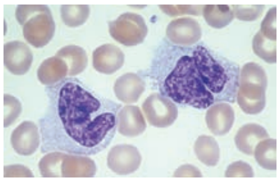 (From Harmening, DM: Clinical Hematology and Fundamentals of Hemostasis, ed. 5. FA Davis, Philadelphia, 2009, with permission.)	Foamy, convoluted nucleus; ground glass–looking cytoplasm	2%–8%
Eosinophil	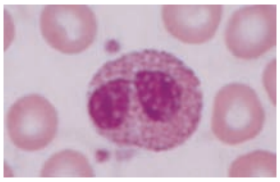 (From Harmening, DM: Clinical Hematology and Fundamentals of Hemostasis, ed. 5. FA Davis, Philadelphia, 2009, with permission.)	Bilobed nucleus; large orange-red granules in cytoplasm	1%–4%

TABLE 12-4—cont'd

White Blood Cell Types in the Circulating Blood and Reference Ranges for the Leukocyte Differential

White Blood Cell	Image of Cell	Description of Morphology	Percentage Normally Present (suggested reference range)
Basophil	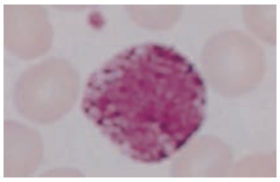 (From Harmening, DM: Clinical Hematology and Fundamentals of Hemostasis, ed. 5. FA Davis, Philadelphia, 2009, with permission.)	Bilobed nucleus (difficult to see clearly); large dark blue granules in cytoplasm	0%–1%
Band cell (immature neutrophil)	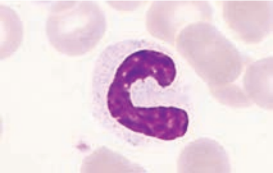 (From Harmening, DM: Clinical Hematology and Fundamentals of Hemostasis, ed. 5. FA Davis, Philadelphia, 2009, with permission.)	C-shaped nucleus; lavender-gray granules in cytoplasm	0%–5%

of the red blood cells including their size, color, and shape) and determination of size are used to identify each type of blood cell. The leukocyte differential, the red blood cell morphology, and a platelet estimate are reported as part of the results. Examination of the blood smear is necessary when the automated results indicate that there is a need to double-check the differential count or cell morphology of the specimen, or when the method used for the CBC does not allow for an automated white blood cell differential.

Test Your Knowledge 12-6

What characteristics of the white blood cells are used to differentiate the different types? (Outcome 12-7)

Blood smears are stained with a **polychromatic stain,** which consists of a combination of an alkaline and acidic component that allows for cells to appear in a variety of colors. Wright's stain is a popular polychromatic stain used for blood smears. Different structures will absorb more of one color or the other, as they will naturally be attracted to the alkaline or acidic components of the stain. The resulting colors and cell structural appearance are used to differentiate the cells. Nuclei are stained blue by the basophilic (alkaline) part of the stain; other structures are stained red orange by the eosinophilic (acid) stain component. Some structures pick up both stain components.

Manual Leukocyte Differential

For performance of a manual differential, the stained blood smear is placed on the stage of microscope and the immersion oil objective (100X) is used to magnify the blood smear. An area where cells appear in a single layer and are not overlapped is located while looking at the slide. The viewing field is moved from side to side across the width of the smear. Typically 100 consecutive white blood cells are observed and accounted for by type. The size of the cell and the structures of the nucleus and the cytoplasm are used to identify each white blood cell type. An electronic or a manual tabulator keeps count. When 100 cells are counted, the percentage of each type is the total number observed.

Test Your Knowledge 12-7

List the five white blood cell types in order from those of the greatest number to those of the least.

(Outcome 12-7)

Red Blood Cell Morphology

As well as counting the different types of WBCs on a differential, it is important to observe the RBCs and pay careful attention to their size and shape. As introduced in Chapter 11, normal red blood cells are circular biconcave disks with pale centers. They are 6 to 8 microns in diameter. They are a soft red color and have no nuclei. Red blood cells are smaller than white blood cells and larger than platelets. The red blood cells on the stained slide should all be similar in size, with natural slight differences in the sizes and shapes. These normal red blood cells are called **normocytic** and **normochromic**. The suffix –cytic refers to the size and shape of the red blood cells and the suffix –chromic refers to the color of the cells when viewed under the microscope.

Normocytic red blood cells are the normal size of 6 to 8 microns. Normochromic red blood cells have a center that is paler than the surrounding cell, which means that they contain a sufficient concentration of hemoglobin. In hematological disorders, the red blood cells may vary in the amount of hemoglobin, the size and shape of the cells, the staining characteristics (pale or dark red) and the structures present within the cells.

Test Your Knowledge 12-8

What is the term used to describe a red blood cell visualized on a stained smear that is paler or lighter in color than a normal erythrocyte? (Outcome 12-9)

Red blood cells that have too little hemoglobin will have larger, paler central areas than normal cells. These cells are described as **hypochromic. Spherocytes** are RBCs that assume a spherical shape and are said to be **hyperchromic,** because they have excess hemoglobin concentration and thus no central pallor. **Polychromatic** (a multicolored appearance achieved using a mixture of varied colored stains) red blood cells appear a little more bluish in hue as they contain RNA that attracts the alkaline component of the stain. **Anisocytosis** is a condition in which there is a variety of sizes of red blood cells present. Cells larger than normal size are considered **macrocytes** and those smaller than normal are **microcytes**.

Test Your Knowledge 12-9

What is the term used when the red blood cells on a smear are varied in shape more than usual?

(Outcome 12-9)

Poikilocytosis describes red blood cells that are present on the slide in a variety of different shapes. Some types of cells are named for their descriptive shape. Examples of these include teardrops, helmet cells, **sickle cells** (arch-shaped cells that result from low oxygen tension and that contain an abnormal substance known as hemoglobin S, and elliptical-shaped cells known as **ovalocytes.** Cells that have a reduced hemoglobin concentration (known as **target cells)** have so little hemoglobin that there is only a slight ring of color at the edges. These target cells have a spot of trapped hemoglobin in the center (the bull's eye of the target). Figure 12-1 includes examples of normal and abnormal red blood cell morphology.

Immature nucleated red blood cells may be present on the slide when there has been an event that puts an extreme demand on the bone marrow for production of new blood cells. These nucleated red blood cells (NRBCs) will have a blue staining nucleus. NRBCs are not normally present in the circulating blood of an adult but they may be seen in fetal blood or that of young infants. If NRBCs are present, it is important to report this as part of the CBC result, as their presence may be clinically significant.

Platelet Morphology and Estimated Count

Platelets are small, oval-shaped, and purple cellular fragments when visualized on the stained blood smear. A platelet estimate may be calculated by counting the number of platelets on a stained blood smear in fields viewed under the microscope where cells are not overlapped and red blood cell morphology is easily visualized. Ten fields are counted, and the average number of platelets per field is multiplied by 20,000 to achieve a platelet estimate. In a normal platelet count, there are 8 to 20 platelets on an oil immersion field.

Manual Blood Cell Counts

Red blood cells, white blood cells, and platelets may be counted manually by a qualified laboratory professional. This is not a CLIA-waived procedure, so in most circumstances it is not a procedure that will be

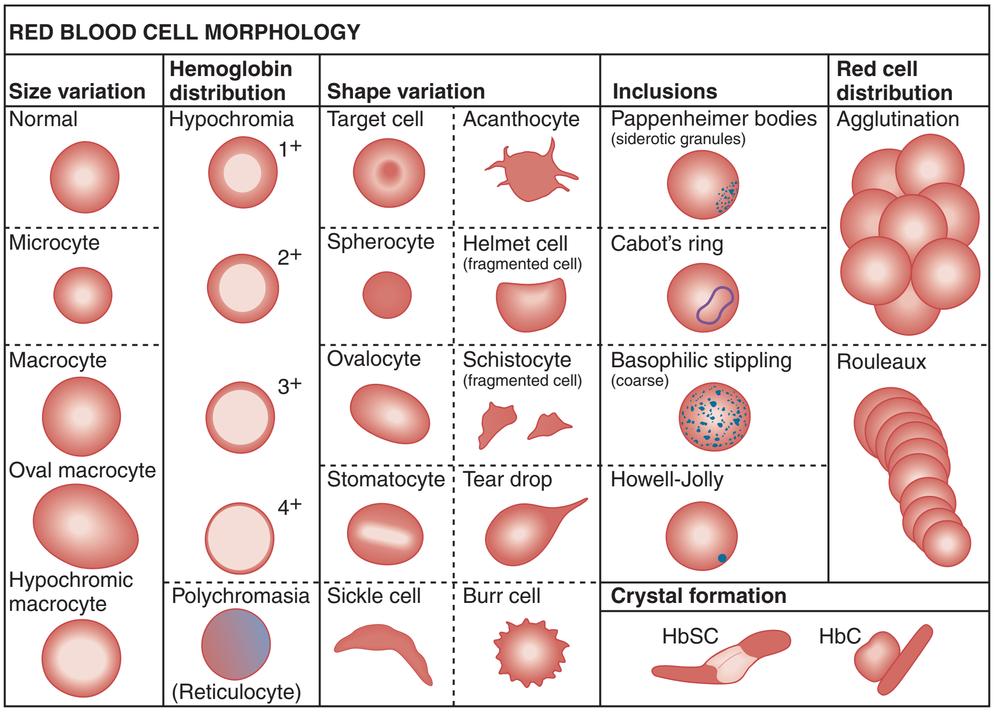
RED BLOOD CELL MORPHOLOGY
Size variation
Hemoglobin distribution
Shape variation
Inclusions
Red cell distribution
Normal
Hypochromia
1+
Target cell
Acanthocyte
Pappenheimer bodies
(siderotic granules)
Agglutination
Microcyte
2+
Spherocyte
Helmet cell
(fragmented cell)
Cabot's ring
Macrocyte
3+
Ovalocyte
Schistocyte
(fragmented cell)
Basophilic stippling
(coarse)
Rouleaux
Oval macrocyte
4+
Stomatocyte
Tear drop
Howell-Jolly
Hypochromic macrocyte
Polychromasia
(Reticulocyte)
Sickle cell
Burr cell
Crystal formation
HbSC
HbC

Figure 12-1 Normal and abnormal red blood cell morphology. *From Harmening, DM: Clinical Hematology and Fundamentals of Hemostasis, ed. 5. FA Davis, Philadelphia, 2009, with permission.*

performed by a medical assistant. A manual cell count may be necessary when the automated system is nonfunctional or in situations in which the cell counts are too low to be measured by the instrument commonly used. For a manual count, blood samples are mixed with a specified amount of diluent (Fig. 12-2A) and added to a **hemocytometer.** The hemocytometer is a chamber of specific dimension and depth that has a grid marked for counting the cells under the microscope (Fig. 12-2B). The cells are counted using the magnification of the microscope. Because of the standardized dilution ratio and dimensions of the hemocytometer, a mathematical calculation is used to determine the number of cells in a milliliter of blood.

AUTOMATED ANALYZERS FOR COMPLETE BLOOD COUNT TESTING

Blood cell analyzers have come into use with advances in technology. In hospitals and reference laboratories, these analyzers allow for large volumes of tests to be performed in a short period of time. Automated instruments in the physician office laboratory have added greater accuracy and precision to the results. Most of these test procedures use **electrical impedance** (a process for counting blood cells that depends on their resistance to the flow of an electrical current) and light-scattering principles to count and differentiate blood cells. Beckman-Coulter, Abbott, ABX, and Bayer-Technicon are companies that produce these automated analyzers for large-scale, time-efficient hematology testing.

In electrical impedance analyzers, the cells pass through an opening with an electrical current flowing through it. The changes of electrical resistance due to the formed elements in the sample are monitored and they are counted as voltage pulses. These pulses are proportional in height to the volume of the cells, which allows them to be identified as white blood cells, red blood cells, or platelets.

Some instruments may also use a beam of light to measure the number and size of the cells in the sample, a process called light scattering. A light-sensitive detector measures the light scattered as the sample is introduced into a chamber and the beam of light is interrupted. The size of the pulse detected is proportional to the size of the cellular element. Many of the analyzers use a combination of these two techniques to count the formed elements in the blood sample. A

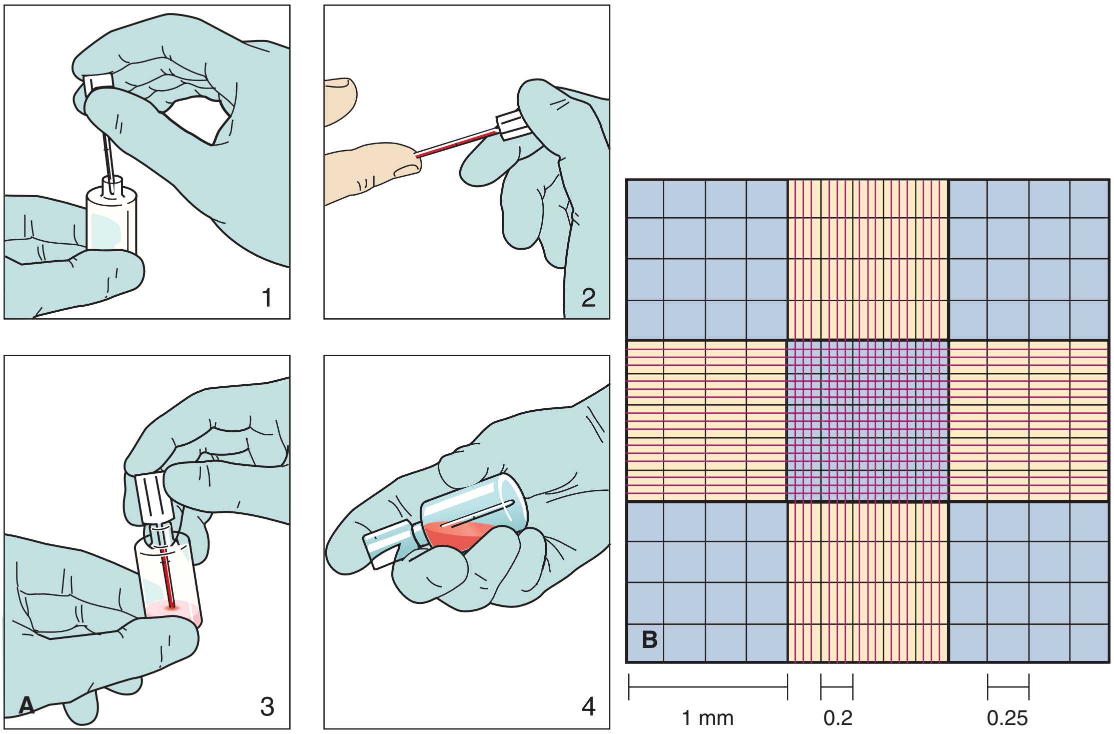

Figure 12-2 (A) Steps involved in using the Unopette system. *From Wedding, ME, and Toenjes, SA: Medical Laboratory Procedures. FA Davis, Philadelphia, 1998, with permission.* (B) Hemocytometer counting chamber. *From Ciesla, B: Hematology in Practice, ed. 2, FA Davis, Philadelphia, 2012, with permission.*

graph of the distribution of cell sizes made during cell counting on an automated analyzer, known as a **histogram,** may also be reported. An example of the results of an automated CBC is shown in Figure 12-3.

The automated white blood cell differential is performed by a combination of techniques. The information gathered using these techniques is combined to classify the type of white blood cell. The cell volume can be determined by electrical impedance, electromagnetic procedures define nuclear characteristics and granular composition, and light-scattering techniques can analyze cell surface and morphological and granular qualities to determine the type of white blood cell. The automated analyzers work well for the vast majority of specimens, but they do have their limitations. Certain samples with abnormal results must have a manual differential performed for accurate results.

The QBC STAR Centrifugal Hematology System (QBC Diagnostics) is an example of an automated instrument that tests one sample at a time (Fig. 12-4). It may be used in the physician office laboratory where automation is desired, but specimen-testing volumes are not high. The QBC STAR system is a CLIA test of moderate complexity, so those performing the testing will participate in proper training of specified quality assurance procedures. The analysis tube used for testing within the instrument may be filled with EDTA-anticoagulated blood or directly from a capillary puncture. The analysis tube contains a dye and a special “float” that is used for the testing process. The tube is filled with blood and placed directly into the analyzer. The testing process involves centrifugation within the instrument and use of the “float” to measure different parameters of the specimen. This instrument is desirable for some office settings because there are no liquid reagents to work with, and the calibration of the instrument is done internally. The analysis tubes contain all the reagents necessary for the procedure. The QBC Star is capable of measuring most of the components of a typical CBC, with the exception of a red blood cell count and some of the indices.

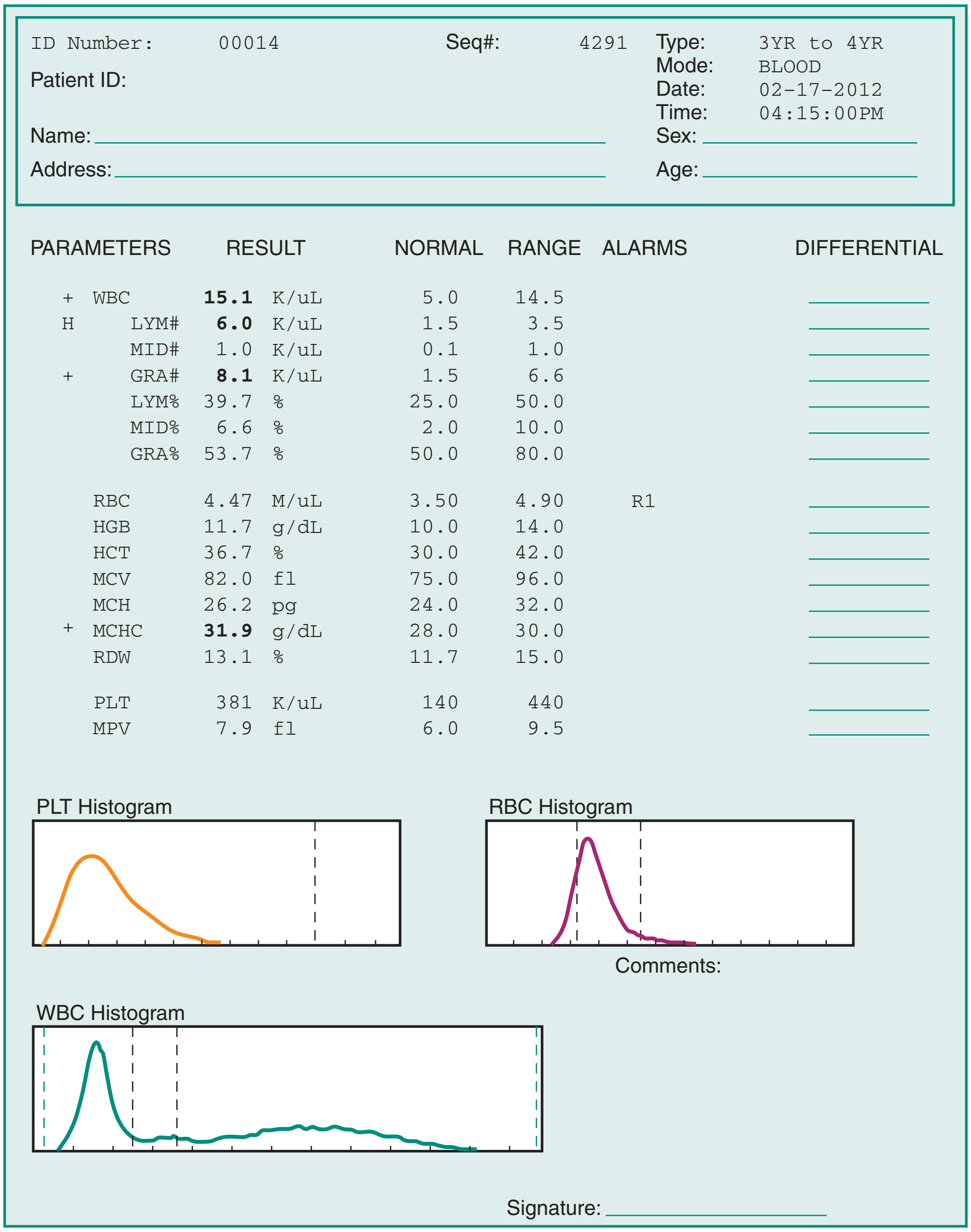

ID Number: 00014

Seq#: 4291

Patient ID:

Name:

Address:

Type: 3YR to 4YR

Mode: BLOOD

Date: 02-17-2012

Time: 04:15:00PM

Sex:

Age:

	PARAMETERS	RESULT		NORMAL	RANGE	ALARMS	DIFFERENTIAL
+	WBC	**15.1**	K/uL	5.0	14.5		
H	LYM#	**6.0**	K/uL	1.5	3.5		
	MID#	1.0	K/uL	0.1	1.0		
+	GRA#	**8.1**	K/uL	1.5	6.6		
	LYM%	39.7	%	25.0	50.0		
	MID%	6.6	%	2.0	10.0		
	GRA%	53.7	%	50.0	80.0		
	RBC	4.47	M/uL	3.50	4.90	R1	
	HGB	11.7	g/dL	10.0	14.0		
	HCT	36.7	%	30.0	42.0		
	MCV	82.0	fl	75.0	96.0		
	MCH	26.2	pg	24.0	32.0		
+	MCHC	**31.9**	g/dL	28.0	30.0		
	RDW	13.1	%	11.7	15.0		
	PLT	381	K/uL	140	440		
	MPV	7.9	fl	6.0	9.5		

PLT Histogram

RBC Histogram

Comments:

WBC Histogram

Signature:

Figure 12-3 CBC result from an automated hematology analyzer, including histograms.

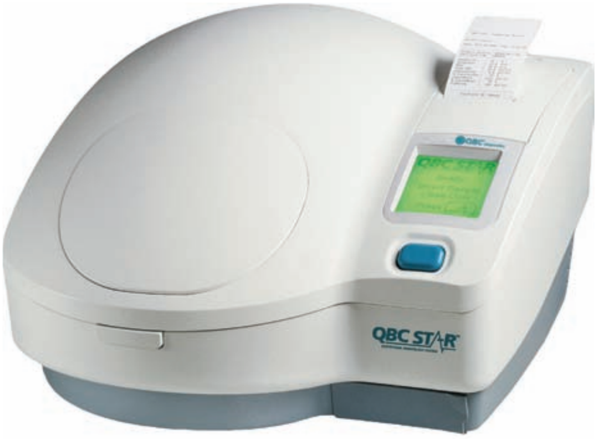

Figure 12-4 QBC STAR hematology analyzer. *Courtesy of QBC Diagnostics.*

SUMMARY

The complete blood count (CBC) is a comprehensive hematology test consisting of a white and red blood cell count, hemoglobin and hematocrit determinations, red blood cell indices, a platelet count, and a white blood cell differential. These tests may be performed manually or by use of automated instrumentation. The three red blood cell indices (MCV, MCH, MCHC) are calculations that numerically describe the size of the red blood cells and the amount of hemoglobin that they contain.

There may be times in which examination of a blood smear is necessary to clarify the differential results or view morphology for other cellular elements in a sample. A stained blood smear is viewed with the microscope for a manual white blood cell differential, red blood cell morphology, and platelet estimation. The complete blood count can be used as a tool to diagnose and monitor disease and to follow the course of treatments.

TIME TO REVIEW

1. Choose the correct list of components of a complete blood count. Outcome 12-2
   a. Hemoglobin, hematocrit, red blood cell count, white blood cell count, erythrocyte sedimentation rate, WBC differential
   b. Hemoglobin, hematocrit, red blood cell count, white blood cell count, red blood cell indices, platelet count, WBC differential
   c. Hemoglobin, hematocrit, red blood cell count, platelet count, erythrocyte sedimentation rate, WBC differential
   d. Hemoglobin, hematocrit, red blood cell count, white blood cell count, hemoglobin electrophoresis, WBC differential, platelet count
2. The reference range for an adult white blood cell count is: Outcome 12-3
   a. 4,300 to 10,800/mm^3
   b. 140,000 to 400,000/mm^3
   c. 5.0 to 6.5 million/mm^3
   d. 12 to 14 g/dL
3. A white blood cell count may be decreased because of: Outcome 12-5
   a. A chronic viral infection
   b. Exposure to toxins
   c. Chemotherapy use
   d. All of the above
4. The reference or normal range for the platelet count is: Outcome 12-3
   a. 4,300 to 10,800/mm^3
   b. 150,000 to 450,000/mm^3
   c. 5.0 to 6.5 million/mm^3
   d. 12 to 14 g/dL
5. This type of white blood cell is the most numerous in a healthy adult and makes up about 60% of the white blood cells: Outcome 12-7
   a. Eosinophil
   b. Monocyte
   c. Neutrophil
   d. Lymphocyte
6. This is the term used to describe a variety of sizes of red blood cells on the stained smear: Outcome 12-9
   a. Anisocytosis
   b. Poikilocytosis
   c. Spherocytosis
   d. Hemolysis
7. This type of cell appears as small, purple, circular, or ovoid fragments on the stained blood smear. Outcome 12-10
   a. White blood cell
   b. Red blood cell
   c. Platelet
   d. Band cell

8. An increased white blood cell count may be evident with which disease? Outcome 12-4

a. Thrombocythemia
b. Anemia
c. Leukopenia
d. Leukemia

9. Microcytosis is a word used to describe: Outcome 12-9

a. Large red blood cells
b. Large white blood cells
c. Small red blood cells
d. Small platelets

10. A hemocytometer is used to: Outcome 12-1

a. Measure a hematocrit level
b. Perform manual white blood cell counts
c. Perform staining of peripheral blood smears
d. Measure hemoglobin concentrations

Case Study 12-1: Clinical significance

Mr. Smith, a 67-year-old patient, presented to the office with symptoms of a cough, a little tightness in his chest, and complaints of a slight fever. The medical assistant took Mr. Smith's vital signs and wrote his results and chief complaint in his chart. When the physician examined the patient, he believed that Mr. Smith was suffering from a respiratory infection. His plan was to prescribe an antibiotic if the infection turned out to be caused by bacteria, so he asked the medical assistant to draw blood for a CBC.

The medical assistant performed a venipuncture on Mr. Smith and processed the sample for pickup by the reference laboratory for testing later that day. She explained to Mr. Smith that the results should be back in the office in the morning and if the CBC results indicate the presence of a bacterial infection, the physician will call in a prescription for an antibiotic to his pharmacy. She told the patient that she would call him in the morning.

The next morning these results were faxed to the office:

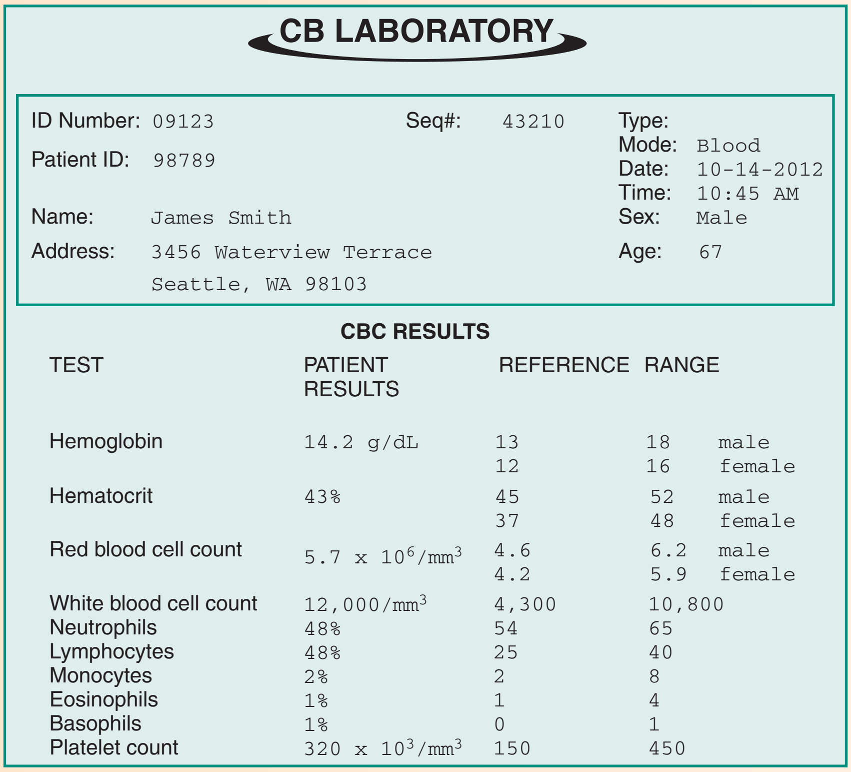

CB LABORATORY

ID Number: 09123
Seq#: 43210
Patient ID: 98789
Name: James Smith
Address: 3456 Waterview Terrace
Seattle, WA 98103
Type:
Mode: Blood
Date: 10-14-2012
Time: 10:45 AM
Sex: Male
Age: 67

CBC RESULTS

TEST	PATIENT RESULTS	REFERENCE	RANGE	
Hemoglobin	14.2 g/dL	13	18	male
		12	16	female
Hematocrit	43%	45	52	male
		37	48	female
Red blood cell count	$5.7 \times 10^6/\text{mm}^3$	4.6	6.2	male
		4.2	5.9	female
White blood cell count	$12{,}000/\text{mm}^3$	4,300	10,800	
Neutrophils	48%	54	65	
Lymphocytes	48%	25	40	
Monocytes	2%	2	8	
Eosinophils	1%	1	4	
Basophils	1%	0	1	
Platelet count	$320 \times 10^3/\text{mm}^3$	150	450	

Continued

The medical assistant highlighted the abnormal results, retrieved James Smith's chart, and took the laboratory report and chart to the physician. The doctor reviewed the results and asked the medical assistant to contact Mr. Smith to let him know that he would not be prescribing antibiotics at this time.

The medical assistant called Mr. Smith and explained that the results were not consistent with a bacterial infection. She told Mr. Smith that no antibiotics would be prescribed at this time, but to check back in a week if the symptoms had not gotten better or if things appeared to be worse.

1. What color tube was used for the blood sample?
2. What out-of-range laboratory results should the medical assistant have highlighted for the physician?
3. Are the abnormal laboratory results consistent with a bacterial infection?

RESOURCES AND SUGGESTED READINGS

"QBS Star"
Provides an overview of the function of the QBC Star hematology analyzer http://www.qbcdiagnostics.com

Chapter 13

Hemoglobin and Hematocrit

Nikki A. Marhefka, EdM, MT(ASCP), CMA (AAMA)

CHAPTER OUTLINE

- Hemoglobin
- Hemoglobinopathies
- Hemoglobin Testing
  - CLIA–Waived Hemoglobin Testing
  - Potential Errors in Hemoglobin Testing
- Hematocrit
  - CLIA–Waived Hematocrit Testing
  - Potential Errors in Hematocrit Testing
- The Relationship of Hemoglobin and Hematocrit Values
- Anemia
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

13-1 Define the key terms.

13-2 Describe the structure and function of a hemoglobin molecule.

13-3 State the type of normal adult hemoglobin and list other types present with various hemoglobinopathies.

13-4 Analyze the general principle of and sources of error associated with CLIA-waived testing of hemoglobin.

13-5 Evaluate the concept of the hematocrit.

13-6 Identify the layers into which the blood sample has separated in a spun hematocrit specimen.

13-7 Examine the general principle of and sources of error associated with CLIA-waived hematocrit procedures.

13-8 Recognize the normal reference ranges for hemoglobin and hematocrit.

13-9 Define *anemia* and examine its clinical significance.

13-10 List the various types of anemia and explain how they are classified.

CAAHEP AND ABHES STANDARDS

CAAHEP 2008 Standards

I.C.I. Anatomy and Physiology #6: Identify common pathology as it relates to the interaction of body systems.
I.C.I. Anatomy and Physiology #7: Analyze pathology as it relates to the interaction of body systems.
I.P.I. Anatomy and Physiology #12: Perform hematology testing.

ABHES 2010 Standards

10. Medical Laboratory Procedures, b. CLIA-waived tests, Graduates: 2. Hematology Testing.

KEY TERMS

- Anemia
- Anisocytosis
- Aplastic anemia
- Bilirubin
- Cyanmethemoglobin
- Erythropoiesis
- Hematocrit (hct)
- Hb A
- Hemoglobin (hgb)
- Hemoglobin and hematocrit (H&H)
- Hemoglobin electrophoresis
- Hemoglobinopathy
- Hemolysis
- Heterozygous
- Homozygous
- Hypochromic
- Interstitial fluid
- Intrinsic factor
- Iron-deficiency anemia
- Macrocytic
- Microcytic
- Packed cell volume
- Pernicious anemia
- Poikilocytosis
- Protoporphyrin
- Sickle cell anemia
- Sickle cell disease
- Sickle cell trait
- Splenomegaly
- Target cell
- Thalassemia
- Urobilinogen

Two of the most important hematology test procedures are for **hemoglobin (hgb)**, a protein responsible for oxygen transport, and the percentage of the space occupied by red blood cells compared to the entire blood volume known as the **hematocrit (hct)**. Measurement of the **hemoglobin and hematocrit** (abbreviated as **H&H**) is commonly performed in all laboratories and physician offices. The results of these two tests may be used to help diagnose or monitor the progress of treatment for **anemia,** a condition in which the oxygen-carrying capacity of the blood is reduced. This chapter presents additional information about the clinical significance of the hemoglobin and hematocrit measurements, and provides details of testing methods for these parameters.

HEMOGLOBIN

As was established in previous chapters, mature red blood cells are filled with fluid as they are released from the bone marrow. The intracellular fluid is primarily made up of hemoglobin. Oxygen attaches to this protein complex for subsequent transport to the tissues of the body, where it is needed for the production of energy. The primary function of hemoglobin (and thus, the red blood cells) is to transport oxygen to the tissues of the body. Hemoglobin is composed of four molecules of heme within a globin protein chain. Each of these heme molecules has one atom of iron, and one molecule of oxygen can attach to each of these iron atoms. Thus, one red blood cell can transport four molecules of oxygen.

Hemoglobin production begins in the erythroid precursor cells of the bone marrow. The heme is synthesized by the mitochondria of the bone marrow cells, and the globin chains are constructed by the ribosomes of the cells as they mature in the bone marrow. Most of the hemoglobin is formed by the early precursor cells, which are nucleated, but the production continues even after these cells have lost their nucleus until they are released into the general circulation (Fig. 13-1).

Test Your Knowledge 13-1

Hemoglobin is made up of one molecule of __________ and four molecules of _________. (Outcome 13-2)

Test Your Knowledge 13-2

What is the function of hemoglobin? (Outcome 13-2)

The hemoglobin molecule becomes oxygenated in the lungs where the oxygen concentration is very high. As the red blood cells leave the pulmonary circulation, the affinity of tissues for the oxygen carried by the hemoglobin molecule will cause the oxygen to diffuse through the red blood cell membrane and across capillary walls to enter the tissues. After the oxygen has left the red blood cells, the cells are capable of transporting carbon dioxide back to the lungs to be expelled. With the delivery of oxygen and transport of

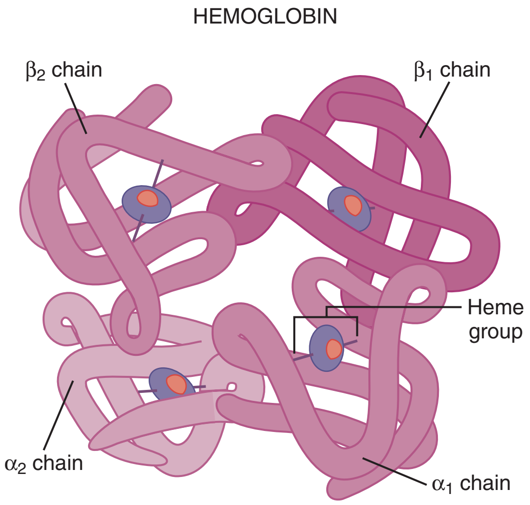

Figure 13-1 A hemoglobin molecule. Note the polypeptide changes with the heme molecules in the center of the complex.

carbon dioxide, the hemoglobin molecules of the red blood cells help the body control the acid-base balance that is necessary for survival.

Mature red blood cells have a life span of approximately 120 days. During their lifetime, they must travel constantly through the cardiovascular system, remaining flexible enough to pass through capillary beds to perform their duties. Eventually, worn-out red blood cells are removed by the tissue macrophage system in the spleen and the hemoglobin is degraded into iron, **protoporphyrin,** and globulin. Protoporphyrin is included in the heme portion of the hemoglobin molecule, and is necessary to make the iron functional and available to attach to the oxygen. The iron released by the breakdown of old red blood cells is removed from the heme, and transported to the bone marrow to be used again. The globulin is degraded into amino acids to be used by the body for creation of protein. Protoporphyrin is converted into **bilirubin,** which combines with albumin for transport to the liver. The liver cells further process the bilirubin and it is excreted as bile, which enters the intestines where bacteria convert it to **urobilinogen** to be excreted in the feces.

HEMOGLOBINOPATHIES

The normal adult hemoglobin is denoted as **Hb A.** Hemoglobin in red blood cells of the fetus and the newborn is Hb F. There are genetic hemoglobin variants present in the population, in which the structure of the globin part of the molecule is altered. The presence of an abnormal hemoglobin is known as a **hemoglobinopathy.** The hemoglobinopathies are a result of genetic changes that may be inherited or caused by mutations. Both the **homozygous** form in which the patient has two identical genes for the same characteristic or the **heterozygous** form that contains two genes for abnormal hemoglobin variants can lead to a mild or severe anemia. If Hb A is present in the cells as well as the abnormal hemoglobin, the abnormality may be masked and no symptoms will be evident. **Sickle cell disease** and **sickle cell trait** are a result of the presence of Hb S. Other hemoglobinopathies may be asymptomatic, whereas others can lead to severe anemia.

POINT OF INTEREST 13-1

Hemoglobinopathies

The hemoglobin molecule must have a very specific structure in order to function optimally. Any change in the structure of the globin chains in the hemoglobin molecule may be classified as a hemoglobinopathy. Many of these abnormalities remain undetected because they are not clinically significant. Others, however, will have a significant effect on the body's ability to transport oxygen.

- **Methemoglobinemia:** For the hemoglobin molecule to function appropriately, it is necessary for the iron molecule to be ferrous. This means that the iron molecule is positively charged with two more protons than electrons present in the molecule. If the iron is present in a state other than this, the hemoglobin molecule has a poor affinity for oxygen. This results in a state of cyanosis for the body, because the red blood cells cannot deliver the oxygen appropriately if they do not pick it up as they should in the lung tissues.
- **Hemoglobinopathies with decreased oxygen affinity:** Certain abnormalities of the hemoglobin molecule may create a situation in which there is a pronounced decrease in the affinity for oxygen. This means that the molecules do not pick up the oxygen as they should in the lungs. This type of hemoglobinopathy is characterized by cyanosis. Hb Seattle and Hb Vancouver are examples of this type of hemoglobinopathy.
- **Increased oxygen affinity:** Erythrocytosis is a common symptom of hemoglobinopathies with increased oxygen affinity. In this situation, the molecule holds on so tightly to the oxygen molecules that they cannot be deposited throughout the

Continued

body. The kidneys will increase their secretion of erythropoetin because of the reduced oxygen in the tissues of the body. Hb Chesapeake is an example of this type of hemoglobinopathy.

- **Hemolytic hemoglobinopathies:** Hb S and Hb C are abnormalities of the hemoglobin molecule structure that affect the integrity of the red blood cell molecule as well as the hemoglobin molecule. These abnormalities occur because of the way that the globin chains are interconnected. The red blood cell membranes have increased hemolysis because the hemoglobin molecule formation is unstable. This has a profound effect on the life span of the red blood cell. The high levels of hemolysis lead to a constant state of anemia, jaundice, and cholelithiasis.

Hb S (sickle cell) also may cause the vessels of the body to become occluded due to the distinct shape of the red blood cells. The abnormal crescent shape becomes lodged in the small capillaries of the body, with multiorgan effects.

Thalassemias are a group of inherited disorders in which abnormal hemoglobin is present in the red blood cells. In thalassemia, anemia results either from a defective production rate of the alpha or beta globin chains that make up the globin portion of the normal hemoglobin molecule. Without these chains, it is not possible for the red blood cell precursors to build Hb A, the normal hemoglobin present in the adult red blood cell. Those who are homozygous for the defective gene have severe hypochromic, microcytic anemia, and a variety of other issues related to the severe anemia (Table 13-1).

TABLE 13-1

Thalassemias

Alpha Thalassemias	Beta Thalassemias
Insufficient alpha globin protein produced	Insufficient beta globin protein
Clinical defect: A gene or genes related to the alpha globin protein of hemoglobin is/are missing or changed	**Clinical defect:** Gene defects affect the production of the beta globin protein of hemoglobin
Frequency: Found in persons of Southeast Asian, Middle Eastern, Chinese, and African descent	**Frequency:** Found in persons of Mediterranean descent, and to a lesser extent in Chinese, other Asians, and African Americans
Subtypes • Major Hydrops fetalis—babies die before or shortly after birth • Hemoglobin H disease, moderate to severe anemia • Minor: Called alpha thalassemia trait, mild anemia. • Alpha thalassemia silent carrier has the abnormal gene, but it is not evident because these genes are heterozygous.	**Subtypes** • Major Cooley's anemia—severe anemia • Intermediate, moderate anemia • Minor: Called beta thalassemia trait

*Thalassemia is a genetic defect in the protein portion of the hemoglobin molecule and affects alpha or beta globin protein production. Both globin proteins are necessary to create healthy hemoglobin.Thalassemia causes the body to make fewer healthy red blood cells because they are hemoglobin deficient and are not as well equipped to transport oxygen as normal hemoglobin.

Test Your Knowledge 13-3

List two abnormal types of hemoglobin. (Outcome 13-3)

When a hemoglobinopathy is suspected, identification of the hemoglobin variant present in the red blood cells can be accomplished through **hemoglobin electrophoresis.** In this technique, blood for examination is placed on a cellulose acetate gel support medium that is capable of transmitting an electrical current. An alkaline electrolyte solution is added, and an electrical current is applied to the support medium. Each hemoglobin subtype moves across the support medium at a different rate and creates a distinctive pattern. The bands of hemoglobin are detected by staining and identified by their placement on the medium as they have moved. This is a sophisticated test that is performed in reference or hospital laboratories.

HEMOGLOBIN TESTING

Hemoglobin testing is one of the most frequently performed laboratory procedures. The results can be used to detect anemia and provide information to determine its severity. Hemoglobin testing may also be used to monitor the treatment for anemia. Hemoglobin

TABLE 13-2

Hemoglobin Reference Ranges (Normal Values)

Age/Gender	Hgb (g/dL)
Adult male	14–18
Adult female	12–16
Newborn	17–23
Two-month-olds	9–14
Child, 1–14 years	11.3–14.4

reference ranges vary by age and gender of the patient (Table 13-2). The normal reference ranges for the adult male are 13 to 18 g/dL, and the reference ranges for the adult female are 12 to 16 g/dL. Women of child-bearing age have a lower normal reference range than do men due to the blood loss experienced as a result of the monthly menstrual cycle. Infant hemoglobin values are significantly lower than those of adults, with an average of 11 to 14 g/dL. Children's hemoglobin values gradually increase from infancy levels to adulthood.

Hemoglobin is usually converted to another compound, called **cyanmethemoglobin,** for measurement. Cyanmethemoglobin is a colored pigment, so the change in color once this conversion has been accomplished can be measured as absorbance by a spectrophotometer, using a wavelength of 540 nm. To release the hemoglobin from the cells to be converted and measured, the red blood cell membranes must be broken or dissolved, a process called **hemolysis.** The cyanmethemoglobin reagents used for hemoglobin testing contain a surfactant that promotes rapid hemolysis and formation of the cyanmethemoglobin moledule. Large complex analyzers used in reference and hospital laboratories as well as CLIA-waived point of care testing methods use this basic procedure.

CLIA-Waived Hemoglobin Testing

The CLIA-waived hemoglobin procedures are performed by handheld, point-of-care instruments. These analyzers have been approved by the U.S. Food and Drug Administration (FDA) for performance of hemoglobin testing. Each instrument determines the hemoglobin values with minimal steps involved in the testing process. The HemoCue Hb instrument (manufactured by HemoCue Inc.), and i-STAT (manufactured by Abbott) are examples of common CLIA-waived hemoglobin analyzers used in physician offices (Fig. 13-2). There are similarities in the testing methods for all the CLIA-waived hemoglobin testing procedures. These include the following:

1. For most of these instruments, blood is collected by capillary puncture and placed on a reaction chamber. The reaction chamber includes the necessary chemicals to lyse the red blood cells and allow the conversion of the hemoglobin to cyanmethemoglobin. For many of these procedures, blood may also be taken from a tube containing anticoagulant.
2. The reaction chamber is placed into the analyzer, and the color intensity is read by a spectrophotometer to determine hemoglobin content.
3. The handheld instruments require very little maintenance or calibration procedures to produce reliable results. However, it is imperative that the operator follow all manufacturer recommendations for calibration and quality control procedures. These results must be logged appropriately, and must all be within an acceptable range before patient testing is performed on the instrument.
4. Test results are available as a digital readout on the instrument. In addition, some of the instruments may be connected to a computer for result storage and printout.
5.
6.

Potential Errors in Hemoglobin Testing

Even though the hemoglobin testing procedure is straightforward, it is still possible to make errors that may have significant impacts on the results. Potential

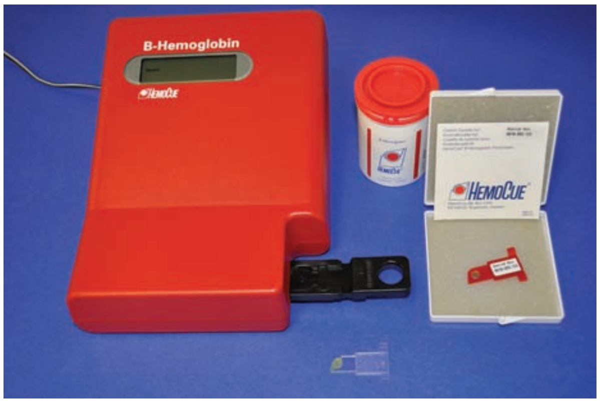

Figure 13-2 A hemoglobinometer.

Procedure 13-1: CLIA-Waived Hemoglobin Testing Using the HemoCue Hemoglobin Analyzer

Hemoglobin analysis is one of the most common CLIA-waived tests performed in physician office laboratories and larger, more complex laboratories. For the HemoCue method, capillary, venous, or arterial blood may be used. Hemoglobin results may be used to screen or monitor the progress of the treatment for anemia.

TASK

Accurately perform hemoglobin testing using a HemoCue hemoglobin analyzer.

CONDITIONS

- Hand-washing supplies and/or alcohol-based hand sanitizer
- Disposable gloves
- Lancet for capillary punctures
- Alcohol prep pad
- Gauze pads
- Biohazardous sharps container
- HemoCue hemoglobin analyzer
- HemoCue microcuvettes
- HemoCue calibration cuvette
- HemoCue control solution
- Disinfectant wipes for cleaning work area

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.12. Perform Hematology Testing

ABHES Standards

- Perform selected CLIA-waived tests that assist with diagnosis and treatment: 2. Hematology testing

Procedure	Rationale
1. Sanitize hands, allow them to dry completely, and apply gloves.	Gloves must be worn for any procedures in which exposure to blood or other potentially infectious materials is anticipated.
2. Turn on the HemoCue instrument, and wait until the digital printout on the screen indicates that the instrument is ready for testing.	The HemoCue instrument will complete a set of internal verification procedures. Once this has been successfully completed, the digital readout on the screen will indicate that the instrument is ready.
3. Verify the expiration printed on the microcuvette container.	Expired supplies are never to be used for control or patient testing.
4. Open the microcuvette bottle, and remove a microcuvette. Close the bottle securely as soon as the desired microcuvette is removed.	The microcuvette container must stay closed securely. These microcuvettes are very sensitive to moisture exposure.
5. Using the quality control material that is designed for this instrument, verify that the instrument is working correctly. Follow the manufacturer's instructions for use and frequency of testing these materials. Fill the microcuvette with the control material, then insert the cuvette into the holder and push into the instrument for analysis. Note the digital readout when ready, and compare the results to those designated as acceptable for this control material.	The quality control material comes with a set of acceptable result ranges. There are materials available to test at low, medium, and high levels of hemoglobin concentration. Verify that the readout on the digital display indicates a result within the acceptable range for that level of control before performing patient testing. Document the result for this quality control material.

Procedure	Rationale
6. Perform a capillary puncture, using appropriate technique. Wipe away the first drop of blood.	Appropriate capillary puncture procedure includes disinfection of the draw site, allowing the alcohol to dry, using an appropriate site, performing the puncture perpendicular to the fingerprints on the finger, and wiping away the first drop of blood that forms. This drop of blood is contaminated with interstitial fluid and if used for the analysis, can produce erroneous results.
7. Touch the open end of the HemoCue microcuvette to the drop of blood. The microcuvette will fill by capillary action to the appropriate level.	Don't touch the capillary puncture site with the microcuvette; it should be placed into the drop of blood. The microcuvette will not overfill. 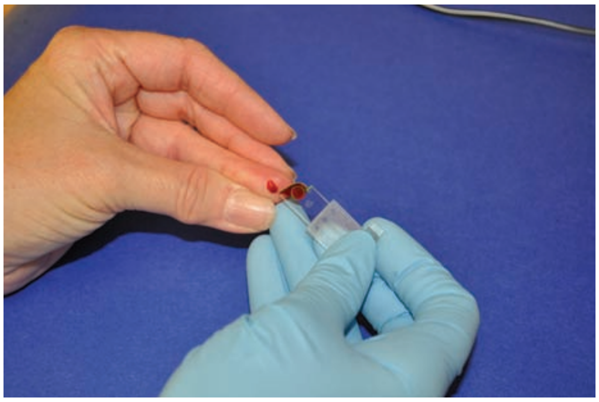
8. Wipe off any excess blood from the bottom and sides of the microcuvette with a gauze pad, then place the microcuvette in the cuvette holder and push the holder into the instrument for analysis.	It is important to wipe away excess blood. However, do not allow the gauze to absorb blood from the interior of the cuvette. 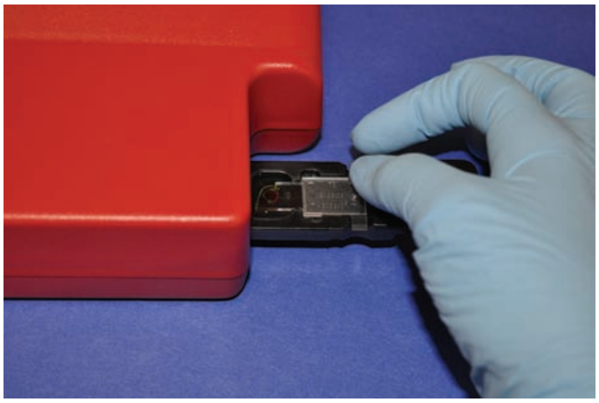 The interior of the microcuvette is coated with a chemical that lyses the red blood cells, and converts the hemoglobin present to cyanmethemoglobin to be measured in the instrument.

Continued

Procedure 13-1: CLIA-Waived Hemoglobin Testing Using the HemoCue Hemoglobin Analyzer—cont'd

Procedure	Rationale
9. The digital display will present a result within 10 to 20 seconds. Document this hemoglobin result.	The result will continue to be displayed until the cuvette holder is pulled out of the instrument. 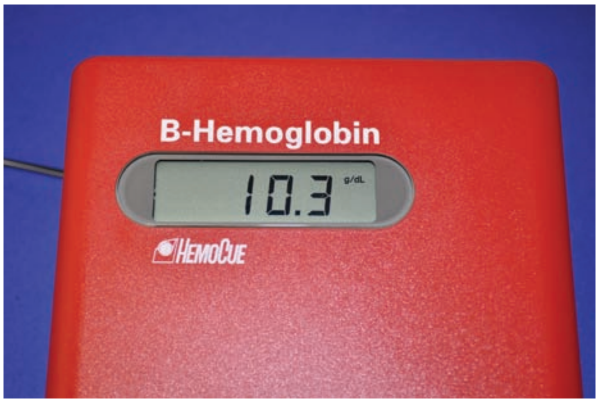
10. After recording the hemoglobin result, pull out the cuvette holder and remove the cuvette. The used cuvette must be discarded as a biohazardous material.	The blood contained in the microcuvette is still considered to be infectious, so it must be disposed of appropriately.
11. Clean and disinfect the work area, then remove gloves and sanitize your hands.	The work area should always be cleaned after use to prepare it for the next user, and hands must always be sanitized afte removing gloves.
12. Document the procedure in the patient's chart if it is available.	In a physician office laboratory, this procedure should be documented in the patient's chart. If the medical assistant is working in a laboratory setting where the chart is not available, this step is not necessary.

Date	
4/2/2012:	*Hemoglobin test performed. QC within range. Hgb 13.3 g/dL* ~~~~~~~~~~~~~~~~
09:30 a.m.	~~~~~~~~~~~~~~~~~~~~~~~~~~ *Connie Lieseke, CMA (AAMA)*

sources of error to take into consideration include the following:

- Adequate blood flow during the capillary collection is necessary to get a homogeneous (well mixed) sample without clots.
- To avoid diluting the blood sample with **interstitial fluid** (the fluid between cells), there should be no squeezing during capillary sample collection and the first drop or two of blood present after the puncture must be wiped away before the blood is applied to the testing chamber or device. If the sample is diluted by interstitial fluid, the hemoglobin value will be erroneously lowered.
- There should be no air bubbles or spaces in the collecting device.

- As is the case with all automated instruments, the reagents and supplies must not be expired when used, and storage requirements for all reagents and samples must be followed as directed.
- The test should be performed within the appropriate time limits after specimen collection as indicated on the instructions for the analyzer.
- All instructions for use of the analyzer must be followed exactly as described.
- Internal and external quality control must be performed at the intervals recommended by the manufacturer, and the results must be within the desired ranges before patient results are reported.

Test Your Knowledge 13-4

Are hemoglobin tests performed on intact red blood cells? (Outcome 13-4)

Test Your Knowledge 13-5

List two potential errors in performing hemoglobin testing. (Outcome 13-4)

HEMATOCRIT

The hematocrit is the percentage of space occupied by red blood cells as compared to the entire blood volume in a sample of whole blood. It may also be called the **packed cell volume.** The hematocrit is an indirect test that provides important information about the concentration of red blood cells in the circulating blood. The space occupied by the red blood cells depends on the size of the cells and the total number of cells present. If there is an appropriate amount of red blood cells with a normal hemoglobin concentration in a specimen, the hematocrit will be in the normal reference range. If the red blood cells are small with insufficient hemoglobin or the number of red blood cells is decreased, the hematocrit value will be below the normal reference range.

The normal reference ranges for hematocrit vary with gender and age of the patient (Table 13-3). Newborns have a very high hematocrit value, but it decreases quickly after birth. The reference range for a child is less than that of an adult, but it continues to increase as they approach adulthood. Adult men have a normal range of 45% to 52%, and adult women have a normal range of 37% to 48%. Hematocrit values are decreased with the various forms of anemia, and may be elevated in situations with an overproduction of red blood cells, such as in polycythemia.

TABLE 13-3

Hematocrit reference ranges (normal values)

Age/Gender	hct Value (%)
Adult male	45–52
Adult female	37–48
Newborn	50–62
Two-month-olds	31–39
Child, 1–6 years	30–40

CLIA-waived procedures for hematocrit testing are manual methods requiring centrifugation of the specimen. Hematocrit values may also be derived by using moderately or highly complex automated instruments. These analyzers will use the hemoglobin concentration and the mean corpuscular volume (MCV) to calculate a hematocrit result.

Test Your Knowledge 13-6

Define *hematocrit.* (Outcome 13-5)

CLIA-Waived Hematocrit Testing

In typical hematocrit testing, the red blood cells are forced to the bottom of the tube by centrifugation (Fig. 13-3). This tube may be known as a capillary tube or a microhematocrit tube because the tube is very small and designed to be spun in a small centrifuge. After centrifugation, the tube is compared to a reading device to measure the hematocrit value. The hematocrit is always reported as a percentage, indicating the volume of red blood cells in the sample as compared to the total blood volume.

There are some variables in the testing process, but the following common components are present in all the CLIA-waived manual hematocrit procedures:

- Two hematocrit tubes are always filled from each patient. Both tubes are centrifuged and results are measured from both tubes. The final result reported will be an average of the two tubes, as long as they are within 2% of each other. If the results from the two tubes are more than 2% apart, the sample will need to be recollected and respun.
- The hematocrit tubes must contain anticoagulant (the tube interior is coated with heparin) if they are to be used for capillary samples drawn directly from the

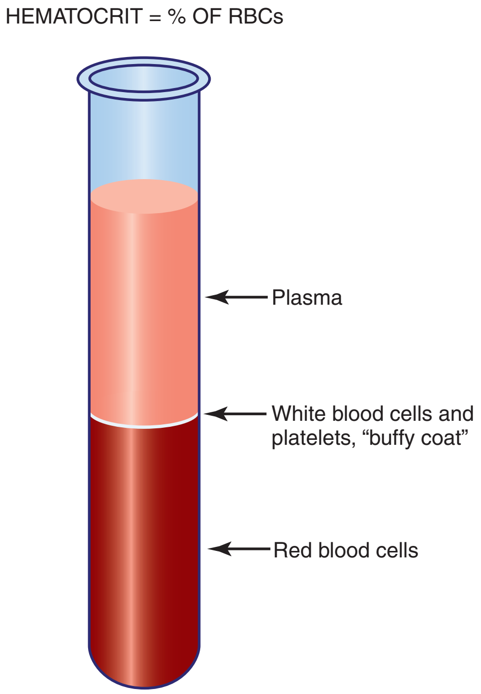

Figure 13-3 Layers of the hematocrit tube. The bottom layer is made up of red blood cells covered with a very thin layer of buffy coat, which represents the white blood cells in the specimen. The uppermost layer is made up of plasma. There is air at the very top of the tube.

patient. If anticoagulant is not used, the blood will clot in the hematocrit tube and the results will be erroneous. It is also possible to use plain tubes (without anticoagulant) if the hematocrit sample is taken from an EDTA lavender top tube.

- Glass tubes should not be used for this process in consideration of employee safety. Safety coated glass tubes are available, which have a glass interior and are covered with a Mylar coating to keep them from breaking. Plastic tubes are also available.
- The sample must be well mixed prior to filling the hematocrit tube (if utilizing an EDTA anticoagulated sample). If a capillary sample is to be used, the site must not be squeezed, as this will add too much interstitial fluid to the sample. The additional fluid may result in a falsely decreased hematocrit result.
- As the tube is filled, air bubbles should be avoided. The addition of bubbles in the specimen can cause the tube to fill insufficiently, in which case it may not be possible to measure the results after centrifugation.
- The end of the hematocrit tube will be plugged with sealing clay to keep the sample from leaving the tube during the centrifugation process.
- Follow the manufacturer's recommendations for the centrifugation times necessary, which may vary with each type of centrifuge.
- Read the results immediately after centrifugation using a built-in reader, or another device designed for hematocrit results. Following centrifugation, the blood in the tube will have been divided into three layers. The hematocrit result is read at the top of the red blood cell column, which is the bottom layer in the tube. There is a thin area called a buffy coat (lighter in color and made of the white blood cells and platelets in the sample) that is visible at the top of the red blood cells, and the plasma makes up the rest of the volume of the sample in the tube (see Fig. 13-3).
- The moderately complex automated procedures available for hematocrit testing are often used in the physician office laboratory. An example of a common instrument is the i-STAT analyzer, made by Abbott Laboratories. The process requires only a few drops of blood. This instrument uses individual cartridges to which the blood sample is added. The hematocrit level is determined by use of an electrical current. As is the case with all CLIA-waived testing procedures, all instructions and quality control procedures recommended by the manufacturer must be followed.

Potential Errors in Hematocrit Testing

The hematocrit testing process may be considered simplistic. However, it is possible to negatively affect the results if not paying close attention to detail. The following are some of the most common errors encountered when using the manual hematocrit procedure:

- If a capillary specimen is used for the procedure, the sample must be collected from free-flowing blood from a skin puncture. If the area is squeezed, interstitial fluid (tissue fluid) may contaminate the sample or the red blood cells can be broken open (hemolysis). The first drop of blood from the skin puncture must be wiped away, as this contains some interstitial fluid as the lancet cut through tissue.
- If the plasma appears pink or reddish, cells may have been hemolyzed during collection. Excessive hemolysis can decrease the value of the hematocrit. Hemolysis may occur during the collection process or as a result of intravascular hemolysis in severe disease states.
- The hematocrit tube must be filled appropriately. If too little blood is added to the tube, it will not be possible to read the result using a reading device after centrifugation. Conversely, the result for a tube that is overfilled cannot be measured using a reading device.

Procedure 13-2: CLIA–Waived Hematocrit Testing Using the StatSpin CritSpin Microhematocrit Centrifuge and Digital Reader

Hematocrit testing is a common CLIA-waived test that is performed in physician office laboratories and larger, more complex laboratories. Spun hematocrit analysis is relatively inexpensive, requires a small amount of blood, and takes only a few minutes. Although capillary spun microhematocrit testing is the most common method, venous blood may also be used for testing. Hematocrit results may be used to screen for or monitor the progress of the treatment for anemia.

TASK

Accurately perform microhematocrit analysis utilizing the StatSpin CritSpin microhematocrit centrifuge and digital reader. The process must be completed within 15 minutes.

CONDITIONS

- Hand-washing supplies and/or alcohol-based hand sanitizer
- Disposable gloves
- Lancet for capillary punctures
- Alcohol prep pad
- Gauze pads
- Biohazardous sharps container
- StatSpin CritSpin microhematocrit centrifuge and digital reader
- Plastic heparinized microhematocrit tubes
- Sealing clay
- HemoCue control solution
- Disinfectant wipes for cleaning work area

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.12. Perform Hematology Testing

ABHES Standards

- Perform selected CLIA-waived tests that assist with diagnosis and treatment: 2. Hematology testing

Procedure	Rationale
1. Sanitize hands, allow them to dry completely, and apply gloves.	Gloves must be worn for any procedures in which exposure to blood or other potentially infectious materials is anticipated.
2. Use the quality control material that is designed for this procedure and this instrument, and verify that the instrument is working correctly. Follow the manufacturer's instructions for use and frequency of testing these quality control materials.	The quality control material comes with a set of acceptable result ranges. There are materials available to test at low, medium, and high levels of hematocrit results. Verify that the result is within the acceptable range for that level of control before performing patient testing. Document the result for this quality control material.
3. Remove two heparinized microhematocrit tubes from the container, and set them on a piece of gauze in the work area.	The tubes need to be within easy reach for the next step of the process. Heparinized tubes must be used for capillary samples because the blood must be anticoagulated to avoid clotting within the tube.

Continued

Procedure 13-2: CLIA-Waived Hematocrit Testing Using the StatSpin CritSpin Microhematocrit Centrifuge and Digital Reader—cont'd

Procedure	Rationale
4. Perform a capillary puncture, using appropriate technique. Wipe away the first drop of blood.	Appropriate capillary puncture procedure includes disinfection of the draw site, allowing the alcohol to dry, using an appropriate site, performing the puncture perpendicular to the fingerprints on the finger, and wiping away the first drop of blood that forms. This drop of blood is contaminated with interstitial fluid and if used for the analysis, can produce erroneous results.
5. Hold the microhematocrit tube by the end with the color-coded band. Touch the opposite end of the tube to the drop of blood. The blood will enter the tube via capillary action. Fill the tube until the blood reaches the colored band at the opposite end.	For best results, do not allow air bubbles to form in the tube. The tube should be filled to the colored band to allow for accurate reading of the results. 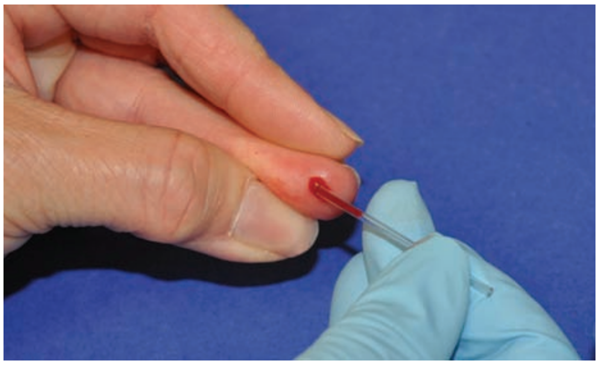
6. Remove the tube from the blood source, and tilt the tube until the blood is halfway between the colored band and the end of the tube.	This step is important to allow for room for the sealing compound at the end of the tube.

Procedure	**Rationale**
7. Hold the microhematocrit tube horizontally and push the end of the tube with the color-coded band into the sealing clay. Twist the tube and remove from the sealing clay. Wipe away any excess blood that may have leaked from the opposite end of the tube.	If the microhematocrit tubes are not sealed in this way, the blood will leak out of the tubes during the spinning process. 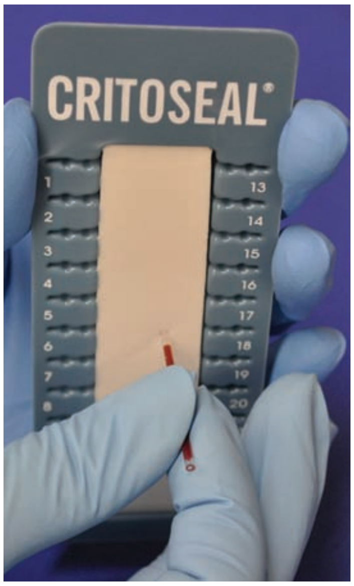 CRITOSEAL® 1 2 3 4 5 6 7 8 13 14 15 16 17 18 19 20
8. Place the tube with the sealed end toward the outer rim of the rotor.	If the sealed end of the tube is not placed to the outer edge of the rotor, all the blood will leak out of the tube when it is spinning.
9. Repeat steps 5 through 8 with another microhematocrit tube. After it is placed in the rotor, screw the rotor cover in place securely.	There should always be two tubes spun for each result to verify the accuracy of the result.
10. Holding the black knob in the center of the rotor, place it in the rotor holder in the centrifuge.	The rotor must be seated securely in place to avoid malfunction of the centrifuge.
11. Close the cover on the centrifuge and start the centrifuge. The StatSpin CritSpin centrifuge requires 2 minutes to fully pack the cells in the microhematocrit tubes.	The time allowed for the centrifugation will vary by method used.
12. When the rotor stops spinning, remove the rotor and place it in the digital reader.	The digital reader is an optional device. A circular or card reader may also be used to read the results.
13. Line up the pointer on the digital reader with the bottom of the cells in the microhematocrit tube and press the “zero” on the reader. Then line up the pointer with the top of the plasma column and set the “100” status on the reader. The final reading is accomplished at the top of the red blood cell column. The result is displayed on the digital reader.	All three steps must be followed to allow for accurate reading of the hematocrit results. Note: If a digital reader is not available, line up the tube with the bottom of the red blood cell column at the “zero” line on a hematocrit reader card, after which the top of the plasma column is to be aligned at the “100” on the card. Read the hematocrit result at the top of the red blood cell column.

Procedure 13-2: CLIA-Waived Hematocrit Testing Using the StatSpin CritSpin Microhematocrit Centrifuge and Digital Reader—cont'd

Procedure	Rationale
14. Record this result.	The result must be recorded before removing the rotor from the system, or going on to read the second tube.
15. Following the same procedure, read the hematocrit result on the second tube.	The results must be within 2% of each other for this to be a valid result. If they are not, the process must be repeated. Report the mean (average) of the two tubes read.
16. Remove the microhematocrit tubes from the rotor and dispose of them as biohazardous sharps.	Although these are plastic tubes, they should still be disposed of in a puncture-resistant container.
17. Clean and disinfect the work area, then remove gloves and sanitize your hands.	The work area should always be cleaned after use to prepare it for the next user, and hands must always be sanitized after removing gloves.
18. Document the procedure in the patient's chart if it is available.	In a physician office laboratory, this procedure should be documented in the patient's chart. If the medical assistant is working in a laboratory setting where the chart is not available, this step is not necessary.

Date	
4/2/2012:	Hematocrit test performed. QC within range. Hct 41%
09:50 a.m.	Connie Lieseke, CMA (AAMA)

The required volume of blood for the hematocrit analysis is determined by following the procedure as written.

- The test may be affected if the timing or the speed of the centrifuge is insufficient. The cells will not pack at the bottom of the tube fully, which may cause the results to be falsely elevated.
- Inaccurate alignment when reading the result will cause errors. The employee must be careful to read only the red blood cell column, and to have the tube placed appropriately against the reading device for accurate results.
- If the sealing clay has not adequately filled the end of the tube, some of the blood will leak out during centrifugation and will affect the result. The end of the tube that is filled with clay is placed to the outside of the centrifuge so that the clay keeps the blood in the tube as the specimen is spun.
- The hematocrit value may be inaccurate immediately after excessive blood loss or following transfusion. Excess plasma may be trapped between the red blood cells in macrocytic, hypochromic, or sickle cell anemias. The cells will not pack as well because of the area occupied by the trapped plasma.

Test Your Knowledge 13-7

What is used to create the layers in the blood sample in manual CLIA-waived hematocrit procedures?

(Outcome 13-6)

THE RELATIONSHIP OF HEMOGLOBIN AND HEMATOCRIT VALUES

Hemoglobin and hematocrit testing is often ordered together as H&H. To verify the validity of the results, it is important to compare the hematocrit and hemoglobin value of the patient's sample. The hematocrit value is approximately three times that of the hemoglobin value in a normal specimen. The hemoglobin value multiplied times three should be within two of the hematocrit results. In addition, the hemoglobin value is about three times that of the red blood cell count. By checking these relationships each time these tests are performed, the medical assistant can verify that the samples are all from the same patient and that the test procedures were performed correctly. Lipemia, hemolysis, and other clinical conditions can make these comparisons inaccurate. If the sample results don't fall within these parameters, the tests should be repeated.

ANEMIA

Anemia is a disorder characterized by the decreased oxygen-carrying capacity of the blood. The following three clinical scenarios may be evident with anemia:

1. A decreased total red blood cell count
2. A decrease in the hemoglobin content of each cell
3. A dysfunctional abnormal hemoglobin content that is not capable of transporting oxygen as it should

To treat anemia, practitioners will attempt to identify the cause. The problem may be chronic inadequate red blood cell production, red blood cell destruction or loss, or the inability of the bone marrow to react to blood loss adequately. Patients with anemia may exhibit signs of fatigue, pallor (pale skin), tachycardia (rapid heartbeats), or dyspnea (difficult breathing). Difficulty with concentration and memory issues may also be common symptoms. These symptoms are all related to the decreased oxygen-carrying capacity of the blood, which starves the cells of the body.

Test Your Knowledge 13-8

Name the disorder where the oxygen-carrying capacity of the blood is reduced. (Outcome 13-9)

Diagnosis and classification of the type of anemia may be accomplished with physical examination results, patient history, and laboratory tests. Patients with anemia may have decreased red blood cell counts, hemoglobin concentrations, or hematocrit values. Some types of anemia exhibit low values in more than one of these parameters, whereas others may demonstrate only one abnormal result. Anemia is often characterized by red blood cell indices that are outside of the normal range as well as specific red blood cell morphology changes.

Red blood cell morphology is observed on the stained blood smear utilized for the white blood cell differential by a trained laboratory technologist. **Anisocytosis** (red blood cells which appear in various sizes) and **poikilocytosis** (red blood cells which appear in various shapes) may be present. The red blood cells of anemic patients will appear abnormal, and may be **microcytic,** (small) **macrocytic** (large), or **hypochromic** (pale in color).

Impaired red blood cell production may be idiopathic (cause unknown) or due to toxins or infection. Anemia may also occur when there is defective synthesis of the heme of the hemoglobin molecule. Acute or chronic hemorrhage may cause anemia when the body is incapable of making enough cells to make up for the loss. Hemolytic anemia occurs with excessive lysis of the red blood cells intravascularly. This can be the result of red blood cell membrane disorders, hemoglobin disorders, metabolic disease, chemical toxins, extensive burns, malaria, or autoimmune destruction of the erythrocytes. A few specific types of anemia are presented below:

- **Iron-deficiency anemia** occurs when there is insufficient iron available for adequate hemoglobin formation. This is a hypochromic, microcytic anemia, as each cell has a reduced amount of hemoglobin. Iron deficiency may be caused by many situations: the body's supply of iron may be low, the absorption of the iron may be impaired, or the demand may be high. As our body breaks down old red blood cells, most of the available iron present in the system is in the tissue macrophage/reticuloendothelial tissue of the spleen. Dietary iron is absorbed in the small intestines.

  Clinical features of iron deficiency anemia may include sores on or a burning sensation of the tongue, ulcers at the corner of the mouth, "pica" (craving for dirt or ice), or chronic gastritis. "Milk anemia" is often a cause in 6- to 24-month-old children, as their diet may not contain sufficient iron. Other causes may include gastrectomy (total or partial) or prolonged treatment of peptic ulcers or acid reflux. Rapid growth during childhood and increased red blood cell production during pregnancy may be causes as well.

- **Aplastic anemias** result from decreased hematopoiesis in the bone marrow that leads to decreased red blood cell production. Only the red blood cells can be affected (**erythropoiesis**) or all blood cell production can be decreased. Radiation, medications, chemicals, or viral infections may be the cause. Chemotherapy agents and radiation treatment for cancer are examples. In most cases, no specific reason for the decreased production of red blood cells is determined. The red blood cell count will be low, and the red blood cell morphology may show anisocytosis and poikilocytosis. If platelet formation is decreased in addition to the low number of red blood cells, bleeding gums and nosebleeds may also be apparent.
- **Pernicious anemia** results from inadequate secretion of **intrinsic factor** by the parietal cells of the stomach lining. Intrinsic factor is necessary for absorption of vitamin B_{12}, which is necessary for red blood cell formation in the bone marrow. The symptoms of pernicious anemia may include pallor, slight jaundice (pale yellow skin), tongue sores, and tingling feelings in the extremities. The red blood cells of the circulating blood are often macrocytic. Autoantibodies to intrinsic factor may be present. The defective absorption of B_{12} may also be associated with celiac disease, malnutrition, or inflammatory disease of the small intestines. Patients with pernicious anemia who lack the intrinsic factor will need to receive B_{12} injections regularly to meet the body's needs.
- **Sickle cell anemia** is caused by a hemoglobinopathy. These patients are homozygous for Hb S; the majority of their hemoglobin is the S variant. The red blood cells of patients with sickle cell anemia are misshapen; they appear crescent or sickle shaped. The red blood cells are destroyed because of their abnormal shape during changes in oxygen tension, The Hb S also causes the cells to be more rigid, which increases their rate of destruction as they try to squeeze through capillaries.

  Sickle cell disease causes **splenomegaly** (an enlarged spleen), as blood pools in the spleen. Patients' hands and feet are sometimes swollen, and patients have increased chance of infection and exhibit joint pain. The abnormally shaped red blood cells become stuck in the blood vessels, causing pain and poor delivery of oxygen to tissues. The red blood cell morphology will include sickle and **target cells** if examined under the microscope.

Test Your Knowledge 13-9

What laboratory tests may be ordered by a physician to verify whether a symptomatic patient may be anemic? (Outcome 13-8)

Test Your Knowledge 13-10

List four examples of specific anemias. (Outcome 13-10)

SUMMARY

Hemoglobin in the erythrocytes is the molecule to which oxygen attaches and travels in the blood to every cell in the body to be available for energy production. The normal hemoglobin of adults is designated as Hb A. The hemoglobin concentration can be measured by analyzers spectrophotometrically, by measuring color change accomplished as the hemoglobin in the red blood cells is changed to cyanmethemoglobin. The type of hemoglobin present in cells can be identified by hemoglobin electrophoresis performed in reference or hospital laboratories.

The hematocrit is the amount of space occupied by red blood cells as compared to the whole sample, expressed as a percentage. Hematocrit values can be calculated by automated analyzers or directly read from a hematocrit tube after spinning in a centrifuge. The hemoglobin and hematocrit tests are the most frequently ordered procedures used to monitor the oxygen-carrying capacities of the body. If these values are below the normal reference range, anemia is suspected, and health-care professionals will use addition information to plan a course of treatment. Anemia is diagnosed by the history and physical examination of the patient, as well as laboratory test results. The results of laboratory tests that describe the volume and other characteristics of erythrocytes are used to classify the type of anemia present and establish a course of treatment.

TIME TO REVIEW

1. A degradation product of hemoglobin is: Outcome 13-1
   - **a.** Anisocytosis
   - **b.** Intrinsic factor
   - **c.** Bilirubin
   - **d.** Polycythemia

2. __________ from the tissues can contaminate hemoglobin or hematocrit specimens and decrease the result. Outcome 13-1
   - a. Anemia
   - b. Interstitial fluid
   - c. Intrinsic factor
   - d. Erythropoiesis

3. Hemoglobin is made of chains of globin and four molecules of: Outcome 13-2
   - a. Heme
   - b. Sodium
   - c. Cyanmethemoglobin
   - d. Urobilinogen

4. __________ attaches to the hemoglobin and travels from higher oxygen concentration levels in the lungs to the cells of the body with increased oxygen need. Outcome 13-2
   - a. Iron
   - b. Vitamin B_{12}
   - c. Glucose
   - d. Oxygen

5. The type of hemoglobin normally seen in adults is ___________. Outcome 13-3
   - a. Hb A
   - b. Hb C
   - c. Hb F
   - d. Hb S

6. Hemoglobin analyzers determine hemoglobin concentration by using: Outcome 13-4
   - a. Electrophoresis
   - b. Spectrophotometry
   - c. Centrifugation
   - d. Visual examination of the cells

7. Restate the adult reference ranges for hemoglobin and hematocrit testing. Outcome 13-8

8. Packed cell volume is a synonym for ____________. Outcome 13-5
   - a. Hemoglobin
   - b. Hematocrit
   - c. Sickle cell trait
   - d. Polycythemia

9. List the layers of the blood sample of a spun hematocrit. Outcome 13-6

10. Anemia may be the result of: Outcome 13-9
    - a. Excessive blood loss
    - b. Inadequate red blood cell production in the bone marrow
    - c. Abnormal hemoglobin molecule structures
    - d. All of the above

11. Name the anemia caused by an insufficient iron supply in the body. Outcome 13-10
    - a. Hemolytic anemia
    - b. Iron-deficiency anemia
    - c. Pernicious anemia
    - d. Sickle cell anemia

12. A substantial decrease in red blood cell production may lead to a(n) __________ anemia. Outcome 13-10
    - a. Iron-deficiency
    - b. Aplastic
    - c. Pernicious
    - d. Sickle cell

Case Study 13-1: Why am I so tired?

A 35-year-old male patient came to the medical office with complaints of being weak, lethargic, and unable to focus on tasks. The health-care provider assessed the symptoms and ordered a hemoglobin and hematocrit test to be performed. The results were Hgb 12.3 g/dL and Hct 38%.

Based on the result, the physician ordered another blood draw for iron studies.

1. What disorder was indicated by the results?
2. Why did the physician order the iron studies?

RESOURCES AND SUGGESTED READINGS

Hemoglobal organization; details about thalassemia http://www.hemoglobal.org/whatisthalassemia.html

Mayo Clinic information about anemia http://www.mayoclinic.com/health/anemia/DS00321

Information about the various HemoCue testing instruments http://www.HemoCue.com

Chapter 14

Erythrocyte Sedimentation Rate

Nikki A. Marhefka, EdM, MT(ASCP), CMA (AAMA)

CHAPTER OUTLINE

Erythrocyte Sedimentation Rate
Plasma Proteins Affecting the Erythrocyte Sedimentation Rate
The Influence of Red Blood Cells on the Erythrocyte Sedimentation Rate
Reference Ranges
Clinical Significance of Erythrocyte Sedimentation Rate Testing
Erythrocyte Sedimentation Rate Determination
    CLIA–Waived Erythrocyte Sedimentation Rate Methods
    Automated Erythrocyte Sedimentation Rate Methods
Potential Sources of Error for the Erythrocyte Sedimentation Rate Procedure
Summary
Time to Review
Case Study
Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

14-1 Define the key terms.

14-2 Analyze the effect of plasma proteins on erythrocyte sedimentation rates.

14-3 Describe the effects of abnormally shaped red blood cells in normal and abnormal erythrocyte sedimentation rates.

14-4 Recognize normal reference ranges for erythrocyte sedimentation rates for males and females.

14-5 List the pathological and nonpathological states that may exhibit an increased sedimentation rate.

14-6 Describe the CLIA-waived testing methods for the erythrocyte sedimentation rate.

14-7 Correctly perform an erythrocyte sedimentation rate.

14-8 Highlight procedural errors that can cause erroneous erythrocyte sedimentation rate results.

CAAHEP AND ABHES STANDARDS

CAAHEP 2008 Standards

I.P.I.12: Anatomy and Physiology: Perform Hematology Testing

I.C.I.6: Identify common pathology related to each body system

ABHES 2010 Standards

Medical Laboratory Procedures; Perform selected CLIA-waived tests that assist with diagnosis and treatment; #2: Hematology Testing

KEY TERMS

- Albumin
- Aliquot
- Erythrocyte sedimentation rate (ESR, sed rate)
- Fibrinogen
- Globulin
- Inflammation
- Pipette
- Rouleaux
- Wintrobe
- Westergren

The erythrocyte sedimentation rate is one of the oldest laboratory tests still performed today. The test provides a measurement indicating how far the red blood cells settle in a specimen over a given period of time. The proteins present in the plasma and the specific conditions of the red blood cells in the specimen influence this rate. The erythrocyte sedimentation rate test may be performed using automated methods in hospital and reference laboratories or using CLIA-waived manual methods, which may be performed in any laboratory. The presence of **inflammation** (the protective reaction of the body's tissues to injury or irritation), acute or chronic infection, or neoplasms may be characterized by elevated ESR results. However, normal physiology may also cause the erythrocyte sedimentation rate to be elevated.

ERYTHROCYTE SEDIMENTATION RATE

The erythrocyte sedimentation rate (also known as the **ESR** or **sed rate**) is used (in addition to other more specific laboratory tests and patient history) to screen for inflammatory processes in the body. The results of an ESR are nonspecific; an abnormal result does not provide information about the type or cause of the inflammation, but only that it is present. In most situations, an EDTA (lavender top) anticoagulated venous blood sample is used for the test. (Some test methods require a different type of tube, specified later in the chapter.) A long, hollow, plastic tube (**pipette**) is filled with blood (an **aliquot**) which has been diluted with a citrate or saline solution, and the tube is placed upright on a stable, vibration-free, level surface. Gravity causes the red blood cells to settle out of solution and fall to the bottom of the pipette. After 1 hour, the amount that the red blood cells settled is measured in millimeters. The results of the ESR are reported in millimeters per hour (mm/hr).

Test Your Knowledge 14-1

What is measured by the erythrocyte sedimentation rate?

(Outcome 14-1)

PLASMA PROTEINS AFFECTING THE ERYTHROCYTE SEDIMENTATION RATE

Blood plasma normally contains the proteins known as **albumin**, **globulin**, and **fibrinogen.** Inflammatory processes or infection may cause an increase in the levels of globulin and fibrinogen in the body. When these protein levels are elevated, this will increase the value of the erythrocyte sedimentation rate. Elevated amounts of globulin and fibrinogen decrease the negative charge on the surface of the red blood cells, which means that the cells are not as repelled from one another as they would be normally. This causes the cells to become somewhat "sticky," allowing the red blood cells to stack on top of one another like coins (Fig. 14-1). This configuration is known as **rouleaux** formation. Rouleaux are not uncommon occurrences, but the increase in plasma protein concentrations will cause it to become more pronounced. The increased rouleaux will cause the ESR result to be elevated because these stacked cell formations fall faster than single cells. The plasma proteins influence the settling most when the patient's hematocrit is between 30% and 40%.

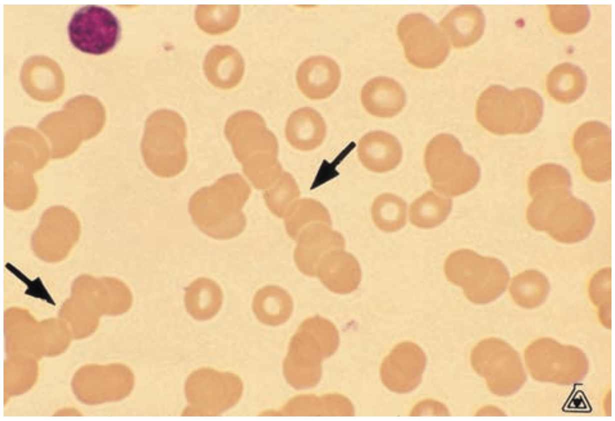

Figure 14-1 Rouleaux formation. Note how the red blood cells "stack" on top of one another, resembling stacks of coins. *From the College of American Pathologists, with permission.*

Test Your Knowledge 14-2

What happens to the red blood cells in the rouleaux formation? (Outcome 14-1)

Test Your Knowledge 14-3

What plasma proteins contribute to the formation of rouleaux when increased in concentration? (Outcome 14-2)

Test Your Knowledge 14-4

Why do red blood cells in rouleaux formation settle faster than single red blood cells? (Outcome 14-3)

THE INFLUENCE OF THE RED BLOOD CELLS ON THE ERYTHROCYTE SEDIMENTATION RATE

When the ratio of red blood cells to plasma is lower than normal, the potential for rouleaux formation increases and will increase the speed that the red blood cells settle. For example, someone with anemia may have an elevated erythrocyte sedimentation rate, because his sample has fewer red blood cells as compared to the total sample volume than would be present normally. The presence of macrocytes (large cells) in the specimen will also cause the ESR to be greater than normal.

Erythrocyte sedimentation rates below the normal range may be present when the sample has a lot of microcytes (small red blood cells) or spherocytes (small, dense red blood cells) present. Sickle cells will also cause a decrease in the ESR, as the cells do not stack together as they would normally, so they will not settle out as expected. Patients with polycythemia will demonstrate a decreased erythrocyte sedimentation rate because of the excess number of red blood cells. Excessive leukocytosis (as in the case of chronic myelogenous leukemia) will also cause a decreased ESR result.

When the ESR values are out of the normal range because of the sizes or shapes of the red blood cells, the results may not change rapidly with treatment. Red blood cells live for approximately 120 days in the circulation, and the ESR result will not be significantly changed until the red blood cells present in the body have been replaced by healthy cells. It also takes time for the protein levels in the blood to change significantly with treatment. This means that the erythrocyte sedimentation rate is used more often for initial screening for inflammation or infection than it is to monitor the immediate effects of treatment.

REFERENCE RANGES

The normal reference ranges for the ESR vary by gender and by age. This is partly caused by the difference in the normal number of circulating red blood cells for men and women. The reference ranges are also elevated in adults over the age of 50. The normal reference ranges are listed in Table 14-1.

Test Your Knowledge 14-5

List normal ESR values for men and women younger than the age of 50. (Outcome 14-4)

CLINICAL SIGNIFICANCE OF ERYTHROCYTE SEDIMENTATION RATE TESTING

Inflammatory processes and acute infection with tissue breakdown increase the amount of fibrinogen and globulins present in the bloodstream. These proteins will interact with the red blood cells and cause the erythrocyte sedimentation rate to be elevated. The ESR can be used to diagnose a disease or condition, monitor an ongoing disease process or treatment, or assist with a differential diagnosis in certain disorders. For instance, the ESR will be elevated when a patient has recently experienced a myocardial infarction, but will not be elevated with angina pectoris. The ESR is not as sensitive to

TABLE 14-1

Reference ranges for the Westergren erythrocyte sedimentation rate

	Male	Female
Younger than 50 years of age	Less than 15 mm/hr	Less than 20 mm/hr
Older than 50 years of age	Less than 20 mm/hr	Less than 30 mm/hr
Older than 85 years of age	Less than 30 mm/hr	Less then 40 mm/hr

changes in treatment as some other laboratory tests may be, but it still may be used to monitor treatment over an extended period of time.

It is important to remember that nonpathological issues may also cause the ESR to appear elevated. Pregnancy is a normal physiological process that will exhibit an elevated ESR. By the 12th week of pregnancy, the ESR result will have increased above the normal range, and will remain so through the first month postpartum. Women who are menstruating may also have an elevated erythrocyte sedimentation rate.

The disorders of polymyalgia rheumatica (an inflammatory condition of the muscles) and temporal arteritis (inflammation and damage to the blood vessels that supply the head) are the only two disease conditions for which an elevated ESR is the primary means of diagnosis. The test may also be used (in addition to other clinical information and laboratory tests) to evaluate septic arthritis, pelvic inflammatory disease, and appendicitis. Rheumatoid arthritis, chronic infection, collagen disease (systemic lupus erythematosus [SLE]), and certain neoplasms (especially prostate or kidney cancer) are characterized by increased ESR results. Erythrocyte sedimentation rates may be used to the monitor the efficacy of therapy in osteomyelitis. The ESR results may be misleading in patients with osteoarthritis, as these patients may have ESR results within normal range even with the presence of inflammation in the joints. The ESR results may also offer valuable information to help determine the risk associated with coronary artery disease.

Test Your Knowledge 14-6

If an ESR is elevated, what process is occurring in the body? (Outcome 14-5)

POINT OF INTEREST 14-1

Seditainer procedure

Becton, Dickinson has created the Seditainer system, a unique setup for the erythrocyte sedimentation procedure. It is designed to measure the erythrocyte sedimentation rate without transferring the blood to a separate tube. This eliminates the potential for exposure to blood products while transferring the specimen, as well as the possibility of adding bubbles to the specimen during the ESR setup. In the Seditainer system, the evacuated tubes used for drawing blood during the venipuncture have a black top, and they contain 1.25 mL of a sodium citrate solution. The tube needs to be filled to the point at which the vacuum is exhausted, and the specimen needs to be well mixed. The tube is placed in a special Seditainer stand, and the knob on the stand is used to adjust the top of the blood column to the "0" level. Each spot in the Seditainer stand is marked with a scale used to read the ESR level at the end of 1 hour. The tubes are then discarded without ever being opened.

ERYTHROCYTE SEDIMENTATION RATE DETERMINATION

There are two methods used for performing the ESR. The oldest method is the **Wintrobe** method. To perform the Wintrobe ESR, well-mixed blood from a tube containing EDTA anticoagulant is transferred directly to a Wintrobe tube using a pipette with a long, slim end. The Wintrobe tube is short, with graduated markings up to 100 mm. The ESR result is read after the sample has been allowed to sit undisturbed for 1 hour. It is important that the ESR is not vibrated or disturbed during the time that the test is performed.

The Wintrobe method is not used as widely as is the **Westergren** method. The Westergren method involves the dilution of the EDTA-anticoagulated blood with a sodium citrate solution or saline before the blood is added to the pipette. The Westergren pipette is also smaller in diameter and taller than the Wintrobe tube.

Most offices use a CLIA-waived modified Westergren method that uses a "closed system," allowing the blood to be added to the pipette with a minimal amount of exposure to the specimen for the employee. The test is performed on an EDTA-anticoagulated venous blood sample. A specified amount of room-temperature, well-mixed blood is diluted with a premeasured citrate solution and introduced into a hollow column. The column is marked off into millimeters for ease of reading the results. The top of the sample must be at the "0" mark when the timing is started as a point of reference. The column is placed in a rack so that it is exactly level, and timing begins immediately. The column must not be disturbed during the timing process, and the specimens must be kept out of direct sunlight. At exactly 1 hour, the top of the red blood cell column is noted. The ESR result is documented as the millimeters per hour that the cells have settled. The result is compared to the reference ranges for the patient, based on gender and age.

During the 60 minutes of the testing process, the settling of the red blood cells occurs in three stages. The first stage takes place within the first 10 minutes of the testing period. Rouleaux formation (the stacking of red blood cells) is beginning to occur so there is little sedimentation, even in abnormal samples. The second stage occurs in the next 40 minutes, during which there is a constant rate of settling as the rouleaux are set up (if present). The "stacks" of cells fall rapidly during this time because of the increased volume to surface area. The final stage occurs in the last 10 minutes of the testing process. At this time the settling is slowing as the cells are forming the packed cell volume at the bottom of the column.

CLIA-Waived Erythrocyte Sedimentation Rate Methods

There are various modified Westergren ESR systems available for use in the laboratory. Sediplast (manufactured by Polymedco) is one of the most common kits. An EDTA-anticoagulated venous blood sample of 0.8 mL is pipetted into a vial containing 0.2 mL of 3.8% sodium citrate diluting solution. The vial is then gently inverted to mix and placed into the accompanying rack. A 100-mm marked tube is gently pushed into the vial until it sits on the bottom of the small vial containing the blood. The tube is designed for autozeroing, as the blood mixture will reach the "0" mark and any extra will enter into an overflow area. The timing must begin immediately and continue for exactly 60 minutes. The sedimentation rate in mm/hr is read at the top of the red blood cell column. This closed system minimizes potential exposure to the sample while setting up the procedure and also prevents the potential for splashing when disposing of the setup after testing. This process is expanded and detailed in Procedure 14-1.

POINT OF INTEREST 14-2

The C-reactive protein

The erythrocyte sedimentation rate is elevated when inflammation occurs within the body. However, the ESR does not react quickly when inflammation first starts, and the result does not change immediately when treatment begins. The C-reactive protein (CRP) is also elevated with systemic inflammation processes in the body. The CRP is also known as an acute phase protein. The level of the CRP is elevated with infection, cardiovascular disease, and the formation of atherosclerosis. Because the level of CRP is affected immediately with treatment that reduces inflamation, this test may be a better choice than the ESR with which to monitor a course of treatment.

Although the Wintrobe method is not used as often as is the modified Westergren method, the Wintrobe method is still a CLIA-waived procedure carried out some laboratories. The Wintrobe method requires a special pipette to add the specimen to the tube used for the test, and the method is not considered as safe for the employee because it provides an opportunity for more potential exposure to bloodborne pathogens as the blood is added to the tube.

Test Your Knowledge 14-7

List two CLIA-waived ESR procedures. (Outcome 14-6)

Automated Erythrocyte Sedimentation Rate Methods

For laboratories that perform high volumes of testing, there are automated ESR analyzers available. These instruments automatically perform the dilutions (if necessary), set up the ESR tubes, and have internal readers that detect the color change present at the top of the red blood cell column when the designated time has passed. Some determine the ESR in less than the 60 minutes required for the manual methods. Polymedco makes the Sedimat and the Sedten analyzers. The Ves-matic, manufactured by Diesse, uses the EDTA lavender top tube directly to perform the ESR measurement. Many of these analyzers are capable of processing multiple ESR tests at one time.

POTENTIAL SOURCES OF ERROR FOR THE ERYTHROCYTE SEDIMENTATION RATE PROCEDURE

The ESR test must be performed very carefully; procedural errors will directly affect the results. The first consideration is the collection of the blood sample to be used. If the blood is allowed to clot, the ESR will be falsely decreased because the fibrinogen in the sample will be used forming the blood clot. Samples will ideally be used for the ESR within a few hours of the blood draw. Most laboratories state that the sample may be used for the test if it is kept at room temperature (20° to 25°C) for up to 24 hours. Refrigeration may cause irreversible changes in the way the cells interact in the sample, so the process must occur within 12 hours if the sample is refrigerated. The blood must be at room temperature at the time the test is performed.

Procedure 14-1: Perform CLIA–Waived Sediplast Westergren Erythrocyte Sedimentation Rate

Erythrocyte sedimentation rates are one of the oldest, easiest, and least expensive testing procedures performed in laboratories. The Westergren method is the most common CLIA-waived manual system employed.

TASK

Perform an erythrocyte sedimentation rate using the Sediplast system, and read and record the results correctly.

CONDITIONS

- Hand-washing supplies and/or alcohol-based hand sanitizer
- Disposable gloves
- Sediplast Modified Westergren ESR reservoir tube, pipette, and rack
- Lavender top (EDTA) tube filled with blood
- Biohazardous disposal bag
- Face shield and laboratory coat
- Transfer pipette
- Test tube rack

CAAHEP/ABHES STANDARDS

CAAHEP 2008 Standards

I.P.I.12 Anatomy and Physiology; Perform Hematology Testing

ABHES 2010 Standards

- Medical Laboratory Procedures, b. Perform selected CLIA–waived tests that assist with diagnosis and treatment;
- #2: Perform Hematology Testing

Procedure	Rationale
1. Sanitize hands (allow them to dry), then apply gloves, laboratory coat, and face shield.	Gloves must be worn for any procedures during which exposure to blood or other potentially infectious materials is anticipated. This procedure has an increased risk for exposure to blood products, so a laboratory coat and face shield should also be worn.
2. Gather testing supplies. Check the expiration of the ESR kit components. Verify that the necessary quality control has been performed prior to testing the patient sample.	Expired products must never be used for testing. If the required quality control has not been performed, this needs to be accomplished (and the results verified) before the patient test results are reported.
3. Verify the test ordered and that the sample is appropriately identified.	The test and patient identity should always be verified before any testing procedure is started.
4. Verify that the EDTA tube is at room temperature.	The blood must always be at room temperature for the ESR test result to be valid.
5. Invert the EDTA tube 8 to 10 times to mix the sample.	The specimen must be well mixed or it will provide erroneous test results.
6. Carefully remove the rubber stopper on the top of the EDTA blood tube. Avoid aerosol formation by removing the top slowly. Place the stopper top down on the tabletop, and place the blood tube in a test tube rack.	If the stopper is removed too quickly, a fine “mist” of blood may be released, increasing the potential for exposure.

Procedure	Rationale
7. Remove the pink cap from the prefilled reservoir vial provided with the test kit. Be careful not to spill any of the solution.	The proper amount of diluent in the tube is vital for the dilution factor to be correct and the results to be valid.
8. Using the transfer pipette, add blood to the reservoir vial until it reaches the fill line. Avoid bubbles or air spaces in the reservoir vial. Recap the EDTA blood tube.	It is important to add the necessary volume of blood to the diluent to allow for the correct dilution ratio and to provide enough specimen for an accurate ESR reading.
9. Replace the pink cap on the reservoir vial and invert fully at least 8 times.	The specimen must be mixed thoroughly with the diluent for the results to be accurate.
10. Place the vial in the rack on a level surface. Insert the marked ESR pipette through the pink cap and push down until it comes to the bottom of the vial. The blood will rise to the "0" mark within the ESR pipette, and any excess blood will enter the overflow reservoir.	The blood must reach the "0" mark or the test will not be valid. This process should be accomplished fairly slowly to make certain that the blood reaches the top of the ESR pipette.
11. Make certain that the rack is in a place where it will not be disturbed and where it is not in direct sunlight. Set a timer for 1 hour.	Vibrations, movement or direct sunlight may affect the ESR results. A timer is necessary for appropriate timing of the procedure. 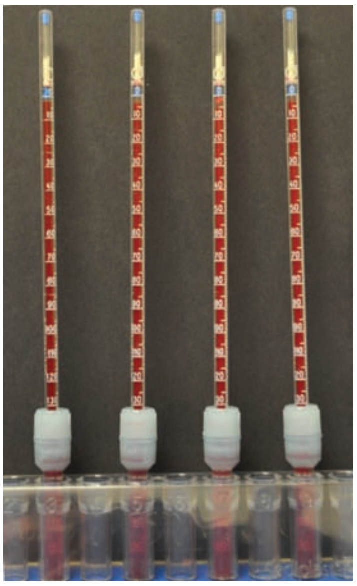 Courtesy of Fisher HealthCare, Inc.

Continued

Procedure 14-1: Perform CLIA-Waived Sediplast Westergren Erythrocyte Sedimentation Rate—cont'd

Procedure	Rationale
12. After 1 hour, note the number at the top of the red blood cell column. This will be in mm marked on the side of the ESR tube. Note this number as the result in mm/hr, and document the result on the log sheet or the patient's chart.	The marks should be relatively easy to read, and are read from the top of the pipette down. 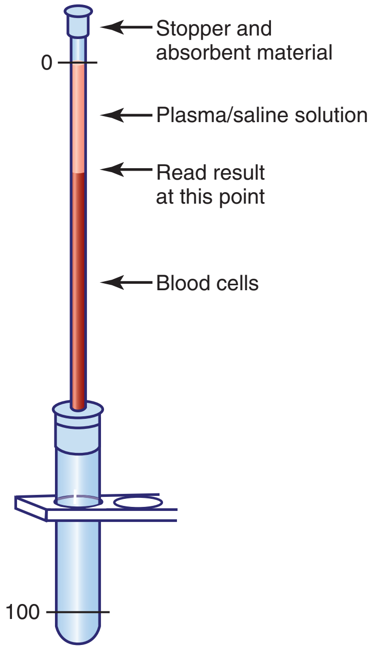 Stopper and absorbent material 0 Plasma/saline solution Read result at this point Blood cells 100 mm 0 10 mm 0 10
13. Dispose of the ESR setup and the transfer pipette in the biohazardous garbage. Disinfect the work area.	There is blood in the transfer pipette and the ESR setup, so both must be disposed of in the biohazardous garbage.
14. Remove gloves and sanitize hands. Remove face shield.	Hands must always be sanitized after removing gloves. The face shield is not necessary for all procedures, so should be removed until needed again.

Procedure	Rationale
15. Document the ESR results in the patient's chart if not completed previously. If necessary, also log the lot number and expiration of the kit components on the log sheet.	All procedures must be appropriately documented in the patient's chart.

Date	
1/18/2020:	ESR 22 mm/hr.
1100 a.m.	Connie Lieseke, CMA (AAMA)

The tube of venous blood must be well mixed so that the aliquot removed for the test is representative of the sample. Before removing the blood for the test, the tube is gently inverted at least eight times. After the blood is added to the diluent, adequate mixing is again imperative. When reading the results, if there is not clear definition at the top of the red blood cells in the plasma, the result is to be read at the point when there is a definitive border to the red blood cells in the column.

There can be no bubbles in the column of blood or diluted blood. The tube (column) must be exactly vertical during the testing process, so the holding rack into which it is placed must be level as well. Many of the racks come with leveling devices at the corners so that this can be verified. Movement (such as vibrations) can alter the results, so it is important that this test is not performed on a counter next to a centrifuge or a printer. The timing must be exact for accurate readings, so a timer is usually employed rather than a watch or a clock.

Test Your Knowledge 14-8

List two testing procedure errors that can affect the value of the ESR. (Outcome 14-7)

SUMMARY

The erythrocyte sedimentation rate is a test that screens for the presence of inflammation. The ESR results can be altered as a result of infection, autoimmune disease, or growth of neoplasms in the body. With inflammation, there is a change in the levels of plasma proteins with an increase in fibrinogen and some globulins. This protein change can cause the red blood cells to form excessive rouleaux, which settle out of solution faster than do single red blood cells. The observance of this sedimentation of red blood cells in 60 minutes constitutes the test procedure. The CLIA-waived procedures include manual Westergren and Wintrobe methods. There are also automated testing systems available for laboratories that perform a high volume of testing or desire the results in less than 1 hour. The erythrocyte sedimentation rate is used as a tool to aid in diagnosis, monitor disease process, follow treatment, and assess prognosis of certain disorders.

TIME TO REVIEW

1. Rouleaux formation is: Outcome 14-1
   a. An abnormal collection of proteins in the circulating blood
   b. The test for inflammation
   c. A group of red blood cells stacking together like coins
   d. None of the above

2. Examples of proteins that are increased in the plasma when the sedimentation rate is elevated include: Outcome 14-2
   a. Albumin and fibrinogen
   b. Fibrinogen and globulins
   c. Globulins and albumin
   d. Globulin and tyrosine

3. A 30-year-old male patient has an ESR of 38 mm/hr. The result of this test for this patient is _______ _______. Outcome 14-4
   a. Within the reference range for the patient's gender and age
   b. Elevated for the patient's gender and age
   c. Within range for his age, but not his gender
   d. Below the reference range

4. An 87-year-old woman has a sedimentation rate of 38 mm/hr. The result for this patient is ______ because of this patient's gender and age. Outcome 14-4
   - a. Normal
   - b. Elevated
   - c. Elevated
   - d. Decreased

5. An elevated ESR indicates the presence of: Outcome 14-5
   - a. Inflammation
   - b. Erythropoiesis
   - c. Albumin
   - d. Renal disease

6. A/an ___________ anticoagulated venous blood sample is used for the erythrocyte sedimentation rate. Outcome 14-6
   - a. Citrate
   - b. Heparin
   - c. EDTA
   - d. Fluoride

7. True or False: Westergren and Wintrobe are CLIA-waived erythrocyte sedimentation procedures. Outcome 14-6

8. ______ is measured to determine the ESR value. Outcome 14-7
   - a. The ratio of proteins in the blood
   - b. The rouleax formation of red blood cells
   - c. The settling of the red blood cells
   - d. The presence of red blood cells in the bone marrow

9. Examples of potential errors when performing an ESR include: Outcome 14-8
   - a. Insufficient sample added to diluent
   - b. Insufficient mixing of sample and diluents
   - c. Reading result at 50 minutes
   - d. All of the above

Case Study 14-1: Cutting corners

Margaret Lindon is a new employee, and she was nervous on her first day alone in the laboratory. One of the first orders she received was to collect a specimen and perform an ESR test. She had been trained on this procedure the week before, and remembered that it was not very difficult. The requisition indicated that a lavender top tube was to be drawn, and Margaret knew where the ESR supplies were kept. As Margaret was collecting the blood sample, the blood flow stopped, and she could not get it to continue. She discontinued the blood draw with a partial lavender top tube filled.

After the patient left, Margaret began to set up the ESR test. She added the blood to the reservoir tube containing the citrate solution, but realized that she did not have quite enough specimen to fill it to the required line. She inserted the ESR pipette, and the blood-citrate solution came almost to the top of the pipette, but was shy by a few millimeters. After 1 hour, Margaret read the results of the ESR, and reported out that it was 28 mm/hr.

1. Was the correct specimen drawn for the ESR order?
2. What were the potential sources of error for this scenario? Was this result of 28 mm/hr a valid result?

RESOURCES AND SUGGESTED READINGS

"Polymedco Products"
Information about the various clinical diagnostic test kits and educational materials provided by Polymedco. http://www.polymedco.com/products-pages-7.php
"Diesse USA Product Line"
Products from Diesse USA including an ESR system. http://www.diesse-usa.com/products.html

Chapter 15

Coagulation Studies

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Purpose of Coagulation Studies
- Mechanisms of Blood Clotting
- Disorders Diagnosed or Monitored With Laboratory Coagulation Tests
  - Thrombosis
  - Atrial Fibrillation
  - Postsurgical Prophylaxis
  - Disseminated Intravascular Coagulation
  - Thrombocytopenia
  - Antiphospholipid Syndrome
  - Hemophilia
- Laboratory Tests Used to Diagnose Coagulation Disorders or Monitor Anticoagulant Therapy
  - Prothrombin and International Normalized Ratio
  - Activated Partial Thromboplastin Time/Partial Thromboplastin Time
  - Platelet Count
  - Fibrinogen
  - Fibrin Degradation Products
  - D-dimer
  - Bleeding Time Test
  - Other Tests Used to Assess Platelet Function
- Specimen Requirements for Coagulation Testing
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

15-1 Define the key terms.

15-2 Explain the purpose of coagulation studies.

15-3 Describe the process of coagulation.

15-4 Compare the formation of a thrombus to that of an embolus.

15-5 Compare and contrast the disorders diagnosed or monitored with laboratory coagulation testing.

15-6 Examine how the prothrombin test and INR calculation are used together to monitor coagulation status for a patient.

15-7 Successfully perform a CLIA-waived PT/INR test.

15-8 Explain appropriate use of a PTT test for monitoring coagulation.

15-9 Evaluate the use of platelet counts in coagulation testing.

15-10 Explain why bleeding time tests are not performed as frequently as PT and PTT tests.

15-11 Describe other tests of platelet function.

15-12 List other coagulation tests that may be ordered.

15-13 List the requirements for coagulation specimen collection and potential sources of error for the various tests.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

None

ABHES Standards

None

KEY TERMS

Afibrogenemia
Aggregate
Antiphospholipid syndrome
Atrial fibrillation
Clotting cascade
Clotting factor
Coagulation
Coumadin
D-dimer
Deep vein thrombosis (DVT)
Disseminated intravascular coagulation (DIC)
Dysfibrinogenemia
Embolus
Endothelial cell lining
Extrinsic pathway
Fibrin degradation products (FDPs)
Fibrinogen
Fibrinolysis
Fibrinolytic drugs
Hemophilia
Hemostasis
Heparin
Hypofibrinogenemia
International normalized ratio (INR)
Intrinsic pathway
Occlude
Partial thromboplastin time (PTT)
Pathological thrombosis
Plasmin
Platelet count
Prophylactic
Prothrombin time (PT)
Stroke
Thrombin
Thrombocytes
Thrombocytopenia
Thrombosis
Viscous
Warfarin
von Willebrand's factor

When most of us receive a paper cut or other minor injury, we take for granted that this minor blood vessel injury will stop bleeding and heal within a short period of time. The process of blood clotting, known as **coagulation,** and the breakdown of these clots within the body are monitored by use of various coagulation testing procedures. Platelets, fibrinogen, thrombin, and 13 other clotting factors are all involved in the process of clot formation, which is necessary for repair of blood vessel damage. The blood vessels also play a role. In addition, chemicals are produced by the body to break down the clots when they are no longer needed. This process is known as **fibrinolysis.** Coagulation studies include quantitative and qualitative analysis of the various components, as well as their function.

PURPOSE OF COAGULATION STUDIES

Coagulation studies are performed in the laboratory when the various elements in the clotting process appear to be imbalanced or malfunctioning. These defects can cause excessive bleeding, bruising, or unwanted blood vessel occlusion (blocking of the vessel) with formation of blood clots in the bloodstream. Anticoagulant medication use is also closely monitored with laboratory coagulation testing. Laboratory tests may include quantitative analysis of certain plasma proteins involved in the clotting process, as well as analysis of platelet count and function. Coagulation studies are often performed prior to surgical procedures to verify that the balance between bleeding and clotting in the body is at a desirable level. **Hemostasis** is the word used to describe the processes needed to balance clot formation and fibrinolysis in the body.

Test Your Knowledge 15-1

What is one reason that coagulation studies may be performed? (Outcome 15-2)

MECHANISMS OF BLOOD CLOTTING

The balance between blood clotting and excessive bleeding involves numerous factors. These include the blood vessels of the body (including the cells that line those vessels); the platelets **(thrombocytes)** circulating in the bloodstream; and numerous coagulation factors present in the circulation as enzymes, substrates, inhibitors, and plasma proteins. The study of coagulation and hemostasis focuses on two major systems. The primary coagulation system includes the action of the platelets and the vascular response to vessel injury. The actions of the plasma proteins and other chemicals in the circulation make up the secondary coagulation system, and complete the coagulation process.

Test Your Knowledge 15-2

Are platelets solely responsible for coagulation when a vessel is damaged? (Outcome 15-3)

When damage occurs to a blood vessel, the **endothelial cell lining** of that artery, vein, or capillary is disrupted. The first response of the body is to constrict the damaged blood vessel and other vessels in the vicinity, limiting the blood flow to the area so that platelet aggregation can begin. In small arterioles, venules, and capillaries, this may stop the bleeding without moving to the next step. In larger vessels, however, this will not be enough.

When intact, the blood vessel lining does not attract platelets; in fact, the two repel each other. However, when the endothelial cell lining is damaged, the exposed collagen layer underneath the endothelial cells attracts the platelets to **aggregate** and begin the process of coagulation. Fibrinogen and *von Willebrand's factor* assist in the adhesion of the platelets to one another. The following are some of the next steps involved in primary coagulation:

1. The platelets in the area of the damaged vessel change shape and begin to adhere tightly to one another.
2. Platelets are filled with granules in their cytoplasm. As they aggregate and change shape, they release the content of these granules into the plasma. This chemical causes the platelet plug to become more stable and the connections between the platelets to be very tight.
3. The platelets release a compound known as *proaccelerin,* or *labile factor,* that works with vitamin K to start the rest of the changes necessary for a permanent fibrin clot to form.

Test Your Knowledge 15-3

What is the body's first response to vessel injury? (Outcome 15-3)

Secondary coagulation includes a series of events described as the **clotting cascade.** This cascade involves continuous feedback and interaction in which each **clotting factor** affects other clotting factors to achieve the formation of a solid fibrin clot over the area of the damaged vessel. All of the clotting factors are produced in the liver and introduced into the bloodstream. Some of the clotting factors are also produced by endothelial cells in the vessel lining and megakaryocytes in the bone marrow. There are at least 10 clotting factors, and they are designated by Roman numerals. They work in two different pathways, the **intrinsic pathway** and the **extrinsic pathway,** to accomplish the same goal. The intrinsic pathway involves chemicals that are all present in the bloodstream naturally, whereas the extrinsic pathway is initiated by release of chemicals into the bloodstream by the subendothelial cells exposed during the vessel injury. Table 15-1 lists the clotting factors with their Roman numeral numbers, alternative names, and the laboratory tests used to monitor their activity in the body.

Test Your Knowledge 15-4

What does the term *clotting cascade* refer to? (Outcome 15-3)

It is important to keep in mind that the balance of bleeding and clotting in the body must be regulated naturally by clotting processes, as well as by fibrinolysis (the breakdown of the formed clots when they are no longer needed) and inhibitors that keep the system from overreacting to damages to the vessel walls. Any natural or acquired interference in this balance will negatively affect hemostasis.

TABLE 15-1

Clotting factors with alternative names and common tests used to monitor levels or detect deficiencies*

Clotting Factor	Alternative Name	Laboratory Test
I	Fibrinogen	Fibrinogen level
II	Prothrombin	Prothrombin time (PT)
III	Thromboplastin	PTT
IV	Ionized calcium	Ionized calcium levels
V	Proaccelerin or labile factor	PT and PTT
VII	Proconvertin or stable factor	PT
VIII	Antihemophilic (also linked to von Willebrand's factor)	PTT
IX	Plasma thromboplastin component	PTT
X	Stuart-Prower	PT and PTT
XI	Plasma thromboplastin antecedent	PTT
XII	Hageman factor	PTT
XIII	Fibrin stabilizing factor	PT and PTT

*There is no factor VI.

DISORDERS DIAGNOSED OR MONITORED WITH LABORATORY COAGULATION TESTS

Coagulation testing is used to diagnose or monitor a variety of diseases or disorders. Some of the results provide information that will help the health-care provider to anticipate problems (such as those performed prior to surgical procedures), whereas other laboratory values are used to achieve a diagnosis or monitor the progress of a disease.

Thrombosis

Thrombosis is the clot formation that occurs naturally in the body. However, there may also be times when **pathological thrombosis** occurs. Clots may form in any vessel of the body related to different risk factors and exhibiting different symptoms dependent on the location of clot formation. Arterial thrombosis can be caused by hypertension, atherosclerosis, blood that is too **viscous,** or thick, and by functional platelet abnormalities. Clots in the arteries usually include a small amount of fibrin and more white cells than thromboses formed in the other vessels.

Pathological thrombosis in the venous system is usually made up of red blood cells and large amounts of fibrin. Deep vein thrombosis (DVT) may be caused by impaired (slow) blood flow, especially in those patients who are recovering from surgery and are therefore inactive. Other factors that may contribute to formation of venous thrombi include impaired breakdown of clots (fibrinolytic deficiency) and inadequate levels of clotting inhibitors that keep the body from excessive thrombosis formation. A very serious complication of deep vein thrombosis occurs when pieces of the blood clot dislodge from the vessel of formation and enter the bloodstream. This is called an **embolus,** and if it becomes lodged in the pulmonary tissues, it may be fatal.

Test Your Knowledge 15-5

How is an embolus different from a thrombus?

(Outcome 15-4)

Atrial Fibrillation

Atrial fibrillation is a condition in which there are erratic electrical signals sent through the atria, which are the upper chambers of the heart. Because of these erratic signals, the heart does not beat with the correct rhythm or at the correct speed. This causes turbulence, or a disrupted blood flow through the heart, and an increased risk of blood clot (thrombus) formation within the atria. This is especially common in the left atria. If a thrombus is present, the patient is at a very high risk of embolus formation as well, as the blood travels through the heart and may pick up pieces of the clot. The emboli often travel to the brain, where they **occlude** (block) the small vessels and cause a **stroke.** A stroke (also known as a cerebrovascular accident) is the result of a blockage of blood flow to an area of the brain. Patients who have chronic atrial fibrillation are often prescribed **warfarin** as a lifelong treatment option, as warfarin is a drug that interferes with the extrinsic clotting factor system of the body, and decreases the risk of blood clots forming in the atria.

Postsurgical Prophylaxis

Because the opportunity for thrombosis formation is heightened when the blood flow to a specific area of the body is slowed, those who are immobile or recovering from surgery are at high risk of blood clot development in their legs, especially. This is called a **deep vein thrombosis,** or **DVT.** Increased passive exercise (when possible) helps to reduce this risk, and is often started the same day or one day postsurgery with the assistance of a member of the physical therapy team. However, this may not always be enough to keep clots from forming in the deep veins of the legs. Many patients who anticipate extended hospitalization after surgery will be placed on heparin during or immediately after the procedure as a **prophylactic** measure, treatment in anticipation of a potential problem, as opposed to treatment after the blood clot has occurred. Heparin is fast acting, is given via IV or injection, and affects the intrinsic clotting system of the body. Heparin is a very good choice for inpatient treatment, but as patients near discharge, they are usually started on oral warfarin treatment, as this is easier to maintain in the home environment.

Disseminated Intravascular Coagulation

Disseminated intravascular coagulation (DIC) occurs when the blood clotting system is activated throughout the body, rather than just at a site with injury to a blood vessel. Small thrombi form everywhere, including within organs and capillary beds. That patient's coagulation factors are used up after a time, as well as the substances that break down blood clots. This often results in uncontrolled bleeding. DIC can be a process that is slow to develop, and in this situation the individual usually has systemic thrombi develop, but no uncontrolled bleeding. In acute (or rapid-acting) DIC, the patient develops the multiple thrombi, but they also may hemorrhage because of the rapid use of the coagulation factors within their body.

There is *always* an underlying cause for patients who exhibit disseminated intravascular coagulation. Many times (especially in acute cases) if the underlying situation is addressed and resolved, the DIC will also resolve. Potential causes of disseminated intravascular coagulation include the following:

- Systemic infection with bacteria or fungi
- Severe trauma
- Solid tumors or hematological cancers
- Obstetrical complications such as placental abruption or amniotic fluid emboli
- Aortic aneurysms or other vascular disorders
- Systemic inflammatory responses, such as those that occur with transfusion reactions

The first course of treatment for DIC is to resolve the underlying problem, if possible. However, in situations in which the initial disease process cannot easily be resolved, or cannot be resolved quickly enough, the symptoms of disseminated intravascular coagulation may be treated with transfusions of platelets, coagulation factors contained in fresh frozen plasma, and fibrinogen as provided in cryofibrinogen precipitate.

Test Your Knowledge 15-6

How are DIC and DVT similar? (Outcome 15-5)

Thrombocytopenia

In **thrombocytopenia,** the patient does not have enough platelets circulating to coagulate effectively when necessary. Certain medications or other underlying hematology disease states may cause this condition. Because of the low numbers of platelets, the clots that form in response to vascular damage will not be adequate. In addition to thrombocytopenia, patients may exhibit disorders of platelet adhesion, where the platelet count is normal for the patient, but their platelets lack the ability to stick to other platelets and form the aggregate plug that is necessary to stop the bleeding with vascular injuries.

Test Your Knowledge 15-7

How is thrombocytopenia different from the other disorders presented thus far in this chapter? (Outcome 15-5)

Antiphospholipid Syndrome

Antiphospholipid syndrome is a condition in which small systemic blood clots form, presenting potential risk for embolism or organ failure. Because the blood clots can form anywhere, the patient is at a high risk for stroke, or total occlusion of blood vessels providing circulation to internal organs.

Hemophilia

Hemophilia is a sex-linked inherited disorder that affects approximately 1 in 5,000 males. There are two types of hemophilia: type A and type B. Those affected with type A hemophilia lack the clotting factor VIII. Those with type B lack the clotting factor IX. Both groups of patients may bleed excessively and experience vascular damage, in addition to internal bleeding in the elbows, knees, and other joints. Not all patients with hemophilia are affected to the same extent; the disorder may express itself with mild to moderate symptoms. It is also possible to develop hemophilia in response to chemotherapy treatment or other disease states. In those situations, the patients actually develop antibodies to the clotting factors, which render them useless when needed within the body. Patients with all types of hemophilia will need to be treated with the missing clotting factors to avoid excessive bleeding. The clotting factors may be derived from human donated plasma, or they may be created in a laboratory without human elements.

LABORATORY TESTS USED TO DIAGNOSE COAGULATION DISORDERS OR MONITOR ANTICOAGULANT THERAPY

As we have learned, certain disease states may cause a patient to be more prone to thrombosis formation, whereas other pathologies may cause excessive bleeding. For those who have disorders causing excessive bleeding, the coagulation tests associated with diagnosis and monitoring of the conditions are usually performed at a hospital or reference laboratory. These tests may include the prothrombin time (PT), the activated partial prothrombin time (PTT), fibrinogen levels, fibrin degradation products (FDPs), and D-dimer. Health-care providers may also request levels of specific clotting factors. Table 15-2 contains information about the tests covered in this chapter, why they are commonly ordered, and the specimen types required. Medical assistants may perform PT/INR tests, and will be involved in the specimen collection process for the tests performed outside the physician office, so they will benefit from the basic knowledge about the specimen requirements and clinical significance of the various test results.

TABLE 15-2

Common coagulation tests, clinical indication or significance, and specimen requirements

Test Name	Abbreviation	Test Use	Specimen Type
Activated partial prothrombin time	PTT or PTT	Screening for coagulation dysfunction; intrinsic system of blood coagulation. Also used to monitor heparin anticoagulant use.	Light blue top tube with citrate anticoagulant; must be filled completely.
Fibrin degradation products	FDP	Monitoring of breakdown products from blood clot dissolution	Special FDP tube for plasma levels; may also use red top tubes for serum levels.
D-dimers	D-dimers	Fibrin breakdown products from blood clot dissolution	Citrated plasma from full light blue top tube. Plasma should be frozen if the test cannot be performed immediately.
Fibrinogen level	Fib	Measures amount of circulating fibrinogen in bloodstream	Plasma from full light blue top citrate tube.
Bleeding time	BT	Platelet function	Test performed on patient
Platelet count	Plt Ct	Assessment of number of platelets available for blood coagulation	Lavender top tube with EDTA anticoagulant. Test performed on whole blood.
Platelet function assay 100	PFA 100	Assessment of platelet function; clotting ability of platelets	Full light blue top citrated tube; whole blood is used for the assay.

TABLE 15-2

Common coagulation tests, clinical indication or significance, and specimen requirements—cont'd

Test Name	Abbreviation	Test Use	Specimen Type
Prothrombin time	Prothrombin time or PT	Screening for coagulation dysfunction; specifically extrinsic blood clotting cascade. Also used for monitoring oral anticoagulant use.	Full light blue top citrate tube or capillary sample if performed with point-of-care testing.
International normalized ratio	INR	Factor used for monitoring oral anticoagulant use; standardizes results using various reagents and testing methods	Full light blue top citrate tube or capillary sample if test is performed with point-of-care testing instrument

Prothrombin and International Normalized Ratio

The **prothrombin time (PT)** is a test that measures the efficacy of the extrinsic coagulation process. Prothrombin is a clotting factor that is manufactured by the liver, and the change of prothrombin to **thrombin** is one of the last steps to forming a clot when a vessel is damaged. The action of clotting factors I, II, V, VII, and IX may be monitored with the prothrombin time test result. The prothrombin time may be elevated in patients who have insufficient amounts of these clotting factors, as well as those who are deficient in vitamin K. Patients suffering from liver disease may also have an increased prothrombin time result because of the inability of the liver to create the clotting factors as needed.

The PT is reported in seconds, as it is a measurement of the amount of time needed for a clot to form in the specimen after the addition of thromboplastin and calcium to a sample of platelet-poor plasma. PT tests are often performed as part of a coagulation screen before surgical procedures or when a patient shows signs of excessive bleeding or bruising. Prothrombin times may be used to screen for initial coagulation disorders. The PT is also used to help monitor oral anticoagulant use, as the result is elevated when a patient is being treated with these oral anticoagulants.

POINT OF INTEREST 15-1

History of warfarin

As the population in the United States grew dramatically in the early 1900s, so did the demands for farming. The Midwest was farmed extensively, growing crops to feed the U.S. population as well as the food to feed dairy and beef cattle. Once the "natural" feed products of the area were exhausted (such as wheat and grasses native to the region), it was necessary to research other crops that might suffice to feed the animals. Sweet clover was introduced, as the protein content and digestibility was similar to that of alfalfa. However, sweet clover was more prone to spoilage than the other food sources had been because of issues related to its high moisture content.

Soon, there were reports of unexplained hemorrhage and sickness in cattle that had been fed from the sweet clover after harvest. Through analysis, it was found that the spoilage of the clover created dicoumarol. Dicoumarol attached itself to the vitamin K present in the intestines of the cattle. Vitamin K is essential for several blood clotting mechanisms, so these cattle would bleed excessively, as the necessary vitamin was no longer usable.

In 1948, warfarin was created as a derivative of dicoumarol, and in 1952 it was introduced to the world as a type of rat poison. A young navy recruit decided that he wanted to end his life, and ingested a large dose of this warfarin rat poison. Remarkably, he did not die, although his blood clotting processes were definitely affected. This situation prompted more research into how warfarin could be used to treat humans, and in 1954, it was first authorized for human use as an oral anticoagulant. In 1955, President Eisenhower was treated with warfarin after a heart attack.

Warfarin (brand name Coumadin) is the most common drug prescribed worldwide. It is taken by those who have a prior history of blood clots or heart disorders such as atrial fibrillation and heart valve deformities, as well as given to surgical patients. The efficacy of warfarin can be affected by ingestion of foods that are high in vitamin K, such as cabbage, cauliflower, spinach, soybeans, and broccoli.

A CLIA-waived PT test procedure is used to monitor the anticoagulant effects of medications such as **warfarin** (Coumadin) in the physician office laboratory. These drugs are used by patients who are at a high risk for blood clot formation. Close monitoring of the effects of oral anticoagulants is necessary, as patients who have too much in their bloodstream run the risk of excessive bleeding, whereas those with too little could develop thrombosis.

A calculation called the **international normalized ratio (INR)** is reported with the PT result. The INR is a factor that allows for standardization of prothrombin tests performed with different methodologies and/or reagents. Prothrombin tests performed with different lot numbers of reagents or different methodologies may differ by 10% to 20% from one laboratory to another. This can be problematic when adjusting oral anticoagulant dosages based on PT test results. In 1983 the World Health Organization proposed a calculation (the INR) to standardize the variables that are present when performing prothrombin testing with various instruments and lot numbers for reagents. For an individual who is not on oral anticoagulants, the INR would be approximately 1, and for those who are being treated with warfarin or Coumadin, the INR should be 2 to 3. Decisions to change medication levels are based on the INR result, not the PT result that may be reported with the INR. Reference ranges for the prothrombin time may vary between laboratories, but they are approximately 11 to 13 seconds.

Test Your Knowledge 15-8

True or False: Is the INR used when screening for potential coagulation defects? (Outcome 15-6)

The Coaguchek S, manufactured by Roche Diagnostics, is a CLIA-waived PT/INR testing method often used in the physician office laboratory. This type of analyzer is a point-of-care test (POCT) that allows for immediate results. Performance of this test in the office can benefit the patient by allowing the results to be available to the health-care provider within minutes. This allows the patient to receive any necessary instruction about medication changes warranted by the results while still in the office. The test requires only 10 microliters of blood (approximately one drop), and the sample may be obtained through a capillary puncture or a venipuncture.

In situations in which the PT and INR are ordered as part of a coagulation screen or when the testing is performed at a reference laboratory, a full, well-mixed, light blue top sodium citrate tube is required for the test.

Procedure 15-1: CLIA-Waived Prothrombin Time/INR Testing Using the CoaguChek S Instrument

TASK

Perform PT/INR testing utilizing the CoaguChek S instrument and document the results appropriately. Complete all steps within 10 minutes.

CONDITIONS

- Hand sanitization supplies
- Gloves
- Completed patient requisition form and patient chart
- CoaguChek S instrument
- Test strips
- Quality control material
- Test strip code chip
- Bulb and capillary provided with instrument for application of sample
- Gauze or cotton balls
- Alcohol pad for sanitization
- Lancet device or venipuncture supplies
- Biohazardous disposal container
- Biohazardous sharps container

Note: This test may be performed on a capillary blood sample or a sample obtained through venipuncture. Be certain to have the necessary supplies available for the method used for sample collection, as well as a biohazardous sharps disposal container and gloves.

CAAHEP/ABHES STANDARDS

CAAHEP 2008 Standards

None

ABHES 2010 Standards

None

Procedure	Rationale
1. Verify test ordered on requisition form. Greet and identify patient.	The requisition should always be examined to be certain which test is ordered, and the patient identity must always be verified before proceeding.
2. Sanitize hands, allow them to dry, and apply gloves.	Gloves protect the hands from exposure to potential bloodborne pathogens. Gloves should always be applied in the presence of the patient.
3. Gather necessary supplies. Test strips must be at room temperature for at least 5 minutes prior to use. Compare the number on the code chip (located in a small compartment at the top of the box of test strips) to the last three numbers after the lot symbol on the test strip pouch. If these numbers do not match, this code chip may not be used with these strips.	Supplies must be within reach before starting the process so that the timing of the testing is accurate. The code chip must be the correct number for the reagent strip for the test result to be accurate.
4. Verify whether quality control (QC) testing needs to be performed on the instrument. If so, perform QC testing and verify validity of results before proceeding with patient testing.	The frequency of the QC testing is based on the manufacturer's and individual laboratory recommendations.
5. Determine the method to be used for specimen collection. a. If a capillary sample is to be used, prepare the lancet. Roche manufactures a capillary tube and bulb to aid with specimen collection, as well as a lancet specifically designed to be used with the CoaguChek system. Assemble the capillary tube and bulb if desired. The blood may also be applied directly to the test strip without the aid of the capillary tube.	The capillary sample is the most common method of testing used in the physician office laboratory. CoaguChek capillary tubes
b. Blood obtained through venipuncture is acceptable, as long as it is collected into a preservative-free plastic syringe using a 21- to 23-gauge needle. The specimen would need to be applied to the strip immediately after the collection has been completed.	Preservatives and/or anticoagulants are not acceptable for this testing method. The blood cannot be allowed to clot before application on the test strip, or the results will be inaccurate.

Continued

Procedure 15-1: CLIA-Waived Prothrombin Time/INR Testing Using the CoaguChek S Instrument—cont’d

Procedure	Rationale
6. Insert the code chip into the machine, and place the instrument on a flat, vibration-free surface. Turn on the power, and make certain that the code displayed on the screen matches the numbers printed on the foil wrapper enclosing the test strip.	If the codes do not match, the instrument will not provide a test result.
7. The test strip icon will flash on the CoaguChek S display screen. Open the foil pouch and insert a test strip into the instrument by sliding it (printed side up) into the tray on the front of the instrument in the direction of the arrows. Continue to push the test strip into the instrument until it stops. Test strips must be used within 4 minutes of removal from the foil package.	For most of the testing process, the CoaguChek S display screen will guide your next steps. CoaguChek XS Plus HAUPTMENU PATIENTEN TEST QC TEST SPEICHER SETUP *Courtesy of Roche Diagnostics.*
8. Prepare the appropriate site for the capillary puncture, or obtain the blood sample via venipuncture.	Once the test strip is placed into the instrument, the time will be limited for the specimen to be collected and applied for the testing process.
9. Monitor the display screen; at this point, a test strip icon should be visible as well as a flashing clock icon. Do not apply the sample until the icon for a blood drop appears on the screen.	If the sample is applied too soon, the instrument will not process the sample and the screen will display an error.

Procedure	Rationale
10. Once the flashing blood icon appears on the screen, a countdown for 180 seconds begins. The instrument will beep every 30 seconds until an adequate sample has been applied to the test strip. The sample must be applied before the countdown reaches 0.	A countdown of 180 seconds should be adequate time to obtain the capillary sample, or to prepare to apply the venipuncture specimen.
11. Utilizing appropriate capillary or venipuncture procedure, obtain the blood sample. Apply a drop of blood to the test strip directly over the flashing yellow light. This flashing light indicates where the sample is to be applied.	For this procedure, the first drop of blood obtained from a capillary sample is to be used; do not wipe away the first drop of blood as you would customarily do for other types of collections. For samples obtained through venipuncture, the first four drops of blood are to be discarded from the end of the needle, after which the next drop is to be applied to the sample area. Do not allow bubbles to form when applying the sample.
12. After specimen application, the screen will display a clock icon. Do not touch the instrument or the test strip during this time. The result will be available on the screen (and stored in the memory of the instrument) within 1 minute.	The clock icon is displayed during the entire testing procedure.
13. Document the result on the instrument log sheet.	Once the test strip is removed from the machine, the result is no longer displayed on the screen, so it is best to record it immediately.
14. Discard the test strip by pulling it out of the instrument and discarding it into a biohazardous waste container.	The test strip is contaminated with blood and must be disposed of accordingly.
15. Sanitize the work area and put away supplies.	Be sure to refrigerate the test strips right after use.
16. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
17. Document the test results in patient chart.	All results must be documented in the patient's chart.

Date	
04/18/2013:	PT/INR test performed from capillary puncture. PT result 23.2 seconds, INR 2.1
	Connie Lieseke, CMA (AAMA)

Activated Partial Thromboplastin Time/ Partial Thromboplastin Time

Activated partial thromboplastin time is abbreviated PTT. The PTT is a measurement of the efficacy of the intrinsic coagulation pathway. This pathway includes the actions of clotting factors VIII, IX, XI, and XII. The PTT is measured in seconds, just as the PT is. Technically it is the amount of time that the blood sample takes to clot when phospholipid and calcium are added to the specimen. The PTT test may be ordered (with the PT) as part of a presurgical screen, when a patient has excessive bleeding or bruising, or to monitor the use of **heparin** therapy. Heparin is a fast-acting anticoagulant that affects the factors in the intrinsic blood-clotting pathway. Heparin is used to treat patients who are at high risk for thrombosis and to treat those who have already developed blood clots, especially in postsurgical settings. Heparin is not a medication commonly used in the home setting, because it must be administered IV or by subcutaneous injection multiple times per day. Reference ranges for the PTT are 25 to 31 seconds, and the specimen must be a full light blue top sodium citrate tube.

Test Your Knowledge 15-9

Which medication is monitored by using the PTT?

a. Coumadin
b. warfarin
c. heparin
d. aspirin (Outcome 15-8)

Platelet Count

One of the first steps in the coagulation process is the aggregation of platelets at the site of the vessel injury. For this reason, it is imperative that platelets are present in sufficient numbers and also that they function appropriately in order for the body to repair vessel damage as it should. Platelets are produced by the bone marrow, and when the blood vessels are injured, they adhere to the site of the injury and to other platelets. This primary "plug" helps to protect the body from excess blood loss, and also starts the clotting cascade, which ends with a more permanent fibrin mesh at the site to allow for the damage to heal.

Insufficient platelet amounts (thrombocytopenia) will put the patient at risk for excessive bleeding with vessel injury. A **platelet count** is performed to monitor the number of platelets present in the general circulation. Platelet counts are routinely performed as part of a complete blood count, and may also be ordered as part of a coagulation screen. The specimen required is a lavender top EDTA tube, and the reference range is 150,000 to 450,000 mm^3.

Fibrinogen

Fibrinogen is a "sticky" substance manufactured by the liver that is an essential part of the blood-clotting process. It is naturally dissolved in plasma, and is the precursor to fibrin. Fibrin is necessary for effective blood clot formation. Fibrinogen is also known as clotting factor I. The presence of adequate levels of fibrinogen is essential to platelet aggregation. Levels of fibrinogen may be decreased in the presence of genetic disorders, resulting in a diagnosis of **afibrinogenemia** or **hypofibrinogenemia. Dysfibrinogenemia** is a rare genetic disease that causes abnormal fibrinogen activity; the fibrinogen created in the liver is not able to convert to fibrin correctly to complete blood clot formation. Low levels of fibrinogen may contribute to excessive bleeding, and conversely, levels that are above the normal range may contribute to formation of small intravascular blood clots. Fibrinogen levels may be elevated in any condition that causes inflammation including pregnancy, acute infection, cancer, coronary heart disease, myocardial infarction (MI), and stroke. These increased levels of circulating fibrinogen will make the patient more prone to pathological blood clot formation, so the fibrinogen level test is often performed with the PT, PTT, and platelet count assays to screen for coagulation dysfunction. Reference ranges for fibrinogen are approximately 20 to 400 mg/dL, but may vary with the method used for testing by any individual laboratory. Fibrinogen requires a full light blue top sodium citrate tube.

Test Your Knowledge 15-10

If fibrinogen levels are decreased, what might a potential outcome be for the patient? (Outcome 15-11)

Fibrin Degradation Products

Clot creation and subsequent breakdown must remain in balance for the body to maintain a healthy coagulation response. Once a damaged vessel has healed, **plasmin** (an enzyme designed to dissolve blood clots) and other chemicals in the plasma immediately begin to dissolve the fibrin mesh forming the clot so that it does not become too large. The by-products of fibrinolysis may be measured in the laboratory with an assay that measures **fibrin degradation products (FDPs).** FDPs are essentially fragments that are released as blood clots are dissolved. These fragments are released when fibrinogen (the precursor of fibrin) or fibrin is broken down by plasmin. The amount of fibrin degradation products increase in disseminated intravascular coagulation, pulmonary embolism, and with some obstetrical complications. In most laboratories, the FDP test is performed on plasma collected in a light blue top tube with sodium citrate anticoagulant. If the test is to be performed within 4 hours, the tube is to remain at room temperature. If the sample must be transported elsewhere for testing and the processing takes more than 4 hours, the tube should be spun down and the plasma must be frozen. Serum FDP assays may also be performed, which may require specific tubes that contain thrombin to clot the blood and chemicals to prevent fibrinolysis within the tube. Normal reference ranges for FDP are based on the testing procedure, and may vary.

D-Dimer

The FDP test measures all the breakdown products created as blood clots dissolve. One of these by-products is very specific for certain cross-linked components of the fibrin clot, and that can be measured by performing a **D-dimer** test. D-dimers are a specific type of breakdown product that indicates the dissolution of fibrin, rather than just fibrinogen. The D-dimer test is currently replacing the FDP analysis in many laboratories. D-dimers are always present in small amounts in the blood, but elevated levels may be indicative of disseminated intravascular coagulation, pulmonary embolism, or deep vein thrombosis. The body increases the amount of D-dimers produced as soon as a clot begins to form because the clot dissolution process starts almost simultaneously to keep the blood clot from growing larger than necessary. Surgical procedures may elevate D-dimer levels, as well as pregnancy, inflammation, some types of cancer, and liver disease. Use of **fibrinolytic drugs** (also called clot busters) will also increase the amount of FDP and D-dimers in the bloodstream, so these tests may be used to monitor fibrinolytic medication therapy as well. The reference ranges for D-dimers are based on the testing method used for the assay. The D-dimer analysis requires platelet-poor citrated plasma, drawn in a light blue top tube. The specimen must be spun in a typical benchtop centrifuge for at least 10 minutes before separation of the plasma from the cells. In addition, there are specialized centrifuges that are specially equipped to create platelet-poor plasma in less than 10 minutes for tests that require this type of specimen.

Bleeding Time Test

A bleeding time test is performed to evaluate how well platelets function to form a plug in the capillary beds with induced damage. It is not a test performed from a blood sample, but is a measurement performed directly on the patient. Because there are other tests used to monitor the different parts of the coagulation cascade, the bleeding time is generally used to identify platelet function abnormalities, and it may also detect deficiencies of **von Willebrand's factor,** which is part of the factor VIII used for blood clotting. The platelet count and all other clotting factors may be within the normal reference ranges, but if the platelets do not perform as they should, the patient may still have issues with prolonged bleeding. In the past, the bleeding time test was a common presurgical laboratory order, but the true ability of the bleeding time to detect potential surgical hemorrhage has been called into question in recent years.

Bleeding times have been performed using various methods since they were first developed approximately 85 years ago. Many of these procedures were not standardized, so it was difficult to reproduce the results or to establish reliable normal reference ranges. Historically, the procedures all included a "stab" into soft tissues of the body (usually the earlobe) and monitoring the wound until bleeding had stopped. A lancet or surgical blade was used, but there was a lack of uniformity concerning the depth or length of the incisions created.

Test Your Knowledge 15-11

List two reasons that the bleeding time test is not considered to be the test of choice for platelet function analysis as it once was. (Outcome 15-10)

Most modern facilities that perform bleeding times today use a modified Ivy method. This method includes the use of the Surgicutt, a spring-loaded lancet device manufactured by the International Technidyne Corporation. The blade in the lancet device has a depth of 1 mm and a width of 0.5 mm. The device is easy to use, and the blade retracts after use so that there is minimal opportunity for accidental employee exposure to patient fluid. The standardized depth and width of the incision has assisted in the determination of reliable reference ranges and safety for the patients. There are also special lancet devices available for use on infants.

Bleeding times are used to determine how long it will take for a patient to stop bleeding after a superficial puncture or incision is made into the skin. The bleeding time procedure requires an incision device (such as the Surgicutt), a sphygmomanometer (blood pressure cuff), an antiseptic wipe, sterile filter paper provided in circles, and an adhesive bandage. Figure 15-1 shows the supplies necessary to perform a bleeding time test. The performance of a bleeding time is a unique procedure and includes the following concepts:

1. It is essential to ask the patient about aspirin use within the past week. Even one aspirin, or other products that include aspirin, can profoundly affect the results obtained with the bleeding time.

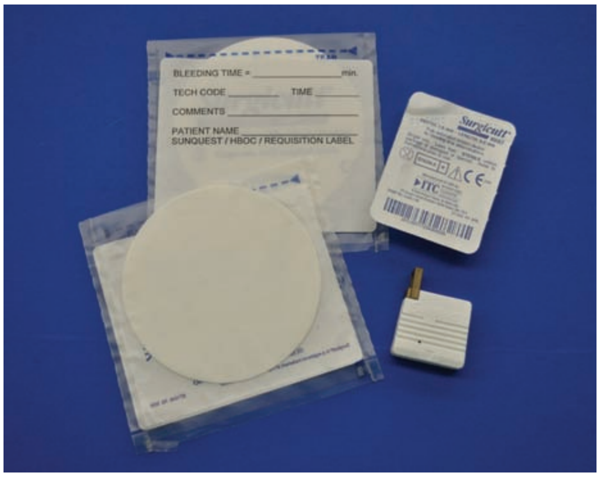
BLEEDING TIME = ________ min.
TECH CODE ________ TIME ________
COMMENTS
PATIENT NAME
SUNQUEST / HBOC / REQUISITION LABEL
Surgicutt
ITC

Figure 15-1 Bleeding time supplies including Surgicutt lancet and filter paper.

The patient should be asked whether he or she is under therapy with heparin or oral blood thinners such as warfarin. If the patient has taken any of these medications, you must check with the physician before performing the test, as the validity of the results may be affected. Long-term use of other NSAIDs (such as ibuprofen) may also increase the bleeding time.

2. The bleeding time is to be performed 2 to 3 inches below the bend of the elbow on the skin of the forearm. Visible veins, moles, tattoos, scars, bruises, shunts, or fistulas and edematous areas should be avoided, as well as areas with excessive hair growth. Clean the site thoroughly with an alcohol antiseptic wipe and allow the alcohol to dry completely before continuing with the procedure.
3. Apply a blood pressure cuff above the elbow to the arm to be used, and inflate the cuff to 40 mm Hg.
4. The bleeding time device (a Surgicutt lancet) is to be applied to the arm in a horizontal manner, and pressure must not be used when making the incision. Start the stopwatch as soon as the incision is made. At 30-second intervals, wick away excess blood with filter paper, taking care not to disturb the actual incision site. Rotate the filter paper as the blood is blotted away each time so that each wicking session is definitive.
5. When no more blood can be blotted away, stop timing the process and release the blood pressure cuff. The time that has elapsed since the incision was made is the bleeding time. Normal results are usually 1 to 9 minutes. If the results are less than 1 minute or more than 9 minutes, some laboratories will require that the test be repeated, as there may be procedural errors that contributed to the result.

Bleeding times are not used as frequently as they once were, and many laboratories no longer offer this test as an option. It has been shown that the correlation of an elevated bleeding time to the risk of excessive surgical bleeding is not always conclusive, and the results for bleeding times are not reproducible. The procedure often causes a scar at the site of the incision, which can be a significant consequence to some patients. In addition, the use of NSAIDs (such as ibuprofen or Advil) may affect the bleeding time, in addition to aspirin or aspirin-containing products. The effect of the drugs on the bleeding time result is not predictable; the bleeding time may be extended, but it is difficult to predict the amount that it will be affected. Other variables may be the result of incorrect technique; if the incision is too deep or too shallow, the results will not be accurate. Also, when blotting away excess blood, if the incision is touched, the bleeding time may be falsely elevated.

It is important to realize that the PTT and PT tests are ordered much more commonly than the bleeding time test. These tests have fewer variables that could cause erroneous results than the bleeding time does, and abnormal results are mapped more specifically to the various coagulation factors. The bleeding time technique also requires specific training, whereas the PT and PTT tests require a more common capillary or venipuncture procedure. In addition, the bleeding time requires 10 to 20 minutes to complete, which may be too time intensive for busy practice or laboratory settings.

Other Tests Used to Assess Platelet Function

Platelet function cannot be measured just by counting the number of cells, and screening for platelet dysfunction is not as precise or easily performed as is a platelet count. The bleeding time test may help to identify gross platelet function abnormalities, but it is not sensitive or specific enough to identify all issues. Many reference and hospital laboratories are now using the Platelet Function Analyzer 100 (PFA-100), manufactured by Siemens Corporation. This machine tries to simulate the bleeding process that occurs in the body by using a tube of blood obtained through a venipuncture. The testing process involves the addition of whole blood into a tube that is coated with additives that promote platelet adherence and aggregation in order to mimic these actions within the human body. The machine measures the amount of time that is necessary for a blood clot to form within this treated tube, and this is reported as a closure time (CT).

There are other types of automated platelet function tests available, but they are not used as widely as is the PFA-100. Controversy still exists about whether these new automated tests are as useful for predicting overall platelet function assessment as the bleeding time once was, even though the bleeding time test is no longer performed in many institutions.

SPECIMEN REQUIREMENTS FOR COAGULATION TESTING

With the exception of CLIA-waived point-of-care PT testing, when collecting samples for coagulation testing, a light blue top tube with sodium citrate is used for most coagulation tests ordered. This includes the fibrinogen test, PTT, PT, and sometimes the FDP test. Some laboratories use a special FDP tube that only requires 2 mL of blood in a 5-mL tube.

- Evacuated light blue top vacuum tubes may include different concentrations of citrate (with the same volume of fluid anticoagulant in the tube), so it is important to clarify which type should be drawn for the testing method used in the laboratory where the testing will occur. The light blue top tubes may be available with 3.2% citrate or 3.5% citrate. In all situations, a ratio of one part anticoagulant to nine parts blood must be maintained, or the result will be erroneous because of specimen dilution.
- The citrate tube must be filled until the vacuum has been exhausted; no short samples can be tested. If the evacuated tube is not filled until the vacuum is exhausted, the clotting time (and subsequently the testing results) will be falsely elevated because of the dilution factor with the anticoagulant in the tube.
- Tubes are spun in the centrifuge before the test is performed, and the plasma is tested immediately or separated from the cells. Ideally, the sample will be tested within 1 hour of collection, or within 4 hours if the specimen has been refrigerated.
- For patients who have a hematocrit measuring above 55%, it is necessary to reduce the amount of anticoagulant used in the tube to avoid prolonged clotting times because of the change in the ratio of plasma to anticoagulant. Laboratories will have an established procedure for the performance of this process.
- It is critical that samples used for coagulation testing remain free of IV fluid or other contaminating substances. Many times the samples are drawn as part of the dialysis process, so it is especially critical to clear the line of all other fluids before the specimen is obtained.
- Clotted or hemolyzed samples will also not be acceptable for coagulation testing, as the results will be inaccurate. To avoid microclot formation in the tube, it is very important that the tubes used for coagulation studies are inverted at least eight times as soon as possible after collection.

Test Your Knowledge 15-12

True or False: Light blue top tubes collected for coagulation studies may be partially filled with no negative effects. (Outcome 15-13)

A lavender top tube with EDTA anticoagulant is used for platelet counts. Platelet counts may be performed as part of a complete blood count, or they may be ordered separately when needed. As in the case with the light blue top citrate tube, it is very important that the lavender top tube be mixed thoroughly right after the blood draw to avoid microclot formation. Platelet counts are performed on whole blood, so the specimen is not centrifuged after collection.

SUMMARY

Coagulation and clot dissolution are vital processes to maintain homeostasis in the human body. The repair of damaged blood vessels within the body is a complex process that includes platelets and a variety of chemical reactions to work efficiently. Coagulation studies determine whether there are enough platelets and other chemicals present for the process to be successful, as well as evaluating the efficiency of the clotting process to predict excessive bleeding or unnecessary blood clot formation. Other disease processes may be complicated by formation of small blood clots, so patients may be medicated with anticoagulant drugs to avoid potential problems. This medication use must be monitored closely with the use of coagulation tests on a regular basis. The anticoagulant medications target the various aspects of the clotting cascade and require different testing mechanisms to monitor their levels. Common tests used to monitor coagulation processes in the body and medications that affect the clotting mechanisms include the PT/INR test, platelet counts and platelet function tests, and the activated partial thromboplastin time. Other tests may be performed to evaluate the amount of clot dissolution or fibrinolysis occurring in the body. Coagulation testing may be performed in the physician office laboratory as well as in hospital and reference laboratories.

TIME TO REVIEW

1. *Aggregate* is a word used to describe: Outcome 15-1
   - **a.** Platelet adhesion and clumping
   - **b.** Consolidation of clotting factors
   - **c.** Deep vein emboli
   - **d.** Fibrinolysis
2. Which of these terms might be used to define FDP? Outcome 15-1
   - **a.** The factors necessary to form a clot
   - **b.** Products produced as a clot dissolves
   - **c.** Platelet count
   - **d.** Clotting time
3. A substance that is viscous may also be described as: Outcome 15-1
   - **a.** Runny
   - **b.** Thin
   - **c.** Syrupy
   - **d.** Dry
4. Fibrinolysis is an important part of hemostasis because: Outcome 15-3
   - **a.** Fibrinolysis helps to control the number of unwanted blood clots in the body
   - **b.** Fibrinolysis is the first step in coagulation
   - **c.** Fibrinolysis helps to keep the blood viscous
   - **d.** None of the above
5. True or False: A diagnosis of atrial fibrillation increases the risk of unwanted blood clot formation. Outcome 15-5
6. True or False: Postsurgical prophylactic treatment is provided only to patients with hemophilia. Outcome 15-5
7. Platelet counts may be ordered as part of a coagulation workup because: Outcome 15-8
   - **a.** Adequate numbers of platelets are necessary for hemostasis
   - **b.** Platelets play a very important role in the repair process for damaged vessels
   - **c.** Extremely high numbers of platelets can contribute to excessive clot formation
   - **d.** All of the above
8. What are two potential sources of error for the CLIA-waived PT/INR test procedure? Outcome 15-7
9. How are D-dimers and fibrin degradation products similar? Outcome 15-12

Case Study 15-1: **Presurgical puzzle**

Mr. Hammond will be having a total hip replacement next week. His physician orders a coagulation screen, including a PT, PTT, platelet count, fibrinogen level, and a bleeding time. All results are within the normal range except for the bleeding time, which at 12 minutes is elevated slightly.

1. What may be an explanation for this situation?

Case Study 15-2: **Relative values?**

Cindy Lou is a new medical assistant who has just started to work in the laboratory. She has experience a high volume of patients this morning, and she feels a bit overwhelmed. Later in the day, the physician office receives a call from the laboratory with abnormal PT and PTT results for one of the patients that Cindy drew that morning. The results for both tests are elevated, but the patient has no signs of excessive bruising or bleeding. These tests were ordered as a screen prior to an elective surgery that the patient was to have done next week.

The physician asks Cindy to contact the laboratory and see if the results for the platelet count are available. Cindy talks to the technician performing the hematology testing, and she is told that the specimen for the platelet count needs to be redrawn because it was clotted.

1. How do you think the specimen condition for the platelet count and the other tube used for the PT/PTT testing might be related?
2. How does this situation with the clotted tube relate to the elevated results for the PT and PTT?

RESOURCES AND SUGGESTED READINGS

"Atrial Fibrillation and Warfarin"
Excellent explanation of the risks involved with atrial fibrillation as related to blood clots, and the use of warfarin for treatment http://www.patient.co.uk/showdoc/23068883

"D-dimers and Fibrin Degradation Products"
Information about coagulation studies performed at Massachusetts General Hospital; great information about the differences between FDP and D-dimers http://www2.massgeneral.org/pathology/coagbook/CO002700.htm

"Disseminated Intravascular Coagulation (DIC); Consumption Coagulopathy; Defibrination Syndrome"
Details about disseminated intravascular coagulation disease. http://www.merck.com/mmpe/sec11/ch136/ch136c.html

"Hemophilia, Factor VIII Deficiency"
Facts about bleeding and clotting disorders http://www.hemophilia.org/NHFWeb/MainPgs/MainNHF.aspx?menuid=179&contentid=45&rptname=bleeding,

"Warfarin and Chinese Medicine"
Details about the development of warfarin, plus lots of information related to herbs and Chinese medications that affect bleeding tendencies. http://www.itmonline.org/arts/warfarin.htm

"ITC, The Point of Care, Surgicutt Bleeding Time Device"
Detailed information about the Surgicutt device and procedure for bleeding time testing. http://www.itcmed.com/Products/Surgicutt

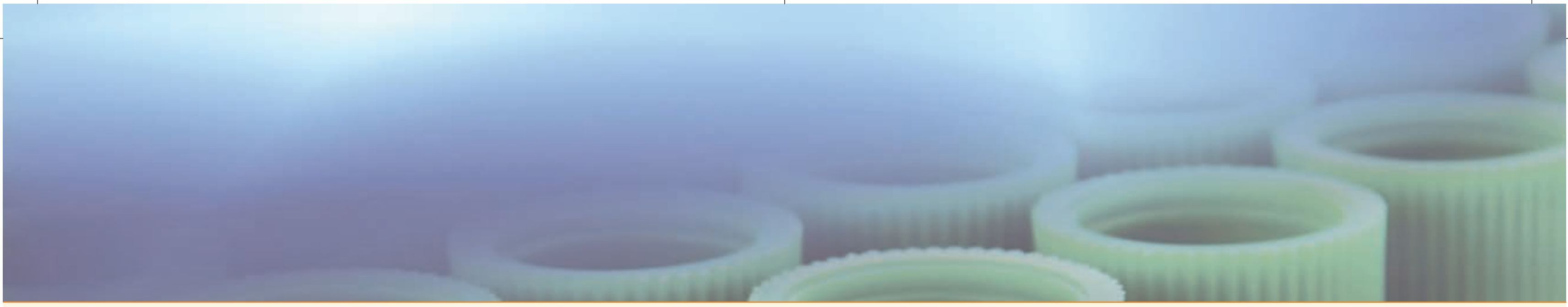

Section III

Hematology and Coagulation

What Does It All Mean?

As you can see, hematology and coagulation testing provides invaluable information to physicians and other primary health-care providers. Now, let us revisit our Case in Point for this section and see what it all means.

Case in Point

At the beginning of this section we met M.J., a 38-year-old male patient who was worked into the doctor's schedule because he felt tired and weak and experienced bleeding gums after brushing his teeth. Patient history revealed that M.J. received a blood transfusion following a tractor accident several weeks prior to this appointment. Upon your encounter with M.J., you noted that he seemed very pale. The doctor ordered a CBC, PT, and PTT to help assess the situation. Upon examination of the results obtained (located on page 258), no wonder M.J. felt so poorly.

M.J. is suffering from a condition known as disseminated intravascular coagulation (DIC). *Disseminated*, in this case, means "widespread over the body." *Intravascular* refers to taking place within blood vessels, and *coagulation* means "the clotting of blood." Under normal conditions, the human body has built-in mechanisms to clot blood when injury occurs and to dissolve clots when appropriate. This balance between clotting and bleeding is known as *hemostasis*. However, when the clotting and bleeding balance is disrupted, individuals experience abnormal bleeding, as noted when M.J. bled while brushing his teeth, and clotting. This imbalance can cause serious injury to the organs of the body. Individuals become anemic, as evidenced by M.J.'s complaints of being tired and weak, as well as the observation of the patient being pale. This anemia results because of the incredible loss of blood, which in turn results in an inability of red blood cells to transport oxygen to the tissues. If the condition is left untreated, death may result. Fortunately, in M.J.'s case, he recognized the abnormal situation taking place in his body and sought help.

Although laboratory results are important in confirming the diagnosis and monitoring of DIC, physicians often make the initial diagnosis based primarily on clinical symptoms. Unfortunately there is no single laboratory test that is diagnostic for DIC; a battery of tests must be performed and the results examined in concert with each other rather than independently. In addition to the tests presented in this case study there are additional laboratory tests that may be performed to provide more information about the clinical situation when DIC is suspected. A further discussion of these additional tests is beyond the scope of this text.

On the Horizon

A wide variety of substances, often referred to as analytes, make up the test menu available in the chemistry section of the clinical laboratory. Likewise, the test methodologies that may be used to determine these analytes are many. Some tests are performed on complex instruments, whereas others are conducted using relatively simple-to-operate equipment. The results obtained from such tests provide healthcare team members with a wealth of knowledge in the diagnosis and monitoring of numerous diseases and conditions.

Relevance for the Medical Assistant

Medical assistants (MAs) typically assume one or both of these roles associated with chemistry laboratory tests: (1) specimen collection and processing and (2) performance of select chemistry tests. In addition to performing CLIA-waived tests, MAs may also perform moderate-complexity tests with appropriate training. Further, MAs may also review patient charts while performing their other duties and in the process may examine laboratory results (from this area as well as all areas of the laboratory). Familiarity of common laboratory tests and their significance allows MAs to take the information obtained from such reviews into account as they work with their patients.

Case in Point

You are working at a health fair being conducted at Acme Innovations, Inc. Part of the health fair consists of collecting blood for a fasting blood glucose level. The following result was obtained on Carrie W., a 34-year-old female Acme employee:

Test	Patient Result	Reference Range
Fasting blood glucose	215 mg/dL	70–100 mg/dL

Questions for Consideration:

- What does the word fasting refer to in this case?
- What is the purpose of the fasting blood glucose test?
- What laboratory test could be ordered to help assess this patient and to monitor Carrie's condition?
- Suppose that Carrie is pregnant and her doctor wishes to assess her for gestational diabetes. What laboratory test could be ordered in this situation?

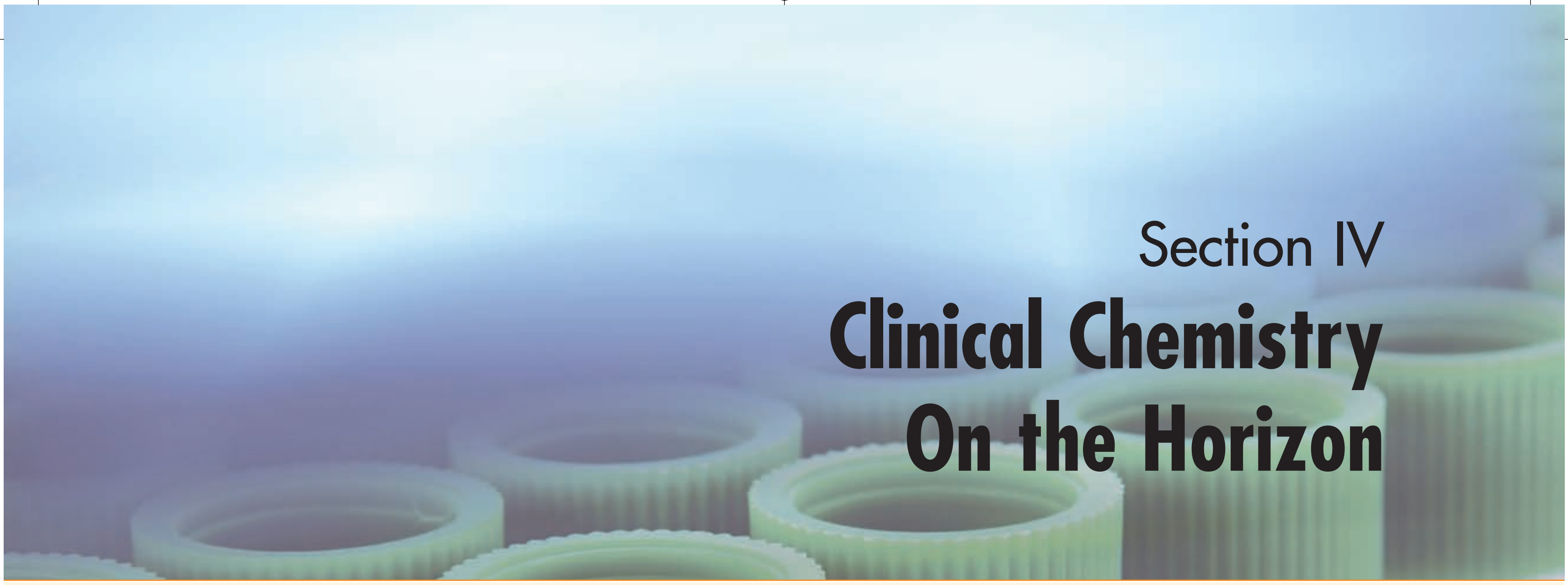

Section IV

Clinical Chemistry On the Horizon

The content in this section is limited to three chapters and serves as an overview of the chemistry section of the clinical laboratory:

Chapter 16: Overview of Clinical Chemistry introduces the reader to the scope of the chemistry section of the clinical laboratory, including the identification and brief description of specimens upon which this testing is performed. A brief overview of some of the more common individual chemical analytes, including glucose and cholesterol as well as groups of tests known as profiles, are covered.

Chapter 17: Glucose Testing covers the physiology and pathophysiology associated with blood glucose. The various glucose tests available, including venous blood samples, hemoglobin A1c, glucose tolerance tests, and capillary puncture glucose testing, are described. Suggested procedures and quality control considerations are also covered.

Chapter 18: Other Select Chemistry Tests covers the physiology and pathophysiology associated with the following representative chemistry tests: cholesterol, lipids, and electrolytes. Proper collection and testing technique, reference ranges, quality control issues, sources of error, and test interpretation are included.

Chapter 16

Overview of Clinical Chemistry

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Clinical Chemistry
- Specimen Types Used for Clinical Chemistry Analysis
  - Plasma
  - Serum
  - Other Specimen Types Used for Clinical Chemistry Testing
- CLIA–Waived Clinical Chemistry Tests
  - Glucose Testing
  - Cholesterol and Lipid Testing
    - *Cholesterol*
    - *Triglycerides*
    - *Lipid Panels*
  - Electrolytes
- Other Common Clinical Chemistry Tests
  - Blood Urea Nitrogen
  - Creatinine
  - Thyroid Panels
  - Comprehensive Metabolic Panel and Basic Metabolic Panel
  - Cardiac Enzymes
  - Hepatic or Liver Profile
- Reference Ranges
- Potential Sources of Error
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

16-1 Define the key terms.

16-2 Relate how serum differs from plasma.

16-3 Describe common substances dissolved in plasma.

16-4 Examine the reasons that glucose testing is performed in the physician office laboratory.

16-5 Cite the tests included in a lipid panel.

16-6 List common analytes included in an electrolyte panel.

16-7 Differentiate which organ system is evaluated using a blood urea nitrogen test and creatinine test.

16-8 Evaluate the purpose of ordering tests in a panel format rather than individually.

16-9 Describe the tests commonly ordered as cardiac enzymes.

16-10 Explain the rationale for serial cardiac enzyme blood specimen collections.

16-11 Restate the liver profile tests most commonly ordered by physicians.

16-12 Outline the parameters that must be taken into account when considering reference ranges for clinical chemistry tests.

16-13 Examine potential sources of error for chemistry testing procedures.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

None

ABHES Standards

None

KEY TERMS

- Analytes
- Atherosclerosis
- Basic metabolic panel (BMP)
- Blood urea nitrogen (BUN)
- Brain natriuretic peptide (BNP)
- Cholesterol
- Clinical chemistry
- Comprehensive metabolic panel (CMP)
- Creatine kinase (CK)
- Creatinine
- Diabetes mellitus
- Electrolytes
- Endogenous cholesterol
- Exogenous cholesterol
- Gestational diabetes
- Glycosylated hemoglobin (Hb A1c)
- High-density lipoprotein (HDL)
- Hemolysis
- Hepatic function panel
- Homeostasis
- Hyperlipidemia
- Hyperthyroidism
- Hypothyroidism
- Icteric
- Ischemia
- Jaundice
- Lipemia
- Lipid
- Lipoproteins
- Low-density lipoprotein (LDL)
- Lytes
- Myocardial infarction
- Myoglobin
- Occlude
- Physiology
- Plaque
- Plasma
- Quality not sufficient (QNS) specimen
- Qualitative
- Quantitative
- Serially
- Serum
- Thyroid-stimulating hormone (TSH)
- Thyroxine (T4)
- Triiodothyronine (T3)
- Troponin
- Very low-density lipoprotein (VLDL)

CLINICAL CHEMISTRY

When scientists first began to study the human body, their focus was on anatomy. Science was not able to analyze what could not be seen with the naked eye, so much of the **physiology** (or function) of the body was unknown. Many hypotheses were formed, but it took a very long time before they could be proved or disproved. It was obvious that the human body functioned best when in a state of **homeostasis** (internal balance), and the scientists could recognize some of the more visible means by which this balance was achieved. It was easy to see how the iris in the eye constricted when exposed to bright light or when perspiration was formed on the forehead in hot weather. However, many of the small changes within the human body were not so obvious; they are based on adjustments to the levels of chemicals needed to keep our body functioning appropriately. It was not until the 19th century that scientists began to realize how closely the appropriate balance of these chemicals in the body was linked to overall health.

Clinical chemistry includes the **quantitative** analysis of the various **analytes** (substances being analyzed; in this case, chemicals) dissolved in the fluids of our bodies. Quantitative tests provide an actual number that represents the amount of a substance present in the body.

Qualitative testing, which indicates the presence or absence of specific chemicals, may also be performed in the clinical chemistry department in the laboratory. Chemical elements are present in our bodies at all times, but increases or decreases in the levels of certain analytes may be indicative of a disease process. Clinical chemistry testing allows the health-care provider to evaluate these changes and use them to diagnose and prescribe treatment. Procedures used to monitor drug levels for some prescription drugs or to identify substances that may be present in the body in the case of environmental or drug poisoning are also performed in the clinical chemistry area of the laboratory.

Test Your Knowledge 16-1

How are qualitative tests different from quantitative tests? (Outcome 16-1)

Reference laboratories may offer hundreds of clinical chemistry tests on various types of specimens. Although all these tests are clinically significant, many of them are only used in the presence of specific disease states. In this chapter, we concentrate on the most commonly performed clinical chemistry tests using blood, urine, and stool specimens.

SPECIMEN TYPES USED FOR CLINICAL CHEMISTRY ANALYSIS

Common specimens in clinical chemistry testing include urine, **serum,** and **plasma.** Many other fluid specimens, including cerebrospinal fluid, synovial fluid taken from the joints of the body, semen, amniotic fluid, and peritoneal fluid from the abdominal cavity, are analyzed for quantitative chemistry. Occasionally chemistry testing may be performed using whole blood. This is usually done in physician office laboratories using CLIA-waived testing methods. In larger laboratories, most clinical chemistry testing is performed on plasma or serum specimens separated from the sample after centrifugation. Serum is obtained from clotted blood and plasma comes from whole blood.

Plasma

As presented in Chapter 8, plasma is the liquid portion of the blood in our body in which the blood cells are suspended. This liquid makes up more than 50% of the total blood volume. We also use the word *plasma* to describe the liquid portion of the blood when collected in a tube containing anticoagulant. Plasma is made up of at least 90% water, in which there are many dissolved substances and gases. There are too many chemical compounds present in plasma to list them all separately, but general categories include the following:

- **Plasma proteins:** Plasma proteins are large molecules that are not excreted as waste products and which are designed to remain in the blood. These include antibodies (immunoglobulins), as well as blood-clotting proteins (fibrinogen and prothrombin) and albumin.
- **Medications:** If an individual must take medication, it is transported throughout the body in the plasma.
- **Electrolytes:** Electrolytes are compounds that have a positive or negative charge when dissolved in water, and are capable of transmitting an electrical impulse within the body.
- **Hormones:** Endocrine glands throughout the body secrete hormones directly into the bloodstream. Hormone tests are commonly ordered to verify the function of the various endocrine glands, such as the pituitary gland and thyroid gland.
- **Lipids:** Fat molecules such as cholesterol are transported through the bloodstream.
- **Enzymes:** Protein-based molecules that catalyze (speed up) chemical reactions are known as enzymes.
- **Food:** When digestion is complete, the food we eat has been broken into components that can be easily used by our body for energy or building blocks for other molecules. These breakdown products are absorbed into our bloodstream and dissolve in the plasma for transportation. Fatty acids, proteins, and simple sugars are carried as food in the plasma. In addition, essential trace elements such as minerals and vitamins are also present.
- **Dissolved gases:** CO_2 is present in plasma as a by-product of metabolism. Other dissolved gases include oxygen (even though it is attached to the red blood cells) and nitrogen.
- **Waste:** Cell metabolism creates waste products that are processed in the liver and removed from the body through the urinary system.

Serum

Serum is the liquid portion of blood that has been collected into a tube that does not contain anticoagulant. Plasma and serum are very similar; however, when a specimen is drawn in a tube without an anticoagulant, the proteins and other blood clotting factors are utilized as the blood clot forms. The lack of coagulation factors dissolved in the liquid limits some of the tests that may be performed on a serum specimen.

Test Your Knowledge 16-2

Which of these are present from a centrifuged tube containing no anticoagulant?

a. Serum
b. Plasma
c. Whole blood
d. None of the above (Outcome 16-2)

Other Specimen Types Used for Clinical Chemistry Testing

As mentioned previously in this chapter, clinical chemistry testing may be performed on numerous specimen types. In the physician office laboratory, whole blood may be used for CLIA-waived procedures such as glucose or cholesterol testing, providing a quantitative result.

Urine chemical testing is performed as part of a routine random urinalysis with semiquantitative results reported that indicate a range for the analyte values. In some disease states, it is necessary to obtain results that are more specific, with a specific number reported for the chemical concentration rather than a range. This type of quantitative urine chemistry testing is performed at reference or large hospital laboratories, but not in physician office laboratories. These quantitative urine tests may be used to screen for kidney damage caused by diabetes, hypertension, or other disease processes. Urine may also be analyzed to research the cause of renal calculi (kidney stone) formation. Common quantitative tests performed on urine specimens include microalbumin, total protein, and creatinine. A timed urine specimen (such as a 24-hour urine specimen) is generally used for these tests.

For all clinical chemistry testing, it is important to keep in mind that reference ranges are based on certain specimen types. For instance, serum and plasma look identical once the liquid has been removed from the blood specimen, but the reference ranges for some chemicals are very different in the two specimen types. It is essential to collect the correct type of tube for the analysis when performing a blood collection, and also to label the specimen type if the liquid is removed from the original collection container to avoid any potential problems. Other types of specimens may have a similar appearance if they are transported in an unlabeled specimen container.

CLIA-WAIVED CLINICAL CHEMISTRY TESTS

Glucose Testing

The most common type of CLIA-waived clinical chemistry test performed in the clinic setting is glucose testing. This test may be used to screen for **diabetes mellitus** (type 2 diabetes) or to monitor the glucose levels of patients who have already been diagnosed and are under treatment. **Gestational diabetes** in pregnant women may also be diagnosed with tests performed in the physician office laboratory. It is beneficial to the patient to have these tests performed on site, as the results are available within minutes. Abnormal glucose levels may be reported to the health-care provider immediately, allowing for additional tests to be ordered or for further patient interaction to be initiated if necessary. CLIA-waived glucose testing methods use whole blood specimens, usually obtained from a capillary puncture. The glucose levels may be ordered as fasting, random, or timed specimens.

Test Your Knowledge 16-3

List two reasons that glucose testing might be performed in a physician office laboratory. (Outcome 16-4)

Diabetic patients may also be monitored with a **glycosylated hemoglobin,** or **Hb A1c,** test. Hemoglobin is present in all the red blood cells of the body, and one subunit of this hemoglobin molecule is hemoglobin A. When hemoglobin A is exposed to excessive levels of glucose, it becomes glycosylated, which means that the glucose molecules bind to the hemoglobin A irreversibly. The Hb A1c test can measure the amount of glycosylated hemoglobin present, and because red blood cells live for approximately 3 months in the body, Hb A1c results may be used indirectly as an indication of the overall blood sugar levels for the past few months. Hb A1c may be performed as a CLIA-waived test in the physician office laboratory or automated testing may be performed in a reference laboratory. Whole blood is used for this analysis; capillary samples are used for the CLIA-waived methods, and an EDTA lavender top tube is used for the reference laboratory methods.

Cholesterol and Lipid Testing

Lipids are substances including oils and fats that are insoluble in water and soluble in organic solvents. **Cholesterol** is an example of a lipid, but there are other types of fats in our bodies as well. Our digestive system breaks down ingested fats into fatty acids, which are transported through the bloodstream to the liver, where they are used to manufacture lipids. Because lipids do not dissolve in water, the body attaches lipids to a protein molecule for transport to the cells of the body to be used. These complexes are called **lipoproteins.** Lipids are used as an alternative source of energy when glucose is not readily available, and they are also necessary for hormone production, cell wall integrity, utilization of fat-soluble vitamins, hunger satiation, maintenance of appropriate body temperature, and skin health.

In the laboratory, the two types of lipids most commonly analyzed are cholesterol and triglycerides. Total blood cholesterol and total triglyceride counts provide valuable information to health-care providers about cardiovascular health and potential risk factors for the future. However, more information can be gained by breaking down the relative amounts of lipids according to subtype. These categories are **HDL (high-density lipoproteins), LDL (low-density lipoproteins),** and **VLDL (very low-density lipoproteins).** Elevated total cholesterol, triglyceride, and LDL levels may are considered to be risk factors for coronary heart disease. Chapter 18 provides more in-depth information about lipid testing.

Test Your Knowledge 16-4

Which of these tests is included in a lipid panel?

a. LDH
b. LDL
c. VLDL
d. ALT

(Outcome 16-5)

Cholesterol

Cholesterol is essential for the human body to function properly. Cholesterol is used for hormone production, absorption of vitamin D, bile production, and cellular membrane structure, and is also used as part of the covering (the myelin sheath) on some of the nerves in the body. The liver creates all the cholesterol that we need to function normally. This is called **endogenous cholesterol** because it is manufactured within the body. **Exogenous cholesterol** comes from our diets. When cholesterol builds up in the body it can lead to **plaque** formation within the blood vessels, a condition known as **atherosclerosis.** Plaque buildup may eventually **occlude** (block) blood vessels, or pieces of plaque may break away from the vessel lining and travel as an embolus, lodging elsewhere in the body and causing a blood clot.

Cholesterol levels ordered individually do not require fasting samples. However, cholesterol levels are often ordered as part of a lipid panel with the triglyceride, HDL, LDL, and VLDL levels. These additional tests do require a fasting sample, so patients should be advised to abstain from eating for 12 hours prior to the blood draw.

Triglycerides

Patients who have a diet that is high in carbohydrates may have elevated triglyceride levels. Most of the fat in our bodies is made up of triglycerides. When a person takes in more calories than are needed for energy, the excess calories are converted into triglycerides and stored as fat for use later as an energy source. Those who have diabetes, hypertension, or excess alcohol intake are especially prone to high triglyceride levels. **Hyperlipidemia** is the word used to describe high levels of lipids (cholesterol and triglycerides) in the plasma. Patients who have elevated triglyceride levels may have plasma or serum that appears “milky” because of all the fat molecules present. The presence of the fat molecules suspended in the blood is called **lipemia.** Because triglyceride levels do rise after ingestion of food, triglyceride testing usually requires a fasting blood draw.

Lipid Panels

A lipid panel commonly includes a total cholesterol level, a triglyceride level, HDL, LDL, VLDL, and a total cholesterol : HDL ratio. These values will provide the health-care provider with valuable information concerning the risk factors for coronary artery disease.

POINT OF INTEREST 16-1
Cholesterol testing

Cholesterol testing is performed to assist the health-care provider in evaluating the patient for potential heart disease. The total cholesterol level consists of high-density lipoproteins (HDL) and low-density lipoproteins (LDL). The HDL is considered to be “good cholesterol” because it actually cleans excess cholesterol from the blood vessels in the body and carries it away in the bloodstream. Ideally, the patient will have a higher concentration of HDL in the bloodstream than LDL. Elevated levels of LDL may lead to heart disease, as the low-density lipoproteins will deposit cholesterol on the walls of the blood vessels.

There is a direct correlation of elevated cholesterol results and/or an imbalance of HDL and LDL results and atherosclerosis. The following individuals are at higher risk than others:

- Smokers
- Those diagnosed with diabetes mellitus

Continued

- Patients with hypertension
- Men over 45 and women over 55
- Family history of heart disease
- Obesity

Laboratory tests usually place the patient in one of three categories:

High risk: Having a total cholesterol level of 240 mg/dL or above

Borderline high: Having a cholesterol result of 200 to 239 mg/dL. Often the health-care provider will order additional testing to see if the elevated level is the result of the presence of the LDL (bad cholesterol) or HDL (good cholesterol).

Desirable or normal: Having a total cholesterol level less than 200 mg/dL is considered normal, and places the patient at a low risk of heart disease.

Triglyceride testing is often performed with the cholesterol assay in the laboratory as part of a lipid panel. Triglycerides are a form of fat, and are often carried by another lipoprotein, very low-density lipoprotein (VLDL). Elevated levels of triglycerides in the bloodstream may also lead to heart disease. Poor diet and low levels of exercise may lead to elevated triglyceride levels.

Electrolytes

Routine blood screening often includes an order for **electrolytes.** Electrolytes (often abbreviated as **lytes**) include several minerals that are essential for normal body function. They carry an electrical charge (positive or negative charge) when dissolved in water. Sodium, potassium, chloride, and CO_2 are usually included in an electrolyte assay. Additional tests may also be included by some laboratories, such those for as magnesium and calcium. Electrolyte assays may provide information to the health-care provider about the acid-base balance of the body, causes for edema, or additional information about the status of kidney dysfunction. Electrolyte levels may also be altered by certain drugs, dehydration, and overhydration. The tests within the electrolyte panel may be ordered individually, or as a group. CLIA-waived automated procedures are available for electrolyte testing, but these tests may also be performed at hospital laboratories and reference laboratories. More information is available about the specific testing methods for electrolytes and the clinical significance of the analytes in Chapter 18.

Test Your Knowledge 16-5

Which of these analytes is not included in an electrolyte panel?

a. Sodium
b. Potassium
c. Chloride
d. Cadmium (Outcome 16-6)

OTHER COMMON CLINICAL CHEMISTRY TESTS

So far, we have focused on clinical chemistry tests that may be performed in the physician office laboratory as well as reference laboratories. In this section of the chapter, we focus on some of the most commonly ordered clinical chemistry tests that are not usually performed in the physician office laboratory. These tests would all be performed on serum or plasma in a hospital or reference laboratory. Many of these commonly performed tests are ordered as profiles or panels rather than individual tests. The panels are beneficial to the health-care provider because the tests included in a panel may evaluate various aspects of a specific organ system; in the case of the **basic metabolic panel (BMP)** and **comprehensive metabolic panel (CMP),** they evaluate parameters of several organ systems. Obtaining all the results at once can allow for a differential diagnosis, and also may save the patient time and money by not having to return to the laboratory for separate blood draws. Charges for the tests included in a Medicare-approved panel may not legally surpass the costs of ordering the tests individually, so it is cost effective to perform multiple tests at one time.

Test Your Knowledge 16-6

List two advantages of ordering tests as a panel rather than individually. (Outcome 16-8)

Blood Urea Nitrogen

As protein is broken down by the body, it produces urea as a by-product. Nitrogen is part of this urea molecule. Urea should be cleared from the blood by the kidneys, so if the level blood urea nitrogen is elevated, it may be an indication that the patient has impaired renal function. **Blood urea nitrogen (BUN)** is a measurement of the amount of urea in the bloodstream. This test is often ordered in conjunction with a creatinine test.

Creatinine

Creatinine is a by-product of muscle metabolism. As in the case with the blood urea nitrogen, creatinine should be cleared from the blood by the kidneys. Elevated creatinine levels may indicate impaired kidney function. Creatinine levels may be performed on urine samples to correlate the levels with those obtained from testing the serum or plasma. This provides more in-depth information about potential kidney dysfunction.

Test Your Knowledge 16-7

True or False: BUN and creatinine tests are routinely ordered to evaluate kidney function. (Outcome 16-7)

Thyroid Panels

The thyroid gland secretes several very important hormones directly into the bloodstream. These hormones, **thyroxine (T4)** and **triiodothyronine (T3)** are essential for regulation of cellular metabolism. The thyroid gland also secretes thyrocalcitonin, which helps to stimulate calcium storage in the bones of the body. Many laboratories offer thyroid panels that include the T3 and T4. Another hormone that affects thyroid function is the **TSH** (**thyroid-stimulating hormone),** which is secreted by the pituitary gland. Thyroid-stimulating hormone is necessary for the thyroid to be "stimulated" to produce the T3 and T4. For the thyroid gland to function normally, there must be a source of iodine in the diet. Iodine can be ingested with iodized salt, seafood, or vegetables that were grown in soil with supplementary iodine. **Hypothyroidism** is a condition in which the thyroid gland does not produce enough of the hormones necessary to stimulate cell metabolism at a normal level. Patients may also suffer from **hyperthyroidism,** in which case the thyroid gland is producing hormones at a level that is above normal, resulting in overstimulation of cell metabolism. TSH levels outside of the normal range may cause under- or overstimulation of the thyroid gland, resulting in an imbalance; for this reason, the TSH may also be included in some thyroid panels.

Comprehensive Metabolic Panel and Basic Metabolic Panel

The CMP is a set of 14 tests that screen for problems with the kidneys, glucose metabolism, liver, and acid-base balance of the body. The CMP may be used as a general screening test during a routine office visit, or the panel may be ordered to monitor medication use or changes in a disease process. The tests included in this panel are BUN, creatinine, sodium, potassium, chloride, carbon dioxide, calcium, albumin, total protein, bilirubin, alkaline phosphatase (often abbreviated as ALP), alanine aminotransferase (also called SGPT or ALT), aspartate aminotransferase (which may also be known as SGOT or AST), and glucose. The CMP is a panel that has been approved for payment by the Center for Medicare & Medicaid Services (CMS), so it is acceptable for it to be ordered as a diagnostic test for patients with Medicare coverage.

The BMP includes eight tests that will help to identify problems with electrolytes, kidney function, glucose metabolism, and acid-base balance. The basic metabolic panel may be used as a diagnostic tool, or used to monitor medication use or patient progress with ongoing treatment. The BMP includes glucose, calcium, sodium, potassium, carbon dioxide, BUN, and creatinine levels. The BMP has also been approved for payment as a panel by the CMS, so patients who have Medicare as their primary insurance may have this test ordered knowing that its cost is reimbursable.

Cardiac Enzymes

Cardiac enzymes are ordered in situations in which the health-care provider suspects damage to the heart, as in the case of **ischemia** (a temporary decrease of blood flow to the heart muscle) or in the case of trauma to the heart muscle from a **myocardial infarction** (heart attack). The profile usually includes a total **creatine kinase (CK)** and a CK-MB, in addition to other tests such as **troponin** or **myoglobin.** Creatine kinase is found in muscles other than the heart muscle; therefore, if the total CK level is elevated, additional testing must be performed to identify the source of the abnormal result. CK isoenzymes are subtypes of the creatine kinase analyte. The CK-MB (creatine kinase myocardial band) isoenzyme is a specific marker for heart muscle damage. The CK-MB levels will start to rise in the bloodstream approximately 4 hours after a myocardial infarction, and will peak approximately 18 hours after the event. Troponin levels will be elevated with heart muscle damage, but not with other types of muscle trauma. Troponin I levels are especially sensitive to myocardial damage. Troponin testing may be performed on a STAT basis when a myocardial infarction is suspected because these levels become elevated sooner and remain elevated longer than the CK-MB after a heart attack. This allows for earlier diagnosis and appropriate treatment. It is recommended that the troponin levels are tested every 2 to 4 hours for the first 24 hours after a cardiac event. Myoglobin may also be used as an indicator for a myocardial infarction, because it will be elevated in the bloodstream sooner after a cardiac event than troponin will be. Myoglobin is found in cardiac and skeletal muscles, so the levels may be elevated in the bloodstream in situations other than myocardial infarction because it is not specific to the heart muscle.

When a patient has chest pain and a myocardial infarction is suspected, the cardiac enzymes are ordered **serially,** because the levels of the various components will change with time if there was damage to the heart muscle. Specimens are usually collected every few hours, and the health-care provider looks for a pattern in the elevation of the different components to rule out or differentially diagnose damage to the heart muscle as a result of a heart attack. If the specimens are not drawn several times within the first 24 hours, it is not possible to establish whether a heart attack really occurred. The myoglobin is usually the first test to be elevated, but most health-care providers like to have more information before proceeding with a differential diagnosis. It is important to remember that other diagnostic procedures (such as an electrocardiogram) will also assist with the diagnosis as well.

Another common cardiac analysis is the **brain natriuretic peptide (BNP)** test. BNP is synthesized in the ventricles of the heart. The BNP secretion rate is increased under conditions of additional myocardial stretch and abnormal wall tension. The test is used to differentiate whether dyspnea (difficulty breathing) is the result of pulmonary conditions or cardiac dysfunction. An excessively elevated BNP level is indicative of heart failure.

Test Your Knowledge 16-8

Why are cardiac enzymes ordered as a series of blood draws rather than just once? (Outcome 16-10)

Hepatic or Liver Profile

Disorders of hepatic (liver) function may be diagnosed using a **hepatic function panel.** Many of the hepatic enzymes are also found in other tissues of the body. The use of a panel may be helpful to the provider for a differential diagnosis. Patients with elevated liver enzymes often exhibit **jaundice** (yellowing of the skin and eyes), nausea and vomiting, or urine that is dark in color. Specimens obtained from patients with liver disease or damage may have plasma or serum that appears **icteric.** Icteric samples have a dark yellow or greenish tint to the fluid portion of the blood, which often correlates to an elevated total bilirubin level. Hepatic function panels may also be used to monitor liver function when patients are taking medication that is known to cause hepatic damage. This panel usually includes at least seven tests that may be elevated with liver dysfunction: ALT (alanine aminotransferase), alkaline phosphatase, AST (aspartate aminotransferase), bilirubin, total protein, and albumin.

Hepatic function panels may include additional tests in some laboratory settings. However, if there are additional tests added to those listed above, Medicare may not pay for the tests performed.

Test Your Knowledge 16-9

Which of these tests are included in a hepatic function panel?

a. AST
b. Total protein
c. ALT
d. None of the above
e. All of the above (Outcome 16-11)

REFERENCE RANGES

It is important to realize that the reference ranges for clinical chemistry tests are based on the technique used by the laboratory performing the test, the age and gender of the patient, the test preparation, and sometimes even by the time of the day that the specimen was drawn. Occasionally there may be a difference in reference ranges for the plasma versus serum samples as well. It is important always to document the details of the blood draw and patient preparation carefully when collecting the blood sample so that the reference ranges will be accurate. It is also essential that the medical assistant learn how to read the laboratory report so that abnormal results are recognized immediately. The reference ranges for the laboratory tests listed in this chapter are included in Table 16-1 and Table 16-2.

POTENTIAL SOURCES OF ERROR

The most common sources of error for clinical chemistry tests occur in the preanalytical area. These include improper patient preparation, inappropriate specimen collection techniques, and errors in specimen processing. Specific sources of specimen processing errors include the following:

- **Hemolysis:** A sample can be hemolyzed (the red blood cells broken open) during specimen collection or afterward during processing. Examples of procedures that may cause **hemolysis** include a traumatic blood draw, using a needle that is too small for the collection process,

TABLE 16-1

Reference ranges for glucose testing

	Fasting Plasma Glucose	2-Hour Postprandial Glucose
Normal	70–100 mg/dL	Less than 140 mg/dL
Prediabetes	101–125 mg/dL	141–199 mg/dL
Diabetes	126 mg/dL or above	200 mg/dL or above

Data from American Diabetes Association.

TABLE 16-2

Reference ranges for common chemical analytes (listed alphabetically)

Analyte	Normal Range (Adults)
Albumin (Alb)	3.5–5 g/dL
Alkaline phosphatase (ALP)	42–136 U/L
Alanine aminotransferase (ALT)	10–35 U/L
Aspartate aminotransferase (AST)	0–35 U/L
Bilirubin, total (TBili)	0.3–1 mg/dL
Blood urea nitrogen (BUN)	10–20 mg/dL
Brain natriuretic peptide (BNP)	0–100 ng/L
Calcium (Ca)	8.2–10.5 mg/dL
Carbon dioxide (CO_2)	22–30 mEq/L
Chloride (Cl)	96–106 mEq/L
Cholesterol, total (Chol)	Less than 200 mg/dL
Creatinine (Creat)	0.6–1.2 mg/dL
Creatine kinase (CK) aka Creatine Phosphokinase (CPK)	55–170 U/L
High-density lipoprotein (HDL)	Greater than 50 mg/dL
Lactate dehydrogenase (LD, LDH)	100–190 U/L
Low-density lipoprotein (LDL)	Less than 100 mg/dL
Myoglobin	Less than 90 μg/L
Potassium (K)	3.5–5.0 mEq/L
Sodium (Na)	136–145 mEq/L
Thyroid-stimulating hormone (TSH)	0.4–4.2 μU/mL
Thyroxine (T4)	4.5–11.2 μg/dL
Triglyceride (Trig)	Less than 150 mg/dL
Triiodothyronine (T3)	75–220 ng/dL

shaking a tube after collection rather than using gentle inversion, or subjecting the whole blood sample to extreme heat or cold temperatures. Hemolysis may also result if a tube without anticoagulant is centrifuged before the specimen is allowed to clot adequately. Any activity that causes the red blood cells in the specimen to be agitated unnecessarily may potentially cause the red blood cells to be broken. Hemolysis causes the serum or plasma to take on a red color, which interferes with many testing methods. Figure 16-1 shows a specimen

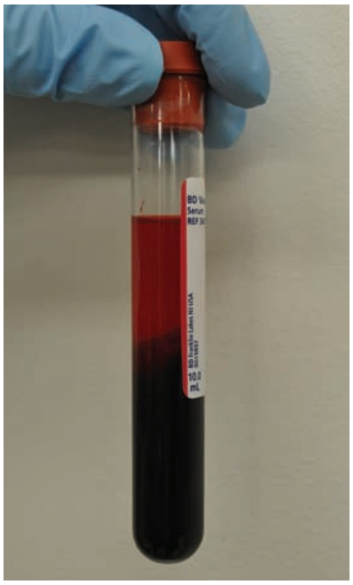

Figure 16-1 A centrifuged blood specimen with hemolysis evident in the serum.

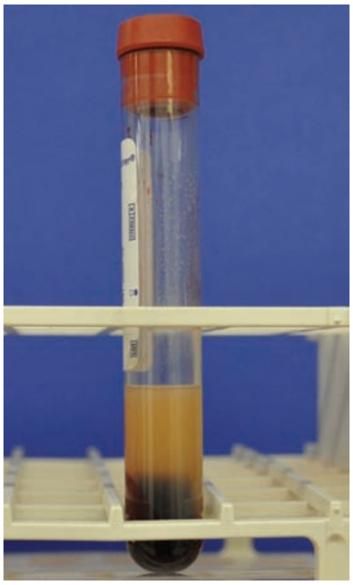

Figure 16-2 A centrifuged blood specimen with lipemia evident in the serum.

with visible hemolysis. Hemolysis may also increase levels of certain analytes in the plasma (potassium or iron, for instance) that are not normally present in high concentrations. The reference ranges for these chemicals are based on nonhemolyzed specimens, so it may appear that the patient needs treatment for an imbalance when the result is erroneous due to hemolysis.

- **Lipemic specimens:** Lipemia is present in a specimen when there are too many lipoproteins in the blood circulation. The excessively high levels of lipoproteins may be caused by hereditary hyperlipidemia or chronic liver disease. Malabsorption disorders may also cause lipemia to be evident. More often, lipemia results when a patient has not followed the instructions for specimen preparation by not fasting as directed before the blood draw. Lipemic specimens have a "milky" appearance that may be visible immediately after collection or after centrifugation. These fatty particles suspended in the specimen will interfere with many clinical chemistry assays. Some testing may be possible if the specimen is "cleared" using a special centrifugation process, but often the laboratory will ask that the specimen be redrawn if the lipemia is severe. Lipemia is evident in the specimen included in Figure 16-2.
- **Specimen collection errors:** The medical assistant who is performing specimen collection needs to keep the difference between plasma and serum samples in mind. If a test calls for a serum sample, a tube without anticoagulant must be used for the collection. In addition, certain test results may also be affected if a tube with serum/plasma separator gel is used for the collection. Specifics about the type of specimens that are acceptable and unacceptable for a particular test would be listed in the laboratory directory, which provides collection specifics for the tests offered by that laboratory. This SST restriction is most common when collecting specimens for medication levels.

Another potential source of error when collecting clinical chemistry specimens is the incorrect timing of the collection. Tests may be ordered as a series of blood draws, as in the case of the cardiac enzymes. Medication levels may also be ordered at specific times as peak and trough draws. It is important to understand the process involved in these timed collections so that the specimens are properly labeled and the reference ranges are matched to the specimen type. In the case of medication levels, it is important to find out if there are restrictions concerning the length of time between the last dose and the specimen collection so that the health-care provider will receive results that allow for proper treatment of the patient.

In addition to the incorrect type of specimen collection and the incorrect timing, another potential source of error is the order in which the tubes are collected. Carryover of anticoagulant from one tube to another may cause changes in the levels of certain analytes when the testing procedures are performed. Potassium levels, for example, may be elevated if the lavender top tube is collected before the green top tube and a carryover of anticoagulant results.

- **Quantity not sufficient:** Laboratory directories provide information about the volume needed for testing procedures, and they may also provide a minimum volume accepted for the test. A **QNS specimen** (one that does not have the quantity necessary for testing) will require a re-collection, which delays potential treatment for the patient. These volume requirements must be taken into consideration before the sample is collected; they can affect the choices made concerning which type of collection setup to use and how many tubes to draw. The stated minimum volumes usually allow the test to be performed *only once;* this means that if there are any errors during the analysis or an extremely high or low result, retesting the sample will not be possible. Sometimes it is necessary to retest the sample using dilutions to obtain a true value, but this would also not be possible if only the minimum volume is submitted to the laboratory.
- **Exposure to light:** Some analytes (bilirubin and ferritin, for example) will deteriorate when exposed to light after collection. These samples need to be covered immediately (usually wrapping foil around the tube will suffice) and centrifuged as soon as possible. The plasma or serum to be analyzed should be separated into a tube designed to protect the specimen from light exposure. Figure 16-3 shows an amber plastic transfer tube that is designed to minimize the light exposure for a specimen.
- **Exposure to air:** Once a specimen has been centrifuged, it may be necessary to remove the rubber top of the tube and separate out the plasma or serum. The exposure to air should be minimized, as the concentration for some analytes will change with the exposure. Alcohol analysis is an excellent example of this; the longer the tube is open, the lower the alcohol concentration may become in the specimen. Carbon dioxide is another common analyte that may decrease with continued exposure to air.
- **Specimen processing:** Careful adherence to specimen processing is essential for clinical chemistry tests. Many of the chemical concentrations will change when left at room temperature for extended periods of time, so researching the specimen requirements prior to processing is important. If there is a question about how to handle a specimen, the laboratory directory should be consulted, and if necessary, the testing laboratory should be contacted for clarification *before* the specimen is collected. Plasma or serum may require freezing or refrigeration within a certain amount of time after collection to maintain the integrity of the sample. It is also necessary to prechill specimen collection tubes for some analytes.

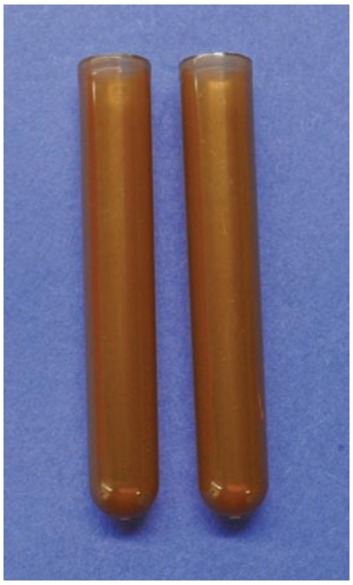

Figure 16-3 Amber transfer tubes designed to minimize specimen exposure to light.

- **Delayed processing:** Blood specimens to be used for clinical chemistry should be centrifuged as soon as possible so that the plasma or serum can be removed from the cells in a timely manner. Serum tubes must be allowed to clot completely before centrifuging. Delayed separation of cells from plasma or serum will allow analytes that are in a higher concentration within the cells than outside the cells to "leak" out of the cells and cause erroneous test results. This can result in an inaccurate diagnosis and/or unnecessary treatment for the patient. Conversely, glucose that is present in the plasma or serum will continue to be used by the cells in the specimen as an energy source if the serum or plasma is not removed from the cells in a timely manner. Therefore, the longer that the cells remain in contact with the fluid portion of the blood, the lower the glucose levels will become in the plasma or serum. This can make it appear that the patient has a very low plasma glucose level when the specimen is tested in the laboratory.

Test Your Knowledge16-10

Which of these potential sources of error are related to high levels of lipids in the bloodstream?

a. Lipemia
b. Hemolysis
c. Erroneous low glucose levels
d. QNS specimens (Outcome 16-13)

SUMMARY

Clinical chemistry is a complex component of laboratory testing. There are hundreds of tests that may fall into this category, and the specimen collection and processing requirements will vary depending on the testing method and laboratory policies. Some clinical chemistry tests may be performed in physician office laboratories; most of these will use whole blood specimens for analysis. Common clinical chemistry tests performed in the physician office laboratory include glucose and cholesterol testing. Clinical chemistry tests performed at reference laboratories are usually performed on plasma or serum specimens rather than whole blood specimens. Many of these tests may be ordered as panels or profiles. There are numerous sources of error in specimen collection and processing for clinical chemistry tests, so the medical assistant has to remain diligent about specific specimen requirements to provide appropriate specimens for meaningful laboratory results.

TIME TO REVIEW

1. Atherosclerosis is: Outcome 16-1
   - a. Hardening of the arteries due to age
   - b. A buildup of waxy plaque on the lining of blood vessels
   - c. A condition related to high glucose levels
   - d. The presence of occult blood in the stool

2. Creatinine is present in the blood as a by-product of: Outcome 16-1
   - a. Protein metabolism
   - b. Heart damage
   - c. Muscle metabolism
   - d. Glucose metabolism

3. Myoglobin and troponin are examples of: Outcome 16-1
   - a. Tests included in a typical hepatic function panel
   - b. Tests included in a typical renal function panel
   - c. Tests included in the BMP
   - d. Tests included in an order for cardiac enzymes

4. True or False: Plasma is the liquid portion of the blood separated after centrifugation from a tube that does not contain anticoagulant. Outcome 16-2

5. True or False: Hemoglobin is a substance that is dissolved in blood plasma. Outcome 16-3

6. Which of these tests are not included in a typical lipid panel? Outcome 16-5
   - a. Total cholesterol
   - b. Total triglyceride
   - c. VLDL
   - d. Creatinine

7. An electrolyte panel typically includes which of these tests? Outcome 16-6
   - a. BUN
   - b. CPK
   - c. ALT
   - d. Potassium

8. True or False: A panel of laboratory tests costs more for the patient than ordering all the tests in the panel individually. Outcome 16-8

9. Which of these tests are not usually included in a cardiac panel? Outcome 16-9
   - a. ALT
   - b. CK-MB
   - c. CK
   - d. Myoglobin

10. Which of these factors may be taken into consideration when evaluating reference ranges for a test? Outcome 16-12
    - a. Age and gender of the patient
    - b. Testing methodology
    - c. Time of day for specimen collection
    - d. All of the above

11. True or False: Hemolysis is avoidable when collecting a blood sample. Outcome 16-13

Case Study 16-1: What order?

Mr. Oliver arrived early for his blood draw one Monday morning. The medical assistant was running late that day, and seemed to be a bit distracted as she prepared for the blood draw. Mr. Oliver's physician had ordered a CBC (to be collected in a potassium EDTA tube) and a potassium level (to be collected in a green top heparinized tube). The MA completed the blood draw, and thanked Mr. Oliver for his time.

The next day, the physician's office called Mr. Oliver and asked him to come in to have his potassium level rechecked, as it was elevated on the test from the day before. This was not an expected result, as usually Mr. Oliver's potassium level was decreased below the reference range. The specimen was drawn and checked on a STAT basis, and the result was in the low end of the reference range. The physician told Mr. Oliver that the previous result was elevated because of a laboratory error.

1. What are two sources of error that could cause the potassium result to be erroneously elevated in this scenario?

RESOURCES AND SUGGESTED READINGS

"Modern Technology Helps Shed Light on Illness in Artists of the Past"
A summary of an article by Paul L. Wolf, MD, which appeared in the November 2005 edition of the *Archives of Pathology and Laboratory Medicine,* a publication of the College of American Pathologists. Very interesting information concerning the way that illnesses and drugs have influenced the work of many famous artists and composers. http://www.newswise.com/articles/view/516145/

"Cholesterol"
Excellent presentation on cholesterol and the effects that it may have on the cardiovascular system; also includes information about lifestyle changes and symptoms of myocardial infarction and stroke http://www.americanheart.org

"Diseases and Conditions Index, Coronary Artery Disease"
National Heart, Lung and Blood Institute; National Institutes of Health
Excellent information about coronary artery disease including illustrations. http://www.nhlbi.nih.gov/health/dci/Diseases/Cad/CAD_WhatIs.html

"Third Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)"
National Heart, Lung and Blood Institute: National Institutes of Health
Includes recommendations for lipid testing and desirable ranges and treatment options. http://www.nhlbi.nih.gov/guidelines/cholesterol/index.htm

Chapter 17

Glucose Testing

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Glucose Utilization and Control Mechanisms
- Pathophysiology of Glucose Metabolism
  - Prediabetes
  - Diabetes
    - *Type 1 Diabetes*
    - *Type 2 Diabetes*
    - *Gestational Diabetes*
- Types of Glucose Tests Performed
  - Fasting Blood Glucose Test
- Random Glucose Test
  - Postprandial Glucose Test
  - Oral Glucose Tolerance Testing Procedures
  - Glycosylated Hemoglobin
  - Blood Ketone Testing
  - Capillary Sample Testing and Correlation to Plasma Glucose Levels
  - Urine Testing for Diabetics
- Glucose Testing Methods
  - Home Glucose Testing Instruments
    - *Quality Control and Common Errors*
  - Laboratory Glucose Testing and Potential Sources of Error
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

17-1 Define the key terms.

17-2 Explain how insulin and glucagon work together to maintain healthy blood sugar levels.

17-3 List three ways that insulin secretion affects the body.

17-4 Compare and contrast type 1 and type 2 diabetes.

17-5 Describe how gestational diabetes is different from type 1 and type 2 diabetes.

17-6 List potential problems that may develop with uncontrolled diabetes.

17-7 Explain why it is important to diagnose and treat gestational diabetes.

17-8 Examine the differences between the random and fasting glucose testing.

17-9 Explain the procedure for postprandial glucose testing.

17-10 Describe the steps involved in a glucose tolerance test.

17-11 Identify the clinical significance of the Hb A1c test.

17-12 Explain how the results for capillary whole blood glucose testing may compare to those tested on a plasma sample.

17-13 Describe common maintenance and quality control issues that may need to be addressed in home and laboratory glucose testing methods.

17-14 Perform CLIA-waived glucose testing.

17-15 Perform a CLIA-waived Hb A1c test.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.13. Perform chemistry testing

ABHES Standards

- 10. Medical Laboratory Procedures, b. CLIA-waived tests
- Graduates: b. Perform selected CLIA-waived tests that assist with diagnosis and treatment, #3 Chemistry Testing

KEY TERMS

Albumin
Autoantibodies
Autoimmune response
Body mass index (BMI)
Carbohydrates
Diabetes
Diabetes insipidus
Diabetes mellitus
Diabetic ketoacidosis
Diabetic neuropathy
Etiology
Fasting blood sugar (FBS)
Fasting plasma glucose (FPG)
Gestational diabetes
Glucagon
Glucose
Glucose challenge test
Glucosuria
Glycated hemoglobin
Glycemic control
Glycogen
Glycogenolysis
Glycolysis
Glycosylated hemoglobin
Hb A1c
Hyperglycemia
Hypoglycemia
Impaired fasting glucose (IFG)
Impaired glucose tolerance (IGT)
Insulin
Insulin dependent
Insulin resistance
Islets of Langerhans
Ketones
Ketonuria
Macrosomia
Metabolic syndrome
Microalbumin
Microalbuminuria
Oral glucose tolerance test (OGTT)
Polydipsia
Polyphagia
Polyuria
Postprandial
Prediabetes
Type 1 diabetes
Type 2 diabetes

History has provided us with many detailed observations of patients with **diabetes.** The ancient physicians documented the symptoms related to diabetes, such as frequent excessive urination and weight loss; however, they were not able to effectively treat the condition. In the first century AD, a physician in ancient Greece named this malady *diabetes,* based on the Greek word for "siphon," as it seemed to him that the liquid taken in by the body was just "siphoned" through and came directly out as urine. Other ancient civilizations provide examples of urine from ill patients that attracted ants because of the sweetness of the fluid. In the 1700s, physicians discovered that the urine and blood from patients exhibiting diabetic symptoms tasted sweet, like sugar, and this became an accepted means of diagnosing diabetes in symptomatic patients.

The treatment of diabetes remained ineffective because physicians could not determine the cause or source of the dysfunction responsible for the symptoms. Some thought there was a problem with digestion in these patients, whereas others thought it must be a disease of the kidneys because there was so much urine produced. It was not until the late 19th century that scientists discovered evidence that the pancreas was involved in the process. In 1889, two physicians who were studying fat digestion and utilization discovered that the removal of the pancreas caused the animals in the experiment to exhibit the symptoms of diabetes. In 1922, Frederick Banting and Charles Best experimented with extracts from the pancreas to isolate **insulin,** an enzyme that is necessary for the proper utilization of glucose by the cells of the body. Insulin became available in the 1920s as injected medication to treat diabetes, with immediate lifesaving results for those afflicted.

In 1935, scientist Roger Himsworth presented evidence that diabetes was actually two separate diseases: "insulin sensitive" (today's type 1 diabetes) and "insulin insensitive" (now known as type 2 diabetes). This discovery provided the opportunity for a deeper understanding of the disease and its treatment, as well as the development of an oral medication for "insulin-insensitive" type of diabetes, which became available in the 1950s.

Even though the cause and types of diabetes are now understood and more treatment options are available, diabetes is still a very serious, widespread, and expensive health problem. According to statistics from the Centers for Disease Control and Prevention (CDC), approximately 8% of the population of the United States has diabetes, and in 2007 it was the sixth leading cause of death. Common complications of diabetes include cardiovascular disease, increased stroke risk, poor healing, hypertension, blindness, kidney disease, nervous system dysfunction, amputations, and dental disease. Diabetes can also contribute to complications of pregnancy and causes increased susceptibility to other illnesses. Because of the severity of this disease and the number of those afflicted, **glucose** testing has become a very common procedure performed in physician office laboratories and reference laboratories. Health-care providers and patients must work closely together to manage the disease and provide appropriate treatment to avoid the onset of life-threatening complications.

GLUCOSE UTILIZATION AND CONTROL MECHANISMS

Glucose is a type of simple sugar that is needed as an energy source by all living things. When **carbohydrates** are digested, glucose enters into the bloodstream as a byproduct. As the blood glucose levels rise, the body reacts in several ways:

- Some of the glucose goes directly to the brain tissues, where it is needed for normal function. Cells of the brain do not require insulin to take in glucose for energy.
- The increased levels of glucose in the blood trigger the pancreas to release insulin, which is produced by clusters of specialized beta cells in the pancreatic **islets of Langerhans.** Insulin is required for glucose to enter most of the cells of the body so that it can be used as an energy source. The cells of the body have a receptor that interacts with the insulin molecule. This interaction allows glucose to pass through the cell membrane. Insulin may be considered as a "key" to open the doors on the cells to allow the glucose to enter. The cells accept as much glucose as needed for normal function.
- Because a meal may provide more glucose than is needed immediately, insulin triggers the body to store the excess glucose in the muscles and liver as **glycogen,** which is essentially a long string of glucose molecules. When the demand for glucose increases past what is in the bloodstream, this glycogen is broken down and released for use as glucose. The increased glucose demand may also result in the transmission of a message to the brain encouraging the body to take in food. Additional unneeded glucose may be stored as fat.

Test Your Knowledge 17-1

What is glycogen? (Outcome 17-1)

Once the glucose has entered the cells of the body, the blood glucose levels begin to return to normal. The goal of the body is to keep a consistent level of glucose in the bloodstream at all times while providing the necessary energy to the cells. Glucose continues to be used by the cells constantly, and with increased activity the glucose demands are heightened. The increased demand causes the glucose levels in the bloodstream to decrease below the normal range as the glucose molecules move into the cells. When this occurs, another cell type (alpha cells) in the pancreatic islets of Langerhans secrete a hormone called **glucagon,** which counteracts the effects of insulin. (See Fig. 17-1 for more details about the pancreas and the islets of Langerhans.) Glucagon causes the pancreas to reduce insulin production, and signals the liver, muscles, and adipose tissues of the body to release some of the glucose that is stored there into the bloodstream to increase blood glucose levels. This process of glycogen breakdown is called **glycogenolysis**. Figure 17-2 demonstrates the balance between glucose, insulin, and glucagon.

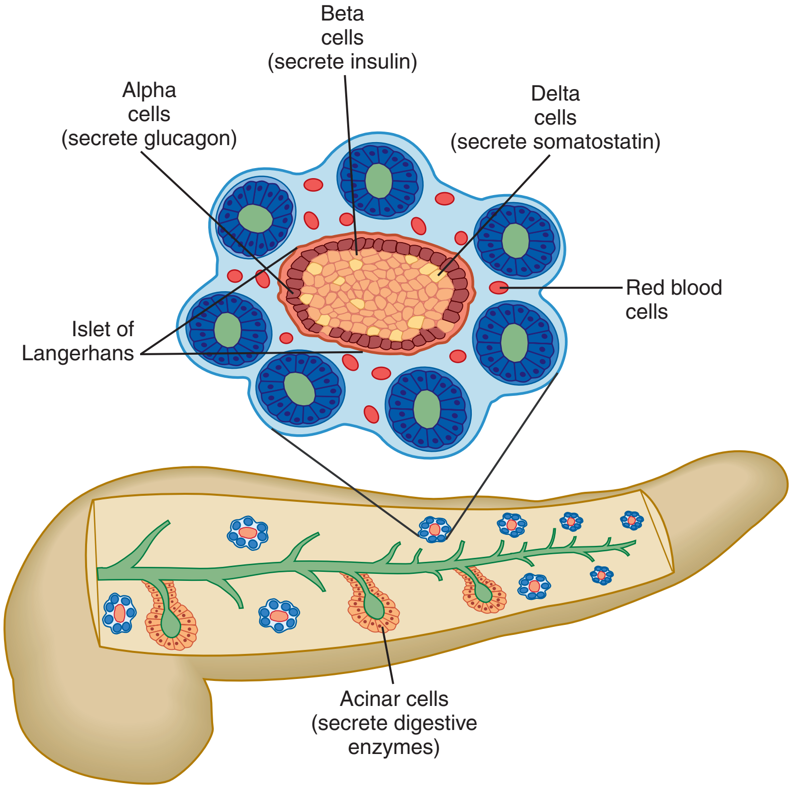
Beta cells (secrete insulin)
Alpha cells (secrete glucagon)
Delta cells (secrete somatostatin)
Red blood cells
Islet of Langerhans
Acinar cells (secrete digestive enzymes)

Figure 17-1 Pancreas and islets of Langerhans; note the alpha and beta cells.

The intricate balance of the blood sugar levels, insulin secretion, and glucagon secretion is essential for the body to function normally. Insulin is required for most of the cells in the body to take in glucose to be used for energy, and insulin also triggers the formation of glycogen and triglycerides from the excess blood glucose so that the body has a backup source of energy between meals or during times of increased need. In addition, the secretion of insulin stimulates the liver and muscle cells to absorb amino acids introduced into the bloodstream during digestion and use these to create proteins. Insulin also stimulates the cells to take in fatty acids from the bloodstream so that there is not an excessive buildup.

Test Your Knowledge 17-2

Which of these processes is triggered by the secretion of insulin?

a. Formation of glycogen
b. Increase of blood glucose
c. Formation of glucose
d. Secretion of glucose (Outcome 17-3)

PATHOPHYSIOLOGY OF GLUCOSE METABOLISM

Prediabetes

According to the American Diabetes Association, there are approximately 57 million Americans that have **prediabetes,** which is defined as a condition in which the blood glucose level is elevated, but not high enough to be indicative of a diagnosis for type 1 or type 2 diabetes. Those who have prediabetes may already be experiencing some of the adverse effects of **hyperglycemia** (elevated blood glucose) and are at a greatly increased risk of developing diabetes within the next few years.

Those with prediabetes have some degree of insulin resistance, usually associated with obesity, a sedentary lifestyle, and poor eating habits. Patients with increased abdominal fat and an elevated **body mass index (BMI)** (a measurement of weight in relationship to height) are at an increased risk, as well as those with a history of type 2 diabetes in their immediate family. In addition, diabetes is more common in certain ethnic groups. Latinos,

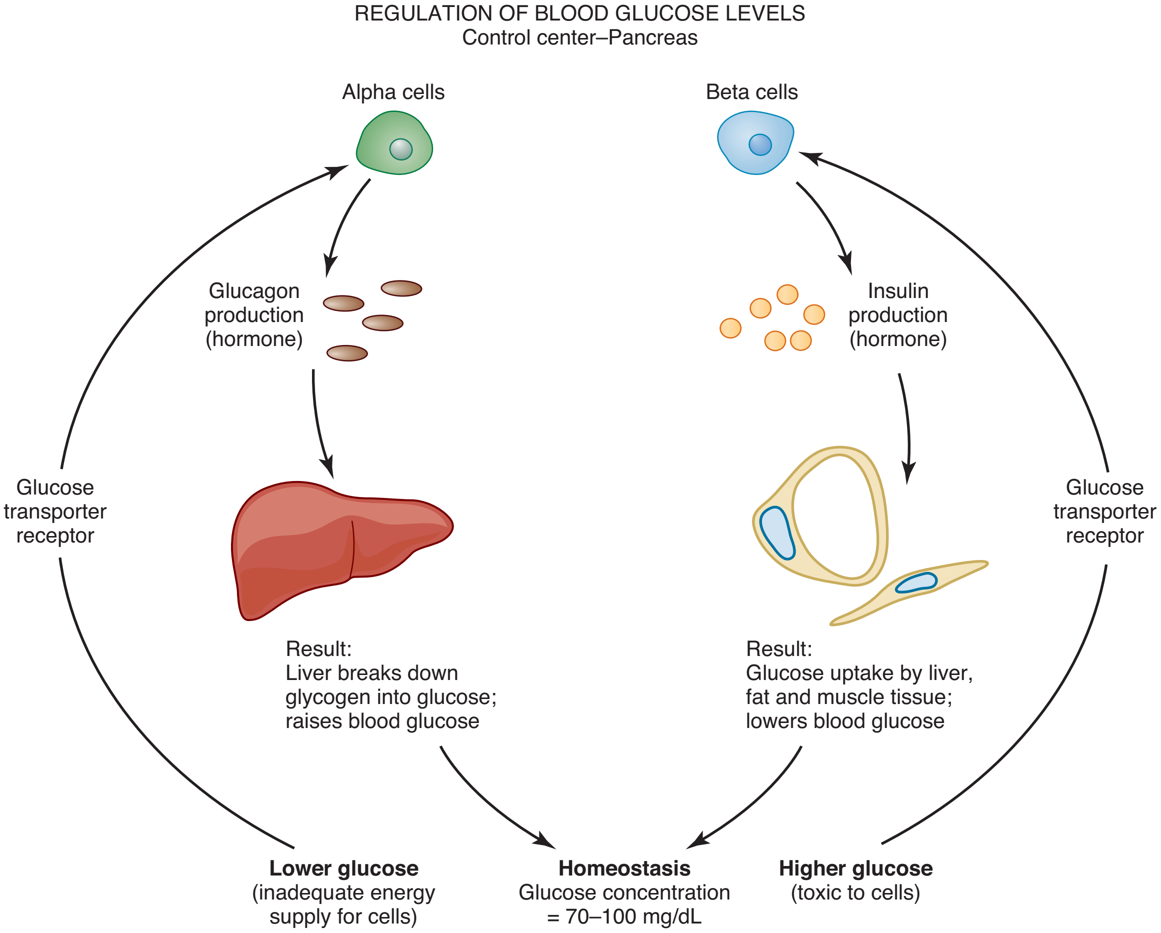
REGULATION OF BLOOD GLUCOSE LEVELS
Control center–Pancreas
Alpha cells
Beta cells
Glucagon production (hormone)
Insulin production (hormone)
Glucose transporter receptor
Glucose transporter receptor
Result:
Liver breaks down glycogen into glucose; raises blood glucose
Result:
Glucose uptake by liver, fat and muscle tissue; lowers blood glucose
Lower glucose
(inadequate energy supply for cells)
Homeostasis
Glucose concentration = 70–100 mg/dL
Higher glucose
(toxic to cells)

Figure 17-2 Representation of the balance between blood glucose, insulin, and glucagon. When blood glucose is increased, the body produces insulin to allow the glucose to enter the cells. This decreases the blood glucose levels, which triggers glucagon to be released if the level drops too low. The glucagon then decreases the absorption of glucose into the cells, and increases the breakdown of glycogen, if necessary.

African Americans, Native Americans, Asian Americans, and Pacific Islanders have an increased presence of diabetes in their populations. The elderly also have a disproportionate amount of diabetes present. Because these groups have more diabetes in their midst, prediabetes is also more likely.

Prediabetes is detected by testing the blood glucose level. The desirable **fasting plasma glucose (FPG)** result tested after approximately 12 hours of fasting is below 100 mg/dL. Prediabetes is indicated by an FPG of 100 to 125 mg/dL, and patients with diabetes usually demonstrate an FPG of 126 mg/dL or above. Prediabetics with an FPG of 100 to 125 mg/dL are said to have **impaired fasting glucose (IFG).** The prediabetes diagnosis may also be based on an abnormal **oral glucose tolerance test (OGTT)** result. The result at 2 hours during this procedure for prediabetics will be between 140 and 200 mg/dL, and for diabetics it will be over 200 mg/dL. (The OGTT procedure is detailed later in this chapter.) Prediabetics with glucose levels at 2 hours between 140 and 200 mg/dL are said to have **impaired glucose tolerance (IGT).**

Those diagnosed with prediabetes do not always proceed to develop diabetes. Studies have shown that if patients with prediabetes make a concentrated effort to increase physical activity and reduce their weight by 5%

to 10%, they may not develop diabetes, or at the very least, they may delay the onset of diabetes for a few years. These lifestyle changes will also stop any undesirable effects of the hyperglycemia on the cells of the body. The American Diabetes Association strongly recommends counseling and follow-up for this group to assist with lifestyle changes.

Diabetes

The word *diabetes* is used to refer to a group of disorders that all exhibit **hyperglycemia**, or elevated blood glucose. These may also be known as different types of **diabetes mellitus,** or "sweet" diabetes, because of the glucose content in the urine of those diagnosed. There are actually three different types of diabetes associated with hyperglycemia, which differ by their **etiology** (cause) and by their treatment. These include **type 1 diabetes,** which previously was known as insulin-dependent diabetes or juvenile-onset diabetes. **Type 2 diabetes** was previously known as non-insulin-dependent diabetes. **Gestational diabetes** is a disorder of glucose metabolism that only affects those who are pregnant. A diagnosis of type 1 or type 2 diabetes is usually only assigned to a patient after confirmation of the abnormal blood levels.

Type 1 Diabetes

Patients with type 1 diabetes have a lack of insulin. The beta cells of the islets of Langerhans in the pancreas have been destroyed and are not capable of creating enough insulin to keep the body healthy. Type 1 diabetes is present in approximately 5% to 10% of all diabetic patients, and presents most often in children or adolescents. Adults may also be afflicted, although this is not as common as it is in those under the age of 20. The destruction of the beta cells responsible for insulin production is most commonly associated with an **autoimmune response,** in which the patient's body develops **autoantibodies** that attack the cells in the islets of Langerhans. There seems to be a genetic predisposition to this condition, although it appears that there is usually some sort of "trigger" that brings on the autoantibody production. Viral infections and environmental factors have been linked to the progression of the disease.

Type 1 diabetes may develop over a few weeks; the onset of the disease is often quite rapid as compared to type 2 diabetes, which has a gradual progression. Type 1 diabetes patients commonly exhibit fatigue, increased urination and increased thirst, nausea and vomiting, and weight loss in spite of an increased appetite. During the initial onset of symptoms, the patient may have periods of hyperglycemia as well as **hypoglycemia** (low blood sugar). If the diagnosis and treatment are not established early in the disease process, the body will react to the shortage of insulin by using fatty acids as an energy source instead of depending on glucose. The use of fatty acids may cause a buildup of **ketones** in the bloodstream. Ketones are a by-product of fat metabolism, and because they are acidic in nature, they will change the pH of the body over a period of time. Initial symptoms of **diabetic ketoacidosis** include abdominal pain, nausea, and vomiting. As the condition progresses, the patient may have increased shallow respirations, which is the body's way of attempting to change the acidic blood pH. The patient may have difficulty thinking and communicating clearly, and may eventually become comatose. It is at this point that the diagnosis of type 1 diabetes is often assigned, as the patient is admitted to the emergency room or hospital once he or she becomes critically ill.

Treatment of type 1 diabetes requires frequent insulin injections. Insulin cannot be administered orally, as it is not effective once it has been digested. There are many different types of insulin; some act quickly and stay in the circulation for a short period of time, and others are designed to be longer lasting. Patients who are **insulin dependent** monitor their blood glucose level several times daily and adjust their insulin intake accordingly.

Type 2 Diabetes

Type 2 diabetes is characterized by a lack of insulin activity on the cells of the body. This may be due to **insulin resistance,** in which the cells have a diminished ability to interact with insulin as they should. Because the insulin no longer acts on the cells of the body to allow the glucose to pass through into the cells, the blood glucose levels remain elevated. Type 2 diabetes may also be the result of a decreased production of insulin by the beta cells of the pancreas. In situations in which the patient exhibits insulin resistance, the body will try to produce more insulin to overcome the resistance, but after a period of time the demand is excessive and the production cannot keep up with the needs of the body. Those who have type 2 diabetes often have high levels of insulin in the bloodstream as well as high levels of glucose. In addition to the high levels of blood glucose that result because of ineffective cellular transport, the body also tries to accommodate the imbalance by breaking down additional glycogen in the liver, which adds additional glucose to the bloodstream and complicates the situation further.

Type 2 diabetes is the most common form of the disease, affecting approximately 95% of all diabetics. The CDC has declared type 2 diabetes to be at an epidemic level, as the numbers of those diagnosed increase every year. The etiology (cause) of type 2 diabetes is complicated. There is a direct correlation with obesity and individuals who have been diagnosed with **metabolic syndrome** (a group of risk factors that occur together and increase the risk of coronary heart disease, stroke, and type 2 diabetes), as well as a family history of type 2 diabetes and a sedentary lifestyle. Women with a history of polycystic ovary syndrome or gestational diabetes also have an increased chance of developing type 2 diabetes. Those older than age 40 are at a higher risk than those who are younger, although there has been a sharp increase in the number of obese children diagnosed with type 2 diabetes in the past few years. Certain ethnic groups are at a higher risk, such as Native Americans and those of Asian or African descent.

Diagnosis of type 2 diabetes may be delayed for years after the first onset of symptoms as this type of diabetes develops slowly. As with type 1 diabetes, many patients diagnosed with type 2 diabetes demonstrate increased urination and increased thirst. Some other common symptoms include unexplained weight loss or weight gain, flu-like symptoms with fatigue and nausea accompanied by a loss of appetite, changes in vision, and poor healing. Patients may also find that they are more susceptible to other illnesses, such as the common cold. Having frequent urinary or yeast infections may also be an indicator of type 2 diabetes. Another clinical indicator may be a change in oral health, because diabetes may cause inflammation that leads to infection in the gums. In addition, some patients experience tingling or loss of sensation in the fingertips or toes, which is caused by damaged nerve endings from the high levels of glucose in the body.

Type 2 diabetes is first treated with lifestyle changes. Most patients diagnosed with this type of diabetes are overweight, so the first goal is to lose approximately 8% to 10% of the total weight. Diabetics are also encouraged to exercise for at least 150 minutes per week to improve their health. Oral medications that enhance the ability of the body to react to insulin or stimulate the beta cells of the pancreas to produce more insulin may also be used. If these measures are ineffective in accomplishing **glycemic control** (appropriate blood glucose levels), the patient with type 2 diabetes may be treated with insulin injections.

Untreated or uncontrolled type 1 or type 2 diabetes may lead to hypertension and other serious cardiovascular diseases, as well as **diabetic neuropathy**; nervous system damage; eye and kidney tissue damage; poor healing; foot and skin complications; and gastroparesis, a disorder in which the stomach takes too long to empty after eating. Those with diabetes are at an increased risk of developing depression. Patients with type 1 diabetes may also have an increased risk of developing celiac disease (a chronic digestive disorder caused by the inability to metabolize gluten) or a painful musculoskeletal condition known as "frozen shoulder," in which the individual loses movement of the shoulder for a period of time.

Test Your Knowledge 17-3

List one way that type 1 and type 2 diabetes are similar. (Outcome 17-4)

Test Your Knowledge 17-4

List two potential consequences of uncontrolled or untreated diabetes. (Outcome 17-6)

Gestational Diabetes

Gestational diabetes (sometimes abbreviated as GDM) is similar to type 2 diabetes because it is a form of glucose intolerance rather than a reduction in insulin production. Approximately 4% of all pregnant women develop gestational diabetes. It is more frequent among Hispanic Americans, African Americans, Native Americans, Asian Americans, indigenous Australians, and Pacific Islanders. Obesity, previous delivery of an infant weighing more than 9 pounds, and a positive family history of diabetes also increase the risk of developing gestational diabetes. Women who have gestational diabetes will usually demonstrate it in subsequent pregnancies, and are at an increased risk of developing type 2 diabetes later in life.

Although the exact cause of gestational diabetes is not clear, it is thought that the insulin resistance present in gestational diabetes is demonstrated as a result of the additional hormones present in the body during pregnancy. The placenta produces high levels of hormones (such as cortisol and estrogen) to support and sustain the pregnancy. These hormones block or reduce the action of insulin on the cells of the mother's body, which causes her blood glucose levels to become elevated.

The high levels of hormones are not usually a problem until late in the pregnancy, as the mother's pancreas normally responds to the additional insulin demands early in the pregnancy. As the pregnancy continues, the additional maternal insulin production cannot overcome the effects of the hormones produced by the placenta, and the mother will exhibit sustained elevated blood glucose levels.

Gestational diabetes is a serious health risk because of the additional glucose in the mother's bloodstream. Glucose passes through the placenta to the fetus, causing a condition known as **macrosomia.** *Macrosomia* means "fat baby." In this condition, the additional glucose passed on to the baby from the mother causes the baby's body to produce excessive fat cells in response to the additional unneeded glucose. The baby may become excessively large, which can cause complications during delivery. Many babies born to mothers with gestational diabetes must be delivered using a cesarean delivery, as their excessive size does not allow a vaginal birth. If a vaginal birth is attempted, these babies are at a high risk of shoulder damage during the delivery. Studies have shown that babies born to mothers with gestational diabetes have increased risk of becoming obese as children, as well as an increased risk of developing type 2 diabetes later in life.

In addition to the risks from excessive size, infants born to mothers with gestational diabetes have an increased risk of breathing problems, as well as extreme hypoglycemia (low blood glucose levels) after birth. During the pregnancy, the infant has been producing enough insulin to process the glucose created by his or her own body, as well as the additional glucose supplied by the circulation of the mother. When the additional maternal glucose supply is discontinued at birth, there is too much insulin in the infant's bloodstream, resulting in low levels of blood glucose. Without careful monitoring, this situation can become life threatening. These babies may also be at an increased risk for other chemical imbalances in the first few days of life.

Test Your Knowledge 17-5

Is gestational diabetes more like type 1 or type 2 diabetes? Why? (Outcome 17-5)

Test Your Knowledge 17-6

How is macrosomia related to gestational diabetes? (Outcome 17-7)

POINT OF INTEREST 17-1

Another type of diabetes

There is another disease that shares the name diabetes, but it is not associated with hyperglycemia. **Diabetes insipidus** is an endocrine disease characterized by a lack of the hormone vasopressin, which acts on the kidneys to reduce urinary output and allow for concentration of the urine. Vasopressin is produced by the pituitary gland, and those with this disease either do not produce enough vasopressin, or they have developed a resistance to the actions of vasopressin to concentrate urine. A patient with diabetes insipidus produces excessive amounts of urine each day, and often sees the physician because of symptoms associated with bed-wetting, dehydration, and/or electrolyte imbalances. Diabetes insipidus has very little in common with the other forms of diabetes discussed in this text, except for the initial common symptoms of increased urination and increased thirst.

TYPES OF GLUCOSE TESTS PERFORMED

The role of the laboratory in diabetes care is to assist the health-care provider in the initial diagnosis, differentiation about the type of diabetes present, and monitoring the progress of the treatment once it has started for the patient. There are several types of glucose test procedures performed that provide necessary information to the health-care provider.

Fasting Blood Glucose Test

The glucose level performed in a laboratory on a fasting blood specimen is known as a fasting plasma glucose (FPG). The term **fasting blood sugar (FBS)** may also be used. Most laboratories perform blood glucose levels on plasma or serum, rather than on whole blood. To prepare for this test, the patient needs to fast for 12 hours and should not smoke, drink anything but water, or take medication before the test. (Sometimes the medication is a *must;* this should be cleared with the physician before the specimen is drawn.) The World Health Organization recommends the use of plasma obtained from a venipuncture for this test. If the test cannot be performed within an hour of the blood draw, a gray top tube should be used for the specimen. The additives in the gray top tube reduce **glycolysis** (utilization of the glucose in the specimen by the living cells present) for up to 24 hours at room temperature. The desirable fasting plasma or serum glucose level is 70 to 100 mg/dL

Patients with FPG of 100 to 125 mg/dL have impaired fasting glucose, and are prediabetic. A result equal to or above 125 mg/dL is indicative of diabetes.

Random Glucose Test

A random blood glucose test is one that is taken at any time of the day without regard for the time of the last meal. A patient who is symptomatic for diabetes, perhaps exhibiting **polyuria,** (excessive urination) **polyphagia,** (excessive hunger), or **polydipsia** (excessive thirst) and unexplained weight loss may have a random glucose test ordered. According to the criteria established by the American Diabetes Association, if the random glucose result for a symptomatic patient is equal to or greater than 200 mg/dL, then the patient is classified as a diabetic.

Test Your Knowledge 17-7

Which of these best describes a random glucose?

a. A glucose level drawn 1 hour after eating
b. A glucose level drawn after 12 hours of fasting
c. A glucose level drawn at any time of the day without any previous preparation
d. A glucose level drawn just before eating

(Outcome 17-8)

Postprandial Glucose Test

Postprandial refers to something done after eating or after mealtime. Glucose testing that is performed postprandial may mean that the blood specimen was collected a specific amount of time after a meal, or it may mean that the specimen was drawn at a certain interval after ingestion of a glucose-rich beverage. Diabetics who are striving to achieve close glycemic control will test their blood glucose levels at home 1 or 2 hours postprandial to see how their body has handled the meal. Most diabetic educational materials provide target ranges for 2-hour postprandial blood glucose levels. Ideally, the blood glucose should not go above 140 mg/dL for nondiabetic patients; for diabetics the level should remain below 180 mg/dL 2 hours after eating.

Initial screening procedures for gestational diabetes may also use postprandial glucose testing. This may be called a **glucose challenge test.** Patients do not need to fast before this procedure. When they arrive at the laboratory or physician office, they are given a drink that contains 50 g of glucose. This drink must be ingested within 5 minutes for the test to be valid. A blood specimen is drawn exactly 1 hour after ingestion of the glucose solution. Patients with gestational diabetes will demonstrate a postprandial blood glucose level at 1 hour that is greater than or equal to 140 mg/dL. If this glucose challenge result is elevated, then the health-care provider will usually order a 3-hour oral glucose tolerance test.

Test Your Knowledge 17-8

True or False: A laboratory postprandial glucose test is drawn at a specific time interval after ingestion of a drink high in glucose or ingestion of a meal.

(Outcome 17-9)

Oral Glucose Tolerance Testing Procedures

According to the National Institutes of Health, the oral glucose tolerance test is more sensitive to the detection of prediabetes than the fasting glucose test level. The OGTT is used primarily for the diagnosis of diabetes mellitus. This test may be performed on a patient with an abnormal FPG, assuming that the result of the FPG was less than 200 mg/dL. If the fasting result was equal to or more than 200 mg/dL, an oral glucose tolerance test is not indicated because the high FPG level is indicative of diabetes. This test may also be ordered for patients who are being evaluated for gestational diabetes, or those who had an abnormal 1-hour postprandial glucose result.

- A patient preparing for an oral glucose tolerance test should have a carbohydrate-rich diet for at least 3 days prior to the scheduled test. Diets that restrict carbohydrate intake may produce erroneous results. The patient must fast for 12 hours prior to the test. During this time, no food or liquid other than water is to be consumed. The patient should also avoid smoking and strenuous exercise.
- Upon arrival at the laboratory or the clinic, the patient will have a fasting plasma glucose specimen drawn. This specimen should be analyzed for glucose content before continuing with the process; if the fasting plasma glucose results are elevated above 126 mg/dL, the test will not continue. Also, some laboratories will collect a urine specimen with each blood draw; this helps to provide more information about the glucose that may be “spilled” into the urine specimen. Usually a blood glucose of 170 mg/dL or greater will cause glucose to be present in the urine specimen, as the body cannot handle these high levels.
- After the fasting specimen has been drawn, the patient will be given a solution with 75 or 100 g of glucose dissolved in water. (For a patient who is not pregnant,

75 g are used; when testing for gestational diabetes, 100 g are used.)

- Blood specimens (and urine if that is the policy of the laboratory) will be collected at least three more times during the procedure: 1 hour, 2 hours, and 3 hours after the initial dose of glucose drink has been ingested. Some laboratories will also collect a blood specimen half an hour after ingestion to provide more information about how the body is utilizing the glucose in the specimen.
- Occasionally, patients will feel nauseous, tired, cold, and perspire while undergoing this test. It is important that patients are kept calm, and that they stay in an area where they can be monitored while waiting. Sometimes a patient will vomit shortly after ingestion of the glucose solution; if this occurs early in the process (prior to 1 hour after drinking the sweetened drink), then the procedure should be discontinued and rescheduled, as the results may be inaccurate. Also, if a patient loses consciousness or exhibits symptoms of slurred speech and impaired thought processes, the test should be discontinued and the physician or emergency personnel should be notified immediately. These could be signs of hypoglycemia that require immediate treatment.

The interpretation of the test results is different when screening for diabetes mellitus than it is for gestational diabetes. For gestational diabetes screening with a 100-g glucose solution, if there are two abnormal results in the various timed blood draws, the patient is diagnosed with gestational diabetes, whether or not she is symptomatic. According to the American Diabetes Association, these results are considered abnormal for a pregnant woman:

- Fasting plasma glucose: greater than or equal to 95 mg/dL
- 1-hour plasma glucose: greater than or equal to 180 mg/dL
- 2-hour plasma glucose: greater than or equal to 155 mg/dL
- 3-hour plasma glucose: greater than or equal to 140 mg/dL

For suspected diabetes mellitus in patients who are not pregnant, a 75-g glucose solution is used. The American Diabetes Association has established these parameters for a diagnosis of diabetes in nonpregnant individuals:

- Fasting plasma glucose: greater than or equal to 126 mg/dL
- 1-hour plasma glucose: greater than or equal to 200 mg/dL
- 2-hour plasma glucose: greater than or equal to 200 mg/dL
- 3-hour plasma glucose: greater than or equal to 200 mg/dL (the 3-hour sample is not always included in the procedure for non-gestational-diabetes patients)

Test Your Knowledge 17-9

How many times (at a minimum) is the blood drawn for a patient completing a 3-hour glucose tolerance test?

(Outcome 17-10)

Glycosylated Hemoglobin

As presented in Chapter 16, elevated blood glucose levels will eventually change the hemoglobin molecule present in the red cells of the body by irreversibly binding glucose to the hemoglobin A subunit. This compound is known as **glycosylated** or **glycated hemoglobin,** and is abbreviated as **Hb A1c.** The typical red blood cell survives in the body for 90 to 120 days. Because the changes to the hemoglobin molecule are constantly occurring whenever the blood glucose level becomes elevated (whether the patient is aware of the elevation or not) the Hb A1c level is an excellent way to measure glycemic control over a 2- to 3-month period. The Hb A1c test is not recommended as a diagnostic tool for diabetes, but is an excellent tool for monitoring the management of the disease.

The American Diabetes Association has provided guidance regarding the frequency of this test and the interpretation of the results. A nondiabetic patient will have Hb A1c levels below 6%. For diabetic patients, the goal is to remain below 7%, as it is understood that there may be spikes in the blood glucose that are unavoidable. As part of the management of diabetes, those who have unstable blood sugars should have their Hb A1c tested quarterly; those who appear to have good glycemic control should be checked at least two times per year. In addition, the recommendation includes performance of an annual lipid profile and urine **microalbumin** testing to screen for renal damage. Microalbumin testing is discussed in more detail later in this chapter.

Test Your Knowledge 17-10

The Hb A1c test may be used to monitor the overall glycemic control for a patient for which of these time intervals?

a. 1 month
b. 6 weeks
c. 3 months
d. 6 months

(Outcome 17-11)

TABLE 17-1

Types of diabetes, diagnostic test procedures, and treatment methods

Type of Diabetes	Symptoms	Clinical Presentation	Diagnostic Test Procedures	Treatment
Diabetes insipidus	Excessive amount of urine production; dehydration and electrolyte imbalance as a result of fluid loss	Lack of vasopressin production by pituitary gland or kidney resistance to vasopressin	Medical history, physical examination, urinalysis, magnetic resonance imaging and/or CAT scan of the brain, fluid deprivation testing	Vasopressin (usually a nasal spray) and fluid replacement if necessary
Type 1 diabetes	Hyperglycemia, fatigue, dehydration, nausea, vomiting, weight loss	Autoantibodies destroy pancreatic cells that produce insulin, resulting in reduced or absence of insulin	Medical history, physical examination, abnormal blood glucose levels with verification of fasting plasma glucose above 126 mg/dL or random plasma glucose above 200 mg/dL	Insulin injections
Type 2 diabetes	May be asymptomatic for years; usually develops slowly. Hyperglycemia, fatigue, dehydration, nausea, vomiting, weight loss	Cells of the body develop resistance to insulin; keeps glucose from entering cells so the plasma glucose levels become elevated	Medical history, physical examination, abnormal blood glucose levels with fasting plasma glucose or glucose tolerance test	Lifestyle changes, oral medication; insulin injections may become necessary if oral medications don't accomplish desired glycemic control
Gestational diabetes	Often asymptomatic; may develop symptoms like those of diabetes mellitus	Increased levels of hormones from pregnancy increase cellular resistance to insulin	Abnormal glucose challenge test and/or abnormal glucose tolerance test	Lifestyle changes, oral medications, or insulin injections (if needed for glycemic control) until birth of infant

POINT OF INTEREST 17-2
New Testing procedures for monitoring diabetes

There is a new procedure available for monitoring the glycemic control of diabetes patients. The GlycoMark blood test measures the level of a substance naturally present at a consistent level in the bloodstream. This molecule is known as 1,5-anhydro-D-glucitol (1,5 AG), which is a monosaccharide that is very close in chemical structure to glucose. When the glucose level of a diabetic patient rises above 180 mg/dL, the kidneys are not able to reabsorb the 1,5 AG molecule because the glucose blocks the sites for this molecule to be reabsorbed. Thus, when the glucose level rises (with poor glycemic control) the 1,5 AG concentration decreases below the reference range.

This procedure provides an opportunity to monitor the glycemic control for diabetic patients for the past 1 to 2 weeks. As a comparison, the Hb A1c test provides a measurement of the glycemic control over the past 3 months. Many physicians feel that this new testing method may allow them to treat their newly diagnosed diabetics more efficiently, because the results will be affected by poor glycemic control very quickly.

Blood Ketone Testing

Many home glucose meters are now capable of testing ketones as well as glucose levels in the bloodstream. Elvated levels of ketones in the blood are indicative of fluctuations in the blood glucose levels. Although close attention to diet, exercise, and medication will help to ensure that a diabetic patient maintains appropriate glycemic control, there may be situations that stress the body and allow the diabetes to become out of control, such as the following:

- Patients with preexisting diabetes who become pregnant
- Patients who are ill for several days, especially if they are vomiting or have diarrhea
- Situations of unusually high stress
- Blood glucose levels equal to or greater than 300 mg/dL
- Symptoms of hyperglycemia such as nausea, fatigue, and vomiting that are not associated with another disease process

Some ketone testing procedures will provide qualitative results indicating the presence or absence of ketones in the blood. Other testing methods provide quantitative results, indicating the level of ketones present. Patients should discuss their situation with a health-care provider before testing for the presence of blood ketones.

Capillary Sample Testing and Correlation to Plasma Glucose Levels

The World Health Organization recommends that the plasma glucose level (rather than whole blood testing) is used for diagnosis and monitoring of diabetic patients. However, this is not always possible, even in a laboratory or clinic setting. Some rural areas in the United States do not have laboratory facilities that meet these standards. Outside of the United States it can become even more problematic. These rural or underserved areas often have a glucose meter available, which is designed to perform whole blood glucose testing rather than plasma glucose testing.

Plasma glucose levels are 10% to 15% higher than those of whole blood obtained via a capillary procedure. Because the glucose levels for diagnosis of prediabetes and diabetes are very specific, it may be necessary to calculate the plasma glucose level using a whole blood capillary result. Many of the new glucose meters include this calculation in their results when reported; for instance, instead of the number on the screen indicating the glucose level in the whole blood specimen that was just tested, a calculation will have already been included and the number is actually indicative of the plasma glucose level. If the meter in use does not use this calculation, many of the glucose meter manufacturers' websites will provide more information to correlate whole blood glucose to plasma glucose levels.

This discrepancy between plasma and whole blood glucose is not often explained to patients who are self-monitoring their levels at home. It is important to mention that their results may not compare to the laboratory results because the reference ranges and testing methods are different. The health-care provider and any diabetes educators who are working with the patients need to help them develop parameters to use for self-monitoring from the testing instrument that they are using at home.

Glucose may also be tested using serum samples. This is not recommended for diabetes screening because the glucose level in the specimen can change slightly during the time period required for the blood to clot. If the specimen is centrifuged immediately after clotting, with immediate separation of the serum from the cells, then the glucose result will be very similar to that of a plasma specimen. However, if there is a delay in the processing of the specimen, the serum glucose result will be much lower than a plasma result would have been for that patient.

Test Your Knowledge 17-11

If a whole blood glucose and a plasma glucose were collected on the same patient at the same time, which of the results would be expected to be higher?

(Outcome 17-12)

Urine Testing for Diabetics

Urine testing is not used as a diagnosis tool for diabetes or prediabetes. The threshold for each individual for the amount of glucose tolerated in the bloodstream may be different. Usually anyone who has a blood glucose level above 170 mg/dL will have glucose present in the urine as well, a condition known as **glucosuria.**

The urine of diabetic patients may also be monitored for other chemical substances. The presence of ketones in the urine **(ketonuria)** may be an indicator of poor glycemic control. When the glucose in the bloodstream cannot enter the cells of the body to be used for energy, the body will break down fatty acids as an alternative energy source. This process is natural, and occurs in our body during fasting periods and times when our energy needs are accelerated. Ketones are a by-product of fatty acid metabolism, and if they are present in the urine, it may indicate excessive use of fatty acids for energy. This means that there is not enough insulin (or the cells are resisting the insulin that is present) to allow the glucose to be metabolized as it should by the cells of the body.

Glucose molecules are large, and are not designed to be passed into the urine in measurable amounts. When a patient has high levels of blood glucose over an extended period of time, these large molecules are forced through the filtration system of the nephron, causing damage because of their large size. The damaged glomerulus will now allow other large molecules present in the blood to enter the urine. The presence of these substances in the urine is considered to be abnormal and indicative of damage to the kidney. One of these large molecules is a plasma protein called **albumin.** Approximately 20% to 40% of all diabetics develop renal disease, and the first sign that the kidneys have been damaged may be the presence of **microalbuminuria.** This means that there are small amounts of albumin present in the urine; the levels are just above normal ranges. When the albumin levels in the urine are still low, it may be possible to eliminate further kidney damage by increasing efforts for glycemic control. If the levels increase to macroalbuminuria (large amounts of albumin in the urine), it is a sign that the renal disease has progressed with more severe damage to the filtration system of the kidneys. It is recommended that all diabetics be tested for urine microalbumin levels at least annually.

GLUCOSE TESTING METHODS

There are many different types of instruments used to perform glucose testing. The equipment varies from machines that analyze hundreds of specimens per hour to those that test one specimen at a time as a CLIA-waived test.

Home Glucose Testing Instruments

Blood glucose monitors are the most common self-testing products sold in the world. They are an essential part of diabetes education and monitoring, and allow diabetic patients a degree of freedom and control of their disease that previously was not possible. Blood glucose meters are all similar in function; a small testing strip is inserted into the instrument, a drop of blood is applied, and a result appears on the screen on the front of the instrument within seconds. However, these instruments may vary in size, complexity, cost of operation, and necessary quality control measures. These are some considerations to take into account when choosing an instrument for home glucose monitoring, including the following:

1. What type of diabetes is being monitored? For those who have type 1 diabetes and use an insulin pump, a continuous glucose monitor may be the most appropriate choice of device. These monitors have a sensor that is placed just under the surface of the skin that samples the blood glucose level 24 hours per day. These results are transmitted to a recording device with a memory, or to an insulin pump. If the continuous glucose monitor is paired with an insulin pump, the amount of insulin injected will be automatically matched to the glucose readings. A health-care provider and/or diabetes educator would be able to help the patient decide if this is the best method for a particular patient.
2. A patient should not take the first testing instrument that is offered and assume that they are all alike. Some instruments require a larger specimen volume than do others, and other machines have a code that must be verified whenever a new bottle of test strips are opened. Some patients prefer instruments that don't need a code to be entered, and some are most interested in a small instrument that can take samples from different areas of the body in addition to the fingertips. If possible, patients should sample several types of instruments when they meet with their health-care provider to discuss the need for blood glucose testing. Diabetes educators are another good resource.
3. Instruments may vary by cost of operation. The blood glucose meters themselves are often quite inexpensive; sometimes they are even free. However, the testing strips can be quite expensive. A typical diabetic patient who is self-monitoring will use at least two or three strips per day. Some insurance plans will help to pay only for specific brands of instruments and strips, so it is best to investigate this early in the process.
4. Patients should be certain that they can read the results displayed on the instrument used. For those with impaired vision, meters are available that display the result in larger numbers. Also, the lighting on the screen may vary, which might make certain models less appealing for some patients with specific vision needs.

Quality Control and Common Errors

Many diabetic patients have been self-monitoring for more than a decade. Unfortunately, they may not have been offered the resources that are now available when they started the process, and their home testing methods may not be providing accurate results. It is important to remember that the glucose meters used by patients at home are an essential tool for their health maintenance. Education about quality control methods for the instrument, storage and use of strips, and record-keeping options should be offered.

- Periodically, patients should bring in their home glucose meter and compare plasma glucose levels from the laboratory to those that they obtain at the same time using their own instrument. Although these results may vary as much as 10% to 15% from one another, this process can identify meters that are not testing accurately, or help the clinic staff to identify problems with the way the patient is performing the test on his or her meter while observing the patient.
- Quality control solution may be purchased and tested periodically to assure that the results obtained by the meter are accurate. Desirable ranges for the quality control solution are usually printed on the box or bottle containing the solution, and the patient can immediately verify the accuracy of the result. Explain to patients that if the machine is not testing within these ranges, they need to correct the situation before continuing use of the meter.
- Cleaning and proper storage of the instrument is essential, but often overlooked.
- Reagent strips must be stored as directed by the manufacturer, and need to be discarded if they are discolored or expired. Many patients leave the bottle open between tests, which causes the strips to absorb moisture and provide erroneous results.
- Reviewing the troubleshooting section of the meter instructions with patients can help them to understand error codes and possible courses of action when the instrument is not performing as expected. Much of this information is available online as well.
- Patients who are self-monitoring blood glucose levels must keep careful records in order for their efforts to be meaningful to their care. Educating the patients about different ways to accomplish this goal can be beneficial. Some of the meters available on the market have the capacity to store a great deal of information that may be downloaded to a computer to be printed or transmitted to a health-care provider. Patients may prefer to keep a handwritten log; if so, they should be provided with blank copies for their use.
- Ask patients about the type of lancet they are using, and how they are disposing of the used devices. Many of the newer automatic lancets may be adjusted according to the depth of the puncture, which is beneficial to those who test often. Lancets should never be used more than once, and should not be disposed of in the regular home trash. Some offices allow the patients to bring in their lancets in an appropriate biohazardous sharps container for disposal; others do not. To assist patients with proper disposal methods, find out what the recommendations are in your area for disposal of this type of waste, and provide that information to your diabetic patients.

Test Your Knowledge 17-12

What is one area in which a medical assistant may need to offer education and assistance for diabetics who are performing home monitoring of their glucose levels?

(Outcome 17-13)

Laboratory Glucose Testing and Potential Sources of Error

When glucose testing is performed in a large laboratory on an automated instrument, plasma or serum is tested. There are several different enzymatic methods available for glucose testing with similar reference ranges based on the type of specimen. Fasting samples will have different results from random specimens or those that were drawn as part of a glucose tolerance test.

Glucose testing may also be ordered on cerebrospinal fluid (CSF). The reference range for this test is approximately two-thirds of the blood glucose level for the same patient collected at the same time as the cerebral spinal fluid sample. When the CSF glucose levels are low as compared to the plasma glucose, it may be indicative of meningitis.

Because a specimen drawn for plasma or serum glucose testing in a laboratory must be processed before it is analyzed, there exists opportunities for error in the preanalytical phase. The following are some considerations that must be addressed to avoid potential problems:

- The timing of the blood draw is critical. For instance, if the specimen is to be drawn fasting, it is imperative that the patient is informed and the preparation is verified before the collection occurs.
- Glucose continues to be metabolized by the cells present in the blood specimen after it is added to the collection tube. If the specimen is not processed by centrifugation and tested within 1 hour of collection, the plasma or serum glucose level may not indicate the actual amount of glucose present in the bloodstream of the patient; instead, it will be indicative of the amount left in the tube after the glucose has been utilized by the cells. If there will be more than a 1-hour delay in testing after collection, a tube with sodium fluoride or potassium oxalate additives should be used for the specimen collection. This slows down the utilization of glucose by the cells, and allows the

plasma glucose to remain stable for approximately 24 hours at room temperature.

- Reference ranges for glucose levels are generally stated as plasma levels. However, serum may also be drawn for glucose testing, as long as the serum is separated from the cells within 1 hour. The reference ranges are usually the same for plasma and serum glucose.
- CLIA-waived testing for blood glucose is not designed to be performed with plasma samples. It is important to read the directions for the CLIA-waived methods to ensure that the results will be accurate.

Procedure 17-1: Performance of Hb A1c Test Using Bayer's A1c Now+ System

TASK

Correctly perform an Hb A1c test using Bayer's A1c Now+ testing system.

CONDITIONS

- Gloves
- Laboratory coat
- Hand sanitization supplies
- A1c Now *plus* monitor
- Sample dilution kit pouch
- Test cartridge pouch
- Whole blood from capillary puncture or well-mixed heparinized blood sample
- Gauze pad
- Biohazardous waste container
- Quality control materials

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P. Anatomy and Physiology, #13 Perform chemistry testing

ABHES Standards

- 10. Medical Laboratory Procedures, b. CLIA-waived tests
- Graduates: b. Perform selected CLIA-waived tests that assist with diagnosis and treatment, #3 Chemistry Testing

Procedure	Rationale
1. Greet and then identify patient using at least two unique identifiers.	All patients must be identified properly before collecting samples or performing laboratory testing.
2. Verify test ordered, and explain procedure to patient.	All laboratory test orders should be verified by checking the chart and/or requisition form more than once. A quick explanation of the procedure for the patient will ensure more cooperation.
3. Verify that all samples, reagents, and monitor are at room temperature.	Test results may not be valid if the sample and/or reagents are not at room temperature.
4. Wash hands and apply gloves.	Hands should always be washed between patients and before starting any procedures. Gloves are appropriate personal protective equipment (PPE) for this procedure.
5. Assemble necessary equipment.	Various steps in this procedure require careful adherence to limits in time. Materials that are organized will assist in the process of following the directions properly.

Continued

Procedure 17-1: Performance of Hb A1c Test Using Bayer's A1c Now+ System—cont'd

Procedure	Rationale
6. Prior to performance of test, verify whether a quality control (QC) specimen needs to be tested, and if so, complete that test *before* the patient's test is performed.	Liquid QC should be used to verify the test results at these times: **a.** With a new shipment **b.** Whenever a new lot number of reagents or QC is put into use **c.** New operator; someone who is being trained on the procedure **d.** Problems with storage, instrument, reagents, etc. **e.** To ensure that storage conditions are fine, QC should be performed at least once monthly.
7. Verify that all lot numbers match on the monitor, the sample dilution kit pouch, and the test cartridge pouch.	If the lot numbers do not match, the monitor will not provide a valid test result upon completion of the test.
8. Verify that the kit has not expired.	The Bayer's A1c Now+ kit may be kept at room temperature for 4 months, but then any unused materials must be discarded. If kept refrigerated, the supplies may all be used until the printed date of expiration.
9. Open the sample dilution pouch and remove the dilution device.	The dilution device should not be removed until just before the test is performed.
10. Perform the capillary puncture following appropriate procedure, or mix the whole blood sample to prepare for next step.	Whole blood specimen may be obtained from fingerstick capillary puncture, or a green top heparinized sample may be used as long as it has not been at room temperature for more than 8 hours, or refrigerated for more than 14 days. Lithium heparin or sodium heparin are acceptable sample types.
11. Add a 5-µL blood sample to blood collection device.	It is not necessary to measure the blood added to the collection device; however, take care that the sample area on the collection device is filled, but not overfilled. If it appears to be underfilled, add more sample; if overfilled, wipe away excess. Do not allow excess blood to remain on the outside of the collection device.
12. Plug blood collection device into the sampler body. Push firmly so that there is no gap on insertion.	If a gap is present, the sample will not combine with the dilution solution inside the sampler body, and the results will be invalid.

Procedure	Rationale
13. Shake the device six to eight times to mix the dilution solution with the sample thoroughly.	Solution must be well mixed to break down the blood sample for appropriate testing.
14. Set sample assembly on the tabletop as the test cartridge is prepared.	
15. Tear open the test cartridge pouch and ensure that the code number on the cartridge matches the code number printed on the instrument.	This system is set up so that the monitor has a definitive number of uses available; all the supplies need to be from the same serial number and cartridge code numbers for the results to be valid.
16. Insert the test cartridge into the monitor until it is seated firmly and an audible click is heard.	If the test cartridge is not seated firmly, no test procedure will occur.
17. Verify that the instrument screen display reads WAIT.	The instrument undergoes a series of internal self-checks before the testing process proceeds.
18. When the instrument screen displays SMPL, pick up the sampler device and remove the base piece, exposing the plunger.	The base piece must be removed so that the sample can be added to the testing device. The sample must be added within 2 minutes.
19. Deliver the sample by pressing the plunger gently and firmly into the corresponding sample application area on the cartridge. Remove after 1 second.	The sample application must occur all at once, but excessive force is not necessary. The application device must be removed for the testing process to proceed.
20. The instrument screen will display a countdown from 5 minutes to 0. Do not move the monitor during this time.	The testing process takes 5 minutes to complete. Moving the monitor may cause inaccurate or invalid test results.
21. The results will display as a percentage on the screen. Record results appropriately in computer, on log sheet, and/or in chart if available.	Results must be recorded immediately so that they are reported correctly.
22. Following office protocol, allow the patient to leave the testing area, or consult with the health-care provider.	If results are far outside of the normal range, it may be office policy that the health-care provider speaks with the patient before he or she leaves the office.
23. Dispose of sample dilution cartridge and test cartridge as biohazardous trash, and disinfect work area.	The work area must always be clean and disinfected after each procedure.
24. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
25. Document the test results in patient chart.	All results must be documented in the patient's chart.

Date	
8/30/2014:	*Hb A1c 6.9% ~~~~~~~~~~~~~~~~~~~~~~ Connie Lieseke, CMA (AAMA)*
12:45 p.m.	*~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~*

Procedure 17-2: Performance of Whole Blood Glucose Testing

TASK

Correctly perform a blood glucose test using the HemoCue Glucose 201 Analyzer.

CONDITIONS

- Gloves
- Laboratory coat
- Hand sanitization supplies
- HemoCue Glucose 201 Analyzer
- HemoCue Glucose microcuvettes
- HemoCue Glucose Control Cuvette
- Capillary puncture supplies
- Biohazard sharps container
- Gauze or laboratory wipes
- Biohazardous waste container
- Quality control materials

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P. Anatomy and Physiology, #13 Perform chemistry testing

ABHES Standards

- 10. Medical Laboratory Procedures, b. CLIA-waived tests
- Graduates: b. Perform selected CLIA-waived tests that assist with diagnosis and treatment, #3 Chemistry Testing

Procedure	Rationale
1. Greet and then identify patient using at least two unique identifiers.	All patients must be identified properly before collecting samples or performing laboratory testing.
2. Verify test ordered, and explain procedure to patient.	All laboratory test orders should be verified by checking the chart and/or requisition form more than once. A quick explanation of the procedure for the patient will ensure more cooperation.
3. Wash hands and apply gloves.	Hands should always be washed between patients and before starting any procedures. Gloves are appropriate personal protective equipment (PPE) for this procedure.
4. Assemble necessary equipment, and verify that all reagents are within the posted expiration dates.	Glucose microcuvettes must be stored refrigerated, as well as the glucose control solution, but they must be removed from the refrigerator and placed within reach before the process begins. Materials that are organized will assist in the process of following the directions properly.
5. Turn on the HemoCue instrument. When the LCD display reads READY, the instrument is ready to be used.	The HemoCue instrument undergoes a series of self-checks when it is turned on.
6. Place the HemoCue Control Cuvette on the cuvette holder and push the holder into the instrument.	The control cuvette is designed to check the function of the instrument and detect any interference with the testing method. It should be used at least once per day.

Procedure	Rationale
7. Verify that the values displayed on the instrument screen are within the acceptable ranges provided with the Control Cuvette.	The ranges are specific per cuvette and are designed specifically for each instrument. If the results are out of range, try cleaning the cuvette and testing again. If they continue to be out of range, the instrument cannot be used for patient testing and the manufacturer must be notified.
8. Prior to performance of test, verify whether a quality control (QC) specimen needs to be tested, and if so, complete that test *before* the patient's test is performed.	Liquid QC should be tested following laboratory protocol and manufacturer's recommendations. Examples of when QC should be used to verify the test results include the following: **a.** With a new shipment **b.** Whenever a new lot number of reagents or QC is put into use **c.** New operator; someone who is being trained on the procedure **d.** Problems with storage, instrument, reagents, etc. **e.** To ensure that storage conditions are fine, QC should be performed at least once monthly
9. Perform a capillary puncture, using appropriate technique. Wipe away the first drop of blood.	The first drop of blood should be discarded, as it may be contaminated with interstitial fluids resulting in an erroneous result.
10. Remove a microcuvette from the container. Hold the open end of the microcuvette to the drop of blood obtained from the capillary puncture. Approximately 5 µL of blood will enter the microcuvette.	The blood will enter the microcuvette by capillary action. It is not possible to overfill the microcuvette. Avoid air bubbles as the microvuvette is filled.
11. Wipe away any excess blood on the outside of the microcuvette.	If excess blood is present on the outside of the microcuvette, it may interfere with the testing process and cause erroneous results or an instrument error while processing the specimen.
12. Place the filled microcuvette on the holder and push it into the instrument for processing.	The microcuvette will fit only when placed on the holder in a certain direction.
13. Within a few seconds, the glucose result will appear on the display screen in mg/dL. Record this result appropriately, in the computer, the patient log sheet, and/or the patient's chart.	The result will stay on the screen until the microcuvette holder is pulled from the instrument for the next sample to be added. It is important to record it as soon as it is evident on the screen.
14. Following office protocol, allow the patient to leave the testing area, or consult with the health-care provider.	If results are far outside of the normal range, it may be office policy that the health-care provider speak with the patient before he or she leaves the office.
15. Pull the microcuvette holder back out, remove the microcuvette, and discard in the biohazardous trash.	The microcuvette contains blood, so it must be discarded as a biohazardous substance.

Continued

Procedure 17-2: Performance of Whole Blood Glucose Testing—cont'd

Procedure	Rationale
16. Turn off the instrument, and disinfect the microcuvette holder.	The microcuvette holder can be completely removed from the instrument for cleaning with alcohol or soap and water.
17. Be sure to dispose of the capillary puncture device in a sharps container and any other equipment that may be contaminated with blood in a biohazardous container.	It is vital to always properly dispose of biohazards.
18. Disinfect the work area.	It is important to keep the work area clean and organized.
19. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
20. Document the test results in the patient's chart.	All results must be documented in the patient's chart.

Date	
8/17/2014:	*Fasting blood glucose 85 mg/dL* *Connie Lieseke, CMA (AAMA)*
10:50 a.m.	

SUMMARY

Diabetes is disease that has reached epidemic proportions. It requires early diagnosis and consistent monitoring for successful treatment. Consequences of untreated diabetes include cardiovascular disease, loss of vision, poor healing, diabetic neuropathy, and permanent renal damage. Prediabetes can now be identified, and early intervention has helped to stop the progression of the disease for those who make appropriate changes to their diet and exercise habits. There are various testing procedures that may be used to diagnose prediabetes and diabetes, and to monitor treatment once it is under way. Type 1 diabetes must be treated with insulin injections, and the dosage is directly related to the amount of glucose present in the body at the time. Type 2 diabetes is managed with lifestyle changes and oral medication. Frequent blood glucose monitoring assists with treatment for this type of diabetes as well. Gestational diabetes is developed by pregnant women and may be identified toward the end of the pregnancy. Gestational diabetes that is untreated may lead to pregnancy complications as well as serious health issues for the unborn child. Several different timed glucose tests allow an opportunity to identify and classify the diabetic patient, including fasting plasma glucose, and postprandial glucose testing. Hb A1c testing is also useful to monitor the progress of the disease. Self-management testing performed at home by diabetic patients helps to achieve glycemic control between visits to the health-care provider.

TIME TO REVIEW

1. Glycogenolysis is: Outcome 17-1
   - **a.** The breakdown of glycogen to be used for energy
   - **b.** The formation of glycogen as a means of storing energy for later use
   - **c.** The breakdown of glucose
   - **d.** The breakdown of glucagon
2. What does a BMI tell us about a patient? Outcome 17-1
   - **a.** A calculation comparing the height and weight of a patient
   - **b.** A calculation used to compare glucose and body weight

c. A calculation of Hb A1c
d. None of the above

3. When is the term *impaired fasting glucose* used? Outcome 17-1
   a. When the fasting glucose level is over 200 mg/dL
   b. When the fasting glucose level is between 100 to 125 mg/dL
   c. When the fasting glucose level is below 100 mg/dL
   d. When the fasting glucose level is between 100 to 124 mg/dL

4. Does the release of glucagon raise or lower blood glucose levels? Outcome 17-2
   a. Glucagon raises blood glucose levels
   b. Glucagon decreases blood glucose levels

5. True or False: Type 1 diabetes is treated with insulin injections whereas type 2 diabetes is treated with oral medications and lifestyle changes. Outcome 17-4

6. True or False: A patient who is diagnosed with gestational diabetes had diabetes before she became pregnant. Outcome 17-5

7. Which of these body systems may be affected by diabetes? Outcome 17-6
   a. Cardiovascular
   b. Integumentary
   c. Nervous
   d. Eye
   e. All of the above

8. How long should a patient go without food and drink before a fasting plasma glucose is drawn? Outcome 17-8
   a. 8 hours
   b. 16 hours
   c. 12 hours
   d. 10 hours

9. True or False: A postprandial glucose sample is always drawn after drinking a glucose-rich beverage. Outcome 17-9

10. What is the correct number of times that a patient drinks the glucose solution while completing a 3-hour glucose tolerance test procedure? Outcome 17-10
    a. One
    b. Two
    c. Three
    d. Four

11. True or False: The Hb A1c test is used to establish a diagnosis of diabetes. Outcome 17-11

12. How may the serum glucose level be affected if the sample is drawn into a tube with no additive and the tube is allowed to sit for 3 hours before centrifugation and processing? Outcome 17-13
    a. The glucose level may increase
    b. The glucose level may decrease
    c. There will be no effect on the glucose level
    d. It will be impossible to test the glucose level

Case Study 17-1: Use of the Hb A1c

Rose Grace is a patient who has been treated successfully for type 2 diabetes for many years with oral medication and lifestyle changes. The past few months have been especially stressful for her, and she has not been monitoring her diet, exercise, or blood glucose levels consistently during this time. A week before her quarterly appointment with her physician she starts to watch her diet carefully and catch up with some of her glucose monitoring to prepare for the visit. When she sees the physician, she is pleased to have her random glucose at 145 mg/dL, and she does not share all the activities of the past few months with her health-care provider, as it appears that her blood glucose is fine. However, when the Hb A1c result is available, Rose's physician calls her and tells her that she needs to come in for another visit to discuss additional treatment options.

1. What did the Hb A1c result tell the health-care provider that was not apparent with the blood test performed at the initial visit?

RESOURCES AND SUGGESTED READINGS

"National Diabetes Information Clearinghouse"
Provides in-depth information about the various types of diabetes, diagnosis, and management http://www.diabetes.niddk.nih.gov/

"American Diabetes Association website"
Information about risk factors, managing diabetes, and support for those who have diabetes; website provides information for providers, patients, and family members http://www.diabetes.org

"Executive Summary: Standards of Medical Care in Diabetes—2009"
Diabetes Care 32, no. S1 (January 2009): S6–S12, doi: 10.2337/dc09-S006.

Standards of diabetic medical care developed by worldwide consensus and adopted by the World Health Organization, National Institutes of Health, and the American Diabetes Association. Includes parameters for diagnosis and treatment.

"2007 Diabetes Fact Sheet"
National Center for Chronic Disease Prevention and Health Promotion;
Provides statistics for diabetes morbidity and mortality in 2007
http://www.cdc.gov/diabetes

Chapter 18

Other Select Chemistry Tests

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Lipid Testing
  - Cholesterol
    - *Cholesterol Metabolism Within the Body*
    - *Types of Cholesterol*
    - *Risk Factors, Desired Ranges, and Clinical Interpretation of Abnormal Lipid Results*
    - *Triglycerides*
    - *Metabolic Syndrome*
    - *Reference Ranges and Clinical Interpretation of Abnormal Results*
  - Lipid Panels
- Electrolytes
  - Function of Electrolytes and Consequences of Electrolyte Imbalances
    - *Testing Preparation and Common Testing Methods*
  - Reference Ranges
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

18-1 Define the key terms.

18-2 Identify dietary sources of cholesterol.

18-3 Compare and contrast the different types of lipoproteins tested in a lipid panel.

18-4 Provide the desired range for cholesterol levels.

18-5 Describe the effects of elevated cholesterol levels on the body.

18-6 Explain how the body uses triglycerides.

18-7 Describe the appearance of lipemia in a blood specimen.

18-8 Evaluate the potential health issues that are caused by high levels of triglycerides.

18-9 List the components of common lipid panels.

18-10 Explain the patient preparation necessary before a specimen is collected for a lipid analysis.

18-11 Perform a CLIA-waived cholesterol test.

18-12 Describe how electrolytes are used by the body.

18-13 Summarize the consequences of untreated electrolyte imbalances.

18-14 List the common analytes included in an electrolyte panel.

18-15 Perform a CLIA-waived electrolyte test.

CAAHEP/ABHES STANDARDS

CAAHEP 2008 Standards

I.P.I.13. Perform Chemistry Testing

I.A.I.2. Use language/verbal skills that enable patients' understanding

ABHES Standards

- 10. Medical Laboratory Procedures, b. CLIA-waived tests
- 10. Medical Laboratory Procedures, b. 3) Chemistry Testing

KEY TERMS

- Acidosis
- Alkalosis
- Anions
- Apolipoproteins
- Buffer
- Cations
- Extracellular
- Fibrates
- Hyperkalemia
- Hypernatremia
- Hypokalemia
- Hyponatremia
- Intracellular
- Ions
- Lipemia
- Lipoproteins
- Malabsorption
- Metabolic syndrome
- Niacin
- Statins

Clinical chemistry includes the quantitative chemical analysis of various body fluids, as introduced in Chapters 16 and 17 in this text. Because this area of the laboratory performs so many different testing procedures, it is not possible to cover them all in detail in this book. This chapter provides more details about lipid testing and provides additional information about the role of electrolytes in the body. Cholesterol analysis and some electrolyte testing are commonly performed in physician office laboratories, as well as hospital and reference laboratories.

LIPID TESTING

Lipid testing includes the quantitative analysis of cholesterol levels, differentiation of the different types of lipoproteins that transport cholesterol in the body, and analysis of total triglyceride levels. Heart disease is the most common cause of death in the United States, and because the presence of elevated cholesterol (hypercholesterolemia) or hyperlipidemia is a risk factor for many serious cardiovascular complications, frequent monitoring of lipid levels is very important.

Cholesterol

Cholesterol is a white, soft, waxy substance that is present in all the cells of humans. Although humans use cholesterol to form cell membranes, create hormones, and perform other vital body functions, too much cholesterol is very damaging to the blood vessels within our body. It is recommended by the National Cholesterol Education Program that all adults have their cholesterol levels checked at least once every 5 years. The cholesterol level for any individual is dependent on many factors, including genetics, sex, diet, level of physical activity, social habits, and age.

Cholesterol Metabolism Within the Body

The cholesterol that is needed by our body is manufactured in the liver, but cholesterol is also ingested with animal products, dairy products, and hidden in processed foods. This excess cholesterol is deposited within the blood vessels, causing plaque buildup. The plaque on the interior of the arteries causes them to harden and lose elasticity, a condition known as *atherosclerosis*. The vessels may eventually fill with sticky, hard cholesterol and become occluded (blocked) so that the blood flow is disrupted. The blood that is trapped by the plaque buildup may become clotted. Pieces of the plaque may also break away from the vessel walls and travel to other areas of the body as emboli.

Test Your Knowledge 18-1

The human body produces adequate cholesterol to meet its needs. Additional cholesterol is added as part of our diets. What is one primary source of ingested cholesterol? (Outcome 18-2)

Types of Cholesterol

Cholesterol is a type of lipid that does not dissolve in blood plasma. It must be attached to another molecule for transport throughout the body. The molecules used for cholesterol transport are also created in the liver, and are called **apolipoproteins.** These are special types of protein that are designed to attach to a lipid molecule, becoming a **lipoprotein.** As introduced in Chapter 16, there are three types of lipoproteins present to transport cholesterol in the bloodstream:

- **High-density lipoprotein (HDL):** A carrier molecule that is known as the "good cholesterol" because it transports cholesterol through the vascular system without contributing to the buildup on the vessel walls. In fact, HDL has been shown to help remove some of the cholesterol that has accumulated in the vascular system. A high level of HDL appears to protect the body from heart disease, and a low level (below 40 mg/dL for men and 50 mg/dL for women) increases the risk for heart disease and stroke.
- **Low-density lipoprotein (LDL):** The major lipoprotein cholesterol transport molecule present in the bloodstream. This is often referred to as the "bad cholesterol" because the lipoprotein has a higher fat content than does the HDL and actually sticks to the blood vessel walls and adheres to the waxy buildup, leading to vessel occlusion. This waxy buildup (plaque) creates a unique environment in which blood cells may bind to one another and form a blood clot. When a blood clot or an occlusion occurs in one of the coronary vessels that supplies blood to the heart, a heart attack may result. Blocked blood flow to the brain results in a stroke (also known as a cerebrovascular accident [CVA]). According to the American Heart Association, LDL cholesterol levels above 160 mg/dL greatly increase the chance of heart disease.
- **Very Low-Density Lipoprotein (VLDL):** Present in smaller amounts than the other transport molecules. Elevated VLDL levels are directly related to the formation of plaque from sticky cholesterol deposits, which increases the risk of heart disease and stroke.

Test Your Knowledge 18-2

Which classifications of lipoproteins are considered as contributors to plaque buildup? (Outcome 18-3)

Risk Factors, Desired Ranges, and Clinical Interpretation of Abnormal Lipid Results

Total cholesterol screening assists with identification of those who are at high risk for cardiovascular disease. According to the Centers for Disease Control and Prevention (CDC), approximately 17% of the adults in the United States have total cholesterol levels above the desired range. Cholesterol levels may be elevated in certain diseases, such as diabetes, hypothyroidism, and kidney or renal disease. However, more commonly the elevation of total cholesterol is the result of risk factors such as a sedentary lifestyle, poor diet, cigarette smoking, heredity, and obesity. The presence of two or more risk factors for a specific patient may double the risk of a heart attack. Native Americans and African Americans are at an exceptionally high risk because of genetic factors.

HDL levels may be increased in women because of the natural presence of estrogen in their bodies. Testosterone has been shown to lower HDL levels. Cigarette smoking, sedentary lifestyles, and obesity all contribute to lower HDL levels. It is possible to change the relative levels of LDL and HDL in the body through diet and exercise. The goal of treatment is to decrease total cholesterol, increase HDL levels and decrease LDL levels. Increased exercise, smoking cessation, and weight loss can elevate HDL levels and decrease LDL levels. Early detection, immediate lifestyle changes, and oral medications may have a profound effect on these analytes. LDL levels may be decreased with the use of a class of medications called **statins.**

Test Your Knowledge 18-3

True or False: Elevated HDL levels are desirable in a healthy adult. (Outcome 18-4)

Cholesterol reference ranges are based on fasting specimens. According to the CDC, the recommended cholesterol levels are the following:

- Total cholesterol levels below 200 mg/dL
- LDL cholesterol level below 100 mg/dL
- HDL cholesterol at 40 mg/dL or above
- Triglycerides below 150 mg/dL

Test Your Knowledge 18-4

Sally Seashore has her fasting cholesterol test performed as a part of her yearly physical. Her total cholesterol level was 205 mg/dL. Is this within the desirable range? (Outcome 18-4)

It is not only adults who have elevated cholesterol levels. Children who have parents or grandparents who developed heart disease before they were age 55, or those who have other risk factors such as race and obesity should have their cholesterol levels monitored closely, as it is now clear that the development of plaque buildup in the blood vessels may begin very early in life.

Test Your Knowledge 18-5

Why are elevated LDL levels considered to be a health risk? (Outcome 18-5)

Procedure 18-1: Cholesterol Testing Using the Cholestech LDX System

TASK

Perform CLIA-waived cholesterol testing using the Cholestech LDX System

CONDITIONS

- Gloves
- Laboratory coat
- Hand sanitization supplies
- Cholestech LDX analyzer
- Cholestech thermal label printer
- Liquid controls (level 1 and level 2)
- Test cassettes in foil wrap
- Optics test cassette and holder
- Capillary puncture supplies
- Cholestech capillary tubes and black plungers for collection and application of blood samples
- Biohazardous sharps container
- Gauze or laboratory wipes
- Biohazardous waste container
- Liquid quality control materials (two levels)
- Acrylic safety shield (if using lithium heparin vacuum tubes rather than capillary samples for the testing process)

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I. Anatomy and Physiology, #13. Perform Chemistry Testing **I.A.I.** Anatomy and Physiology, #2. Use language/verbal skills that enable patients' understanding

ABHES Standards

- 10. Medical Laboratory Procedures, b. CLIA-waived tests
- 10. Medical Laboratory Procedures, b. 3) Chemistry Testing

Procedure	Rationale
1. Greet and then identify patient using at least two unique identifiers.	All patients must be identified properly before collecting samples or performing laboratory testing.
2. Verify test ordered, and explain procedure to patient. Verify whether patient has prepared properly.	All laboratory test orders should be verified by checking the chart and/or requisition form more than once. A quick explanation of the procedure for the patient will ensure more cooperation. Cholesterol testing requires the patient to fast for at least 12 hours prior to the blood draw. It is important that the patient understands the meaning of a 12-hour fast to verify whether he or she has followed the necessary preparation instructions.
3. Wash hands and apply gloves.	Hands should always be washed between patients and before starting any procedures. Gloves are appropriate personal protection equipment (PPE) for this procedure, as well as a laboratory coat.
4. Assemble necessary equipment, and verify that all reagents are within the posted expiration dates. Allow all reagents to come to room temperature.	The Cholestech LDX cassettes are stored in the refrigerator, but must come to room temperature before use. It is recommended that they be removed from the refrigerator for at least 10 minutes prior to use. Organization of materials will make it easier to follow the procedure specified by the manufacturer properly.

Procedure	Rationale
5. Plug in the analyzer and allow it to initialize. Verify that the printer has paper attached to the testing instrument correctly.	The analyzer must be allowed to initialize before use.
6. Perform an optics check using an optics check cassette. When testing the optics test cassette, the following procedures are followed: **a.** Press RUN. The analyzer will then perform a self-test, and display a Self-Test OK message on the screen. The drawer will then automatically open, and a message of Load Cassette and Press Run will appear on the screen. **b.** Use the optics check cassette by inserting the cassette into the open drawer of the instrument. Be careful not to touch the dark areas of the cassette; it should be handled only by the edges and by touching the clear parts of the cassette. **c.** Press RUN again. The drawer will shut and the testing process for the optics check will begin.	The optics check cassette is to be used as directed by the manufacturer and laboratory policy to verify that the four optic channels of the instrument are working correctly. This should be used at least once daily before patient samples or quality control samples are tested. An optics test should also be completed after the machine is moved or cleaned.
d. A series of numbers will be displayed on the screen. These are the results of the optics check. The results must all fall between 80 and 105, and they must also be within the ranges provided with purchase of the optics check to be a valid test.	Acceptable ranges will be provided with the cuvette upon purchase. If the results do not fall within the acceptable range, the instrument may not be used for specimen testing until the cause has been identified and the problem has been identified and solved.
e. Verify that the results are as desired before continuing with the testing process. If the results are not within range, the instrument will display an error code (Optics Test Fail), and it will not be possible to process patient samples until a valid optics check is performed.	Troubleshooting tips will be available in the user manual and/or manufacturer insert.
7. Remove the optics check cuvette from the instrument and place it back in the designated container.	The cuvette must be stored in the designated container to keep it from being damaged. Exposure to moisture or damaging the cuvette may affect the performance.
8. Prior to patient testing, verify whether a quality control (QC) specimen needs to be tested, and if so, complete that process before the patient's test is performed. Patient testing cannot be performed unless the quality control values are within the established ranges.	Liquid QC should be tested following laboratory protocol and manufacturer's recommendations. Examples of when QC may be utilized to verify the test results include the following: **a.** With a new shipment **b.** Whenever a new lot number of reagents or QC is put into use **c.** New operator; someone who is being trained on the procedure **d.** Problems with storage, instrument, reagents, etc.
9. Tear open the foil package and set the testing cassette on a level nonabsorbent surface.	The cassette needs to be ready for the application of the sample as soon as the blood collection is complete.

Continued

Procedure 18-1: Cholesterol Testing Using the Cholestech LDX System—cont'd

Procedure	Rationale
10. Perform a capillary puncture, using appropriate technique. Wipe away the first drop of blood.	The first drop of blood should be discarded, as it may be contaminated with interstitial fluids resulting in an erroneous result.
11. Remove a capillary tube and a black plunger from the container. Insert the black plunger into the capillary tube, and then hold the open end of the capillary tube to the drop of blood obtained from the capillary puncture. The capillary tube will fill with sample, and this should be completed within approximately 10 seconds.	The blood will enter the capillary tube without further action if the capillary tube is kept level while the end is immersed in a drop of blood. Avoid air bubbles as the capillary tube is filled.
12. Introduce the blood into the sample well on the cassette. Be careful not to touch the dark areas of the cassette. Use the black plunger to empty the capillary tube into the testing well.	Blood on other areas of the cassette apart from the sample well may interfere with the testing process.
13. Place the filled cassette into the drawer as soon as the sample has been added. If the drawer has closed, press the RUN button again and the drawer will open.	The cassette needs to be processed immediately for accurate results. Do not allow the filled cassette to be moved excessively. It should be kept horizontal as it is placed in the drawer.
14. Within a few seconds, the cholesterol result will appear on the display screen in mg/dL. Record this result on the patient log sheet.	The result will stay on the screen until the RUN button is pushed to open the drawer for the next test sample to be added. It is important to record the result as soon as it is evident on the screen. The result will print if the instrument is attached to a printer.
15. Following office protocol, allow the patient to leave the testing area, or consult with the health-care provider about the test results.	If results are far outside of the normal range, the office policy may be that the health-care provider speaks with the patient before he or she leaves the office.
16. Discard the capillary tube and plunger and contaminated cassette in the biohazardous garbage.	The capillary tube, plunger, and cassette contain blood, so they must be discarded as a biohazardous substance.
17. Turn off the instrument if there are no further tests needed and disinfect the work area. Discard any other waste appropriately.	The work area should always be disinfected after use.
18. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
19. Document the test results in patient chart.	All results must be documented in the patient's chart.

Date	
8/17/2014:	Fasting cholesterol 195 mg/dL ~~~~~~~~~~~~~~~~~~~~~~~~
2:15 p.m.	~~~~~~~~~~~~~~~~~~~~ Connie Lieseke, CMA (AAMA)

Triglycerides

Triglyceride levels are usually included in lipid panels as an additional indicator for evaluating the risk of cardiovascular disease. Triglycerides are the end product of the fats that we ingest. When a person takes in more calories than are needed for energy, some of the excess calories are also turned into triglycerides and stored as fat for use later as an energy source. These are then stored in the liver and the adipose cells of our body, and may also sometimes be stored in the muscle tissue. A patient who often eats more calories than those used by the body may develop elevated blood triglyceride levels.

Test Your Knowledge 18-6

What leads to a buildup of triglycerides in the adipose tissue of the body?

a. Consumption of excess calories
b. Pancreatitis
c. Elevated LDL levels
d. Plaque buildup (Outcome 18-6)

Metabolic Syndrome

Elevated triglyceride levels unaccompanied by other risk factors are not linked as closely to cardiovascular disease as are high cholesterol levels. However, the high triglyceride levels may be part of a group of conditions that are known as **metabolic syndrome.** Metabolic syndrome includes hypertension, hyperglycemia, excessive fat deposits around the waistline, low HDL levels, and high triglyceride levels. Patients who simultaneously exhibit most of these conditions in the metabolic syndrome have an increased risk for cardiovascular disease, as well as a greater potential for developing diabetes. Oral medications may be prescribed to lower triglyceride levels. **Niacin** is commonly used, as is a classification of drugs called **fibrates.** If the triglyceride levels are elevated as well as total cholesterol and LDL levels, a combination of drugs may be used to lower the various lipid levels.

Reference Ranges and Clinical Interpretation of Abnormal Results

Hyperlipidemia is the word used to describe elevated levels of cholesterol and triglycerides in the bloodstream, and as introduced in Chapter 16, patients with hyperlipidemia may have cloudy or "milky" plasma and/or serum apparent after a sample of their blood is spun in a centrifuge. This appearance is known as **lipemia,** and it is caused by suspended fat particles, which can cause interference for many testing procedures. Large reference laboratories may have specialized centrifuges designed to clear the excess fats from the plasma or serum so that testing can continue on these specimens.

Test Your Knowledge 18-7

Lipemia is a suspension of _____________ , visualized in the liquid portion of the blood after centrifugation:

a. Hemolyzed red blood cells
b. Excess fat particles
c. Plaque
d. Milk (Outcome 18-7)

According to the CDC, triglyceride levels should be drawn as fasting samples, because the blood levels become elevated with the ingestion of food. Triglyceride levels may be categorized as the following:

- Normal: below 150 mg/dL
- Borderline high: 150 to 199 mg/dL
- High: 200 to 499 mg/dL
- Very high: equal to or greater than 500 mg/dL

Lipid Panels

A lipid panel commonly includes a total cholesterol level; a triglyceride level; the HDL, LDL, VLDL; and a total cholesterol/HDL ratio. These values provide the health-care provider with valuable information to determine whether the cholesterol levels for a patient indicate a risk factor for coronary artery disease. The lipid panel also provides information that is useful for monitoring patients who are taking medication to treat high cholesterol or triglycerides. Current recommendations from the National Institutes of Health's National Cholesterol Education Program include the performance of a screening fasting (at least 12 hours) lipid panel at least once every 5 years. For patients with elevated LDL levels and/or triglyceride or total cholesterol levels that are too high, lifestyle changes and possible drug treatment should be started immediately. The fasting lipid panel should be rechecked once every 6 weeks until the established goals have been met for a patient, and a fasting lipid panel should continue to be performed every 4 to 6 months while completing treatment.

POINT OF INTEREST 18-1

Types of fats

Fats are essential to homeostasis, but ingestion of too much can be quite harmful. Most Americans take in an excess of fat on a daily basis. According to the Centers for Disease Control and Prevention (CDC), fats should not make up more than 35% of our caloric intake each day. In addition to controlling the amount of fat ingested, we can benefit from monitoring the types of fats we eat, because certain types of fats cause much more harm to our cardiovascular systems than others. Polyunsaturated and monounsaturated fats are healthier for the body than *trans* fats, saturated fats, and cholesterol. The unhealthy fats are "hidden" in many of the processed foods that we eat. To avoid an unhealthy diet, it is important to be educated about the types of fats in our diets:

- ***Trans* fats:** It is required that all products containing *trans* fats have them listed on their nutrition labels. This type of fat actually starts as a liquid; a process called hydrogenation changes the oils into a solid fat. *Trans* fat helps processed foods last longer without spoilage, but the presence of *trans* fats increases LDL cholesterol levels and decreases HDL cholesterol levels. Many processed baked goods (such as pies, cookies, crackers, and cakes) include *trans* fats to extend their shelf life. *Trans* fats may also be present in solid margarine products. Nutritionists recommend soft margarine over solid margarine whenever possible.
- **Saturated fats:** The current recommendation is to restrict our daily caloric intake from saturated fats to less than 10%. This type of fat comes from meat and high-fat dairy products. It is also present in palm and coconut oils, which can be included in a many processed foods. Choosing low-fat products, fat-free milk, and lean meats can help to control the amount of saturated fats ingested. Check labels to reduce the amount of saturated fats (and *trans* fats) whenever possible. Removing the skin from meats before cooking, and skimming off visible fat from soups and stews can also help to reduce dietary intake. Diets that are high in saturated fat are believed to contribute to cardiovascular disease.
- **Cholesterol:** Cholesterol is usually found in the same type of products as those having high levels of saturated fat. Meat, eggs, and dairy products contain high levels of cholesterol. Steps that lower the intake of saturated fat will also help to control cholesterol in the diet because they appear in similar foods. Remember, genetics also plays a role in total blood cholesterol content; many patients will have high levels of cholesterol even with excellent diets and active lifestyles. In this case, patients will usually need to be treated with oral medications to keep their cholesterol level at a reasonable level.
- **Monounsaturated and polyunsaturated fats:** These fats are considered the most "healthy" fats. They come from ingestion of fish, vegetable oils, and nuts. All sources of fat are high in calories, but polyunsaturated and monounsaturated fats do not pose as high a cardiovascular threat as do the other types.

ELECTROLYTES

In Chapter 16, electrolytes are defined as substances that are capable of conducting an electrical charge when dissolved in water. Particles that are positively or negatively charged because of the loss or gain of electrons are called **ions,** and all electrolytes fall within this description**.** Positively charged ions are called **cations,** and those ions with a negative charge are called **anions.** The most common cations in the body are sodium and potassium. Sodium has a high **extracellular** concentration, which means that it has a higher concentration outside the body's cells, in the plasma. Potassium is higher in concentration inside the cells, in the **intracellular** environment**.** The most common anions present in the body are chloride and bicarbonate. Bicarbonate (HCO_3^-) is an ion created when carbon dioxide reacts with water. Most electrolyte panels measure Na^+, K^+, Cl^- and bicarbonate. Figure 18-1 illustrates the distribution of the most common electrolytes inside and outside red blood cells. Bicarbonate may be measured directly or by analysis of total CO_2, as most of the carbon dioxide present in the blood is found as bicarbonate molecules. Additional electrolytes present in the body include magnesium, calcium, phosphate, and sulfate; however, these are not included in most electrolyte panels.

Test Your Knowledge 18-8

True or False: Electrolytes are neutral; they neither exhibit a positive nor negative charge when dissolved in water.

(Outcome 18-1)

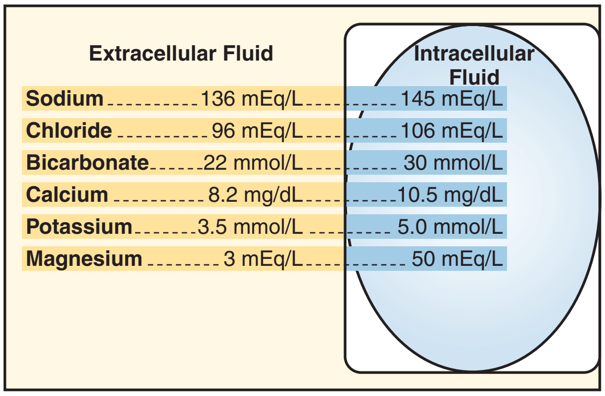
Extracellular Fluid
Intracellular Fluid
Sodium 136 mEq/L 145 mEq/L
Chloride 96 mEq/L 106 mEq/L
Bicarbonate 22 mmol/L 30 mmol/L
Calcium 8.2 mg/dL 10.5 mg/dL
Potassium 3.5 mmol/L 5.0 mmol/L
Magnesium 3 mEq/L 50 mEq/L

Figure 18-1 Major intracellular and extracellular electrolytes.

Function of Electrolytes and Consequences of Electrolyte Imbalances

Appropriate amounts of the various electrolytes in the body are essential to maintain homeostasis and acid-base balance. The electrolytes are also responsible for transmission of nervous impulses, and they play a major role in contraction and relaxation of the muscle fibers present throughout the body. The presence of the various intracellular and extracellular electrolytes in appropriate concentrations also allows for cells to keep their shape so that they can function properly. More specific electrolyte functions include the following:

- **Sodium (Na^+):** Sodium is the electrolyte found in highest concentration outside the cells of the body. We take in sodium through our diets, and the amount retained by the body is controlled by the kidneys. Sodium loss also occurs through perspiration. Sodium is essential for controlling the amount of fluid retained in the cells of the body. When the sodium levels are too low (a condition known as **hyponatremia)** in relation to the amount of fluid present in the body, the patient may exhibit drowsiness and confusion. If the condition worsens, it can lead to muscle twitches, seizures, coma, and even death. Hyponatremia can occur with various renal disorders, endocrine imbalances that the amount of antidiuretic hormone (ADH) produced, excessive water intake, or retention of fluids from various disease processes. Certain prescription drugs may also contribute to low sodium levels. **Hypernatremia** (elevated sodium levels) results when the concentration of sodium is elevated as compared to the amount of fluid in the body. Essentially, a patient with hypernatremia has too little water in the body for the amount of sodium that is present. This can be the result of severe vomiting and/or diarrhea with fluid loss, or excessive sweating without fluid replacement. Hypernatremia may also occur with severe burns, and is a risk for those with disease states that cause excessive urine to be produced. Sodium levels that are outside the reference range may lead to lethargy, confusion, muscle twitches, seizures, coma, and death if not treated promptly.
- **Potassium (K^+):** Potassium is present in high concentrations within the cells of the body, but normally is quite low in the blood plasma. The kidneys control the amount of potassium excreted from the body, and dietary intake helps to keep the necessary balance for healthy function. The presence of potassium is essential for cardiac function, as well as the transmission of nervous impulses to other muscles of the body. **Hyperkalemia** is the word used to describe elevated levels of potassium in the blood, and **hypokalemia** is the word used to describe diminished levels of potassium. Both conditions may be caused by renal dysfunction or certain hormonal imbalances. Reduced dietary intake may also contribute to hypokalemia, as well as use of prescription medications that contribute to increased urination. Diseases that cause gastrointestinal **malabsorption** (problems with the absorption of nutrients from food ingested) will also cause hypokalemia. Potassium imbalances are not tolerated well by the body, and may result in cardiac arrhythmias or other symptoms of muscle weakness.
- **Chloride (Cl^-):** Chloride is an anion, which means that it has a negative charge when dissolved in water. It is a very important extracellular presence, as chloride contributes to homeostasis by participation in acid-base balance. Chloride moves into and out of the cells of the body in response to the transfer of hydrogen and carbon dioxide through the cellular membrane. When chloride levels become imbalanced, the acid-base neutrality of the body's tissues may be affected, and **acidosis** (increase in acidity of the blood) or **alkalosis** (increase in alkalinity of the blood) may result.
- **Bicarbonate (HCO_3^-):** Bicarbonate is formed when carbon dioxide combines with water present in blood plasma. As is the case with chloride, bicarbonate plays

an essential role in acid-base balance of the body. It acts as a **buffer,** which means that the presence of bicarbonate helps to keep the body's pH at a neutral level by interacting with substances capable of changing the pH. Most of the bicarbonate in the body is found outside the cells, but there is a small amount present intracellularly. Bicarbonate may be indirectly measured in the laboratory by analyzing the amount of total CO_2 present in the blood. Abnormal amounts of bicarbonate (directly tested or analyzed by measuring total carbon dioxide levels) may contribute to acidosis or alkalosis.

Test Your Knowledge 18-9

A patient who has been treated with diuretics for the past few months develops an arrhythmia and is seen in the ER. When the patient has his blood tested, which electrolyte level may be outside of the reference range?
(Outcome 18-13)

Testing Preparation and Common Testing Methods

Specimen collection for electrolyte testing does not require any specific preparation, but it is important to collect the sample using appropriate phlebotomy technique to avoid erroneous results. Electrolyte testing performed in hospital and reference laboratories usually requires plasma or serum samples, which need to be separated from the blood cells in the specimen within an hour of collection. Electrolyte results can easily be affected by poor preanalytical procedures. Because potassium is at such a high concentration within blood cells and exhibits such a low extracellular concentration, any specimen collection or processing errors that contribute to hemolysis will cause the plasma potassium values to be falsely elevated. Excessive hemolysis may also cause the plasma sodium and chloride results to be decreased because of a dilution effect from all the intracellular contents released when the cells are broken. Contamination of the specimen with IV fluids will also alter the plasma electrolyte results; the effects will vary depending on the type of fluids being administered. CLIA-waived electrolyte procedures use whole blood for testing. The samples may be obtained by capillary puncture or by venipuncture, using a lithium heparin (light green top tube) or sodium heparin (dark green top tube). It is not possible to determine whether a sample is hemolyzed at the time of the blood draw by looking at the whole blood specimen; hemolysis does not become apparent until the sample has been spun in a centrifuge or until it has sat long enough for the plasma and/or serum to begin to separate visibly from the cells. This means that it is especially critical to avoid damage to the specimen when it will be used for whole blood testing, because it will not be evident that the cells have been hemolyzed, which could cause erroneous test results for potassium and possibly for the other electrolytes. The phlebotomist must implement appropriate venipuncture technique, including the use of the correct needle size, with complete insertion of the bevel into the interior of the vein. The tourniquet should not be applied for longer than 1 minute, and pumping of the fist should not be encouraged. Also, the blood sample should be gently inverted (not shaken).

There are many instruments that have been approved for CLIA-waived electrolyte testing. These instruments are capable of performing all the tests included in most electrolyte panels, including sodium, potassium, chloride, and bicarbonate. The testing instruments are present in physician office laboratories, but may also be used as a STAT testing method in other environments. These include the Abaxis Piccolo Blood Chemistry Analyzer and the Abbott i-STAT System. (See Fig. 18-2.)

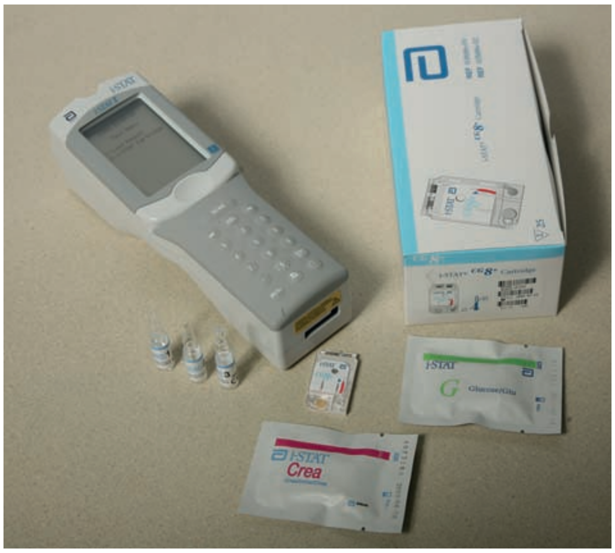

Figure 18-2 Abbott i-STAT analyzer. *Courtesy of Abbott Point of Care, Princeton, NJ.*

Test Your Knowledge 18-10

Which of these tests is not part of a typical electrolyte panel?

a. Sodium
b. Potassium
c. Nitrogen
d. Chloride (Outcome 18-14)

Reference Ranges

It is important to realize that the reference ranges for any analyte may vary according to the method used for testing. The following are approximate ranges established for adults for plasma samples:

- Sodium: 135 to 145 mmol/L
- Potassium: 3.5 to 5.0 mmol/L
- Chloride: 98 to 108 mmol/L
- Bicarbonate: 22 to 30 mmol/L

Procedure 18-2: Electrolyte Testing Using an Abbott i-STAT Chemistry Analyzer

TASK

Perform CLIA-waived whole blood electrolyte testing using an i-STAT analyzer.

CONDITIONS

- Gloves
- Laboratory coat
- Hand sanitization supplies
- I-STAT chemistry analyzer
- I-STAT analyzer electrolyte (panel or individual test) cartridges
- Capillary puncture supplies or venipuncture supplies and green top tubes
- Biohazardous sharps container
- Gauze or laboratory wipes
- Biohazardous waste container
- Quality control materials

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I. Anatomy and Physiology, #13. Perform Chemistry Testing **I.A.I.** Anatomy and Physiology, #2. Use language/verbal skills that enable patients' understanding

ABHES Standards

- 10. Medical Laboratory Procedures, b. CLIA-waived tests
- 10. Medical Laboratory Procedures, b. 3) Chemistry Testing

Procedure	Rationale
1. Greet and then identify patient using at least two unique identifiers.	All patients must be identified properly before collecting samples or performing laboratory testing.
2. Verify test ordered, and explain the testing procedure to the patient.	All laboratory test orders should be verified by checking the chart and/or requisition form more than once. A quick explanation of the procedure for the patient will ensure more cooperation.

Continued

Procedure 18-2: Electrolyte Testing Using an Abbott i-STAT Chemistry Analyzer—cont'd

Procedure	Rationale
3. Assemble necessary equipment, and verify that all reagents are within their expiration dates. For an electrolyte panel to be performed, it may be necessary to use a cartridge that tests for other analytes in addition to those in the electrolyte panel, due to the cartridge packaging.	i-STAT cartridges are to be stored at 2° to 8°C until their printed expiration. Cartridges must be removed from the refrigerator and allowed to be at room temperature for at least 5 minutes before use. A cartridge expires 2 weeks after it is removed from the refrigerator, and it is not to be put back in the refrigerator once it has reached room temperature. Materials that are organized will assist in the process of following the directions properly.
4. Wash hands and apply gloves.	Hands should always be washed between patients and before starting any procedures. Gloves are appropriate personal protective equipment (PPE) for this procedure.
5. Turn on the analyzer. If necessary, use the electronic simulator cartridge to verify that the instrument is operating correctly.	An internal electronic simulation occurs after every 8 hours that the instrument is in use. It may also be necessary to use an external electronic simulator cartridge if the internal test is not successful or if the laboratory policy dictates the use of this additional step. This cartridge is stored at room temperature. To use, insert the cartridge into the instrument with the I facing upward. If the test is successful, the LED screen will display a PASS message.
6. Prior to performance of test, verify whether a quality control (QC) specimen needs to be tested, and if so, complete that test before the patient's test is performed. It is recommended that two levels of quality control be tested every time QC is run.	Liquid QC should be tested following laboratory protocol and manufacturer's recommendations. Examples of when QC should be used to verify the test results include the following: **a.** With a new shipment **b.** Whenever a new lot number of reagents or QC is put into use **c.** New operator; someone who is being trained on the procedure **d.** Problems with storage, instrument, reagents, etc. **e.** To ensure that storage conditions are fine, QC should be performed at least once monthly
7. Remove the cartridge from the foil pouch. Be careful to handle the cartridge only by the edges, and place it on a level, clean surface until use.	If the contact points of the cartridge are scratched or damaged before the test is performed, it may have an adverse effect on the test results. The cartridge should be removed from the pouch before the sample is collected just to save time.
8. Perform a capillary puncture, using appropriate technique. Wipe away the first drop of blood. Use a heparinized capillary tube to collect a blood sample, avoiding the formation of air bubbles in the tube.	The first drop of blood should be discarded, as it may be contaminated with interstitial fluids, resulting in an erroneous result.

Procedure	Rationale
9. Hold the end of the filled capillary tube to the sample well. The well will fill with blood through capillary action. Fill the well to the indicated fill mark. The sample may also be taken from a sodium heparin– or lithium heparin–evacuated blood tube, or from a sterile syringe without additives.	The blood will enter the cartridge well by capillary action. It is acceptable to slightly overfill the well, but if the well is grossly overfilled or if blood is smeared on the cartridge outside of the well, the instrument will give an ERROR code when the assay is performed. Avoid air bubbles as the sample well is filled. If an evacuated blood tube is used, a plain plastic capillary tube can be used to transfer the specimen to the cartridge. If a sterile syringe is used for the venipuncture, it is important to apply the sample immediately to the cartridge to avoid clots in the specimen.
10. The analyzer will request an operator ID and a specimen/patient identification number. Enter these according to laboratory policy.	These identification methods allow for better documentation and tracking.
11. After the analysis has completed, the result will be displayed on the LED screen for 45 seconds. It may be recorded directly on the patient log sheet and transferred to the patient chart when ready.	The result may be recalled before running the next specimen by pushing DIS (display) on the instrument. Sample results are also kept in the analyzer memory, and some laboratories will have the instrument interfaced with their patient data system for direct documentation in the patient record.
12. Following office protocol, allow the patient to leave testing area, or consult with the health-care provider.	If results are far outside of the normal range, office policy may be that the health-care provider speaks with the patient before he or she leaves the office.
13. Remove the used cartridge from the instrument, and discard it as biohazardous trash.	The cartridge contains blood, so it must be discarded as a biohazardous substance.
14. Turn off the instrument or prepare to run the next specimen.	If numerous specimens are waiting to be assayed, leave the instrument turned on between samples.
15. Dispose of the capillary puncture device or used venipuncture supplies in a sharps container immediately after use. Any other equipment that may be contaminated with blood must be disposed of as a biohazard.	It is vital to always properly dispose of biohazards.
16. Disinfect the work area.	It is important to keep the work area clean and organized. Prompt disinfection helps to stop the cycle of infection.
17. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
18. Document the test results in patient chart.	All results must be documented in the patient's chart.

Date	
8/17/2014:	Electrolyte test performed: Sodium 140 mmol/L, Potassium 3.7 mmol/L, Chloride 100 mmol/L, TCO2 25 mmol/L
11:29 a.m.	Connie Lieseke, CMA (AAMA)

SUMMARY

Clinical chemistry is a complex laboratory department where a wide variety of tests may be performed. This chapter focused on some of the clinical chemistry tests that might be performed in a physician office laboratory as well as in larger reference laboratories. These include lipid panels, electrolyte analysis, and fecal occult blood testing. Lipid testing is very important, as elevated levels of total cholesterol, LDL, and triglycerides may be indicative of potential damage to the cardiovascular system and subsequent myocardial infarction or stroke. Electrolytes are responsible for conduction of nerve impulses and maintenance of homeostasis with acid-base balance. Imbalances of sodium, potassium, chloride, and bicarbonate can be life threatening, and may indicate an ongoing disease process elsewhere in the body.

TIME TO REVIEW

1. Hypernatremia is: Outcome 18-1
   - a. Elevated sodium concentration in the blood
   - b. Elevated potassium concentration in the blood
   - c. Decreased sodium concentration in the blood
   - d. Decreased potassium concentration in the blood
2. Statins are used to treat: Outcome 18-1
   - a. Electrolyte imbalances
   - b. Anemia
   - c. Elevated triglyceride levels
   - d. Elevated cholesterol levels
3. True or False: Cholesterol levels are increased by excessive ingestion of carbohydrates. Outcome 18-2
4. True or False: When evaluating cholesterol levels, the total cholesterol is not the only level considered by the practitioner. Outcome 18-9
5. Which type of lipoprotein is considered to be the "good cholesterol"? Outcome 18-3
   - a. HDL
   - b. VLDL
   - c. Triglyceride
   - d. LDL
6. Elevated lipid levels in the bloodstream increase the chance of: Outcome 18-8
   - a. Plaque formation on the blood vessel walls
   - b. Myocardial infarction
   - c. Liver damage
   - d. Electrolyte imbalances
7. Which of these are included in the metabolic syndrome? Outcome 18-1
   - a. Elevated triglyceride levels
   - b. Hypertension
   - c. Hyperglycemia
   - d. Excessive fat deposits around the waistline
   - d. All of the above
8. Which tests are included in a lipid panel? Outcome 18-9
   - a. Total cholesterol, HDL, LDL, VLDL, triglyceride
   - b. HDL, GLDL, LDL, BUN
   - c. Total cholesterol, triglyceride, HDL, anion gap
   - d. None of the above
9. How should a patient prepare for a lipid panel specimen collection? Outcome 18-10
   - a. No preparation is necessary
   - b. The patient must fast for 12 hours
   - c. The patient should follow a carbohydrate-restrictive diet for a week prior
   - d. The patient should follow a protein-rich diet for a week prior
10. What is a consequence of elevated or decreased bicarbonate levels? Outcome 18-13
    - a. Loss of cellular membrane shape
    - b. Cardiac arrhythmia
    - c. Acidosis or alkalosis
    - d. Muscle tremors

Case Study 18-1: **Difficult draw**

Kim Babado, CMA (AAMA), is working in the laboratory area of an internal medicine clinic. She has the responsibility of drawing blood on most of the patients in the practice, performing some point-of-care testing, and processing blood and other samples to prepare them for transport to a large reference laboratory across town. The second patient of the morning has an order for an electrolyte panel. When Kim examines the patient's arm, she has a difficult time finding a suitable site for the blood draw. The patient has very small veins that seem to roll away every time Kim touches them. Kim's first attempt is unsuccessful. On her second try, Kim does successfully puncture the vein, but obtains a slow blood flow that fills the tube approximately half full. After the blood specimen is spun in the centrifuge, Kim notes a light pink shade to the plasma.

1. Do you think that the specimen is appropriate for electrolyte testing?

Case Study 18-2: **Dietary effects**

A patient has a lipid panel drawn on Monday afternoon, and his results include an elevated triglyceride level. The ordering physician asks for the specimen to be redrawn to verify the results before meeting again with the patient.

1. What is a possible explanation for the elevated triglyceride level for this patient?
2. What might be important information to give the patient before the next blood draw?

RESOURCES AND SUGGESTED READINGS

"Colorectal Cancer Screening"
Colorectal cancer screening information and recommendations from the Centers for Disease Control and Prevention http://www.cdc.gov/cancer/colorectal/basic_info/screening/

"Hemoccult's Newest Gold Standard in FOBTs-Hemoccult ICT"
Beckman Coulter Hemoccult ICT test information http://www.hemocultfobt.com/sales/sales_Hemo_ICT.htm

"Hemoccult Fecal Occult Blood Tests"
Provides links to overview for all of Beckman Coulter Hemoccult products http://www.hemoccultfobt.com/

"Tests Granted Waived Status Under CLIA"
List of all CLIA-waived tests, including CPT code for reimbursement http://www.cms.hhs.gov/CLIA/downloads/waivetbl.pdf

"Cholestech"
Detailed information about the Cholestech analyzers available for lipid testing. http://cholestech.com

"Abbott Point of Care"
Detailed information about the Abbott i-STAT analyzer and other point-of-care testing instruments manufactured by Abbott http://abbottpointofcare.com

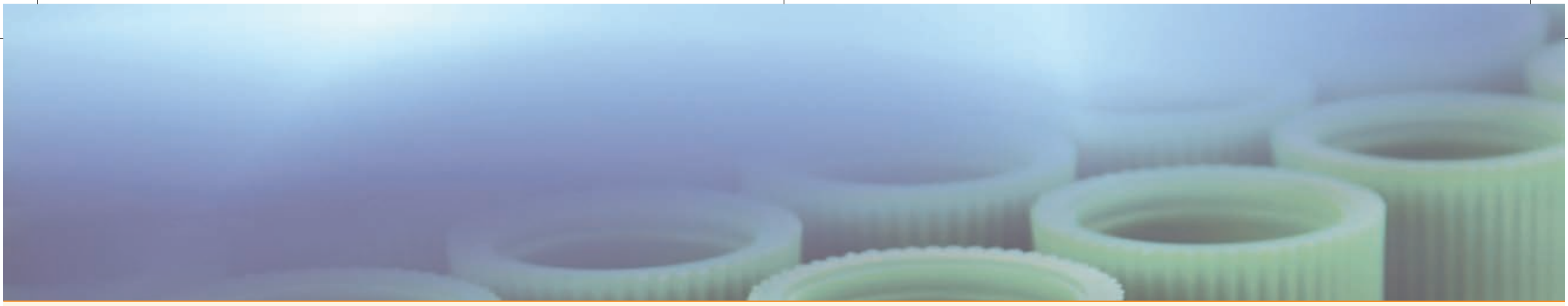

Section IV

Clinical Chemistry

What Does It All Mean?

As the chapters in this section clearly point out, there is a plethora of tests that are typically performed in the clinical chemistry area of the laboratory. These tests may be ordered and performed on an individual basis or may be ordered and tested in groups known as profiles. Physicians and other primary care providers such as nurse practitioners and physician assistants make these important ordering decisions.

Case in Point

At the beginning of this section we met 34-year-old Carrie W. A fasting blood glucose test was performed on her as part of a health fair sponsored by Carrie's employer, Acme Innovations, Inc. Her value was significantly higher than the highest number in the normal range for this test. The fact that this test was done after a period of fasting indicates that her ability to process glucose is likely impaired.

Glucose serves as the body's main source of energy. Under normal circumstances, glucose is metabolized and the resulting products are used by the body for the many functions it performs on a daily basis. An unusually high glucose level suggests that the body is unable to metabolize glucose properly. The most likely condition associated with this scenario is diabetes mellitus. Individuals with diabetes mellitus typically either do not produce the substance insulin that is necessary for glucose metabolism or they produce insulin but its function is impaired. In either event, glucose is not being properly used and eventually builds up in the body, thus causing increased blood glucose levels. Because the body can count on glucose for energy, it resorts to fat. Incomplete fat metabolism ensues. A by-product of incomplete fat metabolism is a substance known as ketone.

Once the body has reached its saturation point of glucose and ketones, excess amounts of both substances are deposited into the urine for disposal from the body. Thus, along with high blood glucose levels, individuals with diabetes mellitus typically also have high levels of glucose and ketones in the urine (urine testing is covered in detail in Chapter 21).

In some instances a change in diet stabilizes the blood glucose levels to the point that the diabetes mellitus can be manageable without taking other steps. However, there are instances when treatment with insulin is indicated.

On the Horizon

Perhaps the easiest, most noninvasive specimen type for analysis from humans is urine. In fact, the concept of analyzing urine has been noted in drawings on ancient history walls. Early analyzers determined that if urine from a diabetic individual is spread on the ground, then because of its sweet smell, ants would gravitate to it. At one point, a group of so-called self-taught urine analyzers, known as the "pisse prophets," marketed their services to individuals. Patients who submitted their urine and paid a fee received an analysis of their urine without ever being examined by a physician.

There have been many advances in urine analysis over the years, one of the more notable being the advent of plastic strips impregnated with chemical pads, known as a dipstick. Today, a test that incorporates the dipstick technology, known as a urinalysis, screens for a number of abnormalities. In many instances, such abnormalities may be detected through urinalysis testing, often before the patient experiences any symptoms. This section focuses on the urinalysis test, as it is crucial in the diagnosis and monitoring of disease as well as in the monitoring of select treatment regimens. The occult blood test is also covered in this section.

Relevance for the Medical Assistant (Health-Care Provider)

As with so many other aspects of laboratory testing, medical assistants and other health-care support personnel are often involved in the collection of samples for urinalysis testing by assisting the patient in the process or providing the patient with collection instructions. Furthermore, medical assistants often perform urinalysis and occult blood testing along with appropriate quality control. Medical assistants are also responsible for completing the documentation associated with this testing.

Continued on page 390.

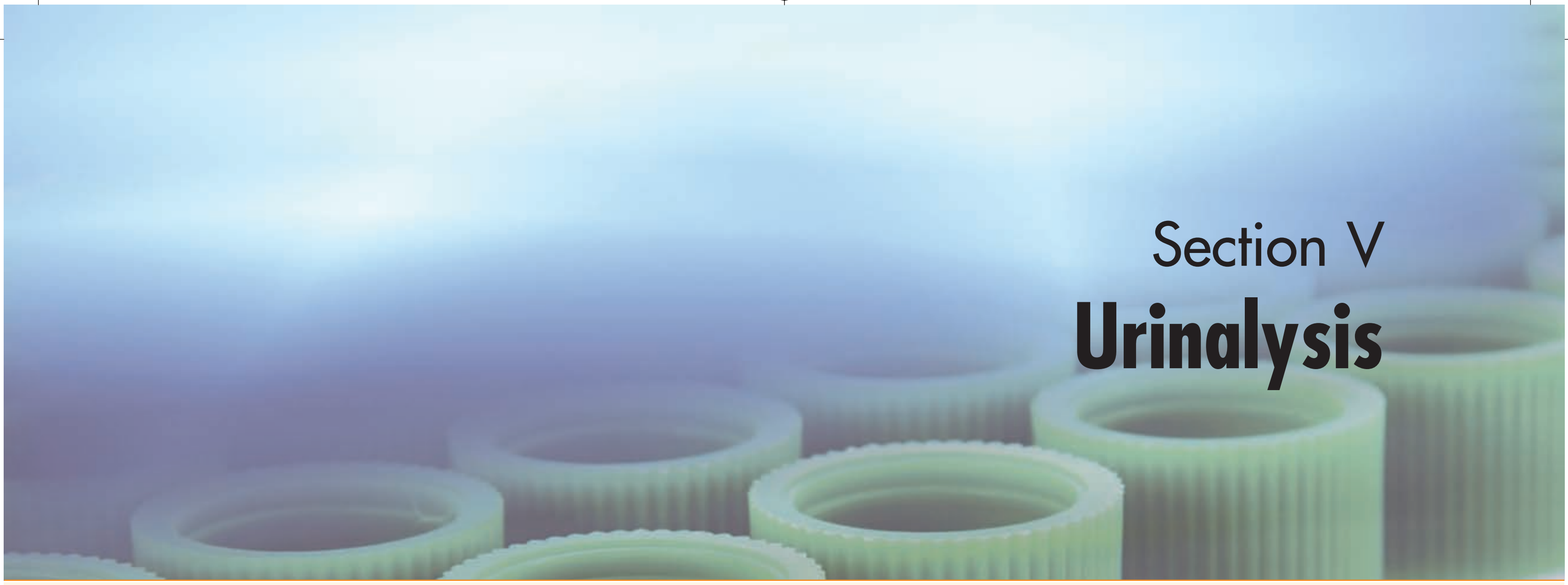

Section V

Urinalysis

On the Horizon

Chapter 19: Urinalysis introduces the reader to key anatomical parts associated with the urinary system. A brief overview of the formation of urine detailing the functions of the anatomical parts of the urinary system is presented. An overview of the routine urinalysis test including quality control is described.

Chapter 20: Physical Characteristics of Urine covers the color, clarity, volume, and odor of urine samples. The measurement of specific gravity and testing for it using a urinometer and refractometer are introduced. Proper technique, normal test reference range, sources of error and interference, and test interpretation are described.

Chapter 21: Chemical Examination of Urine and Feces identifies and describes 10 chemical tests that comprise a typical urine dipstick. The occult blood test is introduced and described. Proper technique, normal test reference range, sources of error and interference, and test interpretation are covered.

Chapter 22: Microscopic Examination of Urine introduces the reader to the most commonly seen microscopic "formed elements" found in urine. Proper technique, normal test reference range, sources of error and interference, and test interpretation are covered.

Continued from page 388.

Case in Point

Three-year-old Johnny is rushed to the doctor's office after he drank antifreeze in the family garage. As the medical assistant on duty at the time, you are asked to give instructions to Johnny's mother so she can help the young boy collect a urine sample. Once the sample is collected, you are asked to conduct the physical and chemical examination portions of a urinalysis test on it. You obtain the following results:

Urinalysis Physical Examination

Component	Result	Reference Range
Color	Amber	Yellow
Clarity	Cloudy	Clear

Urinalysis Chemical Examination (Dipstick)

Component	Result	Reference Range
Glucose	Negative	Negative
Bilirubin	Negative	Negative
Ketone	Negative	Negative
Specific gravity	1.005	1.005–1.030
Blood	Trace	Negative
pH	6.0	4.5–8.0
Protein	100 mg/dL	Negative–trace
Urobilinogen	0.2 mg/dL	0.2–1 mg/dL
Nitrite	Negative	Negative
Leukocyte esterase	Negative	Negative

Questions for Consideration:

- What key points would you make when giving directions to Johnny's mother when she helps Johnny collect a urine sample?
- What is the purpose of the physical examination of urine?
- What is the purpose of the chemical examination of urine?
- What is the third portion of the urinalysis test that has not yet been done?
- What is the purpose of the chemical examination portion of the urinalysis test?

Chapter 19

Urinalysis

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Historical Perspective of Urinalysis
- Anatomy and Physiology of the Urinary System
  - The Kidneys
  - Ureters
  - Bladder
  - Urethra
- Sequence of Urine Production and Excretion
- Clinical Significance of Urine Testing Results
  - Physical Examination
  - Chemical Examination
  - Microscopic Examination
- Quality Assurance for Urine Testing Procedures
- Standard Precautions Used When Analyzing Urine Specimens
- Types of Urine Specimens
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

19-1 Define the key terms.

19-2 Describe the function of the urinary system.

19-3 Differentiate the parts of the urinary system.

19-4 Explain how urine is formed in the kidneys.

19-5 Describe the route followed as urine is excreted from the body.

19-6 Discuss the clinical significance of urinalysis results.

19-7 Compare and contrast the different types of testing involved in routine urinalysis.

19-8 Describe common quality assurance procedures used for the analysis of urine.

19-9 Describe the personal protective equipment necessary to process urine samples.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.C.I.5: Describe the normal function of each body system.

ABHES Standards

None

KEY TERMS

Afferent arteriole
Aliquot
Anterior
Bladder
Bowman's capsule
Calyx
Distal convoluted tubule
Efferent arteriole
Filtration
Glomerular filtration
Glomerulus
Hilum
Kidneys
Loop of Henle
Macroscopic
Microscopic
Micturition
Nephrons
Peritoneum
Pisse prophets
Proximal convoluted tubule
Reabsorption
Reagent dipstick
Renal cortex
Renal medulla
Renal pelvis
Renal tubule
Retroperitoneal
Rugae
Secretion
Specific gravity
Ureters
Urethra
Urinalysis
Urinary meatus
Urine microscopic
Uroscopy

HISTORICAL PERSPECTIVE OF URINALYSIS

Urine has been used as a primary diagnostic tool by those who practice medicine for thousands of years. The first recorded observation of the characteristics of urine was written over 6,000 years ago. Ancient physicians could only make assumptions utilizing the color, appearance, smell, and sometimes taste of urine because their knowledge of human anatomy and physiology was very limited. Urine observation and analysis was known as **uroscopy.** Some of these early diagnostic theories resulting from uroscopy proved to be quite accurate, whereas others were later shown to be complete fallacies.

Unfortunately, the type and amount of information that could be derived from urine became exaggerated, and examination of urine without the presence of the patient or any other clinical information became popular. The analysis of urine lost medical credibility for a time, as the examination of urine became a tool used by uneducated individuals who pretended to diagnose disease and predict the future using urine specimens. This type of uroscopy was a profitable venture until 1637, when Thomas Brian published a book that labeled these practitioners as **pisse prophets.** The book discredited their practices and led to the demise of this profession.

Today, the analysis of urine is a valuable tool used by health-care practitioners to screen for potential abnormalities or monitor disease processes and the efficacy of treatment. A urinalysis may provide important information to assist with the diagnosis of diabetes or kidney damage, or detect a potential urinary tract infection. Urinalysis is a relatively inexpensive, noninvasive CLIA-waived procedure that is commonly performed in physician offices by medical assistants. The analysis of urine includes observation of the physical characteristics of the urine specimen that may be seen or measured with the naked eye. The physical characteristics of the urine include documentation of the color and appearance of the urine specimen. Urinalysis procedures also include procedures such as the measurement of urine **specific gravity** (a measurement of the density or concentration of the urine specimen), or measurement of the specimen volume when needed for reference. Additional analysis for the presence of various chemical components is another part of the testing process. Analysis of urine may also include a **microscopic** component, which uses a microscope to look for formed elements in the urine specimen. Urine microscopic analysis is not a CLIA-waived test, but medical assistants often prepare and set up the specimen on the microscope for examination by a practitioner or other qualified professional. Additional training may allow for medical assistants to perform microscopic examination of urine specimens as well.

Test Your Knowledge 19-1

How are procedures used to document the physical characteristics of the urine different from the microscopic procedures? (Outcome 19-7)

It is important to adhere to the proper collection procedure for urinalysis specimens. Ideally, urinalysis is performed on clean-catch midstream urine specimen samples, and if the procedures for this collection are not followed carefully, erroneous results may be reported to the physician. Inadequate cleansing before collection will introduce contaminants into the specimen that can alter the results. In addition, if a culture is added to the specimen after initial testing, it is important that the specimen be collected using the clean-catch midstream method. Ideally, approximately 50 mL of urine will be submitted for urinalysis testing, but according to the Clinical Laboratory and Standards Institute, the minimum volume for accurate results is 12 mL.

ANATOMY AND PHYSIOLOGY OF THE URINARY SYSTEM

The urinary system includes two **kidneys**, two **ureters**, the **bladder,** and the **urethra.** (These structures can be seen in Fig. 19-1.) The functions of the urinary system include filtration of the blood, excretion of excess waste, and the regulation of fluid and acid-base balance within the body. The kidneys also help to regulate blood pressure, and they produce erythropoietin, a hormone that stimulates production of red blood cells.

Test Your Knowledge 19-2

What is the function of the urinary system? (Outcome 19-2)

The Kidneys

The kidneys are considered the most important organs of the urinary system because they are responsible for the formation of urine. They are bean-shaped structures located to the back of the abdominal cavity on either side of the vertebral column. The **peritoneum,** or membrane that lines the abdominal cavity, covers only the **anterior** (front) side of the kidneys, which means that they are **retroperitoneal,** or behind the peritoneum. The left kidney is located a bit higher than the right kidney. A thick renal capsule consisting of connective tissue and fat surrounds and protects each kidney. The indented (concave) portion of the kidney is known as the **hilum,** the region where the renal artery enters the kidney and the renal vein and ureter leave the kidney.

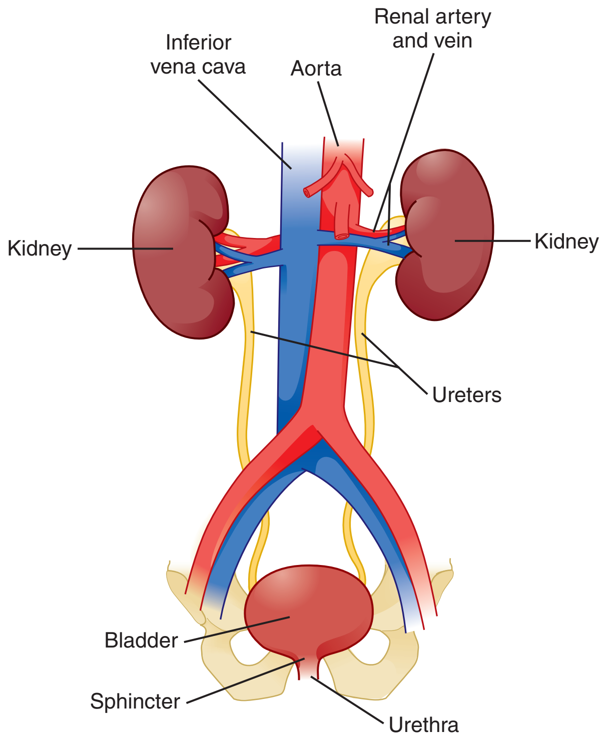

Figure 19-1 The parts of the urinary system, including the kidneys, ureters, urinary bladder, and the urethra.

Test Your Knowledge 19-3

Which of these best describes the kidneys?

a. Located in front of the peritoneum
b. Pear-shaped organs
c. Located on either side of the vertebral column with one slightly lower than the other
d. Covered with hilum (Outcome 19-3)

The kidneys are made up of an outer **renal cortex** and an inner **renal medulla.** The renal medulla may be visualized as a series of structures known as *pyramids,* as is evident in Figure 19-2. Each pyramid drains into a **calyx,** which serves as a collection tubule for the urine as it is created. The area of the kidney where the calyces join is called the **renal pelvis,** which is connected to the ureter leaving the kidney.

Each kidney contains approximately 1 million **nephrons.** Nephrons are known as the functional unit of the

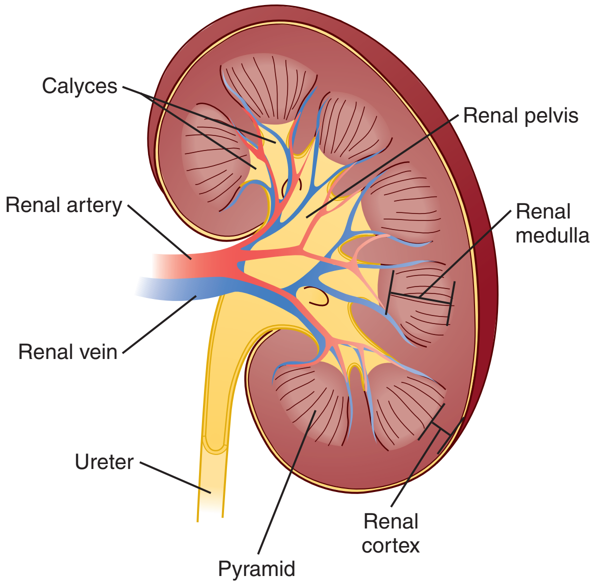

Figure 19-2 A cross section of a kidney, showing the renal vein and artery, ureter, renal pelvis, renal medulla, renal pyramids, and renal cortex.

kidneys, because these microscopic structures filter the blood and create urine utilizing filtration, reabsorption, and secretion, as described later in this chapter. The vast majority of the nephrons in the kidneys are *cortical nephrons,* located in the outer cortex layer of the kidney. All nephrons function to filter blood and create urine, but the cortical nephrons are responsible primarily for removal of waste products and reabsorption of nutrients. *Juxtamedullary nephrons* are longer, and extend deep into the medulla. These nephrons specialize in concentration of the urine.

Test Your Knowledge 19-4

Why are the nephrons called the functional units of the kidneys? (Outcome 19-3)

Almost one-quarter of the body's blood supply is in the kidneys at any given time. The kidneys are incredibly vascular to allow for adequate processing of the blood supply and production of at least 1 liter of urine each day. To understand how this process occurs, it is important to look closely at the anatomy and physiology of the nephrons, as they perform **filtration, reabsorption,** and **secretion** to create urine. A nephron is shown in Figure 19-3.

Blood enters the nephrons through a small vessel called the **afferent arteriole.** The blood continues to flow through a complex tangled tuft of capillaries that makes up the **glomerulus.** The capillaries rejoin, and the blood passes through the **efferent arteriole** as it leaves the glomerulus. The efferent arteriole is much smaller than the afferent arteriole, which causes a great deal of pressure to build up within the glomerulus. This hydrostatic pressure pushes out fluid from the blood as it travels through the tuft of capillaries. The cell lining of these capillaries is different from those in other places in the body, as it contains pores that allow small molecules to pass through and facilitates the passage of fluid at a high rate. This process is called **glomerular filtration.**

The fluid that is forced from the capillaries during glomerular filtration is collected by a tubule that surrounds the tuft of capillaries. This is called the **Bowman's capsule.** The Bowman's capsule is the beginning of the **renal tubule,** a hollow structure that contains the blood filtrate. Not all the fluid forced from the capillaries is excreted from the body as urine; if it were, the body would be depleted of vital electrolytes and fluid very quickly. Approximately 99% of the fluid and electrolytes are reabsorbed by capillaries that surround the renal tubules. Additional substances are also secreted into the filtrate before it leaves the kidneys as urine.

The renal tubules form a V-shaped complex, including the **proximal convoluted tubule,** the **loop of Henle,** and finally the **distal convoluted tubule** on the other side of the V. (See Fig. 19-3.) This complex of tubules is surrounded by capillaries, which allows for essential exchanges to occur between the blood and the filtrate within the tubules. The primary reabsorption of fluid and electrolytes occurs in the proximal convoluted tubule. More water, sodium, and chloride are reabsorbed into the blood from the filtrate in the loop of Henle at the bottom of the V-shaped complex. Most of the final adjustments are accomplished in the distal convoluted tubule to create urine from the glomerular filtrate. Urine that has passed through the renal tubules flows into the collecting duct, which then drains into the calyx area of the kidney. The calyces open up to form the renal pelvis, and the urine leaves the kidneys through the ureter.

Test Your Knowledge 19-5

What processes occur in the renal tubules? (Outcome 19-4)

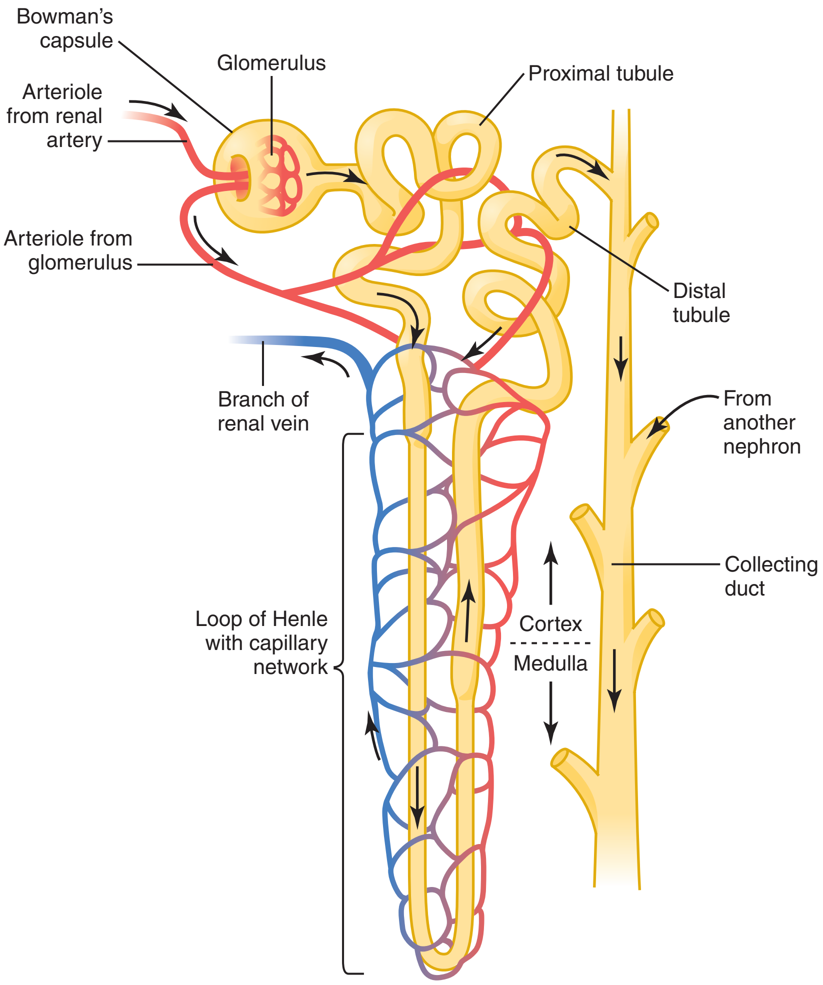
Bowman's capsule
Glomerulus
Proximal tubule
Arteriole from renal artery
Arteriole from glomerulus
Distal tubule
Branch of renal vein
From another nephron
Collecting duct
Loop of Henle with capillary network
Cortex
Medulla

Figure 19-3 The components of the nephron, including the glomerulus, Bowman's capsule, proximal convoluted tubule, loop of Henle, and distal convoluted tubule.

Ureters

The ureters are flexible hollow tubes that connect the kidneys to the bladder. Urine flows from each renal pelvis through the ureter by peristaltic waves created by the muscular ureter walls. Each ureter is approximately 12 in. long, which allows enough length for them to pass behind the abdominal cavity before entering the bladder.

Test Your Knowledge 19-6

Are the ureters involved in the production of urine?
(Outcome 19-4)

Bladder

The bladder is a hollow muscular sac that is designed to store urine. The size and shape of the bladder differ according to the amount of urine being stored at any time. The interior lining of the empty bladder forms **rugae,** which are folds that allow for the bladder to expand when filled with urine. Typically, the adult bladder can hold about 500 mL of urine, but the urge to urinate can be felt when the bladder contains approximately 150 mL of urine. The release of urine from the bladder **(micturition)** involves the interaction of the micturition reflex center in the sacral region of the spinal cord and the voluntary muscles that surround the urethra.

Urethra

When the urine leaves the bladder, it travels through a hollow tube called the urethra as it leaves the body. The urethra varies in length according to the patient's gender. The male urethra is approximately 21 cm in length, whereas the female urethra averages only 4 cm. In males the urethra also is used by the reproductive system to transport semen. The opening at the distal (far) end of the urethra where it leaves the body is called the **urinary meatus.**

SEQUENCE OF URINE PRODUCTION AND EXCRETION

Each part of the urinary system plays an important role in the formation and excretion of urine. The process begins as the blood is filtered in the glomerular capillaries and ends as the urine passes through the urinary meatus.

- Blood is forced through the glomerular capillaries with sufficient pressure to force most of the liquid portion of the blood out through the pores in the capillaries to form the glomerular filtrate.
- The glomerular filtrate is collected in the Bowman's capsule that surrounds the glomerulus.
- The glomerular filtrate continues to pass through the renal tubules, including the proximal convoluted tubule, the loop of Henle, and the distal convoluted tubule.
- The refined filtrate (now known as urine) enters the collecting ducts.
- The collecting ducts join to drain into the calyces, which are part of the renal pelvis.
- The urine leaves the renal pelvis through the hilum into the ureter.
- The ureter transports the urine into the urinary bladder.
- Urine leaves the body through the urethra.
- The distal end of the urethra where the urine leaves the body is known as the urinary meatus.

CLINICAL SIGNIFICANCE OF URINE TESTING RESULTS

Urinalysis is one of the most common tests performed in physician offices and laboratories. The test is fast, simple, and relatively inexpensive to perform. Urine testing is considered reliable to detect or monitor the progress of numerous disease states. Disorders of carbohydrate metabolism, urinary tract infections, and abnormal pH changes in the body may all be detected through analysis of urine. The malfunction of the kidneys or liver may also be evident in a urine specimen. Urinalysis is also commonly performed as a screening procedure for asymptomatic disease. The analysis of urine has several parts: physical examination, chemical testing, and a microscopic examination of the urine sediment for detection of formed elements present in the specimen.

Test Your Knowledge 19-7

True or False: Urinalysis is performed only to confirm a clinical diagnosis. (Outcome 19-6)

POINT OF INTEREST 19-1

Bladder cancer testing

Bladder cancer is one of the most prevalent cancers in the United States. It is most common in men over the age of 60. The risk of developing bladder cancer is greatly elevated in those who smoke, and patients who have experienced chemical exposure (such as dry cleaning chemicals) are also at a high risk for bladder cancer. One of the first symptoms of bladder cancer may be the presence of blood in the urine (hematuria), as well as increased urinary frequency, irritability, and pelvic pain. If detected early, bladder cancer is highly treatable.

As of June 2004 the Department of Health and Human Services U.S. Preventive Services Task Force recommends against routine screening for bladder cancer for asymptomatic patients without the presence of risk factors. However, there are various tests for patients who show symptoms of bladder cancer, or for those who have multiple risk factors. One of these is cystoscopy, which is a visual examination of the bladder using a scope. Biopsies that are obtained during a cystoscopy may also provide valuable information.

In addition to these screening procedures, the health-care provider may choose to test the patient's urine specimen for the presence of the *nuclear matrix protein*, which is often elevated in the presence of bladder cancer. The NMP22 Bladder Chek test (manufactured by Matritech Corporation) is the only U.S. Food and Drug Administration–approved, CLIA-waived test for the screening of urine for potential bladder cancer. For patients who have been previously diagnosed with bladder cancer, the BTA Stat test (manufactured by Polymedco) may be used to detect bladder cancer cells. The BTA Stat test has not been approved for use as a screening mechanism. A positive result with either of these tests will lead to further investigative procedures to establish the presence of cancer cells and define the severity of the condition.

Physical Examination

The first step taken when analyzing a urine specimen is to perform a direct observation of the urine specimen. The color and appearance of the urine specimen may be noted, depending on the policies and procedures of the laboratory. Some laboratories will also allow for documentation of any unusual specimen odor, because an ammonia-like odor may indicate an old urine specimen that has not been properly preserved. In addition, the physical examination may include the documentation of the urine specific gravity, using a refractometer. The specific gravity measurement provides information about the amount of dissolved substances present in the urine specimen. The physical examination, as well as the rest of the procedures included in a urinalysis, should be performed on fresh urine specimens within 1 hour of collection. If there will be a delay in the testing, the specimen needs to be refrigerated at 4° to 6°C or added to a preservative tube as directed by laboratory policy. If refrigerated, the specimen needs to be brought back to room temperature before testing.

Test Your Knowledge 19-8

What is observed and recorded during the physical examination of a urine specimen? (Outcome 19-7)

Chemical Examination

After the initial physical examination has been performed and recorded, the urine specimen will be tested with a **reagent dipstick,** a thin piece of plastic covered with reagent pads that will change color depending on the concentration of specific chemicals in the urine specimen. Reagent dipsticks are commonly used to test for glucose, ketones, pH, bilirubin, urobilinogen, red blood cells, white blood cells, nitrites, albumin, and protein. Specific gravity may also be measured using this method. Chemical testing may be performed manually with visualization of the color changes on the reagent pads, or an instrument may be used to evaluate the changes in color and assign values. More details about the chemical testing of urine are presented in Chapter 21.

Microscopic Examination

The final test performed as part of a urinalysis is the **urine microscopic** examination. An **aliquot** (small sample) of the urine specimen is centrifuged. At least 12 mL of urine are required to obtain an appropriate specimen for microscopic examination. The sediment formed at the bottom of the tube during centrifugation is placed on a slide and viewed under the microscope. The laboratory professional will document the formed elements present when the specimen is examined. Cells commonly present in the urine specimen include squamous epithelial cells as well as red and white blood cells. Bacteria and fungi may be observed along with many other microscopic structures. Chapter 22 focuses more specifically on the microscopic urine examination and the significance of the results.

Test Your Knowledge 19-9

Which part of the urinalysis procedure uses reagent test strips?

a. Observation of color and clarity
b. Microscopic examination
c. Chemical analysis
d. Culture (Outcome 19-7)

POINT OF INTEREST 19-2

HIV urine testing

According to the Centers for Disease Control and Prevention, every year there are more than 16 million people in the United States who are tested for potential HIV infection. Many of these patients will have a rapid test performed in the physician office, using an oral sample (from the mucosal surface of the mouth) or a capillary blood sample. Most patients are not aware that urine may also be tested for HIV antibodies. Urine HIV testing has not been approved as a rapid test method, but Calypte Biomedical Corporation manufactures a urine screening test that has been approved by the U.S. Food and Drug Administration for HIV testing. It is important to remember that urine is not considered to be an infectious material for HIV; however, antibodies may be present in urine. Urine testing for HIV has been shown to be a little less sensitive and specific for HIV antibodies

Continued

than the testing performed on blood or oral samples. Any positive results should be rechecked two times, and if the result is positive, the specimen must be sent out for confirmation with a western blot confirmatory test before the patient is to be considered HIV positive.

QUALITY ASSURANCE FOR URINE TESTING PROCEDURES

Urine analysis involves several different processes, some of which may be performed manually or by means of instrumentation. Quality assurance methods for handling and testing urine specimens include the following:

- Appropriate explanation of collection techniques should be provided to patients. A clean-catch midstream urine collection technique is best for urinalysis testing to provide a sample that is contaminant free.
- A sterile collection container with appropriate labeling should be provided to the patient. This allows the specimen to remain contaminant free in case there is a culture added to the specimen after the urinalysis is performed.
- Testing should occur within 1 hour of collection whenever possible. If this is not feasible, refrigeration or preservation of the sample at 4° to 6°C must occur within 1 hour.
- Calibration of and use of quality control specimens on the refractometer if used for specific gravity testing must be implemented.
- Staff training must be ongoing for standardization of terms used to report the physical examination and other results.
- Use of external quality control specimens must occur; a positive and negative quality control specimen should be processed for all parameters reported on patient specimens. The frequency of testing is dependent on the number of urinalysis tests performed, the manufacturer's recommendations, and the adopted laboratory policies.
- Reagent strips must be stored properly; they are to remain in the original container with the dessicant as packaged, and they should remain out of direct light and be protected from humid environments. The cap must stay on the container to keep moisture from affecting the strips, which would result in erroneous results.
- Centrifuges used to spin urine specimens should be calibrated regularly.
- Commercial quality control specimens should be processed by all who perform urine microscopic examination. Slides and pictures used for proficiency testing may also be useful.

STANDARD PRECAUTIONS USED WHEN ANALYZING URINE SPECIMENS

As introduced in Chapte 3, Standard Precautions combine the principles of Universal Precautions with those of Body Substance Isolation. Universal Precautions state that all blood and other potential infectious body fluid specimens should be treated as if they are infectious for HIV, hepatitis B, and other bloodborne pathogens. Body Substance Isolation procedures specify how employees are to be protected from other pathogens that may be present in a specimen. Essentially, when testing urine using the Standard Precautions, it is important for the employee to be shielded from potential exposure to the pathogens that may be present in the urine specimen. This protection may be accomplished by use of gloves when handling the urine specimens and by wearing a laboratory coat to protect against potential splashes or spills. Eye protection should be worn when removing the cover from the tube to protect against aerosols or splashing.

Test Your Knowledge 19-10

Which of these is a type of personal protective equipment (PPE) that should be worn while a medical assistant performs a urinalysis test?

a. Laboratory coat
b. Gloves
c. Respiratory protection
d. a and b
e. a, b, and c (Outcome 19-9)

TYPES OF URINE SPECIMENS

This chapter and the others in this section focus on the collection, processing, and testing of urinalysis samples. A urinalysis should be performed using a clean-catch midstream urine specimen, preferably collected with the first morning void. However, as presented in Chapter 9, urine specimens may be used for other types of testing as well. These samples differ in their collection requirements, analysis procedures, and the clinical significance of the results. Examples of alternative urine specimens include the following:

- Timed specimens, such as postprandial specimens or those collected during a glucose tolerance testing procedure
- 24-hour urine specimens used for specific chemical analysis
- Postprostatic massage specimens
- Urine collected for *Chlamydia* testing
- Urine collected for drug testing

SUMMARY

For many years analysis of urine specimens has been an invaluable tool for providing important clinical information. The urinary system is complex, and is responsible for the filtration of toxins from the body, as well as the acid-base and overall fluid balance that is necessary for homeostasis. The technological advances in testing methods are dramatic, and urine may now be used to identify and/or monitor the progress of many disease states. Urinalysis is commonly performed in physician offices as well as large reference laboratories. A typical urinalysis test includes macroscopic, physical, chemical, and microscopic components. Quality control plays a very important role in every part of urine testing procedures.

TIME TO REVIEW

1. The renal calyx is: Outcome 19-1
   a. A cup-like extension of the renal pelvis
   b. Part of the nephron
   c. The membrane that covers the kidney
   d. A vessel in the Bowman's capsule
2. Specific gravity is a ratio that compares: Outcome 19-1
   a. The weight of the urine specimen to a "normal" urine specimen
   b. The weight of the urine specimen to the patient's blood
   c. The weight of the urine specimen to distilled water
   d. None of the above
3. The afferent and efferent arterioles: Outcome 19-2
   a. Enter and exit the bladder
   b. Enter and exit the kidney
   c. Enter and exit the ureters
   d. Enter and exit the glomerulus
4. True or False: The urinary system is very important for maintenance of the body's normal pH. Outcome 19-2
5. The urinary bladder: Outcome 19-3
   a. Helps to concentrate the urine
   b. Serves as a storage receptacle
   c. Can stretch to hold more than 1 L of fluid
   d. Is superior to the kidneys in the body
6. The initial filtration of the blood: Outcome 19-4
   a. Occurs in the distal convoluted tubule
   b. Occurs as the blood enters the kidney
   c. Occurs as the blood passes through the glomerulus
   d. Occurs in the ureters
7. True or False: Microscopic examination of urine specimens is an automated procedure. Outcome 19-7
8. What is the recommendation for the use of quality control materials for the reagent strip testing portion of the urinalysis? Outcome 19-8
   a. Test positive controls only
   b. Test negative controls only
   c. Test positive and negative controls
   d. No quality control materials need to be used for this part of the analysis
9. Why should a laboratory coat be worn when performing urinalysis testing? Outcome 19-9
10. True or False: Because blood is not usually present in urine specimens, it is not necessary to wear gloves when processing urine samples. Outcome 19-9

Case Study 19-1: **Reagent storage requirements**

When John arrives at work in the office laboratory on Monday morning, he finds that the lid for the urine reagent strips is not on the container. There was no one in the office over the weekend, and John was not the last person in the laboratory at the end of the day on Friday.

1. Should John use these strips for urine testing? Why or why not?

RESOURCES AND SUGGESTED READINGS

Clinical and Laboratory Standards Institute, Urinalysis: Approved Guideline, ed. 3. CLSI document G16-A3. Wayne, PA, 2009

Approved laboratory standards for collecting, processing, and testing urinalysis samples

Becton, Dickinson, "Best Practice: Tips for Urinalysis." Lab-Notes, 13, no. 2, 2003

Online newsletter from Becton, Dickinson and Company. Includes details about urine processing for urinalysis and urine cultures, including information about preservative tubes http://www.bd.com/vacutainer/labnotes/2003spring/best_practice.asp

Chapter 20

Physical Characteristics of Urine

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

Physical Characteristics of Urine
  Urine Color
  Urine Clarity
  Specimen Volume
  Urine Odor
  Specific Gravity
    Refractometer Measurement Technique
    Urinometer Measurement Technique
    Reagent Strip Methodology
  Procedure for Recording Physical Characteristics of a Urine Specimen
Potential Sources of Error
Summary
Time to Review
Case Study
Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

20-1 Define the key terms.

20-2 Describe the clinical significance of abnormal urine color.

20-3 Explain how urine clarity is interpreted and reported.

20-4 Discuss the clinical significance of urine volume and odor.

20-5 Describe the clinical significance of specific gravity results.

20-6 Compare and contrast the different methods used to perform urine specific gravity testing.

20-7 Provide the units used to report specific gravity results.

20-8 List potential sources of error for urine physical examination procedures.

20-9 Analyze and document the physical characteristics of a urine specimen.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.13. Perform Urinalysis

ABHES Standards

- 10. Medical Laboratory Procedures, b. Perform selected CLIA-waived tests that assist with diagnosis and treatment, #1: Urinalysis

KEY TERMS

Anuria
Amorphous phosphates
Amorphous urates
Clarity
Enuresis
Glycosuria
Hematuria
Meniscus
Myoglobin
Nocturia
Oliguria
Polyuria
Pyuria
Refractive index
Refractometer
Specific gravity
Urinary frequency
Urinary incontinence
Urinometer
Urochrome

PHYSICAL CHARACTERISTICS OF URINE

There are several steps involved in analyzing a urine specimen. Analysis and documentation of the physical properties of urine make up the first procedural step. This analysis includes observation and documentation of the color, appearance, and specific gravity of the urine specimen. Urine odor may also be noted in some cases, but is not part of the standardized reporting procedure. The physical characteristics of urine may provide information for the health-care provider about kidney function and the status of other body systems. Many times the abnormal color or appearance of a urine specimen is the first sign of a serious disorder.

Urine Color

Normally, urine specimens may appear pale yellow (also known as straw) to amber in color. **Urochrome** is the substance that provides the traditional yellow color observed in urine. Urochrome is a pigment that is released continuously as a product of hemoglobin breakdown and normal metabolism. Because the amount of urochrome is consistent, the intensity of the yellow color in the specimen will vary with the patient's hydration status. Urine specimens that are more concentrated (such as the first morning void specimen) will have a darker yellow color than those that are obtained later in the day, as the patient becomes more hydrated.

Test Your Knowledge 20-1
How is urochrome related to urine color?
(Outcome 20-1)

The terms used to describe the color of urine may vary slightly among laboratories, but it is important that they are standardized within each facility. Simple descriptive terms are best, such as straw, yellow, dark yellow, and amber. Employees should be trained to recognize these colors so that the reporting methods are consistent. It is also important to evaluate urine color on a well-mixed specimen in a clear container, so that the true color can be evaluated. Figure 20-1 shows various colors of urine specimens.

A noticeable change in urine color may be caused by a pathological condition. However, certain foods, vitamins, or medications may also cause an abnormal color. Table 20-1 lists potential causes for various urine colors. The most commonly encountered abnormal colors and their significance are listed here:

- *Dark yellow to amber:* Although these colors are technically within the normal range, sometimes urine will

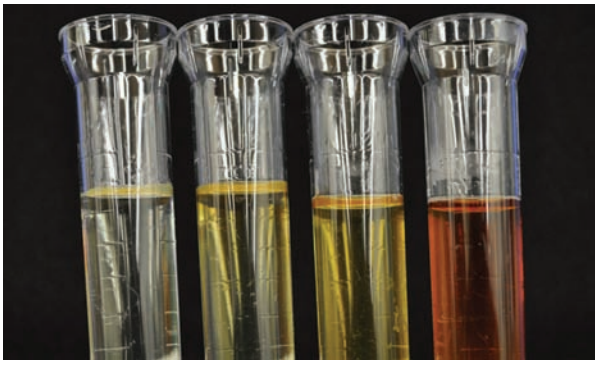

Figure 20-1 Urine color: *left to right,* straw; yellow; dark yellow; amber.

TABLE 20-1

Urine Colors and Possible Causes

Urine Color	Possible Causes	More Information
Colorless to pale yellow	Recent fluid intake, diabetes mellitus, polyuria	Pale yellow and/or colorless urine are not necessarily abnormal, but may be present in some disease states
Dark yellow	First morning specimen, recent strenuous exercise	Usually normal; fluorescent, dark yellow urine may be due to ingestion of multivitamins
Yellow/brown/amber	Hepatitis, dehydration	
Green	Presence of *Pseudomonas* bacterial species in urine specimen	The culture reports will confirm the presence of this bacterium type.
Blue or green	Certain types of drugs or dyes used for contrast studies; ingestion of Clorets breath mints	
Pink or red	Red blood cells, myoglobin, ingestion of fresh beets or fresh blackberries	
Brown or black	Hemoglobin in urine that has been standing or melanin	Melanin may be present in patients with melanoma

appear dark yellow or amber when bilirubin is present in the specimen. This may indicate liver dysfunction, such as in hepatitis, or other conditions that lead to increased red blood cell destruction. Recent strenuous exercise may also cause urine to appear dark yellow.

- *Orange:* Urine may also appear orange when bilirubin levels are elevated in the specimen. However, a more common cause of bright orange urine is the presence of Pyridium (phenazopyridine), a medication used to treat recurrent urinary tract infections. This medication produces a potent orange pigment that will interfere with many of the testing procedures included in a urinalysis.
- *Red to brown:* The presence of red blood cells **(hematuria)** in the urine specimen may cause it to appear pink, red, or even brown. Urinary tract infections or renal dysfunction may allow red blood cells to enter the urinary tract. The presence of hemoglobin from increased red blood cell lysis elsewhere in the body also leads to red urine specimens, although the urine is clearer than it is when intact red blood cells are present. **Myoglobin** is a by-product of muscle destruction, and elevated levels may also cause the urine to appear reddish in color. Many medications, fresh beets, and fresh blackberries may also contribute to a red urine specimen.
- *Blue or green:* Urinary tract infections caused by *Pseudomonas* bacterial genus may cause the patient's urine to appear green in color. Various medications may also cause a green or blue hue.
- *Fluorescence:* Multivitamins may cause the urine to "fluoresce," although the urine color may be normal.
- *Brown or black:* Hemoglobin may turn brown in acidic urine that has been left standing for an extended period of time. Brown or black urine may be evident for patients that have melanoma. This is caused by the presence of melanin and melanogen in the specimen.

Test Your Knowledge 20-2

Ms. Jones has had recurrent urinary tract infections for most of the past year. Her physician has prescribed a medication to treat these infections. On Monday morning, Ms. Jones takes her first dose of the medication. Later that afternoon, Ms. Jones goes to the restroom and is alarmed to see that her urine is bright orange.

Is the bright orange color indicative of a serious problem?

What is the probable cause of the abnormal color in her urine? (Outcome 20-2)

Urine Clarity

The word **clarity** is used to describe the transparency of the urine specimen. A normal urine specimen will be clear when examined in a clear container. The color and the clarity of the specimen are usually evaluated at the same time. As was explained in the section Urine Color, the terms used to describe urine clarity need to be standardized within a laboratory. Common descriptive terms used include clear, hazy, cloudy, or turbid. Figure 20-2 shows the difference in specimen urine clarity. When evaluating urine clarity, it can be helpful to use an object or print behind the specimen; the visibility of the object or print when viewed through the urine specimen will help to determine the clarity that should be reported. Common terms used to describe urine clarity and their clinical significance are listed here:

- *Clear:* A clear specimen has no visible particles and is transparent. Most normal urine specimens appear clear.
- *Hazy:* Hazy specimens may have a few particles floating in the specimen and are not completely transparent. However, it is still relatively easy to see through a hazy specimen. Mucus and other cellular elements may cause the urine to appear hazy.
- *Cloudy:* A cloudy specimen will be difficult to see through because of the particles that are suspended in the specimen. Recently voided specimens may appear cloudy because of **pyuria** (the presence of white blood cells or pus), bacteria, mucus, red blood cells, sperm, yeast, or casts in the specimen. Fat particles suspended in the urine specimen will cause it to be excessively cloudy. Contamination caused by vaginal creams may also cause the urine specimen to appear cloudy. In addition, specimens that have been allowed to stand at room temperature for extended periods of time may develop a cloudy appearance because of the increase in the number of bacteria present in the specimen.

Refrigerated specimens may also become cloudy; this is due to the presence of urates and phosphates in the urine specimen. When the specimen is allowed to cool, these waste products become visible as diffuse (amorphous) particles throughout the specimen, known as **amorphous urates** or **amorphous phosphates.** Amorphous urates are present in acidic urine, and may cause the cloudy urine to appear slightly pink. If there are so many amorphous urates present that the specimen cannot be viewed under the microscope, the urine can be warmed to 60°C to clear the urine. Alkaline urine will result in the formation of amorphous phosphates, with a white, cloudy appearance. The amorphous phosphates can be cleared from the specimen if needed for microscopic examination by adding a drop of acetic acid to the specimen.

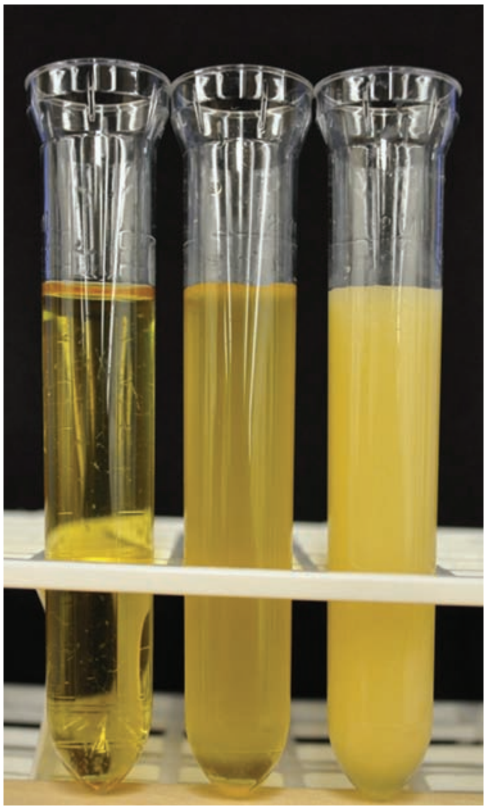

Figure 20-2 Urine clarity: *left to right,* clear; hazy; cloudy.

Urine clarity provides clues concerning the type and number of formed elements present in the specimen. A cloudy specimen has a high concentration of formed elements, whereas a clear specimen usually has a low concentration. The chemical analysis and microscopic examination provide additional information about any chemicals or structures present in the specimen. A clear urine specimen is not always normal, as there may be abnormal chemical constituents present that don't cause the urine to appear abnormally cloudy.

Test Your Knowledge 20-3

Why is it important to place print or something similar behind the urine specimen when evaluating the urine clarity? (Outcome 20-3)

Specimen Volume

The volume of urine produced in a 24-hour period will vary depending on factors such as the hydration status, general renal health, and dietary habits of a patient. The normal adult urine volume for a 24-hour period is between 750 and 2,000 mL per day. Terms used to describe abnormalities in urine volume or urinary frequency are listed below:

- **Anuria:** The absence of urine production
- **Urinary frequency:** An increase in the urge to urinate
- **Oliguria:** Diminished urinary output of less than 400 mL per day
- **Polyuria:** Excessive urination
- **Enuresis:** The inability to control the flow of urine. This term is used for patients who have involuntary urination after the age at which bladder control should have been established.
- **Nocturia**: Excessive or frequent urination at night
- **Urinary incontinence**: Involuntary urination

To perform a complete urinalysis, most laboratories prefer a specimen of approximately 50 mL; the minimum acceptable volume is usually 12 mL. Those performing the testing procedures on urine specimens with a volume less than 12 mL will need to be mindful of the limited availability for rechecking the results of the analysis if needed.

Test Your Knowledge 20-4

A patient has been ill with vomiting and diarrhea for several days. His urine volume over 24 hours is likely to be:

a. Increased
b. Decreased (Outcome 20-4)

Urine Odor

Routine urinalysis procedures do not include the documentation of urine odor. A normal, recently voided urine specimen has an aromatic odor that is not unpleasant. Abnormal urine odors are generally not considered to be clinically significant, but they may still be associated with certain conditions or characteristics of the urine specimen. Urine specimens that are left at room temperature for an extended period of time will develop an ammonia odor as the urea present in the specimen degrades. Bacterial infections of the urinary tract will result in a urine specimen with a strong foul-smelling odor. A urine specimen with a fruity, sweet odor may indicate the presence of ketones in the urine in a diabetic patient. Other metabolic disorders may give the urine a characteristic odor. For example, patients with phenylketonuria will excrete urine with a characteristic mousy or musty odor. Ingestion of certain foods (such as asparagus, garlic, or onions) can also cause the urine specimen to have an unusual smell.

Test Your Knowledge 20-5

True or False: A urine specimen with a strong odor is always indicative of a urinary tract infection.

(Outcome 20-4)

POINT OF INTEREST 20-1

The odoriferous asparagus

The odor associated with a urine specimen may be attributed to its volume and the concentration of the chemicals present in the urine specimen. For example, within a short period of time after ingesting asparagus, many people have reported that their urine has a "sulfur" smell. It is thought that this distinctive odor is caused by the breakdown of amino acid compounds in the asparagus as it is digested.

However, there is a continuing debate over the production of odoriferous urine because it does not appear to be universal. Some studies have shown that less than half of the population actually excretes sulfurous-smelling urine after asparagus ingestion. It was thought that genetically the two groups differed, and they were labeled as "excretors" and "nonexcretors," based on whether they smelled the sulfur in their urine after asparagus ingestion. This belief is still held by many. However, other scientists believe that all individuals actually excrete the sulfurous odor, but only some individuals are able to detect the odor. Those who cannot detect the smell are not aware of its presence, and may have been misclassified as "nonexcretors" in the original studies. There is not enough scientific evidence available to prove or disprove either hypothesis.

Specific Gravity

Specific gravity measurements are one of the physical examinations performed during a routine urinalysis. Urine specific gravity is defined as the density (or weight) of the urine specimen as compared to the density (weight) of distilled water at the same temperature. The density of the urine specimen is dependent on the amount and size of the dissolved substances present in the urine specimen. These include such substances as glucose, proteins, and electrolytes. The kidneys are responsible for selective reabsorption of essential chemicals and fluid after the blood is filtered. Abnormal specific gravity values may be an early indication of renal dysfunction. Low specific gravity urine is considered to be dilute with very few dissolved substances, whereas urine with a high specific gravity indicates that the urine is concentrated with increased amounts of dissolved substances.

Test Your Knowledge 20-6

True or False: Urine specimens with high specific gravity results are always cloudy. (Outcome 20-5)

Because urine specific gravity is measured against distilled water, the results are reported using units that reflect this comparison. Urine specific gravity is reported in units beginning with 1.000 (for distilled water) to amounts above 1.030 (for highly concentrated urine specimens). Urine concentration (as measured by the specific gravity of the urine specimen) may be elevated with congestive heart failure, dehydration, adrenal gland dysfunction, **glycosuria** (the presence of glucose in the urine), high levels of urine protein, or recent use of high molecular weight IV fluids. The urine specific gravity may also be elevated in situations in which the patient has become dehydrated as a result of excessive vomiting or diarrhea. Decreased urine specific gravity may be present in diabetes insipidus, excessive fluid intake, or renal failure. The normal urine specific gravity is 1.005 to 1.030.

Test Your Knowledge 20-7

A urine specific gravity is documented on the patient's chart as 1.25. Is this correct? (Outcome 20-7)

Specific gravity may be measured in several ways, but there are three methods used most often in physician office laboratories. These include the use of a **refractometer**, a **urinometer**, or a reagent dipstick.

Refractometer Measurement Technique

An instrument called a refractometer measures the **refractive index** of a solution. This is a measurement of the velocity of light in air as compared to the velocity of light in a solution. The number of dissolved particles present in the solution affect the velocity and the angle at which the light passes through a urine specimen. Clinical refractometers (see Fig. 20-3) used for urine analysis use a prism to direct one distinct wavelength of light for comparison to a special scale. The concentration of the urine specimen has a direct effect on the angle of the light as it passes through to the prism. The refractive index is not identical to the specific gravity of urine, but refractometers are calibrated to provide specific gravity readings.

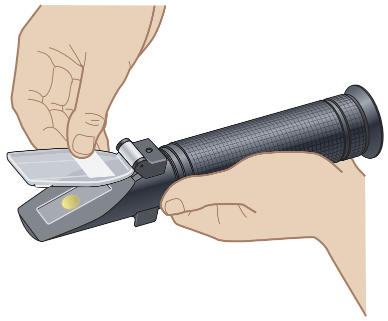

Close the daylight plate gently.

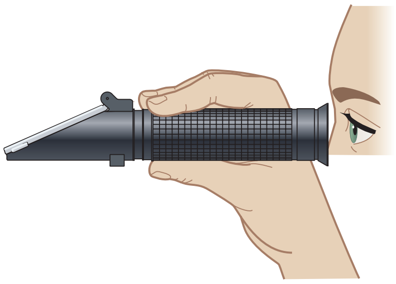

Look at the scale through the eyepiece.

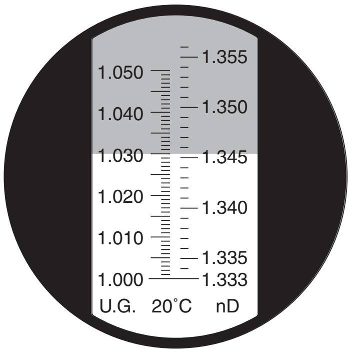

Read the scale where the boundary line intercepts it.

Figure 20-3 Steps to be taken when determining urine specific gravity with a refractometer. *Courtesy of NSG Precision Cells. www.precisioncells.com*

Refractometers are portable, relatively inexpensive, and easily calibrated. This method of measurement uses only a drop of urine, which can be beneficial as compared to other methods. Refractometers must be calibrated regularly with distilled water to achieve a result of 1.000. Five percent NaCl may also be used to check the calibration, and it should read 1.022. Urine quality control samples should also be used to verify that the refractometer is working correctly and that the operator knows how to read the results appropriately. Procedure 20-1 provides detailed instructions for the use of a refractometer to measure urine specific gravity.

Urinometer Measurement Technique

A urinometer is a device that uses a weighted float attached to a scale. This weighted float will displace a volume of liquid relative to its weight. When used with distilled water, the weight will sink until the calibrated scale reads exactly 1.000. Because urine has more dissolved substances present than does distilled water, the urinometer float will displace less fluid. This will result in a higher reading on the urinometer scale. Figure 20-4 includes an example of a urinometer.

Urinometers are rarely used for specific gravity testing. The Clinical and Laboratory Standards Institute does not recommend their use for urine specific gravity, but they may still be used in small physician office laboratories. There are several disadvantages to the urinometer method, including the large volume of urine required for the procedure (10 to 15 mL), the fragility of the weighted float/scale, and the need for a specific container that allows for the displacement of the fluid during the process. In addition, the specific gravity is measured at the **meniscus** (the bottom of the curve at the top of the urine column) of the fluid after the float is added to the urine specimen; the reported results may sometimes be inaccurate because of problems with reading the meniscus properly.

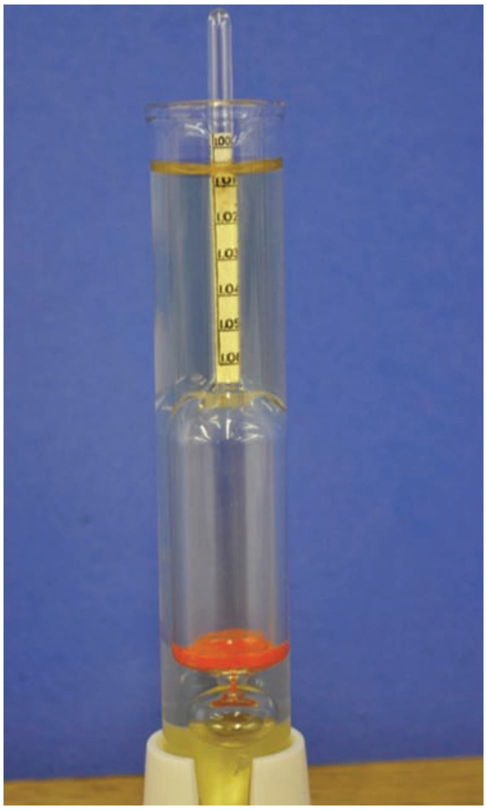

Figure 20-4 A urinometer. The result is read at the top of the urine column (at the meniscus) on the measurement scale printed on the urinometer.

Reagent Strip Methodology

Specific gravity measurements performed by using reagent strips are another indirect measurement of the amount of dissolved substances in the urine specimen. Reagent strip reactions are based on the release of ionic solutes in the specimen. These ions are released for measurement when exposed to the reagent pad on the strip as it is dipped into the urine specimen. A color change occurs on the reagent pads, which can be measured in 0.005 intervals from 1.000 to 1.030. Many facilities have adopted this method of measuring specific gravity because it does not involve an additional instrument for the procedure, as the results are read for specific gravity while reading those for other chemical analytes on the reagent strip.

Test Your Knowledge 20-8

What are three advantages of using the refractometer method for specific gravity testing? (Outcome 20-6)

POTENTIAL SOURCES OF ERROR

Numerous preanalytical variables may affect the physical characteristics of urine specimens. The assessment procedures for color, clarity, and specific gravity may also be performed incorrectly. The following are some of the most common sources of error:

- *Incorrect or missing identification on the specimen:* Urine specimens must be labeled with the patient's name, birth date, and specimen number or other unique identifier. The date and time of collection must also be noted on the requisition form and on the specimen.
- *Inappropriate specimen storage:* Urine specimens are to be analyzed within 2 hours of collection. If this is not possible, the specimen must be refrigerated or appropriately preserved within 2 hours of collection. Ideally, the color and clarity of the specimen should be determined, but if they were not, the specimen must be allowed to reach room temperature again before assessment.

Procedure 20-1: Observation and Documentation of Urine Physical Properties

TASK

The student will perform an assessment of the physical properties of a urine specimen, including the color, appearance, and specific gravity. A handheld refractometer will be used for the specific gravity measurement.

CONDITIONS

- Gloves, laboratory coat, and protective eyewear
- Freshly voided urine specimen
- Transfer tube
- Refractometer
- Graph paper to be used as "print" for urine clarity comparison
- Distilled water
- Transfer pipettes
- Kimwipes
- Permanent fine-tip pen for marking identification on the transfer tube
- Black pen for documentation on the patient's chart
- Patient's chart
- Biohazardous waste container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I. Anatomy and Physiology, #13. Perform Urinalysis

ABHES Standards

- 10. Medical Laboratory Procedures, b. Perform selected CLIA-waived tests that assist with diagnosis and treatment, #1: Urinalysis

Procedure	Rationale
1. Wash hands	Hand washing breaks the cycle of infection.
2. Put on gloves.	Wearing proper personal protective equipment (PPE) protects employees from potential exposure to pathogens present in the specimen.
3. Verify test order and assemble necessary equipment and specimen.	Verification of test order eliminates potential errors, and organization of supplies saves time.
4. Verify specimen labeling and identification.	Urine specimens need to be labeled with the name of the patient, the birth date, and the account or specimen number. The date and time of the collection must also be documented.
5. Verify that specimen collection was less than 1 hour before testing or that it was refrigerated and/or preserved appropriately after collection.	Specimens that are allowed to stay at room temperature for more than 1 hour after collection without the addition of preservative may yield inaccurate results because of changes in pH, bilirubin or urobilinogen concentration, glucose levels, and bacterial contamination.
6. Allow any refrigerated specimens to come to room temperature before proceeding.	Specimens that are colder than room temperature may exhibit crystallization of amorphous urates or phosphates, which can interfere with measurement of all the physical characteristics of the urine specimen.

Procedure	Rationale
7. Mix specimen well.	Urine specimens must be mixed well to judge their color and appearance accurately, as well as the specific gravity reading. Without mixing, the formed elements in the specimen will settle to the bottom of the container and may not be considered when determining whether a specimen has an abnormal color or clarity. These substances will also not be reflected in the specific gravity reading.
8. Observe the color of the specimen by visual observation. Report as straw, light yellow, yellow, dark yellow, or amber.	The terms used to describe specimen color need to be standardized throughout the laboratory.
9. Place a printed piece of paper behind the specimen to measure the clarity. Report as clear, hazy, cloudy, or turbid.	Abnormal urine appearance may be linked to a pathological condition. Use of standardized terms is important throughout the laboratory.
10. Pour a small amount of the urine specimen (less than 50 mL) into a transfer tube that has been labeled with the patient's name and/or ID number.	The urine specimen may need to have a culture performed on it; this decision is often not made until the routine urinalysis has been completed. By pouring off specimen into a transfer tube rather than inserting the pipette directly into the urine container, contamination is avoided.
11. Using distilled water, verify the function of the refractometer: a. Open the refractometer cover and place one drop of distilled water on the prism under the cover using a clean transfer pipette. b. Close the cover, and verify that the sample has spread over the entire prism surface. c. Look at the scale through the eyepiece; bright light in the vicinity helps with visualization. d. Read the scale where the line intercepts it. This is where the colors change with a sharp line of division. e. Wipe the sample away from the face of the prism with a Kimwipe laboratory tissue and water.	The refractometer should read 1.000 when the specific gravity of distilled water is tested. If this is not the reading achieved, follow the manufacturer's directions for calibration of the refractometer before testing the urine specimen. 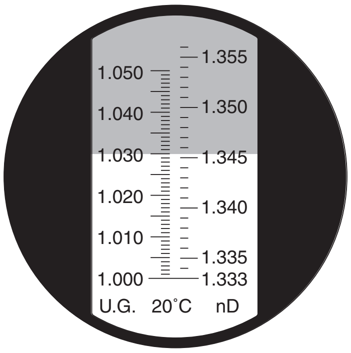 1.050 1.040 1.030 1.020 1.010 1.000 1.355 1.350 1.345 1.340 1.335 1.333 U.G. 20˚C nD Read the scale where the boundary line intercepts it.

Continued

Procedure 20-1: Observation and Documentation of Urine Physical Properties—cont'd

Procedure	Rationale
12. Following laboratory protocol, test a quality control specimen using the refractometer. Use the same procedure outlined in step 10. Chart the results on the quality control log.	The frequency of the quality control measurement will be dependent on the number of tests performed in most laboratories. Be certain that the quality control material reading falls within the acceptable range; if the results are outside the acceptable range, do not use the refractometer for testing the specific gravity of the urine specimen until the accuracy of the instrument has been verified.
13. Using a clean transfer pipette, place one to two drops of the urine specimen (from the transfer tube) on the face of the prism. Following the procedure outlined in step 10 above, read the specific gravity of the specimen.	Reference ranges for urine specific gravity are approximately 1.005 to 1.030.
14. Dispose of the urine left in the transfer tube by pouring it down the sink and flushing it with plenty of water. Dispose of the tube and transfer pipette used for the urine in a biohazard disposal bag. If the urine is not to undergo additional testing immediately, refrigerate it or preserve it accordingly.	
15. Clean the instrument using distilled water and a Kimwipe.	Do not allow the urine specimen to dry on the prism. Do not use harsh cleaners or abrasives to clean the prism or it will become scratched and unusable. Refer to the manufacturer's instructions for more cleaning procedures.
16. Remove gloves and sanitize hands.	Gloves should always be removed immediately after a procedure is performed. Hand sanitization is appropriate before and after gloves are worn.
17. Document results for patient specimen on the chart (if available) and testing log sheet. Be certain to use the correct digits for the documentation.	Documentation needs to occur immediately after the specimen testing is complete.

Date	
10/24/2013	*Urine color: yellow, Clarity: clear, Specific gravity: 1.015*
11:58 a.m.	*Connie Lieseke, CMA (AAMA)*

- *Insufficient specimen volume:* There must be a minimum of 12 to 15 mL of urine submitted for a routine urinalysis to be performed.
- *Use of nonstandardized terms for reporting:* Although there may be various terms used to describe the color and/or clarity of the urine specimen, all laboratory personnel must use the same standard terms within their institution. This allows for normal ranges to be established, and is important when reporting results.
- *Lack of personnel training:* Although the assessment of the urine physical properties is a relatively simple process, there must be documented training for those who perform this task.
- *Use of noncalibrated equipment:* Refractometers must be calibrated regularly using distilled water, and quality control materials should be tested to verify that they are working appropriately.

Test Your Knowledge 20-9

A medical assistant performing urinalysis reported the urine clarity for a specimen as partly hazy. If the laboratory where she works uses the same criteria as that noted in this textbook for specimen clarity, is this the correct way to report this test result? (Outcome 20-8)

SUMMARY

The physical examination of urine will provide information about the color, clarity, and specific gravity of the specimen. Occasionally, an unusual odor may also be noted during this part of the urine analysis. Results that are outside the reference range for these physical characteristics may be the result of pathological conditions, or they may be caused by medication use or ingestion of certain foods. Accurate results must begin with a specimen that has been collected properly, identified appropriately, and preserved or refrigerated within 2 hours of collection. When reporting the color and clarity of a specimen, it is important to use terms that have been approved by the facility so that the reporting methods are standardized. Instruments used for specific gravity readings must be calibrated and quality control material must be tested at regular intervals to verify the accuracy of the measurement. Medical assistants must also be aware of potential sources of error when performing an assessment of the physical properties of a urine specimen to avoid inaccurate results.

TIME TO REVIEW

1. Which of the following words is used to describe blood in the urine? Outcome 20-1
   - a. Glycosuria
   - b. Hematuria
   - c. Pyuria
   - d. Oliguria
2. Which of the following words is used to describe a slightly cloudy urine specimen? Outcome 20-1
   - a. Hazy
   - b. Clear
   - c. Turbid
   - d. Amber
3. Which of the following is an instrument used to measure the specific gravity of a urine specimen? Outcome 20-1
   - a. Hemoglobinometer
   - b. Urochrome meter
   - c. Refractometer
   - d. None of the above
4. Which of the following may cause urine specimens to exhibit an abnormal color? Outcome 20-2
   - a. Ingestion of rare meat
   - b. Ingestion of broccoli
   - c. Aspirin use
   - d. Ingestion of fresh beets
5. Which of the following is used to describe urine clarity when the specimen has particulate matter suspended in it? Outcome 20-3
   - a. Hazy
   - b. Cloudy
   - c. Turbid
   - d. All of the above
6. A urine specimen with an ammonia-like odor may be caused by: Outcome 20-4
   - a. Bacteria in the specimen
   - b. Asparagus ingestion
   - c. Multivitamin use
   - d. Dye studies
7. A urine specimen with a specific gravity reading of 1.035 will demonstrate: Outcome 20-5
   - a. High levels of dissolved substances
   - b. Low levels of dissolved substances
   - c. An amber color
   - d. A foul odor

8. Which of the following is not a potential source of error for urine testing? Outcome 20-8

- **a.** A urine specimen received without a patient name
- **b.** Observation and documentation of urine color and clarity upon specimen receipt
- **c.** Observation and documentation of urine color and clarity immediately after taking the specimen from the refrigerator
- **d.** Use of a container that was not provided by the laboratory for the urine collection

Case Study 20-1: Too much to do

Cindy Collier, CMA (AAMA), is working in the office laboratory for a busy internal medicine clinic. She is about to perform a urinalysis on a specimen that was collected from a patient who is in to see the physician because of blood in her urine. Cindy verifies the identification on the specimen and places it on the counter in the laboratory area. Before she has a chance to do anything else, she is called away to perform a venipuncture and PT/INR on another patient. Approximately 1 hour later, Cindy hurries into the laboratory so that she can start to analyze the urine specimen. She picks up the cup, and documents that the specimen is clear and appears yellow in color. She also notes that there is a red "ring" around the bottom interior of the cup.

1. What did Cindy forget to do before assessing the color and clarity of the specimen?
2. Is her assessment of the specimen's being yellow and clear a correct result for this specimen?

RESOURCES AND SUGGESTED READINGS

Clinical and Laboratory Standards Institute, Urinalysis: Approved Guideline, ed. 3. CLSI document G16-A3. Wayne, PA, 2009
Approved laboratory standards for collecting, processing, and testing urinalysis samples

Chapter 21

Chemical Examination of Urine and Feces

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Urine Analytes and Their Clinical Significance
    - Bilirubin
    - Blood
    - Glucose
    - Ketones
    - Leukocytes
    - Nitrite
    - pH
    - Protein
    - Urobilinogen
    - Specific Gravity
- Potential Sources of Error
- Safety Precautions
- Quality Control Procedures
- Urine Testing Methods
- Confirmatory Urine Testing
- Fecal Occult Blood Testing
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

21-1 Define the key terms.

21-2 Differentiate various disease states related to abnormal urine chemistry results.

21-3 Identify abnormal values for the analytes tested with urine chemistry analysis.

21-4 List potential sources of error for urine chemistry testing, and describe how these errors may be avoided.

21-5 Describe appropriate safety precautions implemented when testing urine.

21-6 Compare the testing methods available for urine chemistry analysis.

21-7 Perform CLIA-waived urine chemistry analysis using a manual and an automated testing method.

21-8 Provide examples of confirmatory tests performed on urine specimens.

21-9 Explain the importance of fecal occult blood testing.

21-10 Detail the necessary patient preparation for fecal occult blood specimen collection.

21-11 Perform a CLIA-waived fecal occult blood test.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

III.P.III.2. Practice Standard Precautions
I.P.I.14. Perform Urinalysis

ABHES Standards

- Apply principles of aseptic techniques and infection control.
- Use standard precautions.
- Perform CLIA-waived tests that assist with diagnosis and treatment, Urinalysis.
- Instruct patients in the collection of a fecal specimen.

KEY TERMS

Acidosis
Alkalosis
Bence Jones protein
Bilirubin
Conjugated bilirubin
Fecal occult blood testing (FOBT)
Glucosuria
Glycosuria
Guaiac method
Hematuria
Hemoglobinuria
iFOB
Intravascular lysis
Jaundice
Ketones
Ketonuria
Leukocyte esterase
Leukocytes
Microalbuminuria
Multiple myeloma
Myoglobin
Myoglobinuria
Nitrite
Proteinuria
Semiquantitative
Sulfosalicylic acid precipitation test
Urobilinogen

The chemical testing of urine specimens is the second component included in a complete urinalysis. There have been numerous changes to chemical urine testing procedures through the years, and the most significant changes occurred with the development of reagent test strips. Since the 1950s it has been possible to test numerous chemical analytes at once with disposable reagent test strips. These strips are made of a plastic with absorbent pads attached. (An example is shown in Fig. 21-1.) The pads are impregnated with various chemicals, and each pad is designed to change color as it reacts with a specific analyte present in the urine specimen. The resulting colors on the pads are interpreted by comparing the individual pad to a chart supplied with the reagent strips. An example of a chart used for comparison is shown in Figure 21-1. By comparing the color changes with the reference chart it is possible to perform a **semiquantitative** measurement, providing an approximate value for each of the chemicals being tested. The results may be reported as the milligrams per deciliter present, or by using a semiquantitative reporting method of trace, 1+, 2+, 3+, or 4+. Some of the results may also be reported as positive or negative, which is an example of a *qualitative* test result.

URINE ANALYTES AND THEIR CLINICAL SIGNIFICANCE

A routine urinalysis usually includes testing for bilirubin, blood, glucose, ketones, leukocytes, nitrites, pH, protein, and urobilinogen. The reagent strips most commonly used include testing for specific gravity as well. Many of these chemical substances are normally present in the urine specimen, but the amount of the individual analyte present may change with certain disease states. Other chemicals are not present in the urine specimen normally, and their detection may

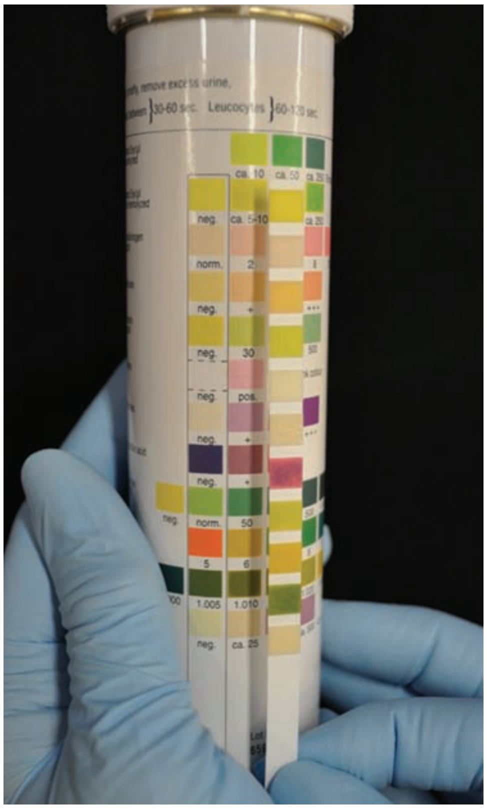

Figure 21-1 Urine chemical reagent testing strip and chart used for comparison when reading urine chemical reagent strip results.

be clinically significant. Chemical analysis of urine specimens may detect dysfunction of carbohydrate metabolism, pH imbalances, red blood cell hemolysis, and liver or kidney problems. The performance of urine chemical testing in the physician office laboratory is beneficial for the patient because abnormalities are quickly detected and appropriate treatment or follow-up testing can be taken care of immediately while the patient is still in the presence of the health-care provider.

There are two primary manufacturers for urine reagent strips. Multistix is manufactured by Siemens Medical Solutions Diagnostics, and Chemstrip is manufactured by Roche Diagnostics. The strips are available with various types of reagent pads, depending on the needs of the health-care provider ordering the urine tests. If testing is performed with an automated system, there may be a recommendation for one brand rather than the other. The manufacturer's insert will include instructions for use and will also include interfering substances for each analyte.

Test Your Knowledge 21-1

True or False: All analytes tested by the urine reagent strips are present in measurable quantities in a normal urine specimen. (Outcome 21-3)

Bilirubin

Normally, **bilirubin** is not present in the urine. The presence of bilirubin in the urine specimen may be an early indication of liver disease or bile obstruction. Bilirubin is a by-product of red blood cell destruction. As red blood cells are broken down at the end of their 120-day life cycle, hemoglobin is released. This decomposition of red blood cells occurs in the spleen and liver. As the hemoglobin molecules released from the red blood cells degrade further, they are split into smaller components. Bilirubin is created from hemoglobin as part of this degradation process. To travel through the bloodstream, bilirubin must be carried by molecules of albumin (a type of protein) because the bilirubin is not water soluble. The bilirubin-albumin molecule is too large to enter the urine within the kidneys. Instead, the bilirubin-albumin complex is returned for processing in the liver, where the bilirubin becomes water soluble and is separated from the albumin. This water-soluble bilirubin is called **conjugated bilirubin.** Conjugated bilirubin is not usually detected in the urine because it is passed from the liver directly into the bile duct to be secreted into the intestines. The intestinal bacteria further alter the bilirubin, changing it into a compound known as **urobilinogen.** Bile obstruction may lead to the presence of conjugated bilirubin in the urine, and liver dysfunction may also cause bilirubin to be detected in the urine specimen.

Bilirubin is has a strong yellow-orange color. When it becomes elevated in the bloodstream, the patient's skin, sclera of the eye, and plasma reflect this intense yellow color, known as **jaundice.** Excessive hemolysis, liver dysfunction, or bile obstruction may cause jaundice. Just as the bilirubin is elevated in the bloodstream with bile obstruction or liver dysfunction, the levels may also be elevated in the urine specimen, allowing for detection with a chemical analysis.

Test Your Knowledge 21-2

What clinical dysfunction will cause a urine specimen to test positive for bilirubin? (Outcome 21-2)

Blood

There are three reasons that a urine sample may show a positive result for blood. The presence of intact red blood cells, known as **hematuria,** will cause a positive reaction. Hematuria will often present with a cloudy, red urine specimen with large amounts of red blood cells present. Microscopic examination of the urine specimen will detect red blood cells in the urine sediment.

The presence of hemoglobin without the presence of intact red blood cells **(hemoglobinuria)** will also cause a positive result for blood in a urine sample. Hemoglobinuria may be the result of red blood cell lysis that occurs in the urinary tract, or it can be caused by **intravascular lysis,** which is the breakdown of red blood cells within the vessels. Intravascular lysis occurs in hemolytic transfusion reactions. Hemoglobinuria may also be present with severe burns, malaria, hemolytic anemias, or with some spider or snake bites that cause hemolysis within the blood vessels. The presence of blood in the urine specimen is not normal, except in the case of specimen contamination with menstrual blood or trauma caused by catheter insertion.

Finally, the presence of **myoglobin** in the urine, known as **myoglobinuria,** will result in a positive result for blood. Myoglobin is present in muscle tissues, where it serves as an oxygen-storing molecule. When present in urine, it may cause the urine to appear red-brown in color, but the specimen clarity will remain clear. Myoglobin may be present in the urine when there is trauma to the muscles, convulsions, alcoholism, electrocution, or excessive exercise. Myoglobin is actually toxic to the kidneys; therefore, high amounts present in the urine may be an indicator of concurrent kidney damage. There are many testing methods and comparisons between urine and plasma that may differentiate a positive urine blood result due to hemoglobin from one that is due to myoglobin, because there are very few intact red blood cells present in the urine in both cases.

Test Your Knowledge 21-3

Which of the following will result in a positive test for blood in the urine specimen?

a. Intact red blood cells
b. Elevated levels of iron in the blood
c. Pus in the urine
d. None of the above (Outcome 21-3)

Glucose

The presence of glucose in the urine is not a normal result. Blood glucose is normally regulated through hormonal changes and kidney function, including the reabsorption of almost all the glucose filtered out by the glomeruli. However, when the blood glucose concentration becomes excessively elevated (at levels between 160 and 180 mg/dL) the glucose concentration is too high for most of it to be reabsorbed by the tubules, and the glucose "spills" into the urine. **Glycosuria** is a word used to describe the presence of sugar in the urine; **glucosuria** is used when the sugar in the urine is identified as glucose. Glucose is the most common cause of glycosuria, and most reagent strip methods test positive only in the presence of glucose; they do not react with other sugars. However, galactose, lactose, fructose, and pentose are other sugars that may be present in the urine. Point of Interest 21-1 provides more information about additional test procedures that may be performed on the urine specimen to allow for detection of sugars other than glucose. Elevated levels of galactose in the urine of infants may be indicative of a serious condition that needs immediate intervention; therefore, alternative testing methods are often used for urine testing on infants to provide an opportunity to detect galactosuria.

Urine glucose testing can be an invaluable part of diabetes screening. Because the symptoms of diabetes are not clearly defined and do not present themselves in the same way for each individual, there are many patients who are unaware that they are diabetic. For this reason, glucose testing is one of the most common tests performed on urine specimens. Because the blood glucose concentration can fluctuate throughout the day, urine glucose testing done for the purpose of diabetes screening should be performed on fasting specimens. Diabetic patients may also use urine glucose testing to monitor their disease progress; this testing is usually performed on postprandial specimens. Glucosuria may also be present when the reabsorption of the kidney tubules has been compromised, as in the case of renal failure.

Test Your Knowledge 21-4

True or False: All diabetic patients will have a positive urine glucose result at all times. (Outcome 21-2)

POINT OF INTEREST 21-1

Galactosemia

The glucose test pad on the urine chemical reagent strip is specific for glucose only; other types of sugars that may be present in the urine are not detected by the enzymatic chemical reaction that occurs when this pad is exposed to the urine specimen. However, for infants, this reagent pad test does not provide enough information. Urine specimens that test negative for an infant should have an additional test performed that will detect other types of sugars that may be present in the urine. When an infant ingests lactose (a type of sugar commonly found both in breast milk and formula), the lactose interacts with a naturally occurring enzyme, lactase. Lactase splits the lactose molecule into two smaller sugar molecules, glucose and galactose. Glucose is easily used by the body for energy, but galactose must be metabolized further before the body can make use of it. Hereditary galactosemia is a condition in which the infant is not capable of metabolizing the galactose. The galactose levels in the bloodstream begin to climb, causing initial symptoms of convulsions, irritability, lethargy, poor weight gain, jaundice, and vomiting. If the elevated levels of galactose continue to increase, the liver, brain, kidneys, and eyes may be affected irreversibly. Detection of galactosemia within the first few days of life is critical, as the treatment involves the complete removal of this sugar type from the diet, and these changes must begin immediately.

To ensure that elevated levels of other sugars such as galactose (also known as reducing substances) are detected in urine specimens for infants, an alternate form of testing is used. The Clinitest test manufactured by Bayer Corporation uses a copper reduction testing method, in contrast to the enzymatic method used on test strip reagent pads. The copper reduction test method will detect the presence of all types of sugar, not just glucose. In this test, a few drops of the urine specimen are mixed with a few drops of water, and a reagent tablet is dropped into the tube. A violent reaction ensues, and the resulting color of the mixture in the tube is compared to a chart (provided with the tablets) to interpret the amount of reducing substances present in the specimen. If the test result for the Clinitest is positive, it is assumed that the sugar present is not glucose, as this would have also caused a positive result on the reagent pad reaction.

Care should be taken when performing the Clinitest procedure. Gloves, eye protection, and skin protection need to be worn, as the reaction that occurs within the tube involves boiling the mixture and it is possible to cause damage to the skin, eyes, or mucous membranes. It is important to dispose of the mixture properly as well, as it contains sodium hydroxide and citric acid. The procedure and precautions outlined in the package insert should be followed carefully.

Ketones

Ketones are a product of fat metabolism, and their presence in urine is not considered a normal result. When fat is metabolized, the process usually continues to the point at which the ketones are fully broken down and undetectable. However, if the glucose in the body cannot be readily used for energy, then fat metabolism is increased, and ketones may be present in the blood and in the urine. Uncontrolled diabetes mellitus, malabsorption syndromes, excessive reduction in carbohydrate intake, and vomiting may all lead to the presence of ketones in the urine **(ketonuria)** because these conditions don't allow glucose to be used efficiently as a source of energy.

The word *ketones* is used to describe three different products released into the bloodstream as fat is metabolized: acetone, acetoacetic acid, and beta-hydroxybutyric acid. Reagent strips test specifically for acetoacetic acid. The other two ketone types are actually created by acetoacetic acid, so detection of this compound will provide a clinically significant result, indicating the presence of any type of ketone product.

Urine ketones are often monitored in patients with type 1 diabetes mellitus. When the body does not have enough insulin, the glucose present in the bloodstream cannot be used, and the body will begin to metabolize increasing amounts of fat for energy. If the ketones can be detected early (in the blood and in the urine), the patient may be aware of an impending crisis and the dosage of insulin can be regulated to allow for more efficient glucose metabolism.

Leukocytes

Leukocytes are white blood cells. Their presence in the urine in small amounts is not an abnormal result, but when the number of leukocytes becomes elevated, it is indicative of inflammation or infection of the urinary tract. Reagent strips test for the presence of **leukocyte esterase,** a substance present in granulocytic white blood cells (those that have granules in their cytoplasm), such as neutrophils, basophils, and eosinophils. Leukocyte esterase is also present in monocytes. Neutrophils are the white blood cell type that is most commonly elevated in urinary tract infections. Leukocytes may be visualized with microscopic examination of the urine, but because leukocyte esterase is present even with lysed white blood cells, the presence of leukocytes in the urine specimen may be overlooked if only the microscopic results are considered in the patient assessment.

Test Your Knowledge 21-5

True or False: The leukocyte esterase result will always be positive when white blood cells are present in the urine specimen. (Outcome 21-3)

Nitrite

Some types of bacteria are capable of converting nitrate, which is normally present in urine, to **nitrite,** which is not usually detectable. *Escherichia coli*, which is the most common cause of bacterial urinary tract infection, is capable of producing nitrites. Many other gram-negative organisms are also capable of this conversion, such as those of the *Klebsiella, Proteus,* and *Serratia* genera. A positive nitrite result is indicative of the presence of bacteria in the urine. However, there are two factors that must be considered when interpreting this result. First of all, not all bacteria are capable of converting nitrate to nitrite, so a negative result does not necessarily mean that the specimen is bacteria free. Second, the bacteria must be present in the urine specimen long enough to accomplish the conversion, and this usually takes 4 to 6 hours. If a nitrite test is to be used as a screening tool for a potential urinary tract infection, a first morning void specimen should be used for the test, because the urine collected with this type of specimen has been in the bladder all night with the bacteria, allowing for an opportunity for the conversion to occur. Specimens collected at other times of the day may or may not have been in the bladder for a long enough period of time for this result to be a positive one, even if the bacteria present are capable of converting nitrate to nitrite.

pH

The pH of normal freshly voided urine may vary from 4.6 to 8.0. The abbreviation pH stands for "parts hydrogen," and the pH of a substance is the measurement of how acidic or alkaline that substance is, based on the concentration of hydrogen ions present. The lungs and the exchange of ions in the urinary tubules of the kidneys primarily control the pH of the human body. The pH of the blood must be maintained between 7.35 and 7.45.

Urinary pH may be abnormal when the body is in a state of **acidosis** (a blood pH of 7.35 or below) or when the blood pH is elevated above 7.45 (known as **alkalosis).** Metabolic issues not related to kidney function may cause an abnormal urine pH. Kidney dysfunction may also lead to alkaline or acidic urine because the exchange of ions occurs in the renal tubules. Medications or nutritive supplements may be used for patients with chronic urinary tract infections to keep the urine slightly acidic, as this environment is not supportive of bacterial growth. This may cause urine pH results that are outside of the normal range, but will not necessarily be interpreted as abnormal for the patient who is undergoing treatment.

Protein

The presence of protein in the urine **(proteinuria)** is most commonly associated with renal disease. A positive result for protein does not always indicate kidney damage, but there should always be additional testing performed to determine whether the proteinuria is indicative of a pathological condition. Normal urine has very little protein present, as most protein molecules are not allowed to enter the urine through the glomerulus because the molecules are too large to leave the plasma. Any that are filtered out are usually reabsorbed by the renal tubules and are not present in the urine specimen.

Protein may be present in the urine when there is no renal disease or damage. This occurs when the total protein level of the plasma is elevated. This elevation is the result of the increased presence of small protein molecules in the plasma that are capable of passing through the pores in the glomerulus. The kidneys will filter out the excess protein, but when they have reached their capacity for protein reabsorption in the renal tubules, protein will be present in the urine specimen. Muscle trauma, fever, or excessive intravascular hemolysis will elevate the blood protein levels. This is usually a transient result that will not continue once the underlying condition has been resolved.

Patients with **multiple myeloma** (a cancer of the bone marrow) may present elevated amounts of **Bence Jones protein** in their plasma. This specific type of protein will be filtered out of the blood, and when the tubular reabsorption capacity has been surpassed, Bence Jones protein will be present in the urine specimen. It is possible to differentiate Bence Jones protein from other types of protein by taking advantage of the different reaction of this specific type of protein when exposed to heat. The Bence Jones protein will coagulate (clump up) when exposed to temperatures between 40°C and 60°C, but it will dissolve back into the solution when exposed to temperatures of 100°C. Other proteins will coagulate at a similar temperature, but will not dissolve at the higher temperatures. The absence of Bence Jones protein in the urine is not a valid means of ruling out multiple myeloma, as many patients with this disease do not excrete levels high enough to be detected in the urine specimen. Protein electrophoresis is a more specific test used for diagnosis.

The most common type of protein present in the urine as a result of damage to the kidneys is albumin. It may be indicative of damage to the glomeruli or to the renal tubules. The presence of albumin may be a transient condition, as in the case of strenuous exercise or dehydration. Patients with hypertension may also present with proteinuria. Positive protein tests in pregnant women (especially in the last trimester of pregnancy) may be indicative of preeclampsia, a serious condition that must be treated immediately. **Microalbuminuria,** or the chronic presence of small amounts of albumin in the urine, is a common occurrence in diabetic patients. It is an indication that the glucose levels in the blood are not stabilized, and the increased workload required of the kidneys to filter out these large molecules has caused the glomeruli to become damaged. The damaged glomerulus allows protein molecules to leak into the urine specimen, presenting as microalbuminuria.

Most reagent strips detect only the presence of albumin. If other types of protein are suspected, it may be necessary to use other testing methods to detect their presence. The **sulfosalicylic acid precipitation test** (SSA test) is a test that will detect all forms of protein in the urine specimen. In this procedure, 3 mL of 3% sulfosalicylic acid is added to an equal volume of centrifuged urine. The specimen is mixed well with the acid, and the degree of turbidity (cloudiness) is measured. The grade or degree of turbidity is correlated to the probable amount of protein present in the specimen. This test may be useful when there is color interference in the urine specimen and a protein test is desired, or when the urine specimen has a very alkaline pH, as this can interfere with the reagent strip results for protein.

Test Your Knowledge 21-6

Which of these are nonpathogenic causes of proteinuria?

a. Fatigue
b. Fever
c. Vomiting
d. Ingestion of red meat (Outcome 21-3)

Urobilinogen

The presence of a small amount of urobilinogen in the urine is normal. However, increased levels may be indicative of liver disease (such as hepatitis or cirrhosis) or hemolytic disorders. Urobilinogen is produced in the intestines by the intestinal bacteria as bilirubin is broken down. Once formed, either it will be reabsorbed into the bloodstream (where it will eventually pass through the kidneys) or it may be excreted in the feces. Some of the urobilinogen in the intestines is further broken down by intestinal bacteria to form urobilin, a pigmented substance that adds color to feces. Urine urobilinogen levels may be elevated because the liver is incapable of processing the urobilinogen in the blood or because there are elevated levels of bilirubin present. In the case of bile duct obstruction, the urine will be positive for bilirubin, but the urobilinogen levels will be normal. Hemolytic diseases may produce a negative bilirubin result with a positive urobilinogen reaction.

Specific Gravity

The clinical significance of the specific gravity test is presented in Chapter 20, as well as several testing methods that are used for specific gravity measurements. Specific gravity tests may be useful in monitoring the hydration status of patients or to measure the ability of the renal tubules to concentrate the urine specimen as needed. The specific gravity pad on the reagent strip takes advantage of the increased amount of ions present in urine specimens having higher concentrations. The greater the number of ions that are present in the urine, the more the color changes when exposed to the specimen. The specific gravity reagent pads are usually sensitive from 1.000 to 1.030 and are semiquantitative. The results may be inaccurate with alkaline urine specimens (falsely decreased results) or when high concentrations of protein are present in the urine specimen (falsely elevated results).

POINT OF INTEREST 21-2

CLIA-waived drug screening tests

Many employers now require their potential employees to have a urine drug screening test performed prior to an offer of employment. These drug screening collections are becoming a common part of the duties for medical assistants and phlebotomists in many offices. In some cases, the drug screenings may be federally mandated. Other employers have developed a drug screening policy to keep the workplace safer. If a drug screening collection is ordered by an employer, the medical assistant will be involved in the collection process, and a chain of custody will be started to document the transportation of the specimen until it is tested in a reference laboratory that specializes in this type of testing. This process requires specific training and a great deal of attention to detail. The employment status of the client depends on these results.

Urine drug screenings may also be ordered to check for illegal drug use in patients for whom this is suspected as part of their clinical diagnosis. These tests may be performed in the office, using CLIA-waived urine drug screening kits. They are available to test for many different drugs of abuse, such as marijuana, cocaine, amphetamines, opiates, oxycodone, and PCP. These tests are generally performed by applying the urine specimen to a reagent stick that will change color if the metabolite of that specific drug is present. In addition, employers who would like to have immediate results may ask for a quick screening test to be performed in the office using these CLIA-waived kits. This is not always conclusive; sometimes the patient may be taking prescription drugs that will cause the screening tests to be positive, because the urine tests are designed to identify the metabolites (breakdown products) of the drug that was ingested, not the drug itself. If a patient has a positive result, it is imperative to follow the office protocol for sending out the specimen for confirmatory testing.

A urine drug screening test may not always detect the presence of the drugs in the system even if the patient has been exposed recently. Each drug has a different time interval for which it can be detected. This can vary depending on the amount of exposure, the size of the individual, his or her unique metabolic rate, the pH of the urine specimen, and the type of drug. Drugs that are lipid soluble (such as marijuana) are detectable for longer periods of time than are many other drugs that are considered to be more dangerous. The product literature for the CLIA-waived testing kit used in the office should be consulted for more information about the sensitivity of the test. Some facilities now request testing of hair samples, as the retention time for the drug in the hair follicles is much longer than its presence in the urine. There are currently no CLIA-waived procedures for hair testing.

In addition, if the urine specimen is very dilute, it is possible that the test may be falsely negative. Most offices that perform drug screening collections and/or testing will have established policies in which specimens that are too dilute will need to be recollected. This decision is based on the specific gravity value of the specimen.

Another CLIA-waived test that may be requested by employers and/or health-care providers tests for the presence of alcohol by testing a saliva sample. Orasure Technologies manufactures a CLIA-waived quantitative saliva alcohol test. If an employer suspects that an employee has been ingesting alcohol while at work, that employer may request this test be performed as a screening tool. A blood alcohol test would be more sensitive, and may also be requested. This would require testing at a reference or hospital laboratory, as there are no CLIA-waived tests available to screen blood for alcohol.

POTENTIAL SOURCES OF ERROR

Some of the most common errors encountered while performing urine chemical testing are the following:

- *Incorrect specimen labeling:* Urine specimens must have the labels affixed to the specimen container, not to the lid of the container. Lids may be removed from multiple specimen containers at once, allowing for confusion when attempting to identify the correct patient name for each specimen. Labeling errors also include specimens that are mislabeled or not labeled at the time of collection.
- *Improper storage:* The reagent strips must be protected from direct light, moisture, and excessive heat. They should also not be stored in an area near volatile liquids, as the pads on the strip may absorb some of the chemicals present in the environment. Strips should not be removed from the bottle until just before use, and the desiccant provided needs to remain in the bottle of strips until they are all used. If the reagent strips

are exposed to moisture, they will become discolored and must be discarded. The lid for the bottle must be replaced immediately after removing reagent strips. The bottles are to be stored at room temperature, and care should be taken not to use the strips after the expiration date printed on the bottle. Reagent strips should never be cut in half or altered in any way.

- *Improper technique:* The manufacturer's instructions state that the urine reagent strip is to be dipped completely into a well-mixed specimen and removed immediately. Excess urine is to be removed from the strip by sliding the edge of the strip along the side of the container as it is withdrawn from the specimen. Inaccurate results are produced when the reagent strip is allowed to remain in the urine for an extended period of time, as the chemicals in the reagent pads will erode, contaminating the urine specimen to produce inaccurate color changes. Once the strip is removed from the urine, it is recommended that any excess urine still present be blotted away with an absorbent pad. Also, many formed elements in the urine specimen will settle to the bottom of the container if the specimen is not well mixed, allowing for potential false-negative results when the strip is inserted into the specimen.
- *Incorrect timing of reactions:* Incorrect timing of reactions is probably the most common error in reagent strip analysis. Each brand of test strips will specify how long the urine should remain in contact with the testing strip before the result is to be read by observing the color change on the pad. This time varies among tests, but all test pads should usually be read within 120 seconds. Unfortunately, the person performing the test commonly dips the reagent strip into the urine specimen, but doesn't follow the timing recommendations when reading the results. This results in an inconsistency within the laboratory and may lead to false-positive test results because of increased reaction times, or false-negative test results because of inadequate reaction times.
- *Incorrect result interpretation:* The reagent strips and color charts used to read the reactions are not interchangeable among manufacturers. For manual procedures, it is important to line up the reagent strip appropriately to the color chart used for comparison, and to report the analyte using the units and reporting methods suggested by each manufacturer. Some tests are reported as qualitative results in which the result is either positive or negative, whereas other tests are reported using a scale of 1+, 2+, and so on. Other analytes are reported using units of measurement, such as 20 mg/dL.
- *Test interference:* Because every reagent pad on the test strip is impregnated with a different chemical, each pad may be affected by different types of interfering substances. Urines that are very alkaline may interfere with several of the pads, as will strong colors (such as those with Pyridium [phenazopyridine] use), the presence of detergents or antiseptics, or high specific gravity. High levels of ascorbic acid (vitamin C) may cause false-negative results for blood, glucose, bilirubin, nitrite and leukocyte esterase. The presence of menstrual blood may cause false-positive results for blood that are not clinically significant. Many diagnostic dyes and medications also can cause interference for some of the testing methods. The manufacturer's insert will provide more detailed information about interfering substances for the various tests.
- *Incorrect urine storage or preparation:* Although this is not technically a potential source of error only for reagent strip testing, it is a common problem that can affect all the urinalysis results. Urine samples should be refrigerated, tested, or properly preserved within 1 hour of collection. Urine that is allowed to remain at room temperature for extended periods of time can undergo chemical changes that were not present at the time of collection.
- *Failure to perform maintenance, calibration, or quality control:* Whether the method used in the laboratory is manual or automated, quality control testing should at least be performed at the frequency specified by the manufacturer. In addition, automated instruments have requirements for calibration and maintenance, and these recommendations must be followed and documented appropriately.

Test Your Knowledge 21-7

What are three ways that reagent strips can be handled incorrectly? (Outcome 21-4)

Test Your Knowledge 21-8

Which of these is an example of an improper urine chemistry testing technique that may affect results?

a. Performing the test without wearing gloves
b. Allowing the reagent strip to remain in the urine for 60 seconds before removal
c. Mixing the specimen just prior to testing
d. Allowing the urine specimen to reach room temperature before testing (Outcome 21- 4)

SAFETY PRECAUTIONS

Urine is not considered to be an infectious agent for HIV and other bloodborne pathogens unless it is visibly contaminated with blood. However, it may still be infectious because it is contaminated with other microorganisms, such as bacteria. Mucous membrane exposure is the route of infection most reasonably anticipated for urine specimens being tested for chemical analytes, because of the potential for splashing the urine into the eyes or mouth as the specimen is handled. Standard precautions should be used when handling urine specimens, and appropriate actions should be taken to avoid splashing, spilling, or formation of aerosols when handling the urine specimen. Application of standard precautions for urine specimens include the following:

- Appropriate hand hygiene procedures before and after application of gloves. Also, if accidental exposure to bare skin does occur, this site should be disinfected as soon as possible.
- Glove use when performing testing procedures. Gloves should be removed when leaving the testing area and approaching a clean area of the laboratory or physician office.
- Wearing of long-sleeved, fluid-resistant laboratory coats that are completely buttoned or snapped.
- Occlusive dressings or bandages used over broken skin.
- Use of eye protection and/or face shield when exposure is anticipated.
- Nail biting, smoking, eating, drinking, or manipulating contact lenses is prohibited when handling urine specimens.
- Covering urine specimens before centrifuging.

Test Your Knowledge 21-9

What is the most common route of accidental exposure to urine specimens? (Outcome 21-5)

QUALITY CONTROL PROCEDURES

The chemical analysis of urine is a CLIA-waived testing procedure. This does not mean, however, that quality control is not an important part of the process. For instance, the package insert for the Bayer Multistix urinalysis test strips suggests that a positive and negative quality control should be tested whenever a new bottle of strips is put into use. This is a minimum frequency suggestion; individual laboratories may require that quality control samples be checked each day or each shift, depending on workload and/or laboratory policy. Most laboratories will require quality control testing daily. The frequency recommendation for automated urine chemistry testing may be different from the advice given for manual urine testing. There may also be unexpected situations (such as the exposure of the strips to extreme heat and/or cold) that will warrant quality control testing before patient specimens can be analyzed.

Commercial quality control materials are available for purchase. Examples of these include Lyphochek Quantitative Urine Control by Bio Rad, and KOVA-Trol (manufactured by Hycor Biomedical, Inc.). It is suggested that the positive control used for urinalysis procedures be a weakly positive specimen, so that the sensitivity of the testing process can be challenged. (Alternatively, three levels of control can be implemented to verify the sensitivity of the testing process.) It is important to note that water should never be used as a negative control because it does not really resemble urine enough to test the procedure. Some laboratories will freeze aliquots of positive and negative urine specimens to be used as quality control specimens; this is appropriate as long as the desired results are verified with multiple testing procedures and acceptable ranges are established for each test. Ideally, each laboratory will also participate in an external quality assessment survey to evaluate its performance as compared to other laboratories utilizing the same testing procedures. This is not a required component for CLIA-waived tests, but the comparison with other laboratories will allow for more confidence in the test results.

It is important to remember that patient testing should be performed only if the urine quality control results are within the acceptable range. If the results are not as expected, then patient results cannot be analyzed until the reason for the discrepancy has been identified. Troubleshooting steps may include using a new bottle of control material or using a new testing strip from a different bottle or lot number.

URINE TESTING METHODS

Urine reagent strip testing is considered one of the easiest laboratory tests performed. However, it is important to follow the instructions provided by the manufacturer precisely the way that they are written to ensure that the patient results are valid. The strips must be stored properly, and must not be exposed to moisture. Reagent strips must be discarded after their expiration date, and must be used only once. In addition, it is important to adhere to the instructions for timing the reactions and interpreting the results. The color changes for each test pad must be read individually at different time intervals, ranging from 30 to 120 seconds.

The manufacturer's insert for the reagent strips also provides information about potential causes of interference. An example is ascorbic acid (vitamin C), which can cause false-negative results for glucose, blood, and other substances on some reagent strips. Staff who perform chemical urine analysis need to be familiar with the package insert provided for the reagent strips used by their facility, and a copy should be available for reference at all times.

Reagent strips are also available to test for one or two chemical analytes, rather than the entire spectrum that is usually included in a routine urinalysis. Ketones and glucose are some of the most common analytes included on these limited test strips. These can be beneficial for those with type 1 diabetes or for monitoring pregnant women for potential gestational diabetes. Other test strips may be used to monitor protein and creatinine levels in the urine, which can help to detect kidney dysfunction in high-risk patients, or can be used to monitor the progress of kidney disease for patients who have already been diagnosed. Microalbuminuria may be detected by strips designed to pick up very low levels of protein in the urine. The presence of microalbuminuria is a significant factor in the progression of diabetes.

Test Your Knowledge 21-10

True or False: Urine reagent strips manufactured by different companies may be available with different test capabilities.

The reference values for urine chemical strips are the same whether the test is performed manually or by automation. These results are summarized on Table 21-1. It is important to remember that these reference values are obtained by following the manufacturer's instructions precisely; special attention should be given to the timing intervals for reading the results.

Automated urinalysis instruments are designed to read reagent strip reactions at the appropriate intervals. The results are usually printed out and can also be sent automatically through a computer interface for many instruments. The use of automated urinalysis instruments may help reduce errors by eliminating potential sources of error with the timing of reactions and the interpretation of color changes on the reagent pads. Some of the automated instruments are designed to read one strip at a time, whereas others are capable of processing numerous reagent strips at once. Bayer and Roche Diagnostics both manufacture automated urinalysis instruments, including the Bayer Clinitek models and the Roche Urisys instruments (Fig. 21-2). Various models are available from both of these manufacturers. Calibration and

TABLE 21-1

Reference ranges for urine chemistry results

Chemical Analyte	Normal Finding
pH	4.6–8.0
Protein	2–8 mg/dL (negative to trace)
Specific gravity	1.005–1.030
Leukocyte esterase	Negative
Nitrite	Negative
Glucose	Negative
Ketones	Negative
Leukocytes	Negative
Blood	Negative
Urobilinogen	0.1–1.0 mg/dL

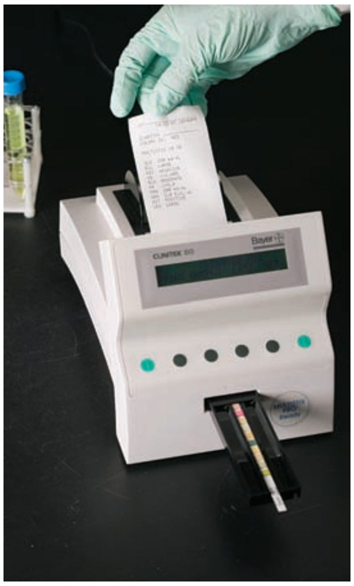

Figure 21-2 Bayer Clinitek urine analyzer with printed results. *From Eagle, S, Brassington, C, Dailey, C, and Goretti, C: The Professional Medical Assistant: An Integrative, Teamwork-Based Approach. FA Davis, Philadelphia, 2009, with permission*

maintenance are necessary for both of these machines, in addition to the testing of quality control specimens at specified intervals. Results are then recorded on a urinalysis report form (Fig. 21-3).

Test Your Knowledge 21-11

What is an advantage of automated urine chemistry testing procedures over manual testing procedures?

(Outcome 21-6)

CONFIRMATORY URINE TESTING

Sometimes the color of a urine specimen may cause color interference with the reagent strip test pads, producing invalid results. It may then be necessary to perform other types of confirmatory tests to provide some reliable information to the practitioner. The following are the most common confirmatory tests performed:

- *Acetest* (Bayer Corporation): The Acetest is used to test for the presence of ketones in the urine specimen. These reagents may be used to test other body fluids if necessary. The Acetest tablet reaction is less affected by the inherent color of the urine specimen, and the color changes are less subtle than they are on the reagent pads on the urine dipsticks. The testing process requires a tablet and an absorbent pad.
- *Ictotest* (Ames Corporation): The Ictotest is used to detect the presence of bilirubin in the urine specimen. This procedure is more sensitive to the presence of bilirubin than the reagent pad test, and is affected less by the strong colors present in some urine specimens. The test involves a tablet and an absorbent pad.

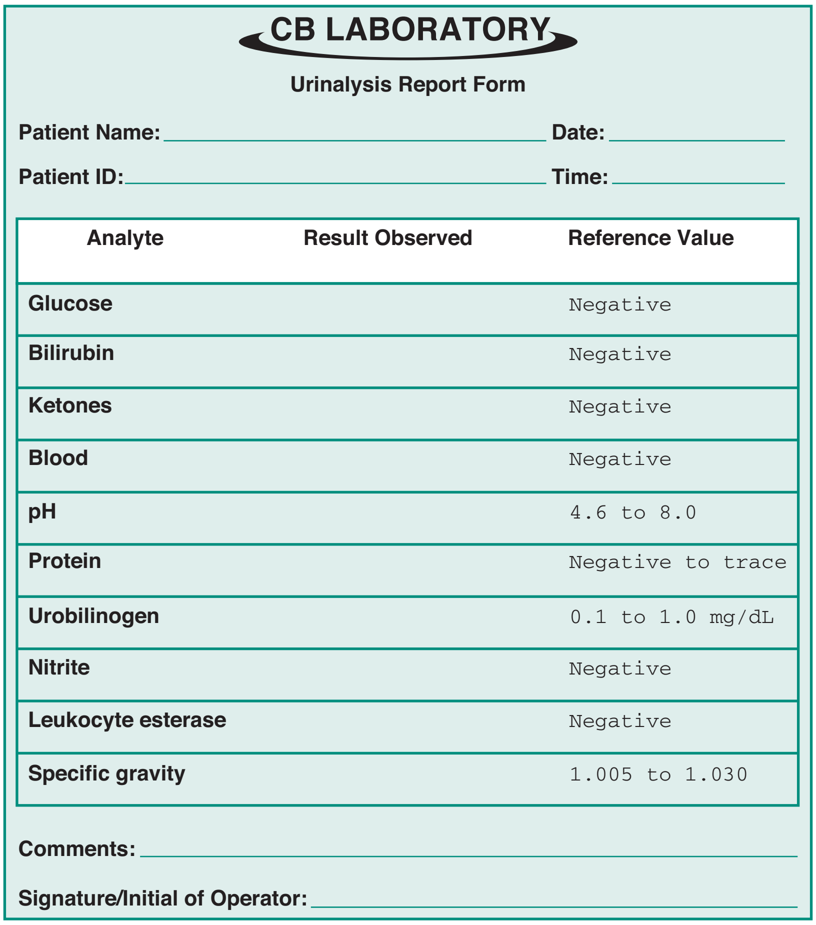

CB LABORATORY

Urinalysis Report Form

Patient Name: ______ Date: ______

Patient ID: ______ Time: ______

Analyte	Result Observed	Reference Value
Glucose		Negative
Bilirubin		Negative
Ketones		Negative
Blood		Negative
pH		4.6 to 8.0
Protein		Negative to trace
Urobilinogen		0.1 to 1.0 mg/dL
Nitrite		Negative
Leukocyte esterase		Negative
Specific gravity		1.005 to 1.030

Comments: ______

Signature/Initial of Operator: ______

Figure 21-3 Urinalysis report form.

Procedure 21-1: CLIA-Waived Chemical Examination of Urine Using Manual Reagent Strip Method

TASK

Performance of a CLIA-waived chemical examination of a urine specimen utilizing the manual reagent strip method.

CONDITIONS

- Gloves, laboratory coat, and protective eyewear
- Urine specimen freshly voided or appropriately preserved
- Clear conical plastic transfer tube
- Reagent strips with color comparison chart for reading results
- Normal and abnormal quality control specimens
- Test tube rack
- Kimwipes or paper towels
- Permanent fine-tip pen for marking identification on the transfer tube
- Urinalysis report form
- Black pen for documentation on the patient's chart
- Patient's chart
- Biohazardous waste container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

III.P. Psychomotor Skills, III. Applied Microbiology/Infection Control, #2. Practice Standard Precautions
I.P. Psychomotor Skills, I. Anatomy and Physiology, #14. Perform Urinalysis

ABHES Standards

- Apply principles of aseptic techniques and infection control.
- Use standard precautions.
- Perform CLIA-waived tests that assist with diagnosis and treatment, Urinalysis.

Procedure	Rationale
1. Assemble necessary equipment. Put on eye protection, or position plastic shield in a way that the procedure may be performed behind the shield.	Organization of supplies saves time and helps to avoid errors. Eye protection shields employees from potential biohazard exposure.
2. Wash hands.	Hand washing breaks the cycle of infection.
3. Verify test order and specimen labeling and identification.	Verification of test order and specimen identification eliminates potential errors.
4. Verify that specimen collection was less than 1 hour before testing or that it was refrigerated and/or preserved appropriately after collection.	Specimens that are allowed to stay at room temperature for more than 1 hour after collection without the addition of preservative may yield inaccurate results because of changes in pH, bilirubin or urobilinogen concentration, glucose levels, and bacterial contamination.
5. Allow any refrigerated specimens to come to room temperature before proceeding.	Specimens that are colder than room temperature may exhibit crystallization of amorphous urates or phosphates, which can interfere with measurement of all the physical characteristics of the urine specimen, as well as cause potential interference with the chemical analysis.

Continued

Procedure 21-1: CLIA-Waived Chemical Examination of Urine Using Manual Reagent Strip Method—cont'd

Procedure	Rationale
6. Mix specimen well with a gentle swirling motion.	Urine specimens must be mixed well to provide accurate chemistry results.
7. Pour some of the urine specimen into a labeled clear conical transfer tube. The transfer tube should be at least half full. Place the transfer tube in the tube rack.	Using a transfer tube for the urine to be chemically analyzed eliminates potential contamination of the original specimen by dipping in the reagent strip. If a urine culture is necessary, the original urine specimen will not be contaminated.
8. Prior to performance of patient testing, verify whether a quality control (QC) specimen needs to be tested, and if so, complete that process before the patient's test is performed. Patient testing cannot be performed unless the quality control values are within the established ranges.	Liquid QC should be tested following laboratory protocol and manufacturer's recommendations. Examples of when QC may be used to verify the test results include the following: **a.** With a new shipment of reagent strips **b.** Whenever a new lot number of reagents or QC is put into use **c.** New operator; someone who is being trained on the procedure **d.** Problems with storage, instrument, reagents, etc.
9. Perform the reagent strip test following these procedures: **a.** Dip the reagent strip into the well-mixed urine in the transfer tube. Make certain that all the reagent pads are moistened. **b.** Remove the reagent strip from the urine immediately, running the straight edge of the reagent strip along the top of the tube to remove excess urine. **c.** Immediately blot the edge of the reagent strip on a paper towel or laboratory wipes. **d.** Begin the timing for the pad development as the reagent strip is removed from the tube. **e.** Observe the color changes on the reagent pad at the appropriate time intervals as directed by the manufacturer. Compare the colors to the provided chart for observation, or to the chart provided on the reagent strip bottle. **f.** Record the results on the urinalysis report form. **g.** Discard the used reagent strip into a biohazard waste container.	Careful adherence to the procedure will allow for accuracy of results.

Procedure	Rationale
10. Dispose of the urine left in the transfer tube by pouring it down the sink and flushing it with plenty of water. Dispose of the tube and transfer pipette used for the urine in a biohazardous disposal bag. If the urine is not to undergo additional testing immediately, refrigerate it or preserve it accordingly.	
11. Sanitize the work area, remove gloves, and sanitize hands.	Gloves should always be removed immediately after a procedure is performed. Hand sanitization is appropriate before and after gloves are worn.
12. Document results for patient specimen on the chart (if available) and testing log sheet. Be certain to use the correct digits for the documentation.	Results must be recorded immediately after the testing process is complete.

Date	
10/24/2014:	*Chemical Urinalysis: Glucose: negative, Bilirubin: negative, Ketones: negative, Blood: negative, pH: 6.0, Protein: negative, Urobilinogen: 0.1 mg/dL, Nitrite: negative, Leukocyte esterase: negative, Specific gravity: 1.020*
11:58 a.m.	*Connie Lieseke, CMA (AAMA)*

Procedure 21-2: Chemical Examination of Urine Using Automated Reagent Strip Method

TASK

Perform a chemical examination of a urine specimen utilizing an automated urinalysis instrument.

CONDITIONS

- Gloves, laboratory coat, and protective eyewear
- Urine specimen freshly voided or appropriately preserved
- Clear conical plastic transfer tube
- Reagent strips and automated testing instrument with paper for result recording
- Normal and abnormal quality control specimens
- Test tube rack
- Kimwipes or paper towels
- Permanent fine-tip pen for marking identification on the transfer tube
- Urinalysis report form
- Black pen for documentation on the patient's chart
- Patient's chart
- Biohazardous waste container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

III.P. Psychomotor Skills, III. Applied Microbiology/Infection Control, #2. Practice Standard Precautions

I.P. Psychomotor Skills, I. Anatomy and Physiology, #14. Perform Urinalysis

ABHES Standards

- Apply principles of aseptic techniques and infection control.
- Use standard precautions
- Perform CLIA-waived tests that assist with diagnosis and treatment, Urinalysis.

Continued

Procedure 21-2: Chemical Examination of Urine Using Automated Reagent Strip Method—cont'd

Procedure	Rationale
1. Assemble necessary equipment. Put on eye protection or position plastic shield in a way that the procedure may be performed behind the shield.	Organization of supplies saves time and helps to avoid errors. Eye protection protects employees from potential biohazard exposure.
2. Wash hands.	Hand washing breaks the cycle of infection.
3. Verify test order and specimen labeling and identification.	Verification of test order and specimen identification eliminates potential errors.
4. Verify that specimen collection was less than 1 hour before testing or that it was refrigerated and/or preserved appropriately after collection.	Specimens that are allowed to stay at room temperature for more than 1 hour after collection without the addition of preservative may yield inaccurate results because of changes in pH, bilirubin or urobilinogen concentration, glucose levels, and bacterial contamination.
5. Allow any refrigerated specimens to come to room temperature before proceeding.	Specimens that are colder than room temperature may exhibit crystallization of amorphous urates or phosphates, which can interfere with measurement of all the physical characteristics of the urine specimen, as well as cause potential interference with the chemical analysis.
6. Mix specimen well with a gentle swirling motion.	Urine specimens must be mixed well to provide accurate chemistry results.
7. Pour some of the urine specimen into a labeled clear conical transfer tube. The transfer tube should be at least $^1/_2$ full. Place the transfer tube in the tube rack.	Use of a transfer tube for the urine to be chemically analyzed eliminates potential contamination of the original specimen by dipping in the reagent strip. If a urine culture is necessary, the original urine specimen will not be contaminated.
8. Prior to performance of patient testing, verify whether a quality control (QC) specimen needs to be tested, and if so, complete that process before the patient's test is performed. Patient testing cannot be performed unless the quality control values are within the established ranges.	Liquid QC should be tested following laboratory protocol and manufacturer recommendations. Examples of when QC may be used to verify the test results include the following: **a.** With a new shipment of reagent strips **b.** Whenever a new lot number of reagents or QC is put into use **c.** New operator; someone who is being trained on the procedure **d.** Problems with storage, instrument, reagents, etc.

Procedure	Rationale
9. Perform the reagent strip test following these procedures: **a.** Dip the reagent strip into the well-mixed urine in the transfer tube. Make certain that all the reagent pads are moistened. **b.** Remove the reagent strip from the urine immediately, running the straight edge of the reagent strip along the top of the tube to remove excess urine. **c.** Immediately blot the edge of the reagent strip on paper towel or laboratory wipes. **d.** Place the reagent test strip onto the appropriate area to be fed into the urinalysis instrument. Input any specimen ID necessary into the instrument, if possible. **e.** Monitor the instrument as the test is completed, watching for possible errors. **f.** Record the results on the urinalysis report form after they are printed. Some models display the results on a screen; if so, it will be necessary for them to be transposed onto the report form. **g.** Discard the used reagent strip into a biohazardous waste container.	Careful adherence to the procedure will allow for accuracy of results.
10. Dispose of the urine left in the transfer tube by pouring it down the sink and flushing it with plenty of water. Dispose of the tube and transfer pipette used for the urine in a biohazardous disposal bag. If the urine is not to undergo additional testing immediately, refrigerate it or preserve it accordingly.	Urine does not need to be disposed of in the biohazardous garbage, but it is important to flush the sink well after disposal.
11. Sanitize the work area, remove gloves, and sanitize hands.	Gloves should always be removed immediately after a procedure is performed. Hand sanitization is appropriate before and after gloves are worn.
12. Document results for patient specimen on the chart (if available) and testing log sheet. Be certain to use the correct digits for the documentation.	Results must be recorded immediately after the testing process is complete.

Date	
10/24/2009:	Chemical urinalysis using Bayer Clinitek 500; Quality control results within range: Glucose: negative, Bilirubin: negative, Ketones: negative, Blood: negative, pH: 6.0, Protein: negative, Urobilinogen: 0.1 mg/dL, Nitrite: negative, Leukocyte esterase: negative, Specific gravity: 1.020
11:58 a.m.	Connie Lieseke, CMA (AAMA)

- *Sulfosalicylic acid precipitation test:* This test involves the addition of an acid to the urine specimen and an observation for the amount of turbidity caused when the acid is added. A semiquantitative assessment of the amount of protein present in the urine specimen may result by measuring the turbidity present. The SSA precipitation test is not affected by the color of the specimen, as the measurement involves the turbidity present, not a color change reaction.
- *Clinitest* (Bayer Corporation): This test is used as a confirmatory test for the presence of glucose and other sugars in the urine specimen. It is not affected as much by the original color of the specimen, although some substances (such as ampicillin and vitamin C) can cause inaccurate results.
- *Microscopic examination:* Chapter 22 details the process used for urine microscopic examinations. In situations in which the color of the specimen prohibits valid results for blood, leukocyte esterase, or nitrites, the examination of the urine sediment for these formed elements may provide additional information for the practitioner.

FECAL OCCULT BLOOD TESTING

Fecal occult blood testing (FOBT) detects blood that is "hidden" in or on feces. Blood in the stool may be clinically significant even though there are only small amounts present. It is not unusual for adults to shed a few milliliters each day in their feces as a result of the natural processes occurring in the stomach and intestines. However, increased blood loss may be due to a pathological condition that needs medical attention. Blood may be present in the stool in the following situations:

- Bleeding hemorrhoids
- Ulcers and/or inflammation of the stomach or duodenum
- Presence of polyps, lesions, or tumors in the intestines
- Nosebleeds
- Bleeding gums
- Diverticulitis, colitis, and/or Crohn's disease
- Parasitic infection
- Colorectal cancer

Stool testing for the presence of occult blood is a simple, noninvasive, and relatively inexpensive screening test. The fecal occult blood test is not specific for any disease process, but a positive result will indicate a need for further investigation with a colonoscopy, proctosigmoidoscopy, or a lower gastrointestinal x-ray that uses barium to visualize the digestive tract. Most commonly this test is ordered as a screen for asymptomatic patients, but it may also be used when there is anemia present with no identifiable cause or when the patient exhibits symptoms of other disorders that may result in a positive screen.

According to the Centers for Disease Control and Prevention (CDC), colorectal cancer is the second leading cancer killer in the United States. Early stages of this disease are usually asymptomatic; therefore, colorectal cancer is often overlooked until it has progressed to a point at which treatment options are limited. The CDC and the American Cancer Society recommend that all adults who are 50 to 75 years of age have a yearly fecal occult blood screening test performed. Those past the age of 75 should discuss the necessity of further testing with their physician. Earlier (before the age of 50) or more frequent screening tests may be necessary if a patient has a family history of colorectal cancer or if the patient has inflammatory bowel disease. For patients who are covered by Medicare, yearly fecal occult blood screening is a service that is covered for payment.

Test Your Knowledge 21-12

Why does the American Cancer Society recommend a yearly fecal occult blood screening? (Outcome 21-9)

There are two types of fecal occult blood screening kits available. The **guaiac method**, usually referred to as the **FOBT** (for fecal occult blood test), uses a cardboard specimen holder containing a special guaiac paper inside. This paper is used for application of the specimen and the developer. If there is blood present in the stool, it oxidizes the guaiac, and when the developer (hydrogen peroxide solution) is added, the surrounding test paper turns blue where blood is present (Fig. 21-4). The guaiac reacts with heme present in the blood cells. This method has been available for many years, and is considered reliable if the patient follows the dietary restrictions and collection procedures as instructed. The guaiac method is not as sensitive or specific as the newer methods, but it is inexpensive, so many facilities still use it for their screening. Hemoccult (manufactured by Beckman Coulter) is most commonly used. Seracult (manufactured by Propper) is another screening kit that is available. To perform the collection process properly using guaiac-based kits, keep the following key points in mind:

1. Seven days prior to and during the collection, all NSAIDs (such as naproxen or ibuprofen) must be discontinued. One adult aspirin per day is allowed. Tylenol is also allowed.

2. Three days before and then during the collection, all vitamin C supplements must be discontinued, as well as citrus fruits or juices. For some test kits, turnips, broccoli, melons, radishes, and fresh fruit should also be discontinued. During this time, ingestion of red meat (beef, lamb, or liver) must be discontinued as well. Anticoagulant use may also cause false-positive test results, but patients must consult their healthcare provider before discontinuing these drugs.
3. The stool specimens should be collected from three consecutive bowel movements, preferably on three consecutive days. Follow directions provided concerning the volume of stool to apply, and to which side of the slide to apply it.

4.

Test Your Knowledge 21-13

Which of these substances should be avoided for 7 days prior to and then during the collection period for a FOBT?

a. Red meat
b. Caffeine
c. Sodium
d. NSAIDs

(Outcome 21-10)

5.

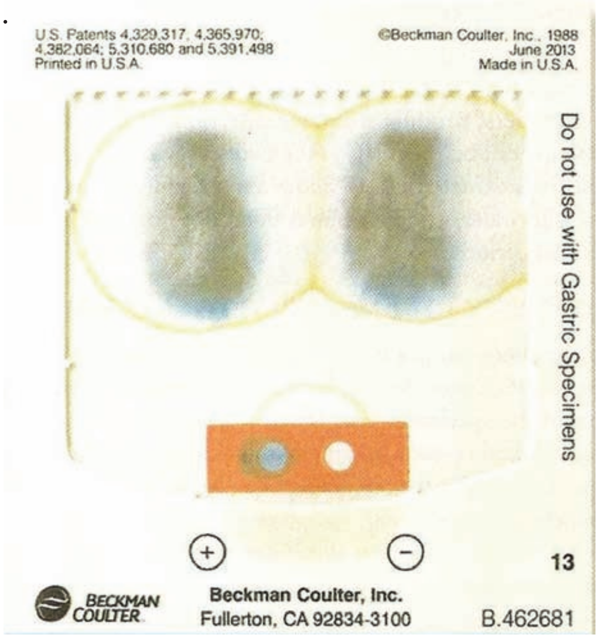
U.S. Patents 4,329,317, 4,365,970; 4,382,064; 5,310,680 and 5,391,498
Printed in U.S.A.
©Beckman Coulter, Inc., 1988
June 2013
Made in U.S.A.
Do not use with Gastric Specimens
+
−
13
BECKMAN COULTER
Beckman Coulter, Inc.
Fullerton, CA 92834-3100
B.462681

Figure 21-4 Positive fecal occult blood test result using guaiac method. Note the blue area around the specimens, which indicates a positive result and the presence of occult blood in the stool. In addition, there is a positive result on the control area of the slide. *Courtesy of Beckman Coulter.*

The other fecal occult blood testing method, the **iFOB,** which stands for immunochemical fecal occult blood, uses an immunochemical test to detect the presence of globin, part of the hemoglobin molecule. This test reacts only with human blood, and is more sensitive and specific than is the guaiac testing method. The iFOB has been shown to detect a positive result with less blood present in the stool, and to have fewer false-positives due to interfering substances. There are fewer dietary restrictions associated with this testing method; therefore, patient compliance may increase. The iFOB test may be available as a kit containing individual testing cartridges or as slides similar to the other type of fecal occult blood tests. It is still recommended that the sample be collected over three consecutive bowel movements over three days, as polyps and lesions in the GI tract may bleed intermittently, and this method of collection increases the opportunity to detect bleeding if present. Hemoccult ICT (manufactured by Beckman Coulter) is one example of this immunochemical testing method. Other manufacturers, such as Quidel, also produce a CLIA-waived rapid test that provides a qualitative result for the presence of fecal occult blood. The immunochemical tests for occult blood are much more expensive than are the guaiac tests, and are not commonly found in the physician office laboratory.

Regardless of the type of test used, the desired result is a negative one, indicating that no measurable occult blood is present in the stool. However, a positive result does not necessarily indicate colorectal cancer, or even a precancerous condition. The fecal occult blood test procedure is to be used as a screening tool only, and further testing is always necessary to identify the cause of any positive results. The test results for the FOBT and the iFOB are reported as a positive or negative result, indicating the presence or absence of occult blood in the specimen.

Procedure 21-3: Fecal Occult Blood Testing Using Guaiac Method

TASK

Provide patient instruction for the collection of the sample necessary and perform a test to detect the presence of fecal occult blood using the Hemoccult Sensa test cards and developer.

CONDITIONS

- Gloves
- Laboratory coat
- Hemoccult Sensa test cards
- Hemoccult Sensa developer
- Stool specimens
- Biohazardous waste container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

III.P. Psychomotor Skills, III. Applied Microbiology/Infection Control, #2. Practice Standard Precautions

ABHES Standards

- Apply principles of aseptic techniques and infection control.
- Use standard precautions
- Instruct patients in the collection of a fecal specimen

Procedure	Rationale
1. Greet and then identify patient using at least two unique identifiers.	All patients must be identified properly before collecting samples or performing laboratory testing.
2. Verify test ordered, and explain the specimen collection requirements to the patient.	All laboratory test orders should be verified by checking the chart and/or requisition form more than once. A thorough explanation of the collection procedure for the patient will ensure more cooperation. Written literature should also be provided so that the patient knows how to prepare appropriately and collect the specimens correctly. Preparation and collection instructions include the following: • Collect and apply samples from bowel movements from three different days to the slide. This increases the opportunity to detect blood that may be intermittently present from polyps or cancer present in the GI system. • Do not collect a sample if frank (obvious) blood is present in the stool. • For best results, collect the sample before the stool makes contact with the water present in the toilet. Stool specimens that are retrieved from toilet water with chemical additives are unacceptable. • Protect slides from heat, bright lights, and exposure to strong household chemicals such as ammonia. • For 7 days before and during the stool collection period, avoid NSAIDs such as aspirin and ibuprofen. One adult aspirin per day is acceptable. • For 3 days before and during the stool collection period, avoid vitamin C in excess of 250 mg/day and red meats. • Eat a well-balanced diet.

Procedure	Rationale
3. Provide the patient with a set of three slides, an applicator, collection tissues, and written instructions for collection. Let the patient know how and when to bring the specimens back to the laboratory. Make certain that the slides are correctly identified with the patient's name and account number.	Slides must be returned for testing within 14 days of collection. It is important that the date of each collection is recorded on the outside front of the slide. To apply the specimen to the slide, follow these steps: **1.** Open the front of the slide. Using the applicator provided, place a small amount of stool specimen as a thin smear covering box A. Using the same applicator, choose a different part of the stool specimen and take another small amount to smear on box B. **2.** Dispose of the applicator and close the front of the slide. **3.** Patients must not open the back of the slide. This is for laboratory use only.
4. When the slides arrive back at the laboratory, wash hands and apply gloves before touching the slides.	Hands should always be washed between patients and before starting any procedures. Gloves are appropriate personal protective equipment (PPE) for this procedure.
5. Assemble the developer, and verify that all reagents are within the posted expiration dates.	The developer should be stored at room temperature and must be used before the expiration date.
6. Verify the patient identification on the cards.	Always verify that the correct specimen is in hand before completing the testing process.
7. Open the back of the slide. Place two drops of Hemoccult developer directly over the specimens in box A and box B. Results are to be read at 60 seconds after application of the developer. Any blue on or at the edges of the slide is considered to be a positive result for the presence of occult blood.	Guaiac occult blood test results should not be read by those who have color blindness, as they may not recognize the development of the blue color with a positive result.
8. Make a note of the result at 60 seconds.	The result must not be read before allowing 60 seconds for development or after 60 seconds has elapsed.
9. Before the results are considered to be valid, the quality control area of the slide must be developed.	Apply one drop of developer between the positive and negative designation of the QC testing area. Read the results at 10 seconds. • The positive area will turn blue, and no blue will develop in the negative area. • If results are not as expected when performing quality control, the patient results cannot be reported.
10. Repeat the testing process and quality control assessment on the other slides.	The procedure is the same for each slide.

Continued

Procedure 21-3: Fecal Occult Blood Testing Using Guaiac Method—cont'd

Procedure	Rationale
11. Dispose of the slide in the biohazardous disposal container.	Biohazardous specimens must be disposed of properly.
12. Put away developer, remove gloves, wash hands, and record specimen results.	Hands must always be washed after removing gloves.

Date	
8/19/2014:	Hemoccult slides tested; negative X3 ~~~~~~~~~~~~~~~~~~~~
11:50 a.m.	~~~~~~~~~~~~~~~~~~~~ Connie Lieseke, CMA (AAMA)

Procedure 21-4: Fecal Occult Blood Testing Using iFOB Quickvue Method

TASK

Perform a fecal occult blood test using the Quickvue immunoassay method.

CONDITIONS

- Gloves
- Laboratory coat
- Hand sanitization supplies
- Specimen pouch containing
  - QuickVue collection tube (covered with identification label) contained in absorbent sleeve
  - Collection paper with adhesive
  - Patient instructions
  - Return mailer
- QuickVue test cassette
- Absorbent laboratory wipes or gauze pads
- Timer
- Quality control material (if necessary)
- Biohazardous waste container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

III.P. Psychomotor Skills, III. Applied Microbiology/Infection Control, #2. Practice Standard Precautions

ABHES Standards

- Apply principles of aseptic techniques and infection control.
- Use standard precautions.
- Instruct patients in the collection of a fecal specimen.

Procedure	Rationale
1. Greet and then identify patient using at least two unique identifiers.	All patients must be identified properly before collecting samples or performing laboratory testing.
2. Verify the test ordered, and explain the collection procedure to the patient. Enter the patient identification information on the collection tube label.	All laboratory test orders should be verified by checking the chart and/or requisition form more than once. The collection process for this procedure is very specific, and must be followed carefully. Patient instructions include the following: • During the collection process, do not allow the fecal specimen to contact the toilet water. The specimen must also be protected from urine.

Procedure	Rationale
	• When ready to begin the collection, attach the collection paper to the toilet seat by removing the tape cover from the ends of the collection paper and attaching the adhesive material to the toilet seat across the back of the toilet. Attach the paper in such a way that it sags in the middle of the toilet. • Defecate on the collection paper. Do not remove the fecal specimen or the collection paper from the toilet seat at this point. • Unscrew the sampler from the collection tube. • Pierce the fecal specimen (while still on the collection paper) with the grooved end of the sampler device at least five times in different areas of the specimen. • Insert the sampler into the collection tube and close the cap securely. • Shake the tube to mix the buffer provided in the tube with the sample just added to the tube. • Flush the remaining collection paper and the fecal sample down the toilet.
3. Explain to the patient how to return the specimen to the laboratory for testing.	The specimens may be mailed using the provided mailer, or they may be delivered back to the laboratory. Specimens may be kept at room temperature for up to 8 days after collection.
4. When the specimen is returned to the laboratory, verify that the identification is appropriate and that the patient followed the correct collection procedure.	The specimen is invalid if the collection procedure is not followed.
5. Sanitize hands and apply gloves.	Hands should always be washed between patients and before starting any procedures. Gloves are appropriate personal protective equipment (PPE) for this procedure.
6. Assemble necessary equipment, and verify that all reagents are within the posted expiration dates.	The only additional equipment for this test would be the timer and the test cassette.
7. Remove the test cassette from the foil wrapper and place it on a level surface.	The test cassette should not be removed until right before use.
8. Remove the absorbent sleeve from the specimen pouch, and remove the collection tube from the absorbent sleeve.	The absorbent sleeve is provided in case of specimen leakage during transport.
9. Shake the collection tube well to mix the sample and the buffer solution.	The specimen and buffer must be well mixed for valid test results.

Continued

Procedure 21-4: Fecal Occult Blood Testing Using iFOB Quickvue Method—cont'd

Procedure	Rationale
10. Prior to performance of test, verify whether a quality control (QC) specimen needs to be tested, and if so, complete that test *before* the patient's test is performed.	External QC material should be tested following laboratory protocol and manufacturer's recommendations. Examples of when QC should be used to verify the test results include the following: **a.** With a new shipment of the testing materials **b.** Whenever a new lot number of reagents or QC is put into use **c.** New operator; someone who is being trained on the procedure **d.** Problems with storage, instrument, reagents, etc. **e.** To ensure that storage conditions are fine, QC should be performed at least once monthly
11. Hold the collection tube upright and remove the light blue cap at the top of the tube.	It is important to hold the tube upright for this step.
12. Completely cover the exposed tip of the collection tube with the absorbent pad.	It is important to completely cover the exposed tip to avoid potential exposure to the specimen during the next step.
13. Without twisting, quickly snap off the tip of the collection container.	The tip will not snap off if a twisting motion is used.
14. Turn the collection tube upside down into a vertical position. Dispense six drops into the well of the testing cassette.	The tube should be held vertically to dispense the correct size of drop into the well.
15. Read the results in 5 to 10 minutes.	Some positive results will be evident at 5 minutes, whereas others may take up to 10 minutes to develop as a positive. Negative results should be reported within 10 minutes. Do not read the results after 10 minutes. Results are interpreted as follows: • Maroon lines will appear at the C and T area of the test cassette when the result is positive for the presence of occult blood. • Negative results will only exhibit a line at the C area. • If a line appears only at the T area of the test cassette or if no lines appear, the test results are invalid.
16. Dispose of the testing cassette and sample tube in the biohazardous disposal container.	These should all be disposed of as biohazardous materials because of the potential for the presence of blood or other biological hazards in the specimen.

Procedure	Rationale
17. Disinfect the work area.	The work area should always be disinfected after potential exposure to biological substances.
18. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
19. Document the test results in patient chart and/or log sheet.	All results must be documented in the patient's chart.

Date	
8/17/2014:	*IFOB test negative for occult blood* ~~~~~~~~~~~~~~~~~~~~
12:50 p.m.	~~~~~~~~~~~~~~~~~~~~ *Connie Lieseke, CMA (AAMA)*

SUMMARY

The comprehensive chemical analysis of urine is a frequently performed procedure as part of a routine urinalysis. Chemical analyses may also be limited to a few analytes, testing for the presence of substances such as glucose, ketones, and microalbumin. Urine reagent test strips contain pads that change color in the presence of various chemicals present in the urine specimen. The amount of color change is observed and recorded, either manually or automatically using an instrument. These color changes provide qualitative or semiquantitative results for the chemical analytes on the strip. Valuable information for the diagnosis and management of various renal diseases and other body system dysfunctions may be obtained by performing chemical urine testing. Fecal occult blood testing is performed as a screening test for the presence of colorectal cancer, allowing for early identification and treatment of that disease.

TIME TO REVIEW

1. What is the difference between glucosuria and glycosuria? Outcome 21-1

2. Jaundice may include : Outcome 21-1
   - **a.** Yellow sclera of the eye
   - **b.** Red, bloody urine
   - **c.** Light-colored feces
   - **d.** Pus in the urine

3. The sulfosalicylic acid precipitation test is used to detect which of these in the urine? Outcome 21-8
   - **a.** Glucose
   - **b.** Ketones
   - **c.** Bilirubin
   - **d.** Protein

4. True or False: A positive blood test on a urine specimen is always indicative of intact red blood cells in the urine. Outcome 21-2

5. The presence of ketones in the urine specimen is indicative of: Outcome 21-2
   - **a.** High blood sugar
   - **b.** Bile obstruction
   - **c.** Increased fat metabolism
   - **d.** Urinary tract infection

6. A positive nitrite result in the urine specimen may be indicative of: Outcome 21-2
   - **a.** Hematuria
   - **b.** White blood cells in the urine specimen
   - **c.** Elevated specific gravity results
   - **d.** Bacteria in the specimen

7. True or False: A urine specimen with a pH of 4.0 is considered to be outside of the reference range. Outcome 21-3

8. True or False: Medication that causes the urine to appear bright orange is likely to cause interference on a urine chemistry test. Outcome 21-4

9. Which of the following may cause test interference on the urine chemistry tests? Outcome 21-4

a. Vitamin C ingestion
b. Dairy products
c. Cold medication
d. Excessive caffeine intake

10. List three safety precautions applicable to urine chemistry testing procedures. Outcome 21-5

11. True or False: A positive fecal occult blood test is always indicative of colorectal cancer. Outcome 21-9

Case Study 21-1: Exceptions to the rule?

Sally Steiner is at the physician's office because symptoms that may be indicative of a urinary tract infection. She has been urinating frequently and experiences pain when voiding. When the medical assistant in the laboratory performs the chemical analysis on her urine specimen, the nitrite test is negative.

1. Does this result eliminate the possibility that there may be bacteria present in the urine specimen? Why or why not?

RESOURCES AND SUGGESTED READINGS

Clinical and Laboratory Standards Institute, Urinalysis: Approved Guideline, ed. 3. CLSI document G16-A3. Wayne, PA, 2009

Approved laboratory standards for collecting, processing, and testing urinalysis samples

"Galactosemia"

The U.S. Library of Medicine, Genetic Home Reference, explains how galactosemia is passed genetically and the effect that it may have on the body http://ghr.nlm.nih.gov/condition=galactosemia

"Urinalysis, a Comprehensive Review"

In-depth article from the American Association of Family Practitioners about the causes of urinalysis results, as well as information about false-positives and false-negatives http://www.aafp.org/afp/2005/0315/p1153.html

Chapter 22

Microscopic Examination of Urine

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Reasons for Performing Urine Microscopic Examinations
- Common Formed Elements in the Urine and Their Clinical Significance
  - Epithelial Cells
  - Mucus
  - Blood Cells
  - Spermatozoa
  - Artifacts
  - Casts
  - Crystals
  - Microorganisms
- Methods Used for Urine Microscopic Examinations
- Reporting Urine Microscopic Results
- Role of Medical Assistants in Microscopic Urine Examination Procedures
- Quality Control and Quality Assurance Procedures for Urine Microscopic Examinations
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

22-1 Define the key terms.

22-2 Explain why urine microscopic examination procedures are performed.

22-3 List the various formed elements identified in urine specimens.

22-4 Describe the clinical significance of the formed elements in a urine specimen.

22-5 Contrast the standard methods used to prepare a urine specimen for microscopic analysis.

22-6 Describe how to focus the urine sediment specimen appropriately on the microscope for viewing.

22-7 Summarize the procedures for reporting formed elements in urine specimens.

22-8 Describe how a medical assistant may be involved in the performance of a urine microscopic analysis.

22-9 Detail what type of quality assurance and quality control procedures are necessary to perform with urine microscopic analysis.

22-10 Prepare a urine specimen for microscopic analysis.

CAAHEP/ABHES STANDARDS

CAAHEP 2008 Standards

I.P.I.14. Perform Urinalysis
III,P.III.2. Practice Standard Precautions

ABHES Standards

- Medical Office Clinical Procedures: b. Apply principles of aseptic techniques and infection control, i. Use standard precautions
- Perform Urinalysis

KEY TERMS

Artifacts
Casts
Crystals
Epithelial cell
Fungus
Hematuria
Mucus
Provider performed microscopy (PPM)
Pyuria
Reflexive
Sediment
Spermatozoa
Supernatant

The final step in a complete urinalysis is the microscopic examination of the urine sediment. This procedure is performed to visualize and identify the formed elements that may be present. These microscopic structures may enter the urine while it is in the kidneys or as the urine passes through the lower urinary tract. The microscopic examination can detect infection, trauma, or damage to the urinary system, or provide information used to monitor certain disease states. In addition, some metabolic disorders may be detected because of the presence of pathological crystals in the urine specimen. Because the presence of the formed elements is not always clinically significant when they are present in small amounts, the microscopic examination provides quantitative information about the elements present in addition to the identification of the structures.

REASONS FOR PERFORMING URINE MICROSCOPIC EXAMINATIONS

The Clinical and Laboratory Standards Institute recommends the performance of a urine microscopic examination whenever the chemical or physical testing is abnormal on the urine specimen. Because of this recommendation, some laboratories always perform urine microscopic examinations when the chemical analysis or physical examination of the urine specimen is abnormal. This is known as **reflexive** testing. The decision for adding a urine microscopic examination may also be based on the appearance of the specimen; for example, if the specimen appears cloudy or bloody, then a microscopic examination may be added. In addition, the urine microscopic examination may be ordered by a healthcare practitioner to screen for a disease process or to monitor the progress of treatment. For example, the practitioner may need quantitative information about the amount of bacteria present in a sample in order to prescribe appropriate medication, or it may be necessary to determine if certain types of crystals are present in the specimen for a definitive diagnosis.

Test Your Knowledge 22-1

Provide one reason that a urine microscopic examination may be performed. (Outcome 22-2)

COMMON FORMED ELEMENTS IN THE URINE AND THEIR CLINICAL SIGNIFICANCE

The formed elements that may be present in a urine specimen include epithelial cells, mucus, blood cells, spermatozoa, artifacts, casts, crystals, and various microorganisms. These structures are identified and the approximate amount present in the urine sediment is estimated and reported. Some of the structures are identified using the low-power objective on the microscope (10X), whereas others require the high-power objective (40X) for visualization.

Many normal urine specimens contain a few formed elements. It is important to appropriately report the quantity observed, because the presence of very small amounts of some structures may have clinical significance. The reference ranges and clinical significance of each formed element vary, as explained in the following sections.

Epithelial Cells

A common cell identified in the urine specimen is an **epithelial cell.** These cells are constantly added to the urine as they are shed from the lining of the urinary tract. There are three types of epithelial cells that may be present, and they are named according to the area where they originate within the urinary tract:

- *Squamous epithelial cells:* These are the most common epithelial cells present in the urine. Their presence is not considered to be clinically significant, as the cells naturally slough off from the lining of the lower urinary tract in both males and females, and from the vagina in females. Squamous epithelial cells are large, flat, and irregular in shape, with a large nucleus and abundant cytoplasm. Their shape may be described as having a "fried egg" appearance. Squamous epithelial cells are often the first formed element that is visible when focusing the microscope. They are reported using the low-power objective as rare, few, moderate, or many per low-power field. Usually there are at least a few squamous epithelial cells present in any urine sediment (Fig. 22-1).
- *Transitional epithelial cells:* These cells originate from the structures of the kidneys and the ureters, bladder, and upper urethra in males. They line the urinary tract in these locations, and their presence in small numbers in the urine is not considered abnormal. However, if the number of transitional epithelial cells is elevated, it may be indicative of a pathological condition such as a malignancy. Transitional epithelial cells may also appear in increased numbers after urinary catheterization or other invasive urological procedures. These cells are identified using the high-power objective and reported as rare, few, moderate, or many per high-power field. A normal urine sample should not exhibit more than the rare transitional epithelial cell.
- *Renal tubular epithelial cells:* As the name implies, the renal tubular epithelial cells come from the renal tubules in the kidneys. The size and shape of the renal tubular epithelial cells differ, depending on the area of the tubules from which they originate. They are smaller than the other epithelial cells, and their appearance can vary from round to columnar. The presence of renal tubular epithelial cells in the urine specimen is clinically significant, as it may indicate conditions in which damage is occurring to the renal tubules. Examples include glomerulonephritis, viral infections, chronic or acute kidney disease, and other disease states that cause necrosis of the kidney tubules. Renal tubular epithelial cells are identified using the high-power objective, and reported as a number per high-power field. The presence of renal tubular epithelial cells in the urine sediment is always considered to be an abnormal result.

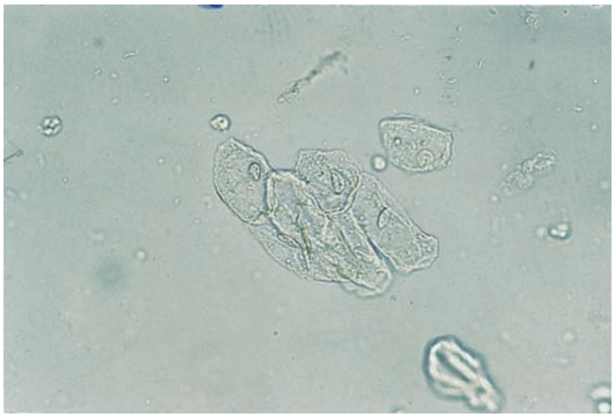

Figure 22-1 Squamous epithelial cells in urine sediment. *From Strasinger, SK, and Di Lorenzo, MS: Urinalysis and Body Fluids, ed. 5. FA Davis, Philadelphia, 2008, with permission.*

Test Your Knowledge 22-2

Is the presence of epithelial cells in a urine specimen always clinically significant? (Outcome 22-4)

Mucus

The presence of **mucus** in the urine specimen is very common. Mucus appears as thread-like strands made of protein. Figure 22-2 shows an example of mucus. The mucous membranes that line the lower urinary tract and the renal epithelial cells normally produce mucus. The strands may be difficult to visualize when the microscopic light source is at its brightest; it is often necessary to reduce light intensity to see mucus clearly. Mucus is not indicative of a pathological condition in the body, and has no clinical significance. Large amounts of mucus in the specimen may produce a positive protein result on the chemical urine analysis. Mucus is identified using the low-power objective, and is reported as rare, few, moderate, or many per low-power field. Normal urine specimens may include a moderate amount of mucus.

Blood Cells

Red blood cells and white blood cells are often identified as part of the urine microscopic examination. Their clinical significance varies with the number present and the type of blood cell visualized.

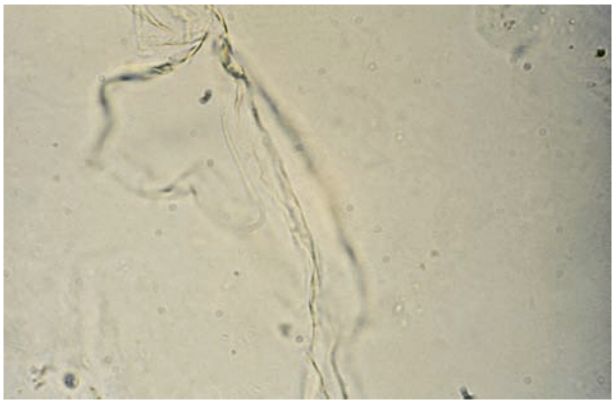

Figure 22-2 Threads of mucus; note the stringy appearance. *From Strasinger, SK, and Di Lorenzo, MS: Urinalysis and Body Fluids, ed. 5. FA Davis, Philadelphia, 2008, with permission.*

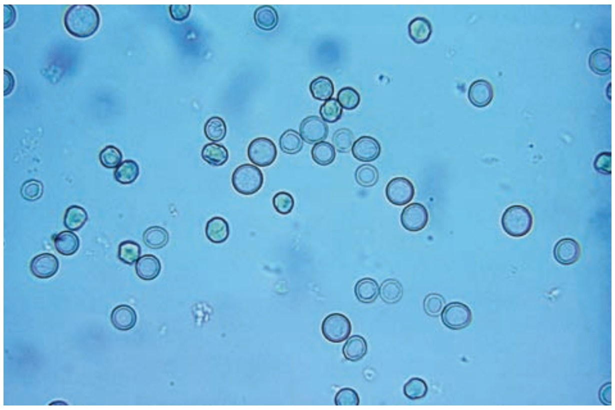

Figure 22-3 Red blood cells (erythrocytes) as they appear in urine sediment. *From Strasinger, SK, and Di Lorenzo, MS: Urinalysis and Body Fluids, ed. 5. FA Davis, Philadelphia, 2008, with permission.*

- *Red blood cells:* It is not uncommon to see one or two red blood cells when examining urine sediment with a high-power magnification. When the number of red blood cells is elevated, it is indicative of damage to the glomeruli of the kidney or trauma to the vessels surrounding the urinary tract. The presence of red blood cells in the urine is a called **hematuria.** The number of red blood cells present in the specimen is relative to the severity of the damage to the glomerulus or the vessels. Red blood cells may also be present in the urine specimen after strenuous exercise.

Red blood cells may be difficult to identify when viewed in the urine specimen. They are colorless, smooth, and have no nucleus. Red blood cells appear as round biconcave disks that are difficult to distinguish from air bubbles, budding yeast, oil droplets, or artifacts in the specimen. The addition of stain to the specimen may aid in the identification and counting of the red blood cells. Figure 22-3 shows how red blood cells appear in the urine. Red blood cells are identified using the high-power objective, and are reported as the number seen per high-power field.

- **White blood cells:** A few white blood cells may be evident in a normal clean-catch midstream urine specimen when viewed under the microscope. Elevated numbers of white blood cells in the urine is called **pyuria,** which indicates infection or inflammation of the urinary tract. Prostatitis, pyelonephritis, and cystitis may cause elevated white blood cell counts in the urine specimen. Pyuria may also be present when there are tumors in the urinary tract, and other systemic inflammatory conditions.

It is easier to identify white blood cells in the urine specimen than it is to identify red blood cells. The white blood cells are larger, have a granular appearance, and contain a nucleus (Fig. 22-4). Most of those present in the urine specimen are neutrophils, although eosinophils or mononuclear cells may be elevated in certain situations. White blood cells are identified using the high-power objective and reported as the number visualized per high-power field.

Spermatozoa

Occasionally, **spermatozoa** may be present in urine specimens from both males and females after sexual intercourse. The spermatozoa will be identified by their distinctive shape, as they have an oval head that is

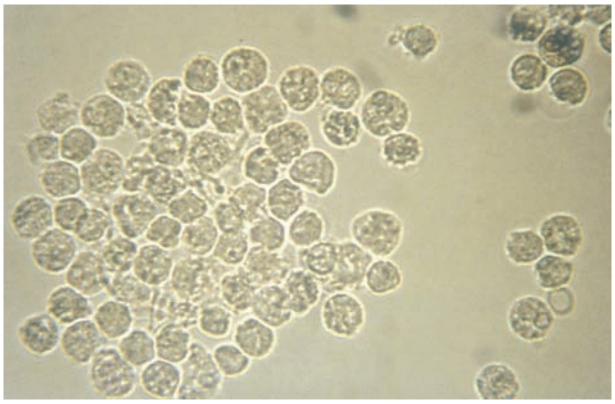

Figure 22-4 White blood cells. Note the presence of the nucleus and the cytoplasmic granules. *From Strasinger, SK, and Di Lorenzo, MS: Urinalysis and Body Fluids, ed. 5. FA Davis, Philadelphia, 2008, with permission.*

POINT OF INTEREST 22-1

Clue cells

Occasionally the urine microscopic report for a female patient may include the presence of *clue cells.* This term is used to describe vaginal epithelial cells that are coated with the *Gardnerella vaginalis* bacterium. The clue cells are similar in appearance to the squamous epithelial cells, except that the clue cell borders are ill defined or completely obliterated with a coating of bacteria. The presence of these cells in the urine specimen may be indicative of an infection known as bacterial vaginosis, and they must be noted if viewed in the urine specimen.

There are numerous types of bacteria present as normal flora in the vagina at all times. Bacterial vaginosis is the result of an overgrowth of *G. vaginalis* when the balance of normal flora is disrupted. The use of broad-spectrum antibiotics may destroy healthy bacteria and allow the *G. vaginalis* to increase in number. Douching, retained tampons, use of intrauterine devices for contraception, and use of products that contain nonoxynol-9 may also cause the bacterial balance to shift in the vaginal tissues.

Bacterial vaginosis is the cause of 60% of vulvovaginal infections. It is important to diagnose and treat this condition promptly because it has been linked to an increased risk of pelvic inflammatory disease (PID). Pelvic inflammatory disease is a serious condition that may lead to infertility as well as being a component of many other disease processes.

somewhat teardrop shaped, and a long tail. Urine is toxic to spermatozoa, so it is unusual for these cells to exhibit the same amount of motility as they would in a semen specimen. The clinical significance of spermatozoa in a urine specimen is limited to situations of male infertility and retrograde ejaculation in which the sperm is released into the bladder instead of into the urethra. The presence of spermatozoa should be reported in any situation, whether or not the clinical significance is clear. Elevated amounts of spermatozoa in the urine specimen may cause the urine dipstick to test positive for protein. The presence of spermatozoa is noted with the rest of the microscopic examination results, but the manner in which is it reported is laboratory specific.

Test Your Knowledge 22-3

Is it possible to see spermatozoa in the urine of a female patient? (Outcome 22-3)

Artifacts

Artifacts are contaminants that are visualized in the urine specimen. These may include fecal material, toilet paper, air bubbles, starch or powder, and clothing fibers. Artifacts are not usually reported as part of the urine microscopic examination, because they are not clinically significant. These structures are difficult to identify and may be confused with other formed elements in the urine specimen because of similarities in their shape and size.

Casts

The kidney tubules normally secrete a substance called the Tamm-Horsfall protein. This protein is not clinically significant unless certain conditions exist within the tubules that allow the Tamm-Horsfall protein to "gel" and form a protein-based mold of the tubule. Reduced urine flow, increased amounts of dissolved substances in the glomerular filtrate, and acidic urine conditions contribute to the formation of these **casts.** The casts assume the shape of the tubule or collecting duct in which they were formed. Any structures or substances that were present within the tubule or collecting duct when the cast was formed are trapped within the cast. These trapped substances may include bacteria, cells, or diffuse granules. Eventually, the protein making up the cast breaks away from the lumen of the tubules and flushes out with the urine. These casts can then be identified in the urine sediment when viewed under the microscope.

Casts are cylindrical because of the shape of the structure in which they were originally formed in the nephron tubules. They are classified using the names of the trapped substances within the protein structures. To visualize casts, the intensity of the light source on the microscope must be decreased, and the urine must be examined with the low-power objective. The presence of a few hyaline casts in a urine specimen is not considered abnormal, but the presence of any other type of cast may indicate damage to the kidney tubules. Some common types of casts identified in the urine sediment include the following:

- *Hyaline casts:* Clear, colorless, transparent. Hyaline casts may be difficult to see and to differentiate from threads of mucus. Figure 22-5 shows a hyaline cast. Strenuous exercise, heat exposure, or dehydration may cause an increased number of hyaline casts. Elevated numbers of hyaline casts may also be indicative of pyelonephritis, glomerulonephritis, or chronic renal disease. It is normal to see a few hyaline casts when

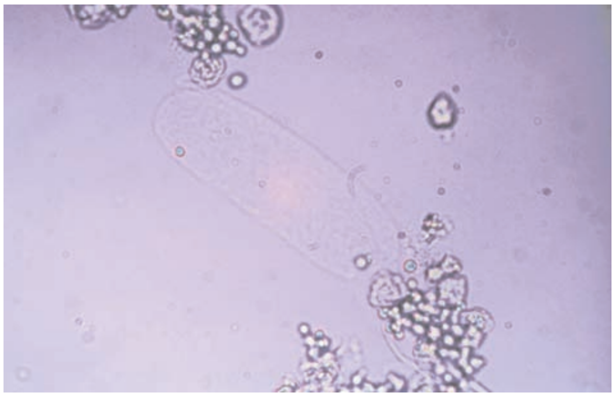

Figure 22-5 Hyaline cast. These casts are "hollow" and have no structures filling their core. *From Strasinger, SK, and Di Lorenzo, MS: Urinalysis and Body Fluids, ed. 5. FA Davis, Philadelphia, 2008, with permission.*

examining the urine under low magnification. Hyaline casts are reported as the number visualized per low-power field.

- *Granular casts:* Granular casts contain fine or coarse granules. These granules are disintegrated structures that were trapped in the renal tubule at the time that the cast was formed. Granular casts may be present in the urine sediment with stress or strenuous exercise, but they may also be evident with glomerulonephritis or pyelonephritis. Granular casts are identified using the low-power objective and reported as the number visualized per low-power field.
- *Cellular casts:* Cellular casts contain renal epithelial cells, red blood cells, or white blood cells. Renal epithelial cell casts are indicative of damage to the renal tubules and are not present in a normal urine specimen. Red blood cell casts may be seen after strenuous exercise, but their presence may also be indicative of bleeding within the nephron structure as in the case of glomerulonephritis. The red blood cell casts have a distinct orange-red color. White blood cell casts are present when there is inflammation or infection within the nephron. There may also be bacteria present in a white blood cell cast. Pyelonephritis is often characterized by the presence of white blood cell casts in the urine sediment. Cellular casts are identified using the low-power objective and are reported as the number visualized per low-power field.
- *Waxy casts:* Waxy casts appear smooth with an irregular outline. They are very refractive when examined with decreased light under low-power magnification. There are no trapped substances within the cast, but the outline is easier to see than that of a hyaline cast. Waxy casts are indicative of an extreme decrease in the urine flow through the tubules that accompanies renal failure, and are not present in a normal urine specimen. They are identified using the low-power objective and reported as the number visualized per low-power field.

Crystals

The presence of **crystals** in normal urine sediment is common. Although many types of crystals are considered normal, it is important to report them when identified because their presence in increased numbers may be clinically significant. Crystals are formed from the dissolved substances (especially salts) in the urine. The precipitation of these dissolved substances into crystals is affected by the pH of the urine, as well as reduced temperature and increased specific gravity. Urine that has been at room temperature or refrigerated for a period of time will often contain abundant crystals in the urine sediment. In addition, highly concentrated urine (urine with a high specific gravity) with many dissolved substances will crystallize readily. The pH of the urine specimen plays a valuable role when attempting to identify crystals under the microscope, as certain types of crystals are evident in acidic urine, whereas others are present in alkaline urine. Various crystals can be seen in Figure 22-6.

Normal crystals may be present in acidic or alkaline urine specimens. These include the following:

- Amorphous (without a distinct shape) urates are present in acidic urine. The refrigeration of the specimen significantly contributes to the formation of amorphous urates in the specimen. The urine sediment may appear pink, and the amorphous crystals will appear as yellow-brown diffuse granules throughout the specimen when viewed under the microscope. Amorphous urate crystals may return to solution with the application of heat or addition of sodium hydroxide to the specimen.
- Uric acid crystals are evident in acidic urine, and appear with numerous shapes and a yellow-brown color. They may be present in increased numbers with leukemia when the patient is receiving chemotherapy, and occasionally in patients with gout.
- Calcium oxalate crystals are occasionally found in acidic urine or, more commonly, in urine with a neutral pH. They are seldom found in alkaline urine. These crystals appear as "envelopes" and are very refractive when viewed under the microscope.

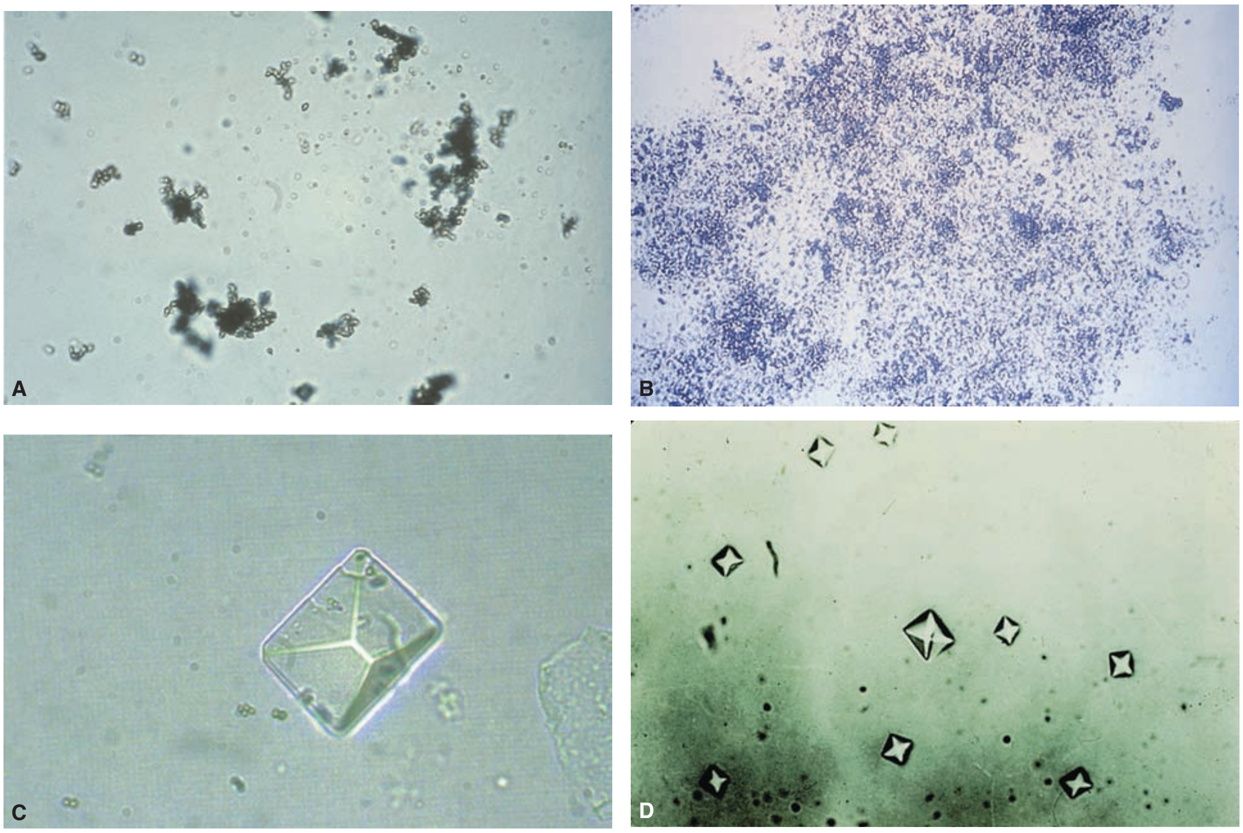

Figure 22-6 Various crystals. A. Amorphous urates. B. Amorphous phosphates. C. Triple phosphate crystals. D. Calcium oxalate crystals. *From Strasinger, SK, and Di Lorenzo, MS: Urinalysis and Body Fluids, ed. 5. FA Davis, Philadelphia, 2008, with permission.*

- Amorphous phosphates are present in alkaline urine. This crystallization is especially evident after refrigeration of the specimen. The amorphous phosphate crystals appear as diffuse granular particles that may obscure the other structures present in the sediment when viewed under the microscope. The urine specimen may take on a white color if amorphous phosphates are present. Amorphous phosphates dissolve with the addition of 10% acetic acid.
- Triple phosphate crystals are present in neutral or alkaline urine. They are colorless, and very refractive when viewed under the microscope. They are often described as a "coffin lid," shaped like a six-sided prism. Their presence is not clinically significant.

The presence of other crystals considered abnormal in the urine sediment may be indicative of metabolic diseases. Use of certain drugs (such as sulfonamides or ampicillin) may also result in crystal formation. Urine specimens with alkaline or neutral pH values rarely contain abnormal crystals. Fortunately, the abnormal urine crystals all have very characteristic shapes that are easily recognizable. Crystals that are considered abnormal include the following:

- Cystine crystals are present in patients with cystinuria, a metabolic disorder in which individuals are not able to reabsorb cystine in the renal tubules as a normal individual would. Cystine crystals are hexagonal.
- Cholesterol crystals are not usually visualized unless the urine specimen has been refrigerated. When they are present, they appear as flat rectangles with notches out of the corners of the structure. Their presence indicates a disorder that includes lipiduria (the presence of lipids in the urine), such as nephrotic syndrome.
- Tyrosine, leucine, and bilirubin crystals may be present in the urine of patients with liver disorders. Tyrosine needles look like clumps of needles, whereas leucine crystals look like yellow-brown, oily spheres. Bilirubin crystals are also described as lumps of needle structures, but they are yellow (like the color of bilirubin).

- Sulfonamide crystals have historically been present in urine sediment in patients undergoing sulfa drug treatment. However, the newer sulfonamide drugs are designed to be more soluble, which means that these crystals are rarely present now. If there are sulfonamide crystals present in the urine specimen, it may mean that the patient is not adequately hydrated. These crystals appear as bundles of needles or designs that resemble sheaves of wheat.
- Ampicillin crystals may also be present in urine when patients receive very large doses of the medication without adequate hydration. They also appear as colorless bundles of needles.
- Hippuric acid is increased in the urine when xylene and toluene are metabolized. Xylene and toluene are common industrial chemicals. Hippuric acid crystals may be an indication of increased exposure to these chemicals, as in workplace exposure or in those who abuse chemicals, such as sniffing glue products.

Microorganisms

When viewed with a compound microscope, various microorganisms may be visualized in the urine sediment. These include bacteria, fungi, and parasites. Smaller microorganisms such as viruses are not visible in the urine sediment with this type of magnification.

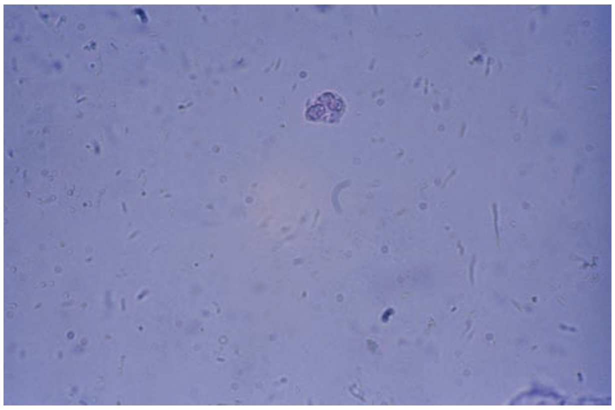

Figure 22-7 Bacteria as they appear in the urine sediment. *From Strasinger, SK, and Di Lorenzo, MS: Urinalysis and Body Fluids, ed. 5. FA Davis, Philadelphia, 2008, with permission.*

- *Bacteria:* It is not uncommon to see a few bacteria in urine specimens that are collected using the clean-catch midstream method. (See Fig. 22-7.) The presence of bacteria is usually the result of external contamination, especially in women. These bacterial contaminants multiply quickly after the urine is collected, and may cause the nitrite on the reagent stick to test positive if the urine is allowed to sit for more than an hour. When present in elevated numbers, the presence of bacteria will cause the urine pH to become more alkaline. (A pH above 8 may indicate the presence of excess bacterial contamination.) Bacteria must be visualized by using the high-power microscope objective because of their small size. The methods of reporting bacteria may vary among labs. Most commonly, bacteria is reported as few, moderate, or many, but a quantitative system of 1+, 2+, or 3+ may be used. Urinary tract infections may be caused by bacilli or cocci. When bacteria are present in the urine specimen in high numbers and white blood cells are evident, the urinalysis is generally followed up with a culture for positive identification of the bacteria and an antibiotic sensitivity.
- *Fungi: Candida albicans* is the most common type of fungi (yeast) present in urine. Yeast is commonly present in the urine of diabetic patients or those who are immunocompromised. Women who have the vaginal infection *candidiasis* may also have yeast present in their urine specimen. Yeast may be difficult to differentiate microscopically from red blood cells, as their size and shape are similar. Advanced yeast infections may present with yeast that is “budding” from stocks or branches as it multiplies and matures. The presence of yeast is reported in the same way that the presence of bacteria is reported.
- *Parasites:* The most common parasite present in the urine specimen is *Trichomonas vaginalis*. This is a sexually transmitted pathogenic microorganism that causes vaginal inflammation in females. Infection in males is often asymptomatic. *T. vaginalis* is often motile when viewed under the microscope in the urine sediment, and it may be recognized by the flagella that provides its movement. The motility of the organism is necessary for positive identification of the parasite. The presence of *T. vaginalis* is determined using the high-power objective and is reported as rare, few, moderate, or many per high-power field.

Test Your Knowledge 22-4

Which of the following microorganisms is not visualized in a urine sediment specimen using a compound microscope?

a. Viruses
b. Bacteria
c. Fungi
d. Parasites (Outcome 22-4)

METHODS USED FOR URINE MICROSCOPIC EXAMINATION

The urine microscopic examination is the part of the routine urinalysis that is the least standardized from one laboratory to another. Each site may vary in the way that the specimen is prepared, in how much urine is used for preparation of the sediment, or by the way that the specimen is visualized during the examination. All procedures include the centrifugation of a specified volume of urine, separation of the **supernatant** (the fluid from the top of the specimen) from the **sediment** (the formed elements that have been compacted to the bottom of the tube), and examination of the urine sediment under a microscope. The formed elements are then examined using the low-power and/or high-power outcomes, and the results are reported.

The results of the microscopic examination may be affected by several factors. The age of the specimen and the conditions under which the specimen was stored can affect the integrity of the structures to be visualized. The volume of specimen centrifuged, the time of centrifugation, and the force of the centrifugation will determine the amount of sediment produced for the microscopic examination. The results may also be affected by the amount of supernatant discarded and the amount of sediment added to the slide for examination. The manner in which the results are reported may also differ according to laboratory protocol. Because of these variables, the Clinical and Laboratory Standards Institute has recommended that the process for preparing and performing urine microscopic examinations be standardized with use of commercial systems that minimize the variables in the process.

The KOVA system (manufactured by Hycor Biomedical, Inc.) is an example of a commercial urinalysis microscopic kit that uses a special pipette to standardize the amount of sediment prepared, has specific instructions concerning the rotations per minute and length of centrifugation to be used, and contains a special slide well used for examination of the sediment that dictates the amount of sediment used for the examination. The KOVA system also includes a stain that enhances the contrast of the formed elements to assist with identification. Procedure 22-1 details the use of the KOVA method for a microscopic examination. There are other commercial systems available, including the Urisystem manufactured by ThermoFisher Scientific and the Count-10 system manufactured by V-Tech, Incorporated.

Test Your Knowledge 22-5

Why might the KOVA system be used to prepare a urine microscopic specimen for examination?

(Outcome 22-5)

Regardless of the system used to process the urine and set up the microscopic examination, it is imperative that a fresh or properly stored urine specimen is used. If the urine specimen is allowed to remain at room temperature for an extended period of time (more than an hour), it may affect the results of the urine microscopic examination, in addition to the chemical analysis of the speicmen. The formed elements present in the specimen will disintegrate and become distorted in specimens that are stored at room temperature for an extended period. The red blood cells, white blood cells, and hyaline casts are most affected. Specimens that are not analyzed within an hour may also exhibit an increase in bacterial growth, accompanied by an increase in pH.

Urine that is refrigerated prior to the performance of a microscopic examination will need to reach room temperature before the procedure is performed. If the urine is still very cold at the time of the examination, the precipitation of crystals will be visualized when viewed under the microscope. The presence of amorphous phosphates and amorphous urates in the refrigerated urine may obscure other clinically significant structures present in the specimen.

Medical assistants are often involved in the preparation of the urine sediment for the microscopic examination. In addition, the medical assistant may be asked to place the slide on the microscope stage and focus it appropriately so that the health-care provider (or another qualified laboratory professional) may perform the examination. To accomplish this, it is important that the medical assistant learn how to focus the microscope so that the formed elements are visible. Procedure 22-2 outlines the important concepts necessary to accomplish this task.

REPORTING URINE MICROSCOPIC RESULTS

As the various formed elements have been presented in this chapter, the reference ranges have been presented as well. It is important to remember that many normal urine specimens will include very few formed elements. If there are structures present in the urine sediment, it is important to report them in a way that is easy to interpret. The exact ranges and terminology used for

Procedure 22-1: Preparation of Urine Sediment for Microscopic Examination

TASK

Prepare a slide for urine microscopic examination using the KOVA system. Complete all steps within 15 minutes.

CONDITIONS

- Hand sanitization supplies
- Gloves
- Laboratory coat
- Face Shield
- Conical tube
- Freshly voided urine specimen
- KOVA System supplies, including tube cap, KOVA pipette, slide, and stain
- Centrifuge
- Test tube rack
- Microscope
- Biohazardous disposal container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.14. Perform Urinalysis **III,P.III.2.** Practice Standard Precautions

ABHES Standards

- Medical Office Clinical Procedures: b. Apply principles of aseptic techniques and infection control, i. Use standard precautions
- Perform Urinalysis

Procedure	Rationale
1. Sanitize hands, allow them to dry, and apply gloves.	Gloves protect the hands from exposure to potential bloodborne pathogens. Gloves should always be applied in the presence of the patient.
2. Gather necessary supplies. Urine must be at room temperature prior to use.	All supplies must be within reach so that the process can be completed in a timely manner. The urine must be at room temperature to avoid excess crystallization of the specimen.
3. Apply face shield and mix urine well by gently rotating the specimen cup or inverting the cup if necessary. Pour 12 mL of the specimen into the KOVA conical tube.	The application of the face shield protects the eyes and mucous membranes from potential splash and aerosol of the specimen. The urine must be well mixed for the results to be accurate, and the KOVA system instructions require 12 mL of urine to be added to the tube.
4. Apply the supplied cap to the tube, and centrifuge the specimen at 1,500 revolutions per minute for 5 minutes.	Centrifugation is necessary for the solid substances in the specimen to settle to the bottom of the tube so that they can be separated from the liquid at the top of the tube. The protective face shield can be removed at this time until the specimen is taken from the centrifuge, at which point it should be put on again.
5. When the centrifuge has come to a complete stop, remove the conical tube carefully.	Take care not to disturb the sediment in the bottom of the tube when handling the specimen.
6. Remove the cap and carefully insert the "petter" pipette included with the system into the urine specimen. Make sure the hook at the top of the pipette is settled into place over the edge of the conical tube.	The KOVA pipette is designed with a special hook device at the top of the apparatus. When the "petter" pipette is inserted correctly into the specimen, the hook will be firmly seated over the side of the conical tube.

Procedure	Rationale
7. With the pipette in place, carefully pour off the supernatant by inverting the tube over the sink to allow the liquid to pour out of the tube.	The KOVA pipette is specially designed with a bulb at the bottom of the pipette that holds the urine sediment in place as the supernatant is poured out of the tube. Approximately 1 mL of sediment is left in the bottom of the tube. 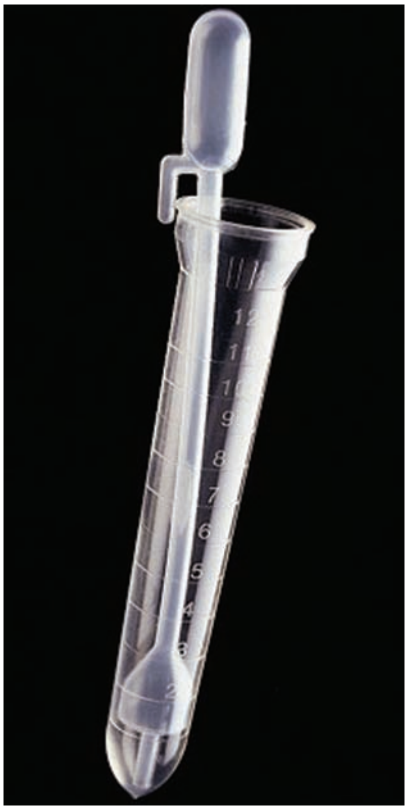 Courtesy of Fisher HealthCare, Inc.
8. Remove the pipette from the tube, but do not discard it at this time. Insert one drop of KOVA stain into the conical urine tube.	Only one drop of stain should be added so that the sediment is not diluted.
9. Reinsert the pipette into the tube and mix the stain with the sediment by application of gentle pressure on the bulb at the top of the pipette. Squeeze and release the bulb until the specimen is uniformly mixed.	The stain and the sediment must be thoroughly mixed, but excessive formation of air bubbles should be avoided.

Continued

Procedure 22-1: Preparation of Urine Sediment for Microscopic Examination—cont'd

Procedure	Rationale
10. Use the pipette to add a drop of the stained sediment to the KOVA slide. The slide is created with an opening at the top where the sediment can be added. The slide must not be overfilled or underfilled. At this point in the procedure, the protective face shield can be removed. Discard the pipette into the biohazardous disposal container.	If the slide is overfilled or underfilled, the results of the microscopic examination may be erroneous. When filling the slide, avoid the formation of air bubbles as much as possible. 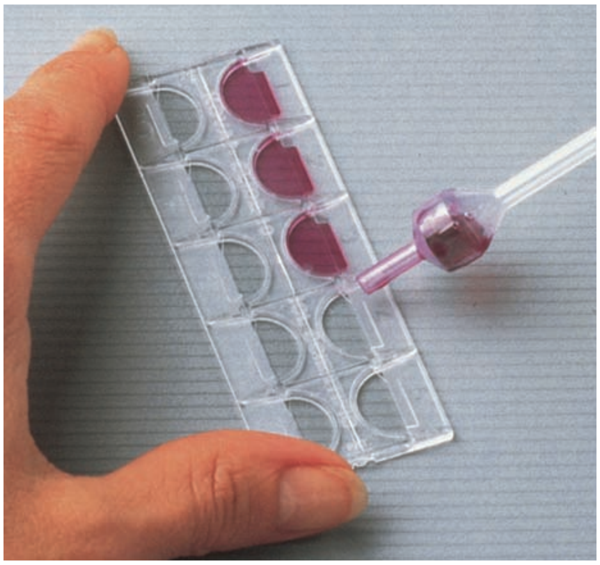 Courtesy of Fisher HealthCare, Inc.
11. Allow the specimen in the slide to settle for at least 1 minute before proceeding to the microscopic examination.	Focusing on the specimen too soon will make it difficult to identify or quantitate the formed elements correctly.
12. Complete the procedure necessary to focus the slide under the microscope.	The medical assistant may be involved in the focusing of the specimen, although he or she will not be performing the actual microscopic examination.

Procedure 22-2: Focusing the Urine Sediment Under the Microscope

TASK

Focus the urine sediment under the microscope to prepare it for identification and quantitation of formed elements. The process should be completed in less than 5 minutes.

CONDITIONS

- Hand sanitization supplies
- Gloves
- Laboratory coat
- KOVA slide with stained sediment added
- Biohazardous disposal container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.14. Perform Urinalysis **III,P.III.2.** Practice Standard Precautions

ABHES Standards

- Medical Office Clinical Procedures: b. Apply principles of aseptic techniques and infection control, i. Use standard precautions
- Perform Urinalysis

Procedure	Rationale
1. Sanitize hands, allow them to dry, and apply gloves.	Gloves protect the hands from exposure to potential bloodborne pathogens. Gloves should always be applied in the presence of the patient.
2. Move the mechanical stage all the way down, away from the microscope outcomes. Place the filled KOVA slide on the mechanical stage of the microscope. Use the clips on the stage to stabilize the slide.	The stage is moved away from the outcomes so that there is room to place the slide on the stage. All microscopic specimens are placed on the stage for examination. The clips allow the slides to be moved as needed for examination.
3. Move the 10X outcome into place for viewing. Minimize the light by using the adjustment knob at the base of the microscope and/or by closing the aperture opening where the light comes through to the slide.	Reduction of the light using the low-power outcome allows for better focus on the formed elements in the specimen.
4. Use the coarse adjustment knob to bring the stage as close as possible to the outcomes.	This will bring the slide into approximate focus when viewed through the eyepieces.
5. Look through the eyepieces, and use the coarse and fine adjustment knobs as needed to bring the specimen into focus. If necessary, move the slide across the stage to find a formed element to use as reference while focusing. Squamous epithelial cells are often present and easily identifiable.	Some specimens have very few formed elements; therefore, it may be necessary to scan several fields before a point of reference may be identified.
6. Inform the health-care practitioner or technologist that the slide is ready to be viewed. The microscopic examination should be completed within 5 minutes of the initial focus so that the specimen does not become dried out on the slide.	If the specimen dries out, it could lead to erroneous results because the concentration of formed elements on the slide can change with evaporation, and the appearance of the elements may change.
7. When the procedure is completed, discard the slide in the biohazardous disposal container, discard the tube with the rest of the stained sediment in the biohazardous waste container, disinfect the work site and microscope as needed, and remove gloves. Sanitize hands.	It is important to disinfect work areas and appropriately dispose of all laboratory specimens for infection control.

reporting may vary among laboratories. The differences in reporting methods may be the result of the system used for the examination (for instance, the commercial systems may have a grid that is used for counting and reporting the elements), or the differences may be the result of laboratory preference. Certain structures are reported as the number present per low-power field, whereas others report the number per high-power field. In addition, some of the formed elements are quantitatively reported as rare, few, moderate, and many, whereas others are reported using real numbers, such as 0 to 5, 5 to 10, and so on. Table 22-1 lists common formed elements and the units used for reporting.

Test Your Knowledge 22-6

True or False: All normal urine specimens contain a high number of formed elements. (Outcome 22-7)

ROLE OF THE MEDICAL ASSISTANT IN MICROSCOPIC URINE PROCEDURES

The examination of urine sediment is not a CLIA-waived procedure. Medical assistants who perform this procedure in a physician office or other laboratory setting must possess the appropriate training and education to perform this moderately complex procedure. Urine microscopy examination may also be performed by the health-care provider in a physician office laboratory, as long as that laboratory has been registered appropriately with the Centers for Medicare & Medicaid Services as a site that is performing **provider performed microscopy (PPM).** As part of the PPM process, the medical assistant will centrifuge the urine sample with added stain if necessary, place the sediment on the slide, and focus the specimen to prepare for the provider's microscopic examination. The medical assistant may also perform routine maintenance on the microscope.

Test Your Knowledge 22-7

List one duty that a medical assistant might perform in reference to the urine microscopic examination. (Outcome 22-8)

QUALITY CONTROL AND QUALITY ASSURANCE PROCEDURES FOR URINE MICROSCOPIC EXAMINATIONS

Quality assurance is a broad term used to describe the policies and procedures that ensure quality laboratory processes and patient care. For urine microscopic examinations, these parameters include the following components of the preanalytical phase of laboratory testing:

- Providing appropriate patient instructions for the collection of the specimen
- Ensuring adequate specimen volume
- Storing specimens correctly
- Using clean supplies
- Calibrating the centrifuge used to spin the specimen
- Separating the urine supernatant correctly
- Setting up the urine sediment correctly on the slide with the appropriate volume
- Focusing the specimen correctly in preparation for examination by a qualified laboratory professional

TABLE 22-1

Common formed elements and method of reporting

Common Formed Elements	Units Used for Reporting
Squamous epithelial cells	Rare, few, moderate, or many report per low-power field (lpf)
Mucus	Not reported by all laboratories; if reported, use rare, few, moderate, or many per low-power field (lpf)
Red blood cells	Average number visualized after counting 10 high-power fields (hpf). Reported as # per hpf
White blood cells	Average number visualized after counting 10 high-power fields. Reported as # per hpf
Spermatozoa	May or may not be reported according to laboratory protocol
Casts	Average number visualized after counting 10 low-power fields (lpf). Reported as # per lpf
Crystals	Reporting methods vary; usually rare, few, moderate, or many per high-power field
Bacteria	Rare, few, moderate, or many per high-power field
Yeast	Rare, few, moderate, or many per high-power field

Quality control refers more specifically to policies and procedures directly relating to the testing process. For urine microscopy, most medical assistants will not directly perform the microscopic quality control procedures because they are not qualified to perform examination of urine sediment under the microscope. However, it is important for medical assistants to know that they are still part of the quality control process. The medical assistant needs to verify that all reagents (such as stains) to be used are not expired. The instructions for use of commercial systems must be followed carefully to achieve desired results. Appropriate maintenance of the microscope is also part of quality control, because it is used directly in the testing process.

Medical assistants who have been properly trained to perform the urine microscopic examination may be responsible for quality control procedures that verify their ability to identify and quantitate the formed elements in the urine specimen correctly. These quality control procedures may include examination of commercially prepared specimens that contain known quantities of various formed elements. The laboratory will also participate in proficiency testing procedures in which results from standardized specimens are compared to those of other laboratories across the country to test for competency.

SUMMARY

The microscopic examination of urine specimens provides valuable information to health-care professionals. As the third part of a complete urinalysis, the knowledge gained from urine microscopy enhances the results obtained from the macroscopic and chemical examination of the specimen. The presence of formed elements in the urine such as blood cells, epithelial cells, crystals, and microorganisms provides valuable information to assist in a diagnosis of urinary tract infections, kidney disease or trauma, or metabolic disorders. The standardization of the specimen processing and microscopic examination procedure allows for more reliability in the results. To perform a urine microscopic examination, medical assistants would need appropriate education and training to meet the requirements for performance of a CLIA moderately complex procedure. However, medical assistants who have not received this additional training may still be responsible for preparing a specimen and focusing it under the microscope for reading by a qualified individual.

TIME TO REVIEW

1. The presence of white blood cells in a urine specimen is called: Outcome 22-1
   - **a.** Pyuria
   - **b.** Hematuria
   - **c.** Hemoglobinuria
   - **d.** Cysturia
2. List two types of formed elements that may be identified in a urine microscopic examination. Outcome 22-3
3. Which of the following may be clinically significant if reported as part of a urine microscopic examination? Outcome 22-4
   - **a.** The presence of a rare white blood cell
   - **b.** Many bacteria
   - **c.** The presence of a rare hyaline cast
   - **d.** The presence of a rare squamous epithelial cell
4. Provide the name of one commercial system used to prepare urine specimens for microscopic examinations. Outcome 22-5
5. When focusing the urine sediment under the microscope, which microscope objective is used initially to bring the speciment into approximate focus? Outcome 22-6
   - **a.** 10X
   - **b.** 40X
   - **c.** 100X
   - **d.** None of the above
6. True or False: The presence of red blood cells is reported as few, moderate, or many per lpf. Outcome 22-7
7. Which of the following represents a typical task performed by a medical assistant in preparing for a urine microscopic examination? Outcome 22-8
   - **a.** Counting the number of white blood cells present in 10 fields using the high-power objective
   - **b.** Identifying the types of crystals present in the specimen
   - **c.** Centrifuging the urine specimen
   - **d.** Processing quality control specimens to verify competency with the procedure
8. True or False: Medical assistants are not responsible for any quality control procedures pertaining to the urine microscopic examination. Outcome 22-9

9. How much supernatant is usually left in the tube with the sediment when preparing the sediment to be placed on a slide for viewing? Outcome 22-10

- **a.** 5 mL
- **b.** 3 mL
- **c.** 0 mL
- **d.** 1 mL or less

Case Study 22-1: When is a specimen too cold?

Cassie Mason is a medical assistant working in a physician office laboratory at which urinalysis tests are performed. Cassie completed all the initial urinalysis test procedures before leaving for lunch. This included notation of the color and appearance of each specimen, as well as the chemical dipstick testing. She cleaned her work area, put all the specimens in the refrigerator, and left the results for the health-care practitioner to review before lunchtime.

When she returned from lunch an hour later, there was a note from the practitioner asking her to set up urine microscopic slides for two of the urine specimens for the practitioner to examine later that afternoon between appointments. About an hour later, Cassie removed the urine specimens from the refrigerator, centrifuged them, and immediately set up the slides and focused them on the microscope. Immediately thereafter, the practitioner came into the laboratory area and looked at the slide of the first specimen. After a cursory glance, she asked Cassie whether this specimen had been refrigerated. When Cassie acknowledged that it had, she told Cassie that she needed to pour another specimen, allow it to sit for at least a half an hour, then recentrifuge, and set up the slide again for viewing.

1. What did the practitioner see in the specimen that made her ask about the refrigeration?
2. How will it help to have the specimen sit at room temperature for a period of time before setting up another slide to be viewed?

Case Study 22-2: Normal results?

A patient drops off a urine specimen that was collected at home to the laboratory. The time of collection noted on the requisition form was several hours before the specimen was delivered to the laboratory. The medical assistant who is working in the laboratory performed a chemical analysis of the specimen, and notes that the pH of the specimen is 8, and that the nitrite result test produces a positive result.

1. In this scenario, what are the three details that provide a possible explanation for the test results obtained?
2. Should this specimen be used for a complete urinalysis?

RESOURCES AND SUGGESTED READING

University of Iowa Carver College of Medicine, practice tests for the Urine Microscopic Examination
Presentation with questions and answers about the urine microscopic examination process http://www.medicine.uiowa.edu/cme/clia/modules.asp?testID=20

"Origin and composition of renal casts"
Explains the origin and composition of the various types of renal casts http://www.agora.crosemont.qc.ca/urinesediments/doceng/doc_016.html

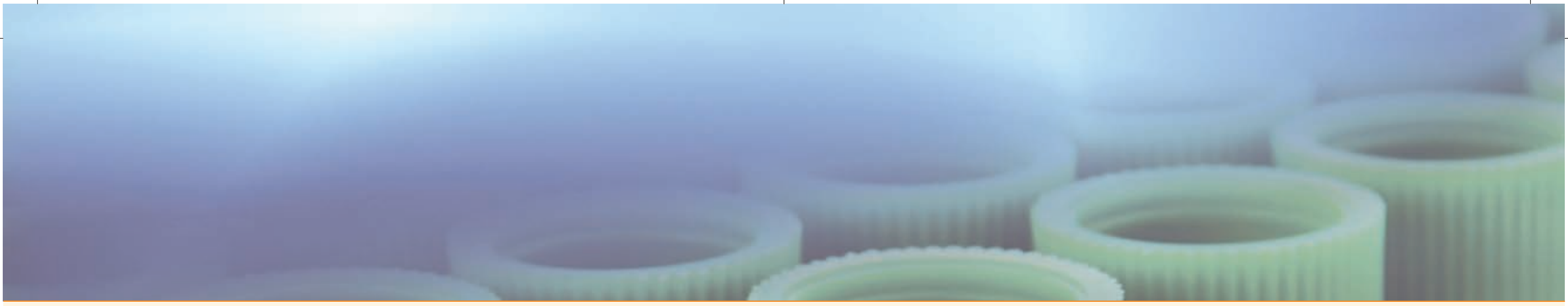

Section V

Urinalysis

What Does It All Mean?

The typical urinalysis test consists of three components: physical examination, chemical examination, and microscopic examination. In many settings only the physical examination and chemical examination are routinely performed. A microscopic examination is only performed when the physical and/or chemical examination results suggest the presence of formed elements. Some situations in which a microscopic examination is typically indicated include the following:

- A sample that is cloudy or turbid
- A sample that is red in color
- A sample that has a large top layer of white or yellow foam
- A sample that has more than trace amounts of protein

It is important to correlate the results obtained in each of the urinalysis components performed. For example, one would expect to detect red blood cells on the dipstick test of a sample noted as being red during physical examination.

Case in Point

Recall Johnny, whom we met at the beginning of this section. He is an active 3-year-old who loves to play in the garage. He was rushed to his doctor's office after drinking antifreeze.

Examination of his urinalysis physical examination results reveals that both the color and clarity of Johnny's sample are abnormal. The fact that his urine is cloudy suggests that formed elements are present and indicate that a microscopic examination should be performed. The most important chemical examination result in this case is protein. Any amount more than a trace is considered abnormal and indicates the need to perform a microscopic examination.

These results are most likely the result of Johnny's ingestion of antifreeze, which is toxic to the kidneys. The kidneys typically respond to such substances by ceasing to produce urine. This stagnate state provides a perfect environment for the development of formed elements known as casts. Casts are cigar-shaped structures that contain a core matrix of a protein known as Tamm-Horsfall. Although this is a different protein from that which the corresponding dipstick test is designed to detect, presence of abnormal amounts of protein suggest that casts may be present. The best next course of action in this case is to perform a microscopic examination on Johnny's urine looking for formed elements, particularly casts.

On the Horizon

Advances in our understanding of how the body's immune system works have exploded in recent years. From this work, important discoveries have emerged. One such example is the fact that we now know that diabetes mellitus is associated with the immune system! No one would have even suspected this just a few short years ago. As a result of discoveries such as this, immunology has truly become a laboratory discipline within itself. Likewise, laboratory testing in this area, also known as immunology, has improved and expanded. In fact, there is much crossover of laboratory disciplines associated with immunology tests. Some laboratories have a section dedicated to performing immunology tests. In other settings, immunologic testing is performed throughout the laboratory.

Relevance for the Medical Assistant (Health-Care Provider)

As with so many other aspects of laboratory testing, medical assistants and other health-care support personnel often participate in immunologic testing by being involved in corresponding sample collection and processing. Furthermore, medical assistants often perform select CLIA-waived immunologic tests along with appropriate quality control. Medical assistants are also responsible for completing the documentation associated with this testing.

Case in Point

Four-year-old P.J. presented to the pediatrician's office where you work accompanied by his very worried father. In your initial conversation with P.J.'s father you learn that little P.J. had been battling a fever and sore throat for more than a week and was not showing any signs of improvement. Upon examination of the patient, the physician noted that the child had numerous yellow "spots" on the tonsillar area of his throat. "No wonder this little guy is so uncomfortable," you thought to yourself. The doctor ordered a strep screen test and requested that you perform the test. You obtain the following results: Rapid strip screen—Negative.

Questions for Consideration:

- What is the specimen type of choice for strep screen testing?
- What is the proper collection protocol for this specimen?
- What is the strep screen test designed to detect?
- What accounts for the negative strep screen results obtained?
- What could be done at this juncture to help determine the microbiological agent responsible for PJ's infection?

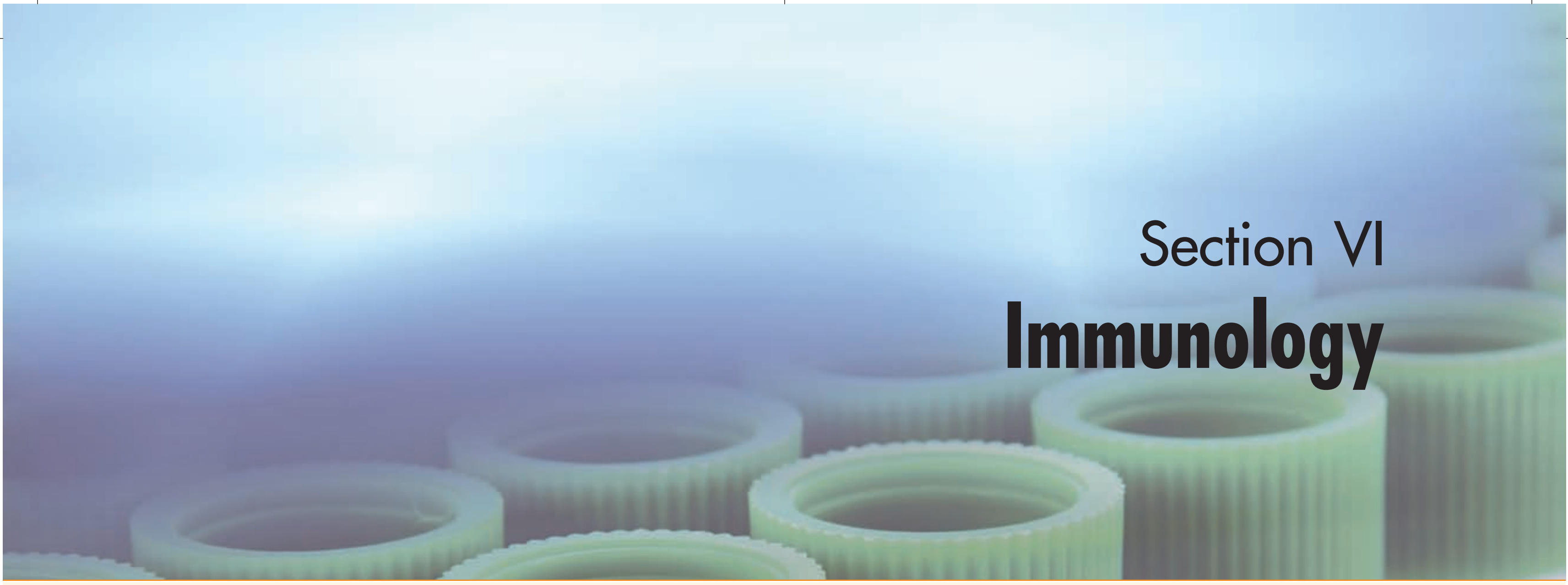

Section VI

Immunology

Chapter 23: Immunology provides the reader with a basic understanding of the immune system. Immunologic principles as they relate to blood typing and hemolytic disease of the newborn are covered. An introduction and overview of common immunology and serology test methods are presented.

Chapter 24: Immunologic-Based Rapid Testing includes a description of the common rapid tests that are routinely performed, including strep screen, infectious mononucleosis testing, pregnancy testing, *Helicobacter pylori* testing, and testing for influenzas A and B and HIV. The nature of each condition associated with these tests is examined. Appropriate follow-up testing is briefly discussed when indicated. Proper technique, normal test reference range, quality control, sources of error and interference, and test interpretation are described.

Chapter 23

Immunology

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

Immunity and Immunology
The Immune Process
  The First Line of Defense: Nonspecific Mechanical and Chemical Barriers
  The Second Line of Defense: Internal Nonspecific Response
  The Third Line of Defense: Acquired or Adaptive Immunity
How Immunity Is Acquired
Failure of Our Immune Systems
Blood Types
Erythroblastosis Fetalis
Immunology Testing Methods
Common Serology Tests Performed in Reference or Hospital Laboratories
CLIA–Waived Tests Commonly Performed in the Physician Office Laboratory
Summary
Time to Review
Case Study
Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

23-1 Define the key terms.

23-2 Explain how serology and immunology are related to immunity.

23-3 Describe the mechanical barriers used by the immune system.

23-4 Explain how the internal defenses at the site of an invasion by foreign substances function.

23-5 Summarize how acquired immunity is activated in the body.

23-6 Distinguish what happens when the immune process malfunctions.

23-7 Examine how vaccinations activate the immune system.

23-8 Demonstrate understanding of the basic ABO and Rh blood typing systems and the role they play in hemolytic disease of the newborn.

23-9 Describe various immunology testing methods and common serology tests performed in hospital and reference laboratories.

23-10 Explain why some serology tests might be performed in physician offices rather than in reference laboratories.

23-11 Explain the clinical significance of the common CLIA-waived serology tests performed in physician offices.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.C.5. Describe the normal function of each body system
I.C.6: Identify common pathology related to each body system
III.C.10: Identify disease processes that are indications for CLIA-waived tests

ABHES Standards

- Identify and apply the knowledge of all body systems; their structure and functions; and their common diseases, symptoms, and etiologies

KEY TERMS

Agglutination
Allergy
Antibody
Antibody-mediated immunity
Antigen
Autoimmune disease
B cells
Cell-mediated immunity
Chemotaxis
Cilia
Complement
ELISA or EIA testing
Erythroblastosis fetalis
Gammaglobulins
Globulin
Hemolytic disease of the newborn (HDN)
Histamine
Immunity
Immunoglobulins
Immunology
Infection
Inflammation
Interferons
In vitro
In vivo
Lateral flow immunoassay
Lysozyme
Macrophage
Mucous membranes
Natural killer (NK) cells
Normal flora
Pathogens
Phagocytes
Phagocytosis
Pus
RhoGam
RIA
Serology
T cells
Vaccine

IMMUNITY AND IMMUNOLOGY

Our bodies are constantly under attack by an army of microorganisms, toxins, allergens and other substances that are recognized as foreign or "nonself." Luckily, our defenses are usually able to protect us from this ongoing attack. The recognition and destruction of these foreign substances by our bodies are known as **immunity**, and the study of this process is called **immunology**.

Immunity includes numerous protective mechanisms within our bodies. In order for these systems to function properly, we must be able to detect what is "self" and what is "nonself." When a foreign (or nonself) invader is detected, the immune process is activated. A foreign substance capable of initiating an immune response is called an **antigen.** The antigen may be a microorganism (such as bacteria or viruses), an allergen (such as pollen or dander), a foreign body (such as a sliver), or even a cancerous cell from within the body.

Antibodies are substances that the body produces in response to certain antigens. The antibody is designed to attach to, or bond with, the specific antigen targeted. Antibodies are made of a type of protein called **globulin,** and because they are part of the immune process, they are known as **immunoglobulins.** The job of immunoglobulins is complex, and includes attachment to and destruction of the invader, as well as remembering the unique identity of the antigen so that if this invader comes into the body again, the body is ready to attack and destroy much faster than it was at the first exposure. Antibodies seek out and irreversibly bind to their specific antigen, rendering it ineffective.

A great deal of research has been performed on the immune responses of our bodies. We are now able to test for specific antibodies and antigens in the laboratory, using blood and other body fluids. This area of the laboratory may be known as the Immunology Department, but more commonly it is referred to as the Serology Department. **Serology** is the study of antigen and antibody reactions in serum, and may include methods such as radioactive laboratory testing, agglutination reactions, or simplistic tests such as those used to detect pregnancy.

Test Your Knowledge 23-1

How are antigens and antibodies related to immunity?

(Outcome 23-2)

Test Your Knowledge 23-2

Which department in the laboratory provides testing methods to monitor the immune system?

(Outcome 23-2)

THE IMMUNE PROCESS

If you've ever watched a movie about the Revolutionary War or Civil War, it is difficult to see the soldiers march forward on the battlefield in straight lines again and again to meet their enemies. Eventually, with a great deal of effort and substantial losses, this method of combat might affect the outcome of the encounter. On the other hand, the battles won with a minimal number of casualties were those that employed *multiple* strategies, or lines of defense.

We wage a similar battle in our bodies every day. The sheer number of foreign substances we encounter is overwhelming. We need more than one line of defense to defeat the invaders with minimal harm to ourselves. A healthy immune system makes use of three basic defense mechanisms (Fig. 23-1). These include mechanical and chemical barriers, internal defenses at the site of the invasion, and acquired or adaptive immunity created by the body in response to a specific antigen. The first two lines of defense are nonspecific, and function in a similar manner with any invader they encounter. These nonspecific mechanisms are protecting us all from birth or very soon afterward. The third defensive mechanism, however, is very specific and is something that we develop throughout our lives.

The First Line of Defense: Nonspecific Mechanical and Chemical Barriers

The first line of defense used by our immune system includes mechanical barriers, chemicals, organisms that are always present in our bodies, and protective reflexes. The goal of these mechanisms is to keep the invaders out of our bodies or prevent them from surviving long enough to cause any further harm.

Intact skin and the **mucous membranes** of our bodies are extremely important because they form a physical or mechanical barrier to many pathogens and other types of invaders. Their secretions also play an important role, by adjusting the pH to deter growth of most pathogens, and by trapping other invaders in a sticky environment where they can be destroyed or immobilized. The **cilia**

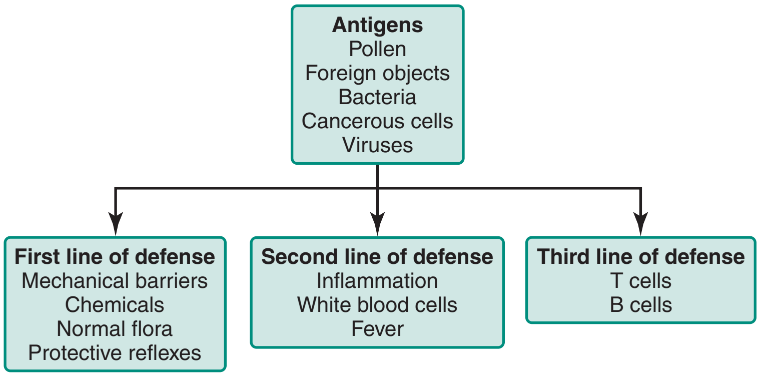

Figure 23-1 Flow chart showing the different kinds of immune responses and their triggers.

lining the mucous membranes constantly work to sweep out any trapped substances. Reflexes, such as sneezing or coughing, then move these potential **pathogens** out of the body. The acidic environment in our stomach and the enzymes in saliva also assist with this first line of defense. Earwax protects our ears, perspiration helps to flush away unwanted bacteria, and an enzyme in tears called **lysozyme** destroys many potential invaders.

In addition to the protective mechanisms already mentioned, our bodies are populated naturally by numerous types of bacteria. Although this might seem to be a bad thing, it is actually quite beneficial to our immune system. The presence of this **normal flora** creates a competitive environment in which most invading bacteria cannot survive. We have normal flora throughout our bodies, including in and on our skin, in our mouths, and in our intestinal tract.

Test Your Knowledge 23-3

How does a break in the skin trigger the first line of defense to respond? (Outcome 23-3)

Test Your Knowledge 23-4

Are sneezes a good thing? Why or why not? (Outcome 23-3)

The Second Line of Defense: Internal Nonspecific Response

If a foreign substance makes its way past the first line of immune defenses, it will be confronted with a wave of new defense mechanisms within the body. These are all part of the internal nonspecific response, also known as the second line of defense. Included in this arsenal are phagocytosis, inflammation, fever, and natural killer cells. This second line of defense is also known as natural or innate immune response.

When cells are injured, they release a chemical into the bloodstream that signals the second line of defense to activate. This signaling is called **chemotaxis,** and it attracts certain types of white blood cells to the scene of the invasion to act as **phagocytes.** The white blood cells (neutrophils and monocytes) are specially programmed to ingest and destroy the invaders. This process is called **phagocytosis.**

When a foreign substance has entered the body, we often observe the surrounding area become red, hot, and swollen. The internal process that causes this appearance is called **inflammation,** and if it is caused by a pathogen, it may be known as **infection.** Inflammation occurs as a result of injury to or irritation of a cell. When the cells become irritated or injured, they release **histamine.** The histamine causes the blood vessels in the area to dilate, or become larger. The blood flow increases to the irritated area, and this causes the heat and redness we see and feel. Histamine also causes the blood vessel walls to become "leaky," so more fluid leaks out of the bloodstream into the surrounding injured area, causing swelling. Pain is evident because the surrounding nerve receptors are stimulated by all the activity.

As the blood supply is increased, more phagocytes are brought to the area. While doing their job, these cells create a lot of debris. The built-up dead phagocytes, fluid, pathogens, and injured blood cells that accumulate are what we recognize as **pus.**

The next symptom we often note in a pathogenic invasion is fever, which occurs because the phagocytes that respond to the scene of the injury release a chemical that signals the brain to increase the body's temperature. The elevation in temperature helps the second line of defense by decreasing the ability of many pathogens to survive, as well as stimulating even more phagocytosis.

As if all the other activity weren't enough, protective proteins also assist at the scene of the invasion. There are two groups of protective proteins, known as **interferons** and **complement** proteins. Interferons are secreted by cells that have been infected by viruses, and they help other cells resist the spread and replication of the virus. Complement proteins are able to destroy bacteria by covering the surface of the bacterial cells and punching holes in the cell membrane, causing the bacteria to die.

The last component of the second line of defense is the **natural killer (NK) cells,** which are a special type of lymphocyte that responds to the invasion. NK cells can attack and destroy human cells that have been infected by pathogens. In addition, natural killer cells may be beneficial by destroying certain cancer cells before they can spread.

The Third Line of Defense: Acquired or Adaptive Immunity

Acquired or adaptive immunity is a defense mechanism that requires the body to "think" about the invaders. It involves three characteristics that are not present in the first or second line of defense. Recognition of an antigen as nonself must first occur, after which the antigen (invader) is identified very specifically. Then the body forms a memory of that antigen, so it will be immediately recognized in case of subsequent exposure. Usually this entire process requires 1 to 2 weeks after the first exposure to a specific antigen.

T cells and **B cells** are specifically programmed lymphocytes. They are both part of this third level of the immune process, but they each perform different functions when exposed to an antigen. T cells are involved in cell-mediated immunity, and B cells are involved in **antibody-mediated immunity,** also known as humoral immunity.

Cell-mediated immunity begins at the site of the invasion (Fig. 23-2). The T cells become "activated" toward a specific antigen when it is presented by a **macrophage.** (A macrophage is a type of phagocyte that leaves the bloodstream and goes into the tissues.) Once activated, this T cell creates four identical copies of itself, which all have specific functions:

- Killer T cells are produced to kill the antigens.
- Helper T cells are designed to stimulate other T cells and help the B cells.
- Suppressor T cells (also known as regulatory T cells) are created to inhibit the activity of the T and B cells when enough have been produced.
- Memory T cells are responsible for remembering the antigen for future encounters, which allows the cell-mediated immunity to occur faster with the next exposure.

Most of the lymphocytes in our circulation are T cells. Although no antibodies are produced by the T cells, they play an important role in many contact sensitivity reactions, as well as viral infections, fungal infections, and destruction of malignant cells.

Antibody-mediated immunity uses the lymphocytic B cells (Fig. 23-3). As with the cell-mediated immunity, the antigen must first be exposed, or introduced, to the B cell by presentation from the macrophage. The helper T cells release a chemical to assist with this recognition process. Once activated, the B cell produces two replicas of itself, which have their own duties:

- Plasma cells are created that secrete antibodies that are specific for a particular antigen. These antibodies

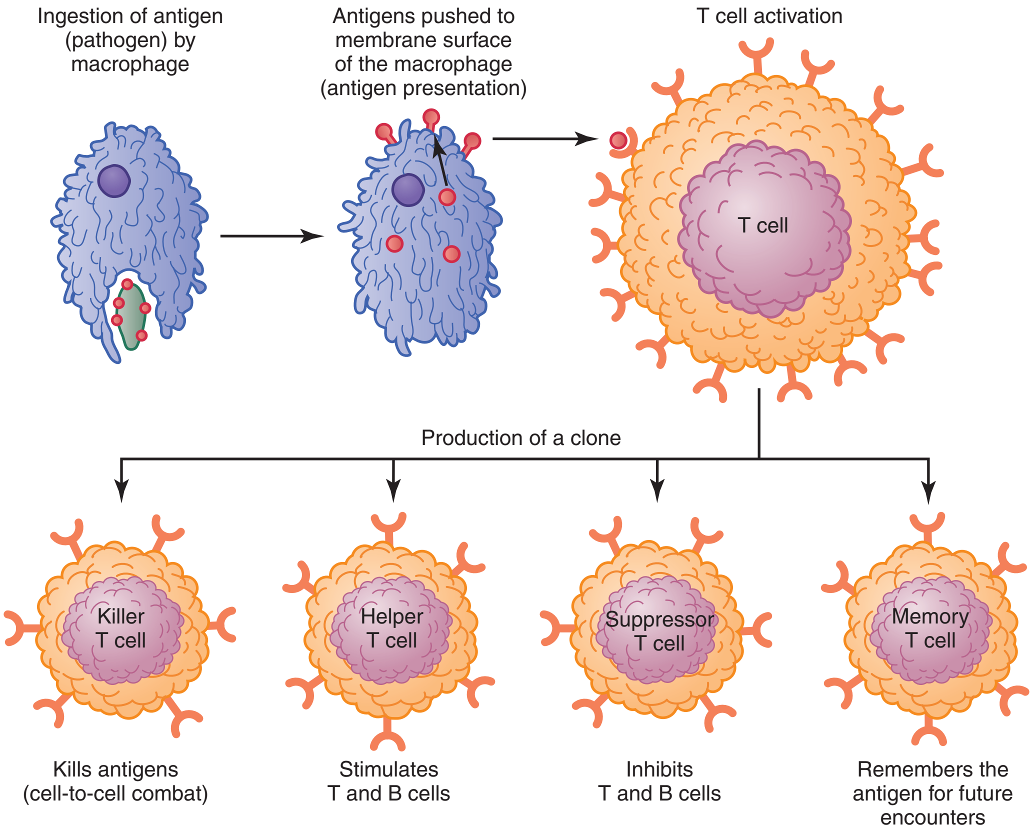
Ingestion of antigen (pathogen) by macrophage
Antigens pushed to membrane surface of the macrophage (antigen presentation)
T cell activation
T cell
Production of a clone
Killer T cell
Helper T cell
Suppressor T cell
Memory T cell
Kills antigens (cell-to-cell combat)
Stimulates T and B cells
Inhibits T and B cells
Remembers the antigen for future encounters

Figure 23-2 Cell-mediated immunity. This drawing represents the T-cell activation and the subsequent steps, creating four subsets of T cells.

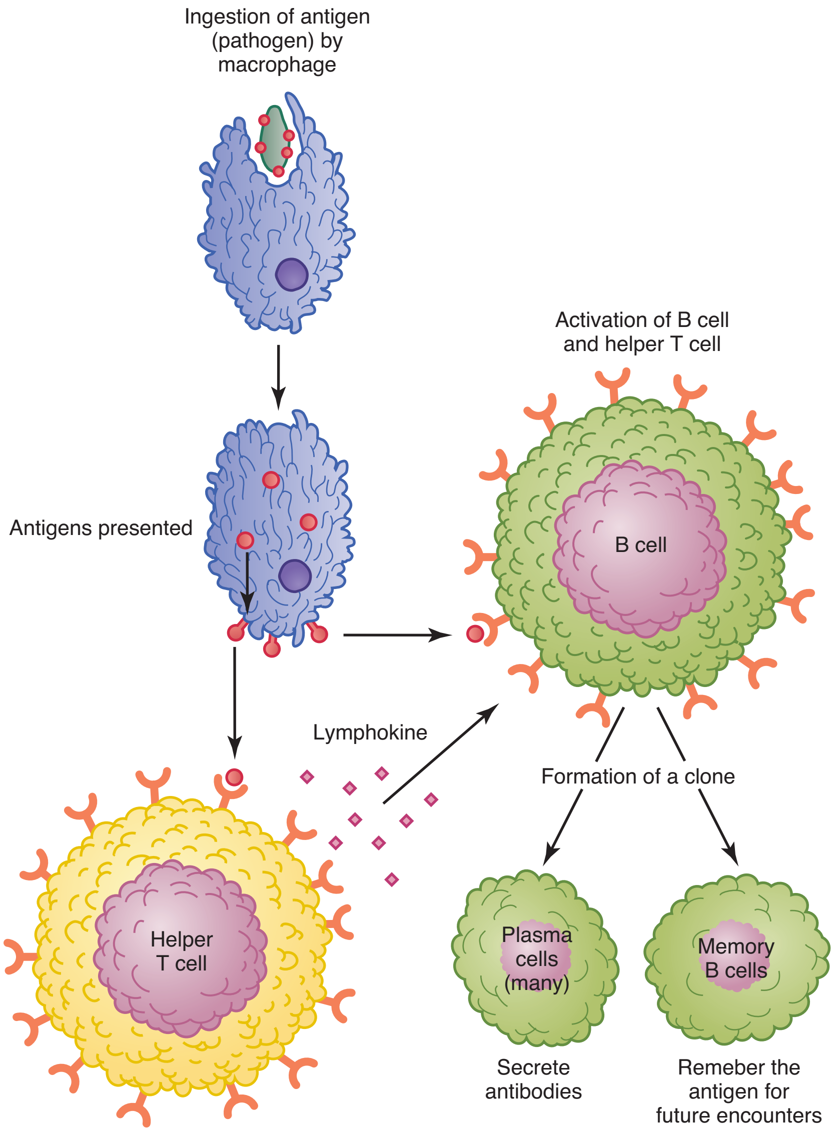
Ingestion of antigen (pathogen) by macrophage
Antigens presented
Activation of B cell and helper T cell
B cell
Lymphokine
Formation of a clone
Helper T cell
Plasma cells (many)
Memory B cells
Secrete antibodies
Remeber the antigen for future encounters

Figure 23-3 Antibody-mediated immunity. B-cell activation leads to antibody creation by plasma cells.

work like a lock and key and cannot be interchanged with other antigens. The antibodies bind to the antigens, blocking their ability to affect the cells of the body.

- Memory B cells form a "picture" of the antigen and remember it for future encounters.

Test Your Knowledge 23-5

True or False: Antigen exposure always triggers all three lines of defense. (Outcome 23-4)

Test Your Knowledge 23-6

Which line of defense is "unique" toward a specific antigen? (Outcome 23-4)

HOW IMMUNITY IS ACQUIRED

There are essentially four ways we can acquire immunity:

- *Natural active immunity acquired from an infection with a pathogen:* This causes the body to produce a memory of the antigen and protects the body on subsequent exposures. For instance, natural active immunity is produced when a person is infected with chicken pox. Once the person recovers, he or she is now immune to this pathogen. This is a long-term type of immunity.

- *Artificial active immunity acquired from a* **vaccine:** The typical antibody response of our body after exposure to an antigen is usually too slow to protect us from the initial infection. However, with subsequent exposures, the protection is so fast that we never become ill. Vaccines do this artificially. They contain an antigen that causes the body to develop antibodies as if it were being exposed for the first time. The vaccine may be a killed or weakened pathogen (known as *attenuated)* or it may just be part of a pathogen. Vaccines are available for many diseases, such as hepatitis A and B, measles, and mumps. The protection is long term with this type of immunity.
- *Natural passive immunity, such as the antibodies a mother passes on to an infant through the placenta or through breast milk:* This is a temporary type of immunity and fades as the child grows.
- *Artificial passive immunity:* This usually occurs with the injection of **gammaglobulins** (artificial antibodies). This immunity is short acting, and protects the individual in an emergency situation. For example, a family member has been diagnosed with hepatitis A, but the rest of the family is unvaccinated for that disease. The other family members are then injected with an "artificial antibody" (gammaglobulin) so that they will be protected until the infected individual recovers from the hepatitis A infection.

Test Your Knowledge 23-7

What does it mean when we say that we are "immune" to a certain disease? (Outcome 23-5)

Test Your Knowledge 23-8

How do vaccines "fool" our immune systems? Which line of defense do they target? (Outcome 23-5)

POINT OF INTEREST 23-1

Vaccines

Vaccines are used to prevent disease by introducing an antigen into the body to induce an immune response. After the vaccine is processed by the body, antibodies are created against that specific antigen. When the patient is exposed again to the disease-causing virus or bacteria, the body immediately produces large amounts of antibody that prevent the disease from developing.

The vaccine antigen may be attenuated, meaning that the antigen it is killed or weakened. Vaccines may also include parts of bacteria or virus particles that are capable of stimulating an immune response but not capable of causing disease. In addition, an inactivated toxin produced by the bacteria (called a toxoid) may be used. The tetanus and diphtheria vaccinations are toxoid vaccines.

FAILURE OF OUR IMMUNE SYSTEMS

Unfortunately, our immune systems don't always work as they should. Sometimes our body's ability to recognize self and nonself is not fully developed, or is somehow dysfunctional. In this case, our body may attack its own cells, causing damage. This is the theory behind many forms of **autoimmune diseases,** such as rheumatoid arthritis. Also, as we age, our immune systems don't seem to detect the malignant or damaged cells in our bodies and destroy them as they should. This allows the cells to reproduce and cause cancer at a higher rate as we age.

In the case of an **allergy,** our immune system becomes hypersensitive after the initial exposure to an antigen. With subsequent exposures to that antigen, the immune response causes inflammation and organ dysfunction. Allergies may range from those with annoying symptoms (such as hay fever) to those that are truly life threatening, such as reactions to insect stings and certain medications. Table 23-1 provides examples of diseases associated with the immune system.

BLOOD TYPES

Blood types also involve antigens and antibodies, but in a different way. The antigens that determine blood types are on the surface of all red blood cells. Blood typing also involves natural and acquired antibodies in the plasma. There are several situations that might warrant a test for blood type, including transfusions, testing to prevent hemolytic disease of the newborn, and for investigative or forensic reasons. There are hundreds of different antigens present on the red blood cells of our bodies, all capable of causing transfusion complications. Fortunately, most serious complications from transfusions are due to the ABO and Rh blood groups, so much of our testing is focused on those groups.

TABLE 23-1

Common immune system disorders

Classification	Examples of Specific Disorders
Hypersensitivity (allergy)	Anaphylactic reactions, dermatitis or eczema, rhinitis, asthma
Autoimmune disorders	Systemic lupus erythematosus, type 1 diabetes, Graves' disease, celiac disease, myasthenia gravis, rheumatoid arthritis, systemic sclerosis
Malignancies of immune cells	Lymphocytic leukemias, lymphoma, multiple myeloma
Congenital immune disorders	Thymic hypoplasia, agammaglobulinemia
Secondary or acquired immunodeficiency	AIDS, transplant anti-rejection medication use, systemic infections, side effect of radiation treatment, overuse of antibiotics

The ABO blood types include A, B, AB, and O, and they represent the antigens present (or absent) on the red blood cells of an individual. If people have a blood type of A, this means they have the A antigens on their cells. If B antigens are present, they are blood type B, and if both are present, the blood type is known as AB. Type O may be thought of as type "zero," as it means neither A or B antigens are present (Fig. 23-4).

The ABO blood typing system is unique, because in the plasma of each person there are natural antibodies to the antigens that are *not* present on their red blood cells. This means a person with a type A blood type has anti-B naturally in his or her plasma, and the plasma of a person with type B blood will contain antibodies to the A antigen. An individual who is type AB has neither anti-A nor anti-B in the plasma, and type O individuals naturally possess both anti-A and anti-B in their plasma. These plasma antibodies are created within the first few months of life as an immune response to dietary antigens and exposure to various environmental factors.

When patients receive blood transfusions, they generally receive a blood product known as packed red blood cells rather than whole blood. To prepare packed red blood cells, the unit of donor blood is centrifuged so that the plasma is separated from the cells, and most of the plasma and the white blood cells are separated from the red blood cells. Removal of the plasma avoids an overload of fluid for the patients needing multiple units to increase their red blood cell and hemoglobin concentration. The absence of white blood cells in the blood product helps to reduce the potential for an immune response.

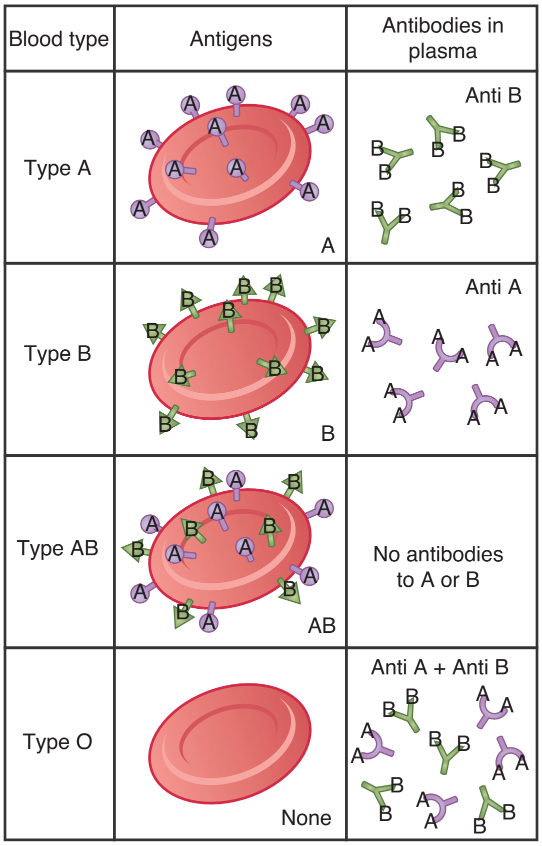

Figure 23-4 Various blood type antigens and antibodies.

It is critical to consider the natural antibodies present in the plasma if a transfusion is needed. Ideally, a person should only be transfused with his or her own blood type. In an emergency situation, a person may be given a different type, but there are only a few alternatives that are safe. If a type B person received type A blood, the natural anti-A antibodies would cause the cells to clump (agglutinate), and eventually hemolyze (rupture). The by-products of this reaction could also clog the capillaries in the kidneys and cause renal failure. Blood type O is known as the universal donor, because there are no ABO antigens on the cells. When type O packed red blood cells are transfused to a patient with any other blood type, they are not rejected by the recipient and the natural anti-A and anti-B that was present in the plasma of the donor are not present in the packed cells to damage the recipient's blood cells. Type AB is known as the universal recipient, because there are no ABO antibodies in the plasma to damage any blood type that is transfused to these recipients. About 40% of the population is type A, 11% is type B, 45% is type O, and 4% is type AB (Table 23-2).

TABLE 23-2

Percentage of various blood types in the U.S. population

Blood Type	Percentage of U.S. Population
A	40%
B	11%
AB	4%
O	45%

Another very important red blood cell antigen is the Rh or D antigen. It was discovered in 1939 with work on rhesus monkeys, hence the designation Rh. Approximately 85% of the population has the D antigen on their cells, and they are known as Rh positive. Those who don't have the Rh (D) antigen on their cells are known as Rh negative.

Unlike the ABO blood types, Rh-negative individuals do not have naturally occurring Rh antibodies in their plasma. However, they can be sensitized to this foreign Rh antigen if exposed. This exposure may occur with a transfusion error, or it may occur with pregnancy and delivery. There is usually no problem evident with the first Rh exposure because the antibody production takes too much time. However, the antigenic response is *very* swift and strong with subsequent exposures to Rh-positive blood.

ERYTHROBLASTOSIS FETALIS

Erythroblastosis fetalis is a severe hemolytic anemia that affects Rh-positive newborns and occasionally other newborns with ABO incompatibility. This condition may also be known as **hemolytic disease of the newborn (HDN).** The Rh status of the blood cells is only an issue when the baby is Rh positive and the mother is Rh negative. During a first pregnancy, this incompatibility does not usually cause harm to the baby, as the mother's response to the Rh antigen does not happen quickly enough for the antibodies to attack the baby's red blood cells. However, during the first pregnancy, and especially during delivery of the baby and the placenta, there is an exchange of fetal blood with the mother, and the mother will become sensitized, which means that she will develop antibodies to the Rh antigens on the baby's red blood cells. If she becomes pregnant again with an Rh-positive baby, her antibodies to the Rh antigens will destroy the baby's red blood cells and the baby may be born with severe anemia, jaundice, and an enlarged spleen and liver, which may be fatal for the newborn.

An injection of **RhoGam** (anti-Rh gammaglobulin) during the pregnancy and shortly after delivery should be given to all Rh-negative mothers. This will keep the mother from becoming sensitized to the Rh antigen so she does not create the anti-Rh antibodies. When she becomes pregnant again with an Rh-positive baby, her body will react as if this were her first exposure, and the Rh incompatibility will not be a problem.

POINT OF INTEREST 23-2

Kidney transplants for patients with incompatible blood types

Many potential kidney donors are rejected because their blood type is not compatible with that of the patients needing a kidney. There is now a process that allows patients to receive a kidney from a donor with an incompatible blood type.

The natural antibodies present in the plasma of the recipient of a kidney with an incompatible blood type usually cause the kidney to be destroyed. The process for ABO-incompatible kidney transplantation includes a procedure called plasmapheresis. The patient has a unit of blood removed from his or her body, and a portion of the liquid plasma is then removed. After each of the plasmapheresis procedures, the patient will receive replacement antibodies that are necessary to fight off other infections. The procedure may need to be repeated numerous times, until the level of ABO antibody detected in the blood is low enough to be considered safe for the transfusion.

Kidney transplant rejection rates for this procedure are not much different from the rate for patients who receive kidneys of the same ABO blood type. The patient will have more plasmapheresis and artificial antibody treatments after the transplant, depending on the individual needs of the patient. In addition, the recipient may also require a splenectomy, because the spleen is the organ in which the ABO antibodies are produced.

Test Your Knowledge 23-9

What antigens are present on the blood cells of an individual with a blood type of AB positive?

(Outcome 23-8)

Test Your Knowledge 23-10

A fellow student just found out that she is expecting a baby. Her blood type is A, Rh negative. Will she be likely to receive a RhoGam injection? Why?

(Outcome 23-8)

IMMUNOLOGY TESTING METHODS

Various immunology testing methods are available for use in the medical office as well as hospital and reference laboratories. They can be performed **in vivo** (inside a living organism) as in allergy or tuberculosis testing; an antigen is placed under the skin, and the patient is monitored for a visible reaction. More often, they are performed **in vitro** (within a laboratory testing environment), where we commonly refer to them as serology tests.

Serology tests must be specific to the antigen or antibody being tested, and they must also be sensitive enough to detect even a small amount of the substance in the specimen. Many medications or other concurrent disease states can interfere with the specificity of a certain test, and sometimes can cause false-positive or false-negative results. The sensitivity of a testing method can be affected by high or low temperatures during shipping or storage, and also by minor errors that occur during the testing procedures. Package inserts should be examined carefully for specimen requirements, interfering substances, storage and handling directions, and testing procedures before performing any sort of serology testing.

All immunology tests are performed to determine the presence of a particular antibody or antigen. The specimen tested may be blood, serum, urine, oral fluids, or other body tissues. Various testing methods are used, including the following:

- Enzyme-linked immunosorbent assay tests **(ELISA or EIA)** tests usually involve a testing surface that is coated with an antibody or antigen bound to an enzyme. After the specimen is added, if the matching antibody or antigen is present, a color change will occur that can be read either by the testing personnel directly or by using an instrument to measure the amount of color change.
- An adaptation of the EIA principle, **lateral flow immunoassay** is used in many CLIA-waived tests. The specimen must be absorbed and must “flow” across the testing surface to reach the area imbedded with the antigen or antibody. If the test is positive, the antigen/antibody complex will cause a color change to occur in the testing area. These tests are read directly by the testing personnel, and have a built-in quality control area as well.
- Radioimmunoassays **(RIAs)** use a radioactively labeled protein as the indicator of a positive result. The antigen/antibody complex formed in a specimen that is positive will combine with the radioactive protein rather than an enzyme as in the EIA method. The final result is obtained from the amount of radioactivity present in the sample at the end of the testing process. These assays are complicated, and are performed only by qualified personnel in hospital or reference laboratories.
- Agglutination is the formation of large insoluble aggregates (clumps) of cells that form when an antigen/antibody complex is present. Agglutination may use red blood cells and antibodies, as in the case of ABO and Rh blood typing, or it may use latex beads, as do some of the mononucleosis testing procedures. This reaction is read by the testing personnel directly, as the complex is visible to the naked eye.

COMMON SEROLOGY TESTS PERFORMED IN REFERENCE OR HOSPITAL LABORATORIES

Serology tests may share common methodology, but the complexity of the testing process can vary greatly. The tests mentioned below are usually performed only in reference or hospital settings, as they require additional training and/or credentials for the person performing the test.

- Testing for *ABO/Rh* major blood antigen groups uses agglutination techniques. Screening is also performed for other common red blood cell antigens, and compatibility tests are performed on the donor blood when a transfusion workup is necessary.
- *The antistreptolysin O (ASO)* test is used to detect antistreptolysin O antibodies, which are present in a patient with prior recent infection by group A streptococcus. The identification of those with prior infection is important, as group A streptococcus infection can lead to other complications such as glomerulonephritis, rheumatic fever, and bacterial endocarditis. The ASO antibody may be present for several months after the initial infection.
- *Cold agglutinins* are antibodies present with certain types of infections, pneumonias, cancers, or inflammatory autoimmune disorders. These antibodies can cause the red blood cells of the body to “clump” together when exposed to cold temperatures. This leads to hemolytic anemia and other complications.

- *The C-reactive protein (CRP)* test verifies the presence of that protein, which is present in the blood with acute systemic inflammation. It may be used as part of the diagnosis for a specific condition related to inflammation, or to monitor the effectiveness of treatment.
- *Hepatitis testing* determines which form of viral hepatitis (A through E) an individual has, and it may also provide more information about where an individual is in the disease process.
- Although a *pregnancy test* can be performed at home or in a physician office laboratory as a CLIA-waived test, it is sometimes necessary to use the more sensitive methods that are available in a larger laboratory. These methods can sometimes detect pregnancy sooner than can the CLIA-waived tests, and they can also quantitate the amount of human chorionic gonadotropin (HCG) present in the specimen. This quantitative analysis can be vital in the case of a suspected ectopic pregnancy, or in the diagnosis and treatment of testicular cancer, in which the hormone is also present in the bloodstream in male patients.
- Rheumatoid arthritis is a crippling and progressive autoimmune disease that affects the joints of the body. Those afflicted will have the *rheumatoid factor* present in their blood, which is a unique antibody used for diagnosis.
- *The most common tests used to detect syphilis* are the rapid plasma reagin (RPR) test and the venereal disease research laboratory (VDRL) test. The tests detect the presence of the microorganism *Treponema pallidum,* the causative agent of syphilis.

Test Your Knowledge 23-11

What do all serology tests have in common?
(Outcome 23-9)

Test Your Knowledge 23-12

True or False: The same disease may be detected by more than one serology test. (Outcome 23-9)

CLIA–WAIVED TESTS COMMONLY PERFORMED IN THE PHYSICIAN OFFICE LABORATORY

Tests performed in a physician office laboratory are very important for appropriate diagnosis and treatment of various diseases, as the results are available quickly within the same office where the patient is being examined. This section covers these tests, as well as their testing procedures, in more detail (Table 23-3).

- *Mononucleosis:* Mononucleosis is caused by the Epstein-Barr virus. The test performed for diagnosis identifies the presence of the heterophile antibody produced with this infection,.
- *Pregnancy:* Pregnancy tests determine the presence or absence of human chorionic gonadotropin (HCG) in urine or serum. CLIA-waived tests are qualitative in nature, with a positive or negative result.
- *Helicobacter pylori: H. pylori* is the bacterium responsible for many gastric and duodenal ulcers and gastritis. The test can be performed on serum and other body tissues.

TABLE 23-3

Commonly performed CLIA–waived immunology tests

Disease/Condition	Causative Agent	Test Performed	Required Specimen
Mononucleosis	Epstein-Barr virus	Mononucleosis screen or heterophile test	Blood
Pregnancy	Human chorionic gonadotropin (HCG)	Pregnancy (HCG) test	Serum or urine
H. pylori	*Helicobacter pylori* bacteria	*H. pylori* test	Blood or tissue sample
HIV	Human immunodeficiency virus	HIV test	Blood, oral fluids
Flu	Influenza virus	Influenza test	Nasal or nasopharyngeal swabs
Bladder cancer	Bladder tumor–associated antigen	BTA test	Urine
Strep throat	Group A streptococcus bacteria	Strep screen	Throat swab

- *HIV:* The human immunodeficiency virus screening test checks for the presence of the antibodies produced with HIV infection. This is designed to be a *screening* test only, and more definitive testing would then be performed on positive samples before the patient would be deemed HIV positive. Various body fluids can be tested for the presence of HIV.
- *Influenza:* The influenza test is performed during the winter months to detect the presence of influenza A or B, which causes a very serious respiratory infection.
- **Bladder tumor–associated antigen:** The screening test for bladder cancer uses a urine specimen.
- *Group A streptococcus infection:* The screening test for active infection with group A streptococcus uses a throat swab. The result is available in minutes after collection, allowing for much faster treatment than was previously available when only a throat culture was used.

Test Your Knowledge 23-13

Why might it be beneficial to perform simple serology tests in the physician office instead of a reference or hospital laboratory? (Outcome 23-10)

Test Your Knowledge 23-14

What is the causative agent of mononucleosis? (Outcome 23-11)

SUMMARY

Our immune systems are complex and protect us in various ways. We have three basic mechanisms that interact to provide this protection. They include various natural barriers, cellular and chemical responses to invaders, and antibody formation. We use vaccinations to protect us by activating our immune systems, and we also have naturally occurring antibodies to ABO blood types that are unlike our own. Blood type antibodies may be protective, but they can also cause serious health issues and must be considered during pregnancy to protect the unborn child.

Laboratory testing for immunity is known as serology. Various tests are performed in the physician office laboratory, as well as in hospital and reference laboratories. These tests may be qualitative or quantitative in nature, and are invaluable to aid in appropriate diagnosis of numerous disease states.

TIME TO REVIEW

1. A substance released in response to antigenic exposure that causes blood vessels to dilate is a(n): Outcome 23-1
   - **a.** Lymphocyte
   - **b.** Histamine
   - **c.** Interferon
   - **d.** Chemotaxis
2. A(n) ______________________ is anything that our body recognizes as nonself. Outcome 23-1
3. How are antigens and antibodies related to serology testing? Outcome 23-2
4. What role does our stomach play in protecting us from antigens? Outcome 23-3
   - **a.** Inflammation destroys the pathogens
   - **b.** The pH of the stomach acid helps to destroy microorganisms that may make their way into our bodies
   - **c.** Antibodies are produced here to destroy the invaders; white blood cells engulf the antigens
   - **d.** The stomach does not play a role in immunity
5. What is phagocytosis? Outcome 23-1
   - **a.** The process by which white blood cells engulf and destroy pathogens
   - **b.** The process that occurs when a noncompatible blood type is given to a patient
   - **c.** A term used to describe the first line of immunity
   - **d.** The production of antibodies
6. True or False: Immunity provided by the mother at birth is lifelong for many diseases. Outcome 23-5
7. True or False: Acquired immunity occurs only as a consequence of vaccinations. Outcome 23-6
8. What is used in vaccinations to activate our immune systems? Outcome 23-6
   - **a.** Immunoglobulin
   - **b.** Killed (attenuated) viruses or parts of viruses
   - **c.** Normal flora
   - **d.** B cells
9. Antibodies to blood type antigens A or B: Outcome 23-8
   - **a.** Occur after exposure to another blood type
   - **b.** Are naturally occurring and do not require exposure to develop

c. Do not exist in the bloodstream
d. Are harmless

10. True or False: Serology tests use a variety of testing methods. Outcome 23-9

11. Which of the following are acronyms for tests used to detect syphilis infection? Outcome 23-9

a. VDRL and RPR
b. HCG and IgG
c. HET and RPR
d. CRP and VDRL

12. Doctor Johnson is about to prescribe a medication to treat acne for Carly Lucas. This medication is not advised for those who are pregnant. Which CLIA-waived test might the physician order to be sure this drug is safe for Carly to take? Outcome 23-9

a. RPR
b. *H. pylori*
c. HCG
d. Mononucleosis

13. True or False: All serology tests performed in physician office laboratories require blood samples as the speciment type. Outcome 23-11

Case Study 23-1: Could this be right?

Dr. Johnson has just handed his medical assistant a requisition form for a blood draw for Mr. Daniels, a 22-year-old male. The physician has ordered a β HCG test. The medical assistant is surprised, because she has only seen HCG tests ordered for females in the past. She is sure there is an error, and asks Mr. Daniels to wait a few minutes until she can confirm the order with the physician.

1. Is this test an appropriate one for a 22-year-old male? If so, why might the physician have ordered the test?

RESOURCES AND SUGGESTED READINGS

“How Your Immune System Works”
Another explanation for how the immune system works http://www.health.howstuffworks.com/immune-system16.htm

“Overcoming Sensitivity Limitations of Lateral-Flow Immunoassays With a Novel Labeling Technique”
Information about various medical devices and testing methods http://www.devicelink.com/ivdt/archive/06/05/010.html

“Quidel Rapid Diagnostic Products”
Links to the Quidel point-of-care test kits, including CLIA–waived tests http://www.quidel.com/products/product_list.php?cat=1&by=brand&group=1

“Tests Waived by the FDA From January 2000 to Present”
List of all CLIA–waived tests; updated regularly http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfClia/testswaived.cfm

“What Does It Mean to Be Rh Negative?”
Further information about Rh-negative immunity issues and the use of RhoGam. This site has information for patients and health-care professionals. http://www.rhogam.com/Patient/WhatRhNegativeMeans/Pages/default.aspx

Chapter 24

Immunological Based Rapid Testing

Constance L. Lieseke, CMA (AAMA), MLT, PBT(ASCP)

CHAPTER OUTLINE

- Immunology Methods and Procedures
- Rapid Testing
  - Advantages of Rapid Testing
  - Common Procedural Elements of Rapid Testing
  - CLIA–Waived Regulations and Their Application to Immunology -Based Rapid Testing
- Common CLIA–Waived Rapid Tests and Their Clinical Significance
  - Group A Streptococcal Screening
  - Mononucleosis Testing
  - Pregnancy Testing
  - *Helicobacter pylori* Testing
  - Influenzas A and B
  - HIV Testing
- Summary
- Time to Review
- Case Study
- Resources and Suggested Readings

Learning Outcomes

After reading this chapter, the successful student will be able to:

24-1 Define the key terms.

24-2 Analyze the advantages of rapid immunology testing.

24-3 Restate the common aspects of various rapid testing procedures.

24-4 Describe potential follow-up testing procedures that may be necessary based on rapid testing results.

24-5 Differentiate the clinical significance of strep screens, mononucleosis testing, rapid urine pregnancy tests, *Helicobacter pylori* testing, BTA tests, influenzas A and B, and HIV testing.

24-7 Perform CLIA-waived rapid testing procedures for a strep screen, a urine pregnancy test, and a test for mononucleosis.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I.15. Perform immunology testing
III.P.III.2. Practice Standard Precautions
III.P.III.3. Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations
III.P.III.8. Perform CLIA-waived microbiology testing
I.C.I.6. Identify common pathology related to each body system
III.C.III.10. Identify disease processes that are indications for CLIA-waived tests

ABHES Standards

- Perform selected CLIA-waived tests that assist with diagnosis and treatment, #6: Kit Testing, a. Pregnancy, b. Quick Strep
- Dispose of Biohazard Materials

KEY TERMS

- Agglutination
- Chromatographic
- Colorimetric
- Epstein-Barr virus (EBV)
- *Helicobacter pylori (H. pylori)*
- Heterophile antibody
- Human chorionic gonadotropin (HCG)
- Immunology
- Lateral flow immunoassay
- Mononucleosis
- Qualitative
- Quantitative
- Quantitative HCG
- Rapid flow immunoassay
- Sensitivity
- Serology
- Specificity
- Streptococcal pharyngitis

Introduced in Chapter 23, **immunology** is defined as the study of the components of the immune system and their function. This includes numerous areas of specialty, such as autoimmune disorders, infectious disease, blood banking, and even tissue typing and organ transplantation. **Serology** describes the area of the laboratory where antigen-antibody reactions are studied. Serology implies that serum is used for the testing method, but immunology (or serology) testing performed today does not just use serum; urine, tissues, whole blood, and other body fluids may also be used. Immunology or immunoassay testing is performed to detect or identify antibodies or antigens in a specimen. The testing process may be **qualitative**, employing a positive or negative result, or **quantitative**, employing a numerical result, including varied levels of test complexity. Diseases such as mononucleosis, influenza, and hepatitis may be diagnosed using immunology tests.

Test Your Knowledge 24-1

How are immunology and serology similar?

(Outcome 24-1)

IMMUNOLOGY METHODS AND PROCEDURES

The majority of the common immunology tests are performed to detect the presence of a specific antigen or antibody in the specimen. The patient's specimen is mixed with the commercial solution that contains the antigen or antibody for that specific disease or condition. If the appropriate counterpart (antigen or antibody) is present in the specimen, a reaction will occur. In CLIA-waived immunology testing methods, a color change is evident when this reaction takes place. The CLIA-waived kits use an absorbent membrane that is treated with a chemical additive. Once the specimen and additional reagents are added to the membrane, a color change will occur, indicating whether the test is positive or negative. Testing methods that include a color change may be called **chromatographic** or **colorimetric** assays.

When choosing a certain test method, laboratories must consider whether the test method demonstrates **specificity** and **sensitivity** for the desired substance. The specificity of a certain method describes the ability to detect only the substance that is being tested for in the specimen. Testing methods that are not very specific will have an increased amount of false-positive test results because substances other than the one desired will also be detected by this method. The sensitivity of a testing method is related to the minimum concentration of the desired substance that will be detected. Ideally, laboratory testing methods will demonstrate a positive result when a very small amount of the substance is present in the specimen so that this presence is not overlooked.

Test Your Knowledge 24-2

If a test reacts positively with the presence of several different antigens, is this described as a lack of specificity or a lack of sensitivity? (Outcome 24-1)

POINT OF INTEREST 24-1

Bladder tumor–associated antigen

Bladder tumor antigens (BTA) are produced by bladder tumor cells and shed into the urine of those individuals with bladder cancer. There are two CLIA-waived rapid tests that are designed to detect the presence of this antigen in the urine. This test has not been approved as a screening test for bladder cancer because of its low specificity. The BTA test may be positive in the presence of other disorders of the renal system, such as renal stones, nephritis, renal cancer, urinary tract infections, or recent trauma to the bladder or urinary tract. The CLIA-waived rapid screening test is used for monitoring residual or recurring bladder cancer after treatment. Positive results are often followed up with urine cytology, cystoscopy, and examination of biopsy samples.

RAPID TESTING

Rapid testing describes immunology and other laboratory tests that have results available in a very short time. Usually a rapid test can be completed and read within 30 minutes. For rapid immunology testing, the most common rapid testing methods used are **rapid flow immunoassay** procedures and **agglutination** tests. Rapid flow immunoassay procedures are testing methods in which the specimen and any additional reagents must be absorbed and "flow" across the testing surface before they can interact with the antigen or antibody present to display a result. This type of testing may also be known as **lateral flow immunoassay** procedures. These testing methods are usually chromatographic, which means that their results are evident as a color change on the surface of the testing area.

When agglutination testing is performed, the sample and reagents are mixed on a flat surface designed to allow full view of the mixture. The antibodies and antigens present in the specimen will aggregate (agglutinate) if they both are present. Agglutination is visible as clumps on the testing surface.

Test Your Knowledge 24-3

Provide two similarities present in most rapid testing procedures. (Outcome 24-3)

Advantages of Rapid Testing

Many physician office laboratories perform rapid testing so that the health-care provider can obtain as much diagnostic information as possible in a short period of time. This can be especially advantageous when the testing is performed before the patient leaves the office. Testing in the physician office laboratory allows the health-care provider to prescribe medication or other treatment immediately without an additional office visit or phone call. Immediate results can save time and money for the patient, and often lead to better compliance with follow-up care.

Large laboratories may also choose to perform rapid testing rather than automated analysis for certain assays. In some situations, these rapid tests are the most cost-effective way to perform a specific assay. For tests that are not performed often, rapid testing kits may also be used to minimize potential waste of outdated reagents and controls. The CLIA-waived tests may be performed by employees with minimal training, allowing the laboratory professionals to perform tests of higher complexity. Rapid testing allows for the results to be available sooner than they might be if the tests were batched (tested together in large groups) and only performed at certain intervals as they often are in reference or hospital laboratories.

Test Your Knowledge 24-4

List two advantages of rapid testing procedures. (Outcome 24-2)

Common Procedural Elements of Rapid Testing

Rapid testing methods differ by manufacturer and by test. However, there are similarities that are present no matter what type of test or which manufacturer is involved. These include the following:

- Relatively few reagents
- Small sample volume
- Detailed package inserts with images to use for result interpretation
- Qualitative results interpreted as positive or negative
- Lateral flow immunoassay tests have internal validation indicators that change color to indicate whether the device is working and the sample and reagents were added appropriately
- Testing methods that are timed for completion and usually take 10 minutes or less to complete

- Kits that include all supplies needed for the test, with the exception of personal protective equipment
- Positive and negative controls included with each kit

CLIA-Waived Regulations and Their Application to Immunology-Based Rapid Testing

Many rapid tests are CLIA-waived procedures. The U.S. Food and Drug Administration (FDA) has classified these as simplistic assays with minimal steps and easy-to-interpret test results. However, those rapid methods that require additional specimen preparation or multiple steps in the testing process may be categorized as moderately complex by CLIA standards. For instance, agglutination tests are relatively simple to perform, but reading the results requires subjective interpretation that places agglutination tests in the CLIA moderately complex category. Some rapid lateral flow immunoassay tests may be CLIA waived if whole blood is used for the testing process, but regarded as moderately complex if plasma or serum is used because of the additional steps required to process the specimen for the procedure. If the laboratory where the testing is performed is authorized to perform only CLIA-waived procedures, it is important to verify whether the testing method is appropriate for the qualifications of the personnel performing the test in the facility before it is used. The vendor should be able to verify whether the test is CLIA waived or of higher complexity, and often the CLIA-waived tests are labeled as such in bold letters to avoid confusion.

The instructions provided by the manufacturer must be closely followed for all CLIA-waived testing procedures, as well as those of moderate complexity. These instructions will specify the temperature at which the testing kit should be stored; the conditions for appropriate specimen collection, preparation, and storage; and the frequency at which quality control samples should be tested, in addition to the step-by-step testing procedure instructions. The package insert will also provide necessary information to interpret the test results correctly.

COMMON CLIA-WAIVED RAPID TESTS AND THEIR CLINICAL SIGNIFICANCE

Common immunology-based CLIA-waived rapid tests include strep screens, pregnancy tests, tests for the infectious mononucleosis, and tests for *Helicobacter pylori.* CLIA-waived rapid tests for influenzas A and B and HIV are also available.

Group A Streptococcal Screening

Group A streptococcal infection is a common cause of **streptococcal pharyngitis,** more commonly known as strep throat. Common symptoms of group A strep throat infection include sore throat, fever, white or yellow deposits on the tonsillar area of the throat, swollen glands at the front of the neck, and an absence of cough. It is important to identify patients with streptococcal infections and treat them appropriately with antibiotics so that a more serious secondary disease does not develop. If a patient with pharyngitis has a negative group A strep screen but exhibits additional symptoms, the health-care provider may order a culture specimen to be set up to verify whether group A strep is present. Group A strep infection may cause inflammation of the joints, formation of scar tissue on the heart as a complication of endocarditis, or a dangerous kidney condition called glomerulonephritis.

Test Your Knowledge 24-5

What is a possible complication of untreated streptococcal infection? (Outcome 24-5)

A group A strep screen is an uncomplicated test that can be completed in less than 10 minutes. A sterile throat swab (provided in the kit) is used to collect a specimen from the tonsillar area and back of the throat. Following the instructions included in the package insert, single drops of the kit reagents are added to the swab, after which the reaction is timed for the appropriate interval, usually 5 minutes. The specimen/reagent mixture will travel across the reaction area of the kit, and a positive result will be evident within the time frame indicated in the manufacturer's insert. The result must be read at the specified time after the addition of the reagent to the sample to be valid.

Mononucleosis Testing

Mononucleosis is an infectious disease caused by the **Epstein-Barr virus,** primarily infecting children and adolescents. It is spread through saliva, and is also called the "kissing" disease. Infectious mononucleosis is

Procedure 24-1: Perform a CLIA-Waived Rapid Strep Screening Procedure

Rapid strep screens are the most common CLIA-waived test performed in physician offices. They may also be performed in hospital or reference laboratories. In addition to patient samples, rapid strep screens may also be used by the microbiology departments in some settings to verify the identity of the growth on a culture specimen.

TASK

Perform a rapid strep screen on an appropriate specimen taken from the throat of a patient.

CONDITIONS

- Hand sanitization equipment
- Gloves
- Rayon collection swab inoculated with throat specimen
- Laboratory coat
- CLIA-waived rapid strep kit
- External quality control materials (if required by manufacturer's recommendations or laboratory policy)
- Timer or watch
- Control log sheet
- Test log sheet
- Manufacturer's insert with reference material for reading results
- Biohazardous waste container

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I. Anatomy and Physiology, #15 Perform immunology testing **III.P.III.** Applied Microbiology/Infection Control, #2 Practice Standard Precautions, #3 Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations, #8 Perform CLIA-waived microbiology testing

ABHES Standards

- Perform selected CLIA-waived tests that assist with diagnosis and treatment, #6: Kit Testing, b. Quick Strep
- Dispose of Biohazard Materials

Procedure	Rationale
1. Verify the test ordered on the requisition form and the identification information on the specimen label.	The test and identification on the specimen label should be verified every time to avoid erroneous results.
2. Gather necessary supplies. Verify the expiration date on the kit and the required frequency for the quality control materials. Refer to the manufacturer's insert, if necessary, to verify the quality control (QC) frequency recommended for the kit.	Expired kits may not be used. The manufacturer's insert will provide minimum frequency for external quality control testing, as well as information about any internal quality control testing. Each laboratory may establish a unique schedule for QC testing; however, it may not be less frequent than that recommended by the manufacturer.
3. Wash hands (if they are visibly soiled) or apply hand sanitizer. Allow hands to dry completely.	Clean hands stop the spread of infection. Hands should be completely dry before attempting to apply gloves, or it will be difficult to put the gloves on the wet hands.

Continued

Procedure 24-1: Perform a CLIA-Waived Rapid Strep Screening Procedure—cont'd

Procedure	Rationale
4. If necessary, perform external quality control testing on the test kit. Log the results. Verify whether the kit is performing correctly before proceeding.	CLIA-waived kits usually come with external quality control testing materials. For those that do not include these materials, they may be purchased separately. Before patient testing may continue, it is important to verify that the positive and negative controls are testing in a valid range. If the controls are out of range, no patient testing can be performed until the problem has been identified and solved.
5. Perform the strep screen test as recommended by the manufacturer.	Test procedures will vary. Some kits use a test tube and dipstick, whereas others use a reagent well that is part of a test cassette. It is imperative that the instructions for the procedure are followed carefully. The number of drops and the timing of the different procedures are very important for accurate results.
6. Read the test result at the appropriate time. Test results are invalid after the time has elapsed, so they must be read immediately. Use the reference material included with the manufacturer's insert to interpret the results correctly. Record the results on the test log sheet and/or in a computer-based reporting system.	The result must be read at the appropriate time for the test to be valid. Reading results too soon or too long after the time interval has elapsed will result in erroneous results. The manufacturer's inserts will provide a chart to be used for reference when interpreting the color development changes for the kit.
7. Dispose of the throat swab and all testing materials in the biohazard disposal container.	All supplies should be disposed of properly for infection control reasons.
8. Remove gloves and wash hands.	Hands must always be washed after gloves are removed.
9. Record the results in the patient's chart, if available.	All results need to be charted immediately for the health-care provider to take appropriate action.

Date	
2/20/2014:	Rapid strep screen performed. Results negative
2:00 p.m.	Connie Lieseke, CMA(AAMA)

most common in those between the ages of 15 and 25. Fatigue, weakness, sore throat, headache, fever, and swollen lymph nodes are common symptoms of this infection. Sometimes the mononucleosis test may be ordered as a **heterophile antibody** test, because of the type of antibody produced by the body when infected with the Epstein-Barr virus. The antigen-antibody tests used to diagnose infectious mononucleosis use erythrocytes from sheep, horse, or cattle cells to interact with the heterophile antibodies.

The CLIA-waived mononucleosis tests demonstrate a positive result when the heterophile antibodies are present rather than reacting specifically with the Epstein-Barr virus. This means that false-positive results are not

POINT OF INTEREST 24-2
RSV

Respiratory syncytial virus (RSV) is a virus that affects the respiratory system. In most cases, an infected individual will recover from this virus within 2 weeks of infection. However, some populations are severely ill with RSV infection. Some infants, young children, and older adults (over the age of 60) may be at risk for serious infection. In the United States, RSV is the most common cause of the inflammation of the small airways in the lung (also known as viral bronchiolitis) and pneumonia in infants younger than 1 year of age. Recent studies have also linked an increase in cases of severe respiratory illness in the older population to RSV infection.

There are rapid tests available for identification of RSV infection in respiratory specimens. Viral cultures are also commonly performed. RSV is especially sensitive to changes in temperature in the laboratory, which means that it may be expressed as a false-negative result if the specimen is not handled correctly. Older children and adults may need a blood test performed, as they historically have low viral levels present in the respiratory specimens.

uncommon. Patients with lupus or lymphoma may also have heterophile antibodies present in their specimen, even though they do not have infectious mononucleosis. The CLIA-waived tests may result in a false-negative result for young children because they often do not create heterophile antibodies even when infected with the Epstein-Barr virus.

Because the symptoms of infectious mononucleosis are very similar to those of streptococcal pharyngitis, the health-care provider may order a throat swab collection for a group A strep screen to be collected at the same time that the sample is obtained for the mononucleosis test. Procedure 24-2 explains the testing process for a CLIA-waived mononucleosis test.

Pregnancy Testing

CLIA-waived tests for pregnancy are designed to react in the presence of **human chorionic gonadotropin (HCG),** a hormone that is found in urine or blood during pregnancy. The developing placenta secretes HCG, and most rapid test kits can detect the presence in the urine or blood just a few days after conception. CLIA-waived HCG tests provide a qualitative positive or negative result. Procedure 24-3 explains how to perform a CLIA-waived pregnancy test in more detail.

Procedure 24-2: CLIA–Waived Mononucleosis Testing Using an OSOM Mononucleosis Testing Kit

Infectious mononucleosis is diagnosed using laboratory data and clinical symptoms of sore throat, fever, and swollen lymph glands. A positive laboratory test indicates the presence of heterophile antibodies, which are present with the Epstein-Barr virus infection in most patients with acute infectious mononucleosis.

TASK

Perform mononucleosis testing using the CLIA-waived OSOM mononucleosis testing kit, manufactured by Genzyme.

CONDITIONS

- Hand sanitization supplies
- Gloves
- Completed patient requisition form and patient chart
- Appropriate venipuncture supplies if a venous blood sample is to be used. EDTA or heparinized blood samples are acceptable.
- Supplies included in the test kit
  - Test stick
  - Test tube
  - Diluent
  - Positive and negative control material
  - Workstation
  - Capillary tubes and bulb for capillary blood samples
- Timer or watch
- Biohazardous disposal container
- Biohazardous sharps container

Continued

Procedure 24-2: CLIA-Waived Mononucleosis Testing Using an OSOM Mononucleosis Testing Kit—cont'd

Note: This test may be performed using a capillary blood sample or a sample obtained through venipuncture. Be certain to have the necessary supplies available for the method used for sample collection, as well as a biohazardous sharps disposal container and gloves. Either EDTA or heparin is an appropriate additive for venous samples for the OSOM mononucleosis test.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I. Anatomy and Physiology, #15 Perform immunology testing **III.P.III.** Applied Microbiology/Infection Control, #2 Practice Standard Precautions, #3 Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations

ABHES Standards

- Perform selected CLIA-waived tests that assist with diagnosis and treatment, #4 Immunology, #6 Kit Testing
- Dispose of Biohazardous Materials

Procedure	Rationale
1. Verify test ordered on requisition. Greet and then identify patient.	The requisition form should always be examined to be certain which test is ordered, and the patient identity must always be verified before proceeding.
2. Sanitize hands, allow them to dry, and apply gloves.	Gloves protect the hands from exposure to potential bloodborne pathogens. Gloves should always be applied in the presence of the patient.
3. Gather necessary supplies.	Supplies must be within reach before staring the process so that the timing of the testing is accurate. Courtesy of Genzyme Corporation.

Procedure	Rationale
4. Verify whether quality control (QC) testing needs to be performed on the kit. If so, perform QC testing and verify validity of results before proceeding with patient testing.	The frequency of the QC testing is based on the manufacturer's and individual laboratory recommendations. Minimum recommendations from the Genzyme Corporation are that the positive and negative QC should be run with each new lot number and each new untrained operator. If the quality control results are not within the established range, patient testing cannot be performed until the problem has been solved.
5. Determine the method to be used for specimen collection. Collect the sample using appropriate techniques.	Capillary or venous whole blood may be used for this testing procedure.
a. If a capillary sample is to be tested, use the capillary tube and bulb supplied with the test kit. Follow appropriate procedures to fill the provided capillary tube with blood.	The capillary sample is the most common method of testing used in the physician office laboratory.
b. Blood obtained through venipuncture is acceptable, as long as it is collected into a tube containing EDTA or heparin.	Always follow appropriate venipuncture collection procedures.
6. Add the sample to the test tube.	For capillary samples, use the bulb provided to add the entire contents of the capillary tube to the test tube. If a venous sample is to be tested, add one drop to the test tube, using the transfer pipettes included in the test kit. Samples must be at room temperature at the time of testing.
7. Add one drop of diluent to the bottom of the test tube. Mix the diluent with the sample by gently moving the tube from side to side.	Slowly add the diluent so that it goes to the bottom of the test tube.
8. Remove a test stick from the container, and immediately recap the container.	The test sticks must not be exposed to moisture. The container must not be left uncapped for an extended period of time.
9. Place the absorbent (testing) end of the test strip into the sample/diluent mixture in the bottom of the tube and leave the strip in place.	One end of the test strip is absorbent, and the other end is designed to be handled during the testing procedure.
10. Set a timer (or note the time on a watch) for 5 minutes.	The test result is to be read at 5 minutes. Positive results may be read earlier, as long as the red control line is present.

Continued

Procedure 24-2: CLIA-Waived Mononucleosis Testing Using an OSOM Mononucleosis Testing Kit—cont'd

Procedure	Rationale
11. At 5 minutes, observe the test stick for the test results. **a.** If a blue line and the red control line are present, the test result is interpreted as a positive. **b.** If the red control line is present, but no blue line is evident, the test is interpreted as a negative result.	It is important to pay close attention to the test stick to interpret the results appropriately. The red control line must always be visible for the test to be valid.
c. If there is no red control line or if the background in the testing area is not clear (white), the test is considered to be invalid, and must be repeated.	Results are not to be reported as invalid. This is an indication that the test must be repeated with careful attention to the procedure. Quality control material should also be used to verify the performance of the test kit.
12. Make note of the test result.	It is not appropriate to chart the result while still wearing gloves that may be potentially contaminated.
13. Discard the pipette or capillary tube, testing strip, and test tube in the biohazardous disposal container.	Because blood was used in the testing procedure, all contaminated material must be disposed of as biohazardous substances.
14. Sanitize the work area and put away supplies.	The testing kit may be stored at room temperature. Follow laboratory protocol for storage of any unused specimens.
15. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
16. Document the test results in the patient's chart.	All results must be documented in the patient's chart.

Date	
04/18/2013:	Mono test performed from capillary puncture. Results negative ~~~~~~~~~~~~~~~~
11:33 a.m.	~~~~~~~~~~~~~~~~~~~~~~~~~~~ Connie Lieseke, CMA (AAMA)

In some situations, the HCG level in the blood or urine is not elevated enough to be detected by the rapid test procedure. This may be the case very early in pregnancy when the placenta has not had an opportunity to secrete enough of the hormone for the test to be positive. When a woman has an ectopic pregnancy, the hormone levels may never be elevated enough to produce a positive rapid test. In this situation, the healthcare provider may order a **quantitative HCG test,** which provides a number representing the level of hormone present in the blood. If the quantitative HCG result does not increase over time, the pregnancy is usually not viable. In a normal pregnancy, the HCG levels should continue to increase until approximately 3 months of gestation, at which time they will start to decline through the rest of the pregnancy.

Helicobacter pylori Testing

In 1982, it was discovered that most duodenal ulcers and many gastric ulcers were caused by a previously unknown bacteria, ***Helicobacter pylori***. Until then, it was believed that ulcers were caused by stress, lifestyle, and food choices. Medications that reduced the production of stomach acid were often used to heal these ulcers, but recurrence was common once the treatment ended. In 1982, pathologist Robin Warren discovered that at least 50% of the biopsies examined from ulcer patients exhibited small, curved bacteria in the area of inflammation. Together with his partner, Barry Marshall, Warren provided evidence that these ulcers were caused by the bacteria he had observed and that the ulcers could be eradicated (without recurrence) with antibiotics and medication to reduce the acidity of the stomach. The bacteria cause inflammation and weaken the stomach lining, allowing stomach acids and bacteria to penetrate the layers and cause a hole or ulcer.

There are now numerous CLIA-waived rapid tests available to detect the presence of *H. pylori*. This bacterium is unique, as it has been isolated only from humans. It is thought that the transmission usually

Procedure 24-3: CLIA–Waived Urine HCG Testing Using the Beckman Coulter Icon 25 Testing Kit

Human chorionic gonadotropin (HCG) is secreted by the placenta shortly after implantation of a fertilized ovum. It may be detected in the urine specimen 7 to 10 days after conception, which may be as early as the first day of a missed menstrual period.

TASK

Perform urine HCG testing using the Beckman Coulter Icon 25 testing kit, and record the results appropriately. Complete all steps within 10 minutes.

CONDITIONS

- Hand sanitization supplies
- Gloves
- Completed patient requisition form and patient chart
- Supplies contained in test kit
- Test devices
- Disposable transfer pipettes
- Commercial quality control material
- Timer or watch
- Urine specimen collected in clean container
- Biohazardous disposal container
- Biohazardous sharps container

Note: The HCG test kit may be used for urine testing or serum testing. The test is only CLIA–waived when performed using urine samples.

CAAHEP/ABHES STANDARDS

CAAHEP Standards

I.P.I. Anatomy and Physiology, #15 Perform immunology testing **III.P.III.** Applied Microbiology/Infection Control, #2 Practice Standard Precautions I#3 Select appropriate barrier/personal protective equipment (PPE) for potentially infectious situations

ABHES Standards

- Perform selected CLIA-waived tests that assist with diagnosis and treatment, #6 Kit Testing, a. Pregnancy
- Dispose of Biohazardous Materials

Procedure	Rationale
1. Verify test ordered on requisition. Greet and then identify patient if present for collection.	The requisition form should always be examined to be certain which test is ordered, and the patient identity must always be verified before proceeding.
2. Sanitize hands, allow them to dry, and apply gloves.	Gloves protect the hands from exposure to potential bloodborne pathogens. Gloves should always be applied in the presence of the patient.

Continued

Procedure 24-3: CLIA–Waived Urine HCG Testing Using the Beckman Coulter Icon 25 Testing Kit—cont'd

Procedure	Rationale
3. Gather necessary supplies. Supplies and samples must be at room temperature for valid test results.	Supplies must be within reach before staring the process so that the timing of the testing is accurate. ICON 25 hCG Product No. 43025 25 Tests BECKMAN COULTER Courtesy of Beckman Coulter.
4. Verify whether quality control (QC) testing needs to be performed. If so, perform QC testing and verify validity of results before proceeding with patient testing.	The frequency of the QC testing is based on the manufacturer's and individual laboratory recommendations. The package insert for this testing procedure does not specify a minimum frequency for quality control testing.
5. Examine the urine specimen. If it appears cloudy, centrifuge the specimen before testing.	Excessive particulate matter needs to be removed from the specimen through centrifugation to ensure valid test results. The specimen should be collected in a clean container.
6. Remove the test device from the wrapper just prior to use.	The test device should not be exposed to room air (and potential exposure to moisture) for an extended period before use, as this may lead to an invalid test result.
7. Place the test device on a flat stable surface. Place three drops of urine in the sample well (marked with an S) using the supplied transfer pipette.	Avoid formation of air bubbles in the specimen during application.
8. Set the timer (or note the time on a watch) to 3 minutes.	The sample results are not valid before 3 minutes have elapsed.

Procedure	Rationale
9. At 3 minutes, observe the reaction area for the presence of red lines.	The test results need to be read precisely at 3 minutes. A delay in reading the results may allow for slight color development to occur, which may be erroneously interpreted as a positive result. 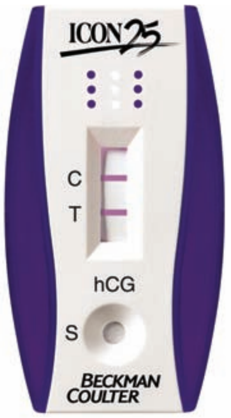 Courtesy of Beckman Coulter.
a. A red line in the Control area in addition to the Test area is to be interpreted as a positive result.	To be a valid test, there must always be a red line in the Control area at three minutes.
b. A red line in the Control area without the presence of a red line in the Test area is to be interpreted as a negative result.	Look carefully at the Test area before reporting a result.
c. If there is no red line in the Control area, the result is not valid, and the test must be repeated.	Invalid tests are not to be reported as such; this is an indication that the test must be repeated with particular attention to procedural details. Quality control material should also be used to verify the performance of the test kit.
10. Make note of the test result.	It is not yet appropriate to chart the result, as the operator is still wearing gloves that may be potentially contaminated.
11. Discard the testing device and transfer pipette into the biohazardous trash.	The testing supplies were exposed to urine.
12. Put away the test kit, and store the specimen according to laboratory protocol.	The test kit may be stored at room temperature.
13. Sanitize the work area.	The work area must always be sanitized before proceeding to other tasks.
14. Remove gloves and sanitize hands.	Hands must always be sanitized after removing gloves.
15. Document the test results in the patient's chart.	All results must be documented in the patient's chart.

Date	
04/23/2013:	Urine HCG test performed. Results positive. Physician notified.
1:15 p.m.	Connie Lieseke, CMA (AAMA)

occurs early in life and that the microorganism is probably passed directly from one person to another. The most common specimen used for *H. pylori* testing is whole blood, but there are also CLIA-waived kits available for testing gastric biopsy specimens. Serum, plasma, and stool specimens may also be tested.

Test Your Knowledge 24-6

True or False: The presence of *Helicobacter pylori* in the stomach is linked to an infection of the lungs with the same bacteria. (Outcome 24-5)

Influenzas A and B

The Centers for Disease Control and Prevention (CDC) estimates that every year 5% to 20% of the U.S. population comes down with the seasonal flu. More than 200,000 people are hospitalized with complications of this infection, and it leads to at least 36,000 deaths each year. Influenza flu viruses cause the flu. There are three types of human flu viruses, but only A and B are of clinical significance. Symptoms of the flu include fever, body aches, dry cough, headache, and sore throat. The duration of the symptoms may vary, but for many individuals the recovery time is a week or more. Human influenza C causes mild upper respiratory symptoms and is not vaccinated against. Yearly vaccination is recommended for the majority of the population against influenzas A and B.

Laboratory testing for identification and treatment of the flu virus is not always necessary. Many times the health-care professional will diagnose the patient with seasonal flu based on a clinical examination, and prescribe rest and symptomatic treatment. However, in situations in which the patient has other complicating factors (such as asthma, heart disease, or low immunity) it may be necessary to verify whether the patient is infected with influenza A or influenza B so that antiviral medications may be administered. Laboratory testing may also be indicated if the patient is living in a controlled, confined environment (such as a long-term care facility) so that the spread of the virus can be controlled with isolation of the patient and administration of antiviral medication.

Influenza cultures will provide detailed, specific information about the virus if it is present, but a culture result will usually take 3 to 10 days to be finalized, during which time the patient may be spreading the virus to others. Rapid CLIA-waived tests are a better option for the initial influenza screening, as the results are usually available within 15 minutes after collection. Examples of CLIA-waived tests include Quick Vue and BinaxNOW.

Many times the symptoms of the seasonal flu may mimic those of bacterial infections. Identification of the flu virus through rapid testing methods will also eliminate the use of unneeded antibiotics when bacterial infection is not present. All CLIA-waived testing methods, as well as the rapid tests that have been categorized as moderately complex use nasopharyngeal swabs for influenzas A and B. Nasal swabs, nasal rinses, or nasal aspirates may also be acceptable for some testing methods. Ideally, the specimen will be collected 3 to 4 days after infection is suspected. A viral culture may be indicated when an unexpected negative result is encountered, especially in situations in which the prevalence of the virus in the community is high. Blood testing may also be performed, using a method called reverse transcription polymerase chain reaction. This method is very sensitive and specific to the virus, and may allow for differentiation of variant strains present in the population.

Test Your Knowledge 24-7

When may a viral culture be indicated after a rapid influenza test is performed? (Outcome 24-4)

HIV Testing

Approximately one individual out of every five infected with HIV is unaware of his or her positive status, according to the CDC. It is also estimated that there are approximately 40,000 new individuals infected by HIV annually in the United States. Historically, many of the individuals tested did not return to the health-care provider to receive their HIV results, so they were not treated appropriately, and continued to potentially spread the disease.

In 2004, the FDA approved several rapid testing kits to be used for HIV testing. This development was significant because it facilitated access to early diagnosis and treatment. The patient is now able to visit a health-care provider for pretest counseling and have the test run while waiting for the results. Before leaving the office, the patient will be provided with a result (although positive results are considered to be preliminary). The patient is also given information about appropriate interpretation of their HIV status and instructions before they leave. Only patients with a preliminary positive HIV screening test are required to return for their confirmatory results.

Test Your Knowledge 24-8

What is an advantage of rapid HIV testing?

(Outcome 24-4)

There are currently four CLIA-waived test procedures available for HIV testing:

1. OraQuick ADVANCED HIV-1/2 Antibody Test, by Orasure Technologies
2. Uni-Gold Recombigen HIV Test, by Trinity BioTech
3. Clearview HIV 1/2 Stat-Pak, by Inverness Professional Diagnostics
4. Clearview Complete HIV 1/2, also by Inverness Professional Diagnostics

All four are CLIA waived for whole blood testing from capillary or venous samples. Several may also be moderately complex if serum or plasma is used for the test procedure. The OraQuick and Clearview methods are capable of detecting two strains of HIV. HIV-1 is the most common strain worldwide, but an additional strain (HIV-2) has been identified, especially in West Africa. It is important to remember that any presumptive positive HIV rapid test result must be confirmed with a western blotting test or a immunofluorescent assay before the patient is considered HIV positive.

The newest CLIA-waived testing procedure for HIV (OraQuick) uses a sample of oral mucosal transudate rather than blood. Oral mucosal transudate is the fluid from the gums and cheek tissue. This method can benefit those patients who are reluctant to have their blood drawn for the test. The sample is obtained by handing the patient a special collection "paddle" and having him or her place it for a few seconds between the cheek and gum, after which it is rubbed over the gum area in the mouth. The end of the paddle that contains the specimen is then placed in a tube that contains the reagent for the test. If HIV antibodies are present in the sample, the reaction indicator on the paddle will demonstrate a positive result after 20 minutes. If the CLIA-waived rapid test is positive, another oral fluid specimen or a blood specimen must be collected for the confirmatory test. Negative test results are interpreted as a negative result for current HIV infection.

Although some of the processes involved in the rapid HIV testing procedures may vary, because a positive HIV test result is clinically significant, all rapid HIV testing procedures include specific federal requirements:

1. All patients must receive an information sheet (included with the testing kit) that explains HIV/AIDS, the testing process, and what the results mean, and provides information about the confirmation of positive results. Patients also sign a consent form before testing.
2. Quality control (QC) testing (positive and negative control samples) must be performed as directed. This includes QC testing whenever a new operator uses the test kit (prior to testing patient samples), whenever a new lot goes into use, and whenever a new shipment is received (even if it is the same lot number). Quality control testing must also be performed if the storage temperature drops below 35°F or rises above 80°F, or if the testing environment drops below 59°F or rises above 99°F. Each laboratory will also specify additional testing intervals for quality control.

SUMMARY

Immunologic-based rapid testing methods are commonly used as diagnostic tools for various diseases. The tests will identify the presence of antigens or antibodies present in whole blood or other body fluids. Many of the rapid testing procedures are CLIA-waived lateral flow immunoassays that indicate a positive result through a color change on the result area of the kit. Some of the rapid tests are used for confirmative diagnostic information so that the patient can be treated appropriately as soon as possible, whereas others may be used as an initial screening test or to follow up after disease treatment. Additional testing procedures are sometimes warranted once the rapid testing procedures are complete. The manufacturer's insert for the CLIA-waived rapid tests must always be followed closely to avoid erroneous results or invalid tests.

TIME TO REVIEW

1. If a test result is reported as positive or negative, the test is ________________. Outcome 24-1
   a. Qualitative
   b. Quantitative

2. How do rapid testing methods save money for patients? Outcome 24-2
   - a. Fewer office visits
   - b. Inexpensive testing methods
   - c. Less expensive antibiotics
   - d. a and b

3. If a rapid test for HIV is positive, what is the required follow-up testing procedure? Outcome 24-4
   - a. HIV culture
   - b. AIDS test
   - c. Immunofluorescent assay or western blot test
   - d. Neutrophil blood count

4. True or False: A negative urine HCG test always means that the patient is not pregnant. Outcome 24-5

5. When performing a group A strep screen, where should the sample be collected from the patient? Outcome 24-6
   - a. Tonsillar area and the back of the throat
   - b. Tonsillar area and the tongue
   - c. Gum line
   - d. Nasopharynx

6. How is an infection of *Helicobacter pylori* treated? Outcome 24-5
   - a. Surgical correction
   - b. Antiseptic suppositories
   - c. Antibiotics and acid reducing medications
   - d. None of the above

7. Which of these may be an order for a mononucleosis test? Outcome 24-5
   - a. BTA
   - b. Influenza A
   - c. Heterophile
   - d. GAS

Case Study 24-1: CLIA-waived testing procedures

James works as a "float" medical assistant at Briarcreek Medical Center. Today he has been assigned to work in the laboratory area at the busy obstetric/gynecology office of Dr. Stanton. The first patient of the day has a requisition with an FSH and a LH ordered, which both require serum for processing. James draws two tiger top tubes, thanks the patient, and cleans up his work area. Right after the patient leaves, Dr. Stanton tells James that he would also like a STAT pregnancy test to be performed on this patient.

1. Is it possible for James to perform a STAT HCG test on the specimen that he has in the laboratory?
2. What may keep James from performing the test?

RESOURCES AND SUGGESTED READINGS

"Tests Granted Waived Status under CLIA" https://www.cms.gov/CLIA/downloads/waivetbl.pdf

"Helicobacter pylori Fact Sheet for Health Care Providers" Information about *Helicobacter pylori* infection and treatment http://www.cdc.gov/ulcer/files/hpfacts.PDF

"Getting Tested" San Francisco AIDS Foundation, "HIV Testing" http://www.sfaf.org/aids101/hiv_testing.html#urine

"Evaluation of Rapid Influenza Diagnostic Tests for Detection of Novel Influenza A (H1N1) Virus—United States, 2009" CDC Morbidity and Mortality Weekly Report, August 8, 2009 http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5830a2.htm

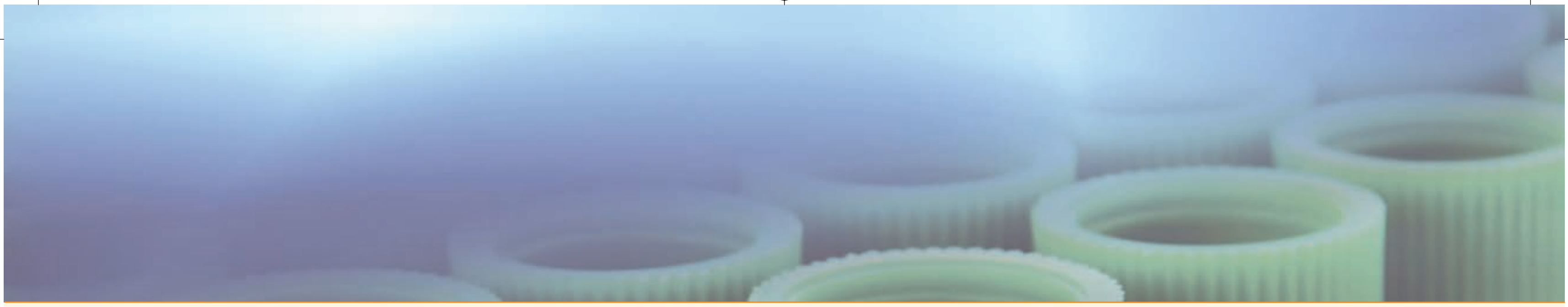

Section VI

Immunology

What Does It All Mean?

As this section clearly points out, immunology is an important aspect of laboratory testing that spans all of the other laboratory disciplines. Immunologic connections are often associated with advances, new discoveries, and findings in the field of medicine. Perhaps the most important concept in immunology testing is antigen-antibody complexes, that is, when antibodies targeted against specific antigens find them and bind together. Visualization of these complexes, as seen in the laboratory, is known as agglutination.

Case in Point

It was noted in the case of little P.J. that his strep screen test was negative, meanng that no agglutination was seen. Most strep screen tests are designed to detect group A streptococci antigens when present. A negative test indicates that the antigens being sought were not present. Because throat infections of this type may be caused by other bacterial or viral agents, the best step to take at this juncture is to submit a new, fresh specimen to the laboratory for microbial culture. Note that, in this case, if the strep screen was submitted to a clinical laboratory, it would likely be performed either in the microbiology area, because the test is looking for a microbiological agent or in a section dedicated to immunologic testing. The location of such tests being performed depends on the organization of the laboratory. Rapid tests such as the strep screen have revolutionized laboratory testing in terms of time and resources. In the case of strep screen results, what used to take 48 hours or more to complete is now available in about 10 minutes. Furthermore, because some of the rapid tests have been deemed CLIA-waived , trained medical assistants and other primary health-care providers can perform these tests in a physician office setting.

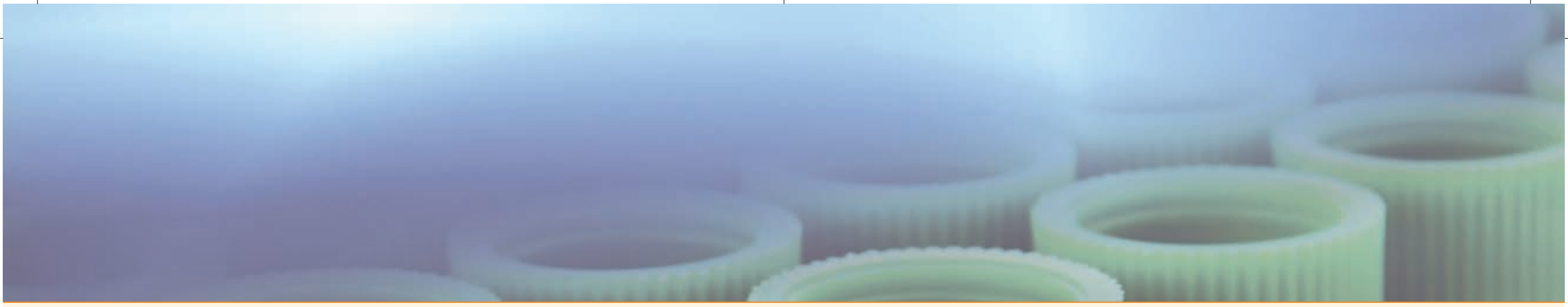

Appendix A

Reference Ranges

Reference Ranges for Glucose Testing Results

Fasting plasma glucose	70–100 mg/dL
2-hr PP glucose	Less than 140 mg/dL
Glycosylated hemoglobin (HbA1c)	
Nondiabetic	5%
Diabetes, well controlled	2.5%–6%
Diabetes, well controlled	>8%

From the American Diabetes Association.

Reference Ranges of Common Chemical Analytes (listed alphabetically)

Alanine aminotransferase (ALT)	10–35 U/L
Albumin (Alb)	3.5–5 g/dL
Alkaline phosphatase (ALP)	42–136 U/L
Amylase	30–70 U/L
Aspartate aminotransferase (AST)	0–35 U/L
Bilirubin, total (TBili)	0.3–1 mg/dL
Blood urea nitrogen (BUN)	10–20 mg/dL
Brain natriuretic peptide (BNP)	0–100 ng/L
Calcium (Ca)	8.2–10.5 mg/dL
Carbon dioxide (CO_2)	22–30 mEq/L
Chloride (Cl)	96–106 mEq/L
Cholesterol, total (Chol)	Less than 200 mg/dL
Creatine kinase (CK), aka Creatine Phosphokinase (CPK)	55–170 U/L
Creatinine (Creat)	0.6–1.2 mg/dL
Erythropoietin	5–35 IU/L
Ferritin, adult	10–150 ng/mL
High-density lipoprotein (HDL)	Greater than 50 mg/dL
Iron, serum adult	35–165 mcg/dL
Lactate dehydrogenase (LD, LDH)	100–190 U/L
Low-density lipoprotein (LDL)	Less than 100 mg/dL
Magnesium	1.6–2.2 mg/dL
Myoglobin	Less than 90 μg/L
Potassium (K)	3.5–5.0 mEq/L
Phosphorus, adult	2.5–4.5 mEq/dL
Phosphorus, child	4.5–6.5 mg/dL
Sodium (Na)	136–145 mEq/L
Thyroid-stimulating hormone (TSH)	0.4–4.2 μU/mL
Thyroxine (T4)	4.5–11.2 μg/dL
Triglyceride (Trig)	Less than 150 mg/dL
Triiodothyronine (T3)	75–220 ng/dL

Reference Ranges for Urine Tests

Appearance and color	Clear; straw to yellow in color
pH	4.6–8.0
Protein	2–8 mg/dL (negative to trace)
Specific gravity	1.005–1.030
Leukocyte esterase	Negative
Nitrite	Negative
Glucose	Negative
Ketones	Negative
Leukocytes	Negative
Blood	Negative
Urobilinogen	0.1–1.0 mg/dL
White blood cells	3–4/hpf
White blood cell casts	Negative
Red blood cells	1–2/hpf
Red blood cell casts	Negative
Crystals	Few/negative

Reference Ranges for Hematology	
Red blood cell, male	4.6–6.2 × 10^6 cells/µL
Red blood cell, female	4.2–5.9 × 10^6 cells/µL
Hemoglobin, male	13–18 g/dL
Hemoglobin, female	12–16 g/dL
Hematocrit, male	45%–52%
Hematocrit, female	37%–48%
White blood cell count	4,300–10,800 cells/mm^3
Neutrophils	54%–65%
Lymphocytes	25%–40%
Monocytes	2%–8%
Eosinophils	1%–4%
Basophils	0%–1%
Platelets	150,000–450,000/mm^3
Reticulocyte count, adult	0.5%–1.5% of all RBCs
Reticulocyte count, child	0.5%–2.0% of all RBCs
Erythrocyte sedimentation rate	
Adults: males younger than 50 years	<15 mm/hr
Adults: males older than 50 years	<20 mm/hr
Adults: females younger than 50 years	<20 mm/hr
Adults: females older than 50 years	<30 mm/hr
Children: 6 months–12 years	3–13 mm/hr

Reference Ranges for Coagulation	
Fibrinogen	150–400 mg/dL
D-dimers	<250 ng/mL
Prothrombin time	11–13 sec
Activated partial thromboplastin time	25–31 sec
Bleeding time (Ivy method)	1–9 min

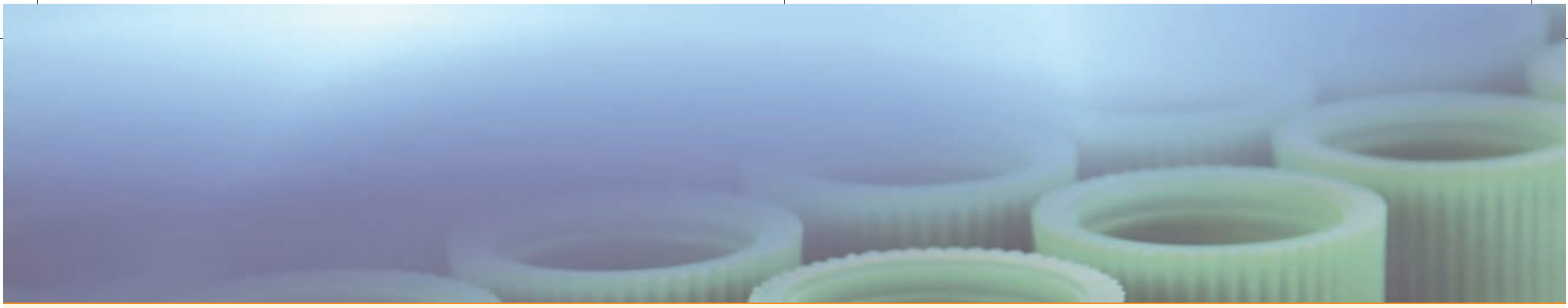

Appendix B
Test Your Knowledge Answers

Chapter 1

1-1. Reference laboratories have the most tests available. (Outcome 1-2)

1-2. The microbiology department would test the patient specimen to identify the microorganisms present in the wound. (Outcome 1-3)

1-3. **b.** The test is being performed to see if the patient is developing diabetes so that the disease can be treated early, or prevented if possible. (Outcome 1-4)

1-4. **False.** Medical assistants may work in specimen processing or as laboratory assistants performing a variety of duties in microbiology, histology, and the like. They may also work as phlebotomists if that laboratory provides this service. Medical assistants may also perform CLIA-waived or moderate-complexity tests with appropriate training. (Outcome 1-5)

1-5. Yes, as long as the training is appropriate and they have been designated as CLIA-waived or moderate-complexity tests. (Outcome 1-5)

1-6. There are various correct answers to this question. One reason is so that the results can be delivered or transmitted to the correct health-care provider upon completion. Another reason may involve reimbursement. Insurance companies require the name of the ordering practitioner for successful payment. (Outcome 1-6)

1-7. The ABN allows patients who have Medicare as their primary insurance to make an informed decision about whether they want to have laboratory tests performed that may not be reimbursed. The form offers an opportunity for the patient to refuse the test, or to accept the fee and have the test performed anyway. (Outcome 1-7)

1-8. To verify specimen requirements, tube stopper colors to be drawn, or any special handling procedures necessary for the test ordered. (Outcome 1-8)

1-9. They both contain important patient information, as well as the name of the ordering practitioner. They are both critical parts of the process, and both must be used for quality patient care. (Outcome 1-9)

1-10. The three phases of laboratory testing include the preanalytical phase, the analytical phase, and the postanalytical phase. (Outcome 1-10)

1-11. The preanalytical phase. This error occurred prior to the performance or reporting of the test results, so it is included with the preanalytical phase of testing. (Outcome 1-11)

Chapter 2

2-1. Yes, a physician assistant may serve as a laboratory director in a physician office laboratory. (Outcome 2-2)

2-2. Yes, this is a role often assumed by a medical technologist. (Outcome 2-3)

2-3. A phlebotomist (in most states) is not required to have specific credentials gained by a national examination or formal education. A phlebotomist is required to have a high school education (or the equivalent) and documented training in phlebotomy. (Outcome 2-4)

2-4. No, they are not limited to drawing blood. They can perform a variety of duties, as long as there is documented training for these expanded duties. (Outcome 2-5)

2-5. CLIA '88 was designed to ensure reliable, accurate, and timely laboratory results. (Outcome 2-6)

2-6. The minimum level of education is as follows:
- **a.** High school graduate (or the equivalent)
- **b.** Same, but additional documented training (on the job is sufficient)
- **c.** At least an associate degree in a laboratory-based curriculum. (Outcome 2-7)

2-7. Assignment of complexity levels takes into account the difficulty of the actual test performance (how many steps involved, etc.) as well as the clinical significance of the test results. (If the test was reported incorrectly, what impact would it have on the patient and his or her treatment?) (Outcome 2-7)

2-8. A clinical laboratory scientist with four years of education could perform any level of testing as designated by CLIA. (Outcome 2-8)

2-9. The U.S. Food and Drug Administration assigns the categories to the laboratory tests. (Outcome 2-9)

Chapter 3

3-1. Viruses, fungi, and parasites are examples of other types of microorganisms. (Outcome 3-2)
3-2. No. (Outcome 3-3)
3-3. They appear different under the microscope. Staphylococci appear in clusters, and diplococci appear in pairs. (Outcome 3-4)
3-4. Bacilli appear as ovals or rod-shaped organisms. (Outcome 3-4)
3-5. Viruses have a different cellular makeup than bacteria; bacteria have cell walls and membranes, and contain various structures that allow them to survive and reproduce. Viruses have genetic material contained in a protein capsule, and depend on the cells of other living organisms to reproduce. Viruses are much smaller than bacteria. Antibiotics work to destroy bacteria, but not on viruses. (Outcome 3-5)
3-6. No, it is just medically aseptic, with most of the bacteria destroyed, with a special emphasis on pathogens. (Outcome 3-6)
3-7. By washing their hands. (Outcome 3-7)
3-8. She has broken the chain at the point of a susceptible host; she is no longer able to be affected by hepatitis A. (Outcome 3-7)
3-9. Everyone is considered infectious. (Outcome 3-8)
3-10. Hand washing. (Outcome 3-8)
3-11. Alcohol-based hand rubs. (Outcome 3-9)
3-12. Before eating or drinking, after removing gloves, between patients, when visibly contaminated. (Outcome 3-10)
3-13. So that the health-care worker doesn't contaminate his or her hands again by touching the dirty handles. (Outcome 3-10)
3-14. She touched the phone with her gloved hand; her gloves may be contaminated with pathogens and should be removed before she touches the phone or puts her hand near her face. (Outcome 3-9)
3-15. A laboratory coat and gloves. (Outcome 3-9)
3-16. Employees. (Outcome 3-11)
3-17. The laboratory assistant may look at the label on the container; it is required to contain the manufacturer's information. (Outcome 3-13)
3-18. There are nine required components. (Outcome 3-12)
3-19. The NFPA labels provide information about flammability, reactivity, health hazards, and other specific hazards for the chemical. They are color coded and have numeric symbols to indicate the severity of the hazard. (Outcome 3-14)
3-20. The letters are assigned to the fire extinguishers to indicate what type of fire they should be used to extinguish. (Outcome 3-14)
3-21. Other potentially infectious materials include semen, vaginal secretions, cerebrospinal fluid, pericardial fluid, pleural fluid, etc. (Outcome 3-16)
3-22. **False.** It was designed to protect health-care employees. (Outcome 3-15)
3-23. Determination of the employee risk of exposure, education, and implementation of appropriate use of personal protective equipment and other means to reduce the potential exposure of the employees; hepatitis B vaccination; postexposure evaluation procedures; and ongoing communication of hazards. (Outcome 3-17)
3-24. No; this is why it is so important to find out how to use the one that is in place at your worksite. (Outcome 3-18)
3-25. No; much of the waste generated in a laboratory is routine office waste, such as paper. (Outcome 3-19)
3-26. No. Hepatitis, or inflammation of the liver, can be caused by a variety of disease states. Viral hepatitis still has several forms; hepatitis A in particular is not a bloodborne pathogen. (Outcome 3-20)
3-27. They are both bloodborne pathogens, they are both caused by viruses, and they both have long-lasting effects on the body that may be or are fatal. (Outcome 3-20)
3-28. She should immediately rinse them out with water. (Outcome 3-21)
3-29. **True.** (Outcome 3-21)

Chapter 4

4-1. Both include documentation, both affect the quality of the test results, both involve employees, both involve written procedures, and both may involve external quality controls. (Outcome 4-2)
4-2. Incorrect laboratory results may result in improper treatment for the patient. Patients may not receive the necessary treatment, or they may be treated incorrectly based on the inaccurate results, leading to serious consequences. (Outcome 4-3)
4-3. No, they do not. They may provide information about whether the instrument is operating correctly, but they do not let you know whether a test is valid. (Outcome 4-4)
4-4. **d, a,** and **c.** (Outcome 4-4)
4-5. Cindy is following the procedure as she was taught; this means that the manufacturer's instructions may have changed since she was trained. Additional explanations

may include the use of blood for a kit for which urine is the required specimen, or using an expired kit or one that has been stored incorrectly. This scenario stresses the need to follow the package insert instructions. (Outcome 4-6)

4-6. Without a date, there is really no reason to document the result, as it is not evident when the test was performed and which patient results were dependent on that result as verification that the testing device or instrument was working correctly. (Outcome 4-7)

4-7. No, it is not used to document qualitative results, only quantitative. It would not be possible to use this sort of chart to document positive and negative answers, as it is designed to show an acceptable range of results for a specific test. (Outcome 4-7)

4-8. A calibration sample has a known value, not a known range of acceptable values. Once the calibration specimen is tested, the instrument is adjusted until it reads precisely the known value. A quality control specimen has a known range of values, and it is designed to show that the machine is within calibration. (Outcome 4-8)

4-9. It is not required for all laboratories. Those that perform only CLIA-waived tests are not required to perform proficiency testing, although some sort of outside verification of performance is recommended. (Outcome 4-8)

4-10. The most important step that a medical assistant can take is to follow the manufacturer's instructions specifically. (Outcome 4-8)

Chapter 5

5-1. Enforcement: Laws are enforced by governmental agencies, and those who are noncompliant may be punished financially or even imprisoned. Ethics are not enforced or punished in the same way.

Creation: Laws are created by government agencies; ethics are either created personally by the individual based on morals and values or they are created by a professional organization. (Outcome 5-2)

5-2. **b.** Implied consent. (Outcome 5-3)

5-3. Yes, she has the right to get a copy of her medical record, although she might have to wait for it to be copied. No, it is not all right for her to take the entire original record from the office. (Outcome 5-4 and Outcome 5-5)

5-4. No; assault means the threat of injury, not the actual injury. (Outcome 5-1)

5-5. An intentional tort means that a wrong was committed by one person against another person intentionally, or on purpose. An unintentional tort is a case of accidental wrongdoing. (Outcome 5-6)

5-6. A medical assistant is certainly liable, and it is possible for him or her to be sued for malpractice. (Outcome 5-7)

5-7. No, the scope of practice may change state to state, depending on the laws for a particular state. (Outcome 5-8)

5-8. No, it also includes the portability of health insurance, and the method in which health-care information is transmitted for reimbursement, as well as standardization of health-care information. (Outcome 5-9)

5-9. Several standards are presented in the AAMA code of ethics, including respecting patients, keeping confidential information confidential unless it is required by law, upholding high principles of the profession, being a life-long learner, and being involved in the community. (Outcome 5-10)

5-10. This is most definitely an example of risk management, as appropriate documentation is very important to avoid potential legal issues. (Outcome 5-11)

Chapter 6

6-1. **d.** Compound microscope. (Outcome 6-2)

6-2. The stage is the surface where the slide is placed for observation/examination. (Outcome 6-3)

6-3. The condenser concentrates the light from the light source. (Outcome 6-3)

6-4. **False;** this outcome should be cleaned last. (Outcome 6-4)

6-5. **c.** Lens paper. (Outcome 6-4)

6-6. The coarse adjustment knob should be used first. (Outcome 6-5)

6-7. No, the stage should not be moved once the specimen is in focus with a lower magnification when changing from one outcome to another. (Outcome 6-5)

6-8. **b.** Electron microscope. (Outcome 6-2)

6-9. Generally, medical assistants do not perform microscopic examinations in which they are identifying different formed elements in a specimen. However, with appropriate training, a medical assistant may be able to perform a normal manual differential count, and also to perform urine microscopic examinations. (Outcome 6-6)

6-10. Answers may vary, but they may include cleaning the centrifuge, using a tachometer to monitor the rpms, checking the centrifuge for any cracks or discolorations, and checking the electrical system for wear. All these procedures would, of course, be documented. (Outcome 6-7)

6-11. Whenever the centrifuge is used, a container of similar size with the same volume of liquid must be placed across from the specimen before turning on the centrifuge. (Outcome 6-7)

6-12. Temperatures should be monitored at least daily. In situations in which the temperature control is critical (as in blood bank facilities), the temperatures must be monitored continuously. In these critical situations, an alarm system is also in place to verify that the temperature stays within the established range for that storage unit. (Outcome 6-8)

6-13. Answers may vary, but could include the following:
- On-site testing allows for a definitive diagnosis to be established as well as a plan of treatment while the patient is still in the office.
- CLIA-waived tests performed in the office may be less expensive.
- Patients are often more compliant because the tests can be performed in the office during an office visit. (Outcome 6-10)

6-14. This method of measurement uses an electrical current. When the specimen is passed through an aperture that runs between the two electrodes for the current, the number, size, and other characteristics of the cells in the specimen can be measured. The amount of disruption (or impedance) to the electrical current provides the information needed for the measurement. (Outcome 6-11)

6-15. **d.** Whole blood. (Outcome 6-13)

6-16. Yes, the results are visible on the screen of the instrument, and often the instrument is capable of printing the results at the time of the test as well. (Outcome 6-13)

6-17. The Cholestech LDX is a common CLIA-waived instrument used for cholesterol testing in the physician office laboratory. (Outcome 6-14)

6-18. No, it is not measured from intact red blood cells. The cells must be broken (or lysed) before the hemoglobin test can be performed. (Outcome 6-15)

6-19. No, the glucometer is an instrument used to test glucose levels in the blood, which is a type of chemistry test, not a hematology test. (Outcome 6-14)

6-20. **d.** Meniscus. (Outcome 6-17)

Chapter 7

7-1. Demographic information, documentation of test ordered, ICD-9 codes, last dose of medication if a drug level is ordered, type of fluid to be tested, wound source, date of testing, name of ordering physician. (Outcome 7-2)

7-2. Both types of panels are designed to assist the healthcare provider with an easier diagnosis by grouping tests that complement one another in a panel. (Outcome 7-3)

7-3. This is not an acceptable order, because standing orders cannot be extended for more than a year. (Outcome 7-4)

7-4. No repeat collections needed, quicker opportunity for diagnosis, faster treatment. (Outcome 7-5)

7-5. No, it should not because there is no way of verifying if it is the armband that belongs to the patient who is in the bed. (Outcome 7-6)

7-6. The medical assistant should ask him if he has had coffee in the past 12 hours before she begins the blood draw. (Outcome 7-8)

7-7. The reference ranges for most laboratory tests are dependent on many factors, including the collection method. Reference ranges for arterial blood are different from those of venous specimens, which are different from capillary specimens. In addition, the reimbursement for the collection may be different for the various types of blood specimen collections. (Outcome 7-10)

7-8. Initials or identification of the person who has collected the specimen, date and time of collection, patient name, and unique identifier (such as patient ID or birth date). (Outcome 7-12)

7-9. The specimen has a special label placed over the top of the container that is tamper evident. Also, the chain of custody form provides documentation of all those who had access to the specimen. The specimen should be kept in a locked area with limited access when it is not in the possession of a laboratory employee. (Outcome 7-13)

Chapter 8

8-1. The heart, arteries, veins, and capillaries. (Outcome 8-1)

8-2. Arteries have a pulse, veins do not. Arteries carry bright red blood that has just been oxygenated, whereas the blood in veins is darker. Veins have valves to help the blood to move in one direction; arteries have blood pulsing through and do not need the valves. Upon palpation, they feel different—the arteries are a bit stiffer; veins are spongier. (Outcome 8-3)

8-3. No. Capillaries are much smaller; they are microscopic in size. (Outcome 8-4)

8-4. The antecubital area. (Outcome 8-5)

8-5. No. They are also performed on the heel when drawing an infant. (Outcome 8-6)

8-6. Not always; there are times when there is no other choice, so certain precautions have to be taken before drawing the blood near a site at which an IV is running. (Outcome 8-7)

8-7. The edema can add additional fluid to the sample, causing erroneous results. There is an increased chance of infection, and the blood draw may be very difficult. (Outcome 8-7)

8-8. Needle, tubes, sharps biohazardous disposal container, gauze, alcohol pads, tube holder or syringe, gloves. (Outcome 8-8)

8-9. No. The ones used for the evacuated tube system have two ends that are sharp; syringe needles have only one end that is sharp, and they screw onto the syringe. Butterfly systems have needles that are much shorter and attach to a long piece of tubing. (Outcome 8-8)

8-10. Because they have the air removed from the tube to create a vacuum and allow the blood to enter without any additional "pulling" action. (Outcome 8-8)

8-11. The stoppers are color coded to indicate the type (or absence of) additives. (Outcome 8-10)

8-12. To avoid potential crossover contamination from one type of tube to another. This contamination may cause erroneous results. (Outcome 8-11)

8-13. Because the reference ranges are different for each type of blood specimen. The capillary specimens generally have more fluid contamination so the results may be much different. (Outcome 8-13)

8-14. No, it is not. The capillary samples don't have the potential contamination issues from each tube that the venipuncture procedures have, so the order of draw is based more on the tubes that might clot accidentally if they are left unattended too long. (Outcome 8-11)

8-15. Proper patient identification procedure is the most important aspect. (Outcome 8-14)

8-16. The specimens must be drawn into a tube that does not contain anticoagulant. After the draw is complete, the sample must sit for at least half an hour undisturbed at room temperature to allow for a clot to form. After a clot has formed, the sample can be centrifuged for at least 10 minutes so that complete separation can occur. Finally, the serum may be separated using various techniques. (Outcome 8-17)

8-17. Serum does not contain clotting factors, as they are used up in the blood clot. Plasma does contain the clotting factors. Serum is taken from a tube that does not contain anticoagulant. Plasma is taken from a tube that contains anticoagulant. (Outcome 8-18)

8-18. False. Tests that require whole blood must be thoroughly mixed prior to testing. (Outcome 8-17)

8-19. d. All of the above. (Outcome 8-19)

8-20. False. The angle at which the needle is inserted for venipuncture plays a significant role in the success of the procedure and avoidance of the negative outcome. (Outcome 8-20)

Chapter 9

9-1. Diagnosis of urinary tract infections, chemical analysis. (Outcome 9-2)

9-2. Labeling procedures, documentation of medication, documentation of collection times for specimen, communication to the patient for collection procedures. (Outcome 9-2)

9-3. To avoid external contaminants which may lead to a misdiagnosis and incorrect patient treatment. (Outcome 9-2)

9-4. The urine specimen container should be labeled before it is given to the patient. (Outcome 9-6)

9-5. A straight or intermittent catheterization procedure is used to collect a urine specimen for culture, if needed. (Outcome 9-3)

9-6. The patient should be told not to urinate directly into the container. (Outcome 9-7)

9-7. The first morning void specimen is collected immediately after waking up. The 24-hour urine specimen is collected over 24 hours. (Outcome 9-3)

9-8. Because it is not possible to control the urination times of an infant, the device has to be ready to collect the specimen whenever it is generated. (Outcome 9-3)

9-9. Because microscopic pathogens may be present, and OSHA recommends that gloves are worn whenever there is a potential for exposure to body fluids. (Outcome 9-4)

9-10. The bacterial count. (Outcome 9-5)

Chapter 10

10-1. Pathogen. (Outcome 10-1)

10-2. Specimen source, patient name, name or ID for collector, date, and time of collection. (Outcome 10-2)

10-3. Ideally, the sample should be collected before the antibiotic therapy has been initiated. (Outcome 10-2)

10-4. It may differ in the form that it takes (liquid versus solid), the container it is provided in, and by how selective it may be for specific types of samples. (Outcome 10-3)

10-5. Bacterial samples will grow more quickly than the other types listed. (Outcome 10-4)

10-6. No, this will contaminate the specimen with organisms that were not present in the area of specimen collection. (Outcome 10-2)

10-7. They both use swabs for their collection, and they are both collected from the same area at the back of the throat. Also, the PPE used for both types of collection is the same. (Outcome 10-5)

10-8. No, it is best to collect two swabs, as one is used for the strep screen and one is used for the culture. If one swab is collected and the culture is set up before the strep screen is performed, there is a chance that the strep screen could produce a false negative because of an insufficient sample volume. If the strep screen is performed first, the culture cannot be set up from the swab afterward, as there are chemicals added to the swab that would affect the bacterial growth. (Outcome 10-5)

10-9. No, they are not. Ideally, for both types of collections, the medical assistant will touch areas that appear most inflamed, and if there are any white pustules, those will be areas to target as well. (Outcome 10-5)

10-10. Many of the pathogens present in the sputum of infected patients are spread by droplets while the patient is coughing or sneezing. Collection of the sputum specimen in the waiting area could potentially expose all those who are present and even those who follow, as the droplets may land on surfaces that they touch. Sputum specimens should be collected in a private area; it is recommended that the patients collect them in the privacy of their home. (Outcome 10-6)

10-11. **a.** Specimens taken from indwelling catheters. (Outcome 10-7)

10-12. The bottles include different nutrition sources and additives. One bottle is designed to encourage the growth of aerobic microorganisms, whereas the other is designed for anaerobic microorganism growth. (Outcome 10-8)

10-13. **b.** Two times (Outcome 10-8)

10-14. The CSF tubes must be marked according to their order of collection. This allows the laboratory to process them appropriately when performing chemistry, hematology, and microbiology testing procedures. (Outcome 10-9)

10-15. **c.** Swabs (Outcome 10-10)

10-16. **False.** It is definitely advisable to wash a wound before collecting a culture specimen, so that any extraneous (transient) microorganisms are washed away from the area. (Outcome 10-11)

10-17. It is important because stool specimens include a great deal of normal flora. The media used for this type of specimen encourages the growth of the pathogenic microorganisms that are present in the GI tract, while discouraging the growth of the normal flora to avoid overgrowth. This is more important for stool specimens than it is for any other type of specimen because the normal flora occurs in such high numbers. (Outcome 10-12)

10-18. It reduces the discomfort of using a dry swab in the eye; these patients usually are already in pain from the eye infection. (Outcome 10-13)

10-19. **a.** Otitis externa. This is often linked to swimming or introduction of liquid to the ear canal, as can occur in bathing or showering. (Outcome 10-14)

10-20. No, fungal infections can occur in any area of the body. (Outcome 10-15)

10-21. KOH preps and wet mounts are similar in that they are used for examination of microorganisms in their living state, without preservative or staining. Ideally, the movement of the organisms can be evaluated, as well as the morphology. They are also similar in that both methods may be used to examine samples taken from the same areas of the body. KOH preps are different from wet mounts because the KOH prep is designed specifically to clear away all elements from the sample that are not fungal in nature; the wet mount allows for examination of everything in the specimen. (Outcome 10-16)

10-22. **a.** The O&P collection procedures include containers with preservative solution, but the stool culture samples procedures do not. (Outcome 10-17)

10-23. To collect the sample for detection of pinworms or their ova, clear adhesive tape is pressed against the rectum to pull out the worms or eggs. The adhesive tape is then placed on the slide to be examined under the microscope. (Outcome 10-18)

10-24. The two different types of stains allow for bacteria to be differentiated with the staining procedure. By using two different colors of stains, it is possible to identify bacteria as gram positive (those that absorb the purple stain in their cell walls) or gram negative (those that absorb the pink counterstain in their cell walls). If only one stain were used, it would not allow a chance to examine the bacteria that did not absorb the initial stain, as they would be colorless. (Outcome 10-19)

10-25. **d.** All of the above. (Outcome 10-20)

10-26. Antibiotic discs are used to apply different types of antibiotics to the bacteria so that the microbiologist can determine which antibiotic may be most effective to treat the infection. The antibiotics that do not allow the bacteria to grow close to their borders are more effective in treating that specific bacteria than are those antibiotics that allow the growth to continue right up to their borders. (Outcome 10-21)

10-27. C&S means "culture and sensitivity." This is a procedure in which the specimen is cultured and the infection-causing pathogen is identified Then an antibiotic sensitivity is performed to determine which medication is most effective for use. A C&S is not performed for viral or fungal cultures, as antibiotics are not used to treat these infections. (Outcome 10-22)

Chapter 11

11-1. Erythrocytes, leukocytes, and thrombocytes (red cells, white cells, platelets) make up the formed elements in blood. (Outcome 11-3)

11-2. Hematopoiesis is the process of the production of blood cells. (Outcome 11-2)

11-3. *Pluripotent* means "having the ability to differentiate into several types of specialized cells." (Outcome 11-2)

11-4. No, the nucleus is expelled from the red blood cell before it enters the general circulation. Mature red blood cells do not include a nucleus. (Outcome 11-3)

11-5. There may be an increase in the number of reticulocytes following blood loss. (Outcome 11-4)

11-6. The neutrophil is the white blood cell that first fights bacterial infection; therefore, the numbers increase as the cells are employed. (Outcome 11-5)

11-7. The eosinophil count may increase in an allergic reaction. It is a granulocytic white blood cell with a bilobed nucleus and bright red-orange granules in the cytoplasm. (Outcome 11-5)

11-8. The lymphocyte has forms that make antibodies and are active in cell-mediated immunity. (Outcome 11-5)

11-9. Platelets stick together to form the platelet plug to fill the break in the vessel, release serotonin to help the vessel contract, and start the chemical reactions of coagulation to increase the stability of the platelet plug. (Outcome 11-7)

11-10. Three CLIA-waived procedures are hemoglobin, hematocrit, and erythrocyte sedimentation rate testing. (Outcome 11-8)

Chapter 12

12-1. The CBC (complete blood count) tests are the hemoglobin, hematocrit, red blood cell count, white blood cell count, and the differential. The erythrocyte indices and a platelet count are often included. (Outcome 12-2)

12-2. The normal range is 4.2 to 6.2 million red blood cells per cubic millimeter of blood. (Outcome 12-3)

12-3. The white blood cell count is increased in bacterial infection and in leukemia. (Outcome 12-4)

12-4. The red blood cell indices describe the size of red blood cells and the amount of hemoglobin that they contain. (Outcome 12-6)

12-5. The appearance of the white blood cell nuclei and the cytoplasm is used for differentiation. In addition, the colors added by the polychromatic stain aid in the identification process. The nuclei, cytoplasm, and granules all absorb different colors. (Outcome 12-7)

12-6. From greatest in number to least, they are neutrophils, lymphocytes, monocytes, eosinophils, and basophils. (Remember—Never let monkeys eat bananas.) (Outcome 12-7)

12-7. The normal red blood cell is circular, with a pale center. (Outcome 12-9)

12-8. A pale red blood cell that has a decreased amount hemoglobin is hypochromic. (Outcome 12-9)

12-9. *Poikilocytosis* is the word used to describe a variety in the shapes of red blood cells on the blood smear. (Outcome 12-9)

Chapter 13

13-1. Hemoglobin is made up of one molecule of globin and four molecules of heme. (Outcome 13-2)

13-2. Oxygen delivery to the tissues of the body. (Outcome 13-2)

13-3. Answers may vary, but examples of abnormal hemoglobin strains are Hb S, Hb C, Hb D, Hb G, Hb J, and Hb M. (Outcome 13-3)

13-4. No, the cells must be lysed first. (Outcome 13-4)

13-5. Errors may include overfilling or underfilling the tube, sealing inadequately with the clay on the end of the tube, using an incorrect speed for the centrifuge, allowing the specimen to sit for too long before reading the result, contaminating with interstitial fluid, and using an incorrect reading technique. (Outcome 13-5)

13-6. The hematocrit is the amount of space occupied by red blood cells in a sample of whole blood. It is reported as a percentage. (Outcome 13-5)

13-7. The centrifuge is used to create the three layers in performing the hematocrit. (Outcome 13-5)

13-8. Anemia is the disorder in which there is low oxygen-carrying capacity of blood. (Outcome 13-9)

13-9. The physician could order a hemoglobin and/or hematocrit. The red blood cell count and indices could also give additional information for anemia. (Outcome 13-8)

13-10. Iron deficiency, pernicious, aplastic, and sickle cell anemias are examples. (Outcome 13-10)

Chapter 14

14-1. The sedimentation rate is a laboratory test where the settling of red blood cells is observed. (Outcome 14-1)

14-2. Rouleaux formation is the stacking of red blood cells on top of one another, as in a stack of coins. It occurs when the red cells are no longer repelled from one another. (Outcome 14-1)

14-3. Increased fibrinogen and certain globulin levels in the plasma decrease the negative charge on the red blood cells, and they are not as repelled from each other and will stack more easily into the rouleaux formation. (Outcome 14-2)

14-4. There is less surface area to the volume of the stacked cells so the stack will settle faster in the plasma than single red blood cells. (Outcome 14-3)

14-5. For those younger than 50 years of age, the normal value for men is less than 15 and the normal value for women is less than 20 mm/hr. (Outcome 14-4)

14-6. The inflammatory process causes the shift in proteins that will lead to an increased ESR. (Outcome 14-5)

14-7. The Sediplast Westergren erythrocyte sedimentation system and the Wintrobe ESR system. (Outcome 14-6)

14-8. Any two of the following procedure errors will affect the value of the ESR: using a clotted sample could cause erroneous results; using an EDTA tube that is not well mixed; using a sample that is at refrigerated temperature; leaving air spaces in the column that is filled with the aliquot of blood; not keeping the column vertical during the hour of the procedure; placing the column in direct sunlight or in an area where there is vibration or movement. Also, reading the test at less than 60 minutes would cause an erroneous result. (Outcome 14-7)

Chapter 15

15-1. It may be done to screen for possible coagulation disorders or to monitor medication use. (Outcome 15-2)

15-2. No; there are other factors that are involved in addition to the platelets. (Outcome 15-3)

15-3. The constriction of blood vessels in the immediate and surrounding area is the body's first response to vessel injury. (Outcome 15-3)

15-4. This term refers to a series of events that seal off vessel injury and secondary coagulation steps that follow the formation of the initial platelet plug. (Outcome 15-3)

15-5. A thrombus is a blood clot that is stationary; an embolus is a part of a blood clot or a whole blood clot that has left the point of formation and has traveled through the body. (Outcome 15-4)

15-6. DIC and DVT are similar because they both involve formation of blood clots. DVT is a deep vein thrombosis that is contained in a particular vessel, whereas DIC is intravascular coagulation that is spread throughout the body. (Outcome 15-5)

15-7. Thrombocytopenia is different because it is a disorder specific to the number of platelets; all the other disorders discussed are in reference to other aspects of the clotting process. (Outcome 15-5)

15-8. No, the INR is used when monitoring a patient who is taking warfarin or Coumadin, not for initial screening. It may be reported as part of the initial screening, but the health-care provider will not use the INR to make a decision on potential coagulation defects without the prothrombin time result as well. (Outcome 15-6)

15-9. **c.** Heparin. (Outcome 15-8)

15-10. The patient may exhibit slow or inadequate blood clotting with blood vessel injury, resulting in excessive bruising and/or bleeding. (Outcome 15-11)

15-11. The answers may vary, but should include the following: results are not reproducible; aspirin has a profound effect; other medications and health conditions may affect the result to a high degree; and the bleeding time result does not seem to correlate well to the true risk of bleeding in surgical situations. (Outcome 15-10)

15-12. **False.** It is imperative that these tubes are filled to the point at which the vacuum is exhausted. (Outcome 15-13)

Chapter 16

16-1. Qualitative tests provide a result that indicates whether the specimen is positive or negative for a specific substance. Quantitative tests provide an actual number that indicates the amount of a substance that is present in the specimen. (Outcome 16-1)

16-2. **a.** Serum. (Outcome 16-2)

16-3. Glucose testing may be performed in a physician office laboratory to screen for the presence of diabetes mellitus or gestational diabetes, or to monitor the progress of a patient who is being treated for these conditions. (Outcome 16-4)

16-4. **b.** LDL. (Outcome 16-5)

16-5. **d.** Cadmium. (Outcome 16-6)

16-6. Panels allow the health-care provider to have access to a variety of test results, which can often aid in a differential diagnosis.

If the provider orders only one test at a time, the patient may need to return to the office several times to have specimens collected, which would be cumbersome and time consuming.

Some panels also allow the health-care provider to evaluate more than one organ system at once, which can help to establish a diagnosis.

Panels may be less expensive in some situations. (Outcome 16-8)

16-7. **True.** This is the most common reason that these tests are ordered. (Outcome 16-7)

16-8. When damage to the heart muscle occurs, the specific cardiac enzymes become elevated at different rates. To determine whether a myocardial infarction occurred, it is essential to collect a series of specimens to see if the enzymes have become elevated. (Outcome 16-10)

16-9. **a, b,** and **c.** (Outcome 16-11)

16-10. **a.** Lipemia. (Outcome 16-13)

Chapter 17

17-1. Glycogen is the storage form of glucose in the body; it is produced when there is excess glucose present in the bloodstream and stored in the liver and the adipose tissues. (Outcome 17-1)

17-2. **a.** Formation of glycogen. (Outcome 17-3)

17-3. Answers may vary and might include the following: Type 1 and type 2 diabetes both exhibit elevated levels of blood glucose. They both may exhibit similar symptoms. Both type 1 and type 2 diabetes are diagnosed using blood glucose levels, in addition to physical examination and medical history. Both types of diabetes are caused by problems with insulin in the body: a decrease in the amount of glucose, or cellular resistance to glucose. (Outcome 17-4)

17-4. Diabetic neuropathy, damage to the optic nerve, cardiovascular damage, irreversible kidney damage, poor healing. (Outcome 17-6)

17-5. Gestational diabetes is more like type 2 diabetes because in both there is adequate insulin present, but the cells develop a resistance to the action of the insulin. In type 1 diabetes, the patient does not have enough insulin produced. (Outcome 17-5)

17-6. Untreated gestational diabetes may lead to macrosomia (an excessively large fetus). (Outcome 17-7)

17-7. **c.** A glucose level drawn at any time of the day without any previous preparation. (Outcome 17-8)

17-8. **True.** A laboratory postprandial test involves the ingestion of a glucose drink. A postprandial blood glucose test done at home is performed a certain amount of time after a meal. (Outcome 17-9)

17-9. The blood is drawn a minimum of four times if the test is completed entirely. This includes a fasting, 1-hour, 2-hour, and 3-hour draw. (Outcome 17-10)

17-10. **c.** 3 months. (Outcome 17-11)

17-11. The plasma glucose will be higher than the whole blood glucose. (Outcome 17-12)

17-12. Medical assistants may assist with this process in many ways. These include helping the patients understand how their plasma glucose and whole blood levels will be different from one another, showing the patients how to use their blood testing equipment properly, helping patients find an instrument that works well for them, explaining how to document blood glucose levels properly, and assisting with proper use and disposal of lancets. MAs may also assist patients with reimbursement for diabetes testing supplies by researching the policies of their insurance. (Outcome 17-13)

Chapter 18

18-1. Ingested cholesterol comes from animal products (like meat), dairy products, and processed foods. (Outcome 18-2)

18-2. LDL and VLDL. (Outcome 18-3)

18-3. Yes, it is desirable to have high HDL levels and low LDL levels, respectively. (Outcome 18-4)

18-4. No, this result is not within the desirable range. The reference range (or desirable range) is less than 200 mg/dL. (Outcome 18-4)

18-5. Elevated LDL levels are considered a health risk because they directly contribute to plaque buildup by depositing cholesterol on the lining of the blood vessels. (Outcome 18-5)

18-6. **a.** Consumption of excess calories. (Outcome 18-6)

18-7. **b.** Excess fat particles. (Outcome 18-7)

18-8. **False.** Electrolytes exhibit a positive or negative charge when dissolved in water. (Outcome 18-1)

18-9. The electrolyte most commonly associated with arrhythmia is potassium. (Outcome 18-13)

18-10. **c.** Nitrogen. (Outcome 18-14)

Chapter 19

19-1. The physical examination of the urine specimen includes observations of the urine that can be seen with the naked eye; the microscopic procedures require the use of a microscope for magnification and are generally performed on urine sediment. (Outcome 19-7)

19-2. The function of the urinary system is the filtration of blood, excretion of waste, and the regulation of the body's fluid and acid-base balance. (Outcome 19-2)

19-3. **c.** Located on either side of the vertebral column with one slightly lower than the other. (Outcome 19-3)

19-4. Because urine is formed in the nephron. (Outcome 19-3)

19-5. Reabsorption and secretion. (Outcome 19-4)

19-6. No, ureters are only involved in the transportation of urine, not in the actual formation of urine. (Outcome 19-4)

19-7. **False.** Urinalysis testing is performed for a variety of reasons, including screening of asymptomatic individuals. (Outcome 19-6)

19-8. The physical examination procedures include observation of color, clarity, and sometimes the odor of the urine. (Outcome 19-7)

19-9. **c.** Chemical analysis. (Outcome 19-7)

19-10. **d, a,** and **b.** (Outcome 19-9)

Chapter 20

20-1. Urochrome is the pigment that gives urine its characteristic yellow color. (Outcome 20-1)

20-2. The bright orange color present in the urine is not clinically significant; it is merely a result of the medication she was given, probably pyridium, which is prescribed to those with recurrent urinary tract infections. (Outcome 20-2)

20-3. Using some sort of print or object behind the urine specimen allows the person performing the test to assess the transparency of the urine with more accuracy; it gives a point of reference. (Outcome 20-3)

20-4. **b.** Decreased. (Outcome 20-4)

20-5. No, there are other causes for strong odors in addition to a urinary tract infection. (Outcome 20-4)

20-6. **False.** A high specific gravity means that the specimen has a lot of dissolved substances; this does not necessarily mean that the urine is cloudy. High levels of large molecules such as glucose and protein do not make the urine cloudy. (Outcome 20-5)

20-7. No, this is not correct. Urine specific gravity must always be reported with three spaces to the right of the decimal point, the first of which is a zero, for example, 1.025. (Outcome 20-7)

20-8. Refractometers take only a drop of urine, they are relatively inexpensive, and they may be portable. (Outcome 20-6)

20-9. No, partly hazy is not a valid description of the clarity of the urine specimen. (Outcome 20-8)

Chapter 21

21-1. **False.** The reference range for many of the analytes is actually a negative result; this means that they are not normally present in the specimen. (Outcome 21-3)

21-2. Bile obstruction or liver dysfunction. (Outcome 21-2)

21-3. **a.** Intact red blood cells. (Outcome 21-3)

21-4. No. If the diabetic has too much glucose in the bloodstream at any given time, the urine may be positive for glucose. However, if the blood glucose level is under control (through diet, exercise, medication, or insulin use), then the urine may be negative for glucose. The values can fluctuate quite a bit throughout the day. (Outcome 21-2)

21-5. **False.** The leukocyte esterase result will not always be positive when white blood cells are present in the urine specimen, because not all white blood cells contain this enzyme. However, the only ones that do not secrete this enzyme are the lymphocytes, and they are not usually the type of white blood cell present in an infection. (Outcome 21-3)

21-6. **b.** Fever. (Outcome 21-3)

21-7. The strips can be exposed to moisture by leaving the lid off the container they are kept in or removing the desiccant packaged with the strips. They can also be exposed to direct light, excessive heat, or volatile liquids. They can be used past their expiration date or when discolored. (Outcome 21-4)

21-8. **b.** Allowing the reagent strip to remain in the urine for 60 seconds before removing the strip. (Outcome 21-4)

21-9. The most common route of exposure is through mucous membrane exposure by splashing or aerosol exposure. (Outcome 21-5)

21-10. **True.** Some only test for one or two analytes (e.g., glucose and ketones), and some do or do not provide an opportunity for specific gravity testing even if they are the more comprehensive strips. (Outcome 21-6)

21-11. Less opportunity for operator error, less time intensive for the person performing the test, and opportunities for multiple testing procedures to occur at once are all advantages of automated urine chemistry testing. (Outcome 21-6)

21-12. Because the FOBT is a means of detecting early, asymptomatic colorectal cancer. (Outcome 21-9)

21-13. **d.** NSAIDs. (Outcome 21-10)

Chapter 22

22-1. It may be added as a reflexive test when the chemical tests are abnormal. The urine microscopic may also be performed when ordered by the practitioner; this may occur if the practitioner suspects the presence of certain formed elements in the urine specimen, such as crystals or casts. (Outcome 22-2)

22-2. No, the presence of squamous epithelial cells is not considered clinically significant. (Outcome 22-4)

22-3. **b** Yes, spermatozoa may be present in the urine of females if they have recently been involved in sexual intercourse. (Outcome 22-3)

22-4. **a.** Viruses. (Outcome 22-4)

22-5. The KOVA system allows for a standardized amount of sediment to be visualized, and it also adds a stain to the sediment before it is placed on the slide for visualization. The stain will allow for more contrast with structures that appear similar, which will aid with identification. (Outcome 22-5)

22-6. **False.** Many normal urine specimens have very few formed elements present in the urine sediment. (Outcome 22-4)

22-7. A medical assistant may be responsible for providing instructions to the patient for specimen collection; accepting and logging in the specimen; pouring off some of the sample into a tube to be centrifuged; centrifuging the urine specimen; processing the centrifuged specimen to place the sediment on a slide; adding stain to the sediment for visualization; focusing the microscope for visualization of the formed elements on the slide. (Outcome 22-8)

Chapter 23

23-1. Antigens stimulate an immune response in our bodies, and antibodies may be produced in response to protect us. (Outcome 23-2)

23-2. It is usually tested in the serology area of the laboratory. (Outcome 23-2)

23-3. The injured cells release a chemical that signals for the internal nonspecific responses to begin. These include histamine production and white blood cell migration. (Outcome 23-3)

23-4. They are good for our immune system. They often occur in response to a foreign object in our nasal passages, and they can expel this before it causes a problem for our body. (Outcome 23-3)

23-5. No. If the first or second line of defense destroys the foreign invader, the subsequent level is never involved in the process. (Outcome 23-4)

23-6. The third line of defense is the one that is very specific toward each antigen. This includes antibody production. (Outcome 23-4)

23-7. It means that we have been exposed (either naturally or artificially) to that disease before, and our body has produced antibodies that will react immediately to protect us in subsequent exposures. (Outcome 23-5)

23-8. Vaccines "fool" our immune system into producing antibodies to a specific antigen without ever becoming ill with that disease. They target the third line of defense. (Outcome 23-5)

23-9. A antigens and B antigens. (Outcome 23-8)

23-10. Yes, she will probably be receiving a RhoGam injection to keep her body from forming anti-Rh that might cause harm to this fetus or those in the future. (Outcome 23-8)

23-11. They involve antigen/antibody reactions as part of the testing procedure. (Outcome 23-9)

23-12. Yes. An example may be the tests for syphilis VDRL and RPR. (Outcome 23-9)

23-13. The results are available quickly so that the physician can take immediate action. They are also usually less expensive than other methods may be for the patient. (Outcome 23-10)

23-14. The Epstein Barr virus. (Outcome 23-11)

Chapter 24

24-1. Immunology is the study of the immune system, including the antigens and antibodies involved in the process. Serology is the study specifically of the antigens and antibodies; therefore, it is part of immunology. They are similar because they both include the study of the reactions of antigens and antibodies. (Outcome 24-1)

24-2. The test is demonstrating a lack of specificity. (Outcome 24-1)

24-3. Most are immunoassays in which the reagents and the sample are mixed and move across the testing surface to produce a chromatographic result indicating a positive or a negative. They are also similar because they usually take less than 15 minutes to obtain a result. (Outcome 24-3)

24-4. Answers may vary, but should include the following: The rapid tests allow for faster treatment and better patient compliance because the patients are in the facility when the diagnosis is made and treatment is prescribed; the tests are usually cost effective, especially so for tests that are seldom performed; the results are ready in a short period of time. (Outcome 24-2)

24-5. Complications include glomerulonephritis, inflammation of the joints, and endocarditis. (Outcome 24-5)

24-6. No, it is not present elsewhere. (Outcome 24-5)

24-7. When there is an unexpected negative result with a high prevalence of influenza in the population. (Outcome 24-4)

24-8. The patient doesn't need to return for results, so it allows the health-care provider to provide education and counseling to the patient at the same time that the test specimen is collected. This helps to control the spread of the disease by those at high risk who are unaware of their HIV-positive status. (Outcome 24-4)

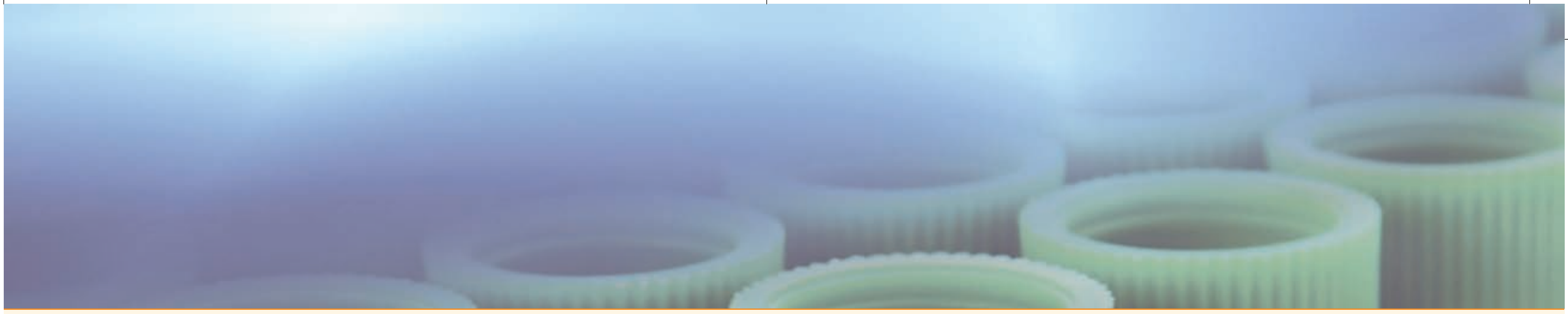

Appendix C

Tube Guide for BD Vacutainer Venous Blood Collection

BD Vacutainer® Venous Blood Collection Tube Guide

BD Helping all people live healthy lives

For the full array of BD Vacutainer® Blood Collection Tubes, visit www.bd.com/vacutainer.
Many are available in a variety of sizes and draw volumes (for pediatric applications).
Refer to our website for full descriptions.

BD Vacutainer® Tubes with BD Hemogard™ Closure	BD Vacutainer® Tubes with Conventional Stopper	Additive	Inversions at Blood Collection*	Laboratory Use
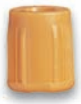 **Gold**	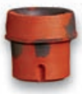 **Red/ Gray**	• Clot activator and gel for serum separation	5	For serum determinations in chemistry. May be used for routine blood donor screening and diagnostic testing of serum for infectious disease.** Tube inversions ensure mixing of clot activator with blood. Blood clotting time: 30 minutes.
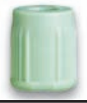 **Light Green**	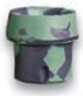 **Green/ Gray**	• Lithium heparin and gel for plasma separation	8	For plasma determinations in chemistry. Tube inversions ensure mixing of anticoagulant (heparin) with blood to prevent clotting.
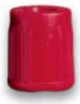 **Red**	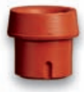 **Red**	• Silicone coated (glass) • Clot activator, Silicone coated (plastic)	0 5	For serum determinations in chemistry. May be used for routine blood donor screening and diagnostic testing of serum for infectious disease.** Tube inversions ensure mixing of clot activator with blood. Blood clotting time: 60 minutes.
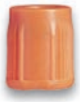 **Orange**		• Thrombin-based clot activator with gel for serum separation	5 to 6	For stat serum determinations in chemistry. Tube inversions ensure mixing of clot activator with blood. Blood clotting time: 5 minutes.
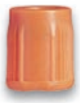 **Orange**		• Thrombin-based clot activator	8	For stat serum determinations in chemistry. Tube inversions ensure mixing of clot activator with blood. Blood clotting time: 5 minutes.
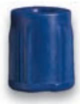 **Royal Blue**		• Clot activator (plastic serum) • K_2EDTA (plastic)	8 8	For trace-element, toxicology, and nutritional-chemistry determinations. Special stopper formulation provides low levels of trace elements (see package insert). Tube inversions ensure mixing of either clot activator or anticoagulant (EDTA) with blood.
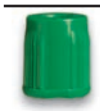 **Green**	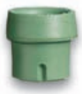 **Green**	• Sodium heparin • Lithium heparin	8 8	For plasma determinations in chemistry. Tube inversions ensure mixing of anticoagulant (heparin) with blood to prevent clotting.
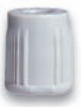 **Gray**	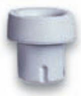 **Gray**	• Potassium oxalate/ sodium fluoride • Sodium fluoride/Na_2 EDTA • Sodium fluoride (serum tube)	8 8 8	For glucose determinations. Oxalate and EDTA anticoagulants will give plasma samples. Sodium fluoride is the antiglycolytic agent. Tube inversions ensure proper mixing of additive with blood.
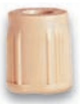 **Tan**		• K_2EDTA (plastic)	8	For lead determinations. This tube is certified to contain less than .01 µg/mL(ppm) lead. Tube inversions prevent clotting.

BD Vacutainer® Tubes with BD Hemogard™ Closure	BD Vacutainer® Tubes with Conventional Stopper	Additive	Inversions at Blood Collection*	Laboratory Use
	Yellow	• Sodium polyanethol sulfonate (SPS) • Acid citrate dextrose additives (ACD): **Solution A -** 22.0 g/L trisodium citrate, 8.0 g/L citric acid, 24.5 g/L dextrose **Solution B -** 13.2 g/L trisodium citrate, 4.8 g/L citric acid, 14.7 g/L dextrose	8 8 8	SPS for blood culture specimen collections in microbiology. ACD for use in blood bank studies, HLA phenotyping, and DNA and paternity testing. Tube inversions ensure mixing of anticoagulant with blood to prevent clotting.
Lavender	**Lavender**	• Liquid K_3EDTA (glass) • Spray-coated K_2EDTA (plastic)	8 8	K_2EDTA and K_3EDTA for whole blood hematology determinations. K_2EDTA may be used for routine immunohematology testing, and blood donor screening.*** Tube inversions ensure mixing of anticoagulant (EDTA) with blood to prevent clotting.
White		• K_2EDTA and gel for plasma separation	8	For use in molecular diagnostic test methods (such as, but not limited to, polymerase chain reaction [PCR] and/or branched DNA [bDNA] amplification techniques.) Tube inversions ensure mixing of anticoagulant (EDTA) with blood to prevent clotting.
Pink	**Pink**	• Spray-coated K_2EDTA (plastic)	8	For whole blood hematology determinations. May be used for routine immunohematology testing and blood donor screening.*** Designed with special cross-match label for patient information required by the AABB. Tube inversions prevent clotting.
Light Blue **Clear**	**Light Blue**	• Buffered sodium citrate 0.105 M (≈3.2%) glass 0.109 M (3.2%) plastic • Citrate, theophylline, adenosine, dipyridamole (CTAD)	3-4 3-4	For coagulation determinations. CTAD for selected platelet function assays and routine coagulation determination. Tube inversions ensure mixing of anticoagulant (citrate) to prevent clotting.
Clear	New **Red/ Light Gray**	• None (plastic)	0	For use as a discard tube or secondary specimen tube.

Note: BD Vacutainer® Tubes for pediatric and partial draw applications can be found on our website.

BD Diagnostics
Preanalytical Systems
1 Becton Drive
Franklin Lakes, NJ 07417 USA

BD Global Technical Services: 1.800.631.0174
BD Customer Service: 1.888.237.2762
www.bd.com/vacutainer

* Invert gently, do not shake

** The performance characteristics of these tubes have not been established for infectious disease testing in general; therefore, users must validate the use of these tubes for their specific assay-instrument/reagent system combinations and specimen storage conditions.

*** The performance characteristics of these tubes have not been established for immunohematology testing in general; therefore, users must validate the use of these tubes for their specific assay-instrument/reagent system combinations and specimen storage conditions.

BD, BD Logo and all other trademarks are property of Becton, Dickinson and Company. © 2010 BD

Printed in USA 07/10 VS5229-13

BD Vacutainer Venous Blood Collection Tube Guide. Courtesy and © 2010 Becton, Dickinson and Company.

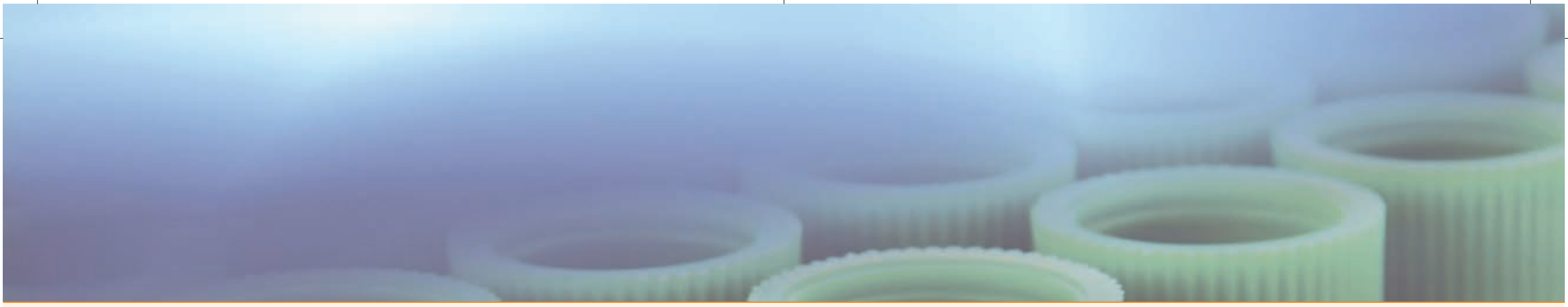

Glossary

1 or 2 hours postprandial (1 hr pp or 2 hr pp) – A specimen drawn 1 or 2 hours after a meal.

24-hour urine collection – A urine specimen collected over a 24-hour period to be used for chemical analysis.

Accuracy – A statistical term representing how close a laboratory result is to the true value.

Acidosis – An increase in the acidity of the blood and body tissues (decrease in pH) because of a buildup of acids in the blood.

Activated partial thromboplastin time (aPTT or PTT) – A test used to measure the time required for a fibrin clot to form after the activation factor, calcium, and a phospholipid mix have been added to the plasma. The results are recorded in seconds.

Advance Beneficiary Notice of Noncoverage (ABN) – A form that must be filled out for Medicare patients when the laboratory test ordered is not likely to be reimbursed.

Aerobes – Bacteria that require oxygen to grow and reproduce.

Aerobic – An organism that requires oxygen to survive and replicate.

Afferent arteriole – The arteriole that transports blood toward the center of the glomerulus of the kidney.

Afibrinogenemia – A condition in which a patient lacks fibrinogen in the bloodstream.

Agar – A solid, jelly-like culture medium usually provided in Petri dishes or tubes.

Agglutination – The clumping of particles (living or nonliving) in fluid. This is often a visible reaction that may be viewed under the microscope.

Aggregate – To come together or cluster in a mass.

Agranular – A classification used to describe white blood cells that do not have granules in their cytoplasm.

AIDS – Acquired immune deficiency syndrome.

Albumin – A plasma protein that functions as a carrier molecule for transporting various chemicals throughout the body. Albumin also functions to prevent plasma from leaking out of the capillaries. Patients with poor glycemic control may have albumin present in their urine, indicating damage to the kidneys.

Aliquot – A small portion of the whole specimen.

Alkalosis – An increase in blood alkalinity resulting from an increased number of alkaline substances or a decrease in the number of acidic substances. The pH will be elevated from the normal range in a state of alkalosis.

Allergy – An immune response to a foreign antigen that results in a hypersensitivity to that antigen. The immune response includes inflammation and dysfunction of body organs or systems.

Ambulatory – Able to walk or transport oneself; not confined to bed.

Ambulatory care – Care provided to patients who are able to ambulate (walk) into the premises. Generally refers to physician offices, urgent care facilities, emergency rooms, and the like.

Amorphous phosphates – A granular diffuse precipitate in alkaline urine that is made up of crystallized calcium and phosphate.

Amorphous urates – A granular diffuse precipitate in acidic urine that is made up of crystallized uric acid.

Anaerobes – Bacteria that do not need oxygen for growth and reproduction.

Anaerobic – An organism that does not need oxygen to survive and replicate. The growth of this type of organism is often actually inhibited by the presence of oxygen in its environment.

Analysis – Assists with differential diagnosis of congestive heart failure or respiratory disease.

Analyte – A chemical substance that is being tested or analyzed.

Anemia – A reduction in the oxygen-carrying capacity of the blood.

Anion – An ion that carries a negative charge. Electrolytes are ions when dissolved in water, and those with a negative charge (such as chloride and bicarbonate) are anions.

Anisocytosis – Condition in which there is inequality in the size of red blood cells.

Antecubital area – The area of the arm that is directly in front of the elbow. This is the preferred area for venipuncture procedures.

Anterior – Directional term meaning "to the front."

Antibiotic sensitivity testing – Performed to determine how effective antimicrobial therapy is against a certain type of bacteria.

Antibody – A protein substance developed in response to antigenic exposure.

Antibody-mediated immunity – A term used to describe an immune response using antibodies; the third line of defense.

Anticoagulants – Substances added to keep blood from coagulating or clotting.

Antigen – A substance that the body detects as "foreign" or "nonself." This may initiate the formation of antibodies.

Antiphospholipid syndrome – A disorder with increased systemic thrombi formation, in which patients have high levels of antibodies against phospholipids in their system.

Anuria – The absence of urine production.

Aperture – An opening or hole through which a controlled amount of a substance, liquid, or light is admitted.

Aplastic anemia – Anemia caused by deficient red cell production in the bone marrow.

Apolipoproteins – Proteins that combine with and transport cholesterol in the bloodstream.

Arm – Part of a microscope that connects the base of the microscope to the tube where the optical lenses are attached.

Arteries – Blood vessels that carry blood away from the heart.

Arterioles – Small arteries that lead into a capillary at its distal end.

Artifacts – Contaminating substances in the urine specimen that have no clinical significance.

Asepsis – A term that means that a surface is without pathogenic microorganisms.

Aseptic – Free of pathogens.

Assault – A term used to describe a threat to inflict injury, with the apparent means to accomplish the threat.

Asymptomatic – Without symptoms or obvious outward signs of disease.

Atherosclerosis – Sclerosis or hardening of the arteries because of the buildup of a waxy, sticky coating on the inside of the vessel walls.

Atrial fibrillation – A disorder in which there is ineffective ejection of blood into the atria. Those with atrial fibrillation have a rapid and irregular heartbeat.

Autoantibodies – An antibody that is produced by the body against the cells of the same body. Autoantibodies are thought to be the cause of the destruction of the insulin-producing islets of Langerhans, leading to type 1 diabetes.

Autoimmune disease – A disease in which the body does not recognize the antigens on the surface of its own cells; it sees these cells as nonself.

Autoimmune response – The response the body when it no longer recognizes all the cells of the body as "self" and begins to damage these cells.

B cells – A type of lymphocyte that remembers an antigen after exposure to it.

Bacilli – Bacteria that are oval or rod shaped when viewed under the microscope.

Bacteria – Single-cell microorganisms that possess a cell wall in addition to the cell membrane that human cells possess.

Balanced – The practice of placing a tube of equal size and similar volume across from the specimen to be processed in the centrifuge. This allows the centrifuge to operate with symmetry, and not be "pulled" to one side or the other as it spins.

Band cell – An immature, unsegmented neutrophil.

Base – The bottom of a microscope that sits on the bench top.

Basic metabolic panel (BMP) – A group of eight tests that have been approved by the CMS for overall health screening and/or for monitoring certain disease states.

Basilic vein – A vein in the antecubital area that is on the inner side (medial) running over the top of the brachial artery. This vein is the last choice of the major veins in this area for venipuncture procedures.

Basophil – A granulocytic white blood cell essential to the nonspecific immune response in inflammation.

Battery – Touching another without his or her consent.

Beaker – A glass container that may be used in the laboratory to hold, mix, or heat chemicals. Beakers usually have a cylinder shape and a flat bottom.

Bence Jones protein – An immunoglobulin (protein) that is present in the renal tubules and excreted in the urine of patients having multiple myeloma.

Bevel – The slanted area of the needle that creates a sharp point at the tip.

Bilirubin – A product produced from the degradation of hemoglobin; it is yellow or orange in color.

Bilobed – Having two lobes or sections.

Binocular – A term used to describe a microscope that has two oculars so that both eyes can be used at once to view the specimen.

Biohazard symbol – Symbol used to mark bags or containers that hold items that are biohazardous in nature.

Biohazardous waste – Liquid or semiliquid blood or other potentially infectious materials. This includes contaminated items that would release blood (or OPIM) in a liquid or semiliquid state if compressed. It also includes items that are caked with dried blood (or OPIM) that are capable of releasing these substances while being handled. This classification also includes pathological or microbiological waste that contains blood or other potentially infectious materials. (See also Regulated waste.)

Bladder – The urinary bladder is a muscular "sac" that holds urine until it is expelled from the body.

Blast – Immature stage in cellular development.

Blood urea nitrogen (BUN) – A by-product of protein metabolism that should be cleared from the blood by the kidneys. Elevated amounts in the blood are indicative of renal dysfunction.

Bloodborne pathogens (BBPs) – Pathogenic microorganisms present in human blood that are transmitted by direct contact to blood or mucous membranes of another person.

Bloodborne Pathogens Standard (1910.1030) – A regulation created by the Occupational Safety and Health Administration to educate and protect health-care workers from occupational exposure to blood or other potentially infectious materials. Originally created in 1988; there have been several addendums since that time.

Body mass index (BMI) – A statistical calculation comparing the weight and height of a person. This calculation is used to estimate whether a patient may be overweight or underweight according to his or her weight.

Bowman's capsule – A portion of the renal tubules that collects the initial urinary filtrate when it is forced from the glomerulus.

Brain natriuretic peptide (BNP) – Enzyme secreted by the muscular walls of the ventricles.

Broth – A liquid culture medium.

Buffer – Compound that reacts with strong acids or bases to minimize pH changes in the body.

Butterfly system – A system used for venipuncture that includes a short needle attached to a length of tubing. The tubing is then attached to a syringe or directly to an evacuated tube for blood withdrawal.

C&S – Culture and sensitivity; includes the growth and identification of the bacterial microorganisms present in a specimen as well as the determination of the antibiotics best suited for treatment of the infection.

Calibration – Standardization of an instrument by comparison of the results produced against those of a known value, and appropriate adjustment of the instrument as needed to achieve optimal results.

Calibration verification – The verification of the success of a calibration process by analyzing quality control samples.

Calyx – A cup-like drainage vessel that collects urine as it is formed in the renal pyramids and transports the urine to the renal pelvis to be expelled from the kidneys.

Candle jar – A setup using a jar and a candle that encourages growth of bacteria that prefer less oxygen in their environment. A candle is placed in the jar with the culture plates and then lit. The jar is closed tightly, and the candle uses up the oxygen in the jar, after which it goes out.

Capillaries – The smallest blood vessels. Capillaries are microscopic in size and connect the arterial and venous systems.

Capillary action – A surface tension that attracts liquid to fill a capillary tube or elevates liquid when touched by a solid.

Capillary puncture – A method of blood collection during which an incision is made in through the skin, which passes through the epidermis into the dermal layer.

Capillary tubes – Small, hollow plastic or glass tubes used to perform microhematocrit testing.

Carbohydrate – An organic molecule that is used as an energy source by the body. Carbohydrates are broken down during digestion to provide glucose, which is easily used by the cells of the body.

Carcinogenic – Anything capable of causing cancer.

Cardiovascular – Pertaining to the heart, arteries, veins, and/or capillaries.

Carriers – Someone who is infected (or carrying) a pathogen in the body without symptoms. Sometimes these individuals may have already undergone the complete disease process but keep the causative agent in the blood, and sometimes they have never been symptomatic.

Casts – Protein "shells" formed in the glomerular tubules. Casts may be filled with white cells, red cells, granules, or other substances. All but hyaline casts may be clinically significant.

Cations – An ion that carries a positive charge. Electrolytes are ions when dissolved in water, and those with a positive charge (such as potassium and sodium) are cations.

CD4 cell – A type of lymphocyte (a white blood cell) that is attacked by HIV.

Cell mediated – A term used to describe all the activities of the T cells during the specific immune response to an antigen.

Centers for Disease Control and Prevention (CDC) – The premier public health agency in the United States. The CDC is part of the Department of Health and Human Services, and it protects the public health through research and implementation of recommendations for education, vaccination, infection control, and proper treatment of disease.

Centers for Medicare & Medicaid Services (CMS) – The organization responsible for enforcement of the CLIA regulations. Its members perform site visits and collect the fees associated with the registration process.

Centrifugal force – An outward force caused by the rotation of an object. The centrifugal force causes the specimen that is being centrifuged to separate into layers based on their relative weight or density.

Centrifuge – An instrument that spins specimens at a high speed to force the contents outward and separate the components as needed.

Cephalic vein – A vein in the antecubital area that is on the outer (lateral) side of this area when the arm is placed in the appropriate position for a venipuncture.

Cerebrospinal fluid (CSF) samples – Used to help diagnose various diseases of the central nervous system.

Chain of Custody – A form used to document all those who have possession of or process a legal specimen through all the steps of collection, testing, and disposal.

Chemotaxis – The process that signals more white blood cells to come to the scene of an invasion by a foreign substance.

Cholesterol – A soft, waxy lipid that is found in various parts of the human body, including the bloodstream.

Chromatographic – Analysis of substances on the basis of color changes that occur during the testing process.

Cilia – Slim, thread-like projections on the inner surface of mucous membranes that sweep away foreign objects such as pollen and dust.

Cirrhosis – A chronic liver disease characterized by scarring and ineffective function.

Clarity – The clearness of appearance. A term used to describe urine transparency.

Clean-catch midstream urine specimen – A procedure to be followed for urine specimen collection in which the potential for contamination with microorganisms from outside the urinary tract is minimized.

CLIA '88 – Clinical Laboratory Improvement Amendment of 1988.

CLIA-waived – Laboratory tests that have been determined by the FDA to fit this category. These tests are very simple to perform, need very little interpretation, and result in minimal clinical significance if they are performed or interpreted incorrectly.

Clinical chemistry – The quantitative or qualitative measurement of clinical compounds in the fluid portions of the body. This includes analysis of blood, urine, CSF, and other fluid specimens from the body.

Clinical significance – Meaningful results or actions affecting the clinical outcome or treatment of a patient.

Clot activators – Substances (such as silicone) that accelerate or aid in the clotting process within an evacuated tube.

Clotting cascade – The clotting process in which each step is dependent on the step prior until the clot has been formed.

Clotting factor – Enzymes and chemicals that interact in the clotting cascade to form a clot.

Coarse adjustment knob – The knob used on a microscope to bring an object into approximate focus.

Cocci – Round bacteria, which can then be classified further by their appearance when examined microscopically.

COLA – An independent company that accredits laboratories.

Colonies – Clusters of microorganisms visible on a culture.

Colorimetric – Analysis of the amount of a substance present in a sample, based on the amount of light absorbed by the substance. True colorimetric analysis often uses an instrument called a colorimeter to read the results. Colorimetric may sometimes be used synonymously with chromatographic.

Competency assessment – A process used to verify whether an employee is performing a testing procedure properly.

Competency testing – See Competency assessment.

Complement – A protein that assists in the second line of defense by destroying pathogens and sending signals to stimulate phagocytosis and inflammation.

Complete blood count (CBC) – A hematology test that includes hemoglobin concentration, hematocrit, red blood cell count, red blood cell indices, white blood cell count, leukocyte differential, and a platelet count.

Complexity – Level of intricacy, difficulty, or complication.

Compound microscope – A microscope that has two means of magnification that are "compounded" or multiplied for a higher level of magnification.

Comprehensive metabolic panel (CMP) – A group of 14 tests ordered together to provide a comprehensive analysis of overall patient health. This panel has been approved for reimbursement by the CMS.

Condenser – A lens on many microscopes that concentrates the light as it comes through an aperture.

Conjugated bilirubin – Bilirubin that has been changed by the liver to be water soluble. Conjugated bilirubin is also known as direct bilirubin.

Contagious – Anything that is capable of being transmitted from one person to another.

Contaminated sharps – Sharps (needles or other sharp objects such as lancets or glassware) that have been contaminated with blood or other potentially infectious materials. Essentially, any contaminated object that can penetrate the skin.

Contraindications – Any reason that a procedure would be inadvisable or inappropriate.

Coumadin – An anticoagulant medication used to prevent or treat blood clots. Coumadin is taken orally, and monitored with the PT/INR test.

Cover slip – A thin piece of glass or plastic that is placed on top of a liquid specimen when viewed under the microscope.

Creatine kinase (CK) – A by-product of muscle metabolism. May also be known as CPK.

Creatinine – A by-product of muscle metabolism that is cleared from the blood by the kidneys. Elevated levels of creatinine in the blood may be indicative of renal dysfunction.

Credentialed – Formal recognition of competence achieved by an examination. ASCP provides credentialing opportunities to laboratory professionals.

Critical results – Results that are very far outside the expected reference range for a specific test and indicate a critical or crisis condition for the patient. These results are considered to be something that must be acted on immediately by notifying the health-care provider who ordered the test. The critical result ranges are established by the practice (in the case of a physician office laboratory) or the pathologist in a larger laboratory.

Crystals – The precipitation of dissolved substances in the urine. Some of these may be clinically significant.

Culture – The growth of microorganisms in special medium that contains chemicals and/or nutrition to encourage and support their growth.

Current procedural terminology (CPT) codes – These codes are developed and monitored by the American Medical Association and are assigned to individual procedures to be used for reimbursement.

Cyanmethemoglobin – A colored compound used to measure hemoglobin. Hemoglobin is converted to this form so that the color change can be measured and quantitated.

Cylinder – A slim, round glass container used to measure liquids in the laboratory setting.

D-dimer – A breakdown product when clots are dissolved in the bloodstream. D-dimers are specifically formed when fibrin is broken down. It is is used as a diagnostic tool for different types of thrombosis.

Deep vein thrombosis (DVT) – Blood clots formed in the deep veins of the extremities, especially the legs.

Diabetes – A general term used to describe a group of diseases that demonstrate excessive urination. This is most commonly used to describe diabetes mellitus, which is diabetes with hyperglycemia.

Diabetes insipidus – Diabetes characterized by excessive urination resulting from the inability of the kidneys to concentrate urine.

Diabetes mellitus – Diabetes that includes high levels of blood glucose. Diabetes mellitus is caused either by a failure of the pancreas to produce insulin (type 1 diabetes) or by the resistance of the cells of the body to the presence of insulin (type 2 diabetes).

Diabetic ketoacidosis – A condition in which the blood pH is acidic, caused by an accumulation of ketones from an extended state of high blood glucose.

Diabetic neuropathy – Nerve damage caused by vascular and metabolic changes in those with diabetes mellitus. This may cause numbness, tingling, or pain, especially in the extremities. The loss of sensation in the feet may lead to serious damage.

Diapedesis – The movement of white blood cells out of small arterioles, venules, and capillaries as part of the inflammatory process.

Differential – A procedure whereby a blood sample is stained and WBCs are identified and counted by type.

Diluent – A liquid (such as water or saline) used to dilute another ingredient.

Diplococci – Round bacteria that appear in pairs when viewed under the microscope.

Disclosure – Sharing of the patient's information.

Disinfection – The application of a substance to materials and surfaces to destroy pathogens.

Disk-diffusion method – A procedure in which antibody-impregnated disks are used to test for antibiotic sensitivity on an agar plate.

Dissecting microscope – A microscope used to achieve a low level of magnification of three-dimensional objects, often those that are untreated and/or unstained.

Disseminated intravascular coagulation (DIC) – A disorder in which the coagulation process is stimulated throughout the body instead of in a localized area.

Distal convoluted tubule – A section of the renal tubule that is surrounded by capillaries in the nephron. The distal convoluted tubule is a twisted tubule that is distal to the Bowman's capsule.

Distal urethra – The part of the urethra that is closest to the outside of the body, or furthest from the bladder.

Diurnal variation – The change seen in a urinary or blood chemical at different times during a 24-hour day.

Documentation – Recording of pertinent information. This may be documentation of patient results, quality control results, or maintenance procedures, in a chart or on maintenance charts or logs.

Duress – The use of threats to accomplish a goal.

Dysfibrinogenemia – Any abnormality in the structure or function of the fibrinogen present in the bloodstream.

Ectoparasites – Parasites evident on the outside of their host.

Edema – A condition in which body tissues contain an excessive amount of fluid. An edema an be localized or systemic, and can result from many disease processes.

Efferent arteriole – The arteriole that transports blood away from the center of the glomerulus. The efferent arteriole is smaller in diameter than the afferent arteriole.

Efficacy – The capacity for producing a desired result or effect; a measurement of how effective a mechanism is toward reaching a desired goal.

Electrical impedance – A process for counting blood cells that depends on their resistance to the flow of an electrical current.

Electrolyte – A substance that is capable of conducting electricity when dissolved in water. In the human body, common electrolytes include sodium, potassium, and chloride.

Electron microscope – A specialized microscope that uses electrons to create a visual image with great detail.

Electronic quality control devices – Devices that are used to verify the accuracy of an instrument. These electronic devices should produce results that are within a narrow range every time they are tested on the instrument.

Emancipated – Not constrained or restricted; in the case of emancipated minors, this is a legal procedure that releases a child from the control of the parents, and releases any responsibility that the parents may have for the child.

Embolus – A mass of undissolved material traveling in the bloodstream. An embolus may be a blood clot that has broken away from its point of origin, or other foreign material such as a cluster of bacteria.

Endogenous cholesterol – Cholesterol that is created by the liver for use in the body.

Endoparasites – Parasitic infection within the host.

Endothelial cell lining – Flat epithelial cells that line the interior of the blood vessels.

Engineering controls – Devices that are used to limit or eliminate the potential for exposure to bloodborne pathogens in the workplace.

Enuresis – Inability to control the flow of urine.

Enzyme-linked immunosorbent assay (ELISA) testing – Also called EIA testing, ELISA testing is a common method of identifying the presence or absence of antibodies and/or antigens in a serum or plasma specimen.

Eosinophil – Granulocytic white blood cells with a lobed nucleus and cytoplasmic granules that stain red/orange with Wright's stain. Eosinophils contribute to the destruction of parasites and to allergic reactions by releasing chemical mediators.

Epidemiology – The study of the spread of disease.

Epithelial cells – Cells from the epithelial lining of the urinary tract. These may be squamous, renal tubular epithelial cells, or transitional epithelial cells. They are classified according to their point of origin.

Epstein-Barr virus (EBV) – The virus that causes infectious mononucleosis.

Erythroblastosis fetalis – A severe hemolytic disease of the newborn characterized by anemia and jaundice.

Erythrocyte – Mature red blood cell.

Erythrocyte sedimentation rate (ESR) – A nonspecific laboratory test of the speed at which erythrocytes settle out of anticoagulated blood.

Erythroid – Pertaining to erythrocytes.

Erythropoiesis – The process that results in the formation of red blood cells.

Erythropoietin – A protein made by the kidneys that stimulates the proliferation of red blood cells.

Ethics – Moral standards and behavioral codes that dictate the behavior of an individual.

Etiology – The cause of a disease.

Eustachian tube – A flexible hollow tube that extends from the middle ear to the nasopharynx.

Evacuated tube system – A system used to draw blood that involves the use of a multipoint needle, a needle holder, and an evacuated tube.

Exogenous cholesterol – Cholesterol that is introduced to the body through ingestion of meat and dairy.

Exposure control plan – A plan that must be created by employers to educate and help protect employees from bloodborne pathogen exposure. This plan includes education, vaccination, safety procedures, and follow-up in case of exposure.

Expressed consent – Consent for a behavior that is expressed fully and personally by the patient; it may be an oral or written expression.

External quality control specimens – Quality control specimens that are purchased from an outside source, or those that are included in a test kit, but are separate from the testing system. These are usually liquid controls.

Extracellular – Outside the cells. Used to describe substances outside the cells of the body, in the plasma or interstitial fluid.

Extrinsic pathway – Part of the blood-clotting cascade that includes chemicals that are secreted into the bloodstream in response to vessel injury.

Exudate – A fluid that is released from the human body. Exudates are classified according to the solid substances present in the drainage, or the level of protein present.

Eyepiece – The part of the microscope for viewing.

Fasting – A period of at least 12 hours during which a patient forgoes food or drink (other than water). Generally done in preparation for a blood specimen to be drawn.

Fasting blood sugar (FBS) – The glucose level performed in a laboratory on a fasting blood specimen.

Fasting plasma glucose (FPG) – The level of glucose tested in the plasma (liquid portion of the blood) after a patient has fasted for at least 8 hours.

Fasting urine specimen – Urine collected during the second morning void. This urine will indicate the state of the body after a prolonged fast.

Fecal occult blood testing (FOBT) – Testing that identifies the presence of occult (hidden) blood in the stool. This test screens for the early stages of colorectal cancer, as well as additional sources of bleeding.

Fecal occult blood – Blood, which cannot be seen, that is present in the feces of an individual.

Fecal-oral route – A method of transmitting infective microorganisms. The pathogen is shed in the feces of the infected individual and then transferred to the skin. It is then spread to food, water, inanimate objects, or other people.

Fibrates – A class of medication used to treat elevated triglyceride levels. Fibrates may also decrease the LDL levels, but to a lesser extent.

Fibrin degradation products (FDPs) – Substances present in the bloodstream when clot dissolution occurs in the body. Fibrin degradation products include the by-products of fibrinogen as well as fibrin that may be involved in the blood clots.

Fibrinogen – A protein that is created in the liver and is present in blood plasma. Fibrinogen is converted into fibrin by the action of ionized calcium.

Fibrinolysis – The breakdown of fibrin in blood clots, allowing them to dissolve. Fibrinolysis occurs spontaneously in the body, and can also be accomplished by use of medications.

Fibrinolytic drugs – Also called clot busters.

Filtration – The process of removing large particles from a solution by allowing the liquid to pass through a semipermeable membrane.

Fine-adjustment knob – The focusing knob used on the microscope to provide fine, detailed focus at higher levels of magnification.

First morning void – A urine specimen collected right after waking from sleep. This specimen usually is more concentrated, and is the recommended collection method for some tests.

Flanges – Flared edges at the bottom of the evacuated tube holder that allow for leverage when inserting or removing a tube.

Flask – A glass container used to mix or store chemicals. Flasks are generally smaller at the top than they are at the bottom.

Fomites – A substance or surface that adheres to and transmits infectious microorganisms.

Formed elements (blood) – The cellular components of blood: erythrocytes, leukocytes, thrombocytes.

Formed elements (urine) – Substances in the urine specimen that may be viewed microscopically, such as cells, bacteria, and fungal elements.

Fraud – A dishonest or deceitful act that is performed with the intent to hide the truth.

Fungal elements – Fungus microorganisms present in the urine specimen. The most common are yeast.

Fungi – A type of microorganism that causes disease in humans. This type of infection is called a mycotic infection. Examples of fungal infections in humans are yeast infections, athlete's foot, and thrush.

Gammaglobulin (also known as immune globulin) – A fluid that contains active antibodies against a particular antigen that may be given to offer short-term, immediate protection to the recipient.

Gauge – A term used to describe the diameter of the hollow interior portion of a needle. Gauge number is inversely related to the size of the needle.

Gestational diabetes – Hyperglycemia detected during pregnancy in women who did not have diabetes before becoming pregnant.

Glass slides – A piece of finely ground glass that is uniform in shape. Specimens are applied to the glass slide to be viewed under the microscope. Some specimens require staining or some sort of treatment before viewing, whereas others are viewed as they appear naturally.

Globulin – One of the group of plasma proteins that controls colloidal osmotic pressure (oncotic pressure) within the capillaries, participates in the immune response, and binds with substances to transport them in blood.

Glomerular filtration – The filtration process that occurs in the glomerulus. Because the afferent arteriole is much larger in diameter than the efferent arteriole, pressure builds up in the glomerulus, forcing filtrate and small molecules from the capillary bed while retaining large molecules in the bloodstream.

Glomerulus – A group of twisted capillaries encircled by the Bowman's capsule in the nephron of the kidney. The hydrostatic pressure in the glomerulus forces out the initial filtrate from the blood, which eventually becomes urine.

Glucagon – A hormone created by the islets of Langerhans in the pancreas that offsets the effects of insulin on the cells of the body.

Glucose – A type of sugar in the bloodstream that results from digestion of carbohydrates. Glucose is the most commonly used energy source in the body.

Glucose challenge – A process in which a patient is given a glucose drink that contains 50 g of glucose to ingest. The blood is drawn for a glucose test 1 hour after the glucose drink has been ingested.

Glucose tolerance test – A test that screens for diabetes (or occasionally hypoglycemia), which involves serial blood draws following ingestion of a drink containing glucose. Sometimes known as an oral glucose tolerance test, or OGTT.

Glucose tolerance urine specimen – A specimen collected as part of a glucose tolerance test. These are tested specifically for the presence of glucose and sometimes for ketones.

Glucosuria – The presence of glucose in the urine.

Glycated hemoglobin – Hemoglobin that is irreversibly changed with the addition of glucose to the molecule. This occurs when there have been elevated blood glucose levels for an extended period of time.

Glycemic control – A term used to describe the balance or control of blood glucose levels. Good glycemic control means that the patient does not have extremely high or low levels of blood glucose.

Glycogen – A storage form of glucose when there are excessive amounts present in the bloodstream. Glycogen is stored in the liver and some of the cells of the body, to be used for energy when glucose levels are low.

Glycogenolysis – The breakdown of glycogen to allow for the release of glucose when needed. This often occurs during periods of fasting, or when there is an elevated need for energy, such as during exercise.

Glycolysis – The first steps of the use of glucose by the cells of the body; a series of reactions that convert glucose to pyruvic acid and energy.

Glycosuria – The presence of sugar in the urine. This may be glucose, pentose, galactose, or fructose. Further testing is necessary to identify the type of sugar present in the specimen.

Glycosylated hemoglobin (HbgA1c) – Hemoglobin that has been changed by the addition of glucose molecules to the protein. The measurement of the percentage of glycosylated hemoglobin in the blood provides an indication of the glucose levels present in the bloodstream over the past 3 months. A subtype of hemoglobin (Hb A) is glycated when the blood glucose is elevated for an extended period of time, as in the case of uncontrolled diabetes mellitus.

Graduated cylinder – A cylinder that is marked in different quantities for measuring liquids.

Gram stain – A staining method that allows the classification of bacteria dependent on the color of the stain that they absorb. Gram-positive bacteria retain the crystal violet dye used in the process, and appear purple. Gram-negative bacteria do not keep the crystal violet dye in their cell membranes, and instead absorb the color from the safranin counterstain. They appear pink when viewed under the microscope.

Granular – A classification used to describe white blood cells that contain granules in their cytoplasm.

Granulocyte – White blood cells that have names ending in –phil (neutrophil, eosinophil, and basophil) and have distinct granules visible in their cytoplasm.

Granulocyte-macrophage colony-stimulating factor (GM-CSF) – A glycoprotein that stimulates the production and functional activity of neutrophils, monocytes, and macrophages.

Guaiac – A type of tree resin used for fecal occult blood testing. Guaiac interacts with the heme in the hemoglobin molecule present in red blood cells. If there is blood present, the interaction makes the guaiac paper turn blue when hydrogen peroxide developer solution is added.

Hanging drop slide – A slide with a depression in the center. This allows a drop of fluid containing the specimen to be suspended from a cover slip over the depression for observation microscopically.

Hazard communication standard – A document created to educate and protect employees from the chemicals in use at their workplace.

Hb A – The normal adult hemoglobin, composed of two alpha and two beta chains.

Hb A1c; HgbA1c – See Glycosylated hemoglobin (HgbA1c).

Health Information Portability and Accountability Act (HIPAA) – A federal law that affects the security, standardization, and confidentiality of health-care information. It also affects the rights of those who have insurance coverage when they change jobs.

Health care–associated infection – An infection contracted by an individual at a health-care facility. Refers specifically to an infection that the individual did not come into contact with in any other way.

Helicobacter pylori (H. pylori) – A type of bacterium that causes gastric and duodenal ulcers.

Hematocrit (Hct) – The volume of erythrocytes packed by centrifugation in a given volume of blood.

Hematology – The study of blood and blood-forming tissues.

Hematoma – A mass of blood caused by damage to a blood vessel.

Hematopoiesis – Production and development of blood cells, normally in the bone marrow.

Hematuria – The presence of blood in a urine specimen.

Hemoconcentration – A condition resulting from a tourniquet that is too tight or is left on too long, causing a relative increase in the cells in the antecubital area and possibly resulting in erroneous test results.

Hemocytometer – A counting chamber for determining the number of cells in a stated volume of blood.

Hemoglobin – The protein in red blood cells that carries oxygen.

Hemoglobin (Hgb) – The iron-containing pigment of red blood cells that carries oxygen from the lungs to the tissues.

Hemoglobin and hematocrit (H&H) – Hemoglobin and hematocrit testing is often ordered together as H&H.

Hemoglobin electrophoresis – The movement of hemoglobin on a medium as a result of changes in electrical potential, used to determine type of hemoglobin present.

Hemoglobinopathies – Any one of a group of genetic diseases caused by or associated with the presence of one or several forms of abnormal hemoglobin in the blood.

Hemoglobinuria – Hemoglobin in the urine.

Hemolysis – The destruction (lysis) of the red blood cell membrane with the release of the fluid contents. A visible reddish tint becomes evident in the plasma if there is marked hemolysis in the specimen.

Hemolytic disease of the newborn (HDN) – A neonatal disease, usually the result of an Rh incompatibility in which the antibodies of the mother cross the placenta and "attack" the neonate. May cause neonatal anemia, jaundice, edema, and enlargement of the liver and spleen.

Hemolytic uremic syndrome – An abnormal reaction to infection with the *E. coli* strain O157:H7. The *E. coli* bacteria secrete a toxin that causes an inflammatory process, resulting in blood clot formation in the capillary vessels of the body. As the red blood cells of the body pass through these clogged capillaries, it causes them to become damaged, which consequently causes damage to the kidneys as they attempt to process these damaged cells. The patient must be treated for the anemia and renal failure immediately.

Hemophilia – Blood-clotting disorders in which the patients have deficiencies of specific clotting factors.

Hemostasis – A stoppage of bleeding or of circulation.

Heparin – A naturally occurring anticoagulant in the human body. Heparin is also provided as a fast-acting anticoagulant, administered via IV or subcutaneously.

Hepatic function panel – A group of tests that are ordered together to monitor liver function.

Hepatitis – Inflammation of the liver. Hepatitis may be caused by a viral infection or other disease state that causes the liver to be inflamed.

Hepatitis A virus (HAV) – A type of virus that causes inflammation of the liver in infected patients. Hepatitis A is transferred via the fecal-oral route, and usually does not cause permanent liver damage. Vaccinations are available for this disease.

Hepatitis B virus (HBV) – A type of virus that causes inflammation of the liver in infected patients. Hepatitis B is bloodborne, and is transferred through sexual contact, contaminated needles, mucous membrane exposure, or other direct blood-to-blood contact. Hepatitis B infection may lead to permanent liver damage. Hepatitis B is the most common bloodborne pathogen to be transmitted to health-care workers. There are vaccinations available for this disease.

Hepatitis C virus (HCV) – A bloodborne virus that causes inflammation of the liver. Hepatitis C infection usually leads to permanent liver damage, and is often asymptomatic until the liver damage has become quite advanced. There is no vaccination available for this disease.

Hepatitis D – A type of virus that causes liver inflammation. This is considered a genetic mutation or variant of the hepatitis B virus, as it is present only in those who are already infected with hepatitis B. There is no vaccination for this disease.

Heterophile antibody – A nonspecific type of antibody that is present in those with infectious mononucleosis. Heterophile antibodies are also present in other inflammatory situations.

Heterozygous – A genetic term in which the two genes for one characteristic are not the same.

High complexity – Laboratory tests that have been determined by the FDA to be complex enough so that the procedures can be performed only by highly trained laboratory professionals.

High-density lipoprotein (HDL) – A lipid subgroup; this is known as the “good” cholesterol because it helps to reduce the plaque buildup on the interior of the blood vessels.

Hilum – The indented (concave) portion of the kidney.

Histamine – A substance produced by the body in response to antigenic exposure that causes dilation of the blood vessels, increased secretion of acid in the stomach, smooth muscle constriction, mucus production, tissue swelling, and itching (during allergic reactions).

Histogram – A graph of the distribution of cell sizes made during cell counting on an automated analyzer.

Homeostasis – The internal balance of the body that is necessary for good health. Homeostasis includes minor adjusting to outside stimuli.

Homozygous – A genetic term in which the two genes for one characteristic are the same.

Hospital laboratories – Laboratories that are usually housed within a hospital and provide services primarily to patients of that institution. Hospital laboratories may be owned by the hospital or contracted to provide their services.

Human chorionic gonadotropin (HCG) – A hormone that is secreted by the trophoblasts of a fertilized ovum early in pregnancy. This hormone maintains the corpeus luteum so that it will secrete estrogen and progesterone. The presence of HCG in blood and urine is indicative of pregnancy.

Human immunodeficiency virus (HIV) – This is a retrovirus that invades the CD4 cells of humans and causes the failure of their immune systems. Infection leads to the eventual development of AIDS. It is a bloodborne pathogen, and may be transmitted through sexual contact, mucous membrane exposure, or parenteral exposure.

Hyperchromic – Pertaining to excess hemoglobin concentration in the erythrocyte.

Hyperglycemia – Blood glucose levels elevated above acceptable reference ranges.

Hyperkalemia – Elevated levels of potassium in the blood plasma.

Hyperlipidemia – Excessive levels of lipid in the bloodstream.

Hypernatremia – Elevated levels of sodium in the blood plasma.

Hyperthyroidism – A condition in which the thyroid gland is overactive.

Hypochromic – A condition of the blood in which the erythrocytes have a reduced hemoglobin content and appear pale in color.

Hypofibrinogenemia – Low levels of circulating fibrinogen in the bloodstream.

Hypoglycemia – Decreased blood glucose levels below acceptable reference ranges.

Hypokalemia – Decreased levels of potassium in the blood plasma.

Hyponatremia – Decreased levels of sodium in the blood plasma.

Hypothyroidism – A condition in which the thyroid gland is underactive.

Icteric – A term used to describe the serum or plasma of a patient with hepatic dysfunction. An icteric specimen has a yellowish green color.

iFOB – An acronym used to describe an immunochemical method of testing for fecal occult blood. It uses an antigen-antibody interaction with the globin portion of the hemoglobin molecule and is an alternative testing procedure to the guaiac method for the presence of fecal occult blood.

Immunity – Protection from diseases.

Immunoglobulins – Plasma proteins that play a role in protection from disease. These may occur on the surface of the lymphocytes, or may be secreted in response to antigenic exposure.

Immunology – The study of the components of the immune system and their function.

Impaired fasting glucose (IFG) – A term used to describe prediabetics who exhibit elevated fasting plasma glucose levels.

Impaired glucose tolerance (IGT) – A term used to describe prediabetics who exhibit elevated glucose levels after drinking a glucose solution.

Implied consent – Consent for a procedure that is implied by the actions of the individual.

International Classification of Diseases, 9th ed. (ICD-9) – Codes assigned by the CDC to symptoms or disease states. These codes must be included for reimbursement for laboratory procedures.

In vitro – Latin for "in glass," referring to testing done outside of the body.

In vivo – Latin for "in the living body," referring to something occurring inside the body, such as antigen-antibody reactions within a living subject.

Incubators – A device by which the temperature, and sometimes humidity, of the environment can be controlled. Incubators are often used to help bacterial specimens grow for testing or to assist with the completion of a testing procedure.

Indwelling catheter – A plastic tube that is allowed to remain in the urinary bladder for a period of time to facilitate drainage of urine as it is formed.

Infection – A disease state caused by pathogenic microorganisms, including bacteria, viruses, parasites, or fungi. Infections may be localized or may spread throughout the body. Symptoms often include pain, fever, redness and swelling, and loss of function.

Infection control – Policies and procedures to help minimize the spread of infection and to analyze data in relation to infection within a health-care facility.

Infectious – A term used to describe a subject who is capable of transmitting disease to others. The transmission may occur directly or indirectly.

Inflammation – A immunological defensive mechanism of the body, including increased blood flow, tissue swelling, increased white blood cell presence, and release of chemical toxins.

Informed consent – Approval that is obtained in writing after the patient has been informed about the specifics of the proposed procedure, the risks of the procedure, alternative treatments, and information about what may happen if the patient decides against the procedure.

Inoculation – Addition of a specimen to a culture medium for growth.

Insulin – A hormone produced by the islets of Langerhans of the pancreas. Insulin is necessary for the cells of the body to absorb glucose to be utilized as energy.

Insulin dependent – Diabetes that requires injections of insulin to control blood glucose levels. Patients with type 1 diabetes are usually insulin dependent.

Insulin resistance – A condition in which cells have a diminished ability to interact appropriately with insulin.

Intentional tort – A wrongful act that is committed by one person against another person deliberately.

Interferons – A group of proteins produced in the body having antiviral properties.

Interleukin – A hematopoietic growth factor affecting growth and differentiation of lymphocytes.

Intermittent (straight) catheter – A catheter (hollow, plastic tube) placed directly into the bladder to facilitate drainage. The catheter is removed as soon as the bladder has drained.

Internal quality control – Quality control that is built into the testing kit, provided by the manufacturer.

International normalized ratio (INR) – A mathematical calculation used to standardize treatment based on prothrombin time (PT) results. PT results may vary, depending on lot numbers of reagents and different instruments used for testing. The INR is a calculation that takes these variables into account and provides a tool for monitoring those patients on anticoagulant therapy. Most patients on anticoagulant therapy should have an INR between 2.0 and 3.0.

Interstitial – Between the cells.

Interstitial fluid – The fluid between cells.

Intracellular – Within the cells.

Intravascular lysis – The destruction of red blood cells within the vessels of the body.

Intrinsic factor – A glycoprotein secreted by the parietal cells of the gastric mucosa, necessary for the absorption of vitamin B_{12}.

Intrinsic pathway – Part of the blood-clotting cascade that uses chemicals that are normally present in the bloodstream.

Ions – An atom or group of atoms that has lost or gained electrons so that it possesses an electrical charge. Electrolytes are substances that become ions when dissolved in water.

Iris diaphragm – A device with a variable diameter that controls the amount of light that is used to view a specimen through a microscope.

Iron-deficiency anemia – Anemia resulting from a greater demand on stored iron than can be supplied.

Ischemia – An insufficient supply of blood to an organ or body part. This term is often used to describe a lack of blood flow to areas of the heart when the patient has atherosclerosis, causing partial blockage of the vessels.

Islets of Langerhans – Specialized groups of pancreatic cells that produce insulin and glucagon. Insulin and glucagon are necessary for controlling the levels of blood glucose in the body.

Jaundice – A yellow staining of body tissues and the fluids of the body as a result of excess bilirubin in the bloodstream in patients with hepatic dysfunction.

Ketones – A chemical that is elevated in the bloodstream with increased fat metabolism. This is usually associated with dysfunctional carbohydrate utilization, as in the case of diabetes mellitus.

Ketonuria – The presence of ketones in the urine.

Kidneys – Organs that create urine.

Kirby-Bauer method – A method used to test specific bacteria for antibiotic sensitivity, utilizing antibody-impregnated disks on special agar plates.

KOH (potassium hydroxide) – A solution used for examination of a sample under the microscope which is performed immediately after collection to search for the presence of fungi.

Labile – Describes a substance or sample that is not fixed, and is easily changeable. These types of specimens generally must be tested immediately to provide accurate results.

Laboratory – Any site where specimens may be collected, processed, or tested.

Laboratory directory – A reference document in which specimen requirements and other useful information might be found. It includes all the tests offered by a particular laboratory.

Laboratory report – A document used to transmit laboratory results and reference ranges.

Laboratory requisition – A form used to order laboratory tests.

Laboratory thermometer – A thermometer used to measure temperatures in a controlled environment, such as a refrigerator, freezer, or incubator.

Lateral flow immunoassay – A test used to determine the presence of an antigen or antibody in a specimen. In this type of test, the sample moves laterally across the testing surface, resulting in a color change if the antibody and/or antigen is present in the specimen. This testing method is used in many CLIA-waived tests.

Laws – Regulations established by a governmental authority that are enforced and punishable for noncompliance.

Leukemia – A class of hematological malignancies in which immortal clones of immature blood cells multiply at the expense of normal blood cells.

Leukocyte – White blood cell.

Leukocyte esterase – An enzyme secreted by granulated white blood cells and monocytes.

Leukocytosis – Abnormal increase in white blood cell (leukocyte) count.

Leukopenia – Abnormal decrease in white blood cell (leukocyte) count.

Liability – Legally responsible for an action.

Libel – A false statement in print.

Light source – The origin of light or illumination for a specimen to be viewed on a microscope. For most compound microscopes, this is a light bulb located in the base of the unit.

Lipemia – The presence of excessive lipids in the blood. This term is used to describe the "milky" appearance of the plasma or serum in specimens with elevated lipid content.

Lipid – A term used to describe fats in the body. Cholesterol and triglycerides are types of lipid molecules.

Lipoproteins – Simple proteins bound to lipids or fats in the bloodstream. Lipoproteins are the method by which cholesterol travels in the bloodstream; they are further differentiated as HDL, LDL. or VLDL.

Lobulated – Having lobes or sections.

Loop of Henle – The middle portion of the renal tubules where exchanges occur as part of the formation of urine from the renal filtrate.

Low-density lipoprotein (LDL) – A subclass of cholesterol present in the body; this is known as the "bad" cholesterol because it contributes to plaque buildup on the blood vessel walls.

Lumbar puncture – A procedure used to obtain samples of cerebrospinal fluid for examination and testing.

Lumen – The term used to describe the interior hollow portion of the needle. This term may be used to describe the inside of any hollow object, such as a vein or artery.

Lymphocyte – A white blood cell responsible for much of the body's immune protection.

Lymphoid – Pertaining to lymphocytes.

Lysed – The action of cell destruction (lysis) by chemical or mechanical methods, releasing the contents of the cells.

Lysozyme – An enzyme found in various cells and body fluids that damages bacterial cell walls.

Lytes – See Electrolytes.

Macrocyte – Cells larger than normal size.

Macrocytic – Abnormally large erythrocytes exceeding 10 μ in diameter.

Macrophage – Cells that are capable of phagocytosis.

Macroscopic – A description of a structure that is large enough to be seen or examined without magnification.

Macrosomia – An abnormally large baby. An infant who has a birth weight above the 90th percentile of the standardized growth curve exhibits macrosomia. This may be caused by gestational diabetes in the mother.

Malabsorption – Inadequate or incomplete absorption of nutrients from the digestive tract.

Malfeasance – Performance of a wrong and unlawful act.

Malpractice – The negligence of a professional person.

Mastectomy – The removal of breast tissue.

Material Safety Data Sheets (MSDS) – A descriptive sheet required by law to accompany chemicals upon delivery. These sheets contain essential safety information about the chemical, as well as contact information for the manufacturer.

Mean – The mathematical average of a set of results.

Mean corpuscular hemoglobin (MCH) – The average weight of hemoglobin in red blood cells in a sample.

Mean corpuscular hemoglobin concentration (MCHC) – The average weight of hemoglobin in a given volume of packed blood cells.

Mean corpuscular volume (MCV) – The average volume of a red blood cell in a sample.

Media – Plural of *medium;* a term used to describe the substance that supports growth of microorganisms.

Median cubital vein – The middle vein of the antecubital area; this is the first vein of choice for venipuncture procedures.

Medical asepsis – A state of cleanliness and disinfection that has reduced the number of microorganisms and pathogens to a significant degree.

Megakaryocyte – A large bone marrow cell with large or multiple nuclei from which platelets are derived.

Meniscus – A curve in the surface of a liquid in a cylinder caused by surface tension between molecules.

Metabolic syndrome – The presence of four interrelated risk factors for atherosclerosis: insulin resistance (hyperglycemia), hypertension, hyperlipidemia, and obesity.

Methicillin-resistant *Staphylococcus aureus* (MRSA) – *Staphylococcus aureus* organisms that show resistance to the methicillin family of antibiotics. This strain is a growing concern to the health-care profession.

MIC – See Minimum inhibitory concentration.

Microalbumin – Small amounts of albumin.

Microalbuminuria – Small amounts of albumin in the urine. This may be an indication of poor blood glucose control and kidney damage because of high glucose levels.

Microcollection containers – Small blood collection containers designed to be used for small amounts of blood. Often used for capillary draws, or situations in which the venipuncture is very difficult.

Microcytes – A small erythrocyte with a diameter of less than 5 μ.

Microcytic – Small erythrocytes less than 5 μ in diameter.

Micrometer (micron, μ) – One millionth of a meter, or 10^{-6}.

Microorganisms – Living organisms that are too small to be seen without the aid of a microscope.

Microscopic – A description of a structure that cannot be seen or examined without the use of magnification.

Micturition – Urination or the voiding of urine.

Minimum inhibitory concentration (MIC) – A term used to describe the lowest concentration of an antibiotic that will inhibit bacterial growth. This term is used to describe a procedure using a plate of wells that have specific antibiotics at known concentrations.

Misfeasance – Performance of a lawful act in an improper manner.

Moderate complexity – Laboratory tests that have been determined by the FDA to be moderately complex in nature. These tests require more quality control procedures, proficiency testing, and comprehensive documented personnel training than the CLIA-waived tests.

Monocular – A term used to describe a microscope that has only one ocular for viewing.

Monocyte – A mononuclear phagocytic white blood cell derived from myeloid stem cells.

Mononucleosis – An infectious disease caused by the Epstein-Barr virus.

Morphology – Description of the structure and form; term used to describe the shape of cells in a complete blood count.

Motility – Spontaneous movement; a term often used when describing microorganisms viewed in their natural state without preservation or staining.

MRSA – Acronym for methicillin-resistant *Staphylococcus aureus*. (See also Methicillin-resistant *Staphylococcus aureus*.)

Mucous membranes – Membranes lining body cavities that open to the outside of the body.

Mucus – A viscous fluid secreted by the epithelial cells.

Multiple myeloma – A malignant condition in which the bone marrow is overcome by nonfunctional cancer cells.

Multisample needle – A needle used with the evacuated tube system having a point at both ends.

Mycoses – Fungal infections.

Mycotic – A term used to describe fungal infections.

Myeloid – Pertaining to the precursor cell for granulocytes, monocytes, erythrocytes, and megakaryocytes.

Myocardial infarction – A blockage of the coronary blood vessels, resulting in damage to and death of the heart muscle.

Myoglobin – A protein that stores and uses oxygen in cardiac and skeletal muscle tissue. The blood myoglobin concentration will become elevated with damage to the heart or the skeletal muscle of the body.

Myoglobinuria – The presence of myoglobin in the urine.

Nasopharyngeal collection technique – A sample taken from the portion of the pharynx just above the soft palate. This specimen is obtained by insertion of a collection swab through the nasal cavity to the back of the throat.

Nasopharynx – The portion of the pharynx located directly behind the nose, just above the soft palate.

National Fire Protection Association (NFPA) – A nonprofit organization established in 1896 whose mission is to use education, codes and standards, research, and training to reduce fire and other hazards. This organization is known worldwide for its fire-prevention efforts.

Natural killer cells – Large lymphocytes that bind to invading cells and inject toxins, causing cell lysis.

Negligence – Occurs when a professional does not act with reasonable care.

Nephrons – The functional unit of the kidney where urine is formed.

Neutrophil – A granulocytic white blood cell responsible for much of the body's protection against infection.

Niacin – A B vitamin used to treat elevated triglyceride and cholesterol levels.

Nitrite – Nitrites are important components of the blood because they dilate blood vessels, reduce blood pressure, and perform other functions.

Nocturia – Excessive or frequent urination at night.

Nonintact skin – Skin that is broken. This may be caused by a scratch or puncture, but it may also be caused by irritation, such as develops with use of gloves and frequent hand washing.

Nonfeasance – The failure to perform a necessary medical act.

Normal flora – Resident bacteria normally present in specific areas of the body. This bacterial presence is not pathogenic.

Normal ranges – A range of results that are expected for a specific test within a certain population. Also referred to as reference ranges.

Normochromic – A red blood cell with sufficient hemoglobin.

Normocytic – A red blood cell of normal size, which is of 6 to 8 μ.

Nosocomial infection – An infection contracted in a hospital or other inpatient care facility.

O&P – Ova and parasites.

Objective lenses – The lenses of the microscope that are closest to the specimen. Compound microscopes have various levels of magnification available by using different lenses.

Occlude – To close up or obstruct, such as when a blood vessel has been blocked by a clot or plaque buildup.

Occupational exposure – Exposure that occurs while performing duties in the workplace. Occupational exposure does not always mean that there was an error in the way a task was performed; some types of exposure are inherent to all careers. For laboratory and healthcare personnel, occupational exposure to bloodborne pathogens is commonplace.

Occupational Safety and Health Administration (OSHA) – A governmental agency within the Department of Labor. OSHA's mission is to ensure the health and safety of all American workers.

Ocular lenses – The lens that is closest to the eyepiece. These lenses provide a low level of magnification that is not adjustable.

Oil immersion lens – The most powerful objective lens (usually 100X magnification) that is used with a drop of oil between the lens and the slide. This allows for the light to be focused fully on the specimen without an opportunity for scatter.

Oliguria – Diminished urinary output of less than 400 mL per day.

Oncologist – A physician who specializes in the field of tumors and/or cancer treatment.

Opportunistic pathogens – Infections that usually do not cause disease in a healthy individual. When the immune system of an individual is compromised, or when there is an imbalance of some sort present, a pathogen may take advantage of the situation and cause an opportunistic infection.

Oral glucose tolerance test (OGTT) – A screening test for diabetes mellitus in which the patient is given a standardized glucose solution to drink, then has blood glucose levels checked at hourly intervals to monitor the response to the ingestion.

Other potentially infectious materials (OPIMs) – This category includes semen, vaginal secretions, cerebrospinal fluid, synovial fluid, pleural fluid, pericardial fluid, peritoneal fluid, amniotic fluid, saliva during dental procedures, breast milk, and any body fluids that are visibly contaminated with blood.

Otitis – Inflammation of the ear.

Otitis externa – Inflammation or infection of the outer ear.

Otitis media – Inflammation or infection of the middle ear.

Ova and parasite (O&P) examination – An examination of a stool sample for the presence of parasitic microorganisms or their eggs (ova).

Ovalocyte – An elliptical red blood cell.

Packed cell volume – A synonym for the hematocrit, the volume of erythrocytes packed by centrifugation in a given volume of blood.

Palpation – Process of examination by the application of the hands or fingers; assessing an area of the body by use of touch.

Panels – Groups of laboratory tests that are ordered together. These are often organ or disease specific. Also known as profiles.

Panic results – See Critical results.

Parasites – Any organism that lives on or within another living organism, often at the expense of the host organism.

Parenteral exposure – Exposure that occurred with a piercing of the skin.

Partial thromboplastin time (PTT) – Measurement of the efficacy of the intrinsic coagulation pathway. This pathway includes the actions of clotting factors VIII, IX, XI, and XII. The PTT is measured in seconds.

Pathogen – A microorganism that is capable of causing disease.

Pathological thrombosis – Formation of thrombi unnecessarily or excessively, not only in response to vessel injury.

Patient identifiers – Personal information that may identify a patient seeking care including demographics, Social Security numbers, birth dates, and the like.

Peak collection – A blood specimen drawn at the time when a medication is at its highest level in the bloodstream.

Percutaneous – Through the skin.

Peritoneum – The membrane that lines the abdominal cavity.

Pernicious anemia – A chronic macrocytic anemia in which the stomach secretes insufficient intrinsic factor for B_{12} absorption.

Personal protective equipment (PPE) – Equipment that is worn as protection for the employee, including goggles, masks, respirators, laboratory coats, or face shields.

Petri dish – A small, round dish with a flat bottom and a lid. Agar is often contained in a Petri dish for culture procedures. Petri dishes may be made of glass or plastic.

Phagocytes – A cell that has the ability to ingest and destroy antigens.

Phagocytosis (phagocytic; phagocytize) – A three-stage process by which phagocytes (certain white blood cells) engulf and destroy microorganisms, antigenic substances, and cellular debris.

Phlebotomist – A person who draws blood from a vein, capillary, or artery.

Physician office laboratories (POL) – Laboratories that are located within a physician's office that perform testing on the specimens collected from the patients of that practice.

Physiology – The study of the function of the human body.

Pinworm – Infection caused by the parasite *Enterobius vermicularis,* which is the most common type of worm infection in the United States.

Pipette – A device used in the laboratory to transfer liquid from one place to another.

Pisse prophets – Nonmedical practitioners who used urine to diagnose and predict disease without scientific basis for the practice.

Plaque – Deposits of cholesterol within blood vessels.

Plasma – The liquid portion of blood within the body. This term is also used to describe the liquid portion of a specimen when it has been drawn into a tube containing anticoagulant.

Plasma separator tubes (PST) – Tubes containing a thixotropic separating gel to form a barrier between the blood cells and the liquid portion of the blood after centrifugation.

Plasmin – An enzyme involved in the breakdown of blood clots.

Platelet – Formed element in the blood that aids in blood clotting.

Platelet count – The quantity of platelets (thrombocytes) present in a blood sample. Adequate numbers are essential for effective coagulation.

Pluripotent – Having the ability to differentiate into several types of specialized cells, a quality that stem cells are said to have.

Poikilocytosis – Variations in the shape of red blood cells.

Point-of-care testing (POCT) – Testing that is performed at the bedside or chairside of a patient, with results available within minutes.

Polychromatic stain – A staining process consisting of both an alkaline stain and an acidic stain that allows for cells to stain a variety of colors.

Polycythemia – Having an excess of red blood cells.

Polydipsia – Excessive thirst.

Polymorphonuclear – Possessing a nucleus consisting of several parts or lobes connected by fine strands.

Polyphagia – Excessive hunger.

Polyuria – Excessive urine production.

Postexposure prophylaxis (PEP) – Protective measures taken immediately after exposure to a chemical or other type of hazard. This may include medication, but may also include other types of precautions taken “just in case” after a chemical or biohazardous exposure.

Postprandial – After eating.

Precision – A statistical term that is used to describe the reproducibility of the results for a test to each other.

Prediabetes – A condition in which a patient shows early evidence of impaired carbohydrate metabolism or impaired secretion of insulin.

Privacy Rule – An expansion of the original HIPAA mandate that applies to confidentiality of health information, and the rights of patients to control access to their own records.

Proficiency testing – A process in which the test results for a specific analyte are compared to those of other laboratories that use the same testing method to verify the accuracy of the test results.

Proficiency testing – A standardized, formal assessment of the quality of test results. These are commonly performed when the laboratory is sent a sample with a known value to see how competent its testing methods are when the specimen is tested.

Profiles – A group of tests often ordered together that test for a specific body system or a specific disease process. Also known as panels.

Prophylactic – A treatment that is used to prevent a disease process or prevent an unwanted outcome.

Prostatitis – Inflammation of the prostate.

Prostatitis specimen – A urine specimen collected to identify a bacterial infection of the prostate.

Prostatitis specimen collection process – A special procedure for collecting urine in which there are three specimens collected in close progression.

Protected health information (PHI) – All information about a patient included in his or her medical record.

Proteinuria – The presence of protein in the urine.

Prothrombin time (PT; protime) – A measurement (in seconds) of the time needed for blood to clot when thromboplastin and calcium are added to a plasma specimen. This test is used to evaluate those who are on oral anticoagulant therapy as well as to screen for potential clotting disorders.

Protime – See Prothrombin time.

Protoporphyrin – A substance included in the heme portion of the hemoglobin molecule. Necessary to make iron functional and available to attach to the oxygen.

Protozoa – Complex single-cell microorganisms, most of which are nonpathogenic.

Provider-performed microscopy (PPM) – Microscopy examinations performed by a health-care provider in the office. To perform these legally, the laboratory will need to be registered with CLIA as a facility that performs PPM.

Provider-performed microscopy procedures (PPMPs) – These are laboratory procedures that are performed by a midlevel provider or a physician using a microscope.

The examinations are performed on specimens that are unfixed and time sensitive, such as urine sediment microscopic examinations and wet mounts. There are specific quality control requirements that accompany this type of laboratory procedure, including maintenance on the microscope used for the examination.

Proximal convoluted tubule – The proximal part of the renal tubule where the renal filtrate first enters after it leaves the Bowman's capsule.

Plasma separator tubes (PSTs) – PSTs contain a thixotropic gel that separates the blood cells from the liquid when spun in a centrifuge.

Pus – The liquid product of inflammation, including dead cells, fluid, and pathogens.

Pyuria – The presence of white blood cells or "pus" in the urine specimen.

Qualitative – A term used to describe results which indicate whether or not a substance is present in a specimen, rather than the amount of the substance that is present. Qualitative tests provide a positive or negative result.

Qualitative result – A test that determines the presence or absence of a substance, having a yes/no or positive/negative result.

Quality assurance – All the processes used to assure quality in laboratory testing, including all phases of the laboratory experience: before, during, and after the test is performed.

Quality control (QC) – Practices used in the laboratory to ensure that the results for a given test are correct.

Quantitative HCG – A test that provides a numeric result for the presence of human chorionic gonadotropin (HCG) in serum. This test may be necessary if an ectopic pregnancy or fetal demise is suspected.

Quantitative result – A test that determines the amount of a specific substance in a sample. The result is expressed as a number or percentage.

Quantitative test – A test that provides a numeric result, indicating the level of the analyte present in the specimen.

Quantity not sufficient (QNS) specimen – A specimen that is not of sufficient quantity for testing.

RACE – An acronym used to remind employees of the order of action in the event of a fire. R: Rescue those who might be in danger; A: Alarm, activate the alarm system; C: Confine the fire as much as possible by closing doors and windows; E: Extinguish the fire if possible.

Radioimmunoassay (RIA) testing – A testing method in which the antigen or antibody in the specimen binds to a radioactive substance to be measured.

Random – Without plan or purpose or pattern. When used in connection with urine specimen collection, *random* refers to specimens collected at unspecified intervals.

Random collection – A specimen that can be drawn at any time of the day without limitations or special preparation.

Rapid flow immunoassay – A term that may be used interchangeably with lateral flow immunoassay. Rapid flow immunoassay also describes test methods in which the specimen/reagents do not flow laterally across the surface.

Reabsorption – The process by which water and other necessary substances are reabsorbed into the body through the capillaries that surround the renal tubules.

Reagent dipstick – A plastic strip that is imbedded with various reagent pads that test for chemicals in urine specimens.

Reagent strips – Strips of material (usually some type of plastic) that have been covered or treated with chemicals. These are designed to change color or produce certain substances as part of a chemical reaction. Reagent strips are designed to be used one time only, and may be used to detect the presence of a certain substance in a specimen or to measure the amount of a specific substance.

Red blood cell (RBC) – Formed element in the blood; primary function of red blood cells is to transport oxygen.

Red blood cell index – A calculated value that helps to classify RBCs in terms of size and hemoglobin concentration.

Reference laboratory – A reference laboratory is a private commercial laboratory that tests a high volume of specimens, and often performs highly specialized tests that are not available at a physician office laboratory or a hospital laboratory. The reference laboratory receives specimens from other laboratories, and hospital or physician office laboratories, not just from their general community. Specimens may be shipped from all over the country.

Reference ranges – The expected ranges for a specific population for an analyte. These ranges can vary by age, gender, or race. Ranges are established by each laboratory, based on testing method.

Reflexive testing – A policy in which follow-up tests are added automatically to a specimen when abnormal results are found with the initial testing ordered. This often applies to culture and urine specimens.

Reflexive – The automatic performance of additional laboratory testing in response to abnormal test results. For urine, the urine microscopic examination is often added on as a reflexive test for abnormal urine chemistry results.

Refractive index – A measurement of how much light is deflected from its normal path when it passes through a different media. The amount of deflection may be affected by the concentration of dissolved substances in the medium.

Refractometer – An instrument used to determine the specific gravity of a liquid.

Regulated waste – See Biohazardous waste.

Renal cortex – The outer layer of the kidney.

Renal medulla – The inner layer of the kidney.

Renal pelvis – A basin-shaped hollow structure that collects urine as it is formed before it exits the kidney through the ureters.

Renal tubule – The area of the nephron where reabsorption and secretion occur as urine is formed.

Reportable condition – A condition or disease that has been deemed a threat to the public.

Resident bacteria – Another term used to describe normal flora.

Resistant – A bacterial microorganism that is not affected by the properties of a specific antibiotic.

Reticulocyte – The last immature stage of red blood cells; when stained with a supravital stain, the darkly staining granules are fragments of the endoplasmic reticulum.

Reticulum – A mesh-like microscopic network in a cell.

Retroperitoneal – Behind the peritoneum.

Revolving nosepiece – The part of a microscope where the objectives attach. The revolving nosepiece can be turned as the different objectives are utilized.

RhoGam – A type of immune globulin used to protect neonates who have Rh-negative mothers.

Risk management – Ways of reducing potential risk to patients, employees, and employers.

Rotations per minute (rpm) – Used to measure the speed at which a centrifuge operates.

Rouleaux – A group of red blood cells that are stuck together, resembling a roll of coins.

Rugae – A fold or crease in the cells lining various areas of the body that allows for more stretching and flexibility of these areas.

Sanitization – To remove microorganisms from equipment or surfaces.

Scabies – An infection of the skin with a parasitic mite that causes intense itching.

Scope of practice – The various activities that a professional is qualified to perform. This may be based on education and/or experience and differs by profession.

Secretion – The final process in urine formation within the renal tubules. During secretion, substances are added to the urine filtrate from the capillaries that are wrapped around the renal tubules.

Sediment – Formed elements in the urine specimen that settle to the bottom of the urine specimen when spun in the centrifuge.

Semiquantitative – A type of testing procedure that provides an approximate amount of a substance.

Sensitive – A bacterial microorganism that is killed or kept from multiplying by the use of a specific antibiotic.

Sensitivity – The minimum concentration of the desired substance that will be detected.

Septicemia – Infection of the bloodstream by a pathogenic microorganism.

Sequela – A secondary complication or infection as a result of the primary infection.

Serially – A series at set intervals. This term is used to describe laboratory samples that are to be collected at specific intervals over a period of time.

Serology – The study of the antigen and antibody components in the liquid portion of the blood or other body fluids.

Serotonin – A chemical vasoconstrictor.

Serum – The liquid portion of the blood without the various clotting factors. Serum is taken from a blood specimen that has been allowed to clot before centrifugation.

Serum separator tubes (SST) – Tubes that contain a thixotropic separating gel that is designed to form a barrier between the blood cells and the liquid portion of the blood after it is centrifuged.

Sharp – Any object that has projections, corners, or edges that are capable of piercing the skin. If the object has come into contact with blood or other potentially infectious materials, it is a biohazardous sharp. These must be disposed of in a special, puncture-proof container.

Shift – A sudden change in quality control values. This may be a shift from one side of the mean to another, or a significant jump in values on the same side of the mean.

Short draw – When there has been a loss of vacuum in the evacuated tube, blood will not enter the tube, or it will fill incompletely. Often, this requires that the specimen be redrawn.

Sickle cell – Red blood cell changed to a crescent shape by low oxygen tension. These cells contain the abnormal hemoglobin, Hb S.

Sickle cell anemia – An inherited autosomal recessive trait in which the individual's red blood cells contain hemoglobin S.

Sickle cell trait – Condition of being heterozygous for Hb S in which each red blood cell has Hb A and Hb S; these cells become sickled under extremely low oxygen tension.

Slander – Untrue statements that are spoken about another person.

Slant – A tube with agar added and allowed to harden in a slanted form to allow for maximum surface area for bacterial growth.

Specific gravity – The weight of a substance as compared to an equal volume of distilled water. The specific gravity of urine provides an indication of the amount of dissolved substances present in the specimen.

Specificity – The ability to detect only the desired substance in the specimen.

Specimens – A piece or part of something that is designed to show the characteristics of the whole.

Spectrophotometer – A machine (or part of a machine) that is used to measure how much light of a certain wavelength has been absorbed by a specimen.

Spectrophotometrically – An estimation of coloring matter in a solution by wavelength of light.

Splenomegaly – An enlarged spleen.

Spermatozoa – A mature male sex cell.

Spherocyte – An erythrocyte that assumes a spheroid shape and has no central pallor.

Spirilla – Bacteria that have a spiral shape when viewed under the microscope.

Spores – A dormant bacterial form that is resistant to heat and moisture. Spores are the form that many bacteria assume when conditions are unfavorable, and this protective form allows them to survive until the conditions become more comfortable for growth and replication.

Sputum – Mucus that is expelled from the lungs and lower respiratory system with a deep cough.

Stage – The portion of a microscope where the glass slide is placed for viewing.

Standard of care – An acceptable level of performance for the employees within a profession.

Standard precautions – Guidelines developed by the CDC to assist with the spread of infectious disease in health-care facilities.

Standing orders – Orders for laboratory tests to be drawn at specific intervals over a period of time.

Staphylococci – Bacteria that are round and appear in grape-like clusters when viewed under a microscope.

STAT – Something that is to be accomplished immediately.

Statins – A classification of drugs used to treat patients with elevated lipid levels. Statins are especially effective for treatment of elevated LDL levels and triglycerides.

Statutes – Laws established by the legislative branch of a government entity.

Stem cells – Precursors of all blood cells that are present in the bone marrow.

Stereomicroscope – A microscope that allows a specimen to be viewed from two different angles to create a three-dimensional image. Also known as a dissecting microscope.

Sterilize – A process by which all microorganisms are destroyed on a surface.

Strep screens – CLIA-waived procedures often performed in a physician office laboratory. These tests employ a throat swab specimen and produce results within minutes. Strep screens test for group A streptococci, and allow for timely treatment of streptococcal throat infections.

Streptococcal pharyngitis – Inflammation and infection of the pharynx and tonsillar area caused by group A streptococcal bacteria.

Streptococci – Bacteria that appear round and in small "chains" when viewed under the microscope.

Stroke – A loss of neurological function that is caused by vessel injury or occlusion in the blood vessels of the brain. Also known as a cerebrovascular accident (CVA).

Sulfosalicylic acid precipitation test – A test that uses turbidity as a measure of the presence of protein in the urine.

Supernatant – The clear part of the urine specimen left at the top of the tube after the sediment has settled out to the bottom of the tube.

Suprapubic aspiration – A procedure used to remove urine from the bladder when traditional catheterization is not an option. Suprapubic aspiration uses a syringe and needle to aspirate urine directly from the bladder. The needle is inserted through the external abdominal wall.

Surgical asepsis – The destruction of all pathogenic organisms before they enter the body. This involves the use of sterile supplies for invasive procedures.

Susceptible – Having very little or no defense from infection by a specific microorganism.

Syringe system – A method of drawing blood that involves a syringe with a needle attached. This method is used when a patient has veins that may not be able to withstand the evacuated tube method, which applies a lot of reverse pressure all at once, which may collapse the vein.

T cells – Specialized lymphocytic white blood cells that are programmed in the thymus to serve as part of the immune response.

Tachometer – An instrument used to measure rotation speed.

Target cell – An erythrocyte that has a reduced hemoglobin concentration and appears with a little color around the edge of the cell and a dark central area.

Thalassemia – Inherited disorders in which abnormal hemoglobin is produced by either a defective production rate of the alpha or beta globin chains or a decreased synthesis of the beta chain.

Thrombin – An enzyme that converts fibrinogen to fibrin, which is necessary for definitive blood clot formation.

Thrombocytes – Platelets.

Thrombocythemia – Overproduction of platelets by the bone marrow.

Thrombocytopenia – An abnormal decrease in the number of platelets.

Thrombocytosis – An increase in the number of platelets.

Thrombopoietin – A growth factor that acts on the bone marrow to stimulate platelet production as well as the production of other cell lines.

Thrombosis – A blood clot formed in response to vascular injury, or pathological formation of a blood clot.

Thrombus – A blood clot that adheres to the wall of a vessel or organ.

Thyroid-stimulating hormone (TSH) – A hormone produced by the pituitary gland that stimulates the thyroid gland to produce thyroxine, thyrocalcitonin, and triiodothyronine.

Thyroxine (T4) – A hormone produced by the thyroid gland that affects the rate of metabolism of the cells throughout the body.

Tort – Wrongs that are committed that do not involve a contractual agreement.

Tourniquet – A stretchy, rubber-like band that is applied to the arm to facilitate the location of veins to be used for blood draws.

Transient bacteria – Bacteria that are not part of the normal flora of the body. These are the bacteria that are picked up on the hands and skin during daily activities and those that are washed off when washing the hands.

Transmissible – Something that is capable of being passed, or transmitted, to another.

Trends – A slow, steady incline or decline in the quality control values over time.

Trichomoniasis – Infection with a parasite that often lives in the vagina in women or urethra in men. This parasite causes discharge, itching, and burning, and may be transmitted through sexual intercourse.

Triiodothyronine (T3) – A hormone produced by the thyroid gland that affects the rate of metabolism of the cells throughout the body.

Troponin – A muscle protein that is found in skeletal and cardiac muscle, but not smooth muscle. An elevated cardiac troponin level is one of the first indicators of heart trauma.

Troubleshooting – Investigative processes employed when the quality control specimens are not within the expected ranges.

Trough collection – A specimen drawn at the time when a medication is at the lowest level in the bloodstream. This is usually drawn immediately before the next dose is due.

Type 1 diabetes – Diabetes that results from inadequate or absent insulin secretion. Extended periods of hyperglycemia result, and treatment involves insulin injections.

Type 2 diabetes – Diabetes that results from insulin resistance. Type 2 diabetes is usually successfully treated with lifestyle changes and oral medications.

Unintentional tort – A wrongful act that is accidentally committed by one person against another person.

Universal precautions – The concept that all blood products and other potentially infectious materials are to be treated as if they are infected with bloodborne pathogens.

Ureters – The tubes that carry urine from the kidney to the bladder. Most individuals have two ureters, one for each kidney.

Urethra – The tube that extends from the bladder to the outside of the body through which urine travels.

Urinalysis – Analysis of a urine specimen. The urinalysis usually includes a macroscopic examination of the specimen for color and clarity, measurement of the specific gravity of the specimen, testing for various chemicals, and may also include microscopic examination for suspended structures in the specimen.

Urinary frequency – An increase in the urge to urinate.

Urinary incontinence – Involuntary urination.

Urinary meatus – The external opening of the urethra.

Urine culture – The process by which a urine specimen is added to a medium containing nutrients that nurture the growth of pathogenic microorganisms.

Urine microscopic – The examination of urine sediment under the microscope for detection and quantification of structures suspended in the urine.

Urinometer – A type of hydrometer used to measure specific gravity of urine. A urinometer has a float that will fall to a depth measured at 1.000 when used with distilled water, but will not fall to this depth in urine specimens with an increase in dissolved substances.

Urobilinogen – A derivative of bilirubin; bilirubin is converted to urobilinogen by intestinal bacteria.

Urochrome – A pigment that gives urine its characteristic yellow color.

Uroscopy – The name originally used to describe the analysis of urine specimens. This name is no longer used routinely.

U.S. Department of Health and Human Services (HHS) – An agency that assumes the overall responsibility for laboratory quality assurance as designated in CLIA '88.

U.S. Food and Drug Administration (FDA) – The federal agency responsible for assigning the level of complexity for various laboratory tests.

Vaccine – A substance containing altered antigenic material used to stimulate an immune response to a particular pathogen and provide protection. Vaccines are usually injected.

Vacuum – A space that has all the air removed, creating a "vacuum" when it is opened (or pierced in the case of the vacuum tubes used for venipuncture).

Valid – A term used to describe a test result that can safely be reported. This usually refers to qualitative tests, in which there is an indicator on the testing device to show that the test result is valid. This is independent of whether the result is positive or negative.

Vasodilation – An enlargement of a vein due to the relaxation of the vessel walls.

Vasovagal syncope – Loss of consciousness, usually related to a stressful or upsetting event, such as a blood draw.

Vector – A carrier of pathogenic microorganisms. The vector is usually unaffected by the pathogen that it carries, but it is able to transmit it to a susceptible individual.

Veins – Blood vessels in the body that carry blood from the capillary beds back to the heart.

Venules – The smallest vessels in the venous system. Capillary vessels gradually enlarge and change in structure to become venules. Venules then become veins.

Virus – A microorganism that is too small to be seen without the aid of a microscope with advanced technology. Viruses consist of genetic material wrapped in a protein coating, and they use the genetic information of the host cell to reproduce themselves.

Viscous – Thick, sticky, or syrupy.

Very low-density lipoprotein (VLDL) – A subclass of cholesterol. Elevated VLDL levels are linked to an increased chance of plaque formation in the blood vessels.

von Willebrand's factor – A glycoprotein that is present in plasma. It carries clotting factor VIII, which is necessary for blood hemostasis. von Willebrand's factor also interacts with other processes in the clotting cascade.

Warfarin – An oral anticoagulant.

Water baths – A tank filled with water to be heated to a specific temperature. Test tubes are suspended in the heated water to speed up certain chemical reactions during testing procedures.

Westergren erythrocyte sedimentation rate (ESR) – A test performed in which the anticoagulated blood is mixed with a sodium citrate or saline solution prior to transfer into the pipette for testing.

Wet mount – A method used to observe living organisms microscopically without staining or preserving the specimen. This is especially useful when observing the movement (or motility) of microorganisms. A wet mount is essentially a suspension of the specimen placed on a slide for observation under the microscope.

White blood cell (WBC) – Formed element in the blood; the primary function of white blood cells is to provide immunity for the body.

Window period – The period of HIV infection when the infected individual does not test positive in typical laboratory tests, but is still capable of transmitting the disease. This period can last for several months after infection.

Winged infusion set – A butterfly setup. The winged infusion set consists of a needle (usually a half inch in length) attached to a piece of plastic tubing. The needle has plastic "wings" on both sides where the fingers grip the device while drawing blood.

Wintrobe erythrocyte sedimentation rate (ESR) – A method of performing the ESR in which the anticoagulated blood is placed directly into a short Wintrobe tube specifically designed to perform the ESR. The tube is graduated in millimeters for reading the results. The tubes are usually made of glass.

Wood's lamp – A lamp used for examination of the skin. The Wood's lamp uses ultraviolet light to detect organisms or areas of the body that fluoresce. Many mycotic infections produce fluorescence.

Work practice controls – Controls that alter the manner in which a task is performed to minimize the risk of exposure to bloodborne pathogens.

Zone of inhibition – The area around a specific antibiotic disk where the growth of the microorganism is absent. The organism on the culture plate is sensitive (susceptible) to the antibiotic that is on the disk.

DANA W. MAYO
RONALD M. PIKE
DAVID C. FORBES

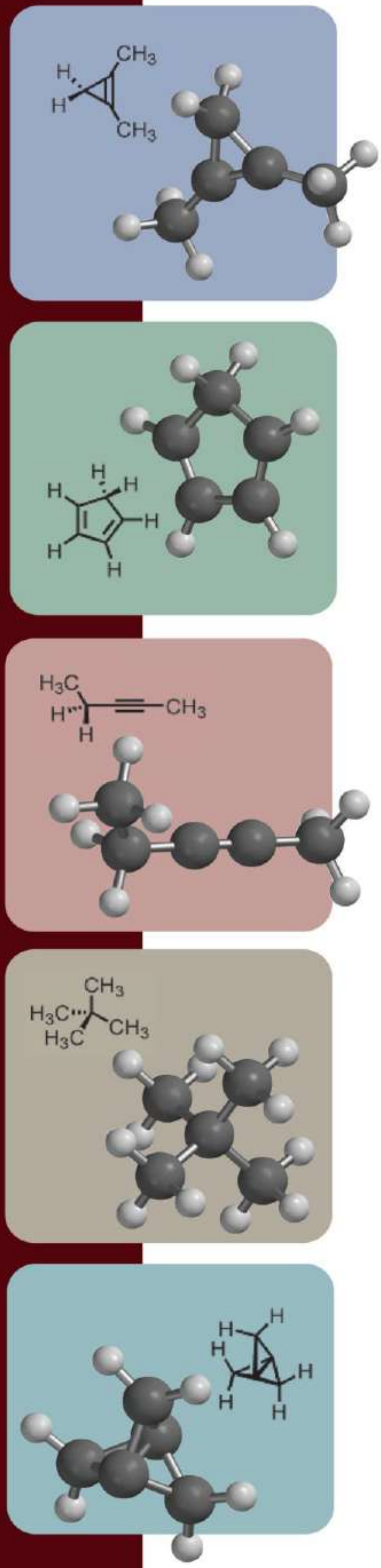
H
H
CH3
CH3
H
H
H
H
H
H
H3C
H
H
CH3
CH3
H3C
CH3
H3C
H
H
H
H
H
H

Microscale Organic Laboratory

WITH MULTISTEP AND MULTISCALE SYNTHESES

FIFTH EDITION

PREFACE

Twenty-five years ago, in 1985, when *Microscale Organic Laboratory (MOL)* was first published (as paperback Xerox copies of an unproofed manuscript!), it was the only microscale organic laboratory text available. In the February 1999 *Book Buyers Guide Supplement to the Journal of Chemical Education,* however, there were seventeen laboratory manuals (of a total of thirty-nine) containing miniaturized, fully microscale, or a mixture of micro and macro experiments. Fast forward ten years and without any doubt, microscale techniques have solidly established their place in chemical education. The number of lab manuals currently in print reflects the growing number of students being introduced to organic chemistry through microscale techniques. While the conversion may not yet be quite as high as the eighty percent predicted by David Brooks back in 1985, a conservative estimate would be that a solid two-thirds majority of sophomore students now work with miniaturized experiments compared with the amounts of material employed in these laboratories in the late 1970s.

The major changes that were made to *MOL* in the fourth edition were very well received by our readers. Indeed, we are now nearing the fine-tuning stage in the evolution of this laboratory text. Hence, *MOL5* on the surface will look very much like *MOL4*. *MOL5,* however, has undergone further significant internal reorganization and rewriting. Many helpful suggestions have been received from reviewers and from instructors who have used previous editions of this text. As a result, some major changes have been made for this new edition:

- A key change to the 5th edition is the modification of the procedural sections to allow for inquiry-based experimentation. Reaction times have been replaced with guidelines, and options on how to best monitor reactions and gauge product purity are left to the discretion of the student or instructor. Many ideas and new approaches can stem from this change—for example, instructional sections can be split into small groups and each group can approach the monitoring of a reaction differently. Using a completely separate set of experiments, discussions can be pursued which focus on reaction purity and evidence which offers the experimentalist sufficient data about what was prepared. Students can then compare notes at the end of the lab period and discuss the various approaches and end results. We hope that this change will allow the lab to become a more interactive experience between groups of students, should that be the wish of the instructor. The opportunity to monitor a reaction rather than assume reaction completion by simply following a time-based instruction and to allow for students to gather additional evidence of product purity empowers the student and adds an element of excitement to the lab experience. Optional inquiry-based guidelines have been added to experiments 5A, 5B, 7, 19B, 24A, and 32. Experiments 11A, 16, and 28 have been modified in a way which focuses on validation of product purity. References to inquiry-based guidelines ? and validation experiences V are noted in the text.
- The use of microwave heating as a tool in synthetic organic chemistry is fast-growing and is becoming an enabling technology. Optional instructions have been added to experiments to allow for the integration of microwave heating as a tool for performing reactions. Since reaction times are shorter than when conventional heating methods are used,

students have the opportunity to supplement these activities with traditional techniques and as stated above engage in discussions comparing the two. Optional microwave heating instructions have been added to experiments 7, 8, 15, 22, and 30. References to microwave use are noted in the text by the use of this icon M.

- A rich collection of end of chapter exercises and the addition of pre and post lab questions provides students with the valuable opportunity to test and practice their own understanding of each laboratory experiment.
- Discussion sections that appear at the beginning of each Experiment have been added, revised, and expanded upon. These discussion sections provide chemical context/background for each experiment, and provide more information regarding the chemical principles involved in each experimental procedure.
- **The Refractive Index material, which was formally in chapter 4, has been moved to the book companion web site: http://www.wiley.com/college/mayo.**
- **Chapter 10W, "Advanced Microscale Organic Laboratory Experiments," (formerly chapter 7) has been moved to the book companion web site: http://www.wiley.com/college/mayo.**

Additional Resources

Text web site-http://www.wiley.com/college/mayo

As with the previous edition, a major portion of the background theoretical discussions have been moved to the text web site, without affecting the operational part of the text. Likewise, the web site has allowed us to move a number of more advanced discussions out of the printed text. Wherever the shift of this material has occurred the move is flagged by reference call-outs using an icon www.

These **web reference discussions** include information on the following topics:

- Microscale lab equipment and techniques
- Semimicroscale distillation
- Reduced pressure distillations with microspinning band columns
- Vacuum pumps and pressure regulation
- Crystallization
- Measurement of Specific Rotation
- Introduction to Infrared Spectroscopy—Introduction to Theory
- Group Frequencies of the Hydrocarbons
- Characteristic Frequencies of the Heteroatom Functional Groups
- Instrumentation—the Infrared Interferometer
- Tables of Derivatives

The majority of the background infrared spectra and the associated discussions used to develop the use of group frequencies from these spectra are also found on the web site, while the text still contains the essential tables of characteristic frequencies that are in every day use in the laboratory. The many compound data tables, used primarily in the chapter on qualitative identification, also reside on the web site. The Classification of Experiments Based on Mechanism is also available on the web site.

CONTENTS

Chapter 1

INTRODUCTION 1

General Rules for the Microscale Laboratory 3
The Organic Chemistry Laboratory 4

Chapter 2

SAFETY 5

Making the Laboratory a Safer Place 5
Nature of Hazards 5
Reduction of Risks 6
Precautionary Measures 7
Thinking About the Risks In Using Chemicals 8
Disposal of Chemicals 8
Material Safety Data Sheets 9
Alternate Sources of Information 12
Estimating Risks from Vapors 13
Microwave Safety 14
Concluding Thoughts 15
General Safety References 16

Chapter 3

INTRODUCTION TO MICROSCALE ORGANIC LABORATORY EQUIPMENT AND TECHNIQUES 18

Microglassware Equipment 19
  Standard Taper Joints 19
  Conical Vials 20
  Condensers 20
  Distillation Heads 20
  Recrystallization Tubes 20
  Miscellaneous Items 20
  Gas Chromatographic Fraction Collection Items 21
Standard Experimental Apparatus 21
  Heating and Stirring Arrangements 21
  Sand Bath Technique—Hot Plate Calibration 21
  Metal Heat-Transfer Devices 22
  Stirring 22
  Reflux Apparatus 23
  Distillation Apparatus 24
  Moisture-Protected Reaction Apparatus 25
  Specialized Pieces of Equipment 26
Microwave Heating as a Tool for Organic Chemistry 27
  Introduction 27
  Applications 32
  Equipment Available 34
  Experimental Protocols 35
Microscale Laws 35
  Rules of the Trade for Handling Organic Materials at the Microscale Level 35
  Rules for Working with Liquids at the Microscale Level 36
  Rules for Working with Solids at the Microscale Level 39
The Laboratory Notebook 40
  Key Components of a Laboratory Experiment Write-up 40
Example of a Laboratory Notebook Entry 41
Calculating Yields 42

Chapter 4

DETERMINATION OF PHYSICAL PROPERTIES 45

Liquids 46
  Ultramicro Boiling Point 46
Density 49
Solids 50
  Melting Points 50

Chapter 5

MICROSCALE LABORATORY TECHNIQUES 55

TECHNIQUE 1 *Gas Chromatography* 55
  GC Instrumentation 56
TECHNIQUE 2 *Simple Distillation* 61
TECHNIQUE 3 *Fractional Distillation* 64
TECHNIQUE 4 *Solvent Extraction* 67
  Intermolecular Properties: Solubility 67

- Partition (or Distribution) Coefficient **70**
  - Extraction **72**
  - Solid–Liquid Extraction **79**
  - Drying Agents **80**
  - Solid-Phase Extraction **83**
- **TECHNIQUE 5** *Crystallization* **85**
  - General Crystallization Procedure **85**
  - Simple Crystallization **87**
  - Filtration Techniques **88**
- **TECHNIQUE 6** *Chromatography* **92**
  - Column, Flash, High-Performance Liquid, and Thin-Layer Chromatography **92**
  - Column Chromatography **92**
  - Flash Chromatography **95**
  - Thin-Layer Chromatography **97**
  - Paper Chromatography **99**
  - High-Performance Liquid Chromatography **100**
- **TECHNIQUE 6B** *Concentration of Solutions* **101**
  - Distillation **102**
  - Evaporation with Nitrogen Gas **102**
  - Removal of Solvent Under Reduced Pressure **102**
- **TECHNIQUE 7** *Collection or Control of Gaseous Products* **105**
  - Water-Insoluble Gases **105**
  - Trapping Byproduct Gases **106**
- **TECHNIQUE 8** *Measurement of Specific Rotation* **108**
  - Theory **108**
  - The Polarimeter **109**
- **TECHNIQUE 9** *Sublimation* **111**
  - Sublimation Theory **112**
  - Experimental Setup **113**
  - Precautions **113**

Chapter 6

MICROSCALE ORGANIC LABORATORY EXPERIMENTS 115

- **EXPERIMENT 1** *Getting to Know You: Measurement of Physical Properties* **116**
  - Discussion **117**
  - Experimental Procedure **118**
    - Melting Point **118**
- **EXPERIMENT 2** *The Separation of a 25-μL Mixture of Heptanal (bp 153 °C) and Cyclohexanol (bp 160 °C) by Gas Chromatography* **123**
  - Discussion **123**
  - Collection Yield **124**
  - Collection Yield **125**
  - Components **126**
  - Experimental Procedure **126**
- **EXPERIMENT 3** *Distillation* **129**
- **Experiment 3A** Simple Distillation at the Semimicroscale Level: Separation of Ethyl Acetate from *trans*-1,2-Dibenzoylethylene **130**
  - Discussion **130**
  - Components **130**
  - Experimental Procedure **131**
- **Experiment 3B** Fractional Semimicroscale Distillation: Separation of Hexane and Toluene **132**
  - Discussion **133**
  - Components **133**
  - Experimental Procedure **133**
- **Experiment 3C** Fractional Semimicroscale Distillation: Separation of Methylpentane and Cyclohexane Using a Spinning-Band Column **135**
  - Discussion **135**
  - Components **136**
  - Experimental Procedure **136**
- **Experiment 3D** Fractional Semimicroscale Distillation: The Separation of 2-Methylpentane and Cyclohexane Using a Spinning Band in a Hickman–Hinkle Still **138**
  - Discussion **139**
  - Components **139**
  - Experimental Procedure **139**
- **EXPERIMENT 4** *Solvent Extraction* **141**
- **Experiment 4A** Determination of Partition Coefficient for the System Benzoic Acid, Methylene Chloride, and Water **141**
  - Discussion **141**
  - Components **144**
  - Experimental Procedure **144**
- **Experiment 4B** Solvent Extraction I: The System; Benzoic Acid, Methylene Chloride, and 10% Sodium Bicarbonate Solution; An Example of Acid–Base Extraction Techniques **146**
  - Reaction **146**
  - Discussion **146**
  - Experimental Procedure **146**
- **Experiment 4C** Solvent Extraction II: A Three-Component Mixture; An Example of the Separation of an Acid, a Base, and a Neutral Substance **147**
  - Discussion **147**
  - Components **148**
  - Experimental Procedure **148**

EXPERIMENT 5 *Reduction of Ketones Using a Metal Hydride Reagent: Cyclohexanol and* cis- *and* trans-4-tert-*Butylcyclohexanol* **151**
  Reaction (Experiment [5A]) **151**
  Discussion **152**
Experiment 5A Cyclohexanol **153**
  Experimental Procedure **153**
Experiment 5B *cis*- and *trans*-4-*tert*-Butylcyclohexanol **158**
  Reaction **158**
  Experimental Procedure **158**
EXPERIMENT 6 *Photochemical Isomerization of an Alkene:* cis-*1,2-Dibenzoylethylene* **163**
  Biologically Important Photochemical Reactions **164**
  Reaction **165**
  Discussion **166**
Experiment 6a Purification of *trans*-1,2-Dibenzoylethylene **166**
  Experimental Procedure **166**
Experiment 6B Isomerization of an Alkene: Thin-Layer Chromatographic Analysis **167**
  Experimental Procedure **167**
Experiment 6C Isomerization of an Alkene: Nuclear Magnetic Resonance Analysis **173**
  Experimental Procedure **173**
EXPERIMENT 7 *The Cannizzaro Reaction with 4-Chlorobenzaldehyde: 4-Chlorobenzoic Acid and 4-Chlorobenzyl Alcohol* **174**
  Reaction **176**
  Discussion **176**
  Experimental Procedure **177**
Experiment 7-1 4-Chlorobenzoic Acid and 4-Chlorobenzyl Alcohol: Preparation Using a Monomode Microwave Apparatus **184**
  Experimental Procedure **184**
Experiment 7-2 4-Chlorobenzoic Acid and 4-Chlorobenzyl Alcohol: Preparation Using a Multimode Microwave Apparatus **185**
  Experimental Procedure **185**
EXPERIMENT 8 *The Esterification Reaction: Ethyl Laurate, Isopentyl Acetate, and the Use of Acidic Resins* **188**
  Reaction **188**
  Discussion **189**
  Lipids **190**
Experiment 8A Ethyl Laurate **199**
  Reaction **199**
  Experimental Procedure **199**
Experiment 8B Isopentyl Acetate: Semimicroscale Preparation **201**
  Reaction **201**
  Experimental Procedure **201**
Experiment 8B-1 Isopentyl Acetate: Preparation Using a Monomode Microwave Apparatus **203**
  Experimental Procedure **203**
Experiment 8B-2 Isopentyl Acetate: Preparation Using a Multimode Microwave Apparatus **205**
  Experimental Procedure **205**
Experiment 8C Esterification by Use of Acidic Resins: Semimicroscale Preparations **206**
  Reaction **207**
  Experimental Procedure **207**
EXPERIMENT 9 *The E1 Elimination Reaction: Dehydration of 2-Butanol to Yield 1-Butene,* trans-*2-Butene, and* cis-*2-Butene* **209**
  The Development of Carbocation Theory **210**
  Reaction **211**
  Discussion **212**
  Experimental Procedure **215**
EXPERIMENT 10 *The E2 Elimination Reaction: Dehydrohalogenation of 2-Bromobutane to Yield 1-Butene,* trans-*2-Butene, and* cis-*2-Butene* **217**
  Reaction **218**
  Discussion **218**
  Experimental Procedure **220**
EXPERIMENT 11 *The Isolation of Natural Products* **224**
Experiment 11A Isolation and Characterization of an Optically Active Natural Product: Usnic Acid **224**
  Lichens and Natural Products **225**
  Discussion **227**
  Experimental Procedure **227**
Experiment 11B Isolation and Characterization of a Natural Product: Caffeine and Caffeine 5-Nitrosalicylate **229**
  Alkaloids **230**
  The Classification of Alkaloids **230**
  Discussion **231**
  Experimental Procedure **233**
  Derivative: Caffeine 5-Nitrosalicylate **235**
  Experimental Procedure **236**
Experiment 11C Isolation of a Natural Product by Steam Distillation: Cinnamaldehyde from Cinnamon **238**
  Essential Oils **239**

Discussion **241**
Component **242**
Experimental Procedure **242**
EXPERIMENT 12 *Reductive Catalytic Hydrogenation of an Alkene: Octane* **244**
Reaction **245**
Discussion **245**
Experimental Procedure **247**
EXPERIMENT 13 *Hydroboration–Oxidation of an Alkene: Octanol* **250**
Reaction **251**
Discussion **251**
Experimental Procedure **254**
EXPERIMENT 14 *Diels–Alder Reaction: 4-Cyclohexene-*cis*-1,2-dicarboxylic Acid Anhydride* **257**
Reaction **258**
Discussion **259**
Experimental Procedure **261**
Optional Semimicroscale Preparation **266**
EXPERIMENT 15 *Diels–Alder Reaction: 9,10-Dihydroanthracene-9,10-α,β-succinic Acid Anhydride* **269**
Reaction **269**
Discussion **270**
Experimental Procedure **271**
Optional Semimicroscale Preparations **272**
Experiment 15-1 9,10-Dihydroanthracene-9,10-α,β-succinic Acid Anhydride: Preparation Using a Monomode Microwave Apparatus **273**
Experimental Procedure **273**
Experiment 15-2 9,10-Dihydroanthracene-9,10-α,β-succinic Acid Anhydride: Preparation Using a Multimode Microwave Apparatus **274**
Experimental Procedure **274**
Experiment 16 *Grignard Reaction with a Ketone: Triphenylmethanol* **275**
Reaction **276**
Discussion **277**
Experimental Procedure **279**
EXPERIMENT 17 *Grignard Reaction with an Aldehyde: 4-Methyl-3-heptanol* **284**
Reaction **284**
Discussion **284**
Experimental Procedure **285**
EXPERIMENT 18 *The Perkin Reaction: Condensation of Rhodanine with an Aromatic Aldehyde to Yield o-Chlorobenzylidene Rhodanine* **289**
Reaction **290**
Discussion **291**
Experimental Procedure **292**
Optional Semimicroscale Preparation **293**
EXPERIMENT 19 *Alkene Preparation by the Wittig Reaction: (E)-Stilbene; Methylene-4-*tert-*butylcyclohexane; and* trans*-9-(2-Phenylethenyl) anthracene* **294**
Reaction **296**
Discussion **296**
Experiment 19A (*E*)-Stilbene by the "Instant Ylide" Method **299**
Reaction **299**
Experimental Procedure **300**
Experiment 19B (*E*)-Stilbene by the Horner–Wadsworth–Emmons Reaction **302**
Reaction **302**
Experimental Procedure **302**
Experiment 19C Methylene-4-*tert*-butylcyclohexane **303**
Reaction **303**
Experimental Procedure **304**
Experiment 19D *trans*-9-(2-Phenylethenyl)-anthracene **306**
Reaction **306**
Experimental Procedure **306**
EXPERIMENT 20 *Aldol Reaction: Dibenzalacetone* **309**
Reaction **309**
Discussion **310**
Experimental Procedure **311**
Optional Semimicroscale Preparation **316**
EXPERIMENT 21 *Quantitative Analysis of Grignard Reagents: 1-Methylbutylmagnesium Bromide and Phenylmagnesium Bromide* **317**
Reaction **318**
Discussion **318**
Experimental Procedure **319**
EXPERIMENT 22 *Williamson Synthesis of Ethers* **321**
Reaction **321**
Discussion **322**
Experiment 22A Propyl *p*-Tolyl Ether **323**
Experimental Procedure **323**
Optional Macroscale Preparation **324**
Experiment 22B Methyl *p*-Ethylphenyl Ether **327**
Reaction **327**
Experimental Procedure **327**
Optional Semimicroscale and Macroscale Preparations **329**

Experiment 22C Butyl *p*-Nitrophenyl Ether: Preparation Using a Monomode Microwave Apparatus **332**
- Reaction **332**
- Experimental Procedure **332**

Experiment 22D Butyl *p*-Nitrophenyl Ether: Preparation Using a Multimode Microwave Apparatus **334**
- Reaction **334**
- Experimental Procedure **334**

EXPERIMENT 23 *Amide Synthesis: Acetanilide and* N,N′-*Diacetyl-1,4-phenylenediamine* **338**
- Reaction **338**
- Discussion **339**

Experiment 23A Acetanilide **341**
- Experimental Procedure **341**
- Optional Semimicroscale Preparation **342**

Experiment 23B *N,N′*-Diacetyl-1,4-phenylenediamine **343**
- Reaction **343**
- Experimental Procedure **343**

EXPERIMENT 24 *Imide Synthesis:* N-*Phenylmaleimide* **346**
- Reaction **346**
- Discussion **347**

Experiment 24A Maleanilic Acid **348**
- Experimental Procedure **348**
- Reaction **348**

Experiment 24B *N*-Phenylmaleimide **349**
- Reaction **349**
- Experimental Procedure **350**

EXPERIMENT 25 *Synthesis of Cyclic Carboxylic Acid Anhydrides: Succinic Anhydride and Phthalic Anhydride* **352**
- Reaction **352**
- Discussion **352**

Experiment 25A Succinic Anhydride **354**
- Experimental Procedure **354**

Experiment 25B Phthalic Anhydride **355**
- Reaction **355**
- Experimental Procedure **355**

EXPERIMENT 26 *Diazonium Coupling Reaction: Methyl Red* **356**
- Reaction **357**
- Discussion **357**
- Experimental Procedure **359**

EXPERIMENT 27 *Friedel–Crafts Acylation: Acetylferrocene and Diacetylferrocene* **361**
- Reaction **362**
- Discussion **362**
- Experimental Procedure **364**

EXPERIMENT 28 *Halogenation: Electrophilic Aromatic Substitution to Yield 4-Bromoacetanilide* **368**
- Reaction **368**
- Discussion **369**
- Experimental Procedure **369**

EXPERIMENT 29 *Nitration: 2,5-Dichloronitrobenzene;* N,N′-*Diacetyl-2,3-dinitro-1,4-phenylenediamine; 5-Nitrosalicylic Acid; and 2- and 4-Nitrophenol* **373**
- General Reaction **374**
- Discussion **374**
- Semimicroscale Preparation of Anhydrous Nitric Acid **375**

Experiment 29A 2,5-Dichloronitrobenzene **376**
- Experimental Procedure **376**
- Reaction **376**
- Experimental Procedure **377**

Experiment 29B *N,N′*-Diacetyl-2,3-dinitro-1,4-phenylenediamine **378**
- Reaction **378**
- Experimental Procedure **378**

Experiment 29C 5-Nitrosalicylic Acid **379**
- Reaction **379**
- Experimental Procedure **380**

Experiment 29D 2- and 4-Nitrophenol **381**
- Reaction **381**
- Experimental Procedure **381**

EXPERIMENT 30 *Nucleophilic Aromatic Substitution: 2,4-Dinitrophenylthiocyanate* **384**
- Reaction **385**
- Discussion **385**
- Experimental Procedure **386**

Experiment 30-1 2,4-Dinitrophenylthiocyanate: Preparation Using a Monomode Microwave Apparatus **387**
- Experimental Procedure **387**

Experiment 30-2 2,4-Dinitrophenylthiocyanate: Preparation Using a Multimode Microwave Apparatus **388**
- Experimental Procedure **388**

EXPERIMENT 31 *Halogenation Using* N-*Bromosuccinimide: 9-Bromoanthracene* **390**
- Reaction **390**
- Discussion **391**
  - Initiation Step **391**
  - Propagation Step **391**
- Experimental Procedure **392**

EXPERIMENT 32 *Hypochlorite Oxidation of an Alcohol: Cyclohexanone* **394**
  Reaction **394**
  Discussion **394**
  Experimental Procedure **395**
EXPERIMENT 33 *Chromium Trioxide–Resin or Hypochlorite Oxidation of an Alcohol: 9-Fluorenone* **398**
Experiment 33A 9-Fluorenone: CrO$_3$ Oxidation of 9-Fluorenol **398**
  Reaction **398**
  Discussion **398**
  Experimental Procedure **399**
Experiment 33B 9-Fluorenone: NaOCl Oxidation of 9-Fluorenol **401**
  Reaction **401**
  Discussion **401**
  Experimental Procedure **401**
EXPERIMENT 34 *Hypochlorite Oxidation of Methyl Ketones by the Haloform Reaction: Benzoic Acid and* p*-Methoxybenzoic Acid* **403**
  Reaction **404**
  Discussion **404**
Experiment 34A Benzoic Acid **405**
  Experimental Procedure **405**
Experiment 34B *p*-Methoxybenzoic Acid **406**
  Reaction **406**
  Experimental Procedure **406**
  Optional Semimicroscale Preparation **407**
EXPERIMENT 35 *Conversion of Cyclohexyl Bromide to Cyclohexene–An E2 Elimination Reaction: Factors Affecting the Rate of a Chemical Reaction* **409**
  Reaction **409**
  Discussion **409**
  Experimental Procedure **414**
  Data Analysis **416**
  Variation of Parameters **416**

Chapter 7

SEQUENTIAL SYNTHESES: THE TRANSITION FROM MACRO TO MICRO **421**
SEQUENCE A *The Synthesis of Hexaphenylbenzene* **424**
EXPERIMENTS A1$_a$, A2$_a$, A3$_a$, A1$_b$, A2$_b$, A3$_b$, and A4$_{ab}$ *The Synthesis of Hexaphenylbenzene from Benzaldehyde:* **427**
Experiment A1$_a$ The Benzoin Condensation of Benzaldehyde: Benzoin **429**
  Reaction **429**
  Discussion **430**
  Semimicroscale Experimental Procedure **431**
  Optional Scales **432**
  Microscale Reaction Procedure **432**
Experiment A2$_a$ Copper(II) Ion Oxidation of Benzoin: Benzil **433**
  Reaction **434**
  Discussion **434**
  Semimicroscale Experimental Procedure **435**
  Optional Microscale Preparation **437**
Experiment A3$_a$ Tetraphenylcyclopentadienone **438**
  Reaction **438**
  Discussion **439**
  Microscale Reaction Procedure (1) **439**
  Microscale Reaction Procedure (2) **440**
Experiment A1$_b$ (*E*)-Stilbene **441**
  Reaction **442**
  Discussion **442**
  Experimental Procedure **443**
Experiment A2$_b$ Bromination of (*E*)-Stilbene: *meso*-Stilbene Dibromide **444**
  Reaction **445**
  Discussion **445**
  Semimicroscale Experimental Procedure **447**
Experiment A3$_b$ Dehydrohalogenation of *meso*-Stilbene Dibromide: Diphenylacetylene **450**
  Reaction **450**
  Discussion **450**
  Semimicroscale Experimental Procedure **451**
  Optional Macroscale and Microscale Preparations **452**
Experiment A4$_{ab}$ Hexaphenylbenzene **453**
  Reaction **454**
  Discussion **454**
  Experimental Procedure **455**
SEQUENCE B *The Stepwise Synthesis of Nylon-6,6* **457**
Experiment B1 Oxidation of Cyclohexanol: Adipic Acid **458**
  Reaction **458**
  Discussion **458**
  Experimental Procedure **460**

Experiment B2 Preparation of an Acid Chloride: Adipoyl Chloride **461**
Reaction **461**
Discussion **462**
Experimental Procedure **462**
Experiment B3 Preparation of a Polyamide: Nylon-6,6 **464**
Reaction **464**
Discussion **464**
Experimental Procedure **465**
SEQUENCE C *The Synthesis of Sulfanilamide* **466**
The Sulfa Drugs **466**
Experiment C1 Acetylation of Aniline: 2,2,2-Trifluoroacetanilide **467**
Reaction **468**
Discussion **468**
Experimental Procedure **469**
Experiment C2 Chlorosulfonation of 2,2,2-Trifluoroacetanilide: *p*-(Trifluoroacetamido)-benzenesulfonyl Chloride **470**
Reaction **471**
Discussion **471**
Experimental Procedure **472**
Experiment C3 Preparation of an Arene Sulfonamide: Sulfanilamide **473**
Discussion **474**
Experimental Procedure **474**
SEQUENCE D *The Synthesis of 2′-Bromostyrene* **476**
The Synthesis of a Fragrance **476**
Experiment D1 The Verley–Doebner Modification of the Knoevenagel Reaction: *trans*-Cinnamic Acid **479**
Reaction **479**
Discussion **479**
Experimental Procedure **482**
Experiment D2 Bromination of *trans*-Cinnamic Acid: *erythro*-2,3-Dibromo-3-phenylpropanoic Acid **483**
Reaction **484**
Discussion **484**
Experimental Procedure **486**
Experiment D3 An Elimination Reaction with *erythro*-2,3-Dibromo-3-phenylpropanoic Acid: 2′-Bromostyrene **488**
Reaction **488**
Discussion **488**
Experimental Procedure **489**
PART A Conditions: Water Solvent **490**
PART B Conditions: Acetone Solvent **491**
SEQUENCE E *The Synthesis of Piperonylonitrile from Piperonyl Alcohol* **493**
Synthetic Perfumes **494**
Experiment E1 Piperonal **495**
Reaction **495**
Discussion **496**
Experimental Procedure **497**
Optional Microscale Preparation **499**
Experiment E2 Piperonal O-(2,4-Dinitrophenyl) oxime **501**
Reaction **501**
Discussion **501**
Experimental Procedure **502**
Optional Microscale Preparation **504**
Experiment E3 Piperonylonitrile **505**
Reaction **505**
Discussion **505**
Experimental Procedure **507**
SEQUENCE F *Introduction to Photochromism: The Synthesis of a Photochromic Imine* **509**
Classes of Photochromic Reactions **509**
Applications of Photochromism **511**
The Reactions of Sequence F **512**
Experiment F1 An Aldol Reaction: trans-4-Nitrochalcone **513**
Reaction **514**
Discussion **514**
Experimental Procedure **515**
Experiment F2 erythro-2,3-Dibromo-3-(4-nitrophenyl) propiophenone **518**
Reaction **518**
Discussion **518**
Experimental Procedure **520**
Experiment F3 trans-2-(4-Nitrophenyl)-3-benzoylaziridine **524**
Reaction **524**
Discussion **524**
Experimental Procedure **527**
Part A **528**
Part B **528**
Experiment F4 A Photochromic Imine: 2-exo-6-exo-2,4-Diphenyl-6-(4-nitrophenyl)-1,3-diazabicyclo [3.1.0] hex-3-ene **531**
Reactions **531**
Discussion **532**
Experimental Procedure **535**

JWCL196_fm_i-xxii.qxd 11/24/09 1:05 PM Page xx

EQA

Chapter 8

SPECTROSCOPIC IDENTIFICATION OF ORGANIC COMPOUNDS 539

Infrared Spectroscopy **539**
Introduction to Group Frequencies: Interpretation of Infrared Spectra **540**
A Survey of Group Frequencies Identified in Organic Molecules **543**
  Group Frequencies of the Hydrocarbons **544**
  Group Frequencies of Carbonyl Groups: C═O **545**
  Group Frequencies of the Heteroatom Functional Groups **547**
  Esters **548**
Infrared Spectroscopy Instrumentation and Sample Handling **551**
  Instrumentation **551**
  Sample Handling in the Infrared **552**
Nuclear magnetic resonance spectroscopy **561**
  Nuclear spin **561**
Instrumentation **562**
Chemical Shift **565**
Spin–Spin Coupling **566**
Intensities **569**
Second-Order Effects **569**
Interpretation of ^{1}H NMR Spectra **571**
^{1}H Chemical Shifts **573**
Spin–Spin Coupling **574**
  Geminal Coupling **574**
  Vicinal Coupling **575**
  Long-Range Coupling **576**
Examples of Complex, Yet First-Order, Coupling **576**
  Ethyl Vinyl Ether **576**
  Allyl Acetate **578**
^{13}C NMR Spectroscopy **581**
Two-Dimensional NMR Spectroscopy **587**
Nuclear Magnetic Resonance Sampling **589**
Ultraviolet–Visible Spectroscopy: Introduction to Absorption Spectroscopy **593**
UV–VIS Spectroscopy **595**
Application to Organic Molecules **597**
Instrumentation **604**
  The Source of Radiation **604**
  The Monochromator **605**
  Sample Compartment **606**
  The Detector **606**
  The Electronics: The Amplifier and Recorder **607**
Sample Preparation **607**
  Criteria for Choosing a Solvent **608**
Mass Spectrometry **609**
Instrumentation **611**
  Ion Source **612**
  Mass Analyzer **613**
  Detector **615**
  Tuning the Mass Spectrometer **615**
  Sample Introduction **616**
  Gas Chromatography/Mass Spectrometry (GC/MS) **616**
  Capillary Columns **616**
  Split Injection **617**
  Split/Splitless Injection **617**
Features of the Mass Spectrum **618**
  Terms **619**
  Isotope Peaks **619**
  Recognizing the Molecular Ion **621**
  Mass Spectral Interpretation **622**
Case Study: Synthesis of Methyl Benzoate **623**

Chapter 9

QUALITATIVE IDENTIFICATION OF ORGANIC COMPOUNDS 629

Organic Qualitative Analysis **629**
Preliminary Tests **631**
  Nonchemical Tests **631**
  Ignition Test **632**
Separation of Impurities **633**
Detection of Elements Other Than Carbon, Hydrogen, or Oxygen **634**
  Sodium Fusion **634**
  Sulfur **635**
  Nitrogen **636**
  The Halogens (Except Fluorine) **636**
Solubility Characteristics **638**
The Classification Tests **640**
  Alcohols **640**
  Periodic Acid: Vicinal Diols **642**
  Aldehydes and Ketones **642**
  Alkanes and Cycloalkanes: Saturated Hydrocarbons **644**
  Alkenes and Alkynes: Unsaturated

CONFIRMING PAGES aptara

  - Hydrocarbons **645**
  - Alkyl Halides **645**
  - Amides, Ammonium Salts, and Nitriles **647**
  - Amines **648**
  - Aromatic Hydrocarbons with no Functional Groups **649**
  - Carboxylic Acids **650**
  - Esters **650**
  - Ethers **651**
  - Methyl Ketones and Methyl Carbinols **651**
  - Nitro Compounds **652**
  - Phenols and Enols **653**
- Preparation of Derivatives **654**
- Carboxylic Acids **655**
  - Preparation of Acid Chlorides **655**
  - Amides **655**
  - Anilides **656**
  - Toluidides **656**
- Alcohols **656**
  - Phenyl- and α-Naphthylurethanes (Phenyl- and α-Naphthylcarbamates) **656**
  - 3,5-Dinitrobenzoates **657**
- Aldehydes and Ketones **658**
  - 2,4-Dinitrophenylhydrazones **658**
  - Semicarbazones **658**
- Amines **658**
  - Primary and Secondary Amines: Acetamides **658**
  - Primary and Secondary Amines: Benzamides **659**
  - Primary, Secondary, and Tertiary Amines: Picrates **659**
- Acid Chlorides and Anhydrides **660**
  - Amides **660**
- Aromatic Hydrocarbons **660**
  - Picrates **660**
- Nitriles **660**
  - Hydrolysis to Amides **660**
- Phenols **661**
  - α-Naphthylurethanes (α-Naphthylcarbamates) **661**
  - Bromo Derivatives **661**
- Aliphatic Hydrocarbons, Halogenated Hydrocarbons, Amides, Nitro Compounds, Ethers, and Esters **662**

Glossary **665**

Index **668**

CH$_4$apter 1

INTRODUCTION

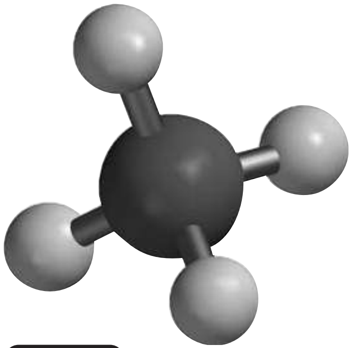

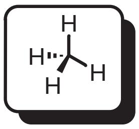

You are about to embark on a challenging adventure—the microscale organic chemistry laboratory!

Your course is going to be quite different from the conventional manner in which this laboratory has been taught in past decades. You will be learning the experimental side of organic chemistry from the microscale level. Although you will be working with very small amounts of materials, you will be able to observe and learn more organic chemistry in one year than many of your predecessors did in nearly two years of laboratory work. You will find this laboratory an exciting and interesting place to be. While we cannot guarantee it for you individually, the majority of students who went through the program during its development found the microscale organic laboratory to be a surprisingly pleasant adventure.

This textbook is centered on helping you develop skills in microscale organic laboratory techniques. Its focus is twofold. For those of you in the academic environment and involved with the introductory organic laboratory, it allows the flexibility of developing your own scaling sequence without being tied to a prescribed set of quantities. For those of you working in a research environment at the advanced undergraduate or graduate level or in the industrial area, this text will provide the foundation from which you can develop a solid expertise in microscale techniques directly applicable to your work. Working at the microscale level is substantially different from using conventional operations in the organic laboratory with multigram quantities of materials.

During the last two decades, the experimental side of organic chemistry has moved ever closer to the microscale level. This conversion started in earnest nearly thirty years ago and has been spurred on by the rapidly accelerating cost of chemical waste disposal. As we have said, you will be working with very small amounts of materials, but the techniques that you will learn and experience you will gain will allow you to accomplish more organic chemistry in the long run than many of your predecessors.

First, we want to acquaint you with the organization and contents of the text. ***With the fifth edition, a continued effort has been made to streamline the basic reference material from the text using our accompanying website (www.wiley.com/college/MOL5). Accordingly, Chapter 10W (formerly Chapter 7 of the fourth edition) in its entirety has been placed online. Throughout this edition, Chapter 10W is identified with a "W" (e.g., Chapter 10W), indicating its location online. Furthermore, an icon will be used in the margin to indicate website material that will be of interest to the user. We hope this treatment of the laboratory will make the more important aspects of the basic text easier to access and will speed your laboratory work along.*** We then give you a few words of advice, which, if they are heeded, will allow you to avoid many of the sand traps you will find as you develop microscale laboratory techniques. Finally, we wax philosophical and attempt to describe what we think you should derive from this experience.

www

www

www

After this brief introduction, the second chapter is concerned with safety in the laboratory. This chapter supplies information that will allow you to estimate

---

Chapter 1: CH$_4$, Methane
a substance of natural origin, known as Marsh Gas to the alchemists.

your maximum possible exposure to volatile chemicals used in the microscale laboratory. Chapter 2 also discusses general safety protocol for the laboratory. It is vitally important that you become familiar with the details of the material contained in this chapter; your health and safety depend on this knowledge.

The next three chapters are concerned primarily with the development of experimental techniques. Chapter 3 describes in detail the glassware employed in microscale organic chemistry: the logic behind its construction, tips on its usage, the common arrangements of equipment, and various other laboratory manipulations, including techniques for transferring microquantities of materials. Suggestions for the organization of your laboratory notebook are presented at the end of this chapter.

Chapter 4 deals with equipment and techniques for determining a number of physical properties of microscale samples. Chapter 5 is divided into nine technique sections. Detailed discussions develop the major areas of experimental technique that are used in the microscale organic laboratory.

Chapters 6, 7, and 10W contain the main experimental sections of this text. Chapter 6 is focused primarily on preparative organic chemistry at the microscale level and consists of 35 experiments. While the number of experiments has not changed with this edition, there have been changes to how the reactions are monitored and conducted. Six experiments (Experiments 5A, 5B, 7, 19B, 24A, and 32) in Chapter 6 have been modified in a way which replaces the posting of a reaction time with the task of monitoring the reaction by TLC until complete. The TLC technique is asked of the experimentalist in three more experiments (Experiments 11A, 16, and 28) in order to provide additional evidence of reaction purity upon recrystallization of the crude reaction mixture. And finally, five experiments (Experiments 7, 8, 15, 22, and 30) in Chapter 6 now have optional exercises which utilize microwave technologies. Additional selections of individual experiments can be drawn from those experiments presented in Chapter 7. Chapter 10W, which is now located online, contains a series of seven experiments of a more sophisticated nature. A number of the experiments contained in Chapters 6 and 10W are of optional scale so that you may also have the opportunity to gain some experience with experimentation at larger scales. Chapter 7 consists of a set of six sequential experiments that are essentially identical to the type of problems tackled by research chemists involved in synthetic organic chemistry. A number of these multistep procedures begin the first step in the experiment with large-scale, multigram quantities of starting material, but require microscale techniques to complete the final step or two. The use of this chapter is most appropriate in the final stages of the course, for example, the latter part of the second semester of a two-semester sequence.

Chapter 8 develops the characterization of organic materials at the microscale level by spectroscopic techniques. The chapter starts with a brief discussion of the interpretation of infrared (IR) group frequencies and is followed by a more detailed treatment of nuclear magnetic resonance (NMR) spectral data, a brief discussion of ultraviolet-visible (UV–vis) spectroscopy, and a brief introduction to the theory, experimental techniques, and applications of mass spectrometry to organic chemistry. A more detailed introduction to the theoretical basis for these spectroscopic techniques is also presented on the accompanying website.

Chapter 9 develops the characterization of organic materials at the microscale level by the use of classical organic reactions to form solid derivatives. Tables of derivative data for use in compound identification by these techniques are discussed and are included on the website as Appendix A.

A list of all the experiments grouped by reaction mechanism is given on the web-site as Appendix B.

The organization of the experimental procedures given in Chapters 6, 7, and 10W is arranged in the following fashion. A short opening statement describing the reaction to be studied is followed by the reaction scheme.

www

Generally, a brief discussion of the reaction follows, including a mechanistic interpretation. In a few cases of particularly important reactions, or where the experiment is likely to precede presentation of the topic in the classroom, a more detailed description is given. The estimated time needed to complete the work, and a table of reactant data come next. For ease in organizing your laboratory time, the experimental section is divided into four subsections: *reagents and equipment, reaction conditions, isolation of product, and purification and characterization.*

We then introduce a series of questions and problems designed to enhance and focus your understanding of the chemistry and the experimental procedures involved in a particular laboratory exercise. Finally, a bibliography offering a list of literature references is given. Although this list comes at the end of the experimental section, we view it as a very important part of the text. The discussion of the chemistry involved in each experiment is necessarily brief. We hope that you will take time to read and expand your knowledge about the particular experiment that you are conducting. You may, in fact, find that some of these references become assigned reading.

A prompt (➨) in the text indicates that experimental apparatus involved with that stage of the experiment are shown in the margin. Important comments are italicized in the text, and **Warnings** and **Cautions** are given in boxes and also indicated in the margins.

In an effort to streamline our treatment of the laboratory we have moved a considerable quantity of material from the previous editions, MOL3 and MOL4, and placed it in easily accessible form on our website (www.wiley.com/college/MOL5). An icon lets you know that supplemental material is available on the website. New to this edition is a detailed listing within the table of contents of all materials available online. We hope this format will make the more important aspects of the basic text easier to access and speed your laboratory work along.

www

GENERAL RULES FOR THE MICROSCALE LABORATORY

1. Study the experiment before you come to lab. This rule is a historical plea from all laboratory instructors. In the microscale laboratory it takes on a more important meaning. You will not survive if you do not prepare ahead of time. In microscale experiments, operations happen much more quickly than in the macroscale laboratory. Your laboratory time will be overflowing with many more events. If you are not familiar with the sequences you are to follow, you will be in deep trouble. Although the techniques employed at the microscale level are not particularly difficult to acquire, they do demand a significant amount of attention. For you to reach a successful and happy conclusion, you cannot afford to have the focus of your concentration broken by having to constantly refer to the text during the experiment. Disaster is ever present for the unprepared.

2. ALWAYS work with clean equipment. You must take the time to scrupulously clean your equipment before you start any experiment. Contaminated glass-ware will ultimately cost you additional time, and you will experience the frustration of inconsistent results and lower yields. Dirty equipment is the primary cause of reaction failure at the microscale level.

3. CAREFULLY measure the quantities of materials to be used in the experiments. A little extra time at the beginning of the laboratory can speed you on your way at the end of the session. A great deal of time has been spent optimizing the conditions employed in these experiments in order to maximize yields. Many organic reactions are very sensitive to the relative quantities of substrate (the material on which the reaction is taking place) and reagent (the reactive substance or substances that bring about the change in the substrate). After equipment contamination, the second-largest cause of failed reactions is attempting to run a reaction with incorrect quantities of the reactants present. Do not be hurried or careless at the balance.

4. Clean means DRY. Water or cleaning solution can be as detrimental to the success of a reaction as dirt or sludge in the system. You often will be working with very small quantities of moisture-sensitive reagents. The glass surface areas with which these reagents come in contact, however, are relatively large. A slightly damp piece of glassware can rapidly deactivate a critical reagent and result in reaction failure. *This rule must be strictly followed.*

5. ALWAYS work on a clean laboratory bench surface, preferably glass!

6. ALWAYS protect the reaction product that you are working with from a disastrous spill by carrying out all solution or solvent transfers over a crystallizing dish.

7. ALWAYS place reaction vials or flasks in a clean beaker when standing them on the laboratory bench. Then, when a spill occurs the material is more likely to be contained in the beaker and less likely to be found on the laboratory bench or floor.

8. NEVER use cork rings to support round-bottom flasks, particularly if they contain liquids. You are inviting disaster to be a guest at your laboratory bench.

9. ALWAYS think through the next step you are going to perform *before* starting it. Once you have added the wrong reagent, it is back to square one.

10. ALWAYS save everything you have generated in an experiment until it is successfully completed. You can retrieve a mislabeled chromatographic fraction from your locker, but not from the waste container!

THE ORGANIC CHEMISTRY LABORATORY

The confidence gained by mastering the microscale techniques described here will pay big dividends as you progress into modern-day experimental chemistry. The organic laboratory has had a reputation of being smelly, long, tedious, and pockmarked with fires and explosions; but present-day organic chemistry is undergoing a revolution at the laboratory bench. New techniques are sweeping away many of the old complaints, as an increasing fraction of industrial and academic research is being carried out at the microscale level.

This book allows the interested participant to rapidly develop the skills needed to slice more deeply into organic chemistry than ever before. The attendant benefits are greater confidence and independence in acquired laboratory techniques. The happy result is that in a microscale-based organic chemistry laboratory, you are more likely to have a satisfying encounter with the experimental side of this fascinating field of knowledge.

C$_2$H$_4$apter 2

SAFETY

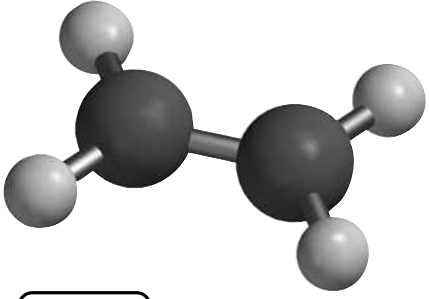

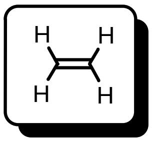

Research laboratories vary widely with respect to facilities and support given to safety. Large laboratories may have several hundred chemists and an extensive network of co-workers, supervisors, safety officers, and hazardous-waste managers. They also, according to government regulations, have an extensive set of safety procedures and detailed practices for the storage and disposal of hazardous wastes. In small laboratories, the individual chemist may have to take care of all these aspects of safety. Some laboratories may routinely deal with very hazardous materials and may run all reactions in hoods. Others may deal mainly with relatively innocuous compounds and have very limited hood facilities.

Our approach is to raise some questions to think about and to suggest places to look for further information. In this chapter, we do not present a large list of safety precautions for use in all situations; rather, we present a list of very basic precautionary measures. A bibliography at the end of the chapter offers a list of selected references. *We urge you to consult these references concerning specific safety regulations.* Many laboratories may have safety guidelines that will supercede this very cursory treatment. This chapter is no more than a starting point.

MAKING THE LABORATORY A SAFER PLACE

Murphy's law states in brief, "If anything can go wrong, it will." Although it is often taken to be a silly law, it is not. Murphy's law means that if sparking switches are present in areas that contain flammable vapors, sooner or later there will be a fire. If the glass container can move to the edge of the shelf as items are moved around or because the building vibrates, at some time it will come crashing to the floor. If the pipet can become contaminated, then the mouth pipetter will eventually ingest a contaminant.

We cannot revoke Murphy's law, but we can do a lot to minimize the damage. We can reduce the incidence of sparks and flames and flammable vapors. We can make sure that if the accident does occur, we have the means to contain the damage and to take care of any injuries that result. All of this means thinking about the laboratory environment. Does your laboratory have or enforce regulations related to important items such as eye, face, and foot protection, safety clothing, respiratory equipment, first aid supplies, fire equipment, spill kits, hoods, and compliance regulations? *Think ahead* about what could go wrong and then *plan* and *prepare* to minimize the chance of an accident and be prepared to respond when one does occur.

NATURE OF HAZARDS

The chemistry laboratory presents a wide assortment of risks. These risks are outlined briefly here so that you can begin to think about the steps necessary to make the laboratory safer:

1. **Physical hazards.** Injuries resulting from flames, explosions, and equipment (cuts from glass, electrical shock from faulty instrumentation, or improper use of instruments).

Chapter 2: C$_2$H$_4$, Ethylene
a substance of natural origin, released by ripening fruit.

2. **External exposure to chemicals.** Injuries to skin and eyes resulting from contact with chemicals that have spilled, splashed, or been left on the bench top or on equipment.
3. **Internal exposure.** Longer term (usually) health effects resulting from breathing hazardous vapors or ingesting chemicals.

REDUCTION OF RISKS

Many things can be done to reduce risks. The rules below may be absolute in some laboratories. In others, the nature of the materials and apparatus used may justify the relaxation of some of these rules or the addition of others.

1. **Stick to the procedures described by your supervisor.** This attention to detail is particularly important for the chemist with limited experience. In other cases, variation of the reagents and techniques may be part of the work.
2. **Wear approved safety goggles.** We can often recover quickly from injuries affecting only a few square millimeters on our bodies, unless that area happens to be in our eyes. Larger industrial laboratories often require that laboratory work clothes and safety shoes be worn. Wear them, if requested.
3. **Do not put anything in your mouth under any circumstances while in the laboratory.** This includes food, drinks, chemicals, and pipets. There are countless ways that surfaces can become contaminated in the laboratory. Since there are substances that must never be pipetted by mouth, one must get into the habit of *never* mouth pipetting anything.
4. **Be cautious with flames and flammable solvents.** Remember that the flame at one end of the bench can ignite the flammable liquid at the other end in the event of a spill or improper disposal. Flames must never be used when certain liquids are present in the laboratory, and flames must always be used with care. Check the *fire diamond* hazard symbol, if available.
5. **Be sure that you have the proper chemicals for your reaction.** Check labels carefully, and return unused chemicals to the proper place for storage. Be sure to replace caps on containers immediately after use. An open container is an invitation for a spill. Furthermore, some reagents are very sensitive to moisture, and may decompose if left open.
6. **Minimize the loss of chemicals to air or water and dispose of waste properly.** Some water-soluble materials may be safely disposed of in the water drains. Other wastes should go into special receptacles. Pay attention to the labels on these receptacles. Recent government regulations have placed stringent rules on industrial and academic laboratories for proper disposal of chemicals. *Severe penalties are levied on those who do not follow proper procedures.* We recommend that you consult general safety references nos. 3 and 4 at the end of the chapter.
7. **Minimize skin contact with any chemicals.** Use impermeable gloves when necessary, and promptly wash any chemical off your body. If you have to wash something off with water, use lots of it. Be sure that you know where the nearest water spray device is located.

NOTE. Do not use latex gloves. They are permeable to many chemicals and some people are allergic to them. A recommended substitute are the various grades of nitrile gloves.

8. **Do not inhale vapors from volatile materials.** Severe illness or internal injury can result.
9. **Tie back or confine long hair and loose items of clothing.** You do not want them falling into a reagent or getting near flames.
10. **Do not work alone.** Too many things can happen to a person working alone that might leave him or her unable to obtain assistance. As in swimming, the "buddy system" is safest.
11. **Exercise care in assembling glass and electrical apparatus.** All operations with glass, such as separating standard taper glassware, involve the risk that the glass may break and that lacerations or punctures may result. Seek help or advice with glassware, if necessary. Special containers should be provided for the disposal of broken glass. Electrical shock can occur in many ways. When making electrical connections, make sure that your hands, the laboratory bench, and the floor are all dry and that *you* do not complete an electrical path to ground. Be sure that electrical equipment is properly grounded and insulated.
12. **Report any injury or accident to the appropriate person.** Reporting injuries and accidents is important so that medical assistance can be obtained if necessary. It also allows others to be made aware of any safety problems; these problems may be correctable.
13. **Keep things clean.** Put unused apparatus away. Immediately wipe up or care for spills on the bench top or floor. This also pertains to the balance area and to where chemicals are dispensed.
14. **Never heat a closed system.** Always provide a vent to avoid an explosion. Provide a suitable trap for any toxic gases generated in a given reaction.
15. **Learn the correct use of gas cylinders.** Even a small gas cylinder can become a lethal bomb if not properly used.
16. **Attend safety programs.** Many industrial laboratories offer excellent seminars and lectures on a wide variety of safety topics. Pay careful attention to the advice and counsel of the safety officer.
17. **Above all, use your common sense.** Think before you act.

PRECAUTIONARY MEASURES

Know the location and operation of safety equipment in the laboratory. Locate the nearest

- Fire extinguisher
- First aid kit
- Eye wash
- Telephone
- Emergency shower
- Exit
- Fire blanket

Know where to call (have the numbers posted) for

- Fire
- Medical emergency
- Spill or accidental release of corrosive or toxic chemicals.

Know where to go

- In case of injury
- To evacuate the building

THINKING ABOUT THE RISKS IN USING CHEMICALS

The smaller quantities used in the microscale laboratory carry with them a reduction in hazards caused by fires and explosions; hazards associated with skin contact are also reduced. However, care must be exercised when working with even the small quantities involved.

There is great potential for reducing the exposure to chemical vapors, but these reductions will be realized only if everyone in the laboratory is careful. One characteristic of vapors emitted outside hoods is that they mix rapidly throughout the lab and will quickly reach the person on the other side of the room. In some laboratories, the majority of reactions may be carried out in hoods. When reactions are carried out in the open laboratory, each experimenter becomes a polluter whose emissions affect nearby people the most, but these emissions become added to the laboratory air and to the burden each of us must bear.

The concentration of vapor in the general laboratory air space depends on the vapor pressure of the liquids, the area of the solid or liquid exposed, the nature of air currents near the sources, and the ventilation characteristics of the laboratory. One factor over which each individual has control is evaporation, which can be reduced by the following practices:

- Certain liquids must remain in hoods.
- Reagent bottles must be recapped when not in use.
- Spills must be quickly cleaned up and the waste discarded.

Chemicals must be properly stored when not in use. Some balance must be struck between the convenience of having the compound in the laboratory where you can easily put your hands on it and the safety of having the compound in a properly ventilated and fire-safe storage room. Policies for storing chemicals will vary from place to place. There are limits to the amounts of flammable liquids that should be stored in glass containers, and fire-resistant cabinets must be used for storage of large amounts of flammable liquids. Chemicals that react with one another should not be stored in close proximity. There are plans for sorting chemicals by general reactivity classes in storerooms; for instance, Flinn Scientific Company includes a description of a storage system in their (2009) chemical catalog.

DISPOSAL OF CHEMICALS

Chemicals must also be segregated into categories for disposal. The categories used will depend on the disposal service available and upon federal, state, and local regulations. For example, some organic wastes are readily incinerated, while those containing chlorine may require much more costly treatment. Other wastes may have to be buried. For safety and economic reasons, it is important to place waste material in the appropriate container. In today's world, it often costs more to dispose of a chemical than to purchase it in the first place! The economic impact of waste generation and disposal is gigantic. Based upon the toxic release inventory from the EPA for the year of 2005, do you realize that the chemical industry in the United States released more than 4 billion pounds of on-site and off-site chemical waste?[1] Each year hundreds of billions of dollars is spent per year in waste treatment, control, and disposal costs!

It is our obligation as chemists to decrease the impact that hazardous chemicals have on our environment. A movement is currently underway (referred to as "Green Chemistry" or "Benign by Design") to accomplish this goal by focusing on the design, manufacture, and use of chemicals and chemical processes that have little or no pollution potential or environmental risk.

MATERIAL SAFETY DATA SHEETS

Although risks are associated with the use of most chemicals, the magnitudes of these risks vary greatly. A short description of the risks is provided by a Material Safety Data Sheet, commonly referred to as an MSDS. All participants of a laboratory experience are strongly encouraged to educate themselves on the risks, large or small, of the chemicals they are scheduled to work with while in lab. The information contained on these sheets can be obtained from a number of locations. While they are normally provided by the manufacturer or vendor of the chemical, and users are required to keep on file the MSDS of each material stored or used, data sheets can be easily obtained online.

As an example, the 1985 MSDS for acetone is shown here. This sheet was provided by the J. T. Baker Chemical Company. Sheets from other sources, especially those online[2], do provide a very illuminating comparison of what has transpired over 20+ years. Much of the information on these sheets is self-explanatory, but let's review the major sections of the acetone example.

www

Section I provides identification numbers and codes for the compound and includes a summary of the risks associated with the use of acetone. Because these sheets are available for many thousands of compounds and mixtures, there must be a means of unambiguously identifying the substance. A standard reference number for chemists is the Chemical Abstracts Service Number (CAS No.).

A quick review of the degree of risks is given by the numerical scale under Precautionary Labeling. This particular scale is a proprietary scale that ranges from 0 (very little or nonexistent risk) to 4 (extremely high risk). The National Fire Protection Association (NFPA) uses a similar scale but the risks considered are different. Other systems may use different scales, and there are some that represent low risks by the highest number! Be sure that you understand the scale being used. Perhaps some day one scale will become standard.

Section II covers risks from mixtures. Because a mixture is not considered here, the section is empty. Selected physical data are described in Section III. Section IV contains fire and explosion data, including a description of the toxic gases produced when acetone is exposed to a fire. The MSDSs are routinely made available to fire departments that may be faced with fighting a fire in a building where large amounts of chemicals are stored.

---

[1] For online access to the EPA Toxic Release Inventory, see: www.epa.gov/tri/

[2] For an online MSDS for acetone, see: www.jtbaker.com/msds/englishhtml/A0446.htm

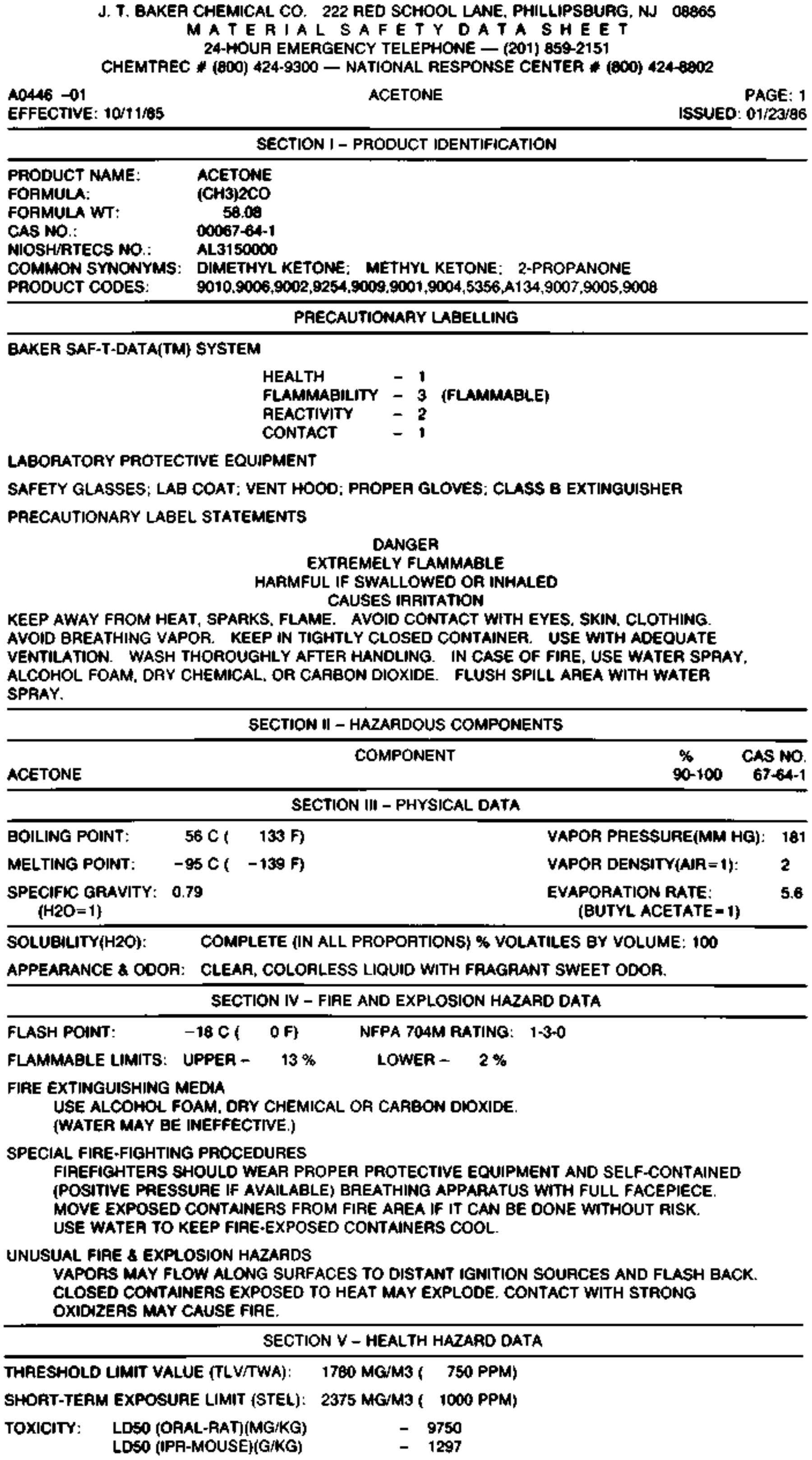

J. T. BAKER CHEMICAL CO. 222 RED SCHOOL LANE, PHILLIPSBURG, NJ 08865
MATERIAL SAFETY DATA SHEET
24-HOUR EMERGENCY TELEPHONE — (201) 859-2151
CHEMTREC # (800) 424-9300 — NATIONAL RESPONSE CENTER # (800) 424-8802

A0446 –01 ACETONE PAGE: 1
EFFECTIVE: 10/11/85 ISSUED: 01/23/86

SECTION I – PRODUCT IDENTIFICATION

PRODUCT NAME: ACETONE
FORMULA: (CH3)2CO
FORMULA WT: 58.08
CAS NO.: 00067-64-1
NIOSH/RTECS NO.: AL3150000
COMMON SYNONYMS: DIMETHYL KETONE; METHYL KETONE; 2-PROPANONE
PRODUCT CODES: 9010,9006,9002,9254,9009,9001,9004,5356,A134,9007,9005,9008

PRECAUTIONARY LABELLING

BAKER SAF-T-DATA(TM) SYSTEM

HEALTH – 1
FLAMMABILITY – 3 (FLAMMABLE)
REACTIVITY – 2
CONTACT – 1

LABORATORY PROTECTIVE EQUIPMENT

SAFETY GLASSES; LAB COAT; VENT HOOD; PROPER GLOVES; CLASS B EXTINGUISHER

PRECAUTIONARY LABEL STATEMENTS

DANGER
EXTREMELY FLAMMABLE
HARMFUL IF SWALLOWED OR INHALED
CAUSES IRRITATION

KEEP AWAY FROM HEAT, SPARKS, FLAME. AVOID CONTACT WITH EYES, SKIN, CLOTHING. AVOID BREATHING VAPOR. KEEP IN TIGHTLY CLOSED CONTAINER. USE WITH ADEQUATE VENTILATION. WASH THOROUGHLY AFTER HANDLING. IN CASE OF FIRE, USE WATER SPRAY, ALCOHOL FOAM, DRY CHEMICAL, OR CARBON DIOXIDE. FLUSH SPILL AREA WITH WATER SPRAY.

SECTION II – HAZARDOUS COMPONENTS

COMPONENT	%	CAS NO.
ACETONE	90-100	67-64-1

SECTION III – PHYSICAL DATA

BOILING POINT: 56 C (133 F)
VAPOR PRESSURE(MM HG): 181
MELTING POINT: −95 C (−139 F)
VAPOR DENSITY(AIR=1): 2
SPECIFIC GRAVITY: (H2O=1) 0.79
EVAPORATION RATE: (BUTYL ACETATE=1) 5.6

SOLUBILITY(H2O): COMPLETE (IN ALL PROPORTIONS) % VOLATILES BY VOLUME: 100

APPEARANCE & ODOR: CLEAR, COLORLESS LIQUID WITH FRAGRANT SWEET ODOR.

SECTION IV – FIRE AND EXPLOSION HAZARD DATA

FLASH POINT: −18 C (0 F) NFPA 704M RATING: 1-3-0

FLAMMABLE LIMITS: UPPER – 13 % LOWER – 2 %

FIRE EXTINGUISHING MEDIA
USE ALCOHOL FOAM, DRY CHEMICAL OR CARBON DIOXIDE.
(WATER MAY BE INEFFECTIVE.)

SPECIAL FIRE-FIGHTING PROCEDURES
FIREFIGHTERS SHOULD WEAR PROPER PROTECTIVE EQUIPMENT AND SELF-CONTAINED (POSITIVE PRESSURE IF AVAILABLE) BREATHING APPARATUS WITH FULL FACEPIECE. MOVE EXPOSED CONTAINERS FROM FIRE AREA IF IT CAN BE DONE WITHOUT RISK. USE WATER TO KEEP FIRE-EXPOSED CONTAINERS COOL.

UNUSUAL FIRE & EXPLOSION HAZARDS
VAPORS MAY FLOW ALONG SURFACES TO DISTANT IGNITION SOURCES AND FLASH BACK. CLOSED CONTAINERS EXPOSED TO HEAT MAY EXPLODE. CONTACT WITH STRONG OXIDIZERS MAY CAUSE FIRE.

SECTION V – HEALTH HAZARD DATA

THRESHOLD LIMIT VALUE (TLV/TWA): 1780 MG/M3 (750 PPM)

SHORT-TERM EXPOSURE LIMIT (STEL): 2375 MG/M3 (1000 PPM)

TOXICITY: LD50 (ORAL-RAT)(MG/KG) – 9750
LD50 (IPR-MOUSE)(G/KG) – 1297

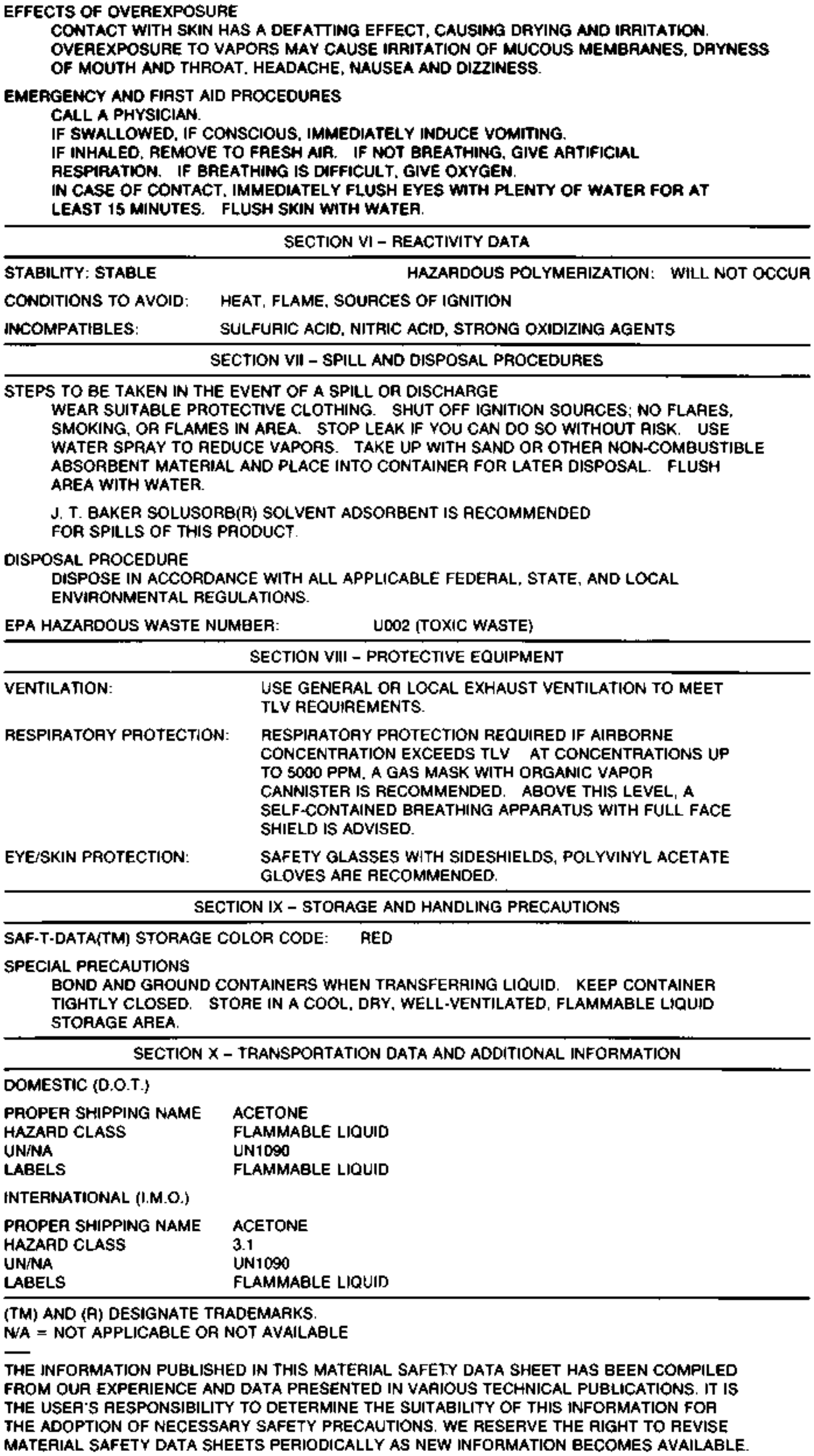

EFFECTS OF OVEREXPOSURE
CONTACT WITH SKIN HAS A DEFATTING EFFECT, CAUSING DRYING AND IRRITATION. OVEREXPOSURE TO VAPORS MAY CAUSE IRRITATION OF MUCOUS MEMBRANES, DRYNESS OF MOUTH AND THROAT, HEADACHE, NAUSEA AND DIZZINESS.

EMERGENCY AND FIRST AID PROCEDURES
CALL A PHYSICIAN.
IF SWALLOWED, IF CONSCIOUS, IMMEDIATELY INDUCE VOMITING.
IF INHALED, REMOVE TO FRESH AIR. IF NOT BREATHING, GIVE ARTIFICIAL RESPIRATION. IF BREATHING IS DIFFICULT, GIVE OXYGEN.
IN CASE OF CONTACT, IMMEDIATELY FLUSH EYES WITH PLENTY OF WATER FOR AT LEAST 15 MINUTES. FLUSH SKIN WITH WATER.

SECTION VI – REACTIVITY DATA

STABILITY: STABLE — HAZARDOUS POLYMERIZATION: WILL NOT OCCUR

CONDITIONS TO AVOID:	HEAT, FLAME, SOURCES OF IGNITION
INCOMPATIBLES:	SULFURIC ACID, NITRIC ACID, STRONG OXIDIZING AGENTS

SECTION VII – SPILL AND DISPOSAL PROCEDURES

STEPS TO BE TAKEN IN THE EVENT OF A SPILL OR DISCHARGE
WEAR SUITABLE PROTECTIVE CLOTHING. SHUT OFF IGNITION SOURCES; NO FLARES, SMOKING, OR FLAMES IN AREA. STOP LEAK IF YOU CAN DO SO WITHOUT RISK. USE WATER SPRAY TO REDUCE VAPORS. TAKE UP WITH SAND OR OTHER NON-COMBUSTIBLE ABSORBENT MATERIAL AND PLACE INTO CONTAINER FOR LATER DISPOSAL. FLUSH AREA WITH WATER.

J. T. BAKER SOLUSORB(R) SOLVENT ADSORBENT IS RECOMMENDED FOR SPILLS OF THIS PRODUCT.

DISPOSAL PROCEDURE
DISPOSE IN ACCORDANCE WITH ALL APPLICABLE FEDERAL, STATE, AND LOCAL ENVIRONMENTAL REGULATIONS.

EPA HAZARDOUS WASTE NUMBER: U002 (TOXIC WASTE)

SECTION VIII – PROTECTIVE EQUIPMENT

VENTILATION:	USE GENERAL OR LOCAL EXHAUST VENTILATION TO MEET TLV REQUIREMENTS.
RESPIRATORY PROTECTION:	RESPIRATORY PROTECTION REQUIRED IF AIRBORNE CONCENTRATION EXCEEDS TLV. AT CONCENTRATIONS UP TO 5000 PPM, A GAS MASK WITH ORGANIC VAPOR CANNISTER IS RECOMMENDED. ABOVE THIS LEVEL, A SELF-CONTAINED BREATHING APPARATUS WITH FULL FACE SHIELD IS ADVISED.
EYE/SKIN PROTECTION:	SAFETY GLASSES WITH SIDESHIELDS, POLYVINYL ACETATE GLOVES ARE RECOMMENDED.

SECTION IX – STORAGE AND HANDLING PRECAUTIONS

SAF-T-DATA(TM) STORAGE COLOR CODE: RED

SPECIAL PRECAUTIONS
BOND AND GROUND CONTAINERS WHEN TRANSFERRING LIQUID. KEEP CONTAINER TIGHTLY CLOSED. STORE IN A COOL, DRY, WELL-VENTILATED, FLAMMABLE LIQUID STORAGE AREA.

SECTION X – TRANSPORTATION DATA AND ADDITIONAL INFORMATION

DOMESTIC (D.O.T.)

PROPER SHIPPING NAME	ACETONE
HAZARD CLASS	FLAMMABLE LIQUID
UN/NA	UN1090
LABELS	FLAMMABLE LIQUID

INTERNATIONAL (I.M.O.)

PROPER SHIPPING NAME	ACETONE
HAZARD CLASS	3.1
UN/NA	UN1090
LABELS	FLAMMABLE LIQUID

(TM) AND (R) DESIGNATE TRADEMARKS.
N/A = NOT APPLICABLE OR NOT AVAILABLE

THE INFORMATION PUBLISHED IN THIS MATERIAL SAFETY DATA SHEET HAS BEEN COMPILED FROM OUR EXPERIENCE AND DATA PRESENTED IN VARIOUS TECHNICAL PUBLICATIONS. IT IS THE USER'S RESPONSIBILITY TO DETERMINE THE SUITABILITY OF THIS INFORMATION FOR THE ADOPTION OF NECESSARY SAFETY PRECAUTIONS. WE RESERVE THE RIGHT TO REVISE MATERIAL SAFETY DATA SHEETS PERIODICALLY AS NEW INFORMATION BECOMES AVAILABLE.

Health hazards are described in Section V. The entries of most significance for evaluating risks from vapors are the Threshold Limit Value (or TLV) and the Short-Term Exposure Limit (STEL). The TLV is a term used by the American Conference of Governmental Industrial Hygienists (ACGIH). This organization examines the toxicity literature for a compound and establishes the TLV. This standard is designed to protect the health of workers exposed to the vapor 8 hours a day, five days a week. The Occupational Safety and Health Administration (OSHA) adopts a value to protect the safety of workplaces in the United States. Their value is termed the Time-Weighted Average (TWA) and in many cases is numerically equal to the TLV. The STEL is a value not to be exceeded for even a 15-minute averaging time. TLV, TWA, and STEL values for many chemicals are summarized in a small handbook available from the ACGIH (2000); they are also collected in the *CRC Handbook of Chemistry and Physics.*

The toxicity of acetone is also described in terms of the toxic oral dose. In this case, the LD_{50} is the dose that will cause the death of 50% of the mice or rats given that dose. The dose is expressed as milligrams of chemical per kilogram of body weight of the subject animal. The figures for small animals are often used to estimate the effects on humans. If, for example, we used the mouse figure of 1297 mg/kg and applied it to a 60-kg chemist, a dose of 77,820 mg (~98.5 mL) would kill 50% of the subjects receiving that dose. As a further example, chloroform has an LD_{50} of 80 mg/kg. For our 60-kg chemist, a dose of 4800 mg (~3 mL) would be fatal for 50% of these cases. The effects of exposure of skin to the liquid and vapor are also described.

Section VI describes the reactivity of acetone and the classes of compounds with which it should not come in contact. For example, sodium metal reacts violently with a number of substances (including water) and should not come in contact with them. Strong oxidizing agents (such as nitric acid) should not be mixed with organic compounds (among other things). The final sections (Sections VII–X) are self-explanatory.

ALTERNATE SOURCES OF INFORMATION

Similar information in a more compact form can be found in the *Merck Index* (Merck). This basic reference work provides information on the toxicity of many chemicals. It often refers one to the *NIOSH Pocket Guide to Chemical Hazards* (National Institute for Occupational Safety and Health). The *Merck Index* also supplies interesting information about the common uses of the chemicals listed, particularly related to the medical area. References to the chemical literature are also provided. The *CRC Handbook of Chemistry and Physics,* which is updated each year, contains a wide range of data (located in tables) in the area of health, safety, and environmental protection. It also includes directions for the handling and disposal of laboratory chemicals. Your laboratory should have a copy of this work. Most chemical supply houses now label their containers with data showing not only the usual package size, physical properties, and chemical formula, but also pictures or codes showing hazard information. Some include a pictogram (for example, see the newer Aldrich Chemical labels on their bottles). The J. T. Baker Company uses the Baker SAF-T-DATA System.

ESTIMATING RISKS FROM VAPORS

Other things (availability, suitability) being equal, one would, of course, choose the least toxic chemical for a given reaction. Some very toxic chemicals play very important roles in synthetic organic chemistry, and the toxicity of the chemicals in common use varies greatly. Bromine and benzene have TLVs of 0.7 and 30 mg/m^3, respectively, and are at the more toxic end of the spectrum of chemicals routinely used. Acetone has a TLV of 1780 mg/m^3. These representative figures do not mean that acetone is harmless or that bromine cannot be used. In general, one should exercise care at all times (make a habit of good laboratory practice) and should take special precautions when working with highly toxic materials.

The TLV provides a simple means to evaluate the relative risk of exposure to the vapor of any substance used in the laboratory. If the quantity of the material evaporated is represented by m (in milligrams/hour) and the TLV is expressed by L (milligrams per cubic meter), a measure of relative risk to the vapor is given by m/L. This quantity represents the volume of clean air required to dilute the emissions to the TLV. As an example, the emission of 1 g of bromine and 10 g of acetone in one hour leads to the values of m/L of 1400 m^3/hour (h) for the bromine and 5.6 m^3/h for acetone. These numbers provide a direct handle on the *relative* risks from these two vapors. It is difficult to assess the absolute risk to these vapors without a lot of information about the ventilation characteristics of the laboratory. If these releases occur within a properly operated hood, the threat to the worker in the laboratory is probably very small. (However, consideration must be given to the hood exhaust.)

Exposure in the general laboratory environment can be assessed if we assume that the emissions are reasonably well mixed before they are inhaled and if we know something about the room ventilation rate. The ventilation rate of the room can be measured by a number of ways.[3] Given the ventilation rate, it might be safe to assume that only 30% of that air is available for diluting the emissions. (This accounts for imperfect mixing in the room.) The effective amount of air available for dilution can then be compared with the amount of air required to dilute the chemical to the TLV.

Let us continue our example. Suppose that the laboratory has a volume of 75 m^3 and an air exchange rate of 2 air changes per hour. This value means that $(75\ \text{m}^3)(2/\text{h})(0.3) = 45\ \text{m}^3/\text{h}$ are available to dilute the pollutants. There may be enough margin for error to reduce the acetone concentration to a low level (5.6 m^3/h is required to reach the TLV), but use of bromine should be restricted to the hood. An assessment of the accumulative risk of several chemicals is obtained by adding the individual m/L $\left(\frac{\text{mg/h}}{\text{mg/m}^3}\right)$ values.

The m/L figures may also be used to assess the relative risk of performing the experiment outside a hood. Since m/L represents the volume of air for each student, this may be compared with the volume of air actually available for each student. If the ventilation rate for the entire laboratory is Q (in cubic meters per minute) for a section of n students meeting for t minutes, the volume for each student is kQt/n cubic meters. Here k is a mixing factor that allows for the fact that the ventilation air will not be perfectly mixed in the laboratory before it is exhausted. In a reasonable worst-case mixing situation a k value of 0.3 seems reasonable. Laboratories with modest ventilation rates supplied by

---

[3] Butcher, S. S.; Mayo, D. W.; Hebert, S. M.; Pike, R. M.,"Laboratory Air Quality, Part I"; *J. Chem. Educ.* **1985,** *62*, A238; and"Laboratory Air Quality, Part II"; *J. Chem. Educ.* **1985,** *62*, A261.

JWCL196_c02_005-017.qxd 11/23/09 8:31 AM Page 14

EQA

15–20 linear feet of hoods can be expected to provide 30–100 m^3 per student over a 3-h laboratory period if the hoods are working properly. Let us take the figure of 50 m^3 per student as an illustration. If the value of m/L for a compound (or a group of compounds in a reaction) is substantially less than 50 m^3, it may be safe to do that series of operations in the open laboratory. If m/L is comparable to or greater than 50 m^3, a number of options are available: (1) Steps using that compound may be restricted to a hood. (2) The instructional staff may satisfy themselves that much less than the assumed value is actually evaporated under conditions present in the laboratory. (3) The number of individual repetitions of this experiment may be reduced. The size of the laboratory section can be reduced or the experiment may be done in pairs or trios.

Conducting reactions in a hood does not automatically convey a stamp of safety. Hoods are designed to keep evaporating chemicals from entering the general laboratory space. For hoods to do their job, there must be an adequate flow of air into the hood, and this air flow must not be disturbed by turbulence at the hood face. A frequently used figure of merit for hood operation is the face velocity of 100 ft/min. This is an average velocity of air entering the hood opening. (In the event that your lab does not have monitoring systems already housed within the hoods, instruments for measuring flow rate are available and can be purchased from most major equipment suppliers.) Even with a face velocity of 100 ft/min, vapors can be drawn out of an improperly designed hood simply by people walking by the opening, or by drafts from open windows.

Hood performance should be checked at regular intervals. The face velocity will increase as the front hood opening is decreased. If an adequate face velocity cannot be maintained with a front opening height of 15 cm, use of the hood for carrying out reactions will be limited. A low face velocity may indicate that the fans and ductwork need cleaning, that the exhaust system leaks (if it operates under lower than ambient pressure), or that the supply of makeup air is not adequate. When the hood system is properly maintained, the height of the hood opening required to provide an adequate face velocity is often indicated with a sticker.

Hoods are often used for storage of volatile compounds. A danger in this practice is that the hood space can become quickly cluttered, making work in the hood difficult, and the air flow may be disturbed. Of course, hoods being used for storage must never be turned off.

MICROWAVE SAFETY*

M

Scientific microwave apparatus is designed for preparative chemistry and is built with safety in mind. *Domestic (household) microwave ovens should not be used for preparative chemistry*. When employing microwave heating, all the safety precautions that are taken when performing a reaction using conventional heating should be adhered to, particularly the fact that reaction vessels can be hot and, when sealed, residual pressure needs to be released carefully at the end of the reaction. There are also some safety precautions that are specific to reactions using microwave heating:

- Adherence to the microwave manufacturer's user manual and guidelines is essential.

---

*This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

CONFIRMING PAGES aptara

- Before running reactions in sealed vessels, it is prudent to check the reaction vessels for cracks or any other signs of damage prior to use.
- Only fill the reaction vessels to manufacturer's specifications; do not overfill the reaction vessels. An approximate "rule of thumb" is to fill the vessel to no further than half its capacity.
- Only seal a closed reaction vessel with the manufacturer's recommended cap. These caps are designed to vent and reseal in the case of an overpressurization during a reaction.
- If the cap is a twist-on type, be sure to use the appropriate tool to tighten the cap to the manufacturer-specified torque.
- It is important to monitor temperature and pressure profiles during the course of a reaction and to set safety limits before starting. A reaction mixture can be heated gradually to the set temperature or, in some cases, chemists prefer to heat up the contents as rapidly as possible. In the case of the latter, care needs to be taken to ensure that the temperature and pressure do not rise uncontrollably. As a result, it is best to use a low initial microwave power.
- Before performing a reaction at elevated temperatures, chemists should consider carefully the stability of the reagents and solvents they use at these temperatures.
- It is important to ensure efficient stirring, especially when using heterogeneous reaction mixtures and metal catalysts or reagents since localized heating can occur, resulting in some cases in melting of the vessel walls and, if under pressure, failure of the vessel.
- The stir bars used for agitation should not be of exactly 3 cm in length since this equates to ¼ wavelength of a microwave at 2450 MHz and thus acts as an antenna.
- Upon completion of a microwave run, the microwave unit will start a cooling process. In the case of the monomode microwave, the pressure sensor will release when the tube is cool enough to handle (50°C). With multimode microwave units, the apparatus is set to cool for a period of time but, at the end of this, the reaction vessels may still be hot. Check the temperature before removing the reaction vessels.
- When opening sealed vessels at the end of a reaction, be sure to point the vessel away from your face and any other person, preferably doing so in a hood. Any remaining pressure will release as soon as the cap is removed. If the tube is very warm, cool it in an ice bath before removing the cap.

CONCLUDING THOUGHTS

This brief chapter touches only a few of the important points concerning laboratory safety. The risk from vapor exposure is discussed in some detail, but other risks are treated briefly. Applications in some laboratories may involve reactions with a risk from radiation or infection or may involve compounds that are unstable with respect to explosion. The chemist must be aware of the potential risks and must be prepared to go to an appropriate and detailed source of information, as needed. The references cited here represent a small fraction of the safety data, texts, and journals available on this subject. It is highly recommended that the library and/or laboratory at your institution have at least this minimal selection. Of course, the selections should be kept up to date!

QUESTIONS

2-1. After bookmarking a reputable MSDS URL, locate a chemical of your choice and print out the data. If the information is not available, go to your stock room and request a copy of the MSDS. Underline on the sheet the CAS No., solubility data, fire and explosion data, reactivity data, and what protective equipment is required when using this chemical. Does your laboratory meet the safety regulations required to use this chemical? Why or why not?

2-2. Think and describe what you would do in each of the following situations which could happen in your laboratory.

(a) You are working at your station and the 100-mL round-bottom flask in which you are running a reaction in ether solvent suddenly catches fire.

(b) The person working across the laboratory bench from you allows hydrogen chloride gas to escape from his or her apparatus.

(c) A reagent bottle is dropped, spilling concentrated sulfuric acid.

(d) A hot solution "bumps," splashing your face.

2-3. You are working in the laboratory using 3.0 mL of benzene in an extraction procedure. An alternative to benzene is toluene. However, three times more toluene is required to perform the extraction. The isolation of the desired product from the extraction solution requires evaporation of the solvent (benzene or toluene). This takes 0.5 h to complete. Calculate the relative risks of using these two solvents. Which solvent would you use and why?

2-4. A laboratory has four hoods; each is 39 in. wide. When the hood door is open to a height of 8 in. and the hoods are operating, the average air velocity through the hood face is 170 ft/min.

(a) Evaluate the total ventilation rate for this room, assuming that there are no other exhausts.

(b) The laboratory is designed for use by 30 students. Evaluate the air available per student if the mixing factor is 0.3 and the experiments last for 3 h.

(c) An experiment is considered in which each student would be required to evaporate 7 mL of methylene chloride (CH_2Cl_2). Estimate the average concentration of methylene chloride. Look up the TLV or the TWA for methylene chloride and consider how the evaporation might be performed.

2-5. An experiment is considered in which 1 mL of diethylamine would be used by each student. The ventilation rate for the laboratory is 5 m^3/min. Look up the TLV (or TWA) for diethylamine, $(C_2H_5)_2NH$. What restrictions might be placed on the laboratory to keep the average concentration, over a 3-h period, less than one-third of the TWA? Assume a mixing factor of 0.3.

GENERAL SAFETY REFERENCES

1. ACS Committee on Chemical Safety, *Safety in Academic Chemical Laboratories, Vol. 1: Accident Prevention for College and University Students; Vol. 2: Accident Prevention for Faculty and Administrators,* 7th ed.; American Chemical Society: Washington, DC, 2003.
2. *Handbook of Laboratory Safety,* 5th ed.; Furr, A. K. Jr., Ed.; CRC Press: Boca Raton, FL, 2000.
3. Committee on Prudent Practices for Handling, Storage, and Disposal of Chemicals in Laboratories, National Research Council, *Prudent Practices for Handling Chemicals in Laboratories;* National Academy Press: Washington, DC, 1995.
4. Armour, M. A. *Hazardous Laboratory Chemicals Disposal Guide,* 3rd ed.; CRC Press: Boca Raton, FL, 2003.
5. *Working Safely with Chemicals in the Laboratory,* 2nd ed.; Gorman, C. E., Ed.; Genium: Schenectady, NY, 1995.
6. Alaimo, R. J., Ed. *Handbook of Chemical Health and Safety,* 1st ed., Oxford University Press: New York, 2001.
7. *The Sigma–Aldrich Library of Regulatory and Safety Data;* Lenga, R. E.; Votoupal, K. L., Eds.; Aldrich Chemical Co., Milwaukee, WI, 1992.

8. Young, J. A., *Improving Safety in the Chemical Laboratory: A Practical Guide,* 2nd ed.; Wiley: New York, 1991.

9. American Chemical Society, *Less Is Better (Laboratory Chemical Management for Waste Reduction)*, American Chemical Society: Washington, DC, 2008.

10. Verschueren, K., *Handbook of Environmental Data on Organic Chemicals,* 3rd ed.; Van Nostrand Reinhold: New York, NY, 1996.

11. Lewis, R. J., Sr., *Hazardous Chemicals Desk Reference,* 6th ed.: Wiley-Interscience: New York, 2008.

12. *NIOSH Pocket Guide to Chemical Hazards,* National Institute for Occupational Safety and Health, U. S. Government Publication Office: Washington, DC, 2008. A CD-ROM version can be obtained from the NIOSH Publications Office (http://wwwn.cdc.gov/pubs/niosh.aspx).

13. *OSHA Regulated Hazardous Substances,* Vols. I and II, Noyes Data Corp.: Park Ridge, NJ, 1990.

14. Lund, G.; Sansone, E. B.“Safe Disposal of Highly Reactive Chemicals”; *J. Chem. Educ.* **1994,** *71,* 972.

15. It is also recommended that one refer to the numerous articles on safety that appear regularly in the *Journal of Chemical Education* (http://jchemed.chem.wisc.edu/) and *The Chemical Educator.*

BIBLIOGRAPHY

ACGIH. *Threshold Limit Values and Biological Exposure Indices.* Available from ACGIH, Kemper Woods Center, 1330 Kemper Meadow Drive, Cincinnati, OH 45240.

Aldrich, Catalog Handbook of Fine Chemicals, 1001 W. St. Paul Ave., Milwaukee, WI, 2009–2010.

Anastas, P. T.; Farris, C.A., Eds., *“Benign by Design—Alternate Synthetic Design for Pollution Prevention,”* ACS Symposium Series 577, American Chemical Society; Washington, DC, 1994.

Anastas, P. T.; Williamson, T.C., Eds.,*“Green Chemistry—Designing Chemistry for the Environment,”* ACS Symposium Series 626, American Chemical Society: Washington, DC, 1996.

———, *Green Chemistry: Frontiers in Benign Chemical Synthesis and Processes,* Oxford University Press: New York, 1998.

Mayo, D. W.; Hebert, S. M.; Pike, R. M.,“Laboratory Air Quality, Part I”; *J. Chem. Educ.* **1985,** *62,* A238;“Laboratory Air Quality, Part II”; *J. Chem. Educ.* **1985,** *62*, A261.

Flinn Scientific Company, *Chemical Catalog/Reference Manual* (2009). Available from Flinn Scientific Co., P.O. Box 219, Batavia, IL 60510.

Handbook of Chemistry and Physics, 89th ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 2008–2009.

The Merck Index, 14th ed.; Budavari, S., Ed.; Merck Research Laboratories Publications: White-house Station, NJ, 2008. For online access to five different platforms of the Merck Index, see: www.merckbooks.com/mindex/online/html

Mollinelli, R. P.; Reale, M. J.; Freudenthal, R. I., *Material Data Safety Sheets,* Hill & Gernett: Boca Raton, FL, 1992.

C_3H_4apter 3

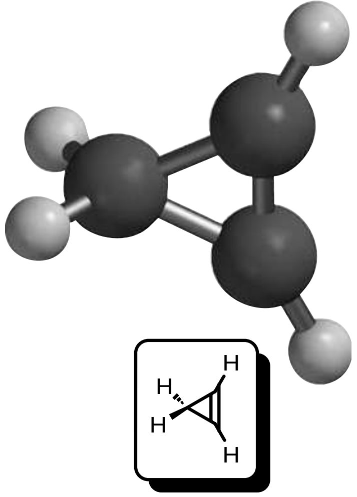

INTRODUCTION TO MICROSCALE ORGANIC LABORATORY EQUIPMENT AND TECHNIQUES

We begin this chapter with a description of the standard pieces of glassware that are generally employed in a microscale laboratory. Modern standard taper glassware is particularly convenient to use and gives the student a sense of the flavor of the research laboratory. It is not essential, however, for the experimental work in an instructional laboratory, and many courses use glassware with alternative connectors. We describe the standard taper glassware as just one example of microscale equipment that is available. The operations carried out in the laboratory will be very similar or identical if, for example, a plastic connector is used to assemble the experimental setup. We next consider a series of standard experimental apparatus setups that use this equipment, and present a short discussion of the role that they play in the laboratory. We end the chapter with a set of laws that govern how one operates in a microscale laboratory (the rules are a bit different than those for a macroscale laboratory) and a set of guidelines for recording your experimental data. The basic individual pieces of equipment are shown in Figures 3.1 to 3.7.

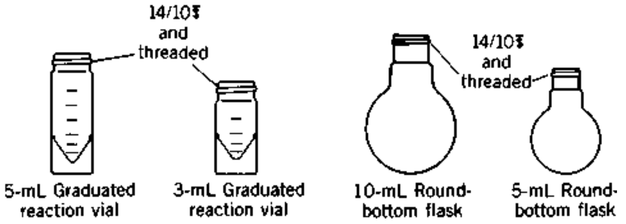
14/10 ₮ and threaded
14/10 ₮ and threaded
5-mL Graduated reaction vial
3-mL Graduated reaction vial
10-mL Round-bottom flask
5-mL Round-bottom flask

Figure 3.1 Reaction flasks.

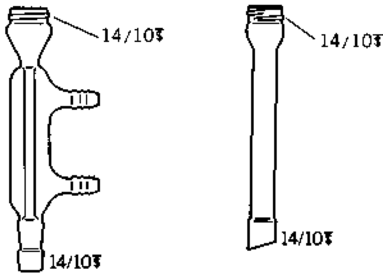
14/10 ₮
14/10 ₮
14/10 ₮
14/10 ₮

Figure 3.2 Condensers.

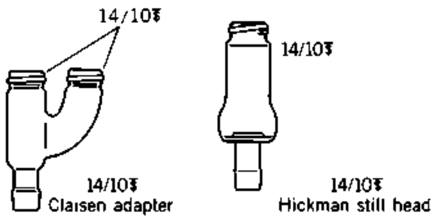
14/10 ₮
14/10 ₮
14/10 ₮
Claisen adapter
14/10 ₮
Hickman still head

Figure 3.3 Distillation heads.

Chapter 3: C_3H_4, Cyclopropene
Demyanov and Doyarenko (1922).

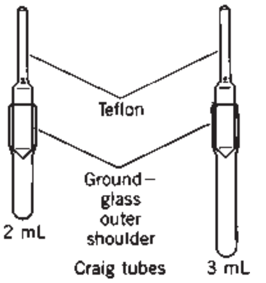
Teflon
Ground-glass outer shoulder
2 mL
Craig tubes
3 mL

Figure 3.4 **Recrystallization tubes.**

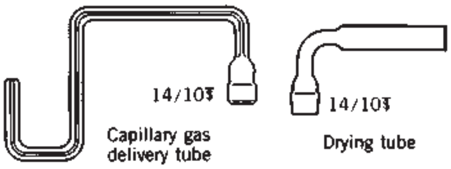
14/10₮
Capillary gas delivery tube
14/10₮
Drying tube

Figure 3.5 **Miscellaneous items.**

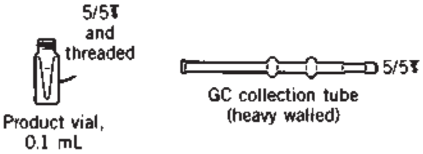
5/5₮ and threaded
Product vial, 0.1 mL
GC collection tube (heavy walled)
5/5₮

Figure 3.6 **Gas chromatographic fraction collection items.**

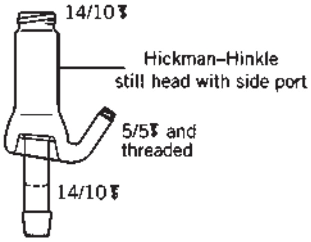
14/10₮
Hickman–Hinkle still head with side port
5/5₮ and threaded
14/10₮

Figure 3.7 **Hickman–Hinkle distillation column.**

MICROGLASSWARE EQUIPMENT

Standard Taper Joints

Standard taper ground-glass joints are the common mechanism for assembling all conventional research equipment in the organic laboratory. The symbol ₮ is commonly used to indicate the presence of this type of connector. Normally, ₮ is either followed or preceded by #/#. The first # refers to the maximum inside diameter of a female (outer) joint or the maximum outside diameter of a male (inner) joint, measured in millimeters. The second number corresponds to the total length of the ground surface of the joint (Fig. 3.8). The advantage of this type of connection is that if the joint surfaces are lightly greased, a vacuum seal is achieved. One of the drawbacks of using these joints is that contamination of the reacting system readily occurs if the solvents present in the reaction vessel dissolve the grease. In small-scale reactions this contamination can be particularly troublesome.

The small joints used in the microscale experimental organic laboratory, however, have the ease of assembly and physical integrity of research-grade, standard taper, ground-glass joints along with a number of important additional features. The joint dimensions are usually ₮ 14/10. The conical vials in which most microscale reactions are carried out use this type of connecting system. Note that in addition to being ground to a standard taper on the inside surface of the throat of the vial, these vials also have a screw thread on the outside surface (Fig. 3.9).

This arrangement allows a standard taper male joint to be sealed to the reaction flask by a septum-type (open) plastic screw cap. The screw cap applies compression to a silicone rubber retaining O-ring positioned on the shoulder of the male joint (Fig. 3.10). The compression of the O-ring thereby achieves a greaseless gas-tight seal on the joint seam, while at the same time clamping the two pieces of equipment together. The ground joint provides both protection from intimate solvent contact with the O-ring and mechanical stability to the connection. The use of this type of connector leads to a further bonus during construction of an experimental setup. Because the individual sections are small, light, and firmly sealed together, the entire arrangement often can be mounted

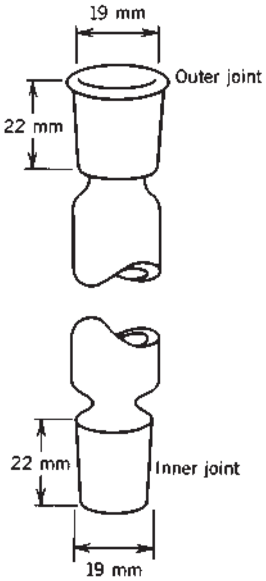
19 mm
Outer joint
22 mm
22 mm
Inner joint
19 mm

Figure 3.8 **Standard taper joints (₮).** *(From Zubrick, James W.* The Organic Chem Lab Survival Manual, *7th ed.; Wiley: New York, 2008. Reprinted by permission of John Wiley & Sons, Inc., New York.)*

Figure 3.9 **Threaded female joint.**

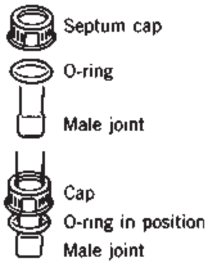
Septum cap
O-ring
Male joint
Cap
O-ring in position
Male joint

Figure 3.10 **Male joint with septum cap and O-ring.**

on the support rack by a single clamp. In conventional systems it is often necessary to use at least two clamps. This can easily lead to strain in the glass components unless considerable care is taken in the assembly process. Clamp strain is one of the major sources of experimental glassware breakage. The ability to single-clamp most microscale setups effectively eliminates this problem.

NOTE. When ground-glass joint surfaces are grease free it is important to disconnect joints soon after use (particularly with basic solutions) or they may become locked or "frozen" together.

Joints of the size employed in these microscale experiments, however, are seldom a problem to separate if given proper care (*keep them clean!*).

Conical Vials

Both the conical vials (3 and 5 mL) and the round-bottom flasks are designed to be connected via an O-ring compression cap installed on the male joint of the adjacent part of the system (see Fig. 3.1).

Condensers

Two types of condensers (air condensers and water-jacketed condensers) are available; in most cases the water-jacketed condenser can work well as an air condenser. Condensers are usually attached to 14/10 ₸-jointed reaction flasks. The upper female joints allow connection of the condenser to the 14/10 ₸ drying tube and the 14/10 ₸ capillary gas delivery tube (see Fig. 3.2).

Distillation Heads

The simple Hickman still is used with an O-ring compression cap to carry out semi-micro simple or crude fractional distillations. The Hickman–Hinkle spinning band still uses a 3-cm fractionating column and routinely develops between five and six theoretical plates. The Hickman–Hinkle still is currently available with 14/10 ₸ joints and can be conveniently operated with the 14/10 ₸ 3- and 5-mL conical vials (see Figs. 3.1, 3.3, and 3.7). The still head is also available with an optional sidearm collection port.

Recrystallization Tubes

Craig tubes are a particularly effective method for recrystallizing small quantities of reaction products. These tubes possess a nonuniform ground joint in the outer section. The substitution of Teflon for glass in the head makes these systems quite durable and much less susceptible to breakage during centrifugation (see Fig. 3.4).

Miscellaneous Items

The Claisen head (see Fig. 3.3) is often used to facilitate the syringe addition of reagents to closed moisture-sensitive systems (such as Grignard reactions) via a septum seal in the vertical upper joint. This joint can also function to position the thermometer (using an adapter) in the well of a Hickman–Hinkle still (see Fig. 3.15). The Claisen adapter is also used to mount the drying tube in a protected position remote from the reaction chamber. The drying tube, in turn, is used to protect moisture-sensitive reaction components from atmospheric water vapor, while allowing a reacting system to remain unsealed. The capillary gas delivery tube is employed in transferring gases formed during reactions to storage containers (see Fig. 3.5 and Chapter 3W, Fig. 3.11W).

www

Gas Chromatographic Fraction Collection Items

For fraction collection the gas chromatographic (GC) collection tube is connected directly to the exit port of the GC detector through a stainless steel standard taper adapter. The collected sample is then transferred to a 0.1-mL conical vial for storage. The system is conveniently employed in the resolution and isolation of two-component mixtures (see Fig. 3.6).

STANDARD EXPERIMENTAL APPARATUS

Heating and Stirring Arrangements

It is important to be able to carry out microscale experiments at accurately determined temperatures. Very often, transformations are successful, in part, because of the ability to maintain precise temperature control. In addition, many reactions require reactants to be intimately mixed to obtain a substantial yield of product. Therefore, the majority of the reactions you perform in this laboratory will be conducted with rapid stirring of the reaction mixture.

Sand Bath Technique—Hot Plate Calibration

A most convenient piece of equipment for heating or stirring or for performing both operations simultaneously on a microscale level is the hot-plate–magnetic stirrer. Heat transfer from the hot surface to the reaction flask is generally accomplished with a crystallizing dish containing a *shallow* layer of sand that can conform to the size and shape of the particular vessel employed. The temperature (external) of the system is monitored by a thermometer embedded in the sand near the reaction vessel.

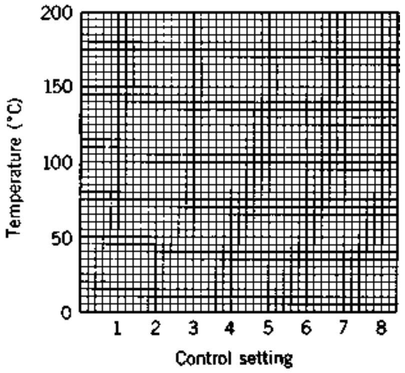

Figure 3.11 **Plot your bath and/or vial temperature (°C) versus hot-plate control setting.**

A successful procedure for determining the temperature inside the vial relative to the bath temperature is to mount a second thermometer in a vial containing 2 mL of high-boiling silicone oil. The vial temperature is then measured at various sand-bath temperatures and the values are entered on graphs of vial temperatures versus hot-plate settings and bath temperatures versus hot-plate settings (see Fig. 3.11 and Chapter 3W, Fig. 3.5W) for your particular hot-plate system (see also section on Metal Heat-Transfer Devices, p. 22). These data will save considerable time when you bring a reaction system to operating temperature. When you first enter the laboratory, it is advisable to adjust the temperature setting on the hot-plate stirrer with the heating device, or bath, in place. The setting is determined from your control setting–temperature calibration curve.This procedure will allow the heated bath to reach a relatively constant temperature by the time it is required. You will then be able to make small final adjustments more quickly, if necessary.

www

NOTE. Heavy layers of sand act as an insulator on the hot-plate surface, which can damage the heating element at high temperature settings. When temperatures over 150 °C are required, it is especially important to use the minimum amount of sand.

Recording the weight of sand used and the size of the crystallizing dish will help to make the graph values more reproducible.

The high sides of the crystallizing dish protect the apparatus from air drafts, and so the dish also operates somewhat as a hot-air bath. Heating can be made even more uniform by covering the crystallizing dish with aluminum foil (see Fig. 3.12 and Chapter 3W, Fig. 3.1W). This procedure works well, but is a bit awkward and is required in only a few instances.

The insulating properties of sand provide a readily available variable heat source because the temperature of the sand is higher deeper in the bath; thus, the depth of sand used in the bath is exceedingly important. **The depth should always be kept to a minimum, in the range of 10–15 mm.** Finally, sand baths offer a significant safety advantage over oil baths. Individual grains of sand are so small that they have little heat capacity and thus are less likely to burn the chemist in the event of a spill.

Metal Heat-Transfer Devices

An alternative to the sand bath is a heat-transfer system that employs copper tube plates or aluminum metal blocks drilled to accommodate the different reaction vials and flasks (Chapter 3W, Fig. 3.3W).

Stirring

Stirring the reaction mixture in a conical vial is carried out with Teflon-coated magnetic spin vanes, and in round-bottom flasks with Teflon-coated magnetic stirring bars (see Fig. 3.12 and Chapter 3W, Fig. 3.1W). It is important to put the reaction flask as close to the center and to the bottom surface of the crystallizing dish as possible when using magnetic stirring. This arrangement is a good practice in general, as it leads to using the minimum amount of sand needed in a sand bath.

If the reaction does not require elevated temperatures, but needs only to be stirred, the system can be assembled without the heat-transfer device (sand bath or metal plate). Some stirred reactions, on the other hand, require cooling.

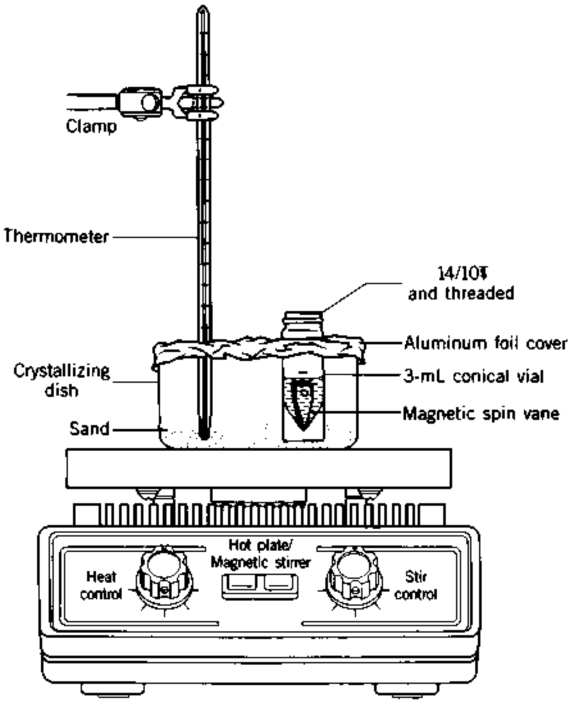

Figure 3.12 Hot-plate–magnetic stirrer with sand bath and reaction vial.

In these cases a crystallizing dish filled with ice water, or with ice water and salt, if lower temperatures are called for, will provide the correct environment.

Reflux Apparatus

To bring about a successful reaction between two substances, it is often necessary to mix the materials together intimately and to maintain a specific temperature. The mixing operation is conveniently achieved by dissolving the materials in a solvent in which they are mutually soluble. If the reaction is carried out in solution under reflux conditions, the choice of solvent can be used to control the temperature of the reaction. Many organic reactions involve the use of a reflux apparatus in one arrangement or another.

What do we mean by *reflux?* The term means to "return," or "run back." This return is exactly how the reflux apparatus functions. When the temperature of the reaction system is raised to the solvent's boiling point (constant temperature), all vapors are condensed and returned to the reaction flask or vial; this operation is not a distillation and the liquid phase remains at a stable maximum temperature. In microscale reactions, two basic types of reflux condensers are utilized: the air-cooled condenser, or air condenser (Chapter 3W, Fig. 3.6W), and the water-jacketed condenser (see Fig. 3.13 and Chapter 3W, Fig. 3.7W). The air condenser condenses solvent vapors on the cool vertical wall of an extended glass tube that dissipates the heat by contact with the laboratory room air. This arrangement functions quite effectively with liquids boiling above 150 °C. Indeed, a simple test tube can act as a reaction chamber and air condenser all in one unit, and many simple reactions can be most easily carried out in test tubes.

www

www

Air condensers can occasionally be used with lower boiling systems; however, the water-jacketed condenser is more often employed in these situations. The water-jacketed condenser employs flowing cold water to remove heat from the vertical column and thus facilitate vapor condensation. It is highly effective at condensing vapor from low-boiling liquids.

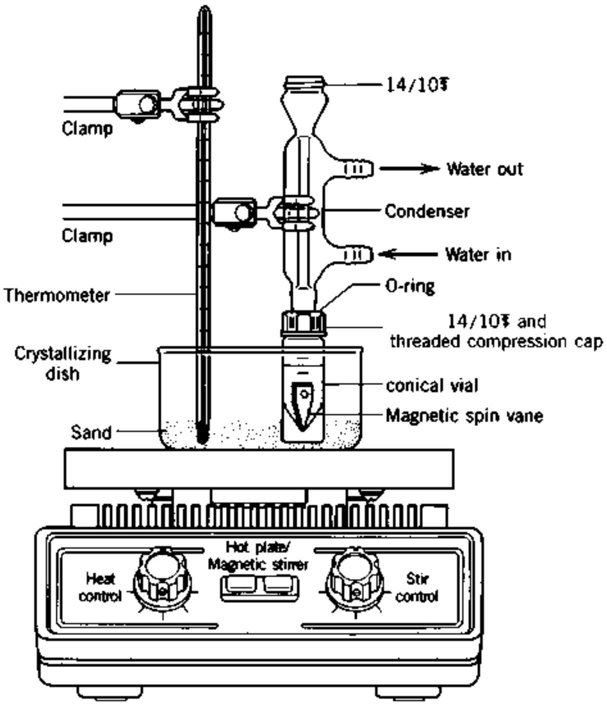

Figure 3.13 Water-jacketed condenser with conical vial, arranged for heating and magnetic stirring.

Both styles of condensers accommodate various sizes of reaction flasks and are available with 14/10 ₮ standard taper joints. The tops of both condenser columns have a female 14/10 ₮ joint.

In refluxing systems that do not require significant mixing or agitation, the stirrer (magnetic spin vane or bar) usually is replaced by a "boiling stone." These sharp-edged stones possess highly fractured surfaces that are very efficient at initiating bubble formation as the reacting medium approaches the boiling point. The boiling stone acts to protect the system from disastrous boilovers and also reduces "bumping."(Boiling stones should be used only once and must **never** be added to a hot solution. In the first case, the vapor cavities become filled with liquid upon cooling, and thus a boiling stone becomes less effective after its first use. In the second case, **adding the boiling stone to the hot solution may suddenly start an uncontrollable boilover**).

Distillation Apparatus

Distillation is a laboratory operation used to separate substances that have different boiling points. The mixture is heated, vaporized, and then condensed; the early fractions of condensate are enriched in the more volatile component. Unlike the reflux operation, in distillations none, or only a portion, of the condensate is returned to the flask where vaporization is taking place. Many distillation apparatus have been designed to carry out this basic operation. They differ mainly in small features that are used to solve particular types of separation problems. In several of the microscale experiments contained in Chapters 6, 7, and 10W *semimicroscale* distillations are required. In carrying out these distillations the choice of still depends to a large degree on the difficulty of the separation required (generally, how close are the boiling points in the mixture to be separated?).

www

The Hickman still head (Fig. 3.14) is ideally suited for simple distillations. This system has a 14/10 ₮ male joint for connection to conical vials or round-bottom

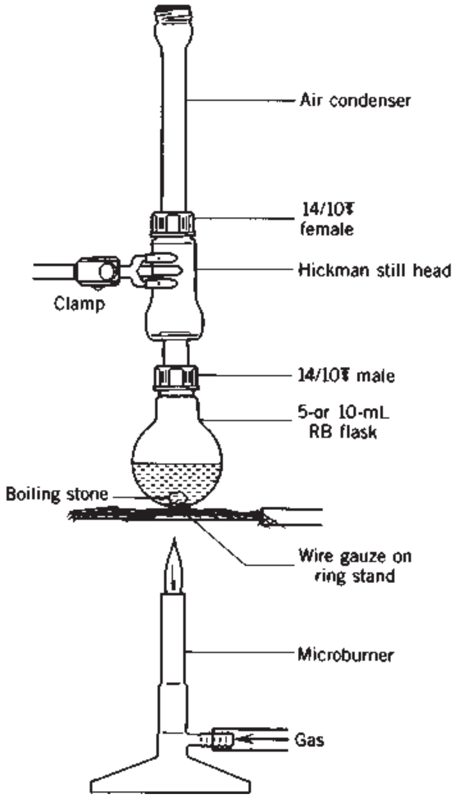

Figure 3.14 Hickman still head and air condenser with 5-mL round-bottom flask, arranged for microburner heating.

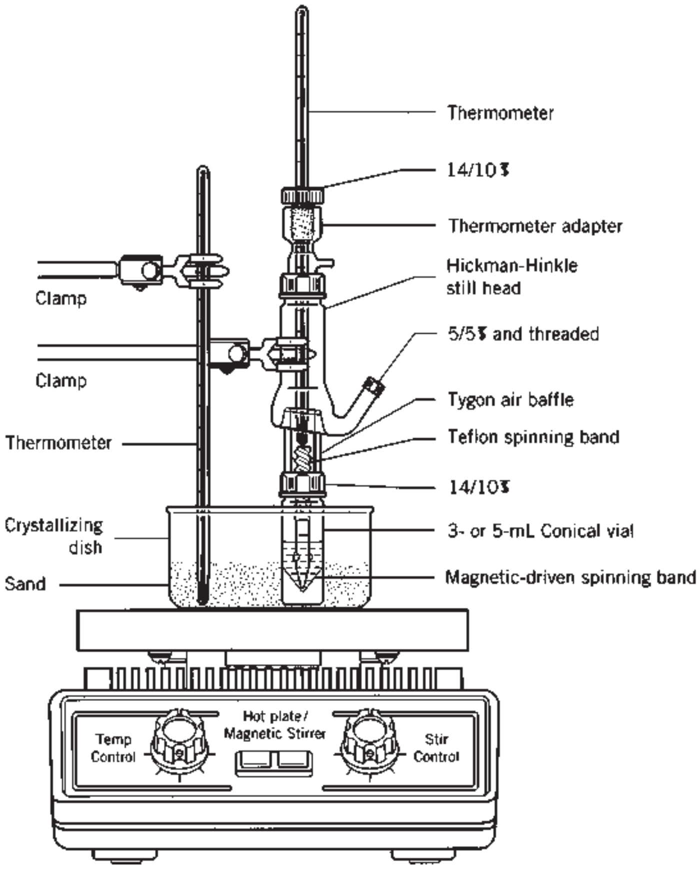
Thermometer
14/10₮
Thermometer adapter
Hickman-Hinkle still head
5/5₮ and threaded
Tygon air baffle
Teflon spinning band
14/10₮
3- or 5-mL Conical vial
Magnetic-driven spinning band
Clamp
Clamp
Thermometer
Crystallizing dish
Sand
Hot plate/ Magnetic Stirrer
Temp Control
Stir Control

Figure 3.15 Hickman–Hinkle still head with side-port 3- or 5-mL conical vial, Teflon spinning band, and thermometer adapter and arranged for heating and magnetic stirring.

flasks. The still head functions as both an air condenser and a condensate trap. For a detailed discussion of this piece of equipment see Experiments [3A] and [3B]. The simple Hickman still has been modified (see Fig. 3.15) with a spinning band. The still continues to function in much the same way as the simple Hickman still, but a tiny Teflon spinning band is now mounted in a slightly extended section between the male joint and the collection collar. When the band is spun at 1500 rpm by a magnetic-stirring hot plate, this still functions as an effective short-path fractional distillation column (see Distillation, Experiment [3D]). In addition, this modified system has a built-in thermometer well that allows reasonably accurate measurement of vapor temperatures plus a sidearm port for removing distillate.

The most powerful microscale distillation system currently available is the 2.5-in. vacuum-jacketed microscale spinning-band distillation column (see Fig. 3.16 and Experiment [3C] for description and details). This still is designed for conventional downward distillate collection and nonstopcock reflux control. The column is rated at ~10 theoretical plates.

For a discussion of reduced pressure distillations see Distillation.

www

Moisture-Protected Reaction Apparatus

Many organic reagents react rapidly and preferentially with water. *The success or failure of many experiments depends to a large degree on how well atmospheric moisture is excluded from the reaction system.* The “drying tube,” which is packed with a desiccant such as anhydrous calcium chloride, is a handy way to carry out a reaction in an apparatus that is not totally closed to the atmosphere, but

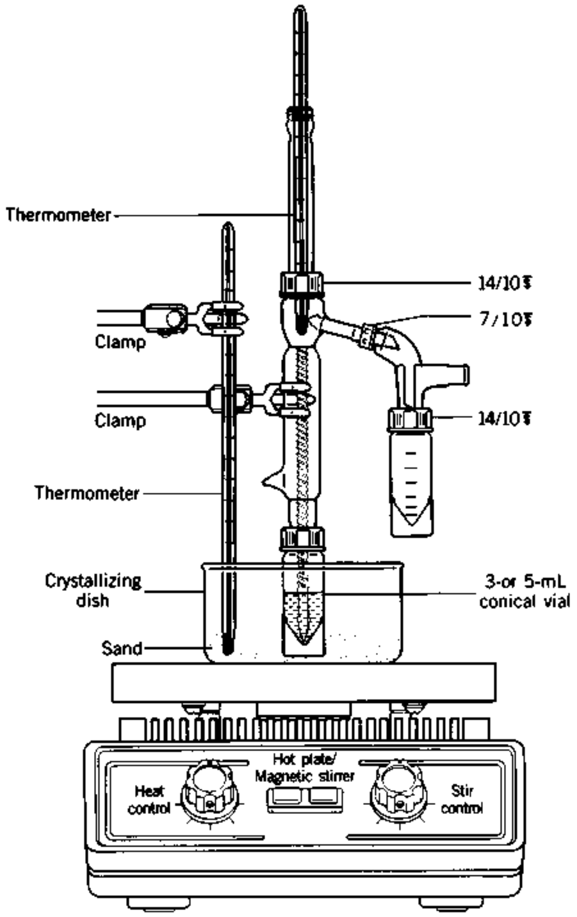
Thermometer
14/10₮
7/10₮
Clamp
Clamp
14/10₮
Thermometer
Crystallizing dish
3-or 5-mL conical vial
Sand
Hot plate/ Magnetic stirrer
Heat control
Stir control

Figure 3.16 Micro spinning band distillation column (2.5 in).

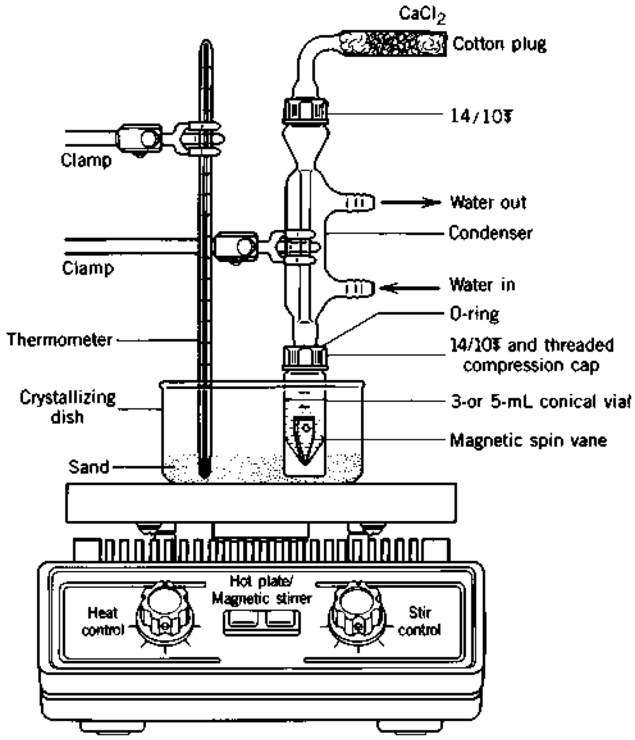
CaCl₂
Cotton plug
14/10₮
Clamp
Water out
Condenser
Clamp
Water in
O-ring
Thermometer
14/10₮ and threaded compression cap
Crystallizing dish
3-or 5-mL conical vial
Magnetic spin vane
Sand
Hot plate/ Magnetic stirrer
Heat control
Stir control

Figure 3.17 Moisture-protected water-jacketed condenser with 3- or 5-mL conical vial, arranged for heating and magnetic stirring.

that is reasonably well protected from water vapor. The microscale apparatus described here are designed to be used with the 14/10 ₮ drying tube. The reflux condensers discussed earlier are constructed with female 14/10 ₮ joints at the top of the column, which allows convenient connection of the drying tube if the refluxing system is moisture sensitive (see Fig. 3.17).

Because many reactions are highly sensitive to moisture, successful operation at the microscale level can be rather challenging. If anhydrous reagents are to be added after an apparatus has been dried and assembled, it is important to be able to introduce these reagents without exposing the system to the atmosphere, particularly when operating in a humid atmosphere. In room-temperature reactions that do not need refluxing, adding anhydrous reagents is best accomplished by use of the microscale Claisen head adapter. The adapter has a vertical screw-threaded standard taper joint that will accept a septum cap. The septum seal allows syringe addition of reagents and avoids the necessity of opening the apparatus to the laboratory atmosphere (see Fig. 3.18).

Specialized Pieces of Equipment

Collection of Gaseous Products. Some experiments lead to gaseous products. The collection, or trapping, of gases is conveniently carried out by using the capillary gas delivery tube. This item is designed to be attached directly to a 1- or 3-mL conical vial (see Fig. 3.19), or to the female 14/10 ₮ joint of a condenser connected to a reaction flask or vial (Chapter 3W, Fig. 3.11W). The tube leads to the collection system, which may be a simple, inverted, graduated cylinder; a blank-threaded septum joint; or an air condenser filled with water

www →

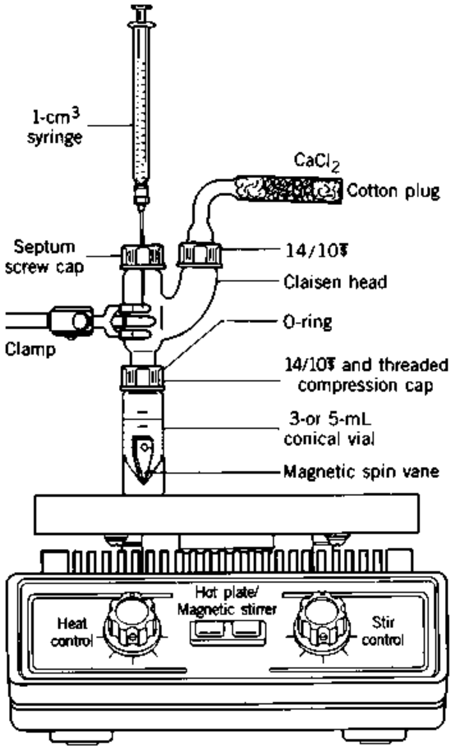
1-cm³ syringe
CaCl₂
Cotton plug
Septum screw cap
14/10\$
Claisen head
Clamp
O-ring
14/10\$ and threaded compression cap
3-or 5-mL conical vial
Magnetic spin vane
Heat control
Hot plate/ Magnetic stirrer
Stir control

Figure 3.18 Moisture-protected Claisen head with 3- or 5-mL conical vial, arranged for syringe addition and magnetic stirring.

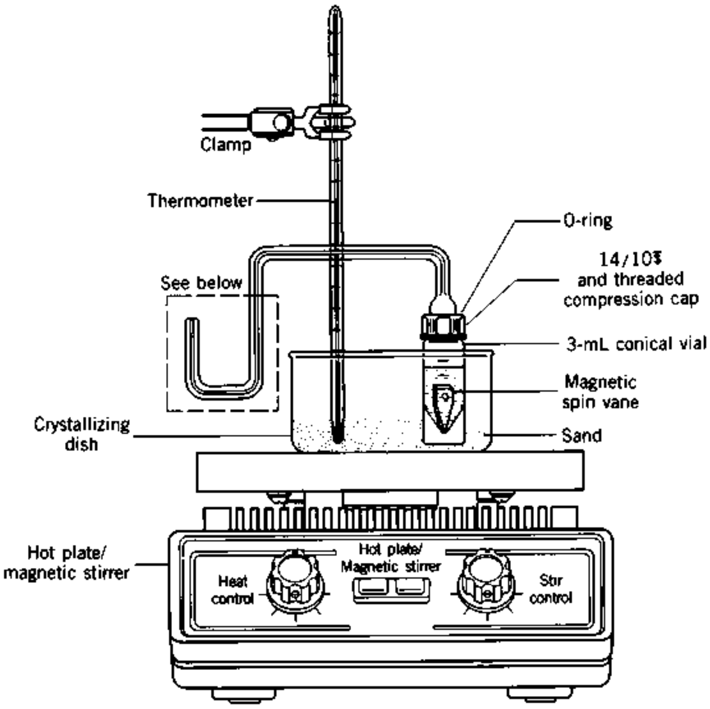
Clamp
Thermometer
O-ring
14/10\$ and threaded compression cap
See below
3-mL conical vial
Magnetic spin vane
Crystallizing dish
Sand
Hot plate/ magnetic stirrer
Heat control
Hot plate/ Magnetic stirrer
Stir control

Figure 3.19 Conical vial (3-mL) and capillary gas delivery tube arranged for heating and magnetic stirring.

(if the gaseous products are not water-soluble). The 0.1-mm capillary bore considerably reduces dead volume and increases the efficiency of product transfer.

Collection of Gas Chromatographic Effluents. The trapping and collection of gas chromatographic liquid fractions become particularly important exercises in microscale experiments. When yields of a liquid product are less than 100 μL conventional distillation, even using microscale equipment, is impractical. In this case, preparative gas chromatography replaces conventional distillation as the route of choice to product purification. A number of the reaction products in Chapters 6, 7, and 10W depend on this approach for successful purification and isolation. The ease and efficiency of carrying out this operation is greatly facilitated by employing the 5/5 \$ collection tube and the 0.1-mL 5/5 \$ conical collection vial (Chapter 3W, Fig. 3.12W).

MICROWAVE HEATING AS A TOOL FOR ORGANIC CHEMISTRY[*]

Introduction

An appliance found in almost all homes is a microwave oven. It is possible to heat food much more quickly and easily using a microwave as compared to the stove top. The observation that microwave energy can be used to heat food was first made by Percy Spencer, an employee of the Raytheon Corporation.[1] His company was a manufacturer of radar sets and, while working on one he noticed that the candy bar he had in his pocket had melted. Intrigued by this, the next day he brought in some popcorn from home and found that if he placed this near his radar set, it popped. In 1945, Raytheon filed a patent for the microwave cooking process, and in 1947 they built the first microwave oven called the Radarange.[2] It was almost 6 feet (1.8 m) high and weighed over 750 pounds (340 kg) and cost between $2,000 and $3,000.[3] The first popular home model was launched in 1967, and current estimates suggest that over 200 million microwave ovens are in use throughout the world today.[4]

---

[*] This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

Just as microwave ovens prove so valuable in the kitchen, it is also possible to use similar technology in preparative chemistry. It was in 1986 that the first reports of microwave heating as a tool for organic chemistry appeared in the scientific literature.[5,6] Two research groups published results they had obtained in their laboratories using kitchen microwave ovens. They said that chemistry that usually takes hours to reach completion using conventional heating could be performed in a matter of minutes in a microwave oven. Since these first reports, the use of microwave heating in organic chemistry has grown rapidly. Today, the technology is used in industry and academic laboratories for performing a wide range of reactions. Microwave heating has opened up a range of new areas in organic chemistry, allowing chemists to perform reactions quickly and easily. As an example, in this book Experiment 7, the Cannizzaro reaction, is performed in *one hour* using conventional heating. In the microwave protocol, the reaction is complete *in just one minute*.

The use of microwave heating addresses a number of the green chemistry principles.[7] Since it is often possible to obtain higher yields using microwave heating as opposed to conventional heating, there will be less waste and unused reagents. Also, since microwave heating is fast, there is often not enough time for products to decompose so this makes the product purification cleaner and easier. Chemists have also used the inherent advantages of microwave heating to their advantage for developing cleaner alternatives to known reactions. Take, for example, the use of water as a solvent instead of organics such as dichloromethane and benzene. Work has shown that water is an excellent solvent for organic chemistry, especially when combined with microwave heating. It is possible to heat water well above its boiling point in a sealed reaction vessel very safely and efficiently using microwaves. At these higher temperatures, water behaves more like an organic solvent. While most organic compounds are not soluble in water at room temperature, they can be in this higher temperature water, or at least partly so. This means that reactions can take place and, when the mixture cools down at the end, the product crystallizes out and is easily removed. As well as allowing for a more environmentally friendly solvent to be used, it also makes purification easy.

---

[1] *Reader's Digest*, August 1958, page 114.

[2] Spencer, P. L. 1945. Method of treating foodstuffs. US Patent 2,495,429, filed October 5, 1945, and published January 24, 1950.

[3] Gallawa, J. C. *Complete Microwave Oven Service Handbook: Operation, Maintenance, Troubleshooting, and Repair.* Prentice Hall, 2000.

[4] US Bureau of Labor Statistics.

[5] Gedye, R.; Smith, F.; Westaway, K.; Ali, H.; Baldisera, L.; Laberge, L.; and Rousell, J. "The Use of Microwave Ovens for Rapid Organic Synthesis," *Tetrahedron Lett.* **1986,** *27,* 279.

[6] Giguere, R. J.; Bray, T. L.; Duncan, S. M.; and Majetich, G. "Application of Commercial Microwave Ovens to Organic Sysnthesis," *Tetrahedron Lett*. **1986,** *27,* 4945.

[7] Anastas, P. T.; Warner, J. C. *Green Chemistry: Theory and Practice*; Oxford University Press, New York, 1998.

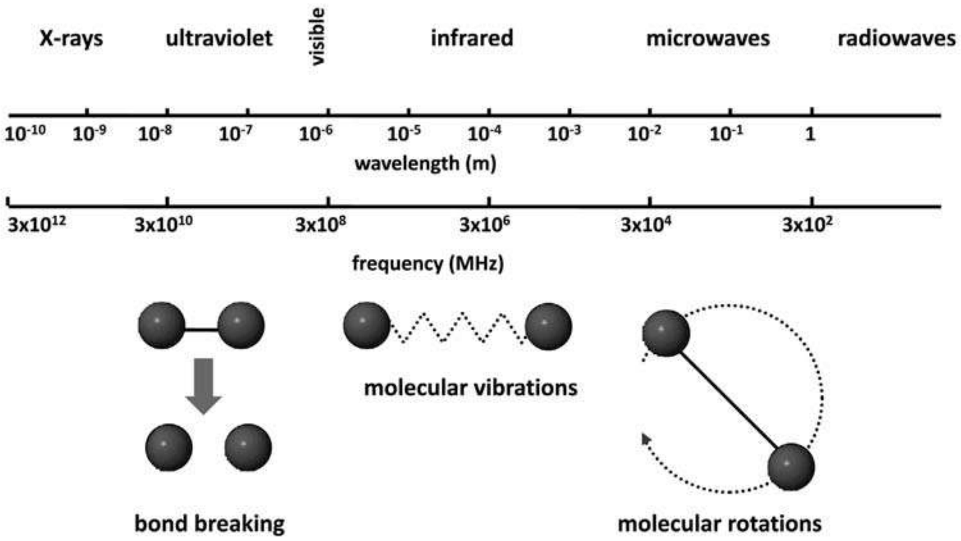
X-rays
ultraviolet
visible
infrared
microwaves
radiowaves
10^{-10}
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
1
wavelength (m)
3×10^{12}
3×10^{10}
3×10^{8}
3×10^{6}
3×10^{4}
3×10^{2}
frequency (MHz)
molecular vibrations
bond breaking
molecular rotations

Figure 3.20 The electromagnetic spectrum.

Before looking at applications in organic chemistry, it is important to appreciate some of the physical chemistry concepts behind microwave heating. The microwave region of the electromagnetic spectrum (Fig. 3.20) is classified as that between 300 and 300,000 megahertz (MHz). Compared to ultraviolet, infrared, and visible light, microwave irradiation is of relatively low energy. As a result, microwaves are not high enough in energy to break chemical bonds. Instead they can only make molecules rotate. This is very different from the more energetic ultraviolet radiation which, when it interacts with molecules, can break bonds, giving rise to the area of chemistry known as photochemistry.

Both home and scientific microwave equipment operates at 2,450 MHz. Interestingly, the microwave region of the electromagnetic spectrum is also used for navigation, communication, and remote sensing purposes. This includes technologies such as global positioning systems, wireless Internet, and Bluetooth as well as radar. As a result, the frequency used in microwave ovens has to be different from those used for these other applications and so is strictly regulated.

Microwaves, like all electromagnetic energy, move at the speed of light and comprise oscillating electric and magnetic fields (Fig. 3.21). These components oscillate at right angles to each other and to the direction of propagation.

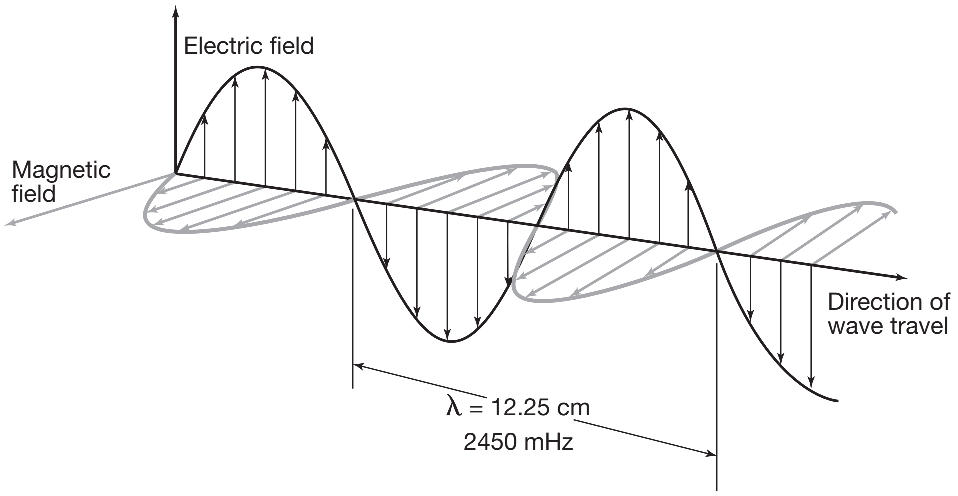
Electric field
Magnetic field
Direction of wave travel
λ = 12.25 cm
2450 mHz

Figure 3.21 Microwave energy comprises electric and magnetic fields.

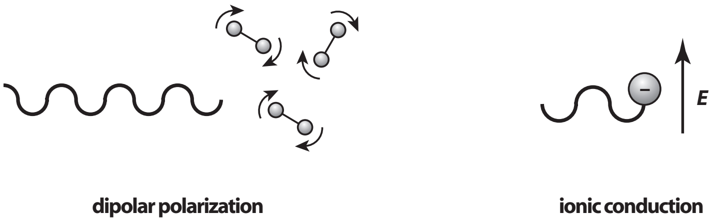

Figure 3.22 The two mechanisms by which microwave energy leads to heating.

There are two ways that microwaves can heat a sample, both involving the interaction of molecules in the sample with the electric field of the microwave irradiation (Fig. 3.22).

If a molecule possesses a dipole moment, then when it is exposed to microwave irradiation, the dipole tries to align with the applied electric field. Since the electric field is oscillating, the dipoles constantly try to realign to follow this. Molecules have time to align with the electric field but not to follow the oscillating field exactly. This continual reorientation of the molecules results in friction and thus heat. This heating method is termed dipolar polarization. If a molecule is ionic, then the electric field component of the microwave irradiation moves the ions back and forth through the sample also colliding them into each other. This movement again generates heat and is termed ionic conduction.

Conventionally, in order to heat reaction mixtures, chemists tend to use a hot plate or sand bath. These can be relatively slow and inefficient ways of transferring heat into a sample because they depend on convection currents and the thermal conductivity of the reaction mixture. Also, the walls of the reaction vessel can be hotter than the contents, thus introducing a thermal gradient. This can mean that reagents or products can be decomposed over time because they sit on the walls of the vessel. When using microwave heating, since the energy interacts with the sample directly, heating can be much more effective. In addition, microwave heating is safer than conventional heating; there are no sand baths or hot plates that can burn the chemist.

Each solvent or reagent in a reaction mixture will interact with microwave energy differently. Although not the only factor in determining the absorbance of microwave energy, the polarity of the solvent is a helpful tool for determining how well it will heat when placed into a microwave field; more polar molecules are affected more and nonpolar less. Solvents can be split into three categories; namely, those that absorb microwaves well, moderately, and poorly (Fig. 3.23). High-absorbing solvents will heat up very fast

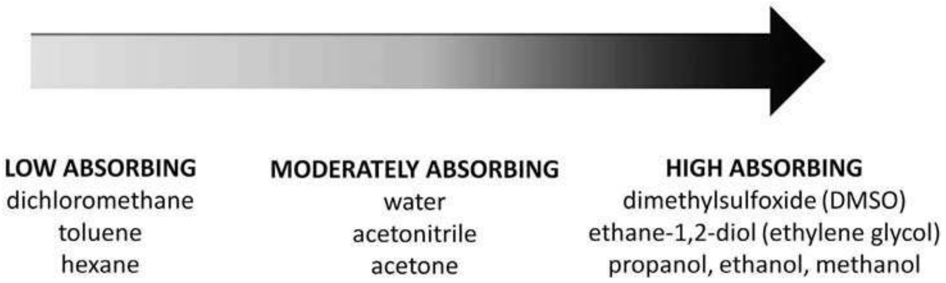

LOW ABSORBING	MODERATELY ABSORBING	HIGH ABSORBING
dichloromethane	water	dimethylsulfoxide (DMSO)
toluene	acetonitrile	ethane-1,2-diol (ethylene glycol)
hexane	acetone	propanol, ethanol, methanol

Figure 3.23 The heating characteristics of common solvents.

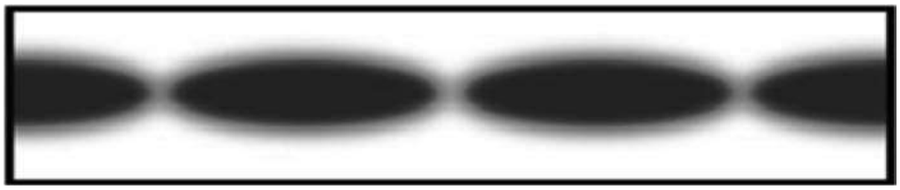

Figure 3.24 Areas of high and low microwave energy are found in the cavity of a multimode microwave unit.

upon microwave irradiation. Lower-absorbing solvents can still be used, but work better if one of the reagents in the reaction mixture is itself a good absorber. Interestingly, water absorbs more weakly than methanol while it is considerably more polar. This can be attributed, at least in part, to that fact that the strong, extensive hydrogen bonding in water goes some way to restricting rotation of molecules when irradiated with microwaves.

Electric power is turned into microwave energy using a magnetron, this in essence being a high-voltage tube in which electrons generated from a heated cathode are affected by magnetic and electric fields in such a way as to produce microwave radiation. As the microwaves come into the cavity (heating chamber) of a home microwave unit, they will move around and bounce off the walls. As they do so, they will generate pockets (called modes) of high energy and low energy as the moving waves either reinforce or cancel out each other (Fig. 3.24). This means that the microwave field in the microwave cavity is not uniform. Instead, there will be hot spots and cold spots; these correspond to the pockets of high and low energy, respectively. Domestic microwave ovens are therefore called"multimode"microwave ovens.

While home microwave ovens are useful for heating food, performing chemical reactions using them presents a number of challenges. They have no accurate temperature measurement device; the microwave field inside the oven is not uniform; and they are not safe for containing hot, flammable, organic solvents. These problems have led to the need for scientific microwave apparatus, specifically designed for performing chemical reactions safely and reliably. Scientific multimode microwave units have been developed for use in preparative chemistry (Fig. 3.26*b*). As well as being built to withstand explosions of reaction vessels inside the microwave cavity, temperature and pressure monitoring has been introduced as is the ability to stir reaction mixtures. It is possible to run a number of reactions at the same time in a multimode microwave oven, the samples being placed into tubes and loaded onto a turntable. As the samples move around, because they are large enough to absorb the microwave energy effectively, heating is fairly uniform.

When performing reactions on a small scale, it is sometimes difficult to heat the small volumes of reagents effectively in a multimode microwave apparatus. This is because, with the hot and cold spots that occur in the cavity of a multimode apparatus, it is hard to get constant microwave energy to irradiate the small sample. To overcome these problems, a smaller, single-mode (often called monomode) microwave apparatus has been developed. The cavity of a monomode microwave apparatus is designed for the length of only one wave (mode) (Fig. 3.25). By placing the sample in the middle of the cavity it can be irradiated constantly with microwave energy. Using a monomode apparatus, it is possible to heat samples of as little as 0.2 mL very effectively. The upper volume limit of the monomode apparatus is determined by the size of the microwave cavity and is in the region of 100 mL (Fig. 3.26*a*).

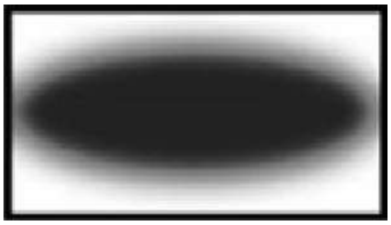

Figure 3.25 The cavity of a monomode microwave unit is designed to fit just one mode.

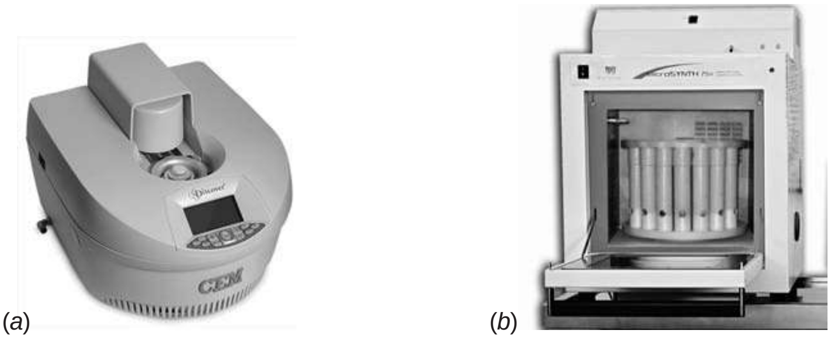

(*a*) (*b*)

Figure 3.26 (a) A monomode microwave unit (reproduced with permission from CEM Corporation) and (b) a multimode microwave unit (reproduced with permission from Milestone srl).

Applications

Many organic reactions require heat in order to proceed. In the lab, this is traditionally done using a hot plate, steam, oil or sand bath or, before that, a Bunsen burner. For those reactions that do require heat, the problem is that these heating sources are inefficient and reactions can often take a long time to reach completion. By using microwave heating, reaction times can be dramatically reduced and product yields can be higher. Shortening the time of known reactions is not the only advantage that microwave heating is having. It is impacting modern organic chemistry by opening up avenues to compounds that were previously not accessible. It is also a cleaner way to do preparative chemistry. Almost any reaction that needs heat can be performed in a microwave (Fig. 3.27). There are a few exceptions, including those that are known to be highly exothermic.

Microwave heating has proven particularly useful in the pharmaceutical industry where compounds need to be made rapidly so they can be screened for activity as drug candidates. In an interesting experiment undertaken by Boehringer Ingelheim Pharmaceuticals, the exact amount of time saved using microwave as opposed to conventional heating was determined.[8] Two

oxidation
reduction
substitution
addition
cycloadditions
rearrangements
ester and amide synthesis
ring-forming
heterocycle synthesis
metal-catalyzed processes

Figure 3.27 Some classes of organic reaction that can be performed using microwave heating.

---

[8] "Timesavings associated with microwave-assisted synthesis: A quantitative approach", C. R. Sarko in *Microwave Assisted Organic Synthesis* edited by J. P. Tierney and P. Lidstrom, Blackwell Publishing, Oxford, 2005.

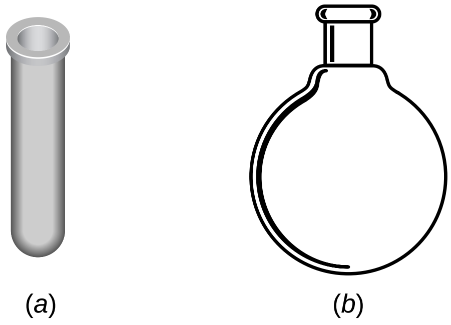
(*a*) (*b*)

Figure 3.28 Reactions can be performed using either a sealed tube or an open, round-bottom flask.

scientists were told to make a series of compounds. One of them used microwave chemistry and the other used conventional methods. Both scientists were given the same preparative route to the molecules to follow. However, after 37 days the chemist using the conventional heating approach concluded that the molecules could not be generated using that route. The microwave chemist on the other hand optimized reaction conditions and produced the final products in two days.

Many reactions utilizing microwave heating have been performed in sealed vessels (Fig. 3.28*a*). These are tubes of varying sizes that can be sealed with a specially designed stopper. Reaction mixtures can then be heated to temperatures well above the boiling point of the solvent inside. This offers a very safe way to perform chemistry at high temperatures and pressures. It is much more convenient than the steel containers used traditionally for this sort of chemistry. Also, it is possible to monitor the temperature and pressure of reaction mixtures very closely, and this means it is possible to report the exact reaction conditions used so that others can use them.

Another option is to use standard laboratory glassware in a microwave. Reactions can be run in round-bottomed flasks equipped with a reflux condenser (Fig. 3.28*b*). The flask sits inside the microwave cavity, and the reflux condenser comes out through the top of the apparatus. Often, just as good results can be obtained using an open vessel arrangement as compared to a sealed tube.

When using a monomode microwave unit, it is possible to perform reactions using sealed tubes of capacity ranging from 0.2–25 mL and open vessels ranging from 10–100 mL. Reactions are performed one at a time. When using sealed tubes, it is possible to automate the unit using robotics so that when one reaction is complete, the tube can be removed from the microwave and the next one put in. This allows for multiple reactions to be performed without the need for the operator to be present. Up to 60 reaction vessels can be lined up and run one after another.

Multimode microwave units can process multiple reaction vessels at the same time. The sealed vessels all sit in a holder (reaction carousel) that fits into the microwave cavity. Working on a scale of up to a few grams, it is possible to process up to 40 reaction vessels at a time. Up to 92 reactions can be run at a time when using microscale quantities of reagents. Another option possible when working in multimode microwave unit is to use one large reaction vessel. This can either be a larger sealed vessel (up to 1 L in volume) or an open round-bottom flask (up to 5 L in volume). This enables chemists to scale up their reactions to make more of their desired compound.

Equipment Available

There are a number of commercially available scientific microwave units. The four major microwave manufacturers are listed here:

Anton Paar is an Austrian company that manufactures a multimode microwave unit called the Synthos 3000. There are a number of reaction carousels that are available with the unit, allowing for reactions to be performed from the microliter scale up to 100 mL. On the small scale, reactions are run in specially designed silicon carbide plates with either 24 or 48 wells. Using plates made from this inert, highly microwave-absorbing material allows for equal heating of all the wells. Larger reactions are performed in glass or quartz tubes sealed with a specially designed stopper.

Biotage, a company based in Sweden, manufactures a monomode microwave unit called the Initator. Using this instrument, it is possible to run reactions on scales from 0.2–20 mL in sealed tubes. It is possible to automate the unit with a robotic arm, thus allowing up to 60 reactions to be run sequentially.

CEM Corporation, a company based in North Carolina, manufactures a range of microwave units. Its monomode microwave apparatus is called the Discover platform. A number of variants are available. It is possible to run reactions from 0.5–60 mL in sealed tubes using this unit, as well as open round-bottom flasks up to 125 mL in capacity. It is possible to automate the unit, allowing reactions to be run sequentially. In addition, an accessory is available for loading reaction vessels with reactive gases such as hydrogen and carbon monoxide, opening the door for performing a wide range of reactions otherwise not possible using microwave heating. CEM also manufactures a multimode microwave unit called the MARS. There are a number of reaction carousels accommodating sealed tubes that can be used with the unit. In addition, open round-bottom flasks up to 5 L in capacity can be placed into the microwave cavity and standard reflux glassware attached. This allows for scale-up of reactions using batch processing.

Milestone, a company based in Italy, manufactures a number of microwave units. The MultiSYNTH has the capability to act as both a monomode and a multimode microwave unit. This means that conditions can be optimized in a small sealed tube using the monomode functionality and then a series of up to 12 small or 6 larger reactions can run in parallel in multimode. The unit can also accommodate a round-bottom flask of capacity up to 1 L, allowing a reaction to be performed at atmospheric pressure. The MicroSYNTH platform is a multimode microwave unit. There are a number of reaction carousels that can be used, allowing for reactions to be performed in parallel using either glass, Teflon, or quartz tubes. In addition, open round-bottom flasks can be placed into the microwave cavity and standard reflux glassware attached.

All the modern scientific microwave units have the capability to measure temperature during the course of a reaction. This can be done remotely using an infrared sensor located in the wall or the bottom of the microwave cavity. In many cases it is also possible to record the temperature inside a reaction vessel using a fiber-optic probe or thermocouple. Pressure measurement is also possible in many cases. The contents of a reaction mixture can be stirred by means of a magnetic stir plate located beneath the microwave cavity and a Teflon stir bar in the vessel.

When running a reaction, key parameters such as temperature, pressure, and microwave power can be measured throughout the run and data saved to

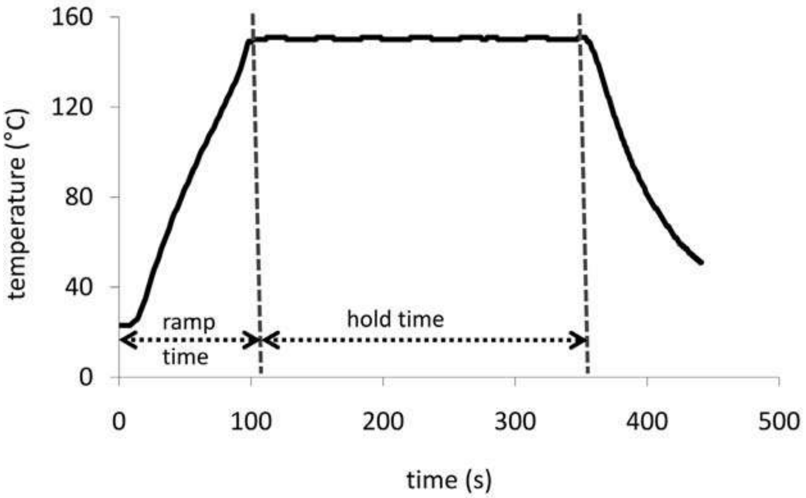

Figure 3.29 Example of a heating profile for a reaction performed using microwave heating.

a computer for use in reports and for reproducing the conditions at a later date (Fig. 3.29). Generally, when programming a protocol into a microwave unit, there are two important stages. The first is entering the ramp time. This is the time that the user wants the microwave to take to reach the target temperature. The second is entering the hold time. This is the time that the user wants the reaction mixture to remain at the target temperature before cooling back to room temperature. The microwave unit will use the requisite microwave power to heat the reaction mixture to temperature and then the power will automatically fluctuate to hold the reaction at the set temperature.

Experimental Protocols

Experimental protocols using microwave heating have been added to Experiments 7, 8, 15, 22, and 30. The experiments in this book can be performed on a range of these commercially available microwave units. The procedures are split into two classes; the first is generally for use with monomode microwave apparatus (Biotage Initiator and CEM Discover) and the second for use with multimode microwave units (Anton Paar Synthos 3000, CEM MARS, and Milestone MicroSYNTH). A modified version of the monomode procedure for use with the Anton Paar Synthos 3000 in conjunction with the silicon carbide plate format for microscale chemistry is added as a footnote in the monomode protocol.

MICROSCALE LAWS

Rules of the Trade for Handling Organic Materials at the Microscale Level

Now that we have briefly looked at the equipment we will be using to carry out microscale organic reactions, let us examine the specific techniques that are used to deal with the small quantities of material involved. Microscale synthetic organic reactions, as defined by Cheronis,[9] start with 15–150 mg of the limiting reagent. These quantities sound small, and they are. Although 150 mg of a light, powdery material will fill half a 1-mL conical vial, you will have a hard time observing 15 mg of a clear liquid in the same container, even with magnification. This volume of liquid, on the other hand, is reasonably easy to observe if it is in a 0.1-mL conical vial. A vital part of the game of working with small amounts of materials is to become familiar with microscale techniques and to practice them as much as possible in the laboratory.

---

[9] Cheronis, N. D. *Semimicro Experimental Organic Chemistry*; Hadrion Press: New York, 1958.

Rules for Working With Liquids at the Microscale Level

1. Liquids are never poured at the microscale level. Liquid substances are transferred by pipet or syringe. As we are working with small, easy-to-hold glass-ware, the best way to transfer liquids is to hold both containers with the fingers of one hand, with the mouths as close together as possible. The free hand is then used to operate the pipet (syringe) to withdraw the liquid and make the transfer. This approach reduces to a minimum the time that the open tip is not in, or over, one vessel or the other. We use three different pipets and two standard syringes to perform most experiments involving liquids. **This equipment can be a prime source of contamination.** Be very careful to thoroughly clean the pipets and syringes after each use.

- **a. Pasteur pipet (often called a capillary pipet).** A simple glass tube with the end drawn to a fine capillary. These pipets can hold several milliliters of liquid (Fig. 3.30*a*) and are filled using a small rubber bulb or one of the very handy, commercially available pipet pumps. Because many transfers are made with Pasteur pipets, it is suggested that several of them be calibrated for approximate delivery of 0.5, 1.0, 1.5, and 2.0 mL of liquid. This calibration is easily done by drawing the measured amount of a liquid from a graduated cylinder and marking the level of the liquid in the pipet. This mark can be made with transparent tape, or by scratching with a file. Indicate the level with a marking pen before trying to tape or file the pipet.
- **b. Pasteur filter pipet.** A very handy adaptation of the Pasteur pipet is a filter pipet. This pipet is constructed by taking a small cotton ball and placing it in the large open end of the standard Pasteur pipet. Hold the pipet vertically and tap it gently to position the cotton ball in the drawn section of the tube (Fig. 3.30*b*). Now form a plug in the capillary section by pushing the cotton ball down the pipet with a piece of copper wire (Fig. 3.30*c*). Finish by seating the plug flush with the end of the capillary (Fig. 3.30*d*). The optimum-size plug will allow easy movement along the capillary while it is being positioned by the copper wire. Compression of the cotton will build enough pressure against the walls of the capillary (once the plug is in position) to prevent the plug from slipping while the pipet is filled with liquid. If the ball is too big, it will wedge in the capillary before the end is reached, and wall pressure will be so great that liquid flow will be shut off. Even some plugs that are loose enough to be positioned at the end of the capillary will still have developed sufficient lateral pressure to make the filling rate unacceptably slow. If the cotton filter, however, is positioned too loosely, it may be easily dislodged from the pipet by the solvent flow. These

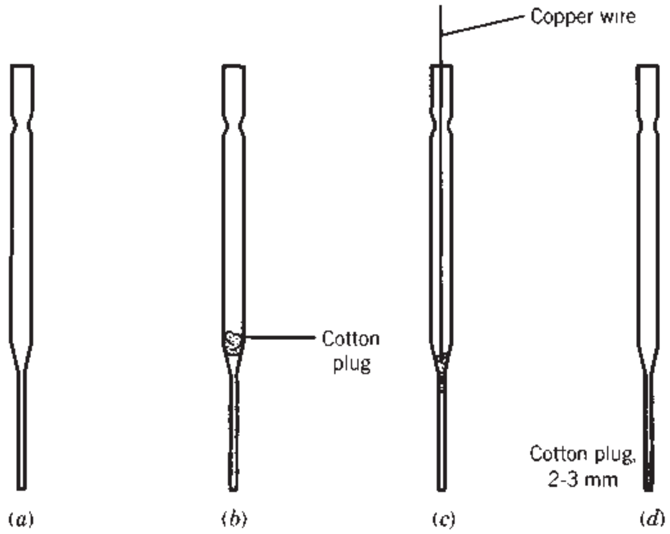

Figure 3.30 Preparation of Pasteur filter pipet.

plugs can be quickly and easily inserted with a little practice. Once in place, the plug is rinsed with 1 mL of methanol and 1 mL of hexane, and dried before use.

There are two reasons for placing the cotton plug in the pipet. First, it solves a particular problem with the transfer of volatile liquids via the standard Pasteur pipet: the rapid buildup of back pressure from solvent vapors in the rubber bulb. This pressure quickly tends to force the liquid back out of the pipet and can cause valuable product to drip on the bench top. The cotton plug tends to resist this back pressure and allows much easier control of the solution once it is in the pipet. The time-delay factor becomes particularly important when the Pasteur filter pipet is employed as a microseparatory funnel (see the discussion on extraction techniques in Technique 4, p. 67).

Second, each time a transfer of material is made, the material is automatically filtered. This process effectively removes dust and lint, which are constant problems when working at the microscale level. A second stage of filtration may be obtained by employing a disposable filter tip on the original Pasteur filter pipet as described by Rothchild.[10]

c. **Automatic pipet (considered the Mercedes–Benz of pipets).** Automatic pipets quickly, safely, and reproducibly measure and dispense specific volumes of liquids. These pipets are particularly valuable at the microscale level, because they generate the precise and accurate liquid measurements that are absolutely necessary when handling only microliters of a liquid. The automatic pipet adds considerable insurance for the success of an experiment, since any liquid can be efficiently measured, transferred, and delivered to the reaction flask.

The automatic pipet consists of a calibrated piston pipet with a specially designed disposable plastic tip. It is possible to encounter

---

[10] Rothchild, R. *J. Chem. Educ.* **1990,** *67,* 425.

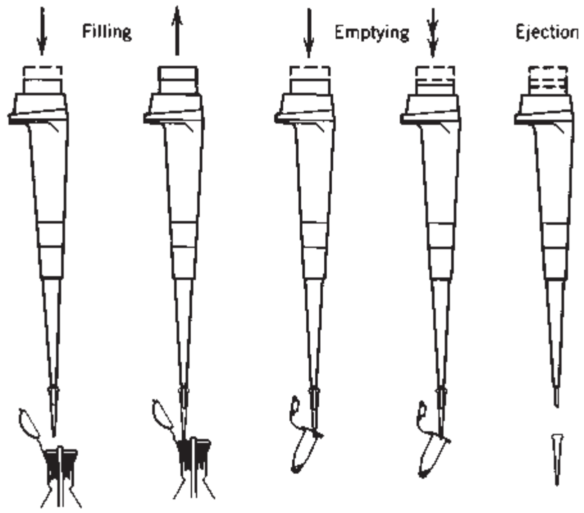

Figure 3.31 Operation of automatic delivery pipet.

any one of three pipet styles: single volume, multirange, or continuously adjustable (see Fig. 3.31). The first type is calibrated to deliver only a single volume. The second type is adjustable to two or three predetermined delivery volumes. The third type is the most versatile; it can be set by the user to deliver any volume within the range of the pipet. Obviously, the price of these valuable laboratory tools goes up with increasing features. Automatic pipets are expensive, and usually must be shared in the laboratory. Treat them with respect!

The automatic pipet is designed so that the liquid comes in contact only with the disposable tip.

- Never load the pipet without the tip in place.
- Never immerse the tip completely in the liquid that is being pipetted.
- Always keep the pipet vertical when the tip is attached.
- If an air bubble forms in the tip during uptake, return the liquid, discard the tip, and repeat the sampling process.

If these three rules are followed, most automatic pipets will give many years of reliable service. A few general rules for improving reproducibility with an automatic pipet should also be followed:

- Try to use the same uptake and delivery motion for all samples. Smooth depression and release of the piston will give the most consistent results. Never allow the piston to snap back.
- *Always* depress the piston to the first stop before inserting the tip into the liquid. If the piston is depressed after submersion, formation of an air bubble in the tip becomes likely, which will result in a filling error.
- *Never* insert the tip more than 5 mm into the liquid. It is good practice not to allow the body of the pipet to contact any surface, or bottle neck, that might be wet with a chemical.
- If an air bubble forms in the tip during uptake, return the fluid, discard the tip, and repeat the sampling process.

d. Syringes. Syringes are particularly helpful for transferring liquid reagents or solutions to sealed reaction systems from sealed reagent or solvent reservoirs. Syringe needles can be inserted through a septum, which avoids opening the apparatus to the atmosphere. Syringes are also routinely employed in the determination of ultramicro boiling points (10-μL GC syringe). It is critically important to clean the syringe needle after each use. Effective cleaning of a syringe requires as many as a dozen flushes. For many transfers, the microscale laboratory uses a low-cost glass 1-mL insulin syringe in which the rubber plunger seal is replaced with a Teflon seal (ACE Glass). For preparative GC injections, the standard 50- or 100-μL syringes are preferred (see Technique 1).

2. Liquid volumes may be converted easily to mass measures by the following relationship:

$$\text{Volume (mL)} = \frac{\text{mass (g)}}{\text{density (g/mL)}}$$

3. Work with liquids in conical vials, and work in a vial whose capacity is approximately twice the volume of the material it needs to hold. The trick here is to reduce the surface area of the flask in contact with the sample to an absolute minimum. A conical vial is thus better than the spherical surface of the conventional round-bottom flask.

Liquids may also be weighed directly. A tared container (vial) should be used. After addition of the liquid, the vial should be kept capped throughout the weighing operation. This procedure prevents loss of the liquid by evaporation. If the density of the liquid is known, the approximate volume of the liquid should be transferred to the container using an automatic delivery pipet or a calibrated Pasteur pipet. Use the above expression relating density, mass, and volume to calculate the volume required by the measured mass. Adjustment of the mass to give the desired value can then be made by adding or removing small amounts of liquid from the container by Pasteur pipet.

NOTE. Before you leave the balance area, be sure to replace all caps on reagent bottles and clean up any spills. A balance is a precision instrument that can easily be damaged by contamination.

Rules for Working With Solids at the Microscale Level

1. General considerations. Working with a crystalline solid is much easier than working with the equivalent quantity of a liquid. Unless the solid is in solution, a spill on a clean glass working surface usually can be recovered quickly and efficiently. *Be careful, however, when working with a solution. Treat solutions as you would a pure liquid.*

2. Transfer of solids. Solids are normally transferred with a microspatula, a technique that is not difficult to develop.

3. Weighing solids at the milligram level. Electronic balances can automatically tare an empty vial. Once the vial is tared, the reagent is added in small portions. The weight of each addition is instantly registered; material is added until the desired quantity has been transferred.

Solids are best weighed in glass containers (vials or beakers), in plastic or aluminum weighing trays (“boats”), or on glazed weighing paper. Filter paper or other absorbent materials are not good choices: small quantities of the weighed material will often stick to the fibers of the paper, and vice versa.

THE LABORATORY NOTEBOOK

Writing is the most important method chemists use to communicate their work. It begins with the record kept in a laboratory notebook. An experiment originally recorded in the laboratory notebook can become the source of information used to prepare scientific papers published in journals or presented at meetings. For the industrial chemist, these written records are critical in obtaining patent coverage for new discoveries.

It is important that you learn to keep a detailed account of your work. A laboratory notebook has several key components. Note how each component is incorporated into the example that follows.

Key Components of a Laboratory Experiment Write-up

1. Date experiment was conducted
2. Title of experiment
3. Purpose for running the reaction
4. Reaction scheme
5. Table of reagents and products
6. Details of procedure used
7. Characteristics of the product(s)
8. References to product or procedure (if any)
9. Analytical and spectral data
10. Signature of person performing the experiment and that of a witness, if required

In reference to point 6, it is the obligation of the person doing the work to list the equipment, the amounts of reagents, the experimental conditions, and the method used to isolate the product. Any color or temperature changes should be carefully noted and recorded.

Several additional points can be made about the proper maintenance of a laboratory record.

11. A hardbound, permanent notebook is essential.
12. Each page of the notebook should be numbered in consecutive order. For convenience, an index at the beginning or end of the book is recommended and pages should be left blank for this purpose.
13. If a page is not completely filled, an "X" should be used to show that no further entry was made.
14. Data are always recorded directly into the notebook, *never* on scrap paper! Always record your data in ink. If a mistake is made, draw a neat line through the word or words so that they remain legible. Do not completely obliterate anything; you might learn from your mistakes, if you can read them later.
15. Make the record clear and unambiguous. Pay attention to grammar and spelling.
16. In industrial research laboratories, your signature, as well as that of a witness, is required, because the notebook may be used as a legal document.

17. Always write and organize your work so that someone else could come into the laboratory and repeat your directions without confusion or uncertainty. *Completeness* and *legibility* are key factors.

Most of you are newcomers to the organic laboratory, and the reactions you will be performing have probably been worked out and checked in detail. Because of this, your instructor may not require you to keep your notebook in such a meticulous fashion. For example, when you describe the procedure (item 6), it may be acceptable to make a clear reference to the material in the laboratory manual and to note any modifications or deviations from the prescribed procedure. In some cases, it may be more practical to use an outline method. In any event, the following example should be studied carefully. It may be used as a reference when detailed records are important in your work. It is more important to record what you observed and what you actually did, than to record what you were supposed to observe and what you were supposed to do.

NOTE. Because of its length, the example here is typed. Notebooks are usually handwritten. Many chemists, however, now use computers to record their data.

The circled numbers refer to the list on p. 40

EXAMPLE OF A LABORATORY NOTEBOOK ENTRY

19 July 2009 ①

PREPARATION OF DIPHENYL SUCCINATE ②

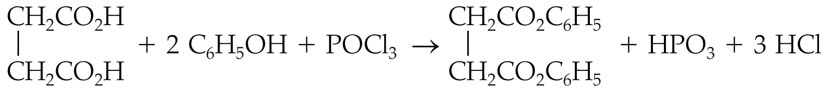

$$\begin{array}{l} CH_2CO_2H \\ | \\ CH_2CO_2H \end{array} + 2\,C_6H_5OH + POCl_3 \rightarrow \begin{array}{l} CH_2CO_2C_6H_5 \\ | \\ CH_2CO_2C_6H_5 \end{array} + HPO_3 + 3\,HCl$$

④

Diphenyl succinate is being prepared as one of a series of dicarboxylic acid esters that are to be investigated as growth stimulants for selected fungi species. ③

This procedure was adapted from that reported by Daub, G. H., and Johnson, W. S. *Organic Syntheses,* Wiley: New York, 1963; Collect. Vol. IV, p. 390. ⑧

Physical Properties of Reactants and Products

Compound	MW[a]	Amounts	mmol	mp (°C)	bp (°C)
Succinic acid	118.09	118 mg	1.0	182	
Phenol	94.4	188 mg	2.0	40–42	182
Phosphorous oxychloride	153.33	84 μL	0.9		105.8
Diphenyl succinate	270.29			121	

[a] MW = molecular weight.

⑤

⑥ In a 3.0-mL conical vial containing a magnetic spin vane and equipped with a reflux condenser protected by a calcium chloride drying tube were placed succinic acid (118 mg, 1.0 mmol), phenol (188 mg, 2.0 mmol), and phosphorous oxychloride (84 μL 0.9 mmol). The reaction mixture was heated with stirring at 115 °C in a sand bath in the **hood** for 1.25 h. It was necessary to conduct the reaction in the **hood,** because hydrogen chloride (HCl) gas evolved during the course of the reaction. The drying tube was removed, toluene (0.5 mL) was added through the top of the condenser using a Pasteur pipet, and the drying tube was replaced. The mixture was then heated for an additional 1 hour at 115 °C.

The hot toluene solution was separated from the red syrupy residue of phosphoric acid using a Pasteur pipet. The toluene extract was filtered by gravity using a fast-grade filter paper and the filtrate was collected in a 10-mL Erlenmeyer flask. The phosphoric acid residue was then extracted with two additional 1.0-mL portions of hot toluene. These extracts were also separated using the Pasteur pipet and filtered, and the filtrate was collected in the same Erlenmeyer flask. The combined toluene solutions were concentrated to a volume of approximately 0.6 mL by warming them in a sand bath under a gentle stream of nitrogen (N_2) gas in the **hood.** The pale yellow liquid residue was then allowed to cool to room temperature; the diphenyl succinate precipitated as colorless crystals. The solid was collected by vacuum filtration using a Hirsch funnel, and the filter cake was washed with three 0.5-mL portions of cold diethyl ether. The product was dried in a vacuum oven at 30°C (3 mm Hg) for 30 min.

⑦ ⑧ The 181 mg (67%) of diphenyl succinate had an mp of 120–121 °C (lit. value 121 °C: *CRC Handbook of Chemistry and Physics,* 89th ed.; CRC Press: Boca Raton, FL, 2008–2009; no. 13559, p. 3–220).

⑨ The IR spectrum exhibits the expected peaks for the compound. [*At this point, the data may be listed, or the spectrum attached to a separate page of the notebook.*]

⑩ Marilyn C. Waris

witnessed by
O. Jeanne d'Arc Mailhiot 19 July 2009

CALCULATING YIELDS

Almost without exception, in each of the experiments presented in this text, you are asked to calculate the percentage yield. For any reaction, it is always important for the chemist to know how much of a product is actually produced (experimental) compared to the theoretical (maximum) amount that could have been formed. The percentage yield is calculated on the basis of the relationship

$$\%\ \text{yield} = \frac{\text{actual yield (experimental)}}{\text{theoretical yield (calculated maximum)}} \times 100$$

The percentage yield is generally calculated on a weight (gram or milligram) or on a mole basis. In the present text, the calculations are made using milligrams.

Several steps are involved in calculating the percentage yield.

Step 1 Write a *balanced* equation for the reaction. For example, consider Experiment [22], the Williamson synthesis of propyl *p*-tolyl ether.

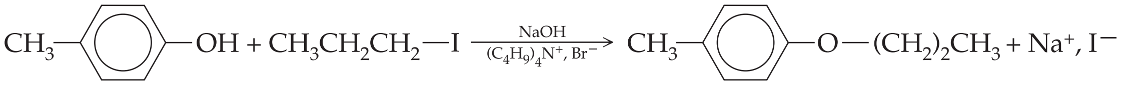

p-Cresol   Propyl iodide   Propyl *p*-tolyl ether

Physical Properties of Reactants				
Compound	**MW (mg/mmol)**	**Amount**	**mmol**	***d*(mg/μL)**
p-Cresol	108.15	160 μL	1.56	1.5312
25% (by weight) NaOH soln	40.0	260 μL	~1.6	
Tetrabutylammonium bromide	322.38	18 mg	0.056	
Propyl iodide	169.99	150 μL	1.54	1.5058

Step 2 Identify the *limiting* reactant. The ratio of reactants is calculated on a millimole (or mole) basis. In the example, 1.56 mmol of *p*-cresol and ca. 1.6 mmol of sodium hydroxide are used, compared to 1.54 mmol of propyl iodide, which is therefore the limiting reagent. The tetrabutylammonium bromide is not considered because it is used as a catalyst—it is neither incorporated into the product nor consumed in the reaction. The calculation of the theoretical yield is thus based on the amount of propyl iodide, 1.54 mmol.

Step 3 Calculate the *theoretical* (maximum) amount of the product that could be obtained for the conversion, based on the limiting reactant. Here, one mole of propyl iodide produces one mole of the propyl *p*-tolyl ether. Therefore, the maximum amount of propyl *p*-tolyl ether (molecular weight = 150.2) that can be produced from 1.54 mmol of propyl iodide is 1.54 mmol, or 231 mg.

Step 4 Determine the *actual* (experimental) yield (milligrams) of product isolated in the reaction. This amount is invariably less than the theoretical quantity, unless the material is impure (one common contaminant is water). For example, student yields for the preparation of propyl *p*-tolyl ether average 140 mg.

Step 5 Calculate the *percentage yield* using the weights determined in steps 3 and 4. The percentage yield is then

$$\%\ \text{yield} = \frac{140\ \text{mg (actual)}}{231\ \text{mg (theoretical)}} \times 100 = 60.6\%$$

As you carry out each reaction in the laboratory, try to obtain as high a percentage yield of product as possible. The reaction conditions in this book's experiments have been carefully developed; if you master the microscale techniques for transferring reagents and isolating products, your yields will be as high as possible.

QUESTIONS

3-1. Factory A produces the wheels that are used for the frames made in Factory B. Factory C relies exclusively on the materials produced in Factories A and B. Assuming all the necessary parts minus the wheels and frames are housed in Factory C, how many bicycles can be completely assembled when Factories A and B provide 36 wheels and 15 frames, respectively? Explain.

3-2. You are provided a vial that contains 180 mg of material. This material represents a 44 percent isolated yield. Calculate the theoretical amount (theoretical yield) that could have been formed.

3-3. The density of 2-methyl-2-butanol is 0.806 g/mL. How many mgs represent an aliquot of 430 μL? How many mmols represent an aliquot of 0.650 mL (2-methyl-2-butanol, formula weight is 88.15 g/mol)?

C_4H_6apter 4

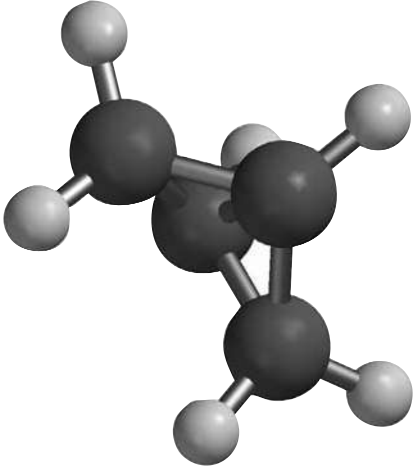

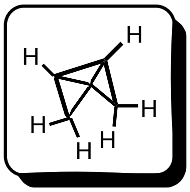

DETERMINATION OF PHYSICAL PROPERTIES

Determination of physical properties is important for substance identification and as an indication of material purity. Historically, the physical constants of prime interest have included boiling point, density, and refractive index in liquids and the melting point in solids. In special cases, optical rotation and molecular weight determinations may be required. Today, with the widespread availability of spectroscopic instrumentation, powerful new techniques may be applied to the direct identification and characterization of materials, including the analysis of individual components of very small quantities of complex mixtures. The sequential measurement of the infrared (IR) and mass spectro-metric (MS) characteristics of a substance resolved "on the fly" by capillary gas chromatography (GC) can be quickly determined and interpreted. This particular combination (GC-IR-MS), which stands out among a number of hyphenated techniques that are now available, is perhaps the most powerful system yet developed for molecular identification. The rapid development of high-field multinuclear magnetic resonance (NMR) spectrometers has added another powerful dimension to identification techniques. NMR sensitivity, however, is still considerably lower than that of either IR or MS. The IR spectrum alone, obtained with one data point per wavenumber can add more than 4000 measurements to the few classically determined properties. *Indeed, even compared to high-resolution MS and pulsed ^{1}H and ^{13}C NMR, the infrared spectrum of a material remains a powerful set of physical properties (transmission elements) available to the organic chemist for the identification of an unknown compound.*[1]

Simple physical constants are determined mainly to assist in establishing the purity of *known* materials. Because the boiling point or the melting point of a material can be very sensitive to small quantities of impurities, these data can be particularly helpful in determining whether a starting material needs further purification or whether a product has been isolated in acceptable purity. Gas (GC), high-performance liquid (HPLC), and thin-layer (TLC) chromatography, however, now provide more powerful purity information when such data are required. When a new composition of matter has been formed, an elemental (combustion) analysis is normally reported if sufficient material is available for this destructive analysis. For new substances we are, of course, interested in establishing not only the identity, but also the molecular structure of the materials. In this situation other modern techniques (such as ^{1}H and ^{13}C NMR spectroscopy, high-resolution MS, and single-crystal X-ray diffraction) can provide sensitive and powerful structural information.

When comparisons are made between experimental data and values obtained from the literature, it is essential that the latter information be obtained

---

Chapter 4: C_4H_6, Bicyclo[1.1.0]butane
Lemal, Menger, and Clark (1963).

[1] Griffiths, P. R.; de Haseth, J. A. *Fourier Transform Infrared Spectrometry,* 2nd ed.; Wiley: New York, 2007.

from the most reliable sources available. This is especially true when considering the volume of misinformation found online. Certainly, judgment, which improves with experience, must be exercised in accepting any value as a standard. In addition to the wealth of data found online (e.g., SciFinder Scholar), the known classical properties of a large number of compounds can be obtained from the *CRC Handbook of Chemistry and Physics* and the *Merck Index.* The *Aldrich Catalog Handbook of Fine Chemicals* is also a readily available, inexpensive source. These reference works list physical properties for inorganic, organic, and organometallic compounds. The *Aldrich Catalog* also references IR and NMR data for a large number of substances. New editions of the *CRC Handbook* and the *Aldrich Catalog* are published each year.

LIQUIDS

Ultramicro Boiling Point

Upon heating, the vapor pressure of a liquid increases, though in a nonlinear fashion. When the pressure reaches the point where it matches the local pressure, the liquid boils. That is, it spontaneously begins to form vapor bubbles, which rapidly rise to the surface. If heating is continued, both the vapor pressure and the temperature of the liquid will remain constant until the substance has been completely vaporized (Fig. 4.1).

Because microscale preparations generally yield about 30–70 μL of liquid products, using only 5 μL or less of material for boiling point measurements is highly desirable. The modification of the earlier Wiegand ultramicro boiling-point procedure[2] to the ultramicro procedure described here has established that reproducible and reasonably accurate (±1 °C) boiling points can be observed on 3–4 μL of most liquids thermally stable at the required temperatures.

Procedure. Ultramicro boiling points can be conveniently determined in standard (90-mm-length) Pyrex glass capillary melting-point tubes. The melting-point tube replaces the conventional 3- to 4-mm (o.d.) tubing used in the Siwoloboff procedure.[3] The sample (3–4 μL) is loaded into the melting-point capillary via a 10-μL syringe and centrifuged to the bottom if necessary. A small glass bell replaces the conventional melting-point tube as the bubble generator in micro boiling-point determinations. It is formed by heating 3-mm (o.d.) Pyrex tubing with a microburner and drawing it out to a diameter small enough to be readily fit inside the melting-point capillary. A section of the drawn capillary is fused and then cut to yield two small glass bells approximately 5 mm long (Fig. 4.2*a*). It is important that the fused section be reasonably large, because it is more than just a seal. The fused glass must add enough weight to the bell that it will firmly seat itself in the bottom of the melting-point tube.

An alternative technique for preparing the glass bells follows: heat the midsection of an open-ended melting point capillary tube and then draw the glass to form a smaller capillary section. This section is then broken approximately in the middle and each open end is sealed. The appropriate length for the bell is then broken off. Thus, two bells are obtained, one from each section. The sealing

---

[2] Wiegand, C. *Angew. Chem.* **1955,** *67,* 77. Mayo, D. W.; Pike, R. M.; Butcher, S. S.; Meredith, M. L. *J. Chem. Educ.* **1985,** *62,* 1114.
[3] Siwoloboff, A. *Berichte* **1886,** *19,* 795.

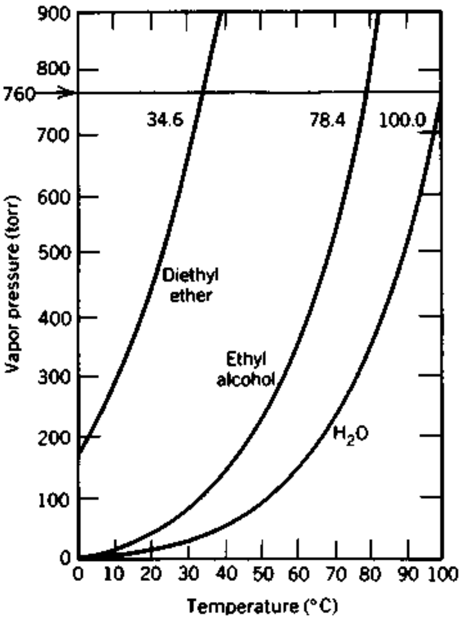

Figure 4.1 Vapor pressure curves. (From Brady, J. E.; Humiston, G. E. *General Chemistry,* 3d ed.; Wiley: New York, 1982. (Reprinted by permission of John Wiley & Sons, New York.)

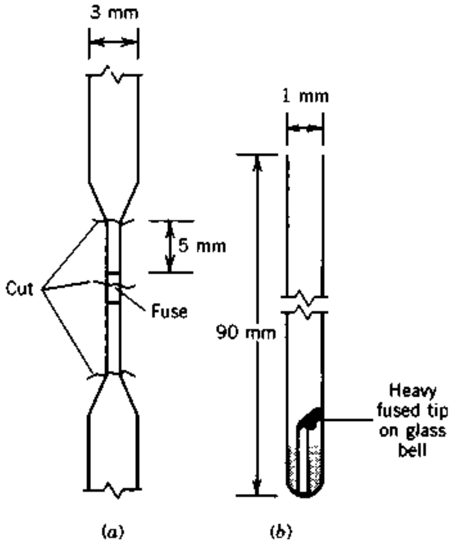

Figure 4.2 (a) Preparation of small glass bell for ultramicro boiling-point determination. (b) Ultramicro boiling-point assembly. (From Mayo, D. W.; Pike, R. M.; Butcher, S. S.; Meredith, M. L. *J. Chem. Educ.* ***1985,*** *62,* 1114.)

process (be sure that a significant section of glass is fused during the tube closure to give the bell enough weight) can be repeated on each remaining glass section and thus a series of bells can be prepared in a relatively short period.

A glass bell is now inserted into the loaded melting-point capillary, open end first (down), and allowed to fall (centrifuged if necessary) to the bottom. The assembled system (Fig. 4.2*b*) is then inserted onto the stage of a Thomas-Hoover Uni-Melt Capillary Melting Point Apparatus (Fig. 4.3)[4] or similar system (such as a Mel-Temp).

The temperature is rapidly raised to 15–20 °C below the expected boiling point (the temperature should be monitored carefully in the case of unknown substances), and then adjusted to a maximum rise rate of 2 °C/min and heated until a *fine stream* of bubbles is emitted from the glass bell. The heat control is then adjusted to drop the temperature. The boiling point is recorded at the point where the last escaping bubble collapses (i.e., when the vapor pressure of the substance equals the atmospheric pressure). The heater is then rapidly adjusted to again raise the temperature at 2 °C/min and induce a second stream of bubbles. This procedure may then be repeated several times. *A precise and sensitive temperature control system is essential to the successful application of this cycling technique, but it is not essential for obtaining satisfactory boiling-point data.*

Utilization of the conventional melting-point capillary as the "boiler" tube has the particular advantage that the boiling point of a liquid can readily be determined using a conventional melting-point apparatus. The illumination and magnification available make the observation of rate changes in the bubble stream easily seen. Inexpensive 10-μL GC injection syringes appear to be the most successful instrument to use for transferring the small quantities of liquids involved. The 3-in. needles normally supplied with the 10-μL barrels

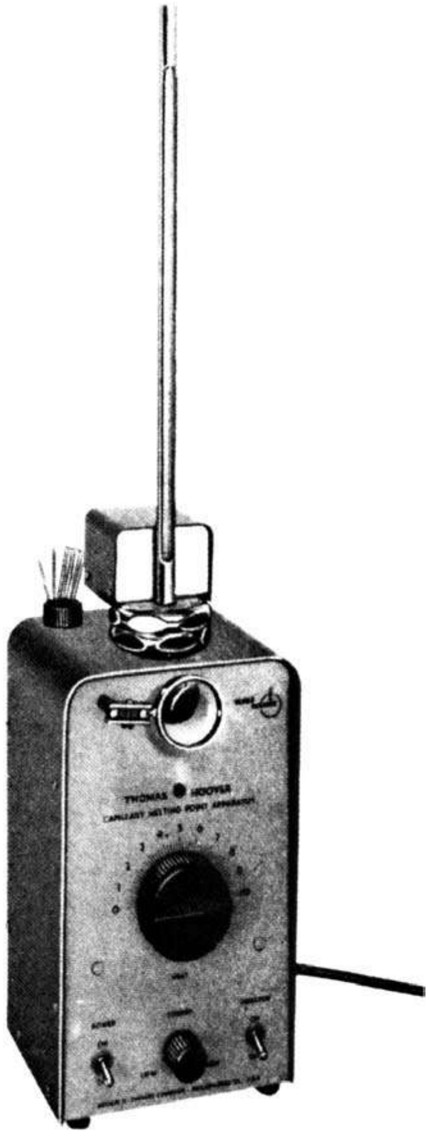

Figure 4.3 Thomas-Hoover melting-point determination device. *(Courtesy of Thomas Scientific, Swedesboro, NJ.)*

---

[4] Thomas Scientific, P.O. Box 99, Swedesboro, NJ 08085.

will not reach the bottom of the capillary; liquid samples deposited on the walls of the tube, however, are easily and efficiently moved to the bottom by centrifugation. After the sample is packed in the bottom of the capillary tube, the glass bell is introduced. The glass bell is necessary because a conventional Siwoloboff fused-capillary insert would extend beyond the top of the melting-point tube; thus, capillary action between the "boiler" tube wall and the capillary insert would draw most of the sample from the bottom of the tube up onto the walls. This effect often precludes the formation of the requisite bubble stream.

Little loss of low-boiling liquids occurs (see Table 4.1). Furthermore, if the boiling point is overrun and the sample is suddenly evaporated from the bottom of the "boiler" capillary, it will rapidly condense on the upper (cooler) sections of the tube. These sections extend above the heat-transfer fluid or metal block. The sample can easily be recentrifuged to the bottom of the tube and a new determination of the boiling point begun. Note that if the bell cavity fills completely during the cooling point of a cycle, it is often difficult to reinitiate the bubble stream without first emptying the entire cavity by overrunning the boiling point.

Observed boiling points for a series of compounds, which boil over a wide range of temperatures, are summarized in Table 4.1.

Materials that are thermally stable at their boiling point will give identical values on repeat determinations. Substances that begin to decompose will give values that slowly drift after the first few measurements. The observation of color and/or viscosity changes, together with a variable boiling point, signal the need for caution in making extended repeat measurements.

Comparison of the boiling points obtained experimentally at various atmospheric pressures with reference boiling points at 760 torr is greatly facilitated by the use of pressure–temperature nomographs such as that shown in Figure 4.4. A straight line from the observed boiling point to the observed pressure will pass through the corrected boiling-point value. These values can be of practical importance when carrying out reduced pressure distillations.

Table 4.1 **Observed Boiling Points (°C)**

Compound	Observed	Literature Value	Reference
Methyl iodide	42.5	42–43	*a*
Isopropyl alcohol	82.3	82.3	*b*
2,2-Dimethoxypropane	80.0	83.0	*c*
2-Heptanone	149–159	151.1	*d*
Cumene	151–153	152.4	*e*
Mesitylene	163	164.7	*f*
p-Cymene	175–178	177.1	*g*
Benzyl alcohol	203	205.3	*h*
Diphenylmethane	263–265	265	*i*

Note. (Observed values are uncorrected for changes in atmospheric pressure (corrections all estimated to be less than ±0.5 °C.)

Source. CRC Handbook of Chemistry and Physics, 89th ed.; CRC Press: Boca Raton, FL, 2008–2009: [a]no. 6307, p. 3–306; [b]no. 9167, p. 3–442; [c]no. 3883, p. 3–190; [d]no. 5689; p. 3–274; [e]no. 6478, p. 3–314; [f]no. 10509, p. 3–540; [g]no. 6509, p. 3–316; [h]no. 780, p. 3–42; [i]no. 4498, p. 3–218.

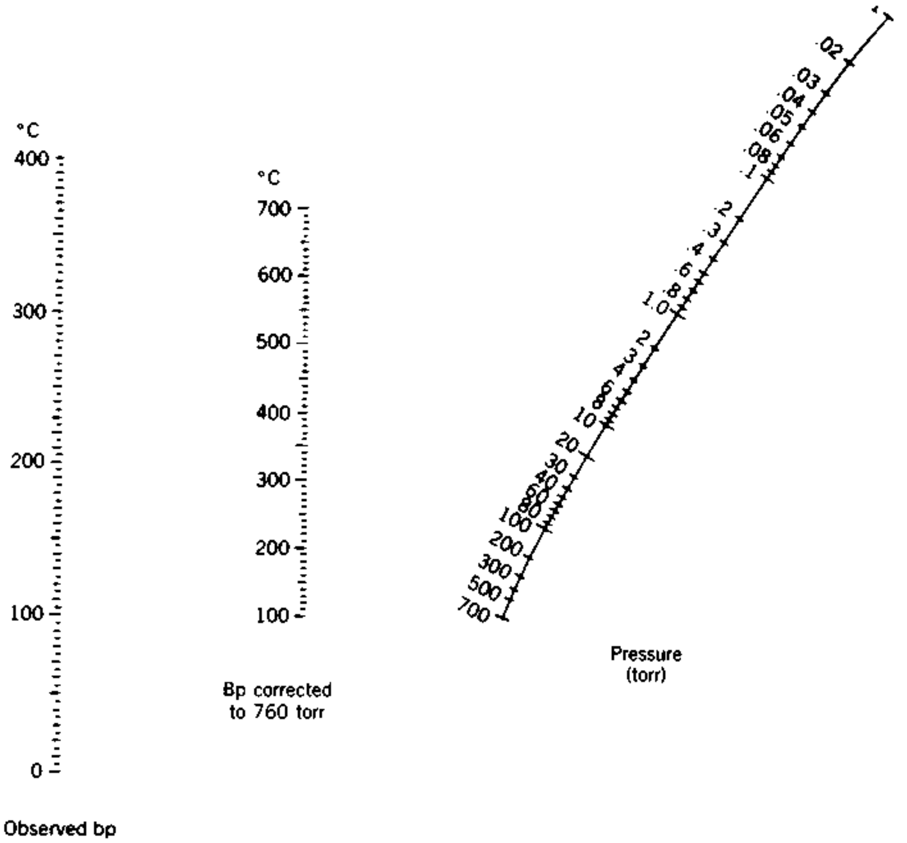

Figure 4.4 Pressure–temperature nomograph.

DENSITY

Density, defined as mass per unit volume, is generally expressed as grams per milli-liter (g/mL) or grams per cubic centimeter (g/cm^3) for liquids. Accurate nondestructive procedures have been developed for the measurement of this physical constant at the microscale level. A micropycnometer (density meter), developed by Clemo and McQuillen requires approximately 2 μL (Fig. 4.5).[5]

This very accurate device gives the density to three significant figures. The system is self-filling, and the fine capillary ends do not need to be capped while temperature equilibrium is reached or during weighing (the measured values tend to degrade for substances boiling under 100 °C and when room temperatures rise much above 20 °C). In addition, the apparatus must first be tared, filled, and then reweighed on an *analytical* balance. A technique that results in less precise densities (good to about two significant figures), but which is far easier to use, is simply to substitute a 50- or 100-μL syringe for the pycnometer. The method simply requires weighing the syringe before and after filling it to a measured volume as in the conventional technique. With the volume and the weight of the liquid known, the density can be calculated. A further advantage of the syringe technique is that the pycnometer is not limited to a fixed volume. Although much larger samples are required, it is not inconvenient to utilize the entire sample obtained in the reaction for this measurement, since the material can be efficiently recovered from the syringe for additional

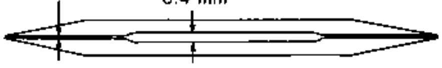

Figure 4.5 Pycnometer of Clemo and McQuillen. (From Schneider, F. L. Monographien aus dem Gebiete der qualitativen Mikroanalyse, *Qualitative Organic Microanalysis,* Vol. II; Benedetti-Pichler, A. A., Ed.; Springer: Vienna, 1964.)

---

[5] Clemo, G. R.; McQuillen, A. *J. Chem. Soc.* **1935,** 1220.

characterization studies. Because density changes with temperature, these measurements should be obtained at a constant temperature.

An alternative to the syringe method is to use *Drummond Disposable Microcaps* as pycnometers. These precision-bore capillary tubes, calibrated to contain the stated volume from end to end (accuracy ±1%), are available from a number of supply houses.[6] These tubes are filled by capillary action or by suction using a vented rubber bulb (provided). The pipets can be obtained in various sizes, but as with the syringe, volumes of 50, 75, or 100 μL are recommended. When using this method, handle the micropipet with forceps and not with your fingers (it's hot). The empty tube is first *tared,* and then filled and weighed again. The difference in these values is the weight of liquid in the pipet. For convenience, the pipet may be placed in a small container (10-mL beaker or Erlenmeyer flask) when the weighing procedure is carried out.

Two inexpensive micropycnometers can also be easily prepared: The first can be made from a Pasteur pipet as reported by Singh et al.[7] The volume of each individual pycnometer can be varied from 20 to 100 μL, or larger if desired. Values to three significant figures are obtained using an *analytical balance,* because evaporation is generally negligible, and if the pycnometer mouth is small.

The second pycnometer, by Pasto and co-workers, is made from a melting-point capillary tube.[8] In both of these techniques, the volume of the pycnometer must be determined. The procedure to determine the density involves the following steps. The empty micropycnometer is tared on an analytical balance, filled with the liquid in question, and reweighed (the difference in weights is the weight of the liquid). The sample is removed and the pycnometer is rinsed with acetone and dried. It is then filled with distilled water and reweighed. From the known[9] density of water at the given temperature the volume of water can be determined and thus the volume of the pycnometer. The volume of the original liquid sample also equals this value. The weight and volume of the sample are used to calculate its density.

SOLIDS

Melting Points

In general, the crystalline lattice forces holding organic solids together are distributed over a relatively narrow energy range. The melting points of organic compounds, therefore, are usually relatively sharp, that is, less than 2 °C. The range and maximum temperature of the melting point, however, are very sensitive to impurities. Small amounts of sample contamination by soluble impurities nearly always will result in melting-point depressions.

The drop in melting point is usually accompanied by an expansion of the melting-point range. Thus, in addition to the melting point acting as a useful guide in identification, it also can be a particularly effective indication of sample purity.

---

[6] Drummond Disposable Microcaps are available from Thomas Scientific, P.O. Box 99, Swedesboro, NJ 08085; and Sargent-Welch Scientific Co., a VWR company, P.O. Box 1026, Skokie, IL 60097.

[7] Singh, M. M.; Szafran, Z.; Pike, R. M. *J. Chem. Educ.* **1993,** *70,* A36; see also Ellefson-Kuehn, J., and Wilcox, C. J. *J. Chem. Educ.* **1994,** *71,* A150; and Singh, M. M.; Pike, R. M.; Szafran, Z. *Microscale and Selected Macroscale Experiments for General and Advanced General Chemistry;* Wiley: New York, 1995.

[8] Pasto, D.; Johnson, C.; Miller, M. *Experiments and Techniques in Organic Chemistry;* Prentice-Hall: Englewood Cliffs, NJ, 1992.

[9] Values for the density of water at various temperatures can be found in the *CRC Handbook of Chemistry and Physics.*

Procedure. In the microscale laboratory, two different types of melting-point determinations are carried out: (1) simple capillary melting points and (2) evacuated melting points.

Simple Capillary Melting Point. Because the microscale laboratory utilizes the Thomas–Hoover Uni-Melt apparatus or a similar system for determining boiling points, melting points are conveniently obtained on the same apparatus. The Uni-Melt system utilizes an electrically heated and stirred silicone oil bath. The temperature readings require no correction in this case because the depth of immersion is held constant. (This assumes, of course, that the thermometer is calibrated to the operational immersion depth.) Melting points are determined in the same capillaries as boiling points. The capillary is loaded by introducing about 1 mg of material into the open end. The sample is then tightly packed (~2 mm) into the closed end by dropping the capillary down a length of glass tubing held vertically to the bench top. The melting-point tube is then ready for mounting on the metal stage, which is immersed in the silicone oil bath of the apparatus. If the melting point of the substance is expected to occur in a certain range, the temperature can be rapidly raised to ~2 °C below the expected value. At that point, the temperature rise should be adjusted to a maximum of 2 °C/min, which is the standard rate of change at which the reference determinations are obtained. The melting-point range is recorded from the temperature at which the first drop of liquid forms (point *e* in Fig. 4.6) to that at which the last crystal melts (point *m* in Fig. 4.6).

Evacuated Melting Points. Many organic compounds begin to decompose at their melting points. This decomposition often begins as the melting point is approached and may adversely affect the values measured. The decomposition can be invariably traced to reaction with oxygen at elevated temperatures. If the melting point is obtained in an evacuated tube, therefore, much more accurate melting points can be obtained. These more reliable values arise not only from increased sample stability, but because several repeat determinations

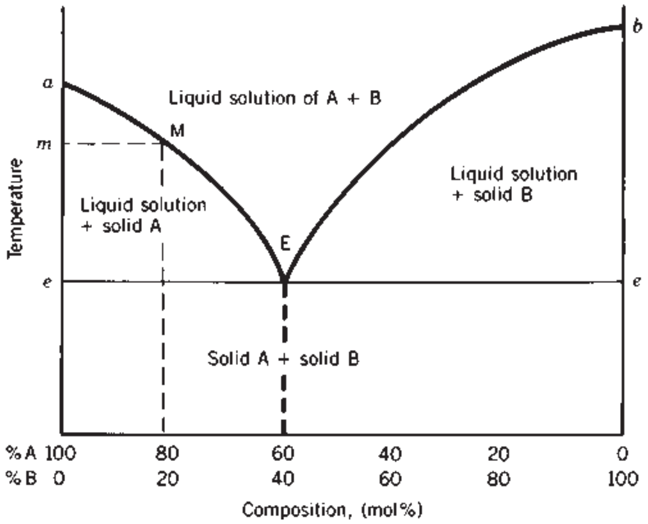

Figure 4.6 Melting point composition diagram for the binary mixture, *A* + *B*. In this diagram, *a* is the melting point of the solid *A*, *b* of solid *B*, *e* of eutectic mixture *E*, and *m* of the 80% *A*:20% *B* mixture, *M*.

can often be made on the same sample. The multiple measurements then may be averaged to provide more accurate data.

Evacuated melting points are quickly and easily obtained with a little practice. The procedure is as follows: Shorten the capillary portion of a Pasteur pipet to approximately the same length as a normal melting-point tube (Fig. 4.7*a*). Seal the capillary end by rotating in a microburner flame. Touch the pipet only to the very edge of the flame, and keep the large end at an angle below the end being sealed (Fig. 4.7*b*). This technique will prevent water from the flame being carried into the tube, where it will condense in the cooler sections. Then load 1–2 mg of sample into the drawn section of the pipet with a microspatula (Fig. 4.7*c*). Tap the pipet gently to seat the solid powder as far down the capillary as it can be worked (Fig. 4.7*d*). Then push the majority of the sample part way down the capillary with the same diameter copper wire that you used to seat the cotton plug in constructing the Pasteur filter pipet (Fig. 4.7*e*). Next, connect the pipet with a piece of vacuum tubing to a mechanical high-vacuum pump. Turn on the vacuum and evacuate the pipet for 30 s (Fig. 4.7*f*). With a microburner, gently warm the surface of the capillary tubing just below the drawn section. On warming, the remaining fragments of the sample (the majority of which has been forced farther down in the tube) will sublime in either direction away from the hot section. Once the traces of sample have been "chased" away, the heating is increased, and the capillary tube is collapsed, fused, and separated from the shank. The shank remains connected to the vacuum system (Fig. 4.7*g*). The vacuum system is then vented and the shank is discarded. The sample is tightly packed into the initially sealed end of the evacuated capillary by dropping it down a section of glass tubing, as in the case of packing open melting-point samples. After the sample is packed (~2 mm in length, see Fig. 4.7*h*), a section of the evacuated capillary about 10–15 mm above the sample is once more gently heated and collapsed by the microburner flame (Fig. 4.7*i*).

This procedure is required to trap the sample below the surface of the heated silicone oil in the melting-point bath, and thus avoid sublimation up the tube to cooler sections during measurement of the melting point. The operation is a little tricky and should be practiced a few times. It is very important that the tubing completely fuse. Now the sample is ready to be placed in the melting-point apparatus. The procedure beyond this point is the same as in the open capillary case, except that after the sample melts, it can be cooled, allowed to crystallize, and remelted several times, and the average value of the range reported. If these values begin to drift downward, the sample can be considered to be decomposing even under evacuated, deoxygenated conditions. In this case the first value observed should be recorded as the melting point and decomposition noted (mp xx dec, where dec = decompose).

Mixture Melting Point. Additional information can often be extracted from the sensitivity of the melting point to the presence of impurities. Where two different substances possess identical melting points (not uncommon), it would be impossible to identify an unknown sample as either material based on the melting point alone. If reference standards of the two compounds are available, however, then mixtures of the unknown and the two standards can be prepared. It is important to prepare several mixtures of varying concentrations for melting-point comparisons, since the point of maximum depression need not occur on the phase diagram at the 50:50 ratio (see Fig. 4.6). In samples that do not exhibit any decomposition at the melting point, the prepared

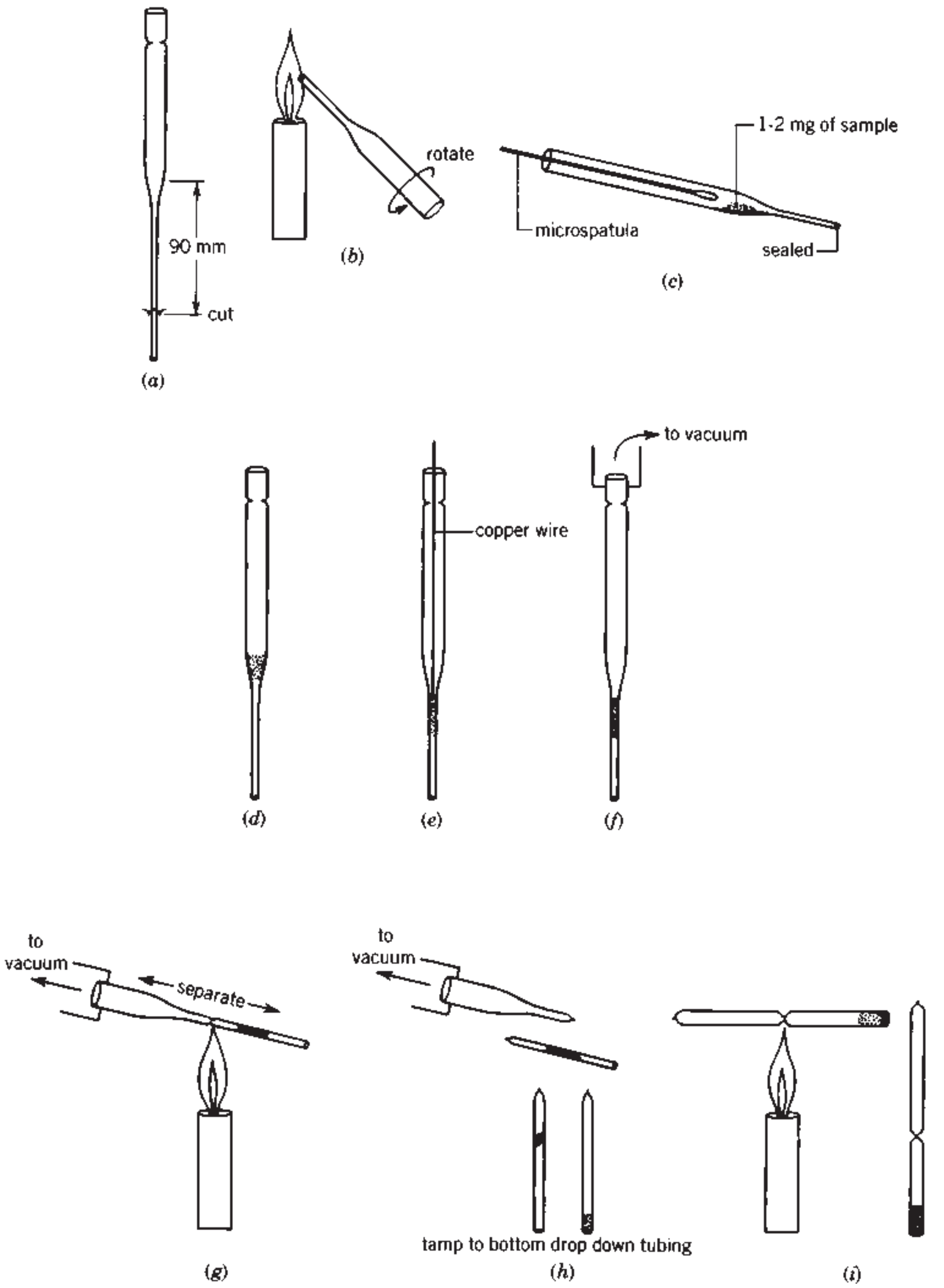
90 mm
cut
(a)
rotate
(b)
1-2 mg of sample
microspatula
sealed
(c)
(d)
copper wire
(e)
to vacuum
(f)
to vacuum
separate
(g)
to vacuum
tamp to bottom
drop down tubing
(h)
(i)

Figure 4.7 **Procedure for obtaining evacuated melting-point capillaries.**

mixtures should be first heated until a homogeneous melt is obtained. Each is then cooled and ground to a fine powder, and the definitive melting point is obtained on the ground sample. The melting points of the unknown and the mixed samples should be obtained simultaneously (the Uni-Melt stage will accept up to seven capillaries at one time). This is desirable because all the samples will then be heated at the same rate. The unknown sample and the mixture of the unknown with the correct reference will have identical values, but the mixture of the reference with a different substance will give a depressed melting point. This procedure is the classical step to positive identification of crystalline solids.

Mixtures of two different compounds only rarely fail to exhibit mixture melting-point depression, but it can happen. Some mixtures may not show a depression or show only a very small one, due to eutectic or compound formation. Elevation of the melting point has also been observed. Therefore, if mixture melting-point data are used for identification purposes, comparison of other physical constants or spectroscopic data is advocated to establish identity beyond any reasonable doubt.

QUESTIONS

4-1. Room temperature is recorded when a density determination for a given substance is performed in the laboratory. Why?

4-2. Describe how you would determine the melting point of a substance that sublimes before it melts.

4-3. You are presented with four vials, each containing a white crystalline solid. Two are unlabeled vials containing pure samples of *trans*-cinnamic acid and urea, respectively. The other two are labeled reference standards for each sample. Devise a method for the proper identification of the unlabeled vials, knowing that the literature melting point for both *trans*-cinnamic acid and urea is 132.5–133 °C.

4-4. In the microscale method of determining boiling points, one heats the liquid until a steady stream of bubbles is observed coming out of the bell. The temperature is then lowered and the boiling point is read just as the bubbles stop. Why is this technique preferable to measuring the boiling point when the bubbles first start to appear?

4-5. What would you expect the observed boiling point to be at 10 torr of a liquid which has a boiling point of 300 °C at 760 torr?

C$_5$H$_6$apter 5

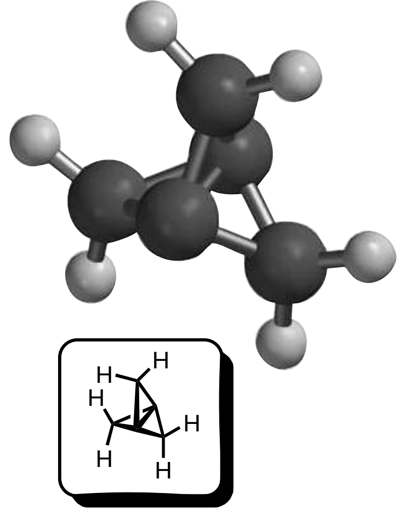

MICROSCALE LABORATORY TECHNIQUES

This chapter introduces the microscale organic laboratory techniques used throughout the experimental sections of the textbook. These must be mastered to be successful when working at this scale. Detailed discussions are given for each individual experimental technique. At the end of each discussion there is a list of the experiments in Chapters 6, 7 and 10W that use the technique. These lists should prove useful to instructors compiling experiments to be covered in the laboratory. The lists will also be handy for students who wish to examine the application of a particular technique to other experiments not covered in their laboratory sequence.

As was the case with the fourth edition, a continued effort has been made to streamline the basic reference material from the text using our accompanying website (www.wiley. com/college/MOL5). The icon at the right is used throughout the text to indicate website material that will be of interest to the user. We hope this treatment of the laboratory will make the more important aspects of the basic text easier to access and will speed your laboratory work along.

www

One of the principal hurdles in dealing with experimental chemistry is the isolation of pure materials. Characterization (identification) of a substance almost always requires a pure sample of the material. This is a particularly difficult demand of organic chemistry because most organic reactions generate several products. We are generally satisfied if the desired product is the major component of the mixture obtained. This chapter places a heavy emphasis on separation techniques.

TECHNIQUE 1

Gas Chromatography

Technique 1 begins the discussion of the resolution (separation) of microliter quantities of liquid mixtures via preparative gas chromatography. Techniques 2 and 3 deal with semimicro adaptations of classical distillation routines that focus on the separation of liquid mixtures involving one to several milliliters of material.

Chromatography methods revolutionized experimental organic chemistry. These methods are by far the most powerful of the techniques for separating mixtures and isolating pure substances, either solids or liquids. Chromatography is the resolution (separation) of a multicomponent mixture (several hundred components in some cases) by distribution between two phases, one stationary and one mobile. The various methods of chromatography are categorized by the phases involved: column, thin-layer, and paper (all solid–liquid chromatography); partition (liquid–liquid chromatography); and vapor phase

---

Chapter 5: C$_5$H$_6$, Propellane
Wiberg and Walker (1982).

(gas–liquid chromatography, or simply gas chromatography). The principal mechanism these separations depend on is differential solubility, or adsorbtivity, of the mixture components in the two phases involved. That is, the components must exhibit different partition coefficients (see also Technique 4 for a detailed discussion of partition coefficients).

Gas chromatography (GC, sometimes called vapor-phase chromatography) is an extraordinarily powerful technique for separating mixtures of organic compounds. The stationary phase in GC is a high-boiling liquid and the mobile phase is a gas (the carrier gas). Gas chromatography can separate mixtures far better than distillation techniques can (see Technique 2 discussion).

Preparative GC separations, which involve perhaps 5–100 μL of material, require relatively simple instrumentation but sacrifice resolution for the ability to separate larger amounts of material.

Analytical GC separations require tiny amounts of material (often 0.1 μL of a very dilute solution), and can separate incredibly complex mixtures. The ability to work with small quantities of materials in analytical GC separations is an advantage at the microscale level. This analytical tool is used to analyze distillation fractions in Experiments [3C] and [3D].

GC Instrumentation

GC instrumentation can range from straightforward and relatively simple systems to systems with complex, highly automated, and relatively expensive components. A diagram of a common and simple GC typically used in an instructional laboratory is shown in Figure 5.1.

Injection Port The analysis begins in a heated injection port. The sample mixture is introduced by syringe through a septum into the high-temperature chamber (injection port) through which the inert carrier gas (the mobile phase) is flowing. Helium and nitrogen are common carrier gases. The solubility of the sample in the carrier gas depends mostly on the vapor pressure of the substances in the sample. Heating the injection port helps to ensure the vaporization of less volatile samples. There are two major constraints on GC:

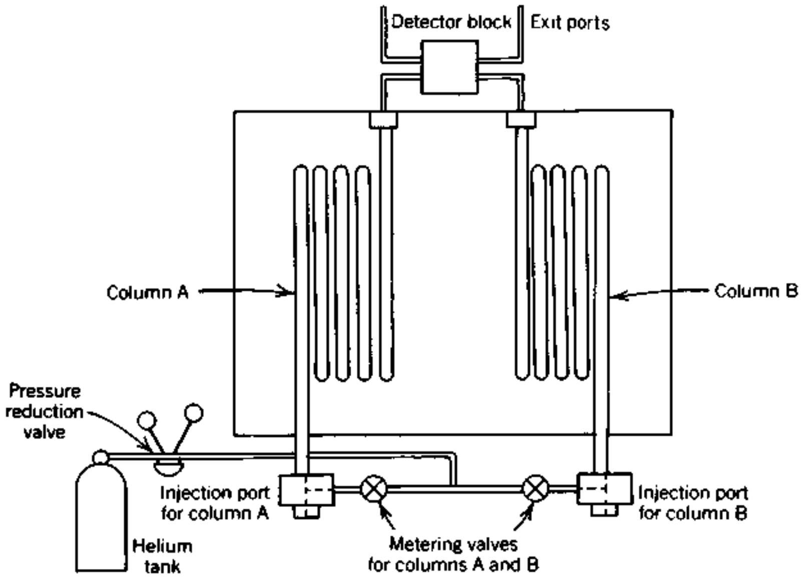
Detector block
Exit ports
Column A
Column B
Pressure reduction valve
Injection port for column A
Injection port for column B
Helium tank
Metering valves for columns A and B

Figure 5.1 Block diagram of a dual-column gas chromatograph showing essential parts. *(Courtesy of GOW-MAC Instrument Co., 277 Brodhead Rd., Bethlehem, PA 18017.)*

The sample must be stable at the temperature required to cause vaporization, and the sample must have sufficient vapor pressure to be completely soluble in the carrier gas at the column operating temperatures.

NOTE. When injecting a sample, always position your thumb or finger over the syringe plunger. This prevents a blow-back of the sample by the carrier gas pressure in the injection port.

Column The vaporized mixture is swept by the carrier gas from the injection port onto the column. Bringing the sample mixture into intimate contact with the column begins the separation process. The stationary liquid phase in which the sample will partially dissolve is physically and/or chemically bonded to inert packing material (often called the support) in the column. Gas-chromatographic columns are available from manufacturers in a variety of sizes and shapes. In the diagram of the GOW-MAC instrument (Fig. 5.1), two parallel coiled columns are mounted in an oven. Considerable oven space can be saved and better temperature regulation achieved if the columns are coiled. Temperature regulation is particularly important, because column resolution degrades rapidly if the entire column is not at the same temperature. Most liquid mixtures need a column heated above ambient temperatures to achieve the vapor pressure the separation requires.

The mixture is separated as the carrier gas sweeps the sample through the column. Columns are usually made from stainless steel, glass, or fused silica. The diameter and length of the column are critical factors in separating the sample mixture.

Packed Columns. In packed columns the liquid (stationary) phase in contact with the sample contained in the mobile gas phase is maximized by coating a finely divided inert support with the nonvolatile liquid. The coated support is carefully packed into the column so as not to develop empty spaces. Packed columns are usually $\frac{1}{4}$ or $\frac{1}{8}$ inch in diameter and range from 4 to 12 feet in length. These columns are particularly useful in the microscale laboratory, since they can be used for both analytical and preparative GC. Simple mixtures of 20–80 μL of material can often be separated into their pure components and collected at the exit port of the detector. Smaller samples (0.2–2.0 μL range) will exhibit better separation.

Capillary Columns. Capillary columns have no packing; the liquid phase is simply applied directly to the walls of the column. These columns are referred to as wall-coated, open-tubular (WCOT) columns. The reduction in surface area (compared to packed columns) is compensated for by tiny column diameters (perhaps 0.1 mm) and impressive lengths (100 m is not uncommon). Capillary columns are the most powerful columns used for analytical separations. Mixtures of several hundred compounds can be completely resolved on a capillary GC column. These columns require a more sophisticated and expensive chromatography instrument. Capillary columns, because of their tiny diameters, can accommodate only very small samples, perhaps 0.1 μL or less of a dilute solution. Capillary columns cannot be used for preparative separations.

Liquid Phase Once the sample is introduced on the column (in the carrier gas), it will undergo partition into the liquid phase. The choice of the liquid phase is particularly important because it directly affects the relative distribution coefficients.

In general, the stationary liquid phase controls the partitioning of the sample by two criteria. First, if little or no interaction occurs between the sample components and the stationary phase, the boiling point of the materials will determine the order of elution. Under these conditions, the highest boiling species will be the last to elute. Second, the functional groups of the components may interact directly with the stationary phase to establish different partition coefficients. Elution then depends on the particular binding properties of the sample components.

Some typical materials used as stationary phases are shown below.

Name	Stationary Phase	Maximum Temperature (°C)	Mechanism of Interaction
Silicone oil DC710, etc.	$R_3Si[OSiR_2]_nOSiR_3$	250	According to boiling point
Polyethylene glycol (Carbowax)	$HO[CH_2CH_2O]_nCH_2CH_2OH$	150	Relatively selective toward polar components
Diisodecyl phthalate	o-$C_6H_4[CO_2\text{-isodecyl}]_2$	175	According to boiling point

Oven Temperature The temperature of the column will also affect the separation. In general, the elution time of a sample will decrease as the temperature is increased. That is, retention times are shorter at higher temperatures. Higher boiling components tend to undergo diffusion broadening at low column temperatures because of the increase in retention times. If the oven temperature is too high, however, equilibrium partitioning of the sample with the stationary phase will not be established. Then the components of the mixture may elute together or be incompletely separated. Programmed oven temperature increases can speed up elution of the higher boiling components, but suppress peak broadening and therefore increase resolution. Temperature-programming capabilities require more sophisticated ovens and controllers.

Flow Rate The flow rate of the carrier gas is another important parameter. The rate must be slow enough to allow equilibration between the phases, but fast enough to ensure that diffusion will not defeat the separation of the components.

Column Length As noted, column length is an important factor in separation performance. As in distillations, column efficiency is directly proportional to column height, which determines the number of evaporation–condensation cycles. In a similar manner, increasing the length of a GC column allows more partition cycles to occur. Difficult-to-separate mixtures, such as the xylenes (very similar boiling points: *o*-xylene, 144.4 °C; *m*-xylene, 139.1 °C; and *p*-xylene, 138.3 °C), have a better chance of being separated on longer columns. In fact, both GC and distillation resolution data are described using the same term, *theoretical plates* (see Technique 2 and Experiments [3C] and [3D]).

Detector and Exit Port A successfully separated mixture will elute as its individual components at the instrument's exit port (also temperature controlled). To monitor the exiting vapors, a detector is placed in the gas stream just before the exit port (Fig. 5.1). After passing through the detector, the carrier gas and the separated sample components are then vented.

One widely used detector is the nondestructive, thermal conductivity detector, sometimes called a hot-wire detector. A heated wire in the gas stream changes its electrical resistance when a substance dilutes the carrier gas and thus

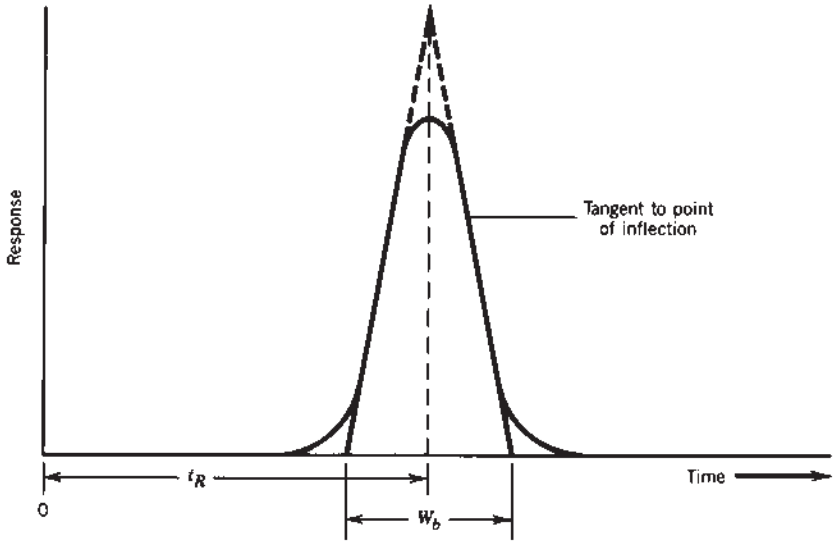

Figure 5.2 Schematic chromatogram.

changes its thermal conductivity. Helium has a higher thermal conductivity than most organic substances. When substances other than helium are present, the conductivity of the gas stream changes, which changes the resistance of the heated wire. The change in resistance is measured by comparing it to a reference detector mounted in a second (parallel) gas stream (Wheatstone bridge). The resulting electrical signal is plotted on a chart recorder, where the horizontal axis is time and the vertical axis is the magnitude of the resistance difference. The plot of resistance difference versus time is referred to as the gas *chromatogram.* The retention time (t_R) is defined as the time from sample injection to the time of maximum peak intensity. The baseline width (W_b) of a peak is defined as the distance between two points where tangents to the points of inflection cross the baseline (Fig. 5.2).

Capillary GC systems, and other GC systems used only for analytical separations, often use a flame-ionization detector (FID). In a flame-ionization detector, the gas eluting from the GC column is mixed with air (or oxygen) and hydrogen, and burned. The conductivity of the resulting flame is measured; it changes with the ionic content of the flame, which is proportional to the amount of carbon (from organic material) in the flame. The advantage of an FID is its high sensitivity; amounts of less than a microgram are easily detected. Its disadvantage is that it destroys (burns) the material it detects.

Theoretical Plates It is possible to estimate the number of theoretical plates (directly related to the number of distribution cycles) present in a column for a particular substance. The parameters are given in the relationship[1]

$$n = 16\left(\frac{t_R}{W_b}\right)^2$$

where the units of retention time (t_R) and baseline width (W_b) are identical (minutes, seconds, or centimeters). As in distillation columns, the larger the number of theoretical plates, *n,* the higher the resolution of the column and the better the separation.

---

[1] Berg, E. W. *Physical and Chemical Methods of Separation;* McGraw-Hill: New York, 1963, p. 111.

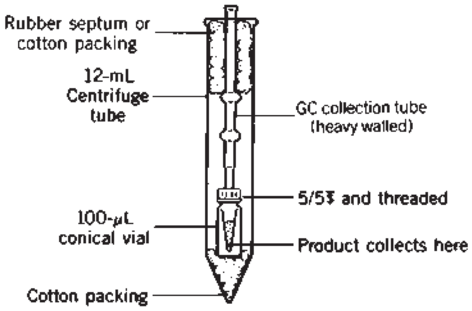

Figure 5.3 Gas chromatographic collection tube and 0.1-mL conical vial.

The efficiency of a system may be expressed as the *height-equivalent theoretical plate* (HETP) in centimeters (or inches) per plate. The HETP is related to the number of theoretical plates n by

$$\text{HEPT} = \frac{L}{n}$$

where L is the length of the column, usually reported in centimeters. The smaller the HETP, the more efficient the column.

www

The number of theoretical plates available in fractional distillation columns is limited by column holdup (see Techniques 2, 3, and website discussion of distillation theory). Thus, distillations of less than 500 μL are generally not practical. Gas-chromatographic columns, on the other hand, operate most efficiently at the microscale or submicroscale levels, where 500 μL would be an order of magnitude (even 3–8 orders of magnitude in the case of capillary columns) too large.

Fraction Collection Sequential collection of separated materials can be made by attaching suitable sample condensing tubes to the exit port (see Fig. 3.6).

Procedure for Preparative Collection. The collection tube (oven dried until 5 min before use) is attached to the heated exit port by the metal 5/5 $ joint. Sample collection is begun 30 s before detection of the expected peak on the recorder (based on previously determined retention values; refer to your local laboratory instructions) and continued until 30 s after the recorder's return to baseline. After the collection tube is detached, the sample can be analyzed directly when collected into a GC NMR collection tube[2] or transferred to the 0.1-mL conical GC collection vial. After the collection tube is joined to the vial (preweighed with cap) by the 5/5 $ joints, the system is centrifuged to force the sample down into the vial (Fig. 5.3). The collection tube is then removed, and the vial is capped and reweighed.

The efficiency of collection can exceed 90% with most materials, even with relatively low-boiling substances. In the latter case, the collection tube, after attachment to the instrument, is wrapped with a paper tissue. As the (oven-dried) tube is being wrapped, it is also being flushed by the carrier gas, which removes any traces of water condensation. The wrapping is then saturated with liquid nitrogen to cool the collection tube.

Preparative GC in the microscale laboratory often replaces the macroscale purification technique of fractional distillation. Distillation is impractical with less than 500 μL of liquid.

---

[2] Bressette, A. R. *J. Chem. Educ.* **2001**, *78,* 366.

Refer to Experiment [2] for specific experimental details on preparative GC applied to the separation of a number of binary (two-component) mixtures. These are designed as practice examples to give you experience with sample collection.

QUESTIONS

5-1. What is the main barrier to separating liquid mixtures of less than 500 μL by distillation?

5-2. A sample mixture of ethyl benzoate (bp 212 °C) and dodecane (bp 216.2 °C) is injected on two GC columns. Column A has DC710 silicone oil as the stationary phase, and column B uses polyethylene glycol as the stationary phase. Which substance would be certain to elute first from column A and would the same material be expected to elute first from column B? Which column, A or B, would be expected to give the better separation of these two substances?

5-3. Question 5-2 refers to separating a mixture of two high boiling liquids by gas chromatography. These materials have similar boiling points. List several GC variables and conditions that would make it easier to separate these substances by gas chromatography.

5-4. Capillary GC columns have better resolution than packed columns even though the enormous surface area provided by the packing material is absent in capillary columns. Why?

5-5. Preparative GC requires packed columns. Why is this technique limited to these lower resolution columns?

NOTE. Gas chromatographic purification of reaction products is suggested in the following experiments: Experiments [2], [3C], [3D], [5A], [5B], [8C], [9], [10], [13], [17], and [32].

Simple Distillation

TECHNIQUE
2

Distillation is the process of heating a liquid to the boiling point, condensing the heated vapor by cooling, and returning either a portion of, or none of, the condensed vapors to the distillation vessel. Distillation differs from reflux (see p. 23) only in that at least some of the condensate is removed from the boiling system. Distillations in which a fraction of the condensed vapors are returned to the boiling system are often referred to as being under "partial reflux." Two types of distillations will be described. Students are encouraged to refer to and study the more detailed discussion of distillation theory. The website also has a detailed discussion of the theory of steam distillation, which is used in Experiments [11C] and [32]. There are times when ordinary distillation may not be feasible for the separation of a liquid from dissolved impurities. The compound of interest may boil at a high temperature that is difficult to control with a simple apparatus, or it may tend to decompose or oxidize at high temperatures. If the compound is only sparingly soluble in water and any small amount of water can be removed with a drying agent, steam distillation may be the technique of choice. Compounds that are immiscible in water have very large positive deviations from Raoult's law. Therefore, the boiling temperature is generally lower than that of water and the compound.

← www

Simple distillation involves the use of the distillation process to separate a liquid from minor components that are nonvolatile, or that have boiling points at least 30–40 °C above that of the major component. A typical setup for a macroscale distillation of this type is shown in Figure 5.4. At the microscale

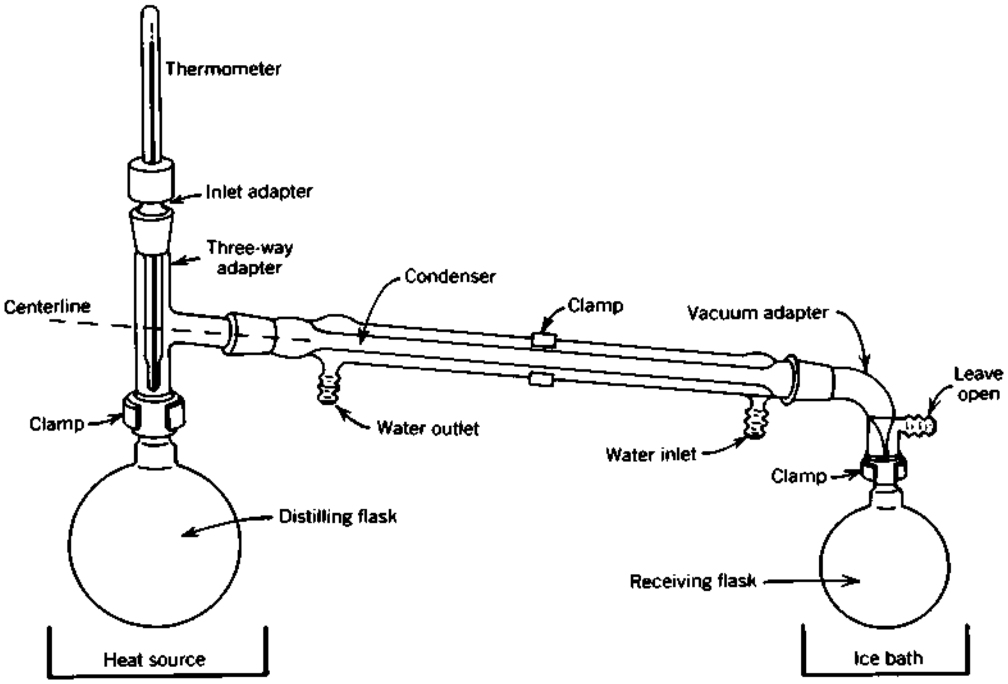

Figure 5.4 A complete simple distillation setup. (From Zubrick, J. W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, 2008. Reprinted by permission of John Wiley & Sons, Inc., New York.)

level, working with volumes smaller than 500 μL, GC techniques (see Technique 1) have replaced conventional microdistillation processes.[3] Semimicroscale simple distillation is an effective separation technique for volumes in the range of 0.5–2 mL. Apparatus that achieve effective separation of mixture samples in this range have been developed. One of the most significant of these designs is the classic Hickman still, shown in Figure 5.5. This still is used in several ways in the experiments described in Chapters 6, 7, and 10W for purifying solvents, carrying out reactions, and concentrating solutions. Experiment [3] introduces the use of the Hickman still.

www

In a distillation where liquid is to be separated from a nonvolatile solute, the vapor pressure of the liquid is lowered by the presence of the solute, but the vapor phase consists of only one component. Thus, except for the incidental transfer of non-volatile material by splashing, the material condensed should consist only of the volatile component (see Experiment [3A]).

We can understand what is going on in a simple distillation of two volatile components by referring to the phase diagrams shown in Figures 5.6 and 5.7. Figure 5.6 is the phase diagram for hexane and toluene. The boiling points of these liquids are separated by 42 °C. Figure 5.7 is the phase diagram for methylcyclohexane and toluene. Here the boiling points are separated by only 9.7 °C.

Imagine a simple distillation of the hexane–toluene pair in which the liquid in the pot is 50% hexane. In Figure 5.6, when the liquid reaches 80.8 °C it will be in equilibrium with vapor having a composition of 77% hexane. This result is indicated by the line *A–B.* If this vapor is condensed to a liquid of the same composition, as shown by line *B–C,* we will have achieved a significant enrichment of the condensate with respect to hexane. This change in composition is referred to

---

[3] Schneider, F. L. In *Monographien aus dem Gebiete der qualitätiven Mikroanalyse,* Vol. II: *Qualitative Organic Microanalysis;* Benedetti-Pichler, A. A., Ed.; Springer-Verlag: Vienna, 1964; p. 31.

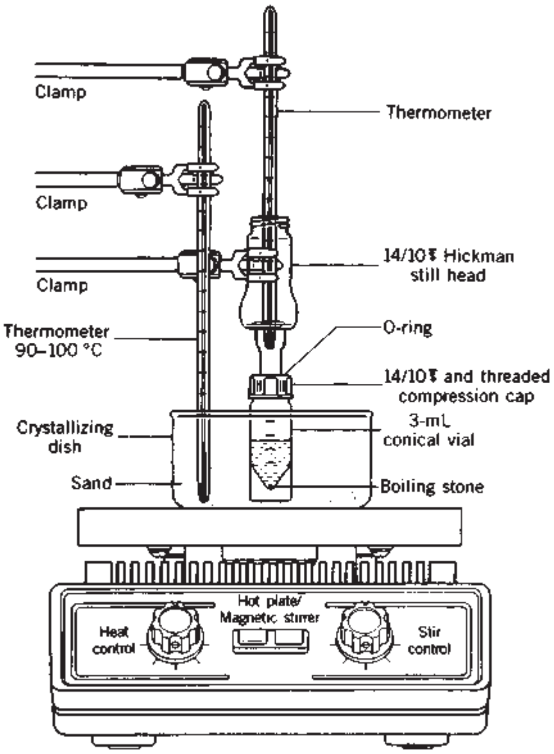
Clamp
Thermometer
Clamp
14/10₮ Hickman still head
Clamp
Thermometer 90–100 °C
O-ring
14/10₮ and threaded compression cap
3-mL conical vial
Crystallizing dish
Sand
Boiling stone
Hot plate/ Magnetic stirrer
Heat control
Stir control

Figure 5.5 Hickman still (14/10 ₮ with conical vial [3 mL]).

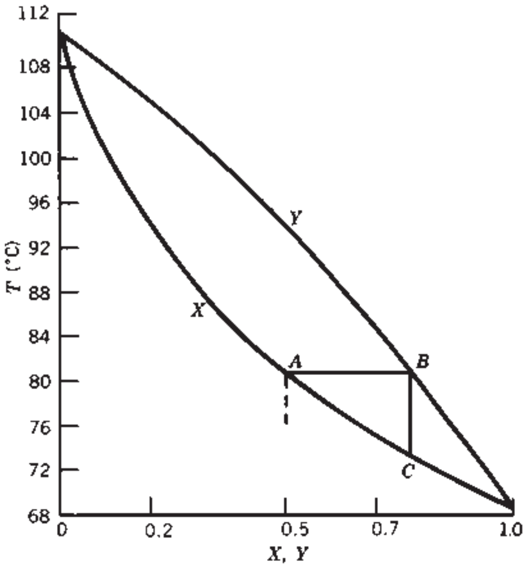

Figure 5.6 Temperature as a function of liquid composition (*X*) and vapor composition (*Y*): hexane and toluene.

as a simple distillation. The process of evaporation and condensation is achieved by the theoretical construct known as a *theoretical plate.* When this distillation is actually done with a Hickman still, some of the mixture will go through one evaporation and condensation cycle, some will go through two of these cycles, and some may be splashed more directly into the collar. A resolution (separation) of between one and two theoretical plates is generally obtained.

Referring to Figure 5.7, if we consider the same process for a 50% mixture of methylcyclohexane and toluene, the methylcyclohexane composition will increase to 58% for a distillation with one theoretical plate. Simple distillation may thus provide adequate enrichment of the MVC (more volatile constituent) if the boiling points of the two liquids are reasonably well separated, as they are for hexane and toluene. If the boiling points are close together, as they are for methylcyclohexane and toluene, the simple distillation will not provide much enrichment.

As we continue the distillation process and remove some of the MVC by condensing it, the residue in the heated flask becomes less rich in the MVC. This means that the next few drops of condensate will be less rich in the MVC. As the distillation is continued, the condensate becomes less and less rich in the MVC.

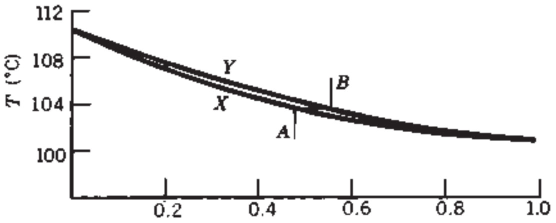

Figure 5.7 Temperature as a function of liquid composition (*X*) and vapor composition (*Y*): methylcyclohexane and toluene.

We can improve on simple distillation by repeating the process. For example, we could collect the condensate until about one-third is obtained. Then we could collect a second one-third aliquot in a separate container. Our original mixture would then be separated into three fractions. The first third would be richest in the MVC and the final third (the fraction remaining in the distillation pot) would be the richest in the least volatile component. If the MVC were the compound of interest, we could re-distill the first fraction collected (from a clean flask!) and collect the first third of the material condensing in that process. This simplest of all fractional distillation strategies is used in Experiment [3B].

QUESTIONS

5-6. What is the major drawback of trying to distill a 500-μL mixture of liquids, all with boiling points below 200 °C?

5-7. How might you separate the mixture discussed in question 5-6 if distillation were unsuccessful? Explain your choice.

5-8. If starting with an equal mixture of hexane and toluene, approximate the composition of hexane if the vapor at 94 °C is condensed to a liquid using the data presented in Figure 5.6.

5-9. Why do simple distillations require that the components of the mixture to be separated have boiling points that are separated by 40 °C or more?

5-10. Which constituent of an equimolar mixture makes the larger contribution to the vapor pressure of the mixture, the higher or lower boiling component? Explain.

NOTE. The following experiments use Technique 2: Experiments [3A], [3B], [11C], [29], and [32].

www → *[3A*$_{adv}$*].*

TECHNIQUE 3

Fractional Distillation

Fractional distillation can occur in a distillation system containing more than one theoretical plate. This process must be used when the boiling points of the components differ by less than 30–40 °C and fairly complete separation is desired. A fractionating column is needed to accomplish this separation. As discussed previously, a liquid– vapor composition curve (Fig. 5.8) shows that the lower boiling component of a binary mixture makes a larger contribution to the vapor composition than does the higher boiling component. On condensation, the liquid formed will be richer in the lower boiling component. This condensate will not be pure, however, and, in the case of components with close boiling points, it may be only slightly enriched. If the condensate is vaporized a second time, the vapor in equilibrium with this liquid will show a further enrichment in the lower boiling component. The trick to separating liquids with similar boiling points is to repeat the vaporization–condensation cycle many times. Each cycle is one *theoretical plate.* Several column designs, which achieve varying numbers of theoretical plates, are available for use at the macro level (Fig. 5.9).

Most distillation columns are designed so that fractionation efficiency is achieved by the very large surface area in contact with the vapor phase (and

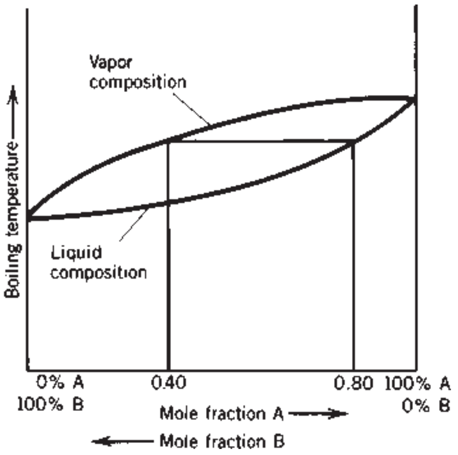

Figure 5.8 Liquid–vapor composition curve.

very similar to the way increased resolution is obtained on a GC column (see Technique 1). This increased surface area can be accomplished by packing the fractionating column with wire gauze or glass beads. Unfortunately, a large volume of liquid must be distributed over the column surface in equilibrium with the vapor. Furthermore, the longer the column the more efficient it becomes (see Technique 1), but longer columns also require additional liquid phase. The amount of liquid phase required to fill the column with a liquid—vapor equilibrium is called *column holdup*. Column holdup is essentially lost from the distillation because this volume can never go past the top of the distillation column; it can only return to the distillation pot upon cooling. Column holdup can be large compared to the total volume of material available for the distillation. With mixtures of less than 2 mL, column holdup precludes the use of the most common fractionation columns. Columns with rapidly spinning bands of metal gauze or Teflon have very low column holdup and have a large number of theoretical plates relative to their height (Fig. 5.10).

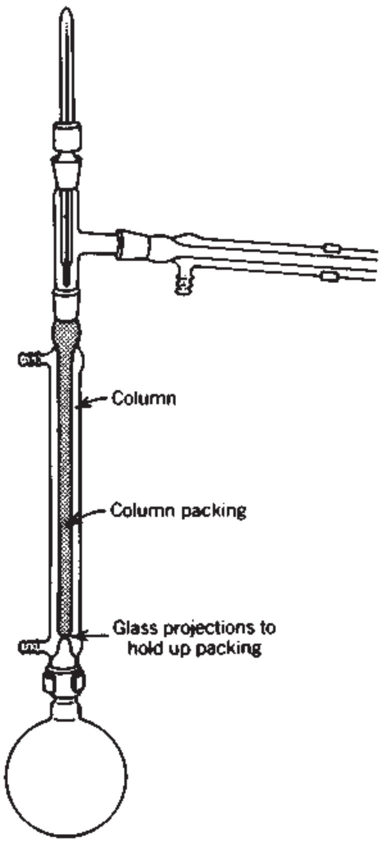

Figure 5.9 A fractional distillation setup. (From Zubrick, J.W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, 2008. Reprinted with permission of John Wiley & Sons, Inc., New York.)

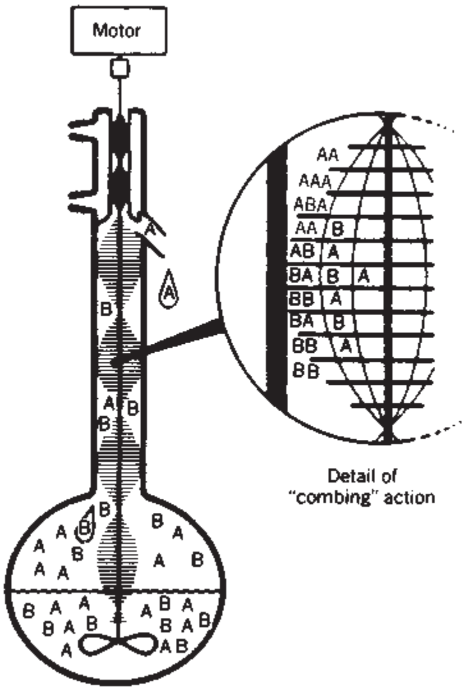

Figure 5.10 Schematic of a metal-mesh, spinning-band still.

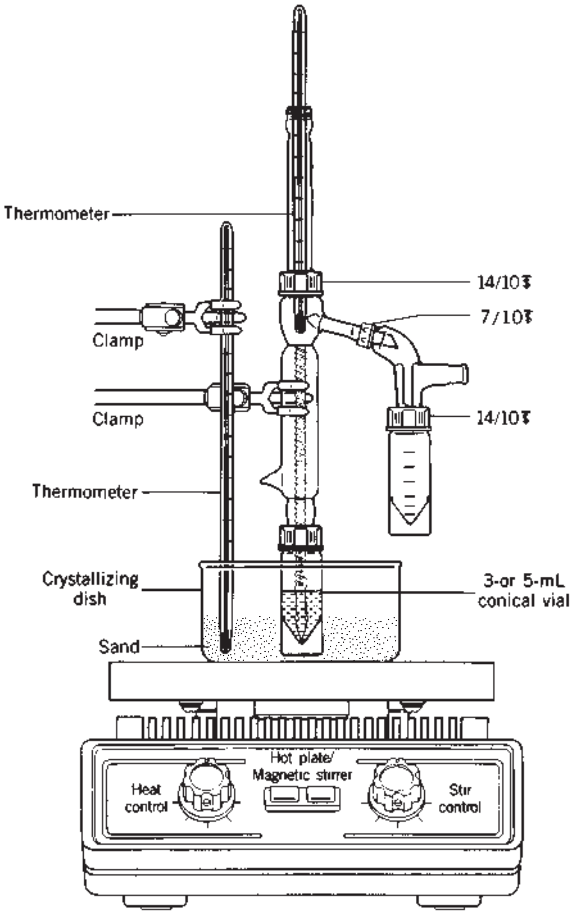
Thermometer
14/10₸
7/10₸
Clamp
Clamp
14/10₸
Thermometer
Crystallizing dish
3-or 5-mL conical vial
Sand
Hot plate/ Magnetic stirrer
Heat control
Stir control

Figure 5.11 Microspinning band distillation column (3 in.).

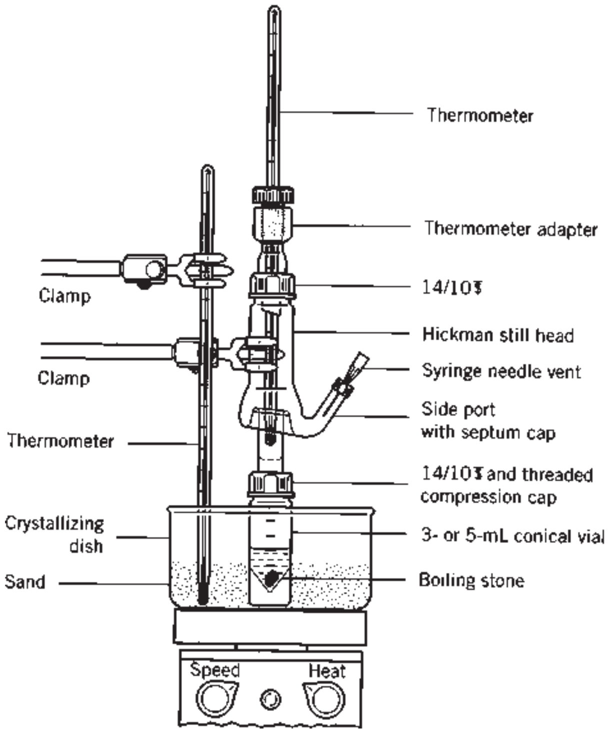
Thermometer
Thermometer adapter
14/10₸
Clamp
Hickman still head
Clamp
Syringe needle vent
Side port with septum cap
Thermometer
14/10₸ and threaded compression cap
Crystallizing dish
3- or 5-mL conical vial
Sand
Boiling stone
Speed
Heat

Figure 5.12 Hickman still with thermometer adapter.

Microscale spinning-band distillation apparatus (Fig. 5.11) can achieve nearly 12 theoretical plates and are simple enough to be used in the instructional laboratory. This still has a Teflon band that fits closely inside an insulated glass tube. The Teflon band has spiral grooves which, when the band is spun (1000–1500 rpm), rapidly return condensed vapor to the distillation pot. A powerful extension of this apparatus uses a short spinning band inside a modified Hickman still head (see Fig. 3.15). These stills are called Hickman–Hinkle stills; 4-cm Hickman–Hinkle columns can have more than 10 theoretical plates. The commercially available 2.5-cm version is rated at 6 theoretical plates. Experiments [3C] and [3D] involve fractional distillation with spinning-band columns.

The thermometer is positioned directly down the center of the distillation column, with the bulb just at the bottom of the well. It is very important to position both the still and the thermometer as vertically as possible; the thermometer must not touch the glass walls of the column (Fig. 5.12). Experiment [3B] uses the Hickman still for fractional distillation. A two-theoretical-plate distillation is obtained with this system on a two-component mixture by carrying out two sequential fractional distillations.

www → For a more detailed discussion of how spinning bands work, see the discussion of distillation.

QUESTIONS

5-11. Why is it very important that the hot vapor in microscale distillations climb the column very slowly?

5-12. Why might Teflon be the material of choice for constructing microscale spinning bands?

5-13. The spinning band overcomes two major problems of microscale distillations by wiping the liquid condensate rapidly from the column walls. What are these problems?

5-14. Why are spinning bands so effective at increasing the number of theoretical plates in distillation columns?

5-15. Why is steam distillation often used to isolate and purify naturally occurring plant substances?

NOTE. The following experiments use Technique 3: Experiments [3C] and [3D].

TECHNIQUE 4

Solvent Extraction

Solvent extraction is frequently used in the organic laboratory to separate or isolate a desired compound from a mixture or from impurities. Solvent extraction methods are readily adapted to microscale work because small quantities are easily manipulated in solution. Solvent extraction methods are based on the solubility characteristics of organic substances in the solvents used in a particular separation procedure. Liquid–liquid and solid–liquid extractions are the two major types of extractions used in the organic laboratory.

Intermolecular Properties: Solubility

Substances vary greatly in their solubility in various solvents, but a useful and generally true principle is that a substance tends to dissolve in a solvent that is chemically similar to itself. In other words, *like dissolves like.* The significant exceptions to this general statement are seen when solubilities are determined by acid–base properties.

Thus, to be soluble in water a compound needs to have some of the molecular characteristics of water. Alcohols, for example, have a hydroxyl group (—OH) bonded to a hydrocarbon chain or framework (R—OH). The hydroxyl group can be thought of as effectively half a water (H_2O) molecule; its polarity is similar to that of water. This polarity is due to the charge separation arising from the different electronegativities of the hydrogen and oxygen atoms. The O—H bond, therefore, is considered to have partial ionic character.

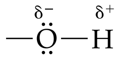

$$—\overset{\delta-}{\ddot{\underset{..}{O}}}—\overset{\delta+}{H}$$

Partial ionic character of the hydroxyl group

This polar, or partial ionic, character leads to relatively strong hydrogen bond formation between molecules with hydroxyl groups. Strong hydrogen bonding (shown here for the ethanol–water system) occurs in molecules that have a hydrogen atom attached to an oxygen, nitrogen, or fluorine atom—all three are quite electronegative atoms.

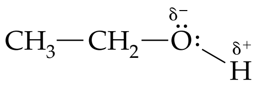

Ethanol

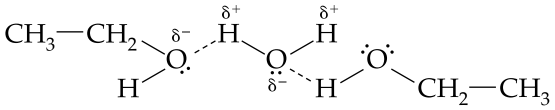

Hydrogen bond formation

Table 5.1 Comparison of Boiling Point Data

Name	Formula	MW	bp (°C)
Ethanol	CH_3CH_2OH	46	78.3
Propane	$CH_3CH_2CH_3$	44	−42.2
Methyl acetate	$CH_3CO_2CH_3$	74	54
Diethyl ether	$(CH_3CH_2)_2O$	74	34.6
Ethene	$CH_2{=}CH_2$	28	−102
Methylamine	CH_3NH_2	31	−6

The hydroxyl end of the ethanol molecule is very similar to water. When ethanol is added to water, therefore, they are miscible in all proportions. That is, ethanol is completely soluble in water and water is completely soluble in ethanol. This degree of solubility occurs because the attractive forces between the two molecules are nearly as strong as those between two water molecules; however, the attraction in the first case is somewhat weakened by the presence of the nonpolar ethyl group, CH_3CH_2—. Hydrocarbon groups attract each other only weakly, as demonstrated by their low melting and boiling points. Three examples of the contrast in boiling points between compounds of different structure, but similar molecular weight, are summarized in Table 5.1. Molecules that attract each other weakly (lower intermolecular forces) have lower boiling points.

Ethanol is completely miscible with water, but the solubility of octanol in water is less than 1%. Why the difference in solubilities between these two alcohols? The dominant structural feature of ethanol is its polar hydroxyl group; the dominant structural feature of octanol is its nonpolar alkyl group:

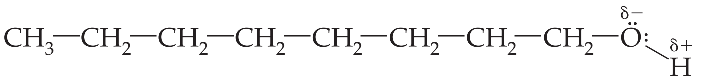

$CH_3{-}CH_2{-}CH_2{-}CH_2{-}CH_2{-}CH_2{-}CH_2{-}CH_2{-}\overset{\delta-}{\ddot{O}}{:}{-}\overset{\delta+}{H}$

Octanol

$CH_3{-}CH_2{-}\underset{\cdot\cdot}{\ddot{O}}{-}CH_2{-}CH_3$

Diethyl ether

As the size of the hydrocarbon section of the alcohol molecule increases, the intermolecular attraction between the polar hydroxyl groups of the alcohol and the water molecules is no longer strong enough to overcome the hydrophobic (lacking attraction to H_2O) nature of the nonpolar hydrocarbon section of the alcohol. On the other hand, octanol has a large nonpolar hydrocarbon group as its dominant structural feature. We might, therefore, expect octanol to be more soluble in less polar solvents, and, in fact, octanol is completely miscible with diethyl ether. Ethers are weakly polar solvents because a C—O bond is much less polar than an O—H bond (carbon is less electronegative than oxygen). Because both octanol and diethyl ether are rather nonpolar, each is completely soluble in the other. For compounds with both polar and nonpolar groups, in general, those compounds with five or more carbon atoms in the hydrocarbon portion of the molecule will be more soluble in nonpolar solvents, such as pentane, diethyl ether, or methylene chloride. Figure 5.13 summarizes the solubilities of a number of straight-chain alcohols, carboxylic acids, and hydrocarbons in water. As expected, most monofunctional compounds with more than five carbon atoms have solubilities similar to the hydrocarbons.

Several additional relationships between solubility and structure have been observed.

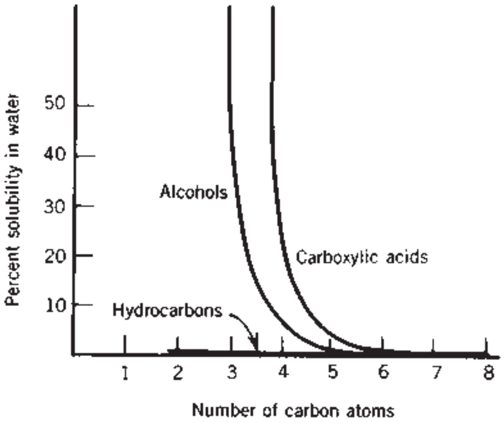

Figure 5.13 Solubility curve of acids, alcohols, and hydrocarbons. (From Kamm, O. *Qualitative Organic Analysis,* 2nd ed.; Wiley: New York, 1932. Reprinted with permission of John Wiley & Sons, New York.)

1. Branched-chain compounds have greater water solubility than their straight-chain counterparts, as illustrated in Table 5.2 with a series of alcohols.

2. The presence of more than one polar group in a compound will increase that compound's solubility in water and decrease its solubility in nonpolar solvents. For example, sugars, such as cellobiose, contain multiple hydroxyl and/or acetal groups and are water soluble and ether insoluble. Cholesterol, which has only a single hydroxyl group on its 27 carbon atoms, is insoluble in water and quite soluble in ether.

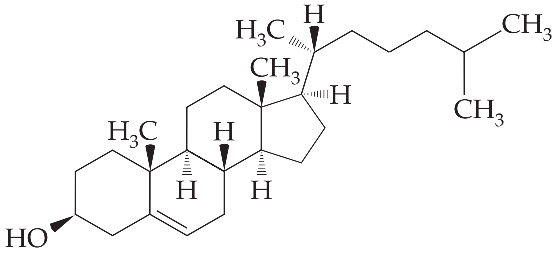

Cholesterol

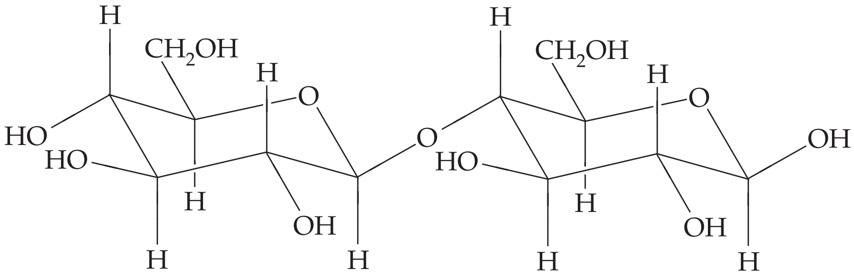

Cellobiose

Table 5.2 Water Solubility of Alcohols		
Name	Structural Formula	Solubility (g/100 g H_2O at 20 °C)
Hexanol	$CH_3(CH_2)_4CH_2OH$	0.6
Pentanol	$CH_3(CH_2)_3CH_2OH$	2.2
2-Pentanol	$CH_3(CH_2)_2CH(OH)CH_3$	4.3
2-Methyl-2-butanol	$(CH_3)_2C(OH)CH_2CH_3$	11.0

Table 5.3 Water Solubility of Amines

Name	Structural Formula	Solubility (g/100 g H_2O at 25 °C)
Ethylamine	$CH_3CH_2NH_2$	∞
Diethylamine	$(CH_3CH_2)_2NH$	∞
Trimethylamine	$(CH_3)_3N$	91
Triethylamine	$(CH_3CH_2)_3N$	14
Aniline	$C_6H_5—NH_2$	3.7
1,4-Diaminobenzene	$H_2N—C_6H_4—NH_2$	3.8

3. The presence of a chlorine atom, even though it lends some partial ionic character to the mostly covalent C—Cl bond, does not normally impart water solubility to a compound. In fact, compounds such as methylene chloride (CH_2Cl_2), chloroform ($CHCl_3$), and carbon tetrachloride (CCl_4) have long been used as solvents for extracting aqueous solutions. The latter two solvents are not often used nowadays, unless strict safety precautions are exercised, because they are potentially carcinogenic.

4. Most functional groups capable of forming a hydrogen bond with water increased the water solubility of a substance. For example, smaller alkyl amines have significant water solubility; the water-solubility data for a series of amines are summarized in Table 5.3.

The solubility characteristics of any given compound govern its distribution (*partition*) between the phases of two immiscible solvents (in which the material has been dissolved) when these phases are intimately mixed.

PARTITION (OR DISTRIBUTION) COEFFICIENT

A given substance X is partially soluble in each of two immiscible solvents. If X is placed in a mixture of these two solvents and shaken, an equilibrium will be established between the two phases. That is, substance X will partition (distribute) itself in a manner that is a function of its relative solubility in the two solvents:

$$X_{\text{solvent 1}} \rightleftharpoons X_{\text{solvent 2}}$$

The equilibrium constant, K_p, for this equilibrium expression is known as the *partition* or *distribution coefficient*:

$$K_p = \frac{[X_{\text{solvent 2}}]}{[X_{\text{solvent 1}}]}$$

The equilibrium constant is thus the ratio of the concentrations of the species, X, in each solvent for a given system at a given temperature. The partition coefficient can be conveniently estimated as the ratio of the solubility of X in solvent 1 vs. solvent 2:

$$K_p = \frac{\text{solubility of } X \text{ in solvent 2}}{\text{solubility of } X \text{ in solvent 1}}$$

When solvent 1 is water and solvent 2 is an organic solvent such as diethyl ether, the basic equation used to express the coefficient K_p is

$$K_p = \frac{(g/100\ mL)_{organic\ layer}}{(g/100\ mL)_{water\ layer}}$$

This expression uses grams per 100 mL for the concentration units. Note that the partition coefficient is dimensionless, so any concentration units may be used if the units are the same for both phases. For example, grams per liter (g/L), parts per million (ppm), and molarity (M) can all be used. If equal volumes of both solvents are used, the equation reduces to the ratio of the weights ($g_{organic}/g_{water}$) of the given species in the two solvents:

$$K_p = \frac{g_{organic\ layer}}{g_{water\ layer}}$$

Determination of the partition coefficient for a particular compound in various immiscible-solvent combinations often can give valuable information for isolating and purifying the compound by using extraction techniques. Thus, liquid–liquid extraction is a common separation technique used in organic as well as analytical laboratories.

Table 5.4 provides some examples of K_p values determined at room temperature for a number of compounds in the water–methylene chloride system.

Let us now look at a typical calculation for the extraction of an organic compound P from an aqueous solution using diethyl ether. We will assume that the $K_{p\ ether/water}$ value (partition coefficient of P between diethyl ether and water) is 3.5 at 20 °C. If a solution of 100 mg of P in 300 μL of water is extracted at 20 °C with 300 μL of diethyl ether, the following expression holds:

$$K_{p\ ether/water} = \frac{C_e}{C_w} = \frac{W_e/300\ \mu L}{W_w/300\ \mu L}$$

where

W_e = weight of P in the ether layer
W_w = weight of P in the water layer
C_e = concentration of P in the ether layer
C_w = concentration of P in the water layers

Since $W_w = 100 - W_e$, the preceding relationship can be written as

$$K_{p\ ether/water} = \frac{W_e/300\ \mu L}{(100 - W_e)/300\ \mu L} = 3.5$$

If we solve for the value of W_e, we obtain 77.8 mg; the value for W_w is 22.2 mg. Thus, we see that after one extraction with 300 μL of ether, 77.8 mg of P (77.8% of the total) is removed by the ether and 22.2 mg (22.2% of the total) remains in the water layer. Is it preferable to make a single extraction with the total quantity of solvent available, or to make multiple extractions with portions of the solvent? The second method is usually more efficient.

Table 5.4 **Representative K_p Values in CH_2Cl_2—H_2O**

Compound	K_p Value
Nitrobenzene	51.5
Aniline	3.3
1,2-Dihydroxybenzene	0.2

To illustrate, consider extracting the 100 mg of P in 300 μL of water with *two 150*-μL portions of diethyl ether instead of one 300-μL portion.

For the first 150-μL extraction,

$$\frac{W_e/150\ \mu\text{L}}{W_w/300\ \mu\text{L}} = \frac{W_e/150\ \mu\text{L}}{(100 - W_e)/300\ \mu\text{L}}$$

Solving for W_e, we obtain 63.6 mg. The amount of P remaining in the water layer (W_w) is then 36.4 mg. The aqueous solution is now extracted with the second portion of ether (150 μL). We then have

$$\frac{W_e/150\ \mu\text{L}}{(36.4 - W_e)/300\ \mu\text{L}} = 3.5$$

As before, by solving for W_e, we obtain 23.2 mg for the amount of P in the ether layer; $W_w = 13.2$ mg in the water layer.

The two extractions, each with 150 μL of ether, removed a total of 63.6 mg + 23.2 mg = 86.8 mg of P (86.8% of the total). The P left in the water layer is then 100 − 86.8, or 13.2 mg (13.2% of the total).

It can be seen from these calculations that the multiple-extraction technique is more efficient. The single extraction removed 77.8% of P; the double extraction (with the same total volume of ether) increased this to 86.8%. To extend this relationship, three extractions with one-third the total quantity of ether in each portion would be even more efficient. You might wish to calculate this to prove the point. Of course, there is a practical limit to the number of extractions that can be performed.

The multiple-extraction example shown here illustrates that several extractions with small volumes is more efficient than a single extraction procedure. This is always true *provided* the partition coefficient is neither very large nor very small. If the partition coefficient K_p for a substance between two solvents is very large ($K_p > 100$) or very small ($K_p < 0.01$), multiple extractions (using the same total amount of solvent) do not significantly increase the efficiency of the extraction process.[4]

Extraction

Liquid–Liquid Extraction. The more common type of extraction, liquid–liquid extraction, is used extensively. It is a very powerful method for separating and isolating materials at the microscale level. It is operationally not a simple process, so attention to detail is critical.

There are several important criteria to consider when choosing a solvent for the extraction and isolation of a component from a solution:

- The chosen extraction solvent must be immiscible with the solution solvent.
- The chosen extraction solvent must be favored by the distribution coefficient for the component being extracted.
- The chosen extraction solvent should be readily separated from the desired component after extraction. This usually means that it should have a low boiling point.
- The chosen organic extraction solvent must not react chemically with any component in the aqueous mixture being extracted.

---

[4] Palleros, D. R. *J. Chem. Educ.* **1995,** *72,* 319.

NOTE. The aqueous phase may be modified, as in acid–base extractions, but the organic solvent does not react with the components in the aqueous mixture. See Experiments [4B, 4C], pp. 146–150.

Microscale Extraction. A capped conical vial or a stoppered centrifuge tube is the best container for most microscale extractions, but a small test tube may be used. Note that a conical vial and a centrifuge tube have the same inner shape. This shape has the advantage that as the lower phase (layer) is withdrawn by pipet, the interface (boundary) between the two liquid phases becomes narrower and narrower, and thus easier to see, at the bottom of a conical container. This is not the case for a test tube. The centrifuge tube has the added advantage that if a solid precipitate must be separated or an emulsion broken up, it can easily be done using a centrifuge.

A good rule of thumb is that the container to be used for the extraction should be *at least three times the volume of liquid* you wish to extract.

Regardless of the container used, in any liquid–liquid extraction, the two immiscible solvents must be completely mixed to maximize the surface area of the interface between the two and allow partitioning of the solute. This can be accomplished by shaking (carefully to avoid leakage around the cap), using a Vortex mixer, or by adding a magnetic spin vane and then stirring with a magnetic stirrer.

Another important rule in the extraction process is that you should *never discard any layer until the isolation is complete.*

Let us consider a practical example. Benzanilide can be prepared by the in situ rearrangement of benzophenone oxime in acid solution:

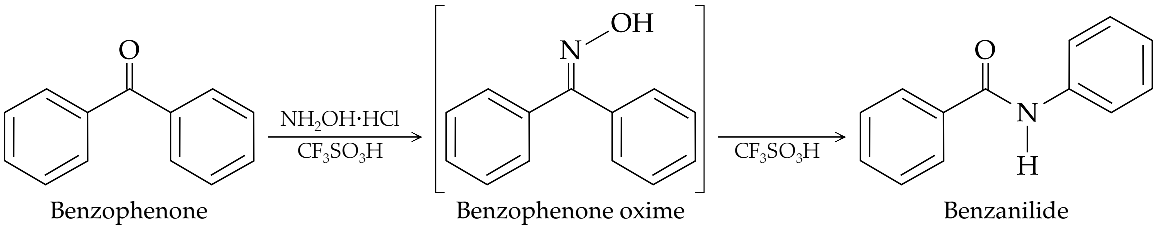

The benzanilide is separated from the reaction mixture by extraction with three 1.0-mL portions of methylene chloride solvent.

NOTE. Saying, for example, "extracted with three 1.0-mL portions of methylene chloride" means that three extractions are performed (one after the other, each using 1.0 mL of methylene chloride) and the three methylene chloride extracts are combined.

A microscale extraction process consists of two parts: (1) mixing the two immiscible solutions, and (2) separating the two layers after the mixing process.

1. Mixing. In the experimental procedure for the isolation of the benzanilide product, methylene chloride (1.0 mL) is added to the aqueous reaction mixture contained in a 5.0-mL conical vial (or centrifuge tube). The extraction procedure is outlined in the following steps:

Step 1. Cap the vial.

Step 2. Shake the vial gently to thoroughly mix the two phases (careful!)

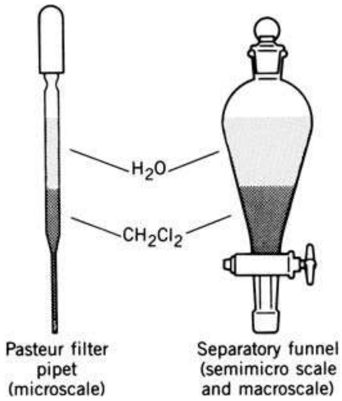
H_2O
CH_2Cl_2
Pasteur filter pipet (microscale)
Separatory funnel (semimicro scale and macroscale)

Figure 5.14 Extraction devices.

NOTE. The mixing may be carried out using a Vortex mixer or magnetic stirrer—see previous discussion.

Step 3. Carefully vent the vial by loosening the cap to release any pressure that may have developed.

Step 4. Allow the vial to stand on a level surface to permit the two phases to separate. A sharp phase interface should appear.

NOTE. For safety reasons it is advisable to place the vial in a small beaker to prevent tipping. If a volatile solvent such as ether is used, it is advisable to place the vial or centrifuge tube in a beaker of ice water to prevent loss of solvent during the transfers.

2. Separation. At the microscale level, the two phases are separated with a Pasteur filter pipet (a simple Pasteur pipet can be used in some situations), which acts as a miniature separatory funnel. The separation of the phases is shown in Figure 5.14.

A major difference between macro and micro techniques is that when microscale volumes are used, as just discussed, the mixing and separation are done in two parts. When macroscale volumes are used in a separatory funnel, mixing and separation are both done in the funnel in one step. The separatory funnel is an effective device for extractions with larger volumes, but it is not practical for microscale extractions because of the large surface areas involved.

Benzanilide is more soluble in methylene chloride than in water. Multiple extractions are performed to ensure complete removal of the benzanilide from the aqueous phase. The methylene chloride solution is the lower layer because it is more dense than water. The following list outlines the general method for an organic solvent more dense than water.

NOTE. (a) One technique is to hold the pipet across the palm of the hand and squeeze the bulb with the thumb and index finger. (b) Remember to have an empty tared vial available in which to place the separated phase. (c) A pipet pump (Figure 5.15) may be used to replace the bulb. One advantage with using a pipet pump is the dispensing of liquids in a more controlled fashion.

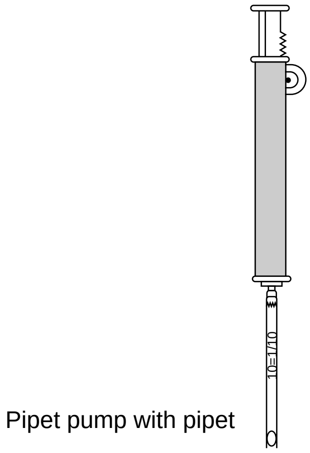

Figure 5.15 Pipet pump with pipet. (Reprinted with permission of John Wiley and Sons, Inc. from Szafran, Z.; Pike, R. M.; Foster, J. C. Microscale *General Chemistry Laboratory,* 2nd ed., p. 36. 2003.)

The recommended procedures are shown in Figures 5.16 and 5.17.

Step 1. Squeeze the pipet bulb to force air from the pipet.

Step 2. Insert the pipet into the vial until it is close to the bottom. Be sure to hold the pipet vertically.

Step 3. Carefully allow the bulb to expand, drawing only the lower methylene chloride layer into the pipet. This should be done in a smooth, steady manner so as not to disturb the interface between the layers. With practice, you can judge the amount that the bulb must be squeezed to just separate the layers. Keep the pipet vertical. *Do not tip the pipet back and allow liquid to enter the bulb! Do not suck liquid into the bulb!*

Step 4. (Step 4 is not shown in the figure.) Holding the pipet vertical, place it over and into the neck of an empty vial (as shown in Fig. 5.16, Step 2), and gently squeeze the bulb to transfer the methylene chloride solution into the vial. A second extraction can now be performed after adding

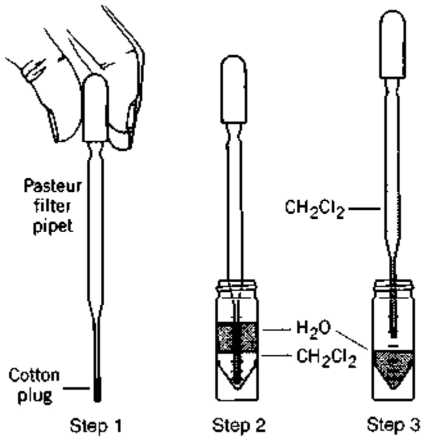

Figure 5.16 Pasteur filter pipet separation of two immiscible liquid phases; the more dense layer contains the product.

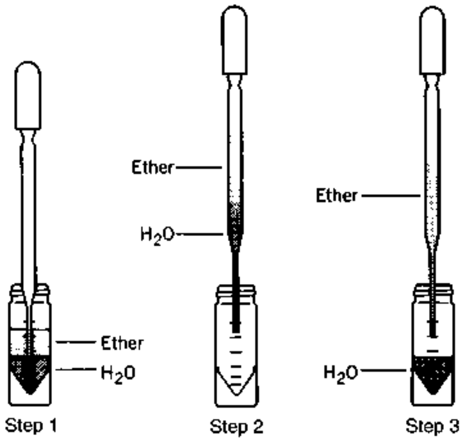

Figure 5.17 Pasteur filter pipet separation of two immiscible liquid phases; the less dense layer contains the product.

another portion of methylene chloride to the original vial. The procedure is repeated. Multiple extractions can be performed in this manner. Each methylene chloride extract is transferred to the same vial—that is, the extracts are combined. The reaction product has now been transferred from the aqueous layer (aqueous phase) to the methylene chloride layer (organic phase), and the phases have been separated.

In a diethyl ether–water extraction, the ether layer is less dense and thus is the upper layer (phase). An organic reaction product generally dissolves in the ether layer and is thus separated from water-soluble byproducts and other impurities. The procedure followed to separate the water–ether phases is identical to that described above for methylene chloride–water systems, except that here the top layer (organic layer) is transferred to the new container. The following list outlines the general method for an organic solvent less dense than water (refer to Fig. 5.17).

Step 1. Squeeze the pipet bulb to force air from the pipet and insert the pipet into the vial until it is close to the bottom. Then, draw **both** phases slowly into the pipet. Keep the pipet vertical. *Do not tip the pipet back and allow liquid to enter the bulb! Do not suck liquid into the bulb!* Try not to allow air to be sucked into the pipet, as this tends to mix the phases in the pipet. If mixing does occur, allow time for the interface to re-form.

Step 2. Return the aqueous layer (bottom layer) to the **original** container by gently squeezing the pipet bulb.

Step 3. Transfer the separated ether layer (top layer) to a new tared vial.

Separatory Funnel—Semimicroscale and Macroscale Extractions. A separatory funnel (Fig. 5.14) is effective for extractions carried out at the semimicroscale and macroscale levels. The mixing and separation are done in the funnel itself in one step. Many of you may be familiar with this device from the general chemistry laboratory. The same precautions as outlined above for microscale extraction should be observed here.

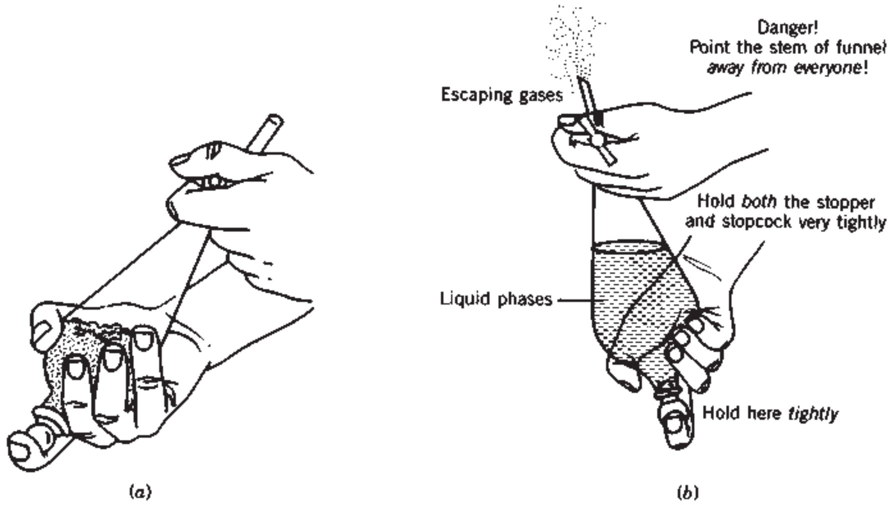
Danger!
Point the stem of funnel *away from everyone!*
Escaping gases
Hold *both* the stopper and stopcock very tightly
Liquid phases
Hold here *tightly*
(*a*)
(*b*)

Figure 5.18 (a) Correct position for holding a separatory funnel while shaking. (b) Correct method for venting a separatory funnel.

NOTE. The funnel size should be such that the total volume of solution is less than half the total volume of the funnel. If the funnel has a ground-glass stopcock and/or stopper, the ground-glass surfaces must be lightly greased to prevent sticking, leaking, or freezing. If Teflon stoppers and stopcocks are used, grease is not necessary because these are self-lubricating.

Step 1. Close the stopcock of the separatory funnel.

Step 2. Add the solution to be extracted, after first making sure that the *stopcock is closed.* The funnel should be supported in an iron ring attached to a ring stand or rack on the lab bench.

Step 3. Add the proper amount of extraction solvent (about one-third of the volume of the solution to be extracted is a good rule of thumb) and place the stopper on the funnel.

Step 4. Remove the funnel from the ring stand, keeping the stopper in place with the index finger of one hand, and holding the funnel in the other hand with your fingers positioned so they can operate the stopcock (Fig. 5.18*a*).

Step 5. Carefully invert the funnel (make sure its stem is pointing up, and not pointing at you or anyone else). Slowly open the stopcock to release any built-up pressure (Fig. 5.18*b*). Close the stopcock and then shake the funnel for several seconds. Position the funnel for venting (make sure the stem is pointing up, and not pointing at you or anyone else). Open the stopcock to release built-up pressure. Repeat this process 2–4 times. Then, close the stopcock and return the funnel upright to the iron ring.

Step 6. Allow the layers to separate and then remove the stopper.

Step 7. Place a suitable clean container just below the tip of the funnel. Gradually open the stopcock and drain the bottom layer into the clean container.

Step 8. Remove the upper layer by pouring it from the top of the funnel. This way it will not become contaminated with traces of the lower layer found in the stem of the funnel.

When aqueous solutions are extracted with a *less dense solvent*, such as ether, the bottom, aqueous layer can be drained *into its original container*. Once the top (organic) layer is removed from the funnel, the aqueous layer can then be returned for further extraction. Losses can be minimized by rinsing the original container with a small portion of the extraction solvent, which is then added to the funnel. When the extraction solvent is denser than the aqueous phase (e.g., methylene chloride), the aqueous phase is the top layer, and therefore is kept in the funnel for subsequent extractions.

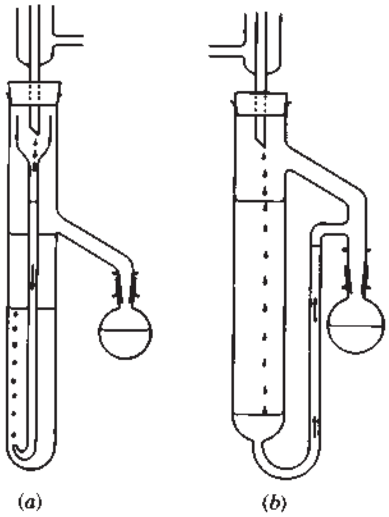

Figure 5.19 Early designs for single-stage extractors: (a) Kutscher–Steudel extractor; (b) Wehrli extractor.

Continuous Liquid–Liquid Extraction. Continuous extraction of liquid–liquid systems is also possible and particularly valuable when the component to be separated is only slightly soluble in the extraction solvent. The advantage of using continuous extraction is that it can be carried out with a limited amount of solvent. In batchwise extractions a prohibitive number of individual extractions might have to be performed to accomplish the same overall extraction. Specialized apparatus, however, is required for continuous liquid–liquid extraction.

Two types of continuous extraction apparatus are often used to isolate various species from aqueous solutions using less dense and more dense immiscible solvents (e.g., diethyl ether and methylene chloride) (Fig. 5.19).

The extraction is carried out by allowing the condensate of the extraction solvent, as it forms on the condenser on continuous distillation, to drop through an inner tube (see Fig. 5.19*a* in the case of the less dense solvent) and to percolate up through the solution containing the material to be extracted. This inner tube usually has a sintered glass plug on its end, which generates smaller droplets of the solvent and thus increases the efficiency of the procedure. The extraction solution is then returned to the original distilling flask. Eventually, in this manner, the desired material, extracted in small increments, is collected in the boiling flask and can then be isolated by concentrating the collected solution. This method works on the premise that fresh portions of the less-dense phase are continuously introduced into the system, and it is often used in those instances where the organic material to be isolated has an appreciable solubility in water. In the case of a more dense extraction solvent (see Fig. 5.19*b*) the system functions in much the same fashion, but in this case the inner tube is removed and the condensed vapors percolate directly through the lighter phase (the phase to be extracted) to form the lower layer. This layer can cycle back to the distillation flask through a small-bore tubing connection from the bottom of the receiver flask to the distillation flask. Continuous liquid–liquid extraction is useful for removing extractable components from those having partition ratios that approach zero. Note that this method requires a very long period of time.

Separation of Acids and Bases. The separation of organic acids and bases is another important and extensive use of the extraction method. The distribution coefficients of organic acids and bases are affected by pH when one of the solvents is water. An organic acid that is insoluble in neutral water (pH 7) becomes soluble when the water is made basic with an aqueous sodium hydroxide solution. The acid and the sodium hydroxide quickly react to form a sodium carboxylate salt, RCO_2^-, Na^+. The salt is, of course, ionic and therefore it readily dissolves in the water. Thus, the acid–base reaction reverses the solubility characteristics of a water-insoluble organic acid.

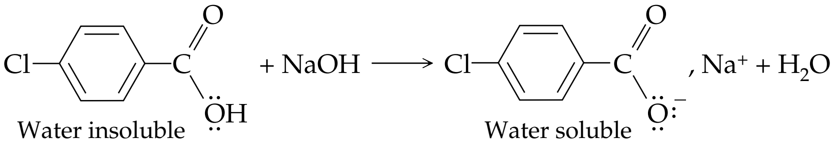
+ NaOH
, Na⁺ + H₂O
Water insoluble
Water soluble

The water phase may then be extracted with an immiscible organic solvent to remove any impurities, leaving the acid salt in the water phase. Neutralizing the water layer with hydrochloric acid (to pH ≤ 7) reprotonates the carboxylate salt to reform the carboxylic acid, and causes the purified water-insoluble organic acid to precipitate (if it's a solid). In a similar fashion, water-insoluble organic bases, such as amines (RNH_2), can be rendered water soluble by treatment with dilute hydrochloric acid to form water-soluble hydrochloride salts (e.g., Experiment [23]).

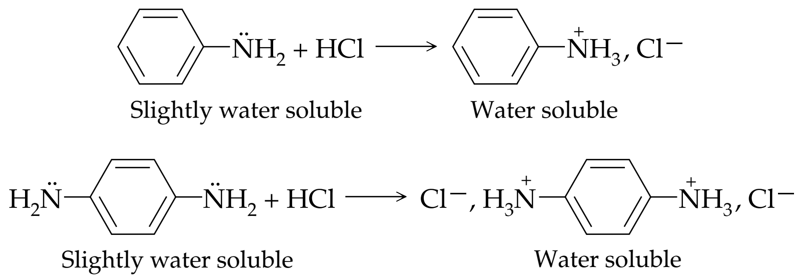
+ HCl
Slightly water soluble
Water soluble
+ HCl
Slightly water soluble
Water soluble

Extraction procedures can be used to separate mixtures of solids. For example, the flow chart below diagrams a sequence used to separate a mixture made up of an aromatic carboxylic acid ($ArCO_2H$), an aromatic base ($ArNH_2$), and a neutral aromatic compound (ArH). Aromatic compounds are discussed here simply because they are likely to be crystalline solids.

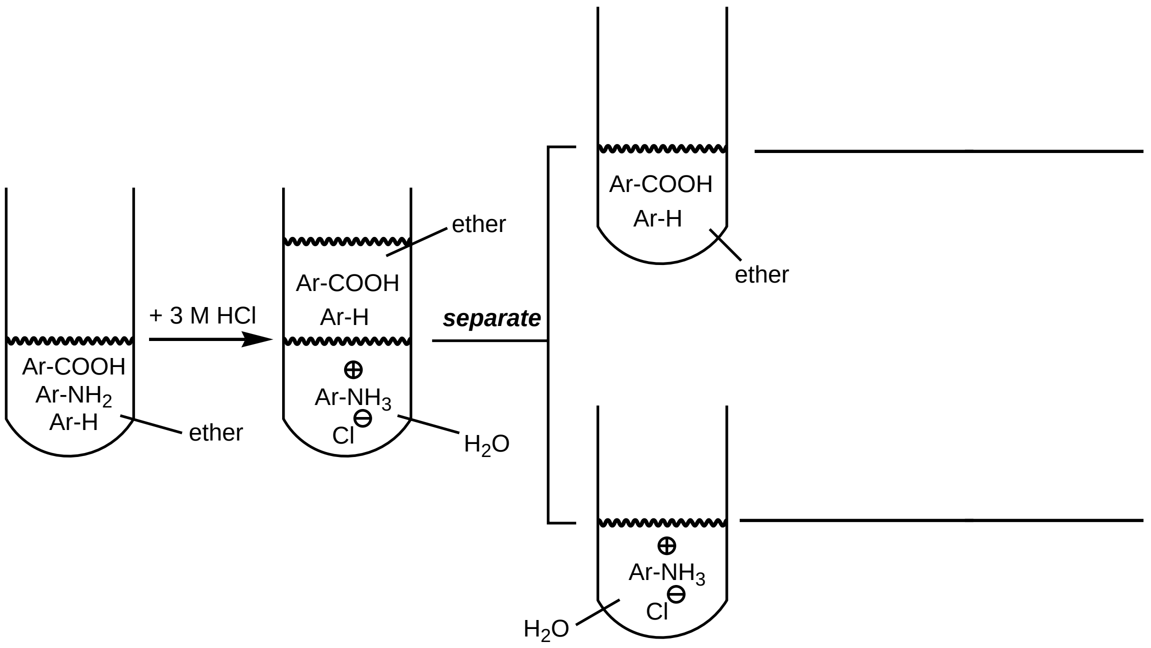
Ar-COOH
Ar-NH2
Ar-H
ether
+ 3 M HCl
ether
Ar-COOH
Ar-H
Ar-NH3
Cl
H2O
separate
Ar-COOH
Ar-H
ether
Ar-NH3
Cl
H2O

In this example, we assume that the organic acid and base are solids. If either or both were liquids, an additional extraction of the final acidic aqueous or alkaline solution with ether, followed by drying and concentration, would be required to isolate the acidic or basic component.

Salting Out. Most extractions in the organic laboratory involve water and an organic solvent. Many organic compounds have partial solubility in both solvents. To extract them from water, the partition coefficient (between the organic solvent and water) can be shifted in favor of the organic layer by saturating the water layer with an inorganic salt, such as sodium chloride. Water molecules prefer to solvate the polar ions (in this case sodium and chloride ions), and thus free the neutral organic molecules to migrate into the organic phase. Another way to think of this is to realize that the ionic solution is more polar than pure water, so the less polar organic molecules are less soluble than in pure water. Forcing an organic material out of a water solution by adding an inorganic salt is called *salting out.*

Salting out can also be effectively used for the *preliminary drying* of the wet organic layer that results from an extraction process. (Diethyl ether, in particular, can dissolve a fair amount of water.) Washing this organic layer with a saturated salt solution removes most of the dissolved water into the aqueous phase. This makes further drying of the organic phase with solid drying agents easier and much more effective (see Drying Agents below).

Solid–Liquid Extraction

The simplest form of solid–liquid extraction involves treating a solid with a solvent and then decanting or filtering the solvent extract away from the solid. An example of this technique (see Experiment [11A]) is extracting usnic acid from

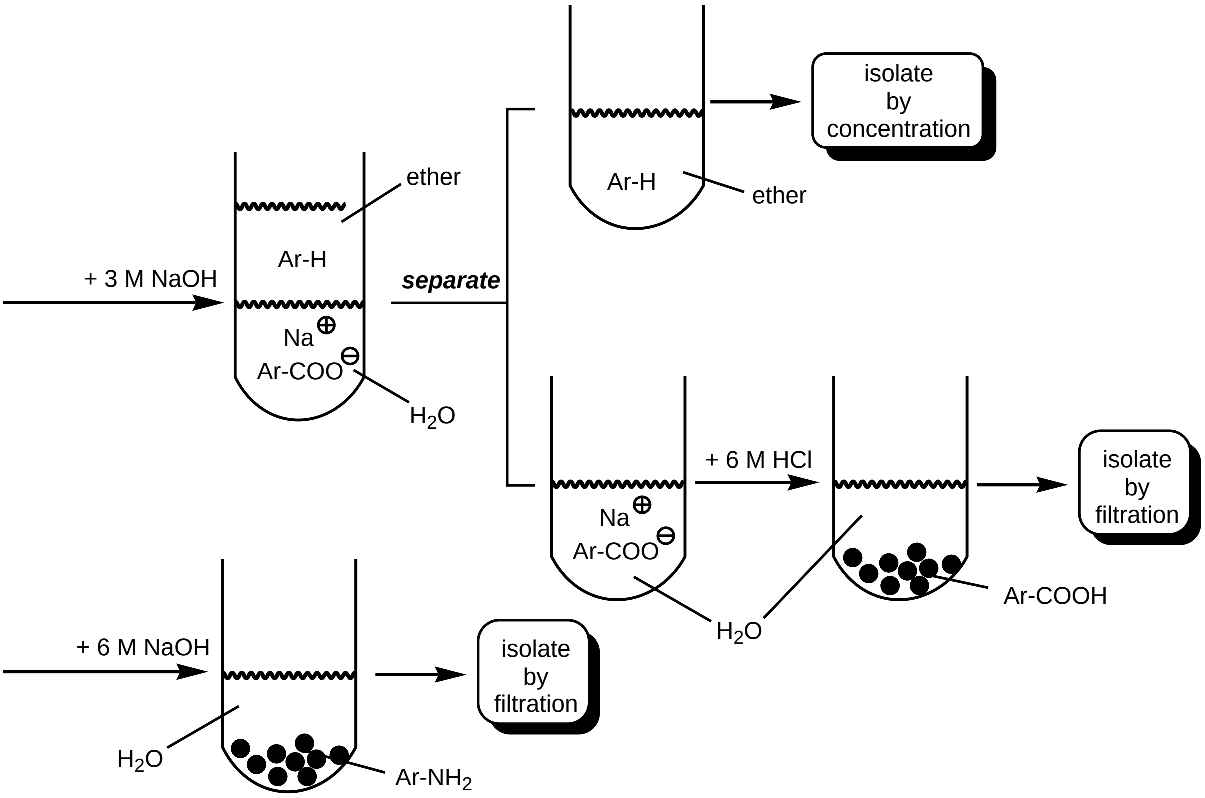

its native lichen with acetone. This type of extraction is most useful when only one main component of the solid phase has appreciable solubility in the solvent. The extraction of caffeine from tea (see Experiment [11B]) is another example of this method; it is accomplished by heating the tea in an aqueous solution of sodium carbonate. This approach works well because the water swells the tea leaves and allows the caffeine to be extracted more readily.

Microscale extractions of trimyristin from nutmeg, and cholesterol from gallstones, have been described.[5] Diethyl ether was used as the solvent in both cases. A packed Pasteur pipet column was used for filtering, drying (nutmeg experiment), and decolorizing (gallstone experiment).

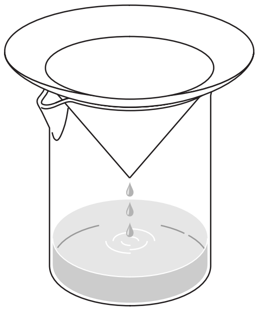

Figure 5.20 A solid–liquid continuous extraction apparatus.

Herrera and Almy described a simple continuous extraction apparatus (Fig. 5.20).[6] The apparatus is constructed from a 50-mL beaker and a paper cone prepared from a 9-cm disk of filter paper (nonfluted), which rests on the lip of the beaker. A small notch is cut in the cone to allow solvent vapor to pass around it. The extraction solvent is placed in the beaker; the solid material to be extracted is placed in the cone. A watch glass containing 2–3 g of ice is placed on top of the assembly to act as the condenser and to hold the paper cone in place. As the ice melts, the water is removed and replaced with fresh ice. The beaker is heated on a hot plate in the hood (some solvent evaporates during the extraction process and may need to be replaced). The concentrated solution collected in the beaker is then cooled and the solid product is isolated by filtration or is recrystallized. This system needs to be attended at all times, but works reasonably well for brief extractions.

Various apparatus have been developed for use when longer extraction periods are required. They all use what is called a *countercurrent process*. The best-known apparatus is the Soxhlet extractor, first described in 1879 (Fig. 5.21).[7] The solid sample is placed in a porous thimble. The extraction-solvent vapor, generated by refluxing the extraction solvent contained in the distilling pot, passes up through the vertical side tube into the condenser. The liquid condensate then drips onto the solid, which is extracted. The extraction solution passes through the pores of the thimble, eventually filling the center section of the Soxhlet. The siphon tube also fills with this extraction solution and when the liquid level reaches the top of the tube, siphoning action returns the thimbleful of extract to the distillation pot. The cycle is automatically repeated many times, concentrating the extract in the distillation pot. The advantage of this arrangement is that the process may be continued automatically and unattended for as long as necessary. The solvent is then removed from the extraction solution collected in the pot, providing the extracted compound(s). Soxhlet extractors are available from many supply houses and can be purchased in various sizes. Of particular interest to us is the microscale variety, which is effective for small amounts of material and is now commercially available.[8]

Drying Agents

Organic extracts separated from aqueous phases usually contain traces of water. Even washing with saturated salt solution (see Salting Out above) cannot

---

[5] Vestling, M. M. *J. Chem. Educ.* **1990,** *67,* 274.
[6] Herrera, A.; Almy, J. *J. Chem. Educ.* **1998,** *75,* 83.
[7] Soxhlet, F. *Dinglers Polytech. J.* **1879,** *232,* 461.
[8] Microscale Soxhlet equipment is available from ACE Glass, Inc., 1430 Northwest Boulevard, Vineland, NJ 08360.

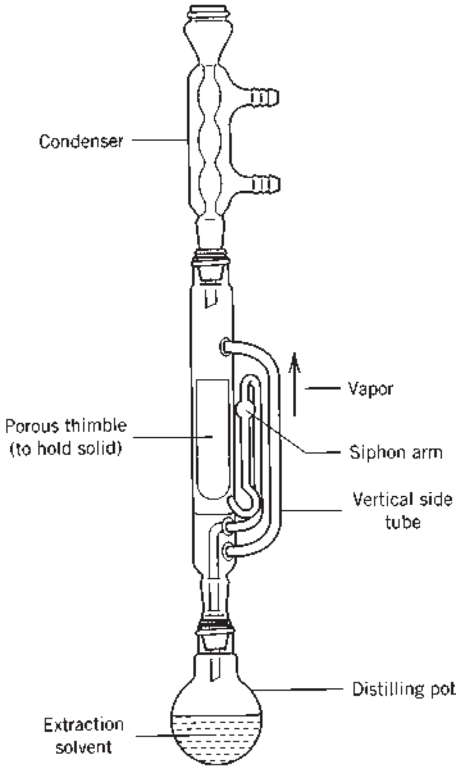

Figure 5.21 Soxhlet extractor.

remove all of the water. Organic extracts must therefore be dried to remove any residual water before the solvent is evaporated or further purification is performed. Organic extracts can be conveniently dried with an anhydrous inorganic salt, such as magnesium sulfate, sodium sulfate, or calcium sulfate. These salts readily absorb water and form insoluble hydrates, thus removing the water from the wet organic phase. The hydrated solid can then be removed from the dried solution by filtration or by decanting (pouring) the solution away from the solid. Although many drying agents are known, not every drying agent can be used in every case. The ideal drying agent should dry the solution quickly, have a high capacity for water, cost little, and not react with the material being dried.[9]

Table 5.5 summarizes the properties of some of the more common drying agents used in the laboratory.

Make sure that the solid drying agent is in its *anhydrous* form. Sodium sulfate is a good general-purpose drying agent and is usually the drying agent of choice at room temperature. Use the granular form, if at all possible.

Magnesium sulfate is supplied as a fine powder (high surface area). It has a high water capacity and is inexpensive; it dries solutions more quickly than does sodium sulfate. The disadvantage of magnesium sulfate is that the desired product (or water molecules) can become trapped on the surface of the fine particles. If it is not thoroughly washed after separation, precious product may

---

[9] Quantitative studies on the efficiency of drying agents for a wide variety of solvents have been reported. See Burfield, D. R.; Smithers, R. H. *J. Org. Chem.* **1983,** *48,* 2420, and references therein. Other useful information can be found in Armarego, W. L. F.; Chai, C. L. L. *Purification of Laboratory Chemicals,* 5th ed.; Elsevier: New York, 2003, and in Ridduck, J. A.; Bunger, W. B.; Sakano, T. K. *Organic Solvents, Physical Properties and Methods of Purification,* 4th ed.; Wiley: New York, 1986.

Table 5.5 **Properties of Common Drying Agents**

Drying Agent	Formula of Hydrate	Comments
Sodium sulfate	$Na_2SO_4 \cdot 10H_2O$	Slow to absorb water and inefficient, but inexpensive and has a high capacity. Loses water above 32 °C Granular form available.
Magnesium sulfate	$MgSO_4 \cdot 7H_2O$	One of the best. Can be used with nearly all organic solvents. Usually in powder form.
Calcium chloride	$CaCl_2 \cdot 6H_2o$	Relatively fast drying, but reacts with many oxygen- and nitrogen-containing compounds. Usually in granular form.
Calcium sulfate	$CaSO_4 \cdot \frac{1}{2} H_2O$	Very fast and efficient, but has a low dehydration capacity.
Silica gel	$(SiO_2)_m \cdot nH_2O$	High capacity and efficient. Commercially available t.h.e. SiO_2 drying agent is excellent.[a]
Molecular sieves	$[Na_{12}(Al_{12}Si_{12}O_{48})] \cdot 27H_2O$	High capacity and efficient. Use the 4-Å size.[b]

[a] Available from EMD Chemicals, 10394 Pacific Center Court, San Diego, CA 92121.

[b] Available from Sigma-Aldrich, Inc., 940 West Saint Paul Ave., Milwaukee, WI 53233.

be lost. Furthermore, it is usually more difficult to remove a finely powdered solid agent, which may pass through the filter paper (if used) or clog the pores of a fine porous filter. A smaller surface area translates into less adsorption of product on the surface and easier separation from the dried solution.

Molecular sieves have pores or channels in their structures. A small molecule such as water can diffuse into these channels and become trapped. The sieves are excellent drying agents, have a high capacity, and dry liquids completely. The disadvantages are that they dry slowly and are more expensive than the more common drying agents.

Calcium chloride is very inexpensive and has a high capacity. Use the granular form. Do not use it to dry solutions of alcohols, amines, or carboxylic acids because it can react with these substances.

Calcium sulfate is often sold under the trade name of Drierite. It is a somewhat expensive drying agent. Do not use the blue Drierite (commonly used to dry gases) because the cobalt indicator (blue when dry, pink when wet) may leach into the solvent.

The amount of drying agent needed depends on the amount of water present, on the capacity of the solid desiccant to absorb water, and on its particle size (actually, its surface area). If the solution is wet, the first amount of drying agent will clump (molecular sieves and t.h.e. SiO_2 are exceptions). Add more drying agent until it appears mobile when you swirl the liquid. A solution that is no longer cloudy is a further indication that the solution is dry. Swirling the contents of the container increases the rate of drying; it helps establish the equilibrium for hydration:

$$\underset{\text{Anhydrous solid}}{\text{Drying agent} + nH_2O} \rightleftharpoons \underset{\text{Solid hydrate}}{\text{Drying agent} \cdot nH_2O}$$

Most drying agents achieve approximately 80% of their drying capacity within 15 min; longer drying times are generally unnecessary. The drying agent may be added directly to the container of the organic extract, or the extract may be passed through a Pasteur filter pipet packed with the drying agent. A funnel fitted with a cotton, glass wool, or polyester plug to hold the drying agent may also be used.

As for the most common question asked with this technique—Is this "dry"?—you should be encouraged to have in your lab a series of flasks which contain a set quantity of solvent and drying agent. The difference with each flask within the series is the percentage of water which allows for a visual comparison of what is and what is not "dry."

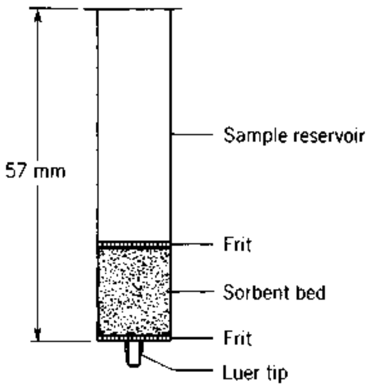

Figure 5.22 Polyethylene solid-phase extraction column.

Solid-Phase Extraction

In the modern research laboratory, the traditional liquid–liquid extraction technique may be replaced by the solid-phase extraction method.[10] The advantages of this newer approach are that it is rapid, it uses only small volumes of solvent, it does not form emulsions, isolated solvent extracts do not require a further drying stage, and it is ideal for working at the microscale level. This technique is finding wide acceptance in the food industry and in the environmental and clinical area, and it is becoming the accepted procedure for the rapid isolation of drugs of abuse and their metabolites from urine. Solid-phase extraction is accomplished using prepackaged, disposable, extraction columns. A typical column is shown in Figure 5.22. The columns are available from several commercial sources.[11]

The polypropylene columns can be obtained packed with 100–1000 mg of 40-μm sorbent sandwiched between two 20-μm polyethylene frits. The columns are typically 5–6 cm long. Sample volumes are generally 1–6 mL.

The adsorbent (stationary phase) used in these columns is a nonpolar adsorbent chemically bonded to silica gel. In fact, they are the same nonpolar adsorbents used in the reversed-phase high-performance liquid chromatography (HPLC). More specifically, the adsorbents are derivatized silica gel where the —OH groups of the silica gel have been replaced with siloxane groups by treating silica gel with the appropriate organochlorosilanes.

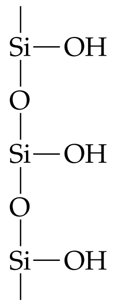

Silica surface

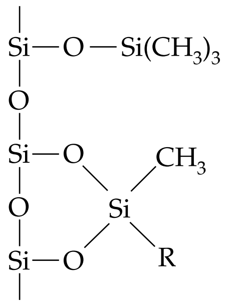

Chemically bonded silica surface

---

[10] For a description of this method see Zief, M.; Kiser, R. *Am. Lab.* **1990,** *22 70;* Zief, M. *NEACT J.* **1990,** *8,* 38; Hagen, D. F.; Markell, C. G.; Schmitt, G. A.; Blevins, D. D. *Anal. Chim. Acta* **1990,** *236,* 157; Arthur, C. L.; Pawliszyn, J. *Anal. Chem.* **1990,** *62,* 2145; Dorsey, J.; Dill, K. A. *Chem. Rev.* **1989,** *89,* 331; Zubrick, J. W. *The Organic Chem Lab Survival Manual,* 7th ed., Wiley: New York, 2008; Simpson, N. J. K, Ed. *Solid-Phase Extraction: Principles, Techniques and Applications,* Marcel Dekker: New York, 2000; "Solid Phase Extraction." Retrieved March 19, 2009 from www.sigmaaldrich.com/analytical-chromatography/sample-preparation/spe.html.

[11] These columns are available from Analytichem International, J. T. Baker, Inc., Supelco, Inc., Aldrich Chemical, Waters Associates and Biotage (a Division of Dyax Corp).

Two of the most popular nonpolar packings are those containing R groups consisting of an octadecyl ($C_{18}H_{37}$—)or phenyl (C_6H_5—) group. These packing materials (stationary phases) can adsorb nonpolar (like attracts like) organic material from aqueous solutions. The adsorbed material is then eluted from the column using a solvent strong (nonpolar) enough to displace it, such as methanol, methylene chloride, or hexane. The analyte capacity of bonded silica gels is about 10–20 mg of analyte per gram of packing.

An example of a typical solid-phase extraction is the determination of the amount of caffeine in coffee using a 1-mL column containing 100 mg of octadecyl-bonded silica. This efficient method isolates about 95% of the available caffeine. The column is conditioned by flushing 2 mL of methanol followed by 2 mL of water through the column. One milliliter of a coffee solution (~0.75 mg of caffeine/mL) is then drawn through the column at a flow rate of 1 mL/min. The column is washed with 1 mL of water and air dried (vacuum) for 10 min. The adsorbed caffeine is then eluted with two 500-μL portions of chloroform.

QUESTIONS

5-16. You are presented a two-phase system. The two liquids are immiscible. The top phase is blue and the bottom, orange. One phase is water. Please devise an experiment to **definitively** differentiate which phase is water.

5-17. Which layer (upper or lower) will each of the following organic solvents usually form when being used to extract an aqueous solution?

toluene   methylene chloride   diethyl ether
hexane   acetone

5-18. Construct a flow chart to demonstrate how you could separate a mixture of 1,4-dichlorobenzene, 4-chlorobenzoic acid, and
4-chloroaniline using an extraction procedure.

5-19. A slightly polar organic compound partitions itself between ether and water phases. The K_p (partition coefficient) value is 2.5 in favor of the ether solvent. What simple procedure could you use to increase this K_p value?

5-20. You weight out exactly 1.00 mg of benzoic acid and dissolve it in a mixture of 2.0 mL of diethyl ether and 2.0 mL of water. After mixing and allowing the layers to separate, the ether layer is removed, dried, and concentrated to yield 0.68 mg of benzoic acid. What is the K_p value (ether/water) for this system?

5-21. If asked to separate an equal mixture of benzoic acid [pK_a = 4.2] and 2-naphthol [pK_a = 9.5] using a liquid–liquid extraction technique, explain why an aqueous solution of $NaHCO_3$[pK_a = 6.4] would be far more effective than the stronger aqueous solution of NaOH[pK_a = 15.7].

BIBLIOGRAPHY

For overviews on extraction methods see the following general references:

Dean, J. R. *Extraction Methods for Environmental Analysis;* Wiley: New York, 1998.
Handley, A. J. *Extraction Methods for Organic Analysis;* CRC Press LLC: Boca Raton, FL, 1999.
Kirk–Othmer Encyclopedia of Chemical Technology, 4th ed.; Wiley: New York, 1993; Vol. 10, p. 125.
Schneider, Frank L. *Qualitative Organic Microanalysis*; Vol. II of *Monographien aus dem Gebiete der qualitätiven Mikroanalysis*, A. Benedetti-Pichler, Ed.; Springer-Verlag: Vienna, 1964, p. 61.
Shugar, G. J. *Chemical Technicians' Ready Reference Handbook*, 3rd ed.; McGraw-Hill: New York, 1990.
Zubrick J. W. *The Organic Chem Lab Survival Manual: A Student's Guide to Techniques*, 7th ed.; Wiley: New York, 2008.
Zubrick, J. W. *The Organic Chem Lab Survival Manual*, 7th ed. Wiley: New York, 2008.

NOTE The following experiments use Technique 4: Experiments [4A], [4B], [5A], [5B], [7], [8A], [11A], [11B], [11C], [12], [13], [16], [17], [19A], [19B], [19C], [22A], [22B], [23], [27],[30], [32], [34A], [34B], [A1$_b$], [D3], [E3], [F1], [F2], [F3], and [F4].

www → *[3A$_{adv}$], [4$_{adv}$], and [6$_{adv}$].*

Crystallization

This discussion introduces the basic technique of purifying solid organic substances by crystallization. The technique of crystallizing an organic compound is fundamental; it must be mastered if you are going to purify solids. *It is not an easy art to acquire.* Organic solids tend not to crystallize as easily as inorganic substances.

Legend has it that an organic chemist resisted an invitation to leave a well-worn laboratory for new quarters because he suspected that the older laboratory (in which many crystallizations had been carried out) harbored seed crystals for a large variety of substances the chemist needed. Carried by dust from the earlier work, these traces of material presumably aided the successful initiation of crystallization of reluctant materials. Further support for this legend comes from the often quoted (but never substantiated) belief that after a material was first crystallized in a particular laboratory, subsequent crystallizations of the material, regardless of its purity or origin, were always easier.

In several areas of chemistry, the success or failure of an investigation can depend on the ability of a chemist to isolate tiny quantities of crystalline substances. Often the compounds of interest must be extracted from enormous amounts of extraneous material. In one of the more spectacular examples, Reed et al. isolated 30 mg of the crystalline coenzyme lipoic acid from 10 tons of beef liver residue.[12]

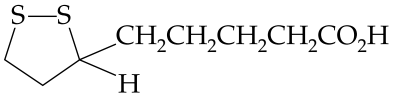

Lipoic acid

General Crystallization Procedure

The following steps are the essentials of crystallization:

Step 1. Select a suitable solvent.

Step 2. Dissolve the material to be purified in the minimum amount of warm solvent. *Remember that most organic solvents are extremely flammable and that many produce very toxic vapor.*

Step 3. Once the solid mixture has fully dissolved, filter the heated solution, and then bring it to the point of saturation by evaporating a portion of the solvent.

Step 4. Cool the warm saturated solution to reduce the solubility of the solute; this usually causes the solid material to precipitate. If the material has a low melting point or is very impure it may come out of solution sometimes as an oil. If so, reheat the solution and allow it to recool slowly.

Step 5. Isolate the solid by filtration, and then remove the last traces of solvent.

The crystallization is successful if the solid is recovered in good yield and is purer than it was before the crystallization. This cycle, from solid state to solution and back to solid state is called *recrystallization* when both the initial and final materials are crystalline.

---

[12] Reed, L. J.; Gunsalus, I. C.; Schnakenberg, G. H. F.; Soper, Q. F.; Boaz, H. E.; Kem, S. F.; Parke, T. V. *J. Am. Chem. Soc.* **1953,** *75,* 1267.

Although the technique sounds fairly simple, in reality it is demanding. Successful purification of microscale quantities of solids will require your utmost attention. Choosing a solvent system is critical to a successful crystallization. To achieve high recoveries, the compound to be crystallized should ideally be very soluble in the hot solvent, but nearly insoluble in the cold solvent. To increase the purity of the compound, the impurities should be either *very soluble* in the solvent at all temperatures or *not soluble* at any temperature. The solvent should have as low a boiling point as possible so that traces of solvent can be easily removed (evaporated) from the crystals after filtration. It is best to use a solvent that has a boiling point at least 10 °C lower than the melting point of the compound to be crystallized to prevent the solute from "oiling out" of solution. Thus, the choice of solvent is critical to a good crystallization. Table 5.6 lists common solvents used in the purification of organic solids.

When information about a suitable solvent is not available, the choice of solvent is made on the basis of solubility tests. Craig's rapid and efficient procedure for microscale solubility testing works nicely; it requires only milligrams of material and a nine-well, Pyrex spot plate.[13]

Place 1–2 mg of the solid in each well and pulverize each sample with a stirring rod. Add 3–4 drops of a given solvent to the first well and observe whether the material dissolves at ambient temperature. If not, stir the mixture for 1.5–2 min and observe and record the results. Repeat this process with the chosen set of solvents, using a separate well for each solubility test. *Keep track of which well contains which solvent.* Place your test plate (containing the samples) on a hot plate (set at its lowest setting) in the **hood;** add additional solvent if necessary. Record the solubility characteristics of the sample in each hot solvent. Cool the plate and see if crystallization occurs in any of the wells. On the basis of your observations, choose an appropriate solvent or a solvent pair (see the following paragraph) to recrystallize your material.

Solubility relationships are seldom ideal for crystallization; most often a compromise is made. If there is no suitable single solvent available, it is possible to use a mixture of two solvents, called a *solvent pair*. A solvent is chosen that will

Table 5.6 Common Solvents

Solvent	bp (°C)	Polarity
Acetone	56	Polar
Cyclohexane	81	Nonpolar
Diethyl ether	35	Intermediate polarity
Ethanol, 95%	78	Polar
Ethyl acetate	77	Intermediate polarity
Hexane	68	Nonpolar
Ligroin	60–90	Nonpolar
Methanol	65	Polar
Methylene chloride	40	Intermediate polarity
Methyl ethyl ketone	80	Intermediate polarity
Petroleum ether	30–60	Nonpolar
Toluene	111	Nonpolar
Water	100	Polar

---

[13] Craig, R. E. R. *J. Chem. Educ.* **1989,** *66,* 88.

readily dissolve the solid. Once the solid is dissolved in the minimum amount of hot solvent, the solution is filtered. A second solvent, miscible with the first, in which the solute has much lower solubility, is then added dropwise to the hot solution to achieve saturation. In general, polar organic molecules have higher solubilities in polar solvents, and nonpolar materials are more soluble in nonpolar solvents (like dissolves like). Table 5.7 lists some common solvent pairs.

It can take a long time to work out an appropriate solvent system for a particular reaction product. In most instances, with known compounds, the optimum solvent system has been established. Most crystallizations are not very efficient because many impurities have solubilities similar to those of the compounds of interest. Recoveries of 50–70% are not uncommon.

Several microscale crystallization techniques are available.

Table 5.7 Common Solvent Pairs

Solvent 1 (more polar)	Solvent 2 (less polar)
Acetone	Diethyl ether
Diethyl ether	Hexane
Ethanol	Acetone
Ethyl acetate	Cyclohexane
Methanol	Methylene chloride
Acetone	Water
Water	Ethanol
Toluene	Ligroin

Simple Crystallization

Simple crystallization works well with large quantities of material (100 mg and up), and it is essentially identical to that of the macroscale technique.

Step 1. Place the solid in a small Erlenmeyer flask or test tube. A beaker is not recommended because the rapid and dangerous loss of flammable vapors of hot solvent occurs much more easily from the wide mouth of a beaker than from an Erlenmeyer flask. Furthermore, solid precipitate can rapidly collect on the walls of the beaker as the solution becomes saturated because the atmosphere above the solution is less likely to be saturated with solvent vapor in a beaker than in an Erlenmeyer flask.

Step 2. Add a minimal amount of solvent and heat the mixture to the solvent's boiling point in a sand bath. Stir the mixture by twirling a spatula between the thumb and index finger. A magnetic stir bar may be used if a magnetic stirring hot plate is used.

Step 3. Continue stirring and heating while adding solvent dropwise until all of the material has dissolved.

Step 4. Add a decolorizing agent (powdered charcoal, ~2% by weight; or better, activated-carbon Norit pellets,[14] ~0.1% by weight), to remove colored minor impurities and other resinous byproducts.

Step 5. Filter (by gravity) the hot solution into a second Erlenmeyer flask (pre-heat the funnel with hot solvent). This removes the decolorizing agent and any insoluble material initially present in the sample.

Step 6. Evaporate enough solvent to reach saturation.

Step 7. Cool to allow crystallization (crystal formation will be better if this step takes place slowly). After the system reaches room temperature, cooling it in an ice bath may improve the yield.

Step 8. Collect the crystals by filtration on a Büchner or Hirsch funnel. Save the mother liquor (this is the term used to describe the solution that was separated from the original crystals) until the identity of the product has been established. In some cases, it is possible to recover more product by concentrating and further cooling the mother liquor. The second crop of crystals, however, is usually not as pure as the first.

Step 9. Wash (rinse) the crystals carefully.

Step 10. Dry the crystals.

---

[14] Available from Sigma-Aldrich Chemical Co., 940 West St. Paul Ave., Milwaukee, WI 53233.

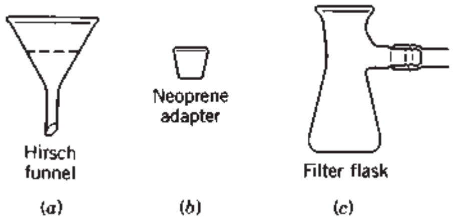

Figure 5.23 Component parts for vacuum filtration.

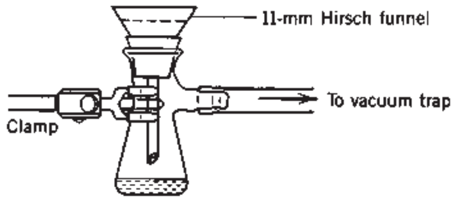

Figure 5.24 Vacuum filtration apparatus.

Filtration Techniques

Use of the Hirsch Funnel. The standard filtration system for collecting products purified by recrystallization in the microscale laboratory is vacuum (suction) filtration with an 11-mm Hirsch funnel. Many reaction products that do not require recrystallization can also be collected directly by vacuum filtration. The Hirsch funnel (Fig.5.23*a*) is composed of a ceramic cone with a circular flat bed perforated with small holes. The diameter of the bed is covered by a flat piece of filter paper of the same diameter. The funnel is sealed into a filter flask with a Neoprene adapter (Fig. 5.23*b*). Plastic and glass varieties of this funnel that have a polyethylene or glass frit are now available. It is still advisable to use the filter paper disk with these funnels to prevent the frit from clogging or becoming discolored. Regardless of the type of filter used, *always wet the filter paper disk with the solvent being used in the crystallization and then apply the vacuum.* This ensures that the filter paper disk is firmly seated on the bed of the filter.

Filter flasks have thick walls, and a side arm to attach a vacuum hose, and are designed to operate under vacuum (see Fig. 5.23*c*). The side arm is connected with *heavy-walled* rubber vacuum tubing to a water aspirator (water pump). The water pump uses a very simple aspirator based on the Venturi effect. Water is forced through a constricted throat in the pump. (See the detailed discussion of the Venturi effect and water pumps in the section on reduced pressure [vacuum] distillations.) When water flows through the aspirator, the resulting partial vacuum sucks air down the vacuum tubing from the filter flask. *Always turn the water on full force.* With the rubber adapter in place, air is pulled through the filter paper, which is held flat on the bed of the Hirsch (or Büchner) funnel by the suction. The mother liquors are rapidly forced into the filter flask, where the pressure is lower, by atmospheric pressure. The crystals retained by the filter are dried by the stream of air passing through them (Fig. 5.24).

www →

When you use a water pump, *it is very important to have a safety trap mounted in the vacuum line leading from the filter flask.* Any drop in water pressure at the pump (easily created by other students on the same water line turning on other aspirators at the same time) can result in a backup of water into the flask. As the flow through the aspirator decreases, the pressure at that point rapidly increases and water is forced up the vacuum tubing toward the filter flask (Fig. 5.24). It is also important to vent the vacuum by opening the vent stopcock or disconnecting the rubber tubing from the filter flask, *before* the water is turned off (see Fig. 5.25).

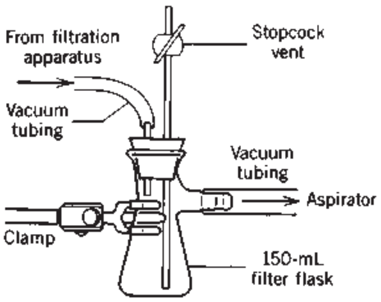

Figure 5.25 Vacuum trap.

In some cases, the precipitate collected on the Hirsch funnel is not highly crystalline. The filter cake may be too thick or pasty to dry simply by pulling air through it. A thin, flexible rubber sheet or a piece of plastic food wrap placed over the mouth of the funnel, such that the suction generated from the vacuum pulls the sheeting down onto the filter cake (collected crystals), will place atmospheric pressure on the solid cake. This pressure can force much of the remaining solvent from the collected material, and thus further dry it. Use a piece of sheeting large enough to cover the entire filter cake. Otherwise, a vacuum may not be created and adequate drying may not occur.

In some instances, substances may retain water or other solvents with great tenacity. To dry these materials, a *desiccator* is often used. This is generally a glass or plastic container containing a *desiccant* (a material capable of absorbing water). The substance to be dried, held in a suitable container, is then placed on a support above the desiccant. This technique is often used in quantitative analysis to dry collected precipitates. Vacuum desiccators are available (Fig. 5.26). If this method of drying is still insufficient, a *drying pistol* can be used (Fig. 5.27). The sample, in an open container (vial) is placed in the apparatus, which is then evacuated. The pistol has a pocket in which a strong adsorbing agent, such as P_4O_{10} (for water), NaOH or KOH (for acidic gases), or paraffin wax (for organic solvents), is placed. The pistol is heated by refluxing vapors that surround the barrel. A simple alternative to this method is the use of a side-armed test tube (Fig. 5.28).

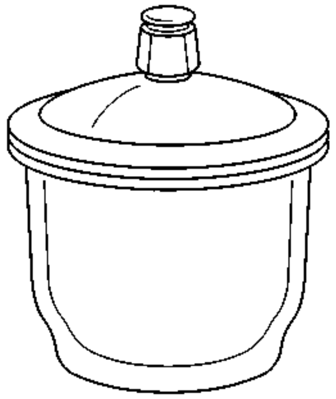

Figure 5.26 Vacuum desiccator.

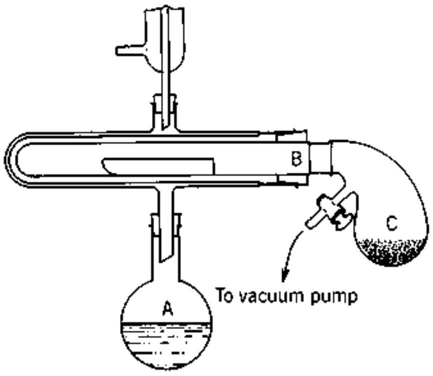

Figure 5.27 Abderhalden vacuum drying apparatus. A, refluxing heating liquid; B, vacuum drying chamber; C, desiccant.

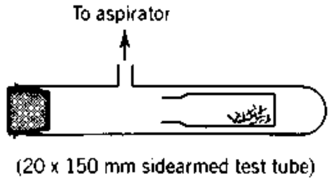

Figure 5.28 Side-arm test tube as a vacuum dryhing apparatus.

A Hirsch-Funnel Alternative—A Nail-Filter Funnel. A nail-filter funnel is a low-cost substitute for a Hirsch funnel (Fig. 5.29).[15] This apparatus is easily assembled from common laboratory glassware. Obtain a soft-glass rod that fits in the stem of a small glass funnel. Cut the rod to a suitable size and heat the tip of one end over a burner flame. When the tip becomes soft, hold the rod vertically and press the hot tip against a cold metal surface to flatten it to form a flat nail-like head. *The nail head should not be perfectly round or it will block the flow of liquid.* Cut the cooled rod to a suitable length and place the "nail" inside the stem of the funnel so that the flattened head of the nail rests on the top opening of the funnel. Cut a piece of filter paper just slightly larger than the nail head, place it on the nail head, and then place the funnel in a filter flask with a neoprene adapter. Be sure to wet the filter paper before filtering.

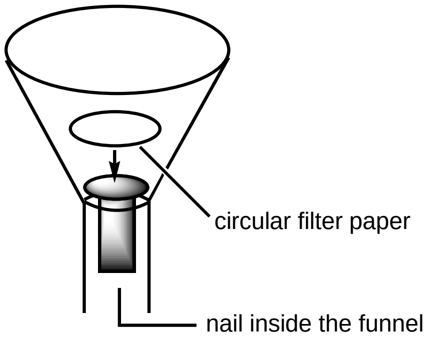

Figure 5.29 Nail-filter funnel.

Craig Tube Crystallizations. The Craig tube[16] is commonly used for microscale crystallizations in the range of 10–100 mg of material (Fig. 5.30). The process consists of the following steps.

Step 1. Place the sample in a small test tube (10 × 75 mm)

Step 2. Add the solvent (0.5–2 mL), and dissolve the sample by heating in the sand bath; add drops of solvent as needed. Rapid stirring with a microspatula (roll the spatula rod between your thumb and index finger) helps dissolve the material and protects against boilover. Add several drops of solvent by Pasteur pipet after the sample has completely dissolved. It will be easy to remove this excess at a later stage, since the volumes involved are very small. The additional solvent ensures that the solute will stay in solution during the hot transfer. Norit charcoal pellets may be added at this stage, if needed to remove colored impurities.

---

[15] Singh, M. M.; Pike, R. M.; Szafran, Z. *Microscale & Selected Macroscale Experiments for General & Advanced General Chemistry;* Wiley: New York, NY, 1995, pp. 47, 63; Claret, P. A. *Small Scale Organic Preparations;* Pitman: London, 1961, p. 15.

[16] Craig, L. C.; Post, O. W. *Ind. Eng. Chem., Anal. Ed.* **1944,** *16,* 413.

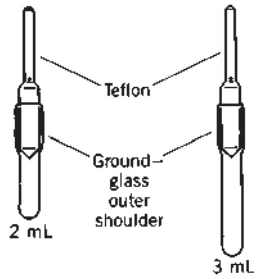

Figure 5.30 Craig tubes.

Step 3. Transfer the heated solution to the Craig tube by Pasteur filter pipet (see Fig. 3.30) that has been preheated with hot solvent. This transfer automatically filters the solution. A second filtration is often necessary if powdered charcoal has been used to decolorize the solution.

Step 4. The hot, filtered solution is then concentrated to saturation by gentle boiling in the sand bath. Constant agitation of the solution with a microspatula during this short period can avoid the use of a boiling stone and prevent boilover. Ready crystallization on the microspatula just above the solvent surface is a good indication that saturation is close at hand.

Step 5. The upper section of the Craig tube (the "head" or stopper) is set in place and the saturated solution is allowed to cool in a safe place. As cooling commences, seed crystals, if necessary, may be added by crushing them against the side of the Craig tube with a microspatula just above the solvent line. A good routine, if time is available, is to place the assembly in a small Erlenmeyer, then place the Erlenmeyer in a beaker, and finally cover the first beaker with a second inverted beaker. This procedure will ensure slow cooling, which will enhance good crystal growth (Fig. 5.31). A Dewar flask may be used when very slow cooling and large crystal growth are required.

Step 6. After the system reaches room temperature, cooling in an ice bath may improve the yield.

Step 7. Remove the solvent by inverting the Craig tube assembly into a centrifuge tube and spinning the mother liquors away from the crystals (Fig. 5.32). *This operation should be carried out with care.* First, fit the head with a thin copper wire (Fig. 5.32), held in place by a loop at the end of the wire that is placed around the narrow part of the neck. Some Teflon heads have a hole in the neck to anchor the wire. The copper wire should not be much longer than the centrifuge tube.

Step 8. Now insert the Craig tube into a centrifuge tube. To do this, hold the Craig tube upright (with the head portion up) and place the centrifuge tube *down over* the Craig tube. Push the Craig tube up with your finger so that the head is against the inverted bottom of the centrifuge tube, and then invert the whole assembly (Fig 5.32).

Step 9. Place the assembly into a centrifuge tube, *balance the centrifuge,* and spin the mother liquors away from the crystals (Fig. 5.32). This

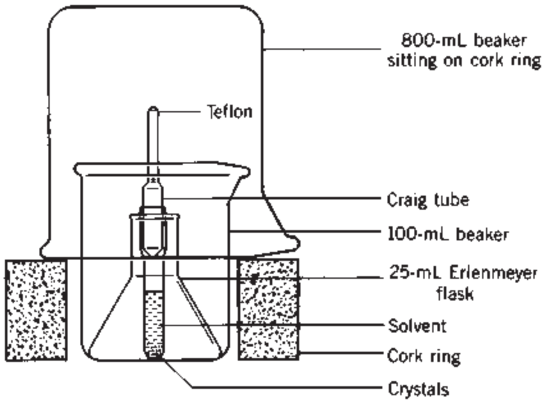

Figure 5.31 Apparatus for slow crystallization.

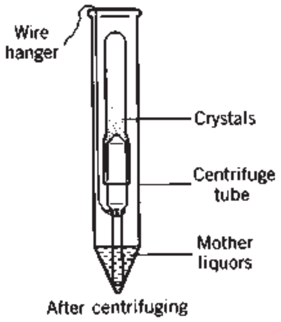

Figure 5.32 Crystal collection with a Craig tube.

replaces the usual filtration step in simple crystallizations. It avoids another transfer of material and also avoids product contact with filter paper.

Step 10. Remove the apparatus from the centrifuge, and then carefully remove the Craig tube from the centrifuge tube. Gently pull upward on the copper wire while at the same time applying downward pressure with your fingers to the bottom of the inverted Craig tube (this will keep the Craig tube assembly together and not let any of the crystalline product fall back into the centrifuge tube and into the mother liquors). Once the Craig tube assembly is removed from the centrifuge tube, turn it so that the neck of the tube is up, and then disassemble it. At this point scrape any crystalline product clinging to the head into the lower section. If the lower section is tared, it can be left to air dry to constant weight or placed in a warm vacuum oven (use a rubber band or thin wire to wrap a piece of filter paper over the open end to prevent dust from collecting on the product while drying).

The cardinal rule in carrying out the purification of small quantities of solids is *Keep the number of transfers to an absolute minimum!* The Craig tube is very helpful in this regard.

For other recrystallization and filtration methods, see reference material online; Chapter 5W, Crystallization.

www

QUESTIONS

5-22. What is the purpose of using activated carbon in a recrystallization procedure?

5-23. List several advantages and disadvantages of using a Craig tube for recrystallization.

5-24. Why is it advisable to use a stemless or a short-stemmed funnel when carrying out a gravity filtration?

5-25. Which of the following solvent pairs could be used in a recrystallization? Why or why not?
(a) Acetone and ethanol
(b) Hexane and water
(c) Hexane and diethyl ether

5-26. You perform a recrystallization on 60 mg of a solid material and isolate 45 mg of purified material. What is the percent recovery? Further concentration of the mother liquor provides an additional 8 mg of material. What is the total percent recovery?

5-27. Describe two techniques that can be used to induce crystallization.

5-28. When would you advise someone to use a solvent pair to carry out a recrystallization?

5-29. You are provided a solid which you suspect is not pure and the accompanying data sheet has very limited information. Two items which are recognizable are that an ethereal solvent has worked well when performing recrystallizations and the literature melting point is 61 °C. When looking at what ethereal solvents you have to choose from, you see two: *t*-butyl methyl ether and diethyl ether. Why would the latter be the far better choice knowing that aside from relative boiling points, the two ethers are very similar when considering their physical properties.

NOTE. The following experiments use Technique 5: Experiments [6], [7], [15], [16], [18], [19A], [19B], [19C], [19D], [20], [23A], [23B], [24A], [24B], [25A], [25B], [26], [28], [29A], [29B], [29C], [29D], [30], [31], [33A], [33B], [34A], [34B], [A1$_a$], [A2$_a$], [A3$_a$], [A1$_b$], [A2$_b$], [A3$_b$], [A4$_{ab}$], [B1], [C2], [D1], [D2], [E1], [E2], [F1], [F2], and [F4].
[2$_{adv}$], [3A$_{adv}$], [3B$_{adv}$], [5$_{adv}$], [6$_{adv}$], and [7$_{adv}$].

www

TECHNIQUE 6

Chromatography

Column, Flash, High-Performance Liquid, and Thin-Layer Chromatography

Technique 6A

The basic theory of chromatography is introduced in Technique 1 in the discussion of gas-phase separations. The word *chromatography* is derived from the Greek word for color, *chromatos*. Tswett discovered the technique in 1903 while studying ways to separate mixtures of natural plant pigments.[17] The chromatographic zones were detected simply by observing the visual absorption bands. Thus, as originally applied, the name was not an inconsistent use of terminology. Today, however, most mixtures are colorless. The separated zones in these cases are detected by other methods.

Two chromatographic techniques are discussed in this section. Both depend on adsorption and distribution between a stationary solid phase and a moving liquid phase. The first is column chromatography, which is used extensively throughout organic chemistry. It is one of the oldest of the modern chromatographic methods. The second technique, thin-layer chromatography (TLC), is particularly effective in rapid assays of sample purity. It can also be used as a preparative technique for obtaining tiny amounts of high-purity material for analysis.

Column Chromatography

Column chromatography, as its name implies, uses a column packed with a solid stationary phase. A mobile liquid phase flows by gravity (or applied pressure) through the column. Column chromatography uses polarity differences to separate materials. A sample on a chromatographic column is subjected to two opposing forces: (1) the solubility of the sample in the elution solvent system, and (2) the adsorption forces binding the sample to the solid phase. These interactions comprise an equilibrium. Some sample constituents are adsorbed more tightly; other components of the sample dissolve more readily in the liquid phase and are eluted more rapidly. The more rapidly eluting materials, thus, are carried further down the column before becoming readsorbed, and thus exit the column before more tightly bound components. The longer the column, the larger the number of adsorption–dissolution cycles (much like the vaporization–condensation cycles in a distillation column), and the greater the separation of sample components as they elute down the column. A molecule that is strongly adsorbed on the stationary phase will move slowly down the column; a molecule that is weakly adsorbed will move at a faster rate. Thus, a complex mixture can be resolved into separate bands of pure materials. These bands of purified material eventually elute from the column and can be collected.

Many materials have been used as the stationary phase in column chromatography. Finely ground alumina (aluminum oxide, Al_2O_3) and silicic acid (silica gel, SiO_2) are by far the most common adsorbents (stationary phases). Many common organic solvents are used as the liquids (sometimes called *eluents*) that act as the mobile phase and elute (wash) materials through the column. Table 5.8 lists the better known column packing and elution solvents.

---

[17] Tswett, M. *Ber. Deut. Botan. Ges.* **1906,** *24*, 235.

Table 5.8 **Column Chromatography Materials**

Stationary Phase		Moving Phase	
Alumina	↑ *Increasing adsorption of polar materials*	Water	↑ *Increasing solvation of polar materials*
Silicic acid		Methanol	
Magnesium sulfate		Ethanol	
Cellulose paper		Acetone	
		Ethyl acetate	
		Diethyl ether	
		Methylene chloride	
		Cyclohexane	
		Pentane	

Silica gel impregnated with silver nitrate (usually 5–10%) is also a useful adsorbent for some functional groups. The silver cation selectively binds to unsaturated sites via a silver-ion π complex. Traces of alkenes are easily removed from saturated reaction products by chromatography with this system (see Experiment [12]). This adsorbent, however, must be protected from light until used, or it will quickly darken and become ineffective.

Column chromatography is usually carried out according to the procedures discussed in the following five sections.

Packing the Column. The quantity of stationary phase required is determined by the sample size. A common rule of thumb is to use a weight of packing material 30– 100 times the weight of the sample to be separated. Columns are usually built with roughly a 10:1 ratio of height to diameter. In the microscale laboratory, two standard chromatographic columns are used:

1. A Pasteur pipet, modified by shortening the capillary tip, is used to separate smaller mixtures (10–100 mg). Approximately 0.5–2.0 g of packing is used in the pipet column (Fig. 5.33*a*).
2. A 50-mL titration buret (modified by shortening the column to 10 cm above the stopcock) is used for larger (50–200 mg) or difficult-to-separate sample mixtures. A buret column uses approximately 5–20 g of packing (Fig. 5.33*b*).

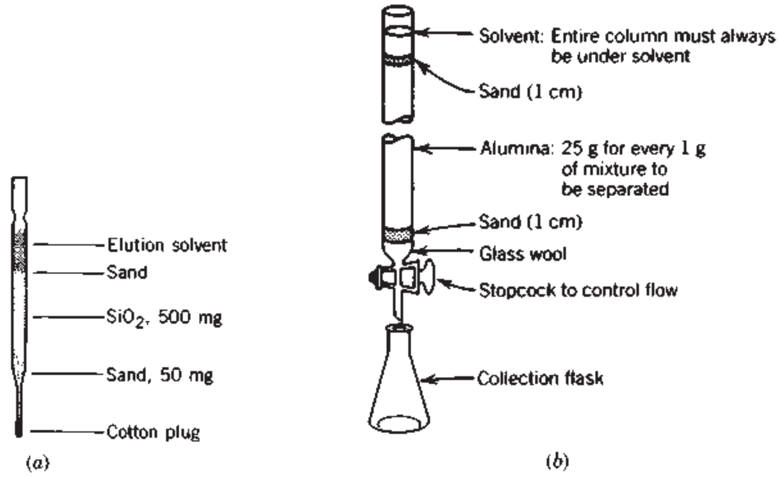
Elution solvent
Sand
SiO2, 500 mg
Sand, 50 mg
Cotton plug
(a)
Solvent: Entire column must always be under solvent
Sand (1 cm)
Alumina: 25 g for every 1 g of mixture to be separated
Sand (1 cm)
Glass wool
Stopcock to control flow
Collection flask
(b)

Figure 5.33 Chromatographic columns: (a) a Pasteur pipet column; (b) a buret column. (From Zubrick, James W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, 2008. Reprinted by permission of John Wiley & Sons, Inc., New York.)

Both columns are prepared by first clamping the empty column in a vertical position and then seating a small cotton or glass wool plug at the bottom. For a buret column, cover the cotton with a thin layer of sand. The Pasteur pipets are loaded by adding the dry adsorbent with gentle tapping, "dry packing." The pipet column (*dry column*) is then premoistened just prior to use.

The burets (*wet columns*) are packed by a slurry technique. In this procedure the column is filled part way with solvent; then the stopcock is opened slightly, and as the solvent slowly drains from the column a slurry of the adsorbent–solvent is poured into the top of the column. The column should be gently tapped while the slurry is added. The solvent is then drained to the top of the adsorbent level and held at that level until used. Alternatively, the wet-packed column can be loaded by sedimentation techniques rather than using a slurry. One such routine is to initially fill the column with the least-polar solvent to be used in the intended chromatographic separation. Then the solid phase is slowly added with gentle tapping, which helps to avoid subsequent channeling. As the solid phase is added, the solvent is slowly drained from the buret at the same time. After the adsorbent has been fully loaded, the solvent level is then lowered to the top of the packing as in the slurry technique.

Sample Application. Using a Pasteur pipet, apply the sample in a minimum amount of solvent (usually the least polar solvent in which the material is readily soluble) to the top of the column. *Do not disturb the sand layer!* Rinse the pipet, and add the rinses to the column just as the sample solution drains to the top of the adsorbent layer.

Elution of the Column. The critical step in resolving the sample mixture is eluting the column. Once the sample has been applied to the top of the column, the elution begins (a small layer of sand can be added to the top of the buret column after addition of the first portion of elution solvent).

NOTE: Do not let the column run dry: This can cause channels to form in the column.

In a buret column, the flow is controlled by the stopcock. The flow rate should be set to allow time for equilibrium to be established between the two phases; this will depend on the nature and amount of the sample, the solvent, and how difficult the separation will be. The Pasteur pipet column is *free flowing* (the flow rate is controlled by the size of the capillary tip and its plug); once the sample is on the column, the chromatography will require constant attention.

If necessary, it is possible to ease this restriction somewhat by modifying the pipet. Place a Tygon connector (short sleeve) at the top of the pipet column. Once the sample is on the column, insert a second pipet into this connector with its tip just below the liquid level on the top of the column. Add additional solvent through the second pipet (use a bulb, if necessary, but remove it before the elution begins), which acts as a solvent reservoir. As the solvent level in the column pipet drops below the tip of the top pipet, air is admitted, and additional solvent is automatically delivered to the chromatographic column. Thus, the solvent head on the column is maintained at a constant volume. The top pipet need be filled only at necessary intervals; larger volumes of solvent can thus be added to this reservoir. This arrangement also prevents dislodging of the absorbent as new solvent is added.

NOTE It is exceedingly important that solvents do not come in contact with the Tygon sleeve holding the second pipet. These sleeves contain plasticizers that will readily dissolve and contaminate the sample.

The choice of solvent is dictated by a number of factors. A balance between the adsorption power of the stationary phase and the solvation power of the elution solvent governs the rate of travel of the material down the column. If the material travels rapidly down the column, then too few adsorption–elution cycles will occur and the materials will not separate. If the sample travels too slowly, diffusion broadening takes over and separation is degraded. Solvent choices and elution rates can strike a balance between these factors and maximize the separation. It can take considerable time to develop a solvent or mixture of solvents that produces a satisfactory separation of a particular mixture.

Fraction Collection. As the solvent elutes from the column, it is collected in a series of "fractions" by using small Erlenmeyer flasks or vials. Under ideal conditions, as the mixture of material travels down the column, it will separate into several individual bands (zones) of pure substances. By careful collection of the fractions, these bands can be separated as they sequentially elute from the column (similar to the collection of GC fractions in the example described in Technique 1). The bands of eluted material can be detected by a number of techniques (weighing fraction residues, colored materials, TLC, etc.).

Column chromatography is a powerful technique for the purification of organic materials. In general, it is significantly more efficient than crystallization procedures. Recrystallization is often best avoided until the last stages of purification, where it will be most efficient. Rely instead on chromatography to do most of the separation.

Column chromatography of a few milligrams of product usually takes no more than 30 min, but chromatographing 10 g of product might easily take several hours, or even all day. Large-scale column chromatography (50–100 g) of a complex mixture could take several days to complete using this type of equipment.

Flash Chromatography

Flash chromatography, first described by Still and co-workers in 1978, is a common method for separating and purifying nonvolatile mixtures of organic compounds.[18] The technique is rapid, easy to perform, relatively inexpensive, and gives good separations. Many laboratories routinely use flash chromatography to separate mixtures ranging from 10 mg to 10 g.

This moderate-resolution, preparative chromatography technique was originally developed using silica gel (40–63 μm). Bonded-phase silica gel of a larger particle size can be used for reversed-phase flash chromatography. Flash chromatography columns are generally packed dry to a height of approximately 6 in. Thin-layer chromatography is a quick way to choose solvents for flash chromatography. A solvent that gives differential retardation factor (DR_f) values (of the two substances requiring seperation) ≥ 0.15 on TLC usually gives effective separation with flash chromatography. Table 5.9 lists typical experimental parameters for various sample sizes, as a guide to separations using flash chromatography. In general, a mixture of organic compounds separable by TLC can be separated preparatively using flash chromatography.

---

[18] Still, W. C.; Kahn, M.; Mitra, A. *J. Org. Chem.* **1978,** *43,* 2923.

Table 5.9 Typical Experimental Parameters

Column Diameter (mm)	Total Volume of Eluent (mL)[a]	Typical Sample Loading (mg) $DR_{f > 0.2}$	Typical Sample Loading (mg) $DR_{f > 0.1}$	Typical Fraction (mL)
10	100–150	100	40	5
20	200–250	400	160	10
30	400–450	900	360	20
40	500–550	1600	600	30
50	1000–1200	2500	1000	50

Source. Data from Majors, R. E., and Enzweiler, T. *LC, GC* **1998,** *6*, 1046.
[a] Required for both packing and elution.

Flash chromatography apparatus generally consists of a glass column equipped to accept a positive pressure of compressed air or nitrogen applied to the top of the column. A typical commercially available arrangement is shown in Figure 5.34.[19]

Generally, a 20–25% solution of the sample in the elution solvent is recommended, as is a flow rate of about 2 in./min. The column must be conditioned, before the sample is applied, by flushing the column with the elution solvent to drive out air trapped in the stationary phase and to equilibrate the stationary phase and the solvent.

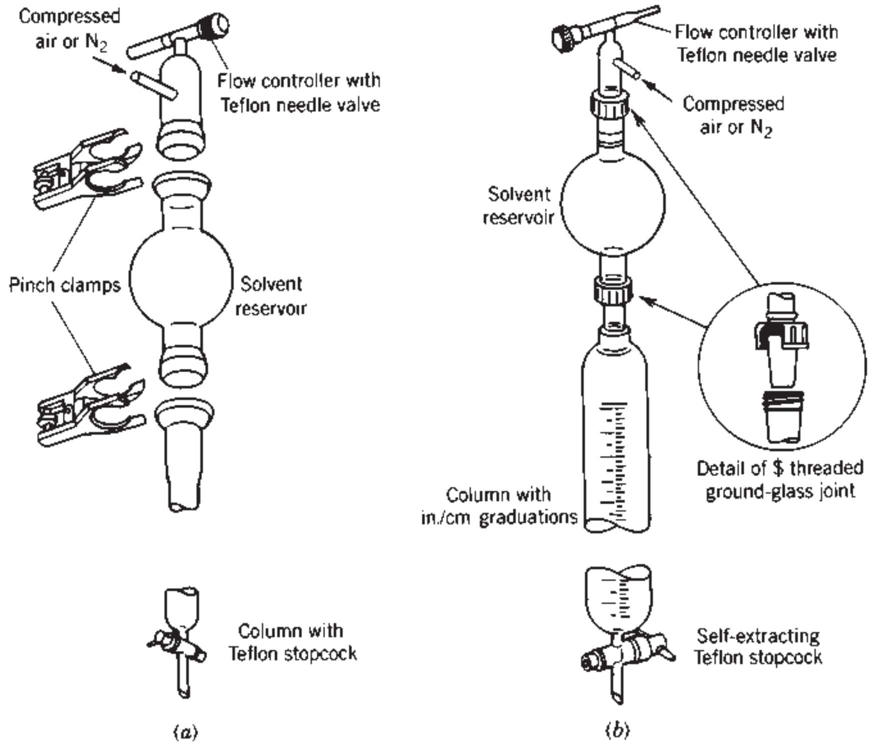
Compressed air or N2
Flow controller with Teflon needle valve
Pinch clamps
Solvent reservoir
Column with Teflon stopcock
(a)
Flow controller with Teflon needle valve
Compressed air or N2
Solvent reservoir
Detail of $ threaded ground-glass joint
Column with in./cm graduations
Self-extracting Teflon stopcock
(b)

Figure 5.34 (a) Conventional column and (b) screw-threaded column.

---

[19] A complete line of glass columns, reservoirs, clamps, and packing materials for flash chromatography is offered by Sigma-Adrich Chemical Co., P.O. Box 355, Milwaukee, WI 53201. Silica gels for use in this technique are also available from Amicon, Danvers, MA; J. T. Baker, Phillipsburg, NJ; EM Science/Merck, Gibbstown, NJ.; ICN Biomedicals, Inc., Cleveland, OH.

Several modifications of the basic arrangement have been reported, especially in regard to the adaptation of the technique to the instructional laboratory. These involve inexpensive pressure control valves, use of an aquarium "vibrator" air pump, and adapting a balloon reservoir to supply the pressurized gas.

At the microscale level, a pipet bulb or pump on the pipet column can be used to supply pressure to the column. If a bulb is used, squeeze it to apply pressure, and *remove the bulb from the pipet before releasing it!* Otherwise, material may be sucked up into the bulb and most likely disturb the column. Reapply the bulb to re-create pressure. If a pump is used, do not back off the pressure once it has been applied.

An improved method, utilizing a capillary Pasteur pipet for introducing the sample onto the chromatographic column approximately doubles the effectiveness (theoretical plates) of the column.[20] Dry-column flash chromatography[21] has been adapted for use in the instructional laboratory.[22] The "column" consists of a dry bed of silica gel in a sintered glass funnel placed in a standard vacuum filtration flask; the solvent is eluted by suction. Small (16 × 150-mm) test tubes inserted into the flask below the stem of the funnel are used to collect the fractions. This technique has been used successfully to separate mixtures ranging from 150 to 1000 mg.

Thin-Layer Chromatography

Thin-layer chromatography (TLC) is closely related to column chromatography, in that the phases used in both techniques are essentially identical. Alumina and silica gel are typical stationary phases, and the usual solvents are the mobile phases. There are, however, some distinct differences between TLC and column chromatography. The mobile (liquid) phase *descends* in column chromatography; the mobile phase *ascends* in TLC. The column of stationary-phase material used in column chromatography is replaced by a thin layer (100 μm) of stationary phase spread over a flat surface. A piece of window glass, a microscope slide, or a sheet of plastic can be used as the support for the thin layer of stationary phase. It is possible to prepare your own glass plates, but plastic-backed thin-layer plates are only commercially available. Plastic-backed plates are particularly attractive because they can easily be cut with scissors into strips of any size. Typical strips measure about 1 × 3 in., but even smaller strips can be satisfactory.

Thin-layer chromatography has some distinct advantages: it needs little time (2– 5 min) and it needs *very* small quantities of material (2–20 μg). The chief disadvantage of this type of chromatography is that it is not very amenable to preparative scale work. Even when large surfaces and thicker layers are used, separations are most often restricted to a few milligrams of material.

NOTE. Do not touch the active surface of the plates with your fingers. Handle them only by the edge.

TLC is performed as follows:

Step 1. Draw a light pencil line parallel to the short side of the plate, 5–10 mm from the edge. Mark one or two points, evenly spaced, on the line.

---

[20] Pivnitsky, K. K. *Aldrichimica Acta* **1989,** *22,* 30.

[21] Harwood, L. M.; *Aldrichimica Acta* **1985,** *18,* 25; Sharp, J. T.; Gosney I.; Rowley A. G. *Practical Organic Chemistry;* Chapman & Hall: New York, 1989.

[22] Shusterman, A. J.; McDougal, P. G.; Glasfeld, A. *J. Chem. Educ.* **1997,** *74,* 1222.

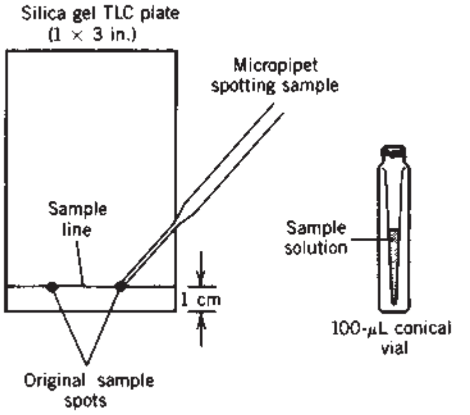
Silica gel TLC plate
(1 × 3 in.)
Micropipet spotting sample
Sample line
1 cm
Sample solution
100-μL conical vial
Original sample spots

Figure 5.35 Applying a sample to a TLC plate.

Place the sample to be analyzed (1 mg or less) in a 100-μL conical vial and add a few drops of a solvent to dissolve the sample. Use a capillary micropipet to apply a small fraction of the solution from the vial to the plate (Fig. 5.35). (These pipets are prepared by the same technique used for constructing the capillary insert for ultramicro boiling-point determinations, see Chapter 4.) Apply the sample to the adsorbent side of the TLC plate by gently touching the tip of the filled capillary to the plate. Remove the tip from the plate before the dot of solvent grows to much more than a few millimeters in diameter. If it turns out that you need to apply more sample, let the dot of solvent evaporate and then reapply more sample to exactly the same spot.

Step 2. Place the spotted thin-layer plate in a screw-capped, wide-mouth jar, or a beaker with a watch glass cover, containing a small amount of elution solvent (Fig. 5.36). It helps if one side of the jar's (beaker's) interior is covered with a piece of filter paper that wicks the solvent up to increase the surface area of the 100 solvent. The TLC plate must be positioned so that the spot of your sample is *above* the solvent. Cap the jar, or replace the watch glass on the beaker, to maintain an atmosphere saturated with the elution solvent. The elution solvent climbs the plate by capillary action, eluting the sample up the plate. *Do not move the developing chamber after the action has started.* Separation of mixtures into individual spots occurs by exactly the same mechanism as in column chromatography. Stop the elution by removing the plate from the jar or beaker when the solvent front nears the top of the TLC plate. Quickly (before the solvent evaporates) mark the position of the solvent front on the plate.

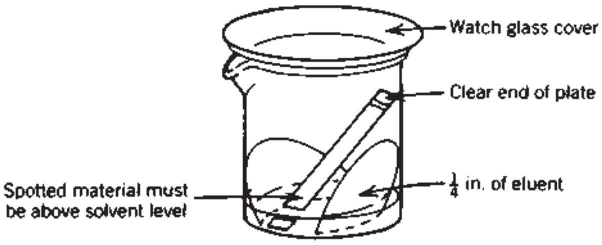
Watch glass cover
Clear end of plate
$\frac{1}{4}$ in. of eluent
Spotted material must be above solvent level

Figure 5.36 Developing a TLC plate. (From Zubrick, J. W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York. 2008. Reprinted by permission of John Wiley & Sons, Inc., New York.)

Step 3. Colorless, separated components of a mixture can often be observed in a developed TLC plate by placing the plate in an iodine-vapor chamber (a sealed jar containing solid I_2) for a minute or two. Iodine forms a reversible complex with most organic substances and dark spots will thus appear in those areas containing sample material. Mark the spots with a pencil soon after removing the TLC plate from the iodine chamber because the spots may fade. Samples that contain a UV-active chromophore (see Chapter 8) can be observed without using iodine. TLC plates are commonly prepared with an UV-activated fluorescent indicator mixed in with the silica gel. Sample spots can be detected with a hand-held UV lamp; the sample quenches the fluorescence induced by the lamp and appears as a dark spot against the fluorescent blue-green background.

Step 4. The TLC properties of a compound are reported as R_f values (retention factors). The R_f value is the distance traveled by the substance divided by the distance traveled by the solvent front (this is why the position of the solvent front should be quickly marked on the plate when the chromatogram is terminated; see Fig. 5.37). TLC R_f values vary with the moisture content of the adsorbent. Thus, the actual R_f of a compound in a given solvent can vary from day to day and from laboratory to laboratory. The best way to determine if two samples have identical R_f values is to elute them together on the same plate.

Thin-layer chromatography is used in a number of applications. The speed of the technique makes it quite useful for monitoring large-scale column chromatography. Analysis of fractions can guide decisions on the solvent–elution sequence. TLC analysis of column-derived fractions can also determine how best to combine collected fractions. Following the progress of a reaction by periodically removing small aliquots for TLC analysis is an extremely useful application of thin-layer chromatography.

Paper Chromatography

The use of cellulose-paper as an adsorbent is referred to as *paper chromatography.* This technique has many of the characteristics of TLC in that sheets or

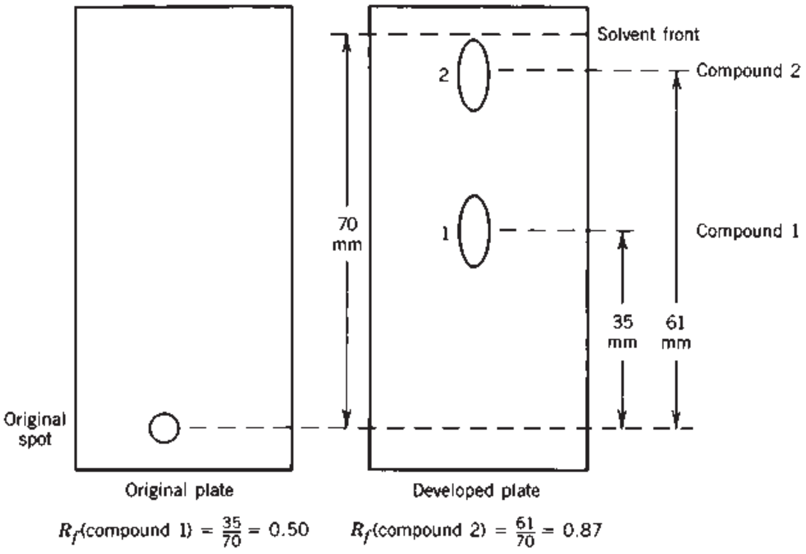

Figure 5.37 Determining R_f values.

strips of filter paper are used as the stationary phase. In this case, however, the paper is usually positioned to hang down from trays holding the paper and the elution solvent. The solvent front, therefore, descends downward rather than upward as in TLC. Paper chromatography has a distinct advantage: It is very amenable to the use of aqueous mobile phases and very small sample sizes. It is primarily used for the separation of highly polar or polyfunctional species such as sugars and amino acids. It has one major disadvantage: It is very slow. Paper chromatograms can easily take three to four hours or more to elute.

High-Performance Liquid Chromatography

Although gas chromatography is a powerful chromatographic method, it is limited to compounds that have a significant vapor pressure at temperatures up to about 200 °C. Thus, compounds of high molecular weight and/or high polarity cannot be separated by GC. High-performance liquid chromatography (HPLC) does not present this limitation.

GC and HPLC are somewhat similar, in the instrumental sense, in that the analyte is partitioned between a stationary phase and a mobile phase. Whereas the mobile phase in GC is a gas, the mobile phase in HPLC is a liquid. As shown schematically in Figure 5.38, the mobile phase (solvent) is delivered to the system by a pump capable of pressures up to about 6000 psi. The sample is introduced by the injection of a solution into an injection loop. The injection loop is brought in line between the pump and the column (stainless steel) by turning a valve; the sample then flows down the column, is partitioned, and flows out through a detector.

The solid phase in HPLC columns used for organic monomers is usually some form of silica gel. "Normal" HPLC refers to chromatography using a solid phase (usually silica gel) that is more polar than the liquid phase, or solvent, so that the less polar compounds elute more rapidly. Typical solvents include ethyl acetate, hexane, acetone, low molecular weight alcohols, chloroform, and acetonitrile. For extremely polar compounds, such as amino

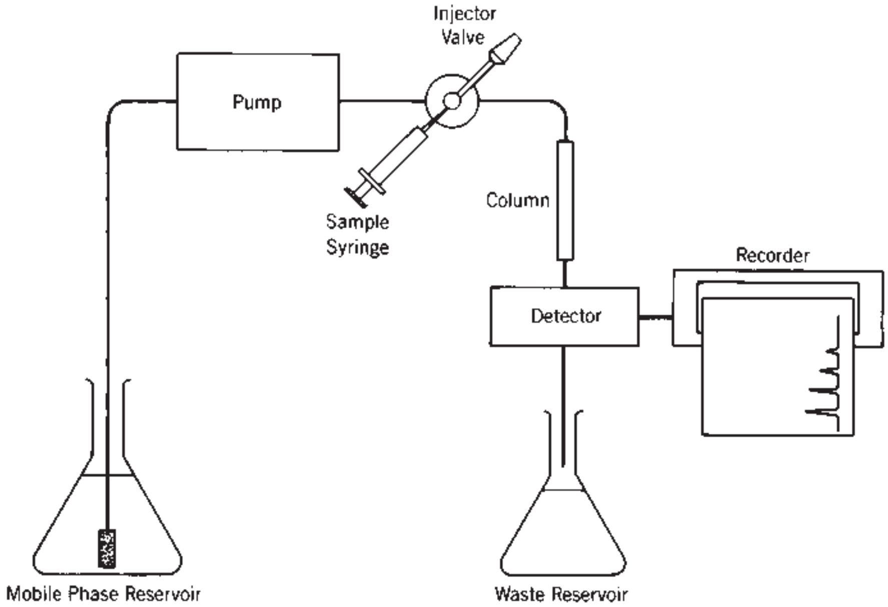

Figure 5.38 High-performance liquid chromatography system block diagram. *(Courtesy of the Perkin-Elmer Corp., Norwalk, CT.)*

acids, "reversed-phase" HPLC is used. Here, the liquid phase is more polar than the stationary phase, and the more polar compounds elute more rapidly. The mobile phase is usually a mixture of water and a water-miscible organic solvent such as acetonitrile, dioxane, methanol, isopropanol, or acetone. The stationary phase is usually a modified silica gel where the —OH groups of the silica gel have been replaced by —OSiR groups; R is typically a linear C_{18} alkyl chain. These so-called "bonded-phase" columns are not capable of handling as much analyte as normal silica gel columns, and are thus easily overloaded and are less useful for preparative work. (For further discussion see Solid-Phase Extraction, Technique 4, page 83.)

A wide variety of detection systems are available for HPLC. UV detection is common, inexpensive, and sensitive. The solvent flowing off the column is sent through a small cell where the UV absorbance is recorded over time. Many detectors are capable of variable wavelength operation so the detector can be set to the wavelength most suitable to the compound or compounds being analyzed. Photodiode array detectors are available; these can obtain a full UV spectrum in a fraction of a second, so that more information can be obtained on each component of a mixture. For compounds that absorb light in the visible (vis) spectrum, many detectors can be set to visible wavelengths. The principal shortcoming of UV-vis detection is that to be detected, compounds being studied must have a UV chromophore, such as an aromatic ring or other conjugated π system (see Chapter 8).

For compounds that lack a UV-vis chromophore, refractive index (RI) detection is a common substitute. An RI detector measures the difference in refractive index between the eluant and a reference cell filled with the elution solvent. Refractive index detection is significantly less sensitive than UV-vis detection, and the detector is quite sensitive to temperature changes during the chromatographic run.

More sophisticated HPLC instruments offer the ability to mix two or three different solvents and to use solvent gradients by changing the solvent composition as the chromatographic run progresses. This allows the simultaneous analysis of compounds that differ greatly in their polarity. For example, a silica gel column might begin elution with a very nonpolar solvent, such as hexane. The solvent polarity is then continuously increased by blending in more and more ethyl acetate until the elution solvent is pure ethyl acetate. This effect is directly analogous to temperature programming in GC.

For analytical work, typical HPLC columns are about 5 mm in diameter and about 25 cm in length. The maximum amount of analyte for such columns is generally less than 1 mg, and the minimum amount is determined by the detection system. High-performance liquid chromatography can thus be used to obtain small amounts of purified compounds for infrared (IR), nuclear magnetic resonance (NMR) or mass spec-trometric (MS) analysis. Larger "semipreparative" columns that can handle up to about 20 mg without significant overloading are useful for obtaining material for ^{13}C NMR spectroscopy or further synthetic work.

HPLC has the advantage that it is rapid, it uses relatively small amounts of solvent, and it can accomplish very difficult separations.

Concentration of Solutions

Technique 6B

The solvent can be removed from chromatographic fractions (extraction solutions, or solutions in general) by a number of different methods.

Distillation

Concentration of solvent by distillation is straightforward, and the standard routine is described in Technique 2 (page 61). This approach allows for high recovery of volatile solvents and often can be done outside a hood. The Hickman still head and the 5- or 10-mL round-bottom flask are useful for this purpose. Distillation should be used primarily for concentration of the chromatographic fraction, followed by transfer of the concentrate with a Pasteur filter pipet to a vial for final isolation.

Evaporation with Nitrogen Gas

A very convenient method for removal of final solvent traces is the concentration of the last 0.5 mL of a solution by evaporation with a gentle stream of nitrogen gas while the sample is warmed in a sand bath. This process is usually done at a hood station where several Pasteur pipets can be attached to a manifold leading to a source of dry nitrogen gas. Gas flow to the individual pipets is controlled by needle valves. *Always test the gas flow with a blank vial of solvent.*

Ruekberg described an alternative way to remove solvent from solutions of compounds that are not readily oxidized.[23] The setup includes an aquarium air pump, a pressure safety valve and ballast container, a drying tube, and a manifold. Blunted hypodermic needles are used in place of Pasteur pipets.

The sample vial will cool as the solvent evaporates, and gentle warming and agitation of the vial will thus help remove the last traces of the solvent. This avoids possible moisture condensation on the sample residue, as long as the gas itself is dry. *Do not leave the heated vial in the gas flow after the solvent is removed!* This precaution is particularly important in the isolation of liquids. Tare the vial before filling it with the solution to be concentrated; constant weight over time is the best indication that all solvent has been removed.

Removal of Solvent Under Reduced Pressure

Concentration of solvent under reduced pressure is very efficient. It reduces the time for solvent removal in microscale experiments to a few minutes. In contrast, distillation or evaporation procedures require many minutes for even relatively small volumes. Several methods are available.

Filter Flask Method. This vacuum-concentration technique can be tricky and should be practiced prior to committing hard-won reaction product to this test. The procedure is most useful with fairly large chromatographic fractions (5–10 mL). The sequence of operations is as follows (see also Fig. 5.39):

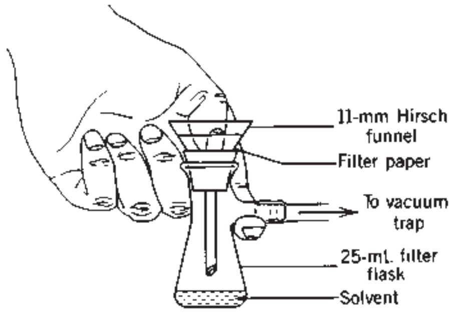

Figure 5.39 Removal of solvent under reduced pressure.

---

[23] Ruekberg, B. *J. Chem. Educ.* **1995,** *72,* A200.

Step 1. Transfer the chromatographic fraction to the 25-mL filter flask.

Step 2. Insert the 11-mm Hirsch funnel and rubber adapter into the flask.

Step 3. Turn on the water pump (with trap) and connect the vacuum tubing to the pressure flask side arm while holding the flask in one hand.

Step 4. Place the thumb of the hand holding the filter flask over the Hirsch funnel filter bed to shut off the air flow through the system (see Fig. 5.39). This will result in an immediate drop in pressure. The volatile solvent will rapidly come to a boil at room temperature. Thumb pressure adjusts air leakage through the Hirsch funnel and thereby controls the pressure in the system. It is also good practice to learn to manipulate the pressure so that the liquid does not foam up into the side arm of the filter flask.

The filter flask must be warmed by the sand bath during this operation; rapid evaporation of the solvent will quickly cool the solution. The air leak used to control the pressure results in a stream of moist laboratory air being rapidly drawn over the surface of the solution. If the evaporating liquid becomes cold, water will condense over the interior of the filter flask and contaminate the isolated residue. Warming the flask while evaporating the solvent will avoid this problem and help speed solvent removal. The temperature of the flask should be checked from time to time by touching it with the palm of the free hand. The flask is kept slightly above room temperature by adjusting the heating and evaporation rates. It is best to practice this operation a few times with pure solvent (blanks) to see whether you can avoid boilovers and accumulating water residue in the flask.

Rotary Evaporator Method. In most research laboratories, the most efficient way to concentrate a solution under reduced pressure is to use a **rotary evaporator.** A commercial micro-rotary evaporator is shown in Figure 5.40.

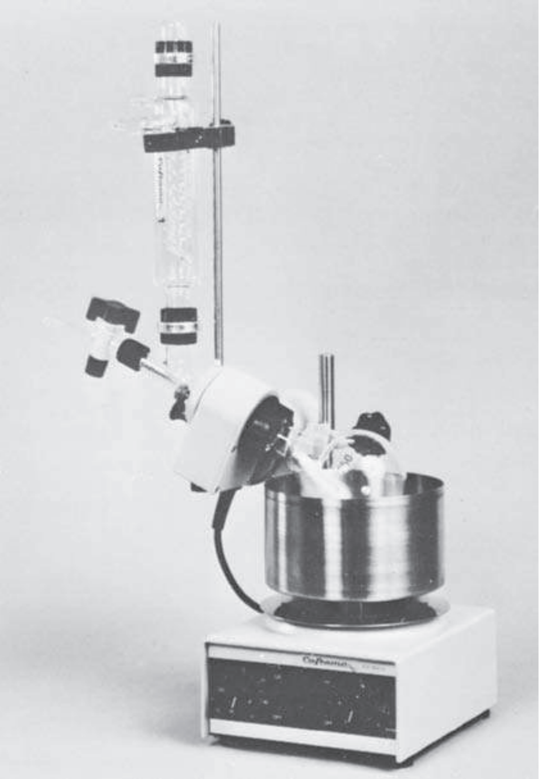

Figure 5.40 Heidolph micro-rotary evaporator. *(Courtesy of Caframo, Ltd., Wiarton, Ontario,Canada.)*

This equipment makes it possible to recover the solvent removed during the operation.

The rotary evaporator is a motor-driven device that rotates the flask containing the solution to be concentrated under reduced pressure. The rotation continuously exposes a thin film of the solution for evaporation. This process is very rapid, even well below the boiling point of the solvent being removed. Since the walls of the rotating flask are constantly rewetted by the solution, bumping and superheating are minimized. The rotating flask may be warmed in a water bath or other suitable device that controls the rate of evaporation. A suitable adapter (a "bump bulb") should be used on the rotary evaporator to guard against splashing and sudden boiling, which may lead to lost or contaminated products.

In microscale work, never pour a recovered liquid product from the rotary flask. Always use a Pasteur pipet.

Hickman Still–Rotary Evaporation Apparatus. A simple microscale rotary evaporator for use in the instructional laboratory consists of a 10-mL round-bottom flask connected to a capped Hickman still (side-arm type), which in turn is attached to a water aspirator (with trap).[24] The procedure involves transferring the solution to be concentrated to the preweighed 10-mL flask. The flask is then attached to a Hickman still with its top joint sealed with a rubber septum and threaded compression cap. The apparatus is connected by the still side arm to the trap–vacuum source with a vacuum hose. With the aspirator on, one shakes the apparatus while warming the flask in the palm of the hand. In this manner, bumping is avoided and evaporation is expedited. The still acts as a splash guard. Heat transfer is very effective, and once the flask reaches ambient temperature, the vacuum is released by venting through the trap stopcock.

QUESTIONS

5-30. When marking the sample line on a TLC plate, why is it inadvisable to use a ball-point pen?

5-31. A series of dyes is separated by TLC. The data are given below. Calculate the R_f value for each dye.

Material	Distance moved (cm)
Solvent	6.6
Bismarck brown	1.6
Lanacyl violet BF	3.8
Palisade yellow 3G	5.6
Alizarine emerald G	0.2

5-32. Why is it important not to let the level of the elution solvent in a packed chromatographic column drop below the top of the solid-phase adsorbent?

5-33. What are some advantages of using column chromatography to purify reaction products in the microscale laboratory?

5-34. Discuss the similarities and dissimilarities of TLC, paper, and column chromatography.

---

[24] Maynard, D. F. *J. Chem. Educ.* **1994,** *71,* A272.

5-35. Discuss the similarities and dissimilarities of HPLC and gas chromatography.

5-36. **(a)** What are the main advantages of using flash chromatography?
(b) How can TLC be used in connection with flash chromatography?

5-37. Using the information presented on the right, please identify and explain which spot has an R_f value of 0.5.

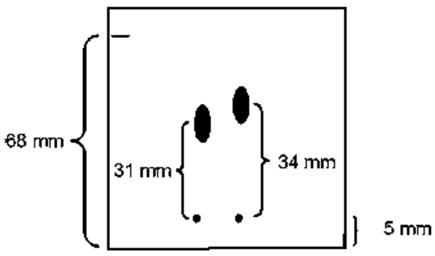

NOTE. The following experiments use Technique 6: Experiments [8A], [8B], [8C], [11C], [12], [13], [16], [17], [19A], [19B], [19C], [19D], [22A], [22B], [27], [29A], [29B], [29C], [29D], [30], [33A], [33B], [35], [A2$_a$], [A1$_b$], [E1], and [E3]. [1A$_{adv}$], [1B$_{adv}$], [4$_{adv}$], and [7$_{adv}$].

www

Collection or Control of Gaseous Products

TECHNIQUE 7

Water-Insoluble Gases

Numerous organic reactions lead to the formation of gaseous products. If the gas is insoluble in water, collection is easily accomplished by displacing water from a collection tube. A typical experimental setup for the collection of gases is shown in Figure 5.41.

As illustrated, the glass capillary efficiently transfers the evolved gas to the collection tube. The delivery system need not be glass; small polyethylene or polypropylene tubing may also serve this purpose. In this arrangement, a syringe needle is inserted through a septum to accommodate the plastic tubing as shown in Figure 5.42. An alternative to this connector is a shortened Pasteur pipet inserted through a thermometer adapter (Fig. 5.42). Another alternative to the syringe needle or glass pipet tip is suggested by Jacob.[25] The lower half of the tapered tip of a plastic automatic delivery pipet

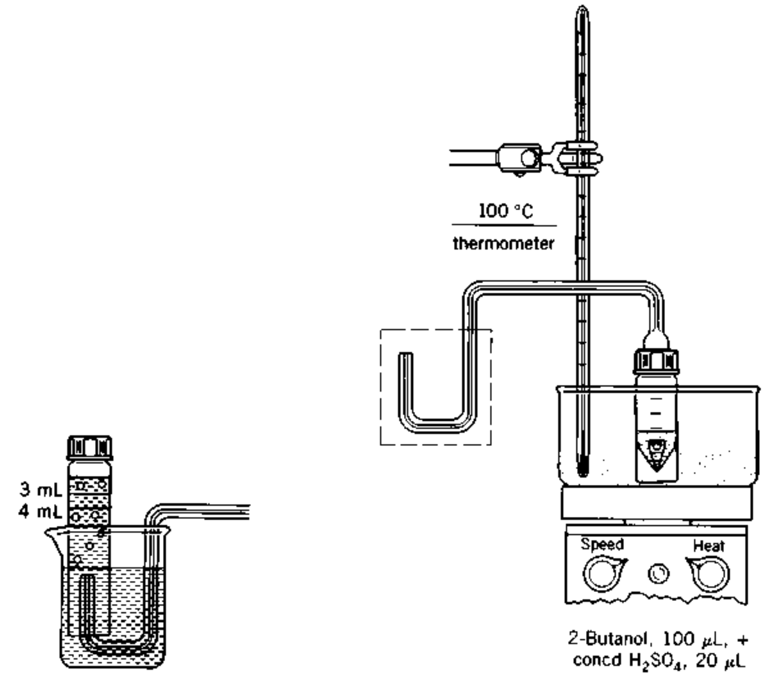

Figure 5.41 Microscale gas collection apparatus.

---

[25] Jacob, L. A. *J. Chem. Educ.* **1992,** *69,* A313.

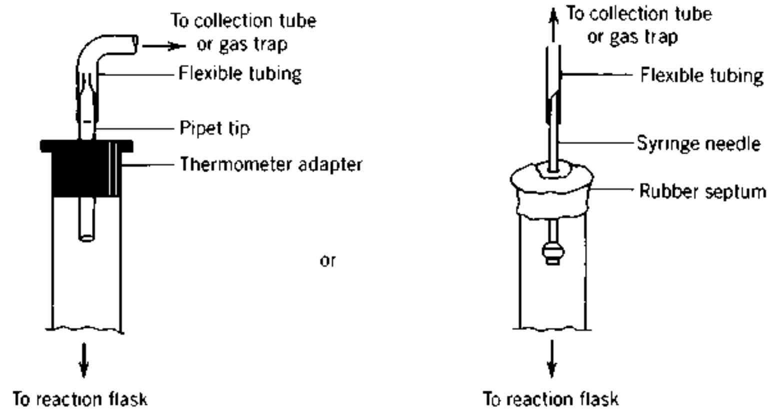

Figure 5.42 Alternative arrangements for controlled gas collection.

tip is cut off to prevent buildup of excess pressure in the reaction vessel. The pipet tip is then inserted through a previously pierced rubber septum or into a thermometer adapter. The narrow end of the tip is then inserted into the plastic tubing.

An example of a reaction leading to gaseous products that can use this collection technique is the acid-catalyzed dehydration of 2-butanol described in Experiment [9]. The products of this reaction are a mixture of alkenes: 1-butene, *trans*-2-butene, and *cis*-2-butene, which boil at −6.3, 0.9, and 3.7 °C respectively and *sec*-butyl ether (2,2′-oxybisbutane) which boils at 123 °C. While all four compounds are formed in the reaction mixture, the setup is designed to collect the gases and thus the three alkenes.

In Figure 5.41 the gas collection tube is capped with a rubber septum. This arrangement allows for convenient removal of the collected gaseous butenes using a gas-tight syringe, as shown in Figure 5.43. In this particular reaction, the mixture of gaseous products is conveniently analyzed at ambient temperature by GC (see Technique 1).

Trapping Byproduct Gases

Some organic reactions release poisonous or irritating gases as byproducts. For example, hydrogen chloride, ammonia, and sulfur dioxide are typical byproducts in organic reactions. In these cases, the reaction is generally run in a **hood.** A gas trap may be used to prevent the gases from being released into the laboratory atmosphere. If the evolved gas is water soluble, the trap technique works well at the microscale level. The evolved gas is directed from the reaction vessel to a container of water or other aqueous solution, wherein it dissolves (reacts). For example, a dilute solution of sodium or ammonium hydroxide is suitable for acidic gases (such as HCl); a dilute solution of sulfuric or hydrochloric acid is suitable for basic gases (such as NH_3 or low molecular weight amines). Various designs are available for gas traps. A simple, easily assembled one for a gas that is very soluble in water is shown in Figure 5.44. Note that the funnel is not immersed in the water. If the funnel is held below the surface of the water and a large quantity of gas is absorbed or dissolved, the water easily could be drawn back into the reaction assembly. If the gas to be collected is not very soluble, the funnel may be immersed just below (1–2 mm) the surface of the water.

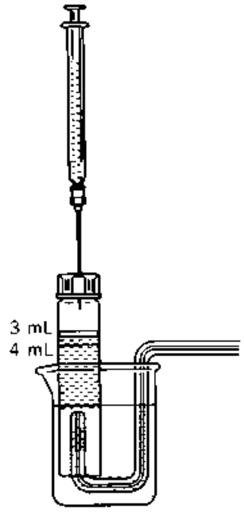

Figure 5.43 Removal of collected gases.

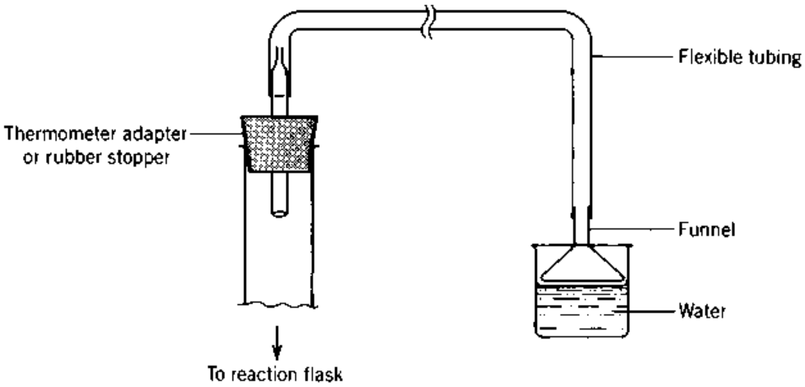

Figure 5.44 Trapping a water-soluble gas.

At the microscale level, small volumes of gases are evolved that may not require the funnel. Three alternatives are available:

a. Fill the beaker (100 mL) in Figure 5.41 with moistened fine glass wool and lead the gas delivery tube directly into the wool.

b. Place moistened glass wool in a drying tube and attach the tube to the reaction apparatus (see Chapter 3W, Fig. 3.10W). However, be careful not to let the added moisture drip into the reaction vessel; place a small section of **dry** glass wool in the tube before the moist section is added.

c. Use a water aspirator. An inverted funnel can be placed over the apparatus opening where the evolved gas is escaping (usually the top of a condenser) and connected with flexible tubing (through a water trap) to the aspirator. A second arrangement is to use a glass or plastic T-tube (open on one end) inserted in the top of the condenser, by use of a rubber stopper, in place of the funnel.[26] If the reaction must be run under anhydrous conditions, a drying tube is inserted between the condenser and T-tube. This arrangement is very efficient, easy to assemble, and inexpensive. The simplest method is to clamp a Pasteur pipet so that its tip is inserted well into the condenser, and connect it (through a water trap) to the aspirator.

QUESTIONS

5-38. In Figure 5.41 why is a septum, not just a plain cap, used on the top of the gas collection tube?

5-39. An evolved gas is directed from the reaction vessel to a container of water or other aqueous solution, wherein it dissolves (reacts). For example, a dilute solution of sodium or ammonium hydroxide is suitable for acidic gases (such as HCl). What solution would be appropriate to trap thiols and sulfides? (*Hint:* Consult a qualitative analysis text.)

5-40. One way to eliminate emissions is to place moistened glass wool in a drying tube, which is then attached to the reaction apparatus (Fig. 3.10W). What precautions must be taken when using this method?

5-41. In the collection of water-insoluble gases with the apparatus shown in Figure 5.41, describe how one might measure the rate at which a gas is evolved from a reaction mixture.

NOTE. The following experiments use Technique 7: Experiments [9], [10], [14], [A2$_a$], and [B2].

---

[26] Horodniak, J. W.; Indicator, N. *J. Chem. Educ.* **1970,** 47, 568.

TECHNIQUE 8

Measurement of Specific Rotation

Solutions of optically active substances, when placed in the path of a beam of polarized light, may rotate the plane of the polarized light. Enantiomers (two molecules that are nonidentical mirror images) have identical physical properties (melting points, boiling points, infrared and nuclear magnetic resonance spectra, etc.) except for their interaction with plane polarized light, their *optical activity*. Optical rotation data can provide important information concerning the absolute configuration and the enantiomeric purity of a sample.

Optical rotation is measured using a *polarimeter*. This technique is applicable to a wide range of analytical problems, from purity control to the analysis of natural and synthetic compounds. The results obtained from measuring the observed angle of rotation α are generally expressed as the *specific rotation* [α].

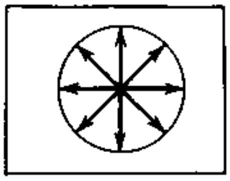

Figure 5.45 Oscillation of the electric field of ordinary light occurs in all possible planes perpendicular to the direction of propagation. (From Solomons, T. W. G. *Organic Chemistry,* 9th ed.; Wiley: New York, 2008. Reprinted by permission of John Wiley & Sons, Inc., New York.)

Theory

Ordinary light behaves as though it were composed of electromagnetic waves in which the oscillating electric field vectors are distributed among the infinite number of possible orientations around the direction of propagation (see Fig. 5.45).

NOTE. A beam of light behaves as though it is composed of two, mutually perpendicular, oscillating fields: an electric field and a magnetic field. The oscillating magnetic field is not considered in the following discussion.

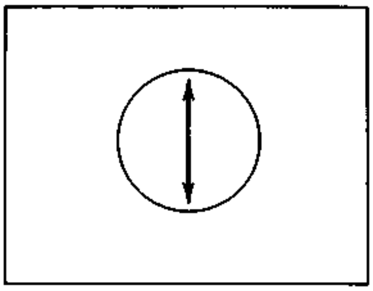

Figure 5.46 The plane of oscillation of the electric field of plane-polarized light. In this example the plane of polarization is vertical. (From Solomons, T. W. G. *Organic Chemistry,* 9th ed.; Wiley: New York, 2008. Reprinted by permission of John Wiley & Sons, Inc., New York.)

The planes in which the electrical fields oscillate are perpendicular to the direction of propagation of the light beam. If one separates one particular plane of oscillation from all other planes by passing the beam of light through a polarizer, the resulting radiation is plane-polarized (Fig. 5.46). In the interaction of light with matter, this plane-polarized radiation is represented as the vector sum of two circularly polarized waves. The electric vector of one of the waves moves in a clockwise direction; the other moves in a counterclockwise direction. Both waves have the same amplitude (Fig. 5.47). These two components add vectorially to produce plane-polarized light.

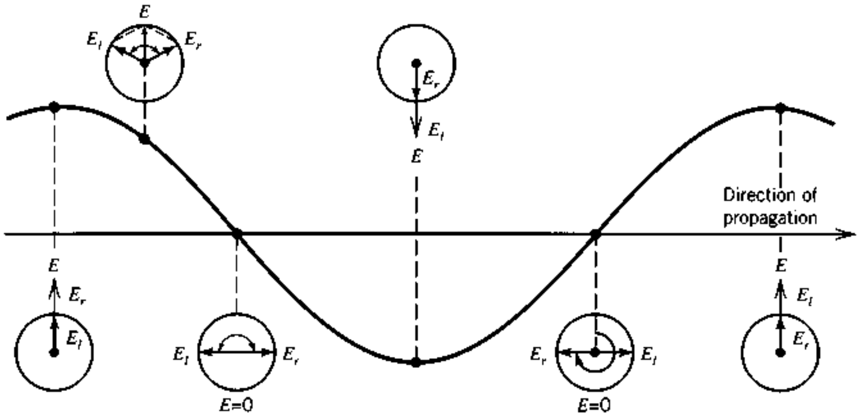

Figure 5.47 A beam of plane-polarized light viewed from the side (sine wave) and along the direction of propagation at specific times (circles) where the resultant vector *E* and the circularly polarized components E_l and E_r are shown. (From Douglas, B., McDaniel, D. H., and Alexander, J. J. *Concepts and Models of Inorganic Chemistry,* 3rd ed. Wiley, New York, 1994. Reprinted by permission of John Wiley & Sons, Inc., New York.)

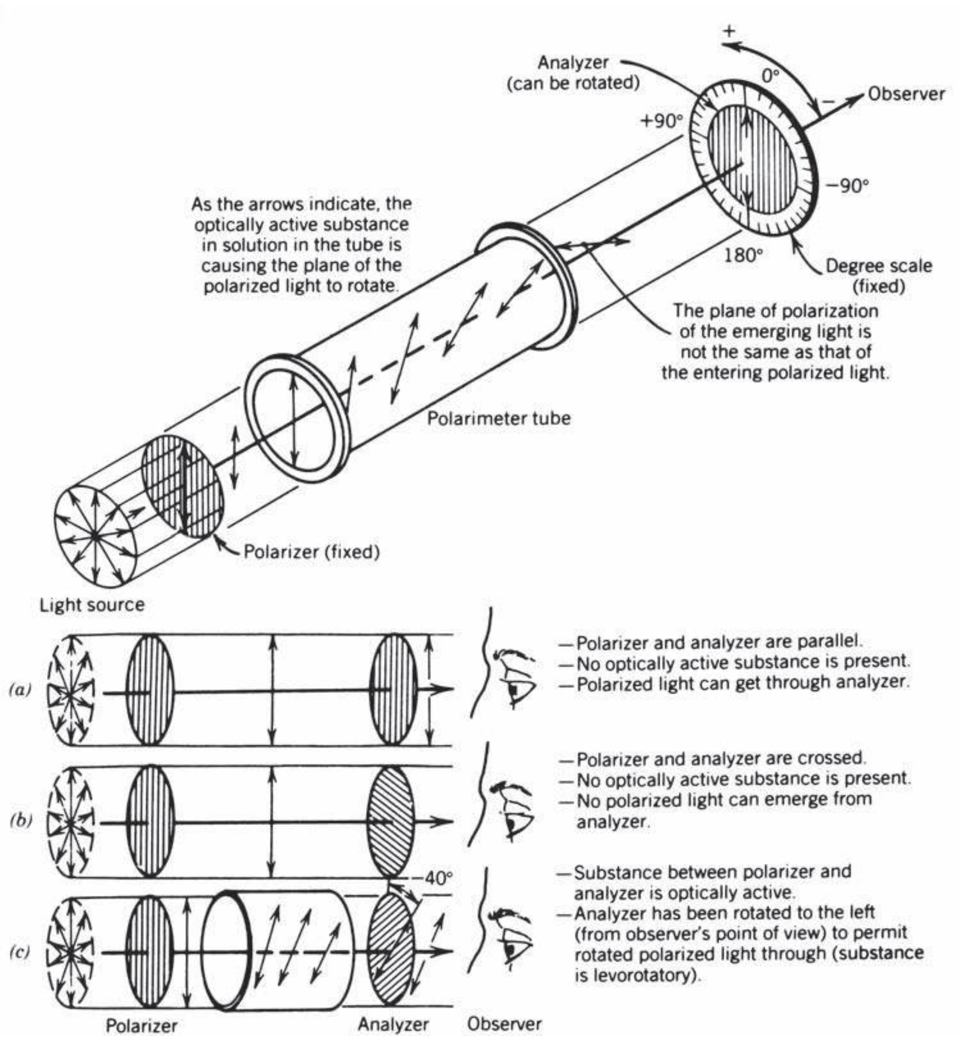
+
Analyzer
(can be rotated)
0°
+90°
−
Observer
−90°
180°
Degree scale
(fixed)
As the arrows indicate, the
optically active substance
in solution in the tube is
causing the plane of the
polarized light to rotate.
The plane of polarization
of the emerging light is
not the same as that of
the entering polarized light.
Polarimeter tube
Polarizer (fixed)
Light source
(a)
—Polarizer and analyzer are parallel.
—No optically active substance is present.
—Polarized light can get through analyzer.
(b)
—Polarizer and analyzer are crossed.
—No optically active substance is present.
—No polarized light can emerge from
analyzer.
−40°
(c)
—Substance between polarizer and
analyzer is optically active.
—Analyzer has been rotated to the left
(from observer's point of view) to permit
rotated polarized light through (substance
is levorotatory).
Polarizer
Analyzer
Observer

Figure 5.48 Operation of a polarimeter. (From Solomons, T. W. G. *Organic Chemistry,* 9th ed. Wiley, New York, 2008. Reprinted by permission of John Wiley & Sons, Inc. New York.)

If the passage of plane-polarized light through a material reduces the velocity of one of the circularly polarized components more than the other by interaction with bonding and nonbonding electrons, the transmitted beam of radiation has its plane of polarization rotated from its *original* position (Figs. 5.48 and 5.49). A **polarimeter** is used to measure this angle of rotation.

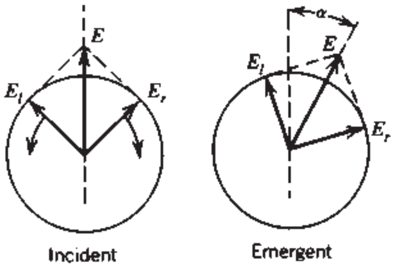
E
E_l
E_r
α
E
E_l
E_r
Incident
Emergent

Figure 5.49 Plane-polarized light before entering and after emerging from an optically active substance. (From Douglas, B., McDaniel, D. H., and Alexander, J. J. *Concepts and Models of Inorganic Chemistry,* 3rd ed. Wiley, New York, 1994. (Reprinted by permission of John Wiley & Sons, Inc., New York.)

The Polarimeter

The polarimeter measures the amount of rotation caused by an optically active compound (in solution) placed in the beam of the plane polarized light. The principal parts of the instrument are diagrammed in Figure 5.48. Two Nicol prisms are used in the instrument. The first prism, which polarizes the original light source, is called the polarizer. The second prism, called the analyzer, is used to examine the polarized light after it passes through the solution being analyzed.

When the axes of the analyzer and polarizer prisms are parallel (0°) and no optically active substance is present, the maximum amount of light is transmitted. If the axes of the analyzer and polarizer are at right angles to each other (90°), no transmission of light is observed. Placing an optically active solution into the path of the plane-polarized light causes one of the circularly polarized components to be slowed more than the other. The refractive indices are, therefore, different in the two circularly polarized beams. Figure 5.48 represents a case in which the left-hand component has been affected the most.

NOTE. In the simplified drawing, Figure 5.48, the effect on only one of the circularly polarized waves is diagrammed. See Figure 5.49 for a more accurate description (view from behind the figure).

This tilts the plane of polarization. The analyzer prism must be rotated to the left to maximize the transmission of light. If rotation is counterclockwise, the angle of rotation is defined as (−) and the enantiomer that caused the effect is called levorota-tory (*l*). Conversely, clockwise rotation is defined as (+), and the enantiomer is dextrorotatory (*d*). Tilting the plane of polarization is called *optical activity.* Note that if a solution of equal amounts of a *d* and an *l* enantiomeric pair is placed in the beam of the polarimeter, no rotation is observed. Such a solution is *racemic*; it is an equimolar mixture of enantiomers.

The magnitude of optical rotation depends on several factors: (1) the nature of the substance, (2) the path length through which the light passes, (3) the wavelength of light used as a source, (4) the temperature, (5) the concentration of the solution used to make the measurement of optical activity, and (6) the solvent used in making the measurement.

The results obtained from the measurement of the observed angle of rotation, α_{obs}, are generally expressed in terms of *specific rotation* [α]. The sign and magnitude of [α] are dependent on the specific molecule and are determined by complex features of molecular structure and conformation; they cannot be easily explained or predicted. The specific rotation is a physical constant characteristic of a substance. The relationship of [α] to α_{obs} is as follows:

$$[\alpha]_{\lambda}^{T} = \frac{\alpha_{obs}}{lc}$$

where

T = temperature of the sample in degrees Celsius (°C),
l = the length of the polarimeter cell in decimeters (1 dm = 0.1 m = 10 cm),
c = concentration of the sample in grams per milliliter (g/mL),
λ = the wavelength of light in nanometers (nm) used in the polarimeter.

These units are traditional, though most are esoteric by contemporary standards. The specific rotation for a given compound depends on both the concentration and the solvent, and thus both the solvent and concentration used must be specified. For example, $[\alpha]_{D}^{25}$ (c = 0.4, $CHCl_3$) = 12.3° implies that the measurement was recorded in a $CHCl_3$ solution of 0.4 g/mL at 25 °C using the sodium D line (589 nm) as the light source.

For increased sensitivity, many simple polarimeters have an optical device that divides the viewed field into three adjacent parts (triple-shadow polarimeter; Fig. 5.50). A very slight rotation of the analyzer will cause one portion to become dimmer and the other lighter (Fig. 5.50*a* and 5.50*c*). The angle of rotation reading (α) is recorded when the field sections all have the same intensity. An accuracy of ± 0.1° can be obtained.

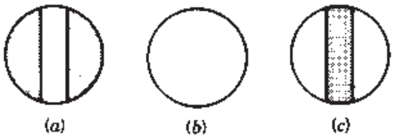

Figure 5.50 View through the eyepiece of the polarimeter. The analyzer should be set so that the intensity in all parts of the field is the same (b). When the analyzer is displaced to one side or the other, the field will appear as in (a) or (c).

Inaccurate Measurements. Several conditions may lead to inaccurate measurements, including trapped air bubbles in the cell, and solid particles suspended in the solution. Filter the solution, if necessary.

High-Performance Polarimeters and Optical Rotary Dispersion. For details of these two related topics refer to online discussion, Technique 8.

www

Applications to Structure Determination in Natural Products. Natural products provide interesting opportunities for measuring optical activity. An excellent example is the lichen metabolite, usnic acid, which can be easily isolated from its native source, "Old Man's Beard" lichens, as golden crystals (see Experiment [11A]).

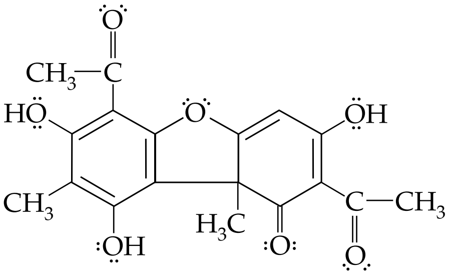

Usnic acid

Usnic acid contains a single stereocenter (stereogenic center, or chiral center) and, therefore, has the possibility of existing as an enantiomeric pair of stereoisomers. Generally, in a given lichen, only one of the stereoisomers (*R* or *S*) is present. Usnic acid has a very high specific rotation (~ ± 460°) which makes it an ideal candidate for optical rotation measurements at the microscale level.

QUESTIONS

5-42. A solution of 300 mg of optically active 2-butanol in 10 mL of water shows an optical rotation of $-0.54°$. What is the specific rotation of this molecule?

5-43. Draw the structure of usnic acid and locate its stereocenter.

5-44. After drawing all stereoisomers of 3-amino-2-butanol, identify the enantiomeric pairs.

5-45. If a solution of an equimolar mixture of an enantiomeric pair is placed in the beam path of the polarimeter, what would you observe?

5-46. The specific rotation of (+)Q is + 12.80°. At identical concentration, solvent, pathlength, and light wavelength, the observed rotation of a solution containing both enantiomers of Q is 6.40°. What are the relative concentrations of each enantiomer in the solution?

NOTE. The following experiment uses Technique 8: Experiment [11A]

Sublimation

TECHNIQUE 9

Sublimation is especially suitable for purifying solids at the microscale level. It is useful when the impurities present in the sample are nonvolatile under the conditions used. Sublimation is a relatively straightforward method; the impure solid need only be heated.

Sublimation has additional advantages: (1) It can be the technique of choice for purifying heat-sensitive materials—under high vacuum it can be effective at low temperatures; (2) solvents are not involved and, indeed, final traces of solvents are effectively removed; (3) impurities most likely to be separated are those with lower vapor pressures than the desired substance and often, therefore, lower solubilities, exactly those materials very likely to be contaminants in a recrystallization; (4) solvated materials tend to desolvate during the process; and (5) in the specific case of water of solvation, it is very effective even with substances that are deliquescent. The main disadvantage of the technique is that it can be less selective than recrystallization when the vapor pressure of the desired material being sublimed is similar to that of an impurity.

Some materials sublime at atmospheric pressure (CO_2, or dry ice, is a well-known example), but most sublime when heated below their melting points under reduced pressure. The lower the pressure, the lower the sublimation temperature. Substances that do not have strong intermolecular attractive forces are excellent candidates for purification by sublimation. Napthalene, ferrocene, and *p*-dichlorobenzene are examples of compounds that are readily sublimed.

Sublimation Theory

Sublimation and distillation are closely related. Crystals of solid substances that sublime, when placed in an evacuated container, will gradually generate molecules in the vapor phase by the process of evaporation (i.e., the solid has a vapor pressure). Occasionally, one of the vapor molecules will strike the crystal surface or the walls of the container and be held by attractive forces. This process, condensation, is the reverse of evaporation.

Sublimation is the complete process of evaporation from the solid phase to condensation from the gas phase to directly form crystals without passing through the liquid phase.

A typical single-component phase diagram is shown in Figure 5.51, which relates the solid, liquid, and vapor phases of a substance to temperature and pressure. Where two of the areas (solid, liquid, or vapor) touch, there is a line, and along each line the two phases exist in *equilibrium*. Line *BO* is the sublimation–vapor pressure curve of the substance in question; only along line *BO* can solid and vapor exist together in equilibrium. At temperatures and pressures along the *BO* curve, the liquid state is thermodynamically unstable. Where the three lines representing pairs of phases intersect, all three phases exist together in equilibrium. This point is called the *triple point*.

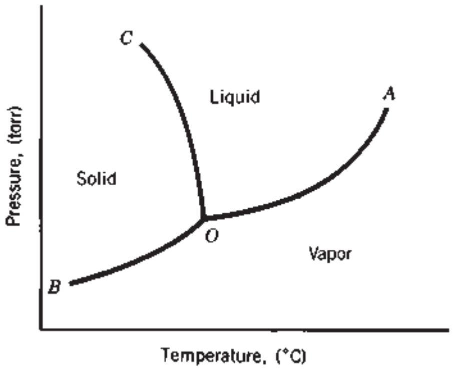

Figure 5.51 Single-component phase diagram.

Many solid substances have a sufficiently high vapor pressure near their melting point that allows them to be sublimed easily under reduced pressure in the laboratory. Sublimation occurs when the vapor pressure of the solid equals the pressure of the sample's environment.

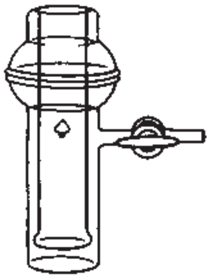

Figure 5.52 Vacuum sublimator. *(Courtesy of ACE Glass Inc., Vineland, NJ.)*

Experimental Setup

Heating the sample with a microburner or a sand bath to just below the melting point of the solid causes sublimation to occur. Vapors condense on the cold-finger surface, whereas any less volatile residue will remain at the bottom of the flask. Apparatus for sublimation of small quantities are now commercially available (Fig. 5.52). Two examples of simple, inexpensive apparatus suitable for sublimation of small quantities of material in the microscale organic laboratory are shown in Figure 5.53.

An example of the purification of a natural product, where the sublimation technique at the microscale level is effective, is the case of the alkaloid caffeine. This substance can be isolated by extraction from tea (see Experiment [11B]).

Precautions

Several precautions should be observed when performing a sublimation:

1. If you use the first setup in Figure 5.53, make sure you attach the hose connections to the cold finger in the proper manner. *The incoming cold water line is attached to the center tube.*
2. If you generate a vacuum using a water aspirator, make sure you place a water trap in the line. *Apply the vacuum to the system before you turn on the cooling water to the condenser.* This will keep moisture in the air in the flask from being condensed on the cold finger. Let the cold finger warm up before releasing the vacuum.
3. After the sublimation is complete, release the vacuum *slowly* so as not to disturb the sublimed material.
4. When using either of the arrangements in Figure 5.53, be careful to avoid loss of purified product as you remove the cold finger from the assembly.
5. The distance between the tip of the cold finger and the bottom of the sublimator should be less than 1 cm in most cases.

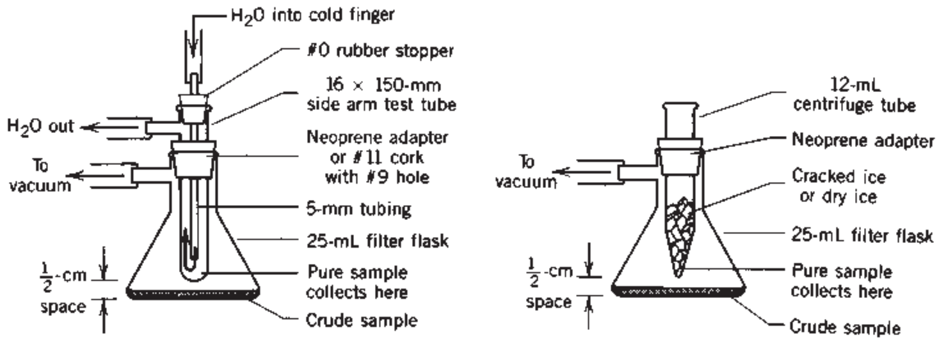
H$_2$O into cold finger
#0 rubber stopper
16 × 150-mm side arm test tube
H$_2$O out
To vacuum
Neoprene adapter or #11 cork with #9 hole
5-mm tubing
25-mL filter flask
$\frac{1}{2}$-cm space
Pure sample collects here
Crude sample
12-mL centrifuge tube
Neoprene adapter
To vacuum
Cracked ice or dry ice
25-mL filter flask
$\frac{1}{2}$-cm space
Pure sample collects here
Crude sample

Figure 5.53 Various sublimation apparatus.

QUESTIONS

5-47. List the advantages and disadvantages of sublimation as a purification technique.

5-48. For a solid compound to evaporate at atmospheric pressure it must have an unusually high vapor pressure. What molecular structural features contribute to this vapor pressure?

5-49. Why apply the vacuum to the sublimation system *before* you turn on the cooling water to the water condenser?

5-50. Why place a water trap in the vacuum line when using an aspirator to obtain the vacuum?

5-51. Why is sublimation particularly useful for purifying deliquescent compounds?

5-52. The 72% recovery after performing a sublimation translates to 32 mg of material. Calculate the amount of crude sample prior to performing the sublimation.

NOTE. The following experiments use Technique 9: Experiments [11B], [25A], and [25B].

C$_6$H$_6$apter 6

MICROSCALE ORGANIC LABORATORY EXPERIMENTS

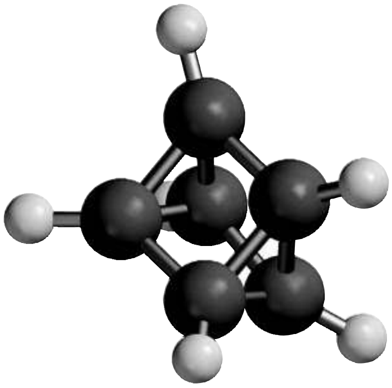

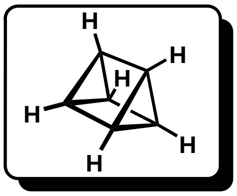

This chapter contains the experimental details of a collection of famous organic reactions that are at the heart of this field of chemistry. One of the great triumphs of the human intellect has been our ability to rationalize the physical transformations of organic materials. The experimental laboratory is the source of information out of which predictive organic theory has been fashioned. The microscale organic laboratory is designed to give you the opportunity to experience, first hand, how organic reactions occur. This program will allow you to see how experimental data are directly related to the development of the structural and mechanistic theory surrounding these transformations. If successful, the microscale organic laboratory should bring the lecture portion of your course to life. What you have been studying in two dimensions in lecture, now becomes alive in three dimensions in the laboratory.

The 35 experiments that make up this chapter focus largely on some of the most important of the fundamental organic reactions that have been discovered over the last two centuries. Because the application of these reactions to synthesis has been extensive, the microscale laboratory program offers a broad and practical introduction to organic chemistry. A significant number of the experiments include optional scale-up procedures to provide laboratory experience at the semimicroscale level, an inquiry-driven format allowing the student to monitor the reaction prior to work-up, validation opportunities to confirm product purity by TLC analysis, and microwave procedures if alternative formats of experimental work are desired. Studying one or two of these modified formats at the beginning of the second semester can be helpful, particularly if some of the multistep syntheses covered in Chapter 7 are planned for study later in that semester. These latter sequences make extensive use of semimicroscale experimentation.

? highlights an inquiry-driven format

V highlights a validation opportunity

M highlights a microwave procedure

The microscale experiments are designed to enhance your ability to master the miniaturized experimental techniques used. A *Prior Reading* section highlights which techniques are to be used in a particular experiment and outlines the pages to refer to in the Techniques section (see Chapter 5). At the beginning of the semester it is important to *review these sections before proceeding* with the assigned experiment (heed the advice in Chapter 1). After examining the written material, it would be particularly helpful if you could see the video we have prepared that is accessible online. It covers the techniques in detail and can be viewed either in your spare time or in a prelab lecture period. As the basic techniques are used repeatedly, you will rapidly gain a firm grasp of the manipulations required and become comfortable working at this scale. You

← www

Chapter 6: C$_6$H$_6$, Prismane
Katz and Acton (1973).

should also take advantage of the large number of graphics in the text margin that detail how the experimental apparatus are assembled. There are prompt signs (➨) in the text near the point where the equipment is to be used.

www →

www →

The important role of spectroscopic techniques in the modern organic laboratory is emphasized in the microscale organic laboratory. This information is set aside as a separate chapter, along with additional valuable details on the website, which are reference materials to be used while undertaking the experimental sections in Chapters 6–7. Numerous cases of detailed analyses of the spectra are found in the *Purification and Characterization* sections of the experiments. Because infrared spectroscopy continues to play a major role in the characterization of reaction products in the introductory laboratory, and because this technique is currently given only a cursory treatment in most lecture texts, we have included a fairly detailed qualitative introduction to the theory of the effect and the instrumentation used in obtaining these observations, principally in Chapter 8 and on the website. *While the IR part of the spectroscopic section (including the on line material) may cover more ground than is normally found in many introductory laboratories, we feel it is important to overcome the black-box attitude that students can rapidly develop toward complex chemical instrumentation when they are turned loose on these powerful instruments with very limited knowledge.* Thus, we hope to be able to accomplish this transformation of attitude by offering students the essential details that will allow them to gain a command of the logic and the mechanics of obtaining and interpreting infrared spectral data. As mentioned earlier, *we have also included, in a number of the experiments, illustrative examples of the detailed spectral analysis* that can be used to examine starting materials, follow the progress of the reaction, and finally to assess the character and purity of the products.

A discussion section precedes each experiment. In a number of cases, especially when named reactions are involved, a brief biographical sketch of the individual so honored is included. At this point we also introduce pertinent information concerning the reaction mechanism, often in considerable detail. When appropriate, the relevance of a reaction to the life sciences and the chemical industry is explored.

Note that *Safety* and *Warning* indicators are highlighted or boxed in the experiments. We urge you to *always* adhere strictly to the safety precautions listed.

The nomenclature of organic compounds is often confusing to the beginning student and even occasionally to the experienced research chemist! To ease your introduction to the name game, the common name (sometimes referred to as the *trivial name*) of the compound to be synthesized is also listed at the beginning of each experiment. In addition, the ***Chemical Abstracts*** (CA) number and CA index name are also given.

Good luck! Enjoy your adventure in transforming small quantities of a large number of organic materials.

EXPERIMENT 1

Getting to Know You: Measurement of Physical Properties

Purpose. This experiment will acquaint you with the experimental techniques used to measure certain classic physical properties of organic substances. These properties include the boiling point, density, and refractive index for liquids, and the melting point for solids. The procedures are outlined in Chapter 4.

A further objective is to study sampling techniques for obtaining the spectral characteristics of organic materials. You will observe that absorption spectra of organic substances contain, without question, the most important collection of physical constants of a material available to the investigator. Spectral information can lead to the elucidation of molecular structure and the rapid identification of unknown substances.

You will also learn how to locate the literature values for these measured properties using various chemical handbooks and online resources.

Prior Reading

Chapter 4: Determination of Physical Properties
- Ultramicro-Boiling Point (pp. 46–48)
- Density (pp. 49–50)
- Refractive Index (online)
- Melting Point (pp. 50–54)
  - Capillary Method (p. 51)
  - Evacuated Technique (pp. 51–52)
  - Mixture Melting Points (pp. 52–54)

Chapter 5, Technique 6: Thin-Layer Chromatography (pp. 97–99)

www

DISCUSSION

Organic compounds have a number of physical properties that allow their precise characterization. These include the classical physical constants: boiling point, density, and refractive index for liquids and melting point for solids. The rapid development of modern chemical instrumentation, however, has also made easily accessible many of the spectral properties of these materials. Spectroscopy, in particular, provides information that is extremely powerful for establishing the structure of unknown molecular systems and for rapidly identifying known materials.

Not only are physical properties used to characterize a specific organic compound, but they are often used to compare one compound to another. Examples of this approach are illustrated in Chapter 9, Qualitative Identification of Organic Compounds. The route to identification of an unknown organic species has become increasingly dependent on the measurement of the physical properties of the pure substance.

The various classical physical constants of a large number of organic substances appear in the *CRC Handbook of Chemistry and Physics, Lange's Handbook, Aldrich Catalog Handbook of Fine Chemicals*, and several online resources. The *Aldrich Catalog* also contains references to published infrared (IR) and nuclear magnetic resonance (NMR) spectral data for many of the compounds. Complementing these databases are several online resources, such as SciFinder Scholar, which allow for users to directly and efficiently access published spectral data. The *CRC Atlas of Spectral Data and Physical Constants for Organic Substances* contains IR and ultraviolet (UV) peak positions. Collections of spectra may be located in the Aldrich Libraries of both IR and NMR spectra, and in the Sadtler Library.

A detailed discussion of the spectral properties of organic compounds is given in Chapter 8 and on the website.

www

EXPERIMENTAL PROCEDURE

NOTE. Your instructor will select which of the physical properties you are to measure. The length of your laboratory period, the size of your section, and the number of instruments available will all play a role in determining how many of these properties will be suggested.

Melting Point

1. Using the melting point apparatus (*your instructor will provide you with the experimental details concerning the operation of the particular instrument to be used in your laboratory)*, determine the melting point of acetanilide and compare your result with the literature values (Lit. values) reported in the CRC and Aldrich references (see pp. 50–54 for the experimental details on how to proceed with this measurement).

Determined value __________

CRC: Lit. value __________; Ed. __________; Page __________;
Compound No. __________

Aldrich: Lit. value __________; Year __________; Page __________;
Compound No. __________

2. Now determine an ***evacuated*** melting point (see pp. 51–52 for the experimental details on how to proceed with this measurement) of caffeine and compare it to the values in the CRC and Aldrich references. *Your instructor will provide you with the experimental details concerning the operation of the particular instrument to be used in your laboratory.*

First, determine the melting point of caffeine in an unevacuated melting-point tube. After observing the melting point, allow the temperature to drop below the melting point and observe whether the sample crystallizes again. If crystallization occurs, observe a second melting point, and then repeat this procedure a third time. Follow the same routine with an evacuated sample, and compare the results of the two sets of melting points. Do you observe any differences between these data sets?

2nd Determined value

Unevacuated	***Evacuated***
1st Determined value __________	1st Determined value __________
2nd Determined value __________	2nd Determined value __________
3rd Determined value __________	3rd Determined value __________

CRC: Lit. value __________; Ed. __________; Page __________;
Compound No. __________

Aldrich: Lit. value __________; Year __________ ; Page __________;
Compound No. __________

3. Why is a second seal required on the evacuated melting-point tube?

4. When are evacuated melting points necessary?

5. Next consider the technique of ***mixture melting points*** (see pp. 52–54 for a discussion of the details of this procedure). This technique is the classical approach for establishing a positive identification of a substance when pure reference standards are available in the laboratory. For two examples where this type of measurement is applied in the microscale organic laboratory, see Experiments [6] and [34A].

6. In the present experiment, you have obtained the melting points of acetanilide and caffeine. Using these values as reference standards observe the melting points of two mixtures: (1) caffeine 75%–acetanilide 25% and (2) caffeine 25%–acetanilide 75%. Should these observations be made in evacuated capillaries or not? *(Your instructor will provide you with the experimental details concerning the operation of the particular instrument to be used in your laboratory.)*

(1) caffeine 75%–acetanilide 25%	(2) caffeine 25%–acetanilide 75%
Determined value__________	Determined value__________

7. (Optional) Use the following compounds and mixtures: caffeine, acetanilide, caffeine 75%–acetanilide 25%, and caffeine 25%–acetanilide 75%; your instructor will provide no less than one duplicate set of vials. For sections that consist of an odd number of students, triplication of one set will work. While the vials will be labeled, the labels will not offer any direction as to the contents nor its match. Once the vials are distributed, your goal is to determine the melting point of your "unknown" and find your match. Caution should be exercised since depending on how the vials are "duplicated," there may be more than one match! That is, when considering a laboratory section consisting of 18 students, the 18 vials may consist of 10 vials of caffeine and 8 vials of acetanilide. This could as well be presented as 4 vials of caffeine, 4 vials of acetanilide, 4 vials of caffeine 75%–acetanilide 25%, and 6 vials of caffeine 25%–acetanilide 75%. As you can surmise, an even more challenging exercise can be crafted upon the creative mixing of the pairings above. Success will come to those who correctly determine the melting point of their unknown and thus multiple melting point determinations may be needed prior to seeking your match.[1]

8. Is it possible to detect the presence of impurities by melting-point measurements? Why?

9. Would it be possible to establish the composition of an unknown binary mixture of two substances from mixture melting-point data? Explain.

Ultramicro-Boiling Point

1. Make several (5) glass bells (p. 46). Your instructor will demonstrate the procedure, or you can view the procedure on video, if available. Place the bells in a small glass vial and store them in your micro kit.

2. Use the technique for determining ultramicro-boiling points (bp) on a *melting-point apparatus hot stage* (very likely this well be the same apparatus you used in melting-point section of the experiment) as discussed in

---

[1] This exercise is based upon discussions led by Jerry R. Mohrig and Christina Noring Hammond during an NSF-sponsored CWCS Workshop on Teaching Guided-Inquiry Organic Chemistry Laboratories (Irvine, CA; July **2005**).

Chapter 4 (p. 47), and complete the following table *(your instructor will provide you with the experimental details concerning the operation of the particular instrument to be used in your laboratory).*

	Propane	Butane	Pentane	Hexane	Heptane
bp (°C)	−42.1	−0.5	36.1	67	?

3. Compare your measured value for heptane to those listed in the *CRC Handbook, Aldrich Catalog*, and online.

Determined value __________; Barometric pressure __________ torr
Corrected value __________

CRC: Lit. value __________; Ed. __________; Page __________;
Compound No. __________

Aldrich: Lit. value __________; Year __________; Page __________;
Compound No. __________

4. Prepare and attach a graph of molecular weight (MW) vs. boiling point (bp) using the above data.

5. Explain the trend in boiling point as the molecular weight increases.

Density

1. Use the syringe method of measuring the density outlined in Chapter 4 (p. 49), and 0.5 mL of liquid, to complete the following table and compare your value to those listed in the CRC and Aldrich references, *(Your instructor will provide you with the experimental details concerning the operation of the particular equipment to be used in this experiment. Pay close attention to the details concerning the operation of the balances.)*

Methylene Chloride **(CH_2Cl_2)**		***Octane*** **C_8H_{18}**
Density (g/mL)	1.33	?
Determined value	__________	__________

CRC: Lit. value __________; Ed. __________; Page __________;
Compound No. __________

Aldrich: Lit. value __________; Year __________; Page __________;
Compound No. __________

2. Methylene chloride is often used as a solvent in the laboratory to extract an *organic species* from an aqueous solution. Will the CH_2Cl_2 normally form the top or bottom layer?

3. If octane were used to extract an aqueous phase, would it form the top or bottom layer? How did you arrive at this answer?

Refractive Index

www → **1.** Using the lens paper disk technique outlined online determine the refractive index of methanol and compare the value you have obtained to that found in the CRC and Aldrich data collections. *(Your instructor will* *provide you with the experimental details concerning the operation of the particular instrument used in your laboratory.)*

Determined value __________

CRC: Lit. value __________; Ed. __________; Page __________;
Compound No. __________

Aldrich: Lit. value __________; Year __________; Page __________;
Compound No. __________

2. Correct your measured value to 20 °C.

Temperature of your measurement __________.

Temperature of literature value __________.

Your measured value corrected to 20 °C __________.

Calculations:

3. Based on your observations, how does refractive index vary with temperature?

Thin-Layer Chromatography

1. Using the thin-layer chromatography technique outlined in Chapter 5 (Technique 6A, pp. 97–99), determine the R_f values for benzaldehyde, 4-chlorobenzaldehyde, and cyclohexanol using as a solvent system ethyl acetate and hexane (1:4). While a UV lamp will be sufficient for the visualization of benzaldehyde and 4-chlorobenzaldehyde when working with TLC plates with a UV-activated fluorescent indicator mixed in with the silica gel, a staining system of *p*-anisaldehyde will work when wanting to visualize cyclohexanol. *(Your instructor will provide you with the experimental details concerning the operation of staining a TLC plate using* p*-anisaldehyde. Please use for preparation of* p*-anisaldehyde solution: 135 mL EtOH, 5 mL concentrated* H_2SO_4*, 1.5 mL glacial acetic acid, and 3.7 mL* p*-anisaldehyde.)*

2. After recording the R_f values using as a solvent system ethyl acetate and hexane (1:4), determine what solvent system is needed to obtain values between 0.3−0.4 for each of the following systems: benzaldehyde, 4-chlorobenzaldehyde, and cyclohexanol.

3. Using as a solvent system ethyl acetate and hexane (1:4), predict which compound from the following set would have a higher R_f value: (a) cyclohexanol and cyclohexane, (b) benzoic acid and benzaldehyde, and (c) caffeine and naphthalene.

4. (Optional) Three stock solutions labeled solution A, solution B, and solution C, will be provided by your instructor. They will be one of the following systems: benzaldehyde, 4-chlorobenzaldehyde, or cyclohexanol. Select one (or more) and determine its identity using the thin-layer chromatography technique having already established R_f values in several solvent systems.

Infrared Spectroscopy Sampling Procedures

Sampling of Liquids

1. Use the technique for obtaining an IR spectrum of a thin-liquid film as described on page 553 (*your instructor will provide you with the experimental* *details concerning the operation of the particular instrument to be used in your laboratory and the type of windows on which your sample will be mounted).*

2. Obtain the spectra of *n*-octane and 1-octanol (use a single scan with Fourier-transform [FT] instruments). Compare your data with the Aldrich or Sadtler reference collections, or those obtained online using for example SciFinder Scholar, and compare the two spectra to each other.

3. What differences do you observe between the two spectra? Can you associate differences in molecular structure to the differences in the spectra (see Infrared Discussions in Chapter 8 and on the website relating to the IR spectra of *n*-hexane and 1-hexanol).

4. Occasionally, when an IR spectrum is obtained, some of the very strong bands will appear with flattened peaks, as if they were totally absorbing the energy at those wavenumber values. The flat bottom of the band, however, does not correspond to 0% transmission on the scale, but will indicate an energy transmission of 5–10% or even higher. Can you explain the odd shape of the band? Is the energy being totally absorbed or not? Explain your answers.

Sampling of Solids

1. Use the technique for obtaining an IR spectrum of a solid melting above 100 °C as described in Chapter 8 (pp. 553–554) *(your instructor will provide you with the experimental details concerning the operation of the particular instrument to be used in your laboratory and the type of KBr die in which your sample will be pressed if an ATR FTIR is unavailable).*

2. Obtain the IR spectrum of caffeine (use four scans with FT instruments). Compare your data with an authentic spectrum of caffeine and with the data given in Experiment [11B]. Your sample may be saved by taping it to a file card with your name, and stored by your instructor in a desiccator. If, later in the year, you isolate caffeine from its natural source (Experiment [11B]), you will be able to compare the material you have extracted and purified from the plant with your own authentic reference spectrum.

3. Occasionally, the spectral region from 4000 to 2000 cm^{-1} in solid samples is steeply sloping to lower transmission values at higher wavenumber values. Is this drop in transmission an absorption of the radiation? If so, to what process can the absorption be ascribed, and if not, what is the cause of the decreased transmission?

4. Compare the spectra of caffeine obtained in your laboratory section. Are the spectra all identical? Where do they differ? To what effect can you ascribe the differences, if there are any? (This is a good open question for the entire lab section.)

Nuclear Magnetic Resonance Spectroscopy Sampling Procedures: For NMR sampling procedures and examples see Chapter 8 and Experiments [5B], [22A], and [28]

QUESTIONS

6-1. An unknown carboxylic acid has a boiling point of 100 °C at 25 torr. Using the pressure–temperature nomograph on page 49, determine its boiling point at 760 torr. Identify the acid from the list in Appendix A, Table 9W.1.

6-2. Discuss with those in your laboratory section the consequences of incorrect sample loading, variable sample size, and rates of heating and how these factors might lead you to obtain an incorrect value of the melting point for your solid sample.

6-3. The mass of a certain volume of an unknown liquid at 25 °C is 234 mg. The mass of an equal volume of water at the same temperature is 201 mg. Calculate the density of the unknown liquid at 25 °C.

6-4. The melting point of pure *trans*-cinnamic acid is 133 °C, and that of 2-acetoxybenzoic acid (aspirin) is 135 °C (Appendix A, Table 9W.2). Given pure reference standards of both acids, describe a melting-point procedure by which you could identify whether an unknown sample melting in this range could be assigned to either structure, or to neither one.

www

6-5. Why is filter paper a poor choice of surface on which to powder or crush a solid crystalline sample before placing it in a capillary melting-point tube?

6-6. Why is the ultramicro-boiling point determined precisely at the point when the last vapor bubble has escaped and the liquid phase begins to rise in the bell cavity?

BIBLIOGRAPHY

For further information on basic laboratory techniques:

Sharp, J. T.; Gosney, I; Rowley, A. G. *Practical Organic Chemistry: A Student Handbook of Techniques,* Chapman Hall: London, 1989.

Shriner, R. L.; Hermann, C. K. F.; Morrill, T. C.; Curtin, D. Y.; Fuson, R. C. *The Systematic Identification of Organic Compounds,* 8th ed.; Wiley: New York, 2003.

Vogel, A. I. *Vogel's Textbook of Practical Organic Chemistry,* 5th ed.; Furnis, B. S.; et al. Eds.; Wiley: New York, 1989.

Zubrick, J. W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, 2008.

EXPERIMENT 2

The Separation of a 25-μL Mixture of Heptanal (bp 153 °C) and Cyclohexanol (bp 160 °C) by Gas Chromatography

Purpose. This experiment illustrates the separation of a 25-μL mixture, consisting of heptanal and cyclohexanol, into the pure components. The volume of the mixture is approximately that of a single drop, and the materials boil within 7 °C of each other. This mixture would be difficult, if not impossible, to separate by the best distillation techniques available. The purity of the fractions collected from the gas chromatograph (GC) can be assessed by boiling points, refractive indexes, or infrared (IR) spectra.

Prior Reading

Technique 1: Microscale Separation of Liquid Mixtures by Preparative Gas Chromatography (pp. 55–61)
Chapter 4: Determination of Physical Properties
Ultramicro-Boiling Point (pp. 46–48)
Refractive Index (online) www
Experiment [1]: Measurement of Physical Properties
Ultramicro-Boiling Point (p. 119)
Chapter 8: Introduction to Infrared Spectroscopy (pp. 539–561)

DISCUSSION

The efficacy of GC separations is highly dependent on the experimental conditions. For example, two sets of experimental data on the heptanal–cyclohexanol mixture are given below to demonstrate the effects of variations in oven temperature on retention times.

In Data Set A, the oven temperature was allowed to rise slowly from 160 to about 170 °C during a series of sample collections. The retention time of heptanal dropped from about 3 min to close to 2 min, whereas the retention time of cyclohexanol was reduced from about 5.5 min to nearly 4 min. The significant decrease in resolution over this series of collections is reflected in the number of theoretical plates calculated, which was over 300 for heptanal and about 500 for cyclohexanol in the first trial, but declined to below 200 for both compounds toward the last run (see Data Set A).

COLLECTION YIELD

Cyclohexanol

Density of cyclohexanol = 0.963 mg/μL.

In 25 μL of 1:1 cyclohexanol–heptanal, we have 12.5 μL of cyclohexanol.

Therefore, 125 μL × 0.963 mg/μL = 12 mg of cyclohexanol injected.

Percent recovered = (8.3 mg/12.0 mg) × 100 = 69% cyclohexanol collected.

Heptanal

Density of heptanal = 0.850 mg/μL.

Therefore, 12.5 μL × 0.85 mg/μL = 10.6 mg of heptanal injected.

Percent recovered = (8.1 mg/10.6 mg) × 100 = 76% heptanal.

In the Data Set B collections, stable oven temperatures and flow rates were maintained, and the data exhibit excellent reproducibility. Oven temperature was held at 155 °C throughout the sampling process. The retention time of heptanal was observed to be slightly longer than 3 min, with a variance of 6 s, whereas the cyclohexanol retention time was found to be slightly longer than 6 min, with a variance of 12 s. The resolution remained essentially constant throughout the series, and the number of theoretical plates

Data Set A

	Heptanal				Cyclohexanol			
Trial No.	Retention Time (min)	Baseline Width (min)	Number of Theoretical Plates	Recovery (mg)	Retention Time (min)	Baseline Width (min)	Number of Theoretical Plates	Recovery (mg)
1	3.1	0.7	314	8.0	5.6	1.0	502	8.0
2	2.9	0.7	275	8.0	5.3	1.0	449	8.0
3	3.0	0.7	294	7.0	5.7	1.0	520	8.0
4	2.8	0.7	256	8.0	5.1	1.1	344	8.0
5	2.5	0.6	278	8.0	4.3	1.1	244	9.0
6	2.7	0.5	467	7.0	4.6	1.0	339	10.0
7	2.5	0.6	278	10.0	4.2	1.0	282	8.0
8	2.2	0.5	310	9.0	3.5	1.0	196	8.0
9	1.8	0.5	207	8.0	3.0	1.0	144	8.0
10	2.3	0.7	173	8.0	3.9	1.0	243	8.0
Av	2.5 ± 0.4	0.6 ± 0.09	285 ± 7	8.1 ± 0.8	4.5 ± 0.9	1.0 ± 0.05	326 ± 129	8.3 ± 0.7

Data Set B

	Heptanal				Cyclohexanol			
Trial No.	Retention Time (min)	Baseline Width (min)	Number of Theoretical Plates	Recovery (mg)	Retention Time (min)	Baseline Width (min)	Number of Theoretical Plates	Recovery (mg)
1	3.5	0.7	400	8.0	6.6	1.1	576	8.0
2	3.2	0.7	334	9.0	6.0	1.1	476	7.0
3	3.5	0.7	400	7.0	6.6	1.2	484	10.0
4	3.2	0.7	334	9.0	6.1	1.0	595	9.0
5	3.1	0.6	427	8.0	6.0	1.1	476	8.0
6	3.2	0.7	334	9.0	6.0	1.1	476	9.0
7	3.3	0.8	272	9.0	6.1	1.1	492	8.0
8	3.1	0.7	313	8.0	6.0	1.1	476	10.0
9	3.2	0.7	334	8.0	6.1	1.1	492	8.0
10	3.2	0.7	334	8.0	6.2	1.1	508	8.0
Av	3.2 ± 0.1	0.7 ± 0.05	348 ± 47	8.3 ± 0.7	6.2 ± 0.2	1.1 ± 0.05	505 ± 44	8.5 ± 1.0

calculated was about 350 for heptanal and about 500 for cyclohexanol (see Data Set B).

COLLECTION YIELD

Cyclohexanol

Density of cyclohexanol = 0.963 mg/μL.

In 25 μL of 1:1 cyclohexanol–heptanal, there are 12.5 μL of cyclohexanol.

Therefore, 125 μL × 0.963 mg/μL = 12 mg of cyclohexanol injected.

Percent recovered = (8.5 mg/12.0 mg) × 100 = 71% cyclohexanol collected.

Heptanal

Density of heptanal = 0.850 mg/μL.

Therefore, 12.5 μL × 0.85 mg/μL = 10.6 mg of heptanal injected.

Percent recovered = (8.3 mg/10.6 mg) × 100 = 78% heptanal.

The results just described demonstrate that the resolution of GC peaks may be very sensitive to changes in retention time resulting from instability in oven temperatures. Since the number of theoretical plates is related to resolution values, significant degradation in column plate values can occur with variations in oven temperatures. When you compare the time and effort required to obtain a two-plate fractional distillation on a 2-mL mixture (see Experiment [3B] and Technique 2) with the speed and ease used to obtain a 500 plate separation on 12.5 μL of cyclohexanol in this experiment, it is hard not to be impressed with the enormous power of this technique.

COMPONENTS

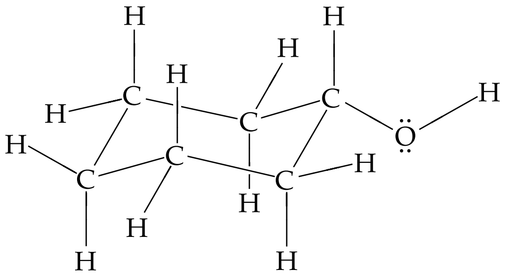

Cyclohexanol

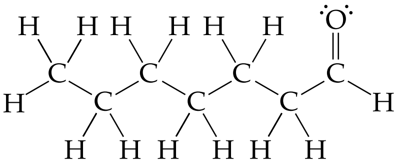

Heptanal

EXPERIMENTAL PROCEDURE

Estimated time for the experiment: 2.0 h.

Physical Properties of Components

Compound	MW	Amount	bp (°C)	Density (*d*)	n_D
Heptanal	114.19	12.5 μL	153	0.85	1.4113
Cyclohexanol	100.16	12.5 μL	160	0.96	1.4641

Reagents and Equipment. The procedure involves injecting a 25-μL mixture of heptanal–cyclohexanol 1:1 (v/v) into a $\frac{1}{4}$-in. × 8-ft stainless-steel column packed with 10% Carbowax 80/100 20M PAW-DMS. Experimental conditions (GOW-MAC series No. 350) are He flow rate, 50 mL/min; chart speed, 1 cm/min; oven temperature, 155 °C.

Procedure for Preparative Collection. The liquid effluents are collected in an uncooled, 4-mm-diameter collection tube (double reservoirs; overall tube length 40–50 mm, see Fig. 6.1)

The collection tube (oven dried until 5 min before use) is attached to the heated exit port by the 5/5 ₮ joint. Sample collection is initiated 0.5 min prior to detection on the recorder of the expected peak (time based on previously determined retention values)[2] and continued until 0.5 min following return to baseline. After the collection tube is detached, the sample is transferred to the 0.1-mL conical GC collection vial. The transfer is facilitated by the 5/5 ₮ joint on the conical vial. After the collection tube is joined to the vial (preweighed with stopper), the system is centrifuged (see Fig. 6.1). The collection tube is then removed and the vial is stoppered and reweighed.

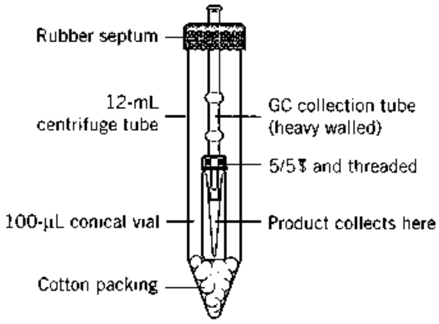

Figure 6.1 Gas chromatographic collection tube and 0.1-mL conical vial.

Characterization. Calculate the percent recovery. These amounts should range between 7 and 10 mg. Determine the boiling point of each fraction and obtain the refractive index or IR spectrum, if time permits. These latter measurements will require most, if not all, of the sample not used for boiling-point determination.

www

Assess the purity and efficiency of the separation from your tabulated data and the GC chromatogram.

---

[2] Refer to your local laboratory instructions.

Alternative Mixture Pairs for Preparative Collection

(all mixtures are 1:1 v/v)

Mixture

a. *Separation* of a 40-μL mixture of[3]
(1S)-(−)-α pinene (bp 156 °C, $n_D = 1.4650$, $d = 0.855$)
(1S)-(−)-β pinene (bp 165 °C, $n_D = 1.4782$, $d = 0.859$)

Components

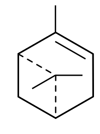
(1*S*)-(−)-α-Pinene

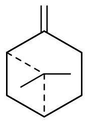
(1*S*)-(−)-β-Pinene

Chromatographic Parameters

A 40-μL injection
Flow rate: 50 mL/min
Column temperature: 120 °C
Column: 20% Carbowax
Elution time
  α-Pinene: ~8 min
  β-Pinene: ~12 min
Average recovery
  α-Pinene: 8.3 μL
  β-Pinene: 10.6 μL

Mixture

b. *Separation* of a 40-μL mixture of
2-Heptanone (bp 149–150 °C, $n_D = 1.4085$, $d = 0.820$)
Cyclohexanol (bp 160–161 °C, $n_D = 1.4641$, $d = 0.963$)

Components

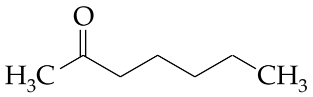
2-Heptanone

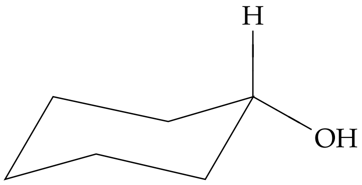
Cyclohexanol

Chromatographic Parameters

A 40-μL injection
Flow rate: 50 mL/min
Column temperature: 145 °C
Column: 20% Carbowax
Elution time
  2-Heptanone: ~5.5 min
  Cyclohexanol: ~10.0 min

---

[3] Refractive index at D line of sodium = n_D and density = d.

Average recovery
2-Heptanone: 8.1 μL (41%)
Cyclohexanol: 11.4 μL (57%)

Mixture

c. *Separation* of a 40-μL mixture of
d-Limonene (bp 175–176 °C, $n_D = 1.4743$, $d = 0.8402$)
Cyclohexanol acetate (bp 173 °C, $n_D = 1.4401$, $d = 0.9698$)

Components

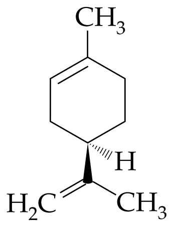

d-Limonene

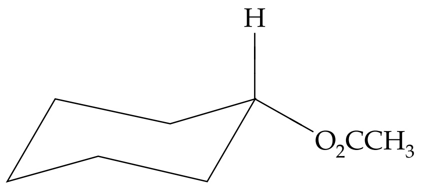

Cyclohexyl acetate

Chromatographic Parameters

A 40-μL injection
Flow rate: 50 mL/min
Column temperature: 170 °C
Column: 20% Carbowax
Elution time
d-Limonene: ~5.5 min
Cyclohexyl acetate: ~7.5 min
Average recovery
d-Limonene: 8.7 μL (44%)
Cyclohexyl acetate: 10.0 μL (50%)

QUESTIONS

6-7. Based on the data presented in the Data Set A chromatographic separation, can you explain why there is such a steep decline in column efficiency with temperature change?

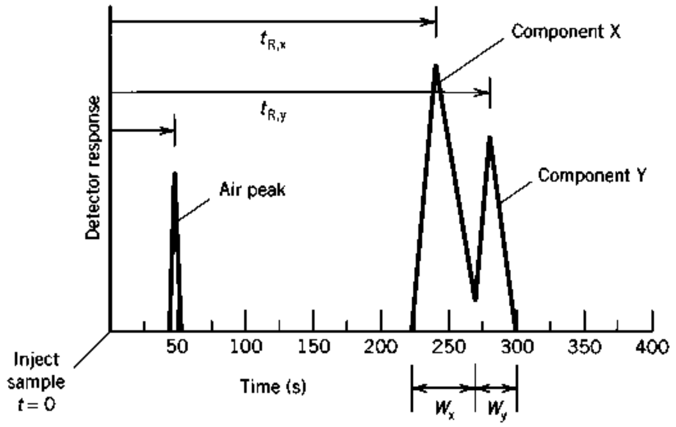

6-8. Consider the following gas chromatogram for a mixture of analytes X and Y:
- **(a)** Calculate the number of theoretical plates for the column in reference to the peaks of each component (X and Y).
- **(b)** If the column is 12 m long, calculate the height equivalent theoretical plate (HETP (in plates per cm) for this column.

6-9. The number of theoretical plates a column has is important, but the crucial factor is the ability to separate two or more substances. That is, how well resolved are the peaks? The resolution of two peaks depends not only on how far apart they are (t_R), but also on the peak width (W). Baseline resolution (R) is defined by the following equation:

$$R = \frac{2\Delta t_R}{W_X + W_Y}$$

Because of the tailing of most species on the column, a value of 1.5 is required to give baseline resolution.
- **(a)** Calculate the resolution for the peaks in Question 6–8.
- **(b)** Do you think a quantitative separation of the mixture is possible based on your answer?
- **(c)** Has baseline resolution been achieved?

6-10. Discuss at least two techniques you might use to increase the resolution of the column in Question 6–9 (without changing the column).

6-11. Retention times for several organic compounds separated on a GC column are given below.

Compound	**t_R (s)**
Air	75
Pentane	190
Heptane	350
2-Pentene	275

- **(a)** Calculate the relative retention of 2-pentene with respect to pentane.
- **(b)** Calculate the relative retention of heptane with respect to pentane.

BIBLIOGRAPHY

Selected references on gas chromatography:

Grob, R. L.; Barry, E. F., Eds.; *Modern Practices of Gas Chromatography;* Wiley: New York, 2004.

Jennings, W.; Mittiefehidt, E.; Stremple, P., Eds.; *Analytical Gas Chromatography;* 2nd ed., Academic Press: New York, 1997.

McNair, H. M.; Miller, J. M. *Basic Gas Chromatography;* Wiley: New York, 1997.

Sadek, P. C. *Illustrated Pocket Dictionary of Chromatography;* Wiley*: New York, 2004.*

Distillation

EXPERIMENT 3

In the following set of experiments, we will examine the applications of a variety of distillation techniques to the purification of liquid mixtures. In Experiments [3A] and [3B] you will conduct simple distillations. In Experiment [3A] a volatile liquid component is separated from a nonvolatile solid. Experiment [3B] illustrates the use of the Hickman still in the separation of hexane and toluene, which have boiling points 42 °C apart. The composition of the fractions is analyzed by refractive index and boiling point. Experiments [3C] and [3D] introduce the use of micro spinning-band distillation columns for the separation of cyclohexane (bp 80.7 °C) and 2-methylpentane (bp 60.3 °C). The composition of the distillate fractions are determined by gas chromatography. The number of theoretical plates is determined for the spinning-band column used. In Experiment [3D] you will be introduced to one of the simplest yet most efficient and powerful distillation techniques for the separation of liquid mixtures at the semimicroscale level, the Hickman–Hinkle still.

Experiment 3A

Simple Distillation at the Semimicroscale Level: Separation of Ethyl Acetate from *trans*-1,2-Dibenzoylethylene

Purpose. *Simple distillation* is examined using the distillation process to separate a liquid ester from minor components that are nonvolatile or that have boiling points much greater (>100 °C) than that of the major component.

Prior Reading

Techniques 2 and 3: Distillation (pp. 61–67 and online)
Distillation Theory (p. 61 and online)
Simple Distillation at the Semimicroscale Level (pp. 61–64)
Chapter 4: Determination of Physical Properties
Ultramicro-Boiling Point (pp. 46–48)
Density (pp. 49–50)
Refractive Index (online)
Evacuated Melting Point (pp. 51–52)

DISCUSSION

Semimicroscale simple distillation can be an effective separation technique with volumes from 0.5 to 2 mL. Apparatus have been developed that achieve effective separation of mixture samples in this range. One of the most significant of these designs is the classic Hickman still, shown in both Figures 5.5 and 6.2 because of its importance to the associated discussions. In this experiment you will effect the separation of a two-component mixture by the use of this still.

The Hickman still is used in several of the microscale experiments to purify solvents, carry out reactions, and concentrate solutions for recrystallization. An introduction to the use of the Hickman still is given in this experiment.

In a distillation where a liquid is separated from a nonvolatile solute, the vapor pressure of the liquid is lowered by the presence of the solute, but the vapor phase consists only of the pure liquid component. Thus, except for the transfer of nonvolatile material by incidental splashing, the material condensed in the collar of the Hickman still should consist only of the volatile component. In the present experiment this component is ethyl acetate. The temperature of the vial and contents being distilled will rise during the distillation process, since the concentration of the impurity is increasing as the volatile liquid is removed. This effect lowers the vapor pressure of the liquid. However, the boiling point of the liquid remains constant, since only the pure liquid component is being vaporized.

COMPONENTS

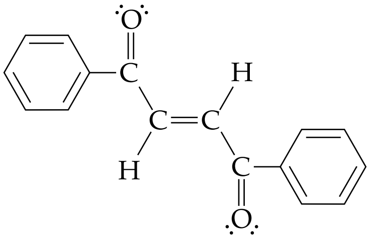

trans-1,2-Dibenzoylethylene

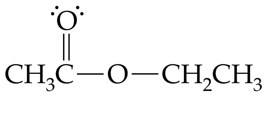

$CH_3C(=O)—O—CH_2CH_3$

Ethyl acetate

EXPERIMENTAL PROCEDURE

Estimated time for the experiment: 2.0 h.

Physical Properties of Components

Compound	MW	Amount	mp (°C)	bp (°C)	*d*	n_D
Ethyl acetate	88.12	1.0 mL		77	0.90	1.3723
trans-1,2-Dibenzoylethylene	236.27	50 mg	111			

Reagents and Equipment. Transfer 1 mL of the yellow stock solution (*trans*-1,2-dibenzoylethylene/ethyl acetate, 50 mg/mL) to a 3-mL conical vial by automatic delivery pipet (remember to place the vial in a small beaker to prevent tipping during the transfer). Place a boiling stone (or a magnetic spin vane if desirable) into the vial and assemble the Hickman still head. The still assembly is mounted in a sand bath on a hot plate (see Fig. 6.2).

Experimental Conditions. The temperature of the bath is raised to 90–100 °C at a rate of 5 °C/min.

CAUTION: Do not let the temperature of the still rise too rapidly.

Once boiling commences, the rate of heating should be lowered to the point where the temperature increases at 2–3 °C/min. A slow distillation rate is very important in establishing equilibrium between the vapor and liquid components in the mixture. Follow the course of the distillation by the rise of condensate on the sides of the Hickman column. When the condensate reaches the trap, adjust the bath temperature so that liquid is removed from the column slowly (~100 μL/3 min). A smooth, slow distillation will provide a cleaner

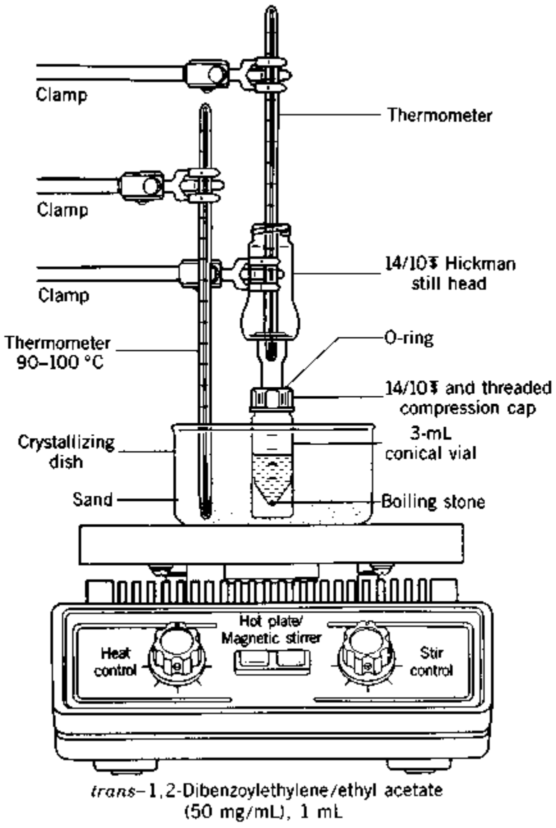
Clamp
Thermometer
Clamp
14/10₮ Hickman still head
Clamp
Thermometer 90–100 °C
O-ring
14/10₮ and threaded compression cap
Crystallizing dish
3-mL conical vial
Sand
Boiling stone
Hot plate/ Magnetic stirrer
Heat control
Stir control
trans-1,2-Dibenzoylethylene/ethyl acetate (50 mg/mL), 1 mL

Figure 6.2 Hickman still (14/10 ₮ with conical vial [3 mL]).

separation of the components, and will also avoid mechanical transfer of nonvolatile components via splattering to the condensate trap (if the condensate appears yellow, contamination has occurred).

Collect approximately 50–150 μL of the ester in the collar of the still (the first fraction collected is often referred to as the forerun; give the temperature range). As the distillation continues, remove the forerun with a Pasteur pipet having a slightly bent tip (bend the tip with a microburner—if you have a Hickman still with a side-arm collection port, the pipet tip will not need to be bent). Place the fraction in a clean, dry, 1-dram, screw-capped vial (use an aluminum foil liner to avoid cap contamination). Label the fraction with a marking pen. Collect a second fraction of ester (400–500 μL, which may require combining two, or even three, collections from the collar; give the temperature range), which should be clear and colorless. Remove and store as before. Discontinue the distillation. Allow the distilling flask to cool slowly by leaving it in the warm sand bath while measuring the physical properties of the distillate fractions.

Characterization. Three physical properties of the ester will be measured to establish the identity and purity of the compound by comparison with known literature values.

www→ Determine the refractive index of the two fractions collected. Compare the experimental values to those found in the literature for ethyl acetate. If the values are within 0.0010 unit of each other, the fractions can be considered to have the same constitution. Are the values for the two fractions the same? If not, which one deviates the most from the reference data? Attempt to explain the result.

Determine the density (see Chapter 4) of the ester, using material contained in the second fraction. This measurement is nondestructive and the material used may be recovered for use in further tests. Compare your results with those values found in the literature.

Determine the boiling point of the second fraction by the ultramicroboiling-point procedure (see Chapter 4). Compare your result with the literature value. Does this fraction appear to be pure ethyl acetate?

In the next step, disconnect and cool the 3-mL conical vial in an ice water bath for 10 min. *trans*-1,2-Dibenzoylethylene will crystallize from the concentrated solution. Remove the remaining solvent from the distillation vial with a Pasteur filter pipet and place the crystals on a porous clay plate to air dry. The melting point of the crystalline material is obtained by the evacuated capillary method and compared with the literature value.

Reference values of the physical constants are available online and in the *CRC Handbook of Chemistry and Physics.* Submit a copy of the table prepared in your laboratory notebook to the instructor, after first tabulating the experimentally measured values of the physical properties, in addition to those reported in the literature for ethyl acetate (see *acetic acid, ethyl ester* if necessary).

Experiment 3B

Fractional Semimicroscale Distillation: Separation of Hexane and Toluene

Purpose. This experiment effects the separation of a binary liquid mixture composed of liquids having boiling points that are relatively far apart, greater than 30 °C. It will help you develop the skills to operate a semimicrodistillation apparatus so that purifications required in later experiments can be successfully carried out.

Prior Reading

Technique 2: Distillation (pp. 61–64)
Distillation Theory (p. 61 and online) www
Simple Distillation at the Semimicroscale Level (pp. 61–64)
Technique 3: Fractional Semimicroscale Distillation (pp. 64–67)
Chapter 4: Determination of Physical Properties
Ultramicro-Boiling Point (pp. 46–48) www
Refractive Index (online)

DISCUSSION

Hexane and toluene are liquid hydrocarbons that have boiling points approximately 40 °C apart. The liquid–vapor composition curve in Figure 5.6 represents this system; it is apparent that a two-plate distillation should yield nearly pure components. The procedure to be outlined consists of two parts. The first deals with the initial distillation (first plate), which separates the liquid mixture into three separate fractions. The second deals with *redistillation* of the first and third fractions (second plate). Exercising careful technique during the first distillation should provide a fraction rich in the lower boiling component, a middle fraction, and a fraction rich in the higher boiling component. Then careful *redistillation* of these fractions can be expected to complete the separation of the two components and to produce fractions of relatively pure hexane and toluene. The Hickman still used in the microscale laboratory is a simple, short-path column, and, therefore, one would not expect complete separation of the hexane and toluene in one cycle.

COMPONENTS

$CH_3CH_2CH_2CH_2CH_2CH_3$

Hexane

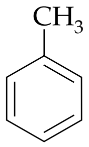

Toluene

EXPERIMENTAL PROCEDURE

Estimated time for the experiment: 2.0 h.

Physical Properties of Components					
Compound	MW	Amount	bp (C°)	*d*	n_D
Hexane	86.18	1.0 mL	69	0.66	1.3751
Toluene	92.15	1.0 mL	111	0.87	1.4961

Reagents and Equipment. Use an automatic delivery pipet to place 1.0 mL of hexane and 1.0 mL of toluene in a clean, dry, stoppered 5-mL conical vial.

Place the vial in a small beaker to prevent tipping. Add a boiling stone (or a magnetic spin vane if desirable) assemble the Hickman still with the thermometer positioned directly down the center of the column (see previous discussion), and mount the system in a sand bath (see Fig. 6.3).

Experimental Conditions. The temperature of the sand bath is raised to 80–90 °C, at a maximum rate of 5 °C/min (>70 °C at 3 °C/min) using a hot plate.

CAUTION: Do not let the temperature of the still rise too rapidly.

Once gentle boiling begins, the heating rate should be lowered to a maximum of 2 °C/min. It is ***absolutely crucial*** that the distillation rate be kept below 100 μL/3min. to achieve the necessary fraction enrichment that will permit good separation during the second stage of the experiment. The distillate is collected in *three fractions* over the temperature ranges (1) 65–85 °C (bath temperature ∼95−110 °C); (2) 85–105 °C (bath temperature ∼140 °C) and (3) 105–110 °C (bath temperature ∼170 °C) in amounts of approximately 800, 400, and 800 μL, respectively. Remove each fraction from the still with a bent-tip Pasteur pipet. Store the liquid condensate (fractions) in clean, dry, 1-dram, screw-cap vials. *Remember to number the vials in order and use an aluminum foil cap liner.*

Characterization of Crude Fractions. For each of the three fractions, record the refractive index. Fraction 1 has been enriched in one of the two components. Which one? Does the refractive index agree with that found in the literature? Fraction 3 has been enriched in the other component. Does the refractive index of that fraction support your first conclusion? If partial enrichment has been achieved, proceed to the second phase of the distillation.

Redistillation of Fraction 1 Redistill fraction 1 in a clean Hickman still with a thermometer arranged as before (Fig. 6.3), using a 3-mL conical vial and the procedure just outlined. Collect an initial fraction over the boiling range 68–71 °C (∼100−200 μL). Remove it from the collar, using the Pasteur pipet, and place it in a 1-dram screw-cap vial.

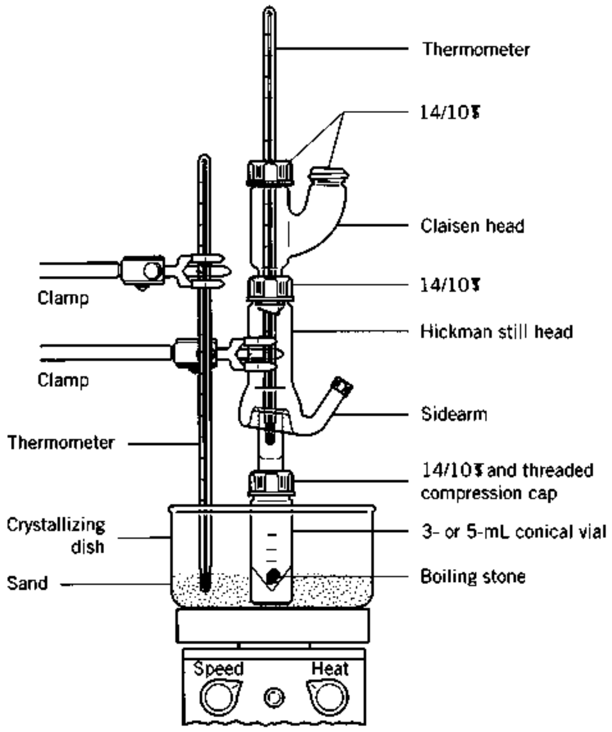

Figure 6.3 Hickman still with Claisen head adapter.

Characterization of Fraction 1 Determine the ultramicro-boiling point and the refractive index of this lower boiling fraction. Compare the experimental values obtained with those of pure hexane reported in the literature.

Redistillation of Fraction 3 Fraction 3 is placed in a clean Hickman still, using a thermometer and a 3-mL conical vial (Fig. 6.3), and redistilled using the procedure outlined. Collect an initial fraction over the boiling range 95–108 °C (~500 μL), and transfer this fraction by Pasteur pipet to a screw-cap vial. Collect a final fraction at 108–110 °C (~250 μL), and transfer the material to a second vial. *This second fraction is the highest boiling fraction to be collected in the three distillations and should be the richest in the high-boiling component*.

Characterization of Fraction 3 Determine the refractive index and boiling point of the second fraction and compare your results with those found in the literature for toluene. *Determine the refractive index and boiling point of pure toluene for comparison purposes.*

Experiment 3C

Fractional Semimicroscale Distillation: Separation of 2-Methylpentane and Cyclohexane Using a Spinning-Band Column

Purpose. The purpose of this experiment is to separate two liquids with boiling points that are relatively similar: less than 20 °C apart, to learn the operation of a high-performance spinning-band distillation column, and to develop the skills for purifying small quantities of liquid mixtures.

Prior Reading

Technique 1: Microscale Separation of Liquid Mixtures by GC (pp. 55–61)
Technique 2: Distillation (pp. 61–64)
    Distillation Theory (p. 61 and online)
    Simple Distillation at the Semimicroscale Level (pp. 61–64)
    Fractional Semimicroscale Distillation (pp. 64–67)
Chapter 4: Determination of Physical Properties
    Ultramicro Boiling Point (pp. 46–48)
    Refractive Index (online)

DISCUSSION

In this experiment, the separation of a 2-mL mixture of 2-methylpentane and cyclohexane using a 2.5-in. spinning-band distillation column is described. The purity of the fractions is determined by gas chromatography and by measurement of the refractive index. Finally, the number of theoretical plates is estimated. You will separate a 50:50 mixture of 2-methylpentane and cyclohexane using the spinning band distillation column shown in Figure 6.4.

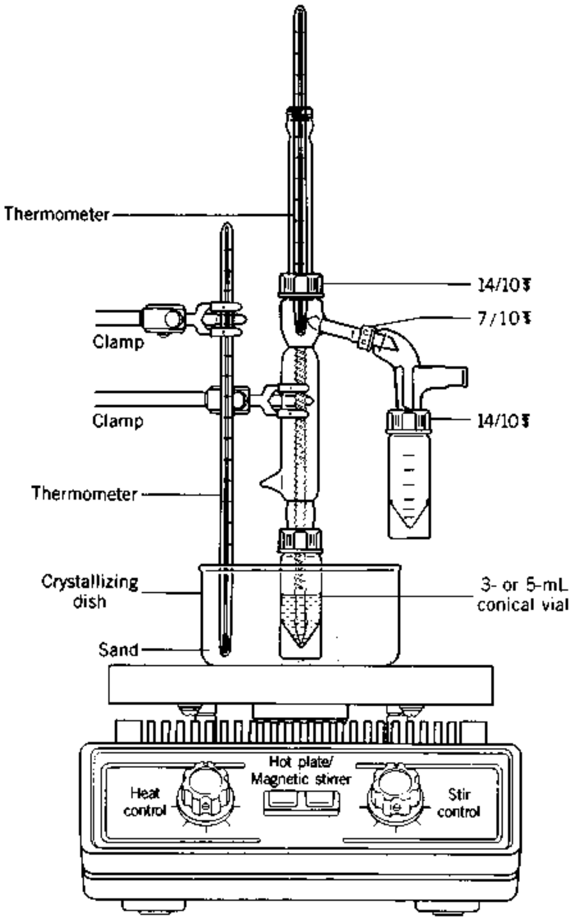
Thermometer
Clamp
Clamp
Thermometer
Crystallizing dish
Sand
14/10 ₸
7/10 ₸
14/10 ₸
3- or 5-mL conical vial
Hot plate/ Magnetic stirrer
Heat control
Stir control

Figure 6.4 Micro spinning-band distillation column (2.5 in.).

COMPONENTS

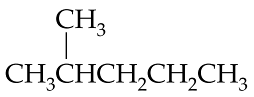

$$\begin{array}{l} CH_3 \\ | \\ CH_3CHCH_2CH_2CH_3 \end{array}$$

2-Methylpentane

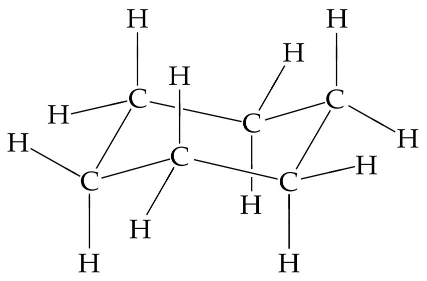

Cyclohexane

EXPERIMENTAL PROCEDURE

Estimated time for the experiment: 3.0 h.

Physical Properties of Components				
Compound	MW	Amount	bp (°C)	n_D
2-Methylpentane	86.18	1.0 mL	60.3	1.3715
Cyclohexane	84.16	1.0 mL	80.7	1.4266

Reagents and Equipment. Assemble the system as shown in Figure 6.4, making sure that the Teflon band is aligned as straight as possible in the column. In particular, the pointed section extending into the vial must be straightened to minimize vibration during spinning of the band.

Place a pipet bulb on the side arm of the collection adapter. This bulb plays an important function in the operation of the column: Attachment of the bulb creates a closed system. Suspension of the thermometer with a septum on the top of the condenser can act to release any buildup of pressure.

CAUTION: The system must be able to vent at the thermometer during operation!

Once the spinning band has been tested and rotates freely, place 1.0 mL of 2-methylpentane and 1.0 mL of cyclohexane in the vial (to be delivered with a Pasteur pipet or an automatic delivery pipet). Reassemble the system and lower the column into the sand bath or copper-tube block. The beveled edge on the air condenser should be rotated 180° from the collection arm.

NOTE. It is important to make an aluminum foil cover for the sand bath; this cover will reflect the heat and hot air away from the collection vial.

Experimental Conditions. Gently heat (copper-tube block, Fig. 3.3W) the vial until boiling occurs. The magnetic stirrer is turned to a low-spin rate when heating commences. When reflux is observed at the base of the column, the magnetic stirrer is adjusted to intermediate spin rate. Once liquid begins to enter the column the spin rate is increased to the maximum (1000–1500 rpm).

www

NOTE. It is absolutely critical that the temperature of the vial be adjusted so that vapors in the column rise very slowly. *It is possible for overheated vapors to be forced through the air condenser.*

When the vapors slowly arrive in the unjacketed section of the column head, the condenser joint acts as a vapor shroud to effectively remove vapors from the receiver-cup area. During this total reflux period, maximum separation of the components is achieved. Once vapor reflux occurs in the head of the column, the system is left for 20–30 min to reach thermal equilibrium. During this period of total reflux, the head thermometer should read about 57–60 °C (at least for most of the equilibration time).

Following the equilibration period, collection of the resolved components may begin. Rotate the air condenser 180° so that the beveled edge is over the collection duct. At this point, manipulation of the pipet bulb allows drainage of the condensate from the side arm. (This procedure is repeated occasionally to continue drainage from the side arm.) Collect six drops ($\sim$0.30 mL). After removing the collection vial, transfer the contents into a covered vial using a Pasteur pipet. Label all fractions. Collect two 0.6-mL fractions (the pipet bulb may be removed during collection of these latter fractions); then turn off the heat and stirring motor, and remove the vial from the sand bath. Transfer the material remaining in the vial, using a Pasteur pipet, to a fourth covered vial.

Characterization of the Fractions. The composition of each of the fractions may be determined by gas chromatographic analysis and measurement of

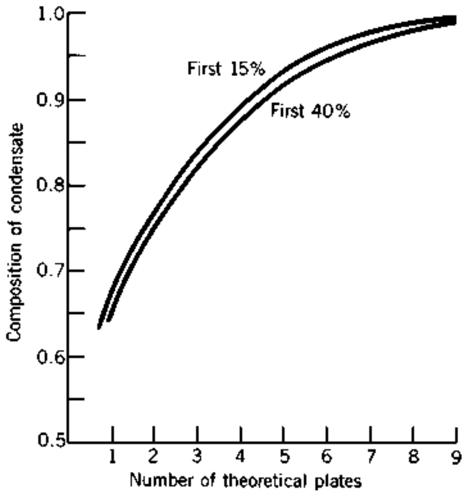

Figure 6.5 Composition of the first 15% and the first 40% of the volume collected in the distillation of a 50% (v/v) mixture of 2-methylpentane and cyclohexane.

the refractive index. A GOW-MAC gas chromatograph should be set up as follows:

Column	DC 710
Injection	10 μL
Temperature	80 °C
Flow rate (He)	55 mL/min
Chart speed	1 cm/min

If we assume that the refractive index is a linear function of the volume fraction, the following relationship gives us the volume fraction of 2-methylpentane in a mixture. The volume fraction is X and the measured refractive index is n_D:

$$X = \frac{1.4266 - n_D}{1.4266 - 1.3715}$$

The curve shown in Figure 6.5 may be used to estimate the number of theoretical plates from the composition (mole fraction) of the *first* 0.30-mL fraction. For example, if the composition of the *first* 0.30 mL is 0.89, we would infer that the system had a resolution equivalent to about four theoretical plates. Note that the number of plates cannot be determined with confidence if the composition is greater than about 0.97. If we really wanted to determine the number of theoretical plates for a system with more than five plates, we could start with a mixture containing only 10 or 20% of the most volatile component (MVC), rather than the 50% used in this example.

Experiment 3D

Fractional Semimicroscale Distillation: The Separation of 2-Methylpentane and Cyclohexane Using a Spinning Band in a Hickman–Hinkle Still

Purpose. In this exercise you will become familiar with a powerful modification of the classic Hickman still: the Hickman-Hinkle spinning band distillation apparatus. This small still is one of the most efficient techniques developed for the purification of small quantities of liquids. You will develop the skills for handling small quantities of liquids and their purification by distillation, and become familiar with these techniques, so that they may be used in the purification of reaction products formed in later experiments.

DISCUSSION

This distillation separates the same two compounds used in Experiment [3C]. The distillate can be analyzed to determine the number of theoretical plates. If careful attention is given to the procedure, the spinning Hickman–Hinkle is capable of more than six theoretical plates.

As in Experiment [3C] the separation of a 2-mL mixture of 2-methylpentane and cyclohexane is achieved. The purity of the fractions can be determined by gas chromatography and by measurement of the refractive index.

COMPONENTS

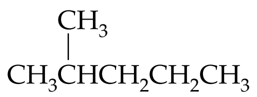

$$CH_3\underset{\displaystyle CH_3}{\overset{|}{C}}HCH_2CH_2CH_3$$

2-Methylpentane

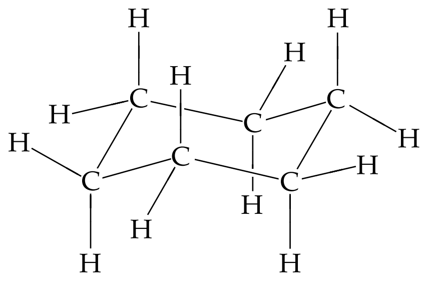

Cyclohexane

EXPERIMENTAL PROCEDURE

Estimated time for the experiment: 3.0 h.

Physical Properties of Components

Compound	MW	Amount	bp (°C)	n_D
2-Methylpentane	86.18	1.0 mL	60.3	1.3715
Cyclohexane	84.16	1.0 mL	80.7	1.4266

Reagents and Equipment. The system is assembled as shown in Figure 3.15 on page 25. In the process, make sure that the Teflon band is aligned as straight as possible in the column. In particular, the pointed section extending into the vial must be straightened to minimize vibration during spinning of the band.

Once the spinning band has been tested and rotates freely, place 1.0 mL of 2-methylpentane and 1.0 mL of cyclohexane in the vial (to be delivered with a Pasteur pipet or an automatic delivery pipet). Reassemble the system and lower the column into the sand bath.

Cover the sand bath with aluminum foil during the distillation to prevent the collar of the still from overheating. It is easier to regulate the temperature of the bath when it is covered. However, for distillations, a more efficient heating technique is the recently developed copper-tube block (Fig. 3.3W).

Experimental Conditions. Gently heat the vial until boiling occurs. When heating commences, turn on the magnetic stirrer at a low setting. When reflux commences at the base of the column the magnetic stirrer is raised to intermediate settings. Once liquid begins to enter the column, the spin rate is increased to the maximum (1000–1500 rpm).

NOTE. It is extremely important that careful temperature control be exercised at this stage so that the condensing vapors ascend the column very slowly.

Vapor-phase enrichment by the most volatile component is limited mainly to this period, as fraction collection commences immediately on arrival of the vapor column at the annular ring. Once condensation occurs, fractions are collected by the same technique used in Experiment [3B]. Characterization of the fractions, however, follows the procedure given in Experiment [3C].

Characterization of the Fractions. The composition of each of the fractions may be determined by gas chromatography, the refractive index, or both. See Experiment [3C] for details.

An alternative approach to the procedures discussed in Experiment [3C] is to establish the fraction volume by weight. The curves shown in Figure 6.5 may again be used to estimate the number of theoretical plates. The volume of the first fraction can be estimated, or determined more accurately by weighing the fraction in a tared screw-cap vial. The composition of this fraction then may be determined and the fraction of the total represented by this portion calculated. If, for example, the first fraction has a volume of 0.4 mL (20% of the total) and has a composition 0.89 by volume of 2-methylpentane, we would infer that the system had a resolution equivalent to about four theoretical plates.

QUESTIONS

6-12. The boiling point of a liquid is affected by several factors. What effect does each of the following conditions have on the boiling point of a given liquid?
- **(a)** The pressure of the atmosphere
- **(b)** Use of an uncalibrated thermometer
- **(c)** Rate of heating of the liquid in a distillation flask

6-13. Calculate the vapor pressure of a solution containing 30 mol% hexane and 70 mol% octane at 90 °C assuming that Raoult's law is obeyed.
Given: vapor pressure of the pure compounds at 90 °C: hexane = 1390 torr, octane = 253 torr.

6-14. In any distillation for maximum efficiency of the column, the distilling flask should be approximately one-half full of liquid. Comment on this fact in terms of (a) a flask that is too full and (b) a flask that is nearly empty.

6-15. Occasionally during a distillation, a solution will foam rather than boil. A way of avoiding this problem is to add a *surfactant* to the solution.
- **(a)** What is a surfactant?
- **(b)** What is the chemical constitution of a surfactant?
- **(c)** How does a surfactant reduce the foaming problem?

6-16. Explain why packed and spinning-band fractional distillation columns are more efficient at separating two liquids with close boiling points than are unpacked columns.

6-17. Explain what effect each of the following mistakes would have had on the simple distillation carried out in this experiment.
- **(a)** You did not add a boiling stone.
- **(b)** You heated the distillation flask at too rapid a rate.

6-18. In the ultramicro-boiling-point determination, why is the boiling point taken just as bubbles cease emerging from the bell?

6-19. Define each of the following terms, which are related to the distillation process:
- **(a)** Distillate
- **(b)** Normal boiling point
- **(c)** Forerun

6-20. How does the refractive index of a liquid vary with temperature? What corrective factor is often used to determine the value at a specific temperature, for example, 20 °C, if the measurement were made at 25 °C?

BIBLIOGRAPHY

General references on distillation:

Lodwig, S. N. *J. Chem. Educ.* **1989,** *66,* 77.

Stichlmair, J. G.; Fair, J. R. *Distillation: Principles and Practice;* Wiley: New York, 1998.

Technique of Organic Chemistry, Vol. IV, *Distillation,* 2nd ed., Perry, E. S.; Weissberger, A., Eds.; Interscience-Wiley: New York, 1967.

Vogel, A. I. *Vogel's Textbook of Practical Organic Chemistry,* 5th ed.; Furnis, B. S., et al. Eds.; Wiley: New York, 1989.

Zubrick, J. W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, 2008.

Solvent Extraction

EXPERIMENT 4

Experiment 4A

Determination of a Partition Coefficient for the System Benzoic Acid, Methylene Chloride, and Water

Purpose. This exercise illustrates the general procedures that are used to determine a partition coefficient at the microscale level. Experience in weighing milligram quantities of materials on an electronic balance, the use of automatic delivery pipets for accurately dispensing microliter quantities of liquids, the transfer of microliter volumes of solutions with the Pasteur filter pipet, and the use of a Vortex mixer, are techniques encountered in this experiment.

Prior Reading

Chapter 3: Experimental Apparatus
    Pasteur Filter Pipet (pp. 36–37)
    Automatic Delivery Pipet (pp. 37–38)
    Weighing of Solids in Milligram Quantities (p. 39)
Technique 4: Solvent Extraction
    Liquid–Liquid Extraction (p. 72)
    Partition Coefficient Calculations (pp. 70–72)
    Drying of the Wet Organic Layer (pp. 80–83)
    Separation of Acids and Bases (pp. 77–79)

DISCUSSION

Solubility. Substances vary greatly in their solubility in various solvents, but based on observations, many of which were made in the very early days of chemical experimentation, a useful principle has evolved that allows the chemist to predict rather accurately the solubility of a particular substance. It is generally true that *a substance tends to dissolve in a solvent that is chemically similar to itself. In other words, like dissolves like.*

Thus, for a particular substance to exhibit solubility in water requires that species to possess some of the characteristics of water. For example, an important class of compounds, the organic alcohols, have the hydroxyl group (—OH) bonded to a hydrocarbon chain or framework (R—OH). The hydroxyl group can be viewed as being effectively one-half a water (HOH) molecule, and it has a similar polarity to that of water. This results from a charge separation arising from the difference in electronegativity between the hydrogen and oxygen and oxygen atoms. The O—H bond, therefore, is considered to have *partial ionic character:*

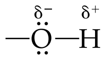

$$-\overset{\delta^-}{\ddot{\underset{..}{O}}}-\overset{\delta^+}{H}$$

Partial ionic character of the hydroxyl group

This *polar,* or partial ionic, character leads to relatively strong hydrogen bond formation between molecules having this entity. Strong hydrogen bonding is evident in molecules that contain a hydrogen atom attached to an oxygen, nitrogen, or fluorine atom, as shown here for the ethanol–water system. This polar nature of a functional group is present when there are sufficient differences in electronegativity between the atoms making up the group:

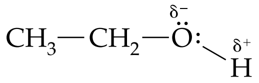

Ethanol

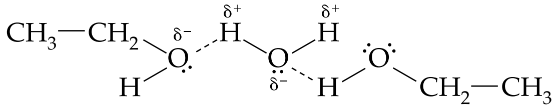

Hydrogen bond formation

In ethanol (CH_3CH_2OH), it is apparent that the hydroxyl end of the molecule is very similar to water (HOH). When ethanol is added to water, therefore, they are miscible in all proportions. That is, ethanol is completely soluble in water and water is completely soluble in ethanol. This solubility results because the attractive forces set up between the two molecules (CH_3CH_2OH and H_2O) are nearly as strong as between two water molecules; however, the attraction in the first case is somewhat weakened by the presence of the nonpolar alkyl ethyl group, CH_3CH_2—. Hydrocarbon groups attract each other only weakly, as evidenced by their low melting and boiling points. Three examples of the contrast in boiling points between compounds of different structure but similar molecular weight are summarized in Table 6.1. Clearly, those molecules that attract each other weakly have lower boiling points.

When we compare the water solubility of ethanol (a two-carbon [C_2] alcohol that, as we have seen, is completely miscible with water) with that of octanol (a straight-chain eight-carbon [C_8] alcohol), we find that the solubility of octanol is less than 2% in water. Why the difference in solubilities between these two alcohols? The answer lies in the fact that the dominant structural feature of octanol has become the non-polar nature of its alkyl group:

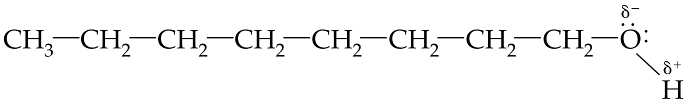

Octanol

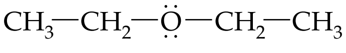

$$CH_3-CH_2-\ddot{\underset{..}{O}}-CH_2-CH_3$$

Diethyl ether

As the bulk of the hydrocarbon section of the alcohol molecule increases, the intramolecular attraction between the polar hydroxyl groups of the alcohol and the water molecules is no longer sufficiently strong to overcome the *hydrophobic character* (lack of attraction to water) of the nonpolar hydrocarbon section of the alcohol. On the other hand, octanol has a large nonpolar hydrocarbon group as its dominant structural feature. We might, therefore, expect octanol to exhibit enhanced solubility in less polar solvents. In fact, octanol is found to be completely miscible with diethyl ether. Ethers are solvents of weak polarity. Since the nonpolar characteristics are significant in both molecules, mutual solubility is observed. It has been empirically demonstrated that, in general, if a compound has both polar and nonpolar groups present in its structure, those compounds having five or more carbon atoms in the hydrocarbon portion of the molecule will be more

Table 6.1 **Comparison of Boiling Point Data**

Name	Formula	MW	bp (°C)
Ethanol	CH_3CH_2OH	46	78.3
Propane	$CH_3CH_2CH_3$	44	−42.2
Methyl acetate	$CH_3CO_2CH_3$	74	54
Diethyl ether	$(CH_3CH_2)_2O$	74	34.6
Ethylene	$CH_2{=}CH_2$	28	−102
Methylamine	CH_3NH_2	31	−6

Table 6.2 **Water Solubility of Alcohols**

Name	Formula	Solubility (g/100 g H_2O)
Pentanol	$CH_3(CH_2)_3CH_2OH$	4.0
2-Pentanol	$CH_3(CH_2)_2CH(OH)CH_3$	4.9
2-Methyl-2-butanol	$(CH_3)_2C(OH)CH_2CH_3$	12.5

Note. Data at 20 °C.

soluble in nonpolar solvents, such as pentane, diethyl ether, or methylene chloride. Figure 5.13, on page 69, summarizes the solubilities of a number of straight-chain alcohols, carboxylic acids, and hydrocarbons in water. As expected, those compounds with more than 5 carbon atoms are shown to possess solubilities similar to those of the hydrocarbons.

Several additional relationships between solubility and structure have been observed and are pertinent to the discussion.

1. Branched-chain compounds have greater water solubility than their straight-chain counterparts, as illustrated in Table 6.2 with a series of alcohols.

2. The presence of more than one polar group in a compound will increase that compound's solubility in water and decrease its solubility in nonpolar solvents. For example, high molecular weight sugars, such as cellobiose, which contain multiple hydroxyl and/or acetal groups, are water soluble and ether insoluble; cholesterol (C_{27}), which possesses only a single hydroxyl group, is water insoluble and ether soluble:

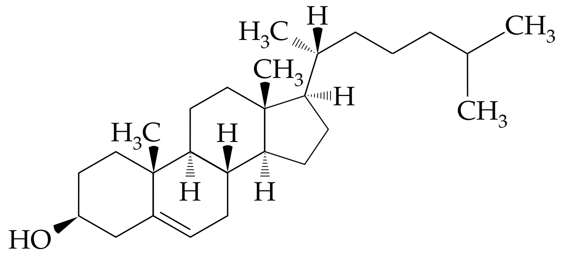

Cholesterol

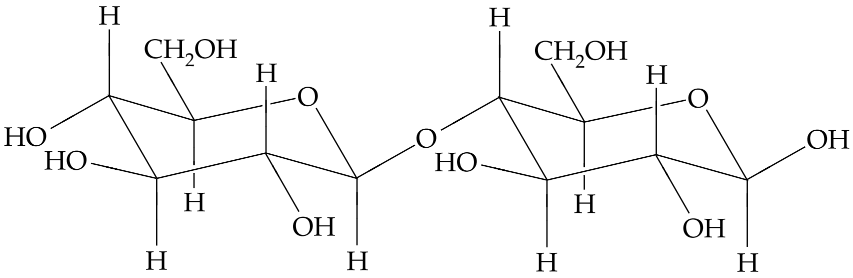

Cellobiose

3. The presence of a chlorine atom, even though it lends some partial ionic character to the covalent C—Cl bond, does not normally impart water solubility to a compound. In fact, such compounds as methylene chloride (CH_2Cl_2), chloroform ($CHCl_3$), and carbon tetrachloride (CCl_4) have long been used as solvents for the extraction of aqueous solutions. It should be noted that use of the latter two solvents is no longer recommended, unless strict safety precautions are exercised, because of their potential carcinogenic nature.

Table 6.3 Water Solubility of Amines

Name	Formula	Solubility (g/100 g H_2O)
Ethylamine	$CH_3CH_2NH_2$	∞
Diethylamine	$(CH_3CH_2)_2NH$	∞
Trimethylamine	$(CH_3)_3N$	91
Triethylamine	$(CH_3CH_2)_3N$	14
Aniline	$C_6H_5NH_2$	3.7
p-Phenylenediamine	$H_2NC_6H_4NH_2$	3.8

Note. Data at 25 °C.

4. Most functional groups that are capable of forming a hydrogen bond with water, if it constitutes the dominant structural feature of a molecule, will impart increased water solubility characteristics to a substance (the five-carbon rule obviously applies here in determining just what is a dominant feature). For example, certain alkyl amines (organic relatives of ammonia) might be expected to have significant water solubility. This finding is indeed the case, and the water-solubility data for a series of amines is summarized in Table 6.3.

The solubility characteristics of any given compound will uniquely govern that substance's distribution (*partition*) between the phases of two immiscible solvents (in which the material has been dissolved) when these phases are intimately mixed. In this experiment we determine the partition coefficient (distribution coefficient) of benzoic acid between two immiscible solvents, methylene chloride and water.

COMPONENTS

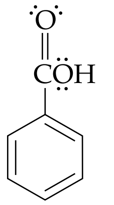

Benzoic acid

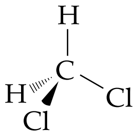

Methylene chloride

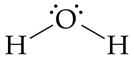

Water

EXPERIMENTAL PROCEDURE

Estimated time of experiment: 1.5 h.

Physical Properties of Components

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Benzoic acid	122.13	50 mg	0.41	122		
Methylene chloride		1.20 mL			40	1.33
Water		600 μL			100	1.00

Equipment Setup and Addition of Reagents. Weigh and add to a 5.0-mL conical vial fitted with a screw cap, 50 mg (0.41 mmol) of benzoic acid. Now add 600 μL of methylene chloride followed by 600 μL of water.

The solvents are transferred to the vial with automatic delivery pipets (use a different pipet or pipet tip for each solvent). *The methylene chloride addition should be carried out in the* ***hood.***

HOOD

Procedure for Establishing Equilibrium Distribution. Cap the vial and shake (or use a Vortex mixer) until the benzoic acid dissolves and the two phases have been thoroughly mixed. Vent the vial (by releasing the Cap-seal) and then allow the two layers to separate.

Carefully draw the lower methylene chloride layer into a Pasteur filter pipet and transfer it to a 5-mL conical vial containing 100 mg of anhydrous, magnesium sulfate. If the amount of the methylene chloride layer is insufficient to properly transfer into the 5-mL conical vial, carefully add more so that a proper transfer can occur. Once complete, recap the vial.

NOTE. If the volume of the methylene chloride layer is so large that it cannot be transferred completely in one operation, a second transfer may be required. Be careful not to over-fill the pipet to insure that solvent does not come in contact with the rubber bulb.The technique of removing the last traces of water from the methylene chloride solution is often referred to as drying the solution. *It can involve any one of a number of insoluble anhydrous salts, which convert the moisture retained in the organic phase to water of crystallization. In this case, we are using sodium sulfate.*

Isolation of the Benzoic Acid. After drying the methylene chloride solution for 10–15 min, transfer the anhydrous solution to a previously **tared** vial (the term **tare** means to preweigh the empty vial) using a Pasteur filter pipet (the use of the filter pipet is a convenient way of separating the solid hydrated sodium sulfate from the dried solution). Rinse the sodium sulfate with an additional 600 μL of methylene chloride and combine the rinse with the solution in the tared 5.0-mL conical vial. Evaporate the solvent under a gentle stream of nitrogen gas in a warm sand bath in the **hood.** (It is important to warm the solution while evaporating the solvent; otherwise the heat of vaporization will rapidly cool the solution. In this latter case, as the cold, solid acid precipitates from the saturated solution, moisture will condense from the air entrained in the evaporation process, and contaminate the surface of the recovered material.)

HOOD

NOTE. If a hot sand bath is used, a boiling stone is placed in the vial before it is tared. The boiling stone will help to avoid explosive, sudden boiling of the solvent when the vial is placed in the sand bath.

Weight Data and Calculations. Weigh the vial and determine the weight of benzoic acid that remains following removal of the methylene chloride. Break up the hard cake of precipitated benzoic acid with a microspatula and *briefly* reheat the vial and contents in a sand bath to remove the last traces of solvent and any water that remains in the system. Cool and reweigh. Repeat this operation until a constant weight is obtained. This weight represents the benzoic acid that dissolved in the *methylene chloride layer*.

The *original weight* of benzoic acid used minus the amount of benzoic acid recovered in the methylene chloride layer equals the weight of the benzoic acid that dissolved and still remains in the *water layer*.

Since equal volumes of both solvents were used, the partition coefficient may be simply determined from the ratio of the weight of benzoic acid in the methylene chloride solvent to the weight of benzoic acid in the water layer.

Calculate the partition coefficient for benzoic acid in the solvent pair used in this exercise.

Experiment 4B

Solvent Extraction I: The System; Benzoic Acid, Methylene Chloride, and 10% Sodium Bicarbonate Solution; An Example of Acid–Base Extraction Techniques

Purpose. This exercise illustrates an extensively used extraction technique in which a reversible reaction is employed to alter the solubility characteristics of the substance of interest.

Prior Reading

- *Chapter 3:* Experimental Apparatus
    - Pasteur Filter Pipet (pp. 36–37)
    - Automatic Delivery Pipet (pp. 37–38)
    - Weighing of Solids in Milligram Quantities (p. 39)
- *Technique 4:* Solvent Extraction
    - Liquid–Liquid Extraction (p. 72)
    - Partition Coefficient Calculations (pp. 70–72)
    - Drying of the Wet Organic Layer (pp. 80–83)
    - Separation of Acids and Bases (pp. 77–79)

REACTION

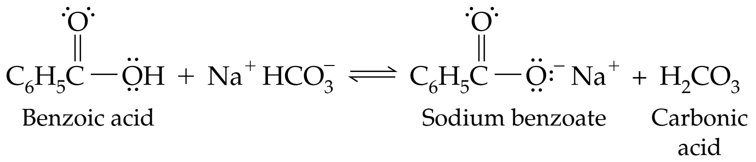

$$C_6H_5C(=O)-OH + Na^+ HCO_3^- \rightleftharpoons C_6H_5C(=O)-O^- Na^+ + H_2CO_3$$

Benzoic acid

Sodium benzoate

Carbonic acid

$$H_2CO_3 \rightleftharpoons CO_2 + H_2O$$

DISCUSSION

Benzoic acid reacts readily with sodium bicarbonate to form sodium benzoate, carbon dioxide, and water. The sodium salt of benzoic acid has highly ionic characteristics and thus, unlike the free acid, the salt is very soluble in water and nearly insoluble in methylene chloride. This salt is characterized by a full ionic bond between the carboxylic acid group of the acid and the sodium ion. It is, therefore, a new substance exhibiting many of the solubility properties commonly associated with inorganic ionic salts.

EXPERIMENTAL PROCEDURE

Estimated time of experiment: 1.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Benzoic acid	122.13	50 mg	0.41	122		
Methylene chloride		1.20 mL			40	1.33
Sodium bicarbonate (10% solution)		600 μL				

Procedure for Establishing Distribution. Repeat the identical procedures carried out in Experiment [4A], but replace the 600 μL of water with 600 μL of 10% sodium bicarbonate solution. A good estimate of the efficiency of the conversion of benzoic acid to the sodium salt of the acid, which because of its ionic character is found almost exclusively in the aqueous phase, can be made by recovering any unreacted acid from the methylene chloride layer and using the distribution coefficient established in Experiment [4A]. Also, be sure to obtain a melting point of any recovered residue (assumed above to be benzoic acid) from the organic phase, since contamination of free acid by the acid salt can be detected by this measurement. Sodium benzoate has a melting point above 300 °C, whereas benzoic acid melts near 122 °C.

Test for a Carboxylic Acid. As illustrated in the above reaction, when a carboxylic acid comes in contact with a solution containing bicarbonate ion, carbon dioxide is generated. Once saturation of the solution by carbon dioxide occurs, bubbles of carbon dioxide gas are observed to form in the liquid phase. This effervescence may be used as a qualitative test for the presence of the carboxylic acid functional group in an unknown substance.

Place 1–2 mL of 10% sodium or potassium bicarbonate on a small watch glass. Add the pure acid, one drop from a Pasteur pipet if the sample is a liquid (~5 mg if it is a solid), to the bicarbonate solution. Evolution of bubbles (CO_2) from the mixture indicate the presence of an acid.

Perform the above test for the presence of carboxyl groups on several organic acids, such as acetic, benzoic, propanoic, or chloroacetic acid.

Solvent Extraction II: A Three-Component Mixture; An Example of the Separation of an Acid, a Base, and a Neutral Substance

Experiment 4C

Purpose. This exercise investigates how solvent extraction techniques can be applied effectively to problems that require the separation of mixtures of organic acids, bases, and neutral compounds in the research or industrial laboratory.

Prior Reading

Chapter 3: Experimental Apparatus
  Pasteur Filter Pipet (pp. 36–37)
  Automatic Delivery Pipet (pp. 37–38)
  Weighing of Solids in Milligram Quantities (p. 39)
Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
  Drying of the Wet Organic Layer (pp. 80–83)
  Separation of Acids and Bases (pp. 77–79)
  Salting Out (p. 79)

DISCUSSION

As implied in the discussions of Experiments [4A] and [4B], the solubility characteristics of organic acids in water can be shown to be highly dependent on the pH of the solution. By extending this extraction approach to include organic bases, it has been possible to develop a general procedure for the separation of mixtures of organic acids, bases, and neutral substances.

NOTE. Refer to Technique 4, p. 78–79 for a flowchart outlining the procedure.

The components of the mixture to be separated in this experiment are benzoic acid, ethyl 4-aminobenzoate (a base), and 9-fluorenone (a neutral compound, which may be prepared in Experiment [33A]).

COMPONENTS

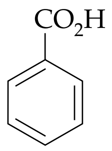

Benzoic acid

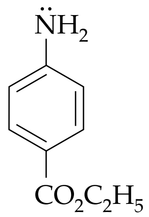

Ethyl 4-aminobenzoate

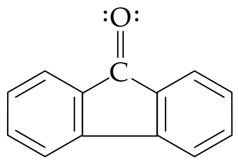

9-Fluorenone

EXPERIMENTAL PROCEDURE

Estimated time of experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Benzoic acid	122.1	50 mg	0.41	122		
Ethyl 4-amino-benzoate	165.19	50 mg	0.31	89		
9-Fluorenone	180.22	50 mg	0.27	84		
Diethyl ether	74.12	4 mL			35	0.7184
3 M HCl		4 mL				
3 M NaOH		4 mL				
6 M HCl						
6 M NaOH						

NOTE. In carrying out the separation, you should keep a record or flowchart of your procedure (as suggested in the prior reading assignment) in your laboratory notebook. You should also be particularly careful to label all flasks.

HOOD

Reagents and Equipment. Weigh and add to a stoppered or capped 15-mL centrifuge tube the following: 50 mg (0.41 mmol) of benzoic acid, 50 mg (0.31 mmol) of ethyl 4-aminobenzoate, and 50 mg (0.27 mmol) of 9-fluorenone. Now, in the **hood,** add 4 mL of diethyl ether using a 10-mL graduated cylinder for the transfer. The solids may be dissolved by either stirring with a glass rod or mixing on a Vortex mixer (capped vial).

Separation of the Basic Component. Cool the solution in an ice bath. Now, using a calibrated Pasteur pipet, add 2 mL of 3 M HCl dropwise to the cooled solution with swirling. Cap and thoroughly mix the resulting two-phase system for several minutes (a Vortex mixer works well). Vent carefully and after the layers have separated remove the bottom (aqueous) layer using a Pasteur filter pipet and transfer this phase to a **labeled,** 10-mL Erlenmeyer flask.

Repeat this step with an additional 2 mL of the 3 M acid solution. As before, transfer the aqueous layer to the same labeled Erlenmeyer flask. Stopper or cap this flask. *Save the ether solution.* The aqueous acidic solution is to be used in the next step.

NOTE. A small amount of crystalline material may form at the interface between the layers. A second extraction generally dissolves this material.

Isolation of Ethyl 4-Aminobenzoate: The Basic Component. To the aqueous acidic solution, separated and set aside in the previous step, add 6 M NaOH dropwise until the solution is distinctly alkaline to litmus paper. Cool the flask in an ice bath for about 10–15 min. Collect the solid precipitate that forms in the basic solution by reduced-pressure filtration using a Hirsch funnel. Wash the precipitate with two 1-mL portions of distilled water. Air-dry the washed microcrystals by spreading them on a clay plate, filter paper, or in a vacuum drying oven. Weigh the material and calculate the percent recovery. Obtain the melting point of the dry ethyl 4-aminobenzoate and compare your result to the literature value. This material is used as a topical anesthetic.

Separation of the Acidic Component. Add 2 mL of 3 M NaOH to the ether solution that was set aside earlier in the experiment. At this point, if necessary, add additional ether (~1–2 mL) so that the volume of the organic layer is at least equal to, or somewhat larger than, that of the aqueous phase. This adjustment in volume should allow an efficient distribution to take place when the two phases are mixed. Then carry out the extraction as before, allow the layers to separate, and finally transfer the bottom aqueous basic layer to a **labeled,** 10-mL Erlenmeyer flask.

Repeat this routine and again remove the aqueous layer and transfer it to the same Erlenmeyer flask. Stopper this flask (containing the extracted aqueous basic phase) and set it aside for later use.

Separation of the Neutral Component. Wash (extract) the remaining ether solution contained in the centrifuge tube with two 1-mL portions of water. Separate the lower aqueous layer in each sequence. *Save the aqueous wash layer temporarily; it will be discarded at the very end of the experiment. (It is good practice to never discard any layer until you have recovered or accounted for all of the material.)* Now add about 300 mg of anhydrous granular sodium sulfate to the wet ether (ether saturated with water) solution. Cap the tube and set it aside while working up the alkaline extraction solution. This procedure will allow sufficient time for the traces of moisture to be removed from the ether solution by hydration of the insoluble drying agent. If the sodium sulfate initially forms large clumps, you may add a further quantity of the anhydrous salt.

Isolation of Benzoic Acid: The Acidic Component. Add 6 M HCl dropwise to the aqueous alkaline solution, which was separated and set aside earlier, until the solution becomes distinctly acidic to litmus paper. Then cool the flask in an ice bath for about 10 min. If only a small amount (10–25 mg) of precipitate is obtained on acidification, add a small amount of a saturated aqueous solution of sodium chloride (salting out effect; see Prior Reading assignment) to help promote further precipitation of the benzoic acid. Collect the precipitated benzoic acid by filtration under reduced pressure using a Hirsch funnel. Wash (rinse) the filter cake (precipitated acid) with two 1-mL portions of cold distilled water. Dry the solid product using one of the techniques described earlier for ethyl 4-aminobenzoate. Weigh the dry benzoic acid and calculate your percent recovery.

Obtain a melting point of this material and compare your result to the literature value.

Isolation of 9-Fluorenone: The Neutral Component. Use a Pasteur filter pipet to transfer the dried ether solution collected earlier to a tared 10-mL Erlenmeyer flask containing a boiling stone. Rinse the drying agent with an additional 1 mL of ether and combine the ether wash with the anhydrous organic phase.

HOOD

Concentrate the ether solution on a warm sand bath using a *slow* stream of nitrogen gas in the **hood.** Obtain the weight of the residue (9-fluorenone) after removal of the solvent and calculate the percent recovery. Obtain a melting point of the material and compare your result to the literature value.

QUESTIONS

6-21. **(a)** Explain why diethyl ether would be expected to be a satisfactory solvent for the straight-chain hydrocarbons hexane and heptane. **(b)** Explain why *t*-butyl methyl ether would not be expected to be an ideal solvent for the polyhydroxylated carbocycles glucose and fructose.

6-22. The solubility of *p*-dibromobenzene in benzene is 80 μg/100 μL at 25 °C. Would you predict the solubility of this compound to be greater, less, or approximately the same in acetone solvent at this temperature? Explain.

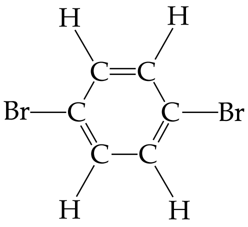

p-Dibromobenzene

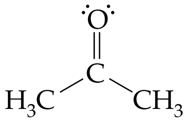

Acetone

6-23. **(a)** Each of the solvents listed below are used in experiments in this text to extract organic compounds from aqueous solutions.

(i) Methylene chloride
(ii) Pentane
(iii) Toluene
(iv) Diethyl ether

Will the organic phase be the upper or lower layer when each of these solvents is mixed with water? Explain your answer for each case. **(b)** If you placed an ice cube in each of the solvents i–iv listed above in (a), would you expect an ice cube placed in each to float or sink? Explain your answer for each case.

6-24. A 36-mg sample of an organic compound (MW = 84) is dissolved in 10 mL of water. This aqueous solution is extracted with 5.0 mL of hexane. Separation and analysis of the aqueous phase shows that it now contains 12 mg of the organic compound. Calculate the partition coefficient for the compound.

6-25. A qualitative method often used to determine whether an organic compound contains oxygen is to test its solubility in concentrated sulfuric acid. Almost all oxygen-containing compounds are soluble in this acid. Explain.

6-26. In the discussion of multiple extractions (p. 71), it was suggested that in the example given you might extend the relationship to the next step by using one-third of the total quantity of the ether solvent in three portions. The reason for increasing the number of extractions was to determine whether this expansion would increase the efficiency of the process even further. To determine if this next step is worth the effort, perform the calculations for the extraction of 100 mg of P in 300 μL of water with three 100-μL portions of ether. Assume the partition coefficient is 3.5 (as before).

(a) Compare the amounts of P extracted from the water layer using one, two, or three extractions.
(b) Do you think that the additional amount of P extracted from the water layer using three extractions is justified? Might it be justified if P were valuable and you were working on the industrial scale of 100 kg of P in 3000 L of water?

BIBLIOGRAPHY

Lo, T. C.; Baird, M. H. I.; Hanson, C., Eds. *Handbook of Solvent Extraction;* Krieger: Melbourne, FL, 1991.

Rydberg, J.; Cox, M.; Musikas, C.; Choppin, G. R., Eds., *Solvent Extraction Principles and Practices,* 2nd ed.; C.H.I.P.S.: Weimar, Texas, 2004.

Shriner, R. L.; Hermann, C. K. F.; Morrill, T. C.; Curtin, D. Y.; Fuson, R. C. *The Systematic Identification of Organic Compounds*, 8th ed.; Wiley: New York, 2003.

Vogel, A. I. *Vogel's Textbook of Practical Organic Chemistry*, 5th ed.; Furnis, B. S., et al. Eds.; Wiley: New York, 1989.

Zubrick, J. W. *The Organic Chem Lab Survival Manual*, 7th ed.; Wiley: New York, 2008.

EXPERIMENT 5

Reduction of Ketones Using a Metal Hydride Reagent: Cyclohexanol and cis- *and* trans-4-tert-*Butylcyclohexanol*

Common name: cyclohexanol
CA number: [108-95-0]
CA name as indexed: cyclohexanol

Common name: 4-*tert*-butylcyclohexanol
CA number: [98-52-2]
CA name as indexed: cyclohexanol, 4-(1,1-dimethylethyl)-

Purpose. The reduction of a ketone carbonyl to the corresponding alcohol is carried out using sodium borohydride, a commercially available metal-hydride reducing agent. The alcoholic reaction products are isolated by extraction techniques and purified by preparative gas chromatography. Cis and trans diastereomers are formed in the reduction of the 4-*tert*-butylcyclohexanone. These diastereomeric products can be separated during the preparative GC isolation. The stereochemistry of the structures can be deduced, once the mixture is separated into its pure components, using either NMR or IR spectroscopy.

Prior Reading

- *Technique 1:* Gas Chromatography (pp. 55–61)
- *Technique 4:* Solvent Extraction
    - Liquid–Liquid Extraction (p. 72)
    - Drying of a Wet Organic Layer (pp. 80–83)
    - Concentration of Solutions (pp. 101–104)
- *Technique 6:* Thin-Layer Chromatography (pp. 97–99)
- *Chapter 8:* Infrared Spectroscopy (pp. 539–554)
    - Nuclear Magnetic Resonance Spectroscopy (pp. 561–587)

REACTION (EXPERIMENT [5A])

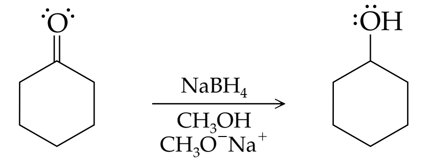

Cyclohexanone Cyclohexanol

DISCUSSION

An important route for the synthesis of primary and secondary alcohols is the reduction of aldehydes and ketones, respectively. Reduction involves the addition of the equivalent of molecular hydrogen H—H to the carbonyl group (C═O).

A variety of pathways have been discovered to accomplish this conversion, but the most common method used in the research laboratory involves complex metal-hydride reagents. Two reagents that enjoy wide application are lithium aluminum hydride ($LiAlH_4$) and sodium borohydride ($NaBH_4$).

Lithium aluminum hydride is a powerful reducing agent that reacts not only with aldehydes and ketones, but with many other carbonyl containing functional groups as well. Because the first of the four deliverable hydrides available is the most reactive hydride, it will attack esters, lactones, carboxylic acids, anhydrides, and amides. It also reduces noncarbonyl systems, such as, alkyl halides, alkyl azides, alkyl isocyanates, and nitriles. Note that $LiAlH_4$ can be used safely only in aprotic solvents (a solvent that does not contain an ionizable [acidic] proton), such as diethyl ether or tetrahydrofuran (THF). In protic solvents, lithium aluminum hydride reacts *violently* with the acidic hydrogen of the solvent to rapidly generate hydrogen gas:

$$LiAlH_4 + 4\ CH_3OH \rightarrow LiAl(OCH_3)_4 + H_2$$

CAUTION: The hydrogen gas often ignites. This particular hydride reagent should not be used unless specific instructions are made available for its proper use under anhydrous conditions.

Sodium borohydride is a much more selective reducing reagent partly due to the fact that the fourth hydride, not first as is the case with lithium aluminum hydride, is the most reactive, and when used in excess, it is a much milder reagent than $LiAlH_4$. For this reason sodium borohydride is usually used for the reduction of aldehydes and ketones. It does not react with the vast majority of the less reactive organic functional groups, such as C═C, C≡C, nitro, cyano, and even some carbonyl-containing systems, such as, amides and carboxylic acids. Sodium borohydride does react at an appreciable rate with water, but only slowly with aqueous alkaline solution (no available protons), methanol, α, β-unsaturated ketones, and esters. For small-scale reactions an excess of reagent is generally used to compensate for the amount of borohydride that reacts with the protic solvent (methanol). This approach is preferred to that of using a solvent in which the sodium borohydride is less soluble (it is insoluble in ether), because the reaction is driven more rapidly to completion under the former conditions. On the other hand, sodium borohydride can react rapidly with strong acids to generate hydrogen gas. This reaction may be used to advantage as a source of in situ hydrogen for the reduction of C═C bonds (see Experiment [12]). The relatively high cost of the metal hydride reducing agents is offset by their low molecular weight (more moles per gram) and the fact that 1 mol of reducing agent reduces 4 mol of aldehyde or ketone.

The key step in the reduction of a carbonyl group by sodium borohydride is the transfer of a hydride ion (:H^-) from boron to the carbon atom of the polarized carbonyl group:

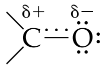

Table 6.4 Reduction of 4-***tert***-Butylcyclohexanone

Reagent	Trans (%)	Cis (%)
Sodium borohydride	80	20
Lithium aluminum hydride	92	8
Lithium tri-*sec*-butylborohydride	7	93

In the reaction, the electron-rich hydride ion is acting as a *nucleophile* (nucleus-seeking), which attacks the *electrophilic* (electron-seeking) carbon atom of the carbonyl group:

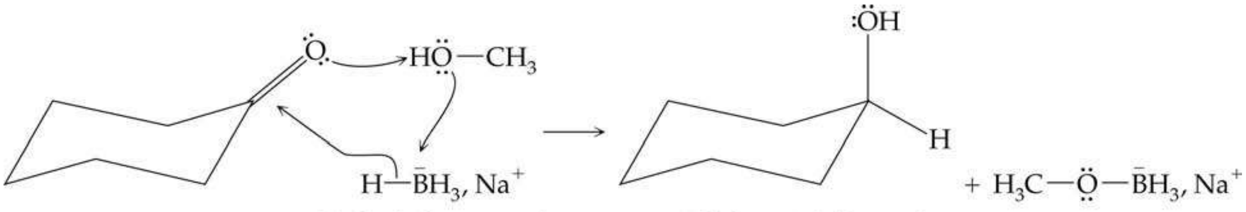

The hydride may make an equatorial (e) or axial (a) attack, depending on steric factors.

The overall reduction process requires two hydrogen atoms, but only one comes directly from the borohydride reagent. The other hydrogen atom is derived from the protic solvent (methanol).

In the 4-*tert*-butylcyclohexanone example, the steric environment is different on either face of the carbonyl group. In this case, the hydride reducing agent attacks more rapidly from the axial direction, and thus the equatorial alcohol (axial H) is the major product. This reaction pathway is preferred with the relatively small sodium borohydride and lithium aluminum hydride reagents. When one stereoisomeric product is preferentially formed, the reaction is called a *stereoselective* reaction.

Sterically large hydride reducing reagents, such as lithium tri-*sec*-butylborohydride, are forced to make an equatorial attack, due to steric factors (these hydride reagents run into the 1,3-diaxial hydrogen atoms), and thus, the cis isomer of 4-*tert*-butylcyclohexanol is the major product. The data are summarized in Table 6.4, which relates the stereochemistry of the reduction to the metal hydride reagent used.

Cyclohexanol

Experiment 5A

The reaction is shown above.

EXPERIMENTAL PROCEDURE

Estimated time for the experiment: 1.5 h.

For the GC analysis, 15 min per student.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Cyclohexanone	98.15	100 μL	0.97	156	0.95	1.4507
Methanol	32.04	250 μL	10.3	65	0.79	1.3288
Sodium borohydride reducing solution		300 μL				

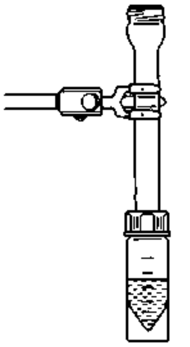
Cyclohexanone, 100 μL + CH_3OH, 250 μL + $NaBH_4$ solution, 300 μL

Reagents and Equipment. With the aid of an automatic delivery pipet, place 100 μL (95 mg, 0.97 mmol) of cyclohexanone in an oven-dried and tared (preweighed) 5.0-mL conical vial equipped with an air condenser. Now add 250 μL of methanol and gently stir the contents in the vial using a glass stirring rod (or a magnetic spin vane if desirable) to obtain a homogeneous solution (←). In the **hood,** add 300 μL of sodium borohydride reducing solution dropwise, with stirring, to the solution of the ketone.

HOOD

NOTE. The cyclohexanone, methanol, and sodium borohydride solutions are dispensed using automatic delivery pipets. Weigh the cyclohexanone after delivery to get an accurate weight for the yield calculations.

The stock reducing solution should be prepared just prior to conducting the experiment.

INSTRUCTOR PREPARATION. In a 10-mL Erlenmeyer flask place 50 mg of anhydrous sodium methoxide and 2.5 mL of methanol. To this solution add 100 mg of sodium borohydride. Stopper the flask tightly and swirl the contents gently to dissolve the solid phase (100 μL of this solution provides ~2.0 mg of $NaOCH_3$ and 4.0 mg of $NaBH_4$).

NOTE. Test for activity of the reducing solution: Add 1–2 drops of the freshly prepared reducing solution to about 200 μL of concentrated hydrochloric acid. Generation of hydrogen gas bubbles is a positive test.

Reaction Conditions. After allowing the resulting solution to stand 10 min at room temperature, begin monitoring the resulting solution by TLC. Use as a solvent system ethyl acetate:hexane (1:4), the R_f of cyclohexanol is 0.4 (R_f of cyclohexanone is 0.6) when stained with a solution of *p*-anisaldehyde (135 mL ethanol, 5 mL H_2SO_4, 1.5 mL glacial acetic acid, and 3.7 mL *p*-anisaldehyde).

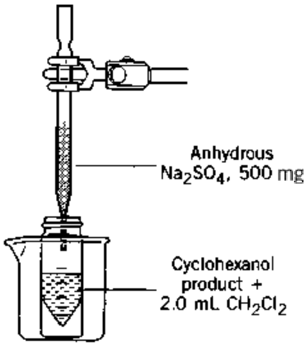

Isolation of Product. When the reaction is complete as judged by TLC, use a calibrated Pasteur pipet to add dropwise 1.0 mL of cold dilute hydrochloric acid (0.1 M HCl). Extract the aqueous mixture with three 0.5-mL portions of methylene chloride. Upon each addition of methylene chloride, cap the vial, shake it gently, and then carefully vent it by loosening the cap (a Vortex mixer may be used if available). After the layers have separated, remove the bottom methylene chloride layer using a Pasteur filter pipet and transfer it to a Pasteur filter pipet containing about 500 mg of anhydrous sodium sulfate.

Collect the dried eluate in a tared 5.0-mL conical vial containing a boiling stone. Use an additional 0.5 mL of methylene chloride to rinse the sodium sulfate and collect the rinse in the same conical vial (←).

HOOD

An additional rinse of the sodium sulfate may be made if desired. Remove the methylene chloride solvent by careful evaporation in the **hood** by gentle warming in a sand bath (constantly agitate the surface of the solution with a microspatula to prevent superheating and subsequent boilover). *In this instance, do not use a stream of nitrogen gas to hasten the evaporation. The volatility of the product alcohol is such that a substantial loss of product will occur if this technique is used.*

Purification and Characterization. The crude cyclohexanol reaction product remaining after evaporation of the methylene chloride solvent is usually of sufficient purity for direct characterization.

Determine the weight of the liquid residue and calculate the percent crude yield. Determine the refractive index (3 μL, optional) and boiling point (4 μL) of the material. Compare your values with data given in the literature.

Obtain the infrared spectrum of the crude (dry) cyclohexanol product by the capillary film sampling technique. Compare the spectrum of the starting ketone in Figure 6.6 to that of your isolated material. Is there evidence of the unreacted starting material in your product? The spectrum of cyclohexanol crude product is shown in Figure 6.7.

NOTE. Most of the infrared spectra referred to in the experimental analysis sections are derived by Fourier transform and plotted on a slightly different scale than the other spectra presented in the text. These spectra utilize a 12.5-cm^{-1}/division format below 2000 cm^{-1} and undergo a 2:1 compression above 2000 cm^{-1} (25 cm^{-1}/division).

Infrared Analysis: A Comparison Reactant and Product. The key absorption bands to examine in the spectrum of cyclohexanone occur at 3420, 3000–2850, 1715, and 1425 cm^{-1}. The lack of significant absorption between 3100–3000 and 1400–1350 cm^{-1} also should be noted. The sharp weak band at 3420 cm^{-1} is not a fundamental vibration (not O—H or N—H stretching), but arises from the first

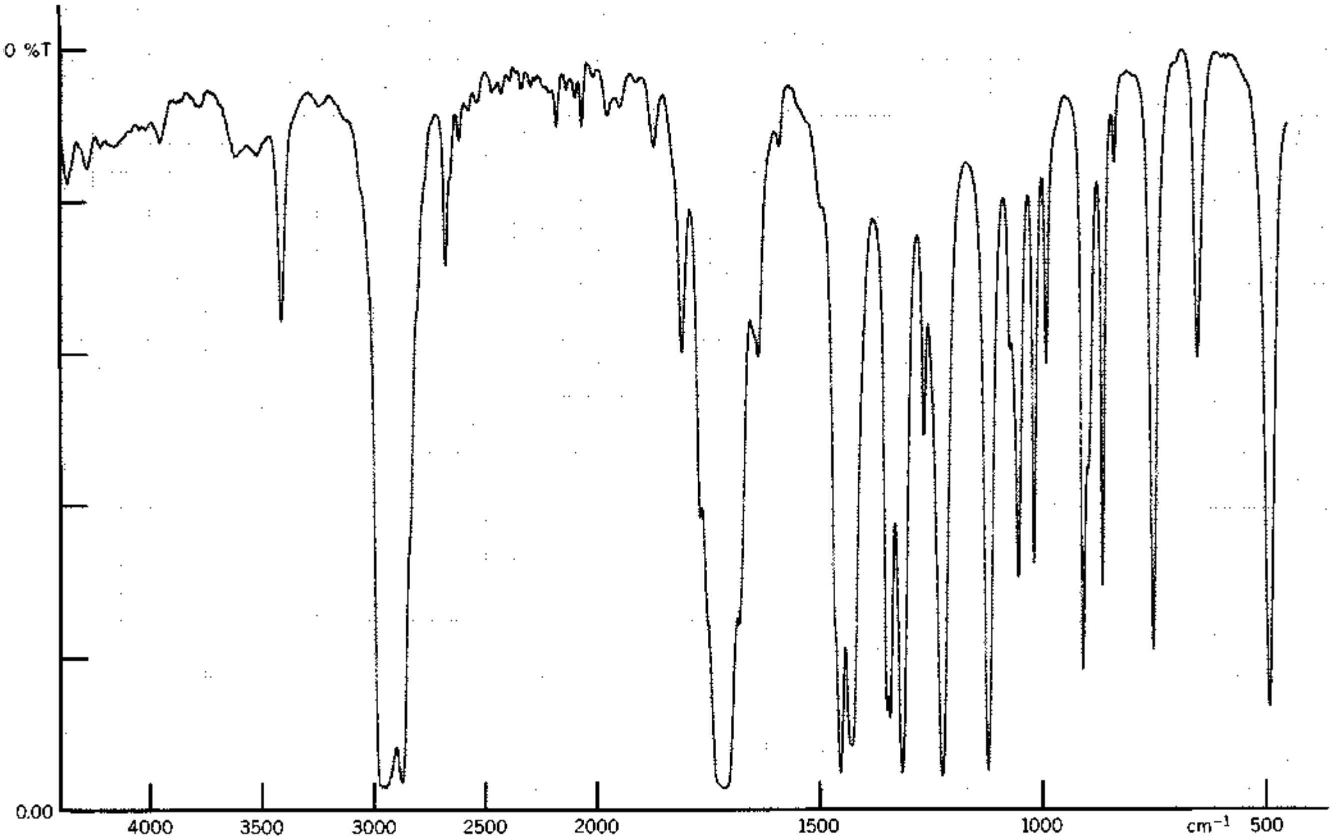

Sample Cyclohexanone

%T x ABS Background Scans 4 Scans 16

Acquisition & Calculation Time 42 sec Resolution 4.0 cm^{-1}

Sample Condition liquid, neat Cell Window KBr

Cell Path Length capillary film Matrix Material

Figure 6.6 IR spectrum: cyclohexanone.

overtone of the very intense carbonyl stretching mode found at 1715 cm^{-1}. The overtone does not fall exactly at double the frequency of the fundamental, but usually occurs slightly below that value because of anharmonic effects. The lack of absorption in the region near 3100–3000 cm^{-1} and the presence of a series of very strong absorption bands at 3000–2850 cm^{-1} indicates that the only C—H stretching modes present are part of sp^3 systems. Thus, the spectrum is typical of an aliphatic ketone. The occurrence of a 1425-wavenumber band suggests the presence of at least one methylene group adjacent to the carbonyl group, while the 1450-wavenumber band requires other methylene groups more remote to the C═O group. The lack of absorption in the 1400- to 1375-wavenumber region indicates the absence of any methyl groups (a good indication of a simple aliphatic ring system) and further suggests that the absorption at 1450 cm^{-1} must arise entirely from methylene scissoring modes. The value of 1715 cm^{-1} for the C═O stretch supports the presence of a six-membered ring.

Now examine the spectrum of your reaction product (a typical example is given in Fig. 6.7). The spectrum is rather different from that of the starting material. The major changes are a new very strong broad band occurring between 3500 and 3100 cm^{-1} and the large drop in the intensity of the band found at 1715 cm^{-1}.

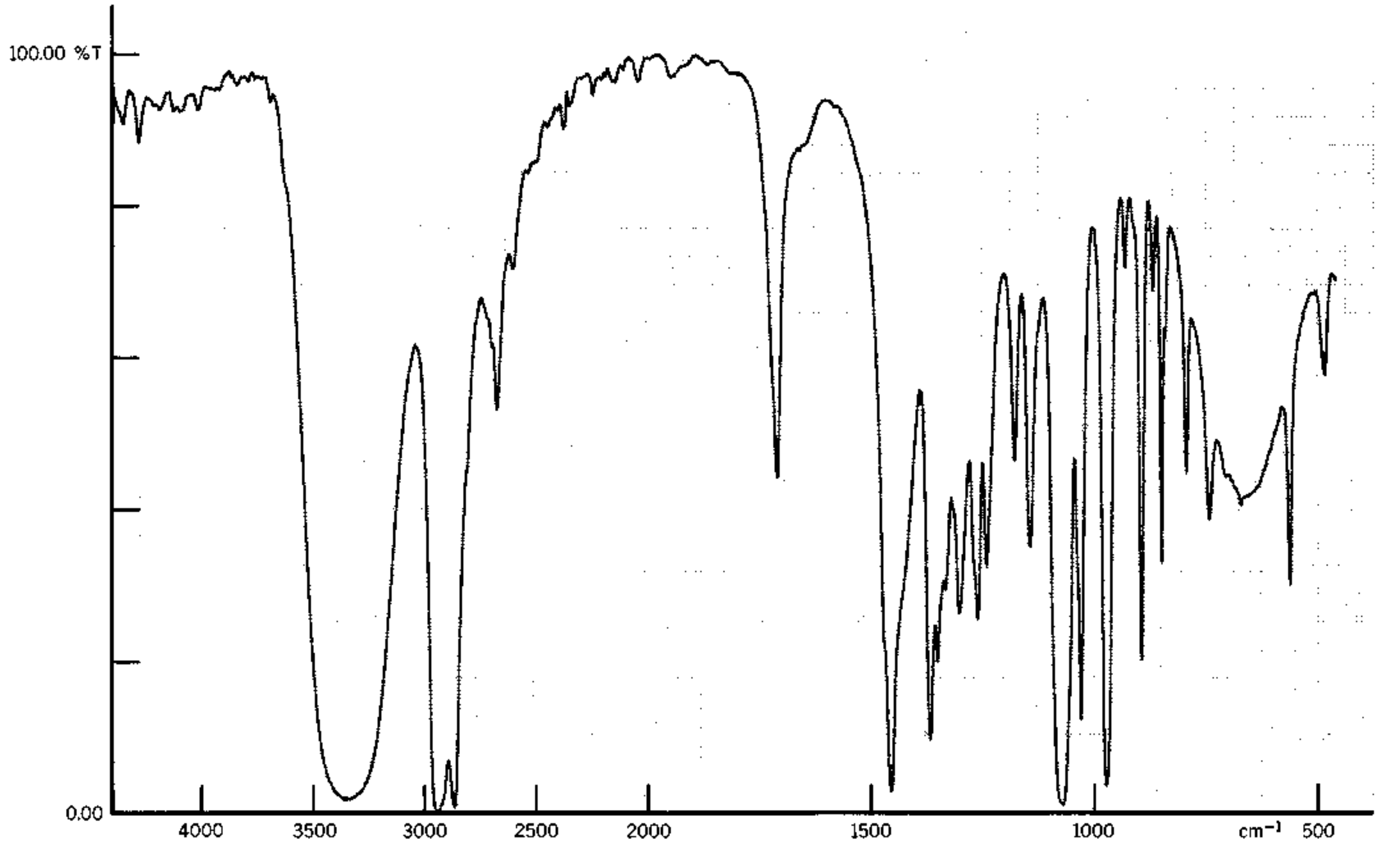

Sample Cyclohexanol (crude product)
%T X ABS Background Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm - 1
Sample Condition liquid, neat
Cell Window KBr
Cell Path Length capillary film
Matrix Material

Figure 6.7 **IR spectrum: cyclohexanol (crude product).**

These changes indicate the reductive formation of an alcohol group from the carbonyl system. The band centered near 3300 cm^{-1} results from the single polarized O—H stretching mode. The drop in intensity of the 1715-wavenumber band indicates the loss of the carbonyl function. The exact amount of cyclohexanone remaining can be determined by carrying out a Beer's law type analysis, but in this case we will use gas chromatographic techniques to determine this value. Other bands of interest in the spectrum of cyclohexanol occur at 1069 and 1031 cm^{-1}. These bands can be assigned, respectively, to the equatorial and axial C—O stretching of the rotational conformers of this alicyclic secondary alcohol. A broad band (width ~300 cm^{-1}) can be found near 670 cm^{-1}. This absorption arises from an O—H bending, out-of-plane mode, of the associated alcohol. This band is generally identified only in neat samples where extensive hydrogen-bonding occurs. Also note that the band at 1425 cm^{-1} has vanished because there are no methylene groups alpha to carbonyl systems in the product.

Separation of Small Quantities of Cyclohexanone from Cyclohexanol

Now proceed with purification of the reaction product by preparative gas chromatography. Use the following conditions and refer to Experiment [2] for the collection technique. If time permits, or perhaps in a later laboratory period, determine the infrared spectrum of the purified product. Describe and explain the changes observed in the new spectrum compared to that of the crude product.

Example

9:1 (v/v) cyclohexanol/cyclohexanone
10% Carbowax 20 M (stationary phase)
Injection volume: 15 μL
Temperature: 130 °C
He flow rate: 50 mL/min
Column: $\frac{1}{4}$-in. × 8-ft stainless steel
Chart speed: 1 cm/min

	Cyclohexanol		Cyclohexanone
Run	Retention Time (min)	Yield (mg)	Retention Time (min)
1	15.3	6.5	11.6
2	17.3	7.3	12.5
3	17.2	8.5	12.6
4	16.0	9.6	12.0
5	14.6	7.3	11.2
6	14.5	8.7	11.2
7	15.5	8.9	11.7
8	15.5	8.9	11.8
9	16.4	8.8	12.3
10	15.4	8.4	12.7
Av	15.8 ± 1	8.3 ± 0.9	12.0 ± 0.6

$$\text{Cyclohexanol injected} = 0.9(15\ \mu\text{L})(0.963\ \text{mg}/\mu\text{L}) = 13.0\ \text{mg}$$
$$\text{Percent yield} = 8.3/13.0 \times 100 = 63.8\%$$

NOTE. Collection efficiencies approaching 90% can be obtained by cooling the collection tube. Liquid nitrogen-soaked tissues work best, but methanol-soaked tissues or ice water can give a significant improvement.

Chemical Tests. Several chemical tests (see Chapter 9) may also be used to establish that an alcohol has been formed by the reduction of a ketone. Perform the ceric nitrate and 2,4-dinitrophenylhydrazine test on both the starting ketone and the alcohol product. Do your results demonstrate that an alcohol was obtained? You may also wish to prepare a phenyl- or α-naphthylurethane derivative of the cyclohexanol. Before the development of chemical instrumentation, the formation of solid derivatives was used extensively to identify reaction products.

Experiment 5B

cis- and *trans*-4-*tert*-Butylcyclohexanol

REACTION

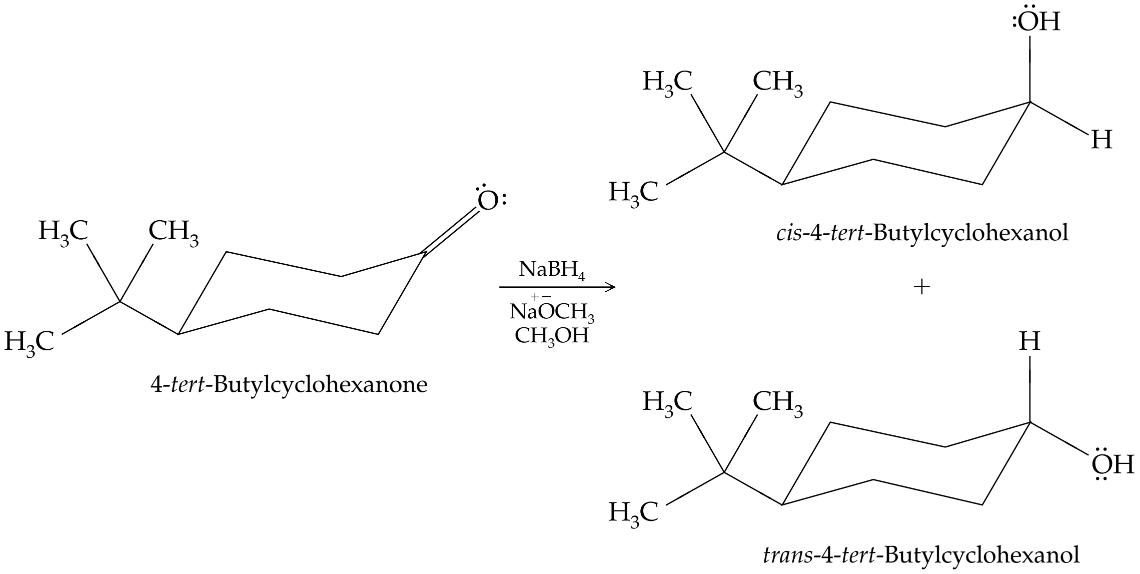
4-*tert*-Butylcyclohexanone
NaBH$_4$
NaOCH$_3$
CH$_3$OH
cis-4-*tert*-Butylcyclohexanol
+
trans-4-*tert*-Butylcyclohexanol

EXPERIMENTAL PROCEDURE

Estimated time of the experiment: 2.0 h.
For the GC analysis, 15 min per student.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
4-*tert*-Butylcyclohexanone	154.25	50 mg	0.33	47–50	
Methanol	32.04	50 μL			65
Sodium borohydride reducing solution		100 μL			

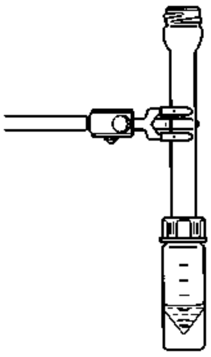

4-*tert*-Butylcyclohexanone, 50 mg + CH$_3$OH, 50 μL + NaBH$_4$ solution, 100 μL

Reagents and Equipment. In a tared 3.0-mL conical vial equipped with an air condenser weigh and place 50 mg (0.33 mmol) of 4-*tert*-butylcyclohexanone followed by 50 μL of methanol (⬅). Gently stir using a glass stir rod (or a magnetic spin vane if desirable) the contents of the vial to obtain a homogeneous solution.

Now slowly add 100 μL of the sodium borohydride reducing solution while stirring.

NOTE. The liquid reagents are dispensed by use of automatic delivery pipets. The preparation of the reducing solution is given in Experiment [5A], Reagents and Equipment (p. 154).

Reaction Conditions. After allowing the resulting solution to stand 10 min at room temperature, begin monitoring the resulting solution by TLC. Using as a solvent system of ethyl acetate:hexane (1:4), the R_f of 4-*tert*-butylcyclohexanol is 0.3 (R_f of 4-*tert*-butylcyclohexanone is 0.6) when stained with a solution of *p*-anisaldehyde (135 mL ethanol, 5 mL H_2SO_4, 1.5 mL glacial acetic acid, and 3.7 mL *p*-anisaldehyde).

Isolation of Product. When the reaction is complete as judged by TLC, work up the resulting solution using the procedure described in Experiment [5A], Isolation of Product (p. 154), with the exception that 250 mg of sodium sulfate is placed in the Pasteur filter pipet (➨).

In the **hood,** remove the dried methylene chloride solvent from the final solution by directing a gentle stream of nitrogen gas onto the liquid surface while at the same time externally warming the vial in a sand bath. The use of a heating bath will help to avoid moisture condensation on the residue during solvent evaporation.

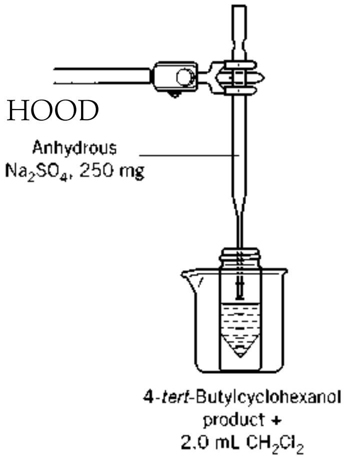

Purification and Characterization. The product mixture remaining after removal of the methylene chloride is normally of sufficient purity for direct characterization. Weigh the solid product and calculate the percent yield. Determine the melting point of your material. 4-*tert*-Butylcyclohexanol (mixed isomers) has a melting point of 62–70 °C.

Obtain the IR and NMR spectra of the crude mixture of isomers. Infrared sampling in this instance is best accomplished by the capillary film-melt (use the heat lamp) technique (see Chapter 8 and online IR discussions).

www

Infrared Analysis. Refer to the discussion in Experiment [5A] for an interpretation of the absorption bands found at 3435, 3000–2850, 1717, and 1425 cm^{-1} in the starting material (Fig. 6.8), and at 3250, 3000–2850 (1717 variable relative intensity—may be quite weak, why?), 1069, and 1031 cm^{-1} in the crude alcohol (Fig. 6.9).

In addition, the ketone has bands at 1396 (weak) and 1369 (strong), and the alcohol has bands at 1399 (weak) and 1375 (strong) cm^{-1}. These two pairs of bands establish the presence of the tertiary butyl group in these compounds.

Note that a weak band (3495 cm^{-1}) is present on the high wavenumber side of the 3250-cm^{-1} O—H stretching mode and that even in neat samples of the tertiary butyl derivative, the 670-wavenumber band, clearly evident in cyclohexanol, is difficult to observe.

The mixture of two diastereomeric alcohols that have been synthesized provides an ideal opportunity to introduce you to nuclear magnetic resonance (NMR) spectroscopy. This technique is an extremely powerful tool for the discrimination and characterization of diastereomeric compounds. As you will see, the two diastereomers have quite different NMR spectra. An interpretation of these spectra will allow you to determine the ratio of the two isomers and to make an unambiguous assignment of their relative stereochemistry.

This experiment gives you an opportunity to obtain and interpret NMR data if you have access to NMR equipment. The two diastereomeric alcohols exhibit different splitting patterns for the proton on the carbon bearing the —OH

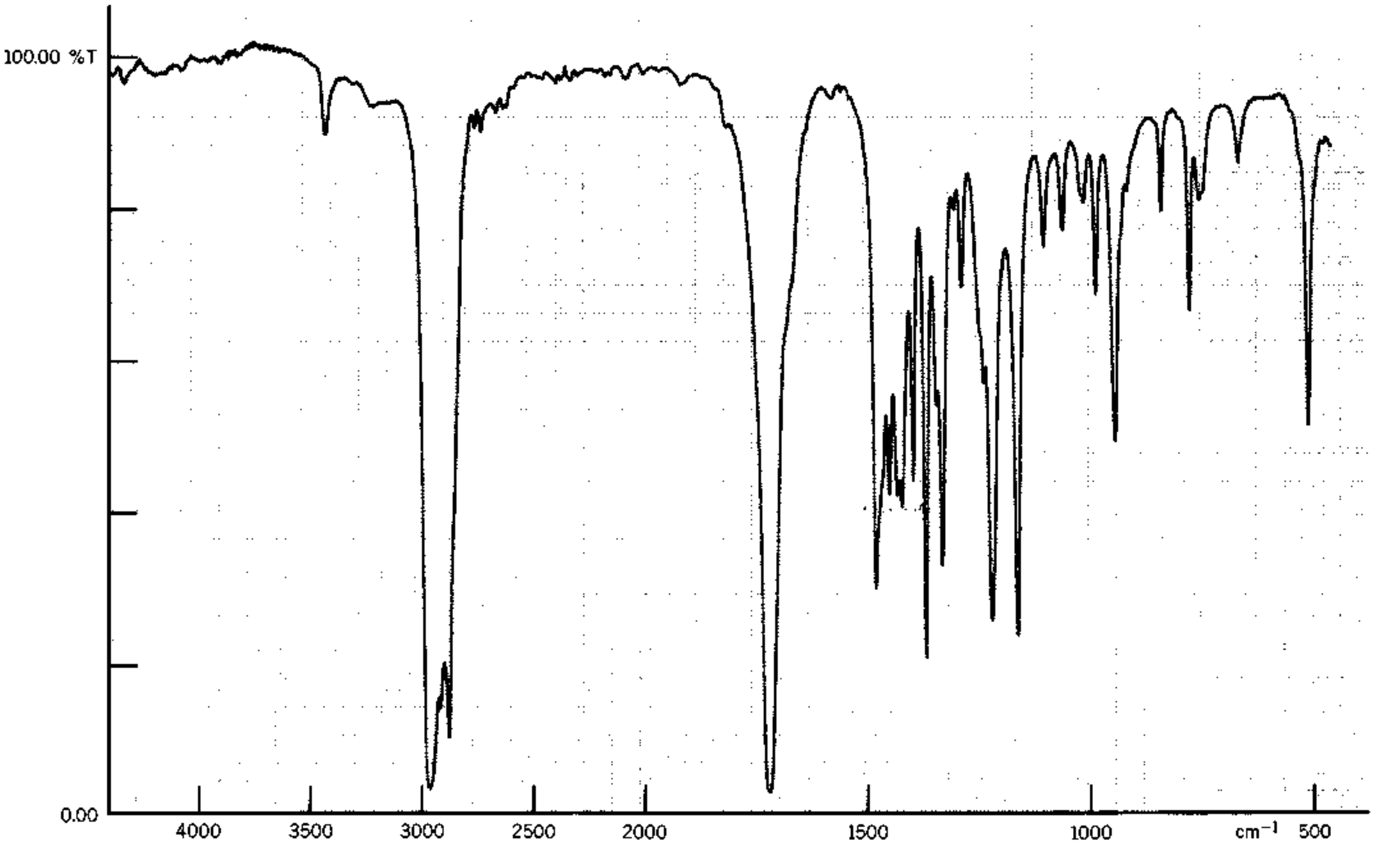

Sample	4-tert-Butylcyclohexanone		
%T X ABS — Background Scans	4	Scans	16
Acquisition & Calculation Time	42 sec	Resolution	4.0 cm⁻¹
Sample Condition	solid - melt	Cell Window	KBr
Cell Path Length	capillary film	Matrix Material	

Figure 6.8 IR spectrum: 4-*tert*-butylcyclohexanone.

group. These signals can be integrated to determine the ratio of the diastereomeric alcohols in the sample.

Nuclear Magnetic Resonance Analysis. Refer to the expanded NMR spectrum in Figure 6.10. The signals at about 4.04 and 3.52 ppm correspond to the proton on the carbon bearing the O—H group in the two diastereomers of 4-*tert*-butylcyclohexanol shown. On closer inspection, the downfield signal (4.04 ppm) is a pentet and the upfield signal (3.52 ppm) is a triplet of triplets. The pentet implies that the proton in question is coupled with equal coupling constants (J) to four adjacent protons. The triplet of triplets implies that the proton in question is coupled to two adjacent protons with a large coupling constant and to two other adjacent protons with a smaller coupling constant. Specifically, the proton in the first case must be equatorial and the proton in the second case must be axial, because the dihedral angle between an equatorial proton and each of the four adjacent protons is the same, about 60°, $J = 2$–3 Hz. When a proton is axial, the dihedral angle to the two adjacent equatorial

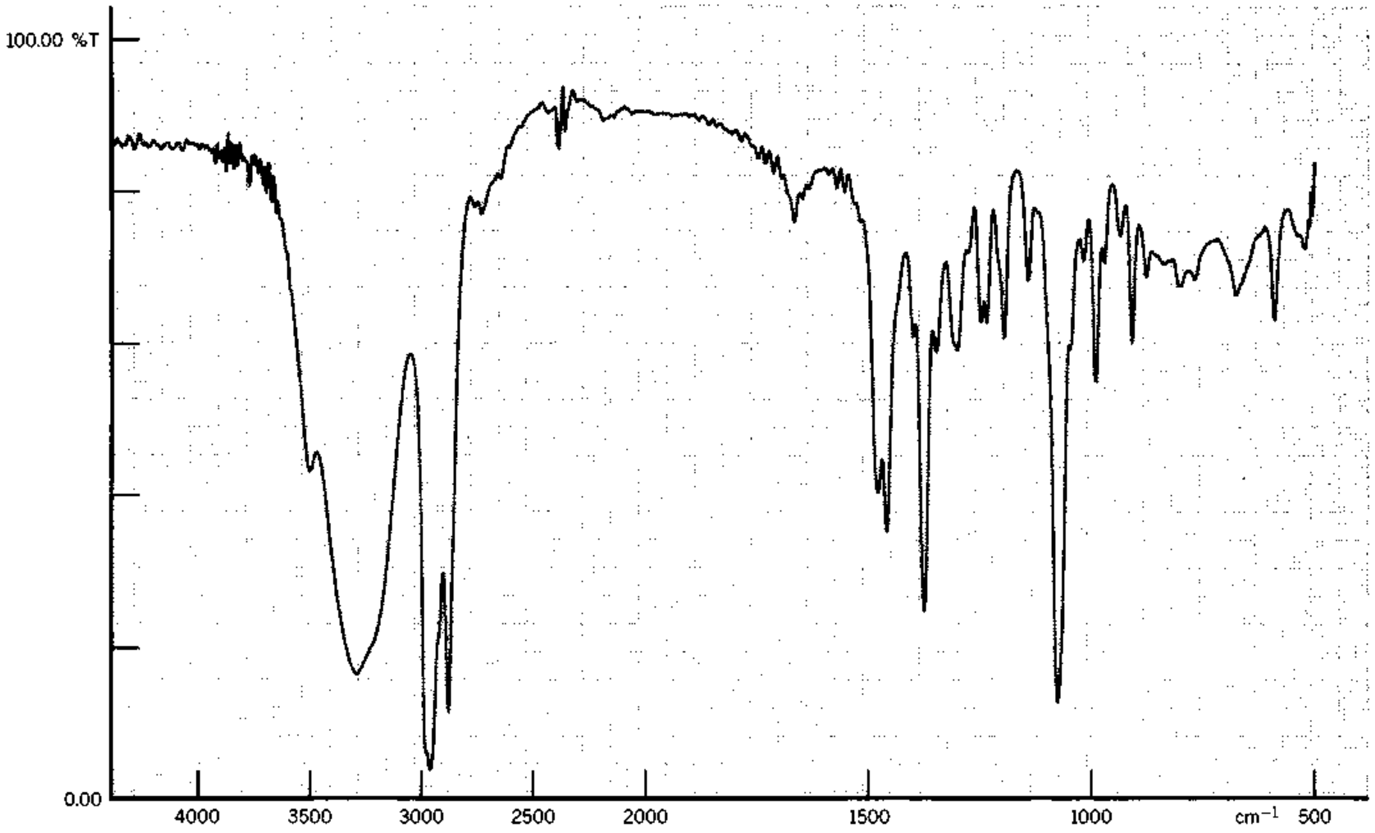

Sample 4-tert-Butylcyclohexanol (crude product and mixed isomers)

%T x ABS __ Background Scans 4	Scans 16
Acquisition & Calculation Time 42 sec	Resolution 4.0 cm - 1
Sample Condition solid	Cell Window
Cell Path Length	Matrix Material KBr

Figure 6.9 IR spectrum: 4-*tert*-butylcyclohexanol (crude product and mixed isomers).

protons is about 60° ($J \approx 3$ Hz) and the dihedral angle to the two adjacent axial protons is about 180° ($J \approx 13$Hz), thus producing a triplet of triplets.

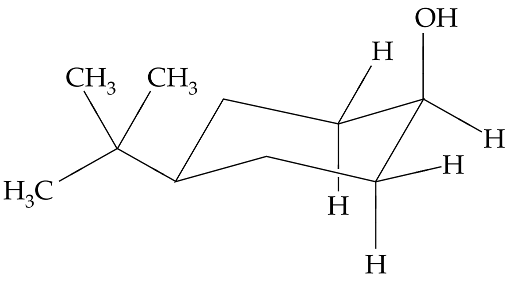

Cis

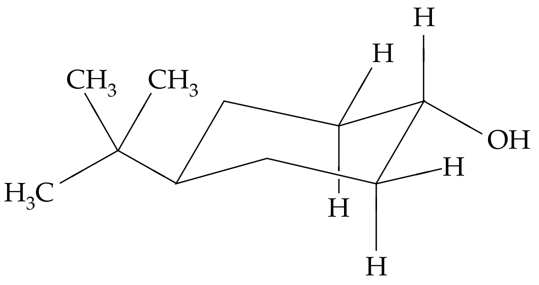

Trans

Gas Chromatographic Analysis. The cis and trans isomers of 4-*tert*-butylcyclohexanol may be separated by gas chromatography using a $\frac{1}{4}$-in. × 8-ft,

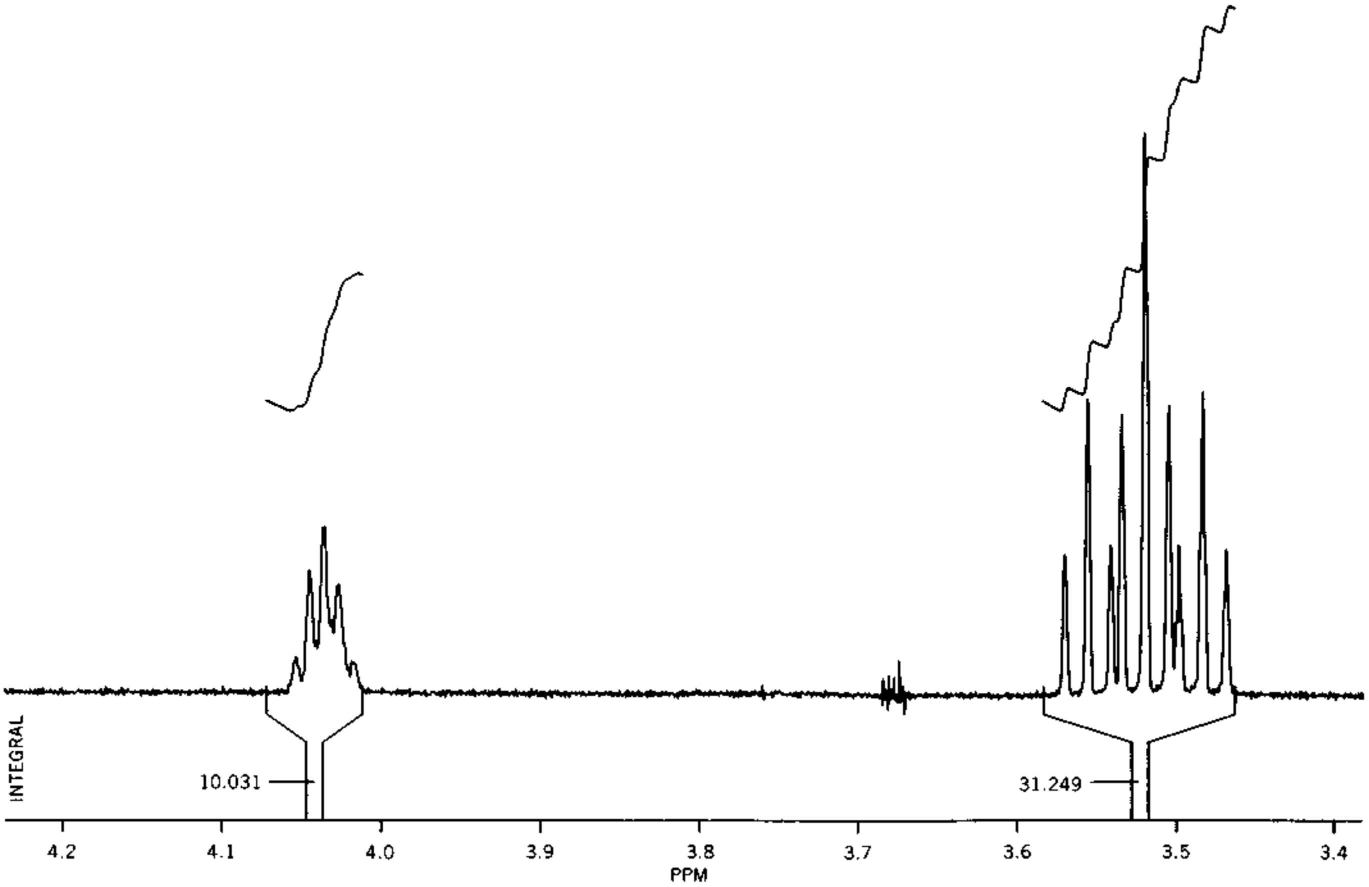

Figure 6.10 NMR spectrum: *cis*- and *trans*-4-*tert*-butylcyclohexanol.

10% FFAP column at 110 °C. Prepare a methylene chloride solution of the alcohol mixture having a concentration of 0.5 mg/μL and inject a 5.0-μL sample into the GC apparatus. At a flow rate of 50 mL/min (He), the retention time of the cis isomer is ~13 min, and the retention time of the trans isomer is ~16 min.

Determine the percentage of each isomer present in the sample by determining the area under each curve.

NOTE. If a Carbowax column (170 °C) is used, the starting ketone has a retention time similar to that of the cis alcohol. Therefore, if the reaction does not go to completion, the apparent cis/trans ratio may not be accurate. It has recently been demonstrated that 10% FFAP columns will resolve all components present in the product mixture. Thus, the starting ketone and the cis isomer concentrations may be effectively established in addition to the trans isomer.

The latter separation scheme was developed by T. J. Dwyer and S. Jones at the University of California, San Diego.

NOTE. Several techniques may be used to calculate the area under a curve (mm^2), but the following method is satisfactory for your needs and gives reproducible results of ±3–4%: Multiply the peak height (mm) by the width at one-half height (mm), measured from the base line of the curve.

QUESTIONS

6-27. Suggest a chemical test that would allow you to distinguish between *tert*-butyl alcohol and 1-butanol, both of which give a positive ceric nitrate test.

6-28. Which of the isomeric butanols ($C_4H_{10}O$) can be prepared by reduction of a ketone with sodium borohydride?

6-29. Why are there axial and equatorial hydroxyl isomers for 4-*tert*-butylcyclohexanol, but not for cyclohexanol itself?

6-30. What aldehyde or ketone would you reduce to prepare the following alcohols?
(a) Benzyl alcohol **(b)** 3,3-Dimethyl-2-butanol **(c)** 3-Methyl-1-butanol

6-31. The *cis*- and *trans*-4-*tert*-butylcyclohexanol prepared in Experiment [5B] each have a plane of symmetry. Draw this symmetry element for each of the diastereomers.

6-32. In the spectrum of the crude product obtained from the reduction of 4-*tert*-butylcyclohexanone, the fingerprint region appears to possess bandwidths that are slightly broader than those found in cyclohexanol. Explain.

6-33. The reduction of 4-*tert*-butylcyclohexanone is a stereoselective reaction. Which isomer predominates? If you do not have NMR data available, it is still possible to arrive at a rough estimate of the product ratio. How would you go about this measurement? Suggest a value.

6-34. In the spectrum of the crude 4-*tert*-butylcyclohexanols one observes: (a) a weak band (3495 cm^{-1}) located on the high side of the 3250-cm^{-1} O—H stretching mode. (b) Even in neat samples of this alcohol, the 670-cm^{-1} band, clearly evident in cyclohexanol, is difficult to observe. Explain these observations. Is the same effect operating in both cases?

6-35. Sketch the proton NMR spectrum you would expect to observe for the following compounds.
(a) Acetone
(b) 1,1,2-Tribromoethane
(c) Propyl chloride
(d) 2,4-Dimethyl-3-pentanone
(e) 1-Bromo-4-methoxybenzene

BIBLIOGRAPHY

General references on metal hydride reduction:

Itsuno, S. *Org. React.* **1998,** *52,* 395.
Seyden-Penne, J. *Reductions by Alumino- and Borohydrides in Organic Systhesis;* VCH-Lavoisier: Paris, 1997.
Walker, E. R. H. *Chem. Soc. Rev.* **1976,** *5,* 23.

Sodium borohydride as a reducing agent:

Brown, H. C. *Organic Synthesis via Boranes,* Wiley: New York, 1975.
Cragg, G. M. W. *Organoboranes in Organic Synthesis;* Marcel Dekker: New York, 1973.
Ems-Wilson, J. *J. Chem. Educ.* **1996,** *73,* A171.
Fieser, L. F.; Fieser, M. *Reagents for Organic Synthesis;* Wiley: New York, 1967; Vol. I, p. 1050 and subsequent volumes.
Paquette, L., Ed. *Encyclopedia of Reagents for Organic Synthesis;* Wiley: New York, 2004.

Lithium aluminum hydride as a reducing agent:

Brown, W. G. *Org. React.* **1951,** *6,* 469.
Fieser, L. F.; Fieser, M. *Reagents for Organic Synthesis;* Wiley: New York, 1967; Vol. I, p. 581 and subsequent volumes.

***trans*-4-*tert*-Butylcyclohexanol has been prepared from the ketone using $LiAlH_4$ as the reducing agent:**

Eliel, E. L.; Martin, R. J. L.; Nasipuri, D. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 175.
See *Organic Syntheses,* Coll. Vols., for the use of these reagents in specific reductions.

EXPERIMENT 6

Photochemical Isomerization of an Alkene: cis-1,2-Dibenzoylethylene

Common names: *cis*-1,2-dibenzoylethylene
cis-1,4-diphenyl-2-butene-1,4-dione
CA number: [959-27-3]
CA name as indexed: 2-butene-1,4-dione, 1,4-diphenyl-, (Z)-

Purpose. This exercise illustrates the ease of cis–trans isomerization in organic molecules and, specifically in this case, demonstrates the isomerization of a trans alkene to the corresponding cis isomer via photochemical excitation.

BIOLOGICALLY IMPORTANT PHOTOCHEMICAL REACTIONS

A number of important biochemical reactions are promoted by the adsorption of UV-vis radiation. Vitamin D_3, which regulates calcium deposition in bones, is biosynthesized in just such a photochemical reaction. This vitamin is formed when the provitamin, 7-dehydrocholesterol, is carried through fine blood capillaries just beneath the surface of the skin and exposed to sunlight. The amount of radiation exposure, which is critical for the regulation of the concentration of this vitamin in the blood stream, is controlled by skin pigmentation and geographic latitude. Thus, the color of human skin is an evolutionary response to control the formation of vitamin D_3 via a photochemical reaction.

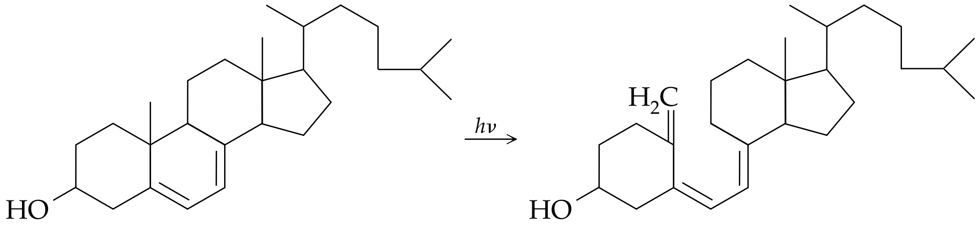

7-Dehydrocholesterol

Vitamin D_3

Another set of significant photochemical reactions in human biochemistry is contained in the chemistry of vision. These reactions involve vitamin A_1 (retinol), which is a C_{20} compound belonging to a class of compounds known as diterpenes. These compounds are molecules formally constructed by the biopolymerization of four isoprene, $CH_2{=}C(CH_3){-}CH{=}CH_2$, molecules. Retinol (an all-trans pentaene) is first oxidized via liver enzymes (biological catalysts) to vitamin A aldehyde (*trans*-retinal). The *trans*-retinal, which is present in the light-sensitive cells (the retina) of the eye, undergoes further enzymatic transformation (retinal isomerase) to give *cis*-retinal (a second form of vitamin A aldehyde) in which one of the double bonds of the all-trans compound is isomerized.

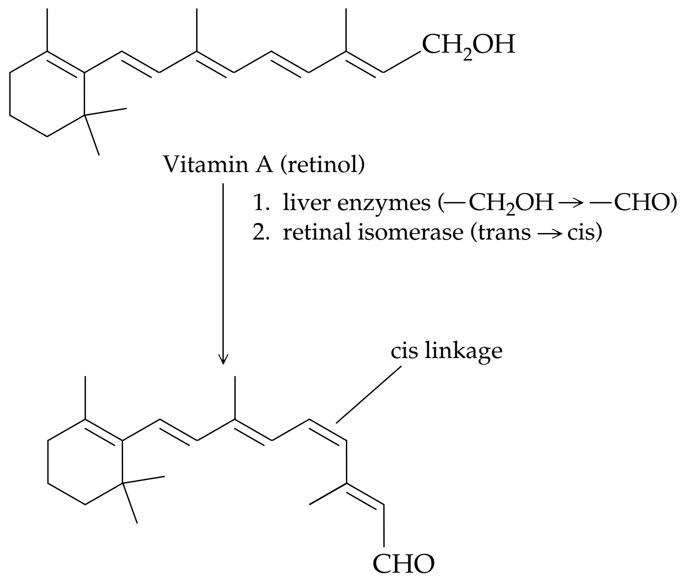
CH2OH
1. liver enzymes (—CH2OH → —CHO)
2. retinal isomerase (trans → cis)
cis linkage
CHO

Vitamin A (retinol)

Vitamin A aldehyde (*cis*-retinal)

The cis isomer of vitamin A aldehyde (retinal) possesses exactly the correct dimensions to become coupled to opsin, a large protein molecule (MW ~38,000) (coupling involves a reaction of the retinal aldehyde group, —C(H)═O, with an amine group [—NH_2] of the protein to form an imine linkage [RCH═NR]), to generate a light-sensitive substance, rhodopsin. This material is located in the rodlike structures of the retina. When the protonated form of rhodopsin (—CH═N^+HR), which absorbs in the blue-green region of the visible spectrum near 500 nm, is exposed to radiation of this wavelength, isomerization of the lone cis double bond of the diterpene group occurs and *trans*-rhodopsin is formed:

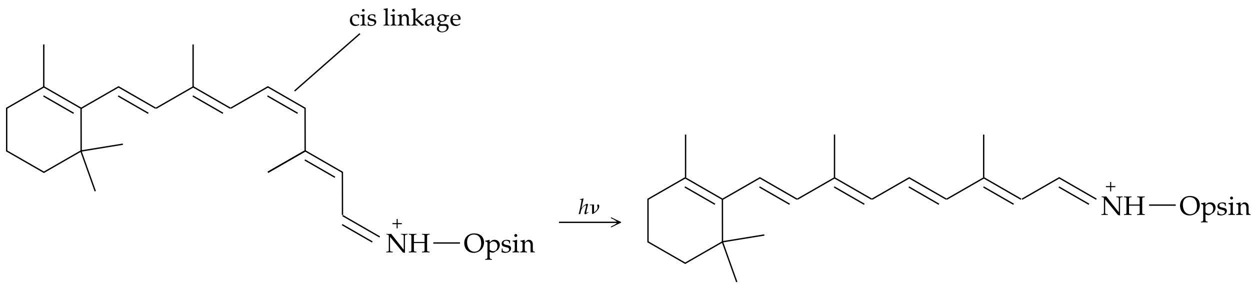

cis-Rhodopsin *trans*-Rhodopsin

This photoreaction (a fast reaction, 10^{-12} s) involves a significant change in the geometry of the diterpene group, which eventually (10^{-9} s) results in both a nerve impulse and the separation of *trans*-retinal from the opsin. The trans isomer is then enzymatically reisomerized back to the cis compound, which then starts the initial step of the visual cycle over again.

There are two interesting facts about this reaction: (1) This reaction is incredibly sensitive. A single photon will cause the visual nerve to fire. (2) All known visual systems use *cis*-retinal, regardless of their evolutionary trail.

The photoreaction that we study next is very similar to the cis–trans double-bond isomerism found in the vitamin A visual pigments. The only difference is that in our case we will be photochemically converting a trans double-bond isomer to the cis isomer.

Prior Reading

Technique 5: Crystallization
    Introduction (pp. 85–87)
    Use of the Hirsch Funnel (pp. 88–89)
    Craig Tube Crystallization (pp. 89–91)
    Recrystallization Pipet

Technique 6A: Thin-Layer Chromatography (pp. 97–99)

REACTION

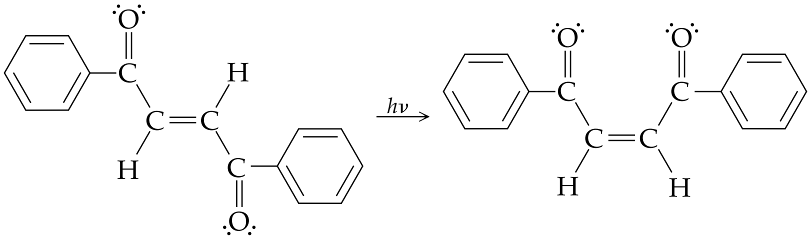

trans-1,2-Dibenzoylethylene *cis*-1,2-Dibenzoylethylene

DISCUSSION

The π bond of an alkene (C═C) is created by overlap of the sp^2 hybridized carbons' p orbitals. Rotation about the axis of the C═C requires a good deal of energy because it destroys the p orbital overlap, and thus the π bond. Unless the material is irradiated with light of the appropriate wavelength, absorption of the radiation does not take place, and the isomers do not interconvert unless sufficient thermal energy (60–65 kcal/mol, typically >200 °C) is supplied to break the π bond. If one of the electrons in the π bond (π molecular orbital) is photochemically excited into an antibonding, π*, molecular orbital, as occurs in this experiment, the π bond is weakened significantly. Rotation about this bond can then occur rapidly at room temperature.

It is the high-energy barrier to rotation about the C═C that gives rise to the possibility of alkene stereoisomers. The cis and trans isomers of an alkene system are called *diastereoisomer*s, or *diastereomers*. Like all stereoisomers, these isomers differ only in the arrangement of the atoms in space. These isomers have all the same atoms bonded to each other. Diastereomers are not mirror images of each other and, of course, are not superimposable (identical). This particular type of diastereomer often was referred to in the older literature as a "geometric isomer." Diastereomers, therefore, would be expected to possess different physical properties, such as melting points, boiling points, dipole moments, densities, and solubilities, as well as different spectroscopic properties. Because of these differences in physical properties, diastereomers are amenable to separation by chromatography, distillation, crystallization, and other separation techniques. We use both chromatographic and crystallization methods in the present experiment.

The course of the isomerization may be followed using *thin-layer chromatography* (Experiment [6B]) or by spectroscopic techniques using NMR analysis (Experiment [6C]).

The photochemical isomerization of a diazabicyclohexene system is presented in Experiment [F4].

Experiment 6A

Purification of *trans*-1,2-Dibenzoylethylene

EXPERIMENTAL PROCEDURE

Estimated time to complete the purification: 1.5 h of laboratory time.

Physical Properties of Components

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
trans-1,2-Dibenzoylethylene	236.27	100 mg	0.42	111	
Ethanol (95%)		6.0 mL			78.5
Methylene chloride		4.0 mL			40

Purification Conditions. Purify the starting alkene by recrystallization.

Weigh and add to a 10-mL Erlenmeyer flask 100 mg (0.42 mmol) of *trans*-1,2-dibenzoylethylene and 3.0 mL of methylene chloride.

NOTE. If the melting point of the alkene was not supplied to you, set aside a small sample (1–2 mg) of the weighed sample so that the evacuated melting point of this initial material can be obtained later.

JWCL196_c06_115-187.qxd 11/23/09 10:42 AM Page 167

EQA

Add decolorizing charcoal pellets (10 mg) to this solution and swirl the mixture gently for several minutes. Use a Pasteur filter pipet to transfer the methylene chloride solution to a second 10-mL Erlenmeyer flask containing a boiling stone (remember to hold the necks of the two flasks close together with the fingers of one hand during the transfer).

NOTE. If powdered charcoal is used instead of pellets, filter the solution using a Pasteur pipet packed with Celite, sand, and a cotton plug (Technique 6A). Rinse the filter paper with an additional 1 mL of methylene chloride and collect this rinse in the same flask.(➡)

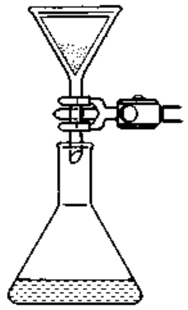

Trans isomer + CH_2Cl_2, 3 mL

Concentrate the filtered solution to dryness in a warm (50 °C) sand bath under a slow stream of nitrogen gas in the **hood.**

HOOD

Now add 95% ethanol (1–3 mL) to the flask and dissolve the yellow solid residue by warming in a sand bath, adding hot ethanol dropwise, until a homogeneous solution is obtained (➡).

Allow the solution to cool slowly to room temperature over a period of 15 min and then place it in an ice bath for an additional 10 min. Collect the yellow needles by vacuum filtration using a Hirsch funnel (➡) and then air-dry them on a porous clay plate or on filter paper.

Weigh the product and calculate the percent recovery. Determine the evacuated melting point and compare your result with both the literature value and that obtained with the material saved prior to recrystallization. If Experiment [3A] has been completed, compare the melting point of that sample of the trans alkene, which was obtained by concentration, via distillation, of an ethyl acetate solution. In the latter experiment, a simple crystallization was performed without the aid of decolorizing charcoal.

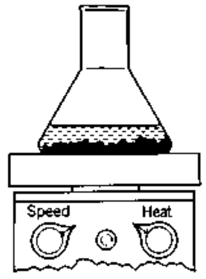

trans-1,2-Dibenzoylethylene, 150 mg + CH_2Cl_2, 3.0 mL, charcoal pellets

Experiment 6B

Isomerization of an Alkene: Thin-Layer Chromatographic Analysis

EXPERIMENTAL PROCEDURE

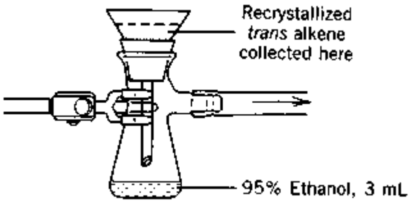

Estimated time to complete the experiment: 2.5 h of laboratory time. The reaction requires approximately 1 h of irradiation; the actual time to completion is quite sensitive to both radiation flux and temperature. These factors are largely determined by the distance the reaction vessel is positioned from the source of radiation.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
trans-1,2-Dibenzoylethylene	236.27	25 mg	0.11	111	
Ethanol (95%)		3.5 mL			78.5

Reagents and Equipment. To a 13 × 100-mm test tube weigh and add 25 mg (0.11 mmol) of recrystallized *trans*-1,2-dibenzoylethylene and 3.0 mL of 95% ethanol.

Reaction Conditions. Use a sand bath to warm the mixture ***gently*** until a homogeneous solution is obtained. Stopper the test tube *loosely,* or cover it with

GENTLY

CONFIRMING PAGES aptara

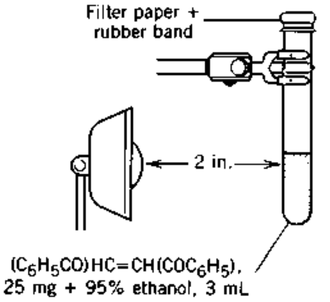
Filter paper + rubber band
2 in.
$(C_6H_5CO)HC{=}CH(COC_6H_5)$,
25 mg + 95% ethanol, 3 mL

filter paper held in place by a rubber band, and then place it approximately 2–4 in. from a 275-W sun lamp. Irradiate the solution for approximately 1 h (⬅).

NOTE. If a lower wattage lamp is used, longer irradiation times or shorter distances will be necessary. In either case, solvent evaporation can be significantly reduced by directing a flow of cool air (fan) over the reaction tubes. An alternative procedure is to allow the (sealed) tube to stand in sunlight at room temperature for several days.

The progress of the isomerization may be followed by thin-layer chromatography (TLC) analysis.

INFORMATION. The TLC analysis is carried out using Eastman Kodak silica gel–polyethylene terephthalate plates with a fluorescent indicator. ***Activate the plates at an oven temperature of 100 °C for 30 min*** *and then place them in a desiccator to cool until needed. After spotting, elute the plates using methylene chloride as the solvent. Visualize the spots with a UV lamp. The course of the reaction is followed by removing small samples (2–3 drops) of solution from the* ***hot*** *test tube at set time intervals with a Pasteur pipet and placing them in separate $\frac{1}{2}$-dram vials. See Technique 6A for the method of TLC analysis and the determination of R_f values. Approximate R_f values: trans = 0.72; cis = 0.64.*

HOT

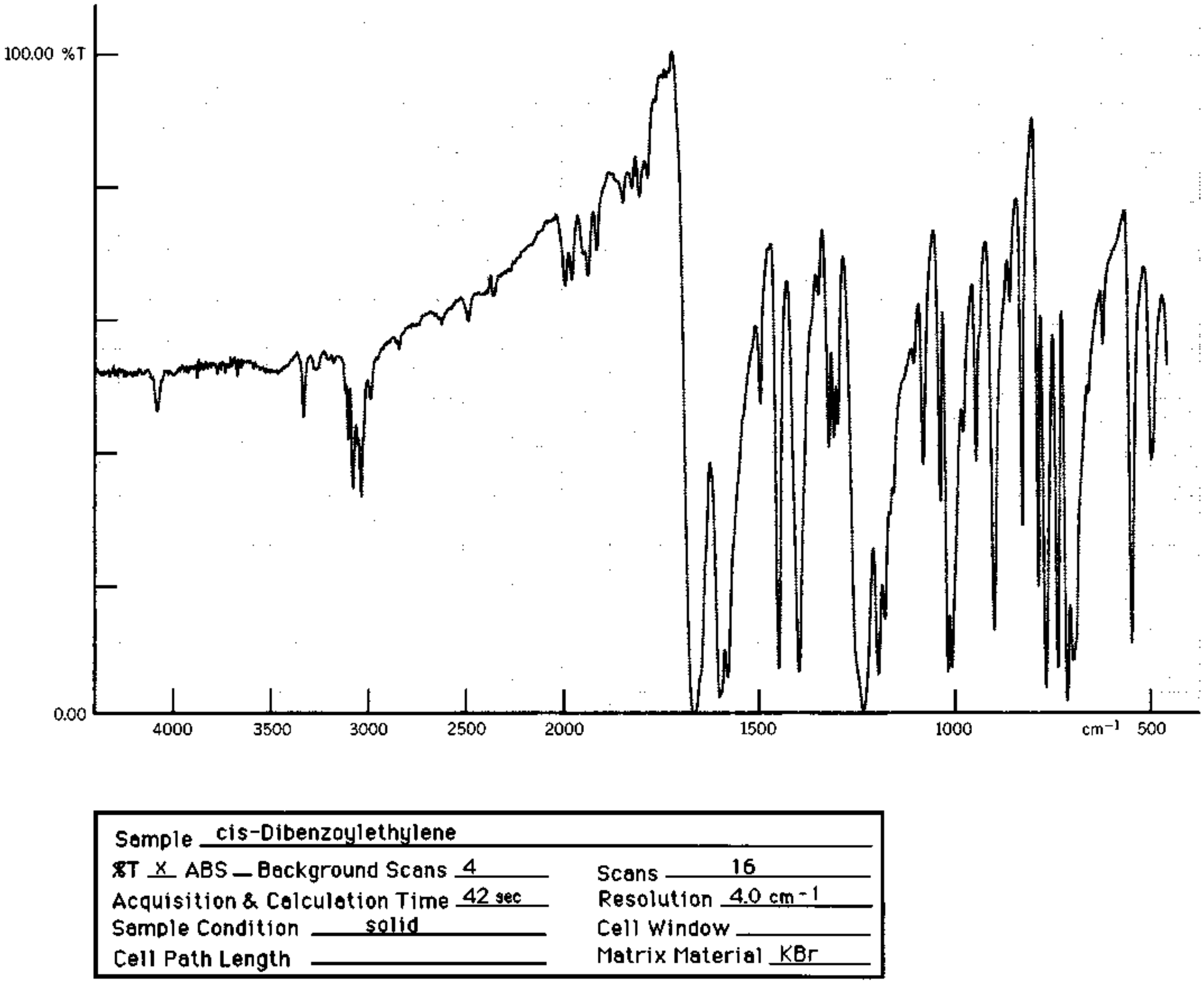

Sample cis-Dibenzoylethylene
%T X ABS __ Background Scans 4
Acquisition & Calculation Time 42 sec
Sample Condition solid
Cell Path Length
Scans 16
Resolution 4.0 cm-1
Cell Window
Matrix Material KBr

Figure 6.11 IR spectrum: *cis*-dibenzoylethylene.

HOT

Isolation of Product. Remove the **hot** test tube from the light source and allow the solution to cool to room temperature.

Place the resulting mixture in an ice bath to complete crystallization of the *colorless cis*-1,2-dibenzoylethylene product. Collect the crystals by vacuum filtration using a Hirsch funnel, wash them with 0.5 mL of cold 95% ethanol, and then air-dry them on a porous clay plate or on filter paper.

Purification and Characterization. Weigh the dried product and calculate the percent yield. Determine the melting point (evacuated) and compare your result with the literature value. The purity of the crude isolated product may be further determined by TLC analysis (if not used above). Finally, if necessary, further purify a portion of the isolated product by recrystallization from 95% ethanol using a Craig tube.

Obtain IR spectra of the cis and trans isomers and compare them to Figures 6.11 and 6.12. Alternatively, or in addition to the IR analysis, the UV-vis spectra may be observed in methanol solution and the results compared to Figures 6.13 and 6.14.

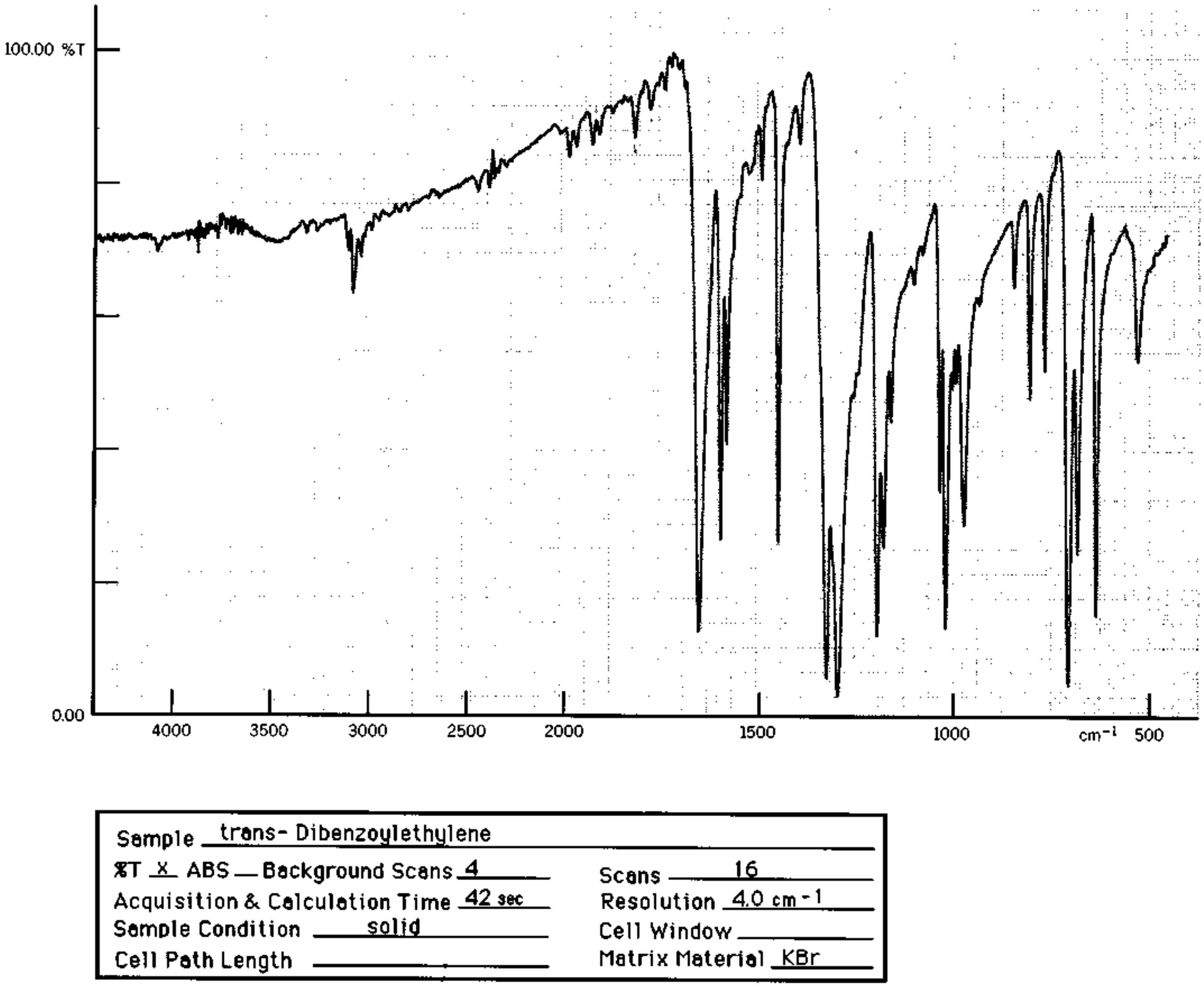

Sample trans- Dibenzoylethylene

%T x ABS ___ Background Scans 4

Scans 16

Acquisition & Calculation Time 42 sec

Resolution 4.0 cm ⁻¹

Sample Condition solid

Cell Window ___

Cell Path Length ___

Matrix Material KBr

Figure 6.12 IR spectrum: *trans*-dibenzoylethylene.

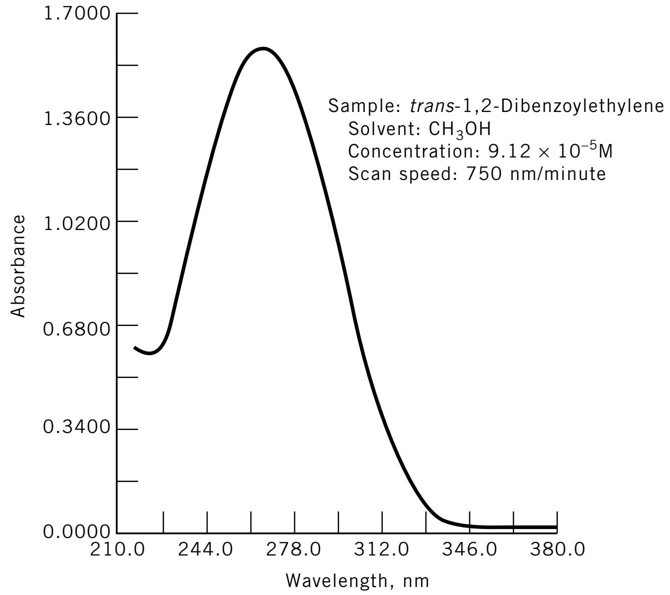

Figure 6.13 UV spectrum: *trans*-dibenzoylethylene.

Ultraviolet-Visible Analysis. The bright yellow color of the *trans*-dibenzoylethylene rapidly fades as the conversion to the colorless cis compound progresses under irradiation. This visual observation may be supported by an examination of the absorption spectra of the isomers in a methanol solution from 225 to 400 nm (see Figs. 6.13 and 6.14). The λ_{max} of the trans isomer drops from 268 to 259 nm in the cis compound. This shift to shorter wavelengths is just enough to move the long-wavelength end of the absorption band in the trans isomer out of the visible region and into the near-ultraviolet (thus, the cis compound does not absorb light to which the eye is sensitive and the compound appears colorless). This observation is consistent with the theory that the spatial contraction of extended π systems moves the

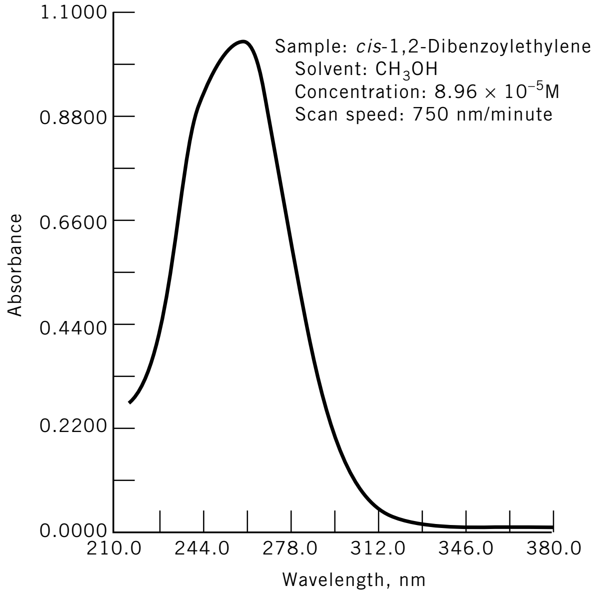

Figure 6.14 UV spectrum: *cis*-dibenzoylethylene.

associated electronic transitions to higher energy gaps (higher frequencies or shorter wavelengths), which is exactly what takes place during the trans–cis isomerization.

Infrared Analysis. The infrared spectral changes are consistent with the proposed reaction product. Consider the spectrum of the trans starting material (Fig. 6.12). It contains the following:

1. The macro group frequency train (see Strategies for Interpreting Infrared Spectra, p. 542, for conjugated aromatic ketones is 3080 and 3030 (C—H, aromatic), 1652 (doubly conjugated carbonyl), 1599, and 1581 (ν_{8a}, ν_{8b} degenerate ring stretch, strong intensity of 1581-wavenumber peak confirms ring conjugation), 1495 and 1450 (ν_{19a}, ν_{19b} degenerate ring stretch, ν_{19a} weak) cm^{-1}.

2. The monosubstituted phenyl ring macro group frequency train is 1980(d), 1920(d),[4] 1820, 1780 (mono combination-band pattern), 708 (C—H out-of-plane bend, C═O conjugated), 686 (ring puckering) cm^{-1}.

3. The presence of the trans double bond is indicated by the single medium intensity 970 cm^{-1} band, as the C—H stretching region is overlapped by the aromatic ring C—H stretches.

The spectrum of the cis photoproduct (Fig. 6.11) possesses the same macro group frequencies as the starting material:

1. The conjugated aromatic ketone frequency train is 3335 (overtone of C═O stretch), 3080 and 3040, 1667, 1601, 1581, 1498 (weak), and 1450 cm^{-1}.

2. The monosubstituted phenyl ring macro frequency train is assigned peaks of 1990, 1920, 1830, 1795, 710, and 695 cm^{-1}.

3. The cis double bond is clearly present and utilizes a macro frequency train of 3030 (overlapped by aromatic C—H stretch), 1650 (C═O stretch, shoulder on the low wavenumber side of conjugated carbonyl; resolved in spectra run on thin samples), 1403 (═C—H out-of-phase, in-plane bending mode, strong band not present in trans compound), 970 (trans, in-phase, out-of-plane bend missing), 820 (cis, in-phase out-of-plane bend; band not found in spectrum of trans isomer) cm^{-1}.

Discuss the similarities and differences of the experimentally derived spectral data to the reference spectra (Figs. 6.11 and 6.12).

Optional Isolation of the Thermodynamically Most Stable Reaction Product Under the Conditions Used to Carry Out the Above Reaction. If the photo reaction exposure is continued, the cis isomer, which is formed quickly, slowly undergoes conversion to a more stable, new product, ethyl 4-phenyl-4-phenoxy-3-butenoate. This conversion of the intermediate cis isomer may be followed conveniently by thin-layer chromatography. Maximum yields are obtained over approximately 24 h under the above reaction conditions. The product has $R_f = 0.8$ in 1:1 hexane/methylene chloride. Evaporation of the solvent yields a yellow oil.

---

[4] d = double peak.

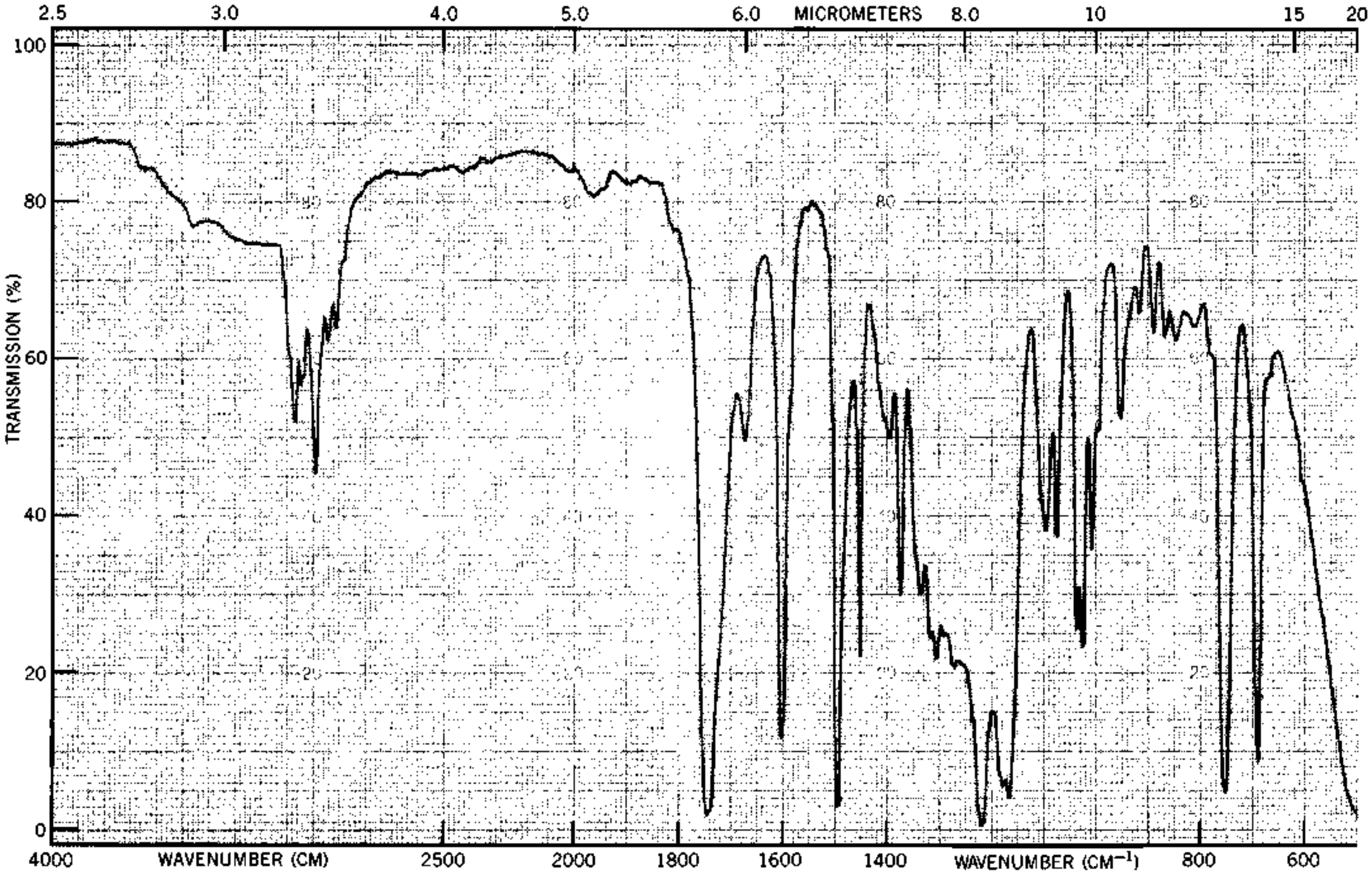

Scan Time: 12 min
Slit Program: normal
Concentration: neat
Cell Path length: capillary film
Operator: E.M.
Date 7/20/87
Sample: Ethyl 4-phenyl-4-phenoxy-3-butenoate

Figure 6.15 IR spectrum: ethyl 4-phenyl-4-phenoxy-3-butenoate.

The infrared spectrum of this new material (Fig. 6.15) indicates that significant changes have occurred in the structure of the material. The proposed structure involves a rather spectacular molecular rearrangement of the cis-unsaturated ketone to yield an ethyl ester containing a phenoxy-substituted double bond. Can you rationalize the data to fit this structure? Suggest possible macro group frequencies that are present in the IR spectrum of the rearranged product.

Mixture Melting-Point Measurements. Observe a series of evacuated mixture melting points (see Chapter 4) with isomer ratios of 75:25, 50:50, and 25:75. These values will allow you to *estimate* the eutectic temperature of this system. This same technique is used in Experiment [29A] to aid in identifying the isolated product, 2,5-dichloronitrobenzene, which has a melting point only 3 °C higher than 1,4-dichlorobenzene, the starting material in this nitration reaction.

Isomerization of an Alkene: Nuclear Magnetic Resonance Analysis

Experiment 6C

EXPERIMENTAL PROCEDURE

Estimated time of experiment: 0.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
trans-1,2-Dibenzoylethyene	236.27	5 mg	0.02	111	
Chloroform-*d*	120.39	500 μL			62

Reagents and Equipment. Prepare a sample of 4–5 mg of recrystallized *trans*-1,2-dibenzoylethylene in about 500 μL of $CDCl_3$ in an NMR tube. Using a fine capillary, spot a silica gel TLC plate (see Experiment [6B]) with a sample of this solution. Use TLC to track the results of the NMR experiment.

Obtain an NMR spectrum of this solution.

Reaction Conditions

NOTE. A 15-min exposure is normally adequate but, as can be seen by TLC analysis, complete isomerization of this alkene may not occur in this short a time.

Clamp the NMR tube 3–4 in. (see comments above, Experiment [6A]) from a 275-W sun lamp for 15–20 min (see Experiment [6B]).

Analysis of the Results. After irradiation, spot the original TLC plate with this irradiated solution and elute the plate with methylene chloride/hexane (1:1) solvent (see Experiment [6B]). Obtain an ^{1}H NMR spectrum of the irradiated solution.

Does the TLC analysis correlate with the NMR data? Is there evidence that the isomerization did occur?

The data in the following table were taken on a 300-MHz NMR instrument. Which signal in each isomer comes from the hydrogen atoms attached to the C═C group?

^{1}H Chemical Shift Data

Trans Isomer (ppm)	Cis Isomer (ppm)
8.06 doublet	7.9 doublet
8.01 singlet	7.55 triplet
7.63 triplet	7.43 triplet
7.52 triplet	7.14 singlet

QUESTIONS

6-36. What properties should an ideal recrystallization solvent have?

6-37. What is meant by the term *solvent pair* when it is used in reference to the recrystallization of solids?

6-38. What is the purpose of adding powdered charcoal to the solvent system during a recrystallization sequence?

6-39. Why do we obtain larger crystals if a solution of an organic compound is allowed to cool slowly?

6-40. List several advantages of using the Craig tube to purify crystalline products in contrast to using the Hirsch funnel filtration method.

6-41. The stereochemistry of the more highly substituted alkenes is difficult to define using the cis and trans designations. Therefore, a more systematic manner of indicating stereochemistry in these systems has been developed that uses the *E* and *Z* nomenclature system. Draw the structures of the *E* and *Z* stereoisomers of 1,4-diphenyl-2-butene-1,4-dione used in this experiment. In this case, which is cis and which is trans?

6-42. The cis H—C═C—H in-phase out-of-plane bending frequency normally is found in the range 740–680 cm^{-1} and is broad. In addition, this vibrational mode can be quite variable in intensity. On conjugation with C═O groups, the band rises to near 820 cm^{-1}.

- **(a)** Can you explain the underlying cause of this observation? (*Hint:* See the discussion of cis double-bond group frequencies in Chapter 8.)
- **(b)** Can you explain why a broad band near 690 cm^{-1} assigned to this out-of-plane mode is found in the cis product obtained in this reaction?

6-43. In *trans*-1,2-dibenzoylethylene, even though the conjugated C═O group frequency coincides directly with the expected C═C stretching frequency, we would not have expected to observe the latter stretching mode in the infrared spectrum. Why?

6-44. Explain how this photochemical isomerization allows the production of the thermodynamically *less* stable cis isomer. In other words, why is the trans isomer exclusively converted to the cis isomer during short reaction periods and not vice versa? Is it possible, under these conditions, that the trans and cis isomers are in equilibrium with one another?

BIBLIOGRAPHY

The photochemical isomerization was adapted from the following references:

Pasto, D. J.; Ducan, J. A.; Silversmith, E. F. *J. Chem. Educ.* **1974,** *51,* 277.
Silversmith, E. F.; Dunsun, F. C. *J. Chem. Educ.* **1973,** *50,* 568.

Reviews on photochemical isomerization reactions may be found in the following references.

Arai, T.; Tokumaru, K. *Chem. Rev.* **1993,** *93,* 23.
Coxton, J. M.; Halton, B. *Organic Photochemistry,* 2nd ed.; Cambridge University Press: New York, 1987.
Coyle, J. D. *Introduction to Organic Photochemistry;* Wiley: New York, 1991.
Kagan, J. *Organic Photochemistry: Principles and Applications;* Academic Press: Orlando, FL, 1993.
Kopecky, J. *Organic Photochemistry: A Visual Approach;* Wiley/VCH, New York, 1992.
Prasad, P. N. *Introduction to Biophotonics;* Wiley: New York, 2003.

EXPERIMENT 7

The Cannizzaro Reaction with 4-Chlorobenzaldehyde: 4-Chlorobenzoic Acid and 4-Chlorobenzyl Alcohol[5]

Common name: 4-chlorobenzoic acid
CA number: [74-11-3]
CA name as indexed: 4-chlorobenzoic acid
Common name: 4-chlorobenzyl alcohol
CA number: [873-76-7]
CA name as indexed: benzenemethanol, 4-chloro-

---

[5] Portions of this experiment were previously published: Mayo D. W.; Butcher, S. S.; Pike R. M.; Foote, C. M.; Hotham, J. R.; Page, D. S. *J. Chem. Educ.* **1985,** *62,* 149.

Stanislao Cannizzaro (1826–1910): This famous reaction is named for Stanislao Cannizzaro. He was born in Palermo, Italy, where he became interested in chemistry at an early age and studied with Professor Piria at the University of Pisa. He arrived in Paris after participating in the 1847 Sicilian uprising. By 1851 he had moved to Egypt where he became Professor of Chemistry at the National School of Alexandria. It was here that he discovered the reaction that carries his name. In the process he also was the first person to identify benzyl alcohol, which he obtained from the reaction of potash (KOH) with benzaldehyde. It is interesting to note that benzyl alcohol was the first alcohol of the aromatic series to be isolated and characterized. Cannizzaro devoted extensive work to the study of aromatic alcohols and he was the first to propose the term *hydroxyl* for the —OH group. Later he demonstrated the conversion of aromatic alcohols of the benzyl class into their corresponding halides and further to their phenyl acetic acid derivatives. In 1855 he moved to Genoa and in 1861 he moved back to his birthplace to become Professor of Chemistry at Palermo. Finally, 10 years later at the young age of 45, he assumed the chemistry chair at the University of Rome, which he held for 39 years! Upon his move to Rome his old political interests surfaced and he became a member of the Italian Senate that same year. He later carried the honor of serving as vice president of that august body. Cannizzaro was also honored by the Royal Society of London in 1891 when he was awarded the Copley Medal for his investigations of atomic and molecular weights. He was quick to recognize the importance of Avogadro's hypothesis and his studies played a key role in placing this hypothesis on a sound scientific basis, which ultimately was the spark that ushered chemistry into the modern era.[6]

Purpose. This experiment illustrates the simultaneous oxidation and reduction of an aromatic aldehyde to form the corresponding benzoic acid and benzyl alcohol. The experiment further demonstrates the techniques for separation of a carboxylic acid from a neutral alcohol. For a detailed discussion of this extraction procedure, see Experiment [4C].

Prior Reading

Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
  Separation of Acids and Bases (pp. 77–79)
Technique 5: Crystallization
  Use of the Hirsch Funnel (pp. 88–89)
  Craig Tube Crystallization (pp. 89–91)
Technique 6A: Thin-Layer Chromatography (pp. 97–99)
Chapter 8: Infrared Spectroscopy (pp. 539–561)

---

[6] See Newell, L. C. *J. Chem. Educ.* **1926,** *3,* 1361; Parravano, N. *J. Chem. Educ.,* **1927,** *4,* 836; Tilden, W. A. *J. Chem. Soc.* **1912,** 1677; *Dictionary of Scientific Biography;* Gillespie, C. C., Ed.; Scribner's: New York, 1971, Vol. III, p. 45; Surrey, A. R. *Name Reactions in Organic Chemistry;* Academic Press: New York, NY, 1954; p. 27.

REACTION

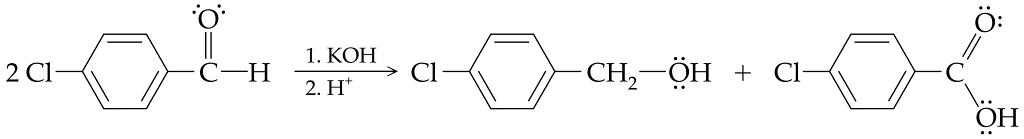

4-Chlorobenzaldehyde 4-Chlorobenzyl alcohol 4-Chlorobenzoic acid

DISCUSSION

The carbonyl group of an aldehyde represents the intermediate stage of oxidation (ox) (or reduction [red]) between an alcohol and a carboxylic acid:

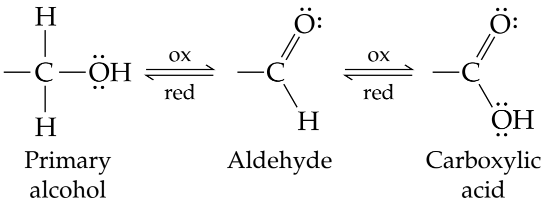

It is not surprising then, to find a reaction in which a specific type of aldehyde will undergo an oxidation–reduction sequence to form the corresponding alcohol and carboxylic acid. Such a reaction is the **Cannizzaro reaction.** In the presence of hydroxide ion, aldehydes that lack (acidic) *alpha (α)-hydrogen* atoms undergo a self-oxidation–reduction reaction. Thus, under the influence of strong base, one molecule of the aldehyde reduces a second molecule of aldehyde to the primary alcohol, and, in the process, is itself oxidized to the corresponding carboxylate anion. Aldehydes with α-hydrogen atoms, however, do not undergo this reaction because in the presence of base these α-cabon atoms are deprotonated. The resulting enolate generally leads to an aldol reaction (see Experiments [20] and [A3$_a$]).

The first step in the mechanism is the nucleophilic attack of the hydroxide anion on the carbonyl group of the aldehyde. This attack is followed by the key step in the reaction, the transfer of a hydrogen atom with its pair of electrons (a hydride ion) to the carbonyl group of a second molecule of aldehyde:

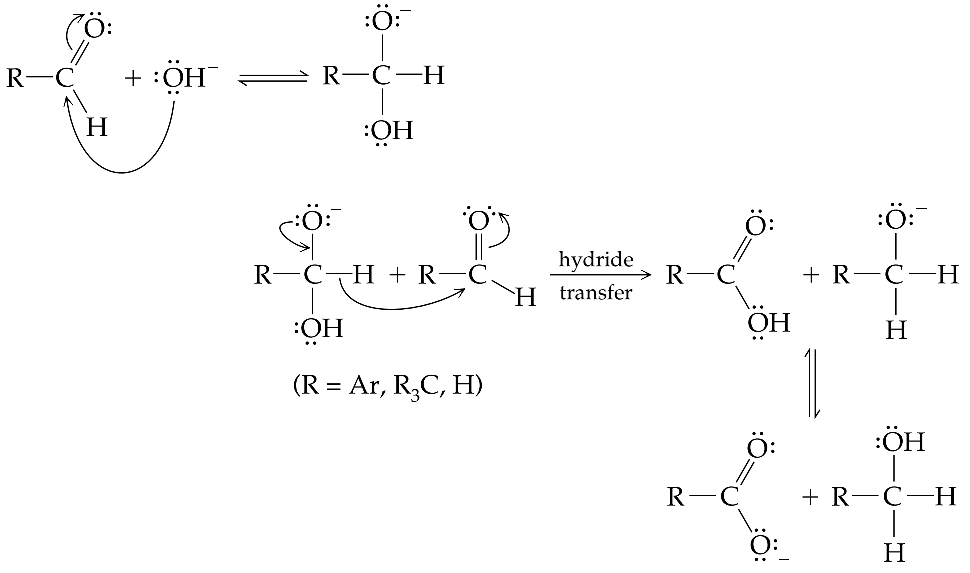

The strong electron-donating character of the negatively charged oxygen atom in the first intermediate anion greatly facilitates the ability of the aldehydic hydrogen to be transferred as a hydride ion, that is, *with* its pair of electrons (:H^-) to a second molecule of aldehyde. As seen in the preceding diagram, this nucleophilic attack on the carbon of a carbonyl group leads to the formation of a carboxylic acid and an alkoxide anion. The equilibrium established in the final step lies far to the right, and involves a fast acid–base reaction that yields both an alcohol and a carboxylate anion. *Thus, even though the Cannizzaro reaction is an equilibrium reaction, it proceeds nearly to completion.* The proposed mechanism is supported by evidence obtained by carrying out the reaction in D_2O. It was found that under these conditions the product alcohol did not contain any α-deuterium substitution. This result suggests that the transferred hydride ion must come from the aldehyde group and not from the solvent.

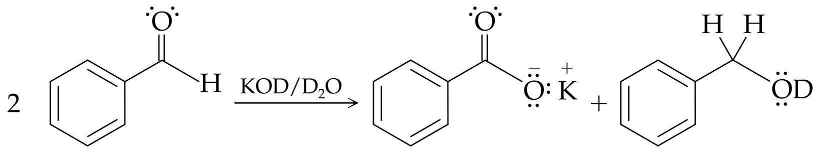

Since few aldehydes lack α-hydrogen atoms, the Cannizzaro reaction is of limited use in modern synthetic sequences, although a variation called the *crossed* Cannizzaro is occasionally used for the reduction of aromatic aldehydes. In this reaction excess formaldehyde is mixed with the aromatic aldehyde and becomes preferentially activated by base since it is present in excess. Thus, the aldehyde present to the lesser extent is reduced to the primary alcohol in high yield. For example,

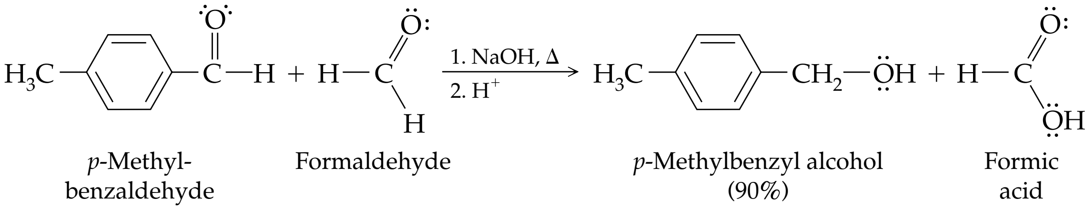

In addition, it is known that α-ketoaldehydes undergo an internal Cannizzaro reaction to yield α-hydroxy carboxylic acids:

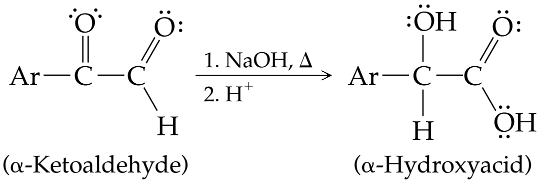

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: two laboratory periods.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	n_D
4-Chlorobenzaldehyde	140.57	150 mg	1.1	47.5		
Methanol	32.04	400 μL			65	1.3288
Potassium hydroxide (11 M)	56.11	550 μL	6.05			

JWCL196_c06_115-187.qxd 11/23/09 10:42 AM Page 178

EQA

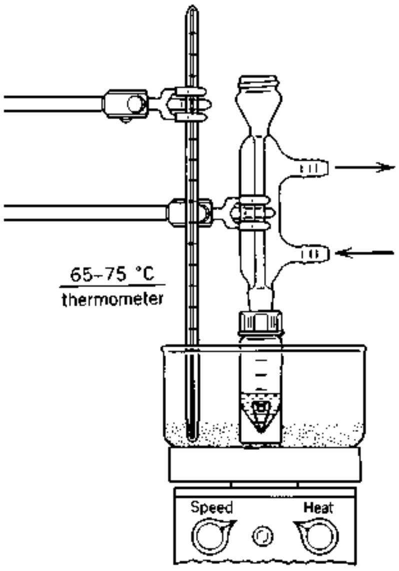

NOTE. This reaction may be carried out in a centrifuge tube containing a boiling stone. The tube should be covered by a piece of filter paper held in place by a rubber band.

Reagents and Equipment. Weigh and add to a 5.0-mL conical vial containing a magnetic spin vane and equipped with a reflux condenser, 150 mg (1.1 mmol) of 4-chlorobenzaldehyde followed by 0.4 mL of methanol (⬅). With gentle swirling, add 0.55 mL of a 11 M aqueous solution of potassium hydroxide.

NOTE. It is convenient to dispense the methanol and KOH solution using automatic delivery pipets. The glass equipment should NOT be rinsed or cleaned with acetone followed by air drying. The residual acetone undergoes an aldol reaction that forms high-boiling contaminants, which interfere with the isolation of the desired products. Also, because ***strong alkali*** *is being used in the reaction, it is important to lightly grease the ground joint of the condenser.*

CAUTION: The concentrated methanolic KOH solution is very caustic. Do not allow it to come into contact with the skin or eyes.

Reaction Conditions. Once the reaction mixture reaches the temperature range of 65–75 °C for a period of 30 min, begin monitoring the reaction progress by TLC. Using a solvent system of ethyl acetate:hexane (1:4) and visualization by a UV lamp, the R_f values of 4-chlorobenzalcohol and 4-chlorobenzoic acid are 0.4 and 0.3 respectively (R_f of 4-chlorobenzaldehyde is 0.6).

NOTE. Initially the mixture is very thick. It will not stir well until the reaction temperature is reached.

Isolation and Purification. When the reaction is complete as judged by TLC, cool the reaction mixture to room temperature and add 2.0 mL of chilled distilled water. Extract the resulting solution with three 0.5-mL portions of methylene chloride, using a Pasteur filter pipet to transfer the extracts to a 3.0-mL conical vial. On each addition of methylene chloride, cap the vial, shake gently, and then carefully vent by loosening the cap. A Vortex mixer, if available, is convenient for this extraction step. After separation of the layers, remove the lower methylene chloride layer using a Pasteur filter pipet.

IMPORTANT. Save both the alkaline and methylene chloride phases for further workup.

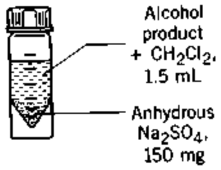
Alcohol product + CH_2Cl_2, 1.5 mL
Anhydrous Na_2SO_4, 150 mg

1. 4-Chlorobenzyl alcohol. Wash the combined methylene chloride extracts with two 0.25-mL portions of saturated sodium bicarbonate solution followed by one 0.5-mL portion of distilled water. Remove the aqueous upper phase (Pasteur filter pipet) used in each washing step and save this combined material in a separate 10-mL Erlenmeyer flask. (*This material will be discarded at the end of the experiment [see Rule 10 for the Microscale Laboratory, Chapter 1, p. 4.]*) Dry the methylene chloride layer over 150 mg of granular anhydrous sodium sulfate (⬅).

After drying the solution, transfer it by use of a Pasteur filter pipet to a tared 3.0-mL conical vial. Be particularly careful not to let any granules of the hydrated sodium sulfate adhere to the surface of the pipet and become transferred with the dried organic phase. Rinse the sodium sulfate drying agent with 0.3 mL of fresh methylene chloride, and combine the rinse with the dried organic phase. Evaporate the methylene chloride using a stream of dry nitrogen

CONFIRMING PAGES aptara

gas in a warm sand bath in the **hood.** If formation of the alcohol took place during the reaction, crude 4-chlorobenzyl alcohol will remain in the vial following removal of the solvent.

HOOD

To purify the crude alcohol, recrystallize the residue using a solution of 4% acetone in hexane (0.25 mL). Collect the product under reduced pressure using a Hirsch funnel, and wash the filter cake (*product crystals packed on the funnel are often referred to as filter cake*) with 0.2 mL of ice cold hexane to give the desired 4-chlorobenzyl alcohol. Air-dry the product on a porous clay plate or on filter paper.

Weigh the 4-chlorobenzyl alcohol and calculate the percent yield. Determine the melting point and compare your value with that found in the literature. Obtain the IR spectrum and compare it with that in Figure 6.16.

This point in the procedure is a logical place to divide the experiment into two laboratory sessions, if this seems appropriate. A second melting point may be obtained at the beginning of the next laboratory period after the sample of aromatic

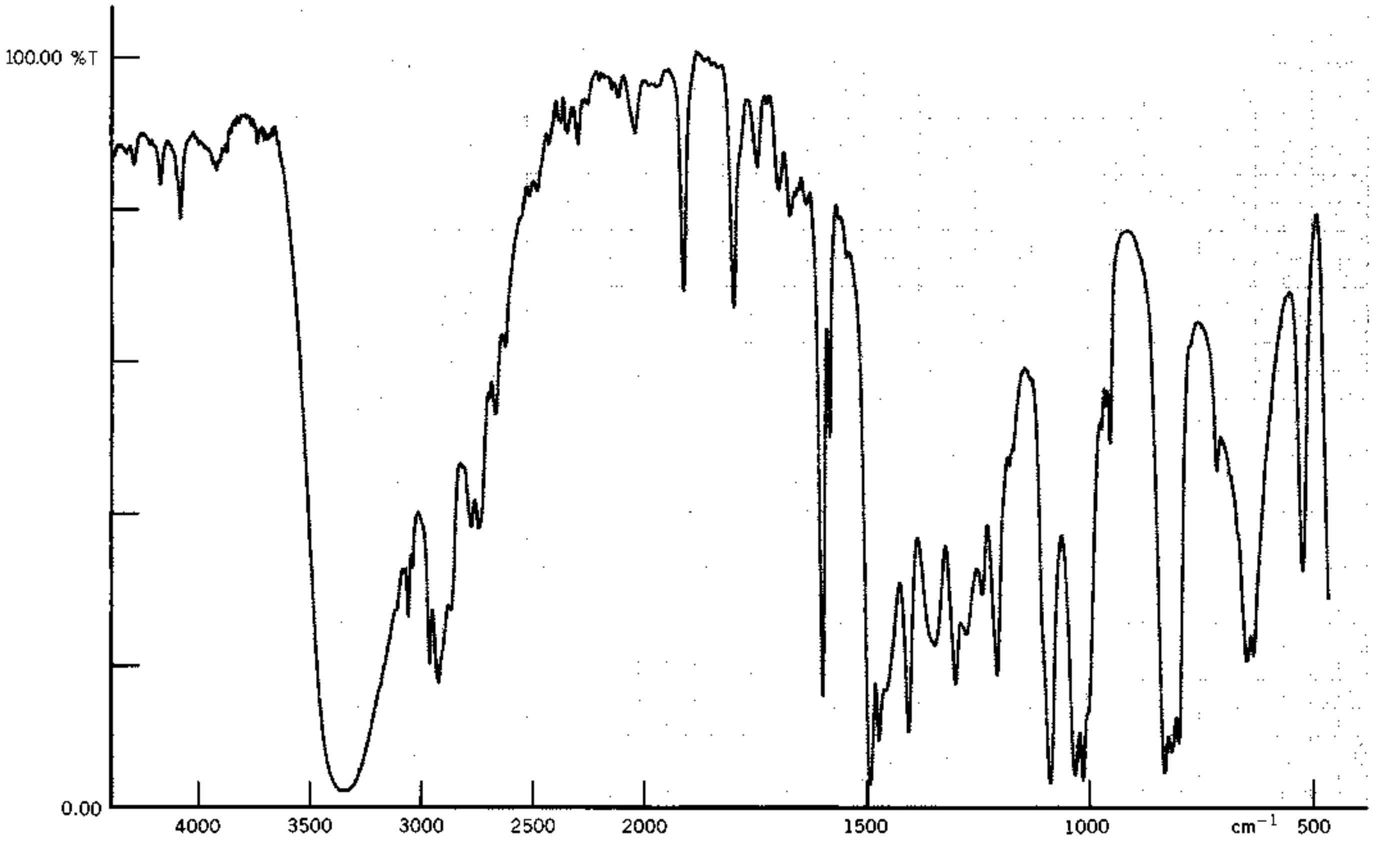

Sample 4-Chlorobenzyl alcohol
%T X ABS __ Background Scans 4
Acquisition & Calculation Time 42 sec
Sample Condition solid - melt
Cell Path Length capillary film
Scans 16
Resolution 4.0 cm-1
Cell Window KBr
Matrix Material

Figure 6.16 IR spectrum: 4-chlorobenzyl alcohol.

alcohol has had additional time to air dry. A convenient way to let the sample dry for several days is to transfer the crystals from the porous plate and spread them over the bottom of a small Erlenmeyer flask (either 5 or 10 mL). Then, cover the mouth of the flask with filter paper held in place by a rubber band. This procedure will allow the last traces of moisture and solvent to evaporate, and at the same time protect the sample from dirt and dust particles. This is a good technique for drying most crystalline samples. If moisture is a particularly difficult problem, the sample stored in this fashion can be placed in a desiccator in the presence of a hygroscopic material, such as calcium chloride.

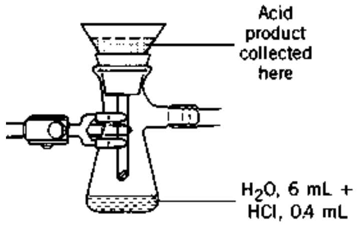

2. 4-Chlorobenzoic acid. Dilute the alkaline phase, obtained and saved during the original extraction procedure, by adding 2.0 mL of water. The dilute aqueous phase is then acidified by the addition of 0.4 mL of concentrated hydrochloric acid. Collect the voluminous white precipitate of the product that forms on addition of the acid, under reduced pressure by use of a Hirsch funnel. Rinse the filter cake with 2.0 mL of distilled water (⬅).

NOTE. It is advisable to check the pH of the solution to insure that it is acidic. Blue litmus or pH paper may be used. Remove a small drop of solution using a glass stirring rod and touch the rod to the paper. ***Never*** *place the paper into the reaction solution.*

Air dry the solid product on a porous clay plate or on filter paper to obtain the crude 4-chlorobenzoic acid product. If a white precipitate is not obtained upon acidification, add a small amount of a saturated sodium chloride solution (salting out technique [see Technique 4, p. 79]) to aid the process.

To purify the crude acid, recrystallize with methanol in a Craig tube. Weigh the dried material and calculate the percent yield. Determine the melting point and compare your value to that found in the literature. Obtain the IR spectrum in a potassium bromide disk and compare your spectrum with that in Figure 6.17.

Infared Analysis. The molecule is 4-chlorobenzaldehyde. The infrared spectrum of this aromatic aldehyde (Fig. 6.18, page 182) is rich and interesting. The aromatic aldehyde macro group frequency train (see Strategies for Interpreting Infrared Spectra) consists of peaks near 3090, 3070, 2830 and 2750, 1706, 1602, 1589, and 1396 cm^{-1}.

a. 3070 cm^{-1}: a C—H stretch on sp^2 carbon

www

b. 2840 and 2740 cm^{-1}: This pair of bands is a famous example of powerful Fermi coupling (see Chapter 8 and IR discussions). The unperturbed C—H stretching of the aldehyde C—H group would be expected to occur near 2790 cm^{-1} (the C—H stretch on an sp^2 carbon would have been expected to be found at higher values, but the oxygen system in this case significantly lowers the observed values). The in-plane bending mode of this C—H group occurs at 1390 cm^{-1}. Thus, the first harmonic should fall near 2780 cm^{-1}, very close (~10 cm^{-1}) to the stretching frequency of this oscillator. The essential conditions are, therefore, satisfied and strong Fermi coupling occurs and gives rise to the split peaks at 2840 and 2740 cm^{-1}. The latter band is moved well away from the normal sp^3 C—H symmetric stretching modes. Thus, the lower

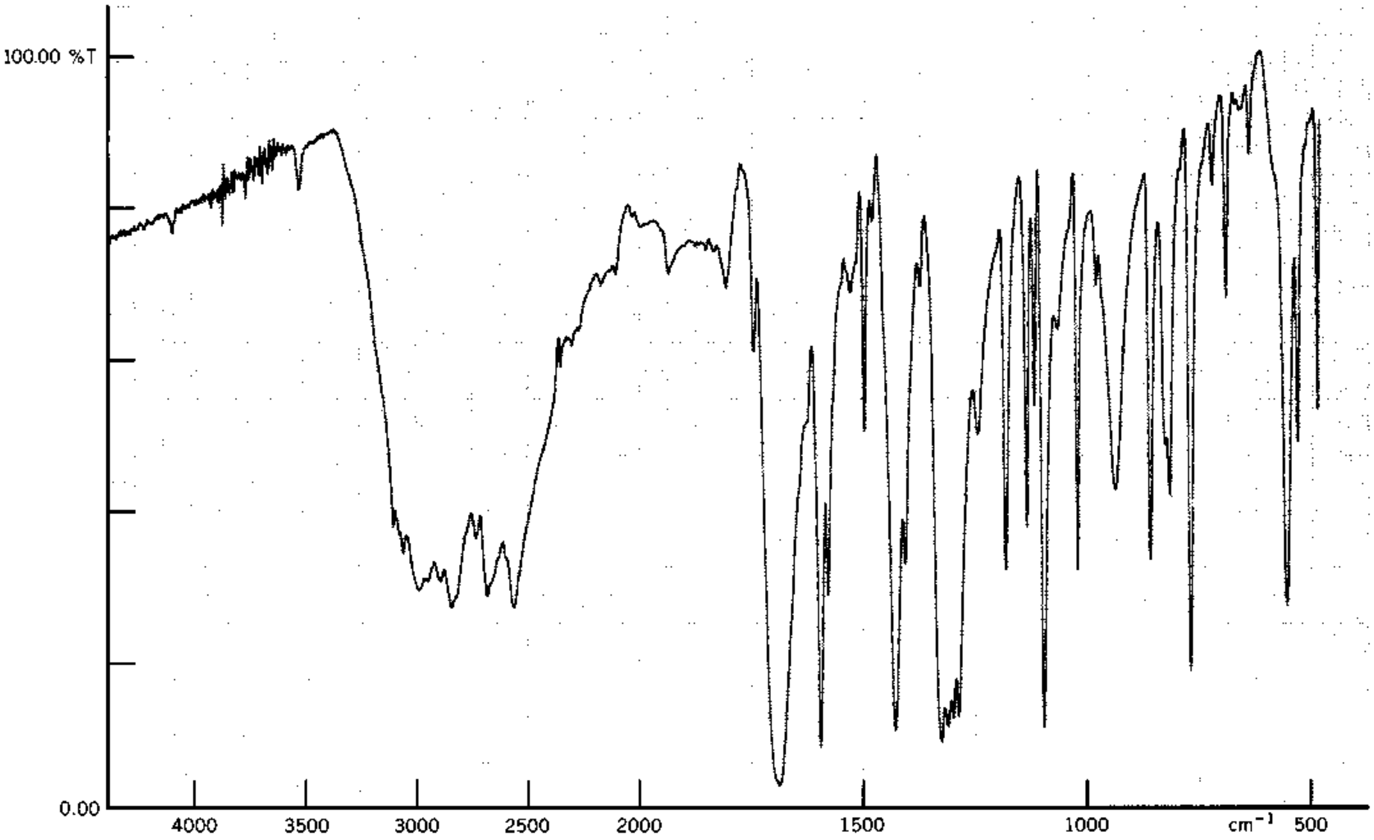

Sample 4-Chlorobenzoic acid

%T x ABS — Background Scans 4

Acquisition & Calculation Time 42 sec

Sample Condition solid

Cell Path Length

Scans 16

Resolution 4.0 cm⁻¹

Cell Window

Matrix Material KBr

Figure 6.17 IR spectrum: 4-chlorobenzoic acid.

wavenumber component of the Fermi-coupled aldehyde C—H mode leads to easy identification of an aldehyde, even when it represents a very small fraction of an aliphatic system. In the present case where no aliphatic C—H oscillators are present, both components are observed.

c. 1704 cm^{-1}: The carbonyl stretch of the aldehyde group. The frequency observed in aliphatic aldehydes falls in the range 1735–1720 cm^{-1}, but when conjugated, the value drops 15–25 cm^{-1} (see Chapter 8 and online IR discussions) and is found in the range 1720–1700 cm^{-1}.

d. 1601 and 1578 cm^{-1}: This pair of bands is related to the degenerate ring stretching vibrations v_{8a} and v_{8b} of benzene (see also Infrared discussions in Experiment [1B$_{adv}$]).

e. 1390 cm^{-1}: The aldehyde C—H in-plane bending vibration. The first harmonic of this vibration is Fermi coupled to the aldehyde C—H stretching mode, as discussed earlier.

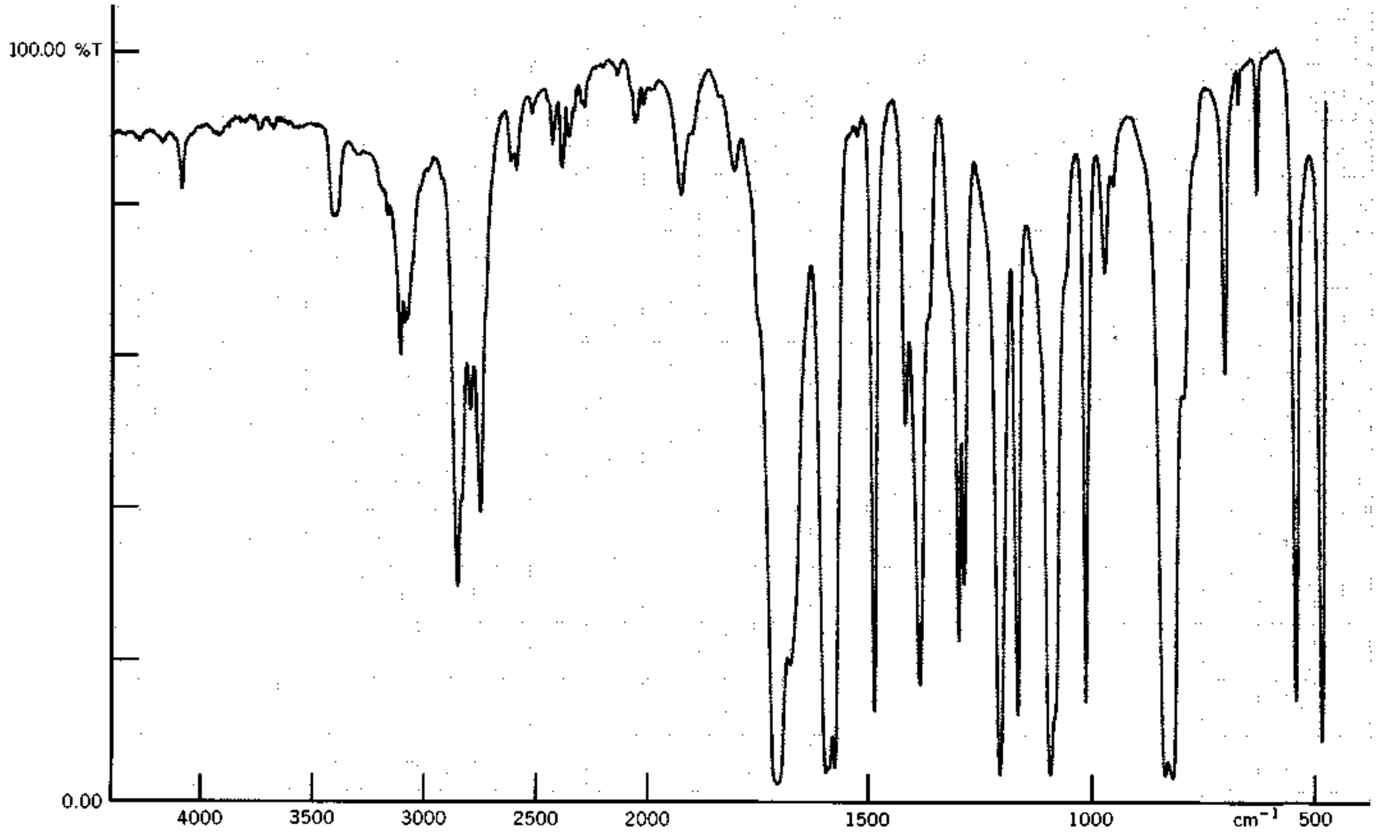

Sample 4-Chlorobenzaldehyde
%T X ABS — Background Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm-1
Sample Condition solid - melt
Cell Window KBr
Cell Path Length capillary film
Matrix Material

Figure 6.18 IR spectrum: 4-chlorobenzaldehyde.

This compound also possesses a second powerful macro group frequency train that assesses the substitution pattern of the aromatic ring system (see Chapter 8, IR discussions). The para-disubstituted benzene ring macro group frequency train requires peaks in the following regions: 1950, 1880, 1800, 1730, 750, and 690 cm^{-1}.

www→

a. 1905, 1795 cm^{-1}: This pair of weak bands, with the higher wavenumber band more intense than the lower member, arises from combination bands (see Chapter 8, IR discussions) which involve the out-of-plane bending frequencies of the ring C—H bonds (see below). The exact wavenumber positions are not very important, but the overall shape of the pattern can be used to determine the ring substitution pattern (see Chapter 8 and online IR discussions).

www→

b. 832 cm^{-1}: This strong band is very characteristic of para-disubstituted benzene rings. The 832-cm^{-1} peak arises from the in-phase out-of-plane bending vibration of the two pairs of C—H groups on opposite sides of the six-membered ring (see Chapter 8, and online IR discussions).

www→

The C—Cl stretching mode when substituted on an aromatic ring often becomes heavily coupled with ring vibrations and, as in this case, does not give rise to an identifiable group frequency. The presence of this group must be determined by other methods, such as a Beilstein or sodium fusion test (see Chapter 9, pp. 634–638). The presence of a hidden group (very likely halogen), however, is strongly indicated. Because the system must be para substituted (see above paragraphs), one of the substituents on the ring must exhibit very little identifiable absorption from 4000 to 500 cm^{-1}.

The reaction involves the formation of two products, one neutral and one acidic, both of which incorporate large portions of the substrate molecule. These materials can be characterized by their IR spectra. The neutral product is proposed to be 4-chlorobenzyl alcohol. An examination of the spectrum (Fig. 6.16) supports the presence of two macro group frequency trains. One (a) is the same one found in the starting aldehyde, the para-disubstituted benzene ring frequency train. The other (b) is a primary aliphatic alcohol macro frequency train. Macro (a) has been expanded to include all aromatic ring-specific group frequencies, as the alcoholic side chain is decoupled from the ring and the carbon–chlorine frequencies fall outside the range of the instrumentation normally used in these measurements.

The macro group frequencies for 4-chlorobenzyl alcohol are as follows:

a. 3055 (aromatic C—H stretch), 1906 and 1793 (para combination band pattern), 1596 and 1583 (ring stretch degenerate pair, ν_{8a} and ν_{8b}; 1583 intensity indicates weak conjugation with the ring), 1495 and 1475 (ring stretch degenerate pair, ν_{19a} and ν_{19b}), 834 (C—H, out-of-plane bend) cm^{-1}.

b. 3340 (broad, O—H stretch), 2965–2925 (C—H, aliphatic sp^3), 1450–1300 (broad, O—H bend, associated), 1015 (C—O, stretch, primary alcohol), 630 (broad, weak O—H bend, associated) cm^{-1}.

The acidic product generated in the reaction is assumed to be 4-chlorobenzoic acid. This material also is a para-substituted aromatic compound; thus, the spectrum of this compound (Fig. 6.17) possesses a macro frequency train (a), similar to those of the aldehyde and alcohol. In addition, this benzoic acid derivative exhibits an extended aromatic acid macro group frequency train (b). The macro frequencies are as follows:

a. 1935 and 1795 (para combination band pattern), 852 (ring C—H, in-phase, out-of-plane bend) cm^{-1}.

b. 340–2200 (very broad, very strong, O—H stretch, associated acid), 1683 (C—O, stretch, out-of-phase, associated acid dimer), 1596 and 1578 (degenerate ring stretch), 1500 and 1432 (degenerate ring stretch), 1320–1280 (C—O, stretch), 885 (O—H, out-of-plane, ring dimer bend)cm^{-1}.

The carbon–chlorine stretch is not observed.

Examine the spectra of the reaction products you have obtained, in a potassium bromide matrix. Discuss the similarities and differences of the experimentally derived spectral data to the reference spectra (Figs. 6.16–6.18).

It is a useful exercise to compare the results of organic qualitative analysis reaction tests, which were used historically to classify the preceding compounds, with the data now available from modern spectroscopic instrumentation.

Chemical Tests. Perform each of the following tests (see Chapter 9). Do the results confirm that you have isolated an aromatic carboxylic acid and an aromatic alcohol?

1. The ignition test.
2. The Beilstein or the sodium fusion test for the presence of halogen.
3. The ceric nitrate and/or the Jones oxidation test for the alcohol.
4. The solubility of the acid in sodium bicarbonate and sodium hydroxide solutions. Is carbon dioxide evolved in the bicarbonate test?

If you were to prepare a derivative for the alcohol and acid products, which one would you choose? See Chapter 9, Preparation of Derivatives.

Experiment 7-1*

M

4-Chlorobenzoic Acid and 4-Chlorobenzyl Alcohol: Preparation Using a Monomode Microwave Apparatus

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	n_D
4-Chlorobenzaldehyde	140.57	150 mg	1.1	47		
Methanol	32.04	0.4 mL			65	1.3288
Potassium hydroxide (11 M)	56.11	0.4 mL				

Reagents and Equipment. This experiment is designed for use in the CEM Discover and Biotage Initator microwave units.

In a 10.0-mL glass microwave reaction vessel containing a magnetic stir bar, place 150 mg (1.1 mmol) of 4-chlorobenzaldehyde followed by 0.4 mL of methanol. With gentle swirling, add 0.4 mL of an 11 M aqueous solution of potassium hydroxide.

HOOD

CAUTION: Dispense liquid reagents in the **hood** using automatic delivery pipets. Concentrated methanolic potassium hydroxide is *highly caustic*. Since the reaction requires heating reagents to above their boiling point in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential*.

Reaction Conditions. Place the reaction vessel in the microwave cavity and, depending on the equipment used, position the pressure device on top. Program the microwave unit to heat the reaction mixture to 120 °C using no more that 50 W of microwave power, and hold at this temperature for 1 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the tube from the microwave unit.

---

* This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

Isolation of Product. Add 2.0 mL of cold distilled water to the reaction vessel. Swirl or stir on a stirplate to dissolve the solid material. Transfer the reaction mixture with a Pasteur pipet into a 5.0-mL conical vial. Rinse the microwave reaction vessel with 0.5 mL of dichloromethane and add the washings to the conical vial. Cap the vial, shake gently, and then carefully vent by loosening the cap. After allowing separation of the layers, transfer the bottom (organic) layer to a 3.0-mL conical vial with a Pasteur filter pipet. Extract the aqueous solution in the 5-mL conical vial with two additional 0.5 mL portions of dichloromethane. After each addition carefully cap the vial, shake gently, and vent the solution by loosening the cap. On each occasion, transfer the bottom (organic) layer to the 3.0-mL conical vial with a Pasteur filter pipet. At the end of the extractions, label the 5.0-mL conical vial containing the aqueous alkaline layer "alkaline phase." This phase will be used later in the experiment.

IMPORTANT. Save both the aqueous and dichloromethane layers for further workup.

The remainder of the work-up for the two products is the same as for Experiment 7 (pp. 176–184).

4-Chlorobenzoic Acid and 4-Chlorobenzyl Alcohol: Preparation Using a Multimode Microwave Apparatus

Experiment 7-2*

M

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	n_D
4-Chlorobenzaldehyde	140.57	465 mg	3.3	47		
Methanol	32.04	2.5 mL			65	1.3288
Potassium hydroxide (11 M)	56.11	2.5 mL				

Reagents and Equipment. This experiment is designed for use in the CEM MARS, Milestone START, and Anton Paar Synthos 3000 microwave units. When using the Anton Paar Synthos 3000 unit with the 24-position silicon carbide plate rotor containing glass vials, the reagent and solvent quantities cited in the monomode procedure should be used in conjunction with the reaction conditions here in the multimode procedure.

In a microwave reaction vessel containing a magnetic stir bar, place 465 mg (3.3 mmol) of *p*-chlorobenzaldehyde and 2.5 mL of methanol. With stirring, slowly add 2.5 mL of an 11 M aqueous solution of potassium hydroxide. Cap the vessel with the microwave pressure cap and adjust the tightness to the manufacturer-specified level. Place the sealed vessel into its outer protective jacket.

---

*This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

HOOD

CAUTION: Dispense liquid reagents in the **hood** using automatic delivery pipets. Concentrated methanolic potassium hydroxide is *highly caustic*. Since the reaction requires heating reagents to above their boiling point in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential*.

Reaction Conditions. Insert the loaded vessels into the reaction carousel ensuring they are evenly spaced and then place the carousel into the microwave cavity. If provided by the manufacturer, connect a temperature probe to the control vessel. Program the microwave unit to heat the reaction vessels to 120 °C and hold at this temperature for 1 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the carousel from the microwave unit.

Isolation or Product. Add 5 mL of distilled water to the reaction vessel to dissolve any solids, transfer the solution to a 60-mL separatory funnel, and clamp the funnel to a ring stand. Rinse the microwave reaction vessel with an additional 5 mL of distilled water and add it to the separatory funnel. Extract the alkaline solution with three 3.0-mL portions of dichloromethane. After each addition of dichloromethane carefully cap and invert the funnel. Immediately vent the funnel by opening the stopcock. Close the stopcock, mix the two layers several times by inverting the funnel, repeatedly vent, and then place the funnel back on the ring stand and remove the stopper. Drain the lower (organic) layer into a 25-mL Erlenmeyer flask after each extraction. At the end of the extractions, pipet the alkaline layer into a separate 25-mL Erlenmeyer flask and label it "alkaline phase." (*This phase will be used in part 2.*)

IMPORTANT. Save both the aqueous and organic layers for further workup.

1. 4-chlorobenzyl alcohol. Pipet the organic layer back into the separatory funnel. Wash the combined organic layers with two 5-mL portions of saturated sodium bicarbonate solution. Remove the lower (organic) layer by draining it through the stopcock into a 25-mL Erlenmeyer flask and pipet the remaining aqueous phase into a 25-mL Erlenmeyer flask after each washing. Save the aqueous waste until the experiment is complete and then discard as directed. Return the organic layer to the separatory funnel by pipet. Wash the organic layer once with 5 mL of distilled water. Dry the lower (organic) layer by draining it into a clean 25-mL Erlenmeyer flask containing 200 mg of anhydrous sodium sulfate. Transfer the anhydrous dichloromethane solution, using a Pasteur filter pipet, to a clean tared 10-mL pear-shaped flask. Remove the dichloromethane on a rotary evaporator or by evaporation in the **hood** using a gentle stream of nitrogen gas with warming in a sand bath. A white precipitate will remain in the flask. Reweigh the flask plus product and calculate the crude yield. HOOD

To further purify the crude alcohol, recrystallize the residue using a solution of 4% acetone in hexane (0.75 mL). Collect the product under reduced pressure using a Hirsch funnel, and wash the filter cake with 0.5 mL of ice cold hexanes. Air-dry the product on a porous clay plate.

Weigh the 4-chlorobenzyl alcohol and calculate the % yield. Determine the melting point of the product and compare it to the literature value. Obtain the IR spectrum and compare it with that in Figure 6.16.

2. 4-chlorobenzoic acid. Cool the "alkaline phase" in an ice bath and add dropwise 4 mL of 6M aqueous hydrochloric acid. A white precipitate of the product will form. Allow the solution to chill for a further 10 minutes and then collect the product under reduced pressure using a Hirsch funnel. Rinse the filter cake with 2.0 mL of cold distilled water.

Air dry the solid on a porous clay plate or on filter paper to obtain the crude 4-chlorobenzoic acid. If the product is very pasty it may be left to dry overnight. The product can be further purified by recrystallization with methanol. Weigh the dried material and calculate the % yield. Determine the melting point of your product and compare your value to the literature value. Obtain the IR spectrum and compare it to the spectrum given in Figure 6.17.

QUESTIONS

6-45. The discussion mentions that the crossed Cannizzaro reaction can be realized when one of the components is formaldehyde. Predict the product(s) of the reaction below and give a suitable name to the reactants and products.

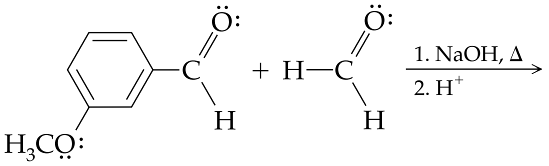

6-46. One group of investigators has suggested that a dianion might be the source of hydride in the Cannizzaro reaction. Explain why this species would be a better source of hydride than the species in the mechanism presented in the discussion.

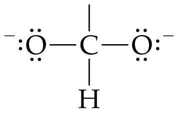

$$^{-}{:}\ddot{\underset{\cdot\cdot}{O}}-\underset{\displaystyle H}{\overset{|}{\underset{|}{C}}}-\ddot{\underset{\cdot\cdot}{O}}{:}^{-}$$

6-47. The Cannizzaro reaction is an oxidation–reduction sequence. What type of reagent is formaldehyde acting as in Question 6.45?

6-48. Propose a mechanism for the internal Cannizzaro reaction depicted below.

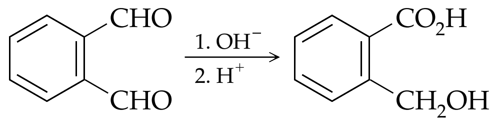

o-Phthalaldehyde

o-(Hydroxymethyl) benzoic acid

BIBLIOGRAPHY

Cannizzaro, S. *Annalen* **1853,** *88,* 129.

Reviews on the Cannizzaro reaction:

Bergens, S. H.; Fairlie, D. P.; Bosnich, B. *Organometallics* **1990,** *9,* 566.

Geissman, T. A. *Org. React.* **1944,** *2,* 94.

McDonald, R. S.; Sibley, C. E. *Can. J. Chem.* **1981,** *59,* 1061.

Swain, C. G.; Powell, A. L.; Sheppard, L. A.; Morgan, C. R. *J. Am. Chem. Soc.* **1979,** *101,* 3576.

Examples of the Cannizzaro Reaction:

Davidson, D.; Weiss, M. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 590.

Esteb, J. J.; O'Reilly, S.; Richter, J. M. *J. Chem. Educ.* **2004,** *81,* 1794.

Wilson, W. C. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 256.

EXPERIMENT 8

The Esterification Reaction: Ethyl Laurate, Isopentyl Acetate, and the Use of Acidic Resins

Common names: ethyl laurate, ethyl dodecanoate
CA number: [106-33-2]
CA name as indexed: dodecanoic acid, ethyl ester

Common names: isopentyl acetate, isoamyl acetate
CA number: [123-92-2]
CA name as indexed: 1-butanol, 3-methyl-, acetate

Common name: butyl acetate
CA number: [123-86-4]
CA name as indexed: acetic acid, butyl ester

Common names: pentyl acetate, amyl acetate
CA number: [628-63-7]
CA name as indexed: acetic acid, pentyl ester

Common name: hexyl acetate
CA number: [142-92-7]
CA name as indexed: acetic acid, hexyl ester

Common name: octyl acetate
CA number: [112-14-1]
CA name as indexed: acetic acid, octyl ester

Prior Reading

Technique 4: Solvent Extraction
- Liquid–Liquid Extraction (p. 72)
- Drying of the Wet Organic Layer (pp. 80–83)

Technique 6A: Chromatography
- Packing the Column (p. 93)
- Elution of the Column (p. 94)

Purpose. This exercise explores the classic reactions of carboxylic acids (RCO_2H) with alcohols ($R'OH$), in the presence of acid catalyst, to yield esters (RCO_2R') plus water (H_2O, a small stable molecule). The physical properties of these esterification products are examined and the techniques of distillation and column chromatography are applied to the purification of these materials.

REACTION

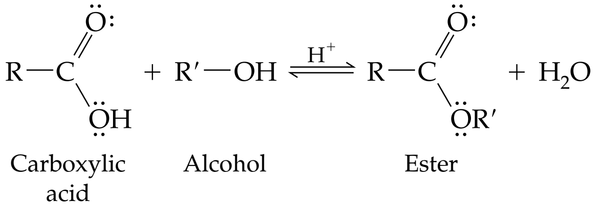

DISCUSSION

Esters are among the most important of the carboxylic acid (and alcohol) derivatives. Substances possessing this functional group are widely distributed in nature in the form of waxes, essential oils, fatty acid esters, and aromas. The ester functionality plays a significant role in biochemistry, both in primary metabolism and in a variety of substances exhibiting remarkable physiological activity in humans (hormones and neurotransmitters). Esters find extensive use in commercial products from fingernail polish remover and artificial sweeteners, to polymeric fibers, plasticizers, and surfactants.

Biosynthesis of Esters. Fatty acids are naturally occurring, long, straight-chain, C_{12}–C_{40} carboxylic acids; most contain an even number of carbon atoms.Their biosynthesis provides an important and interesting example of a primary metabolic pathway in which a special type of ester is the essential link between the enzyme and the substrate (acetic acid). The enzyme-bound substrate grows by repeated addition of two-carbon (C_2) units and, when eventually released from the enzyme, has undergone an extension of the fatty acid hydrocarbon chain.

The first step in fatty acid biosynthesis involves the formation of a thiol ester, acetyl coenzyme A (acetyl CoA), from acetic acid (present in the primary metabolic pool) and the thiol group (mercapto, or —SH group) of the coenzyme (HSCoA). A thiol ester is an ester in which the single-bonded oxygen (from the alcohol component) is replaced by a sulfur atom,

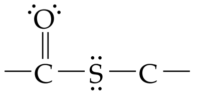

$$-\overset{\overset{\displaystyle \ddot{O}}{\|}}{C}-\ddot{\underset{\cdot\cdot}{S}}-C-$$

from the coenzyme. Coenzymes are loosely bound, nonprotein factors attached to the enzyme that play an important role in the catalytic function of the enzyme. These coenzymes are distinguished, in an ill-defined manner, from prosthetic groups, which are intimately attached to active sites of enzymes. Part of the role of the CoA is to facilitate the transfer of the substrate (the C_2 unit) to a new thiol group of the enzyme (protein), where the next stage of the biosynthesis takes place:

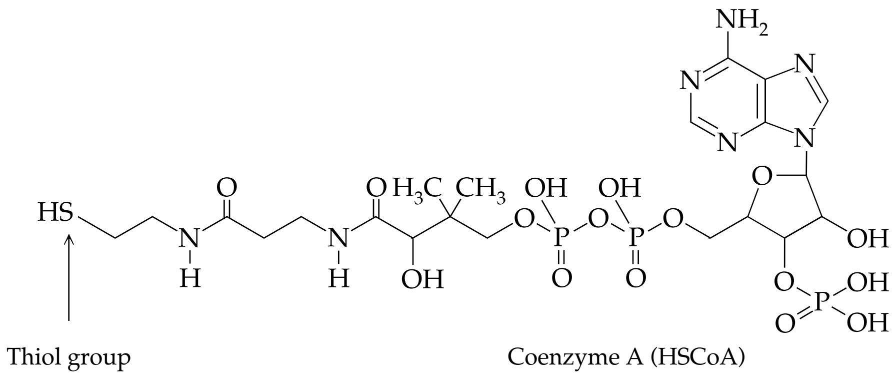

Coenzyme A (HSCoA)

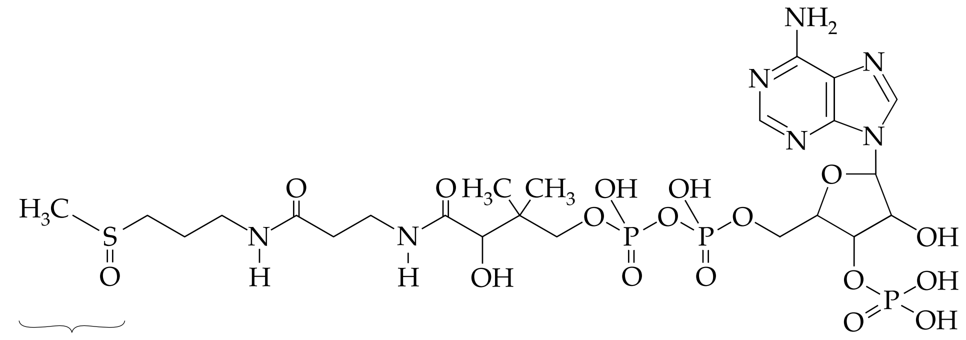

Acetyl thiolester

Acetyl coenzyme A ($CH_3COSCoA$)

This reactive thiol ester is capable of undergoing aldol-type (see Experiment [20]) condensations under physiological conditions. AcetylCoA is first carboxylated with the help of the enzyme, acetyl CoA-carboxylase, to yield a thiolmalonyl derivative. The resulting intermediate possesses an activated methylene group

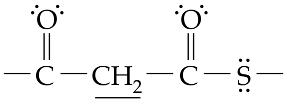

$$-\overset{\ddot{O}}{\overset{\|}{C}}-\underline{CH_2}-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{S}-$$

www

which with further enzymatic support undergoes a Claisen condensation (see Experiment [$3A_{adv}$]) with an acetyl group that has been also transferred via acetyl CoA to an appropriate acyl-carrier protein, fatty acid synthetase. Reduction, dehydration, further reduction, and finally hydrolysis of the thiol ester, yields the fatty acid extended by a C_2 hydrocarbon group.

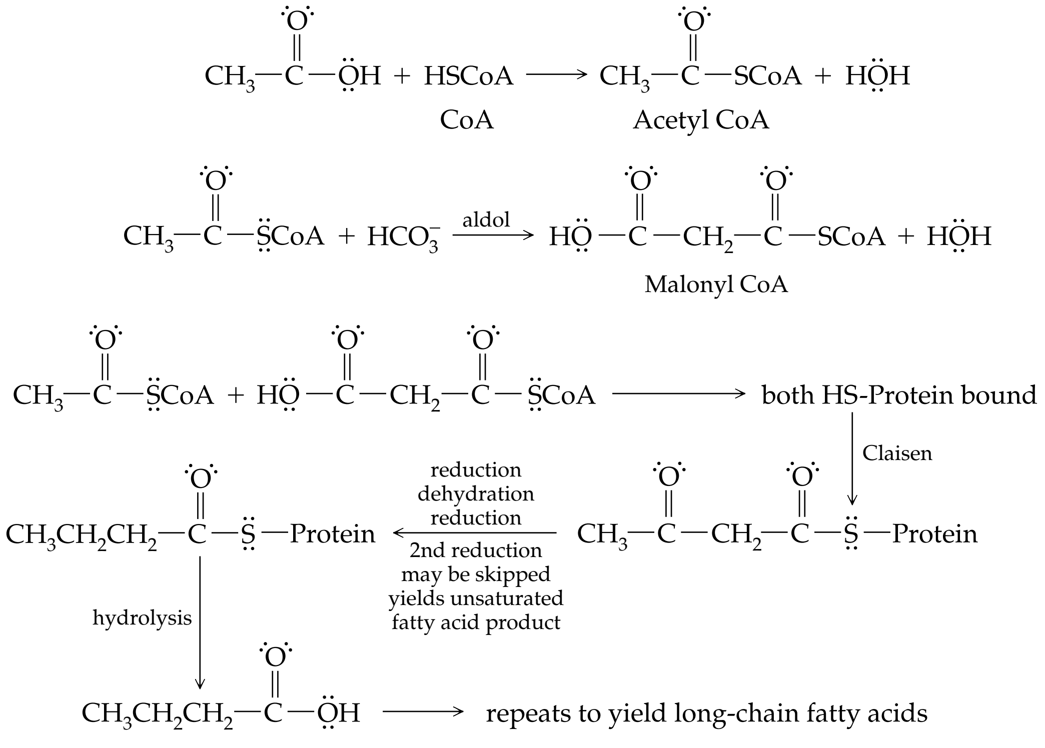

$$\underset{\text{CoA}}{CH_3-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{O}H + HSCoA} \longrightarrow \underset{\text{Acetyl CoA}}{CH_3-\overset{\ddot{O}}{\overset{\|}{C}}-SCoA} + H\ddot{O}H$$

$$CH_3-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{S}CoA + HCO_3^- \xrightarrow{\text{aldol}} \underset{\text{Malonyl CoA}}{H\ddot{O}-\overset{\ddot{O}}{\overset{\|}{C}}-CH_2-\overset{\ddot{O}}{\overset{\|}{C}}-SCoA} + H\ddot{O}H$$

$$CH_3-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{S}CoA + H\ddot{O}-\overset{\ddot{O}}{\overset{\|}{C}}-CH_2-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{S}CoA \longrightarrow \text{both HS-Protein bound}$$

$$\text{both HS-Protein bound} \xrightarrow{\text{Claisen}} CH_3-\overset{\ddot{O}}{\overset{\|}{C}}-CH_2-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{S}-\text{Protein}$$

$$CH_3-\overset{\ddot{O}}{\overset{\|}{C}}-CH_2-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{S}-\text{Protein} \xrightarrow[\substack{\text{2nd reduction}\\ \text{may be skipped}\\ \text{yields unsaturated}\\ \text{fatty acid product}}]{\substack{\text{reduction}\\ \text{dehydration}\\ \text{reduction}}} CH_3CH_2CH_2-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{S}-\text{Protein}$$

$$CH_3CH_2CH_2-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{S}-\text{Protein} \xrightarrow{\text{hydrolysis}} CH_3CH_2CH_2-\overset{\ddot{O}}{\overset{\|}{C}}-\ddot{O}H \longrightarrow \text{repeats to yield long-chain fatty acids}$$

LIPIDS

Oils, Fats, Waxes, and Aromas. Fatty acids derived from primary metabolism play a key role in the formation of naturally occurring oils, fats, and waxes. Fats and oils are esters of these acids with a triol, glycerol ($HOCH_2CHOHCH_2OH$).

The common fats and oils are formed from mixtures of C_4–C_{26} saturated fatty acids with the vast majority derived from C_{12}–C_{18} acids. The oils are more likely to include significant contributions from mixtures of unsaturated fatty acids.

Since fats and oils are triesters of glycerol, they are generally called *triglycerides*. In plants and animals, triglycerides function as energy reserves that can be used in primary metabolism when food (energy) is not available to the organism.

Although the fats have high molecular weights, they are generally found to be very low-melting solids, particularly if they contain unsaturated fatty acids. The bent chains, which result from incorporating cis-alkene (C═C) groups into the chain, prevent close packing in a solid and, as a result, such molecules exhibit lower melting points. For example, compare oleic acid ($C_{18}H_{34}O_2$, mp 4 °C) to its saturated analog, stearic acid($C_{18}H_{36}O_2$, mp 69–70 °C). The former melts more than 60 °C lower because of the cis carbon–carbon double bond.

Oleic acid is the simplest of the unsaturated C_{18} fatty acids, because it has a single C═C group located in the middle of the chain. This unsaturated fatty acid is the most widely distributed of all fatty acids. It is the dominant component (76–86%) of the triglycerides in olive oil. Highly saturated fats, on the other hand, are generally solids at room temperature because the straight-chain fatty acids pack together well (see Fig. 6.19 on next page).

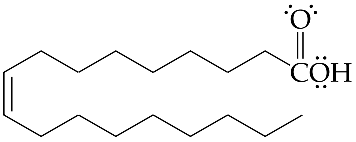

Oleic acid
[(Z)-9-Octadecenoic acid]

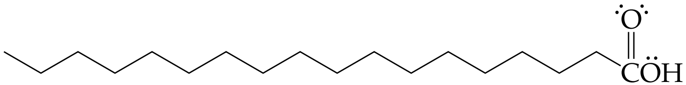

Stearic acid
[Octadecanoic acid]

Many vegetable and fish oils are liquid triglycerides. Because these organisms operate at ambient temperatures, evolution dictated that low-melting fats were required to avoid solidification. In warm-blooded animals, higher melting fats can be tolerated and are used.

The cheap and plentiful unsaturated oils can be converted to solid fats by hydrogenation of the alkene groups, which gives straight-chain alkyl groups. As consumers historically have desired to cook with solid, white, and creamy fats (such as lard) derived from animal triglycerides (low in unsaturation), hydrogenation of vegetable oils, such as peanut, soybean, and cottonseed oils, has been carried out on a large scale (this process is referred to as **hardening** the fat).

Unfortunately, the relationship between saturated fats in the human diet and the formation of cholesterol (a simple lipid, see below) plaque and coronary heart disease has been established. The dietary switch to less saturated fats is currently underway.

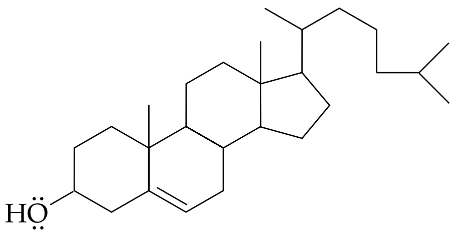

Cholesterol

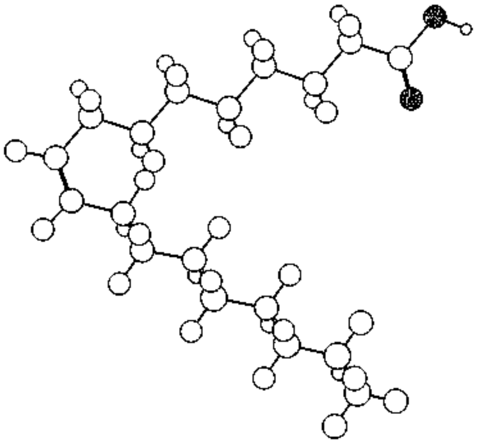

Oleic acid — ball and stick model

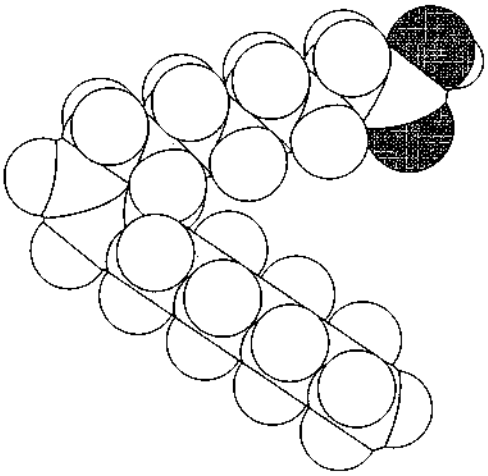

Oleic acid — space-filling model

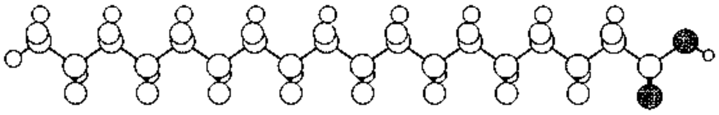

Stearic acid — ball and stick model

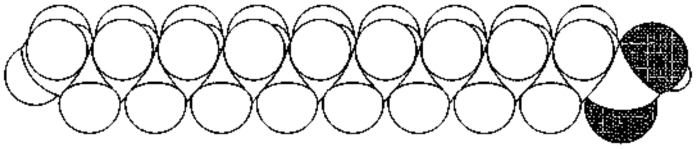

Stearic acid — space-filling model

Figure 6.19 Molecular models of fatty acids.

The triglycerides obtained from animal fats have been used for a very long time as a source of soap. When fats are boiled with lye (sodium hydroxide) the ester linkages are cleaved by a process known as *saponification* (the term originates from the Latin word for soap, *sapon*, as does the modern French word for soap, *savon*) to yield the sodium salt of the fatty acid and the esterifying alcohol (glycerol).

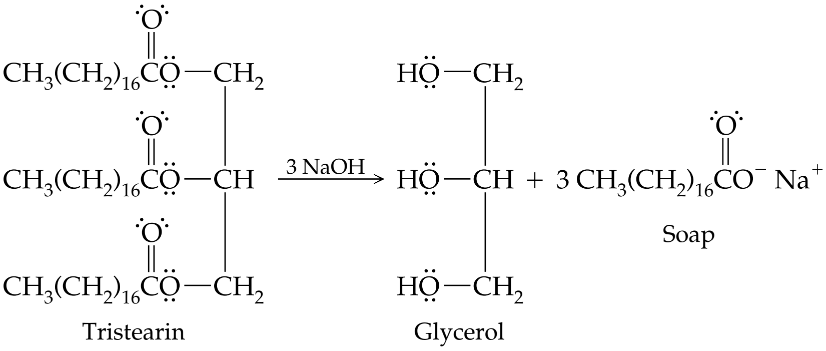

Saponification of a triglyceride found in animal fat

Salts of fatty acids function effectively as soaps because one end of the straight-chain system has the highly polar carboxylate ion and is readily solvated in water. The rest of the fatty acid molecule has all the characteristics of a nonpolar hydrocarbon and readily dissolves in hydrocarbons, such as greases and oils. We refer to the polar end (head) as being hydrophilic (attracted to water) and the hydrocarbon end (tail) as being lipophilic (attracted to oils). When dispersed in an aqueous solution, fatty acids tend to form micelles (spherical clusters of molecules). The lipophilic ends of the fatty acids occupy the interior of the cluster, while the polar ends, which are heavily solvated by water molecules, form the outer surface of the spherical micelle. Micelles absorb the hydrocarbon chains of the triglycerides, and thus soaps break up and help to dissolve the fats and oils that tend to coat skin, clothes, and the surfaces of eating and cooking utensils (see Figs. 6.20 and 6.21).

Waxes are naturally occurring esters of fatty acids (in waxes, chain lengths can reach as high as C_{36}) and a variety of other alcohols that often possess relatively complicated structures (steroid alcohols) and/or long chains. For example, *n*-octacosanol, $CH_3(CH_2)_{26}CH_2OH$, has been isolated from the esters in wheat

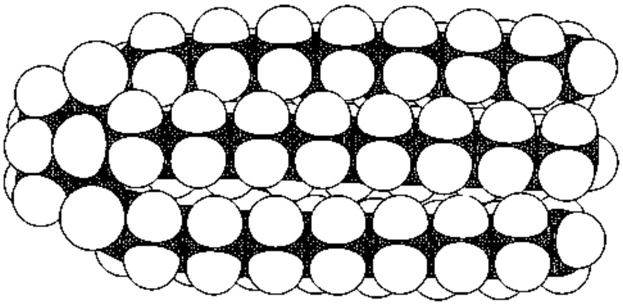

A space-filling model of a saturated triglyceride

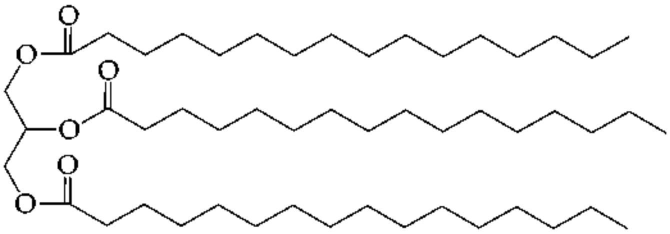

A saturated triglyceride

Figure 6.20 A space-filling model of a saturated triglyceride.

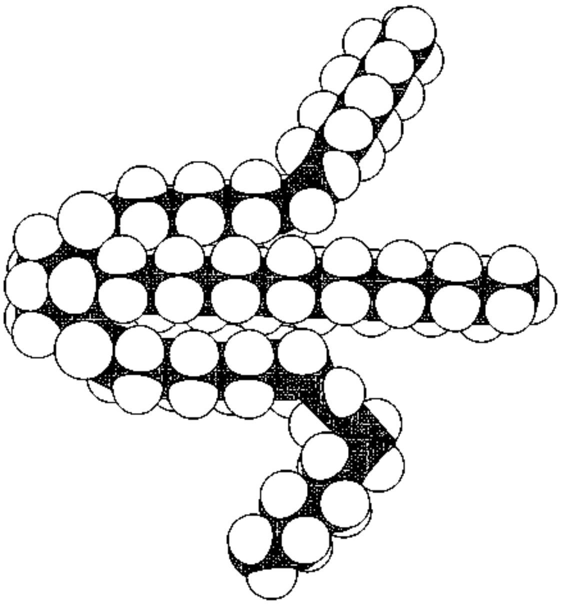

A space-filling model of an unsaturated triglyceride.

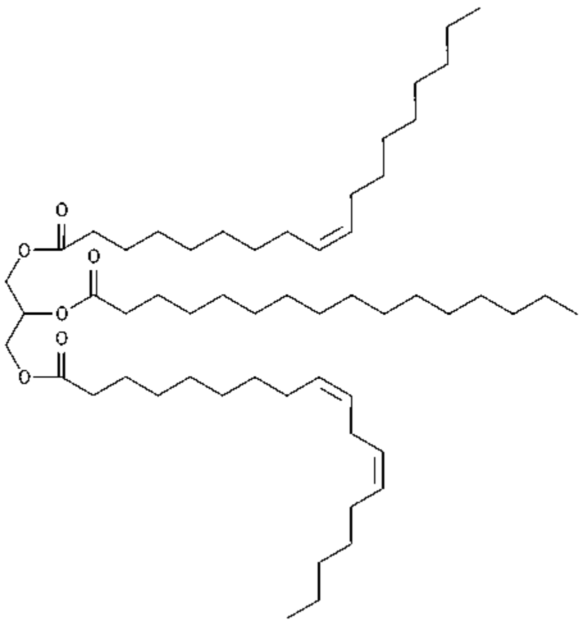

An unsaturated triglyceride

Figure 6.21 A space-filling model of an unsaturated triglyceride.

waxes, and a component of carnauba wax (traditionally an automobile wax) has 62 carbons, $CH_3(CH_2)_{33}CO_2(CH_2)_{26}CH_3$. The biological role of carnauba wax is as a leaf coating involved in the conservation of plant moisture. Animal waxes include cetyl palmitate (spermaceti) found in sperm whales and beeswax (one constituent of which has been identified as $CH_3(CH_2)_{29}CO_2(CH_2)_{29}CH_3$, which is used in the construction of the honeycomb).

Lower molecular weight, naturally occurring esters make major contributions to the pleasant aromas of fruits and flowers. These odors have been shown to be composed generally of complex mixtures of materials that have been separated only since the development of modern chemical instrumentation. Single components, however, may play a dominant role in an individual plant or animal. Propyl acetate (pears), ethyl butyrate (pineapples), and 3-methylbutyl acetate (bananas) are examples of simple esters responsible for a particular plant odor. Odors derived from esters are not limited just to esters of straight-chain carboxylic acids, as is demonstrated by oil of wintergreen, methyl salicylate:

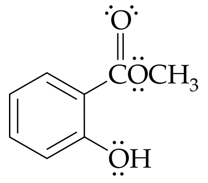

Oil of wintergreen

Phospholipids. *Lipid* is a term applied to those natural substances that are more soluble in nonpolar solvents than in water. In its most general sense, it is a broad definition that includes fats, waxes, hydrocarbons, and so on. In biochemistry, lipids are more narrowly defined as substances that yield fatty acids upon hydrolysis.

Another class of glycerides are those substances in which one of the fatty acid groups has been replaced by a phosphoric acid residue: the phospholipids, or more accurately, the phosphoglycerides. The phosphate group is almost always further esterified, usually with a biological amino alcohol, such as choline (the lecithins) or ethanolamine (the cephalins):

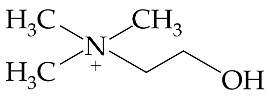

Choline

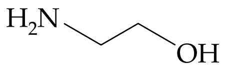

Ethanolamine

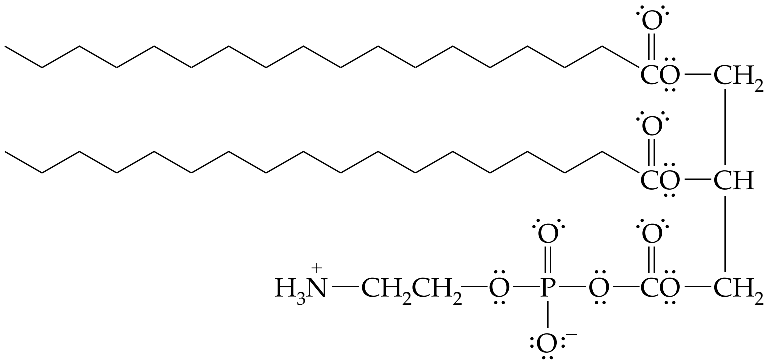

Phosphoglyceride, a cephalin

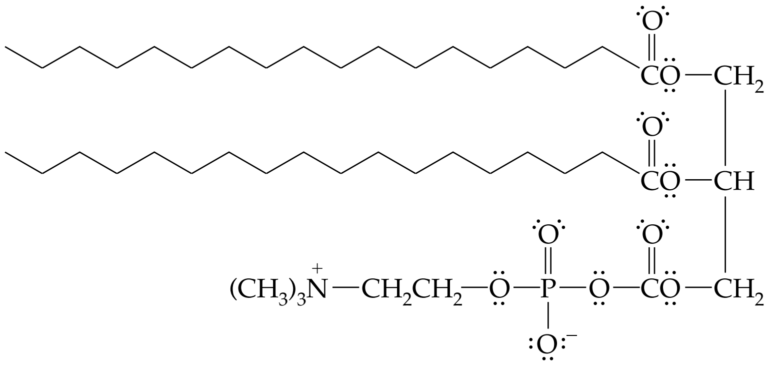

Phosphoglyceride, a lecithin

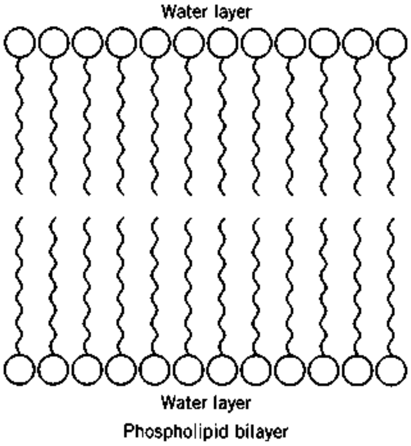

Figure 6.22 Phospholipid bilayer.

These latter groups significantly increase the polarity of the glycerol section of the molecule so that phosphoglycerides undergo strong self-association. In aqueous solutions, this intermolecular attraction can lead to lipid bilayer formation (Fig. 6.22). In a lipid bilayer, the molecules organize themselves to form sheets that contain a double layer of the molecules formed by tail-to-tail association within the interior of the sheet; the outer surface of the lipid bilayer contains the polar heads, which are heavily solvated by water molecules. This association of phosphoglycerides is the key feature in the construction of cell membranes. Thus, esters must have played a vital role at the very earliest stages as cell structures evolved in the development of living systems.

Preparation of Esters. Esters are generally synthesized by one of four fundamental routes:

1. Esterification of a carboxylic acid with an alcohol in the presence of an acid catalyst
2. Alcoholysis of acid chlorides, anhydrides, or nitriles
3. Reaction of a carboxylate salt with an alkyl halide or sulfate
4. Transesterification reactions

The first of these pathways, known as Fischer esterification, is the method used for the preparation of ethyl laurate in Experiment [8A] and of isopentyl acetate in Experiment [8B]. A modern variation of the Fischer esterification is used in Experiment [8C]. The development of this esterification reaction represents just one of a number of major discoveries in organic chemistry by Emil Fischer.

Emil Fischer (1852–1919)[7] In 1874 Fischer obtained his Ph.D. from the University of Strasbourg, studying with Adolf von Baeyer. He later had appointments as Professor of Chemistry at Erlangen, Würzburg, and Berlin universities.

In 1875, at the age of 23, and one year after completing his graduate studies, he synthesized phenyl hydrazine (C_6H_5—$NHNH_2$) for the first time. This highly reactive reagent later played a key role in Fischer's work on elucidating structures of a large majority of the sugars (carbohydrates), an entire class of important and complex organic molecules. Sugars, or carbohydrates, represent the prime pathway for the storage of radiant energy from the sun, through photosynthesis, as chemical energy. In the short period from 1891 to 1894, Fischer established not only the basic structures, but also the configurations of all the known sugars. In addition, he predicted all the theoretically possible isomers and, in the process, developed a method of representing the three-dimensional molecular structures in two-dimensional drawings that became known as Fisher projection formulas. These representations are still in use today, and have been widely applied beyond sugar chemistry. This work by Fischer led directly to proving the existence of the asymmetric carbon atom, a concept proposed by Vant Hoff and LeBel in 1874.

Fischer was also active in the area of protein chemistry. He demonstrated that amino acids are the basic subunits from which proteins are constructed.

---

[7] See *Chem. Ind.* **1919,** *42,* 269; Darmstaedter, L.; Oester, R. E. *J. Chem. Educ.* **1928,** *5,* 37; Ratman, C.V. *ibid.* **1942,** *38,* 93; Kauffman, G. B.; Priebe, P. M. *ibid.* **1990,** *67,* 93; *Chem. Eng. News* **1992,** June p. 25. Recommended reading, "The Emil Fischer–William Ramsay Friendship: The Tragedy of Scientists in War." *J. Chem. Educ.* **1990,** *67,* 451.

Fischer devised methods for the synthesis of many of the known amino acids. Perhaps his most ingenious contribution was the "lock and key" hypothesis of how proteins bind with substrates of complementary shapes. This work ultimately led to our understanding of how enzymes, the catalysts of biochemical reactions, function.

Fischer carried out extensive work on the chemistry of purine and on those compounds containing its nucleus. Purine is one of the two nitrogen base ring systems present in DNA. Fischer synthesized approximately 150 members of this class of heterocyclic compounds (including the first synthesis of the alkaloid caffeine (see Experiment [11B]), uric acid, and the xanthines (also see Experiment [11B]). He developed a general synthesis of another nitrogen heterocycle, indole, which was so effective that it has become one of the classic synthetic methods of organic chemistry and is known today as the "Fischer indole synthesis":

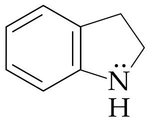

Indole

Fischers work essentially laid the foundation of modern biochemistry. Regarded as the greatest organic chemist of his time, Fischer became the second chemist to receive the Nobel Prize (1902). Depressed by the loss of his young wife at the age of 33, by the loss of two of his three sons (one by suicide, the other in World War I), suffering from the advanced stages of intestinal cancer, and saddened by the socioeconomic conditions of postwar Germany, Fischer committed suicide.

Mechanism of the Fischer Esterification Reaction. The Fischer esterification proceeds by nucleophilic attack of the alcohol on the protonated carbonyl group of the carboxylic acid to form a tetrahedral intermediate. Collapse of the tetrahedral intermediate regenerates the carbonyl group and produces the ester and water. The overall sequence is outlined here:

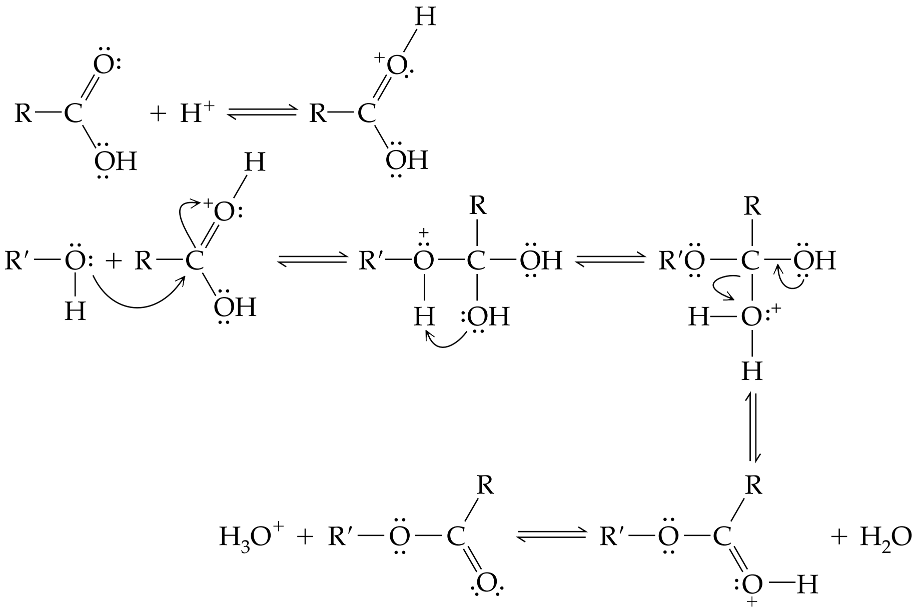

In the Fischer esterification with primary alcohols, the products are only slightly favored by the equilibrium and, therefore, to obtain substantial yields of the ester, the equilibrium must be shifted toward the products. This result can be accomplished in a number of ways. For example, an excess of the starting alcohol can be used to shift the position of equilibrium toward the products. This technique is used in the preparation of ethyl laurate (Experiment [8A]). An analogous alternative is to use an excess of the carboxylic acid. A third option to drive the reaction is the removal of one or both of the products (the ester or water) as they are formed during the reaction. The preparation of isopentyl acetate, synthesized in Experiment [8B], depends on two of these strategies: (1) the reaction is run in an excess of the carboxylic acid (it doubles as the solvent) and (2) the water generated as one of the products is removed by a drying agent. The acid catalyst used in Fischer esterifications is generally dry hydrogen chloride, concentrated sulfuric acid, or a strong organic acid, such as *p*-toluenesulfonic acid.

When the carboxyl group and hydroxyl group are present in the same molecule, an intramolecular esterification may occur and a cyclic ester (called a lactone) may be formed. Lactonization requires an acceptable conformation: the two groups must be close and spatially positioned to react. Ring closure is especially favorable if lactone formation yields five- or six-membered (stable and rapidly formed) ring systems.

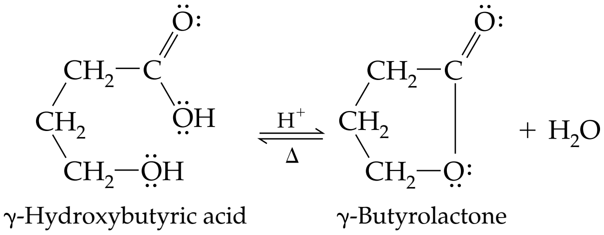

As noted earlier, Fischer esterification is an equilibrium reaction and is thus reversible. Thus, heating an ester in aqueous solution, in the presence of an acid catalyst, regenerates the corresponding carboxylic acid and alcohol. This latter reaction is called *acid hydrolysis* of an ester. The rate-determining step in both the forward esterification reaction and the reverse reaction, acid hydrolysis, is the formation of the tetrahedral intermediate. It is, therefore, evident that the rate of the reaction will be determined by the ease with which the nucleophile (alcohol on esterification and water on hydrolysis) approaches the carbonyl group. Steric and electronic factors have been shown to have large effects on the rate of esterification. An increase in the number of bulky substituents substituted on the α and β positions of the carbonyl-containing compound decreases the rate (steric effects). Electron-withdrawing groups near the carbonyl group, on the other hand, tend to increase the rate by increasing the electrophilicity (partial positive charge) of the carbonyl carbon (electronic effects). Conversely, electron-donating groups act to retard the rate of esterification (electronic effects).

You are to synthesize the ethyl ester of lauric acid in Experiment [8A]. Lauric acid, $CH_3(CH_2)_{10}CO_2H$ (dodecanoic acid), is one of the four most common fatty acids found in naturally occurring triglycerides. It is named for the laurel botanical family from which it was first isolated in 1842. It is the most abundant of the fatty acids isolated from the vegetable oils of palm kernel oil (52%), the seed fat of *Elaeis guineensis;* of coconut oil (48%), *Cocos nucifera;* and of babassu oil (46%), *Attalea funifera*.

In the preparation given for ethyl laurate (Experiment [8A]), acetyl chloride is used to generate the HCl catalyst in situ. Notice that the other product of this step is a molecule of the desired ester:

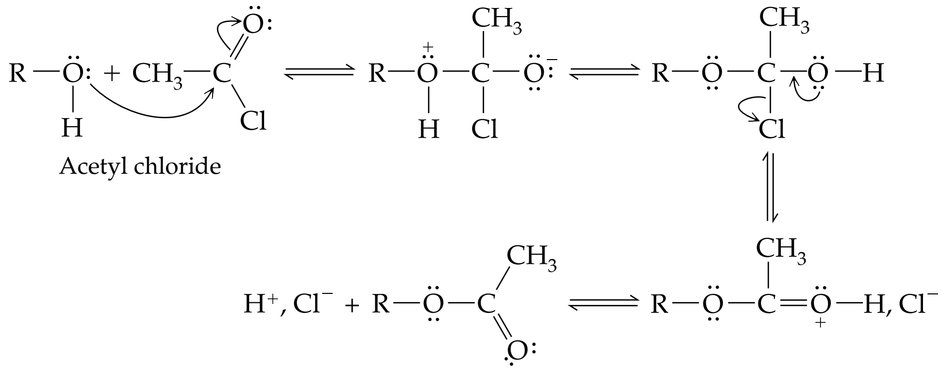

In Experiment [8B] you will be synthesizing the isopentyl alcohol (3-methylbutanol) ester of acetic acid (the basic building block of the fatty acids), isopentyl acetate (isoamyl acetate). This low molecular weight ester has a distinct banana- or pear-like odor, and the liquid product is often referred to as banana oil or pear oil (see above). Isopentyl acetate has a wide variety of uses: as a flavoring agent in mineral waters and syrups; a solvent for oil paints, tannins, nitrocellulose, lacquers, and a number of other commercial products; a perfume ingredient in shoe polish; and in the manufacture of artificial silk, leather, and pearls. You are very likely to find this experiment to be a pleasant olfactory experience!

In Experiment [8C] you will use a polymer-bound acid reagent to catalyze the esterification reaction. Polymer-bound reagents are becoming increasingly useful in organic synthesis; both in the research laboratory and in industrial-scale reactions.

Ethyl Laurate

Experiment 8A

REACTION

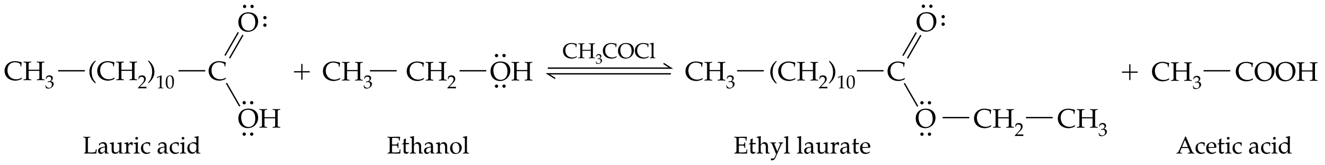

$$CH_3-(CH_2)_{10}-C(=O)-OH + CH_3-CH_2-OH \xrightleftharpoons{CH_3COCl} CH_3-(CH_2)_{10}-C(=O)-O-CH_2-CH_3 + CH_3-COOH$$

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 3.0 h.

Physical Properties of Reactants							
Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Lauric acid	200.3	70 mg	0.35	44			
Ethanol	46.07	1.0 mL	17		78.5	0.71	1.3611
Acetyl chloride	78.50	30 μL	0.42		50.9	1.11	1.3898

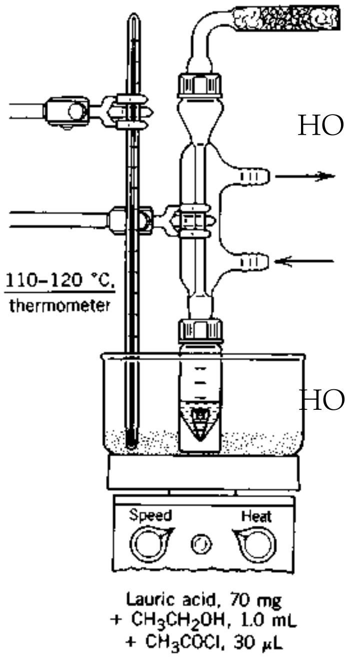

Reagents and Equipment. Weigh and add 70 mg (0.35 mmol) of lauric acid to a 3.0-mL conical vial containing a magnetic spin vane. Then equip the reaction vial with a reflux condenser that is protected by a calcium chloride drying tube (⬅). Use a graduated 1.0-mL pipet to add 1.0 mL of absolute ethanol to the vial. In the **hood,** add 30 μL (0.43 mmol) of acetyl chloride, using an automatic delivery pipet. (Remember not to turn on the cooling water in the condenser until the system is completely assembled. Otherwise, the cold inner surface will condense moisture from the laboratory atmosphere. It takes only a very small amount of water on the condenser walls to completely deactivate your acid chloride catalyst.) HOOD

HOOD

CAUTION: Acetyl chloride is an irritant. Dispense this reagent in the ***hood*** using an automatic delivery pipet. (Be sure to quickly reassemble the reaction vial and condenser following addition of the acid chloride, because this reagent will rapidly react with moist laboratory air and lose its activity.)

Reaction Conditions. Heat the reaction mixture for 1 h at gentle reflux with stirring, using a sand bath temperature of 110–120 °C. Cool the resulting mixture to room temperature.

Isolation of Product. After removing the reflux condenser and drying tube, remove the spin vane with forceps. Add a boiling stone to the cooled vial and then concentrate the reaction solution to a volume of about 0.25 mL by warming in a sand bath in the **hood.** Remove the vial from the sand bath and allow the vial to cool. Use a 1.0-mL graduated pipet to add 0.5 mL of diethyl ether and 0.25 mL of 5% sodium bicarbonate solution to the concentrated product mixture. Cap the conical vial and shake gently (or mix on a Vortex mixer). Loosen the cap carefully to vent the two-phase mixture. Remove the bottom aqueous layer using a Pasteur filter pipet and set it aside in a labeled Erlenmeyer flask. Extract the ether phase with three additional 0.25-mL portions of 5% sodium bicarbonate solution. Save the aqueous wash after each extraction (you may combine them with the initial aqueous basic phase in the Erlenmeyer) and do not discard them until you have successfully purified and characterized the ethyl laurate. HOOD

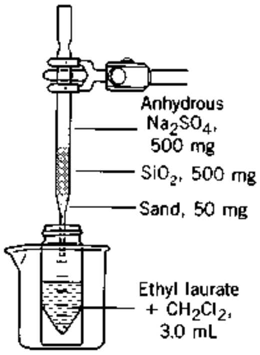

Purification and Characterization. Dry and purify the wet, crude ether solution of ethyl laurate by column chromatography. In a Pasteur filter pipet, place 500 mg of activated silica gel followed by 500 mg of anhydrous sodium sulfate (⬅). Wet the column first with 0.5 mL of methylene chloride and then transfer the crude ether solution of ethyl laurate to the column using a Pasteur filter pipet. Use a **tared** 5-mL conical vial containing a boiling stone as a collection flask for the column eluant. Rinse the reaction vial with methylene chloride (0.5 mL). Transfer the rinse to the column. Repeat both rinse and transfer with a second aliquot of methylene chloride (0.5 mL). Add an additional 1.0 mL of methylene chloride directly to the column to ensure complete elution of the ester.

HOOD Remove the ether-methylene chloride solvent by evaporation in the **hood** by using a stream of nitrogen gas with gentle warming in a sand bath. Make sure the vial remains warm to your fingers during the evaporation process—a warm vial ensures that condensation of moisture on the liquid product is avoided during solvent concentration.

The recovered ethyl laurate is a clear, viscous, pleasant-smelling ester, and is usually fully characterized without further purification. Weigh the product and calculate the percent yield. Determine the refractive index (optional) and boiling point and compare your data to the literature values. Obtain an IR spectrum of the ester using the capillary film technique and compare it to the spectrum of an authentic sample or to one found in the literature (*Aldrich Library of IR spectra* and/or SciFinder Scholar).

Chemical Tests. Does this ester give a positive hydroxamate test (see Chapter 9)? Check the solubility of ethyl laurate in water. Would you have predicted the result? Is the ester soluble in 85% phosphoric acid or in concentrated sulfuric acid?

Isopentyl Acetate: Semimicroscale Preparation

Experiment 8B

Isopentyl acetate is prepared by the following procedure.

REACTION

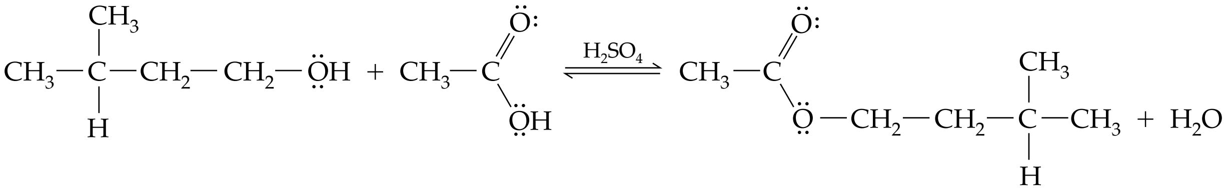

Isopentyl alcohol Acetic acid Isopentyl acetate

EXPERIMENTAL PROCEDURE

Estimated time to complete the experient: 3.0 h.

Physical Properties of Reactants						
Compound	**MW**	**Amount**	**mmol**	**bp (°C)**	***d***	**n_D**
Isopentyl alcohol	88.15	800 μL	7.4	132	0.81	1.4053
Acetic acid	60.1	1.5 mL	26.2	116	1.05	1.3720
Sulfuric acid, concd		4 drops				

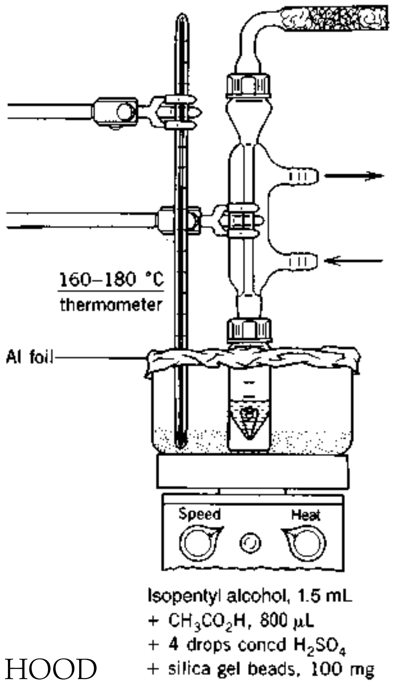
160–180 °C
thermometer
Al foil
Speed
Heat
Isopentyl alcohol, 1.5 mL
+ CH_3CO_2H, 800 μL
+ 4 drops concd H_2SO_4
+ silica gel beads, 100 mg
HOOD

Reagents and Equipment. In a 5.0-mL conical vial containing a magnetic spin vane and equipped with a reflux condenser protected by a calcium chloride drying tube, place 800 μL (7.4 mmol) of isopentyl alcohol, 1.5 mL (26.2 mmol) of glacial acetic acid, 4 drops (Pasteur pipet) of concentrated sulfuric acid, and approximately 100 mg of silica gel beads (➡).

> CAUTION: Cap the vial immediately after addition of each reagent. Dispense the reagents in the ***hood*** using automatic delivery pipets. Concentrated acetic and sulfuric acids are *corrosive*.

NOTE. The silica gel beads (t.h.e. desiccant) are used to absorb the water as it is generated in the reaction.

Reaction Conditions. Heat and stir the reaction mixture using a sand bath temperature of 160–180 °C for 1 h. Cool the resulting mixture to room temperature and remove the spin vane with forceps. Add 0.5 mL of diethyl ether to increase the volume of the organic phase.

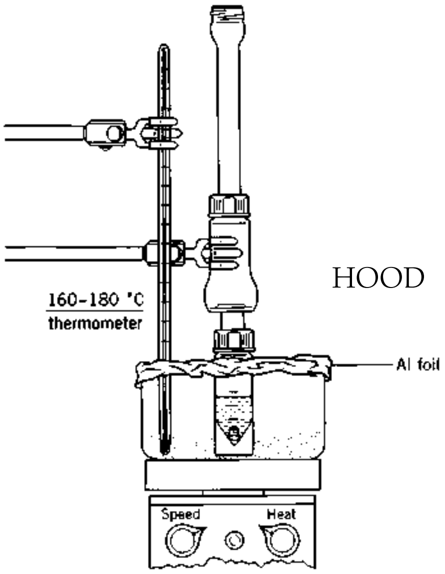

Distillation of reaction products

HOOD

Isolation of Product. Extract the crude organic product with three 2-mL portions of 5% sodium bicarbonate solution, followed by 1 mL of water. During each extraction, cap the vial, shake gently, vent carefully, and then allow it to stand so the layers may separate. A Vortex mixer may be used to good advantage in this sequence. Remove the aqueous layer and be sure not to discard it until the final product is purified and characterized.

Following the aqueous extraction, dry the organic phase by adding anhydrous sodium sulfate to the vial. Transfer the anhydrous solution, using a Pasteur filter pipet, to a clean, dry 3-mL conical vial containing a boiling stone. Remove the added ether by evaporation in the **hood** by using a gentle stream of nitrogen gas with gentle warming in a sand bath. Next, attach the vial containing the crude ester residue to a Hickman still equipped with an air condenser and arranged in a sand bath for distillation (⬅). Cover the sand bath with aluminum foil during the procedure.

Purification and Characterization. Distill the isopentyl acetate product at a sand bath temperature of 160–180 °C. As the material collects in the collar of the still, transfer it by Pasteur pipet (9 in.) to a **tared** $\frac{1}{2}$-*dram* vial.

www

Weigh the clear, viscous, pleasant-smelling isopentyl acetate, and calculate the percent yield. Determine the refractive index (optional) and boiling point, and compare your values with those found in the literature.

Obtain an IR spectrum of the ester using the capillary film technique and compare it with that shown in Figure 6.23.

Infrared Analysis. The conversion of a branched-chain aliphatic primary alcohol to an acetate ester results in significant changes in the infrared spectrum of the molecule. These changes are similar to those observed in straight-chain systems. The two macro group frequency trains in the present example are

1. Isopentyl alcohol (Fig. 6.24): 3350 (broad), 3000–2850, 1460–1300, 1060, and 660 cm^{-1}
2. Isopentyl acetate (Fig. 6.23): 3490 (weak), 3000–2850, 1746 (strong), 1367 (broad), 1250, and 1040 cm^{-1}

If the macro group frequencies are further constrained to include only those systems in which the chain branching involves the presence of an isopropyl group, two additional bands near 1385 and 1365 cm^{-1} are required. These peaks, which are present in both the alcohol and the ester, arise from the spatially coupled and split symmetric methyl bending vibrations of the aliphatic backbone. In both isopentyl alcohol and isopentyl acetate, this pair of bands are found at identical locations, 1386 and 1367 cm^{-1}.

Chemical Tests. Does the product give a positive hydroxamate test for an ester (Chapter 9)? Check the solubility of this ester in water, ether, and 85% H_3PO_4. In which solubility group (see Chapter 9, pp. 638–639) does isopentyl acetate fall?

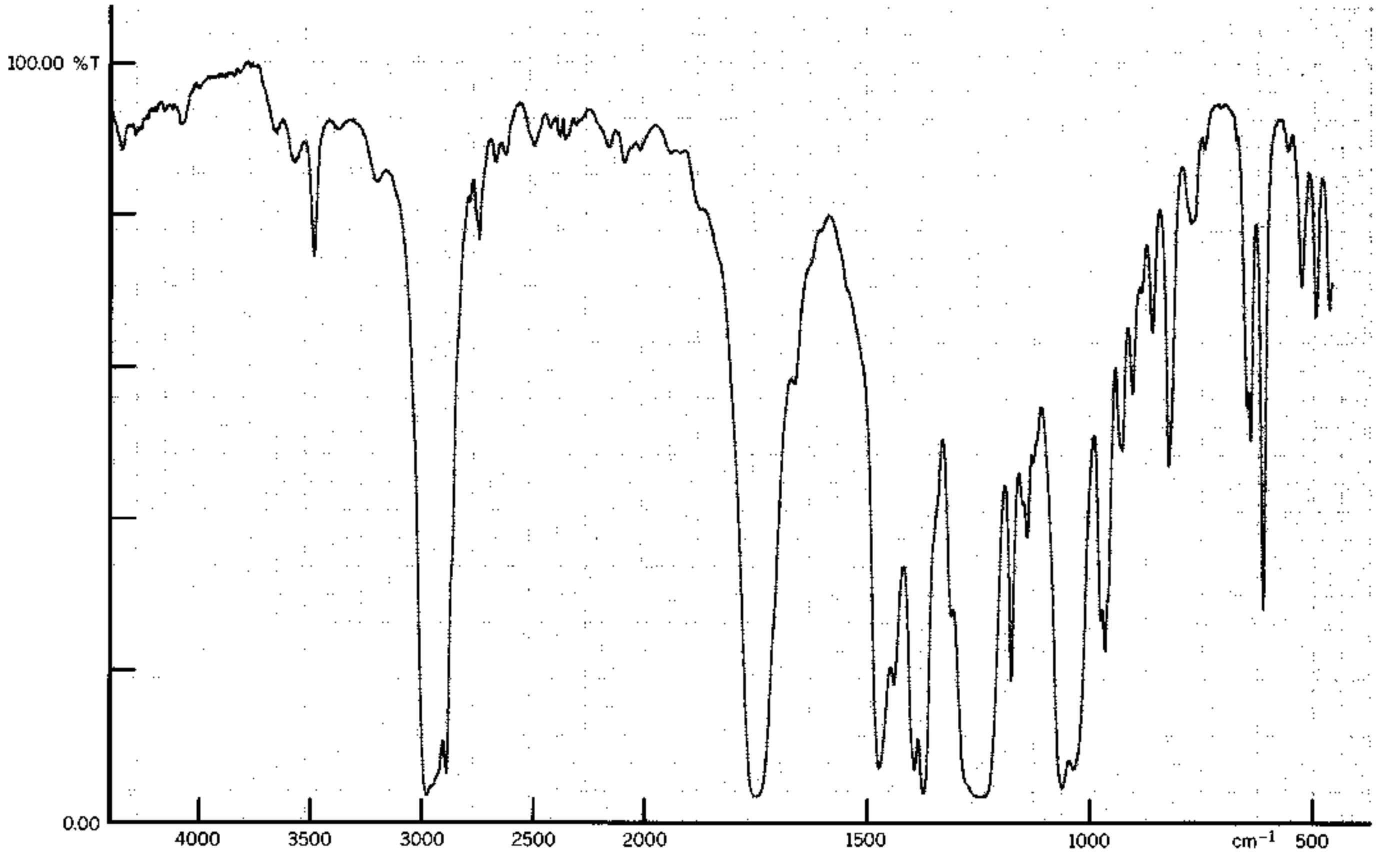

Sample Isopentyl acetate
%T x ABS — Background Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm-1
Sample Condition liquid, neat
Cell Window KBr
Cell Path Length capillary film
Matrix Material

Figure 6.23 **IR spectrum: isopentyl acetate.**

Isopentyl Acetate: Preparation Using a Monomode Microwave Apparatus

Experiment 8B-1*

M

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.25 h.

Physical Properties of Reactants						
Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Isopentyl alcohol	88.15	0.8 mL	7.4	132	0.81	1.4053
Acetic acid	60.1	1.5 mL	26.2	116	1.05	1.3720
Sulfuric acid, conc		4 drops				

*This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

JWCL196_c06_188-254.qxd 11/17/09 1:32 PM Page 204

EQA

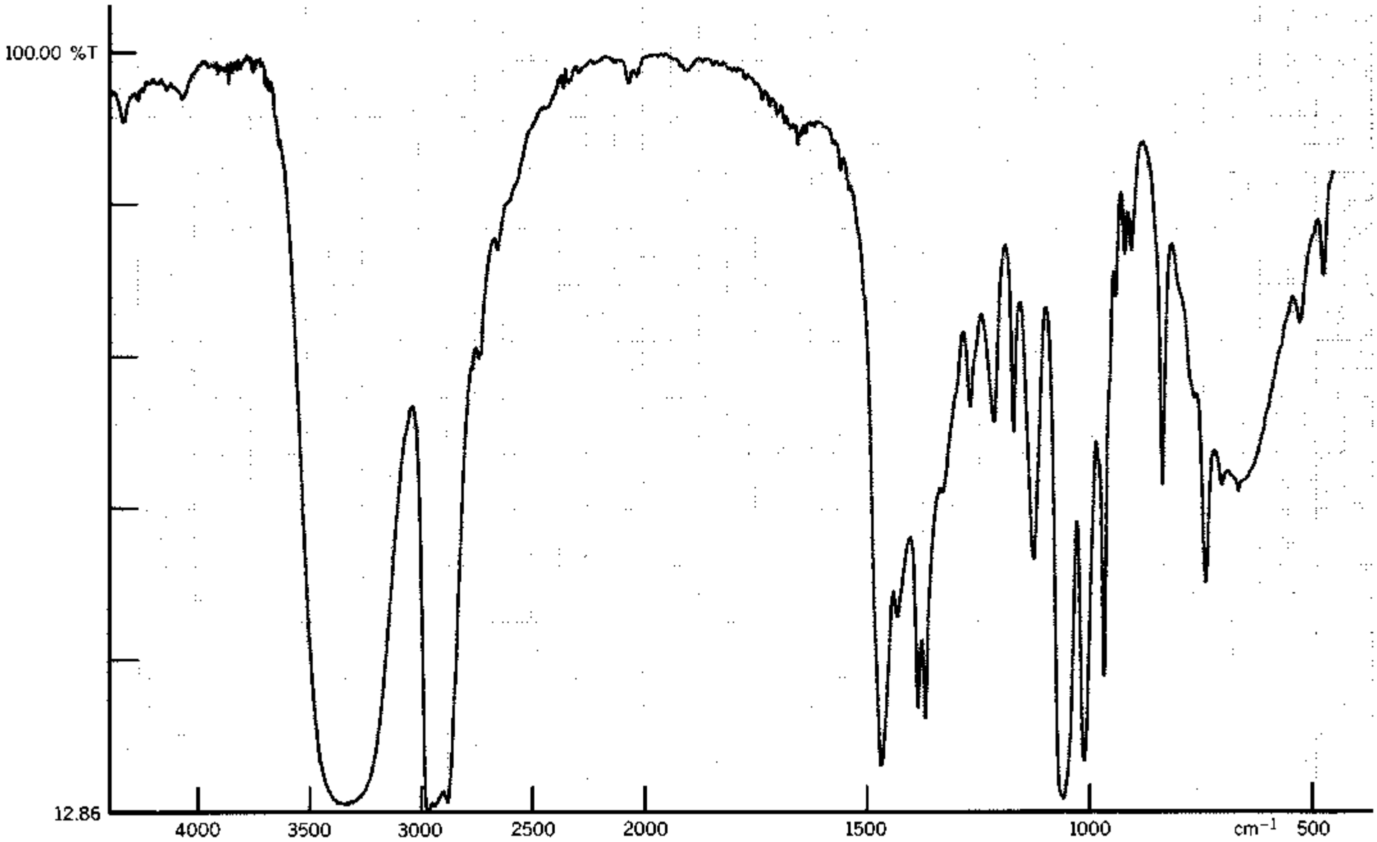

Sample Isopentyl alcohol

%T x ABS — Background Scans 4	Scans 16
Acquisition & Calculation Time 42 sec	Resolution 4.0 cm⁻¹
Sample Condition liquid, neat	Cell Window KBr
Cell Path Length capillary film	Matrix Material

Figure 6.24 IR spectrum: isopentyl alcohol.

Reagents and Equipment. This experiment is designed for use in the CEM Discover and Biotage Initator microwave units.

In a 10.0-mL glass microwave reaction vessel containing a magnetic stir bar, place 0.8 mL (7.4 mmol) of isopentyl alcohol, 1.5 mL (26.2 mmol) of glacial acetic acid, 4 drops (Pasteur pipet) of concentrated sulfuric acid, and approximately 100 mg of silica beads. Immediately cap the vessel with the microwave pressure cap.

HOOD

CAUTION: Dispense reagents in the **hood** using automatic delivery pipets. Concentrated acetic and sulfuric acids are *corrosive*. Since the reaction requires heating reagents to above their boiling point in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential.*

Reaction Conditions. Place the reaction vessel in the microwave cavity and, depending on the equipment used, position the pressure device on top. Program the microwave unit to heat the reaction mixture to 130 °C and hold at this temperature for 15 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the tube from the microwave unit. Transfer the reaction mixture with a Pasteur pipet into a 5.0-mL conical vial. Rinse

CONFIRMING PAGES aptara

the microwave reaction vessel with 0.5 mL of diethyl ether and add the washings to the conical vial.

Isolation of Product. The remainder of the procedure is identical to experiment 8B (p. 202).

Isopentyl Acetate: Preparation Using a Multimode Microwave Apparatus

Experiment 8B-2*

M

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.25 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Isopentyl alcohol	88.15	1.6 mL	14.7	132	0.81	1.4053
Acetic acid	60.1	3 mL	52.4	116	1.05	1.3720
Sulfuric acid, conc		8 drops				

Reagents and Equipment. This experiment is designed for use in the CEM MARS, Milestone START, and Anton Paar Synthos 3000 microwave units. When using the Anton Paar Synthos 3000 unit with the 24-position silicon carbide plate rotor containing glass vials, the reagent and solvent quantities cited in the monomode procedure should be used in conjunction with the reaction conditions here in the multimode procedure.

In a microwave reaction vessel containing a magnetic stir bar, place 1.6 mL (14.7 mmol) of isopentyl alcohol, 3.0 mL (52.4 mmol) of glacial acetic acid, 8 drops (Pasteur pipet) of concentrated sulfuric acid, and approximately 200 mg of silica beads. Immediately cap the vessel with the microwave pressure cap and adjust the tightness to the manufacturer-specified level. Place the sealed vessel into its outer protective jacket.

HOOD

CAUTION: Dispense reagents in the **hood** using automatic delivery pipets. Concentrated acetic and sulfuric acids are *corrosive*. Since the reaction requires heating reagents to above their boiling point in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential.*

Reaction Conditions. Insert the loaded vessels into the reaction carousel ensuring they are evenly spaced and then place the carousel into the microwave cavity. If provided by the manufacturer, connect a temperature probe to the control vessel. Program the microwave unit to heat the reaction vessels to 120 °C and hold at this temperature for 5 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the carousel from the microwave unit.

Isolation of Product. Place 10 mL of 10% sodium bicarbonate solution in a 30-mL separatory funnel. Clamp the funnel to a ring stand. Pipette the reaction mixture from the microwave reaction vessel to the separatory funnel. An effervescent reaction will occur as the excess acetic acid is neutralized. After the effervescence has subsided, add water to dissolve any solids. Add 5 mL of diethyl ether and carefully cap and invert the funnel. Immediately vent the funnel by opening the stopcock. Close the stopcock, place the funnel back on the ring stand, and remove the stopper. Drain the lower (aqueous) layer into a 50-mL Erlenmeyer flask. Extract the crude organic layer with a further two 5-mL portions of 10% sodium bicarbonate solution, followed by 5 mL of water. During each extraction, cap and invert the funnel several times releasing the pressure by opening the stopcock and then allow the funnel to stand so the layers will separate. Remove the lower (aqueous) layer after each extraction into the 50-mL Erlenmeyer flask. Save the aqueous waste until the experiment is complete and then discard as directed. Dry the organic layer by pipetting it into a clean 25-mL Erlenmeyer flask containing 200 mg of anhydrous sodium sulfate. Transfer the anhydrous solution, using a Pasteur filter pipet, to a clean tared 10-mL pear-shaped flask containing a boiling stone. Remove the ether on a rotary evaporator or by evaporation in the **hood** using a gentle stream of nitrogen gas with warming in a sand bath. Next, attach the flask containing the crude ester residue to a Hickman still equipped with an air condenser and arrange in a sand bath for distillation. Cover the sand bath with aluminum foil during the procedure.

---

*This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

HOOD

Purification and Characterization. Distill the isopentyl acetate product at a sand bath temperature of 160–180 °C. As the material collects in the collar of the still, transfer it by Pasteur pipet (9 in.) to a **tared** $\frac{1}{2}$-*dram* vial.

Weigh the clear, viscous, pleasant-smelling isopentyl acetate, and calculate the percent yield. Determine the refractive index (optional) and boiling point, and compare your values with those found in the literature.

www

Obtain an IR spectrum of the ester using the capillary film technique and compare it with that shown in Figure 6.23.

Experiment 8C

Esterification by Use of Acidic Resins: Semimicroscale Preparations

Modern synthetic reactions are making increased use of reagents that are heterogeneous in character. The use of resins as the support material for reactive compounds has become very popular because of their ease of removal and reliability. In this experiment you will utilize one of the recent additions to the arsenal of resin catalysts, Nafion 417. This catalyst is a powerfully acidic resin. It can approach the acidities of 100% sulfuric acid and of trifluoromethanesulfonic acid in trifluoroacetic anhydride solution. The "superacidity" of the sulfonic acid group in Nafion 417 is attributable to the electron-withdrawing ability of the perfluorocarbon backbone of the resin to which it is attached.

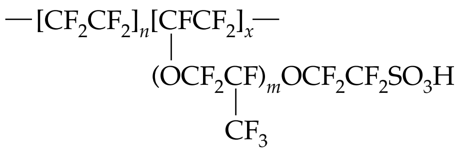

$$-[CF_2CF_2]_n[\underset{\underset{(OCF_2\underset{\underset{CF_3}{|}}{C}F)_mOCF_2CF_2SO_3H}{|}}{C}FCF_2]_x-$$

Nafion 417 resin with an equivalent weight of approximately 1200 contains tetrafluoroethylene (the monomer used to make Teflon) and perfluorovinyl ether units in a ratio of 7:1. These resins have been used successfully by Olah, as in this experiment, for the esterification of carboxylic acids.[8]

---

[8] Olah, G. A.; Keumi, T.; Meidar, D. *Synthesis* **1978,** 929.

REACTION

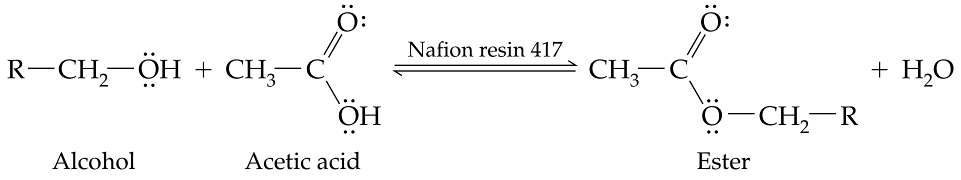

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
1-Butanol	74	365 μL	4.0	118	0.810	1.3990
Isopentyl alcohol	88	500 μL	4.6	132	0.81	1.4053
1-Pentanol	88	500 μL	4.6	138	0.811	1.4090
1-Hexanol	102	570 μL	4.55	156	0.814	1.4180
1-Octanol	130	700 μL	4.45	196	0.827	1.4290
Acetic acid	60	275 μL	4.8	116	1.049	1.3720
Nafion 417 resin	Equivalent weight 1100			1.5 × 0.5-cm strip		

Reagents and Equipment. In a 5.0-mL conical vial containing a magnetic spin vane and equipped with a reflux condenser protected by a calcium chloride drying tube, place the alcohol (*see table for correct molar quantities for your particular alcohol*), 275 μL (4.8 mmol) of glacial acetic acid, and a strip of Nafion 417 resin (1.5 × 0.5 cm). (If the strip of resin is placed so as to encircle the spin vane, mixing of the reaction solution is particularly effective.)

NOTE. The Nafion 417 resin has a significant number of hydrophilic centers that function in exactly the same way as the silica gel does in Experiment [8B]. Thus, Nafion 417, in addition to functioning as an acid catalyst, also acts as a desiccant, which helps to drive the esterification to completion by absorbing water as it is generated in the reaction.

Reaction Conditions. The reaction mixture is heated with stirring at a sand bath temperature of 160–180 °C for 30 min. The mixture is then cooled to room temperature.

Isolation of Product. Remove the Nafion strip and the spin vane (with forceps) from the cool reaction mixture and carefully (to avoid foaming) rinse them with 1 mL of 5% $NaHCO_3$. The rinse is added to the vial (to form a two-phase system) and swirled with the reaction mixture. The bicarbonate wash is then separated and the aqueous phase is transferred to a 10-mL Erlenmeyer flask for temporary storage. The reaction mixture is washed with a further 1-mL portion of 5% $NaHCO_3$, and then by two 1-mL distilled water washes (2 × 1 mL). All the aqueous washes are transferred to the storage Erlenmeyer. Dry the crude and wet reaction product by passing it down a Pasteur pipet containing a cotton plug and 1 g of anhydrous granular sodium sulfate (Na_2SO_4). Collect the product residue from the column in a tared 3-mL conical vial and weigh it to determine the crude yield. Obtain an IR spectrum and measure the refractive index (optional) of the crude material. Does the IR

www

spectrum indicate the presence of any unreacted starting material in the crude product? What would be the most likely contaminant?

Purification and Characterization. A 25-μL sample of the crude ester is purified by prep-GC (see GC conditions for each product mixture and Experiment [2] for details on the collection procedure). Following collection of the purified ester, obtain an IR spectrum and boiling point, and compare these with the data obtained on the crude residue. The ester is the first compound to elute from the column. It may be followed by a small amount of a second component. You can calculate the purity of the crude product by measuring the area under the two elution bands. (Only in the case of the butyl acetate synthesis does the workup procedure distort the apparent crude yields: Can you explain why this occurs?) If you want to identify the contaminant (second band) directly, it may be possible to also collect a second fraction from the GC if the contaminant comprises a reasonable fraction of the product mixture (5–10%). It may, however, require a second injection of 30–40 μL.

Gas Chromatographic Conditions	
General Stainless steel columns in 8ft × $\frac{1}{4}$ in. packed with 20% Carbowax Flow rates: 50 mL/min (He); sample size: 25 μL	
Specific	
Ester 1	**Ester 2**
Butyl acetate Column temperature: 120 °C Retention time Ester: 8.5 min Impurity: 11 min	*Pentyl acetate* Column temperature: 130 °C Retention time Ester: 8.5 min Impurity: 11 min
Ester 3	**Ester 4**
Isopentyl acetate Column temperature: 140 °C Retention time Ester: 6 min Impurity: 8 min	*Hexyl acetate* Column temperature: 160 °C Retention time Ester: 10 min Impurity: 12 min
Ester 5	
Octyl acetate Column temperature: 185 °C Retention time Ester: 10 min Impurity: 13 min	

QUESTIONS

6-49. In the preparation of the esters given in this experiment, the reaction product was extracted with 5% sodium bicarbonate solution ($NaHCO_3$) in the isolation step. Why? What gas was evolved during this washing step? Write a balanced equation for the reaction that produced it.

6-50. **(a)** Why is a large excess of acetic acid used in the preparation of isopentyl acetate?
(b) Write a mechanism for the preparation of isopentyl acetate using isopentyl alcohol and acetic acid.

6-51. Concentrated sulfuric acid is used as a catalyst for the esterification of acetic acid in the preparation of isopentyl acetate. Why is the sulfuric acid needed if another acid, acetic acid, is already present?

6-52. Fatty acids are long-chain carboxylic acids, usually of 12 or more carbon atoms, isolated from saponification of fats and oils (esters of glycerol). Draw the structure of each of the fatty acids named below and also determine its common name:
(a) Hexadecanoic acid *(Z)*-9-octadecenoic acid
(b) Octadecanoic acid *(Z,Z)*-9,12-octadecadienoic acid

6-53. Write a mechanism for the acid-catalyzed transesterification reaction of ethyl acetate with 1-butanol, which gives butyl acetate.

6-54. In the infrared spectra of acetates, two intense bands are usually observed in the 1270- to 1000-wavenumber region. These peaks are related to the stretching vibrations of the two C—O bonds of the ester group. Should we expect to be able to assign the C—O bonds to individual peaks, and if so, which mode is associated with which band? Why? (*Hint:* The peak that occurs at higher wavenumbers is consistently close to 1250 cm^{-1}.)

BIBLIOGRAPHY

These references are selected from the large number of examples of esterification given in *Organic Syntheses:*

Bailey, D. M.; Johnson, R. E.; Albertson, N. F. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 618.
Bowden, E. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 414.
Eliel, E. L.; Fisk, M. T.; Posser, T. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 169.
Emerson, W. S.; Longley, R. I., Jr. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 302.
Fuson, R. C.; Wojcik, B. H. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 260.
McCutcheon, J. W. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 526.
Mic'ovic', V. M. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 264.
Peterson, P. E.; Dunham, M. *Organic Syntheses* Wiley: New York, 1988; Collect. Vol. VI, p. 273.
Stevenson, H. B.; Cripps, H. N.; Williams, J. K. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 459.

For an overall review of esterification see

Euranto, E. K. in *The Chemistry of Carboxylic Acids and Esters;* Patai, S. Ed.; Interscience: New York, 1969, p. 505.

EXPERIMENT 9

The E1 Elimination Reaction: Dehydration of 2-Butanol to Yield 1-Butene, trans*-2-Butene, and* cis*-2-Butene*

Common name: 1-butene
CA number: [06-98-9]
CA name as indexed: 1-butene

Common name: *trans*-2-butene
CA number: [624-64-6]
CA name as indexed: 2-butene, (*E*)-

Common name: *cis*-2-butene
CA number: [590-18-1]
CA name as indexed: 2-butene, (*Z*)-

Purpose. This experiment illustrates the variety of pathways that are available to acid-catalyzed elimination reactions of secondary (2°) alcohols via carbocation intermediates. The dehydration of 2-butanol forms a mixture of gaseous alkene products. The alkenes formed in this reaction are separated and identified by using one of the most powerful instrumental techniques available to the modern research chemist for the separation of complex mixtures: gas chromatography (GC).

THE DEVELOPMENT OF CARBOCATION THEORY

The dehydration reaction that you are about to study is representative of the large collection of reactions that are classified as E1 elimination reactions. These reactions all form an intermediate in which one of the carbon atoms bears, if not a *full* positive charge, at least a significant fractional positive charge. It is this fleeting, high-energy intermediate, an aliphatic carbocation, that makes these reactions so interesting.

The development of bonding theory in organic chemistry during the late nineteenth and early twentieth centuries did not accept the existence of carbocations, except in a few esoteric instances. This position was reasonable because aliphatic compounds show little ionic character. The first proposal that these nonpolar substances might actually form cations came from Julius Stieglitz (1867–1937), of the University of Chicago, in a paper published in 1899. Eight years later, James F. Norris (1871–1940) at Massachusetts Institute of Technology produced compelling evidence that these substances might be intermediates in reactions of certain *tert*-butyl halides.

These two papers were the origin of the concept that organic carbon cations, which in those days became known as *carbonium ions*, were far more widespread in organic reactions than previously anticipated. Indeed, these early investigations caused more controversy and more experimental work to unambiguously prove the existence of what are now called *carbocations* than any other single problem in American chemistry.

It was, however, the English chemists Arthur Lapworth, Sir Robert Robinson, C. K. Ingold (University of London), and E. D. Hughes who, from 1920 to 1940, undertook a massive effort to develop the experimental and theoretical data to place these early postulates on a solid scientific foundation. Between 1920 and 1922, Hans Meerwein in Bonn, Germany, demonstrated that carbon rearrangements in the camphene series could be best explained by postulating the presence of carbocation intermediates. Perhaps the most important contribution to the entire subject was published in 1932 by Frank C. Whitmore of Pennsylvania State College, where he was Dean of the School of Chemistry and Physics. His paper, "The Common Basis of Intramolecular Rearrangements," brought together a vast array of data in a beautifully consistent interpretation that essentially cemented the carbocation into contemporary organic chemical theory.

A reaction not too distant (in fact, rather close if you overheat your own!) from the one that is to be carried out below, but one that also included a rearrangement along with the dehydration, was studied by Dorothy Bateman and C. S. "Speed" Marvel at the University of Illinois as early as 1927.

Within 2 years of the publication of the Whitmore paper, Robinson proposed the formation of the steroids (including cholesterol) from squalene (a C_{30} polyunsaturated polyisoprene molecule) via an incredible series of intermediates and rearrangements. Later, following the elucidation of the structure of lanosterol, R. B. Woodward and K. Bloch made a brilliant proposal that at once rationalized the biosynthetic origin of both lanosterol and cholesterol and implicated lanosterol as an intermediate in cholesterol biosynthesis. Their mechanism involved the concerted (bonds made and broken simultaneously) cyclization of four rings, as well as four rearrangements following the generation of the initial carbocation intermediate, to ultimately yield lanosterol.[9]

The elucidation of these biochemical pathways is further evidence of the major impact that our understanding of carbocation chemistry has had on related fields, such as biochemistry.

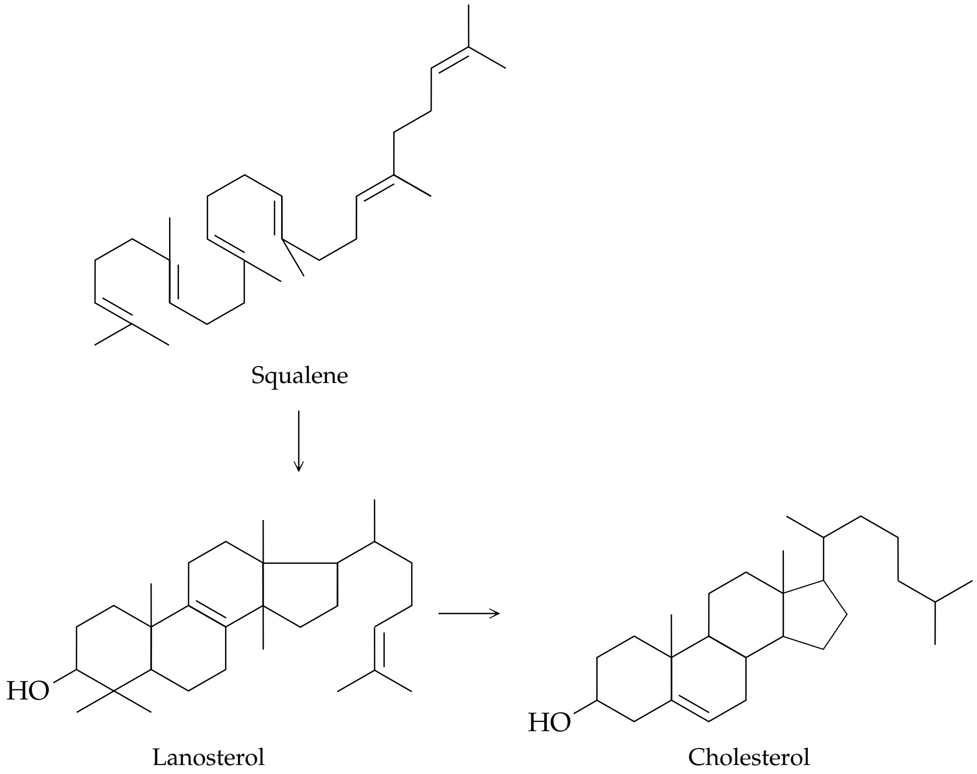

Prior Reading

Technique 1: Gas Chromatography (pp. 55–61)
Technique 7: Collection or Control of Gaseous Products (pp. 105–107)

REACTION

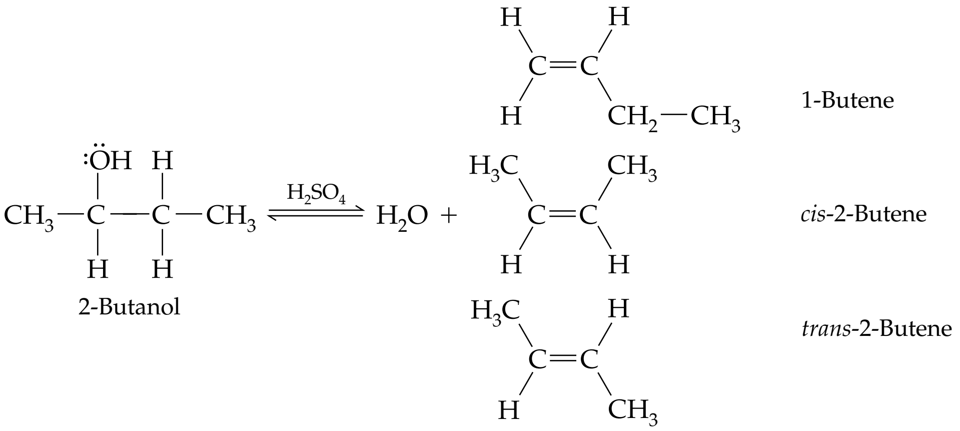

---

[9] Stieglitz, J. *Am. Chem. J.* **1899,** *21,* 101. Norris, J. F. *Am. Chem. J.* **1907,** *38,* 627. Meerwein, H.; van Emster, K. *Berichte* **1922,** *55,* 2500. Whitmore, F. C. *J. Am. Chem. Soc.* **1932,** *54,* 3274. Bateman, D. E.; Marvel, C. S. *J. Am. Chem. Soc.* **1927,** *49,* 2914. Robinson, R. *Chem. Ind.* **1934,** *53,* 1062. Woodward, R. B.; Bloch, K. *J. Am. Chem. Soc.* **1953,** *75,* 2023. See also: Tarbell, D. S.; Tarbell, T. *The History of Organic Chemistry in the United States, 1875–1955;* Folio: Nashville, TN, 1986.

DISCUSSION

The formation of an alkene (or alkyne) frequently involves loss of a proton and a leaving group from adjacent carbon atoms. The generalized reaction scheme is shown below.

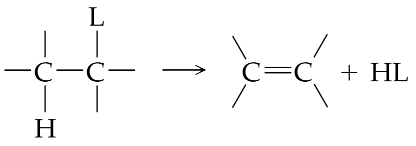

Such a reaction is called an *elimination reaction*, because a small molecule (HL) is *eliminated* from the organic molecule. If the molecule eliminated is water, the reaction may also be referred to as a *dehydration*. One of the common synthetic routes to alkenes is the dehydration of an alcohol.

Dehydration of alcohols is an acid-catalyzed elimination reaction. Experimental evidence shows that alcohols react in the order tertiary (3°) > secondary (2°) > primary (1°); this reactivity relates directly to the stability of the carbocation intermediate formed in the reaction. Generally, sulfuric or phosphoric acid is used as the catalyst in the research laboratory. A Lewis acid, such as aluminum oxide or silica gel, is usually the catalyst of choice at the fairly high temperatures used in industrial scale reactions.

The mechanism for this reaction is classified as E1 (elimination, unimolecular). The elimination, or dehydration reaction, proceeds in several steps:

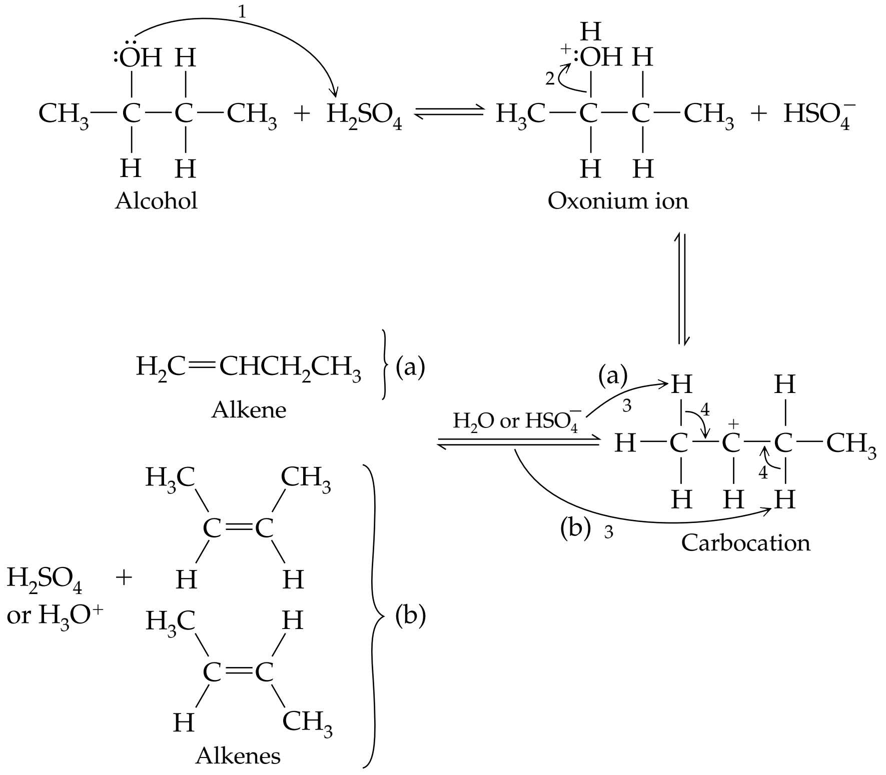

The first step (1) involves the very rapid, though reversible, protonation of the oxygen atom of the alcohol to form an *oxonium ion* (an oxygen cation with a full octet of electrons). This protonation step is important because it produces a good leaving group, water. Without acid, the only available leaving group is hydroxide ion (HO^-) which, as a strong base, is a poor leaving group.

This finding is the reason why acid plays such an important role in the mechanism of this dehydration. The second step (2) of the reaction is the dissociation of the oxonium ion to form an intermediate *carbocation* and water. This step is the rate-determining (and therefore the slowest) step of the reaction. In the third step (3), the carbocation is deprotonated by a ubiquitous water molecule (or other base, such as bisulfate ion, present in the system) in another rapid equilibration. The carbocation gains stability (lower energy) by releasing a proton (H^+) from a carbon atom adjacent (α) to the carbocation (route (a) or (b) shown) to the attacking base. Thus, in step 4 the catalyst is regenerated as a protonated molecule of water (H_3O^+), and the electron pair, previously comprising the C—H bond adjacent to the carbocation, flows toward the positive charge, generating a stable and neutral alkene. A variety of isomeric products may be formed, since different protons adjacent to the carbocation may be removed, and different conformations of the intermediate carbocation are also possible (see routes (a) and (b) above and further discussion below).

E1 elimination reactions, as we have just seen, involve equilibrium conditions and, thus, to maximize the yield (drive the reaction to completion), the alkene is usually removed from the reaction while it is in progress. A convenient technique for accomplishing this task is distillation, which is often used because alkenes *always* have a lower boiling point than the corresponding alcohol. In the present reaction, the alkenes are gases and, therefore, are easily removed and collected as described in the experimental section.

Many 1° (primary) alcohols also undergo dehydration, but usually by a different route, the E2 (elimination, bimolecular) mechanism. This step is governed mainly by the fact that the 1° carbocation that would be required in an E1 process is a relatively unstable (very high energy) intermediate. In this case, attack by the base occurs directly on the oxonium ion:

$$CH_3CH_2CH_2\underset{\cdot\cdot}{\ddot{O}}H \xrightleftharpoons{H_2SO_4} CH_3CH_2CH_2\underset{\cdot\cdot}{\overset{+}{O}}H_2 + HSO_4^-$$

Oxonium ion

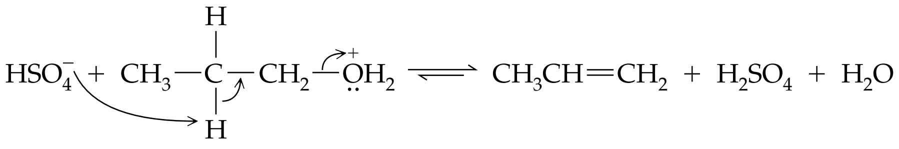

$$HSO_4^- + CH_3-\underset{H}{\overset{H}{C}}-CH_2-\underset{\cdot\cdot}{\overset{+}{O}}H_2 \rightleftharpoons CH_3CH{=}CH_2 + H_2SO_4 + H_2O$$

Oxonium ion

E1 elimination is usually accompanied by a competing S_N1 substitution reaction that involves the *same* carbocation intermediate. Since both reaction mechanisms are reversible in this case, and since the alkene product gases are easily removed from the reaction, the competing substitution reaction (which predominantly regenerates the starting alcohol by attack of water, as a nucleophile, at the carbocation) is not troublesome and the equilibrium eventually leads to gas evolution.

As discussed earlier, many alcohols can dehydrate to yield more than one isomeric alkene. In the present reaction involving the dehydration of 2-butanol, at least three alkenes are usually formed—1-butene, *trans*-2-butene, and *cis*-2-butene:

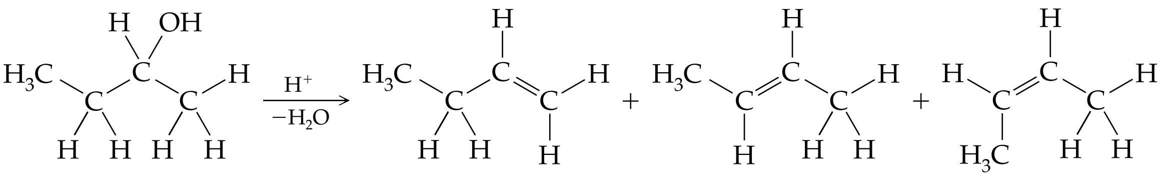

1-Butene *trans*-2-Butene *cis*-2-Butene

We would expect the alkene generated in the largest amount to be the one possessing the highest degree of substitution, since it is the most stable product (lowest energy). This is exactly what is observed: more than 90% of the products are isomers of 2-butene. Because trans alkenes are thermodynamically more stable than their cis counterparts, and since the reaction is reversible, one might expect the dominant isomer in the 2-butene mixture to be trans. Indeed, nearly twice as much trans as cis isomer is formed. An empirical rule, originally formulated by the Russian chemist Alexander Zaitsev (or Saytzeff), for base-catalyzed E2 eliminations states that the alkene with the largest number of alkyl substituents on the double bond will be the major product. This rule can also correctly predict the relative ratio of substituted alkenes to be expected from a given E1 elimination reaction, and it obviously applies in the case of 2-butanol.

Rearrangements of alkyl groups (such as methide, $^{-}{:}CH_3$ and ethide, $^{-}{:}CH_2CH_3$) and hydrogen (as hydride, $:H^{-}$) are often observed during the dehydration of alcohols, especially in the presence of very strong acid where carbocations can exist for longer periods of time. For example, when 3,3-dimethyl-2-butanol is treated with sulfuric acid, the elimination reaction yields the mixture of alkenes shown below. Can you predict which alkene is the principal product?

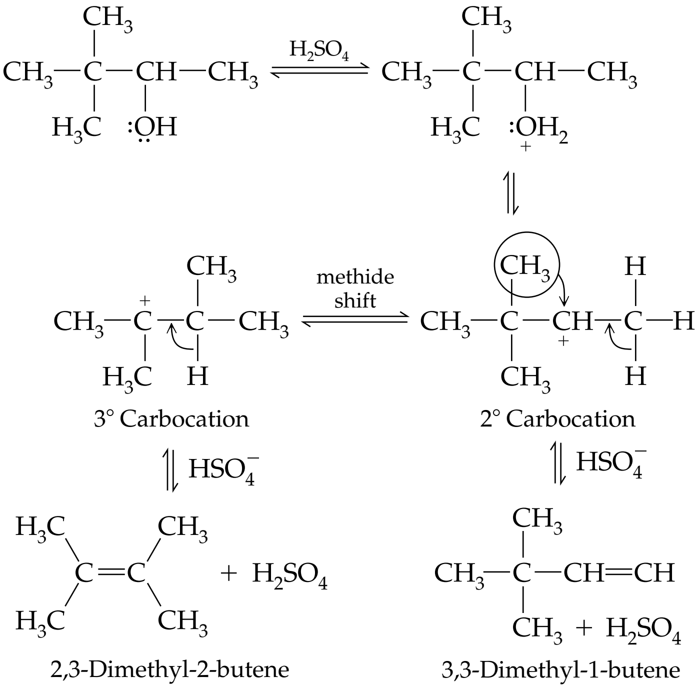

Reactions carried out under these conditions are very susceptible to alkide or hydride shifts if a more stable carbocation intermediate can be formed. In the example above, a 2° carbocation rearranges into the more stable 3° carbocation. This intramolecular rearrangement involves the transfer of an entire alkyl group (in this case a methyl substituent), together with its bonding pair of electrons, to an adjacent carbon atom. This migration, or shift, of the methyl group to an adjacent carbocation is called a *1,2-methide shift*. It commonly occurs in aliphatic systems involving carbocation intermediates that have alkyl substituents adjacent to the cation. Hydride shifts are also frequently observed and actually appear in a wider range of molecules.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2 h.

NOTE. Staggering the starting times of the reaction will allow more students easier access to the GC instrument when the product gases are analyzed.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
2-Butanol	74.12	100 μL	1.1	99.5	0.81	1.3978
Concd sulfuric acid	98.08	50 μL				

Reagents and Equipment. Assemble the gas collection apparatus shown in the figure *before* the reactants are mixed (➨).

The capacity of the gas collection reservoir is determined by the following procedure. Seal the collection tube with a septum cap and invert it. Then add 3.0 mL of water and mark the 3.0-mL level. Finally, add an additional 1.0 mL, and also mark the 4.0-mL level.

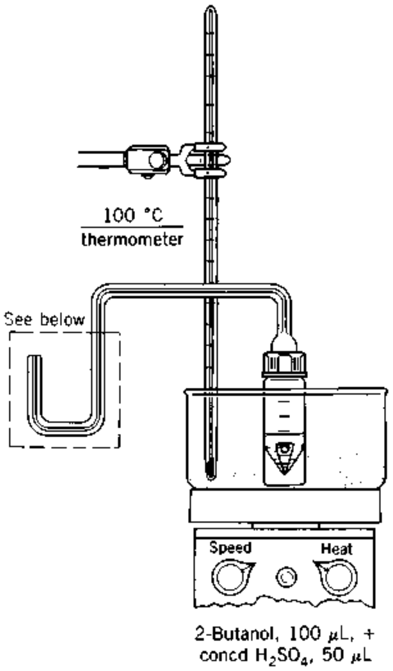

NOTE. Use of a Teflon-lined septum cap on the gas collection reservoir is necessary to prevent loss of the collected butene gases by permeation.

To position the gas collection reservoir, carry out the following steps: (1) fill the reservoir with water; (2) place your finger (index finger is usually used) over the open end of the reservoir; (3) invert it; (4) place it, with the open end down, into a beaker (250 mL) filled with water. When your finger is removed, the column of water should remain in the reservoir.

Place 100 μL (81 mg, 1.1 mmol) of 2-butanol and 50 μL of concentrated sulfuric acid in a **clean,** dry, 1.0-mL conical vial containing a magnetic spin vane.

CAUTION: Sulfuric acid is a strong, corrosive material. Contact with the skin or eyes can cause severe burns. It is best to dispense the reactants using automatic delivery pipets.

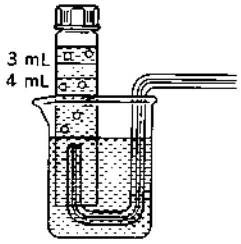

Cap-seal the 1-mL vial to the gas delivery tube and position the tube under water into the open end of the gas collection reservoir as shown in the figure. Clamp the reservoir in place.

Reaction Conditions. Heat the reaction mixture, with gentle stirring, using a sand bath until the evolution of gas takes place (sand bath temperature ~110–120 °C). The mixture should be warmed slowly through the upper temperature range to prevent foaming.

Isolation of Product. Collect about 3–4 mL of gas in the collection reservoir. Remove the gas delivery tube from the gas collection reservoir, and then from the water bath, before removing the reaction vial from the heat. Use this sequence of steps in shutting down the reaction to prevent water from being sucked back into the hot reaction flask while it is cooling down.

CAUTION: If water is drawn back into hot concentrated sulfuric acid, a very dangerous situation can occur, particularly if larger quantities of the reagents than recommended above are used in the experiment.

NOTE. Do not remove the gas collection reservoir from the water bath. The beaker containing the reservoir is carried to the gas chromatograph for analysis of the collected gas.

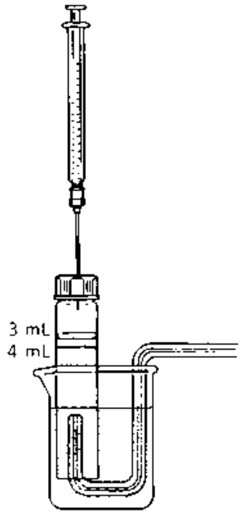

With the aid of a gas-tight syringe, withdraw a 0.5-cm^3 sample through the rubber septum for GC analysis (⬅).

Purification and Characterization. Analyze the collected gas by GC without further purification.

Gas Chromatographic Conditions
Column: 0.25 in. × 8 ft packed with 20% silicone DC 710
Room temperature
Flow rate: 20 mL/min (He gas)
Sample size: 0.5 cm^3 of collected gas

Record the literature values of the physical properties of the products.

The butanes have been determined to elute from the DC 710 column in the following order: 1-butene, *trans*-2-butene, and *cis*-2-butene. *If the reaction mixture is heated above the recommended temperatures, a rearrangement can occur to yield isobutene, which will be detected as an additional isomeric product.*

If we make the assumption that the amount of each substance in the gaseous mixture is proportional to the area under its corresponding GC peak, determining the relative amounts of the three components of the gas sample becomes a straightforward calculation. The accuracy is, of course, dependent on how well the three peaks are resolved in the gas chromatogram.

NOTE. In calculating relative quantities of alkenes formed in the reaction, several techniques may be used to quantitatively determine the composition of the gas mixture if an integrating recorder is not available. Two methods are described here:

1. *Determination of the areas under the peaks gives reproducible results of ±3–4% when these areas are assumed equal to the peak heights (mm) × the peak widths at half-height (mm), measured from the baseline of the curve.*

2. *An other way to determine the areas under the peaks is to cut out the peaks from the chromatogram and weigh them on an analytical balance (sensitivity to 0.1 mg). The weights of the peaks are directly proportional to the relative amount of each compound in the gas sample.*[10]

QUESTIONS

6-55. Gas chromatographic analysis of a mixture of organic compounds gave the following peak areas (cm^2): hexane = 2.7, heptane = 1.6, hexanol = 1.8, and toluene = 0.5.

(a) Calculate the mole percent composition of the mixture. Assume that the response of the detector (area per mole) is the same for each component.

(b) Calculate the weight percent composition of the mixture, using the same assumptions as in (a).

---

[10] See Pecsok, R. L. *Principles and Practices of Gas Chromatography;* Wiley: New York, 1961; p. 145.

6-56. At the end of the experiment we note that if the mixture is heated strongly, rearrangement can occur and isobutene (2-methyl-2-propene) is also formed. Suggest a mechanism to account for the formation of this compound.

6-57. When *tert*-pentyl bromide is treated with 80% ethanol, the following amounts of alkene products are detected on analysis:

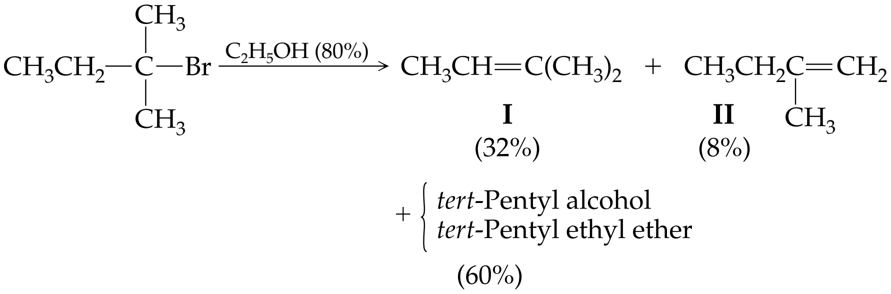

$$CH_3CH_2-\underset{CH_3}{\overset{CH_3}{C}}-Br \xrightarrow{C_2H_5OH\ (80\%)} \underset{\substack{\textbf{I}\\(32\%)}}{CH_3CH{=}C(CH_3)_2} + \underset{\substack{\textbf{II}\\(8\%)}}{CH_3CH_2\underset{CH_3}{C}{=}CH_2}$$

$$+\left\{\begin{array}{l}\textit{tert}\text{-Pentyl alcohol}\\ \textit{tert}\text{-Pentyl ethyl ether}\end{array}\right. \quad (60\%)$$

Offer an explanation of why compound **I** is formed in far greater amount than the terminal alkene.

6-58. The —$^{+}SR_2$ group is easily removed in elimination reactions, but the —SR group is not. Explain.

6-59. Why is sulfuric acid, rather than hydrochloric acid, used to catalyze the dehydration of alcohols?

BIBLIOGRAPHY

Several dehydration reactions of secondary alcohols using sulfuric acid as the catalyst are given in *Organic Syntheses:*

Adkins, H.; Zartman, W. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 606.

Bruce, W. F. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 12.

Coleman, G. H.; Johnstone, H. F. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 183.

Grummitt, O.; Becker, E. I. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 771.

Norris J. F. *Organic Syntheses*; Wiley: New York, 1941; Collect. Vol. I, p. 430.

Wiley, R. H.; Waddey, W. E. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 560.

For an overview of elimination reactions:

March, J. *Advanced Organic Chemistry,* 4th ed.; Wiley: New York, 1992, p. 982.

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed.; Wiley: New York, 2007, Chap 17.

For details on carbonium ion formation:

Olah, G. A.; Schleyer, P. von R., Eds. *Carbonium Ions;* Wiley: New York, Vol. 1, 1968; Vol. II, 1970; Vol. III, 1972.

Olah, G. A.; Prakash, G. K. S. *Carbocation Chemistry;* Wiley: New York, 2004.

Experiment [9] is adapted from the method given by: Helmkamp, G. K.; Johnson, H. W. Jr. *Selected Experiments in Organic Chemistry,* 3rd ed.; W. H. Freeman: New York, 1983; p. 99.

EXPERIMENT 10

The E2 Elimination Reaction: Dehydrohalogenation of 2-Bromobutane to Yield 1-Butene, trans-2-Butene, and cis-2-Butene

Common name: 1-butene
CA number: [06-98-9]
CA name as indexed: 1-butene

Common name: *trans*-2-butene
CA number: [624-64-6]
CA name as indexed: 2-butene, (*E*)-

Common name: *cis*-2-butene
CA number: [590-18-1]
CA name as indexed: 2-butene, (*Z*)-

Purpose. This reaction illustrates the base-induced dehydrohalogenation of alkyl halides with strong base and is used extensively for the preparation of alkenes. The stereo- and regiochemical effects of the size of the base is investigated, and the product mixture is analyzed by the use of gas chromatography.

Prior Reading

Technique 1: Gas Chromatography (pp. 55–61)
Technique 7: Collection or Control of Gaseous Products (pp. 105–107)

REACTION

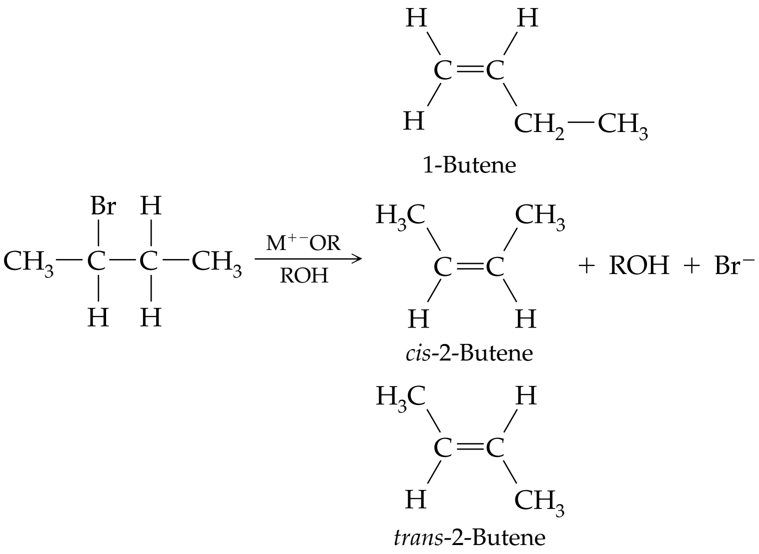

DISCUSSION

Base-induced elimination (dehydrohalogenation) of alkyl halides is a general reaction and is an excellent method for preparing alkenes. This process is often referred to as β *elimination*, since a hydrogen atom is always removed β to the halide (leaving group):

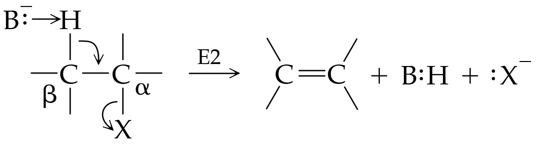

A high concentration of a strong base in a relatively nonpolar solvent is used to carry out the dehydrohalogenation reaction. Such combinations as sodium methoxide in methanol, sodium ethoxide in ethanol, potassium isopropoxide in isopropanol, and potassium *tert*-butoxide in *tert*-butanol or dimethyl sulfoxide (DMSO) are often used.

Elimination reactions almost always yield an isomeric mixture of alkenes, where this is possible. Under the reaction conditions, the elimination is *regioselective* and follows the Zaitsev rule when more than one route is available for the elimination of HX from an unsymmetrical alkyl halide. That is, the reaction proceeds in the direction that yields the most highly substituted alkene. For example,

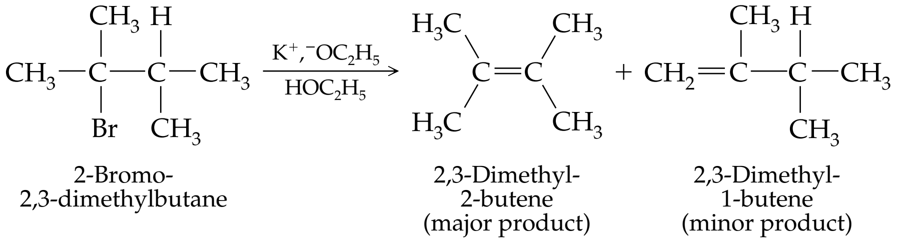
K$^+$,$^-$OC$_2$H$_5$
HOC$_2$H$_5$
2-Bromo-2,3-dimethylbutane
2,3-Dimethyl-2-butene (major product)
2,3-Dimethyl-1-butene (minor product)

In cases where cis or trans alkenes can be formed, the reaction exhibits stereo-selectivity, and the more stable trans isomer is the major product.

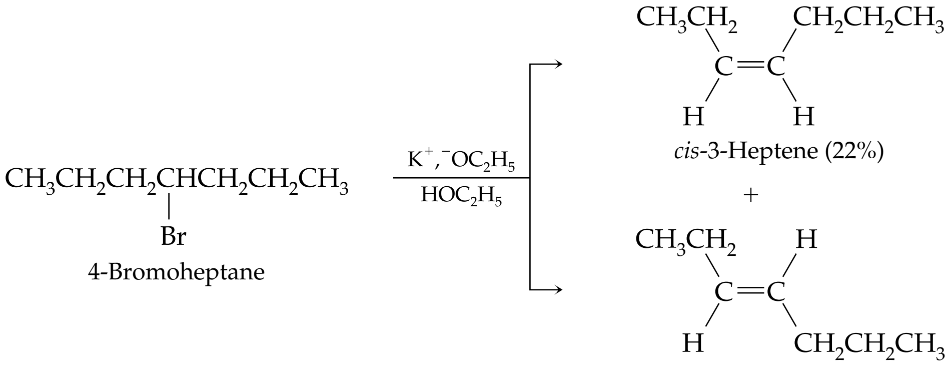

cis-3-Heptene (22%)

trans-3-Heptene (78%)

Experimental evidence indicates that the five atoms involved in the E2 elimination reaction must lie in the same plane; the anti-periplanar conformation is preferred. This conformation is necessary for the orbital overlap that must occur for the π bond to be generated in the alkene. The sp^3-hybridized atomic orbitals on carbon that comprise the C—H and C—X σ bonds broken in the reaction develop into the p orbitals comprising the π bond of the alkene formed:

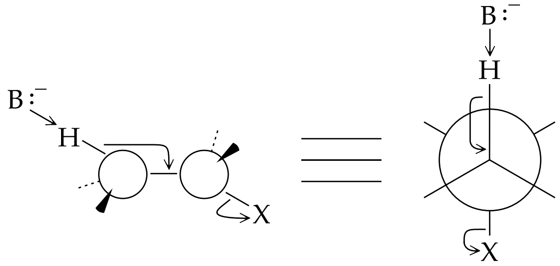

Anti-periplanar conformation

There is a smooth transition between reactant and product. Analogous to the S_N2 reaction, no intermediate has been isolated or detected. Furthermore, no rearrangements occur under E2 conditions. This situation is in marked contrast to E1 elimination reactions, where carbocation intermediates are generated and rearrangements are frequently observed (see Experiment [9]).

The alkyl halide adopts the anti-periplanar conformation in the transition state, and experimental evidence demonstrates that if the size of the base is increased, then it must be difficult for the large base to abstract an internal β-hydrogen atom. In such cases, the base removes a less hindered β-hydrogen atom, leading to a predominance of the thermodynamically **less stable (terminal) alkene** in the product mixture. This type of result is often referred to as **anti-Zaitsev** or **Hofmann elimination.** Thus, in the reaction of 2-bromo-2,3-dimethylbutane given above, the 2,3-dimethyl-1-butene would be the major product (anti-Zaitsev) if the conditions used a bulkier base. The anti-periplanar arrangements are illustrated in the Newman projections below.

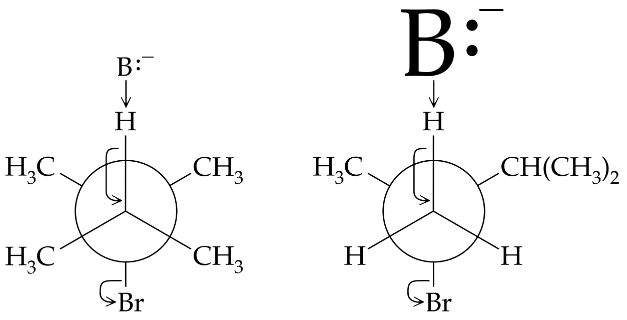

Zaitsev product
internal alkene

Anti-Zaitsev product
terminal alkene

Two anti-periplanar conformations
of 2-bromo-2,3-dimethylbutane

Dehydrohalogenation of alkyl halides in the presence of strong base (E2) is often accompanied by the formation of substitution (S_N2) products. The extent of the competitive substitution reaction depends on the structure of the alkyl halide. Primary alkyl halides give predominantly substitution products (the corresponding ether), secondary alkyl halides give predominantly elimination products, and tertiary alkyl halides give exclusively elimination products. For example, the reaction of 2-bromopropane with sodium ethoxide proceeds as follows:

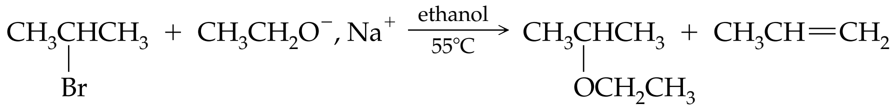

$$\underset{\text{Br}}{\mathrm{CH_3\underset{|}{C}HCH_3}} + \mathrm{CH_3CH_2O^-, Na^+} \xrightarrow[55°\mathrm{C}]{\text{ethanol}} \underset{\mathrm{OCH_2CH_3}}{\mathrm{CH_3\underset{|}{C}HCH_3}} + \mathrm{CH_3CH{=}CH_2}$$

2-Bromopropane — Sodium ethoxide — Ethyl isopropyl ether (21%) — Propene (79%)

In general, for the reaction of alkyl halides with strong base,

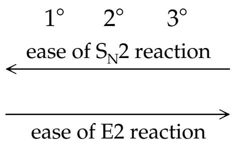

EXPERIMENTAL PROCEDURE

Estimated time of experiment: 2.5 h, if two reactions are run by each student.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
2-Bromobutane	137.03	100 μL	0.92	91.2	1.26	1.4366
Methanol	32.04	3.5 mL		64.9	0.791	1.3288
2-Propanol	60.09	3.5 mL		82.4	0.785	1.3776
2-Methyl-2-propanol (*tert*-butanol)	74.12	3.5 mL		82–83	0.786	1.3838
3-Ethyl-3-pentanol	116.20	3.5 mL		140–142	0.839	1.4266
Sodium	22.98	60 mg	2.6	883	0.97	
Potassium	39.10	60 mg	1.5	760	0.86	

Table 6.5 **Reagent Combinations**

Alcohol Solvent	Metal	Alkoxide Base Produced
Methanol	Sodium	Sodium methoxide
2-Propanol	Potassium	Potassium 2-propoxide
2-Methyl-2-propanol (*tert*-butanol)	Potassium	Potassium 2-methyl-2-propoxide (potassium *tert*-butoxide)
3-Ethyl-3-pentanol	Potassium	Potassium 3-ethyl-3-pentoxide

Reagents and Equipment. The combinations of reagents in Table 6.5 may be used to prepare the alkoxide base. Students should compare results to observe a total picture of the effect.

Preparation of the Alkoxide Base. Measure and add to a 5.0-mL conical vial containing a magnetic spin vane 3–3.5 mL of the anhydrous alcohol to be used (see Table 6.5). Add a 60-mg piece of potassium (or sodium) metal and immediately attach the vial to a reflux condenser protected by a calcium chloride drying tube. Place the arrangement in a sand bath and with stirring heat the mixture gently (~50 °C) (⬇).

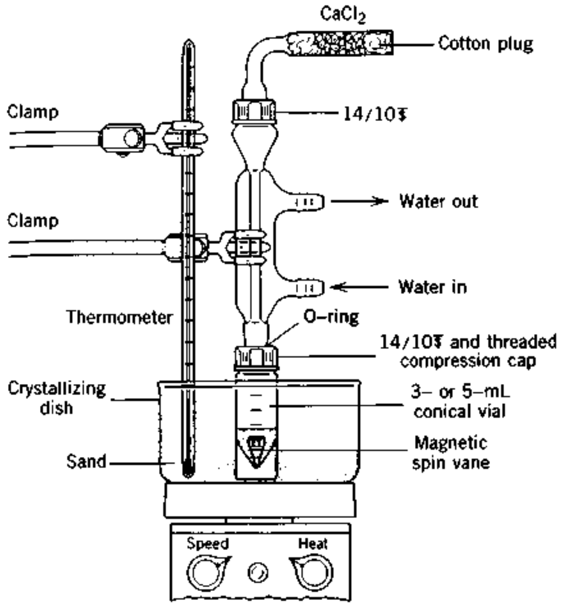

Assembly for preparation of alkoxide base

NOTE. If the sodium/methanol combination is used, it is not necessary to heat the mixture. A fairly vigorous reaction occurs at room temperature. It is recommended that the instructor cut the Na/K metal before commencing the laboratory.

CAUTION: Handle sodium and potassium with care. These metals react vigorously with moisture and are kept under paraffin oil or xylene. Remove the small piece of the metal from the oil using a pair of forceps or tongs—***never use your fingers!*** Dry the metal quickly by pressing it with filter paper (to soak up the oil), and immediately add it to the alcohol in the reaction vial. Any residual pieces of sodium/potassium should be stored in a bottle marked "sodium/ potassium residues." **Never** throw small pieces of these metals in the sink or in water. To destroy, in the **hood,** add small amounts of the metal to absolute ethanol.

HOOD

When all the metal has reacted, remove the assembly from the sand bath and cool to near room temperature (do not remove the drying tube.).

Table 6.6 Temperature Conditions

Base	Temperature (°C)
$NaOCH_3$	100–110
$KOCH(CH_3)_2$	130–140
$KOC(CH_3)_3$	140–150
$KOC(CH_2CH_3)_3$	175–180

Reaction Conditions. Remove the drying tube from the condenser and use a calibrated Pasteur pipet to introduce 100 μL of 2-bromobutane down through the condenser into the vial. Replace the drying tube and place the assembly in the **preheated** sand bath (see Table 6.6). Remove the drying tube and attach the gas delivery tube to the top of the condenser so that the open end of the tube is beneath the water level of the reservoir (⬇). If the connection to the top of the condenser is not made with an O-ring cap seal connection, lightly grease the ground-glass joint to insure a gas-tight seal. After about 10–15 air bubbles emerge, place the water-filled gas collection tube over the open end of the gas delivery tube.

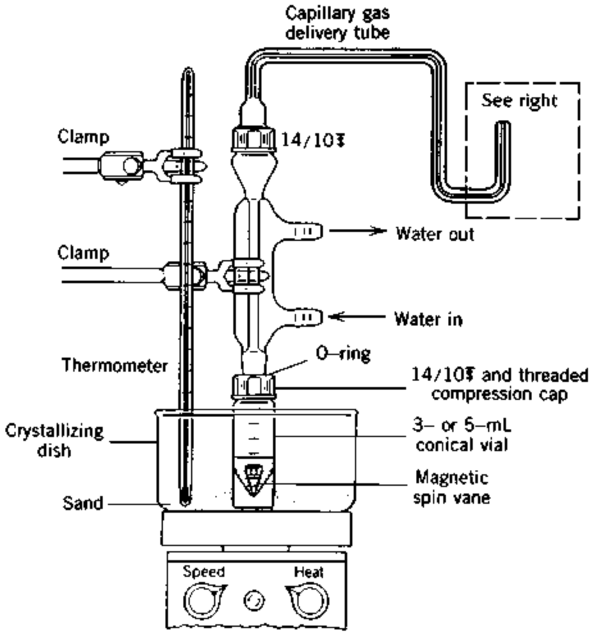
Capillary gas delivery tube
See right
Clamp
14/10 ₮
Water out
Clamp
Water in
Thermometer
O-ring
14/10 ₮ and threaded compression cap
Crystallizing dish
3- or 5-mL conical vial
Magnetic spin vane
Sand
Speed
Heat

Assembly for gas delivery

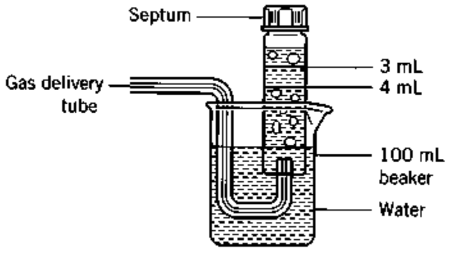
Septum
Gas delivery tube
3 mL
4 mL
100 mL beaker
Water

Arrangement for gas collection

Isolation of Product. Collect about 6–7 mL of gas in the collection reservoir and then use a hypodermic syringe to withdraw a 0.7- to 0.8-mL sample through the rubber septum for GC analysis.

NOTE. Remove the gas delivery tube from the collecting reservoir and then from the water before discontinuing the heat on the reaction vial. This order of events prevents water from being sucked back into the reaction flask.

Purification and Characterization. The collected gas is analyzed by gas chromatography without further purification.

Gas Chromatographic Conditions
Column: $\frac{1}{4}$ in. × 8 ft packed with 20% DC 710
Room temperature
Flow rate (He gas): 30 mL/min
Sample size: 0.7–0.8 mL of collected gas
Chart speed: 1 cm/min

Assuming that the amount of each substance in the gas is proportional to the areas of its corresponding peak, determine the ratio of the three components in the gas sample.

Area Under a Curve. *Several techniques may be used. The following method gives reproducible results of ±3–4%:* Area = peak height (mm) × width at half-height (mm), measured from the baseline of the curve.

The order of elution of the butenes is 1-butene, *trans*-2-butene, and *cis*-2-butene. Record the literature values of their physical properties.

QUESTIONS

6-60. Outline a complete mechanistic sequence for the reaction of 2-bromobutane with potassium 2-methyl-2-propoxide in 2-methyl-2-propanol solvent to form the three alkenes generated in the reaction (1-butene, *trans*-2-butene, and *cis*-2-butene). Include a clear drawing of the anti-periplanar transition state for the formation of each alkene.

6-61. Does the mixture of gases collected in this experiment consist only of alkenes? If not, what other gases might be present?

6-62. Predict the predominant alkene product that would form when 2-bromo-2-methylpentane is treated with sodium methoxide in methanol. If the base were changed to $KOC(CH_2CH_3)_3$ would the same alkene predominate? If not, why? What would be the structure of this alternate product, if it formed?

6-63. Predict the more stable alkene of each of the following pairs:
(a) 1-Hexene or *trans*-3-hexene
(b) *trans*-3-Hexene or *cis*-3-hexene
(c) 2-Methyl-2-hexene or 2,3-dimethyl-2-pentene

6-64. Starting with the appropriate alkyl halide and base–solvent combination, outline a synthesis that would yield each of the following alkenes as the major or only product:
(a) 1-Butene
(b) 3-Methyl-1-butene
(c) 2,3-Dimethyl-1-butene
(d) 4-Methylcyclohexene

6-65. When *cis*-1-bromo-4-*tert*-butylcyclohexane reacts with sodium ethoxide in ethanol, it reacts rapidly to yield 4-*tert*-butylcyclohexene. Under similar conditions, *trans*-1-bromo-4-*tert*-butylcyclohexane reacts very slowly. Using conformational structures, explain the difference in reactivity of these cis–trans isomers.

BIBLIOGRAPHY

Several dehydrohalogenation reactions of alkyl halides using alkoxide bases are given in *Organic Syntheses:*

Allen, C. F.; Kalm, M. J. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 398.

McElvain, S. M.; Kundiger, D. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 506.

Paquette, L. A.; Barrett, J. H. *Organic Syntheses*; Wiley: New York, 1973; Collect. Vol. V, p. 467.

Schaefer, J. P.; Endres, L. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 285.

For an overview of elimination reactions:

March, J. *Advanced Organic Chemistry,* 4th ed.; Wiley: New York, 1992, p. 982.

Smith M. B.; March, J. *Advanced Organic Chemistry*, 6th ed.; Wiley: New York, 2007, Chap. 17.

Experiment [10] is adapted from the method given by:

Leone, S. A.; Davis, J. D. *J. Chem. Educ.* **1992,** 69, A175.

EXPERIMENT 11

The Isolation of Natural Products

These experiments are designed to acquaint you with the procedures used to isolate naturally occurring and often biologically active organic compounds. These substances are known as *natural products* because they are produced by living systems. The particular natural products you are going to study come from the plant kingdom. At the end of the nineteenth century more than 80% of all medicines in the Western world were natural substances found in roots, barks, and leaves. There was a widespread belief at that time that in plants there existed cures for all diseases. As Kipling wrote, "Anything green that grew out of the mold/ Was an excellent herb to our fathers of old." Even as the power of synthetic organic chemistry has grown during this century, natural materials still constitute a significant fraction of the drugs used in modern medicine. For example, in the mid-1960s when approximately 300 million new prescriptions were written each year, nearly half were for substances of natural origin. These materials have played a major role in successfully combating the worst of human illnesses, from malaria to high blood pressure; diseases that affect hundreds of millions of people.

Unfortunately, during the latter half of this century a number of very powerful natural products that subtly alter the chemistry of the brain have been used in vast quantities by our society. The ultimate impact on civilization is of grave concern. Evidence clearly demonstrates that these natural substances disrupt the exceedingly complex and delicate balance of biochemical reactions that lead to normal human consciousness. How well the brain is able to repair the damage from repetitive exposure is unknown. We are currently conducting experiments to answer that question.

The natural products that you may isolate in the following experiments include a bright-yellow crystalline antibiotic (Experiment [11A]), a white crystalline alkaloid that acts as a stimulant in humans (Experiment [11B]), and an oily material with a pleasant odor and taste (Experiment [11C]).

Experiment 11A

Isolation and Characterization of an Optically Active Natural Product: Usnic Acid

Common name: usnic acid

CA number: [7562-61-0]

CA name as indexed: 1,3(2*H*,9b*H*)-dibenzofurandione, 2,6-diacetyl-7, 9-dihydroxy-8,9b-dimethyl-

Purpose. In this exercise you will extract the active principle, usnic acid, from one of the lichens that produce it. Usnic acid is a metabolite found in a variety of lichens. For this experiment we utilize a local (in Maine) species of lichen, *Usnea hirta* (often referred to as *old-man's beard)*, which is a fruticose lichen (a lichen that possesses erect, hanging, or branched structures). The extraction technique illustrated here is often used to isolate natural products from their native sources (see also Experiment [11B], on page 229, for another extraction strategy). Because usnic acid possesses a single chiral center (stereocenter) and only one of the enantiomers is produced in old-man's beard, this experiment also functions as an introduction to the methods used to measure the specific rotation of optically active substances.

LICHENS AND NATURAL PRODUCTS

Lichens, of which there are estimated to be greater than 15,000 species, are an association between an algae and a fungus that live together in an intimate relationship. This association is often called symbiosis. Symbiosis requires that two different organisms live together in both close structural proximity and interdependent physiological combination. The term ordinarily is applied to situations where the relationship is advantageous, or even required, for one or both, but not harmful to either. In the case of lichens, the algae can be grown independently of the fungi that obtain nutrients from the algae cells. The fungi are, therefore, considered to be parasitic and their contribution to the union has been viewed historically only as an aid in the absorption and retention of water and perhaps to provide a protective structure for the algae. It appears, however, that the fungi may play a far more important role in the life of the lichen than earlier appreciated. The fungi appear to generate a metabolite, usnic acid, which is the most common substance found in these primitive systems. This acid can comprise up to 20% of the dry weight of some lichens! Even more intriguing is the original belief that usnic acid appears to have *no biological function* in these plants. Why would a living system channel huge amounts of its precious energy into making an apparently useless substance? Recently, with our increased understanding of the role of chemical communication substances in ecology, it has been recognized that usnic acid very likely makes a major symbiotic contribution as a *chemical defense agent*. Indeed, in 1945 Burkholder demonstrated that several New England lichens possess antibiotic properties, and usnic acid was subsequently shown to be the active agent against several kinds of bacteria, including staphylococcus. The Finnish company, Lääke Oy Pharmaceutical, has prepared from reindeer lichen a broad-spectrum usnic acid antibiotic for treating tuberculosis and serious skin infections. There is, in fact, evidence that lichens were used in medicine by the ancient Egyptians, and from 1600 to 1800 C.E. these plants were considered an outstanding cure for tuberculosis. Usnic acid has been investigated for use as an antibiotic by the U.S. Public Health Service. It proved to be effective in dilutions between 1 part in 100,000 and 1 part in 1,000,000 against several Gram-positive organisms.

This widespread lichen metabolite is the material isolated in this experiment. Usnic acid was first isolated and identified in 1843 by Rochleder, but a molecule of this complexity was beyond the structural knowledge of organic chemistry in those days. The structure was finally determined in 1941 by Schöpf, and in 1956 it was synthesized in the laboratory by Sir D. H. R. Barton (Nobel Laureate). Barton's route involved a spectacular one-step dimerization of a simple precursor, a synthesis that very closely mimicked the actual biogenetic pathway (see chemistry). The key step was the one-electron (1 e^-) oxidation of methylphloracetophenone, which leads directly to the dimerization. The mechanism of this reaction, both in the plant and in the laboratory synthesis, is essentially identical to the oxidative coupling of 2-naphthol to give 1,1′-bi-2-naphthol, which is explored in detail in Experiment [5_{adv}]:

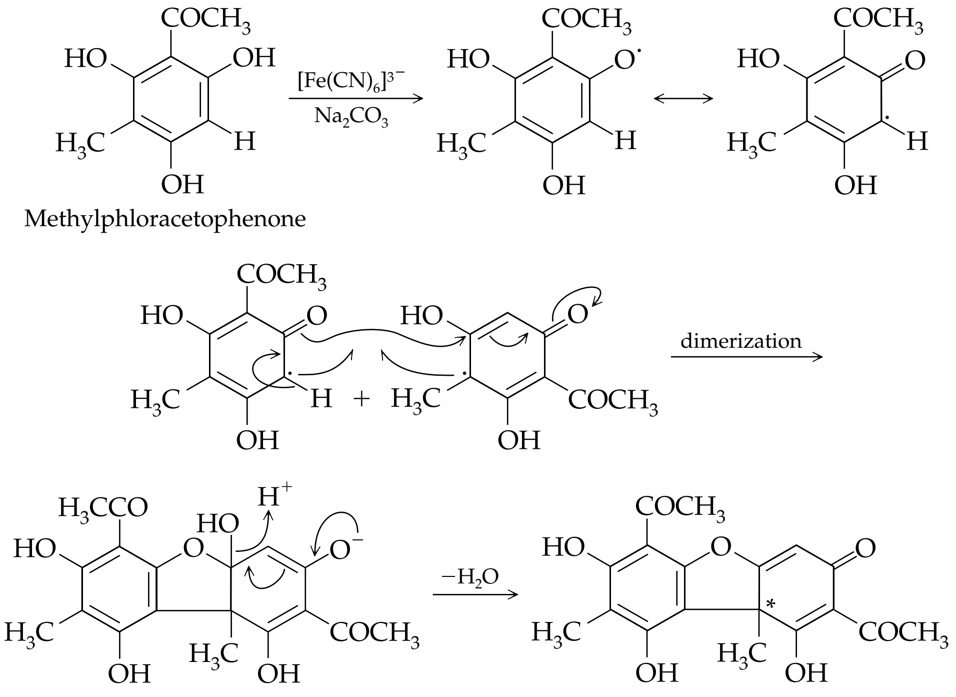

Methylphloracetophenone

Usnic Acid

The chiral center (stereocenter) (*) is bonded to a highly conjugated aromatic ring system (see structure), which gives rise to a very large specific rotation. This enhanced interaction with polarized radiation makes this compound a particularly interesting molecule to examine for optical activity. The production of a single enantiomer in the natural product, which, as discussed above, is formed by an oxidative coupling process, implies that there must be an intimate association between the substrate and an enzyme (a biological catalyst that itself is optically active) during the crucial coupling process.[11]

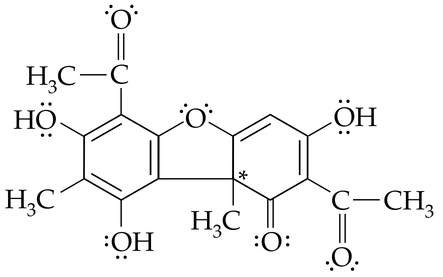

Prior Reading

Technique 6A: Thin-Layer Chromatography (pp. 97–99)

Technique 8: Measurement of Specific Rotation
Optical Rotation Theory (pp. 108–111)

---

[11] Dean, F. M.; Halewood, P.; Mongkolsuk, S.; Roberston, A.; Whally, W. B. *J. Chem. Soc.* **1953,** 1250. Kreig, M. B. *Green Medicine;* Rand McNally: New York, 1964. Lewis, W. H.; Elvin-Lewis, M. P. F. *Medical Botany;* Wiley: New York, 1977. Hendrickson, J. B. *The Molecules of Nature;* W. A. Benjamin, New York: 1965. Richards, J. H.; Hendrickson, J. B. *The Biosynthesis of Steroids, Terpenes, and Acetogenins;* W. A. Benjamin: New York, 1964. Schöpf, C.; Ross, F. *Annalen* **1941,** *546,* 1 (see further references cited in Experiment [5_{adv}]).

DISCUSSION

Nonracemic solutions of chiral substances, when placed in the path of a beam of polarized light, may rotate the plane of the polarized light clockwise or counterclockwise and are thus referred to as *optically active*. This angle of optical rotation is measured using a *polarimeter*. This technique is applicable to a wide range of analytical problems varying from purity control to the analysis of natural and synthetic products in the medicinal and biological fields. The results obtained from the measurement of the observed angle of rotation (α_{obs}) are generally expressed in terms of *specific rotation* $[\alpha]$. The sign and magnitude of $[\alpha]$ are dependent on the specific molecule and are determined by complex features of molecular structure and conformation, and thus cannot be easily explained or predicted. The relationship of $[\alpha]$ to α_{obs} is $[\alpha]_\lambda^T = \dfrac{\alpha_{\text{obs}}}{l \cdot c}$ where T is the temperature of the sample **in degrees Celsius** (°C), l is the length of the polarimeter cell **in decimeters** (1 dm = 0.1 m = 10 cm), c is the concentration of the sample **in grams per milliliter** (g/mL), and λ is the wavelength of the light used in the polarimeter **in nanometers** (nm). These units are traditional, though most are esoteric by contemporary standards. Thus, the specific rotation for a given compound is normally reported in terms of temperature, wavelength, concentration, and the nature of the solvent. For example; $[\alpha]_\text{D}^{25} = +12.3°$ ($c = 0.4$, $CHCl_3$) implies that the measurement was recorded in a $CHCl_3$ solution of 0.4 g/mL at 25 °C using the sodium D line (589 nm) as the light source. Unless indicated, the pathlength is assumed to be 1 dm in these observations.

Usnic acid contains a single stereocenter (see structure), and therefore it can exist as a pair of enantiomers. In nature, however, only one of the enantiomers (*R* or *S*) would be expected to be present. Usnic acid has a very high specific rotation, $[\alpha]_\text{D}^{25} = +488°$ ($c = 0.4$, $CHCl_3$), which will give a large α_{obs} even at low concentrations, and for this reason it is an ideal candidate to measure rotation in a microscale experiment.

Racemic (equimolar amounts of each enantiomer) usnic acid has been resolved (separated into the individual enantiomers) through preparation and separation of the diastereomeric (−) brucine salts. This procedure was the route followed to obtain an authentic synthetic sample for comparison with the natural material. The separation was required because the dimerization step in the synthesis, which was carried out in the absence of enzymatic, or other chiral, influence, gave a racemic product.

A common method of extracting chemical constituents from natural sources is presented in this experiment. In this case, only one chemical compound, the usnic acid, is significantly soluble in the extraction solvent, acetone. For this reason, the isolation sequence is straightforward.

EXPERIMENTAL PROCEDURE

Isolation of Usnic Acid. Estimated time for completion of the experiment: 2.5 h.

Physical Properties of Components

Compound	MW	Amount	bp (°C)
Lichen		1.0 g	
Acetone	58.08	15.0 mL	56.2

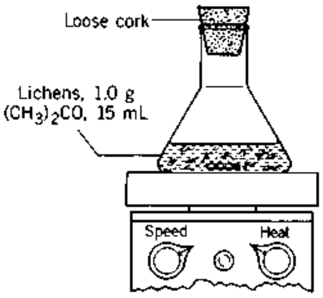

Reagents and Equipment. Weigh and place about 1.0 g of oven-dried (40 °C) crushed or cut-up lichens and 15.0 mL of acetone in a 50-mL Erlenmeyer flask containing a magnetic stirrer. Loosely cap the flask with a cork stopper (⬅). The lichens used in this experiment are *Usnea hirta*.

Reaction Conditions. Stir or occasionally swirl the mixture for no less than 30 min at room temperature. If necessary, periodically push the lichens below the surface of the acetone solvent using a glass rod.

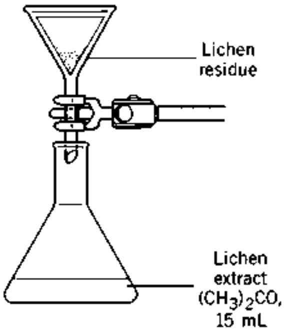

HOOD

Isolation of Product. Prior to filtering the resulting mixture, remove a 0.5 mL aliquot for final analysis. Filter by gravity and collect the filtrate in a 25-mL Erlenmeyer flask (⬅). A Pasteur filter pipet may be used to make this transfer, if desired. In the **hood,** remove the acetone solvent under a slow stream of air or nitrogen on a warm sand bath nearly to dryness. Allow the remainder of the acetone to evaporate at room temperature to obtain the crude bright yellow or orange usnic acid crystals.

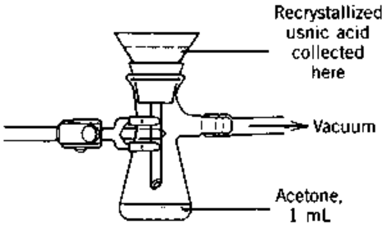

Purification and Characterization. Recrystallize all but 5 mg of the crude extract from acetone/95% ethanol (10:1). Dissolve the crystals in the minimum amount of hot acetone, keeping the recrystallization vessel hot, and add the appropriate volume of 95% ethanol. Allow the mixture to cool to room temperature and then place the flask in an ice bath to complete the recrystallization. Collect the golden-yellow crystals by vacuum filtration (⬅) and wash them with *cold* acetone. Dry the crystals on a porous clay plate or on a sheet of filter paper. As an alternative and more efficient procedure, the crude material may be recrystallized using a Craig tube, avoiding the filtration step with the Hirsch funnel.

V

Weigh the yellow needles of usnic acid and calculate the percentage of the acid extracted from the dry lichen. Determine the melting point (use the evacuated melting point technique) and compare your value to that found in the literature. Using a solvent system of ethyl acetate:hexane (1:4) and a UV lamp for visualization, compare the crude extract with the purified usnic acid by TLC (R_f value for usnic acid is 0.32). Obtain an IR spectrum and compare it with that of an authentic sample or one from the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Chemical Tests. Chemical tests can assist in establishing the nature of the functional groups in usnic acid. Perform the 2,4-dinitrophenylhydrazine test and the ferric chloride test (see Chapter 9). Are the results significant?

Determination of the Specific Rotation. Though usnic acid is an optically active compound with a very high specific rotation, a low-volume, long-pathlength cell must be used to successfully determine its specific rotation with microscale quantities.

Dissolve usnic acid (80 mg) in 4.0 mL of tetrahydrofuran (THF) solvent and transfer the solution to the polarimeter cell using a Pasteur pipet.

NOTE. To obtain this quantity (80 mg) of usnic acid will very likely require pooling the recrystallized product of eight or nine students. Spectral grade THF should be used as the solvent. Many of the early specific rotation values on these substances were recorded with chloroform as the solvent, but, because it possesses some toxicity, it is now avoided if possible.

Place the cell in the polarimeter and measure the angle of rotation. Calculate the specific rotation using the equation given in the discussion section.

QUESTIONS

6-66. Determine the correct *R* or *S* designation for each of the following molecules:

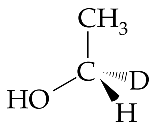

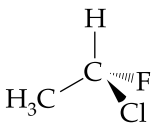

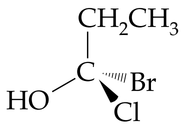

6-67. The structure originally proposed for cordycepic acid, which has [α] = +40.3°, was

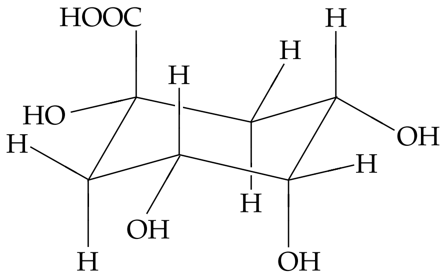

Why is this not a plausible structure?

6-68. A sample of 150 mg of an organic compound is dissolved in 7.5 mL of water. The solution is placed in a 20-cm polarimeter tube and the rotation is measured in a polarimeter. The rotation observed is +2.676°. Distilled water, in the same tube, gave a reading of +0.016°. Calculate the specific rotation for the compound.

6-69. Compound A is optically active and has the molecular formula $C_5H_{10}O$. On catalytic hydrogenation (addition of hydrogen) of A, compound B is obtained. Compound B has the molecular formula $C_5H_{12}O$ and is optically inactive. Give the structure for compounds A and B.

6-70. Which of the following compounds have a meso form?
- **(a)** 2,3-Dibromopentane
- **(b)** 2,4-Dibromopentane
- **(c)** 2,3-Dibromobutane

BIBLIOGRAPHY

This experiment is adapted from that given by:

Todd, *D. Experimental Organic Chemistry;* Prentice-Hall: Englewood Cliffs, NJ, **1979,** p. 57.

Synthesis of usnic acid:

Barton, D. H. R.; DeFlorin, A. M.; Edwards, O. E. *J. Chem. Soc.* **1956,** 530.

Penttila, A., Fales, H. M. *Chem. Commun.* **1966,** 656.

A large-scale method of isolation of usnic acid has been reported:

Stark, J. B.; Walter, E. D.; Owens, H. S. *J. Am. Chem. Soc.* **1950,** *72,* 1819.

Optical, crystallographic and X-ray diffraction data have been reported for usnic acid:

Jones, F. T.; Palmer, K. J. *J. Am. Chem. Soc.* **1950,** *72,* 1820.

Isolation and Characterization of a Natural Product: Caffeine and Caffeine 5-Nitrosalicylate

Experiment 11B

Product

Common names: caffeine, 1,3,7-trimethyl-2,6-dioxopurine

CA number: [58-08-2]

CA name as indexed: 1*H*-purine-2,6-dione, 3,7-dihydro-1,3,7-trimethyl-

Purpose. To extract the active principle, an alkaloid, caffeine, from a native source, tea leaves. Caffeine is a metabolite (a product of the living system's biochemistry) found in a variety of plants. We will use ordinary tea bags as our source of raw material. This experiment illustrates an extraction technique often used to isolate water-soluble, weakly basic natural products from their biological source (see also Experiment [11A] for another extraction strategy). The isolation of caffeine will also give you the opportunity to use sublimation as a purification technique, since caffeine is a crystalline alkaloid that possesses sufficient vapor pressure to make it a good candidate for this procedure. In addition, the preparation of a derivative of caffeine, its 5-nitrosalicylate salt, will be carried out. This latter conversion takes advantage of the weakly basic character of this natural product.

ALKALOIDS

Caffeine belongs to a rather amorphous class of natural products called alkaloids. This collection of substances is unmatched in its variety of structures, biological response on nonhost organisms, and the biogenetic pathways to their formation.

The history of these fascinating organic substances begins at least 4000 years ago. They were incorporated into poultices, potions, poisons, and medicines, but no attempt was made to isolate and identify the substances responsible for the physiological response until the very early 1800s.

The first alkaloid to be obtained in the pure crystalline state was morphine. Friedrich Wilhelm Sertürner (1783–1841) isolated morphine in 1805. He recognized that the material possessed basic character and he, therefore, classified it as a vegetable alkali (that is, a base with its origin in the plant kingdom). Thus, compounds with similar properties ultimately became known as alkaloids. The term "alkaloid" was introduced for the first time by an apothecary, Meissner, in Halle in 1819.

Sertürner, also a pharmacist, lived in Hamelin, another city in Prussia. He isolated morphine from opium, the dried sap of the poppy. Since the analgesic and narcotic effects of the crude resin had been known for centuries, it is not surprising that, with the emerging understanding of chemistry, the interest of Sertürner became focused on this drug, which is still medicine's major therapy for intolerable pain. He published his studies in detail in 1816 and very quickly two French professors, Pierre Joseph Pelletier (1788–1842) and Joseph Caventou (1795–1877) at the Ecole de Pharmacie in Paris, recognized the enormous importance of Sertürner's work.

In the period from 1817 to 1820, these two men and their students isolated many of the alkaloids, which continue to be of major importance. Included in that avalanche of purified natural products was caffeine, which they obtained from the coffee bean. This substance is the target compound that you will be isolating directly from the raw plant in this experiment. A little more than 75 years later, caffeine was first synthesized by Fischer in 1895 from dimethylurea and malonic acid.

THE CLASSIFICATION OF ALKALOIDS

These compounds are separated into three general classes of materials.

1. **True alkaloids:** These compounds contain nitrogen in a heterocyclic ring; are almost always basic (the lone pair of the nitrogen is responsible for

this basic character); are derived from amino acids in the biogenesis of the alkaloid; invariably are toxic and possess a broad spectrum of pharmacological activity; are found in a rather limited number of plants (of the 10,000 known genera only 8.7% possess at least one alkaloid); and normally occur in a complex with an organic acid (this helps to make them rather soluble in aqueous media). As we will see, there are numerous exceptions to these rules. For example, there are several very well-known quaternary alkaloids. These are compounds in which the nitrogen has become tetravalent and positively charged (as in the ammonium ion). Thus, they are not actually basic.

2. **Protoalkaloids:** These compounds are simple amines, derived from amino acids, in which the basic nitrogen atom is not incorporated into a ring system; they are often referred to as *biological amines*. An example of a protoalkaloid is mescaline.
3. **Pseudoalkaloids:** These compounds contain nitrogen atoms usually *not* derived from amino acids. There are two main classes into which pseudoalkaloids are divided, the steroidal alkaloids and the *purines*. Caffeine has been assigned to this latter class of alkaloids.

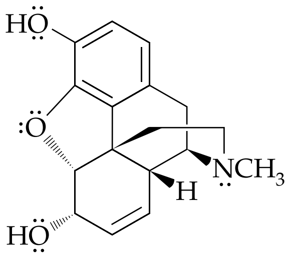

Morphine

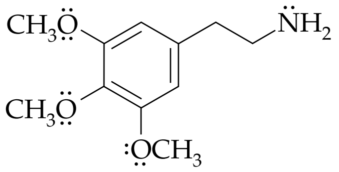

Mescaline

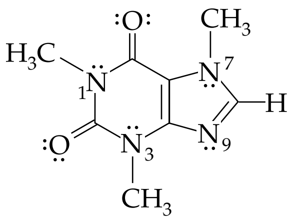

Caffeine

Prior Reading

Technique 4: Solvent Extraction
Solid–Liquid Extraction (p. 79)
Liquid–Liquid Extraction (p. 72)
Technique 9: Sublimation
Sublimation Theory (pp. 112–113)

DISCUSSION

Caffeine (1,3,7-trimethylxanthine) and its close relative theobromine (3,7-dimethyl-xanthine) both possess the oxidized purine skeleton (xanthine). These compounds are often classified as pseudoalkaloids, since only the nitrogen atom at the 7 position can be traced to an amino group originally derived from an amino acid (in this case glycine). This classification emphasizes the rather murky problem of deciding just what naturally occurring nitrogen bases are *true* alkaloids. We will simply treat caffeine as an alkaloid.

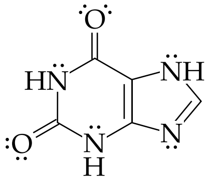

Xanthine

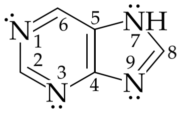

Purine

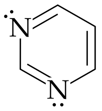

Pyrimidine

Although the pyrimidine ring (present in caffeine's purine system) is a significant building block of nucleic acids, it is rare elsewhere in nature.

These two methylated xanthines are found in quite a number of plants and have been extracted and widely used for centuries. Indeed, they very likely have been, and remain today, the predominant stimulant consumed by humans. Every time you make a cup of tea or coffee, you perform an aqueous extraction of plant material (tea leaves, *Camellia sinenis*, 1–4%, or coffee beans, *Coffea* spp., 1–2%) to obtain a dose of 25–100 mg of caffeine. Caffeine is also the active substance (~2%) in maté (used in Paraguay as a tea) made from the leaves of *Ilex paraguensis*. In coffee and tea, caffeine is the dominant member of the pair, whereas in *Theobroma cacao*, from which we obtain cocoa, theobromine (1–3%) is the primary source of the biological response. Caffeine acts to stimulate the central nervous system with its main impact on the cerebral cortex, and as it makes one more alert, it is no surprise that it is the chief constituent in No-Doz® pills.

Caffeine is readily soluble in hot water, because the alkaloid is often bound in thermally labile, partially ionic complexes with naturally occurring organic acids, such as with 3-caffeoylquinic acid in the coffee bean. For this reason it is relatively easy to separate caffeine from black tea leaves by aqueous extraction.

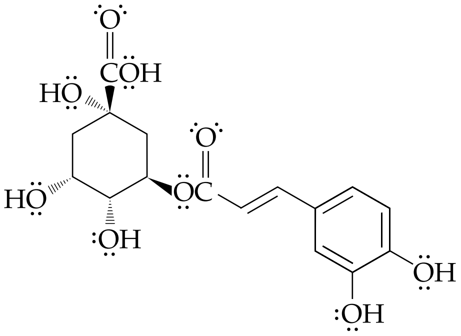

3-Caffeoylquinic acid

Other substances, mainly tannic acids, are also present in the tea leaves and are also water soluble. The addition of sodium carbonate, a base, during the aqueous extraction helps to increase the water solubility of these acidic substances by forming ionic sodium salts and liberating the free base.

Subsequent extraction of the aqueous phase with methylene chloride, in which free caffeine has a moderate solubility, allows the transfer of the caffeine from the aqueous extract to the organic phase. At the same time, methylene chloride extraction leaves the water-soluble sodium salts of the organic acids behind in the aqueous phase.

Extraction of the tea leaves directly with nonpolar solvents (methylene chloride) to remove the caffeine gives very poor results—since, as we have seen, the caffeine is bound in the plant in a partially ionic complex that will not be very soluble in nonpolar solvents. Thus, water is the superior extraction solvent for this alkaloid. The water also swells the tea leaves and allows for easier transport across the solid–liquid interface.

Following extraction and removal of the solvent, sublimation techniques are applied to the crude solid residues to purify the caffeine. This technique is especially suitable for the purification of solid substances at the microscale level, if they possess sufficient vapor pressure. Sublimation techniques are particularly advantageous when the impurities present in the sample are nonvolatile under the conditions used.

Sublimation occurs when a substance goes directly from the solid phase to the gas phase upon heating, bypassing the liquid phase. Sublimation is technically a straightforward method for purification in that the materials need only be heated and therefore, mechanical losses can be kept to a minimum (the target substance must, of course, be thermally stable at the required temperatures). Materials sublime only when heated *below* their melting points, and reduced pressure is usually required to achieve acceptable sublimation rates. Obviously, substances that lend themselves best to purification by sublimation are those that do not possess strong intermolecular attractive forces. Caffeine and ferrocene (used as a reactant in Experiment [27]) meet these criteria because they present large flat surfaces occupied predominantly with repulsive π electrons. For other isolation methods, see the discussion of *solid-phase* extraction in Technique 4, where caffeine is extracted from coffee beans (pp. 83–84).

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Constituents

Compound	MW	Amount	mmol	mp (°C)
Tea		1.0 g		
Water		10 mL		
Sodium carbonate	105.99	1.1 g	10	851

Reagents and Equipment. Carefully open a commercial tea bag (2.0–2.5 g of tea leaves) and empty the contents. Weigh out 1.0 g of tea leaves and place them back in the empty tea bag. Close and secure the bag with staples.

Weigh, and add to a 50-mL Erlenmeyer flask, 1.1 g (0.01 mol) of anhydrous sodium carbonate followed by 10 mL of water (➡). Heat the mixture with occasional swirling on a hot plate to dissolve the solid. Now add the 1.0 g of tea leaves (in the tea bag) to the solution. Place the bag in the flask so that it lies flat across the bottom.

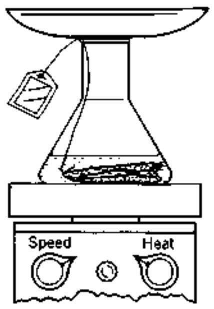

Anhydrous Na_2CO_3, 1.1 g + H_2O, 10 mL + tea bag with tea leaves, 1.0 g

Reaction Conditions. Place a small watch glass over the mouth of the Erlenmeyer flask and then heat the aqueous suspension to **gentle** boiling for 30 min on the hot plate.

GENTLY

Isolation of Product. Cool the flask and contents to room temperature. Transfer the aqueous extract from the Erlenmeyer flask to a 12- or 15-mL centrifuge tube using a Pasteur filter pipet. In addition, gently squeeze the tea bag by pressing it against the side of the Erlenmeyer flask to recover as much of the basic extract as possible. Set aside the tea bag and its contents.

Extract the aqueous solution with 2.0 mL of methylene chloride.

NOTE. The tea solution contains some constituents that may cause an emulsion. If you obtain an emulsion during the mixing of the aqueous and organic solvent layers (by shaking or using a Vortex mixer), it can be broken readily by centrifugation.

Separate the lower (methylene chloride) layer (check to make sure that the lower layer is, indeed, the organic layer by testing the solubility of a few drops of it in a test tube with distilled water) using a 9-in. Pasteur pipet. Drain the wet extracts through a filter funnel containing a small plug of cotton that is covered with about 2.0 g of anhydrous sodium sulfate, previously "moistened" with a small amount of methylene chloride (⬅). (The organic phase will be saturated with water following the extraction, therefore it is referred to as "wet." It also may contain a few droplets of the aqueous phase, which become entrained during the phase separation; this can be particularly troublesome if an emulsion forms during the mixing.)

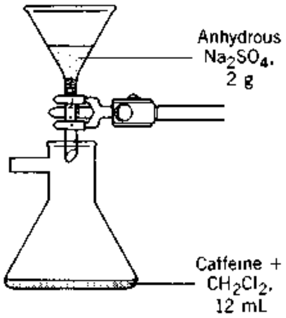

Collect the dried filtrate in a 25-mL filter flask. Extract the remaining aqueous phase with four additional 2.0-mL portions of methylene chloride (4 × 2 mL). Each extraction (an extraction is often referred to as a washing) is separated, dried as above, and transferred to the same filter flask. Finally, rinse the sodium sulfate with an additional 2.0 mL of methylene chloride and combine this wash with the earlier organic extracts.

HOOD

Add a boiling stone to the flask and concentrate the solution to dryness in the **hood** by warming the flask in a sand bath. The crude caffeine should be obtained as an off-white crystalline solid.

Purification and Characterization. Purify the crude solid caffeine by sublimation.

Assemble a sublimation apparatus as shown in Figure 5.53, on page 113; either arrangement is satisfactory. Using an aspirator, apply a vacuum to the system through the filter flask (remember to install a water-trap bottle between the sidearm flask and the aspirator). After the system is evacuated, run cold water gently through the cold finger or add ice to the centrifuge tube. By cooling the surface of the cold finger *after* the system has been evacuated, you will minimize the condensation of moisture on the area where the sublimed sample will collect.

NOTE. Less caffeine will be lost if the bottom of the cold finger is positioned less than 5 mm from the bottom of the filter flask.

Once the apparatus is evacuated and cooled, begin the sublimation by gently heating the flask with a microburner or sand bath. If you use a gas burner, always keep moving the flame back and forth around the bottom and sides of the flask.

NOTE. Be careful. Do not melt the caffeine. If the sample does begin to melt, remove the flame for a few seconds before heating is resumed. Overheating the crude sample will lead to decomposition and the deposition of impurities on the cold finger. High temperatures are not necessary since the sublimation temperature of caffeine (and of all solids that sublime) is below the melting point. It is generally worthwhile to carry out sublimations as slowly as possible, as the purity of the material collected will be enhanced.

When no more caffeine will sublime onto the cold finger, remove the heat, shut off the aspirator and the cooling water to the cold finger, and allow the apparatus to cool to room temperature under reduced pressure. Once cooled, carefully vent the vacuum and return the system to atmospheric pressure. *Carefully* remove the cold finger from the apparatus.

NOTE. If the removal of the cold finger is done carelessly, the sublimed crystals may be dislodged from the sides and bottom of the tube and drop back onto the residue left in the filter flask.

Scrape the caffeine from the cold finger onto weighing paper using a microspatula and a sample brush. Weigh the purified caffeine and calculate its percent by weight in the original tea leaves. Determine the melting point and compare your value to that in the literature.

If your melting point apparatus uses capillary tubes to determine the melting point, an evacuated sealed tube is necessary, since caffeine sublimes; the melting point is above the sublimation temperature (see Chapter 4). The melting point may be obtained using the Fisher–Johns apparatus without this precaution.

Obtain an IR spectrum and compare it with that of an authentic sample.

Chemical Test. Does the soda lime or the sodium fusion test (see Chapter 9) confirm the presence of nitrogen in your caffeine product?

DERIVATIVE: CAFFEINE 5-NITROSALICYLATE

It is not completely surprising to find that caffeine in the coffee bean is bound in a thermally labile complex with acid, since this alkaloid is a weakly basic substance possessing a base strength somewhat greater than that of an aryl amide. Because the purine ring system of caffeine has little reactive functionality, the formation of simple chemical derivatives is limited. The ease of association with high-melting carboxylic acids, however, offers a route to a variety of materials with well-defined melting points.

One such acid is 5-nitrosalicylic acid, which is prepared by nitration of salicylic acid in Experiment [29]. The purified caffeine obtained in this experiment may be further characterized by preparation of the 5-nitrosalicylate complex. This association is similar to the natural one formed with 3-caffeoylquinic acid:

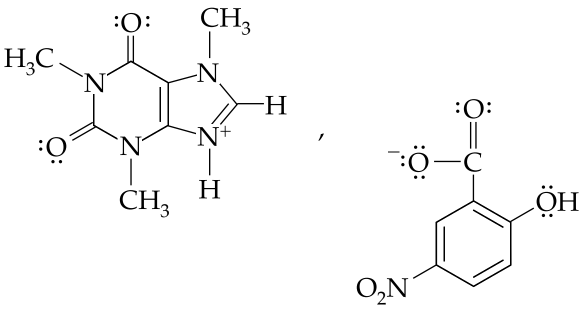

Caffeine 5-nitrosalicylate

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 0.5 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
Caffeine	194.20	11 mg	0.06	238	
5-Nitrosalicylic acid	183.12	10 mg	0.06	229–30	
Petroleum ether (60–80 °C)		0.5 mL			60–80
Ethyl acetate	88.12	0.7 mL			77
Caffeine 5-nitrosalicylate	377.32			180	

Reagents and Equipment. Weigh and add to a 3.0-mL conical vial, containing a magnetic spin vane and equipped with an air condenser, 11 mg (0.06 mmol) of caffeine, 10.0 mg (0.06 mmol) of 5-nitrosalicylic acid, and 0.7 mL of ethyl acetate.

Reaction Conditions. *Gently* warm the mixture on a hot plate, with stirring, to dissolve the solids. Add 0.5 mL of petroleum ether (bp 60–80 °C) to the warm ethyl acetate solution, mix, and warm for several seconds. Remove the spin vane using forceps.

Isolation of Product. Cool the mixture to room temperature and then place it in an ice bath for 10–15 min. Collect the crystals under reduced pressure using a Hirsch funnel, and then wash the filter cake with 0.5 mL of cold ethyl acetate. Dry the product on a porous clay plate or filter paper.

Purification and Characterization. The product normally is sufficiently pure for direct characterization. Weigh the caffeine 5-nitrosalicylate and calculate the percent yield. Determine the melting point and compare your value with that reported above.

Obtain an IR spectrum and compare it with those shown in Figures 6.25 and 6.26. The infrared spectrum reveals some of the details of the derivative formation (see below).

Infrared Analysis. The infrared spectrum of 5-nitrosalicylic acid (Fig. 6.26 on page 238) is characteristic of aromatic carboxylic acids (see discussion in Experiment [7]). Note that (1) the substitution of the ring is revealed by the presence of the 1,2,4-combination band pattern with peaks at 1940, 1860, and 1815 cm^{-1}; (2) the strongest band in the spectrum below 1750 cm^{-1} is assigned to the symmetric stretch of the —NO_2 group found at 1339 cm^{-1}, and (3) the conjugated carboxyl C═O stretch is located at 1675 cm^{-1}.

In the spectrum of the complex (Fig. 6.25) we do not find evidence for ionized carboxylate. This group, if present, would give rise to two very strong *broad* bands at 1600–1550 and 1400–1330 cm^{-1}. What is observed is the carboxylate C═O stretch at 1665 cm^{-1} overlapped with a caffeine band. Evidence for very strong hydrogen bonding, however, is indicated by the series of very broad bands extending from 3550 to 2000 cm^{-1}. The complex association, therefore, very likely does not involve a complete proton transfer as is indicated in the above simplified chemical structures of the complex. The unambiguous formation of the 5-nitrosalicylic acid complex is best ascertained by the identification of the presence of the —NO_2 symmetric stretching vibration from a strong band located at 1345 cm^{-1}. Caffeine does not possess a band in this region.

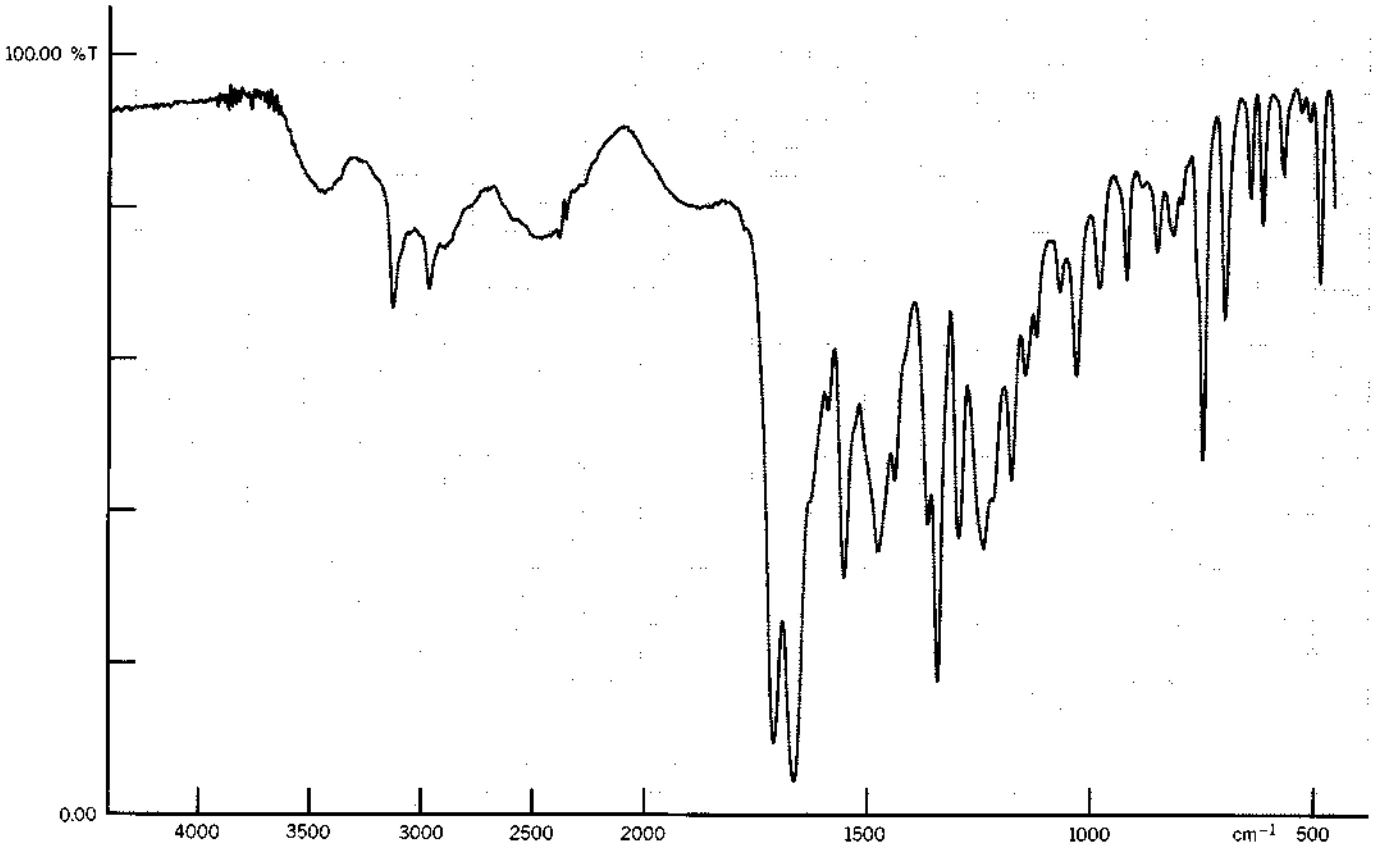

Sample	Caffeine 5-nitrosalicylate		
%T x ABS __ Background Scans	4	Scans	16
Acquisition & Calculation Time	42 sec	Resolution	4.0 cm-1
Sample Condition	solid	Cell Window	
Cell Path Length		Matrix Material	KBr

Figure 6.25 IR spectrum: caffeine 5-nitrosalicylate.

QUESTIONS

6-71. Compounds such as naphthalene and 1,4-dichlorobenzene find use as mothballs since they sublime at a slow rate at atmospheric pressure. Explain this behavior in terms of the structure of the molecules.

6-72. **(a)** How many peaks would you expect to find in the NMR spectrum of caffeine? **(b)** What characteristic absorption bands would you expect to find in the infrared spectrum of caffeine? **(c)** Are the NMR spectra (proton and carbon) of caffeine when compared to caffeine 5-nitrosalicylate as characteristic to the comparisons found using IR?

6-73. The vapor pressures of 1,2-diphenylethane, *p*-dichlorobenzene, and 1,3,5-trichlorobenzene are 0.06, 11.2, and 1.4 torr, respectively, at their melting point (52–54 °C). Which compound is likely to be sublimed most rapidly at a reduced pressure of 15 torr and a temperature of 40 °C?

6-74. To color spots on TLC plates for easier visualization after elution with solvent, the plates can be "developed" in a sealed chamber containing solid iodine. Explain how the solid–vapor equilibrium operates in this instance.

6-75. Caffeine is soluble in ethyl acetate. Do you think that the purity of your product could be checked by TLC using ethyl acetate as an elution solvent? Explain.

6-76. The infrared spectrum of 5-nitrosalicylic acid (Fig. 6.26) possesses the typical broad medium band found in *acid dimers* (908 cm^{-1}). In the caffeine 5-nitrosalicylate complex, however, this band is missing. Suggest a reason why the 908-wavenumber peak vanishes.

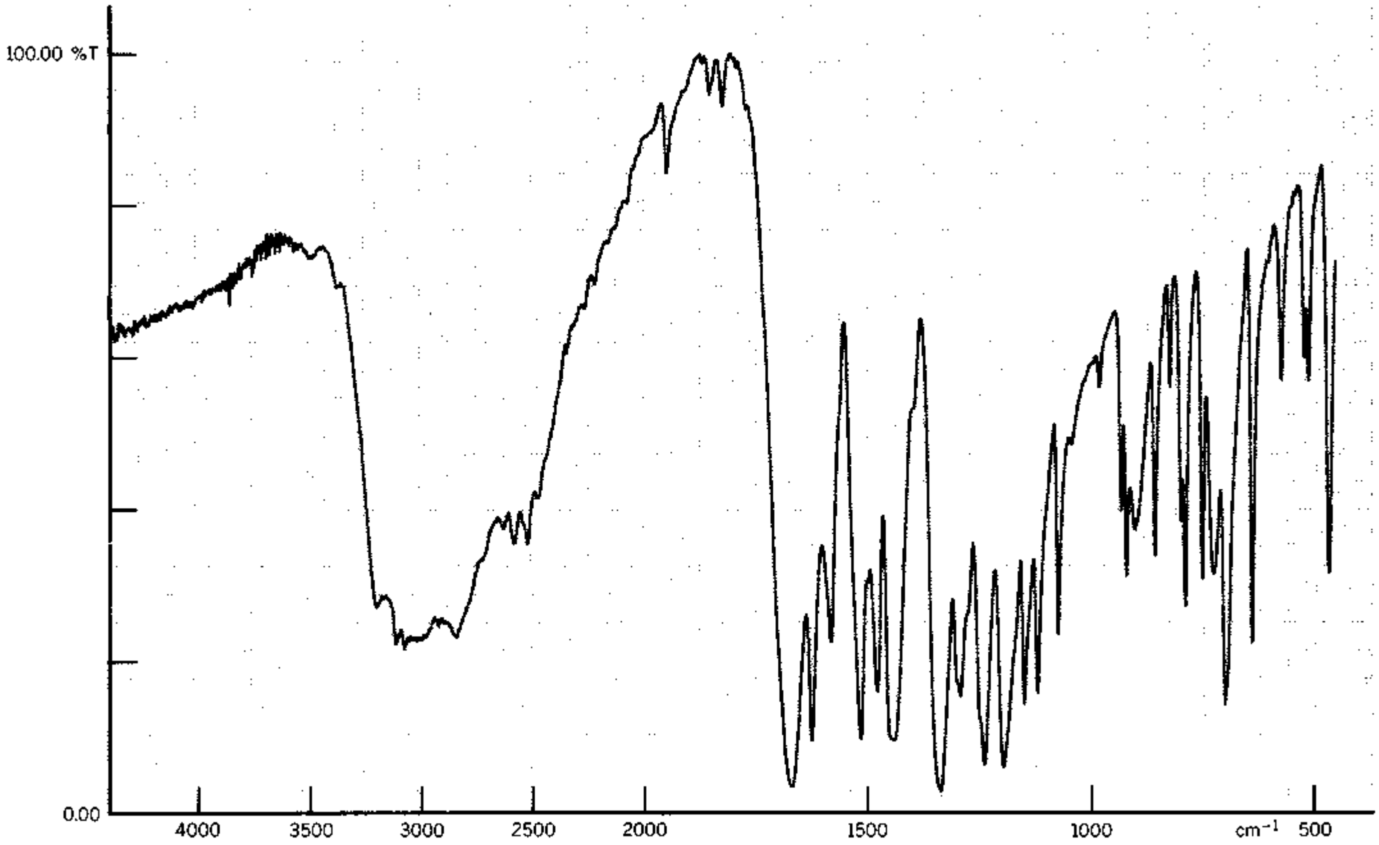

Sample 5-Nitrosalicylic acid

%T x ABS — Background Scans 4

Scans 16

Acquisition & Calculation Time 42 sec

Resolution 4.0 cm^{-1}

Sample Condition solid

Cell Window

Cell Path Length

Matrix Material KBr

Figure 6.26 IR spectrum: 5-nitrosalicylic acid.

BIBLIOGRAPHY

Several extraction procedures for isolating caffeine from tea, some designed for the introductory organic laboratory, have been reported.

Ault, A.; Kraig, R. *J. Chem. Educ.* **1969,** *46,* 767.
Mitchell, R. H.; Scott, W. A.; West, P. R. *J. Chem. Educ.* **1974,** *51, 69.*
Moyé, A. L. *J. Chem. Educ.* **1972,** *49,* 194.
Murray, S. D.; Hansen, P. J. *J. Chem. Educ.* **1995,** *72,* 851.
Onami, T.; Kanazawa, H. *J. Chem. Educ.* **1996,** *73,* 556.

For additional information on the Fisher-Johns melting point apparatus see

Morhrig, J. R.; Neckers, D. C. *Laboratory Experiments in Organic Chemistry,* 3rd ed.; Van Nostrand: New York, **1979,** p. 92.
Zubrick, J. W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, **2008,** p. 94, Fig. 12.5.

Experiment 11C

Isolation of a Natural Product by Steam Distillation: Cinnamaldehyde from Cinnamon

Common name: cinnamaldehyde

CA number: [14371-10-9]

CA name as indexed: 2-propenal, 3-phenyl-, (*E*)-

Purpose. In this exercise you will extract oil of cinnamon from a native plant source, such as *Cinnamomum zeylanicum*, and then purify the principal flavor and odor component of the oil, cinnamaldehyde. The experiment demonstrates the importance of steam distillation techniques (at the semimicro level) to the collection of essential oils.

Prior Reading

Technique 2: Simple Distillation at the Semimicroscale Level (pp. 61–64)
Technique 3: Steam Distillation
Technique 4: Solvent Extraction
    Liquid–Liquid Extraction (p. 72)
    Drying of the Wet Organic Layer (pp. 80–83)
Technique 6: Chromatography
    Concentration of Solutions with Nitrogen Gas (p. 102)

ESSENTIAL OILS

Let us begin by defining what we mean by the term "metabolite." The metabolism of an organism is composed of the biochemical reactions and pathways in that living system. The products (most of them organic molecules) derived from this array of molecular transformations are the *metabolites*. This vast collection of substances that are generally referred to as *natural products* are, in fact, the metabolites of the natural living world.

Natural products are divided into two large families of compounds. Those metabolites that are *common* to the large majority of all organisms are known as the *primary metabolites*. In general, they have well-defined roles in the biochemistry of the system. For example, the amino acids are the building blocks for protein synthesis in all organisms. The second great category of natural products is known as the *secondary metabolites*. Individual secondary metabolites are far less widely distributed in nature and may be unique to single species (or even limited to a variety of a particular species). While the biochemical role of some of these compounds was established early and easily, the majority of these materials were believed to be of little importance to the functioning of the living system, and their presence was unexplained until very recently. With the development of chemical communication theory over the last few decades, however, the vitally important roles of many of the secondary metabolites in the life cycles of their particular host organisms have been revealed.

You may have had the opportunity to become acquainted with secondary metabolites in Experiments [11A] and [11B]. In Experiment [11A] an acetogenin (this term refers to the biochemical origin of this material from eight acetic acid residues), usnic acid, was isolated from a lichen where it can occur in dramatically high concentrations. Only recently has the role of usnic acid as a defense mechanism come to be fully appreciated.

In Experiment [11B], the alkaloid caffeine was obtained from tea. This compound is a very unusual example of a purine ring system in a secondary metabolite. The ecological significance of the presence of caffeine in both tea and coffee seeds has been established and it has been shown that caffeine acts against both predators and competitors.

We will now examine, in Experiment [11C], a third class of secondary metabolites: the *essential oils*. The majority of these materials are high-boiling liquids that can be extracted from plant material via steam distillation techniques. The value of codistilling high-boiling substances was learned early in the days of alchemy. Because these oils often gave pleasant odors and flavors, they were considered to be the "essence" of the original plant material. Eventually, they became known as essential oils.

These materials were used as flavorings, perfumes, and medicines, and as both insect repellents and attractants. By the early 1800s, as it became possible to establish the carbon/hydrogen ratio in organic substances, many of the oils possessing pine-type odors (the oil of turpentine) were shown to have identical C/H ratios. These materials ultimately became known as terpenes. The terpenes all have their origin in mevalonic acid, from which they utilize, as their building block, a branched five-carbon unit as in isoprene. The terpenes of the essential oils occur as C_{10} (monoterpenes), C_{15} (sesquiterpenes), C_{20} (diterpenes), C_{30} (triterpenes), and C_{40} (tetraterpenes) compounds. Today, this collection of substances represents a large fraction of the known secondary metabolites, including the steroids. Terpenes may be polymerized, extending to much higher molecular weights, with between 1000 and 5000 repeating isoprene units (MW = 60,000–350,000) to yield polymers known as the natural rubbers.

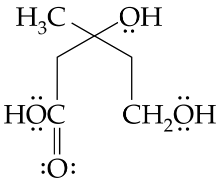
Mevalonic acid

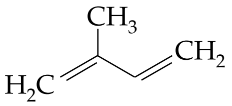
Isoprene

As we have seen, many of the compounds found in the essential oils possess pleasing properties of taste and odor, and we now know that many of these systems contain either ketone or aldehyde functional groups. Our senses of taste and smell, however, possess a wide range of responses to the shape and dimensions of the carbon skeleton supporting the main functionality that triggers the odor signal. Thus, our sense of odor may involve simultaneous multiple stimulations by many different molecular species or, as in a number of cases, the principal response may be to a single component. Since the shape of the odor- or taste-inducing molecule plays a significant role in the effect, it is not surprising that chirality can have a dramatic impact on our perception of a particular odor.

One of the classic examples of this type of response is the case of the cyclic ketone carvone, which contains a single stereocenter (*). The *S* enantiomer is the principal odor and flavor component in caraway seed, whereas the *R* enantiomer gives rise to the odor and flavor of spearmint!

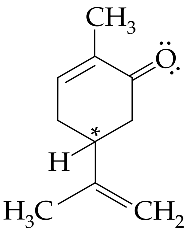
Carvone

What we find pleasant may be offensive to others. A constituent of the oil of lemon grass is citronellal, a C_{10} unsaturated aldehyde. While we find this compound to have a pleasant fragrance, it is a potent alarm signal in ants that is shunned by many other insects. Thus, this terpene aldehyde has been used effectively by both ants and humans as an insect repellent.

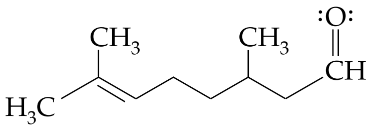

Citronellal

In this experiment we will isolate the principal component of the oil of cinnamon, another naturally occurring aldehyde, cinnamaldehyde. The oil is first extracted from the dried parts of the *Cinnamomum* plant by steam distillation. Although this aromatic aldehyde is a component of an essential oil, it is not formed from mevalonic acid and is not a terpene. Cinnamaldehyde is also not an acetogenin nor is it related to usnic acid. The origin of this fragrant material is shikimic acid, which is part of the plant's *primary metabolism*.

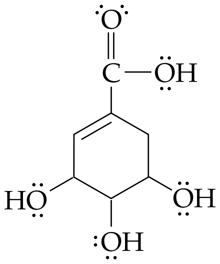

Shikimic acid

Cinnamaldehyde's formation from shikimic acid utilizes one of only two biogenetic routes in nature that lead to the aromatized benzene ring (the other pathway is found in the acetogenins and produces secondary metabolites like usnic acid; see Experiment [11A]). The shikimic acid route contributes to a class of metabolites called the phenylpropanes (Ph—C_3), of which cinnamaldehyde is one of a limited number of simple end products. Another close relative is, for example, eugenol from oil of cloves.

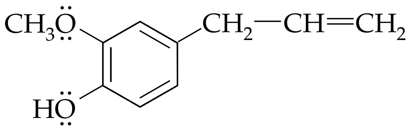

Eugenol

The principal metabolic fate of the phenylpropanes is the formation of lignin polymers that are the fundamental basis of the structural tissue in all plants. Thus, cinnamaldehyde itself is a relatively rare example of a primary metabolite which has been expressed in an *essential oil*.

DISCUSSION

Steam Distillation. The process of steam distillation can be a valuable technique in the laboratory for the separation of thermally labile, high-boiling substances from relatively nonvolatile materials. For steam distillation to be successful, however, the material to be isolated must be nearly immiscible with water. (For details of the theory of steam distillation refer to Technique 3 discussions).

www

Steam distillation is in essence the codistillation (or simultaneous distillation) of two immiscible liquid phases. By definition, one of these liquid phases is water and the other phase is usually a mixture of organic substances that have a low solubility in water. Though steam distillation is widely used as a separation technique for natural products, and occasionally for the isolation and/or purification of synthetic products that decompose at their normal boiling points, it has several limitations. For example, it is not the method of choice when a dry product is required or if the compound to be isolated reacts with water. Obviously, steam distillation is not feasible if the compound to be isolated decomposes upon contact with steam at 100 °C.

Although cinnamaldehyde decomposes at its normal boiling point, it may be extracted from the plant without degradation by boiling water. Steam distillation, therefore, is the method of choice for the isolation of this pleasant smelling and tasting aldehyde.

COMPONENT

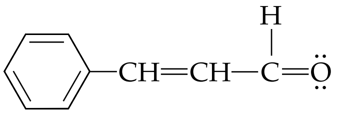

Cinnamaldehyde

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Components

Compound	MW	Amount
Cinnamon		1 g
Water	18.02	4.0 mL

Components and Equipment. Place 1 g of chopped stick cinnamon (or powder) and 4 mL of water in a 10-mL round-bottom flask containing a boiling stone. Attach the flask to a Hickman still head equipped with an air condenser.

Distillation Conditions. Place the apparatus on a **sand bath maintained at 150–160 °C.** Use an aluminum foil shield or baffle (not shown in drawing) to cover the sand bath. This procedure will prevent the collection collar of the Hickman still from overheating.

NOTE. The mixture tends to foam during the distillation, so care must be taken to prevent contamination of the distillate by cinnamon particles, especially if powdered cinnamon is used. If stick cinnamon is used, first evacuating the flask containing the water and cinnamon for a few minutes and then returning the system to atmospheric pressure will allow water to fill the pores of the bark and greatly aids in reducing the foaming problem (*See* Taber, D. F.; Weiss, A. J. *J. Chem. Educ.* **1998,** *75,* 633).

Position a thermometer in the throat of the Hickman still to record the distillation temperature, which should be very close to 100 °C. In early Hickman stills (without sidearm access) this thermometer makes pipetting of the condensate difficult. It is suggested that in the latter setup, the thermometer be removed after the initial distillation temperature is established and recorded.(⬇)

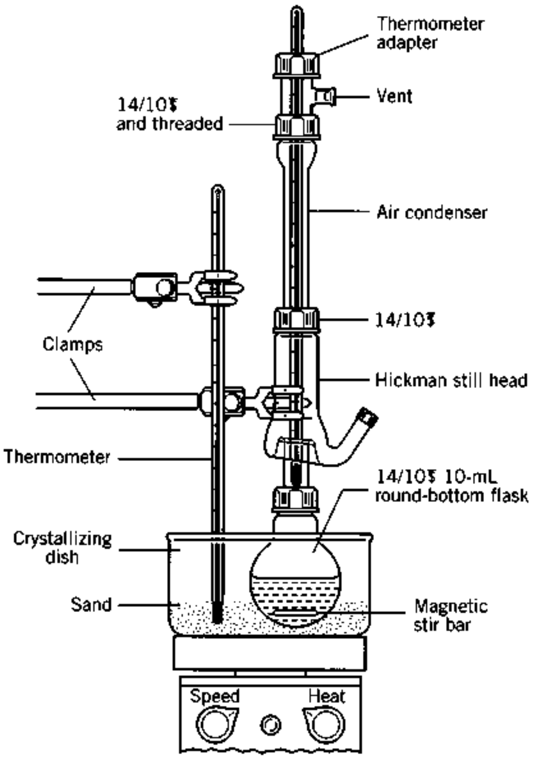

Cinnamon stick + H_2O, 4 mL

Isolation of Cinnamaldehyde. Remove the milky cinnamaldehyde–water (two-phase) distillate that collects in the collar of the still using a 9-in. Pasteur pipet (or 6-in. in the case of sidearm stills). Transfer this material to a 12- or 15-mL centrifuge tube. Continue the distillation for approximately 1 h or until about 5–6 mL of distillate is collected in the centrifuge tube.

NOTE. Add additional water during the course of the distillation to maintain the original volume in the flask. Add this water using a 9-in. Pasteur pipet inserted down the neck of the Hickman still, after first removing the thermometer if it is still in place (the thermometer need not be replaced in the still following the addition).

Extract the combined distillate fractions, which you collected in the centrifuge tube, with three successive 2-mL portions of methylene chloride (3 × 2 mL). Use the first portion of methylene chloride to rinse the collection collar of the Hickman still. After each extraction, transfer the lower methylene chloride layer (Pasteur filter pipet) to a 25-mL Erlenmeyer flask.

Dry the combined extracts over anhydrous sodium sulfate.

Purification and Characterization. Transfer the dried methylene chloride solution **in at least two portions,** using a Pasteur filter pipet, to a **tared** 5-mL conical vial. Evaporate the solvent in a warm sand bath using a slow stream of nitrogen gas. After all the solution has been transferred and the solvent evaporated, rinse the sodium sulfate with an additional two 0.5-mL portions of methylene chloride. Transfer the rinses to the same vial and concentrate as before.

Weigh the flask and calculate the percentage of crude cinnamaldehyde extracted from the original sample of cinnamon.

Record the infrared spectrum and compare it to that reported in the literature (*Aldrich Library of Infrared Spectra* and/or SciFinder Scholar).

To further characterize the aldehyde, prepare its semicarbazone derivative (see Chapter 9, Preparation of Derivatives: Aldehydes and Ketones). The semicarbazone derivative has a melting point of 215 °C.

QUESTIONS

6-77. List several advantages and disadvantages of steam distillation as a method of purification.

6-78. Explain why the distillate collected from the steam distillation of cinnamon is cloudy.

6-79. Calculate the weight of water required to steam distill 500 mg of bromobenzene at 95 °C. The vapor pressure of water at this temperature is 640 torr; that of bromobenzene is 120 torr.

6-80. Steam distillation may be used to separate a mixture of *p*-nitrophenol and *o*-nitrophenol. The ortho isomer distills at 93 °C; the para isomer does not. Explain.

6-81. A mixture of nitrobenzene and water steam distills at 99 °C. The vapor pressure of water at this temperature is 733.2 torr. What weight of water is required to steam distill 300 mg of nitrobenzene?

BIBLIOGRAPHY

Several steam distillation procedures have been reported for isolation of essential oils from native plant sources:

Bell, C. E., Jr.; Taber, D. I.; Clark, K. A. *Organic Chemistry Laboratory: Standard and Microscale Experiments,* 3rd ed.; Harcourt College Pubs.: Orlando, FL, 2001.

Taber, D. F.; Weiss, A. J. *J. Chem. Educ.* **1998,** *75,* 633.

Vogel, A. I. *Vogel's Textbook of Practical Organic Chemistry,* 5th ed.; Furnis, B. S., et al., Eds.; Wiley: New York, 1989.

Zubrick, J. W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, 2008.

EXPERIMENT 12

Reductive Catalytic Hydrogenation of an Alkene: Octane

Common name: octane
CA number: [1111-65-9]
CA name as indexed: octane

Purpose. This experiment shows you how to reduce the carbon–carbon double bond of an alkene by addition of molecular hydrogen (H_2). You will gain an understanding of the important role that metal catalysts play in the stereospecific reductions of alkenes (and alkynes), to form the corresponding alkanes, by the activation of molecular hydrogen. You can observe the powerful influence on column chromatography of heavy metal ions, such as silver (Ag^+), which lead to effective separation of mixtures of alkenes from alkanes. Finally, you will appreciate the enormous importance and breadth of application of these reduction reactions in both industrial synthesis and basic biochemistry (for important examples see Experiment [8])

REACTION

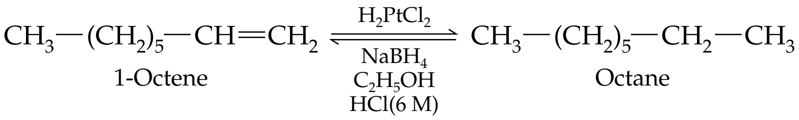

$$CH_3-(CH_2)_5-CH{=}CH_2 \xrightleftharpoons[\substack{NaBH_4 \\ C_2H_5OH \\ HCl(6\ M)}]{H_2PtCl_2} CH_3-(CH_2)_5-CH_2-CH_3$$

1-Octene Octane

Prior Reading

Technique 4: Solvent Extraction
Liquid–Liquid Extraction (p. 72)
Technique 6: Chromatography
Column Chromatography (pp. 92–95)
Concentration of Solutions (pp. 101–104)

DISCUSSION

The addition of hydrogen to an alkene (or to put it another way, the saturation of the double bond of an alkene with hydrogen) to produce an alkane is an important reaction in organic chemistry. Alkanes are also called saturated hydrocarbons, because the carbon skeletons of alkanes contain the greatest possible number of hydrogen atoms permitted by tetravalent carbon atoms; alkenes are thus unsaturated hydrocarbons. Hydrogenation reactions have widespread use in industry. For example, we all consume vegetable fats hardened by partial hydrogenation (see Experiment [8]) of the polyunsaturated oils that contained several carbon–carbon double bonds per molecule as isolated from their original plant sources. Partially hydrogenated fats represented a consumer market of over 2.9 billion pounds in 1992. Since that time major restaurants, fast food chains, and food producers have greatly lowered the amount of trans fats in their produts. Furthermore, batch processes involving hydrogenation reactions are nondiscriminating when considering the geometric isomers cis and trans.

The hydrogenation reaction is exothermic; the energy released is approximately 125 kJ/mol for most alkenes, but on the kinetic side of the ledger, this reductive pathway requires a significant activation energy to reach the transition state.Thus, alkenes can be heated in the presence of hydrogen gas at high temperatures for long periods without any measurable evidence of alkane formation. However, when the reducing reagent (H_2) and the substrate (the alkene) are in intimate contact with each other in the presence of a finely divided metal catalyst, rapid reaction does occur at room temperature. Under these conditions, successful reduction is generally observed at pressures of 1–4 atm. For this reason, this reaction is often referred to as low-pressure catalytic hydrogenation. These reactions are called heterogeneous reactions, since they occur at the boundary between two phases—in this case a solid and a liquid.

The main barrier to the forward progress of the reaction is the very strong H—H bond that must be broken. Molecular hydrogen, however, is adsorbed by a number of metals in substantial quantities; indeed, in some instances the amount of hydrogen contained in the metal lattice can be greater than that in an equivalent volume of pure liquid hydrogen! In this adsorption process the H—H bond is broken or severely weakened. (This adsorption process, which necessarily involves a large exchange of chemical energy between the metal lattice and the adsorbed hydrogen, may in some, as yet unexplained way, be related to the "cold fusion" problem in which palladium saturated with a heavy isotope of hydrogen [deuterium] allegedly exhibits apparent excess thermal energies on electrolysis.)

The π system of the alkene is also susceptible to adsorption onto the metal surface and when this occurs the barrier to reaction between the alkene and the activated hydrogen drops dramatically.

Catalytic hydrogenations are also a representative example of a class of organic reactions known as *addition reactions*, which are reactions in which two new substituents are *added* to a molecule (the alkene substrate in this case) across a π system. Usually addition is 1,2, but in extended π systems such as 1,3 dienes, the addition may occur 1,4. In catalytic hydrogenations, formally, one hydrogen atom of a hydrogen molecule adds to each carbon of the alkene linkage, C═C. It is not at all clear that both hydrogen atoms must come from the same original hydrogen molecule even though they are added stereospecifically in syn (cis) fashion while both systems are coordinated with the metal surface. A representation of the stereochemistry of this addition is given below.

The metals most often used as catalysts in low-pressure (1–4 atm) hydrogenations in the laboratory are nickel, platinum, rhodium, ruthenium, and palladium. In industry, high-pressure, large-scale processes are more likely to be found. For example, Germany had little or no access to naturally formed petroleum deposits during World War II, but did possess large coal mines. The Germans mixed powdered coal with heavy tar (from previous production runs) and 5% iron oxide and heated this in the presence of H_2 at a pressure of 3000 lb/in^2, at about 500 °C for 2 h to yield synthetic crude oil. Thirteen German plants operating in 1940 produced 24 million barrels that year, with an average of 1.5–2 tons of coal ultimately converted to about 1 ton of gasoline.

In the present experiment the metal catalyst, platinum, is generated in situ by the reaction of chloroplatinic acid with sodium borohydride. The reduced platinum metal is formed in a colloidal suspension, which provides an enormous surface area, and therefore excellent conditions, for heterogeneous catalysis. The molecular hydrogen necessary for the reduction can also be conveniently generated in situ by the reaction of sodium borohydride with hydrochloric acid:

$$4\,\mathrm{NaBH_4} + 2\,\mathrm{HCl} + 7\,\mathrm{H_2O} \longrightarrow \mathrm{Na_2B_4O_7} + 2\,\mathrm{NaCl} + 16\,\mathrm{H_2}$$

This reduction technique does not require equipment capable of safely withstanding high pressures. The use of chloroplatinic acid, therefore, is particularly attractive for saturating easily reducible groups, such as unhindered alkenes or alkynes, in the laboratory. The potential limitation to the use of this reagent is that other reducible functional groups, such as aldehydes and ketones, normally inert to catalytic hydrogenations of alkenes and alkynes, may be reduced by the sodium borohydride. Thus, with chloroplatinic acid and sodium borohydride, we accept, as a compromise, a more limited set of potential reactants (substrates) for the convenience inherent in the reagent.

The platinum catalyst generated in the reaction medium adsorbs both the internally generated molecular hydrogen and the target alkene on its surface. The addition of the hydrogen molecule (H_2) (evidence strongly suggests that it is actually atomic hydrogen that attacks the alkene π system) to the alkene system while they are both adsorbed on the metal surface results in the reduction of the substrate and the formation of an alkane. The addition is, as mentioned earlier, syn (or cis), since both hydrogen atoms add to the same face of the alkene plane. The mechanistic sequence is outlined below:

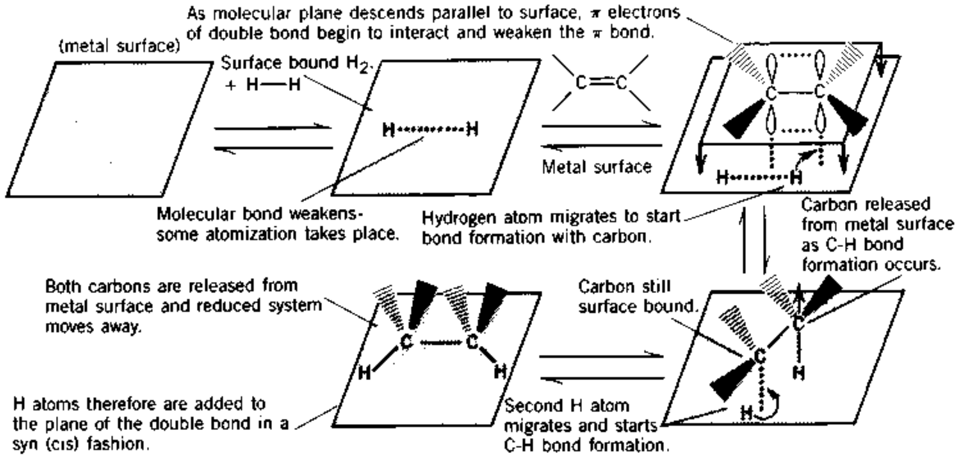
As molecular plane descends parallel to surface, π electrons of double bond begin to interact and weaken the π bond.
(metal surface)
Surface bound H_2.
+ H—H
H••••••••H
C=C
Metal surface
H••••••H
Molecular bond weakens-some atomization takes place.
Hydrogen atom migrates to start bond formation with carbon.
Carbon released from metal surface as C-H bond formation occurs.
Both carbons are released from metal surface and reduced system moves away.
Carbon still surface bound.
C—C
H
H
H
H
H atoms therefore are added to the plane of the double bond in a syn (cis) fashion.
Second H atom migrates and starts C-H bond formation.

This experiment also provides an opportunity for you to study a powerful aspect of column chromatography in which heavy metal ions have a particularly important role in the purification of the product. Unreacted alkene has the potential to be a problem contaminant during the isolation and purification of the relatively low-boiling saturated reaction product.The successful removal of the remaining 1-octene from the desired *n*-octane in the product mixture is achieved by using column chromatography with silver nitrate/silica gel as the stationary phase. Complex formation between the silver ion (on the silica gel surface) and the π system of the unreacted alkene acts to retard the rate of elution of the alkene relative to that of the alkane.

The ability of alkenes to form coordination complexes with certain metal ions having nearly filled d orbitals was established some time ago. In the case of the silver ion complex with alkenes, the orbital nature of the bonding is believed to involve a σ bond formed by overlap of the filled π orbital of the alkene with the free s orbital of the silver ion plus a π bond formed by overlap of the vacant antibonding π* orbitals of the alkene together with the filled d orbitals of the metal ion.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
1-Octene	112.22	120 μL	0.76	121.3	0.72	1.4087
Ethanol (absolute)	46.07	1.0 mL		78.5		
Chloroplatinic acid (0.2 M)	517.92	50 μL				
Sodium borohydride (1 M)	37.83	125 μL				
Dilute HCl (6 M)		100 μL				

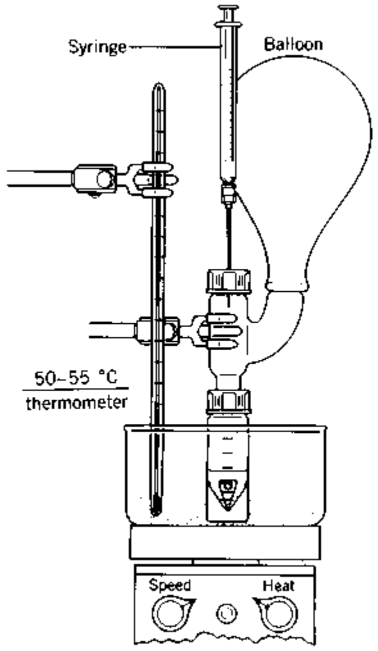

Step I:
0.2 *M* H_2PtCl_6 solution,
50 µL +
CH_3CH_2OH, 1.0 mL
+ $NaBH_4$ solution, 125 µL
Step II:
6 *M* HCl, 100 µL
+ 1-octene, 120 µL
+ $NaBH_4$ solution, 1.0 mL

Reagents and Equipment. Equip a 5.0-mL conical vial containing a magnetic spin vane with a Claisen head fitted with a rubber balloon and Teflon-lined rubber septumcap (good *GC septa* work best—this is an important point, because several injections through the septum are required and the seal must remain gas-tight). Mount the assembly in a sand bath on a magnetic stirring hot plate (⬅).

NOTE. 1. No residual acetone (perhaps from cleaning the equipment) can be present since it reacts with the $NaBH_4$. 2. The balloon must make a gas-tight seal to the Claisen head, so be sure to secure it with copper wire or a rubber band.

Remove the 5.0-mL vial from the Claisen head and add the following reagents. *Recap the vial after each addition.*

a. Add 50 µL of a 0.2 M solution of chloroplatinic acid (H_2PtCl_6) (automatic delivery pipet).
b. Add 1.0 mL of absolute ethanol (calibrated Pasteur pipet).
c. Add 125 µL of the sodium borohydride reagent (automatic delivery pipet).

NOTE. Reattach the vial ***immediately*** *to the Claisen head after the $NaBH_4$ solution is added.*

Stir the mixture vigorously. The solution should turn black immediately as the finely divided platinum catalyst is formed.

INSTRUCTOR PREPARATIONS. 1. The 0.2 M H_2PtCl_6 solution is prepared by adding 41 mg (0.1 mmol) of the acid to 0.5 mL of deionized water. 2. The sodium borohydride reagent is prepared by adding 0.38 g (0.01 mol) of $NaBH_4$ to a solution of 0.5 mL of 2.0 M aqueous NaOH in 9.5 mL of absolute ethanol.

After 1 min, use a syringe to add 100 µL of 6 M HCl solution through the septum cap. In a like manner using a fresh syringe, add *immediately* to the acid solution, a solution of 120 µL (86 mg, 0.76 mmol) of 1-octene dissolved in 250 µL of absolute ethanol. (This solution is conveniently prepared in a 1-mL conical vial; the reagents are best dispensed using automatic delivery pipets.)

Now add dropwise (clean syringe) 1.0 mL of the $NaBH_4$ reagent solution over a 2-min interval.

NOTE. At this point the balloon should inflate and remain inflated for at least 30 min. If it does not, the procedure must be repeated.

Reaction Conditions. Stir the reaction mixture vigorously at a sand bath temperature of 50 °C for 45 min.

Isolation of Product. Cool the reaction to ambient temperature and dropwise add 1 mL of water. Extract the resulting mixture in the reaction vial with three 1.0-mL portions of pentane. Transfer each pentane extract to a stoppered 25-mL Erlenmeyer flask containing 0.5 g of anhydrous sodium sulfate.

The reaction mixture is extracted as follows. Upon addition of each portion of pentane, cap the vial, shake, vent carefully, and then allow the layers to separate. A Vortex mixer may be used if available. The transfers must be made using a Pasteur filter pipet because the pentane solvent is particularly volatile.

Using a Pasteur filter pipet, transfer the dried solution to a second 25-mL Erlenmeyer flask. Rinse the drying agent with an additional 1.0 mL of pentane (calibrated Pasteur pipet) and add the rinse to this second flask. Add a boiling stone and concentrate the solution to a volume of about 1.0–1.5 mL by warming it gently in a sand bath in the **hood.**

HOOD

Purification and Characterization. The saturated product, *n*-octane, is purified by column chromatography. In a Pasteur filter pipet place about 50 mg of sand, 500 mg of 10% silver nitrate on activated silica gel (200 mesh), and then 50 mg of anhydrous sodium sulfate (➡).

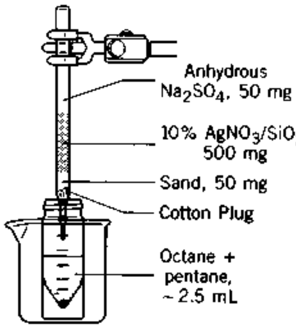
Anhydrous Na_2SO_4, 50 mg
10% $AgNO_3/SiO_2$, 500 mg
Sand, 50 mg
Cotton Plug
Octane + pentane, ~2.5 mL

CAUTION: Silver nitrate stains the skin. Protective gloves should be worn during this operation. The ~10% silver nitrate/silica gel used in this separation is commercially available.

Wet the column with 0.5 mL of pentane (calibrated Pasteur pipet) and then transfer the concentrated crude product, as obtained above, to the column by Pasteur filter pipet. Elute the octane from the column using 1.5 mL of pentane and collect the eluate in a tared 5.0-mL conical vial containing a boiling stone.

Fit the vial with an air condenser and then place the assembly in the **hood** in a sand bath maintained at a temperature of 90–100 °C to evaporate the pentane solvent.

HOOD

OPTIONAL. The evaporation is continued until a constant weight of product is obtained. This procedure is the best approach, but has to be done very carefully or a considerable amount of product can be lost.

Record the weight of product and calculate the percent yield. Determine the boiling point and refractive index (optional) of your material and compare your results with those reported in the literature for octane. Obtain an IR spectrum. Compare your results with those reported in the literature (*Aldrich Library of IR Spectra* and/or SciFinder Scholar). Also, compare your IR spectrum to that of the 1-octene starting material. Can you establish from the above data if your sample is contaminated by traces of the pentane extraction solvent? If not, how would you go about determining the presence of this potential impurity?

www

QUESTIONS

6-82. Squalene, first isolated from shark oil and a biological precursor of cholesterol, is a long-chain aliphatic alkene ($C_{30}H_{50}$). The compound undergoes catalytic hydrogenation to yield an alkane of molecular formula $C_{30}H_{62}$. How many double bonds does a molecule of squalene have?

6-83. A chiral carboxylic acid A ($C_5H_6O_2$) reacts with 1 mol of hydrogen gas on catalytic hydrogenation. The product is an achiral carboxylic acid B ($C_5H_8O_2$). What are the structures of compounds A and B?

6-84. Two hydrocarbons, A and B, each contain six carbon atoms and one $C{=}C$. Compound A can exist as both *E* and *Z* isomers but compound B cannot. However, both A and B on catalytic hydrogenation give only 3-methylpentane. Draw the structures and give a suitable name for compounds A and B.

6-85. What chemical test would you use to distinguish between the 1-octene starting material and the octane product?

6-86. Give the structure and names of five alkenes having the molecular formula C_6H_{12} that produce hexane on catalytic hydrogenation.

BIBLIOGRAPHY

Selected references on catalytic hydrogenation:

Birch, A. J.; Williamson, D. H. *Org. React.* **1976,** *24,* 1.

Carruthers, W. *Some Modern Methods of Organic Synthesis,* 3rd ed.; Cambridge Univ. Press: New York, 1986; p. 411.

De, S; Gambhir, G.; Krishnamurty, H. G. *J. Chem. Educ.* **1994,** *71,* 922.

Kabalka, G. W.; Wadgaonkar, P. P.; Chatla, N. *J. Chem. Educ.* **1990,** *67,* 975.

Parker, D. in *The Chemistry of the Metal-Carbon Bond,* Hartley, F. R., Ed.; Wiley: New York, 1987, Vol 4, Chapter 11, p. 979.

Pelter, A.; Smith, K.; Brown, H. C. *Borane Reagents (Best Synthetic Methods);* Academic Press: Orlando, FL, 1988.

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed.; Wiley: New York, 2007, Chap. 15, p. 1065.

Selected examples of catalytic hydrogenation of alkenes in *Organic Syntheses:*

Adams, R.; Kern, J. W.; Shriner, R. L. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 101.

Bruce, W. F.; Ralls, J. O. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 191.

Cope, A. C.; Herrick, E. C. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 304.

Herbst, R. M.; Shemin, D. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 491.

Ireland, R. E.; Bey, P. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 459.

McMurry, E. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 781.

Meyers, A. I.; Beverung, W. N.; Gault, R. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 371.

EXPERIMENT 13

Hydroboration–Oxidation of an Alkene: Octanol

Common name: octanol
CA number: [111-87-5]
CA name as indexed: 1-octanol

Purpose. The oxidation of an alkene to an alcohol is investigated via the in situ formation of the corresponding trialkylborane, followed by the oxidation of the carbon–boron bond with hydrogen peroxide. The conditions required for hydroboration (a reduction) of unsaturated hydrocarbons are explored. Alkylboranes are particularly useful synthetic intermediates for the preparation of alcohols. The example used in this experiment is the conversion of 1-octene to 1-octanol in which an anti-Markovnikov addition to the double bond is required to yield the intermediate, trioctylborane. Since it is this alkyl borane that subsequently undergoes oxidation to the alcohol, hydroboration offers a synthetic pathway for introducing substituents at centers of unsaturation that are not normally available to the anti-Markovnikov addition reactions that are based on radical intermediates.

Prior Reading

Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
Technique 6: Chromatography
  Column Chromatography (pp. 92–95)
  Concentration of Solutions (pp. 101–104)

REACTION

$$\underset{\text{1-Octene}}{3\,CH_3{-}(CH_2)_5{-}CH{=}CH_2} \xrightarrow{THF\cdot BH_3} \underset{\text{Trioctylborane}}{[CH_3(CH_2)_7]_3B} \xrightarrow[OH^-]{H_2O_2} \underset{\text{1-Octanol}}{3\,CH_3{-}(CH_2)_7{-}OH}$$

DISCUSSION

The course of this reaction depends (1) on the *stereospecific* reductive addition of diborane (B_2H_6, introduced as the borane • tetrahydrofuran complex ($BH_3 \cdot THF$)) to an alkene to form an intermediate trialkylborane and (2) on oxidation of the borane with alkaline hydrogen peroxide to yield the corresponding alcohol.

The first step in the reaction sequence is generally called a *hydroboration*.[12] The addition of diborane is a rapid, quantitative, and general reaction for all alkenes (as well as alkynes) when carried out in a solvent that can act as a Lewis base. The ether solvation of the diborane, for example, is the key to the success of this reaction. In the absence of a Lewis base, borane (BH_3) exists as a dimer (B_2H_6), which is much less reactive than the monomer (BH_3). Borane, however, does exist in coordination with ether type solvents. It is the monomer (BH_3) that functions as the active reagent in the reductive addition.

As depicted in the following mechanism, the boron hydride rapidly adds successively to three molecules of the alkene to form a trialkylborane.

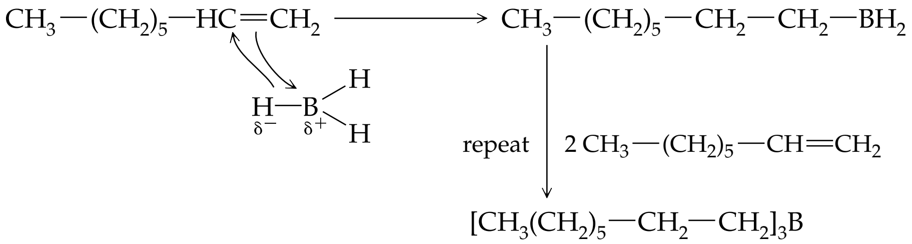

Note that the transition state of this addition reaction is generally considered to be a *four-center* one, and that the 1-octene substrate is oriented such that the boron becomes bonded to the least-substituted carbon atom of the double bond. Thus, the reaction can be classified as *regioselective*, and it will be sensitive to substitution on the carbon–carbon double bond.

In the developing transition state, the alkene π electrons (the least tightly held, and most nucleophilic) flow to the electron-deficient boron atom (the vacant p orbital is the electrophile). The formation of the *transition state* is controlled in large part by the polarization of the alkene π system during the early stages of formation of the transition state. At this point a partial positive charge begins to form on the more highly substituted carbon (the more stable carbocation), and a partial negative charge on the least substituted carbon. The orientation of the polarization, therefore, is to a large extent controlled by the electron-releasing effects of the alkyl substituents on the alkene, which enables the more highly substituted of the sp^2 carbon atoms to better accommodate the positive charge. As the reaction proceeds, the boron acquires a partial negative charge in response to the incoming electron

---

[12] For references relating to the use of diborane as a hydroboration reagent, see Experiment [1_{adv}].

www

density. The ease of hydride (:H^-) transfer from the boron to the more highly substituted carbon atom of the alkene, therefore, increases. Thus, hydroboration involves simultaneous hydride release and boron–carbon bond formation, and is a concerted reaction. The reaction can be conveniently considered as passing through a four-centered transition state, wherein the atoms involved undergo simultaneous changes in bonding (i.e., electron redistribution [see below]).

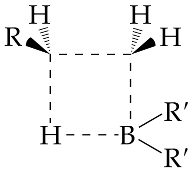

Hydroboration, as we have seen, can be classified as a **concerted addition reaction** in which no intermediate is formed. The mechanism is characteristic of a group of reactions called **pericyclic** (from the Greek, meaning *around the circle*) **reactions,** which involve a cyclic shift of electrons in and around the *transition state*. The mechanism proposed is further supported by the fact that rearrangements are not normally observed in hydroboration reactions, which implies that there are no carbocationic intermediates.

When alkenes with varying degrees of substitution undergo hydroboration, the boron ends up on the least substituted sp^2 carbon atom. While it might appear from the products that the regioselectivity is controlled by steric factors, this assumption is probably too simplistic. Steric and electronic factors both favor, and are both likely responsible for, the observed regioselectivity in hydroboration reactions.

Accumulated evidence demonstrates that the reaction occurs by **syn** addition, which is a consequence of the four-centered transition state. Therefore, the new C—B and C—H bonds are necessarily formed on the same face of the C═C bond, as shown in the following example:

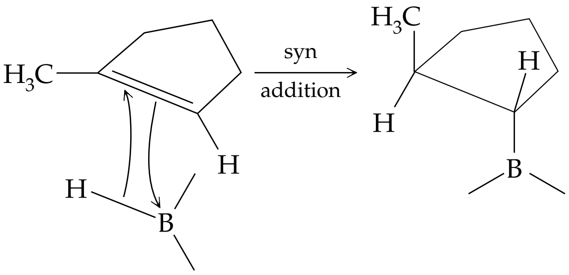

Organoboranes are important in organic synthesis as reactive intermediates. Reactions have been developed by which the boron atom may be replaced by a wide variety of functional groups, such as —H, —OH, —NH_2, —Br, —I, and —COOH. The present experiment demonstrates the conversion of an organoborane to an alcohol by oxidation with alkaline hydrogen peroxide. It is not necessary to isolate the organoborane prior to its oxidation. This simplification is particularly fortuitous in this case, since most alkylboranes, when not in solution, are pyrophoric (spontaneously flammable in air).

With regard to the second stage of the hydroxylation process, there is now conclusive evidence that oxidation of the C—B bond proceeds with retention of configuration at the carbon atom bearing the boron. That is, the hydroxyl group that replaces the boron atom has the identical orientation in the molecule as the boron:

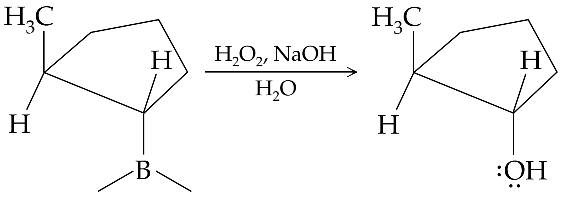

Thus, in unsymmetrical alkenes the hydroboration–oxidation sequence of reactions leads to the addition of the elements of H—OH to the original C═C in an anti-Markovnikov manner.

In the oxidation step a hydroperoxide anion (HOO^-) is generated in the alkaline medium. This species makes a nucleophilic attack on the boron atom to form a boron hydroperoxide. A 1,2 migration of an alkyl group from boron to oxygen occurs to yield a boron monoester (a borate). Hydrolysis of the boron triester, generated by successive rearrangement of all three alkyl groups, produces the desired alcohol. The mechanism of the oxidation sequence is given below.

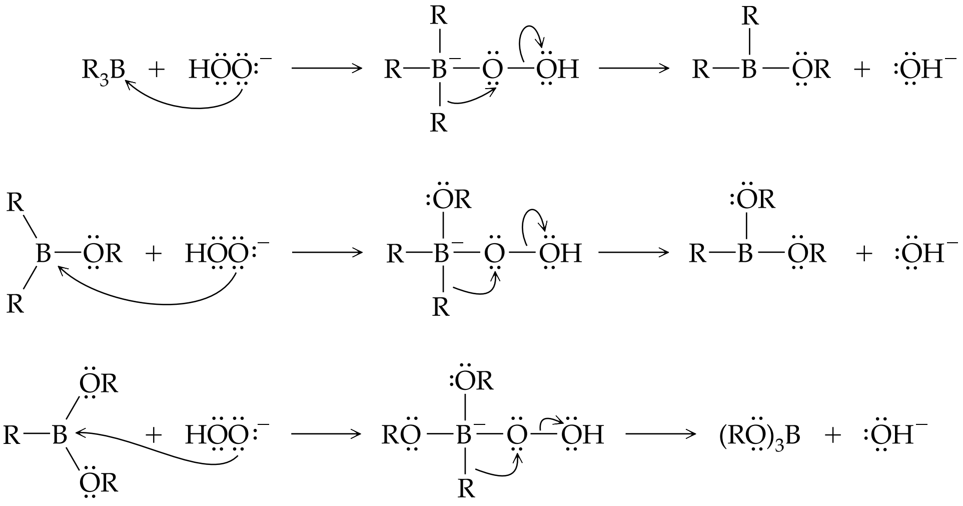
R3B + HÖÖ:⁻ ⟶ R—B⁻—Ö—ÖH ⟶ R—B—ÖR + :ÖH⁻
R
R
R
R—B—ÖR + HÖÖ:⁻ ⟶ R—B⁻—Ö—ÖH ⟶ R—B—ÖR + :ÖH⁻
R
:ÖR
R
:ÖR
R—B + HÖÖ:⁻ ⟶ RÖ—B⁻—Ö—ÖH ⟶ (RÖ)3B + :ÖH⁻
ÖR
ÖR
:ÖR
R

$$B(OR)_3 + OH^- \xrightarrow{H_2O} 3\,ROH + B(OH)_4^-$$

In the final step, alkaline hydrolysis of the trialkyl borate ester yields 3 moles of the alcohol.

The effective use of B_2H_6 in the hydroboration reaction was discovered in 1955 by H. C. Brown, and is just one of the many important hydride reagents developed by Professor Brown and his coworkers at Purdue University.

Herbert Charles Brown (1912) Brown obtained his B.S. in chemistry from the University of Chicago (1936) and his Ph.D. from the same institution in 1938. He later became a Professor of Chemistry at Wayne State and Purdue Universities.

Working with H. I. Schlesinger at the University of Chicago, Brown developed practical routes for the synthesis of diborane (B_2H_6). He discovered that diborane reacted rapidly with LiH to produce lithium borohydride ($LiBH_4$), discovering and opening a synthetic route to the metal borohydrides. These compounds proved to be powerful reducing agents. Later, he developed an effective route to $NaBH_4$, which led to the commercial production of this material. Metal borohydrides, particularly $LiAlH_4$ (developed by Schlesinger and

Albert Finholt), have revolutionized how organic functional groups are reduced in both the research laboratory and the industrial plant.

In 1955, Brown discovered that alkenes can be converted to organoboranes by reaction with diborane (actually the monomer in ether solution) and with organoboranes containing a B—H bond (the hydroboration reaction). The organoboranes are valuable intermediates in organic synthesis because the boron substituent can be quickly and quantitatively replaced by groups such as —OH, —H, —NH_2, or —X (halogen). Thus, organoboranes have become an attractive pathway for the preparation of alcohols, alkanes, amines, and organohalides.

Brown's investigation of the addition compounds of trimethyl borane, diborane, and boron trifluoride with amines has provided a quantitative estimation for steric strain effects in chemical reactions. He also investigated the role of steric effects in solvolytic, displacement, and in elimination reactions. His results demonstrate that steric effects can assist, as well as hinder, the rate of a chemical reaction.

Brown has published over 700 scientific papers and is the author of several texts. For his extensive work on organoboranes, Brown (with G. Wittig [organophosphorus compounds]) received the Nobel Prize in Chemistry in 1979.[13]

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 4.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	d	n_D
1-Octene	112.22	210 μL	1.34	121	0.72	1.4087
Borane • THF (1M)		500 μL	0.50			
Sodium hydroxide (3 M)	40.00	300 μL				
Hydrogen peroxide (30%)	34.01	300 μL				

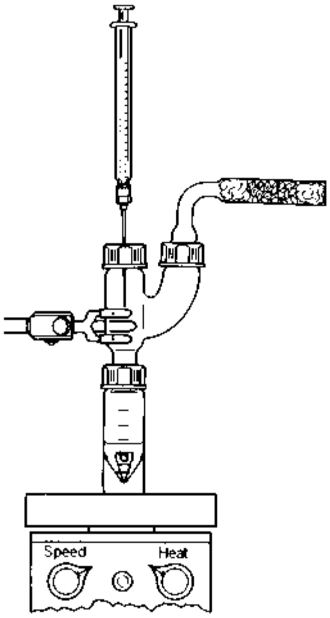

1-Octene, 210 μL + 1M BH_3•THF, 500 μL

Reagents and Equipment. Equip a 5.0-mL conical vial, containing a spin vane, with a Claisen head fitted with a rubber septum and calcium chloride drying tube (⬅). Through the rubber septum add 210 μL (150 mg, 1.34 mmol) of 1-octene (in one portion) with a 1.0-cm^3 syringe.

NOTE. Dry the glassware and syringe in a 100 °C oven for at least 30 min before use. An alternate method is to "flame-out" the glassware with a microburner and a flow of dry nitrogen, and to then add the drying tube and caps.

Cool the reaction vessel in an ice bath and, using the same syringe, add 500 μL (0.5 mmol) of the 1 M borane · THF solution through the septum over a 5-min period.

CAUTION: The BH_3 • THF reagent reacts *violently* with water.

---

[13] See *McGraw-Hill Modern Scientists and Engineers;* S. P. Parker, Ed.; McGraw-Hill: New York, 1980, Vol. 1, p. 150.

Reaction Conditions. Allow the reactants to warm to room temperature and then stir for 45 min. Using a Pasteur pipet, carefully add two drops of water to hydrolyze any unreacted borane complex.

NOTE. At this stage of the procedure the vial may be removed from the Claisen head, capped, and allowed to stand until the next laboratory period.

If the experiment is not interrupted, proceed by first removing the reaction vial from the Claisen head. Then use a graduated 1.0-mL pipet to add 300 μL of 3 M NaOH solution, followed by the dropwise addition of 300 μL of 30.0% hydrogen peroxide solution over a 10-min period using another graduated 1.0-mL pipet. Stir the reaction vial gently after each addition.

CAUTION: Hydrogen peroxide blisters the skin. Concentrated solutions of hydrogen peroxide can explode!

Attach the vial to a reflux condenser and warm the reaction mixture with stirring, for 1 h in a sand bath at 40–50 °C (➡).

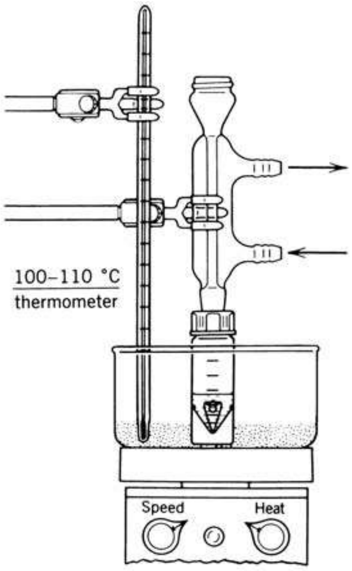

3 *M* NaOH, 300 μL +
30% H_2O_2, 300 μL +
$[CH_3(CH_2)_7]_3B$ intermediate +
0.5 mL THF

Cool the resulting two-phase mixture to room temperature and use forceps to remove the spin vane. Add 0.5 mL of diethyl ether to establish a reasonable volume for extraction of the organic phase.

Isolation of Product. Using a Pasteur filter pipet, separate the bottom aqueous layer and transfer it to a 3.0-mL reaction vial. *Save* the organic phase in the 5.0-mL conical vial.

Extract the aqueous phase placed in the 3.0-mL conical vial with two 1.0-mL portions of diethyl ether. Upon the addition of each portion of ether, cap, shake, and carefully vent the vial and allow the layers to separate. The top ether layer is then separated using a Pasteur filter pipet and the ether extracts are combined with the previously saved organic phase in the 5.0-mL conical vial. *If a solid forms during the extraction, add a few drops of 0.1 M HCl.*

Extract the combined organic phases with 750 μL of 0.1 M HCl solution, followed by extraction with several 0.5-mL portions of distilled water or until the aqueous extract is neutral to pH paper. Transfer the neutral organic phase to a 10-mL Erlenmeyer flask. Rinse the conical vial with a further 0.5 mL of ether and combine the rinse with the ether solution in the Erlenmeyer flask. Add granular anhydrous sodium sulfate (Na_2SO_4) (200 mg) to the combined organic phases and let it stand with occasional swirling for 20 min. If large clumps of drying agent form, add an additional 100 mg of Na_2SO_4. The solution should be clear at the end of the drying period. Then transfer the solution to a tared 3.0-mL conical vial in 1-mL aliquots, add a boiling stone to the vial, and concentrate by warming in a sand bath (60–65 °C) in the **hood** to yield the crude product residue. HOOD

Weigh the vial and calculate the crude yield.

Gas Chromatographic Analysis. This crude product is easily analyzed by gas chromatography. The procedure involves the injection of 10 μL of the liquid material onto a $\frac{1}{4}$-in. × 8-ft steel column packed with 10% Carbowax 80/100 20M PAW-DMS. Experimental conditions are He flow rate, 50 mL/min; chart speed, 1 cm/min; temperature, 190 °C.

The liquid components elute in the order: unreacted 1-octene, a small amount of 2-octanol, and the major product, 1-octanol. Approximate retention times (conditions above) are 1.3, 3.1, and 4.2 min, respectively.

Collect the eluted 1-octanol in an uncooled, 4-mm-diameter collection tube. Transfer the collected material to a 0.1-mL conical vial (see Technique 1). This material may then be analyzed by IR spectroscopy and/or used in the procedure described in the following section on Purification and Characterization.

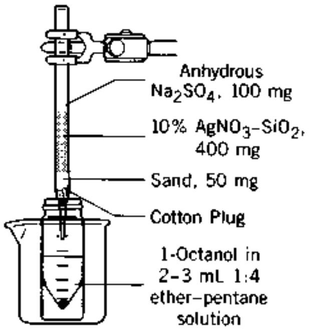
Anhydrous Na_2SO_4, 100 mg
10% $AgNO_3$–SiO_2, 400 mg
Sand, 50 mg
Cotton Plug
1-Octanol in 2–3 mL 1:4 ether–pentane solution

Purification and Characterization. Pack a Pasteur filter pipet with 400 mg of 10% silver nitrate-treated activated silica gel followed by 100 mg of anhydrous sodium sulfate. Dissolve the organic residue isolated above in 500 μL of pentane (spectral or HPLC grade) and then transfer this solution by Pasteur filter pipet to the column (⬅). Elute the material from the column with 3 mL of a 1:4 diethyl ether–pentane solution. Collect the eluate in a tared 5.0-mL conical vial containing a boiling stone.

HOOD

Concentrate the collected eluate to a constant weight by warming on a sand bath (60–65 °C.) in the **hood.** Weight the octanol product and calculate the percent yield.

This product may again be analyzed by gas chromatography. Follow the procedure and experimental conditions outlined above. The 1-octene impurity should have been removed during the column chromatography step. The small percentage of 2-octanol byproduct, however, more than likely will still be detected. The actual percent composition of the mixture can be calculated by determination of the areas under the chromatographic peaks (see Technique 1).

Obtain an IR spectrum of the alcohol and compare your result to that recorded in the literature (*Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Chemical Tests. A positive ceric nitrate test (Chapter 9) should confirm the presence of the alcohol grouping. The ignition test may be used to establish that the material is an aliphatic species. The phenyl or α-naphthylurethane derivative may also be prepared to further characterize the alcohol (Chapter 9).

It might also be of interest to determine the solubility characteristics of this C_8 alcohol in water, ether, concentrated sulfuric acid, and 85% phosphoric acid (Chapter 9). Do your results agree with what you would predict for this alcohol?

What chemical tests would you perform to determine the difference between the starting alkene and the alcohol product?

QUESTIONS

6-87. Using the hydroboration reaction, outline a reaction sequence for each of the following conversions:
  (a) 1-Butene to 1-butanol
  (b) 1-Methylcyclohexene to *trans*-2-methylcyclohexanol
  (c) 2-methylpropene to 2-methyl-1-propanol

6-88. When diborane (B_2H_6) dissociates in ether solvents, such as tetrahydrofuran (THF), a complex between borane (BH_3) and the ether is formed. For example,

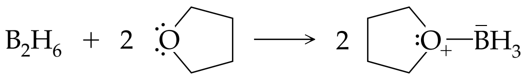

  (a) In the Lewis sense, what is the function of BH_3 as it forms the complex? Explain.
  (b) Write the Lewis structure for BH_3. Diagram its expected structure indicating the bond angles in the molecule.

6-89. In reference to question 6-88a:
- **(a)** Explain why borane (BH_3) reacts readily with the π-electron system of an alkene.
- **(b)** Explain why diborane (B_2H_6) reacts only very slowly with C═C groups.

6-90. In an unsymmetrical alkene, the boron atom adds predominantly to the least substituted carbon atom. For example, 2-methyl-2-butene gives the products indicated below:

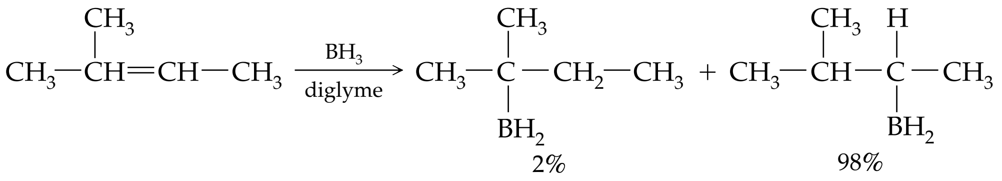

$$CH_3-\overset{CH_3}{\overset{|}{CH}}=CH-CH_3 \xrightarrow[\text{diglyme}]{BH_3} \underset{2\%}{CH_3-\underset{BH_2}{\underset{|}{\overset{CH_3}{\overset{|}{C}}}}-CH_2-CH_3} + \underset{98\%}{CH_3-\overset{CH_3}{\overset{|}{CH}}-\underset{BH_2}{\underset{|}{\overset{H}{\overset{|}{C}}}}-CH_3}$$

(Diglyme: $CH_3\ddot{\underset{..}{O}}-CH_2CH_2-\ddot{\underset{..}{O}}-CH_2CH_2-\ddot{\underset{..}{O}}CH_3$, diethyleneglycol dimethyl ether)

Offer a reasonable explanation to account for the ratio obtained.

NOTE. The above solvent (diglyme) has been shown to cause a significant increase in the number of miscarriages by workers who come in contact with it.

6-91. An advantage of the hydroboration reaction is that rearrangement of the carbon skeleton does not occur. This lack of migration contrasts with results obtained upon the addition of hydrogen chloride to the double bond. For example,

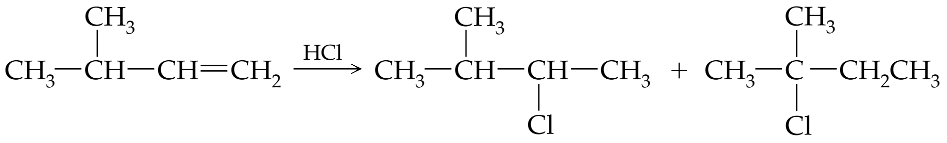

$$CH_3-\overset{CH_3}{\overset{|}{CH}}-CH=CH_2 \xrightarrow{HCl} CH_3-\overset{CH_3}{\overset{|}{CH}}-\underset{Cl}{\underset{|}{CH}}-CH_3 + CH_3-\underset{Cl}{\underset{|}{\overset{CH_3}{\overset{|}{C}}}}-CH_2CH_3$$

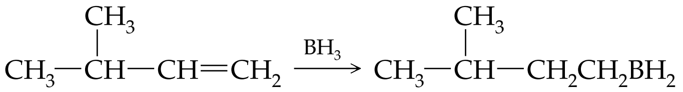

$$CH_3-\overset{CH_3}{\overset{|}{CH}}-CH=CH_2 \xrightarrow{BH_3} CH_3-\overset{CH_3}{\overset{|}{CH}}-CH_2CH_2BH_2$$

Offer an explanation for the difference in these results.

BIBLIOGRAPHY

Brown, H. C.; Subba, B. C. *J. Am. Chem. Soc.* **1956,** *78,* 5694.

Matteson, D. S. in *The Chemistry of the Metal-Carbon Bond,* Hartley, F. R., Ed.; Wiley: New York, 1987, Vol: 4, Chapter 3, p. 307.

Pelter, A.; Smith, K.; Brown, H. C. *Boron Reagents: Best Synthetic Methods;* Academic Press: New York, 1988.

Smith, K. *Organoboron Chemistry in Organometallics in Synthesis —A Manual;* Schlosser, M., Ed.; Wiley: New York, 1994.

Also see

Kabalka, G. W.; Maddox, J. T.; Shoup, T.; Bowers, K. R. *Organic Syntheses* **1996,** *73,* 116.

Experiment [13] was based on the work reported by

Kabalka, G. W. J. *Chem. Educ.* **1975,** *52,* 745; also see Kabalka, G. W; Wadgaonkar, P. P.; Chatla, N. *J. Chem. Educ.* **1990,** *67,* 975.

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Wiley-Interscience: New York, 2007, Chap. 15, p. 1075.

Diels–Alder Reaction: 4-Cyclohexene-cis-1,2-dicarboxylic Acid Anhydride

EXPERIMENT 14

Common name: 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride

CA number: [85-43-8]

CA name as indexed: 1,3-isobenzofurandione, 3a,4,7,7a-tetrahydro-

Purpose. This experiment demonstrates the use of the Diels–Alder reaction in the preparation of six-membered carbocyclic rings. The cyclic products are obtained by reaction of a conjugated diene with an alkene. The illustration given here involves the treatment of 1,3-butadiene (generated in situ) with maleic anhydride to form the corresponding Diels–Alder product. These *add*ition prod*uct*s are often called *adducts*.

Prior Reading

Technique 7: Collection or Control of Gaseous Products (pp. 105–107)
Chapter 8: Infrared Spectroscopy (pp. 539–561)
Nuclear Magnetic Resonance Spectroscopy (pp. 561–593)

REACTION

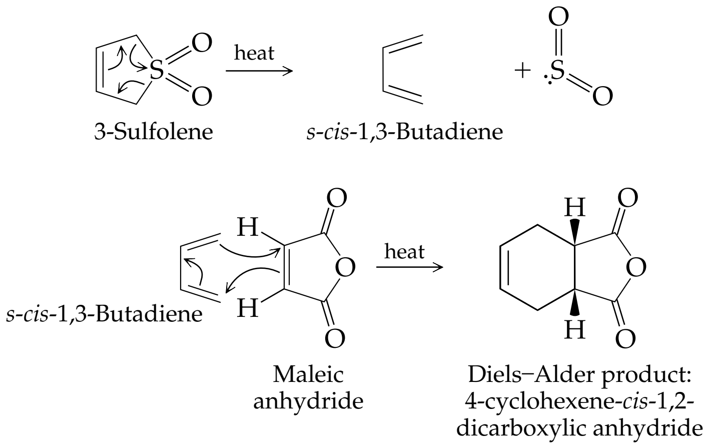

Otto Paul Hermann Diels (1876–1954) Diels obtained his Ph.D. in 1899 while studying with Emil Fischer at the University of Berlin. He later became Associate Professor of Chemistry at the University of Berlin, and in 1916 he moved to the University of Kiel. In 1906 Diels discovered carbon suboxide gas (C_3O_2), obtained from the dehydration of malonic acid. He did extensive studies on saturated fats and fatty acids. Diels also developed the use of selenium as a mild dehydrogenation agent. This latter work led to the commercial production of polyunsaturated oils.

In the same year that he identified carbon suboxide, Diels began to investigate cholesterol with E. Abderhalden. The structure of this lipid had not yet been determined. He was the first to study the products of the selenium dehydrogenation of cholesterol and isolated a hydrocarbon ($C_{18}H_{16}$), which became known as "Diels' hydrocarbon." This substance proved to possess the basic steroidal ring structure; its subsequent synthesis by Diels in 1935 (almost 30 years after he started this work) led to the rapid elucidation of the structures of a vast array of steroidal sex hormones, saponins, cardiac glycosides, bile pigments, and adrenal cortical hormones, such as cortisone.

Diels, however, is best known for his discovery (with his student Kurt Alder) of the reaction that now bears his name (it is now known as the *Diels–Alder cycloaddition reaction*), which was first published in 1928 (see Diels, O.; Alder, K. *Liebigs Ann. Chem.* **1928,** *460,* 98). This reaction involves the 1,4 addition of dienophile reagents to diene substrates to produce six-membered cycloalkenes. The reaction has found extensive application in the synthesis of terpenes and other natural products, since six-membered rings abound in the metabolites of living systems, and because for some time it was one of the few methods available for the synthesis of these cyclic structures. Because of the impact of their work in the field of organic synthesis, Diels shared the 1950 Nobel Prize (in chemistry) with Alder. He also was the author of a popular textbook (*Einfuhrung in die organische Chemie*), first published in 1907, which went through 19 editions by 1962.[14]

DISCUSSION

The Diels–Alder reaction is one of the most useful synthetic reactions in organic chemistry because, in a *single* step, it produces two new carbon–carbon bonds and up to four stereocenters. It is an example of a [4 + 2] cycloaddition reaction (4 π electrons + 2 π electrons) between a *conjugated* 1,3-diene and an alkene (dienophile; to have an affinity for dienes, from the Greek *philos,* meaning loving), which leads to the formation of cyclohexenes. Alkynes may also be used as dienophiles, in which case the reaction produces 1,4-cyclohexadienes. The reaction proceeds faster if the dienophile bears electron-withdrawing groups and if the diene bears electron-donating groups. Thus, α,β-unsaturated esters, ketones, nitriles, and so on, make excellent dienophiles, which are often used in the Diels–Alder reaction. By varying the nature of the diene and dienophile, a very large number of compounds can be prepared. Unsubstituted alkenes, such as ethylene, are poor dienophiles and react with 1,3-butadiene only at elevated temperatures and pressures. These high activation energies (slow reactions) pose a particular problem for the Diels–Alder reaction. The Diels–Alder reaction is a reversible, equilibrium reaction that is not very exothermic. Since the equilibrium constant (K_{eq}) is temperature dependent [$K_{eq} = e^{-\Delta G/RT}$], K_{eq} decreases with increasing temperature and eventually can become quite small at the high temperature needed for the reaction of an unactivated dienophile to proceed at a reasonable rate. Elevating the temperature will increase the rate of the reaction, but this will also reduce the amount of product formed, and therefore the lowest possible temperature must often be used.

The reaction is a thermal *cycloaddition* (a ring is formed), which occurs in one step and is thus a *concerted* reaction. Both new C—C single bonds and the new C═C π bond are formed simultaneously, as the three π bonds in the reactants break. The electron flow for the reaction is shown below. The reaction is thus classified as a *pericyclic* reaction (from the Greek meaning "around the circle").

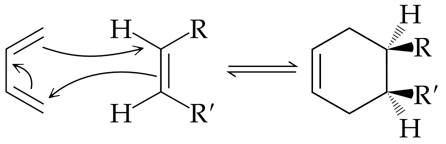

The diene component must be in the *s-cis* (the "*s*" refers to the conformation about a single bond) conformation to yield the cyclic product with the cis C═C required by the six-membered ring. For this reason, cyclic dienes usually react more readily than acyclic species. For example, 1,3-cyclopentadiene, which is locked in the *s-cis* configuration, reacts with maleic anhydride about 1000 times faster than 1,3-butadiene, which prefers an *s-trans* conformation.

---

[14] See Newett, L. C. *J. Chem. Educ.* **1931,** *8,* 1493; *Dictionary of Scientific Biography,* C. C. Gillespie, Ed.; Scribner's: New York, 1971, Vol. IV, p. 90; *McGraw-Hill Modern Scientists and Engineers,* S. P. Parker, Ed.; McGraw-Hill: New York, 1980, Vol.1, p. 289.

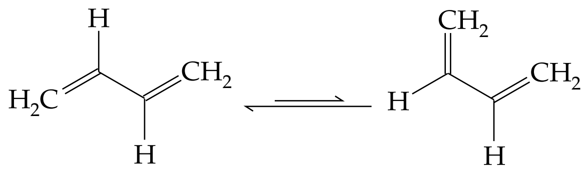

s-trans-1,3-Butadiene *s-cis*-1,3-Butadiene

The reaction is highly stereospecific and the orientation of the groups on the dienophile are *retained* in the product; thus, the addition must be suprafacial–suprafacial. That is, by having the stereochemical information preserved, both new bonds are formed on the same face of the diene and on the same face of the dienophile. Thus, two groups that are cis on the dienophile will be cis in the product (and trans will give a trans product).

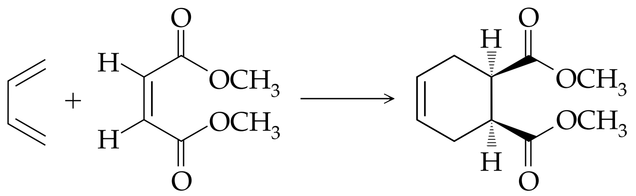

Carbomethoxy groups cis Carbomethoxy groups cis

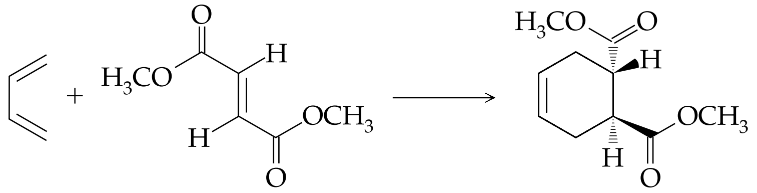

Carbomethoxy groups trans Carbomethoxy groups trans

The reaction of cyclopentadiene with maleic anhydride demonstrates the further stereochemical consequence of the relative orientation of the reactants in Diels–Alder reactions. In this situation, there are two possible ways in which the reactants may bond. This reaction leads to the formation of two products: the endo and exo stereoisomers:

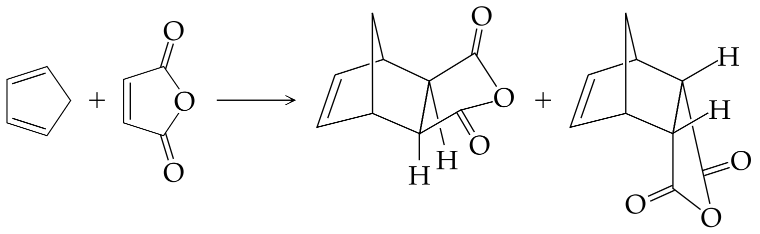

Generally, the endo form of the product predominates (endo-rule which is a result of secondary orbital overlap between the diene and dienophile), but endo/exo ratios may vary, depending on several steric and electronic factors and with reaction conditions.

The diene used in this experiment, 1,3-butadiene, is a gas at room temperature (bp −5 °C), which makes it a difficult reagent to measure and handle in the laboratory. Fortunately, 1,3-butadiene can be generated in situ from 3-sulfolene, a solid reagent that is easily handled. In an example of a retro-cycloaddition reaction, 3-sulfolene decomposes at a moderate temperature to yield sulfur dioxide and 1,3-butadiene.

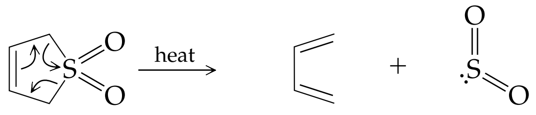

3-Sulfolene s-*cis*-1,3-Butadiene

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

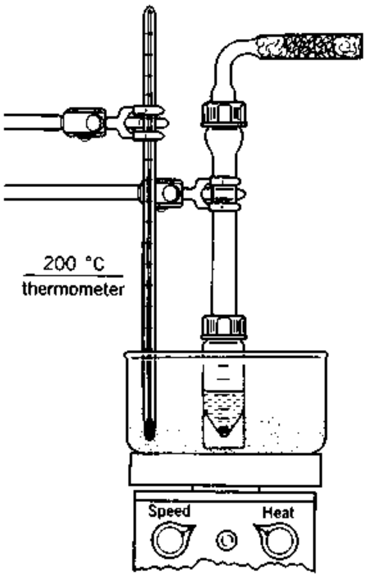

3-Sulfolene, 170 mg,
+ maleic anhydride, 90 mg
+ xylene, 80µL

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
3-Sulfolene	118.15	170 mg	1.42	66	
Maleic anhydride	98.06	90 mg	0.92	60	
Xylene		80 µL			137–140

Reagents and Equipment. Weigh and place 80 µL of xylene, 90 mg (0.92 mmol) of maleic anhydride, and 170 mg (1.42 mmol) of 3-sulfolene in a 3.0-mL conical vial equipped with an air condenser protected with a calcium chloride drying tube and containing a boiling stone (➡).

NOTE. The maleic anhydride should be finely ground and protected from moisture to prevent hydrolysis to the corresponding acid. A mixture of xylenes in the boiling-point range 137–140 °C will suffice. Use freshly distilled solvent or dry it over molecular sieves before use. Dispense the xylene in the ***hood*** *using an automatic delivery pipet. In large laboratory sections it is recommended that the evolved* SO_2 *be trapped (see Prior Reading).* HOOD

Reaction Conditions. Heat the reaction mixture to reflux using a sand bath for 20 min. Avoid overheating.

CAUTION: The reaction is exothermic. Avoid overheating. Sulfur dioxide is evolved in the process and adequate ventilation should be provided.

Carefully remove the **hot** conical vial from the sand bath and allow the contents to cool to room temperature. HOT

Isolation of Product. Add 0.5 mL of toluene to the cooled solution, and then add petroleum ether (60–80 °C) dropwise until a slight cloudiness persists. Roughly 0.25–0.35 mL of petroleum ether will be needed.

Reheat the solution until it becomes clear and then cool it in an ice bath. *During the recrystallization step, the sides of the vial may have to be scratched with a glass rod to induce crystallization.*

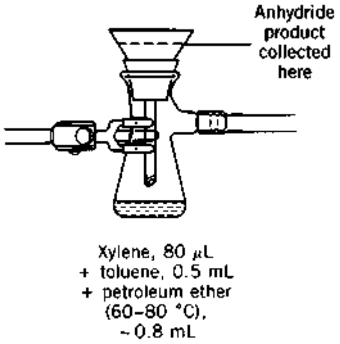
Anhydride product collected here
Xylene, 80 µL + toluene, 0.5 mL + petroleum ether (60–80 °C), ~0.8 mL

Collect the crystalline product by vacuum filtration and wash the filter cake on the Hirsch funnel with 0.5 mL of *cold* petroleum ether (60–80 °C) (⬅).

NOTE. Do not wash with an excess of the petroleum ether, or a loss of product will result.

Purification and Characterization. The product is of sufficient purity for direct characterization. Weigh the material and calculate the percent yield. Determine the melting point and compare your result with the value found in the literature. Obtain an IR spectrum of the material.

The reaction involves two reactants (butadiene and maleic anhydride), which both contribute functional groups to the product. The infrared spectrum of the isolated material reflects this observation. Compare the infrared spectrum of your product with that of the reference spectrum.

Infrared Analysis. The spectrum of one of the starting materials, 3-sulfolene (Fig. 6.27), is representative of an alkene sulfone. The macro group frequency train for an unconjugated five-membered ring alkene fits the data reasonably

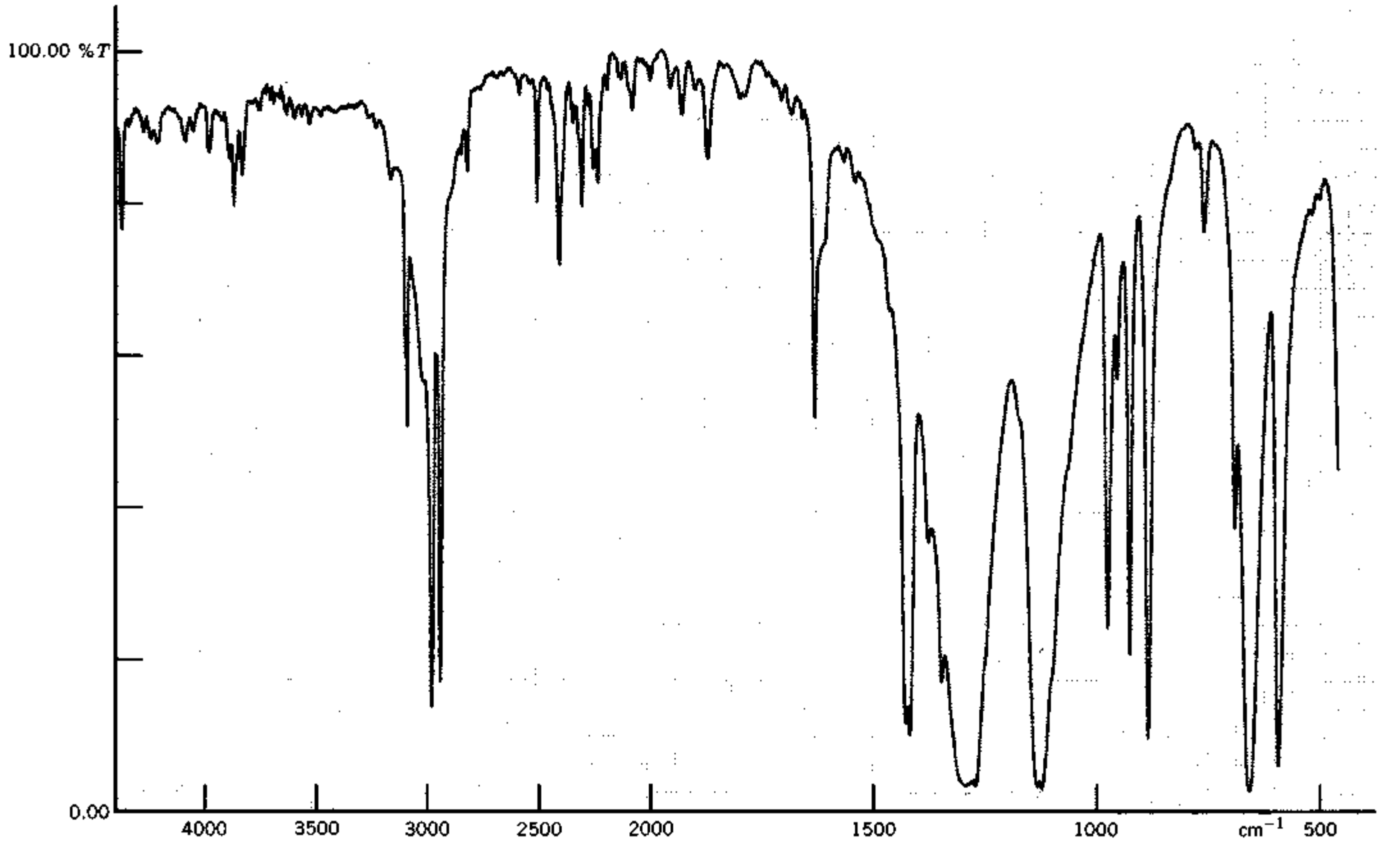

Sample 3-Sulfolene
%T X ABS — Background Scans 4 Scans 16
Acquisition & Calculation Time 42 sec Resolution 4.0 cm⁻¹
Sample Condition solid Cell Window
Cell Path Length Matrix Material KBr

Figure 6.27 **IR spectrum: 3-Sulfolene.**

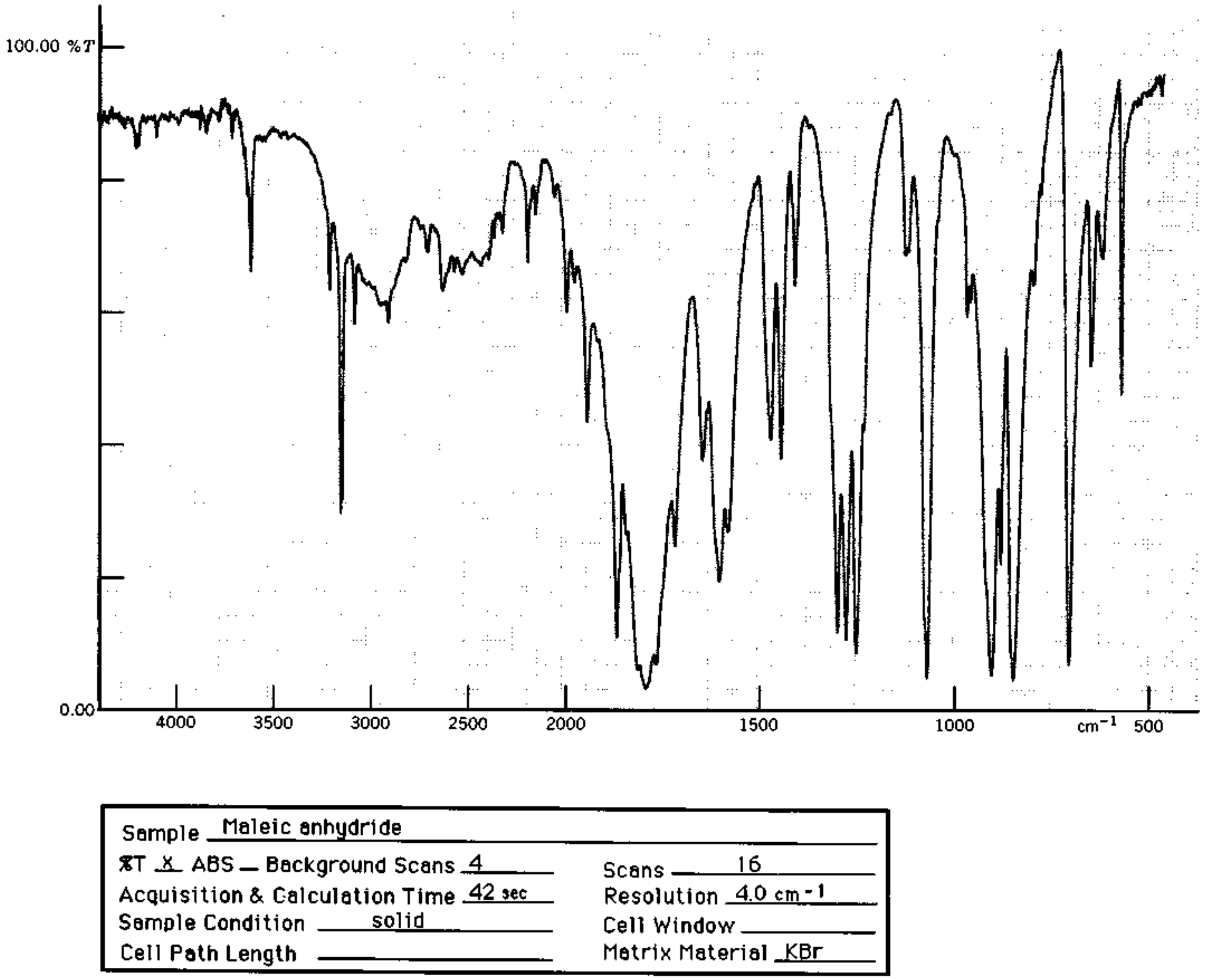

Sample	Maleic anhydride		
%T X ABS — Background Scans	4	Scans	16
Acquisition & Calculation Time	42 sec	Resolution	4.0 cm-1
Sample Condition	solid	Cell Window	
Cell Path Length		Matrix Material	KBr

Figure 6.28 IR spectrum: maleic anhydride.

well: 3090 (═C—H, stretch), 1635 C═C, stretch), and 657 (H—C ═C—H, cis, out-of-plane bend) cm^{-1}. The presence of the sulfone group is convincingly identified by the very strong bands at 1287 and 1127 cm^{-1}, which are assigned to the coupled in-phase and out-of-phase S—O stretching vibrations. The antisymmetric and symmetric C—H stretching vibrations of the methylene groups are observed at 2975 and 2910 cm^{-1}. Note that the influence of the heterocyclic five-membered ring system has raised all of the C—H stretching modes by 40–50 cm^{-1}.

The spectrum of the other reactant, maleic anhydride (Fig. 6.28), possesses all the peaks representative of a conjugated five-membered ring anhydride: 3110 (═C—H, stretch), 1858 (C═O, in-phase stretch, weak), 1777 (C═O, out-of-phase stretch, strong), 1595 (C═C, stretch, weak), 1060 (C—O, antisymmetric stretch), 899 (C—O, symmetric stretch), 835 (H—C═C—H, out-of-plane bend, normally near 700 cm^{-1}, but raised by conjugation to carbonyls). The in-phase stretch of the anhydride carbonyls is particularly weak as the five-membered ring system forces the two oscillators into nearly opposing positions.

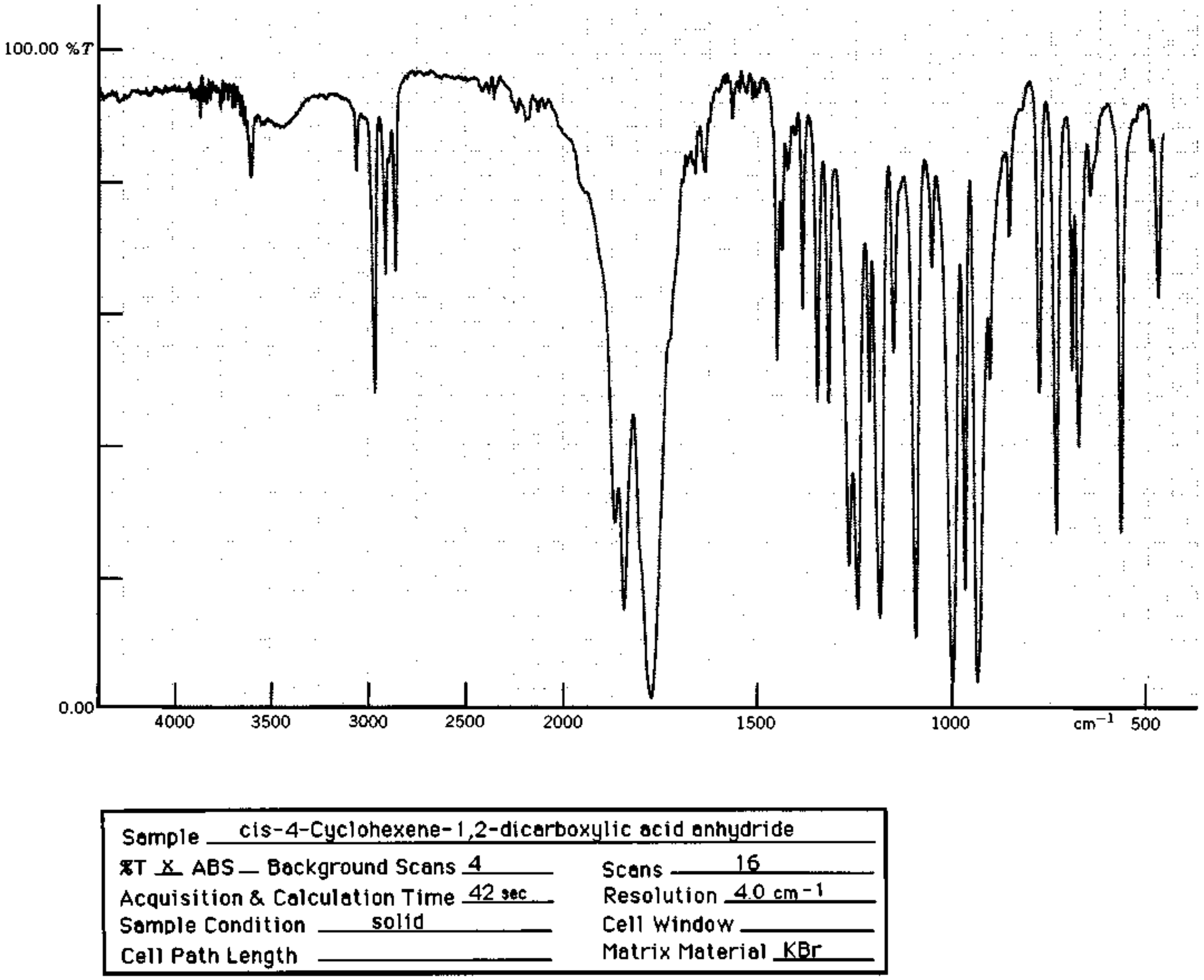

Sample	cis-4-Cyclohexene-1,2-dicarboxylic acid anhydride		
%T X ABS __ Background Scans	4	Scans	16
Acquisition & Calculation Time	42 sec	Resolution	4.0 cm-1
Sample Condition	solid	Cell Window	
Cell Path Length		Matrix Material	KBr

Figure 6.29 IR spectrum: 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride.

The Diels–Alder product exhibits all of the spectral properties of a nonconjugated, alkenyl five-membered anhydride (Fig. 6.29). The macro group frequency train for the anhydride portion of the molecule involves peaks at 1844, 1772, 998, and 935 cm^{-1}:

a. **1844 cm^{-1}:** This weak band is similar to the 1858-wavenumber band found in maleic anhydride, and involves the in-phase stretch of the two coupled carbonyl groups. The band is somewhat more intense in the case of this saturated five-membered ring example, because the ring system is more easily distorted from planarity. This band often exhibits some secondary splitting on the high-wavenumber side (see discussion of the 935-wavenumber band below).

b. **1772 cm^{-1}:** This band arises from the out-of-phase stretch of the anhydride carbonyls and is the most intense band in the spectrum. The separation, or splitting, between the two carbonyl stretching modes of anhydrides is relatively constant and falls in the range of 60–90 cm^{-1}.

JWCL196_c06_255-308.qxd 11/17/09 1:36 PM Page 265

EQA

c. **998 and 935 cm^{-1}:** These two bands are identified with the antisymmetric and symmetric C—O—C stretching modes. It is the overtone of the symmetric stretch that is Fermi coupled (see IR discussions) to the symmetric carbonyl stretch (935 × 2 = 1870 cm^{-1}. The frequency match is variable, with the harmonic generally falling on the high-wavenumber side of the fundamental.

The alkenyl section of the product possesses the following macro group frequency bands: 3065 (═C—H stretch), 1635 (C═C stretch, unsubstituted C═C, 5,6-membered fused ring system), and 685 (H—C═C—H, cis, out-of-plane bend) cm^{-1}.

The saturated C—H section of the cyclohexene ring possesses two identical methylene groups that contain the short macro frequency train of 2972, 2929 and 2860, and 1447 cm^{-1}:

a. **2972 cm^{-1}:** Antisymmetric C—H stretch of the —CH$_2$— group.
b. **2929 and 2860 cm^{-1}:** Split symmetric C—H stretch of the —CH$_2$— group. This mode is split by Fermi coupling (see Chapter 8) with the overtone of the symmetric methylene scissoring vibration found at 1447 cm^{-1} (1447 × 2 = 2894). The uncoupled fundamental mode would be expected to occur at or near 2890 cm^{-1}, and thus, falls very close to the predicted location of the overtone.
c. **1447 cm^{-1}:** Symmetric deformation vibration (scissoring motion) of the methylene group present in the cyclohexene system.

Examine the spectrum of your reaction product. Discuss the similarities and differences of the experimentally derived spectral data to the reference spectra (Figs. 6.27–6.29).

Nuclear Magnetic Resonance Analysis. The 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride provides an ideal example of the utility of ^{13}C NMR when only limited information is available from the ^{1}H NMR spectrum. The 300-MHz ^{1}H spectrum is shown in Figure 6.30, page 266. Due in part to the presence of two stereocenters, as well as long-range coupling through the π system of the alkene, the entire ^{1}H spectrum is second order and no information is available from the coupling constants because the spectrum is too complex. Limited assignments to peaks could be made on the basis of chemical shift, but it would be difficult to make any statements regarding the purity of your sample based on the ^{1}H NMR spectrum, since an impurity could well be hidden beneath any of the complex signals.

On the other hand, the fully ^{1}H-decoupled ^{13}C spectrum (Figure 6.31, page 267) of 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride is much less complex. Because of the mirror plane of symmetry in the compound, there are only four unique carbon atoms and thus only four peaks are seen in the ^{13}C NMR spectrum. The 1:1:1 triplet centered at 77 ppm is due to the solvent, $CDCl_3$. Although no ^{1}H–^{13}C coupling is seen as a result of the ^{1}H decoupling, ^{2}H–^{13}C coupling is observed because ^{1}H and ^{2}H (D) resonate at different frequencies.

Sample preparation for ^{13}C NMR is essentially the same as for ^{1}H NMR spectroscopy except that significantly more material is required to obtain a ^{13}C NMR spectrum in a reasonable amount of time. In this case, acceptable signal-to-noise levels can be obtained on 40 mg of material in about 10 min. Although tetramethylsilane (TMS) can be added as a reference, it is often

CONFIRMING PAGES aptara

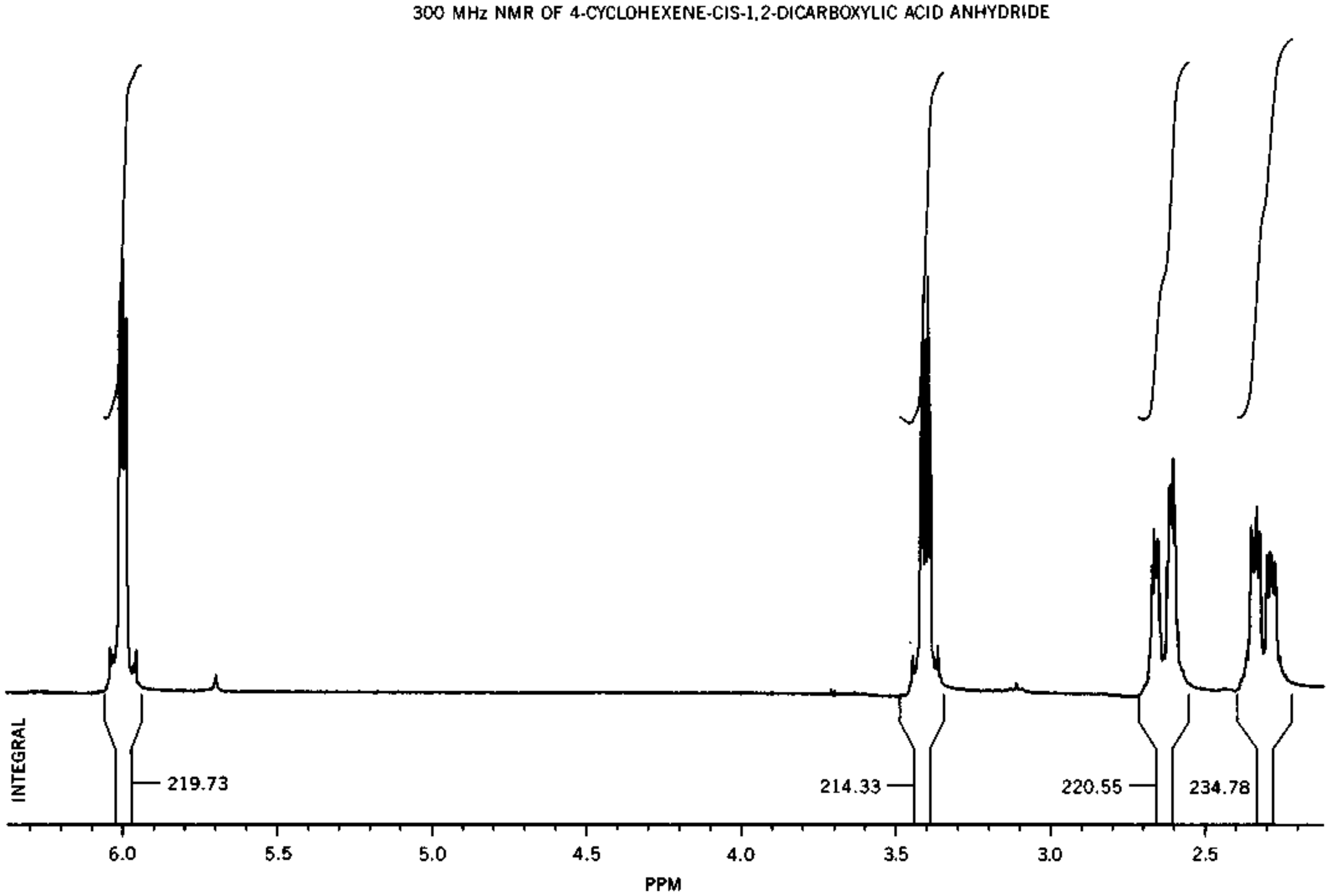

Figure 6.30 **^{1}H-NMR spectrum: 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride.**

more convenient to reference the spectrum relative to the known chemical shift of the solvent signal. Do not hesitate to use all of your material for the ^{13}C spectrum; it can be easily recovered later by emptying the NMR tube into a small vial, rinsing once with solvent, and evaporating the solvent in a hood under a gentle stream of dry nitrogen.

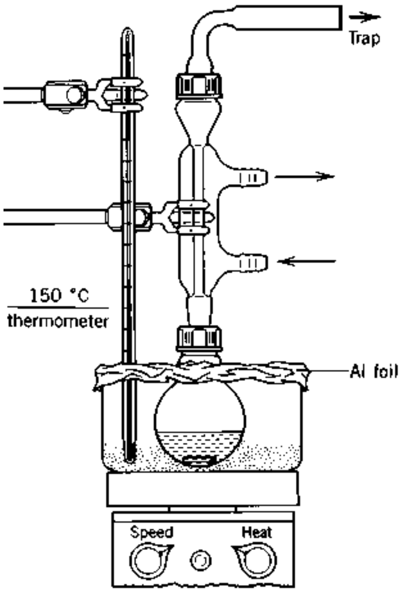

Reflux conditions
10-mL RB flask

Chemical Tests. Selected chemical classification tests can also be used to aid in characterization of this compound. Is the compound soluble in water? If so, does the aqueous solution turn blue litmus paper red? Is the compound soluble in 5% NaOH and 5% $NaHCO_3$? Is there evidence of CO_2 evolution with the bicarbonate solution? If so, what does this test indicate? Give the structure of the product formed when the material is added to the sodium hydroxide solution.

Perform the Baeyer test for unsaturation (Chapter 9). Is there evidence for the presence of a $\mathrm{C{=}C}$ bond?

OPTIONAL SEMIMICROSCALE PREPARATION

The microscale Diels–Alder addition may be scaled up by a factor of 5. The procedure is similar to that outlined above, with the exceptions noted below.

1. Use a 10-mL round-bottom flask containing a magnetic stirrer, equipped with a reflux condenser protected by a calcium chloride drying tube. Place an insulating aluminum foil guard between the bottom of the reflux condenser and the round-bottom flask (⬅).

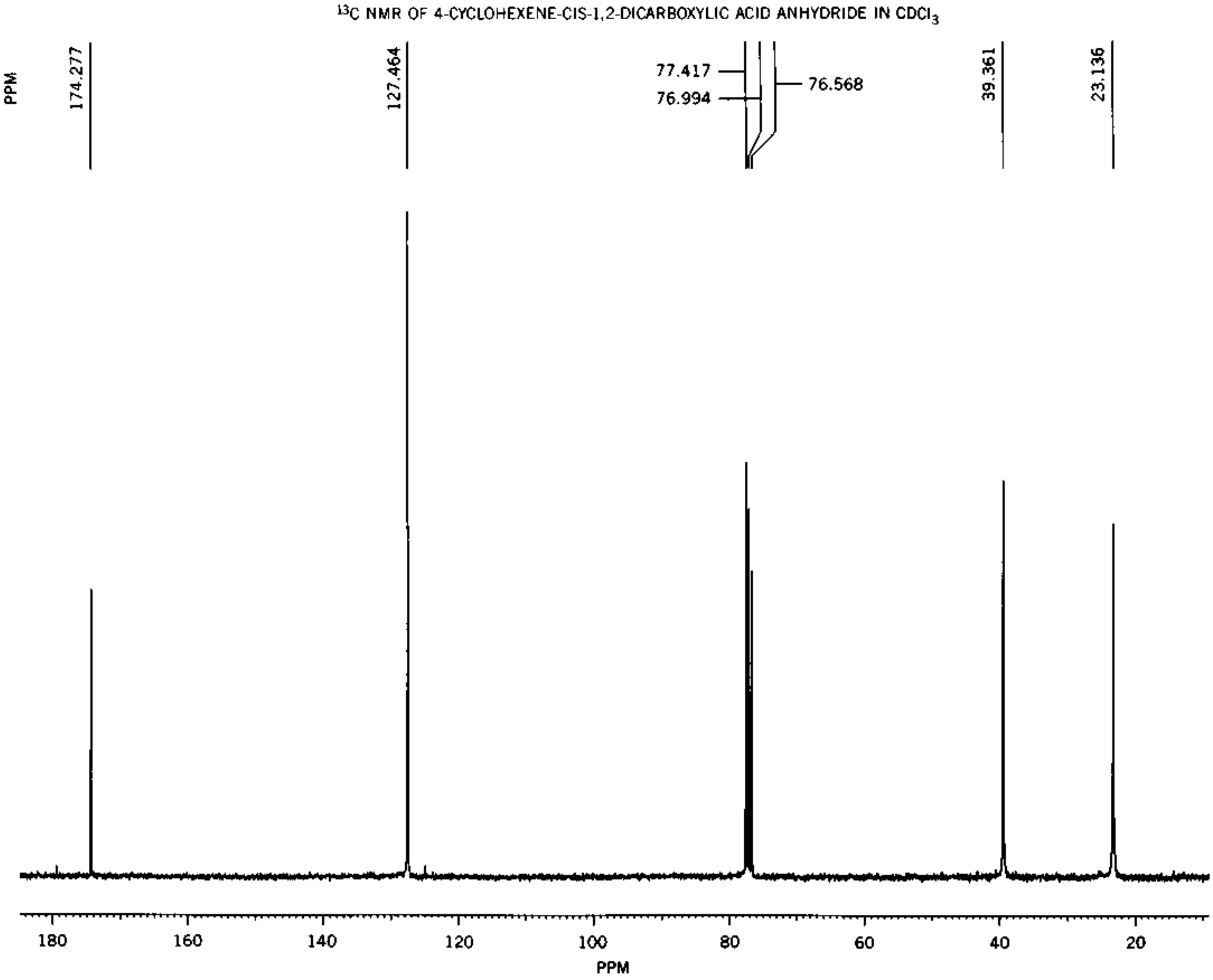

Figure 6.31 ^{13}**C-NMR spectrum: 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride.**

CAUTION: At this scale, due to the generation of a considerably larger quantity of sulfur dioxide than at the microscale level, the reaction must be run in the ***hood*** or provisions should be made to trap the evolved gas.

HOOD

2. Increase the reagent and solvent amounts about 5-fold:

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
3-Sulfolene	118.15	850 mg	7.2	66	
Maleic anhydride	98.06	450 mg	4.6	60	
Xylene		400 μL			137–140

3. Once in solution, heat the reaction mixture to reflux for 20 min. *Avoid overheating.* Then, cool the reaction to room temperature.

4. After cooling, add 5 mL of toluene and transfer the resulting clear solution to a 25-mL Erlenmeyer flask. The toluene solution may require heating and stirring to dissolve any crystals that may have formed. Now add petroleum ether (60–80 °C) (2.5–3.5 mL).

5. After isolation by vacuum filtration, wash the product with 5 mL of cold petroleum ether (60–80 °C).

6. Characterize the product as outlined in the microscale procedure (above).

QUESTIONS

6-92. Predict the product in each of the following Diels–Alder reactions.

(a)

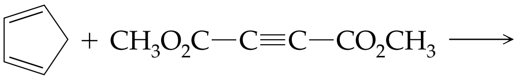

(d)

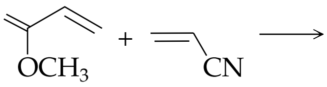

(b)

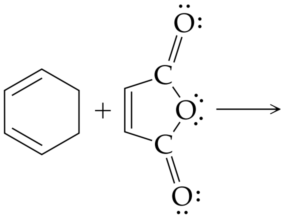

(e)

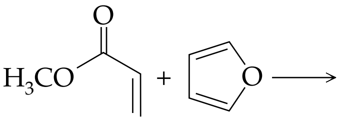

(c)

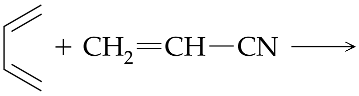

(f)

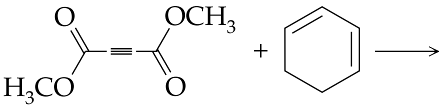

6-93. Cyclopentadiene reacts as a diene in the Diels–Alder reaction a great deal faster than does 1,3-butadiene. Explain.

6-94. Predict the diene and dienophile that would lead to each of the following products.

(a)

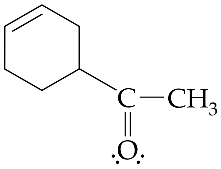

(c)

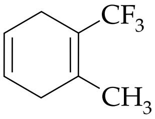

(e)

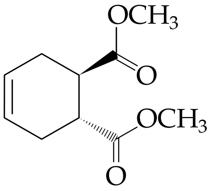

(b)

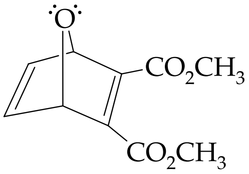

(d)

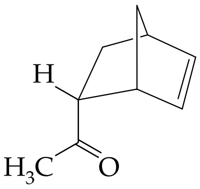

(f)

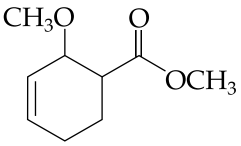

6-95. The decomposition of 3-sulfolene to form 1,3-butadiene generates 1 mol of sulfur dioxide gas per mol of 1,3-butadiene. Substantial quantities of SO_2 would be generated if this decomposition were carried out on a large scale. Suggest a method for trapping the gas to prevent its escape into the environment.

6-96. Two structural isomers are formed when 2-methyl-1,3-butadiene reacts with ethyl acrylate (ethyl 2-propenoate). Draw structures for these isomers.

6-97. In general, on going from the unsaturated fully conjugated anhydrides to the saturated systems, the two C—O stretching modes begin to coalesce, and in many cases only a single band is observed. In maleic anhydride, these modes occur at 1060 and 899 cm^{-1}, while in *cis*-4-cyclohexene-1,2-dicarboxylic acid anhydride they are found at 998 and 935 cm^{-1}. Explain this observation.

6-98. Acyclic anhydrides exhibit a reversal of the intensity relationship of the carbonyl stretching vibrations found in the cyclic anhydrides. That is, the lower wavenumber band is now the weaker member of the pair. Explain why this intensity exchange occurs.

6-99. If in Experiment [14] you had prepared the other possible diastereomer of the product, 4-cyclohexene-*trans*-1,2-dicarboxylic acid anhydride, how many lines would you expect to see in the ^{13}C NMR spectrum? Note that this isomer does not have a mirror plane of symmetry.

6-100. Does the ^{13}C NMR spectrum unambiguously demonstrate the position of the carbon–carbon double bond? How?

6-101. In ^{1}H spectra, the relative intensities of lines within a triplet are 1:2:1. The signal for $CDCl_3$ is a triplet with relative intensities of 1:1:1. Why?

BIBLIOGRAPHY

The preparation of 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride by the reaction of 1,3-butadiene with maleic anhydride is recorded:

Cope, A. C.; Herrick, E. C. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 890.

Other Diels–Alder reactions reported in *Organic Syntheses* include

Carlson, R. M.; Hill, R. K. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 196.

Greico, P.; Larsen, S. D. *Organic Syntheses;* Wiley: New York, 1933; Collect, Vol. VIII, p. 31.

Jung, M. E.; McCombs, C. A. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 445.

Kozmin, S. A.; He, S.; Rawal, V. H. *Organic Syntheses;* Wiley: New York, 2004; Collect. Vol. X, p. 442.

For selected reviews of the reaction see

Butz, L. W.; Rytina, A. W. *Org. React.* **1949,** *5,* 136.

Ciganek, E. *Org. React.* **1984,** *32,* 1.

Pindur, U.; Lutz, G.; Otto, C. *Chem. Rev.* **1993,** *93,* 741.

Rappoport, Z., Ed. *The Chemistry of Dienes and Polyenes;* Wiley: New York, 1997, Vol. I.

Weinreb, S. M. *Comp. Org. Syn.* **1991,** *5,* 513.

The conditions of this reaction were adapted from those reported by

Sample, T. E., Jr.; Hatch, L. F. *J. Chem. Educ.* **1968,** *45,* 55.

Diels–Alder Reaction: 9,10-Dihydroanthracene-9,10-α,β-succinic Acid Anhydride

EXPERIMENT 15

Common name: 9,10-dihydroanthracene-9,10-α,β-succinic acid anhydride
CA number: [85-43-8]
CA name as indexed: 1,3-isobenzofurandione, 3a,4,7,7a-tetrahydro-

Purpose. The Diels–Alder reaction is investigated. You will explore the role of an aromatic ring system as the diene substrate in this addition reaction. The reaction studied in this experiment is an example of a 1,4 addition by an activated alkene dienophile across the 9,10 positions of anthracene.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch funnel (pp. 88–89)

REACTION

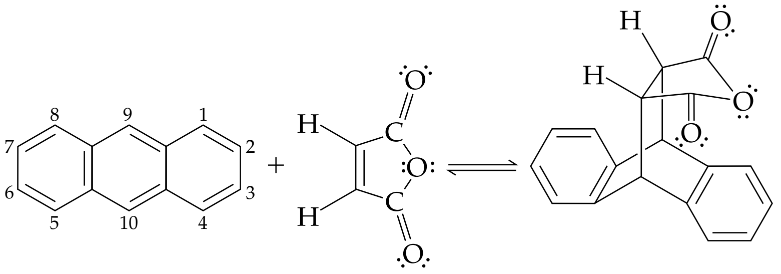

Anthracene

Maleic anhydride

9,10-Dihydroanthracene-9,10-α,β -succinic acid anhydride

This famous class of reactions are named for **Otto Paul Hermann Diels** and **Kurt Alder,** who were primarily responsible for its development. Diels and Alder received the Nobel Prize in 1950 for this work. See Experiment [14] for a biography of Diels. A short biography of his student Kurt Alder follows.[15]

Kurt Alder (1902–1958) Alder obtained his Ph.D. in 1926 while studying with Otto Diels at the University of Kiel. His dissertation was titled *Causes of the Azoester Reaction*. He became Professor of Chemistry at the University of Kiel in 1934 and later became head of the Chemical Institute at the University of Cologne (1940). For a few years (1936–1940) he was Research Director of the Baeyer dye works.

Together with Diels, Alder was responsible for the development of what came to be known as the Diels–Alder reaction. This reaction typically involves the reaction of a 1,3-conjugated diene with an activated alkene (dienophile) to form a six-membered cycloalkene. While reactions of this type had been reported as early as 1893, Diels and Alder were the first to recognize their great versatility. Alder continued to focus his academic research in this area following his graduate work with Diels. Over a number of years, Alder carried out a systematic study of the reactivity of a large number of dienes and dienophiles and established the structure and stereochemistry of many new adducts. He also expanded his doctoral research, studying the condensation of azoesters with dienes to yield the corresponding heterocyclic adducts.

Alder demonstrated that successful addition required that the diene double bonds possess an *s*-cis conformation (*s* refers to the single bond connecting the two double bonds). Furthermore, he realized that the bridged ring adducts formed by using cyclic dienes were closely related to natural products, such as camphor, and that this reaction offered a powerful route for the synthesis of a wide variety of naturally occurring compounds, particularly the terpenes. The Diels–Alder reaction also has been invaluable in the industrial synthesis of thousands of new organic materials from insecticides and dyes to lubricating oils and pharmaceuticals.

Alder investigated autooxidation and polymerization processes particularly during his industrial years. For example, he was involved in an extensive study of polymerizations related to the formation of Buna-type synthetic rubbers.[16]

DISCUSSION

This experiment is a further example of the Diels–Alder reaction. *For a discussion of the basic aspects of this reaction see Experiment [14]*. In the present case, the central ring of anthracene is shown to possess the characteristic properties of a diene system. Thus, this aromatic compound reacts to form stable Diels–Alder adducts with many dienophiles at the 9 and 10 positions (the two positions on the central ring where new bonds can be made without destroying the aromaticity of the other two rings). Maleic anhydride, a very reactive dienophile, is used here in the reaction with anthracene. Note, that as this reaction is reversible, it is usually best carried out at the lowest possible temperatures consistent with an acceptable reaction rate (see Experiment [14]).

---

[15] For references to the Diels–Alder reaction, see Experiment [14].

[16] See Allen, C. F. H. *J. Chem. Educ.* **1933,** *10*, 494; *Dictionary of Scientific Biography,* C. C. Gillespie, Ed.; Scribner's: New York, 1970, Vol. I, p. 105; *McGraw-Hill Modern Scientists and Engineers,* S. P. Parker, Ed., McGraw-Hill: New York, 1980, Vol. 1, p. 8.

Higher molecular weight polynuclear aromatic hydrocarbons (PAHs) containing the anthracene nucleus have also been found to react with maleic anhydride. These ring systems, however, can differ widely in reaction rates. Typical examples of those systems that undergo the Diels–Alder reaction are 1,2,5,6-dibenzanthracene (a), 2,3,6,7-dibenzanthracene (pentacene) (b), and 9,10-diphenylanthracene (c).

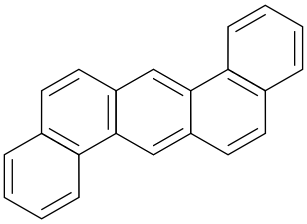
(a)

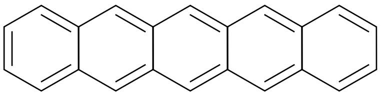
(b)

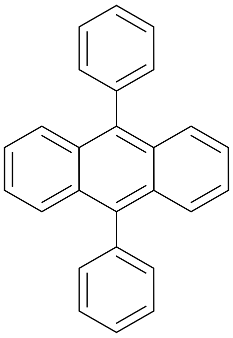
(c)

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
Anthracene	178.24	80 mg	0.44	216	
Maleic anhydride	98.06	40 mg	0.40	60	
Xylenes		1.0 mL			137–140
9,10-Dihydroanthracene-9,10-α,β-succinic acid anhydride	276			261–262	

Reagents and Equipment. Weigh and place 80 mg (0.44 mmol) of anthracene and 40 mg (0.40 mmol) of maleic anhydride in a 3.0-mL conical vial containing a boiling stone and equipped with an air condenser protected by a calcium chloride drying tube. Now, while in the **hood,** add 1.0 mL of xylene to the solid mixture using an automatic delivery pipet (➡).

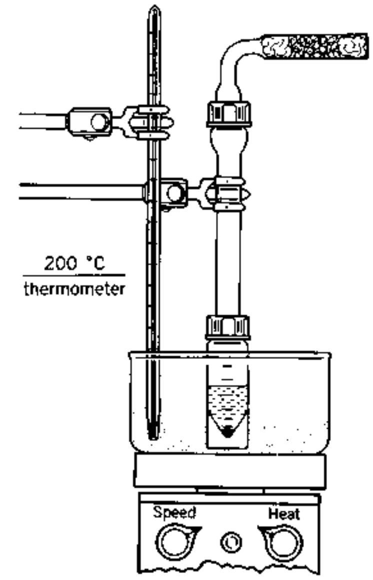

HOOD

Anthracene, 80 mg
+ maleic anhydride, 40 mg
+ xylene, 1.0 mL

NOTE. High-purity grades of anthracene and maleic anhydride are strongly recommended. Anthracene may be recrystallized from 95% ethanol. A mixture of xylenes with a boiling-point range of 137–140 °C is sufficient, but the solvent (xylenes) should be dried over molecular sieves before use.

Reaction Conditions. Heat the reaction mixture at reflux for 30 min in a sand bath at about 200 °C. During this time the initial yellow color of the reaction mixture gradually disappears. (Why?) Allow the resulting bleached solution to cool to room temperature and then place it in an ice bath for 10 min to complete the crystallization of the product.

Isolation of Product. Collect the crystals by vacuum filtration using a Hirsch funnel and wash the filter cake with two 300-μL portions of cold ethyl acetate (➡). Partially dry the filter cake under suction using plastic food wrap

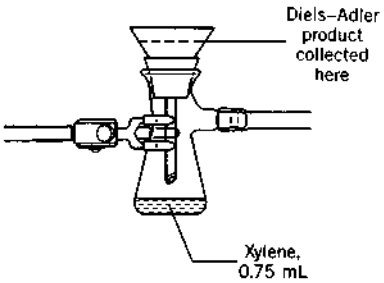

(see Prior Reading). Transfer the filtered, washed, and partially dried product to a porous clay plate or filter paper to complete the drying.

Purification and Characterization. The product is often of sufficient purity for direct characterization at this point. The adduct, however, may be further purified by recrystallization from ethyl acetate using a Craig tube. Weigh the dried Diels–Alder product and calculate the percent yield. Determine the melting point and compare your result with the value listed above.

NOTE. It is suggested that a hot-stage melting point apparatus, such as the Fisher–Johns, be used due to the high melting point of the product.

Chemical Tests. The *ignition test* is often used as a preliminary method to categorize hydrocarbon materials (see Chapter 9). Aromatic compounds give a yellow, sooty flame. Perform this test on anthracene. Based on your results can anthracene be classified as aromatic?

Carry out the ignition test on the following compounds and determine whether they should be classified as aromatic materials.

a. Toluene
b. Octane
c. Isopropyl alcohol
d. Nitrobenzene
e. *trans*-Stilbene

OPTIONAL SEMIMICROSCALE PREPARATIONS

This Diels–Alder reaction may be scaled up by a 5- or 50-fold increase in reactants. The scaled-up procedures are nearly identical in either case to that given for the microscale preparation; changes are noted below.

1. Fivefold Scaleup

a. In place of the conical vial as the reaction vessel, use a 10-mL round-bottom flask containing a magnetic stirrer connected to the water-jacketed reflux condenser that is protected by a calcium chloride drying tube (⬅).

b. The reagent and solvent quantities are given in the following table:

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
Anthracene	178.24	400 g	2.2	216	
Maleic anhydride	98.06	200 g	2.0	60	
Xylene	106.16	5.0 mL			137–140
Ethyl acetate	88.11	3.0 mL			77

c. The reaction flask is **heated at 185–190 °C** bath temperature for 30 min to obtain optimized yields.

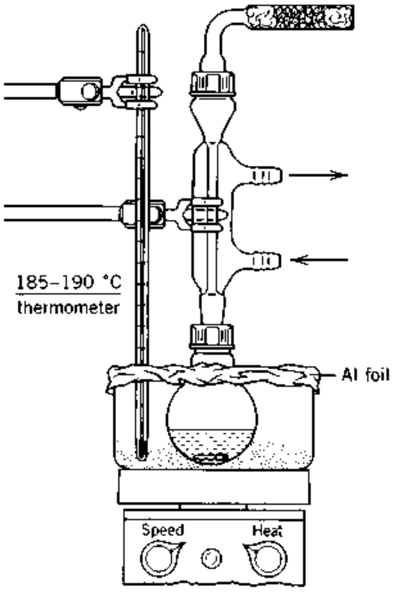

Reflux conditions
10- or 100-mL RB flask

2. Fiftyfold Scaleup

a. In place of the conical vial as the reaction vessel, use a 100-mL round-bottom flask connected to a water-cooled reflux condenser that is fitted with a calcium chloride drying tube.

b. The required reagent and solvent quantities are given in the following table:

Compound	MW	Amount	mol	mp (°C)	bp (°C)
Anthracene	178.24	4.0 g	0.2	216	
Maleic anhydride	98.06	2.0 g	0.2	60	
Xylene	106.16	50 mL			137–140
Ethyl acetate	88.11	30 mL			77

c. Heat the reaction mixture at **vigorous** reflux for **2 h.**

d. Recrystallize the crude product from ethyl acetate.

9,10-Dihydroanthracene-9,10-α,β-succinic Acid Anhydride: Preparation Using a Monomode Microwave Apparatus

Experiment 15-1*

M

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1 h.

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
Anthracene	178.24	178 mg	1.00	216	
Maleic Anhydride	98.06	98 mg	1.00	60	
Toluene		2 mL			111
9,10-Dihydroanthracene-9,10-α,β-succinic Acid Anhydride	276			261–262	

Reagents and Equipment. This experiment is designed for use in the CEM Discover and Biotage Initator microwave units.

In a 10.0-mL glass microwave reaction vessel containing a magnetic stir bar, place 178 mg (1.00 mmol) of anthracene, 98 mg (1.00 mmol) of maleic anhydride, and 2 mL of toluene. Immediately cap the vessel with a microwave pressure cap.

CAUTION: Since the reaction requires heating toluene (solvent for the reaction) to above its boiling point in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential.*

Reaction Conditions. Place the reaction vessel in the microwave cavity and, depending on the equipment used, position the pressure device on top. Program the microwave unit to heat the reaction mixture at maximum power to 180 °C and hold at this temperature for 10 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the tube from the microwave unit. Allow the resulting solution to cool to room temperature and then place it in an ice bath for 10 min to complete the crystallization of the product.

*This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

Isolation of Product. The remainder of the procedure is identical to Experiment 15 (pages 271–272).

Experiment 15-2*

M

9,10-Dihydroanthracene-9,10-α,β-succinic Acid Anhydride: Preparation Using a Multimode Microwave Apparatus

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1 h.

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
Anthracene	178.24	214 mg	1.20	216	
Maleic Anhydride	98.06	118 mg	1.20	60	
Toluene		5 mL			111
9,10-Dihydroanthracene-9,10-α,β-succinic Acid Anhydride	276			261–262	

Reagents and Equipment. This experiment is designed for use in the CEM MARS, Milestone START, and Anton Paar Synthos 3000 microwave units. When using the Anton Paar Synthos 3000 unit with the 24-position silicon carbide plate rotor containing glass vials, the reagent and solvent quantities cited in the monomode procedure should be used in conjunction with the reaction conditions here in the multimode procedure.

In a microwave reaction vessel containing a magnetic stir bar, add 214 mg (1.20 mmol) of anthracene, 118 mg (1.20 mmol) of maleic anhydride, and 5 mL of toluene. Immediately cap the vessel with the microwave pressure cap and adjust the tightness to the manufacturer-specified level. Place the sealed vessel into its outer protective jacket.

CAUTION: Since the reaction requires heating toluene (solvent for the reaction) to above its boiling point in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential*.

Reaction Conditions. Insert the loaded vessels into the reaction carousel ensuring they are evenly spaced and then place the carousel into the microwave cavity. If provided by the manufacturer, connect a temperature probe to the control vessel. Program the microwave unit to heat the reaction vessels to 180 °C using maximum power and hold at this temperature for 10 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the carousel from the microwave unit. Allow the resulting solution to cool to room temperature and then place it in an ice bath for 10 min to complete the crystallization of the product.

Isolation of Product. The remainder of the procedure is identical to Experiment 15 (pages 271–272).

---

*This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

QUESTIONS

6-102. Given the data tabulated below for the rate of reaction of maleic anhydride with a series of substituted 1,3-butadienes, offer a reasonable explanation to account for the trend in the rates.

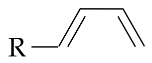

R	*k* (rel) at 25 °C
—H	1
—CH_3	4.2
—$C(CH_3)_3$	<0.05

6-103. Predict the structure of the product formed in the following reactions.

(a)

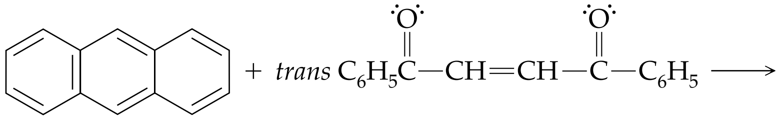

$$+\ \textit{trans}\ C_6H_5\overset{\overset{\displaystyle \ddot{O}}{\|}}{C}-CH{=}CH-\overset{\overset{\displaystyle \ddot{O}}{\|}}{C}-C_6H_5 \longrightarrow$$

(b)

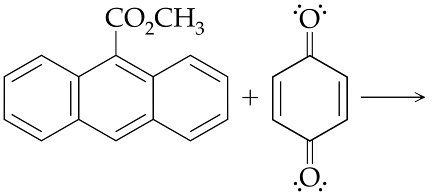

6-104. Offer an explanation of why anthracene preferentially forms a Diels–Alder adduct at the 9,10 positions.

6-105. Experiment [31] demonstrates the monobromination of anthracene to yield one specific monosubstituted product. Other conditions may be used to form other monobrominated anthracenes. Draw the structures of, and name, the possible monobromo-substituted anthracenes that could be prepared in the laboratory.

6-106. There are four reasonable resonance structures for anthracene. Draw them.

6-107. A large number of polycyclic benzenoid aromatic hydrocarbons are known. One of these, benz[*a*]pyrene, is a powerful carcinogen found in tobacco smoke. From the literature, locate and then draw the structure of this hydrocarbon. Can you suggest other sources where this material might be expected to be present?

6-108. Anthracene undergoes a *suprafacial* $[_{\pi}4_{s} + {}_{\pi}4_{s}]$ cycloaddition reaction at the 9,10 positions when subjected to irradiation to form a cyclic dimer. *Suprafacial* is a term used to describe in detail that the addition of the dienophile to the 1,3-π-system has occurred all from the same side (or face) of this planar structure. It comes from the Latin meaning above, or the dorsal side, and it is symbolized by the subscript "s" in the terminology: $[_{\pi}4_{s} + {}_{\pi}4_{s}]$. This is a theoretically forbidden thermal pericyclic reaction, but it can occur under photochemical conditions. Draw the structure of this product.

BIBLIOGRAPHY

For a review of diastereoselectivity in the reaction see

Coxon, J. M. et al., *Diastereofacial Selectivity in the Diels–Alder Reaction* in *Advances in Detailed Reaction Mechanisms* **1994,** *3,* 131.

Waldmann, H. *Synthesis* **1994,** 535.

Grignard Reaction with a Ketone: Triphenylmethanol

EXPERIMENT 16

Common name: triphenylmethanol
CA number: [76-84-6]
CA name as indexed: benzenemethanol, α,α-diphenyl-

Purpose. The techniques required to prepare Grignard reagents are developed. The reaction of these reagents with ketones to form *tertiary* alcohols is investigated. You will gain experience, first hand, working with these highly air- and moisture-sensitive materials. And you will observe the formation of this famous reagent, in which magnesium metal is transformed (heterogeneous conditions) into organometallic salts of enormous value to the synthetic chemist.

www This exercise is the first of a number of experiments in this chapter in which the Grignard reaction is studied at the microscale level. Even if you do not actually do more than one of them, it is well worth the time to read through the other Experiments [17], [21], and [4_{adv}] to get a sense of the breadth of the applications of the reaction.

Prior Reading

Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
  Drying of the Wet Organic Layer (pp. 80–83)
Technique 5: Crystallization
  Craig Tube Crystallizations (pp. 89–91)
Technique 6A: Chromatography
  Thin-Layer Chromatography (pp. 97–99)

REACTION

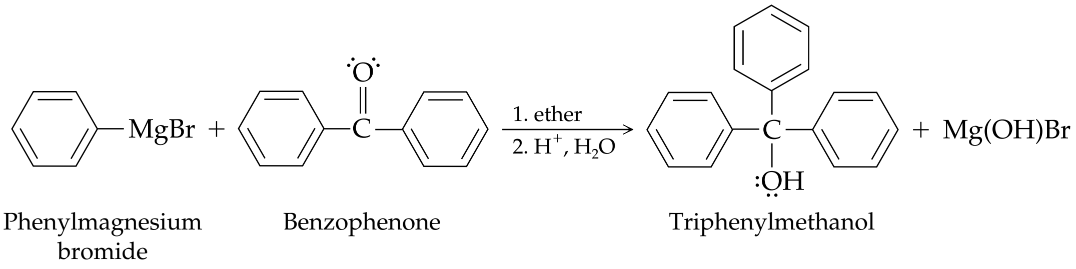

François Auguste Victor Grignard (1871–1935) Born in Cherbourg, Grignard was professor of Chemistry at the Universities of Lyons and Nancy. After studying for one year with Bouveault, Grignard became a graduate student of Phillippe Antoine Barbier, a professor at the University of Lyons. Barbier, who was working in the area of terpene chemistry, had found that magnesium could be used in place of zinc in the reaction of methyl iodide with an unsaturated ketone (methylheptenone) to yield the corresponding tertiary alcohol. This route was much preferred since the zinc reagents were difficult to work with because they were pyrophoric (spontaneously flammable in air). The use of magnesium in the formation of tertiary alcohols was reported in 1899. Barbier suggested to Grignard that it might be interesting to further investigate the reaction of magnesium with alkyl halides. This study was to form the basis of Grignard's doctoral dissertation. Grignard discovered that treatment of alkyl iodides with magnesium in *diethyl ether* produced an alkylmagnesium iodide by a spontaneous reaction at ambient temperatures. His initial results, reported in 1900, were followed by seven papers the following year. His doctoral thesis on organomagnesium compounds and their application to synthetic organic chemistry was presented in 1901, when Grignard was 30 years old.

Grignard continued this work, having recognized the enormous potential of the alkyl magnesium halides in organic synthesis. These species are now known as *Grignard reagents* and when the reagent is used in synthesis, the reaction is called a *Grignard reaction*. These reagents have found great utility in the preparation of many kinds of organic compounds, including alcohols, ketones, esters, and carboxylic acids. As mentioned above, these reagents contain a carbon–metal bond, and therefore they are classed in a large group of substances called *organometallic* compounds.

For this work, Grignard received the Nobel Prize in 1912. In 1919 Grignard returned to Lyons where he succeeded Barbier as chairman of the Department. By the end of his life, the scientific literature contained over 6000 papers dealing with Grignard reagents and their application.

Grignard also did extensive work in the areas of the terpenes, quantitative ozonolysis of alkenes, aldol reactions, catalytic hydrogenation, and dehydrogenation and cracking of hydrocarbons.[17]

DISCUSSION

Grignard reagents possess significant nucleophilic character because of the highly polarized carbon–metal bond that results in considerable carbanionic character at carbon. Grignard discovered that these reactive materials readily attack the electrophilic carbon of a carbonyl group. It is this direct attack on carbon by a carbon nucleophile, resulting in carbon–carbon bond formation, that makes these such important reactions. Furthermore, as the carbonyl is the most ubiquitous functionality in all of organic chemistry, Grignard reagents have found great utility and widespread use in organic synthesis.

The formation of the organomagnesium halide (Grignard reagent) involves a heterogeneous reaction between magnesium metal and an alkyl, alkenyl, or aryl halide in ether solution. The solvent may be any one of a number of ethers, but diethyl ether and tetrahydrofuran are by far the most popular.

$$\mathrm{R(Ar)}\text{—}\mathrm{X} + \mathrm{Mg} \xrightarrow{\text{ether}} \mathrm{R(Ar)}\text{—}\mathrm{Mg}\text{—}\mathrm{X}$$

R = alkyl, alkenyl
and Ar = aryl

The reaction between an alkyl, alkenyl, or aryl halide and magnesium takes place on the surface of the metal and is an example of a *heterogeneous* (across two phases) reaction. The reactivity of the alkyl halides is in the order $\mathrm{Cl} < \mathrm{Br} < \mathrm{I}$; fluorides do not generally react. Substituted alkyl halides react in the order $1° > 2° > 3°$ alkenyl and aromatic halides also form Grignard reagents to varying degrees.

It is important to understand the role of the ether solvent in the formation of the Grignard reagents. The reaction at the surface of the metal is essentially an oxidation–reduction reaction. The metal is partially oxidized to the greater than $1+$ state and the organohalide is reduced to a halide ion and a highly polarized carbon–metal bond with the magnesium. The overall reaction can be viewed as forming the species, $\mathrm{R}\overset{\delta-\;\delta+}{\text{—}}\mathrm{Mg}^{+}\mathrm{X}^{-}$ as the Grignard reagent. This highly polarized material is insoluble in most nonpolar organic solvents. The reaction will proceed at the surface of the metal until a layer of the insoluble organometallic reagent has formed. At this point, the surface reaction with the magnesium will immediately cease. If protic solvents are used, they would instantly react with the highly basic Grignard reagent, R—MgX, to form the corresponding hydrocarbon, R—H. Thus, the use of either nonpolar or protic solvents does not lead to successful Grignard reagent formation. Why, then, is ether, a relatively nonpolar solvent, essential to the preparation of Grignard reagents?

---

[17] See Gordon, N. E. *J. Chem. Educ.* **1930,** *7,* 1487; Rheineoldt, H. *J. Chem. Educ.* **1950,** *27,* 476; Kauffman, G. B. *J. Chem. Educ.* **1990,** *67,* 569; Gilman H. *J. Am. Chem. Soc.* (Proc.) **1937,** *59,* 17; Gibson, C. S.; Pope, W. J. *J. Chem. Soc.* **1937,** 171; *Dictionary of Scientific Biography,* C. C. Gillespie, Ed., Scribner's: New York, 1972, Vol. V, p. 540.

The magnesium is essentially divalent, and electron deficient, when it reacts with the halide to form the RMgX species. A full octet around the metal atom requires two additional pairs of electrons. It is the energy gained by filling this octet that drives the coordination of the magnesium with two molecules of the ether solvent. This association in turn dramatically increases the solubility of the Grignard reagent in the relatively nonpolar ether solvent, and thus promotes further Grignard reagent formation.

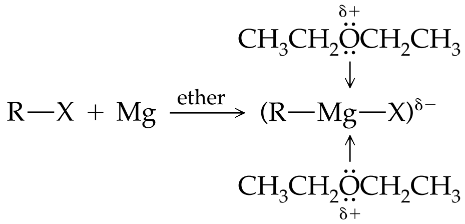

$$\mathrm{R{-}X} + \mathrm{Mg} \xrightarrow{\text{ether}} \begin{array}{c} \overset{\delta+}{\mathrm{CH_3CH_2\ddot{O}CH_2CH_3}} \\ \downarrow \\ (\mathrm{R{-}Mg{-}X})^{\delta-} \\ \uparrow \\ \underset{\delta+}{\mathrm{CH_3CH_2\ddot{O}CH_2CH_3}} \end{array}$$

This interaction of RMgX with ether solvent also may be described as a Lewis acid–base interaction in which the coordinating solvent molecules are usually not written. When a Grignard reagent is described, it is important to remember that this vital solvation is always taking place.

The reactions of Grignard reagents with different types of carbonyl groups yield a number of important functional groups. For example, reaction with formaldehyde yields 1° alcohols; with higher aldehydes, 2° alcohols; with ketones, 3° alcohols; with esters, 3° alcohols; with acyl halides, ketones; with *N,N*-dialkylformamides, aldehydes; and with carbon dioxide, carboxylic acids.

In this experiment you will study the addition of the aryl Grignard reagent (phenylmagnesium bromide) to a diaryl ketone (benzophenone) to yield the corresponding tertiary (3°) alcohol. Because it is possible to vary both the structure of the Grignard reagent and the ketone, a wide variety of 3° alcohols may be obtained by this synthetic route.

The mechanism, as discussed above, can be thought of as involving rapid nucleophilic attack by the Grignard reagent at the carbon of the carbonyl group. Hydrolysis of the resulting alkoxide ion intermediate with dilute acid yields the desired alcohol. The reaction sequence is outlined here:

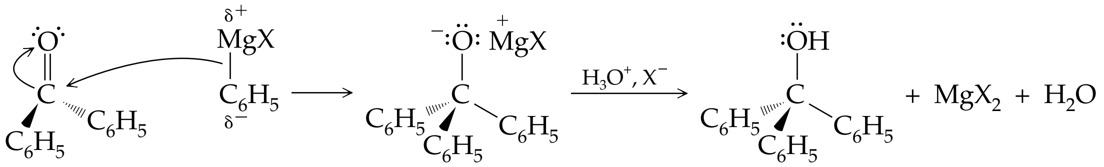

By using Grignard reagents, it is theoretically possible to synthesize a very large number of alcohols. Indeed, there is often more than one synthetic pathway open to a desired product. The choice of route is generally dictated by the availability of starting materials and the associated costs of these compounds.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: Two laboratory periods when starting with bromobenzene. If using commercial phenylmagnesium bromide, one laboratory period (experiment starting on p. 280). **(The Benzophenone Reagent)**

CAUTION: Ether is a flammable liquid. All flames must be extinguished during the time of this experiment.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Bromobenzene	157.02	76 μL	0.72		156	1.50	1.5597
Diethyl ether	74.12	1.3 mL			34.5	0.73	
Magnesium	24.3	18 mg	0.74				
Iodine	253.8	1 crystal					
Benzophenone	182.21	105 mg	0.58	48			

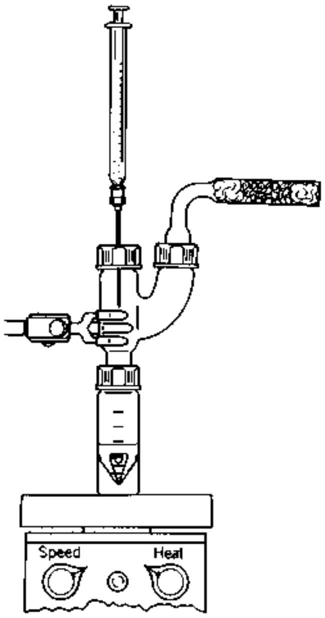

Step I:
BrC_6H_5, 76 μL
+ Mg, 18 mg
+ I_2, 1 crystal
+ $(CH_3CH_2)_2O$, 700 μL
Step II:
$(C_6H_5)_2CO$, 105 mg
+ $(CH_3CH_2)_2O$, 600 μL

Reagents and Equipment

Preparation of Phenylmagnesium Bromide

NOTE. All the glassware used in the preparation of the Grignard reagent should be cleaned and dried in an oven at 110 °C for at least 30 min. After removal from the drying oven, the hot glassware should be placed in a desiccator and cooled before being assembled. Flame drying of the apparatus with a microburner annealing flame (high gas mixture) is an alternative, if ovens are not available. This latter method preferably should be carried out prior to assembly and, as with the oven-dried equipment, the hot glassware should be placed in a desiccator to cool. Care must be taken when flaming the reaction vial, because thermal shock can easily crack this heavy-walled vessel. If the flame-drying procedure is performed on the assembled apparatus, care also must be taken not to overheat the O-rings and plastic Capseals.

In a 3.0-mL conical vial containing a magnetic spin vane and equipped with a Claisen head fitted with a calcium chloride drying tube and a rubber septum, weigh and place 18 mg (0.74 mmol) of polished magnesium ribbon, a small crystal of iodine, and 100 μL of anhydrous ether (using an automatic delivery pipet in the **hood**)(➡).

HOOD

NOTE. Scrape a 2- to 3-in. piece of magnesium ribbon clean of surface oxide (MgO) coating and then cut it into 1-mm long sections.

Alternate Procedure: Place the (polished) magnesium metal and an iodine crystal in the reaction vial and quickly assemble the apparatus. Warm the mixture (using a microburner or hot plate) **gently** until evidence of purple iodine vapor is observed.

GENTLY

NOTE. If a microburner is used in this step, all flames must be extinguished before proceeding.

Now add 100 μL of anhydrous ether using a 1.0-mL syringe inserted through the septum.

In a dry, screw-capped vial, prepare a solution of 76 μL (113 mg, 0.72 mmol) of bromobenzene in 400 μL of anhydrous diethyl ether. *Automatic delivery* pipets are used to transfer these reagents in the **hood.** Recap the vial until just before injecting its contents into the reaction vial.

HOOD

Draw the bromobenzene–ether solution into a 1.0-mL syringe and insert the syringe through the rubber septum on the Claisen head. Place an additional 300 μL of anhydrous diethyl ether rinse in the empty vial (which contains traces of bromobenzene), cap it, and set it aside for later use. With slow stirring, add 6–8 drops of the bromobenzene solution to induce the initial formation of Grignard reagent. The evolution of tiny bubbles from the surface of the magnesium (the heat of reaction is vaporizing the low-boiling ether solvent) is evidence of successful reaction initiation.

Once the reaction gives evidence of initiation, add the remainder of the bromobenzene dropwise *slowly* over a 3- to 5-min period. *(In macroscale reactions it is extremely important to make sure that initiation of the reaction has occurred prior to adding large quantities of the organohalide, as Grignard reactions often go through an induction period before starting up. If significant quantities of the halide are present when the reaction commences, the sudden and rapid rate of reaction can produce a very rapid evolution of heat, and the reaction may well erupt out of control.) Warm the reactants gently to maintain gentle reflux.*

After adding the bromobenzene, draw the ether rinse from the capped vial into the syringe and also add it to the reaction vial through the septum in a single portion.

Heat the resulting heterogeneous reaction mixture gently with stirring for 15 minutes.

CAUTION: *Do not overheat!* This overheating will cause loss of ether solvent and promote formation of byproducts. Small fragments of magnesium may remain at the end of the reaction. Maintain no more than a gentle reflux at all times. If solvent volume decreases rapidly, check for leaks around the Capseals and add additional anhydrous ether through the septum to make up for the lost volume of ether.

Cool the gray-brown mixture of the Grignard reagent, phenylmagnesium bromide, to room temperature.

The Benzophenone Reagent. Prepare a solution of 105 mg (0.58 mmol) of benzophenone in 300 μL of anhydrous diethyl ether in a dry vial with a cap. *The ether is measured using a graduated 1-mL syringe and is dispensed in the* ***hood.***

HOOD

Draw the solution immediately into a 1.0-mL syringe, and then insert the syringe needle through the rubber septum on the Claisen head. Place an additional 300 μL of the anhydrous diethyl ether in the empty vial, cap it, and set it aside for later use.

Reaction Conditions. *Carefully*, with stirring, add the benzophenone solution to the Grignard reagent (1.2 equiv) over a period of approximately 30 s or at a rate that maintains the temperature of the ether solvent at a no more than gentle reflux.

Upon completion of this addition, add the rinse from the capped vial, in like manner, in a single portion.

Stir the reaction mixture for 2–3 min and then allow it to cool to room temperature. Remove the reaction vial from the Claisen head and cap it.

During this cooling period the reaction mixture generally solidifies. *Once the reaction vial is detached from the Claisen head, it is recommended that the vial be placed in a 10-mL beaker to prevent loss of product by accidental tipping.*

NOTE. If the laboratory is done in two periods, you should stop either at this point, or following the hydrolysis sequence described in the next step.

Isolation of Product. Hydrolyze the magnesium alkoxide salt by the *careful*, dropwise addition of 3 M HCl from a Pasteur pipet, while at the same time using a small stirring rod to break up the solid residue. Continue the addition until the aqueous phase tests acidic with litmus paper. A two-layer reaction mixture forms (ether–water) as the solid gradually dissolves.

CAUTION: The addition of the acid may be accompanied by the evolution of heat and some frothing of the reaction mixture. An ice bath should be handy to cool the solution, if necessary. Additional ether may be added, if required, to maintain the volume of the organic phase. Check the acidity of the mixture periodically. The total reaction mixture must be acidic; both insufficient or excess amounts of hydrochloric acid will result in a decreased yield of product during the subsequent workup.

Now remove the magnetic spin vane with forceps and set it aside to be rinsed with an ether wash. Cap the vial tightly, shake, carefully vent, and allow the layers to separate.

Using a Pasteur filter pipet, transfer the lower aqueous layer to a clean 5.0-mL conical vial.

NOTE. Save the ether layer—it should contain your product.

Wash the acidic aqueous layer with three 0.5-mL portions of diethyl ether (calibrated Pasteur pipet). Rinse the spin vane with the first portion of ether as it is added to the vial. Cap the vial, shake (or use a Vortex mixer, if available), vent carefully, and allow the layers to separate. After each extraction, combine the organic phase with the ether solution saved above. The bottom (aqueous) layer is set aside in a 10-mL Erlenmeyer flask until the experiment is completed.

Now extract the combined ether layers with 0.5 mL of cold water to remove any acidic residue. Combine the aqueous rinse with the previously extracted and stored aqueous layers in a 10-mL Erlenmeyer flask. Dry the ether solution (capped) over 250–300 mg of anhydrous granular sodium sulfate for approximately 10 min. Stir the drying agent intermittently with a glass rod or swirl the flask. If large clumps of sodium sulfate begin to develop and the solution remains cloudy, it may be necessary to transfer the ether extracts to another vial for a second treatment with the drying agent, or you may be able to simply add more Na_2SO_4 to the original vial. The ether solution should be *clear* following treatment with the anhydrous sodium sulfate. Make all these transfers with Pasteur filter pipets.

Transfer the dried ether solution to a *previously tared* Craig tube containing a boiling stone. Carry out the transfer in 0.5-mL portions, concentrating each ether aliquot by warming the vial in a sand bath in the **hood** between the transfers. Rinse the vial and drying agent with an additional 0.5 mL of ether, add the rinse to the Craig tube, and, finally, concentrate the solution to dryness to yield the product residue.

HOOD

Determine the weight of the crude triphenylmethanol product.

Purification and Characterization. The major impurity usually present in the triphenylmethanol is biphenyl, which is formed by a coupling reaction,

$$2RX + Mg \xrightarrow{\text{ether}} R{-}R + MgX_2$$

The purity of your crude product may be determined using *thin-layer chromatography.*

TLC CONDITIONS. Use Eastman Kodak fluorescent silica gel sheets (1 × 4 cm). Develop the plates with methylene chloride and visualize the spots by UV light. Reference R_f *values for triphenylmethanol and biphenyl are about 0.6 and 0.9, respectively.*

The coupled byproduct can be separated from the desired alcohol by taking advantage of the difference in solubilities of the hydrocarbon and the alcohol in ligroin. Ligroin, a nonpolar alkyl solvent, readily dissolves the nonpolar biphenyl, whereas the polar tertiary alcohol is much less soluble.

Add 0.5 mL of cold ligroin to the crude product contained in a Craig tube and scrape and agitate the solid material into a suspension with a small stirring rod. Swirl and stir the solid product with the solvent for several minutes.

Recover the solid triphenylmethanol using the Craig tube in the usual manner. *Save* the ligroin solution that will contain any biphenyl by transferring it, using a Pasteur filter pipet, to a tared 10-mL Erlenmeyer flask.

HOOD

Repeat the above extraction with a second 0.5-mL portion of ligroin, again stirring the solid suspension and combining the recovered ligroin solution with that saved above. Place the Erlenmeyer flask (cover the mouth with filter paper held by a rubber band) in the **hood** overnight, or warm it in a sand bath to allow the ligroin to evaporate. Estimate the amount of biphenyl (and any other impurities) produced in the reaction.

Heat the Craig tube containing the solid triphenylmethanol in a 100 °C oven for 5 min, and then place the crystals on a clay plate to complete the drying process. A further check of the product purity should be carried out by TLC, as described above. Comparisons of the two, pure and impure product, will validate that upon trituration followed by recrystallization, TLC analysis offers the experimentalist sufficient evidence of product purity.

Weigh the purified triphenylmethanol product and calculate the percent yield. Determine the melting point of the material and compare your result to that recorded in the literature. If desired, the product may be purified further by recrystallization from isopropanol using the Craig tube.

Characterization of the triphenylmethanol is best done by obtaining the IR and NMR spectra and comparing the spectral data to that of an authentic sample or to the published spectra in *The Aldrich Library of IR Spectra* and *The Aldrich Library of NMR Spectra,* respectively.

Triphenylmethanol (0.1 g/L in methanol) and biphenyl (0.05 g/L in methanol) also have markedly different UV spectra. This electronic absorption data can further help to establish the identity of the products formed in this Grignard reaction.

UV Spectral Data

Triphenylmethanol	λ_{max} 240 nm (log ε_{max} 3.16, dioxane)
	λ_{max} 253 nm (log ε_{max} 3.26, dioxane)
	λ_{max} 260 nm (log ε_{max} 3.28, dioxane)
Biphenyl	λ_{max} 247 nm (log ε_{max} 4.24, ethanol)

QUESTIONS

6-109. Predict the product formed in each of the following reactions and give each reactant and product a suitable name:

(a) $CH_3CH_2MgBr + CH_2O \xrightarrow[2.\ H^+]{1.\ ether}$

(b) $p\text{-}CH_3C_6H_4MgBr + CH_3CH_2CHO \xrightarrow[2.\ H^+]{1.\ ether}$

(c) $C_6H_5MgBr + D_2O \xrightarrow{ether}$

(d)
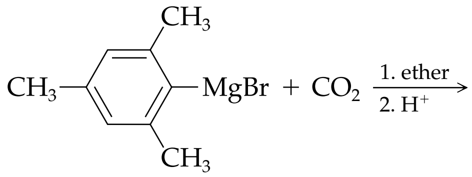

6-110. Using the Grignard reaction, carry out the following transformations. Any necessary organic or inorganic reagents may be used. Name all reactants and products.

(a)
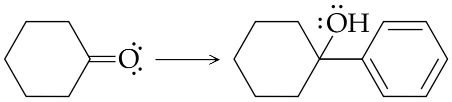

(b)
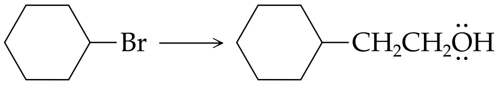

(c)
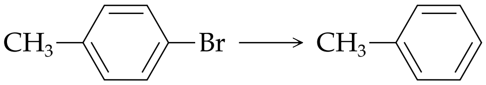

6-111. Outline a synthetic reaction scheme for the preparation of triphenylmethanol from
(a) Methyl benzoate
(b) Diethyl carbonate

6-112. In the experiment, ligroin may be used as a solvent for the separation of the product from biphenyl.
(a) What is ligroin?
(b) Can you suggest an alternative solvent that might be used in this step?

6-113. Give the reaction scheme, showing the products formed (before hydrolysis), when one equivalent of ethylmagnesium bromide is treated with one equivalent of 5-hydroxy-2-pentanone. Does addition of two equivalents of the Grignard reagent to this ketone yield a different product(s)? If so, give the structure(s).

BIBLIOGRAPHY

The alternative method of preparing the Grignard reagent is adapted from

Eckert, T. S. *J. Chem. Educ.* **1987,** *64,* 179.

General references on Grignard reagents:

Coates, G. E.; Green, M. L. H.; Wade, K. *Organometallic Compounds,* 3rd ed.; Methuen: London; Vol. II, 1968.
Grignard, V. *Compt. Rend.* **1900,** *130*, 1322.
Huryn, D. M. *Comp. Org. Syn.* **1991,** *1,* 49.
Lai, Y. H. *Synthesis* **1981,** 585.
Raston, C. L.; Salem, G. in *The Chemistry of the Metal–Carbon Bond;* Hartley, F. R., Ed.; Wiley: New York, 1987, Vol. 4, p. 159.
Shirley, D. A. *Org. React.* **1954,** *8,* 28.

A synthesis of triphenylmethanol (triphenylcarbinol) is reported in *Organic Syntheses:*

Bachmann, W. E.; Hetzner, H. P. *Organic Syntheses*; Wiley: New York, 1955; Collect. Vol. III, p. 839.

The preparation of a series of tertiary alcohols by the addition of Grignard reagents to diethyl carbonate is reported in *Organic Syntheses:*

Moyer, W. W.; Marvel, C. S. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, pp. 602–603. Also see Braun, M; Gräf, S; Herzog, S. *Organic Syntheses* **1995,** *72,* 32.

EXPERIMENT 17

Grignard Reaction with an Aldehyde: 4-Methyl-3-heptanol

Common name: 4-methyl-3-heptanol
CA number: [14979-39-6]
CA name as indexed: 3-heptanol, 4-methyl-

Purpose. You will carry out a classic method for the synthesis of *secondary* alcohols: the addition of a Grignard reagent to an aldehyde (other than formaldehyde).[18]

Prior Reading

Technique 4: Solvent Extraction
    Liquid–Liquid Extraction (p. 72)
    Drying of the Wet Organic Layer (pp. 80–83)
Technique 6: Chromatography
    Column Chromatography (pp. 92–95)
Technique 1: Gas Chromatography (pp. 55–61)

REACTION

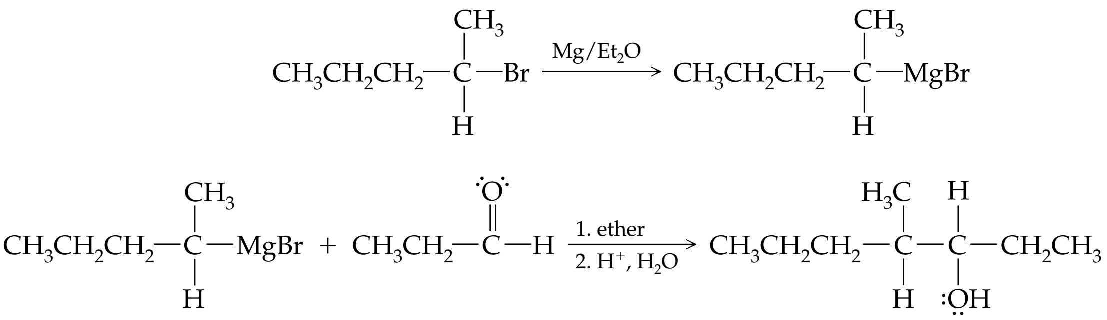

$$CH_3CH_2CH_2-\overset{CH_3}{\underset{H}{C}}-Br \xrightarrow{Mg/Et_2O} CH_3CH_2CH_2-\overset{CH_3}{\underset{H}{C}}-MgBr$$

$$CH_3CH_2CH_2-\overset{CH_3}{\underset{H}{C}}-MgBr + CH_3CH_2-\overset{O}{\overset{\|}{C}}-H \xrightarrow[2.\ H^+,\ H_2O]{1.\ ether} CH_3CH_2CH_2-\overset{H_3C}{\underset{H}{C}}-\overset{H}{\underset{OH}{C}}-CH_2CH_3$$

1-Methylbutylmagnesium bromide   Propanal   4-Methyl-3-heptanol

NOTE. See Experiment [16] for a biography of François Auguste Victor Grignard, Nobel Laureate, who discovered and developed the Grignard reagents. This experiment also contains further details about the mechanism and use of these reagents that have had such a powerful influence on synthetic organic chemistry.

DISCUSSION

In this experiment, the addition of a nucleophilic Grignard reagent (1-methylbutylmagnesium bromide), to the electrophilic carbonyl carbon of an aldehyde (propanal) is described. The product obtained is a 2° alcohol, 4-methyl-3-heptanol. Because it is possible to vary the structure of both the Grignard reagent and the aldehyde, a wide variety of 2° alcohols can be prepared by this route. Primary alcohols result when formaldehyde is used as the aldehyde. Secondary alcohols may also be obtained with these reagents when ethyl formate, an ester, acts as the electrophile. This latter reaction, however, requires two molar equivalents of the Grignard reagent. The mechanism for the reaction of an aldehyde with a Grignard reagent follows.

---

[18] For references relating to the preparation of Grignard reagents, see Experiment [16].

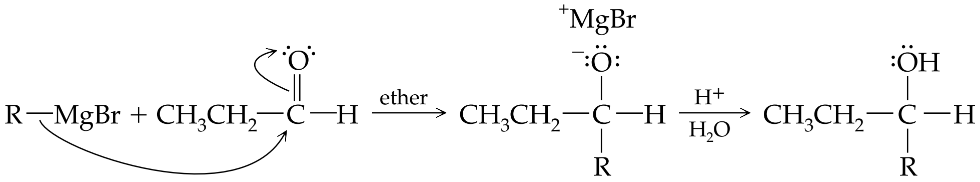

$$R{-}MgBr + CH_3CH_2{-}\overset{\overset{\ddot{O}}{\|}}{C}{-}H \xrightarrow{\text{ether}} CH_3CH_2{-}\overset{\overset{^{+}MgBr}{^{-}:\ddot{O}:}}{\underset{R}{C}}{-}H \xrightarrow[H_2O]{H^+} CH_3CH_2{-}\overset{:\ddot{O}H}{\underset{R}{C}}{-}H$$

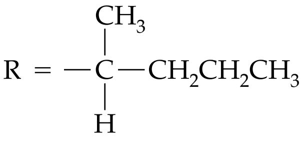

$$R = -\overset{CH_3}{\underset{H}{C}}-CH_2CH_2CH_3$$

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 3.0 h. The chromatographic separation requires approximately an additional 15 min per student.

CAUTION: Ether is a flammable liquid. All flames must be extinguished during this experiment.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
2-Bromopentane	151.05	125 μL	1.0	117	1.2	1.4413
Diethyl ether	74.12	700 μL		34.5		
Magnesium	24.31	36 mg	1.48			
Iodine	253.81	1 crystal				
Propanal	58.08	50 μL	0.69	49	0.81	1.3636

Reagents and Equipment

Preparation of 1-Methylbutylmagnesium Bromide

NOTE. All the glassware used in the preparation of the Grignard reagent should be cleaned and dried in an oven at 110 °C for at least 30 min. After removal from the drying oven, the hot glassware should be placed in a desiccator and cooled before being assembled. Flame drying of the apparatus with a microburner annealing flame (high gas mixture) is an alternative, if ovens are not available and if ether is not in use. This latter method preferably should be carried out prior to assembly, and as with the oven-dried equipment, the hot glassware should be placed in a desiccator to cool. Care must be taken when flaming the reaction vial, as thermal shock can easily crack this heavy-walled vessel. If the flame-drying procedure is performed on the assembled apparatus, care also must be taken not to overheat the O-rings and plastic Capseals.

NOTE. Scrape a 2- to 3-in. piece of magnesium ribbon clean of oxide coating and cut it into sections 1 mm in length. This freshly cut material should be handled only with forceps.

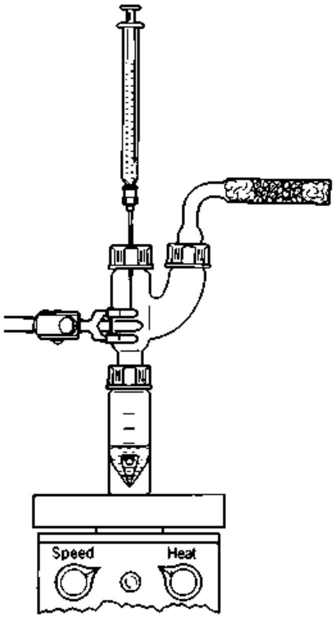

Step I:
Mg, 36 mg + I_2, 1 crystal
+ $(CH_3CH_2)_2O$, 500 μL
+ $CH_3CHBr(CH_2)_2CH_3$, 125 μL
Step II:
CH_3CH_2CHO, 50 μL
+ $(CH_3CH_2)_2O$, 200 μL

Prepare a 3.0-mL conical vial containing a magnetic spin vane and equipped with a Claisen head fitted with a calcium chloride drying tube and a rubber septum. Weigh and place 36 mg (1.5 mmol) of magnesium in the vial, and then add a small crystal of iodine, followed by 100 μL of anhydrous ether (⬅).

Alternative Procedure: Place the magnesium metal and the iodine crystal in the vial and assemble the apparatus. *Gently* warm the mixture (microburner or hot plate) until evidence of purple vapor from the iodine is seen.

NOTE. If a microburner is used in this step, all flames must be extinguished before proceeding.

Now add the 100 μL of anhydrous ether, using a 1.0-mL syringe inserted through the septum.

Prepare a solution of 125 μL (153 mg, 1.0 mmol) of 2-bromopentane in 300 μL of anhydrous diethyl ether in a dry, screw-capped vial. *Use an automatic delivery pipet to deliver these reagents.*

After the assembly has cooled to room temperature, draw the 2-bromopentane solution into a 1.0-mL syringe and then insert the syringe needle through the rubber septum on the Claisen head. Place an additional 100 μL of diethyl ether in the empty vial, cap it, and set it aside for later use.

While stirring the heterogeneous mixture, add 6–8 drops of the 2-bromopentane–ether solution to initiate the formation of the Grignard reagent. The evolution of tiny bubbles from the surface of the magnesium is evidence of reaction.

SLOWLY When the reaction has started, **slowly** add the remainder of the 2-bromopentane–ether solution dropwise over a 3- to 5-min period. Warm the reactants slightly. Upon completion of this addition, draw the rinse in the capped vial into the syringe and add it through the septum in a single portion to the reaction vial. Gently warm the resulting solution for 15 min.

CAUTION: *Do not overheat.* Overheating will cause loss of ether solvent. Small fragments of magnesium may remain at the end of the addition of the alkyl halide.

Cool the gray-colored solution of Grignard reagent to room temperature.

The Propanal Reagent. Prepare a solution of the aldehyde by weighing 50 μL (40 mg, 0.7 mmol) of propanal into a tared, oven-dried, capped vial followed by the addition of 100 μL of anhydrous diethyl ether. The propanal is the limiting reagent and therefore an accurate weight should be recorded for the yield calculations. *Dispense the aldehyde and diethyl ether by automatic delivery pipets in the* **hood.** HOOD

Immediately draw the aldehyde solution into a 1.0-mL syringe and insert the syringe needle through the rubber septum on the Claisen head.

Place an additional 100 μL of the anhydrous diethyl ether in the empty vial, cap it, and set it aside for later use.

Reaction Conditions. Now add the propanal solution *carefully*, with stirring, to the Grignard reagent over a period of about 30 s at such a rate as to keep the ether solvent at a steady reflux.

Following this addition, add the rinse in the capped vial in one portion in a similar manner.

Stir the reaction mixture for 5 min and then allow it to cool to room temperature. Remove the conical vial and cap. It is recommended that the vial be placed in a beaker to prevent tipping and loss of product.

NOTE. If the laboratory is done in two periods, you should stop at this point (recap the vial for storage) or after the hydrolysis sequence in the next step.

Isolation of Product. Hydrolyze the magnesium alkoxide salt by the *careful*, dropwise addition of 2–3 drops of water from a Pasteur pipet. Stir the resulting mixture for 5 min. A two-phase (ether–water) reaction mixture develops as the magnesium salt is hydrolyzed.

CAUTION: The addition of water causes the evolution of heat. An ice bath should be handy to cool the solution if it begins rapid reflux.

Now add 2–3 drops of 3 M HCl. Remove the vial, cap it, and allow it to stand at room temperature for 5 min. Test the aqueous layer with litmus paper. The solution should be slightly acidic. *Too much or too little aqueous HCl will cause problems in the subsequent workup*.

Remove the magnetic spin vane with forceps and set it aside to be rinsed with an ether wash. Cap the vial tightly, shake (or use a Vortex mixer), vent *carefully,* and allow the layers to separate.

Using a Pasteur filter pipet, transfer the aqueous (lower) layer to a clean 5.0-mL conical vial. *Save* the ether layer since it contains the crude reaction product.

Now wash the aqueous layer, previously transferred to the 5.0-mL vial, with three 0.5-mL portions of diethyl ether. Rinse the magnetic spin vane with the first portion as it is added to the vial. Upon addition of each portion of ether (using a calibrated Pasteur pipet), cap the vial, shake (or use a Vortex mixer), vent carefully, and allow the layers to separate. With the aid of a Pasteur filter pipet, remove each ether layer and combine it with the ether solution retained above. After the final extraction, save the aqueous (lower) layer until you have isolated and characterized the final product. Extract the combined ether fractions with 0.5 mL of cold water to remove any acidic material. Save the aqueous rinse until you have isolated and characterized the final product.

Dry the ether solution by transferring it, using a Pasteur filter pipet, to a shortened Pasteur filter pipet containing 500 mg of anhydrous sodium sulfate. Collect the eluate in a tared 10 × 75-mm test tube. In the **hood,** remove the ether solvent from the eluate by warming in a sand bath to concentrate the solution to a weight less than 90 mg.

HOOD

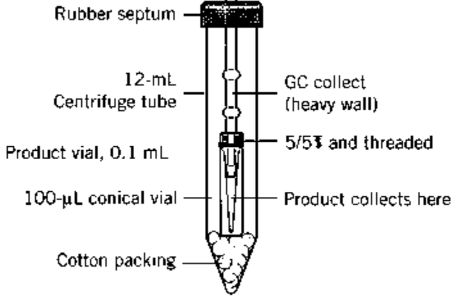

Purification and Characterization. The product, 4-methyl-3-heptanol, is isolated and purified using gas chromatography.

Use a 100-μL syringe to inject the entire sample of crude material obtained above onto the GC column. Collect the components of interest as they elute from the column, using the chromatography technique described in the Prior Reading section (➨).

Gas Chromatographic Conditions

GOW-MAC Series-150, Thermal Conductivity Detector
20% Carbowax 20M column, $\frac{1}{4}$ in. × 8 ft
Temperature, 145 °C
Flow rate of 50 mL/min (He gas)

The retention time for 4-methyl-3-heptanol under these conditions is 7–8 min (2 min after any other peak).

Determine the weight of the 4-methyl-3-heptanol collected and calculate the percent yield. Determine the boiling point and refractive index (optional) of the alcohol, and compare these with values in the literature.

Obtain an IR spectrum of the product as a thin film and compare it with that recorded in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar). If enough of the hydrocarbon byproduct is collected, also obtain an IR spectrum to aid in its identification.

Chemical Tests. The ignition test should indicate that this compound is an aliphatic species. Does your result confirm this fact? Perform the ceric nitrate test to demonstrate the presence of the —OH group and the Lucas test to demonstrate that a secondary alcohol has been prepared. If you were required to prepare a solid derivative of this alcohol, which one would you select? It may be of interest to determine the solubility of this product in water, ether, concentrated sulfuric, and 85% phosphoric acids. Do your results agree with what you would predict? What test(s) would you perform to establish that one of the starting reagents was an aldehyde?

QUESTIONS

6-114. Show how one could carry out each of the following transformations using the Grignard reaction. Any necessary organic or inorganic reagents may be used. Name each reactant and product.

(a) $CH_3(CH_2)_2CH_2Br \rightarrow (CH_3CH_2CH_2CH_2)_2CH\underset{\cdot\cdot}{\ddot{O}}H$

(b)

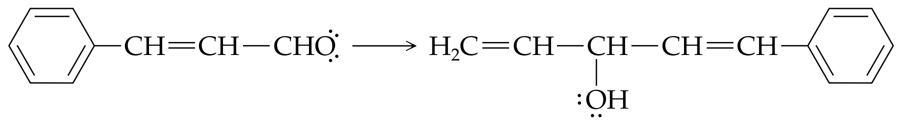

(c)

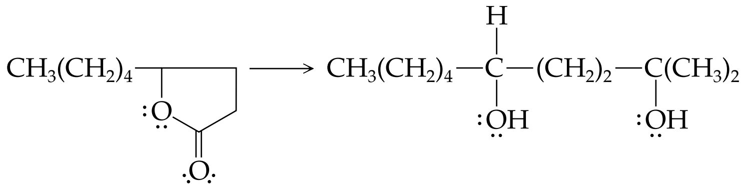

(d)

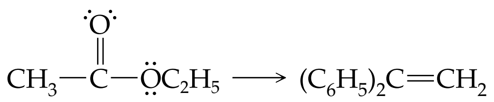

$CH_3-\overset{\overset{\ddot{O}}{\|}}{C}-\underset{\cdot\cdot}{\ddot{O}}C_2H_5 \longrightarrow (C_6H_5)_2C=CH_2$

6-115. Explain why Grignard reagents cannot be prepared from an organic halide that also contains a hydroxyl (—OH), a carboxyl (—CO_2H), a thiol (—SH) or an amino (—NH_2) group.

6-116. What would be the final product of the reaction between methyl benzoate and two equivalents of ethylmagnesium bromide?

6-117. Consider the same reaction as in Question 6-116 except that in this case it is carried out with ethyl benzoate. What product would be expected in this case?

6-118. Grignard reagents may be used to prepare other organometallic reagents, for example, ethylmagnesium bromide reacts with cadmium chloride to yield diethylcadmium:

$$2\ CH_3CH_2MgCl + CdCl_2 \rightarrow (CH_3CH_2)_2Cd + 2\ MgCl_2$$

Indicate the products from each of the following reactions and name each organometallic product:

$$4\ CH_3MgCl + SiCl_4 \rightarrow$$

$$2\ C_6H_5MgCl + HgCl_2 \rightarrow$$

BIBLIOGRAPHY

List of secondary alcohol preparations presented in *Organic Syntheses*:

Boeckman, R. K., Jr.; Blum, D. M.; Ganer, B.; Halvey, N. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 1033.

Coburn, E. R. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 696.

Coleman, G. H.; Craig, D. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 179.

Drake, N. L. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 406.

Overberger, C. G.; Saunders, J. H.; Allen, R. E.; Gander, R. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 200.

Skattebol, L.; Jones, E. R. H.; Whiting, M. C. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 792.

Trust, R. I.; Ireland, R. E. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 606.

EXPERIMENT 18

The Perkin Reaction: Condensation of Rhodanine with an Aromatic Aldehyde to Yield o-Chlorobenzylidene Rhodanine

Common name: *o*-chlorobenzylidene rhodanine
CA number: [6318-36-1]
CA name as indexed: 4-thiazolidinone, 5-[(2-chlorophenyl)methylene]-2-thioxo-

Purpose. This experiment explores the use of the interesting heterocyclic compound, rhodanine, as the source of an active (acidic) —CH_2— group. (The methylene group contained in the thiazolidinone ring system possesses the capacity to participate in base-catalyzed condensation reactions similar to those of the **aldol reaction.**) You will carry out a base-catalyzed condensation reaction with an aromatic aldehyde. You will examine the properties of this condensation product, which has the capacity to function as an intermediate in a number of synthetic pathways. Indeed, one of these routes yields the important class of aromatic amino acids, the phenylalanines.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallization (pp. 89–91)

REACTION

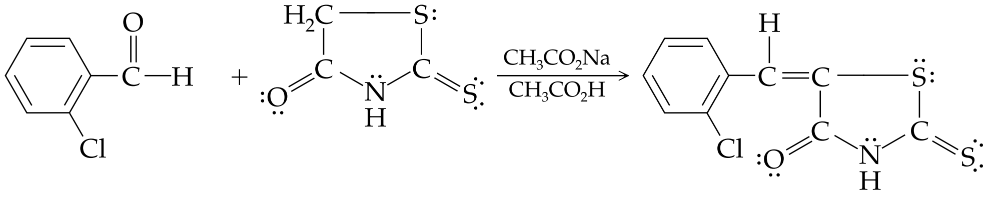

o-Chlorobenzaldehyde Rhodanine *o*-Chlorobenzylidene rhodanine

Sir William Henry Perkin (1838–1907) Perkin came under the influence of the German chemist, A. W. von Hofmann, at the Royal College of Chemistry (London), where Perkin was a student. In his second year at the school, he gained the title of Hofmann's honorary assistant, and following a publication the next year, when he was 17, he was advanced to the rank of assistant. At home, he set up his own laboratory where he worked evenings and on vacations. This laboratory is where he undertook, at Hofmann's suggestion, the synthesis of quinine.

Working with his friend Arthur Church in his home laboratory, he and Church prepared one of the first azo dyes derived from naphthalene (nitrosonaphthlene). This effort resulted in his first patent (with Church). In 1856 at the age of 18, Perkin discovered the first commercially significant synthetic coal-tar dye, *mauve* or *aniline purple*. He also obtained a patent on the synthetic method for preparing this material. The method involved the treatment of an aniline salt with bichromate of potash ($K_2Cr_2O_7$) (this was an outgrowth of his quinine studies). Against Hofmann's wishes, Perkin withdrew from college, designed a factory, and started commercial production of this dye, which was marketed as Perkin's Tyrian Purple. The name *Tyrian Purple* was originally used for a prized purple dye from the Phoenician city of Tyre, which today is part of Lebanon. Tyrian Purple was produced in small quantities from a material isolated from a snail. The production of this dye became a lost art during the Dark Ages, but a species of mollusk containing this very rare pigment was rediscovered in Ireland in 1684. The natural dye is obtained by the air oxidation of a colorless fluid expressed from the glands of the snail. It required the contents of 10,000 snails to obtain a single gram of the fluid.

Perkin's dye synthesis was the beginning of the coal-tar dyestuffs industry. He later developed and manufactured magenta (violet dye) and alizarin (a red dye). The rapid acceptance of these dyes by fabric manufacturers was demonstrated by the fact that annual alizarin production reached 220 tons per year by 1871. At age 37 and a wealthy man, Perkin sold his commercial holdings and devoted the rest of his life to pure chemical research.

Perkin's later years were as productive as his earlier ventures. He developed methods for the preparation of aminoacetic acid. He established the structural relationships between tartaric, fumaric, and maleic acids, and synthesized cinnamic acid. This last endeavor led to the development of what is now known as the *Perkin reaction*. The reaction is widely used to prepare unsaturated acids from aromatic aldehydes. These studies led to his synthesis of coumarin (actually it was the first condensation product he obtained with the classic reaction; see also Experiment [3A$_{adv}$]). Other areas investigated by Perkin dealt with the relationship of physical properties and chemical structure.

Perkin was knighted in 1906 in recognition of his contributions to chemistry and to the practical application of many of his discoveries. In the United

States, The Society of Chemical Industry awards a Perkin Medal each year to an outstanding industrial chemist.[19]

DISCUSSION

The classic Perkin reaction is the base-catalyzed condensation of an aromatic aldehyde with a carboxylic acid anhydride to yield and an α,β-unsaturated carboxylic acid. The initial stages of the condensation can be viewed as an aldol-type reaction (see Experiments [20], [$3A_{adv}$], [$3B_{adv}$], and [A3a].

www

A variation of the Perkin reaction is the condensation of aromatic aldehydes with rhodanine, which plays a similar role to that of the anhydride in the original reaction. Rhodanine, a derivative of the thiazolidinone ring system was first synthesized in 1935 by Percy Julian (a future president of Howard University) and Bernard Sturgis (an undergraduate at DePauw University at the time). This heterocyclic molecule has an active (acidic) methylene group that can be deprotonated with a relatively mild base (in this case acetate, CH_3COO^- or AcO^-, ion) to generate the nucleophile that attacks the carbonyl group of the aldehyde.

Under the conditions used in the current experiment, dehydration–elimination rapidly follows the initial nucleophilic addition with formation of the benzylidene intermediate. The mechanism is shown here:

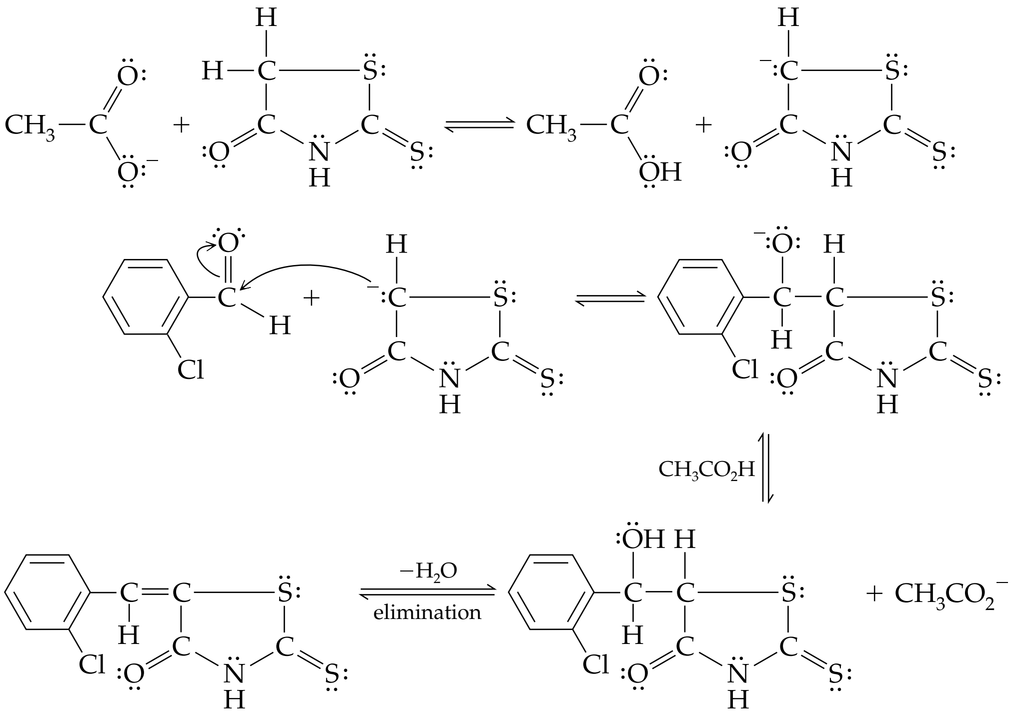

While the resulting rhodanine derivatives have exhibited antibacterial, antitubercular, antimalarial, antifungal, and antiparasitic activity, the principal focus of attention on these interesting compounds has been as reactive synthetic intermediates. For example, these particular compounds can be converted, in high yield, in three steps, to nitriles in which the side chain of the original aldehyde has been extended by an additional carbon atom:

---

[19] See Edelstein, S. M. *American Dyestuff Reporter* **1956,** *45,* 598; Mendola, R. *J. Chem. Soc.* **1908,** *93,* 2214; Levinstein, H. *Chem. Ind.* **1938,** 1137; Rose, R. E. *Ind. Eng. Chem.* **1938,** *16,* 608; *Dictionary of Scientific Biography,* C. C. Gillespie, Ed., Scribner's: New York, 1974, Vol. X, p. 515.

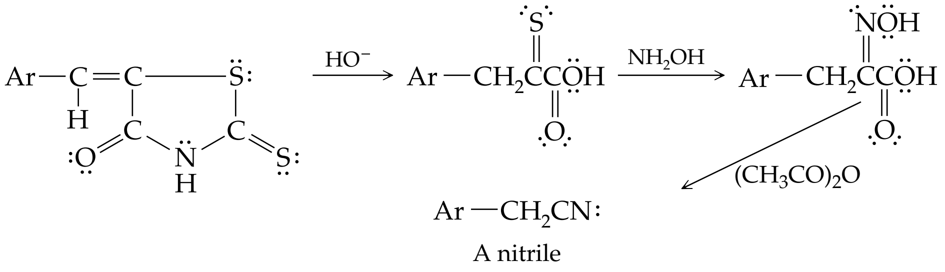

EXPERIMENTAL PROCEDURE

Estimated time of the experiment: 1.5 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Rhodanine	133.19	30 mg	0.23	170			
Sodium acetate	82.03	52 mg	0.63	324			
Acetic acid, glacial	60.05	1.0 mL			118		
o-Chlorobenzaldehyde	140.57	58 mg	0.41		212	1.25	1.5662
o-Chlorobenzylidene rhodanine	259.76			191			

HOOD

Reagents and Equipment. In a 3.0-mL conical vial containing a boiling stone and equipped with an air condenser, weigh out and place 30 mg (0.23 mmol) of rhodanine and 52 mg (0.63 mmol) of anhydrous sodium acetate. Now, in the **hood,** add 1.0 mL of glacial acetic acid dispensed from a graduated pipet. To this mixture, measure and add 58 mg (0.41 mmol) of *o*-chlorobenzaldehyde (⬅).

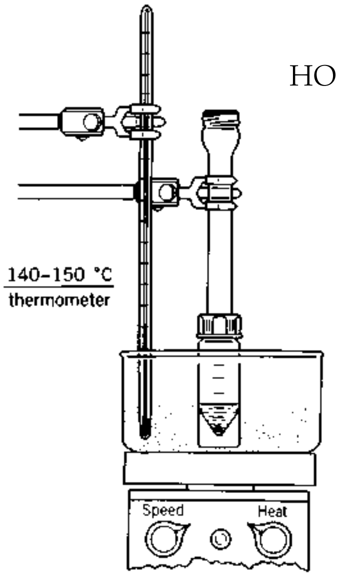

Rhodanine, 30 mg
+ NaOAc, 52 mg
+ CH_3CO_2H, 1.0 mL
+ o-ClC_6H_4CHO, 58 mg
Note *It is important that the vial be immersed in sand to the level of the reaction mixture.*
Ac = acetyl group

NOTE. Dry the sodium acetate in the oven for 1 h before use. The aldehyde ***must*** *be free from the corresponding acid or lower yields of product will result. It is recommended that the purity of the aldehyde be checked by IR analysis. The reaction vial may be weighed before and after the addition of aldehyde to obtain an accurate weight.*

Reaction Conditions. Heat the reaction mixture in a sand bath at 140–150 °C for 30 min.

NOTE. Immerse the vial in the sand up to the level of the top of the reaction mixture.

As the reaction progresses, the mixture becomes homogeneous and turns yellow. At the end of the reaction period the resulting solution is cooled to room temperature. When the conical vial has reached ambient temperature, place it in an ice bath to complete crystallization of the condensation product.

Isolation of Product. Collect the yellow crystals by vacuum filtration using a Hirsch funnel. Rinse the reaction vial with two 1.0-mL portions of cold glacial acetic acid transferred by a calibrated Pasteur pipet, and then use each rinse to wash the filter cake on the Hirsch funnel (➨). Complete the removal of the acid solvent by air drying the crystals on a porous clay plate.

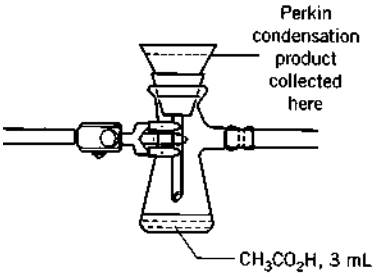

Purification and Characterization. Weigh the dried product and calculate the percent yield of the crude material. Recrystallization (Craig tube) of a 5-mg portion of the benzylidene product from 0.5 mL of glacial acetic acid yields fine, bright yellow needles that can be used to complete the characterization of this interesting substance.

Determine the melting point and compare your results with the value given in the table at the beginning of the experiment.

Chemical Tests. This compound is an interesting reaction product because it contains three different types of heteroatoms: chlorine, nitrogen, and sulfur. The sodium fusion test (see Chapter 9) may be used to substantiate the presence of these elements.

It would also be of interest to establish if a positive Beilstein test for chlorine is found when in the presence of sulfur and nitrogen, or whether the soda–lime test for nitrogen is observed with the elements of chlorine and sulfur also located within this molecule (see Chapter 9).

Does ignition of the material (see Chapter 9) indicate that the compound contains an aromatic ring?

OPTIONAL SEMIMICROSCALE PREPARATION

This experiment may be scaled up and carried out at a level 5 or 10 times greater than the amounts used in the above microscale experiment. The data given below are for the 10-fold procedure. The procedure is identical to that given above in the microscale section, but with the following modifications:

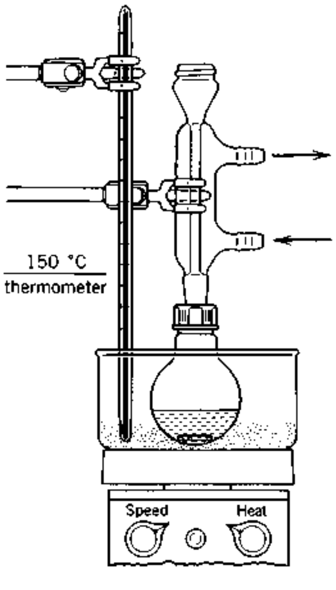

10-mL RB flask

1. A 10-mL round-bottom flask containing a spin bar and fitted with an air cooled condenser is used in place of the conical reaction vial (➨).

2. The reagent and solvent amounts are given in the following table.

Physical Properties of Reactants and Products

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Rhodanine	133.19	300 mg	2.3	170			
Sodium acetate	82.03	520 mg	6.3	324			
Acetic acid, glacial	60.05	5.0 mL		118			
o-Chlorobenzaldehyde	140.57	580 mg	4.1		212	1.25	1.5662
o-Chlorobenzylidene rhodanine	259.76		191				

3. After the product is air dried on a clay plate until no acetic acid (vinegar) odor remains, it is placed in a 100 °C oven to dry overnight. It may also be dried in a vacuum drying oven until no acetic acid (vinegar) odor remains.

QUESTIONS

6-119. In which of the following two species, (a) or (b), is the underlined hydrogen atom more acidic? Explain.

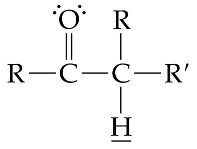

(a)

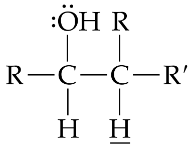

(b)

6-120. A number of compounds similar to rhodanine and possessing active methylene groups have been used in the Perkin condensation. Two are shown below:

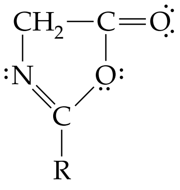

(An azlactone)

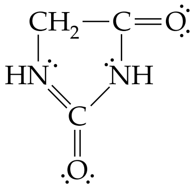

Hydantoin

Draw the structure of the product that would be formed if each underwent the Perkin condensation with *p*-chlorobenzaldehyde.

6-121. *para*-Nitrobenzaldehyde reacts at a faster rate than benzaldehyde in the Perkin reaction, while *p-N,N*-dimethylaminobenzaldehyde is much less reactive toward the same nucleophile. Explain.

6-122. Explain the fact that the $C{=}O$ group in 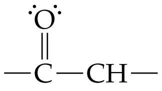 $-\overset{\overset{\displaystyle O}{\|}}{C}-CH-$ is effective in increasing the acidity of the α-hydrogen atom.

6-123. A large number of condensations that are closely related to the Perkin reaction. Among these are the aldol (Experiments [20], [3B$_{adv}$], and [A3a]), Knoevenagel, Claisen (Experiment [3A$_{adv}$]), and Dieckmann condensations. What general class of compounds can be prepared using each of these well-known reactions?

www →

BIBLIOGRAPHY

For a review on the general Perkin reaction see

Johnson, J. R. *Org. React.* **1942,** *1,* 210.
Koepp, E.; Vögtle, F. *Synthesis* **1987,** 177
Rosen, T. *Comp. Org. Syn.* **1991,** *2,* 395.

For examples of the general Perkin reaction see

Corson, B. B. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 229.
Herbst, R. M.; Shemin, D. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 1.
Johnson, J. R. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 426.
Thayer, F. K. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 398.
Weiss, R. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 61.

For references using rhodanine in the Perkin reaction see

Andreasch, R. *Monatsh. Chem.* **1928,** *49,* 122.
Brown, F. C. *Chem. Rev.* **1961,** *61,* 463.
Campbell, N.; McKail, J. E. *J. Chem. Soc.* **1948,** 1215.
Foye, W. O.; Tovivich, P. *J. Pharm. Sci.* **1977,** *66,* 1607.
Julian, P. L.; Sturgis, B. M. *J. Am. Chem. Soc.* **1935,** *57,* 1126.

EXPERIMENT 19

Alkene Preparation by the Wittig Reaction: (E)-Stilbene; Methylene-4-tert-butylcyclohexane; and trans-9-(2-Phenylethenyl)anthracene

Common names: *(E)*-stilbene; *trans*-1,2-diphenylethene
CA number: [103-30-0]
CA name as indexed: benzene, 1,1′-(1,2-ethenediyl)bis-, *(E)*-

Common name: 1-methylene-4-*tert*-butylcyclohexane
CA number: [13294-73-0]
CA name as indexed: cyclohexane, 1-(1,1-dimethylethyl)-4-methylene-

Common name: *trans*-9-(2-phenylethenyl)anthracene
CA number: [42196-97-4]
CA name as indexed: anthracene, 9-(2-phenylethenyl)-, *(E)*-

Purpose. The conditions under which the Wittig reaction is carried out are investigated. The Wittig reaction involves the reaction of a *phosphorus ylide* (Experiments [19A] and [19C]) with an aldehyde or ketone, and is used extensively in organic synthesis to synthesize alkenes. The use of the Horner–Wadsworth–Emmons modified Wittig reaction between an aldehyde and a phosphonate ester is investigated (Experiments [19B] and [19D]) using phase-transfer catalysis.

Prior Reading

- *Technique 4:* Solvent Extraction
  - Liquid–Liquid Extraction (p. 72)
  - Drying of a Wet Organic Layer (pp. 80–83)
  - Concentration of Solutions (pp. 101–104)
- *Technique 5:* Crystallization
  - Use of the Hirsch Funnel (pp. 88–89)
  - Craig Tube Crystallization (pp. 89–91)
- *Technique 6:* Chromatography
  - Packing the Column (p. 93)
  - Elution of the Column (p. 94)
  - Thin-Layer Chromatography (pp. 97–99)

Georg Friedrich Karl Wittig (1897–1987) Wittig obtained his Ph.D. in 1923 at the University of Marburg under von Auwers. He later became Professor of Chemistry at the Universities of Braunschweig, Freiburg, Tübingen, and Heidelberg. His initial research work was concerned with the concept of ring strain, diradical formation, and valance tautomerism. However, he soon became involved in carbanion chemistry. Wittig discovered halogen–metal exchange reactions and then moved extensively into research involving the chemistry of the ylides. In 1953, he discovered the reactive phosphonium ylides and subsequently their reaction with aldehydes and ketones to give alkenes under very mild conditions; this reaction is now known as the *Wittig reaction*. This discovery allowed the introduction of the C$=$C linkage at a specific location in the product. In recognition for this work, Wittig received the Nobel Prize in 1979 (with H. C. Brown [organoboron compounds]).

In other work, Wittig postulated the dehydrobenzene intermediate and proved its existence through trapping reactions of the Diels–Alder type. He discovered sodium tetraphenylborate, which is now used in the analytical determination of potassium and ammonium ions. His later work involved the chemistry of metalated Schiff bases. This work subsequently led to the development of the concept of directed aldol condensations.

Wittig coined a large number of technical terms—valence tautomerism, ylide, onium complexes, halogen–metal exchange, dehydrobenzene, and umpolung (polarity reversal). Wittig died in 1987 at the age of 90.[20]

REACTION

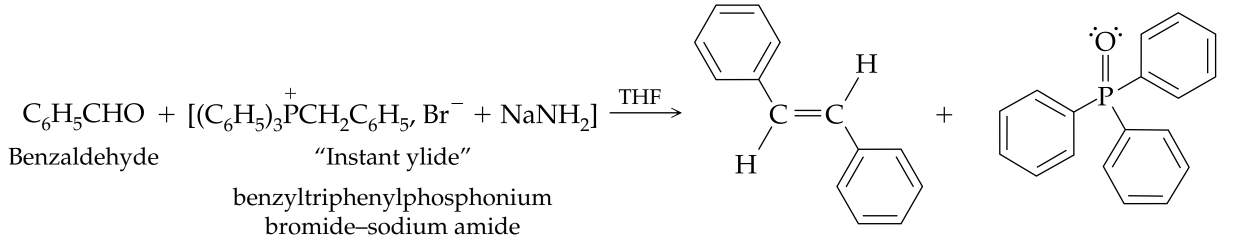
$C_6H_5CHO + [(C_6H_5)_3\overset{+}{P}CH_2C_6H_5, Br^- + NaNH_2] \xrightarrow{THF}$
Benzaldehyde
"Instant ylide"
benzyltriphenylphosphonium
bromide–sodium amide
H
C═C
H
+
O
P

E-Stilbene

Triphenylphosphine oxide

DISCUSSION

The Wittig reaction constitutes a valuable method for the preparation of alkenes. The major advantages of this approach are that (1) there is no ambiguity in the location of the C═C generated by the reaction, as there can be in many elimination reactions, and (2), there is no potential for rearrangements, as there can be in E1 elimination reactions.

An *ylide* is a neutral species whose Lewis structure contains opposite charges on adjacent atoms. The atoms involved are carbon and an element from either group 15 (VA) or 16 (VIA) of the periodic table, such as N, P, or S. The Wittig reaction uses phosphorus ylides, which are obtained by deprotonation of a phosphonium salt with a strong base. Phosphorus ylides are relatively stable, but reactive species, for which the following resonance structures may be written; the phosphorus atom can exceed an octet by accommodating electron donation into its 3d orbitals.

$$[R_3\overset{+}{P}\text{—}\overset{-}{C}H_2\colon \longleftrightarrow R_3P{=}CH_2]$$

The phosphonium salts are available through a nucleophilic displacement reaction of triphenylphosphine with various alkyl halides. Triphenylphosphine is a good nucleophile, and thus most phosphonium salts are easily prepared.

$$(C_6H_5)_3P\colon + RCH_2X \longrightarrow (C_6H_5)_3\overset{+}{P}CH_2R, X^-$$
$$(X = I, Br, Cl)$$

The hydrogen atoms on the carbon attached to the resulting phosphorus cation are somewhat acidic because they are adjacent to a positive charge, a significant electron-withdrawing group. Thus, treatment of the phosphonium salt with a strong base, such as butyllithium in THF or sodium hydride in DMSO, removes one of these protons and produces the ylide.

$$(C_6H_5)_3\overset{+}{P}CH_2R, X^- + C_4H_9Li \xrightarrow{THF} \underset{\text{An ylide}}{(C_6H_5)_3\overset{+}{P}\text{—}\overset{-}{\ddot{C}}H\text{—}R} + C_4H_{10} + LiX$$

---

[20] See Parker, S. P. Ed., *McGraw-Hill Modern Scientists and Engineers,* McGraw-Hill: NY, 1980, Vol. 3, p. 341; *The Annual Obituary;* St. James Press: Chicago, 1987, p. 460; Oesper, R. E. *J. Chem. Educ.* **1954,** *31*, 357.

The ylides are generally not isolated, but rather generated in situ and reacted directly with the carbonyl compound.

The "instant ylides" are solid-phase mixtures of a phosphonium salt with sodium amide ($NaNH_2$, a strong base and the conjugate base of ammonia) and are now commercially available. In the solid phase, no reaction between the strong base and the phosphonium salt occurs. Thus, to generate the desired ylide, the "instant ylide" mixture need only be placed in a suitable solvent. This process is a marked advantage over the usual methods used to obtain these species. Ylides, because of the significant negative charge density on carbon, are nucleophilic enough to attack electrophiles as reactive as the carbon of a carbonyl group. When the ylide is reacted with an aldehyde or a ketone, an intermediate *oxaphosphetane* (a four-membered ring containing an oxygen and a phosphorus atom) is formed, which decomposes to give the alkene product. The mechanistic sequence is outlined here.

Generation of the Ylide

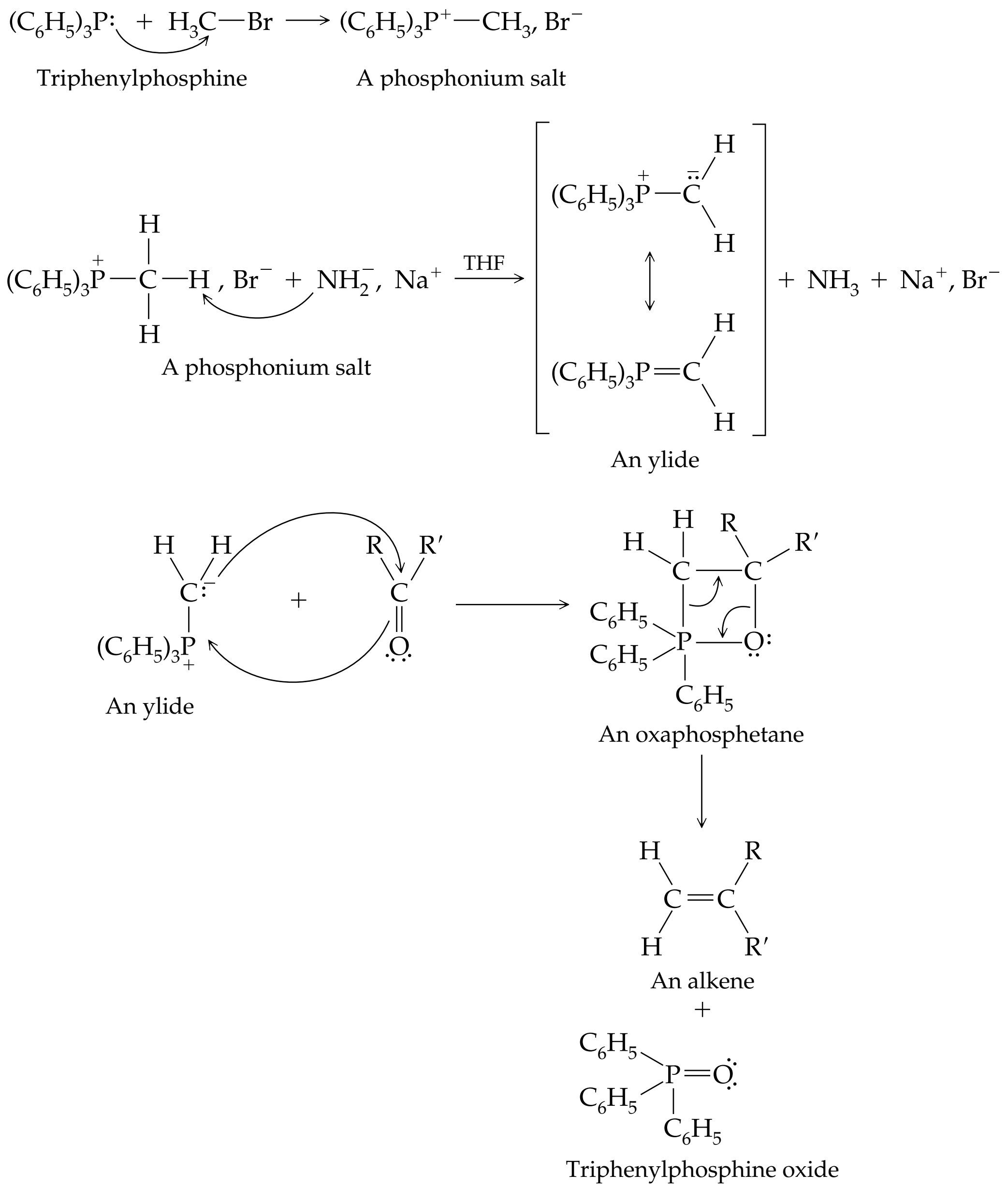
(C6H5)3P: + H3C—Br → (C6H5)3P+—CH3, Br−
Triphenylphosphine
A phosphonium salt
(C6H5)3P+—CH(H)—H, Br− + NH2−, Na+
THF
A phosphonium salt
+ NH3 + Na+, Br−
An ylide
An ylide
An oxaphosphetane
An alkene
+
Triphenylphosphine oxide

The Wittig reaction is very general. The significant limitation is that the carbonyl compound cannot contain acidic hydrogen atoms, such as in an alcohol or carboxyl group, or electrophiles more reactive than the aldehyde or ketone itself. The Wittig reaction can be quite stereoselective for the formation of either *Z* or *E* alkenes, but the factors involved in such stereoselectivity are sometimes difficult to predict and often difficult to explain.

An important modification (often called the Horner–Wadsworth–Emmons reaction) of the Wittig reaction makes use of phosphonate esters, $RPO(OR')_2$. It is highly stereoselective for the formation of *E*-alkenes. The reaction and mechanism are depicted below for the preparation of *(E)*-stilbene. Instead of using a phosphorus cation to stabilize the negative charge, as in the phosphonium ylide above, a phosphonate ester group is used to stabilize an adjacent carbanion.

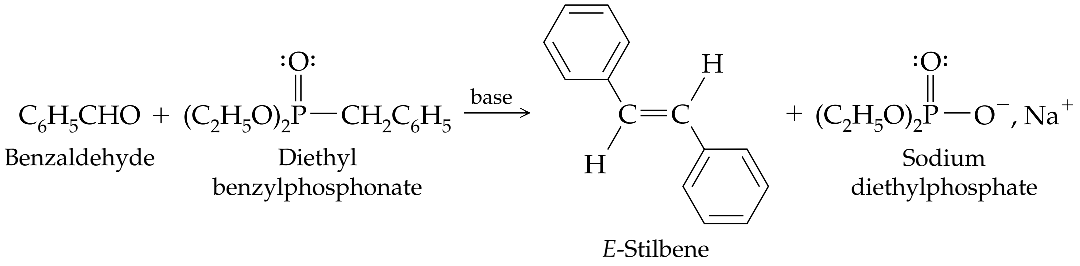
C_6H_5CHO + $(C_2H_5O)_2P(=O)—CH_2C_6H_5$
base
+ $(C_2H_5O)_2P(=O)—O^-$, Na^+
Benzaldehyde
Diethyl benzylphosphonate
E-Stilbene
Sodium diethylphosphate

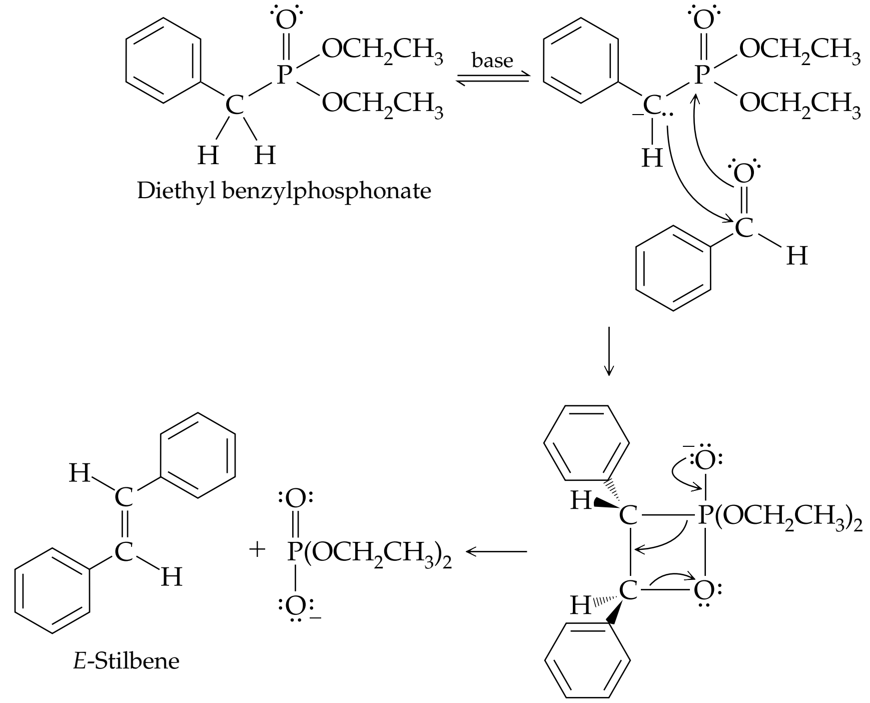
OCH_2CH_3
OCH_2CH_3
Diethyl benzylphosphonate
base
$P(OCH_2CH_3)_2$
E-Stilbene

The use of the phosphonate ester (Horner–Wadsworth–Emmons reaction) allows much easier separation of the product alkene, since the sodium phosphate byproduct is water soluble; the byproduct of the Wittig reaction, triphenylphosphine oxide, is not water soluble. In the Horner–Wadsworth–Emmons modification, a conjugated, or electron-withdrawing, substituent (such as a phenyl or carbonyl group) on the nucleophilic carbon is used to assist in the stabilization of the carbanion. This modification (Experiment [19B]) may be used as an alternative to Experiment [19A] for the preparation of *(E)*-stilbene. The "instant-ylide" Wittig reaction yields predominantly the *E* isomer of stilbene (70%). The Horner–Wadsworth–Emmons reaction yields exclusively the *E* isomer. Both procedures are given below. The synthesis described in Experiment [19D] also uses the Horner–Wadsworth–Emmons modification.

Horner–Wadsworth–Emmons reactions lend themselves to the use of *phase-transfer catalysis* in Experiments [19B] and [19D]. Phase-transfer catalysis allows the use of an aqueous base (NaOH in H_2O) with the organic compounds dissolved in an organic solvent (hexane in Experiment [19B], methylcyclohexane in Experiment [19D]) immiscible with water. The reaction system, as the name implies, involves two phases—an aqueous phase and an organic phase. The phase-transfer catalyst plays a very important role: without it, no reaction would occur, since the initial reactants (hydroxide ion and diethyl benzylphosphonate) are dissolved in different, immiscible phases, and NaOH is insoluble in hexane (or methylcyclohexane), and diethyl benzylphosphonate is insoluble in water.

The key features of the phase-transfer catalyst, Aliquat 336, are that it is a quaternary ammonium salt, with long-chain alkyl groups attached to the nitrogen.

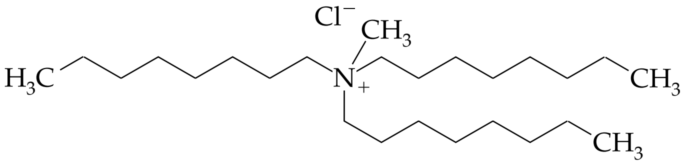

Aliquat 336

The phase-transfer catalyst is soluble in both water and the organic solvent. It is water soluble because it is an ion, and it is hexane soluble because of the three long-chain alkyl groups. Thus, the phase-transfer catalyst distributes itself in both phases, and freely shuttles back and forth through the phase boundary between solvent layers. In aqueous NaOH, the chloride anion exchanges with hydroxide anion, as the counterion to the ammonium cation. When it does this, the catalyst carries the hydroxide ion from the aqueous phase, as an ion-pair, across the phase boundary into the organic phase, where the base then reacts with the diethyl benzylphosphonate. The Horner–Wadsworth–Emmons reaction then occurs, producing the alkene and diethyl phosphate anion. This anion becomes associated with the ammonium cation of the phase-transfer catalyst and is transported to the aqueous layer, where the catalyst picks up another hydroxide ion and repeats the entire process.

(*E*)-Stilbene by the "Instant Ylide" Method

Experiment 19A

CAUTION: Tetrahydrofuran is a flammable liquid. All flames must be extinguished in the laboratory when this solvent is used.

REACTION

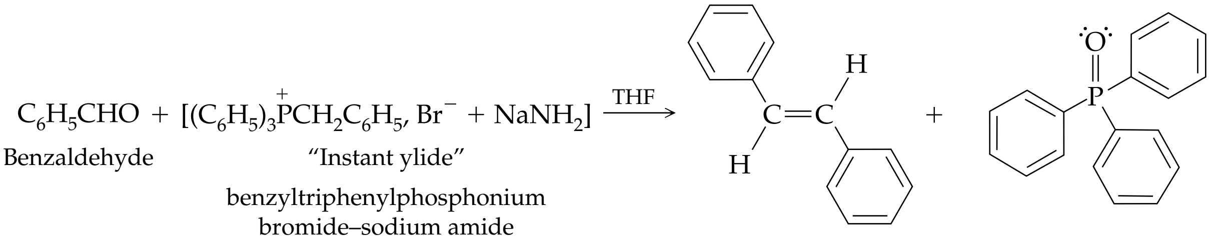

$$C_6H_5CHO + [(C_6H_5)_3\overset{+}{P}CH_2C_6H_5, Br^- + NaNH_2] \xrightarrow{THF}$$

E-Stilbene Triphenylphosphine oxide

EXPERIMENTAL PROCEDURE

Estimated time to complete the reaction: Two 3-h laboratory periods.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Benzaldehyde	106.13	100 μL	0.95	178	1.04	1.5463
"Instant ylide" benzyltriphenyl phosphonium bromide–sodium amide		600 mg	~1.2			
Tetrahydrofuran	72.12	1.0 mL		67		

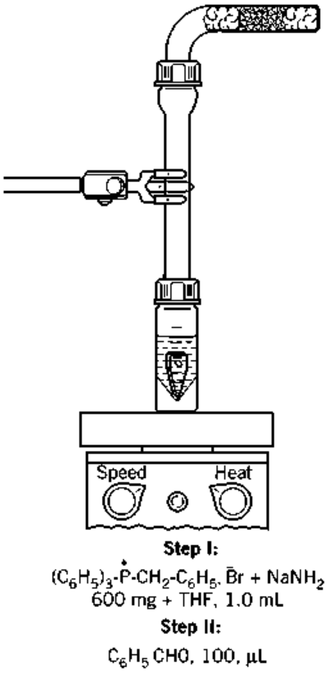
Speed
Heat
Step I:
$(C_6H_5)_3\text{-}\overset{+}{P}\text{-}CH_2\text{-}C_6H_5, \bar{B}r + NaNH_2$
600 mg + THF, 1.0 mL
Step II:
C_6H_5CHO, 100, μL

Reagents and Equipment. Weigh and place 600 mg (~1.2 mmol) of benzyltriphenylphosphonium bromide–sodium amide *(instant ylide)* mixture in a dry 5.0-mL conical vial containing a magnetic spin vane. Add *freshly distilled* dry tetrahydrofuran (1.0 mL) using a calibrated Pasteur pipet, and immediately attach the vial to an air condenser protected by a calcium chloride drying tube (⬅). Stir the mixture for 15 min at room temperature. During this time, the mixture turns orange. Following generation of the ylide, remove the air condenser for as short a time as possible and quickly add 100 μL (0.95 mmol) of benzaldehyde *(freshly distilled)* to the reaction flask, using an automatic delivery pipet, and immediately reattach the air condenser.

Reaction Conditions. Stir the resulting heterogeneous mixture at room temperature for an additional 15 min. The system develops a light brown color during this time.

Isolation of Product. Work up the reaction by adding 1.0 mL of a 25% aqueous NaOH solution (calibrated Pasteur pipet), and transfer the resulting mixture to a 15-mL centrifuge tube using a Pasteur pipet. Rinse the reaction vial with three 2.0-mL portions of diethyl ether, each of which is also transferred to the centrifuge tube (Pasteur filter pipet). Partially neutralize the resulting two-phase mixture by careful addition of 3.0 mL of 0.1 N HCl. Transfer the ether layer (top) by Pasteur filter pipet to a short microcolumn prepared from a Pasteur filter pipet containing 1.5 g of anhydrous sodium sulfate (⬅).

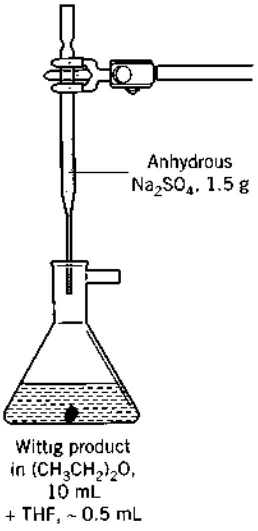

Collect the dried eluate in a 25-mL Erlenmeyer flask containing a boiling stone. Extract the remaining aqueous layer with two additional 2.0-mL portions of ether and transfer these ether extracts, as before, to the microcolumn containing anhydrous sodium sulfate. Concentrate the eluate solution (~10 mL) to dryness using a gentle stream of nitrogen, or by warming it in a sand bath in the **hood** to give a white solid.

HOOD

Purification and Characterization. Purify the crude product by chromatography, using a silica gel column.

The triphenylphosphine oxide is relatively insoluble in the hexane solvent. First, separate the triphenylphosphine oxide byproduct from the mixture of stilbenes by extracting the solid obtained above with hexane, using the following procedure.

To the 25-mL Erlenmeyer flask containing the crude product, add 2.0 mL of hexane and agitate the solution with swirling. Some breakup of the material with a microspatula may be necessary. Transfer the hexane solution by Pasteur filter pipet to a 10-mL Erlenmeyer flask containing a boiling stone. Extract the remaining crude solid with two additional 2.0-mL portions of hexane, and transfer these extracts to the 10-mL Erlenmeyer flask as before. Concentrate the hexane solution (~6 mL) to about 1.5 mL using a gentle stream of nitrogen gas, or by warming on a sand bath in the **hood.**

HOOD

Pack a short (1 × 10 cm) chromatography column with 2.0 g of activated silica gel, and premoisten the column with ligroin (60–80 °C) (➨). Add the above hexane solution directly to the column. Now elute the column with 15 mL of a 1:1 ligroin (60–80 °C)/methylene chloride solution. Collect the eluate, in a single fraction, in a tared 25-mL **filter flask** containing a boiling stone. Concentrate the solution to dryness under reduced pressure in a warm sand bath to yield the pure product mixture (➨).

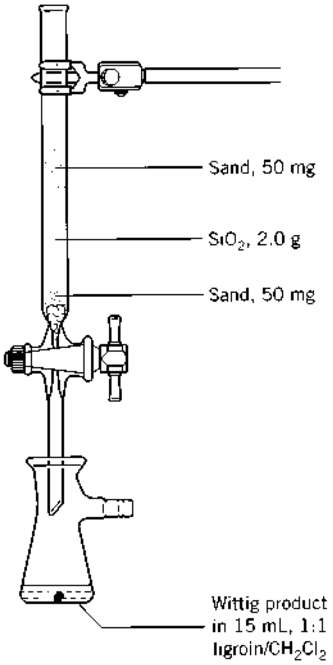

NOTE. A 25-mL sidearm filter flask, equipped with a Hirsch funnel and filter paper disks to control the pressure, is a convenient system for the removal of a small volume (5–20 mL) of solvent. A rotary evaporator, if available, is a nice alternative.

Weigh the crude product residue and calculate the percent yield.

The isolated product mixture may be analyzed by gas chromatography and/or thin-layer chromatography. *This constitutes the second week of laboratory for this experiment.*

Gas Chromatographic Analysis. Dissolve the crude product mixture isolated above in the minimum amount of 1:1 ligroin (60–80 °C)/methylene chloride solution (~0.5–0.75 mL). Inject a 10-μL sample into a gas chromatograph, set up according to the following conditions:

Column: $\frac{1}{4}$-in. × 8-ft, 20% Carbowax 20M on Chromasorb P (80/100 mesh)

Temperature: 220 °C

Flow rate: 30 mL/min (He gas)

Chart speed: 1 cm/min

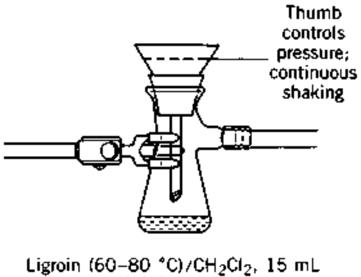

Ligroin (60–80 °C)/CH_2Cl_2, 15 mL + Wittig product

The compounds elute in the following order: *(Z)*-stilbene, followed by *(E)*-stilbene. Measure the ratio of peak heights and calculate the isomeric composition of your mixture.

Thin-Layer Chromatographic Analysis. Spot a TLC plate with a sample from the product solution used for GC analysis, and also with a standard solution.

Use hexane as the elution solvent, silica gel (containing a fluorescent indicator) as the stationary phase, and UV light for visualization.

Typical R_f values are *(E)*-stilbene 0.21; *(Z)*-stilbene, 0.27.

Concentrate the ligroin–methylene chloride product mixture as before. Separate, and purify, the *(E)*-stilbene by recrystallization from a minimum amount of 95% ethanol, using the Craig tube.

Weigh the dried and purified *(E)*-stilbene product and calculate the percent yield. The purity of this material may be determined by TLC, using the conditions outlined above.

Obtain a melting point and IR spectrum of the material and compare your results with those reported in the literature.

(E)- and *(Z)*-Stilbene also exhibit different absorptions in the ultraviolet region. The data are summarized below:

(Z)-Stilbene: (1-mm cell)

λ_{max} 223 nm (ε_{max} = 20,600 methanol, 0.05 g/L)

λ_{max} 276 nm (ε_{max} = 10,900, methanol, 0.1 g/L)

(E)-Stilbene: (0.05 g/L, 1 mm cell)

λ_{max} = 229 nm (ε_{max} = 21,000, methanol)
λ_{max} = 294 nm (ε_{max} = 33,200, methanol)
λ_{max} = 307 nm (ε_{max} = 32,100, methanol)

These ultraviolet absorption data illustrate an interesting example of steric effects on the absorption pattern exhibited by geometrical isomers of an alkene. There is significant steric hindrance between the two phenyl groups in the *Z* isomer, which causes the phenyl groups to twist out of coplanarity with the alkene. Thus, conjugation is diminished. This result is reflected in the lower intensity of the 276-nm band as compared to the 294-nm band in the *E* isomer.

Chemical Tests. Further characterization may be accomplished by performing the Br_2/CH_2Cl_2 test for unsaturation. Note that the dibromo compound is prepared in Experiment [$A2_b$]. It may be used here as a reference sample in the characterization of *(E)*-stilbene. The ignition test (Chapter 9) may be used to confirm the presence of the aromatic portion of the molecule.

Experiment 19B

(*E*)-Stilbene by the Horner–Wadsworth–Emmons Reaction

REACTION

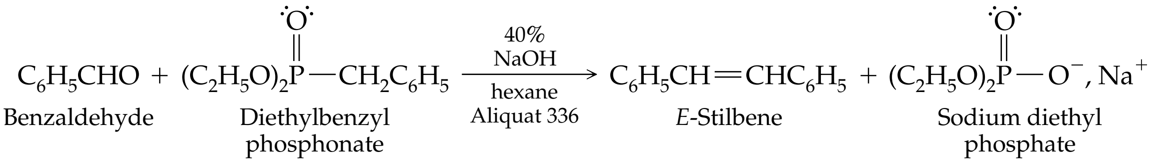

$$C_6H_5CHO + (C_2H_5O)_2\overset{\overset{\ddot{O}}{\|}}{P}-CH_2C_6H_5 \xrightarrow[\text{hexane, Aliquat 336}]{\text{40\% NaOH}} C_6H_5CH{=}CHC_6H_5 + (C_2H_5O)_2\overset{\overset{\ddot{O}}{\|}}{P}-O^-, Na^+$$

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Benzaldehyde	106.12	100 μL	0.98	178	1.04	1.5463
Diethyl benzylphosphonate	228.23	200 μL	0.96	106–108	1.095	
Aliquat 336 (tricaprylmethylammonium chloride)	404.17	88 mg	0.22			
Hexane	86.18	2.0 mL				
40% Sodium hydroxide		2.0 mL				

Reagents and Equipment. Weigh and place 88 mg (100 μL) of tricaprylmethylammonium chloride (Aliquat 336) in a 10-mL round-bottom flask containing a magnetic stirrer. Add 100 μL (0.98 mmol) of benzaldehyde, 200 μL (0.96 mmol) of diethyl benzylphosphonate, 2.0 mL of hexane, and 2 mL of 40% sodium hydroxide solution. Attach the flask to a reflux condenser.

HOOD

NOTE. The benzaldehyde, diethyl benzylphosphonate, hexane, and NaOH solution are dispensed in the ***hood*** *using automatic delivery pipets. Aliquat 336 is very viscous and is best measured by weighing. A medicine dropper is used to dispense this material. It is advisable to lightly grease the bottom joint of the condenser since strong base is being used.*

Reaction Conditions. Warm the two-phase mixture to reflux on a sand bath (temperature ~90–100 °C) for 1 h. Stir the reaction mixture vigorously during this period. Allow the resulting solution to cool to nearly room temperature. Crystals of product may appear as cooling occurs. Confirm by TLC (plates containing a fluorescent indicator) that the reaction is complete using as a solvent system ethyl acetate:hexane (0.1:9.9) and a UV lamp for visualization. The R_f value of benzaldehyde (0.4) is lower than both the R_f values of *E*- and *Z*-stilbene (➡).

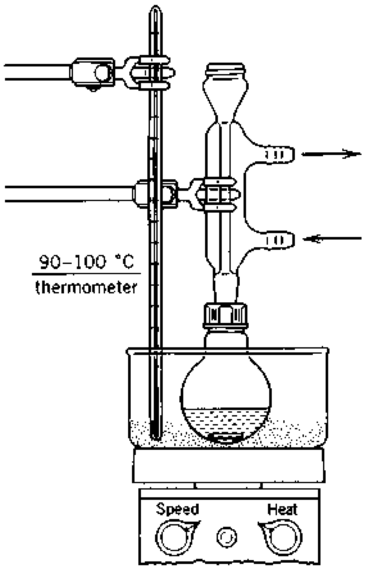

C_6H_5CHO, 100 μL
diethyl benzylphosphonate, 200 μL,
+ Aliquat 336, 88 mg
40% NaOH, 2 mL, hexane, 2 mL
10-mL RB flask

Isolation of Product. Once all the evidence shows that benzaldehyde is no longer present, add methylene chloride (700 μL), which will dissolve any crystalline material that may have formed. Then, using a Pasteur pipet, transfer the contents of the round-bottom flask to a 15-mL centrifuge tube. Rinse the flask with an additional 300 μL of methylene chloride and transfer the rinse to the same centrifuge tube. Remove the aqueous layer carefully, using a Pasteur filter pipet, and save it in a vial until you have successfully isolated and characterized the product.

NOTE: Use care when determining which layer is the "aqueous layer." At room temperature, the density of a 40% sodium hydroxide solution is 1.40 g/mL whereas the density of methylene chloride is 1.33 g/mL.

Wash the organic layer with two 1-mL portions of water. Stir the mixture with a small glass rod after each addition (or a Vortex mixer may be used) and then remove the water layer and add it to the one collected before. Dry the methylene chloride solution by addiing a small amount of anhydrous sodium sulfate. Use a Pasteur filter pipet to transfer the dried solution to a 10-mL Erlenmeyer flask. Wash the sodium sulfate remaining in the centrifuge tube with two 1-mL portions of methylene chloride. Remove these washings, using the Pasteur filter pipet, and transfer them to the same Erlenmeyer flask.

Concentrate the solution, which contains the desired product, to dryness on a warm sand bath under a stream of nitrogen.

Purification and Characterization. The *(E)*-stilbene obtained is, in most cases, sufficiently pure for characterization. However, this should be confirmed by thin-layer chromatography as outlined earlier (see Purification and Characterization, Experiment [19A]). If only a trace of the *Z* isomer is detected, recrystallize the product directly, using absolute ethanol. Collect the recrystallized material by vacuum filtration using a Hirsch funnel. Maintain the vacuum for an additional 10 min to partially dry the crystalline product. Then place it on a clay plate, or on filter paper, and allow it to dry thoroughly. As an alternative, the product may be dried in a vacuum drying oven (or pistol) for 10–15 min at 30 °C (1–2 mm Hg).

Weigh the dried *(E)*-stilbene product and calculate the percent yield. The *(E)*-stilbene may be characterized as outlined in Experiment [19A].

Methylene-4-*tert*-butylcyclohexane

Experiment 19C

REACTION

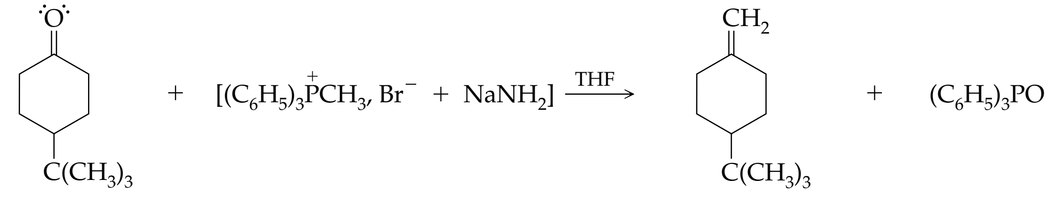

4-*tert*-Butylcyclohexanone

"Instant ylide"
methyltriphenylphosphonium
bromide–sodium amide

Methylene-
4-*tert*-butylcyclohexane

Triphenylphosphine
oxide

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 4.0 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	n_D
4-*tert*-Butylcyclohexanone	154.26	100 mg	0.64	50		
"Instant ylide" (methyltriphenylphosphonium bromide–sodium amide)		320 mg	~0.72			
Tetrahydrofuran		1.0 mL			67	
Methylene-4-*tert*-butylcyclohexane	151.27				185	1.4630

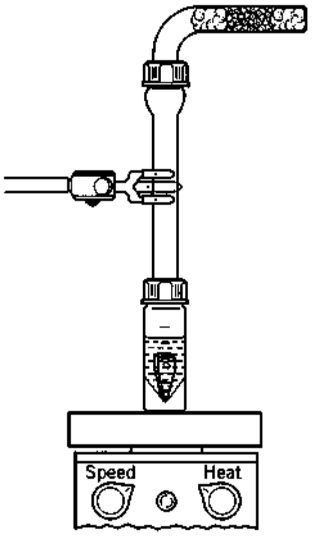

Step I:
$(C_6H_5)_3\overset{+}{P}\text{-}CH_3$, $\bar{B}r$ + $NaNH_2$,
320 mg + THF, 1.0 mL
Step II:
4-*tert*-Butycyclohexanone, 100 mg

Reagents and Equipment. In a dry 5.0-mL conical vial containing a magnetic spin vane, and equipped with an air condenser protected by a calcium chloride drying tube, weigh and place 320 mg (~0.72 mmol) of methyltriphenylphosphonium bromide–sodium amide (instant ylide) mixture (⬅). Now add *freshly distilled* tetrahydrofuran (1.0 mL), using a calibrated Pasteur pipet, and stir the mixture for 15 min at room temperature. During this period it turns a bright yellow color.

Following generation of the ylide, weigh and add 100 mg (0.64 mmol) of 4-*tert*-butylcyclohexanone to the reaction flask.

Reaction Conditions. Stir the resulting heterogeneous mixture, at room temperature, for an additional 90 min. The mixture develops a light-tan color over this period of time.

Isolation of Product. Work up the reaction by adding 1.0 mL of a 25% aqueous NaOH solution (calibrated Pasteur pipet), and then transfer the resulting mixture to a 12-mL centrifuge tube using a Pasteur filter pipet. Rinse the reaction flask with three 2.0-mL portions of diethyl ether. Transfer each rinse to the same centrifuge tube. Partially neutralize the resulting two-phase system by the careful addition of 2.0 mL of 0.1 N HCl and mix briefly.

Separate the ether layer by Pasteur filter pipet and place it in a 25-mL Erlenmeyer flask. Extract the remaining aqueous layer with three additional 5-mL portions of diethyl ether. Separate these extracts as before and combine them with the original ether layer.

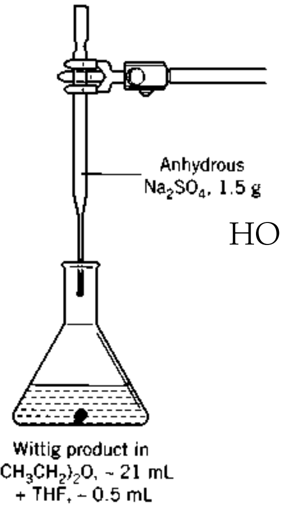

Wittig product in
$(CH_3CH_2)_2O$, ~ 21 mL
+ THF, ~ 0.5 mL

HOOD

Transfer the combined ether fractions, by Pasteur filter pipet, to a short microcolumn prepared from a Pasteur filter pipet containing 1.5 g of anhydrous sodium sulfate (⬅). Collect the dried eluate in a 25-mL Erlenmeyer flask containing a boiling stone. Concentrate this solution to dryness, using a gentle stream of nitrogen, or by warming on a sand bath in the **hood,** to yield a colorless liquid residue.

Purification and Characterization. Purify the crude product isolated above by chromatography on an alumina column (see figure on next page).

HOOD

Pack a short (1 × 10 cm) buret column with 4.0 g of activated basic alumina and premoisten the alumina with ligroin. Dissolve the crude material in 0.5 mL of 9:1 ligroin (60–80 °C)/methylene chloride solvent and transfer the solution to the column using a Pasteur filter pipet. Elute the product in a single fraction using 8.0 mL of 9:1 ligroin (60–80 °C)/methylene chloride. Collect the eluate in a tared 25-mL filter flask containing a boiling stone. Evaporate the solvent in the **hood** under vacuum, with swirling, in a warm sand bath, leaving a liquid residue. Product loss may occur if the concentrated residue is heated excessively (see figure on page 306).

The product is of sufficient purity for characterization. Weigh the methylene-4-*tert*-butylcyclohexane and calculate the percent yield. Determine the boiling point and compare your result with that listed in the Reactants and Product table.

Obtain an IR spectrum of the alkene product, using the capillary film technique. Compare your result with Figure 6.32.

Infrared Analysis. The spectrum (Fig. 6.8 p. 160) of the ketone was discussed in Experiment [5B]. The major change observed in the spectrum of the product (Fig. 6.32), when compared with the starting material, is the loss of the carbonyl absorption band (1717 cm^{-1}), and its replacement by a less-intense new band at 1654 cm^{-1}. This latter absorption is associated with the stretching motion of the C═C system exocyclic to the six-membered ring. The key bands associated with group frequencies present in the reaction product are identified at 3075, 3000–2850, 1783, 1654, 1394, 1368, and 883 cm^{-1}.

The presence of these bands confirms the projected structure of the Wittig reaction product. The data may be interpreted as follows: the bands at 3075 (sharp spike), 1783 (weak), 1654 (sharp–medium), and 883 (strong) cm^{-1} form a frequency train that is defined as a "terminal alkene macrofrequency." The overall interpretation requires the presence of all four bands in the spectrum if the specific assignment of any one of them is to be correct. Thus, these four data

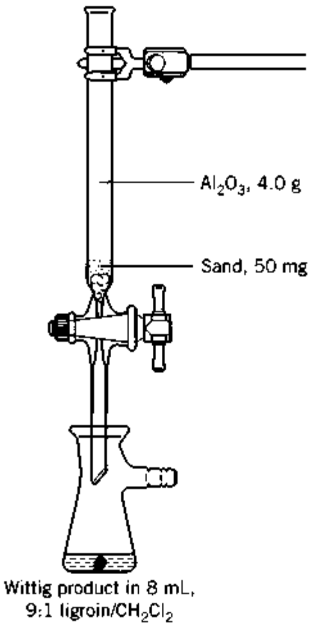

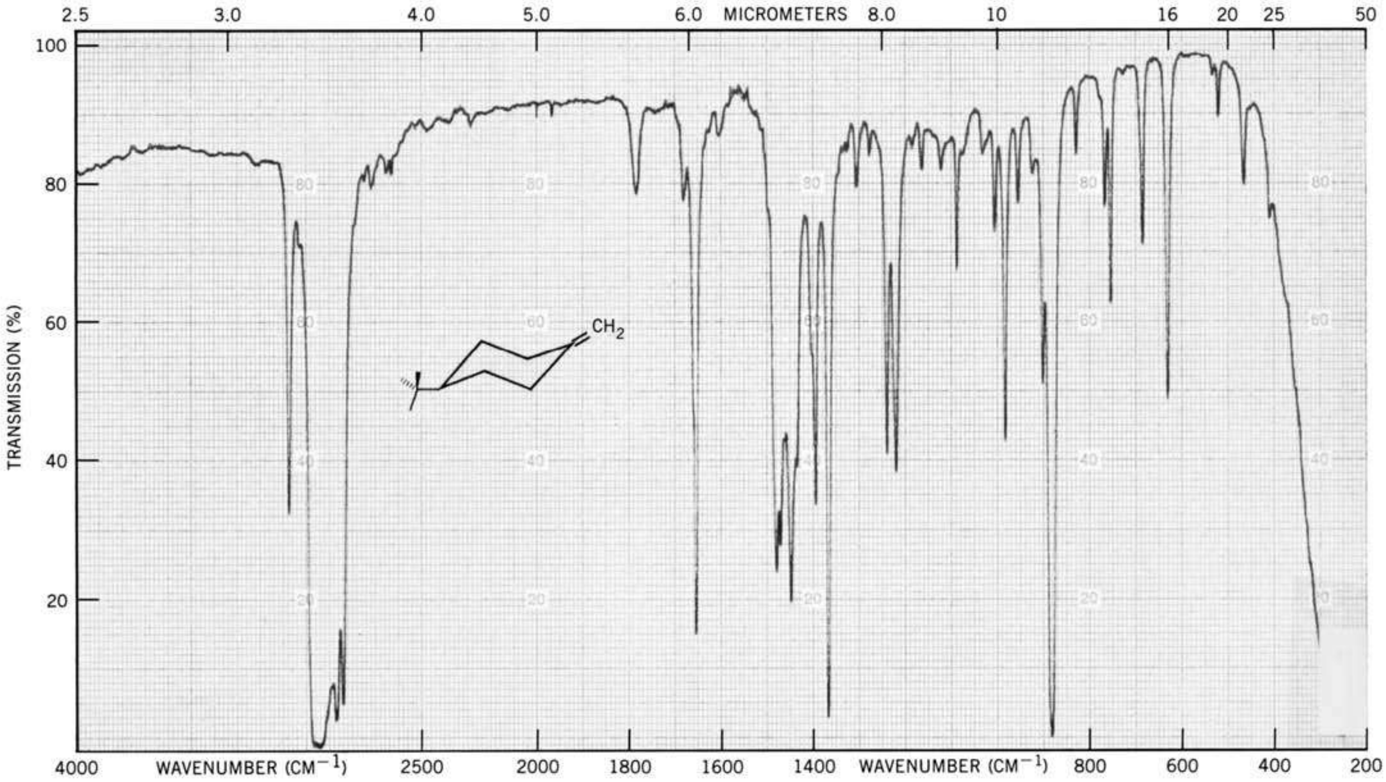

ABSCISSA EXPANSION SUPPRESSION	ORDINATE EXPANSION % T ✓ ABS	SCAN TIME 12 min RESPONSE SLIT PROGRAM N	REP. SCAN SINGLE BEAM TIME DRIVE PRE SAMPLE CHOP OPERATOR CD DATE 8/15/86
SAMPLE 4-t-Butylmethylene cyclohexane ORIGIN	REMARKS 283B	SOLVENT neat CONCENTRATION	CELL PATH KBr 0.015mm REFERENCE

Figure 6.32 IR spectrum: methylene-4-*tert*-butylcyclohexane.

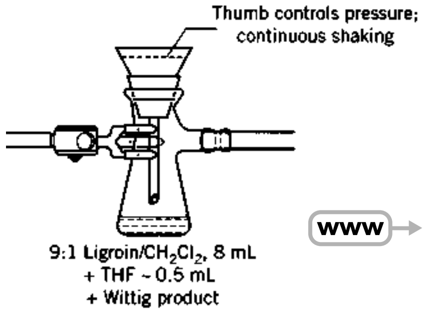
Thumb controls pressure; continuous shaking
9:1 Ligroin/CH_2Cl_2, 8 mL
+ THF ~ 0.5 mL
+ Wittig product

points lead to a structural interpretation with very high confidence limits. The 3075-wavenumber band arises from the coupled antisymmetric stretch of the two C—H oscillators on the terminal alkenyl methylene group. The weak 1783-wavenumber peak is an overtone of the strong band observed near 883 cm^{-1} (883 × 2 = 1766 cm^{-1}) and is unusually intense for an isolated harmonic. The fundamental can be assigned as the ${=}CH_2$ out-of-plane wag (C—H deformation). In this fairly rare example, the observed harmonic occurs at a higher wavenumber value than twice the fundamental frequency. A situation of this type is termed *negative anharmonicity* (see Infrared Discussions). Finally, the $C{=}C$ stretching mode is assigned to the 1654-wavenumber band, which is consistent with the requirements of this frequency train as it is found below 1660 cm^{-1} (see Infrared Discussions of cis, terminal, or vinyl carbon–carbon double bonds). The bands identified at 1396 and 1368 cm^{-1} indicate that the tertiary butyl group, as expected, has been preserved during the conversion of the ketone to a terminal alkene (see also Discussion, Experiment [5B]).

Experiment 19D

trans-9-(2-Phenylethenyl)anthracene

REACTION

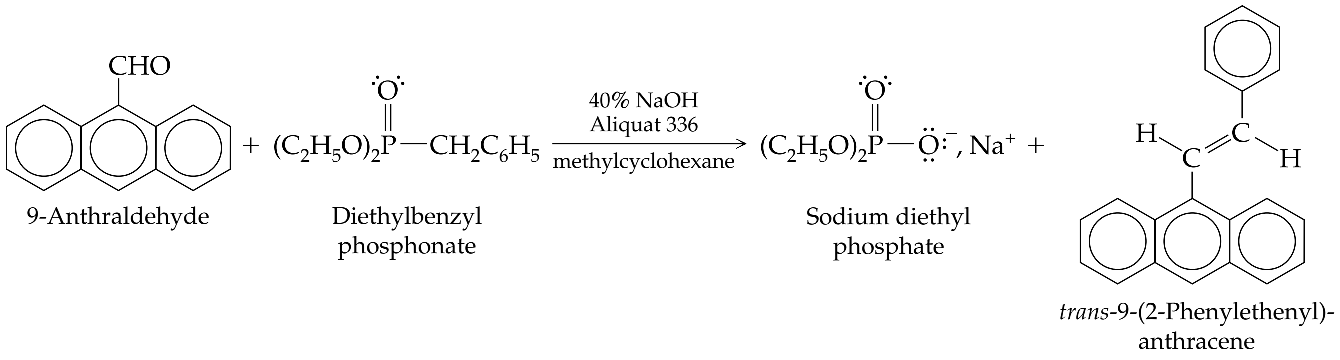
CHO
9-Anthraldehyde
+ $(C_2H_5O)_2P(=O)—CH_2C_6H_5$
Diethylbenzyl phosphonate
40% NaOH
Aliquat 336
methylcyclohexane
$(C_2H_5O)_2P(=O)—O^-$, Na^+ +
Sodium diethyl phosphate
H
C
C
H
trans-9-(2-Phenylethenyl)-anthracene

EXPERIMENTAL PROCEDURE

Estimated time to complete the reaction: 2.5 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
9-Anthraldehyde	206.24	106 mg	0.51	104–105			
Diethyl benzylphosphonate	228.23	120 μL	0.58		106–8 @ 1mm	1.095	1.4970
Aliquat 336 (tricaprylmethyl-ammonium chloride)	404.17	2 drops					
Methylcyclohexane	98.2	2.0 mL			101	0.77	1.4215
40% Aqueous sodium hydroxide		2.0 mL					
trans-(2-Phenylethenyl) anthracene	280.4			130–132			

Reagents and Equipment. Weigh and place 106 mg (0.51 mmol) of 9-anthraldehyde in a 10-mL round-bottom flask containing a magnetic stirrer. Add 2 drops of Aliquat 336 (tricaprylmethylammonium chloride), 120 μL (0.58 mmol) of diethyl benzylphosphonate, 2.0 mL of methylcyclohexane and 2 mL of 40% sodium hydroxide solution. Attach the vial to a reflux condenser (➨).

NOTE. The diethyl benzylphosphonate, methylcyclohexane, and NaOH solution are dispensed in the ***hood*** *using automatic delivery pipets. The Aliquat 336 is very viscous. A medicine dropper is used to dispense this material. It is advisable to lightly grease the bottom joint of the condenser, since strong base is being used.*

HOOD

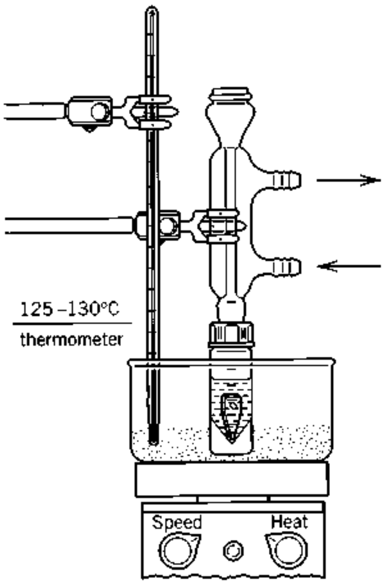

9-Anthraldehyde, 106 mg,
+ diethylbenzyl phosphonate, 120 μL,
+ Aliquat 336, 2 drops, 40% NaOH, 2 mL
+ methylcyclohexane, 2 mL

Reaction Conditions. Heat the two-phase mixture, on a sand bath at a temperature of about 125–130 °C, for 45 min. Stir the reaction mixture vigorously during this period; the upper (organic) layer turns deep red. Allow the mixture to cool to room temperature.

Isolation of Product. Transfer the two-phase solution to a 15-mL glass centrifuge tube using a Pasteur filter pipet. Rinse the flask with 1 mL of methylene chloride and transfer this rinse to the centrifuge tube. Carefully remove the aqueous layer, using a Pasteur filter pipet, and save it in a vial until you have successfully isolated and characterized the product. Wash the organic layer with two 2-mL portions of water. Stir the mixture with a small glass rod after each addition (or a Vortex mixer may be used) and then remove the water layer and save it in the same vial as before. Dry the methylene chloride solution by adding sodium sulfate. Using a Pasteur filter pipet, transfer the dried solution to a tared 10-mL Erlenmeyer flask containing a boiling stone. Wash the sodium sulfate remaining in the centrifuge tube with two 1-mL portions of methylene chloride. Also transfer these washings to the Erlenmeyer flask.

Concentrate the solution, which contains the desired product, almost to dryness on a sand bath under a slow stream of nitrogen gas. Now add approximately 1–2 mL of 2-propanol (isopropanol) to the flask and allow the resulting solution to stand at room temperature for 10–15 min, and then place it in an ice bath to complete the crystallization of the product.

Purification and Characterization. Collect the yellow crystals by vacuum filtration, using a Hirsch funnel, and wash the filter cake with two 1-mL portions of *cold* methanol. Maintain the vacuum for an additional 10 min to partially dry the crystalline product. Then place the material on a clay plate, or on filter paper, and allow it to dry thoroughly. As an alternative, the product may be dried in a vacuum drying oven (or pistol) for 10–15 min at 30 °C (1–2 mm).

The purity of the product may be checked using *thin-layer chromatography* (see Technique 6A). Dissolve a small amount of the starting aldehyde and the product in ethanol and apply a sample to the TLC plate. Use toluene as the elution solvent, silica gel (containing a fluorescent indicator) as the stationary phase, and UV light for visualization. Typical R_f values are 0.92 for *trans*-9-(2-phenylethenyl)anthracene and 0.50 for 9-anthraldehyde.

Weigh the product and calculate the percent yield. A portion of the material may be further purified by recrystallization from 2-propanol using the Craig tube.

Determine the melting point and compare it with the value given in the Reactants and Product table. Obtain the IR spectrum of the alkene product. The material should have a strong absorption band at 962 cm^{-1}, which confirms the presence of a trans double bond.

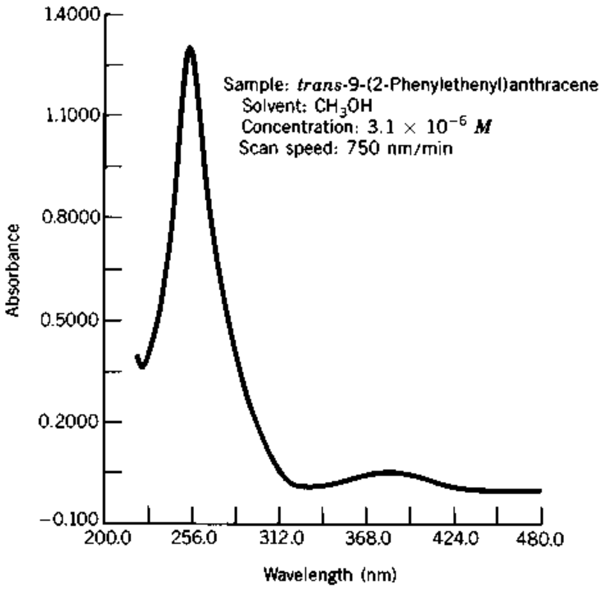

Figure 6.33 UV-visible spectrum: *trans*-9-(2-phenylethenyl)anthracene.

This material has a characteristic UV spectrum in methanol. The following data were obtained at a concentration of 3.1×10^{-6} M (see Fig. 6.33):

λ_{max} 255 nm (ε_{max} = 35,207, methanol)
λ_{max} 383 nm (ε_{max} = 3618, methanol)

QUESTIONS

6-124. Complete each of the following reactions by giving a suitable structure for the species represented by the letters. Give a suitable name for compound B in each reaction.

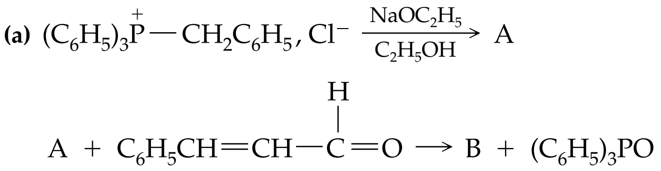

(a) $(C_6H_5)_3\overset{+}{P}-CH_2C_6H_5,\ Cl^- \xrightarrow[C_2H_5OH]{NaOC_2H_5} A$

$A + C_6H_5CH{=}CH-\overset{\overset{\displaystyle H}{|}}{C}{=}O \longrightarrow B + (C_6H_5)_3PO$

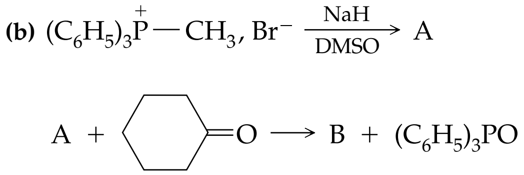

(b) $(C_6H_5)_3\overset{+}{P}-CH_3,\ Br^- \xrightarrow[DMSO]{NaH} A$

6-125. Why is it important that any aldehyde used in the Wittig reaction be free of carboxylic acid impurities?

6-126. Reaction of triphenylphosphine with benzyl bromide produces the corresponding phosphonium salt. Suggest a suitable mechanism for this reaction.

6-127. Heteroatoms other than P are also capable of stabilizing the negative charge on C to yield ylides. For example, nitrogen is capable of forming such a system:

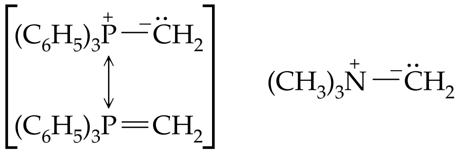

$$\left[\begin{array}{c}(C_6H_5)_3\overset{+}{P}-\overset{-}{\ddot{C}}H_2\\ \updownarrow\\ (C_6H_5)_3P{=}CH_2\end{array}\right]\qquad (CH_3)_3\overset{+}{N}-\overset{-}{\ddot{C}}H_2$$

A phosphorus ylide A nitrogen ylide

Why are resonance structures not drawn for the nitrogen ylide as in the phosphorus system?

6-128. Would you expect that the sulfonium salt, $(C_6H_5)_2CH_3S^+,Br^-$, is capable of forming an ylide when reacted with a strong base? If so, would its structure be best represented in a manner resembling the P or N ylide? Explain.

6-129. Explain why the C=C stretching mode gives rise to a rather weak IR band in 1-methylcyclohexene, while in its isomer, methylenecyclohexene, the band is of medium to strong intensity.

6-130. Predict the C═C stretching frequencies of the alkenes formed when cyclopentanone, cyclobutanone, and cyclopropanone undergo the Wittig reaction with methyltriphenylphosphonium bromide–sodium amide reagent.

6-131. What concentration of a given compound (mol/L) in methanol, having a molecular weight of 165, would have been used if the solution gave an absorbance of 0.68 with a calculated $\varepsilon_{max} = 14{,}800$?

BIBLIOGRAPHY

Of the many articles on the Wittig reaction, four are listed:

Maryanoff, B. E.; Reitz, A. B. *Chem. Rev.* **1989,** *89,* 863.
Silversmith, E. F. *J. Chem. Educ.* **1986,** *63,* 645.
Vedejs, E.; Marth, C. F. *J. Am. Chem. Soc.* **1990,** *112,* 3905.
Wadsworth, W. S., Jr. *Org. React.* **1977,** *25,* 73.

Selected references pertaining to the Horner–Wadsworth–Emmons modification of the Wittig reaction:

Boutagy, J.; Thomas, R. *Chem. Rev.* **1974,** *74,* 87.
Denmark, S. C.; Chen, C-T. *J. Am. Chem. Soc.* **1992,** *114,* 10674.
Horner, L.; Hoffmann, H.; Wippel, H. G.; Klahre, G. *Chem. Ber.* **1959,** *92,* 2499.
Wadsworth, W. S., Jr.; Emmons, W. D. *J. Am. Chem. Soc.* **1961,** *83,* 1733.

Below are selected examples of the Wittig reaction in *Organic Syntheses:*

Campbell, T. W.; McDonald, R. N. *Organic Syntheses*; Wiley: New York, 1973; Collect. Vol. V, p. 985.
Jorgenson, M. J.; Thacher, A. F. *Organic Syntheses*; Wiley: New York, 1973; Collect. Vol. V, p. 509.
Lang, R. W.; Hansen, H. *Organic Syntheses*; Wiley: New York, 1990; Collect. Vol. VII, p. 232.
McDonald, R. N.; Campbell, T. W. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 499.
Nagata, W; Wakabayashi, T.; Hayase, Y. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 358.
Wadsworth, W. S., Jr.; Emmons, W. D. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 547.
Wittig, G.; Schöllkopf, U. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 751.

For a historical perspective of the Wittig reaction, see:

Hofmann, R. W. "Wittig and His Accomplishments: Still Relevant Beyond His 100th Birthday." *Angew. Chem. Int. Ed.* **2001,** *40,* 3915.

EXPERIMENT 20

Aldol Reaction: Dibenzalacetone

Common names: dibenzalacetone, dibenzylideneacetone
CA number: [35225-79-7]
CA name as indexed: 1,4-pentadien-3-one, 1,5-diphenyl-, *(E,E)*-

Purpose. The synthetically useful **aldol reaction** is investigated as a method of forming carbon–carbon bonds. It is a general reaction of aldehydes that may also be extended to ketones. The specific case outlined in this experiment is known as the Claisen–Schmidt reaction. Experiments [$A3_a$] and [F1] provide other examples of the aldol condensation.

Prior Reading

Technique 5: Crystallization
- Introduction (pp. 85–87)
- Use of the Hirsch Funnel (pp. 88–89)
- Craig Tube Crystallization (pp. 89–91)

REACTION

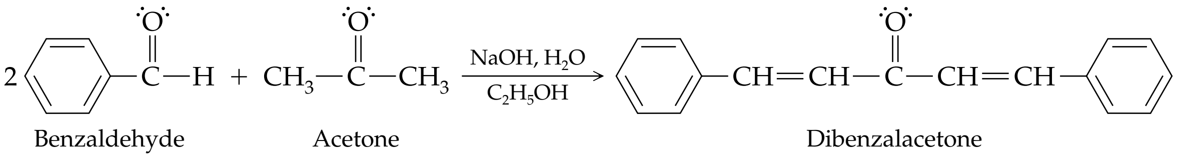

$$2\,C_6H_5{-}\overset{\overset{\large\ddot{O}}{\|}}{C}{-}H + CH_3{-}\overset{\overset{\large\ddot{O}}{\|}}{C}{-}CH_3 \xrightarrow[C_2H_5OH]{NaOH,\ H_2O} C_6H_5{-}CH{=}CH{-}\overset{\overset{\large\ddot{O}}{\|}}{C}{-}CH{=}CH{-}C_6H_5$$

JWCL196_c06_309-375.qxd 11/17/09 1:42 PM Page 310

EQA

DISCUSSION

The aldol reaction (aldol condensation) is one of the fundamental reactions of organic chemistry, since it leads to the formation of a new carbon–carbon bond and is broadly applicable. A condensation reaction is one in which two molecules are joined with the concomitant expulsion of a small stable molecule, usually water or an alcohol. The aldol reaction may be used to condense various combinations of aldehydes and ketones. The mixed aldol condensation of an aldehyde having no α-hydrogen atom with a ketone is specifically known as the Claisen–Schmidt reaction. This variation of the aldol condensation is illustrated here in the synthesis of dibenzalacetone.

The reaction conditions of this aldol condensation favor the formation of the product, dibenzalacetone. This product is insoluble in the aqueous ethanol solvent and precipitates from the reaction as it is formed, whereas the starting materials and the intermediate, benzalacetone, are all soluble in aqueous ethanol. These experimental conditions assist in driving the equilibrium reaction to completion.

The aldol condensation involves generation of an *enolate* by removal of an acidic proton from a carbon to the carbonyl group of an aldehyde or ketone, and subsequent nucleophilic addition of this enolate to the carbonyl carbon of an aldehyde or ketone. The reaction is usually base catalyzed and involves several mechanistic steps.

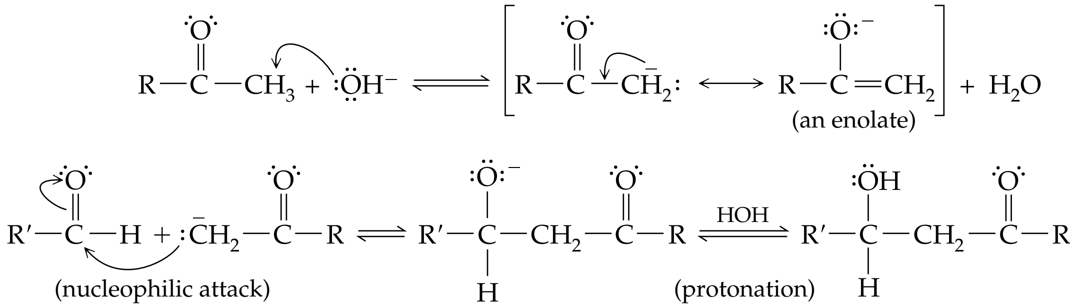

The reaction involves several steps: (1) base-catalyzed generation of an enolate, (2) nucleophilic attack of this anion on a carbonyl carbon, and (3) protonation of the resulting anion to yield the initial aldol product, a β-hydroxy carbonyl compound. Note that each step in the sequence is in equilibrium and the entire reaction is, therefore, reversible. Treatment of the β-hydroxy carbonyl compound with base causes the reverse aldol (**retro-aldol**) reaction to occur.

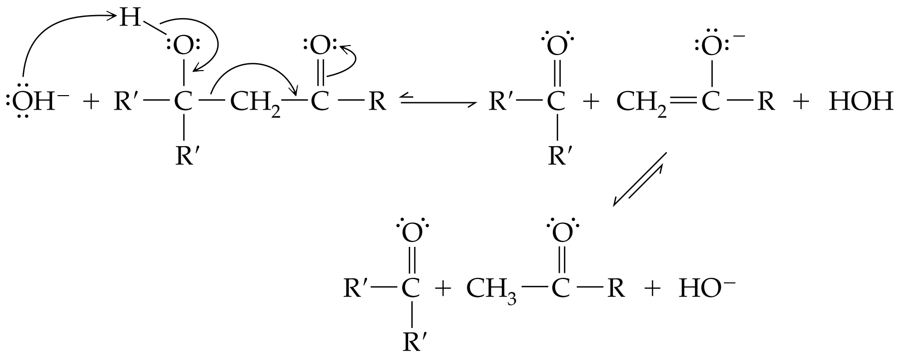

The β-hydroxy carbonyl product may be isolated in most cases, if desired, as the subsequent dehydration is generally much slower than the addition reaction that precedes it. The final stage, as in the present reaction, is a hydroxide-catalyzed dehydration of this initial product by way of its enolate. Though hydroxide ion (HO^-) is generally not a good leaving group, the hydrogen α to the ketone is quite acidic, the elimination produces a rather stable and conjugated α,β-unsaturated ketone, and under strongly basic conditions the hydroxide ion is an adequate leaving group.

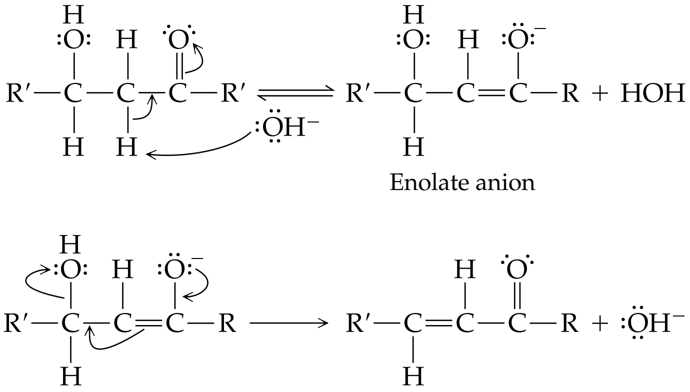

In the present reaction, a double aldol condensation occurs, which yields the dibenzalacetone product. An additional example of the aldol reaction is shown in Experiment [$A3_a$], where tetraphenylcyclopentadienone is prepared.

NOTE. This reaction may be carried out in a 10 × 75-mm test tube. However, the reagents must be stirred efficiently with a glass rod at frequent intervals. If a larger test tube is used, a small magnetic stirring bar or vane is more efficient as an agitator.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Benzaldehyde	106.13	80 μL	0.79	178	1.04	1.5463
Acetone	58.08	29 μL	0.40	56	0.79	1.3588
NaOH catalyst solution		1 mL				

NOTE. It is recommended that the purity of the benzaldehyde be checked by IR. The presence of benzoic acid in the benzaldehyde can substantially lower the yield of product. The benzaldehyde may be purified by distillation under reduced pressure (bp 178–179 °C; 57–59 °C @ 8 torr).

The benzaldehyde may be added to the vial by weight, or by volume (using an automatic delivery pipet). To prevent loss of acetone by evaporation, fit the reaction vial with a septum cap and cool it in an ice bath. Add the acetone (by volume) through the septum using a GC syringe.

Stoichiometric quantities of the reagents are used. An excess of benzaldehyde results in a more intractable product; excess acetone favors the formation of benzalacetone.

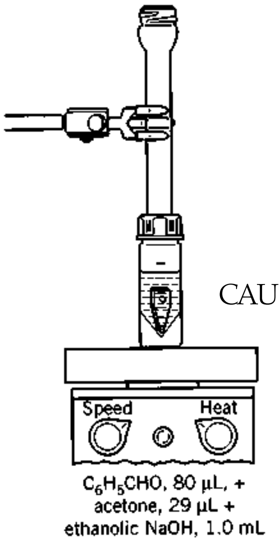

Reagents and Equipment. In a 3.0-mL conical vial containing a magnetic spin vane and equipped with an air condenser place 80 μL (84 mg, 0.79 mmol) of benzaldehyde and 29.0 μL (23 mg, 0.40 mmol) of acetone (←).

Now add to this reaction mixture 1.0 mL of the aqueous, ethanolic sodium hydroxide catalyst solution, delivered from a calibrated Pasteur pipet.

CAUSTIC

INSTRUCTOR PREPARATION. The catalyst solution is prepared by dissolving 0.4 g of sodium hydroxide (which is ***caustic****) in 4.0 mL of water. To this solution add 3.0 mL of 95% ethanol.*

Reaction Conditions. Stir the reaction mixture at room temperature for 30 min. During this time the solid yellow product precipitates from solution.

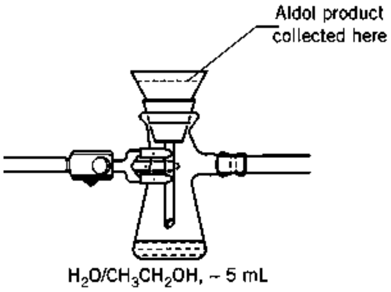

Isolation of Product. Collect the crude yellow dibenzalacetone by vacuum filtration using a Hirsch funnel (←). Remove the magnetic spin vane from the reaction vial with forceps. *Some of the product adheres to the magnetic spin vane. This material should be removed by carefully scraping the vane with a microspatula. The material is added to the product collected by filtration.*

Wash the filter cake with three 1.0-mL portions of water. The filtrate should be nearly neutral, as indicated by pH test paper. If not, repeat the washing until the test indicates that the filtrate is neutral.

NOTE. It is essential to remove the NaOH completely. If it is not removed, the recrystallization step will be difficult.

Air-dry the product by maintaining the suction on the Hirsch funnel for approximately 10 min. During this operation a piece of plastic food wrap may be placed over the mouth of the Hirsch funnel to aid in the drying process (see Prior Reading).

NOTE. An alternative procedure may be used to isolate and purify the dibenzalacetone. Transfer the reaction product directly to a large Craig tube. The washings and the ethanol recrystallization step (see Purification and Characterization section) may then be carried out with no transfer of the material. In this way, loss of product is minimized.

Purification and Characterization. The crude dibenzalacetone may be purified by recrystallization from 95% ethanol, using a Craig tube.

Weigh the dried product and calculate the percent yield. Determine the melting point and compare your result with those in the literature.

Obtain an infrared spectrum of the material and compare your spectrum with that shown in Figure 6.34.

A comparison of the infrared spectra of the starting reagents with that of the product is given below.

Infrared Analysis

Acetone: This simple aliphatic ketone possesses the short macro group frequency train: 3415 (overtone of C═O stretch, $2 \times 1712 = 3424$ cm^{-1}), 3000–2850 (sp^3 C—H stretch), 1712 (C═O stretch), and 1360 cm^{-1} (symmetric

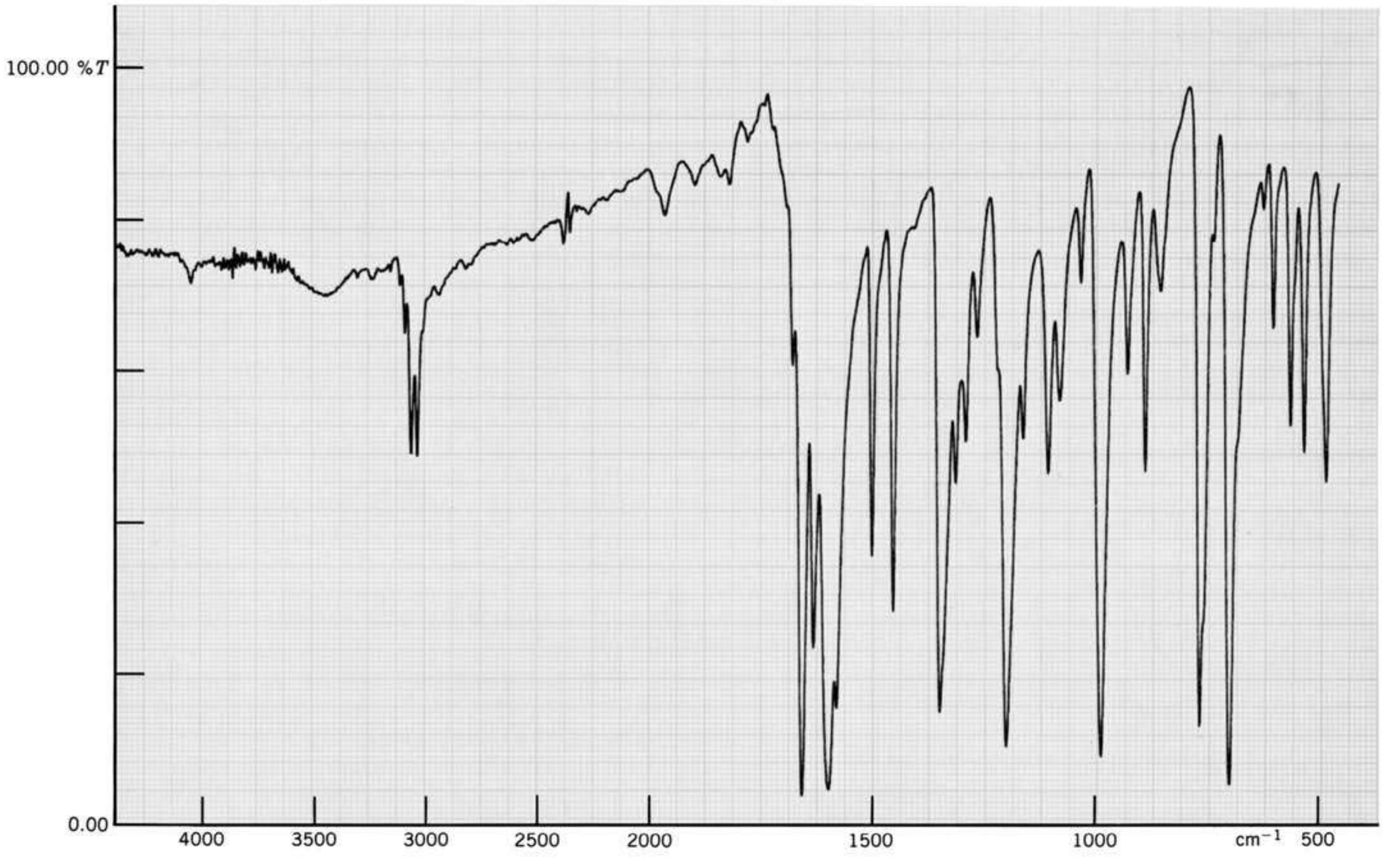

Sample Dibenzalacetone

%T x ABS __ Background Scans 4

Scans 16

Acquisition & Calculation Time 42 sec

Resolution 4.0 cm-1

Sample Condition solid

Cell Window

Cell Path Length

Matrix Material KBr

Figure 6.34 IR spectrum: dibenzalacetone.

methyl bend α to a carbonyl. See discussion of methyl bends in acetates in Experiment [8B]).

Benzaldehyde: The infrared spectrum of this aromatic aldehyde (Fig. 6.35) is rich and interesting. The aromatic aldehyde macro group frequency consists of the following peaks: 3070, 2830 and 2750, 1706, 1602, 1589, and 1396 cm^{-1}.

a. 3070 cm^{-1}: C—H stretch on sp^2 carbon.

b. 2830 and 2750 cm^{-1}: This pair of bands is another example of Fermi coupling (see discussion in Experiment [7] and Infrared Discussions).

c. 1706 cm^{-1}: The carbonyl stretch of the aldehyde group. The frequency observed in aliphatic aldehydes falls in the range 1735–1720 cm^{-1}, but when conjugated, the value drops (see Infrared Discussions) and is found in the range 1720–1700 cm^{-1}.

d. **1602 and 1589 cm^{-1}:** This pair of bands is related to the degenerate ring stretching vibrations ν_{8a} and ν_{8b} of benzene (see infrared discussions in Chapter 10W, Experiment [1B$_{adv}$] and Infrared Discussions).

e. **1396 cm^{-1}:** The aldehyde C—H in-plane bending vibration. The first harmonic of this vibration is Fermi coupled to the aldehyde C—H stretching mode.

Benzaldehyde also possesses a second powerful macro group frequency train that reflects the substitution pattern of the aromatic ring system (see Infrared Discussions). The monosubstituted benzene ring macro group frequency train contains peaks in the following regions: 1950, 1880, 1800, 1730, 750, and 690 cm^{-1}.

a. **1990, 1920, 1830, and 1770 cm^{-1}:** this series of four weak bands with generally decreasing intensity from high-to-low wavenumber values (the third peak near 1800 cm^{-1} may be intensified if the ring is conjugated as in this case) arise from combination bands (see Infrared Discussions), which involve the out-of-plane bending frequencies of the ring C—H bonds (see below). The exact wavenumber positions are not very important, but the overall shape of the pattern can be used to determine the ring substitution pattern (see Infrared Discussions).

b. **750 and 690 cm^{-1}:** this pair of bands is very characteristic of monosubstituted phenyl groups. The 750-wavenumber peak arises from the all-in-phase out-of-plane bending vibration of the five C—H groups adjacent to each other on the ring (see Infrared Discussions). The 690 cm^{-1} companion band involves an out-of-plane displacement of the ring carbon atoms (see Infrared Discussions).

Dibenzalacetone: The infrared spectrum of the product (Fig. 6.35) contains many features of the starting materials, plus new and shifted bands unique to the newly formed structure.

a. The monosubstituted aromatic macro group frequency remains: 1960, 1890, 1815, 1770, 760, and 690 cm^{-1}. In addition, the two pairs of degenerate ring stretching bands are present: 1598, 1580 and 1500, 1450 cm^{-1}. The 1500-wavenumber band is rather weak in benzaldehyde, but intensifies in the product.

b. The ketone carbonyl remains the most intense band in the spectrum but is shifted to 1653 cm^{-1} because of the carbonyl's conjugation to two aromatic rings. The aldehyde's carbonyl stretching mode at 1706 cm^{-1} has vanished.

c. The methyl C—H stretching bands between 3000 and 2850 cm^{-1}, and the coupled aldehyde C—H stretching peaks at 2830 and 2750 have disappeared.

d. New bands appear at 3058 and 3035 cm^{-1} (alkene C—H, stretch) which are overlapped with the aromatic ring C—H stretching) peaks (3075 cm^{-1}). In addition, bands appear at 1627 (conjugated C═C) and 985 cm^{-1} (trans-substituted C═C, C—H in-phase out-of-plane bend).The latter band occurs slightly above its usual location near 965 cm^{-1}. This rise in frequency is the result of conjugation of the double bond to the carbonyl group. For an example of a much more dramatic rise in frequency, refer to the discussion of the cis C—H bending modes in the spectrum of maleic anhydride (Experiment [14]).

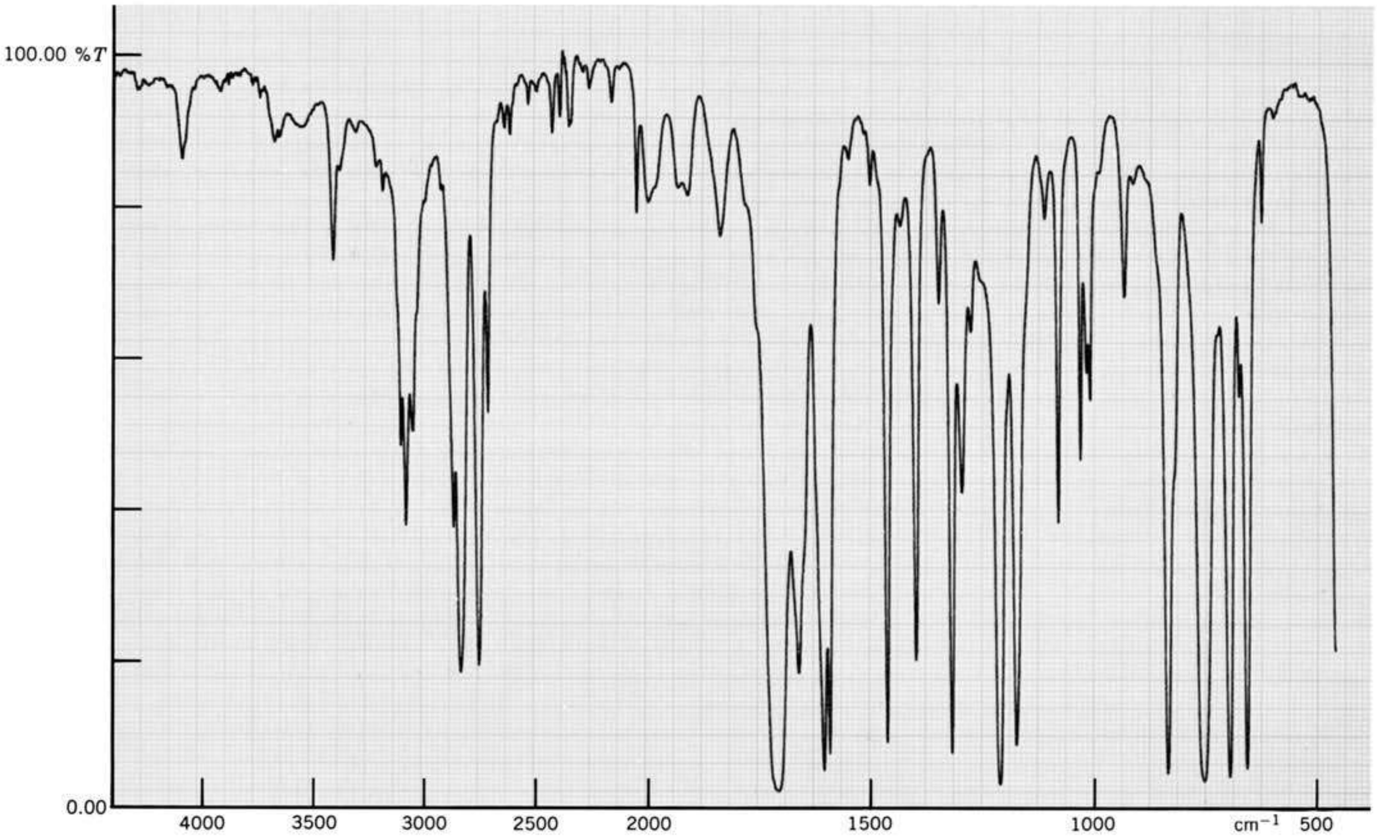

Sample Benzaldehyde
%T X ABS _ Background Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm-1
Sample Condition liquid, neat
Cell Window KBr
Cell Path Length capillary film
Matrix Material

Figure 6.35 IR spectrum: benzaldehyde.

If performed in a potassium bromide matrix, examine the spectrum of your product. Discuss the similarities and differences of the experimentally derived spectral data to the reference spectra (Figs. 6.34 and 6.35).

Dibenzalacetone is a compound that can be characterized by classical chemical tests.

Chemical Tests. An ignition test (See Table 9.1, p. 633) should indicate that dibenzalacentone contains an aromatic ring system. Perform the test to confirm this.

Several classification tests might also be of assistance in classifying this compound. Does the 2,4-dinitrophenylhydrazine test for an aldehyde or ketone give a positive result? Isolate the 2,4-dinitrophenylhydrazone derivative and determine its melting point. Does it correspond to the literature value of 180 °C? What further test could be run to determine whether the carbonyl is present as an aldehyde or ketone?

A test for unsaturation should be enlightening. Would you perform the bromine–methylene chloride or the Baeyer test (Chapter 9, Classification Tests). Did the correct test give a positive result?

OPTIONAL SEMIMICROSCALE PREPARATION

This experiment may be scaled up to be carried out at 5 or 10 times the amounts used in the above micro preparation. The data summarized below are for the 10–fold procedure.

The procedure is identical to the above with the following exceptions.

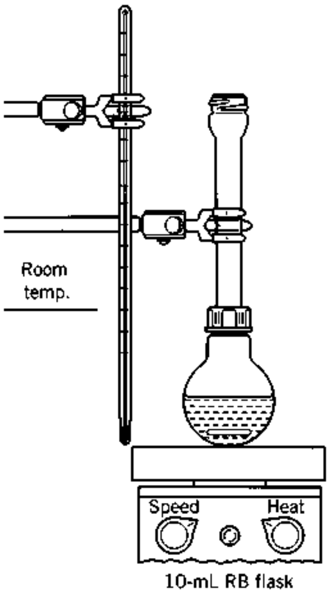

1. Use a 10-mL round-bottom flask fitted with an air condenser (⬅).
2. Increase the reagent and solvent amounts.

Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	d	n_D
Benzaldehyde	106.13	800 μL	7.9	178	1.04	1.5463
Acetone	58.08	300 μL	4.0	56	0.79	1.3588
NaOH catalyst solution		5 mL				

3. The collected filter cake is transferred to a 10-mL beaker, stirred with 5.0 mL of water, and then recollected by vacuum filtration. Repeat this process, usually about three times, until the filtrate is neutral to litmus paper.

QUESTIONS

6-132. A key step in the total synthesis of the hydrocarbon azulene follows. Outline a suitable mechanism to account for the reaction.

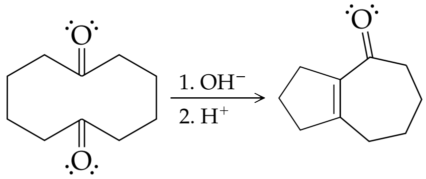

6-133. The aldol reaction has been utilized extensively for the generation of five- and six-membered rings. Suggest a suitable mechanism for the cyclization reactions shown below.

(a)

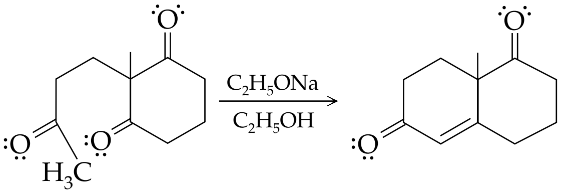

(c)

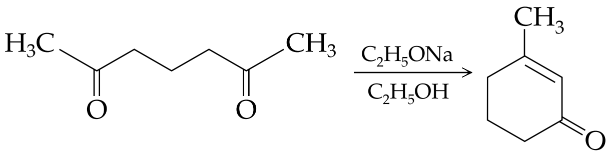

(b)

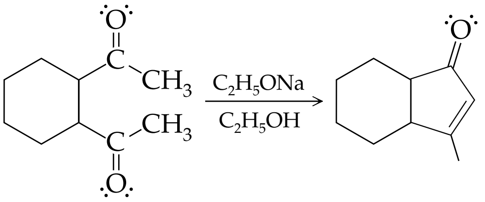

6-134. Predict the major organic product formed in each of the following reactions.

(a) $CH_3CH_2NO_2 + CH_2O \xrightarrow{NaOH}$

(b)

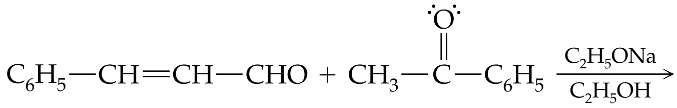

$C_6H_5{-}CH{=}CH{-}CHO + CH_3{-}\overset{\overset{\textstyle\ddot{O}}{\|}}{C}{-}C_6H_5 \xrightarrow[C_2H_5OH]{C_2H_5ONa}$

(c) $C_6H_5CHO + C_6H_5CH_2CN \xrightarrow[C_2H_5OH]{C_2H_5ONa}$

6-135. "Crossed"or "mixed"aldol condensations are practical for synthesis, if one of the aldehydes (or ketones) has no α-hydrogen atoms. Explain.

6-136. Give several examples of aldehydes or ketones that could be used in a "crossed"aldol condensation with propanal. Assign structures and names to the products that could be formed and point out any side reactions that might occur.

6-137. In the aldol condensation using the conditions of this experiment, why might it be essential that the benzaldehyde contain no benzoic acid?

BIBLIOGRAPHY

Review articles:

Cowden, C. J.; Paterson, I. *Org. React.* **1997,** *51,* 1.
Mahrwald, R. *Modern Aldol Reactions,* Vols. 1, 2; Wiley-VCH: New York, 2004.
Mestres, R."A Green Look at the Aldol Reaction."*Green Chemistry;* **2004,** *12,* 586.
Mukaiyama, T. *Org. React.* **1982,** *28,* 203.
Mukaiyama, T.; Kobazachi, S. *Org. React.* **1994,** *46,* 1.
Patai, S.; Rappoport, Z., Eds. *The Chemistry of the Enones;* Wiley: New York, 1989, Part I, Part II.
Smith, M. B.; March. J. *Advanced Organic Chemistry,* 6th ed.; Wiley: New York, 2007, Chap. 16, p. 1339.

Examples of the aldol and Claisen–Schmidt reactions:

Auerbach, R. A.; Crumrine, G. S.; Ellison, D. L.; House, H. O. *Organic Syntheses*; Wiley: New York, 1988; Collect. Vol. VI, p. 692.
Conrad, C. R.; Dolliver, M. A. *Organic Syntheses*; Wiley: New York, 1943; Collect. Vol. II, p. 167.
Hill, G. A.; Bramann, G. M. *Organic Syntheses*; Wiley: New York, 1941; Collect. Vol. I, p. 81.
Kohler, E. P.; Chadwell, H. M. *Organic Syntheses*; Wiley: New York, 1941; Collect. Vol. I, p. 78.
Leuck, G. J.; Cejka, L. *Organic Syntheses*; Wiley: New York, 1941; Collect. Vol. I, p. 283.
Russell, A.; Kenyon, R. L. *Organic Syntheses*; Wiley: New York, 1955; Collect. Vol. III, p. 747.
Wittig, G.; Hesse, A. *Organic Syntheses*; Wiley: New York, 1988; Collect. Vol. VI, p. 901.

EXPERIMENT 21

Quantitative Analysis of Grignard Reagents: 1-Methylbutylmagnesium Bromide and Phenylmagnesium Bromide

Common names: 1-methylbutylmagnesium bromide, 2-pentylmagnesium bromide
CA number: [57325-22-1]
CA name as indexed: magnesium, bromo(1-methylbutyl)-

Common name: phenylmagnesium bromide
CA number: [100-58-3]
CA name as indexed: magnesium, bromophenyl-

Purpose. In this experiment you will generate a Grignard reagent, a common and synthetically useful source of a nucleophilic carbanion, and use an aqueous titration method to determine the amount prepared.

Prior Reading

Moisture-Protected Reaction Apparatus (pp. 25–26)

REACTION

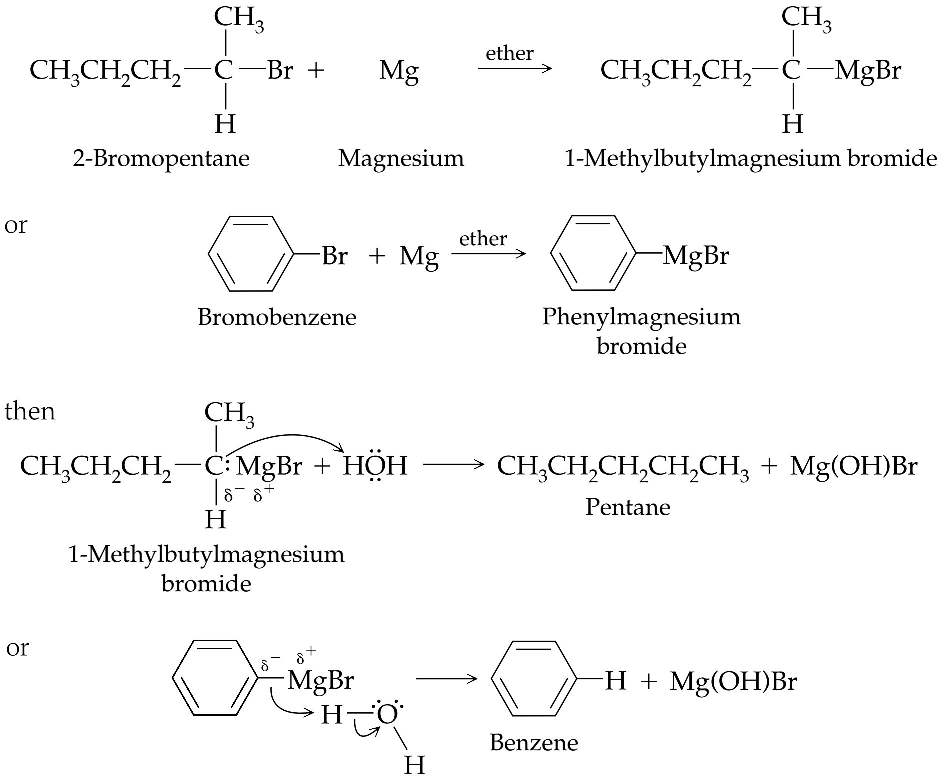
CH3
CH3CH2CH2—C—Br + Mg
H
ether
CH3CH2CH2—C—MgBr
2-Bromopentane
Magnesium
1-Methylbutylmagnesium bromide
or
Br + Mg
ether
MgBr
Bromobenzene
Phenylmagnesium bromide
then
CH3CH2CH2—C: MgBr + HÖH
δ− δ+
CH3CH2CH2CH2CH3 + Mg(OH)Br
1-Methylbutylmagnesium bromide
Pentane
or
δ− δ+
MgBr
H—Ö
H
H + Mg(OH)Br
Benzene

DISCUSSION

This experiment demonstrates the formation of a Grignard reagent (1-methylbutylmagnesium bromide or phenylmagnesium bromide) and a titration method by which the amount of the reagent prepared can be analyzed.

The discovery by Victor Grignard in 1900 that organic halides react with magnesium metal to give organomagnesium compounds was a landmark in organic chemistry. Grignard reagents are among the most useful and versatile reagents in organic synthesis.

The reaction of the Grignard reagent with water is the basis of the analytical method used in this experiment.

$$\mathrm{RMgX + HOH \longrightarrow RH + Mg(OH)X}$$

In the Grignard reagent, the carbon atom bound to the electropositive magnesium atom has a high negative charge density, which is responsible for the strong nucleophilic and basic character exhibited by this organometallic reagent. The carbon, acting as a base, can abstract even a weakly acidic proton from protic reagents, such as water, carboxylic acids, alcohols, and so on. In this process, the corresponding hydrocarbon (the conjugate acid of the R group carbanion) and the basic magnesium halide species are produced. This reaction sequence can be used in the laboratory as a synthetic method to convert organohalides to hydrocarbons. In the examples given in this experiment, 1-methylbutylmagnesium bromide would yield pentane, while phenylmagnesium bromide gives benzene upon protonation.

Titration of the Mg(OH)X species with standardized acid solution makes it possible to determine the amount of Grignard reagent originally formed in the solution.

$$2\ \mathrm{Mg(OH)X + H_2SO_4 \longrightarrow\ + 2\ HOH + MgSO_4 + MgX_2}$$

An excess of the sulfuric acid is generally added to ensure that the Mg(OH)X is completely reacted. The excess acid is then neutralized with standard sodium hydroxide solution. The difference between the total amount of sulfuric acid used and the amount of sodium hydroxide required corresponds to the number of equivalents of the acid actually used to neutralize the Mg(OH)X species. This value is then directly related to the equivalents of Grignard reagent by the reactions given above.

Preparation of the Grignard Reagent[21]

CAUTION: Ether is a flammable liquid. Extinguish all flames during this experiment.

NOTE. All the glassware used in the experiment should be cleaned, dried in an oven at 110 °C for at least 30 min, and then cooled in a desiccator before use.

EXPERIMENTAL PROCEDURE

Estimated time for the experiment: 1.5 h.

Part 1 1-Methylbutylmagnesium Bromide

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
2-Bromopentane	151.05	125 µL	1.0	117	1.21	1.4413
Magnesium	24.31	36 mg	1.5			
Iodine	253.81	1 crystal				
Diethyl ether	74.12	400 µL		34.5		

Prepare the Grignard reagent exactly as described in Experiment [16]. The reagents, amounts of reagents, order of addition, workup manipulations, and precautions are the same. The equipment is also identical (➡).

Cool the gray-colored Grignard reagent mixture to room temperature, and then assay it by the titration method outlined in Part 2.

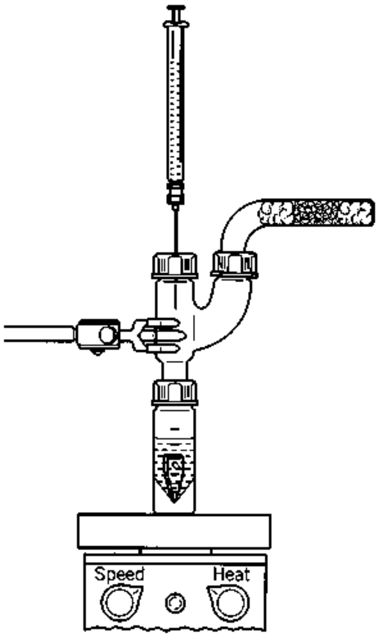

Mg, 36 mg + I_2, 1 crystal + $(CH_3CH_2)_2O$, 400 µL + $CH_3CHBr(CH_2)_2CH_3$, 125 µL

Part 2 Phenylmagnesium Bromide[22]

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Bromobenzene	157.02	76 µL	0.72	156	1.50	1.5597
Diethyl ether	74.12	1.3 mL		34.5	0.73	
Magnesium	24.3	17.5 mg	0.73			
Iodine		1 crystal				

---

[21] For references relating to the preparation of Grignard reagents, see Experiment [16].

[22] This Grignard reagent is prepared exactly as described in Experiment [16]. The reagents, amount of reagents, order of addition, workup manipulations, and precautions are the same. The equipment is also identical.

www *NOTE. If not used for analysis, the solution of Grignard reagent may be treated with various reagents to prepare a wide variety of compounds. For example, see Experiments [16], [17], and $[4_{adv}]$.*

Analysis of the Grignard Reagent: Place 10 mL of freshly boiled distilled water and one drop of phenolphthalein indicator in a 50-mL Erlenmeyer flask. Using a syringe, transfer the cool Grignard reagent solution to the Erlenmeyer flask. Rinse the reaction vial with 0.5 mL of diethyl ether and add the rinse to the Erlenmeyer flask.

NOTE. The addition of water to the Grignard reagent results in the hydrolysis of this reagent to form the corresponding hydrocarbon and a basic magnesium halide, Mg(OH)X. The water is initially boiled to remove any dissolved carbon dioxide that might interfere with the titration.

Make sure that in the transfer of the Grignard reagent solution all small pieces of unreacted magnesium are excluded.

Analysis by Titration: Add, from a 10-mL buret, 5 or 6 mL of standard 0.2 N H_2SO_4 solution to the ethereal Grignard solution. *The resulting solution should be acidic and colorless. If not, add an additional portion of the acid.*

HOOD

Add a boiling stone to the flask and heat the mixture at a sand bath temperature of 90–95 °C in the **hood** for 5 min.

While the solution is still warm, add a drop of phenolphthalein indicator solution and neutralize the excess acid by back titration with 0.1 M NaOH solution. Back titration produces a very light-colored, pink end point. *It may be necessary to add an additional drop of acid, and then more base, to get the best possible end point.*

Data and Calculations: The difference between the initial and final buret readings is the volume of standard acid and base used in the titration of the Grignard reagent.

From the data, calculate the equivalents of Grignard reagent formed. Also, as a percentage, determine the amount of Grignard reagent analyzed compared to its theoretical yield of formation.

QUESTIONS

6-138. Technical grade ether often contains ethanol. Would you recommend this material as a suitable solvent for the preparation of Grignard reagents? If not, why not?

6-139. It is likely that the amount of Grignard reagent your analysis indicates was formed is greater than the amount of Grignard reagent actually present just before you added water. Explain.

6-140. What hydrocarbon would you expect to obtain by the action of water on each of the Grignard reagents listed below?

- **(a)** Butylmagnesium bromide
- **(b)** *sec*-Butylmagnesium bromide
- **(c)** Isobutylmagnesium bromide
- **(d)** *tert*-Butylmagnesium bromide

6-141. What product would each of the Grignard reagents in question 6-140 yield when treated with D_2O?

6-142. **(a)** What product would each of the Grignard reagents listed in Question 6–140 yield when treated with ethanol?

- **(b)** With isopropyl alcohol?
- **(c)** Explain why 4-hydroxycyclohexanone is not a viable candidate when considering Grignard reagents?

6-143. The solubility of Grignard reagents in ether plays a crucial role in their formation. The reagents are soluble because the magnesium is coordinated to the ether oxygen in a Lewis acid–base interaction. Each ether molecule donates an electron pair to the magnesium to complete an octet.

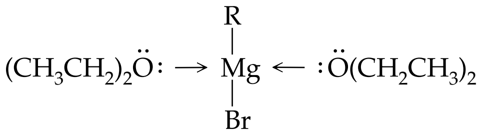

$$(CH_3CH_2)_2\ddot{O}\colon \rightarrow \underset{\displaystyle Br}{\overset{\displaystyle R}{Mg}} \leftarrow \colon\ddot{O}(CH_2CH_3)_2$$

Grignard reagents are normally insoluble in hydrocarbon solvents. However, they can be rendered soluble by the addition of a tertiary amine to the hydrocarbon–Grignard reagent mixture. Explain.

BIBLIOGRAPHY

For the many references related to the preparation and use of the Grignard reagent cited in *Organic Syntheses*, see the Reaction Indexes in Collected Volumes I–X under Grignard Reactions.

EXPERIMENT 22

Williamson Synthesis of Ethers

Common names: propyl *p*-tolyl ether, 4-propoxytoluene
CA number: [5349-18-8]
CA name as indexed: benzene, 1-methyl-4-propoxy-

Common names: methyl *p*-ethylphenyl ether, *p*-ethylanisole
CA number: [1515-95-3]
CA name as indexed: benzene, 1-ethyl-4-methoxy-

Common names: butyl *p*-nitrophenyl ether
CA number: [7244-78-2]
CA name as indexed: benzene, 1-butoxy-4-nitro-

Purpose. The conditions under which ethers are prepared are explored by the well-known Williamson ether synthesis. You will prepare alkyl aryl ethers by S_N2 reactions of alkyl halides with substituted phenoxide anions. The use of phase-transfer catalysis is demonstrated.

Prior Reading

Technique 4: Solvent Extraction
Liquid–Liquid Extraction (p. 72)
Technique 6: Chromatography
Column Chromatography (pp. 92–95)
For Optional Scaleup: Separatory Funnel Extraction (pp. 75–77)

REACTION

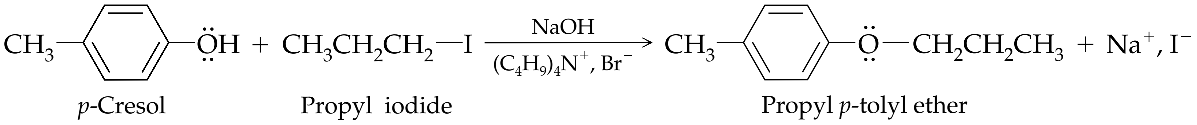

DISCUSSION

The compounds whose preparations are described in Experiments [22A], [22B], [22C], and [22D] are alkyl aryl ethers. The general method of preparation is the Williamson synthesis, an S_N2 reaction specifically between a phenoxide ion (ArO^-) nucleophile and an alkyl halide. This reaction is often used for the synthesis of symmetrical and unsymmetrical ethers where at least one of the ether carbon atoms is primary or methyl, and thus amenable to an S_N2 reaction. Elimination (E2) is generally observed if secondary or tertiary halides are used, since phenoxide ions are also bases.

The conditions under which these reactions are conducted lend themselves to the use of phase-transfer catalysis. The reaction system involves two phases: the aqueous phase and the organic phase. In the present case, the alkyl halide reactant acts as the organic solvent, as does the product formed. The phase-transfer catalyst plays a very important role. In effect, it carries the phenoxide ion, as an ion-pair, from the aqueous phase, across the phase boundary into the organic phase, where the S_N2 reaction then occurs. The ether product and the corresponding halide salt of the catalyst are produced in this reaction. The halide salt then migrates back into the aqueous phase, where the halide ion is exchanged for another phenoxide ion, and the process repeats itself. The catalyst can play this role, since the large organic groups (the four butyl groups) allow the solubility of the ion-pair in the organic phase, while the charged ionic center of the salt renders it soluble in the aqueous phase. For further discussions of phase-transfer catalysis, see Experiments [19B] and [19D].

In the reactions described below, the mechanism is a classic S_N2 process, and involves a backside nucleophilic attack of the phenoxide anion on the alkyl halide.

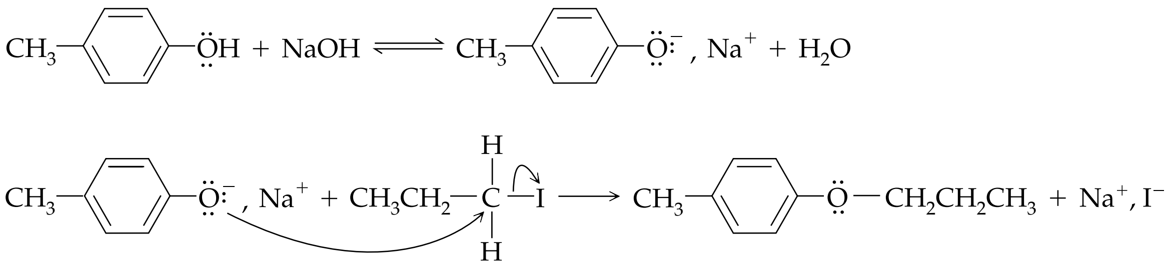

It is of interest to contrast the acidity of phenols with that of simple alcohols. A phenol is more acidic than an alcohol. In a typical aliphatic alcohol (e.g., ethanol) loss of the proton forms a strong anionic base, alkoxide ion (ethoxide ion).

$$R-CH_2-OH \rightleftharpoons H^+ + \underset{\text{Alkoxide ion}}{R-CH_2-O^-}$$

The strongly basic characteristics of the alkoxide species are due to the fact that the negative charge is localized on the oxygen atom. Ethanol has a $pK_a = 16$. In contrast, the conjugate base of a phenol can delocalize its negative charge.

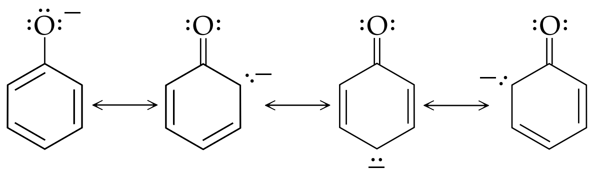

Phenoxide ion

The phenoxide ion is stabilized by this resonance delocalization; therefore, it is a weaker base than the alkoxide ion. Conversely, the phenol is a stronger acid than a typical aliphatic alcohol. Phenol has a pK_a = 10 and is thus 1 million times more acidic than ethanol.

Propyl *p*-Tolyl Ether

Experiment 22A

The reaction for Experiment [22A] is shown above.

EXPERIMENTAL PROCEDURE

Estimated time of the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
p-Cresol	108.15	160 μL	1.56	32–34	202	1.02	1.5312
25% NaOH solution		260 μL					
Tetrabutylammonium bromide	322.38	18 mg	0.056	103–104			
Propyl iodide	169.99	150 μL	1.54		102	1.75	1.5058

Reagents and Equipment. Weigh and place 160 μL (168 mg, 1.56 mmol) of *p*-cresol in a 5.0-mL conical vial containing a magnetic spin vane. Now add 260 μL of 25% aqueous sodium hydroxide and thoroughly mix the resulting solution (➡). To this solution weigh and add the tetrabutylammonium bromide ($Bu_4N^+Br^-$) catalyst (18 mg), followed by 150 μL (262 mg, 1.54 mmol) of propyl iodide. Immediately attach the vial to a reflux condenser.

NOTE. Warm the cresol in a hot water bath to melt it. Dispense this reagent and the propyl iodide in the **hood** *using an automatic delivery pipet.*

HOOD

CAUTION: Propyl iodide is a cancer suspect agent.

Reaction Conditions. Place the reaction vessel in a sand bath and stir vigorously at 110–115 °C for 45–60 min.

Isolation of Product. Cool the resulting two-phase mixture to room temperature, and remove the spin vane with forceps. Rinse the spin vane with 1.0 mL of diethyl ether, adding the rinse to the two-phase mixture. Cap the vial, agitate, vent, and transfer the bottom aqueous layer, using a Pasteur filter pipet, to a 3.0-mL conical vial. A Vortex mixer, if available, can be used to good advantage in this extraction step. Wash this aqueous fraction with 1.0 mL of diethyl ether. Save this and all subsequent aqueous fractions together in a small Erlenmeyer flask until your final product has been isolated and characterized. Now transfer this diethyl ether wash to the 5-mL conical vial containing the ether solution of the product. Extract the resulting ether solution with a 400-μL portion of 5% aqueous sodium hydroxide solution. Cap the vial, agitate, vent, and remove and save the bottom aqueous layer, using a Pasteur filter pipet. Wash the product–ether solution with 200 μL of water. Remove, and save, the aqueous phase to obtain the crude, wet ether solution of the product. Add a boiling stone to the vial and concentrate the solution in a warm sand bath under a gentle stream of nitrogen to isolate the crude product.

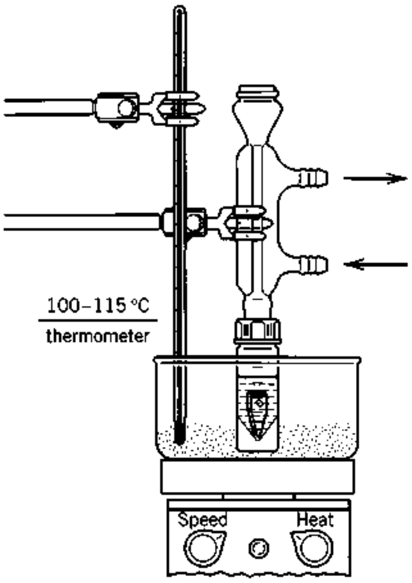

p-$CH_3C_6H_4OH$, 160 μL +
25% NaOH, 260 μL +
$Bu_4\overset{+}{N}\,\overset{-}{Br}$, 18 mg
+ PrI, 150 μL

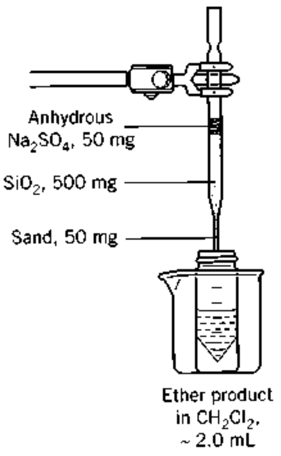
Anhydrous Na_2SO_4, 50 mg
SiO_2, 500 mg
Sand, 50 mg
Ether product in CH_2Cl_2, ~ 2.0 mL

Purification and Characterization. The crude product is purified by chromatography on silica gel. Prepare a microchromatographic column by placing 500 mg of activated silica gel in a Pasteur filter pipet, followed by 50 mg of anhydrous sodium sulfate (⬅). Dissolve the crude product in 250 μL of methylene chloride and transfer the resulting solution to the dry column by use of a Pasteur pipet. Elute the material with 2.0 mL of methylene chloride and collect the eluate in a tared 3.0-mL conical vial containing a boiling stone. Evaporate the solvent by placing the vial in a sand bath maintained at a temperature of 60–65 °C.

Weigh the pure propyl *p*-tolyl ether and calculate the percent yield. Determine the boiling point and density (optional) and compare the experimental values with those in the literature.

Obtain an IR spectrum of the compound and compare it to that shown in Figure 6.36 for 4-propoxytoluene (propyl *p*-tolyl ether).

Nuclear Magnetic Resonance Analysis. If facilities permit, you can obtain both ^1H and ^{13}C NMR spectra of your propyl *p*-tolyl ether in $CDCl_3$ and compare your spectra with those in Figures 6.37 and 6.38. There are two extraneous peaks in the ^1H spectrum: the small singlet at 7.24 ppm is due to residual $CHCl_3$ in the $CDCl_3$, and the small singlet at 1.55 ppm is probably due to a trace amount of H_2O in either the sample or the NMR solvent. The 1:1:1 triplet at 77 ppm in the ^{13}C spectrum is from the $CDCl_3$ solvent.

Since the ^1H spectrum is entirely first order, it can be readily interpreted. You should be able to use the splitting patterns to assign peaks to each of the different groups of protons in the molecule. The integration can assist you.

Chemical Tests. Qualitative chemical tests can also be used to assist in characterizing this compound as an ether. Perform the ignition test (Table 9.1) to determine whether the material contains an aromatic ring.

A key factor to investigate is the solubility characteristics of this material (see Chapter 9). Determine its solubility in water, 5% sodium hydroxide, 5% hydrochloric acid, concentrated sulfuric acid, and 85% phosphoric acid. Do the results place this compound in the solubility class of an ether containing less than 8, or more than 8, carbon atoms? Does the ferrox test (Chapter 9) confirm the presence of oxygen in the compound?

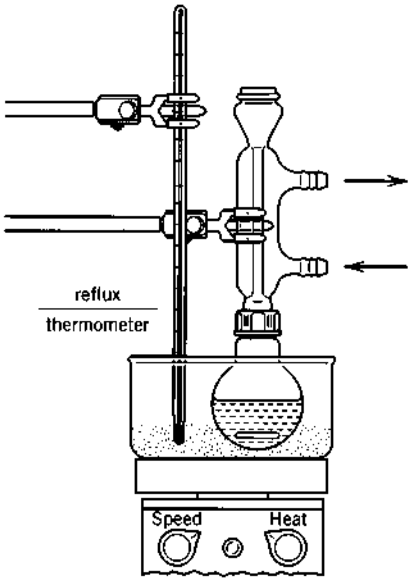

Reflux conditions
100-mL RB flask

OPTIONAL MACROSCALE PREPARATION

This ether may be prepared on a larger scale (~100-fold increase) using a procedure similar to that just outlined, with the following modifications.

1. Use a 100-mL round-bottom flask containing a magnetic stirrer and equipped with a reflux condenser (⬅).

2. The reagent and solvent amounts are summarized in the following table. Note that propyl bromide is used in place of propyl iodide at this scale.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
p-Cresol	108.15	16.3 g	0.15	32–34	202	1.02	1.5312
NaOH	40.0	6.05 g	0.15				
Tetrabutyl ammonium bromide	322.38	0.41 g	0.0012	103–104			
Propyl bromide	122.99	12.71 g	0.1		71	1.35	1.4341
Water		25 mL					

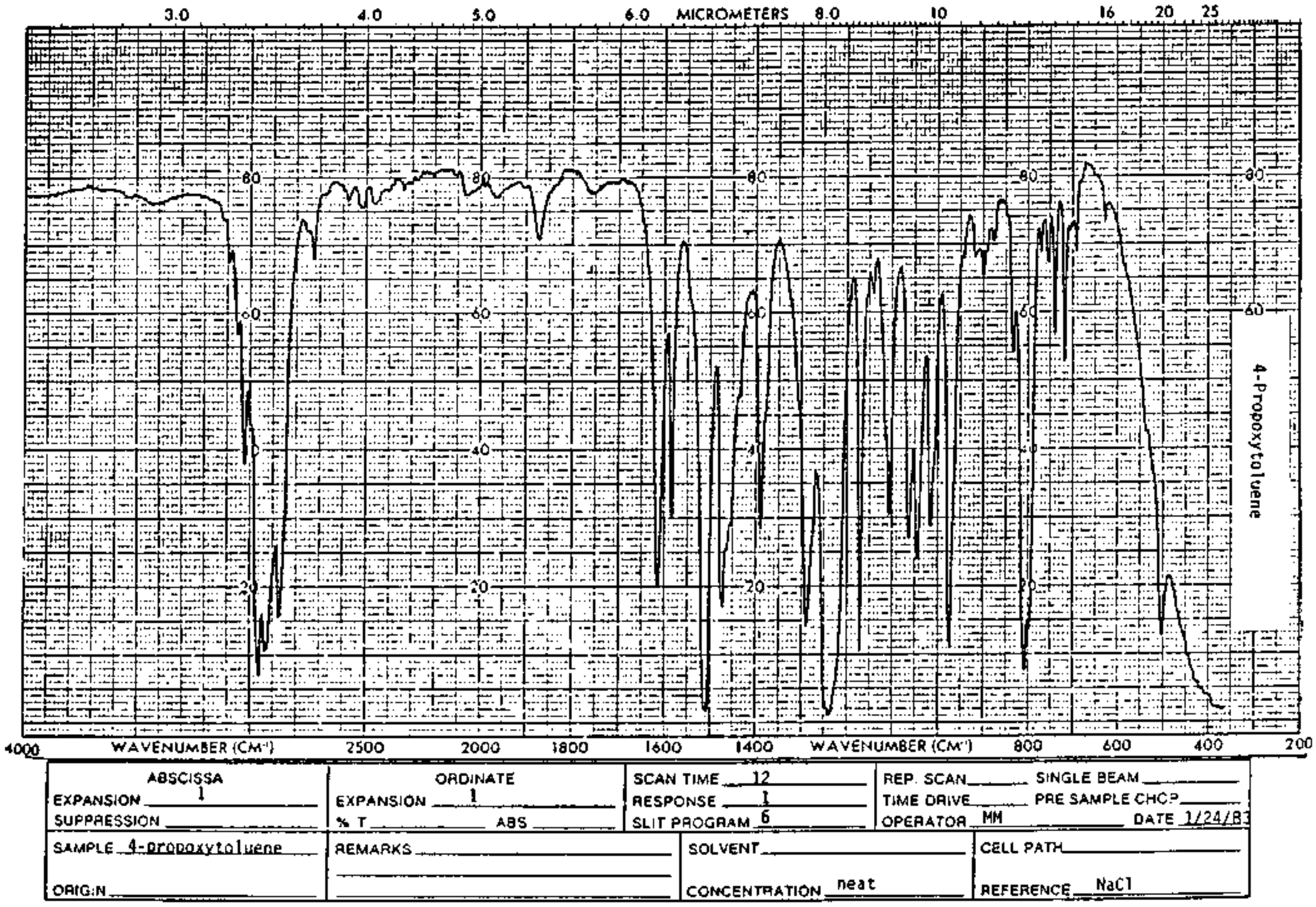

ABSCISSA
EXPANSION 1
SUPPRESSION

ORDINATE
EXPANSION 1
% T ABS

SCAN TIME 12
RESPONSE 1
SLIT PROGRAM 6

REP. SCAN SINGLE BEAM
TIME DRIVE PRE SAMPLE CHOP
OPERATOR MM DATE 1/24/83

SAMPLE 4-propoxytoluene
ORIGIN
REMARKS
SOLVENT
CONCENTRATION neat
CELL PATH
REFERENCE NaCl

Figure 6.36 IR spectrum: propyl *p*-tolyl ether.

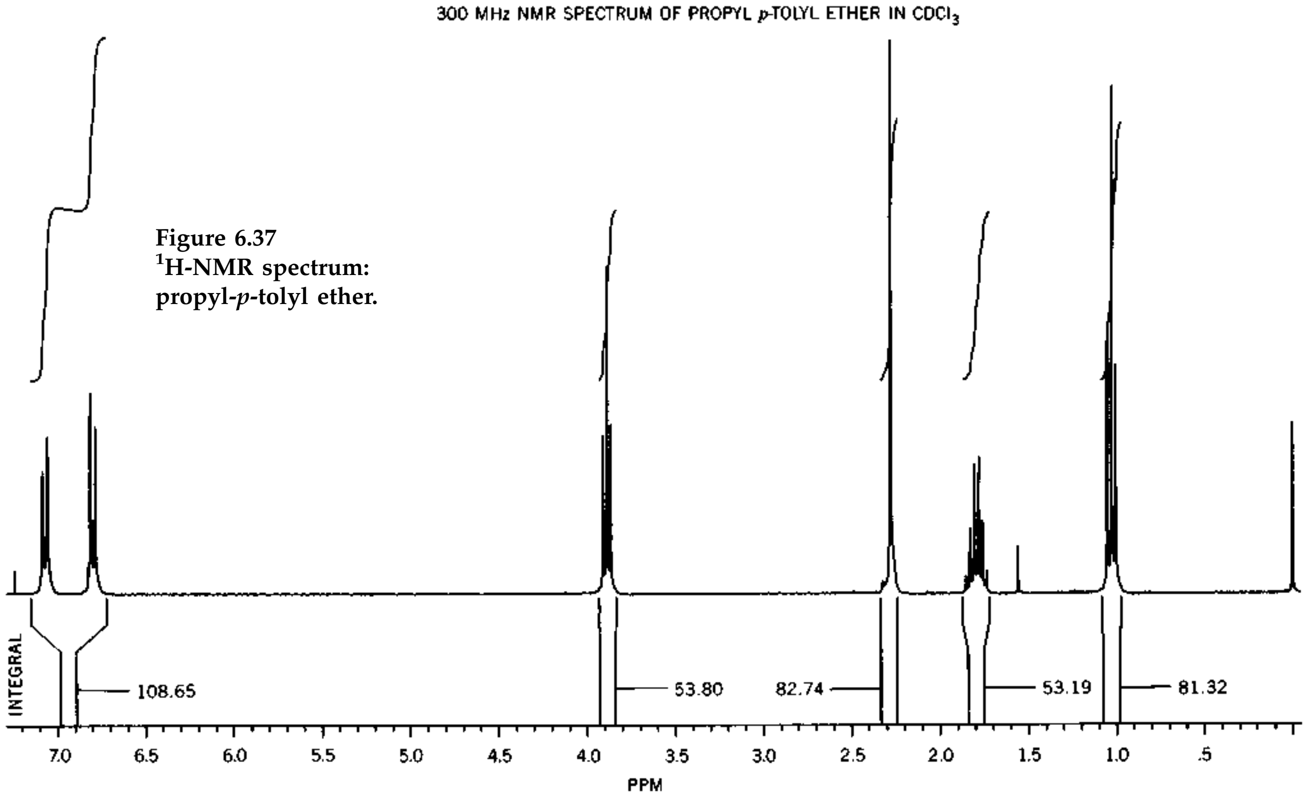

Figure 6.37
^{1}H-NMR spectrum:
propyl-*p*-tolyl ether.

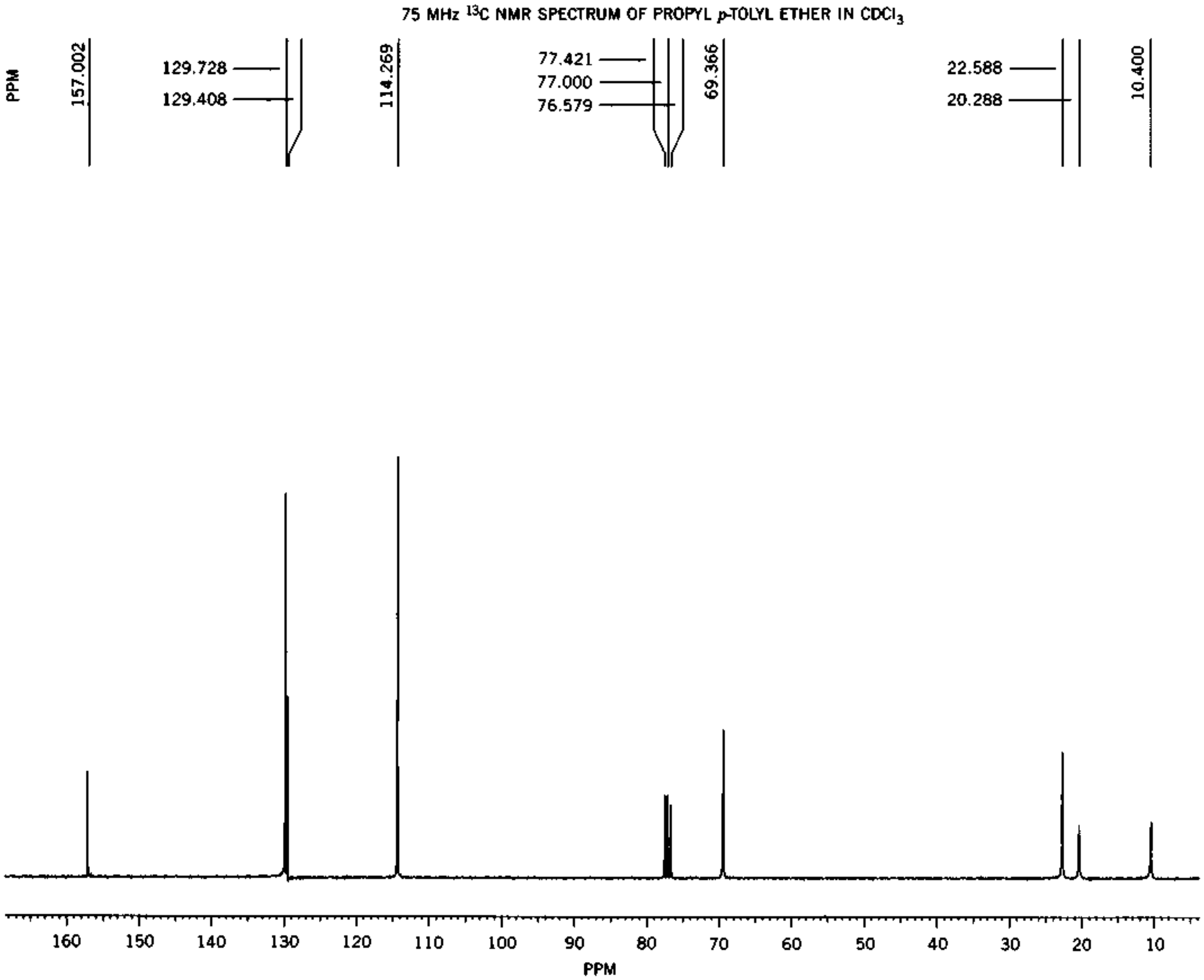

Figure 6.38 ^{13}**C-NMR spectrum: propyl *p*-tolyl ether.**

3. Stir the reaction mixture at reflux temperature for 90 min.

Isolation of Product. Allow the solution to cool to room temperature, transfer it to a 125-mL separatory funnel, and remove the aqueous layer. Store the aqueous layer in a 125-mL Erlenmeyer flask. Wash the organic layer successively with 5% NaOH solution (20 mL) and distilled H_2O (20 mL). After each washing, remove the aqueous phase and add it to the 125-mL Erlenmeyer flask, which should be kept until the final product has been purified and characterized. Collect the remaining deep-red organic layer in a 125-mL Erlenmeyer flask, and dry it over anhydrous sodium sulfate.

Remove the drying agent by filtration through a glass wool plug and collect the product ether in a tared container. Weigh, calculate the percent yield, and then purify and characterize a small amount of the material as described in the microscale procedure.

Methyl *p*-Ethylphenyl Ether

Experiment 22B

REACTION

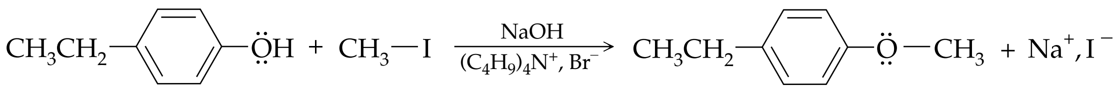

4-Ethylphenol Methyl iodide Methyl *p*-ethylphenyl ether

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
4-Ethylphenol	122.17	150 mg	1.2	42–45			
25% NaOH solution		250 μL					
Tetrabutylammonium bromide	322.38	15 mg	0.047	103–104			
Methyl iodide	141.94	90 μL	1.45		41–43	2.28	1.5304

Reagents and Equipment. Use the same apparatus as in Experiment [22A] for this synthesis. Weigh and add 150 mg (1.2 mmol) of 4-ethylphenol to the reaction vial followed by 250 μL of 25% aqueous sodium hydroxide solution (➡). Stir the mixture at room temperature until dissolution occurs. The phase-transfer catalyst (tetrabutylammonium bromide ($Bu_4N^+Br^-$), 15 mg, 0.05 mmol) is now added, followed by 90 μL (205 mg, 1.45 mmol) of methyl iodide.

NOTE. Methyl iodide is toxic and must be dispensed in the ***hood.*** *Dispense both the alkaline solution and the methyl iodide using an automatic delivery pipet. Because of its volatility, methyl iodide is used in slight excess.*

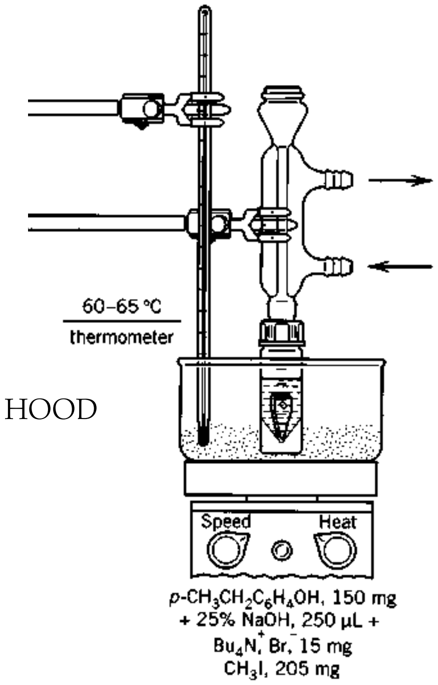
60–65 °C
thermometer
HOOD
Speed
Heat
p-$CH_3CH_2C_6H_4OH$, 150 mg
+ 25% NaOH, 250 μL +
$Bu_4\overset{+}{N}$, $\bar{Br}$, 15 mg
CH_3I, 205 mg

Reaction Conditions. Place the reaction assembly in a sand bath maintained at 60–65 °C and stir the mixture for 1 h.

Isolation of Product. Work up the resulting product mixture as described in Experiment [22A].

Purification and Characterization. Purify the crude product by chromatography on silica gel as described in Experiment [22A], Purification and Characterization (➡).

Weigh the pure methyl *p*-ethylphenyl ether and calculate the percent yield. Determine the boiling point and density (optional). Compare your results with the values reported in the literature.

Obtain the IR spectrum of the compound and compare it with that in Figure 6.39.

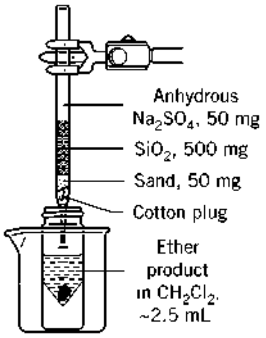
Anhydrous Na_2SO_4, 50 mg
SiO_2, 500 mg
Sand, 50 mg
Cotton plug
Ether product in CH_2Cl_2, ~2.5 mL

Nuclear Magnetic Resonance Analysis. If facilities permit, you can obtain both ^1H and ^{13}C NMR spectra of your methyl *p*-ethylphenyl ether in

$CDCl_3$, and compare your spectra with those in Figures 6.40 and 6.41. There are two extraneous peaks in the ^1H spectrum: The small singlet at 7.24 ppm is due to residual $CHCl_3$ in the $CDCl_3$ and the small singlet at 1.55 ppm is probably due to a trace amount of water in either the sample or the NMR solvent. The 1:1:1 triplet at 77 ppm in the ^{13}C spectrum is from the $CDCl_3$ solvent.

Since the ^1H spectrum is entirely first order, it can be readily interpreted. You should be able to use the splitting patterns to assign peaks to each of the different groups of protons in the molecule. The integration can assist you.

Chemical Tests. Qualitative chemical tests can also be used to assist in characterizing this compound as an ether. Perform the ignition test (Table 9.1) to determine whether the material contains an aromatic ring.

A key factor to investigate is the solubility characteristics of this material (see Chapter 9). Determine its solubility in water, 5% sodium hydroxide, 5% hydrochloric acid, concentrated sulfuric acid, and 85% phosphoric acid. Do the results place this compound in the solubility class of an ether containing fewer than 8, or more than 8, carbon atoms?

Does the ferrox test (Chapter 9) confirm the presence of oxygen in the compound?

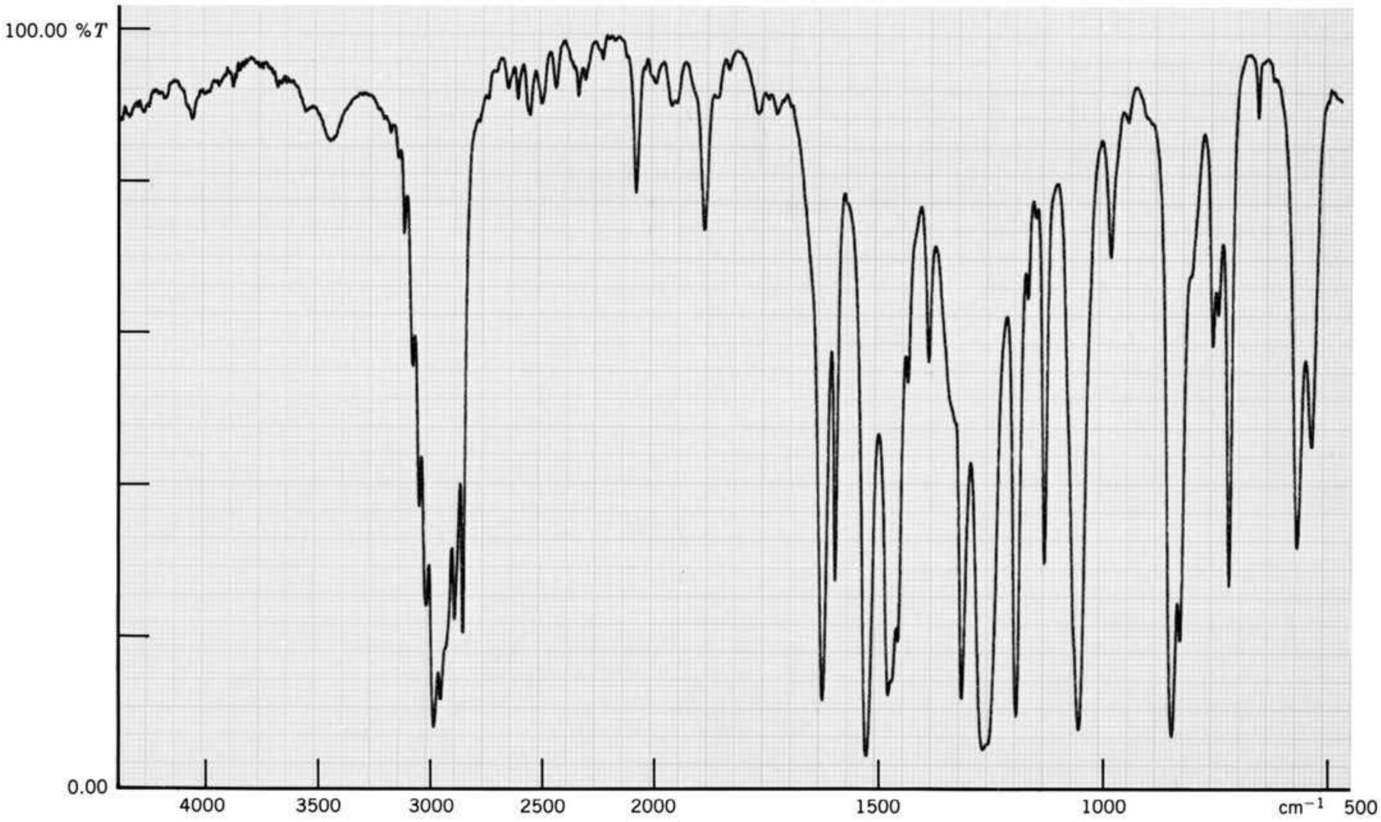

Sample 4-Ethylanisole	
%T x ABS _ Background Scans 4	Scans 4
Acquisition & Calculation Time 42 sec	Resolution 4.0 cm-1
Sample Condition liquid, neat	Cell Window KBr
Cell Path Length capillary film	Matrix Material

Figure 6.39 IR spectrum: methyl *p*-ethylphenyl ether.

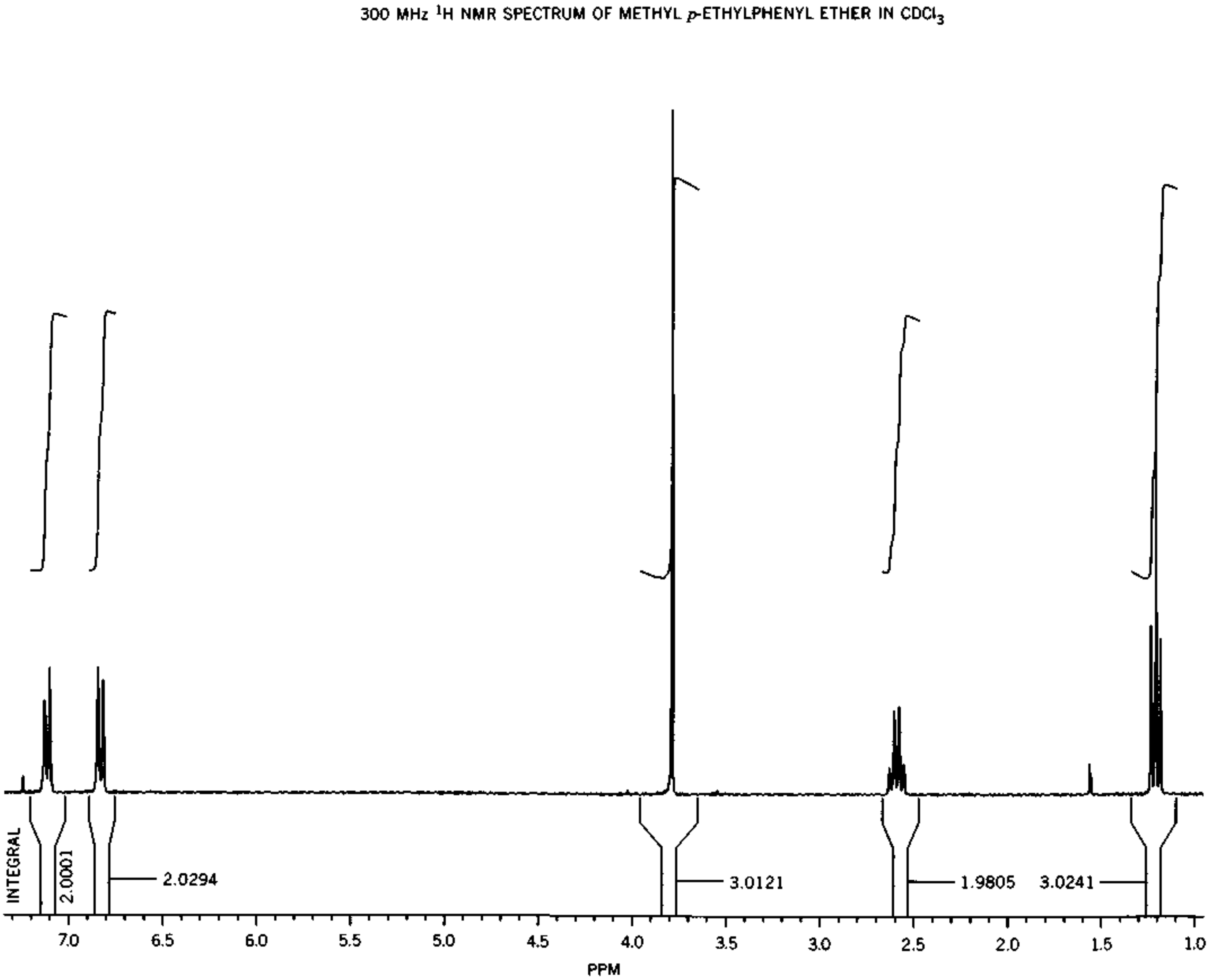

Figure 6.40 **^{1}H-NMR spectrum: methyl *p*-ethylphenyl ether.**

OPTIONAL SEMIMICROSCALE AND MACROSCALE PREPARATIONS

If desired, this experiment can be scaled up by a factor of 10 or more.

Tenfold Scaleup. The reagent and solvent amounts are given in the following table.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
4-Ethylphenol	122.17	1.5 g	12.3	42–45			
NaOH	40.0	625 mg	15.6				
Tetrabutylammonium bromide	322.38	150 mg	0.47	103–104			
Methyl iodide	141.94	2.05 g	14.5		41–43	2.28	1.5304
Water		2.5 mL					

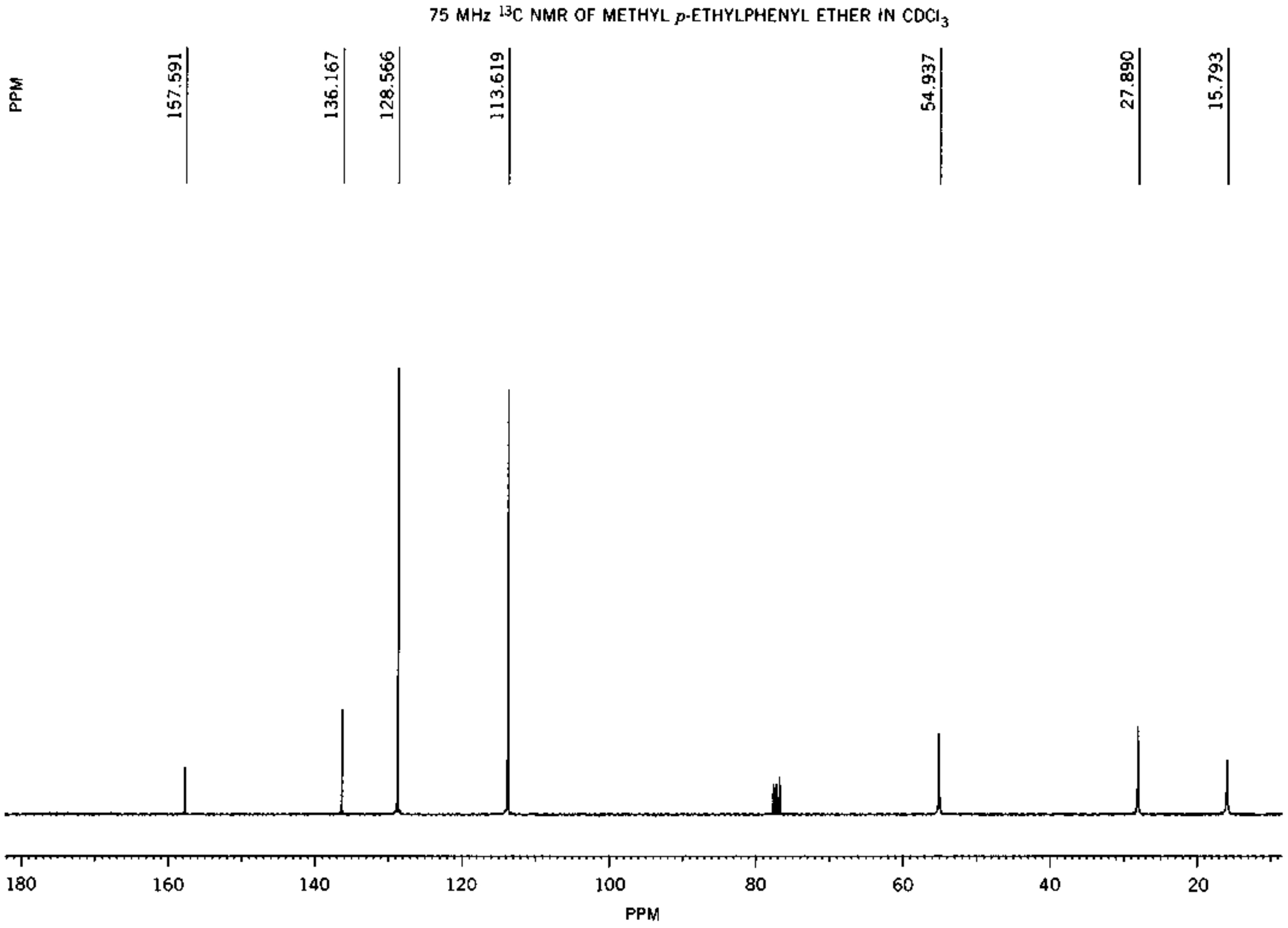

Figure 6.41 ^{13}**C-NMR spectrum: methyl *p*-ethylphenyl ether.**

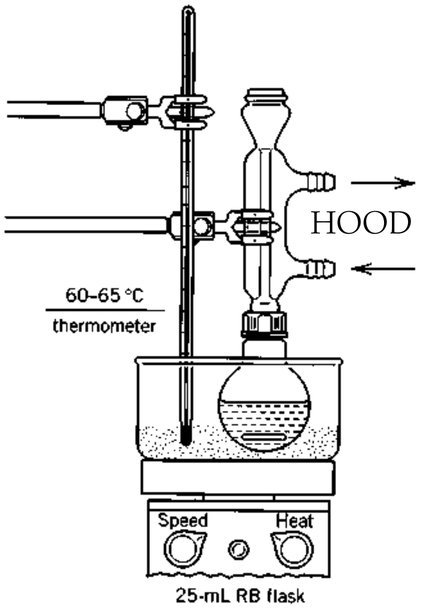

Reagents and Equipment. In a 25-mL round-bottom flask equipped with a stirring bar and a reflux condenser, place 2.5 mL of water and 1.5 g (12 mmol) of *p*-ethylphenol (the phenol will not be water soluble). Cool the mixture by immersing the flask in a beaker of cold water, and then, with stirring, **cautiously** add 625 mg of sodium hydroxide.

After dissolution of the sodium hydroxide, weigh and add 150 mg of the phase-transfer catalyst, tetrabutylammonium bromide, followed by 2.05 g (900 μL) of methyl iodide (in the **hood**) using an automatic delivery pipet (⬅).

Reaction Conditions. Place the reaction vessel in a sand bath, and heat the reaction mixture, with stirring, at 60–65 °C for 1 h.

Isolation of Products. Cool the resulting two-phase mixture to room temperature and then transfer it by Pasteur pipet to a 15-mL centrifuge tube. Wash the reaction flask with three 1-mL portions of diethyl ether and transfer the washings to the centrifuge tube. Cap the tube, shake, and vent (this operation may be done using a Vortex mixer if available), and allow the layers to separate. Remove the lower aqueous layer using a Pasteur filter pipet, and save it in a 10-mL Erlenmeyer flask until the final product has been purified and characterized.

NOTE. Do not remove any precipitated material that settles between the two layers.

Extract the ether layer with one 2-mL portion of 5% sodium hydroxide solution and then with 1 mL of water. Add these washings to the 10-mL Erlenmeyer flask.

Purify the crude product by chromatography on silica gel. Prepare a column by placing 5 g of activated silica gel in a 25-mL buret, followed by 0.5 g of anhydrous sodium sulfate. Dissolve the wet ether extract obtained above in 2.5 mL of methylene chloride and transfer the resulting solution by Pasteur pipet to the dry column. Elute the sample with an additional 20 mL of methylene chloride. Collect the eluate in a tared 50-mL Erlenmeyer flask containing a boiling stone. Evaporate the solvent by placing the flask in a sand bath maintained at 60–65 °C in the **hood.** Use a gentle stream of nitrogen or dry air to hasten the process.

HOOD

Weigh the resulting product and calculate the percent yield. Characterize the product as outlined above in the microscale procedure.

Eightyfold Scaleup. The reagent and solvent amounts are given in the following table.

Physical Properties of Reactants

Compound	MW	Amount	mol	mp (°C)	bp (°C)	*d*	n_D
4-Ethylphenol	122.17	12.5 g	0.102	42–45			
NaOH	40.0	4.0 g	0.10				
Tetrabutylammonium bromide	322.38	0.40 g	0.0012	103–104			
Methyl iodide	141.94	14.5 g	0.102		41–43	2.28	1.5304
Water		25 mL					

Reagents and Equipment. Use a 100-mL round-bottom flask containing a magnetic stirring bar and equipped with a reflux condenser (➡).

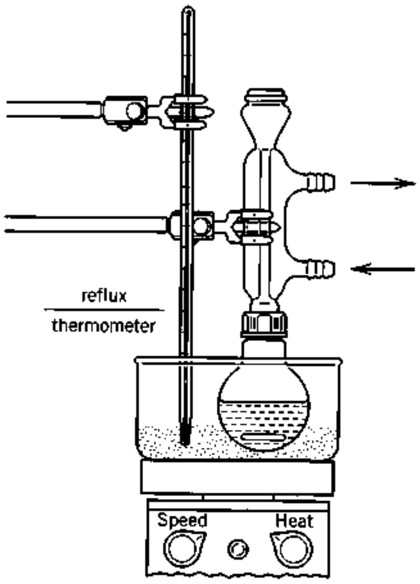

Reflux conditions
100-mL RB flask

Reaction Conditions. Heat the reaction mixture at reflux, with stirring, for 90 min.

Isolation of Product. Allow the two-phase mixture to cool to room temperature. Transfer this mixture to a 125-mL separatory funnel and remove the aqueous layer. Store this in a 125-mL Erlenmeyer flask until the final product has been purified and characterized. Wash the organic layer successively with 5% sodium hydroxide solution (20 mL) and distilled water (20 mL). After each washing remove the aqueous layer and add it to the 125-mL Erlenmeyer flask. Finally, collect the reddish-brown organic layer in another 125-mL Erlenmeyer flask and dry it over anhydrous sodium sulfate.

Remove the drying agent by filtration through a glass wool plug and collect the product ether in a tared container. Weigh, calculate the percent yield, and then characterize the material as described in the microscale procedure.

Experiment 22C*

Butyl *p*-Nitrophenyl Ether: Preparation Using a Monomode Microwave Apparatus

M

REACTION

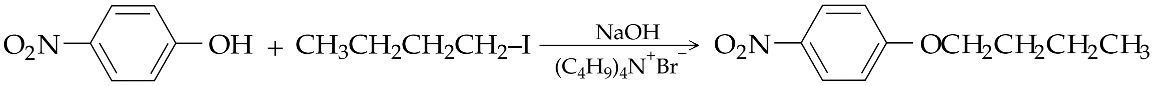

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2 h.

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
p-nitrophenol	139	215 mg	1.55	110–115			
1-iodobutane	184	0.17 mL	1.49		130–131	1.617	1.498
Tetrabutylammonium bromide	322	20 mg	0.06	103–104			
15% NaOH solution		2 mL					

Reagents and Equipment. This experiment is designed for use in the CEM Discover and Biotage Initator microwave units.

In a 10.0-mL glass microwave reaction vessel containing a magnetic stir bar, place 215 mg (1.55 mmol) of *p*-nitrophenol, 20 mg (0.06 mmol) of tetrabutylammonium bromide, 2 mL of 15% sodium hydroxide solution, and 0.17 mL (1.49 mmol) of 1-iodobutane. Immediately cap the vessel with the microwave pressure cap.

HOOD

CAUTION: 1-iodobutane is toxic and must be dispensed in the **hood.** Dispense both the 1-iodobutane and alkaline solution using an automatic delivery pipet. Since the reaction requires heating the reaction mixture to above the boiling point of some components in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential*.

Reaction Conditions. Place the reaction vessel in the microwave cavity and, depending on the equipment used, position the pressure device on top. Program the microwave unit to heat the reaction mixture to 150 °C using no more that 50 W of microwave power, and hold at this temperature for 5 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the tube from the microwave unit.

Isolation of Product. Transfer the reaction mixture with a Pasteur pipet into a 30-mL separatory funnel. Clamp the funnel to a ring stand. Rinse the microwave reaction vessel with 2.0 mL of diethyl ether and add the washings to the separatory funnel. Carefully cap and invert the funnel. Immediately vent the funnel by opening the stopcock. Close the stopcock, place the funnel back on the ring stand and remove the stopper. Drain the lower (aqueous) layer into a 50-mL Erlenmeyer flask. Extract the crude organic layer with an additional two 5-mL portions of 5% sodium hydroxide solution, followed by 5 mL of water. During each extraction, cap and invert the funnel several times and each time release the pressure by opening the stopcock and then allow the funnel to stand so the layers will separate. Remove the lower

---

*This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

(aqueous) layer after each extraction into the 50-mL Erlenmeyer flask. Save the aqueous waste until the experiment is complete and then discard as directed. Dry the organic layer by pipetting it into a clean 25-mL Erlenmeyer flask containing 200 mg of anhydrous sodium sulfate. Transfer the anhydrous solution, using a Pasteur filter pipet, to a clean tared 10-mL pear-shaped flask. Remove the ether on a rotary evaporator or by evaporation in the **hood** using a gentle stream of nitrogen gas with warming in a sand bath to isolate the crude product. Reweigh the flask and calculate the crude yield.

HOOD

Purification and Characterization. The crude product can be further purified by recrystallization from 95% ethanol using a Craig tube.

Weigh the pure butyl *p*-nitrophenyl ether and calculate the percent yield. Determine the melting point and compare the experimental values with those in the literature.

Obtain an IR spectrum of the compound and compare it to that shown in Figure 6.42 for 4-butoxy nitrobenzene (butyl *p*-nitrophenyl ether).

Nuclear Magnetic Resonance Analysis. If facilities permit, you can obtain both ^{1}H and ^{13}C NMR spectra of your butyl *p*-nitrophenyl ether in $CDCl_3$, and compare your spectra with those in Figures 6.43 and 6.44.

^{13}C NMR ($CDCl_3$): δ 164.3, 141.4, 125.8, 114.5, 68.6, 31.1, 19.2, 13.8

^{1}H NMR ($CDCl_3$): δ 8.2 (d, J = 9.1, 2H), 7.93 (d, J = 9.2, 2H), 4.05 (t, J = 6.5, 2H), 1.8 (m, J = 8.0, J = 6.5, 2H), 1.6 (m, J = 8.0, J = 7.4, 2H), 1.0 (t, J = 7.4, 3H)

IR (neat, HATR) 3114, 2955, 2873, 1594, 1509, 1261, 1175, 1109, 846 cm^{-1}

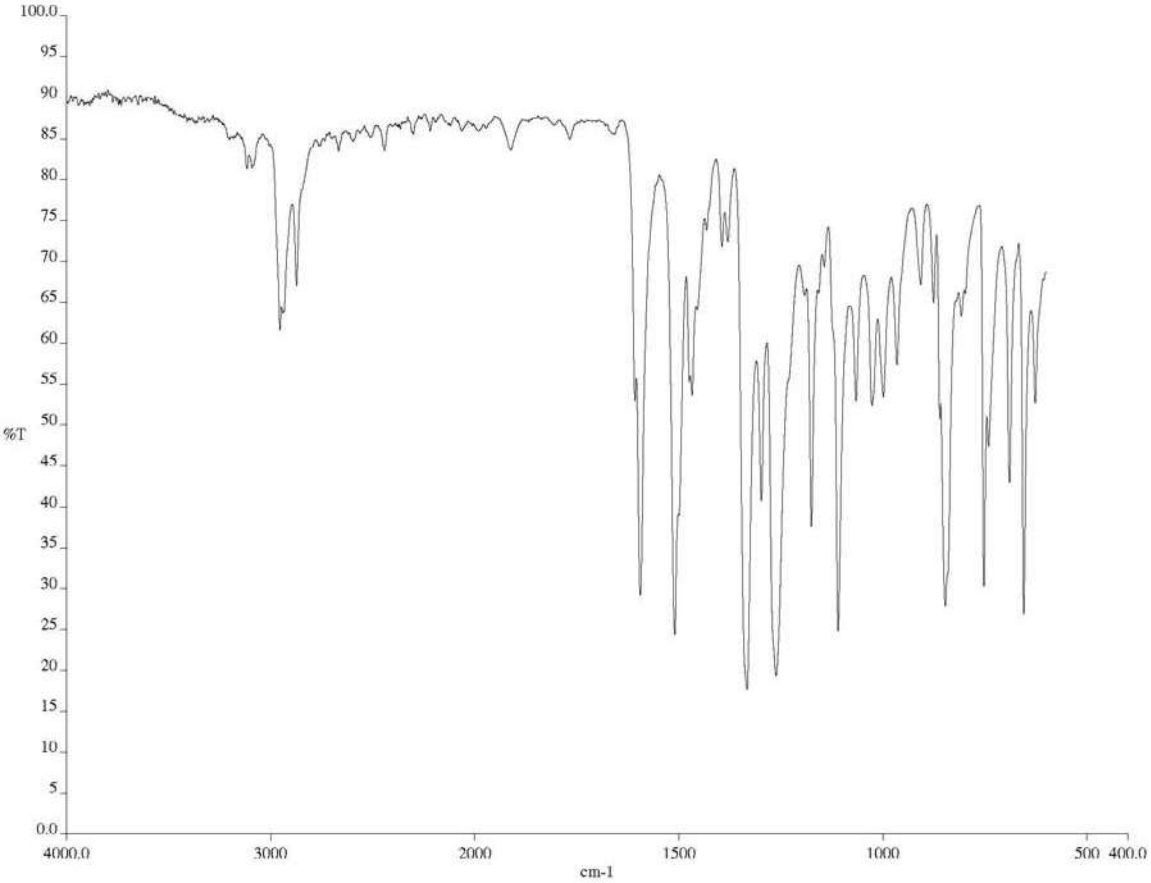

c:\pel_data\spectra\cm-3 p-butoxynitronenzene.sp

Figure 6.42 HATR-IR spectrum: butyl *p*-nitrophenyl ether.

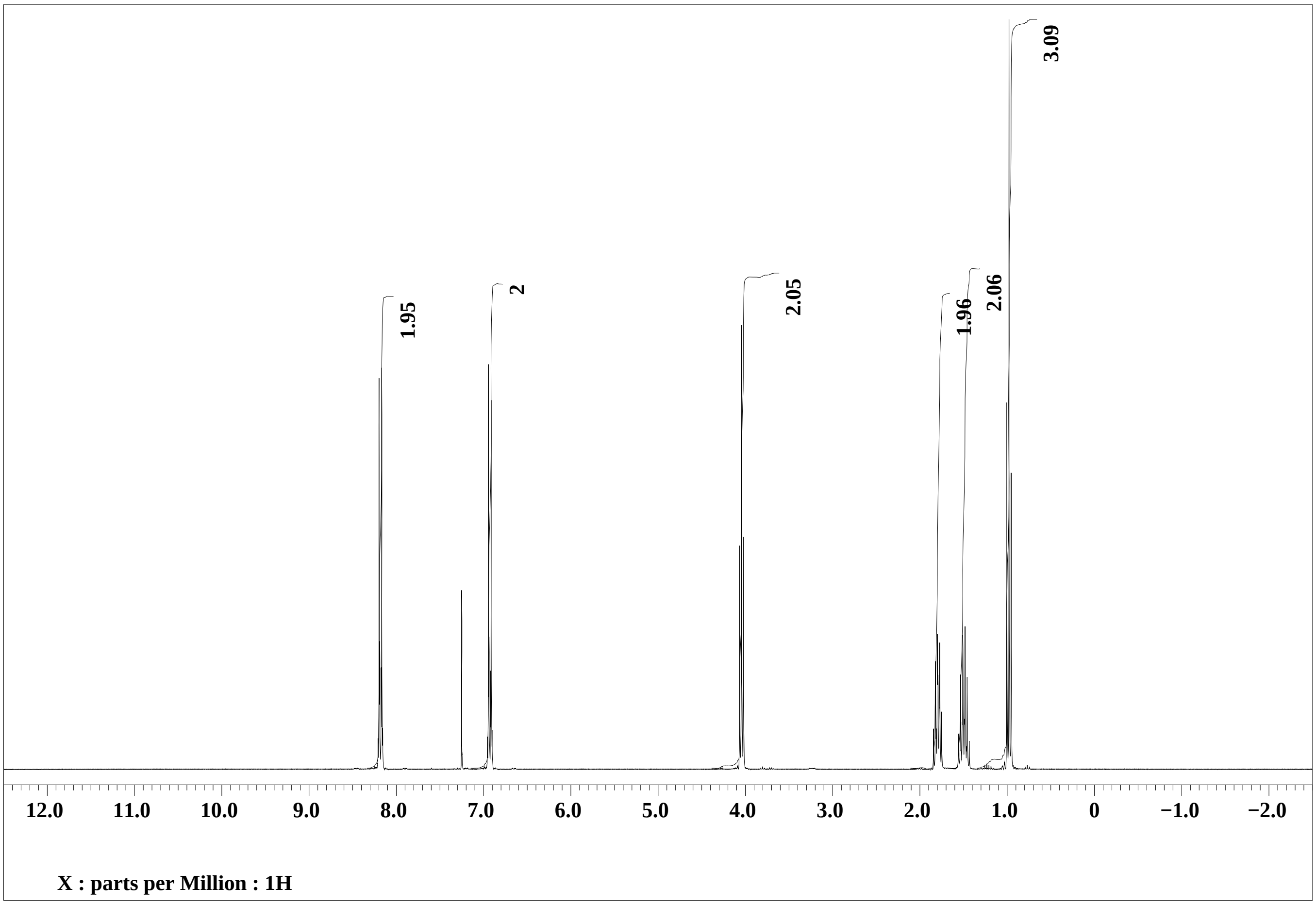

Figure 6.43 ^{1}H-NMR spectrum: butyl *p*-nitrophenyl ether in $CDCl_3$.

Experiment 22D*

Butyl *p*-Nitrophenyl Ether: Preparation Using a Multimode Microwave Apparatus

M

REACTION

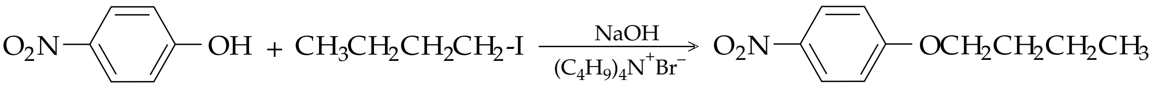

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2 h.

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
p-nitrophenol	139	215 mg	1.55	110–115			
1-iodobutane	184	0.17 mL	1.49		130–131	1.617	1.498
Tetrabutylammonium bromide	322	20 mg	0.06	103–104			
15% NaOH solution		5 mL					

[*] This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

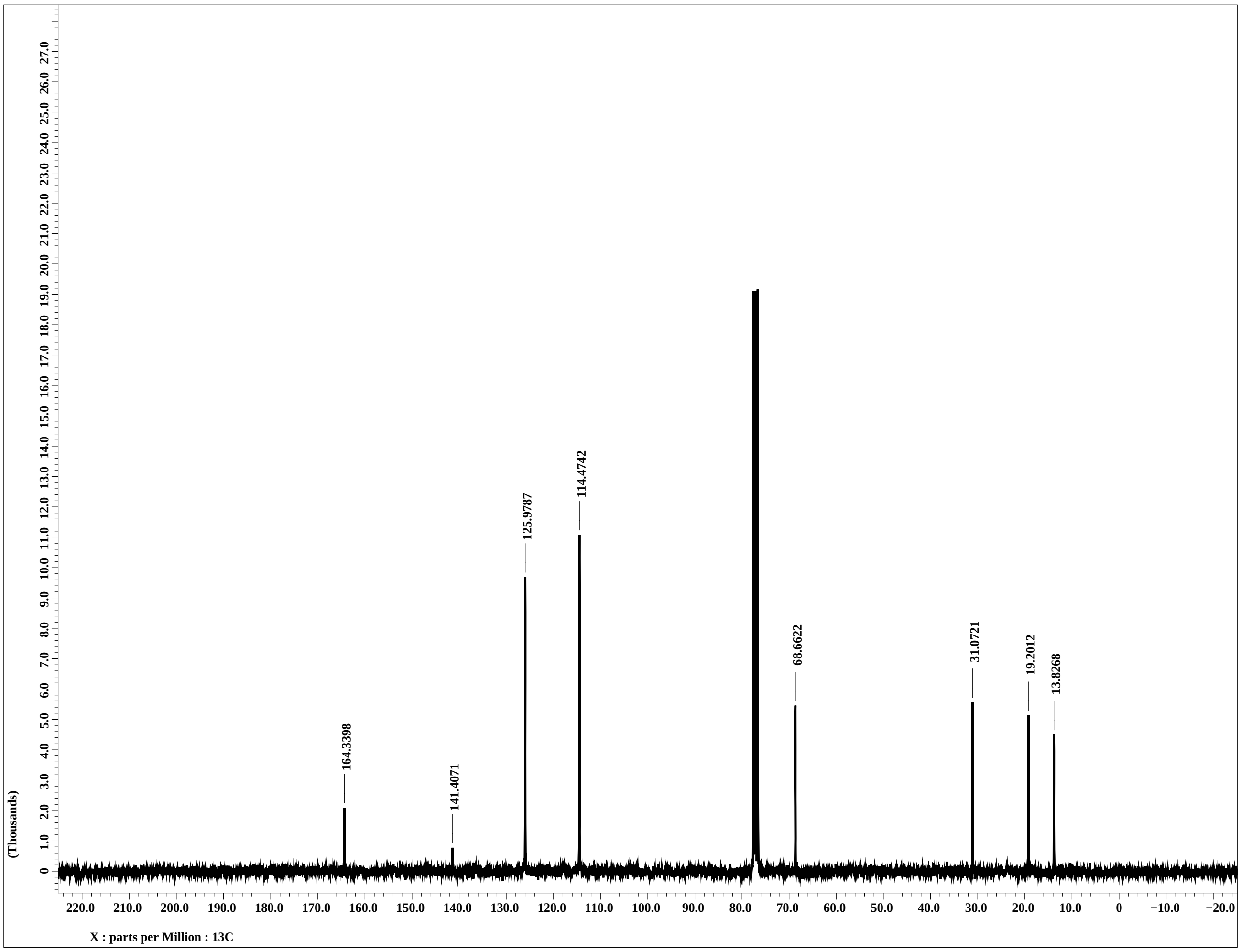

Figure 6.44 ^{13}**C-NMR spectrum: butyl *p*-nitrophenyl ether in CDCl$_3$.**

Reagents and Equipment. This experiment is designed for use in the CEM MARS, Milestone START, and Anton Paar Synthos 3000 microwave units. When using the Anton Paar Synthos 3000 unit with the 24-position silicon carbide plate rotor containing glass vials, the reagent and solvent quantities cited in the monomode procedure should be used in conjunction with the reaction conditions here in the multimode procedure.

In a microwave reaction vessel containing a magnetic stir bar, place 215 mg (1.55 mmol) of *p*-nitrophenol, 20 mg (0.06 mmol) of tetrabutylammonium bromide, 5 mL of a 15% sodium hydroxide solution, and 0.17 mL (1.49 mmol) of 1-iodobutane. Immediately cap the vessel with the microwave pressure cap and adjust the tightness to the manufacturer-specified level. Place the sealed vessel into its outer protective jacket.

CAUTION: 1-iodobutane is toxic and must be dispensed in the **hood.** Dispense both the 1-iodobutane and alkaline solution using an automatic delivery pipet. Since the reaction requires heating the reaction mixture to above the boiling point of some components in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential.*

HOOD

Reaction Conditions. Insert the loaded vessels into the reaction carousel ensuring they are evenly spaced and then place the carousel into the microwave cavity. If provided by the manufacturer, connect a temperature probe to the control vessel. Program the microwave unit to heat the reaction vessels to 150 °C and hold at this temperature for 5 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the carousel from the microwave unit.

Isolation of Product. Transfer the reaction mixture with a Pasteur pipet into a 30-mL separatory funnel. Clamp the funnel to a ring stand. Rinse the microwave reaction vessel with 2.0 mL of diethyl ether and add the washings to the separatory funnel. Carefully cap and invert the funnel. Immediately vent the funnel by opening the stopcock. Close the stopcock, place the funnel back on the ring stand and remove the stopper. Drain the lower (aqueous) layer into a 50-mL Erlenmeyer flask. Extract the crude organic layer with an additional two 5-mL portions of 5% sodium hydroxide solution, followed by 5 mL of water. During each extraction, cap and invert the funnel several times and each time release the pressure by opening the stopcock and then allow the funnel to stand so the layers will separate. Remove the lower (aqueous) layer after each extraction into the 50-mL Erlenmeyer flask. Save the aqueous waste until the experiment is complete and then discard as directed. Dry the organic layer by pipetting it into a clean 25-mL Erlenmeyer flask containing 200 mg of anhydrous sodium sulfate. Transfer the anhydrous solution, using a Pasteur filter pipet, to a clean tared 10-mL pear-shaped flask. Remove the ether on a rotary evaporator or by evaporation in the **hood** using a gentle stream of nitrogen gas with warming in a sand bath to isolate the crude product. Reweigh the flask and calculate the crude yield.

HOOD

Purification and Characterization. The crude product can be further purified by recrystallization from 95% ethanol using a Craig tube.

Weigh the pure butyl *p*-nitrophenyl ether and calculate the percent yield. Determine the melting point and compare the experimental values with those in the literature.

Obtain an IR spectrum of the compound and compare it to that shown in Figure 6.42 for 4-butoxy nitrobenzene (butyl *p*-nitrophenyl ether).

Nuclear Magnetic Resonance Analysis. If facilities permit, you can obtain both ^{1}H and ^{13}C NMR spectra of your butyl *p*-nitrophenyl ether in $CDCl_3$, and compare your spectra with those in Figures 6.43 and 6.44.

^{13}C NMR ($CDCl_3$): δ 164.3, 141.4, 125.8, 114.5, 68.6, 31.1, 19.2, 13.8

^{1}H NMR ($CDCl_3$): δ 8.2 (d, $J = 9.1$, 2H), 7.93 (d, $J = 9.2$, 2H), 4.05 (t, $J = 6.5$, 2H), 1.8 (m, $J = 8.0$, $J = 6.5$, 2H), 1.6 (m, $J = 8.0$, $J = 7.4$, 2H), 1.0 (t, $J = 7.4$, 3H)

IR (neat, HATR) 3114, 2955, 2873, 1594, 1509, 1261, 1175, 1109, 846 cm^{-1}

QUESTIONS

6-144. Sulfides are often prepared using an S_N2 reaction. For example,

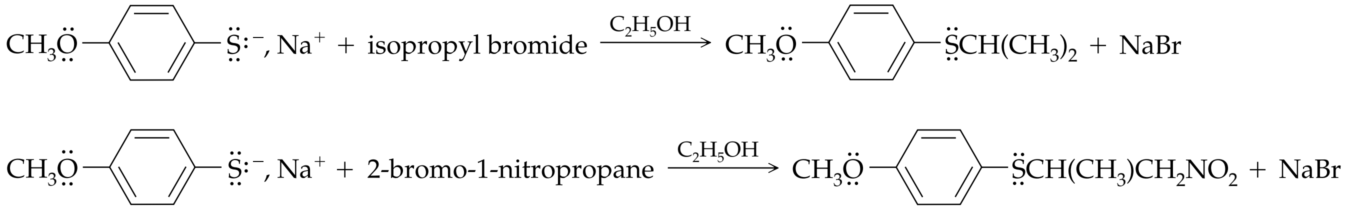

The reaction with isopropyl bromide is 16 times faster than the reaction with 2-bromo-1-nitropropane. Explain.

6-145. If 3-bromo-1-propanol is treated with NaOH, a compound of molecular formula C_3H_6O is formed. Suggest a structure for this product.

6-146. Arrange the substituted phenols given below in order of increasing reactivity toward ethyl iodide in the Williamson reaction. After arranging and explaining your order, does this order match that of rates of deprotonation?

(a) 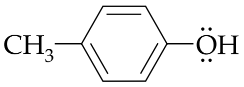

(b) 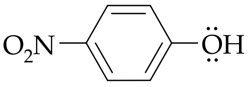

(c) 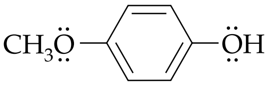

6-147. *trans*-2-Chlorocyclohexanol reacts readily with NaOH to form cyclohexene oxide, but the cis isomer will not undergo this reaction. Explain.

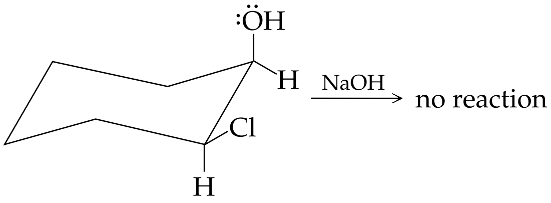

cis-2-Chlorocyclohexanol

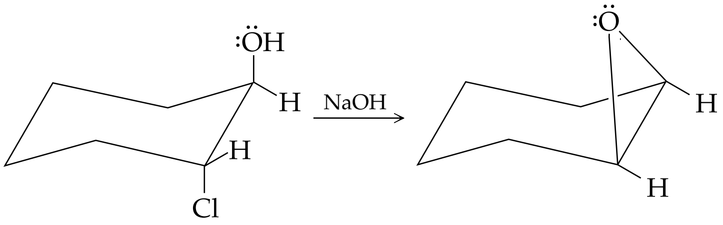

trans-2-Chlorocyclohexanol

Cyclohexene oxide

6-148. *tert*-Butyl ethyl ether might be prepared two ways using different starting materials.

$$\begin{tikzcd}
(CH_3)_3C\ddot{O}:^-, K^+ + CH_3CH_2Cl \arrow[dr] & \\
& (CH_3)_3C—\ddot{O}—CH_2CH_3 \\
CH_3CH_2\ddot{O}:^-, K^+ + (CH_3)_3CCl \arrow[ur] &
\end{tikzcd}$$

Which route would you choose to prepare the above ether, and why?

6-149. What product(s) would you expect to form when tetrahydrofuran is treated with excess hydroiodic acid (HI)?

6-150. Write a suitable mechanism for the cleavage of butyl isopropyl ether with HI at 100 °C to form exclusively isopropyl alcohol and 1-iodobutane. Explain why butyl alcohol and isopropyl iodide are not formed in the reaction.

6.151. There are only four lines in the aromatic region of the fully ^1H decoupled ^{13}C NMR spectrum of propyl *p*-tolyl ether (110–160 ppm), yet there are six aromatic carbon atoms. Explain.

BIBLIOGRAPHY

Selected review articles on the Williamson synthesis.

Dermer, O. C. *Chem. Rev.* **1934,** *14,* 409.

Feuer, H.; Hooz, J. in *The Chemistry of the Ether Linkage;* Patai, S., Ed.; Wiley: New York, 1967, p. 446.

Smith, M. B.; March J. *Advanced Organic Chemistry,* 6th ed.; Wiley-Interscience: New York, 2007, Chap. 10, p. 529.

Examples of the Williamson reaction in *Organic Syntheses* include

Allen, C. F. H.; Gates, J. W., Jr. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 140. *ibid.,* p. 418.

Boehme, W. R. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 590.

Fuson, R. C.; Wojcik, B. H. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 260.
Gassman, P. G.; Marshall, J. L. *Organic Syntheses;* Wiley New York, 1973; Collect. Vol. V, p. 424.
Gokel, G. W.; Cram, D. J.; Liotta, C. L.; Harris, H. P.; Cook, F. L. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 301.
Kuryla, W. C.; Hyve, J. E. *Organic Syntheses;* Wiley: New York, 1973; Collect, Vol. V, p. 684.
Mirrington, R. N.; Feutrill, G. I. *Organic Syntheses;* Willey: New York, 1988; Collect. Vol. VI, p. 859.
Pedersen, C. J. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 395.
Vyas, G. N.; Shah, N. M. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 836.

Review articles on phase transfer catalysis:

Dehmlow, E. V.; Dehmlow, S. S. *Phase Transfer Catalysis,* 3rd ed.; VCH: New York, 1993.
Gokel, G. W.; Weber, W. P. *J. Chem. Educ.* **1978,** *55,* 350; *ibid.,* 429.
Smith, M. B.; March. J. *Advanced Organic Chemistry*, 6th ed.; Wiley-Interscience: New York, 2007, Chap. 10, p. 508.

The procedures used in these experiments for the preparation of the ethers were adapted from the work of

McKillop, A.; Fiaud, J. C.; Hug, R. P. *Tetrahedron* **1974,** *30,* 1379.
Rowe, J. E. *J. Chem. Educ.* **1980,** *57,* 162.

EXPERIMENT 23

Amide Synthesis: Acetanilide and N,N′*-Diacetyl-1,4-phenylenediamine*

Common name: acetanilide
CA number: [103-84-4]
CA name as indexed: acetamide, *N*-phenyl-
Common name: *N,N′*-diacetyl-1,4-phenylenediamine
CA number: [140-50-1]
CA name as indexed: acetamide, *N,N′*–1,4-phenylenebis-

Purpose. You will carry out one of the major synthetic routes used in the preparation of amides; the method involves the reaction of ammonia, or a primary or secondary amine, with an active acylating reagent. You will also explore the use of acetic anhydride as an acylating agent. The acetanilide product (Experiment [23A]) may be used in Experiment [28].

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallizations (pp. 89–91)

REACTION

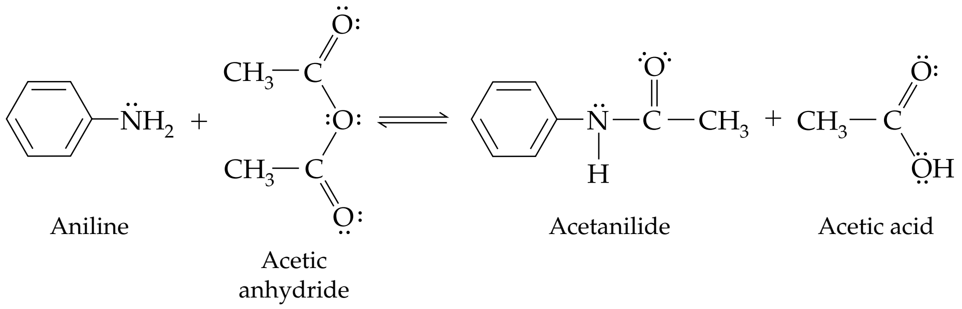

DISCUSSION

A number of important chemical and biochemical synthetic sequences are initiated by the addition of a nitrogen nucleophile to a carbonyl carbon atom to yield **carboxylic amides.** Amides are classified as **primary (1°), secondary (2°)** or **tertiary (3°)** based on the number of carbon atoms attached to the nitrogen.

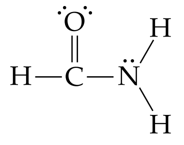

1° Amide
Methanamide
(formamide)

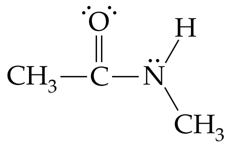

2° Amide
N-Methylethanamide
(*N*-methylacetamide)

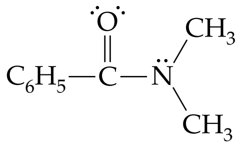

3° Amide
N,*N*-Dimethylbenzamide

Cyclic amides are called **lactams** and are classified by ring size. Imides contain a nitrogen bonded to two carbonyl groups and are nitrogen analogues of anhydrides.

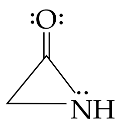

An α-lactam

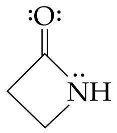

A β-lactam

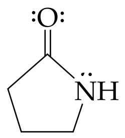

A γ-lactam

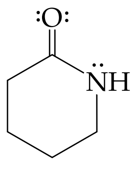

A δ-lactam

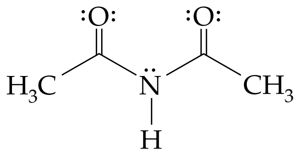

An imide

Amides appear in such diverse compounds as penicillin V (a β-lactam and an amide) and polypeptides (α-amino acids linked by amide bonds); and an imide, 1,2-benzenecarboxylic imide, is used in the Gabriel synthesis of amines. An imide is prepared in Experiment [24] and anhydrides are prepared in Experiments [25A] and [25B]. The polyamide polymer, nylon, is prepared in Chapter 7, Sequence B.

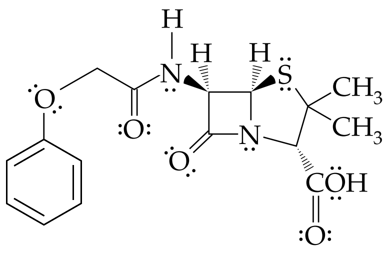

Penicillin V

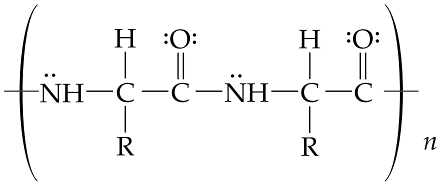

A polypeptide

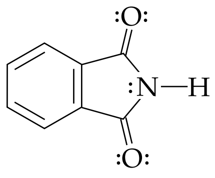

1,2-Benzenecarboxylic imide

The experiments outlined here illustrate the preparation of 2° amides. The process involves the attack of a primary amine on the acetyl group of acetic anhydride. Ammonia or secondary amines also react readily with this reagent to yield 1° and 3° amides, respectively. The mechanism shown here is an example of the attack of a nucleophilic reagent on a carbonyl carbon of the anhydride:

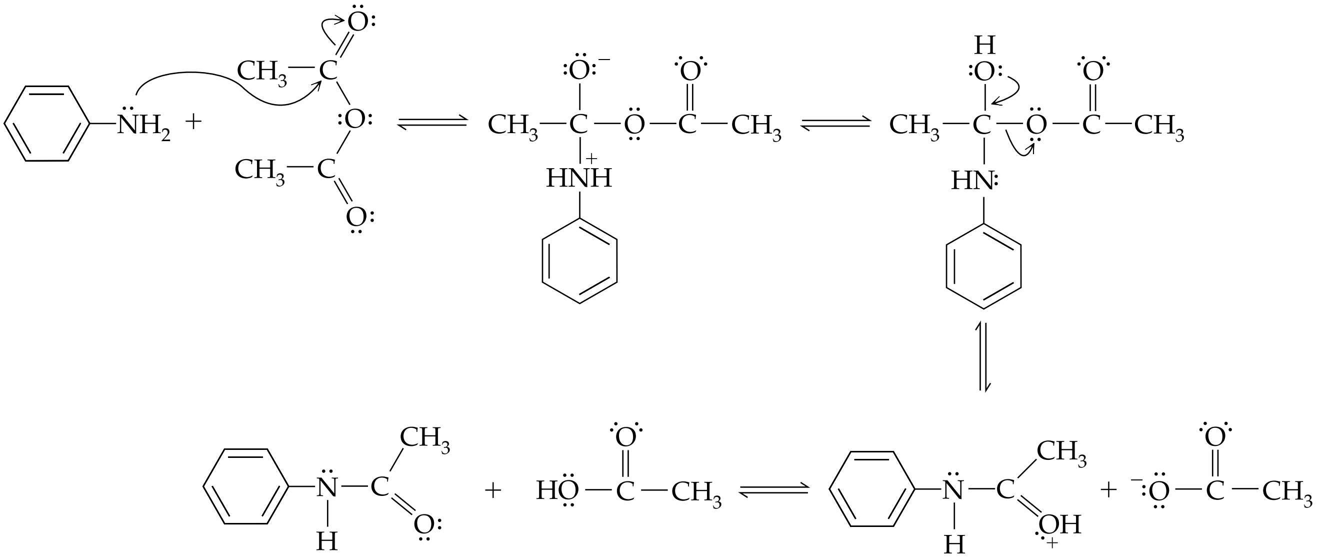

In the preparations given below, the amine reagents (aniline and *p*-phenylenediamine) are purified as their hydrochloride salts. Arylamines are relatively weak bases (K_b values in the order of 10^{-10}) but when treated with a strong acid, such as HCl, they are completely protonated, yielding the corresponding water-soluble hydrochloride salt:

$$\underset{\text{Aniline}}{C_6H_5{-}NH_2} + HCl \rightarrow \underset{\substack{\text{Anilinium}\\ \text{hydrochloride salt}\\ \text{(water soluble)}}}{C_6H_5{-}NH_3^+, Cl^-}$$

As directed in the experiment, decolorizing charcoal is added to the aqueous solution of the arylamine salt. The charcoal absorbs impurities and subsequent filtration of the mixture, which removes the charcoal, yields an aqueous solution of the purified arylamine salt.

The second stage of the reaction sequence requires that a solution of sodium acetate be added to the reaction mixture after initial addition of acetic anhydride to the purified anilinium hydrochloride salt solution:

$$C_6H_5{-}NH_3^+, Cl^- + CH_3COO^-, Na^+ \rightleftharpoons C_6H_5{-}NH_2 + CH_3COOH + Na^+, Cl^-$$

Addition of the sodium acetate solution serves to liberate the arylamine so that the desired nucleophilic substitution reaction may occur; ammonium cations are not nucleophilic, since they are positively charged and do not even possess a lone pair of electrons.

Sodium acetate is the conjugate base of acetic acid, which is a weak acid. Furthermore, the anilinium ion ($pK_a = 4.6$) is a slightly stronger acid than acetic acid ($pK_a = 4.8$). Thus, the equilibrium reaction is shifted to the right, producing the arylamine.

In Experiment [23B], the *p*-phenylenediamine forms the corresponding dihydrochloride salt, $Cl^- \cdot H_3\overset{+}{N}{-}C_6H_4{-}\overset{+}{N}H_3 \cdot Cl^-$. As in Experiment [23A], the aqueous solution of the salt is purified with charcoal, and upon addition of the sodium acetate solution, the free amine is regenerated.

This overall process illustrates an important transformation for most amines. These amines can be converted to water-soluble ionic salts by reaction with acids and can be recovered from these acid salts by treatment with base. This technique was used in Experiment [4C] to extract ethyl 4-aminobenzoate, as its water-soluble salt, from a mixture.

Acetanilide

Experiment 23A

The reaction is shown above.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Aniline	93.13	100 µL	1.09		184	1.02	1.5863
Concd HCl		3 drops					
Sodium acetate trihydrate	136.08	150 mg	1.10	58			
Acetic anhydride	102.09	150 µL	1.59		140	1.08	1.3901

Reagents and Equipment. In the **hood,** place 100 µL of aniline in a tared 10 × 75-mm test tube (standing in a small beaker or Erlenmeyer flask). Fit the tube with a cork stopper.

HOOD

CAUTION: Aniline is a toxic material and a cancer suspect agent.

NOTE. Dispense the aniline using an automatic delivery pipet. Again weigh the test tube and container to determine the exact amount of aniline delivered.

Now using a 1.0-mL graduated pipet add, with swirling, 0.5 mL of water; then, in the **hood,** add 3 drops of concentrated hydrochloric acid using a Pasteur pipet. Add 10 mg of powdered decolorizing charcoal, or the pelletized form (Norit), to the resulting solution.

HOOD

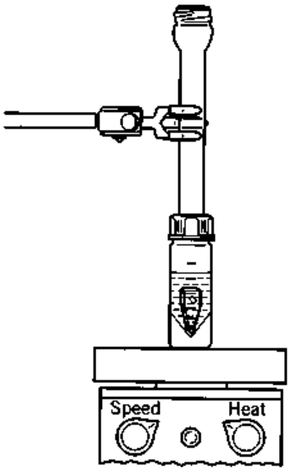

$C_6H_5NH_2$, 100 µL
+ H_2O, 1.0 mL
+ concd HCl, 3 drops
+ NaOAc·$3H_2O$, 150 mg
+ $(CH_3CO)_2O$, 150 µL

Using a Pasteur pipet, transfer the well-mixed suspension to a 25-mm funnel fitted with fast-grade filter paper to remove the charcoal by gravity filtration. *Wet the filter paper in advance with distilled water and blot the excess water from the stem of the funnel.*

Collect the filtrate in a 3.0-mL conical vial. Use an additional 0.5 mL of water to rinse the test tube and the collected charcoal. Combine the rinse with the original filtrate. Place a magnetic spin vane in the vial and attach it to an air condenser (➡).

NOTE. If the pelletized form of charcoal is used, transfer through a Pasteur filter pipet directly to the 3.0-mL conical vial should be sufficient.

NOTE. Tap all of the filtrate from the funnel stem into the collecting vial. As a result of this purification step, a clear, colorless solution of aniline hydrochloride should be obtained.

Dissolve 150 mg (1.10 mmol) of sodium acetate trihydrate in 0.5 mL of distilled water in a 10 × 75-mm test tube. Cap the tube and set the solution aside for use in the next step.

HOOD

Remove the air condenser, and then use an automatic delivery pipet in the **hood** to add, with stirring, 150 μL of acetic anhydride to the solution of aniline hydrochloride, followed quickly by addition (Pasteur pipet) of the previously prepared solution of sodium acetate. Reattach the air condenser.

Reaction Conditions. The reaction is very rapid and the product begins to precipitate immediately upon mixing of the reagents. Stir to thoroughly mix the reagents. Allow the reaction mixture to stand at room temperature for approximately 5 min and then place it in an ice bath for an additional 5–10 min to complete the crystallization process.

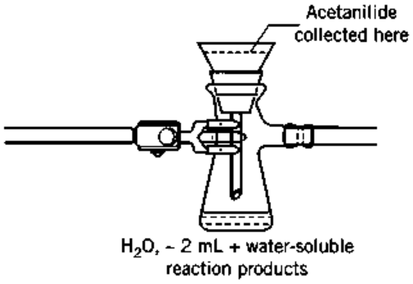
Acetanilide collected here
H_2O, ~ 2 mL + water-soluble reaction products

Isolation of Product. Collect the acetanilide product by filtration under reduced pressure using a Hirsch funnel (⬅). Rinse the conical vial with two 0.5-mL portions of water (using calibrated Pasteur pipet) and use the rinse to wash the collected filter cake. Place a piece of plastic food wrap over the mouth of the funnel and continue the suction for 5–8 min (see Prior Reading). The snow-white crystals are further dried on a porous clay plate or on filter paper in a desiccator.

Purification and Characterization. Further purification of the product is generally not required. However, the acetanilide may be recrystallized from hot water or from ethanol–water using the Craig tube.

NOTE. Acetanilide (150 mg) can be recrystallized from approximately 3 mL of water, or from 2 mL of ethanol–water (1:10 v/v), with better than 80% recovery.

Weigh the dried crystals and calculate the percent yield. Determine the melting point of the material and compare your result to that reported in the literature.

Obtain an IR and/or NMR spectrum of the product and compare it with that of an authentic sample or to that recorded in the literature (*The Aldrich Library of IR Spectra, The Aldrich Library of NMR Spectra,* and/or SciFinder Scholar).

Chemical Tests. Characterization of the product may be enhanced by performing several chemical tests given in Chapter 9.

Check the solubility of acetanilide in water. Is the aqueous solution acidic, basic, or does it remain neutral as indicated by pH paper? Does the ignition test indicate that an aromatic group is present? Does the soda lime or sodium fusion test indicate the presence of nitrogen? Does the hydroxamate test for amides give a positive result?

OPTIONAL SEMIMICROSCALE PREPARATION

This experiment can be scaled up to be run at five times the amounts used in the above microscale preparation.

The procedure is identical to the above with the following exceptions.

1. Use two 15 × 100-mm test tubes.

2. Carry out the reaction in a 10-mL round-bottom flask containing a magnetic spin bar and fitted with an air condenser (➡).

3. Increase the amounts of all reagents and the solvent by a factor of 5.

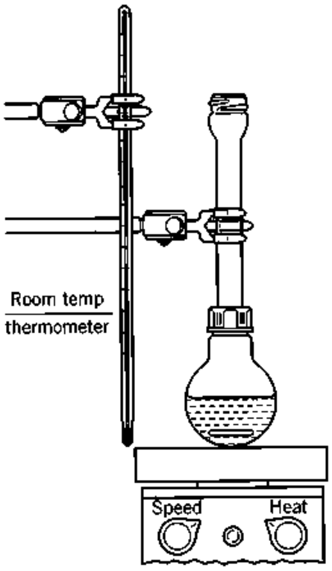

A 10-mL RB flask

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Aniline	93.13	500 μL	5.45		184	1.02	1.5863
Concd HCl		0.75 mL					
Sodium acetate trihydrate	136.08	750 mg	5.50	58			
Acetic anhydride	102.09	750 μL	7.93		140	1.08	1.3901

N,N′-Diacetyl-1,4-phenylenediamine

Experiment 23B

REACTION

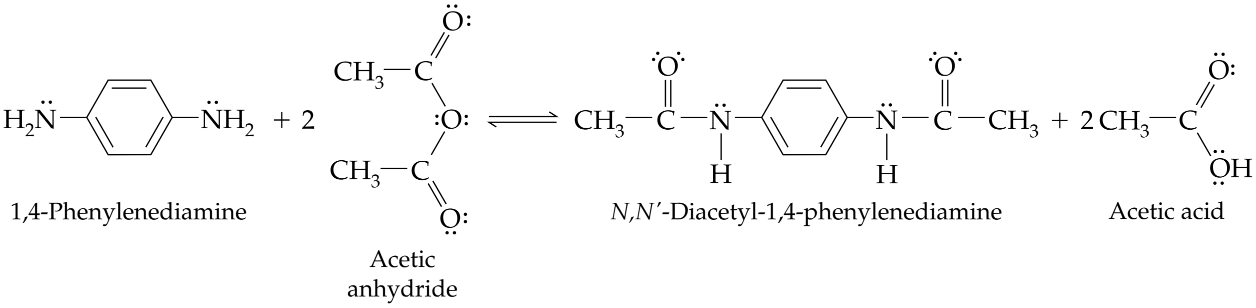

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
1,4-Phenylenediamine	108.14	117 mg	1.08	138			
Concd HCl		6 drops					
Sodium acetate trihydrate	136.08	300 mg	2.20	58			
Acetic anhydride	102.09	350 μL	3.71		140	1.08	1.3901

Reagents and Equipment. Weigh and place 117 mg (1.08 mmol) of 1,4-phenylenediamine in a 10 × 75-mm test tube (standing in a small beaker or Erlenmeyer flask).

CAUTION: This reagent is toxic and a cancer suspect agent.

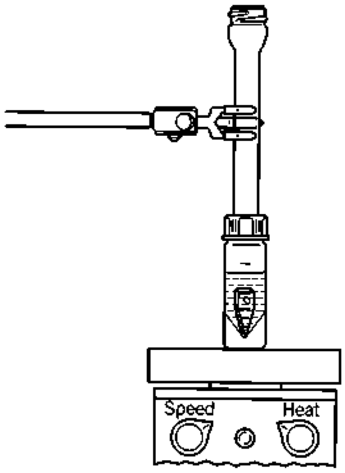

p-$NH_2C_6H_4NH_2$, 117 mg
+ H_2O, 2.5 mL
+ concd HCl, 6 drops
+ NaOAc, 300 mg
+ $(CH_3CO)_2O$, 350 μL

With gentle swirling, add 1.0 mL of distilled water and, using a Pasteur pipet, 6 drops of concentrated hydrochloric acid. After dissolution, add 30 mg of either powdered decolorizing charcoal (Norit) or the pelletized form. Transfer the well-mixed suspension, by use of a Pasteur pipet, to a 25-mm funnel fitted with fast-grade filter paper previously wet with water. The charcoal is removed by gravity filtration. Collect the filtrate, which is clear to slightly yellow, in a 5-mL conical vial containing a magnetic spin vane. Use three 0.5-mL portions of water (calibrated Pasteur pipet) to rinse the test tube, and in turn use the rinse to wash the collected charcoal. The rinse is combined with the original filtrate. *Blot the excess water from the stem of the funnel*. Attach the vial to an air condenser (⬅).

NOTE. If the pelletized form of charcoal is used, transfer through a Pasteur filter pipet directly to the 5.0-mL conical vial should be sufficient.

Dissolve 300 mg (2.20 mmol) of sodium acetate trihydrate in 0.5 mL of distilled water in a 10 × 75-mm test tube. Stir the mixture with a spatula to aid the dissolution process. Cap the tube and set it aside for use in the next step.

Reaction Conditions. Remove the air condenser, and use an automatic delivery pipet to add 350 μL of acetic anhydride to the solution of 1,4-phenylenediamine dihydrochloride, and stir the mixture briefly using a magnetic stirrer.

HOOD

NOTE. Dispense the acetic anhydride in the **hood** *using an automatic delivery pipet. A slight amount of white precipitate may be observed at this stage.*

Now add the previously prepared sodium acetate solution by Pasteur pipet to the reaction mixture with stirring. Reattach the air condenser.

Isolation of Product. The reaction is very rapid and the desired product begins to precipitate almost immediately. After stirring briefly, allow the mixture to stand at room temperature for a few minutes and then place it in an ice bath for an additional 5–10 min.

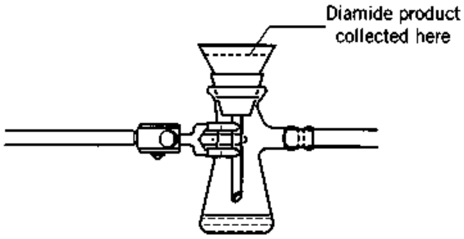

H_2O, ~3.5 mL + water-soluble reaction products

Purification and Characterization. Collect the crude *N,N′*-diacetyl-1,4-phenylenediamine by vacuum filtration using a Hirsch funnel (⬅). Rinse the vial with two 0.5-mL portions of water and use the rinse to wash the filter cake. Place a piece of plastic food wrap over the mouth of the funnel and continue the suction for an additional 5–8 min. Place the collected material on a porous clay plate or filter paper to dry further.

NOTE. If this material is to be used in Experiment [29B], recrystallization from methanol is suggested.

Weigh the dried crystals and calculate the percent yield. Determine the melting point and compare it with the value in the literature. Obtain an IR and/or the NMR spectrum of the product. The infrared spectrum is shown in Figure 6.45.

Chemical characterization tests might also be run as outlined in Experiment [23A].

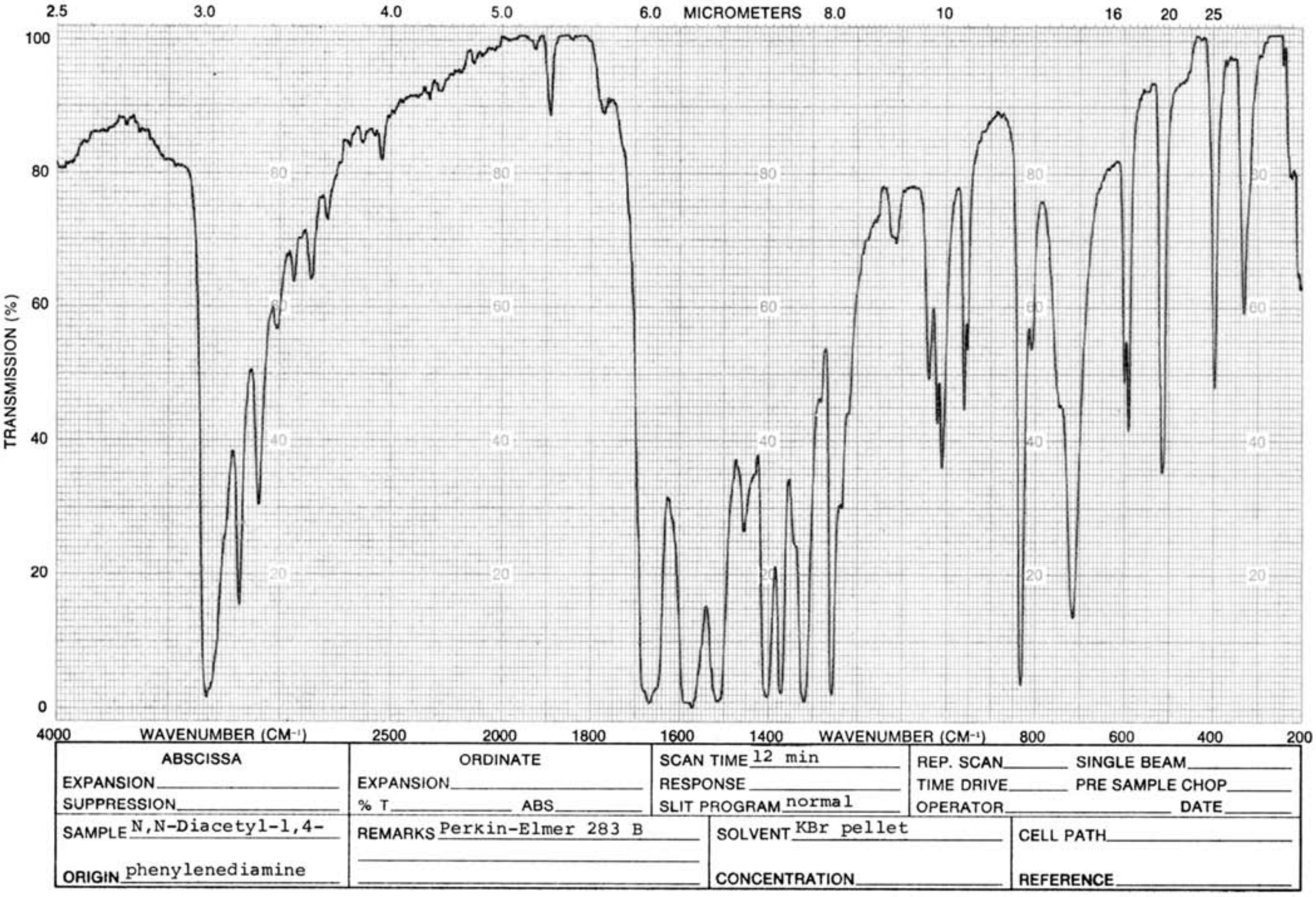

ABSCISSA
EXPANSION
SUPPRESSION

ORDINATE
EXPANSION
% T ABS

SCAN TIME 12 min
RESPONSE
SLIT PROGRAM normal

REP. SCAN SINGLE BEAM
TIME DRIVE PRE SAMPLE CHOP
OPERATOR DATE

SAMPLE N,N-Diacetyl-1,4-
ORIGIN phenylenediamine

REMARKS Perkin-Elmer 283 B

SOLVENT KBr pellet
CONCENTRATION

CELL PATH
REFERENCE

Figure 6.45 IR spectrum: *N,N′*-diacetyl-1,4-diphenylenediamine.

QUESTIONS

6-152. What is the function of the sodium acetate in the reactions outlined in this experiment?

6-153. Which is the stronger base: aniline or cyclohexylamine? Explain.

6-154. Arrange the following substituted anilines in increasing order of reactivity toward acetic anhydride:

(a)

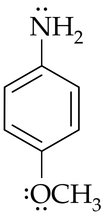

(b)

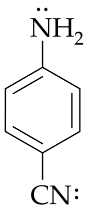

(c)

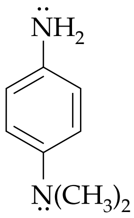

6-155. Suggest a mechanism for the preparation of acetic anhydride from acetic acid and acetyl chloride in the presence of pyridine (an amine base).

6-156. Anhydrides generally react more slowly with an amine than acid chlorides, though both reactions produce amides. Explain this observation.

JWCL196_c06_309-375.qxd 11/17/09 1:42 PM Page 346

EQA

6-157.

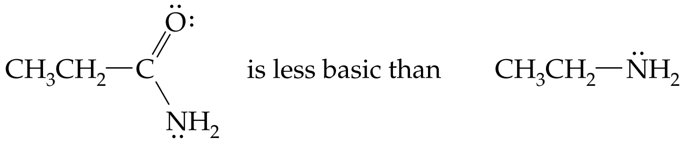

$CH_3CH_2{-}C(=O)NH_2$ is less basic than $CH_3CH_2{-}NH_2$

Explain.

BIBLIOGRAPHY

Review articles:

Beckwith, A. L. J. In *The Chemistry of the Amides;* J. Zabicky, Ed.; Wiley: New York, 1970; p. 73.

Satchell, D. P. N. *Q. Rev.* **1963,** *17,* 160.

Smith, M. B.; March. J. *Advanced Organic Chemistry,* 6th ed.; Wiley: New York, 2007. Chap. 16, p. 1429.

Selected acylation reactions between anhydrides and amines in *Organic Syntheses:*

Cava, M. P.; Deana, A. A.; Muth, K.; Mitchell, M. J. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 944.

Fanta, P. E.; Tarbell, D. S. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 661.

Herbst, R. M.; Shemin, D. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 11.

Jacobs, T. L.; Winstein, S.; Linden, G. B.; Robson, J. H.; Levy, E. F.; Seymour, D. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 456.

Noyes, W. A.; Porter, P. K. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 457.

Wiley, R. H.; Borum, O. H. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 5.

EXPERIMENT 24

Imide Synthesis: N-*Phenylmaleimide*

Common name: *N*-phenylmaleimide
CA number: [941-69-5]
CA name as indexed: 1*H*-pyrrole-2,5-dione, 1-phenyl-

Purpose. To extend the amide synthesis (Experiment [23]) to the preparation of imides. In this experiment, the condensation of a cyclic anhydride with aniline to form an imide is described. The initial reaction to give the carboxylic amide is followed by an intramolecular condensation to produce the desired imide derivative.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallizations (pp. 89–91)
Technique 6A: Thin-Layer Chromatography (pp. 97–99)

REACTION

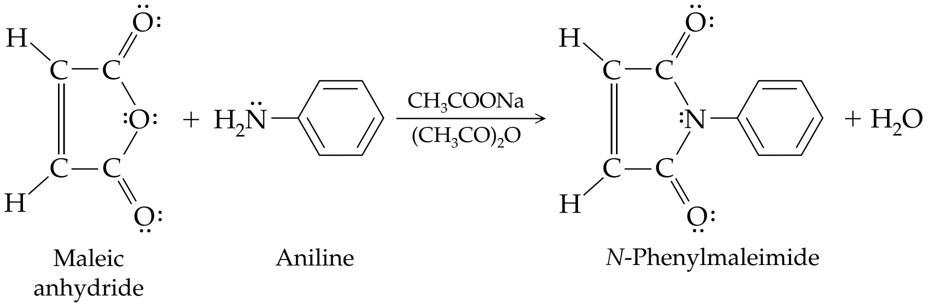

CONFIRMING PAGES aptara

JWCL196_c06_309-375.qxd 11/17/09 1:42 PM Page 347
EQA

DISCUSSION

Imides are diacyl derivatives of ammonia or primary amines. The reaction is similar in its scope and mechanism to the acetylation of aniline or 1,4-phenylenediamine presented in Experiments [23A] and [23B]. As illustrated in the present experiment, cyclic anhydrides produce cyclic imides. Cyclic anhydrides are prepared in Experiments [25A] and [25B]. Derivatives of imides have been suggested for use in the treatment of arthritis, tuberculosis, and epilepsy. Several also have been found to be growth stimulants. Imide-based polymers are used in many applications, including fire-resistant woven fabrics. The *N*-phenylmaleimide prepared in this experiment is also a good dienophile in the Diels–Alder reaction (see Experiments [14] and [15]), and in fact has been used as a reagent to characterize 1,3 dienes.

The first step in the reaction between the primary amine and the cyclic anhydride is an addition–elimination reaction, which involves a nucleophilic attack by the amine on a carbonyl carbon of the maleic anhydride. This results in the formation of an amide and a carboxylic acid, which are linked together to constitute maleanilic acid. Since anhydrides are considerably more reactive toward nucleophiles, to promote ring closure to the imide, the carboxylic acid is then converted to another (mixed) anhydride by reaction with acetic anhydride. This anhydride then undergoes an intramolecular nucleophilic addition–elimination by the amide nitrogen, which gives the desired imide, *N*-phenylmaleimide. The second acylation of a nitrogen nucleophile is much slower than the first. That is, attack of an amide nitrogen on the carbonyl carbon of the anhydride is slower than the attack of the amine nucleophile on the anhydride carbonyl carbon. The mechanistic sequence is given below (R = phenyl):

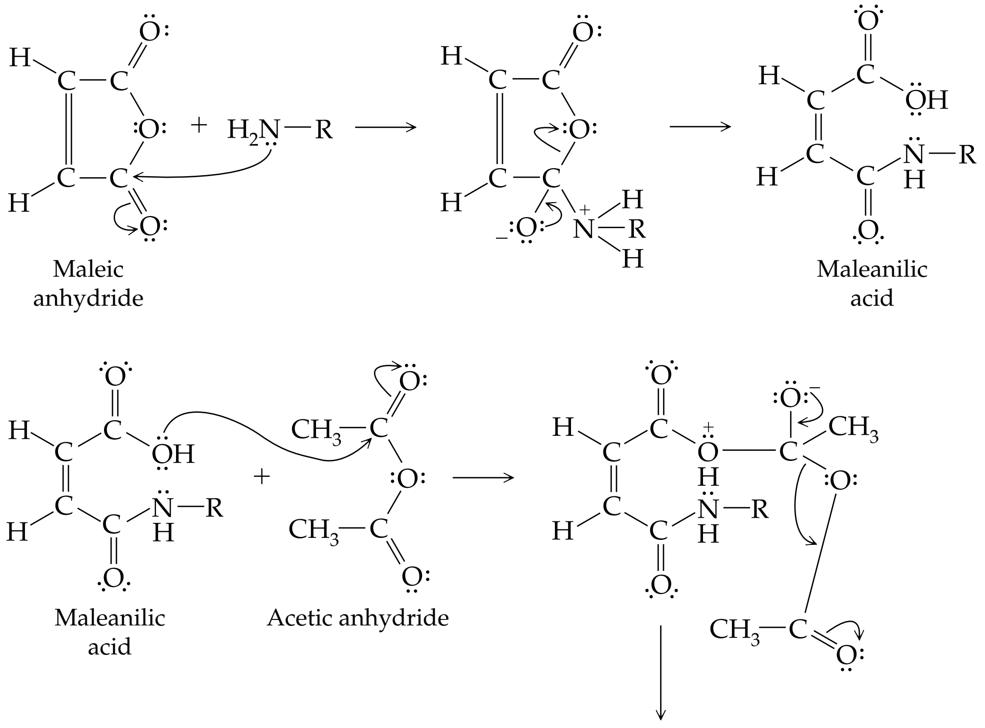

(see next page)

CONFIRMING PAGES aptara

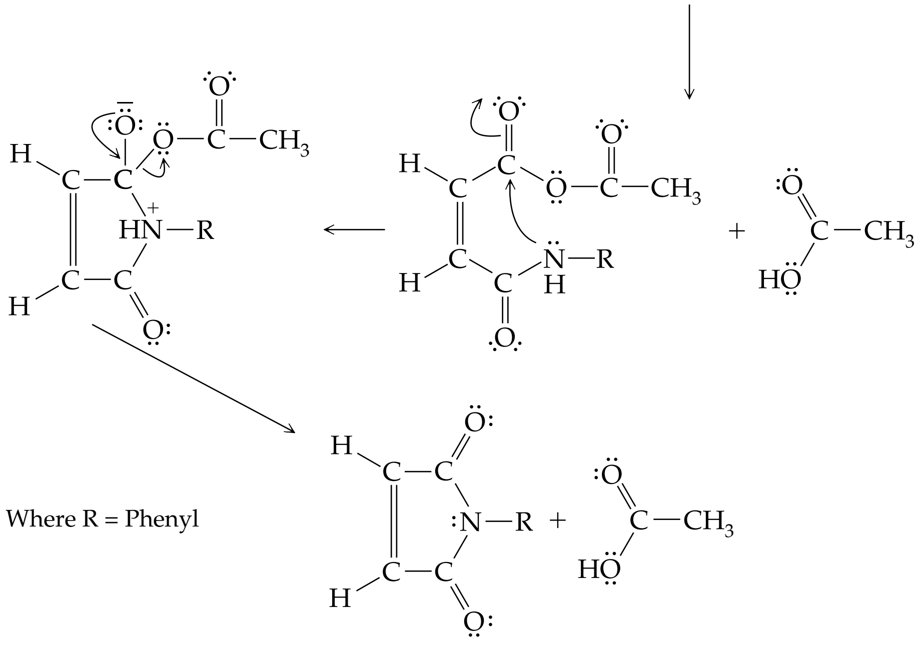

Where R = Phenyl

N-Phenylmaleimide

Experiment 24A

Maleanilic Acid

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

REACTION

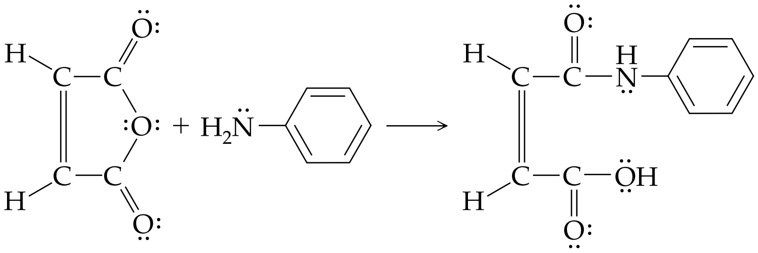

Maleic anhydride

Aniline

Maleanilic acid

NOTE. It is recommended that the purity of the maleic anhydride be checked by mp. The presence of maleic acid can substantially lower the yield of product.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Maleic anhydride	98.06	60 mg	0.61	60			
Acetonitrile	41.05	1.2 mL			81		
Aniline	93.13	56 μL	0.62		184	1.02	1.5863

Reagents and Equipment. In a 3.0-mL conical vial containing a magnetic spin vane, and equipped with an air condenser protected with a drying tube, place 60 mg (0.61 mmol) of maleic anhydride and 1.0 mL of anhydrous acetonitrile (➡). Stir the mixture at room temperature until all the maleic anhydride has dissolved.

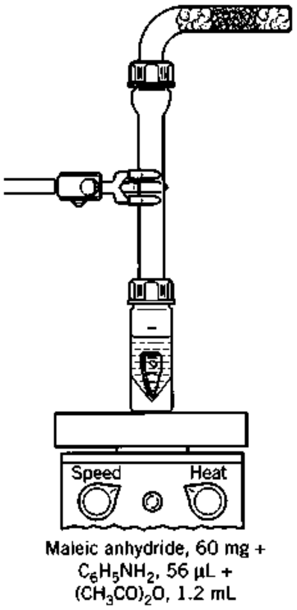

Maleic anhydride, 60 mg + $C_6H_5NH_2$, 56 μL + $(CH_3CO)_2O$, 1.2 mL

HOOD

CAUTION: Dispense this reagent in the **hood** using a calibrated Pasteur pipet.

In a separate, dry $\frac{1}{2}$-*dram* vial prepare a solution of 56 μL (57 mg, 0.62 mmol) of aniline in 100 μL of anhydrous acetonitrile.

HOOD

CAUTION: Dispense these reagents in the **hood** using a calibrated Pasteur pipet. Aniline is highly toxic and is a cancer suspect agent.

Using a Pasteur pipet, add the aniline–acetonitrile solution in one portion to the stirred maleic anhydride–acetonitrile solution. Rinse the $\frac{1}{2}$-*dram* vial with 100 μL of anhydrous acetonitrile and also transfer this rinse to the reaction solution.

Reaction Conditions. After stirring the reaction mixture at room temperature for 15 min, begin monitoring the system by TLC (plates with fluorescent indicator). Using as a solvent system ethyl acetate:hexane (3:2) and a UV lamp for visualization, the R_f value for maleanilic acid is 0.2. The R_f values for aniline and maleic anydride are 0.73 and 0.67, respectively. Once complete as judged by TLC, cool the reaction mixture in an ice bath for 5–10 min.

?

Isolation of Product. Collect the deposit of fine, cream-colored powder by vacuum filtration using a Hirsch funnel (➡). Wash the maleanilic acid crystals with 0.5 mL of cold diethyl ether (calibrated Pasteur pipet), and air-dry them in the funnel for 5 min while maintaining the suction.

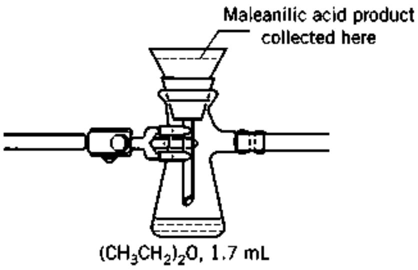

$(CH_3CH_2)_2O$, 1.7 mL

Purification and Characterization. Weigh the maleanilic acid and calculate the percent yield. Determine the melting point and compare your value to that given in Cava et al. (Bibliography section). Obtain an IR spectrum using the KBr pellet technique and compare it with an authentic sample. The air-dried product is suitable for use in the next step without further purification.

N-Phenylmaleimide

Experiment 24B

REACTION

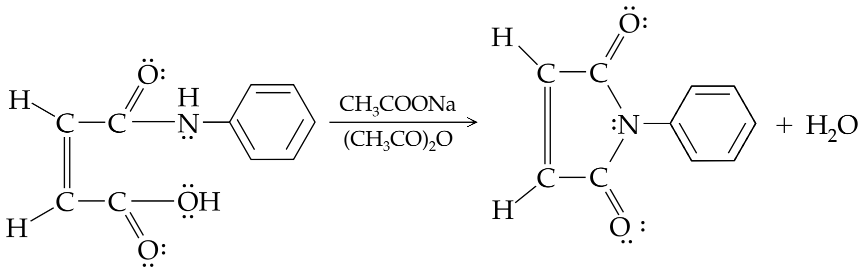

Maleanilic acid
N-Phenylmaleimide

EXPERIMENTAL PROCEDURE

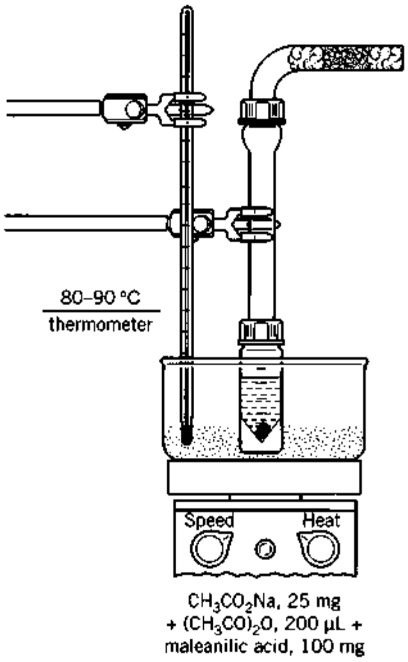

HOOD

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants							
Compound	MW	Amount	mmol	mp (°C)	bp (°C)	d	n_D
Maleanilic acid	191.18	100 mg	0.52	201–202			
Sodium acetate	82.03	25 mg	0.30	324			
Acetic anhydride	102.09	200 μL	2.12		140	1.08	1.3901

Reagents and Equipment. In a 3.0-mL conical vial containing a magnetic spin vane, and equipped with an air condenser protected by a drying tube, place 25 mg (0.30 mmol) of anhydrous sodium acetate and 200 μL (216 mg, 2.12 mmol) of acetic anhydride (⬅).

> CAUTION: Acetic anhydride is corrosive and a lachrymator. It should be dispensed in the **hood** by use of an automatic delivery pipet.

Now add 100 mg (0.52 mmol) of maleanilic acid (prepared in Experiment [24A]) to the reaction vial.

Reaction Conditions. Heat the reaction mixture, with stirring, at a sand bath temperature of 80–90 °C for 30 min. Then cool the resulting mixture to room temperature, add 1.0 mL of cold water (calibrated Pasteur pipet), stir for a few minutes, and then place the vial in an ice bath for 5–10 min to complete crystallization.

Isolation of Product. Collect the solid product by vacuum filtration using a Hirsch funnel and then wash the filter cake with three 0.5-mL portions of cold water (calibrated Pasteur pipet) (⬅). Cover the mouth of the funnel with plastic food wrap and continue the suction for an additional 5–10 min.

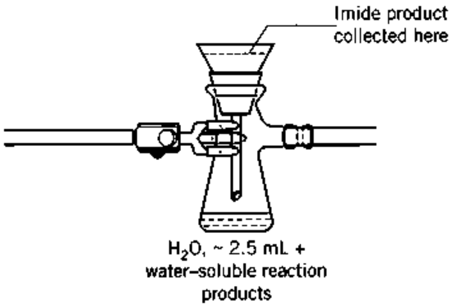

Purification and Characterization. Recrystallize the crude *N*-phenylmaleimide from cyclohexane using the Craig tube, to yield canary-yellow needles. After drying the product on filter paper, or on a porous clay plate, weigh the crystals and calculate the percent yield. Determine the melting point and compare your result with the value given by Cava et al. (Bibliography section). Obtain an IR spectrum and compare it with that of an authentic sample or with that shown in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

QUESTIONS

6-158. As stated in the discussion section, the second step in the reaction to form the imide is much slower than that of the first stage (formation of the acid amide). Explain.

6-159. Phthalimide has a $K_a = 5 \times 10^{-9}$. Write an equation for the reaction of phthalimide with potassium amide (a strong base) in *N*,*N*-dimethylformamide (DMF) solvent. Name the product.

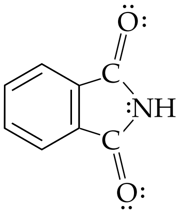

6-160. Predict which of the following species is the most acidic. Explain.

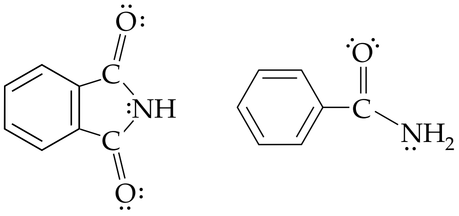

Phthalimide Benzamide

6-161. The phthalimide anion is a strong nucleophile. It can react easily with primary alkyl halides to form substituted phthalimides. One advantage when working with substituted phthalimides is that when treated with hydrazine, primary amines are furnished through this alkylation protocol (Gabriel synthesis). For both synthetic pathways (a) and (b), suggest a suitable mechanism.

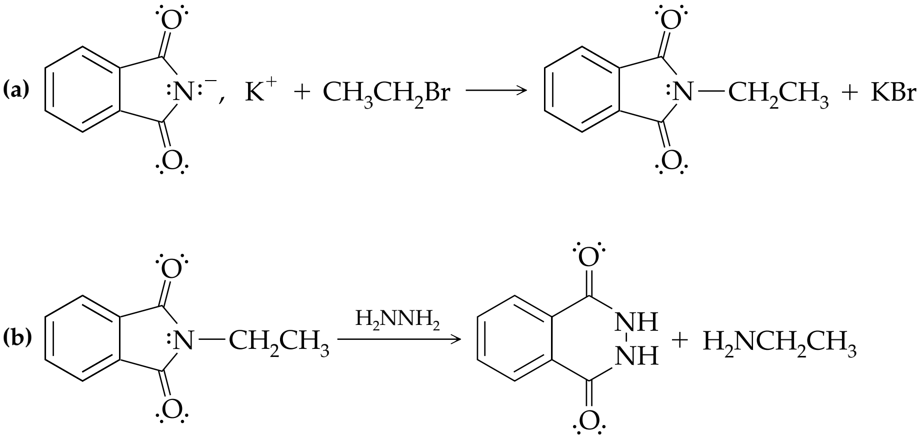

6-162. *N*-Phenylmaleimide, the product prepared in Experiment [24B], can act as a dienophile in the Diels–Alder reaction (see Experiments [14] and [15]). Draw the structure of the product that would be formed by the treatment of *N*-phenylmaleimide with (a) 3-sulfolene under the conditions given in Experiment [14] and (b) furan.

BIBLIOGRAPHY

Review articles on cyclic imides:

Benjamin, E.; Hijji, Y. *Molecules* **2008,** *13,* 157.

Hargreaves, M. K.; Pritchard, J. G.; Dave, H. R. *Chem. Rev.* **1970,** *70,* 439.

Naik, S.; Bhattacharjya, G.; Talukdar, R.; Patel, B. K. *J. Org. Chem.* **2004,** *69,* 1254.

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed.; Wiley-Interscience: New York, 2007, Chap. 16, p. 1429.

Wheeler, O. H.; Rosado, O. In *The Chemistry of the Amides;* Zabicky, J., Ed.; Wiley: New York, 1970, p. 335.

Selected imide preparations in *Organic Syntheses* include

Cava, M. P.; Deana, A. A.; Muth, K.; Mitchell, M. J. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 944.

Noyes, W. A.; Porter, P. K. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 457.

Smith, L. I.; Emerson, O. H. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 151.

Soine, T. O.; Buchdahl, M. R. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 106.

For the preparation of maleanilic acid also see

Ram, R. N.; Varsha, K. *J. Chem. Educ.* **1990,** *67,* 985.

EXPERIMENT 25

Synthesis of Cyclic Carboxylic Acid Anhydrides: Succinic Anhydride and Phthalic Anhydride

Common name: succinic anhydride
CA number: [108-30-5]
CA name as indexed: 2,5-furandione, dihydro-

Common names: phthalic anhydride, benzene-1,2-dicarboxylic anhydride
CA number: [85-44-9]
CA name as indexed: 1,3-isobenzofurandione

Purpose. One of the important methods for preparing cyclic carboxylic acid anhydrides is carried out. The reaction demonstrates the use of acetic anhydride, an important industrial and research chemical, as a dehydrating agent.

Prior Reading

Technique 5: Crystallization
  Use of the Hirsch Funnel (pp. 88–89)
  Craig Tube Crystallizations (pp. 89–91)
Technique 9: Sublimation (pp. 111–114)

REACTION

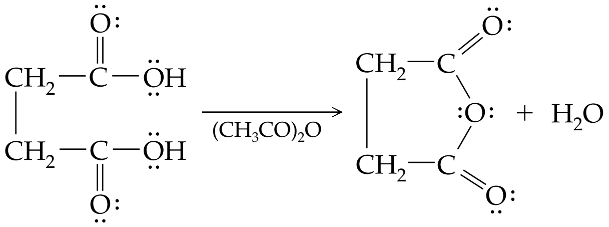

Succinic acid Succinic anhydride

DISCUSSION

Five- and six-membered cyclic anhydrides can be easily formed when the corresponding dicarboxylic acid is heated in the presence of a dehydrating agent. One of the most commonly used dehydrating agents is acetic anhydride. The formation of an anhydride from its corresponding acid by reaction with another anhydride is referred to as *anhydride exchange*.

The similarity of this reaction for the preparation of anhydrides to that for the synthesis of imides (Experiment [24]) should be noted. The ring closure to form the imide is mechanistically related to that of anhydride formation: in one case an amide nitrogen makes a nucleophilic attack on a carbonyl, while in the other, an acid oxygen acts as the nucleophile.

The mechanistic sequence for anhydride exchange is

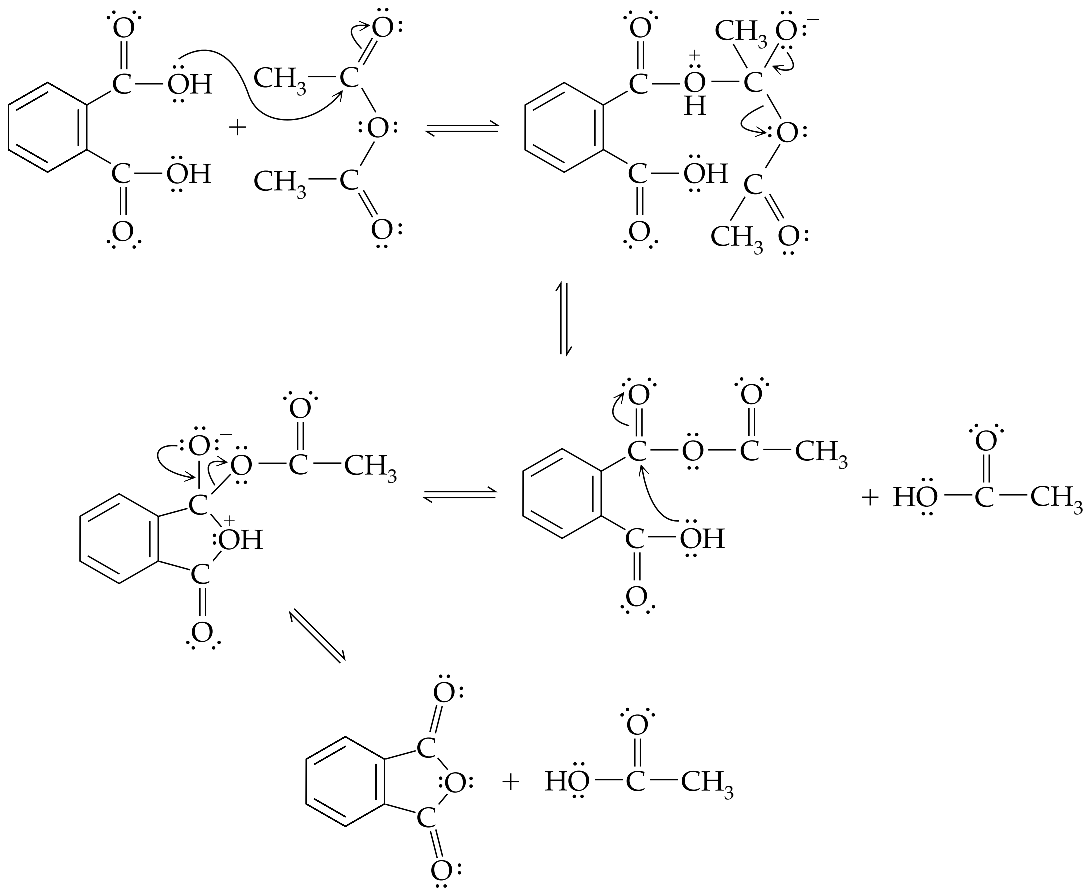

It is possible to prepare five- and six-membered cyclic anhydrides in the absence of acetic anhydride by direct dehydration at elevated temperatures. Maleic anhydride, for example, is easily obtained by this method in greater than 85% yield. Heating of succinic acid at 300 °C yields succinic anhydride in 95% yield.

This equilibrium is driven toward the products since formation of three molecules (two molecules of acetic acid and one molecule of anhydride) is entropically favored over the two molecules of reactants. The equilibrium could be further driven toward the products by distilling off the more volatile acetic acid as it is formed.

Acetic anhydride is an important industrial reagent. Over one-half its annual production of approximately 750,000 tons is used for the manufacture of cellulose acetate. Cellulose acetate is a widely used textile fiber and is the chief component of cigarette filters. Acetic anhydride is the acetylation reagent used for the production of aspirin (acetylsalicylic acid). Succinic anhydride finds use in the succinylation of gelatin used as a blood plasma substitute, as a food preservative in chicken against *Salmonella,* and as a dog food preservative.

Phthalic anhydride finds extensive use in plasticizer formulations for many resins, and in the manufacture of dyes.

Experiment 25A

Succinic Anhydride

The reaction is shown on p. 352.

EXPERIMENTAL PROCEDURE

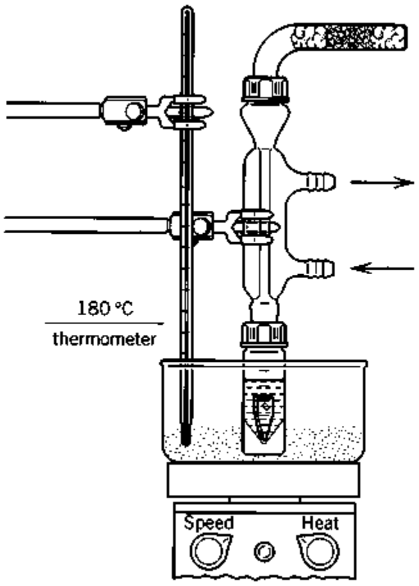

Succinic acid, 150 mg + $(CH_3CO)_2O$, 200 μL

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Succinic acid	118.09	150 mg	1.27	188			
Acetic anhydride	102.09	200 μL	2.12		140	1.08	1.3901

Reagents and Equipment

NOTE. All equipment must be dried in an oven (110 °C) for 30 min before use.

Weigh and place 150 mg (1.3 mmol) of succinic acid in a 1.0-mL conical vial containing a magnetic spin vane. Add 200 μL (2.12 mmol) of acetic anhydride and then attach the vial to a reflux condenser protected by a calcium chloride drying tube (⬅).

HOOD

CAUTION: Acetic anhydride is moisture sensitive and an irritant. Dispense it in the **hood** using an automatic delivery pipet.

Reaction Conditions. Heat the reaction mixture, with stirring, in a sand bath at a temperature of 180 °C for 45 min, timing the reaction from the point at which the succinic acid is completely dissolved.

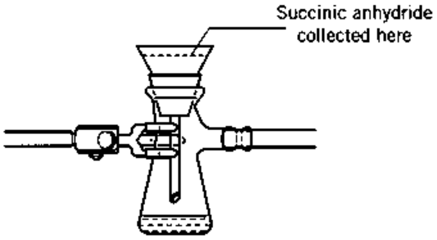

$(CH_3CH_2)_2O$, 1.5 mL + CH_3CO_2H, ~ 100 μL

Isolation of Product. Cool the mixture to room temperature. A voluminous precipitate of succinic anhydride deposits. Further cool the vial in an ice bath for 5 min, and collect the solid material by vacuum filtration using a Hirsch funnel (⬅). Wash the white needles with three 0.5-mL portions of diethyl ether (calibrated Pasteur filter pipet) and then place them on a porous clay plate or filter paper to dry.

Purification and Characterization. The succinic anhydride crystals should be sufficiently pure for characterization. These crystals may, however, be recrystallized from absolute ethanol using a Craig tube.

Weigh the product and calculate the percent yield. Determine the melting point and obtain an infrared spectrum. Compare your results to those listed in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar). What characteristic absorptions do you observe for the anhydride group in the carbonyl region of the spectrum?

Phthalic Anhydride

Experiment 25B

REACTION

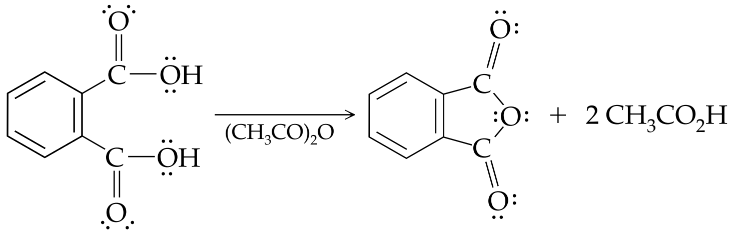

Phthalic acid

Phthalic anhydride

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants							
Compound	MW	Amount	mmol	mp (°C)	bp (°C)	d	n_D
Phthalic acid	166.14	100 mg	0.60	210			
Acetic anhydride	102.09	200 μL	2.12		140	1.08	1.3901

Reagents and Equipment. Weigh and add 100 mg (0.60 mmol) of phthalic acid to a 3.0-mL conical vial containing a magnetic spin vane. Add 200 μL (2.1 mmol) of acetic anhydride and then attach the vial to a reflux condenser protected by a calcium drying tube (➨).

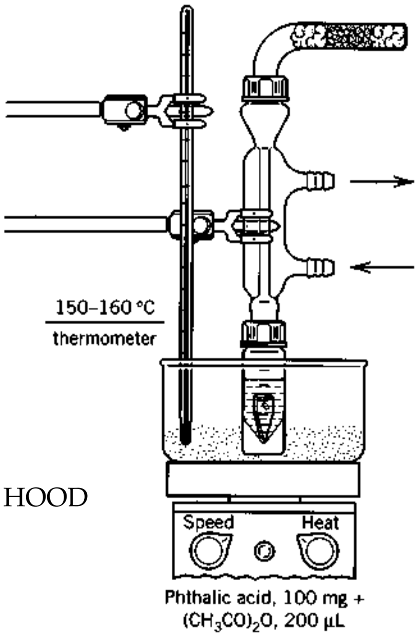

> CAUTION: Acetic anhydride is moisture sensitive and an irritant. Dispense it in the **hood** using an automatic delivery pipet.

Reaction Conditions. Heat the reaction solution, with stirring, at a sand bath temperature of 150–160 °C for 30 min.

NOTE. Position the vial firmly on the bottom of the sand bath vessel to maintain this reaction temperature.

Isolation of Product. Cool the mixture to room temperature, whereupon the product crystallizes from solution. Cool the vial and contents in an ice bath for 10 min and collect the solid by vacuum filtration using a Hirsch funnel (➨). Rinse the filter cake carefully by dropwise addition of 0.5 mL of cold hexane (Pasteur pipet) and continue the suction for several minutes. Complete the drying of the solid product by placing the crystals on a porous clay plate or on filter paper.

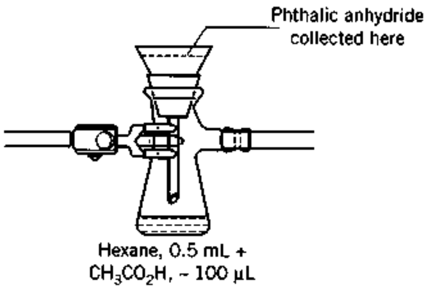

Purification and Characterization. The phthalic anhydride should be sufficiently pure for characterization. It may be purified further by sublimation or by recrystallization from absolute ethanol using the Craig tube.

Weigh the product and calculate the percent yield. Determine the melting point and obtain an infrared spectrum. Compare your results with those reported in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar). What characteristic absorptions do you observe for the anhydride group in the carbonyl region of the spectrum?

QUESTIONS

6-163. As stated in the discussion, direct dehydration can be used as a method for the preparation of five- and six-membered cyclic anhydrides. Propose a suitable mechanism for the reaction below:

$$\text{maleic acid} \xrightarrow{\Delta} \text{maleic anhydride} + H_2O$$

6-164. Propose a suitable mechanism for the formation of the mixed anhydride obtained in the following reaction:

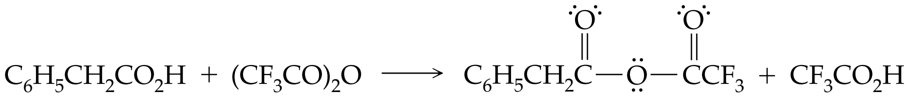

$$C_6H_5CH_2CO_2H + (CF_3CO)_2O \longrightarrow C_6H_5CH_2C(=\ddot{O})-\ddot{O}-C(=\ddot{O})CF_3 + CF_3CO_2H$$

6-165. There are two stereoisomeric 1,3-cyclobutane dicarboxylic acids. One can form a cyclic anhydride, the other cannot. Draw the structures of these compounds and indicate which one can be converted to a cyclic anhydride. Explain.

6-166. When maleic acid is heated to about 100 °C it forms maleic anhydride. However, fumaric acid requires a much higher temperature (250–300°C) before it dehydrates. In addition, it forms only maleic anhydride. Explain.

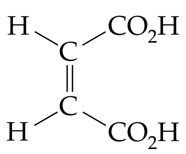

Maleic acid

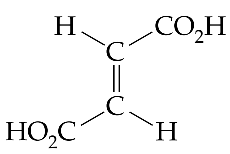

Fumaric acid

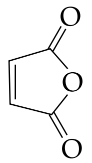

Maleic anhydride

6-167. What product would you expect to obtain from reaction of one equivalent of propanol with phthalic anhydride?

BIBLIOGRAPHY

Selected references from *Organic Syntheses* in which anhydrides are prepared, using acetic anhydride as the dehydrating agent, include

Cason, J. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 630.

Clarke, H. T.; Rahrs, E. J. *Organic Syntheses;* Wiley: New York, 1944; Collect. Vol. I, p. 91.

Grummitt, O.; Egan, R.; Buck, A. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 449.

Horning, E. C.; Finelli, A. F. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 790.

Nicolet, B. H.; Bender, J. A. *Organic Syntheses;* Wiley: New York, 1944; Collect. Vol. I, p. 410.

Shriner, R. L.; Furrow, C. L. Jr. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 242.

The synthesis of succinic anhydride is described in

Fieser, L. F.; Martin, E. L. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 560.

EXPERIMENT 26

Diazonium Coupling Reaction: Methyl Red

Common names: methyl red
CA number: [493-52-7]
CA name as indexed: benzoic acid, 2-[[4-(dimethylamino)phenyl]azo]-

Purpose. In this experiment you will learn the process of generating arenediazonium salts in solution. The arenediazonium salt generated will be used in an electrophilic aromatic substitution reaction (diazo coupling) to prepare an azobenzene derivative. Many azobenzene derivatives, including the one prepared here, have extensively conjugated π-electron systems. Because these are highly colored compounds, they are generally referred to as azo dyes.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallization (pp. 89–91)

REACTION

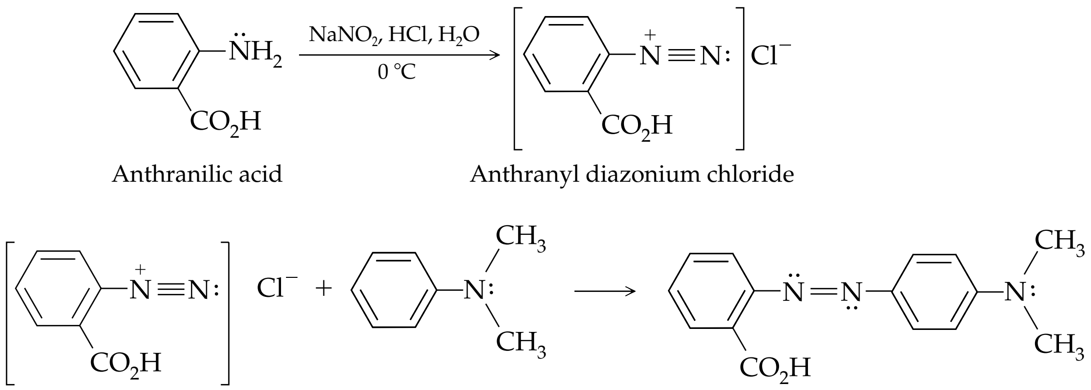

NaNO$_2$, HCl, H$_2$O
0 °C
$\ddot{\mathrm{N}}\mathrm{H}_2$
CO$_2$H
$\overset{+}{\mathrm{N}}\equiv\mathrm{N}$:
Cl$^-$
CH$_3$
N:

Anthranilic acid

Anthranyl diazonium chloride

Anthranyl diazonium chloride

N,*N*-Dimethylaniline

Methyl red

DISCUSSION

The coupling of a diazonium salt to a suitable aromatic substrate is an example of an aromatic electrophilic substitution reaction. When primary aromatic (and also aliphatic) amines ($\mathrm{ArNH_2}$) are treated with nitrous acid ($\mathrm{NaNO_2 + HCl \rightarrow HONO}$), they are converted into diazonium cations, $\mathrm{ArN_2}^+$. In solution, nitrous acid (HONO) is in equilibrium with its anhydride, dinitrogen trioxide ($\mathrm{N_2O_3}$), which is the actual diazotizing agent. The primary amine reacts with the dinitrogen trioxide to form a nitrosamine:

$$\mathrm{Ar{-}NH_2 + N_2O_3} \longrightarrow \underset{\text{(a nitrosamine)}}{\mathrm{Ar{-}NH{-}N{=}\underset{\cdot\cdot}{\dot{O}}}} + \mathrm{HONO}$$

The nitrosamine is in equilibrium with its tautomer, a diazoic acid. The diazoic acid then undergoes dehydration to form the diazonium salt. Diazonium salts are explosive when dry, and therefore are generally not isolated.

$$\underset{\text{(a nitrosamine)}}{\mathrm{Ar{-}\ddot{N}H{-}\ddot{N}{=}\ddot{O}{:}}} + \mathrm{H_3O{:}^+} \rightleftharpoons \underset{\text{(a diazoic acid)}}{\mathrm{Ar{-}\ddot{N}{=}\ddot{N}{-}\underset{\cdot\cdot}{\dot{O}}H}} + \mathrm{H_3O{:}^+}$$

$$\mathrm{Ar{-}\ddot{N}{=}\ddot{N}{-}\underset{\cdot\cdot}{\dot{O}}H} + \mathrm{H_3O{:}^+} \longrightarrow \underset{\text{Diazonium ion}}{\mathrm{Ar{-}\overset{+}{N}{\equiv}N{:}}} + 2\,\mathrm{H_2\ddot{O}{:}}$$

Reaction of the diazonium salt with various aromatic compounds leads to the formation of azo derivatives by what is generally called a "coupling reaction," but is mechanistically simply an ordinary electrophilic aromatic substitution reaction. The mechanism of the reaction is given here:

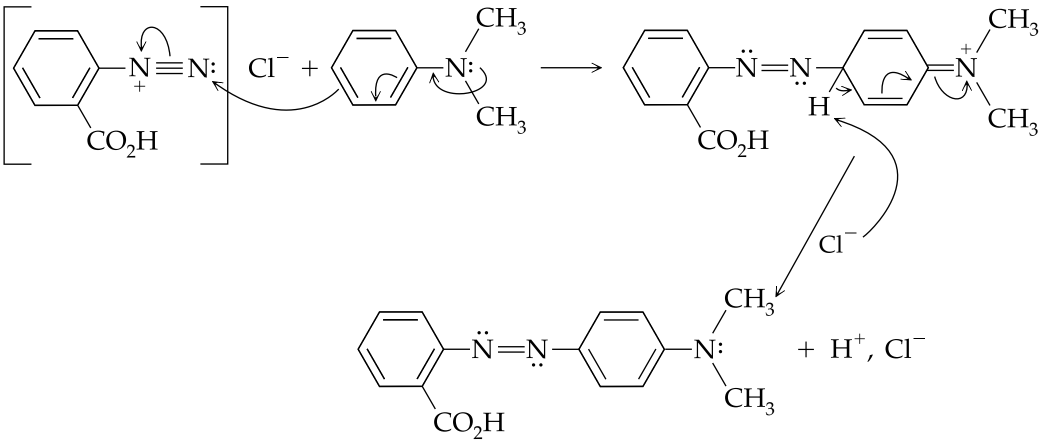
CO$_2$H
Cl$^-$ +
CH$_3$
CH$_3$
H
Cl$^-$
+ H$^+$, Cl$^-$

Azo dyes find use as acid–base indicators. For example, Methyl Red prepared in this experiment, Methyl Orange, and Congo Red are well-known acid–base indicators. Azo dyes are commonly used in the textile, food, and cosmetic industries; FD&C Yellow No. 6, a yellow azo dye is used to color candy, ice cream, beverages, and so on. Several azo dyes (including Butter Yellow and FD&C Red No. 2) have been banned by the FDA from use in foods, drugs, and cosmetics in the United States because of suspected carcinogenic properties.

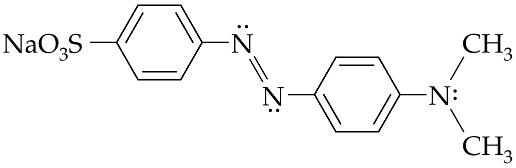
NaO$_3$S
CH$_3$
CH$_3$

Methyl Orange

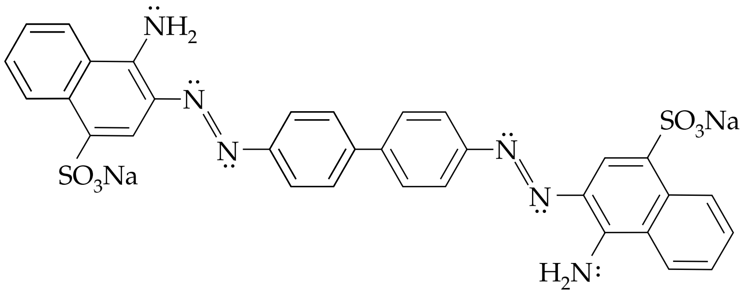
NH$_2$
SO$_3$Na
SO$_3$Na
H$_2$N

Congo Red

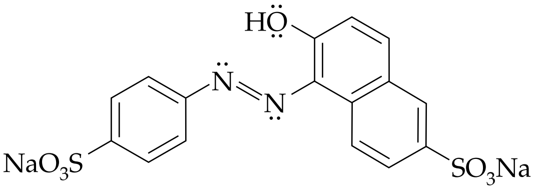
HO
NaO$_3$S
SO$_3$Na

FD&C Yellow No. 6

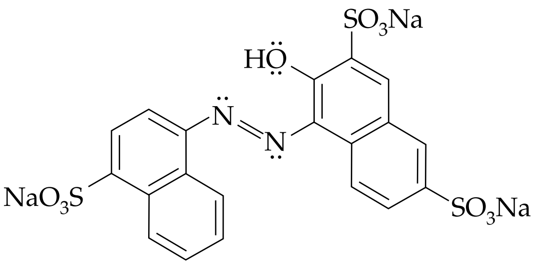
SO$_3$Na
HO
NaO$_3$S
SO$_3$Na

FD&C Red No. 2

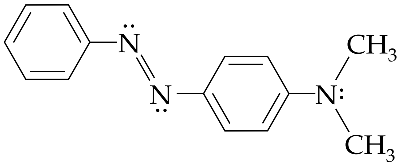

Butter Yellow

EXPERIMENTAL PROCEDURE

Estimated time for the completion of the experiment: 3.0 h.
The reaction is shown on p. 357.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Anthranilic acid	137.14	65 mg	0.47	146–147			
Concd HCl		150 μL					
Water		800 μL					
Sodium nitrite	69.0	36 mg	0.52	271			
N,N-Dimethylaniline	121.18	89 μL	0.71		194	0.96	1.5582
Sodium acetate	82.03	68 mg	0.83	324			
10% aq. NaOH		100 μL					

CAUTION: When dry, benzenediazonium 2-carboxylate detonates violently upon being scraped or heated. It must, therefore, be kept in solution at all times.

Reagents and Equipment. Equip a 3.0-mL conical vial with a magnetic spin vane and an air condenser (➨). Weigh and add 65 mg (0.48 mmol) of anthranilic acid to the vial. Now add a solution of 150 μL of concentrated hydrochloric acid dissolved in 400 μL of water to the vial, using a Pasteur pipet.

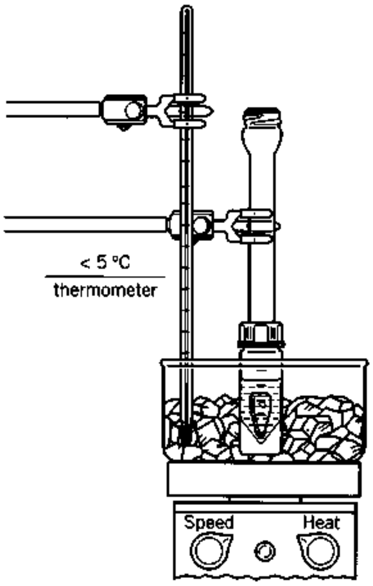

Addition 1:
Anthranilic acid, 65 mg + concd HCl, 150 μL + H_2O, 600 μL + $NaNO_2$, 36 mg
Addition 2:
$C_6H_5N(CH_3)_2$, 85 mg
Addition 3:
CH_3CO_2Na, 68 mg + H_2O, 200 μL

CAUTION: When preparing the acid solution, the acid must be added *to* the water. Dispense these reagents using automatic delivery pipets.

If necessary, warm the mixture, with stirring, on a hot plate magnetic stirrer to obtain a homogeneous solution. Cool the solution in an ice bath, with stirring, for 10 min.

In a 10 × 75-mm test tube, or a small vial, prepare a solution of 36 mg (0.52 mmol) of sodium nitrite dissolved in 200 μL of water. Cool this solution in an ice bath.

Reaction Conditions. When both solutions in the ice bath are cooled to a temperature below 5 °C slowly add (dropwise) the nitrite solution to the stirred anthranilic acid solution, while maintaining the temperature below 5 °C. This transfer is accomplished using a Pasteur pipet. *The solution must be kept cool so that the diazonium salt will not hydrolyze to the corresponding phenol.*

After a period of 4–5 min, check the clear solution of anthranyldiazonium chloride for the presence of excess nitrous acid by placing a drop of the solution on a piece of potassium iodide–starch test paper. If an excess is present, the test paper gives an immediate blue color. If no color is obtained, prepare additional nitrite solution and add as before until a positive test is observed.

Remove the air condenser from the reaction vial containing the solution of anthranyldiazonium chloride. Fairly rapidly, add 89 μL (85 mg, 0.71 mmol) of *N,N*-dimethylaniline (automatic delivery pipet). Reattach the air condenser.

CAUTION: This aniline derivative is toxic and should be dispensed in the **hood.**

HOOD

Stir the solution for an additional 15 min, keeping the temperature below 5 °C.

Prepare a solution of 68 mg (0.83 mmol) of sodium acetate dissolved in 200 μL of water in a 10 × 75-mm test tube. Transfer this solution (Pasteur pipet) to the reaction mixture. Make this addition *without* removing the air condenser. Maintain the resulting solution at 5 °C with stirring, for an additional 20 min.

Remove the reaction vial from the ice bath and allow it to stand for 15 min in order to warm to ambient temperature.

Now add 100 μL of 10% aqueous NaOH solution (automatic delivery pipet) to the solution. Allow the reaction mixture to stand at room temperature for about 30 min. *The formation of the azo compound is a very slow reaction, but the rate of formation is increased by raising the pH of the solution.*

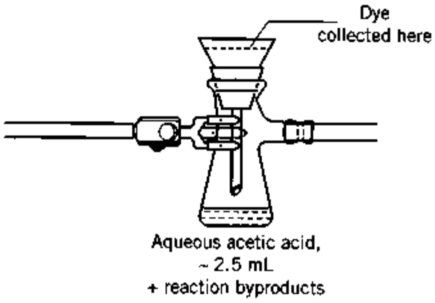

Isolation of Product. Collect the precipitate of crude Methyl Red dye by vacuum filtration using a Hirsch funnel (⬅). Rinse the reaction flask with 0.5 mL of water and use this rinse to wash the crystals. Then wash the crystals with 0.5 mL of 3 M acetic acid, to remove unreacted *N,N*-dimethylaniline from the product, followed by another wash with 0.5 mL of water. *This last wash is usually pale pink in color.*

NOTE. Dispense the small amounts of water and acetic acid using a calibrated Pasteur pipet.

Purification and Characterization. Dissolve the crude product in 500 μL of methanol. If necessary, warm the mixture in a beaker of hot water to aid in the dissolution. Cool the solution in an ice bath and collect the resulting crystals of Methyl Red by vacuum filtration using a Hirsch funnel. Dry the material on filter paper or under vacuum at room temperature.

Weigh the product and calculate the percent yield. Determine the melting point and compare it to the value given in the literature. If further purification is desired, recrystallize the material from toluene using a Craig tube.

QUESTIONS

6-168. In the experiment, a point is made that the formation of the azo compound is a slow reaction, but that the rate is increased by raising the pH of the solution. Why is this necessary? In other words, how does the pH of the solution affect the reactivity of the *N,N*-dimethylaniline reagent?

6-169. In relation to Question 6-168, diazonium salts couple with phenols in slightly alkaline solution. What effect does the pH of the solution have on the reactivity of the phenol?

6-170. Starting with the appropriate aromatic amine and using any other organic or inorganic reagent, outline a synthetic sequence for the preparation of the following azo dyes:

(a)

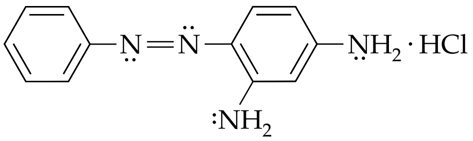

Chrysoidine

(b)

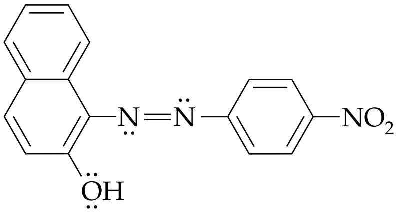

Para Red

(c)

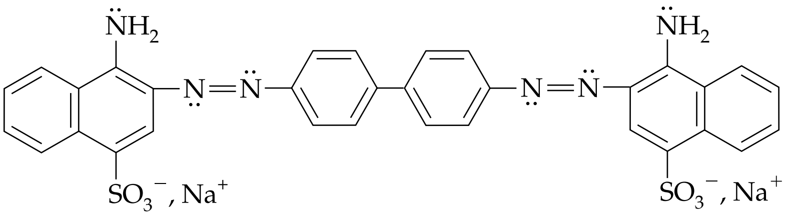

Congo Red

6-171. What is the main structural feature of the azo dyes that causes them to be colored compounds?

6-172. Methyl Orange is an acid–base indicator. In dilute solution at pH > 4.4, it is yellow.

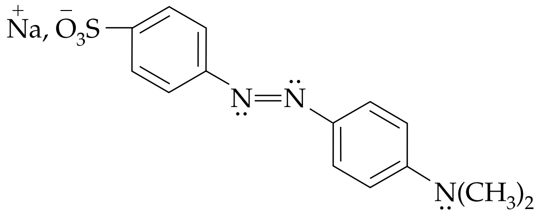

At pH = 3.2 the solution appears red. Draw a structure of the species that is formed at the lower pH if the acid proton adds to the azo nitrogen atom adjacent to the aromatic ring containing —SO_3^- group. Why does the proton add to this particular nitrogen when two other nitrogen atoms are available in the molecule?

BIBLIOGRAPHY

For reviews on diazo compounds and azo dyes see

Gordon, P. F.; Gregory, P. *Organic Chemistry of Colour;* Springer: New York, 1983, p. 95.

Patai, S., Ed. *The Chemistry of Diazonium and Diazo Groups;* Wiley: New York, 1978, Chapters 8, 11, 14.

Saunders, H.; Allen, R. L. M. *Aromatic Diazo Compounds,* 3rd ed.; Edward Arnold: London, 1985.

Smith, M. B.; March. J. *Advanced Organic Chemistry,* 6th ed.; Wiley-Interscience: New York, 2007, Chap. 11, p. 691.

Zollinger, H. *Color Chemistry;* VCH: New York, 1987, p. 85.

Zollinger, H. *Diazo Chemistry I;* VCH: New York, 1994.

Selected coupling reactions with diazonium salts from *Organic Syntheses:*

Cleland, G. H. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 21.

Conant, J. B.; Lutz, R. E.; Corson, B. B. *Organic Syntheses;* 1941; Collect. Vol. I, p. 49.

Fieser, L. F. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 35. ibid., p. 39.

Hartwell, J. L.; Fieser, L. F. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 145.

Santurri, P.; Robbins, F.; Stubbins, R. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 341.

The synthesis of Methyl Red is also given in *Organic Syntheses*:

Clarke, H. T.; Kirner, W. R. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 374.

The present experiment is an adaptation of that given in

Vogel, A. I. *A Textbook of Practical Organic Chemistry,* 5th ed.; Furnis, B. S., *et al.* Eds.; Wiley: New York, 1989.

EXPERIMENT 27

Friedel–Crafts Acylation: Acetylferrocene and Diacetylferrocene

Common name: acetylferrocene
CA number: [1271-55-2]
CA name as indexed: ferrocene, acetyl-

Common names: diacetylferrocene, 1,1′-diacetylferrocene
CA number: [1273-94-5]
CA name as indexed: ferrocene, 1,1′-diacetyl-

Purpose. To investigate the conditions under which the synthetically important Friedel–Crafts acylation (alkanoylation) reaction is carried out. The reaction described here illustrates electrophilic aromatic substitution on an aromatic ring contained in an organometallic compound. The highly colored products are easily separated by both thin-layer and dry-column chromatography.

Prior Reading

Technique 4: Solvent Extraction
- Liquid–Liquid Extraction (p. 72)
- Drying of the Wet Organic Layer (pp. 80–83)

Technique 6: Chromatography
- Column Chromatography (pp. 92–95)
- Thin-Layer Chromatography (pp. 97–99)
- Concentration of Solutions (pp. 101–104)

REACTION

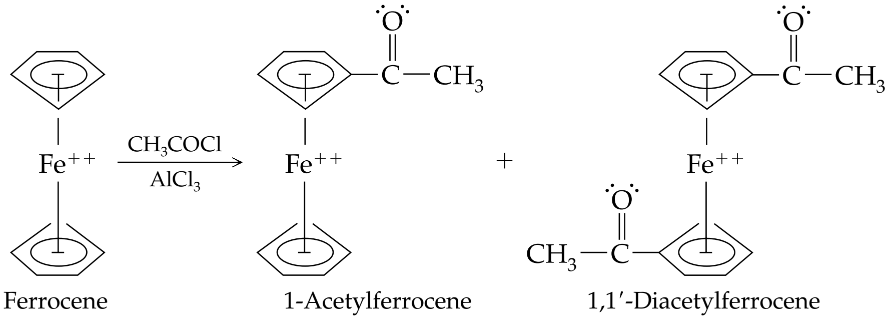

DISCUSSION

The generation of the appropriate electrophile (carbocation, carbocation complex, or acylium ion) in the presence of an aromatic ring system (nucleophile) can lead to alkylation or acylation of the aromatic ring. This set of reactions, discovered by Charles Friedel and James Crafts in 1877, originally used aluminum chloride as the catalyst. The reaction is now known to be catalyzed by a wide range of Lewis acids, including ferric chloride, zinc chloride, boron trifluoride, and strong acids, such as sulfuric, phosphoric, and hydrofluoric acids.

Alkylation is accomplished by use of haloalkanes, alcohols, or alkenes; any species that can function as a carbocation precursor. The alkylation reaction is accompanied by two significant and limiting side reactions: polyalkylation, due to ring activation by the added alkyl groups, and rearrangement of the intermediate carbocation. These lead to diminished yields, and mixtures of products that can be difficult to separate as shown here:

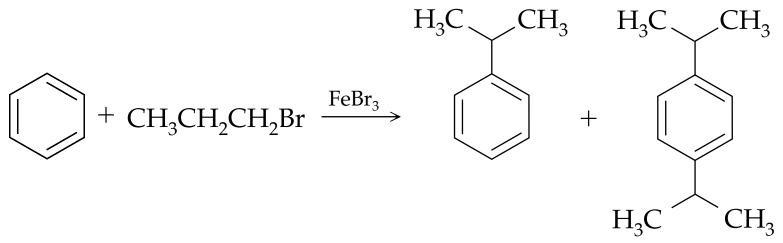

Acylation reactions generally do not suffer from these limitations, and can be conducted using acid chlorides or anhydrides as the electrophilic reagents. Since the introduction of a carbonyl group onto the aromatic ring in an acylation reaction deactivates the ring, the problem of multiple substitution is avoided. The acylium cation, since it is resonance stabilized, is unlikely to rearrange.

The mechanism involves three steps: (1) formation of a cationic electrophile, (2) nucleophilic attack on this electrophile by an aromatic ring, and (3) loss of a proton from the resulting cation to regenerate the aromatic ring system. The mechanism shown here represents the $AlCl_3$ catalyzed generation of the acylium ion electrophile from acetyl chloride (ethanoyl chloride), followed by subsequent nucleophilic attack by the ferrocene ring system:

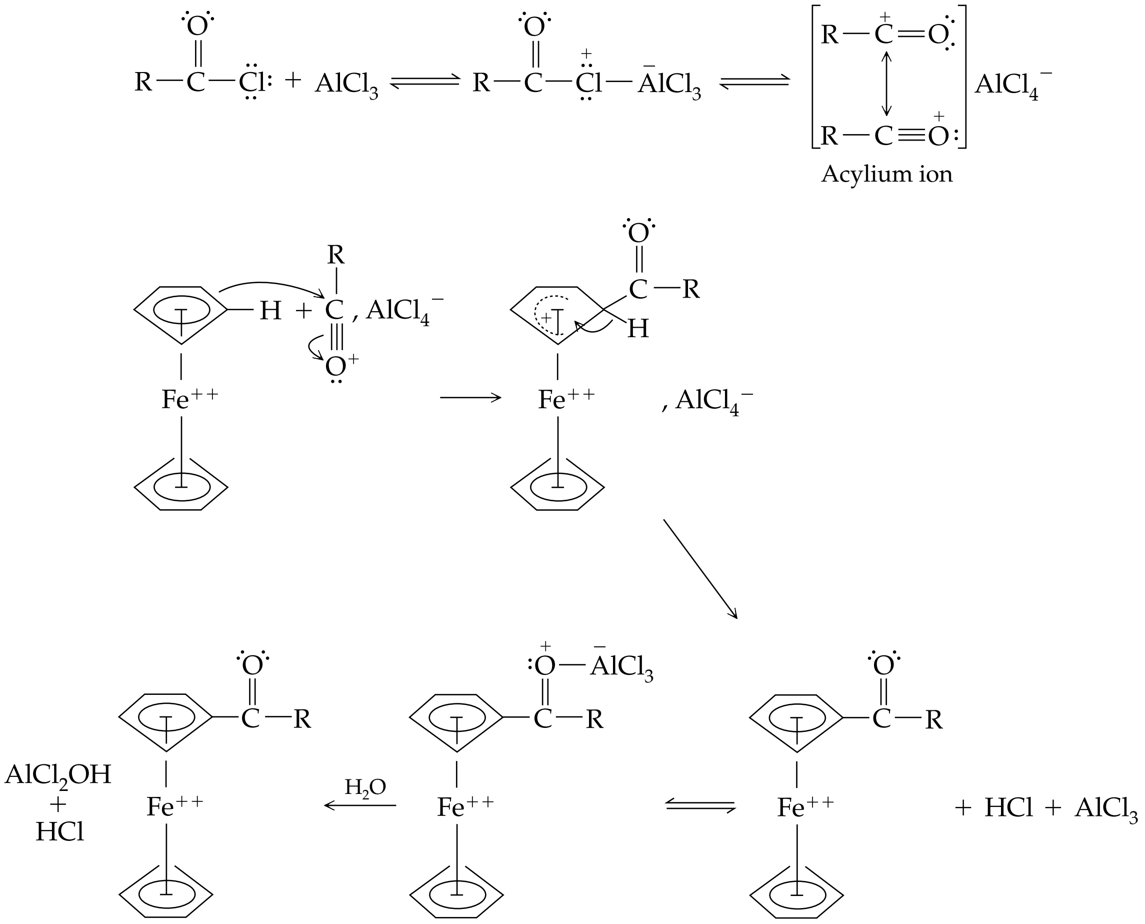

The present experiment also demonstrates the practical value of monitoring reaction progress by TLC analysis.

Charles Friedel (1832–1899) Friedel was Professor of Chemistry at the Sorbonne. He did extensive work on ketones, lactic acid, and glycerol and he discovered isopropyl alcohol. He is best known for his studies of the use of aluminum chloride in the synthesis of aromatic products (**Friedel–Crafts reaction,** 1877). Friedel prepared a series of esters of silicic acid and demonstrated the analogy between the compounds of carbon and silicon, meanwhile confirming the atomic weight of silicon. He determined the vapor densities and molecular weights of the chlorides of aluminum, iron, and gallium.[23]

James Mason Crafts (1839–1917) Crafts was Professor of Chemistry at Cornell University and later at Massachusetts Institute of Technology, where he eventually became President. Crafts studied with Bunsen (Germany) and Wurtz (France) and also worked on the organic compounds of silicon. Crafts was, of course, the codiscoverer of the **Friedel–Crafts reaction.** He also carried out investigations in the area of thermochemistry, catalytic effects in concentrated solutions, and determination of the densities of the halogens at high temperatures.[24]

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: two 3.0-h laboratory periods.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Aluminum chloride	133.34	150 mg	1.12	190			
Acetyl chloride	78.50	80 μL	1.12		51	1.11	1.3898
Ferrocene	186.04	100 mg	0.54	173			
Methylene chloride		4.0 mL			40		

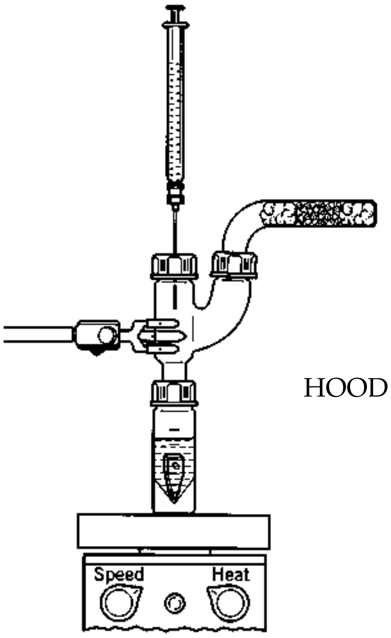

$AlCl_3$, 150 mg + CH_2Cl_2, 5.5 mL + CH_3COCl, 80 μL + ferrocene, 100 mg

Reagents and Equipment

NOTE. Dry the glassware in an oven at 110 °C for 30 min and allow it to cool in a desiccator before starting the experiment.

Equip a tared 5.0-mL conical vial containing a magnetic spin vane with a Claisen head protected by a calcium chloride drying tube and a septum cap. Weigh and add 150 mg (1.12 mmol) of fresh, anhydrous aluminum chloride (⬅).

HOOD

Using a calibrated Pasteur pipet, add 2.5 mL of methylene chloride to the reaction vial. In the **hood,** with swirling, add 80 μL (1.12 mmol) of acetyl chloride from an automatic delivery pipet. Use a syringe to add a solution of 100 mg (0.54 mmol) of ferrocene dissolved in 1.5 mL of methylene chloride to the resulting mixture.

NOTE. Use a capped vial and recap it between the addition of each reagent. After addition of the acetyl chloride, attach the vial to the Claisen head and add the ferrocene solution through the septum as shown in the figure. Do this in one or two portions, depending on whether a 1- or 2-mL syringe is used.

---

[23] See *Berichte* **1899,** *32,* 372; Crafts, J. M. *J. Chem. Soc.* **1900,** *77,* 993; *Bull. Soc. Chim. Fr.* **1900,** *23,* 1; Béhal, A. *ibid.,* **1932,** *51,* 1423; Willemart, A. *J. Chem. Educ.* **1949,** *26,* 3.

[24] Ashdown, A. A. *J. Chem. Educ.* **1928,** *5,* 911; Talbot, H. P. *J. Am. Chem. Soc.* **1917,** *39,* 171; Richards, T. W. *Proc. Am. Acad.* **1917–1919,** *53,* 801; Cross, C. R. *J. Natl. Acad. Sci*. **1914,** *9,* 159.

At this stage, the reaction mixture turns a deep-violet color.

CAUTION: It is important to minimize the exposure to moist air during these transfers. Both the aluminum chloride and the acetyl chloride are highly moisture sensitive so that rapid, yet accurate, manipulations are necessary to minimize deactivation of these reagents, which leads to poor results. In addition, both chemicals are irritants. Avoid breathing the vapors or allowing the reagents to come in contact with skin. These reagents must be dispensed in the **hood.** HOOD

TLC Sample Instructions. Obtain an aliquot for TLC analysis by removing a small amount of the reaction mixture by touching the open end of a Pasteur pipet to the surface of the solution. First, remove the cap from the straight neck of the Claisen head, and then insert the pipet down the neck so as to touch the surface of the solution. Dissolve this aliquot in about 10 drops of cold methylene chloride in a small capped vial. Mark the vial and *save* it for TLC analysis.

Reaction Conditions. Following the addition of the ferrocene solution, note the time, and begin stirring. Allow the reaction to proceed at room temperature for 15 min.

Isolation of Product. Quench the reaction by transferring the mixture by Pasteur pipet to a 15-mL capped centrifuge tube (or a 15-mL screw-capped vial) containing 5.0 mL of ice water. Cool the tube in an ice bath and neutralize the resulting solution by dropwise addition (calibrated Pasteur pipet) of about 0.5 mL of 25% aqueous sodium hydroxide.

NOTE. Avoid an excess of base. Use litmus or pH paper to confirm the neutralization.

Now extract the mixture with three 3-mL portions of methylene chloride. Cap the tube, shake, vent, and allow the layers to separate (a Vortex mixer may be used in this step). Remove the lower (methylene chloride) layer using a Pasteur filter pipet. Combine the methylene chloride extracts in a 25-mL Erlenmeyer flask, and dry the wet solution over about 200 mg of granular anhydrous sodium sulfate for 20 min. Transfer the dried solution to a tared 10-mL Erlenmeyer flask, using a Pasteur filter pipet, in aliquots of 4 mL each. After each transfer, concentrate the solution, in the **hood,** under a stream of dry nitrogen gas in a warm sand bath to a volume of about 0.5 mL. Rinse the drying agent with an additional 2.0 mL of methylene chloride and combine this rinse with the concentrate. Remove several drops of this solution by Pasteur pipet and place them in a capped vial containing 10 drops of *cold* methylene chloride. Mark the vial and *save* it for TLC analysis. Remove the remaining solvent by warming in a sand bath in the **hood** to yield the crude, solid product (~130 mg). Weigh the residue.

HOOD

HOOD

*OPTION. As an option, the combined extracts and washes may be left to evaporate in the **hood** in a 25-mL Erlenmeyer flask, with the mouth covered by filter paper, until the following week.*

HOOD

If the reaction is performed over a 2-week period, this is a convenient point at which to stop. However, if time permits, perform the TLC analysis now.

Thin-Layer Chromatographic Analysis. Use TLC to analyze the two samples saved above. Also analyze a standard mixture of the substituted ferrocenes (supplied by the instructor) at the same time. Use the developed TLC plates as a guide to determine the product mixture obtained in the reaction and as an aid in determining the appropriate elution solvent required for separation of the mixture by dry-column chromatography.

INFORMATION. Good results have been achieved by conducting the TLC analysis with Eastman Kodak silica gel–polyethylene terephthalate plates (#13179). Activate the plates at an oven temperature of 100 °C for 30 min. Place them in a desiccator for cooling and storing until used. Elute the plates using pure methylene chloride as the elution solvent. Visualization of unreacted ferrocene can be enhanced with iodine vapor. See Prior Reading for methods of TLC analysis and determination of R_f *values.*

HOOD

Purification and Characterization. Now purify the reaction products formed in the reaction by dry-column chromatography. The term *dry-column chromatography* refers to the fact that the column is packed with dry alumina, rather than with a slurry (see Prior Reading). Dissolve the solid product residue isolated above in 0.5 mL (calibrated Pasteur pipet) of methylene chloride in a small vial. Mix this solution with 300 mg of alumina (activity III, see Glossary) in a tared vial, and evaporate the solvent under a stream of dry nitrogen in the **hood** to give a product–alumina mixture. Assemble a chromatographic buret column in the following order (bottom to top): prewashed cotton plug, 5 mm of sand, 60–80 mm of alumina (~5.0 g, activity III), *one-half* of the product–alumina mixture, and 10 mm of alumina (←).

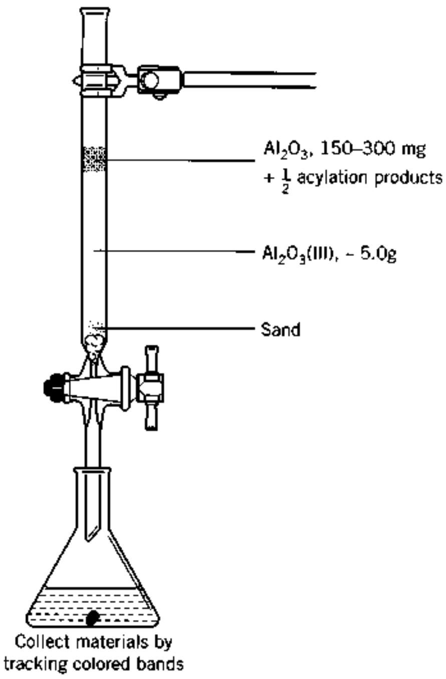
Al_2O_3, 150–300 mg
+ $\frac{1}{2}$ acylation products
Al_2O_3(III), ~ 5.0g
Sand
Collect materials by tracking colored bands

NOTE. This procedure prevents overloading of the chromatographic column during the separation of the reaction products. If the yield of crude reaction products exceeds 75 mg (the usual case), introduce one-half of the alumina–product mixture to the column. If the crude products, however, are obtained in quantities of less than 75 mg, add the entire alumina–product mixture to the top of the column. If only one-half of the alumina–crude ferrocene acylation product mixture is placed on the column, it is important to reweigh the tared vial to establish a reasonably accurate estimate of the overall yields obtained in the reaction.

HOOD

Given the polar nature of the alumina, the products will elute in order of increasing polarity: ferrocene followed by acetylferrocene, followed by diacetylferrocene. Begin elution of the column with pure hexane if TLC analysis indicates that unreacted ferrocene is present in the product mixture. Be sure to add the initial solvent down the side of the column so as not to disturb the alumina bed. During elution, the ferrocenes will separate into two or three bands of different colors on the column. The volume of each eluted fraction should be in the range of 2–5 mL if the band is carefully tracked down the column. Once the ferrocene band has been collected, continue the elution with a 1:1 mixture of CH_2Cl_2/hexane to obtain the monosubstituted product. Further elution with a 9:1 mixture of CH_2Cl_2/CH_3OH will elute the disubstituted material. Collect and save each chromatographic band separately. Store the solvent that elutes without color in an Erlenmeyer flask or beaker until you have isolated all your product. In the **hood,** remove the solvent under a stream of dry nitrogen gas, using a warm sand bath. During concentration of the solvent, spot each fraction on a TLC plate to verify the separation and purity of its contents.

Determine the melting point of each of the isolated products and compare your results to those reported in the literature.

Obtain an IR spectrum of each material and compare the results with an authentic sample or spectra found in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar). Interpretation of the spectra allows an unambiguous determination of substitution based on the presence or absence of absorption in the 1100- to 900-wavenumber region of the spectrum.

Characterization of the Fractions (Total Sample) Isolated From the Reaction Workup

Acetylferrocene: mp __________ °C; __________ mg, __________ mmol

Diacetylferrocene: mp __________ °C; __________ mg, __________ mmol

Total: __________ mmol, __________ % yield

QUESTIONS

6-173. In the formation of diacetylferrocene, the product is always the one in which each ring is monoacetylated. Why is no diacetylferrocene produced in which both acetyl groups are on the same aromatic ring?

6-174. Ferrocene cannot be nitrated using the conventional HNO_3–H_2SO_4 mixed acid conditions, even though nitration is an electrophilic aromatic substitution reaction. Explain.

6-175. In contrast to nitration (Question 6-174), ferrocene undergoes the acetylation and sulfonation reaction. Explain.

6-176. The bonding in ferrocene involves sharing of the 6 electrons from each cyclopentadienyl ring with the iron atom. Based on the electronic configuration of the iron species in the compound, show that a favorable 18-electron inert gas configuration is established at the iron atom.

6-177. In a manner similar to that in Question 6-176, predict whether ruthenocene and osmocene (the ferrocene analogues of ruthenium and osmium) would be stable compounds? Explain.

6-178. Would you predict that bis(benzene)chromium (0) would be a stable compound? Explain.

6-179. Predict the major product(s) in each of the following Friedel–Crafts reactions. Name each product. If the reaction does not occur, offer a reasonable explanation for that fact.

(a)

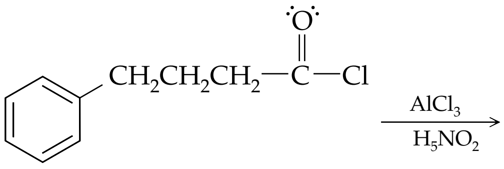

(b)

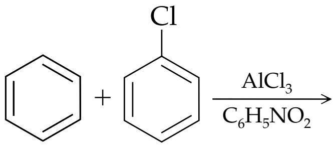

(c)

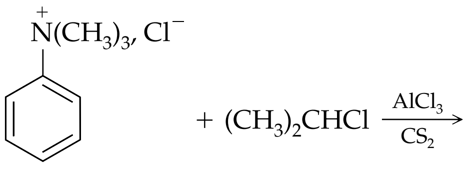

(d)

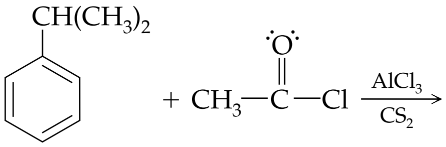

(e)

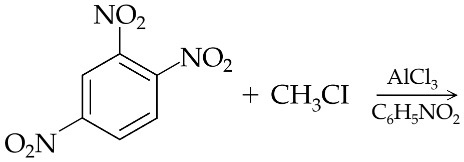

(f)

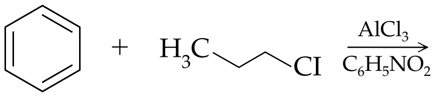

BIBLIOGRAPHY

The acetylation of ferrocene has been monitored using chromatographic techniques. Several references are listed:

Amenta, D. S.; DeVore, T. C.; Gallaher, T. N.; Zook, C. M. *J. Chem. Educ.* **1996,** *73,* 575.

Bohen, J. M.; Joullié, M. M.; Kaplan, F. A. *J. Chem. Educ.* **1973,** *50,* 367.

Nerwith, T. L.; Srouji, N. *J. Chem. Educ.* **1995,** *72;* 455.

Several reviews on the Friedel–Crafts reaction, selected from many cited in the literature:

Gore, P. H. *Chem. Rev.* **1955,** *55,* 229.

Harada, T; Ohno, T.; Kobayashi, S.; Mukaiyama, T. *Synthesis* **1991,** 1216.

Heaney, H. *Comp. Org. Syn.* **1991,** *2,* 733.

Olah, G. A., Ed., *Friedel–Crafts and Related Reactions;* Interscience: New York, 1963–1965; Vols. I–IV.

Smith, M. B.; March. J. *Advanced Organic Chemistry,* 6th ed.; Wiley-Interscience: New York, 2007, Chap. 11, p. 719.

Several of the large number of Friedel–Crafts acetylation reactions given in *Organic Syntheses:*

Adams, R.; Noller, C. R. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 109.

Arsenijivic, L; Arsenijivic, V.; Horeau, A.; Jacques, J. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 34.

Fieser, L. F. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 517.

Fieser, L. F. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 6.

Marvel, C. S.; Sperry, W. M. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 95.

Olson, C. E.; Bader, A. F. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 898.

Sims, J. J.; Selman, L. H.; Cadogan, M. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 744.

EXPERIMENT
28

Halogenation: Electrophilic Aromatic Substitution to Yield 4-Bromoacetanilide

Common names: 4-bromoacetanilide, *p*-bromoacetanilide
CA number: [103-88-8]
CA name as indexed: acetamide, *N*-(4-bromophenyl)-

Purpose. This experiment extends our understanding of the experimental conditions under which electrophilic aromatic substitution reactions are carried out (also see Experiments [27] and [29A]–[29D]). It deals with electrophilic aromatic halogenation. The directive influence of the acetamido, —$NHCOCH_3$, group on the bromination of acetanilide is explored.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch funnel (pp. 88–89)
Craig Tube Crystallization (pp. 89–91)
Technique 6A: Thin-Layer Chromatography (pp. 97–99)

REACTION

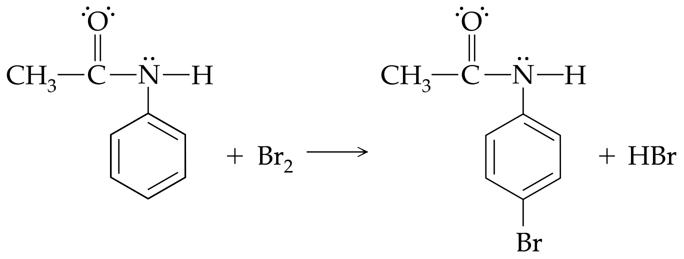

Acetanilide Bromine 4-Bromoacetanilide

DISCUSSION

Aromatic compounds may be brominated by treatment with bromine in the presence of a Lewis acid catalyst, such as ferric chloride. For very electron-rich aromatic rings, such as arylamines, the reaction may proceed in the absence of a catalyst. With amines or phenols, in many cases, it is difficult to stop the bromination at monosubstitution, and all open ortho and para positions are brominated. For this reason, primary aromatic amines are often converted to a corresponding amide derivative, if a monobrominated product is desired. This strategy is demonstrated in the present experiment. The —$NHCOCH_3$ group is a less powerful *o*,*p*-directing group than —NH_2 due to the presence of the electron-withdrawing carbonyl group, which renders the ring less nucleophilic. Electrophilic substitution by bromine is still, however, effectively directed electronically to the ortho and para positions on the ring. The acetamido group, —$NHCOCH_3$, effectively blocks the ortho positions by steric hindrance. For these reasons, only para substitution is observed. The acetanilide used in this experiment may be prepared using the procedure described in Experiment [23A].

The mechanism of the bromination reaction is a classic illustration of an electrophilic substitution on an aromatic ring. The mechanism shown below is presented as proceeding without the aid of a catalyst:

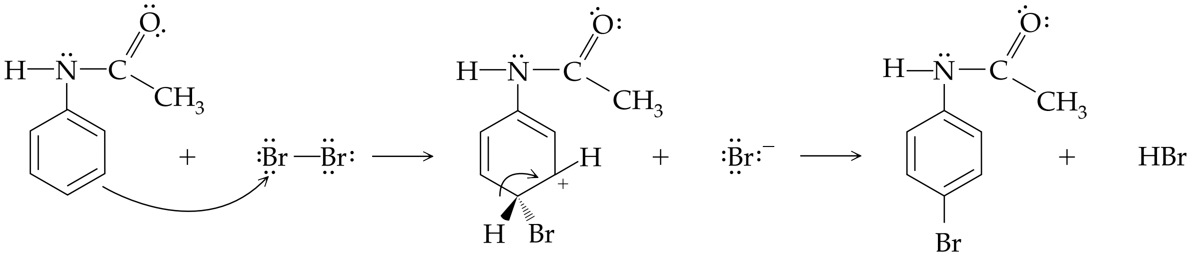

EXPERIMENTAL PROCEDURE

Estimated time for completion of the experiment: 1.5 h.

INSTRUCTOR PREPARATION. For the bromine/AcOH solution. Prepare the bromine–acetic acid reagent by mixing 2.5 mL of bromine with 5.0 mL of glacial acetic acid in the ***hood.***

HOOD

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
Acetanilide	135.17	25 mg	0.19	114	
Glacial acetic acid	60.05	4 drops			118
Bromine–acetic acid solution		3 drops			

Reagents and Equipment. Weigh and place 25 mg (0.19 mmol) of acetanilide in a 3.0-mL conical vial fitted with a cap. Add between 8–10 drops of glacial acetic acid using a medicine dropper. Stir with a glass rod to help

HOOD

dissolve the acetanilide. Once in solution, add a magnetic spin. Now, in the **hood,** add to the clear solution three drops of the bromine–acetic acid solution. Cap the vial immediately.

HOOD

CAUTION: Bromine is a severe irritant. Bromine burns can be severe and require a long time to heal. Always wear plastic gloves and dispense the bromine solution in the **hood.**

Reaction Conditions. Allow the reddish-brown solution to stand at room temperature for 10 min. During this period, yellow-orange colored crystals precipitate from the solution.

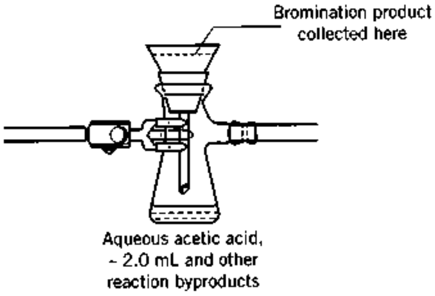
Bromination product collected here
Aqueous acetic acid, ~ 2.0 mL and other reaction byproducts

Isolation of Product. Using a calibrated Pasteur pipet add 0.5 mL of water to the reaction mixture with swirling, followed by 5 drops of aqueous sodium bisulfite solution (33%). This treatment destroys the unreacted bromine (and its residual color) and results in white crystals. Cool the reaction mixture in an ice bath for 10 min to maximize the yield of product.

Collect the crude reaction mixture by vacuum filtration using a Hirsch funnel (⬅). Once collected, add 0.25 mL of cold water (using a calibrated Pasteur pipet) to the 3.0-mL conical vial. Transfer this rinse to the filter cake and repeat two more times (rinse followed by washing of the filter cake with 0.25-mL portions of cold water). Once complete, partially dry the filter cake by drawing air through the crystals under reduced pressure for approximately 5 min. A sheet of plastic food wrap over the funnel mouth aids this process (see Prior Reading).

V

Purification and Characterization. Purify all but 10 mg of the crude 4-bromoacetanilide by recrystallization from 95% ethanol using the Craig tube. Weigh the dried product and calculate the percent yield. Determine the melting point, compare both the crude and recrystallized materials by TLC analysis (plates with fluorescent indicator, R_f value for 4-bromoacetanilide is 0.32 using as solvent system ethyl acetate:hexane (2:3), and a UV lamp for visualization purposes), and compare your observed melting point to the value given in the literature.

Obtain an IR spectrum of the material and compare it with one found in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar). If possible, obtain ^{1}H and/or ^{13}C NMR spectra of your material in DMSO-d_6.

Nuclear Magnetic Resonance Analysis. Figures 6.46 and 6.47 are, respectively, the ^{1}H and ^{13}C NMR spectra of *p*-bromoacetanilide in DMSO-d_6. These can be used to compare with the NMR spectra you may obtain of your product.

In the ^{13}C NMR spectrum, the DMSO-d_6 appears as a septet at 39.7 ppm. The resonance from the methyl group of the *p*-bromoacetanilide occurs at 24 ppm and the amide carbonyl carbon resonates at 169 ppm. The carbon atoms of the benzene ring are observed between 110 and 140 ppm.

In the ^{1}H spectrum, the peak from trace amounts of DMSO-d_5 is seen at about 2.6 ppm. The peak at 3.4 ppm is probably due to water or another impurity in the sample. Note the two small peaks located equidistant to the tall singlet near 2.0 ppm. The small "satellite" peaks are the result of the 1.1% of the methyl groups that have ^{13}C instead of ^{12}C, and thus here coupling between the carbon and the protons is observed. The two doublets for the aromatic protons are observed near 7.5 ppm. The amide NH proton, which is probably hydrogen bonded to the basic (Lewis) sulfoxide functional group in DMSO-d_6,

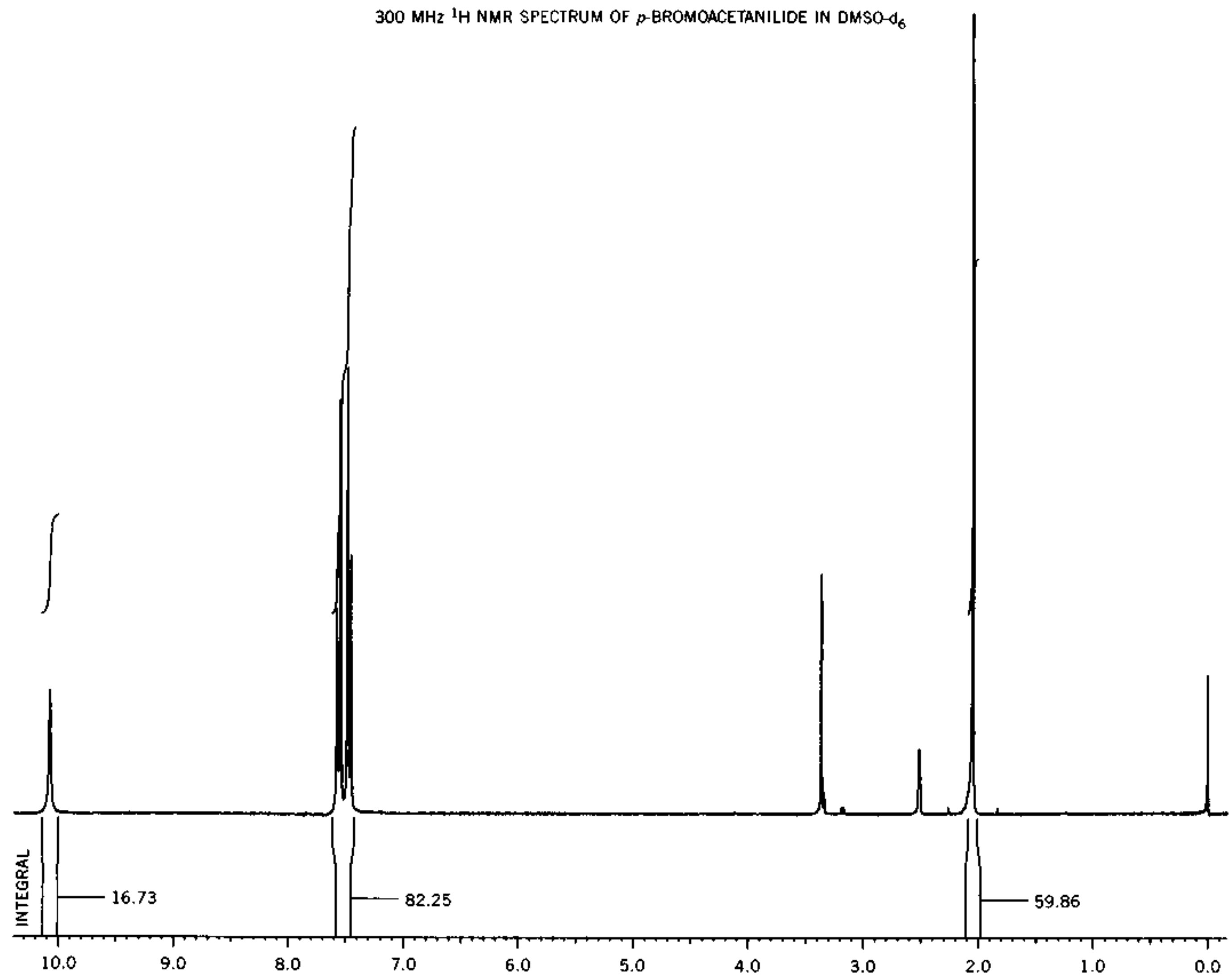

Figure 6.46 ^{1}H-NMR spectrum: *p*-bromoacetanilide.

occurs rather downfield, near 10.1 ppm.This chemical shift may vary in your sample due to subtle differences in concentration, temperature, and moisture content of your DMSO-d_6.

Chemical Tests. Chemical classification tests may also be performed on the amide product.The ignition and the Beilstein test (Chapter 9) are used to confirm the presence of the aromatic ring and the halogen group, respectively. Does the hydroxamate test for amides (Chapter 9) give a positive result?

QUESTIONS

6-180. Use resonance structures to show why the group shown is a less powerful ortho–para directing group than the —NH_2 group:

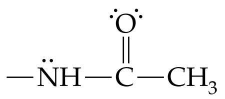

$$-\ddot{N}H-\overset{\overset{\displaystyle \ddot{O}}{\|}}{C}-CH_3$$

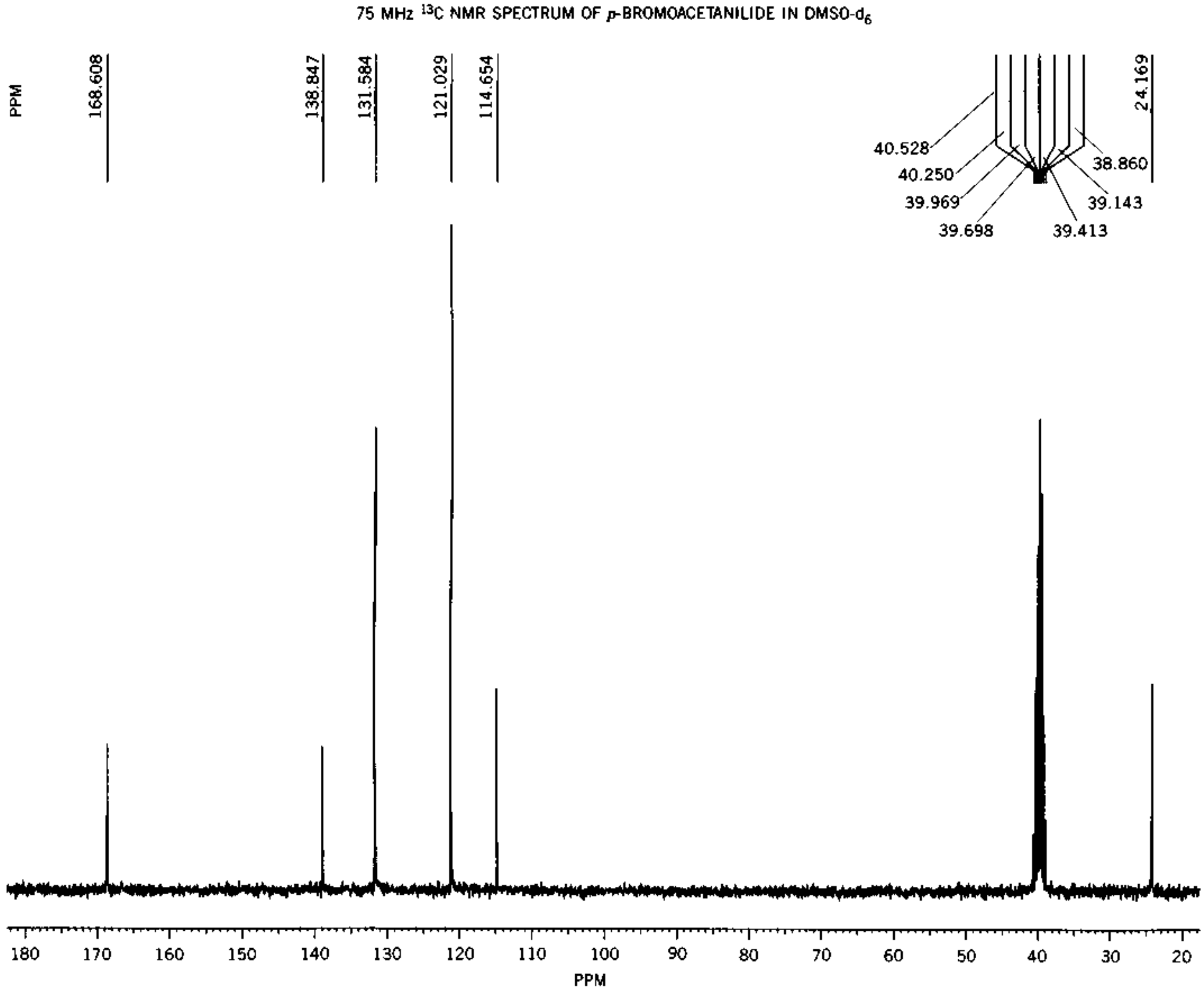

Figure 6.47 ^{13}C-NMR spectrum: *p*-bromoacetanilide.

6-181. Benzene is brominated in the presence of $FeBr_3$ catalyst:

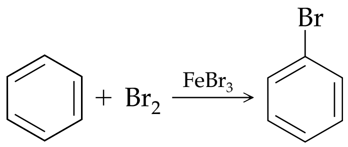

Suggest an appropriate mechanism for this reaction.

6-182. Draw the structure of the major monobrominated product(s) formed when each of the following compounds is reacted with Br_2 in the presence of $FeBr_3$:

(a)

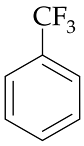

(b)

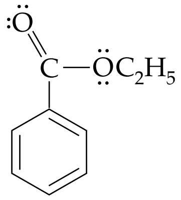

(c)

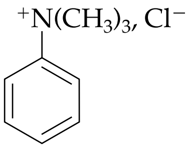

6-183. Arrange the following compounds in order of increasing reactivity toward electrophilic aromatic substitution. Explain the reason(s) for your decisions.

(a)

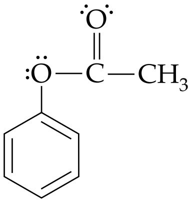

(b)

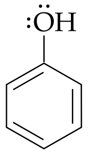

(c)

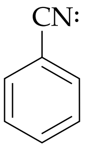

6-184. In the experiment, sodium bisulfite solution is added at the end to destroy the unreacted bromine. What reaction is occurring here? Is HSO_3^- acting as an oxidizing or reducing agent? Write a balanced equation as part of your answer.

6-185. Both the ^1H and ^{13}C NMR spectra (Figs. 6.46 and 6.47) provide unambiguous evidence that the bromination of acetanilide gave exclusively para substitution. Explain.

BIBLIOGRAPHY

For a general review of bromination see

Smith, M. B.; March. J. *Advanced Organic Chemistry,* 6th ed.; Wiley-Interscience: New York, 2007, Chap. 11, p. 698.

For use of an alternate reagent for the bromination of acetanilide see

Schatz, P. F. *J. Chem. Educ.* **1996,** *73,* 267.

The following references are selected from a large number of examples given in *Organic Syntheses* that illustrate electrophilic aromatic substitution using bromine:

Adams, R.; Marvel, C. S. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 128.

Coleman, C. H.; Talbot, W. F. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 592.

Hartman, W. W.; Dickey, J. B. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 173.

Johnson, J. R.; Sandborn, L. T. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 111.

Robison, M. M.; Robison, B. L. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 947.

Sandin, R. B.; McKee, R. A. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 100.

Smith, L. I. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 95.

Nitration: 2,5-Dichloronitrobenzene; N,N′-Diacetyl-2,3-dinitro-1, 4-phenylenediamine; 5-Nitrosalicylic Acid; and 2- and 4-Nitrophenol

EXPERIMENT 29

Common name: 2,5-dichloronitrobenzene
CA number: [89-61-2]
CA name as indexed: benzene, 1,4-dichloro-2-nitro-

Common name: *N,N*′-diacetyl-2,3-dinitro-1,4-phenylenediamine
CA number: [7756-00-5]
CA name as indexed: acetamide, *N,N*′-(2,3-dinitro-1,4-phenylene)bis-

Common names: 5-nitrosalicylic acid, anilotic acid
CA number: [96-97-9]
CA name as indexed: benzoic acid, 2-hydroxy-5-nitro-

Common names: 2-nitrophenol, *o*-nitrophenol
CA number: [88-75-5]
CA name as indexed: phenol, 2-nitro-

Common names: 4-nitrophenol, *p*-nitrophenol
CA number: [100-02-7]
CA name as indexed: phenol, 4-nitro-

Purpose. Aromatic nitration is an important synthetic reaction. This experiment explores two methods used for placing a nitro group on an aromatic ring system via an electrophilic aromatic substitution reaction. In Experiments [29A], [29B], and [29C] anhydrous nitric acid is used as the nitrating agent. In Experiment [29D], nitration is accomplished using a $SiO_2 \cdot HNO_3$ reagent.

Prior Reading

- *Technique 2:* Simple Distillation at the Semimicroscale Level (pp. 61–64)
- *Technique 5:* Crystallization
  - Use of the Hirsch Funnel (pp. 88–89)
  - Craig Tube Crystallization (pp. 89–91)
- *Technique 6:* Chromatography
  - Column Chromatography (pp. 92–95)
  - Thin-Layer Chromatography (pp. 97–99)
  - Concentration of Solutions (pp. 101–104)
- *Chapter 4:* Mixture Melting Points (pp. 52–54)

GENERAL REACTION

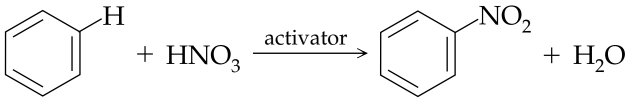

DISCUSSION

The nitration reactions described in this experiment all demonstrate one of the classic electrophilic aromatic substitution reactions. Nitration has been used extensively in organic synthesis since a nitro group on an aromatic ring may be readily reduced to an amino group.

Once introduced onto the aromatic ring, the electron-withdrawing nitro group deactivates the ring toward further reactions with electrophiles. For example, bromination of nitrobenzene leads only to *m*-nitrobromobenzene; no dibromonitrobenzene is readily formed. However, when activating groups (π-electron donors) are present on the ring, it is possible to nitrate the ring twice. This phenomenon can be illustrated by comparing the results of the nitration of 1,4-dichlorobenzene (Experiment [29A]) with that of *N,N′*-diacetyl-1,4-phenylenediamine (Experiment [29B]). Because of the presence of the activating acetamido (CH_3CONH—) groups, the dinitro derivative forms readily.

In Experiment [29C] (the preparation of 5-nitrosalicylic acid), the directing influences of the 1-CO_2H and 2-OH substituents on the entering —NO_2 group are illustrated. In this example, these two groups compliment each other since they both direct the entering nitro group to the 5 position. The 5 position and the 3 position are both electronically favored since the —CO_2H group is meta directing; the —OH group is ortho–para directing. The nitro group ends up at the 5 position, and not at the 3 position, due to steric effects.

The use of a silica gel-based reagent to accomplish nitration under fairly mild conditions is illustrated in Experiment [29D]. The nitrating reagent, $SiO_2 \cdot HNO_3$ is prepared by treatment of silica gel with nitric acid. In the experiment, phenol is nitrated to produce a mixture of products. *Thin-layer chromatography* is used to analyze the mixture, and the ortho and para nitrated phenols are separated by *column chromatography* using a silica gel column. If unreacted phenol is detected in the TLC analysis, an extraction technique is used to separate it from the para isomer. This separation technique is based on the fact that a nitrated phenol is more acidic than phenol itself.

It is generally accepted that the nitronium ion (NO_2^+) is the electrophile that adds to the aromatic ring. The overall mechanism for nitration follows:

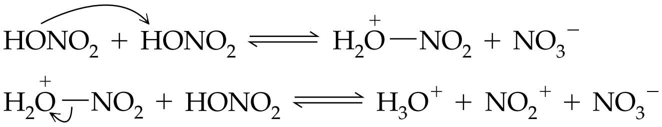

$$HONO_2 + HONO_2 \rightleftharpoons H_2\overset{+}{O}{-}NO_2 + NO_3^-$$
$$H_2\overset{+}{O}{-}NO_2 + HONO_2 \rightleftharpoons H_3O^+ + NO_2^+ + NO_3^-$$

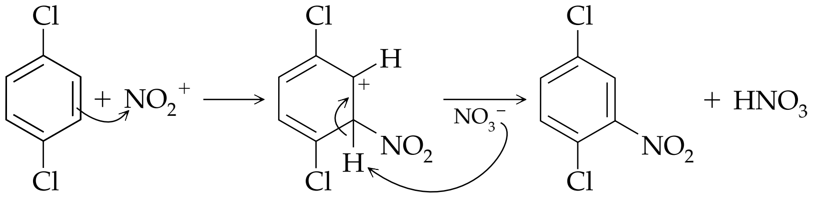

This mechanism illustrates two HNO_3 molecules reacting to generate the nitronium ion as when using the anhydrous nitric acid reagent. Sulfuric acid is often used to enhance the production of NO_2^+ as shown here:

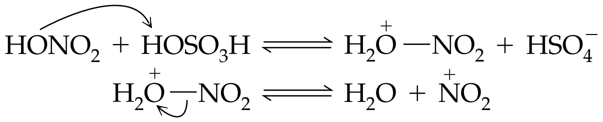

$$HONO_2 + HOSO_3H \rightleftharpoons H_2\overset{+}{O}{-}NO_2 + HSO_4^-$$
$$H_2\overset{+}{O}{-}NO_2 \rightleftharpoons H_2O + \overset{+}{N}O_2$$

Thus a commonly used nitrating reagent is a mixture of concentrated sulfuric and nitric acids.

SEMIMICROSCALE PREPARATION OF ANHYDROUS NITRIC ACID

Anhydrous nitric acid (HNO_3) is prepared by the following procedure.

CAUTION: The reagents and the product of this preparation are highly corrosive. The distillation must be conducted in a **hood.** Appropriate gloves are strongly suggested. Prevent contact with eyes, skin, and clothing. Any spill should be neutralized using solid sodium carbonate or bicarbonate.

HOOD

EXPERIMENTAL PROCEDURE

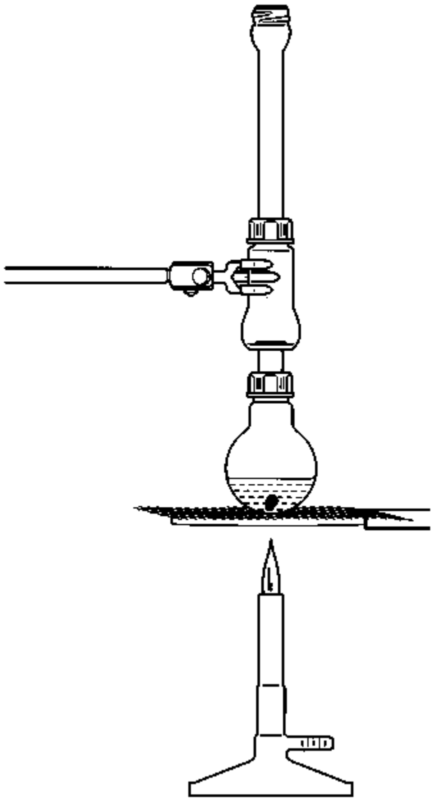

Concd HNO_3, 0.7 mL
+ concd H_2SO_4, 1.0 mL

Estimated time of preparation: 0.5 h.

NOTE. Use this anhydrous nitric acid immediately for the nitration experiments given below. The amount obtained at the scale used here is sufficient for the preparation of two of the nitro compounds described in this experiment.

Physical Properties of Reactants and Product

Compound	MW	Amount	bp (°C)	*d*
Concd nitric acid (68%)		0.7 mL	120.5	1.41
Concd sulfuric acid (96–98%)		1.0 mL	338	1.84
Anhydrous nitric acid	63.01		83	1.40

Reagents and Equipment. Using two clean, dry 1.0-mL graduated pipets, add 0.7 mL of concentrated nitric acid, followed by 1.0 mL of concentrated sulfuric acid, to a 10-mL round-bottom flask containing a boiling stone. Swirl the flask gently to mix the reagents. Attach the flask to a Hickman still fitted with an air condenser (⬅).

NOTE: It is useful to invert a 10-mL beaker over the air condenser to help contain the acid vapors.

CAUTION: Sulfuric acid can cause severe burns. Nitric acid is a strong oxidizing agent. Prevent contact with eyes, skin, and clothing. A spill can be neutralized using sodium carbonate or bicarbonate.

Reaction Conditions. Heat the acid solution very gently with a microburner, keeping the microburner in constant motion, until approximately 0.2 mL of anhydrous nitric acid has been collected as distillate in the collar of the still.

Purification and Characterization. Use the anhydrous nitric acid as collected. No further purification is required.

NOTE. Anhydrous nitric acid (white fuming nitric acid) is a colorless liquid, bp 83°C It is estimated that the nitric acid obtained in this preparation is at least 99.5–100% pure. If it is necessary to store the distillate, remove the acid from the collar of the still (Pasteur pipet) and place it in a 1.0-mL conical vial fitted with a glass stopper. It may be necessary to slightly bend the end of the pipet in a flame so that it can reach the collar of a still that does not have a side port. The anhydrous nitric acid is colorless or faintly yellow.

Experiment 29A

2,5-Dichloronitrobenzene

REACTION

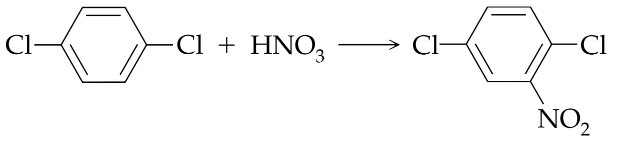

1,4-Dichlorobenzene 2,5-Dichloronitrobenzene

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 0.5 h.

Physical Properties of Reactants						
Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
1,4-Dichlorobenzene	147.01	38 mg	0.26	53		
Anhydrous nitric acid	63.01	100 μL	2.4		83	1.50

Reagents and Equipment. Equip a 3.0-mL conical vial with an air condenser (➨). Weigh and add 38 mg (0.26 mmol) of 1,4-dichlorobenzene, followed by 100 μL of anhydrous nitric acid delivered from a calibrated Pasteur pipet (9 in.).

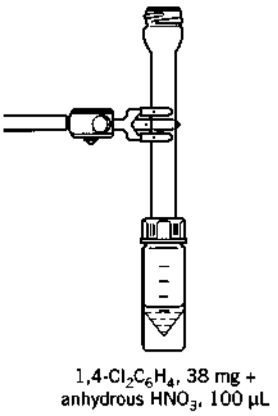

> CAUTION: The nitric acid reagent is highly corrosive. Prevent contact with eyes, skin, and clothing. A spill is neutralized using solid sodium carbonate or bicarbonate.

Reaction Conditions. Allow the resulting solution to stand at room temperature for a period of 15 min. Next add 1.0 mL of water (calibrated Pasteur pipet) dropwise, while stirring with a thin glass rod, and then place the vial in an ice bath to cool.

Isolation of Product. Collect the crystalline precipitate by vacuum filtration using a Hirsch funnel (➨). Wash the filter cake with four 1.0-mL portions of water (calibrated Pasteur pipet) and then place it on a porous clay plate or on filter paper to dry.

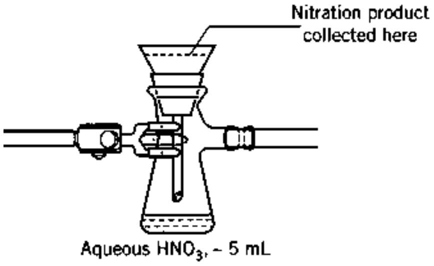

Purification and Characterization. The product, consisting of fine, white needles, is sufficiently pure for characterization. It may be recrystallized from ethanol–water, using a Craig tube, if desired.

Weigh the 2,5-dichloronitrobenzene and calculate the percent yield. Determine the melting point and compare your result to that reported in the literature. Notice that the starting material and the nitrated product have very close melting points. It is recommended that a mixed melting point be carried out to establish that the desired product has been isolated (see Prior Reading).

A further or alternative check on the purity of the material may be made using thin-layer chromatography (TLC).

INFORMATION. Carry out the TLC analysis with Eastman Kodak silica gel–polyethylene terephthalate plates (#13179). Activate the plates at an oven temperature of 100 °C for 30 min. Place them in a desiccator for cooling and storing until used. Elute the plates using hexane solvent. Visualization is accomplished with UV light. See Prior Reading for the methods of TLC analysis and determination of R_f *values.*

Chemical Tests. Additional chemical tests (Chapter 9) may also be performed to further characterize the product. Does the ignition test confirm the presence of the aromatic ring? Does the Beilstein test detect the presence of chlorine? Can the sodium fusion test detect the presence of nitrogen? Can the specific presence of the nitro group be detected by reaction with ferrous hydroxide solution?

Experiment 29B

N,N′-Diacetyl-2,3-dinitro-1,4-phenylenediamine

REACTION

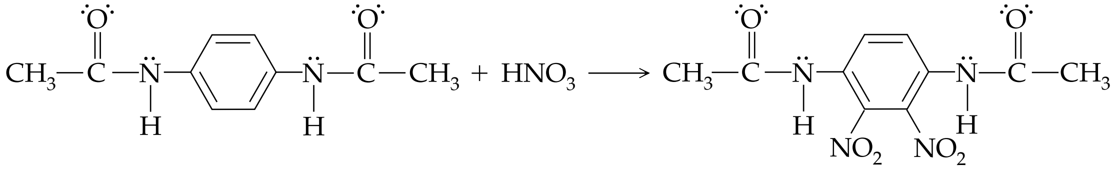

N,N′-Diacetyl-1,4-phenylenediamine

N,N′-Diacetyl-2,3-dinitro-1,4-phenylenediamine

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 0.5 h.

Physical Properties of Reactants and Product						
Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
N,N′-Diacetyl-1,4-phenylenediamine	192	48 mg	0.25	312–315		
Anhydrous nitric acid	63.01	100 µL	2.4		83	1.50
N,N′-Diacetyl-2,3-dinitro-1,4-phenylenediamine	282			257		

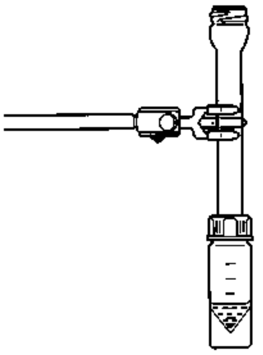

CAUTION

Reagents and Equipment. To a 3.0-mL conical vial equipped with an air condenser, weigh and add 48 mg (0.25 mmol) of *N,N′*-diacetyl-1,4-phenylenediamine. Now, using **caution,** add dropwise 100 µL of anhydrous nitric acid delivered from a calibrated Pasteur pipet (9 in.) (⬅). *The N,N′-diacetyl-1,4-phenylenediamine is prepared by the procedure outlined in Experiment [23B].*

EXOTHERMIC

CAUTION: The reaction is highly ***exothermic.*** A vigorous reaction occurs if the acid is added too rapidly. The nitric acid reagent is highly corrosive; prevent contact with eyes, skin, and clothing. A spill is neutralized using sodium carbonate or bicarbonate.

Reaction Conditions. Allow the resulting solution to stand at room temperature for a period of 10 min. Add 1.0 mL of water from a calibrated Pasteur pipet dropwise and then place the vial in an ice bath to cool.

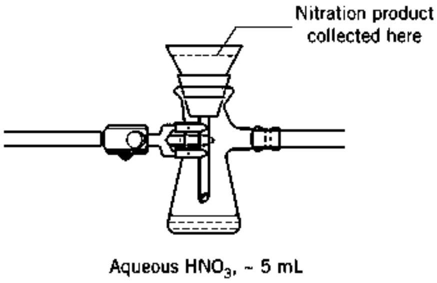

Isolation of Product. Collect the resulting yellow precipitate by vacuum filtration using a Hirsch funnel (⬅). Wash the filter cake with four 1.0-mL portions of water from a calibrated Pasteur pipet. Maintain suction to aid in drying of the product. A piece of plastic food wrap over the mouth of the funnel can aid this process (see Prior Reading). Dry the material further on a porous clay plate, on filter paper, or under vacuum at room temperature.

Purification and Characterization. The product needs no further purification. Weigh the dried material and calculate the percent yield.

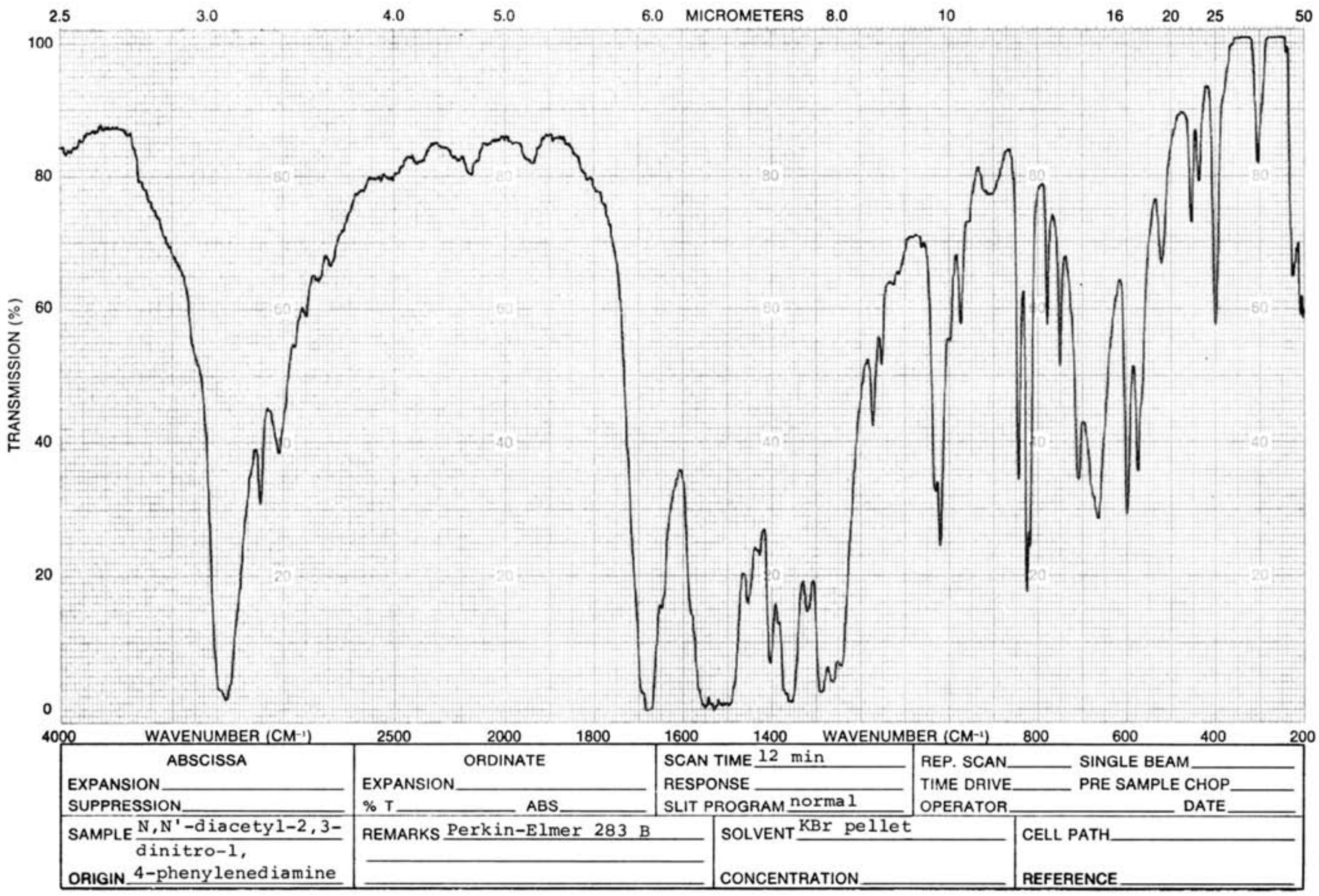

ABSCISSA	ORDINATE	SCAN TIME 12 min	REP. SCAN ___ SINGLE BEAM ___
EXPANSION ___	EXPANSION ___	RESPONSE ___	TIME DRIVE ___ PRE SAMPLE CHOP ___
SUPPRESSION ___	% T ___ ABS ___	SLIT PROGRAM normal	OPERATOR ___ DATE ___
SAMPLE N,N'-diacetyl-2,3-dinitro-1,	REMARKS Perkin-Elmer 283 B	SOLVENT KBr pellet	CELL PATH ___
ORIGIN 4-phenylenediamine		CONCENTRATION ___	REFERENCE ___

Figure 6.48 IR spectrum: *N,N′*-diacetyl-2,3-dinitro-1,4-phenylenediamine.

Determine the melting point and compare your result with the melting point listed in the above Reactant Product table. Obtain an IR spectrum of your product and compare it with that shown in Figure 6.48.

5-Nitrosalicylic Acid

Experiment 29C

This material may be used to prepare the caffeine 5-nitrosalicylate derivative (see Experiment [11B]).

REACTION

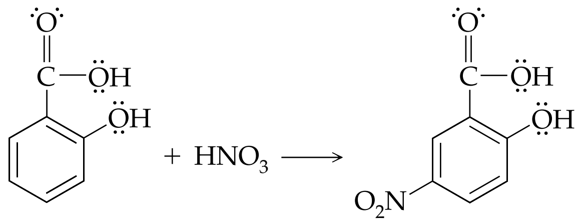

Salicylic acid

5-Nitrosalicylic acid

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 0.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Salicylic acid	138.12	50 mg	0.36	159		
Anhydrous nitric acid	63.01	100 μL	2.4		83	1.50

HOOD

CAUTION: This reaction should be conducted in a **hood.**

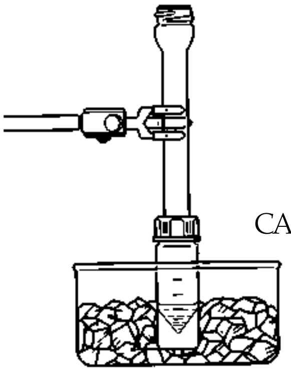

Salicylic acid, 50 mg + anhydrous HNO_3, 100 μL

CAUTION

Reagents and Equipment. Weigh and add 50 mg (0.36 mmol) of salicylic acid to a 3.0-mL conical vial equipped with an air condenser. Place the vial in an ice bath to cool. Also place a stoppered conical vial containing 100 μL of freshly prepared **ANHYDROUS** nitric acid (use **caution**) from a Pasteur pipet in the ice bath (⬅).

CAUTION: The nitric acid reagent is highly corrosive; prevent contact with eyes, skin, and clothing. A spill is neutralized using solid sodium carbonate or bicarbonate.

TOXIC

Reaction Conditions. **Carefully** add the cold nitric acid dropwise from a calibrated Pasteur pipet (9 in.) to the salicylic acid. Keep the vial containing the salicylic acid in the ice bath during the addition. The evolution of a red-brown gas (NO_2, which is **toxic**) is observed during the addition, if the acid is not pure.

EXOTHERMIC

CAUTION: The reaction is highly **exothermic.** A very vigorous reaction occurs if the acid is added too rapidly.

Allow the vial to stand in the ice bath for an additional 20 min and then add 1.0 mL of distilled water dropwise from a calibrated Pasteur pipet to the reaction mixture.

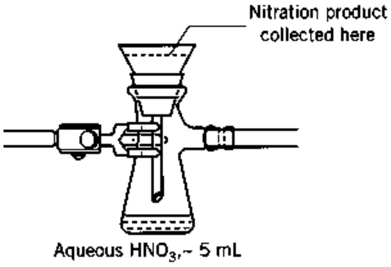

Isolation of Product. Collect the orange-pink solid by vacuum filtration using a Hirsch funnel (⬅). Wash the filter cake with four 1.0-mL portions of cold water from a calibrated Pasteur pipet. Maintain suction to aid in drying of the product. A piece of plastic food wrap over the mouth of the funnel can aid this process (see Prior Reading). Dry the material further on a porous clay plate, on filter paper, or under vacuum at room temperature.

Purification and Characterization. Recrystallize the product using a Craig tube by dissolving the material in the *minimum* amount of absolute ethanol, followed by the dropwise addition of water until precipitation occurs. Cool the mixture in an ice bath and separate the light-yellow crystals. Dry them on a porous clay plate or under vacuum as noted above.

Weigh the crystals and calculate the percent yield. Determine the melting point and compare your value to that listed in the literature. Obtain an IR spectrum of the material and compare it with that recorded in the literature.

2- and 4-Nitrophenol

Experiment 29D

INSTRUCTOR PREPARATION. In a 250-mL Erlenmeyer flask containing a magnetic stirring bar, weigh and place 20.0 g of silica gel (70–230 mesh; the removal of fines is not necessary). Now add 50 mL of 7.5 M nitric acid and stir the mixture for 3 h at room temperature. Remove the nitrated silica gel by gravity filtration (do not rinse), place it on a clay plate, and allow it to air dry in a ***hood*** *overnight. Store the product in an airtight container.*

HOOD

Determine the nitric acid content of the silica gel by titration of a water suspension of the gel with a 0.1 M NaOH solution. The acid content of the gel should be in the range of 16–20% by weight.

Nitration of Phenol

REACTION

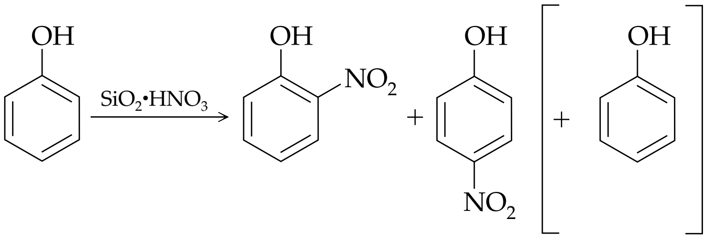

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants					
Compound	**MW**	**Amount**	**mmol**	**mp (°C)**	**bp (°C)**
Phenol	94.11	240 mg	2.55	40.5–41.5	182
$SiO_2 \cdot HNO_3$		1.0 g			
Methylene chloride		5 mL			40

Reagents and Equipment. Weigh and add 240 mg (2.55 mmol) of phenol to a 10-mL round-bottom flask containing a stir bar. Now add 5 mL of methylene chloride. To this solution, weigh and add 1.0 g of nitrated silica gel (~16% HNO_3). Attach the flask to an air condenser (➨).

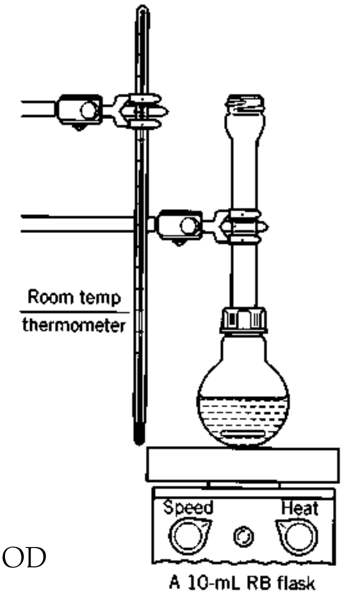

CAUTION: Phenol is highly toxic and corrosive. Prevent contact with eyes and skin. It is best dispensed by warming the container of phenol in a warm water bath and then using an automatic delivery pipet. This should be done in the **hood.**

HOOD

Reaction Conditions. Stir the resulting mixture at room temperature for 5 min.

Isolation of Product. Separate the silica gel from the reaction mixture by gravity filtration through a filter funnel containing a small plug of glass wool. Collect the filtrate in a 10-mL Erlenmeyer flask containing a boiling stone. Wash the collected silica gel with two 0.5-mL portions of methylene chloride, and collect these washings in the same Erlenmeyer flask.

HOOD

Concentrate the filtrate to a volume of about 1.0 mL using a warm sand bath under a slow stream of nitrogen in the **hood.**

Use thin-layer chromatography to obtain an analysis of the product mixture (see Prior Reading). Use methylene chloride as the elution solvent, silica gel (with a fluorescent indicator) as the stationary phase, and UV light for visualization. Typical R_f values are 0.04 for 4-nitrophenol, 0.15 for phenol, and 0.58 for 2-nitrophenol.

NOTE. 2,4-Dinitrophenol has a typical R_f *value of 0.33 under the chromatography conditions above; it is not usually formed under these reaction conditions.*

Characterization and Purification. Column chromatography is now used to separate the mixture of products. Pack (dry) a 1.0-cm-diameter buret column with 7.5 g of activated silica gel. Place the above product solution on the column using a Pasteur pipet and then elute the column with 25 mL of 60:40 methylene chloride/pentane solvent. Collect the first 20 mL of eluate in a tared 25-mL Erlenmeyer flask. Concentrate this fraction to dryness in a warm sand bath under a slow steam of nitrogen to isolate the 2-nitrophenol. Weigh the product.

Now elute the column with about 30 mL of 1:1 ethyl acetate/methylene chloride solvent. Collect the first 20 mL of eluate in a tared 25-mL Erlenmeyer flask containing a boiling stone. Concentrate this fraction as above to a volume of about 2 mL. Use thin-layer chromatography to determine the purity of the product (see conditions outlined above). The main constituent of this fraction is 4-nitrophenol. However, the presence of unreacted phenol or possibly quinone is often detected. Now concentrate the product solution to dryness and weigh the 4-nitrophenol isolated.

If the TLC analysis indicates impurity, 4-nitrophenol may be purified as follows. Dissolve the solid residue in 7 mL of saturated sodium bicarbonate solution and transfer the resulting solution to a 15-mL centrifuge tube. Extract this solution twice with 2-mL portions of methylene chloride (a Vortex mixer may be used to good advantage here). Remove the methylene chloride layers using a Pasteur pipet and save them (together) in a small Erlenmeyer flask until you have isolated and characterized the product.

Cool the resulting aqueous solution in an ice bath and add 6 M HCl dropwise, with mixing (glass rod or Vortex mixer), until it becomes neutral or slightly acidic toward litmus or pH paper. *Do this step carefully. Too vigorous a reaction may result in loss of product.* Now extract the resulting solution with three 2-mL portions of methylene chloride. Following each extraction, remove the methylene chloride using a Pasteur filter pipet and transfer it to a 10-mL Erlenmeyer flask. Dry the wet solution over anhydrous sodium sulfate and then transfer it by Pasteur filter pipet to a tared 10-mL Erlenmeyer flask containing a boiling stone. Concentrate this solution using a warm sand bath under a slow stream of nitrogen gas to yield the 4-nitrophenol product. Weigh the product. The purity of the material may be checked again, using

TLC as outlined above. It may also be recrystallized from water using the Craig tube, if necessary.

To characterize the 2- and 4-nitrophenols, determine their melting points and obtain their infrared spectra. Compare your results to those reported in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Chemical Tests. Chemical classification tests (see Chapter 9) are also of value to establish the identity of these compounds. Perform the ferric chloride test for phenols. Is a positive result obtained for each compound? Does the sodium fusion test detect the presence of nitrogen? The test for nitro groups might also be performed. The phenyl- or α-naphthylurethane derivatives of the phenols may be prepared to further establish their identity (see Chapter 9, Preparation of Derivatives).

QUESTIONS

6-186. Predict the most likely mononitration product from each of the following compounds. Explain the reasons for your choice.

(a)

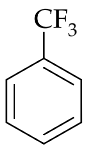

(b)

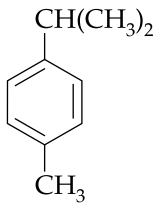

(c)

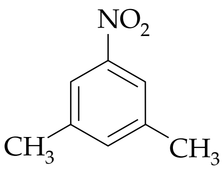

(d)

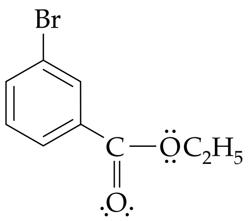

(e)

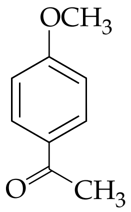

(f)

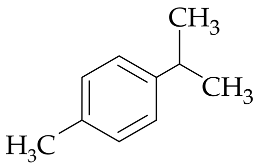

(g)

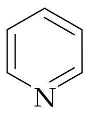

(h)

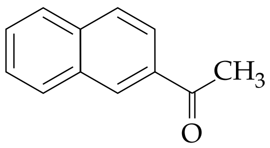

6-187. Write equations to show how nitronium ions might be formed using a mixture of nitric and sulfuric acids.

6-188. Which ring of phenyl benzoate would you expect to undergo nitration more readily? Explain.

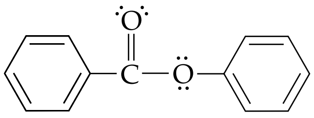

Phenyl benzoate

6-189. Arrange the following compounds in order of increasing reactivity toward nitration. Explain.
- **(a)** Acetanilide
- **(b)** Acetophenone
- **(c)** Bromobenzene
- **(d)** Toluene

6-190. Offer a reasonable explanation of why nitration of 1,4-dichlorobenzene yields the mononitro derivative while *N,N′*-diacetyl-1,4-phenylenediamine forms the dinitro compound.

6-191. Explain why *p*-nitrophenol is a stronger acid than phenol itself. Would *p*-methoxyphenol be a stronger or weaker acid than phenol? Explain.

BIBLIOGRAPHY

Related to anhydrous nitric acid:

Stern, S. A.; Mullhaupt, J. T.; Kay, W. B. *Chem. Rev.* **1960,** *60,* 185.

The preparation used in this experiment is an adaptation of that reported by

Cheronis, N. D.; Entrikin, J. B. *Semimicro Qualitative Organic Analysis;* Crowell: New York, 1947, p. 258.

The preparation of the nitrated silica gel was adapted from that reported by

Tapia, R.; Torres, G.; Valderrama, J. A. *Synth. Commun.* **1986,** *16,* 681.

For a general review of nitration of aromatic species see

Olah, G. A.; Malhotra, R.; Narang, S. C. *Nitration: Methods and Mechanism;* VCH: New York, 1989.

Schefield, K. *Aromatic Nitration*; Cambridge Univ. Press: New York, 1980.

Smith, M. B.; March. *J. Advanced Organic Chemistry,* 6th ed.; Wiley-Interscience: New York, 2007, Chap. 11, p. 685.

These references are selected from a large number of examples of nitration given in *Organic Syntheses*. None of these examples uses anhydrous nitric acid or the nitrated silica gel as the nitrating agent.

Braum, C. E.; Cook, C. D. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 711.

Fanta, P. E.; Tarbell, D. S. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 661.

Fetscher, C. A. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 735.

Fitch, H. M. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 658.

Howard, J. C. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 42.

Huntress, E. H.; Shriner, R. L. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 459.

Mendenhall, G. D.; Smith, P. A. S. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 829.

Newman, M. S.; Boden, H. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 1029.

EXPERIMENT 30

Nucleophilic Aromatic Substitution: 2,4-Dinitrophenylthiocyanate

Common name: 2,4-dinitrophenylthiocyanate
CA number: [1594-56-5]
CA name as indexed: thiocyanic acid, 2,4-dinitrophenyl ester

Purpose. The conditions under which a specific type of **nucleophilic aromatic substitution reaction** can occur are investigated. You will carry out the conversion in a two-phase solvent system through the use of a phase-transfer catalyst. The experiment is an example of nucleophilic aromatic substitution on an aromatic ring having two strongly electron-withdrawing substituents. In this experiment, a bromine substituent (leaving group) is replaced by a thiocyanate, $-S-C{\equiv}N$, group (nucleophile).

Prior Reading

- *Technique 4:* Solvent Extraction
  - Liquid–Liquid Extraction (p. 72)
  - Drying of the Wet Organic Layer (pp. 80–83)
- *Technique 5:* Crystallization
  - Craig Tube Crystallization (pp. 89–91)
- *Technique 6:* Chromatography
  - Column Chromatography (pp. 92–95)
  - Concentration of Solutions (pp. 101–104)

REACTION

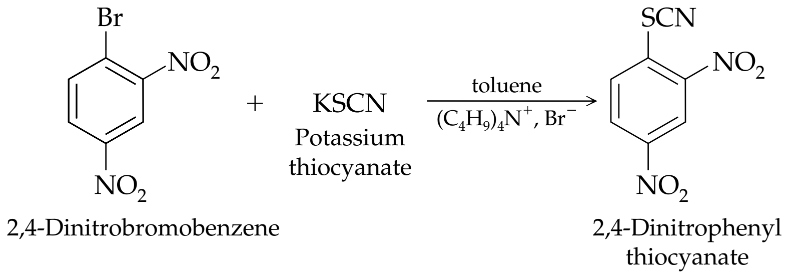

DISCUSSION

Nucleophilic aromatic substitution reactions generally take place only if activating groups are present at the positions ortho and/or para to a good leaving group. Activating groups, in this case, are groups that are electron withdrawing and render the aromatic ring more electron-deficient, and thus more susceptible to nucleophilic attack. On the other hand, π-electron donating groups hinder the reaction. Groups such as $-NO_2$, $-SO_2CH_3$, $-{}^+N(CH_3)_3$, $-CF_3$, $-CN$, $-SO_3^-$, $-Br$, $-Cl$, $-I$, $-CO_2^-$, especially if one $-NO_2$ group is already present, are strongly activating. These groups not only make the ring more susceptible to nucleophilic attack, they assist in the resonance stabilization of the intermediate cyclohexadienyl anion. In the reaction illustrated in this experiment, two nitro groups are present on the ring, both ortho and para to the departing bromine substituent. While functional groups containing a carbonyl group do, in principle, activate the aromatic ring toward nucleophilic attack, the carbonyl groups themselves are usually more likely to react with a nucleophile than is the aromatic ring.

The conditions under which this reaction is conducted lend themselves to the use of phase-transfer catalysis. The system involves two phases: the aqueous phase and the organic phase (toluene). The phase-transfer catalyst plays a very important role. As an ammonium cation, it carries the SCN^- ion from the aqueous phase into the organic phase where the substitution reaction then occurs. The product and the corresponding bromide salt of the phase-transfer catalyst are produced in this conversion. The bromide salt of the phase-transfer catalyst then migrates back into the aqueous phase, and the process repeats itself. The catalyst can play this role since the large organic groups (the four butyl groups) increase the solubility of the phase-transfer catalyst in the organic phase, while the charged ionic center of the catalyst renders it water soluble. Phase-transfer catalysts can also be used in many other reactions, including the preparation of ethers by the Williamson method (see Experiments [22A] and [22B]), and the Horner–Wadsworth–Emmons modification of the Wittig reaction (see Experiments [19B] and [19D]).

The reaction in this experiment proceeds by an *addition–elimination two-step* sequence, of which the first step is generally rate determining. A tetrahedral intermediate (cyclohexadienyl anion) is formed by the attack of the nucleophile on the carbon atom to which the leaving group is attached. In the subsequent step, the leaving group then departs with its bonding electrons, regenerating the aromatic system. It is important to note that, in contrast to nucleophilic substitution reactions on alkyl carbon atoms, the reaction does not proceed by an S_N2 mechanism, as in this case an intermediate is formed. The mechanism is outlined on the next page.

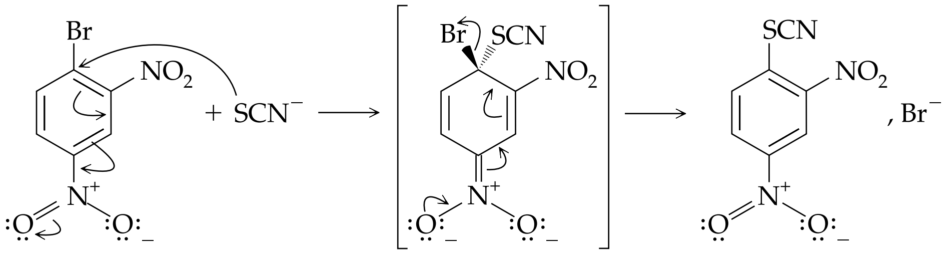

EXPERIMENTAL PROCEDURE

Estimated time to complete the reaction: 2.0 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
2,4-Dinitrobromobenzene	247.01	50 mg	0.20	75	
Tetrabutylammonium bromide	322.38	5 mg	0.02	103–104	
Aqueous potassium thiocyanate (50%)		150 μL			
Toluene	92.15	350 μL			111
2,4-Dinitrophenyl-thiocyanate	244			138–139	

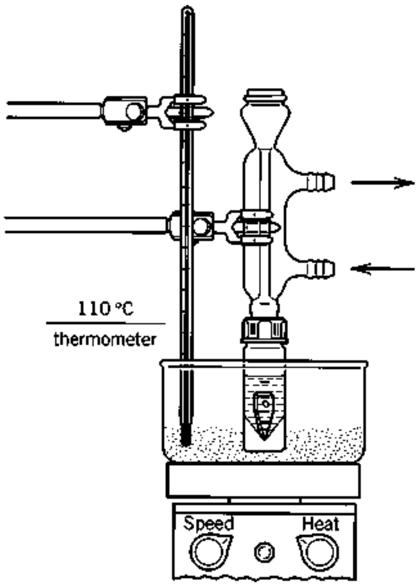

2,4-$(NO_2)_2C_6H_3Br$, 50 mg
+ $C_6H_5CH_3$, 350 μL +
$Bu_4N^+Br^-$, 5 mg +
50% aqueous KSCN, 150 μL

Reagents and Equipment. Assemble a 3.0-mL conical vial containing a magnetic spin vane and equipped with a reflux condenser. Weigh and place in the vial 50 mg (0.20 mmol) of 2,4-dinitrobromobenzene followed by 350 μL of toluene (⬅). To this solution add 5.0 mg (0.02 mmol) of tetrabutylammonium bromide and 150 μL of a 50% aqueous potassium thiocyanate (wt/wt) solution.

HOOD

> CAUTION: Tetrabutylammonium bromide is an irritant. ***Handle with care!*** It is also hygroscopic and should be protected from moisture prior to its use. Dispense the toluene and the thiocyanate solution in the **hood** using automatic delivery pipets.

Reaction Conditions. Heat the resulting mixture at a sand bath temperature of 100 °C with stirring for 1 h. Allow the solution to cool to room temperature.

Isolation of Product. Use a Pasteur filter pipe to separate the toluene layer from the aqueous layer (**saving** both layers) and place the toluene layer in a 3.0-mL capped vial. Extract the toluene layer with two 1.0-mL portions of water and combine the water extracts with the original water layer you saved. Now extract the combined aqueous layers with two 0.5-mL portions of toluene, and combine these toluene extracts with the original toluene layer.

NOTE. The volumes of liquid used in the extraction are measured using calibrated Pasteur pipets. For each extraction, shake the vial, vent it carefully, and allow the layers to separate. The transfers are made using Pasteur filter pipets.

Transfer the toluene solution by Pasteur pipet to a Pasteur filter pipet microcolumn containing 700 mg of anhydrous sodium sulfate previously wetted with toluene (➡). Allow the solution to elute through the column, and collect the dried toluene eluate in a 10-mL Erlenmeyer flask containing a boiling stone. Concentrate this solution by warming in a sand bath under a gentle stream of nitrogen gas in the **hood.** A yellow, solid product is obtained.

HOOD

HOOD

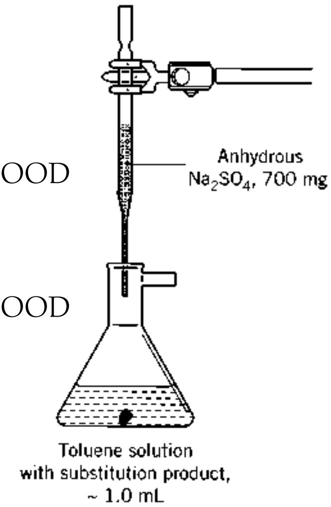
Anhydrous Na_2SO_4, 700 mg
Toluene solution with substitution product, ~ 1.0 mL

Purification and Characterization. The product is generally nearly pure. Re-crystallize the material from chloroform, using a Craig tube, if desired.

CAUTION: Chloroform is toxic! Dispense and use only in the **hood.** Do not breathe the vapors.

Weigh the dried product and calculate the percent yield. Determine the melting point and compare it with that listed in the Reagent and Product table. Obtain an IR spectrum and compare your result to that of an authentic sample, or to one in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar)

2,4-Dinitrophenylthiocyanate: Preparation Using a Monomode Microwave Apparatus

Experiment 30-1*

M

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1 h.

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
1-Bromo-2,4-dinitrobenzene	247.01	67 mg	0.27	71–73	
Aqueous potassium thiocyanate (50%)		0.2 mL			
Ethanol	276	0.8 mL			78
2,4-Dinitrophenylthiocyanate	244			138–139	

Reagents and Equipment. This experiment is designed for use in the CEM Discover and Biotage Initator microwave units.

In a 10.0-mL glass microwave reaction vessel containing a magnetic stir bar, add 67 mg (0.27 mmol) of 2,4-dinitrobromobenzene, 0.2 mL of 50% aqueous potassium thiocyanate (wt/wt) solution, and 0.8 mL of ethanol. Immediately cap the vessel with the microwave pressure cap.

---

*This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

HOOD

CAUTION: Dispense the potassium thiocyanate solution in the **hood** using an automatic delivery pipet. Since the reaction requires heating the reaction mixture to above the boiling point of some components in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential*.

Reaction Conditions. Place the reaction vessel in the microwave cavity and, depending on the equipment used, position the pressure device on top. Program the microwave unit to heat the reaction mixture to 125 °C and hold at this temperature for 5 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the tube from the microwave unit. Allow the resulting solution to cool to room temperature and then place it in an ice bath for a 10-min period to complete the crystallization of the product.

Isolation of Product. Collect the yellow crystals by vacuum filtration using a Hirsch funnel and wash the filter cake with cold 95% ethanol (1–2 mL). Partially dry the filter cake under suction. Transfer the filtered, washed, and partially dried product to a porous clay plate or filter paper to complete drying.

Characterization. Weigh the dried product and calculate the percent yield. Determine the melting point and compare it with the list in the Reagent and Product table. Obtain an IR spectrum of your product, and compare it to that of an authentic sample, or to the spectrum shown in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Experiment 30-2*

M

2,4-Dinitrophenylthiocyanate: Preparation Using a Multimode Microwave Apparatus

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1 h.

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
1-Bromo-2,4-dinitrobenzene	247.01	296 mg	1.20	71–73	
Aqueous potassium thiocyanate (50%)		1.0 mL			
Ethanol	276	4 mL			78
2,4-Dinitrophenylthiocyanate	244			138–139	

Reagents and Equipment. This experiment is designed for use in the CEM MARS, Milestone START, and Anton Paar Synthos 3000 microwave units. When using the Anton Paar Synthos 3000 unit with the 24-position silicon carbide plate rotor containing glass vials, the reagent and solvent quantities cited in the monomode procedure should be used in conjunction with the reaction conditions here in the multimode procedure.

In a microwave reaction vessel containing a magnetic stir bar, add 296 mg (1.2 mmol) of 2,4-dinitrobromobenzene, 1.0 mL of 50% aqueous potassium

[*] This section has been written by Dr. Nicholas E. Leadbeater from the Department of Chemistry at the University of Connecticut, and Dr. Cynthia B. McGowan from the Department of Chemistry at Merrimack College, MA.

thiocyanate (wt/wt) solution, and 4.0 mL of ethanol. Immediately cap the vessel with the microwave pressure cap and adjust the tightness to the manufacturer-specified level. Place the sealed vessel into its outer protective jacket.

CAUTION: Dispense the potassium thiocyanate solution in the **hood** using an automatic delivery pipet. Since the reaction requires heating the reaction mixture to above the boiling point of some components in sealed vessels, *adherence to the microwave manufacturer's guidelines is essential.*

HOOD

Reaction Conditions. Insert the loaded vessels into the reaction carousel ensuring they are evenly spaced and then place the carousel into the microwave cavity. If provided by the manufacturer, connect a temperature probe to the control vessel. Program the microwave unit to heat the reaction vessels to 125 °C and hold at this temperature for 5 min. After heating, allow the reaction mixture to cool to 50 °C or below before removing the carousel from the microwave unit. Allow the resulting solution to cool to room temperature and then place it in an ice bath for 10 min to complete the crystallization of the product.

Isolation of Product. Collect the yellow crystals by vacuum filtration using a Hirsch funnel and wash the filter cake with cold 95% ethanol (3–5 mL). Partially dry the filter cake under suction. Transfer the filtered, washed and partially dried product to a porous clay plate or filter paper to complete drying.

Characterization. Weigh the dried product and calculate the percent yield. Determine the melting point and compare it with the list in the Reagent and Product table. Obtain an IR spectrum of your product, and compare it to that of an authentic sample, or to the spectrum shown in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

QUESTIONS

6-192. Explain why the first reaction below requires substantially more vigorous reaction conditions than the second reaction:

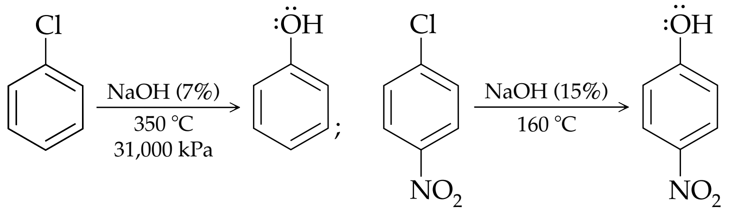

6-193. Explain why the intermediate cyclohexadienyl anion in the nucleophilic aromatic substitution reaction (see Discussion section) is not aromatic, even though it has the same number of π electrons (6) as the starting benzene derivative.

6-194. Complete each of the following reactions and name the expected product of each:

(a)

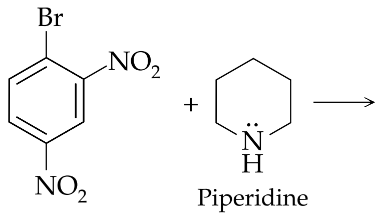

(b)

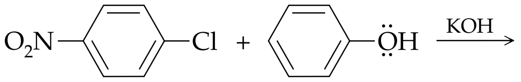

(c)

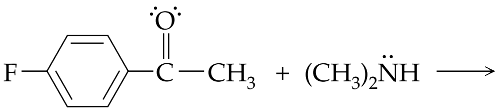

6-195. Compare the potential energy diagram of an S_N2 substitution reaction on an aliphatic halide to that of the nucleophilic substitution reaction carried out in this experiment. Discuss the main differences in the diagrams in terms of the mechanisms.

6-196. Would you expect the rate of the reaction in this experiment to depend on the concentrations of *both* the thiocyanate ion *and* 2,4-dinitrobromobenzene? Explain.

6-197. Show the intermediates, including all the resonance hybrid structures for the cyclohexadienyl anion, that are formed in the following aromatic nucleophilic substitution reactions:

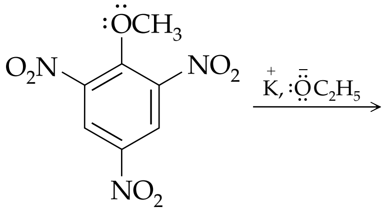

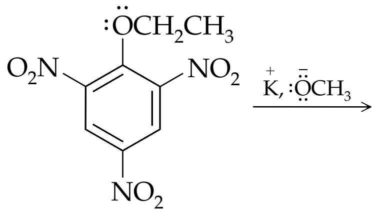

6-198. On workup of each of the reactions given in Question 6-197, what product(s) would you expect to form? If more than one, indicate their relative amounts.

BIBLIOGRAPHY

Selected examples of nucleophilic aromatic substitution reactions in *Organic Syntheses* include

Brewster, R. Q.; Groening, T. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 445.

Bunnett, J. F.; Conner, R. M. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 478.

Hartman, W. W. *Organic Syntheses;*; Wiley: New York, 1941; Collect. Vol. I, p. 175.

Hartman, W. W.; Byers, J. R.; Dickey, J. B. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 451.

Kharasch, N.; Langford, R. B. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 474.

Reverdin, F. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 219.

Sahyun, M. R. V.; Cram, D. J. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 926.

Skorcz, J. A.; Kuminski, F. E. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 263.

Smith M. B.; March. J. *Advanced Organic Chemistry*, 6th ed.; Wiley-Interscience: New York, 2007, Chap. 13, p. 870.

EXPERIMENT 31

Halogenation Using N*-Bromosuccinimide: 9-Bromoanthracene*

Common name: 9-bromoanthracene
CA number: [1564-64-3]
CA name as indexed: anthracene, 9-bromo-

Purpose. A free radical halogenation reaction is run using *N*-bromosuccinimide (NBS), a highly specific brominating agent. This material has the advantage that it is a solid and is therefore easier to handle than bromine, which is a toxic liquid.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallization (pp. 89–91)

REACTION

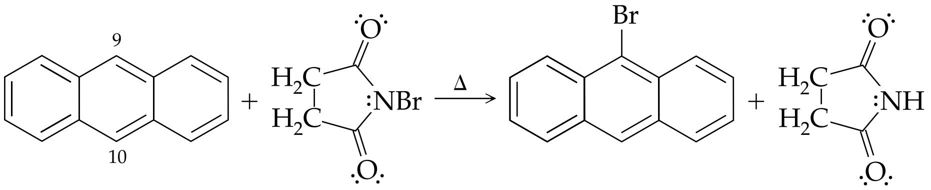

Anthracene *N*-Bromosuccinimide 9-Bromoanthracene Succinimide

DISCUSSION

N-Bromosuccinimide (NBS) is a highly specific brominating agent. When this reagent is used, anthracene is brominated in the 9-position. *N*-Bromosuccinimide may also be used to brominate positions α to (1) a carbonyl group, (2) a triple bond, (3) an alkene (allylic position), and (4) aromatic rings (benzylic position). Other polynuclear hydrocarbons that have been brominated using NBS include naphthalene, phenanthrene, and acenaphthene.

In the preparation of 9-bromoanthracene, the reaction progress is easily followed, since NBS (a reactant) and succinimide (a product) are both nearly insoluble in carbon tetrachloride. The NBS is more dense than the carbon tetrachloride solvent, and as the reaction proceeds this solid disappears from the bottom of the reaction flask and the less dense succinimide forms and floats on the surface of the reaction solution.

Free radicals have been identified in the mechanism of bromination using *N*-bromosuccinimide. In fact, the reaction proceeds only under conditions likely to produce radicals: photochemical conditions, by heating, or in the presence of a free radical initiator. The NBS reagent provides a steady source of small amounts of Br_2 via the rapid reaction of NBS with traces of hydrogen bromide.

The initiation step in bromination with NBS is the formation of a bromine radical by the homolytic dissociation of the Br_2 molecule.

Initiation Step

$$\mathrm{Br_2} \xrightarrow[\text{free radical initiators}]{h\nu \text{ or}} 2\ \mathrm{Br}\cdot$$

The bromine radical then abstracts a hydrogen atom from the 9-position of the anthracene.

Propagation Step

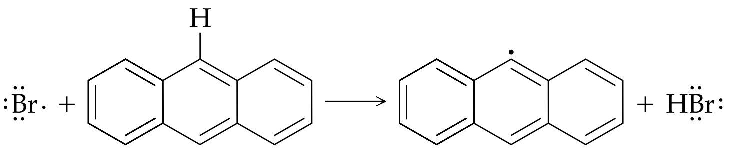

The HBr so formed then reacts with NBS to produce a bromine molecule and succinimide:

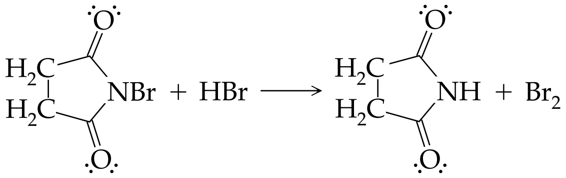

The Br_2 molecule then reacts with the anthracenyl radical formed above to yield the product and a bromine radical:

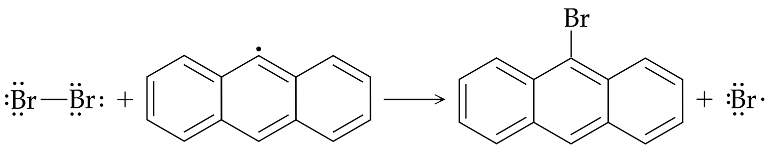

This bromine radical starts the sequence over again. That is, a chain reaction is initiated. In the present reaction, a trace amount of an iodine–carbon tetrachloride solution is added to the reaction mixture. The iodine acts as a moderator or a retarder in the reaction (see Dauben et al. in the Bibliography section), and thus only the monobrominated product is formed; 9,10-dibromoanthracene is not generated under these conditions.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
Anthracene	178.24	50 mg	0.28	216	
N-Bromosuccinimide	177.99	50 mg	0.28	180–183	
Carbon tetrachloride	153.82	0.4 mL			77
Iodine–CCl_4 solution		1 drop			

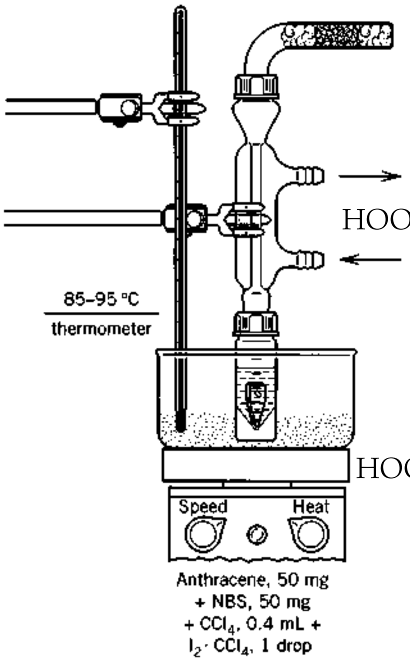

Reagents and Equipment. Weigh and add 50.0 mg (0.28 mmol) of anthracene and 50 mg (0.28 mmol) of *N*-bromosuccinimide to a 3.0-mL conical vial containing a magnetic spin vane. To this mixture add 0.4 mL of carbon tetrachloride followed by one drop of I_2–CCl_4 solution delivered from a Pasteur pipet.

HOOD

CAUTION: Carbon tetrachloride is a cancer suspect agent. Dispense it in the **hood** using an automatic delivery pipet.

Attach the vial to a reflux condenser fitted with a calcium chloride drying tube (⬅).

HOOD

INSTRUCTOR PREPARATION I_2–CCL_4 SOLUTION. Iodine (0.2 g, 0.01 mol) is dissolved in 10 mL of carbon tetrachloride. Place the solution in a ***hood*** *for student use.*

Reaction Conditions. Heat the reaction mixture, with stirring, to reflux in a sand bath at 85–95 °C for 1 h. During this time the solution turns brown and crystals of succinimide appear at the surface of the reaction solution.

Isolation of Product. Collect the succinimide product by vacuum filtration of the warm solution using a Hirsch funnel (⬅). Wash the filter cake of succinimide with three or four 0.5-mL portions of cyclohexane from a calibrated Pasteur pipet. Combine the washings with the original filtrate.

HOOD

Concentrate the filtrate to dryness in a **hood** under reduced pressure to obtain yellow-green crystals of 9-bromoanthracene. *Accelerate evaporation of the solvent by immersing the flask in warm water.*

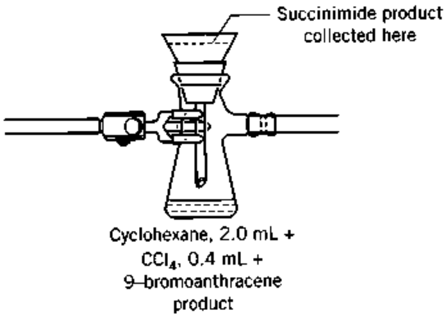

Purification and Characterization. Weigh the air-dried succinimide and calculate the percent yield. Determine the melting point and compare your value to that in the literature. Obtain an IR spectrum and compare your spectrum with that shown in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

The crude 9-bromoanthracene is purified by recrystallization from 95% ethanol using the Craig tube. Weigh the dried product and calculate the percent yield. Determine the melting point and compare your result with that listed in the literature.

Chemical Tests. The ignition test should establish the presence of the aromatic nature of the substituted anthracene compound. Does the Beilstein test indicate the presence of a halogen?

QUESTIONS

6-199. Predict and give a suitable name for the product(s) formed in the following reactions with NBS:

(a) 1-Propene + NBS $\xrightarrow[h\nu]{CCl_4}$

(b)

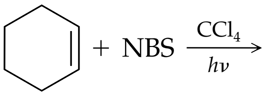

(c) $CH_3CH_2CH_2{-}CH{=}CH{-}CH_3 + NBS \xrightarrow[h\nu]{CCl_4}$

(d)

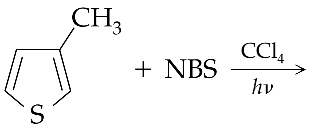

(e)

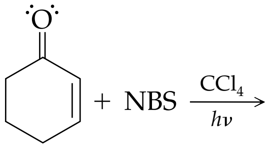

6-200. When 1-octene is treated with NBS, three monobromo straight-chain alkenes with the molecular formula $C_8H_{15}Br$ are isolated from the reaction mixture. Identify these compounds and give each a suitable name.

6-201. Benzyl bromide ($C_6H_5CH_2Br$) can be prepared by treating toluene with NBS in the presence of a peroxide initiator. Suggest a suitable mechanism to account for this reaction.

6-202. The benzyl radical has unusual stability. Account for this fact by drawing appropriate resonance structures.

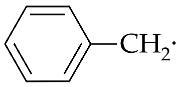

6-203. Suggest a suitable mechanism for the following reaction:

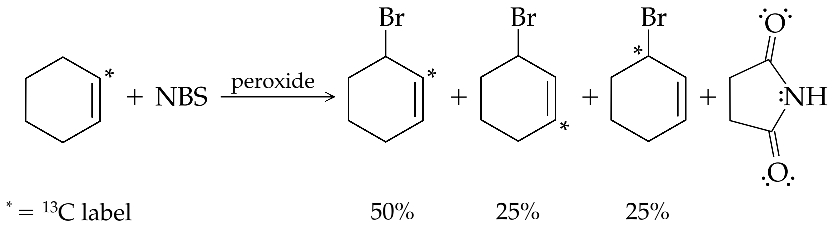

BIBLIOGRAPHY

Dauben, H. J. Jr.; McCoy, L. L. *J. Am. Chem. Soc.* **1959,** *81,* 4863.
Djerassi, C. *Chem. Rev.* **1948,** *43,* 271.
Horner, L.; Winkelmann, E. H. *Angew. Chem.* **1959,** *71,* 349.
Horner, L.; Winkelmann, E. H. *Newer Methods Prep. Org. Chem.* **1964,** *3,* 151.

For an overview of the scope of the reaction see

Smith, M. B.; March. *J. Advanced Organic Chemistry,* 6th ed.; Wiley-Interscience: New York, **2007,** Chap. 11, p. 699.
Also see Mauthen, H. A. *J. Org. Chem.* **1992,** *57,* 2740 for alternative brominating agents.

Brominations in *Organic Syntheses* using NBS:

Amat, M.; Hadida, S.; Sathyanarayana, S.; Bosch, J. *Organic Syntheses* **1997,** *74,* 248.

Campaigne, E.; Tullar, B. F. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 921.

Corbin, T. F.; Hahn, R. C.; Schechter, H. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 328.

Greenwood, F. L.; Kellert, M. D.; Sedlak, J. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 108.

Kalir, A. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 825.

Nakagawa, N.; Saegusa, J.; Tomozuka, M.; Ohi, M.; Kiuchi, M.; Hino, T.; Ban, Y. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 462.

EXPERIMENT 32

Hypochlorite Oxidation of an Alcohol: Cyclohexanone

Common name: cyclohexanone
CA number: [108-94-1]
CA name as indexed: cyclohexanone

Purpose. The oxidation of a secondary alcohol to a ketone is explored using the hypochlorite reagent. Steam distillation is used to separate the product from the reaction mixture and chromatographic techniques are used to purify and isolate the cyclohexanone product.

Prior Reading

- *Technique 1:* Gas Chromatography (pp. 55–61)
- *Technique 2:* Simple Distillation (pp. 61–64)
  - Steam Distillation
- *Technique 4:* Solvent Extraction
  - Liquid–Liquid Extraction (p. 72)
  - Drying of the Wet Organic Layer (pp. 80–83)
  - Salting Out (p. 79)
- *Technique 6A:* Thin-Layer Chromatography (pp. 97–99)

REACTION

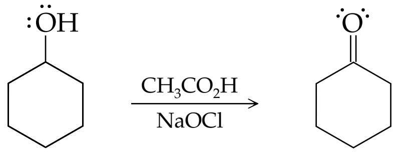

Cyclohexanol   Cyclohexanone

DISCUSSION

Sodium hypochlorite solutions (1.8–2.0 M) sold as liquid bleach are often described as having 11.5–12.5% available chlorine. The term "available chlorine" compares the oxidizing capacity of the solution relative to that of the same weight of chlorine, Cl_2.

Sodium hypochlorite solutions are used extensively in swimming pool sanitation and as a bleach in the pulp and textile industries. A less-concentrated product (5% available chlorine) is used in laundries and as household bleach. Consumption statistics for 1982 indicate that 210×10^3 tons of sodium hypochlorite were consumed in the United States alone. World consumption of sodium hypochlorite for household use is estimated to be 426×10^3 metric tons annually in 2005. The reaction described in this experiment illustrates the use of liquid bleach (11.5–12.5% available chlorine) as an oxidizing agent in the organic laboratory.

Sodium hypochlorite is prepared commercially by passing chlorine gas through a solution of aqueous sodium hydroxide:

$$Cl_2 + NaOH \rightleftharpoons NaOCl + NaCl$$

The actual oxidizing agent in the present experiment is the chloronium ion (Cl^+), which is reduced in the reaction to chloride ion (Cl^-). The cyclohexanol acts as a reducing agent, and thus becomes oxidized to cyclohexanone:

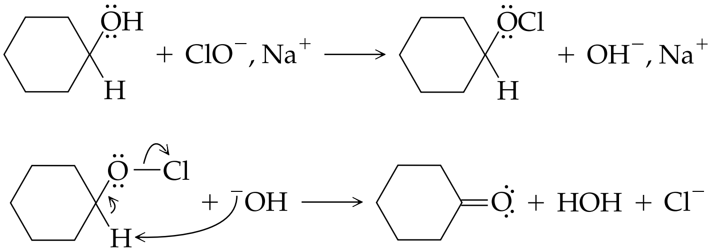

Chromium-based oxidants are commonly used to accomplish these transformations in the research laboratory (see Experiments [33A] and [E1]); the hypochlorite oxidation presents an opportunity to perform an oxidation using the much less costly and much less toxic hypochlorite solution as the oxidant.

In this experiment, a Hickman still is used to isolate the crude cyclohexanone product from the reaction mixture in an example of the steam distillation technique (see Prior Reading). The crude mixture collected in the collar of the still consists of cyclohexanone, water, and acetic acid. If any unreacted cyclohexanol is present in this mixture, it is removed in the subsequent chromatographic purification sequence using alumina. Gas chromatographic analysis may be used to determine the purity of the cyclohexanone product.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 3.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Cyclohexanol	100.16	100 mg	1.0	161	0.96	1.4641
Glacial acetic acid	60.05	250 μL		118		
Sodium hypochlorite soln (~12.5% available chlorine)		2 mL				

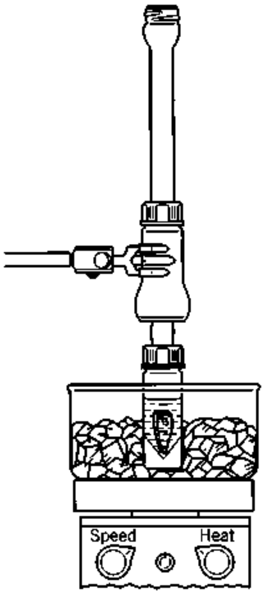

Cyclohexanol, 100 mg +
CH_3CO_2H, 250 μL +
11.5–12.5% NaOCl, 2 mL

Reagents and Equipment. Weigh and place 100 mg (1.0 mmol) of cyclohexanol in a 5.0-mL conical vial containing a magnetic spin vane. Add 250 μL of glacial acetic acid and then attach the vial to a Hickman still (⬅). *Dispense the glacial acetic acid in the **hood** by use of an automatic delivery pipet.*

HOOD

Cool the resulting solution in an ice bath and add dropwise, with stirring, 2.0 mL of aqueous sodium hypochlorite (NaOCl) solution (~12.5% available chlorine, 1.8–2.0 M) by use of a graduated pipet. Remove the ice bath following the addition. *Add the NaOCl solution by inserting the pipet down the neck of the still just into the throat of the vial.*

NOTE. Solid calcium hypochlorite (65% available chlorine) may be used as a source of oxidant in place of the sodium hypochlorite solution.

Reaction Condition. After 30 min, begin monitoring the reaction mixture by TLC. Using as a solvent system ethyl acetate:hexane (1:4), the R_f value of cyclohexanone is 0.6 (R_f value of cyclohexanol is 0.4) when stained with a solution of *p*-anisaldehyde (135 mL ethanol : 5 mL H_2SO_4 : 1.5 mL glacial acetic acid : 3.7 mL *p*-anisaldehyde). *If incomplete as judged by TLC, maintain an excess of hypochlorite oxidizing agent for an additional 30 min and check for the disappearance of starting material by TLC. Overoxidation is not likely. Monitor the aqueous layer periodically using KI–starch paper. If a positive test is not obtained (the paper should turn blue if an oxidant is present), add additional sodium hypochlorite solution (1–3 drops) to ensure that an excess of the oxidizing agent is present. Use a Pasteur pipet inserted down the neck of the still to add the reagent and a clean pipet to remove a drop of solution for testing with KI–starch paper.*

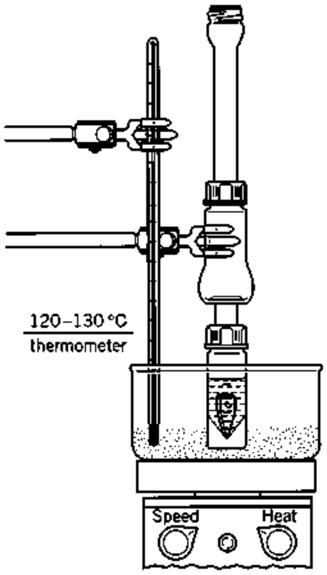

Cyclohexanol, 100 mg +
CH_3COOH, 250 μL +
11.5–12.5% NaOCl, 2 ml +
$NaHSO_3$(sat soln),
~8–10 drops

Isolation of Product. Using a Pasteur pipet, add a saturated, aqueous sodium bisulfite solution dropwise to the reaction mixture until the solution gives a negative KI-starch test.

Separate the crude product from the reaction mixture by steam distillation. With stirring, heat the mixture in a sand bath at a temperature of 120–130 °C (⬅). Collect the first 0.5–1.0 mL of distillate in the ring collar of the condenser. Using a Pasteur pipet, transfer this material to a 3.0-mL conical vial containing a spin vane and remove the heat from the still pot. After the distillation apparatus has cooled, rinse the condenser collar with 0.5 mL of diethyl ether and also transfer this rinse to the conical vial.

NOTE. The distillate collected in the condenser collar consists of cyclohexanone, water, and acetic acid.

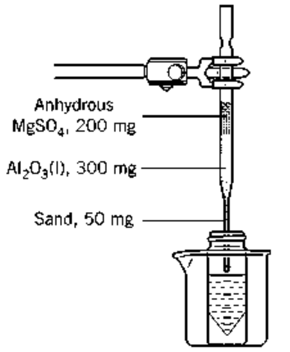

Oxidation product
in $(CH_3CH_2)_2O$,
~1.0 mL

To neutralize the acetic acid present in the separated product mixture, add anhydrous sodium carbonate (~100 mg) in small portions, with stirring, to the solution until evolution of CO_2 gas ceases. Now add 50 mg of NaCl to the mixture. The resulting two-phase system is stirred until all the solid material dissolves.

Use a Pasteur filter pipet to separate the ether layer containing the cyclohexanone product from the aqueous phase and transfer it to a microscale drying and chromatography column. Assemble the column using a Pasteur filter pipet packed first with about 50 mg of sand, then with 300 mg of alumina (Activity I, see Glossary), followed by 200 mg of anhydrous magnesium sulfate (⬅). Wet the column with diethyl ether before the transfer.

Collect the eluate in a tared 3.0-mL conical vial containing a boiling stone. Extract the aqueous layer remaining in the conical vial with three 0.5-mL portions of ether, and also pass each extract through the column and combine these fractions with the original eluate.

Fit the 3.0-mL conical vial with an air condenser and remove the ether by gentle warming on a sand bath in the **hood.**

HOOD

Purification and Characterization. The liquid residue of cyclohexanone isolated upon evaporation of the ether solvent is sufficiently pure for characterization. It may be further purified by prep-GC (see below). Weigh the product and calculate the percent yield.

The purity of the cyclohexanone product may be determined using *gas chromatographic analysis (see Prior Reading and Experiment [5A] to review this technique and the experimental conditions for its use)*. Determine the boiling point and density of your product; compare your results to those given in the literature.

Obtain the IR spectrum of your product and compare it with one in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Chemical Tests. Chemical classification tests may also be run to assist in the characterization of this material. The 2,4-dinitrophenylhydrazine test (Chapter 9) should give a positive result. Isolation of this derivative and the determination of its melting point would further establish the identity of the product as cyclohexanone.

QUESTIONS

6-204. In the experiment, why is a solution of sodium bisulfite added to the reaction product mixture (see Isolation of Product)? Write a reaction to account for what is happening. Is the bisulfite ion acting as an oxidizing or reducing agent?

6-205. In the isolation of the cyclohexanone product, 50 mg of sodium chloride is added to the water–cyclohexanone–diethyl ether mixture. Explain how the addition of sodium chloride aids in isolation of the cyclohexanone product.

6-206. Predict the product(s) for each of the following oxidation reactions. Give a suitable name for each reactant and product.

(a)

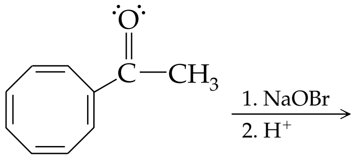

(b) $CH_3CH_2\ddot{\underset{..}{O}}H \xrightarrow[2.\ H^+]{1.\ NaOI}$

(c)

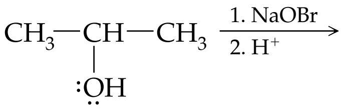

$CH_3-\underset{:\ddot{O}H}{\underset{|}{CH}}-CH_3 \xrightarrow[2.\ H^+]{1.\ NaOBr}$

6-207. 2,3-Dimethyl-2,3-butanediol (*pinacol*), upon heating in aqueous acid, rearranges to form 3,3-dimethyl-2-butanone (*pinacolone*). Suggest a mechanism for this reaction.

6-208. What chemical tests might be used to distinguish between pentanal and 2-pentanone? Between benzyl alcohol and diphenylmethanol?

BIBLIOGRAPHY

As a general reference for the use of sodium hypochlorite as an oxidant, see

Fieser, L. F; Fieser, M. *Reagents for Organic Synthesis;* Wiley: New York, 1967; Vol. I, p. 1084.

Nohrig, J. R.; Nienhuis, D. M.; Linck, C. F.; Van Zoeren, C.; Fox, B. G. *J. Chem. Educ.* **1985,** *62,* 519.

Experiment [32] is based on work reported by

Stevens, R.V.; Chapman, K. T.; Weller, H. N. *J. Org. Chem.* **1980,** *45,* 2030.

See also Zuczek, N. M.; Furth, P. S. *J. Chem. Educ.* **1981,** *58,* 824.

EXPERIMENT 33

Chromium Trioxide–Resin or Hypochlorite Oxidation of an Alcohol: 9-Fluorenone

Common name: 9-fluorenone
CA number: [486-25-9]
CA name as indexed: 9*H*-fluoren-9-one

Purpose. The experiment investigates the use of a polymer-bound chromium trioxide oxidizing agent in the oxidation of a secondary (2°) alcohol to a ketone. An alternative oxidizing agent, sodium hypochlorite, may be used instead. The progress of the reaction is monitored by thin-layer chromatography (TLC). The product ketone can be characterized by formation of its 2,4-dinitrophenylhydrazone (2,4-DNP) derivative.

Prior Reading

Technique 5: Crystallization
- Use of the Hirsch Funnel (pp. 88–89)
- Craig Tube Crystallizations (pp. 89–91)

Technique 6: Chromatography
- Thin-Layer Chromatography (pp. 97–99)
- High-Performance Liquid Chromatography (pp. 100–101)
- Concentration of Solutions (pp. 101–104)

Experiment 33A

9-Fluorenone: CrO_3 Oxidation of 9-Fluorenol

REACTION

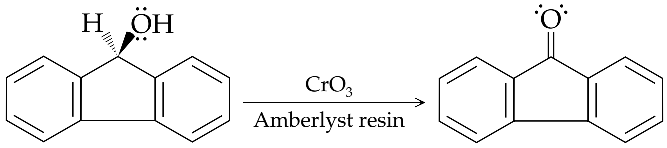

9-Fluorenol

9-Fluorenone

DISCUSSION

This experiment illustrates the oxidation of a 2° alcohol to a ketone. The oxidizing agents commonly used for this purpose are sodium dichromate or chromic oxide in sulfuric acid. In the present case, a convenient and advantageous polymer-bound chromium trioxide reagent is used. It is not only easy to prepare, but is also easy to separate from the product mixture and can be recycled. Today, the use of polymer-bound reagents in organic synthesis is developing at a rapid pace. The mechanism of the oxidation is outlined here:

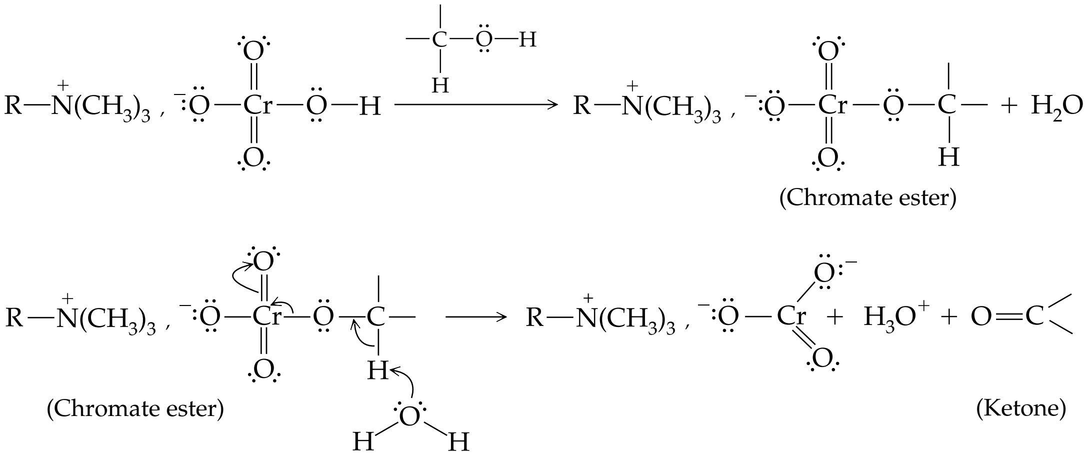

As seen, the oxidation proceeds through the formation of a chromate ester. Note that the oxidation state of chromium, Cr(VI), at this stage remains the same as in the starting reagent. The second stage is equivalent to an E2 elimination reaction, with the water molecule acting as a base. The donation of an electron pair to the metal atom changes its oxidation state to Cr(IV).

Note that a solution of chromic oxide in aqueous sulfuric acid (the Jones reagent) is used as a test reagent for 1° and 2° alcohols. A positive test is observed when the clear orange test reagent gives a greenish opaque solution upon addition of the alcohol (see Chapter 9).

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 3–4 h.

INSTRUCTOR PREPARATION. Prepare the CrO_3 *resin by adding 35 g of Amberlyst A-26 resin to a solution of 15 g of* CrO_3 *in 100 mL of water. Stir the mixture for 30 min at room temperature. Then collect the resin by vacuum filtration using a Büchner funnel and successively rinse it with water and acetone. Partially dry the material on the Büchner funnel by drawing air through the resin under vacuum for 1 h. Allow it to air-dry overnight.*

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
9-Fluorenol	182.23	100 mg	0.55	154	
Chromic oxide–resin		500 mg			
Toluene	92.15	3.5 mL			111

NOTE. CrO_3 *has been listed as a known carcinogen (see* The Merck Index, *12th ed.; 1996; p. 375, no. 2293).*

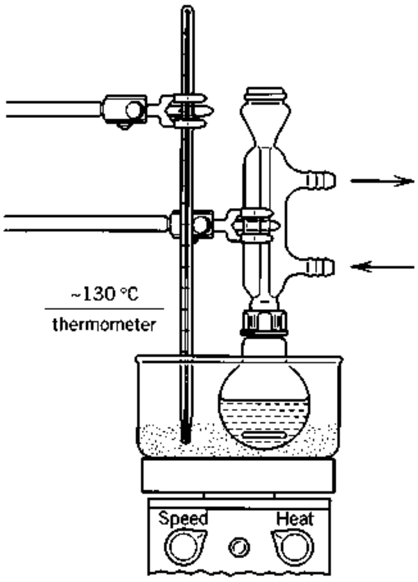

9-Fluorenol, 100 mg + CrO_3 resin, 500 mg $CH_3C_6H_5$, 3.5 mL, 10-mL RB flask

Reagents and Equipment. Weigh and add 100 mg (0.55 mmol) of 9-fluorenol to a 10-mL round-bottom flask containing a magnetic stirring bar. Add 3.5 mL of toluene, and sample the resulting solution for TLC analysis (see Prior Reading). Now add 500 mg of the oxidizing resin and attach the flask to a reflux condenser (⬅). If the longest Pasteur pipet available to you will not reach down the condenser and allow you to sample the solution, consider placing a Claisen head between the condenser and the flask (using the neck above the bend to connect the condenser. If you do this, be sure to cap the other neck of the Claisen head, and be sure to open it for the minimum possible amount of time when sampling—otherwise toluene vapors will escape into the air you breathe.

Reaction Conditions. Heat the contents of the flask to reflux, with stirring, using a sand bath temperature of approximately 130 °C.

Sample the solution for TLC analysis after a period of 5 min and every 15–20 min thereafter. In this manner, the progress of the reaction may be monitored until the conversion is complete (~35–40 min). This point is reached when the TLC analysis shows that the 9-fluorenol has been completely consumed.

SUGGESTED TLC CONDITIONS. Use Eastman Kodak fluorescent silica gel sheets (1 × 6 cm). Elute them with methylene chloride solvent and visualize the spots by UV light. Determine reference R_f values for 9-fluorenol and 9-fluorenone using known reference samples under the same conditions.

Isolation of Product. Cool the reaction mixture and remove the resin by gravity filtration through a cotton plug placed in a small funnel. Transfer the solution to the filter funnel using a Pasteur pipet, and collect the filtrate in a tared 10-mL Erlenmeyer flask containing a boiling stone. Rinse the reaction flask and resin with two 1.0-mL portions of methylene chloride using a calibrated Pasteur pipet. Combine the rinse with the original filtrate.

HOOD In the **hood** remove the solvent from the filtrate under a stream of nitrogen by warming in a sand bath to yield a yellow colored residue of crude 9-fluorenone product.

Purification and Characterization. Obtain the weight of the crude product and calculate the percent yield.

Recrystallize a 30- to 50-mg portion of the 9-fluorenone from hexane (~1.0 mL of hexane/50 mg of ketone) using a Craig tube. Determine the melting point of this purified material and compare your data to those reported in the literature. Calculate the percent recovery on recrystallization.

NOTE. High-performance liquid chromatography (see Prior Reading) is effective in separating the pure ketone. Use a C_{18} reversed-phase column and methanol–water as the elution solvent.

Chemical Tests. The 2,4-dinitrophenylhydrazine test for aldehydes and ketones (see Chapter 9) may be done to confirm the presence of the carbonyl group. The isolation of this derivative and the determination of its melting point (Lit. value = 283 °C) would aid in establishing the identity of the product.

Does the ignition test indicate that this compound is aromatic?

9-Fluorenone has a characteristic UV spectrum. The following data were obtained at a concentration of 8.66×10^{-6} M (see Chapter 8, Ultraviolet–Visible Discussions):

$$\lambda_{\max} 248 \text{ nm} (\varepsilon_{\max} = 5.27 \times 10^4, \text{ methanol})$$
$$\lambda_{\max} 255 \text{ nm} (\varepsilon_{\max} = 7.83 \times 10^4, \text{ methanol})$$
$$\lambda_{\max} 290 \text{ nm} (\varepsilon_{\max} = 3.93 \times 10^3, \text{ methanol})$$

9-Fluorenone: NaOCl Oxidation of 9-Fluorenol

Experiment 33B

REACTION

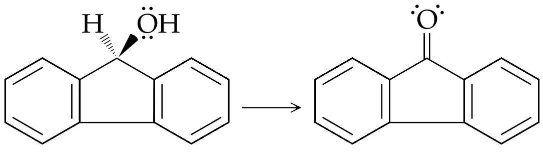

9-Fluorenol 9-Fluorenone

DISCUSSION

This experiment is a further illustration of the oxidation of a 2° alcohol to a ketone. The reagent used is sodium hypochlorite solution. This reagent is also used in oxidation reactions described in Experiments [32] and [34A, B]. The chemistry of the reagent, sodium hypochlorite, is discussed in Experiment [32]. This alternative reagent for the oxidation of 9-fluorenone offers a material that is cheap (Clorox, a household bleach), easy to handle, and environmentally safe.

As in experiment [33A], the progress of the reaction is followed using TLC.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 3 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
9-Fluorenol	182.23	50 mg	0.27	154	
Acetone	58.08	3 mL			56.5
Acetic acid, glacial	60.05	120 μL		16.7	118
Sodium hypochlorite (5.25%)	74.44	~1.3 mL			

Reagents and Equipment. Weigh and add 50 mg (0.27 mmol) of 9-fluorenol to a 10-mL round-bottom flask containing a magnetic stir bar. Add 3 mL of acetone and *sample the resulting solution (2–4 μL) for TLC analysis* (see Prior Reading). Add 120 μL of glacial acetic acid and then attach the flask to an air condenser (➡).

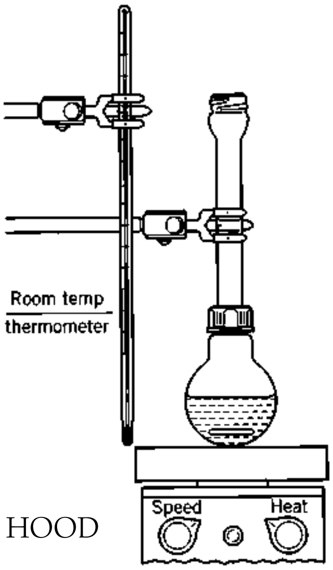

A 10-mL RB flask

CAUTION: Dispense the glacial acetic acid in the **hood** by use of an automatic delivery pipet. This acid burns skin!

HOOD

Use a graduated 1-mL pipet to slowly add, with gentle stirring, 0.4 mL of aqueous sodium hypochlorite solution (5.25%—commercial bleach). Add the NaOCl solution by inserting the pipet down the neck of the condenser just to the throat of the flask.

Reaction Conditions. Gently stir the reaction mixture at room temperature. After a period of 5 min, sample the solution (2–5 μL) for TLC analysis. *Sample the reaction solution by inserting a 9-in. Pasteur pipet down the air condenser to obtain a drop of solution by capillary action.* If starting material is present as shown by TLC analysis, add an additional 0.4 mL of the hypochlorite solution as before. Stir for 5 min and sample again. Continue this process until TLC analysis shows that the 9-fluorenol has been completely consumed. Approximately 1.2–1.4 mL of hypochlorite solution should be sufficient, if the reagent is fresh.

SUGGESTED TLC CONDITIONS. Use Eastman Kodak Fluorescent silica gel sheets (1.0 cm × 6 cm). Elute the plates with 30% acetone/70% hexane solvent, and visualize the spots by UV light. R_f *values are 0.56 for 9-fluorenol and 0.80 for 9-fluorenone.*

Isolation of Product. Using a Pasteur pipet, transfer the reaction solution to a stoppered 15-mL centrifuge tube. Rinse the flask with 300 μL of acetone and transfer the rinse to the same centrifuge tube. Extract the solution with two 2-mL portions of hexane. Transfer the hexane extracts (upper layer) to a second centrifuge tube. Wash the combined hexane extracts with one 1-mL portion of 5% sodium bicarbonate solution followed by one 2-mL portion of water. Dry the hexane solution by addition of a small amount of anhydrous sodium sulfate. Use a Pasteur filter pipet to transfer approximately 3.0 mL of the dried solution to a tared 5-mL conical vial. Concentrate the solution, which contains a portion of the desired product, to dryness on a warm sand bath under a stream of nitrogen or dry air in the **hood.** Transfer the remaining hexane solution to the same vial. Rinse the sodium sulfate with an additional 1 mL of hexane and transfer this rinse to the same vial. Concentrate the solution as before.

HOOD

Purification and Characterization. See this section in Experiment [33A].

QUESTIONS

6-209. Suggest a suitable mechanism for the reaction of 9-fluorenone with 2,4-dinitrophenylhydrazine to form the corresponding 2,4-dinitrophenylhydrazone.

6-210. It is also possible to characterize 9-fluorenone by preparation of an oxime or semicarbazone. Formulate equations showing clearly the formation of these two derivatives and name each reagent used in the preparation.

6-211. As indicated in the Discussion, a solution of chromic oxide in aqueous sulfuric acid is used as a test reagent for 1° and 2° alcohols.

- **(a)** What is this test (consult Chapter 9)?
- **(b)** Predict which of the following alcohols will give a positive test with the chromic oxide reagent. Give the structure for each of the alcohols: 1-heptanol, 2,2,3-trimethyl-3-pentanol, cholesterol, 3-methyl-2-butanol, and 4-*tert*-butylcyclohexanol.

6-212. There are actually **two** isomeric 2,4-dinitrophenylhydrazones of 2-pentanone. Draw the structures of these isomers.

6-213. 2-Pentanone, in reference to Question 6-212, also forms a derivative on treatment with semicarbazide:

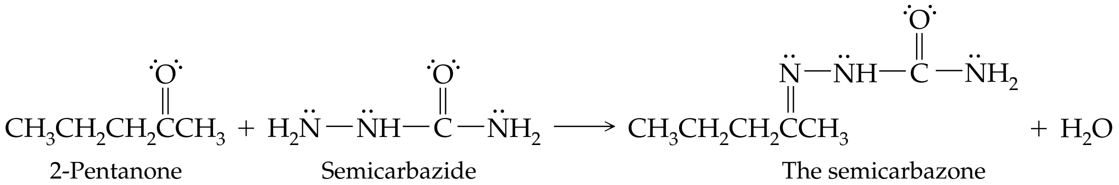

$$CH_3CH_2CH_2\overset{\overset{\cdot\cdot}{O}\,\cdot\cdot}{\overset{\|}{C}}CH_3 + H_2\ddot{N}-\ddot{N}H-\overset{\overset{\cdot\cdot}{O}\,\cdot\cdot}{\overset{\|}{C}}-\ddot{N}H_2 \longrightarrow CH_3CH_2CH_2\overset{\ddot{N}-\ddot{N}H-\overset{\overset{\cdot\cdot}{O}\,\cdot\cdot}{\overset{\|}{C}}-\ddot{N}H_2}{\overset{\|}{C}}CH_3 + H_2O$$

2-Pentanone Semicarbazide The semicarbazone

Note that semicarbazide has **two** —NH_2 groups that might react with the carbonyl of the ketone to form the semicarbazone. Explain why it reacts as depicted above.

BIBLIOGRAPHY

Experiment [33] is an adaptation of that reported by

Wade, L. G., Jr.; Stell, L. M. *J. Chem. Educ.* **1980,** *57,* 438; Experiment [33B] from the work of Jones, C. S.; Albizati, K. *J. Chem. Educ.* **1994,** *71,* A271.

Introduction to the use of polymer-bound reagents:

Hodge, P. *Chem. in Britain* **1978,** *14,* 237.
Leznoff, C. C. *Acc. Chem. Res*. **1978,** *1,* 327.
McKillopp, A.; Young, D. W. *Synthesis* **1979,** 401.
Smith, K., Ed. *Solid Supports and Catalysts ln Organic Synthesis;* Ellis Harwood: Chichester, UK, 1992.

Selected chromate oxidations reported in *Organic Syntheses* include

Boeckman, R. K., Jr.; Blum, D. M.; Ganem, B.; Halvey, N. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 1033.
Conant, J. B.; Quayle, O. R. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 211.
Eisenbraun, E. J. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 310.
Fieser, L. F. *Organic Syntheses;* Wiley: New York, **1963;** Collect. Vol. IV, p. 189.
Krumpolc, M.; Rocek, J. *Organic Syntheses;* Wiley: New York, 1990; Collect. Vol. VII, p. 114.
Zibuck, R.; Streiber, J. *Organic Syntheses;* **1993,** *71,* 236.

For reviews of chromium catalyzed oxidations in organic synthesis see

Cainelli, G.; Cardillo, G. *Chromium Oxidations in Organic Chemistry;* Springer-Verlag: New York, 1984.
Muzart, J. *Chem. Rev.* **1992,** *92,* 113.

For an overview of the oxidation of alcohols using sodium hypochlorite solutions, see

Mohrig, J. R.; Nienhuis, D. M.; Linck, C. F.; Van Zoeren, C.; Fox, B. G. *J. Chem. Educ.* **1985,** *62,* 519.

EXPERIMENT 34

Hypochlorite Oxidation of Methyl Ketones by the Haloform Reaction: Benzoic Acid and p-Methoxybenzoic Acid

Common name: benzoic acid
CA number: [65-86-0]
CA name as indexed: benzoic acid

Common names: *p*-methoxybenzoic acid, 4-methoxybenzoic acid, *p*-anisic acid
CA number: [100-09-4]
CA name as indexed: benzoic acid, 4-methoxy-

Purpose. The well-known haloform reaction is explored as a synthetic route to the preparation of organic acids. You will investigate the use of a basic aqueous solution of hypochlorite ion as a source of molecular chlorine (Cl_2).

Prior Reading

Technique 4: Solvent Extraction
Liquid–Liquid Extraction (p. 72)
Separatory Funnel Extraction (Scaleup) (pp. 75–77)
Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallizations (pp. 89–91)

REACTION

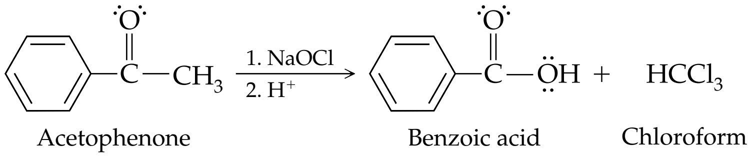

DISCUSSION

The reaction of methyl ketones with a halogen in an alkaline medium is known as the haloform reaction. In this experiment, halogen and base are present because of the equilibrium reaction shown here:

$$H_2O + NaOCl + NaCl \rightleftharpoons 2\,NaOH + Cl_2$$

In the haloform reaction, two products are formed: (1) a haloform ($CHCl_3$, $CHBr_3$, or CHI_3, depending on the halogen used); and (2) the carboxylic acid having one less carbon atom than the starting ketone. It is the formation of the carboxylic acid that gives the reaction its synthetic utility. An advantage of this oxidation is that it does not affect carbon–carbon double bonds, as illustrated here:

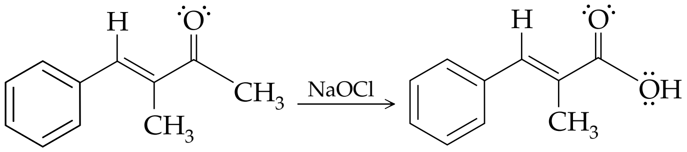

If I_2 in a basic solution is used, iodoform (CHI_3) is generated in the reaction. This compound is a yellow solid that precipitates from the reaction medium. This observation has been used extensively as a chemical test **(iodoform test)** for methyl ketones and the $RCH(OH)CH_3$ structural group (see Chapter 9).

In the present experiment, a solution of sodium hypochlorite (NaOCl) is used as the source of the chlorine. Sodium hypochlorite solutions are marketed with the hypochlorite concentration described in terms of the "available chlorine" content, which is a term comparing the oxidizing potential of the solution with that of the equivalent mass of chlorine. For the purposes of this experiment, a solution of NaOCl that has 5% available chlorine is needed, and common household bleach will suffice. For another example of an oxidation with an aqueous hypochlorite solution, see Experiment [32].

The reaction takes place in two stages. In the first stage, the methyl group is trihalogenated in a stepwise fashion. In the second stage, hydroxide ion attacks the carbonyl carbon of the trihaloketone to generate the haloform along with the alkali metal salt of the acid; acidification yields the carboxylic acid. The mechanistic sequence follows:

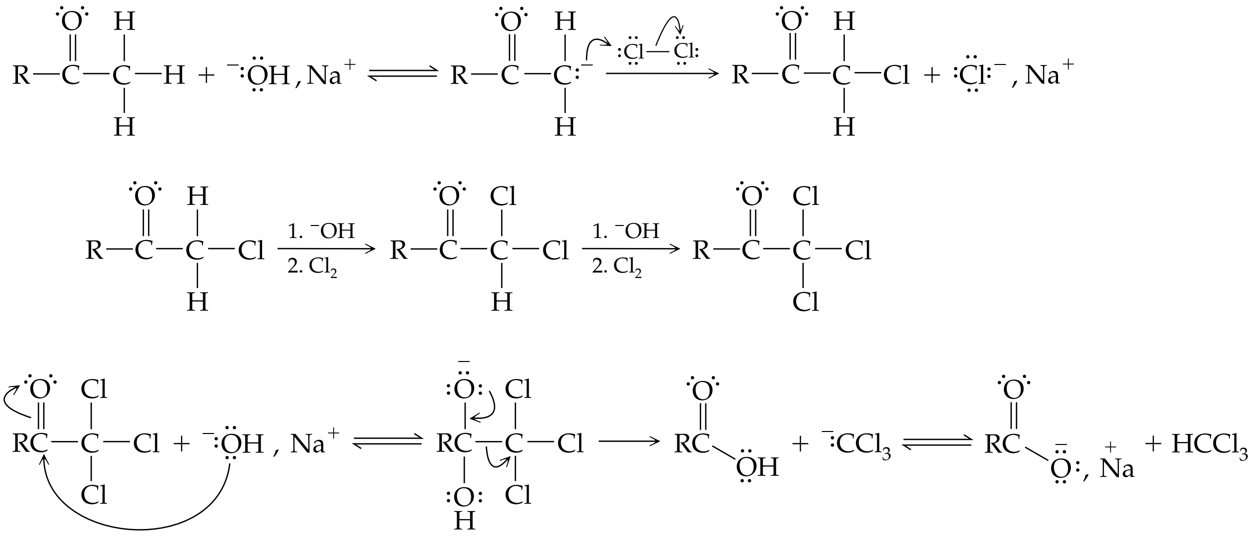

Benzoic Acid

The reaction is shown above.

Experiment 34A

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Acetophenone	120.16	60 μL	51	20.7	202.6	1.03	1.5372
Aqueous NaOCl (household bleach, 5% available chlorine)		2.1 mL					
Sodium sulfite	120.6	15 mg					

Reagents and Equipment. To a 5-mL conical vial containing a magnetic spin vane and equipped with an air condenser, add 6 μL (63 mg, 51 mmol) of acetophenone and 2.1 mL of household bleach (NaOCl, 5% available chlorine) (➡).

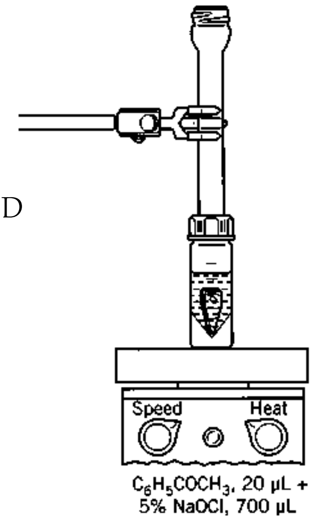

$C_6H_5COCH_3$, 20 μL + 5% NaOCl, 700 μL

CAUTION: Both reagents are irritants and should be dispensed in the **hood** using automatic delivery pipets.

HOOD

Reaction Conditions. Stir the mixture at room temperature for 30 min.

Isolation of Product. Add approximately 15 mg of sodium sulfite to destroy any unreacted bleach, and stir the reaction mixture briefly. Extract the resulting mixture with two 0.5-mL portions of ether (calibrated Pasteur filter pipet). Separate each portion of ether extract using a Pasteur filter pipet.

NOTE. The ether layer is the top layer; the lower, alkaline, layer contains the acid product. The ether extraction removes the chloroform generated in the reaction and any unreacted acetophenone. Save the ether extracts in a small vial until you have successfully isolated and characterized the final product.

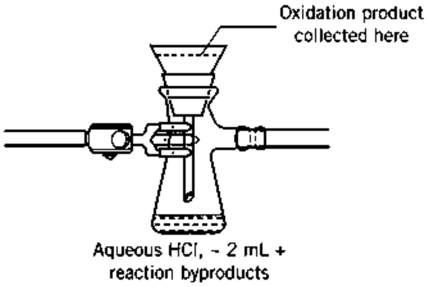
Oxidation product collected here
Aqueous HCl, ~ 2 mL + reaction byproducts

Acidify the aqueous layer (check with pH paper) by adding dropwise 3 M HCl from a Pasteur pipet. A thick, white precipitate of benzoic acid then appears. Collect the solid by vacuum filtration using a Hirsch funnel, and wash the filter cake with three 0.5-mL portions of water (calibrated Pasteur pipet) (⬅). Maintain the vacuum for approximately 5 min by covering the funnel with plastic food wrap (see Prior Reading) to partially dry the product. Transfer the material to a porous clay plate or filter paper to complete the drying process.

Purification and Characterization. The product is of sufficient purity for characterization. Weigh the material and calculate the percent yield.

Determine the melting point and compare your result with the value listed in the literature. If desired, obtain an IR spectrum and compare it with an authentic sample of benzoic acid, or with that given in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Chemical Tests. Chemical classification tests may also be used to assist in product analysis. You might wish to perform the following:

1. The ignition test should indicate the presence of the aromatic ring (Chapter 9).

2. The solubility of the compound in water, 5% NaOH, and 5% $NaHCO_3$ should be checked (see Chapter 9). Does the water solution turn blue litmus paper red? Is there evidence of CO_2 evolution when the benzoic acid is added to the bicarbonate solution? If positive results are obtained in these tests, how do they confirm that a carboxylic acid is present?

3. The preparation of an amide derivative (see Chapter 9) would also aid in establishing the identity of the acid product.

Experiment 34B

p-Methoxybenzoic Acid

REACTION

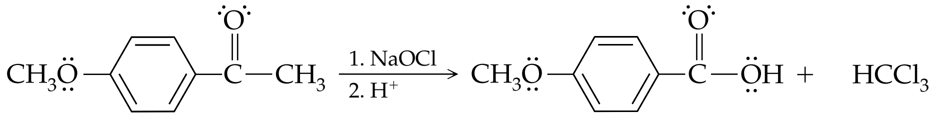

p-Methoxyacetophenone *p*-Methoxybenzoic acid Chloroform

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
p-Methoxyacetophenone	150.8	29 mg	0.19	38–39	258
Aqueous NaOCl (household bleach, 5% available chlorine)		700 µL			
Sodium sulfite	120.6	5 mg			

Reagents and Equipment. Weigh and add 29 mg (0.19 mmol) of *p*-methoxyacetophenone to a 3.0-mL conical vial containing a magnetic spin vane. Now add 700 μLof household bleach (5% available chlorine), and then attach the vial to an air condenser (➡).

CAUTION: The bleach is an irritant to skin and eyes. It should be dispensed using an automatic delivery pipet in the **hood.**

HOOD

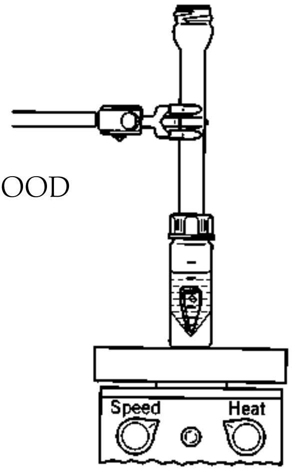

p-$CH_3OC_6H_4COCH_3$,29 mg
+ 5% NaOCl, 700 μL

Reaction Conditions. Stir the mixture for 30 min with very gentle heating. *Use the lowest possible setting on a hot plate magnetic stirrer.*

Isolation of Product. Weigh and add approximately 5 mg of sodium sulfite. Stir the reaction medium briefly and then cool it to room temperature. Extract the resulting mixture with two 0.5-mL portions of ether (calibrated Pasteur pipet). Separate each portion of ether extract using a Pasteur filter pipet.

NOTE. The ether layer is the top layer; the bottom aqueous layer contains the product. The ether extraction removes the chloroform generated in the reaction and any unreacted p*-methoxyacetophenone. Save the ether extracts in a small vial until you have successfully isolated and characterized the final product.*

Acidify the aqueous layer by the dropwise addition (check with pH paper) of 3 M HCl from a Pasteur pipet until a thick, white precipitate of *p*-methoxybenzoic acid is formed. Cool the mixture in an ice bath for 5 min and collect the solid by vacuum filtration using a Hirsch funnel (➡). Wash the filter cake with three 0.5-mL portions of cold water (calibrated Pasteur pipet). Maintain the vacuum for approximately 5 min and cover the funnel with plastic food wrap (see Prior Reading) to partially dry the product. Transfer the material to a porous clay plate or filter paper to complete the drying process.

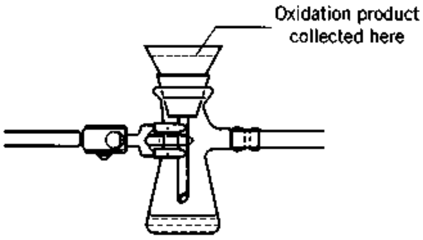

Aqueous HCl, ~2 mL +
reaction byproducts

Purification and Characterization. The product is reasonably pure but may be recrystallized from ethanol–water, if desired, using a Craig tube. Weigh the product and calculate the percent yield. Determine the melting point and obtain an IR spectrum. Compare your results with those reported in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar). Chemical characterization tests can be used to assist in the classification of this product, as outlined in Experiment [34A].

OPTIONAL SEMIMICROSCALE PREPARATION

This preparation may be scaled up by a 15-fold increase in reagent amounts. The procedure is similar to that given above for the micropreparation with the exceptions noted below. *A major difference is that a separatory funnel is used in the extraction step.*

1. Use a 25-mL round-bottom (RB) flask containing a magnetic stirring bar and equipped with an air condenser. The reaction conditions are as given in the microscale experiment (➡).

2. The reagent and solvent amounts are increased 15-fold.

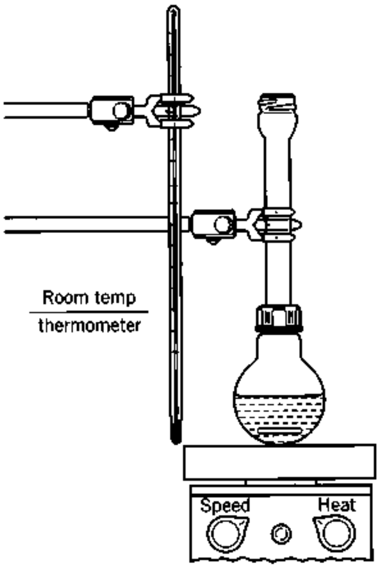

25-mL RB flask

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
p-Methoxyacetophenone	150.8	435 mg	2.88	38–39	258
Aqueous NaOCl (household bleach, 5% available chlorine)		10.5 mL			
Sodium sulfite	120.6	75 mg			

3. Add the sodium sulfite, stir, cool, and then transfer the product mixture to a 125-mL separatory funnel. Extract this mixture with two 7.5-mL portions of diethyl ether. Save the ether extracts in a small Erlenmeyer flask until you have successfully isolated and characterized the final product. The product is in the aqueous layer.

NOTE. Carry out the extraction by first returning the lower aqueous layer to the reaction flask and then removing the ether layer. Then return the aqueous layer to the separatory funnel and rinse the reaction flask with the second portion of ether. This ether rinse is then added to the separatory funnel to complete the extraction.

4. Collect the product by vacuum filtration, then wash the filter cake with three 7-mL portions of cold water.

5. The pale-yellow product is characterized as described in the microscale procedure.

QUESTIONS

6-214. In the haloform reaction, once the first α-hydrogen atom is replaced by a halogen atom, each successive hydrogen is more easily substituted until the trihalo species is obtained. Explain.

6-215. The haloform reaction using I_2 and NaOH is referred to as the "iodoform" test for methyl ketones (see Chapter 9). The test also gives positive results for compounds containing the $—CH(OH)CH_3$ group. This results from the oxidation of the alcohol to the methyl ketone in the first stage. Write a balanced equation for the conversion of $C_6H_5CHOHCH_3$ to the methyl ketone in the presence of I_2 and NaOH. Identify which species is being oxidized and which is being reduced.

6-216. If you were carrying out an industrial scale synthesis in which one step involved a haloform reaction to convert a methyl ketone into the corresponding acid having one less carbon atom, would you use NaOH and Cl_2, NaOH and Br_2, or NaOH and I_2 as the reagent? Give reasons for your choice.

6-217. Can you explain the fact that even though dibenzoylmethane ($C_6H_5COCH_2COC_6H_5$) is not a methyl ketone, it gives a positive iodoform test when treated with the NaOH and I_2.

6-218. Do you think that acetaldehyde would give a positive iodoform test? Explain your reasoning.

6-219. A water-soluble phenol is also acidic toward litmus paper as is a water-soluble carboxylic acid. How would you distinguish the difference between an aromatic acid and a phenol using a chemical test?

BIBLIOGRAPHY

Selected examples of the haloform reaction from *Organic Syntheses*:

Newman, M. S.; Holmes, H. L. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 428.

Sanborn, L. T.; Bousquet, E. W. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 526.

Smith, L. I.; Prichard, W. W.; Spillane, L. J. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 302.

Smith, W. T.; McLeod, G. L. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 345.

Staunton, J.; Eisenbraun, E. J. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 8.

For a review of oxidation see

Fuson, R. C.; Bull, B. A. *Chem Rev.* **1934,** *15,* 275.
Kurti, L.; Czako, B. *Application of Named Reactions in Organic Synthesis,* Elsevier: Amsterdam, 2005.
Smith, M. B.; March. J. *Advanced Organic Chemistry,* 6th ed.; Wiley-Interscience: New York, 2007, Chap. 12, p. 842.
Wiberg, K. B. In *Oxidation in Organic Chemistry;* Trahanovsky, W. S., Ed.; Academic Press: New York, 1978, Vol. 5, Part A, Chapter 2.

Conversion of Cyclohexyl Bromide to Cyclohexene–An E2 Elimination Reaction: Factors Affecting the Rate of a Chemical Reaction

EXPERIMENT 35

Common name: cyclohexene
CA number: [110-83-8]
CA name as indexed: cyclohexene

Purpose. A dehydrohalogenation reaction is designed and carried out for conversion of cyclohexyl bromide (bromocyclohexane) to cyclohexene using a strong base. The relative rate of this reaction is followed with gas chromatography. You will explore the effect of varying several parameters that alter the rate of the reaction such as temperature, concentration of the base, and the nature of the leaving group.

Prior Reading
- *Technique 1:* Gas Chromatography (pp. 55–61)
- *Technique 4:* Solvent Extraction (p. 67)
- *Chapter 3:* Reflux and reflux apparatus (pp. 23–24)
  - Automatic delivery pipets: (pp. 37–38)
  - Theoretical yield calculations: (p. 42)
- *Chapter 6:* Experiment [10]: (pp. 217–224)
- *Lecture text on kinetic theory and mechanisms of E2 reactions*

REACTION

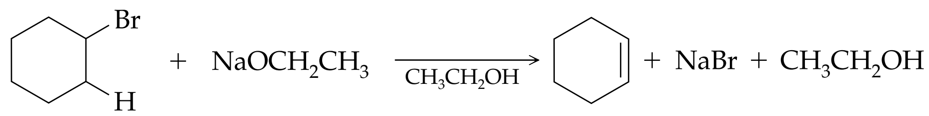

Cyclohexyl bromide Sodium ethoxide Cyclohexene

DISCUSSION

The rates of chemical reactions and the factors that affect them are of great importance in chemistry. This area of chemistry is termed *kinetics*. Most of the ionic reactions observed in inorganic chemistry take place so fast that their rates cannot be easily measured, but in many organic reactions the study of chemical kinetics is particularly important. This is because measurement of reaction rates is a powerful guide for probing reaction mechanisms.

These rates can vary greatly. Five factors are involved in establishing the rate at which a reaction may proceed: (1) the nature of the reacting substances, (2) the states of the reactants, (3) the temperature of the reactants, (4) the concentration of the reactants, and (5) the presence of a catalyst. The rate of a chemical reaction is measured by the decrease in concentration of a reactant or the increase in concentration of a product in a unit period of time.

The relationship between the rate of a chemical reaction and the change in concentration of the reactants can be described quantitatively based on experimental measurements. Each reaction has its own *rate equation*. In general, a rate equation has the form

$$\text{rate} = k[\text{A}]^a[\text{B}]^b[\text{C}]^c \qquad 6\text{–}1$$

in which [A], [B], and [C] represent molar concentrations of reactants (or products or other substances), *k* is the rate constant for a certain reaction at a particular temperature, and the exponents *a, b,* and *c* are positive and usually integers. Both *k* and the exponents must be determined experimentally by observing the variation in the rate of the reaction as the concentrations of the reactants are varied. It is important to recognize that the rate equation depends on the mechanism of the reaction and on the individual steps therein.

In the general rate equation, the exponent *a* is called the order of the reaction with respect to the reactant A. Likewise, the exponent *b* is the order of the reaction with respect to B, and so on. The sum of the exponents is the overall order of the reaction. In the E2 reaction, investigated in the present experiment, we have (see above) a dehydrohalogenation following this mechanism.

Experimentation shows us that if the concentration of the reactant (cyclohexyl bromide) is doubled, the reaction rate doubles. The rate is also doubled if the concentration of the reagent, sodium ethoxide ($NaOC_2H_5$), is doubled. Thus, the concentration of both species present affects the rate of the reaction. The rate expression for this type of reaction (E2) can be quantitatively expressed as

$$\text{rate} = k[\text{C}_6\text{H}_{11}\text{Br}]^1[\text{NaOC}_2\text{H}_5]^1 \qquad 6\text{–}2$$

This expression indicates that the rate equation is *first order* in both the substrate (cyclohexyl bromide) and the basic reagent (sodium ethoxide). The overall rate expression is *second order* (the sum of 1 + 1 = 2).

NOTE. The powers of 1 are not usually expressed, but they are shown here for clarity.

Kinetics and Thermodynamics. Chemical reactions are governed by two basic relationships.

1. **Chemical Thermodynamics:** These phenomena deal with the changes in energy that occur when molecules react. Thus, thermodynamics ultimately controls the *extent* to which a reaction will go to completion.
2. **Chemical Kinetics:** As outlined above, kinetics concern the *speed* at which a reaction will go to completion.

All chemical reactions are reversible. When the reactants and products reach the stage where changes in concentration can no longer be measured, the reaction is considered to be in a **state of equilibrium.** If the equilibrium lies far to the side of product formation, the reaction is said to have essentially *gone to completion*. Thus, when the equilibrium constant, *K*, is large, the reaction is said to have a large *driving force.* When a reaction satisfies this latter condition, a certain amount of *energy* is necessarily released when equilibrium is reached.

In terms of thermodynamics, for a reaction to take place spontaneously, the *free energy*, ΔG, of the products must be lower than the free energy of the reactants; that is, ΔG must be negative. Free energy is made up of two components: enthalpy H (the difference in bond energies between reactants and products) and entropy S (energy related to the randomness of the system). These quantities are related by the expression

$$\Delta G = \Delta H - T\Delta S \qquad 6\text{–}3$$

In many reactions, however, it is the enthalpy that mainly determines whether the reaction can take place spontaneously.

A negative ΔG value it does not necessarily mean that the reaction will take place in a reasonable length of time (that is reach equilibrium). A negative ΔG value is necessary, but not sufficient, to guarantee a spontaneous reaction. The *free energy of activation,* $\Delta G^{\ddagger}$, must also be considered. This condition is illustrated in the potential energy diagram (Fig. 6.49) for a one-step reaction of molecule A with molecule B to form molecules C and D (with no reaction intermediate assumed). This energy profile is typical of the base-catalyzed dehydrohalogenation reaction being studied in the current experiment, the E2 elimination reaction.

The generalized reaction profile is as shown in Figure 6.49

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C} + \mathrm{D} \qquad 6\text{–}4$$

In the diagram the horizontal axis, called the reaction coordinate, indicates the progression of the reaction. $\Delta G^{\ddagger}$, is the free energy of activation for the forward reaction. The point at the top of the curve is called the *transition state*. Bond breaking and bond formation are involved over the range of this transition. Energy must be supplied through collision of molecules A and B (this involves both velocity and orientation in the collision) to move the system into and through the transition state and allow the products to form. As the products form, the activation energy is recovered plus additional energy (mainly from bond formation) so that the total energy evolved, $\Delta G_r^{\ddagger}$; is lower than that of the initial mixture.

The interpretation of the rate constant in terms of experimentally derived energy data is facilitated by the relationship known as the **Arrhenius equation** which rather successfully predicts the temperature dependence of the rate constants:

$$k = A \exp\left(-\frac{E_a}{RT}\right) \qquad 6\text{–}5$$

where

R is the gas constant with the value 8.314 J $\mathrm{mol}^{-1}\ \mathrm{K}^{-1}$

T is the temperature on the Kelvin scale

E_a is the Arrhenius activation energy in joules/mole

A is a constant (for a given reaction) called the *frequency factor*

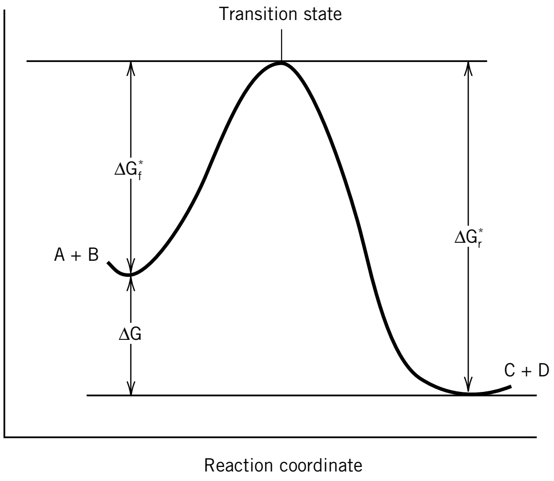

Figure 6.49 Activation energy.

The *frequency factor* is related to the frequency of collisions and the orientation of the reacting molecules. A indicates how many collisions have the correct orientation to lead to products. The rest of the equation, $e^{-E_a/RT}$, gives the fraction of collisions in which the energy of the reacting species is greater than the activation energy for the reaction.

Also note that we can rewrite equation 6–5 as $-E_a = RT \ln(k/A)$, which closely resembles the expression between the equilibrium constant and the Gibbs free energy change, $-\Delta G° = RT \ln k$, and underlines the treatment of E_a as an energy term.

Thus, the Arrhenius equation describes quantitatively the discussion given here. For two reactions *at the same temperature,* the reaction with the higher activation energy has the lower rate constant and thus the slower rate. In other words, there is a smaller fraction of molecules possessing sufficient energy to react. An *increase in temperature* results in more molecules having enough energy to react and thus a faster rate. A change in conditions in which the frequency factor A is increased (by an increase in collisions in which the orientation of the molecules is right for reaction) also results in an increase in rate.

In the present experiment, the temperature is held relatively constant at the boiling point of the reaction mixture. Remember that a boiling solvent has a constant temperature at its boiling point. This is critical in reaction rate investigations for, as noted above, temperature change can exhibit a profound effect on rate. This phenomenon will be investigated in the second half of the experiment.

Geometrical Considerations in the E2 Elimination Reaction. Base-induced elimination of alkyl halides (dehydrohalogenation) is a general reaction and is an excellent method for preparing alkenes. This process is often referred to as β-*elimination,* since a hydrogen atom is always removed β to the halide (leaving group):

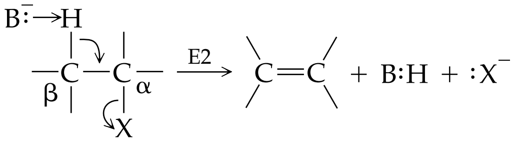

A high concentration of a strong base in a relatively nonpolar solvent is used to carry out the dehydrohalogenation reaction (see Experiment [10] for further examples).

Experimental evidence indicates that the five atoms involved in the E2 elimination reaction must lie in the same plane; the *antiperiplanar* conformation is preferred. This conformation is necessary for the orbital overlap that must occur for the π bond to be generated in the alkene. The sp^3-hybridized atomic orbitals on carbon that comprise the C—H and C—X σ bonds broken in the reaction develop into the p orbitals comprising the π bond of the alkene being formed. The reaction is termed *stereospecific.*

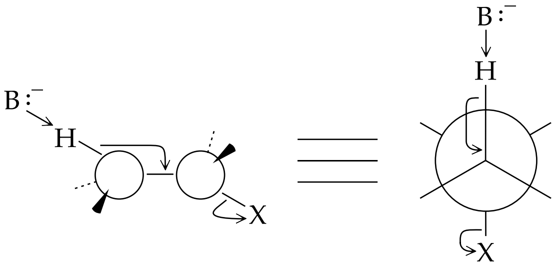

Anti-periplanar conformation

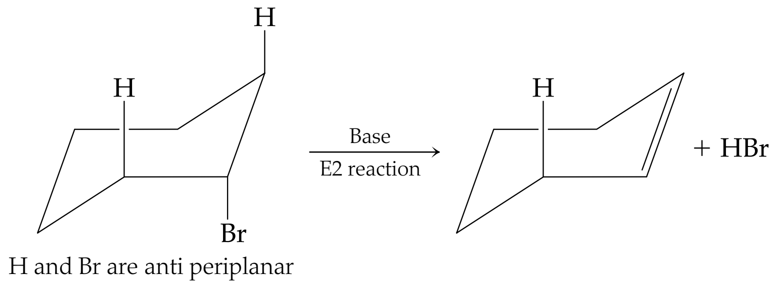

Figure 6.50 Geometric requirements for E2 elimination reaction in cyclohexanes.

The *antiperiplanar* conformation is particularly important in cyclohexane rings. This configuration can be met only if the hydrogen and the halide groups are *trans diaxial* to one another (see Fig 6.50). If either the hydrogen or halide group is *equatorial* the reaction does not occur. Furthermore, while there are exceptions to the relative disposition of the β-hydrogen and leaving group (Cope Elimination, for example), the number of systems which do not rely upon an antiperiplanar conformation are few and far between.

There is a smooth transition between reactant and product. Analogous to the S_N2 reaction, *no intermediate has been isolated or detected*. The reaction takes place in a single step through a high-energy *transition state* in which the double bond begins to form at the same time the hydrogen and halide groups leave. Figure 6.49 illustrates the energy profile for the reaction.

Evidence in support this mechanism is obtained from the measurement of the reaction kinetics of these reactions. Since both the alkyl halide and base concentrations appear to enter into the rate determining step, E2 reactions follow the second-order rate law:

$$\text{Rate} = k[\text{RX}]\,[\text{Base}] \qquad \textit{6–6}$$

As discussed above, a change in concentration of the base or of the alkyl halide should, therefore, affect the overall rate of the reaction. This dependence on concentration can be demonstrated by changing the base concentration. The current reaction, as it proceeds, will be sampled at 30-min intervals to track the generation of the alkene product, cyclohexene. Monitoring the formation of the product is carried out by removing a small aliquot sample from the reaction vessel and immediately quenching (stopping) the reaction with 1 M HCl. Rapid neutralization of the reacting medium (in the aliquot) with acid blocks the reaction. Each quenched sample is then extracted with methylene chloride to isolate the product and any unreacted starting material (cyclohexyl bromide). The rapid analysis of the small quantities of isolated reactants and products is accomplished by GC. Methylene chloride is effective as the extracting solvent, since it elutes from the GC column well before the compounds of interest (cyclohexene and cyclohexyl halide). From the chromatographic data obtained it is possible to determine if the amount of cyclohexyl halide has decreased and if the amount of cyclohexene has begun to increase. The results are reported as the percent cyclohexyl halide reacted.

EXPERIMENTAL PROCEDURE

Estimated time of experiment: two laboratory periods.

NOTE. It is recommended that the experiment be done in two parts. Use the first period to work out the procedure (equipment setup, sampling, GC analysis, etc.) and the second period to study the effect of several parameters on the rate of the reaction.

All drops are measured using a Pasteur pipet.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Cyclohexyl bromide	163.06	0.6 mL	4.87	166.2	1.3359	1.4957
Sodium ethoxide (21% in ethanol)		2 mL				

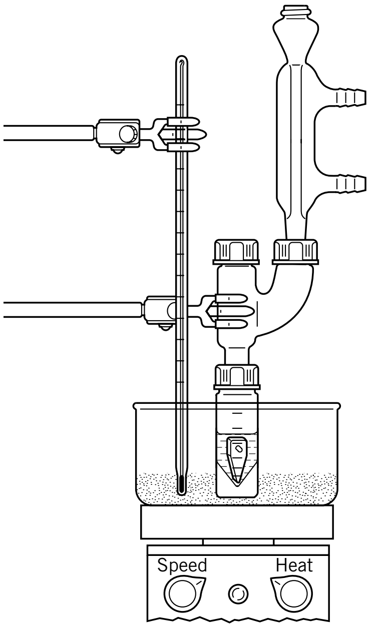

Cyclohexyl bromide, 0.6 mL Sodium ethoxide (21% in EtOH), 2.0 mL

Reagents and Equipment. Using a 10-mL graduated cylinder, measure 2 mL of a 21 wt% solution of sodium ethoxide in ethanol and add it to a 5-mL conical vial containing a magnetic spin vane (use a precalibrated Pasteur pipet for the transfer). Then equip the reaction vial with a Claisen head to which is connected a reflux condenser (⬅).

Reaction Conditions. Heat the reagent, with stirring, to reflux using a sand bath temperature of approximately 100 °C. Remove the cap from the Claisen head and use an automatic delivery pipet to add, in one portion, 0.6 mL (4.87 mmol) of cyclohexyl bromide. *Note the time of addition in your laboratory notebook.* Allow a few seconds for intimate mixing of the reactants to take place and then **immediately** remove 10 drops (about 0.4–0.5 mL) of the reaction solution using a Pasteur pipet, *recap the Claisen head,* and place the aliquot in a small sample vial containing 4–5 drops of 1 M HCl.

NOTE. You should notice that after a short time solid NaBr begins to form in the reaction vial.

To the sample vial containing the initial aliquot, add 10 drops of methylene chloride. Cap the vial and shake gently (or mix on a Vortex mixer). Loosen the cap carefully to vent the two-phase mixture. Remove the bottom, brown organic layer using a Pasteur filter pipet and transfer it to a second vial containing a small amount of anhydrous sodium sulfate. Cap the vial and label it as the *time-zero* sample.

*NOTE. For separation of the layers, it is recommended that **both** layers be drawn into the pipet. The bottom layer is then transferred to the second vial.*

Sampling Procedure. Two further samples for analysis are removed, *as outlined above*, one at 30 min into the reaction and the second after 1 h.

Each time the aliquot is treated

1. Remove 10 drops (0.4–0.5 mL) of reaction solution (recap the Claisen head).
2. Quench the reaction with acid (4–5 drops of 1 M HCl).
3. Extract the product and unreacted cyclohexyl bromide with methylene chloride, followed by drying the methylene chloride solution over anhydrous sodium sulfate.

NOTE. Having the 5 drops of 1 M HCl in each of three sample vials and sodium sulfate in each of three second sample vials (a total of six vials) before you add the cyclohexyl bromide to initiate the reaction will help you facilitate this sampling procedure. Carefully label all the vials!

At the end of the sampling period you should have three samples for analysis by GC: time-zero, 30 min, and 60 min.

Gas Chromatographic Analysis. The composition of the aliquot samples may be determined by gas chromatographic analysis.

Parameters for the GOW-MAC 350 Column

Packed Column:	DC 710 ($\frac{1}{4}$ *inch* × 8 *ft*).
Injection:	10 μL
Temperature:	132 °C
Flow rate (He):	50 mL/min
Chart speed:	1 cm/min

Gas chromatograph data may be used to determine the change in concentrations of cyclohexene and cyclohexyl bromide in the reacting medium during the 1-h reaction period. The samples can be injected after the extracted solution has had a 15- to 20-min period to dry over anhydrous sodium sulfate. *Each sample must elute from the GC before another sample is injected.* Each sample takes approximately 20 min to elute from the column, so start sample injections as soon as possible. For a typical chromatogram see Figure. 6.51.

INSTRUCTOR PREPARATION. For each chromatograph used, a mixture of cyclohexene and cyclohexyl bromide (approximately 80:30 by volume) should be injected once the parameters of the column are set. This is to establish the retention times of the components in the mixture.

NOTE. If the laboratory is done in two parts, it is advisable to stop at this point.

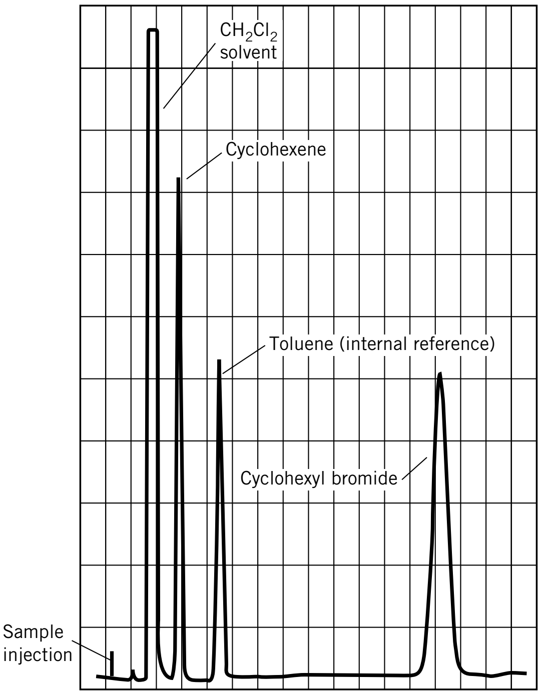

Figure 6.51 Example cyclohexene–cyclohexyl bromide GC.

DATA ANALYSIS

The percentage of each component (cyclohexene and cyclohexyl bromide) in the sample can be determined by measuring the area under the curves. How would you expect the areas under the cyclohexene peak and the cyclohexyl bromide peak to vary with time?

The boiling point of cyclohexene (82.9 °C) is close to that of the refluxing ethanol (78.2 °C). Therefore, when samples are removed from the reaction mixture some of the cyclohexene may be lost, since some of it will be in the vapor phase in the reaction vial. How would this loss affect the observed concentrations over the reaction period?

The *time-zero* fraction should be analyzed, since the reaction starts immediately upon addition of the cyclohexyl bromide to the basic solution.

NOTE. Several alternative techniques may be used to quantitatively determine the composition of the sample mixture from the GC data, if an integrating recorder is not available.

Two methods are described here:

1. Determination of the areas under the peaks gives reproducible results of ±3–4% when these areas are assumed equal to the peak heights (mm) × the peak widths at half-height (mm), measured from the baseline of the curve.

2. An alternative method for determinating these areas is to cut out the peaks from the chromatogram and weigh them on an analytical balance (sensitivity to 0.1 mg). The weights of the peaks are directly proportional to the relative amount of each compound in the sample (assuming that the detector is similar sensitivity to the two components).

The percentage of the individual components is then calculated by dividing each area by the total sum of the two areas and multiplying by 100.

Attach a copy of your chromatogram to your Data Analysis Sheet. Label the cyclohexene and cyclohexyl bromide peaks. Show the area measurements on the chromatogram and all the calculations in your report. What conclusions can you draw based on your results? Include an explanation of the conclusions drawn.

VARIATION OF PARAMETERS

The effect of varying several parameters that influence the *rate* of the reaction can also be investigated. Of particular interest for study are the effect of changes in the concentration of the base, the reaction temperature, and the nature of the leaving group.

INSTRUCTOR NOTE. It is suggested that the class be divided into groups, with each group assigned to investigate a different parameter. The results of each group are then shared by the entire class and each student then submits a report.

The parameters to be investigated are

1. A control (see conditions below).
2. Halving the base concentration.
3. Room temperature reaction.

4. Replacing cyclohexyl bromide with cyclohexyl chloride.
5. Replacing cyclohexyl bromide with cyclohexyl iodide.

Determine the percentage of unreacted cyclohexyl bromide in the sample mixture from the chromatogram for parameters 1, 2, and 3. Determine the percentage of cyclohexyl chloride and iodide in parameters 4 and 5, respectively. A comparison of unreacted starting halide with the control reaction shows the effect on the relative rate of dehydrohalogenation of changing the conditions under which the reaction is carried out. Tabulate your results.

In this set of experiments track only the disappearance of cyclohexyl bromide, and not the appearance of cyclohexene. As mentioned previously, the amount of cyclohexene observed in successive aliquots tends to drop below the predicted values because of the sampling techniques employed. As the reaction progresses, however, the amount of cyclohexyl bromide continues to decrease at the predicted rate (see below).

Use of an Internal Standard. How can we quantitatively track the disappearance of cyclohexyl bromide? As we have seen, the ratio of cyclohexyl bromide to cyclohexene cannot be determined accurately because of the inaccuracies in measuring the cyclohexene formed in the reaction. Also, there are small inaccuracies in the amount of the reaction mixture removed from the system for any given sample. These problems are solved by *standardizing* the samples. An *internal standard* is added to the reaction mixture. Ideally, this standard should not interfere with the reaction being studied and the substance must be easy to track by GC. In these experiments we will use *toluene*. Toluene is not very volatile (bp 110.6 °C) and has a retention time on the GC that does not interfere with the resolution of any other peaks on the chromatogram. It will be added *in a known amount* to the reaction mixture before you begin. For each gas chromatogram we can then measure *the area of the standard's peak* and use it to determine the amount of cyclohexyl bromide in the reaction mixture at that point.

Experimental Conditions

NOTE. Use the same experimental setup and workup procedure as in the previous experiment. A sample for analysis will be taken only at the 30-min mark. You may wish to take two samples so as to determine whether you can duplicate your results. The GC is setup at the same operating conditions. Drops are measured using a Pasteur pipet.

1. **Control**

Reaction conditions

2 mL of 21 wt% solution of sodium ethoxide in ethanol
300 μL toluene
Heat to reflux and stir.
Add 0.6 mL cyclohexyl bromide.

Time the reaction

At the 30-min mark, remove 20 drops of reaction solution and add it to 10 drops of 1 M HCl. Extract the mixture with 20 drops of methylene chloride and dry the methylene chloride layer over anhydrous sodium sulfate (15 min). Analyze the sample by GC.

2. Room Temperature Conditions

Reaction conditions

2 mL of 21 wt% solution of sodium ethoxide in ethanol
300 μL of toluene
Stir at room temperature **(do not heat the hot plate).**
Add 0.6 mL of cyclohexyl bromide.

Time the reaction

At the 30-min mark, remove 20 drops of reaction solution and add it to 10 drops of 1 M HCl. Extract the mixture with 20 drops of methylene chloride and dry the methylene chloride layer over anhydrous sodium sulfate (15 min). Analyze the sample by GC.

3. Variation of the Concentration of Base

Reaction conditions

1 mL of 21 wt% solution of sodium ethoxide in ethanol
1 mL of absolute (anhydrous) ethanol
300 μL of toluene
Heat to reflux and stir.
Add 0.6 mL cyclohexyl bromide.

Time the reaction

At the 30-min mark, remove 20 drops of reaction solution and add it to 10 drops of 1 M HCl. Extract the mixture with 20 drops of methylene chloride and dry the methylene chloride layer over anhydrous sodium sulfate (15 min). Analyze the sample by GC.

4. Variation of Substrate—Cyclohexyl Chloride

Reaction conditions

2 mL of 21 wt% solution of sodium ethoxide in ethanol
300 μL of toluene
Heat to reflux and stir.
Add 0.58 mL cyclohexyl chloride.

Time the reaction

At the 30-min mark, remove 20 drops of reaction solution and add it to 10 drops of 1 M HCl. Extract the mixture with 20 drops of methylene chloride and dry the methylene chloride layer over anhydrous sodium sulfate (15 min). Analyze the sample by GC.

5. Variation of Substrate—Cyclohexyl Iodide

Reaction conditions

2 mL of 21 wt% solution of sodium ethoxide in ethanol
300 μL of toluene
Heat to reflux and stir.
Add 0.63 mL of cyclohexyl iodide.

Time the reaction

At the 30-min mark, remove 20 drops of reaction solution and add it to 10 drops of 1 M HCl. Extract the mixture with 20 drops of methylene chloride and dry the methylene chloride layer over anhydrous sodium sulfate. Analyze the sample by GC.

Data Analysis

To calculate the % of cyclohexyl bromine (chloride or iodide) reacted, follow the outline given here.

Given:

a. The amount of toluene added as the internal standard:

259 mg

Calculation (show here):

______________________ × __________
Volume of toluene used × its density

b. The amount of cyclohexyl bromide used in the reaction:

794 mg

Calculation (show here):

________________________________ × __________
Volume of cyclohexyl bromide used × its density

If cyclohexyl chloride or iodide is used in place of the bromide:
Chloride: 581 mg
Iodide: 1029 mg

NOTE. The amount of cyclohexyl chloride and iodide used is the same molar amount as the cyclohexyl bromide.

The Calculations Using an Internal Standard. The calculations allow us to compare the amount (%) of cyclohexyl bromide (or chloride or iodide) reacted at the 30-min time point for each of the different reaction parameters.

Step 1. The amount of cyclohexyl bromide (chloride or iodide) **left** (unreacted) at the 30-min mark:

$$\frac{\text{Area of } C_6H_5\text{—Br, —Cl, or —I peaks}}{\text{Area of toluene peak}} \times [259 \text{ mg of toluene standard}] = \underline{\qquad} \text{ mg unreacted}$$

Step 2. The amount of cyclohexyl bromide (chloride or iodide) **reacted** at the 30-min mark:

[Amount of C_6H_5—Br, —Cl, or —I used in the reaction (in mg)] −
[the amount left after 30 minutes (in mg), see step 1] = ______ mg reacted

Step 3. The percent of cyclohexyl bromide (chloride or iodide) **reacted** at the 30-min mark:

$$\frac{\text{Amount reacted (in mg)}}{\text{Amount used (in mg)}} \times 100 = \underline{\qquad\qquad} \text{ \% reacted}$$

Attach a copy of your chromatogram to your Data Analysis Sheet. Indicate which parameter you worked on. Label the toluene and cyclohexyl bromide, chloride, or iodide peaks. Show the area measurements on the chromatogram. Show all the above calculations in your report. Draw conclusions as to the effect on the rate of the reaction for each of the parameters investigated in the laboratory. Include an explanation of the conclusions drawn.

QUESTIONS

6-220. What reaction occurs between $NaOCH_2CH_3$ and HCl that stops the dehydrohalogenation reaction?

$$NaOCH_2CH_3 + HCl \rightarrow$$

6-221. Would you expect the replacement of cyclohexyl bromide with cyclohexyl iodide to result in a decrease, increase, or no change in the rate of the reaction? Explain.

6-222. In the second part of the experiment, why was it necessary to use toluene as an internal standard? What are the requirements for an internal standard? How would the loss of some toluene in the sampling procedure (from over heating the reaction system) affect the data?

6-223. A negative ΔG value does not necessarily mean that a reaction will take place in a reasonable length of time. A negative ΔG value is necessary, but not sufficient to guarantee a spontaneous reaction. The *free energy of activation,* $\Delta G^{\ddagger}$, must be considered. Explain this statement.

6-224. In E2 elimination reactions involving cyclohexane rings, why is it critical that the β-hydrogen and the leaving group be *trans diaxial* to one another?

6-225 Please explain why *trans*-1-bromo-4-methycyclohexane is more likely to undergo an E2 reaction when compared to *trans*-1-bromo-4-*tert*-butylcyclohexane even when working with warm toluene.

6-226. What effect on the rate of reaction would be observed if some of the ethanol solvent escaped from the reaction flask (as volatile vapor) while samples are being removed? Explain.

BIBLIOGRAPHY

For a discussion of transition states, see Laidler, K. J. *J. Chem. Educ.* **1988,** *65,* 540.

For texts on Chemical Kinetics see, for example (a) Laidler, K. J. *Chemical Kinetics;* 3rd ed.; Benjamin/Cummings: Menlo Park, CA, 1987; (b) Logan, S. R. *Fundamentals of Chemical Kinetics;* Addison-Wesley Longman: Menlo Park, CA, 1996; and (c) House, J. *Principles of Chemical Kinetics;* 2nd ed.; WCB/McGraw-Hill: Burr Ridge, IL, 2007.

Physical Chemistry texts generally present a kinetic analysis of equilibria. For example, see

(a) Atkins, P. W.; DePaula, J. *Physical Chemistry,* Vol. 1, 8th ed.; Freeman: New York, 2006. (b) Laidler, K. J.; Meiser, J. H.; Sanebuary, B.C. *Physical Chemistry,* 4th ed.; Houghton Mifflin: Boston, 2003; (c) Silbey, R. J.; Alberty, R. A.; Bawendi, M. G. *Physical Chemistry,* 4th ed.; Wiley: New York, 1997; and (d) Engle, T.;

Reid, P. *Physical Chemistry;* Pearson Benjamin Cummings: San Francisco, 2006.

For an introduction to the E2 mechanism see your lecture text or Smith, M. B.; March. J. *Advanced Organic Chemistry,* 6th ed.; Wiley: New York, 2007, Chap. 17, p. 1478. or Carey, F. A.; Sundberg, R. J. *Advanced Organic Chemistry,* 4th ed.; Plenum: New York, 2008, Chapter 6.

C_7H_6apter 7

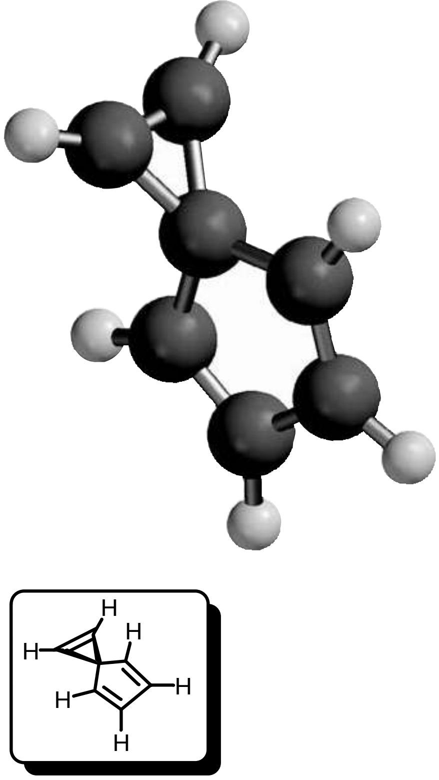

SEQUENTIAL SYNTHESES: THE TRANSITION FROM MACRO TO MICRO

The synthesis of a vast array—now numbering in the millions—of new organic molecules in academic and industrial laboratories over the past 100 years is one of the great achievements of modern science. Many of these new compounds have had profound effects on our way of life, both good and bad. A great challenge in the next century will be how society applies these powerful materials, and the new molecules yet to be born, to the common good.

Our ability to synthesize highly complex organic substances has taken a number of dramatic jumps during this century, and has resulted in a bewildering collection of substances that have been devised, synthesized, and applied to practically every facet of our lives. Many of these materials are now vital to our daily life (consider penicillin) and we all too often take them for granted. In just the last 30–40 years, new advances in pharmaceutical compounds have saved, extended, and made more comfortable the lives of hundreds of millions of people. The list could go on and on, including textiles, surfactants, plastics, and synthetic oils, to name only a few.

Historically, the synthesis of complex organic substances was primarily driven by the need to obtain large quantities of biologically active material that occurred as the product of plant or animal metabolism, but that could be obtained only in very small quantities from nature. For example, the synthesis of the adrenal cortex hormone, cortisone, was a major breakthrough for hormone therapy. The synthesis of this material initially required 33 steps. That is, the research chemist carried out a sequence of 33 reactions in which stable isolable intermediates were formed sequentially, leading ultimately to the desired cortisone molecule. Industrial sequences of this length are now rare, but those requiring three to six steps are common.

The strategy of the synthetic chemist is to devise a route whereby the desired compound can be prepared efficiently, using inexpensive, readily available starting materials in the fewest steps, and have a minimal impact on the environment when considering the waste(s) generated. For each individual step, the yield of intermediate should be as high as possible with a minimum of side reactions. In industry the overall cost of the proposed synthetic sequence must be considered, including the time involved, type of equipment required, and safety factors. Today, with the worldwide demand for organic materials in vast quantities (e.g., petroleum products) it is becoming increasingly important to assess the impact on our environment of these synthetic materials prior to large-scale production.

Chapter 7: C_7H_6, [1,2]Spirene
Simmons and Fukunaga (1967).

Because organic transformations almost always take place with some loss of material (similar bond energies lead to alternative reaction pathways and easy byproduct formation), the yield of intermediate from each individual step can have a significant impact on the overall yield of the final product. *In a multistep synthesis, the overall yield is the mathematical product of the yields of the individual steps.* For example, if we assume that in a five-step sequence for the preparation of a new dye, each step takes place in 85% yield, the overall yield would be $(0.85)^5 \times 100$ or 44% (in a 33-step synthesis with an 85% yield for every step, the overall yield would be ~0.5%). This property of organic synthesis emphasizes why a synthetic route devised to produce a particular molecular structure must be carefully planned to minimize losses at each stage of the chosen pathway. This problem also illustrates why the initial steps of a sequence usually use larger quantities of reactants (macro or semimicro quantities), and why in the last stages, experience at running reactions at the microscale level can be invaluable.

In this chapter we describe a set of six sequential experiments. These experiments vary in the number of intermediates that are required from seven to three, and they vary in the complexity of the chemistry, from straightforward extensions of Chapter 6 to relatively demanding experimentation similar to that described in Chapter 10W.

The target molecules include

- The drug **sulfanilamide** (the first of the antibiotics), which is obtained in a novel three-step sequence not usually found in the introductory laboratory:

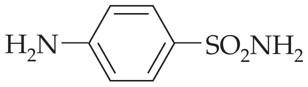

p-Aminobenzenesulfonamide
(sulfanilamide)

- The industrially important polymer, **nylon-6,6** (the first of the commercial synthetic textile fibers), formed in three steps closely paralleling the original synthesis:

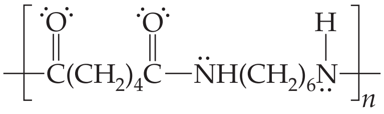

$$\left[\overset{\ddot{O}:}{\overset{\|}{C}}(CH_2)_4\overset{\ddot{O}:}{\overset{\|}{C}}-\ddot{N}H(CH_2)_6\overset{H}{\overset{|}{\underset{..}{N}}}\right]_n$$

Nylon-6,6

- The synthesis of **2′-bromostyrene,** which requires three steps. This compound is an interesting substance because of its commercial use as a fragrance in soap products:

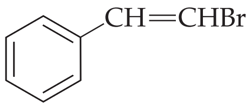

2′-Bromostyrene

- The synthesis of **piperonylonitrile** is an example of a novel conversion, in three steps, of an aromatic aldehyde to an aromatic nitrile:

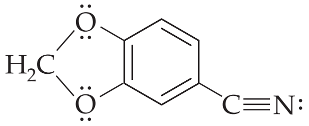

Piperonylonitrile

- The synthesis of **hexaphenylbenzene** requires two parallel three-step sequences to obtain two intermediates that then react with each other to give the target compound after a seven-step synthesis. The final product is a most unusual organic material with one of the highest melting points observed for an organic substance:

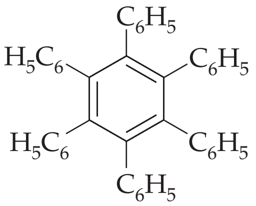

Hexaphenylbenzene

- The synthesis of a **photochromic imine** in four steps yields perhaps the most intriguing substance in the entire chapter. The ability of this material to turn a deep-blue color when exposed to light, and then to lose its color when placed back in the dark makes this structure one of the most interesting of the sequential products. It also involves the most challenging chemistry of the multistep syntheses:

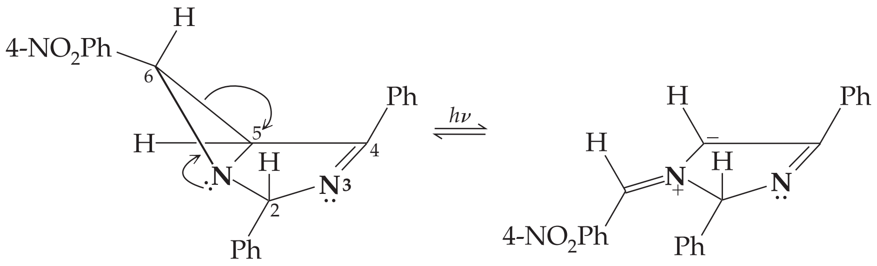

2-*exo*-6-*exo*-2,4-Diphenyl-
6-(4-nitrophenyl)-
1,3-diazabicyclo[3.1.0]hex-3-ene

Blue azomethine ylide
(syn configuration)

In a number of the sequences, the stereochemistry of the reactions is vitally important and controlled by the mechanisms that are operating under the prescribed conditions. This aspect of the transformations is discussed in detail. By performing one or all six of these multistep sequences you will have a chance to challenge your experimental technique under conditions essentially identical to those found in the modern synthetic organic research laboratory. You will quickly recognize the reason why laboratory technique is so vital to the success of multistep syntheses.

SEQUENCE A

The Synthesis of Hexaphenylbenzene

The molecule to be prepared in the Sequence A synthesis is hexaphenylbenzene (**I**):

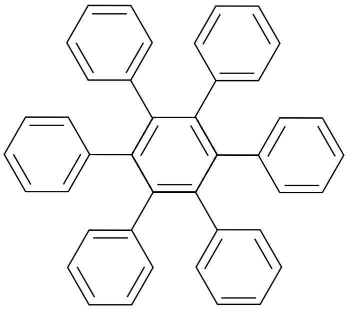

I
Hexaphenylbenzene

This system, which contains seven aromatic rings, was first made by Dilthey at the University of Bonn in 1933 by using a classic Diels–Alder reaction with exactly the same two reactants as you will generate in Sequence A. The Bonn group showed that an earlier claim, by Durand, to have prepared this compound via a massive Grignard attack by phenylmagnesium bromide on hexachlorobenzene, had not actually yielded hexaphenylbenzene. The compound isolated by Durand melted at 266 °C while Dilthey's material melted at 421–422 °C. Hexaphenylbenzene was later synthesized photochemically by Büchi at the Massachusetts Institute of Technology (MIT) in 1962. The MIT group improved the purity of the isolated material, and reported a melting point of 439–441 °C. Fieser, at Harvard University, refined Dilthey's synthetic route, and published the definitive preparation in *Organic Syntheses* in 1966. Fieser obtained melting points *without decomposition* in evacuated melting point capillaries (see Chapter 4) in the range 454–456 °C.

This hydrocarbon system possesses a number of interesting structural and physical properties. First, we should note that it has a relatively high molecular weight, near 534, and a molecular formula of $C_{42}H_{30}$. Hexaphenylbenzene exhibits an extremely high melting point for a nonionic organic material. For example, of the 15,000 plus substances listed in the Table of Physical Constants for Organic Compounds in the *CRC Handbook of Physics and Chemistry,* only two materials melt above 450 °C (and both of these compounds decompose at their melting point), and only 10 melt above 400 °C. Indeed, hexaphenylbenzene melts above hexabenzobenzene (**II**), the completely delocalized seven-ring fused system, mp = 438–440 °C.

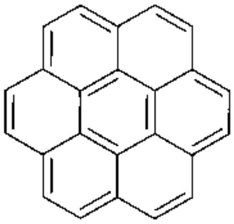

II
Hexabenzobenzene
(coronene)

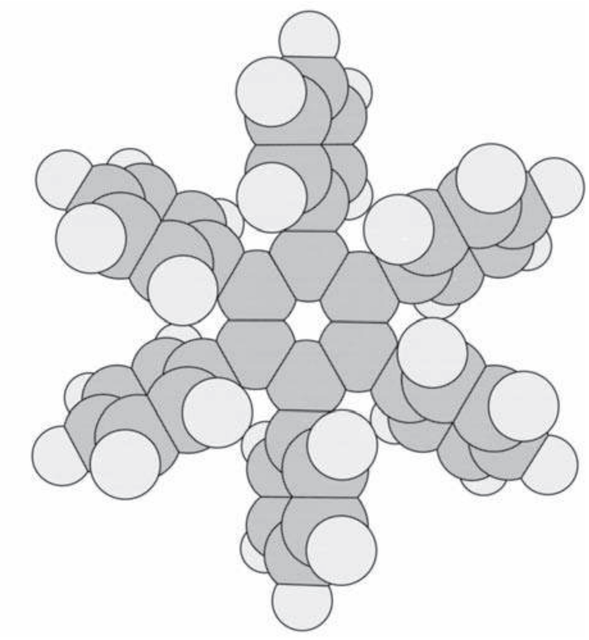

Figure 7.1 Molecular model of one of the chiral rotamers of hexaphenylbenzene.

From molecular modeling studies of hexaphenylbenzene, it appears that the six substituent rings surrounding the central system will be sterically restricted from lying in the plane of that ring by ortho-position interaction (**III**) and Figure 7.1 shows a molecular model of one of the chiral rotamers of hexaphenylbenzene.

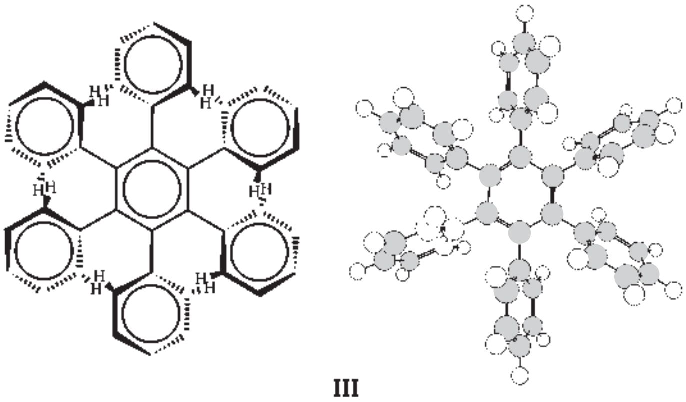

III
Hexaphenylbenzene with steric interactions between ortho positions

The twisted structure presents a particularly interesting problem in stereoisomerism. It is clear that when all the rings are coplanar (dihedral angle of 0°) we have maximum overlap of the π system and delocalization energy, but we also have a maximum of steric repulsion energies. On the other hand, when the dihedral angle approaches 90°, all delocalization is blocked, though steric repulsion between rings is at a minimum. It would seen reasonable to expect the system to reach some energetic compromise between these two extreme orientations. If this is the case, is it possible to establish the angle at which the external rings are positioned? An X-ray crystallographic study carried out by Bart in 1968 on solid crystalline hexaphenyl-benzene did, in fact, show that in the crystal lattice the phenyl groups are twisted 65° out of the plane of the central ring. In the crystal lattice, the molecule adopts a conformation, similar to a six-bladed propeller, which is chiral. That is, in the solid state hexaphenylbenzene can exist in two enantiomeric forms. Indeed, if this molecule happened to undergo resolution of the optical isomers during crystallization, in much the same fashion as Pasteur's tartaric acid salts, it should be possible to mechanically separate the racemate into crystals in which all the rings are tilted only in one possible conformation or in the other. These enantiomers, however, should possess a relatively low barrier to rotation, so that when dissolved in solution rapid racemization, via rotation of the rings (propeller blades) to the opposite pitch, would be expected.

This does seem to be the case. In 1977, Gust at Arizona State University showed that if a derivative of hexaphenylbenzene were synthesized in which two different groups were substituted on adjacent rings in the ortho positions (e.g., a methyl group and a methoxy group, **IV**), two possible sets of diastereomers would result:

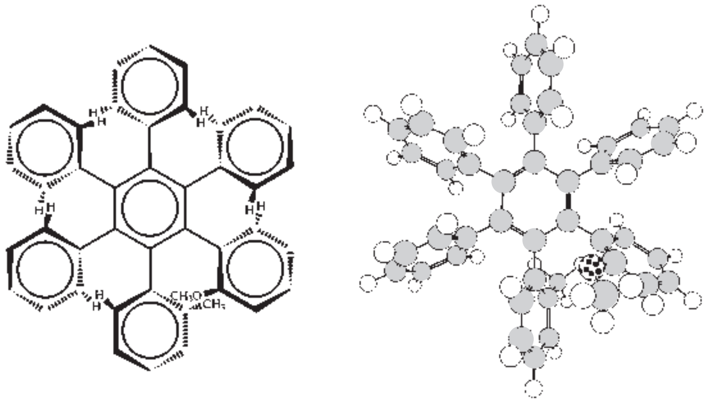

IV
Substituted hexaphenylbenzene with steric interactions between ortho position

It is presumed that it would be impossible for the rings to rotate the two ortho substituents past one another, but that other rotations may or may not be facile. If a large rotational barrier were present, the external rings would remain tilted at 65° with the same pitch. If this were the case, we would expect four diastereomers, and thus four different C—CH_3 resonances in the ^1H NMR spectrum. If rapid interconversion of the tilted conformers occurred, we would expect that the two bulky ortho groups would restrict full rotation of those two rings, even though the barrier to pitch inversion is low. Thus, in this latter case we would expect two diastereomeric pairs of enantiomers (one with the two ortho groups up and one with one up and one down), and two different C—CH_3 resonances in the NMR. When the compound was synthesized, and its ^1H NMR spectrum obtained, two resonances for methyl groups attached to aromatic rings and two O—CH_3 resonances were observed. These two diastereomers, **V**, were separated by column chromatography, and it was found that they slowly interconverted upon being heated to 215 °C. Thus, on the NMR time scale, it would appear that in hexaphenylbenzene, a reasonably rapid inter-conversion of the propeller conformations is taking place in solution at room temperature:

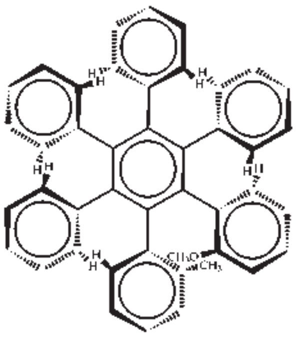

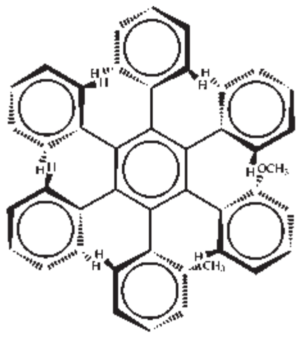

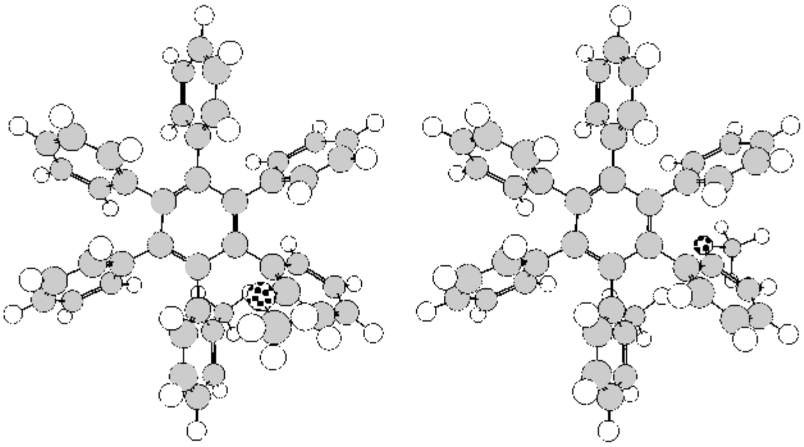

V

ortho-Disubstituted hexaphenylbenzene exists as two diastereomers

The Synthesis of Hexaphenylbenzene from Benzaldehyde:

EXPERIMENTS A1_a, A2_a, A3_a, A1_b, A2_b, A3_b, and A4_{ab}

As shown on the flow charts below and on page 428, the total synthesis of hexaphenylbenzene involves two parallel sets of three reactions each, the **a** series and the **b** series. The two series culminate in synthesizing tetraphenylcyclopentadienone (**a** series) and diphenylacetylene (**b** series), which are then reacted together (Experiment [A4_{ab}]) to produce the final product, hexaphenylbenzene.

The **a** series uses benzaldehyde as a starting material, which is first converted to the α-hydroxyketone benzoin in Experiment [A1_a]. Benzoin is then oxidized to benzil (Experiment [A2_a]). Benzil (along with diphenylacetone) is used in the construction of tetraphenylcyclopentadienone in Experiment [A3_a], the third and last of the Sequence A_a intermediates.

The **b** series of synthetic experiments also begins with benzaldehyde, which is converted in Experiment [A1_b] into (*E*)-stilbene. (*E*)-Stilbene is the precursor to *meso*-1,2-dibromo-1,2-diphenylethane (*meso*-stilbene dibromide) prepared in Experiment [A2_b]. This dibromide is in turn converted

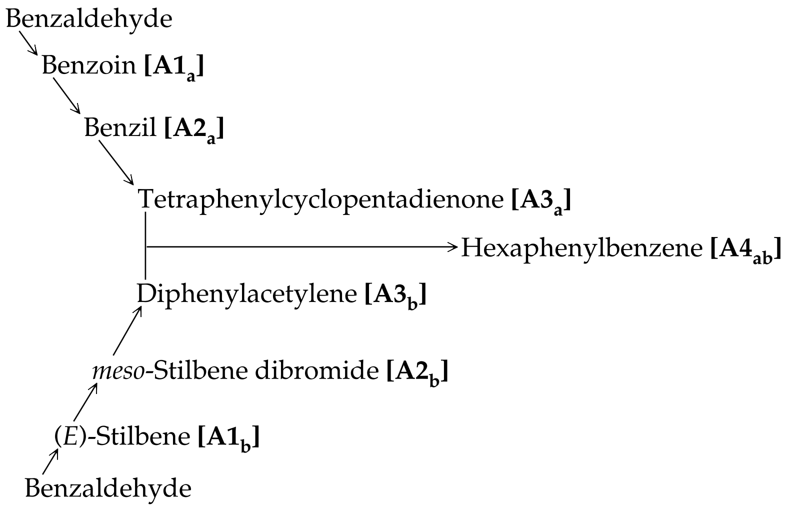
Benzaldehyde
Benzoin [A1_a]
Benzil [A2_a]
Tetraphenylcyclopentadienone [A3_a]
Hexaphenylbenzene [A4_{ab}]
Diphenylacetylene [A3_b]
meso-Stilbene dibromide [A2_b]
(*E*)-Stilbene [A1_b]
Benzaldehyde

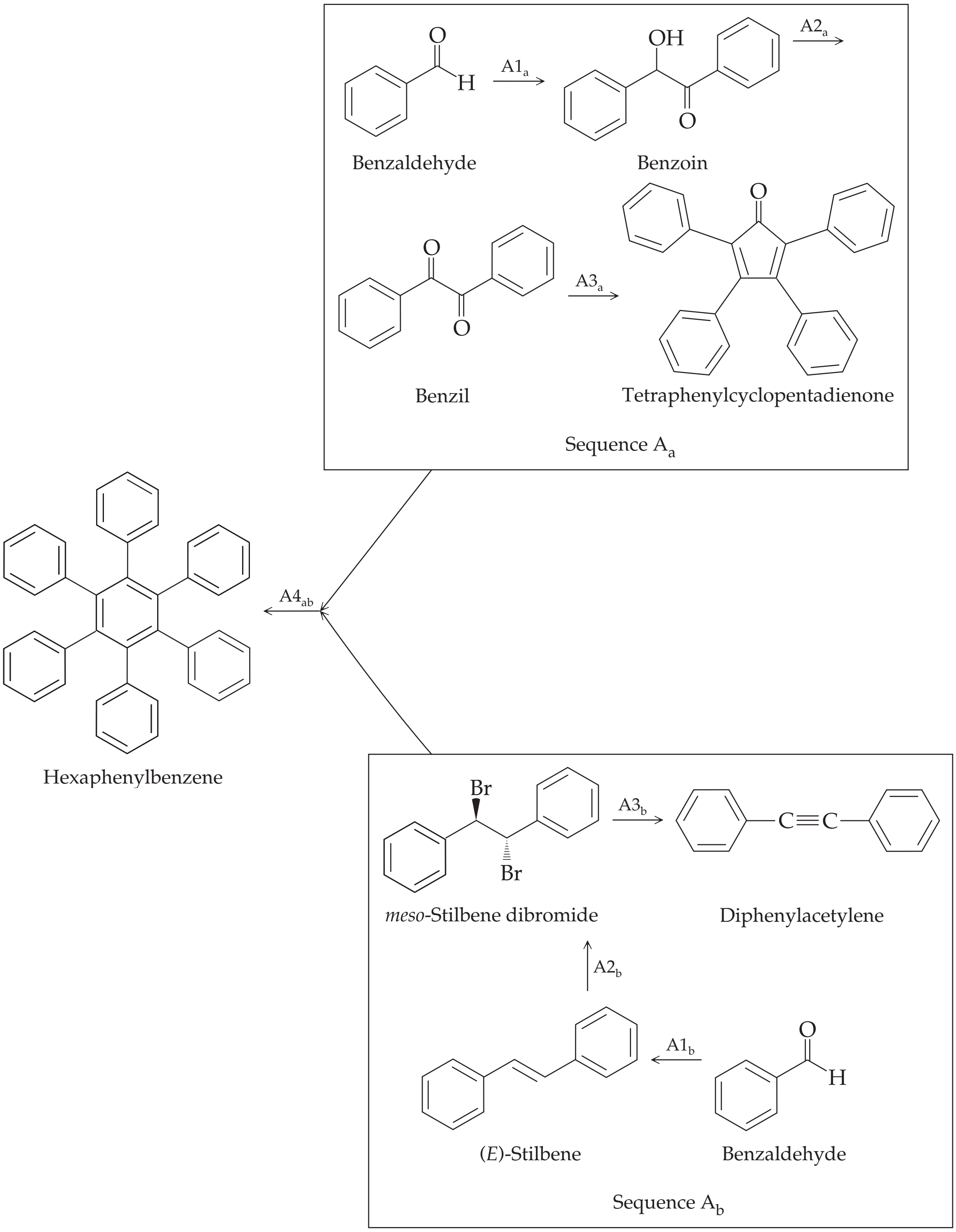
O
H
A1a
OH
O
A2a
Benzaldehyde
Benzoin
O
O
A3a
O
Benzil
Tetraphenylcyclopentadienone
Sequence Aa
A4ab
Hexaphenylbenzene
Br
Br
A3b
C≡C
meso-Stilbene dibromide
Diphenylacetylene
A2b
A1b
O
H
(E)-Stilbene
Benzaldehyde
Sequence Ab

by the double dehydrohalogenation reaction in Experiment [A3_b] into diphenylacetylene.

The diphenylacetylene is then reacted with the tetraphenylcyclopentadiene, synthesized in the last step of the **a** series, in a Diels–Alder reaction to produce the final product of the sequences, hexaphenylbenzene, in Experiment [A4_{ab}].

This preparation of hexaphenylbenzene demonstrates the manner in which a variety of basic organic reactions can be integrated to prepare a desired end product.

BIBLIOGRAPHY

Bart, J. C. *J. Acta Crystallogr., Sect. B* **1968,** *24,* 1277.

Büchi, G.; Perry, C. W.; Robb, E. W. *J. Org. Chem.* **1962,** *27,* 4106.

Dilthy, W.; Schommer, W.; Trosken, O. *Berichte* **1933,** *66B,* 1627.

Durand, J. F.; Hsun, L. W. *C. R. Hebd. Seances Acad. Sci.* **1931,** *191,* 1460.

Fieser, L. F. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 604.

Gust, D. *J. Am. Chem. Soc.* **1977,** *99,* 6980.

The Benzoin Condensation of Benzaldehyde: Benzoin

Experiment A1$_a$

Common name: benzoin
CA number: [579-44-2]
CA name as indexed: ethanone, 2-hydroxy-1,2-diphenyl-

Purpose. One of the classic reactions of organic chemistry, the *benzoin condensation,* is carried out. You will examine the properties of the α-hydroxyketone product of this well-known reaction. The particular α-hydroxyketone generated in this experiment is the compound from which the reaction gains its name, *benzoin*.

This reaction provides quantities of benzoin for use in the multistep synthesis of hexaphenylbenzene (see Experiment [A4$_{ab}$]). Benzoin is synthesized in this first step of the **a** series of the Sequential Experiments. In this sequence of reactions, benzoin is converted by oxidation (Experiment [A2$_a$]) to benzil and then to tetraphenylcyclopentadienone (Experiment [A3$_a$]). The latter compound undergoes a Diels–Alder addition with diphenylacetylene (Experiment [A3$_b$]) to give hexaphenylbenzene (Experiment [A4$_{ab}$]).

NOTE. If the benzoin product is to be used in the reaction sequence, it is recommended that one of the semimicro procedures be followed so that sufficient material will be available for the subsequent steps. The conditions for a one-step microscale experiment are listed following the two semimicroscale procedures.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallizations (pp. 89–91)

REACTION

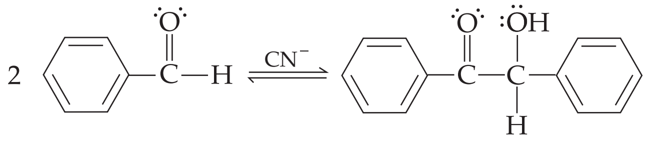

Benzaldehyde Benzoin

DISCUSSION

Aromatic aldehydes, in the presence of catalytic cyanide ion, dimerize to form the corresponding α-hydroxyketone (acyloin). This reaction, which is reversible, is known as the benzoin condensation, even though we now know that it is not actually a condensation reaction, because no water or alcohol is produced. Cyanide ion is a specific catalyst for the reaction with aromatic aldehydes, and can function in this capacity because it is a good nucleophile, it stabilizes the intermediate carbanion, and it is a good leaving group. In the mechanism outlined below, the nucleophilic cyanide ion attacks a molecule of the aromatic aldehyde to form the conjugate base of a cyanohydrin. The effect of the —CN group is to increase the acidity of the α-hydrogen atom, thus allowing the formation of the anion (**I**):

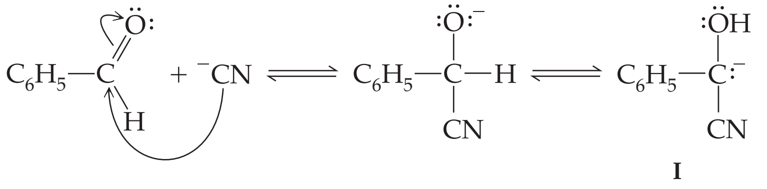

(1)

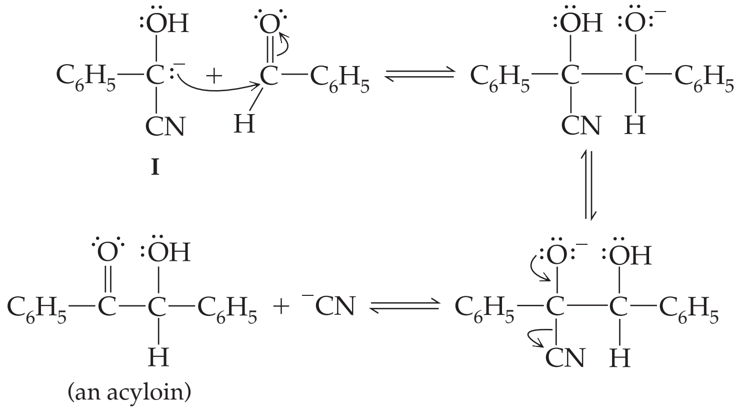

(2)

Once generated, the nucleophilic carbanion (**I**) attacks a second molecule of the aromatic aldehyde to yield a substituted cyanohydrin. This species can then be stabilized by loss of cyanide ion to form the α-hydroxyketone product.

The electronic effects of various substituents on the aromatic ring have been investigated. Because, in this reaction, the same aldehyde functions as both the nucleophile and the electrophile, electronic effects that enhance one of these functions are likely to inhibit the other. If a strongly electron-donating group is in the para position of the ring, the reaction fails due to the increase in electron density at the carbonyl carbon, brought about by the presence of the electron donor, which renders the carbonyl carbon less electrophilic. The benzoin condensation is also hindered by strong electron-withdrawing groups on the ring. The presence of a para-nitro group decreases the electron density on the carbonyl carbon atom in the cyanohydrin anion, making its carbanion less nucleophilic, which drastically retards the rate of addition of the anion to the second molecule of aldehyde.

The cyanide-ion catalysis works only for aromatic aldehydes. Aliphatic aldehydes can, however, be condensed to α-hydroxyketone in the presence of thiazolium salts, such as *N*-dodecylthiazolium bromide:

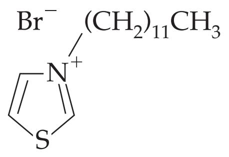

N-Dodecylthiazolium bromide

SEMIMICROSCALE EXPERIMENTAL PROCEDURE

(The microscale reaction quantities are increased by a factor of 2.)

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants						
Compound	**MW**	**Amount**	**mmol**	**bp (°C)**	***d***	**n_D**
Benzaldehyde	106.13	400 µL	3.92	178	1.04	1.5450
Sodium cyanide (0.54 M) in ethanol (95%)		2 mL				

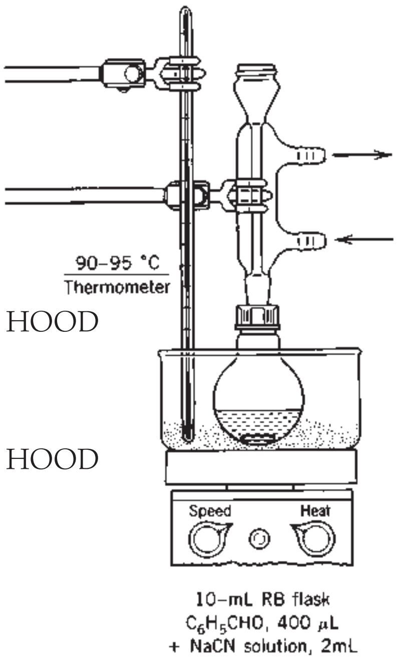

HOOD

HOOD

Reagents and Equipment. Using an automatic delivery pipet in the **hood,** place 400 µL (416 mg, 3.92 mmol) of fresh, acid-free, benzaldehyde in a weighed 10-mL round-bottom flask containing a magnetic spin bar. Reweigh the flask and record the weight. Now add 2 mL of a 0.54 M solution of sodium cyanide in ethanol (in the **hood**). Remember to use a fresh tip on the automatic delivery pipet.

NOTE. The 0.54 M NaCN solution should be prepared by the instructor.

CAUTION: Sodium cyanide (NaCN) is extremely toxic.

Attach the flask to a reflux condenser and mount the assembly in a sand bath on a magnetic stirring hot plate (➡).

Reaction Conditions. Heat the mixture with stirring in a sand bath at 90–95 °C. Maintain this temperature for 30 min. The reaction solution turns yellow, and may then become cloudy within approximately 5 min.

NOTE. Do not overheat the reaction mixture. If the solution begins to darken, immediately remove the vial from the sand bath.

Isolation of Product. At the end of the reflux period, cool the solution to room temperature and then place it in an ice bath for 10 min. Collect the benzoin product by filtration under reduced pressure using a Hirsch funnel (➡). Wash the filter cake on the Hirsch filter bed with two 1-mL portions of cold water, and air-dry the material under suction using plastic food wrap (see Prior Reading) for 5 min. The crude material is further dried on a porous clay plate or on filter paper.

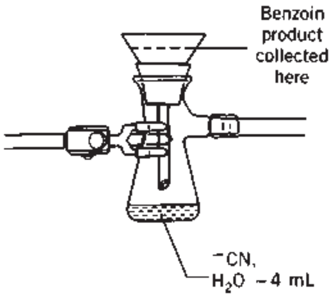

Purification and Characterization. This crude material may be purified by recrystallization from methanol or ethanol (95%) in a Craig tube.

Weigh the benzoin product and calculate the percent yield. Determine the melting point and compare your value to that found in the literature.

Obtain the IR spectrum of the compound. Compare your spectrum to that recorded in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

This compound also has a characteristic ultraviolet spectrum showing a peak of maximum absorption (λ_{max}) at 247 nm (ε_{max} = 13,200, at a concentration of 6.0×10^{-5} M in ethanol) characteristic of the benzoyl group, $C_6H_5C{=}O$.

Chemical Tests. Benzoin contains an aromatic ring. Confirm this fact by performing the ignition test (Chapter 9). To confirm the presence of the alcohol and ketone functions in benzoin carry out the chromic anhydride test for the —OH group and the 2,4-dinitrophenylhydrazine test for the C═O group. Isolate the solid 2,4-dinitrophenylhydrazone derivative and compare its melting point to the literature value.

There is a specific test for the presence of benzoin. Place a few crystals of your material in 800 μL of 95% ethanol. The addition of a few drops of 10% sodium hydroxide solution produces a purple coloration. The color fades when the solution is shaken in air but reappears if the solution is allowed to stand.

OPTIONAL SCALES

These procedures are identical to that given above with the following exceptions:

Fivefold Scaleup

1. Increase the scale of the reaction by a factor of 5 compared to the microscale procedure.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Benzaldehyde	106.13	1.0 mL	9.8	178	1.04	1.5450
Sodium cyanide (0.54 M) in ethanol (95%)		5.0 mL				

2. Use a 10-mL round-bottom flask.

3. The product is washed with two 2-mL portions of cold water.

MICROSCALE REACTION PROCEDURE

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Benzaldehyde	106.13	200 μL	1.96	178	1.04	1.5450
Sodium cyanide (0.54 M) in ethanol (95%)		1.0 mL				

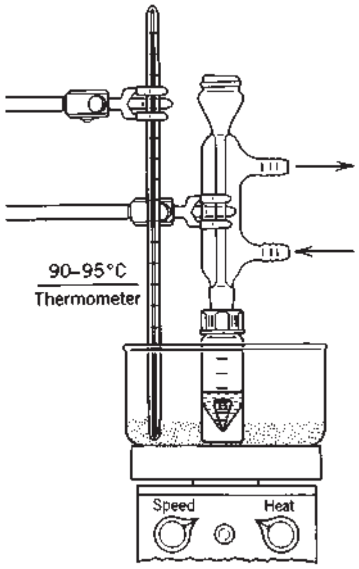

C_6H_5CHO, 200 μL
+ NaCN solution, 1 mL

1. The product is washed with two 0.5-mL portions of cold water.

2. Use a 5-mL conical vial (⬅).

QUESTIONS

7-1. The benzoin produced in this experiment contains a chiral carbon atom (a stereocenter), but the product itself is not optically active. Explain.

7-2. The cyanide ion is a highly specific catalyst for the benzoin condensation. Can you list three functions this ion performs in this catalytic role?

7-3. Can you suggest a reason why *p*-cyanobenzaldehyde does not undergo the benzoin condensation to yield a symmetrical benzoin product?

7-4. Tollens' reagent is used as a qualitative test for the presence of the aldehyde functional group (see Chapter 9, Classification Tests). However, benzoin, which does not contain an aldehyde, gives a positive test with this reagent. Explain.

7-5. Show how one might accomplish the conversion of *p*-methylbenzaldehyde to each of the following compounds.

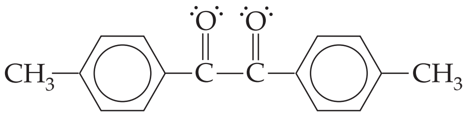

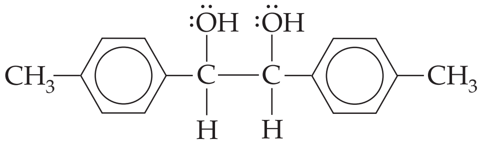

BIBLIOGRAPHY

Adams, R.; Marvel, C. S. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 88.

Hassner, A.; Rai, K. M. L. *Comp. Org. Syn.* **1991,** *1,* 541.

Ide, W. S.; Buck, J. S. *Org. React.* **1948,** *4,* 269.

Kuebrich, I. P.; Schowen, R. L.; Wang, M.; Lupes, M. E. *J. Am. Chem. Soc.* **1971,** *93,* 1214.

Stetter, H.; Kuhlmann, H. *Organic Syntheses;* Wiley: New York, 1990; Collect. Vol. VII, p. 95.

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Wiley: New York, 2007, Chap. 16, p. 1396 and references therein.

For the use of thiamine hydrochloride as a catalyst in the reaction see

Pasto, D. J.; Johnson, C. R.; Miller, M. J. *Experiments and Techniques in Organic Chemistry;* Prentice-Hall: New Jersey, 1992, p. 500.

Copper(II) Ion Oxidation of Benzoin: Benzil

Experiment A2$_a$

Common name: benzil
CA number: [134-81-6]
CA name as indexed: ethanedione, diphenyl-

Purpose. Benzil is the second of three synthetic intermediates in the **a** series of Sequential Reactions, which lead to the synthesis of hexaphenylbenzene. Benzoin, the starting material for this step, is prepared in Experiment [A1$_a$]. Benzil, the product formed in the present reaction, is used in the synthesis of tetraphenylcyclopentadienone in Experiment [A3$_a$]. Tetraphenylcyclopentadienone is then converted to hexaphenylbenzene in Experiment [A4$_{ab}$].

This experiment also affords an excellent opportunity for you to investigate the use of a soluble, metal–ion catalyst as an oxidizing agent. In this case, a secondary alcohol is oxidized to a ketone. Because nitrogen gas is formed as a byproduct, the progress and rate of the reaction can be followed by measuring the evolution of nitrogen.

NOTE. If you plan to continue the synthetic sequence to hexaphenylbenzene, the semimicroscale procedure described below should be used. If you wish to study this reaction as an individual microscale experiment, those conditions follow the semimicro discussion.

Prior Reading

Technique 5: Crystallization
- Use of the Hirsch Funnel (pp. 88–89)
- Craig Tube Crystallizations (pp. 89–91)

Technique 6: Chromatography
- Column Chromatography (pp. 92–95)
- Thin-Layer Chromatography (pp. 97–99)
- Concentration of Solutions (pp. 101–104)

Technique 7: Collection or Control of Gaseous Products (pp. 105–107)

REACTION

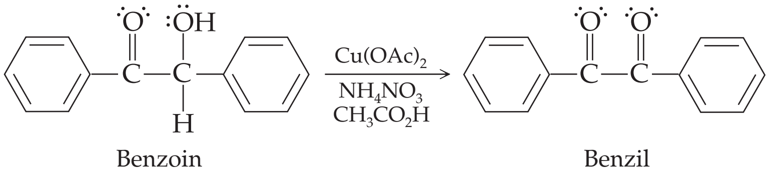

DISCUSSION

Benzil, a diketone, is obtained by the catalytic oxidation of benzoin using Cu^{2+} ion as the catalytic oxidant. The reaction is general for α-hydroxyketones, and is the basis of Fehling's qualitative test for certain sugars. The mechanism of the oxidation shows the catalytic effect of the Cu^{2+} ion as it is continuously reduced and reoxidized in the sequence outlined below. A key ingredient is the nitrate ion, which oxidizes the Cu^{+} to the Cu^{2+} oxidation state, and is in turn reduced to nitrite ion. The nitrite ion in the presence of acid and ammonium ion, decomposes to yield nitrogen gas and water.

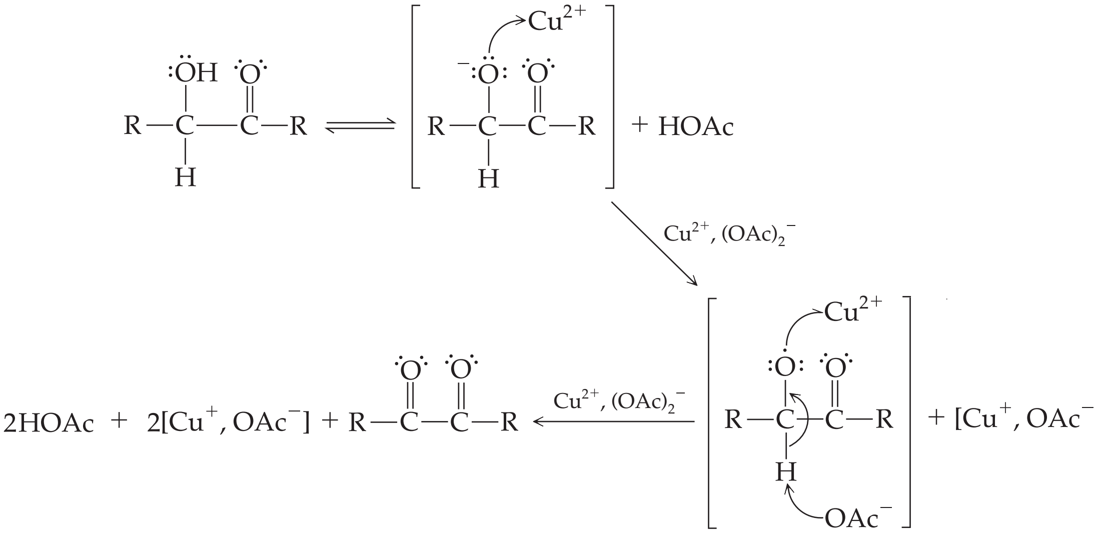

and

$$2\,Cu^{+} + 2\,H^{+} + NO_3^{-} \rightarrow 2\,Cu^{2+} + NO_2^{-} + H_2O$$

$$NH_4NO_2 \xrightarrow{H^{+}} N_2 + 2\,H_2O$$

SEMIMICROSCALE EXPERIMENTAL PROCEDURE

(The microscale reaction is increased fourfold.)

Estimated time of the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)
Benzoin	212.25	400 mg	1.88	137
Cupric acetate solution		1.4 mL		

INSTRUCTOR PREPARATION. Prepare the catalyst solution by dissolving 0.1 g of cupric acetate and 5 g of ammonium nitrate in 7.0 mL of deionized water (may require warming), followed by addition of 28 mL of glacial acetic acid. Place the container in the ***hood*** *and dispense the solution by use of an automatic delivery pipet.*

Reagents and Equipment. Equip a 5.0-mL conical vial, containing a magnetic spin vane, with a reflux condenser, to which is attached a gas exit delivery tube(➡). Weigh and add to the vial 400 mg (1.88 mmol) of benzoin, followed by 1.4 mL of cupric acetate catalyst solution (using a calibrated Pasteur pipet).

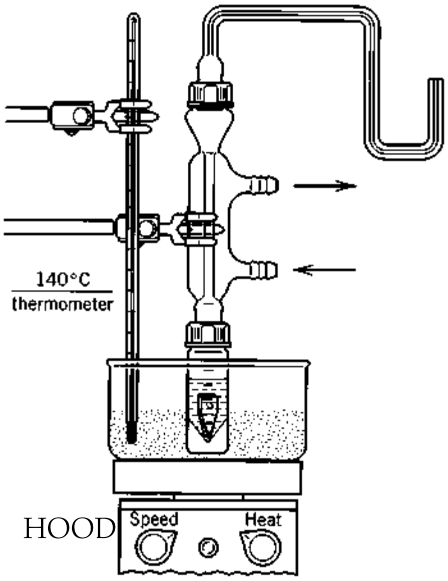

Benzoin, 400 mg
+ cupric acetate catalyst, 1.4 mL,
5–mL conical vial

Arrange the gas delivery tube so that it fits into an inverted 100-mL graduated cylinder that is filled with, and immersed in, a beaker of water (➡). This arrangement facilitates the measurement of the nitrogen gas evolved during the course of the reaction.

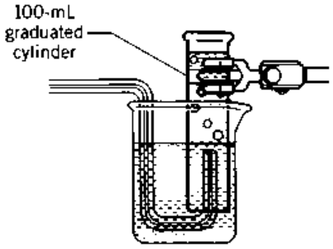

NOTE. It is absolutely necessary that all connections be tight to prevent leakage of the gas evolved. Lightly grease the ground-glass joint on the gas delivery tube.

Reaction Conditions. Heat the reaction mixture with stirring at a sand bath temperature of 140–145 °C for about 1 h or until the collected gas volume remains constant. *As the benzoin dissolves, the reaction mixture turns green and evolution of nitrogen gas commences. The theoretical volume of gas from the oxidation of 400 mg of benzoin is 42.4 mL at standard temperature and pressure (STP).*

Isolation of Product. If the gas delivery tube is used, *disconnect* it from the top of the condenser *before* removing the reaction vial from the heat source.

Cool the reaction mixture to room temperature, add 2 mL of cold water (using a calibrated Pasteur pipet), and then place the reaction vial in an ice bath for 10 min. Collect the yellow crystals of benzil by vacuum filtration using a Hirsch funnel (➡). Rinse the reaction vial and crystals with two additional 2-mL portions of cold water.

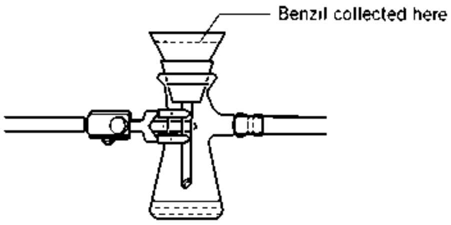

Aqueous acetic acid,
~7.5 mL and other
reaction by-products

Purification and Characterization. Purify the crude product by recrystallization from methanol, or from 95% ethanol. Dry the recrystallized yellow benzil on a porous clay plate, or on filter paper. *The benzil obtained after recrystallization often contains a small amount of benzoin impurity. It may* *be purified by chromatography on silica gel using the procedure outlined below.* Before carrying out the column chromatography purification of benzil, the purity of the initial product may be assessed using thin-layer chromatography (see Prior Reading). Use methylene chloride as the elution solvent, silica gel as the stationary phase, and UV light for visualization. Typical R_f values are benzil, 0.62; benzoin, 0.33.

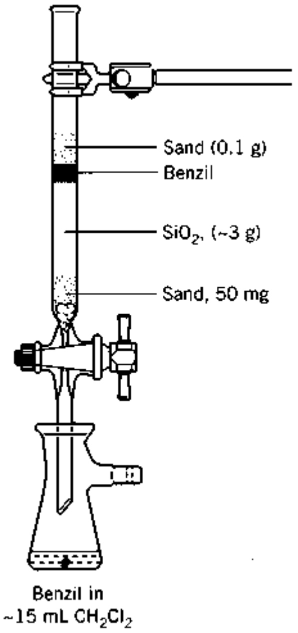

Pack a 1.0-cm-diameter column to a height of 5 cm with a slurry of activated silica gel in methylene chloride (⬅). Introduce the sample of benzil to the column, followed by 100 mg of sand. Use approximately 10–15 mL of methylene chloride to elute the benzil, which is easily identified on the column because of its yellow color. Concentrate the eluate collected in a 25-mL filter flask to obtain the pure benzil (see Prior Reading). A rotary evaporator is an effective alternative.

Weigh the dried product and calculate the percent yield. Determine the melting point and compare your value to that reported in the literature.

Obtain an IR spectrum of the product and compare it with that of the starting material and to that reported in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Benzil also has a characteristic UV spectrum (see Fig. 7.2). It exhibits a wavelength maximum (λ_{max}) at 259 nm (ε_{max} = 16,329 methanol). It is of interest to compare this absorption spectrum of benzil with that of benzoin (Experiment [$A1_a$]). If the melting point and infrared spectrum compare reasonably closely to the literature values, this material may be used in the synthesis of tetraphenylcyclopentadienone (Experiment [$A3_a$]). If the melting point is low, check the product's purity by thin-layer chromatography.

Chemical Tests. Ketones and aldehydes are often characterized by the preparation of a solid derivative. To assist in the characterization of benzil, prepare its 2,4-dinitrophenylhydrazone or semicarbazone (Chapter 9). The melting points of these derivatives are listed in Appendix A, Table A.5.

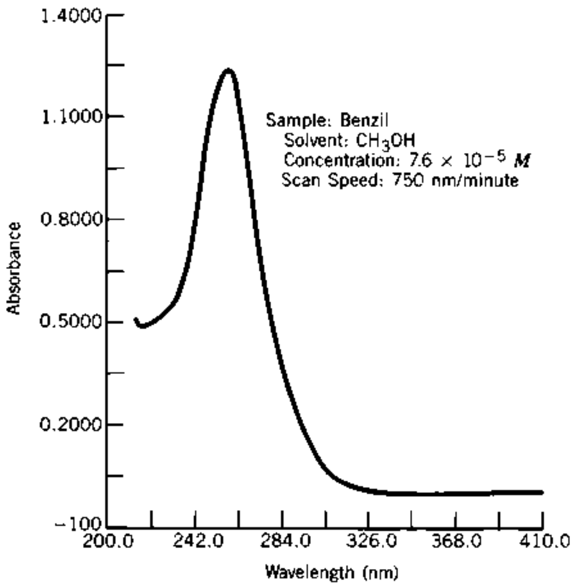

Figure 7.2 UV-visible spectrum: benzil.

OPTIONAL MICROSCALE PREPARATION

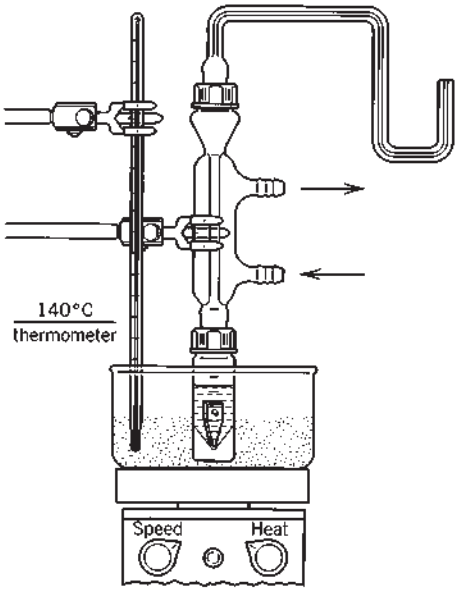

Benzoin, 100 mg + oxidation catalyst, 350 µL, 3-mL conical vial

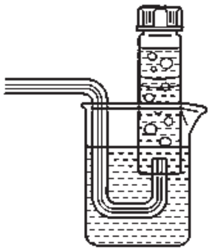

The microscale procedure is similar to that outlined above for the semimicroscale preparation, with the following exceptions:

1. Use a 3.0-mL conical vial containing a magnetic spin vane fitted with a reflux condenser to which is attached a gas delivery tube (➨).
2. Collect the gas in an inverted calibrated collection tube (➨).
3. Decrease the amount of the reagents and solvents fourfold.

Physical Properties of Reactants				
Compound	**MW**	**Amount**	**mmol**	**mp (°C)**
Benzoin	212.25	100 mg	0.47	137
Cupric acetate solution		350 µL		

4. Heat the reaction mixture at about 140 °C until the evolution of gas has ceased.
5. Add 0.5 mL of cold water to the cooled reaction product and, after filtration, wash the material with two 0.5-mL portions of cold water.
6. Purify and characterize the benzil product as described in the semimicroscale procedure, but using a Craig tube for the recrystallization.

QUESTIONS

7-6. In the directions given for the experiment, it is emphasized that the gas delivery tube must be disconnected from the top of the condenser before removing the reaction vial from the heat source. Why is this necessary?

7-7. What qualitative chemical tests would you perform to distinguish between benzoin and benzil? (See Chapter 9.)

7-8. 1,2-Dicarbonyl compounds, such as benzil, can be characterized by reaction with *o*-phenylenediamine to form a substituted quinoxaline:

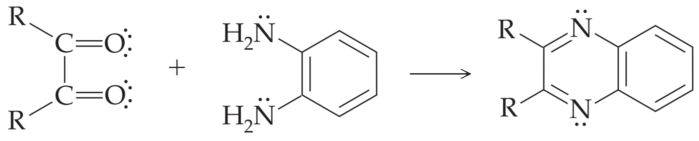

o-Phenylenediamine (a quinoxaline)

(a) Write the structure for the quinoxaline derivative obtained when benzil is the reactant. Do you think this compound would be colored? If so, why?

(b) Suggest a suitable mechanism for the formation of the quinoxaline compounds based on the reaction scheme shown above.

(c) What reagent would you react with *o*-phenylenediamine to prepare the unsubstituted compound quinoxaline? Show a reaction scheme giving the structure of reactants and products.

7-9. Make a sketch of the ^1H NMR spectrum that would be observed for benzil.

7-10. Suggest a method for the synthesis of $C_6H_5CH(OH)CH(OH)C_6H_5$ from benzil. Discuss the stereochemical aspects of this 1,2 diol.

7-11. Based on the ultraviolet data given for benzil in the Purification and Characterization section of this experiment, what concentration of the benzil must have been used if a 1-cm cell was used, and a maximum absorption of 0.5 was observed?

BIBLIOGRAPHY

The synthesis of benzil is reported in *Organic Syntheses:*

Adams, R.; Marvel, C. S. *Organic Syntheses* **1921,** *1,* 25.
Clarke, H. T.; Dreger, E. E. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 87.

For further information on the oxidation see

Depreux, P.; Bethegines, G.; Marcinal-Lefebure, A. *J. Chem. Educ.* **1988,** *65,* 553.
Haines, A. H. *Methods for the Oxidation of Organic Compounds;* Academic Press: New York, Vol. 1, 1985, Vol. 2, 1988.
Hudlicky, M. *Oxidations in Organic Chemistry;* American Chemical Society: Washington, DC, 1990.
Mijs, W. J.; de Jonge, C. R. H. I., Eds. *Organic Synthesis by Oxidation with Metal Compounds;* Plenum: New York, 1987.
Nigh, W. G. In *Oxidation in Organic Chemistry;* Trahanovsky, W. S. Ed.; Academic Press: New York, 1973, Vol. 5, Part B, Chapter 1.
Weiss, M.; Appel, M. *J. Am. Chem. Soc.* **1948,** *70,* 3666.

Experiment A3$_{\text{a}}$

Tetraphenylcyclopentadienone

Common name: tetraphenylcyclopentadienone
CA number: [479-33-4]
CA name as indexed: 2,4-cyclopentadien-1-one, 2,3,4,5-tetraphenyl-

Purpose. The cyclic dienone product of this aldol condensation is the *third* intermediate in the **a** series of Sequential Reactions, which lead to the target molecule, hexaphenylbenzene. It is the last intermediate in the **a** series, and when reacted with the *last* intermediate in the **b** series (Experiment [A4$_{\text{ab}}$]) will give the target molecule.

In addition to supplying a key intermediate in the **a** series synthetic sequence, the experiment illustrates the use of the aldol condensation for the synthesis of a five-membered carbocyclic ring. The product also is a good demonstration of the impact of extended conjugation on the absorption of visible light. Starting with bright-yellow benzil, we form an even more extended π system in this experiment, and as a result, the tetraphenyldienone product absorbs strongly over a significant portion of visible spectrum and is thus deeply colored.

NOTE. If you plan to continue the synthetic sequence to hexaphenylbenzene, the first microscale procedure described on page 439 should be used. If you wish to study this reaction as an individual microscale experiment, follow the second set of conditions.

Prior Reading

Experiment [20] (pp. 309–317)
Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallizations (pp. 89–91)

REACTION

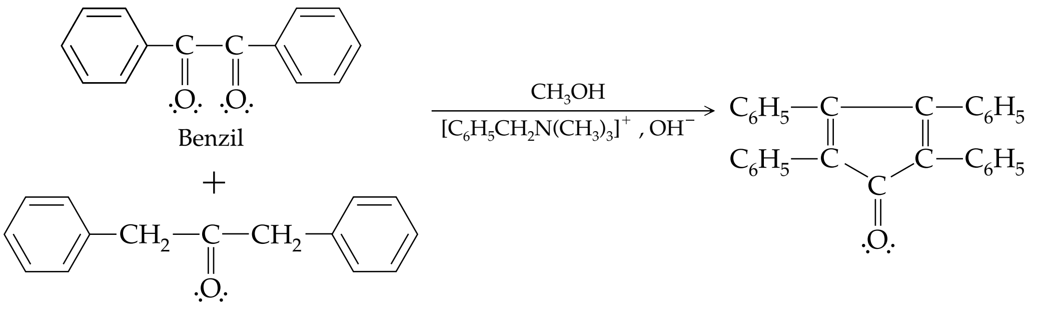

1,3-Diphenylacetone

Tetraphenylcyclopentadienone

DISCUSSION

This experiment is a further example of the aldol condensation reaction (see Experiment [20] for discussion and Experiment [F1] for another example).[1] The reaction carried out here differs from the earlier example in that in this case two ketones, one of which has no α-hydrogen atoms, are the reactants. It is also different because the reagents selected lead to the formation of a carbocyclic ring system. The initially formed aldol product undergoes an elimination reaction to yield a material that has a highly conjugated system of double bonds. In general, the more extended the conjugation in a molecule, the less energy is required to promote the π electrons to a higher energy level. In this case, energy in the visible region of the spectrum is absorbed, which results in a product possessing a deep purple color.

The mechanism involves a sequence of two aldol condensations. The first is intermolecular; the second is intramolecular. The mechanism is similar to that outlined in Experiment [20].

The product of this reaction is the dienone intermediate used in the final step of the synthesis of hexaphenylbenzene (see Experiment [A4$_{ab}$]).

MICROSCALE REACTION PROCEDURE (1)

(The first microscale reaction is increased by a factor of 2.)

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
1,3-Diphenylacetone	210.28	100 mg	0.48	35	
Benzil	210.23	100 mg	0.48	95	
Triethylene glycol	150.18	0.5 mL			278
Benzyltrimethylammonium hydroxide (40% solution in MeOH)		100 µL			

Reagents and Equipment. Measure and place 100 mg (0.48 mmol) of 1,3-diphenylacetone and 100 mg (0.48 mmol) of benzil, followed by 0.5 mL of triethylene glycol, in a 3.0-mL conical vial containing a magnetic spin vane. Equip the vial with an air condenser (➨).

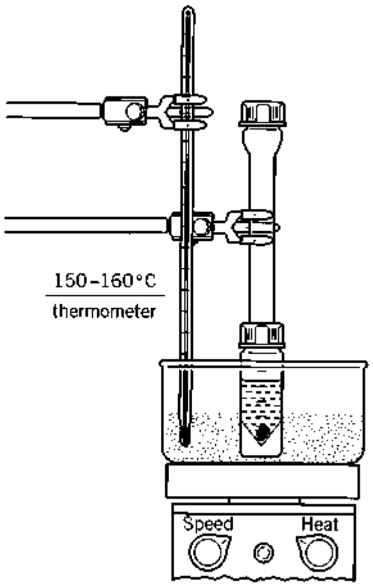

Benzil, 100 mg + 1,3–diphenylacetone, 100 mg + triethylene glycol, 0.5 mL + [Immediately upon removal from the sand bath 100 µL of 40% benzyltrimethylammonium hydroxide in methanol is added.]

NOTE. The benzil used in this reaction must be free of benzoin impurity. If benzil is prepared according to Experiment [A2$_a$], it should be purified by the chromatographic procedure cited therein if benzoin is detected by thin-layer chromatography.

Reaction Conditions. Heat the mixture with stirring in a sand bath at 150–160 °C for 5–10 min. The benzil should dissolve during this time.

Next, remove the apparatus containing the homogeneous reaction solution from the heat source, and immediately add 100 µL of a 40% benzyltrimethylammonium hydroxide–methanol solution to the hot reactants, with gentle shaking. As cooling occurs, dark purple crystals of tetraphenylcyclopentadienone appear. *Cooling may be accelerated by placing the vial under a stream of cold water.*

---

[1] For references to the aldol condensation reaction, refer to Experiment [20].

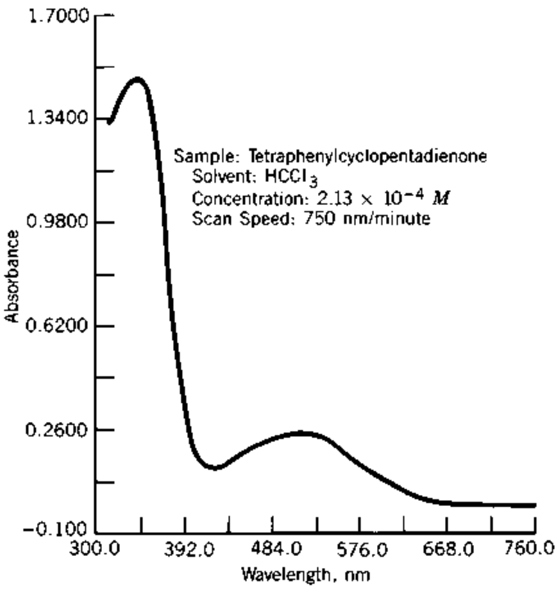

Figure 7.3 UV-visible spectrum: tetraphenylcyclopentadienone.

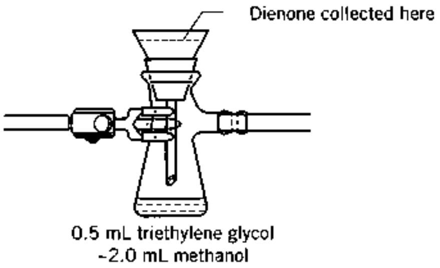

Isolation of Product. Add 1.4 mL of cold methanol with stirring (glass rod), and then cool the mixture in an ice bath for 5–10 min. Collect the dark crystals by filtration under reduced pressure using a Hirsch funnel. Rinse the reaction vial and crystals with a few drops of cold methanol. Continue dropwise addition of cold methanol to the crystals on the filter bed until the product appears purple and is no longer brownish in color. Finally, allow the crystalline product to air-dry on a porous clay plate (or in an oven at about 80 °C for 1 h) (⬅).

Purification and Characterization. The product is often of sufficient purity for direct use in the preparation of hexaphenylbenzene (Experiment [A4$_{ab}$]). If further purification is required, the intermediate dienone may be recrystallized from triethylene glycol using a Craig tube.

Weigh the tetraphenylcyclopentadienone product and calculate the percent yield. Determine the evacuated melting point and compare it with the literature value. Obtain an IR spectrum of the material and compare it with that of an authentic sample.

The comparison of the UV-visible spectra of benzil (see data given in Experiment [A2$_a$]) and the product may be used to demonstrate the shift of absorption bands with increased conjugation in the molecule.

The UV-visible data for tetraphenylcyclopentadienone are summarized as follows and as shown in Figure 7.3 (see Chapter 8).

$$\lambda_{max}\ 510\text{ nm } (\varepsilon_{max} = 1080, \text{ chloroform})$$

$$\lambda_{max}\ 345\text{ nm } (\varepsilon_{max} = 6380, \text{ chloroform})$$

NOTE. If you have synthesized the tetraphenylcyclopentadienone from benzaldehyde, calculate the overall yield to this point in the synthesis of hexaphenylbenzene. Base these calculations on the amount of benzaldehyde you started with.

MICROSCALE REACTION PROCEDURE (2)

When the microscale procedure is used as a single-step experiment, and not as part of the synthesis of hexaphenylbenzene, the scale is conveniently reduced to one-half that outlined above, with the following experimental modifications.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
1,3-Diphenylacetone	210.28	50 mg	0.24	35		
Benzil	210.23	50 mg	0.24	95		
Triethylene glycol	150.18	0.25 mL			278	1.124
Benzyltrimethylammonium hydroxide (40% solution in MeOH)		50 μL				

1. The reaction is carried out in a 3-mL conical vial.
2. The reaction mixture is maintained at 155–165 °C for 10 min.
3. Following addition of the base, the vial is reheated to 150–160 °C (2–3 min) and then allowed to cool.

QUESTIONS

7-12. The above reaction carried out to construct the intermediate dienone parallels an earlier example of an aldol condensation in Experiment [20] in which another dienone was synthesized. In both cases all the reactants possessed carbonyl groups. What further structural similarities between the key reactants were required so that both pathways would lead to aldol condensations?

7-13. Outline a complete mechanistic sequence to account for the formation of the tetraphenylcyclopentadienone compound.

7-14. Cyclopentadienone is unstable and rapidly undergoes the Diels–Alder reaction with itself. Write the structure for this Diels–Alder addition product.

7-15. The Diels–Alder addition product of Question 7-14 undergoes a fragmentation reaction on heating to produce a bicyclotrienone compound plus carbon monoxide. Suggest a structure for this product.

7-16. Using the Hückel [$4n + 2$] rule for aromaticity, predict which of the following species might be expected to show aromatic properties:

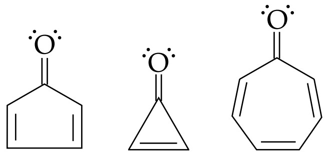

7-17. Based on Questions 7-14 and 7-15, why is tetraphenylcyclopentadienone such a stable compound?

BIBLIOGRAPHY

An *Organic Syntheses* preparation of tetraphenylcyclopentadienone is available:

Johnson, J. R.; Grummitt, O. *Organic Syntheses;* Wiley, New York, 1955; Collect. Vol. III, p. 80.

For preparing this material under microwave conditions see

Elder, J. W. *J. Chem. Educ.* **1994,** *71,* A142.

(*E*)-Stilbene

Experiment A1$_b$

Common names: (*E*)-stilbene; *trans*-1,2-diphenylethene

CA number: [103-30-0]

CA name as indexed: benzene, 1,1′-(1,2-ethenediyl)bis-, (*E*)-

Purpose. The purpose of this experiment is to prepare a sufficient quantity of (*E*)-stilbene to complete a multistep sequence of synthetic reactions to obtain the target compound, hexaphenylbenzene (Experiment [$A4_{ab}$]).

A further purpose of experiment [$A1_b$] is to investigate the use of the Horner–Wadsworth–Emmons modified Wittig reaction to complete the synthetic objective. You will also study a reaction that involves the condensation of an aldehyde and a phosphonate ester, using a highly effective *phase-transfer catalyst*.

NOTE. (E)-*stilbene is the first intermediate to be synthesized in the* ***b*** *series of Sequential Reactions en route to hexaphenylbenzene. If you plan to carry out the entire sequence to hexaphenylbenzene, the semimicroscale procedure presented here should be used. Conditions to run the reaction as an individual microscale experiment are given in detail in Experiment [19B].*

Prior Reading

- *Technique 4:* Solvent Extraction
    - Liquid–Liquid Extraction (pp. 72–75)
    - Drying of a Wet Organic Layer (pp. 80–83)
- *Technique 5:* Crystallization
    - Use of the Hirsch Funnel (pp. 88–89)
    - Craig Tube Crystallization (pp. 89–91)
- *Technique 6:* Chromatography
    - Packing the Column (pp. 93–94)
    - Elution of the Column (pp. 94–95)
    - Thin-Layer Chromatography (pp. 97–99)
    - Concentration of Solutions (pp. 101–104)

REACTION

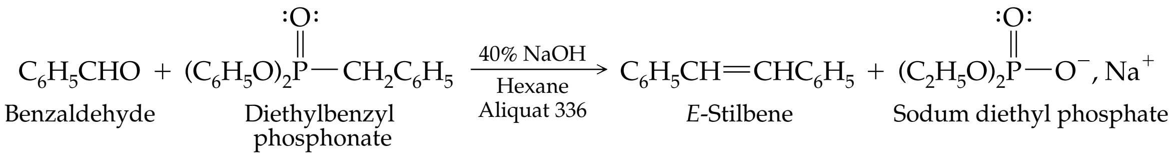

$$C_6H_5CHO + (C_6H_5O)_2\overset{\overset{:O:}{\|}}{P}-CH_2C_6H_5 \xrightarrow[\text{Hexane, Aliquat 336}]{40\%\ \text{NaOH}} C_6H_5CH{=}CHC_6H_5 + (C_2H_5O)_2\overset{\overset{:O:}{\|}}{P}-O^-, Na^+$$

DISCUSSION

For a discussion of the Wittig reaction and a list of references, including the mechanism and modifications, see Experiment [19]. The role of the phase-transfer catalyst in the Horner–Wadsworth–Emmons modification of the Wittig reaction is also discussed in some detail in that experiment.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	d	n_D
Benzaldehyde	106.12	600 μL	5.88	178	1.04	1.5463
Diethyl benzylphosphonate	228.23	1.20 mL	5.57	106–108 (@1 mm)	1.095	1.4970
Aliquat 336 (tricaprylmethylammonium chloride)	404.17	528 mg	1.30			
Hexane	86.18	12.0 mL				
40% Sodium hydroxide		12.0 mL				

Reagents and Equipment. Weigh and place 528 mg (600 μL of tricaprylmethylammonium chloride (Aliquat 336) in a 50-mL round-bottom flask containing a magnetic stirrer. Add 600 μL (5.88 mmol) of *freshly distilled* benzaldehyde, 1.20 mL (5.57 mmol) of diethyl benzylphosphonate, 12.0 mL of hexane, and 12 mL of 40% sodium hydroxide solution. Attach the flask to a reflux condenser (➡).

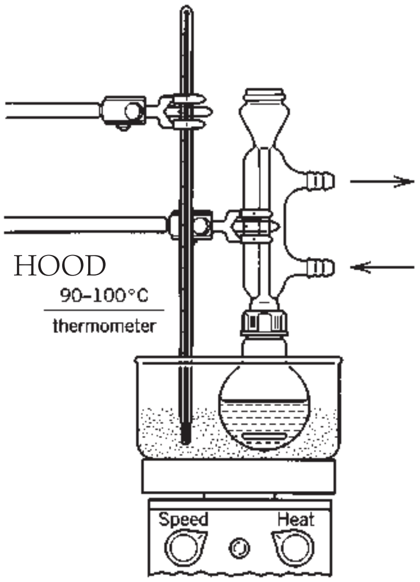

C_6H_5CHO, 600 μL + diethyl benzylphosphonate, 1.2 mL + Aliquat 336, 528 mg + 40% NaOH, 12 mL, hexane, 12 mL, 50-mL RB flask

NOTE. The benzaldehyde, (automatic delivery pipet), diethyl benzylphosphonate (2-mL glass pipet), hexane, and NaOH solution are dispensed in the ***hood.***

Aliquat 336 is very viscous and is best measured by weighing. A medicine dropper is used to dispense this material.

It is advisable to lightly grease the bottom joint of the condenser because strong base is being used in the reacting medium. At the end of the reaction, it is also important to loosen the Cap-seal and twist the joint to make sure it is free to rotate as the apparatus is cooling.

Reaction Conditions. Heat the two-phase mixture at reflux on a sand bath at about 90–100 °C for 1 h. Stir the reaction mixture vigorously during this period. Allow the resulting orange solution to cool to nearly room temperature. Crystals of product may appear as cooling occurs.

Isolation of Product. Add 4.0 mL of methylene chloride, which will dissolve any crystalline material that may have formed. Now transfer the contents of the round-bottom flask to a 125-mL separatory funnel. Rinse the flask with an additional 2.0–3.0 mL of methylene chloride, and transfer this rinse to the separatory funnel using a Pasteur filter pipet. Remove the aqueous layer carefully (*because the densities of the aqueous and organic layers are rather close, it is wise to test the solubility of a few drops of the bottom layer in water to ascertain which phase is the aqueous one in the separatory funnel*), and then wash the remaining organic layer with two 6.0-mL portions of water. Save the combined aqueous extracts in a 50-mL Erlenmeyer flask until you have successfully isolated and characterized the product. Now transfer the remaining wet methylene chloride solution to a 50-mL Erlenmeyer flask. Dry the solution by addition of anhydrous sodium sulfate (~1–2 g). Use a 50-mL glass pipet to transfer the dried solution to a second 50-mL Erlenmeyer flask. Wash the sodium sulfate remaining in the first Erlenmeyer flask with two 3-mL portions of methylene chloride. Remove these washings, using a Pasteur filter pipet, and transfer them to the same Erlenmeyer flask.

You may concentrate the solution, which contains the desired product, to dryness on a warm sand bath under a gentle stream of nitrogen gas, but a more efficient and rapid procedure is to remove the solvent by rotary evaporation (see Concentration of Solutions, Technique 6B).

Purification and Characterization. Recrystallize the crude (*E*)-stilbene from 95% ethanol in a small Erlenmeyer flask or a test tube. Collect the recrystallized material by vacuum filtration using a Hirsch funnel. Maintain the vacuum for an additional 10 min to partially dry the crystalline product. Now place the material on a clay plate, or on filter paper, and allow it to air-dry thoroughly. As an alternative procedure, the final traces of water may be removed by placing the sample (use an open test tube with the mouth covered by filter paper retained by a rubber band) in a vacuum drying oven (or pistol) for 10–15 min at 30 °C (1–2 mm).

The recrystallized compound, generally, is sufficiently pure to use in the next step of the **b** series of Sequential Reactions, Experiment [$A2_b$].

Weigh the (*E*)-stilbene and calculate the percent yield. Obtain a melting point and IR spectrum of the material, and compare your results with those reported in the literature.

Further characterization of the (*E*)-stilbene, including thin-layer chromatography, gas chromatography, and ultraviolet–visible spectroscopy may be carried out as outlined in Experiment [19A].

Chemical Tests. Further characterization may be accomplished by performing the Br_2/CH_2Cl_2 test for unsaturation. Note that the dibromo compound is prepared in Experiment [$A2_b$]. It may be used here as a derivative to characterize the (*E*)-stilbene. The ignition test (see Chapter 9) may be used to confirm the presence of the aromatic portion of the molecule.

QUESTIONS

7-18. Give the structure of the phosphorus ylide and carbonyl compound you might use to prepare the following alkenes:
(a) Methylenecyclohexane **(b)** 2-Methyl-2-hexene **(c)** $C_6H_5CH{=}C(CH_3)_2$

7-19. Trimethylphosphine is less expensive than triphenylphosphine. However, it cannot be used in the preparation of most phosphorus ylides. Explain.

7-20. How, starting from triphenylphosphine, $(C_6H_5)_3P$:, can you prepare the following ylide:
$(C_6H_5)_3P{=}C(CH_3)CH_2CH_2CH_3$

7-21. Consult *The Aldrich Library of IR Spectra* (and/or SciFinder Scholar) for IR spectra of (*E*)- and (*Z*)-stilbene. Which absorption bands are most useful in determining the difference between the two compounds?

7-22. Compare the mechanisms of the aldol condensation (Experiment [20]) with that of the Wittig reaction. Point out any similarities and/or differences.

Experiment $A2_b$

Bromination of (*E*)-Stilbene: *meso*-Stilbene Dibromide

Common names: *meso*-stilbene dibromide, *meso*-1,2-dibromo-1, 2-diphenylethane

CA number: [13440-24-9]

CA name as indexed: benzene, 1,1′-(1,2-dibromo-1,2-ethanediyl)bis-, (*R**,*S**)-

Purpose. You will synthesize the *second* intermediate in the **b** series of Sequential Reactions by carrying out the bromination of (*E*)-stilbene to obtain *meso*-stilbene dibromide. This product is the precursor to diphenylacetylene, the next synthetic intermediate in the **b** series. A further purpose of this experiment is to demonstrate the *stereospecific* addition of bromine to alkenes.

NOTE. If you plan to continue the synthetic sequence to hexaphenylbenzene, the semimicroscale procedure described below should be used. If you wish to study this reaction as an individual microscale experiment, those conditions and other scaleup options follow the semi-micro discussion.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)

REACTION

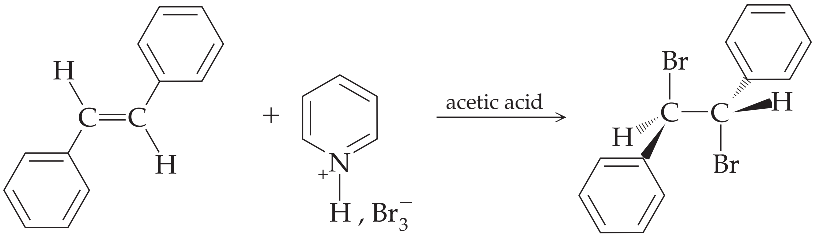

(*E*)-Stilbene

Pyridinium bromide perbromide

meso-Stilbene dibromide

DISCUSSION

The bromination of alkenes is an example of an electrophilic addition reaction (also see Experiments [D2] and [F2]).

In the present reaction, bromination of (*E*)-stilbene yields *meso*-stilbene dibromide. Thus, this reaction is classed as *stereospecific* because the other possible diastereomers are not formed.

The reaction proceeds in two stages. The first step involves the formation of an intermediate cyclic *bromonium ion*. The concept of a three-membered cyclic intermediate was first proposed as early as 1937. Subsequent studies have provided solid evidence that cyclic halonium ions do, indeed, exist. For example, stable solutions of cyclic bromonium ions in liquid SO_2 (−60 °C) have been prepared as SbF_6^- salts. Two examples are given here:

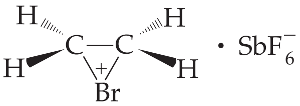

Ethylene bromonium ion salt

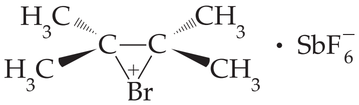

Tetramethylethylene bromonium ion salt

Nuclear magnetic resonance spectroscopic measurements have provided powerful evidence that these and other selected alkenes form stable *bridged* bromonium ion salts. A solid bromonium ion tribromide salt of adamantylidene adamantane has been isolated, and its structure determined by X-ray crystallography:

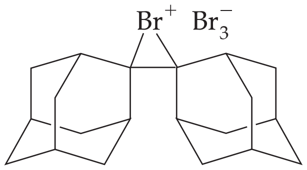

Tribromide salt of adamantylidene adamantane

The bromine molecule (Br_2) is normally symmetrical. However, as it approaches the nucleophilic and electron-rich π bond of the alkene, it becomes polarized by induction and can then function as the electrophile in an addition reaction. The result is the generation of a cyclic bromonium ion:

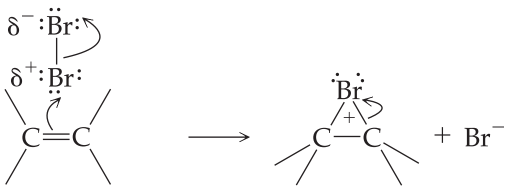

Induced polarization of Br_2 as it approaches the alkene

Bromonium ion

In the present reaction, both the bromine and the *(E)*-stilbene are achiral. However, the bromonium ion that is produced is chiral. In this ion, the bromine atom bridges both carbon atoms of the original carbon–carbon double bond to form a three-membered ring intermediate. The generation of a cyclic species has a profound effect on the *stereochemistry* of the second step of the bromine addition.

The second stage of the bromination involves nucleophilic attack by bromide ion on the intermediate bromonium ion. Since the nucleophile must approach from the face opposite the leaving group, bond formation involves inversion of configuration at the carbon center under attack in the second stage of the bromination reaction.

Note that *either* carbon can be approached by the nucleophile (one attack is shown). This second step is a classic backside S_N2 sequence. The bromination of cyclic alkenes provides further evidence that this type of halogenation is an anti addition, with the bromine atoms introduced trans to one another:

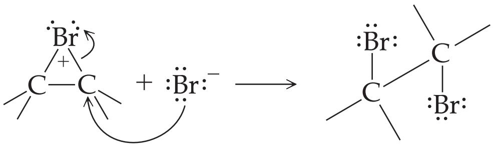

Bromonium ion

Dibromo product

It is important to realize that if two different groups are present on one or both of the sp^2 carbon atoms of the alkene linkage, chiral carbon centers are generated when bromine is added, though if a chiral product were formed from achiral reagents, one would expect it to be racemic. In the case of *(E)*-stilbene, two chiral centers are generated. However, due to the symmetry of the reactants and the stereoselectivity of the reaction, only the meso diastereomer is formed:

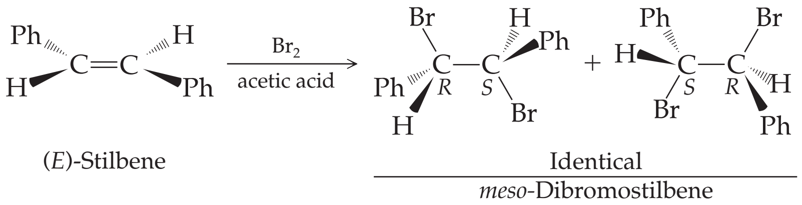

Refer to the Discussion section of Experiment [D2] for further information on the stereochemistry of bromination reactions.

Bromination of alkenes using a Br_2–CCl_4 solution (a red-brown color) is frequently used as a qualitative test for the presence of unsaturation in a compound. Rapid loss of color from the reagent solution is a positive test (see Chapter 9). Pyridinium bromide perbromide, a solid brominating agent, is used as a source of bromine in this experiment. The material is more convenient to handle than liquid bromine (see Experiment [D2]).

SEMIMICROSCALE EXPERIMENTAL PROCEDURE

(The microscale reaction is increased by a factor of 2.6.)

Estimated time to complete the reaction: 1.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
(*E*)-Stilbene	180.25	600 mg	3.3	122–124		
Glacial acetic acid		12 mL			118	1.049
Pyridinium bromide perbromide	319.83	1.2 g	3.7	205		

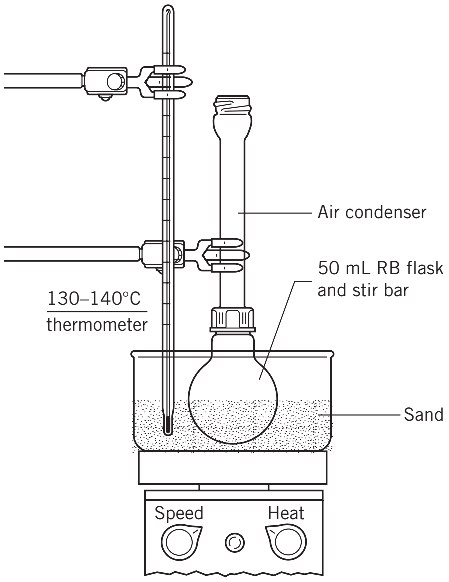

(*E*)-Stilbene, 600 mg + glacial acetic acid, 12 mL + pyridinium bromide perbromide, 1.2 g

Reagents and Equipment. In a 50.0-mL round-bottom flask containing a magnetic spin bar and equipped with an air condenser, weigh and place 600 mg (3.3 mmol) of *(E)*-stilbene. Next add 6 mL of glacial acetic acid (using a graduated cylinder), and warm the resulting mixture in a sand bath at 130–140 °C with stirring until the solid dissolves (~5 min) (➡).

CAUTION: Glacial acetic acid is corrosive and toxic. It is dispensed in the ***hood*** using an automatic delivery pipet.

HOOD

Remove the condenser from the flask, and to the warm solution, in the **hood,** add 1.2 g (3.7 mmol) of pyridinium bromide perbromide in one portion. Wash down any perbromide adhering to the sides of the flask with an additional 6 mL of acetic acid using a Pasteur pipet. Reattach the air condenser.

HOOD

CAUTION: The brominating agent is a mild lachrymator. It should be dispensed in the ***hood.*** An alternative solid brominating agent is tetra-*N*-butylammonium tribromide.

HOOD

Reaction Conditions. With stirring, heat the reaction mixture at a sand bath temperature of 130–140 °C for an additional 5–6 min. (The product often begins to precipitate during this period.)

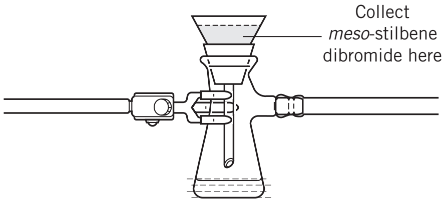

Isolation of Product. Remove the reaction flask from the heat source and allow it to cool to approximately 40–50 °C (water bath). Add 12 mL of water, with swirling, and then place the flask in an ice bath for 5–8 min. Collect the crystalline solid by vacuum filtration using a Hirsch funnel (⬅).

Purification and Characterization. Wash the material with three 2-mL portions of cold water to obtain white crystals, and then with two 2-mL portions of acetone. Air-dry the product on a clay plate or on filter paper.

Weigh the *meso*-stilbene dibromide and calculate the percent yield. Determine the evacuated melting point, and compare your result with the literature value. Obtain IR and NMR spectra and compare them with those reported in the literature (*The Aldrich Library of IR Spectra*, *The Aldrich Library of NMR Spectra,* and/or the corresponding spectral data available online (e.g., SciFinder Scholar)).

Generally, the material is sufficiently pure to be used in the next stage of the **b** series of Sequential Reactions, the preparation of diphenylacetylene. If desired, a small portion (~10–20 mg) may be recrystallized from hot xylene using the Craig tube.

Chemical Tests. You may wish to perform several classification tests on the product (see Chapter 9). Carry out the ignition test to confirm the presence of an aromatic group. The Beilstein test can be used to detect the presence of bromine. The silver nitrate test for alkyl halides should also give a positive result.

Optional Macroscale and Microscale Preparations

Macroscale Reaction Procedure. The procedure is similar to that for the 2.6-fold scaleup preparation with the following exceptions:

1. The reagent and solvent amounts are increased approximately 4.3-fold over the microscale preparation.

Physical Properties of Reactants						
Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
(*E*)-Stilbene	180.25	1.0 g	5.5	122–124		
Glacial acetic acid		12 mL			118	1.049
Pyridinium bromide perbromide	319.83	2.0 g	6.2	205		

2. After cooling the reaction mixture, add 20 mL of water to assist in precipitating the product. Wash the collected crystals with three 3-mL portions of cold water followed by two 3-mL portions of acetone.

Microscale Reaction Procedure. The procedure is similar to that for the 2.6-fold scaleup preparation with the following modifications:

1. Use a 10-mL round-bottom flask

2. The reagent and solvent amounts are *decreased* by a factor of approximately 2.6.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
(*E*)-Stilbene	180.25	230 mg	1.28	122–124		
Glacial acetic acid		4.2 mL			118	1.049
Pyridinium bromide perbromide	319.83	450 mg	1.4	205		

3. Add 2.2 mL of glacial acetic acid at the same time as the addition of the (*E*)-stilbene.

4. An additional 2 mL of glacial acetic acid is added with the brominating reagent.

5. The reaction mixture is diluted with 4.5 mL of water, swirled, and placed in an ice bath for 5–8 min.

6. The filter cake is washed with three 2-mL portions of cold water, followed by two 2-mL portions of acetone.

QUESTIONS

7-23. Using suitable structures, draw the sequence for the addition of bromine to (*Z*)-stilbene.

7-24. Are the results different for the answer in Question 7-23 than for the result in this experiment? If so, how? What is the stereochemical relationship between the products formed in the two reactions?

7-25. Bromine undergoes addition to ethylene in the presence of a high concentration of Cl^- ion to give 1-bromo-2-chloroethane, as well as 1,2-dibromoethane. Chloride ion does not add to the C═C unless bromine is present. Suggest a suitable mechanism to explain these results. Is the rate of bromination significantly affected by the presence of the Cl^- ion?

7-26. Offer an explanation for the fact that bromine adds to 2,3-dimethyl-2-butene 920,000 times faster than to ethylene, to produce the respective dibromides.

7-27. A student adds a few drops of Br_2–CCl_4 solution to an unknown organic compound. The color of the bromine solution disappears. The student reports that the unknown contains a C═C. Would you arrive at the same conclusion? If not, why not?

BIBLIOGRAPHY

A large number of examples of the bromination of alkenes appear in *Organic Syntheses*. Selected references are given below:

Allen, C. F. H.; Abell, R. D.; Normington, J. B. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 205.

Cromwell, N. H.; Benson, R. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 105.

Fieser, L. F. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 195.

Khan, N. A. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 969.

McElvain, S. M.; Kundiger, D. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 123.

Paquette, L. A.; Barrett, J. H. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 467.

Rhinesmith, H. S. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 177.

Snyder, H. R.; Brooks, L. A. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 171.

Also see

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Wiley: New York, 2007, Chap. 15, p. 999 and references therein.

Experiment $A3_b$

Dehydrohalogenation of *meso*-Stilbene Dibromide: Diphenylacetylene

Common names: diphenylacetylene, diphenylethyne
CA number: [501-65-5]
CA name as indexed: benzene, 1,1′-(1,2-ethynediyl)bis-

Purpose. The product formed in this multiple elimination reaction is the third intermediate in the **b** series of Sequence A, and is one of the immediate precursors to our target molecule, hexaphenylbenzene. You will investigate the synthesis and properties of alkynes and become familiar with E2 elimination reactions.

NOTE. If you plan to continue the synthetic sequence to hexaphenylbenzene, the semimicroscale procedure described below should be used. If you wish to study this reaction as an individual microscale experiment, those conditions and other scaleup options follow the semi-micro discussion.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)

REACTION

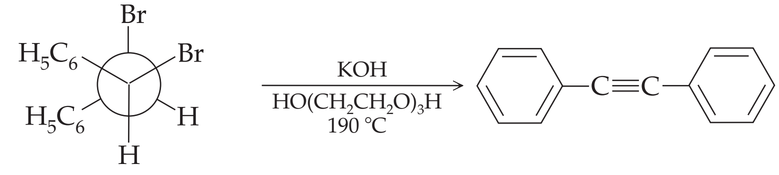

meso-Stilbene didromide Diphenylacetylene

DISCUSSION

This reaction illustrates the double dehydrohalogenation of a *vicinal* dibromo compound to form an alkyne. It is a useful reaction for the synthesis of alkynes, because the starting dibromides are readily available from alkenes (see, e.g., Experiment [$A2_b$]).

The double dehydrohalogenation reaction is usually run in the presence of a strong base, such as KOH or $NaNH_2$, and proceeds in two stages. In the first, an intermediate bromoalkene is formed, which can be isolated under more mildly basic conditions. In fact, this reaction is a valuable route to vinyl halides. The mechanism of elimination involves the abstraction of the proton on the carbon atom β to the halogen. The E2 mechanism, which operates under these strongly basic conditions, is fastest when it involves removal of a proton, H^+, antiperiplanar to the leaving group, Br^-. The E2 sequence of bond breakage and formation involves a smooth transition from reactant to product without the formation of an intermediate (concerted mechanism). The general mechanism is shown here:

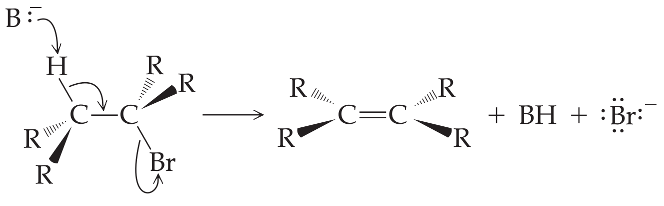

This type of elimination reaction is stereospecific because the geometry of the transition state requires that the H, both Cs, and the Br all lie in the same plane.

If *meso*-stilbene dibromide is treated with KOH in ethanol solvent, it is possible to isolate the monodehydrohalogenation product, the bromoalkene.

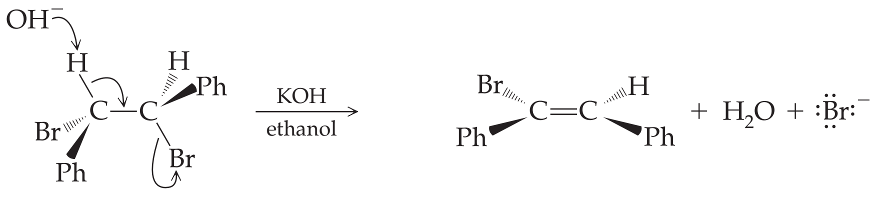

meso-Stilbene dibromide (*E*)-1-Bromo-1,2-diphenylethylene

The second stage of the reaction involves a higher activation energy, and therefore it requires higher temperatures to proceed. In the presence of a strong base near 200 °C, the bromoalkene undergoes an E2 elimination to form the triple bond. Part of the reluctance to eliminate, in this particular case, results from the fact that the elimination proceeds by a syn pathway:

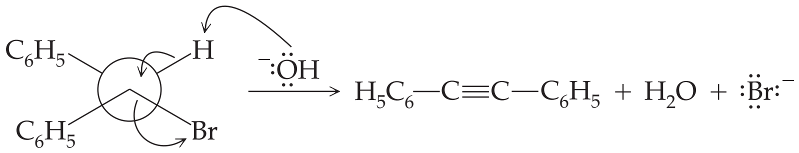

Thus, the stereochemistry of the reactant used necessitates somewhat higher temperatures for the second elimination reaction.

SEMIMICROSCALE EXPERIMENTAL PROCEDURE

(The microscale reaction is increased by a factor of 5.)

Estimated time to complete the reaction: 1.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
meso-Stilbene dibromide	340.07	400 mg	1.2	241 dec		
Potassium hydroxide	56.11	387 mg	6.9	360		
Triethylene glycol	150.18	2 mL			278	1.124

Reagents and Equipment. Weigh and place 400 mg (1.2 mmol) of *meso*-stilbene dibromide and 387 mg (6.9 mmol) of KOH flakes in a 10-mL Erlenmeyer flask containing a magnetic stir bar. Using a graduated cylinder, measure and add 2 mL of triethylene glycol to the flask.

Reaction Conditions. Place the reaction flask in a ***preheated*** sand bath set at a temperature of 190–195 °C, and stir the reaction for 7–8 min.

Isolation of Product. Allow the resulting dark-colored reaction mixture to cool to approximately 40–50 °C (water bath), and then add 5.0 mL of water. Now place the flask in an ice bath for 15 min. Collect the solid product by filtration under reduced pressure using a Hirsch funnel (➡).

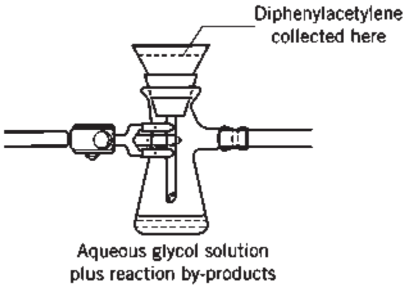
Diphenylacetylene collected here
Aqueous glycol solution plus reaction by-products

Purification and Characterization. Rinse the product crystals with two 1-mL portions of cold 70% ethanol and air-dry them on a porous clay plate or on filter paper. These crystals can be recrystallized from 95% ethanol (~2.0 mL). If desired, the product can be further purified by a second crystallization from 95% ethanol using the Craig tube.

Weigh the recrystallized product and calculate the percent yield. Determine the melting point and compare your result with the literature value. Obtain IR and NMR spectra of the material and compare them with those recorded in the literature (*The Aldrich Library of IR Spectra, The Aldrich Library of NMR Spectra,* and/or the corresponding spectral data available online (e.g., SciFinder Scholar)).

NOTE. If you have synthesized the diphenylacetylene from benzaldehyde, calculate the overall yield to this point in the synthesis of hexaphenylbenzene. Base these calculations on the starting amount of benzaldehyde.

Chemical Test. The ignition test for aromatic groups indicates the presence of the phenyl groups. Decolorization of a Br_2–CH_2Cl_2 solution should give a positive test for unsaturation (see Chapter 9).

OPTIONAL MACROSCALE AND MICROSCALE PREPARATIONS

Macroscale Reaction Procedure. (This reaction is scaled up by a factor of 10 over the microscale procedure.)

The procedure is similar to that outlined above with the following exceptions:

HOOD

1. Carry out the reaction in a 25-mL Erlenmeyer flask containing a boiling stone. Run the reaction in the **hood.**

2. Increase the reagent and solvent amounts approximately twofold over the semimicroscale procedure, as indicated here.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
meso-Stilbene dibromide	340.07	800 mg	2.4	241 dec		
Potassium hydroxide	56.11	756 mg	13	360		
Triethylene glycol	150.18	4 mL			278	1.124

3. After cooling the reaction mixture, add 10 mL of water.

4. Rinse the product crystals with two 1-mL portions of cold 70% ethanol. They can be recrystallized from 95% ethanol (~5.0 mL).

Microscale Reaction Procedure. The procedure is similar to that outlined above with the following exceptions:

1. Carry out the reaction in a 3-mL conical vial containing a boiling stone.
2. The reaction is heated in a sand bath at 190 °C for 5 min.
3. Decrease the amounts of reagents and solvents as given here.

Physical Properties of Reactants						
Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
meso-Stilbene dibromide	340.07	80 mg	0.24	241 dec		
Potassium hydroxide	56.11	75 mg	1.3	360		
Triethylene glycol	150.18	0.4 mL			278	1.124

4. After cooling of the reaction mixture to 40–50 °C (in a water bath), add 1.0 mL of water, and place the reaction vessel in an ice bath for 15 min.

5. Rinse the product crystals with one 0.25-mL portion of cold 70% ethanol. The alkyne can be recrystallized from 95% ethanol (~0.5 mL).

QUESTIONS

7-28. Both (*E*)- and (*Z*)-2-chlorobutenedioic acids dehydrochlorinate to give acetylene dicarboxylic acid:

$$HO_2C-C(Cl)=CH-CO_2H \rightarrow HO_2C-C\equiv C-CO_2H$$

The *Z* acid reacts about 50 times faster than the *E* acid. Explain.

7-29. Compounds containing a carbon–carbon triple bond undergo the Diels–Alder reaction. Formulate the product formed by the reaction of (*E*,*E*)-1,4-diphenyl-1,3-butadiene with diethyl acetylenedicarboxylate.

7-30. Alkynes can be hydrated in the presence of acid and $HgSO_4$ by electrophilic addition of a molecule of water to the triple bond. The reaction proceeds by way of a carbocation intermediate. Hydration of acetylene (ethyne) produces acetaldehyde (ethanal). Outline the steps that occur in this transformation.

7-31. Use the IR tables to locate the absorption bands of the stretching frequencies of the alkyne C—H bond, the alkyne C≡C bond, and the alkene C—H bond. Using these data, explain how you would distinguish between 1-butyne, 2-butyne, and 2-butene.

BIBLIOGRAPHY

For a review on the preparation of alkynes see

Jacobs, T. L. *Org. React.*, **1949,** *5,* 1.

Stang, P. J.; Diederich, F., Eds. *Modern Acetylene Chemistry;* VCH: New York, 1995.

A large number of elimination reactions leading to the formation of acetylenes appear in *Organic Syntheses*. Selected references are given below:

Abbott, W. T. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 515.

Campbell, K. N.; Campbell, B. K. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 763.

Guha, P. C.; Sankaren, D.K. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 623.

Hessler, J. C. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 438.

Khan, N. A. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 967.

Le Coq, A.; Gorgues, A. *Organic Syntheses* **1980,** *59,* 10.

The synthesis of diphenylacetylene has been reported:

Smith, L. I.; Falkof, M. M. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 350.

Hexaphenylbenzene

Experiment A4$_{ab}$

Common name: hexaphenylbenzene
CA number: [992-04-1]
CA name as indexed: 1,1′:2′,1″-terphenyl, 3′,4′,5′,6′-tetraphenyl-

Purpose. This reaction completes the Sequence A experiments. The Diels–Alder reaction is used to form six-membered aromatic rings. You will carry out the decarbonylation and aromatization of an intermediate bicyclic Diels–Alder adduct. You will examine the properties of our synthetic target molecule, hexaphenylbenzene.

NOTE. If you plan to continue the synthetic sequence to hexaphenylbenzene, you should have enough of the two starting reactants to carry out the first true microscale reaction used in Sequence A. The details of this interesting Diels–Alder addition, first carried out in 1933, are described below. You may, of course, wish to study this reaction as an individual microscale experiment, in which case use the same conditions given here.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallizations (pp. 89–91)
Experiment [14]: (pp. 257–269)
Experiment [15]: (pp. 269–275)

REACTION

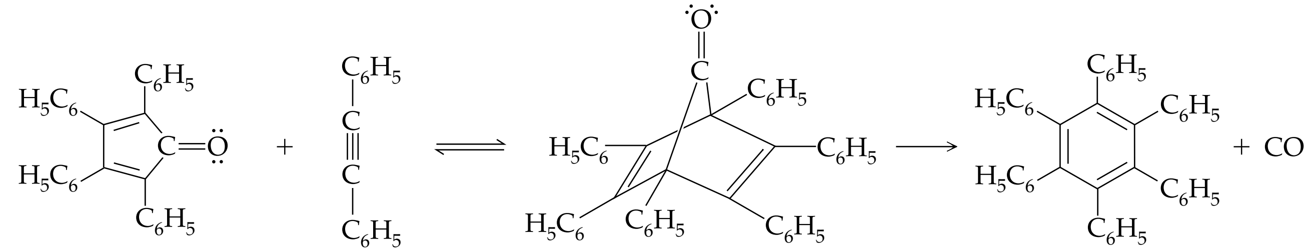

Tetraphenylcyclopentadienone Diphenylacetylene Hexaphenylbenzene

DISCUSSION

This experiment (Experiment [$A4_{ab}$]) completes the Sequence A set of seven experiments that lead to the synthesis of hexaphenylbenzene. As we have noted earlier (see introduction to Sequence A), this compound is a rather unique and interesting organic system possessing a number of unusual properties. For example, it has one of the highest known melting points for a nonionic organic molecule, 465 °C, and it is perhaps even more intriguing that it melts without decomposition. Indeed, its melting point exceeds that of all 15,000 organic compounds listed in the *CRC Handbook* for 1991–1992. Hexaphenylbenzene also contains particularly novel stereochemistry as discussed in the introduction.

The Diels–Alder reaction is one of the most useful synthetic tools in organic chemistry. It is an example of a cycloaddition reaction between a conjugated diene and a dienophile, which leads to the formation of six-membered rings. Here, the initial bicyclic Diels–Alder adduct can undergo a reaction that is the reverse of a concerted cycloaddition reaction between a benzene ring and the lone electron pair on the carbon of carbon monoxide. This retro cycloaddition is thermodynamically favored here because the retro reaction generates an aromatic system, along with the quite stable carbon monoxide molecule. Under the high-temperature conditions used in this experiment, the initial bicyclic Diels–Alder adduct is quickly decarbonylated to yield hexaphenylbenzene and is not itself isolated.

By varying the nature of the diene and dienophile, a very large number of structures can be prepared using the Diels–Alder reaction. In the majority of cases, carbocyclic rings are generated, but ring closure can also occur with reactants containing heteroatoms. This leads to the synthesis of compounds containing heterocyclic rings. For further and more detailed discussion of the Diels–Alder reaction see Experiments [14] and [15].

In the present reaction, an excess of diphenylacetylene is used to ensure that all the tetraphenylcyclopentadienone is consumed in the reaction, since diphenylacetylene is far easier to separate from hexaphenylbenzene in the purification steps.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.0 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)
Tetraphenylcyclopentadienone	384.48	100 mg	0.26	220–221
Diphenylacetylene	178.23	100 mg	0.56	61
Hexaphenylbenzene	534.66			465

Reagents and Equipment. In a 13 × 100-mm Pyrex test tube, place 100 mg (0.26 mmol) of tetraphenylcyclopentadienone and 100 mg (0.56 mmol) of diphenylacetylene. Then transfer to the test tube about 1 mL of high-boiling silicone oil (calibrated Pasteur pipet). Clamp the test tube at a slight angle, facing it away from both yourself and your laboratory neighbors.

Reaction Conditions. Bring the mixture gently to a boil over a 3- to 5-min period by heating the test tube with the moving flame of a microburner. On melting, the reagents dissolve in the hot silicone oil to yield a dark red-purple solution. Continue to gently boil the solution for an additional 10 min. During this latter period, the deeply colored solution fades and hexaphenylbenzene begins to separate from solution as tan crystals.

Isolation of Product. After heating for 15 min, cool the test tube to room temperature and add 4 mL of hexane, with stirring, to dilute the silicone oil and any unreacted starting material. The crude, precipitated hexaphenylbenzene is then collected by filtration on a Hirsch funnel (➡).

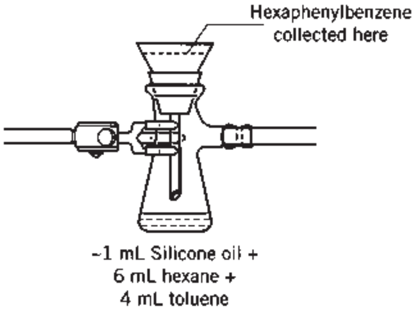

Purification and Characterization. Wash the filter cake with 2 mL of hexane to yield tan crystals of the addition product. Then wash it twice with 2-mL portions of cold toluene to yield white crystalline hexaphenylbenzene. Air-dry the product on filter paper or a porous clay plate. Weigh the Diels–Alder adduct and calculate the percent yield. Obtain an IR spectrum and compare it to that of an authentic sample. The melting point of this material is well over 400 °C, therefore, melting point determinations with apparatus that use oil baths should **not** be attempted. The best approach, if a melting point is required, is to carry out an evacuated melting point determination with one of the metal heating block systems that accept the normal capillaries, *but remember to first check the maximum temperature reading on the thermometer used in the apparatus* (see evacuated melting points, Chapter 4, pp. 51–52).

CAUTION

CAUTION

If necessary, the product can be recrystallized in a Craig tube from diphenyl ether (5–10 mg maximum, since this very high-melting material is very insoluble even in this high-boiling aromatic ether; recrystallization is rather difficult).

This experiment completes the seven-step Sequence A synthesis of hexaphenylbenzene from benzaldehyde. Calculate the overall yield based on both the earlier calculations for each pathway (for the diene and the dienophile used in the final Diels–Alder reaction), see Experiments [$A3_a$] and [$A3_b$], and the actual yield for this step in the synthesis.

QUESTIONS

7-32. What starting materials would you use to prepare each of the following compounds by the Diels–Alder reaction?

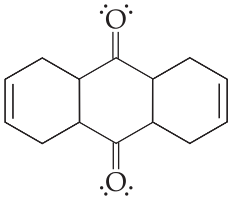

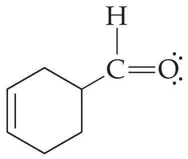

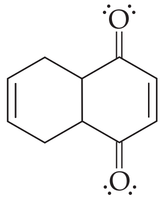

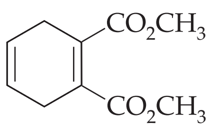

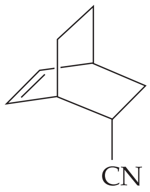

7-33. Diels–Alder reactions with benzene are rare, and require a very reactive (electron-deficient) dienophile, because benzene is a rather unreactive diene. Two are shown below. Give the structures of the product produced in each reaction.

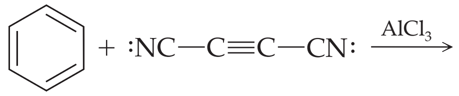

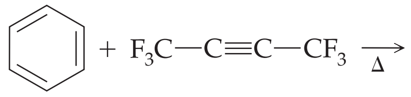

7-34. Shown below are two heteroatom compounds that undergo the Diels–Alder reaction. Formulate the product obtained in each reaction.

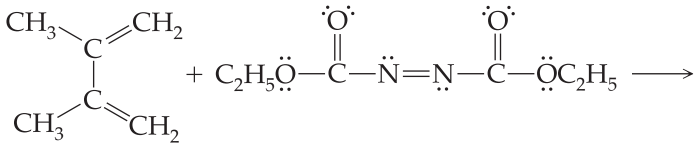

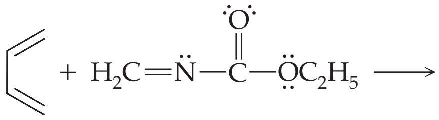

BIBLIOGRAPHY

Review articles:

Bastide, J.; Henri-Rousseau, O. In *The Chemistry of the Carbon–Carbon Triple Bond;* Patai, S., Ed.; Wiley: New York, 1978, Part 1, p. 447.

Butz, L. W.; Rytina, A. W. *Org. React.* **1949,** *5,* 136.

Carruthers, W. *Cycloaddition Reactions in Organic Synthesis;* Pergamon Press: New York, 1990.

Holm, H. L. *Org. React.* **1948,** *4,* 60.

Kloetzel, M. C. *Org. React.* **1948,** *4,* 1.

Norton, J. A. *Chem. Rev.* **1942,** *31,* 319.

Sauer, J. *Angew. Chem. Int. Ed. Engl.* **1966,** *5,* 211; *Ibid.,* **1967,** *6,* 16.

Stang, P. J.; Diederich, F., Eds. *Modern Acetylene Chemistry;* VCH: New York, 1995.

An *Organic Syntheses* preparation using tetraphenylcyclopentadienone in a Diels–Alder reaction to obtain tetraphenylphthalic anhydride has been recorded:

Grummitt, O. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 807.

The Stepwise Synthesis of Nylon-6,6

Purpose. The important industrial polymer, nylon-6,6, is prepared by the technique of step-growth polymerization. The physical properties of the polymer are examined. The two monomers used in the polymerization are synthesized.

Background of an Industrial Polymer. The type of polymerization used in the nylon preparation described in this series of experiments is called "step-growth" polymerization. The technique uses two different difunctional monomers that undergo ordinary organic reactions. In the present case an acid chloride is treated with an amine to produce an amide linkage.

Nylon is a polyamide. In industry it is produced by reaction of two difunctional monomers (or comonomers): a dicarboxylic acid and a diamine. The polymer that you are going to study is of great historical significance in polymer chemistry, because it was the first of the polyamides to be recognized as possessing excellent physical properties for forming very strong fibers. Nylon-6,6 was, in fact, the first commercially produced synthetic polyamide. The "6,6" nomenclature refers to the number of carbon atoms in each of the two comonomers. Industrially, nylon-6,6 is prepared from 1,6-hexanediamine (hexamethylenediamine) and hexanedioic acid (adipic acid):

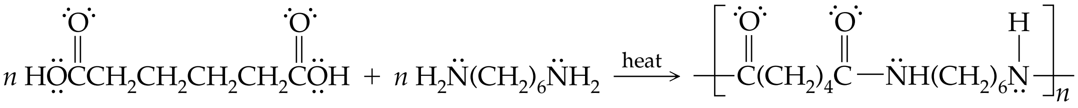

$$n\,\mathrm{HO\overset{O}{\overset{\|}{C}}CH_2CH_2CH_2CH_2\overset{O}{\overset{\|}{C}}OH} + n\,\mathrm{H_2N(CH_2)_6NH_2} \xrightarrow{\text{heat}} \left[\mathrm{\overset{O}{\overset{\|}{C}}(CH_2)_4\overset{O}{\overset{\|}{C}}{-}NH(CH_2)_6\overset{H}{\overset{|}{N}}}\right]_n$$

Adipic acid 1,6-Hexanediamine Nylon-6,6

In the industrial process, the diacid and diamine are mixed to form the corresponding amine salt (hexamethylene diammonium adipate), which is then heated under steam pressure (250 psig) at 275 °C to form the amide bonds. The resulting polymer has an average molecular weight of about 10,000, with an average of over 400 repeating monomer units in each molecule of polymer and a melting point of about 150 °C. Fibers can be drawn from the melted polymer by a "cold-drawing" technique. This method of drawing fibers physically orients the polymer molecules into linear chains that are stabilized by the presence of hydrogen bonds between C═O and the N—H groups of adjacent chains, and the strength of the fiber is thereby increased. The synthetic polyamide linkages in the various forms of nylon are very similar (identical in some cases) to those found in proteins. For example, silk fibers gain their great strength from this type of interaction.

Numerous combinations of diacids and diamines have been evaluated as fiber materials. However, only a few have reached commercial production, which depends on low-cost, easy-to-access intermediates, and satisfactory general and physical properties of the polymer. One such group of materials are the "Aramid" class of fibers, which are prepared from aromatic monomers (see Experiment [B3], Question 7-47). One trade name for a material prepared from these type of fibers is Nomex. It has a high degree of heat and flame resistance. Race car driving suits are made from it, and it is also used as an insulator in the space shuttles.

Nylon-6,6 was first synthesized in 1899 by Gabriel and Maas in Germany. It was not until 1929, however, that the substance was shown to possess practical commercial properties. It was Carother's research program on polyamides at DuPont that made the major discoveries that initiated the world's polymer industry. DuPont began production of nylon in October 1939, and the first nylon stockings were manufactured in May 1940. By 1950, 14 chemical plants in 10 countries produced 55,000 metric tons of polyamide fiber. By 1980 worldwide production had expanded to 3.05×10^6 tons, with about one-third of the polymer synthesized in the United States.

Thus, you should appreciate that the chemical industry carries out organic reactions on massive quantities of material for use in today's highly technological society. The discovery and characterization of these materials all starts in the research laboratory, with many of them initially prepared in microscale quantities. One of the great triumphs of our technology has been the successful scaleup of synthetic organic reactions, but that is a story for another day.

Experiment B1

Oxidation of Cyclohexanol: Adipic Acid

Common name: adipic acid
CA number: [124-04-9]
CA name as indexed: hexanedioic acid

Purpose. You will carry out the nitric acid oxidation of cyclohexanol to obtain adipic acid, an intermediate in the route to prepare nylon, a polyamide.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Standard Experimental Apparatus: Reflux Apparatus (pp. 23–24)

REACTION

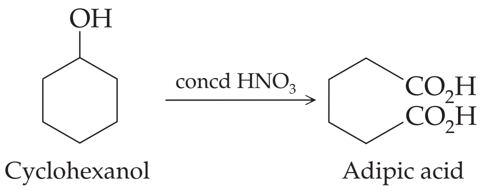

DISCUSSION

Industrially, the production of adipic acid is a two-step sequence. The main route involves the oxidation of cyclohexane with air to form a mixture of cyclohexanol and cyclohexanone. This mixture is then further oxidized using nitric acid, oxygen, and a Cu–V catalyst to yield the acid.

Ninety percent of all the synthetically produced adipic acid is used in the manufacture of nylon-6,6. In the United States, 1.9×10^6 tons/year of adipic acid were manufactured in 1992. In 2004, DuPont alone, produced 1.1×10^6 tons of adipic acid accounting for 38% of the world's total. In the early years of nylon production, adipic acid was also used to prepare the 1,6-hexanediamine (more commonly known as hexamethylenediamine) comonomer. Treatment of the adipic acid with ammonia gave hexanedinitrile, which, on catalytic hydrogenation, produced the diamine. This monomer is now generally obtained from 1,3-butadiene. A recent DuPont industrial process involves direct *regioselective* addition of two molecules of HCN to the diene, in the presence of a transition metal catalyst, to produce the dinitrile intermediate.

The oxidation of cyclohexanol by concentrated nitric acid is mechanistically complex. A reasonable mechanistic route to the dicarboxylic acid is given here. The first stage of the oxidation is considered to proceed by a mechanism similar to that found in chromic acid oxidations of alcohols (see Experiment [33]). The reaction here involves the initial formation of a nitrate ester intermediate, which, under the reaction conditions, cleaves by proton abstraction to form the ketone. This reaction is accompanied by reduction of the nitrate to nitrite. The proton transfer may involve a cyclic intramolecular rearrangement during the oxidation–reduction cleavage step. A likely mechanism is outlined below:

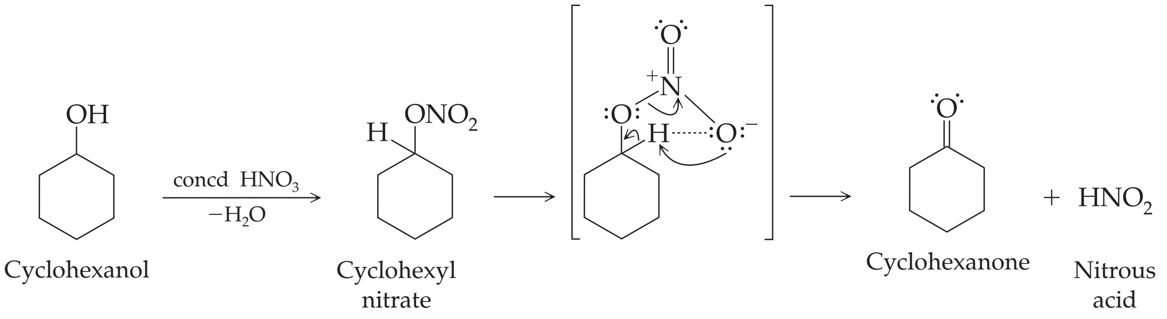

The next stage of the reaction can be viewed as a further oxidation to yield a diketone. This stage is initiated by nucleophilic attack on a nitronium ion (NO^+) derived from either the nitric or nitrous acid. The nucleophile is the enol tautomer of the ketone, and the reaction forms an α-nitrosoketone, which is in tautomeric equilibrium with a mono-oxime. This species rapidly hydrolyzes under acidic conditions to yield an α-diketone intermediate. This sequence is shown here:

$$H^+ + HNO_2 \rightleftharpoons NO^+ + H_2O$$

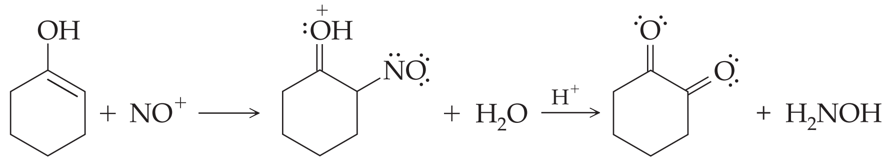

Under strongly acidic conditions, the diketone (these highly electrophilic systems are reactive toward weak nucleophiles) likely undergoes nitrate addition, which is followed by attack of water, ring opening, and reduction of nitrate to nitrite. All this activity ultimately leads to the formation of the desired compound, the open-chain dicarboxylic acid, adipic acid:

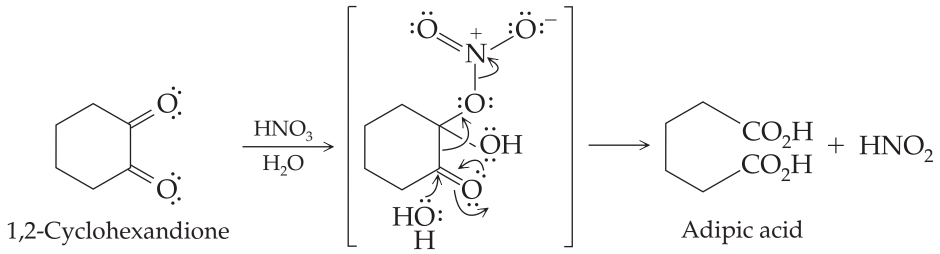

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Cyclohexanol	100.16	1.0 mL	9.6	25.1	161.1	0.96
Concd HNO_3	31	2.0 mL				

Reagents and Equipment. Using a graduated glass pipet, measure and add 2.0 mL of concentrated nitric acid to a 10-mL round-bottom flask containing a boiling stone.

HOOD

CAUTION: Nitric acid is very corrosive. Dispense the material in the ***hood.***

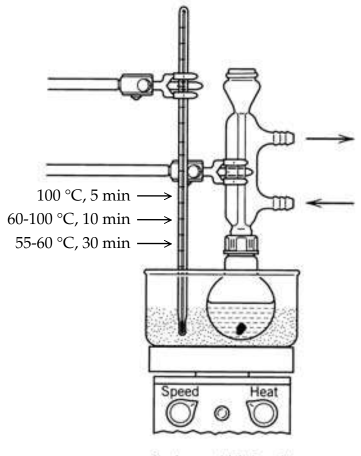

2 mL concd Nitric acid,
1 mL cyclohexanol

Attach the flask to a water-cooled reflux condenser, place the assembly in a sand bath, and heat the acid solution to 55–60 °C. Now add dropwise, using a calibrated 9-in. Pasteur pipet inserted down the throat of the reflux condenser, 1.0 mL of cyclohexanol at a rate of one drop every 30 s. *(Gently swirl the reaction mixture in the bath after each addition.)* The slow addition is necessary to control the reaction temperature (⬅).

Reaction Conditions. Maintain the sand bath temperature at 55–60 °C for 30 min after the addition of cyclohexanol is complete. Now gradually raise the sand bath temperature to 100 °C over 10 min, and then maintain this temperature for an additional 5 min.

Isolation of Product. Remove the assembly from the hot sand bath and allow the solution to cool to room temperature. Detach the round-bottom flask, and clamp it in an ice bath for 5–10 min. Collect the light yellow crystals by vacuum filtration using a Hirsch funnel.

CAUTION: This solution is still strongly acidic.

Wash the product crystals with 200 μL portions of *ice-cold* water until the crystals turn white.

Purification and Characterization. Dry the adipic acid crystals in an oven at 110–125 °C for 10 min. Weigh the product and calculate the percent yield. Determine the melting point and obtain an IR spectrum. Compare your results to those recorded in the literature.

Chemical Test. Add several crystals (~5 mg) of the adipic acid to 1 mL of a 10% aqueous solution of sodium bicarbonate. Does evolution of CO_2 indicate the presence of a carboxylic acid?

QUESTIONS

7-35. Which of the following compounds is the stronger acid: $CF_3CH_2CO_2H$ or $CH_3CH_2CO_2H$? Explain.

7-36. What spectroscopic method would you use to unambiguously distinguish between the following isomeric acids? Give an explanation of how the spectra are interpreted to give you an assignment for each compound.

(a) $CH_3(CH_2)_3CO_2H$
(b) $(CH_3)_2CHCH_2CO_2H$
(c) $(CH_3)_3CCO_2H$

7-37. Indicate how you would use both ^{1}H and ^{13}C NMR spectroscopy to tell the difference between the following isomeric carboxylic acids:

(a) $HO_2CCH_2CH_2CO_2H$
(b) $CH_3CH(CO_2H)_2$

7-38. The two carboxyl groups in 3-chlorohexanedioic acid are not equivalent and thus have different dissociation constants. Which carboxylic group is more acidic? Explain.

7-39. Write an equation for the formation of the salt that could be formed by reaction of one molecule of adipic acid with two molecules of ethylamine.

BIBLIOGRAPHY

For detailed information on the production and use of adipic acid see

Kirk-Othmer Encyclopedia of Chemical Technology, 5th ed., Vol. 1, Willey-VCH: New York, 2004.

For oxidation methods in organic chemistry see

Haines, A. H. *Methods for the Oxidation of Organic Compounds;* Academic Press: New York, Vol. 1, 1985, Vol. 2., 1988.

Hudlicky, M. *Oxidations in Organic Chemistry;* American Chemical Society: Washington, DC, 1990.

Mijs, W. J.; de Jonge, C. R. H. I. *Organic Synthesis by Oxidation with Metal Compounds;* Plenum: New York, 1987.

Taber, D.F. *Organic Synthesis: State of the Art 2003–2005;* Wiley: New York, 2006.

———.*Organic Synthesis: State of the Art 2006–2007;* Wiley: New York 2008.

The synthesis of adipic acid is given in *Organic Syntheses:*

Ellis, B. A. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 18.

Preparation of an Acid Chloride: Adipoyl Chloride

Experiment B2

Common name: adipoyl chloride
CA number: [111-50-2]
CA name as indexed: hexanedioyl dichloride

Purpose. Adipic acid is converted to its corresponding acid chloride by reaction with thionyl chloride. The experiment will help you further understand the nucleophilic substitution reaction pathway by which carbonyl-containing compounds undergo reaction.

Prior Reading

Standard Experimental Apparatus: Reflux Apparatus (pp. 23–24)
Collection or Control of Gaseous Products: (pp. 105–107)

REACTION

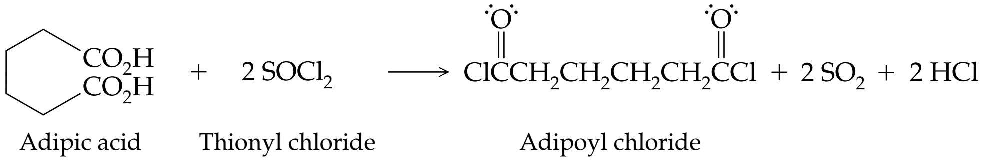

DISCUSSION

Carboxylic acids react with thionyl chloride ($SOCl_2$) to produce the corresponding acid chlorides, as shown in the above reaction. Thionyl chloride is an attractive reagent due to its low cost, and the fact that both byproducts produced in the reaction are gases. Thus, the reaction is driven to completion by the evolution of HCl and SO_2, and a nearly pure acid chloride is obtained. The major drawback to the reaction is that it produces a strong acid (HCl) and thus cannot be used with compounds that are acid sensitive. Oxalyl chloride is often used as an alternative reagent.

The reaction proceeds by a nucleophilic acyl substitution pathway. With thionyl chloride, a chlorosulfite intermediate is generated. Thus, the —OH group is converted into a relatively good leaving group. The chlorosulfite intermediate then undergoes attack by the chloride ion at the carbonyl carbon to yield the final product. The sequence is shown here:

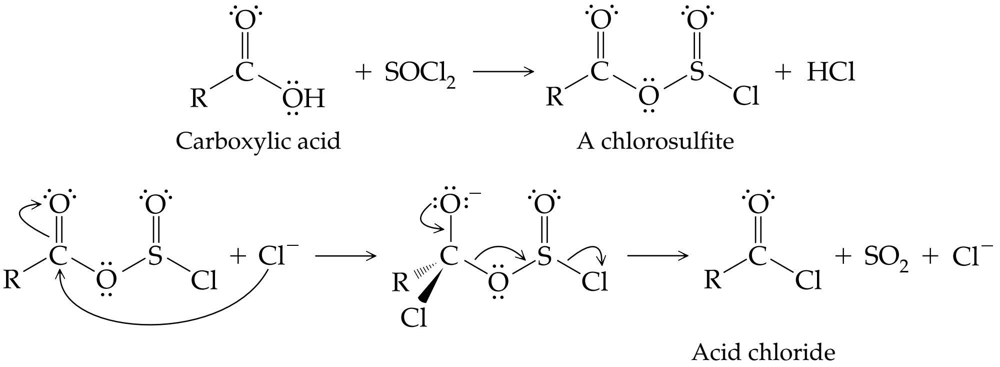

Acid halides are important intermediates and they are used extensively for the conversion of carboxylic acids into other derivatives. For example, acid halides can be used to prepare (in addition to amides): anhydrides, esters, aldehydes, and ketones. Acid halides readily undergo reaction with water (hydrolysis) to form the corresponding carboxylic acid. For this reason the reaction system must be protected from atmospheric moisture when acid halides are formed and/or used.

In the present sequence leading to the formation of nylon, the adipoyl chloride provides a reactive species, which, when treated with a diamine, forms the desired amide linkage inherent to nylon.

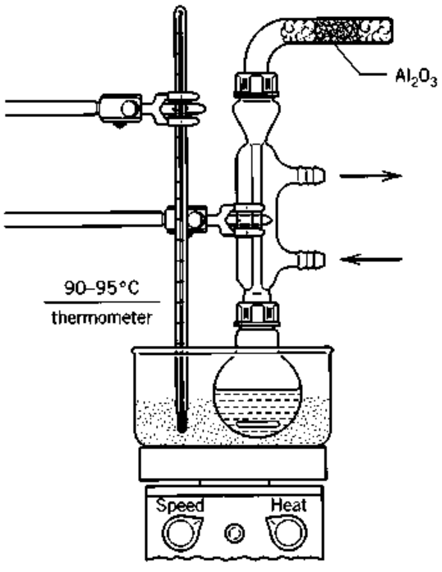

Adipic acid, 500 mg +
thionyl chloride, 0.75 mL
reflux conditions,
10-mL RB flask

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Adipic acid	146.14	500 mg	3.4	153	265	1.35
Thionyl chloride	118.97	0.75 mL	10.4	−105	78.8	1.66

Reagents and Equipment. Place and dry in an oven at about 125 °C for 30 min, a 10-mL round-bottom flask, a water-cooled reflux condenser, and a drying tube packed with glass wool and alumina. After removing them from the oven, place these items in a desiccator and allow them to cool to room temperature. Measure and place 500 mg (3.4 mmol) of adipic acid and 0.75 mL (10.4 mmol) of thionyl chloride into the 10-mL flask (⬅).

NOTE. Store the dry adipic acid and the container of thionyl chloride in a desiccator. Dispense the thionyl chloride in the ***hood*** *using an automatic delivery pipet.* HOOD

CAUTION: Thionyl chloride is both corrosive and a lachrymator.

Add a magnetic stir bar to the flask, quickly attach the flask to the reflux condenser protected by the drying tube, and mount the assembly in a sand bath.

NOTE. If a larger scale reaction is to be run outside of the ***hood,*** *it will be necessary to construct a gas trap to control the evolution of the SO_2 and HCl gases (see Prior Reading).* HOOD

Reaction Conditions. With stirring, heat the reaction mixture to 90–95 °C within 5 min. Continue to heat the system within this temperature range for an additional 1 h.

NOTE. Heating the solution above 95 °C causes decomposition of the product. The presence of unreacted adipic acid in the flask indicates incomplete reaction.

Isolation and Characterization. The progress of the reaction may be followed by IR analysis. With a glass capillary, remove a small sample from the flask and obtain the spectrum of the material using the capillary film technique. Remove the sample from the instrument sampling compartment immediately following the spectral scan to prevent the HCl gas buildup from damaging the instrument. The reaction is considered incomplete if the IR spectrum displays a weak band on the low-wavenumber side of the acid halide carbonyl peak.

If the reaction is determined to be incomplete after 1 h, add an additional 200 μL of $SOCl_2$ and continue to heat the mixture until IR analysis (the disappearance of the weak side band) indicates that completion has occurred. The adipoyl chloride is quite labile, and therefore, it is not purified further, but it is used directly in the preparation of nylon as described in Experiment [B3].

QUESTIONS

7-40. Diagram a complete mechanistic sequence showing the hydrolysis reaction of ethanoyl chloride with water to form ethanoic acid.

7-41. Give an explanation of why acid chlorides are more reactive toward nucleophilic substitution than are the corresponding ethyl esters. *Hint:* Consider the nature of the leaving group and the rate-determining step in an addition–elimination sequence.

7-42. What major organic product would you expect to be formed when acetyl chloride reacts with each of the following reagents?

(a) H_2O **(c)** 1-Butanol and pyridine
(b) NH_3 (excess) **(d)** CH_3COOH^-, Na^+

7-43. Acid chlorides are used extensively as electrophiles in the Friedel–Crafts reaction to prepare aromatic ketones. The reaction involves the treatment of an aromatic hydrocarbon with an acyl chloride in the presence of a Lewis acid, such as aluminum chloride (see Experiment [27]). Using this reaction, outline the reaction sequence you would use to prepare

(a) Ethyl phenyl ketone
(b) Benzophenone

7-44. Formulate a suitable mechanism to account for the reaction of thionyl chloride with carboxylic acids to yield the corresponding chlorosulfite (see the reaction diagrammed in the discussion section). *Hint:* The first step is the nucleophilic attack of the oxygen of the —OH group of the acid on the sulfur atom of the S═O group of the thionyl chloride.

BIBLIOGRAPHY

Adipoyl chloride is prepared as an intermediate in several preparations reported in *Organic Syntheses:*

Fuson, R. C.; Walker, J. T. *Organic Syntheses*; Wiley: New York, 1943; Collect. Vol. II, p. 169.

Guha, P. C.; Sankaran, O. K. *Organic Syntheses*; Wiley: New York, 1955; Collect. Vol. III, p. 623.

See your organic textbook for an introduction to this reaction and its scope in synthesis. For example,

Solomons, T. W. G.; Fryhle, C. B. *Organic Chemistry*, 9th ed., Wiley: New York, 2008, p. 794.

For a review on the preparation and reactions of acyl halides see

Ansell, M. F. in *The Chemistry of Acyl Halides;* Patai, S., Ed.; Wiley: New York, 1972, p. 35.

Experiment B3

Preparation of a Polyamide: Nylon-6,6

Common names: nylon-6,6, polyhexamethylene adipamide
CA number: [32131-17-2]
CA name as indexed: poly[imino(1,6-dioxo-1,6-hexanediyl)imino-1,6-hexanediyl]

Purpose. The polyamide, nylon, is prepared by the step-growth condensation polymerization of adipoyl chloride with 1,6-hexanediamine. An interfacial (emulsion) polymerization technique is used to generate nylon fibers.

REACTION

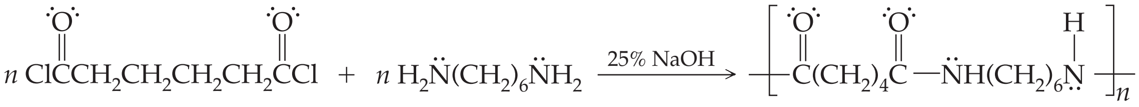

$$n\,\mathrm{Cl}\overset{\overset{O}{\|}}{C}CH_2CH_2CH_2CH_2\overset{\overset{O}{\|}}{C}Cl + n\,H_2N(CH_2)_6NH_2 \xrightarrow{25\%\ NaOH} \left[\overset{\overset{O}{\|}}{C}(CH_2)_4\overset{\overset{O}{\|}}{C}-NH(CH_2)_6\overset{\overset{H}{|}}{N} \right]_n$$

Adipoyl chloride 1,6-Hexanediamine Nylon-6,6

DISCUSSION

The preparation of nylon outlined in this experiment is not the industrial method (see initial discussion). The use of the reactive diacid chloride reagent allows one to carry out the step-growth polymerization reaction under very mild conditions more convenient to the instructional laboratory. The interfacial (emulsion) polymerization technique used consists of dissolving the adipoyl chloride reagent in a water-immiscible solvent (cyclohexane) and bringing this solution into contact with an aqueous solution of the diamine. A thin film forms at the *interface* of the two solvents as the condensation reaction proceeds. A "rope" of nylon polymer can be pulled from the interface of the two solvents because the film is continuously generated as the reaction occurs. This polymer has an average molecular weight of ~10,000! In this particular experiment, about 5–7 meter lengths of nylon polymer can be obtained. This particular synthesis of nylon is often used in lecture demonstrations and chemical magic shows.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 0.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	*d*
Adipoyl chloride	183.05	~622 mg	3.4		
1,6-Hexanediamine (5% aq)	116.21	8 mL		41–42	1.259
Cyclohexane	84.16	8 mL			

Reagents and Equipment. Transfer the clear solution of adipoyl chloride prepared in Experiment [B2] to a 50-mL beaker using a Pasteur pipet. Rinse the flask with 2 mL of cyclohexane and transfer this rinse to the same beaker. Add an additional 8.0 mL of cyclohexane. Now slowly add 8.0 mL of a 5% aqueous solution of 1,6-hexanediamine containing eight drops of 25% NaOH solution.

NOTE. Add the solution using a Pasteur pipet, taking care to run it down the side of the beaker.

Isolation and Characterization. Using a copper wire bent into a small hook, hook the film in the center of the beaker and draw up the nylon fiber from the solution interface. A slow, steady pull will result in long strands of the polymer. Wash the fibers in a beaker of water before handling them.

QUESTIONS

7-45. Explain why the 25% NaOH solution is added to the reaction mixture.

7-46. Draw a structure to illustrate the hydrogen bonding that may occur when two polymer molecules of this polyamide are cold-drawn together.

7-47. Predict the structure of the polymer that would result in the condensation of the following reactants. These monomers are used to produce the polyamide Nomex, a high-melting material used as an insulator in space shuttles and as the fire-resistant fabric in clothing worn in race cars.

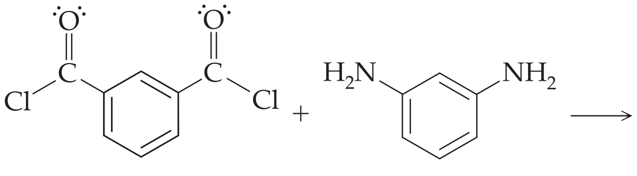

7-48. Amides undergo hydrolysis to carboxylic acids on treatment with alkali. Diagram a suitable mechanism for the conversion of benzamide to benzoic acid using sodium hydroxide as the base.

BIBLIOGRAPHY

For detailed information on the production and use of nylon see

Heckert, W. W. *J. Chem. Educ*. **1953,** *30,* 166.

Kirk–Othmer Encyclopedia of Chemical Technology, 4th ed., Vol. 19, Wiley: New York, 1996, pp. 454, 470, 485.

For information on the interfacial polymerization technique see

Sprague, B. S.; Singleton, R. W. *Text. Res. J.* **1965,** *35,* 999.

Morgan, P. W.; Kwolek, S. L. *J. Chem. Educ.* **1959,** *35,* 182.

Nikonov, V. Z.; Savinov, V. M. In *Interfacial Synthesis;* Millich, F.; Carraher, C. E., Jr., Eds.; Marcel Dekker: New York, 1977, Vol. II, Chap. 15.

Odian, G. *Principles of Polymerization,* 4th ed.; Wiley-Interscience: New York, 2004.

SEQUENCE C

The Synthesis of Sulfanilamide

Purpose. You will carry out the multistep synthesis of the epoch-making antibacterial drug, sulfanilamide. You will learn the techniques for handling moisture sensitive materials and investigate the strategies involved in solving a synthetic organic problem.

THE SULFA DRUGS

The sulfanilamides were the first effective systemic bactericides. The discovery, in the early 1930s, of these substances transformed medical care. Indeed, a major fraction of the world's population is alive today as a result of these compounds and their microbiological successors. Early recognition of the importance of these materials was underlined when their discoverer was honored by receiving the 1938 Nobel Prize in Medicine. Even 60 years later, "sulfa" drugs are still some of the most effective antibacterial substances available to physicians today.

These compounds, to which human existence owes a considerable debt of gratitude, are all derivatives of the parent material, *p*-aminobenzenesulfonamide (sulfanilamide):

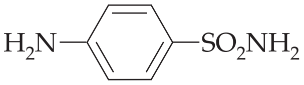

p-Aminobenzenesulfonamide
(sulfanilamide)

This weakly basic compound contains two —NH_2 groups located in quite different environments, so that each one possesses quite different chemical properties and reactivities. One is substituted directly on the aromatic ring; the other is contained in a sulfonamide group. A wide variety of these derivatives can be prepared by introducing different substituents on the nitrogen atom of the sulfonamide functional group. Any attempt at modifying the ring —NH_2 group, however, was found to completely destroy the biological activity.

Prontosil, *p*-[(2,4-diaminophenyl)azo]benzenesulfonamide, was first synthesized in Germany in 1932 as a product of azo-dye research at I. G. Farbenindustrie. It was soon tested for chemotherapeutic activity, because it bound very strongly to protein fibers. In 1933, it became the first drug to effectively cure blood stream infections. The early investigations of the biological activity of Prontosil required direct testing on animal infections, since the drug appeared to be inactive in vitro.

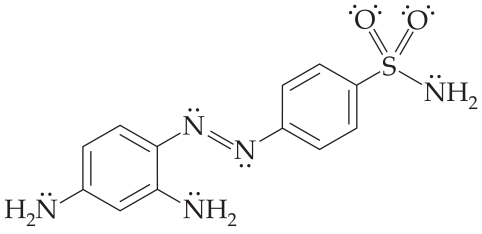

Prontosil

In 1935, Trefouels (Pasteur Institute) established that Prontosil breaks down in the body to yield, as one of the metabolites, sulfanilamide. It was shown later that year, by Fourneau, that the biological activity of Prontosil was entirely contained in the sulfanilamide section of the molecule, and that no biological activity was lost during the cleavage. This later discovery, and the fact that sulfanilamide itself had been earlier synthesized by Gelmo in 1908 (unfortunately by not exploring the biological activity of the substance, he missed the opportunity at a Nobel Prize), voided any patent protection, and led to a worldwide effort to synthesize a wide variety of these materials. By 1944 over 5000 sulfanilamide derivatives had been prepared and tested.

Currently the sulfa drugs are still very important therapeutic agents, but have to a large extent been replaced by antibiotics, such as the penicillins. Their low cost and general effectiveness for urinary tract infections, however, still make sulfanilamide derivatives attractive alternatives. These compounds have also found a valuable role in veterinary medicine.

Production of sulfanilamide reached a peak of 9000 metric tons/year in 1943; present production is about one-half that amount.

The following experiments in this sequence outline a novel and efficient procedure for the preparation of this important substance. The chemistry involves the trifluoroacetylation of aniline (Experiment [C1]), the chlorosulfonation of the resulting acetanilide (Experiment [C2]), which is followed by concomitant aminolysis and hydrolysis of the sulfonyl chloride intermediate (Experiment [C3]) to produce the final product, sulfanilamide:

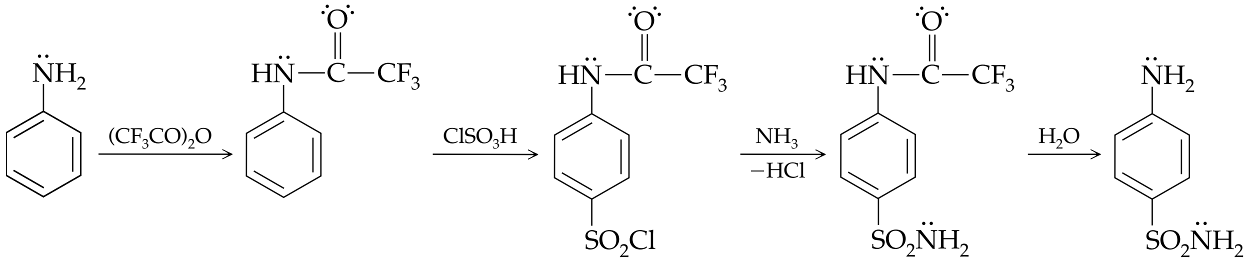

Acetylation of Aniline: 2,2,2-Trifluoroacetanilide

Experiment C1

Common name: 2,2,2-trifluoroacetanilide

CA number: [404-24-0]

CA name as indexed: acetamide, 2,2,2-trifluoro-*N*-phenyl-

Purpose. The goal of this experiment is to protect the easily oxidized amine group of aniline from electrophilic attack during the sulfonation of the benzene ring in the next step of the synthesis. This deactivation is accomplished by reducing the electron density on the amine nitrogen (and thus its nucleophilicity) by acetylating aniline with the strongly electron-withdrawing trifluoroacetyl group.You will become familiar with techniques for handling moisture-sensitive materials.

Prior Reading

Standard Experimental Apparatus: Moisture-protected Claisen head with 3- or 5-mL conical vial, arranged for syringe addition (pp. 25–26)

Technique 6B: Concentration of Solutions:
Evaporation with Nitrogen Gas (p. 102)

REACTION

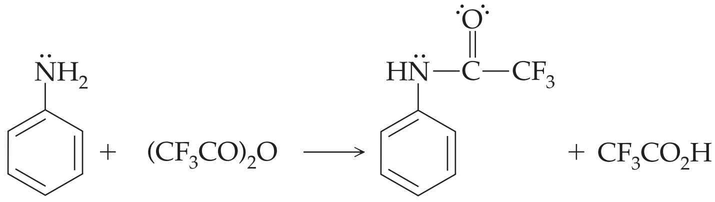

Aniline

Trifluoroacetic anhydride

DISCUSSION

Acid anhydrides react with amines, or ammonia, to yield amides. The reaction of anhydrides is similar to that of acid chlorides (see Experiment [B3]), although the anhydrides normally react at slower rate.

Acetic anhydride, $CH_3C(O)OC(O)CH_3$, is an important industrial reagent used to prepare acetamides, $CH_3C(O)NR_2$, from a variety of amines. For example, acetaminophen, an over-the-counter analgesic, is synthesized by the reaction of *p*-hydroxy-aniline with acetic anhydride:

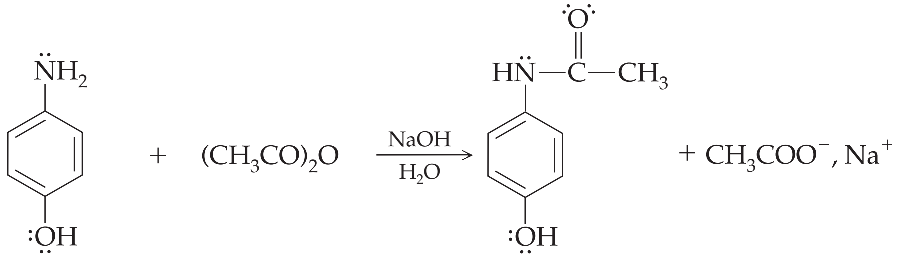

p-Hydroxyaniline

Acetic anhydride

Acetaminophen

In the present reaction, *trifluoro*acetic anhydride is used in place of acetic anhydride. The halogenated anhydride is considerably more reactive because of the presence of the three fluorine atoms on the α-carbon atom of the anhydride. The electron-withdrawing power of the fluorine substituents helps to enhance nucleophilic attack on the anhydride by the weak aromatic nucleophile. Once the amide is formed, the trifluoromethyl group makes the amide a better protecting group than the acetyl group, as the trifluoroacetyl group renders the amine lone electron pair less nucleophilic. Furthermore, the byproduct of the reaction, trifluoroacetic acid, is fairly volatile (bp 72 °C) and is therefore easily removed from the reaction mixture by evaporation. This enhanced reactivity helps avoid the use of an added base, as is required in the preparation of acetaminophen. Chemically, the procedure is simplified and the intermediate trifluoroacetanilide is usually obtained in a relatively pure condition. *This highly reactive anhydride, however, requires considerable care in its use.*

CAUTION: No moisture can remain on the surface of the clean glassware or hydrolysis may occur with explosive force!

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 0.75 h.

Physical Properties of Reactants

Compound	MW	Mass	Volume	mmol	mp (°C)	bp (°C)	*d*
Aniline	93.13	235 mg	230 μL	2.5	−6.2	184.3	1.02
Trifluoroacetic anhydride	201.04	744 mg	500 μL	3.54	−6.5	39.5	1.49
Methylene chloride	84.93		500 μL			40	1.33

Reagents and Equipment. Place and dry in an oven at about 125 °C for 30 min, a 5-mL conical vial, a Claisen head, a drying tube packed with glass wool and calcium chloride, a 1-mL glass syringe, and a $\frac{1}{2}$-dram screw-cap vial. After removal from the oven, place these items in a desiccator and allow them to cool to room temperature. Measure and place 230 μL (2.5 mmol) of aniline and 500 μL of methylene chloride in the 5-mL conical vial. Immediately attach the vial to the Claisen head equipped with a septum cap and protected by the drying tube (see Prior Reading). Cool the vial in a cold water bath.

NOTE. Dispense the aniline and methylene chloride in the ***hood*** *using automatic delivery pipets.* HOOD

CAUTION: Aniline is a highly toxic substance and a cancer suspect agent.

In the **hood,** place 500 μL of trifluoroacetic anhydride and 500 μL of methylene chloride in the dried $\frac{1}{2}$-dram vial. HOOD

NOTE. The anhydride hydrolyzes rapidly in moist air, so make the transfer quickly.

Now draw this solution into the dried syringe, and insert the needle through the septum cap of the Claisen head. Dropwise, **slowly** add this solution to the cold aniline–methylene chloride solution. SLOWLY

NOTE. Heat is produced in the reaction. If the addition is too rapid, the methylene chloride starts to reflux and fumes are emitted through the drying tube.

Reaction Conditions. After the addition is complete, allow the reaction mixture to stand at room temperature for 10 min.

Isolation of Product. Remove the reaction vial and, in the **hood,** concentrate the solution on a warm sand bath under a slow stream of nitrogen gas. The crude 2,2,2-trifluoroacetanilide intermediate is obtained as a white powder. Cap the vial. HOOD

Purification and Characterization. The crude material is not purified further, but rather used directly in the next reaction of the sequence without further characterization.

QUESTIONS

7-49. Outline a complete mechanistic sequence for the reaction to prepare acetaminophen as presented in the discussion section of this experiment.

7-50. In the preparation of amides, acid chlorides or anhydrides may be used to react with the selected amine. It is known that acid chlorides are more reactive than the corresponding anhydrides in this type of nucleophilic acyl substitution reaction. Offer a reasonable explanation for this observation.

7-51. Offer an explanation of why 2,2,2-trifluoroacetic anhydride is more reactive toward aniline than acetic anhydride in the nucleophilic acyl substitution reaction presented in this experiment.

7-52. Offer a reasonable explanation of why acetamide has a higher melting and boiling point then *N,N*-dimethylacetamide even though it has a lower molecular weight:

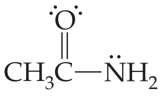

$CH_3\overset{\overset{\ddot{O}}{\|}}{C}-\ddot{N}H_2$

Acetamide
mp 82 °C
bp 221 °C

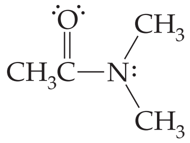

N,N-Dimethylacetamide
mp −20 °C
bp 166 °C

7-53. Outline a simple chemical test that would distinguish between ethyl benzoate and benzamide (*Hint:* See Chapter 9).

BIBLIOGRAPHY

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Part 2; Wiley: New York, 2007, Chap. 16, p. 1429 and references therein.

Selected acylation reactions in *Organic Syntheses* between anhydrides and amines:

Herbst, R. M.; Shemin, D. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 11.

Noyes, W. A.; Porter, P. K. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 457.

Wiley, R. H.; Borum, O. H. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 5.

The use of the perfluoroacetic anhydride reagent in this synthesis is reported in

Hurdis, E. C.; Yang, J. W. *J. Chem. Educ.* **1969,** *46,* 697.

Experiment C2

Chlorosulfonation of 2,2,2-Trifluoroacetanilide: *p*-(Trifluoroacetamido)benzenesulfonyl Chloride

Common name: *p*-(trifluoroacetamido)benzenesulfonyl chloride
CA number: [31143-71-2]
CA name as indexed: benzenesulfonyl chloride,
4-[(trifluoroacetyl)amino]-

Purpose. The substitution of the aniline ring is carried out by introduction of the sulfonyl group. This substitution reaction is the second stage on the route to preparing sulfanilamide. *p*-(Trifluoroacetamido)benzenesulfonyl chloride is prepared by treatment of the intermediate, 2,2,2-trifluoroacetanilide, with chlorosulfonic acid. You will gain experience at handling highly reactive moisture-sensitive reagents.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)

REACTION

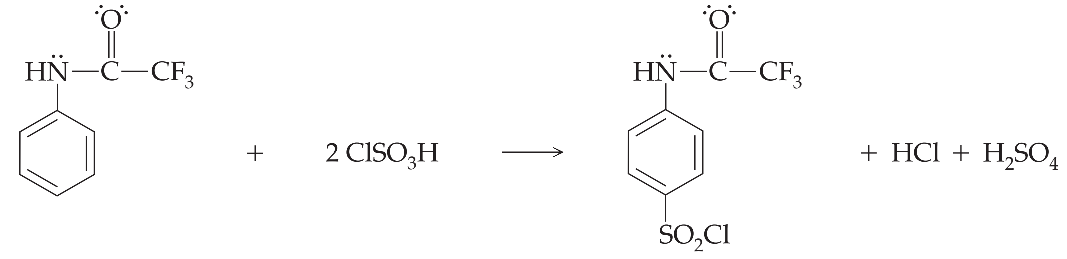

2,2,2-Trifluoroacetanilide Chlorosulfonic acid *p*-(Trifluoroacetamido)-benzenesulfonyl chloride

DISCUSSION

As shown here, the sulfonyl chloride group (—SO_2Cl) can be conveniently introduced to an aromatic ring via an electrophilic aromatic substitution reaction using chlorosulfonic acid. The reaction is usually referred to as *chlorosulfonation*. It has been determined that two equivalents of the acid are required per equivalent of the aromatic compound. In the initial attack the system first forms the corresponding sulfonic acid, which in turn is converted to the sulfonyl chloride. It is believed that the initial stage of the reaction involves SO_3 as the electrophile. It is likely that this reagent results from the establishment of the equilibrium reaction shown here:

$$ClSO_3H \rightleftharpoons SO_3 + HCl$$

The intermediate sulfonic acid is converted to its sulfonyl chloride derivative by reaction with a second equivalent of the chlorosulfonic acid:

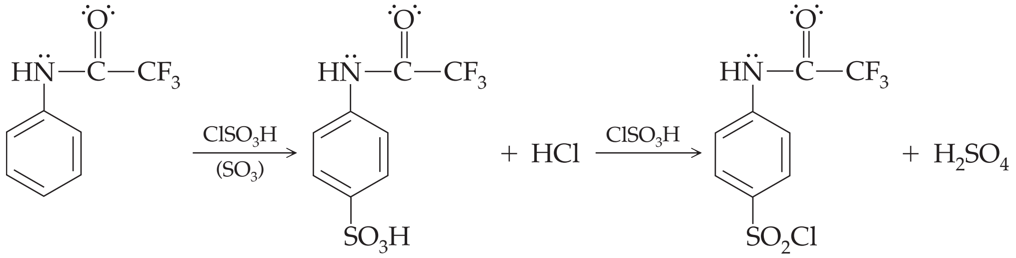

Product isolation is achieved by transferring the solution of the sulfonyl chloride onto crushed ice to precipitate the insoluble acid chloride. This isolation procedure can be used because sulfonyl chlorides are much less susceptible to hydrolysis than their carboxylic acid chloride counterparts (see Experiment [B2]). Sulfonyl chlorides *do* undergo this hydrolysis reaction, albeit rather slowly, resulting in the formation of the corresponding sulfonic acid.

Notice that the para-substituted product is formed in the reaction. Why is this observed given that two other isomers are possible (the meta- and/or ortho-products)? (See Question 7-57.)

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 0.75 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
2,2,2-Trifluoroacetanilide	189.06	478 mg (theoretical value)	2.5	87.6		
Chlorosulfonic acid	116.52	900 μL	13.7	−80	158	1.77

HOOD

NOTE. All reagent transfers must be made and the reaction must be conducted in the ***hood.***

Reagents and Equipment. Place and dry in an oven at about 125 °C for 30 min, an air condenser and a 9-in. Pasteur pipet calibrated to deliver 0.9 mL. After removal from the oven, place these items in a desiccator and allow them to cool to room temperature.

SLOWLY

Attach the vial containing the 2,2,2-trifluoroacetanilide (prepared in Experiment [C1]) to the *dry* air condenser. Measure, using the 9-in. Pasteur pipet, 0.9 mL of chlorosulfonic acid, and **slowly** add this reagent to the flask by inserting the pipet down the neck of the air condenser.

HOOD

> CAUTION: Chlorosulfonic acid is a very corrosive substance. It reacts violently with water and causes serious burns on contact with the skin. Dispense the reagent in the ***hood.*** Leave all equipment used to transfer this material in the *hood* until it is obvious that the residual reagent has reacted with the moist air (white fumes subside), and leave these materials in contact only with *glass* surfaces.

NOTE. An alternative procedure for making the transfer is to measure the amount of this reagent using an oven-dried 1-mL glass pipet and then transfer it to an oven-dried screw-cap vial. The material is then transferred to the reaction vial as described above, using the dry 9-in. pipet.

HOOD

Reaction Conditions. Place the reaction assembly on a sand bath in the **hood** and heat the mixture at 60–70 °C for 10 min.

SLOWLY
HOOD

Isolation and Characterization. Allow the reaction mixture to cool to room temperature, and then place the flask in an ice bath. Using a second, dry 9-in. Pasteur pipet, transfer the cold reaction solution **slowly** to a 10-mL beaker containing ice ($\sim\frac{1}{2}$ full) in the **hood.**

CAUTION
HOOD

> CAUTION: This is a very vigorous reaction. Use ***caution.*** Leave all equipment in the ***hood*** until it is obvious the residual reagent has reacted with the moist air (white fumes subside). Make sure that only glass surfaces are in contact with the reagent residues.

A precipitate of the product, *p*-(trifluoroacetamido)benzenesulfonyl chloride, forms at this stage. Collect the tannish-white precipitate by vacuum filtration and wash the filter cake with three 1-mL portions of ice cold water.

Allow the material to air-dry. Determine the weight of the crude product, and calculate the percent yield from aniline. Determine its melting point, and compare it to the reported value of 142–145 °C.

Chemical Test. Carry out a sodium fusion test to confirm the presence of nitrogen, sulfur, and chlorine in the product (see Chapter 9).

QUESTIONS

7-54. Outline a suitable mechanism to illustrate the reaction of benzenesulfonyl chloride with excess ethylamine to form the corresponding sulfonamide.

7-55. In reference to Question 7-54, why is an excess of ethylamine used?

7-56. Account for the fact that the amide group (—$NHCOCF_3$) is ortho and para directing.
Use resonance structures to illustrate your written explanation.

7-57. As mentioned in the discussion, offer an explanation of why only the para isomer is obtained in the chlorosulfonation reaction carried out in this experiment.

BIBLIOGRAPHY

For a summary of electrophilic aromatic substitution reactions see

Hapworth, J. D.; Waring, D. R.; Waring, M. J. *Aromatic Chemistry* Wiley: New York, 2003, Chap. 2.

For the use of chlorosulfonic acid in aromatic substitution reactions see

Gilbert, E. E. *Sulfonation and Related Reactions;* Wiley: New York, 1965, p. 84.

Rathke, M. E.; Millard, A. A. *Organic Syntheses* **1978,** *58,* 32.

Reid, J. R.; Dufresne, R. F.; Chapman, J. J. *Organic Syntheses* **1997,** *74,* 217.

Scheifele, H. J., Jr.; De Tar, D. F. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 34.

Smiles, S.; Stewart, J. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 8.

Webb, C. N. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 85.

Also see

Smith, M. B.; March, J. *Advanced Organic Chemistry*, 6th ed., Wiley: New York, 2007, Chap. 14, p. 974 and references therein.

Preparation of an Arene Sulfonamide: Sulfanilamide

Experiment C3

Common names: sulfanilamide, *p*-aminobenzenesulfonamide
CA number: [63-74-1]
CA name as indexed: benzenesulfonamide, 4-amino-

Purpose. The sequential synthesis is completed and the sulfa drug, *sulfanilamide*, is obtained by treatment of the sulfonyl chloride intermediate prepared in Experiment [C2], with an aqueous ammonia solution.[2] This interesting product is fully characterized.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallization (pp. 89–91)

---

[2] This experiment is adapted from the work of S. Danishefsky, Yale University (personal communication).

DISCUSSION

The final step in the synthesis of the target molecule, sulfanilamide, is the conversion of the sulfonyl chloride intermediate to a sulfonamide and the removal of the 2,2,2-trifluoroacetyl protecting group. These transformations are accomplished in one step because both reactions take place upon heating the protected sulfonyl chloride with aqueous ammonia. The reaction sequence is shown here:

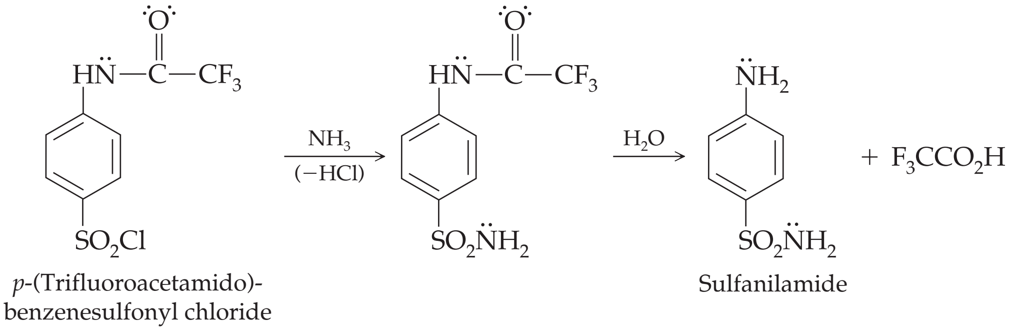

We can now see several reasons for first protecting the reactive amino group on the aniline molecule. First, to attempt the introduction of the sulfonyl group directly on aniline would very likely lead to a species (*p*-aminobenzenesulfonyl chloride) that would react with itself (act as a difunctional monomer) to produce a sulfonamide polymer:

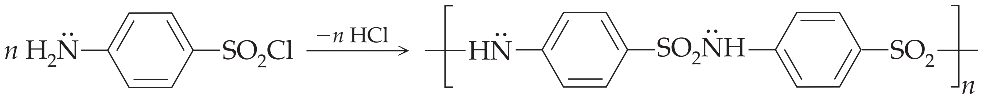

Second, because the sulfonyl group is introduced using chlorosulfonic acid, an acidic medium is present. Under these conditions, because free HCl is available (see discussion in Experiment [C2]), the amino group would be protonated and become a meta directing group. This situation would lead to the formation of the wrong isomer or, at best, a mixture of isomers.

Third, as mentioned above, the amino group is highly prone to oxidation and the unavoidable presence of traces of the strong oxidant, SO_3, would be expected to lead to considerable oxidation of the aromatic amine.

The 2,2,2-trifluoroacetyl group meets all the requirements of a good protecting group. First, the acetylation reaction goes rapidly in nonaqueous media to give a clean, dry product. Second, the protecting group is easily and rapidly removed in the final stage without disturbing the other functional groups in the molecule.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 0.75 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)
(2,2,2-Trifluoroacetamido)-benzenesulfonyl chloride	287	400 mg	1.39	142–145
Aqueous ammonia		600 μL		

CAUTION: All reagent transfers and the *reaction* must be conducted in the ***hood.***

HOOD

Reagents and Equipment. Weigh and place in a 10-mL Erlenmeyer flask 400 mg (1.39 mmol) of the *p*-(2,2,2-trifluoroacetamido)benzenesulfonyl chloride, which was prepared in Experiment [C2]. In a 10 × 75-mm test tube, prepare a solution of 0.6 mL of fresh, concentrated aqueous ammonia (ammonium hydroxide) and 0.4 mL of deionized water. Add this solution to the solid sulfonyl chloride in the Erlenmeyer flask. Now add a boiling stone. Use a glass rod to break up any lumps of the solid that may form.

CAUTION: Dispense the ammonia solution in the ***hood.***

HOOD

CAUTION: If the sulfonyl chloride reagent contains acidic impurities, a vigorous reaction may occur when the reagents are mixed.

Reaction Conditions. Place the Erlenmeyer flask on a hot (100–110 °C) sand bath in the **hood** and heat the mixture until the solid material dissolves. Agitate with a glass rod, if necessary, to assist the dissolution process. Now heat the solution to boiling for an additional minute.

HOOD

Isolation and Characterization. Remove the flask from the sand bath and allow the reaction mixture to cool to room temperature. Place the flask in an ice bath for 15–20 min. During this time light-yellow crystals of product precipitate.

Collect the crystalline product by vacuum filtration, and wash the crystals with three 0.5-mL portions of ice-cold water. Air-dry the sulfanilamide, weigh, and calculate a crude yield, both for this step and for the overall sequence based on the amount of aniline used in the first step. Determine the melting point of this material.

Recrystallize the crude material from water using the Craig tube. Air-dry the white crystals overnight, or in an oven (110 °C) for 10–15 min. Determine the melting point and compare your value to that reported in the literature. To further characterize the material obtain an IR and NMR spectrum and compare them to those found in the literature (*The Aldrich Library of Infrared Spectra, The Aldrich Library of NMR Spectra,* and/or the corresponding spectral data available online (e.g., SciFinder Scholar)).

Chemical Test. Carry out a sodium fusion test to confirm the presence of nitrogen and sulfur in the product (see Chapter 9).

QUESTIONS

7-58. Based on the results of this experiment, which is more nucleophilic: hydroxide ion (HO^-) or ammonia (NH_3)? Explain.

7-59. Outline a suitable mechanism to account for the conversion of the sulfonyl chloride group to the sulfonamide group by reaction with ammonia.

7-60. Outline a synthesis for the sulfa drug, sulfathiazole, starting from benzene. The primary aminothiazole ring system is available:

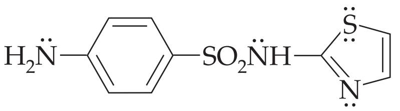

Sulfathiazole

7-61. Sulfonyl chlorides also react readily with alcohols to yield sulfonate esters:

$$\mathrm{ArSO_2Cl} + \mathrm{HOR} \xrightarrow[\text{or } 3^\circ \text{ amine}]{\text{pyridine}} \mathrm{ArSO_2OR} + \mathrm{HCl}$$

Propose a suitable mechanism to account for this reaction.

7-62. In reference to the Hinsberg test used to distinguish between primary, secondary, and tertiary amines (see Chapter 9), primary amines yield a sulfonamide that is soluble in aqueous sodium hydroxide, whereas secondary amines give a sulfonamide that is not soluble in this reagent. Offer an explanation for this observation.

BIBLIOGRAPHY

For an overall view of the sulfa drug story see

Kirk–Othmer Encyclopedia of Chemical Technology, 4th ed., Wiley: New York, 1992, Vol. 2, p. 876.

Examples of treatment of sulfonyl chlorides with amines:

De Boer, Th. J.; Backer, H. J. *Organic Syntheses*; Wiley: New York, 1963; Collect. Vol. IV, p. 943.

Reid, J. R.; Dufresne, R. F.; Chapman, J. J. *Organic Syntheses* **1997,** *74,* 217.

Scheifele, H. J., Jr.; De Tar, D. F. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 34.

The use of the perfluoroacetic anhydride reagent in this synthesis is reported in

Hurdis, E. C.; Yang, J. W. *J. Chem. Educ.* **1969,** *46,* 697.

SEQUENCE D

The Synthesis of 2′-Bromostyrene

Purpose. A three-step synthesis of 2′-bromostyrene is carried out starting with benzaldehyde. You will gain experience working with semimicroscale quantities of organic materials. (See Sequence A, where benzaldehyde is also used as the starting material, in that case for the formation of two different intermediates in the seven-step synthesis of hexaphenylbenzene.) You will become familiar with the stereo-chemistry involved in the addition of molecular halogen to an alkene (see also Experiment [A2$_b$]). You will observe the elimination of a hydrogen halide promoted by a concerted decarboxylation reaction (see also Experiment [A3$_b$]). As an option, ^{1}H NMR spectroscopy will be used to determine the cis/trans ratio of isomers in the product (see also Experiment [5B]).

THE SYNTHESIS OF A FRAGRANCE

Numerous classes of organic compounds have characteristic odors. For example, volatile esters have pleasant odors (see Experiment [8B]) and are often used in perfumes and artificial flavorings where they contribute to the fragrance of the material; acid chlorides have sharp penetrating odors (see Experiment [B2]); the alkyl amines have a "fishy" smell (see Experiment [B3]); benzaldehyde has the odor of bitter almond oil (see Experiment [20]); and cinnamaldehyde, the odor of cinnamon (see Experiment [11C]). Experiment [11C] also contains an expanded discussion of naturally occurring fragrant materials, called *essential oils*. As chain length increases, the odors of the short-chain alkyl carboxylic acids progress from the sharp, irritating odors of formic and acetic acids to the very rank, disagreeable odors of butyric (rancid butter), valeric, and caproic (dirty socks and goats) acids. Low molecular weight thiols (mercaptans), sulfides, and disulfides have intensely disagreeable odors. Examples of these materials are the active principals in the chemical defense spray used by the skunk. Via evolution, this animal has developed a highly effective mixture of 3-methylbutane-1-thiol, *trans*-2-butene-1-thiol, and *trans*-2-buten-1-yl methyl disulfide, which is able to deter most skunk predators, not just humans. Our sense of smell can detect one part of ethanethiol in 50 billion parts of air. This sounds spectacular, but humans have developed other highly sensitive detectors (the retinas of our eyes) to a band of electromagnetic radiation in the region 400–750 nm (the visible region of the spectrum). We have, as a consequence, become less dependent on our sense of chemical communication (our sense of smell). When we compare our detection limit of mercaptans to the male silkworm's detection of the female sex *pheromone*, we find that the *Bombyx mori* respond to threshold levels close to 1 million times lower (in the concentration range of 100 molecules per milliliter)!

Clearly, chemical communication (odor) plays a critical role in mediating animal and plant behavior within and between species. These metabolites are often stored and released when the animal (or plant) responds to certain stimuli. Historically, these substances served as the major communication link between living systems during the evolutionary development of single cell organisms. Thus, they predated hormones, which function in a similar fashion, but within a host collection of cells. While these substances may be used by a given insect species as a sex attractant, pheromones can also mediate other intraspecies functions, such as acting as trail guides and alarm signals. Thus, pheromones are substances that are used to communicate information between individuals of the same species to their adaptive advantage.The activity of a pheromone frequently depends on the configuration around a double bond (*E* or *Z*) as well as the absolute configuration around an *R* or *S* chiral center, just as humans select for these structural features (see discussion in Experiment [11D]). In a number of cases, the response has been shown to depend on the ratio of the isomers. There are several other classes of chemical communication substances used by plants and animals. The *allomones* are interspecies materials, which, on release, are used to the adaptive advantage of the host but to the disadvantage of the receiver. A typical function of an allomone is as a chemical defense agent; the thiols and disulfides used by the skunk are excellent examples. Three typical chemical communication substances are shown here:

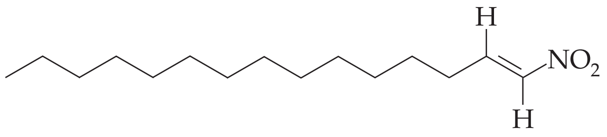

Termite defense allomone

and

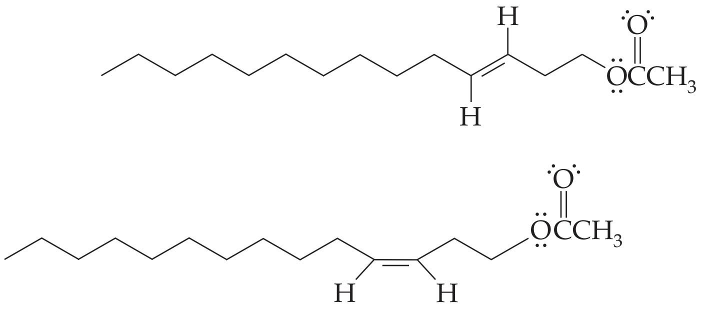

Sex pheromones of the European corn borer (a mixture of cis + trans acetates)

The synthetic route undertaken in the Sequence D experiments leads ultimately to a mixture of the diastereomeric cis–trans isomers of 2′-bromostyrene:

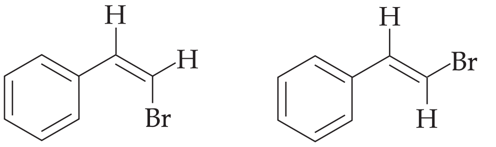

cis-2′-Bromostyrene *trans*-2′-Bromostyrene

This material is often used as an additive to impart a pleasant fragrance to soaps, since it has a very pleasant aroma, similar to that of hyacinth. It is not, however, a naturally occurring material. As a nice illustration of the fact that stereoisomeric compounds can have markedly different physiological effects, it has been demonstrated that a single diastereomer, *trans*-2′-bromostyrene, is responsible for the hyacinth-like odor.

The chemistry involved in this particular synthesis of 2′-bromostyrene contains a number of fundamental organic reactions and is rather interesting to study. The first stage (Experiment [D1]) involves the condensation of benzaldehyde with malonic acid to produce the intermediate, *trans*-cinnamic acid. This reaction is mechanistically similar to the well-known *aldol* reaction, and is often referred to as the *Knoevenagel* condensation:

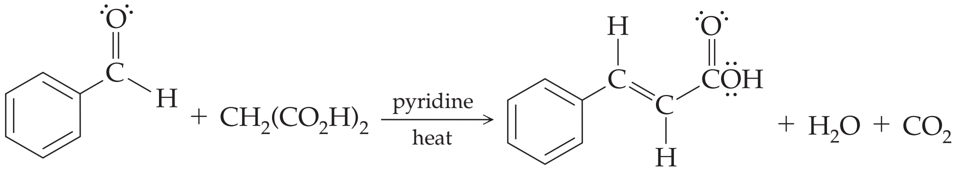

Benzaldehyde Malonic acid *trans*-Cinnamic acid

The second stage (Experiment [D2]) involves the addition of bromine to the intermediate formed in the first step, *trans*-cinnamic acid. The product of this halogenation is *erythro*-2,3-dibromo-3-phenylpropanoic acid. The stereochemistry involved in the formation of this second intermediate is a result of the nature of the anti addition of molecular bromine to a trans alkene. The details are given in the discussion in Experiment [D2].

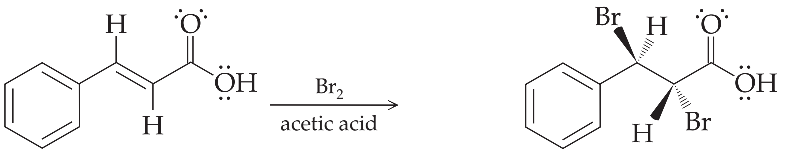

trans-Cinnamic acid *erythro*-2,3-Dibromo-3-phenylpropanoic acid

In Experiment [D3], the third step in Sequence D, the target molecule, 2′-bromostyrene, is generated by a novel elimination reaction that involves the loss of both CO_2 and HBr from the dibromophenylpropanoic acid. This stage of the synthesis offers an excellent opportunity to study the effect of solvent on the stereochemical course of a reaction by spectral analysis. In this case NMR (or infrared) data can be used to establish whether a cis–trans mixture is obtained, or just one pure isomer (cis):

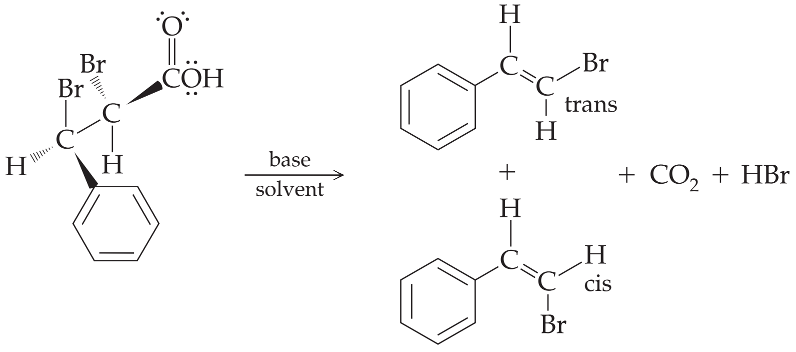

erythro-2,3-Dibromo-3-phenylpropanoic acid

cis- + *trans*-2′-Bromostyrene

The Verley–Doebner Modification of the Knoevenagel Reaction: *trans*-Cinnamic Acid

Experiment D1

Common name: *trans*-cinnamic acid
CA number: [140-10-3]
CA name as indexed: 2-propenoic acid, 3-phenyl-, (*E*)-

Purpose. *trans*-Cinnamic acid is prepared as the first intermediate in the Sequence D set of reactions that ultimately lead to 2′-bromostyrene. You will review the chemistry associated with an important variety of aldol-type condensation, the modified Knoevenagel reaction.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)

REACTION

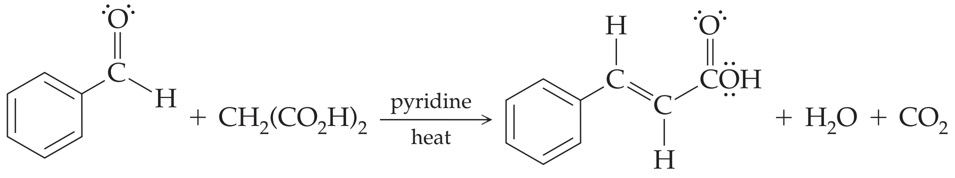

Benzaldehyde Malonic acid *trans*-Cinnamic acid

DISCUSSION

Those reactions that are called aldols derive their name from the early nineteenth-century organic literature. The term was first applied to a self-addition product of acetaldehyde that forms under basic conditions. The product, 3-hydroxybutanal, can form in yields as high as 50% in the presence of sodium hydroxide. This substance eventually became referred to as *aldol,* because it was both an aldehyde and an alcohol.

The aldol reaction can be defined in the broad sense as a reaction in which a nucleophile-generated alpha to an electron-withdrawing functional group (in the large majority of cases, carbonyl groups are responsible for the α-hydrogen acidity, although they are not required) adds to a carbonyl group (it may be self-condensation, as in the example of acetaldehyde, but it may also involve attack on another carbonyl-containing species, if it is present). Several experiments in this text illustrate this type of reactivity. Experiments [20] and [A3$_a$] are classic aldol reactions in which conditions are vigorous enough to result in elimination of the β-hydroxyl group to yield α,β-unsaturated ketones as products. In the first step of Sequence D, we now add the Knoevenagel reaction as another condensation possessing a mechanism similar to that of the aldol.

Emil Knoevenagel (1865–1921) Knoevenagel was born in Hanover, Germany. He was the son of a chemist and started his studies at the Technical Institute at Hanover. Later he studied with both Victor Meyer and Gattermann at Göttingen where he received a Ph.D. in 1889. When Victor Meyer moved to Heidelberg, Knoevenagel accompanied him. In 1896 he was appointed assistant professor of organic chemistry at Heidelberg and, in the same year, published his studies on the reaction that now bears his name. He eventually became professor of organic chemistry in 1900.

Knoevenagel had a variety of interests, including stereoisomerism. He worked extensively with aldol-type reactions, and pioneered the use of amine bases to promote these condensations. He was particularly interested in pyridine chemistry, and was the first to synthesize the pyridine ring by heating hydroxylamine with 1,5-diketones.

Following World War I, in which he saw active service, he continued his studies, only to die suddenly, at the age of 56, during an appendectomy.

Interestingly, although Knoevenagel demonstrated the effectiveness of amine bases in promoting aldol-type reactions and though, as noted above, he was particularly interested in pyridine, he overlooked this material's potential application to aldol condensations. It was left to Verley (1899) and Doebner (1900) to introduce successful modifications of the condensation in which pyridine appears to play a number of roles: as a solvent, as a base, and assisting in the decarboxylation.

The Knoevenagel reaction in its simplest form is the condensation of malonic esters (or their analogues) with aldehydes or ketones in the presence of an amine base catalyst plus a small amount of a carboxylic acid (or amino acid) cocatalyst. The condensation products are often α,β-unsaturated carbonyl compounds. For example,

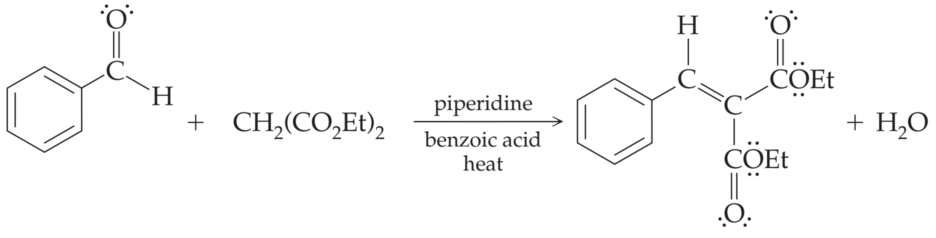

Benzaldehyde Diethyl malonate Diethyl benzalmalonate

Other substances possessing an acidic methylene group have been incorporated in aldol-type reactions. These materials include ethyl cyanoacetate and ethyl acetoacetate, as well as phenylacetonitrile, benzyl ketones, and aliphatic nitro compounds. The reaction is often run in benzene or toluene solvent, so that the water formed can be continuously removed. Other cocatalysts, besides carboxylic acids, are various ammonium salts, such as ammonium acetate and piperidinium acetate. The Knoevenagel reaction gives highest yields when aldehydes are used as the electrophile, although selected ketones can sometimes give acceptable yields. One of the more important properties of these reactions from a synthetic perspective is that they offer a route to the formation of C—C bonds.

In the formation of the first synthetic intermediate in Sequence D, the very effective Verley–Doebner modification of the fundamental Knoevenagel condensation is used. This modification uses malonic acid in place of the conventional ester to promote enolization. In addition, the heterocyclic amine, pyridine, functions as both the base catalyst and the solvent. A cocatalyst, β-alanine (an amino acid), is also introduced. Mechanistically, the reaction closely resembles the aldol condensation in that in both cases a carbanion is generated by abstraction, by base, of a proton *alpha* to a carbonyl group. The resulting carbanion is stabilized as an enolate anion (see below).

The enolate anion then acts as a nucleophile and attacks the carbonyl carbon of a second carbonyl-containing molecule in the reaction. (An example would be the aldol reaction in Experiment [20].) The intermediate aldol product, the β-hydroxyacid, undergoes rapid dehydration under these reaction conditions, to give the α,β-unsaturated diacid shown here:

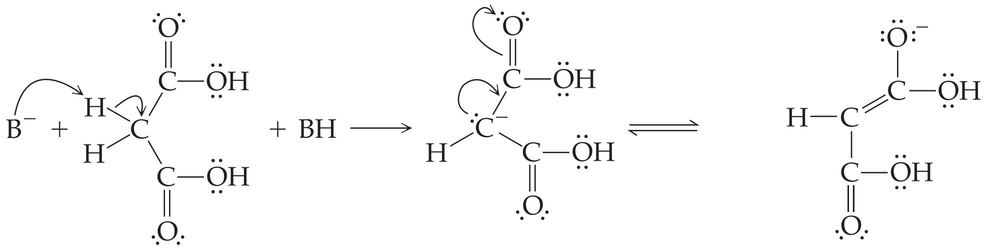

(acidic hydrogen)

(carbanion)

(enolate anion)
(one resonance structure shown)

Conjugate addition (1,4 addition) of pyridine and ionization of a carboxylic acid are followed by decarboxylation and concomitant elimination of pyridine to yield the α,β-unsaturated carboxylic acid as shown here.

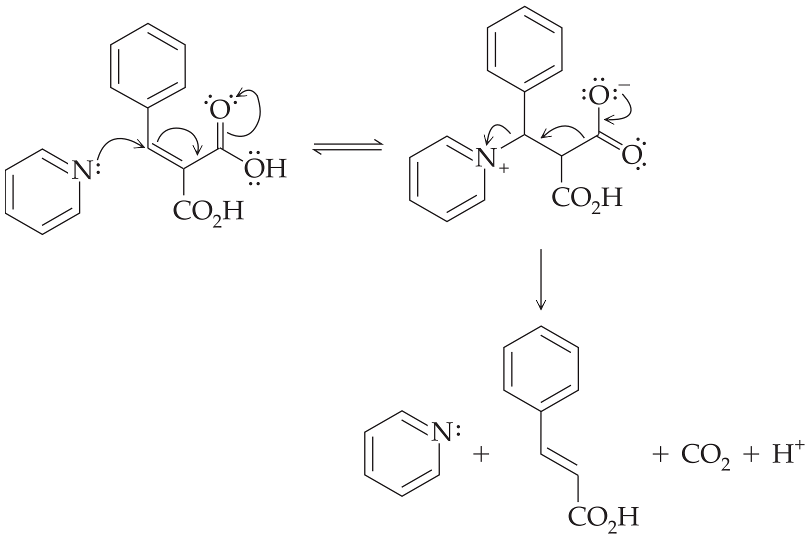

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Benzaldehyde	106.12	580 μL (603 mg)	5.7	−26	179	1.04
Malonic acid	104.06	1.5 g	14.4	135 dec		
Pyridine	79.10	3 mL			115	0.978
β-Alanine	89.09	120 mg	1.35	205 dec		

Reagents and Equipment. In a 10-mL round-bottom flask containing a magnetic stir bar, weigh and place 1.5 g (14.4 mmol) of malonic acid followed by 120 mg of β-alanine. Now, in the **hood,** add 3 mL of pyridine and 580 μL of freshly distilled benzaldehyde.

HOOD

NOTE. Dispense the benzaldehyde using an automatic delivery pipet. Measure the pyridine using a 10-mL graduated cylinder.

CAUTION: Pyridine is a toxic amine with a strong unpleasant odor. Both reagents should be stored and dispensed in the **hood.**

HOOD

Attach the flask to a water-jacketed reflux condenser, and place the assembly in a sand bath on a magnetic stirring hot plate.

Reaction Conditions. Heat the mixture with stirring at a sand bath temperature of about 130 °C for 1.5 h.

Isolation of Product. Allow the reaction mixture to cool to room temperature and, in the **hood,** remove the flask. Transfer the reaction solution using a Pasteur pipet to a 50-mL Erlenmeyer flask containing 12 mL of ice-cold water. Rinse the round-bottom flask with an additional 3 mL of ice-cold water, and transfer the rinse to the same Erlenmeyer flask in like manner. Now add, in small portions, 6 M HCl (~8 mL) until a white precipitate forms and the solution tests weakly acidic to pH paper. Collect the precipitate of *trans*-cinnamic acid by vacuum filtration using a Hirsch funnel. Wash the white crystals with three 2-mL portions of cold water. To partially dry the material, continue the suction for an additional 15 min (remember to cover the Hirsch funnel with filter paper to protect the filter cake from contamination with dust from the laboratory atmosphere during the extended air-drying period).

Purification and Characterization. Oven-dry the partially dried product at 100 °C overnight. Weigh the crude acid and calculate the percent yield. Determine the melting point. Compare your result with the literature value. This material is likely to be of sufficient purity to be used in the next step of Sequence [D], Experiment [D2]. If your material melts below the literature value, for characterization recrystallize approximately 30 mg from hot water using the Craig tube. Dry as before and redetermine the melting point. To further characterize the material, obtain IR and ^{1}H NMR spectra. Compare your spectra with those recorded in the literature (*The Aldrich Library of IR Spectra, The Aldrich Library of NMR Spectra,* and/or the corresponding spectral data available online (e.g., SciFinder Scholar)).

NOTE. Approximately 500 mg of purified product with a melting point within 2–3 °C of the literature value is suggested for continuing the sequence on to Experiment [D2].

Chemical Tests. Add several crystals (~5 mg) of the *trans*-cinnamic acid to 1 mL of 5% sodium bicarbonate on a watch glass. Does evolution of CO_2 indicate the presence of a carboxylic acid? Does the material give a positive bromine test for unsaturation (see Chapter 9).

QUESTIONS

7-63. Write a complete mechanism for the addition of diethyl malonate to ethanal in the presence of base to form a β-hydroxy ester.

7-64. Outline a synthesis that forms at least one C—C bond for each of the following compounds:

- **(a)** $CH_3CH_2CH_2CH_2CH_2CH{=}CHCO_2H$
- **(b)** $CH_3CH_2CH_2CH{=}C(CN)(CO_2CH_2CH_3)$

7-65. As mentioned in the discussion, ketones generally give poor yields in the Knoevenagel reaction with diethyl malonate. However, the reaction with ketones gives good yields with ethyl cyanoacetate and malononitrile. Explain.

7-66. Give the structure of the products of the Knoevenagel reaction for the following pairs of reactants:

- **(a)** Cyclopentanone + malononitrile (dicyanomethane)
- **(b)** Benzaldehyde + ethyl acetoacetate
- **(c)** Propanal + nitromethane

BIBLIOGRAPHY

Knoevenagel, E. *Chem. Ber.* **1898,** *27,* 2345.

Descriptions of the Knoevenagel reaction:

Cope, A. C. *J. Am. Chem. Soc.* **1937,** *59,* 2327.

Cope, A. C.; Hofmann, C. M.; Wyckoff, C.; Hardenberg, E. *J. Am. Chem. Soc.* **1941,** *63,* 3452.

House, H. O. *Modern Synthetic Reactions;* Benjamin: Menlo Park, CA, 1972, pp. 649–653.

Johnson, J. R. *Org. React.* **1942,** *1,* 210.

Jones, G. *Org. React.* **1967,** *15,* 204.

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Wiley: New York, 2007, Chap. 16, p. 1358 and references therein.

Tietze, L. F.; Beifuss, U. *Comp. Org. Syn.* **1991,** *2,* 341.

Selected references of the use of the Knoevenagel reaction recorded in *Organic Syntheses:*

Allen, C. F. H.; Spanger, F. W. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 377.

Cope, A. C.; Hancock, E. M. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 399.

Horning, E. C.; Koo. J.; Fish, M. S.; Walker, G. N. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 408.

McElvain, S. M.; Clemens, D. H. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 463.

This step in the sequence was adapted from the work of

Kolb, K. E.; Field, K. W.; Schatz, P. F. *J. Chem. Educ.* **1990,** *67,* A304.

Bromination of *trans*-Cinnamic Acid: *erythro*-2,3-Dibromo-3-phenylpropanoic Acid

Experiment D2

Common name: *erythro*-2,3-dibromo-3-phenylpropanoic acid
CA number: [31357-31-0]
CA name as indexed: Benzenepropanoic acid, α,β-dibromo-, (*R**,*S**)-

Purpose. The bromination of *trans*-cinnamic acid is carried out to obtain *erythro*-2,3-dibromo-3-phenylpropanoic acid, the direct precursor to 2′-bromostyrene, which is the synthetic target of Sequence D. You will review the stereospecificity of the addition of molecular bromine to an alkene.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)

REACTION

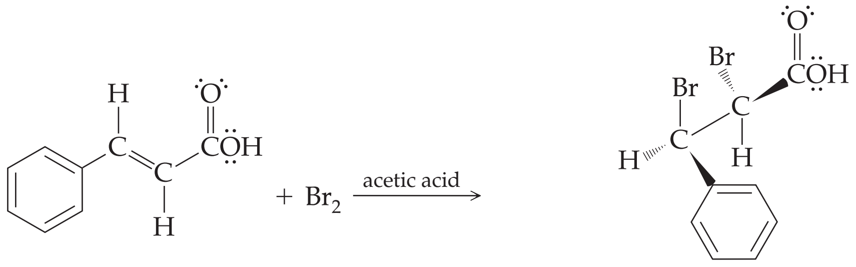

trans-Cinnamic acid

erythro-2,3-Dibromo-3-phenylpropanoic acid

DISCUSSION

The bromination of alkenes is known to be a **stereospecific** reaction. An example and detailed discussion of this addition to *trans*-stilbene is given in Experiment [$A2_b$], and another example is illustrated in Experiment [F2]. In the present case, bromine undergoes a similar addition, and the halogenated product obtained is the erythro diastereomer of 2,3-dibromo-3-phenylpropanoic acid, which is produced as a racemic mixture of the two enantiomers.

The term *erythro* is derived from carbohydrate chemistry, and is used to describe the *relative* configurations of the two adjacent chiral centers (stereocenters). Specifically, the term erythro describes structures whose Fischer projections place identical substituents (on adjacent carbon atoms) on the *same side* of the (otherwise identical) Fischer projection. The opposite of erythro is *threo*, which would describe structures with identical substituents (on adjacent carbons) on opposite sides of the (otherwise identical) Fischer projection. Because erythro and threo describe only the relative configurations of the two chiral centers with respect to one another, there are two enantiomers of each erythro diastereomer, and of each threo diastereomer, as shown here both in Fischer projections and more familiar illustrations:

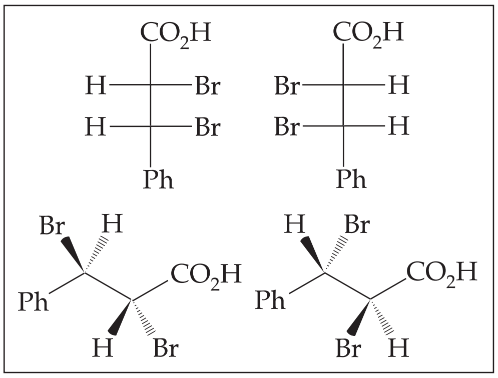

Enantiomers of the erythro diastereomer
of 2,3-dibromo-3-phenylpropanoic acid

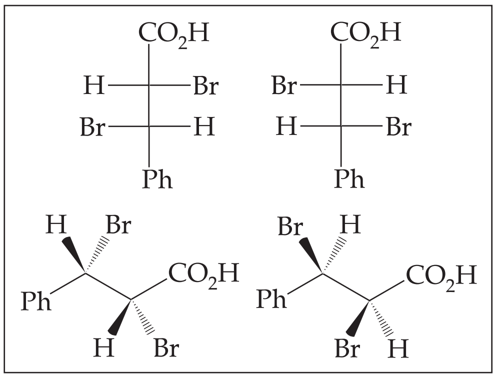

Enantiomers of the threo diastereomer
of 2,3-dibromo-3-phenylpropanoic acid

The result of the halogenation of *trans*-cinnamic acid, as in the case of (*E*)-stilbene in Experiment [A2$_b$], is an anti addition of molecular bromine. The enantiomeric products are shown here:

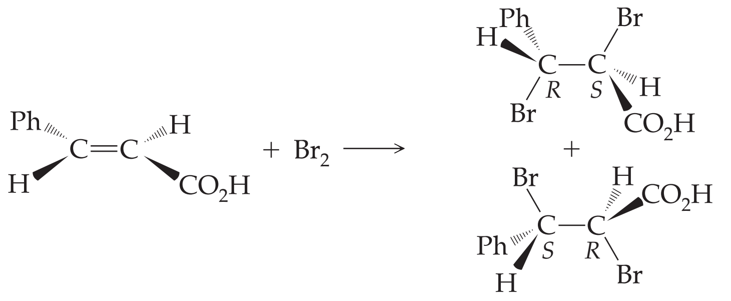

trans-Cinnamic acid

erythro-(2*R*,3*S*- and 2*S*,3*R*)-2,3-dibromo-3-phenylpropanoic acid

What mechanism best explains this observed stereochemistry? The starting un-saturated acid and bromine are both *achiral*, and thus the product, if a chiral molecule, must be racemic. The outer-shell electrons of molecular bromine, however, are highly polarizable and molecular bromine reacts as an electrophile with the nucleophilic π electrons of the alkene group of cinnamic acid, in a reaction that is effectively a nucleophilic substitution reaction on a bromine atom. The product is a carbocation, which is, however, stabilized by the ability of the bromine atom to donate a lone pair to the carbocation. This interaction results in the formation of a *chiral*, cyclic *bromonium ion* intermediate. This intermediate ion is a rigid structure in which the ring section is only open to further attack by a nucleophile (S_N2 attack by the Br^- ion generated in the first stage) from the side opposite the bromine atom.

Because both cinnamic acid and molecular bromine are achiral, two enantiomeric bromonium ions are formed at equal rates. Ring opening of each bromonium ion may preferentially occur as shown here at the carbon bearing the phenyl group, since this carbon bears more fractional positive charge than the carbon adjacent to the carboxyl group. Because the two reaction pathways shown here are enantiomeric, they proceed at equal rates to produce racemic *erythro*-2,3-dibromo-3-phenylpropanoic acid:

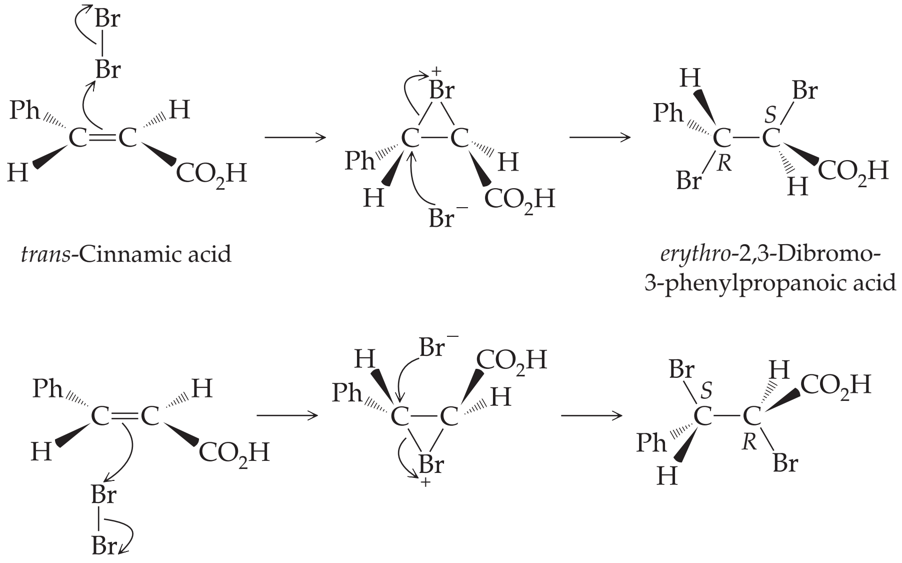

trans-Cinnamic acid

erythro-2,3-Dibromo-3-phenylpropanoic acid

Recent investigations have shed further light on the nature of the cyclic intermediate. Stable solutions of cyclic bromonium ions in liquid SO_2 have been prepared as SbF_6^- salts. For example,

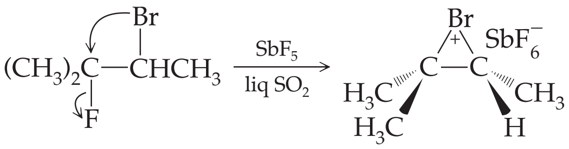

See the discussion in Experiment [A2$_b$] for further evidence relating to these intermediate species.

The *erythro*-2,3-dibromo-3-phenylpropanoic acid product prepared in this experiment has a melting point of 203–204 °C. The corresponding threo diastereomer has been synthesized and its racemate melts at 91–93 °C. Thus, the experimental results support the stereospecific nature of the bromine addition that is rationalized by the proposed mechanism.

NOTE. Pyridinium bromide perbromide, a solid reagent, is used as a source of bromine in this experiment. This material is much easier to handle than liquid bromine (also see Experiment [A2$_b$]).

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
trans-Cinnamic acid	148.2	500 mg	3.37	133		
Pyridinium bromide perbromide (tech)	319.84	1.2 g				
Acetic acid	60.05	12 mL	16.2		116–118	1.05
erythro-2,3-Dibromo-3-phenylpropanoic acid	308.0			203–204		

Reagents and Equipment. Weigh and place 500 mg (3.37 mmol) of *trans*-cinnamic acid in a 25-mL round-bottom flask containing a magnetic stir bar. Now weigh and place (in the **hood**) 1.2 g of pyridinium bromide perbromide in the same flask. HOOD

HOOD CAUTION: Pyridinium bromide perbromide is a corrosive reagent and a lachrymator. Dispense this material in the ***hood.***

Using a graduated cylinder, measure 12 mL of glacial acetic acid and add this to the solid reagents. Attach the flask to a water-cooled reflux condenser and place the assembly in a sand bath.

Reaction Conditions. Heat the mixture with stirring in a sand bath at 120–125 °C for 45 min.

Isolation of Product. Cool the orange solution to room temperature. With a Pasteur pipet, transfer the solution in small portions to a 50-mL Erlenmeyer flask. Rinse the reaction flask with an additional 2 mL of acetic acid, and transfer this rinse to the same Erlenmeyer flask.

Cool the solution in an ice bath for 5 min, and then slowly add 15 mL of cold water in three 5-mL portions. Swirl the contents of the flask between additions. A pale yellow precipitate should form. Collect the crude product by vacuum filtration using a Hirsch funnel. Wash the filter cake with three 2-mL portions of cold water, during which time the solid product should become white.

Purification and Characterization. Dry the product in an oven at 110 °C overnight. Weigh the product and calculate the crude yield. Determine the melting point and compare it to that listed in the Reactant and Product table.

A 20- to 30-mg sample of the *erythro*-2,3-dibromo-3-phenylpropanoic acid may be recrystallized (Craig tube) from chloroform (in the **hood**), if desired.

HOOD

Chemical Test. Perform the Beilstein test to detect the presence of a halogen (see Chapter 9).

NOTE. Approximately 300 mg of purified (washed) product, with a melting point within 3–5 °C of the literature value, is suggested for continuing the sequence on to Experiment [D3].

QUESTIONS

7-67. The product prepared in this experiment has 2 chiral centers (stereocenters), which give rise to 4 stereoisomers. Draw each of these stereoisomers, and indicate the relationships between each of the stereoisomers.

7-68. The addition of bromine to ethylene in the presence of high concentrations of chloride ion in an inert solvent results in the formation of 1,2-dibromoethane and 1-bromo-2-chloroethane. No 1,2-dichloroethane is obtained. Account for these results using a suitable mechanistic sequence.

7-69. 2,3-Dibromobutane contains 2 chiral centers. Therefore, the possibility of 4 stereoisomers exists. However, the addition of bromine to an equimolar mixture of *cis*- and *trans*-2-butene generates only three stereoisomers. Explain.

7-70. Cyclohexene undergoes bromination in methanol solvent to give *trans*–1-bromo-2-methoxycyclohexane. Draw a suitable mechanism to account for this.

BIBLIOGRAPHY

***erythro*-2,3-Dibromo-3-phenylpropanoic acid has been previously prepared:**

Corvari, L.; McKee, J. R.; Zanger, M. *J. Chem. Educ.* **1991,** *68,* 161.
Murahashi, S.; Naota, T.; Tanigawa, Y. *Organic Syntheses;* Wiley: New York, 1990; Collect. Vol. VII, p. 172.
Sudborough, J. J.; Thompson, K. J. *J. Chem. Soc.* **1902,** *83*, 666.

The bromination of the ethyl ester of *trans*-cinnamic acid has been reported:

Abbott, T. W.; Althousen, D. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 270.
McElvain, S. M.; Kundiger, *D. Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 123.

Additional selected references of bromination of alkenes from *Organic Syntheses:*

Allen, C. F. H.; Edens, C. O., Jr. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 731.
Cromwell, N. H.; Benson, R. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 105.
Fieser, L. F. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 195.

Also see

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Wiley: New York, 2007, Chap. 15, p. 999 and references therein.

Experiment D3

An Elimination Reaction with *erythro*-2,3-Dibromo-3-phenylpropanoic Acid: 2′-Bromostyrene

Common names: (*Z*)-β-Bromostyrene, *cis*-2′-bromostyrene
CA number: [588-73-8]
CA name as indexed: benzene, (2-bromoethenyl)-, (*Z*)-
Common names: (*E*)-β-bromostyrene, *trans*-2′-bromostyrene
CA number: [588-72-7]
CA name as indexed: benzene, (2-bromoethenyl)-, (*E*)-

Purpose. An elimination reaction is carried out using *erythro*-2,3-dibromo-3-phenylpropanoic acid, the direct precursor to 2′-bromostyrene. The influence of solvents and base on the course of the elimination reaction is illustrated. You will consider factors that control the stereospecificity of a reaction. You will explore the option of using NMR spectroscopy to establish the cis/trans isomer ratio in the 2′-bromostyrene product.

Prior Reading

Technique 4: Solvent Extraction
Liquid–Liquid Extraction (p. 72)
Technique 6B: Concentration of Solutions (pp. 101–104)

REACTION

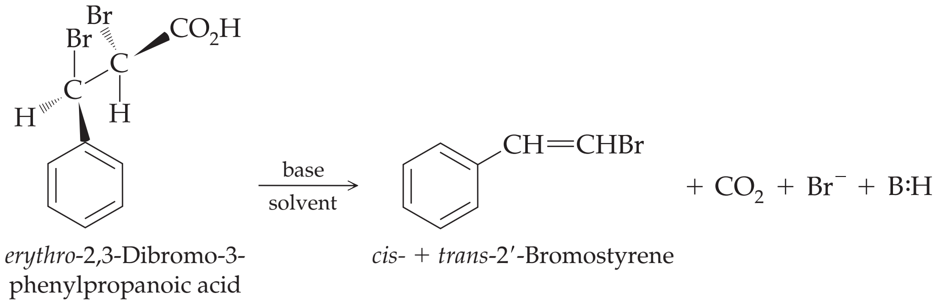

erythro-2,3-Dibromo-3-phenylpropanoic acid

cis- + *trans*-2′-Bromostyrene

DISCUSSION

The course of the elimination depicted in the above reaction is both base and solvent dependent. In this experiment you will have the opportunity to investigate this solvent dependence. In Part A, you will study the effect of carbonate base in aqueous solvent. Under these conditions, the reaction sequence proceeds via an E1 mechanism in which the first step involves the ionization of one of the C—Br bonds. Clearly, of the two possibilities, halide formation that results in a resonance stabilized benzylic carbocation will be highly favored over development of a positive charge on a carbon alpha to a carboxyl group (see Question 7-71). Formation of the carbocation intermediate is then rapidly followed by decarboxylation. Although the conjugated section of the positively charged intermediate is planar, free rotation or partially hindered rotation remains possible about the C—C bond, leading to the remaining tetrahedral carbon. Thus, both *cis*- and *trans*-2′-bromostyrene may be obtained during the subsequent decarboxylation. The proposed sequence is shown here:

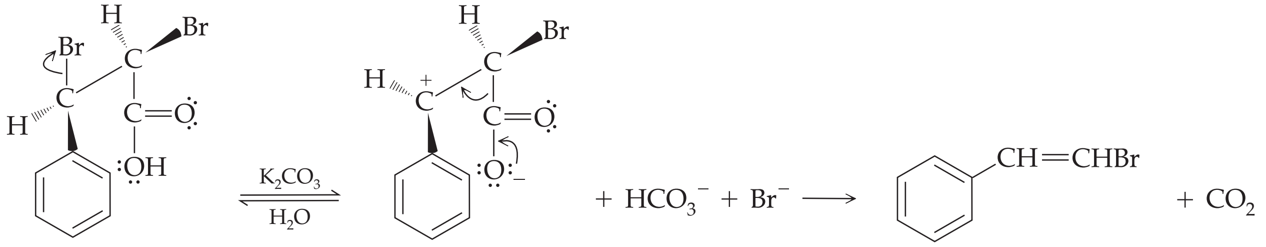

erythro-2,3-Dibromo-3-phenylpropanoic acid (benzylic carbocation) *cis*- + *trans*-2′-Bromostyrene

In Part B of the experiment, the solvent is changed from water (used in Part A) to a much less polar solvent, acetone. Under these nonaqueous conditions, elimination also occurs, but in this case the experimental data demand a different mechanism, one involving a smooth concerted electron flow without short-lived intermediate formation. Furthermore, when carried out in acetone, the reaction yields exclusively *cis*-2′-bromostyrene. Details of the mechanism follow:

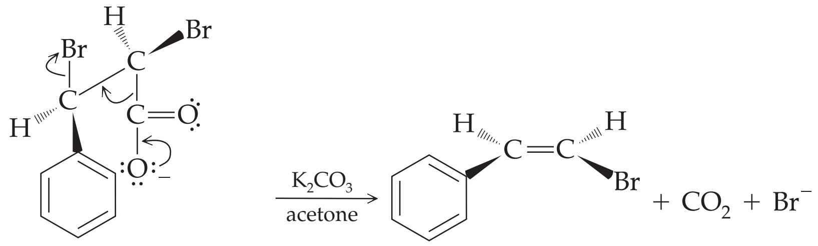

erythro-2,3-Dibromo-3-phenylpropanoic acid anion *cis*-2′-Bromostyrene

Under these conditions, therefore, the reaction becomes stereospecific. The mechanism is classified as an E2 elimination. Note that when the two leaving groups are anti to one another, as is thermodynamically preferred, elimination yields the cis alkene.

In the two sets of reaction conditions to be studied in Parts A and B of this experiment, potassium carbonate is used as the base in both cases and it is the solvent that is varied, as mentioned above. In another experimental procedure, sodium azide is used as the base and *N,N*-dimethylformamide (DMF) is used as the solvent, and higher selectivity for the formation of the cis isomer is reported (see Bibliography, Corvari et al.).

EXPERIMENTAL PROCEDURE

Estimated time to complete the reaction: 1.5 h, Part A; 2.0 h, Part B.

Physical Properties of Reactants				
Compound	**MW**	**Amount**	**mmol**	**mp (°C)**
erythro-2,3-Dibromo-3-phenylpropanoic acid	308.0	300 mg	0.97	203–204
Potassium carbonate	38.21	300 mg	2.1	891

PART A Conditions: Water Solvent

Reagents and Equipment. Weigh and place 300 mg (0.97 mmol) of *erythro*-2,3-dibromo-3-phenylpropanoic acid (prepared in Experiment [D2]) and 300 mg (2.1 mmol) of potassium carbonate in a 25-mL Erlenmeyer flask. Add 5 mL of distilled water.

Reaction Conditions. Warm the mixture in a 120 °C sand bath until the solids dissolve (5–10 min) and then heat for an additional 10 min. A yellow oil should separate from the aqueous phase during this period.

Isolation of Product. Cool the two-phase system to room temperature and, using a Pasteur pipet, transfer the contents of the flask to a 12-mL centrifuge tube. Rinse the Erlenmeyer flask with a 2-mL portion of methylene chloride, which is then transferred to the centrifuge tube. Extract the aqueous phase with the CH_2Cl_2 rinse and follow this extraction with two more extractions of the aqueous phase with 2-mL portions of methylene chloride. Transfer the methylene chloride extracts (the organic phase should be the lower phase, but check solubility in water to be sure) to a 25-mL Erlenmeyer flask containing enough anhydrous sodium sulfate to dry the solution. Allow the extracts to dry for about 10 min.

Transfer a portion of the dried solution, using a Pasteur filter pipet, to a *tared* 5-mL conical vial containing a boiling stone. Concentrate this portion on a warm sand bath under a gentle stream of nitrogen gas in the **hood.** HOOD

Now transfer the remaining dried solution to the same tared vial and concentrate as before. Finally, rinse the sodium sulfate drying agent with an additional 1-mL portion of methylene chloride, transfer it to the vial and concentrate again to yield the *cis/trans*-2′-bromostyrene as a yellow oil.

NOTE. Exercise care during the concentration steps. Product yield will be greatly reduced by overheating or leaving the solution on the sand bath for too long a period.

Purification and Characterization. No further purification of the product is usually necessary. Note the hyacinth-like odor of the liquid. Determine the refractive index and compare it to the literature value (lit. value = 1.6070 at 20 °C, 78% trans isomer). Obtain IR and ^{1}H NMR spectra of the oil. In the infrared, the cis isomer shows a phenyl C—H bend at 770 cm^{-1}, while the trans isomer shows this bending mode at 731 and the alkene C—H out-of-plane mode is found at 941 cm^{-1} (see Bibliography, Strom et al.). Compare your results with the spectrum recorded in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar). www →

NMR Analysis. The ^{1}H NMR spectrum of a mixture of *cis*- and *trans*-2′-bromostyrene is shown in Figure 7.4. The doublets corresponding to the two olefinic protons of the major isomer can be clearly seen, centered at 6.75 and 7.1 ppm. One of the corresponding doublets from the minor isomer is evident at 6.42 ppm. This doublet has a coupling constant of 8.0 Hz. The two larger doublets from the major isomer have a coupling constant of 14.0 Hz, and thus, in this sample, the major isomer must be *trans*-2′-bromostyrene.

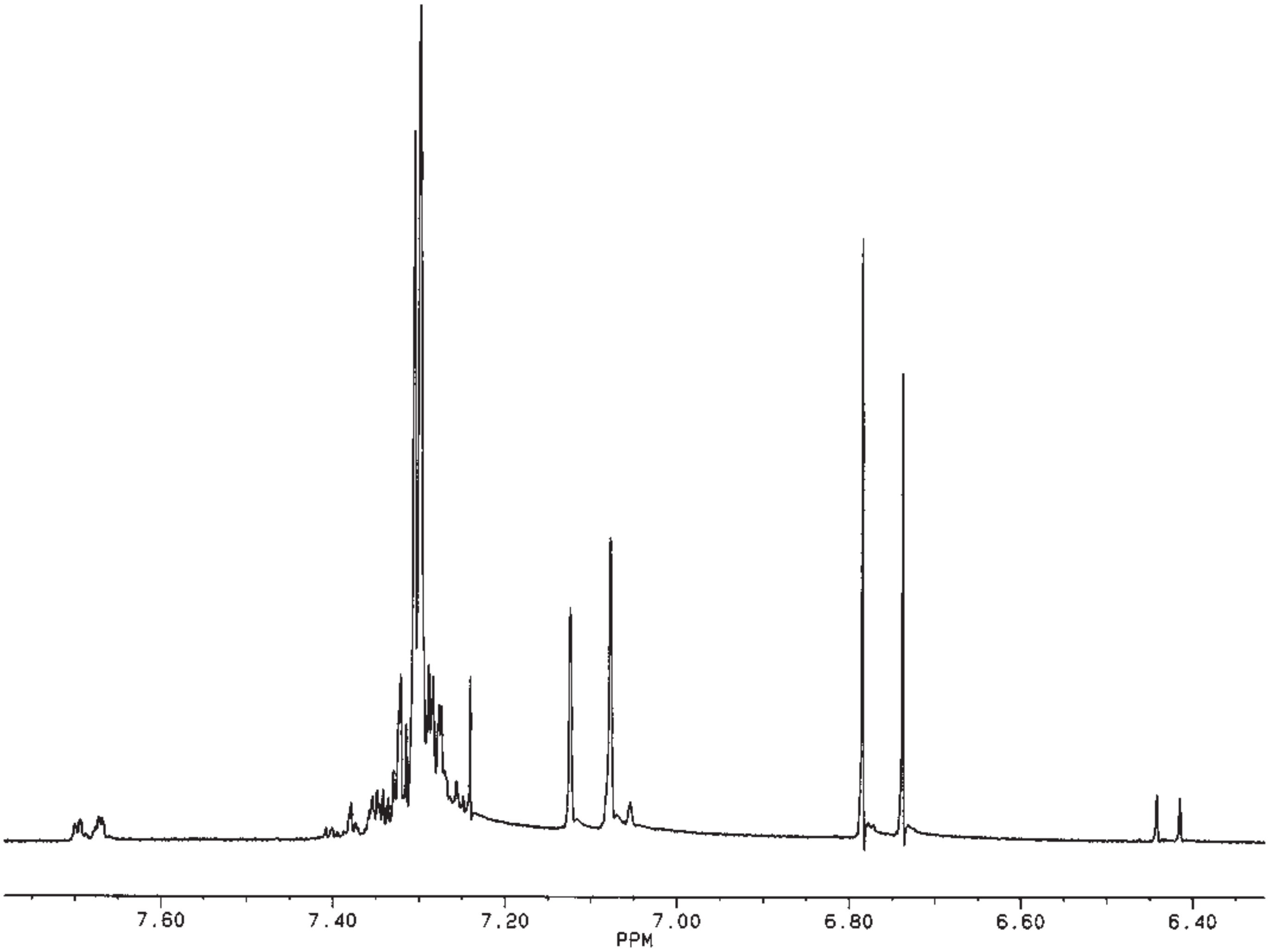

Figure 7.4 **^{1}H NMR spectrum: Mixture of *cis*- and *trans*-2′-bromostyrene.**

We can readily locate the second doublet from the minor, cis isomer by examining the ^{1}H–^{1}H COSY two-dimensional NMR spectrum (see Chapter 8, NMR discussions and Figure 7.5). The spectrum here is shown at a large vertical scale in order to discern the minor (cis) isomer. By looking below the doublet at 6.42 ppm, a cross-peak can be seen at about 7.1 ppm. Careful inspection reveals that these signals must be slightly upfield from the doublet at 7.1 ppm from the major (trans) isomer. Referring back to the ^{1}H NMR spectrum (Figure 7.4) one can see one peak of the hidden doublet just emerging from the upfield end of the doublet from the major isomer. Thus, the second doublet from the minor (cis) isomer must coincidentally be obscured by the NMR signal from the major isomer.

PART B Conditions: Acetone Solvent

Reagents and Equipment. Weigh and place 300 mg (0.97 mmol) of *erythro*-2,3-dibromo-3-phenylpropanoic acid (prepared in Experiment [D2]) and 300 mg (2.1 mmol) of potassium carbonate in a 10-mL round-bottom flask containing a magnetic stir bar. Add 5 mL of acetone that has been dried over sodium sulfate. Attach the flask to a reflux condenser and assemble the apparatus on an 80 °C sand bath.

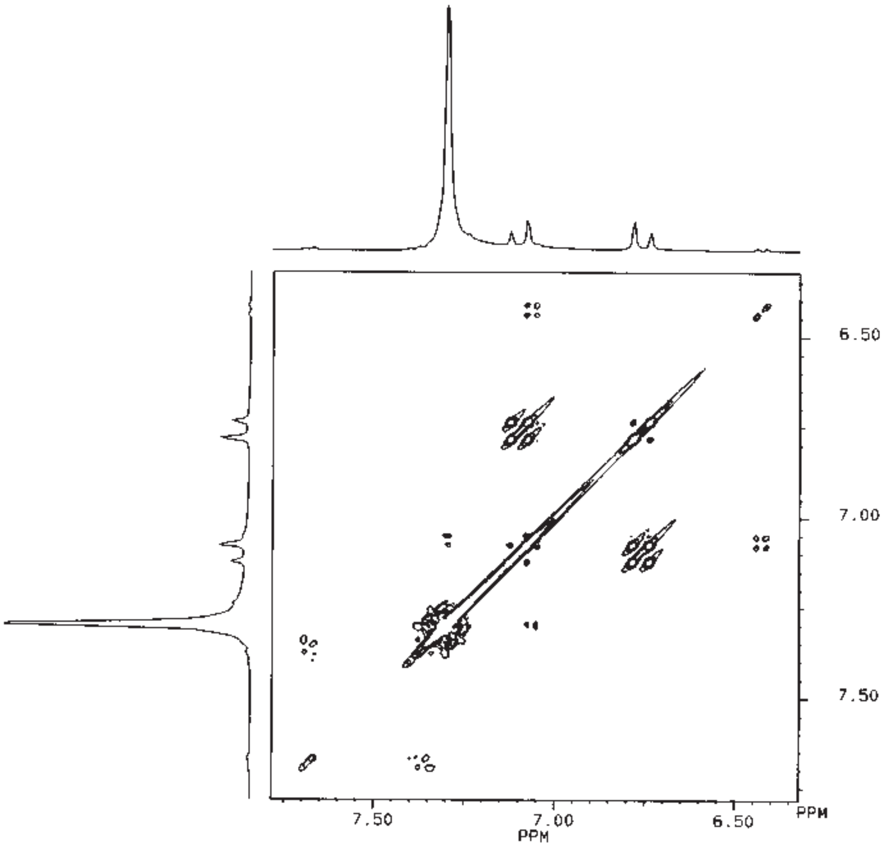

Figure 7.5 **^{1}H–^{1}H COSY NMR spectrum: Mixture of *cis*- and *trans*-2′-bromostyrene.**

Reaction Conditions. Heat (bath temperature ~80 °C) the mixture, with stirring, at reflux for 1 h.

Isolation of Product. Allow the solution to cool to room temperature and then concentrate it to dryness in the **hood** on a warm sand bath under a gentle stream of nitrogen gas. **Do not overheat!** HOOD

Now add 5 mL of distilled water to dissolve the solid cake. A yellow oil appears at this point, forming a two-phase system. Using a Pasteur pipet, transfer this aqueous–oily mixture to a 12-mL centrifuge tube. Rinse the round-bottom flask with a 2-mL portion of methylene chloride, which is then transferred to the centrifuge tube. Extract the aqueous phase with the CH_2Cl_2 rinse and follow this extraction with two more extractions of the aqueous phase with 2-mL portions of methylene chlo-ride. Transfer the organic extracts to a second 25-mL Erlenmeyer flask using a Pasteur filter pipet and dry the combined extracts over anhydrous sodium sulfate for at least 10 min.

Transfer the dried solution in 2- to 3-mL portions to a *tared* 5-mL conical vial containing a boiling stone. Concentrate the solution in the **hood** on a *warm* sand bath under a gentle stream of nitrogen gas. Rinse the sodium sulfate with an additional 1 mL of methylene chloride and combine this rinse with the dried solution. Concentrate to yield a yellow oil, *cis*-2′-bromostyrene. HOOD

NOTE. Exercise care during the concentration steps. Product yield will be greatly reduced by overheating or leaving the solution on the sand bath for too long a period.

Purification and Characterization. No further purification of the product is usually necessary. Weigh the product and calculate the crude percent yield. Obtain IR and ^{1}H NMR spectra of the oil. In the NMR, the peaks due to both isomers, if present, can be discerned (see Discussion, Part A). Compare your IR and NMR results with the spectra obtained for the cis–trans mixture formed in Part A and to those recorded in the Corvari et al. reference (see Bibliography) for pure *cis*-2′-bromostyrene. This reference also gives conditions under which the purity of the material may be more carefully determined by gas chromatography.

Chemical Tests. Several tests can be run to determine the presence of unsaturation, aromatic character, and the presence of bromine. Select the appropriate tests from Chapter 9. Do your results confirm the presence of these groups?

QUESTIONS

7-71. 1-Chloro-1-phenylethane ionizes easily under E1 conditions to form a benzylic carbocation intermediate. The ion is stabilized due to the delocalization of the positive charge to the aromatic ring. Draw resonance structures that indicate the stability of this ion.

7-72. Comment on the fact that *erythro*-2,3-dibromo-3-phenylpropanoic acid undergoes elimination by an E1 pathway in water solvent (Part A conditions), but by an E2 pathway (Part B conditions) when acetone is used as the solvent.

7-73. Explain in terms of the mechanism why conditions in Part A lead to a mixture of cis–trans isomers but that those used in Part B give only the cis isomer.

7-74. A benzylic carbocation is generated under the conditions of Part A. Would the presence of a para CH_3O— group on the benzene ring increase or decrease the stability of the benzylic carbocation? Explain.

7-75. Consider the stereochemistry of the carbocation intermediate formed under the conditions of Part A. Because the starting material used in this experiment, *erythro*-2,3-dibromo-3-phenylpropanoic acid, is racemic, the enantiomer of the carbocation shown must also be generated. Does the other enantiomer lead to the same diastereomer (cis) of the product or does it lead to the trans diastereomer?

BIBLIOGRAPHY

The synthesis of 2′-bromostyrene has been reported:

Corvari, L.; McKee, J. R.; Zanger, M. *J. Chem. Educ.* **1991,** *68,* 161.
Cristol, S. J.; Norris, W. P. *J. Am. Chem. Soc.* **1953,** *75,* 2645.
Grovenstein, E., Jr.; Lee, D. E. *J. Am. Chem. Soc.* **1953,** *75*, 2639.
L'abbé, G.; Miller, M. J.; Hassner, A. *Chem. Ind. (London)* **1970,** 1321.
Mestdagh, H.; Puechberty, A. *J. Chem. Educ.* **1991,** *68,* 515.
Murahashi, S.; Naota, T.; Tanigawa, Y. *Organic Syntheses;* Wiley: New York, 1990; Collect. Vol. VII, p. 172.
Strom, L. A.; Anderson, J. R.; Gandler, J. R. *J. Chem. Educ.* **1992,** *69,* 588.

For an overview of elimination reactions see

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Wiley: New York, 2007, Chap. 17, p. 1477 and references therein.

The Synthesis of Piperonylonitrile from Piperonyl Alcohol

SEQUENCE E

Purpose. The aromatic nitrile, piperonylonitrile, is obtained via a multistep synthesis that does not depend on potentially dangerous diazonium intermediates. The selective oxidation of a primary alcohol to an aldehyde is investigated. You will explore, at the same time, the use of resin-bound reagents that simplify the isolation of products. You will carry out a novel elimination reaction using an oxime to obtain the target nitrile.

Synthetic Perfumes

The first step in this sequence of reactions is the oxidation of piperonyl alcohol to obtain piperonal, an aromatic aldehyde. The reaction is selective and only the aldehyde is obtained; no oxidation to the carboxylic acid is observed:

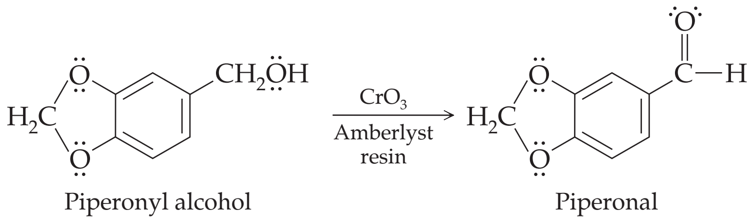

Piperonal has an agreeable odor and is marketed to the perfume industry as *heliotropine*. This commercial term comes from *heliotrope*, which was a name given by early herbalists to any plant that turned to the sun (from the Greek *helios,* the sun, and *trepein,* to turn), such as the sunflower. The name now is more narrowly defined as applying to those plants of the genus *Heliotropium* and specifically to *H. peruvianum*, a common species that is widely cultivated. The fragrance of this particular species also is referred to as *heliotrope.* Since piperonal possesses a very similar fragrance, the origin of the commercial terminology is clear. The commercial source of piperonal, however, is safrole, the formaldehyde ketal (acetal) of 3,4-dihydroxyallylbenzene. Safrole is the chief constituent of the oil of sassafras and is itself obtained commercially from oil of camphor. When safrole is heated in the presence of a strong base, isomerization of the side-chain double bond yields isosafrole. Subsequent oxidation of the isomerized double bond yields piperonal. For more detailed discussions of odor–fragrance molecules and their role in living systems see Sequence D and Experiment [11C].

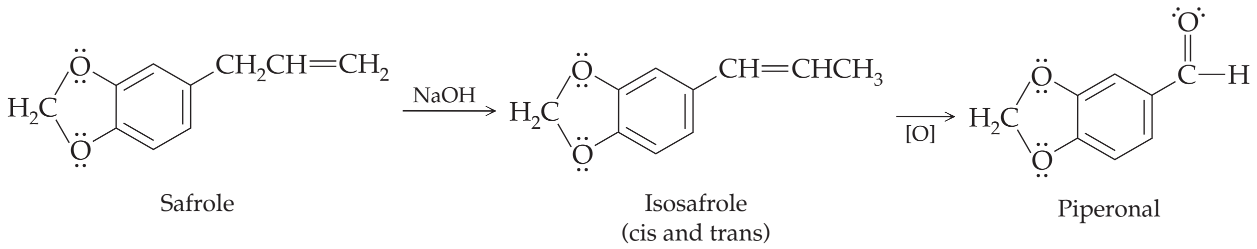

The second step of the sequence is the conversion of the intermediate aldehyde to a substituted oxime. By design, the oxime intermediate is constructed with a good leaving group as the oxygen substituent on the nitrogen atom. The second stage, therefore, creates a molecular system prone to undergo the subsequent desired elimination reaction:

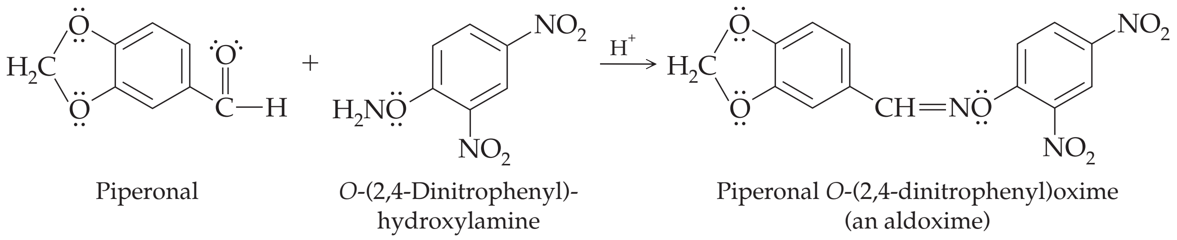

Treatment of the oxime intermediate with alcoholic KOH causes an elimination reaction that yields the target nitrile product, piperonylonitrile:

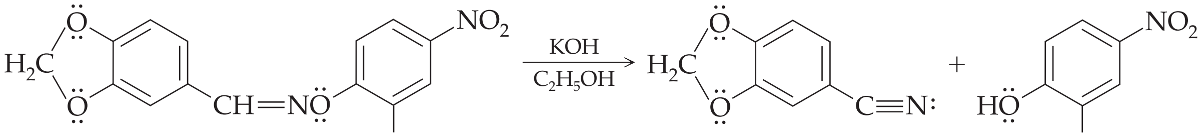

Piperonal *O*-(2,4-dinitrophenyl)oxime (an aldoxime)   Piperonylonitrile   2,4-Dinitrophenol

The present synthetic sequence offers an alternative approach to the preparation of aromatic nitriles. The classic route to this class of compounds is to use the Sandmeyer reaction. The key steps in this latter pathway involve the diazotization of an aromatic amine followed by reaction of the diazonium salt, which is not isolated (it is explosive!), with CuCN to give the aromatic nitrile directly. Aromatic halides (Br or Cl) may also be prepared in this manner using the corresponding copper(I) halides, as shown here:

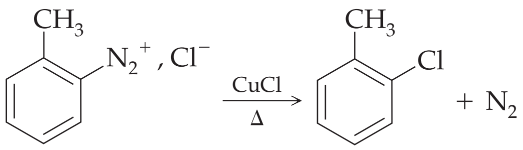

Piperonal

Experiment E1

Common names: piperonal, 3,4-(methylenedioxy)benzaldehyde
CA number: [120-57-0]
CA name as indexed: 1,3-benzodioxole-5-carboxaldehyde

Purpose. The selective oxidation of a primary (1°) alcohol to an aldehyde is investigated. The use of polymer-bound chromium trioxide as an oxidizing agent is explored. You will prepare the intermediate aldehyde, piperonal, as the first step in the sequence to synthesize piperonylonitrile.

Prior Reading

Technique 5: Crystallization
  Use of the Hirsch Funnel (pp. 88–89)
Technique 6: Chromatography
  Column Chromatography (pp. 92–95)
  Concentration of Solutions (pp. 101–104)

REACTION

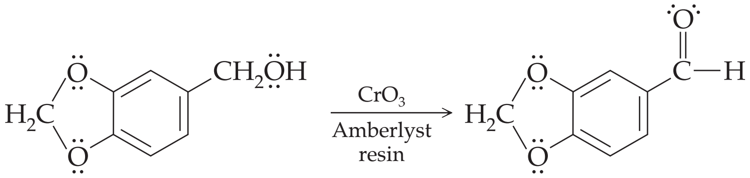

Piperonyl alcohol   Piperonal

DISCUSSION

Many of the preferred reagents for the oxidation of primary alcohols to aldehydes (secondary alcohols to ketones) contain the transition metal chromium in its highest oxidation state, VI. Upon reaction with an alcohol, the yellow-orange chromium(VI) species is reduced to the blue-green chromium(III) state. Normally the reaction is carried out in aqueous acid solution using the sodium dichromate salt, $Na_2Cr_2O_7$, or the oxide, CrO_3. A typical reaction is shown here:

$$3\ CH_3CH_2CH_2OH + 4\ H_2SO_4 + Na_2Cr_2O_7 \longrightarrow 3\ CH_3CH_2CHO + Cr_2(SO_4)_3 + Na_2SO_4 + 7\ H_2O$$

In this case, the aldehyde, propanal, can be isolated in moderate yield because it is relatively volatile and can be removed from the reaction mixture by distillation as it is formed. If not removed, the aldehyde, which is in equilibrium with the corresponding hydrate (geminal diol), is oxidized further to the carboxylic acid:

$$\underset{\text{1-Propanol}}{CH_3CH_2CH_2OH} \xrightarrow[H_2SO_4]{Na_2Cr_2O_7} \underset{\text{Propanal}}{CH_3CH_2CHO} \overset{H_2O}{\rightleftharpoons} \underset{\text{1,1-Dihydroxypropane}}{CH_3CH_2CH(OH)_2} \xrightarrow[H_2SO_4]{Na_2Cr_2O_7} \underset{\text{Propanoic acid}}{CH_3CH_2COOH}$$

Because the hydrate is the species oxidized to the acid, oxidation of the aldehyde can largely be prevented by running the oxidation in a nonaqueous medium.

A number of selective oxidizing agents, such as pyridinium chlorochromate, are frequently used for this purpose. With this reagent, the oxidation stops at the aldehyde stage, because the oxidation, as pointed out earlier, is conducted in a nonaqueous solution. Thus, decanal can be obtained from 1-decanol using this reagent (methylene chloride solvent) in 92% yield. Pyridinium chlorochromate (PCC) is a solid, yellow salt prepared from chromium trioxide as shown here:

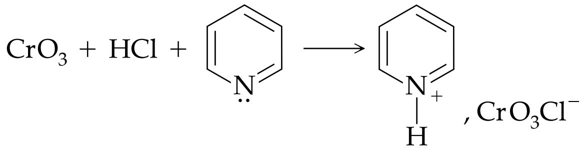

Pyridine

Pyridinium chlorochromate

An alternate reagent is the chromium trioxide(pyridine)$_2$ complex, $CrO_3 \cdot (py)_2$ (where py = pyridine), often referred to as the Collins reagent:

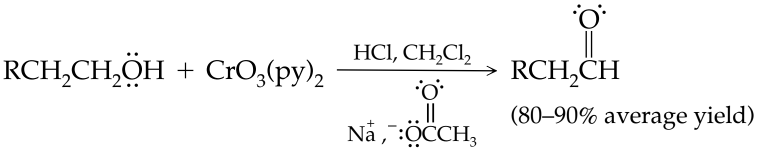

$$RCH_2CH_2\ddot{O}H + CrO_3(py)_2 \xrightarrow[Na^+,\ ^-:\ddot{O}C(=O)CH_3]{HCl,\ CH_2Cl_2} RCH_2CH{=}O$$

(80–90% average yield)

The chief difficulty with this reagent is that the complex is highly hygroscopic. However, it can be prepared in situ, thus avoiding this major drawback. Pyridinium dichromate and chromium trioxide · 3,5-dimethylpyrazole are also effective as selective oxidizing agents for these reactions.

In the present experiment, the selective oxidation of a 1° alcohol to an aldehyde is carried out using a chromium trioxide polymer-bound oxidizing agent. This reagent is also used in Experiment [33A] to oxidize a 2° alcohol to a ketone. As noted there, the reagent is easy to prepare and has the advantage that the toxic reduced chromium species is bound to a polymer resin and can thus be easily separated from the reaction mixture by simple filtration.

As shown here, the oxidation proceeds through the formation of a chromate ester. Note that the oxidation state of chromium, Cr(VI), at this stage remains the same as in the starting reagent. The second stage is equivalent to an E2 elimination reaction, with the water molecule acting as a base. The donation of an electron pair to the metal atom changes its oxidation state to Cr(IV):

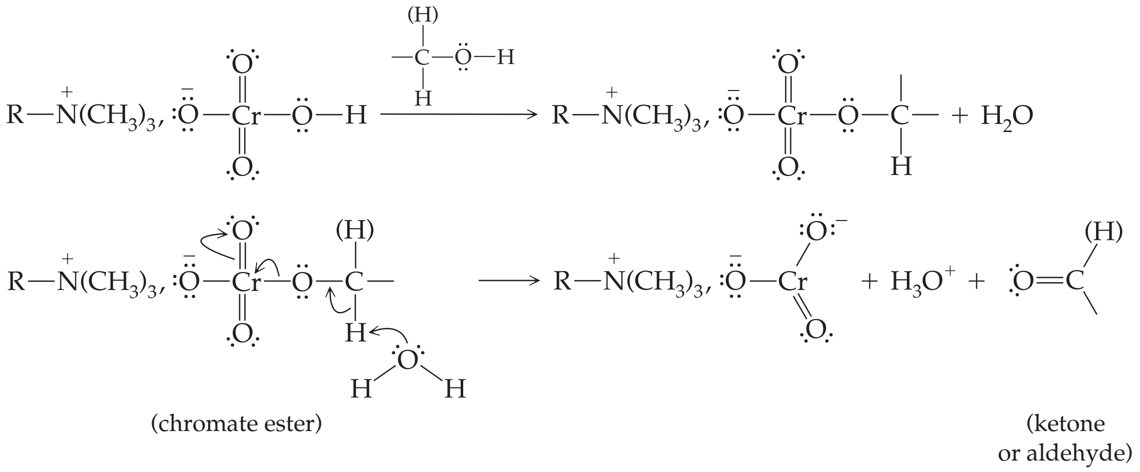

NOTE. If you plan to carry out the total sequence of reactions, the scaled-up procedure should be followed. If an individual oxidation reaction is to be studied, the microscale procedure may be used.

EXPERIMENTAL PROCEDURE

Semimicroscale Preparation. (This procedure is scaled up to 10 times the amounts of the microscale preparation.)

Estimated time to complete the experiment: 3.5 h.

Physical Properties of Reactants						
Compound	**MW**	**Amount**	**mmol**	**mp (°C)**	**bp (°C)**	***d***
Piperonyl alcohol	152.16	1.0 g	6.57	58		
Chromic oxide resin		7.5 g				
Dioxane	88.12	10 mL			101	1.034

Reagents and Equipment. Weigh and place 1.0 g (6.57 mmol) of piperonyl alcohol and 7.5 g of freshly prepared chromic oxide resin in a 25-mL round-bottom flask containing a magnetic stirring bar. Using a graduated cylinder in the **hood,** measure and add to the flask 10 mL of dioxane. Attach the flask to a reflux condenser (➡).

CAUTION: Dioxane is a cancer suspect agent. It should be dispensed and handled in the **hood.** That is, the entire experiment should be set up and run in the **hood,** at least until the solvent is removed.

INSTRUCTOR PREPARATION. The CrO_3 resin is prepared by adding 35 g of Amberlyst A-26 resin to a solution of 15 g of CrO_3 in 100 mL of water. Stir the mixture

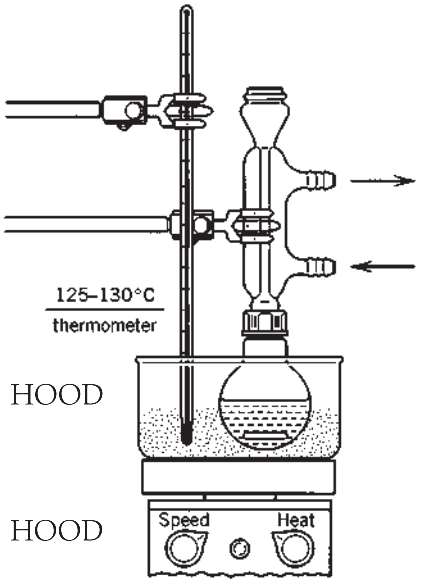

25–mL RB flask
piperonyl alcohol, 1.0 g
+ chromic oxide oxidizing resin, 7.5 g
+ dioxane, 10 mL

for 30 min at room temperature. Collect the resin by filtration under reduced pressure and successively rinse the material with water and acetone. Partially dry the resin on a Büchner funnel by drawing air through it for 1 h, then air-dry the material overnight.

Reaction Conditions. With stirring, heat the reaction mixture at reflux in HOOD the **hood** for a period of 1.0 h using a sand bath temperature of 125–130 °C.

Isolation of Product. Cool the reaction product to room temperature. Transfer the solution by Pasteur filter pipet to a funnel fitted with fast-grade filter paper. Collect the filtrate in a 25-mL Erlenmeyer flask containing approximately 1 g of anhydrous sodium sulfate. Rinse the reaction flask and resin (include the walls of the condenser) with three 1.0-mL portions of methylene chloride solvent (calibrated Pasteur pipet). Transfer each rinse by Pasteur filter pipet to the funnel. Combine each rinse filtrate with the original filtrate.

Now transfer the solution in three portions by Pasteur filter pipet to a 10-mL round-bottom flask and concentrate the solution following each addition using a rotary evaporator. If a rotary evaporator is not available, transfer the solution to a 25-mL Erlenmeyer flask containing a boiling stone and remove HOOD the solvent in the **hood** by warming the flask on a sand bath at 125 °C. A gentle stream of nitrogen aids this concentration process.

The crude piperonal is obtained as an oil and is purified by column chromatography.

Purification and Characterization. Preweigh (and number 1–10) ten 25-mL Erlenmeyer flasks containing a boiling stone. Wet-pack a 25-mL buret using hexane and silica gel in the following manner. Place a cotton plug in the buret followed by a 1-in. layer of sand. Add 15 mL of hexane to the column, and then slowly add 8 g of silica gel while tapping the sides of the buret to release any air bubbles. Finally, add an additional 1 in. of sand to the top of the column.

Allow the hexane to drain from the buret until it reaches the top of the sand level in the column. Collect the eluted hexane in a waste container.

Dilute the crude piperonal oil with 5 drops of methylene chloride and then, using a Pasteur filter pipet, add this solution to the top of the column. Again, drain a portion of the elution solvent until the solvent head is level with the sand at the top of the column. Rinse the sides of the column with 0.5 mL of 4:1 methylene chloride/hexane solution, and again drain the column to the same sand level.

Now add, in order, the following amounts of elution solvents: three 10-mL portions of 4:1 methylene chloride/hexane; three 10-mL portions of methylene chloride; and three 10-mL portions of 9:1 methylene chloride/ethyl acetate.

Collect 10-mL fractions in each of the preweighed Erlenmeyer flasks. After collecting fraction No. 9, use flask No. 10 to collect a final fraction as the column completely drains. (Fraction No. 1 can be added directly to the waste container.)

Concentrate each of the collected fractions to dryness on a warm sand HOOD bath under a gentle stream of nitrogen gas in the **hood.** Weigh the flasks to determine the total amount of piperonal product. Now preweigh an additional 25-mL Erlenmeyer flask containing a boiling stone. To each of the flasks containing white or light yellow fanlike crystals, add sufficient methylene chloride to just dissolve them and then transfer the resulting solutions to the tared Erlenmeyer flask using a Pasteur filter pipet. Concentrate this final solution as before, and weigh the flask to obtain the total amount of piperonal isolated.

NOTE. If a rotary evaporator is available, the eluate can be concentrated in a 10-mL round-bottom flask. Transfer of each fraction to a single evaporation flask shortens the above procedure.

Determine the melting point of your product and compare your value to that found in the literature. Obtain IR and/or ^{1}H NMR spectra and compare your results to those reported in the literature (*The Aldrich Library of IR Spectra, The Aldrich Library of NMR Spectra,* and/or the corresponding spectral data available online (e.g., SciFinder Scholar)).

Chemical Tests. Chemical classification data may also be useful. The ignition test, the Tollens test, and the 2,4-dinitrophenylhydrazine test should all give a positive result for the piperonal compound.

NOTE. Approximately 300 mg of purified product with a melting point within 2–4 °C of the literature value is suggested for continuing the sequence on to Experiment [E2].

OPTIONAL MICROSCALE PREPARATION

The microscale preparation is similar to that of the scaled-up synthesis outlined above, with the following exceptions.

*NOTE. Even in the case of the microscale preparation this reaction should be **entirely** carried out in the **hood** at all times until the dioxane solvent has been removed.*

HOOD

1. The reagent and solvent amounts are one-tenth of those used above.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
Piperonyl alcohol	152.16	100 mg	0.66	58	
Chromic oxide resin		750 mg			
Dioxane	88.12	1.0 mL			101

2. A 5-mL round-bottom flask is used.

3. The reaction mixture is heated for 1 h.

4. After the reaction product is cooled to room temperature, transfer the solution by Pasteur filter pipet to a funnel containing a loose cotton plug covered with 500 mg of anhydrous sodium sulfate. Collect the filtrate in a 10-mL Erlenmeyer flask. Rinse the reaction flask and resin (include the walls of the condenser) with three 0.5-mL portions of methylene chloride solvent (calibrated Pasteur pipet). Transfer each rinse by Pasteur filter pipet to the funnel and, finally, rinse the sodium sulfate with an additional 0.1 mL of methylene chloride. Combine each rinse filtrate with the original filtrate. Concentrate the solution as described.

5. Purify the crude oil using chromatography as follows:

(a) Pack a short buret column with approximately 1.75 g of silica gel, and transfer the crude product by Pasteur filter pipet to the column. Rinse the flask with 1.0 mL of methylene chloride (calibrated Pasteur pipet) and also transfer the rinse to the column.

(b) Now add methylene chloride to the column, 2.0 mL at a time (calibrated Pasteur pipet), and collect the eluate in three tared 10-mL Erlenmeyer flasks. Set aside the first 4.0 mL of eluate collected in the first flask. Collect the next 6.0 mL in one flask; this fraction contains the bulk of the product. Also collect one additional fraction of 3.0 mL in the third flask.

HOOD

(c) Concentrate the second fraction (6.0 mL) in the **hood** using a warm sand bath with a slow stream of nitrogen to assist solvent removal. *Do not forget to add a boiling stone to the flask.*

(d) Weigh the flask and contents. Reheat for 1 min, cool, and weigh again. If the two weights are within 2.0 mg, the product is quite pure. If not, repeat the evaporation process until a constant weight is obtained.

(e) Cool the product. If it does not solidify, place it in an ice bath and scratch the sides and bottom with a glass rod to induce crystallization.

A small amount of additional product can be isolated from the third fraction in a like manner.

QUESTIONS

7-76. Primary alcohols can be oxidized to aldehydes and carboxylic acids. Often it is difficult to stop at the aldehyde oxidation state. One method frequently used to accomplish this, for those that boil below about 100 °C is to remove the product from the reaction mixture as it is formed. This method is based on the fact that aldehydes have a lower boiling point than their corresponding alcohols. Explain this difference in boiling point.

7-77. A specific oxidizing agent for the conversion of primary alcohols to aldehydes is pyridinium chlorochromate, abbreviated as ${}^{+}\text{py} \cdot \text{CrO}_3\text{Cl}^{-}$. Generally, the oxidation is run in methylene chloride solvent. For example,

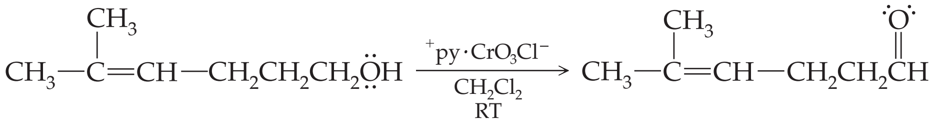

$$\mathrm{CH_3{-}\overset{\overset{\displaystyle CH_3}{|}}{C}{=}CH{-}CH_2CH_2CH_2\ddot{\underset{\cdot\cdot}{O}}H} \xrightarrow[\substack{\mathrm{CH_2Cl_2}\\ \mathrm{RT}}]{{}^{+}\mathrm{py}\cdot \mathrm{CrO_3Cl^-}} \mathrm{CH_3{-}\overset{\overset{\displaystyle CH_3}{|}}{C}{=}CH{-}CH_2CH_2\overset{\overset{\displaystyle \ddot{O}}{\|}}{C}H}$$

Oxidation of this alcohol with the conventional $Na_2Cr_2O_7$–H_2SO_4–water system produces the carboxylic acid. Offer an explanation for the difference in these results.

7-78. In reference to Question 7-77, can you see another advantage of the pyridine chlorochro-mate reagent over that of the conventional conditions?

7-79. An excellent way to identify the presence of an aldehyde group is by ^{1}H NMR. The chemical shift of the aldehydic proton occurs in the 9- to 10-ppm region, where few other proton signals occur. Why is this chemical shift so far downfield?

7-80. A series of compounds with increasing boiling point is listed below. Offer an explanation for this trend.

(a) Butane (8 °C)
(b) Propanal (56 °C)
(c) 1-Propanol (97 °C)

BIBLIOGRAPHY

The present procedure is based on the work reported by

Cainelli, G.; Cardillo, G.; Orena, M.; Sandri, S. *J. Am. Chem. Soc.* **1978,** *98,* 6737.

Several preparations involving the oxidation of a 1° alcohol to an aldehyde are given in *Organic Syntheses:*

Collins, J. C.; Hess, W. W. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 644.

Hurd, C. D.; Meinert, R. N. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 541.

Ratcliffe, R. W. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 373.

Also see

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Wiley: New York, 2007, Chap. 19, p. 1750 and references therein.

Piperonal *O*-(2,4-Dinitrophenyl)oxime

Experiment E2

Common name: piperonal *O*-(2,4-dinitrophenyl)oxime
CA number: [17188-61-3]
CA name as indexed: 1,3-benzodioxole-5-carboxaldehyde, *O*-(2,4-dinitrophenyl)oxime

Purpose. You will carry out the second step in the sequence of synthetic reactions leading to an aromatic nitrile, piperonylonitrile. You will prepare a specific oxime derivative of the aldehyde obtained in Experiment [E1]. This oxime derivative is derivatized on oxygen so as to allow the oxygen to function as a good leaving group, which will allow an elimination reaction to form a nitrile in the final step of the synthetic sequence leading to piperonylonitrile.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)

REACTION

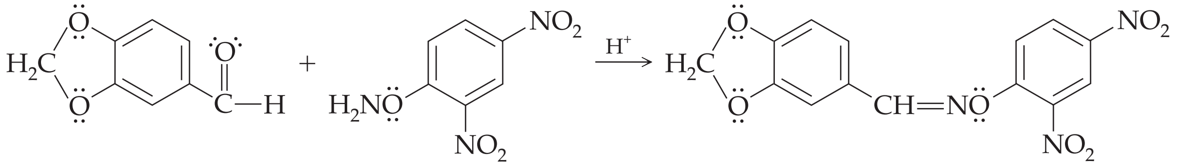

Piperonal   *O*-(2,4-Dinitrophenyl)-hydroxylamine   Piperonal *O*-(2,4-dinitrophenyl)oxime

DISCUSSION

The preparation of this oxime is the second step in this sequence to obtain the aromatic nitrile target molecule. In Experiment [E2] you are going to convert the aromatic aldehyde formed in Experiment [E1] into an *O-phenylated oxime,* which on treatment with base in Experiment [E3] yields the desired nitrile via an elimination reaction. Oxime formation is also involved in the well-known Beckmann rearrangement (see Experiment [6_{adv}]), which is used for the synthesis of amides.

The generation of the oxime intermediate involves a nucleophilic addition of the amine group of the hydroxylamine to the carbonyl carbon, followed by a dehydration to form the carbon–nitrogen double bond (and the oxime group). A general mechanism for the reaction is given here.

The first step is a nucleophilic addition to the carbonyl group. Rapid proton transfer gives an intermediate that generally is not isolated:

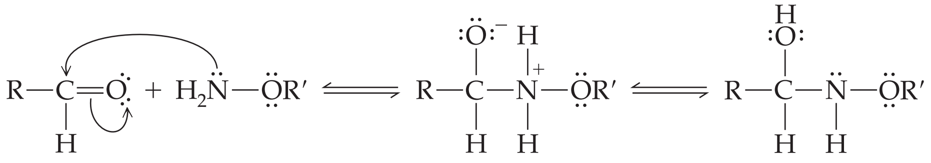

$$\mathrm{R}-\underset{\mathrm{H}}{\underset{|}{\mathrm{C}}}=\ddot{\mathrm{O}}\!: + \mathrm{H_2}\ddot{\mathrm{N}}-\ddot{\mathrm{O}}\mathrm{R'} \rightleftharpoons \mathrm{R}-\overset{:\ddot{\mathrm{O}}:^-}{\overset{|}{\underset{\mathrm{H}}{\underset{|}{\mathrm{C}}}}}-\overset{\mathrm{H}}{\overset{|}{\underset{\mathrm{H}}{\underset{|}{\mathrm{N}^+}}}}-\ddot{\mathrm{O}}\mathrm{R'} \rightleftharpoons \mathrm{R}-\overset{\mathrm{H}}{\overset{:\ddot{\mathrm{O}}:}{\overset{|}{\underset{\mathrm{H}}{\underset{|}{\mathrm{C}}}}}}-\underset{\mathrm{H}}{\underset{|}{\ddot{\mathrm{N}}}}-\ddot{\mathrm{O}}\mathrm{R'}$$

The next stage is an acid-catalyzed dehydration reaction in which water is eliminated to produce the oxime. This stage has been shown to be the rate-determining step in oxime formation.

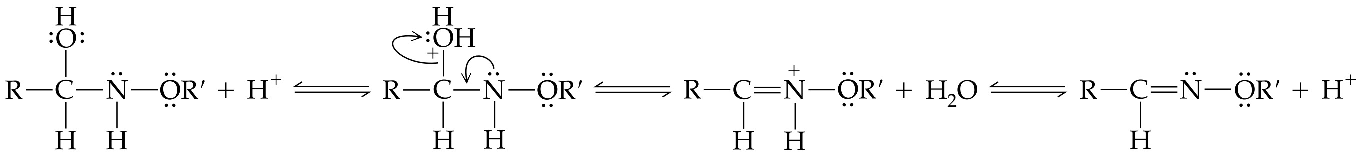

Oxime
(an aldoxime)

NOTE. If you are conducting the total sequence of reactions to obtain piperonylonitrile, follow the scaled-up procedure. If the individual reaction is to be studied, follow the microscale procedure.

EXPERIMENTAL PROCEDURE

Semimicroscale Preparation. (This procedure is scaled up to 10 times the amounts of the microscale preparation.)

Estimated time to complete the experiment: 1 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Piperonal	150.14	300 mg	2.0	37	263	
O-(2,4-Dinitrophenyl) hydroxylamine	199.12	400 mg	2.0	111–112		
Ethanol	46.07	30 mL			78	0.789
HCl (12 M)		2 mL				
Piperonal *O*-(2,4-dinitrophenyl)oxime	330.24			194–195		

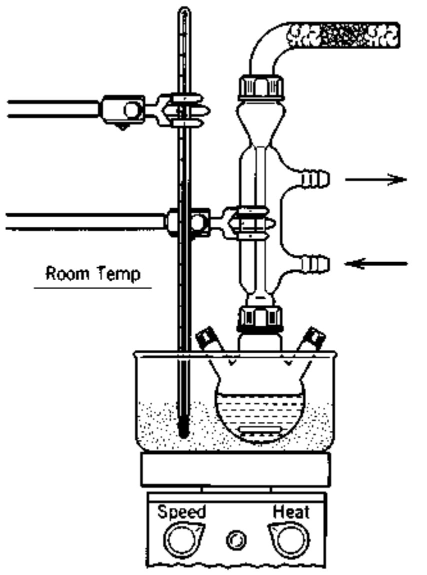

100-mL RB three necked flask containing *O*-(2,4-dinitrophenyl)hydroxylamine, 400 mg + absolute ethanol, 30 mL + piperonal, 300 mg + 12 *M* HCl, 20 mL

Reagents and Equipment. To a 100-mL three-necked round-bottom flask containing a magnetic stir bar and equipped with a reflux condenser protected by a calcium chloride drying tube and two caps or glass stoppers, weigh and place 400 mg (2.0 mmol) of *O*-(2,4-dinitrophenyl)hydroxylamine, followed by 30 mL of absolute ethanol (⬅).

NOTE. The three-necked flask may be replaced with a conventional round-bottom 100-mL flask. In this case, addition of reagents can be carried out in the first instance by removing the condenser and in the second case by addition directly down the condenser with a 9-in. Pasteur pipet.

www → *NOTE. The preparation of O-(2,4-dinitrophenyl)hydroxylamine is described in the Instructor's Manual, which is available from the publisher.*

Attach the flask to the condenser and warm (60–65 °C) the contents, with stirring, in a sand bath until a homogeneous solution is obtained. Now add 300 mg (0.2 mmol) of piperonal via one of the unused flask necks. After the aldehyde has dissolved, add slowly, with stirring, 2 mL of 12 M HCl through an unused flask neck using a Pasteur pipet. Stir the reaction mixture for an additional 1 min after the addition is complete.

Reaction Conditions. The oxime forms immediately. Complete the precipitation of the product by cooling the reaction flask in an ice bath for 10 min.

Isolation of Product. Collect the solid product by filtration under reduced pressure using a Hirsch funnel. Rinse the flask with two 3-mL portions of cold absolute ethanol, using the rinse to wash the crystals on the filter funnel. Now rinse the crystals with three additional 3-mL portions of cold absolute ethanol (➨).

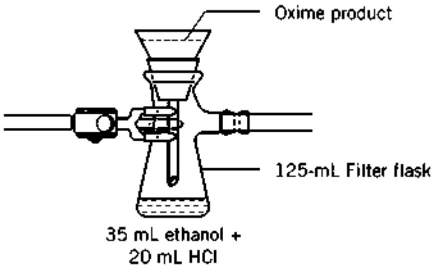

NOTE. Approximately 50–70 mg of purified product with a melting point within 2–4 °C of the literature value is the minimum quantity and quality suggested for continuing the sequence in Experiment [E3].

Air-dry the product on a clay plate or on filter paper. Collect and refrigerate the filtrate for at least 24 h. This procedure generally produces another crop of oxime crystals. This second crop, collected by the same technique, may be combined with the initial product, if its melting point is above 180 °C.

Purification and Characterization. Determine the melting point of the oxime. It is of sufficient purity to proceed with the next step of the sequence if its mp is 187 °C or greater. If necessary, recrystallize the material from acetic acid or hot acetone.

Obtain an IR spectrum of the oxime and compare it with that of the reference standard shown in Figure 7.6.

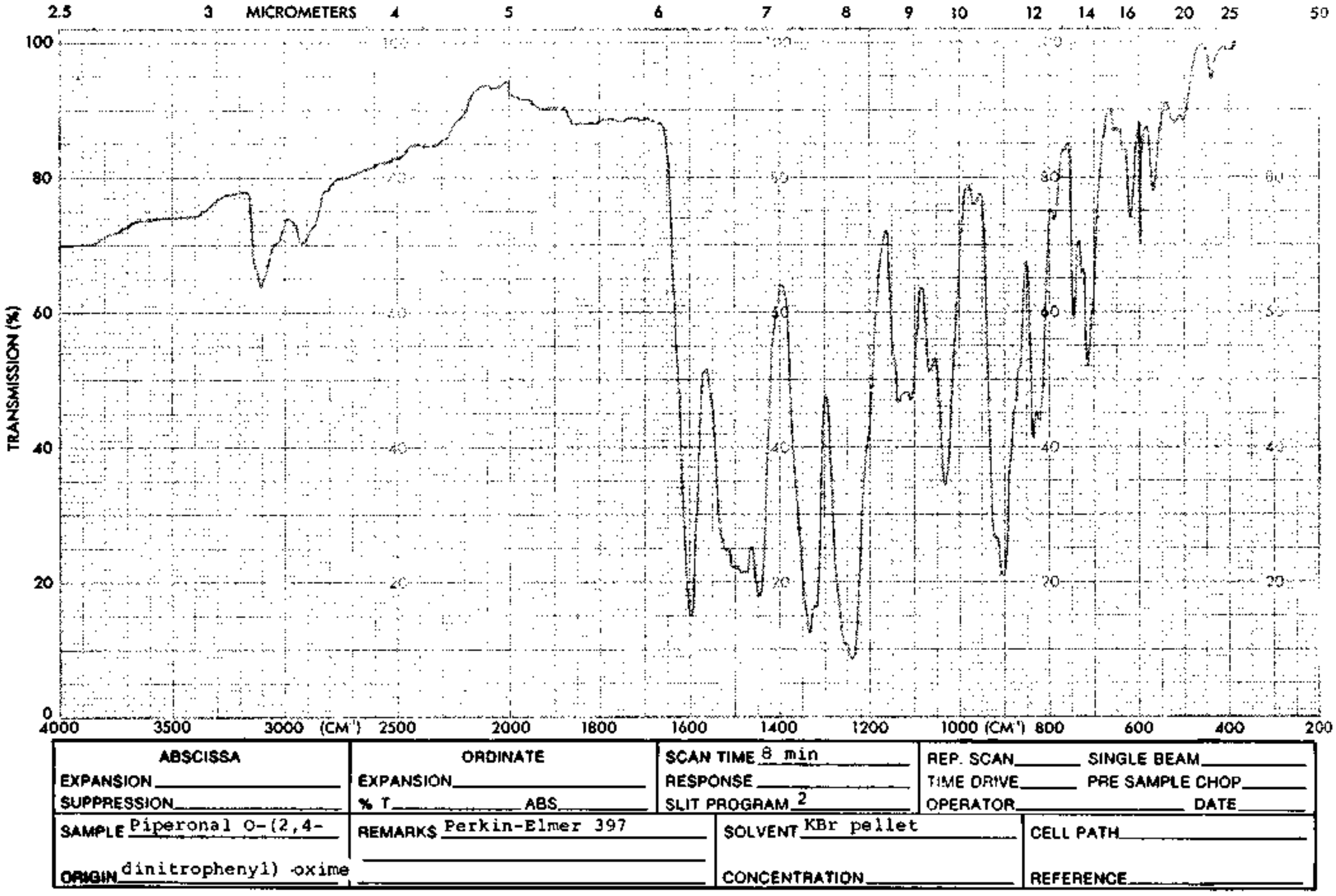

EXPANSION
SUPPRESSION
EXPANSION
% T ABS
SCAN TIME 8 min
RESPONSE
SLIT PROGRAM 2
REP. SCAN SINGLE BEAM
TIME DRIVE PRE SAMPLE CHOP
OPERATOR DATE
SAMPLE Piperonal O-(2,4-
ORIGIN dinitrophenyl) oxime
REMARKS Perkin-Elmer 397
SOLVENT KBr pellet
CONCENTRATION
CELL PATH
REFERENCE

Figure 7.6 IR spectrum: piperonal *O*-(2,4-dinitrophenyl)oxime.

OPTIONAL MICROSCALE PREPARATION

The microscale preparation is similar to that of the scaled-up synthesis outlined above, with the following exceptions:

1. The reagent and solvent amounts are one-tenth of those used above.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Piperonal	150.14	30 mg	0.2	37	263	
O-(2,4-Dinitrophenyl)hydroxylamine	199.12	40 mg	0.2	111–112		
Ethanol	46.07	3.0 mL			78	0.789
HCl (12 M)		2 drops				
Piperonal *O*-(2,4-dinitrophenyl)oxime	330.24			194–195		

2. A 5-mL conical vial is used.

3. The hydrochloric acid is added dropwise through the top of the condenser using a 9-in. Pasteur pipet, after removing the drying tube. As the acid is delivered, the tip of the pipet should be held just above the surface of the solution. The drying tube is then reattached, and the contents are mixed by swirling the flask.

QUESTIONS

7-81. Referring to the first step in the mechanism of oxime formation outlined in the discussion section, offer an explanation of why a high acid concentration would hinder the formation of the intermediate generated in this first stage.

7-82. Oximes prepared from aldehydes or ketones by reaction with hydroxylamine can be reduced to primary amines in high yields. One can use various reagents for the reduction step, including Ni–H_2 in methanol or $LiAlH_4$ in ether. Using a reduction reaction, carry out the following transformations:

(a) 3-Pentanone to 3-aminopentane
(b) Propanal to 1-aminopropane
(c) Benzaldehyde to benzylamine

7-83. Oximes prepared from unsymmetrical ketones are likely to exist as mixtures of geometrical isomers. For example,

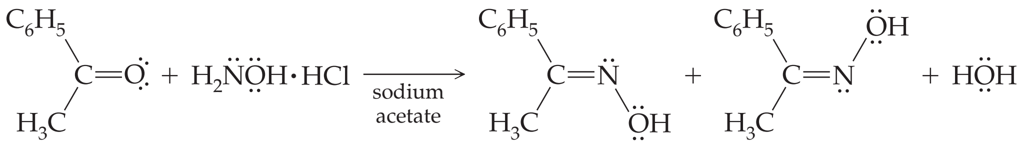

(*E*)- and (*Z*)-isomeric oximes

Which isomer is designated as *E*? How do you account for both of these isomers being formed?

BIBLIOGRAPHY

For the preparation of oximes using hydroxylamine hydrochloride as the reagent see

Bousquet, E. W. *Organic Syntheses;* Wiley: New York, 1943, Collect. Vol. II, p. 313.

Fuson, R. C.; Curtin, D. Y.; Morrill, T. C.; Hermann, C. K. F.; Shriner, R. L. *The Systematic Identification of Organic Compounds*, 7th ed.; Wiley: New York, 1998.

Hach, C. C.; Banks, C. V.; Diehl, H. *Organic Syntheses;* Wiley: New York, 1963, Collect. Vol. IV, p. 229.

Lachman, A. *Organic Syntheses;* Wiley: New York, 1943, Collect. Vol. II, p. 70.

Pasto, D. J.; Johnson, C. R.; Miller, M. J. *Experiments and Techniques in Organic Chemistry;* Prentice Hall: Englewood Cliffs, NJ, 1992, p. 332.

Semon, W. W.; Damerell, V. R. *Organic Syntheses;* Wiley: New York, 1943, Collect. Vol. II, p. 204.

Smith, M. B.; March, J. *Advanced Organic Chemistry*, 6th ed., Wiley: New York, 2007, Chap. 16, p. 1286 and references therein.

Piperonylonitrile

Experiment E3

Common names: piperonylonitrile, 3,4-methylenedioxybenzonitrile
CA number: [4421-09-4]
CA name as indexed: 1,3-benzodioxole-5-carbonitrile

Purpose. The piperonal *O*-(2,4-dinitrophenyl)oxime intermediate, prepared in the previous experiment (Experiment [E2]), is converted into the target molecule, piperonylonitrile. This completes the set of Sequence E reactions for the conversion of a substituted benzyl alcohol into an aromatic nitrile. You will investigate the use of a novel elimination reaction to convert an oxime derivative to a nitrile.

Prior Reading

Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
Technique 6: Chromatography
  Column Chromatography (pp. 92–95)
  Concentration of Solutions (pp. 101–104)

REACTION

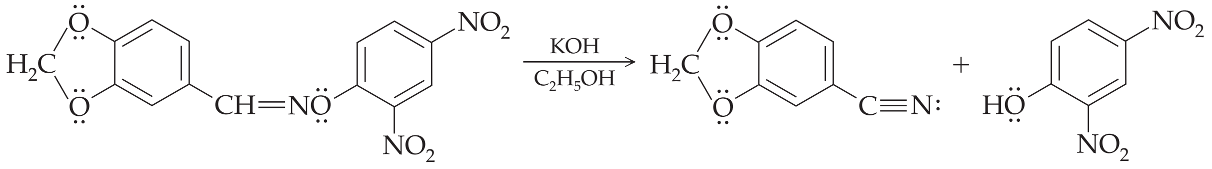

Piperonal *O*-(2,4-dinitrophenyl)oxime Piperonylonitrile 2,4-Dinitrophenol

DISCUSSION

The Sequence E reactions illustrate the oxidation of a benzylic alcohol to its corresponding aldehyde and the subsequent two-step conversion of this aldehyde via an intermediate aldoxime to an aromatic nitrile. This synthetic route is an attractive alternative to the preparation of aromatic nitriles via the Sandmeyer reaction.

Aromatic nitriles are usually prepared by the diazotization of the corresponding aromatic amine followed by treatment of the diazonium salt with copper(I) cyanide. This sequence is an example of the Sandmeyer reaction, a specific example of which is shown here:

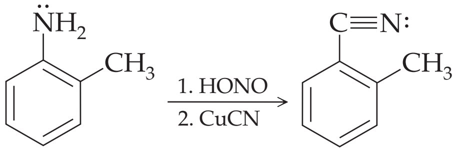

(64–70%)

This particular variation of the Sandmeyer reaction is useful because it allows the conversion of an aromatic amine, readily available by reduction of the corresponding nitro compound, to a reactive carbon substituent. This substitution involves replacement of the C—N bond with a C—C bond.

In the present reaction, treatment of an *O*-phenylated oxime with base yields the nitrile by an elimination reaction. The role of the dinitrophenyl group is to enhance the oxime oxygen to function as a good leaving group. Thus, the phenoxy system departs as the conjugate base of the relatively acidic 2,4-dinitrophenol. The mechanistic sequence is outlined below:

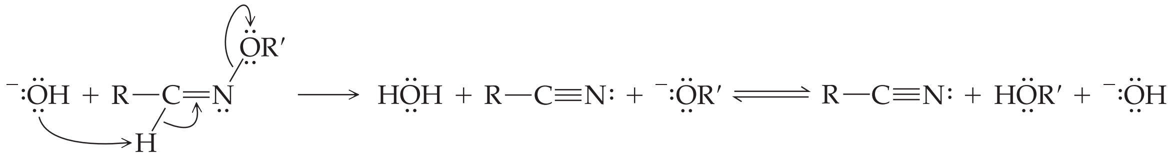

Oximes derived from aldehydes (aldoximes) can be dehydrated to nitriles by many different dehydrating reagents; acetic anhydride is one of the most common reagents used:

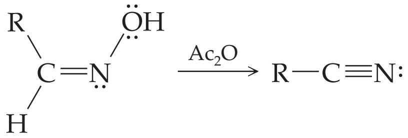

The reaction proceeds more rapidly when the hydrogen and the hydroxyl group are trans to one another (see also Experiments [10], [D3], and [$A3_b$]). Various derivatives of the hydroxylamine other than the ethers, RCH═NOR, illustrated in the present reaction, also undergo the conversion to nitriles. Among these are the RCH═NOCOR and $RCH{=}NOSO_2Ar$ compounds. In some cases it is also possible to convert aldehydes to nitriles in one step by refluxing the reagents in concentrated hydrochloric acid (or by reaction with sodium formate in formic acid or sodium acetate in acetic acid) as follows:

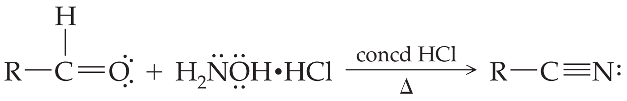

$$R{-}\overset{\overset{\displaystyle H}{|}}{C}{=}\ddot{O}\!: + H_2\ddot{N}\ddot{O}H{\bullet}HCl \xrightarrow[\Delta]{\text{concd HCl}} R{-}C{\equiv}N\!:$$

When oximes are treated with strong acid they are converted to amides by a rearrangement sequence known as the Beckmann rearrangement. This reaction is illustrated in Experiment [6_{adv}].

Nitriles are synthetically versatile functional groups because they are readily converted to carboxylic acids by hydrolysis under acidic or basic conditions, reduced with $LiAlH_4$ to form primary amines, and reaction with Grignard reagents leads to the formation of ketones. These reactions are illustrated here:

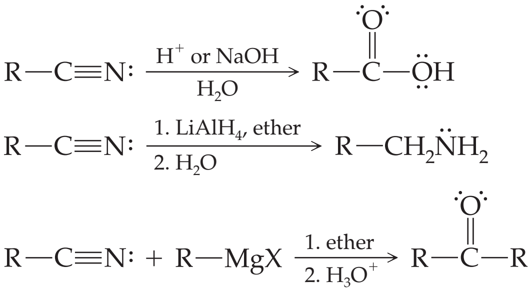

$$R{-}C{\equiv}N\!: \xrightarrow[H_2O]{H^+ \text{ or NaOH}} R{-}\overset{\overset{\displaystyle \ddot{O}}{\|}}{C}{-}\ddot{O}H$$

$$R{-}C{\equiv}N\!: \xrightarrow[\text{2. } H_2O]{\text{1. } LiAlH_4\text{, ether}} R{-}CH_2\ddot{N}H_2$$

$$R{-}C{\equiv}N\!: + R{-}MgX \xrightarrow[\text{2. } H_3O^+]{\text{1. ether}} R{-}\overset{\overset{\displaystyle \ddot{O}}{\|}}{C}{-}R$$

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Piperonal *O*-(2,4-dinitrophenyl)oxime	330.24	50 mg	0.15	194–195		
Ethanol		5 mL			78.5	0.789
Potassium hydroxide (0.2 M)		2 mL				
Piperonylonitrile	147.13			92–93		

Reagents and Equipment. Weigh and place 50 mg (0.15 mmol) of piperonal *O*-(2,4-dinitrophenyl)oxime in a 10-mL round-bottom flask containing a magnetic stir bar. Now add 5.0 mL of 95% ethanol and 2.0 mL of 0.2 M ethanolic KOH. Attach the flask to a reflux condenser (➡).

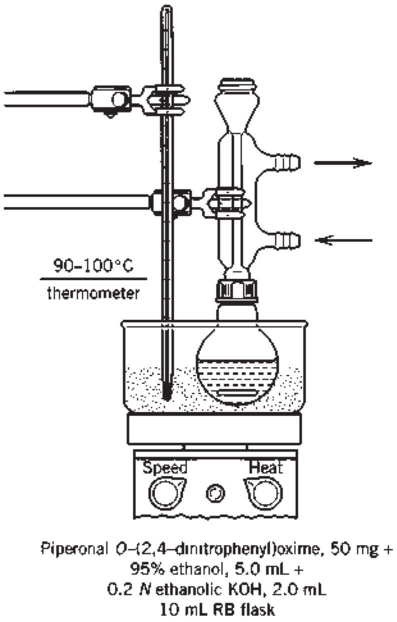
90–100°C
thermometer
Speed
Heat
Piperonal O–(2,4–dinitrophenyl)oxime, 50 mg +
95% ethanol, 5.0 mL +
0.2 N ethanolic KOH, 2.0 mL
10 mL RB flask

NOTE. The oxime is prepared in Experiment [E2]. The 0.2 M ethanolic KOH is prepared using 95% ethanol.

Reaction Conditions. Slowly heat the reaction mixture, while stirring, to reflux by use of a sand bath (100–110 °C) and maintain the mixture at this temperature (gentle reflux) for 1 h. During the initial warming period, the solution turns a deep yellow.

NOTE. If the laboratory time is not sufficient to continue the isolation and purification of the product after the heating is terminated, cool the solution and remove the reaction vial. Cap the vial and store it until the next laboratory period.

Isolation of Product. Remove the reaction flask and concentrate the reaction mixture to a volume of 0.5 mL or less with a gentle stream of nitrogen gas and/or warming in a sand bath in the **hood.** *This concentration process takes a considerable length of time.* HOOD

Now prepare an alkaline solution by diluting 1 mL of 5% aqueous NaOH with 5 mL of distilled water. Use this solution to transfer the concentrated reaction mixture to a 12-mL centrifuge tube in the following manner.

Add a 2-mL portion of the alkaline solution to the reaction flask, mix by swirling, and then transfer the resulting suspension to the centrifuge tube using a Pasteur pipet. Repeat this operation twice, using 2 mL of the alkaline solution each time.

Extract the resulting suspension with four 2-mL portions of methylene chloride (calibrated Pasteur pipet). Remove the methylene chloride extract (bottom layer) using a Pasteur filter pipet, and place the combined fractions in a 25-mL Erlenmeyer flask. Dry the solution over granular anhydrous sodium sulfate (0.5 g).

By use of a Pasteur filter pipet, transfer the dried solution to a second 25-mL Erlenmeyer flask containing a boiling stone. Rinse the drying agent with two 1-mL portions of methylene chloride. Combine the rinses with the original solution. Remove the solvent in the **hood** under a gentle stream of nitrogen and/or by warming in a sand bath to obtain the crude piperonylonitrile. HOOD

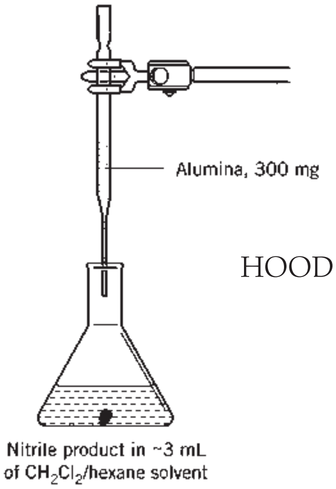

Purification and Characterization. Purify the crude product by column chromatography using a Pasteur filter pipet filled with 300 mg of alumina (neutral, activity 1, see Glossary). Wet the column with 1.0 mL of 1:1 methylene chloride/hexane solution.

Dissolve the residue of crude nitrile in a minimum amount of 1:1 methylene chloride/hexane solvent, and transfer the resulting solution by Pasteur pipet to the column. Elute the nitrile from the column with 2.0 mL of the 1:1 CH_2Cl_2/hexane solvent and collect the eluate in a 10-mL Erlenmeyer flask containing a boiling stone (⬅).

HOOD

Evaporate the solvent under a gentle stream of nitrogen while warming in a sand bath in the **hood.** Dry the white needles of piperonylonitrile on a porous clay plate or on filter paper.

Weigh the product and calculate the percent yield. Determine the melting point and compare it with the literature value.

Obtain an IR spectrum and ^{1}H NMR spectra of the sample and compare your results with those in the literature (*The Aldrich Library of IR Spectra, The Aldrich Library of NMR Spectra,* and/or the corresponding spectral data available online (e.g., SciFinder Scholar)).

QUESTIONS

7-84. The Sandmeyer reaction is based on the replacement of the diazonium group in aryldiazonium salts by chloro, bromo, or cyano groups. Copper salt reagents are used:

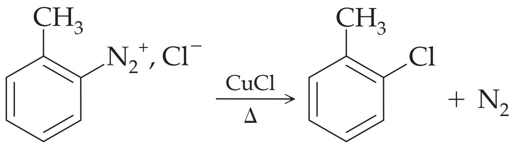

Show how one could carry out the following transformations using the Sandmeyer reaction:

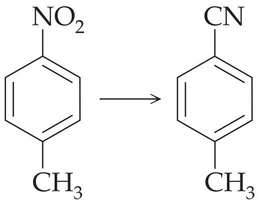

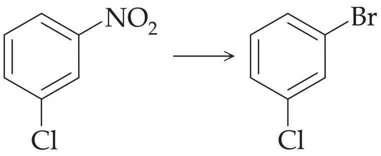

7-85. When CuCN is used in the Sandmeyer reaction, the preparation is generally carried out in a neutral medium. Can you offer an explanation of why this is done?

7-86. Outline a synthetic route for the preparation of nitriles using a carboxylic acid as the starting material.

7-87. A yellow azo dye once used to color margarine has been outlawed because it is carcinogenic. Outline a synthesis of this dye, butter yellow, starting from benzene and *N,N*-dimethylaniline:

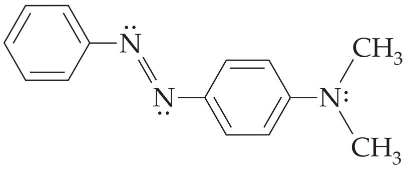

Butter Yellow

BIBLIOGRAPHY

The procedure outlined above for the preparation of piperonylonitrile is based on the work of

Miller, M. J.; Loudon, G. M. *J. Org. Chem.* **1975,** *40*, 126.

Several examples of the conversion of an aromatic aldehyde to the corresponding nitrile are given in *Organic Syntheses:*

Back, J. S.; Ide, W. S. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 622.

Clarke, H. T.; Nagy, S. M. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 690.

For a one-step conversion of aromatic aldehydes to the corresponding nitrile see

DeMott, J. M. Jr.; Kelley, C. J. *J. Chem. Educ.* **2001,** *78*, 780. This procedure works for the conversion of the present experiment as verified by DeMott, J. M., Jr. (Massachusetts College of Pharmacy and Allied Health, Boston), private communication.

There are numerous references to the use of the Sandmeyer reaction in *Organic Syntheses*. Several are cited here:

Bigelow, L. A. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, pp. 135, 136.

Cleland, G. H. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 21.

Grundstone, F. D.; Tucker, S. H. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 160.

Hartwell, J. L. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 185.

Marvel, C. S. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 170.

Rutherford, K. G.; Redmond, W. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 133.

Smith, M. B.; March, J. *Advanced Organic Chemistry*, 6th ed., Wiley: New York, 2007, Chap. 14, p. 984 and references therein.

SEQUENCE F

Introduction to Photochromism: The Synthesis of a Photochromic Imine

The photochromic effect is a property of relative rarity in both organic and inorganic molecular structures. When it is present, a material is found to exhibit reversible color change upon exposure to radiation:

$$\begin{array}{c}\text{Photochromic}\\ \text{Substance A}\end{array} \underset{\Delta \text{ or } h\nu_B}{\overset{h\nu_A}{\rightleftharpoons}} \begin{array}{c}\text{Photochromic}\\ \text{Substance B}\end{array}$$

The photoproduct generally reverts to the initial system via thermal pathways, but there are examples where the reverse reaction is induced by radiation of a different wavelength from that driving the forward reaction, or by both thermal and photochemical processes. Generally, sensitivity to thermal effects controls the concentrations obtained from the forward reactions and, therefore, their effectiveness in producing the product.

CLASSES OF PHOTOCHROMIC REACTIONS

Cis–Trans Isomerizations. Experiment [6] studies the photochromic properties of *trans*-dibenzoylethylene. In this case, the highly conjugated bright-yellow trans diastereomer is rapidly isomerized under intense sunlamp visible radiation, via excited electronic states, to the colorless cis alkene. A π electron is promoted to an anti-bonding π^* molecular orbital, which destroys the π bond, and thus permits facile rotation about the remaining σ bond and formation of the cis alkene:

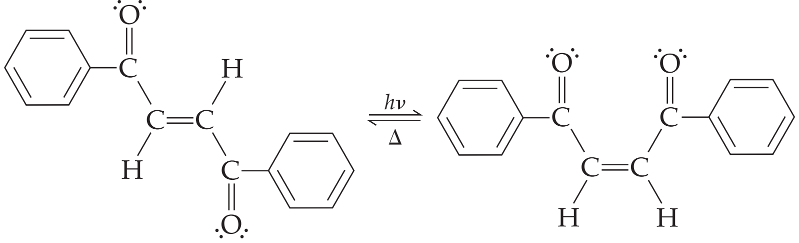

trans-1,2-Dibenzoylethylene
(bright yellow)

cis-1,2-Dibenzoylethylene
(colorless)

This is an endothermic reaction that does not reach equilibrium, but goes to completion because the cis isomer's electronic transitions are shifted to the ultraviolet and do not absorb visible radiation. The cis alkene is a structurally shorter chromophore than its trans isomer, and it is also likely to experience steric crowding with resultant distortion of the π system. The cis isomer, therefore, absorbs at shorter wavelengths (higher energies) and has a lower molecular extinction coefficient (weaker) than the trans isomer. Thus, once formed, the cis isomer is trapped. Upon heating, however, the cis isomer undergoes exothermic isomerization back to the original, more stable, trans alkene under equilibrium conditions.

A good example of photochromic behavior is the highly colored *cis*- and *trans*-azobenzene. In this case the π → π* transition is promoted by ultraviolet light so that nonequilibrium isomerization to the cis isomer requires UV irradiation. The cis isomer is considerably less stable, however, and it undergoes relatively easy reversion back to the trans isomer by other mechanisms (the thermal conversion of visible radiation absorbed by the colored cis compound is an additional isomerization route apparently open to the cis isomer):

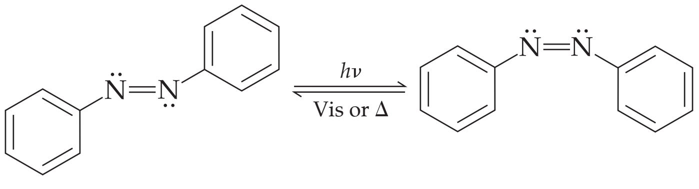

trans-Azobenzene *cis*-Azobenzene

Tautomerism. A number of proton and valence tautomers are subject to photo-chemical induction. One example is 2-methylbenzophenone (**I**), a colorless compound that can be photochemically induced to tautomerize to a system with extended conjugation. The tautomer (**II**) is a yellow material that will revert to the colorless form under thermal conditions:

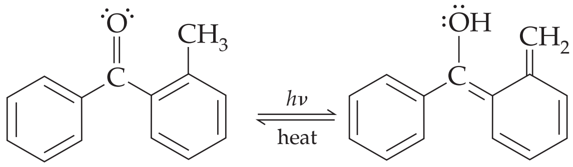

I
2-Methylbenzophenone
(colorless)

II
(enol, yellow)

Homolysis and Heterolysis of Bonds. Photochromic homolysis has been observed with materials such as 2,3,4,4-tetrachloro-1-(4*H*)-naphthalenone (**III**):

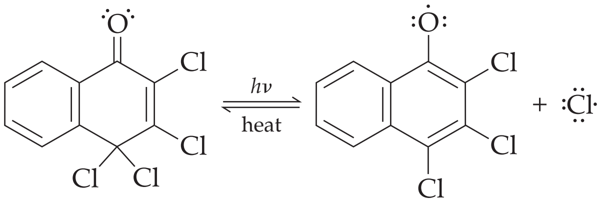

III
(colorless)

(violet)

Excitation leading to heterolysis and zwitterion formation has been observed in numerous spiropyrans, as shown here for **IV**. These compounds undergo ring opening to yield a zwitterion (**V**):

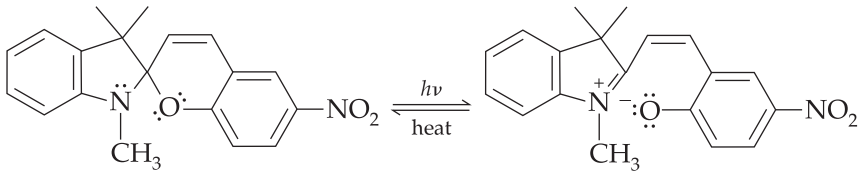

IV
(colorless)

V
(violet)

The spiropyran example in which the photochromism develops following the transformation of a colorless isomer to a violet system is closely related to the structural isomerism observed in the target photochromic substance, a diazabicyclo[3.1.0]hex-3-ene (**VI**) synthesized in Sequence F. In both instances, the photoisomerization involves heterocyclic ring opening with formation of zwitterion **VII**:

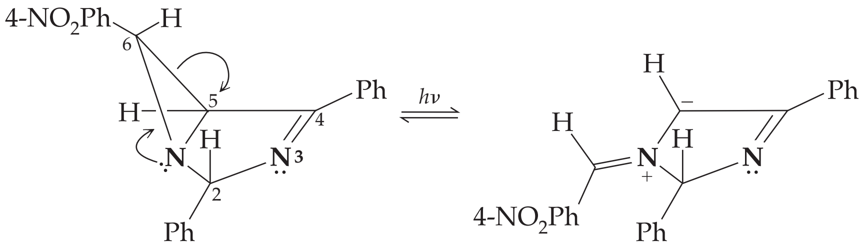

VI
2-*exo*-6-*exo*-2,4-Diphenyl-
6-(4-nitrophenyl)-
1,3-diazabicyclo[3.1.0]hex-3-ene

VII
Blue azomethine–ylide
(syn configuration)

The diazabicyclo[3.1.0]hexenes form a series of compounds of which many exhibit photochromic properties.

APPLICATIONS OF PHOTOCHROMISM

One successful application of inorganic photochromic systems has been in the manufacture of sunglasses. When a particular silver salt is incorporated in the lenses, the glass will darken on exposure to sunlight in order to protect the eyes, but then bleach quickly when the light intensity drops, so that the same glasses can be used at night or indoors.

The cis–trans isomerism of azobenzene has been incorporated in a novel chemical method for information storage at ultrahigh densities with nondestructive readout. A device based on this chemical information storage system would have, it is estimated, the capacity of 100 million bits per square centimeter.

There also have been some investigations by the military directed toward developing camouflage "photochromic paints."

In most applications the ability of the system to undergo essentially endless recycling is an important factor. Thus, the degradation response time impacts heavily on the effectiveness of the system. In this regard, the valence bond tautomeric isomerizations would appear to possess the most promising properties, while those mechanisms that involve fragmentation open the system up to the possibility of irreversible byproduct formation. The heterolytic processes described in the photochromism of the target compound (internal ring opening) would appear to fall in between the two boundaries.

THE REACTIONS OF SEQUENCE F[3]

The synthesis begins with an aldol reaction between 4-nitrobenzaldehyde and acetophenone in Experiment [F1] to yield 4-nitrochalcone:

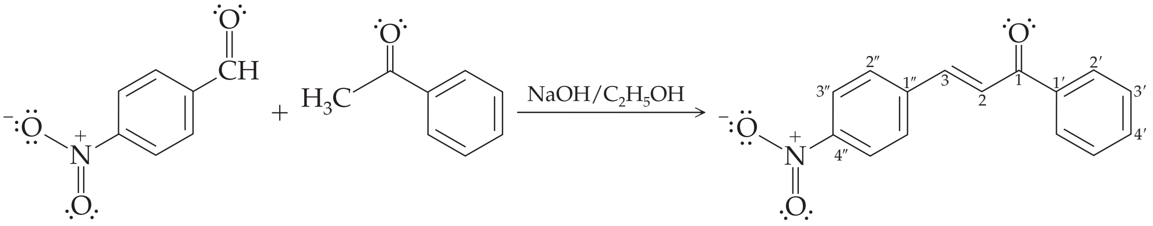

p-Nitrobenzaldehyde Acetophenone *trans*-4-Nitrochalcone

The chalcone is brominated in the second step (Experiment [F2]) to yield *erythro*-2,3-dibromo-3-(4-nitrophenyl)propiophenone:

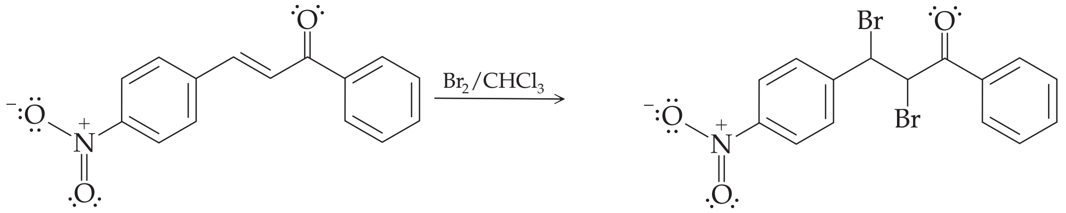

trans-4-Nitrochalcone *erythro*-2,3-Dibromo-3-(4-nitrophenyl)propiophenone

When this halogenated compound is treated with ethanolic ammonium hydroxide for several days, as in Experiment [F3], the system undergoes several reactions (dehydrohalogenation, amination, and ring closure) to ultimately yield a trans-substituted aziridine product:

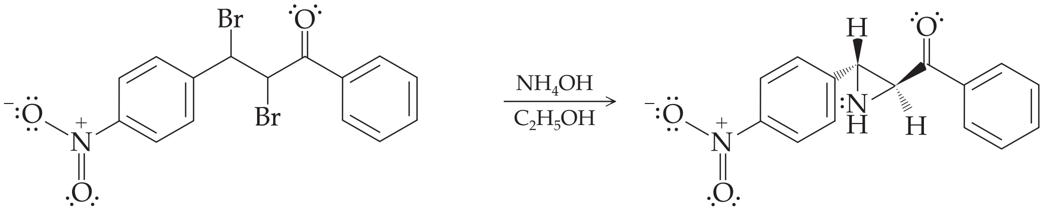

erythro-2,3-Dibromo-3-(4-nitrophenyl)propiophenone *trans*-2-(4-Nitrophenyl)-3-benzoylaziridine

The photochromic substance, a diazabicyclo[3.1.0]hex-3-ene, is obtained in the fourth step (Experiment [F4]) when the aziridine is treated with benzaldehyde, anhydrous ammonia, and ammonium bromide. The reaction requires several days to go to completion.

---

[3] This synthetic sequence is based on a set of experiments first developed for the undergraduate instructional laboratory by Professor R. Marshall Wilson and Laboratory Director D. L. Lieberman of the University of Cincinnati. We are grateful to Paulette Messier of Bowdoin College and Dr. Joanne M. Holland of Sepracor, Inc. for further development and optimization work.

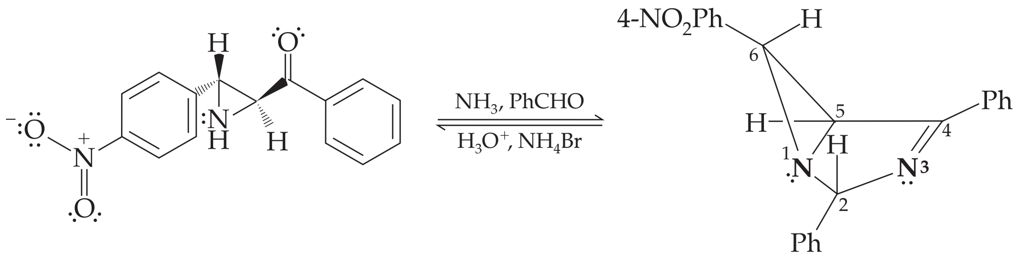

trans-2-(4-Nitrophenyl)-3-benzoylaziridine

2-*exo*-6-*exo*-2,4-Diphenyl-6-(4-nitrophenyl)-1,3-diazabicyclo[3.1.0]-hex-3-ene

The product is a colorless crystalline substance that possesses the property of turning a deep blue when exposed to indoor light:

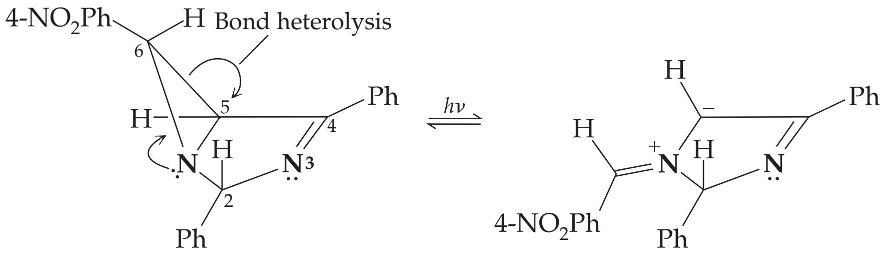

2-*exo*-6-*exo*-2,4-Diphenyl-6-(4-nitrophenyl)-1,3-diazabicyclo[3.1.0.]hex-3-ene

syn-Azomethine-ylide

The mechanism involved in the photochromic isomerization reaction is relatively rare, which makes these substances all the more interesting to study.

BIBLIOGRAPHY

Coyle, J. D.; Hill, R. R.; Roberts, D. R. *Light, Chemical Change and Life: A Source Book in Photochemistry;* The Open University: Milton Keynes, England, 1982, pp. 306–309.

de la Mare, P. D. H.; Suzuki, H. *J. Chem. Soc.* **1968,** 648.

Coxon, J. M.; Halton, B. *Organic Photochemistry;* 2nd ed., Cambridge University Press: New York, 1987.

Kopecky, J. *Organic Photochemistry;* VCH: New York, 1992.

Also see

Horspool, W. M.; Lenci, F., Eds., *CRC Handbook of Photochemistry and Photobiology,* 2nd ed., Vols. 1, 2, CRC Press: Boca Raton, FL, 2003.

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed., Wiley: New York, 2007, Chap. 7, p. 328 and references therein.

An Aldol Reaction: *trans*-4-Nitrochalcone

Experiment F1

Common name: 4-nitrochalcone
CA number: [2960-55-6]
CA name as indexed: 2-propen-1-one, 3-(4-nitrophenyl)-1-phenyl-, (*E*)-

Purpose. We prepare the first of three intermediates on the synthetic pathway to our target molecule, a photochromic imine. A base-catalyzed aldol reaction is carried out in which an aromatic aldehyde is condensed with an aryl alkyl ketone. This addition reaction is followed by dehydration to form an α,β-unsaturated ketone; this particular product is commonly called a chalcone. This intermediate is isolated and purified for use as the starting material in the next stage of the synthesis. You will carry out a semimicroscale reaction to gain experience at conducting larger-scale organic reactions.

Prior Reading

Technique 4: Solvent Extraction
Liquid–Liquid Extraction (p. 72)
Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)

REACTION

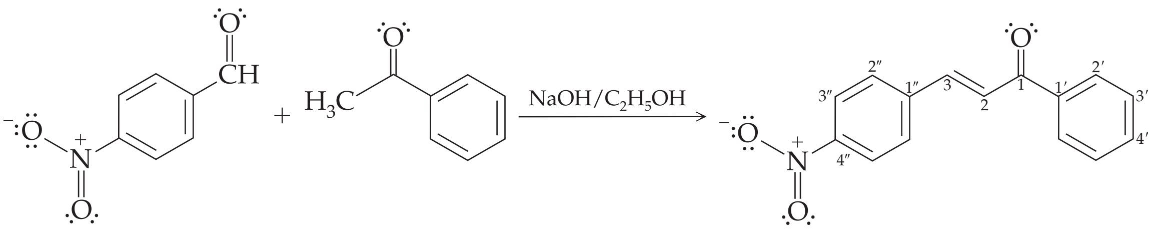

4-Nitrobenzaldehyde Acetophenone *trans*-4-Nitrochalcone

DISCUSSION

The aldol reaction (aldol condensation) is one of the fundamental reactions of organic chemistry because it leads to the formation of a new carbon–carbon bond (see Experiment [20] for a very similar example of the Claisen–Schmidt type of aldol reaction). In this version, the condensation of 4-nitrobenzaldehyde (an aldehyde without an α-hydrogen atom) with acetophenone (a ketone) gives *trans*-4-nitrochalcone. The aldol condensation of the unsubstituted aromatic aldehyde, benzaldehyde with acetophenone, yields *trans*-1,3-diphenyl-2-propenone (PhCH═CHCOPh), which has the common name, chalcone. Thus, the substituted derivatives of this system are known collectively as chalcones.

The extended conjugation in the product favors the formation of the chalcone product. Furthermore, this product is insoluble in the aqueous ethanol solvent and rapidly precipitates from the reaction medium as it is formed, whereas the starting materials are all soluble in aqueous ethanol. Thus, the experimental conditions assist in driving this equilibrium reaction to completion.

The classic aldol condensation involves generation of an *enolate* by removal of an acidic proton from a carbon alpha to the carbonyl group of an aldehyde or ketone, and subsequent nucleophilic addition of this enolate to the carbonyl carbon of an aldehyde or ketone. This reaction is base catalyzed and involves the following mechanistic steps:

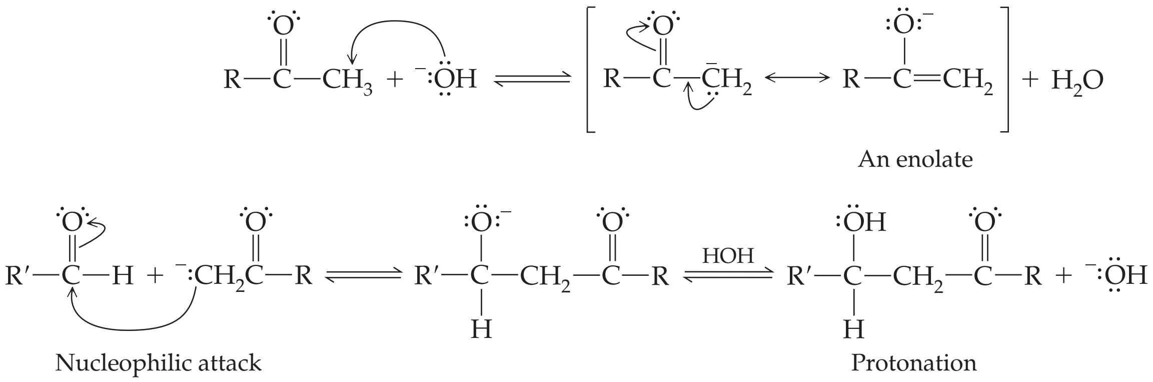

The reaction involves (a) base-catalyzed generation of the enolate, (b) nucleophilic attack of this anion on a carbonyl carbon, and (c) proton transfer to the resulting anion to yield the initial aldol product, a β-hydroxycarbonyl compound. The β-hydroxycarbonyl product may be isolated in many cases, if desired, since the subsequent dehydration is generally much slower than the addition reaction that precedes it. The final stage of the aldol reaction, as in the present reaction, is a hydroxide-catalyzed dehydration of the initial product by way of the enolate. Though hydroxide ion (HO^-) is generally not a good leaving group, the H alpha to the carbonyl in the β-hydroxyketone is quite acidic. In addition, the elimination produces a highly conjugated α,β-unsaturated ketone. Under these strongly basic conditions, the hydroxide ion becomes an adequate leaving group. In these systems, both during the loss of the proton in the formation of the enolate anion and during the loss of hydroxide to yield the α,β-unsaturated ketone, the molecular conformations involved favor development of the more-stable trans product:

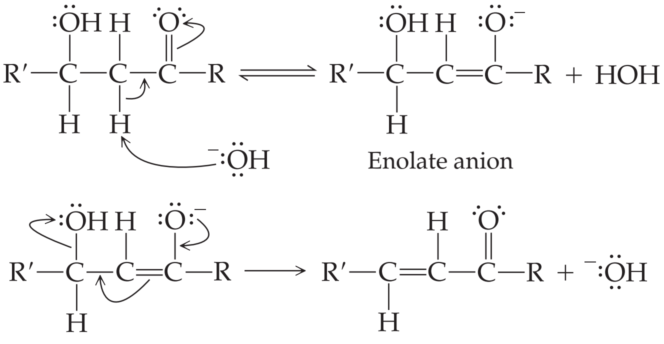

Trans α,β-unsaturated ketone

In the present experiment an aldol condensation yields a benzalacetophenone (chalcone) product. In Experiment [20], a nearly identical double aldol reaction yields dibenzalacetone. A further example of a double aldol reaction is found in Experiment [$A3_a$], where tetraphenylcyclopentadienone is the product of the reaction of benzil and 1,3-diphenylacetone.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 3.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
Acetophenone	120.16	488 μL	4.16	20.5	202.6	
Ethanol (95%)		20 mL			78.5	0.789
4-Nitrobenzaldehyde	151.12	500 mg	3.31	106		
Sodium hydroxide (aq, 10%)		2 mL				

Reagents and Equipment. Weigh and place 500 mg (3.31 mmol) of 4-nitrobenzaldehyde in a 50-mL Erlenmeyer flask containing a magnetic stir bar. Now add 20 mL of 95% ethanol and 488 μL of acetophenone.

Reaction Conditions. Warm the reaction mixture, while stirring in a sand bath at 65–70 °C, until the aldehyde dissolves to yield a clear light yellow solution.

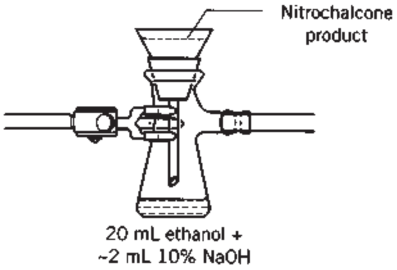

At this point, cool the Erlenmeyer flask in an ice bath for at least 5 min and then, while continuing to cool the system, *carefully* add dropwise with stirring, 2 mL of 10% (aq) sodium hydroxide over a second 5-min period. During this time, the reaction mixture often turns a dark orange color and considerable precipitation may occur. Cool the Erlenmeyer flask for an additional 15 min following the last addition of base.

Isolation of Product. Collect the solid tan precipitate, which formed during the reaction, on the filter bed of a Hirsch funnel under reduced pressure (⬅).

If solid continues to form in the filtrate, refilter the reaction solution and combine the second collection of crystals with the first batch. You want to maximize your total yield of the aldol product, because this is the first step of a four-step synthesis. Thus, you will need efficient recovery of product at each intermediate stage of the synthesis to successfully obtain a reasonable quantity of the photochromic target molecule. Rinse the Erlenmeyer flask once or twice with ice-cold water to effect as closely as possible a quantitative transfer of the chalcone to the Hirsch funnel.

Purification and Characterization. Wash the tan filter cake containing the reaction product, dropwise with an ice-cold 80:20 ethanol/water solution until the product appears as pale yellow crystals. The wash dissolves and removes a red-brown amorphous material that contaminates the crude product in many instances. Transfer the purified and partially dried chalcone to a watch glass, and then place it in a desiccator for final drying. Characterize the anhydrous intermediate accurately by an evacuated melting point and infrared spectrum; the latter can be compared to that recorded in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar). The chalcone normally is of sufficient purity to be carried directly on to Experiment [F2].

NOTE. Approximately 400 mg of purified product with a melting point within 2–4 °C of the literature value is the minimum quantity of intermediate suggested for continuing the sequence on to Experiment [F2].

QUESTIONS

7-88. In a number of cases it is possible to successfully isolate the β-hydroxyketone intermediate prior to the dehydration that forms the α,β-unsaturated ketone.
  (a) Does the isolation of the β-hydroxyketone suggest which step in the aldol condensation is the rate-determining step in this case?
  (b) If so, which one *is* the rate-determining step?
  (c) Suggest what reaction conditions may have a significant impact on determining which step becomes rate determining.
  (d) What structural changes might lead to a change in the rate-determining step?

7-89. Ketones also undergo the aldol condensation, although a successful reaction often requires "enhanced" conditions, since the addition involves an unfavorable equilibrium constant. This is the situation in the reaction in which 4-nitrochalcone is synthesized. With the odds against it, why is the reaction successful in this case?

7-90. If you had obtained both the *cis*- and *trans*-chalcone products, and had purified them by recrystallization, how could you instantly know which one was cis and which one was trans without any further characterization?

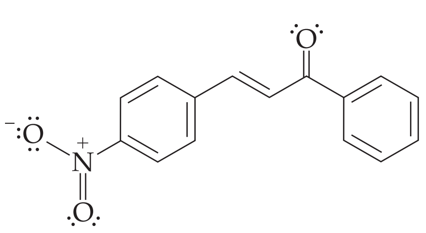

Trans

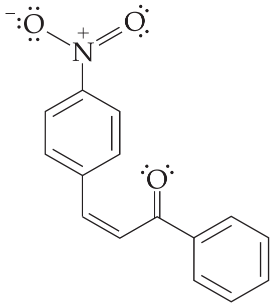

Cis

7-91. The mixed aldol reaction between propionaldehyde and acetone gives an 85% yield of 4-hydroxy-2-hexanone when run in THF at −78 °C with lithium diisopropylamide (LDA, a powerful base). The reaction is carried out by first adding the ketone to the base, cooling the solution, and then adding the aldehyde.

- **(a)** Why does this mixed system give essentially a single product?
- **(b)** Why is there no self-condensation of the acetone?
- **(c)** Why does the system not rapidly go on to dehydrate?
- **(d)** Why is the ketone added to the base rather than vice versa?

7-92. Give the aldol product or products from the following reaction:

$$CH_3CH_2CH_2CH_2CH_2CHO + CH_3CH_2CH_2CH_2CHO \xrightarrow{\text{base}}$$

7-93. The chalcone structure is particularly interesting. It was known in nature long before it was synthesized in the laboratory. This structure is incorporated biosynthetically into a large class of over 300 natural pigments called flavonoids. These substances heavily contribute to the spectacular New England autumn colors and many flower pigments. Flavonoids arise from chain extension of shikimic acid–derived cinnamic acids (see Experiment [10C] for a more detailed discussion of biological origin of these materials). A typical example of a flavone would be luteolin (5,7,3′,4′-tetrahydroxyflavone, **I**), the orange-yellow pigment of the snapdragon:

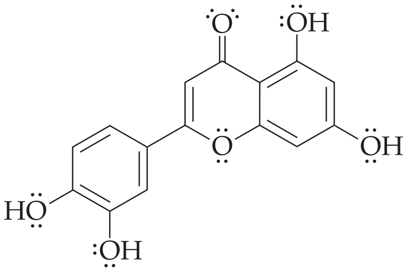

I
Luteolin

Flavonoids may be synthesized by using reactions similar to those used in the chalcone synthesis. For example, the basic flavone structure (**II**) can be simply derived from a Claisen condensation between ethyl benzoate and 2-methoxyacetophenone, followed by treatment with HI:

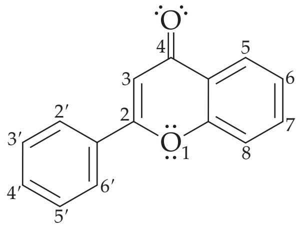

II
Flavone (skeleton)

Show the mechanistic route leading to this flavone from the ester and the ketone.

BIBLIOGRAPHY

Selected examples of the Claisen–Schmidt reaction from *Organic Syntheses* are given here:

Conrod, C. R.; Dolliver, M. A. *Organic Syntheses;* Wiley: New York, 1943, Collect. Vol. II, p. 413.

Kohler, E. P.; Chadwell, H. M. *Organic Syntheses;* Wiley, New York, 1941, Collect. Vol. I, p. 78.

Leuck, G. J.; Cejka, L. *Organic Syntheses;* Wiley: New York, 1941, Collect. Vol. I, p. 283.

Wawzonek, S.; Smolin, E. M. *Organic Syntheses;* Wiley: New York, 1955, Collect. Vol. III, p. 715.

Also see

Heathcock, C. H. In *Asymmetric Synthesis;* Morrison, J. D., Ed.; Academic Press: 1984, Vol. III.

Heathcock, C, H. In *Comprehensive Carbanion Chemistry;* Durst, T.; Buncel, E., Eds.; Elsevier: New York, 1984, Vol. II.

Nielson, A. T.; Houlihan, W. J. *Org. React.* **1969,** *16*, 1.

Smith, M. B.; March, J. *Advanced Organic Chemistry*, 6th ed., Wiley: New York, 2007, Chap. 16, p. 1344 and references therein.

Experiment F2

erythro-2,3-Dibromo-3-(4-nitrophenyl)propiophenone

Common name: *erythro*-2,3-dibromo-3-(4-nitrophenyl)propiophenone
CA number: [24213-17-0]
CA name as indexed: 1-propanone, 2,3-dibromo-3-(4-nitrophenyl)-1-phenyl-, (*R**,*S**)-

Purpose. You will prepare the appropriate dibromide, an intermediate in your synthetic sequence, to act as the precursor to the aziridine ring system. You will carry out a semimicroscale halogenation with bromine as the active reagent. A further purpose of this experiment is to demonstrate the stereospecific addition of bromine to alkenes.

Prior Reading

Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
  Concentration of Solutions (pp. 101–104)
Technique 5: Crystallization
  Use of the Hirsch Funnel (pp. 88–89)

REACTION

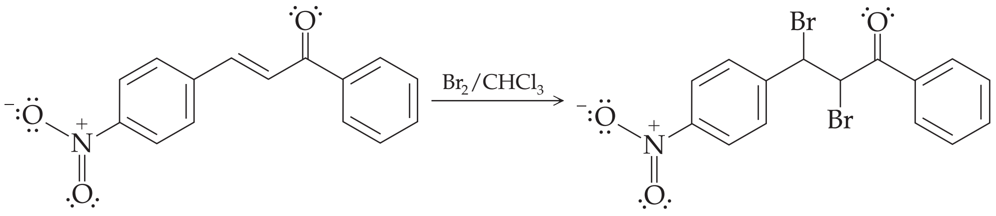

trans-4-Nitrochalcone

erythro-2,3-Dibromo-3-(4-nitrophenyl)propiophenone

DISCUSSION

The bromination of alkenes is an example of an electrophilic addition reaction. (See Experiments [$A2_b$] and [D2] for detailed discussions of the mechanism involved in this reaction. In particular, refer to Experiment [D2], which very closely resembles this reaction, for a discussion of the erythro and threo nomenclature used in this experiment.) In the present reaction, bromination of 4-nitrochalcone yields *erythro*-2,3-dibromo-3-(4-nitrophenyl)propiophenone.

This reaction is *stereospecific* because the other possible diastereomer [threo or (*R**, *R**)] is not formed:

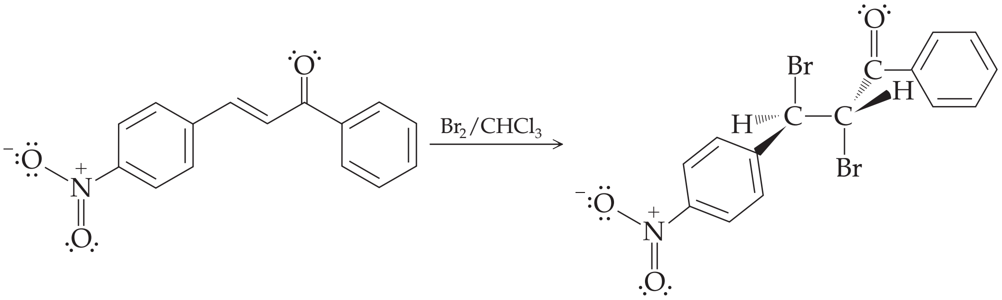

trans-4-Nitrochalcone

erythro-2,3-Dibromo-
3-(4-nitrophenyl)propiophenone

The reaction proceeds in two steps. The first involves the formation (from either side of the plane of the double bond; attack from below is shown here) of an intermediate cyclic *bromonium ion*:

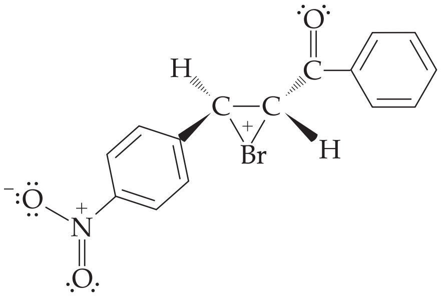

The bromine molecule (Br_2) is normally symmetrical. However, as it approaches the electron-rich, and nucleophilic, π bond of the alkene, it becomes polarized by induction and then functions as the electrophile in an addition reaction. The result is the generation of a cyclic bromonium ion:

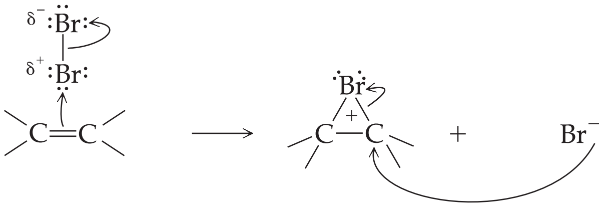

Induced polarization
of Br_2 as it approaches
the alkene

Bromonium ion

Nucleophilic
attack by bromide

In the present reaction, both the bromine and the 4-nitrochalcone are achiral, though the bromonium ion that is produced is chiral. Because it results from the reaction of achiral molecules in an achiral environment, the bromonium ion must be racemic. In this ion, the bromine atom bridges both carbon atoms of the original carbon–carbon double bond to form a three-membered ring intermediate. The generation of this high-energy cyclic species has a profound effect on the *stereochemistry* of the second step of the bromine addition: the ring restricts rotation about the C—C single bond in the carbocation.

The second stage of the bromination involves nucleophilic attack by bromide ion on the intermediate bromonium ion. Since the nucleophile must approach from the face opposite the leaving group, bond formation involves inversion of configuration at the carbon center under attack in the second stage of the bromination reaction.

Note that *either* carbon can be approached by the nucleophile (one attack is shown). This second step is a classic backside S_N2 type displacement:

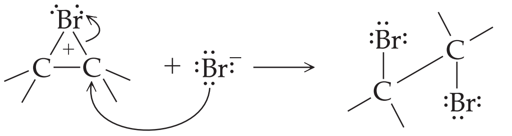

In the case with 4-nitrochalcone, two chiral centers are generated in the bromonium ion and we might, therefore, expect that two diastereomeric pairs would be formed. However, due to the stereoselectivity of the reaction, only a single diastereomer is generated, as a racemic pair of enantiomers (refer to Experiment [D2] for a further discussion of the stereochemistry of this halogenation).

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
trans-4-Nitrochalcone	253	400 mg	1.58	161–164		
Chloroform		~8 mL			61.7	1.492
Bromine/chloroform (2.5%)	179.8	(200 μL Br_2)	3.91		58.8	
erythro-2,3-Dibromo-3-(4-nitrophenyl)propiophenone	413			151–153		

HOOD

Reagents and Equipment. Weigh and place 400 mg (1.58 mmol) of 4-nitrochalcone in a tared 50-mL round-bottom flask containing a magnetic stir bar. Now add about 8 mL of chloroform (dispensed in the **hood**). Connect the flask to a water-jacketed reflux condenser fitted with a drying tube.

HOOD

INSTRUCTOR PREPARATION. The active brominating reagent in this reaction is liquid Br_2 dissolved in chloroform. Prepare a solution of 200 μL (624 mg) of bromine dissolved in 8 mL of chloroform multiplied by the number of students carrying out the experiment. The reagent should be prepared, dispensed, and added to the reaction in the ***hood.***

HOOD

CAUTION: Bromine is a highly reactive substance. Even in chloroform solution you must handle it with care. Be very careful not to get this reagent on your skin. All transfers of the reagent should be made in the ***hood.***

Chloroform itself is highly toxic and a cancer suspect agent. Handle it with respect.

Reaction Conditions. Warm the round-bottom flask in a sand bath between 60 and 70 °C, with stirring, until the 4-nitrochalcone dissolves to yield a clear light yellow solution. Once dissolution has occurred, continue to maintain the bath temperature at 60–70 °C and add the bromine reagent. The addition is carried out dropwise using a 9-in. Pasteur pipet inserted down the condenser (briefly remove the drying tube during this operation), with stirring, over a 10-min period. By the end of the Br_2 addition the solution turns a dark orange. Continue to heat the stirred reaction mixture for an additional 20 min (➡).

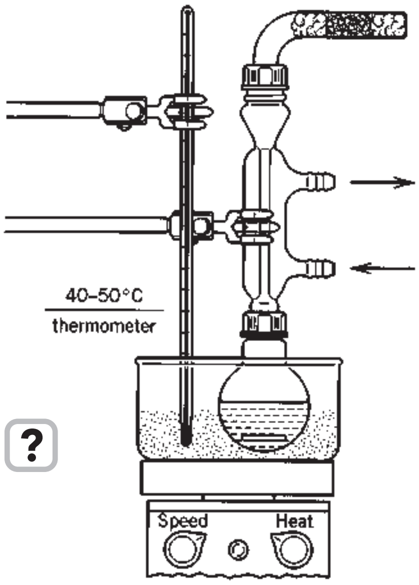

400 mg, 4-nitrochalcone +
16 mL, chloroform +
200 µL, bromine +
50-mL RB flask

After cooling the reaction mixture to room temperature, remove a small aliquot and spot it on a silica gel TLC plate next to a reference spot of the 4-nitrochalcone starting material. Elute the plate with 50:50 methylene chloride/hexane and visualize the spots by UV. (A number of unidentified byproducts are formed in small quantities during the reaction that are often observed on the TLC plates.) There also will probably be a trace of unreacted 4-nitrochalcone left in the reaction solution. The major (and highest R_f) spot ($R_f = 0.3$) is the brominated product. Because an excess of bromine was used in the reaction, however, only small quantities of 4-nitrochalcone are normally detected at this stage. If significant unreacted substrate remains (i.e., more than a faint or weak spot on TLC), add an additional small amount of the bromine reagent and then reanalyze the reaction mixture again by TLC.

Isolation of Product. Once it is established that the reaction is largely complete, remove the solvent and excess reagent from the reaction mixture by rotary evaporation. Weigh the crude residue.

NOTE. At this point the crystalline residue that remains following rotary evaporation may be stored until the next laboratory period by first flushing the round-bottom flask with dry nitrogen (or argon) and then quickly sealing the flask with a ground-glass standard-taper stopper that is sealed with Parafilm. The flask should be labeled and given to your instructor for storage in the freezer. While it is possible to safely interrupt the workup at this stage, most organic materials are much more stable when they are stored in as pure a state as possible. ***If time permits, you are urged to finish the workup of the bromination.***

Purification and Characterization. The crude material, a yellow-orange solid is now partially purified by column chromatography.

The chromatographic column (short buret) is packed with silica gel (10 g) after first positioning a plug of cotton and about 1 cm of sand at the bottom. Then a portion of 50:50 methylene chloride/hexane is added to the column (~15 mL), followed by 10 g of silica gel. The solid substrate is slowly added while the column is tapped to promote even settling of the packing material. During this process the column stopcock is slightly opened to create a slow drip rate of the packing solvent out of the column. As a result of this drainage more solvent may be required to keep the solvent level above that of the silica gel during the settling operation. Finally, carefully drain any excess solvent to the top of the column and close the stopcock after the packing procedure is complete.

The crude *erythro*-2,3-dibromo-3-(4-nitrophenyl)propiophenone is then dissolved in a minimum amount of methylene chloride (~10 mL) and applied to the top of the column by slowly pipetting the solution down the side of the column without disturbing the silica gel (as this solution is added it is also slowly drained onto the column by cracking open the stopcock). As the final quantity of crude product drains to the top of the column, elution is started with 50:50 methylene chloride/hexane solution (again by careful addition so as not to disturb the upper layers of silica gel containing the adsorbed reaction products). Collect 3 × 30-mL fractions in 50-mL Erlenmeyer flasks (labeling each flask with the fraction number). You may observe some yellow zones of material slowly moving down the column during the elution. This colored material usually does not begin to elute with this chromatographic scheme. Once the fractions have been collected, use TLC analysis (silica gel plates, 50:50 methylene chloride/hexane) to determine the composition of the fractions. The purest fractions are combined to give sufficient material to continue on to the next step (Experiment [F3]) in the sequence. Separate and remove the solvent (N_2 and warm sand bath). Determine the weight and melting point (evacuated) of your brominated product. The *erythro*-2,3-dibromo-3-(4-nitrophenyl)-propiophenone should appear as white to light yellow needles. It may be further recrystallized from 95% ethanol if desired.

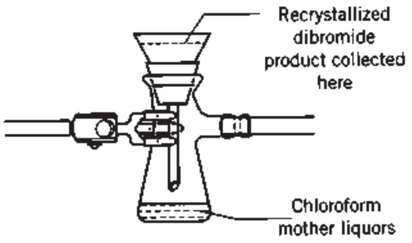
Recrystallized dibromide product collected here
Chloroform mother liquors

If the melting point is only a few degrees low (~142–147 °C), increased purity often is quickly obtained by simply adding a few milliliters of ice cold chloroform to the product residue that is cooled in an ice bath. Triturate the residue for a few seconds with the cold solvent. Withdraw the solvent by Pasteur filter pipet leaving the washed crystals behind (this treatment may be repeated if necessary). Remove traces of the solvent remaining on the residue by a short rotary evaporation. Recheck the melting point (evacuated) to determine if product purity is improved enough to continue on to the third step (⬅).

Weigh your purified *erythro*-2,3-dibromo-3-(4-nitrophenyl)propiophenone intermediate and calculate the percent yield based on both the starting 4-nitrochalcone and the 4-nitrobenzaldehyde (the starting material used in Experiment [F1]). Compare your spectrum to that of a reference standard shown in Figure 7.7.

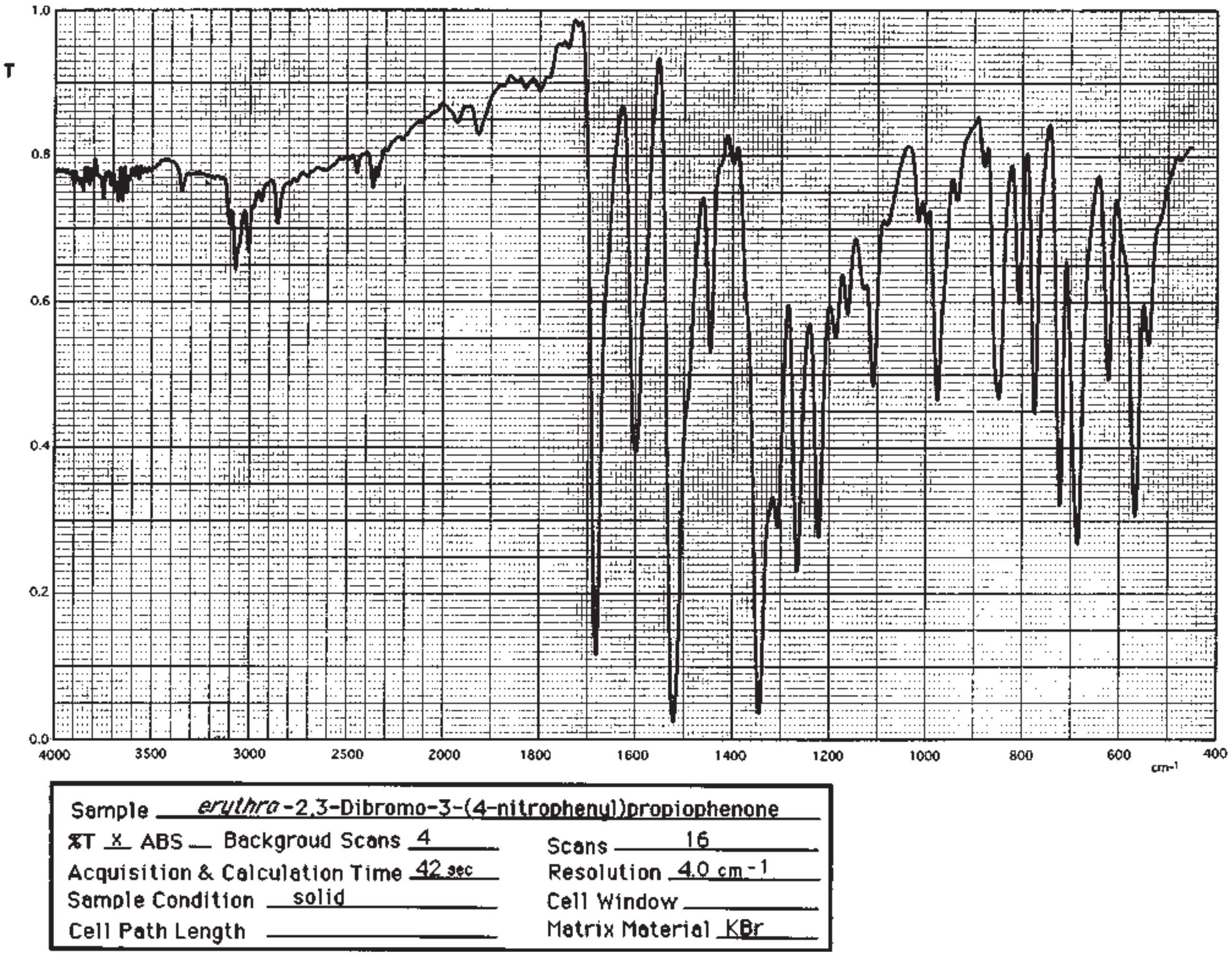

Sample *erythro*-2,3-Dibromo-3-(4-nitrophenyl)propiophenone

%T X ABS __ Backgroud Scans 4

Scans 16

Acquisition & Calculation Time 42 sec

Resolution 4.0 cm-1

Sample Condition solid

Cell Window

Cell Path Length

Matrix Material KBr

Figure 7.7 IR spectrum: *erythro*-2,3-dibromo-3-(4-nitrophenyl)propiophenone.

OPTIONAL. Obtain the ^{1}*H NMR spectrum of the product in* $CDCl_3$*. Identify the resonances for the two protons attached to the halogenated carbon atoms, and ascertain that the product is a single diastereomer. This information will be useful in Experiment [F3].*

NOTE. The erythro-2,3-dibromo-3-(4-nitrophenyl)propiophenone is a particularly sensitive substance and will decompose in contact with air at room temperature over several days. If you do not have time to continue on to Experiment [F3] during this laboratory period you should store your purified material as described above for the crude product.

If you have time to start the reaction, you are urged to continue on to the next step. *Once you have the pure dibromide in hand, it takes only a relatively short time to set up and get the next reaction running. Since this latter reaction will be left to run for a week, you have a lot to gain by getting the third step started during the period when the second intermediate is worked up.*

NOTE. Approximately 450 mg of purified product with a melting point within 2–4 °C of the literature value is the minimum quantity of intermediate suggested for continuing the sequence on to Experiment [F3].

QUESTIONS

7-94. A considerable excess of Br_2 in chloroform is required to successfully drive the halogenation of 4-nitrochalcone to completion. Offer a suggestion as to the role of the excess reagent.

7-95. What product(s) would you expect to obtain from the bromination (Br_2 in CCl_4) of cyclobutene?

7-96. In the bromination of 4-nitrochalcone a racemic dibromide is formed. A second diastereomeric dibromide is structurally possible, but it is not formed in the reaction.
- **(a)** Give stereochemically detailed drawings of the stereoisomer(s) isolated from the reaction mixture.
- **(b)** Give stereochemically detailed drawings of the stereoisomer(s) that is/are not formed.
- **(c)** Why is this reaction stereoselective?
- **(d)** How could the second diastereomer be synthesized if its preparation was required?
- **(e)** Assign, by the *R* and *S* convention, the stereocenters in each of the diastereomers.

7-97. Show the stereoisomer(s) generated by bromination of each of the enantiomers of *cis*- and *trans*-4-bromo-2-pentene. Show the relationship (enantiomer, diastereomer, etc.) and assign *R* and *S* centers in all product(s) and starting materials.

7-98. Show, with the correct absolute configuration, the stereoisomer(s) formed on bromination of *(S)*–4-*tert*-butyl-1-cyclohexene. Would you expect them to be formed in equal amounts?

BIBLIOGRAPHY

A large number of examples of the bromination of alkenes appear in *Organic Syntheses*. Selected references are given below:

Allen, C. F. H.; Abell, R. D.; Normington, J. B. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 205.

Cromwell, N. H.; Benson, R. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 105.

Fieser, L. F. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 195.

Khan, N. A. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 969.

McElvain, S. M.; Kundiger, D. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 123.

Paquette, L. A.; Barrett, J. H. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 467.

Rhinesmith, H. S. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 177.

Snyder, H. R.; Brooks, L. A. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 171.

Also see

Smith, M. B.; March, J. *Advanced Organic Chemistry*, 6th ed., Wiley: New York, 2007, Chap. 15, p. 999 and references therein.

Experiment F3

trans-2-(4-Nitrophenyl)-3-benzoylaziridine

Common names: *trans*-2-(4-nitrophenyl)-3-benzoylaziridine
CA number: [76336-95-3]
CA name as indexed: methanone, [3-(4-nitrophenyl)-2-aziridinyl]phenyl-, *trans*-

Purpose. The third intermediate on the pathway to the target photochromic imine is synthesized. A heterocyclic three-membered ring, an aziridine derivative, is formed. This is the first ring of the diazabicyclohexene system that you will ultimately convert into the photochromic imine. You will study a process that involves three reactions and the formation of two intermediates en route to the final product. You will study a number of interesting stereoselective reactions. You will work with organic reactions that require several days to come to completion.

Prior Reading

Technique 4: Solvent Extraction
Liquid–Liquid Extraction (p. 72)

REACTION

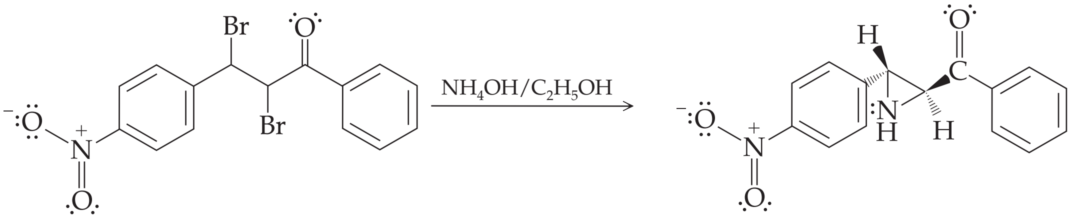

erythro-2,3-Dibromo-3-(4-nitrophenyl)propiophenone

trans-2-(4-Nitrophenyl)-3-benzoylaziridine

DISCUSSION

The conversion of *erythro*-2,3-dibromo-3-(4-nitrophenyl)propiophenone to a substituted aziridine involves a number of interesting steps and intermediates. The *first stage* of the reaction involves attack on the halogenated intermediate by base (concd ammonium hydroxide, $NH_4OH \rightleftharpoons NH_3 + H_2O$). Under the highly polar conditions, the reaction likely proceeds via an E1cb mechanism involving the initial attack by ammonia, acting as a base, on the **2** proton to yield the resonance stabilized anion:

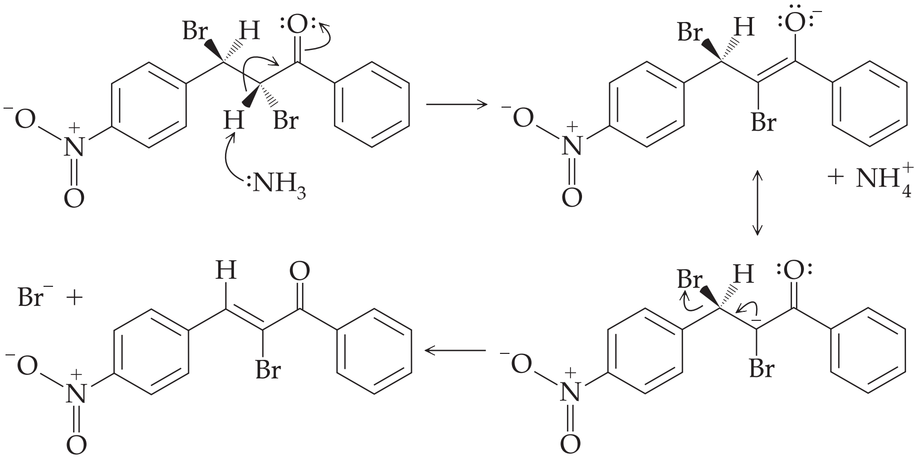

(*Z*)-2-Bromo-4′-nitrochalcone

You should refer to the detailed discussions in Experiments [9] and [10], which describe in detail the chemistry associated with the E1 and E2 elimination mechanisms. The E1cb mechanism, like the E1 mechanism, involves a two-step process, but in this case the order of charge development is the reverse of the E1 mechanism. Here proton abstraction precedes loss of the leaving group. A generalized scheme is shown here:

E1cb Mechanism

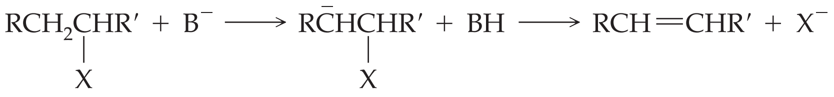

$$\mathrm{RCH_2\underset{\underset{X}{|}}{C}HR' + B^- \longrightarrow R\bar{C}H\underset{\underset{X}{|}}{C}HR' + BH \longrightarrow RCH{=}CHR' + X^-}$$

The mechanisms of a large majority of elimination reactions can be explained by invoking various positions along the continuum between the three elimination mechanisms mentioned here; E1cb and E1 are at the extremes of the continuum, and the E2 mechanism lies exactly halfway between the two.

In the present case, carbanion (or near-carbanion) formation following α-proton abstraction appears to be favored, and formation of the anionic intermediate has the attractive feature that it then allows rotation about the incipient π bond. Stereoelectronic requirements of the elimination mechanism require the carbon–bromine bond to be parallel to the p orbital of the adjacent enolate (or α-keto carbanion, depending on which resonance structure is being discussed), just as the proton being removed and the leaving group prefer to be anti in an E2 elimination. Two possible conformations (or rotamers) meet this requirement. The most stable one will be that with the carbonyl and nitrophenyl groups anti to one another, and elimination of bromide ion from this conformation leads to the alkene with the carbonyl and nitrophenyl groups trans, which is the *Z* alkene, as illustrated here:

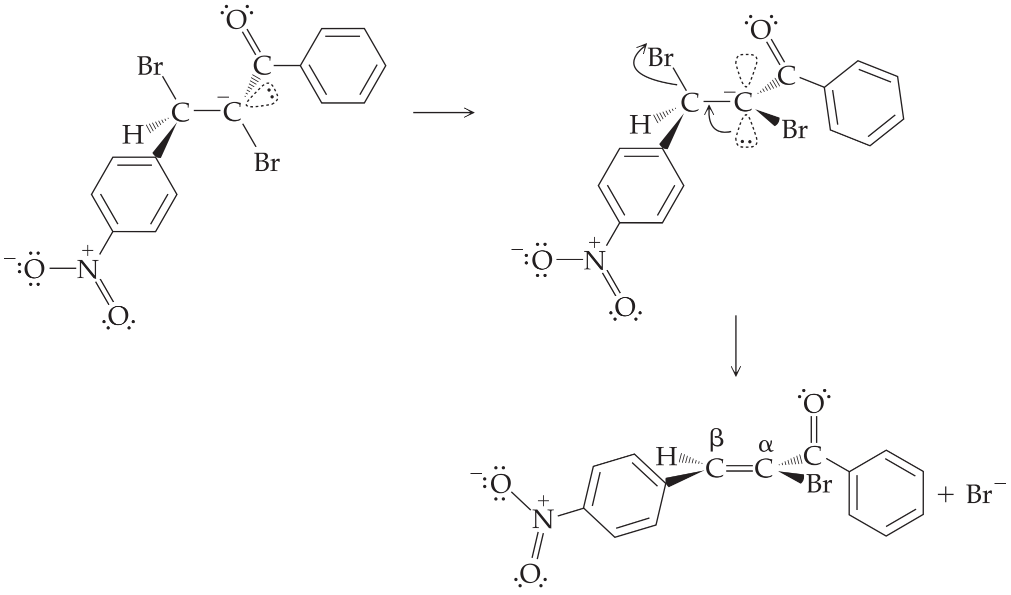

(*Z*)-2-Bromo-4′-nitrochalcone

Thus, the first stage of the reaction primarily leads to formation of an α-bromo-α,β-unsaturated ketone, (*Z*)-2-bromo-4′-nitrochalcone. This alkene could be easily isolated from the reaction medium, if required, since the second step in the reaction occurs at a considerably slower rate than the first step. Thus, appreciable concentrations of this unsaturated intermediate are obtained.

The *second stage* of the reaction involves 1,4 addition of the amine (in this case ammonia in equilibrium with the ammonium hydroxide) to the unsaturated ketone.

There are many examples of nucleophilic reagents that add to α,β-unsaturated aldehydes and ketones in a manner in which the addition is formally 1,4. This result is called conjugate addition. Under basic conditions, these transformations involve initial attack by the nucleophile to the β-carbon atom, followed by electrophilic addition (normally of a proton) on the carbonyl oxygen; the nucleophile and electrophile add at the 1 and 4 positions relative to one another. The enolate formed in the early stages of the reaction is generally quickly protonated to give an enol. The enol will subsequently tautomerize to the ketone. A general mechanistic scheme is shown below for the 1,4 addition of water to an α,β-unsaturated carbonyl system.

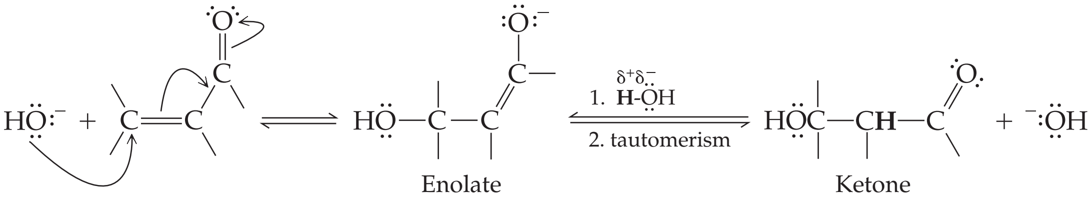

In the present case, conjugate addition has the potential to result in two diastereomers. As in the case of the starting dibromide, these stereoisomers can be either erythro or threo. The distribution between these two products is important because it ultimately determines the ratio of products in the next step (third) of the reaction.

From the geometry of the aziridine final product, and the fact that the internal nucleophilic substitution reaction will proceed with inversion of configuration, we can reliably postulate that conjugate addition must eventually result in the erythro diastereomer. In the conjugate addition of ammonia to 2-bromo-4′-nitrochalcone, the stereochemistry of the product is determined by which face of the resulting enolate receives the proton. The proton source may be the —NH_3^+ group resulting from the initial step of the conjugate addition reaction (as shown here), or it may be an ammonium ion or a molecule of ethanol or water hydrogen bonded to the amino group. A further consideration may be the geometry of the resulting enolate, since there is the potential to generate either the *Z* or the *E* enolate; the *E* enolate is illustrated here because it would appear to be the least hindered of the two diastereomeric enolates:

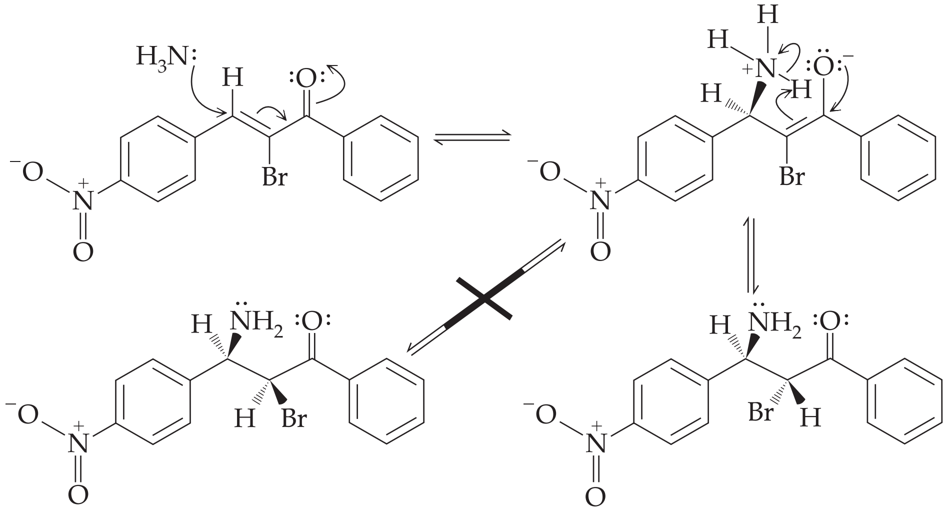

Threo

Erythro

It is also possible that both the threo and erythro diastereomers are produced at comparable rates and that the erythro is thermodynamically preferred by an equilibrium process between the two.

A comparison of the most-favored conformations, based on relative group sizes, for the erythro and threo diastereomers, indicates that the conformation **A** for the erythro isomer would be less sterically crowded than conformation **B** for the threo isomer, as shown here.

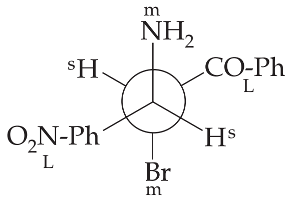

A
(*erythro*)

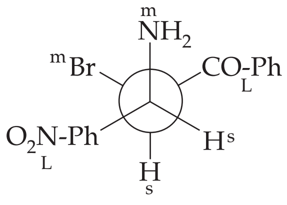

B
(*threo*)

L = large, m = medium, s = small
Ph = phenyl

The *final stage* on the route to the aziridine product involves an internal S_N2 ring closure in which the primary amine group attacks the α-carbon atom holding the remaining bromine from the backside to close an aziridine ring and displace bromide with inversion of configuration at the αcarbon. This displacement has been shown to go exclusively by this mechanism in the case of the β-amino-α-bromoketones; an S_N1 reaction is less likely because of the instability of an α-keto carbocation. For a detailed discussion of the mechanism of the classic S_N2 substitution reaction, refer to Experiment [22].

The stereochemistry of aziridine ring substitution, as pointed out above, is controlled by the product distribution in the β-amino-α-bromoketone intermediate, which in this case favors the erythro configuration. Thus, inversion of configuration during ring formation leads to the trans-substituted aziridine ring in the present example:

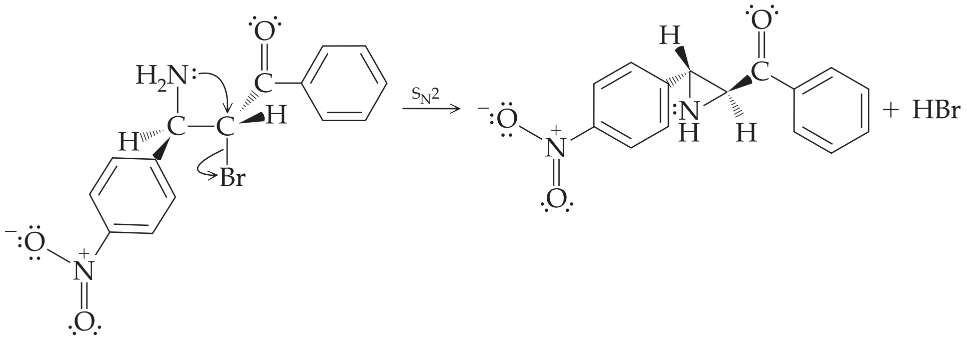

erythro-3-Amino-2-bromo-3-(4-nitrophenyl)propiophenone

trans-2-(4-Nitrophenyl)-3-benzoylaziridine

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 3.5 h.

Estimated time to complete Part A of the experiment: 1.5 h.

Estimated time to complete Part B of the experiment: 2.0 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
erythro-2,3-Dibromo-3-(4-nitrophenyl)propiophenone	411	450 mg	1.10	151–153		
Ethanol (95%)		4.4 mL			78.5	0.789
Ammonium hydroxide (concd)		1.3 mL				
trans-2-(4-Nitrophenyl)-3-benzoylaziridine	268.3			142–143		

PART A

HOOD

Reagents and Equipment. Weigh and place 450 mg (1.10 mmol) of the *erytho*-2,3-dibromo-3-(4-nitrophenyl)propiophenone intermediate synthesized in Experiment [F2] in a labeled 25-mL round-bottom flask containing a magnetic stir bar. In the **hood,** add 4.4 mL (9.8 mL/g of chalcone) of 95% ethanol (graduated cylinder) and 1.3 mL (2.8 mL/g of chalcone) of concd ammonium hydroxide (automatic delivery pipet or 2-mL glass pipet) to the flask. Stopper the flask, swirl to mix the contents, and then seal it with Parafilm. (A polypropylene standard taper stopper is preferred for sealing the vessel for long periods in the presence of base.)

HOOD

CAUTION: Concentrated ammonium hydroxide is a strongly caustic reagent. You must handle it with care. Be particularly alert not to get this reagent on your skin or breathe the vapors. All transfers of the reagent should be made in the ***hood.***

Reaction Conditions. Stir the reaction mixture for 15 min. Then wrap the flask with aluminum foil and continue to stir the system for 24 h. The reaction may be worked up at this point or stored in the dark for 1 week if necessary. Do not expect the chalcone to immediately dissolve in the reaction medium; it will do so over the course of several hours as the aziridine product begins to precipitate. If you wish, you can occasionally magnetically stir the reaction mixture for 15–20 min during this intervening period.

This completes Part A of the experiment.

PART B

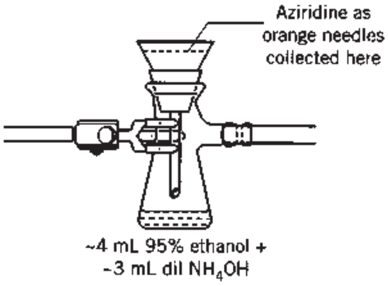
Aziridine as orange needles collected here
~4 mL 95% ethanol + ~3 mL dil NH_4OH

Isolation of Product. Add 15 mL of ice-cold water to the reaction mixture and swirl for 5 min. Collect the solid residue formed in the reaction by vacuum filtration on a Hirsch funnel. The product will appear as pale orange, very fine needles. There may also be a small amount of a more powdery yellow material that should dissolve in three 1.0-mL washes with ice-cold water. Use the first wash to rinse the Erlenmeyer flask and transfer the rinse to the Hirsch funnel. Fine orange needles of the aziridine product will be deposited on the filter bed (⬅).

Purification and Characterization. Collect the orange needles, dry them to constant weight (under reduced pressure), and determine an evacuated melting point. Recrystallize the crude product from hot methanol (~10–20 mL). After crystallization has begun, cool the system further in an ice bath for

10–15 min to complete the collection. The purified aziridine is obtained as long, shiny, pale orange needles via Hirsch filtration.

Weigh the *trans*-2-(4-nitrophenyl)-3-benzoylaziridine and calculate the percentage yield based on both the starting materials: *erythro*-2,3-dibromo-3-(4-nitrophenyl)propiophenone (Experiment [F3]), 4-nitrochalcone (Experiment [F2]), and 4-nitrobenzaldehyde (Experiment [F1]). Recheck the evacuated melting point, to see if further purification is required, and obtain an IR spectrum. Compare the spectrum to that of a reference standard shown in Figure 7.8.

OPTIONAL. Obtain the ^{1}H NMR spectrum of the aziridine. Establish from this spectrum and the NMR data obtained in Experiment [F2] if any unreacted dibromochalcone still contaminates the aziridine sample that has been purified for use in the preparation of the photochromic target molecule. Determine the diastereomeric purity of your product, and determine that the aziridine product is indeed trans substituted.

NOTE. Approximately 50 mg of purified product with a melting point within 2–4 °C of the literature value is the minimum quantity of intermediate suggested for continuing the sequence on to Experiment [F4].

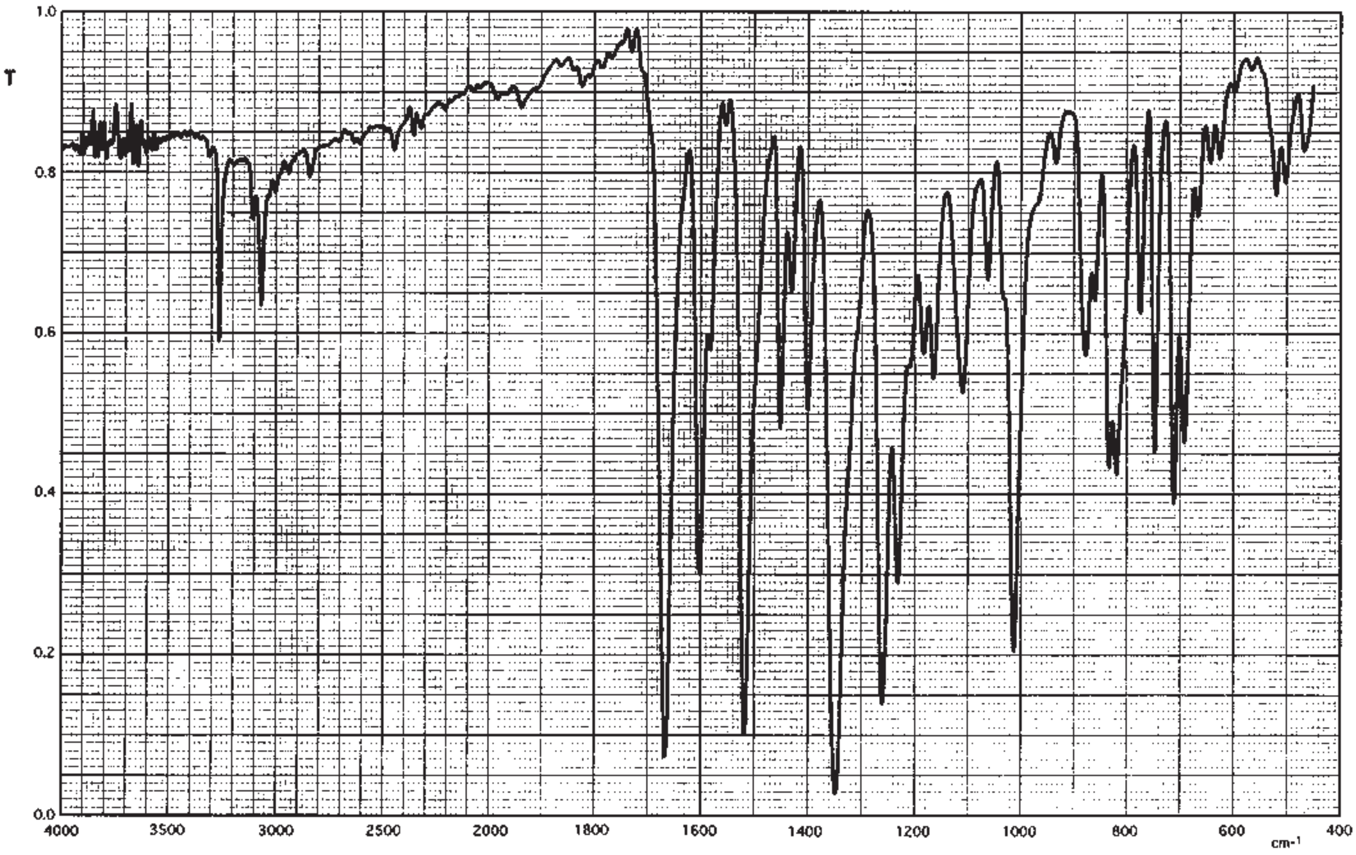

Sample *trans*-2-(4-Nitrophenyl)-3-benzoylaziridine
%T x ABS __ Backgroud Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm-1
Sample Condition solid
Cell Window ____
Cell Path Length ____
Matrix Material KBr

Figure 7.8 IR spectrum: *trans*-2-(4-nitrophenyl)-3-benzoylaziridine.

QUESTIONS

7-99. The synthesis of 1-cyclohexyl-2-phenyl-3-(4-phenylbenzoyl)aziridine from cyclohexylamine and 4′-phenylchalcone dibromide, gives a mixture of 47% of the cis isomer and 44% of the trans isomer. Chromatography on activated alumina yielded, in the first eluates, a crystalline material, mp = 118–119 °C while the final eluates produced a higher melting material, mp = 144–146 °C. The lower melting substance exhibited a $\tilde{\nu}_{C=O}$ = 1656 cm^{-1} while the higher melting compound had $\tilde{\nu}_{C=O}$ = 1686 cm^{-1}. Which product is the trans-substituted aziridine and which is the cis compound?

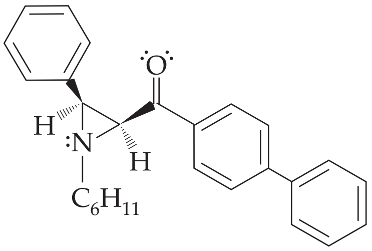

Cis

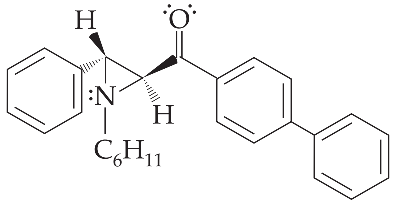

Trans

7-100. *N*-Haloaziridines have been found to possess an extremely high nitrogen inversion barrier. The isolation of the cis (**I**) and trans (**II**) isomers of 1-chloro-2-methylaziridine has been accomplished and they represent the first isolated inversion isomers of trivalent nitrogen. These inversion isomers appear to have remarkable stability. For example, 1-chloro-2,2-dimethylaziridine (**III**) retains configurational stability at temperatures as high as about 135 °C! Make a list of the potential stereoisomerism available to structures **I**, **II**, and **III**. Use perspective drawings to illustrate each isomer and label the stereocenters by the *R* and *S* convention:

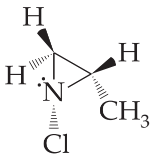

I

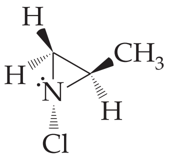

II

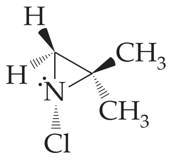

III

7-101. Another route to the effective synthesis of aziridines (**III**) is through the ring closure of the β-amino alcohol (**I**). The alcohol must first be converted to the β-amino hydrogen sulfate (**II**), which is the actual species that undergoes cyclization with strong base. Why is it necessary to convert the hydroxyl group into a hydrogen sulfate group prior to base treatment?

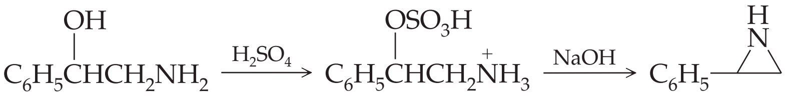

I **II** **III**

7-102. Aziridines are relatively reactive systems and undergo nucleophilic ring opening with the accompanying release of ring strain. Give a Fischer projection drawing of the expected product(s) of the reaction of *N*-ethyl-(2*S*,3*S*)-*trans*-2,3-dimethylaziridine (**I**) in aqueous ethyl amine:

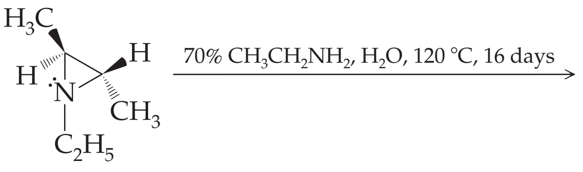

I

BIBLIOGRAPHY

Cromwell, N. H.; Cram, D. J. *J. Am. Chem. Soc.* **1943,** *65,* 301.
Cromwell, N. H.; Mercer, G. D. *J. Am. Chem. Soc.* **1957,** *79,* 3819.
Cromwell, N. H.; Barker, N. G.; Wankel, R. A.; Vanderhorst, P. J.; Olson, F. W.; Anglin, J. H. *J. Am. Chem. Soc.* **1951,** *73,* 1044.
Cromwell, N. H.; Hudson, G. V.; Wankel, R. A.; Vanderhorst, P. J. *J. Am. Chem. Soc.* **1953,** *75,* 5384.
Cromwell, N. H.; Cahoy, R. P.; Franklin, W. E.; Mercer, G. D. *J. Am. Chem. Soc.* **1957,** *79,* 922.
Do Minh, T.; Trozzolo, A. M. *J. Am. Chem. Soc.* **1972,** *94,* 4046.
Heine, H. W.; Hanzel, R. P. *J. Org. Chem.* **1969,** *34,* 171.
Heine, H. W.; Weese, R. H.; Cooper, R. A.; Durbetaki, A. J. *J. Org. Chem.* **1967,** *32,* 2708.
Heine, H. W.; Smith, III, A. B.; Bower, J. D. *J. Org. Chem.* **1968,** *33,* 1097.
Padwa, A.; Clough, S.; Glazer, E. *J. Am. Chem. Soc.* **1970,** *92,* 1778.
Smith, M. B.; March, J. *Advanced Organic Chemistry*, 6th ed., Wiley: New York, 2007, Chap. 10, p. 557 and references therein.

A Photochromic Imine: 2-*exo*-6-*exo*-2,4-Diphenyl-6-(4-nitrophenyl)-1,3-diazabicyclo[3.1.0]hex-3-ene

Experiment F4

Common name: *exo*-2,4-diphenyl-6-(*trans*-4-nitrophenyl)-1,3-diazabicyclo[3.1.0]hex-3-ene

CA number: [36799-57-2]

CA name as indexed: 1,3-diazabicyclo[3.1.0]hex-3-ene, 6-(4-nitrophenyl)-2,4-diphenyl-, (2α,5β,6β)-

Purpose. This experiment completes the synthesis of the photochromic imine, which is incorporated into the diazabicyclo[3.1.0]hex-3-ene skeleton. A rare molecular system is obtained in which an aziridine ring is fused to another heterocyclic ring. You will explore the exceedingly interesting photochromic properties of the target molecule. Microscale techniques are used during the conversion and isolation of this light-sensitive material.

Prior Reading

Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
Technique 5: Crystallization
  Craig Tube Crystallization (pp. 89–91)

REACTIONS

The Photosensitive Compound

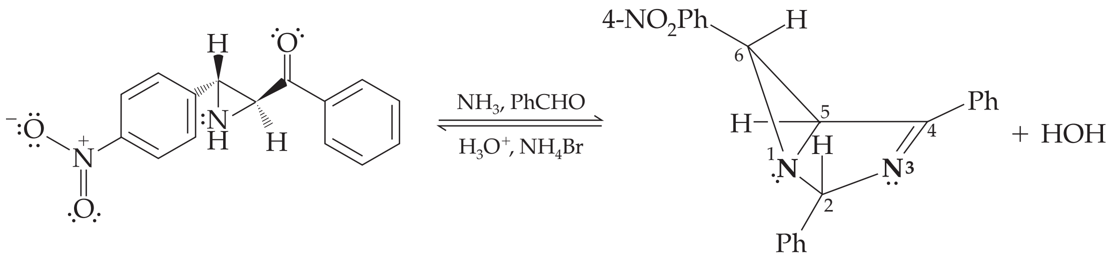

trans-2-(4-Nitrophenyl)-3-benzoylaziridine

2-*exo*-6-*exo*-2,4-Diphenyl-6-(4-nitrophenyl)-1,3-diazabicyclo[3.1.0]-hex-3-ene

Photoproduct

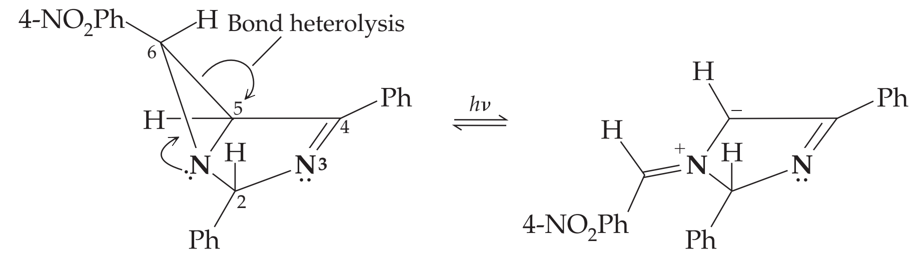
4-NO$_2$Ph
H
Bond heterolysis
6
5
4
Ph
H
N
H
N3
2
Ph
hν
H
Ph
H
N
H
N
4-NO$_2$Ph
Ph

2-*exo*-6-*exo*-2,4-Diphenyl-6-(4-nitrophenyl)-1,3-diazabicyclo[3.1.0]hex-3-ene

syn-Azomethine-ylide

DISCUSSION

This reaction completes the synthesis of the photochromic target molecule, which possesses the 1,3-diazabicyclo[3.1.0]hex-3-ene ring system. The substituted aziridine ring system formed in Experiment [F3] is condensed with benzaldehyde and ammonia to yield this bicyclic system. The reaction may be viewed as proceeding under anhydrous conditions via an initial reaction between the aromatic aldehyde and the base, aided by the ammonium bromide catalyst, to generate an imine as indicated in the following scheme:

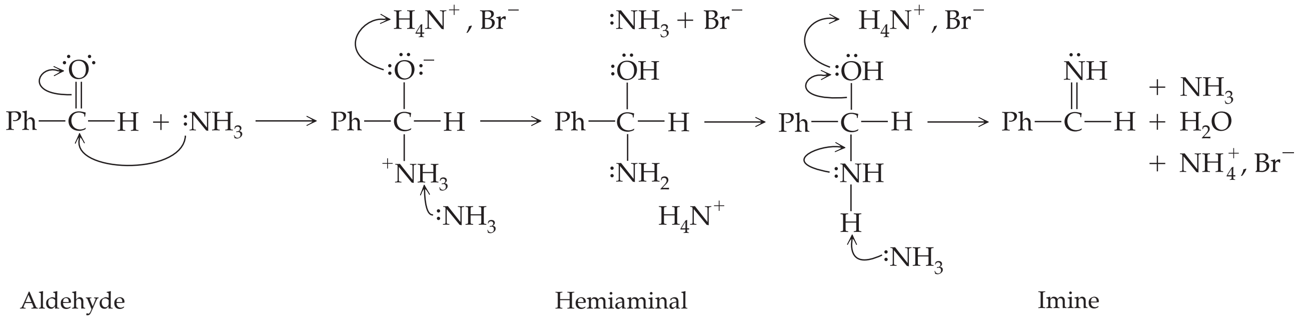
Ph—C—H + :NH$_3$
H$_4$N$^+$, Br$^-$
:NH$_3$ + Br$^-$
H$_4$N$^+$, Br$^-$
NH
+ NH$_3$
+ H$_2$O
+ NH$_4^+$, Br$^-$
:NH$_3$
H$_4$N$^+$
:NH$_3$
Aldehyde
Hemiaminal
Imine

The imine is subsequently attacked by the aziridine nucleophile to yield an aminal, again catalyzed by the ammonium bromide as shown in the following scheme:

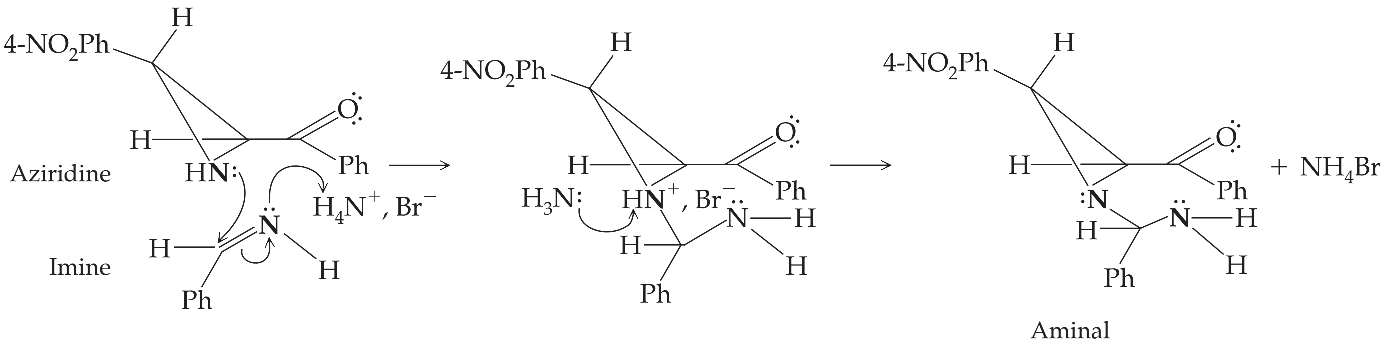
4-NO$_2$Ph
H
O
H
Aziridine
HN:
Ph
H$_4$N$^+$, Br$^-$
H
N
Imine
H
Ph
H$_3$N:
HN$^+$, Br$^-$
N—H
+ NH$_4$Br
:N
N—H
Ph
Aminal

Finally, the aminal undergoes an internal condensation involving ring closure by nucleophilic attack of the primary amine group on the carbon of the carbonyl followed by dehydration of the hemiaminal to yield the diazabicyclo[3.1.0]hex-3-ene ring system, again catalyzed by ammonium bromide. A reasonable scheme is shown below:

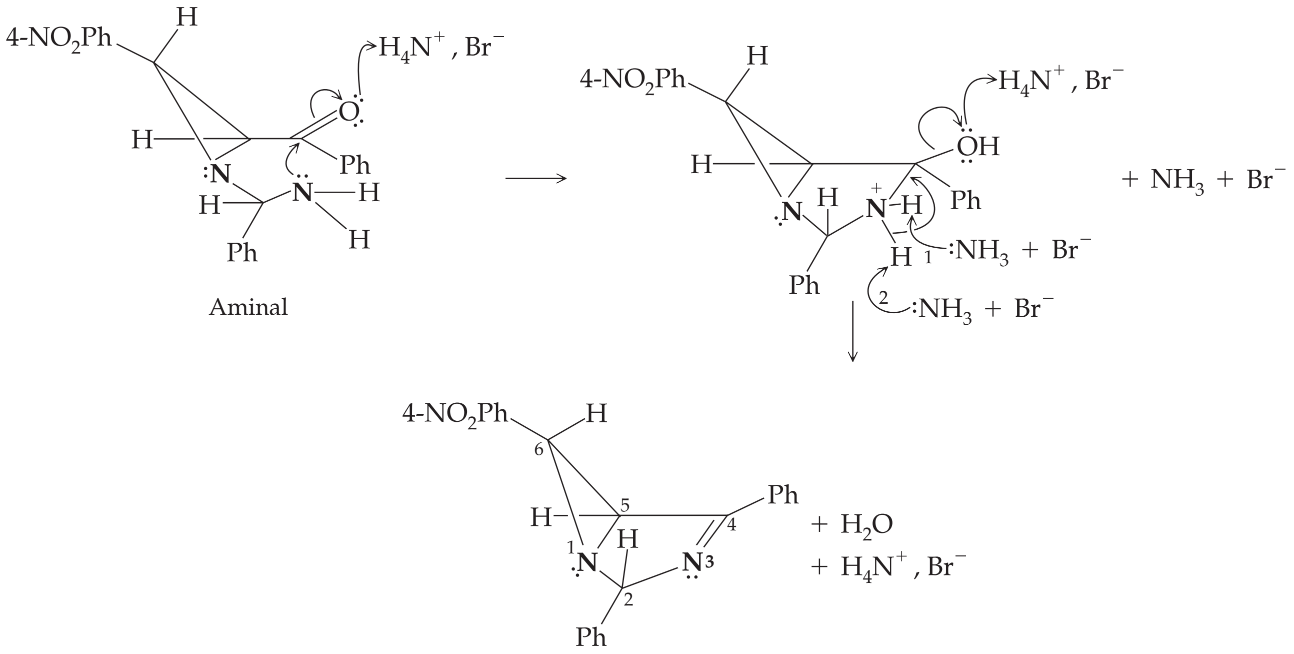
4-NO₂Ph
H
H₄N⁺, Br⁻
O
H
N
Ph
H
N
H
Ph
H
Aminal
4-NO₂Ph
H
H₄N⁺, Br⁻
OH
H
N
H
N-H
Ph
H
1
:NH₃ + Br⁻
2
:NH₃ + Br⁻
Ph
+ NH₃ + Br⁻
4-NO₂Ph
H
6
5
Ph
H
4
1
N
H
N3
2
Ph
+ H₂O
+ H₄N⁺, Br⁻

2-*exo*-6-*exo*-2,4-Diphenyl-6-(4-nitrophenyl)-1,3-diazabicyclo[3.1.0]-hex-3-ene

The photochromic compound appears to interconvert in the solid state to a bright blue azomethine–ylide (**I**) by an electrocyclic ring cleavage. An *ylide* is a neutral species whose Lewis structure contains opposite charges on adjacent atoms. The atoms involved are carbon, and an element from either group 5A or 6A of the periodic table, such as N, P, or S. Of considerable interest is the fact that the photochemical process has been shown to give exclusively the syn isomer.

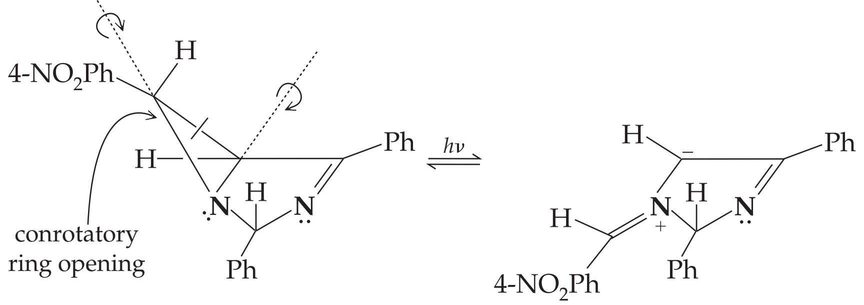
4-NO₂Ph
H
H
Ph
N
H
N
Ph
conrotatory
ring opening
hν
H
Ph
H
N
H
N
Ph
4-NO₂Ph

I
Blue azomethine-ylide
(syn configuration)

This result is consistent with a *conrotatory ring opening* involving a symmetry-allowed concerted transformation in the *ground state*. Since orbital symmetry requires a *disrotatory ring opening* from an *excited state* in these aziridines, the photochemically induced formation of the *syn*-azomethine–ylide (which is isoelectronic with the allyl anion) has been proposed to proceed via a dark reaction in which *electronically excited* states internally convert to *vibrationally excited* ground states. Indeed, evidence supporting this mechanism comes from the thermochromic behavior (a *ground-state* process) of the close relative 6-*exo*-2-dimethyl-4-diphenyl-6-(4-nitrophenyl)-1,3-diazabicyclo [3.1.0]-hex-3-ene (**II**), which when heated to 150 °C turns the same bright blue color (**III**).

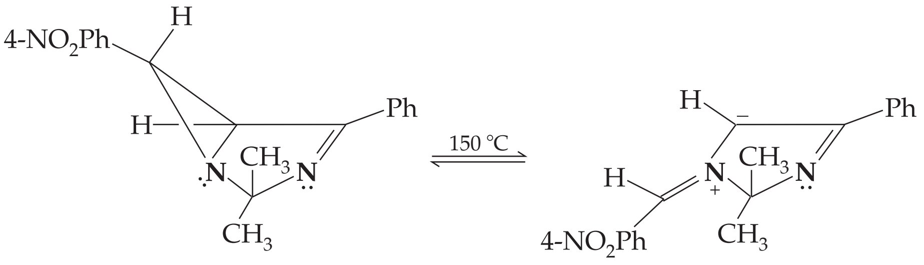

II

III
Blue azomethine-ylide
(syn configuration)

Exposure to light from tungsten lamps will not photochemically induce the ring-opening step. Most laboratories, however, are illuminated by fluorescent lights that emit small amounts of short wavelength radiation at the edge of the ultraviolet region (a wavelength long enough to not be absorbed by Pyrex glass), which will initiate the photochemical reaction. It is important to recognize that the azomethine-ylide photoproduct is a highly reactive species. Thus, if it is formed in solution where it can easily interact with other species, it rapidly decays to various byproducts, generally turning the solution yellow (not blue). It appears that if the ylide is produced in an environment that isolates it from other molecules, the system is stable and, given enough time, it will slowly revert to the diazabicyclic starting material.

The crystalline state is the ideal solution to this problem. Thus, when the nearly colorless, solid imine is irradiated, it turns bright blue. The neighbors to any ylide in the crystal can be either the ylide itself or the diazabicyclo starting material. If colored material is then placed in the dark it will slowly revert to the colorless form once again. These molecular systems have been successfully cycled between the colorless and colored states many hundreds of times with little degradation of the crystalline material. Obviously, it is exceedingly important that (a) you do not expose the imine to light when it is in solution during recrystallization (or at least that contact with light is kept to a minimum; working in a red-light darkroom would be ideal), and (b) recrystallizations should be carried out as quickly as possible, but obviously with care. Remember, you have a lot of time invested in this product, so work as quickly as is consistent with avoiding a costly spill.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*
trans-2-(4-Nitrophenyl)-3-benzoylaziridine	268.3	40 mg	0.15	142–143		
Ethanol, absolute		800 μL			78.5	0.789
Benzaldehyde	106.13	140 μL	1.37		178	1.045
Ammonium bromide	97.95	14 mg	0.14	452 sub		
Ammonia	17.03	Excess			−33	

Reagents and Equipment. Weigh and place 40 mg (0.15 mmol) of the *trans*-2-(4-nitrophenyl)-3-benzoylaziridine synthesized in Experiment [F3] in a 15-mL screwcapped centrifuge tube containing a magnetic spin vane. Dissolve the aziridine (20 μL of EtOH/mg of aziridine) and benzaldehyde (140 μL, 1.37 mmol) in 800 μL of absolute ethanol. To this solution add 14 mg (0.14 mmol) of ammonium bromide (0.35 mg of NH_4Br/mg of aziridine). Stir this mixture for 1 min and saturate the system with anhydrous ammonia (NH_3) in the **hood** (see your instructor for directions on this addition). HOOD

CAUTION: Anhydrous ammonia is a dangerous substance, particularly under pressure. The addition of this material to the reaction must be carried out in the ***hood*** under the *direct* supervision of the laboratory instructor. HOOD

Reaction Conditions. Gently bubble the ammonia gas through the reaction mixture until the system is saturated (until the reaction mixture cools off, ~5 min). Tightly cap the tube (Teflon liner), wrap it in aluminum foil, and stir the mixture for a minimum of 24 h. As in Experiment [F3], the reaction takes place at room temperature over a period of several hours. You can safely store the sealed reaction tube in your locker, protected from light with aluminum foil, for up to 1 week if necessary.

Isolation of Product. After the 24-h period, remove the supernatant liquid by centrifuging the tube and then transferring the liquid by means of a 9-in. Pasteur filter pipet to a Craig tube. Dry the remaining crystals in a stream of nitrogen gas. Remove a small sample of the solid material from the centrifuge tube on a glass rod or spatula and expose this material to direct sunlight or fluorescent light. If the solid material slowly turns blue, you have successfully synthesized the photochromic product. Concentration of the solution in the Craig tube may yield further quantities of the azomethine product. Weigh the tube and determine your percentage crude yield.

Purification and Characterization. Recrystallize the crude product in the dark or red light (best) or with the laboratory lights off from hot 95% ethanol. Dissolution may require as much as 10 mL of solvent and is best carried out in the centrifuge tube. After cooling the centrifuge tube in ice and scratching the sides with a glass rod to induce crystallization, centrifuge the system, and remove the mother liquors with a Pasteur filter pipet. Dry the white (or near white) crystals using a stream of nitrogen gas (crystals should be white if the recrystallization is done without fluorescent lights or sunlight).

NOTE. The photoproduct is an azomethine–ylide that is highly reactive. It is wise, therefore, to reduce the time that the material is in solution to a minimum, and to protect, while in solution, the contents of the capped centrifuge tube from light as much as possible. Once in the crystalline state, the large majority of the photochromic reactions take place in an environment in which the azomethine–ylide is protected from further reaction and thus, given the opportunity (in the dark), the ylide will slowly recycle back to the diazabicyclic precursor with little loss.

After removing the mother liquors and drying the purified photochromic product, determine an evacuated melting point. (The diazabicyclic intermediate substance is oxidatively sensitive and will decompose during atmospheric melting point measurements.) Be alert to color changes during the melting point determination.

Compare your results with the literature value of 169–172 °C. Obtain an infrared spectrum (if using a KBr disk, prepare in the dark or red light) of the white form of the product. Compare your spectral data to that of the reference standard shown in Figure 7.9. Then expose the disk to a bright fluorescent

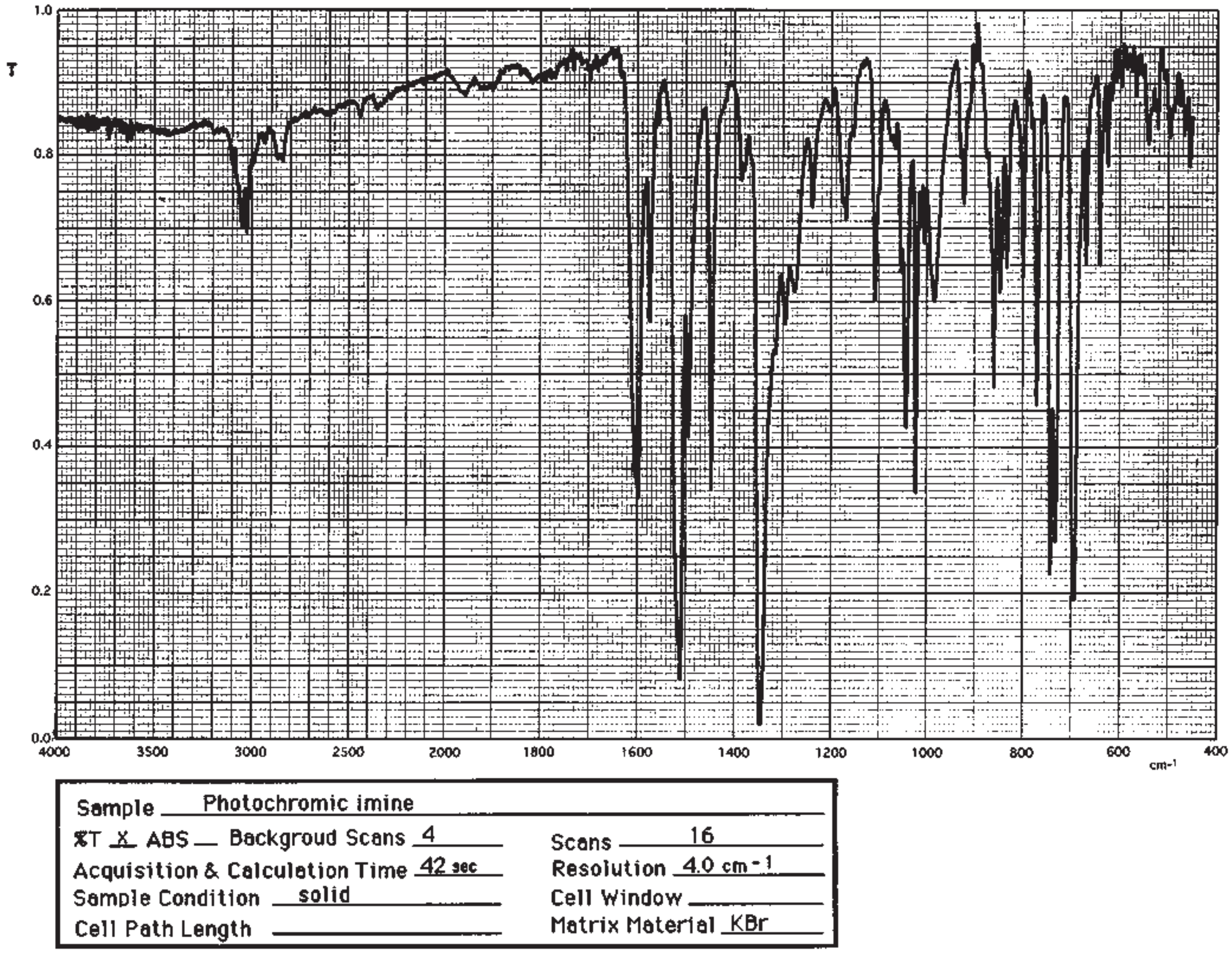

Sample Photochromic imine
%T _X_ ABS ___ Backgroud Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm-1
Sample Condition solid
Cell Window ______
Cell Path Length ______
Matrix Material KBr

Figure 7.9 IR spectrum: photochromic imine.

light for 1 min and redetermine the infrared spectrum. What does a comparison of the two experimentally derived spectra tell you about this photochemically induced reaction?

Store the remainder of your azomethine–ylide in a clean, sealed vial, flushed with N_2, and protected from light. By your next laboratory period this material should be colorless.

QUESTIONS

7-103. The heavily halogenated napthalenones were synthesized well over 100 years ago, but the structural details were sorted out less than 20 years ago. The compound, 2,3,4,4-tetrachloro-1-(4*H*)-naphthalenone (**I**), was discussed in the introduction to this sequence as an example of a substance that gains its photochromic activity by *bond homolysis.* In the synthesis of **I** two tetrachloro isomers were originally isolated. Compound **I** has photochromic activity while the second isomer (**II**) is simply a yellow-colored material. The structure of **II** has been determined to be 2,2,3,4-tetrachloro-1-(2*H*)-naphthalenone. A key piece of physical evidence that allowed the assignment of the structures was the infrared frequencies of the carbonyl groups. These frequencies were found to be 1701 and 1675 cm^{-1}. Which wavenumber value belongs with which structure?

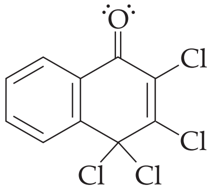

I

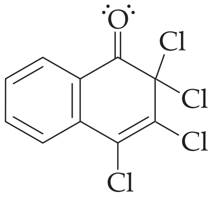

II

Explain your reasoning.

7-104. Explain why the neutral aziridine ring system is isoelectronic with the cyclopropyl carbanion.

7-105. As shown in Question 7-104, the aziridine ring system is isoelectronic with the cyclopropyl anion. Based on the theory of electrocyclic reactions these *even-number* π-electron-pair systems would be expected to undergo photochemical ring-cleavage in *disrotatory* fashion. Huisgen et al. have shown that in dimethyl 1-(4-methoxy-phenyl)aziridine-*trans*-2,3-dicarboxylate (**I**), upon photochemical excitation the ring opens to a 1,3-dipolar azomethine–ylide in a *disrotatory* cleavage that can be trapped by the addition of dimethyl acetylenedicarboxylate, which acts as a dipolarophile:

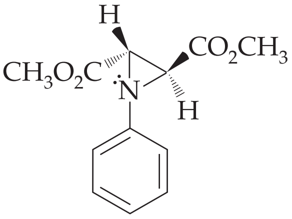

I

(a) Give a perspective drawing of the expected adduct.

(b) Give a perspective drawing of the expected adduct that would be formed from the blue azomethine ylide formed from our photochromic imine.

(c) Is the relative stereochemistry the same in the two adducts in *a* and *b*? Explain.

7-106. The dramatic change in color that occurs when the diazabicyclo[3.1.0]hex-3-ene ring system isomerizes on exposure to long wavelength UV radiation is related to what structural or electronic changes (or both) in the photochemically induced system?

7-107. Discuss the following:

(a) What role does ammonium bromide play in the conversion of the trans-substituted aziridine to the diazabicyclo[3.1.0]hex-3-ene derivative?

(b) Why is this particular bromide salt used?

BIBLIOGRAPHY

DoMinh, T.; Trozzolo, A. M. *J. Am. Chem. Soc.* **1972,** *94,* 4046.

Heine, H. W.; Weese, R. H.; Cooper, R. A.; Durbetaki, A. J. *J. Org. Chem.***1967,** *32,* 2708.

Hermann, H.; Huisgen, R.; Mäder, H. *J. Am. Chem.* **1971,** *93,* 1779.

Huisgen, R.; Scheer, S.; Huber, H. *J. Am. Chem. Soc.* **1967,** *89,* 1753.

Marckwald, W. Z. *Phys. Chem. (Leipzig)* **1899,** *30*, 140.

Padwa, A; Glazer, E. *J. Am. Chem. Soc.* **1972,** *94,* 7788.

Padwa, A.; Clough, S; Glazer, E. *J. Am. Chem. Soc.* **1970,** *92,* 1778. *ibid*., **1970**, *92*, 6997.

Turner, A. B.; Heine, H. W.; Irving, J.; Bush, **J. B.** *J. Am. Chem. Soc.* **1965**, *87*, 1050.

Ullman, E. F.; Henderson, W. A. *J. Am. Chem. Soc.* **1964**, *86,* 5050.

Woodward, R. B.; Hoffmann, R. *Angew. Chem. Int. Ed. Engl.* **1969**, *8*, 781.

Zincke, T.; and Kegel, O. *Berichte* **1888**, *21*, 1030.

For an overview of photochemistry see

Smith, M. B.; March, J. *Advanced Organic Chemistry*, 6th ed., Wiley: New York, 2007, Chap, 7, p. 329 and references therein.

Also see references cited in Experiment [6].

$\mathrm{C_8H_8}$apter 8

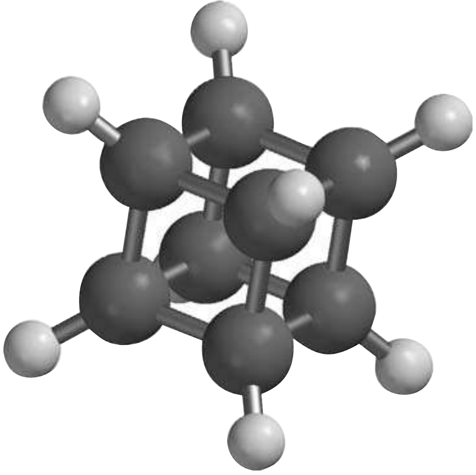

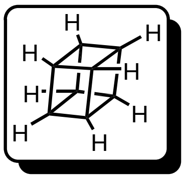

SPECTROSCOPIC IDENTIFICATION OF ORGANIC COMPOUNDS

INFRARED SPECTROSCOPY

The wavelike character of electromagnetic radiation can be expressed in terms of velocity *v,* frequency ν, and wavelength λ of sinusoidally oscillating electric and magnetic vectors traveling through space (Fig. 8.1). Frequency is defined as the number of waves passing a reference point per unit time, usually expressed as cycles per second (s^{-1}) or hertz (Hz). The velocity of the wave, therefore, equals the product of frequency and wavelength:

$$v = \nu\lambda$$

If the wavelength (the distance between the wave maxima or alternate nodes) is measured in centimeters, *v* is expressed in centimeters per second (cm/s). For radiation traveling in a vacuum, *v* becomes a constant, *c* ($c \sim 3 \times 10^{10}$ cm/s), for all wavelengths. When electromagnetic radiation traverses other media, however, the velocity changes. The ratio of the speed in a vacuum, *c,* to the matrix velocity, *v,* is termed the *refractive index, n,* of the material:

$$n = \frac{c}{v}$$

Since *n* is frequency dependent, the frequency at which the refractive index is measured must be specified. Frequency, however, has been shown to be independent of the medium and, therefore, remains constant. Wavelength thus varies inversely with *n.*

$$\lambda = \frac{c}{nv}$$

Since the velocity of electromagnetic radiation in a vacuum is normally greater than that in any other medium, *n* will generally be greater than 1 at all frequencies. Thus, the wavelength must become shorter for a particular frequency when measured in any matrix.

Frequency can be considered to be a more fundamental property of radiation because it is independent of the medium. This property also requires that the energy *E* associated with the radiation be matrix independent because *E* is directly proportional to frequency by

$$E = h\nu$$

Chapter 8: $\mathrm{C_8H_8}$, Cubane
Eaton and Cole (1964).

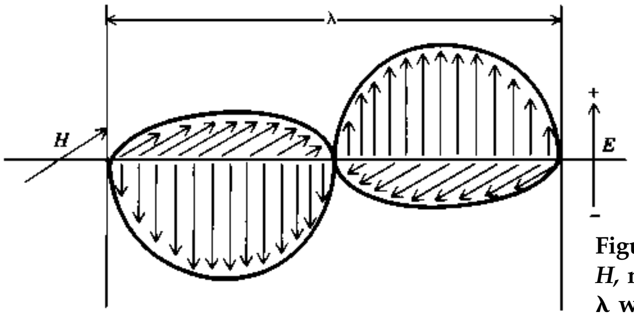

Figure 8.1 Electromagnetic wave. *H,* magnetic field; *E,* electric field; λ wavelength.

where E equals the energy of a photon, which is related to frequency ν by Planck's constant (h) (6.6×10^{-27} erg s or 6.6×10^{-34} J s).

The vibrational states present in molecules can be excited by absorption of photons. The nuclear masses and bond force constants determine the separation of these states and, therefore, the energies of the photons involved in the absorption process. The corresponding radiation frequencies fall predominantly in the IR region (10^{14}–10^{12} Hz) of the electromagnetic spectrum.

The IR spectrum is currently measured in *wavenumbers,* $\tilde{\nu}$, which are units proportional to frequency and energy. The wavenumber is defined as

$$\tilde{\nu} = \frac{\nu}{c} = \frac{E}{hc}$$

and as

$$\nu = \frac{c}{n\lambda} \qquad \text{then in air} \qquad \tilde{\nu} = \frac{\sim 1}{\lambda}$$

The wavenumber, as expressed in units of reciprocal centimeters (cm^{-1}) (the number of waves per centimeter), offers several advantages:

1. Wavenumbers are directly proportional to frequency and are expressed in much more convenient numbers (in this region of the spectrum), 5000–500 cm^{-1}.

2. As shown above, wavenumbers are easily converted to wavelength values. The reciprocal of $\tilde{\nu}$ and conversion of centimeters to wavelength units are all that is required (this is particularly handy because much early IR data were recorded linearly in wavelength). The wavelength unit employed in most of these spectra was the micron, μ (1×10^{-4} cm). The micron has been replaced by a unit expressed in meters, the micrometer, μm ($1\ \mu\text{m} = 1 \times 10^{-6}$ m).

3. Because the wavenumber is directly proportional to frequency and energy, the use of wavenumbers allows spectra to be displayed linearly in energy, which is a distinct aid in sorting out related vibrational transitions. For an introductory discussion of vibational energy see Chapter 8W, IR section, Part I.

www →

INTRODUCTION TO GROUP FREQUENCIES: INTERPRETATION OF INFRARED SPECTRA

Studies of the vibrational spectra of thousands of molecules have revealed that many of the normal modes associated with particular atomic arrangements may be transferred from one molecule to another. A *normal mode* of vibration is one of the residue *fundamental* vibrations of a molecular system in which the atomic displacements are all related by simple harmonic motion to the overall total vibrational motion (or vibrational energy) of the molecule. There are $3N - 6$ (where N = the number of atoms) *normal modes* (or fundamental vibrations or vibrational degrees of freedom—all these terms are essentially synonymous) present in all nonlinear molecules. Linear molecules have only $3N - 5$ normal modes—in this case there is one more normal mode of vibrational energy present because a rotational degree of freedom has been lost. Rotation around the molecular axis involves no energy because the atomic nuclei are assumed to be point sources of matter. (For an introductory discussion of vibrational energy see Chapter 8W, IR section, Part I B.)

www

Operating under selection rules these normal modes of vibration give rise to absorption bands in the infrared region of the spectrum (see, for example, the infrared spectrum of *n*-hexane, Fig. 8.2, p. 544). In the analysis these modes are often assigned numbers. For example, the 30 modes of benzene (where $N = 12$ in the $3N - 6$ expression) can be assigned 1 through 30 or the numbering can be done using any one of a number of different criteria. Subscripts *a* and *b* are often used to indicate doubly degenerate modes, that is, modes that have identical energies (and thus required to have the same frequency. One of the numbering systems for the benzene ring is used here when the aromatic ring stretching vibrations are identified (see Table 8.6 and Chapter 8W, IR section, Part II D, for a more detailed discussion of normal modes).

www

Many of these vibrational frequencies are associated with small groups of atoms that are essentially uncoupled from the rest of the molecule. The absorption bands that result from these modes, therefore, are characteristic of the small group of atoms regardless of the composition of other parts of the molecule. These vibrations are known as the *group frequencies.* Interpretation of infrared spectra of complex molecules based on group frequency assignments is an extremely powerful aid in the elucidation of molecular structure.

The following four factors make significant contributions to the development of a good group frequency from a molecular vibration:

1. The group has a large dipole-moment change during vibrational displacement. This change in moment is formally related to the efficiency of absorption of radiation during the molecular displacement by the expression $I\alpha(\delta\mu/\delta Q)^2$ (where I = intensity, μ = electric dipole moment, and Q = the normal coordinate [a mathematical description of the vibration]). Thus, if $\delta\mu/\delta Q$ is large, there is a large absorption of infrared radiation which gives rise to very intense bands (the intensity is dependent on the square of the moment change at that vibrational frequency).

2. The presence of a large force constant, so that for many of these groups the stretching frequency occurs at high values above the fingerprint region.

3. The fundamental mode occurs in a frequency range that is reasonably narrow (little coupling), but sensitive enough to the local environment to allow for considerable interpretation of the surrounding structure.

4. The range of frequencies is determined by a number of factors that are now well understood in terms of the mass, geometric, electronic, intramolecular, and intermolecular effects (for an introductory discussion of these effects see Chapter 8W, IR section, Part III A).

www

Strategies for Interpreting Infrared Spectra

1. Divide the spectrum into two parts at 1350 cm^{-1}.

2. Above 1350 cm^{-1}, absorption bands have a high probability of being good group frequencies. The interpretation is usually reliable and free from ambiguities. We can be much more confident of our assignments in this region even with rather weak bands.

3. Because of the reliability of the high-wavenumber region, we always begin the interpretation of a spectrum at this end.

4. Bands below 1350 cm^{-1} may be either group frequencies or fingerprint frequencies.

5. Below 1350 cm^{-1}, group frequencies are less easily assigned. In addition, even if a reliable group frequency occurs in this region, absorption at that frequency is not necessarily a result of that mode. That is, fingerprint bands can also randomly occur in the same location as reliable group frequency bands and the observer cannot usually distinguish which type of band is present.

6. To make more confident assignments below 1350 cm^{-1}, it is helpful to be able to associate a secondary property, such as band shape, with the particular mode. For example, it helps to know whether the band is very intense, broad, sharp, occurs as a characteristic doublet, gives the correct frequency shift on isotopic substitution, or the like.

7. A good rule to remember is that in the fingerprint region the **absence** of a band is more important than the presence of a band. If a band is absent, you can conclude with confidence that a reliable group frequency assigned to this region is absent and therefore the group must be absent from the sample. At the same time you also know that no interfering fingerprint bands occur in the region.

8. Before beginning the interpretation, note the sampling conditions and determine as much other information about the sample as possible—such as molecular weight, melting point, boiling point, color, odor, elemental analysis, solubility, and refractive index.

www

9. In the interpretation try to assign the most intense bands first. These bands very often will be associated with a polar functional group.

10. Do not try to assign all the bands in the spectrum. Fingerprint bands are unique to a particular system. Occasionally, intense bands will be fingerprint-type absorptions; these bands, generally, will be ignored in the interpretation. Fingerprint bands do, however, play an important role when infrared data are employed for identification purposes.

11. The correlation chart (back endpaper) can act as a helpful quick aid for checking potential assignments. It is not a substitute for understanding the theory and operation of group frequency logic. *The use of the correlation chart without a good knowledge of group frequencies is the shortest path to disaster!*[1]

12. Try to utilize the so-called *macro group frequency* approach. That is, if the functionality or molecular structural group requires the presence of more than a single group frequency vibrational mode, make sure that all modes are correctly represented. The *macro frequency train* represents a very powerful approach to the interpretation of relatively complex spectral data. This technique is at the core of current work on the automatic computer interpretation of infrared spectra. Contained in the product characterization section of the experiments given in Chapters 6 and 10W (online) are 12 detailed discussions (in Chapter 6, Experiments [5A], [5B], [6], [7], [8], [11], [14], [19], and [20]; in Chapter 10W (online), Experiments [1_{adv}], [4_{adv}] and [6_{adv}]) that demonstrate the operational use of the *macros.* Careful reference to these discussions will be very helpful in the initial stages of learning these interpretation techniques. It should become relatively easy to extend this interpretive approach to other reactions by reference to common infrared library files. Practice using *macro group frequencies* will pay big dividends in the laboratory. This last suggestion is perhaps the most important strategy to master in learning to interpret infrared spectra.

---

[1] Bellamy, L. J. *The Infrared Spectra of Complex Molecules,* 3rd ed.; Chapman & Hall: London, 1975, p. 3.

A SURVEY OF GROUP FREQUENCIES IDENTIFIED IN ORGANIC MOLECULES

The useful group frequencies are listed in the following sets of tables.

NOTE: A detailed description of the associated fundamental vibrational modes, diagrams of the actual displacements of the atoms, along with associated spectra, may be found at Chapter 8W, IR section, Part II. It is highly advisable to study this material.

In the following tables the vibration motion of the localized sections of the molecules assigned to a particular group's frequencies is often described using the following terms:

1. *Symmetric stretch* or *symmetric bend (deformation):* Here the local group retains its symmetry during displacement. The symmetric bend of the methylene group, CH_2, is often termed the *scissoring* bend, while the symmetric bend of a methyl group, CH_3, is termed an *umbrella* mode—both descriptions imply the type of displacements that are taking place in the vibration.
2. *Antisymmetric stretch* or *bend* (deformation): Here the vibrating system loses its symmetry during the vibration. The displacements involve a reflective (mirror image) displacement during the opposite phase of the simple harmonic vibration and the motion is termed antisymmetric rather than asymmetric. Antisymmetric bends (deformations) are often classified as *twisting, rocking,* and *wagging* vibrations.
3. Some vibrations involving planar sections of the molecules are referred to as *in-plane* or *out-of-plane.* They can be either symmetric or antisymmetric in nature and if they involve the bending of all the displaced bonds of a set of atoms moving together in the same direction they will be termed *all-in-phase.*
4. *Degenerate* vibrations are defined as the case where two or more molecular vibrations are required to occur at the same frequency (see Chapter 8W, IR sections, Part I B and Part I D3 for more details).
5. *Overtones* (integral multiples of the fundamental mode frequency) and *sum tones* (the sum of two different fundamental modes) are forbidden bands that are almost always very weak. Occasionally these bands are good group frequencies.

Group Frequencies of the Hydrocarbons

Alkanes	Alkynes
Alkenes	Arenes

Alkanes. The C—H vibrational modes of the alkanes (or mixed compounds containing alkyl groups) that are characteristic and reliable group frequencies are summarized in Table 8.1 (also see Chapter 8W, IR section, Part II A). www

These modes give rise to characteristic bands found in the infrared spectrum of alkanes, such as in the spectrum of *n*-hexane shown in Figure 8.2.

Alkenes C═C Stretching. It is possible to classify open-chain unsaturated systems into two groups, those with C═C stretching modes falling above 1660 cm^{-1} and those with modes falling below 1660 cm^{-1} as shown in Table 8.2 (see also Chapter 8W, IR section, Part II B1). www

Alkenes C—H. Several fundamental modes associated with the alkene C—H groups are group frequencies and are summarized in Table 8.3.

Alkynes. The group frequencies of the alkynes are summarized in Table 8.4 (see also Chapter 8W, IR section, Part II C). www

Table 8.1 **Alkane Vibrational Normal Modes**

C—H Vibrational Modes	$\tilde{\nu} \pm 10$ (cm^{-1})
Methyl groups	
Antisymmetric (degenerate) stretch	2960
Symmetric stretch	2870
Antisymmetric (degenerate) deformation	1460
Symmetric (umbrella) deformation	1375
Methylene groups	
Antisymmetric stretch	2925
Symmetric stretch	2850
Symmetric deformation (scissors)	1450
Rocking mode (all-in-phase)	720

Table 8.2 **Substitution Classification of C═C Stretching Frequencies**

C═C Normal Modes	$\tilde{\nu}$ (cm^{-1})
Trans-, tri-, tetrasubstituted	1680–1665
Cis-, vinylidene- (terminal 1,1-disubstituted), vinyl-substituted	1660–1620

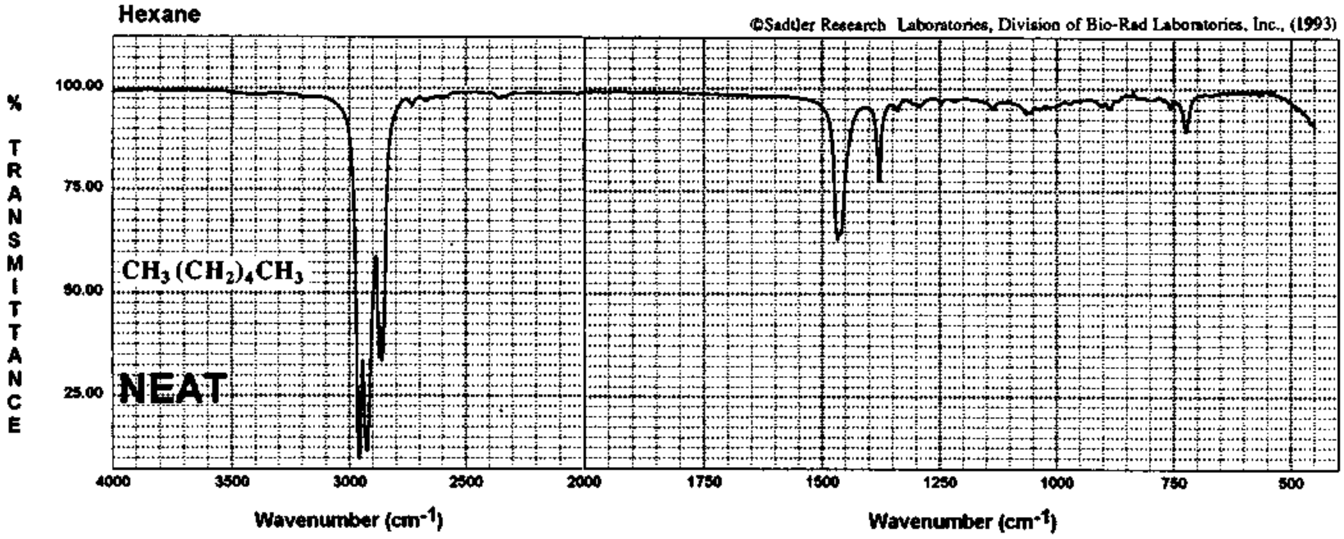

Figure 8.2 IR spectrum: *n*-hexane.

Table 8.3 **Alkene Vibrational Normal Modes**

C—H Vibrational Modes	$\tilde{\nu} \pm 10$ (cm^{-1})
Stretching modes	
Antisymmetric stretch (=CH_2)	3080
Symmetric stretch (=CH_2)	3020
Uncoupled stretch (=CH_2)	3030
Out-of-plane bending modes	
Vinyl group	
Trans hydrogen atoms (in-phase)	990
Terminal hydrogen atoms (wag)	910
Vinylidene group (=CH_2)	
Terminal (wag)	890
Trans alkene	
Trans hydrogen atoms (in-phase)	965
Cis alkene	
Cis hydrogen atoms (in-phase)	~700
Trisubstituted alkene	
Uncoupled hydrogen atom	820
Tetrasubstituted alkene: no vibrational modes seen in IR	

Table 8.4 **Alkyne Vibrational Normal Modes**

C≡C, C—H Vibrational Modes	$\tilde{\nu} \pm 10$ (cm^{-1})
Triple-bond stretch (monosubstituted)	2120
Triple-bond stretch (disubstituted)	2225
R—C≡C—H bond stretch (monosubstituted)	3300

Arenes. The group frequencies of the *phenyl* group can be classified as carbon–hydrogen vibrations consisting of stretching and out-of-plane bending modes, plus carbon–carbon ring stretching and out-of-plane bending modes. The in-plane bending modes in both cases are not effective group frequencies.

The wavenumber values for the all-in-phase C—H bending vibrations are presented in Table 8.5.

The generalized group frequencies of the arenes are summarized in Table 8.6 (see also Chapter 8W, IR section, Part II D).

www

Group Frequencies of Carbonyl Groups: C=O

The carbonyl group is perhaps the single most important functional group in organic chemistry. It is certainly the most commonly occurring functionality. Infrared spectroscopy can play a powerful role in the characterization of the carbonyl because this group possesses all of the properties that give rise to an excellent group frequency. (Table 8.7; for an in depth discussion see Chapter 8W, IR section, Part III A.)

www

Table 8.5 **Arene Out-of-Ring-Plane C—H Deformation Modes**

Arene Fundamentals (C—H bend) (Number of C—H groups directly adjacent)	$\tilde{\nu}$ Range (cm^{-1})
5	770–730
4	770–735
3	810–750
2	860–800
1	900–845

Table 8.6 Arene Group Frequencies

Arene Fundamentals	$\tilde{\nu}$ Range (cm^{-1})
C—H stretch	3100–3000
C=C ring stretch (ν_{8a})	1600 ± 10
C=C ring stretch (ν_{8b})	1580 ± 10
C=C ring stretch (ν_{19a})	1500 ± 10
C=C ring stretch (ν_{19b})	1450 ± 10
C—H out-of-plane bend (1H)	900–860
C—H out-of-plane bend (2H)	860–800
C—H out-of-plane bend (3H)	810–750
C—H out-of-plane bend (4H)	770–735
C—H out-of-plane bend (5H)	770–730
C—C ring out-of-plane bend (1; 1,3; 1,3,5-substituted)	690 ± 10
C—H out-of-plane bend sum tones	2000–1650

The major factors perturbing carbonyl frequencies can be summarized as follows:

Factors That Raise the C=O Frequency

1. Substitution with electronegative atoms
2. Decrease in C—CO—C internal bond angle

Factors That Lower the C=O Frequency

1. Conjugation
2. Hydrogen bonding

Several of these factors may be operating simultaneously, so careful judgment as to the contribution of each individual effect must be exercised in

Table 8.7 Carbonyl Group Vibrational Frequencies

Compound	$\tilde{\nu}$ (cm^{-1})
Ketones, aliphatic, open-chain (R_2CO)	1725–1700
Ketones, conjugated	1700–1675
Ketones, cyclic	*a*
Acyl halides	>1800
Esters, aliphatic	1755–1735
Esters, conjugated	1735–1720
Esters (conjugated to oxygen)	1780–1760
Lactones	*a*
Anhydrides: aliphatic, open-chain	1840–1810 and 1770–1740
Carboxylic acids, aliphatic	1725–1710
Amides	(see Tables 8.22–8.24)
Lactams	*a*
Aldehydes	1735–1720

www

[a] See Chapter 8W, IR section, Part III A.

JWCL196_c08_539-628.qxd 11/18/09 12:44 AM Page 547

EQA

Table 8.8 **Vibrational Normal Modes of the Hydroxyl Group**

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
3500–3200	Very strong	O—H stretch (only strong when hydrogen bonded
1500–1300	Medium strong	O—H in-plane bend (overlap CH_2, CH_3 bend)
1260–1000	Strong	C—C—O antisymmetric stretch
650	Medium	O—H out-of-plane bend

Table 8.9 **Substitution Effects on C—O Stretch of Aliphatic Alcohols**

Type of Alcohol	$\tilde{\nu}_{C—O}$ (cm^{-1})
RCH_2—OH (primary)	1075–1000
R_2CH—OH (secondary)	1150–1075
R_3C—OH (tertiary)	1200–1100
C_6H_5—OH (phenol)	1260–1180

predicting carbonyl frequencies. This judgment develops rapidly with practice at interpretation.

Group Frequencies of the Heteroatom Functional Groups

(Alkanes)
Alcohols
Acyl halides
Amines, primary
Isocyanates
Aldehydes
Carboxylic acids
Nitriles
Thiols
Ketones
Anhydrides
Amides, primary
Halogens
Esters
Ethers
Amides, secondary
Phenyl

Hexane. Refer to Table 8.1 (see also Chapter 8W, IR section, Part II A, and Fig. 8.2).

www

Alcohols. A very intense band appears at ~3350 cm^{-1}, which is assigned to the stretching mode of the O—H group (Table 8.8; also see Chapter 8W, IR section, Fig. W8.24).

www

Of particular importance is a strong band in the spectrum of aliphatic alcohols usually located near 1060 cm^{-1}. This absorption has been identified as the C—O stretching mode. The vibrational displacements of this fundamental are similar to the antisymmetric stretch of water (see Chapter 8W for a detailed discussion of the vibrational modes of the water molecule). Since the vibration involves significant displacement of the adjacent C—C oscillator, the vibration will be substitution sensitive. These latter shifts can be of value in determining the nature of the alcohol (primary, secondary, or tertiary, see Table 8.9).

www

Aldehydes. The aldehyde functional groups gives rise to several good group frequencies (Table 8.10; also see Chapter 8W, IR section, Fig. W8.25).

www

Ketones. The only group frequency mode associated directly with aliphatic ketones is the stretching frequency ($\tilde{\nu}_{C=O}$~1720 cm^{-1}), which occurs within the expected region as discussed above. There are, however, several other related bands (Table 8.11; also see Chapter 8W, IR section, Fig. W8.26).

www

Table 8.10 **Vibrational Normal Modes of the Aliphatic Aldehyde Group**

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
2750–2720	Weak to medium	C(O)—H stretch (see also online Chapter 8W, IR section)
1735–1720	Very strong	C=O stretch
1420–1405	Medium	CH_2 symmetric bend, —CH_2—α to —CHO
1405–1385	Medium	C—H in-plane bend

www

CONFIRMING PAGES aptara

Table 8.11 Normal Vibrational Modes of Aliphatic Ketones

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
3430–3410	Very weak	Overtone of carbonyl stretch
1725–1700	Very strong	C═O
1430–1415	Medium	—CH_2— symmetric bend, —CH_2—α to ketone C═O

Table 8.12 Vibrational Normal Modes of the Aliphatic Ester Group

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
1755–1735	Very strong	C═O stretch
1370–1360	Medium	CH_3 symmetric bend α to ester C═O
1260–1230	Very strong	C—CO—O antisymmetric stretch —acetates
1220–1160	Very strong	C—CO—O antisymmetric stretch —higher esters
1060–1030	Very strong	O—CH_2—C antisymmetric stretch —1° acetates
1100–980	Very strong	O—CH_2—C antisymmetric stretch —higher esters (may overlap with upper band)

Table 8.13 Vibrational Normal Modes of the Acyl Halide Group

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
1810–1800	Very strong	C═O stretch, acyl chlorides
1415–1405	Strong	—CH_2—symmetric bend, α to —COCl carbonyl

Esters

The very strong band found at ~1745 cm^{-1} is typical of the carbonyl frequency of an aliphatic ester, particularly aliphatic acetate esters (Table 8.12; also see Chapter 8W, IR section, Fig. W8.27).

Acyl Halides. The carbonyl stretching mode dominates the spectrum in aliphatic acyl halides. In acyl chlorides it is an extremely intense band occurring near 1800 cm^{-1} (Table 8.13; also see Chapter 8W, IR section, Fig. W8.28).

Carboxylic Acids. Acids, observed in the solid or pure liquid states, often possess a very intense band with a width at one-half peak height of about 1000 cm^{-1}, which covers the region 3500–2200 cm^{-1}. This absorption is characteristic of very strongly hydrogen-bonded carboxylic acid groups (Table 8.14; also see Chapter 8W, IR section, Fig. W8.29).

Anhydrides. The coupling of the anhydride carbonyls through the ether oxygen splits the carbonyls (in the aliphatic case $\tilde{\nu}_{c=o}$ = ~1830, 1760 cm^{-1}) by about 70 cm^{-1} (Table 8.15; also see Chapter 8W, IR section, Fig. W8.30).

Ethers. The large intensity associated with antisymmetric C—O—C stretching mode relative to the other bands occurring in this part of the fingerprint region, particularly in aliphatic compounds, makes it possible, in most cases, to assign

Table 8.14 Vibrational Normal Modes of the Carboxylic Acid Group

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
3500–2500	Very very strong	O—H stretch intensified by hydrogen bonding
2800–2200	Very weak	Overtone and sum tones
1725–1710	Very strong	C═O antisymmetric hydrogen-bonded dimer stretch
1450–1400	Strong	CH_2—CO—O antisymmetric stretch mixed with O—H bend
1300–1200	Strong	CH_2—CO—O antisymmetric stretch mixed with O—H bend
950–920	Medium	Out-of-plane O—H bend, acid dimer

Table 8.15 Vibrational Normal Modes of the Anhydride Group

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
1840–1810	Very strong	C═O in-phase stretch
1770–1740	Very strong	C═O out-of-phase stretch
1420–1410	Strong	—CH_2— symmetric bend α to C═O
1100–1000	Very strong	C—O stretch, mixed modes

Table 8.16 Vibrational Normal Modes of the Ether Group

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
1150–1050	Strong	C—O—C antisymmetric stretch, mixed mode

with confidence the observed strong band (Table 8.16; also see Chapter 8W, IR section, Fig. W8.31).

www

Primary Amines. The spectra of these bases usually possess two bands ($\tilde{\nu}_{N—H}$ = ~3380, ~3300 cm^{-1}) of medium-to-weak intensity. These bands are assigned to the antisymmetric and symmetric N—H stretching modes, respectively, of the primary amino group (Table 8.17; also see Chapter 8W, IR section, Fig. W8.32).

www

Nitriles. The very strong triple bond present in the nitrile group (as in the case of the alkynes) contributes to an unusually high stretching frequency, and the polar character of the group gives rise to very intense bands (Table 8.18; also see Chapter 8W, IR section, Fig. W8.33).

www

Primary Amides. The highly polar amide group leads to very strong hydrogen bonding, which in turn leads to greatly intensified N—H antisymmetric and symmetric stretching modes ($\tilde{\nu}_{N—H}$ = ~3375, ~3200 cm^{-1}; see Table 8.19; also see Chapter 8W, IR section, Fig. W8.34).

www

Secondary Amides. The single N—H group present in secondary amides gives rise to a very strong band near about 3300 cm^{-1}, which is indicative of strong hydrogen bonding. The drop in frequency from that of the primary —NH_2 scissoring mode near 1600 cm^{-1} allows for confident assignment of substitution on secondary amide groups (Table 8.20; also see Chapter 8W, IR section, Fig. W8.35).

www

Table 8.17 **Vibrational Normal Modes of the Primary Amine Group**

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
3400–3200	Weak to medium	NH_2 stretch doublet, (antisymmetric and symmetric modes)
1630–1600	Medium	NH_2 symmetric bend
820–780	Medium	NH_2 wag

Table 8.18 **Vibrational Normal Modes of the Nitrile Group**

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
2260–2240	Strong	$C{\equiv}N$ stretch, aliphatic
2240–2210	Strong	$C{\equiv}N$ stretch, conjugated

Table 8.19 **Vibrational Normal Modes of the Primary Amide Group**

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
3400–3150	Very strong	—NH_2 antisymmetric and symmetric stretching modes, hydrogen bonded
1680–1650	Very strong	$C{=}O$ stretch, hydrogen bonded
1660–1620	Strong	—NH_2 symmetric bend (overlap with $C{=}O$ stretch)
1430–1410	Strong	—CH_2—symmetric bend α to amide carbonyl
750–650	Medium	—NH_2 wag

Table 8.20 **Vibrational Normal Modes of the Secondary Amide Group**

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
3350–3250	Strong	—NH stretch, intensified by hydrogen bonding
3125–3075	Medium	Overtone N—H bend (see also)
1670–1645	Very strong	$C{=}O$ stretch, hydrogen bonded
1580–1550	Strong	N—H in-plane bend (see also)
1415–1405	Strong	—CH_2— symmetric bend α to amide $C{=}O$
1325–1275	Medium	C—N stretch mixed with N—H in-plane bend
725–680	Medium	N—H out-of-plane bend

Table 8.21 **Vibrational Normal Modes of the Amide Carbonyl: Solution and Solid-Phase Data**

Amide	Dilute Solution (cm^{-1})	Solid (cm^{-1})
R—CO—NH_2 (primary)	~1730	~1690–1650
R—CO—NHR (secondary)	~1700	~1670–1630
R—CO—NR_2 (tertiary)	~1650	~1650

Studies of amide carbonyl frequencies in dilute nonpolar solution indicate that hydrogen-bonding effects are largely responsible for the low frequencies observed with primary and secondary amides, but play no role in tertiary amides (see Table 8.21).

JWCL196_c08_539-628.qxd 11/16/09 4:35 PM Page 551

EQA

Table 8.22 Vibrational Normal Mode of the Isocyanate Group

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
2280–2260	Very strong	—N═C═O antisymmetric stretch

Table 8.23 Vibrational Normal Mode of the Thiol Group

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
2580–2560	Weak	S—H stretch

Table 8.24 Vibrational Normal Mode of the Alkyl Chloro Group

$\tilde{\nu}$ (cm^{-1})	Intensity	Mode Description
750–650	Strong	C—Cl stretch (see also)

Table 8.25 Vibrational Normal Modes of the Aryl Chloro Group

$\tilde{\nu}$ (cm^{1})	Intensity	Mode Description
3080	Medium	C—H stretch, bonded to ring carbon
1585	Strong	ν_{8a} ring stretching
1575	Weak	ν_{8b} ring stretching
1475	Strong	ν_{19a} ring stretching
1450	Strong	ν_{19b} ring stretching
747	Strong	C—H all-in-phase, out-of-plane bend
700	Strong	C—Cl stretch
688	Strong	Ring deformation
1945, 1865, 1788, 1733	All weak	Sum tones, out-of-plane C—H bends, pattern matches monosubstitution of ring

Isocyanates. The range of stretching frequencies observed for alkyl-substituted isocyanates is very narrow, $\tilde{\nu} = 2280 - 2260\,\mathrm{cm}^{-1}$, which implies little coupling to the rest of the system (Table 8.22; also see Chapter 8W, IR section, Fig. W8.36).

Thiols. Although weak absorption is associated with the S—H stretching fundamental the band is generally found in a very open region of the infrared spectrum (Table 8.23; also see: Chapter 8W, IR section, Fig. W8.37).

Alkyl Halides. The massive halogen atom is connected to the alkyl section by a fairly weak but highly polarized bond, which dictates that the C—X stretching frequency appears as an intense band at low frequencies (Table 8.24; also see Chapter 8W, IR section, Fig. W8.38).

Aryl Halides (Chlorobenzene). The final system to be considered in this section is the aryl halide, chlorobenzene. Based on the above assignments the group frequencies of the complete hydrocarbon portion and the heteroatom functional group can be assigned as in Table 8.25 (also see Chapter 8W, IR section, Fig. W8.39).

INFRARED SPECTROSCOPY INSTRUMENTATION AND SAMPLE HANDLING

Instrumentation

The workhorse infrared instrument used for routine characterization of materials in the undergraduate organic laboratory is the optical-null double-beam grating spectrometer (Fig. 8.3*a*). For a discussion of double-beam spectrometers, see the UV-vis instrumentation discussion (p. 604). Although many

CONFIRMING PAGES aptara

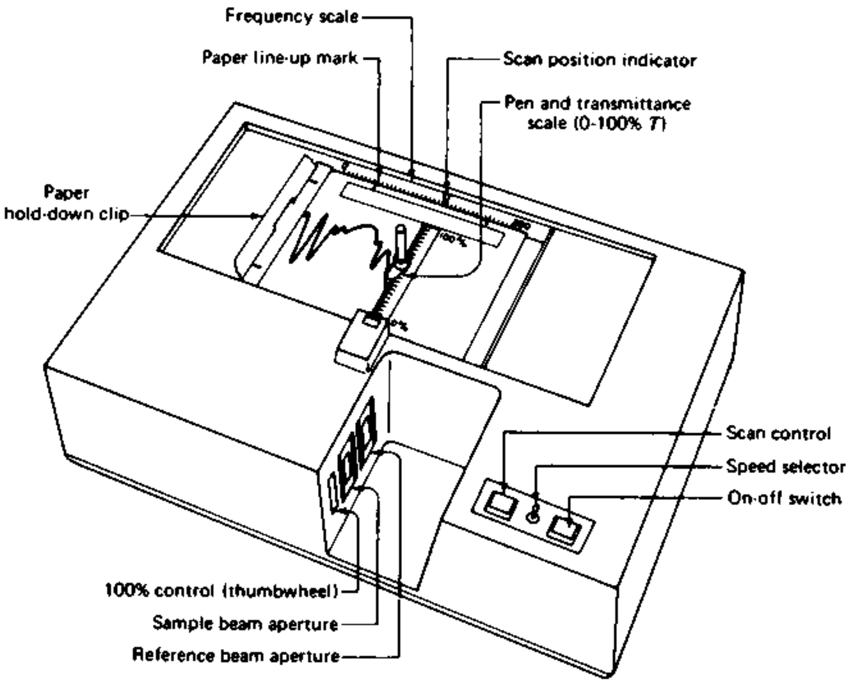
Frequency scale
Paper line-up mark
Scan position indicator
Pen and transmittance scale (0-100% T)
Paper hold-down clip
Scan control
Speed selector
On-off switch
100% control (thumbwheel)
Sample beam aperture
Reference beam aperture

Figure 8.3*a* The Perkin–Elmer model 710B IR spectrometer. From Zubric, James W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, 2008. (Reprinted by permission of John Wiley & Sons, Inc., New York.)

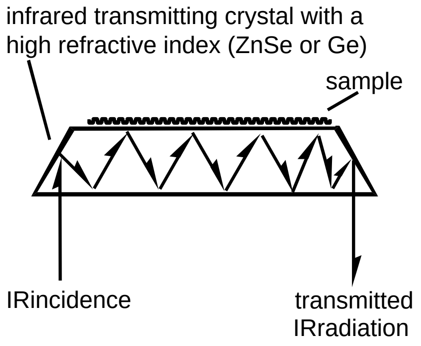
infrared transmitting crystal with a high refractive index (ZnSe or Ge)
sample
IRincidence
transmitted IRradiation

Figure 8.3*b* Attenuated Total Reflectance (ATR).

undergraduate instructional laboratories still utilize this type of instrumentation, the winds of change are blowing. ATR (attenuated total reflectance) FT-IRs and lower-cost FT-IR spectrometers, which both depend on computer manipulation of the spectral data, are becoming the infrared instructional instrumentation of choice. Two of the many benefits when using the more expensive ATR FT-IR spectrometer are faster sampling and spectral reproducibility. Even though ATR plates are prone to contamination, one key advantage when working with an ATR crystal (Fig. 8.3*b*) is that no sample preparation is required. This avoids the need to use KBr or mineral oil (Nujol) when working with solids! Many of the spectra utilized in the interpretive discussions (Chapter 8W, IR section) were generated on a prototype of this kind of infrared instrumentation, the Perkin–Elmer model 1600.

This instrument can acquire 16 scans and carry out the required calculations in 42 s. While the spectrum is being printed out(~40 s) the data on a second sample can be acquired. The 42-s acquisition data are significantly superior to those recorded by dispersive instruments that take from 5 to 8 min to scan a sample from 4000 to 600 cm^{-1}.

A short description of FT-IR spectrometers is included in the discussion of instrumentation on the website (Chapter 8W, IR section, Part IV).

Sample Handling in the Infrared

If not working with an ATR FT-IR, the standard techniques of sample preparation employed to obtain infrared spectra of microscale laboratory products are the use of capillary films with liquids on NaCl (or AgCl) plates and the use of KBr disks and melts in solids. This, of course, assumes that for a spectrum to be obtained in the infrared region, the sample must be mounted in a cell that is transparent to the radiation. With an ATR FT-IR spectrometer, liquids and solids are placed in direct contact with the ATR crystal. Since glass and quartz absorb in this spectral region, cells constructed of these materials cannot be

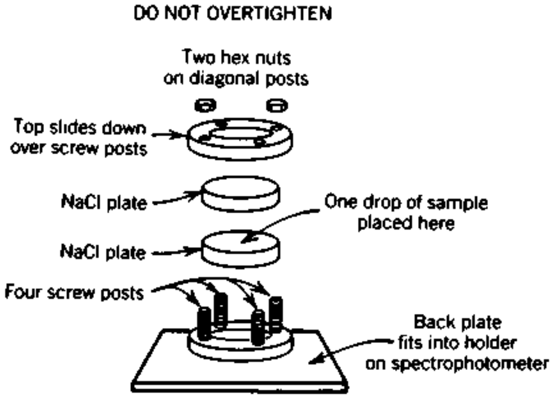
DO NOT OVERTIGHTEN
Two hex nuts on diagonal posts
Top slides down over screw posts
NaCl plate
One drop of sample placed here
NaCl plate
Four screw posts
Back plate fits into holder on spectrophotometer

Figure 8.4 IR salt plates and holder. From Zubric, James W. *The Organic Chem Lab Survival Manual,* 7th ed.; Wiley: New York, 2008. (Reprinted by permission of John Wiley & Sons, Inc., New York.)

used when working with spectrometers not utilizing ATR technology. Accordingly, alkali metal halides have large spectral regions of transmission in the infrared, as do silver halides. Sodium chloride is the most commonly used material in cell windows in infrared sampling.

Liquid Samples. For materials boiling above 100 °C, the procedure is very simple. Using a syringe or Pasteur pipet, place 3–5 μL of sample on a polished plate of sodium chloride or silver chloride or directly on the ATR crystal. If working with a NaCl or AgCl plate, cover it with a second plate of the same material and clamp it in a holder that can be mounted vertically in the instrument. Be sure that the plates are clean when you start and when you are through! Obviously, the sodium chloride plates cannot be cleaned with water. Silver chloride is very soft and scratches easily; it also must be kept in the dark when not in use because it darkens quickly in direct light. Spectra obtained in this fashion are referred to as *capillary film spectra* (Fig. 8.4).

Solution Spectra and the Spectra of Materials Boiling Below 100 °C. These samples generally require a sealed cell constructed of either sodium chloride or potassium bromide windows. Such cells are expensive and need careful handling and maintenance. They are assembled as shown in Figure 8.5.

Solid Samples using non-ATR Spectrometers. Solid powders could be mounted on horizontal sodium chloride plates, and the beam diverted

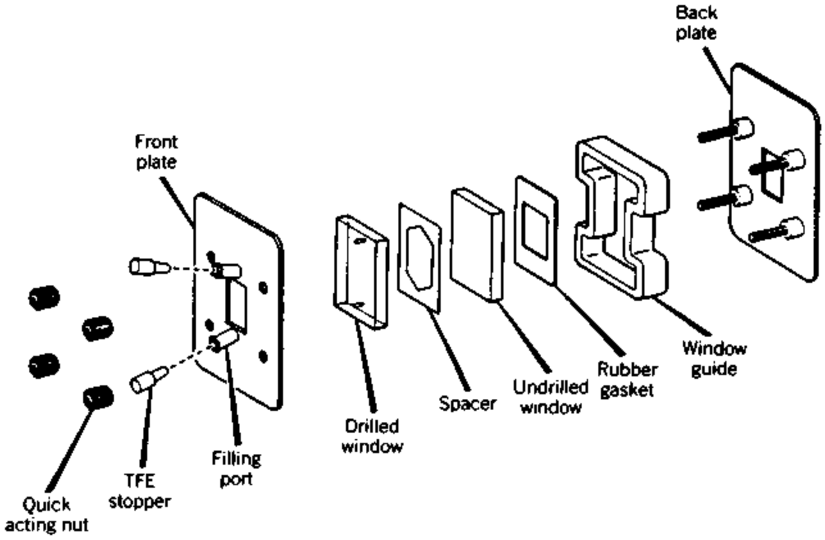
Back plate
Front plate
Window guide
Rubber gasket
Undrilled window
Spacer
Drilled window
Filling port
TFE stopper
Quick acting nut

Figure 8.5 Sealed demountable cell or demountable cell with ports. (Courtesy of the Perkin-Elmer Corp., Norwalk, CT.)

through the sample by mirrors. This procedure would make sample preparation very easy for solids. Unfortunately, powders tend to scatter the entering radiation very efficiently by reflection, refraction, and molecular scattering. Some of these effects become rapidly magnified at higher frequencies, since they vary with the fourth power of the frequency. Thus, in solid-sample scattering a lot of energy is scattered away from the sample beam. This results in poor absorption spectra, as the instrument is forced to operate at very low energies. The detector cannot differentiate between a drop in energy from absorption or one derived from scattering.

For materials melting below 80 °C the simplest technique is to mount the sample between two salt plates and **gently** warm with a heat lamp until melting occurs. With the fast acquisition times of interferometers, the melting point range is now as high as 100 °C and the spectrum can be obtained so rapidly that the sample does not have time to cool and crystallize. (Heated cells are used in research laboratories, but they are rather expensive and difficult to maintain.) For substances melting above 100 °C, the sampling routine most often employed to avoid scattering problems is the potassium bromide (KBr) disk. Potassium bromide is transparent to infrared radiation in the region of interest. Most important, however, the KBr makes a much better match of the refractive indexes between the sample and its matrix than does air. Thus, reflection and refraction effects at the crystal faces of the sample are greatly suppressed.

In the KBr method the sample (2–3 mg) is finely ground in a mortar, the finer the better for lower reflection or refraction losses. Then, 150 mg of previously ground and dried KBr is added to the mortar and quickly mixed by stirring, not grinding, it with the sample. (Potassium bromide is very hygroscopic and will rapidly pick up water while being ground in an open mortar.) When mixing is complete the mixture is transferred to a die and pressed into a solid disk. Potassium bromide will flow under high pressure and seal the solid sample in a glasslike matrix. Several styles of dies are commercially available. For routine use a die consisting of two stainless steel bolts and a barrel is the simplest to operate (see Fig. 8.6). The ends of the bolts are polished flat to form the die faces. The first bolt is seated to within a turn or two of the head. Then the sample mixture is added (avoid breathing over the die while adding the sample). The second bolt is firmly seated in the barrel, and then the clamped assembly is tightened by a torque wrench to 240 in./lb. After standing for 1.5 min, the two bolts are removed, leaving the KBr disk mounted in the center of the barrel, which can then be mounted in the instrument. After the spectrum of the sample is run, the disk can be retrieved and the sample recovered if necessary (Fig. 8.6). *Always clean the die immediately after use. KBr is highly corrosive to steel.*

When infrared spectra are obtained, it is important to establish that the wavenumber values have been accurately recorded. Successful interpretation of the data often depends on very small shifts in these values. Calibration of the frequency scale is usually accomplished by obtaining the spectrum of a reference compound, such as polystyrene film. To save time, record absorption peaks only in the region of particular interest (this applies only to dispersive instrument derived data).

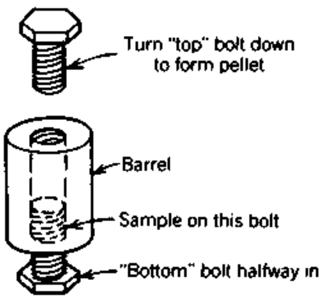

Figure 8.6 The KBr pellet mini-press. From Zubric, James W. *The Organic Chem Lab Survival Manual,* 7th ed.: Wiley: New York, 2008. (Reprinted by permission of John Wiley & Sons, Inc., New York.)

NOTE. Most of the infrared spectra found in the experimental sections of the text, which are Fourier-transform derived (Perkin–Elmer 1600), have been plotted on a slightly different scale than the other spectra presented in the text and on the website. The former spectra utilize a 12.5-cm^{-1}*/mm format below 2000* cm^{-1} *and undergo a 2:1 compression above 2000* cm^{-1} *(25* cm^{-1}*/mm).*

QUESTIONS

NOTE. Some of the following questions assume that the student is familiar with the infrared material contained on the website available to refer to if needed.

www

8-1. The form of the C—H out-of-plane bending vibrations of the vinyl group are shown below:

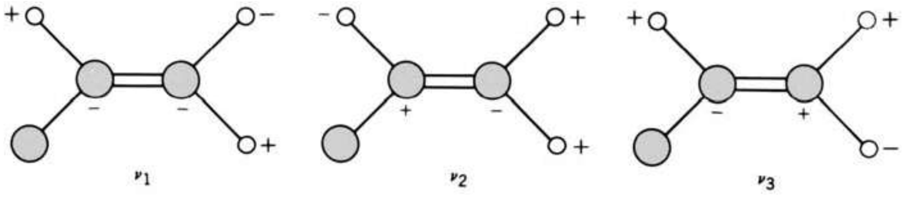

The first two vibrational modes give rise to excellent group frequencies, while the third fundamental does not lend itself to these correlations.

- **(a)** Explain the factors that lead to the third vibrational mode being such a poor group frequency.
- **(b)** Predict the location in the spectrum of the third fundamental vibration.

8-2. In the figure below, the mass of the terminal hydrogen atoms on the acetylene is hypothetically varied from zero to infinity. The response of the C—H symmetric stretching (3374 cm^{-1}) and triple-bond stretching (1974 cm^{-1}) modes to the change in mass is shown.

- **(a)** Calculate the expected deuterium isotope shift for the C—H symmetric stretching mode. Is the hypothetical value close to the calculated value? Explain.
- **(b)** Explain why the triple-bond stretching frequency is approximately 100 cm^{-1} higher for high-mass terminal isotopes (>100) than for the low-mass terminal isotopes (<3).

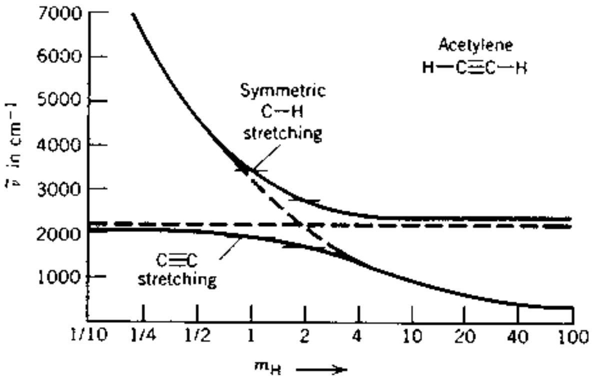

8-3. Acetylene has two C—H groups. It will have two C—H stretching frequencies, the in-phase and out-of-phase stretching modes. The in-phase (symmetric) stretch occurs at 3374 cm^{-1} and the out-of-phase stretch at 3333 cm^{-1}. Explain why the in-phase vibration is located at a higher frequency than the out-of-phase stretch.

8-4. The carbonyl stretching frequencies of a series of benzoyl derivatives are listed below:

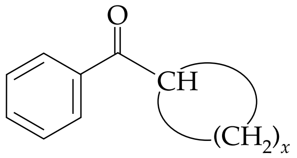

x	$\tilde{\nu}_{C=O}$ (CCl_4)
2	1677
3	1686
4	1687
5	1686

Consider the $\tilde{\nu}_{C=O}$ of acetone at 1715 cm^{-1} as a reference frequency and identify the factors affecting $\tilde{\nu}_{C=O}$ in the series of compounds listed.

8-5. Explain how mass effects act to lower the carbonyl frequency, as well as how inductive and hyperconjugation effects act to raise the carbonyl frequency of aldehydes relative to ketones.

8-6. The carbonyl stretching frequency of aliphatic carboxylic acids in dilute solution is located near 1770 cm^{-1}. This frequency is much higher than the carbonyl frequency of these substances when measured neat (~1720 cm^{-1}). Also, it is considerably higher than the corresponding simple aliphatic ester value (1745 cm^{-1}). Explain.

8-7. In a number of cases, dipolar interactions control the frequency shifts found in carbonyl stretching vibrations. The table lists wavenumber shifts in going from neat to dilute nonpolar solutions. Explain the observed values.

Carbonyl Dipolar Interactions[a]

Compound	$\Delta\tilde{\nu}$ (cm^{-1})
Acetyl chloride	15
Phosgene	13
Acetone	21
Acetaldehyde	23
N,N-Dimethylformamide	50

[a] Shift measured between dilute nonpolar solution and neat sample.

8-8. The antisymmetric $-CH_2-CO-O-$stretching vibration in carboxylic acids is heavily mixed with the in-plane bending mode of the O—H group. In alcohols these two vibrations seldom show evidence of mechanical coupling. Explain.

8-9. Conjugation of the functional group in alkyl isocyanates has little impact on the antisymmetric —N═C—O stretching vibration located near 2770 cm^{-1} Explain.

8-10. In the infrared spectrum of 2-aminoanthraquinone (**I**) two carbonyl stretching frequencies are observed at 1673.5 and 1625 cm^{-1}:

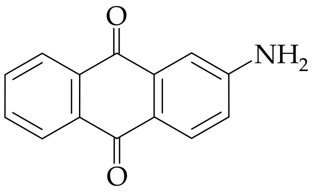

2-Aminoanthraquinone
I

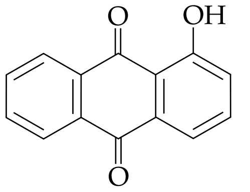

1-Hydroxyanthraquinone
II

(a) Assign carbonyl bands in the infrared spectrum to the carbonyl groups in structure **I** and explain your reasoning.

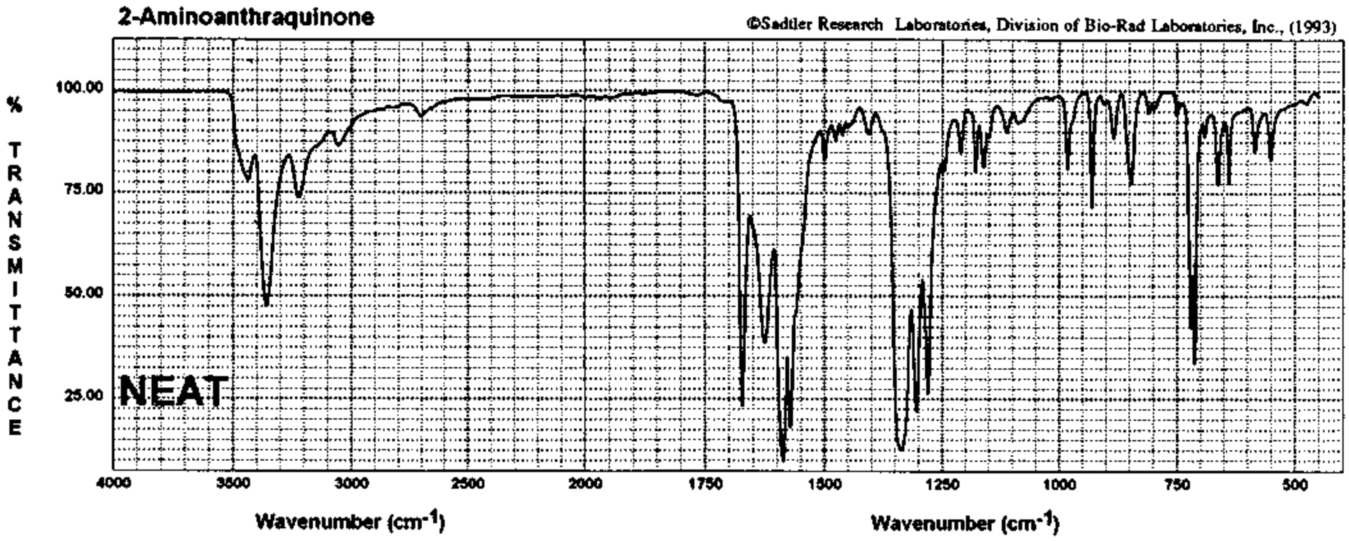

IR spectrum: 2-Aminoanthraquinone.

(b) The infrared spectrum of 1-hydroxyanthraquinone (**II**) also exhibits two carbonyl frequencies, which are located at 1675 and 1637 cm^{-1}. Assign the carbonyl groups to the related absorption bands. Explain your reasoning.

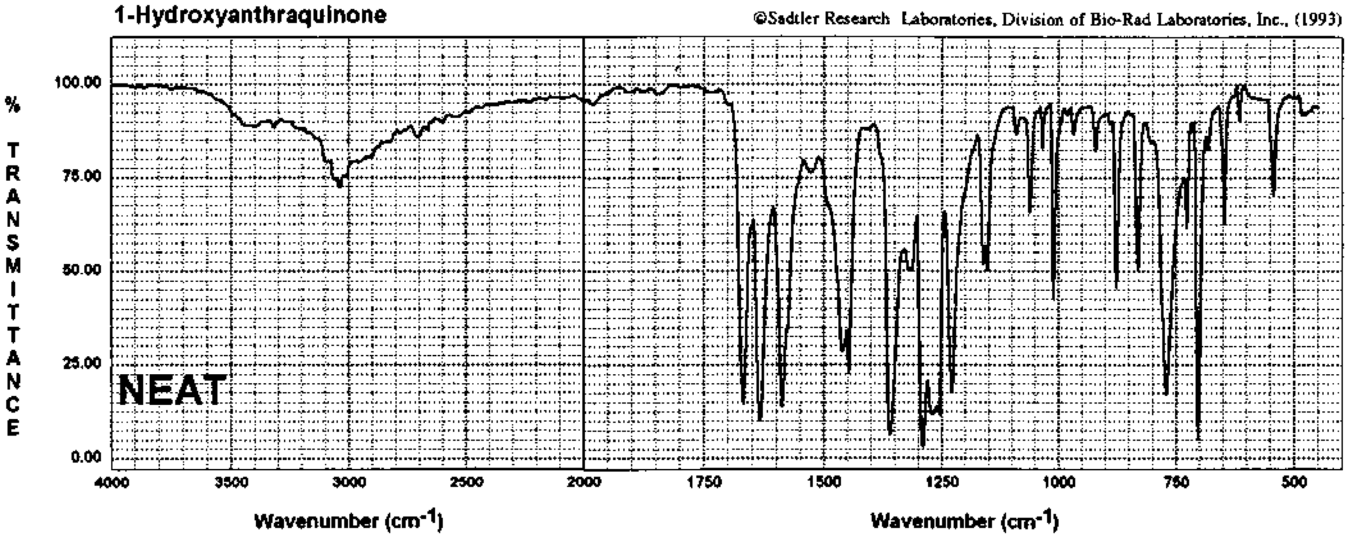

IR spectrum: 1-Hydroxyanthraquinone.

(c) The spectrum of 2-hydroxyanthraquinone exhibits a single carbonyl stretching frequency near 1673 cm^{-1}. Explain why a single carbonyl band would be expected in the system and why this vibration is located at 1673 cm^{-1}.

8-11. Suggest a possible structure for the hydrocarbon C_6H_{14}, which has the infrared spectrum shown here: Is there more than one correct structure?

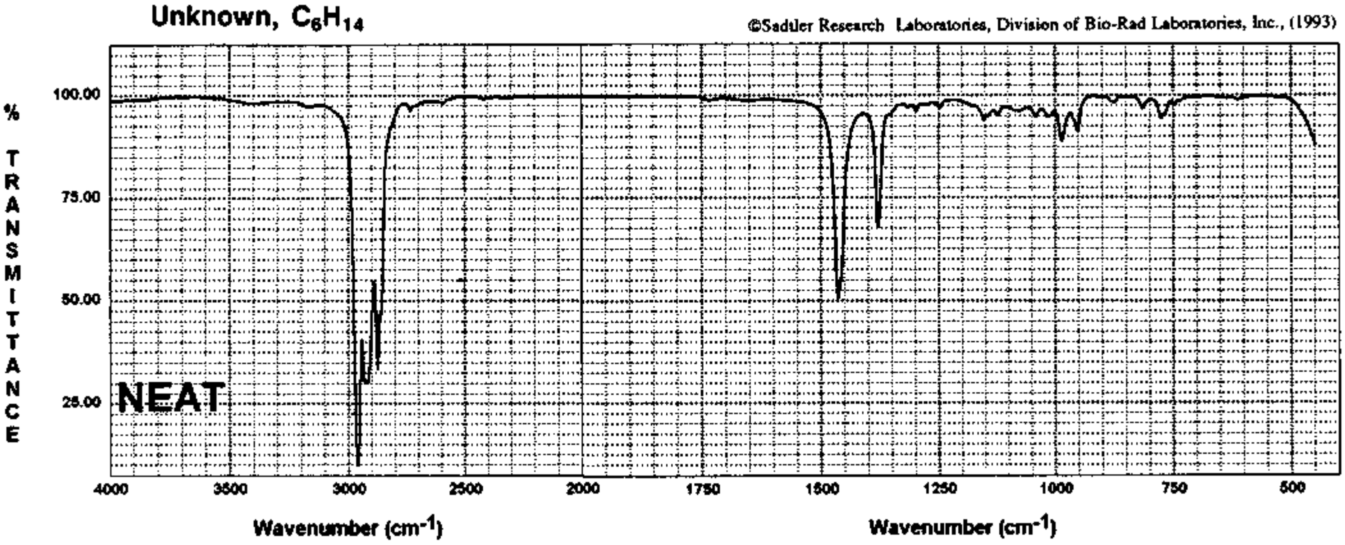

IR unknown spectrum: C_6H_{14}.

8-12. The hydroxylamine **I** can be oxidized by MnO_2 to the amide oxohaemanthidine (**II**). In dilute solution the carbonyl absorption band of **II** occurs at 1702 cm^{-1}. Explain this observation.

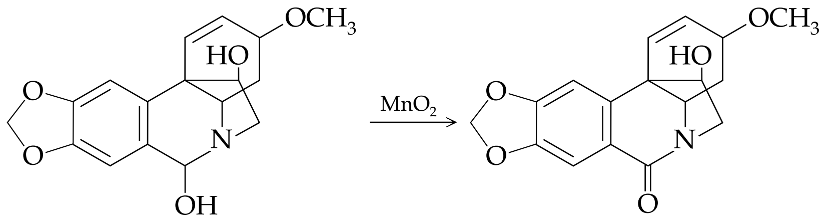

I **II**

8-13. Identify the following alkenes. All samples were obtained from distillation cuts in the C_6 boiling range.

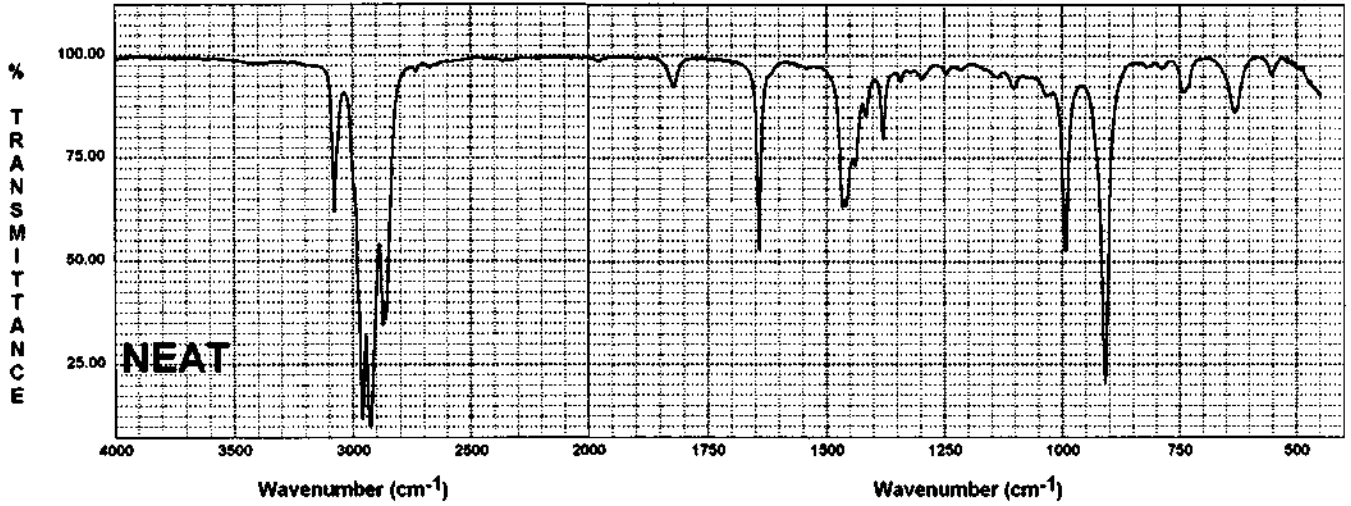

©Sadtler Research Laboratories, Division of Bio-Rad Laboratories, Inc., (1993)

Infrared unknown spectrum *a*.

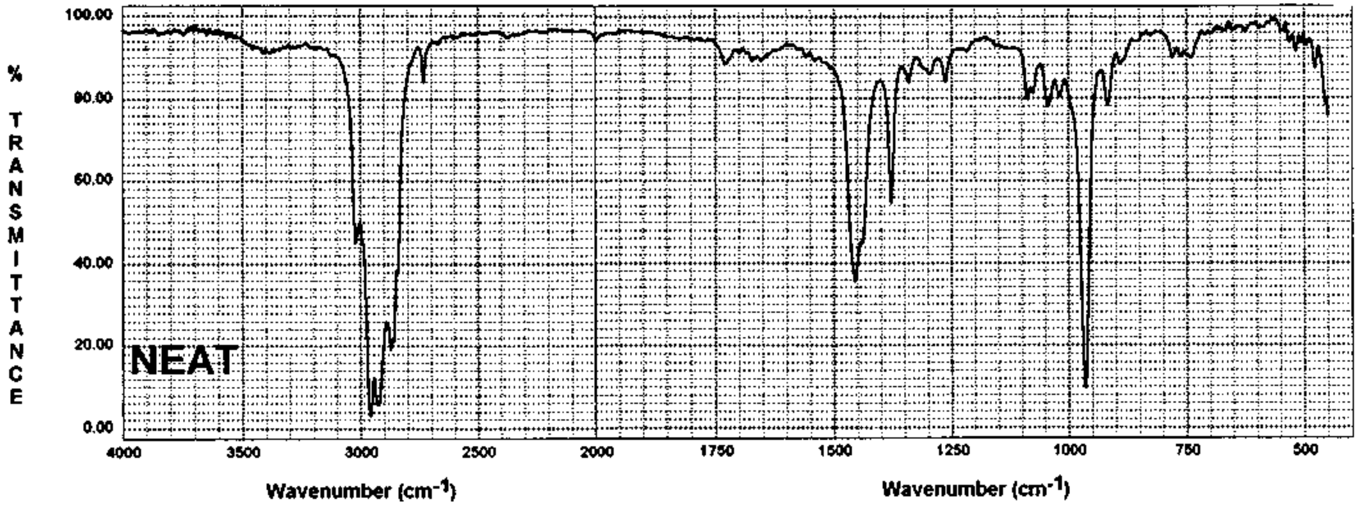

©Sadtler Research Laboratories, Division of Bio-Rad Laboratories, Inc., (1993)

Infrared unknown spectrum *b*.

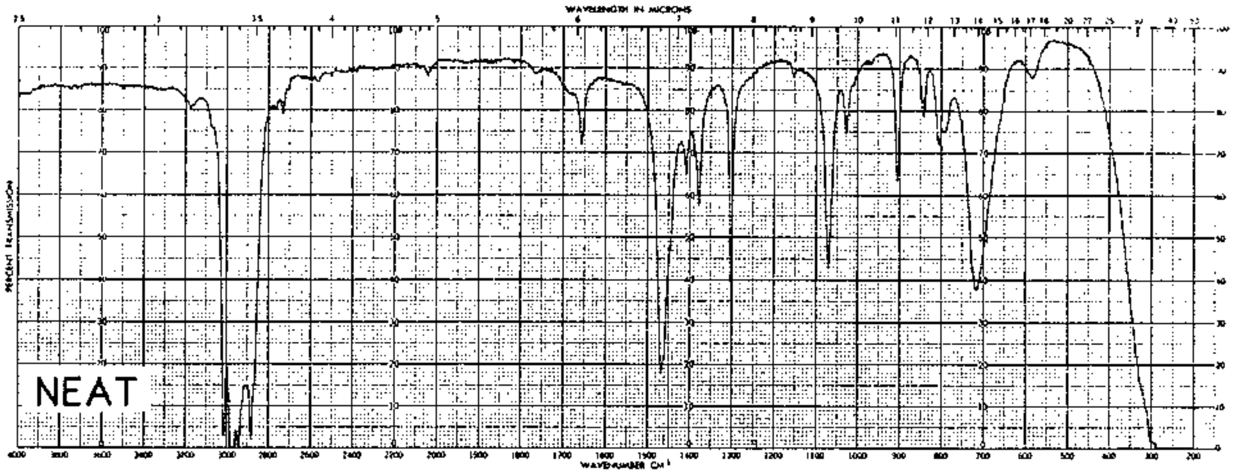

Infrared unknown spectrum *c*. *(Courtesy of Bowdoin College.)*

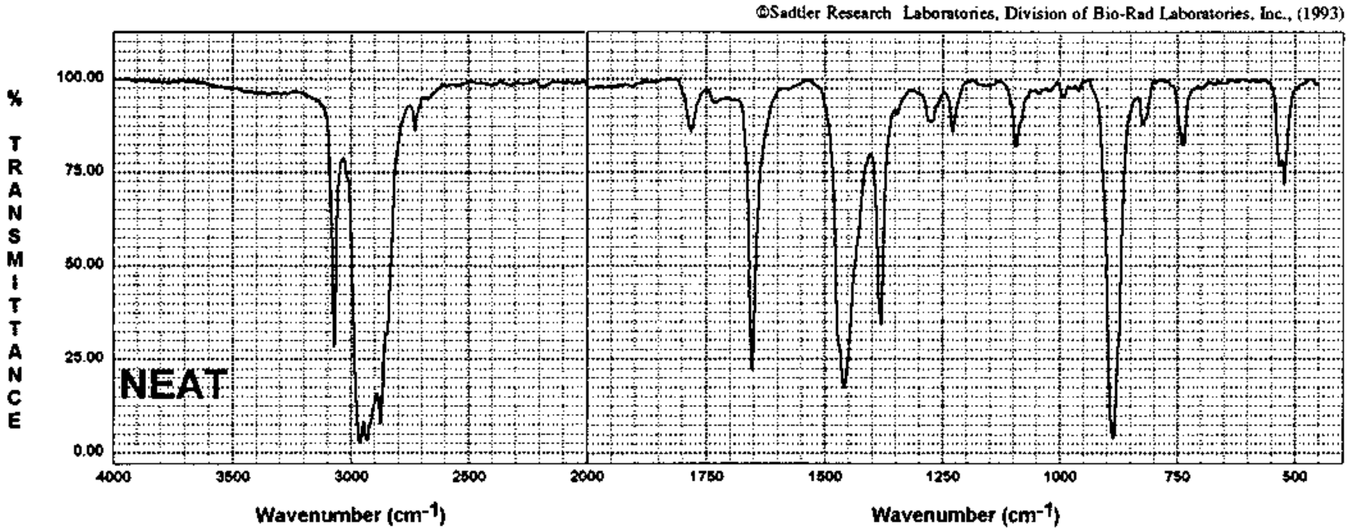

Infrared unknown spectrum *d*.

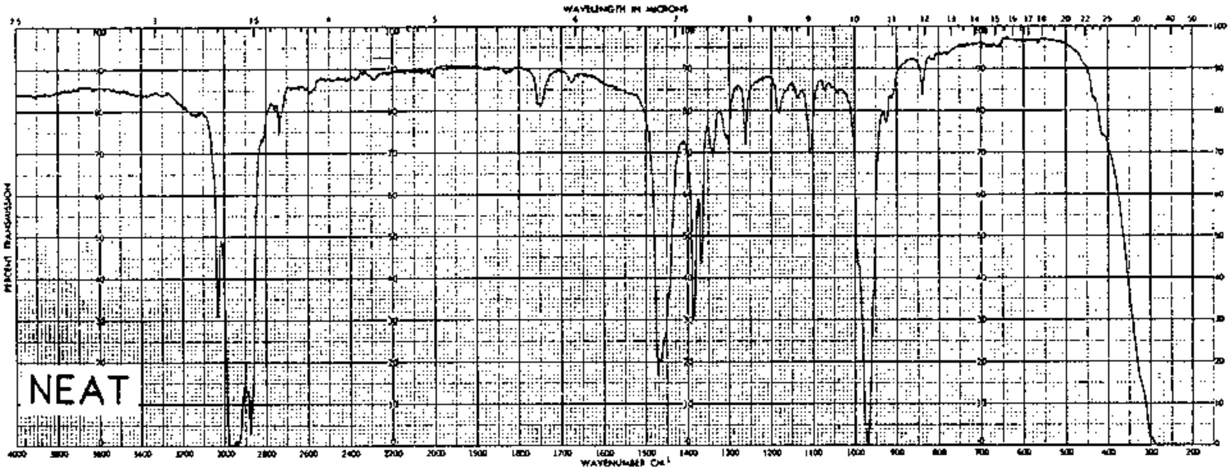

Infrared unknown spectrum *e*. *(Courtesy of Bowdoin College.)*

8-14. The infrared spectra of the three xylene (dimethylbenzene) isomers, and an additional aromatic hydrocarbon, are given below. Assign the spectra to the isomers and suggest a potential structure for the remaining unknown substance.

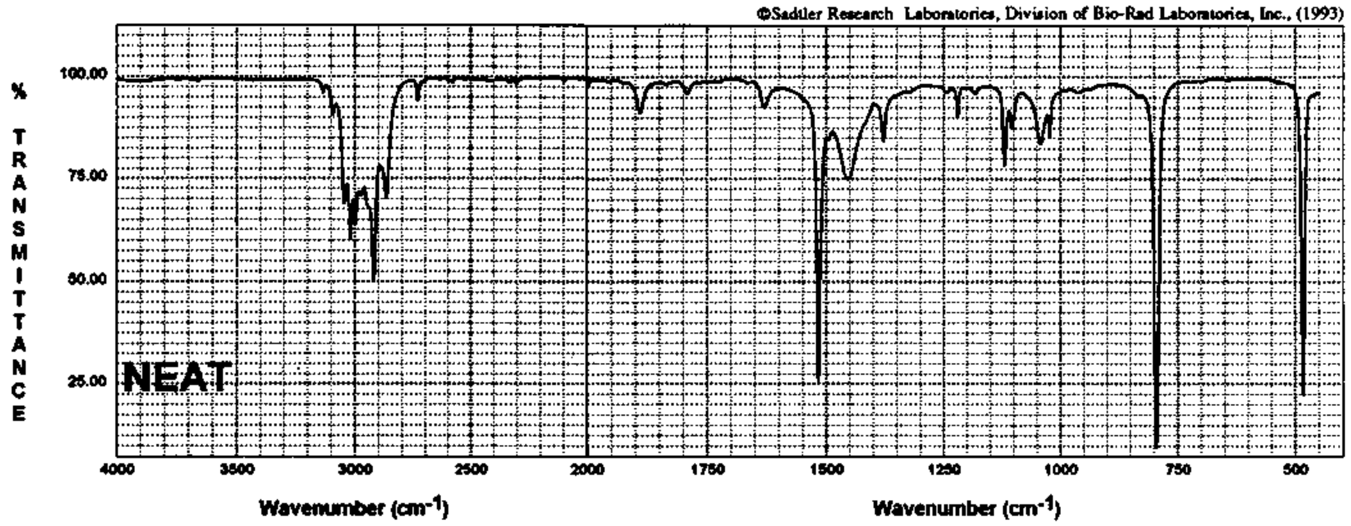

Infrared unknown spectrum *a*.

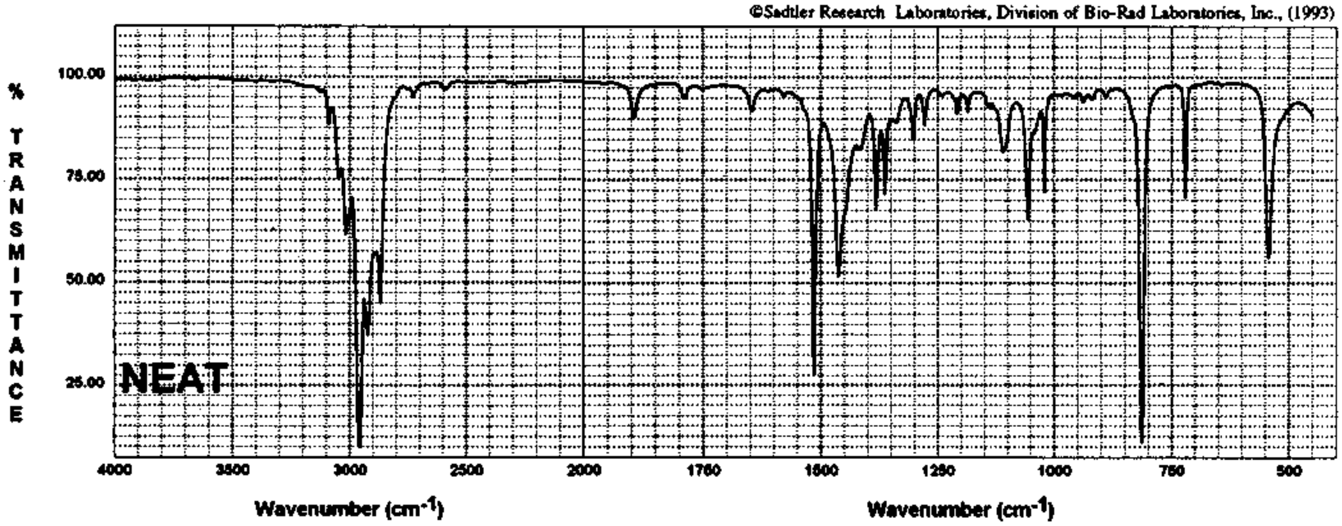

Infrared unknown spectrum *b*.

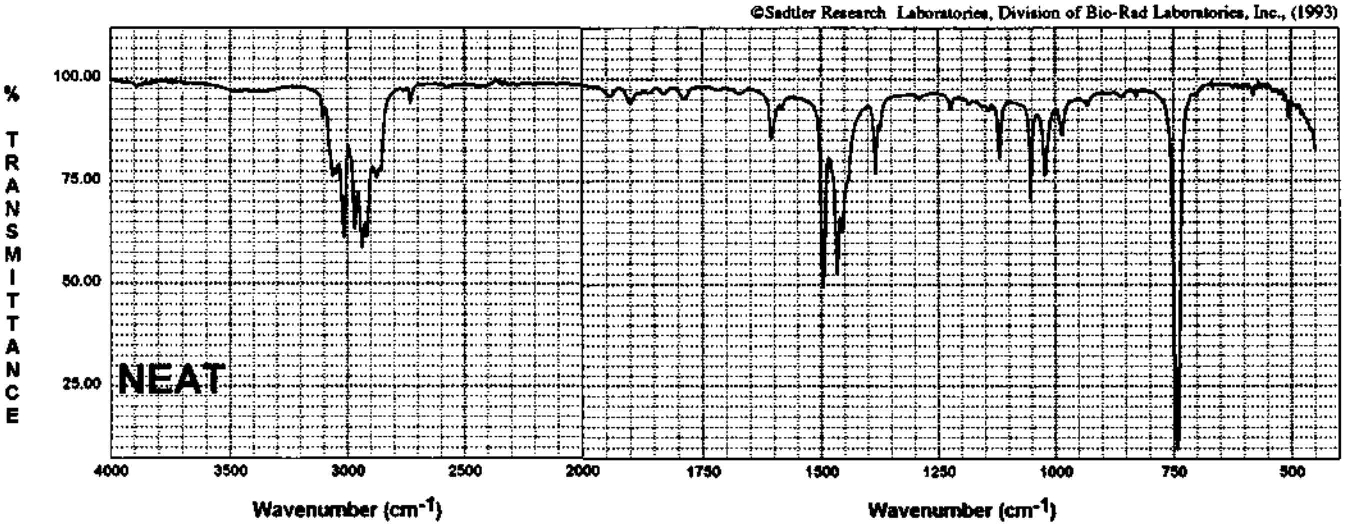

Infrared unknown spectrum *c*.

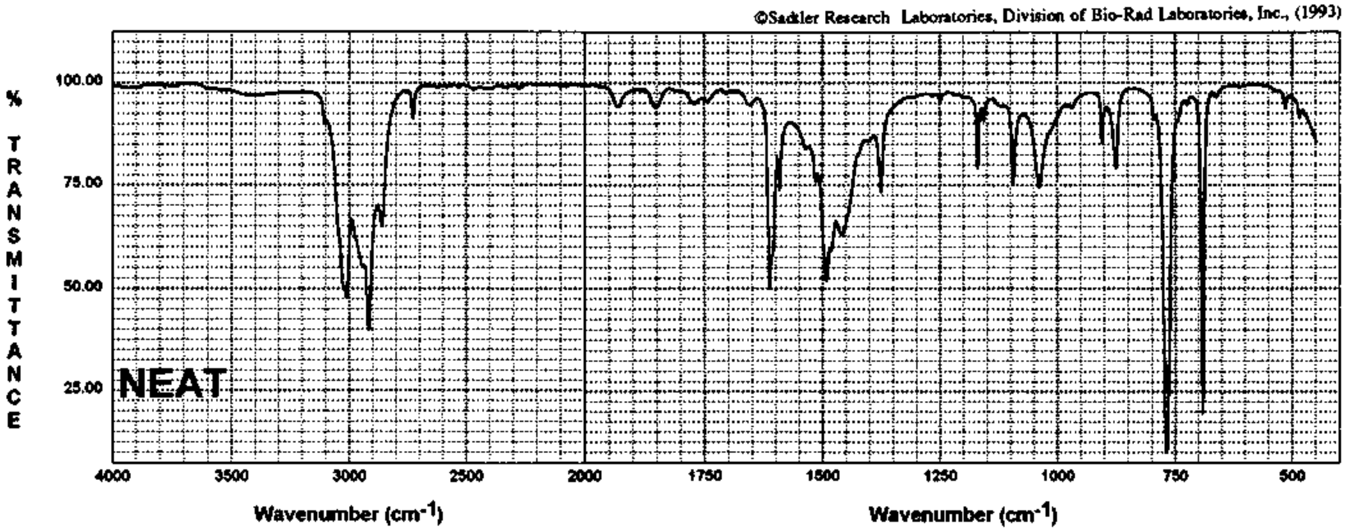

Infrared unknown spectrum *d*.

8-15. The C—H stretching mode of chloroform ($CHCl_3$), which occurs at 3022 cm^{-1}, is one of the rare exceptions to the 3000-cm^{-1} rule. What is the rule? Suggest an explanation for this exception.

BIBLIOGRAPHY

Aldrich Library of Infrared Spectra, Aldrich Chemical Co., Inc., 940 West Saint Paul Avenue, Milwaukee, WI 53233. 3rd ed., 1981, 12,000 spectra, 8 per page, in one volume arranged by chemical type.

Aldrich Library of FT-IR Spectra, Aldrich Chemical Co., Inc., 940 West Saint Paul Avenue, Milwaukee, WI 53233. 2nd ed. 1997, 18,000 spectra, 4 per page, in three volumes arranged by chemical type.

API collection (American Petroleum Institute). About 4500 spectra. M.C.A. collection (Manufacturing Chemists'Association). About 3000 spectra. Chemical Thermodynamics Property Center, Texas A&M University Department of Chemistry, College Station, TX 77843.

Coblentz Society Infrared Spectra Collections for ACS/Labs. (http://www.coblentz.org). An extensive collection of critically evaluated spectra.

Grasselli, J. G.; Ritchey, W. M., Eds., *Atlas of Spectral Data and Physical Constants for Organic Compounds;* CRC Press: 2000 Corporate Blvd. NW, Boca Raton, FL 33431. 2nd ed., 1975.

Sadtler Library. About 95,000 spectra of single compounds; about 12,000 spectra of commercial products. Sadtler Research Labs, 3316 Spring Garden Street, Philadelphia, PA 19014.

For reviews on ATR FT-IR spectroscopy, see

Grdadolnik, J. *Acta Chimica Slovenica.* **2002,** *49*, 631.

Hind, A. R.; Bhargava, S. K.: McKinnon, A. *Adv. Colloid Interface Sci.* **2001,** *93*, 91.

Milosevic, M. *Applied Spectroscopy Reviews* **2004,** *39,* 365.

For reports on the use of ATR FT-IR in the undergraduate laboratory, see

Schuttlefield, J. D.; Grassian,V. H. *J. Chem. Educ.* **2008,** *85*, 279.

Schuttlefield, J. D.; Larsen, S. C.; Grassian,V. H. *J. Chem. Educ.* **2008,** *85*, 282.

NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

Nuclear spin

Nuclear spin is an energy property intrinsic to a nucleus and analogous to the electron spin that plays such an important role in determining electron configurations. Nuclear spin values are quantized, as are electron spins, and are represented by I, the nuclear spin quantum number. Nuclear spin quantum numbers range from 0 through $\frac{7}{2}$ in increments of $\frac{1}{2}$. The nuclei of greatest interest to organic chemists, the ^{1}H, ^{13}C, ^{19}F, and ^{31}P nuclei, have spins of $\frac{1}{2}$; the ^{12}C, ^{16}O, and ^{32}S nuclei have spins of 0 (and thus cannot be observed by nuclear magnetic resonance spectroscopy, NMR); the ^{2}H (deuterium, D) and ^{14}N nuclei have spins of 1. Since any spinning charged particle (or body) produces a magnetic moment, a nucleus with a nonzero spin quantum number has a magnetic moment, μ.

Nuclear spin values are quantized because the nuclear angular momentum, and thus the nuclear magnetic moment, is quantized. When placed in an external magnetic field, nuclei orient their magnetic moments in certain ways with respect to the magnetic field, which is assumed to be aligned with the z axis of a Cartesian coordinate system. These orientations are referred to as the z components of the nuclear magnetic moment, μ_z. For a nucleus with a spin of $\frac{1}{2}$, μ_z may be $+\frac{1}{2}$ or $-\frac{1}{2}$. In general, for a nucleus of spin I, the μ_z takes quantized values from $[-I, -I+1, \ldots, I-1, I]$; or $(2I + 1)$ different values in all. For this discussion we will limit ourselves to nuclei with spin $\frac{1}{2}$, since this is easier to describe, and since most nuclei of interest in organic chemistry are of spin $\frac{1}{2}$.

When placed in a static magnetic field of strength H_0, the magnetic moment, μ_z, of the spinning nucleus precesses about the magnetic field at a frequency, ν, such that $\nu = \gamma H_0/2\pi$, where H_0 is the strength of the applied magnetic field, and γ is a characteristic property of the nucleus known as the gyromagnetic ratio. When a nucleus of spin $\frac{1}{2}$ is placed in a magnetic field, the energies of the $\mu_z = +\frac{1}{2}$ and $-\frac{1}{2}$ states are separated, since in one spin state the nuclear magnetic moment is aligned with the applied magnetic field, and in the other spin state the nuclear magnetic moment is opposed to the applied magnetic field.

The amount of separation of the two energy states, ΔE, is proportional to the magnetic field, and is given by the following expression:

$$\Delta E = \frac{h\gamma H_0}{2\pi} = h\nu$$

When nuclei in the magnetic field are exposed to radiation of the proper frequency, transitions between the two energy states are stimulated, and the nucleus is said to be in *resonance,* or to *resonate.* This transition occurs when the frequency and the energy difference are related by the Planck relation, $\Delta E = h\nu$, and thus the sample will absorb energy of frequency ν. The study of these energy changes is known as nuclear magnetic resonance, or NMR, spectroscopy.

INSTRUMENTATION

In an NMR spectrometer, the magnetic field is provided by a large permanent magnet, electromagnet, or superconducting electromagnet. Commercially available NMR spectrometers have magnets with field strengths that range from 1.4 to 16.3 tesla (the earth's magnetic field at its surface is roughly 5×10^{-5} tesla), and thus operate at frequencies from 60 to 700 MHz for protons. In general, most spectrometers with an operating frequency above 100 MHz use a superconducting electromagnet.

Traditionally, NMR spectra were acquired either by holding the applied magnetic field constant and sweeping the radio frequency (rf), or by holding the rf constant and sweeping the applied magnetic field. Energy absorption by the sample was detected, and the result was the NMR spectrum, a plot of intensity (of energy absorption) versus frequency (or field). This instrumental technique is referred to as continuous wave, or CW, spectroscopy (Fig. 8.7). Over the last 20 years, however, it has been commonly replaced by pulsed, or Fourier transform (FT), NMR spectroscopy. Among many other benefits, FT-NMR spectroscopy allows very rapid acquisition of spectral data, which permits analysis of small samples and rare nuclei, such as ^{13}C.

The basic principles of FT-NMR spectroscopy can be qualitatively explained as follows. Take, for example, an NMR spectrum that contains a single peak at a given frequency. The graph of this spectrum (Fig. 8.8) is a plot of intensity versus frequency. The same information can be conveyed by a plot of intensity versus time that shows a cosine wave at the frequency described by the graph of the usual NMR spectrum. This is shown in Figure 8.9, and for a spectrum with a single frequency, this plot of intensity versus time is almost as easy to interpret as the usual NMR spectrum shown in Figure 8.8.

Of course, it would be very difficult to determine the frequencies of many superimposed cosine waves from this kind of plot, and it would be at best awkward to interpret a complex NMR spectrum presented in such a fashion (Fig. 8.10). The use of the Fourier transform allows us mathematically to

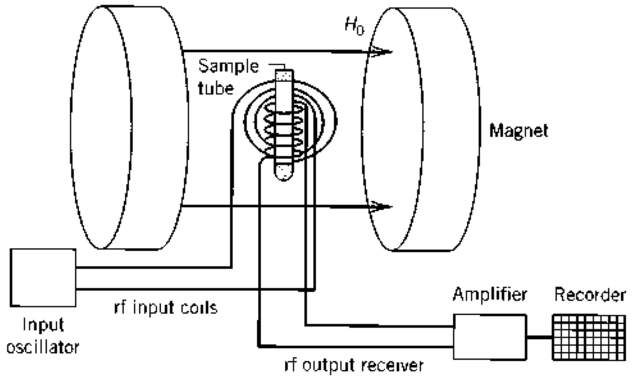
H_0
Sample tube
Magnet
Input oscillator
rf input coils
Amplifier
Recorder
rf output receiver

Figure 8.7 Schematic of NMR spectrometer. (Reprinted with permission of John Wiley & Sons, New York.)

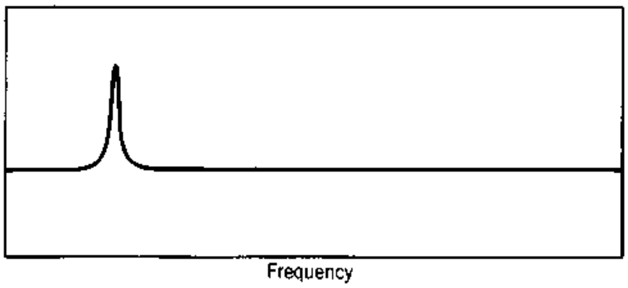

Figure 8.8 Intensity versus frequency (usual NMR spectrum).

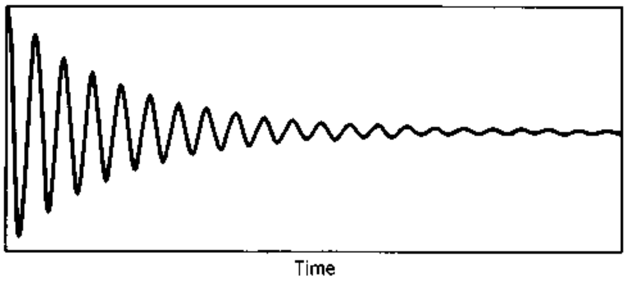

Figure 9 Intensity versus time.

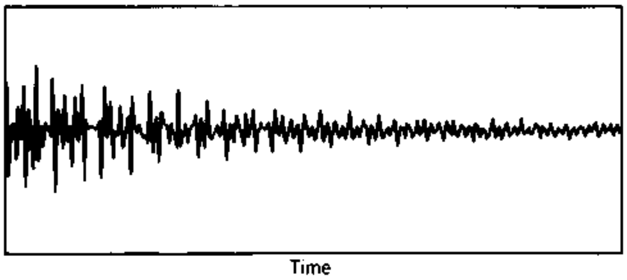

Figure 8.10 Three-signal NMR spectrum: intensity versus time.

interconvert these time domain (Fig. 8.10) and frequency domain spectra (Fig. 8.11). Fourier transform of the apparently complex spectrum in Figure 8.10 gives the spectrum in Figure 8.11. It is then easy to see that there are actually only three different resonance signals contained in the time domain of the data of Figure 8.10.

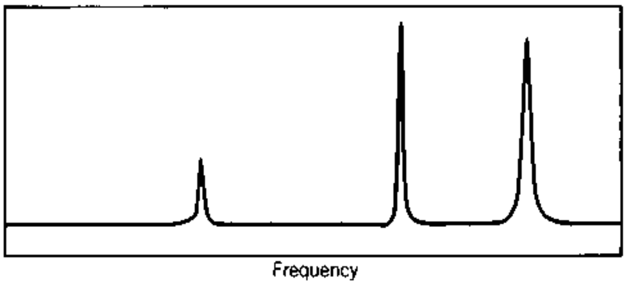

Figure 8.11 Three-signal NMR spectrum: intensity versus frequency.

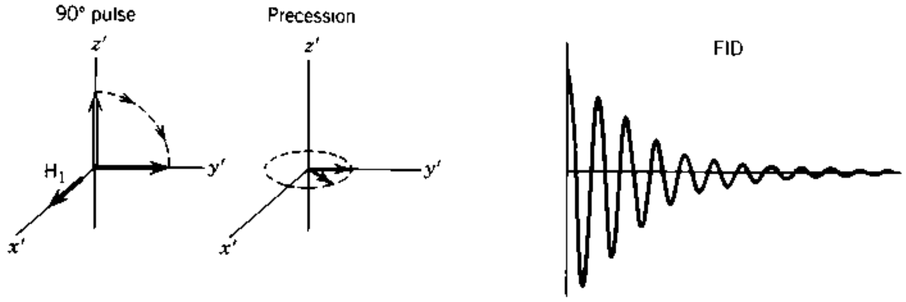

Figure 8.12 Basic pulsed NMR experiment.

Fourier transform NMR spectra are obtained by applying a short (~1−10 ms), high-powered pulse of rf energy to the sample (Fig. 8.12). This pulse affects all the nuclei to be observed. Before the pulse is applied, the equilibrium net nuclear magnetization is aligned with the applied magnetic field, along the *z* axis. The coordinate system is presumed to be rotating about the *z* axis at the frequency of the rf pulse. The pulse, applied down the *x* axis, applies a torque to the nuclear magnetic moments and rotates them into the *xy* plane. At this point the pulse is turned off and the nuclear magnetic moments return to their equilibrium alignments. In the process, they precess about the *z* axis (applied magnetic field) in the *xy* plane and induce a current in a detector coil, which can be thought of as being aligned with the *y* axis. This current varies in a sinusoidal manner, and the observed frequency will be the difference between the resonance frequency of the nuclei and the frequency of the rf pulse.

The detected signal, which is called the free induction decay (FID), is digitized and stored. For small organic molecules in a nonviscous solution, the FID will disappear after a few seconds, which corresponds to the time it takes the nuclei to regain equilibrium alignments after the rf pulse. Thus, an entire ^{1}H NMR spectrum can be obtained in approximately 2 s, in contrast to the 10–15 min usually needed to obtain a CW spectrum. A major advantage of FT-NMR is that many spectra of a sample can be rapidly obtained and added together to increase the signal-to-noise (*S/N*) ratio. Noise is presumably random about some zero level, so when many spectra are added together, the noise level is reduced, while real signals are reinforced when added. The *S/N* ratio is proportional to the square root of the number of spectra added together. Thus, one can obtain the ^{1}H spectrum of a 10-μmol sample in a minute or two. With FT-NMR it is possible to obtain spectra of isotopes that are insensitive and/or of low natural abundance, as well as spectra of large biological molecules in dilute solution. By adding a few hundred spectra together, an adequate ^{13}C NMR spectrum of a 100-μmol sample can be obtained in about 20–30 min.

CHEMICAL SHIFT

In a molecule, the magnetic field at a nucleus depends not only on H_0, the field generated by the instrument (the external field), but also on the magnetic fields associated with the electron density near the nucleus. Electrons are influenced by the external field in such a way that their motion generates a small magnetic field that opposes the applied field, and reduces the actual field experienced at the nucleus. This reduction is very small (relative to the external field) and is on the order of 0.001%, 10 ppm, for most protons, and about 200 ppm for ^{13}C nuclei. Reduction of the external field is known as *shielding*, and it gives rise to differences in the energy separation for nuclei in different electronic environments in a molecule. The differences in the energy separation are known as *chemical* shifts.

The magnitude of the chemical shift depends on the nature of the valence and inner electrons of the nucleus and even on electrons that are not directly associated with the nucleus. Chemical shifts are influenced by inductive effects, which reduce the electron density near the nucleus and reduce the shielding. The orientation of the nucleus relative to π electrons also plays an important role in determining the chemical shift. A proton located immediately outside a π-electron system (as in the case of the protons on benzene rings) will be significantly deshielded. In most molecules the chemical shift is determined by a combination of these factors. Chemical shifts are difficult to predict using theoretical principles, but have been well studied and can usually be easily predicted empirically upon comparison to reference data.

In an NMR spectrum, the absorption of rf energy is detected, as in Figure 8.13, where the energy absorption is shown for increasing frequency. In this example we illustrate the case with two different nuclei, A and X. Since A and X are different, they absorb energy at different frequencies while in the same applied magnetic field.

The spectrum would be displayed as in Figure 8.14. The difference in the resonances is known as a *chemical shift* and is expressed in parts per million (ppm). The use of frequency units is cumbersome and is complicated because

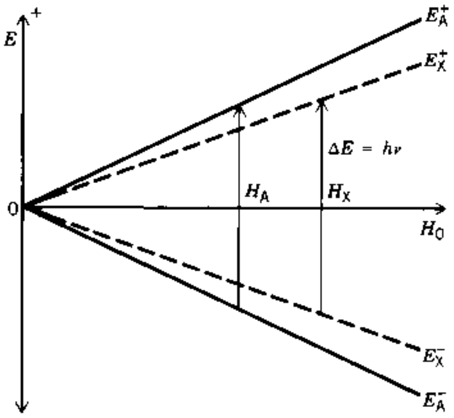

Figure 8.13 The energy splitting for two chemically different protons. The differences between the *A* energy levels (solid lines) and the *X* levels (dashed lines) have been amplified for illustrative purposes. At 60 MHz, nucleus *A* absorbs energy at field H_A and nucleus *X* absorbs energy at field H_X. Nucleus *X* is said to be more strongly shielded than *A*. The resonance for *X* is said to occur upfield of that for *A*.

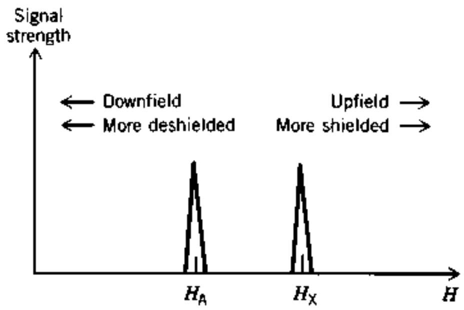

Figure 8.14 The spectrum for the system in Figure 8.13 as it would be displayed. It is conventional to display the spectrum with magnetic field strength increasing to the right so that upfield (and more strongly shielded) is toward the right and downfield (and deshielded) is toward the left.

NMR spectrometers of different magnetic field strengths (and thus operating frequencies) are used. The use of ppm units allows direct comparisons of spectroscopic data obtained on different instruments, when chemical shifts are referenced relative to a reference compound whose chemical shift is arbitrarily defined as 0 ppm. The accepted reference standard for ^1H and ^{13}C NMR in organic solutions is tetramethylsilane (TMS), $(CH_3)_4Si$. The chemical shift relative to TMS is symbolized by δ.

Tetramethylsilane is used as a reference substance for a number of reasons. It is more strongly shielded (Si is more electropositive than C) than most other protons and carbon atoms, and its resonance is thus well removed from other areas of interest in the NMR spectrum. Tetramethylsilane is inert and thus unlikely to react with the compound being analyzed, it is volatile (bp 26 °C) and thus easily removed after a sample has been analyzed, and its 12 identical protons per molecule provide a strong signal per molecule of TMS.

SPIN–SPIN COUPLING

In a molecule with several protons, the exact frequency at which a proton resonates depends not only on the chemical shift of that proton, but also on the spin states of nearby protons. This occurs because the magnetic moments of the nearby protons can either shield or deshield the proton in question from the applied magnetic field, depending on the orientation of the nearby magnetic moments relative to the applied magnetic field. The extent of this perturbation is independent of the applied magnetic field strength. The effect of the spin state of one nucleus on the resonance of another is known as *coupling* or *splitting.*

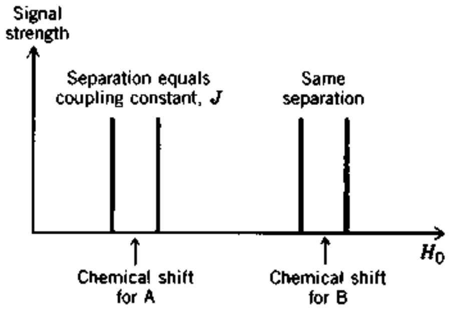

Figure 8.15 Spectrum of two chemically different protons that are coupled.

The spectra resulting from spin–spin coupling depend on the types of nuclei, the distance and geometry between the nuclei, the nature of the bonding, the electronic environment, and the total number of spin states possible. The latter may be illustrated by looking at the spectrum of an imaginary compound that has protons H_A and H_X on adjacent carbons, connected by three bonds: H_A—C—C—H_X (Fig. 8.15). In the first approximation we would expect one resonance for H_A and one resonance for H_X, and the spectrum would resemble that shown in Figure 8.14. In the presence of coupling, the resonance for H_A splits into two signals, one of which corresponds to H_X, having $\mu_z = +\frac{1}{2}$ and the other to $\mu_z = -\frac{1}{2}$. The coupling effect is symmetric in that the H_X resonance also splits into two resonances, one for each spin state of H_A. The magnitude of the separation of the H_A pair (a doublet) or the H_X pair (also a doublet) is known as the coupling constant, or *J.* It is usually expressed in frequency units (Hz), since *J* is independent of the magnetic field strength.

A simple way to explain this is to consider the effect the two possible spin states of H_X have on the resonance frequency of H_A. The equilibrium population distribution of the two spin states in H_X is very close to 1:1, since ΔE is only about 10^{-3} cal/mol. Since H_X has a magnetic moment, there are then two slightly different magnetic fields at H_A. We thus see two signals for H_A, one for those H_A nuclei adjacent to H_X nuclei with μ_z aligned with the applied magnetic field, and one signal for those H_A nuclei adjacent to H_X nuclei with μ_z aligned opposed to the applied field. Coupling between protons connected by more than three bonds does occur, but its magnitude, *J*, is usually small and often not directly observed in a usual NMR spectrum.

The splitting becomes more interesting when there are several nuclei of one type. 1,1-Dibromoethane (CH_3CHBr_2) has three equivalent protons in the

Individual μ_z	**SUM**
(+)(+)(+)	$+\frac{3}{2}$
(+)(+)(−) or (+)(−)(+) or (−)(+)(+)	$+\frac{1}{2}$
(+)(−)(−) or (−)(+)(−) or (−)(−)(+)	$-\frac{1}{2}$
(−)(−)(−)	$-\frac{3}{2}$

Figure 8.16 Possible combinations of spin states for a methyl group.

methyl group and one proton on the C-1 atom. The methyl group exhibits rapid internal rotation so that its three protons are equivalent. The chemical shift for the C-1 proton is 5.86 ppm and that for the methyl protons is 2.47 ppm. Here we can see an example of decreased shielding resulting from the presence of electronegative substituents. Equivalent protons do not couple with one another (this is an important rule in interpreting spectra), but the methyl protons will affect the proton on C-1, and vice versa.

To analyze the splitting pattern, we need to consider the orientations of the nuclear magnetic moments, with respect to the applied magnetic field, for all three methyl protons. Since each of the three protons may have two spin states that are of nearly equal probability, there are $2^3 = 8$ possible combinations of spin states in all for the methyl protons. The net sums of these may have only four different values, as shown in Figure 8.16. The symbol (+) is used to represent $\mu_z = +\frac{1}{2}$ for a single proton and (−) is used to represent $\mu_z = -\frac{1}{2}$. Thus (+)(+)(−) means that protons 1 and 2 have $\mu_z = +\frac{1}{2}$, while proton 3 has $\mu_z = -\frac{1}{2}$.

The number of different μ_z states is $(2N + 1)$, where N is the number of equivalent nuclei (of spin $\frac{1}{2}$) Thus the three methyl protons can generate four slightly different magnetic fields, and the proton on the C-1 of CH_3CHBr_2 sees (in different molecules) four different magnetic fields. Since these different magnetic fields are not of equal probability, but rather are populated in a ratio of 1:3:3:1, the four signals we see for the proton on C-1 when coupled to the methyl group are of intensities 1:3:3:1, and are referred to as a *quartet.* This is shown schematically in Figure 8.17.

Since the proton on C-1 has two possible spin states of nearly equal probability, the protons of the methyl group experience two slightly different magnetic fields and are observed in the spectrum as two slightly separated signals of equal intensity, or a *doublet.* The separation between each of the C-1 proton signals is the coupling constant, *J*, and will equal the *J* of the methyl signal. The proposed spectrum is shown in Figure 8.17*b.* The coupling constant in this case is about 7 Hz. The 60-MHz NMR spectrum of 1,1-dibromoethane is shown in Figure 8.17*c*.

NOTE. The net effect of spin–spin coupling is that a proton (or group of equivalent protons) adjacent to N *other protons will be observed as a multiplet with* (N + *1) lines.*

A proton, or group of equivalent protons, may be coupled to more than one group of nuclei. The spectrum of 1-nitropropane ($CH_3CH_2CH_2NO_2$) is

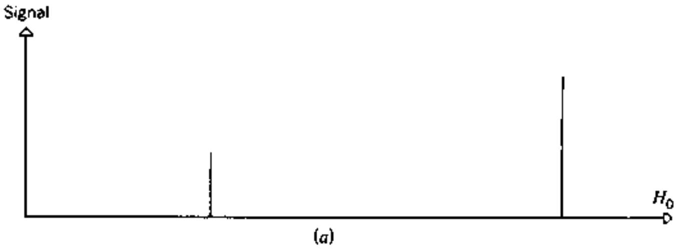

Figure 8.17*a* **The spectrum without any spin–spin coupling.**

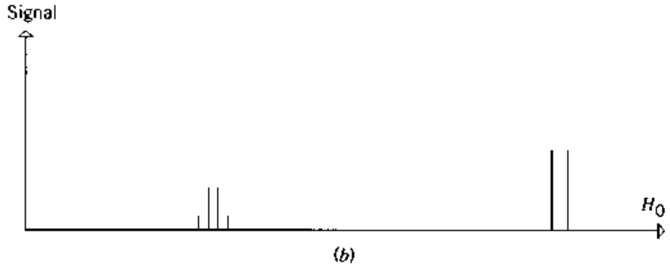

Figure 8.17*b* **A "stick figure" spectrum indicating the expected intensities.**

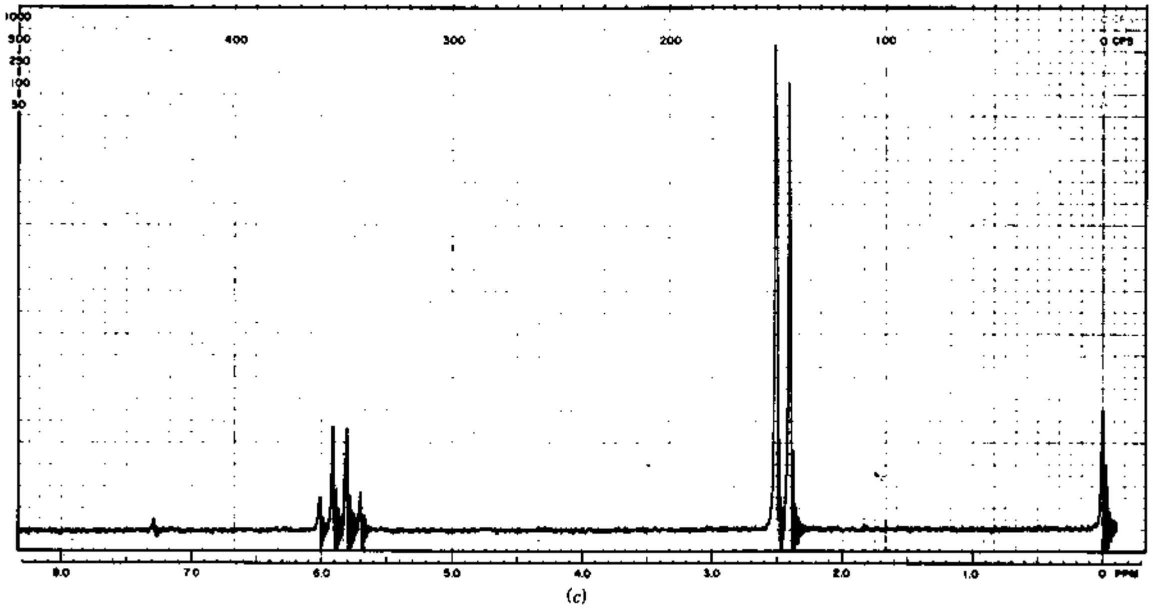

Figure 8.17*c* **The actual 60-MHz spectrum of 1,1-dibromoethane. The TMS signal at 0 ppm is seen as well as a weak signal at 7.3 ppm, which is not from this molecule.**

shown in Figure 8.18. The signal from the central methylene (CH_2) group is seen at about 2.0 ppm. Because the methylene group is adjacent to (and thus coupled to) five protons, its signal is a (5 + 1) or six-line multiplet—a *sextet.*

Nuclei with spins of 1 or greater exhibit more complex spin–spin coupling, since they can exist in more than two different spin states. For a nucleus coupled to *N* nuclei of spin *I*, a multiplet of 2*IN* lines will be observed. Nuclei of spin zero do not couple.

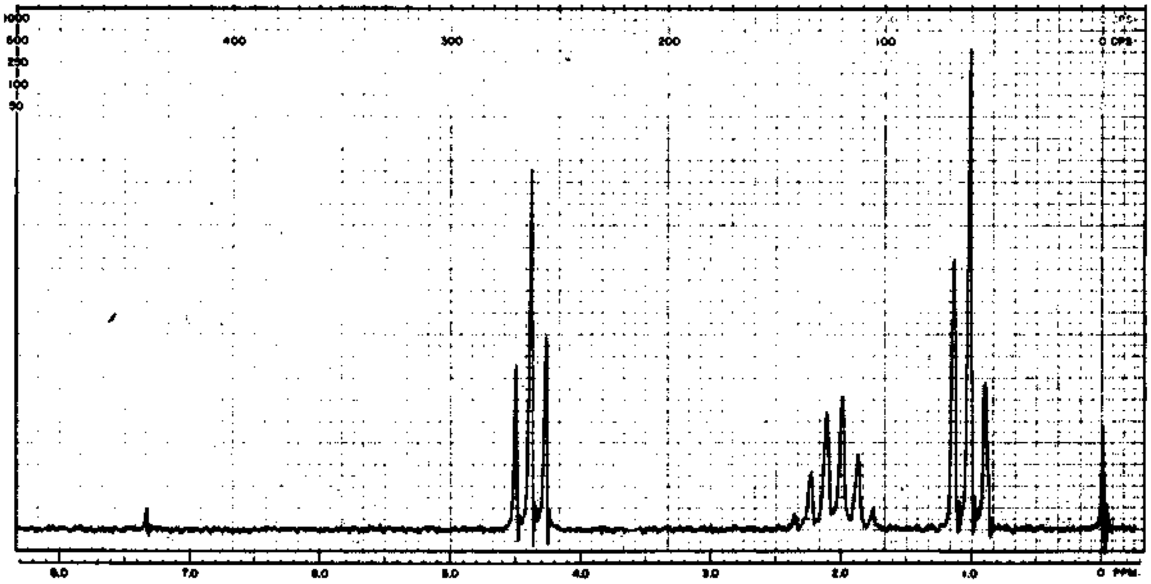

Figure 8.18 The 60-MHz spectrum of 1-nitropropane. (Courtesy of Varian Associates, Palo Alto, CA.) **Starting from the right, the TMS signal at 0 ppm is seen. Next is a 1:2:1 triplet at 1.03 ppm. This triplet results from the protons on C-3 and their coupling with the two protons on C-2. Next is the sextet centered at 2.07 ppm. This multiplet is from the protons on C-2 and their coupling with the protons on C-1 and C-3. Finally, we have the signal from the protons closest to the nitro group centered at 4.38 ppm. These protons appear as a 1:2:1 triplet due to their coupling with the protons on C-2.**

INTENSITIES

The area under an NMR peak is proportional to the number of nuclei giving rise to that signal. The intensity of a resonance is thus best determined by the *integral* of the NMR spectrum over a resonance, or group of resonances. Nuclear magnetic resonance spectrometers can measure the integral, though integration data from an FT spectrometer are less reliable than those from a CW spectrometer. In more complex spectra the intensities are useful as a measure of the number of protons of a given type. For instance, in the above case the integral over both peaks of the methyl group doublet will be three times the integral over the quartet of the proton on C-1. Integration can thus often provide useful information for determining the identity of a compound.

SECOND-ORDER EFFECTS

So far, all of our examples have consisted of first-order spectra. First-order spectra are those multiplets interpretable through elementary coupling analysis, such as that above; second-order spectra are those that are not interpretable in this manner. These highly symmetric and fairly simple first-order spectra are generally observed when the chemical-shift differences (expressed as a frequency) are much greater than the coupling constant. Second-order effects occur when the coupling constants become comparable to or greater than the chemical-shift differences. Thus, spectra obtained on instruments with higher magnetic fields are more likely to be first order, since the frequency differences between given signals increase with increasing magnetic fields. However, the chemical-shift differences (in ppm) remain the same regardless of magnetic field strength.

Second-order effects may be understood in qualitative terms by considering the limiting cases. Let us consider the hypothetical disubstituted ethylene shown in Figure 8.19, where R and M are substituents that might be identical or may have very different effects on the alkenyl protons. In Figure 8.19*a* the spectrum is shown for the situation in which R and M have very different effects. In this case we will observe a first-order spectrum consisting of two doublets. The coupling constant is the separation in the doublets, and the chemical shift of each nucleus is the geometric midpoint of each doublet.

In Figure 8.19*b*, groups R and M are identical. H_A and H_B are identical in this case and only a single resonance is observed (coupling between equivalent nuclei is not observed).

In Figure 8.19*c* the difference in the chemical environment of H_A and H_B is very slight. The spectrum shown may be seen as intermediate between the limiting cases in Figure 8.19*a* and Figure 8.19*b*. Note that there is a "leaning in" of the doublets as the central members increase in intensity at the expense of the outer members. A full continuum of behavior may be expected with cases observed in which the outer members are lost in the noise and the central members take the appearance of a doublet. This would be one example of a class of spectra known as "deceptively simple spectra."

The second-order spectra of systems with more than two protons are difficult to describe even in qualitative terms. Second-order spectra may well display more lines than one would predict from simple coupling theory. Also, in second-order spectra, the coupling constants and the chemical-shift differences may not be obtainable as simple differences in the positions of spectral lines. Thus, spectra obtained at high frequencies (and magnetic fields) are often more useful. As the operating frequency of the instrument is increased, the chemical-shift differences (in frequency terms) increase while the spin–spin coupling remains constant. Thus, the complicating second-order effects are likely to be less noticeable in high-field spectra. The reader is referred to more extensive treatments of NMR for a discussion of second-order cases.

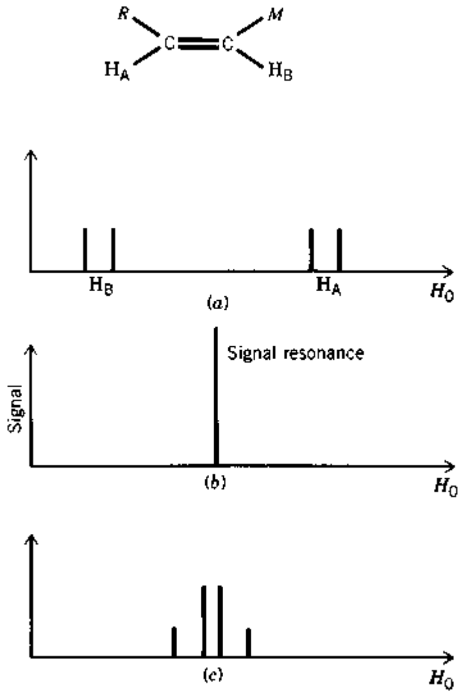

Figure 8.19 Second-order effects. (a) The chemical-shift difference is much larger than the coupling constant and a first-order spectrum is observed. (b) Protons A and B are equivalent and a single resonance is observed. (c) The chemical-shift difference is of the same order of magnitude or less than the coupling constant. Note the "leaning in" of the peak intensities in this spectrum relative to that in part *a*.

INTERPRETATION OF ^{1}H NMR SPECTRA

The first issues that must be addressed are molecular symmetry and the magnetic equivalence or nonequivalence of protons or other functional groups. Even if two protons, or groups, are chemically equivalent, they may or may not be magnetically equivalent. Although molecular symmetry can often simplify NMR spectra, one must be able to discern which protons or groups are equivalent by symmetry. The two most useful symmetry properties (or symmetry operators) are the plane of symmetry and the axis of symmetry.

A plane of symmetry is simply a mirror plane such that one half of the molecule is the mirror image of the other half, as in *meso*-pentane-2,4-diol:

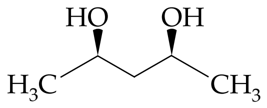

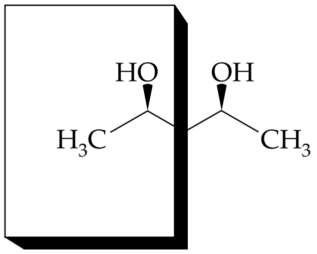

meso-Pentane-2,4-diol

The methyl groups are identical by symmetry, and one would expect this stereoisomer to show one methyl doublet in its ^{1}H NMR spectrum and one signal for a methyl group in its ^{13}C spectrum.

Consider the other diastereomer of pentane-2,4-diol, the chiral *d,l* isomer. This isomer has an axis of symmetry. If the molecule is rotated 180° about an axis in the plane of the paper passing through the central carbon, the molecule can be converted into itself:

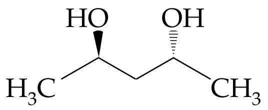

d,l-Pentane-2,4-diol

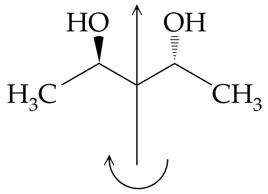

Here, too, the methyl groups are identical by symmetry, and one would expect this stereoisomer to show one methyl doublet in its ^{1}H NMR spectrum and one signal for a methyl group in its ^{13}C spectrum.

It is, however, relatively simple to use NMR spectroscopy to distinguish between these stereoisomers. To do this, look at the two methylene protons on the central carbon, C-3, of each isomer:

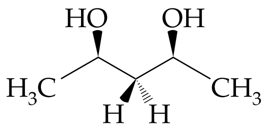

The plane of symmetry in the meso isomer bisects each of the two protons. In the *d,l* isomer, the axis of symmetry interconverts the two protons. Thus, in the *d,l* isomer, the two methylene protons are equivalent by symmetry, but they are not equivalent in the meso isomer. This can also be seen by inspecting the molecule. On the left, one H is syn to both —OH groups and the other is anti to both —OH groups. On the right, each H atom is syn to one —OH group and anti to the other.

The more rigorous way to determine equivalence or nonequivalence is to determine whether the two protons (or groups) are homotopic (identical), diastereotopic, or enantiotopic. To compare two protons, we use the usual Cahn–Ingold–Prelog system for the nomenclature of stereoisomers. We artificially distinguish the relative priority of two protons by a method such as drawing them in different colors or pretending that one is deuterium (as long as the molecule does not contain D). We draw the two possibilities (i.e., the first H as D and then the second H as D) and then determine the stereochemical relationship between the two:

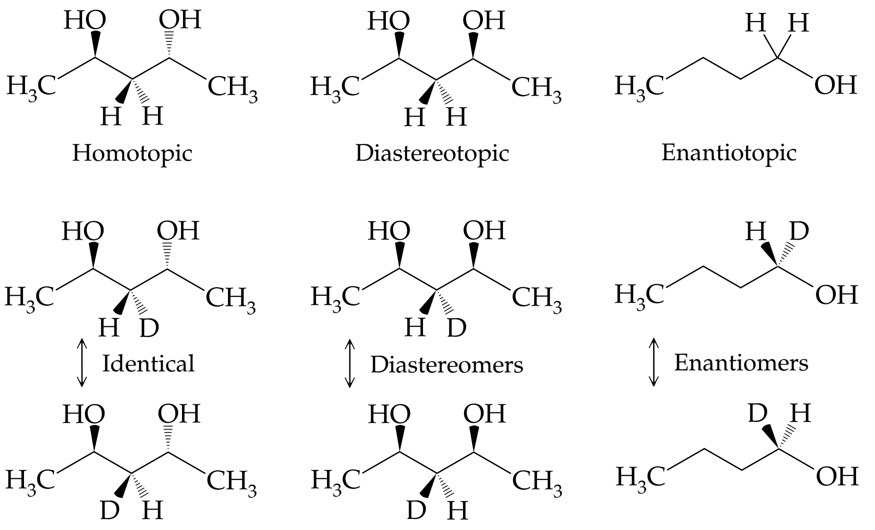

If the two are identical, the two protons are identical, or *homotopic*. If the two structures are diastereomers, the two protons are *diastereotopic,* and if the two structures are enantiomers, the two protons are *enantiotopic.* Diastereotopic protons, or groups, will be magnetically nonequivalent. Enantiotopic protons, or groups, will be magnetically equivalent only in an achiral environment and may appear nonequivalent in a chiral environment, such as a chiral solvent or in a biological sample. Homotopic protons may or may not be magnetically equivalent. Of course, it is possible for magnetically nonequivalent signals to be so close to one another in the NMR spectrum as to overlap (accidentally degenerate).

Homotopic protons may be magnetically nonequivalent if the two protons have different coupling constants to the same third proton. The most common example of this occurs in para-substituted benzenes:

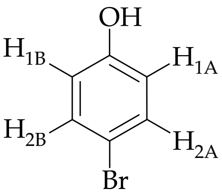

By symmetry, H_{1A} and H_{1B} are equivalent. These protons are not, however, magnetically equivalent because H_{1A} and H_{1B} have different coupling constants to, for example, H_{2A}, and the spectrum of this molecule may well be more complex than one would at first expect.

The equivalence or nonequivalence of functional groups, as well as protons, can easily be determined. The ^{1}H NMR spectrum of menthol shows three methyl doublets, since the two methyls in the isopropyl group are diastereotopic. The ^{13}C spectrum of menthol shows three distinct resonances for the three different methyl groups:

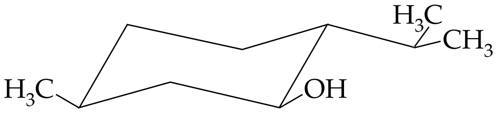

Menthol

^{1}H CHEMICAL SHIFTS

Figure 8.20 summarizes the chemical shifts of protons in a large range of chemical environments. It is, however, a bit dangerous to use figures such as

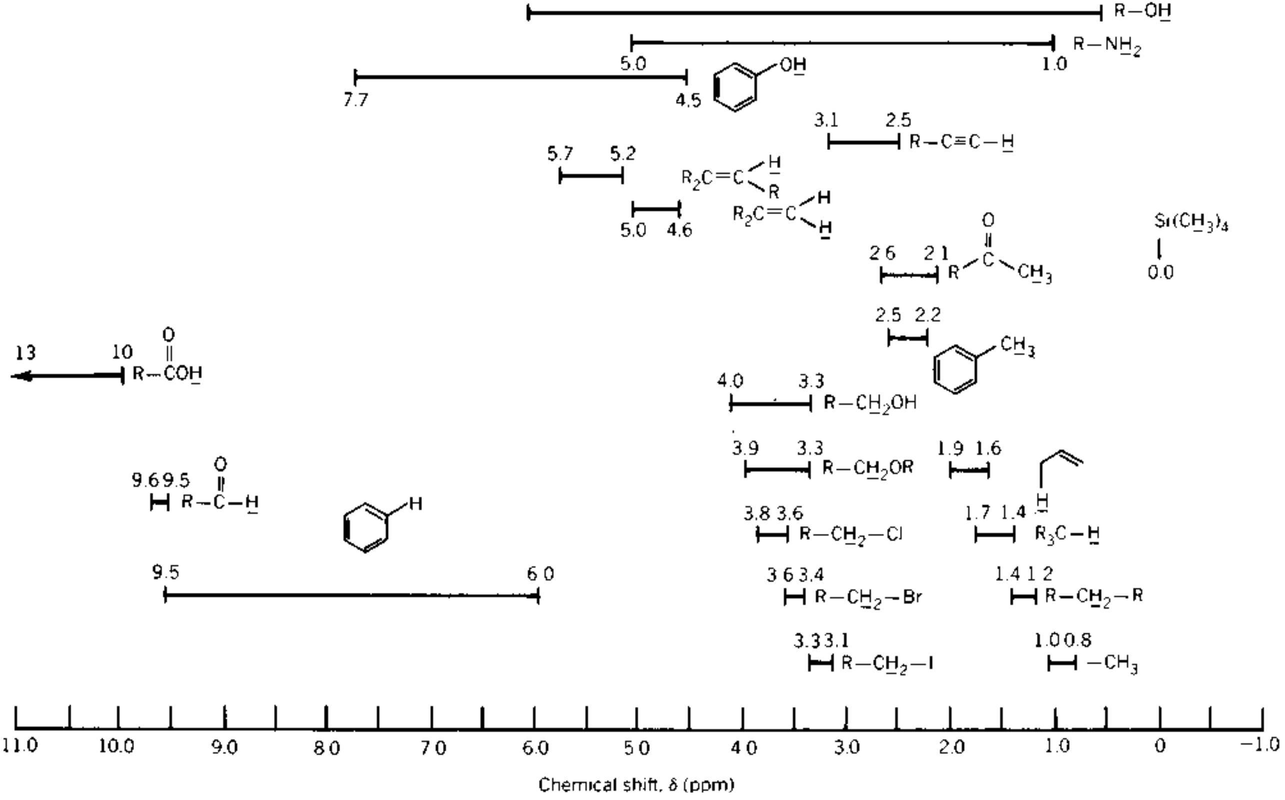

Figure 8.20 NMR ^{1}H chemical shifts. (From Zubrick, J. W. *The Organic Lab Survival Manual,* 7th ed.; Wiley: New York, 2008. Reprinted by permission of John Wiley & Sons, New York.)

this one without understanding some of the factors that underlie shielding and the chemical shift. To give some flavor of the factors that determine chemical shifts and the range of values observed, we will briefly examine chemical shifts in methyl groups and chemical shifts for protons on sp^2 carbon atoms.

Methyl groups bonded to an sp^3 carbon generally have chemical shifts in the range 0.8–2.1 ppm as long as there is no more than one electron-withdrawing group attached to the carbon. The shifts generally increase as the strength of the electron withdrawing group increases, or as more electron-withdrawing groups are added. Groups that inductively withdraw electrons reduce the electron density near the methyl group protons. This results in less shielding and a downfield shift of the methyl resonance. This effect is clearly seen in the spectra of 1-nitropropane and 1,1-dibromoethane (Figure 8.17*c* and Figure 8.18), respectively. The chemical shifts for methyl groups bonded to sp^2 carbon atoms fall in the range 1.6–2.7 ppm.

In the case of a proton bonded to an sp^2 carbon, the location of the proton relative to the π cloud plays an important role in determining the chemical shift. In unconjugated alkenes the chemical shifts fall in the range 5–6 ppm. Where more than one proton is bonded to an alkene, complex second-order spectra can be expected at low operating frequencies since the coupling constants are usually fairly large relative to the difference in resonance frequencies. In aldehydes, RC**H**O, the increased electronegativity of the oxygen increases the deshielding and the chemical shift falls in the range 9–10.5 ppm.

The chemical shift in an aromatic system is generally greater than that for alkenes. For example, the chemical shift of benzene is 7.37 ppm, which is substantially greater than the 5.6 ppm for the alkenyl protons of cyclohexene. Much of this difference results from the "ring current" effect and the orientation of the proton relative to the aromatic π electrons. If the ring substituents are not strongly electron withdrawing or electron donating, such as alkyl groups, the chemical shift for ring protons will not be shifted greatly from that of benzene itself. Furthermore, these substituents generate only small chemical-shift differences among the ring protons. Thus, the 60-MHz spectra for toluene (methylbenzene) appears to have a single resonance in the aromatic region at about 7.1 ppm. If, on the other hand, the substituents are electron withdrawing, the ortho and para ring protons will be somewhat deshielded relative to benzene. Pi-electron-donating substituents, such as a methoxy group, will increase the shielding of groups ortho and para to it.

SPIN–SPIN COUPLING

Coupling information is the primary reason that ^1H NMR is such a powerful tool for organic structure determination. Since coupling information is transmitted through bonds, coupling provides information about nearby protons and can often be used to deduce stereochemistry.

The sign of the coupling constant (usually symbolized as J) may be positive or negative. However, first-order spectra are not sensitive to the sign of the coupling constant. In second-order cases, the sign of J may be determined by a detailed analysis of the spectrum, though the sign of J is generally of little value for organic structure determination.

Geminal Coupling

Nonequivalent protons attached to the same carbon (geminal protons) will couple with one another. These coupling constants tend to be large (>10 Hz) for sp^3 carbon atoms and small (<4 Hz) for sp^2 carbon atoms. Geminal coupling constants tend to decrease with decreasing ring size, because of hybridization changes at carbon, and with the increasing electronegativity of the substituents on a given methylene group.

Vicinal Coupling

Vicinal coupling describes the coupling over three bonds observed between protons attached to two bonded carbon atoms, H—C—C—H. Vicinal coupling constants (*J* values) can range from near 0 to greater than 15 Hz, depending on the stereochemical relationship (dihedral angle) between the coupled protons, the hybridization of the carbon atoms, and the electronegativity of other substituents. For vicinal protons on sp^3 carbon atoms, the coupling constant is related to the dihedral angle and is expressed graphically by the Karplus curve (Fig. 8.21).

Though the magnitude of vicinal coupling is very sensitive to the angle of rotation about the central bond, in many simple cases nearly all coupling constants are equal. This situation is often the case if internal rotation about a C—C single bond can occur on a time scale that is very short relative to the NMR time scale, such as in acyclic systems. In these cases, the effect of internal rotation is completely blurred as far as NMR is concerned, and only an average coupling constant is observed. Vicinal coupling constants in freely rotating alkyl groups are usually observed in the 6.5- to 8-Hz range.

When the central C—C bond between two coupled protons is a double bond, rotation is restricted and separate coupling constants for cis and trans protons may be observed. Cis coupling constants fall in the range 5–12 Hz, whereas trans coupling constants range from 12 to 20 Hz. As a result of these large coupling constants, second-order effects are often observed in substituted alkenes in instruments of lower field strengths.

When rotation about carbon–carbon single bonds is restricted, or when stereochemistry dictates significant conformational preferences, nonaveraged coupling constants may be observed that can complicate the appearance of the NMR spectrum. For example, two diastereotopic protons of a methylene (CH_2) group may well each couple to a given third proton with different coupling constants. The familiar coupling explained at the elementary level suggests that if one proton is adjacent to, for example, two others, the NMR signal of the first proton will be a triplet. This simplification will be true only if the two coupling constants are identical. A triplet is merely a doublet of doublets with equal coupling constants, which gives rise to the familiar triplet with intensities of 1:2:1 (Fig. 8.22*a*). A doublet of doublets, on the other hand, gives rise to a four-line multiplet with peaks of roughly equal intensity (Figure 8.22*b*).

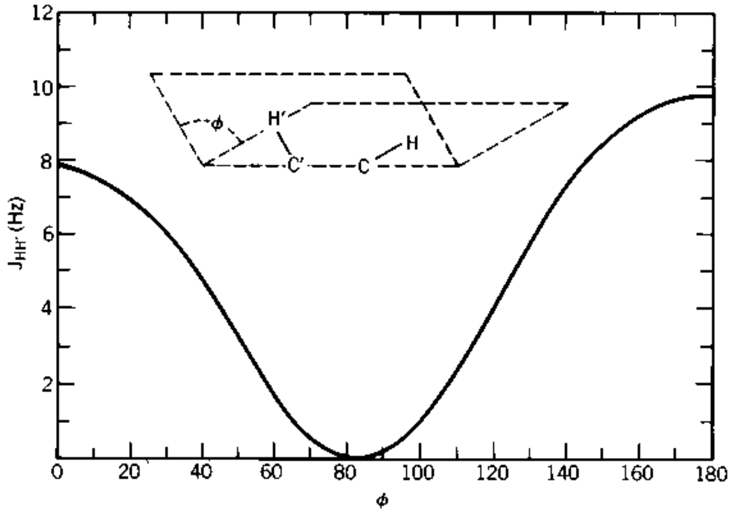

Figure 8.21 The vicinal Karplus correlation showing the relationship between dihedral angle and coupling constants for vicinal protons.

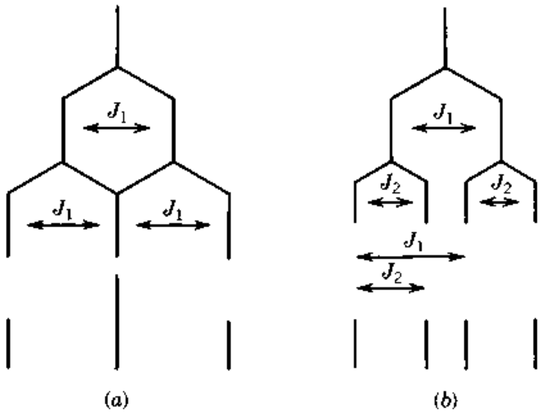

Figure 8.22 **(a) Triplet equals doublet of doublets with equal *J* values; (b) doublet of doublets.**

Long-Range Coupling

Longer range coupling involving four or more bonds is common in allylic systems and in aromatic rings and other conjugated π systems. These coupling constants are generally smaller than the values considered above (i.e., <3 Hz).

EXAMPLES OF COMPLEX, YET FIRST-ORDER, COUPLING

Ethyl Vinyl Ether

The coupling constants of even a seemingly complex multiplet can be discerned in a relatively simple manner. First, the total width (outside peak to outside peak) of a first-order multiplet is equal to the sum of all the coupling constants, keeping in mind that, for example, a triplet of $J = 7$ Hz is really a doublet of doublets with both J values equal to 7 Hz. The expansion of the proton spectrum of ethyl vinyl ether is presented as an example in Figure 8.23. Integration data are displayed between the spectrum and the horizontal axis in Figure 8.23*a*.

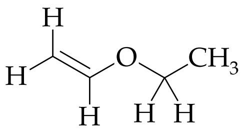

Ethyl vinyl ether

Consider the multiplet centered at 6.45 ppm (Fig. 8.23*b*). By measuring the distance (in Hz) from either outside line to the next inner line, which is 6.8 Hz, the first coupling constant is determined. Then, by measuring from the outside line to the second line in, the second coupling constant is found to be 14.4 Hz. We know that this is the last coupling constant to be found for several reasons. First, if we measure from the outside line to the third line in, we get a value, 21.2 Hz, which is equal to the sum of the previously determined coupling constants. Second, the width of the multiplet (the same measurement in this simple case) is equal to our two coupling constants. Thus, the NMR signal at 6.45 ppm is a doublet of doublets with $J = 14.4$ and 6.8 Hz.

The two doublets of doublets at 4.15 and 3.96 ppm (Fig. 8.23*c*) must be coupled to one another because they both have the coupling constant of 1.9 Hz in common. This geminal coupling constant is typical of the terminal methylene of an alkene.

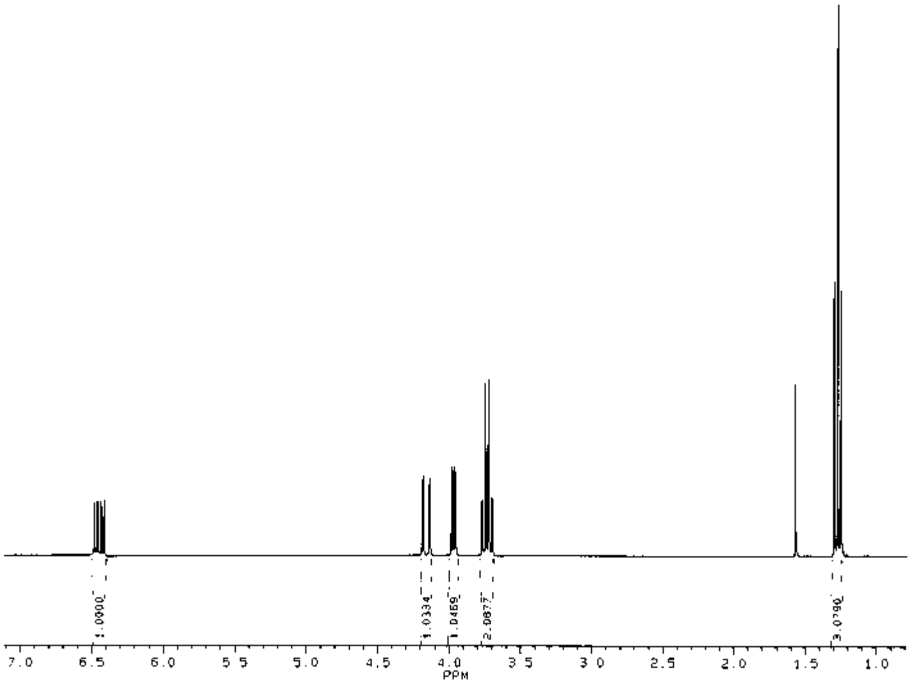

Figure 8.23*a* **Ethyl vinyl ether in CDCl$_3$.**

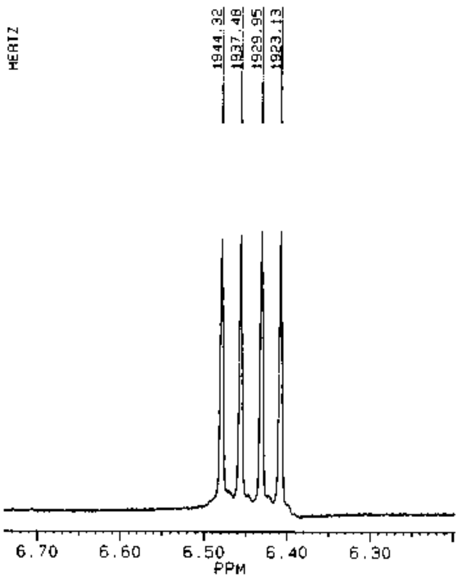

Figure 8.23*b* **Ethyl vinyl ether in CDCl$_3$ (expansion).**

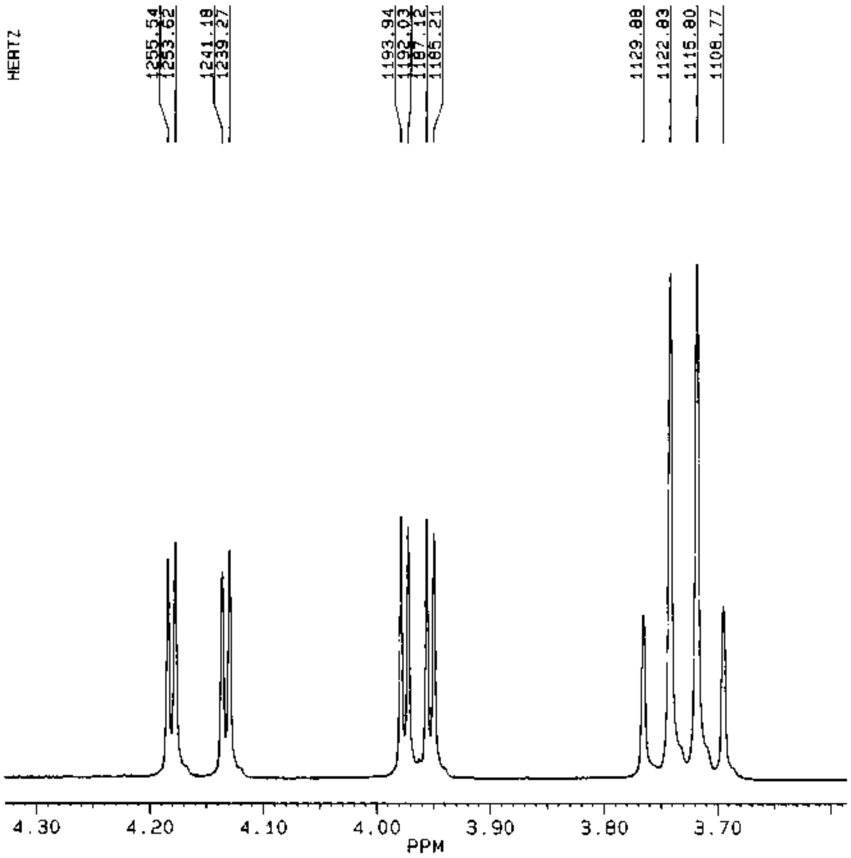

Figure 8.23*c* **Ethyl vinyl ether in CDCl$_3$ (expansion).**

Since the proton at 3.96 is coupled to the proton at 6.45 ppm by $J = 6.8$ Hz, and the proton at 4.15 ppm is coupled to the one at 6.45 ppm by $J = 14.4$ Hz, the proton at 3.96 ppm must be cis, and the proton at 4.15 ppm must be trans, to the alkene proton at 6.45 ppm.

The simple coupling observed for the ethyl group in ethyl vinyl ether can be readily assigned. The triplet at about 1.25 ppm, which integrates for three protons, is due to the methyl group; it is a triplet because the equivalent protons of the methyl group are coupled to the two protons on the adjacent carbon with equal coupling constants. The O—C$\mathbf{H}_2$ protons are observed in the NMR spectrum as the quartet at about 3.75 ppm; they are a quartet because they are coupled equally to the three equivalent protons of the methyl group:

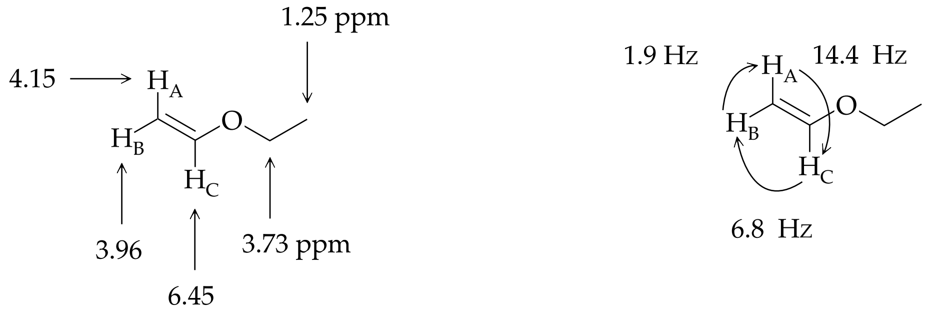

Chemical shifts Ethyl vinyl ether Coupling constants

Allyl Acetate

For a more complex example, refer to the ^{1}H NMR spectrum of allyl acetate (the NMR signal for the methyl group has been omitted) in Figure 8.24*a*.

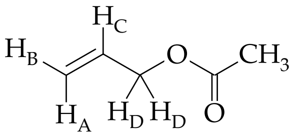

Protons A, B, and C are all chemically distinct, and the two protons labeled D are equivalent to one another by symmetry (the plane of the paper). The multiplet at 4.58 ppm (Fig. 8.24*d*) corresponds to H_D and is a doublet of triplets. The coupling constant for the triplet is 1.4 Hz and the coupling

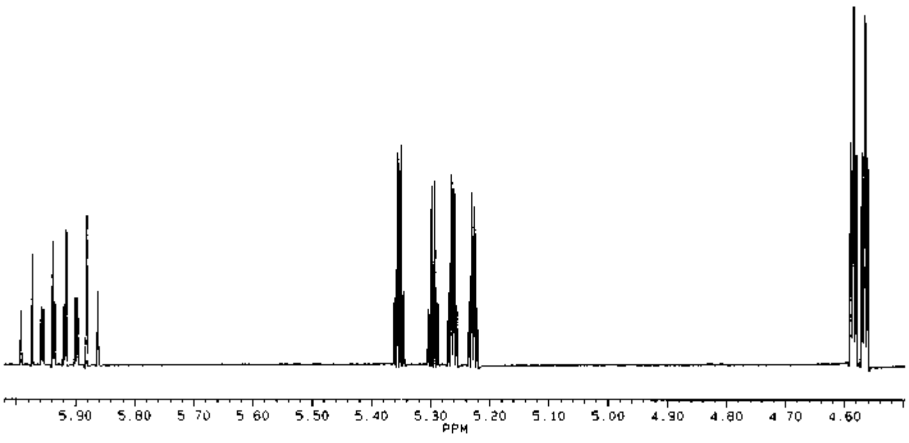

Figure 8.24*a* **Allyl acetate in $CDCl_3$.**

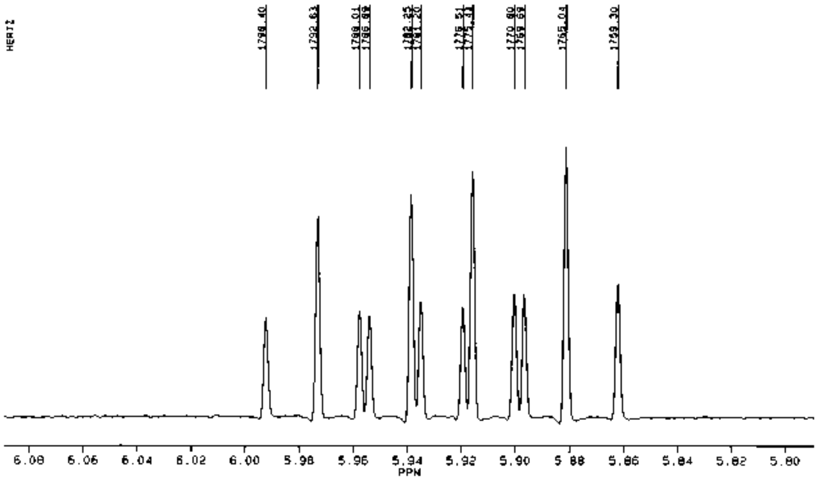

Figure 8.24*b* **Allyl acetate in $CDCl_3$ (expansion).**

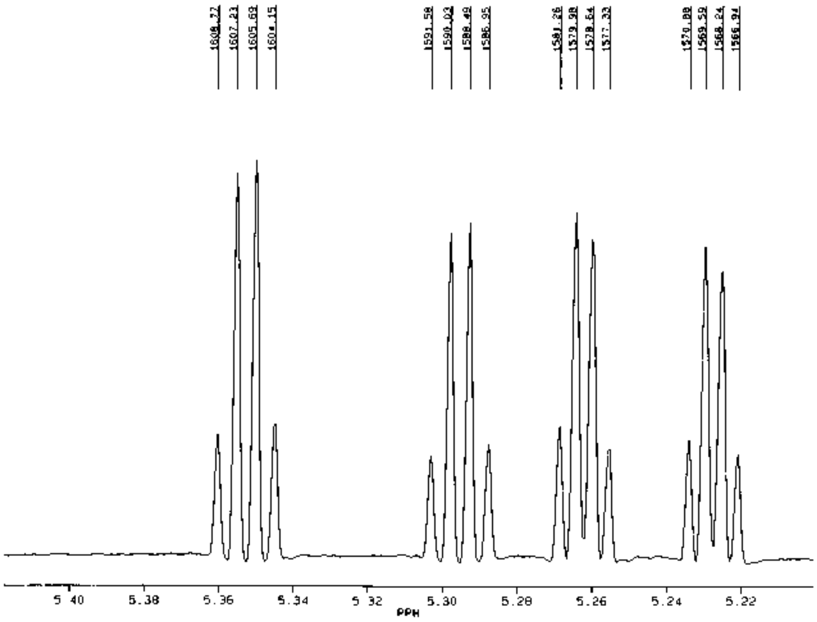

Figure 8.24*c* **Allyl acetate in $CDCl_3$ (expansion).**

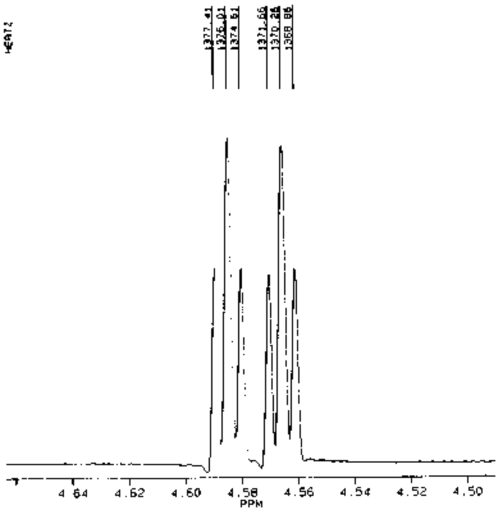

Figure 8.24*d* **Allyl acetate in $CDCl_3$ (expansion).**

constant for the doublet is 5.8 Hz, which can be measured between any two corresponding peaks in the two triplets.

The four quartets around 5.3 ppm (Fig. 8.24*c*) are actually two doublets of quartets at 5.32 and 5.24 ppm and correspond to H_A and H_B in the structure above. At 5.32 ppm, the multiplet is a doublet of quartets, $J = 17.2, 1.5$ Hz. At 5.24 ppm, we have another doublet of quartets, $J = 10.4, 1.3$ Hz. We see quartets because the long-range allylic coupling to the two H_D signals gives a triplet that has a coupling constant J that is approximately equal to the geminal coupling constant (~1.4 Hz) between H_A and H_B. Since NMR line widths are naturally several tenths of a hertz, it is not possible to distinguish between coupling constants such as these that differ only by 0.2 Hz. We can unambiguously distinguish H_A and H_B by the magnitudes of their coupling constants to H_C, which are 17.2 and 10.4 Hz. Since trans coupling constants are larger than cis coupling constants, H_A must have the 17.2-Hz coupling constant to H_C and is thus assigned to the signal centered at 5.32 ppm. Since H_B is coupled to H_C by $J = 10.4$ Hz it is assigned to the signal centered at 5.24 ppm.

Finally, we already know what the multiplet for H_C should look like, since we know all of its coupling constants. It is coupled to the two H_D protons with a coupling constant of 5.8 Hz, to H_A with $J = 17.2$, Hz, and to H_B with $J = 10.4$ Hz. The multiplet for H_C at 5.93 ppm should be, therefore, a doublet of doublets of triplets with $J = 17.2$, 10.4, and 5.8 Hz, respectively. There should be $2 \times 2 \times 3 = 12$ lines and the width should be $17.2 + 10.4 + (2 \times 5.8) = 39.2$ Hz. There are indeed 12 lines (Fig. 8.24*b*) and the distance between the outside peaks is 39.1 Hz, which is a perfectly reasonable deviation from the ideal:

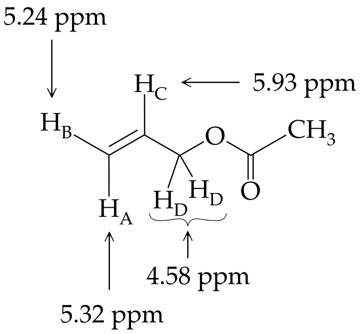

Allyl acetate, chemical shift assignments

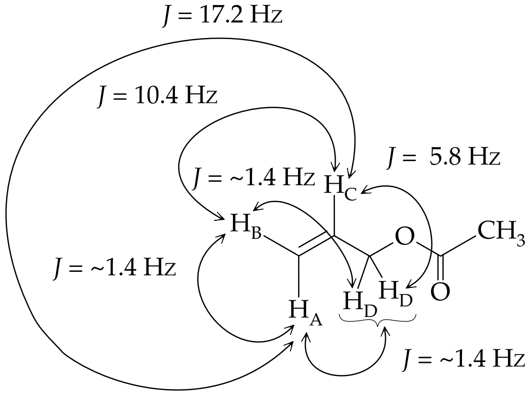

Allyl acetate, proton-proton coupling constants

Since cyclohexane rings are often held in at most two potential conformations (both chairs), coupling constants may allow the determination of relative stereochemistry. On cyclohexane rings in chair conformations, the axial–axial coupling constants for vicinal protons (180° dihedral angle) are on the order of 9–12 Hz.

Equatorial–equatorial and equatorial–axial coupling constants (60° dihedral angles) are on the order of 2–4 Hz. Thus it is often a relatively simple matter to determine stereochemical relationships on a six-membered ring using NMR spectroscopy.

Take, for example, the two diastereomers of 4-*tert*-butylcyclohexanol:

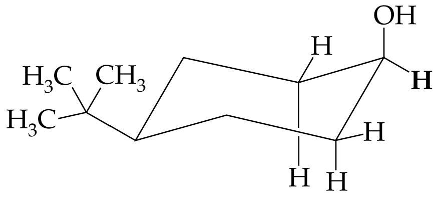

Cis

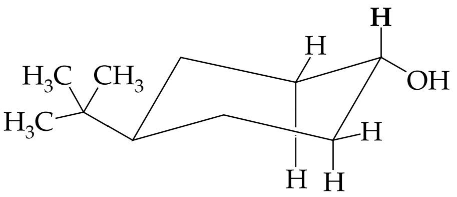

Trans

We know that the very large *tert*-butyl group will effectively always be equatorial. By examining the coupling to the methine proton of the alcohol, it is simple to determine whether that proton is axial or equatorial. It is also possible to distinguish between these stereoisomers by using chemical-shift information (in one isomer the alcohol methine is seen at ~3.52 ppm and in the other at ~4.04 ppm), but use of coupling information provides a far more definitive and unambiguous determination of stereochemistry.

The alcohol methine proton in the cis isomer is equatorial and thus has a 60° dihedral angle to all four adjacent protons that give rise to a pentet (which is really a doublet of doublets of doublets of doublets with equal coupling constants) with a coupling constant *J* = ~3 Hz, which is seen in Figure 8.25.

The alcohol methine in the trans isomer is axial and thus has a 180° dihedral angle to each of the two adjacent axial protons and a 60° dihedral angle to each of the two adjacent equatorial protons. This arrangement gives rise to a triplet of triplets with *J* = ~13 and 3 Hz, which is shown in Figure 8.26.

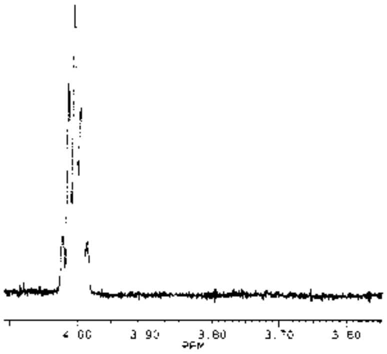

Figure 8.25 ***cis*-4-*tert*-Butylcyclohexanol.**

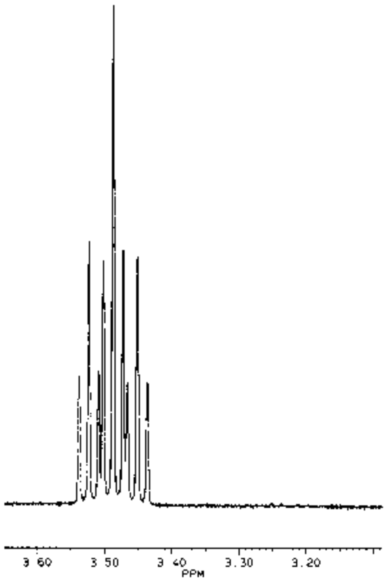

Figure 8.26 ***trans*-4-*tert*-Butylcyclohexanol.**

^{13}C NMR SPECTROSCOPY

With the advent of Fourier transform (FT) NMR spectrometers, ^{13}C NMR spectroscopy is now available as a simple and routine tool for the structure determination of organic molecules. Since ^{13}C is of low natural abundance (1.1%), addition of many spectra is required to obtain acceptable signal-to-noise (*S/N*) levels. With modern spectrometers, ^{13}C spectra can often be acquired simply by issuing software commands; in some instruments a different probe is inserted into the magnet. Since ^{13}C resonates at roughly 25% of the proton operating frequency of a spectrometer system, an instrument that acquires ^{1}H spectra at 300 MHz will be reset to about 75 MHz for ^{13}C work.

Generally, ^{13}C NMR spectra are acquired while the entire ^{1}H frequency range is irradiated by a second rf coil inside the probe assembly. These spectra are referred to as broadband-decoupled ^{13}C spectra and they do not show the effect of spin–spin coupling to ^{1}H nuclei. Such decoupling is done because ^{1}H–^{13}C coupling constants can be quite large (a few hundred Hz) relative to chemical-shift differences, which leads to multiplets split over a large portion of the spectrum and subsequent confusion (Fig. 8.27). It is often simpler to see a single line for each distinct carbon atom in a molecule. Furthermore, irradiation of the ^{1}H spectrum results in signal enhancement of the ^{13}C signals of the attached carbon atoms. This enhancement is the nuclear Overhauser effect (NOE).

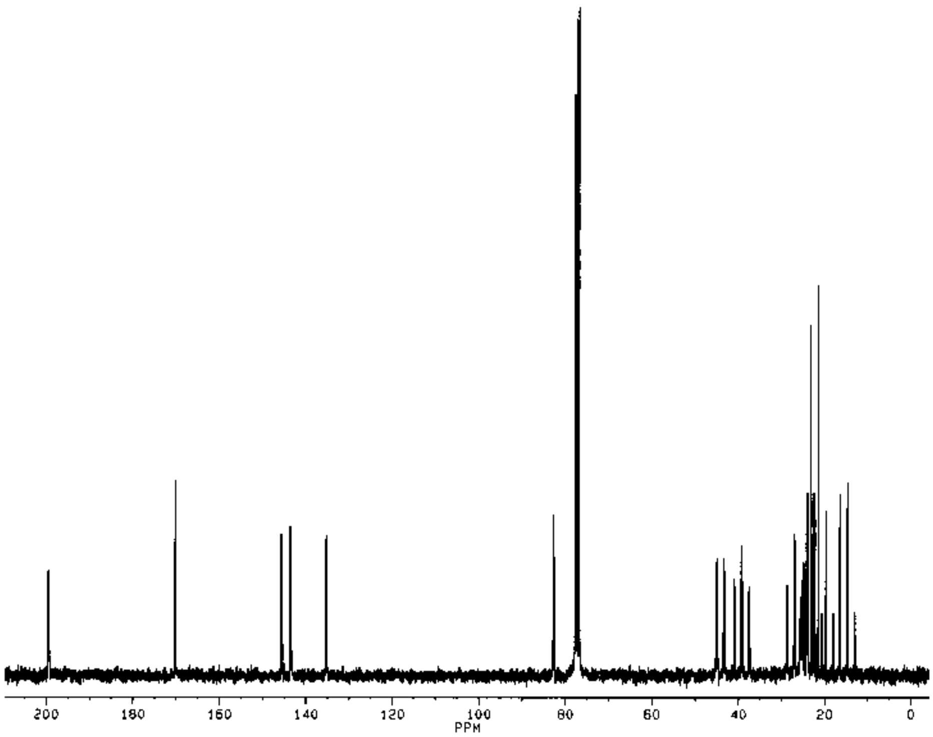

Figure 8.27 **Fully ^{1}H-coupled ^{13}C NMR spectrum of 5-(1-acetoxy-1-methylethyl)-2-methyl-2-cyclohexenone in $CDCl_3$.**

The ^{13}C NMR chemical shifts follow the same rough trends as seen in ^{1}H chemical shifts. ^{13}C chemical shifts, however, are not nearly as amenable to prediction based on the electronegativity of substituents as are ^{1}H chemical shifts. The ^{13}C chemical shifts are, in general, less sensitive to substituent electronegativities, and are far more sensitive to steric effects than are ^{1}H chemical shifts. A brief listing of approximate ^{13}C chemical shifts is provided in Table 8.26; a more extensive and thorough listing is available in the Silverstein et al. reference (Bibliography). As in ^{1}H NMR spectroscopy, TMS ($Si(CH_3)_4$) is used as the internal reference and the chemical shift of TMS is defined as zero. Except for functional groups such as acetals and ketals, sp^{3}-hybridized carbon atoms appear upfield (to the right) of 100 ppm, and sp^{2}-hybridized carbon atoms appear downfield of 100 ppm. Common carbonyl-containing functional groups appear downfield of 160 ppm. Aldehydes and ketones appear at 195–220 ppm; esters, amides, anhydrides, and carboxylic acids appear at 165–180 ppm.

Typical ^{13}C NMR spectroscopy provides an NMR spectrum that is not amenable to integration because of the NOE and insufficient relaxation delays. Therefore, the number of carbon atoms giving rise to a given signal cannot generally be determined by these techniques. It is possible to obtain ^{13}C NMR spectra that can be accurately integrated (inverse-gated decoupling), but this experiment requires a great deal of acquisition time to achieve adequate signal-to-noise levels.

Information about C—H coupling can be readily obtained, however. Fully coupled ^{13}C NMR spectra are not very useful for structure determination because C—H couplings are large (~120–270 Hz, depending mainly on the hybridization at carbon) and multiplets tend to overlap. Furthermore, when the hydrogen atoms are not irradiated, there is no NOE, and the signal-to-noise ratio suffers significantly. The most common use for coupling information is to

Table 8.26 **Approximate ^{13}C NMR Chemical Shifts**

Functional Group	Carbon[a]	Chemical Shift/ δ (ppm)
Alkyl carbon atoms		~5–45
	1° R—**C**H$_3$	~5–30
	2° R—**C**H$_2$—R′	~15–35
	3° R—**C**HR′R″	~20–40
	4° R**C**R′R″R‴	~25–45
Alkenyl carbon atoms		~110–150
	H$_2$**C**═C	~100–125
	HR**C**═C	~125–145
	RR′**C**═C	~130–150
Aromatic carbon atoms		~120–160
Alkynyl carbon atoms	**C**≡C	~65–90
Nitriles	R—**C**≡N	~115–125
Alcohols and ethers	**C**—OH(R)	~50–75
	C—O (epoxides)	~35–55
Amines	**C**—N	~30–55
Alkyl halides	**C**—X	~0–75
Carbonyl groups	**C**═O	~165–220
Ketones, aldehydes	R**C**OR′, R**C**HO	195–220
Carboxylic acids, esters	R**C**O$_2$H, R**C**O$_2$R′	165–180
Amides, anhydrides	R**C**ON, R**C**O$_2$OCR′	160–175

[a] R = alkyl group.

determine the number of protons attached to a given carbon atom. This can be done in a variety of ways, some of which do not actually display the carbon signals as multiplets due to coupling to attached protons.

Single-frequency off-resonance decoupling (SFORD) is a useful technique for determining the number of hydrogen atoms attached to a given carbon. The decoupler is tuned off to one side of the proton spectrum and the sample is irradiated at a single frequency giving rise to ^{13}C spectra that show C—H couplings as a fraction of their actual values and that show a partial NOE. The apparent C—H coupling is dependent on both the actual coupling constant and the difference between the decoupler frequency and the resonance frequency of the hydrogen in question. The major disadvantage of SFORD is its low signal-to-noise ratio, which is due to two factors. First, there is only a partial NOE. Second, when NMR signals are split into multiplets, the signal intensity becomes distributed among several peaks. Thus, SFORD spectra require significantly more spectral acquisitions than do fully decoupled ^{13}C NMR spectra, and to some extent have been replaced with distortionless enhancement by polarization transfer (DEPT) spectra.

DEPT ^{13}C NMR spectroscopy provides a rapid way of determining the number of hydrogen atoms attached to a given carbon atom. DEPT spectra result from a multiple-pulse sequence that terminates in a "read pulse," which can be varied according to the spectrum desired. In DEPT spectra, all peaks are singlets; quaternary carbon atoms (without attached hydrogen atoms) are not seen in any DEPT spectra. In DEPT-135° spectra, CH and CH$_3$ groups appear as singlets of positive intensity, and CH$_2$ groups appear as negative peaks. The

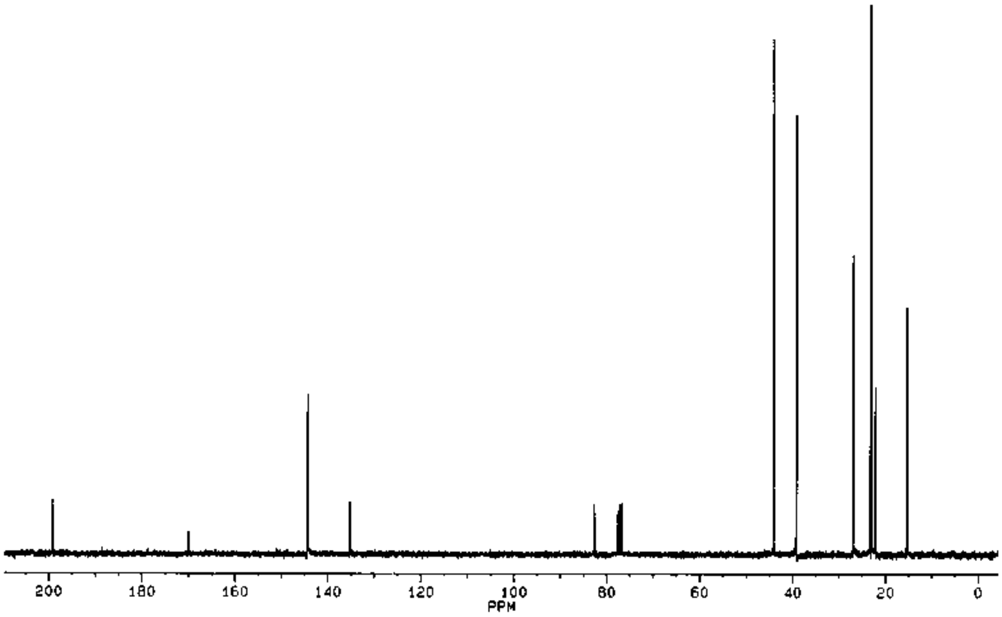

Figure 8.28 Broadband ^{1}H-decoupled ^{13}C NMR spectrum of 5-(1-acetoxy-1-methylethyl)-2-methyl-2-cyclohexenone in $CDCl_3$.

DEPT-90° spectra show only CH groups, and thus allow CH_3 and CH groups to be distinguished. In combination with a routine fully decoupled spectrum, DEPT spectra allow unambiguous assignment of the number of hydrogen atoms attached to each carbon. In practice, such spectral editing techniques are not perfect, and small residual peaks are often seen where, in principle, there should be none; these are usually small enough to be readily distinguished from the ″real″ peaks.

The fully coupled ^{13}C NMR spectrum of the acetoxy-enone (**I**) is shown in Figure 8.27, the broadband decoupled spectrum in Figure 8.28, and the SFORD spectrum in Figure 8.29. The DEPT-135° spectrum is shown in Figure 8.30, and the DEPT-90° spectrum in Figure 8.31. The 1:1:1 triplet centered at 77 ppm is due to the solvent, $CDCl_3$.

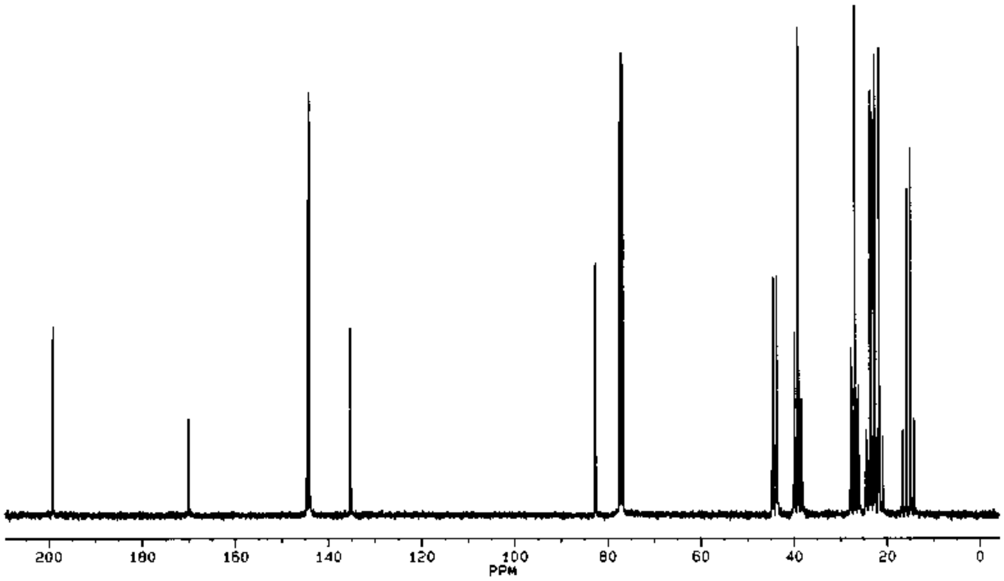

Figure 8.29 Single-frequency off-resonance decoupled (SFORD) ^{13}C NMR spectrum of 5-(1-acetoxy-1-methylethyl)-2-methyl-2-cyclohexenone in $CDCl_3$.

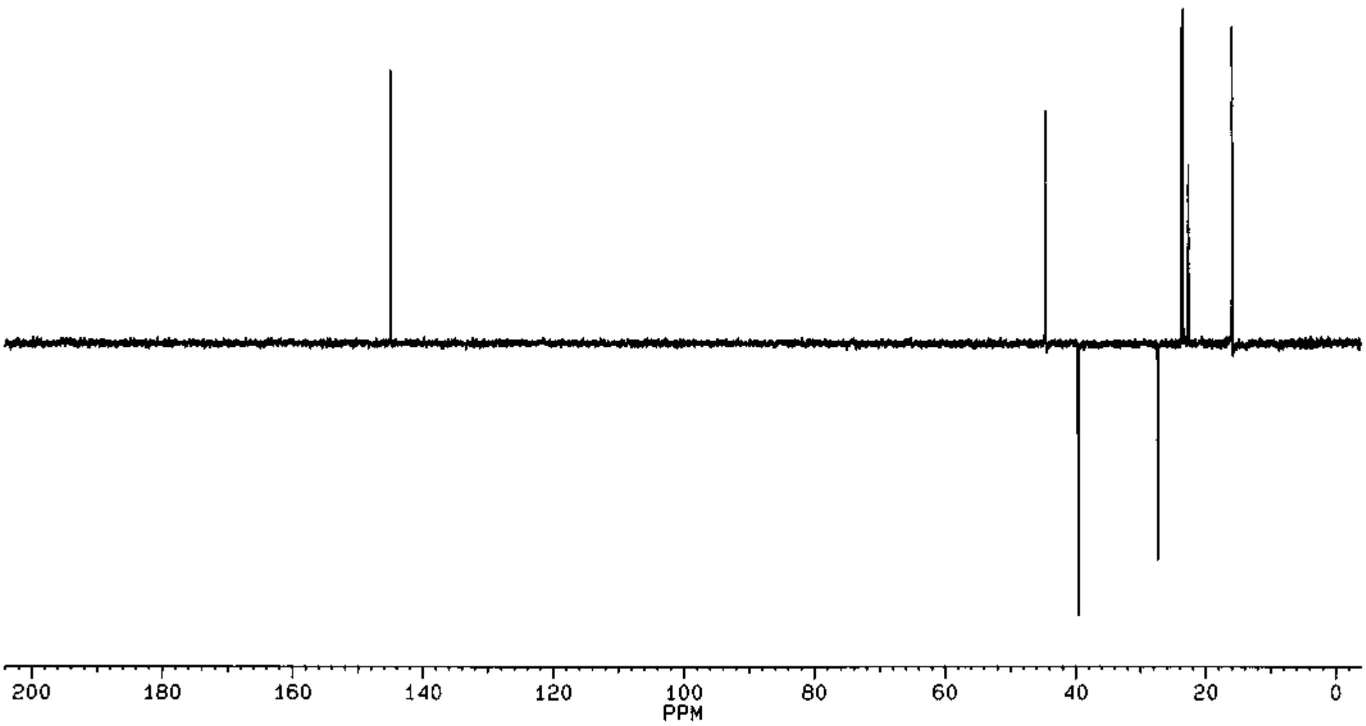

Figure 8.30 The DEPT-135° spectrum of 5-(1-acetoxy-1-methylethyl)-2-methyl-2-cyclohexenone in $CDCl_3$.

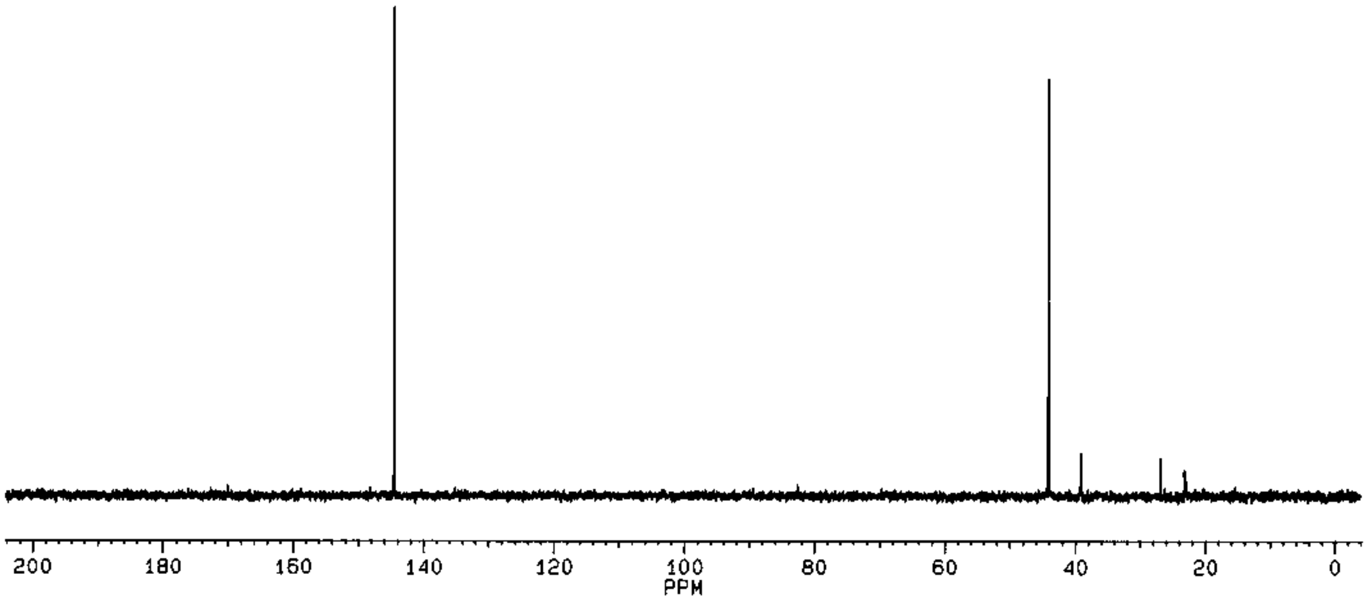

Figure 8.31 The DEPT-90° spectrum of 5-(1-acetoxy-1-methylethyl)-2-methyl-2-cyclohexenone in $CDCl_3$.

Interpretation and assignment of the ^{13}C NMR spectrum are much easier when we unambiguously know how many protons are attached to each carbon.

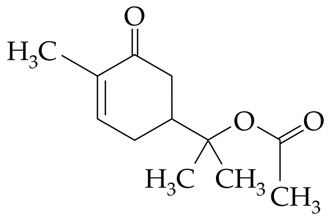

I

The ^{13}C NMR is often better than ^{1}H NMR for distinguishing functional groups because typical ^{13}C chemical shifts are in the range 0–200 ppm relative to TMS, as compared to 0–10 ppm for proton chemical shifts. Coupling between adjacent ^{13}C nuclei is not observed (except in isotopically enriched samples) because the probability of having two rare isotopes adjacent to one another is very small. Because of the absence of decoupling, ^{13}C spectra are less complex than ^{1}H spectra, and ^{13}C spectra are often better suited for the detection and identification of isomeric or other impurities in a sample; it is easy for small peaks to be concealed underneath a complex second-order multiplet in the ^{1}H NMR spectrum.

The 300-MHz ^{1}H spectrum of 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride is shown in Figure 8.32. Owing in part to the presence of two stereocenters, as well as to long-range coupling through the π system of the alkene, the entire ^{1}H spectrum is second order at this field strength, and no information is available from the coupling constants because the spectrum is too complex. Limited assignments to peaks could be made on the basis of chemical shift, but it would be difficult to make any statements regarding purity of our sample based on the ^{1}H NMR spectrum, because an impurity could easily be hidden underneath any of the complex signals.

The ^{13}C spectrum of 4-cyclohexene-*cis*-1,2-dicarboxylic acid anhydride (Figure 8.33) is much less complex. Because of the mirror plane of symmetry in the compound, there are only four different carbon atoms and thus only four lines are seen in the fully decoupled ^{13}C NMR spectrum. This simplicity makes it easy to detect isomeric or other impurities in this sample. These impurities were not as easy to detect in the ^{1}H NMR spectrum. The 1:1:1 triplet centered at 77 ppm is due to the solvent, $CDCl_3$. Although no ^{1}H–^{13}C

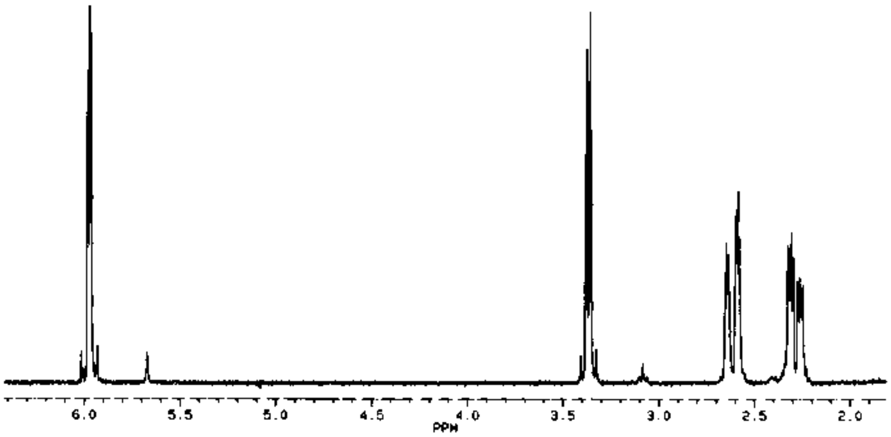

Figure 8.32 The ^{1}H NMR spectrum of 4-cyclohexene-cis-1,2-dicarboxylic acid anhydride in $CDCl_3$.

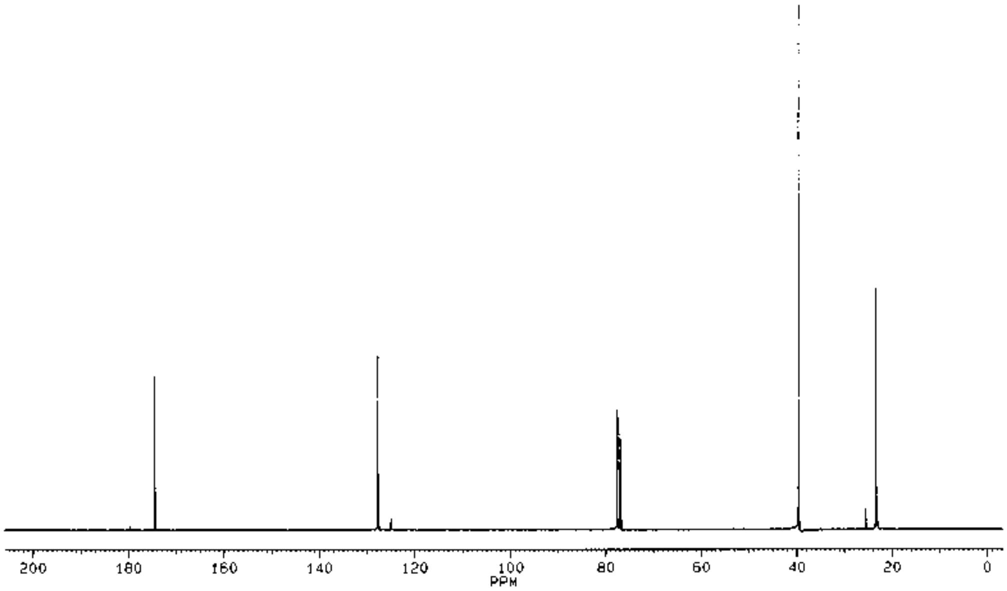

Figure 8.33 Broadband ^{1}H-decoupled ^{13}C NMR spectrum of 4-cyclohexene-cis-1,2-dicarboxylic acid anhydride in $CDCl_3$.

coupling is observed because of the ^{1}H broadband decoupling, ^{2}H–^{13}C coupling is observed because ^{1}H and ^{2}H resonate at different frequencies.

TWO-DIMENSIONAL NMR SPECTROSCOPY

Two significant developments in NMR spectroscopy are the use of Fourier transform techniques, and the development of two-dimensional (2D) NMR spectroscopy. Two-dimensional spectra are obtained using a sequence of rf pulses that includes a variable delay or delays. A set of FIDs is acquired and stored. The variable delay is incremented by a small amount of time and a new set of FIDs are obtained and stored, and so on. At the end, the resulting matrix of FID data is Fourier transformed twice: once with respect to the acquisition time (as in normal FT-NMR) and second with respect to the time of the variable delay in the pulse sequence. The resulting data represent a surface and are presented as a contour plot of that surface.

The most useful 2D spectra for organic compound identification are called correlation spectra. Correlation Spectroscopy (COSY) spectra are presented as a contour plot with routine proton spectra along both of the axes, as shown in the COSY spectrum of ethyl vinyl ether in Figure 8.34. The spectra along the axes are low-digital-resolution spectra and appear to be a bit different from those generated as usual NMR spectra. Note that the 2D spectrum is symmetric about the diagonal that runs from the lower left corner to the upper right corner. Every peak is represented by a peak along the diagonal and, in fact, the diagonal *is* the normal proton spectrum. Where the contour plot indicates a peak other than on the diagonal, the interpretation is that the corresponding peaks on the two axes represent protons coupled to one another.

For example, draw a line down from the signal at 1.2 ppm on the horizontal axis. This line encounters the diagonal at 1.2 ppm and then there is a peak at 3.7 ppm. By drawing a horizontal line over to the spectrum on the vertical axis, we can see that the peaks at 1.2 and 3.7 ppm are coupled to one another; these are the signals from the methyl triplet and methylene quartet of the ethyl group.

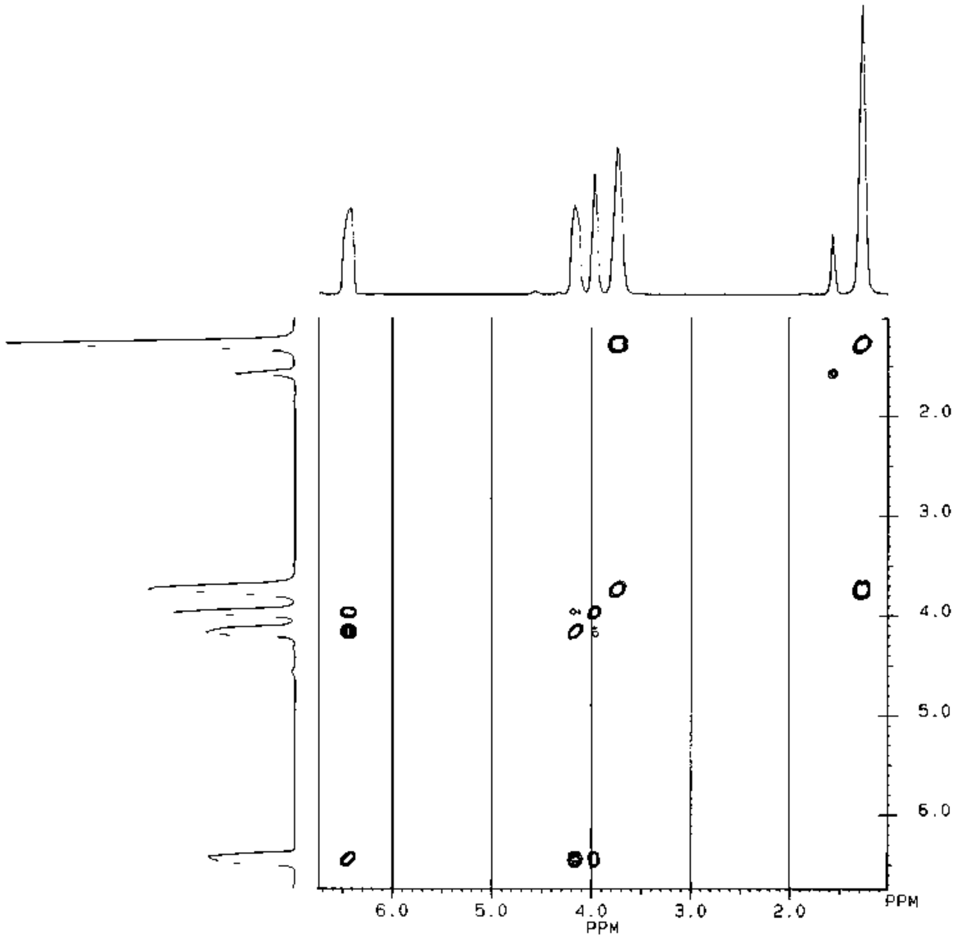

Figure 8.34 The COSY NMR spectrum of ethyl vinyl ether.

We can also see that neither of these is coupled to the remainder of the spectrum, which of course is the vinyl group. Each proton in the vinyl group is coupled to each of the others, as we can see in the COSY spectrum. For example, the signal at 4.2 ppm is coupled to both the signal at 4.0 ppm and the one at 6.4 ppm.

It is possible to obtain 2D spectra that correlate the spectra of different nuclei, such as ^{1}H and ^{13}C. The heteronuclear correlation spectrum of ethyl vinyl ether is shown in Figure 8.35. The ^{13}C spectrum is along the horizontal axis and the ^{1}H spectrum is along the vertical axis. The peaks of the 2D spectrum indicate that the corresponding peaks on the axes represent a carbon and a proton (or protons) that are directly bonded. By using this spectrum we can easily verify the ^{13}C and ^{1}H chemical-shift assignments below:

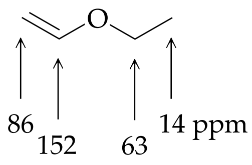

^{13}C chemical shifts

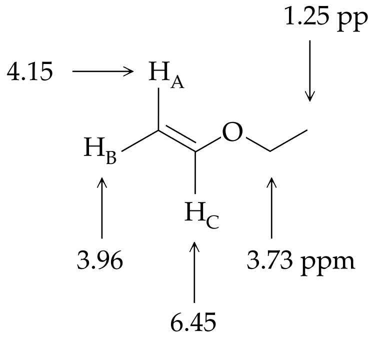

^{1}H chemical shifts

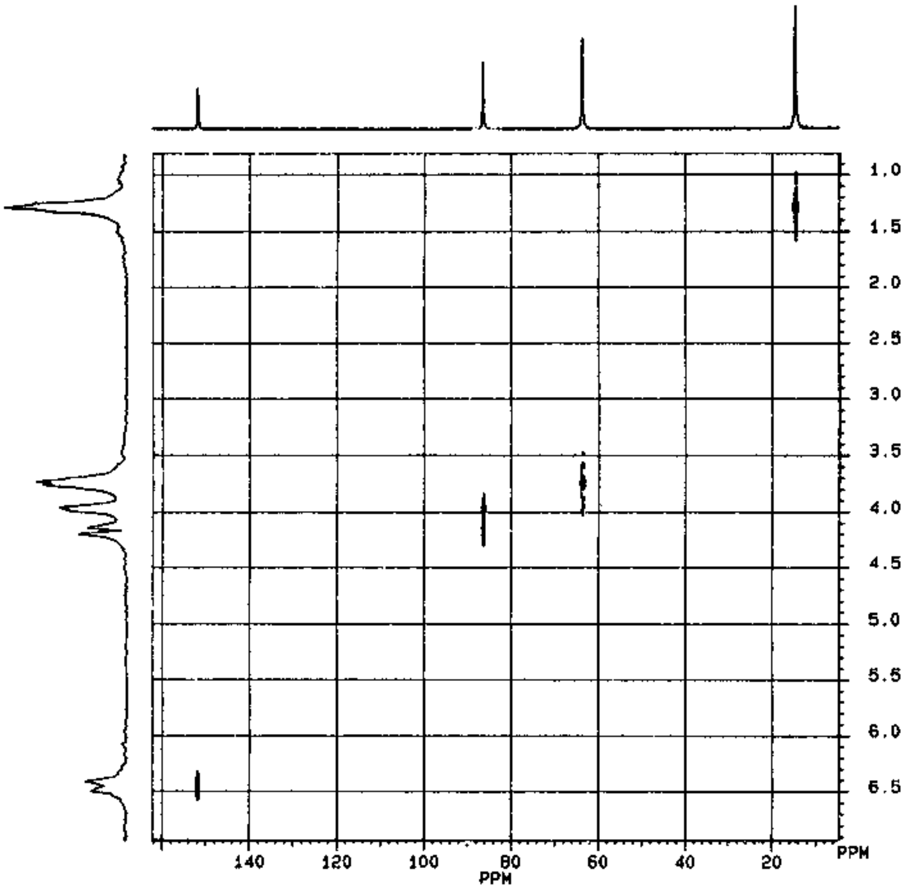

Figure 8.35 Ethyl vinyl ether ^1H–^{13}C correlation spectrum.

There are many other powerful 2D NMR spectroscopic techniques that can provide a wealth of information about molecular structure, even in organic molecules as large as proteins and nucleic acids. A few of the good texts that provide further information about these powerful tools are listed in the Bibliography.

NUCLEAR MAGNETIC RESONANCE SAMPLING

It is usually simple to prepare a sample of a small organic molecule (MW 500) for NMR analysis. For ^1H NMR spectroscopy, the sample size compatible with CW spectrometers is in the range 30–50 mg dissolved in about 0.5 mL of solvent. Fourier transform spectrometers require only 2–3 mg of sample in the same volume of solvent. Most samples are measured in solution in thin-walled tubes 5 mm in diameter and 18–20 cm long. NMR sample tubes are expensive and delicate; they must be as perfectly straight, and have as perfectly concentric walls, as possible. The sample tube is filled to a depth of about 3 cm. Filling the tube to this depth maximizes sample concentration in the active part of the NMR probe, and thus the strength of the signal. Adding more solvent just wastes sample by dilution. The sample tube is spun about its axis in the instrument to average out small changes in the tube walls and in the magnetic field strength over the sample volume. There must be enough solvent in the tube to ensure that the vortex, or whirlpool, created when the tube is spun, does not extend down into the portion of the tube where the rf coils in the NMR probe are active. Many instruments use a depth gauge that shows exactly where this area is. Glass microcells are now commercially available. An inexpensive microcell technique for CW-NMR spectrometers using ordinary 5-mm NMR tubes has been described.[2]

The most practical NMR solvent is deuterochloroform ($CDCl_3$); it is relatively cheap and dissolves many different compounds. Handle this solvent with care, in the **hood,** because it is **toxic!** Many other deuterated solvents are commercially available, including acetone, methanol, and water. The universally accepted internal reference compound employed in making these measurements is tetramethylsilane, $Si(CH_3)_4$ (TMS). The most convenient source of TMS is commercially available $CDCl_3$, which contains about 1% TMS for use with CW spectrometers (commercially available 0.03% solutions are more appropriate for FT spectrometers).

HOOD

The most significant problem in sample preparation is the exclusion of small pieces of dust and dirt, because they may contain magnetic material, which will result in poor spectra. Scrupulously clean samples of liquids can often simply be added to the NMR tube, followed by the solvent. Liquids containing visible impurities, and solids, are best prepared by dissolving the sample in about 0.3–0.4 mL of solvent in a small vial. The solution can then be filtered into the NMR tube through a Pasteur pipet plugged with a small piece of cotton, and the pipet can be rinsed with the NMR solvent to achieve the appropriate volume in the NMR tube. Since TMS is volatile (bp = 26.5 °C), the tube should be capped following addition, or the TMS will evaporate. If several people in a laboratory section are going to be obtaining NMR spectra, you will find that the spectrometer will be easier to tune with each new sample if all the sample tubes are filled to exactly the same level.

At this point, all specific instructions are dependent on the NMR spectrometer available to you. In general, the sample is inserted into the magnet and the magnetic field is adjusted very slightly (called shimming or tuning) to obtain a magnetic field that is as homogeneous as possible throughout the sample volume observed by the spectrometer. These adjustments are accomplished by energizing a collection of small electromagnetic coils around the sample, a process that is often done by the spectrometer's computer. Symptoms of a poorly shimmed spectrometer include broad and/or asymmetric peaks. The best place to check for this is on the TMS peak because it is the one peak in the spectrum you *know* should be narrow and symmetric. Once the spectrometer is tuned, the spectrum is obtained and plotted, and the sample is removed from the magnet.

Your NMR sample can be easily recovered by emptying the NMR tube into a small vial, rinsing the tube once or twice with (a nondeuterated) solvent, and then evaporating the solvent in a **hood** under a gentle stream of dry nitrogen.

HOOD

---

[2] Yu, S. J. *J. Chem. Educ.* **1987,** *64,* 812.

QUESTIONS

Several 60-MHz ^{1}H NMR spectra are given below (Figs. 8.36–8.40) along with the molecular formula of the compound.[3] You should be able to account for at least one acceptable structure and for all of the observed resonances.

8-16. C_4H_8O. Spectrum *a:*

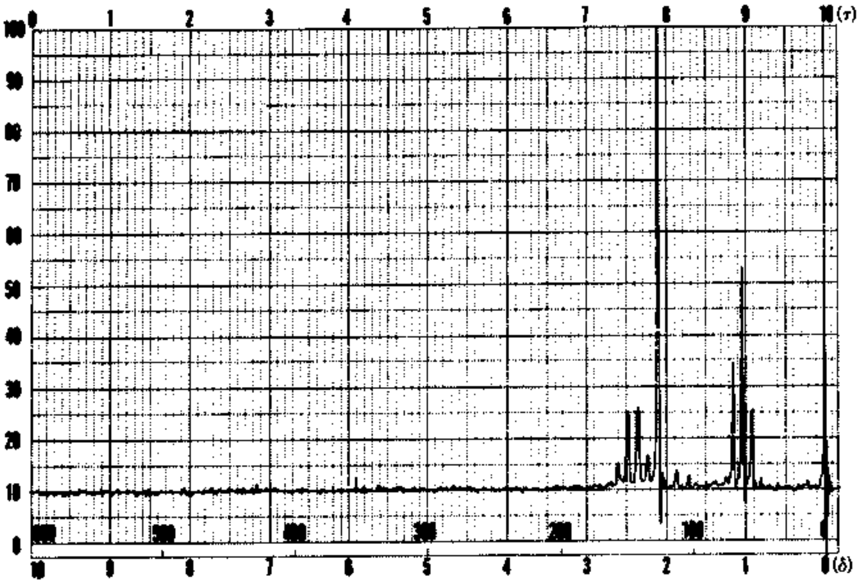

Figure 8.36 **NMR unknown spectrum *a*.**

8-17. $C_3H_6O_2$. Spectra *b* and *c*. Two compounds with the same empirical formula:

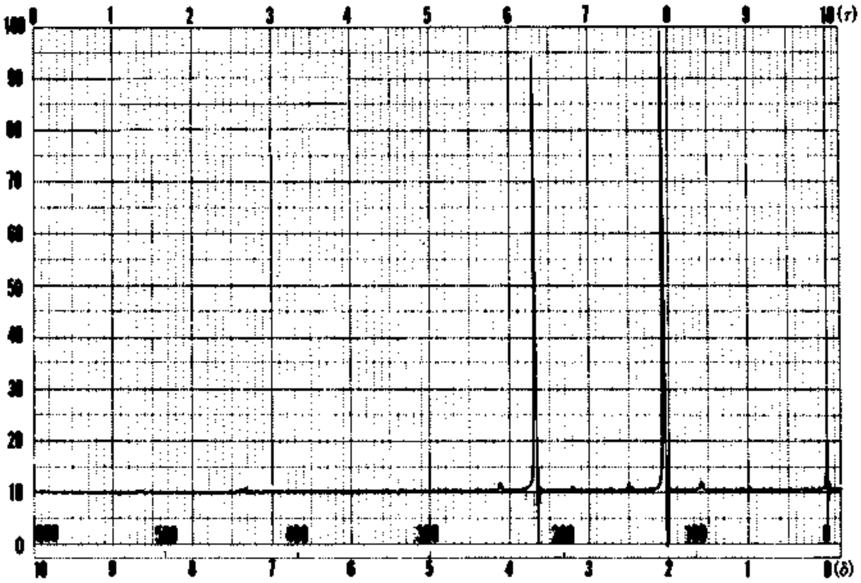

Figure 8.37*a* **NMR unknown spectrum *b*.**

---

[3] From Pouchert, C. J. *The Aldrich Library of NMR Spectra,* 2nd ed.; Aldrich Chemical Co.: Milwaukee, WI, 1983.

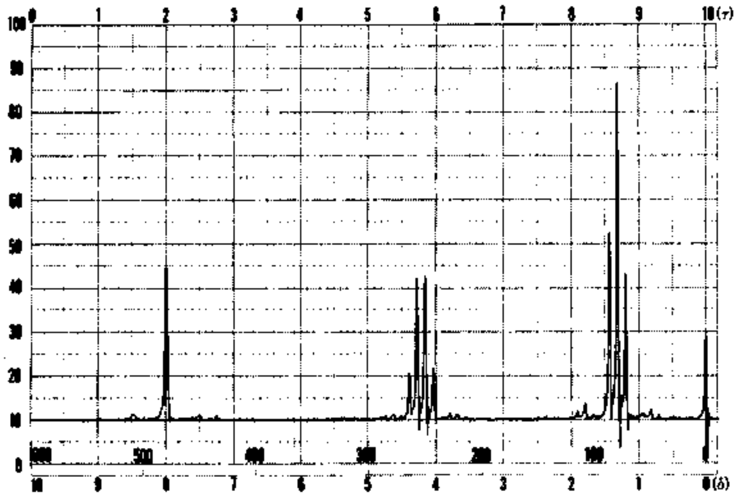

Figure 8.37*b* **NMR unknown spectrum *c*.**

8.18 C_4H_8O. Spectrum *d*. Also give some thought to the weak resonances at 0.5 and 1.8 ppm:

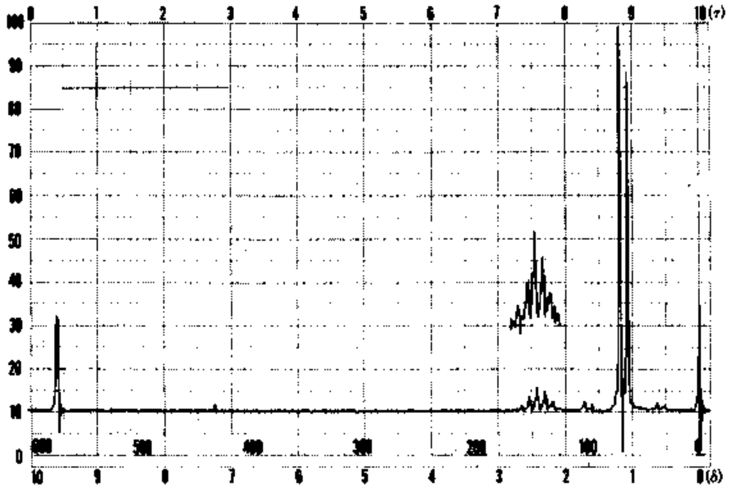

Figure 8.38 **NMR unknown spectrum *d*.**

8-19 C_7H_7Cl. Spectrum *e:*

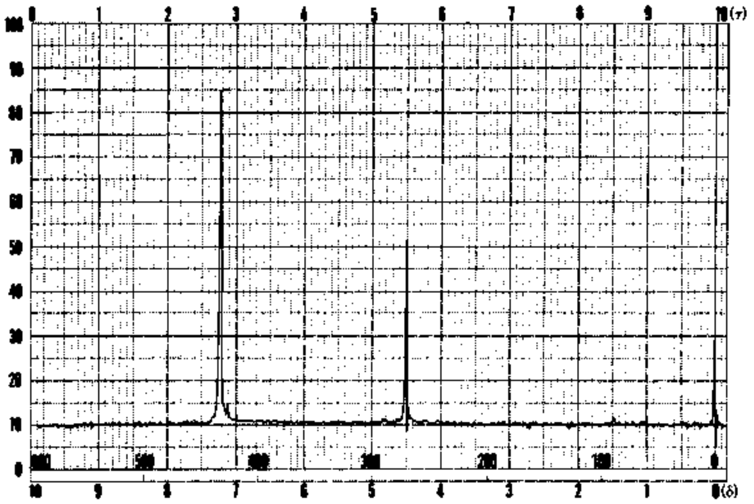

Figure 8.39 **NMR unknown spectrum *e*.**

8-20. $C_8H_{10}O$. Spectrum *f*:

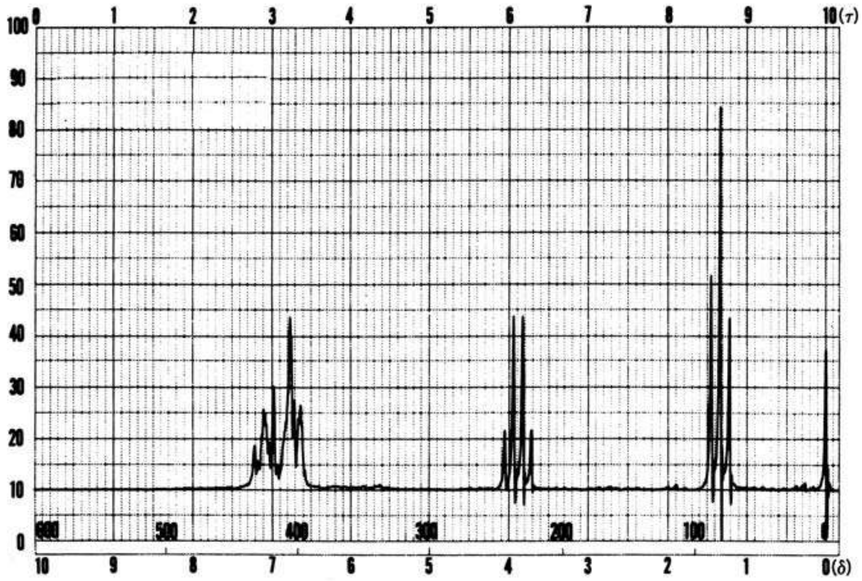

Figure 8.40 NMR unknown spectrum *f*.

BIBLIOGRAPHY

Nuclear magnetic resonance theory and principles of interpretation:

Abraham, R. J.; Fisher, J.; Loftus, P. *Introduction to NMR Spectroscopy;* Wiley: London, 1988.

Cooper, J. W. *Spectroscopic Techniques for Organic Chemists;* Wiley: New York, 1980.

Field, L. D.; Sternhell, S.; Kalman, J. R. *Organic Structures from Spectra*, 3rd ed.; Wiley: London, 2002.

Richards, S. A. *Laboratory Guide to Proton NMR Spectroscopy;* Blackwell Scientific Publications: London, 1988.

Silverstein, R. M.; Webster, F. X.; Kiemle, D. J. *Spectrometric Identification of Organic Compounds,* 7th ed.; Wiley: New York, 2005.

Sorrell, T. N. *Interpreting Spectra of Organic Molecules;* University Science Books: Mill Valley, CA, 1988.

Advanced theory and spectroscopic techniques:

Atta-ur-Rahman *Nuclear Magnetic Resonance. Basic Principles;* Springer-Verlag: New York, 1986.

Claridge, T. D. W. *High-Resolution NMR Techniques in Organic Chemistry. Tetrahedron Organic Chemistry Series Volume 19;* Pergamon: Oxford, UK, 1999.

Derome, A. E. *Modern NMR Techniques for Chemistry Research;* Pergamon Press: Oxford, UK, 1987.

Duddeck, H.; Dietrich, W.; Toth, G. *Structure Elucidation by Modern NMR. A Workbook,* 3rd ed. Springer-Verlag: New York, 1998.

Sanders, J. K. M.; Hunter, B. K. *Modern NMR Spectroscopy. A Guide for Chemists;* 2nd ed.; Oxford University Press: Oxford, UK, 1993.

Sanders, J. K. M.; Constable, E. C.; Hunter, B. K. *Modern NMR Spectroscopy. A Workbook of Chemical Problems;* 2nd ed.; Oxford University Press: Oxford, UK, 1993.

Libraries of NMR spectra:

Bhacca, N. S.; Hollis, D. P.; Johnson, L. F.; Pier, E. A.; Shoolery, J. N. *NMR Spectra Catalog;* Varian Associates: Palo Alto, CA, 1963.

Pouchert, C. J. *The Aldrich Library of NMR Spectra,* 2nd ed.; Aldrich Chemical Co.: Milwaukee, WI, 1983.

Pouchert, C. J.; Behnke, J. *The Aldrich Library of* ^{13}*C and* ^{1}*H FT-NMR Spectra;* Aldrich Chemical Co.: Milwaukee, WI, 1992.

Online libraries of NMR spectra:

Online Resource Guide (NMR and Other Spectra (http://www.library.illinois.edu/chx/onlineresources/nmr.html))

SciFinder Scholar (https://scifinder.cas.org/)

ULTRAVIOLET–VISIBLE SPECTROSCOPY: INTRODUCTION TO ABSORPTION SPECTROSCOPY

In an atom, molecule, or ion, a limited number of electronic energy states are available to the system because of the quantized nature of the energies involved. The absorption of a photon by the system can be interpreted as corresponding to the occupation of a new energy state by an electron. The difference in energy between these two states may be expressed as ΔE:

$$\begin{array}{l} \rule{2em}{0.4pt}\ \text{Upper state (excited electronic state, } E_1\text{)} \\ \updownarrow\ \Delta E \\ \rule{2em}{0.4pt}\ \text{Lower state (ground electronic state, } E_0\text{)} \end{array}$$

where the energy of the photon, E, is related to the frequency of the radiation by the Planck equation,

$$E = h\nu_i$$

where h is Planck's constant, 6.626×10^{-34} J s, and ν_i is the frequency in hertz. In the case above, $\Delta E = E_1 - E_0 = h(\nu_1 - \nu_0) = h\nu_i$.

Thus, when a frequency match between the radiation and an energy gap (ΔE) in the substance occurs, a transition between the two states involved may be induced. The system can either absorb or emit a photon corresponding to ΔE, depending on the state currently occupied (emission would occur if the system relaxed from an upper-level excited state to a lower state). All organic molecules absorb photons with energies corresponding to the visible or ultraviolet regions of the electromagnetic spectrum, but to be absorbed, the incident energy in this frequency range must correspond to an available energy gap between an electronic ground state and an upper-level electronic excited state. The electronic transitions of principal interest to the organic chemist are those that correspond to the excitation of a single electron from the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO). As we will see, this will be the molecule's absorption occurring at the longest wavelength in the electronic absorption spectrum; it is, therefore, the most easily observed.

Electromagnetic radiation can be defined in terms of a frequency ν, which is inversely proportional to a wavelength λ times a velocity c ($\nu = c/\lambda$, where c is the velocity of light in a vacuum, 2.998×10^8 m/s, and $c = \nu\lambda$ is the wave velocity). Thus,

$$\Delta E = h\nu = \frac{hc}{\lambda} = hc\tilde{\nu}$$

where $\tilde{\nu}$ is the wavenumber, defined as the reciprocal of the wavelength $(1/\lambda)\times$ the velocity of light.

Most ultraviolet and visible (UV and vis) spectra are recorded linearly in wavelength, rather than linearly in frequency or in units proportional to frequency (the wavenumber) or in energy values. Wavelength in this spectral region is currently expressed in nanometers (nm, where 1 nm $= 10^{-9}$ m) or angstrom units (Å, where 1 Å $= 10^{-10}$ m). The older literature is full of UV–vis spectra in which wavelength is plotted in millimicrons (mμ), which are also equivalent to 10^{-9} m. For a further discussion of the relationship between frequency, wavelength, wavenumber, and refractive index, see the discussion on infrared spectroscopy.

It is unfortunate that because of instrumentation advantages this region of the spectrum is most often plotted in units that are nonlinear in energy (note the inverse relationship of E to λ) A convenient formula for expressing the relationship of wavelength and energy in useful values is

$$E = 28{,}635/\lambda \text{ kcal/mol} \qquad (\lambda \text{ in nm})$$

or in terms of wavenumbers

$$E = (28.635 \times 10^{-4})\tilde{\nu} \qquad (\tilde{\nu} \text{ in cm}^{-1})$$

Table 8.27 Spectroscopic Wavelength Ranges

Region	Wavelength (m)	Energy (kJ/mol)	Change Excited
Gamma ray	Less than10^{-10}	$> 10^6$	Nuclear transformation
X-ray	$10^{-8}-10^{-10}$	10^4-10^6	Inner shell electron transitions
Ultraviolet (UV)	$4 \times 10^{-7}-1 \times 10^{-8}$	10^3-10^4	Valence shell electrons
Visible (vis)	$8 \times 10^{-7}-4 \times 10^{-7}$	10^2-10^3	Electronic transitions
Infrared (IR)	$10^{-4}-2.5 \times 10^{-6}$	1–50	Bond vibrations
Microwave	$10^{-2}-10^{-4}$	10–1000	Molecular rotations
ESR	10^{-2}	10	Electron spin transitions
NMR	$0.5-5$	0.02–0.2	Nuclear spin transitions

The electromagnetic spectrum and the wavelength ranges corresponding to a variety of energy-state transitions are listed in Table 8.27. Infrared, UV–vis, and rf are of particular interest to the organic chemist because the excitation of organic substances by radiation from these regions of the spectrum can yield significant structural information about the molecular system being studied.

The absorption of rf energy by organic molecules immersed in strong magnetic fields involves exceedingly small energy transitions (~0.05 cal/mol), which correspond to nuclear spin excitations and result in NMR spectra. When a molecule absorbs microwave radiation, the energy states available for excitation correspond to molecular rotations and involve energies of roughly 1 cal/mol. With relatively simple molecules (in the gas phase) possessing a dipole moment (required for the absorption process) the analysis of the microwave spectrum can yield highly precise measurements of the molecular dimensions (bond lengths and angles). Unfortunately, relatively few organic systems exhibit pure rotational spectra that can be rigorously interpreted.

Absorption of radiation in the infrared region of the spectrum involves the excitation of vibrational energy levels and corresponds to energies in the range of about 1–12 kcal/mol. The excitation of electronic states requires considerably higher energies, from a little below 40 to nearly 300 kcal/mol. The corresponding radiation wavelengths would fall across the visible (400–800 nm), the near-UV (200–400 nm), and the far- (or vacuum) UV (100–200 nm) regions.The long-wavelength visible and near-UV regions of the spectrum hold information of particular value to the organic chemist. Here the energies correspond to the excitation of loosely held bonding (π) or lone-pair electrons. The far-UV region, however, involves high-energy transitions associated with the inner-shell and σ-bond electronic energy transitions. This region is difficult to access because atmospheric oxygen begins to absorb UV radiation below 190 nm, which requires working in evacuated or purged instruments (which is why this region is often referred to as the vacuum UV).

UV–VIS SPECTROSCOPY

As we have seen, the application of electronic absorption spectroscopy in organic chemistry is restricted largely to excitation of ground-state electronic levels in the near-UV and vis regions. When photons of these energies are absorbed, the excited electronic states that result have bond strengths appreciably less than their ground-state values, and the internuclear distances and bond angles will be altered within the region of the molecules where the electronic excitation occurs (see Figure 8.41). It is normally reasonable to assume that nearly all of the molecules are present in the ground vibrational state within the ground electronic state. The upper electronic state also contains a set of vibrational levels and any of these may be open to occupation by the excited electron (see Figure 8.41). Thus, an electronic transition from a particular ground-state level can be to any number of upper-level vibrational states on the excited electronic state.

The shape of an electronic absorption band will be determined to a large extent by the spacing of the vibrational levels and the distribution of band intensity over the vibrational sublevels. In most cases these effects lead to broad absorption bands in the UV–vis region.

The wavelength maximum at which an absorption band occurs in the UV–vis region is generally referred to as the λ_{max} of the sample (where wavelength is determined by the band maximum).

The quantitative relationship of absorbance (the intensity of a band) to concentration is expressed by the Beer–Lambert equation:

$$A = \log\frac{I_0}{I} = \varepsilon c l$$

where

A = absorbance, expressed as I_0/I
I_0 = the intensity of the incident light
I = the intensity of the light transmitted through the sample
ε = molar absorbtivity, or the extinction coefficient (a constant characteristic of the specific molecule being observed); values for conjugated dienes typically range from 10,000 to 25,000
c = concentration (mol/L)
λ = length of sample path (cm)

The calculated extinction coefficient and solvent are usually listed with the wavelength at the band maximum. For example, data for methyl vinyl ketone (3-buten-2-one) would be reported as follows:

$$\lambda_{\text{max}}\ 219\ \text{nm}\ (\varepsilon = 3600,\ \text{ethanol})$$
$$\lambda_{\text{max}}\ 324\ \text{nm}\ (\varepsilon = 24,\ \text{ethanol})$$

Typical UV–vis spectra are shown in Experiments [6], [19D], [A2$_a$], and [A3$_a$]. As part of the characterization data, UV–vis information is also given in Experiments[16], [A1$_a$], [19A], and [33A].

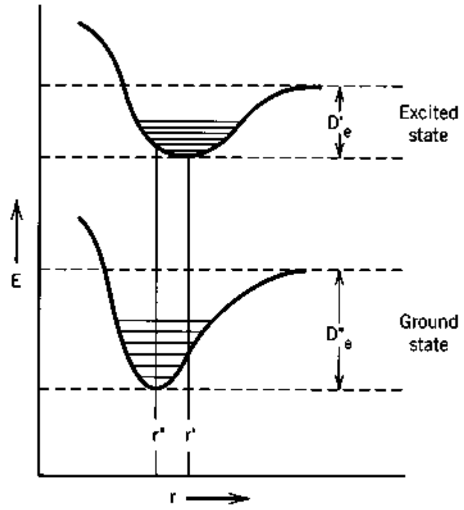

Figure 8.41 Two electronic energy levels in a diatomic molecule.

APPLICATION TO ORGANIC MOLECULES

In organic compounds containing *conjugated* systems of π electrons, a particular *chromophore* present can often be identified by the use of UV–vis spectroscopy. A *chromophore* is, in this case, a group of atoms able to absorb light in the UV–vis region of the spectrum. Since the electronic transitions involved are limited primarily to π-electron (and lone-pair) systems, this type of spectroscopy is less commonly used than the other modern spectroscopic techniques—which, in fact, it predates by several decades. Ultraviolet–visible spectroscopy, however, can play a valuable role in certain situations. For example, if a research problem involves synthesizing a series of derivatives of a complex organic molecule that possesses a strong chromophore, the UV–vis spectrum will be highly sensitive to structural changes involving the arrangement of the π-electron system (see, e.g., Experiment [6]).

In a conjugated alkene, such as 1,3-butadiene, the long-wavelength photon absorbed corresponds to the energy required for the excitation of a π electron from the HOMO, π_2 to the LUMO, π^*_3. For these alkenes, this transition is represented as $\pi \rightarrow \pi^*$; that is, an electron is promoted from a π (bonding) molecular orbital to a π^* (antibonding) orbital. This type of excitation is depicted below for both ethylene and 1,3-butadiene. Note that as a consequence of extending the chromophore and raising the energy of the highest occupied level in butadiene, the energy gap between the HOMO and LUMO levels of ethylene is larger than that in the conjugated system. Thus, the photon required for excitation of ethylene has a higher energy (higher frequency = shorter wavelength, $\lambda_{max} = 171$ nm) than the photon absorbed by 1,3-butadiene($\lambda_{max} = 217$ nm):

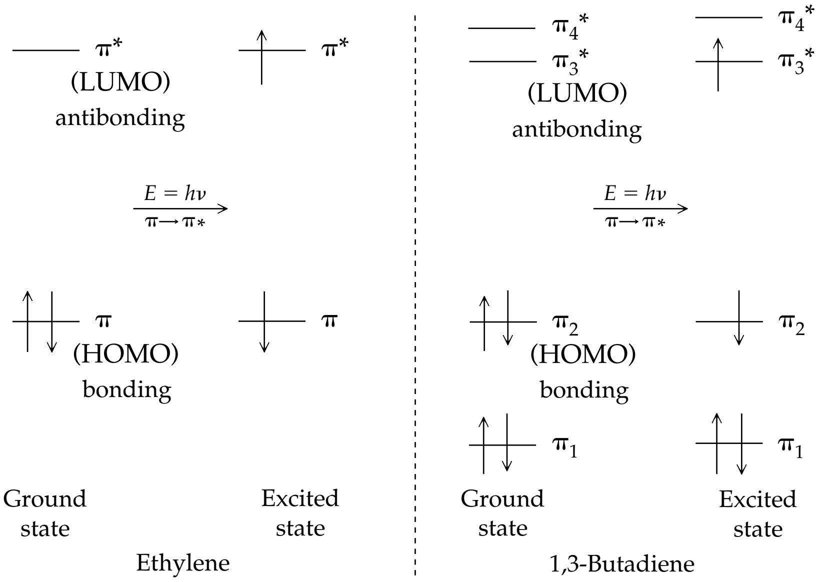

If we continue to extend the chromophore, the decrease of the energy gap between the HOMO and LUMO levels also continues. This drop in ΔE is then reflected in a drop in energy of the photon required to excite the $\pi \rightarrow \pi^*$ transition. This effect is illustrated in Table 8.28.

As the extension of the chromophore continues, the λ_{max} of the $\pi \rightarrow \pi^*$ transition will eventually shift into the visible region. At this point the substance exhibits color. Because the absorbed wavelength is coming from the blue end of the visible spectrum, these compounds will appear yellow. The color will deepen and become red as the energy of the photon required for

Table 8.28 Absorption Maxima of Conjugated Alkene

Name	Structure	λ_{max} (nm)
Ethylene	$CH_2{=}CH_2$	165
1,3-Butadiene	$CH_2{=}CH{-}CH{=}CH_2$	217
1,3,5-Hexatriene	$CH_2{=}CH{-}CH{=}CH{-}CH{=}CH_2$	268
1,3,5,7-Octatetraene	$CH_2{=}CH{-}CH{=}CH{-}CH{=}CH{-}CH{=}CH_2$	290

electronic excitation continues to drop. For example, tetraphenylcyclopentadienone is purple (Experiment [A3$_a$]); the dye, methyl red, is deep red (Experiment [26]); and *trans*-9-(2-phenylethenyl)anthracene is golden yellow (Experiment [19D]).

Compounds that contain a carbonyl chromophore 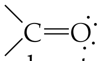 also absorb radiation in the UV region. A π electron in this unsaturated system undergoes a $\pi \rightarrow \pi^*$ transition. However, unless the carbonyl is part of a more extended chromophore, such as an α,β-unsaturated ketone system, the $\pi \rightarrow \pi^*$ transition requires a fairly high-energy photon for excitation, usually below 190 nm in the far-UV and similar to the energy required for excitation of a carbon–carbon double bond. The edge of the $\pi \rightarrow \pi^*$ absorption band may just barely be observed on instrumentation designed for near-UV studies. This partially observed absorption band is generally referred to as *end absorption.* In the case of carbonyls, however, the heteroatom also loosely holds two pairs of nonbonding electrons that are often termed *lone-pair* electrons. These nonbonding electrons reside in orbitals (*n*) that are higher in energy than the bonding π orbital, but lower in energy than the antibonding π* orbital. Thus, while a transition from an *n* level to a π* level is formally forbidden, in fact, weak bands are observed at λ_{max} in the near-UV that have their origin in the excitation of a lone-pair electron by an $n \rightarrow \pi^*$ transition. An energy diagram of a typical carbonyl system follows:

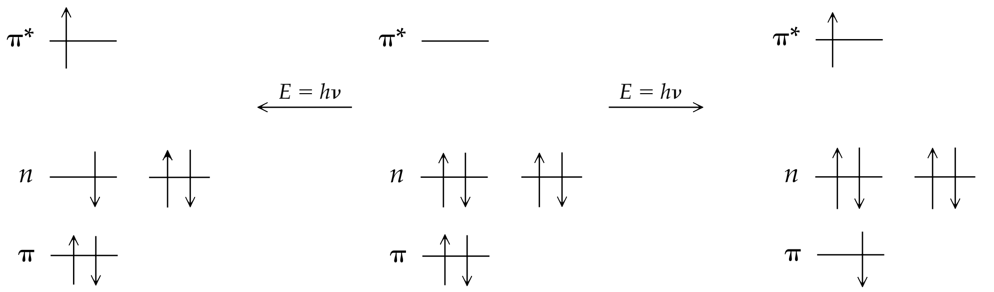

Excited state resulting from promotion of a nonbonding electron to the antibonding π* orbital

Ground state

Excited state resulting from promotion of a π electron to the antibonding π* orbital

Thus, those substances that contain the carbonyl chromophore absorb radiation of wavelengths that corresponds to both the $n \rightarrow \pi^*$ and the $\pi \rightarrow \pi^*$ transitions. For a simple ketone, such as acetone (CH_3COCH_3), the $\pi \rightarrow \pi^*$ transition is found in the far-UV and the $n \rightarrow \pi^*$ in the near-UV. When the carbonyl becomes part of an extended chromophore, such as in methyl vinyl ketone (3-buten-2-one), the spectra reveal that these two transitions have shifted to longer wavelengths—a bathochromic shift (see Fig. 8.42 for the

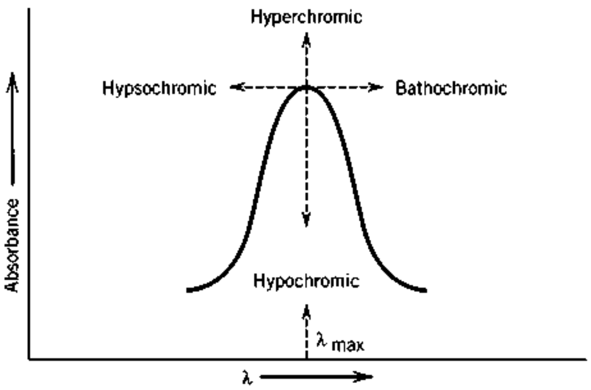

Figure 8.42 Terms describing direction of wavelengths and intensity shifts.

definition of terms used in UV–vis spectra to indicate the direction of wavelength and intensity shifts):

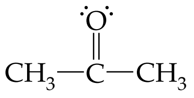

$CH_3-C(=O)-CH_3$

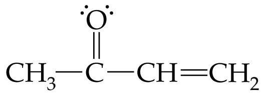

$CH_3-C(=O)-CH=CH_2$

	CH_3COCH_3		$CH_3COCH=CH_2$	
$n \Longrightarrow \pi^*$	λ_{max} 270 nm	ε_{max} 16	λ_{max} 324 nm	ε_{max} 24
$\pi \Longrightarrow \pi^*$	λ_{max} 187 nm	ε_{max} 900	λ_{max} 219 nm	ε_{max} 3600

Saturated systems containing heteroatoms with nonbonded electrons also exhibit weak absorption bands, often as end absorptions, which have their origin in forbidden $n \rightarrow \sigma^*$ transitions. When these heteroatomic groups are attached to chromophores, both the wavelength and the intensity of the absorption can be altered. These are often referred to as *auxochromes* and *auxochromic shifts.*

Often, model compounds containing a chromophore of interest are referred to as an aid in the interpretation of the UV–vis spectrum of a new structure. Substantial collections of data have been developed for a wide variety of chromophores as an aid to this type of correlation. A number of empirical correlations, such as the Woodward–Fieser rules, of substituent effects on λ_{max} values are available. The Woodward–Fieser rules are a set of empirical correlations derived from studies of UV–vis spectral data. Using these rules it is possible to predict with reasonable accuracy the λ_{max} for *new* systems containing various substituents on *known* chromophores. The rules are summarized in Table 8.29.

Examples of homoannular and heteroannular dienes are shown below:

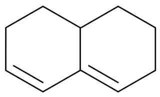

A heteroannular diene

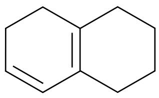

A homoannular diene

An example illustrating the use of these rules follows. Calculate the wavelength at which the following steroidal methyl sulfide will absorb:

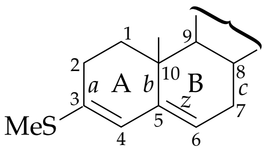

Table 8.29 **Woodward–Fieser Rules for Conjugated Dienes**

Functionality	Increment (nm)
Base value for homoannular diene	253
Base value for heteroannular diene	214
Add:	
For each double bond extending conjugation	+30
For double bond outside of ring (exocyclic)	+5
For alkoxy groups	+6
For *S*-alkyl groups	+30
For Cl, Br groups	+5
For dialkylamino groups	+60
For parts of rings attached to butadiene fragment	+5

The base value for the diene is 214 nm, because the system is heteroannular (if a homoannular diene were present it would take precedence over the heteroannular diene; see the following example). There are three ring residues (or alkyl substituents) attached to the chromophore. Through hyperconjugation, the π system is slightly extended by this type of substitution. The residues are labeled *a, b,* and *c.* Each of these substituents is assumed to add 5 nm to the λ_{max} of the parent heteroannular diene, for a total of 15 nm. The 5,6-double bond in the B ring marked *z* is exocyclic to the A ring, so empirically we add an additional 5 nm. Finally, for the thiomethyl substituent at the 3 position we add 30 nm. The total is 214 + 15 + 5 + 30 = 264 nm.

Thus we have

Predicted value λ_{max} (calcd) = 264 nm

Observed value λ_{max} (obsd) = 268 nm (ε = 22,600)

As another example, consider ergosta-3,5,7,9-tetraene-3-acetate (**I**):

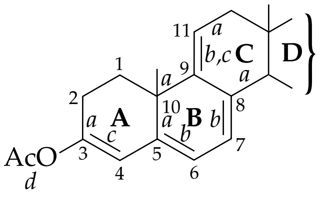

I

Prediction of λ_{max} for a homoannular diene

Parent homoannular diene in ring **B**		253 nm
Increments for		
Double bond extending conjugation	*c* [2 × 30]	60
Alkyl substituent or ring residue	*a* [5 × 5]	25
Exocyclic double bond	*b* [3 × 5]	15
Polar substituents	*d* [0]	0
	λ_{calcd}	353

Predicted value λ_{max} (calcd) = 353 nm

Observed value λ_{max} (obsd) = 355 nm (ε = 19,700)

Table 8.30 Conjugated Carbonyl Systems

α,β-Unsaturated Functionality			Base Value (nm)
Acyclic or six-membered or higher cyclic ketone			215
Five-membered ring ketone			202
Aldehydes			210
Carboxylic acids and esters			195
			Increment (nm)
Extended conjugation			+30
Homoannular diene			+39
Exocyclic double bond			+5
	Substituent Increment (nm)		
Substituent	α	β	δ
Alkyl	+10	+12	+18 (γ and higher)
Hydroxyl	+35	+30	+50
Alkoxy	+35	+30	+31 (γ + 17)
Acetoxy	+6	+6	+6
Dialkylamino		+95	
Chloro	+15	+12	
Bromo	+25	+30	
Alkylthio		+85	
Solvent			**Solvent Increment (nm)**
Water			−8
Ethanol			0
Methanol			0
Chloroform			+1
Dioxane			+5
Ether			+7
Hexane			+11
Cyclohexane			+11

There are additional rules for carbonyl-containing compounds, such as ketones, aldehydes, carboxylic acids, and so on, and for aromatic compounds. Table 8.30 lists the parameters for conjugated carbonyl systems. Note that in contrast to the conjugated diene compounds, in which we are observing $\pi \rightarrow \pi^*$ transitions, the $n \rightarrow \pi^*$ transitions of the carbonyl λ_{max} chromophore are often solvent dependent. Thus, solvent effects will have to be considered when predicting λ_{max} values in these systems.

An example of λ_{max} (ethanol) calculation for a carbonyl system is presented here:

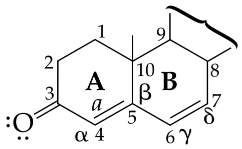

The base value for the α,β-unsaturated six-membered ring ketone system is 215 nm. Extended conjugation adds an additional 30 nm. The presence of an exocyclic double bond, marked *a*, extends the λ_{max} another +5 nm. There is asubstituent on the β-carbon atom (+12 nm) and on the δ-carbon atom (+18 nm). There is no solvent effect because the spectrum was obtained in ethanol (0 shift). The total is 215 + 30 + 5 + 12 + 18 = 280 nm:

Predicted value	λ_{lmax} (calcd) = 280 nm
Observed value	λ_{lmax} (obsd) = 284 nm

The Woodward–Fieser rules work well for systems with four or fewer double bonds. For more extensively conjugated systems, λ_{max} values are more accurately predicted using the Fieser–Kuhn equation:

$$\text{Wavelength} = 114 + 5\,M + n(48.0 - 1.7n) - 16.5\,R_{\text{endo}} - 10R_{\text{exo}}$$

where

n = number of conjugated double bonds
M = number of alkyl substituents in the conjugated system
R_{endo} = number of rings with endocyclic double bonds in the system
R_{exo} = number of rings with exocyclic double bonds in the system

Sample calculation: Find the UV λ_{max} of β-carotene:

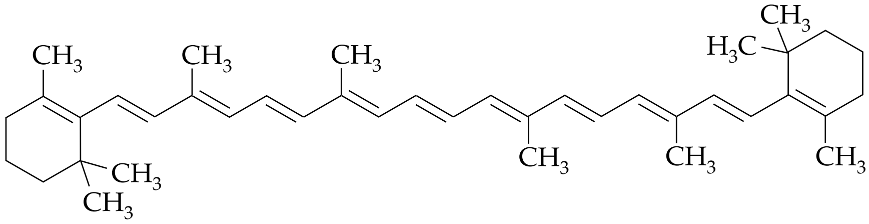

β-Carotene

In the structure there are 11 conjugated double bonds, $n = 11$. There are 6 alkyl groups and 4 ring residues on the conjugated system, $M = 10$. Both rings have an endocyclic double bond, $R_{\text{endo}} = 2$ Neither ring has any exocyclic double bonds, therefore $R_{\text{exo}} = 0$. Substituting in the equation gives

$$\begin{aligned}\text{Wavelength} &= 114 + 5(10) + 11[48 - 1.7(11)] - 16.5(2) - 10(0)\\ &= 114 + 50 + 322.3 - 33 - 0 = 453\text{ nm}\end{aligned}$$

Predicted value	λ_{lmax} (calcd) = 453 nm
Observed value	λ_{max} (obsd) = 455 nm

Two examples of this correlation scheme (see Table 8.31) are

1. 6-Methoxytetralone

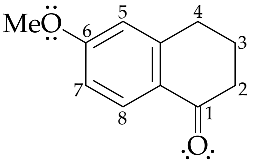

Predicted λ_{max} is calculated by taking

Parent value, 246 nm + one *o*-ring residue, 3 + one *p*-OMe, 25 = 274 nm

Predicted value	λ_{max} (calcd) = 274 nm
Observed value	λ_{max} (obsd) = 276 nm (ε = 16,500)

Table 8.31 **The Benzoyl Chromophore**

Parent Chromophore C_6H_5—CO—R

Function	Wavelength (nm)
R = alkyl or ring residue	246
R = H	250
R = OH, *O*-alkyl	230

Substituent	Substituent Increment (nm) *o*-	*m*-	*p*-
Alkyl or ring residue	3	3	10
—OH, —OCH_3, —*O*-alkyl	7	7	25
—O^- (*p*-sensitive to steric effects)	11	20	78
—Cl	0	0	10
—Br	2	2	15
—NH_2	13	13	58
—NHAc	20	20	45
—$NHCH_3$			73
—$N(CH_3)_2$	20	20	85

Note. Spectra obtained in alcohol solvents.

2. 3-Carboethoxy-4-methyl-5-chloro-8-hydroxytetralone

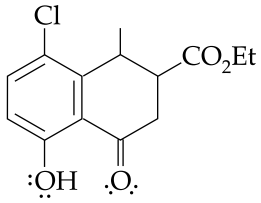

Parent value, 246 nm + one *o*-ring residue, 3 + one *o*-OH, 7 + 0 *m*-Cl = 256 nm

Predicted value λ_{max} (calcd) = 256 nm
Observed value λ_{max} (obsd) = 257 nm (ε = 8000)

In summary, UV–vis spectra can make substantial contributions to understanding the molecular structure of organic substances that possess *chromophores:*

1. Interpretation of ultraviolet–visible spectra often can be a powerful approach for identifying the molecular structure of that section of a new substance that contains the chromophore.

2. The λ_{max} increases within a series of compounds that contain a common chromophore that is lengthened (increased conjugation) over the series. The intensity of the absorption (ε_{max}) also generally becomes greater as conjugation increases, but can be very sensitive to steric effects (see Experiment [6]).

3. The λ_{max} is sensitive to hyperconjugation by alkyl substituents, conformational changes that restrict π-system overlap, configurational, or geometric isomerization in which π systems are perturbed, and structural changes, such as the isomerization of a double bond from an *exocyclic* to an *endocyclic* position and changes in ring size.

Table 8.32 Absorption Maxima of Several Unsaturated Molecules

Compound	Structure	λ_{max} (nm)	ε_{max}
Ethylene	$CH_2{=}CH_2$	171	15,530
1,3-Butadiene	$CH_2{=}CH{-}CH{=}CH_2$	217	21,000
Cyclopentadiene	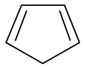	239	3,400
1-Octene	$CH_3(CH_2)_5CH{=}CH_2$	177	12,600
trans-Stilbene	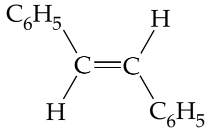	295	27,000
cis-Stilbene	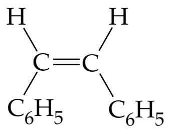	280	13,500
Toluene	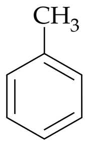	189	55,000
		208	7,900
		262	260
4-Nitrophenol	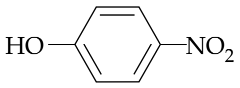	320	9,000
3-Penten-2-one	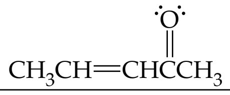 $CH_3CH{=}CH\overset{\overset{\ddot{O}}{\|}}{C}CH_3$	220	13,000
		311	35

4. In many instances, accurate prediction of the λ_{max} of a new molecular system can be made based on empirical correlations of the parent chromophore giving rise to the absorption.

Table 8.32 lists the λ_{max} values of a number of common organic molecules.

INSTRUMENTATION

The acquisition of UV–vis absorption spectra for use in the elucidation of organic molecular structure is now carried out with instrumentation that is typically an automatic-recording photoelectric spectrophotometer. The optical components of one of the classic spectrophotometers is given in Figure 8.43. This system is typical of a high-quality double-beam double-monochromator instrument. The instrument consists of a number of components: the radiation source, monochromator, sample compartment, detector, amplifier, and recorder.

The Source of Radiation

Radiant energy may be generated by either a deuterium discharge lamp or a tungsten– halogen lamp depending on the spectral region to be observed. Deuterium is generally preferred over hydrogen since the intense radiating ball of plasma is slightly larger in the case of deuterium, and therefore source

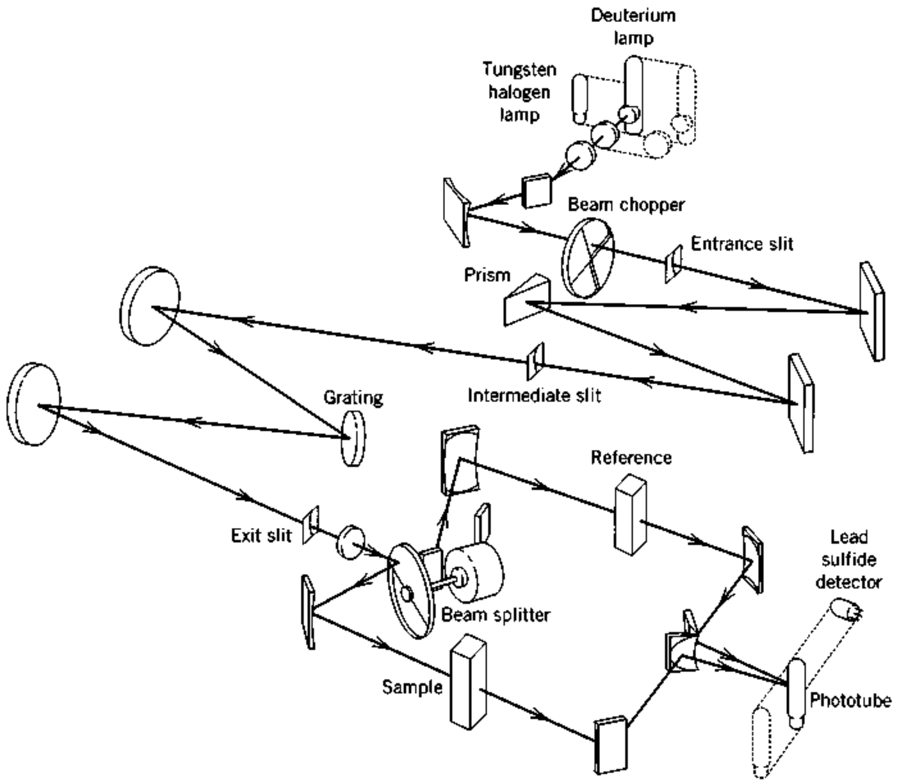

Figure 8.43 Schematic optical diagram of a double beam-in-time spectrophotometer with double monochromation (Cary Model 17D). (Courtesy of Varian Associates, Inc.)

brightness is enhanced by a factor of about 4. Below 360 nm, deuterium gas emits an intense continuum band that covers a major portion of the UV region. With special windows the short wavelength cutoff can be extended down to about 160 nm well out into the vacuum-UV. Emission line spectra limit the long wavelength use of these lamps to about 380 nm. The lamps of choice for the region above 350 nm (the visible) are incandescent filament lamps, because they emit a broad band of radiation from 350 nm on the short wavelength end all the way to about 2.5 μm (the near-IR) on the long wavelength side. Most of the radiation emitted falls outside the visible, peaking at about 1 μm in the near-IR. Nevertheless, tungsten lamps are *the* choice for measurements in the visible region, because they are extremely stable light sources.

Thus, radiation sources must possess two basic characteristics: (1) they must emit a sufficient level of radiant energy over the region to be studied so that the instrument detection system can function, and (2) they must maintain constant power during the measurement period. Source power fluctuations can result in spectral distortion.

The Monochromator

As the name implies, a monochromator (making a single color or hue) functions to isolate a single frequency from the source band of radiation. In practice we settle for isolating a small collection of overlapping frequencies surrounding the monochrome radiation we wish to observe. Thus, the monochromator section of the instrument takes all the source radiation in at one end and releases a very narrow set of bands of radiation at the other end. This function is accomplished, as shown in Figure 8.43, by focusing the entering radiation on an entrance slit that forms a narrow image of the source. After passing through the entrance slit, the spreading radiation is collimated by being reflected off a parabolic mirror, and is converted into parallel light rays (just as in a search light). The collimated radiation is then directed to the dispersing agent, which is usually a quartz prism (quartz is transparent to UV, glass is not) or diffraction grating. The dispersing device spreads the different wavelengths of collimated light out in space. After emerging from the prism the dispersed radiation is redirected to either the same or a new collimator mirror and refocused as an image of the source on the exit slit of the monochromator. The exit slit has only a small fraction of the original radiation focused on it, and allows it to pass through in the image of the source. The remaining frequencies lie at different angles on either side of the exit slit. By mechanically turning the prism or grating, and thus changing the angle of the dispersing device with respect to the exit slit, all of the narrowly dispersed bands of radiation can be passed out of the monochromator in sequential fashion.

Instruments that are designed to reduce unwanted radiation to an absolute minimum will place two monochromators in tandem with an intermediate slit connecting the dispersing systems. In the case illustrated in Figure 8.43 the first monochromator uses a prism, while the second uses a grating. The two monochromators, however, must be in perfect synchronization or no light at all will be transmitted.

Sample Compartment

After leaving the monochromator the radiation is directed to the sample compartment by a rotating sector mirror, where it is alternately focused on the substance to be examined (which is contained in a cell with quartz windows) and a reference cell (which holds the pure solvent used to dissolve the sample). The system now has two beams, hence the name *double-beam spectrophotometer.* After passing through the sample where the absorption of radiation may occur, the beams are recombined.

The sampling position could be placed either before or after the monochromator. In infrared instruments (such as the PE Model 710B, Fig. 8.3*a*) it was generally found before the monochromator until the introduction of interferometers. In UV systems, the sampling area is most often placed after the monochromator, and for good reasons. If the sample were placed before the monochromator, it would be exposed to the entire band of high-energy UV radiation being emitted by the source over the entire sampling period. By positioning the sample after the monochromator, at any one time the sample sees only the very small fraction of the dispersed radiation passed by the exit slit. Thus, sample stability is greatly protected by this arrangement. Remember that near-UV radiation carries photons with energies that approach those of the bond energies of organic molecules.

The Detector

The recombined beams are then focused on the detector. Detectors function as transducers because they convert electromagnetic radiation into electrical current. There are a number of radiation-sensitive transducers available as detectors for these instruments. One is the photomultiplier tube. These detectors operate with photocathodes that emit electrons in direct proportion to the number of photons striking the photosensitive surface and possess very large

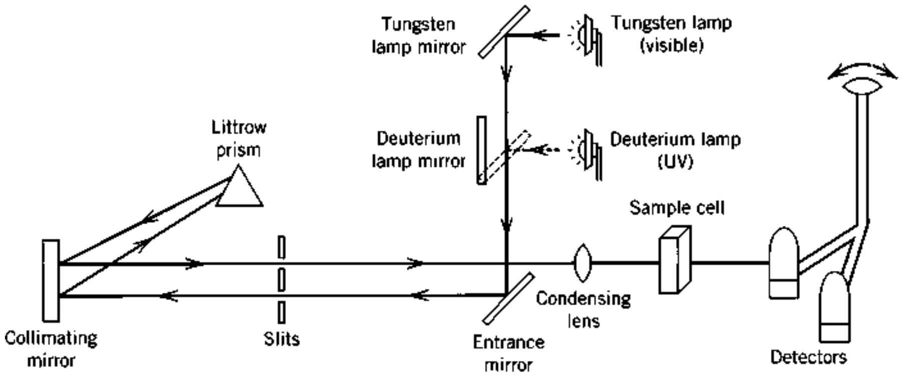

Figure 8.44 UV–visible single-beam spectrometer.

internal amplification. Thus, they operate at low power levels. One particular advantage of the photomultiplier is that you can adjust their sensitivity over a wide range simply by adjusting the supply voltage.

The Electronics: The Amplifier and Recorder

In double-beam instruments, the two signals generated by the sample and reference beams (each referenced against a dark current) in the detector are amplified and the ratio of the sample signal to the reference signal is plotted on a recorder. The simplest of the absorption spectrometers are the single-beam instruments (see Fig. 8.44). These spectrometers are generally employed for problems involving simple one-component analyses. The photometric accuracy of scanned spectra should not be of paramount importance with these systems. Single-beam spectrometers require extremely stable sources and detectors.

SAMPLE PREPARATION

Ultraviolet spectra are usually obtained on samples in solution using quartz cells. Quartz is used because it is transparent to both UV and visible light. For spectra restricted to the visible region, Pyrex cells are satisfactory (and a good deal less expensive), but because Pyrex absorbs UV radiation, these cells cannot be employed for measurements in this region.

Solution cells usually have a horizontal cross section of 1 cm^2 and require about 3 mL of sample solution. Cells must be absolutely clean, and it is advisable to rinse the cell several times with the solvent used to dissolve the sample. A background spectrum of the solvent-filled cell (*without* a reference sample) can easily be obtained at this time and used as a check against contamination of either the cell or the solvent or both.

Because the intensities of electronic transitions vary over a very wide range, the preparation of samples for UV–vis spectra determination is highly concentration dependent. Intense absorption can result from the high molecular extinction coefficients found in many organic chromophores. The sampling of these materials requires very dilute solutions (on the order of 10^{-6}–10^{-4} M). These solutions can be conveniently obtained by the technique of *serial dilution.* In this method a sample of the material to be analyzed is accurately weighed, dissolved in the chosen solvent, and

Table 8.33 **Solvents Used in the Near-UV**

Solvent	Cutoff Wavelength (nm)
Acetonitrile	190
Chloroform	245
(toxic, substitute CH_2Cl_2)	235
Cyclohexane	205
1,4-Dioxane	215
(toxic, substitute EtOEt)	218
95% Ethanol	205
n-Hexane	195
Methanol	205
Isooctane	195
Water	190

Note. Since these solvents have no color, they are transparent in the visible.

diluted to volume in a volumetric flask. Sample weights of 4–5 mg in 10-mL volumetric flasks are typical. An aliquot is then taken from this original solution, transferred to a second volumetric flask, and diluted as before. This sequence is repeated until the desired concentration is obtained.

Numerous choices of solvent are available (a list is given in Table 8.33) and most of them are available in "spectral grade." The most commonly used solvents are water, 95% ethanol, methanol, and cyclohexane.

Criteria for Choosing a Solvent

- The most important factor is solubility of the sample. UV–vis spectra can be very intense, so even low solubility may be quite acceptable in sample preparation.
- The wavelength cutoff for the solvent may be important if the sample absorbs below about 250 nm.
- Sample–solvent molecular interactions must be considered. An example of these effects would be hydrogen bonding of protic solvents with carbonyl systems. Hydrocarbon chromophores are less influenced by solvent character than are the more polar chromophores.

BIBLIOGRAPHY

American Petroleum Research Institute Project 44 *Selected Ultraviolet Spectral Data,* Vols. I–IV; Thermodynamics Research Center, Texas A&M University: College Station, TX, 1945–1977 (1178 compounds).

Feinstein, K. *Guide to Spectroscopic Identification of Organic Compounds;* CRC Press: Boca Raton, FL, 1995.

Field, L. D.; Sternhell, S.; Kalman, J. R. *Organic Structures from Spectra,* 4th ed.; Wiley: New York, 2008.

Grasselli, J. G.; Ritchey, W. M. *Atlas of Spectral Data and Physical Constants for Organic Compounds,* 2nd ed.; CRC Press: Cleveland, OH, 1975.

Harwood, L. M.; Claridge, T. D. W. *Introduction to Organic Spectroscopy;* Oxford University Press: New York, 1997.

Hesse, M; Meier, H; Zeeh, B. *Spectroscopic Methods in Organic Chemistry,* 2nd ed; Thieme Medical Publishing: New York, 2008.

Lang, L., Ed.; *Absorption Spectra in the Ultraviolet and Visible Region,* Vols. 1–20; Academic Press: New York, 1961–1975; Vols. 21–24; Kreiger: New York, 1977–1984.

Pavia, D. L.; Lampman, G. M.; Kriz, G. S. *Introduction to Spectroscopy: A Guide for Students of Organic Chemistry,* 3rd ed.; Saunders College: Philadelphia, 2000.

Pretsch, E.; Clerc, J. T. *Spectra Interpretation of Organic Compounds;* VCH: Wiley: New York, 1997.

Rouessac, F.; Rouessac, A. *Chemical Analysis: Modern Instrumental Methods and Techniques,* 2nd ed.; John Wiley & Sons; New York, 2007.

Standard Ultraviolet Spectra; Sadtler Research Laboratories: Philadelphia.

UV Atlas of Organic Compounds, Vols. I–IV; Butterworths: London, 1966–1971.

Williams, D. H.; Fleming, I. *Spectroscopic Methods in Organic Chemistry,* 6th ed.; WCB/McGraw-Hill: New York, 2007.

MASS SPECTROMETRY*

In comparison with other forms of spectroscopy, such as NMR, IR, or UV–vis, mass spectrometry is unique in terms of how we generate and interpret the spectrum. Instead of monitoring the absorption of electromagnetic radiation in terms of frequency or wavelength, a mass spectrum can be thought of as a snapshot of a rather unconventional organic reaction involving one energetic reactant that decomposes to give a variety of reaction products. By characterizing the composition of this "reaction mixture" we are able to learn the identity of the starting reactant.

The reaction gets started when a molecule in the gas phase is converted to a radical cation by an energetic collision with an electron, as shown below for N_2O:

$$N_2O_{(g)} + e^- \longrightarrow N_2O^{+\cdot} + 2e^-$$

The fact that we have formed a positive ion (cation) with an unpaired electron (radical) becomes important for understanding the decomposition reactions. The process of forming the radical cation yields a collection of energized *molecular ions* that contain a range of internal energies. The molecular ion is produced in a low-pressure environment where it is unable to bump into other molecules. Fragmentation (bond breaking) results in the formation of charged and neutral products. Ideally, intramolecular rearrangements, which could complicate determination of the molecules original structure, do not occur. Depending on characteristics of the molecule and the amount of energy deposited, a variety of fragmentation reactions can take place. For example, the molecular ion ($N_2O^{+\cdot}$) may fragment to give the following products through one- or two-step reactions:

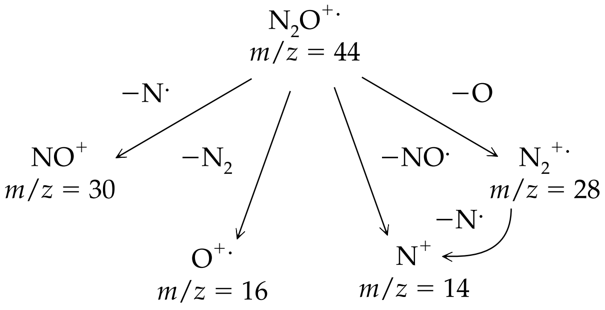

---

[*] This section has been written by Elizabeth A. Stemmler, Professor of Chemistry, Bowdoin College.

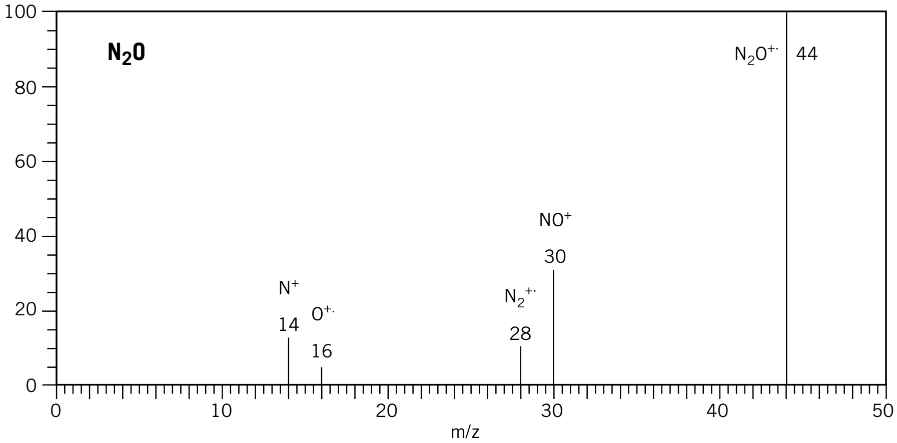

Figure 8.45 Electron ionization mass spectrum of nitrous oxide, N_2O.

Mass spectrometry derives its name from its ability to distinguish the molecular ion and the different charged reaction products based on the ratio of the ion mass to its charge (*m/z* ratio). In most cases, the charge, *z*, is equal to 1 and we can easily tell the difference between the molecular ion (*m/z* = 44) and products, such as N_2^+ (*m/z* 28). In addition, a mass spectrometer can be used to determine the relative amounts of molecular ions and fragment ions present after the reaction has had a little (very little!) time to proceed. A mass spectrum, typically shown in a bar-graph format, is a display of the relative number of each type of ion plotted as a function of the *m/z* ratio (see Fig. 8.45). Instead of reporting the actual number of each type of ion, we normalize the data and give the most abundant ion a value of 100%. Note that only charged species are detected by a mass spectrometer. The neutral products are not observed and their identity must be inferred.

Mass spectrometry is useful to organic chemists because of the information it provides about molecular structure. For example, if the molecular ion is present, that peak can be used to determine the molecular weight (MW) of the neutral molecule. With precise measurement of the *m/z* ratio (to ± 0.0001,for example), the elemental formula of the molecular ion can be determined. For example, N_2O and CO_2 both have a molecular weight of 44; however, measurement of their exact masses (44.0011 and 43.9898, respectively) can be used to assign their elemental formula. Mass spectrometry was used to originally determine the exact mass of each element (see Table 8.34), and these exact masses, *not* the atomic weights, are used to calculate mass.

Even when precise mass measurements are not available, the products in the mass spectrum may provide enough information to determine the structure of the neutral molecule. Interpretation of a mass spectrum involves working backward from the observed charged fragments to a proposed molecular structure. For example, CO_2 or N_2O (same MW) could be distinguished by an examination of the mass spectrum. The fragment ions at *m/z* 14 and 30 in Figure 8.45 clearly eliminate CO_2 as a possible structure. There are no combinations of carbon (mass = 12) and oxygen (mass = 16) that could produce these ions. In addition, the mass spectrum allows us to distinguish between the isomers NNO and NON. What would the mass spectrum of NON show? We would expect to see only a peak at *m/z* 30 (NO^+) and no signal at *m/z* 28 (N_2^+). The mass spectrum would thus support your chemical intuition that NON is an unlikely and unstable molecular structure.

Table 8.34 **Exact Masses and the Atomic Weights for Isotopes of Some Common Elements**

Element	Nuclide	Mass	Atomic Weight[a]
Hydrogen	^{1}H	1.0078	1.0079
	D(^{2}H)	2.0141	
Carbon	^{12}C	12.00000 (std)	12.011
	^{13}C	13.0034	
Nitrogen	^{14}N	14.0031	14.0067
	^{15}N	15.0001	
Oxygen	^{16}O	15.9949	15.9994
	^{17}O	16.9991	
	^{18}O	17.9992	
Fluorine	^{19}F	18.9984	18.9984
Silicon	^{28}Si	27.9769	28.0855
	^{29}Si	28.9765	
	^{30}Si	29.9738	
Phosphorus	^{31}P	30.9738	30.9738
Sulfur	^{32}S	31.9721	32.066
	^{33}S	32.9715	
	^{34}S	33.9679	
Chlorine	^{35}Cl	34.9689	35.4527
	^{37}C1	36.9659	
Bromine	^{79}Br	78.9183	79.904
	^{81}Br	80.9163	
Iodine	^{127}I	126.9045	126.904

[a] Average mass of the naturally occurring isotopes of the element; *not* used for mass calculations in mass spectrometry.

As you will see below, mass spectral interpretation is not always as straightforward as the case given above. Like the outcome of an organic reaction, a mass spectrum will reflect the outcome of competing sequential and simultaneous reaction pathways. For some molecules, very little fragmentation will take place and only the molecular ion is observed. For other less stable molecules, we may have complete conversion of the molecular ion to products, although, because of the high energy required for their formation, we will rarely see complete fragmentation down to products at the atomic level. The interpretation of a mass spectrum requires developing an understanding of important, characteristic reaction pathways and an appreciation of factors influencing ion stability. In many ways, the interpretation of mass spectra provides a place to apply principles of organic chemistry to a unique kind of chemical reaction.

INSTRUMENTATION

All mass spectrometric instruments contain regions where ionization, mass analysis, and ion detection take place. Mass spectrometry takes place at low pressure; all of the mass spectrometric components are contained in a vacuum

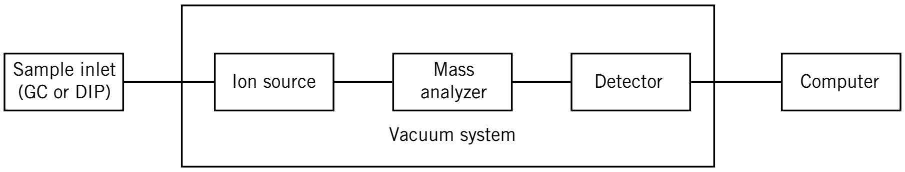
Sample inlet
(GC or DIP)
Ion source
Mass
analyzer
Detector
Computer
Vacuum system

Figure 8.46 Block diagram of components of a mass spectrometer.

system at pressures of 10^{-7} to 10^{-5} torr. Because the instrument must be sealed from the atmosphere to maintain the low pressure, and because samples must be converted to the gas phase prior to ionization, all mass spectrometers have a region devoted to sample introduction. In this region the sample—in the form of a solid, liquid, or gas—is transferred to the low pressure of the mass spectrometer, while preventing the introduction of air. A block diagram of a basic mass spectrometer is shown in Figure 8.46. Ions are generated and fragment in the ion source; the molecular ion and fragments are separated, based upon *m/z* ratios, in the mass analyzer; and the ion signals are converted by the detector into a signal that may be input to a computer.

Ion Source

The ion source is the region where ions are generated. Mass spectrometrists have many ways of creating ions from different types of samples, including biological materials or the surface of a particle. In our discussions, we will focus only on the most common ionization method, electron ionization (EI). In an EI ion source (Fig. 8.47), we send current through a wire, called a filament. As the filament gets hot, electrons are emitted from the surface. The electrons are produced in an electric field, which results in electron acceleration through

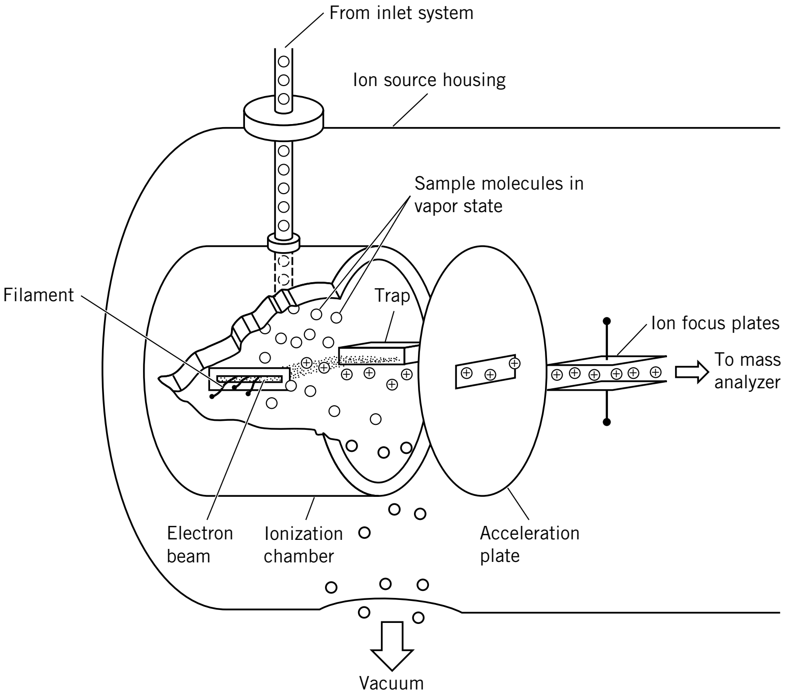
From inlet system
Ion source housing
Sample molecules in
vapor state
Filament
Trap
Ion focus plates
To mass
analyzer
Electron
beam
Ionization
chamber
Acceleration
plate
Vacuum

Figure 8.47 Schematic diagram of an electron ionization source. (From Watson, J.T. *Introduction to Mass Spectrometry,* 3rd ed.; Lippincott-Raven Publishers: Philadelphia, 1997, p. 140.)

the ion source region where the sample vapor is found. If you work with an electron energy that is too low (below the ionization potential for the molecule), no ions will be produced. As the electron energy increases, the molecular ion ($M^{+\cdot}$) will appear. With further increases in electron energy, fragment ions are observed. Formation of doubly charged ions ($M^{2+\cdot}$) occurs, but the ion intensities are very low:

$$\begin{tikzcd}
 & M^{+\cdot} + 2e^{-} & \\
M + e^{-}(70\text{ eV}) \arrow[ur] \arrow[r] \arrow[dr] & M^{2+} + 3e^{-} & \text{(rare)} \\
 & M^{-\cdot} & \text{(not observed)}
\end{tikzcd}$$

Mass spectra, by convention, are measured with 70-eV electrons. At this energy, the ion intensity is high and the distribution of products remains relatively constant with small changes in electron energy. Formation of negative molecular ions with 70-eV electrons does not occur.

Another important role for the ion source is directing the ions toward the mass analyzer. The ions are pushed and pulled as they pass through one or more metal plates that have a hole in the center for ion transmission. These plates accelerate the ions and keep them directed at the mass analyzer. Depending on the mass analyzer in use, the ions are accelerated toward the analyzer with high or low energy.

Mass Analyzer

Ion formation and fragmentation in the source is followed by mass analysis. Mass analyzers are used to separate ions based on their mass-to-charge ratios. Organic chemists commonly use two types of mass analyzers: magnetic sector instruments (low- and high-resolution) and quadrupole instruments. Magnetic sectors separate ions based on dispersion of the ions into beams with different *m/z* ratios; quadrupoles are mass filtering devices.

In a magnetic sector instrument, ions are accelerated out of the ion source into a magnetic field with high (kilovolt) kinetic energies. The magnet field, applied perpendicular to the path of the ions, exerts a force that causes the ions to follow a curved path through the magnet (Fig. 8.48). The extent to which the path is bent depends on the mass-to-charge ratio (more specifically, the momentum) of the ions. Light ions are bent more than heavier ions. If the path that the ions must travel is fixed, ions that are too light or too heavy will run into the walls of the mass analyzer, where they are neutralized, and will then be pumped away by the vacuum system. Only the ions with the correct radius (correct *m/z* ratio) will make it to the detector. To measure a complete mass spectrum, the magnetic field strength is varied to bring ions of different *m/z* ratio to focus on the detector.

High-resolution magnetic sector instruments incorporate an additional energy analyzer prior to mass analysis by the magnetic sector. This more precisely defines the kinetic energies of ions entering the magnetic sector, which improves the mass resolution. High-resolution instruments require more expertise to operate and are less common because of their expense, but they can provide the precise and accurate mass measurements needed to determine elemental composition.

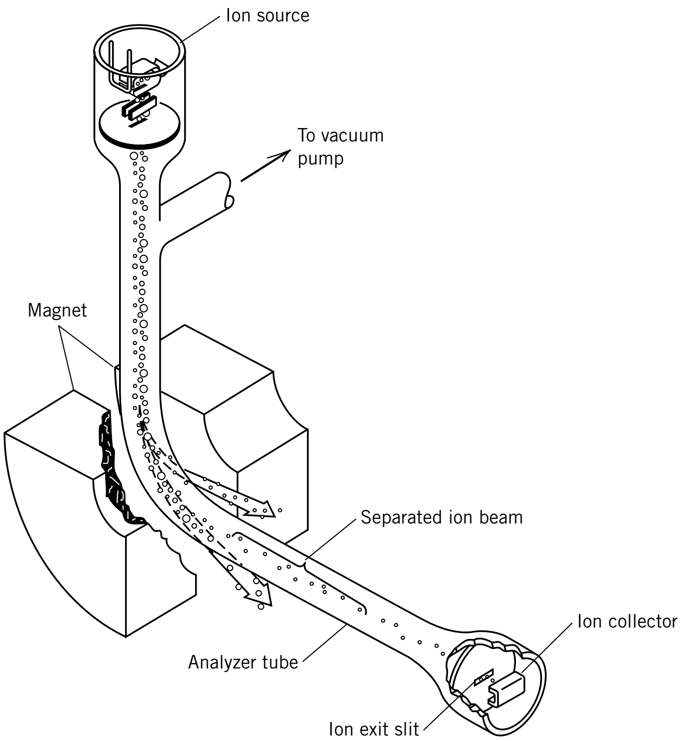
Ion source
To vacuum pump
Magnet
Separated ion beam
Ion collector
Analyzer tube
Ion exit slit

Figure 8.48 Schematic diagram of a magnetic sector mass analyzer. (From McLafferty, F.W.; Turevek, F. *Interpretation of Mass Spectra,* 4th ed.; University Science Books: Sausalito, CA, 1993, p. 8.)

A quadrupole is composed of a set of four rods to which (electric) potentials are applied (Fig. 8.49). To allow ions of a particular *m/z* ratio through the rods, a constant positive potential is applied to two opposing rods (the x-rods), while the remaining two rods experience a constant negative potential (the y-rods). In addition, each set of rods experiences a time-varying potential that causes the rod potentials to vary between positive and negative potentials, with the signals 180° out-of-phase. When the x-rods are positive, the y-rods

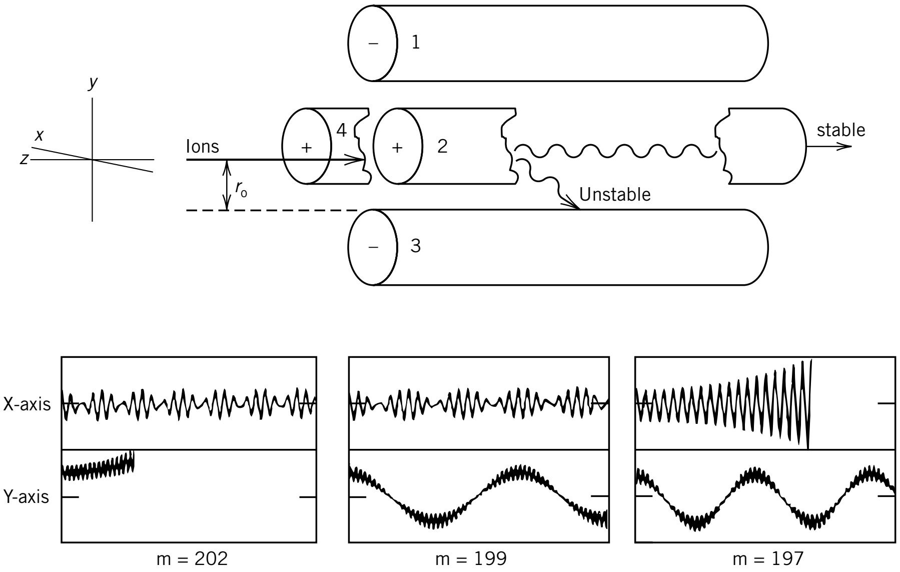
y
x
z
Ions
r_0
−
1
+
4
+
2
stable
Unstable
−
3
X-axis
Y-axis
m = 202
m = 199
m = 197

Figure 8.49 Schematic diagram of a quadrupole mass analyzer. *X* and *Y* axis trajectories for m/z 202, 199, and 197. (From Steel, C.; Henchman, M. *J. Chem. Educ.* **1998,** *75,* 1049–1054.)

are negative. Mass filtering occurs when a group of ions enters the analyzer. Ions that have *m/z* ratios that are too low or too high will experience unstable trajectories through the rods and will strike a rod, become neutralized, and be pumped out of the system. Only ions with an appropriate *m/z* ratio will have a stable trajectory and will make it through the rods to the detector. Figure 8.49 shows the trajectories of three ions with respect to the x- and y-rods. Only the ion with $m/z = 199$ makes it through the quadrupole. To change the *m/z* ratio of the ions that are transmitted, the magnitude of the constant and time-varying potentials are changed. Mass spectrometers that fit on a laboratory benchtop have a mass range of $m/z = 10$ to 650. With the quadrupole mass analyzer a mass spectrum can be measured rapidly (roughly 1 scan per second), which is important when capillary GC columns are used for sample introduction.

Detector

Ions can be detected directly through the current produced when they strike a plate; however, we usually make this signal larger through the use of electron multiplier detectors.

Tuning the Mass Spectrometer

Before the mass spectrometer can be used to collect mass spectra, the instrument must be tuned and calibrated. The tuning procedure involves setting voltages associated with the ion source, lenses, and detector (to optimize sensitivity), and selecting values for potentials applied to the quadrupole (to set the instrument resolution). These tasks are accomplished while a calibration standard is continuously added to the instrument. A common calibration standard is perfluorotributylamine (PFTBA), $(CF_3CF_2CF_2CF_2)_3N$. Usually ions at *m/z* 69, 219, and 502 are monitored (Fig. 8.50).

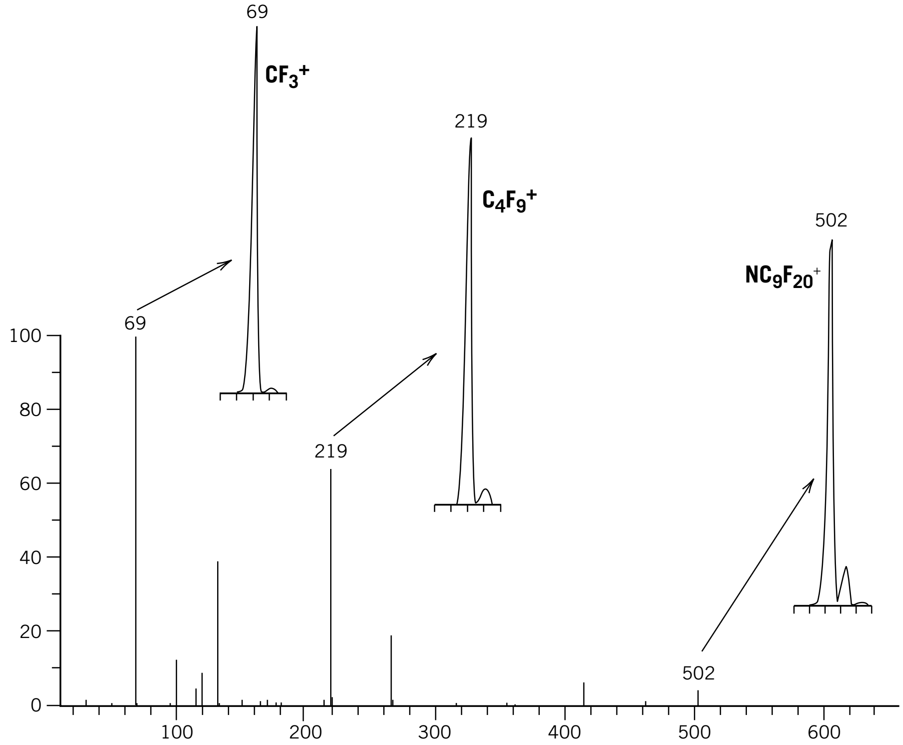

Figure 8.50 Electron ionization mass spectrum of perfluorotributylamine (PFTBA). Inserts show the peak profiles for *m/z* 69, 219, and 502.

Sample Introduction

Samples analyzed by EI mass spectrometry must be converted to gas phase. For pure gases or volatile liquids the samples may be introduced directly through a small orifice that allows an appropriate amount of material into the vacuum chamber. A small amount of a solid sample can be placed in a melting point capillary tube and inserted into the mass spectrometer at the end of a metal rod, called a *direct insertion probe* (DIP). The temperature at the tip of the probe can be varied to promote sublimation of the sample. Another common method of sample introduction is gas chromatography, which is the ideal choice for samples that are impure.

Gas Chromatography/Mass Spectrometry (GC/MS)

While the goal of a synthetic organic reaction is the production of one pure product in high yield, organic reactions often produce a mixture of reaction products. Chromatographic separation of those products is a useful complement to the mass spectral analysis. The components of the mixture elute from the chromatographic column, ideally, as pure peaks. The mass spectrometer, which is scanning rapidly, is then able to collect a few spectra for each eluting peak. Both the chromatographic retention time and the mass spectrum can be used to help identify components of the mixture. Because the compounds are detected with little bias for one type of compound over another, GC/MS has provided organic chemists with a powerful tool to characterize reaction mixtures and assess product purity.

Sensitivity is another distinguishing feature of mass spectrometry. This sensitivity has allowed mass spectrometers to act as detectors for capillary columns, which can separate mixture containing hundreds of compounds, when less than a nanogram(10^{-9} g) of each compound is injected.

Capillary Columns

Most GC/MS instruments use capillary columns for chromatographic separation. Capillary columns are very long (15- to 30-m), open tubes of fused silica that are coated with a thin coating of the stationary phase (Fig. 8.51). A carrier gas, typically helium, is used as the mobile phase. Capillary column diameters are commonly in the range of 0.25 to 0.53 mm, and the coating of the stationary phase is in the range of 0.25 to 1 μm Thicker coatings are used for the separation of low-boiling compounds.

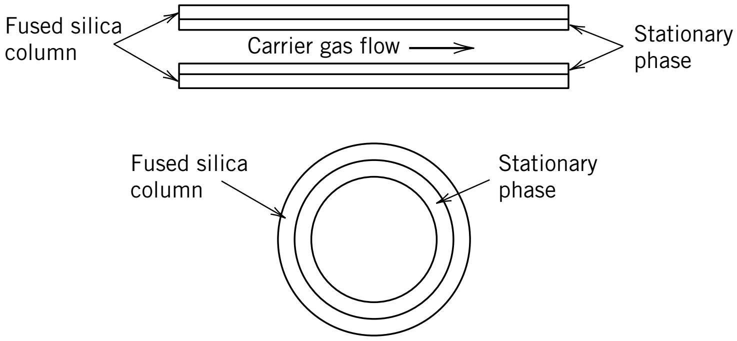

Figure 8.51 Longitudinal and radial cross sections of a capillary column.

Chromatographic resolution increases as a function of the square root of the column length, and the extraordinary length of capillary columns means that most simple mixtures are easily resolved on just a few types of stationary phases. One common nonpolar stationary phase is poly(dimethylsiloxane) (R = CH_3) which can be made slightly more polar by the incorporation of phenyl groups (typically 5% phenyl) in place of methyl groups:

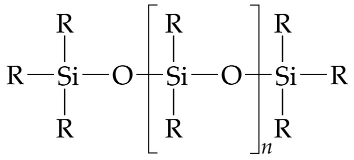

$$R{-}\overset{R}{\underset{R}{Si}}{-}O{-}\left[\overset{R}{\underset{R}{Si}}{-}O\right]_n{-}\overset{R}{\underset{R}{Si}}{-}R$$

These two stationary phases interact with solutes primarily through dispersion interactions, and compounds elute as a function of boiling point. More polar stationary phases are also available. Because capillary columns are used to separate compounds with a wide range of boiling points, we often make use of a technique called *temperature programming.* This techniques allows you to start with a low oven temperature, to optimize the elution of low boiling components, and then increase the oven temperature at a controlled rate, to decrease stationary phase interactions for high-boiling compounds. The increase in temperature decreases retention times and produces narrower peaks for compounds that would otherwise require a long time to elute as a very broad peak.

While samples may be directly injected onto a packed column, the small diameter of the capillary column presents a problem. In addition, it is easy to overload the capillary column with sample (Table 8.35). Two techniques for getting the sample into the column are split and split/splitless injection.

Split Injection

In the split injector the sample is injected into the heated injection port and the evaporated sample is mixed with the carrier gas. The sample/carrier gas mixture is then split between the column and a vent, and a fraction of the sample (determined from the column and vent flow) is introduced to the column (Fig. 8.52*a*). This technique is used to introduce concentrated samples.

Split/Splitless Injection

Splitless injections are used to introduce dilute solutions. The sample is injected into the heated injection port, which is in the "purge off" mode. In this

Table 8.35 **Sample Capacity as a Function of Column Diameter and Stationary Phase Thickness**

Column Diameter (mm)	Stationary Phase Thickness (μm)	Approximate Capacity (ng/component)
0.25	0.10 (thin)	25
	0.25 (most common)	80
	1.0 (thick)	250
0.53	0.10 (thin)	53
	1.0 (most common)	530
	5.0 (thick)	2600

Source: Alltech Capillary Instruction Manual, Bulletin No. 242; Alltech Associates, Inc.: Deerfield, IL, 1991, p. 9.

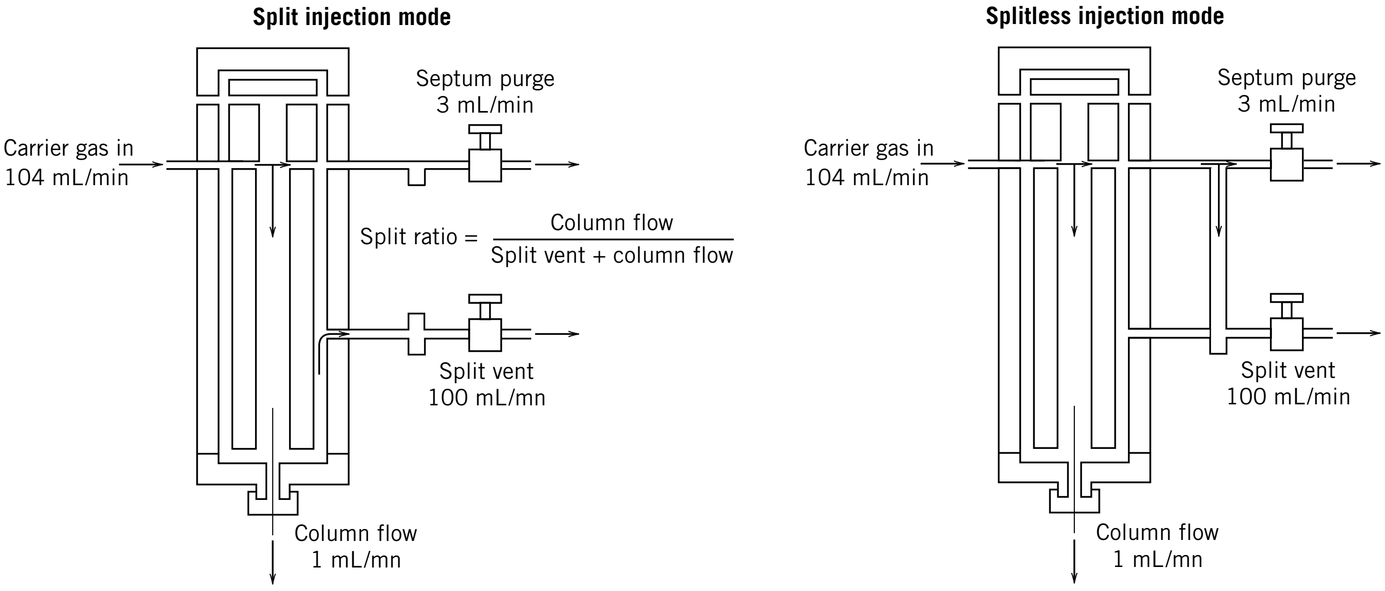
Split injection mode
Septum purge
3 mL/min
Carrier gas in
104 mL/min
Split ratio = Column flow / Split vent + column flow
Split vent
100 mL/mn
Column flow
1 mL/mn
Splitless injection mode
Septum purge
3 mL/min
Carrier gas in
104 mL/min
Split vent
100 mL/min
Column flow
1 mL/mn

(*a*) (*b*)

Figure 8.52 Schematic diagram showing the operation of a split/splitless injection port.

mode, carrier gas flows through the injector directly to the column (Fig. 8.52*b*). This flow rate is very low (0.5–3 mL/min). Of critical importance to splitless injection is the "solvent effect." The oven temperature is maintained below the solvent boiling point, and the vaporized solvent condenses in the column inlet. This condensed solvent acts like a thick layer of stationary phase and traps sample components. After this concentration period (typically 1 min), the injector is changed to the "purge on" mode. This purge sweeps excess solvent (and other volatile components) out of the injector. Purging too early risks venting volatile components, while purging too late increases interference from the solvent tail.

FEATURES OF THE MASS SPECTRUM

A low-resolution mass spectrum can provide many pieces of information that help an organic chemist determine the structure of a molecule. One of the most useful pieces of information is the compound's nominal molecular weight, MW, as determined by identification of the molecular ion, $M^{+\cdot}$. In addition, by careful examination of the region around the molecular ion for the presence of isotopes, we can learn more about the elemental formula for the molecule. The mass spectrum also reveals information about the molecular structure through the appearance of groups of ions characteristic of certain compound types. With more experience and an understanding of mass spectral fragmentation pathways, a molecular structure can be proposed by gathering all information from the spectrum and determining if a proposed structure is consistent with the observed ions. Here we will provide a limited introduction to this process with an emphasis on identification of the molecular ion. We will present one case study to show how mass spectrometry, coupled with gas chromatography, can be used to characterize the products of a synthetic organic reaction.

Terms

The *molecular ion*, represented by $M^{+\cdot}$, is the intact molecule with one electron missing. This should be the peak in the spectrum with the largest mass, but it is not always observed. All spectra have an ion that we call the *base peak*. This is the most abundant peak in the spectrum (*m/z* 44 in the spectrum of N_2O or *m/z* 69 in the spectrum of PFTBA). In the next section you will find that more than one peak may correspond to the $M^{+\cdot}$ or fragment ion when that ion contains elements with different isotopes. We use the term *nominal mass* to describe the mass of the molecule in terms of the most abundant (and, generally, the lightest) isotopes of the element. Relative isotopic abundances for common elements are summarized in Table 8.36.

Isotope Peaks

The mass spectrum for $N(C_4F_9)_3$ (PFTBA; MW = 671) is shown in Figure 8.50. The peaks at *m/z* 69, 219, and 502 are shown above the spectrum as they were measured by the instrument; the spectrum shows their bar-graph representation. These peaks correspond to CF_3^+, $C_4F_9^+$, and $NC_9F_{20}^+$. If you look carefully at the enlarged peaks, you will notice smaller peaks that appear one mass unit above that of the ion. We call these $[A + 1]^+$ peaks. Where do these peaks come from and why does the abundance increase with the mass of the fragment?

If you look at Table 8.36, you will find that fluorine is an isotopically pure element; however, 1.1% of carbon is the ^{13}C isotope. You may recall that it is this low abundance of ^{13}C that you measure with ^{13}C NMR. For nitrogen there is a small amount of ^{15}N. When a molecule contains more than one atom of an isotopically impure element, you increase the chance of finding the higher mass isotope in the molecule. For example, the ^{13}C peaks for *m/z* 69, 219, and 502 of PFTBA (Fig. 8.50) are 1.1, 4.4, and 10.3% of the ^{12}C peaks. The relative intensity increases because there is a higher statistical probability of finding one ^{13}C when the ion has nine vs. one carbon atom. The $[A + 1]^+$ peak intensity from ^{13}C is equal to n times 1.1% the height of the peak A^+, where n is the number of carbon atoms. In addition, we need to add contributions from other A + 1 elements, like nitrogen (0.4%). With precise ion intensity measurements, the relative abundance of the [A + 1] peak can be used to determine the number of carbons present in an organic molecule.

Table 8.36 **Relative Isotope Abundances of Common Elements**

Element	Mass	%	Mass	%	Mass	%
Carbon	12	100	13	1.1		
Hydrogen	1	100	2	0.015		
Nitrogen	14	100	15	0.37		
Oxygen	16	100	17	0.04	18	0.2
Fluorine	19	100				
Silicon	28	100	29	5.10	30	3.4
Phosphorus	31	100				
Sulfur	32	100	33	0.79	34	4.4
Chlorine	35	100			37	32.0
Bromine	79	100			81	97.3
Iodine	127	100				

Mass spectra get very interesting when chlorine or bromine is present. Both elements exist as mixtures of the A and A + 2 isotopes (Table 8.36). The characteristic isotope distribution for bromine, with nearly equal abundances for ^{79}Br and ^{81}Br, is apparent in the spectrum of bromobenzene (Fig. 8.53*a*). The $M^{+\cdot}$ is observed as a cluster of peaks, with *m/z* 156 containing the lightest isotopes (^{12}C and ^{79}Br). Note that the fragment at *m/z* 77 results from loss of Br, and consequently no Br isotope peaks are observed. When more atoms of A + 2 elements are present, characteristic peak distributions are produced (Fig. 8.54). For example, an ion that contains two chlorine atoms will show three peaks, A, [A + 2], and [A + 4], with a 100:65:10.6 intensity distribution. Working from Figure 8.54, you can use a pattern recognition approach to determine the number of chlorine or bromine atoms present in an ion. For example, the spectrum in Figure 8.53*b* shows two ion clusters that suggest the presence of chlorine. A close examination of the distributions indicates that two chlorines are found in the *m/z* 84 cluster, while the *m/z* 49 cluster contains one chlorine. When looking at the mass difference between these ions we work with the *nominal* mass (mass of the ion that has the lightest isotopes). The mass difference, defined by the nominal mass of each cluster, is 35, which corresponds to the mass of one ^{35}Cl (Fig. 8.53*b*).

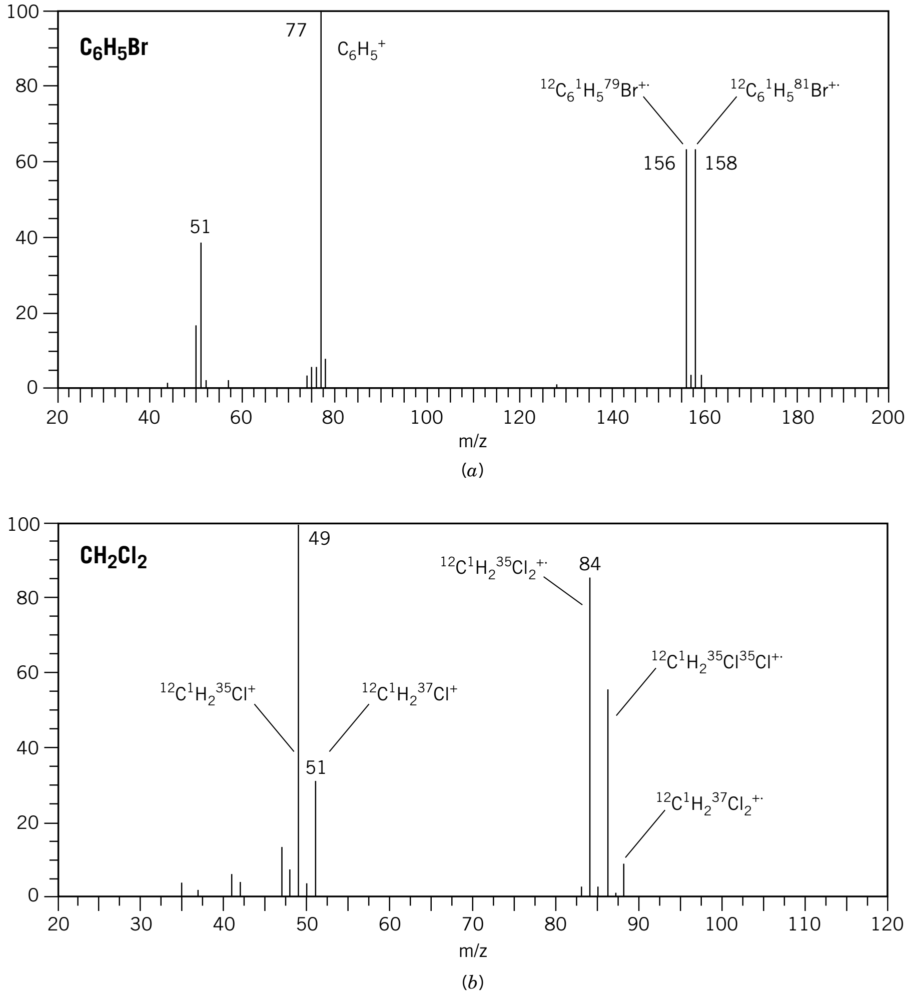

(*a*)

(*b*)

Figure 8.53 Electron ionization mass spectrum of (a) bromobenzene and (b) dichloromethane.

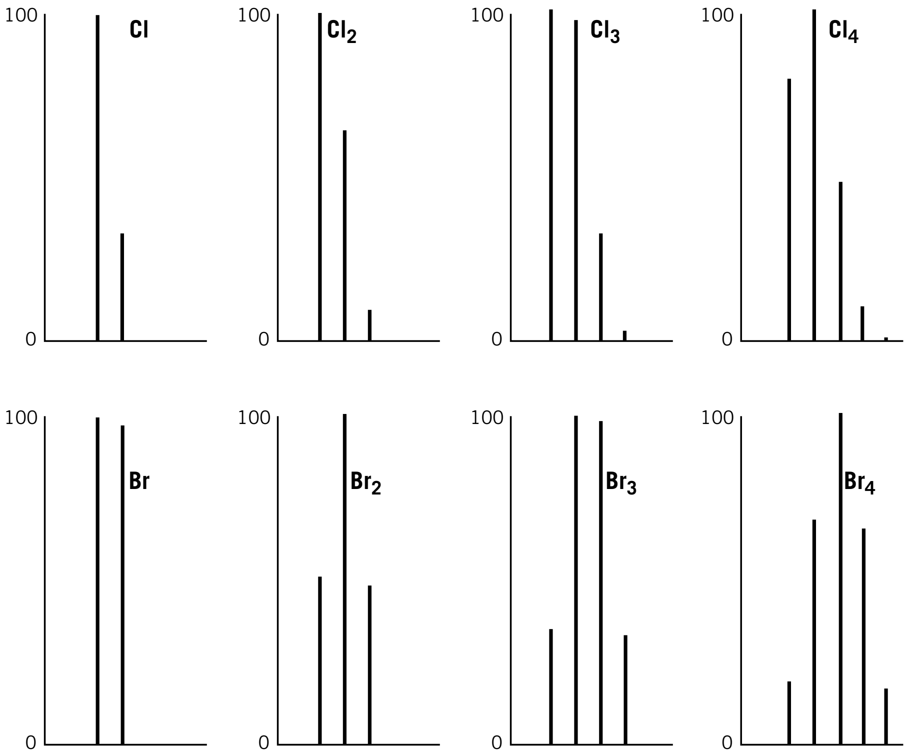

Figure 8.54 **Isotope peak distributions for ions containing chlorine and bromine atoms.**

Recognizing the Molecular Ion

Compound molecular weight is a valuable piece of information that is not always available from the NMR or IR spectrum. In this section we discuss some things to consider as you examine a mass spectrum and attempt to identify the $M^{+\cdot}$ ion. If you look back at the mass spectra that have appeared above, you will find examples of spectra where the molecular ion is the base peak in the spectrum. In some other cases the molecular ion may be weak or not observed at all! For example, PFTBA fragments extensively and the molecular ion does not appear in the spectrum. The following are some things to consider as you attempt to identify the molecular ion.

The molecular ion should be the highest mass peak in the spectrum. When you have tentatively identified a peak as the molecular ion, you should then determine the masses lost from the molecular ion to give high-mass fragments. For example, in Figure 8.53*b* we found a mass difference of 35 between $M^{+\cdot}$ and the first fragment. A listing of common losses from $M^{+\cdot}$ can be found in Table 8.37. If an observed fragment corresponds to an unreasonable loss, such as $M - 12$, this strongly suggests that your tentative identification of the molecular ion is incorrect. Remember that it is always possible that no molecular ion is present.

Another useful feature to consider is the *nitrogen rule*. For most elements found in organic molecules, the compound molecular weight will be *even* if the compound has an even number of nitrogen atoms (remember, zero is an even number). In contrast, the mass will be *odd* if the compound contains an odd number of nitrogen atoms. If you are sure that your product could not contain nitrogen, then an odd mass ion could not correspond to the molecular ion.

Mass spectrometrists also use "softer" ionization techniques to obtain MW information. These techniques included measuring EI mass spectra at lower

Table 8.37 Some Reasonable[a] Losses From $M^{+\cdot}$

Fragment[b]	Radical Lost	Neutral Loss
M-1	H	
M-2		H_2
M-15	CH_3	
M-18		H_2O
M-28		CO or C_2H_4
M-29	C_2H_5	
M-31	OCH_3	
M-32		CH_3OH
M-43	C_3H_7	

[a] Unreasonable losses include [M-4] to [M-14]; [M-21] to [M-25].

[b] For a more complete listing see McLafferty and Turecek (1993).

ionization energies and using a higher pressure ionization technique called *chemical ionization.*

Mass Spectral Interpretation

The following list contains some factors to consider when interpreting mass spectra. To make best use of this summary, the interested reader should consult the text by McLafferty and Turecek, which is considered by many to be the best resource to learn mass spectral interpretation.

1. Using the considerations described above, identify the molecular ion.
2. If possible, determine the elemental composition for $M^{+\cdot}$ and other important peaks using isotopic abundances. In particular, look for isotope peaks from "M + 2" elements like Cl, Br, S, and Si (Table 8.36 and Fig. 8.54). If you are able to establish a molecular formula, calculate "rings + π bonds":

$$\text{For } C_XH_YN_ZO_N \text{ (rings} + \pi \text{ bonds)} = X - \frac{Y}{2} + \frac{Z}{2} + 1$$

NOTE. An even-electron ion, with no unpaired electrons, will have a fractional value. If halogens are present, they are counted as hydrogens.

3. Is the molecular weight odd? If so, this indicates an odd number of nitrogen atoms (for organic molecules).
4. Consider the general appearance of the EI mass spectrum: Is it "aliphatic" (lots of fragmentation) or "aromatic" character (minimal fragmentation)?
5. Look for important low-mass ions (Table 8.38).
6. In the region near M^+, identify fragments lost from the molecular ion (neutral losses) (Table 8.37). Look for intense high-mass ions that may indicate a characteristic, stable fragment ion.
7. Postulate a structure by assembling the various mass fragments/neutral losses. Do the observed fragment ions make sense in terms of fragment/neutral loss stability considerations? Does the structure make sense in

Table 8.38 Some Common Ion Series

Ion Series	Compound Type
m/z 15, 29, 43, 57, 73	Aliphatic hydrocarbons
m/z 38, 39, 50–52, 63–65, 75–78	Aromatic hydrocarbons (not all peaks in ranges will be observed)
m/z 30, 44, 58	Amines
m/z 31, 45, 59	Alcohols

Note. For a more complete listing see McLafferty and Turecek (1993).

terms of other information, such as the reaction conditions, NMR or IR spectra?

8. Verify a postulated structure by comparing the spectrum with a reference spectrum. The reference spectrum may be found in the literature or it may be measured by purchasing or synthesizing a standard of the postulated structure.

CASE STUDY: SYNTHESIS OF METHYL BENZOATE*

To illustrate how gas chromatography and mass spectrometry can be used to characterize the products of an organic reaction, we consider the synthesis of methyl benzoate using a base-catalyzed esterification of benzoic acid. The reaction proposed for this synthesis involved deprotonation of benzoic acid by *n*-butyllithium in dry tetrahydrofuran (THF), followed by the addition of methyl iodide, with dimethylformamide (DMF) added to promote the S_N2 displacement of iodide by the benzoate anion:

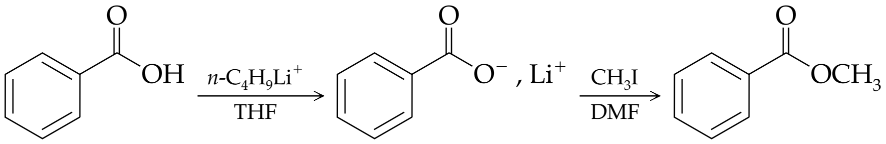

We isolated the reaction products from the reaction mixture by quenching the reaction with water, adding saturated $NaHCO_3$, and extracting the neutral products with diethyl ether. The ethereal solution containing the reaction products was then analyzed by capillary column GC/MS, and the chromatogram shown in Figure 8.55 was produced. The chromatogram displays total ionization as a function of time. The total ionization is a summation of all the ions detected in one scan of the mass spectrometer (one spectrum) plotted as a point as a function of time. The display is often called the TIC (total ionization chromatogram), and the peak areas should reflect the relative amounts of each compound detected by the mass spectrometer.

The chromatogram shown in Figure 8.55 is not quite what we would hope to see. Instead of detecting a single chromatographic peak for our product, we see three peaks. To determine if we made *any* methyl benzoate, and to determine what other components are present in our mixture, we examine the mass

---

*The synthetic work presented below was carried out by Joshua Pacheco, Bowdoin College Class of 1999.

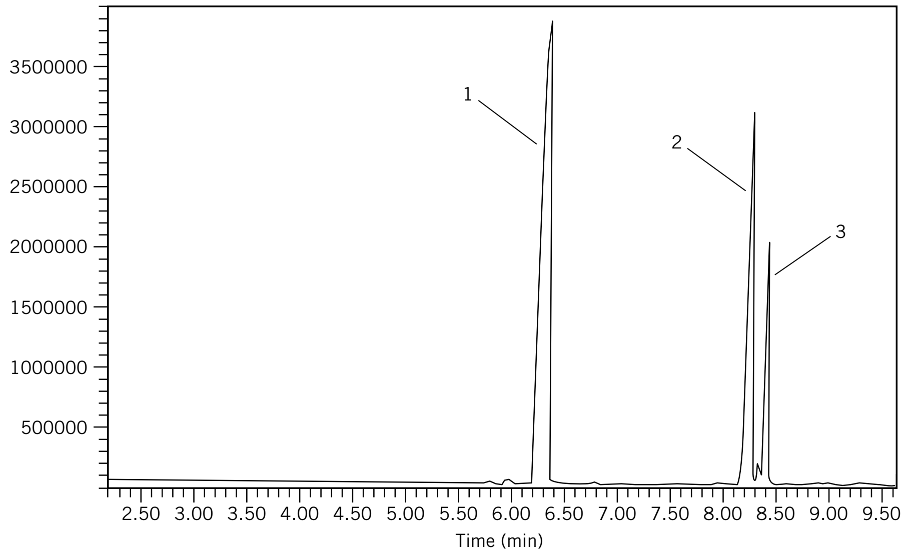

Figure 8.55 **Total ionization chromatogram of a reaction mixture.**

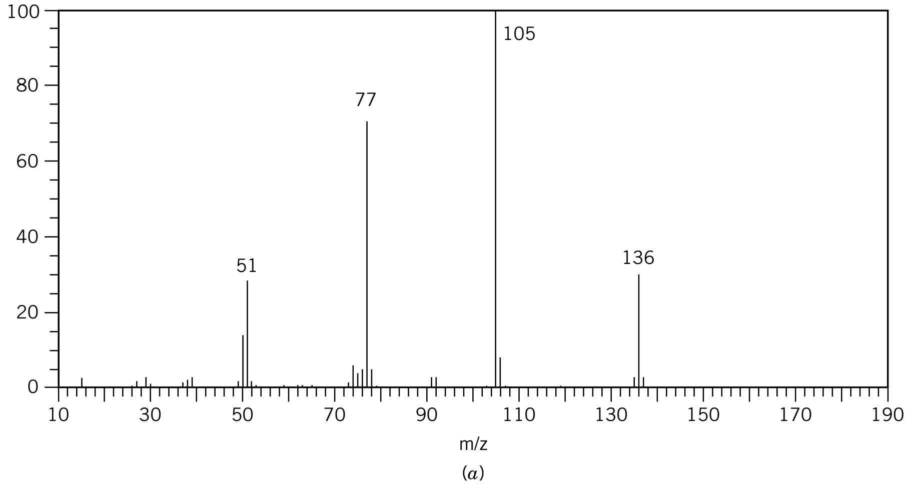

Figure 8.56*a* **Electron ionization mass spectrum of (*a*) peak 1.**

spectrum for each peak. Because we are eager to determine if we made any methyl benzoate, we start by trying to locate a chromatographic peak that has a mass spectrum that corresponds to methyl benzoate. Even if we are not sure what the mass spectrum will look like, we can try to find a spectrum that shows a molecular ion, $M^{+\cdot}$, that corresponds to the molecular weight of methyl benzoate ($C_8H_8O_2$, MW = 136). The mass spectrum for peak **1** (Fig. 8.56*a*) shows an ion at *m/z* 136 that appears as the highest mass peak in the spectrum. Let's now take a closer look at the mass spectrum to see if the fragment ions are consistent with the structure of methyl benzoate.

The base peak in the spectrum appears at *m/z* 105. This intense peak results from a loss of 31 from the molecular ion, which corresponds to loss of OCH_3. This is a predicted loss. Upon ionization, we expect one of the nonbonding electrons on the carbonyl oxygen to be lost, and can consider the charge and unpaired electron to be localized on that oxygen. The unpaired electron initiates a cleavage that results in loss of OCH_3 radical. The ions at *m/z* 77 and 51, which are characteristic of aromatic rings, may form by cleavage on the other side of the carbonyl group. The ions at *m/z* 105 and 77 may

JWCL196_c08_539-628.qxd 11/16/09 4:35 PM Page 625

EQA

undergo another fragmentation, but the loss of another radical species is not generally observed from ions of this type, where all electrons are now paired up. Instead, even-electron neutrals, such as CO or C_2H_2, are lost to give fragments *m/z* 77 and 51, respectively:

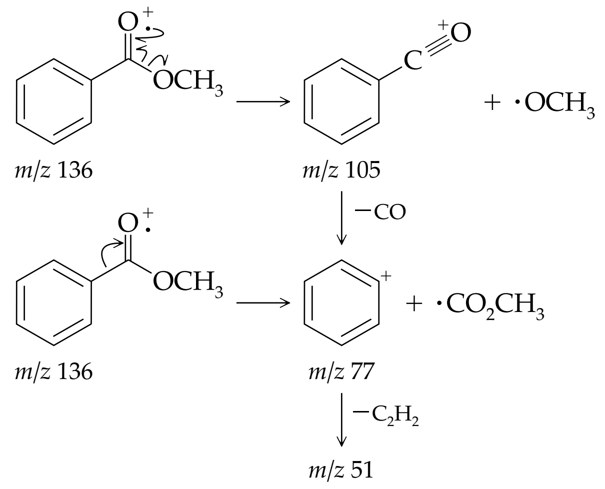

Thus, the mass spectrum for peak **1** is consistent with the structure of methyl benzoate. We could further confirm our identification by consulting a library of mass spectra. If we had any pure methyl benzoate around, we could also prepare a standard and use both the GC retention time and the mass spectrum of the standard as a means of confirming the compound identification.

We can now move on to some other peaks in the chromatogram. The mass spectra for peaks **2** and **3** (Fig. 8.56*b, c*) have many similar features to those of methyl benzoate. Both **2** and **3** show ions at *m/z* 51, 77, and 105. We can conclude that both compounds have a carbonyl group attached to an aromatic ring. We can next consider identification of the molecular ion. For peak **2**, the highest mass ion is *m/z* 162. To determine if this is a reasonable assignment for M^+, we determine losses from *m/z* 162. The ions at *m/z* 133, 120, and 105 could result from losses of 29, 42, and 57, respectively. None of these losses are unreasonable. We next want to consider two important pieces of information. First, if we assume that the MW is 162, we must add

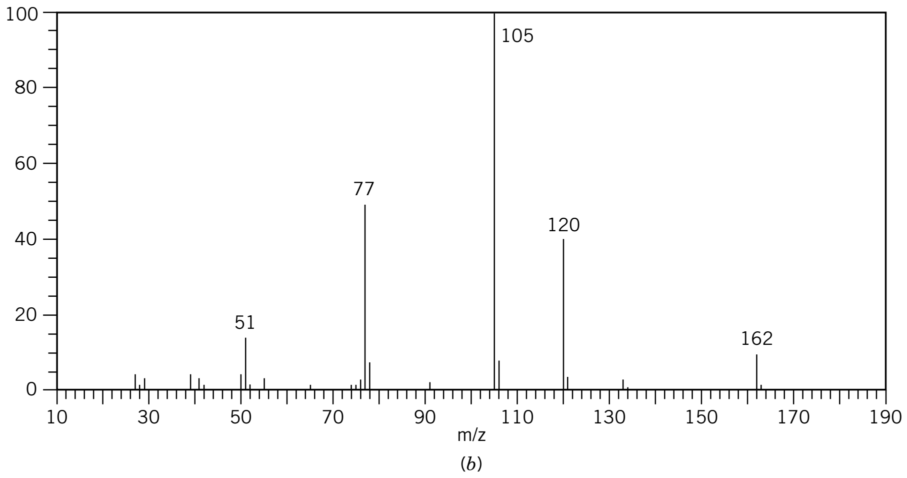

Figure 8.56*b* Electron ionization mass spectrum of peak 2, and (c).

CONFIRMING PAGES aptara

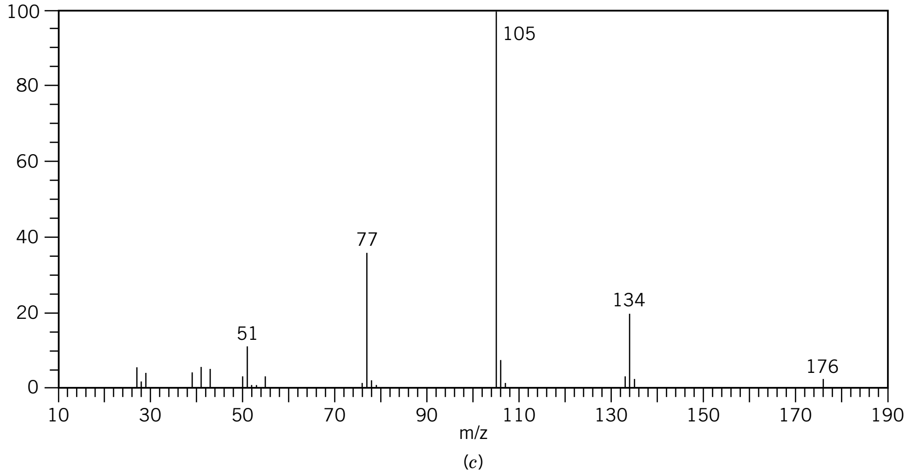

Figure 8.56*c* **Electron ionization mass spectrum of peak 3.**

57 to the carbonyl substituted aromatic ring to make our molecule. Addition of a butyl group is a logical choice:

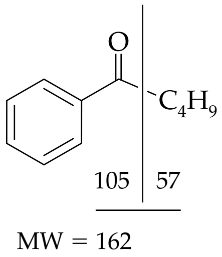

Next we take note of the peak at *m/z* 120. The even mass of this ion, resulting from an even mass molecular ion, indicates that it is a special ion! Even-mass fragments generally result from rearrangement reactions, and rearrangements involving hydrogen transfers to carbonyl groups can produce particularly informative product ions. If a butyl group is attached to the carbonyl, the following fragmentation pathway will occur, which nicely explains the *m/z* 120 peak:

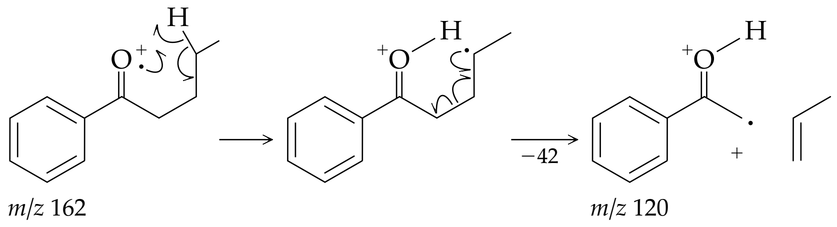

To determine if this assignment makes sense, we go back to consideration of our reaction. How could this product be generated? If we assume that there is some unreacted *n*-butyllithium around after the methyl benzoate has been formed, then the following reaction is possible. Thus, we can feel quite confident in our assignment for peak **2**:

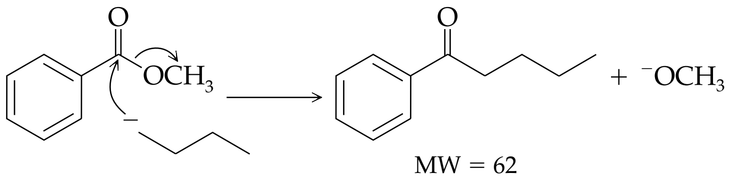

Moving on to peak **3**, we detect an ion at *m/z* 134. Let's start by assuming that this is our molecular ion. The *m/z* 105 peak would result from a loss of 29 (C_2H_5) from *m/z* 134, which is a reasonable loss from $M^{+\cdot}$. This *suggests*, erroneously, that ethyl benzoate is our product:

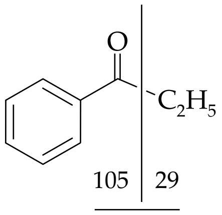

MW = 134

Why is this identification incorrect? If we look back at the chromatogram in Figure 8.55 the chromatographic retention times tell us that something is amiss. Remember that compounds elute from the column in approximate order of increasing boiling point. We would expect that the butyl phenyl ketone would elute *after*, not before, the ethyl phenyl ketone! In addition, you would be hard pressed to propose a mechanism for formation of ethyl benzoate in the context of this reaction. Let's go back and take another look at the mass spectrum. A careful examination shows a small peak at *m/z* 176. We may have incorrectly identified the molecular ion! If $M^{+\cdot}$ is *m/z* 176, this gives losses of 42 and 71 to form *m/z* 134 and 105, respectively. Now it looks like we have a pentyl group attached to the carbonyl, which agrees nicely with the chromatographic retention times. The *m/z* 134 ion becomes one of our special, even mass ions. What does this ion reveal about the pentyl group? The fact that the ion results from loss of 42, and not 58, clearly indicates that this is not an *n*-pentyl group. What makes the most sense is the branching shown below:

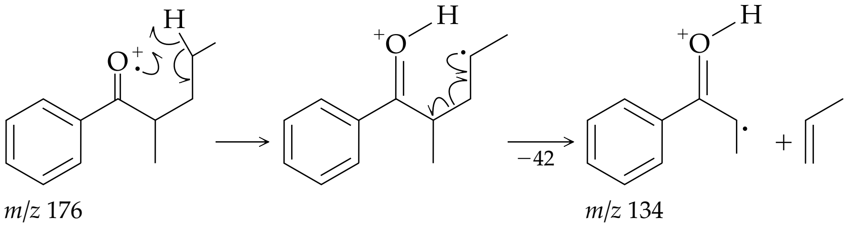

In terms of the chemistry of the reaction, this product also makes sense if deprotonation by *n*-butyllithium occurs α to the carbonyl, followed by reaction with methyl iodide:

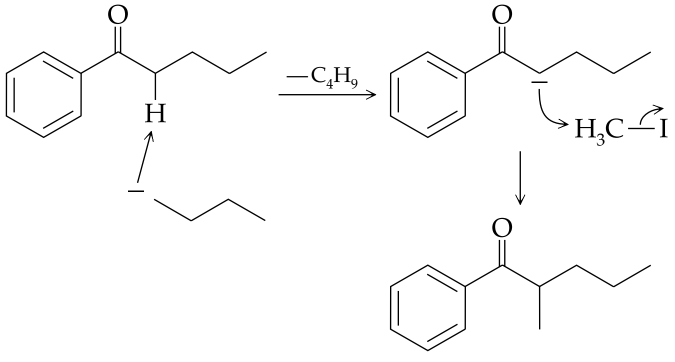

MW = 176

BIBLIOGRAPHY

Mass Spectral Interpretation:

Beynon, J. H. *The Mass Spectra of Organic Molecules;* Elsevier: Amsterdam, 1968.

Biemann, K. *Mass Spectrometry. Organic Chemical Applications;* McGraw-Hill: New York, 1962.

Budzikiewicz, H.; Djerassi, C.; Williams, D. H. *Mass Spectrometry of Organic Compounds;* Holden-Day: San Francisco, 1967.

McLafferty, F. W.; Turecek, F. *Interpretation of Mass Spectra,* 4th ed.; University Science Books: Sausalito, CA, 1993.

Silverstein, R. M.; Webster, F. X.; Kiemle, D. *Spectrometric Identification of Organic Compounds,* 7th ed.; Wiley: New York, 2005.

Watson, J. T.; Sparkman, O. D. *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation,* 4th ed.; Wiley: New York 2007 (also a good introduction to instrumentation).

Theory and Instrumentation:

Dass, C. *Fundamentals of Contemporary Mass Spectrometry;* Wiley: New York, 2007.

De Hoffmann, E.; Stroobant, V. *Mass Spectrometry: Principles and Applications,* 3rd ed.; Wiley: New York, 2007.

Kitson, F. G.; Larsen, B. S.; McEwen, C. N. *Gas Chromatography and Mass Spectrometry: A Practical Guide;* Academic Press: San Diego, 1996.

Message, G. M. *Practical Aspects of Gas Chromatography/Mass Spectrometry;* Wiley: New York, 1984.

Quadrupole Mass Spectrometry and Its Applications; Dawson, P. H., Ed., Elsevier: New York, 1976.

Libraries:

McLafferty, F. W.; Stauff, D. B.; *The Important Peak Index of the Registry of Mass Spectral Data,* Wiley: New York, 1991.

McLafferty, F. W.; Stauffer, D. B. *Wiley/NBS Registry of Mass Spectral Data;* Wiley: New York, 1989.

NIST/EPA/MSDC Mass Spectral Library; National Institute of Standards and Technology. You may also access some mass spectral data through the internet: *WebBook.nist.gov*

C_9H_{12}apter 9

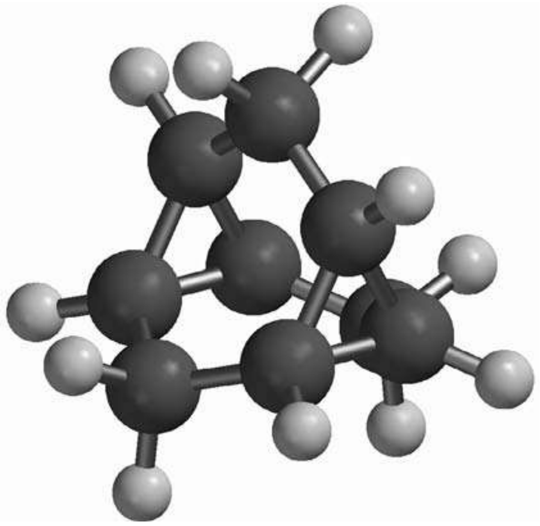

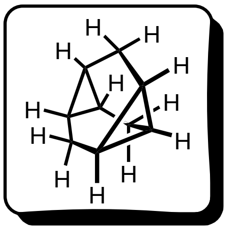

QUALITATIVE IDENTIFICATION OF ORGANIC COMPOUNDS

ORGANIC QUALITATIVE ANALYSIS

One of the exciting challenges that a chemist faces on a regular basis is identifying organic compounds. This challenge is an excellent way for a student to be initiated into the arena of chemical research. Millions of organic compounds are recorded in the chemical literature. At first glance it may seem a bewildering task to attempt to identify one certain compound from this vast array, but most of these substances can be grouped, generally by functional groups, into a comparatively small number of classes. In addition, chemists have an enormous database of chemical and spectroscopic information, which has been correlated and organized over the years, at their disposal. Determination of the physical properties of a molecule, the functional groups present, and the reactions the molecule undergoes has allowed the chemist to establish a systematic, logical identification scheme.

Forensic chemistry, the detection of species causing environmental pollution, the development of new pharmaceuticals, progress in industrial research, and development of polymers all depend to a large extent on the ability of the chemist to isolate, purify, and identify specific chemicals. The objective of organic qualitative analysis is to place a given compound, through screening tests, into one of a number of specific classes, which in turn greatly simplifies the *identification* of the compound. This screening is usually done by using a series of preliminary observations and chemical tests, in conjunction with the instrumental data that developments in spectroscopy have made available to the analyst. The advent of infrared (IR) and nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) have had a profound effect on the approach taken to identify a specific organic compound. Ultraviolet (UV) spectra may also be utilized to advantage with certain classes of materials.

The systematic approach taken in this text for the identification of an unknown organic compound is as follows:

1. Preliminary tests are performed to determine the physical nature of the compound.

2. The solubility characteristics of the unknown species are determined. This identification can often lead to valuable information related to the structural composition of an unknown organic compound.

Chapter 9: C_9H_{12}, Triasterene (*L. aster,* star)
Musso and Biethan (1964)

3. Chemical tests, mainly to assist in identifying elements other than C, H, or O, may also be performed.

4. Classification tests are carried out to detect *common functional groups* present in the molecule. Most of these tests may be done using a few drops of a liquid or a few milligrams of a solid. An added benefit to the student, especially in relation to the chemical detection of functional groups, is that a vast amount of chemistry can be *observed* and *learned* in performing these tests. The successful application of these tests requires that you develop the ability to think in a logical manner and to interpret the significance of each result based on your observation. Later, as the *spectroscopic techniques* are introduced, *the number of chemical tests performed are usually curtailed.*

5. The spectroscopic method of analysis is utilized. As your knowledge of chemistry develops, you will appreciate more and more the revolution that has taken place in chemical analysis over the past 25–30 years and the powerful tools now at your disposal for the identification of organic compounds. In the introductory laboratory, the techniques of IR, NMR, and UV–vis spectroscopy, and mass spectrometry are generally explored.

It is important to realize that *negative* findings are often as important as *positive* results in identifying a given compound. Cultivate the habit of following a *systematic pathway or sequence* so that no clue or bit of information is lost or overlooked along the way. It is important also to develop the *attitude* and *habit* of planning ahead. Outline a logical plan of attack, depending on the nature of the unknown, and follow it. As you gain more experience in this type of investigative endeavor, the planning stage will become easier.

At the *initial* phase of your training, the unknowns to be identified will be relatively pure materials and will all be known compounds. The properties of these materials are recorded in the literature, and/or in the tables on the website; see Chapter 9W. Later, perhaps, mixtures of compounds or samples of commercial products will be assigned for separation, analysis, and identification of the component compounds.

www →

Record all observations and results of the tests in your laboratory notebook. Review these data as you execute the sequential phases of your plan. This method serves to keep you on the path to success.

A large number of texts have been published on organic qualitative analysis. Several references are cited here.

BIBLIOGRAPHY

Cheronis, N. D.; Entriken, J. B. *Identification of Organic Compounds: A Student's Text Using Semimicro Techniques;* New York: Interscience, 1963.

Cheronis, N. D.; Ma, T. S. *Organic Functional Group Analysis by Micro and Semimicro Methods;* Interscience: New York, 1964.

Cheronis, N. D.; Entrikin, J. B.; Hodnett, E. M. *Semimicro Qualitative Organic Analysis,* 3rd ed.; Interscience: New York, 1965.

Feigl, F.; Anger, V. *Spot Tests in Organic Analysis,* 7th ed.; Elsevier: New York, 1966.

Pasto, D. J.; Johnson, C. R.; Miller, M. J. *Experiments and Techniques in Organic Chemistry;* Prentice Hall: Englewood Cliffs, NJ, 1992.

Schneider, F. L. In *Qualitative Organic Microanalysis,* Vol. II of *Monographien aus dem Gebiete der qualitativen Mikroanalyse;* Benedetti-Pichler, A. A., Ed.; Springer-Verlag: Vienna, 1964.

Shriner, R. L.; Hermann, C.K.F.; Morrill, T.C.; Curtin, D.Y.; Fuson, R.C. *The Systematic Identification of Organic Compounds,* 8th ed.; Wiley: New York, 2003.

Vogel, A. I. *Qualitative Organic Analysis,* Part 2 of *Elementary Practical Organic Analysis;* Wiley: New York, 1966.

———. Vogel's Textbook of *Practical Organic Chemistry, Including Qualitative Organic Analysis*, 5th ed.; London: Longman Group, 1989.

PRELIMINARY TESTS

Preliminary tests help you select a route to follow to ultimately identify the unknown material at hand. These tests frequently consume material, so, given the amounts of material generally available at the micro- or semimicroscale level, judicious selection of the tests to perform must be made (in some tests, the material analyzed may be recovered). Each preliminary test that can be conducted with *little expenditure of time and material* can offer valuable clues as to which class a given compound belongs.

Nonchemical Tests

Physical State. If the material is a *solid*, a few milligrams of the sample may be viewed under a magnifying glass or microscope, which may give some indication as to the homogeneity of the material. Crystalline shape is often an aid in classifying the compound.

Determine the melting point, using a small amount of the solid material. A narrow melting point range (1–2 °C is a good indication that the material is quite pure. If a broad range is observed, the compound must be recrystallized from a suitable solvent before proceeding. If the material undergoes decomposition on heating, try an evacuated (sealed-tube) melting point. If any evidence indicates that sublimation is occurring, an evacuated melting point should be run. Furthermore, this result indicates that sublimation might be used to purify the compound, if necessary.

If the material is a *liquid*, the boiling point is determined by the ultramicro method. If sufficient material is on hand and the boiling point reveals that the material is relatively pure (narrow boiling point range), the *density* and the *refractive index* can provide valuable information for identification purposes.

www

Color. Since the majority of organic compounds are colorless, examination of the color can occasionally provide a clue as to the nature of the sample. Use caution, however, since tiny amounts of some impurities can color a substance. Aniline is a classic example. When freshly distilled it is colorless, but on standing a small fraction oxidizes and turns the entire sample a reddish-brown color.

Colored organic compounds contain a *chromophore*, usually extended conjugation in the molecule. For example, 1,2-dibenzoylethylene (Experiments [3A] and [6]) is yellow; 5-nitrosalicylic acid (Experiment [29C]) is light yellow; tetraphenylcyclopentadienone (Experiment [A3$_a$]) is purple.

Can you identify the chromophores that cause these compounds to be colored? Note that a colorless liquid or white solid would not contain these units. Thus, compounds containing these groupings would be excluded from consideration as possible candidates in identification of a given substance.

Odor. Detection of a compound's odor can occasionally be of assistance, since the vast majority of organic compounds have no definitive odor. You should become familiar with the odors of the common compounds or classes. For example, aliphatic amines have a fishy smell; benzaldehyde (like nitrobenzene and benzonitrile) has an almond odor; esters have fruity odors (Experiments [8A–C]). Common solvents, such as acetone, diethyl ether, and toluene, all have distinctive odors. Butyric and caproic acids have rancid odors. Low molecular weight mercaptans (—SH) have an intense smell of rotten eggs. In many cases, extremely small quantities of certain relatively high molecular weight compounds

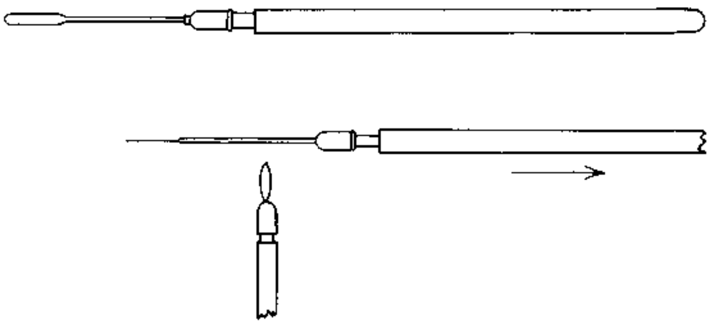

Figure 9.1 Heating on the microspatula. *(Courtesy of Springer-Verlag, Vienna.)*

can be detected by their odor. For example, a C_{16} unsaturated alcohol released by the female silk worm moth elicits a response from male moths of the same species at concentrations of 100 molecules/cm^3. Odors are an important facet of chemical communication between plants and animals and often result in a spectacular behavioral response (see also Experiment [8]).

Odor detection in humans involves your olfactory capabilities and thus can be a helpful lead, *but very rarely can this property be used to strictly classify or identify a substance.* As mentioned above, contamination by a small amount of an odorous substance is always a possibility.

CAUTION

CAUTION: You should be very cautious when detecting odors. Any odor of significance can be detected several inches from the nose. Do not place the container closer than this to your eyes, nose, or mouth. Open the container of the sample and gently waft the vapors toward you.

Ignition Test

CAUTION

CAUTION: Make sure you are wearing safety glasses.

Valuable information can be obtained by carefully noting the manner in which a given compound burns. The ignition test[1] is carried out by placing 1–2 mg of the sample on a spatula, followed by heating and ignition with a microburner flame. Do not hold the sample directly in the flame; heat the spatula about 1 cm from the flat end and move the sample slowly into the flame (see Fig. 9.1).

Important observations to be made concerning the ignition test are summarized in Table 9.1.

As the sample is heated, you should make the following observations:

1. Any melting or evidence of sublimation: This observation gives an approximate idea of the melting point by the temperature necessary to cause melting.

2. Color of the flame as the substance begins to burn (see Table 9.1).

3. Nature of the combustion (flash, quiet, or an explosion). Rapid, almost instantaneous combustion indicates high hydrogen content. Explosion indicates

---

[1] For an extensive discussion on examination of ignition residues see Feigl, F.; Anger, V. *Spot Tests in Organic Analysis,* 7th ed.; Elsevier: New York; 1966, p. 51.

Table 9.1 **Ignition Test Observations**

Type of Compound	Example	Observation
Aromatic compounds, unsaturated, or higher aliphatic compounds	Toluene	Yellow, sooty flame
Lower aliphatic compounds	Hexane	Yellow, almost nonsmoky flame
Compounds containing oxygen	Ethanol	Clear bluish flame
Polyhalogen compounds	Chloroform	Generally do not ignite until burner flame applied directly to the substance
Sugars and proteins	Sucrose	Characteristic odor
Acid salts or organometallic compounds	Ferrocene	Residue

Source. Cheronis, N. D.; Entrikin, J. B. *Semimicro Qualitative Analysis;* Interscience: New York, 1947, p. 85.

the presence of nitrogen- or nitrogen–oxygen-containing groups, for example, nitro groups (Experiment [29]).

4. Nature of the residue, if present, after ignition.

a. If a black residue remains and disappears on further heating at higher temperature, the residue is carbon.

b. If the residue undergoes swelling during formation, the presence of a carbohydrate or similar compound is indicated.

c. If the residue is black initially but still remains after heating, an oxide of a heavy metal is indicated.

d. If the residue is white, the presence of an alkali or alkaline earth carbonate or SiO_2 from a silane or silicone is indicated.

SEPARATION OF IMPURITIES

If the preliminary tests outlined above indicate that the unknown in question contains impurities, it may be necessary to carry out one of several purification steps. These techniques are discussed in earlier chapters and are summarized below for correlation purposes:

1. For a liquid, distillation is generally used (see Techniques 2 and 3).
2. For a solid, recrystallization is generally used (see Technique 5).
3. Extraction is used if the impurity is insoluble in a solvent in which the compound itself is soluble (see Technique 4).
4. Sublimation is a very efficient technique if the compound sublimes (see Technique 9).
5. Chromatography (gas, column, and thin-layer) is often used (see Techniques 1 and 6A).

These techniques may be applied to the separation of mixtures as well.

DETECTION OF ELEMENTS OTHER THAN CARBON, HYDROGEN, OR OXYGEN

Other than C, H, and O, the elements that are most often present in organic compounds are nitrogen, sulfur, and the halogens (F, Cl, Br, or I). To detect the presence of these elements, the organic compound is generally fused with metallic sodium. This reaction converts these heteroatoms to the water-soluble inorganic compounds, NaCN, Na_2S, and NaX. Inorganic qualitative analysis tests enable the investigator to determine the presence of the corresponding anions:

$$\begin{matrix}\text{Organic compound}\\\text{containing}\end{matrix}\left\{\begin{matrix}\mathrm{C}\\\mathrm{H}\\\mathrm{O}\\\mathrm{N}\\\mathrm{S}\\\mathrm{X}\end{matrix}\right\}\xrightarrow[\Delta]{\mathrm{Na}}\left\{\begin{matrix}\mathrm{NaCN}\\\mathrm{Na_2S}\\\mathrm{NaX}\end{matrix}\right\}$$

Sodium Fusion[2]

HOOD *NOTE. The fusion reaction is carried out in the* ***hood.*** *Make sure you are wearing safety glasses. All reagents must be of analytical grade, and deionized water must be used.*

CAUTION

CAUTION: Sodium metal can cause serious burns and it reacts violently with water.

In a small (10 × 75-mm) test tube (soft glass preferred), supported in a transite board (see Fig. 9.2), place about 25–30 mg of clean sodium metal (about one-half the size of a pea).

CAUTION

CAUTION: Use forceps to make this transfer; never touch sodium metal with your fingers.

Heat the tube with a flame until the sodium melts and sodium vapor is observed rising in the tube (see Fig. 9.2).

Mix a small sample of your unknown compound (1–2 drops of a liquid; 6–10 mg if a solid) with about 15–25 mg of *powdered* sucrose.[3] Gentle mixing of solids may be done on filter paper or glassine weighing paper; liquids can be mixed on a watch glass. Add this mixture to the tube, being careful not to get any material on the sides of the test tube.

NOTE. The addition of sucrose to the sample helps reduce various nitrogen or sulfur compounds. It also absorbs volatile materials so that they may undergo the desired reaction before significant vaporization can occur.

Now heat the tube gently to initiate the reaction with sodium. Remove the flame until the reaction subsides, and then heat to redness for 1–2 min. Allow the

---

[2] See Campbell, K. N.; Campbell, B. K. *J. Chem. Educ.* **1950,** *27,* 261 for a discussion of the procedure.

[3] Ordinary confectioner's sugar purchased at the supermarket can be used.

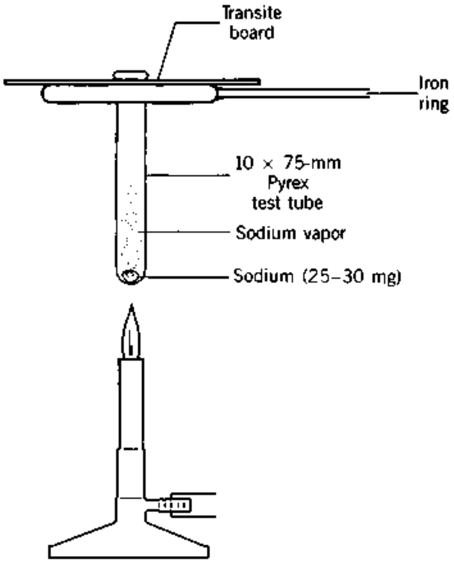

Figure 9.2 Apparatus for sodium fusion.

tube and contents to cool to room temperature. Then, *and only then,* **cautiously** add several drops of methanol (using a Pasteur pipet) to decompose any unreacted metallic sodium. Gently warm the mixture to drive off the excess methanol.

Reheat the tube to a bright red. While the tube is still red hot, lift the transite board and test tube from the iron ring and place the tube in a small beaker (30 mL) containing about 15 mL of deionized water (the transite board acts as a cover on the beaker).

CAUTION: The soft-glass tube usually cracks and breaks during this operation.

CAUTION

Break up the tube with a glass rod, heat the solution to boiling and filter it by gravity into a clean 50-mL Erlenmeyer flask. Wash the filter paper with an additional 2.0 mL of distilled water and combine this wash with the original filtrate.

NOTE. If a Pyrex test tube is used, after the unreacted sodium metal is completely destroyed by adding methanol, add 2 mL of deionized water directly to the tube and contents. Place a glass stirring rod in the tube and heat the solution to boiling with stirring and then filter as described above. Dilute the filtrate with deionized water to about 5 mL.

Using the Fusion Solution. The clear, colorless fusion solution is used to test for the presence of CN^- (nitrogen), S^{2-} (sulfur), and X^- (halogens, except F^-) as described in the following sections.

Sulfur

1. Place 2–3 drops (Pasteur pipet) of the fusion solution on a white spot plate, followed by 2 drops of water. Now add 1 drop of dilute (2%) aqueous sodium nitroprusside solution. The formation of a deep blue-violet color is a positive test for sulfur:

$$\mathrm{Na_2S} + \underset{\text{Sodium nitroprusside}}{\mathrm{Na_2Fe(CN)_5NO}} \longrightarrow \underset{\text{Blue-violet complex}}{\mathrm{Na_4[Fe(CN)_5NOS]}}$$

2. Place 3–4 drops (Pasteur pipet) of the fusion solution on a white spot plate followed by 1–2 drops of acetic acid. Now add 1 drop of 1% lead(II) acetate solution. The formation of a black precipitate (lead sulfide) indicates the presence of sulfur.

Nitrogen[4]

Reagents

1. A 1.5% solution of *p*-nitrobenzaldehyde in 2-methoxyethanol
2. A 1.7% solution of *o*-dinitrobenzene in 2-methoxyethanol
3. A 2.0% solution of NaOH in distilled water

NOTE. All reagent drops are dispensed using Pasteur pipets.

On a white spot plate, place together: 5 drops of reagent **1,** 5 drops of reagent **2,** and 2 drops of reagent **3.** Stir this mixture gently with a glass rod.

Now add 1 drop of the fusion solution. The formation of a deep-purple color is a positive test for the presence of CN^- ion; a yellow or tan coloration is negative. If a positive result is obtained, nitrogen is present in the sample.

The test is valid in the presence of halogens (NaX) or sulfur (Na_2S). It is much more sensitive than the traditional Prussian Blue test.[5]

The Soda Lime Test. In a 10 × 75-mm test tube, mix about 50 mg of soda lime and 50 mg of MnO_2. Add 1 drop of a liquid unknown or about 10 mg of a solid unknown. Place over the mouth of the tube a moist strip of Brilliant Yellow paper (moist, red litmus paper is an alternative). Using a test tube holder, hold the tube at an incline (*pointing away from you and others*) and heat the contents gently at first and then quite strongly. Nitrogen-containing compounds will usually evolve ammonia.

A positive test for nitrogen is the deep red coloration of the Brilliant Yellow paper (or blue color of the litmus paper).

The Halogens (Except Fluorine)

Using the Fusion Solution. In a 10 × 75-mm test tube containing a boiling stone, place 0.5 mL (calibrated Pasteur pipet) of the fusion solution. Carefully acidify this solution by the dropwise addition of dilute HNO_3 (nitric acid), delivered from a Pasteur pipet (test acidity with litmus paper). If a positive test for nitrogen or sulfur was obtained, heat the resulting solution to a gentle boil (stir with a microspatula to prevent boilover) for 1 min over a microburner in the **hood** to expel any HCN or H_2S that might be present. Then cool the tube to room temperature.

HOOD

To the resulting cooled solution, add 2 drops (Pasteur pipet) of aqueous 0.1 M $AgNO_3$ solution.

---

[4] Adapted from Guilbault, G. G.; Kramer, D. N. *Anal. Chem.* **1966,** *39,* 834. *Idem. J. Org. Chem.* **1966,** *31,* 1103. See also Shriner, R. L.; Fuson, R. C.; Morrill, T. C. *The Systematic Identification of Organic Compounds*, 6th ed.; Wiley: New York, 1980, p. 80.

[5] See Vogel, A. I. *Elementary Practical Organic Chemistry*, Part 2, 2nd ed.; Wiley: New York, 1966, p. 37.

A heavy curdy-type precipitate is a positive test for the presence of Cl^-, Br^-, or I^- ion. A faint turbidity is a negative test.

AgCl precipitate is white.
AgBr precipitate is pale yellow.
AgI precipitate is yellow.
AgF is not detected by this test since it is relatively soluble in water.

The silver halides have different solubilities in dilute ammonium hydroxide solution.

Centrifuge the test tube and contents and remove the supernatant liquid using a Pasteur filter pipet. Add 0.5 mL (calibrated Pasteur pipet) of dilute ammonium hydroxide solution to the precipitate and stir with a glass rod to determine whether the solid is soluble.

AgCl is soluble in ammonium hydroxide due to the formation of the complex ion, $[Ag(NH_3)_2]^+$.
AgBr is slightly soluble in this solution.
AgI is insoluble in this solution.

A Further Test. Once the presence of a halide ion has been established, a further test is available to help you distinguish between Cl^-, Br^-, and I^- ions.[6]

As described above, acidify 0.5 mL of the fusion solution with dilute HNO_3. To this solution, add 5 drops (Pasteur pipet) of a 1.0% aqueous $KMnO_4$ solution and shake the test tube for about 1 min.

Now add 10–15 mg of oxalic acid, enough to decolorize the excess purple permanganate, followed by 0.5 mL of methylene chloride solvent. Stopper, shake, and vent the test tube and allow the layers to separate. Observe the color of the CH_2Cl_2 (lower) layer.

A clear methylene chloride layer indicates Cl^- ion.
A brown methylene chloride layer indicates Br^- ion.
A purple methylene chloride layer indicates I^- ion.

The colors may be faint and should be observed against a white background.

The Beilstein Test[7]. In the Beilstein test organic compounds that contain chlorine, bromine, or iodine, and hydrogen decompose on ignition in the presence of copper oxide, to yield the corresponding hydrogen halides. These hydrogen halides react to form the volatile cupric halides that impart a green or blue-green color to a nonluminous flame. It is a very sensitive test, but some nitrogen-containing compounds and some carboxylic acids also give positive results.

Pound one end of a 4-in. long copper wire to form a flat surface that can act as a spatula. The other end of the wire is stuck in a cork stopper to serve as an insulated handle.

Heat the flat tip of the wire in a flame until coloration of the flame is negligible.

---

[6] For a further test to distinguish between the three halide ions see Shriner, R. L.; Fuson, R. C.; Morrill, T. C. *The Systematic Identification of Organic Compounds,* 6th ed.; Wiley: New York, 1980, p. 81. Also see this reference (p. 85) for a specific test for the F^- ion.
[7] Beilstein, F. *Berichte* **1872,** *5,* 620.

On the **cooled** flat surface of the wire, place a drop (Pasteur pipet) of liquid unknown or a few milligrams of solid unknown. Gently heat the material in the flame. The carbon present in the compound will burn first, so the flame will be luminous, but then the characteristic green or blue-green color may be evident. It may be fleeting, so watch carefully.

It is recommended that a known compound containing a halogen be tested so that you become familiar with the appearance of the expected color.

Fluoride ion is not detected by this test, since copper fluoride is not volatile.

SOLUBILITY CHARACTERISTICS

Determination of the solubility characteristics of an organic compound can often give valuable information as to its structural composition. It is especially useful when correlated with spectral analysis. Several schemes have been proposed that place a substance in a definite group according to its solubility in various solvents. The scheme presented below is similar to that outlined in Shriner et al.[8]

There is no sharp dividing line between soluble and insoluble, and an arbitrary ratio of solute to solvent must be selected. We suggest that a compound be classified as soluble if its solubility is greater than 15 mg/500 μL of solvent.

Carry out the solubility determinations, at ambient temperature, in 10 × 75-mm test tubes. Place the sample (~15 mg) in the test tube and add a total of 0.5 mL of solvent in three portions from a graduated or calibrated Pasteur pipet. Between addition of each portion, stir the sample vigorously with a glass stirring rod for 1.5–2 min. If the sample is water soluble, test the solution with litmus paper to assist in classification according to the solubility scheme that follows.

NOTE. To test with litmus paper, dip the end of a small glass rod into the solution and then gently touch the litmus paper with the rod. ***Do not dip the litmus paper into the test solution.***

In doing the solubility tests follow the scheme in the order given. *Keep a record of your observations.*

Step I Test for water solubility. If soluble, test the solution with litmus paper.

Step II If water soluble, determine the solubility in diethyl ether. This test further classifies water-soluble materials.

Step III Water-insoluble compounds are tested for solubility in a 5% aqueous NaOH solution. If soluble, determine the solubility in 5% aqueous $NaHCO_3$. The use of the $NaHCO_3$ solution aids in distinguishing between strong (soluble) and weak (insoluble) acids.

Step IV Compounds insoluble in 5% aqueous NaOH are tested for solubility in a 5% HCl solution.

---

[8] Shriner, R. L.; Fuson, R. C.; Morrill, T. C. *The Systematic Identification of Organic Compounds,* 6th ed.; Wiley: New York, 1980.

JWCL196_c09_629-664.qxd 11/16/09 9:48 PM Page 639

EQA

Step V Compounds insoluble in 5% aqueous HCl are tested with concentrated H_2SO_4. If soluble, further differentiation is made using 85% H_3PO_4, as shown in the scheme.

Step VI Miscellaneous neutral compounds containing oxygen, sulfur, or nitrogen are normally soluble in strong acid solution:

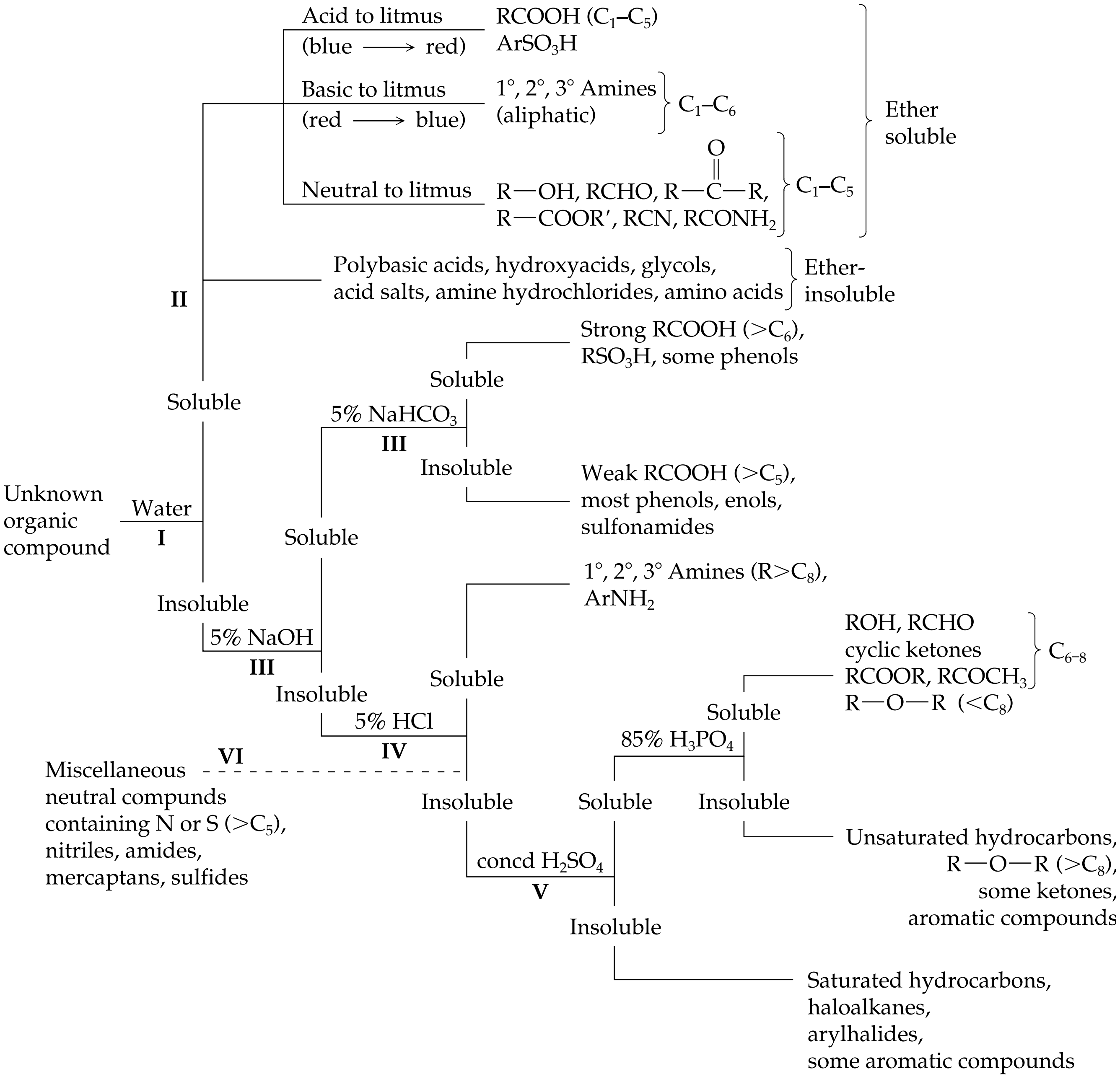
Acid to litmus
(blue ⟶ red)
RCOOH (C_1–C_5)
$ArSO_3H$
Basic to litmus
(red ⟶ blue)
1°, 2°, 3° Amines
(aliphatic)
C_1–C_6
Ether
soluble
Neutral to litmus
R—OH, RCHO, $R-\overset{O}{\overset{\|}{C}}-R$,
R—COOR′, RCN, $RCONH_2$
C_1–C_5
Polybasic acids, hydroxyacids, glycols,
acid salts, amine hydrochlorides, amino acids
Ether-
insoluble
II
Strong RCOOH (>C_6),
RSO_3H, some phenols
Soluble
Soluble
5% $NaHCO_3$
III
Insoluble
Weak RCOOH (>C_5),
most phenols, enols,
sulfonamides
Unknown
organic
compound
Water
I
Soluble
1°, 2°, 3° Amines (R>C_8),
$ArNH_2$
Insoluble
ROH, RCHO
cyclic ketones
RCOOR, $RCOCH_3$
R—O—R (<C_8)
$C_{6–8}$
5% NaOH
III
Soluble
Insoluble
5% HCl
IV
Soluble
85% H_3PO_4
Miscellaneous
VI
neutral compunds
containing N or S (>C_5),
nitriles, amides,
mercaptans, sulfides
Insoluble
Soluble
Insoluble
Unsaturated hydrocarbons,
R—O—R (>C_8),
some ketones,
aromatic compounds
concd H_2SO_4
V
Insoluble
Saturated hydrocarbons,
haloalkanes,
arylhalides,
some aromatic compounds

To classify a given compound, it may not be necessary to test its solubility in every solvent. *Do only those tests that are required to place the compound in one of the solubility groups.* Make your observations with care, and proceed in a logical sequence as you make the tests.

CONFIRMING PAGES aptara

JWCL196_c09_629-664.qxd 11/16/09 9:48 PM Page 640

EQA

THE CLASSIFICATION TESTS[9]

NOTE. For all tests given in this section, drops of reagents are measured out using Pasteur pipets.

Alcohols

Ceric Nitrate Test

INSTRUCTOR PREPARATION. The reagent is prepared by dissolving 4.0 g of ceric ammonium nitrate [$(NH_4)_2Ce(NO_3)_6$] in 10 mL of 2 M HNO_3. Warming may be necessary.

Primary, secondary, and tertiary alcohols with fewer than 10 carbon atoms give a positive test as indicated by a change in color from *yellow* to *red*:

$$\underset{\text{Yellow}}{(NH_4)_2Ce(NO_3)_6} + RCH_2OH \longrightarrow \underset{\text{(Red complex)}}{[\text{alcohol} + \text{reagent}]}$$

Place 5 drops of test reagent on a white spot plate. Add 1–2 drops of the unknown sample (5 mg if a solid). Stir with a thin glass rod to mix the components and observe any color change.

1. If the alcohol is water insoluble, 3–5 drops of dioxane may be added, but run a blank to make sure the dioxane is pure. Efficient stirring gives positive results with most alcohols.

2. Phenols, if present, give a brown color or precipitate.

Chromic Anhydride Test: The Jones Oxidation

INSTRUCTOR PREPARATION. The reagent is prepared by slowly adding a suspension of 1.0 g of CrO_3 in 1.0 mL of concentrated H_2SO_4 to 3 mL of water. Allow the solution to cool to room temperature before using.

The Jones oxidation test is a rapid method to distinguish primary and secondary alcohols from tertiary alcohols. A positive test is indicated by a color change from *orange* (the oxidizing agent, Cr^{6+}) while the oxidizing agent is itself reduced to the *blue green* (Cr^{3+}):

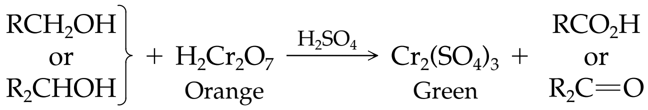

$$\left.\begin{matrix} RCH_2OH \\ \text{or} \\ R_2CHOH \end{matrix}\right\} + \underset{\text{Orange}}{H_2Cr_2O_7} \xrightarrow{H_2SO_4} \underset{\text{Green}}{Cr_2(SO_4)_3} + \begin{matrix} RCO_2H \\ \text{or} \\ R_2C{=}O \end{matrix}$$

The test is based on oxidation of a primary alcohol to an aldehyde or acid, and of a secondary alcohol to a ketone.

On a white spot plate, place 1 drop of the liquid unknown (10 mg if a solid). Add 10 drops of acetone and stir the mixture with a thin glass rod. Add

---

[9] For a detailed discussion of classification tests see (a) Shriner, R. L.; Fuson, R. C.; Morrill, T. C. *The Systematic Identification of Organic Compounds,* 6th ed.; Wiley: New York, 1980, p. 138; (b) Pasto, D. J.; Johnson, C. R.; Miller, M. J. *Experiments and Techniques in Organic Chemistry;* Prentice Hall: Englewood Cliffs, NJ, 1992.

1 drop of the test reagent to the resulting solution. Stir and observe any color change within a 2-second time period.

1. Run a blank to make sure the acetone is pure.

2. Tertiary alcohols, unsaturated hydrocarbons, amines, ethers, and ketones give a negative test within the 2-s time frame for observing the color change. Aldehydes, however, give a positive test, since they are oxidized to the corresponding carboxylic acids.

The HCl/ZnCl$_2$ Test: The Lucas Test

INSTRUCTOR PREPARATION. The Lucas reagent is prepared by dissolving 16 g of anhydrous ZnCl$_2$ in 10 mL of concd HCl while it is cooling in an ice bath.

The Lucas test is used to distinguish between primary, secondary, and tertiary monofunctional alcohols having fewer than six carbon atoms:

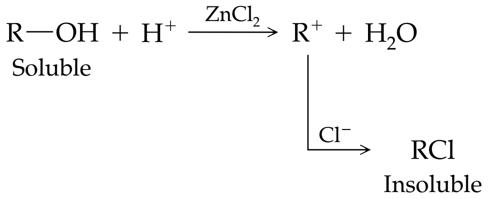

The test requires that the alcohol initially be soluble in the Lucas test reagent solution. As the reaction proceeds, the corresponding alkyl chloride is formed, which is insoluble in the reaction mixture. As a result, the solution becomes cloudy. In some cases a separate layer may be observed.

1. Tertiary, allyl, and benzyl alcohols react to give an immediate cloudiness to the solution. You may be able to see a separate layer of the alkyl chloride after a short time.

2. Secondary alcohols generally produce a cloudiness within 3–10 min. The solution may have to be heated to obtain a positive test.

3. Primary alcohols having less than six carbon atoms dissolve in the reagent but react very, very slowly. Those having more than six carbon atoms do not dissolve to any significant extent, no reaction occurs, and the aqueous phase remains clear.

4. A further test to aid in distinguishing between tertiary and secondary alcohols is to run the test using concentrated hydrochloric acid. Tertiary alcohols react immediately to give the corresponding alkyl halide, whereas secondary alcohols do not react under these conditions.

In a small test tube prepared by sealing a Pasteur pipet off at the shoulder (➨), place 2 drops of the unknown (10 mg if a solid) followed by 10 drops of the Lucas reagent.

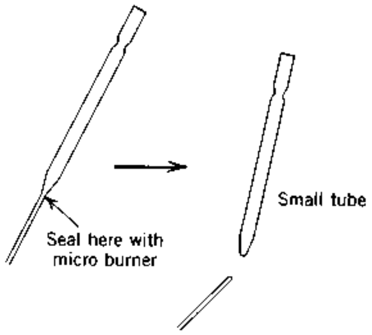

Shake or stir the mixture with a thin glass rod and allow the solution to stand. Observe the results. Based on the times given above, classify the alcohol.

Additional points to consider:

1. Certain polyfunctional alcohols also give a positive test.

2. If an alcohol having three or fewer carbons is expected, a 1-mL conical vial equipped with an air condenser should be used to prevent low molecular weight alkyl chlorides (volatile) from escaping and thus remaining undetected.

The Iodoform Test. This test is positive for compounds that on oxidation generate methyl ketones (or acetaldehyde) under the reaction conditions. For example, methyl carbinols (secondary alcohols having at least one methyl group attached to the carbon atom to which the —OH is attached), acetaldehyde, and ethanol give positive results.

For the test see Methyl Ketones and Methyl Carbinols (p. 651).

Periodic Acid: Vicinal Diols

INSTRUCTOR PREPARATION. This reagent solution is prepared by dissolving 250 mg of periodic acid (H_5IO_6) in 50 mL of deionized water.

Vicinal diols (1,2 diols) are differentiated from the simple alcohols by the characteristic reaction below. Metaperiodic acid (HIO_4) selectively oxidizes 1,2-diols to give carbonyl compounds:

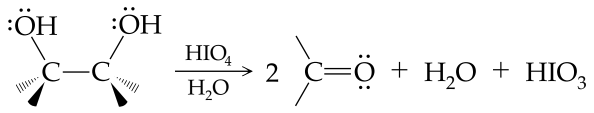

1,2-Diol

The test is based on the *instantaneous* formation of a white precipitate of silver iodate ($AgIO_3$) following addition of silver nitrate:

$$HIO_3 + AgNO_3 \longrightarrow HNO_3 + AgIO_3\downarrow$$

Place 2 mL of the periodic acid reagent solution in a small test tube.

Add 2 drops of concentrated nitric acid and mix the solution thoroughly. Add 2 drops of a liquid unknown (~2–5 mg of a solid) and mix again. Now add 2–3 drops of 5% aqueous silver nitrate solution. An *instantaneous white precipitate* constitutes a positive test.

α-Hydroxyaldehydes, α-hydroxyketones, α-hydroxyacids, 1,2-diketones, and α-aminoalcohols also give a positive test.

Aldehydes and Ketones

The 2,4-Dinitrophenylhydrazine Test

INSTRUCTOR PREPARATION. The reagent solution is prepared by dissolving 1.0 g of 2,4-dinitrophenylhydrazine in 5.0 mL of concentrated sulfuric acid. This solution is slowly added, with stirring, to a mixture of 10 mL of water and 35 mL of 95% ethanol. After mixing, filter the solution.

Aldehydes and ketones react rapidly with 2,4-dinitrophenylhydrazine to form 2,4-dinitrophenylhydrazones. These derivatives range in color from *yellow* to *red*, depending on the degree of conjugation in the carbonyl compound:

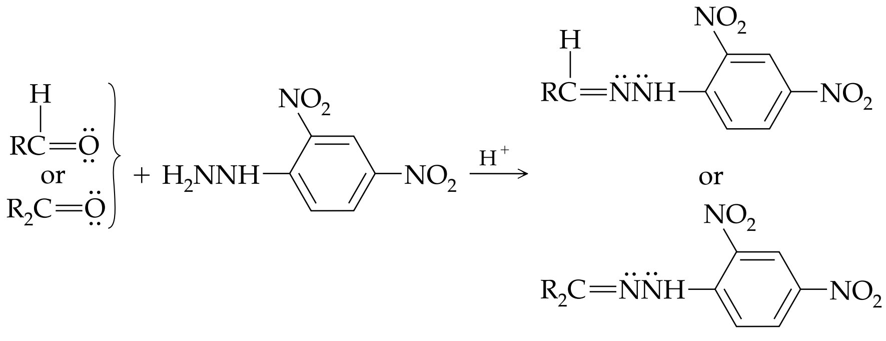

2,4-Dinitrophenylhydrazine

Yellow-to-red precipitate

On a white spot plate place 7–8 drops of 2,4-dinitrophenylhydrazine reagent solution.

Then add 1 drop of a liquid unknown. If the unknown is a solid, add 1 drop of a solution prepared by dissolving 10 mg of the material in 10 drops of ethanol. The mixture is stirred with a thin glass rod. The formation of a red-to-yellow precipitate is a positive test.

NOTE. The reagent, 2,4-dinitrophenylhydrazine, is orange-red and melts at 198 °C (dec). Do not mistake it for a derivative!

Reactive esters or anhydrides react with the reagent to give a positive test. Allylic or benzylic alcohols may be oxidized to aldehydes or ketones, which in turn give a positive result. Amides do not interfere with the test. Be sure that your unknown is pure and does not contain aldehyde or ketone impurities.

Phenylhydrazine and *p*-nitrophenylhydrazine are often used to prepare the corresponding hydrazones. These reagents also yield solid derivatives of aldehydes and ketones.

Silver Mirror Test for Aldehydes: Tollens Reagent. This reaction involves the oxidation of aldehydes to the corresponding carboxylic acid, using an alcoholic solution of silver ammonium hydroxide. A positive test is the formation of a *silver* mirror, or a black precipitate of finely divided silver:

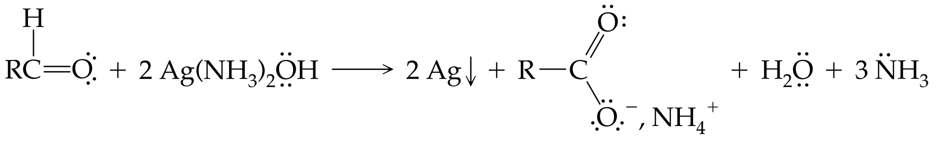

The test should be run only after the presence of an aldehyde or ketone has been established.

In a small test tube prepared from a Pasteur pipet (see the Lucas test) place 1.0 mL of a 5% aqueous solution of $AgNO_3$, followed by 1 drop of aqueous 10% NaOH solution. Now add concentrated aqueous ammonia, drop by drop (2–4 drops) with shaking, until the precipitate of silver oxide just dissolves. Add 1 drop of the unknown (10 mg if a solid), with shaking, and allow the reaction mixture to stand for 10 min at room temperature. If no reaction has occurred, place the test tube in a sand bath at 40 °C for 5 min. Observe the result.

Additional points to consider:

1. Avoid a large excess of ammonia.

2. Reagents must be well mixed. Stirring with a thin glass rod is recommended.

3. *This reagent is freshly prepared for each test. It should not be stored since decomposition occurs with the formation of AgN_3, which is explosive.*

CAUTION

4. This oxidizing agent is very mild and thus alcohols are not oxidized under these conditions. Ketones do not react. Some sugars, acyloins, hydroxylamines, and substituted phenols do give a positive test.

Chromic Acid Test

INSTRUCTOR PREPARATION. The reagent is prepared by dissolving 1 g of chromium trioxide in 1 mL of concd H_2SO_4, followed by 3 mL of H_2O.

Chromic acid in acetone rapidly oxidizes aldehydes to carboxylic acids. Ketones react very slowly, or not at all.

In a 3-mL vial or small test tube, place 2 drops of a liquid unknown (~10 mg if a solid) and 1 mL of spectral-grade acetone. Now add several drops of the chromic acid reagent.

A green precipitate of chromous salts is a positive test. Aliphatic aldehydes give a precipitate within 30 s; aromatic aldehydes take 30–90 s.

The reagent also reacts with primary and secondary alcohols (see Chromic Anhydride Test: Jones Oxidation, p. 640).

Bisulfite Addition Complexes

INSTRUCTOR PREPARATION. The reagent is prepared by mixing 1.5 mL of ethanol and 6 mL of a 40% aqueous solution of sodium bisulfite. Filter the reagent before use, if a small amount of the salt does not dissolve.

Most aldehydes react with a saturated sodium bisulfite solution to yield a crystalline bisulfite addition complex:

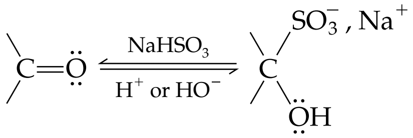

The reaction is reversible and thus the carbonyl compound can be recovered by treatment of the complex with aqueous 10% $NaHCO_3$ or dilute HCl solution.

Place 50–75 μL of the liquid unknown in a small test tube and add 150 μL of the sulfite reagent and mix thoroughly.

A crystalline precipitate is a positive test.

Alkyl methyl ketones and unhindered cyclic ketones also give a positive test.

Alkanes and Cycloalkanes: Saturated Hydrocarbons

Iodine Charge-Transfer Complex. Alkanes exhibit a *negative* iodine charge-transfer complex test. Species containing π electrons or nonbonded electron pairs produce a brown solution. This color formation is due to the charge-transfer complex between iodine and the available electrons:

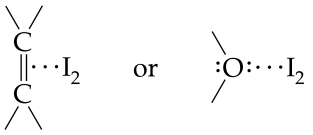

Solutions of iodine and nonparticipating compounds are violet in color.

On a white spot plate, place a small crystal of iodine. Now add 2–3 drops of a liquid unknown. Alkanes give a *negative* test (violet color).

The test is run only on liquid unknowns. Saturated hydrocarbons, fluorinated and chlorinated saturated hydrocarbons, and aromatic hydrocarbons and their halogenated derivatives all give violet solutions. All other species give a positive test (brown solution).

Concentrated Sulfuric Acid. Saturated hydrocarbons, halogenated saturated hydrocarbons, simple aromatic hydrocarbons, and their halogenated derivatives are insoluble in *cold* concentrated sulfuric acid.

CAUTION In a small test tube, using **caution,** place 100 μL of *cold* concentrated sulfuric acid. Now add 50 μL of an unknown. A resulting heterogeneous solution (the unknown does *not* dissolve) is a positive test for a saturated hydrocarbon.

Alkenes, and compounds having a functional group containing a nitrogen or oxygen atom, are soluble in cold, concentrated acid.

Alkenes and Alkynes: Unsaturated Hydrocarbons

Bromine in Methylene Chloride. Unsaturated hydrocarbons readily add bromine (Br_2). An example of this reaction is given in Experiment [F2]:

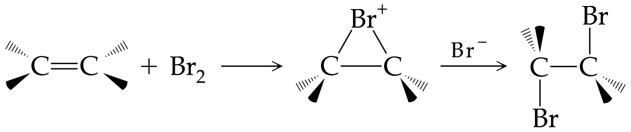

The test is based on the decolorization of a red-brown bromine–methylene chloride solution.

CAUTION: Bromine is highly toxic and can cause burns.

CAUTION

In a 10 × 75-mm test tube, or in a small tube prepared from a Pasteur pipet (see Lucas test), place 2 drops of a liquid unknown (~15 mg if a solid) followed by 0.5 mL of methylene chloride. Add dropwise, in the **hood** with shaking, a 2% solution of bromine in methylene chloride solvent. The presence of an unsaturated hydrocarbon will require 2–3 drops of the reagent before the reddish-brown color of bromine persists in the solution.

HOOD

Additional points to consider:

1. Methylene chloride is used in place of the usual carbon tetrachloride (CCl_4) because it is less toxic.

2. Phenols, enols, amines, aldehydes, and ketones interfere with this test.

Permanganate Test: Baeyer Test for Unsaturation. Unsaturation in an organic compound can be detected by the decolorization of permanganate solution. The reaction involves the cis hydroxylation of the alkene to give a 1,2 diol (glycol):

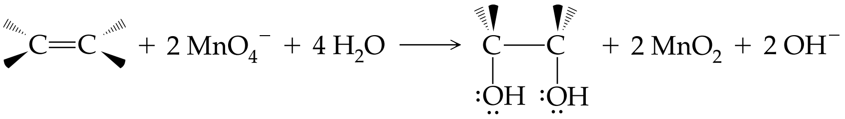

On a white spot plate, place 0.5 mL of *alcohol-free* acetone, followed by 2 drops of the unknown compound (~15 mg if a solid). Now add dropwise (2–3 drops), with stirring, a 1% aqueous solution of potassium permanganate ($KMnO_4$). A positive test for unsaturation is the discharge of purple permanganate color from the reagent and the precipitation of brown manganese oxides.

Any functional group that undergoes oxidation with permanganate interferes with the test (phenols, aryl amines, most aldehydes, primary and secondary alcohols, etc.).

Alkyl Halides

Silver Nitrate Test. Alkyl halides that undergo the S_N1 substitution reaction react with alcoholic silver nitrate ($AgNO_3$) to form a precipitate of the corresponding silver halide.

Secondary and primary halides react slowly or not at all at room temperature. However, they do react at elevated temperatures. Tertiary halides react immediately at room temperature.

In a 1.0-mL conical vial place 0.5 mL of 2% ethanolic $AgNO_3$ solution and 1 drop of unknown (~10 mg if a solid). A positive test is indicated by the appearance of a precipitate within 5 min. If no reaction occurs, add a boiling stone and equip the vial with an air condenser. Heat the solution at *gentle* reflux for an additional 5 min using a sand bath. Cool the solution.

If a precipitate is formed, add 2 drops of dilute HNO_3. Silver halides will not dissolve in nitric acid solution.

Additional points to consider:

1. The order of reactivity for R groups is allyl ≅ benzyl > tertiary > secondary > > > primary. For the halide leaving groups the order is I > Br > Cl.

2. Acid halides, α-haloethers, and 1,2-dibromo compounds also give a positive test at room temperature. Only activated aryl halides give a positive test at elevated temperatures.

Sodium Iodide in Acetone

INSTRUCTOR PREPARATION. The reagent is prepared by dissolving 3 g of sodium iodide (NaI) in 25 mL of acetone. Store in a dark bottle.

Primary alkyl chlorides and bromides can be distinguished from aryl and alkenyl halides by reaction with sodium iodide in acetone (Finkelstein reaction):

$$\mathrm{R{-}X + NaI \xrightarrow{\text{acetone}} R{-}I + NaX}\downarrow$$
$$\mathrm{X = Cl, Br}$$

Primary alkyl bromides undergo an S_N2 displacement reaction within 5 min at room temperature, and primary alkyl chlorides only at 50 °C.

In a 1.0-mL conical vial, place 1 drop of a liquid unknown (~10 mg if a solid) and 3 drops of acetone. To this solution add 0.5 mL of sodium iodide–acetone reagent.

A positive test is the appearance of a precipitate of NaX within 5 min. If no precipitate is observed, add a boiling stone and equip the vial with an air condenser. Warm the reaction mixture in a sand bath at about 50 °C for 5 min. Cool to room temperature and determine whether a reaction has occurred.

Additional points to consider:

1. Benzylic and allylic chlorides and bromides, acid chlorides and bromides, and α-haloketones, α-haloesters, α-haloamides, and α-halonitriles also give a positive test at room temperature.

2. Primary and secondary alkyl chlorides, and secondary and tertiary alkyl bromides, react at 50 °C under these conditions.

3. If the solution turns red brown in color, iodine is being liberated.

Amides, Ammonium Salts, and Nitriles

Hydroxamate Test for Amides. Unsubstituted (on nitrogen) amides, and the majority of substituted amides, will give a positive hydroxamate test:

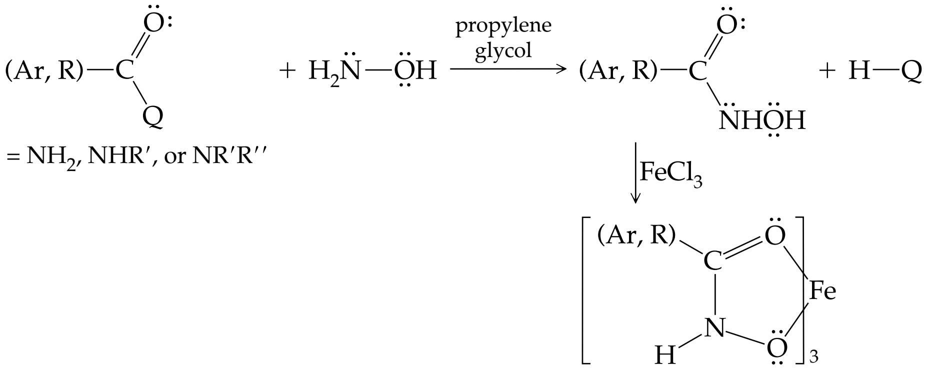

where Q = NH_2, NHR′, or NR′R′′

Red violet

The hydroxamic acid is identified by formation of a red-to-purple color in the presence Fe^{3+} of ion, as for the test with esters (see page 650).

In a 3.0-mL conical vial containing a boiling stone and equipped with an air condenser place 1 drop of a liquid unknown (~10 mg if a solid), followed by 0.5 mL of 1 M hydroxylamine hydrochloride–propylene glycol solution. Heat the resulting mixture to reflux temperature (~190 °C) using a sand bath, and reflux for 3–5 min. Cool the solution to room temperature, and add 2 drops of 5% aqueous $FeCl_3$ solution. The formation of a red-to-purple color is a positive test.

Alkaline Hydrolysis. Ammonium salts, amides, and nitriles undergo hydrolysis in alkaline solution to form ammonia gas, or an amine:

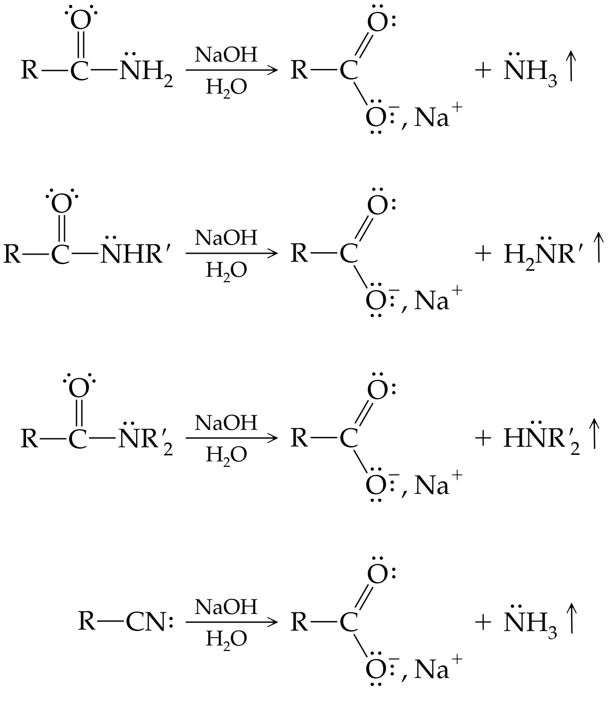

$$R{-}C({=}O){-}NH_2 \xrightarrow[H_2O]{NaOH} R{-}C({=}O){-}O^-, Na^+ + NH_3\uparrow$$

$$R{-}C({=}O){-}NHR' \xrightarrow[H_2O]{NaOH} R{-}C({=}O){-}O^-, Na^+ + H_2NR'\uparrow$$

$$R{-}C({=}O){-}NR'_2 \xrightarrow[H_2O]{NaOH} R{-}C({=}O){-}O^-, Na^+ + HNR'_2\uparrow$$

$$R{-}CN \xrightarrow[H_2O]{NaOH} R{-}C({=}O){-}O^-, Na^+ + NH_3\uparrow$$

Detection of ammonia from ammonium salts, primary amides, and nitriles, by use of a color test using copper sulfate solution, constitutes a positive test for these functional groups. The same test may also be used for secondary and tertiary amides that can generate low molecular weight (volatile) amines upon hydrolysis.

In a 1.0-mL conical vial containing a boiling stone, and equipped with an air condenser, place 1–2 drops of the unknown liquid (~10 mg if a solid) and 0.5 mL of 20% aqueous NaOH solution. Heat this mixture to *gentle* reflux on a sand bath. Moisten a strip of filter paper with 2 drops of 10% aqueous copper sulfate solution and place it over the top of the condenser. Formation of a *blue* color (copper ammonia [or amine] complex) is a positive test.

The filter paper may be held in place using a small test tube holder or other suitable device.

Amines

Copper Ion Test. Amines will give a blue-green coloration or precipitate when added to a copper sulfate solution. In a small test tube, place 0.5 mL of a 10% copper sulfate solution. Now add 1 drop of an unknown (~10 mg if a solid). The blue-green coloration or precipitate is a positive test. Ammonia will also give a positive test.

Hinsberg Test. The Hinsberg test is useful for distinguishing between primary, secondary, and tertiary amines. The reagent used is *p*-toluenesulfonyl chloride in alkaline solution.

Primary amines with fewer than seven carbon atoms form a sulfonamide that is soluble in the alkaline solution. Acidification of the solution results in the precipitation of the insoluble sulfonamide:

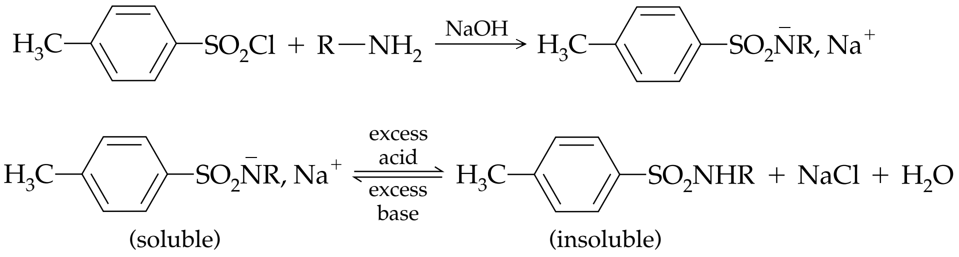
H$_3$C—⬡—SO$_2$Cl + R—NH$_2$
NaOH
H$_3$C—⬡—SO$_2$N̄R, Na$^+$
H$_3$C—⬡—SO$_2$N̄R, Na$^+$
excess acid
excess base
H$_3$C—⬡—SO$_2$NHR + NaCl + H$_2$O
(soluble)
(insoluble)

Secondary amines form an insoluble sulfonamide in the alkaline solution:

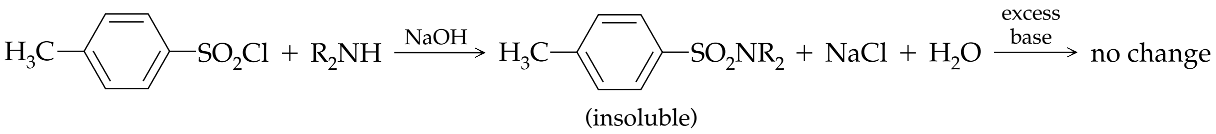
H$_3$C—⬡—SO$_2$Cl + R$_2$NH
NaOH
H$_3$C—⬡—SO$_2$NR$_2$ + NaCl + H$_2$O
excess base
no change
(insoluble)

Tertiary amines normally give no reaction under these conditions:

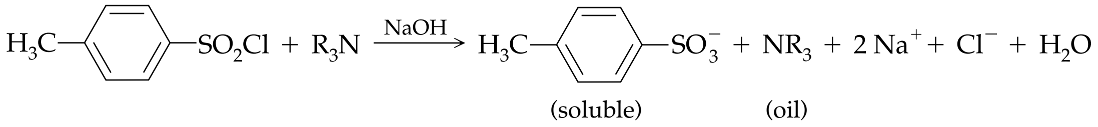
H$_3$C—⬡—SO$_2$Cl + R$_3$N
NaOH
H$_3$C—⬡—SO$_3^-$ + NR$_3$ + 2 Na$^+$ + Cl$^-$ + H$_2$O
(soluble)
(oil)

In a 1.0-mL conical vial containing a boiling stone, and equipped with an air condenser, place 0.5 mL of 10% aqueous sodium hydroxide solution, 1 drop of the sample unknown (~10 mg if a solid), followed by 30 mg of HOOD *p*-toluenesulfonyl chloride (in the **hood**). Heat the mixture to reflux for 2–3 min on a sand bath, and then cool it in an ice bath. Test the alkalinity of the solution using litmus paper. If it is not alkaline, add additional 10% aqueous sodium hydroxide dropwise.

Using a Pasteur filter pipet, separate the solution from any solid that may be present. Transfer the solution to a clean 1.0-mL conical vial and **save.**

SAVE

NOTE. If an oily upper layer is obtained at this stage, remove the lower alkaline phase using a Pasteur filter pipe and ***save.*** *To the remaining oil add 0.5 mL of cold water and stir vigorously to obtain a solid material.*

SAVE

If a solid is obtained, it may be (1) the sulfonamide of a secondary amine; (2) recovered tertiary amine, if the original amine was a solid; or (3) the insoluble salt of a primary sulfonamide derivative, if the original amine had more than six carbon atoms.

Additional points to consider:

1. If the solid is a tertiary amine, it is soluble in aqueous 10% HCl.
2. If the solid is a secondary sulfonamide, it is insoluble in aqueous 10% NaOH.
3. If no solid is present, acidify the alkaline solution by addition of 10% aqueous HCl. If the unknown amine is primary, the sulfonamide will precipitate.

Bromine Water. Aromatic amines, since they possess an electron-rich aromatic ring, can undergo electrophilic aromatic substitution with bromine, to yield the corresponding arylamino halide(s). Therefore, if elemental tests indicate that an aromatic group is present in an amine, treatment with the bromine water reagent may indicate that the amine is attached to an aromatic ring.

For the test, see Phenols and Enols (p. 653).

Aromatic Hydrocarbons with NO Functional Groups

Fuming Sulfuric Acid. Simple aromatic hydrocarbons are insoluble in sulfuric acid (H_2SO_4) but are soluble in fuming sulfuric acid. If these hydrocarbons contain more than two alkyl substituents, they may be sulfonated under these conditions.

In a small test tube place 100 μL of fuming sulfuric acid, using **caution.** Now add 50 μL of the unknown suspected to be aromatic. A resulting homogeneous solution is a positive test.

CAUTION

Azoxybenzene and Aluminum Chloride. This color test is run only on those aromatic compounds that are insoluble in sulfuric acid (see previous test). The color produced in this test results from the formation of a complex of $AlCl_3$ and a *p*-arylazobenzene derivative:

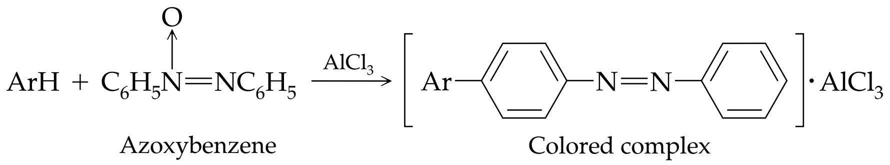

In a small dry test tube, place 250 μL of the aromatic unknown. Add a small crystal of azoxybenzene and about 12 mg of anhydrous aluminum chloride. If a color is not produced immediately, warm the mixture for a few minutes.

Aryl halides and other simple aromatic hydrocarbons give a deep-orange to dark-red color or precipitate. Polynuclear aromatic hydrocarbons, such as naphthalenes and anthracenes, give brown colors. Aliphatic hydrocarbons give no color, or at most a light yellow tint.

Carboxylic Acids

The presence of a carboxylic acid is detected by its solubility behavior. An aqueous solution of the acid will be acidic to litmus paper (or pH paper may be used). Since a sulfonic acid would also give a positive test, the test for sulfur (sodium fusion) is used to distinguish between the two types of acids. A water-soluble phenol is acidic toward litmus paper but also would give a positive ferric chloride test.

Carboxylic acids also react with a 5% solution of sodium bicarbonate (see Experiment [4B]).

Place 1–2 mL of the bicarbonate solution on a watch glass and add 1–2 drops of the acid (~10 mg if a solid). Gas bubbles of CO_2 constitute a positive test.

Esters

Hydroxamate Test. Carboxylic esters can be identified by conversion to hydroxamic acid salts. Acidification of this salt produces the corresponding hydroxamic acid (RCONHOH), which is identified by formation of a red-to-purple color in the presence of Fe^{3+} ion:

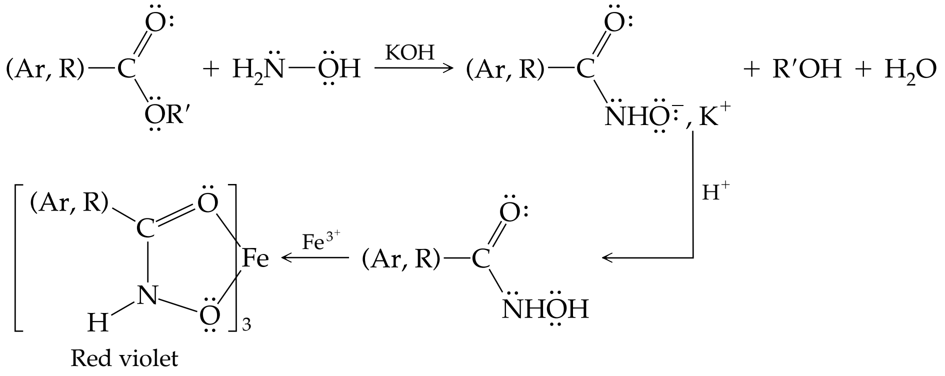

In a 3.0-mL conical vial containing a boiling stone, and equipped with an air condenser, place 1 drop of the liquid unknown (~10 mg if a solid) followed by 0.5 mL of 1.0 M ethanolic hydroxylamine hydrochloride solution. Add 10% methanolic KOH to this solution (dropwise) until the resulting solution has pH ~10 (pH paper). Heat this mixture to reflux temperature using a sand bath for 5 min, cool to room temperature, and acidify to pH = 3–4 by dropwise addition of 5% aqueous HCl solution. Now add 2 drops of 5% aqueous $FeCl_3$ solution. The formation of a red-to-purple color is a positive test.

Additional points to consider:

1. It is suggested that a blank be run for comparison purposes.
2. Acid chlorides, anhydrides, lactones, and imides also give a positive test.

Saponification. This well-known reaction of esters can often be used to classify these compounds. It also may lead to a useful derivative if the corresponding carboxylic acid is isolated.

In a 3.0-mL conical vial containing a magnetic spin vane, place 100 μL of the liquid unknown (~150 mg if a solid) and add 1 mL of 6 M NaOH solution. Attach the vial to a reflux condenser. Now place the vial in a sand bath on a magnetic stirring hot plate and, with stirring, heat the mixture at reflux for 0.5 h, or until the solution becomes homogeneous.

A positive test is the disappearance of the organic layer (if the original unknown was water insoluble) or the lack of the usually pleasant aroma of the unknown ester.

High-boiling esters (bp $>$ 200 °C) are usually not saponified under these conditions due to their low solubility in the aqueous solvent.

Ethers

CAUTION: Upon standing, ethers may form peroxides. Peroxides are very explosive. To test for the presence of these substances, use starch–iodide paper that has been moistened with 6 M HCl. Peroxides cause the paper to turn blue. To remove peroxides from ethers, pass the material through a short column of highly activated alumina (Woelm basic alumina, activity grade 1).[10] Always retest for peroxides before using the ether.

Ferrox Test. The ferrox test is a color test sensitive to oxygen, which may be used to distinguish ethers from hydrocarbons that, like most ethers, are soluble in sulfuric acid.

In a dry 10 × 75-mm test tube using a glass stirring rod, grind a crystal of ferric ammonium sulfate and a crystal of potassium thiocyanate. The ferric hexathiocyanatoferrate that is formed adheres to the rod.

In a second clean 10 × 75-mm test tube, place 2–3 drops of a liquid unknown. If dealing with a solid, use about 10 mg and add toluene until a saturated solution is obtained. Now, using the rod with the ferric hexathiocyanatoferrate attached, stir the unknown. *If the unknown contains oxygen, the ferrate compound dissolves and a reddish-purple color is observed.*

Some high-molecular-weight ethers do not give a positive test.

Bromine Water. Since the aromatic ring is electron rich, aromatic ethers can undergo electrophilic aromatic substitution with bromine to yield the corresponding aryl ether–halide(s). Therefore, if elemental tests indicate that an aromatic group is present in an ether, treatment with the bromine water reagent may substantiate the presence of an aryl ether.

For the test see Phenols and Enols (p. 653).

Methyl Ketones and Methyl Carbinols

Iodoform Test

INSTRUCTOR PREPARATION. Dissolve 3 g of KI and 1 g I_2 in 20 mL of water.

---

[10] Pasto, D. J.; Johnson, C. R.; Miller, M. J. *Experiments and Techniques in Organic Chemistry;* Prentice Hall: Englewood Cliffs, NJ, 1992, p. 33.

The iodoform test involves hydrolysis and cleavage of methyl ketones to form a yellow precipitate of iodoform (CHI_3):

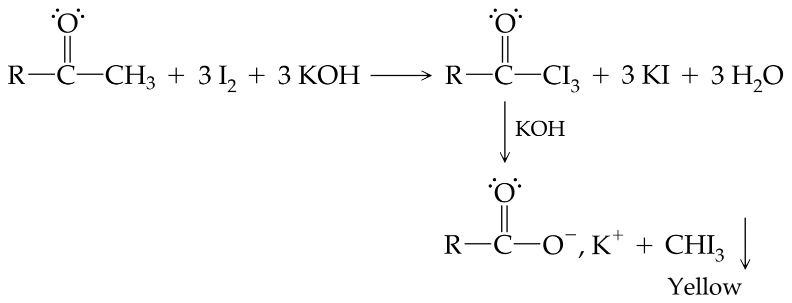

$$R-\overset{\ddot{O}}{\overset{\|}{C}}-CH_3 + 3\,I_2 + 3\,KOH \longrightarrow R-\overset{\ddot{O}}{\overset{\|}{C}}-CI_3 + 3\,KI + 3\,H_2O$$

$$\Big\downarrow \text{KOH}$$

$$R-\overset{\ddot{O}}{\overset{\|}{C}}-O^-, K^+ + \underset{\text{Yellow}}{CHI_3\downarrow}$$

It is also a positive test for compounds that, upon oxidation, generate methyl ketones (or acetaldehyde) under these reaction conditions. For example, methyl carbinols (secondary alcohols having at least one methyl group attached to the carbon atom to which the —OH unit is linked), acetaldehyde, and ethanol give positive results.

In a 3.0-mL conical vial equipped with an air condenser, place 2 drops of the unknown liquid (10 mg if a solid), followed by 5 drops of 10% aqueous KOH solution.

HOOD

NOTE. If the sample is insoluble in the aqueous phase, either mix vigorously or add dioxane (in the ***hood****) or bis(2-methoxyethyl) ether to obtain a homogeneous solution.*

Warm the mixture on a sand bath to 50–60 °C and add the KI–I_2 reagent dropwise until the solution becomes dark brown in color (~1.0 mL). Additional 10% aqueous KOH is now added (dropwise) until the solution is again colorless.

CAUTION

CAUTION: Iodine is highly toxic and can cause burns.

After warming for 2 min, cool the solution and determine whether a yellow precipitate (CHI_3, iodoform) has formed. If a precipitate is not observed, reheat as before for another 2 min. Cool and check again for the appearance of iodoform.

Additional points to consider:

1. The iodoform test is reviewed elsewhere.[11]
2. An example of the general haloform reaction, using bleach to oxidize a methyl ketone, is given in Experiment [34].

Nitro Compounds

Ferrous Hydroxide Test. Many nitro compounds give a positive test based on the following reaction:

$$R-NO_2 + 4\,H_2O + 6\,Fe(OH)_2 \longrightarrow R-NH_2 + \underset{\text{Red-brown}}{6\,Fe(OH_3)\downarrow}$$

The nitro derivative oxidizes the iron(II) hydroxide to iron(III) hydroxide; the latter is a red-brown solid.

In a 1.0-mL conical vial place 5–10 mg of the unknown compound, followed by 0.4 mL of freshly prepared 5% aqueous ferrous ammonium sulfate solution. After mixing, add 1 drop of 3 M sulfuric acid followed by 10 drops of 2 M methanolic KOH. Cap the vial, shake vigorously, vent, and then allow it to stand over a 5-min period. The formation of a red-brown precipitate, usually within 1 min, is a positive test for a nitro group.

---

[11] Fuson, R. C.; Bull, B. A. *Chem. Rev.* **1934,** *15,* 275.

Sodium Hydroxide Color Test. Treatment of an aromatic nitro compound with 10% sodium hydroxide solution may often be used to determine the number of nitro groups present on the aromatic ring system.

Mononitro compounds produce no color (a light yellow may be observed).
Dinitro compounds produce a bluish-purple color.
Trinitro compounds produce a blood-red color.

The color formation is due to formation of Meisenheimer complexes (for a discussion, see Pasto et al.[12]).

To run the test, dissolve 10 mg of the unknown (1–2 drops if a liquid) in 1 mL of acetone in a small test tube. Now add about 200 μL of 10% NaOH solution and shake. Observe any color formation.

If amino, substituted amino, or hydroxyl groups are present in the molecule, a positive color test is not obtained.

Phenols and Enols

Ferric Ion Test. Most phenols and enols form colored complexes in the presence of ferric ion, Fe^{3+}:

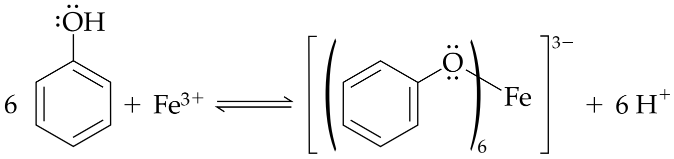

Phenols give red, blue, purple, or green colors. Sterically hindered phenols may give a negative test. Enols generally give a tan, red, or red-violet color.

On a white spot plate place 2 drops of water, or 1 drop of water plus 1 drop of ethanol, or 2 drops of ethanol, depending on the solubility characteristics of the unknown. To this solvent system add 1 drop (10 mg if a solid) of the substance to be tested. Stir the mixture with a thin glass rod to complete dissolution. Add 1 drop of 2.5% aqueous ferric chloride ($FeCl_3$) solution (light yellow in color). Stir and observe any color formation. If necessary, a second drop of the $FeCl_3$ solution may be added.

Additional points to consider:

1. The color developed may be fleeting or it may last for many hours. A slight excess of the ferric chloride solution may or may not destroy the color.
2. An alternative procedure using $FeCl_3$–CCl_4 solution in the presence of pyridine is available.[13]

---

[12] Pasto, D. J.; Johnson, C. R.; Miller, M. J. *Experiments and Techniques in Organic Chemistry;* Prentice Hall: Englewood Cliffs, NJ, 1992, p. 321.
[13] Soloway, S.; Wilen, S. H. *Anal. Chem.* **1952,** *4,* 979.

Bromine Water. Phenols, substituted phenols, aromatic ethers, and aromatic amines, since the aromatic rings are electron rich, undergo aromatic electrophilic substitution with bromine to yield substituted aryl halides. For example,

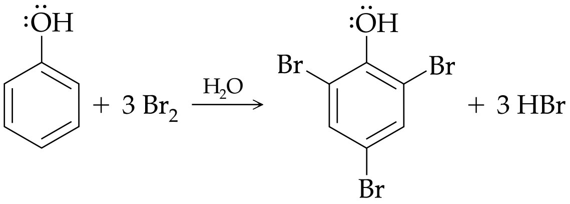

HOOD CAUTION: The test should be run in the **hood.**

In a small test tube, place 1–2 drops of the unknown (~20 mg if a solid) and add 1–2 mL of water. Check the pH of the solution with pH paper. In the HOOD **hood,** add saturated bromine water dropwise until the bromine color persists. A precipitate generally forms.

A positive test is the decolorization of the bromine solution, and often the formation of an off-white precipitate. If the unknown is a phenol, this should cause the pH of the original solution to be less than 7.

PREPARATION OF DERIVATIVES

Based on the preliminary and classification tests carried out to this point, you should have established the type of functional group (or groups) present (or lack of one) in the unknown organic sample. The next step in qualitative organic analysis is to consult a set of tables containing a listing of known organic compounds sorted by functional group and/or by physical properties or by both. Using the physical properties data for your compound, you can select a few possible candidates that appear to "fit" the data you have collected. On a chemical basis, the final step in the qualitative identification sequence is to prepare one or two *crystalline derivatives* of your compound. Selection of the specific compound, and thus final confirmation of its identity, can then be made from the extensive derivative tables that have been accumulated. With the advent of spectral analysis, the preparation of derivatives is often not necessary, but the wealth of chemistry that can be learned by the beginning student in carrying out these procedures is extensive and important. The preparation of selected derivatives for the most common functional groups are given below. Condensed tables of compounds and their derivatives are summarized www → on the website, in Chapter 9W. For extensive tables and alternative derivatives that can be utilized, see the following Bibliography.

BIBLIOGRAPHY

Pasto, D. J.; Johnson, C. R.; Miller, M. J. *Experiments and Techniques in Organic Chemistry*; Prentice Hall: Englewood Cliffs, NJ, 1992.

Rappoport, Z. *Handbook of Tables for Organic Compound Identification,* 3rd ed.; CRC Press: Boca Raton, FL, 1967.

Shriner, R. L.; Hermann, C. K. F.; Morrill, T. C.; Fuson, R. C. *The Systematic Identification of Organic Compounds,* 8th ed.; Wiley: New York, 2003.

NOTE. In each of the procedures outlined below, drops of reagents are measured using Pasteur pipets.

JWCL196_c09_629-664.qxd 11/16/09 9:48 PM Page 655

EQA

CARBOXYLIC ACIDS[14]

Preparation of Acid Chlorides

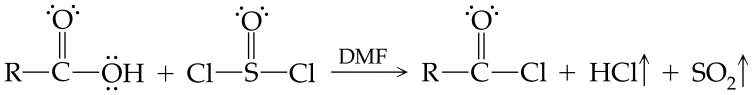

$$\mathrm{R{-}\overset{\overset{\ddot{O}}{\|}}{C}{-}\ddot{O}H + Cl{-}\overset{\overset{\ddot{O}}{\|}}{S}{-}Cl \xrightarrow{DMF} R{-}\overset{\overset{\ddot{O}}{\|}}{C}{-}Cl + HCl\uparrow + SO_2\uparrow}$$

Weigh and place 20 mg of the unknown acid in a dry 3.0-mL conical vial containing a boiling stone and fitted with a cap. Now, in the **hood,** add 4 drops of thionyl chloride and 1 drop of *N,N*-dimethylformamide (DMF). Immediately attach the vial to a reflux condenser that is protected by a calcium chloride drying tube.

HOOD

> CAUTION: This reaction is run in the **hood** since hydrogen chloride and sulfur dioxide are evolved. Thionyl chloride is an irritant and is harmful to breathe. Immediately recap the vial after each addition until the vial is attached to the reflux condenser.

HOOD

Allow the mixture to stand at room temperature for 10 min, heat it at gentle reflux on a sand bath for 15 min, and then cool it to room temperature. Dilute the reaction mixture with 5 drops of methylene chloride solvent.

The acid chloride is not isolated but is used directly in the preparations that follow.

Amides

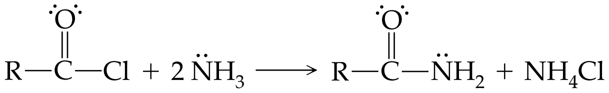

$$\mathrm{R{-}\overset{\overset{\ddot{O}}{\|}}{C}{-}Cl + 2\,\ddot{N}H_3 \longrightarrow R{-}\overset{\overset{\ddot{O}}{\|}}{C}{-}\ddot{N}H_2 + NH_4Cl}$$

Cool the vial in an ice bath and add 10 drops of concentrated aqueous ammonia, in the **hood** via Pasteur pipet, *dropwise*, with stirring. *It is convenient to make this addition down the neck of the air condenser.* The amide may precipitate during this operation. After the addition is complete, remove the ice bath and stir the mixture for an additional 5 min. Now add methylene chloride (10 drops) and stir the resulting mixture to dissolve any precipitate. Separate the methylene chloride layer from the aqueous layer using a Pasteur filter pipet and transfer it to another Pasteur filter pipet containing 200 mg of anhydrous sodium sulfate. Collect the eluate in a Craig tube containing a boiling stone. Extract the aqueous phase with an additional 0.5 mL of methylene chloride. Separate the methylene chloride layer as before and transfer it to the same column. Collect this eluate in the same Craig tube. Evaporate the methylene chloride solution using a warm sand bath in the **hood** under a gentle stream of nitrogen gas. Recrystallize the solid amide product using the Craig tube. Dissolve the material in about 0.5 mL of ethanol, add water (dropwise) until the solution becomes cloudy, cool the Craig tube in an ice bath, and collect the crystals in the usual manner. Dry the crystalline amide on a porous clay plate and determine the melting point.

HOOD

HOOD

---

[14] See Tables 9W.1 and 9W.2.

www

CONFIRMING PAGES aptara

Anilides

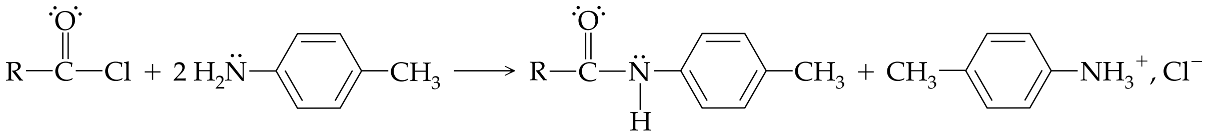

In a 3.0-mL conical vial containing a magnetic spin vane, and equipped with an air condenser, place 5 drops of aniline and 10 drops of methylene chloride. Cool the solution in an ice bath and transfer the acid chloride solution (prepared above) via Pasteur pipet, *dropwise*, with stirring, to the aniline solution in the **hood.** *It is convenient to make this addition down the neck of the condenser.* After the addition is complete, remove the ice bath and stir the mixture for an additional 10 min. HOOD

Transfer the methylene chloride layer to a 10 × 75-mm test tube, and wash it with 0.5 mL of H_2O, 0.5 mL of 5% aqueous HCl, 0.5 mL of 5% aqueous NaOH, and, finally, 0.5 mL of H_2O. For each washing, shake the test tube and remove the top aqueous layer by Pasteur filter pipet. Transfer the resulting wet methylene chloride layer to a Pasteur filter pipet containing 200 mg of anhydrous sodium sulfate. Collect the eluate in a Craig tube containing a boiling stone. Rinse the original test tube with an additional 10 drops of methylene chloride. Collect this rinse and pass it through the same column. Both eluates are combined.

Evaporate the methylene chloride solvent on a warm sand bath under a gentle stream of nitrogen gas in the **hood.** Recrystallize the crude anilide from an ethanol–water mixture using the Craig tube. Dissolve the material in about 0.5 mL of ethanol, add water (dropwise) to the cloud point, cool in an ice bath, and collect the crystals in the usual manner. Dry the purified derivative product on a porous clay plate, and determine its melting point. HOOD

Toluidides

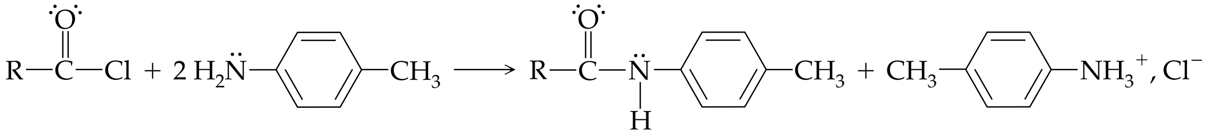

The same procedure described for the preparation of anilides is used, except that *p*-toluidine replaces the aniline.

ALCOHOLS[15]

Phenyl- and α-Naphthylurethanes (Phenyl- and α-Naphthylcarbamates)

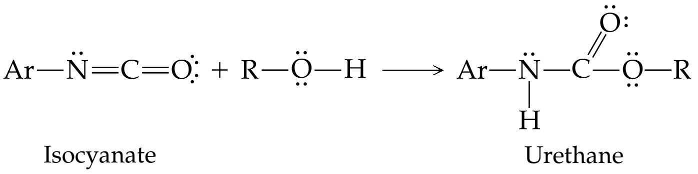

---

[15] See Table 9W.3.

NOTE. For the preparation of these derivatives, the alcohols must be anhydrous. Water hydrolyzes the isocyanates to produce arylamines that react with the isocyanate reagent to produce high-melting, disubstituted ureas.

In a 3.0-mL conical vial containing a boiling stone and equipped with an air condenser protected by a calcium chloride drying tube place 15 mg of an anhydrous alcohol or phenol. Remove the air condenser from the vial and add 2 drops of phenyl isocyanate or α-naphthyl isocyanate. Replace the air condenser immediately. If the unknown is a phenol, add 1 drop of pyridine in a similar manner.

CAUTION: This addition must be done in the **hood.** The isocyanates are lachrymators! Pyridine has the characteristic strong odor of an amine.

If a spontaneous reaction does not take place, heat the vial at about 80–90 °C using a sand bath, for a period of 5 min. Then cool the reaction mixture in an ice bath. It may be necessary to scratch the sides of the vial to induce crystallization. Collect the solid product by vacuum filtration, using a Hirsch funnel, and purify it by recrystallization from ligroin. For this procedure, place the solid in a 10 × 75-mm test tube and dissolve it in 1.0 mL of warm (60–80 °C) ligroin. If diphenyl (or dinaphthyl) urea is present (formed by reaction of the isocyanate with water), it is insoluble in this solvent. Transfer the warm ligroin solution to a Craig tube using a Pasteur filter pipet. Cool the solution in an ice bath and collect the resulting crystals in the usual manner. After drying the product on a porous clay plate, determine the melting point.

3,5-Dinitrobenzoates

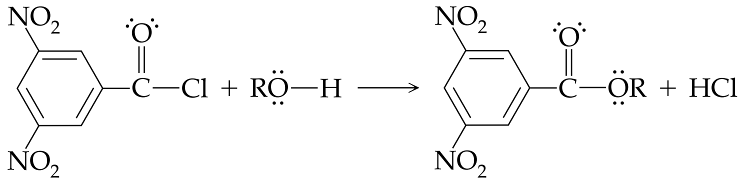

3,5-Dinitrobenzoyl
chloride

NOTE. The dinitrobenzoyl chloride reagent tends to hydrolyze on storage to form the corresponding carboxylic acid. Check its melting point before use (3,5-dinitrobenzoyl chloride, mp = 74 °C; 3,5-dinitrobenzoic acid, mp = 202 °C)

In a 3.0-mL conical vial containing a boiling stone, and equipped with an air condenser protected by a calcium chloride drying tube, place 25 mg of pure 3,5-dinitrobenzoyl chloride and two drops of the unknown alcohol. Heat the mixture to about 10 °C below the boiling point of the alcohol (but not over 100 °C) on a sand bath for a period of 5 min. Cool the reaction mixture, add 0.3 mL of water, and then place the vial in an ice bath to cool. Collect the solid ester by vacuum filtration, using a Hirsch funnel, and wash the filter cake with three 0.5-mL portions of 2% aqueous sodium carbonate (Na_2CO_3) solution, followed by 0.5 mL of water. Recrystallize the solid product from an ethanol–water mixture using a Craig tube. Dissolve the material in about 0.5 mL of ethanol, add water (dropwise) until the solution is just cloudy, cool in an ice bath, and collect the crystals in the usual manner. After drying the product on a porous clay plate, determine the melting point.

ALDEHYDES AND KETONES[16]

2,4-Dinitrophenylhydrazones

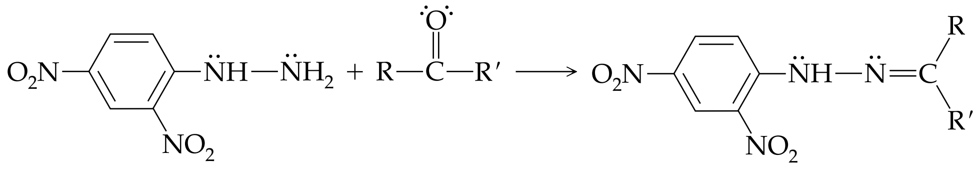

2,4-Dinitrophenylhydrazine A 2,4-dinitrophenylhydrazone

The procedure outlined in the Classification Test Section for aldehydes and ketones (p. 642) is used. Since the derivative to be isolated is a solid, it may be convenient to run the reaction in a 3-mL vial or in a small test tube. Double the amount of the reagents used. If necessary, the derivative can be recrystallized from 95% ethanol.

The procedure is generally suitable for the preparation of phenylhydrazone and *p*-nitrophenylhydrazone derivatives of aldehydes and ketones.

Semicarbazones

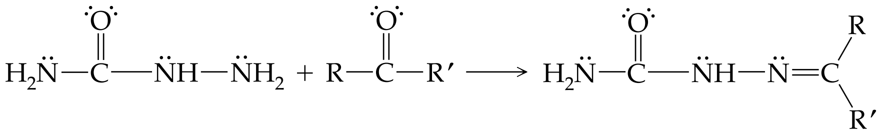

Semicarbazide A semicarbazone

In a 3.0-mL conical vial place 12 mg of semicarbazide hydrochloride, 20 mg of sodium acetate, 10 drops of water, and 12 mg of the unknown carbonyl compound. Cap the vial, shake vigorously, vent, and allow the vial to stand at room temperature until crystallization is complete (varies from a few minutes to several hours). Cool the vial in an ice bath if necessary. Collect the crystals by vacuum filtration, using a Hirsch funnel, and wash the filter cake with 0.2 mL of cold water. Dry the crystals on a porous clay plate. Determine the melting point.

AMINES[17]

Primary and Secondary Amines: Acetamides

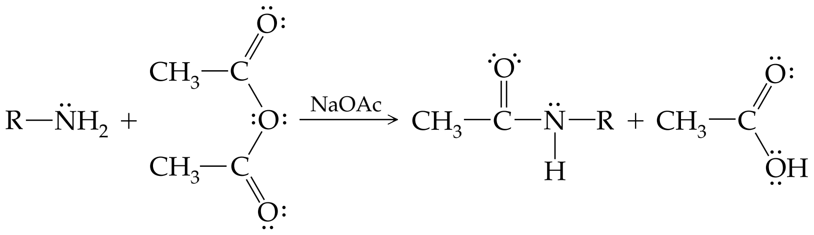

In a 3.0-mL conical vial equipped with an air condenser, place 20 mg of the unknown amine, 5 drops of water, and 1 drop of concentrated hydrochloric acid.

---

[16] See Tables 9W.4 and 9W.5.

[17] See Tables 9W.6 and 9W.7.

In a small test tube, prepare a solution of 40 mg of sodium acetate trihydrate dissolved in 5 drops of water. Stopper the solution and set it aside for use in the next step.

Warm the solution of amine hydrochloride to about 50 °C on a sand bath. Then cool it, and add 40 μL of acetic anhydride in one portion (in the **hood**) through the condenser by aid of a 9-in. Pasteur pipet. In like manner, *immediately* add the sodium acetate solution (prepared previously). Swirl the contents of the vial to ensure complete mixing.

HOOD

Allow the reaction mixture to stand at room temperature for about 5 min, and then place it in an ice bath for an additional 5–10 min. Collect the white crystals by vacuum filtration, using a Hirsch funnel, and wash the filter cake with two 0.1-mL portions of water. The product may be recrystallized from ethanol–water using the Craig tube, if desired. Dry the crystals on a porous clay plate and determine the melting point.

Primary and Secondary Amines: Benzamides

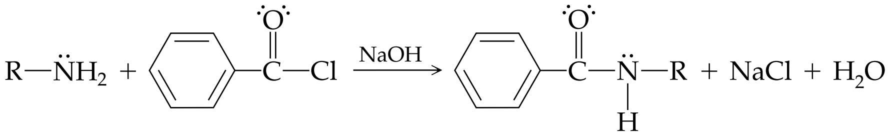

In a 3.0-mL conical vial in the **hood** place 0.4 mL of 10% aqueous NaOH solution, 25 mg of the amine, and 2–3 drops of benzoyl chloride. Cap and shake the vial over a period of about 10 min. Vent the vial periodically to release any pressure buildup.

HOOD

Collect the crystalline precipitate by vacuum filtration, using a Hirsch funnel, and wash the filter cake with 0.1 mL of dilute HCl followed by 0.1 mL of water. It is generally necessary to recrystallize the material from methanol or aqueous ethanol using the Craig tube. Dry the product on a porous clay plate and determine the melting point.

Primary, Secondary, and Tertiary Amines: Picrates

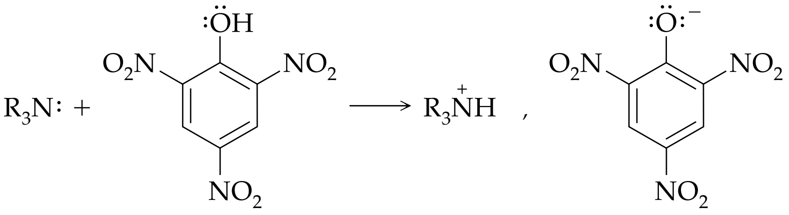

Picric acid

Picrate salt

In a 3.0-mL conical vial containing a boiling stone and equipped with an air condenser, place 15 mg of the unknown amine and 0.3 mL of 95% ethanol.

NOTE. If the amine is not soluble in the ethanol, shake the mixture to obtain a saturated solution and then transfer this solution, using a Pasteur filter pipet, to another vial.

Now add 0.3 mL of a saturated solution of picric acid in 95% ethanol.

CAUTION: Picric acid explodes by percussion or when rapidly heated.

CAUTION

Heat the mixture at reflux, using a sand bath, for about 1 min and then allow it to cool slowly to room temperature. Collect the yellow crystals of the picrate by vacuum filtration, using a Hirsch funnel. Dry the material on a porous clay plate and determine the melting point.

ACID CHLORIDES AND ANHYDRIDES[18]

Amides

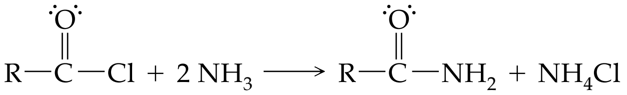

$$R-\overset{\overset{\ddot{O}}{\|}}{C}-Cl + 2\,NH_3 \longrightarrow R-\overset{\overset{\ddot{O}}{\|}}{C}-NH_2 + NH_4Cl$$

HOOD

In a 10 × 75-mm test tube, place 0.4 mL of ice cold, concentrated ammonium hydroxide solution. To this solution, in the **hood,** slowly add, with shaking, about 15 mg of the unknown acid chloride or anhydride. Stopper the test tube and allow the reaction mixture to stand at room temperature for about 5 min. Collect the crystals by vacuum filtration, using a Hirsch funnel, and wash the filter cake with 0.2 mL of ice-cold water. Recrystallize the material, using a Craig tube, from water or an ethanol–water mixture. Dry the purified crystals on a porous clay plate and determine the melting point.

AROMATIC HYDROCARBONS[19]

Picrates

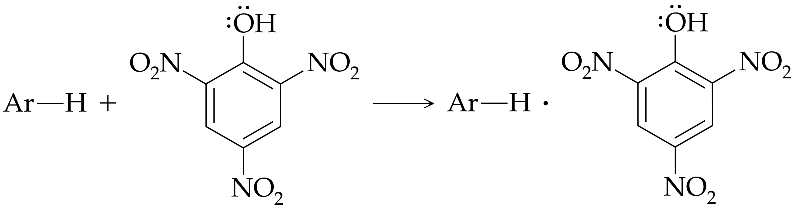

Picric acid

Picrate complex

The procedure outlined on page 659 is used to prepare these derivatives.

NITRILES[20]

Hydrolysis to Amides

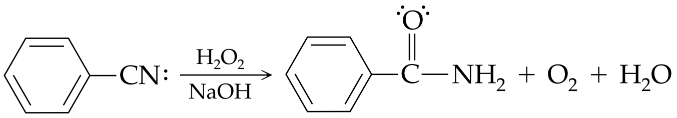

Conversion of nitriles to water-insoluble amides, by hydrolysis with alkaline hydrogen peroxide, is a possible method of characterization for these compounds. It is especially useful for aromatic nitriles.

---

[18] See Table 9W.8.

[19] See Table 9W.9.

[20] See Table 9W.13.

In a 5-mL conical vial containing a magnetic spin vane, weigh and place about 50 mg of the nitrile and 500 μL of a 1 M NaOH solution. Cool the mixture in a water bath and, with stirring, add dropwise 500 μL of 12% H_2O_2 solution. Attach the vial to an air condenser and warm the solution on a sand bath while stirring at 50–60 °C for approximately 45 min. To the cooled reaction mixture add 1–2 mL of cold water, and then collect the solid amide by vacuum filtration. Wash the product with two 1-mL portions of cold water, and recrystallize the amide from aqueous ethanol using the Craig tube. Dry the solid and determine the melting point.

PHENOLS[21]

α-Naphthylurethanes (α-Naphthylcarbamates)

The procedure outlined under Alcohols: Phenyl-, and α-Naphthylurethanes is used to prepare these derivatives (p. 656).

Bromo Derivatives

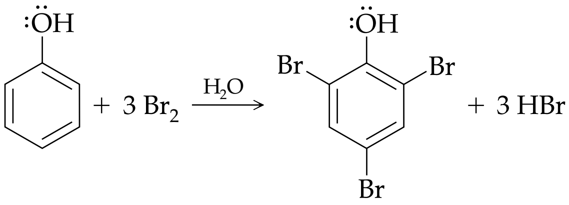

INSTRUCTOR PREPARATION. The brominating reagent is prepared by adding 1.0 mL (3 g) of bromine in the ***hood*** *to a solution of 4.0 g of KBr in 25 mL of water.* HOOD

In a 1.0-mL conical vial, place 10 mg of the unknown phenol followed by 2 drops of methanol and 2 drops of water. To this solution, in the **hood,** add 3 drops of brominating agent from a Pasteur pipet. HOOD

Continue the addition (dropwise) until the reddish-brown color of bromine persists. Now add water (4 drops), cap the vial, shake, vent, and then allow it to stand at room temperature for 10 min. Collect the crystalline precipitate by vacuum filtration using a Hirsch funnel and wash the filter cake with 0.5 mL of 5% aqueous sodium bisulfite solution. Recrystallize the solid derivative from ethanol, or from an ethanol–water mixture, using a Craig tube. Dissolve the material in about 0.5 mL of ethanol, add water until it becomes cloudy, cool in an ice bath, and collect the crystals in the usual manner. Dry the purified product on a porous clay plate and determine the melting point.

---

[21] See Table 9W.10.

www

ALIPHATIC HYDROCARBONS, HALOGENATED HYDROCARBONS, AMIDES, NITRO COMPOUNDS, ETHERS, AND ESTERS[22]

These compounds do not give derivatives directly, but are usually converted into another material that can then be derivatized. The procedures are, for the most part, lengthy, and frequently give mixtures of products. It is recommended that compounds belonging to these classes be primarily identified using spectroscopic methods. Measurement of their physical properties is also of utmost importance.

QUESTIONS

9-1. The following six substances have approximately the same boiling point and are all colorless liquids. Suppose you were given six unlabeled bottles, each of which contained one of these compounds.

Explain how you would use simple chemical tests to determine which bottle contained which compound.

Ethanoic acid
Propyl butanoate
1-Butanol
Toluene
Diisobutylamine
Styrene

9-2. A colorless liquid (C_4H_6O) with a boiling point of 81 °C was found to be soluble in water and also in ether. It gave a negative test for the presence of halogens, sulfur, and nitrogen. It did, however, give a positive test with the Baeyer reagent and also gave a positive test with the 2,4-dinitrophenylhydrazine reagent. It gave negative results when treated with ceric nitrate solution and with Tollens reagent. Treatment with ozone followed by hydrolysis in the presence of zinc gave formaldehyde as one of the products.

What is the structure and name of the colorless liquid?

9-3. A colorless liquid, compound A (C_3H_6O), was soluble in water and ether, and had a boiling point of 94–96 °C. It decolorized a Br_2–CH_2Cl_2 solution and gave a positive ceric nitrate test. On catalytic hydrogenation it formed compound B (C_3H_8O), which did not decolorize the above bromine solution, but did give a positive ceric nitrate test. Treatment of compound A with ozone, followed by hydrolysis in the presence of zinc, gave formaldehyde as one of the products. Compound A formed an α-napthylurethane with a melting point of 109 °C.

What are the names and structures of compounds A and B?

9-4. A compound of formula $C_{14}H_{12}$ gave a positive Baeyer test and burned with a yellow, sooty flame. Treatment with ozone followed by hydrolysis in the presence of zinc gave formaldehyde as one of the products. Also isolated from the ozonolysis reaction was a second compound, $C_{13}H_{10}O$, which burned with a yellow, sooty flame, and readily formed a semicarbazone with a melting point of 164 °C. The ^{1}H NMR spectrum of this compound ($C_{13}H_{10}O$) showed only complex multiplets that were near 7.5 ppm; the fully ^{1}H-decoupled ^{13}C NMR spectrum showed only 5 peaks.

What are the structures and names of the two compounds?

9-5. Compound A ($C_7H_{14}O$) burned with a yellow, nonsooty flame and did not decolorize a bromine–methylene chloride solution. It did give a positive 2,4-dinitrophenylhydrazine test, but a negative Tollens test. Treatment of the compound with lithium aluminum hydride followed by neutralization with acid, produced compound B, which gave a positive Lucas test in about 5 min. Compound B also gave a positive ceric nitrate test. The ^{1}H NMR spectrum for compound A gave the following data:

1.02 ppm	9H, singlet
2.11 ppm	3H, singlet
2.31 ppm	2H, singlet

Give suitable structures for compounds A and B.

---

www → [22] See Tables 9W.11, 9W.12, and 9W.14–9W.17.

9-6. A friend of yours, who is a graduate student attempting to establish the structure of a chemical species from field clover, isolated an alcohol that was found to have an optical rotation of +49.5°. Chemical analysis gave a molecular formula of $C_5H_{10}O$. It was also observed that this alcohol readily decolorized Br_2–CH_2Cl_2 solution. On this basis, the alcohol was subjected to catalytic hydrogenation and it was found to absorb 1 mol equivalent of hydrogen gas. The product of the reduction gave a positive ceric nitrate test, indicating that it, too, was an alcohol. However, the reduced compound was optically inactive.

Your friend has come to you for assistance in determining the structures of the two alcohols. What do you believe the structures are?

9-7. An unknown compound burned with a yellow, nonsmoky flame and was found to be insoluble in 5% sodium hydroxide solution but soluble in concentrated sulfuric acid. Measurement of its boiling point gave a range of 130–131 °C. Combustion analysis gave a molecular formula of C_5H_8O. It was found to give a semicarbazone with a melting point of 204–206 °C. However, it gave a negative result when treated with Tollens reagent and it did not decolorize the Baeyer reagent. It also gave a negative iodoform test.

Identify the unknown compound.

9-8. An unknown organic carboxylic acid, mp = 139–141 °C, burned with a yellow, sooty flame. The sodium fusion test showed that nitrogen was present. It did not react with *p*-toluenesulfonyl chloride, but did give a positive test when treated with 5% aqueous ferrous ammonium sulfate solution, acidified with 3 M H_2SO_4, and then followed by methanolic KOH solution. A 200-mg sample of the acid neutralized 12.4 mL of 0.098 M sodium hydroxide solution.

Identify the acid.

Does your structure agree with the calculated equivalent weight?

9-9. An unknown organic liquid, compound A, was found to burn with a yellow, sooty flame and give a positive Lucas test (~5 min). Upon treatment with sodium dichromate–sulfuric acid solution it produced compound B, which also burned with a yellow, sooty flame. Compound B gave a positive 2,4-dinitrophenylhydrazine test, but a negative result when treated with the Tollens reagent. However, compound B did give a positive iodoform test.

The 1H NMR spectrum for compound A showed the following:

1.4 ppm	3H (doublet)	4.8 ppm	1H (quartet)
1.9 ppm	1H (singlet)	7.2 ppm	5H (complex multiplet)

Give the structures and suitable names for compounds A and B.

9-10. A hydrocarbon, compound A (C_6H_{10}), burned with a yellow, almost nonsmoky flame. On catalytic hydrogenation over platinum catalyst it absorbed 1 mol of hydrogen to form compound B. It also decolorized a Br_2–CH_2Cl_2 solution to yield a dibromo derivative, compound C. Ozonolysis of the hydrocarbon gave only one compound, D. Compound D gave a positive iodoform test when treated with iodine–sodium hydroxide solution. On treatment of compound D with an alcoholic solution of silver ammonium hydroxide, a silver mirror was formed within a few minutes.

Identify the hydrocarbon A and compounds B–D.

9-11. A high-boiling liquid, bp = 202–204 °C burns with a yellow, sooty flame. Sodium fusion indicates that halogens, nitrogen, and sulfur are not present. The compound is not soluble in water, dilute sodium bicarbonate solution, or dilute hydrochloric acid. However, it proved to be soluble in 5% aqueous sodium hydroxide solution. The compound gives a purple color with ferric chloride solution and a precipitate when reacted with bromine–water. Treatment with hydroxylamine reagent did not give a reaction, but a white precipitate was obtained when the compound was treated with α-napthylisocyanate. On drying, this white, solid derivative had an mp = 127–129 °C. Identify the original liquid and write a structure for the solid derivative.

After identifying the unknown liquid, can you indicate what the structure of the precipitate obtained on reaction with bromine might be?

9-12. A colorless liquid, bp = 199–201 °C, burns with a yellow, sooty flame. The sodium fusion test proved negative for the presence of halogens, nitrogen, and sulfur. The compound was not soluble in water, 5% aqueous sodium hydroxide, or 5% hydrochloric acid. However, it dissolved in sulfuric acid with evolution of heat. It did not give a precipitate with 2,4-dinitrophenylhydrazine solution, and it did not decolorize bromine–methylene chloride solution. The unknown liquid did give a positive hydroxamate test and was found to have a saponification equivalent of 136.

Identify the unknown liquid.

9-13. Your friend of Question 9-6 still needs your help. A week later a low-melting solid, compound A, was isolated, which combustion analysis showed had composition $C_9H_{10}O$. The substance gave a precipitate when treated with 2,4-dinitrophenylhydrazine solution. Furthermore, when reacted with iodoform reagent, a yellow precipitate of CHI_3 was observed. Acidification of the alkaline solution from the iodoform test produced a solid material, compound B.

Reduction of compound A with $LiAlH_4$ gave compound C ($C_9H_{12}O$). Compound C also gave compound B when treated with iodoform reagent. Vigorous oxidation of compound A, B, or C with sodium dichromate–sulfuric acid solution gave an acid having an mp = 121–122 °C.

Your friend needs your assistance in determining the structures for compounds A, B, and C. Can you identify the three compounds?

9-14. An organic compound ($C_9H_{10}O$) showed strong absorption in the IR spectrum at 1735 cm^{-1} and gave a semicarbazone having a melting point of 198 °C. It burned with a yellow, sooty flame and also gave a positive iodoform test. The ^1H NMR spectrum of the compound provided the following information:

2.11 ppm	3H (singlet)
3.65 ppm	2H (singlet)
7.20 ppm	5H (complex multiplet)

Identify the unknown organic compound.

9-15. An unknown compound (A) was soluble in ether but only slightly soluble in water. It burned with a clear blue flame and combustion analysis showed it to have the molecular formula of $C_5H_{12}O$. It gave a positive test with the Jones reagent producing a new compound (B) with a formula of $C_5H_{10}O$. Compound B gave a positive iodoform test and formed a semicarbazone. Compound A on treatment with sulfuric acid produced a hydrocarbon (C) of formula C_5H_{10}. Hydrocarbon C readily decolorized a Br_2–CH_2Cl_2 solution, and on ozonolysis, produced acetone as one of the products.

Identify the structure of each of the lettered compounds.

9-16. Compound A (C_7H_{14}) decolorized a Br_2–CH_2Cl_2 chloride solution. It reacted with $18H_3 \cdot$THF reagent, followed by alkaline peroxide solution, to produce compound B. Compound B, on treatment with chromic acid–sulfuric acid solution, gave carboxylic acid C, which could be separated into two enantiomers. Compound A, on treatment with ozone, followed by addition of hydrogen peroxide, produced compound D. Compound D was identical to that material isolated from the oxidation of 3-hexanol with chromic acid–sulfuric acid reagent.

Identify the structures of compounds A, B, C, and D.

9-17. Compound A (C_8H_{16}) decolorized a bromine–methylene chloride solution. Ozonolysis produced two compounds, B and C, which could be separated easily by gas chromatography. Both B and C gave a positive 2,4 dinitrophenylhydrazine test. Carbon–hydrogen analysis and molecular weight determination of B gave a molecular formula of $C_5H_{10}O$. The ^1H NMR spectrum revealed the following information for B:

0.92 ppm	3H, triplet	2.17 ppm	3H, singlet
1.6 ppm	2H, pentet	2.45 ppm	2H, triplet

Compound C was a low-boiling liquid (bp 56°C) The ^1H NMR of this material showed only one singlet.

Identify compounds A, B, and C.

C$_{10}$H$_{10}$apter 10W

ADVANCED MICROSCALE ORGANIC LABORATORY EXPERIMENTS

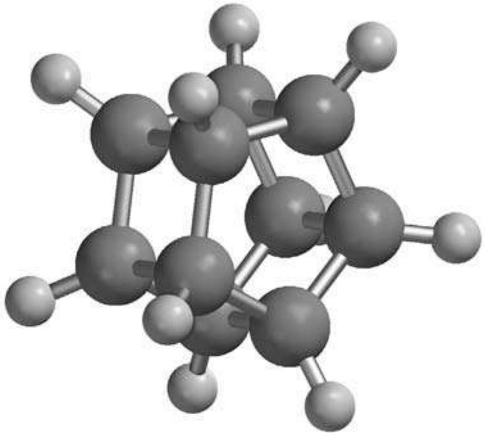

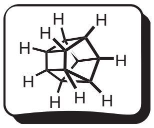

From a theoretical perspective, the chemistry described in this chapter is more demanding of the student–investigator than that described in Chapter 6.

1. The organic reactions performed are less familiar and are not as likely to be found in most introductory texts.
2. The mechanisms proposed for these systems are more involved and not generally developed at the introductory level.
3. Many of the reagents used are rarely used in the introductory organic laboratory.

Thus, the experiments contained in Chapter 10W are specifically tailored to challenge the more advanced undergraduate students, those who are already able to access the chemical literature. This chapter can also offer a special laboratory experience for those few beginning students who are particularly interested in the subject and wish to spend extra preparation time. The techniques involved are not, in most instances, any different from those used in the introductory microscale laboratory reactions described in Chapter 6, and therefore the manipulations involved should not be considered a barrier to undertaking any of the advanced experiments. The reaction conditions, however, are less forgiving to slight deviations from the suggested ones and the ultimate success of the transformations is less secure.

Thus, the experimentation contained in Chapter 10W can be adapted to several levels of undergraduate laboratory programs. For example, the study of these reactions can potentially make significant contributions to advanced undergraduate programs where microscale techniques are being introduced to research-oriented students for the first time.

The formats for the discussions and experimental procedures are similar to those used in Chapter 6. The reactions selected for study in this chapter include

- An unusual borane reduction of a carbonyl directly to a methylene group.
- The trapping of an α,β-unsaturated ketone as its enol acetate, by acylium ion formation with chlorotrimethylsilane and acetic anhydride.

Chapter 10W: C$_{10}$H$_{10}$, Housane
Eaton, Or, Branca and Shankar (1968).

- The synthesis of a heterocyclic ring using diethyl carbonate and sodium hydride. The discovery of the medicinal properties of this class of heterocycles led directly to modern anticoagulation therapy.
- The synthesis of an isotopically labeled molecule through the use of a unique Grignard cross-coupling reaction in the presence of dichloro[1,3-bis(diphenylphosphino)propane]nickel(II).
- An oxidative coupling of a naphthol by ferric chloride in a reaction that mimics nature's method of coupling phenolic substances into important pigments, such as hypericin (**I**):

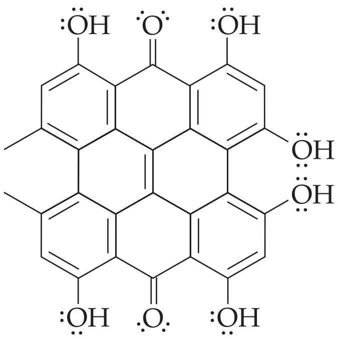

I

- An important molecular rearrangement of oximes to amides discovered by Ernst Otto Beckmann in 1886 and so named in his honor. The modern version uses one of the most powerful acid reagents used in organic chemistry, triflic acid.

The study of the reactions outlined in Chapter 10W should help to facilitate the student's smooth transition into the organic research laboratory. We hope you find the collection as exciting as we did.

EXPERIMENT [1adv]

Diborane Reductions: Thioxanthene and Xanthene

Common name: thioxanthene
CA number: [261-31-4]
CA name as indexed: 9*H*-thioxanthene
Common name: xanthene
CA number: [92-83-1]
CA name as indexed: 9*H*-xanthene

Purpose. This experiment investigates an unusual example of a hydroboration in which a carbonyl group is directly and fully reduced to a methylene group by borane (BH_3). This example is an atypical reduction of an aldehyde or ketone, since the use of this reagent usually leads to the formation of the corresponding alcohol. You will explore the mechanistic rationale for this unexpected product.

Prior Reading

Standard Experimental Apparatus: Moisture Protected Reaction Apparatus (pp. 23–24)

Technique 6: Chromatography
Packing the Column (p. 93)
Elution of the Column (p. 94)

Technique 6B: Concentration of Solutions
Removal of Solvent Under Reduced Pressure (pp. 102-104)

NOTE. See Experiment [13] for a biography of Herbert C. Brown, Nobel Laureate, who discovered and developed the boron hydride reagents. This experiment also contains further details about the use of these powerful reagents, which have revolutionized reduction reactions in organic chemistry.

DISCUSSION

Borane is a useful and selective reducing agent. It is prepared by the reaction of boron trifluoride etherate with sodium borohydride. The borane produced, as the etherate, may be distilled as the dimer, which is a colorless, toxic gas (B_2H_6). Collection of the dimer distillate in tetrahydrofuran (THF) again forms the monomer, in this case as the $BH_3 \cdot$ THF complex. The latter is commercially available as a 1.0 M solution.

$$3\ NaBH_4 + 4\ BF_3 \cdot O(CH_2CH_3)_2 \longrightarrow 3\ NaBF_4 + 2\ B_2H_6\uparrow(\text{gas})$$

Borane complexes can also be formed with other ethers, such as diethyl ether (as just discussed) or diglyme (diethylene glycol dimethyl ether). These complexes form readily because the ether, acting as a Lewis base (electron donor), can satisfy the electron-deficient boron atom, which acts as a Lewis acid (electron acceptor). Borane reacts rapidly with water, and therefore procedures using the $BH_3 \cdot$THF complex must be conducted under anhydrous conditions.

Borane is a Lewis acid that is attacked by electron-rich centers. Thus, when aldehydes or ketones are treated with the $BH_3 \cdot$ THF complex, the borate ester ($H_2B—OR$) is rapidly formed, which, upon hydrolysis, gives the corresponding alcohol. The reduction of the carbonyl group is believed to take place by addition of the oxygen atom to the electron-deficient boron atom, followed by irreversible transfer of hydride ion from the now anionic boron to the carbon atom of the (former) carbonyl:

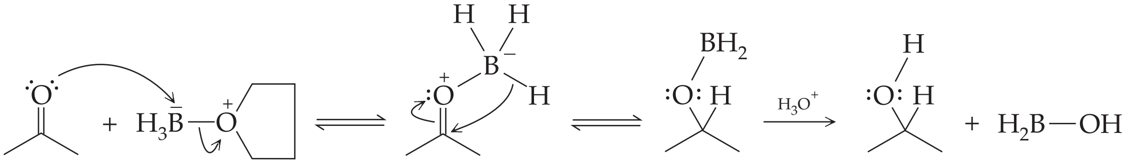

In the case of the xanthone and thioxanthone ring systems, the corresponding alcohol is not formed. The intermediate borate ester undergoes an elimination reaction, forming a borate anion and a resonance-stabilized carbocation. The second stage of the reaction is initiated by displacement of THF from a second $BH_3 \cdot THF$ complex by a lone pair from either the oxygen atom of the xanthene carbocation or the sulfur atom of the thioxanthene carbocation. This *new complex* then undergoes an internal hydride (:H^-) transfer from boron to the C-9 ring position (this is the second hydride attack on this position in the overall reaction) to form a stable oxonium (or sulfonium) ion intermediate. On aqueous–alcohol workup of the reaction mixture the oxonium (sulfonium) product is quickly hydrolyzed to yield xanthene (or thioxanthene) possessing a fully reduced methylene group, —CH_2—, at the 9 position.

Conventional methods for the reduction of a carbonyl to a methylene group that do not require the conjugative assistance of a heteroatom are the well-known Clemmensen (Zn(Hg), HCl), and Wolff–Kishner (H_2NNH_2/KOH) reductions, and the desulfurization of the corresponding thioacetal with Raney nickel.

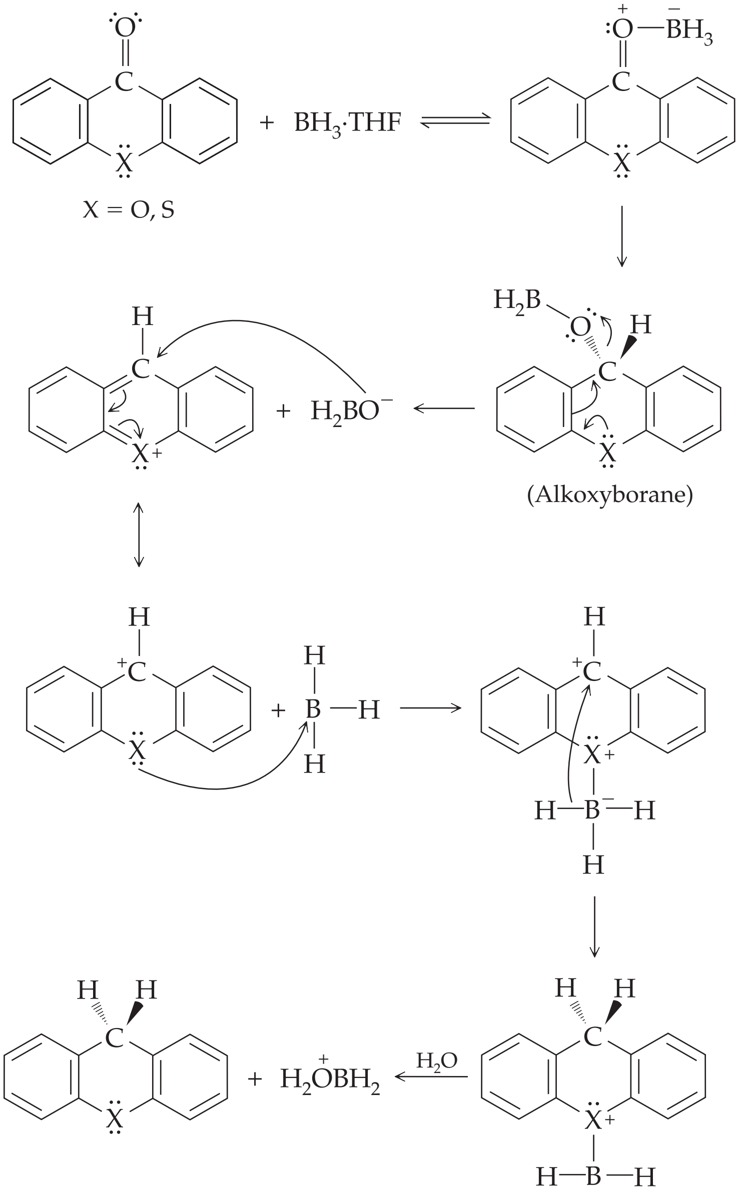

Thioxanthene

Experiment [$1A_{adv}$]

REACTION

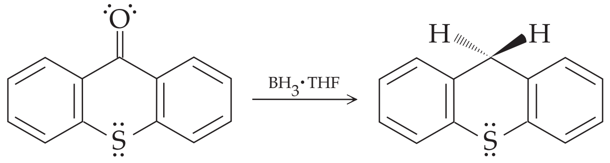

Thioxanthone Thioxanthene

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 1.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	d	n_D
Thioxanthone	212.28	50 mg	0.24	209			
Tetrahydrofuran	72.12	1.7 mL			67	0.889	1.407
Borane · THF, 1 M		1.0 mL	1.0				

Reagents and Equipment. Attach a 5.0-mL conical vial containing a magnetic spin vane to a Claisen head. Then fit the Claisen head with a nitrogen inlet tube (prepared from a syringe) and a rubber septum. Weigh and add 50 mg (0.24 mmol) of thioxanthone to the conical vial.

Flush the reaction vial with a gentle stream of nitrogen gas for several minutes, add 1.7 mL of *dry* (see Note) tetrahydrofuran (THF) through the septum (syringe), and then place a small balloon over the Claisen-head outlet so as to maintain a dry atmosphere in the system (➡).

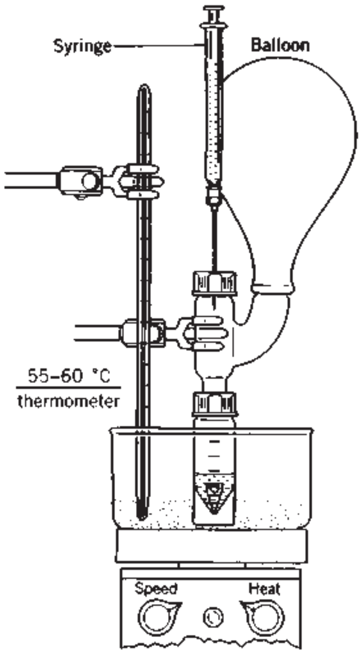

INSTRUCTOR'S NOTE. The THF must be ***absolutely*** *dry. It is recommended that HPLC grade reagent be used. If you do not have a fresh bottle, distill it once from calcium hydride (or sodium benzophenone ketyl) and store it over molecular sieves. It may be stored and used safely for up to a week without adversely affecting the yield of product.*

Heat the mixture with stirring in a sand bath at 55–60 °C until the thioxanthone dissolves, yielding a yellow solution. Then, with continued stirring, add 1.0 mL of a 1.0 M solution of $BH_3 \cdot THF$ in one portion through the rubber septum with a 1.0-mL syringe.

CAUTION: The $BH_3 \cdot THF$ solution reacts *violently* with water.

Reaction Conditions. Heat the solution with stirring in a sand bath at 55–60 °C for 5 min. The solution should become colorless during this time.

Isolation of Product. Quench the reaction by *carefully* adding dropwise approximately 10 drops of 95% ethanol (Pasteur pipet), with stirring, or until the observed foaming subsides. *The aqueous alcohol is added to decompose any unreacted* $BH_3 \cdot$ *THF reagent and to hydrolyze the sulfonium ion intermediate.*

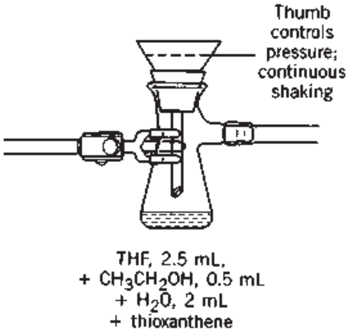

After the solution has cooled to room temperature, transfer the mixture by Pasteur pipet to a 25-mL filter flask containing a boiling stone. Carefully remove roughly one-half of the tetrahydrofuran solvent under reduced pressure with continuous swirling of the flask (see Prior Reading) (⬅). Then, use a calibrated Pasteur pipet to add two 1.0-mL portions of water to the reaction mixture. Carefully remove the remaining tetrahydrofuran and ethanol solvent under reduced pressure with continuous swirling of the flask. As the tetrahydrofuran and ethanol evaporate, white crystals of thioxanthene appear, and a slurry of these crystals in water will remain in the flask after the tetrahydrofuran and ethanol are removed. Collect the product crystals under reduced pressure by use of a Hirsch funnel, and wash the filter cake on the Hirsch filter bed with two 1.0-mL portions of water. Dry the crystals in air on a porous clay plate or on filter paper.

Purification and Characterization. The thioxanthene product is essentially pure as isolated. It may be recrystallized from an ethanol–chloroform mixture, if necessary. Weigh the product and calculate the percent yield. Determine the melting point of the material and compare it with the value reported in the literature. Obtain IR spectra of thioxanthone and thioxanthene and compare them to each other as well as to those given in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Experiment [1B$_{adv}$]

Xanthene

REACTION

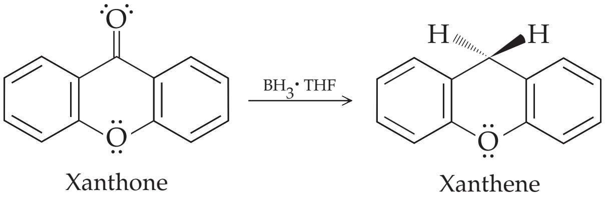

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 3.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	d	n_D
Xanthone	196.22	50 mg	0.26	174			
Tetrahydrofuran	72.12	0.7 mL			67	0.889	1.407
Borane · THF, 1 M		0.75 mL	0.75				

Reagents and Equipment. Using the experimental apparatus described in Experiment [$1A_{adv}$], weigh and place in the reaction flask 50 mg (0.26 mmol) of xanthone followed by 0.7 mL of dry THF (see Experiment [$1A_{adv}$], Reagents and Equipment). Maintain a dry nitrogen atmosphere in the system by the same procedure as used in Experiment [$1A_{adv}$]. Heat the reaction mixture, with stirring, to 55–60 °C using a sand bath. After the xanthone dissolves, use a 1.0-mL syringe to add, in one portion, 0.75 mL of a 1.0 M solution of $BH_3 \cdot THF$ through the rubber septum on the screw-cap port of the Claisen head.

CAUTION: The $BH_3 \cdot THF$ solution reacts *violently* with water.

Reaction Conditions. Stir the reaction solution in a sand bath at 55–60 °C for a period of 1 h.

Isolation of Product. While stirring the warm reaction mixture, quench the reaction by carefully adding (with a Pasteur pipet) 95% ethanol—approximately 10 drops of or until the observed foaming subsides. *The alcohol is added to decompose any unreacted $BH_3 \cdot THF$ reagent and to hydrolyze the oxonium ion intermediate.*

Transfer the solution by Pasteur pipet to a 25-mL filter flask containing a boiling stone. Remove roughly one-half of the tetrahydrofuran solvent under reduced pressure with continuous swirling of the flask (➡). Now add two 1.0-mL portions of water (calibrated Pasteur pipet) to the solution. Carefully remove the **remaining** tetrahydrofuran–ethanol solvent under reduced pressure with continuous swirling of the flask. As the tetrahydrofuran and ethanol are removed and the solution becomes more concentrated, white crystals of xanthene appear as a slurry in the remaining water. Collect the crystals under reduced pressure using a Hirsch funnel.

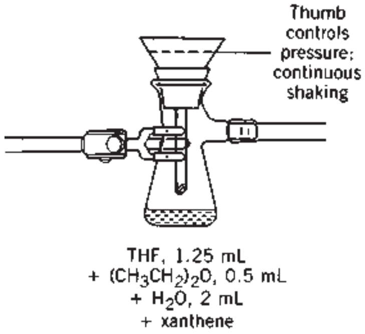
Thumb controls pressure; continuous shaking
THF, 1.25 mL
+ $(CH_3CH_2)_2O$, 0.5 mL
+ H_2O, 2 mL
+ xanthene

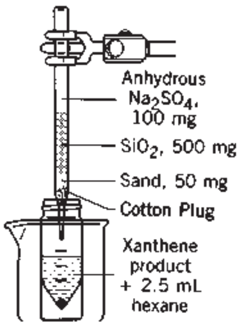
Anhydrous Na_2SO_4, 100 mg
SiO_2, 500 mg
Sand, 50 mg
Cotton Plug
Xanthene product + 2.5 mL hexane

HOOD

Purification and Characterization. Purify the crude xanthene by column chromatography. Place 0.5 g of activated silica gel followed by 0.1 g of anhydrous sodium sulfate in a Pasteur filter pipet (➡).

Wet the column with a small amount of hexane, and then place a solution of the crude xanthene, dissolved in 0.25 mL of methylene chloride, on the column using a Pasteur filter pipet. Elute the xanthene by adding additional hexane (~2.5 mL). Collect the eluate in a tared 5-mL conical vial containing a boiling stone. Remove the hexane solvent by evaporation in the **hood** while warming in a sand bath to yield pure xanthene. This compound may be recrystallized from ethanol if further purification is found to be necessary after characterization of the product.

After air-drying, weigh the solid and calculate the percent yield of xanthene. Determine the melting point and compare it with the value found in the literature.

Obtain an IR spectrum of your xanthene, and compare the spectrum to that shown in Figure 10.1W.

In this experiment, the carbonyl group of an aromatic ketone was reduced to a methylene group. Examine the infrared spectra of the starting material and of the reduction product to see what evidence is present to indicate that the desired reaction has occurred.

Infrared Analysis. We will first consider the spectrum of xanthone (Fig. 10.2W). The macro group frequency associated with six-membered carbocyclic aromatic ring systems applies in this instance (peaks at 3100–3000, 1600, 1585, 1500, and 1450 cm^{-1}). This frequency train involves the bands located at

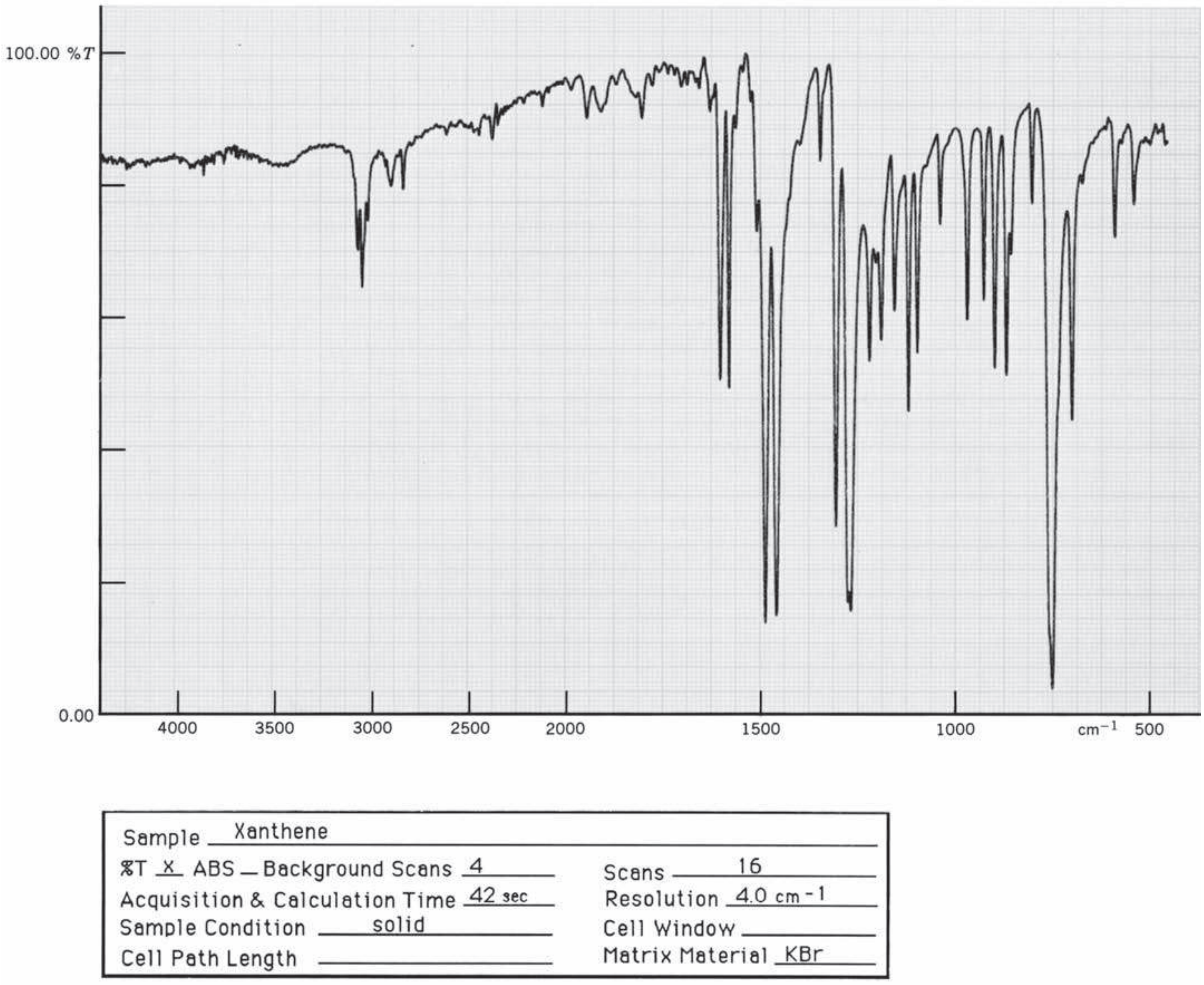

Sample Xanthene
%T X ABS — Background Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm-1
Sample Condition solid
Cell Window
Cell Path Length
Matrix Material KBr

Figure 10.1W IR spectrum: xanthene.

3085–3020, 1610–1570, 1484, and 1460 cm^{-1}. These peaks are assigned as follows:

a. **3085–3020 cm^{-1}:** C—H stretch on sp^2-hybridized carbon. The breadth and complexity of this set of absorption bands indicates the presence of a fairly complex aromatic system.

b. **1610–1570 cm^{-1}:** The peaks observed in this region are related to the two degenerate fundamental stretching motions, ν_{8a} and ν_{8b}, of the simple aromatic ring system. Normally, ν_{8a} is found near 1600 cm^{-1} and is considerably more intense than ν_{8b}, which is located near 1580 cm^{-1}. In xanthone the situation is more complicated, because the central γ-pyrone ring introduces a pseudoaromatic six-membered ring system. Thus, this ring system might be expected to possess somewhat shifted fundamental frequencies for the carbon rings. It is not unexpected, then, that we observe a band system of four major components in this region.

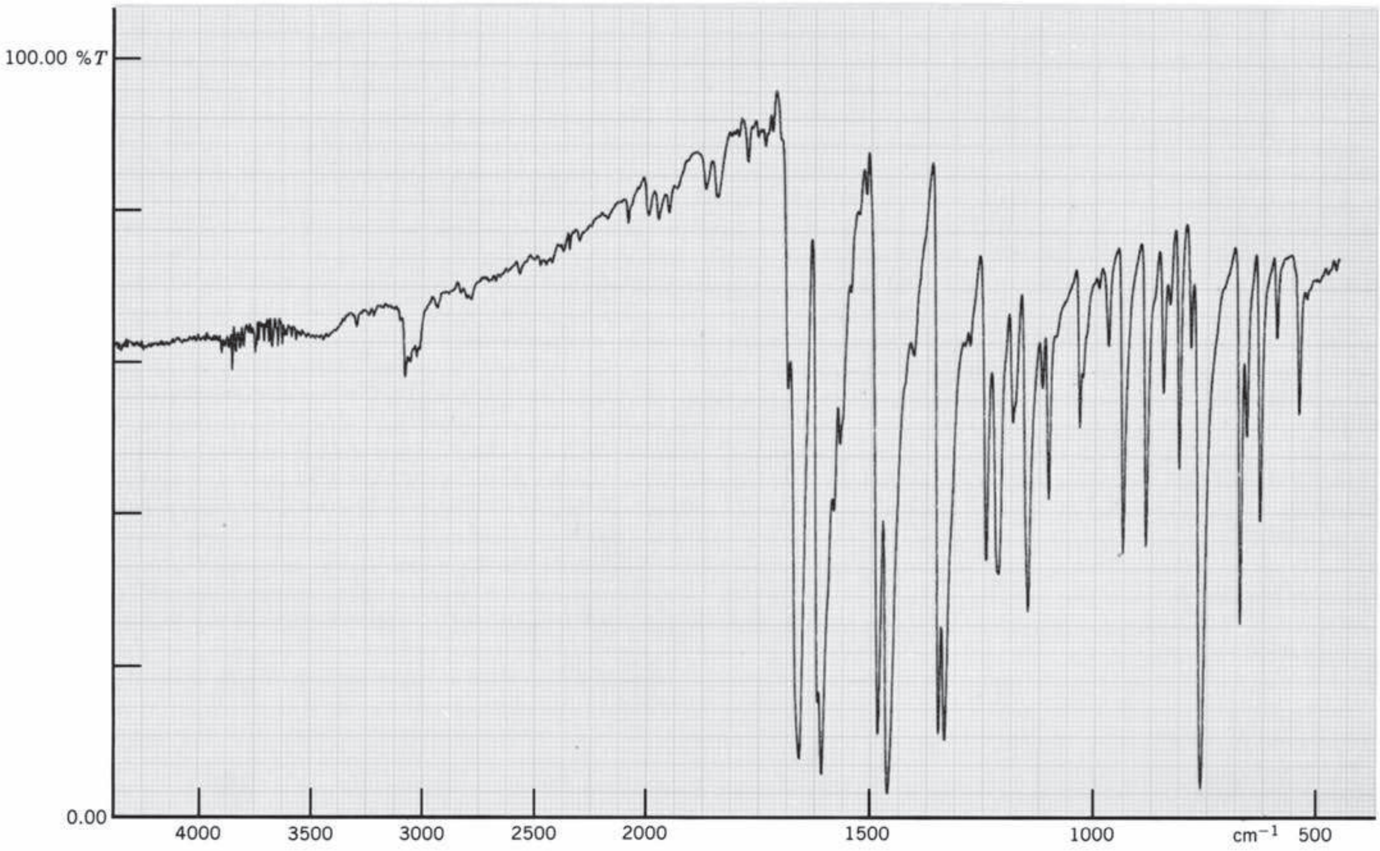

Sample Xanthone
%T X ABS _ Background Scans 4 Scans 16
Acquisition & Calculation Time 42 sec Resolution 4.0 cm - 1
Sample Condition solid Cell Window
Cell Path Length Matrix Material KBr

Figure 10.2W IR spectrum: xanthone.

c. 1484 and 1460 cm^{-1}: Aromatic ring stretch related to ν_{19a} and ν_{19b}. These frequencies are less disturbed by the presence of the pyrone system and occur near their expected locations of 1500 and 1450 cm^{-1}.

The very strong band at 756 cm^{-1}, in the absence of strong absorption near 700 cm^{-1}, is supporting evidence for the presence of four adjacent ring C—H groups, which implies ortho substitution.

The intense band observed at 1656 cm^{-1} is strong evidence for the presence of a highly conjugated carbonyl group, which, of course, is consistent with the structure of the starting material.

The IR spectrum of the product (Fig. 10.1W) supports the complete reduction to the methylene system. The macro group frequency train defined for six-membered carbocyclic aromatic systems still applies. The expected frequencies are very close to the observed values:

a. 3075–3025, 1603, 1589, 1490, and 1457 cm^{-1}.

b. The carbonyl band has vanished, and in its place two weak bands have arisen near 2902 and 2840 cm^{-1}. These latter peaks are assigned to the antisymmetric and symmetric stretching modes of the newly formed methylene group.

c. The most intense band in the spectrum occurs at 750 cm^{-1} (four in a row, C—H all-in-phase, out-of-plane bend) and indicates that the basic substitution pattern has not changed on the ring system during the reaction.

QUESTIONS

10W-1. In the reaction performed in this experiment, assume that the first stage of the reaction is the rate-determining step. Would you then predict that the relative rate of reduction of the carbonyl group in compound A to the methylene group to be faster or slower than that of xanthone under the conditions of this experiment? Explain.

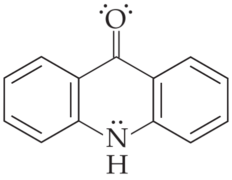

Compound A

10W-2. The reduction of aldehydes or ketones to the methylene group occurs with hydride reagents only when some special feature of the substrate promotes cleavage of the C—OH linkage. Suggest a suitable mechanism by which the reductions given below might occur.

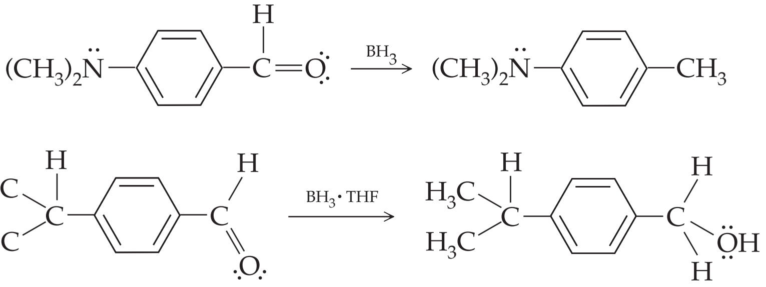

10W-3. Borane also forms complexes with sulfides and amines. Draw a suitable structure to represent the complex formed between BH_3 and dimethyl sulfide, and also that formed between BH_3 and triethylamine.

10W-4. Using your lecture textbook as a reference, find three different methods for the conversion of 4-methylcyclohexanone to methylcyclohexane.

10W-5. The infrared spectrum of the xanthene reduction product contains evidence that demonstrates that conjugation of the rings is still maintained following removal of the carbonyl group. What is this evidence?

BIBLIOGRAPHY

The reactions outlined in Experiments [$1A_{adv}$] and [$1B_{adv}$] were adapted from the work of

Wechter, W. J. *J. Org. Chem.* **1963,** *28,* 2935.

Borane as a reducing agent:

Carruthers, W.; Coldham, I. *Modern Methods of Organic Synthesis,* 4th ed.; Cambridge University Press: New York, 2004, Chap. 7.

Fieser, L. F.; Fieser, M. *Reagents For Organic Synthesis;* Wiley: New York, 1967, Vol. I, p. 199. Subsequent volumes of this series have further examples of diborane as a reducing agent.

Matteson, D. S. in Hartley, F. R.; Patai, S, Eds., *The Chemistry of the Metal-Carbon Bond,* Vol 4, Wiley: New York, 1987, pp. 307–409.

Pelter, A.; Smith, K.; Brown, H. C. *Borane Reagents;* Academic Press: New York, 1988.

Borane as a hydroborating agent:

Brown, H. C. *Hydroboration;* Benjamin: New York, 1962.

Brown, H. C. *Boranes in Organic Chemistry;* Cornell University Press: New York, 1972.

Brown, H. C. *Organic Synthesis Via Boranes;* Wiley: New York, 1975.

See Zaidlewicz, M. for an instructive presentation on *Hydroboration* in *Kirk-Othmer Encyclopedia of Chemical Technology,* 5th ed.; Vol. 13, Wiley-VCH: New York, 2005.

Smith, K. *Organoboron Chemistry in Organometallics in Synthesis—A Manual;* Schlosser, M. M., Ed.; Wiley: New York, 1994.

Zweifel, G.; Brown, H. C. *Org. React.* **1963,** *13,* 1.

Heterocyclic Ring Synthesis: Benzimidazole

EXPERIMENT
[2_{adv}]

Common name: benzimidazole
CA number: [51-17-2]
CA name as indexed: 1*H*-benzimidazole

Purpose. This experiment investigates conditions under which one of the important heterocyclic ring systems, the benzimidazole ring, may be formed. The method used involves the condensation of a 1,2-diaminobenzene with formic acid. The simplest possible benzimidazole ring system, benzimidazole itself, is prepared.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallizations (pp. 89–90)

REACTION

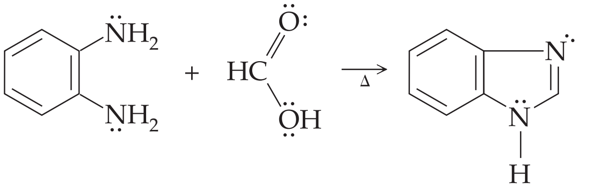

o-Phenylenediamine Formic acid 1-*H*-Benzimidazole

DISCUSSION

This experiment illustrates the classic method of forming the benzimidazole ring system. This heterocycle is generally prepared from 1,2-diaminobenzene (*o*-phenylenediamine) derivatives by reaction with carboxylic acids, or their derivatives, under acidic conditions. The ring system is aromatic; thus it is difficult to oxidize or reduce, and it is stable to both acids and bases. It is an important heterocyclic ring system that occurs in vitamin B_{12} and in many other biologically important compounds. Benzimidazole itself inhibits the growth of certain yeasts and bacteria.

The parent compound of this class of heterocyclic compounds is imidazole. This ring system exhibits basic properties and is protonated to give a conjugate acid with $pK_a = 6.95$. Once the imidazole ring is protonated, the two nitrogen atoms are indistinguishable, because the resonance forms of the protonated species are equivalent. As a resonance stabilized intermediate, the imidazole scaffold offers the synthetic organic chemist a multitude of opportunities when considering its role as part of an ionic liquid. As the term implies, an ionic liquid is a salt. However, when the salt is appropriately modified, it can exist as a liquid at room temperature. As a salt, the system exhibits a significantly lower vapor pressure when compared to standard organic solvents, and thus it can serve the role of a solvent for many synthetic transformations. Over the past decade, ionic liquid technologies have witnessed an unprecedented surge in interest, especially among industrial applications.

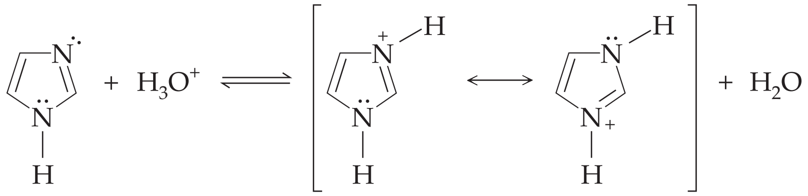

The reaction to form the ring system used in this experiment proceeds in two stages. The first involves the in situ formation of an *N*-substituted formamide, via the usual nucleophilic addition–elimination reaction. The second involves an intramolecular nucleophilic addition to a carbonyl group and subsequent elimination of water to form the unsaturated heterocyclic ring. The sequence is outlined below:

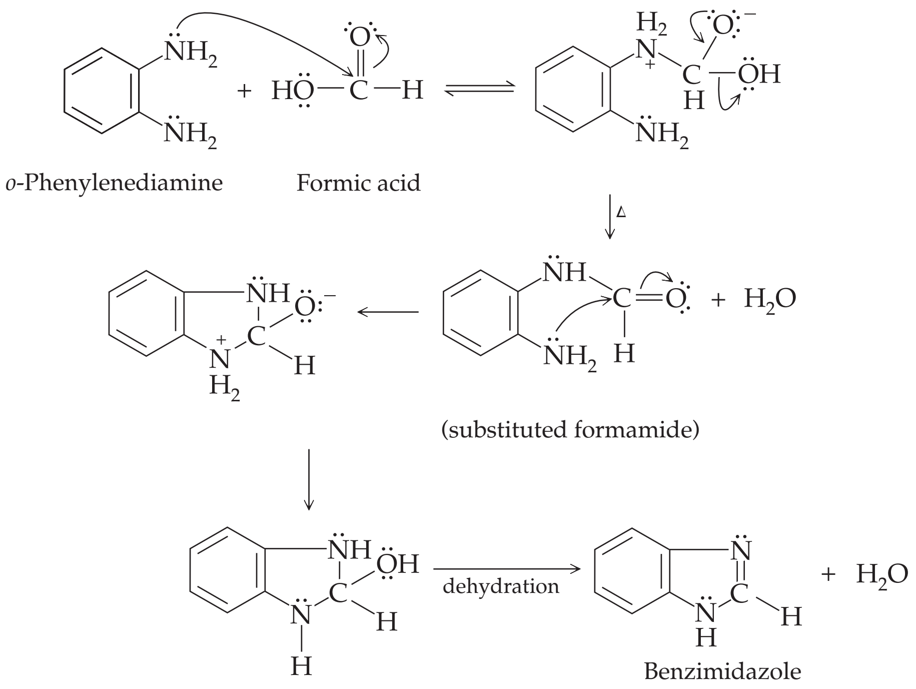

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0 h.

Physical Properties of Reactants							
Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
o-Phenylenediamine	108.1	108 mg	1.0	102			
90% Formic acid	46.03	64 μL	1.7		101	1.22	1.3714

Reagents and Equipment. Weigh and add 108 mg (1.0 mmol) of *o*-phenylenediamine to a 3.0-mL conical vial containing a magnetic spin vane. Now add 64 μL (79 mg, 1.7 mmol) of 90% formic acid, and attach the vial to a reflux condenser protected by a calcium chloride drying tube (➡).

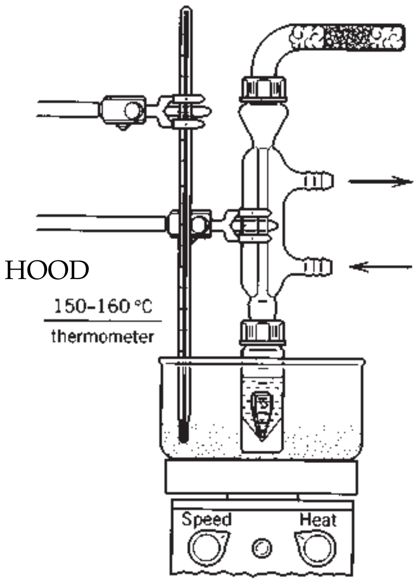

o-$(H_2N)_2C_6H_4$, 108 mg
+ 90% HCO_2H, 64 μL

CAUTION: *o*-Phenylenediamine is toxic and is a cancer suspect agent. Formic acid is very corrosive to the skin and should be dispensed in the **hood** using an automatic delivery pipet.

Reaction Conditions. Heat the reaction mixture, with stirring, at a sand bath temperature of 150–160 °C for 1 h.

Isolation of Product. Allow the reaction mixture to cool to room temperature. Add 630 μL of 10% aqueous sodium hydroxide solution (automatic delivery pipet). Crude benzimidazole precipitates at this point. Collect the product by vacuum filtration using a Hirsch funnel and wash the filter cake with three 0.5-mL portions of cold water (calibrated Pasteur pipet) (➡).

Purification and Characterization. Recrystallize the crude material from water, using a Craig tube. Dry the product in a desiccator or in a vacuum drying apparatus (see Prior Reading). Weigh the crystals and calculate the percent yield. Determine the melting point and compare your result with that listed in the literature.

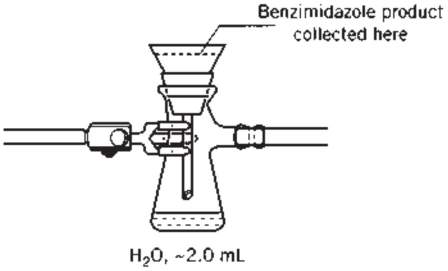

The UV spectrum of benzimidazole in 95% ethanol has been reported.[1]

Chemical Tests. Several tests may be run to assist in the identification of this material. Does the ignition test confirm the presence of the aromatic ring system? Does the soda lime or sodium fusion test indicate that nitrogen is present? Is the material soluble in 10% hydrochloric solution?

QUESTIONS

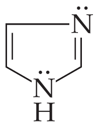

I

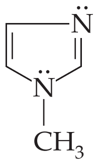

II

10W-6. The parent compound of the imidazole series, imidazole (**I**) itself, was first prepared in 1858: Can you account for the fact that it has a very high boiling point (256 °C), whereas 1-methyl imidazole (**II**) has the somewhat lower boiling point of 199 °C?

10W-7. The imidazole ring system has a great deal of aromatic character. Can you formulate two resonance structures that account for this characteristic?

10W-8. Imidazole is a weak acid, and thus reacts with strong bases to form the corresponding anion. Show this reaction, and draw resonance structures that account for the stability of the conjugate base.

10W-9. Suggest a mechanism for the dehydration involved in the last step in the synthesis of benzimidazole.

10W-10. Imidazole, acting as a nucleophile, catalyzes the hydrolysis of phenyl acetate by attack on the carbonyl carbon atom of the ester. The imidazole displaces the phenoxide anion and forms acetyl imidazole. In turn, the acetyl imidazole is quite unstable in water and hydrolyzes to form acetic acid, and regenerates the imidazole molecule. Write a suitable mechanism outlining these steps.

---

[1] Steck, E. A.; Nackod, F. C.; Ewing, G. W.; Gorman, N. H. *J. Am. Chem. Soc.* **1948,** *70,* 2406.

BIBLIOGRAPHY

The conditions of this reaction were adapted from those reported by

Wagner, E. C.; Millett, W. H. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 65.

For reviews see

Eichen, T.; Hauptmann, S. *The Chemistry of the Heterocycles,* 2nd ed.; Wiley-VCH: New York, 2003, Chap. 5.

Elderfield, R. C., Ed.; Heterocyclic Compounds (Five-membered Heterocyclic Compounds Containing Two Hetero Atoms and their Benzo Derivatives), Vol 5, Wiley: New York, 1957.

Grimmett, M. R. *Imidazole and Benzimidazole Synthesis,* Academic Press: New York, 1997.

Wright, J. B. *Chem. Revs.* **1951,** *48,* 397.

For applications involving ionic liquids see

Davis, J. H., Jr. *Chem. Letters* **2004,** *33,* 1072.

Seddon, K. R. *J. Chem. Technol. Biotechnol.* **1997,** *68,* 351.

Wasserscheid, P.; Welton, T. *Ionic Liquids in Synthesis;* Wiley-VCH Verlag: Stuttgart, Germany, 2002.

Welton, T. *Chem. Rev.* **1999,** *99,* 2071.

EXPERIMENT [3adv]

Heterocyclic Ring Synthesis: 4-Hydroxycoumarin and Dicoumarol

Common name: 4-hydroxycoumarin
CA number: [1076-38-6]
CA name as indexed: 2*H*-1-benzopyran-2-one, 4-hydroxy-

Common names: dicoumarol, dicoumarin
CA number: [66-76-2]
CA name as indexed: 2*H*-1-benzopyran-2-one, 3,3′-methylenebis[4-hydroxy]-

Purpose. You will synthesize a material, dicoumarol, that was the prototype for the oral anticoagulants widely used in medicine to lower blood coagulation rates. A carbon nucleophile is added to a carbonyl carbon to form a C—C bond. You will utilize a Claisen condensation reaction to prepare a β-ketoester, which, upon cyclization, forms a lactone, 4-hydroxycoumarin. Further condensation of two mole equivalents of 4-hydroxycoumarin with formaldehyde yields dicoumarol.

Prior Reading

Technique 2: Simple Distillation at the Semimicroscale Level (pp. 61–64)
Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
Technique 5: Crystallization
  Use of the Hirsch Funnel (pp. 88–89)
  Craig Tube Crystallizations (pp. 89–90)

REACTION

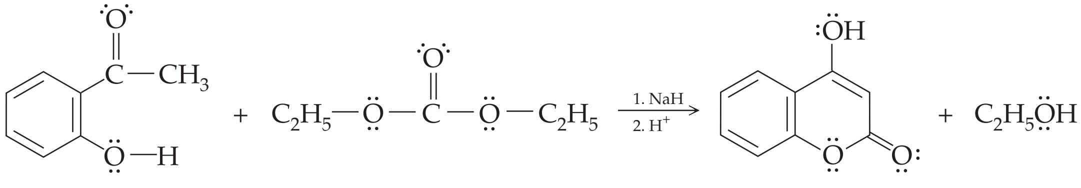

o-Hydroxyacetophenone Diethyl carbonate 4-Hydroxycoumarin Ethanol

DISCUSSION

This reaction sequence illustrates the well-known Claisen condensation, which is widely used to form C—C bonds. The bond formation is brought about by the nucleophilic attack of an enolate anion on the carbonyl carbon of an ester. The enolate is generated by removal of a slightly acidic hydrogen from the α-carbon atom of a ketone, nitrile, or ester, using a relatively strong base. The reaction mechanism is shown here. The methyl ketone is deprotonated by the base, sodium hydride. Sodium hydride (NaH) provides a basic and nonnucleophilic source of hydride ion (H^-). The resulting enolate then attacks the ester, diethyl carbonate. The β-ketoester product thus formed is esterified in an *intramolecular* reaction with the phenolic —OH group to form the lactone product (**I**). Note that the methylene hydrogen atoms in this lactone (**I**) are quite acidic because they are adjacent to two carbonyl groups. In the basic medium of this reaction, the sodium enolate of the lactone (**II**) will be formed. This enolate is water soluble, because it is a salt, which explains why the aqueous solution must be acidified to precipitate the neutral 4-hydroxycoumarin when isolating the product.

1. 4-Hydroxycoumarin

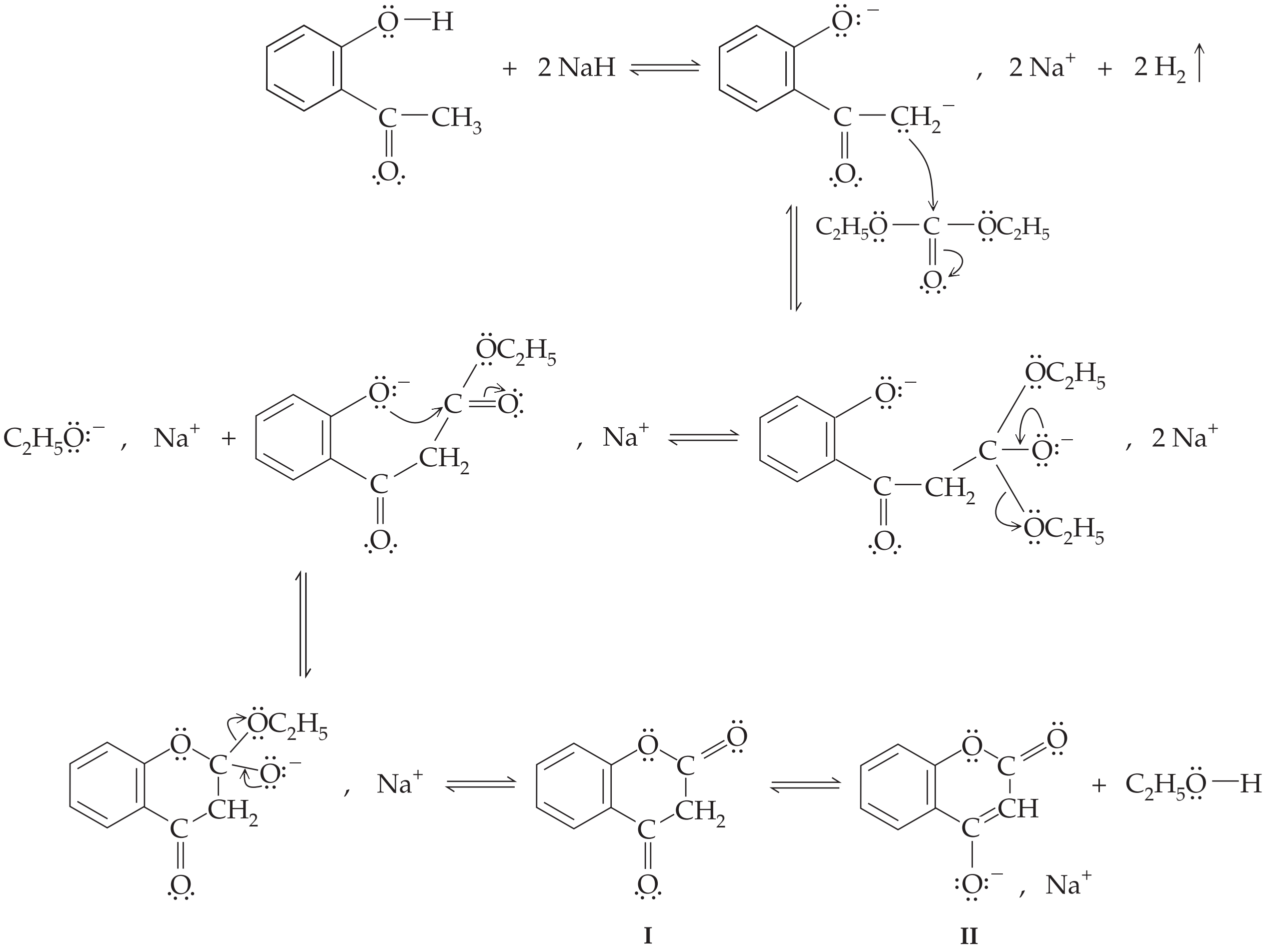

2. Dicoumarol

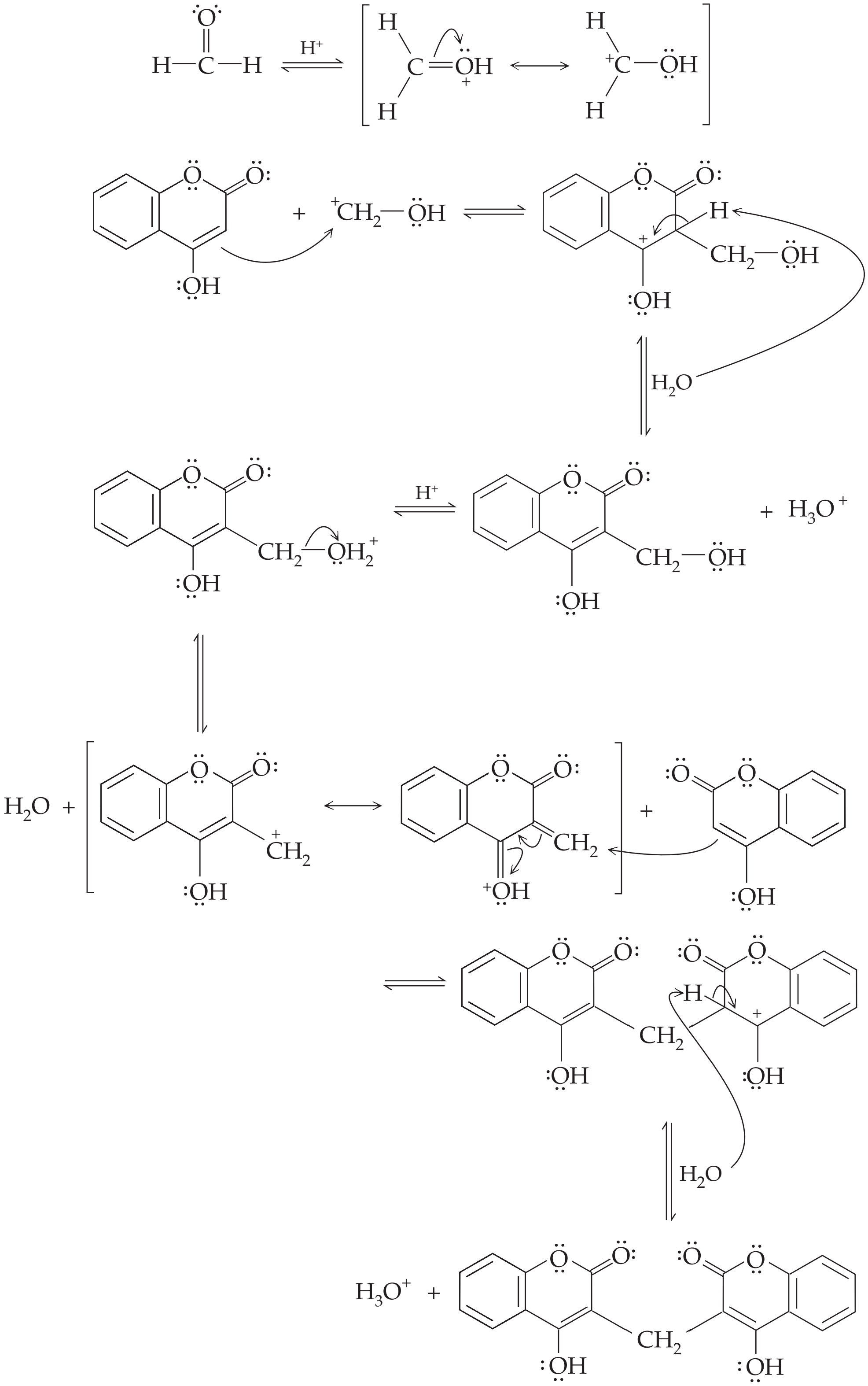
H⁺
H₂O
H⁺
+ H₃O⁺
H₂O +
+
H₂O
H₃O⁺ +

The aldol condensation of 4-hydroxycoumarin with formaldehyde provides an α,β-unsaturated carbonyl compound, which then undergoes a conjugate (1,4-) addition of a second molecule of 4-hydroxycoumarin. This reaction could be catalyzed by either trace base or trace acid; the acid-catalyzed reaction is shown and discussed here. The enol portion of 4-hydroxycoumarin is the nucleophile in an aldol reaction with protonated formaldehyde. The resulting product dehydrates to provide the α,β-unsaturated carbonyl compound, which, after protonation renders it a more reactive electrophile, then reacts with another nucleophilic molecule of 4-hydroxycoumarin in a conjugate addition reaction. This product, upon loss of a proton to the aqueous solvent, leads to dicoumarol. This substance is present in moldy sweet clover. It is a blood anticoagulant and its ingestion leads to the hemorrhagic sweet clover disease that kills cattle.

4-Hydroxycoumarin

Experiment [$3A_{adv}$]

The reaction is shown on pages 679–680.

EXPERIMENTAL PROCEDURE

Estimated time to complete the reaction: 4.0 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	bp (°C)	*d*	n_D
Sodium hydride (60% in oil dispersion)	24.0	85 mg	2.13			
Toluene	92.15	6.0 mL		111	0.86	1.4960
o-Hydroxyacetophenone	136.16	133 μL	1.1	218	1.13	1.5584
Diethyl carbonate	118.13	333 μL	2.75	126	0.96	1.3845

Reagents and Equipment.

NOTE. All equipment used in this reaction must be thoroughly dried in an oven at 110 °C for 30 min just prior to use. Upon removal from the drying oven, it should be allowed to cool to ambient temperature in a desiccator.

Weigh and place 85 mg (2.13 mmol) of sodium hydride (60% dispersion in oil) in a 10-mL round-bottom flask containing a magnetic stirring bar. Now add 3.0 mL of dry toluene. Attach the flask to a Hickman still fitted with an air condenser protected by a calcium chloride drying tube. Wrap the Hickman still 14/10 TS male joint with Teflon tape to prevent joint freeze-up (➡).

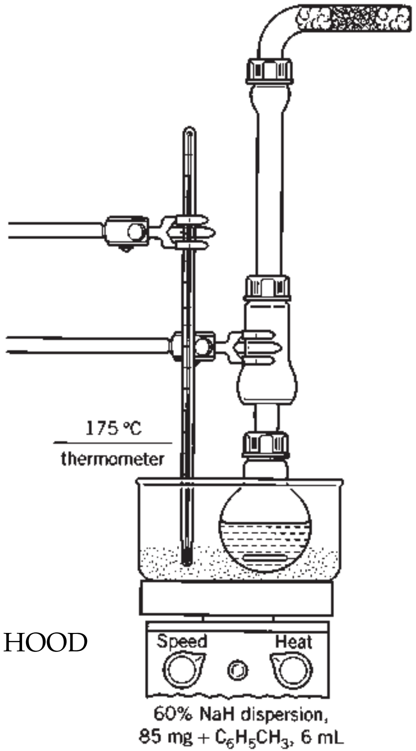

60% NaH dispersion, 85 mg + $C_6H_5CH_3$, 6 mL + *o*-$HOC_6H_4COCH_3$, 150 mg + $(CH_3CH_2O)_2CO$, 342 mg

CAUTION: Sodium hydride (NaH) is a flammable solid. Dispense in the **hood.** Toluene is distilled and stored over molecular sieves. Also dispense this solvent in the **hood.**

In rapid order, place 133 μL (150 mg, 1.1 mmol) of *o*-hydroxyacetophenone, 3.0 mL of dry toluene, and 333 μL (324 mg, 2.75 mmol) of diethyl carbonate in a stoppered 10-mL Erlenmeyer flask.

NOTE. Dispense the small volumes of the liquid reagents using an automatic delivery pipet. Be sure to dry the removable plastic tips in the oven before use. Use a 10-mL graduated cylinder to measure the toluene.

Remove the air condenser from the distilling head and add the *o*-hydroxyacetophenone solution, with stirring, to the reaction flask as rapidly as possible using a Pasteur pipet. The resulting solution turns yellow. Immediately reattach the air condenser.

Reaction Conditions. Place the reassembled apparatus in a sand bath and rapidly raise the temperature of the bath to about 175 °C.

CAUTION: To avoid excessively high temperatures, calibrate the hot plate temperature control before conducting the experiment.

Collect the ethanol and toluene distillate (1.0 mL in 2–3 fractions) in the collar of the still, and then remove the apparatus from the heat source. Allow the reaction solution to cool to room temperature, and then add 3.0 mL of water with stirring.

Isolation of Product. Transfer the resulting two-phase solution, using a Pasteur pipet, to a 15-mL centrifuge tube. Rinse the reaction flask with an additional 2.0 mL of water and add this to the centrifuge tube. Separate the toluene layer using a Pasteur filter pipet and transfer it to a second 15-mL centrifuge tube. Then extract this organic phase with 3.0 mL of water, and add the water extract to the original water phase. *This aqueous solution contains the water-soluble sodium enolate* (**II**). Cool the combined aqueous layers in an ice bath and acidify them by drop-wise addition of concentrated hydrochloric acid delivered from a Pasteur pipet. The solid product precipitates from the aqueous phase. Add acid until the yellow color of the solution disappears (~ 10 drops). Collect the product by vacuum filtration using a Hirsch funnel (⬅).

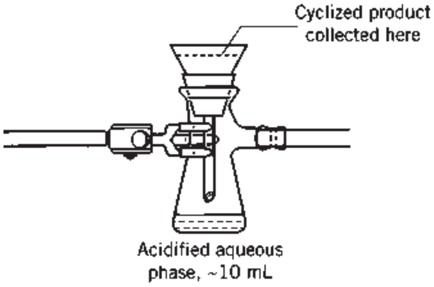
Cyclized product collected here
Acidified aqueous phase, ~10 mL

Purification and Characterization. Recrystallize the crude 4-hydroxycoumarin from 50% ethanol using a Craig tube. Dry the material on a porous clay plate or in a vacuum apparatus (see Prior Reading).

Weigh the product and calculate the percent yield. Determine the melting point and obtain an IR spectrum. Compare your results with those reported in the literature.

Chemical Tests. You may wish to perform the ignition test to establish that the compound is aromatic. Does the ferric chloride test (Chapter 9) for phenols give a positive result?

Experiment [3B$_{adv}$]

Dicoumarol

REACTION

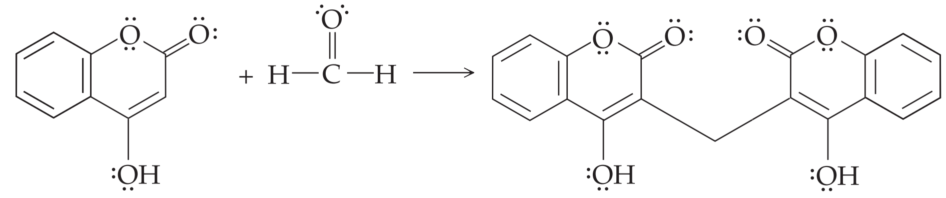

4-Hydroxycoumarin Formaldehyde Dicoumarol

EXPERIMENTAL PROCEDURE

Estimated time to complete the reaction: 0.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)	bp (°C)
4-Hydroxycoumarin	162.15	50 mg	0.31	213–214	
Water		15 mL			100
Formaldehyde (37% in H_2O)		0.5 mL	6		

Reagents and Equipment. Weigh and place 50 mg (0.31 mmol) of 4-hydroxycoumarin, followed by 15 mL of water, in a 50-mL Erlenmeyer flask containing a boiling stone. Heat the mixture to boiling on a hot plate. Now add 0.5 mL (~ 500 mg, ~ 6 mmol) of formaldehyde (37% aqueous solution) to the resulting solution.

CAUTION: Formaldehyde is a cancer suspect agent. Dispense in the **hood.**

HOOD

Reaction Conditions. White crystals of the product form immediately on addition of the formaldehyde solution. Cool the flask in an ice bath.

Isolation of Product. Collect the solid product by vacuum filtration using a Hirsch funnel (➡). Wash the filter cake with three 1-mL portions of ice water. Remove the crystals and dry them on filter paper or on a porous clay plate.

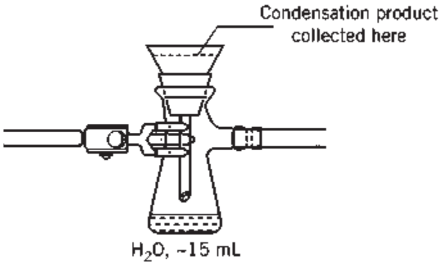

Purification and Characterization. Recrystallize the material from a mixture of toluene/cyclohexanone (~ 2:1) using a Craig tube. After drying the isolated material, determine its melting point and compare your value with that in the literature. Obtain an IR spectrum and compare your spectrum with that in the literature (*The Aldrich Library of IR Spectra* and/or Sci Finder Scholar).

QUESTIONS

10W-11. Why must water be excluded from the reaction that generates 4-hydroxycoumarol?

10W-12. In the first step of the above reaction, formation of 4-hydroxycoumarin, what is the purpose of removing the ethanol generated in the reaction?

10W-13. The Dieckmann condensation is simply an intramolecular Claisen condensation. For example,

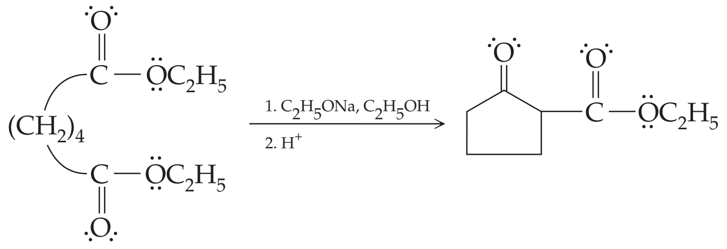

Suggest a suitable mechanism to account for the formation of the cyclic β-ketoester.

10W-14. Predict the product formed in each of the following reactions:

(a) Acetophenone + diethyl carbonate $\xrightarrow[\text{2. H}^+]{\text{1. NaH, toluene}}$

(b) Acetophenone + ethyl formate $\xrightarrow[\text{2. H}^+]{\text{1. C}_2\text{H}_5\text{ONa, ethanol}}$

10W-15. Suggest a synthesis for each of the following compounds using the Claisen condensation. Any necessary organic or inorganic reagents may be used.

(a)

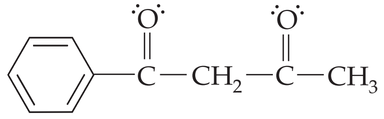

(b)

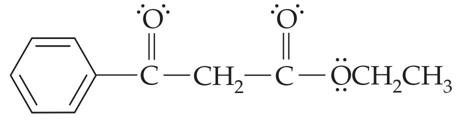

(c)

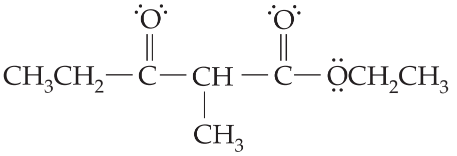

$CH_3CH_2-\overset{O}{\overset{\|}{C}}-\underset{CH_3}{\underset{|}{CH}}-\overset{O}{\overset{\|}{C}}-OCH_2CH_3$

(d)

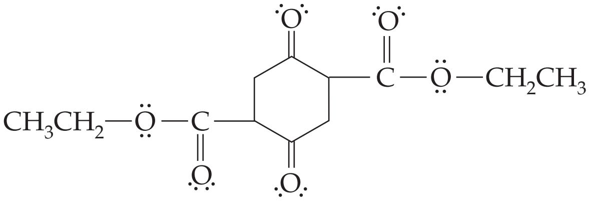

10W-16. Draw structures for the two possible mono enols of 1,3-cyclohexanedione. Explain which one is more stable.

BIBLIOGRAPHY

A review of the Claisen condensation is given in

Davis, B. R.; Garratt, P. *J. Comp. Org. Syn.* **1991,** *2,* 795.
Garst, J. F. J. *Chem. Educ.* **1979,** *56,* 721.
Hauser, C. R.; Swamer, F. W.; Adams, J. T. *Org. React.* **1954,** *8,* 59.

See your organic textbook for an introduction to this reaction and its scope in synthesis. For example,

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed.; Wiley: New York, 2007, Chap. 16, p. 1452.
Solomons, T. W. G.; Fryhle, C. B. *Organic Chemistry,* 9th ed.; Wiley: New York, 2008, p. 842.

Selected references of Claisen condensations from *Organic Syntheses* include

Ainsworth, C. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 536.
Floyd, D. E.; Miller. S. E. *Organic Syntheses;* Wiley: New York, 1963; Collect. Vol. IV, p. 141.
John, J. P.; Swaminathan, S.; Venkataraman, P. S. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 747.
Magnani, A.; McElvain, S. M. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 251.
Riegel, E. R.; Zwilgmeyer, F. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 126.
Snyder, H. R.; Brooks, L. A.; Shapiro, S. H. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 531.

For the synthesis of coumarin see:

Sethna, S.; Phadke, R. *Org. React.* **1953,** *7,* 1.

It may be of interest in a broad aspect to review

Corey, E. J.; Czakó, B; Kürti, L. *Molecules and Medicine,* Wiley: New York, 2007.

EXPERIMENT [4_{adv}]

Grignard and Aryl Halide Cross-Coupling Reaction: 1-Methyl-2-(methyl-d_3)-benzene

Common names: 1-methyl-2-(methyl-d_3)-benzene, α,α,α-d_3-o-xylene
CA number: [25319-53-3]
CA name as indexed: benzene, 1-methyl-2-(methyl-d_3)-

Purpose. This experiment illustrates a selective cross-coupling reaction between an alkyl Grignard reagent[2] and an aryl halide in the presence of a nickel phosphine catalyst to form a C—C bond. The reaction is also used to demonstrate the technique of "labeling" specific hydrogen atoms with an isotope for identification purposes.

Prior Reading

Technique 4: Solvent Extraction
  Liquid–Liquid Extraction (p. 72)
  Drying of the Wet Organic Layer (pp. 80–82)
Technique 6: Chromatography
  Column Chromatography (pp. 92–93)
  Concentration of Solutions (pp. 101–104)

REACTIONS

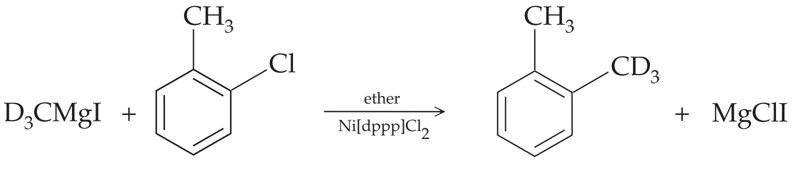

o-Chlorotoluene 1-Methyl-2-(methyl-d_3)-benzene Magnesium chloroiodide

Alternatively, the use of methyl iodide in the formation of the Grignard reagent produces *o*-xylene in step 2.

DISCUSSION

Before the discovery of the nickel-phosphine catalyst, the cross-coupling of Grignard reagents with organic halides was seldom used in synthetic practice. This fact was mainly due to the formation of homocoupling products along with significant elimination side reactions. With the use of this new catalyst, the reaction now has wide application in the synthesis of unsymmetrical alkanes and alkenes.

In this experiment we will synthesize an isotopically labeled *o*-xylene. The *o*-xylene product will be labeled with a trideuteromethyl group. Deuterium, one of the hydrogen **isotopes,** can be used as a **label** in organic compounds. Its incorporation into a molecule can be detected by IR, NMR, or MS. Deuterium labeling is a particularly powerful way of investigating infrared spectra because the resultant frequency shifts (which are inversely proportional to the square root of the mass ratio; see Infrared Discussions, Chapter 8) are the largest obtainable with stable isotopes. Isotopic labeling is often used to follow hydrogen-deuterium exchange reactions, such as enolizations, to study biological reactions, and for gaining insight into organic reaction mechanisms. The nickel catalyst used in this experiment is [1,3-bis(diphenylphosphino)propane]nickel(II) chloride, mercifully abbreviated as Ni[dppp]Cl_2.

---

[2] For general references on Grignard reagents refer to Experiment [16].

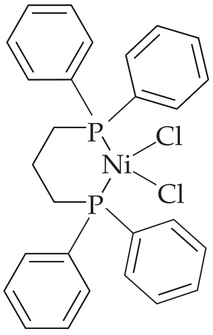

[1,3-Bis(diphenylphosphino)propane]nickel(II) chloride
(Ni[dppp]Cl_2)

It has been proposed that the catalytic ability of the dihalodiphosphinenickel reagent stems from its ability to react with a Grignard reagent to form a diorgano-nickel complex (A).[3] This complex is then converted to the halo-organonickel complex B, by reacting it with an *organic halide:*

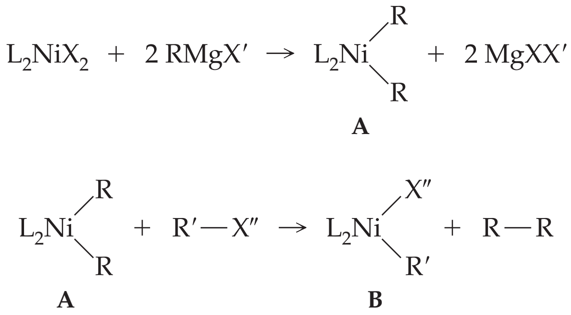

$$\mathrm{L_2NiX_2} + 2\,\mathrm{RMgX'} \rightarrow \underset{\mathbf{A}}{\mathrm{L_2Ni}\begin{matrix}\mathrm{R}\\ \mathrm{R}\end{matrix}} + 2\,\mathrm{MgXX'}$$

$$\underset{\mathbf{A}}{\mathrm{L_2Ni}\begin{matrix}\mathrm{R}\\ \mathrm{R}\end{matrix}} + \mathrm{R'{-}X''} \rightarrow \underset{\mathbf{B}}{\mathrm{L_2Ni}\begin{matrix}\mathrm{X''}\\ \mathrm{R'}\end{matrix}} + \mathrm{R{-}R}$$

(where L_2 = organodiphosphine ligand;
R = alkyl or aryl; R′aryl or vinyl;
X, X′ and X″ = halo

Reaction of complex B with the Grignard reagent forms a new diorganonickel complex C, from which the cross-coupling product is formed by attack of the organic halide. In the reaction, complex B is regenerated and thus the catalytic cycle is completed:

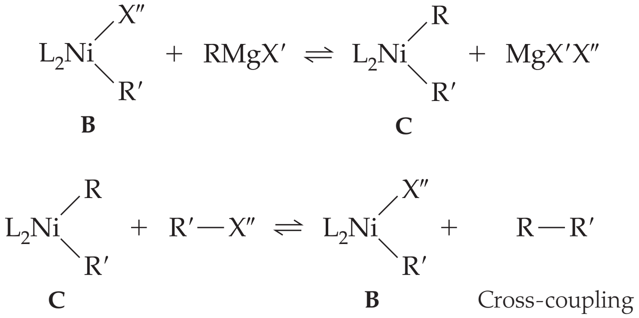

$$\underset{\mathbf{B}}{\mathrm{L_2Ni}\begin{matrix}\mathrm{X''}\\ \mathrm{R'}\end{matrix}} + \mathrm{RMgX'} \rightleftharpoons \underset{\mathbf{C}}{\mathrm{L_2Ni}\begin{matrix}\mathrm{R}\\ \mathrm{R'}\end{matrix}} + \mathrm{MgX'X''}$$

$$\underset{\mathbf{C}}{\mathrm{L_2Ni}\begin{matrix}\mathrm{R}\\ \mathrm{R'}\end{matrix}} + \mathrm{R'{-}X''} \rightleftharpoons \underset{\mathbf{B}}{\mathrm{L_2Ni}\begin{matrix}\mathrm{X''}\\ \mathrm{R'}\end{matrix}} + \underset{\text{Cross-coupling product}}{\mathrm{R{-}R'}}$$

---

[3] Tamao, K.; Sumitani, K.; Kumada, M. *J. Am. Chem. Soc.* **1972,** *94,* 4374.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 5.0 h.

Physical Properties of Reactants and Product

Compound	MW	Amount	mmol	mp (°C)	bp (°C)	*d*	n_D
Magnesium	24.31	65 mg	2.7	649			
Diethyl ether	74.12	700 μL			34.5	0.71	1.3526
Methyl-d_3-iodide	144.96	250 μL	3.93		42	2.28	1.5262
Ni[dppp]Cl_2	542.1	10 mg	0.02				
o-Chlorotoluene	126.59	150 μL	1.28		159	1.08	1.5268
1-Methyl-2-(methyl-d_3)-benzene	109.17				144		1.5055

NOTE. All the glassware used in the experiment should be cleaned, dried in an oven at 110 °C for at least 30 min, and then cooled in a desiccator before use.

Reagents and Equipment. Equip a 5.0-mL conical vial containing a magnetic spin vane with a reflux condenser protected by a calcium chloride drying tube.

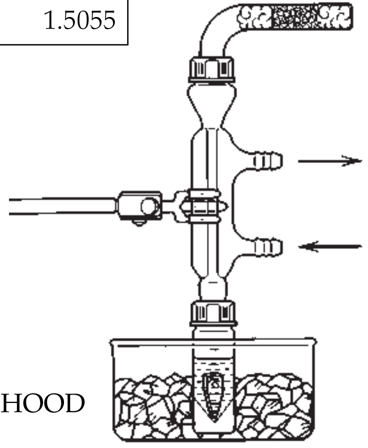

Step I:
Mg, 65 mg +
$(CH_3CH_2)_2O$, 0.5 mL +
CD_3I, 250 μL

Step 1. Scrape a 4- to 5-in. piece of magnesium ribbon to remove the oxide coating. Then cut it into 1-mm-long sections. *Using forceps,* weigh and add 65 mg (2.7 mmol) of the magnesium sections to the conical vial. In the **hood,** use an automatic delivery pipet to dispense 300 μL of anhydrous diethyl ether into the conical vial containing the magnesium. Reassemble the apparatus.

Using automatic delivery pipets, prepare a solution of 250 μL of methyl-d_3 iodide in 200 μL of anhydrous diethyl ether in a capped vial.

CAUTION: Methyl-d_3 iodide is a suspected carcinogen.

Draw this solution into a Pasteur filter pipet. Remove the drying tube and insert the pipet down the length of the condenser, allowing the pipet bulb to rest on the condenser lip.

Reaction Conditions. Place the reaction vial in an ice bath and add the methyl-d_3 iodide dropwise with stirring. Withdraw the pipet, reinstall the drying tube, and stir the mixture for 20 min (➡).

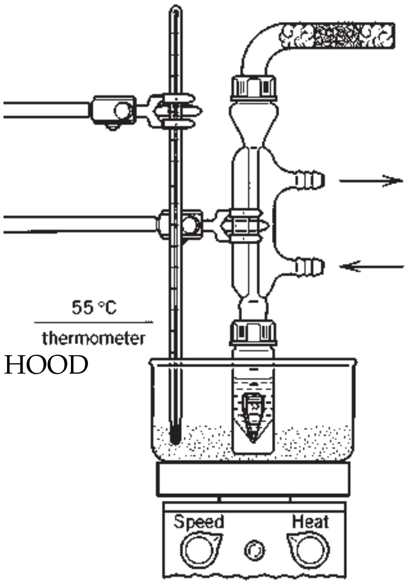

Step II:
Ni[dppp]Cl_2, 10 mg
+ o-$CH_3C_6H_4Cl$, 150 μL
+ $(CH_3CH_2)_2O$, 200 μL

Step 2. In this step, all liquids must be dispensed in the **hood** using automatic delivery pipets. Weigh and add to a capped 0.5-dram vial, 10 mg (0.02 mmol) of Ni[dppp]Cl_2.

CAUTION: Ni[dppp]Cl_2 is a cancer suspect agent.

Now add 150 μL of *o*-chlorotoluene and 200 μL of anhydrous diethyl ether. Draw this solution into a Pasteur filter pipet and, in one portion, add it to the reaction vial through the condenser as described above. Reinsert the drying tube. Place this mixture in a preheated sand bath at a temperature of 55 °C. Heat with stirring for 2 h (➡).

On addition of the catalyst, the solution turns green. After approximately 30 min of heating, the mixture becomes dark brown.

Isolation of Product. Cool the reaction mixture in an ice bath and quench the reaction by the dropwise addition (calibrated Pasteur pipet) of 1.5 mL of 1.0 M HCl solution. *The acid should be added slowly because frothing occurs.*

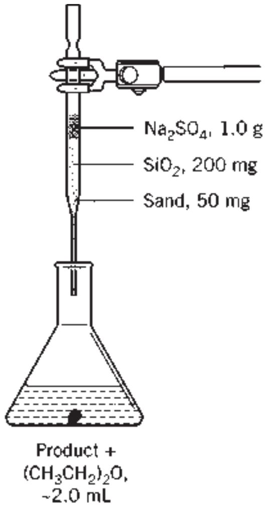

NOTE. The following extracting solutions are measured using calibrated Pasteur pipets. For each extraction operation, the vial is capped, shaken, and vented, and the layers are allowed to separate. The aqueous (lower) phase is then removed.

Add 1.0 mL of ether to the solution. Using a Pasteur filter pipet, remove the aqueous layer. Extract the remaining ether phase with 1.0 mL of water followed by 1.0 mL of saturated sodium bicarbonate solution and then 1.0 mL of deionized water. Then extract with two 1.0-mL portions of 1.0 N sodium thiosulfate (Na_2SO_3) solution and then with 1.0 mL of water.

Purification and Characterization. Isolate and purify the reaction product using column chromatography. In a modified Pasteur filter pipet, place 200 mg of activated silica gel (100 mesh) followed by 1.0 g of anhydrous sodium sulfate. Wet the column with 0.5 mL of ether using a calibrated Pasteur pipet (⬅).

Transfer the ether solution of product to the column by Pasteur pipet, and elute the material from the column with two 0.5-mL portions of ether. Collect the eluate in a tared 10-mL Erlenmeyer flask and then transfer the eluate to a 5-mL conical vial for concentration. Evaporate the solvent by warming the vial in a sand bath in the **hood** to yield the "labeled" *o*-xylene. HOOD

Weigh the product and calculate the percent yield. Determine the boiling point, density, and refractive index (optional) and compare your values with those listed for *o*-xylene in the literature. Obtain an IR spectrum of the crude (dry) reaction product using the capillary film sampling technique and analyze it following the analysis given below.

Infrared Analysis. The spectrum of *o*-chlorotoluene (Fig. 10.3W) nicely mimics the large macro group frequency train for an ortho-substituted dialkylbenzene system, which has the following frequency train: 3100–3000, 3000–2850, 1950, 1920, 1880, 1840, 1790, 1690, 1600, 1580, 1500, 1450, 1380, and 750 cm^{-1}.

a. **3072 and 3026 cm^{-1}:** C—H stretch on aromatic ring.
b. **3000–2850 cm^{-1}:** C—H stretch on alkyl substituents.
c. **1955, 1915, 1880, 1835, 1795, and 1690 cm^{-1}:** Combination band pattern for ortho-disubstituted rings. The fit in this case is extremely good.
d. **1596 and 1577, 1475 and 1450 cm^{-1}:** Two degenerate pairs of ring stretching modes. The 1450 wavenumber band is overlapped by the antisymmetric methyl deformation vibrations.
e. **1382 cm^{-1}:** Symmetric methyl deformation mode.
f. **749 cm^{-1}:** All in-phase out-of-plane bend of four adjacent C—H groups on the aromatic ring.

The C—Cl stretch falls below the region of measurement in these aromatic systems.

The IR spectrum of 1-methyl-2-(methyl-d_3)-benzene is shown in Figure 10.4W. The reaction has replaced the —Cl group with a —CD_3 group. Thus, the only spectral changes expected to be observed will involve the vibrations of the

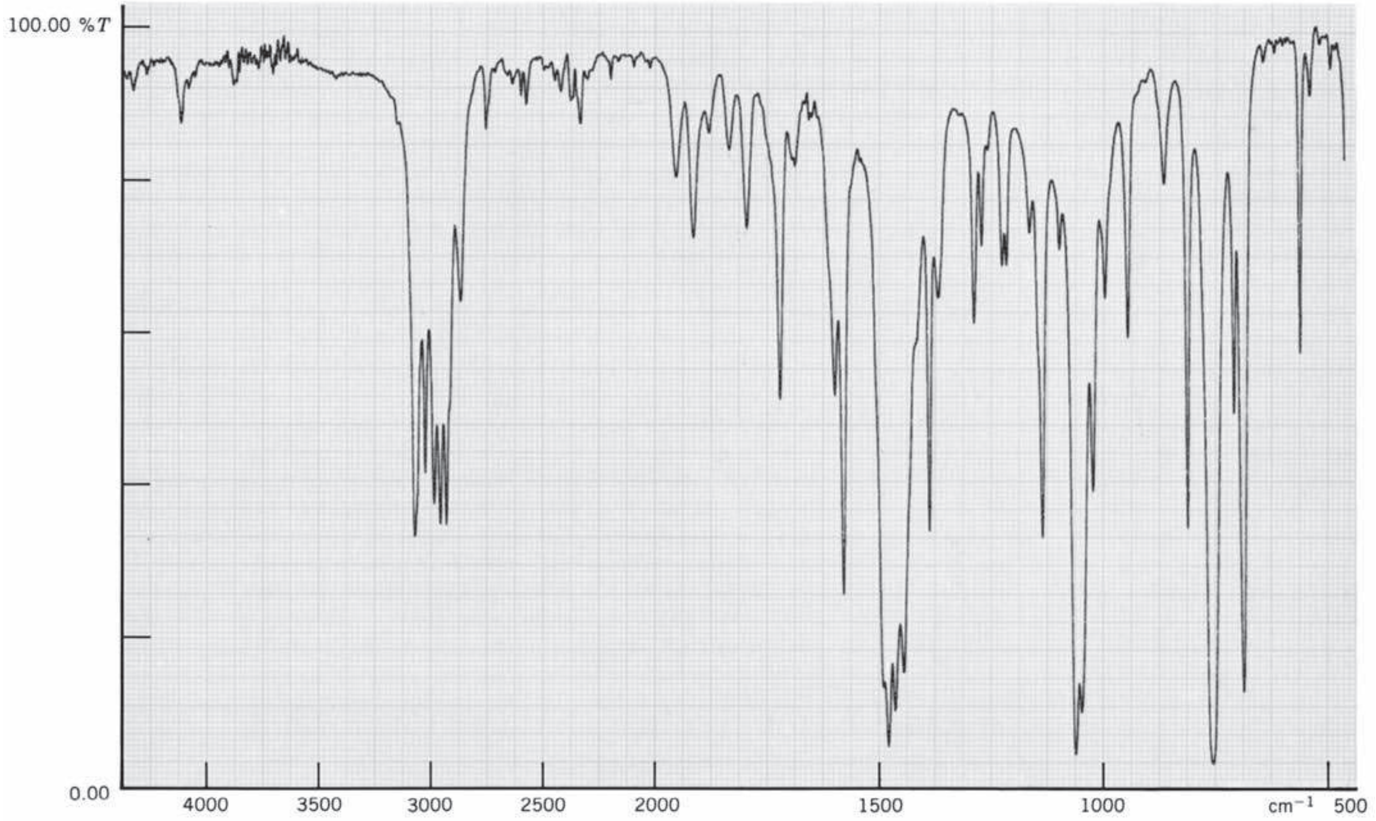

Sample 2-Chlorotoluene
%T X ABS ___ Background Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm-1
Sample Condition liquid, neat
Cell Window KBr
Cell Path Length capillary film
Matrix Material ___

Figure 10.3W IR spectrum: 2-chlorotoluene.

labeled group. To a first approximation, the frequency should shift by a factor of $1/\sqrt{2} = 0.707$, but in practice the full shift is not observed because most vibrations are not pure modes. A factor of 0.75–0.72 is generally observed. In 1-methyl-2-(methyl-d_3)-benzene, we have two sets of doublets, with the doublet centers at 2220 and 2090 cm^{-1}. Because we would expect the C—D stretching modes to give rise to only two band systems, coupling with overtones of lower lying fundamental vibrations is likely occurring. The corresponding C—H stretching modes also evidence some coupling. If we use the major band centers observed at 2965 and 2870 cm^{-1}, however, as the C—H stretching values, the observed ratios are $2220/2965 = 0.749$ and $2090/2870 = 0.728$, results that are quite reasonable to expect for this type of isotopic shift. The bending modes are moved into the cluttered fingerprint region and do not easily lend themselves to analysis.

Examine the spectrum of the reaction product you have obtained as a capillary film. Discuss the similarities and differences of the experimentally derived spectral data to the reference spectra (Figs. 10.3W and 10.4W).

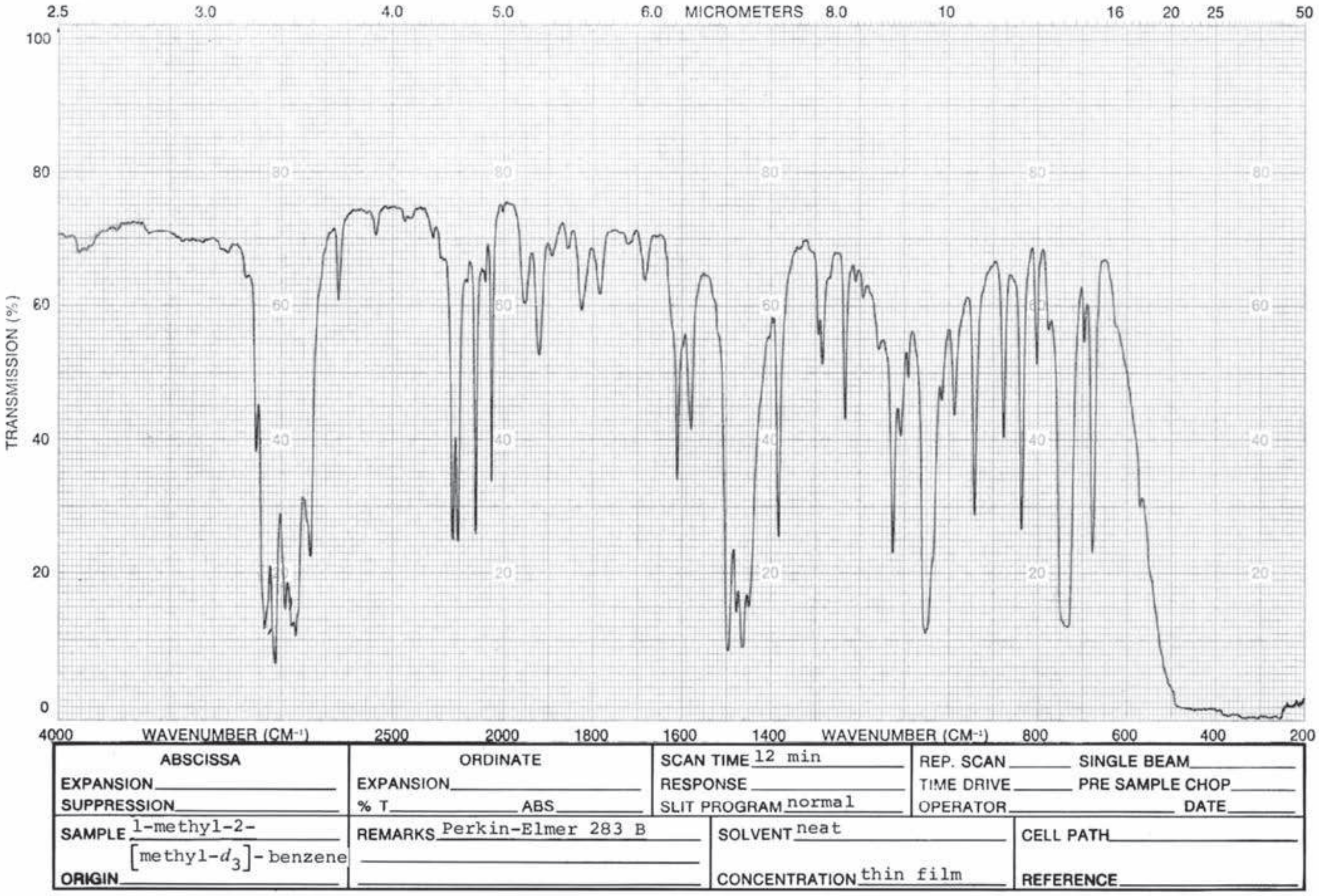

ABSCISSA
EXPANSION
SUPPRESSION

ORDINATE
EXPANSION
% T ABS

SCAN TIME 12 min
RESPONSE
SLIT PROGRAM normal

REP. SCAN SINGLE BEAM
TIME DRIVE PRE SAMPLE CHOP
OPERATOR DATE

SAMPLE 1-methyl-2- [methyl-d_3]- benzene
ORIGIN

REMARKS Perkin-Elmer 283 B

SOLVENT neat
CONCENTRATION thin film

CELL PATH
REFERENCE

Figure 10.4W IR spectrum: 1-methyl-2-(methyl-d_3)-benzene.

QUESTIONS

10W-17. Predict the reagents that could be used to prepare each of the compounds by the Grignard coupling reaction. Give a suitable name to each reactant.

(a)

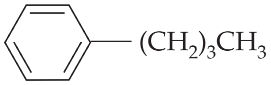

(c)

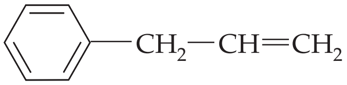

(b)

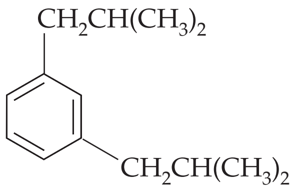

(d)

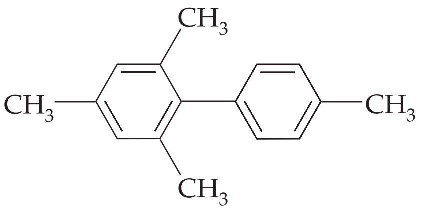

10W-18. The Wurtz coupling reaction involves the treatment of haloalkanes with an active metal, such as sodium.

(a) Butyl bromide + Na
(b) 1-Bromo-3-chorocyclobutane + Na
What products would be formed in the above reactions? Name the products.

10W-19. In reference to Question 10W-18, explain why the coupling reaction of *n*-butyl bromide and *n*-propyl bromide under conditions of the Wurtz reaction is synthetically inefficient for the preparation of heptane.

10W-20. Describe what you would expect to see in the ^{1}H and ^{13}C NMR spectra of 1-methyl-2-(methyl-d_3)-benzene.

10W-21. The spectrum of *o*-chlorotoluene given in Figure 7.3 has two additional peaks that are not present in the library reference standard spectrum. Both these bands are weak. One occurs near 1714 cm^{-1} and the other at 1362 cm^{-1}. Can you offer an explanation for the presence of the extra absorption bands? How successful do you think the coupling reaction would be if carried out using this sample as the starting material?

10W-22. In the spectrum of *o*-chlorotoluene, the lower wavenumber band of the 1596, 1577 pair is the more intense member. In the labeled product, the 1610, 1580 pair has the higher wavenumber peak more intense. Explain this reversal of intensities.

BIBLIOGRAPHY

The present coupling reaction is based on the work reported by

Kumada, M.; Tamao, K.; Sumitani, K. *Organic Syntheses;* Wiley: New York, 1988; Collect. Vol. VI, p. 407, and references cited therein.

Selected Grignard coupling reactions presented in *Organic Syntheses*:

Gilman, H.; Catlin, W. E. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 471.

Gilman, H.; Robinson, J. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 47.

Lespieau, R.; Bourguel, M. *Organic Syntheses;* Wiley: New York, 1941; Collect. Vol. I, p. 186.

Smith, L. I. *Organic Syntheses;* Wiley: New York, 1943; Collect. Vol. II, p. 360.

Turk, A.; Chanan, H. *Organic Syntheses;* Wiley: New York, 1955; Collect. Vol. III, p. 121.

The reaction was first described by

Mayo, D. W.; Bellamy, L. J.; Merklin, G. T.; Hannah, R. W. *Spectrochim. Acta* **1985,** *41A,* 355.

Oxidative Coupling of 2-Naphthol: 1,1′-Bi-2-Naphthol

EXPERIMENT [5 ADV]

Common name: 1,1′-bi-2-naphthol
CA number: [602-0-5]
CA name as indexed: [1,1′-binaphthalene]-2,2′-diol

Purpose. The coupling reaction that aromatic phenols undergo in the presence of transition metal oxidants is investigated. This reaction mimics the biogenetic process that occurs in nature.

Prior Reading

Technique 5: Crystallization
Use of the Hirsch Funnel (pp. 88–89)
Craig Tube Crystallizations (pp. 89–90)

REACTION

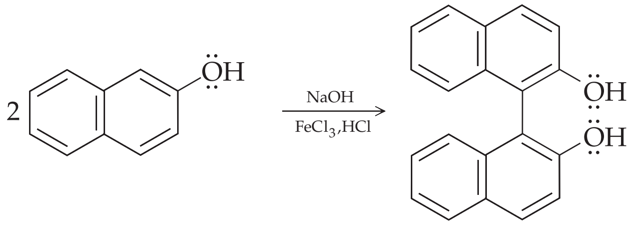

2-Naphthol 1,1'-Bi-2-naphthol

DISCUSSION

This reaction illustrates oxidative coupling of phenols, which is an important biogenetic pathway in nature, leading to the formation of many different natural products.

The coupling reaction involves oxidation of 2-naphthol by electron transfer to give an aryloxy radical, which then dimerizes to yield the product. The mechanism is shown here:

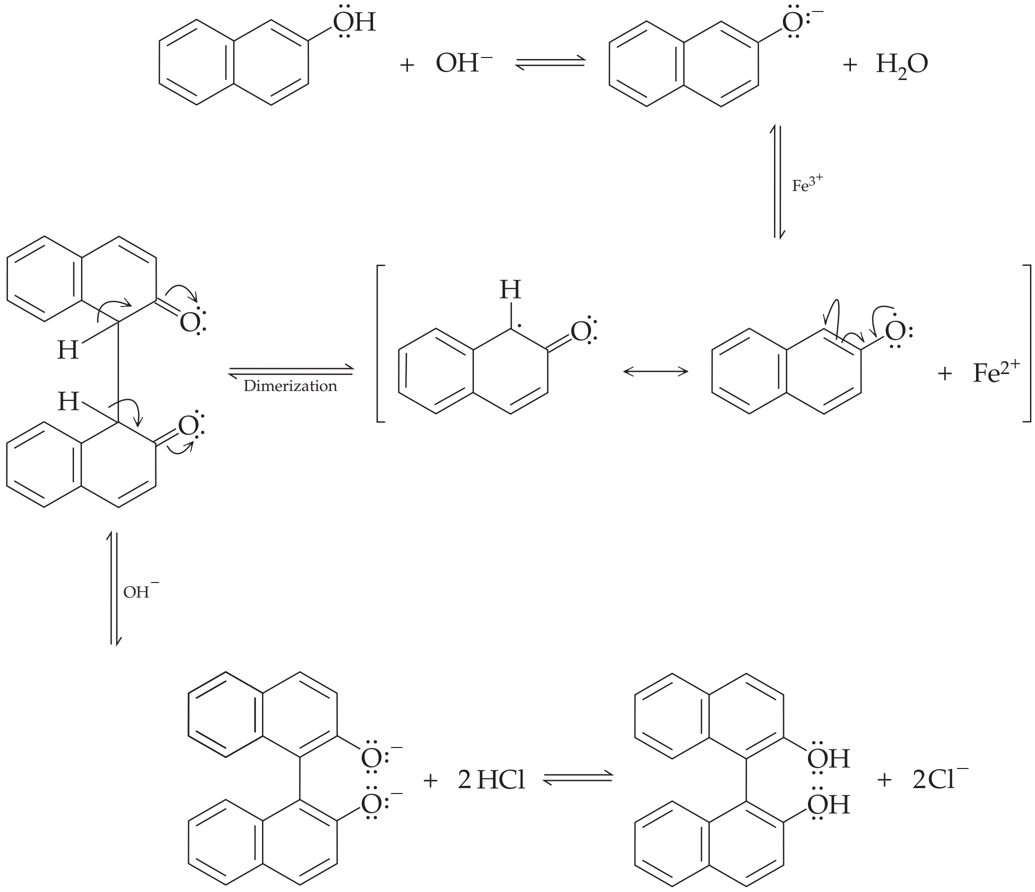

This binaphthol compound, by virtue of restricted rotation about the single bond joining the two naphthalene groups, is a chiral compound. That is, the molecule cannot readily exist in a planar form because of the steric interference of the bulky —OH substituents, and, in fact, the two enantiomers have been separated. Such enantiomeric compounds contain no stereocenter, but rather have an axis of chirality, or a *chiral axis*, which in this case contains the bond joining the two napthalene rings (see Fig. 10.5W)[4]:

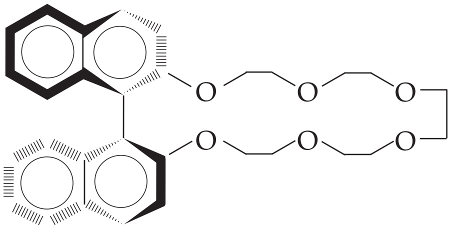

---

[4] Donald J. Cram (Nobel Laureate, 1987) and coworkers have incorporated this binaphthyl group into cyclic crown ethers.

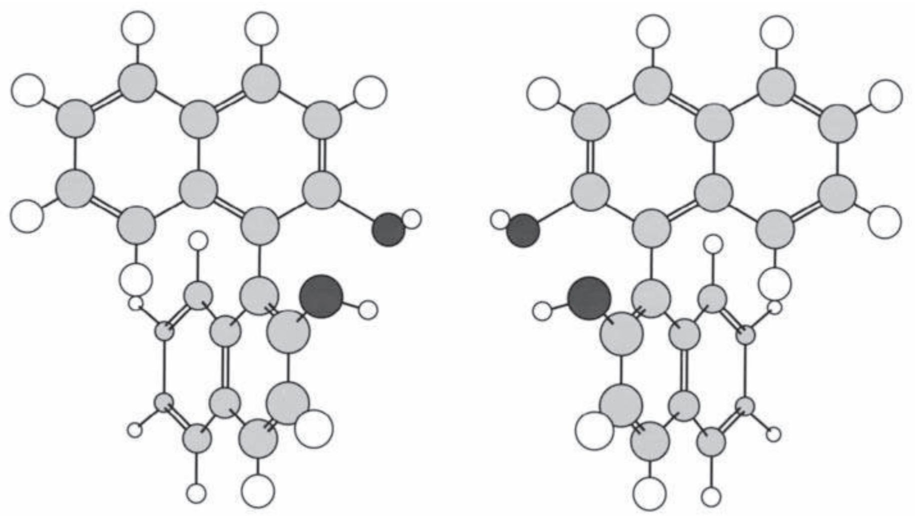

Figure 10.5W Enantiomers of 1,1′-bi-2-naphthol.

An investigation of the complexation of this class of molecules with various ionic species may lead to important insights into the catalytic nature of enzymes. For example, the crown ether similar to that above containing *two* 2,2′-substituted 1,1′-binaphthyl groups as chiral barriers complexes preferentially with one enantiomer of some primary amine salts:

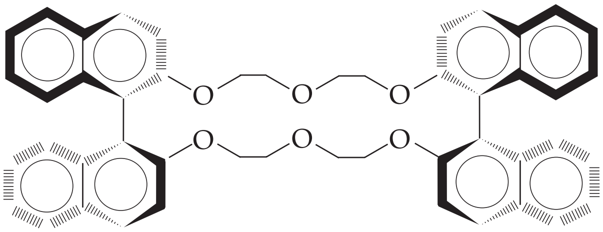

Complexation of enantiomerically pure binaphthols with metal hydride reducing agents allow enantiospecific reductions of some carbonyl compounds.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.0. h.

Physical Properties of Reactants				
Compound	**MW**	**Amount**	**mmol**	**mp (°C)**
2-Naphthol	144.19	100 mg	0.69	123–124
Sodium hydroxide	40.00	30 mg	0.75	318.4
Ferric chloride · $6H_2O$	270.30	297 mg	1.1	37
HCl (concd)	36.46	100 μL		
Water	18.06	3 mL		

Reagents and Equipment. Weigh and add 100 mg (0.69 mmol) of 2-naphthol and 30 mg (0.75 mmol) of sodium hydroxide to a 10-mL round-bottom flask containing a stirring bar.

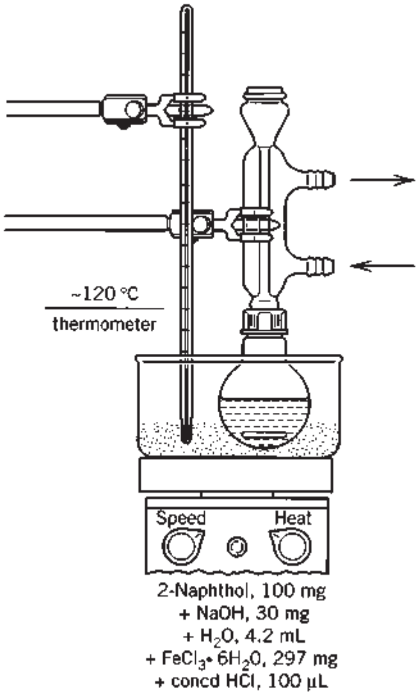

CAUTION: Sodium hydroxide is a corrosive and toxic chemical. Do not allow it to touch your skin or to come in contact with your eyes.

Add 3.0 mL of water. Attach the flask to a reflux condenser and place the assembly in a sand bath on a magnetic stirring hot plate (⬅).

Heat the reaction mixture to reflux, with stirring, using a sand bath temperature of 120 °C.

In a 10-mL Erlenmeyer flask prepare a solution of 180 mg (1.1 mmol) of *anhydrous* ferric chloride (MW 162) *or* 297 mg (1.1 mmol) of ferric chloride *hexahydrate* (MW 270), 1.0 mL of water (calibrated Pasteur pipet), and 100 μL of concentrated hydrochloric acid.

CAUTION: Be careful when mixing acid with water. ***Add the acid to the water.*** Avoid contact with the skin. Dispense the acid with an automatic delivery pipet.

Using a Pasteur pipet, transfer the ferric chloride solution, through the top of the condenser, to the reaction flask. Rinse the Erlenmeyer flask with 200 μL of water, and add this rinse to the reaction flask as before.

Reaction Conditions. Heat the resulting mixture at reflux, with stirring, for 45–60 min using a sand bath temperature of 120 °C. Allow the mixture to cool to room temperature, and then place the flask in an ice bath to complete crystallization of the product.

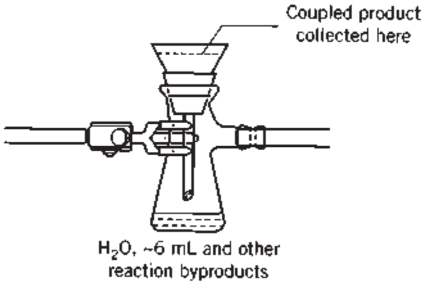
Coupled product collected here
H_2O, ~6 mL and other reaction byproducts

Isolation of Product. Collect the solid product by vacuum filtration using a Hirsch funnel, and wash the filter cake with two 1.0-mL portions of cold water (calibrated Pasteur pipet) (⬅).

Purification and Characterization. Recrystallize the crude product from 95% ethanol using a Craig tube, and dry the resulting crystals on a porous clay plate or on filter paper.

Weigh the crystals and calculate the percent yield. Determine the melting point and compare it with the literature value. Obtain the IR spectrum and compare it to that shown in the literature (*The Aldrich Library of IR Spectra* and/or SciFinder Scholar).

Chemical Tests. Chemical classification tests (Chapter 9) may be used to assist in characterization of this material. Perform the ignition test and the ferric chloride test.

QUESTIONS

10W-23. In the oxidative coupling reaction of 1-naphthol, it is *possible* to obtain three products:

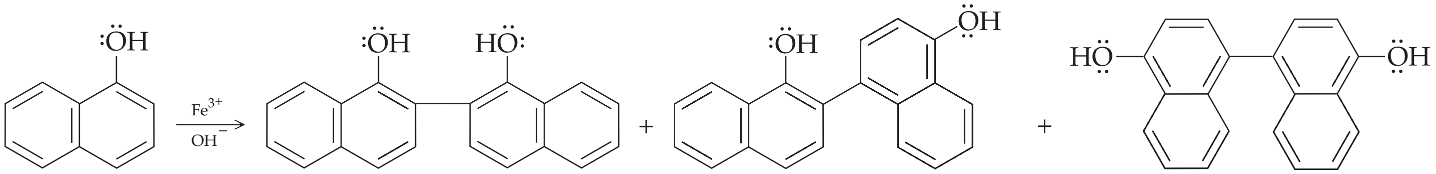

Account for the formation of this mixture by suggesting a mechanistic sequence similar to that presented in the discussion section of this experiment.

10W-24. Predict the structure of the diphenylquinone product, $C_{28}H_{40}O_2$, formed by the oxidative coupling of 2,6-di-*tert*-butylphenol with oxygen in the presence of base.

10W-25. Substituted phenols, such as BHT, are used as antioxidants in processed foods:

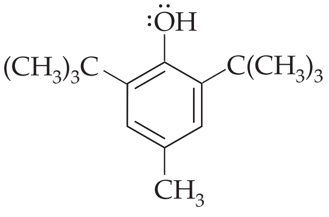

BHT

The role of the antioxidant is to stop spoilage caused by the free radical reactions brought about by reaction of oxygen with compounds containing C═C bonds.

(a) Give a suitable chemical name for BHT.

(b) Can you suggest why this compound is an effective antioxidant?

10W-26. Naphthalene is the simplest of the polycyclic aromatic hydrocarbons and can be represented by three resonance structures. Draw them. Indicate which of the structures are equivalent.

10W-27. Some relatively simple chiral compounds contain a chiral axis. Use molecular models to convince yourself that the two molecules below are enantiomers, even though they contain no stereocenters:

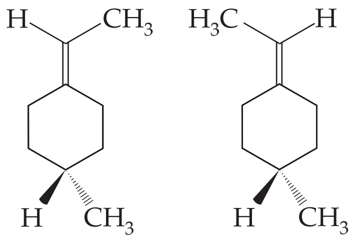

BIBLIOGRAPHY

For references related to the oxidative coupling of phenolic derivatives, see

Dewar, M. J. S.; Nakaya, T. J. *Am. Chem. Soc.* **1968,** 90, 7134.
Scott, A. I. Q. *Rev.* **1965,** *19,* 1.

For an overview of Cram's work on the binaphthyl crown ethers, see

Cram, D. J.; Cram, J. M. *Science* **1974,** 183, 803.

For the resolution of 1,1′-bi-2-naphtol, see

Dongwei, C; Hughes, D. L; Verhoeven, T. R.; Reider, P. J. *Organic Synthesis,* Collect. Vol 10, p. 93. Also see Kazlauskas, P. J. *Organic Synthesis,* Collect. Vol 9, p. 77.

For the use of 1,1′-bi-2-naphthol as a chiral element for asymmetric catalysts, see

Chem. Eng. News: **1995,** 9 Oct, p. 74.
Noyori, R.; Takaya, H. *Acc. Chem. Res.* **1990,** *23,* 345.

Beckmann Rearrangement: Benzanilide

EXPERIMENT
[6 adv]

Common name: benzanilide
CA number: [93-98-1]
CA name as indexed: benzamide, *N*-phenyl-

Purpose. You will carry out the *Beckmann rearrangement* in which a ketone is converted, via an oxime, to an amide. The reaction is an example of a group of reactions in which migration to an electron-deficient nitrogen occurs.

Prior Reading

Technique 4: Solvent Extraction
Liquid–Liquid Extraction (p. 72)
Technique 5: Crystallization
Craig Tube Crystallizations (pp. 89–90)

REACTION

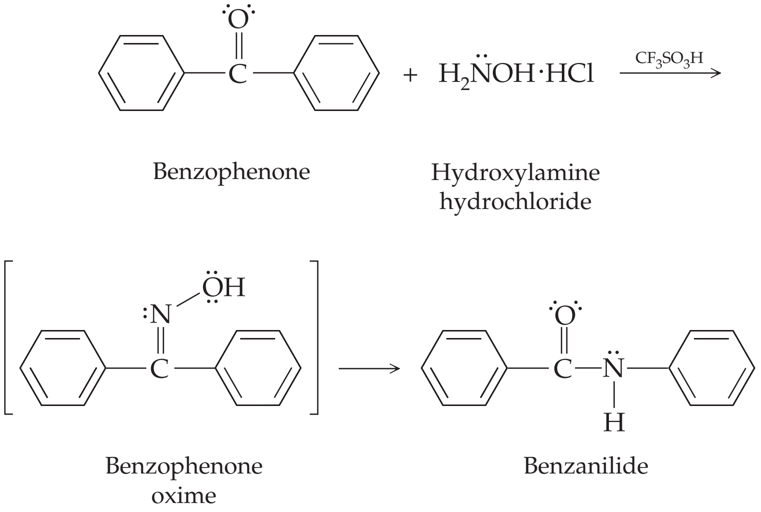

DISCUSSION

The Beckmann rearrangement was discovered in 1886 by E. Beckmann and is the reaction of an oxime of a ketone, in the presence of acid, to yield the corresponding amide or amides. It is a very general reaction and a wide variety of reagents have been used to cause the rearrangement to take place. These include sulfuric, hydrochloric, and polyphosphoric acids, phosphorus pentachloride, and aromatic sulfonyl chlorides. The Beckmann rearrangement is the nitrogen analogue, both functionally and mechanistically, of the Baeyer–Villiger oxidation, which converts a ketone to an ester.

In the present case, only one amide is formed, because benzophenone is a symmetrical ketone. Because the oximes are prepared from ketones, the reaction was often used, before the advent of modern spectroscopic techniques, to determine the structure of the starting ketone. This structure determination was accomplished by the subsequent identification of the acid and amine obtained upon hydrolysis of the amide product of the rearrangement.

As depicted in the following mechanism, the acid catalyst converts the oxime's hydroxyl group to a good leaving group (water, a small neutral molecule). The acid used in this experiment is a very strong acid, trifluoromethanesulfonic acid, usually called *triflic acid*. The first part of the overall reaction in this experiment constitutes the formation of the oxime, which is followed by the actual migration of an aryl (or alkyl) group from the former carbonyl carbon to the nitrogen atom, the Beckmann rearrangement:

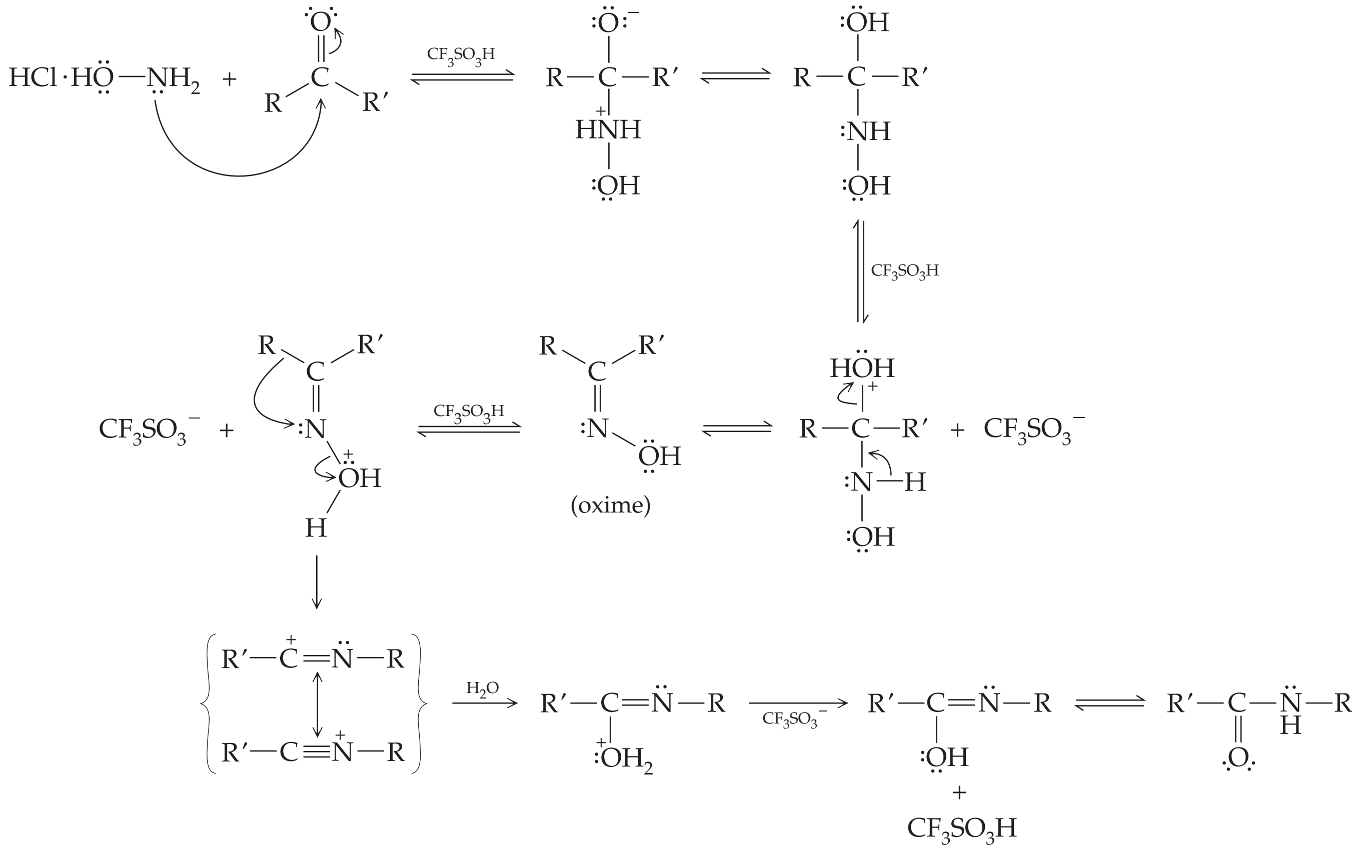

There are two significant points concerning the stereo- and regiochemistry of the reaction: (1) the group that migrates is the one *anti* to the hydroxyl group on the C═N bond and (2), the stereochemistry, if any, of the migrating group is retained.

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 2.5 h.

Physical Properties of Reactants

Compound	MW	Amount	mmol	mp (°C)
Benzophenone	182.21	100 mg	0.55	48.1
Hydroxylamine hydrochloride	69.49	51 mg	0.73	155–57
Triflic acid–formic acid solution		500 μL		

Reagents and Equipment. Weigh and place 100 mg (0.55 mmol) of benzophenone and 51 mg (0.73 mmol) of hydroxylamine hydrochloride in a 5.0-mL conical vial containing a magnetic spin vane. Using an automatic delivery pipet in the **hood,** now add 500 μL of triflic acid–formic acid solution to this mixture. Attach the vial to a re-flux condenser and mount the assembly in a sand bath on a magnetic stirring hot plate (➡).

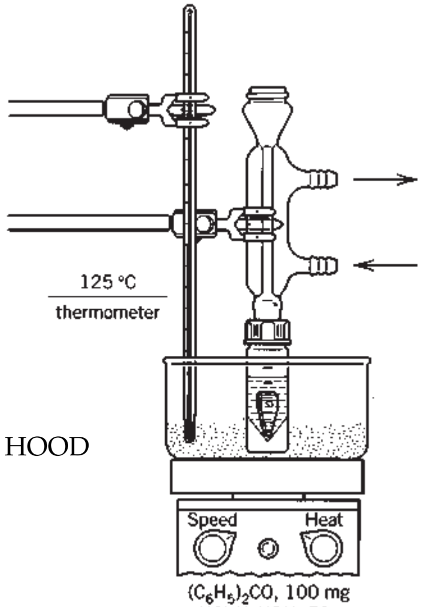

$(C_6H_5)_2CO$, 100 mg
+ $HCl \cdot H_2NOH$, 51 mg
+ $CF_3SO_3H–HCO_2H$, 0.5 mL

CAUTION: Triflic acid (trifluoromethanesulfonic acid) is one of the strongest acids known. It is very corrosive and toxic!

INSTRUCTOR PREPARATION. The acid solution is prepared by adding 2 drops of triflic acid to 5.0 mL of 90% formic acid.

Reaction Conditions. With stirring, heat the reaction mixture at reflux for 1 h in a sand bath at 125 °C. Then cool the resulting solution to room temperature.

Isolation of Product. To the cooled reaction solution add 1.0 mL of water (using a calibrated Pasteur pipet), and extract the resulting mixture with three 1.0-mL portions of methylene chloride. Separate the organic phase using a Pasteur filter pipet and transfer it to a 10-mL Erlenmeyer flask. For each extraction, after addition of the methylene chloride, cap the vial, shake, vent, and then allow the layers to separate.

Dry the combined methylene chloride extracts over granular, anhydrous sodium sulfate. Using a Pasteur filter pipet, transfer the dried solution to a clean, dry 10-mL Erlenmeyer flask containing a boiling stone. Evaporate the solvent in the **hood** under a gentle stream of nitrogen on a warm sand bath to yield the crude reaction product.

HOOD

Characterization and Purification. Transfer the crude benzanilide from the Erlenmeyer flask to a Craig tube, and recrystallize the material from 95% ethanol.

Weigh the product and calculate the percent yield. Determine the melting point and compare your result to that listed in the *CRC Handbook* or found using SciFinder Scholar. Obtain an infrared spectrum.

Infrared Analysis. In this experiment, an aromatic ketone has been rearranged to a secondary amide. By examining the infrared spectra of starting material and product, we can confirm this molecular transformation.

The spectrum of benzophenone (Fig. 10.6W) possesses two macro group frequency trains:

1. **Conjugated aromatic ketone:** This ketone is defined by the bands at 3325 (overtone of ketone carbonyl stretch), 1663 (C═O stretch, doubly conjugated), 1601 and 1580 (ν_{8a} and ν_{8b} degenerate ring stretch), 1492 and 1450 cm^{-1} (ν_{19a} and ν_{19b} degenerate ring stretch). The intensification of the 1580-wavenumber peak confirms the conjugation of the carbonyl to the ring. The 1500-wavenumber ring stretch, which is generally a bit variable in intensity, is quite weak in this case (2 different benzene rings).
2. **Monosubstituted phenyl group:** This group is defined by weak bands located at 1969, 1913, 1823, 1724 cm^{-1}, and strong bands recorded at 701 and 640 cm^{-1}. For discussions of these modes see Chapter 8, and Experiment [20]. In the case of phenyl rings conjugated to carbonyl groups, the 750- and 690-wavenumber bands often appear on the low side of their usual range and can be down as much as 40–50 cm^{-1}, as occurs here.

The rearrangement product has incorporated a heteroatom into its structure, but the carbonyl group and the ring systems have survived. The infrared spectrum of benzanilide (Fig 10.7W) must be consistent with the formation of a secondary amide. We expect to observe a macro group frequency for the presence of monosubstituted phenyl groups, plus a second macro group frequency for a secondary aromatic amide.

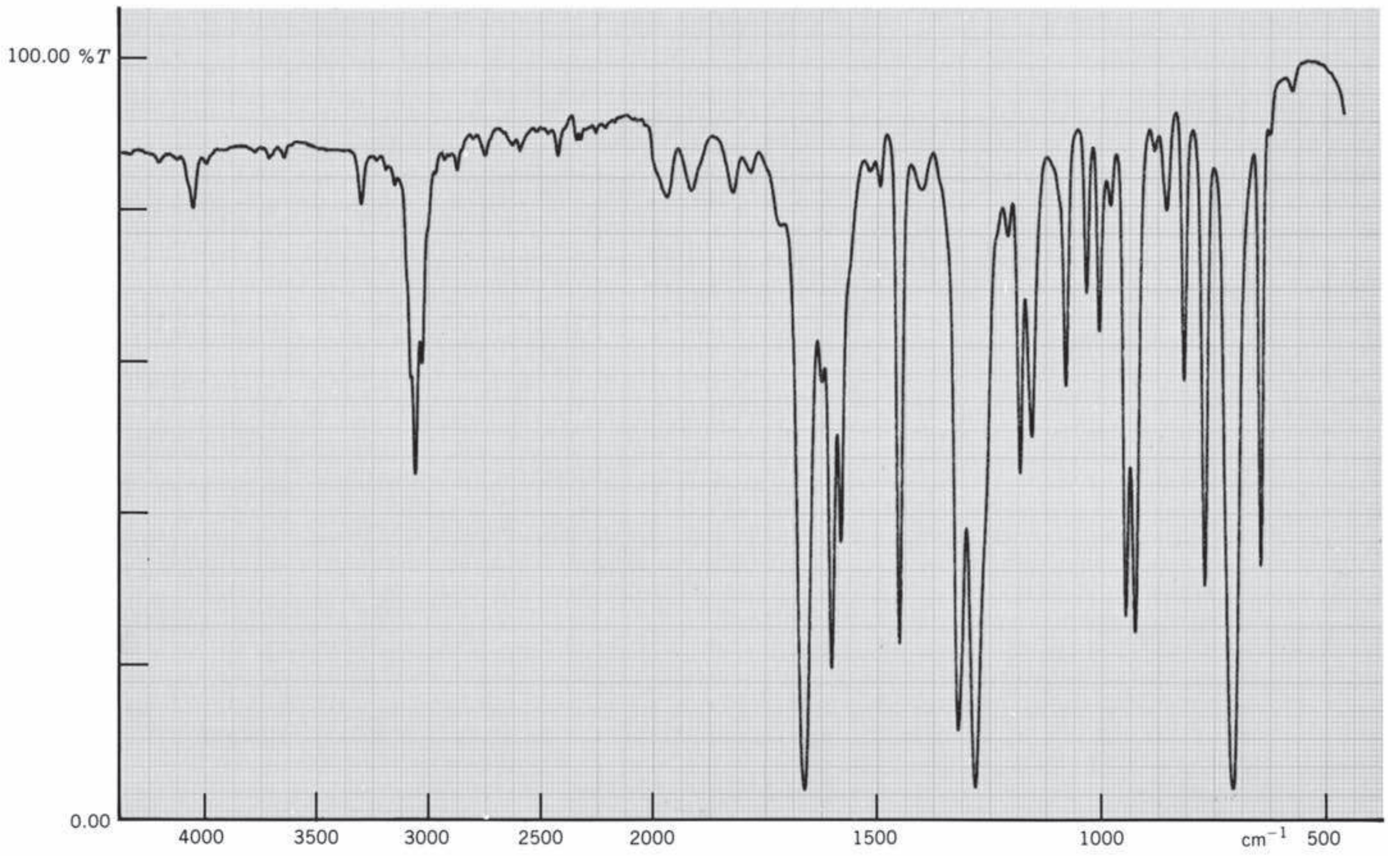

Sample Benzophenone
%T X ABS _ Background Scans 4
Scans 16
Acquisition & Calculation Time 42 sec
Resolution 4.0 cm-1
Sample Condition solid - melt
Cell Window KBr
Cell Path Length capillary film
Matrix Material

Figure 10.6W IR spectrum: benzophenone.

The first macro group frequency is similar to that of the starting material with bands occurring at

a. **1950, 1912, 1840, and 1725 cm^{-1}:** All four combination bands are doubled in this case. The values given are for the centers of the doublets.
b. **752 and 718 cm^{-1}:** C—H out-of-plane bend. The lower value likely corresponds to the ring conjugated directly with the carbonyl group.
c. **693 cm^{-1}:** Ring out-of-plane bend (puckering) vibration.

The second macro group frequency for the aromatic secondary amide uses the following bands: 3350, 3060, 1659, 1602, 1587, 1539, 1322, and 680 (broad) cm^{-1}.

a. **3350 cm^{-1}:** This strong band corresponds to the highly hydrogen-bonded amide N—H stretch.
b. **3060 cm^{-1}:** C—H stretch on sp^2 carbon, aromatic C—H.

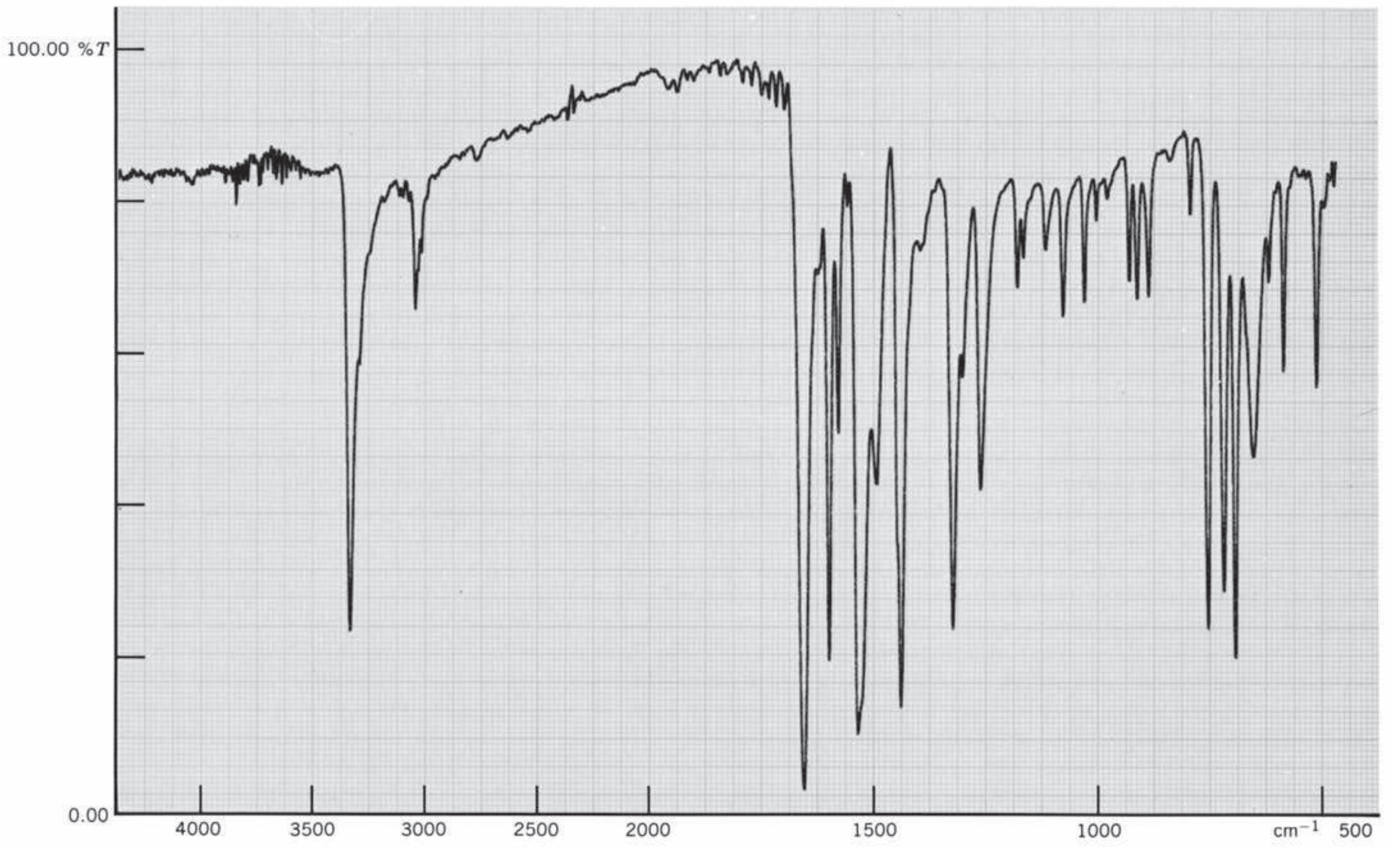

Sample Benzanilide

%T x ABS _ Background Scans 4

Scans 16

Acquisition & Calculation Time 42 sec

Resolution 4.0 cm-1

Sample Condition solid

Cell Window

Cell Path Length

Matrix Material KBr

Figure 10.7W IR spectrum: benzanilide.

- **c. 1659 cm^{-1}:** C═O stretch of a conjugated and hydrogen-bonded amide. It is the most intense band in the spectrum.
- **d. 1602 and 1587 cm^{-1}:** Ring stretch degenerate pair, ν_{8a} and ν_{8b}.
- **e. 1539 and 1322 cm^{-1}:** These two bands involve both N—H bending (in-plane) and C—N stretch. The two fundamentals fall somewhere between 1450–1400 cm^{-1}. Thus, they can mechanically interact and split apart with one component falling at 1539 cm^{-1} and the other near 1322 cm^{-1}. Since the fundamental frequencies will vary somewhat from molecule to molecule, the interaction term that is sensitive to the frequency match (see Infrared discussions, Chapter 8) will also vary in magnitude, and thus the wavenumber separation will be affected.
- **f. 680 cm^{-1}:** Broad and weak. This ill-defined band arises from the out-of-plane bend of the N—H group. It is similar to the O—H bend found in alcohols in this spectral region (see infrared discussion in Experiments [5A], [5B], and [8B]).

Examine the IR spectrum of your reaction product. Discuss the similarities and differences of the experimentally derived spectral data to the reference spectra (Figs. 10.6W and 10.7W).

Chemical Tests. Chemical classification tests (Chapter 9), such as the ignition test and the soda lime test (or sodium fusion test), may also be conducted to further establish the identity of the product. The hydroxamate test may be used to establish the presence of the amide functional group.

QUESTIONS

10W-28. Draw the structure of the product expected in each of the Beckmann rearrangements presented below. Give a suitable name for each product.

(a)

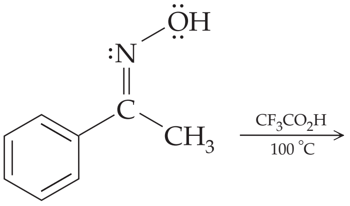

(c)

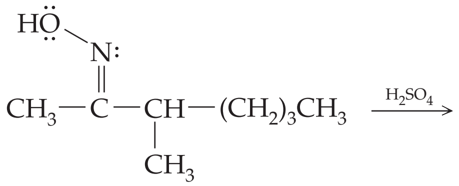

$$CH_3-\overset{\overset{\displaystyle \|}{N-OH}}{C}-\underset{\underset{\displaystyle CH_3}{|}}{CH}-(CH_2)_3CH_3 \xrightarrow{H_2SO_4}$$

(b)

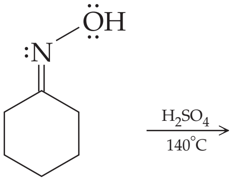

10W-29. The picryl iminoethers (an iminoether is C=N—O—C) of oximes undergo the Beckmann rearrangement without an acid catalyst:

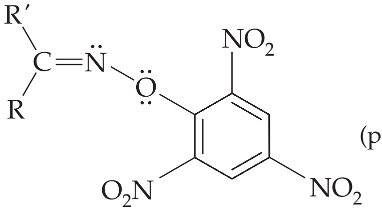

(picryl ester)

Explain why a catalyst is not required in this rearrangement.

10W-30. Compounds A and B undergo the Beckmann rearrangement upon gentle heating of the solid compounds:

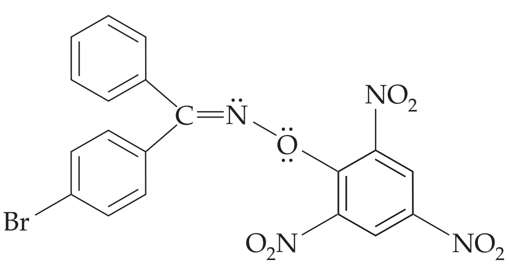

A

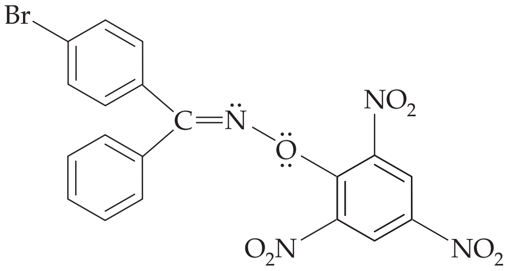

B

(a) Write the structure of the products expected.

(b) If a mixture of the two iminoethers, A and B, are heated in the same reaction flask, what products would be formed? Explain.

10W-31. When acetophenone oxime was allowed to rearrange in ^{18}O-enriched solvent, the amide product contained the same percentage of ^{18}O as the solvent. Explain this observation.

10W-32. Oximes are usually crystalline materials and have been prepared as a means of identifying liquid ketones or aldehydes. It has been found in the preparation of these derivatives that if the acid concentration is too high (low pH), then the oxime does not form. Explain.

10W-33. The mechanical coupling between the N—H bending vibration and the C—N stretching vibration depends on a close frequency match. Normally the C—C, C—N, and C—O stretching vibrations are found in the 1200- to 800-wavenumber region. Why, in the case of the amides, do we find the C—N stretch approaching 1400 cm^{-1}?

BIBLIOGRAPHY

Reviews on the Beckmann rearrangement:

Donaruma, L. G.; Heldt, W. Z. *Org. React.* **1960,** *11,* 1.

Gawley, R. E. *Org. React.* **1988,** *35,* 1.

Jochims, J. C.; Hehl, S.; Herzberger, S. *Synthesis* **1990,** 1128, and references therein.

McCarthy, C. G. in *The Chemistry of the Carbon–Nitrogen Double Bond;* Patai, S., Ed.; Interscience: New York, 1970; p. 408.

Yamabe, S.; Tauchida, N.; Yamazaki, H. *J. Org. Chem.* **2005,** *70,* 10638.

An example of the Beckmann rearrangement using different reagents is given in

Deluca, L.; Giacomelli. G.; Porcheddu, A. *J. Org. Chem.* **2002,** *67,* 6072.

Furuya, Y.; Ishihara, K.; Yanamoto, H. *J. Am. Chem. Soc.* **2005,** *127,* 11240.

Harada, T.; Ohno, T.; Kobayashi, S.; Mukaiyama, T. *Synthesis* **1991,** 1216.

Ohno, M.; Naruse, N.; Terasawa, I. *Organic Syntheses;* Wiley: New York, 1973; Collect. Vol. V, p. 266.

The present reaction was adapted from the work of

Ganboa, I.; Palomo, C. *Synth. Commun.* **1983,** *13,* 941.

EXPERIMENT [7$_{adv}$]

Preparation of an Enol Acetate: Cholesta-3,5-dien-3-ol Acetate

Common name: cholesta-3,5-dien-3-ol acetate
CA number: [2309-32-2]
CA name as indexed: cholesta-3,5-dien-3-ol, acetate

Purpose. This experiment investigates the conditions under which the enol acetate of an α,β-unsaturated ketone is prepared. The combined use of acetic anhydride with chlorotrimethylsilane is an illustration of a technique for the generation of *acylium ions*, a reactive electrophile for acylation reactions.

Prior Reading

Technique 5: Crystallization
- Craig Tube Crystallizations (pp. 89–90)

Technique 6: Chromatography
- Column Chromatography (pp. 92)
- Thin-Layer Chromatography (pp. 97–99)
- Removal of Solvent Under Reduced Pressure (pp. 102–104)

REACTION

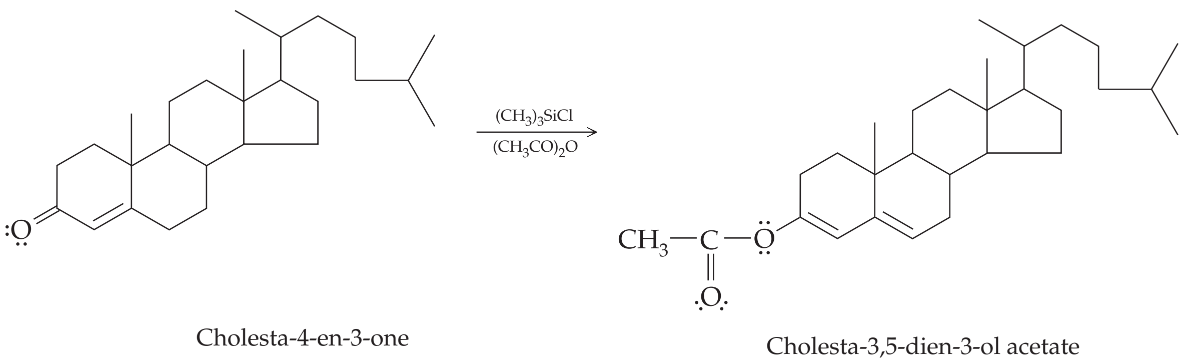
(CH3)3SiCl
(CH3CO)2O
CH3—C—O
O
Cholesta-4-en-3-one
Cholesta-3,5-dien-3-ol acetate

DISCUSSION

Acylium ions are useful reactive intermediates in organic synthesis because their high reactivity allows reactions with relatively weak nucleophiles, in this case the carbonyl oxygen of an α,β-unsaturated enone. The product in this experiment, a dienol acetate of a cholesterol derivative, is itself a useful intermediate in the synthesis of steroids such as cortisone.

Acetic anhydride and chlorotrimethylsilane react, as shown here, to generate a small equilibrium concentration of acylium ion $(CH_3C{\equiv}O^+)$along with trimethylsilyl acetate, $CH_3CO_2Si(CH_3)_3$. The acylium ion is a much more reactive electro philic acyl group than neutral sources of an electrophilic acyl group, such as acetic anhydride or acetyl chloride, CH_3COCl.

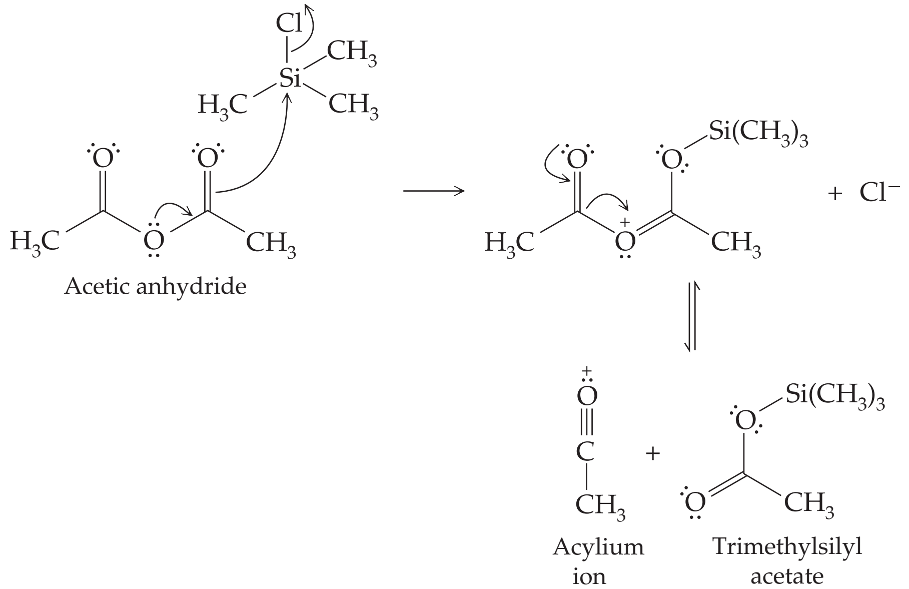
Cl
CH3
Si
H3C
CH3
O
O
H3C
O
CH3
Acetic anhydride
Si(CH3)3
O
O
+
H3C
O
CH3
+ Cl−
+
O
C
CH3
+
Si(CH3)3
O
O
CH3
Acylium ion
Trimethylsilyl acetate

The acylium ion thus generated is a very reactive electrophile, capable of reacting with the relatively weakly nucleophilic lone pairs on the carbonyl oxygen of the cholesta-4-en-3-one. The resulting oxonium ion renders the γ-protons of the enone quite acidic, and thus they are readily removed by a very weak base, such as chloride ion, to generate the product, cholesta-3,5-dien-3-ol acetate. The proposed mechanism for this transformation is shown here:

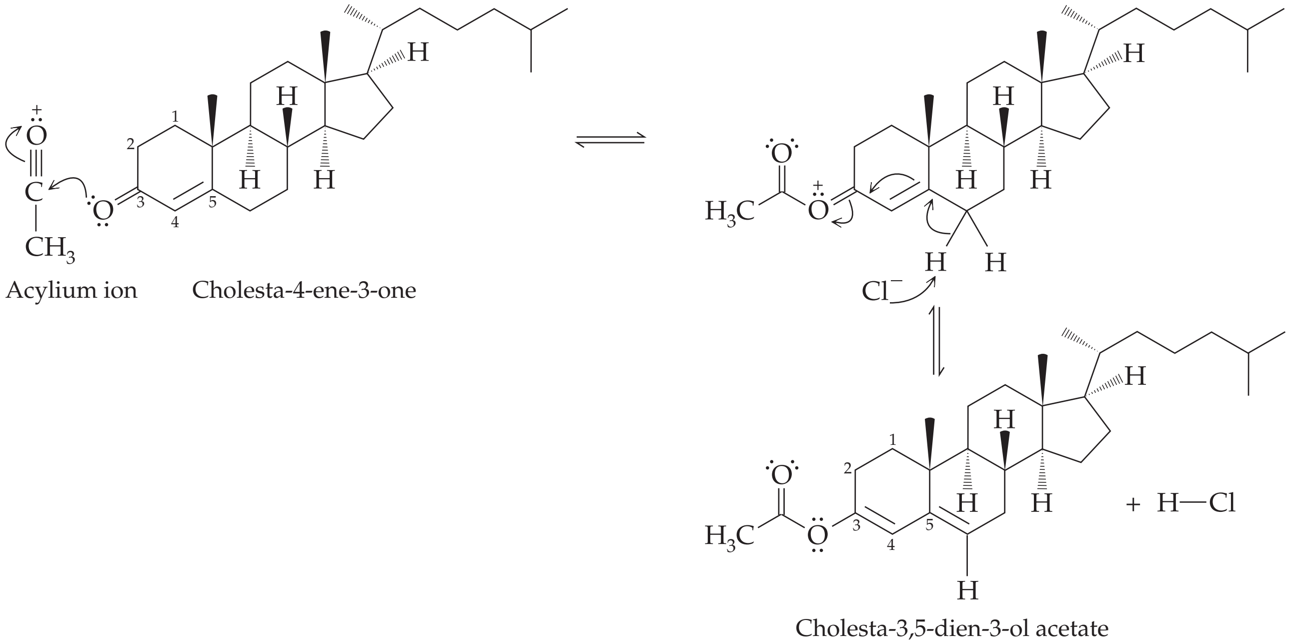

EXPERIMENTAL PROCEDURE

Estimated time to complete the experiment: 4.0 h.

Physical Properties of Reactants and Product							
Compound	**MW**	**Amount**	**mmol**	**mp (°C)**	**bp (°C)**	***d***	n_D
Cholesta-4-en-3-one	384.65	96 mg	0.25	81–82			
Acetic anhydride	102.09	1.0 mL	10.6		140	1.08	1.3901
Chlorotrimethylsilane	108.64	200 μL	1.58		57	0.86	
Cholesta-3,5-dien-3-ol acetate	425.68			79–80			

NOTE. The glass equipment should be dried in an oven (110 °C) and cooled to room temperature in a desiccator before use.

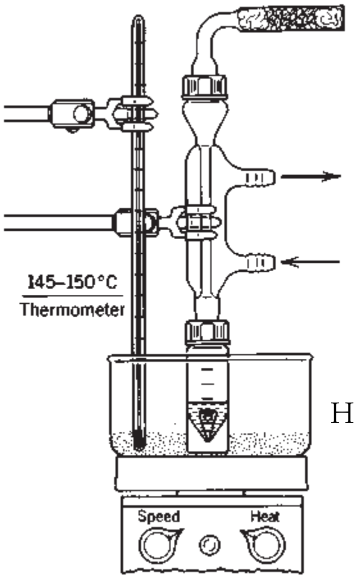

Cholesta-4-ene-3-one, 96 mg + $(CH_3CO)_2O$, 1.0 mL + $(CH_3)_3SiCl$, 200 μL

Reagents and Equipment. Weigh and place 96 mg (0.25 mmol) of cholesta-4-en-3-one in a dry 5.0-mL conical vial containing a magnetic spin vane. Now, in the **hood,** add 1.0 mL of acetic anhydride and 200 μL of chlorotrimethylsilane. Immediately attach the vial to a reflux condenser protected by a calcium chloride drying tube (⬅). HOOD

NOTE. The quality of the acetic anhydride has a significant influence on the reaction. For best results, the anhydride should be distilled and stored over molecular sieves before use. It is convenient to dispense the anhydride and the chlorotrimethylsilane (which hydrolyzes rapidly in moist air) using automatic delivery pipets (in the **hood***). An alternative is to place each reagent in a small bottle sealed with a septum cap. The reagents are then removed through the septum using a 1-mL* HOOD *syringe with a Teflon-tipped plunger while the bottle is connected to another needle providing dry nitrogen. The chlorotrimethylsilane (bp 57 °C) may also be purified by distillation in the* **hood.**

HOOD

Reaction Conditions. Heat the reaction mixture with stirring at reflux for 1–2 h in a sand bath (145–150 °C). Follow the course of the reaction using TLC.

?

TLC DIRECTIONS. Use activated silica gel plates with 1:1 methylene chloride/hexane as the elution solvent. Visualization of the separated components is achieved by placing the plate in a closed jar containing a few crystals of iodine (see Prior Reading for further details).

Isolation of Product. Allow the reaction mixture to cool to room temperature and then place it in an ice bath for 15–30 min. A solid product forms during this time. Collect this solid by vacuum filtration using a Hirsch funnel, and wash the filter cake with 15 mL of 5% aqueous sodium bicarbonate, followed by 5 mL of cold water (➡).

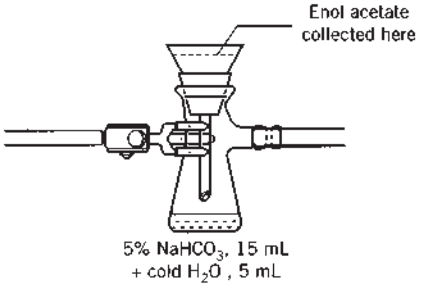

Purification and Characterization. Purify the crude product by column chromatography. In a 1 × 10-cm buret, place 2.0 g of activated silica gel (100 mesh) packed wet (slurry) with methylene chloride (➡). Dissolve the product in about 1.0 mL of hexane and transfer the solution by Pasteur pipet to the column. Elute the material from the column using approximately 10 mL of 1:1 methylene chloride/hexane solvent. Collect the eluate in a tared 25-mL filter flask.

Remove the solvent by warming in a sand bath under reduced pressure to give the solid product, cholesta-3,5-dien-3-ol acetate (⬇). A rotary evaporator, if available, is a more rapid alternative. Recrystallize this material from methanol using a Craig tube, and dry the resulting crystals on a clay plate. Weigh the product and calculate the percent yield. Determine the melting point and compare it with the literature value shown in the Reactant and Product table. Obtain an IR spectrum of the material and compare it with that of an authentic sample, as well as with the spectrum of the starting material.

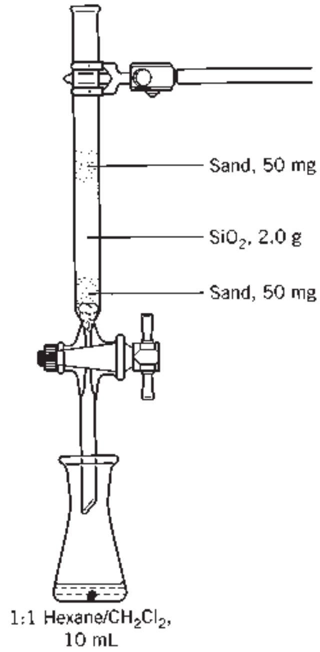

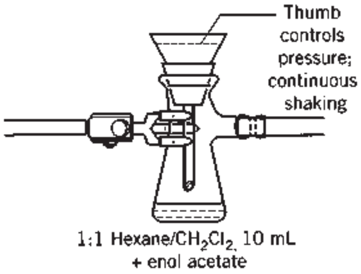
Thumb controls pressure; continuous shaking
1:1 Hexane/CH$_2$Cl$_2$ 10 mL
+ enol acetate

QUESTIONS

10W-34. Reaction of $(CH_3)_3SiCl$ with alcohols produces a trimethylsilyl ether. For example,

$$CH_3CH_2CH_2\ddot{\underset{\cdot\cdot}{O}}H + (CH_3)_3SiCl \xrightarrow[\text{ether}]{(C_2H_5)_3N:} CH_3CH_2CH_2\ddot{\underset{\cdot\cdot}{O}}Si(CH_3)_3 + (C_2H_5)_3\overset{+}{N}H,\ Cl^-$$

The trimethylsilyl ether is more volatile than the corresponding alcohol, and is often used to facilitate GC analysis. Explain.

10W-35. Chlorotrimethylsilane reacts with enolate anions to form stable silyl enol ethers. For example,

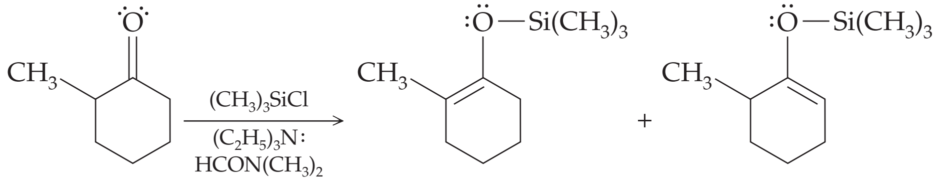

Assuming that the reaction is run under equilibrium conditions, predict which of the above silyl enol ethers is formed in the largest amount and why.

10W-36. In this experiment, the protons at *both* the γ and α′ positions relative to the intermediate oxonium ion are quite acidic. Removal of a proton from the γ position results in the formation (pathway *a* below) of the actual product, cholesta-3,5-dien-3-ol acetate (see Discussion section). Removal of a proton from the α′ position would result in the formation (pathway *b* below) of an isomeric product, cholesta-2,4-dien-3-ol acetate. Because the reaction is conducted under equilibrium conditions, the product obtained must be the thermodynamically more stable of the two possibilities. Why is cholesta-2,4-dien-3-ol acetate less stable than cholesta-3,5-dien-3-ol acetate?

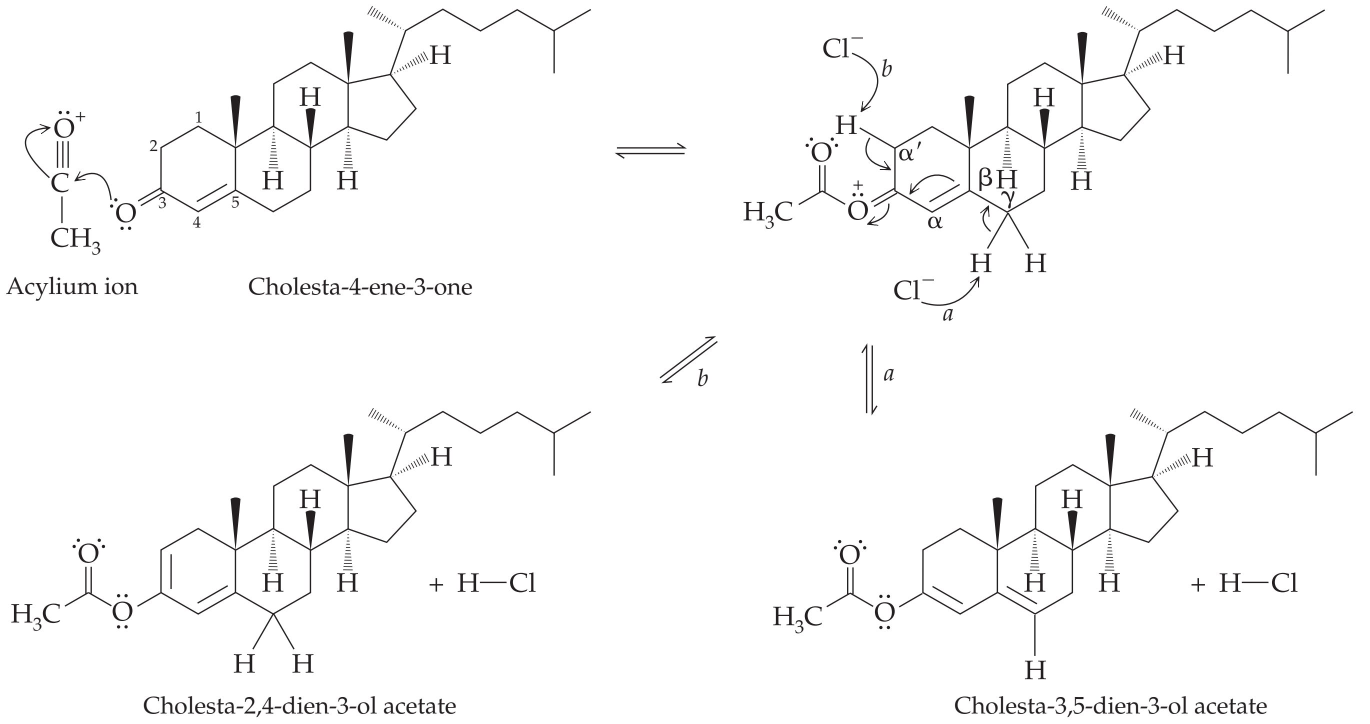

Acylium ion

Cholesta-4-ene-3-one

Cholesta-2,4-dien-3-ol acetate

Cholesta-3,5-dien-3-ol acetate

10W-37. There are two reasons why the enols of β-dicarbonyl compounds, such as pentane-2,4-dione are relatively stable. One is that they are conjugated and the π-electron overlap due to the conjugation provides additional stability. What is the other reason that these species are relatively stable? Would you expect the enol of cyclohexane-1,3-dione to be as stable as the enol of pentane-2,4-dione?

BIBLIOGRAPHY

The present experiment is adapted from the work reported by

Chowdhury, P. K.; Sharma, R. P.; Barua, J. N. *Tetrahedron Lett.* **1983,** *24,* 3383.

For information on the formation and reactivity of enols, see

House, H. O. *Modern Synthetic Reactions,* 2nd ed.; Benjamin: Reading, MA, 1972; Chapter 9, p. 492.

Rappoport, Z., Ed.; *The Chemistry of the Enols;* Wiley: New York, 1990.

Smith, M. B.; March, J. *Advanced Organic Chemistry,* 6th ed.; Wiley: New York, 2007, Chap. 12, p. 800.

GLOSSARY

Absorb To take up matter (to dissolve), or to take up radiant energy.

Activated complex An unstable combination of reacting molecules that is intermediate between reactants and products.

Activation energy The minimum energy, $\Delta G^{\ddagger}$, necessary to form an activated complex in a reaction. Or the difference in energy levels between the ground state and transition state.

Active methylene A methylene group with hydrogen atoms rendered acidic due to the presence of an adjacent (α) electron withdrawing group, such as a carbonyl group.

Activity (of alumina) A measure of the degree to which alumina adsorbs polar molecules.The activity (adsorbtivity) of alumina may be reduced by the addition of small amounts of water. Thus, the amount of water present in a sample of alumina determines the activity grade. Alumina of a specific activity can be prepared by dehydrating alumina at 360 °C for about 5–6 h and then allowing the dehydrated alumina to absorb a suitable amount of water. The Brockmann scale of alumina activity is based on the amount of water (weight percent) that the alumina contains: grade I = 0%, grade II = 3%, grade III = 6%, grade IV = 10%, and grade V = 15%. For further information, see Brockmann, H.; Schodder, H. *Chem Ber.* **1941,** *74,* 73.

Adsorb The process by which molecules or atoms (either gas or liquid) adhere to the surface of a solid.

Aliphatic Term used to refer to nonaromatic species, such as alkanes, alkenes, alkynes.

Aliquot A portion.

Alkaloid A naturally occurring compound that contains a basic amine functional group. They are found particularly in plants.

Anilide A compound that contains a C_6H_5NHCO group. An amide formed by acylation of aniline (aminobenzene).

Bimolecular reaction The collision and combination of two reactants to give an activated complex in a reaction.

Capillary action The action by which the surface of a liquid,where it contacts a solid, is elevated or depressed because of the relative attractions of the molecules of the liquid for each other and for the solid. It is particularly observable in capillary tubes, where it determines the ascent (descent) of the liquid above (below) the level of the liquid in which the capillary tube is immersed.

Catalyst A substance that changes the speed of a chemical reaction without affecting the yield or undergoing permanent chemical change itself.

Characterize To conclusively identify a compound by the measurement of its physical, spectroscopic, and other properties.

Condensation reaction A condensation reaction is an addition reaction that produces water (or another small neutral molecule such as CH_3OH or NH_3) as a byproduct.

Dehydrohalogenation A reaction that involves loss of HX from a halide by treatment with strong base.

Deliquescent Liquefying by the absorption of water from the surrounding atmosphere.

Dihedral angle The angle between two intersecting planes. In organic chemistry the term dihedral angle (or torsional angle) is used to describe the angle between two atoms (or groups) bonded to two adjacent atoms, such as H—C—C—H, and can be determined from a molecular model by looking down the axis of the bond between the two central atoms.

Dipole The separation of charge in a bond or in a molecule with a positively and negatively charged end.

Eluant A mobile phase in chromatography.

Eluate The solution that is eluted from a chromatographic system.

Elute To cause elution.

Elution The flow, in chromatography, of the mobile phase through the stationary phase.

Emulsion A suspension composed of immiscible drops of one liquid in another liquid (e.g., oil and vinegar in salad dressing).

Enol A functional group composed of a hydroxyl group bonded to an alkene.

Enolate The conjugate base of a enol, that is, a negatively charged oxygen atom bonded to an alkene. An enolate results from deprotonation α to a carbonyl group.

Enthalpy change (ΔH) The heat lost or absorbed by a system under constant pressure during a reaction.

Entropy (S) The randomness, or amount of disorder of a system.

Entropy change (ΔS) The change in the amount of disorder.

Filter cake The material that is separated from a liquid, and remains on the filter paper, after a filtration.

Free energy change (ΔG) A predictor of the spontaneity of a chemical reaction at constant temperature. $\Delta G = \Delta H - T\,\Delta S$

Glacial acetic acid Pure acetic acid containing less than 1% water.

Heterocycle A cyclic molecule whose ring contains more than one kind of atom.

Heterolysis Cleavage of a covalent bond in a manner such that both the bond's electrons end up on one of the formerly bonded atoms.

Homogeneous Consisting of a single phase.

Homolysis Cleavage of a covalent bond in a manner such that the bond's electrons are evenly distributed to the formerly bonded atoms.

Hydroboration Addition of borane (BH_3) or an alkyl borane to a multiple bond.

Hydrogenation Addition of hydrogen to a multiple bond.

Hygroscopic Absorbs moisture.

In situ In chemistry, the term usually refers to a reagent or other material generated directly in a reaction vessel and not isolated.

Kinetics Referring to the rate of a reaction.

Lachrymator A material that causes the flow of tears.

Ligroin A solvent composed of a mixture of alkanes.

Mechanism A complete description of how a reaction occurs.

Metabolism The chemical processes performed by a living cellular organism.

Metabolites The compounds consumed and produced by metabolism.

Methine A CH group (with no other hydrogen atoms attached to the carbon atom).

Methylene A CH_2 group (with no other hydrogen atoms attached to the carbon atom).

Mother liquor The residual, and often impure, solution remaining from a crystallization.

Olefin An older term for an alkene.

Optical isomers Enantiomers. Isomers that have a mirror-image relationship.

Order of reaction With respect to one of the reactants, the order of a reaction is equal to the power to which the concentration of that reactant is raised in the rate equation.

Oxonium ion A trivalent oxygen cation with a full octet of electrons (e.g., H_3O^+).

Paraffins An older name for alkanes.

Phase transfer catalysts Agents that cause the transfer of ionic reagents between phases, thus catalyzing reactions.

Plasticizer A substance added to a polymer to make it more flexible or to prevent embrittlement.

Polymer A compound of high molecular mass that is built up of a large number of repeating simple molecules, or monomers.

Racemic Consisting of an equimolar mixture of two enantiomers.

Rate equations Equations giving the relationship between reaction rate and the concentrations of the reactants.

Reaction mechanism The stepwise sequence of elementary reactions in an overall reaction.

Reagent A chemical or solution used in the laboratory to detect, measure, react with, or otherwise examine other chemicals, solutions, or substances.

Reflux The process by which all vapor evaporated or boiled from a vessel is condensed and returned to that vessel.

Rotamers Conformational isomers that can be interconverted by rotation about one or more single bonds (e.g., *gauche* and *anti* butane).

Spontaneous process A physical or chemical change that occurs without the net addition of energy. $\Delta G < 0$ for a spontaneous process.

Sublimation The passing of a solid directly into vapor state without first melting.

Tare A tared container is one whose weight has been measured. The term may also refer to the process of zeroing a balance after a container has been placed on the weighing platform.

Thermodynamics The chemical science that deals with the energy transfers and transformations that accompany chemical and physical changes.

Transition state A combination of reacting molecules that is intermediate between reactants and products.

Triturate To grind to a fine powder. (Or, washing solid organic products in a solvent in which the desired product has little solubility.)

Vapor pressure The pressure exerted by a vapor in equilibrium with a liquid or solid at a given temperature.

Ylide A neutral dipolar molecule in which negative and positive charges are on adjacent atoms.

Zwitterion A neutral molecule containing separated opposite formal charges.

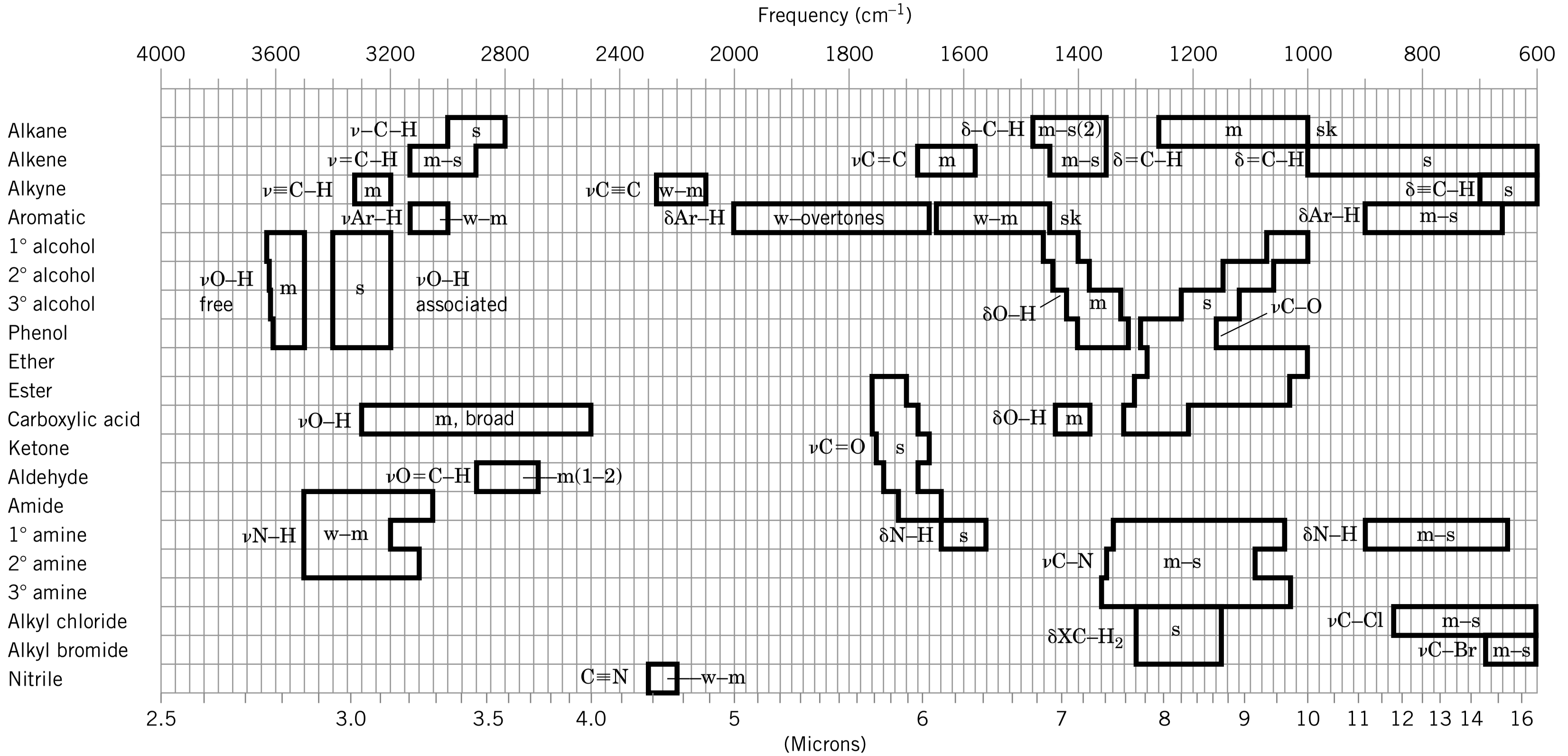

Typical IR absorption frequencies for common functional groups.

Absorptions are as follows: ν = stretching; δ = bending; w = weak; m = medium; s = strong; sk = skeletal

From *Multiscale Organic Chemistry: A Problem-Solving Approach* by John W. Lehman © 2002.

Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.

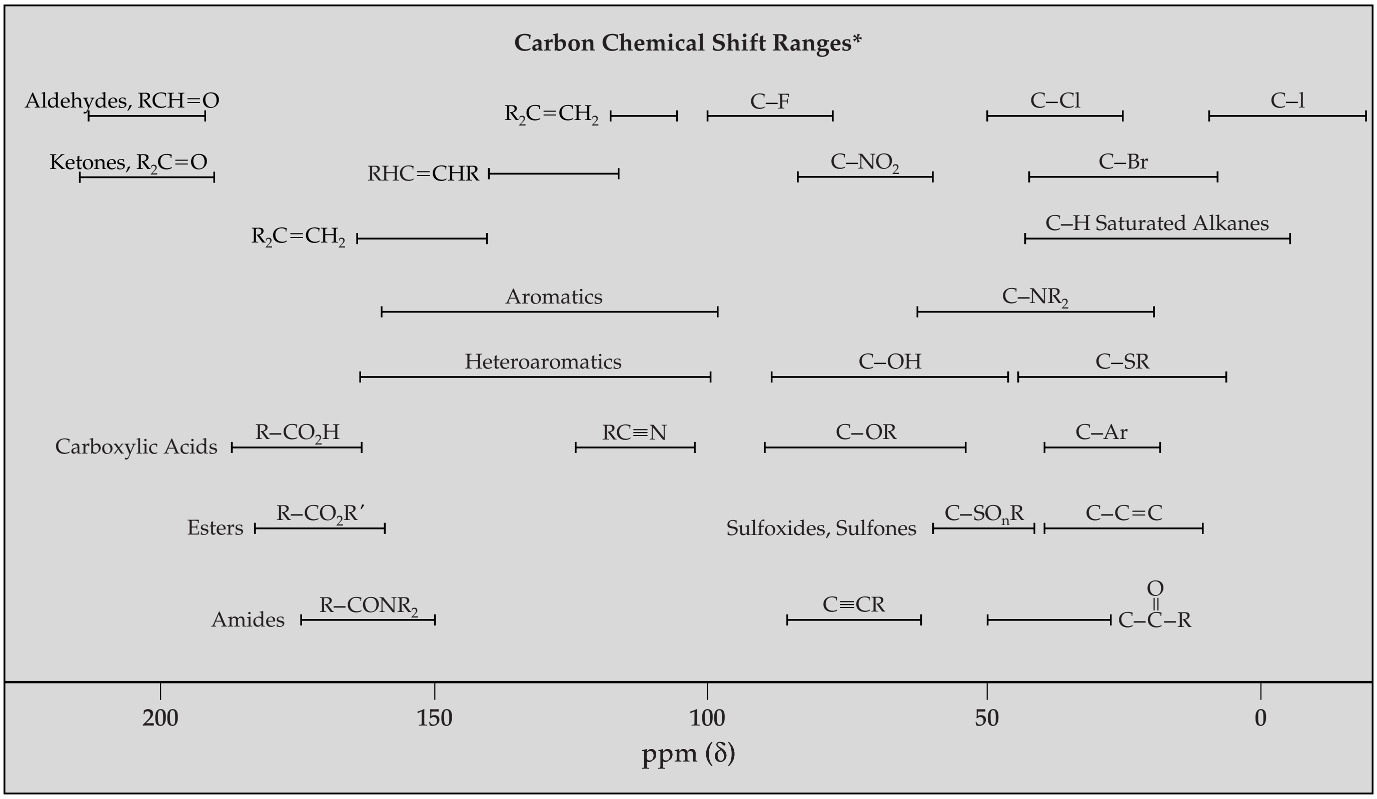
Carbon Chemical Shift Ranges*

*For samples in CDCl$_3$ solution. The δ scale is relative to TMS at δ=0. Organic Chemistry Michigan State University Source, Dept. of Chemistry, Michigan State University.

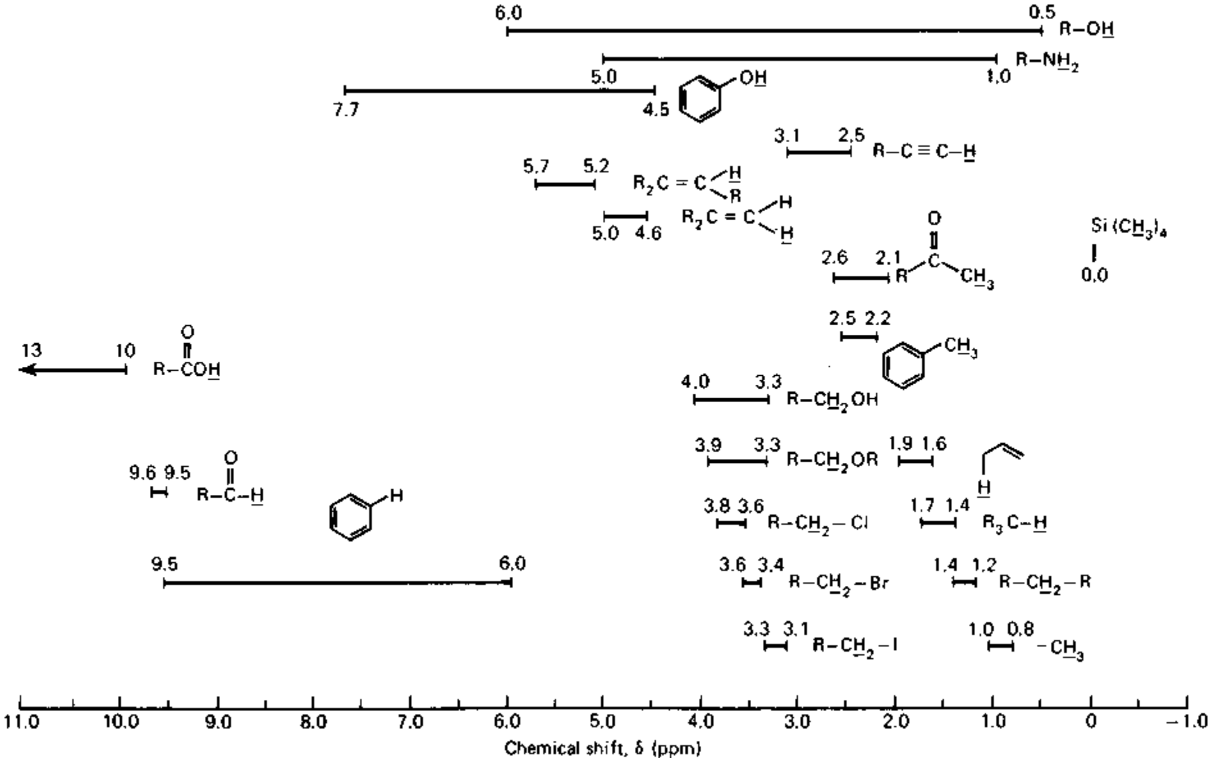

^{1}H NMR Chemical Shift Ranges

PERIODIC TABLE OF THE ELEMENTS

Atomic number→ 6
Symbol → C
Name (IUPAC) → Carbon
Atomic mass → 12.011

IUPAC recommendations→
Chemical Abstracts Service group notation →

1 IA	2 IIA	3 IIIB	4 IVB	5 VB	6 VIB	7 VIIB	8 VIIIB	9 VIIIB	10 VIIIB	11 IB	12 IIB	13 IIIA	14 IVA	15 VA	16 VIA	17 VIIA	18 VIIIA
1 **H** Hydrogen 1.0079																	2 **He** Helium 4.0026
3 **LI** Lithium 6.941	4 **Be** Berylium 9.0122											5 **B** Boron 10.811	6 **C** Carbon 12.011	7 **N** Nitrogen 14.007	8 **O** Oxygen 15.999	9 **F** Fluorine 18.998	10 **Ne** Neon 20.180
11 **Na** Sodium 22,990	12 **Mg** Magnesium 24.305											13 **Al** Aluminum 26.982	14 **Si** Silicon 28.086	15 **P** Phosphorus 30.974	16 **S** Sulfur 32.065	17 **Cl** Chlorine 35.453	18 **Ar** Argon 39.948
19 **K** Potassium 39.098	20 **Ca** Calcium 40.078	21 **Sc** Scandium 44.956	22 **Ti** Titanium 47.867	23 **V** Vanadium 50.942	24 **Cr** Chromium 51.996	25 **Mn** Manganese 54.938	26 **Fe** Iron 55.845	27 **Co** Cobalt 58.933	28 **Ni** Nickel 58.693	29 **Cu** Copper 63.546	30 **Zn** Zinc 65.409	31 **Ga** Gallium 69.723	32 **Ge** Germanium 72.64	33 **As** Arsenic 74.922	34 **Se** Selenium 78.96	35 **Br** Bromine 79.904	36 **Kr** Krypton 83.798
37 **Rb** Rubidium 85.468	38 **Sr** Strontium 87.62	39 **Y** Yttrium 88.906	40 **Zr** Zirconium 91.224	41 **Nb** Niobium 92.906	42 **Mo** Molybdenum 95.94	43 **Tc** Technetium (98)	44 **Ru** Ruthenium 101.07	45 **Rh** Rhodium 102.91	46 **Pd** Palladium 106.42	47 **Ag** Silver 107.87	48 **Cd** Cadmium 112.41	49 **In** Indium 114.82	50 **Sn** Tin 118.71	51 **Sb** Antimony 121.76	52 **Te** Tellurium 127.60	53 **I** Iodine 126.90	54 **Xe** Xeno 131.29
55 **Cs** Caesium 132.91	56 **Ba** Barium 137.33	57 ***La** Lanthanum 138.91	72 **Hf** Hafnium 178.49	73 **Ta** Tantalum 180.95	74 **W** Tungsten 183.84	75 **Re** Rhenium 186.21	76 **Os** Osmium 190.23	77 **Ir** Iridium 192.22	78 **Pt** Platinum 195.08	79 **Au** Gold 196.97	80 **Hg** Mercury 200.59	81 **Tl** Thallium 204.38	82 **Pb** Lead 207.2	83 **Bi** Bismuth 208.98	84 **Po** Polonium (209)	85 **At** Astatine (210)	86 **Rn** Radon (222)
87 **Fr** Francium (223)	88 **Ra** Radium (226)	89 **#Ac** Actinium (227)	104 **Rf** Rutherfordium (261)	105 **Db** Dubnium (262)	106 **Sg** Seaborgium (266)	107 **Bh** Bohrium (264)	108 **Hs** Hassium (277)	109 **Mt** Meitnerium (268)	110 **Uun** (281)	111 **Uuu** (272)	112 **Uub** (285)		114 **Uuq** (289)				

*Lanthanide Series	58 **Ce** Cerium 140.12	59 **Pr** Praseodymium 140.91	60 **Nd** Neodymium 144.24	61 **Pm** Promethium (145)	62 **Sm** Samarium 150.36	63 **Eu** Europium 151.96	64 **Gd** Gadolinium 157.25	65 **Tb** Terbium 158.93	66 **Dy** Dysprosium 162.50	67 **Ho** Holmium 164.93	68 **Er** Erbium 167.26	69 **Tm** Thulium 168.93	70 **Yb** Ytterbium 173.04	71 **Lu** Lutetium 174.97
# Actinide Series	90 **Th** Thorium 232.04	91 **Pa** Protactinium 231.04	92 **U** Uranium 238.03	93 **Np** Neptunium (237)	94 **Pu** Plutonium (244)	95 **Am** Americium (243)	96 **Cm** Curium (247)	97 **Bk** Berkelium (247)	98 **Cf** Californium (251)	99 **Es** Einsteinium (252)	100 **Fm** Fermium (257)	101 **Md** Mendelevium (258)	102 **No** Nobelium (259)	103 **Lr** Lawrencium (262)

Common Organic Solvents: Table of Properties

Solvent	formula	MW	boiling point (°C)	melting point (°C)	density (g/mL)	solubility in water (g/100g)	Dielectric Constant	flash point (°C)
acetic acid	$C_2H_4O_2$	60.05	118	16.6	1.049	Miscible	6.15	39
acetone	C_3H_6O	58.08	56.2	−94.3	0.786	Miscible	20.7(25)	−18
acetonitrile	C_2H_3N	41.05	81.6	−46	0.786	Miscible	37.5	6
benzene	C_6H_6	78.11	80.1	5.5	0.879	0.18	2.28	−11
1-butanol	$C_4H_{10}O$	74.12	117.6	−89.5	0.81	6.3	17.8	35
2-butanol	$C_4H_{10}O$	74.12	98	−115	0.808	15	15.8(25)	26
2-butanone	C_4H_8O	72.11	79.6	−86.3	0.805	25.6	18.5	−7
t-butyl alcohol	$C_4H_{10}O$	74.12	82.2	25.5	0.786	Miscible	12.5	11
carbon tetrachloride	CCl_4	153.82	76.7	−22.4	1.594	0.08	2.24	—
chlorobenzene	C_6H_5Cl	112.56	131.7	−45.6	1.1066	0.05	5.69	29
chloroform	$CHCl_3$	119.38	61.7	−63.7	1.498	0.795	4.81	—
cyclohexane	C_6H_{12}	84.16	80.7	6.6	0.779	<0.1	2.02	−20
1,2-dichloroethane	$C_2H_4Cl_2$	98.96	83.5	−35.3	1.245	0.861	10.42	13
diethyl ether	$C_4H_{10}O$	74.12	34.6	−116.3	0.713	7.5	4.34	−45
diethylene glycol	$C_4H_{10}O_3$	106.12	245	−10	1.118	10	31.7	143
diglyme (diethylene glycol dimethyl ether)	$C_6H_{14}O_3$	134.17	162	−68	0.943	Miscible	7.23	67
1,2-dimethoxy-ethane (glyme, DME)	$C_4H_{10}O_2$	90.12	85	−58	0.868	Miscible	7.2	−6
dimethylether	C_2H_6O	46.07	−22	−138.5	NA	NA	NA	−41
dimethyl-formamide (DMF)	C_3H_7NO	73.09	153	−61	0.944	Miscible	36.7	58
dimethyl sulfoxide (DMSO)	C_2H_6OS	78.13	189	18.4	1.092	25.3	47	95
dioxane	$C_4H_8O_2$	88.11	101.1	11.8	1.033	Miscible	2.21(25)	12
ethanol	C_2H_6O	46.07	78.5	−114.1	0.789	Miscible	24.6	13
ethyl acetate	$C_4H_8O_2$	88.11	77	−83.6	0.895	8.7	6(25)	−4
ethylene glycol	$C_2H_6O_2$	62.07	195	−13	1.115	Miscible	37.7	111
glycerin	$C_3H_8O_3$	92.09	290	17.8	1.261	Miscible	42.5	160
heptane	C_7H_{16}	100.20	98	−90.6	0.684	0.01	1.92	−4
Hexamethylphosphoramide (HMPA)	$C_6H_{18}N_3OP$	179.20	232.5	7.2	1.03	Miscible	31.3	105
Hexamethylphosphorous triamide (HMPT)	$C_6H_{18}N_3P$	163.20	150	−44	0.898	Miscible	??	26
hexane	C_6H_{14}	86.18	69	−95	0.659	0.014	1.89	−22
methanol	CH_4O	32.04	64.6	−98	0.791	Miscible	32.6(25)	12
methyl *t*-butyl ether (MTBE)	$C_5H_{12}O$	88.15	55.2	−109	0.741	5.1	??	−28
methylene chloride	CH_2Cl_2	84.93	39.8	−96.7	1.326	1.32	9.08	1.6
N-methyl-2-pyrrolidinone (NMP)	CH_5H_9NO	99.13	202	−24	1.033	10	32	91
nitromethane	CH_3NO_2	61.04	101.2	−29	1.382	9.50	35.9	35
pentane	C_5H_{12}	72.15	36.1	−129.7	0.626	0.04	1.84	−49
Petroleum ether (ligroine)	—	—	30–60	−40	0.656	—	—	−30
1-propanol	C_3H_8O	88.15	97	−126	0.803	Miscible	20.1(25)	15
2-propanol	C_3H_8O	88.15	82.4	−88.5	0.785	Miscible	18.3(25)	12
pyridine	C_5H_5N	79.10	115.2	−41.6	0.982	Miscible	12.3(25)	17
tetrahydrofuran (THF)	C_4H_8O	72.11	66	−108.4	0.886	30	7.6	−21
toluene	C_7H_8	92.14	110.6	−93	0.867	0.05	2.38(25)	4
triethyl amine	$C_6H_{15}N$	101.19	88.9	−114.7	0.728	0.02	2.4	−11
water	H_2O	18.02	100.00	0.00	0.998	—	78.54	—
water, heavy	D_2O	20.03	101.3	4	1.107	Miscible	??	—
o-xylene	C_8H_{10}	106.17	144	−25.2	0.897	Insoluble	2.57	32
m-xylene	C_8H_{10}	106.17	139.1	−47.8	0.868	Insoluble	2.37	27
p-xylene	C_8H_{10}	106.17	138.4	13.3	0.861	Insoluble	2.27	27

T = 20 °C unless specified otherwise.
Source: http://virtual.yosemite.cc.ca.us/smurov/orgsoltab.htm

SECOND EDITION

BIOCHEMISTRY LABORATORY

MODERN THEORY AND TECHNIQUES

RODNEY BOYER

TABLE OF CONTENTS

Preface xiii
Acknowledgments xvii
About the Author xix

Chapter 1 INTRODUCTION TO THE BIOCHEMISTRY LABORATORY 1

A. Safety in the Laboratory 2
Safety First 2
Material Safety Data Sheets 2
Safe Practices in the Biochemistry Laboratory 3

B. Keeping Records and Communicating Experimental Results 6
The Laboratory Notebook 6
Details of the Experimental Write-Up 7
Communicating Results from Biochemistry Research 9

C. Using Biochemical Reagents and Solutions 14
Water Purity 14
Cleaning Laboratory Glassware 15
Solutions: Concentrations and Calculations 15
Preparing and Storing Solutions 17

D. Quantitative Transfer of Liquids 18
Pipets and Pipetting 18
Automatic Pipetting Devices 21

E. Statistical Analysis of Experimental Data 23
Defining Statistical Analysis 23
The Mean, Sample Deviation, and Standard Deviation 24
Spreadsheet Statistics 28
Statistical Analysis in Practice 28
Study Problems 30 • Further Reading 32

Chapter 2 USING THE COMPUTER AND INTERNET FOR RESEARCH IN BIOCHEMISTRY 35

A. What Is Research and How Is It Done in Biochemistry? 35
What Is Research? 35
The Scientific Method 36

B. Using Computers in Biochemistry 38

Accessing the Internet 39
The World Wide Web 40
C. Web Sites Useful in Biochemistry 40
Directories, Library Resources, Databases, and Tools 40
Viewing Structures of Biomolecules 43
Searching the Biochemical Literature 44
Literature Searches on the Web 45
Sequence Homology in Proteins 47
Virtual Biochemistry Laboratories 47
Study Problems 48 • Further Reading 49 • Computer Glossary 50

Chapter 3 GENERAL LABORATORY PROCEDURES 53
A. pH, Buffers, Electrodes, and Biosensors 53
Measurement of pH 54
Using the pH Electrode 54
Biochemical Buffers 56
Selection of a Biochemical Buffer 57
Buffer Dilutions 63
The Oxygen Electrode 64
Biosensors 66
B. Measurement of Protein Solutions 67
The Biuret and Lowry Assays 67
The Bradford Assay 69
The BCA Assay 70
The Spectrophotometric Assay 70
C. Measurement of Nucleic Acid Solutions 71
The Spectrophotometric Assay 71
Other Assays for Nucleic Acids 72
D. Techniques for Sample Preparation 73
Dialysis 73
Ultrafiltration 74
Lyophilization and Centrifugal Vacuum Concentration 77
E. Radioisotopes in Biochemistry 80
Origin and Properties of Radioactivity 80
Detection and Measurement of Radioactivity 85
Radioisotopes and Safety 90
Study Problems 91 • Further Reading 92

Chapter 4 **CENTRIFUGATION TECHNIQUES IN BIOCHEMISTRY 95**

A. Basic Principles of Centrifugation 96

B. Instrumentation for Centrifugation 99
  Low-Speed Centrifuges 99
  High-Speed Centrifuges 101
  Ultracentrifuges 105

C. Applications for Centrifugation 106
  Preparative Techniques 106
  Analytical Measurements 108
  Care of Centrifuges and Rotors 112
    Study Problems 113 • Further Reading 114

Chapter 5 **PURIFICATION AND ANALYSIS OF BIOMOLECULES BY CHROMATOGRAPHY 115**

A. Introduction to Chromatography 116
  Partition versus Adsorption Chromatography 117

B. Planar Chromatography (Paper and Thin-Layer Chromatography) 118
  Preparation of the Sorbent 118
  Solvent Development 119
  Detection and Measurement of Components 120
  Applications of Planar Chromatography 121
  Advanced Planar Chromatography 121

C. Column Chromatography 122
  Operation of a Chromatographic Column 123
  Packing the Column 124
  Loading the Column 125
  Eluting the Column 125
  Collecting the Eluent 126
  Detection of Eluting Components 126

D. Ion-Exchange Chromatography 126
  Ion-Exchange Resins 127
  Selection of the Ion Exchanger 128
  Choice of Buffer 130
  Preparation of the Ion Exchanger 130
  Using the Ion-Exchange Resin 130
  Storage of Resins 131

E. Gel-Exclusion Chromatography 132
Theory of Gel Filtration 132
Physical Characterization of Gel Chromatography 133
Chemical Properties of Gels 133
Selecting a Gel 135
Gel Preparation and Storage 136
Operation of a Gel Column 136
Applications of Gel-Exclusion Chromatography 138
F. High-Performance Liquid Chromatography (HPLC) 140
Instrumentation 142
Stationary Phases in HPLC 144
Chiral Chromatography 148
The Mobile Phase 150
Sample Preparation and Selection of HPLC Operating Conditions 150
FPLC—A Modification of HPLC 150
Perfusion Chromatography 151
G. Affinity Chromatography and Immunoadsorption 152
Chromatographic Media 153
The Immobilized Ligand 154
Attachment of Ligand to Matrix 154
Immunoadsorption 156
Experimental Procedure for Affinity Chromatography 157
H. Membrane-Based Chromatography 159
Study Problems 161 • Further Reading 163

Chapter 6 CHARACTERIZATION OF PROTEINS AND NUCLEIC ACIDS BY ELECTROPHORESIS 165
A. The Theory of Electrophoresis 166
Introduction 166
Theory and Practice 166
B. Methods of Electrophoresis 167
Polyacrylamide Gel Electrophoresis (PAGE) 167
Discontinuous Gel Electrophoresis 172
Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (SDS-PAGE) 174
Nucleic Acid Sequencing Gels 176

Agarose Gel Electrophoresis 177
Pulsed Field Gel Electrophoresis (PFGE) 180
Isoelectric Focusing of Proteins 182
Two-Dimensional Electrophoresis (2-DE) of Proteins 184
Capillary Electrophoresis (CE) 185
Immunoelectrophoresis (IE) 186

C. Practical Aspects of Electrophoresis 188
Instrumentation 188
Reagents 189
Staining and Detecting Electrophoresis Bands 189
Protein and Nucleic Acid Blotting 192
The Western Blot 194
Analysis of Electrophoresis Results 197
Study Problems 198 • Further Reading 199

Chapter 7 SPECTROSCOPIC ANALYSIS OF BIOMOLECULES 201

A. Ultraviolet-Visible Absorption Spectrometry 202
Wavelength and Energy 202
Light Absorption 204
Electronic Transitions in Biomolecules 205
The Absorption Spectrum 207
The Beer-Lambert Law 207
Instrumentation 209
Applications of UV-VIS Spectroscopy 212

B. Fluorescence Spectrometry 220
Principles 220
Quantum Yield 221
Instrumentation 222
Applications of Fluorescence Spectroscopy 223
Difficulties in Fluorescence Measurements 224

C. Nuclear Magnetic Resonance Spectroscopy 225
NMR Theory 226
NMR in Biochemistry 226
NMR and Protein Structures 227

D. Mass Spectrometry 230
Ionization and Analysis of Proteins 230
MS Applications in Biochemistry 232

E. X-Ray Crystallography 233

Methodology of X-ray Crystallography 233

Study Problems 234 • Further Reading 235

Chapter 8 BIOMOLECULAR INTERACTIONS: LIGAND BINDING AND ENZYME REACTIONS 239

A. Ligand-Macromolecule Interactions (Molecular Recognition) 239

Properties of Noncovalent Binding Interactions 240

Quantitative Characterization of Ligand Binding 242

Scatchard's Equation 244

Cooperative Binding of Ligands 245

Experimental Measurement of Ligand-Binding Interactions 245

The Bradford Protein Assay as an Example of Ligand Binding 247

Computer Software for Analysis of LM Binding 249

B. Biological Catalysis (Enzymes) 250

Classes of Enzymes 250

Kinetic Properties of Enzymes 252

Significance of Kinetic Constants 254

Inhibition of Enzyme Activity 255

Units of Enzyme Activity 256

Specific Activity 258

Design of an Enzyme Assay 258

Kinetic versus Fixed-time Assay 259

Applications of an Enzyme Assay 260

Computer Software for Analysis of Enzyme Kinetic Data 262

Study Problems 262 • Further Reading 264

Chapter 9 MOLECULAR BIOLOGY I: STRUCTURES AND ANALYSIS OF NUCLEIC ACIDS 267

A. Introduction to the Nucleic Acids 268

Chemical Components of DNA and RNA 268

DNA Structure and Function 270

RNA Structure and Function 272

B. Laboratory Methods for Investigation of DNA and RNA 275

Isolation of Chromosomal DNA 275

Isolation of Plasmid DNA 277

Characterization of DNA 279
Ethidium Bromide Binding and Fluorescence 280
Agarose Gel Electrophoresis 282
Sequencing DNA Molecules 282
Isolation and Characterization of RNA 284
Study Problems 285 • Further Reading 286

Chapter 10 MOLECULAR BIOLOGY II: RECOMBINANT DNA, MOLECULAR CLONING, AND ENZYMOLOGY 289
A. Recombinant DNA Biotechnology 290
Molecular Cloning 290
Steps for Preparing Recombinant DNA 292
Cloning Vectors 294
B. Important Enzymes in Molecular Biology and Biotechnology 297
The Restriction Endonucleases 297
Applications of Restriction Enzymes 298
Practical Aspects of Restriction Enzyme Use 299
The Polymerase Chain Reaction 301
C. Nucleic Acid Blotting 304
Study Problems 304 • Further Reading 305

Chapter 11 PROTEIN PRODUCTION, PURIFICATION, AND CHARACTERIZATION 307
A. Procedures for the Purification of Proteins 308
Composition of Proteins 308
Amount of Protein versus the Purity of Protein versus Expense 308
Basic Steps in Protein Purification 309
Preparation of the Crude Extract 311
Stabilization of Proteins in a Crude Extract 312
Separation of Proteins Based on Solubility Differences 315
Selective Techniques in Protein Purification 316
B. Production of Proteins by Expression of Foreign Genes 317
Gene Expression in Prokaryotic Organisms 317
Gene Expression in Eukaryotic Cells 320
C. Protein Characterization 322
D. Determination of Primary Structure 323

Amino Acid Composition 323
Sequencing DNA Instead of the Protein 328
Study Problems 328 • Further Reading 328
Appendix I List of Software Programs and Web Sites Useful for Each Chapter 331
Appendix II Properties of Common Acids and Bases 334
Appendix III Properties of Common Buffer Compounds 335
Appendix IV pK_a Values and pH_I Values of Amino Acids 337
Appendix V Molecular Weight of Some Common Proteins 338
Appendix VI Common Abbreviations Used in This Text 339
Appendix VII Units of Measurement 342
Appendix VIII Table of the Elements 344
Appendix IX Answers to Odd-Numbered Study Problems 348

Index 353

CHAPTER

1 INTRODUCTION TO THE BIOCHEMISTRY LABORATORY

A. Safety in the Laboratory 2
B. Keeping Records and Communicating Experimental Results 6
C. Using Biochemical Reagents and Solutions 14
D. Quantitative Transfer of Liquids 18
E. Statistical Analysis of Experimental Data 23

Study Problems 30
Further Reading 32

Welcome to the biochemistry laboratory! You are reading this book for one of the following reasons: (1) you are enrolled in a formal biochemistry lab course at a college or university and you will use the book as a guide to procedures; or (2) you have started a research project in biochemistry and desire an understanding of the theories and techniques you will use in the lab; or (3) you have started a job in a biochemistry lab and wish to review theory and techniques. Whether you are a novice or experienced in biochemistry, I believe you will find the subject matter and lab work to be exciting and dynamic. Most of the experimental techniques and skills that you have acquired and mastered in other laboratory courses will be of great value in your work. However, you will be introduced to many new concepts, procedures, and instruments that you have not used in chemistry or biology labs. Your success in biochemistry lab activities will depend on your mastery of these specialized techniques, use of equipment, and understanding of chemical/ biochemical principles.

For many students, the primary reason to enroll in a formal biochemistry lab course is to learn how to do **research** in the discipline. Most students have observed graduate and undergraduate students participating in research activities at their institutions, but observers may not be entirely familiar with the step-by-step process. The questions “What is research?” and “How is it done in biochemistry?” will be answered in Chapter 2, Section A, p. 35. For now, we will define research simply as “hunting for the truth.”

As you proceed through this text, you will no doubt compare your activities with previous laboratory experiences. In organic lab, you ran reactions, isolated and purified several hundred milligrams or a few grams of solid or liquid products, and characterized the materials by infrared spectroscopy, gas chromatography, nuclear magnetic resonance spectroscopy, mass spectrometry, and other techniques. In biochemistry lab, you will work with milligram or even microgram quantities, and in most cases the biomolecules will be extracted from biological sources and dissolved in solution, so you will never really “see” the materials under study. But you will observe the dynamic chemical and biological changes brought about by biomolecules. The procedures and instruments introduced in the lab will be your “eyes” and will monitor the occurrence of biochemical events.

This chapter is an introduction to procedures that are of utmost importance for the safe and successful completion of a biochemical project. It is recommended that you become familiar with the following sections before you begin laboratory work.

A. SAFETY IN THE LABORATORY

Safety First

The concern for laboratory safety can never be overemphasized. Most students who are involved in biochemistry laboratory activities have progressed through several years of college lab work without even a minor accident. This record is, indeed, something to be proud of; however, it should not lead to overconfidence. You must always be aware that chemicals used in the laboratory are potentially toxic, irritating, and flammable. However, such chemicals are a hazard only when they are mishandled or improperly disposed of. It is my experience as a lab instructor for 30 years that accidents happen least often to those who come to each lab session mentally prepared and with a complete understanding of the experimental procedures to be followed. Because dangerous situations can develop unexpectedly, though, you must be familiar with general safety practices, facilities, and emergency actions. When we work in the lab, we must also have a special concern for the safety of lab mates. Carelessness on the part of one person can often cause injury to others.

Material Safety Data Sheets

The procedures in this book are designed and described with an emphasis on safety. However, no amount of planning or pretesting of procedures substitutes for awareness and common sense on the part of the student. All chemicals used in the procedures outlined here must be handled with care and respect. The use of chemicals in all U.S. workplaces, including academic research and teaching labs, is regulated by the Federal Hazard Communication Standard, a document written by the Occupational Safety and Health Administration (OSHA).[1] Specifically, the OSHA standard requires all workplaces where chemicals are used to do the following: (1) develop a written hazard communication program; (2) maintain files of **Material Safety Data Sheets (MSDS)** on all chemicals used in that workplace (an MSDS is a detailed description of the properties of a chemical substance, the potential health hazards, and the safety precautions that must be taken when handling it); (3) label all chemicals with information regarding hazardous properties and procedures for handling; and (4) train employees in the proper use of these chemicals. Several states have passed "right-to-know" legislation that amends and expands the federal OSHA standard. If you have an interest in or concern about any chemical used in the laboratory, the MSDS may be obtained from your instructor or laboratory director or from the Internet (www.sigma-aldrich.com, for example). The actual form of an MSDS for a chemical may vary, but certain specific information must be present. Figure 1.1 is a partial copy of the MSDS for glacial acetic acid, a reagent often used in biochemical research. All chemical reagent bottles in a workplace, lab, or stockroom must be labeled to identify potential hazardous materials and to specify personal protection necessary for handling. One standard hazard communication system used for this purpose is the Hazardous Materials Identification System (HMIS® III Wall Chart containing an HMIS® III Label in the lower left-hand corner is shown in Figure 1.2). The health, flammability, physical hazard, and personal protection codes for the chemical reagent are summarized on the bottle label for quick identification.

Safe Practices in the Biochemistry Laboratory

It is easy to overlook some of the potential hazards of working in a biochemistry laboratory. Students often have the impression that they are working less with chemicals and more with natural biomolecules; therefore, there is less need for caution. However, this is not true; many reagents used are flammable and toxic. In addition, materials such as fragile glass (disposable pipets), sharp objects (needles, razor blades, etc.), and potentially infectious biological materials (blood, bacteria, viruses) must be used and disposed of with caution. The extensive use of electrical equipment, including hot plates, stirring motors, and high-voltage power supplies for electrophoresis, presents special hazards.

Proper disposal of all waste chemicals, sharp objects, and infectious agents is essential not only to maintain safe laboratory working conditions, but also to protect the general public and your local environment. Some of the liquid chemical reagents and reaction mixtures from experiments are relatively safe and may be disposed of in the laboratory drainage system without causing environmental damage. However, special procedures must be followed for the use and disposal of most reagents and materials. Often this means that your instructor will provide detailed information on proper use procedures. In some cases, proper

---

[1] *Federal Register,* Vol. 48, Nov. 25, 1983, p. 53280; Vol. 50, Nov. 27, 1985, p. 48758.

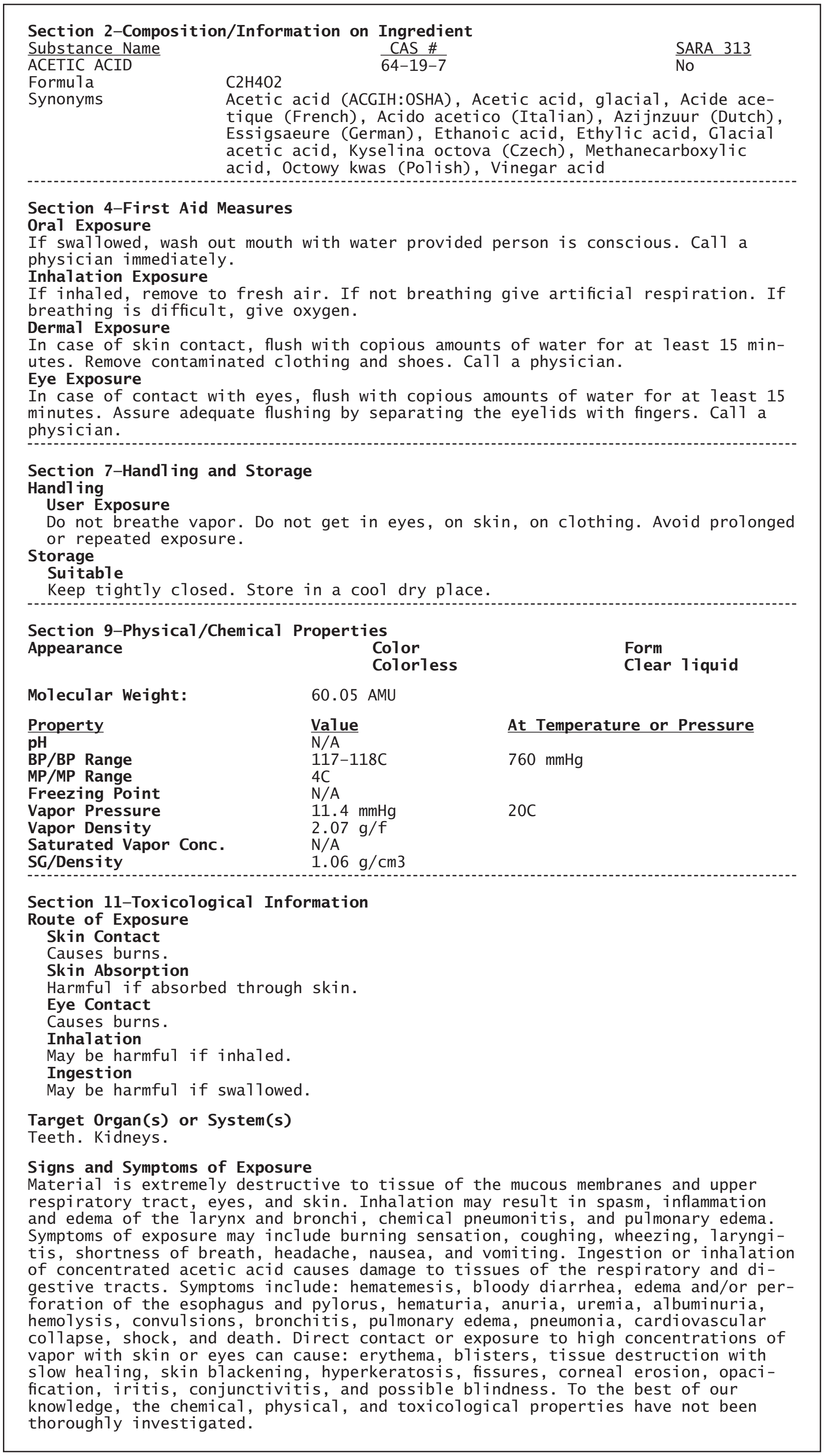

Section 2–Composition/Information on Ingredient

Substance Name	CAS #	SARA 313
ACETIC ACID	64-19-7	No

Formula	C2H4O2
Synonyms	Acetic acid (ACGIH:OSHA), Acetic acid, glacial, Acide acetique (French), Acido acetico (Italian), Azijnzuur (Dutch), Essigsaeure (German), Ethanoic acid, Ethylic acid, Glacial acetic acid, Kyselina octova (Czech), Methanecarboxylic acid, Octowy kwas (Polish), Vinegar acid

Section 4–First Aid Measures

Oral Exposure

If swallowed, wash out mouth with water provided person is conscious. Call a physician immediately.

Inhalation Exposure

If inhaled, remove to fresh air. If not breathing give artificial respiration. If breathing is difficult, give oxygen.

Dermal Exposure

In case of skin contact, flush with copious amounts of water for at least 15 minutes. Remove contaminated clothing and shoes. Call a physician.

Eye Exposure

In case of contact with eyes, flush with copious amounts of water for at least 15 minutes. Assure adequate flushing by separating the eyelids with fingers. Call a physician.

Section 7–Handling and Storage

Handling

User Exposure

Do not breathe vapor. Do not get in eyes, on skin, on clothing. Avoid prolonged or repeated exposure.

Storage

Suitable

Keep tightly closed. Store in a cool dry place.

Section 9–Physical/Chemical Properties

Appearance	Color	Form
	Colorless	**Clear liquid**

Molecular Weight: 60.05 AMU

Property	Value	At Temperature or Pressure
pH	N/A	
BP/BP Range	117–118C	760 mmHg
MP/MP Range	4C	
Freezing Point	N/A	
Vapor Pressure	11.4 mmHg	20C
Vapor Density	2.07 g/f	
Saturated Vapor Conc.	N/A	
SG/Density	1.06 g/cm3	

Section 11–Toxicological Information

Route of Exposure

Skin Contact

Causes burns.

Skin Absorption

Harmful if absorbed through skin.

Eye Contact

Causes burns.

Inhalation

May be harmful if inhaled.

Ingestion

May be harmful if swallowed.

Target Organ(s) or System(s)

Teeth. Kidneys.

Signs and Symptoms of Exposure

Material is extremely destructive to tissue of the mucous membranes and upper respiratory tract, eyes, and skin. Inhalation may result in spasm, inflammation and edema of the larynx and bronchi, chemical pneumonitis, and pulmonary edema. Symptoms of exposure may include burning sensation, coughing, wheezing, laryngitis, shortness of breath, headache, nausea, and vomiting. Ingestion or inhalation of concentrated acetic acid causes damage to tissues of the respiratory and digestive tracts. Symptoms include: hematemesis, bloody diarrhea, edema and/or perforation of the esophagus and pylorus, hematuria, anuria, uremia, albuminuria, hemolysis, convulsions, bronchitis, pulmonary edema, pneumonia, cardiovascular collapse, shock, and death. Direct contact or exposure to high concentrations of vapor with skin or eyes can cause: erythema, blisters, tissue destruction with slow healing, skin blackening, hyperkeratosis, fissures, corneal erosion, opacification, iritis, conjunctivitis, and possible blindness. To the best of our knowledge, the chemical, physical, and toxicological properties have not been thoroughly investigated.

FIGURE 1.1
Partial MSDS for glacial acetic acid. MSDS information is subject to change. For the most current version, contact Sigma-Aldrich. *Courtesy of Sigma-Aldrich Corp., St. Louis, MO;* www.sigma-aldrich.com.

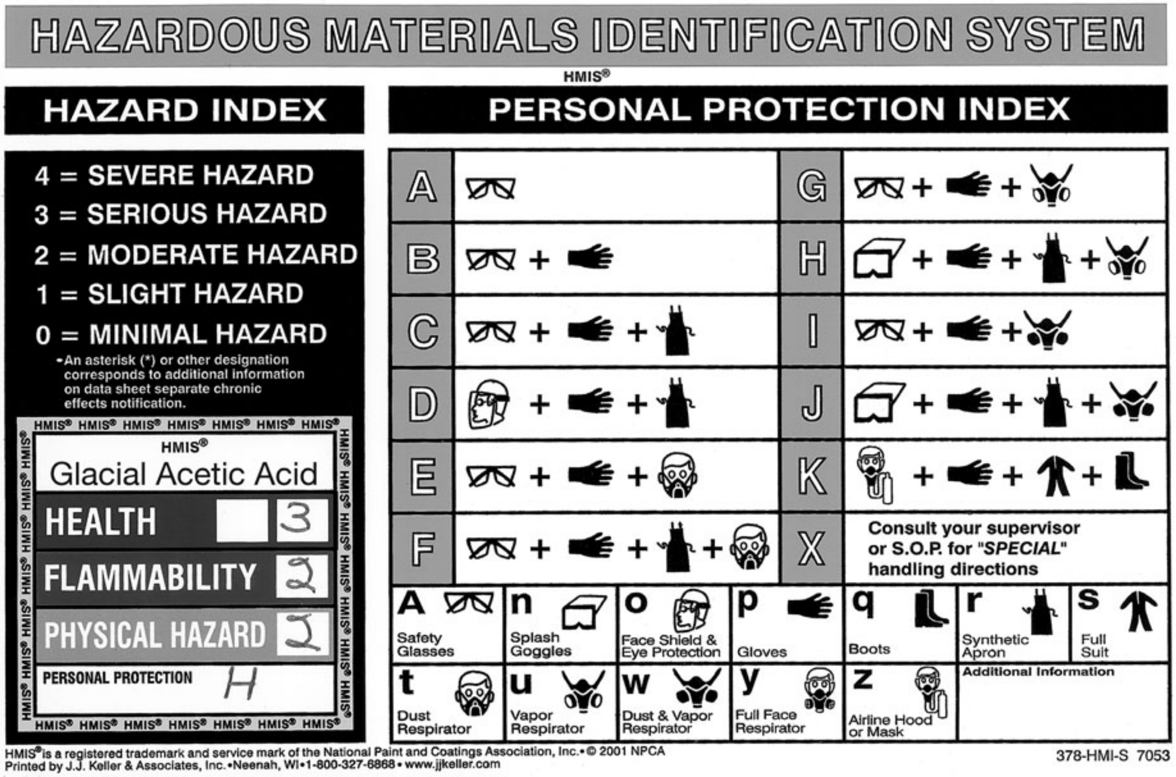
HAZARDOUS MATERIALS IDENTIFICATION SYSTEM
HMIS®
HAZARD INDEX
PERSONAL PROTECTION INDEX
4 = SEVERE HAZARD
3 = SERIOUS HAZARD
2 = MODERATE HAZARD
1 = SLIGHT HAZARD
0 = MINIMAL HAZARD
•An asterisk (*) or other designation corresponds to additional information on data sheet separate chronic effects notification.
HMIS®
Glacial Acetic Acid
HEALTH 3
FLAMMABILITY 2
PHYSICAL HAZARD 1
PERSONAL PROTECTION H
A
B
C
D
E
F
G
H
I
J
K
X
Consult your supervisor or S.O.P. for "SPECIAL" handling directions
A Safety Glasses
n Splash Goggles
o Face Shield & Eye Protection
p Gloves
q Boots
r Synthetic Apron
s Full Suit
t Dust Respirator
u Vapor Respirator
w Dust & Vapor Respirator
y Full Face Respirator
z Airline Hood or Mask
Additional Information
HMIS® is a registered trademark and service mark of the National Paint and Coatings Association, Inc.• © 2001 NPCA
Printed by J.J. Keller & Associates, Inc.•Neenah, WI•1-800-327-6868•www.jjkeller.com
378-HMI-S 7053

FIGURE 1.2 HMIS® III Wall Chart containing an HMIS® III label (lower left-hand corner) for glacial acetic acid displaying the hazard indices for the chemical. HMIS® III is a registered trademark of the National Paint & Coatings Association, Inc. (NPCA). *It is used with permission and may not be further reproduced. NPCA has granted an exclusive license to produce and distribute HMIS® III materials to J. J. Keller & Associates, Inc. Those wishing to utilize the HMIS® III system should contact J. J. Keller at 1-800-327-6868 or* www.jjkeller.com.

disposal will require the collection of waste materials from each laboratory worker, and the institution will be responsible for removal. For each procedure described in this book, appropriate handling of all reagents, materials, and equipment will be recommended.

It is essential that all students be aware of the potential hazards of working in a biochemistry laboratory. A set of rules is an appropriate way to communicate the importance of practicing safe science. The general rules outlined in Figure 1.3 serve as guidelines. Your institution and instructor may have their own list of rules or may want to add guidelines for specific activities. Rules of laboratory safety and chemical handling are not designed to impede productivity, nor should they instill a fear of chemicals or of laboratory procedures. Rather, their purpose is to create a healthy awareness of potential laboratory hazards, to improve the efficiency of each student worker, and to protect the general public and the environment from waste contamination. The list of references at the end of this chapter includes books, journal articles, manuals, and Web sites describing proper and detailed safety procedures.

1. Some form of eye protection is required at all times. Safety glasses with wide side shields or goggles are recommended, but normal eyeglasses with safety lenses may be permitted under some circumstances. Your instructor will inform you of the type of eye protection required. A statement regarding the wearing of contact lenses in the laboratory has been made by the American Chemical Society.[1] In general, contact lenses may be acceptable for wear in the laboratory, but the student must, of course, also wear safety glasses with side shields or goggles like all other students. Students who wear contacts must report to their lab instructor to determine the local rules for the lab.
2. Wear appropriate clothes — comfortable, well-fitting, older clothes that cover most of your skin. Sandals or bare feet are never allowed.
3. Never work alone in the laboratory.
4. Be familiar with the properties of all chemicals used in the laboratory. This includes their flammability, reactivity, toxicity, and proper disposal. This information may be obtained from your instructor, from the HMIS® III label, and from the MSDS. Always wear disposable gloves when using potentially dangerous chemicals or infectious agents.
5. Be familiar with your local rules for the safe handling and disposal of all non-chemical hazards. These include broken glass, “sharps” (needles, syringes, etc.), and biohazards (blood, bacteria, etc.).
6. Be especially careful with electrical equipment like stirrers, hot plates, and power supplies (electrophoresis, etc.). Always unplug before handling and avoid contact with water.
7. If open flames like those of Bunsen burners are necessary, make sure there are no flammable solvents in the area.
8. Eating, drinking, and smoking in the laboratory are strictly prohibited at all times.
9. Unauthorized experiments are not allowed.
10. Mouth suction should never be used to fill pipets or to start siphons.
11. Become familiar with the location and use of standard safety features in your laboratory. All laboratories should be equipped with fire extinguishers, eyewashes, safety showers, fume hoods, chemical-spill kits, fire blankets, first-aid supplies, and containers for chemical disposal. Receptacles should also be available for disposal of dangerous materials like glass, biohazards, and sharps. Any questions regarding the use of these features should be addressed to your instructor, teaching assistant, or lab director.
12. Report all chemical spills, presence of biohazards, accidents, and injuries (even minor) to your instructor.

---

[1] American Chemical Society, Washington, DC; Committee on Chemical Safety, 1998.

FIGURE 1.3 Guidelines for safety in the biochemistry laboratory.

B. KEEPING RECORDS AND COMMUNICATING EXPERIMENTAL RESULTS

The Laboratory Notebook

The biochemistry laboratory experience is not finished when you complete the experimental procedure and leave the laboratory. All scientists, including students, have the obligation to prepare and present written and oral reports on the results of their experimental work. Because these reports may be read and heard by many other professional scientists, they must be completed in a clear, concise, orderly, and accurate manner. Reports are most easily prepared outside of the lab using notes taken in a laboratory notebook *while* the experiment is in progress. These notes usually include procedural details, preparation of all reagents and solutions, setup of equipment, collection of data, and your thought processes and observations during the experiment. Experiments are often complex and move rather quickly, and it would be impossible to write down your data and observations after you have completed the experiments and left the lab. It is also not a good practice to record results on scraps of paper or on paper towels that may easily become lost or destroyed. The lab notebook will also come in handy if you need to troubleshoot or repeat an experiment because of inconsistent results.

Your instructor may have his or her own rules for preparation of the lab notebook, but here are some useful guidelines:

- The notebook should be hardbound with quadrille-ruled (gridded) pages; writing should be done with pen. This provides a permanent, durable record and the potential for construction of tables, graphs, charts, etc. Number each page of the book.
- Save the first few pages of the book for construction of a table of contents. Keep this up-to-date by entering the name of each experiment and page number.
- Use the right-hand pages only for writing your experimental notes. The left-hand pages may be used as scratch paper for your own personal notes, reminders, or calculations not appropriate for the main entry.
- Each entry for an experiment or project must begin with a title and date. The general outline required by many instructors for the written material is shown in Figure 1.4 and described below. Note that Parts I–IIc are labeled Prelab and should be completed before you begin the actual procedures in the lab.

Details of the Experimental Write-Up (see Figure 1.4)

Below is an outline that may be used as a guideline to write a complete report on an experiment.

Prelab

I. Introduction
  (a) Objective or purpose
  (b) Theory
II. Experimental
  (a) Table of materials and reagents
  (b) List of equipment
  (c) Flowchart
  (d) Record of procedure
III. Data and Calculations
  (a) Record of all raw data including printouts
  (b) Method of calculation with statistical analysis
  (c) Present final data in tables, graphs, or figures when appropriate.
IV. Results and Discussion
  (a) Conclusions
  (b) Compare results with known values
  (c) Discuss the significance of the data
  (d) Was the original objective achieved?
  (e) Literature references

FIGURE 1.4 General outline for experimental write-up in your notebook.

Introduction

This section begins with a three- or four-sentence statement of the objective or purpose of the experiment. When preparing this statement, ask yourself, "What are the goals of this experiment?" This statement is followed by a brief discussion of the theory behind the experiment. If a new technique or instrumental method is introduced, give a brief description of the method. Include chemical and biochemical reactions and structures of reagents when appropriate.

Experimental

Begin this section with a list of all reagents and materials used in the experiment. The sources of all chemicals and the concentrations of solutions should be listed. Instrumentation is listed with reference to company name and model number. A flowchart to describe the stepwise procedure for the experiment should be included after the list of equipment.

The write-up to this point is to be completed as a Prelab assignment. The experimental procedure followed is then recorded in your notebook as you proceed through the activities. The details should be sufficient so that a fellow student could use your notebook to repeat the experiment. You should include observations, such as color change or gas evolution, made during the experiment. If you obtain a computer printout of numbers, a spectrum from a spectrophotometer, or a photograph, these records must be saved with the notebook.

Data and Calculations

All raw data from the experiment are to be recorded directly in your notebook, not on separate sheets of paper that can easily become lost. Calculations involving the data must be included for at least one series of measurements. All data numbers should be analyzed by appropriate statistical methods using computer programs as described in Chapter 1, Section E, p. 23.

For many experiments, the clearest presentation of data is in tabular or graphical form. A graph may be prepared directly on the gridded pages of your notebook, or by computer software.

Results and Discussion

This important section of your write-up answers the questions "Did you achieve your proposed goals and objectives?" and "What is the significance of the data?" Any conclusion that you make must be supported by experimental results. It is often possible to compare your data with known values and results from the literature. If problems were encountered in the experiments, these should be outlined with possible remedies for future experiments.

All library references (books, journal articles, and Web sites) that were used to complete the experiment should be listed at the end of the write-up. It is especially important to report references used for laboratory procedures. The standard format to follow for a reference listing is shown at the end of this chapter in the Further Reading section.

Everyone has his or her own writing style. Because there is always room for improvement, it is imperative that you continually try to enhance your writing skills. When your instructor reviews your notebook, he or she should include helpful writing tips. References at the end of this chapter provide further instructions in scientific writing. Your instructor may accept, as a final report, the experimental write-up as described above and in Figure 1.4. However, he or she may request that you present your experimental results in one of the more formal written or oral modes described next.

Several electronic lab notebook (ELN) software packages are available commercially. For example, CERF by Resentris is an ELN designed for biology. The use of ELN's is still somewhat controversial, so you should ask your instructor if they are allowed in your lab. The primary purpose for maintaining a lab notebook, especially in academic and industrial/pharmaceutical labs, is to have a permanent and complete record for scientific and legal reasons. Lab notebooks are legal documents that are essential in patent lawsuits in order to prove authenticity and ownership of a discovery. Thus, it must be proven that the ELN is accurate, confidential, and maintained by authorized scientists, something that may be difficult with some current computer systems.

Communicating Results from Biochemistry Research

A scientific project is not complete until its discoveries have been communicated to colleagues around the world. The three most important methods or tools for communication are: the **scientific paper,** the **oral presentation,** and the **poster.** Although there are many differences in how to prepare for these three common methods of introducing new biochemical information, they all have one thing in common—the sharing of experimental results and conclusions. The distinct rules and traditions of each of the methods will be described and compared here.

The Scientific Paper

A paper published in a biochemical journal is a formal way to report research results to colleagues in the international biochemical community. Before writing such a document, one must first determine the journal to which the article will be submitted. There are hundreds of journals that accept manuscripts in the field of biochemistry (see Figure 1.5). Some have very high rank, prestige, and status based on the significance of research results published, reputation of authors, numbers of citations, whether or not manuscripts are peer-reviewed, and numbers of readers. Most journals are peer-reviewed, which indicate that before a manuscript is published, it is studied by members of the journal's editorial board to assure that the manuscript is scientifically significant, that it appears to be accurate, and that it is useful and of value to readers of the journal. Some journals accept manuscripts in all areas of biochemistry, but the manuscripts undergo rigorous peer-review by scientists with a certain specialty in the field. Other reputable journals are more specialized and accept peer-reviewed articles only in certain areas of biochemistry. Perhaps the best advice is to submit the manuscript to the most prestigious journal that has a large audience interested in his or her specialized topic. Publishing a paper in a reputable, peer-reviewed journal offers historic permanence for one's work, status, and exposure as a scientist; however, because the lag time between acceptance and publication of

Accounts of Chemical Research
Analytical Biochemistry
Annual Review of Biochemistry
Archives of Biochemistry and Biophysics
Biochemical and Biophysical Research Communications
Biochemical and Molecular Medicine
Biochemical Journal
Biochemistry
Biochemistry and Molecular Biology Education
Biochimica et Biophysica Acta: General Subjects; Molecular and Cell Biology; Protein Structure and Molecular Enzymology
BioEssays
Bioorganic Chemistry
Biophysical Journal
Canadian Journal of Biochemistry
Cell
ChemBioChem
Chemistry and Biology
Current Opinion in Structural Biology
DNA Research Online
Electrophoresis
European Journal of Biochemistry
FASEB Journal
Glycobiology
Glycoconjugate Journal
Journal of Biochemistry
Journal of Biological Chemistry
Journal of Cell Biology
Journal of Cell Biology Education
Journal of Chemical Education
Journal of Lipid Research
Journal of Molecular Biology
Journal of Neurochemistry
Journal of Plant Physiology
Macromolecules
Methods: A Companion to Methods in Enzymology
Molecular and Cellular Biochemistry
Molecular and Cellular Proteomics
Nature
Nature Reviews Molecular Cell Biology
Nature Structural Biology
Nucleic Acid Research
Phytochemistry
Proceedings of the National Academy of Sciences USA
Prostaglandins, Leukotrienes and Essential Fatty Acids
Protein Science
Proteomics
RNA
Science
Scientific American
Trends in Biochemical Sciences

FIGURE 1.5
Some journals that publish research articles in biochemistry. Web sites for the journals may be found using a search engine.

a manuscript can sometimes stretch up to one year, the data reported can become outdated or insignificant.

Most research journals are now available in electronic form and accessible through college/university or corporate libraries. The subscription costs are usually borne by the academic institutions or corporations and available to students, faculty, and research staff.

Your instructor may require that you write up the results from an experiment in the form of a journal article, so it is important to understand the conventions used in preparing a manuscript for publication. Most biochemical journal articles have the same basic organization: Title, Abstract, Introduction, Experimental Methods, Results, Discussion, and References. The specific requirements for each of these sections vary among the many journals, so it is important to review several articles in different journals to get a flavor of what is expected. All scientific journals publish "Instructions to Authors," which are available on their Web site. Although your instructor will most likely expect you to follow the requirements of a specific journal, it is instructive to study articles in the following high-ranking journals that publish biochemistry topics: *The Journal of Biological Chemistry* (published by the American Society for Biochemistry and Molecular Biology), *Biochemistry* (published by the American Chemical Society), *Science*, and *Nature*. Use a search engine such as Google, Yahoo, etc. to find Web sites for other journals.

The Oral Presentation

The purpose and mechanics of an oral presentation are quite different from preparing and publishing a paper. You may write a paper over a period of days, weeks, and even months, and the published work is available as a permanent record for readers to study and reference anytime in the future. In an oral presentation, you have a fleeting moment to present data and attempt to convince your audience of the importance of your work. One advantage of the oral presentation, however, is that it provides an opportunity for more direct contact with your audience than does a paper; thus the opportunity exists for immediate questions and feedback.

Presentations usually range from 15 to 60 minutes. Shorter presentations cover a much smaller unit of a research project, whereas 60-minute talks (often called seminars) can give a broader exposure to the research area.

Scientific presentations involve mixed media—oral and visual. The important verbal points are reinforced with the use of a visual aid such as a figure, graph, or other element. Scientific presenters today most often use PowerPoint, computer software that projects electronic slides onto a screen, although overhead transparencies are also acceptable and efficient. Whatever the type of visual aid, the slides must be carefully constructed with special concern for the total number of slides and the amount of information on each. Some presenters use the approximate ratio of one-to-two slides per minute of presentation.

The organization of a talk is similar to that of a paper—Introduction, Experimental Methods, Results, Discussion, Conclusions, Questions/Comments. If your instructor expects you to present a talk, he or she will provide specific information regarding length of time, range of topic, type of visual aids, multimedia, etc.

The Scientific Poster

The scientific poster is a communication method that may be considered a hybrid, as it combines elements of the oral presentation (verbal expression and visual aids) with elements of a paper (printed text and figures). The poster has become the primary medium by which new scientific information is exchanged at all professional conferences, including local, regional, national, and international meetings. At meetings, posters that consist of text and figures arranged in panels on a thin piece of cardboard (average 3′ × 5′) are set up in designated areas during specified times (usually for a day or two), and there is often an official time when the presenter is to be in attendance. The poster, however, may be available to readers for long periods of time in the absence of the presenter. Some of the specific characteristics that describe a poster include (Figure 1.6):

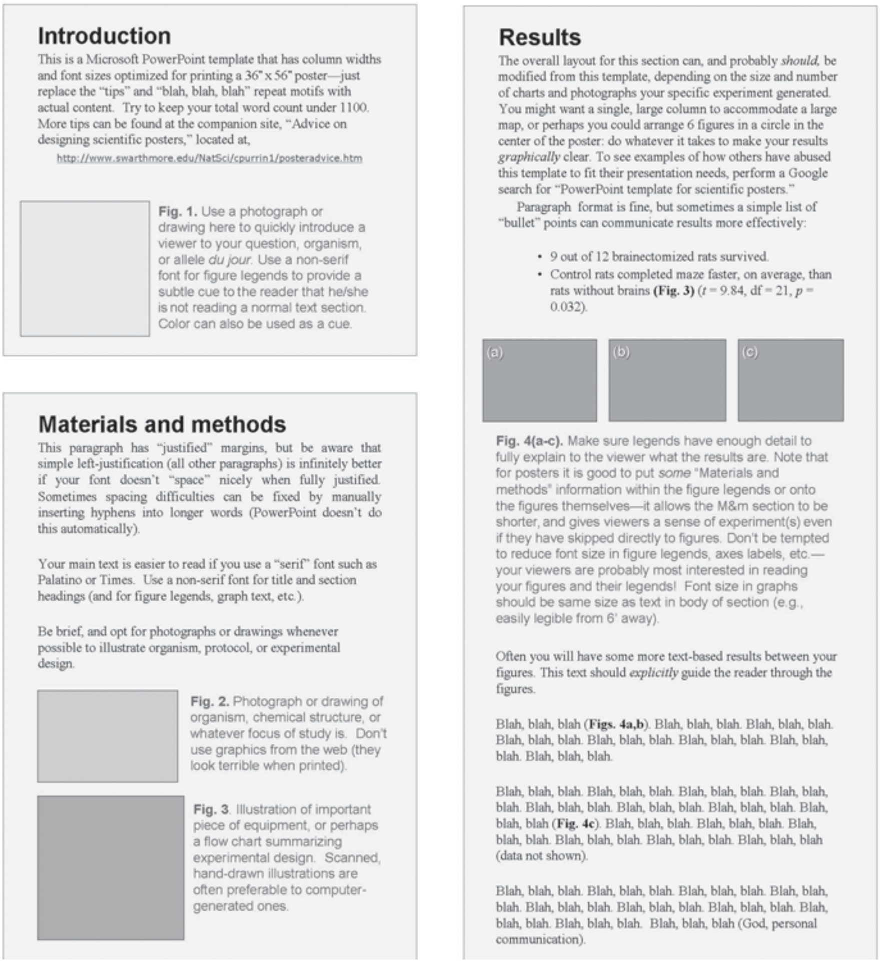

Introduction

This is a Microsoft PowerPoint template that has column widths and font sizes optimized for printing a 36" x 56" poster—just replace the "tips" and "blah, blah, blah" repeat motifs with actual content. Try to keep your total word count under 1100. More tips can be found at the companion site, "Advice on designing scientific posters," located at,

http://www.swarthmore.edu/NatSci/cpurrin1/posteradvice.htm

Fig. 1. Use a photograph or drawing here to quickly introduce a viewer to your question, organism, or allele *du jour*. Use a non-serif font for figure legends to provide a subtle cue to the reader that he/she is not reading a normal text section. Color can also be used as a cue.

Materials and methods

This paragraph has "justified" margins, but be aware that simple left-justification (all other paragraphs) is infinitely better if your font doesn't "space" nicely when fully justified. Sometimes spacing difficulties can be fixed by manually inserting hyphens into longer words (PowerPoint doesn't do this automatically).

Your main text is easier to read if you use a "serif" font such as Palatino or Times. Use a non-serif font for title and section headings (and for figure legends, graph text, etc.).

Be brief, and opt for photographs or drawings whenever possible to illustrate organism, protocol, or experimental design.

Fig. 2. Photograph or drawing of organism, chemical structure, or whatever focus of study is. Don't use graphics from the web (they look terrible when printed).

Fig. 3. Illustration of important piece of equipment, or perhaps a flow chart summarizing experimental design. Scanned, hand-drawn illustrations are often preferable to computer-generated ones.

Results

The overall layout for this section can, and probably *should*, be modified from this template, depending on the size and number of charts and photographs your specific experiment generated. You might want a single, large column to accommodate a large map, or perhaps you could arrange 6 figures in a circle in the center of the poster: do whatever it takes to make your results *graphically* clear. To see examples of how others have abused this template to fit their presentation needs, perform a Google search for "PowerPoint template for scientific posters."

Paragraph format is fine, but sometimes a simple list of "bullet" points can communicate results more effectively:

- 9 out of 12 brainectomized rats survived.
- Control rats completed maze faster, on average, than rats without brains **(Fig. 3)** (t = 9.84, df = 21, p = 0.032).

Fig. 4(a-c). Make sure legends have enough detail to fully explain to the viewer what the results are. Note that for posters it is good to put *some* "Materials and methods" information within the figure legends or onto the figures themselves—it allows the M&m section to be shorter, and gives viewers a sense of experiment(s) even if they have skipped directly to figures. Don't be tempted to reduce font size in figure legends, axes labels, etc.—your viewers are probably most interested in reading your figures and their legends! Font size in graphs should be same size as text in body of section (e.g., easily legible from 6' away).

Often you will have some more text-based results between your figures. This text should *explicitly* guide the reader through the figures.

Blah, blah, blah (**Figs. 4a,b**). Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah.

Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah (**Fig. 4c**). Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah (data not shown).

Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah (God, personal communication).

FIGURE 1.6 Template for a poster in biochemistry. *Courtesy of Professor Colin Purrington, Department of Biology, Swarthmore College*, www.swarthmore.edu/NatSci/cpurrin1/posteradvice.htm.

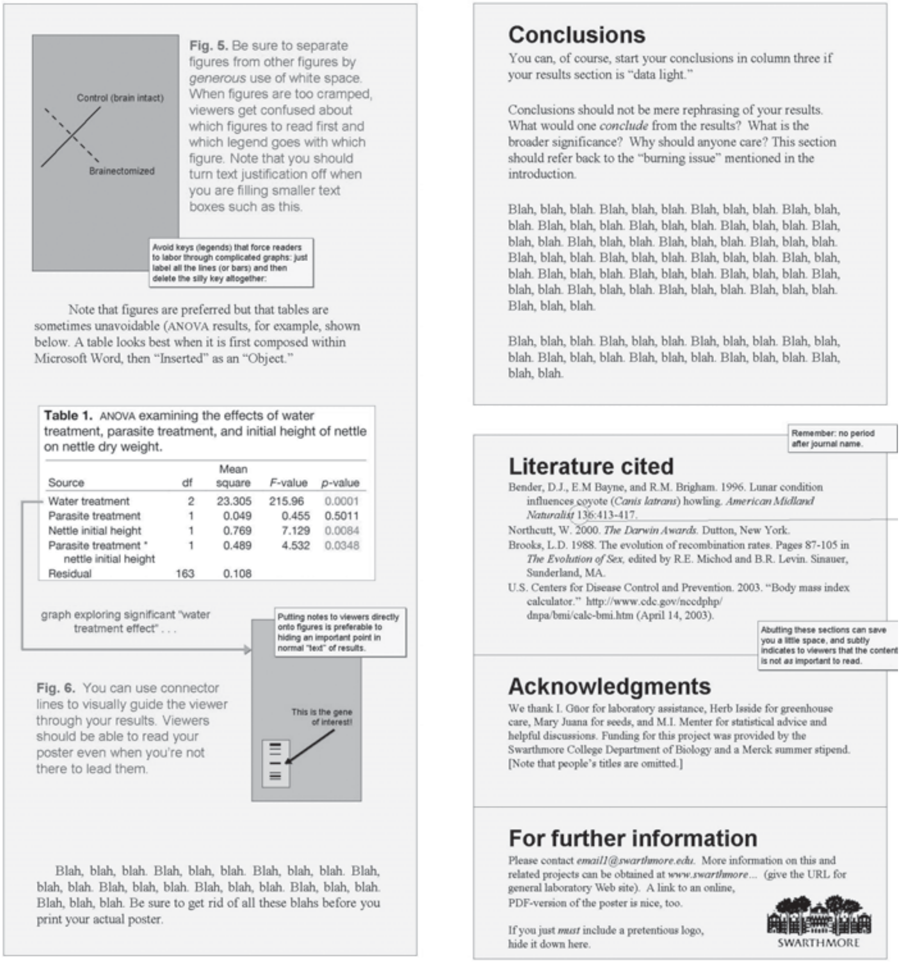
Control (brain intact)
Brainectomized
This is the gene of interest!

Fig. 5. Be sure to separate figures from other figures by *generous* use of white space. When figures are too cramped, viewers get confused about which figures to read first and which legend goes with which figure. Note that you should turn text justification off when you are filling smaller text boxes such as this.

Avoid keys (legends) that force readers to labor through complicated graphs: just label all the lines (or bars) and then delete the silly key altogether:

Note that figures are preferred but that tables are sometimes unavoidable (ANOVA results, for example, shown below. A table looks best when it is first composed within Microsoft Word, then "Inserted" as an "Object."

Table 1. ANOVA examining the effects of water treatment, parasite treatment, and initial height of nettle on nettle dry weight.

Source	df	Mean square	*F*-value	*p*-value
Water treatment	2	23.305	215.96	0.0001
Parasite treatment	1	0.049	0.455	0.5011
Nettle initial height	1	0.769	7.129	0.0084
Parasite treatment * nettle initial height	1	0.489	4.532	0.0348
Residual	163	0.108		

graph exploring significant "water treatment effect" . . .

Putting notes to viewers directly onto figures is preferable to hiding an important point in normal "text" of results.

Fig. 6. You can use connector lines to visually guide the viewer through your results. Viewers should be able to read your poster even when you're not there to lead them.

Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Be sure to get rid of all these blahs before you print your actual poster.

Conclusions

You can, of course, start your conclusions in column three if your results section is "data light."

Conclusions should not be mere rephrasing of your results. What would one *conclude* from the results? What is the broader significance? Why should anyone care? This section should refer back to the "burning issue" mentioned in the introduction.

Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah.

Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah. Blah, blah, blah.

Remember: no period after journal name.

Literature cited

Bender, D.J., E.M Bayne, and R.M. Brigham. 1996. Lunar condition influences coyote (*Canis latrans*) howling. *American Midland Naturalist* 136:413-417.

Northcutt, W. 2000. *The Darwin Awards*. Dutton, New York.

Brooks, L.D. 1988. The evolution of recombination rates. Pages 87-105 in *The Evolution of Sex*, edited by R.E. Michod and B.R. Levin. Sinauer, Sunderland, MA.

U.S. Centers for Disease Control and Prevention. 2003. "Body mass index calculator." http://www.cdc.gov/nccdphp/dnpa/bmi/calc-bmi.htm (April 14, 2003).

Abutting these sections can save you a little space, and subtly indicates to viewers that the content is not *as* important to read.

Acknowledgments

We thank I. Güor for laboratory assistance, Herb Isside for greenhouse care, Mary Juana for seeds, and M.I. Menter for statistical advice and helpful discussions. Funding for this project was provided by the Swarthmore College Department of Biology and a Merck summer stipend. [Note that people's titles are omitted.]

For further information

Please contact *email1@swarthmore.edu*. More information on this and related projects can be obtained at *www.swarthmore*... (give the URL for general laboratory Web site). A link to an online, PDF-version of the poster is nice, too.

If you just *must* include a pretentious logo, hide it down here.

SWARTHMORE

FIGURE 1.6
Continued.

- usually composed of small units of a research project and most often based on preliminary results and conclusions.
- contains many of the same organizational elements as a paper or talk—Title/Authors, Abstract, Introduction, Methods, Results, Discussion, Conclusions, References—but in a much briefer form.
- often enhanced with a brief, oral summary given by the presenter. Only the main points such as the purpose, results, conclusions, and future experiments for the project should be included in this concise summary. As a presenter, during your official time at the poster you will be visited by individuals or small groups who will spend an average of about 10–12 minutes at your poster.
- must be completely self-explanatory, as you will not always be present to answer a reader's questions.
- the environment is usually interactive and informal, allowing for one-on-one contact with other researchers.

- poster sessions at scientific meetings are very democratic and inclusive, as the presenters and audience may consist of all levels of scientists—tenured research professors including Nobel Prize winners, as well as undergraduate research students.

Your instructor may request that you prepare a poster for local display at your institution or for presentation at a regional or national meeting. Specific details will not be given here, as all organizations sponsoring poster sessions at meetings publish their own rules and regulations (regarding poster size, font sizes, etc.) for preparing posters. Many colleges and universities now schedule local meetings where students may obtain experience preparing and presenting posters about their research results. Attend one of these local meetings or walk around the halls of your chemistry and biology departments looking for posters made by research students at your institution. These may serve as very good models for your own creation. You may also find useful information about the specific details of poster construction by searching the Internet. Some helpful Web sites with poster templates are listed in the Further Reading section at the end of this chapter.

C. USING BIOCHEMICAL REAGENTS AND SOLUTIONS

Water Purity

Water is the most common and widely-used substance in the biochemistry laboratory. Applications of water usage include: (1) solvent for preparing most buffer and reagent solutions; (2) column chromatography; (3) high-performance liquid chromatography; (4) tissue culture; and (5) washing glassware. Both the quality and quantity of water required must be considered for each lab application. Ordinary tap water is relatively abundant, but its quality is very low. It contains a variety of impurities including particulate matter (sand, silt, etc.); dissolved organics, inorganics, and gases; and microorganisms (bacteria, viruses, protozoa, and algae). In addition, the natural degradation of microorganisms leads to the presence of byproducts called pyrogens. Tap water should never be used for the preparation of reagent solutions or for any sensitive procedures. For most laboratory procedures, it is recommended that some form of purified water be used. The purity of water is usually measured in terms of **resistivity** (the ability of a liquid to restrict the flow of an electric current). Units for resistivity are Megohms × cm ($M\Omega$.cm) with a relative scale of 1.0 $M\Omega$.cm to 18.2 $M\Omega$.cm (the highest theoretical purity).

There are five basic water purification technologies—distillation, ion-exchange, activated carbon adsorption, reverse osmosis, and membrane filtration. Most academic and industrial research laboratories are equipped with "in-house" purified water, which typically is produced by a combination of the above purifying processes and piped throughout all the labs in a building. The water quality necessary will depend on the solutions to be prepared and on the biochemical procedures to be investigated. For most procedures carried out in the biochemistry lab, water purified by ion-exchange, reverse osmosis, or distillation is usually acceptable. Of these three processes, distillation is the slowest, least energy-efficient, least pure (best is 1.0 $M\Omega$.cm), and most high-maintenance—especially in areas with hard water (needs regular de-scaling). Distilled water must also be stored to prevent contamination by microbes. For special procedures such as buffer standardization, liquid chromatography, and tissue culture, ultra-pure water, which is usually bottled and available commercially, should be used. Water that is purified only by ion-exchange will be low in metal-ion concentration, but may contain certain organics that are washed from the ion-exchange resin. These contaminants will increase the UV-absorbance properties of water. If sensitive UV-spectroscopic measurements are to be made, distilled water (especially glass-distilled) is better than de-ionized. If large volumes of high-purity water are required (18.2 MΩ.cm), reverse osmosis should be the choice.

Cleaning Laboratory Glassware

The results of your experimental work will depend, to a great extent, on the cleanliness of your equipment, especially glassware used for preparing and transferring solutions. There are at least two important reasons for this: (1) many of the chemicals and biochemicals will be used in milligram, microgram, or even nanogram amounts. Any contamination, whether on the inner walls of a beaker, in a pipet, or in a glass cuvette, could be a significant percentage of the total experimental sample; (2) many biochemicals and biochemical processes are sensitive to one or more of the following common contaminants: metal ions, detergents, and organic residues. In fact, the objective of many experiments is to investigate the effect of a metal ion, organic molecule, or other chemical agent on a biochemical process. Contaminated glassware will virtually ensure failure in these activities.

The preferred method for cleaning glassware is to begin with hot tap water. Rinse the glassware at least 10 times with this; then rinse 4–6 times with distilled or de-ionized water. Occasionally it is necessary to use a detergent for cleaning. Use a dilute detergent solution (0.5% in water) followed by 5–10 water rinses with distilled or de-ionized water.

Dry equipment is required for most processes carried out in the biochemistry laboratory. When you needed dry glassware in the organic laboratory, you probably rinsed the glassware with acetone, which rapidly evaporated, leaving a dry surface. Unfortunately, this technique coats the surface with an organic residue consisting of nonvolatile contaminants found in the acetone. Because this residue could interfere with your experiments, it is best to refrain from acetone washing. Glassware and plasticware should be rinsed well with purified water and dried in an oven designated for glassware, not one used for drying chemicals.

Never clean cuvettes or any optically polished glassware with ethanolic KOH or other strong base, as this will cause etching. All glass cuvettes should be cleaned carefully with hot tap water or 0.5% detergent solution, in a sonicator bath or in a cuvette washer, followed by thorough rinsing with purified water.

Solutions: Concentrations and Calculations

The concentrations for solutions used in the biochemistry laboratory may be expressed in several different units. The most common units are:

- ***Molarity (M):*** concentration based on the number of moles of solute per liter of solution. A 1 *M* solution of the amino acid alanine (MW = 89.1) contains 1 mole, or 89.1 g, of alanine in a solution volume of 1 liter. In biochemistry, it is more common to use concentration ranges that are millimolar (mM, 1×10^{-3} M), micromolar (μM, 1×10^{-6} M), or nanomolar (nM, 1×10^{-9} M). A 1 mM solution of alanine contains 0.089 g, or 89 mg (89.1×0.001), of alanine in a solution volume of 1 liter. How many grams of alanine are present in 100 mL of the 1 mM alanine solution? (Ans: 0.0089 g). How many milligrams of alanine are present in 100 mL? (Ans: 8.9 mg).

- ***Percent by weight (% wt/wt):*** concentration based on the number of grams of solute per 100 g of solution. A 5% wt/wt solution of alanine contains 5 g of alanine in 100 g of solution. How many grams of alanine are present in 10 g of this solution? (Ans: 0.5 g).

STUDY EXERCISE 1.1 Preparation of Molar Solutions

a. Many solutions you use will be based on molarity. For practice, assume you require 1 liter of solution that is 0.1 M (100 mM) glucose:

$$\begin{aligned} \text{MW of glucose} &= 180.2 \\ \text{1 mole of glucose} &= 180.2\ \text{g} \\ \text{0.1 mole of glucose} &= 18.02\ \text{g} \end{aligned}$$

To prepare a 0.1 M glucose solution, weigh 18.02 g of glucose and transfer to a 1-liter volumetric flask. Add about 700–800 mL of purified water and swirl to dissolve. Then add water so that the bottom of the meniscus is at the etched line on the flask. Stopper and mix well. The flask must be labeled with solution contents (0.1 M glucose), date prepared, and name of preparer.

b. Assume that you need only 250 mL of 0.10 M glucose. Explain how you would prepare the solution. Emphasize any changes from Part (a).

STUDY EXERCISE 1.2 Concentration Unit Conversions

It is often necessary in your biochemistry lab work to convert one concentration unit to another. For example, you may need to know the concentration of the 0.1 M glucose solution (Study Exercise 1.1) in concentration terms of mg/mL. Here are some basic calculations for practice.

a. Convert the concentration units of 0.1 M glucose to the units of mg of glucose in 1 mL (mg/mL).
According to the procedure described in Study Exercise 1.1(a), the 1000-mL solution of glucose contains 18.02 g of glucose, or is 18.02 g/1000 mL. This is equivalent to 1.80 g/100 mL or 0.018 g/mL. Therefore, the 0.1 M glucose solution concentration is equivalent to about 18 mg/mL.

b. Convert the concentration units of 1 M alanine to the units of g/100 mL.

c. Convert the concentration units of 1 M alanine to the units of % wt/vol.

d. Calculate the concentration of a 0.1 M glucose solution to the units of % wt/vol.

e. Calculate the molar concentration of an ethanol solution that was prepared by adding 10 mL of 100% ethanol to a 100-mL volumetric flask followed by adding water to the line. The density of ethanol is 0.789.

- ***Percent by volume (% wt/vol):*** concentration based on the number of grams of solute per 100 mL of solution. A 10% wt/vol solution of alanine contains 10 g of alanine in 100 mL of solution. How many grams of alanine are present in 50 mL of this solution? (Ans: 5 g).
- ***Weight per volume (wt/vol):*** concentration based on the number of grams, milligrams, or micrograms of solute per unit volume; for example, mg/mL, g/L, mg/100 mL, etc. A solution of alanine, concentration wt/vol = 5 g/L, contains 5 g of alanine in a liter of solution. How many grams of alanine would be present in 2 liters of this solution? In 10 mL? (Ans: 10 g; 0.005 g).

Preparing and Storing Solutions

In general, solid solutes should be weighed on weighing paper or plastic weighing boats, with the use of an electronic analytical or top-loading balance. Liquids are more conveniently dispensed by volumetric techniques; however, this assumes that the density is known. If a small amount of a liquid is to be weighed, it should be added to a tared flask by means of a disposable Pasteur pipet with a latex bulb. The hazardous properties of all materials should be known before use (read MSDS) and the proper safety precautions obeyed.

The storage conditions of reagents and solutions in the biochemistry lab are especially critical. Although some will remain stable indefinitely at room temperature, it is good practice to store all solutions in a closed container. Often it is necessary to store some solutions in a refrigerator at 4°C. This inhibits bacterial growth and slows decomposition of the reagents. Some solutions may require storage below 0°C. If these are aqueous solutions or others that will freeze, be sure there is room for expansion inside the container. Stored solutions must always have a label containing the name and concentration of the solution, the date prepared, and the name of the preparer.

All stored containers, whether at room temperature, 4°C, or below freezing, must be properly sealed. This reduces contamination by bacteria and vapors in the laboratory air (carbon dioxide, ammonia, HCl, etc.). Volumetric flasks, of course, have glass stoppers, but test tubes, Erlenmeyer flasks, bottles, and other containers should be sealed with screw caps, corks, or hydrocarbon foil (Parafilm). Remember that hydrocarbon foil, a wax, is dissolved by solutions containing nonpolar organic solvents like chloroform, diethyl ether, and acetone.

Bottles of pure chemicals and reagents should also be properly stored. Many manufacturers now include the best storage conditions for a reagent on the label. The common conditions are: store at room temperature; store at 0–4°C; store below 0°C; or store in a desiccator at room temperature, 0–4°C, or below 0°C. Many biochemical reagents form hydrates by taking up moisture from the air. If the water content of a reagent increases, the molecular weight and purity of the reagent change. For example, when the coenzyme nicotinamide adenine dinucleotide (NAD^+) is purchased, the label usually reads "Anhydrous molecular weight = 663.5; when assayed, contained 3 H_2O per mole." The actual molecular weight that should be used for solution preparation is 663.5 + (18)(3) = 717.5. However, if this reagent is stored in a moist refrigerator or freezer outside a desiccator, the moisture content may increase to an unknown value.

D. QUANTITATIVE TRANSFER OF LIQUIDS

Practical biochemistry is highly reliant on analytical methods. Many analytical techniques must be mastered, but few are as important as the quantitative transfer of solutions. Some type of pipet will almost always be used in liquid transfer. Because students may not be familiar with the many types of pipets and the proper techniques in pipetting, this instruction is included here.

Pipets and Pipetting

Pipet Fillers

Figure 1.7 illustrates the various types of pipets and fillers. The use of any pipet requires some means of drawing reagent into the pipet. ***Liquids should never be drawn into a pipet by mouth suction on the end of the pipet!*** Small latex bulbs are available for use with disposable pipets (see Figure 1.7A). For volumetric and graduated pipets, two types of bulbs are available. One type (see Figure 1.7B) features a special conical fitting that accommodates common sizes of pipets. To use these, first place the pipet tip below the surface of the liquid. Squeeze the bulb with the left hand (if you are a right-handed pipettor) and then hold it tightly to the end of the pipet. Slowly release the pressure on the bulb to allow liquid to rise to 2 or 3 cm above the top graduated mark. Then, remove the bulb and

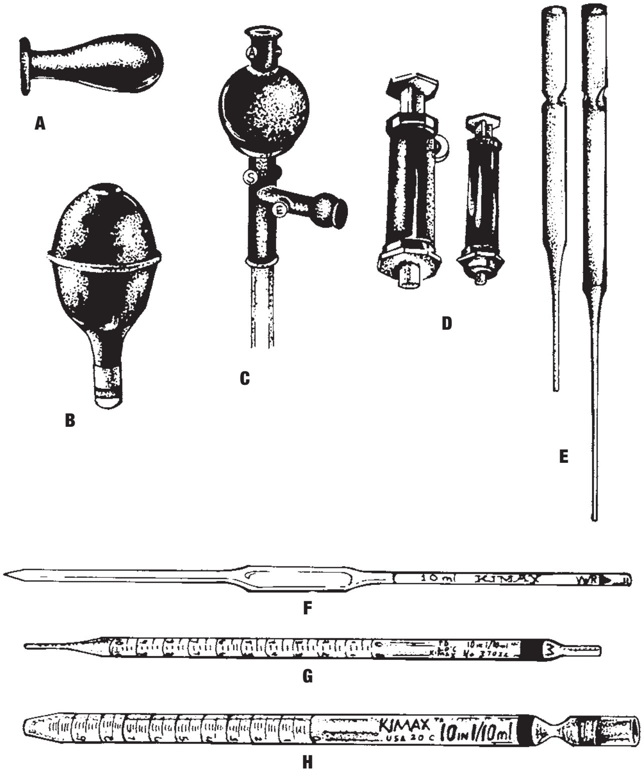

FIGURE 1.7 Examples of pipets and pipet fillers. *Courtesy of Sargent-Welch, VWR International;* www.vwr.com *A Latex bulb, B Pipet filler, C Mechanical pipet filler, D Pipettor pump, E Pasteur pipet, F Volumetric pipet, G Mohr pipet, H Serological pipet.*

quickly grasp the pipet with your index finger over the top end of the pipet. The level of solution in the pipet will fall slightly, but should not fall below the top graduated mark. If it does fall too low, use the bulb to refill.

Safety Pipet Fillers

Mechanical pipet fillers (made of silicone and sometimes called safety pipet fillers, propipets, or pi-fillers) are more convenient than latex bulbs. As shown in Figure 1.7C,D, these fillers are equipped with a system of hand-operated valves and can be used for the complete transfer of a liquid. The use of a safety pipet filler is outlined in Figure 1.8. ***Never allow any solvent or solution to enter the pipet bulb!*** To avoid this, two things must be kept in mind:

1. Always maintain careful control while using valve S to fill the pipet.
2. Never use valve S unless the pipet tip is below the surface of the liquid. If the tip moves above the surface of the liquid, air will be sucked into the pipet and solution will be flushed into the bulb.

Disposable Pasteur Pipets

Often it is necessary to perform a semi-quantitative transfer of a small volume (1–10 mL) of liquid from one vessel to another. Because pouring is not efficient, a **Pasteur pipet** with a small latex bulb may be used (see Figure 1.7A, E). Pasteur pipets are available in two lengths (15 and 23 cm) and hold about 2 mL of solution. These are especially convenient for the transfer of nongraduated amounts to and from test tubes. Typical recovery while using a Pasteur pipet is 90 to 95%. If dilution is not a problem, rinsing the original vessel with a solvent will increase the transfer yield. Used disposable pipets should be discarded in special containers for broken glass.

Calibrated Pipets

Although most quantitative transfers are now done with automatic pipetting devices, which are described later in the chapter, instructions will be given for the use of all types of pipets. If a quantitative transfer of a specific and accurate volume of liquid is required, some form of calibrated pipet must be used.

- ***Volumetric pipets*** (see Figure 1.7F) are used for the delivery of liquids required in whole-milliliter amounts (1, 2, 5, 10, 15, 20, 25, 50, and 100 mL). To use these pipets, draw liquid with a latex bulb or mechanical pipet filler to a level 2–3 cm above the fill line. Release liquid from the pipet until the bottom of the meniscus is directly on the fill line. Touch the tip of the pipet to the inside of the glass wall of the container from which it was filled. Transfer the pipet to the inside of the second container and release the liquid. Hold the pipet vertically, allow the solution to drain until the flow stops, and then wait an additional 5–10 seconds. Touch the tip of the pipet to the inside of the container to release the last drop from the outside of the tip. Remove the pipet from the container. Some liquid may still remain in the tip. Most volumetric pipets are calibrated as “TD” (to deliver), which

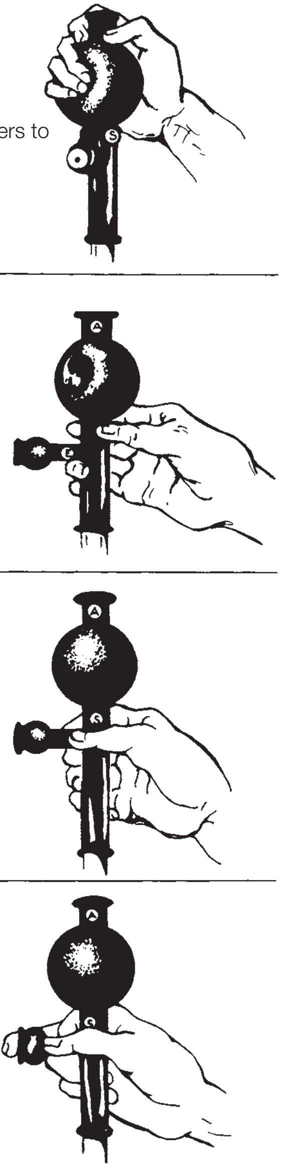

1. Using thumb and forefinger, press on valve A and squeeze bulb with other fingers to produce a vacuum for aspiration. Release valve A, leaving bulb compressed.

2. Insert pipet into liquid. Press on valve S. Suction draws liquid to desired level.

3. Press on valve E to expel liquid.

4. To deliver the last drop, maintain pressure on valve E, cover E inlet with middle finger, and squeeze the small bulb.

FIGURE 1.8
How to use a safety pipet filler.
Courtesy of Sargent-Welch, VWR International.

means the intended volume is transferred *without final blow-out*; that is, the pipet delivers the correct volume.

Fractional volumes of liquid are transferred with **graduated pipets,** which are available in two types:

- ***Mohr pipets*** (see Figure 1.7G) are available in long- or short-tip styles. Long-tip pipets are especially attractive for transfer to and from vessels with small openings. Virtually all Mohr pipets are TD, and they are available in many sizes (0.1 to 10 mL). The marked subdivisions are usually 0.01 or 0.1 mL, and the markings end a few centimeters from the tip. Selection of the proper size is especially important. For instance, do not try to transfer 0.2 mL with a 5- or 10-mL pipet. Use the smallest pipet that is practical. The use of a Mohr pipet is similar to that of a volumetric pipet. Draw the liquid into the pipet with a pipet filler to a level about 2 cm above the 0 mark. Lower the liquid level to the 0 mark. Remove the last drop from the tip by touching it to the inside of the glass container. Transfer the pipet to the receiving container and release the desired amount of solution. The solution should not be allowed to move below the last graduated mark on the pipet. Touch off the last drop.

- ***Serological pipets*** (see Figure 1.7H) are similar to Mohr pipets, except that they are graduated downward to the very tip and are designed for blow-out. Their use is identical to that of a Mohr pipet except that the last bit of solution remaining in the tip must be forced out into the receiving container with a rubber bulb. This final blow-out should be done after 15–20 seconds of draining.

Cleaning and Drying Pipets

Special procedures are required for cleaning glass pipets. Immediately after use, every pipet should be placed, tip up, in a vertical cylinder containing warm tap water or a dilute detergent solution (less than 0.5%). The pipet must be completely covered with solution. This ensures that any reagent remaining in the pipet is forced out through the tip. If reagent solutions are allowed to dry inside a pipet, the tips can easily become clogged and are very difficult to open. After several pipets have accumulated in the water or detergent solution, the pipets should be transferred to a pipet rinser. Pipet rinsers continually cycle fresh water through the pipets. Immediately after detergent wash, tap water may be used to rinse the pipets, but distilled water should be used for the final rinse. Pipets may then be dried in an oven.

Automatic Pipetting Devices

For most quantitative transfers, including many repeated small-volume transfers, a mechanical microliter pipettor (i.e., Eppendorf type, Pipetman) is ideal. This allows accurate, precise, and rapid dispensing of fixed volumes from 1 to 10,000 μL (0.001 to 10 mL). The pipet's push-button system can be operated with one hand, and it is fitted with detachable polypropylene tips (see Figure 1.9). The advantage of polypropylene tips is that the amount of reagent film remaining in the pipet after delivery is much less than for glass tips. Mechanical pipettors are available in many different sizes. Newer models offer continuous volume adjustment, so a single model can be used for delivery of specific volumes within a certain range. Multichannel pipettors are available, which fill several containers or wells at once (see Figure 1.10).

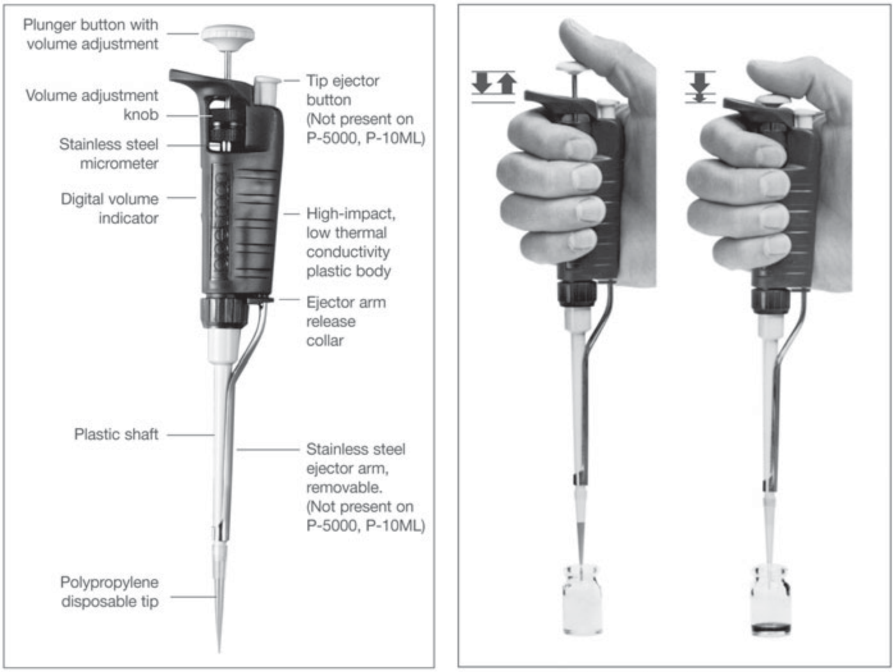
Plunger button with volume adjustment
Volume adjustment knob
Stainless steel micrometer
Digital volume indicator
Tip ejector button (Not present on P-5000, P-10ML)
High-impact, low thermal conductivity plastic body
Ejector arm release collar
Plastic shaft
Stainless steel ejector arm, removable. (Not present on P-5000, P-10ML)
Polypropylene disposable tip

Pipetman P–200 **Operating Pipetman**

FIGURE 1.9 How to use an adjustable pipetting device. Set the desired volume with the digital micrometer or plunger button. Attach a new disposable tip to the shaft of the pipet. Press tip on firmly with a slight twisting motion. Depress the plunger to the first positive stop, immerse the disposable tip into the sample liquid to a depth of 2–4 mm, and allow the pushbutton to return slowly to the up position and wait 1–2 seconds. To dispense sample, place the tip end against the side wall of the receiving vessel and depress the plunger slowly to the first stop. Wait 2–3 seconds, and then depress the plunger to the second stop to achieve final blow-out. Withdraw the device from the vessel carefully with the tip sliding along the inside wall of the vessel. Allow the plunger to return to the up position. Discard the tip by depressing the tip ejector button. *Photos courtesy of Rainin Instrument Company, Inc., Woburn, MA. Pipetman is a registered trademark of Gilson Medical Electronics. Exclusive license to Rainin Instrument Company, Inc.,* www.rainin.com.

To use the pipettor, choose the proper size and place a polypropylene pipet tip firmly onto the cone, as shown in Figure 1.9. Tips for pipets are available in several sizes, for 1–20, 20–250, 200–1000, 1000–5000, and 10,000-μL capacities. Details of the operation of an adjustable pipet are shown in Figure 1.9.

For rapid and accurate transfer of volumes greater than 5 mL, automatic repetitive dispensers are commercially available. These are particularly useful for the transfer of corrosive materials. The dispensers, which are available in several sizes, are simple to use. The volume of liquid to be dispensed is mechanically set; the syringe plunger is lifted for filling and pressed downward for dispensing. Hold the receiving container under the spout while depressing the plunger. Touch off the last drop on the inside wall of the receiving container.

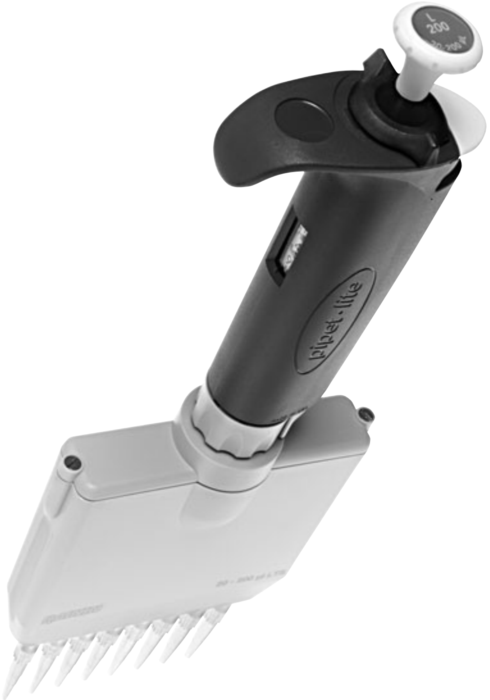

FIGURE 1.10 A multichannel pipettor. *Photo courtesy of Rainin Instrument, LLC., Oakland, CA. Pipet-Lite is a registered trademark of Rainin Instrument Company.* www.rainin.com.

E. STATISTICAL ANALYSIS OF EXPERIMENTAL DATA

The purpose of most biochemistry laboratory exercises is to observe and measure characteristics of a biomolecule or a biological process. The characteristic is often quantitative—a single number or a group of numbers. These measured quantities may be the molecular weight of a protein, the pH of a buffer solution, the absorbance of a colored solution, the rate of an enzyme-catalyzed reaction, the concentration of a protein in solution, or the radioactivity associated with a molecule. If you measure a quantitative characteristic many times under identical conditions, a slightly different result will most likely be obtained each time.

For example, if a radioactive sample is counted twice under identical experimental and instrumental conditions, the second measurement immediately following the first, the probability is very low that the numbers of counts will be identical. If the absorbance of a solution is determined several times at a specific wavelength, the value of each measurement will surely vary from the others. If an assay for cholesterol is performed several times on a blood serum sample from the same individual, the values will probably be close, but not all will be the same (see Study Problem 1.13). Which measurements, if any, are correct? Before this question can be answered, you must understand the source and treatment of numerical variations in experimental measurements.

Defining Statistical Analysis

An **error** in an experimental measurement is defined as a deviation of an observed value from the true value. There are two types of errors, **determinate** and **indeterminate.** Determinate errors are those that can be controlled by the experimenter and are associated with malfunctioning equipment, improperly designed experiments, and variations in experimental conditions. These are sometimes called human errors because they can be corrected or at least partially alleviated by careful design and performance of the experiment. Indeterminate errors are those that are random and cannot be controlled by the experimenter. Specific examples of indeterminate errors are variations in radioactive counting and small differences in the successive measurements of glucose in a serum sample.

Two statistical terms involving error analysis that are often used and misused are **accuracy** and **precision.** Precision refers to the extent of agreement among repeated measurements of an experimental value. Accuracy is defined as the difference between the experimental value and the true value for the quantity. Because the true value is seldom known, accuracy is better defined as the difference between the experimental value and the accepted true value. Several experimental measurements may be precise (that is, in close agreement with each other) without being accurate.

If an infinite number of identical, quantitative measurements could be made on a biosystem, this series of numerical values would constitute a **statistical population.** The average of all of these numbers would be the **true value** of the measurement. It is obviously not possible to achieve this in practice. The alternative is to obtain a relatively small **sample of data,** which is a subset of the infinite population data. The significance and precision of these data are then determined by statistical analysis.

Most quantitative biological measurements can be made in duplicate, triplicate, or even quadruplicate, but it would be impractical and probably a waste of time and materials to make numerous determinations of the same measurement. Rather, when you perform an experimental measurement in the laboratory, you will collect a small sample of data from the population of infinite values for that measurement. To illustrate, imagine that an infinite number of experimental measurements of the pH of a buffer solution are made, and the results are written on slips of paper and placed in a container. It is not feasible to calculate an average value of the pH from all of these numbers, but it is possible to draw five slips of paper, record these numbers, and calculate an average pH. By doing this, you have collected a sample of data. By proper statistical manipulation of this small sample, it is possible to determine whether it is representative of the total population and the amount of confidence you should have in these numbers.

The Mean, Sample Deviation, and Standard Deviation

Radioactive decay with emission of particles is a random process. It is impossible to predict with certainty when a radioactive event will occur. Therefore, a series of measurements made on a radioactive sample will result in a series of different count rates, but they will be centered around an **average** or **mean** value of counts per minute. Table 1.1 contains such a series of count rates obtained with a scintillation counter on a single radioactive sample. A similar table could be prepared for other biochemical measurements, including the rate of an

TABLE 1.1 The Observed Counts and Sample Deviation from a Typical Radioactive Sample

Counts per Minute	Sample Deviation $x_i - \bar{x}$
1243	+21
1250	+28
1201	−21
1226	+4
1220	−2
1195	−27
1206	−16
1239	+17
1220	−2
1219	−3
Mean = 1222	

enzyme-catalyzed reaction or the protein concentration of a solution as determined by the Bradford method. The arithmetic average, or mean, of the numbers is calculated by totaling all the experimental values observed for a sample (the counting rates, the velocity of the reaction, or protein concentration) and dividing the total by the number of times the measurement was made. The mean is defined by Equation 1.1.

>> $$\bar{x} = \frac{\sum^{n} x_i}{n} \quad \textbf{(Eq. 1.1)}$$

where

$\bar{x}$ = arithmetic average or mean
x_i = the value for an individual measurement
n = the total number of experimental determinations
$\sum$ = sum of all the values

The mean counting rate for the data in Table 1.1 is 1222. If the same radioactive sample were again counted for a series of ten observations, that series of counts would most likely be different from those listed in the table, and a different mean would be obtained. If we were able to make an infinite number of counts on the radioactive sample, then a **true mean** could be calculated. The true mean would be the actual amount of radioactivity in the sample. Although it would be desirable, it is not possible experimentally to measure the true mean. Therefore, it is necessary to use the average of the counts as an approximation of the true mean and to use statistical analysis to evaluate the precision of the measurements (that is, to assess the agreement among the repeated measurements).

Because it is not usually practical to observe and record a measurement many times as in Table 1.1, what is needed is a way to determine the reliability of an observed measurement. This may be stated in the form of a question. How close is the result to the true value? One approach to this analysis is to calculate the **sample deviation,** which is defined as the difference between the value for an observation and the mean value, $\bar{x}$ (Equation 1.2). The sample deviations are also listed for each count in Table 1.1.

>> $$\text{Sample deviation} = x_i - \bar{x} \qquad \textbf{(Eq. 1.2)}$$

A more useful statistical term for error analysis is **standard deviation,** a measure of the spread of the observed values. Standard deviation, s, for a sample of data consisting of n observations may be estimated by Equation 1.3.

>> $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \qquad \textbf{(Eq. 1.3)}$$

It is a useful indicator of the probable error of a measurement. Standard deviation is often transformed to **standard deviation of the mean** or **standard error.** This is defined by Equation 1.4, where n is the number of measurements.

>> $$s_m = \frac{s}{\sqrt{n}} \qquad \textbf{(Eq. 1.4)}$$

It should be clear from this equation that as the number of experimental observations becomes larger, s_m becomes smaller, or the precision of a measurement is improved.

Standard deviation may also be illustrated in graphical form (see Figure 1.11). The shape of the curve in Figure 1.11 is closely approximated by the **Gaussian distribution** or **normal distribution curve.** This mathematical

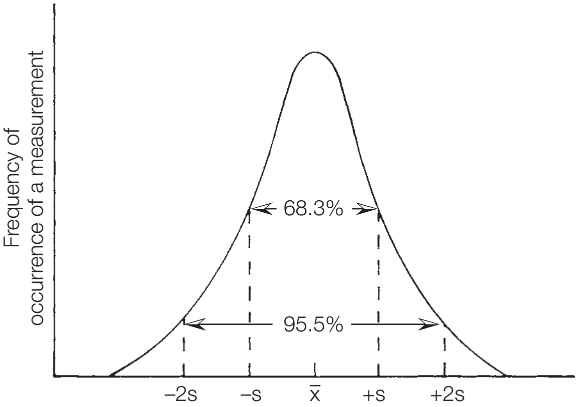

FIGURE 1.11 The normal distribution curve.

treatment is based on the fact that a plot of relative frequency of a given event yields a dispersion of values centered about the mean, $\bar{x}$. The value of $\bar{x}$ is measured at the maximum height of the curve. The normal distribution curve shown in Figure 1.11 defines the spread or dispersion of the data. The probability that an observation will fall under the curve is unity, or 100%. By using an equation derived by Gauss, it can be calculated that for a single set of sample data, 68.3% of the observed values will occur within the interval $\bar{x} \pm s$, 95.5% of the observed values within $\bar{x} \pm 2s$, and 99.7% of the observed values within $\bar{x} \pm 3s$. Stated in other terms, there is a 68.3% chance that a single observation will be in the interval $\bar{x} \pm s$.

For many experiments, a single measurement is made, so a mean value, $\bar{x}$, is not known. In these cases, error is expressed in terms of s, but is defined as the **percentage proportional error,** $\%E_x$, in Equation 1.5.

>> $$\%E_x = \frac{100k}{\sqrt{n}} \qquad \textbf{(Eq. 1.5)}$$

The parameter k is a proportional constant between E_x and the standard deviation. The percent proportional error may be defined within several probability ranges. **Standard error** refers to a confidence level of 68.3%; that is, there is a 68.3% chance that a single measurement will not exceed the $\%E_x$. For standard error, $k = 0.6745$. **Ninety-five hundredths error** means there is a 95% chance that a single measurement will not exceed the $\%E_x$. The constant k then becomes 1.45.

The previous discussion of standard deviation and related statistical analysis placed emphasis on estimating the reliability or precision of experimentally observed values. However, standard deviation does not give specific information about how close an experimental mean is to the true mean. Statistical analysis may be used to estimate, within a given probability, a range within which the true value might fall. The range or confidence interval is defined by the experimental mean and the standard deviation. This simple statistical operation provides the means to determine quantitatively how close the experimentally determined mean is to the true mean. **Confidence limits** (L_1 and L_2) are created for the sample mean as shown in Equations 1.6 and 1.7.

>> $$L_1 = \bar{x} + (t)(s_m) \qquad \textbf{(Eq. 1.6)}$$

>> $$L_2 = \bar{x} - (t)(s_m) \qquad \textbf{(Eq. 1.7)}$$

where

t = a statistical parameter that defines a distribution between a sample mean and a true mean

The parameter t is calculated by integrating the distribution between percent confidence limits. Values of t are tabulated for various confidence limits (Table 1.2). Each column in the table refers to a desired confidence level (0.05 for 95%, 0.02 for 98%, and 0.01 for 99% confidence). The table also includes the term **degrees of freedom,** which is represented by $n - 1$, the number of experimental observations minus 1. The values of $\bar{x}$ and s_m are calculated as previously described in Equations 1.1 and 1.4.

TABLE 1.2 Values of *t* for Analysis of Statistical Confidence Limits

	Probability of Larger Value of t, Sign Ignored						
d.f.	**0.05**	**0.02**	**0.01**	**d.f.**	**0.05**	**0.02**	**0.01**
1	12.706	31.821	63.657	14	2.145	2.624	2.977
2	4.303	6.096	9.925	15	2.131	2.602	2.947
3	3.182	4.541	5.841	16	2.120	2.583	2.921
4	2.776	3.747	4.604	17	2.110	2.567	2.898
5	2.571	3.365	4.032	18	2.101	2.552	2.878
6	2.447	3.143	3.707	19	2.093	2.539	2.861
7	2.365	2.998	3.499	20	2.086	2.528	2.845
8	2.306	2.896	3.355	21	2.080	2.518	2.831
9	2.262	2.821	3.250	22	2.074	2.508	2.819
10	2.228	2.764	3.169	23	2.069	2.500	2.807
11	2.201	2.718	3.106	24	2.064	2.492	2.797
12	2.179	2.681	3.055	25	2.060	2.485	2.787
13	2.160	2.650	3.012				

Spreadsheet Statistics

It is common practice today to use computer spreadsheet programs for statistical analysis of biochemical data. A spreadsheet provides a means to collect and enter data in the form of numbers and text. Perhaps the most versatile and easy-to-use spreadsheet software is Microsoft Excel, although more specialized statistical software programs including SPSS and SyStat are also very useful (see Chapter 2 and Appendix I). Using Excel to estimate statistical terms for experimental data is relatively straightforward. Launching the Microsoft Excel program on your computer brings up the Excel spreadsheet, which consists of rows (number headings) and columns (letter headings). More detailed instructions for the statistical applications of Excel are found at www.microsoft.com.

Statistical Analysis in Practice

The equations for statistical analysis that have been introduced in this section are of little value if you have no understanding of their practical use, meaning, and limitations. A set of experimental data will first be presented, and then several statistical parameters will be calculated using the equations. This example will serve as a summary of the statistical formulas and will also illustrate their application.

STUDY EXERCISE 1.3 Statistical Analysis of Data

Ten identical protein samples were analyzed by the Bradford method for protein analysis. The following values for protein concentration were obtained.

Observation Number	Protein Concentration (mg/mL), x
1	1.02
2	0.98
3	0.99
4	1.01
5	1.03
6	0.97
7	1.00
8	0.98
9	1.03
10	1.01

Sample mean

$$\bar{x} = \frac{\sum x}{n} = \frac{10.02}{10} = 1.00\ \text{mg/mL}$$

Sample deviation

$$\text{Sample deviation} = x_i - \bar{x}$$

Observation	$x_i - \bar{x}$
1	+0.02
2	−0.02
3	−0.01
4	+0.01
5	+0.03
6	−0.03
7	0.00
8	−0.02
9	+0.03
10	+0.01

Calculation of the sample deviation for each measurement gives an indication of the precision of the determinations.

(*Continued*)

Standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

$$s = 0.02$$

The mean can now be expressed as $\bar{x} \pm s$ (for this specific example, 1.00 ± 0.02 mg/mL). The probability of a single measurement falling within these limits is 68.3%. For 95.5% confidence (2*s*), the limits would be 1.00 ± 0.04 mg/mL.

Standard error of the mean

$$s_m = \frac{s}{\sqrt{n}}$$

$$s_m = \frac{0.02}{\sqrt{10}}$$

$$s_m = 0.006$$

This statistical parameter can be used to gauge the precision of the experimental data.

Confidence limits

The desired confidence limits will be set at the 95% confidence level. Therefore, we will choose a value for t from Table 1.2 in the column labeled $t_{0.05}$ and $n - 1 = 9$.

$$\begin{aligned}
L_1 &= \bar{x} + (t_{0.05})(s_m) \\
L_2 &= \bar{x} - (t_{0.05})(s_m) \\
s_m &= 0.006 \\
t_{0.05} &= 2.262 \\
\bar{x} &= 1.00 \\
L_1 &= 1.00 + (2.262)(0.006) \\
L_1 &= 1.01 \\
L_2 &= 0.99
\end{aligned}$$

We can be 95% confident that the true mean falls between 0.99 and 1.01 mg/mL.

Study Problems

1. What personal protection items must be worn when handling glacial acetic acid?
2. Define each of the following terms.
   (a) OSHA
   (b) MSDS
   (c) Flowchart
   (d) Pasteur pipet
   (e) Purified water
   (f) Error
   (g) Standard deviation
   (h) Molarity
3. Draw a schematic picture of your biochemistry lab and mark locations of the following safety features: eyewashes, first-aid kit, shower, fire extinguisher, chemical-spill kits, and direction to nearest exit.

4. Describe how you would prepare a 1-liter aqueous solution of each of the following reagents:
(a) 1 *M* glycine
(b) 0.5 *M* glucose
(c) 10 m*M* ethanol
(d) 100 n*M* hemoglobin

5. Why is de-ionized (ion-exchange) water not suitable as a solvent for use in UV absorbance measurements?

6. Describe how you would prepare just 10 mL of each of the solutions in Problem 4.

7. If you mix 1 mL of the 1 *M* glycine solution in Problem 4a with 9 mL of water, what is the final concentration of this diluted solution in m*M*?

8. Convert each of the concentrations below to m*M* and μM
(a) 10 mg of glucose per 100 mL
(b) 100 mL of a solution 2% in alanine

9. You have just prepared a solution by weighing 20 g of sucrose, transferring it to a 1-liter volumetric flask, and adding water to the line. Calculate the concentration of the sucrose solution in terms of m*M*, mg/mL, and % (wt/vol).

10. The concentrations of cholesterol, glucose, and urea in blood from a fasting individual are listed below in units of mg/100 mL (sometimes called mg%). These are standard concentration units used in the clinical chemistry lab. Convert the concentrations to m*M*.
(a) cholesterol—200 mg%
(b) glucose—75 mg%
(c) urea—20 mg%

11. The following optical rotation readings were taken by a polarimeter on a solution of an unknown carbohydrate. Use Excel or other statistical software to estimate statistical terms.
(a) Calculate the sample mean.
(b) Calculate the standard deviation.

α_{obs} **(degrees)**			
+3.24	+3.20	+3.17	+3.25
+3.15	+3.21	+3.23	
+3.30	+3.19	+3.20	

(c) Calculate the 95% confidence levels for the measurement.

12. Describe how you would prepare 100 mL of a single solution containing all of the following reagents at the designated concentrations.
(a) 0.1 *M* NaCl
(b) 0.05 *M* glucose
(c) 5% wt/vol alanine
(d) 1 mg/mL urea

13. How would you prepare 1 liter of a 1 *M* ethanol solution in water? You do not have an analytical balance available, only a 1-liter volumetric flask and a 100-mL graduated cylinder. The density of ethanol is 0.789 g/mL.

14. A technician at a clinical laboratory received a blood serum sample for cholesterol analysis. In order to check the reliability of the procedure, the technician repeated the assay five times and obtained the results below. Use a spreadsheet statistical analysis to estimate the mean, standard deviation, variance, and standard error of the mean for the data. Total cholesterol in blood serum sample in mg/100 mL (mg%): 157, 154, 155, 152, 155.

Further Reading

PIPETTING

en.wikipedia.org/wiki/pipette
Pipets.

http://www.biology.lsu.edu/introbio/tutorial/Pipets/1208.pipet.html
Pipets and pipetting.

http://www.gilson.com
Information on automatic pipets, procedures for use, and hints.

http://www.rainin.com
Instruction manuals for operation of the Pipetman.

www.newton.dep.anl.gov/askasci/chem03/chem03588.htm
Proper pipet usage.

http://www.umd.umich.edu/casl/natsci/slc/slconline/MICRPIP/index.html
Pipets and pipetting.

PREPARATION OF SOLUTIONS

M. Caspers and E. Roberts-Kirchhoff, ***Biochem. Mol. Biol. Educ.*** **31,** 303–307 (2003). "An Undergraduate Biochemistry Laboratory Course with an Emphasis on Research Experience."

S. Kegley and J. Andrews, ***The Chemistry of Water*** (1997), University Science Books (Sausalito, CA). A discussion of water purity and analysis.

J. Risley, ***J. Chem. Educ.*** **68,** 1054–1055 (1991). "Preparing Solutions in the Biochemistry Lab."

www.labwater.com
Water purification.

http://www.thermo.com/com/cda/landingpage/0,10255,626,00.html
Water purification from Thermo Scientific.

SAFETY

R. Alaimo, ***Handbook of Chemical Health and Safety***, 1st ed. (2001), Oxford University Press (Cary, NC).

American Chemical Society, ***Safety in Academic Chemical Laboratories, Volume I: Accident Prevention for College and University Students; Volume II: Accident Prevention for Faculty and Administrators,*** 7th ed. (2003), ACS (Washington, DC).

M. Armour, ***Hazardous Laboratory Chemicals Disposal Guide,*** 3rd ed. (2003), CRC Press (Boca Raton, FL).

K. Barker, Editor, ***At the Bench: A Laboratory Navigator*** (2005), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY).

National Research Council, ***Prudent Practices in the Laboratory: Handling and Disposal of Chemicals*** (1995), National Academy Press (Washington, DC).

P. Patnaik, ***A Comprehensive Guide to the Hazardous Properties of Chemical Properties,*** 3rd ed. (2007), John Wiley & Sons (Hoboken, NJ).

www.sigma-aldrich.com
MSDS documents for all biochemical products sold.

http://www.osha.gov
Review of functions and regulatory procedures of OSHA.

www.jjkeller.com
Supplier of HMIS labels and information.

http://sitemaker.umich.edu/chemistry.ion/files/safety_rules_complete.doc
Safety in the biochemistry lab.

http://www.jce.divched.org/JCEWWW/Features/CERS
Chemical Education Resource Shelf, Safety and Waste Disposal in the list of textbooks and software for use in chemistry/biochemistry courses.

STATISTICAL ANALYSIS OF DATA

R. Boyer, ***Modern Experimental Biochemistry,*** 3rd ed. (2000), Benjamin-Cummings (San Francisco), pp. 18–25.

P. Meier and R. Zund, ***Statistical Methods in Analytical Chemistry,*** 2nd ed. (2000), John Wiley & Sons (New York).

C. S. Tsai, ***An Introduction to Computational Biochemistry*** (2002), John Wiley & Sons (New York), pp. 11–40.

M. Whitlock and D. Schluter, ***The Analysis of Biological Data,*** 1st ed. (2009), Roberts and Company (Greenwood Village, CO).

http://www.graphpad.com/prism/Prism.htm
Software for statistics and curve fitting.

http://www.statistics.com
Click on free Web-based software for data analysis.

www.sciencesoftware.com/product.php?productid=131
Statistics and curve fitting.

http://www.ruf.rice.edu/~bioslabs/tools/stats/stats.html
Statistics in the lab.

WRITING LABORATORY REPORTS AND COMMUNICATING SCIENCE

K. Barker, Editor, ***At the Bench: A Laboratory Navigator*** (2005), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY). Covers lab orientation, keeping a notebook, and lab procedures.

H. Beall and J. Trimbur, ***A Short Guide to Writing about Chemistry,*** 2nd ed. (2001), Benjamin-Cummings (San Francisco).

R. Boyer, ***Biochem. Mol. Biol. Educ.*** **31,** 102–105 (2003). "Concepts and Skills in the Biochemistry/Molecular Biology Lab."

M. Cargill and P. O'Connor, ***Writing Scientific Research Articles: Strategy and Steps*** (2009), Wiley-Blackwell.

A. Coghill and L. Garson, ***The ACS Style Guide: A Manual for Authors and Editors,*** 3rd ed. (2006). Oxford University Press (Cary, NC).

R. Day and B. Gastel, ***How to Write and Publish a Scientific Paper,*** 6th ed. (2006), Greenwood Publishing Group (Westport, CT).

H. Ebel, C. Bliefert, and W. Russey, ***The Art of Scientific Writing: From Student Reports to Professional Publications in Chemistry and Related Fields,*** 3rd ed. (2004), John Wiley & Sons (New York).

P. Frey, ***Biochem. Mol. Biol. Educ.*** **31,** 237–241 (2003). "Guidelines for Writing Research Papers."

M. Katz, ***From Research to Manuscript: A Guide to Scientific Writing,*** 2nd ed. (2009), Springer (New York).

J. Kovac and D. Sherwood, ***Writing Across the Chemistry Curriculum: An Instructor's Handbook*** (2001), Prentice-Hall (Upper Saddle River, NJ).

J. Matthews and R. Matthews, ***Successful Scientific Writing: A Step-by-Step Guide for the Biological and Medical Sciences,*** 3rd ed. (2007), Cambridge University Press (New York).

T. O'Connor and G. Holmquist, ***Biochem. Mol. Biol. Educ. 37***, 344–348 (2009). "Algorithm for Writing a Scientific Manuscript."

A. Penrose and S. Katz, ***Writing in the Sciences: Exploring Conventions of Scientific Discourse,*** 2nd ed. (2004), Benjamin-Cummings (San Francisco).

L. Tomaska, ***ASBMB Today***, pp. 20–21, May 2007, "Teaching Undergraduates to Write Scientific Papers."

I. Valiela, ***Doing Science: Design, Analysis, and Communication of Scientific Research*** (2001), Oxford University Press (New York).

http://blog.makezine.com/science_room/general/maintaining_a_laboratory_notebook
Maintaining a lab notebook.

http://www.swarthmore.edu/NatSci/cpurrin1/notebookadvice.htm
Advice on keeping a lab notebook.

http://rescentris.com/?gclid=CNWE2M3qhZ4CFRhfagodfQ4SpQ
CERF electronic lab notebook.

http://www.ce.umn.edu/~smith/supplements/poster/guide.htm
Preparing professional scientific posters.

http://www.swarthmore.edu/NatSci/cpurrin1/posteradvice.htm
Advice and template for designing scientific posters.

http://www.ruf.rice.edu/~bioslabs/tools/report/reportform.html
Writing a research paper.

http://www.thomsonscientific.com/cgi-bin/jrnlst/jlresults.cgi?PC=BB
List of biochemistry and biophysics journals.

http://www.biomedcentral.com/browse/journals
List of biochemistry, molecular biology, and medical journals.

http://www.lib.iastate.edu/collections/eresourc/biochem.html
List of electronic journals of interest to biochemists and molecular biologists.

http://helios.hampshire.edu/~apmNS/design/RESOURCES/HOW_READ.html
How to read a scientific paper.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

CHAPTER 2

USING THE COMPUTER AND INTERNET FOR RESEARCH IN BIOCHEMISTRY

A. What Is Research and How Is It Done in Biochemistry? 35
B. Using Computers in Biochemistry 38
C. Web Sites Useful in Biochemistry 40

Study Problems 48
Further Reading 49
Computer Glossary 50

A. WHAT IS RESEARCH AND HOW IS IT DONE IN BIOCHEMISTRY?

What Is Research?

Virtually all scientists are engaged in investigative activities that are broadly called **research.** We begin this discussion on research by defining the term and continuing with details of the steps that are usually followed in the process.

According to the Merriam-Webster Collegiate Dictionary, research is:

1. to search or investigate exhaustively
2. the collecting of information about a particular subject
3. studious inquiry or examination

A more realistic and functional definition comes from librarians:

> “Research is a process of investigation. An examination of a subject from different points of view. It’s not just a trip to the library to pick up a stack of materials, or

> picking the first five hits from a computer search. Research is a hunt for the truth. It is getting to know a subject by reading up on it, reflecting, playing with the ideas, choosing the areas that interest you and following up on them. Research is the way you educate yourself."[1]

The ultimate goal of all research is to seek the truth; however, the approach that is used can vary from discipline to discipline. For biochemists and other molecular life scientists, research is usually focused on elucidating the laws of nature or, in other words, addressing questions about natural phenomena. The approach taken in biochemistry research is application of the historic **"scientific method"** of discovery.

The Scientific Method

The stepwise process for a scientific research project follows the general scheme shown in Figure 2.1. One of the early originators of this organized body of techniques is thought to be Ibn al-Haythan (965–1039 A.D.; Latinized name, Alhazen), a physicist studying optics and living in Basra, Iraq. Many modifications of the process have been made over the centuries, but the key scientific premise has remained—**data must be collected by experiment and observation.**

Each of the steps in Figure 2.1 actually consists of several activities:

1. ***Select the topic.*** The topic for study is often one of strong interest to the scientist and one with which he or she may already have some familiarity. It is sometimes helpful to phrase the problem to solve as a question to which we do not yet know the answer. Some examples of research topics in biochemistry might be: (a) what is the detailed molecular mechanism for a particular enzyme?, (b) what is the base sequence of a particular gene?, (c) what is the function of a protein that has just been isolated?, (d) how can a new method for analysis of fatty acids in blood serum be designed?, and (e) how does a particular drug act?
2. ***Collect current information and resources about the topic.*** This step always begins by reading the **primary literature** in the field of study. (Primary literature refers to articles in research journals in which scientists describe their experiments, report data, and discuss important results.) When you do extensive background research, you start to become an expert in the field. The computer is often heavily used in this step. Refer to Part C (p. 40) in this chapter to learn how to use the Internet in your study of the biochemical literature. Background reading of the literature may also be done by consulting research journals in a library; however, journals are very expensive, so libraries may not have all journals available.
3. ***Propose hypotheses that may answer questions about your topic.*** The dictionary defines hypothesis as "a tentative assumption made in order to draw out and test its logical or empirical consequences." A hypothesis is an important and useful tool in research that helps lead you to answers to your questions. Proposing a hypothesis is guessing, but it is an "educated guess,"

---

[1] http://www.usg.edu/galileo/skills/unit01/infoage01_03.phtml

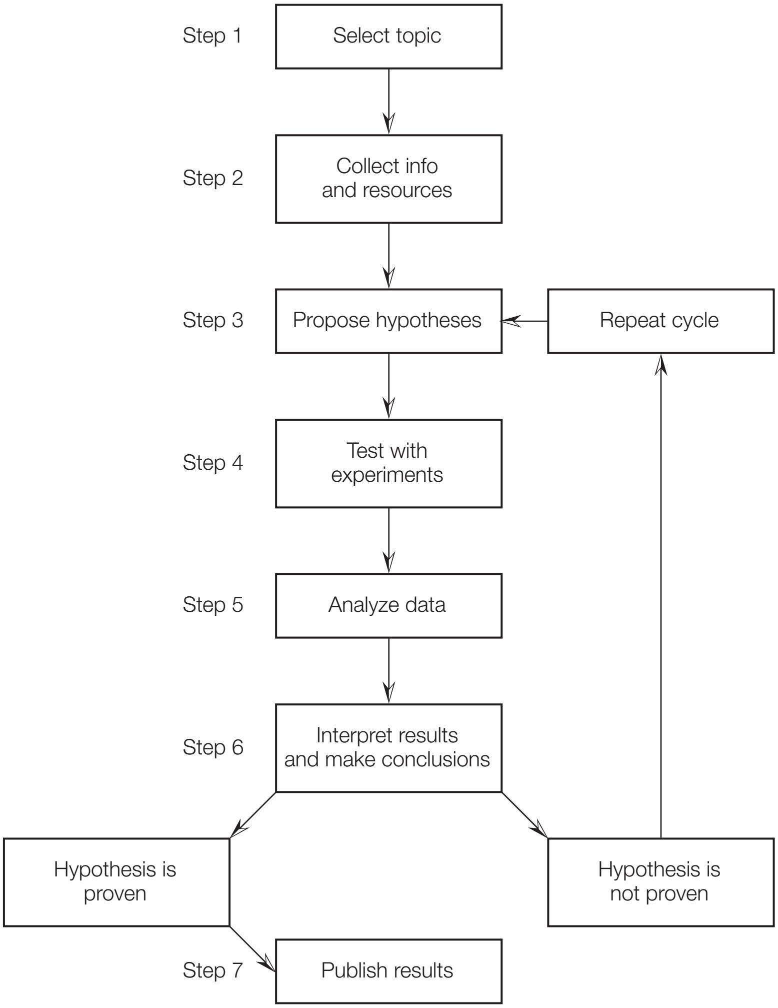
Step 1
Select topic
Step 2
Collect info and resources
Step 3
Propose hypotheses
Repeat cycle
Step 4
Test with experiments
Step 5
Analyze data
Step 6
Interpret results and make conclusions
Hypothesis is proven
Hypothesis is not proven
Step 7
Publish results

FIGURE 2.1 The steps followed in a scientific research project. The pathway is based on the premises of the historic scientific method.

as your reading of the literature will have enhanced your knowledge. Perhaps the most important reason for having hypotheses is that they tell you what experimental directions to take next on your project. (A hypothesis cannot be proven by experimental data, only supported. This will be discussed in more detail later.)

4. ***Design and perform experiments; collect data.*** For hypotheses to be tested properly, it is essential that you design experiments that provide results that are informative. Experimental results must be observable, empirical, and measurable. The best experiments result from knowing your subject well and from asking good questions about what you need to find out about your research problem. It should be obvious that experiments must be completed with accuracy and care and that safe lab techniques are used. Experiments should be repeated several times, with the same results obtained each time before they can be accepted as valid.

5. ***Analyze data.*** Data from experiments may be quantitative or qualitative. Both types can be analyzed by using a computer to prepare figures, tables, and graphs. If your results are primarily numerical, then it will be necessary to carry out calculations followed by proper statistical analysis using Excel and other software (see Chapter 1, Section E, p. 23).
6. ***Interpret results and draw conclusions.*** Are your experimental results and interpretations consistent with at least one of your hypotheses? If so, then that hypothesis can be supported by the results. It could be a likely answer to your initial question. A hypothesis cannot be proven, only supported. **Any hypothesis that is not consistent with experimental results must be rejected.** If all hypotheses are inconsistent with results, then new hypotheses must be proposed and tested. This is shown in Figure 2.1 as a branch from Step 6 to Step 3. This is called an **iteration,** or repeating process. Research is a constantly ongoing, evolving process. It is always in a state of flux, but scientists hope that each step they take moves them closer to an answer to their original question.
7. ***Publish results in a scientific journal.*** This step makes your work available so it may be studied by scientists around the world. Scientists then have the opportunity to repeat your experiments, to check their validity, and to design new experiments to test the hypothesis. Making sure that your experiments are reproducible, especially by scientists in other labs, is an important element of the scientific method.

One problem that can sometimes occur with the scientific method is a lack of objectivity. Scientists can become biased, adopt a favorite hypothesis, and design only experiments supporting that hypothesis. A balanced and objective approach is a necessity in scientific research. In addition to objectivity, another criterion that scientists must adhere to is ethical behavior. Unethical practices, such as have been observed in the form of scientific misconduct (plagiarism, falsifying data, etc.), are a danger to scientific progress. (See the Wikipedia Web site for examples of ethics in research and scientific misconduct.) The currently controversial so-called "climategate" issue is one that some scientists say lacks objectivity and is marred by unnecessary secrecy and a failure to fully share data. The last Web site at the end of this chapter contains an article criticizing the scientific process used in studying changes in the world's climate.

B. USING COMPUTERS IN BIOCHEMISTRY

The modern computer has revolutionized the way we live. Not surprisingly, the computer has also changed the way we do scientific research. The computer has become an essential tool for the investigation of all aspects of biomolecules—their structure, function, reactions, and information. The need for computers in biochemistry and the related molecular life sciences is growing because of two factors: (1) the fields are becoming increasingly quantitative, so there is a requirement for complex and accurate calculations—the computer with its access to **Internet** software and application programs is ideal for this work; and (2) because of the ease of determining protein and nucleic acid sequences and structures, there has been a proliferation of biological information that needs to be organized, stored, and made readily available to researchers. Initially, computers in biochemistry were used in the routine tasks of word processing, graphing, and statistical analysis of research data. But when connected to the Internet, the computer's use is greatly expanded to include some of the following functions: (1) searching the biochemical literature for pertinent books and journal articles, (2) using software for analyzing experimental data, (3) accessing biological **databases** that provide nucleic acid and protein sequences and structures, (4) seeking research methodology, procedures, and protocols for lab work, and many more uses.

This application of computer technology to the analysis, management, and manipulation of biochemical data is sometimes considered a part of the broad field of **computational biochemistry.** An important subfield of computational biochemistry that we will use extensively in this text is **bioinformatics,** defined as the application of computer technology to the storage and use of biological data, especially protein and nucleic acid sequence and structure.

Your first encounter with a computer in the laboratory was probably while you were using an instrument that had a computer to control its operation, to collect data, and to analyze data. All major pieces of scientific equipment, including UV-VIS spectrophotometers, high-performance liquid chromatographs, gas chromatographs, nuclear magnetic resonance spectrometers, mass spectrometers, and DNA sequencers, are now computer controlled. But your use of the computer will not end when you leave the lab, as you will use it to prepare laboratory reports including graphical and statistical analysis of your experimental data.

The primary purpose of this chapter is to learn how to solve problems in biochemistry using computer and software technology. Here we will outline general Internet resources that may be used in later chapters. Each chapter of this text will also describe more specific Internet resources that can be applied to the topics in that chapter. It is important for your education and career that you become knowledgeable and skilled in the use of the computer and Internet. Many specialized and some new terms are introduced in this chapter. Most of the words in bold in Sections B and C on computers are defined in a glossary at the end of the chapter.

Accessing the Internet

It is probably a correct assumption that those reading this book either have their own computer system or have access to computers where they study or work. Therefore, details for setting up and using a computer will not be provided here. Questions regarding computer use may be addressed to your instructor, supervisor, or local computer services office. If you are a student at an academic institution or a worker in a biotech-related lab, you will most likely be provided access to the Internet either through **Ethernet** or by a wireless system. Ethernet is a widely-used, wired networking system appropriate for local area networks (LAN). (Low-cost adaptors, called wireless bridges, are available to make any wired Ethernet-equipped device a part of a wireless network.) Some recommendations for specific hardware and software will be given in this book, but one must be aware that new products and important upgrades are continually being developed. Many of the software programs described here are **freeware,** software available without charge.

For word processing (writing lab reports, etc.), the most popular program is Microsoft Word. Software specialized for scientific and technical writing is available, but probably not necessary at this level. Current software programs for statistical analysis, graphing, and spreadsheets with graphing capability include Excel, Sigmaplot, SPSS, and SyStat (see Chapter 1, Section E, p. 23).

In order to access the Internet, you will need a **Web browser,** which is an interface program that reads **hypertext** and displays **Web pages** on your computer. A survey of computer users in 2010 found the most popular Web browsers to be **Firefox** (Mozilla), **Chrome** (Google), **Internet Explorer** (Microsoft), and **Safari** (Apple).

The World Wide Web

A heavily used and rapidly growing component of the Internet is the **World Wide Web** (WWW, also called "the Web"). This facility, which was launched in 1992, permits the transfer of data as pages in **multimedia form** consisting of text, graphs, figures, audio, and video. Each **Web site,** which has its own address (called a **Uniform Resource Locator, URL),** is composed of **Web pages.** The pages are linked together by **hypertext** pointers so that data stored on computers (**servers**) in different locations may be retrieved via the network by your computer. Web documents are written in a special coded language called **Hypertext Markup Language (HTML).** Typing the URL into the browser dialogue box will bring you to the **home page** or starting point for the Web site. One important feature you may note is that some text on the page is emphasized with underlining, in a different color, or perhaps highlighted in some other way. If you click the mouse on this text (called **hyperlinks**), your computer will connect to another related Web page that provides information on the hyperlink. This feature greatly enhances the use of the Web because related Web sites are connected or linked together and may be quickly accessed by a click of the mouse rather than by typing in a different URL.

C. WEB SITES USEFUL IN BIOCHEMISTRY

Directories, Library Resources, Databases, and Tools

Because of the complexity and the enormous number of Web sites available, one can easily become overwhelmed. Therefore, three tables, Tables 2.1–2.3, are presented that list Web sites selected for study because they are thought to be especially instructive for beginning students in biochemistry and other molecular life sciences. Each site in the tables has a short descriptive note that defines its contents. Table 2.1 consists of Web addresses that contain directories, catalogs, and library resources. Another way to describe Table 2.1 is to say "these addresses are Web sites that help you find other Web sites." This is a good place to start if you want to become familiar with the many resources available. Table 2.2 lists Web sites that have databases and tools that are more specialized in their function. Table 2.3 lists databases that allow searches using the name of a biochemical or class of biochemicals. These Web sites are especially valuable when seeking information and properties of biomolecules, including nomenclature, physical properties, structural data, reaction characteristics, spectroscopic information, and even chromatographic data.

TABLE 2.1 Web Directories and Library Resources

Name	Description	URL
Scirus	For scientific information	http://www.scirus.com/srsapp
Intute	Finding the best Web sites for biochemistry	http://www.intute.ac.uk/biologicalsciences
BioChemWeb	The virtual library of biochemistry, molecular biology, and cell biology	http://www.biochemweb.org
Protocol-Online	Your lab's reference book	http://www.protocol-online.org
University of New Mexico Center for Advanced Research Computing	Best biotech sites in the world	http://www.hpc.unm.edu/~aroberts/main/top5%25.htm
University of Sussex	Chemistry and Biochemistry Library Guide	http://guides.lib.sussex.ac.uk/chemistry
Iowa State University	Biochemistry Library Guide	http://www.lib.iastate.edu/collections/eresourc/biochem.html
Penn. State University	Biochemistry Library Guide	http://www.libraries.psu.edu/psul/researchguides/agbiohealth/bmb.html
Alverno College	Library Resources	http://depts.alverno.edu/library/science.html

TABLE 2.2 Biochemical Databases and Tools

Name	Description	URL
Protein Data Bank (PDB)	Protein structures determined by X-ray and NMR	http://www.rcsb.org/pdb
European Bioinformatics Institute (EBI)	DNA sequences	http://www.ebi.ac.uk
National Center for Biotechnology Information (NCBI)	Variety of databases and resources	http://www.nlm.nih.gov
Swiss-Protein	Protein sequences and analysis	http://www.expasy.ch/tools
The Institute for Genomic Research	Collection of genomic databases	http://www.ebi.ac.uk/genomes/index.html
REBASE—The Restriction Enzyme Database	Restriction enzyme directory and action	http://rebase.neb.com
NCBI Molecules to Go	Structures of biomolecules	http://molbio.info.nih.gov/cgi-bin/pdb
PyMol Molecular Viewer	Structures of biomolecules	http://pymol.org
Entrez browser of NCBI	Database searching including PubMed literature	http://www.ncbi.nlm.nih.gov/Entrez
Protein Information Resource (PIR)	Database searching for proteins	http://pir.georgetown.edu
Munich Information Center for Protein Sequences (MIPS)	Protein sequences	http://www.helmholtz-muenchen.de/en/mips

(continues)

TABLE 2.2 Continued

Name	Description	URL
Journal of Chemical Education Resource Shelf	List of all current biochemistry texts	http://www.jce.divched.org/JCEWWW/Features/CERS
Chimera Molecular Modeling System (UCSF)	Advanced molecular modeling	http://www.cgl.ucsf.edu/chimera
BiomoleculesAlive (ASBMB)	Instructional material	http://www.biomoleculesalive.org
ChemDraw	Chemical structure drawing program	http://www.cambridgesoft.com
MEDLINE (PubMed)	U.S. National Library of Medicine	http://www.nlm.nih.gov

TABLE 2.3 Databases of Biochemical Compounds

Name	Description	URL
International Union of Biochemistry and Molecular Biology (IUBMB)	Nomenclature	http://www.chem.qmul.ac.uk/iubmb
IUBMB Enzyme List	Catalog of enzymes	http://www.chem.qmul.ac.uk/iubmb/enzyme
ChemBioFinder	Structures and properties	http://www.chembiofinder.com
Worthington Enzyme Manual	Properties of enzymes	http://www.worthington-biochem.com/index/manual.html
Enzyme Database of ExPASy	Enzyme names and numbers	http://www.expasy.ch/enzyme
BRENDA	A comprehensive enzyme information system	http://www.brenda-enzymes.info

An annoying aspect of using Web sites is that they can disappear from the Internet. In fact, this may happen to some of the sites in this book. This process, sometimes called “Web rot,” can happen when a link becomes broken or irrelevant over time. If the reason for a Web site disappearance is due to a new address, you will usually be directed to the new site.

Many of the current Web sites you will need for your work and study are listed in the tables here; however, what about new Web sites that have been established since publication of this book? Millions of new sites are created every year. To access these new sites, you need the help of a **search engine,** a searchable directory that organizes Web pages by subject classification. The most widely used search engines are **Google** (http://www.google.com) and **Yahoo** (http://www.yahoo.com). Use of a search engine requires the input of a keyword or terms for searching. As you surf the Web, you may find sites you wish to save and review at a later date. Different Web browsers use different terms to define this saving process. Some use **favorites** and some use **bookmarks.**

STUDY EXERCISE 2.1 Using Directory Web Sites

List five of the "Best Biotech Web Sites in the World" found on the Web site from The University of New Mexico Center for Advanced Research Computing.

Solution:

4Biotech, Access Excellence, AgBiotechNet, biochemist.com, BiochemNet

STUDY EXERCISE 2.2 Using Protocol-Online

Study the home page for Protocol-Online. What are the subtopics available under the main heading of "Biochemistry"?

Solution:

On the home page, click on "Biochemistry." Here you will find hyperlinks to three subtopics: amino acids and proteins, enzyme analysis, and lipids.

STUDY EXERCISE 2.3 IUBMB Enzyme Web Site

Use the IUBMB enzyme site to find the E.C. (Enzyme Commission) number for the enzyme cellulase.

Solution:

Cellulase catalyzes the hydrolysis of the glycosidic bonds between glucose in the substrate cellulose. The E.C. number is 3.2.1.4.

Viewing Structures of Biomolecules

Determining the structure of biomolecules has now become a relatively routine activity in research labs. Thousands of structures experimentally determined by X-ray crystallography, NMR, and electron microscopy are now available on selected Web sites. The structures include those of proteins, nucleic acids, and nucleic acid/protein complexes. Some of the most accessible and useful sites for viewing structures include:

- Protein Data Bank (http://www.rcsb.org)
- Molecules to Go (http://molbio.info.nih.gov/cgi-bin/pdb)
- PyMol Molecular Viewer (http://pymol.org)

The Protein Data Bank and Molecules to Go sites are freeware, but PyMol is a user-sponsored system and requires a subscription. Academic institutions may purchase a site license for an individual, lab and classroom, or complete department.

To obtain a structure, access the Web site home page and type the name of the biomolecule desired into the dialogue box. Instructions for further analysis of the structure are given on all sites.

STUDY EXERCISE 2.4 Protein Data Bank

Use the Protein Data Bank Web site to find a structure for the enzyme cellulase. What kinds of secondary structures do you see in the cellulase molecule? What organism was the source of the cellulase you looked at?

Solution:

As of July 20, 2010, there were structures of 177 different cellulases on the PDB site. On the same date, there was a total of 66,633 structures of biomolecules on the PDB site. Each cellulase at the site will have different percentages of α-helices and β sheets. Describe the structure you looked at.

Searching the Biochemical Literature

An introduction to scientific journals and publishing a research article in a journal was given in Chapter 1 (see Section B, p. 6). Figure 1.5 provides an extensive list of journals that publish research articles in biochemistry. Generating research ideas and maintaining a research lab require an extensive knowledge of the biochemical literature (see Step 2, Figure 2.1). Much of this literature is now accessible on the Internet, but occasionally it may be necessary to use reference books in the library. The biochemical literature is massive and expanding rapidly. It is almost a full-time job just to maintain a current awareness of a specialized research area. There are few disciplinary boundaries in the study of biochemistry. The biochemical literature overlaps into the biological sciences, the physical sciences, the basic medical sciences, and information technology. The intent of the following discussion is to bring order to the many sources of information including textbooks, reference books, research journals, and Internet Web sites that are available.

Textbooks

The student's first exposure to biochemistry is usually a formal class accompanied by the reading of a general textbook of biochemistry. By providing an in-depth survey of biochemistry, textbooks allow students to build a strong foundation of important principles and concepts. By the time most books are in print, the information is one to two years old, but textbooks should still be considered the starting point for mastery of the fundamentals of biochemistry. Many textbooks are now available on-line and thus can be occasionally updated by their authors, or a Web site may be established for addition of new material.

Reference Books

For more specialized and detailed biochemical information that is not offered by textbooks, reference books must be used. Reference works range from general surveys to specialized series. The best works are multivolume sets that continue publication of volumes on a periodic basis. Each volume usually covers a specialized area with articles written by recognized authorities in the field. It should be noted that reference articles of interest to biochemists are often found in publications that are not strictly biochemical. The best known and most widely used review publication is *Annual Review of Biochemistry*. Each volume in this series, which was introduced in 1932, contains several detailed and extensive articles written by experts in the field. For shorter reviews emphasizing current topics, *Trends in the Biochemical Sciences* (TIBS) is widely read. The active researcher has a continuing need for new methods and techniques. Several publications specialize in providing details of research methods, and many research methods are now available on the Web.

Research Journals

The core of the biochemical literature consists of peer-reviewed research journals (primary literature). It is essential for a practicing biochemist to maintain a knowledge of biochemical advances in his or her field of research and related areas. Scores of research journals are published with the intent of keeping scientists up to date. Some research journals have achieved an especially excellent reputation, and articles therein are considered to be of the highest quality. A recent ranking of the biochemical journals, based on the number of citations received, produced the following order for the top six: *Journal of Biological Chemistry, Biochimica et Biophysica Acta, Biochemistry, Proceedings of the National Academy of Sciences of the United States of America, Biochemical Journal*, and *Biochemical and Biophysical Research Communications.* The core journals used by an individual depend on the area of specialty and are best determined from experience (see Figure 1.5, p. 10).

With the expansion of scientific information has come the need for efficient storage and use of research journals. Most publishers now provide journals on-line, and libraries pay the subscription cost for the journals they want. Students, faculty, and researchers can then obtain articles free of charge. If your institution does not subscribe to a journal you need, it is possible to read article titles and abstracts on journal Web sites at no cost. If the complete text for an article is needed, usually there is a charge. The URLs for journals may be obtained by using a search engine.

Literature Searches on the Web

As you study and work in biochemistry, you will often need to complete a thorough literature search on some specialized area or topic. It is not practical to survey the hundreds of books, journals, and reports that may contain information related to the topic. Two publications that provide brief summaries of published articles, reviews, and patents are *Chemical Abstracts* and *Biological Abstracts.* If you are not familiar with the use of these abstracts, ask your instructor or reference librarian for assistance. Both of these abstracts are available on-line.

Research articles of interest to biochemists may appear in many types of research journals. Research libraries do not have the funds necessary to subscribe to every journal, nor do scientists have the time to survey every current journal copy for articles of interest. The computer revolution has reached into the chemical and biochemical literature, and most college and university libraries now subscribe to computer bibliographic search services. One such service is STN International, the scientific and technical information network. This on-line system allows direct access to some of the world's largest scientific databases. The STN databases of most value to life scientists include BIOSIS Previews/RN (produced by Bio Sciences Information Service; covers original research reports, reviews, and U.S. patents in biology and biomedicine) and CA (produced by Chemical Abstracts service, covers research reports in all areas of chemistry including biochemistry). These networks provide on-line service, and their databases can be accessed from personal computers in the office, laboratory, or library. Subscription costs are usually paid by the library.

Excellent Web sites for searching the biochemical literature using your personal computer are available as freeware. The most widely used site is:

- **PubMed** from the National Center for Biotechnology Information (http://www.pubmed.gov)

PubMed is a service of the NIH National Library of Medicine. In addition to the direct URL listed above, it may be accessed through the NCBI site (http://www.ncbi.nlm.nih.gov), National Library of Medicine site (http://www.nlm.nih.gov) or many other database Web sites. To begin a search, enter search terms in the dialogue box at the top of the home page. Search terms may be topics, author name, or journal name. Results are reported in a table, beginning with the most recent articles, with article titles, author(s), and journal reference.

STUDY EXERCISE 2.5 Searching the Biochemical Literature on PubMed

To illustrate the use of this search service, point your Web browser to the appropriate URL above for the PubMed home page. Many features on display are available, but the most basic is the search capability. For bibliographic searching, you may enter in the dialogue box a search term, author name, or journal name. For example, type in "human alpha-lactalbumin," a calcium-binding protein found in milk. Clicking on "Search" will then provide more than 1100 citations (or articles). The lists are composed of article title, author(s), and reference in reverse chronological order. By clicking on the highlighted title (in hypertext), you can retrieve the abstract of the article. Another useful and time-saving feature is the hypertext (see "Related Articles"). Clicking on this will provide a list of papers related to the specific citation. The 1100 papers or so that you obtained in your original search are too many to screen; you may change the search parameters to reduce the number. For example, in the citation listing you see some interesting articles describing that complexes formed between human α-lactalbumin and oleic acid have been reported to display apoptotic activity against tumor cells. You can narrow down the citation list by searching the topic "complexes between human alpha-lactalbumin and oleic acid." This yields about 12 results, a reasonable number of abstracts to read. Practice using PubMed with some terms you find interesting.

Sequence Homology in Proteins

Another activity that is valuable in biochemical research is to seek proteins that have different functions, but similar sequences (sequence homology). These investigations can be done at a couple of Web sites:

- Protein Data Bank (http://www.rcsb.org)
- National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov)

Both Web sites compare biomolecules to seek to find regions of sequence similarity.

STUDY EXERCISE 2.6 Using Databases

The application of the primary databases and structural analytical tools will be introduced using a protein, α-lactalbumin from human milk. We will search databases to find and view its primary and secondary structure and also determine if there are other proteins with a similar amino acid sequence and structure. After completion of these exercises, you will be able to apply these computer tools to proteins of your own choice. Point your Web browser to the Protein Data Bank (PDB) and the Research Collaboratory for Structural Bioinformatics (http://www.rcsb.org). Become acquainted with the PDB by viewing the home page and perhaps clicking on some hyperlinks. Type “human alpha-lactalbumin” in the dialogue box at the top of the page and click on “Search.” Your query will find at least 14 structure hits. An appropriate structure to work with is number 1B90: human α-lactalbumin, low temperature form. Click on the highlighted title for a structure and more information about the protein. On the top tool bar, click on “Sequence.” The protein has one polypeptide chain and 123 amino acid residues. The sequence is given using one-letter abbreviations for amino acids. The sequence also shows locations of α-helices (41%) and beta sheets (9%). For a comparison of this protein with others in the database, click on ”Sequence Similarity.” The resulting table provides the number of proteins with % sequence similarity ranging from 100% down to 30%. For example, there are three proteins that have 90% sequence homology and 577 proteins that have 40% sequence homology. Click on the 40% column to see a list of the proteins, most of which are lysozymes. It is interesting to note that the proteins α-lactalbumin and lysozyme have similar primary, secondary, and tertiary structures, but quite different biochemical activities. The two proteins may have been derived from a common ancestral gene. You could also have used the BLAST tool at the NCBI site for the same analysis of sequence.

Virtual Biochemistry Laboratories

For some colleges and universities around the world, offering a “real” biochemistry laboratory course for students is not a possibility. Some of the reasons for this include a lack of expensive scientific instruments, equipment, and reagents; a scarcity of appropriate lab space; a lack of staff expertise; and an inability to fit scheduled lab time into a tight curriculum. Alternative activities for students are now available in the form of “virtual biochemistry labs” on the Web (Table 2.4). Although simulated labs should never be considered as a replacement or substitute

TABLE 2.4 Virtual Biochemistry Labs

Name	Description	URL
St. Olaf College	Virtual Lab Manual	http://www.stolaf.edu/people/muth/Labmanual.html
Carnegie Mellon University	Biology labs	http://telstar.ote.cmu.edu/biology/lab/index.html
Lund University	Labs in bioanalytical chemistry	http://www.drylabs.com
Changbioscience	Virtual labs for biologists	http://www.changbioscience.com/virtualab.html
HHMI Interactive	Virtual labs in biology	http://www.hhmi.org/biointeractive

for the real thing, especially for student majors in biochemistry, molecular biology, other biological sciences, and chemistry, they may be suitable for some students who will never work in a lab or manipulate lab data. In addition, viewing experiments and procedures in virtual labs may actually offer benefits to students who will perform similar experiments in the lab. The virtual lab may be considered a "prelab," where students will become acquainted with equipment, techniques, and procedures and can more efficiently conduct real experiments in the lab. This experience is especially important when students are dealing with expensive equipment and reagents. The virtual labs may also be considered a safety feature—students can receive training regarding the use of dangerous reagents and procedures.

Study Problems

1. Use PubMed or an enzyme Web site to answer the following questions about the enzyme tyrosinase.
   (a) What are sources of the enzyme besides mushroom?
   (b) What metal ion is present in the native enzyme?
   (c) Find two references for the study of the inhibition of tyrosinase. What inhibitor molecules have been investigated?
   (d) Find another substrate for the enzyme besides L-DOPA.
   (e) Search for the three-dimensional structure of the enzyme.
2. Use PubMed bibliographic searches to find two recent research articles authored by Thomas R. Cech, who won the Nobel Prize in Chemistry for the discovery of catalytic RNA (ribozymes). Write brief summaries of the articles.
3. The technique of immobilized metal-ion affinity chromatography (IMAC) is widely used to purify proteins. Find two proteins that have recently been purified by this procedure.
4. The Western blot procedure is now used to test human serum for the presence of antibodies to the AIDS virus. Find two publications that describe procedures for this assay.
5. Use the REBASE site to determine the specificity of the restriction enzyme *Hind*III.
6. Use the techniques in Study Exercise 2.6 to find the amino acid sequence of the protein recombinant goat α-lactalbumin. What is the percent of α-helices and β-sheet secondary structure?
7. What is the biological function of α-lactalbumin?

8. Study the nucleotide sequence for the human α-lactalbumin gene. Hint: Begin at the NCBI home page and enter Entrez. Click on "Nucleotides" and do a search for the protein. Review the GenBank report for a position of introns and exons. Obtain a FASTA report, download the files, and complete a BLAST search for related sequences.

9. The December 2009 issue of *The Scientist* magazine listed the ten most exciting biotech tools to be introduced in 2009. One of the top ten is "protein expression using synthesized genes." Write a paragraph of about 100 words that describes the new laboratory technique. Use computer bibliographic searches to learn about this new technology.

10. Use the BLAST tool to compare the amino acid sequences for human α-lactalbumin and lysozyme. Repeat the process with the use of BLAST to compare the nucleotide sequences for the genes coding for α-lactalbumin and lysozyme.

Further Reading

A. Baxevanis and B. Ouellette (Editors), ***Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins,*** 3rd ed. (2004), Wiley Interscience.

J. Boyle, ***Biochem. Mol. Biol. Educ.* 32,** 236–238 (2004). "Bioinformatics in Undergraduate Education."

S. Cooper, ***Biochem. Mol. Biol. Educ.* 29,** 167–168 (2001). "Integrating Bioinformatics into Undergraduate Courses."

L. Furge et al., ***Biochem. Mol. Biol. Educ.* 37,** 26–36 (2009). "Vertical and Horizontal Integration of Bioinformatics Education."

J. Gu and P. Bourne (Editors), ***Structural Bioinformatics (Methods of Biochemical Analysis),*** 2nd ed. (2009), Wiley-Blackwell.

C. Hodgman, A. French, and D. Westhead, ***Instant Notes in Bioinformatics,*** 2nd ed. (2009), Taylor & Francis.

H. Höltje, et al., ***Molecular Modeling: Basic Principles and Applications,*** 3rd ed. (2008), John Wiley & Sons.

R. Kaspar, ***Biochem. Mol. Biol. Educ.* 30,** 36–39 (2002). "Integrating Internet Assignments into a Biochemistry/Molecular Biology Laboratory Course."

L. Kirkup, ***Data Analysis with Excel*** (2002), Cambridge University Press (Cambridge, U.K.).

A. Lesk, ***Introduction to Bioinformatics,*** 3rd ed. (2008), Oxford University Press (New York).

J. Pevsner, ***Bioinformatics and Functional Genomics,*** 2nd ed. (2009), Wiley-Blackwell.

J. Ramsden, ***Bioinformatics: An Introduction,*** 2nd ed. (2009), Springer-Verlag (New York).

S. Rovner, ***Chem. Eng. News***, July 27, 2009, pp. 54–57, "Digital Textbooks."

I. Schomburg et al., ***Trends Biochem. Sci.* 27,** 54–56 (2002). "BRENDA: A Resource for Enzyme Data and Metabolic Information."

D. Sheehan, ***Physical Biochemistry: Principles and Applications, 2nd ed.*** (2009), Wiley-Blackwell (Hoboken, NJ). Bioinformatics—Chapter 9, pp. 305–346.

C. St. Clair and J. Visick, ***Exploring Bioinformatics: A Project-Based Approach*** (2009), Jones & Bartlett.

C. Tsai, ***Computational Biochemistry*** (2002), Wiley-Liss (New York).

L. Wang, ***Chem. Eng. News***, September 7, 2009, pp. 97–99, "Learning Chemistry Online."

D. Weisman, ***Biochem. Mol. Biol. Educ.* 38,** 4–9 (2010). "Incorporating a Collaborative Web-based Virtual Laboratory in an Undergraduate Bioinformatics Course."

X. Zhang, ***Biochem. Mol. Biol. Educ.* 37,** 16–23 (2009), "Using Arabidopsis Genetic Sequences to Teach Bioinformatics."

M. Zvelebil and J. Baum, ***Understanding Bioinformatics*** (2007), Garland Science.

http://www.jce.divched.org/JCEWWW/Features/CERS
The Journal of Chemical Education, Chemical Education Resource Shelf. For a complete list of current general biochemistry texts and lab manuals, scroll down to "Biochemistry."

http://www.biointeractive.org
Virtual biology labs sponsored by Howard Hughes Medical Institute.

http://www.vts.intute.ac.uk/
Helping you find the best Web sites for study and research.

http://www.usg.edu/galileo/skills/unit01/infoage01_03.phtml
What is research? The Online Library Learning Center of the University System of Georgia.

http://ugresearch.adm.ohio-state.edu/research.html
What is undergraduate research? From The Ohio State University Undergraduate Research Office.

http://en.wikipedia.org/wiki/Research
Research defined; from Wikipedia.

http://www.sciencebuddies.org/mentoring/project_scientific_method.shtml
Steps of the scientific method.

http://en.wikipedia.org/wiki/Scientific_method
The scientific method defined.

http://en.wikipedia.org/wiki/Scientific_misconduct
Ethics and scientific misconduct.

http://helios.hampshire.edu/~apmNS/design/RESOURCES/HOW_READ.html
How to read a scientific paper.

http://www.changbioscience.com/virtualab.html
Bioinformatics tools and resources for biomedical scientists and students.

http://www.ft.com/cms/s0/8aefbf52-d9e1-11de-b2d5-00144feabdc0.html
Comment on the "Climategate" issue.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

Computer Glossary

bioinformatics the use of computer technology to solve biological problems.
biological databases computer sites that organize, store, and disseminate files that contain information consisting of literature references, nucleic acid sequences, protein sequences, and protein structures.
bookmark a function to save a Web site address for later use (Netscape Navigator).
computational biochemistry the application of computer technology to the analysis, management, and manipulation of biochemical data.
Ethernet a family of frame-based computer networking technologies for local area networks.
favorites the Internet Explorer form of a bookmark.
freeware software that is provided free of charge by its developer.

home page the beginning page for access to a Web site.

HTML HyperText Markup Language; a special, coded language that is used to write Web pages.

hyperlink link or connection between related Web pages.

hypertext a language that connects similar documents on the Web.

Internet the worldwide matrix that allows all computers and networks to communicate with each other.

multimedia several forms of media including text, graphics, video, and audio.

search engine a searchable directory that organizes Web pages by subject classification.

server a computer that acts as the storage site for retrievable data.

URL uniform resource locator; a standard address form that identifies the location of a document on the Internet.

Web site a collection of documents (Web pages) on a server.

WWW World Wide Web (“the Web”); a component of the Internet that uses a hypertext-based language to provide resources.

This page intentionally left blank

CHAPTER

3

GENERAL LABORATORY PROCEDURES

A. pH, Buffers, Electrodes, and Biosensors 53
B. Measurement of Protein Solutions 67
C. Measurement of Nucleic Acid Solutions 71
D. Techniques for Sample Preparation 73
E. Radioisotopes in Biochemistry 80

Study Problems 91
Further Reading 92

All biochemistry experimental activities, whether in the teaching or research laboratory or in an industrial biotech lab, are replete with techniques that must be carried out on an almost daily basis. This chapter outlines the theoretical and practical aspects of some of these general and routine procedures, including preparation of buffers, use of buffers, pH measurements, dialysis, membrane filtration, lyophilization, quantitative methods for protein and nucleic acid measurement, centrifugal vacuum concentration, and measurement of radioactive samples.

A. pH, BUFFERS, ELECTRODES, AND BIOSENSORS

Most biological processes in the cell take place in a water-based environment. Water is an **amphoteric** substance; that is, it may serve as a proton donor (acid) or a proton acceptor (base). Equation 3.1 shows the ionic equilibrium of water.

>> $$H_2O \rightleftharpoons H^+ + OH^- \qquad \textbf{(Eq. 3.1)}$$

In pure water, $[H^+] = [OH^-] = 10^{-7}M$; in other words, the pH or $-\log[H^+]$ is 7. [The pH scale runs from 0 (very acidic) to 14 (very basic)]. Acidic and basic molecules, when dissolved in water in a biological cell or test tube, react with either hydrogen ions or hydroxide ions to shift the equilibrium of Equation 3.1 and result in a pH change of the solution.

Biochemical processes occurring in cells and tissues depend on strict regulation of the hydrogen ion concentration. Natural acids and bases are often generated in cells by normal biological processes and they must be neutralized by buffers. Biological pH is maintained at a constant value by naturally-occurring buffer systems such as phosphate. When biological processes are studied *in vitro,* artificial media must be prepared that mimic the cell's natural environment. Because of the dependence of biochemical reactions on pH, the accurate determination of hydrogen ion concentration has always been of major interest. Today, we consider the measurement and control of pH to be a simple and rather mundane activity. However, an inaccurate pH measurement or a poor choice of buffer can lead to failure in the biochemistry laboratory. You should become familiar with several aspects of pH measurement, electrodes, and buffers. A table of commonly used acids and bases is shown in Appendix II.

Measurement of pH

A pH measurement is usually taken by immersing a glass or plastic combination electrode into a solution and reading the pH directly from a meter. At one time, pH measurements required two electrodes, a pH-dependent glass electrode sensitive to H^+ ions and a pH-independent calomel reference electrode. The potential difference that develops between the two electrodes is measured as a voltage, as defined by Equation 3.2.

>> $$V = E_{\text{constant}} + \frac{2.303RT}{F}\text{pH} \qquad \textbf{(Eq. 3.2)}$$

where

V = voltage of the completed circuit
E_{constant} = potential of reference electrode
R = the gas constant
T = the absolute temperature
F = the Faraday constant

A pH meter is standardized with buffer solutions of known pH before a measurement of an unknown solution is taken. It should be noted from Equation 3.2 that the voltage depends on temperature. Hence, pH meters must have some means for temperature correction. Older instruments usually have a knob labeled "temperature control," which is adjusted by the user to the temperature of the measured solution. Newer pH meters automatically display a temperature-corrected pH value.

Using the pH Electrode

Most pH measurements today are obtained using a single **combination electrode** (Figure 3.1). Both the reference and the pH-dependent electrode are contained in

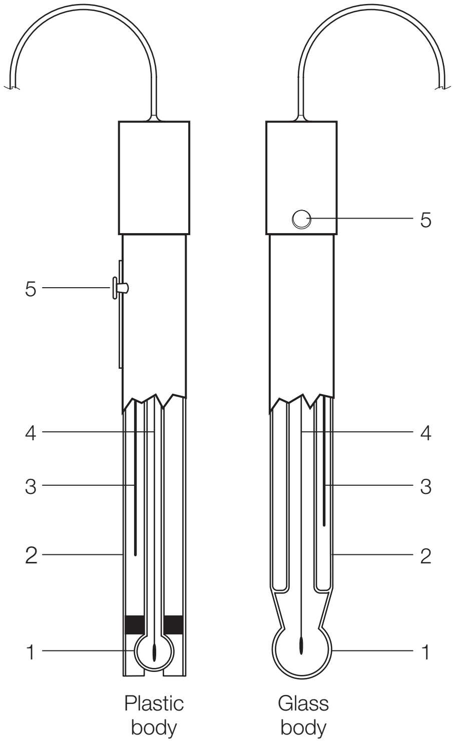

Electrodes are housed in either plastic or an all-glass body configuration. They can be either single cells or, as shown in the diagram, combined into one body for ease of use. Regardless of the configuration, there are several features common to all electrodes.

1. Sensing membrane glass: Performs actual measurement.
2. Reference junction: Acts as a liquid-path electrical conductor.
3. Internal reference: Supplies a constant equilibrium voltage.
4. pH internal element: Supplies a voltage based on the pH value of the sample.
5. Reference fill hole: Used to replace the reference electrolyte solution.

FIGURE 3.1 The combination pH electrode. *Courtesy of Hanna Instruments;* www.hannainst.com.

a single glass or plastic tube. Although these are more expensive than dual electrodes, they are much more convenient to use, especially for smaller volumes of solution. Using a pH meter with a combination electrode is relatively easy, but certain guidelines must be followed. A pH meter not in use is left in a “standby” position. Before use, check the level of saturated KCl in the electrode. If it is low, check with your instructor for the filling procedure. Turn the temperature control, if available, to the temperature of the standard calibration buffers and the test solutions. Be sure the function dial is set to pH. Lift the electrode out of the storage solution, rinse it with distilled water from a wash bottle, and *gently* clean and dry the electrode with a tissue. Immerse the electrode in a standard buffer. Common standard buffers are pH 4, 7, and 10 with an accuracy of ±0.02 pH unit. The standard buffer should have a pH within two pH units of the expected pH of the test solution. The bulb of the electrode must be completely covered with solution. Turn the pH meter to “on” or “read” and adjust the meter with the “calibration dial” (sometimes called “intercept”) until the proper pH of the standard buffer is indicated on the dial. Turn the pH meter to the standby position. Remove the electrode and again rinse with distilled water and carefully blot dry with tissue. Immerse the electrode in a standard buffer of different pH and turn the pH meter to “read.” The dial should read within ±0.05 pH unit of the known value. If it does not, adjust to the proper pH and again check the first standard pH buffer. Clean the electrode and immerse it in the test solution. Record the pH of the test solution.

As with all delicate equipment, the pH meter and electrode must receive proper care and maintenance. All electrodes should be kept clean and stored in solutions suggested by manufacturers. Glass electrodes are fragile and expensive, so they must be handled with care. If pH measurements of protein solutions are often taken, a protein film may develop on the electrode; it can be removed by soaking in 5% pepsin in 0.1 *M* HCl for 2 hours and rinsing well with water. The pH instrument should always be set on "standby" when the electrode is not in a solution.

Measurements of pH are always susceptible to experimental errors. Some common problems are:

1. ***The Sodium Error*** Many glass combination electrodes are sensitive to Na^+ as well as H^+. The sodium error can become quite significant at high pH values, where 0.1 *M* Na^+ may decrease the measured pH by 0.4 to 0.5 unit. Several things may be done to reduce the sodium error. Some commercial suppliers of electrodes provide a standard curve for sodium error correction. Newer electrodes that are virtually Na^+ impermeable are now commercially available. If neither a standard curve nor a sodium-insensitive electrode is available, potassium salts may be substituted for sodium salts.
2. ***Concentration Effects*** The pH of a solution varies with the concentration of buffer ions or other salts in the solution. This is because the pH of a solution depends on the *activity* of an ionic species, not on the concentration. **Activity,** you may recall, is a thermodynamic term used to define species in a nonideal solution. At infinite dilution, the activity of a species is equivalent to its concentration. At finite dilutions, however, the activity of a solute and its concentration are not equal.

   It is common practice in biochemical laboratories to prepare concentrated "stock" solutions and buffers. These are then diluted to the proper concentration when needed. Because of the concentration effects described above, it is important to adjust the pH of these solutions *after* dilution.
3. ***Temperature Effects*** The pH of a buffer solution is influenced by temperature. This effect is due to a temperature-dependent change of the dissociation constant (pK_a) of ions in solution. The pH of the commonly used buffer Tris is greatly affected by temperature changes, with a $\Delta pK_a/C°$ of −0.031. This means that a pH 7.0 Tris buffer made up at 4°C would have a pH of 5.95 at 37°C. The best way to avoid this problem is to prepare the buffer solution at the temperature at which it will be used and to standardize the electrode with buffers at the same temperature as the solution you wish to measure.

Biochemical Buffers

Buffer ions are used to maintain solutions at constant pH values. The selection of a buffer for use in the investigation of a biochemical process is of critical importance. Before the characteristics of a buffer system are discussed, we will review some concepts in acid-base chemistry.

Weak acids and bases do not completely dissociate in solution, but exist as equilibrium mixtures (Equation 3.3).

>>

$$HA \underset{k_2}{\overset{k_1}{\rightleftharpoons}} H^+ + A^- \qquad \textbf{(Eq. 3.3)}$$

HA represents a weak acid and A^- represents its conjugate base; k_1 represents the rate constant for dissociation of the acid and k_2 the rate constant for association of the conjugate base and hydrogen ion. The **equilibrium constant,** K_a, for the weak acid HA is defined by Equation 3.4.

>> $$K_a = \frac{k_1}{k_2} = \frac{[H^+][A^-]}{[HA]}$$ **(Eq. 3.4)**

which can be rearranged to define $[H^+]$ (Equation 3.5).

>> $$[H^+] = \frac{K_a[HA]}{[A^-]}$$ **(Eq. 3.5)**

The $[H^+]$ is often reported as pH, which is $-\log [H^+]$. In a similar fashion, $-\log K_a$ is represented by pK_a. Equation 3.5 can be converted to the $-\log$ form by substituting pH and pK_a:

>> $$pH = pK_a + \log \frac{[A^-]}{[HA]}$$ **(Eq. 3.6)**

Equation 3.6 is the familiar **Henderson-Hasselbalch equation,** which defines the relationship between pH and the ratio of acid and conjugate base concentrations. The salt of the acid (A^-) is also referred to as the **proton acceptor (A)** and the acid (HA) as the **proton donor (D).** The Henderson-Hasselbalch equation is of great value in buffer chemistry because it can be used to calculate the pH of a solution if the molar ratio of buffer ions $([A^-]/[HA])$ and the pK_a of HA are known. Also, the molar ratio of HA to A^- that is necessary to prepare a buffer solution at a specific pH can be calculated if the pK_a is known.

A solution containing both HA and A^- has the capacity to resist changes in pH; i.e., it acts as a **buffer.** If acid (H^+) were added to the buffer solution, it would be neutralized by A^- in solution:

>> $$H^+ + A^- \longrightarrow HA$$ **(Eq. 3.7)**

Base (OH^-) added to the buffer solution would be neutralized by reaction with HA:

>> $$OH^- + HA \longrightarrow A^- + H_2O$$ **(Eq. 3.8)**

The most effective buffering system contains equal concentrations of the acid, HA, and its conjugate base, A^-. According to the Henderson-Hasselbalch equation (3.6), when $[A^-]$ is equal to [HA], pH equals pK_a. Therefore, the pK_a of a weak acid-base system represents the center of the buffering region. The effective range of a buffer system is generally two pH units, centered at the pK_a value (Equation 3.9).

>> $$\text{Effective pH range for a buffer} = pK_a \pm 1$$ **(Eq. 3.9)**

Selection of a Biochemical Buffer

Virtually all biochemical investigations must be carried out in buffered aqueous solutions. The natural environment of biomolecules and cellular organelles is under strict pH control. When these components are extracted from cells, they are most stable if maintained in their normal pH range, usually 6 to 8. An artificial buffer system is found to be the best substitute for the natural cell milieu. It should also be recognized that many biochemical processes (especially some enzyme processes) produce or consume hydrogen ions. The buffer system neutralizes these solutions and maintains a constant chemical environment.

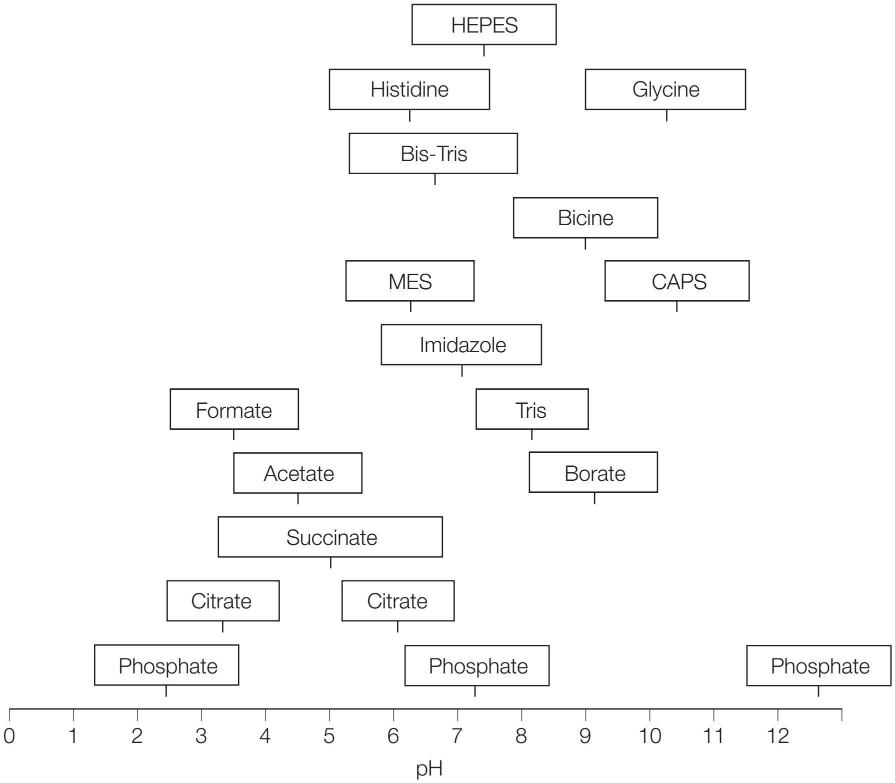

FIGURE 3.2 Effective buffering ranges of several common buffers. See Table 3.1 for abbreviations.

Although most biochemical solutions require buffer systems effective in the pH range 6 to 8, there is occasionally a need for buffering over the pH range 2 to 12. Obviously, no single acid–conjugate base pair will be effective over this entire range, but several buffer systems are available that may be used in discrete pH ranges. Figure 3.2 compares the effective buffering ranges of common biological buffers. It should be noted that some buffers (phosphate, succinate, and citrate) have more than one pK_a value, so they may be used in different pH regions. Many buffer systems are effective in the usual biological pH range (6 to 8); however, there may be major problems in their use. Several characteristics of a buffer must be considered before a final selection is made. The molecular weights and pK values of several common buffer compounds are listed in Appendix III. Following is a discussion of the advantages and disadvantages of the commonly used buffers.

Phosphate Buffers

The phosphates are among the widely used buffers. These solutions have a high buffering capacity and are very useful in the pH range 6.5 to 7.5. Because phosphate is a natural constituent of cells and biological fluids, its presence affords a more "natural" environment than many buffers. Sodium or potassium phosphate solutions of all concentrations are commonly prepared with the use of the Henderson-Hasselbalch equation. The major disadvantages of phosphate solutions are: (1) precipitation or binding of common biological cations (Ca^{2+} and Mg^{2+}), (2) inhibition of some biological processes, including some enzymes, and (3) limited useful pH range.

STUDY EXERCISE 3.1 Henderson-Hasselbalch Equation

Here we will describe the calculation method for preparation of a sodium phosphate buffer of 0.05 *M* total phosphate concentration, pH of 7.0, and a total, final volume of 1 liter. The Henderson-Hasselbalch equation (Equation 3.6) will be used to calculate the conjugate acid and conjugate base concentrations. At pH 7.0, the two forms of phosphate present are:

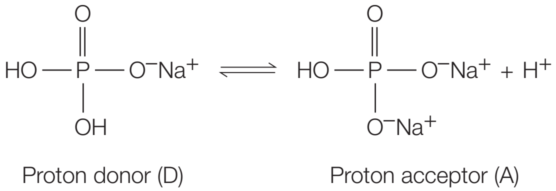

We now use the Henderson-Hasselbalch equation with pH = 7.0 and pK = 7.21 (from Appendix III) to calculate the molar ratio:

$$[A]/[D] = 0.62$$

This gives only one equation, but two unknowns, the concentrations of A and D; therefore, we need another quantitative relationship between [A] and [D]. We know that the total phosphate concentration ($[A] + [D]$) is 0.05 *M*, so:

$$[A] + [D] = 0.05$$

Using these two equations for A and D, we determine that we need 0.031 mole of D and 0.019 mole of A. The appropriate phosphate reagent forms available commercially are:

D = donor: monobasic sodium phosphate monohydrate, $NaH_2PO_4 \bullet H_2O$ (MW = 138); 0.031 mole.

A = acceptor: dibasic sodium phosphate heptahydrate, $Na_2HPO_4 \bullet 7H_2O$ (MW = 268); 0.019 mole.

Converting the number of moles of each reagent to grams:

$$\text{grams of D} = 0.031 \times 138 = 4.28\ \text{g}$$
$$\text{grams of A} = 0.019 \times 268 = 5.09\ \text{g}$$

To prepare the solution, dissolve the appropriate amount of each reagent in about 975 mL of purified water in a 1-liter beaker. Check the pH of the solution and adjust to 7.0 by careful, dropwise addition (with stirring) of dilute NaOH or dilute phosphoric acid. Transfer the solution to a 1-liter volumetric flask, add water to the mark, mix well, and do a final check of the pH.

Zwitterionic Buffers (Good's Buffers)

In the mid-1960s, N. E. Good and his colleagues recognized the need for a set of buffers specifically designed for biochemical studies (Good and Izawa, 1972). He and others noted major disadvantages of the established buffer systems. Good outlined several characteristics essential in a biological buffer system:

1. pK_a between 6 and 8.
2. Highly soluble in aqueous systems.
3. Exclusion or minimal transport by biological membranes.
4. Minimal salt effects.
5. Minimal effects on dissociation due to ionic composition, concentration, and temperature.
6. Buffer–metal ion complexes nonexistent or soluble and well defined.
7. Chemically stable.
8. Insignificant light absorption in the ultraviolet and visible regions.
9. Readily available in purified form.

Good investigated a large number of synthetic zwitterionic buffers and found many of them to meet these criteria. Table 3.1 lists several of these buffers and their properties. Good's buffers are widely used, but their main disadvantage is high cost. Some zwitterionic buffers, such as Tris, HEPES, and PIPES, have been shown to produce radicals under a variety of experimental conditions, so they should be avoided if biological redox processes or radical-based reactions are being studied. Radicals are not produced from MES or MOPS buffers.

The use of the synthetic zwitterionic buffer Tris [Tris(hydroxymethyl)-aminomethane] is now probably greater than that of phosphate. It is useful in the pH range 7.5 to 8.5. Tris is available in a basic form as highly purified crystals, which makes buffer preparation especially convenient. Although Tris is a primary amine, it causes minimal interference with biochemical processes and does not precipitate calcium ions. However, Tris has several disadvantages, including: (1) pH dependence on concentration, since the pH decreases 0.1 pH unit for each 10-fold dilution; (2) interference with some pH electrodes; and (3) a large $\Delta pK_a/C^\circ$ compared to most other buffers. Most of these drawbacks can be minimized by: (1) adjusting the pH *after* dilution to the appropriate concentration, (2) purchasing electrodes that are compatible with Tris, and (3) preparing the buffer at the temperature at which it will be used.

STUDY EXERCISE 3.2 Buffer Made by Titration

We will illustrate the preparation of a Tris buffer using the titration method. (In Study Exercise 3.1, we used the calculation method to prepare the phosphate buffer.) Because Tris is commercially available in highly purified crystals, Tris Base (MW = 121.1), the appropriate amount may be weighed, dissolved in water, and titrated to the desired pH with an acid, usually HCl. For 1 liter of 0.1 *M* Tris-HCl buffer, pH 7.0, dissolve 12.11 g (0.1 mole) of Tris Base in about 975 mL of purified water in a 1-liter beaker. Adjust the pH (originally about 10) to 7.0 by careful, dropwise addition of concentrated hydrochloric acid. This should be done in a 1-liter beaker with gentle stirring. Transfer the solution to a 1-liter volumetric flask, add water to the mark, mix well, and make a final check of the pH.

As you can see by comparing the two procedures, the titration method for buffer preparation is faster and much more convenient than the calculation method. Is it possible to prepare the phosphate buffer in Study Exercise 3.1 by using the titration method? If so, describe the reagent(s) needed and the procedure you would use.

Carboxylic Acid Buffers

The most widely used buffers in this category are acetate, formate, citrate, and succinate. This group is useful in the pH range 3 to 6, a region that offers few other buffer choices. All of these acids are natural metabolites, so they may interfere with the biological processes under investigation. Also, citrate and succinate may interfere by binding metal ions ($Fe^{3+}, Ca^{2+}, Zn^{2+}, Mg^{2+}$, etc.). Formate buffers are especially useful because they are volatile and can be removed by evaporation under reduced pressure.

Borate Buffers

Buffers of boric acid are useful in the pH range 8.5 to 10. Borate has the major disadvantage of complex formation with many metabolites, especially carbohydrates.

Amino Acid Buffers

The most commonly used amino acid buffers are glycine (pH 2 to 3 and 9.5 to 10.5), histidine (pH 5.5 to 6.5), glycine amide (pH 7.8 to 8.8), and glycylglycine (pH 8 to 9). These provide a more "natural" environment similar to cellular components and extracts; however, they may interfere with some biological processes, as do the carboxylic acid and phosphate buffers.

Buffer Dilutions

In biochemistry lab activities, buffers of many different concentrations and volumes are required. It is usually not practical, economical, or convenient to prepare all needed buffers from scratch. However, it is possible to have available a few concentrated stock buffer solutions that may be diluted to produce a new buffer solution of desired concentration and amount.

Dilutions are usually defined in the following way: A dilution of 1 mL to 10 mL (also 1:10; or 1 to 10) means that you take 1 mL of stock solution and add water to a final total volume of 10 mL. It is important to note that the ***amount*** of the buffer components taken from the stock solution is equal to the ***amount*** of buffer components in the diluted solution. However, the ***concentration*** of the buffer components has been changed. If the stock solution above had a buffer concentration of 0.5 *M*, then the 1:10 diluted solution has a buffer concentration of 0.05 *M*.

Two kinds of dilutions, linear and serial, are used most often in biochemistry.

- ***Linear dilutions*** Table 3.2 shows the results of a linear dilution process in which a stock solution of 1 *M* is used to produce samples of different concentrations. Note that there is a linear (progressive) decrease in the concentration of the diluted samples, but the final volumes remain the same (10 mL). The volume of stock solution to use in each case is calculated from the following relationship:

$$\text{volume (stock)} = \frac{\text{volume (diluted)} \times \text{concentration (diluted)}}{\text{concentration (stock)}}$$

TABLE 3.2 Example of a Linear Dilution Process

Stock (*M*)	Final dilution (*M*)	Stock (mL)	Water (mL)*	Final volume (mL)
1.0	0.80	8.0	2.0	10
1.0	0.60	6.0	4.0	10
1.0	0.40	4.0	6.0	10
1.0	0.20	2.0	8.0	10
1.0	0.00	0.0	10.0	10

*Because all solutions are not additive, this number is the approximate amount of water needed to bring the final volume to 10 mL. It may be necessary to add a few drops of water to bring the final volume to 10 mL.

Sample calculation for the first diluted solution in Table 3.2:

$$\text{volume (stock)} = \frac{10\text{ mL} \times 0.80\,M}{1.0\,M} = 8.0\text{ mL}$$

- ***Serial dilutions*** Here dilution starts with a stock buffer solution and then each diluted solution produced is used to prepare the next, etc. For example:

  2.0 mL of Solution I is diluted to a final volume of 100 mL (Solution II)
  2.0 mL of Solution II is diluted to a final volume of 100 mL (Solution III)
  2.0 mL of Solution III is diluted to a final volume of 100 mL (Solution IV)

  If Solution I has a buffer concentration of 10 *M* and water is used for dilution in each case, what is the buffer concentration of Solution II?

  The equation above for linear dilution may also be used here:

$$\text{For Solution II: concentration (diluted)} = \frac{2.0\text{ mL} \times 10\,M}{100\text{ mL}} = 0.2\,M$$

STUDY EXERCISE 3.3 Serial Dilutions

What are the concentrations of Solutions III and IV made from the above serial dilution?

The Oxygen Electrode

Second in popularity only to the pH electrode is the oxygen electrode. This device is a polarographic electrode system that can be used to measure oxygen concentration in a liquid sample or to monitor oxygen uptake or evolution by a chemical or biological system.

The basic electrode was developed by L. C. Clark, Jr. in 1953. The popular Clark-type oxygen electrode with biological oxygen monitor is shown in Figure 3.3. The electrode system, which consists of a platinum cathode and silver anodes, is molded into an epoxy block and surrounded by a Lucite holder. A thin Teflon membrane is stretched over the electrode end of the probe and held in place with an O-ring. The membrane separates the electrode elements, which are bathed in an

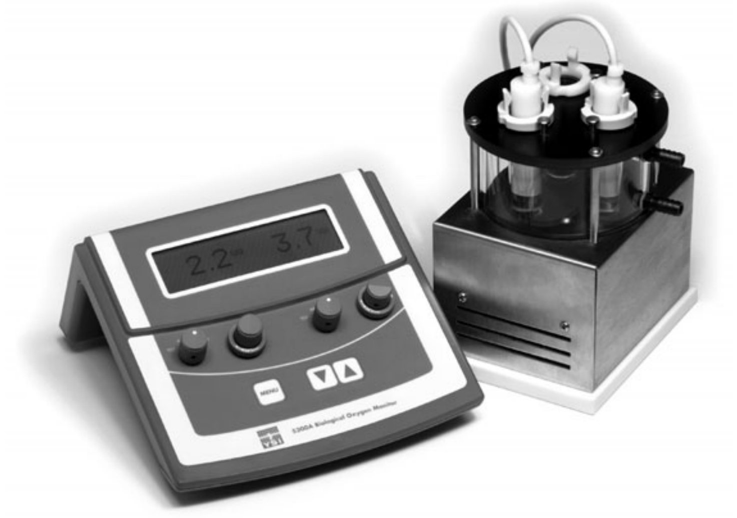

FIGURE 3.3 Biological oxygen monitor, YSI Model 5300A. *Courtesy of YSI Life Sciences; www.ysi.com.*

electrolyte solution (saturated KCl), from the test solution in the sample chamber. The membrane is permeable to oxygen and other gases, which then come into contact with the surface of the platinum cathode. If a suitable polarizing voltage (about 0.8 volt) is applied across the cell, oxygen is reduced at the cathode.

$$\begin{aligned}&O_2 + 2e^- + 2H_2O \longrightarrow [H_2O_2] + 2OH^-\\&[H_2O_2] + 2e^- \longrightarrow 2OH^-\\&\text{Total}\quad O_2 + 4e^- + 2H_2O \longrightarrow 4OH^-\end{aligned}$$

The reaction occurring at the anode is

$$4Ag^\circ + 4Cl^- \longrightarrow 4AgCl + 4e^-$$

The overall electrochemical process is

$$4Ag^\circ + O_2 + 4Cl^- + 2H_2O \longrightarrow 4AgCl + 4OH^-$$

The current flowing through the electrode system is, therefore, directly proportional to the amount of oxygen passing through the membrane. According to the laws of diffusion, the rate of oxygen flowing through the membrane depends on the concentration of dissolved oxygen in the sample. Therefore, the magnitude of the current flow in the electrode is directly related to the concentration of dissolved oxygen. The function of the electrode can also be explained in terms of oxygen pressure. The oxygen concentration at the cathode (inside the membrane) is virtually zero because oxygen is rapidly reduced. Oxygen pressure in the sample chamber (outside the membrane) will force oxygen to diffuse through the membrane at a rate that is directly proportional to the pressure (or concentration) of oxygen in the test solution. The current generated by the electrode system is electronically amplified and transmitted to a meter, which usually reads in units of % oxygen saturation.

Many biochemical processes consume or evolve oxygen. The oxygen electrode is especially useful for monitoring such changes in the concentration of dissolved oxygen. If oxygen is consumed, for example, by a suspension of mitochondria or by an oxygenase enzyme system, the rate of oxygen diffusion through the probe membrane will decrease, leading to less current flow. A linear relationship exists between the electrode current and the amount of oxygen consumed by the biological system.

The use of an oxygen electrode is not free of experimental and procedural problems. Probably the most significant source of trouble is the Teflon membrane. A torn, damaged, or dirty membrane is easily recognized by recorder noise or electronic "spiking." Temperature control of the sample chamber and electrode is essential because the rate of oxygen diffusion through the membrane is temperature dependent. The sample chamber is equipped with a bath assembly for temperature regulation and magnetic stirrer in order to maintain a constant oxygen concentration throughout the solution. Air bubbles in the sample chamber lead to considerable error in measurements and must be removed.

Biosensors

Electrodes or electrode-like devices are currently being developed for the specific measurement of physiologically important molecules such as urea, carbohydrates, enzymes, antibodies, and metabolic products. This type of device, now referred to as a **biosensor,** is an analytical tool or system consisting of an immobilized biological material (such as an enzyme, antibody, whole cell, organelle, or combinations thereof) in intimate contact with a suitable transducer device that will convert the biochemical signal into a quantifiable electric signal. The important components of a biosensor as shown in Figure 3.4 are: (1) a reaction center consisting of a membrane or gel containing the biochemical system to be studied, (2) a transducer, (3) an amplifier, and (4) a computer system for data acquisition and processing. When biomolecules in the reaction center interact, a physicochemical change occurs. This change in the molecular system, which may be a modification of concentration, absorbance, mass, conductance, or redox state, is converted into an electrical signal by the transducer. The signal is then amplified and displayed on a computer screen. Each biosensor is specifically designed for a particular type of molecule or biological

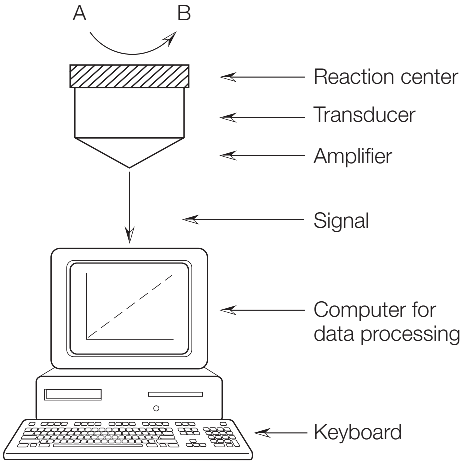

FIGURE 3.4 Components of a biosensor system for determining the identity and concentration of biomolecule A. A is stoichiometrically converted to B by a specific enzyme in the reaction center. The concentration of B is measured by the sensor.

interaction; therefore, details of construction and function vary. Some specific examples of biosensors include:

1. Electrodes based on enzyme activity. These are selective and sensitive devices that may be used to measure substrate concentrations. A biosensor based on glucose oxidase is used to measure the concentration of glucose by detecting the production of H_2O_2.

$$\text{Glucose} + O_2 \rightleftharpoons \text{gluconic acid} + H_2O_2$$

2. Optical biosensors that respond to the absorption of light from a laser. The optical response is transmitted to a computer by a fiber-optic system.

B. MEASUREMENT OF PROTEIN SOLUTIONS

Biochemical research often requires the quantitative measurement of protein concentrations in solutions. Several techniques for such measurement have been developed; however, most have limitations because either they are not sensitive enough or they are based on reactions with specific amino acids in the protein. Since the amino acid content varies from protein to protein, no single assay will be suitable for all proteins. In this section, we discuss five assays: three older, classical methods that are occasionally used today and two other methods that are widely used. In four of the methods, chemical reagents are added to protein solutions to develop a color whose intensity is measured in a spectrophotometer. A “standard protein” of known concentration is also treated with the same reagents and a calibration curve is constructed. The other assay relies on a direct spectrophotometric measurement. None of the methods is perfect because each is dependent on the amino acid content of the protein. However, each will provide a satisfactory result if the proper experimental conditions are used and/or a suitable standard protein is chosen. Other important factors in method selection include the sensitivity and accuracy desired, the presence of interfering substances, and the time available for the assay. The various methods are compared in Table 3.3.

The Biuret and Lowry Assays

When substances containing two or more peptide bonds react with the biuret reagent, alkaline copper sulfate, a purple complex is formed. The colored product is the result of coordination of peptide nitrogen atoms with Cu^{2+}. The amount of product formed depends on the concentration of protein.

In practice, a calibration curve must be prepared by using a standard protein solution. The very best protein to use as a standard is a purified preparation of the protein to be assayed. Since this is rarely available, the experimenter must choose another protein as a relative standard. A relative standard must be selected that provides a similar color yield. An aqueous solution of bovine serum albumin (BSA) is a commonly used standard. Various known amounts of this solution are treated with the biuret reagent, and the color is allowed to develop. Measurements of absorbance at 540 nm (A_{540}) are made against a blank containing biuret reagent and buffer or water. The A_{540} data

TABLE 3.3 Methods for Protein Measurement

Method	Sensitivity	Time	Principle	Interferences	Comments
Biuret	Low 1–20 mg	Moderate 20–30 min	Peptide bonds + alkaline $Cu^{2+} \rightarrow$ purple complex	Zwitterionic buffers, some amino acids	Similar color with all proteins. Destructive to protein samples.
Lowry	High ~5 μg	Slow 40–60 min	(1) Biuret reaction (2) Reduction of phosphomolybdate-phosphotungstate by Tyr and Trp	Ammonium sulfate, glycine, zwitterionic buffers, mercaptans	Time-consuming. Intensity of color varies with proteins. Critical timing of procedure. Destructive to protein samples.
Bradford	High ~1 μg	Rapid 15 min	λ_{max} of Coomassie dye shifts from 465 nm to 595 nm when protein-bound	Strongly basic buffers; detergents Triton X-100, SDS	Stable color that varies with proteins. Reagents commercially available. Destructive to protein samples. Discoloration of glassware.
BCA	High 1 μg	Slow 60 min	(1) Biuret reaction (2) Copper complex with BCA; $\lambda_{max} = 562$ nm	EDTA, DTT, ammonium sulfate	Compatible with detergents. Reagents commercially available. Destructive to protein samples.
Spectrophotometric (A_{280})	Moderate $50 - 1000\ \mu$g	Rapid 5–10 min	Absorption of 280-nm light by aromatic residues	Purines, pyrimidines, nucleic acids	Useful for monitoring column eluents. Nucleic acid absorption can be corrected. Nondestructive to protein samples. Varies with proteins.

are plotted versus protein concentration (mg/mL) or amount of protein (mg). Unknown protein samples are treated with biuret reagent and the A_{540} is measured after color development. The protein concentration is determined from the standard curve. The biuret assay has several advantages, including speed, similar color development with different proteins, and few interfering substances. Its primary disadvantage is its lack of sensitivity.

The **Lowry protein assay** is one of the more sensitive assays and has been widely used. The Lowry procedure can detect protein levels as low as 5 μg. The principle behind color development is identical to that of the biuret assay except that a second reagent (Folin-Ciocalteu) is added to increase the amount of color development. Two reactions account for the intense blue color that develops: (1) the coordination of peptide bonds with alkaline copper (biuret reaction), and (2) the reduction of the Folin-Ciocalteu reagent (phosphomolybdate-phosphotungstate) by tyrosine and tryptophan residues in the protein. The procedure is similar to that of the biuret assay. A standard curve is prepared with bovine serum albumin or other pure protein, and the concentration of unknown protein solutions is determined from the graph.

The obvious advantage of the Lowry assay is its sensitivity, which is up to 100 times greater than that of the biuret assay; however, more time is required for the Lowry assay. Since proteins have varying contents of tyrosine and tryptophan, the amount of color development changes with different proteins, including the bovine serum albumin standard. Because of this, the Lowry protein assay should be used only for measuring changes in protein concentration, not absolute values of protein concentration.

The Bradford Assay

The many limitations of the biuret and Lowry assays have encouraged researchers to seek better methods for quantitation of protein solutions. The **Bradford assay,** based on protein binding of a dye, provides numerous advantages over other methods. The binding of Coomassie Brilliant Blue dye (Figure 3.5) to protein in acidic solution causes a shift in wavelengh of maximum absorption (λ_{max}) of the dye from 465 nm to 595 nm. The absorption at 595 nm is directly related to the concentration of protein. The dye reagent interacts primarily with positively-charged (basic) amino acid residues (arginyl, lysyl, and histidyl) in proteins. In addition, there are probably weaker interactions between the dye molecules and hydrophobic, aromatic amino acid residues (tyrosinyl, phenylalanyl, and tryptophanyl). Proteins that have more than an average number of these interacting residues show higher sensitivity and thus higher absorbance values at 595 nm.

In practice, a calibration curve is prepared by plotting μg of standard protein versus absorbance at 595 nm, which is then used to estimate the amount of protein in unknown samples (Figure 3.6). Recommended standard proteins include bovine gamma globulin (IgG), lysozyme, and ovalbumin because they have closer to the average or typical number of amino acid residues that bind the dye (see above). Although BSA is often recommended in the literature, it is not a good choice for a standard for the Bradford assay, as it has a high content of reacting amino acid residues. The 595 nm absorbance of BSA is about 2.1 times that of IgG (Figure 3.6).

The assay requires only a single reagent, an acidic solution of Coomassie Brilliant Blue G-250. After addition of dye solution to a protein sample, color

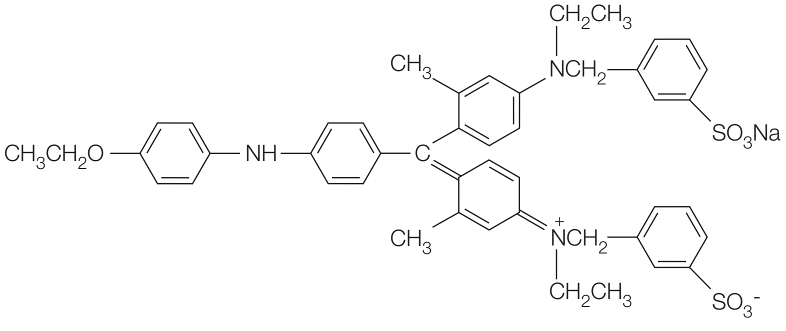

FIGURE 3.5 Coomassie Brilliant Blue G-250 dye is used to measure protein concentration in the Bradford assay.

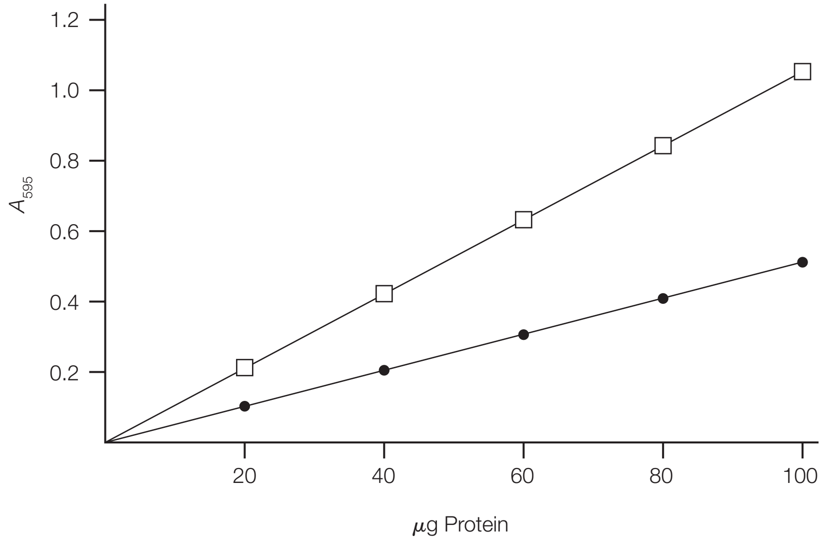

FIGURE 3.6 Student-generated, standard calibration curves for the Bradford protein assay using bovine gamma globulin (IgG, •) and bovine serum albumin (BSA, □) as the protein standards.

development is complete in 2 minutes and the color remains stable for up to 1 hour. The sensitivity of the Bradford assay rivals and may surpass that of the Lowry assay. With a microassay procedure, the Bradford assay can be used to determine proteins in the range of 1 to 20 μg. The Bradford assay shows significant variation with different proteins, but this also occurs with the Lowry assay and can be avoided by use of the proper protein standard. The Bradford method not only is rapid, but also has very few interferences by nonprotein components. The only known interfering substances are the detergents Triton X-100 and sodium dodecyl sulfate. The many advantages of the Bradford assay have led to its wide adoption in biochemical research laboratories.

The BCA Assay

The **BCA protein assay** is based on chemical principles similar to those of the biuret and Lowry assays. The protein to be analyzed is reacted with Cu^{2+} (which produces Cu^{+}) and bicinchoninic acid (BCA). The Cu^{+} is chelated by BCA, which converts the apple-green color of the free BCA to the purple color of the copper-BCA complex. Samples of unknown protein and relative standard are treated with the reagent, the absorbance is read in a spectrophotometer at 562 nm, and concentrations are determined from a standard calibration curve. This assay has the same sensitivity level as the Lowry and Bradford assays. Its main advantages are its simplicity and its usefulness in the presence of 1% detergents such as Triton or sodium dodecyl sulfate (SDS).

The Spectrophotometric Assay

Most proteins have relatively intense ultraviolet light absorption centered at 280 nm. This is due to the presence of aromatic tyrosine and tryptophan residues in the protein. However, the amount of these amino acid residues varies in different proteins, as was pointed out earlier. If certain precautions are taken, the A_{280} value of a protein solution is proportional to the protein concentration. In practice, the procedure is simple and rapid. A protein solution is transferred to a quartz cuvette and the A_{280} is read against a reference cuvette containing the protein solvent only (buffer, water, etc.). For the average protein, a concentration of 1 mg/mL will yield an absorbance at 280 nm of about 1. If the extinction coefficient is known for the protein, its exact concentration may be calculated from the absorbance at 280 nm.

Cellular extracts contain many other compounds that absorb in the vicinity of 280 nm. Nucleic acids, which can be common contaminants in a protein extract, absorb strongly at 280 nm ($\lambda_{max} = 260$). Early researchers developed a method to correct for this interference by nucleic acids. Mixtures of pure protein and pure nucleic acid were prepared, and the ratio A_{280}/A_{260} was experimentally determined. The following empirical equation may be used for protein solutions containing up to 20% nucleic acids.

$$\text{Protein concentration (mg/mL)} = 1.55A_{280} - 0.76A_{260}$$

Although the spectrophotometric assay of proteins is fast, relatively sensitive, and requires only a small sample size, it is still only an estimate of protein concentration unless its extinction coefficient is known. It has certain advantages over the colorimetric assays in that most buffers and ammonium sulfate do not interfere and the procedure is nondestructive to protein samples. The spectrophotometric assay is particularly suited to the rapid and continuous measurement of protein elution from a chromatography column, where only protein concentration *changes* are required.

C. MEASUREMENT OF NUCLEIC ACID SOLUTIONS

The Spectrophotometric Assay

Solutions of nucleic acids strongly absorb ultraviolet light with a λ_{max} of 260 nm. The intense absorption is due primarily to the presence of aromatic rings in the purine and pyrimidine bases. The concentration of nucleic acid in a solution can be calculated if one knows the value of A_{260} of the solution. ***A solution of double-stranded DNA at a concentration of 50 µg/mL in a 1-cm quartz cuvette will give an A_{260} reading of about 1.0. A solution of single-stranded DNA or RNA that has an A_{260} of 1.0 in a cuvette with a 1-cm path length has a concentration of about 40 µg/mL.***

The spectrophotometric assay for nucleic acids is simple, rapid, nondestructive, and very sensitive; however, it should be noted that it measures both DNA and RNA. The minimum concentration of nucleic acid that can be accurately determined is 2.5–5.0 µg/mL.

Measuring the ultraviolet absorption of DNA and RNA solutions can also provide a useful structural probe and an indication of purity. DNA structural changes such as helix unwinding may be detected by measuring absorbance changes of DNA solutions at 260 nm. The ratio of A_{260}/A_{280} is often used as a relative measure of the purity of nucleic acid solutions. Pure DNA with little or no protein present has an absorbance ratio of about 1.8. Pure RNA has a ratio of about 2. Absorbance ratios less then these numbers indicate contamination of the solution by protein material.

STUDY EXERCISE 3.4 Nucleic Acid Concentrations

Two solutions, one of double-stranded DNA and one of single-stranded RNA, were transferred to 1-cm quartz cuvettes and the absorbance (A) at 260 nm of each of the solutions was found to be 0.15. What are the concentrations of the DNA and RNA solutions? Hint: Recall that a standard solution of double-stranded DNA at a concentration of 50 μg /mL in a 1-cm quartz cuvette will yield an A_{260} reading of about 1.0. A solution of single-stranded DNA or RNA that has an A_{260} of 1.0 in a cuvette with a 1-cm path length has a concentration of about 40 μg /mL.

Other Assays for Nucleic Acids

The spectrophotometric assay for nucleic acids, DNA and RNA, is simple and rapid, but it does not have a sufficient level of sensitivity for many applications. Other assays for DNA and RNA depend on the enhanced fluorescence of dyes when bound to nucleic acids. Features of fluorescence assays are compared to the spectrophotometric assay in Table 3.4.

The fluorescence dye bisbenzimidazole, when bound to DNA, emits light at 458 nm that is about five times greater in intensity than the emission by free, unbound dye. The increase in fluorescence is linear with increasing amounts of DNA in the concentration range of 0.01 μg/mL to 5 μg/mL. One special advantage of this assay is that it can be used to quantify DNA in cell suspensions and crude homogenates. It does not respond to RNA or other biomolecules in the extracts. The assay is fast and requires only a fluorimeter, an instrument found in most laboratories.

TABLE 3.4 Measurement of Nucleic Acid Solutions

Method	Sensitivity	Time	Principle	Interferences	Comments
Spectrophotometric (A_{260})	High 2.5 μg	Rapid 5–10 min	Absorption of 260 nm light by aromatic bases	Proteins	Protein absorption may be corrected. A_{280}/A_{260} indicates purity. Nondestructive. Measures both DNA and RNA.
Bisbenzimidazole	Very high 10 ng	Rapid 5–10 min	Enhancement of fluorescence of dye when bound to DNA	None known	Not reliable with plasmids or small DNA. Does not measure RNA. May be used with crude extracts. Nondestructive.
EtBr-Agarose	Very high 1–10 ng	Moderate 30–45 min	Fluorescence of bound EtBr	None known	Measures DNA and RNA.

Another simple and relatively inexpensive fluorescence method uses ethidium bromide–agarose plates. The fluorescence emission of ethidium bromide (EtBr) at 590 nm is enhanced about 25-fold when the dye interacts with DNA. DNA samples for measurement are spotted onto 1% agarose containing (EtBr) in a Petri dish or spot plates. The samples are exposed to UV light and photographed, and the light intensity is compared to standard samples of DNA. This method is sensitive (in the 1–10 ng range), but it detects both DNA and RNA.

D. TECHNIQUES FOR SAMPLE PREPARATION

Dialysis

One of the oldest procedures applied to the purification and characterization of biomolecules is **dialysis,** an operation used to separate dissolved molecules on the basis of their molecular size. The technique involves sealing an aqueous solution containing both macromolecules and small molecules in a porous membrane. The sealed membrane is placed in a large container of low-ionic-strength buffer. The membrane pores are too small to allow diffusion of macromolecules of molecular weight greater than about 10,000. Smaller molecules diffuse freely through the openings (Figure 3.7). The passage of smaller molecules continues until their concentrations inside the dialysis tubing and outside in the large volume of buffer are equal. Thus, the concentration of small molecules inside the membrane is reduced. Equilibrium is volume dependent and is reached after 4 to 6 hours. Of course, if the outside solution (dialysate) is replaced with fresh buffer after equilibrium is reached, the concentration of small molecules inside the membrane will be further reduced by continued dialysis.

Dialysis membranes are available in a variety of materials and sizes. The most common materials are collodion, cellophane, and cellulose. Recent

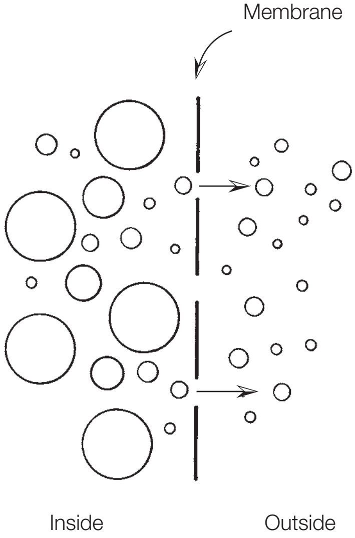

FIGURE 3.7
Diffusion of smaller molecules through dialysis membrane pores.

modifications in membrane construction make a range of pore sizes available. Spectrum Laboratories offers Spectra/Por membrane tubing with complete molecular weight cutoffs ranging from 100 to 300,000.

Dialysis is most commonly used to remove salts and other small molecules from solutions of macromolecules. During the separation and purification of biomolecules, small molecules are added to selectively precipitate or dissolve the desired molecule. For example, proteins are often precipitated by addition of organic solvents or salts such as ammonium or sodium sulfate. Since the presence of organics or salts usually interferes with further purification and characterization of the molecule, they must be removed. Dialysis is a simple, inexpensive, and effective method for removing all small molecules, ionic or nonionic.

Dialysis is also useful for removing small ions and molecules that are weakly bound to biomolecules. Protein cofactors such as NAD, FAD, and metal ions can often be dissociated by dialysis. The removal of metal ions is facilitated by the addition of a chelating agent (EDTA) to the dialysate.

Dialysis of small-volume samples (less than 3 mL) is often tedious, inconvenient, and inefficient. Commercial vials for dialysis are now available for sample sizes from 0.05 to 3.0 mL, with molecular weight cutoffs ranging from 3500 to 14,000 daltons. For example, GeBAflex tubes, offered by Gene Bio-Applications, may be used for dialysis (the removal of small molecules like urea, detergents, and ethidium bromide), buffer exchange, sample concentration, and electroelution of protein and nucleic acid samples (see Chapter 6).

Ultrafiltration

Although dialysis is still used occasionally as a purification tool, it has been largely replaced by ultrafiltration and gel filtration (see Chapter 5, Section E, p. 132). The major disadvantage of dialysis that is overcome by the newer methods is that it may take several days of dialysis to attain a suitable separation. The other methods require 1 to 2 hours or less.

Ultrafiltration involves the separation of molecular species on the basis of size, shape, and/or charge. The solution to be separated is forced through a membrane by an external force. Membranes may be chosen for optimum flow rate, molecular specificity, and molecular weight cutoff. Two applications of membrane filtration are obvious: (1) desalting buffers or other solutions, and (2) clarification of turbid solutions by removal of micron- or submicron-sized particles. Other applications are discussed below.

Ultrafiltration membranes have molecular weight cutoffs in the range of 100 to 1,000,000. They are usually composed of two layers: (1) a thin (0.1–0.5 μm), surface, semipermeable membrane made from a variety of materials including cellulose acetate, nylon, and polyvinylidene, and (2) a thicker, inert, support base (Figure 3.8). These filters function by retaining particles on the surfaces, not within the base matrix.

Membrane filters of these materials can be manufactured with a predetermined and accurately controlled pore size. Filters are available with a mean pore size ranging from 0.025 to 15 μm. These filters require suction, pressure, or centrifugal force for liquid flow. A typical flow rate for the commonly used 0.45$-\mu$m membrane is 57 mL min^{-1} cm^{-2} at 10 psi.

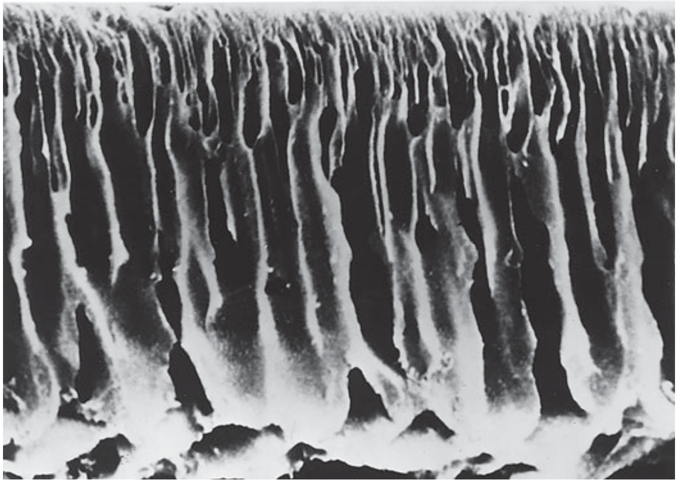

FIGURE 3.8 Electron micrograph of an ultrafiltration membrane showing the two layers. Particles greater than 0.1 μm in diameter are retained on the surface or within pores. *Courtesy of the Millipore Corporation;* www.millipore.com.

Ultrafiltration devices are available for macroseparations (up to 50 L) or for microseparations (milli- to microliters). For solutions larger than a few milliliters, gas-pressurized cells or suction-filter devices are used. For concentration and purification of samples in the milli- to microliter range, disposable filters are available. These devices, often called microconcentrators, offer the user simplicity, time saving, and high recovery. The sample is placed in a reservoir above the membrane and centrifuged. The time and centrifugal force required depend on the membrane, with spin times varying from 30 minutes to 2 hours and forces from $1000 \times g$ to $7500 \times g$. Figure 3.9 outlines the use of a centrifuge microfilter.

The principles behind ultrafiltration are sometimes misunderstood. The nomenclature implies that separations are the result of physical trapping of the particles and molecules by the filter. With polycarbonate and fiberglass filters, separations are made primarily on the basis of physical size. Other filters (cellulose nitrate, polyvinylidene fluoride, and to a lesser extent cellulose acetate) trap particles that cannot pass through the pores, but also retain macromolecules by adsorption. In particular, these materials have protein- and nucleic acid-binding properties. Each type of membrane displays a different affinity for various molecules. For protein, the relative binding affinity is polyvinylidene fluoride > cellulose nitrate > cellulose acetate. We can expect to see many applications of the "affinity membranes" in the future as the various membrane surface chemistries are altered and made more specific. Some applications are described in the following pages.

Clarification of Solutions

Because of low solubility and denaturation, solutions of biomolecules or cellular extracts are often turbid. This is a particular disadvantage if spectrophotometric analysis is desired. The transmittance of turbid solutions can be greatly increased by passage through a membrane filter system.

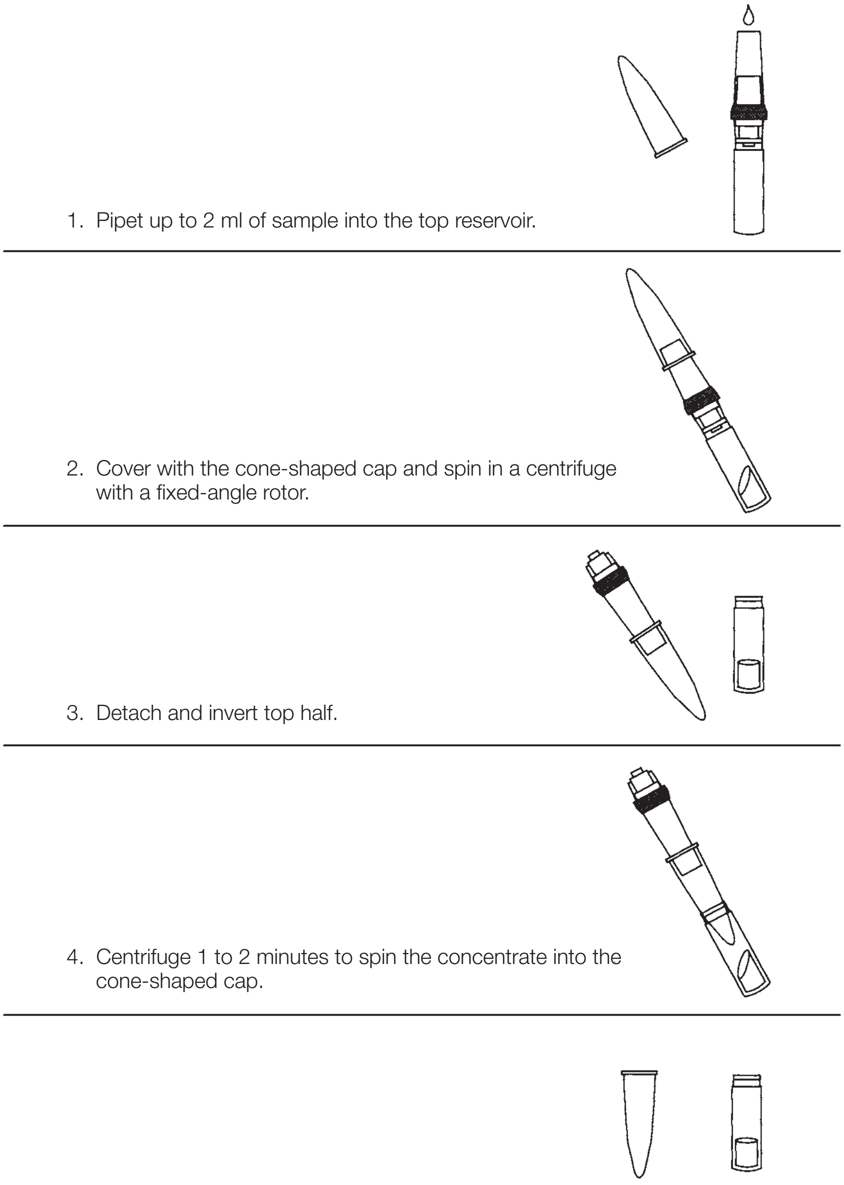

1. Pipet up to 2 ml of sample into the top reservoir.
2. Cover with the cone-shaped cap and spin in a centrifuge with a fixed-angle rotor.
3. Detach and invert top half.
4. Centrifuge 1 to 2 minutes to spin the concentrate into the cone-shaped cap.
5. Concentrate is now completely accessible for further analysis. Both concentrate and filtrate can be stored in collection cups.

FIGURE 3.9 Use of a centrifuge microfilter. *Courtesy of the Millipore Corporation;* www.millipore.com.

This simple technique may also be applied to the sterilization of nonautoclavable materials such as protein and nucleic acid solutions or heat-labile reagents. Bacterial contamination can be removed from these solutions by passing them through filter systems that have been sterilized by autoclaving.

Collection of Precipitates for Analysis

The collection of small amounts of very fine precipitates is the basis for many chemical and biochemical analytical procedures. Membrane filtration is an ideal method for sample collection. This is of great advantage in the collection of radioactive precipitates. Cellulose nitrate and fiberglass filters are often used to collect radioactive samples because they can be analyzed by direct suspension in an appropriate scintillation cocktail.

Harvesting of Bacterial Cells from Fermentation Broths

The collection of bacterial cells from nutrient broths is typically done by batch centrifugation. This time-consuming operation can be replaced by membrane filtration. Filtration is faster than centrifugation, and it allows extensive cell washing.

Concentration of Biomolecule Solutions

Protein or nucleic acid solutions obtained from extraction or various purification steps are often too dilute for further investigation. Since they cannot be concentrated by high-temperature evaporation of solvent, gentler methods have been developed. One of the most effective is the use of ultrafiltration pressure cells, as shown in Figure 3.10. A membrane filter is placed in the bottom and the solution is poured into the cell. High pressure, exerted by compressed nitrogen (air could cause oxidation and denaturation of biomolecules), forces the flow of small molecules, including solvent, through the filter. Membranes are available in a number of sizes, allowing a large variety of molecular weight cutoffs. Larger molecules that cannot pass through the pores are concentrated in the sample chamber. This method of concentration is rapid and gentle and can be performed at cold temperatures to ensure minimal inactivation of the molecules. One major disadvantage is clogging of the pores, which reduces the flow rate through the filter; this is lessened by constant but gentle stirring of the solution.

Lyophilization and Centrifugal Vacuum Concentration

Although ultrafiltration is being used more and more for the concentration of biological solutions, the older technique of **lyophilization** (freeze-drying) is still used. There are some situations (storing or transporting biological materials) in which lyophilization is preferred. Lyophilization is a drying technique that uses the process of sublimation to change a solvent (water) in the frozen state directly to the vapor state under vacuum. The product after lyophilization is a fluffy matrix that may be reconstituted by the addition of liquid. This is one of the most effective methods for drying or concentrating heat-sensitive materials. In practice, a biological solution to be concentrated is “shell-frozen” on the walls of a round-bottom or freeze-drying flask. Freezing of the solution is accomplished by placing the flask (half full with sample) in a dry ice–acetone bath and slowly rotating it as it is held at a 45° angle. The aqueous solution freezes in layers on the wall of the flask. This provides a large surface area for evaporation of water. The flask is then connected to the lyophilizer, which consists of a refrigeration unit and a vacuum pump (see Figure 3.11). The combined unit maintains the sample

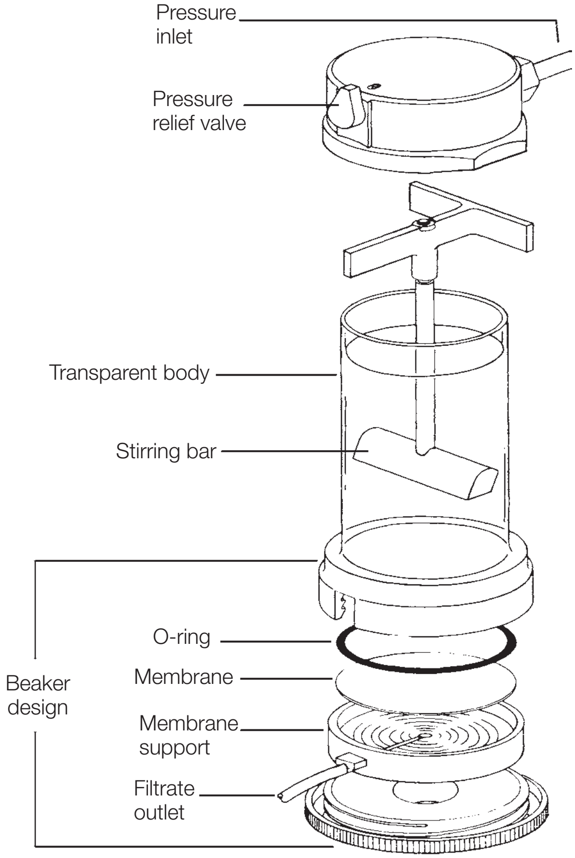

FIGURE 3.10 Schematic of an ultrafiltration cell. *Courtesy of Millipore/Amicon Corporation;* www.millipore.com.

at −40°C for stability of the biological materials and applies a vacuum of approximately 5 to 25 mm Hg on the sample. Ice formed from the aqueous solution sublimes and is pumped from the sample vial. In fact, all materials that are volatile under these conditions (−40°C, 5 to 25 mm Hg) will be removed, and nonvolatile materials (proteins, buffer salts, nucleic acids, etc.) will be concentrated into a light, sometimes fluffy precipitate. Most freeze-dried biological materials are stable for long periods of time and some remain viable for many years.

As with any laboratory method, there are precautions for and limitations of lyophilization that must be understood. Only aqueous solutions should be lyophilized. Organic solvents lower the melting point of aqueous solutions and increase the chances that the sample will melt and become denatured during freeze-drying. There is also the possibility that organic vapors will pass through the cold trap into the vacuum pump, where they may cause damage.

A new and increasingly popular technique for sample concentration and drying is **centrifugal vacuum concentration.** The method may be used to dry a wide variety of biological samples. The process starts with a sample dissolved in a solvent (water or organic). The solvent is evaporated under vacuum in a centrifuge, thus producing a pellet in the bottom of the container. The most widely used

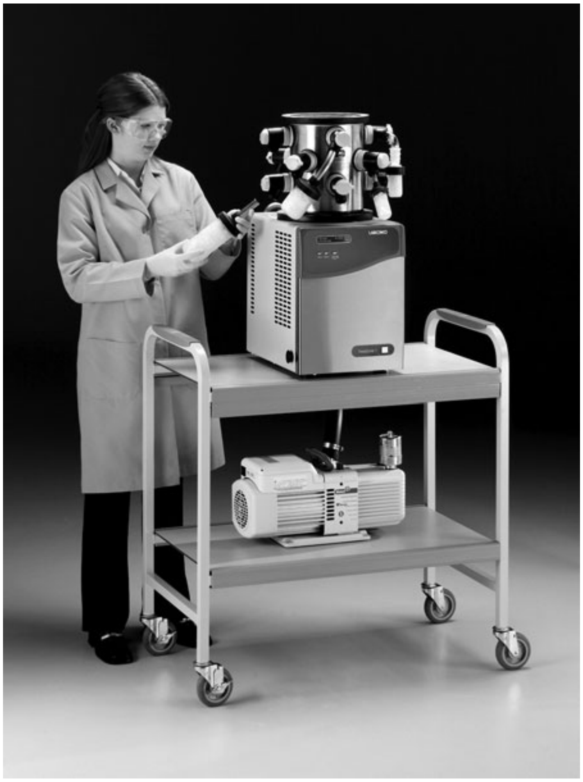

FIGURE 3.11 A lab technician prepares a sample for lyophilization (freeze-drying) with the Labconco FreeZone benchtop freeze dryer, Model # 7740020. © *Labconco Corporation. Photo used with permission from Labconco Corporation.* www.labconco.com.

instrument is the SpeedVac, available through Savant (Figure 3.12). This method is better than freeze-drying because it is faster, it does not require a prefreezing step, it provides 100% sample recovery, and it may be used with solvents other than water. The SpeedVac is usually used for relatively small volumes—1–2-mL samples.

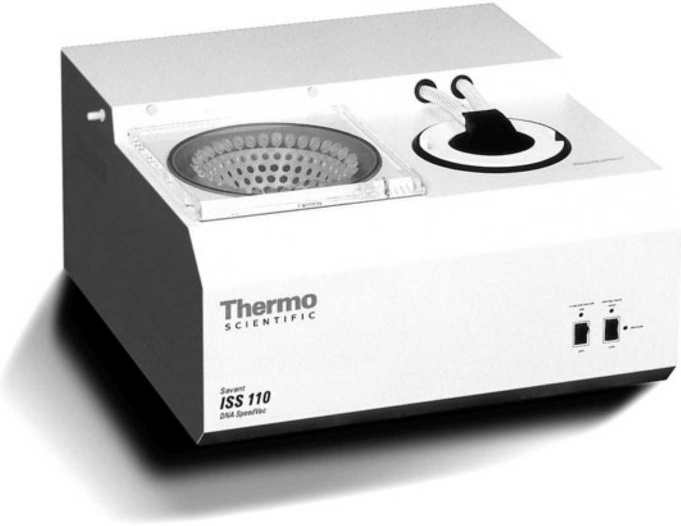

FIGURE 3.12 The SpeedVac centrifugal vacuum concentrator. *Image provided by Thermo Fisher Scientific Inc., all rights reserved.* www.thermofisher.com.

E. RADIOISOTOPES IN BIOCHEMISTRY

The use of radioactive isotopes in experimental biochemistry has provided us with a wealth of information about biological processes. In the earlier days of biochemistry, radioisotopes were primarily used to elucidate metabolic pathways. In modern biochemistry, few experimental techniques offer such a diverse range of applications as that provided by radioactive isotopes. The measurement of radioactivity is a tool that has the potential for use in all areas of experimental biochemistry, including enzyme assays, biochemical pathways of synthesis and degradation, analysis of biomolecules, measurement of antibodies, binding and transport studies, clinical biochemistry, and many others. Radioisotope applications have two major advantages over other analytical methods: (1) sensitive instrumentation allows the detection of minute quantities of radioactive material (some radioactive substances can be measured in the picomole, 10^{-12}, range), and (2) radioisotope techniques offer the ability to differentiate physically between substances that are chemically indistinguishable. The major disadvantages of radioisotope use are, of course, the safety and environmental concerns that must be considered when working with radioactive materials. These concerns have greatly slowed the development of new procedures using radioisotopes, as some countries and U.S. states require that all who work with radioactive materials, including students, must have completed an intensive training course. Because of the potential hazards and problems associated with using and disposing of radioactive waste materials, radioisotopes are generally not used in undergraduate biochemistry teaching labs. They are used in research, in industrial biotechnology applications, and in medicine only when there are no other techniques that can be applied to the problem. They are then handled only by experienced personnel.

The difficult problems and safety issues that arise with radioisotope use have encouraged scientists to search for alternate techniques. The last two Web site references in the Radioisotopes section at the end of the chapter describe the search for "green biochemistry" techniques that avoid the use of radioisotopes. One reference describes the design of protein kinase assays that avoid the use of phosphorus-32. Another describes the synthesis of probes for radioisotope-free photoaffinity labels for proteins.

Origin and Properties of Radioactivity

Introduction

Radioactivity results from the spontaneous nuclear disintegration of unstable isotopes. The hydrogen nucleus, consisting of a proton, is represented as ${}^{1}_{1}\mathrm{H}$. Two additional forms of the hydrogen nucleus contain one and two neutrons; they are represented as ${}^{2}_{1}\mathrm{H}$ and ${}^{3}_{1}\mathrm{H}$. These **isotopes** of hydrogen are commonly called deuterium and tritium, respectively. All isotopes of hydrogen have an identical number of protons (constant charge, +1), but they differ in the number of neutrons. Only tritium is radioactive.

The stability of a nucleus depends on the ratio of neutrons to protons. Some nuclei are unstable and undergo spontaneous nuclear disintegration

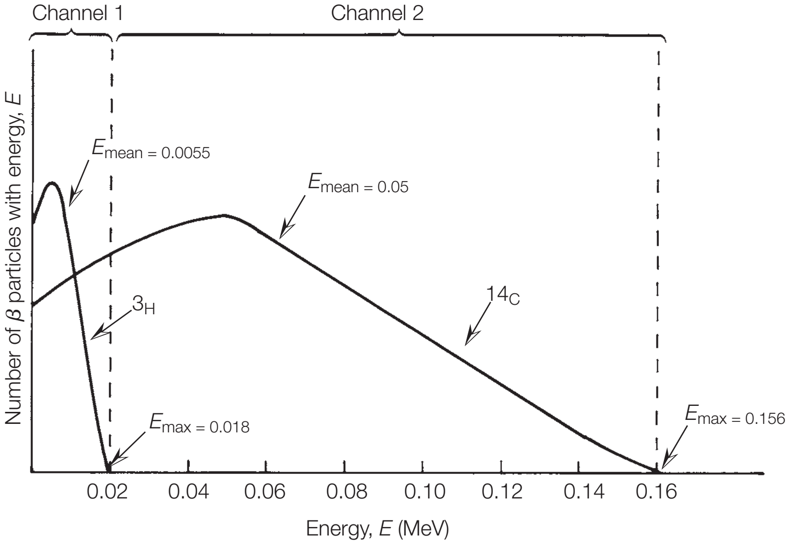

FIGURE 3.13 Energy spectra for the β emitters ^{3}H and ^{14}C. The dashed lines indicate the upper and lower limits for discrimination counting.

accompanied by emission of particles. Unstable isotopes of this type are called **radioisotopes.** Three main types of radiation are emitted during nuclear decay: α particles, β particles, and γ rays. The α particle, a helium nucleus, is emitted only by elements with a mass number greater than 140. These elements are seldom used in biochemical research.

Most of the commonly used radioisotopes in biochemistry are β emitters. The β particles exist in two forms (β^+, positrons, and β^-, electrons) and are emitted from a given radioisotope with a continuous range of energies. However, an average energy (called the *mean* energy) of β particles from an element can be determined. The mean energy is a characteristic of that isotope and can be used to identify one β emitter in the presence of a second (see Figure 3.13). Equations 3.10 and 3.11 illustrate the disintegration of two β emitters. Here $\overline{v}_e$ represents the antineutrino and v_e the neutrino.

>> $$^{32}_{15}P \longrightarrow ^{32}_{16}S + \beta^- + \overline{v}_e \qquad \textbf{(Eq. 3.10)}$$

>> $$^{65}_{30}Zn \longrightarrow ^{65}_{29}Cu + \beta^+ + v_e \qquad \textbf{(Eq. 3.11)}$$

A few radioisotopes of biochemical significance are γ emitters. Emission of a γ ray (a photon of electromagnetic radiation) is often a secondary process occurring after the initial decay by β emission. The disintegration of the isotope ^{131}I is an example of this multistep process.

>> $$^{131}_{53}I \longrightarrow \beta^- + ^{131}_{54}Xe \longrightarrow ^{131}_{54}Xe + \gamma \qquad \textbf{(Eq. 3.12)}$$

Each radioisotope emits γ rays of a distinct energy, which can be measured for identification of the isotope.

The spontaneous disintegration of a nucleus is a first-order kinetic process. That is, the rate of radioactive decay of N atoms ($-dN/dt$, the change of N with time, t) is proportional to the number of radioactive atoms present (Equation 3.13).

>> $$-\frac{dN}{dt} = \lambda N \qquad \textbf{(Eq. 3.13)}$$

The proportionality constant, λ, is called the **disintegration** or **decay constant.** Equation 3.13 can be transformed to a more useful equation by integration within the limits of $t = 0$ and t. The result is shown in Equation 3.14.

>> $$N = N_0 e^{-\lambda t} \qquad \textbf{(Eq. 3.14)}$$

where

$N_0 =$ the number of radioactive atoms present at $t = 0$
$N =$ the number of radioactive atoms present at time $= t$

Equation 3.14 can be expressed in the natural logarithmic form as in Equation 3.15.

>> $$\ln \frac{N}{N_0} = -\lambda \mathrm{t} \qquad \textbf{(Eq. 3.15)}$$

Combining Equations 3.14 and 3.15 leads to a relationship that defines the **half-life.** Half-life, $t_{1/2}$, is a term used to describe the time necessary for one-half of the radioactive atoms initially present in a sample to decay. At the end of the first half-life, N in Equation 3.15 becomes $\frac{1}{2}N_0$, and the result is shown in Equations 3.16 and 3.17.

>> $$\ln \frac{\frac{1}{2}N_0}{N_0} = -\lambda t_{1/2}$$

$$\ln \tfrac{1}{2} = -\lambda t_{1/2} \qquad \textbf{(Eq. 3.16)}$$

>> $$\mathrm{t}_{1/2} = \frac{0.693}{\lambda} \qquad \textbf{(Eq. 3.17)}$$

Equations 3.15 and 3.17 allow calculation of the ratio N/N_0 at any time during an experiment. This calculation is especially critical when radioisotopes with short half-lives are used.

Isotopes in Biochemistry

The properties of several radioisotopes that are important in biochemical research are listed in Table 3.5. Note that many of the isotopes are β emitters; however, a few are γ emitters.

The half-life, defined in the previous section and listed for each isotope in Table 3.5, is an important property to consider when one is designing experiments using radioisotopes. Using an isotope with a short half-life (for example, ^{24}Na with $t_{1/2} = 15$ hrs) is difficult because the radioactivity lost during the course of the experiment is significant. Quantitative measurements made before and after the experiment must be corrected for this loss of activity. Radioactive phosphorus,

TABLE 3.5 Properties of Biochemically Important Radioisotopes

Isotope	Particle Emitted	Energy of Particle (MeV)	Half-Life, $t_{1/2}$
^{3}H	β^-	0.018	12.3 yrs
^{14}C	β^-	0.155	5570 yrs
^{22}Na	β^+	0.55	2.6 yrs
	γ	1.28	
^{24}Na	β^-	1.39	15 hrs
	γ	1.7, 2.75	
^{32}P	β^-	1.71	14.2 days
^{35}S	β^-	0.167	87.1 days
^{36}Cl	β^-	0.714	3×10^5 days
^{40}K	β^-	1.33	1.3×10^9 yrs
^{45}Ca	β^-	0.25	165 days
^{59}Fe	β^-	0.46	45 days
	γ	1.1	
^{60}Co	β^-	0.318	5.3 yrs
	γ	1.3	
^{65}Zn	β^+	0.33	245 days
	γ	1.14	
^{90}Sr	β^-	0.54	29 yrs
^{125}I	γ	0.035	60 days
	γ	0.027	
^{131}I	β^-	0.61, 0.33	8.1 days
	γ	0.64, 0.36, 0.28	
^{137}Cs	β^-	0.51	30 yrs
	γ	0.66	
^{226}Ra	α	4.78	1600 yrs
	γ	0.19	

^{32}P, an isotope of significant value in biochemical research, has a relatively short half-life (14 days), so if quantitative measurements are made they must be corrected as described in Equations 3.16 and 3.17.

Mention should also be made of short-lived isotopes that are important in biotechnology and medical biochemistry. The isotopes ^{11}C, ^{13}N, ^{15}O, and ^{18}F, which are positron emitters, are crucial for use in positron emission tomography (PET).

TABLE 3.6 Units of Radioactivity

Unit's Name	Multiplication Factor (relative to curie)	Activity (dps)
Curie (Ci)	1.0	3.70×10^{10}
Millicurie (mCi)	10^{-3}	3.70×10^{7}
Microcurie (μCi)	10^{-6}	3.70×10^{4}
Nanocurie (nCi)	10^{-9}	3.70×10
Becquerel (Bq)	2.7×10^{-11}	1.0
Mega becquerel (MBq)	2.7×10^{-5}	1.0×10^{6}

Units of Radioactivity

The basic unit of radioactivity is the **curie,** Ci, named in honor of Marie and Pierre Curie. One curie is the amount of radioactive material that emits particles at a rate of 3.7×10^{10} disintegrations per second (dps), or 2.2×10^{12} min^{-1} (dpm). Amounts that large are seldom used in experimentation, so subdivisions are convenient. The **millicurie** (mCi, 2.2×10^{9} min^{-1}) and **microcurie** (μCi, 2.2×10^{6} min^{-1}) are standard units for radioactive measurements (see Table 3.6). The radioactivity unit of the meter-kilogram-seconds (MKS) system is the **becquerel** (Bq). A becquerel, named in honor of Antoine Becquerel, who studied uranium radiation, represents one disintegration per second. The two systems of measurement are related by the equation 1 curie $= 3.70 \times 10^{10}$ becquerels. Since the becquerel is such a small unit, radioactive units are sometimes reported in MBq (mega, 10^{6}) or TBq (tera, 10^{12}). Both unit systems are in common use today, and radioisotopes received through commercial sources are labeled in curies and becquerels.

The number of disintegrations emitted by a radioactive sample depends on the purity of the sample (number of radioactive atoms present) and the decay constant, λ. Therefore, radioactive decay is also expressed in terms of **specific activity,** the disintegration rate per unit mass of radioactive atoms. Typical units for specific activity are mCi/mmole and μCi/μmole.

Although radioactivity is defined in terms of nuclear disintegrations per unit of time, rarely does one measure this absolute number in the laboratory. Instruments that detect and count emitted particles respond to only a small fraction of the particles. Data from a radiation counter are in counts per minute (cpm), which can be converted to actual disintegrations per minute if the counting efficiency of the instrument is known. The percent efficiency of an instrument is determined by counting a standard compound of known radioactivity and using the ratio of detected activity (observed cpm) to actual activity (disintegrations per minute). Equation 3.18 illustrates the calculation using a standard of 1 μCi.

>>

$$\begin{aligned}\%\ \text{efficiency} &= \frac{\text{observed cpm of standard}}{\text{dpm of } 1\mu\text{Ci of standard}} \times 100 \\ &= \frac{\text{observed cpm}}{2.2 \times 10^{6}\ \text{dpm}} \times 100 \qquad \textbf{(Eq. 3.18)}\end{aligned}$$

Detection and Measurement of Radioactivity

Since most of the radioisotopes used in biochemical research are β emitters, only methods that detect and measure β particles will be emphasized.

Liquid Scintillation Counting

Samples for liquid scintillation counting consist of three components: (1) the radioactive material; (2) a solvent, usually aromatic, in which the radioactive substance is dissolved or suspended; and (3) one or more organic fluorescent substances. Components 2 and 3 make up the liquid scintillation system or cocktail. β particles emitted from the radioactive sample interact with the scintillation system, producing small flashes of light, or scintillations. The light flashes are detected by a photomultiplier tube (PMT). Electronic pulses from the PMT are amplified and registered by a counting device called a scaler. A schematic diagram of a typical scintillation counter is shown in Figure 3.14.

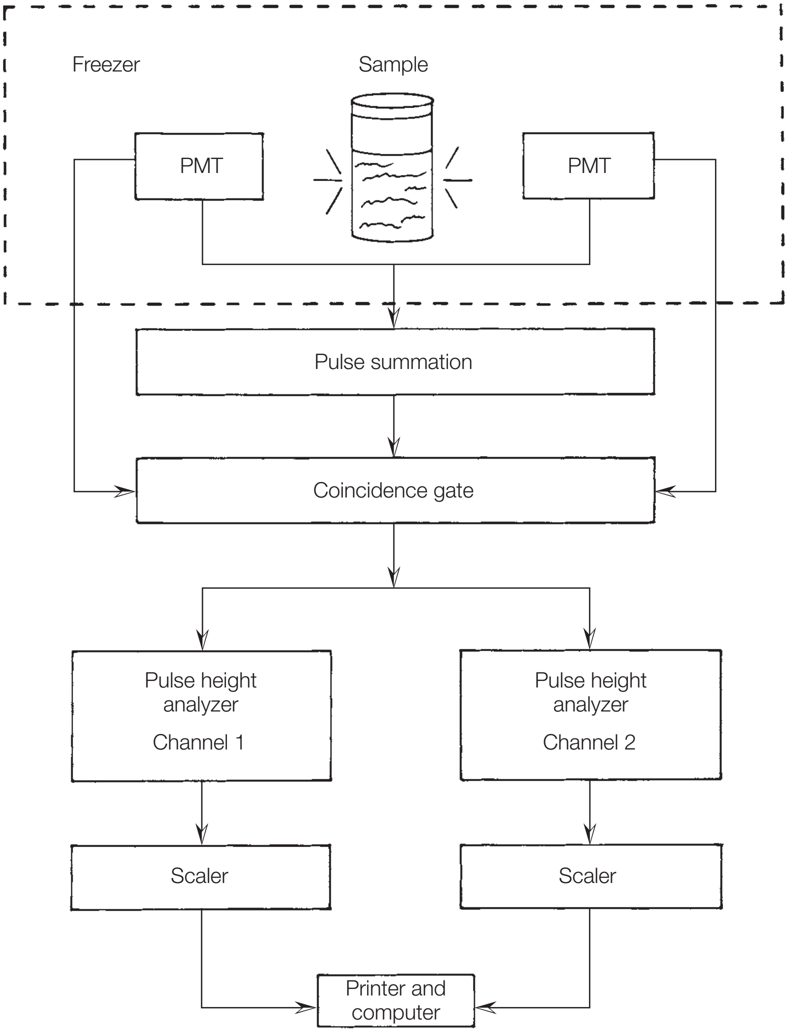
Freezer
Sample
PMT
PMT
Pulse summation
Coincidence gate
Pulse height analyzer
Channel 1
Pulse height analyzer
Channel 2
Scaler
Scaler
Printer and computer

FIGURE 3.14 Diagram of a typical scintillation counter, showing coincidence circuitry.

The scintillation process, in detail, begins with the collision of emitted β particles with solvent molecules, S (Equation 3.19).

>> $$\beta + S \rightarrow S^* + \beta \text{ (less energetic)} \qquad \textbf{(Eq. 3.19)}$$

Contact between the energetic β particles and S in the ground state results in transfer of energy and conversion of an S molecule into an excited state, S*. Aromatic solvents are most often used because their electrons are easily promoted to an excited state orbital (see discussion of fluorescence, Chapter 7, p. 220). The β particle after one collision still has sufficient energy to excite several more solvent molecules. The excited solvent molecules normally return to the ground state by emission of a photon, $S^* \longrightarrow S + hv$. Photons from the typical aromatic solvent are of short wavelength and are not efficiently detected by photocells. A convenient way to resolve this technical problem is to add one or more fluorescent substances (fluors) to the scintillation mixture. Excited solvent molecules interact with a **primary fluor,** F_1, as shown in Equation 3.20.

>> $$S^* + F_1 \longrightarrow S + F_1^* \qquad \textbf{(Eq. 3.20)}$$

Energy is transferred from S* to F_1, resulting in ground state S molecules and excited F_1 molecules, $F_1^* \cdot F_1^*$ molecules are fluorescent and emit light of a longer wavelength than S* (Equation 3.21).

>> $$F_1^* \longrightarrow F_1 + hv_1 \qquad \textbf{(Eq. 3.21)}$$

If the light emitted during the decay of F_1^* is still of a wavelength too short for efficient measurement by a PMT, a **secondary fluor,** F_2, that accepts energy from $F^*{}_1$ may be added to the scintillation system. Equations 3.22 and 3.23 outline the continued energy transfer process and fluorescence of F_2^*.

>> $$F_1^* + F_2 \longrightarrow F_1 + F_2^* \qquad \textbf{(Eq. 3.22)}$$

>> $$F_2^* \longrightarrow F_2 + hv_2 \qquad \textbf{(Eq. 3.23)}$$

The light, hv_2, from F_2^* is of longer wavelength than hv_1 from F_1^* and is more efficiently detected by a PMT. Two widely used primary and secondary fluors are 2,5-diphenyloxazole (PPO) with an emission maximum of 380 nm and 1,4-bis-2-(5-phenyloxazolyl)benzene (POPOP) with an emission maximum of 420 nm. A more efficient but more expensive fluor is 2-(4′-*t*-butylphenyl)-5-(4″-biphenyl)-1,3,4-oxadiazole (Butyl-PBD).

The most basic elements in a liquid scintillation counter are the PMT, a pulse amplifier, and a counter, called a **scaler.** This simple assembly may be used for counting; however, there are many problems and disadvantages with this setup. Many of the difficulties can be alleviated by more sophisticated instrumental features. Some of the problems and practical solutions are outlined below.

Thermal Noise in Photomultiplier Tubes

The energies of the β particles from most β emitters are very low. This, of course, leads to low-energy photons emitted from the fluors and relatively low-energy electrical pulses in the PMT. In addition, photomultiplier tubes produce thermal background noise with 25 to 30% of the energy associated with the fluorescence-emitted photons. This difficulty cannot be completely eliminated, but its effects can be lessened by placing the samples and the PMT in a freezer at −5 to −8°C in order to decrease thermal noise.

A second way to help resolve the thermal noise problem is to use two photomultiplier tubes for detection of scintillations. Each flash of light that is detected by the photomultiplier tubes is fed into a **coincidence circuit.** A coincidence circuit counts only the flashes that arrive simultaneously at the two photodetectors. Electrical pulses that are the result of simultaneous random emission (thermal noise) in the individual tubes are very unlikely. A schematic diagram of a typical scintillation counter with coincidence circuitry is shown in Figure 3.14.

Counting More Than One Isotope in a Sample

The basic liquid scintillation counter with coincidence circuitry can be used only to count samples containing one type of isotope. Many experiments in biochemistry require the counting of just one isotope; however, more valuable experiments can be performed if two radioisotopes can be simultaneously counted in a single sample (double-labeling experiments). The basic scintillation counter just described has no means of discriminating between electrical pulses of different energies.

The size of the current generated in a photocell is nearly proportional to the energy of the β particle initiating the pulse. Recall that β particles from different isotopes have characteristic energy spectra with an average energy (see Figure 3.13). Modern scintillation counters are equipped with **pulse height analyzers** that measure the size of the electrical pulse and count only the pulses within preselected energy limits set by **discriminators.** The circuitry required for pulse height analysis and energy discrimination of β particles is shown in Figure 3.14. Discriminators are electronic "windows" that can be adjusted to count β particles within certain energy or voltage ranges called **channels.** The channels are set to a **lower limit** and an **upper limit,** and all voltages within those limits are counted. Figure 3.13 illustrates the function of discriminators for the counting of ^{3}H and ^{14}C in a single sample. Discriminator channel 1 is adjusted to accept typical β particles emitted from ^{3}H and channel 2 is adjusted to receive β particles of the energy characteristic of ^{14}C.

Quenching

Any chemical agent or experimental condition that reduces the efficiency of the scintillation and detection process leads to a reduced level of counting, or **quenching.** Quenching may be caused by a decrease in the amount of fluorescence from the primary and secondary scintillators (fluors) or a decrease in light activating the PMT. There are four common origins of quenching.

Color quenching is a problem if chemical substances that absorb photons from the secondary fluors are present in the scintillation mixture. Since the secondary fluors emit light in the visible region between 410 and 420 nm, colored substances may absorb the emitted light before it is detected by the photocells. Radioactive samples may be treated to remove colored impurities before mixing with the scintillation solvent.

Chemical quenching occurs when chemical substances in the scintillation solution interact with excited solvent and fluor molecules and decrease the efficiency of the scintillation process. To avoid this type of quenching, the sample can be purified or the fluors can be increased in concentration. Modern scintillation counters have computer programs to correct for color and chemical quenching.

Point quenching occurs if the radioactive sample is not completely dissolved in the solvent. The emitted β particles may be absorbed before they have a chance to interact with solvent molecules. The addition of solubilizing agents such as Cab-O-Sil or Thixin decreases point quenching by converting the liquid scintillator to a gel.

Dilution quenching results when a large volume of liquid radioactive sample is added to the scintillation solution. In most cases, this type of quenching cannot be eliminated, but it can be corrected by one of the techniques discussed below.

Since quenching can occur during all experimental counting of radioisotopes, it is important to be able to determine the extent of the reduced counting efficiency. Two methods for quench correction are in common use.

In the **internal standard ratio** method, the sample, X, under study is first counted; then a known amount of a standard radioactive solution is added to the sample, and it is counted again. The absolute activity of the original sample, A_X, is represented by Equation 3.24.

>> $$A_X = A_s \frac{C_X}{C_s} \tag{Eq. 3.24}$$

where

A_X = activity of the sample, X
C_X = radioactive counts from sample
A_s = activity of the standard
C_s = radioactive counts from standard

The absolute value of C_s is determined from Equation 3.25.

>> $$C_T = C_X + C_s \tag{Eq. 3.25}$$

where

C_T = total radioactive counts from sample and standard

Equation 3.24 can then be modified to Equation 3.26.

>> $$A_X = \frac{A_s C_X}{C_T - C_X} \tag{Eq. 3.26}$$

The internal standard ratio method for quench correction is tedious, time-consuming, and destroys the sample, so it is not an ideal method. Scintillation counters are equipped with a standard radiation source inside the instrument, but outside the scintillation solution. The radiation source, usually a gamma emitter, is mechanically moved into a position next to the vial containing the sample, and the combined system of standard and sample is counted. Gamma rays from the standard excite solvent molecules in the sample, and the scintillation process occurs as previously described. However, the instrument is adjusted to register only scintillations due to γ particle collisions with solvent molecules. This method for quench correction, called the **external standard method,** is fast and precise.

Scintillation Cocktails and Sample Preparation

Many of the quenching problems discussed earlier can be lessened if an experiment is carefully planned. Many sample preparation methods and scintillation liquids are available for use. Only a few of the more common techniques will be mentioned here. Most liquid or solid radioactive samples can be counted by mixing them with a **scintillation cocktail,** a mixture of solvent and fluor(s). The radioactive sample and scintillation solution are placed in a glass, polyethylene, polypropylene, polyester, or polycarbonate vial for counting. Traditional counting vials include a 20-mL-standard vial, a 6-mL miniature vial, a 1-mL Eppendorf tube, or a 200-μL microfuge tube. If glass is used, the solution must contain a low potassium content because naturally occurring ^{40}K is a β emitter. Two major types of solvent systems are available: (1) those that are immiscible with water, in which only organic samples may be used, and (2) those that dissolve aqueous samples. During the early years of liquid scintillation counting, the most common solvents used were toluene and xylene for organic samples and dioxane for aqueous samples. These solvents provided high counting efficiency. However, on the downside, they were flammable (flash points between 4 and 25°C), highly toxic, and presented major disposal costs and problems. In addition, they penetrated plastic counting vials, releasing vapors and liquids into the laboratory. A new environmental awareness has prompted the development of biosafe liquid scintillation cocktails. The use of highly alkylated aromatic solvents has produced cocktails that are fully biodegradable (converted to CO_2 and H_2O), have higher flash points (above 120°C), and offer high counting efficiency. The newer solvents may be disposed of after use by incineration, which greatly reduces costs compared to road transport to a disposal site. Occupational risks are also lessened because the solutions are defined by the Environmental Protection Agency as nontoxic.

Solid radioactive samples or those that are insoluble in either type of solvent may be quantified by collection onto small pieces of a solid support (filter paper or cellulose membrane) and added directly to the cocktail for counting. The efficiency of counting these samples depends on the support, but is usually less than that of counting a homogeneous sample.

Applications of Radioisotopes

The applications of radioisotopes in biochemical measurements are too numerous to outline here. Excellent reference books and Web sites, some of which are listed at the end of this chapter, provide experimental procedures for radioisotope utilization. The technique of **autoradiography,** however, should be mentioned here. This procedure allows the detection and localization of a radioactive substance (molecule or atom) in a tissue, cell, or cell organelle. Briefly, the technique involves placing a radioactive material directly onto or close to a photographic emulsion. Radioisotopes in the material emit radiation that impinges on the photographic plate, activating silver halide crystals in the emulsion. Upon development of the plate, a pattern or image is displayed that yields information about the location and amount of radioactive material in the

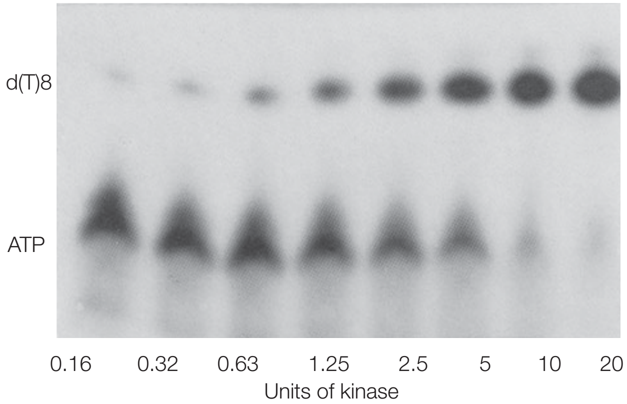

FIGURE 3.15 Autoradiograph showing the transfer of ^{32}P from γ-^{32}P-labeled ATP to the 5′ end of a polynucleotide, $(dT)_8$, catalyzed by polynucleotide kinase. Reagents were incubated for 30 min and electrophoresed on a 20% polyacrylamide, 7 *M* urea gel and autoradiographed. Note that as more units of kinase are added, more ^{32}P-labeled polynucleotide is made. The dark regions indicate the presence of ^{32}P. *Source: Copyright © New England Biolabs. Reprinted with permission.*

sample. A densitometer may be used to scan the autoradiograph and quantify the amount of radioactivity in each region. One of the most common uses of autoradiography in biochemistry and molecular biology is in the detection and quantification of ^{32}P-labeled nucleic acids on polyacrylamide or agarose gels after electrophoresis (Figure 3.15). Autoradiography has also been useful in concentration and localization studies of biomolecules in cells and cell organelles.

Statistical Analysis of Radioactivity Measurements

Radioactive decay is a random process. It is impossible to predict when a radioactive event will occur. Therefore, counting measurements provide only average rates of decay. However, counting measurements may be treated by Gaussian distribution analysis to determine an average counting rate, standard deviation, percent confidence level, and other statistical parameters. An introduction to statistical analysis of radioactivity counting data and other experimental measurements is given in Chapter 1, Section E, p. 23.

Radioisotopes and Safety

Safety should be of major concern in performing any chemistry experiment, but when radioisotopes are involved, special precautions must be taken. Many chemistry and biochemistry departments have specially equipped laboratories for radioisotope work. Alternatively, your instructor should set aside a specific area of your laboratory for using radioactive materials. In either situation, specific guidelines should be mandated and enforced by all institutions.

Study Problems

1. You need to prepare a buffer for biochemistry lab. The required solution is 0.5 *M* sodium phosphate, pH 7.0. Use the Henderson-Hasselbalch equation to calculate the number of moles and grams of monobasic sodium phosphate (NaH_2PO_4) and dibasic sodium phosphate (Na_2HPO_4) necessary to make 1 liter of the solution.
2. Design a "shortcut" method for preparing the phosphate buffer in Study Problem 1. Hint: You need only NaH_2PO_4, a solution of NaOH, and a pH meter.
3. Describe how you would prepare a 0.1 *M* glycine buffer of pH 10.0. You have available isoelectric glycine and sodium glycinate.
4. Describe how you would prepare a 0.20 *M* Tris-HCl buffer of pH 8.0. The only Tris reagent you have available is Tris base. What other reagent do you need, and how would you use it to prepare the solution?
5. Below is a table prepared by a biochemistry student to construct a standard curve for protein analysis. The Bradford assay was used with IgG (0.1 mg/mL) as standard protein. Complete the table by filling in the weight of IgG in each tube and the approximate A_{595} that will be obtained for each tube. Assume the procedure was conducted correctly.

	Tube No.					
Reagents	**1**	**2**	**3**	**4**	**5**	**6**
H_2O (mL)	1.0	0.9	0.8	0.6	0.2	—
IgG volume (mL)	—	0.1	0.2	0.4	0.8	1.0
IgG weight (μg)	0					
Bradford reagent (mL)	5.0	→				
A_{595}	0.00	0.05				

6. Assume that you use the standard curve produced in Study Problem 5 to measure the concentration of an unknown protein. The A_{595} for 1.0 mL of the unknown was 0.52. Prepare a standard curve from the data in Problem 5 and estimate the concentration of unknown protein in the sample in μg/mL.
7. A solution of purified DNA gave in the spectrophotometric assay an A_{260} of 0.35 when measured in a 1-cm quartz cuvette. What is the concentration of the DNA in μg/mL?
8. An RNA sample that you isolated has an absorbance reading at 260 nm of 0.25. What is the concentration of RNA in the sample? Assume the measurement was taken in a 1-cm quartz cuvette.
9. You have just completed an experiment to isolate DNA from *E. coli*. Absorbance measurements on the DNA sample were taken at 260 nm and 280 nm. The 260/280 ratio was 1.6. Is the DNA sample pure? If not, what could be the impurity?
10. Compare the techniques of lyophilization and centrifugal vacuum concentration. Give advantages and disadvantages of each.
11. What amino acid residues are detected when the spectrophotometric assay is used to quantify proteins? Are those amino acids present in the same quantity in all proteins? Explain how this may affect measurement of proteins by this method.
12. What biomolecules would interfere with the measurement of nucleic acids using the spectrophotometric (A_{260}) assay?
13. The concentration of a purified DNA solution is 35 μg/mL. Predict the value of absorbance of the solution at 260 and 280 nm.
14. In Section B of this chapter, you read that the Bradford assay has a higher sensitivity to some proteins than other proteins and hence gives higher absorbance readings at

595 nm. For example, the absorbance readings for bovine serum albumin are more than twice those for bovine gamma globulin (Figure 3.6). Explain why higher absorbance readings are obtained for some proteins. HINT: Use the bioinformatics Web sites given in Chapter 2.

15. In the Bradford assay of a protein, what is the most ideal protein to use as a standard?

16. Prepare a linear dilution table similar to Table 3.2. Assume that the stock solution is 2.0 *M* and you require 20-mL diluted samples of final concentrations: 1.66 *M*, 1.33 *M*, 1.0 *M*, 0.66 *M*, 0.33 *M*, and 0.00 *M*.

17. Select a buffer system that can be used at each of the following pH values.
   (a) pH 3.5
   (b) pH 6.0
   (c) pH 9.5

18. Describe how you would prepare 1 liter of the following buffer: 0.025 *M* Tris-chloride, pH 8.0, containing 0.05 *M* glucose and 0.01 *M* EDTA (MW EDTA disodium salt, dihydrate = 372.2).

19. Calculate λ, the decay constant for ^{32}P.

20. Calculate the half-life of an isotope from the following experimental measurements. At time = 0, the activity is 300 disintegrations per minute. After 1 hour, the activity is 200 disintegrations per minute.

21. The following serial dilution is carried out in the lab:
   1.0 mL of Solution A is diluted to a final volume of 10 mL (Solution B).
   1.0 mL of Solution B is diluted to a final volume of 10 mL (Solution C).
   1.0 mL of Solution C is diluted to a final volume of 10 mL (Solution D).
   Assume that Solution A has a concentration of 2.0 *M*. Calculate the concentrations of Solutions B, C, and D.

22. You need to prepare a buffer solution at each of the following concentrations:

1 *M*, 0.1 *M*, 0.01 *M*, and 0.001 *M*

Describe how you would use a single stock solution of the buffer (5.0 *M*) to prepare 10 mL of each diluted buffer solution.

23. How long will it take a sample of ^{32}P that contains 65,000 disintegrations per minute to decay to 1500 disintegrations per minute?

24. List five radioisotopes that may be measured using a liquid scintillation counter.

Further Reading

GENERAL REFERENCES

S. Sawhney and R. Singh (Editors), ***Introductory Practical Biochemistry,*** (2005), Alpha Science International (Oxford, U. K.).

D. Sheehan, ***Physical Biochemistry: Principles and Applications,*** 2nd ed. (2009), Wiley-Blackwell (Hoboken, NJ).

J. Walker, Editor, ***The Protein Protocols Handbook,*** 2nd ed. (2002), Humana Press (Totowa, NJ).

K. Wilson and J. Walker, Editors, ***Principles and Techniques of Practical Biochemistry,*** 5th ed. (2000), Cambridge University Press (Cambridge, U.K.).

MEASUREMENT OF PROTEINS AND NUCLEIC ACIDS

M. Bradford, ***Anal. Biochem.*** 72, 248–254 (1976). "A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principles of Ligand-Dye Binding."

P. Smith et al., ***Anal. Biochem.*** 150, 76–85 (1985). "Measurement of Protein Using Bicinchoninic Acid."

J. Walker, Editor, ***The Protein Protocols Handbook,*** 2nd ed. (2002), Humana Press, Inc. (Totowa, NJ), pp. 3–21.

http://www.ruf.rice.edu/~bioslabs/methods/protein/abs280.html
Measurement of protein solutions using absorbance at 280 nm.

http://www.ruf.rice.edu/~bioslabs/methods/protein/protcurve.html
Use of colorimeteric assays for proteins.

http://www.ruf.rice.edu/~bioslabs/methods/protein/bradford.html
The Bradford assay.

http://www.bio-rad.com/cmc_upload/Literature/55685/4110065A.pdf
The Bradford assay.

http://bioinformatics.biomed.drexel.edu/CfIB_Resources_files/Week2/RNA%20Spectrophotometry.doc
Analysis of nucleic acids.

http://www.turnerbiosystems.com/doc/appnotes/PDF/DNA_RNA_080409.pdf
Nucleic acid and protein quantitation methods.

http://www.protocol-online.org/prot/General_Laboratory_Techniques/Quantitation/index.html
Quantitation of proteins and nucleic acids.

http://everything2.com/title/Hoechst+dye
Bis-benzimidazole dye for DNA.

pH, BUFFERS, ELECTRODES, AND BIOSENSORS

K. Barker, ***At the Bench: A Laboratory Navigator,*** 2nd ed. (2005), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY).

S. Borman, ***Chem. Eng. News,*** (2006), "A Boost for Biosensors," August 7, pp. 48–49.

R. Boyer, ***Concepts in Biochemistry,*** 3rd ed. (2006), John Wiley & Sons (Hoboken, NJ), pp. 36–62.

R. Curtright et al., ***Biochem. Mol. Biol. Educ.* 32,** 71–77 (2004). "Facilitating Student Understanding of Buffering by an Integration of Mathematics and Chemical Concepts."

R. Garrett and C. Grisham, ***Biochemistry,*** 4th ed. (2010), Brooks/Cole (Belmont, CA).

N. Good and S. Izawa, in ***Methods in Enzymology,*** Vol. XXIV, A. San Pietro, Editor (1972), Academic Press (New York), pp. 53–68. "Hydrogen Ion Buffers for Photosynthesis Research."

D. E. Gueffroy, Editor, ***Buffers—A Guide for the Preparation and Use of Buffers in Biological Systems*** (1993), Calbiochem-Novabiochem Corp., P.O. Box 12087, La Jolla, CA 92039–2087.

D. Nelson and M. Cox, ***Lehninger Principles of Biochemistry***, 5th ed. (2009), Freeman (New York).

J. Risley, ***J. Chem. Educ.* 68,** 1054 (1991). "Preparing Solutions in the Biochemistry Lab."

D. Voet and J. Voet, ***Biochemistry***, 3rd ed. (2005) John Wiley & Sons (Hoboken, NJ).

D. Voet, J. Voet, and C. Pratt, ***Fundamentals of Biochemistry***, 3rd ed. (2008), John Wiley & Sons (Hoboken, NJ).

K. Wilson and J. Walker, Editors, ***Principles and Techniques of Practical Biochemistry,*** 5th ed. (2000), Cambridge University Press (Cambridge, U.K.), pp. 37–42.

http://www.hannainst.com
Complete information on the selection, use, and care of pH meters and electrodes.

http://www.ysi.com
Information on the biological oxygen monitor (oxygen electrode).

http://www.ornl.gov/sci/biosensors
Virtual poster session on biosensors in biomedicine, sponsored by the Oak Ridge National Lab.

http://en.wikipedia.org/wiki/Biosensor
Introduction to biosensors.

http://en.wikipedia.org/wiki/Buffer_solution
Buffers.

http://www.ysilifesciences.com/index.php?page=ysi-5300a-biological-oxygen-monitor
Figure 3.3 Model 5300A oxygen monitor.

RADIOISOTOPES

G. Choppin, J. Liljenzin, and J. Rydberg, ***Radiochemistry and Nuclear Chemistry,*** 3rd ed. (2002), Butterworth-Heinemann (Burlington, MA).
K. Leiser, ***Nuclear and Radiochemistry: Fundamentals and Applications,*** 2nd ed. (2002), John Wiley & Sons (New York).
R. Slater, Editor, ***Radioisotopes in Biology,*** 2nd ed. (2002), Oxford University Press (Oxford).
K. Wilson and J. Walker, Editors, ***Principles and Techniques of Practical Biochemistry,*** 5th ed. (2000), Cambridge University Press (Cambridge, U.K.), pp. 687–728.

http://beckman.com
Click on LS 6500 liquid scintillation counters.

http://en.wikipedia.org/wiki/Scintillation_counter
Scintillation counters.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2721227
Protein kinase assays avoiding radioisotope use.

http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TF9-4FDMYBX-7&_user=10&_rdoc=1&_fmt=&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=de3dc23002975c4d48063d0dfdcd6d30
Design of radioactive-free photoaffinity labels for proteins.

TECHNIQUES FOR SAMPLE PREPARATION

http://www.millipore.com/lab_filtration/clf/capability
Membranes and ultrafiltration.

http://www.thermo.com/com/cda/product/detail/0,1055,23574,00.html
or
http://www.thermo.com/eThermo/CMA/PDFs/Various/File_10350.pdf
Savant SpeedVac concentrators.

http://www.labconco.com
Freeze dryers.

http://www.labconco.com/_Scripts/editc25.asp?catid=345
Freeze dryers.

http://www.spectrapor.com
Dialysis membranes.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

CHAPTER

4

CENTRIFUGATION TECHNIQUES IN BIOCHEMISTRY

A. Basic Principles of Centrifugation 96
B. Instrumentation for Centrifugation 99
C. Applications of Centrifugation 106

Study Problems 113
Further Reading 114

A centrifuge of some kind is found in every biochemistry laboratory. Centrifuges have many applications, but they are used primarily for the preparation of biological samples and for the analysis of the physical properties of biomolecules, organelles, and cells. Centrifugation is carried out by spinning a biological sample at a high rate of speed, thus subjecting it to an intense force (artificial gravitational field). Most centrifuge techniques fit into one of two categories—**preparative centrifugation** or **analytical centrifugation.** A preparative procedure is one that can be applied to the separation or purification of biological samples (cells, organelles, macromolecules, etc.) by sedimentation. Analytical procedures are used to measure physical characteristics of biological samples. For example, the purity, size, shape, and density of macromolecules may be defined by centrifugation.

In this chapter, we will first explore the principles and theory underlying centrifugation techniques in order to provide a fundamental background. We will then turn to a discussion of the application of these techniques to the isolation and characterization of biological molecules and cellular components.

A. BASIC PRINCIPLES OF CENTRIFUGATION

A particle—whether it is a floating solid, a precipitate, a macromolecule, or a cell organelle—is subjected to a centrifugal force when it is rotated at a high rate of speed. The **centrifugal force,** F, is defined by Equation 4.1.

>> $$F = m\omega^2 r \qquad \textbf{(Eq. 4.1)}$$

where

F = intensity of the centrifugal force
m = effective mass of the sedimenting particle
ω = angular velocity of rotation in rad/sec
r = distance of the migrating particles from the central axis of rotation

The force on a sedimenting particle increases with the velocity of the rotation and the distance of the particle from the axis of rotation. A more common measurement of F, in terms of the earth's gravitational force, g, is **relative centrifugal force,** RCF, defined by Equation 4.2.

>> $$\text{RCF} = (1.119 \times 10^{-5})(\text{rpm})^2(r) \qquad \textbf{(Eq. 4.2)}$$

This equation relates RCF to revolutions per minute of the sample. Equation 4.2 dictates that the RCF on a sample will vary with r, the distance of the sedimenting particles from the axis of rotation (see Figure 4.1). It is convenient to determine RCF by use of the nomogram in Figure 4.2. It should be clear from Figures 4.1 and 4.2 that, since RCF varies with r, it is important to define r for an experimental run. Often an average RCF is determined using a value for r midway between the top and bottom of the sample container. The RCF value is reported as "a number times gravity, g."

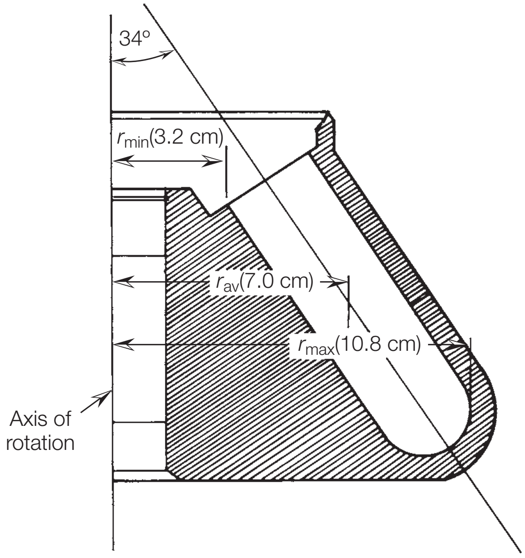

FIGURE 4.1 A diagram illustrating the variation of RCF with r, the distance of the sedimenting particles from the axis of rotation. *Courtesy of Beckman Coulter;* www.beckman.com.

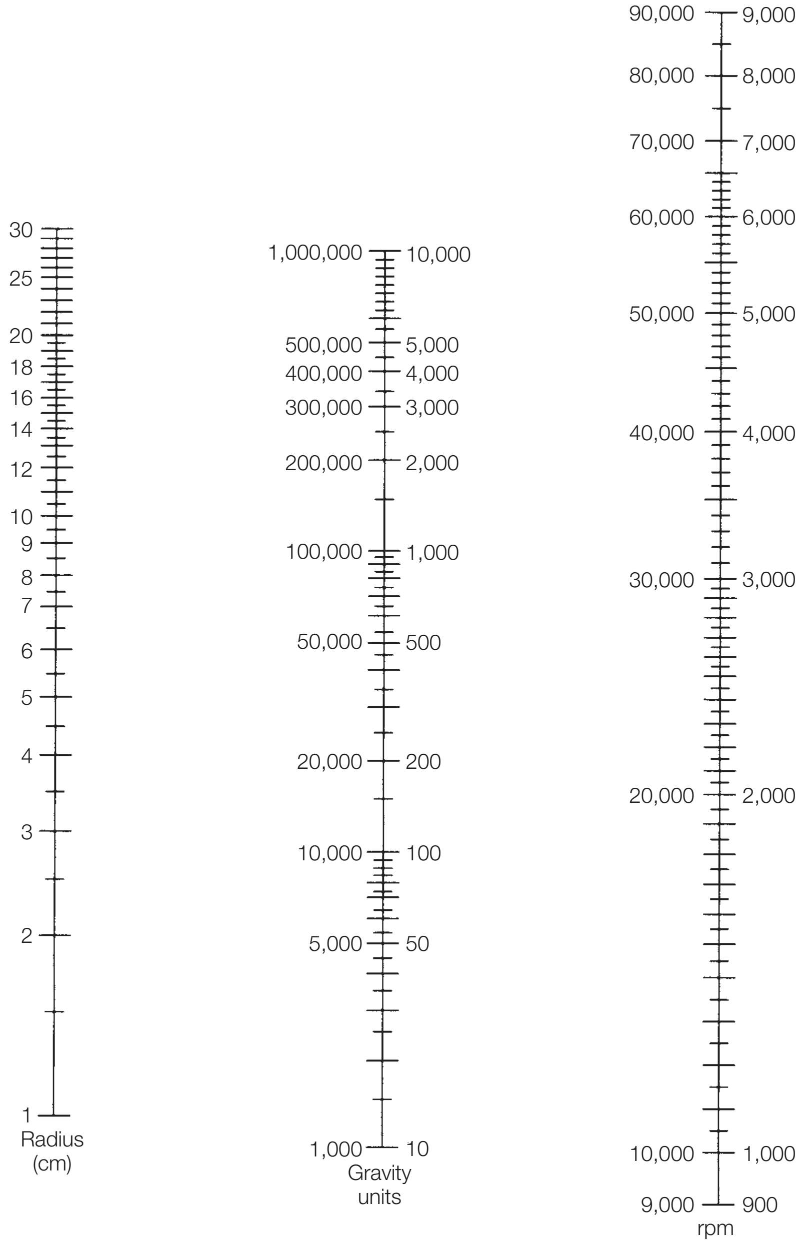

FIGURE 4.2 A nomogram for estimating RCF. To use, place a straight edge connecting the instrumental revolutions (right column) with the distance from the center of rotation (left column). The RCF is read where the straight edge crosses the middle column. *Courtesy of Beckman Coulter;* www.beckman.com.

STUDY EXERCISE 4.1 Relative Centrifugal Force

To illustrate the use of the nomogram in Figure 4.2, consider a centrifuge tube with an average r of 7 cm (see Figure 4.1) being rotated at 20,000 rpm. Use the nomogram to determine the relative centrifugal force (RCF) in gravity (g) units.

Solution: Follow the instructions in the caption for Figure 4.2 to estimate the RCF to be about $32{,}000 \times g$.

Although this introduction outlines the basic principles of centrifugation, it does not take into account other factors that influence the rate of particle sedimentation. Centrifuged particles migrate at a rate that depends on their mass, shape, and density, as well as the density of the medium. The centrifugal force felt by a particle is defined by Equation 4.1. The term m is the **effective mass** of the particle, that is, the actual mass, m_0, minus a correction factor for the weight of water displaced (buoyancy factor) (Equation 4.3).

>> $$m = m_0 - m_0\,\bar{v}\rho \qquad \textbf{(Eq. 4.3)}$$

where

m = effective mass of the sedimenting particle
m_0 = actual mass of the particle
$\bar{v}$ = partial specific volume, the volume change occurring when a particle is placed in a large excess of solvent
ρ = density of the solvent or medium

Equations 4.1 and 4.3 may be combined to describe the centrifugal force on a particle (Equation 4.4).

>> $$F = m_0(1 - \bar{v}\rho)\omega^2 r \qquad \textbf{(Eq. 4.4)}$$

As particles sediment under the influence of the centrifugal field, their movement is countered by a resistance force, the **frictional force.** The frictional force is defined by Equation 4.5.

>> $$\text{Frictional force} = fv \qquad \textbf{(Eq. 4.5)}$$

where

f = frictional coefficient
v = velocity of the sedimenting particle (sedimentation velocity)

The **frictional coefficient,** f, depends on the size and shape of the particle, as well as the viscosity of the solvent. The frictional force increases with the velocity of the particle until a constant velocity is reached. At this point, the two forces are balanced (Equation 4.6).

>> $$m_0(1 - \bar{v}\rho)\omega^2 r = fv = f\left(\frac{dr}{dt}\right) \qquad \textbf{(Eq. 4.6)}$$

The rate of sedimentation, sometimes called **sedimentation velocity,** v, is defined by Equation 4.7.

>> $$v = \frac{dr}{dt} = \frac{m_0(1 - \bar{v}\rho)\omega^2 r}{f} \quad \textbf{(Eq. 4.7)}$$

It is cumbersome and sometimes impractical to express sedimentation velocity in terms of ρ, $\bar{v}$, and f, since these factors are difficult to measure. A new term, **sedimentation coefficient,** s (the ratio of sedimentation velocity to centrifugal force) is introduced by rearranging Equation 4.7 to Equation 4.8.

>> $$s = \frac{v}{\omega^2 r} = \frac{m_0(1 - \bar{v}\rho)}{f} \quad \textbf{(Eq. 4.8)}$$

The term s is most often defined under standard conditions, 20°C and water as the medium, and denoted by $s_{20,\,\mathrm{w}}$. The s value is a physical characteristic used to classify biological macromolecules and cell organelles. Sedimentation coefficients are in the range of 1×10^{-13} to $10{,}000 \times 10^{-13}$ second. For numerical convenience, sedimentation coefficients are expressed in Svedberg units, S, where $1\,\mathrm{S} = 1 \times 10^{-13}$ second. Human hemoglobin has an s value of 4.5×10^{-13} second or 4.5 S. The value of S for several biomolecules, bacterial cells, and cell organelles is shown in Figure 4.3. Note in the figure that there appears to be a direct relationship between the S value and the molecular weight or particle size. This, however, is not always true, as in the case of nonspherical molecules.

STUDY EXERCISE 4.2 Sedimentation Coefficient

The goal of many centrifugation experiments is the measurement of s. This value is important because it can be used to calculate the size (molecular weight, kilobase pairs, etc.) of a molecule or cell organelle. The units of s are not obvious from Equation 4.8. Dimensional analysis shows the following: v in cm/sec, ω in radians/sec, r in cm, m_0 in grams, $\bar{v}$ in cm^3/g, ρ in g/cm^3, and f in g/sec. Determine the units for s. (Ans. = second)

B. INSTRUMENTATION FOR CENTRIFUGATION

The basic centrifuge consists of two components, an electric motor with drive shaft to spin the sample and a **rotor** to hold tubes or other containers of the sample. A wide variety of centrifuges is available, ranging from a low-speed centrifuge used for routine pelleting of relatively heavy particles to sophisticated instruments that include accessories for making analytical measurements during centrifugation. Here we will describe three types, the low-speed or clinical centrifuge; the high-speed centrifuge, including the "microfuge"; and the ultracentrifuge. Major characteristics and applications of each type are compared in Table 4.1.

Low-Speed Centrifuges

Most laboratories have a standard low-speed centrifuge used for routine sedimentation of relatively heavy particles. The common centrifuge has a maximum speed in the range of 4000 to 5000 rpm, with RCF averaging 3000 × g.

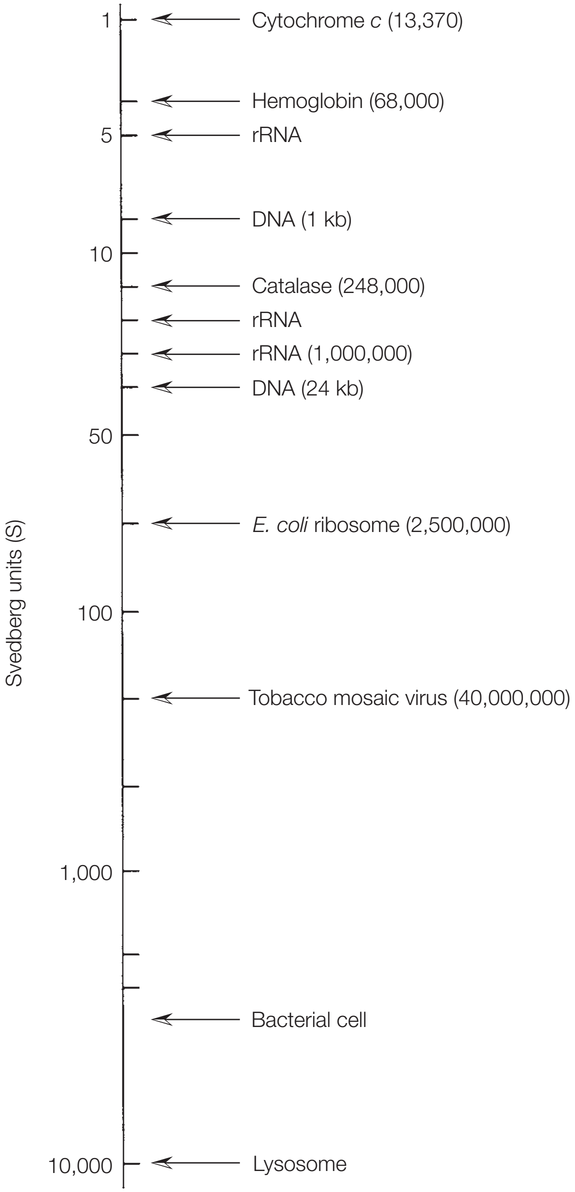

FIGURE 4.3 Range of S values for biomolecules, cell organelles, and cells. kb = kilobase pairs. Molecular weight, or kb, is shown in parentheses.

These instruments usually operate at room temperature with no means of temperature control of the samples. Two types of rotors, **fixed angle** and **swinging bucket,** may be used in the instrument. Centrifuge tubes or bottles that contain 12 or 50 mL of sample are commonly used. Low-speed centrifuges are especially useful for the rapid sedimentation of coarse precipitates or red blood cells. The sample is centrifuged until the particles are tightly packed into a **pellet** at the bottom of the tube. (This technique is sometimes

TABLE 4.1 Types of Centrifuges and Applications

	Type of Centrifuge			
Characteristic	**Low Speed**	**Microfuge (Medium Speed)**	**High Speed**	**Ultracentrifuge**
Range of speed (rpm)	1–6000	1000–15,000	1000–25,000	20–150,000
Maximum RCF (g)	6000	12,000	50,000	1,500,000
Refrigeration	Some	Some	Yes	Yes
Applications				
Pelleting of cells	Yes	Yes	Yes	Yes
Pelleting of nuclei	Yes	Yes	Yes	Yes
Pelleting of organelles	No	No	Yes	Yes
Pelleting of ribosomes	No	No	No	Yes
Pelleting of macromolecules	No	No	No	Yes
Analytical techniques	No	No	No	Yes

called **pelleting.**) The upper liquid portion, the **supernatant,** is then separated by decantation.

High-Speed Centrifuges

For more sophisticated biochemical applications, higher speeds and temperature control of the rotor chamber are essential. A typical high-speed centrifuge is shown in Figure 4.4. The operator of this instrument can carefully control speed and temperature, which is especially important for carrying out reproducible

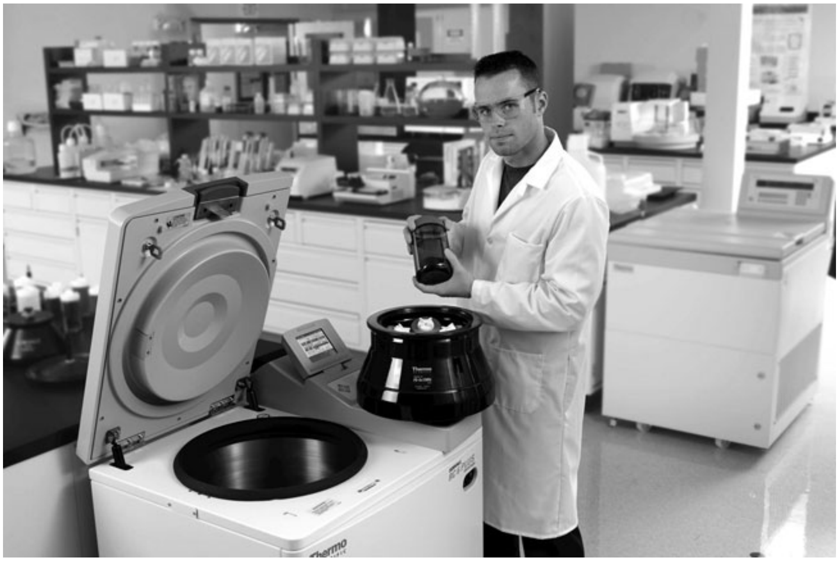

FIGURE 4.4 A lab technician loads samples into a Sorvall RC-6 Plus Superspeed centrifuge, a typical high-speed centrifuge. *Image provided by Thermo Fisher Scientific Inc., all rights reserved.* www.thermofisher.com.

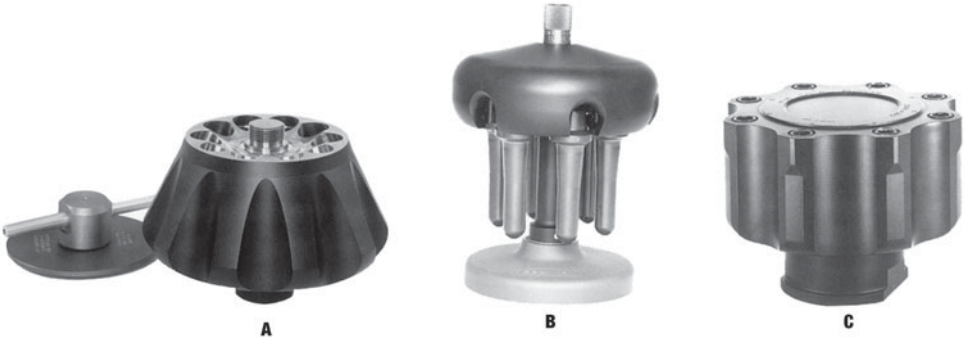

FIGURE 4.5 Rotors for a high-speed centrifuge. **A** Fixed-angle; **B** Swinging-bucket; **C** Vertical. *Courtesy of Beckman Coulter.*

centrifugations of temperature-sensitive biological samples. Rotor chambers in most instruments are maintained at or near 4°C.

Three types of rotors are available for high-speed centrifugation: the fixed-angle, the swinging-bucket, and the vertical rotor (Figure 4.5A–C). Fixed-angle rotors are especially useful for differential pelleting of particles (Figure 4.6A). In swinging-bucket rotors (Figure 4.5B), the sample tubes move to a position perpendicular to the axis of rotation during centrifugation, as shown in Figure 4.7. These are used most often for density gradient centrifugation (see below). In the vertical rotor (Figure 4.5C), the sample tubes remain in an upright position (Figure 4.8). These rotors are used often for gradient centrifugation. Prior to the

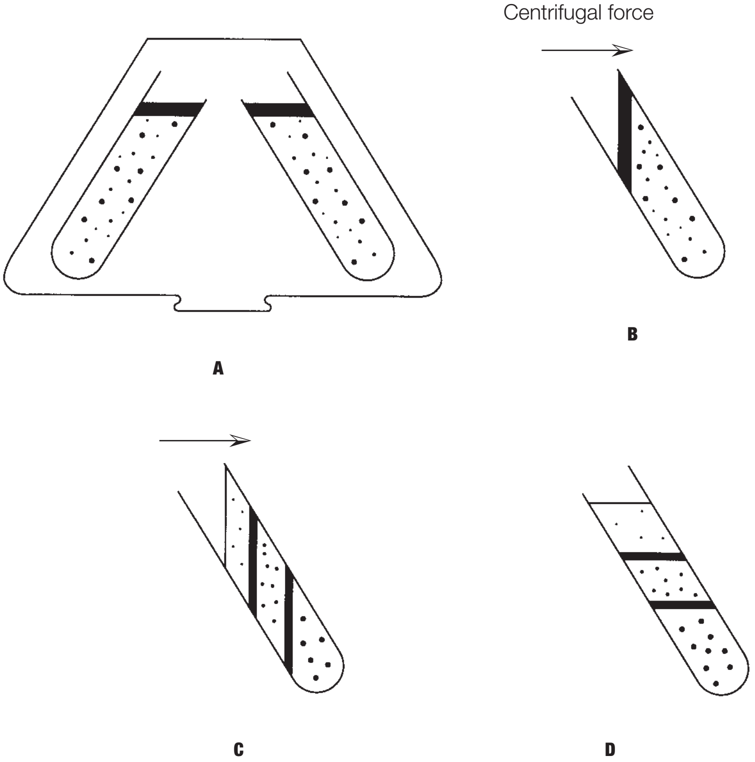

FIGURE 4.6 Operation of a fixed-angle rotor. **A** Loading of sample. **B** Sample at start of centrifugation. **C** Bands form as molecules sediment. **D** Rotor at rest showing separation of two components.

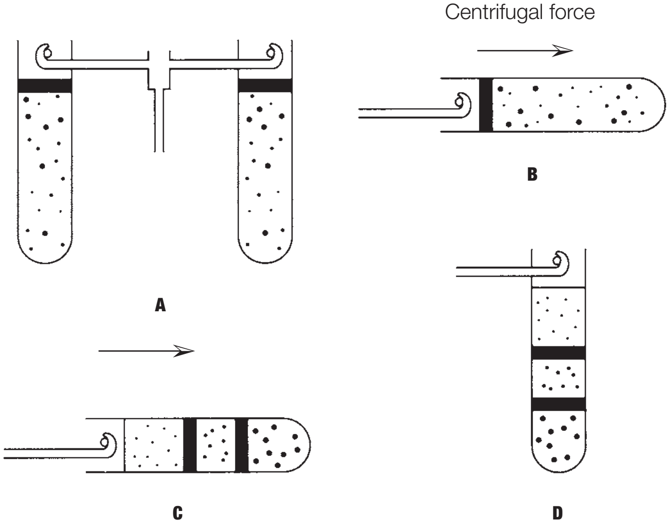
Centrifugal force
A
B
C
D

FIGURE 4.7 Operation of a swinging-bucket rotor. **A** Loading of sample. **B** Sample at start of centrifugation. **C** Sample during centrifugation separates into two components. **D** Rotor at rest.

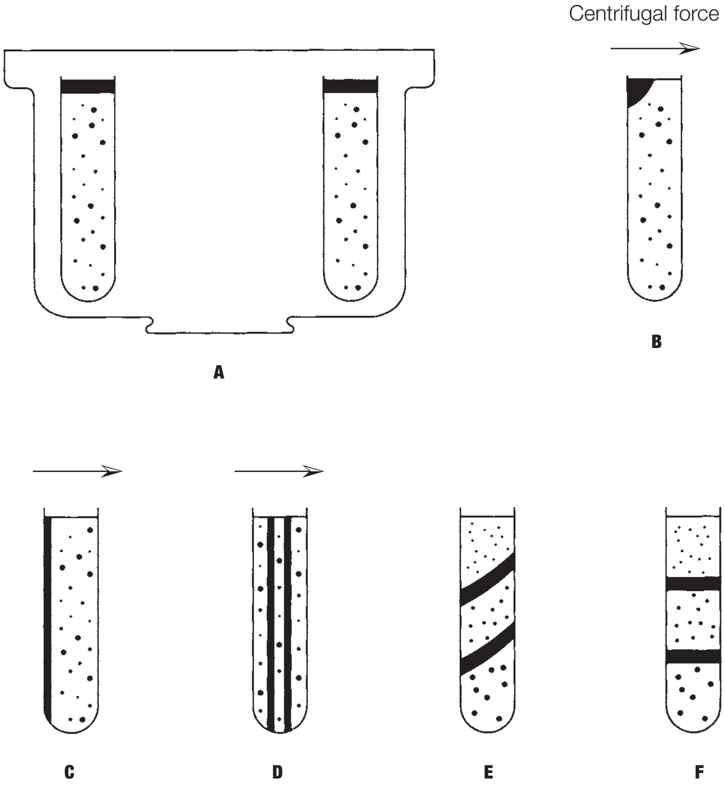
Centrifugal force
A
B
C
D
E
F

FIGURE 4.8 Operation of a vertical rotor. **A** Loading of sample. **B** Beginning of centrifugation. **C, D** During centrifugation. **E** Deceleration of sample. **F** Rotor at rest.

early 1990s, rotors were constructed from metals such as aluminum and titanium. Although metal rotors have great strength, they do have several disadvantages: they are very heavy to handle, they are not corrosion-resistant, and they become fatigued with use. Rotors are now available that are fabricated from carbon-fiber composite materials. They have several advantages over heavy metal rotors. These new rotors are 60% lighter than comparable aluminum and titanium rotors. Because of the lighter weight, acceleration and deceleration times are reduced; thus, centrifuge run times are shorter. This also results in lower service and maintenance costs. Instruments are equipped with a brake to slow the rotor rapidly after centrifugation.

Widely used in the category of medium-speed centrifuges is the **microfuge** (Table 4.1 and Figure 4.9). These instruments, which are designed for the bench-top, are used for rapid pelleting of small samples. Fixed-angle rotors are able to hold up to eighteen 1.5- or 0.5-mL tubes. The maximum speed of most commercial microfuges is between 12,000 and 15,000 rpm, which delivers a force of 11,000−12,000 × g. Some instruments can accelerate to full speed in 6 seconds and decelerate within 18 seconds. Most instruments have a variable speed control and a momentary pulse button for minispins.

The preparation of biological samples almost always requires the use of a high-speed centrifuge. Specific examples will be described later, but high-speed centrifuges may be used to sediment: (1) cell debris after cell homogenization, (2) ammonium sulfate precipitates of proteins, (3) microorganisms, and (4) cellular organelles such as chloroplasts, mitochondria, and nuclei.

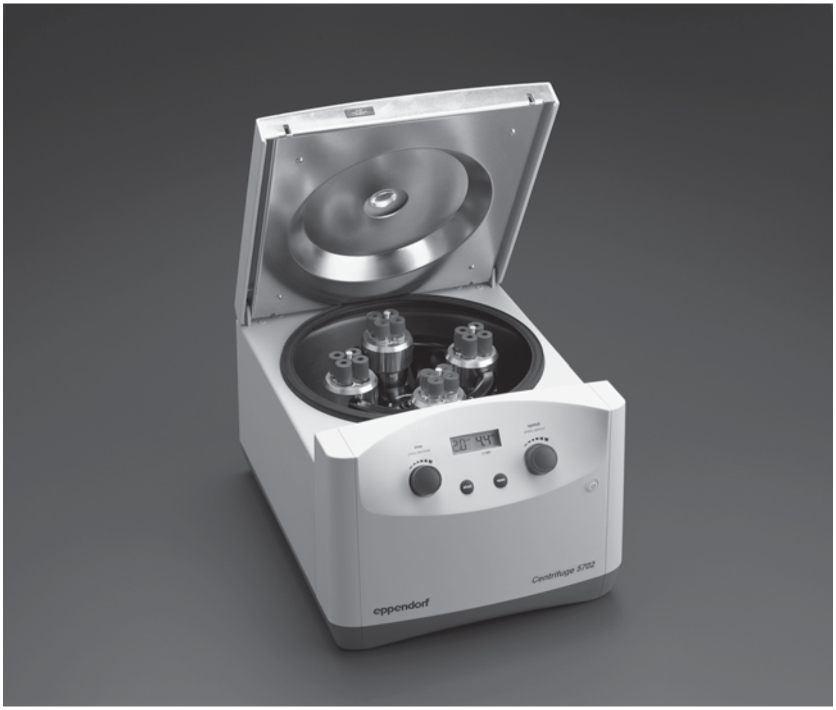

FIGURE 4.9 A variable-speed, general-purpose microcentrifuge—Eppendorf® Centrifuge Model 5702. *Courtesy of Metrohm, Inc.* http://www.metrohmusa.com/home.

Ultracentrifuges

The most sophisticated of the centrifuges are the **ultracentrifuges.** Because of the high speeds attainable (see Table 4.1), intense heat is generated in the rotor. The spin chamber must thus be refrigerated and placed under a high vacuum to reduce friction. The sample in a cell or tube is placed in a rotor, which is then driven by an electric motor. Although it is rare, metal rotors sometimes break into fragments when placed under high stress. The rotor chamber on all ultracentrifuges is covered with a protective steel armor plate. The drive shaft of the ultracentrifuge is constructed of a flexible material to accommodate any "wobble" of the rotor due to imbalance of the samples. It is still important to counterbalance samples as carefully as possible before inserting them in the rotor.

The previously discussed centrifuges—the low, medium, and high speed—are of value only for preparative work, that is, for the isolation and separation of precipitates and biological samples. Ultracentrifuges can be used both for preparative work and for analytical measurements. Thus, two types of ultracentrifuges are available: **preparative models,** primarily used for separation and purification of samples for further analysis, and **analytical models,** which are designed for performing physical measurements on the sample during sedimentation. Two of the most versatile models are Beckman Optima MAX and TLX microprocessor-controlled tabletop ultracentrifuges (Figure 4.10). With a typical fixed-angle rotor, which holds six 0.2- to 2.2-mL samples, the instruments can generate 150,000 rpm and an RCF of up to 1,500,000 $\times g$.

Analytical ultracentrifuges have the same basic design as preparative models, except that they are equipped with optical systems to monitor directly the sedimentation of the sample during centrifugation. The first commercial instrument of this type was the Beckman Model E, introduced in 1947.

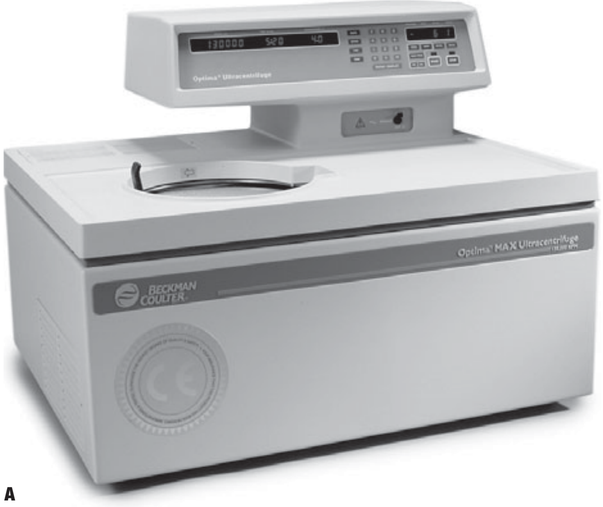

FIGURE 4.10
A Outside and **B** inside of the Beckman Optima TLX ultracentrifuge. *Courtesy of Beckman Coulter. Beckman Coulter and the stylized logo are registered trademarks of Beckman Coulter, Inc. Optima is a trademark of Beckman Coulter, Inc.*

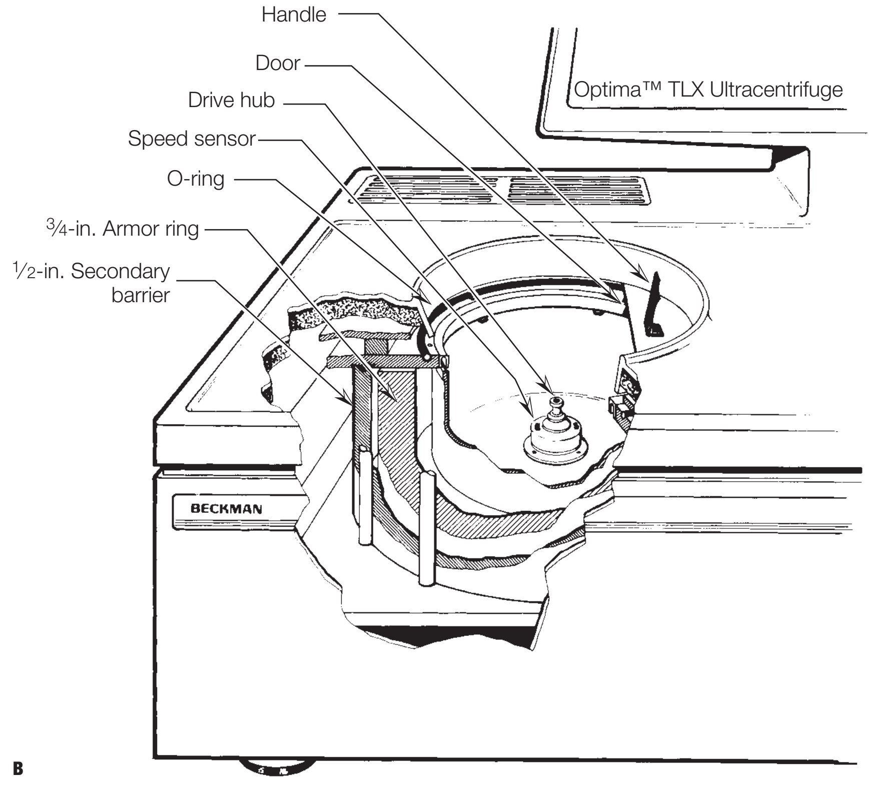

FIGURE 4.10
Continued

For analysis, a sample of nucleic acid or protein (0.1 to 1.0 mL) is sealed in a special analytical cell and rotated. Light is directed through the sample parallel to the axis of rotation, and measurements of absorbance by sample molecules are made. (The Beckman instrument can scan the sample over the wavelength range 190 to 800 nm.) If sample molecules have no significant absorption bands in the wavelength range, then optical systems that measure changes in the refractive index may be used. Optical systems aided by computers are capable of relating absorbance changes or index of refraction changes to the rate of movement of particles in the sample. The optical system actually detects and measures the front edge or moving boundary of the sedimenting molecules. These measurements can lead to an analysis of concentration distributions within the centrifuge cell. Applications of these measurements will be discussed in the next section.

C. APPLICATIONS OF CENTRIFUGATION

Preparative Techniques

Centrifuges in undergraduate biochemistry laboratories are used most often for preparative-scale separation procedures. This technique is quite straightforward, consisting of placing the sample in a tube or similar container, inserting

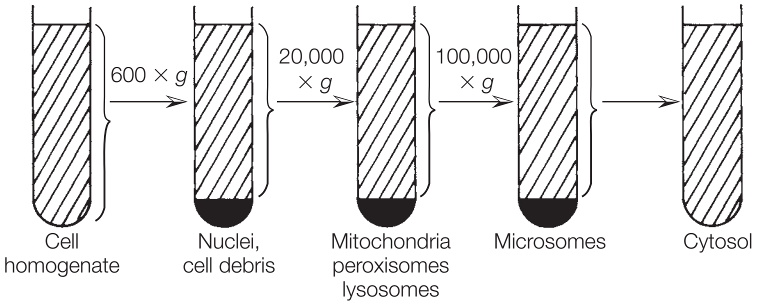

FIGURE 4.11 Fractional centrifugation of a cell homogenate. See text for details.

the tube in the rotor, and spinning the sample for a fixed period. **(Filled centrifuge tubes or bottles must be weighed and balanced before centrifugation. Tubes across from each other in the rotor should have approximately the same weight.)** The sample is removed and the two phases, pellet and supernatant (which should be readily apparent in the tube), may be separated by careful decantation. Further characterization or analysis is usually carried out on the individual phases. This technique, called **velocity sedimentation centrifugation,** separates particles ranging in size from coarse precipitates to cellular organelles. Relatively heavy precipitates are sedimented in low-speed centrifuges, whereas lighter organelles such as ribosomes require the high centrifugal forces of an ultracentrifuge.

Much of our current understanding of cell structure and function depends on separation of subcellular components by centrifugation. The specific method of separation, called **fractional centrifugation,** consists of successive centrifugations at increasing rotor speeds. Figure 4.11 illustrates the fractional centrifugation of a cell homogenate, leading to the separation and isolation of the common cell organelles. For most biochemical applications, the rotor chamber must be kept at low temperatures to maintain the native structure and function of each cellular organelle and its component biomolecules. A high-speed centrifuge equipped with a fixed-angle rotor is most appropriate for the first two centrifugations at 600 × *g* and 20,000 × *g*. After each centrifuge run, the supernatant is poured into another centrifuge tube, which is then rotated at the next higher speed. The final centrifugation at 100,000 × *g* to sediment microsomes and ribosomes must be done in an ultracentrifuge. The 100,000 × *g* supernatant, the **cytosol,** is the soluble portion of the cell and consists of soluble proteins and smaller molecules.

STUDY EXERCISE 4.3 Fractional Centrifugation

The enzymes required for the metabolic process of glycolysis are located in the cytoplasm of the cell. Describe how you could use centrifugation to prepare a cell extract containing these enzymes separated from cell organelles.

Analytical Measurements

A variety of analytical measurements can be made on biological samples during and after a centrifuge run. Most often, the measurements are taken to determine molecular weight, density, and purity of biological samples. All analytical techniques require the use of an ultracentrifuge and can be classified as **differential** or **density gradient.**

Differential Centrifugation

Differential methods involve sedimentation of particles in a medium of homogeneous density. Although the technique is similar to preparative differential centrifugation as previously discussed, the goal is to measure the sedimentation coefficient of a particle. The underlying principles of this technique are illustrated in Figure 4.12A. During centrifugation, a moving boundary is generated between pure solvent and sedimenting particles. An analytical ultracentrifuge is capable of detecting and measuring the rate of movement of the boundary. Hence, the sedimentation velocity, v, can be experimentally determined. By using Equation 4.8, the sedimentation coefficient, s, can be calculated. The value of s for a sedimenting particle is related to the molecular weight of that particle by the Svedberg equation. The Svedberg equation is derived from Equation 4.8 by recognizing that the frictional force, f, may be defined by Equation 4.9.

>> $$f = \frac{RT}{ND} \qquad \textbf{(Eq. 4.9)}$$

where

R = the gas constant, 8.3×10^7 g cm^2/sec/deg/mole
T = the absolute temperature
D = the diffusion coefficient of the solute in units of cm^2/sec
N = Avogadro's number

Thus, Equation 4.8 may be transformed into Equation 4.10.

>> $$\mathrm{s} = \frac{m_0\,(1 - \bar{v}\rho)}{RT/ND} \qquad \textbf{(Eq. 4.10)}$$

$$RTs = m_0\,ND\,(1 - \bar{v}\rho)$$

Since molecular weight, MW, is equal to m_0N, Equation 4.10 is converted to Equation 4.11, the Svedberg equation. In this book, two terms are used to define the molecular masses of biomolecules:

Molecular weight (MW): defined as the ratio of the particle mass to 1/12th the mass of a C-12 atom. MW has no units because it is a ratio.

Molecular mass: defined as the mass of one molecule, or the molar mass divided by Avogadro's number. Molecular masses are expressed in the units of daltons (D, 1/12th the mass of a C-12 atom) or kilodaltons (kD).

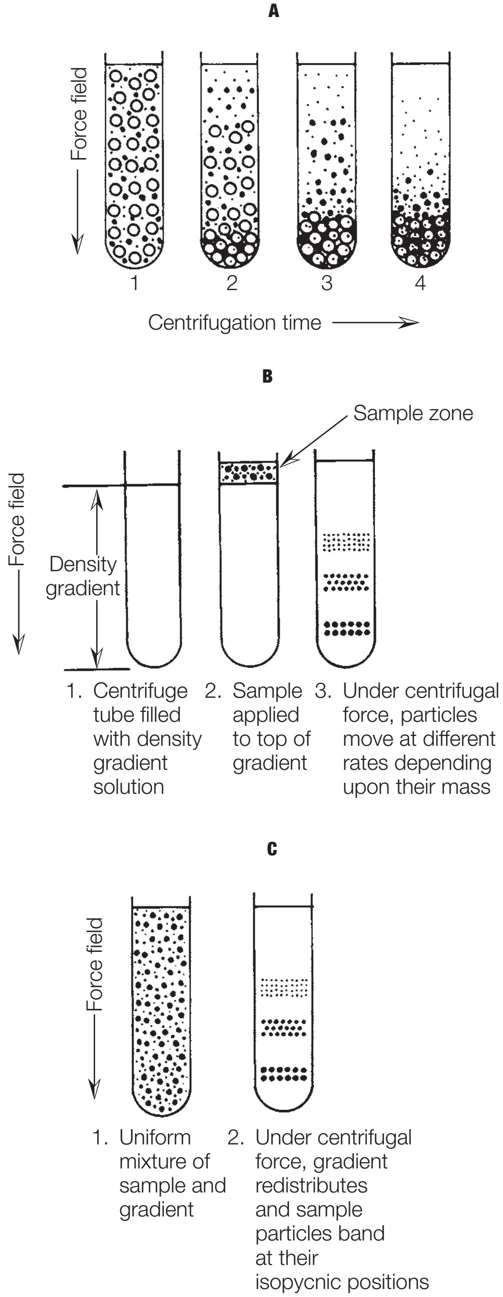
A
Force field
1
2
3
4
Centrifugation time
B
Sample zone
Force field
Density gradient
1. Centrifuge tube filled with density gradient solution
2. Sample applied to top of gradient
3. Under centrifugal force, particles move at different rates depending upon their mass
C
Force field
1. Uniform mixture of sample and gradient
2. Under centrifugal force, gradient redistributes and sample particles band at their isopycnic positions

FIGURE 4.12 A comparison of differential and density gradient measurements. **A** Differential centrifugation in a medium of unchanging density. **B** Zonal centrifugation in a prepared density gradient. **C** Isopycnic centrifugation; the density gradient forms during centrifugation.

>>

$$MW = \frac{RTs}{D(1 - \bar{v}\rho)} \quad \textbf{(Eq. 4.11)}$$

This equation provides an accurate calculation of molecular weight and is applicable to macromolecules such as proteins and nucleic acids. However, its usefulness is limited because diffusion coefficients are difficult to measure and are not readily available in the literature.

An alternative method sometimes used to determine molecular weights of macromolecules is **sedimentation equilibrium.** In the previous example, using the Svedberg equation, the sample is rotated at a rate sufficient to sediment the particles. Here, the sample is rotated at a lower rate, and the particles sediment until they reach an equilibrium position at the point where the centrifugal force is equal to the frictional component opposing their movement (see Equation 4.6). The molecular weight is then calculated using Equation 4.12.

>>

$$\mathrm{MW} = \frac{RT}{(1 - \bar{v}\rho)\omega^2}\left(\frac{1}{rc}\right)\left(\frac{dc}{dr}\right) \qquad \textbf{(Eq. 4.12)}$$

where

$\frac{dc}{dr}$ = change in particle concentration as a function of distance from the rotation center, as measured in the ultracentrifuge

and MW, R, T, $\bar{v}$, ρ, ω, and r have the definitions previously given. This technique is time consuming, as the system may require 1 to 2 days to reach equilibrium. Also, the use of Equation 4.12 is complicated by the difficulty of making concentration measurements in the ultracentrifuge.

Differential ultracentrifugation methods may also be used in analysis of the purity of macromolecular samples. If a single sharp moving boundary is observed in a rotating centrifuge cell, it indicates that the sample has one component and therefore is pure. In an impure sample, each component would be expected to form a separate moving boundary upon sedimentation.

Density-Gradient Centrifugation

In differential procedures, the sample is uniformly distributed in a cell before centrifugation, and the initial concentration of the sample is the same throughout the length of the centrifuge cell. Although useful analytical measurements can be made with this technique, it has disadvantages when applied to impure samples or samples with more than one component. Large particles that sediment faster pass through a medium consisting of solvent and particles of smaller size. Therefore, clear-cut separations of macromolecules are seldom obtained. This can be avoided if the sample is centrifuged in a fluid medium that gradually increases in density from top to bottom. This technique, called **density gradient centrifugation,** permits the separation of multi-component mixtures of macromolecules and the measurement of sedimentation coefficients. Two methods are used, **zonal centrifugation,** in which the sample is centrifuged in a preformed gradient, and **isopycnic centrifugation,** in which a self-generating gradient forms during centrifugation.

Zonal Centrifugation

Figure 4.12B outlines the procedure for zonal centrifugation of a mixture of macromolecules. A density gradient is prepared in a tube prior to centrifugation. This is accomplished with the use of an automatic gradient mixer. Solutions of low-molecular-weight solutes such as sucrose or glycerol are allowed to flow into the centrifuge cell. The sample under study is layered on top of the gradient and placed in a swinging-bucket rotor. Sedimentation in an ultracentrifuge results in movement of the sample particles at a rate dependent on their individual s values. As shown in Figure 4.12B, the various types of particles sediment as *zones* and remain separated from the other components. The centrifuge run is terminated before any particles reach the bottom of the gradient. The various zones in the centrifuge tubes are then isolated by collecting fractions from the bottom of the tube and analyzing them for the presence of macromolecules. The zones of separated macromolecules are relatively stable in the gradient because it slows diffusion and convection. The gradient conditions can be varied by using different ranges of sucrose concentration. Sucrose concentrations up to 60% can be used, with a density limit of $1.28\ g/\mathrm{cm}^3$. The zonal method can be applied to the separation and isolation of macromolecules (preparative ultracentrifuge) and to the determination of s (analytical ultracentrifuge).

Isopycnic Centrifugation

In the isopycnic technique, the density gradient is formed during the centrifugation. Figure 4.12C outlines the mechanism of isopycnic centrifugation. The sample under study is dissolved in a solution of a dense salt such as cesium chloride or cesium sulfate. The cesium salts may be used to establish gradients to an upper density limit of $1.8\ g/\mathrm{cm}^3$. The solution of biological sample and cesium salt is uniformly distributed in a centrifuge tube and rotated in an ultracentrifuge. Under the influence of the centrifugal force, the cesium salt redistributes to form a continuously increasing density gradient from the top to the bottom. The macromolecules of the biological sample seek an area in the tube where the density is equal to their respective densities. That is, the macromolecules move to a region where the sum of the forces (centrifugal and frictional) is zero (Equation 4.6). The macromolecules either sediment or float to this region of equal density. Stable zones or bands of the individual components are formed in the gradient (see Figure 4.12C). These bands can be isolated as previously described. Cesium salt gradients are especially valuable for separation of nucleic acids.

Density gradients are widely used in separating and purifying biological samples. In addition to this preparative application, measurements of s can be made. Gradient techniques used to isolate and purify subcellular components such as microsomes, ribosomes, lysosomes, mitochondria, peroxisomes, chloroplasts, and others. After isolation, they may be biochemically characterized as to their protein, lipid, nucleic acid, and enzyme contents.

Nucleic acids, in particular, have been extensively studied by density gradient techniques. Both RNA and DNA are routinely classified according to their s values. The different structural forms of DNA discussed in Chapter 9 can be determined by density gradient centrifugation.

A summary of centrifugation techniques and applications described in this chapter is provided in Table 4.2. Space does not allow for an exhaustive review of centrifuge applications. Interested students should consult the references at the end of the chapter for recent developments.

TABLE 4.2 A Summary of Centrifuge Techniques and Applications

Centrifuge Techniques	Applications
Preparative	
Velocity sedimentation centrifugation (pelleting)	Separation and isolation of particles in a solution. May be applied to precipitates, cell organelles, cells, or biomolecules.
Fractional centrifugation	Isolation of particles, based on size, by successive centrifugation at increasing rotor speeds. May be used to separate cell organelles (Figure 4.11).
Analytical	
Differential centrifugation	Sedimentation of particles in a medium of homogeneous density. Used to measure the sedimentation coefficient, *s*, and MW of a particle (Figure 4.12A).
Sedimentation equilibrium centrifugation	Used to determine MW of a macromolecule or other particle.
Density gradient centrifugation	
Zonal	Gradient is present in the tube before centrifugation and sample is layered on top. Used to isolate purified molecules and determine *s*.
Isopycnic	Gradient is formed during centrifugation. Used to isolate purified molecules and determine *s*.

Care of Centrifuges and Rotors

Centrifuge equipment represents a sizable investment for a laboratory, so proper maintenance is essential. In addition, poorly maintained equipment is especially dangerous. Since many kinds of instruments are now available, specific instructions will not be given here, but general guidelines are outlined below.

1. Carefully read the operating manual or receive proper instructions before you use any centrifuge.
2. Check the rotor chamber for cleanliness and for damage. Clean with soap and warm water, rinse with distilled water, and dry.
3. Select the proper operating conditions on the instrument. If refrigeration is necessary, set the temperature to the appropriate level and allow 1 to 2 hours for temperature equilibration.
4. Select the proper rotor. Many sizes and types are available. Follow guidelines already stated in this chapter or consult your instructor.
5. Be sure the rotor is clean and undamaged. Observe any nicks, scratches, or other damage that may cause imbalance. If dirty, the rotor should be cleaned with warm water and a mild, nonbasic detergent. A soft brush can be used inside the cavities. Rinse well with distilled water and dry. Scratches should not be made on the surface coating, as corrosion may result.
6. Filled centrifuge tubes or bottles should be weighed carefully and balanced before centrifugation.
7. Rotor manufacturers provide a maximum allowable speed limit for each rotor. Do not exceed that limit.

8. Keep an accurate record of centrifuge and rotor use. Just as your automobile needs service after a certain number of miles, the centrifuge should be serviced after certain intervals of use. Centrifuge maintenance is usually determined by hours of use and total revolutions of the rotor. It is also essential to maintain a record of the use for each rotor. Metal rotors weaken with use, and the maximum allowable speed limit decreases. Rotor manufacturers usually provide guidelines for decreasing the allowable speed for a rotor (derating a rotor).
9. If an unusual noise or vibration develops during centrifugation, immediately turn the centrifuge off.
10. Carefully clean the rotor chamber and rotor after each centrifugation.

Study Problems

1. Which of the following factors will have an effect on the sedimentation rate of a particle during centrifugation?
   (a) Mass of the sedimenting particle
   (b) Angular velocity of rotation
   (c) Atmospheric pressure
   (d) Density of the solvent
2. You wish to centrifuge a biological sample so that it experiences an RCF of 100,000 × *g*. At what rpm must you set the centrifuge assuming an average *r* value of 4?
3. Cytochrome *c* has an *s* value of 1×10^{-13} second and hemoglobin an *s* value of about 4.5×10^{-13} second. Which protein has the larger molecular weight?
4. An enzyme has a sedimentation coefficient of 3.5 S. When a substrate molecule is bound into the active site of the enzyme, the sedimentation coefficient decreases to 3.0 S. Explain this change.
5. A protein with a molecular weight of 100,000 shows a single boundary when centrifuged in aqueous buffer. If the protein is centrifuged in a medium of the same buffer plus 6 *M* urea, two boundaries are observed, one corresponding to a molecular weight of 10,000, the other to a molecular weight of 30,000. The area of the slower peak is two-thirds that of the faster. Describe the subunit structure of the protein.
6. Describe how you would design a centrifuge experiment to isolate sediments containing cell nuclei.
7. Explain the following observation. The density of DNA in 7 *M* CsCl containing 0.15 *M* $MgCl_2$ is less than the density of the same DNA in pure 7 *M* CsCl.
8. Assume that you have centrifuged in a density gradient a sample of DNA that contained both closed, circular DNA and supercoiled DNA. Would you expect to see two bands in the sedimentation pattern? Explain.
9. Assume that a centrifuge is operating at 43,000 rpm. What is the relative centrifugal force at a distance from the central axis of 6 cm?
10. You are spinning a sample at 20,000 rpm in a centrifuge. Use Equation 4.2 to calculate the RCF at three locations in the tube, $r = 3.2$ cm, $r = 7.0$ cm, and $r = 10.8$ cm (see Figure 4.1). Compare and explain the differences in the RCF values at the three locations.
11. Could you use a low-speed centrifuge to sediment mitochondria? Explain.
12. *T*he sedimentation coefficient (S, in Svedberg units) for an unknown protein you are studying is between 1 and 2. If you isolate the protein from a cell homogenate by carrying out fractional centrifugation (Figure 4.11), in what fraction do you expect to find the unknown protein?

Further Reading

J. Berg, J. Tymoczko, and L. Stryer, ***Biochemistry, 6th ed.*** (2007), W. H. Freeman (New York). Centrifugation in biochemistry.

D. Cantor and P. Schimmel, ***Biophysical Chemistry,*** Part II (1980), W. H. Freeman (San Francisco), pp. 591–641. Excellent, but advanced, treatment of theory and applications of analytical centrifugation.

R. Garrett and C. Grisham, ***Biochemistry, 4th ed.*** (2010), Cengage Learning (Belmont, CA). Techniques of centrifugation.

J. Graham and D. Rickwood, ***Biological Centrifugation*** (2001), Springer Verlag (New York). Applications of centrifugation in biology.

E. Holme and H. Peck, ***Analytical Biochemistry, 3rd ed.*** (1998), Addison-Wesley Longman (Essex, U.K.), pp. 153–163. Principles of centrifugation.

N. Price, R. Dwek, and M. Wormald, ***Principles and Problems in Physical Chemistry for Biochemists, 3rd ed.*** (2002), Oxford University Press (Oxford).

S. Sawhney and R. Singh (Editors), ***Introductory Practical Biochemistry,*** (2005), Alpha Science International (Oxford, U. K.).

D. Sheehan, ***Physical Biochemistry: Principles and Applications, 2nd ed.*** (2009), Wiley-Blackwell (Hoboken, NJ), pp. 266–278. Centrifugation.

I. Tinoco, K. Sauer, J. Wang, and J. Puglisi, ***Physical Chemistry: Principles and Applications in Biological Sciences, 4th ed.*** (2002), Prentice Hall (Upper Saddle River, NJ). Chapter 6 covers centrifugation and sedimentation.

D. Voet and J. Voet, ***Biochemistry, 3rd ed.*** (2005), John Wiley & Sons (Hoboken, NJ). Centrifugation in protein purification.

D. Voet, J. Voet, and C. Pratt, ***Fundamentals of Biochemistry, 3rd ed.*** (2008), John Wiley & Sons (Hoboken, NJ). Ultracentrifugation of proteins and nucleic acids.

K. Wilson and J. Walker, Editors, ***Principles and Techniques of Practical Biochemistry, 5th ed.*** (2000), Cambridge University Press (Cambridge), pp. 263–311. “Techniques in Centrifugation.”

http://www.beckman.com
Click on “Centrifuges” and follow links to “Applications,” “Literature,” “Rotor Calculations,” etc.

http://www.metrohmusa.com
Eppendorf centrifuges.

http://www.thermo.com/com/cda/landingpage/0,,128,00.html
Thermo Fisher Scientific Sorvall centrifuges.

http://www.piramoon.com
New carbon-fiber rotors by Fiberlite Centrifuge, Inc.

Additional Web sites for this chapter are available on the book’s Companion Website, which can be found at http://www.chemplace.com.

CHAPTER

5

PURIFICATION AND ANALYSIS OF BIOMOLECULES BY CHROMATOGRAPHY

A. Introduction to Chromatography 116
B. Planar Chromatography (Paper and TLC) 118
C. Column Chromatography 122
D. Ion-Exchange Chromatography 126
E. Gel-Exclusion Chromatography 132
F. High-Performance Liquid Chromatography (HPLC) 140
G. Affinity Chromatography and Immunoadsorption 152
H. Membrane-Based Chromatography 159

Study Problems 161
Further Reading 163

The primary goal of biochemical research is to understand the molecular nature of life processes. The molecular details of a biological process cannot be fully elucidated until the interacting molecules have been isolated and characterized. Therefore, our understanding of the mechanisms of life processes has increased at about the same pace as the development of techniques for the separation and characterization of biomolecules.

Chromatography, the most important technique for isolating and purifying biomolecules, was developed by Mikhail Tswett, an Italian-born, Russian botanist. In 1902, Tswett began his studies on the isolation and characterization of the colorful pigments in plant chloroplasts. He prepared separating columns by packing fine powders like sucrose and chalk (calcium carbonate) into long glass tubes. He then poured petroleum ether-derived plant extracts through the columns. As he continued eluting the columns with solvent, he noted the formation of yellow and green zones. Tswett had invented "chromatography," which he defined in his 1906 publication as “a method in which the components of a mixture are separated on an adsorbent in a flowing solvent.” In addition to introducing a new technique, Tswett also showed by these experiments that chlorophyll exists in different forms. From such humble beginnings, chromatography has developed into the ultimate tool for not only the isolation and purification, but also for the characterization, of biomolecules. Chromatography, which has now been expanded into multiple forms, continues to be the most effective technique for separating and purifying all types of biomolecules. In addition, it is widely used as an analytical tool to measure biophysical and other quantitative properties of molecules.

A. INTRODUCTION TO CHROMATOGRAPHY

All types of chromatography are based on a very simple concept: The sample to be examined is allowed to interact with two physically distinct entities—a **mobile phase** and a **stationary phase** (see Figure 5.1). The sample most often contains a mixture of several components to be separated. The molecules targeted for analysis are called **analytes**. The mobile phase, which may be a liquid or gas, moves the sample components through a region containing the solid or liquid stationary phase, which is called the **sorbent**. Because it varies from one chromatographic method to another, the stationary phase will not be described in detail at this time. However, it may be considered as having the ability to “bind” some types of analytes. The molecular components in the sample distribute themselves between the mobile phase and sorbent and thus have the opportunity to interact intimately with the stationary phase. If some of the sample molecules (analytes) are preferentially bound by the sorbent, they spend more time in the sorbent and are retarded in their movement through the chromatographic system. Molecules that show weak affinity for the sorbent spend more time with the mobile phase and are more easily removed or **eluted** from the system. The many interactions that occur between analytes and the stationary phase sorbent bring about a separation of

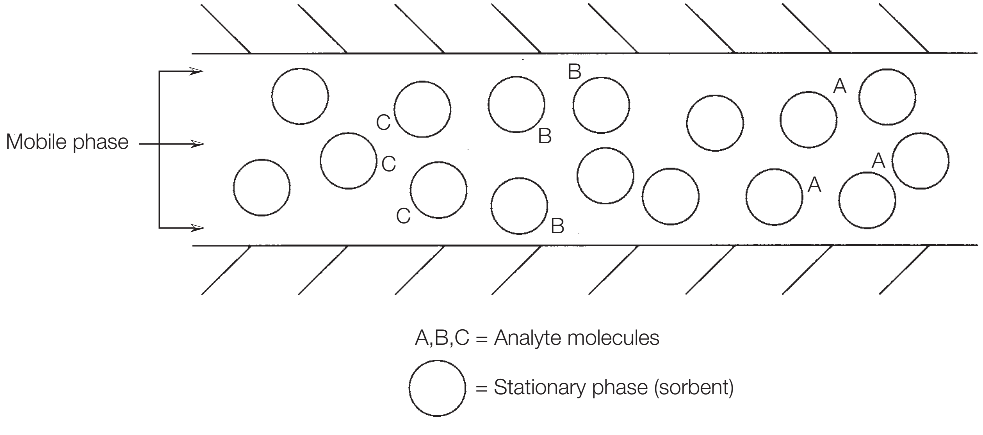

FIGURE 5.1 A representation of the principles of chromatography.

molecules because of different affinities for the stationary phase. The general process of moving a sample mixture through a chromatographic system is called **development**.

The mobile phase can be collected as a function of time at the end of the chromatographic system. The mobile phase, now called the **effluent**, contains the purified analytes. If the chromatographic process has been effective, fractions or "cuts" that are collected at different times will contain the different components of the original sample. *In summary, molecules are separated because they differ in the extent to which they are distributed between the mobile phase and the stationary phase.*

Throughout this chapter and others, biochemical techniques will be designated as **preparative** or **analytical**, or both. A preparative procedure is one that can be applied to the purification of a relatively large amount of a biological material (mg or g). The purpose of such an experiment would be to obtain purified material for further characterization and study. Analytical procedures are used most often to determine the purity of a biological sample; however, they may be used to evaluate any physical, chemical, or biological characteristic of a biomolecule or biological system.

Partition versus Adsorption Chromatography

Chromatographic methods are divided into two types according to how analytes bind to or interact with the stationary phase. **Partition** chromatography is the distribution of an analyte between two liquid phases. This may involve direct extraction using two liquids, or it may use a liquid immobilized on a solid support as in the case of paper, thin-layer, and gas–liquid chromatography. For partition chromatography, the sorbent in Figure 5.1 consists of inert solid particles coated with liquid adsorbent. The distribution of analytes between the two phases is based primarily on solubility differences. The distribution may be quantified by using the **partition coefficient**, K_D (Equation 5.1).

>>

$$K_D = \frac{\text{concentration of analyte in sorbent}}{\text{concentration of analyte in mobile phase}} \qquad \textbf{(Eq. 5.1)}$$

Adsorption chromatography refers to the use of a stationary phase or support, such as an ion-exchange resin, that has a finite number of relatively specific binding sites for analytes. There is not a clear distinction between the processes of partition and adsorption. All chromatographic separations rely, to some extent, on adsorptive processes. However, in some methods (paper, thin-layer, and gas chromatography), these specific adsorptive effects are minimal and the separation is based primarily on nonspecific solubility factors. Adsorption chromatography relies on relatively specific interactions between the analytes and binding sites on the surface of the sorbent. The attractive forces between analyte and support may be ionic, hydrogen bonding, or hydrophobic interactions. Binding of analyte is, of course, reversible.

Because of the different interactions involved in partition and adsorption processes, they may be applied to different separation problems. Partition processes are the most effective for the separation of small molecules, especially those in homologous series. Partition chromatography has been widely used for the separation and identification of amino acids, carbohydrates, and fatty acids. Adsorption techniques, represented by ion-exchange chromatography, are most effective when applied to the separation of macromolecules including proteins and nucleic acids.

In the rest of the chapter, various chromatographic methods will be discussed. You should recognize that no single chromatographic technique relies solely on adsorption or partition effects. Therefore, little emphasis will be placed on classification of the techniques; instead, theoretical and practical aspects will be discussed.

B. PLANAR CHROMATOGRAPHY (PAPER AND THIN-LAYER CHROMATOGRAPHY)

Because of the similarities in the theory and practice of these two procedures, they will be considered together. Both are examples of partition chromatography. In paper chromatography, the cellulose support is extensively hydrated, so distribution of the analyte occurs between the immobilized water (sorbent) and the mobile developing solvent. The initial stationary liquid phase in thin-layer chromatography (TLC) is the solvent used to prepare the thin layer of adsorbent. However, as developing solvent molecules move through the sorbent, polar solvent molecules may bind to the immobilized support and become the sorbent.

Preparation of the Sorbent

The support medium may be a sheet of cellulose or a glass or plastic plate covered with a thin coating of silica gel, alumina, or cellulose. Large sheets of cellulose chromatography paper are available in different porosities. These may be cut to the appropriate size and used without further treatment. The paper should never be handled with bare fingers. Although thin-layer plates can easily be prepared, it is much more convenient to purchase ready-made plates. These are available in a variety of sizes, materials, and thicknesses of stationary support. They are relatively inexpensive and have a more uniform support thickness than handmade plates.

Figure 5.2 outlines the application procedure. The sample to be analyzed is usually dissolved in a volatile solvent. A very small drop of solution is spotted onto the plate with a disposable microcapillary pipet and allowed to dry; then the spotting process is repeated by superimposing more drops on the original spot. The exact amount of sample applied is critical. There must be enough sample so the developed spots can be detected, but overloading will lead to “tailing” and lack of resolution. Finding the proper sample size is a matter of trial and error. It is usually recommended that two or three spots of different concentrations be applied for each sample tested. Spots should be applied along a very faint line drawn with a pencil and ruler. TLC plates should not be heavily scratched or marked. Identifying marks may be made on the top of the chromatogram, where solvent does not reach.

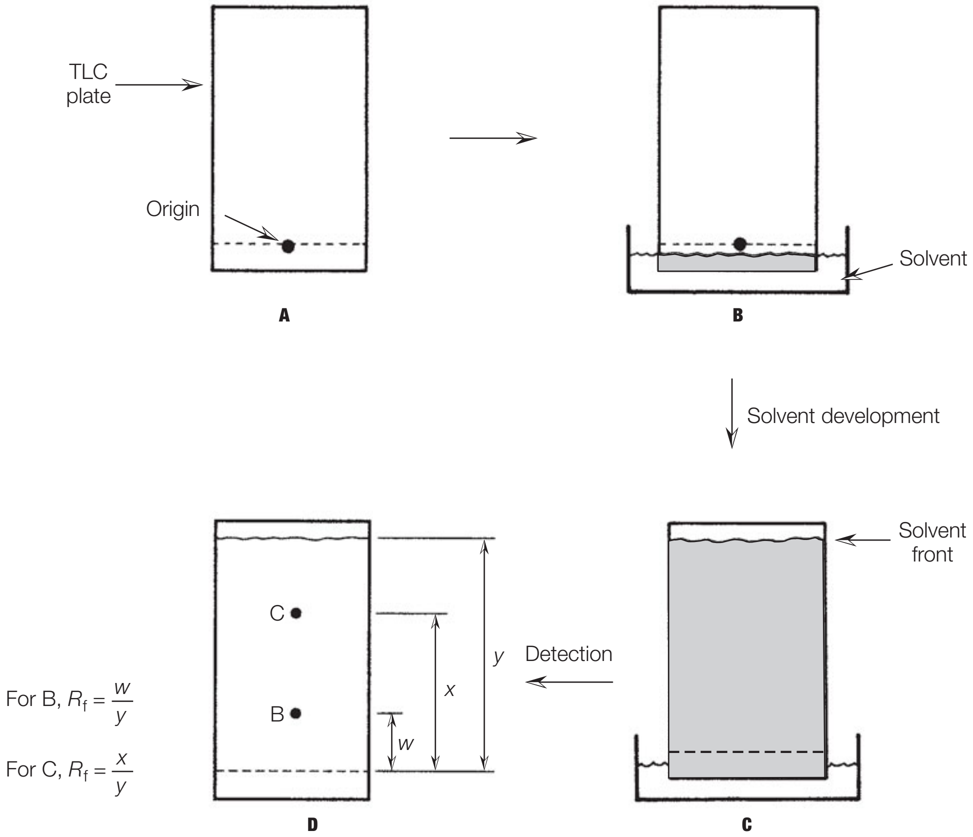

FIGURE 5.2 The procedure of paper and thin-layer chromatography. **A** Application of the sample. **B** Setting plate in solvent chamber. **C** Movement of solvent by capillary action. **D** Detection of separated components and calculation of R_f.

Solvent Development

A wide selection of solvent systems is available in the biochemical literature. If a new solvent system must be developed, a preliminary analysis must be done on the sample with a series of solvents. Solvents can be rapidly screened by developing several small chromatograms (2×6 cm) in small sealed bottles containing the solvents. For the actual analysis, the sample should be run on a larger plate with appropriate standards in a development chamber (Figure 5.3). The chamber must be airtight and saturated with solvent vapors. Filter paper on two sides of the chamber, as shown in Figure 5.3, enhances vaporization of the solvent.

Paper chromatograms may be developed in either of two types of arrangements—ascending or descending solvent flow. Descending solvent flow leads to faster development because of assistance by gravity, and it can offer better resolution for compounds with small R_f values because the solvent can be allowed to run off the paper. R_f values cannot be determined under these conditions, but it is useful for qualitative separations.

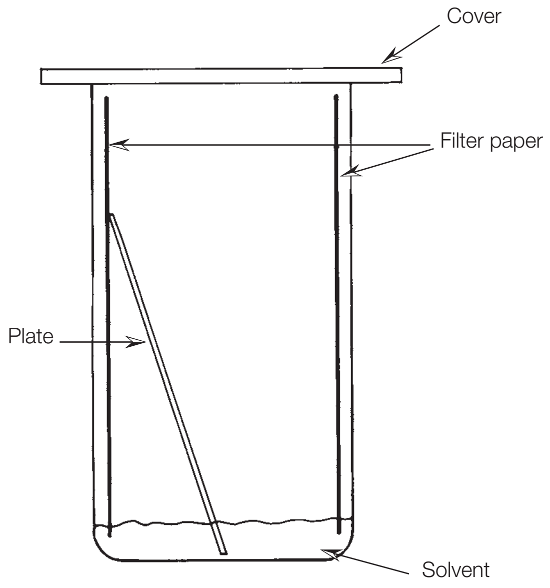

FIGURE 5.3 A typical chamber for paper and thin-layer chromatography.

Two-dimensional chromatography is used for especially difficult separations. The chromatogram is developed in one direction by a solvent system, air dried, turned 90°, and developed in a second solvent system.

Detection and Measurement of Components

Unless the components in the sample are colored, their location on a chromatogram will not be obvious after solvent development. Several methods can be used to locate the spots, including fluorescence, radioactivity, and treatment with chemicals that develop colors. Substances that are highly conjugated may be detected by fluorescence under a UV lamp. Chromatograms may be treated with different types of reagents to develop a color. **Universal reagents** produce a colored spot with any organic compound. When a solvent-developed plate is sprayed with concentrated H_2SO_4 and heated at 100°C for a few minutes, all organic substances appear as black spots. A more convenient universal reagent is I_2. The solvent-developed chromatogram is placed in an enclosed chamber containing a few crystals of I_2. The I_2 vapor reacts with most organic substances on the plate to produce brown spots. The spots are more intense with unsaturated compounds.

Specific reagents react with a particular class of compound. For example, rhodamine B is often used for visualization of lipids, ninhydrin for amino acids, and aniline phthalate for carbohydrates.

The position of each component of a mixture is quantified by calculating the distance traveled by the component relative to the distance traveled by the solvent. This is called **relative mobility** and symbolized by R_f. In Figure 5.2D, the R_f values for components B and C are calculated. The R_f for a substance is a constant for a certain set of experimental conditions. However, it varies with solvent, type of stationary support (paper, alumina, silica gel), temperature, humidity, and other environmental factors. R_f values are always reported along with solvent and temperature.

Applications of Planar Chromatography

Thin-layer chromatography is now more widely used than paper chromatography. In addition to its greater resolving power, TLC is faster and plates are available with several sorbents (cellulose, alumina, silica gel).

Partition chromatography as described in this section may be applied to two major types of problems: (1) identification of unknown samples, and (2) isolation of the components of a mixture. The first application is, by far, the more widely used. Paper chromatography and TLC require only a minute sample size, the analysis is fast and inexpensive, and detection is straightforward. Unknown samples are applied to a plate along with appropriate standards, and the chromatogram is developed as a single experiment. In this way, any changes in experimental conditions (temperature, humidity, etc.) affect standards and unknowns to the same extent. It is then possible to compare the R_f values directly.

Purified substances can be isolated from developed chromatograms; however, only tiny amounts are present. In paper chromatography, the spot may be cut out with scissors and the piece of paper extracted with an appropriate solvent. Isolation of a substance from a TLC plate is accomplished by scraping the solid support from the region of the spot with a knife edge or razor blade and extracting the sorbent with a solvent. "Preparative" thin-layer plates with a thick coating of sorbent (up to 2 mm) are especially useful because they have higher sample capacity.

STUDY EXERCISE 5.1 Planar Chromatography

A mixture containing five amino acids (Ala, Asp, Gly, Phe, Pro) was analyzed using two methods of planar chromatography, paper and cellulose-coated thin layer. The solvent system was *n*-propanol/water (70/30 v/v). Predict the order of the mobility of the amino acids (low R_f to high R_f) on the chromatograms.

Solution: In cellulose planar chromatography, the amino acids interact with two phases, the sorbent (extensively hydrated cellulose, which is very polar) and the mobile phase (*n*-propanol/water, which is less polar than the sorbent). The more polar the amino acid, the stronger it will interact with the hydrated cellulose, thus the slower it will move with solvent during development (lower R_f). The order may be predicted by looking at the polarity of each amino acid side chain and arranging the amino acids in order of decreasing polarity. The correct order of migration (low to high R_f) is: Asp, Gly, Ala, Pro, Phe.

Advanced Planar Chromatography

The applications of planar chromatography listed above require only minimal equipment and supplies that are relatively inexpensive. It is likely that every undergraduate student majoring in biochemistry or molecular biology has completed such an experiment in a science lab. If more complex, sophisticated chromatographic analyses are needed, such as in an academic or biotech research laboratory, then advanced equipment and specialized techniques must be used. Most new advances in planar chromatography have focused on TLC, as paper chromatography is very limited in its applications and flexibility. The major characteristics in TLC that have been improved include standardized methodology, instrumentation, and more effective stationary phases. These changes now make possible the advent of high-performance TLC (HPTLC).

One of the biggest problems with TLC analysis is that experimental conditions are difficult to duplicate. Separation of the components of a mixture is dependent on environmental conditions such as temperature, humidity, and extent of solvent saturation in the chamber. These conditions can be controlled by running the chromatography experiment in a specialized, commercially available enclosure called a **plate development chamber**. This leads to much more reproducible and standardized results.

Improvements are also being made in the development of new stationary phases. The most widely used stationary phase in TLC is silica, which separates molecules on the basis of polarity. (The more polar the component, the stronger it interacts with the very polar silica; hence, it migrates more slowly than a nonpolar component.) As silica separates primarily by polarity, this limits the types of molecules that can be separated. There is now strong interest in the development of reverse stationary phases such as C-18 functionalized silica to separate nonpolar molecules. Another limitation of silica is that it can be used only to separate biomolecules less than 2000-3000 in molecular weight. New and more porous stationary phases are being developed using photopolymerization techniques. These new stationary phases are able to separate protein mixtures containing insulin (MW = 5700), cytochrome *c* (MW = 13,000), lysozyme (MW = 14,600), and myoglobin (MW = 16,900).

C. COLUMN CHROMATOGRAPHY

Adsorption chromatography in biochemical applications usually consists of a solid stationary phase and a liquid mobile phase. The most useful technique is column chromatography, in which the stationary phase is confined to a glass or plastic tube and the mobile phase (a solvent or buffer) is allowed to flow through the solid adsorbent. A small amount of the sample to be analyzed is layered on top of the column. The sample mixture enters the column of adsorbing material and the molecules present are distributed between the mobile phase and the stationary phase. The various components in the sample have different affinities for the two phases and move through the column at different rates. Collection of the liquid phase emerging from the column yields separate fractions containing the individual components in the sample.

Specific terminology is used to describe various aspects of column chromatography. When the actual adsorbing material is made into a column, it is said to be **poured** or **packed**. Application of the sample to the top of the column is **loading** the column. Movement of solvent through the loaded column is called **developing** or **eluting** the column. The **bed volume** is the total volume of solvent

TABLE 5.1 Adsorbents Useful in Biochemical Applications

Adsorbing Materials	Uses
Alumina	Small organics, lipids
Silica gel	Amino acids, lipids, carbohydrates
Fluorisil (magnesium silicate)	Neutral lipids
Calcium phosphate (hydroxyapatite)	Proteins, polynucleotides, nucleic acids
Cellulose	Proteins

and adsorbing material taken up by the column. The volume taken up by the liquid phase in the column is the **void volume**. The **elution volume** is the amount of solvent required to remove a particular analyte from the column. This is analogous to R_f values in planar chromatography.

In adsorption chromatography, solute molecules take part in specific interactions with the stationary phase. Herein lies the great versatility of adsorption chromatography. Many varieties of adsorbing materials are available, so a specific sorbent can be chosen that will effectively separate a mixture. There is still an element of trial and error in the selection of an effective stationary phase. However, experiences of many investigators are recorded in the literature and are of great help in choosing the proper system. Table 5.1 lists the most common stationary phases employed in adsorption column chromatography.

Adsorbing materials come in various forms and sizes. The most suitable forms are dry powders or a slurry form of the material in an aqueous buffer or organic solvent. Alumina, silica gel, and fluorisil do not normally need special pretreatment. The size of particles in an adsorbing material is defined by **mesh size**. This refers to a standard sieve through which the particles can pass. A 100-mesh sieve has 100 small openings per square inch. Adsorbing material with high mesh size (400 and greater) is extremely fine and is most useful for very high resolution chromatography. Table 5.2 lists standard mesh sizes and their most appropriate applications. For most biochemical applications, 100 to 200 mesh size is suitable.

Operation of a Chromatographic Column

A typical column setup is shown in Figure 5.4. The heart of the system is, of course, the column of adsorbent. In general, the longer the column, the better the resolution of components. However, a compromise must be made because flow

TABLE 5.2 Mesh Sizes of Adsorbents and Typical Applications

Mesh Sizes	Applications
20–50	Crude preparative work, very high flow rate
50–100	Preparative applications, high flow rate
100–200	Analytical separations, medium flow rate
200–400	High-resolution analytical separations, low flow rate

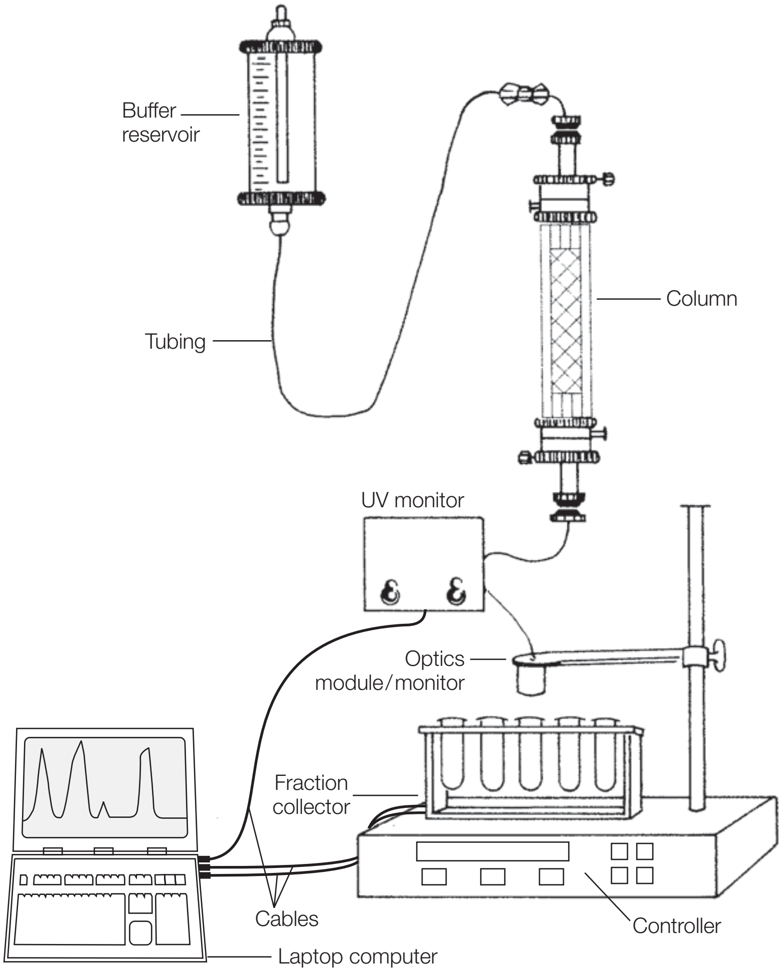
Buffer reservoir
Column
Tubing
UV monitor
Optics module/monitor
Fraction collector
Cables
Controller
Laptop computer

FIGURE 5.4 Setup for the operation of a chromatography column.

rate decreases with increasing column length. The actual size of a column depends on the nature of the adsorbing material and the amount of chemical sample to be separated. For preparative purposes, column heights of 20 to 50 cm are usually sufficient to achieve acceptable resolution. Column inside diameters may vary from 0.5 to 5 cm.

Packing the Column

Once the adsorbing material and column size have been selected, the column is poured. If the tube does not have a fritted disc in the bottom, a small piece of glass wool or cotton should be used to support the column. Most columns are packed by pouring a slurry of the sorbent into the tube and allowing it to settle by gravity into a tight bed. The slurry is prepared with the solvent or buffer that will be used as the initial developing solvent. Pouring of the slurry must be continuous to avoid formation of sorbent layers. Excess solvent is eluted from the bottom of the column while the sorbent is settling. The column must never run dry. Additional slurry is added until the column bed reaches the desired height. The top of the settled adsorbent is then covered with a small circle of filter paper or glass wool to protect the surface while the column is loaded with sample or the eluting solvent is changed.

Sometimes it is necessary to pack a column under pressure (5 to 10 psi). This leads to a tightly packed bed that yields more reproducible results, especially with gradient elution (see the following subsections).

Loading the Column

The sample to be analyzed by chromatography should be applied to the top of the column in a concentrated form. If the sample is solid, it is dissolved in a minimum amount of solvent; if already in solution, it may be concentrated by ultrafiltration as described in Chapter 3, p. 74. After the sample is loaded onto the column with a graduated or disposable pipet, it is allowed to percolate into the adsorbent. A few milliliters of solvent are then carefully added to wash the sample into the column material. The column is then filled with eluting solvent.

Eluting the Column

The chromatography column is developed by continuous flow of a solvent. Maintaining the appropriate flow rate is important for effective separation. If the flow rate is set too high, there is not sufficient time for complete equilibration of the analytes with the two phases. Too low a flow rate allows diffusion of analytes, which leads to poor resolution and broad elution peaks. It is difficult to give guidelines for the proper flow rate of a column, but, in general, a column should be adjusted to a rate slightly less than "free flow." Sometimes it is necessary to find the proper flow rate by trial and error. One problem encountered during column development is a changing flow rate. As the solvent height above the column bed is reduced, there is less of a "pressure head" on the column, so the flow rate decreases. This can be avoided by storing the developing solvent in a large reservoir and allowing it to enter the column at the same rate as it is emerging from the column (see Figure 5.4).

Adsorption columns are eluted in one of three ways. All components may be eluted by a single solvent or buffer. This is referred to as **continual elution**. In contrast, **stepwise elution** refers to an incremental change of solvent to aid development. The column is first eluted with a volume of one solvent and then with a second solvent. This may continue with as many solvents or solvent mixtures as desired. In general, the first solvent should be the least polar of any used in the analysis, and each additional solvent should be of greater polarity or ionic strength. Finally, adsorption columns may be developed by **gradient elution** brought about by a gradual change in solvent composition. The composition of the eluting solvent can be changed by the continuous mixing of two different solvents to gradually change the ratio of the two solvents. Alternatively, the concentration of a component in the solvent can be gradually increased. This is most often done by addition of a salt (KCl, NaCl, etc.). Devices are commercially available to prepare predetermined, reproducible gradients.

Collecting the Eluent

The separated components emerging from the column in the eluent are usually collected as discrete fractions. This may be done manually by collecting specified volumes of eluent in Erlenmeyer flasks or test tubes. Alternatively, if many fractions are to be collected, a mechanical fraction collector is convenient and even essential. An automatic fraction collector (see Figure 5.4) directs the eluent into a single tube until a predetermined volume has been collected or until a preselected time period has elapsed; then the collector advances another tube for collection. Specified volumes are collected by a drop counter activated by a photocell, or a timer can be set to collect a fraction over a specific period.

Detection of Eluting Components

The completion of a chromatographic experiment calls for a means to detect the presence of analytes in the collected fractions. The detection method used will depend on the nature of the analytes. Smaller molecules such as lipids, amino acids, and carbohydrates can be detected by spotting fractions on a thin-layer plate or a piece of filter paper and treating them with a chemical reagent that produces a color. The same reagents that are used to visualize spots on a thin-layer or paper chromatogram are useful for this. Proteins and nucleic acids are conveniently detected by spectroscopic absorption measurements at 280 and 260 nm, respectively. Enzymes can be detected by measurements of catalytic activity associated with each fraction. Research-grade chromatographic systems are equipped with detectors that continuously monitor some physical property of the eluent and display the separation results on a computer screen (see Figure 5.4). The newest advance in detectors is the diode array (see Chapter 7, pp. 210–211). Most often the eluent is directed through a flow cell where absorbance or fluorescence characteristics can be measured. The detector is connected to a recorder or computer for a permanent record of spectroscopic changes. When the location of the various analytes is determined, adjacent fractions containing identical components are pooled and stored for later use.

D. ION-EXCHANGE CHROMATOGRAPHY

Ion-exchange chromatography is a form of adsorption chromatography in which ionic analytes display reversible electrostatic interactions with a charged stationary phase. The chromatographic setup is identical to that described in the last section and Figure 5.4. The column is packed with a stationary phase consisting of a synthetic resin that is tagged with ionic functional groups. The steps involved in ion-exchange chromatography are outlined in Figure 5.5. In stage 1, the insoluble resin material (positively charged) in the column is surrounded by buffer counterions. Loading of the column in stage 2 brings analytes of different charge into the ion-exchange medium. Solutes entering the column may be negatively charged, positively charged, or neutral under the experimental conditions. Analytes that have a charge opposite to that of the resin bind tightly but reversibly to the stationary phase (stage 3). The strength of binding depends on the size of the charge and the charge density (amount of charge per unit volume of molecule) of the analyte. The greater the charge or the charge density, the

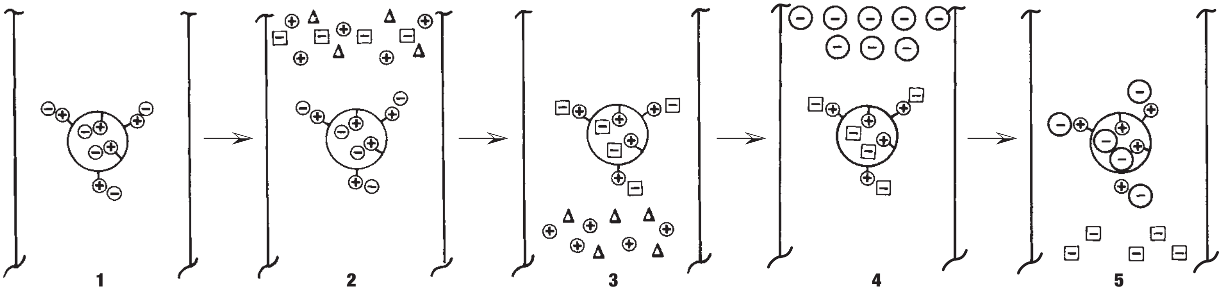

FIGURE 5.5 Illustration of the principles of ion-exchange chromatography. See text for explanation.

stronger the interaction. Neutral analytes (Δ) or those with a charge identical to that of the resin show little or no affinity for the stationary phase and move with the eluting buffer. The bound analytes can be released by eluting the column with a buffer of increased ionic strength or pH (stage 4). An increase in buffer ionic strength releases bound analytes by displacement. Increasing the buffer pH decreases the strength of the interaction by reducing the charge on the analyte or on the resin (stage 5).

The following sections will focus on the properties of ion-exchange resins, selection of experimental conditions, and applications of ion-exchange chromatography.

Ion-Exchange Resins

Ion exchangers are made up of two parts—an insoluble, three-dimensional matrix and chemically bonded charged groups within and on the surface of the matrix. The resins are prepared from a variety of materials, including polystyrene, acrylic resins, polysaccharides (dextrans), agarose, and celluloses. An ion exchanger is classified as **cationic** or **anionic** depending on whether it exchanges cations or anions. A resin that has negatively charged functional groups exchanges positive ions and is a **cation exchanger**. Each type of exchanger is also classified as **strong** or **weak** according to the ionizing strength of the functional group. An exchanger with a quaternary amino group is, therefore, a **strongly basic anion exchanger**, whereas primary or secondary aromatic or aliphatic amino groups would lead to a **weakly basic anion exchanger**. A **strongly acidic cation exchanger** contains the sulfonic acid group. Table 5.3 lists the common ion exchangers according to each of these classifications.

The ability of an ion exchanger to adsorb counterions is defined quantitatively by **capacity**. The **total capacity** of an ion exchanger is the quantity of charged and potentially charged groups per unit weight of dry exchanger. It is usually expressed as milliequivalents of ionizable groups per milligram of dry weight, and it can be experimentally determined by titration. The capacity of an ion exchanger is a function of the porosity of the resin. The resin matrix contains covalent cross-linking that creates a "molecular sieve." Ionized functional groups within the matrix are not readily accessible to large molecules that cannot fit into the pores. Only surface charges would be available to these

TABLE 5.3 Ion-Exchange Resins

Name	Functional Group	Matrix	Class
Anion Exchangers			
AG 1	Tetramethylammonium	Polystyrene	Strong
AG 3	Tertiary amine	Polystyrene	Weak
DEAE-Sephacel	Diethylaminoethyl	Sephacel	Weak
PEI-cellulose	Polyethyleneimine	Cellulose	Weak
DEAE-Sephadex	Diethylaminoethyl	Dextran	Weak
QAE-Sephadex	Diethyl-(2-hydroxyl-propyl)-aminoethyl	Dextran	Strong
DEAE-Sepharose	Diethylaminoethyl	Agarose	Weak
Cation Exchangers			
AG 50	Sulfonic acid	Polystyrene	Strong
Bio-Rex 70	Carboxylic acid	Acrylic	Weak
CM-Sephacel	Carboxymethyl	Sephacel	Weak
P-Cellulose	Phosphate	Cellulose	Intermediate
CM-Sephadex	Carboxymethyl	Dextran	Weak
SP-Sephadex	Sulfopropyl	Dextran	Strong
CM-Sepharose	Carboxymethyl	Agarose	Weak
SP-Sepharose	Sulfonic acid	Agarose	Strong

molecules for exchange. The purely synthetic resins (polystyrene and acrylic) have cross-linking ranging from 2 to 16%, with 8% being the best for general purposes.

With so many different experimental options and resin properties to consider, it is difficult to select the proper conditions for a particular separation. The next section will outline the choices and offer guidelines for proper experimental design.

Selection of the Ion Exchanger

Before a proper choice of ion exchanger can be made, the nature of the molecules to be separated must be considered. For relatively small, stable molecules (amino acids, lipids, nucleotides, carbohydrates, pigments, etc.) the synthetic resins based on polystyrene are most effective. They have relatively high capacity for small molecules because the extensive cross-linking still allows access to the interior of the resin beads. For separations of peptides, proteins, nucleic acids, polysaccharides, and other large biomolecules, one must consider the use of fibrous cellulosic ion exchangers and low-percent cross-linked dextran or acrylic exchangers. The immobilized functional groups in these resins are readily available for exchange even to larger molecules.

The choice of ion exchanger has now been narrowed considerably. The next decision is whether to use a cationic or anionic exchanger. If the analyte has only one type of charged group, the choice is simple. A molecule that has a positive charge will bind to a cationic exchanger and vice versa. However, many biomolecules have more than one type of ionizing group and may have both negatively and positively charged groups (they are amphoteric). The net charge on such molecules depends on pH. At the isoelectric point, the substance has no net charge and would not bind to any type of ion exchanger.

In principle, amphoteric molecules should bind to both anionic and cationic exchangers. However, when one is dealing with large biomolecules, the pH **range of stability** must also be evaluated. The range of stability refers to the pH range in which the biomolecule is not denatured. Figure 5.6 shows how the net charge of a hypothetical protein changes as a function of pH. Below the isoelectric point (pH_I), the molecule has a net positive charge and would be bound to a cation exchanger. Above the isoelectric point, the net charge is negative, and the protein would bind to an anion exchanger. Superimposed on this graph is the pH range of stability for the hypothetical protein. Because it is stable in the range of pH 7.0–10.0, the ion exchanger of choice is an anionic exchanger. In most cases, the isoelectric point of the protein is not known. The type of ion exchanger must be chosen by trial and error as follows. Small samples of the protein mixture in buffer are equilibrated for 10 to 15 minutes in separate test tubes, one with each type of ion exchanger. The tubes are then centrifuged or let stand to sediment the ion exchanger. Check each supernatant for the presence of the desired analyte (A_{260} for nucleic acids, A_{280} for proteins, catalytic activity for enzymes, etc.). If a supernatant has a relatively low level of added protein, that ion exchanger would be suitable for use. This simple test can also be extended to

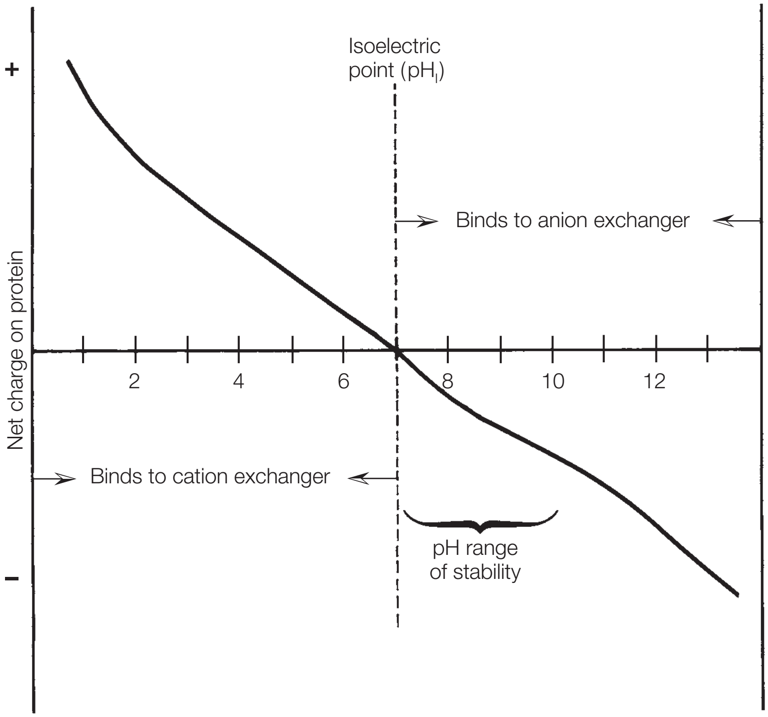

FIGURE 5.6 The effect of pH on the net charge of a protein.

find conditions for elution of the desired macromolecule from the ion exchanger. The ion exchanger charged with the macromolecule is treated with buffers of increasing ionic strength or changing pH. The supernatant after each treatment is analyzed as before for release of the macromolecule.

Choice of Buffer

This decision includes not just the buffer substance, but also the pH and the ionic strength. Buffer ions will, of course, interact with ion-exchange resins. Buffer ions with a charge opposite to that on the ion exchanger compete with analyte for binding sites and greatly reduce the capacity of the column. Cationic buffers should be used with anionic exchangers; anionic buffers should be used with cationic exchangers.

The pH chosen for the buffer depends first of all on the range of stability of the macromolecule to be separated (see Figure 5.6). Second, the buffer pH should be chosen so that the desired macromolecule will bind to the ion exchanger. In addition, the ionic strength should be relatively low to avoid “damping” of the interaction between analyte and ion exchanger. Buffer concentrations in the range 0.05 to 0.1 *M* are recommended.

Preparation of the Ion Exchanger

The commercial suppliers of ion exchangers provide detailed instructions for the preparation of the adsorbents. Failure to pretreat ion exchangers will greatly reduce the capacity and resolution of a column. Most new ion-exchange resins are commercially available in slurry form and are ready to use with a minimum number of pretreatment steps.

Sometimes pretreatment steps do not remove the small particles that are present in most ion-exchange materials. If left in suspension, these particles, called fines, result in decreased resolution and low column flow rates. The fines are removed from an exchanger by suspending the swollen adsorbent in a large volume of water in a graduated cylinder and allowing at least 90% of the exchanger to settle. The cloudy supernatant containing the fines is decanted. This process is repeated until the supernatant is completely clear. The number of washings necessary to remove most fines is variable, but for a typical ion exchanger 8 to 10 times is probably sufficient.

Using the Ion-Exchange Resin

Ion exchangers are most commonly used in a column form. The column method discussed earlier in this chapter can be directly applied to ion-exchange chromatography.

An alternative method of ion exchange is **batch separation**. This involves mixing and stirring equilibrated exchanger directly with the analyte mixture to be separated. After an equilibration time of approximately 1 hour, the slurry is filtered and washed with buffer. The ion exchanger can be chosen so that the desired analyte is adsorbed onto the exchanger or remains unbound in solution. If the latter is the case, the desired material is in the filtrate. If the desired analyte is bound to the exchanger, it can be removed by suspending the exchanger in a buffer of greater ionic strength or different pH. Batch processes have some advantages over column methods. They are rapid, and the problems of packing, channeling, and dry columns are avoided.

Another development in ion-exchange column chromatography allows the separation of proteins according to their isoelectric points. This technique, **chromatofocusing**, involves the formation of a pH gradient on an ion-exchange column. If a buffer of a specified pH is passed through an ion-exchange column that was equilibrated at a second pH, a pH gradient is formed on the column. Proteins bound to the ion exchanger are eluted in the order of their isoelectric points. In addition, protein band concentration (focusing) takes place during elution. Chromatofocusing is similar to isoelectric focusing, introduced in Chapter 6, in which a column pH gradient is produced by an electric current.

Storage of Resins

Most ion exchangers in the dry form are stable for many years. Aqueous slurried ion exchangers are still useful after several months. One major storage problem with a wet exchanger is microbial growth. This is especially true for the cellulose and dextran exchangers. If it is necessary to store pretreated exchangers, an antimicrobial agent must be added to the slurry. Sodium azide (0.02%) is suitable for cation exchangers, and phenylmercuric salts (0.001%) are effective for anion exchangers. Since these preservative reagents are toxic, they must be used with caution.

STUDY EXERCISE 5.2 Cation-Exchange Chromatography

A mixture of amino acids (Lys, Ala, Asp) is subjected to cation exchange chromatography at pH 3.0. Predict the order of elution of the three amino acids and explain your answer. **Hint:** Draw the structures of the three amino acids as they would exist at pH 3. Study the structures and determine how each would interact with the negatively charged ion-exchange resin. Determine the relative strength of binding for each amino acid.

STUDY EXERCISE 5.3 Anion-Exchange Chromatography

A mixture of three proteins was subjected to ion-exchange chromatography using CM-cellulose as the stationary phase. Predict the order of elution of the proteins, assuming that the mixture was applied at low ionic strength and eluted with buffers of increasing ionic strength. The proteins in the mixture are listed below with isoelectric pH (pH_I).

Pepsinogen, 1.0
Cytochrome *c*, 10.6
Myoglobin, 6.8

Hint: The column separates proteins on the basis of net charge. The more negatively charged a protein, the more weakly it will bind to an anion exchange resin like CM-cellulose, and the faster it will elute. At any pH, the protein pepsinogen will have the most negative net charge of any of the three proteins.

STUDY EXERCISE 5.4 Ion-Exchange Chromatography

What kind of ion-exchange resin would be most effective for purifying RNA molecules, an anion exchanger or cation exchanger?

E. GEL-EXCLUSION CHROMATOGRAPHY

The chromatographic methods discussed up to this point allow the separation of molecules according to polarity and charge. The method of **gel-exclusion chromatography** (also called gel filtration, molecular-sieve chromatography, or gel-permeation chromatography) exploits the physical property of molecular size to achieve separation. The molecules of nature range in molecular weight from less than 100 to as large as several million. It should be obvious that a technique capable of separating molecules of molecular weight 10,000 from those of 100,000 would be very popular among research biochemists. Gel-filtration chromatography has been of major importance in the purification of thousands of proteins, nucleic acids, enzymes, polysaccharides, and other biomolecules. In addition, the technique may be applied to molecular weight determination and quantitative analysis of molecular interactions. In this section, the theory and practice of gel filtration will be introduced and applied to several biochemical problems.

Theory of Gel Filtration

The operation of a gel filtration column is illustrated in Figure 5.7. The stationary phase consists of inert particles that contain small pores of a controlled size. Microscopic examination of a particle reveals an interior resembling a sponge. A solution containing analytes of various molecular sizes is allowed to pass through the column under the influence of continuous solvent flow. Analytes larger than the pores cannot enter the interior of the gel beads, so they are limited to the space between the beads. The volume of the column accessible to very large molecules is, therefore, greatly reduced. As a result, they are not slowed in their progress through the column and elute rapidly in a single zone. Small molecules capable of diffusing in and out of the beads have a much larger volume available to them. Therefore, they are delayed in their journey through the column bed. Molecules of intermediate size migrate through the

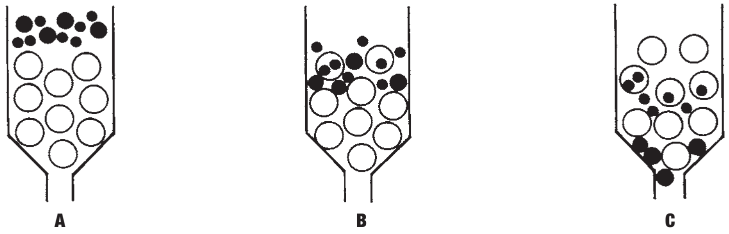

FIGURE 5.7 Separation of molecules by gel filtration. **A** Application of sample containing large and small molecules. **B** Large molecules cannot enter gel matrix, so they move rapidly through the column. **C** Elution of the large molecules first and then smaller molecules.

column at a rate somewhere between those for large and small molecules. Therefore, the order of elution of the various analytes is directly related to their molecular dimensions.

Physical Characterization of Gel Chromatography

Several physical properties must be introduced to define the performance of a gel and solute behavior. Some important properties are:

1. ***Exclusion Limit*** This is defined as the molecular mass of the smallest molecule that cannot diffuse into the inner volume of the gel matrix. All molecules above this limit elute rapidly in a single zone. The exclusion limit of a typical gel, Sephadex G-50, is 30,000 daltons. All analytes having a molecular size greater than this value would pass directly through the column bed without entering the gel pores.
2. ***Fractionation Range*** Sephadex G-50 has a fractionation range of 1500 to 30,000 daltons. Analytes within this range would be separated in a somewhat linear fashion.
3. ***Water Regain and Bed Volume*** Gel chromatography media are often supplied in dehydrated form and must be swollen in a solvent, usually water, before use. The weight of water taken up by 1 g of dry gel is known as the water regain. For G-50, this value is 5.0 ± 0.3 g. This value does not include the water surrounding the gel particles, so it cannot be used as an estimate of the final volume of a packed gel column. Most commercial suppliers of gel materials provide, in addition to water regain, a bed volume value. This is the final volume taken up by 1 g of dry gel when swollen in water. For G-50, bed volume is 9 to 11 mL/g dry gel.
4. ***Gel Particle Shape and Size*** Ideally, gel particles should be spherical to provide a uniform bed with a high density of pores. Particle size is defined either by mesh size or bead diameter (μm). Both the degree of resolution afforded by a column and the flow rate depend on particle size. Larger particle sizes (50 to 100 mesh, 100 to 300 μm) offer high flow rates but poor chromatographic separation. The opposite is true for very small particle sizes ("superfine," 400 mesh, 10 to μm). The most useful particle size, which represents a compromise between resolution and flow rate, is 100 to 200 mesh (50 to 150 μm).
5. ***Void Volume*** This is the total space surrounding the gel particles in a packed column. This value is determined by measuring the volume of solvent required to elute a solute that is completely excluded from the gel matrix. Most columns can be calibrated for void volume with a dye, blue dextran, which has an average molecular mass of 2,000,000 daltons.
6. ***Elution Volume*** This is the volume of eluting buffer necessary to remove a particular analyte from a packed column.

Chemical Properties of Gels

Four basic types of gels are available: **dextran**, **polyacrylamide**, **agarose**, and **combined polyacrylamide-dextran**. The first gels to be developed were those based on a natural polysaccharide, dextran. These are supplied by Pharmacia under the trade name Sephadex. Table 5.4 gives the physical properties of the

TABLE 5.4 Properties of Gel-Filtration Media

Name	Fractionation Range for Proteins (daltons)	Water Regain (mL/g dry gel)	Bed Volume (mL/g dry gel)
Dextran (Sephadex)[1]			
G-10	0–700	1.0 ± 0.1	2–3
G-15	0–1500	1.5 ± 0.2	2.5–3.5
G-25	1000–5000	2.5 ± 0.2	4–6
G-50	1500–30,000	5.0 ± 0.3	9–11
G-75	3000–80,000	7.5 ± 0.5	12–15
G-100	4000–150,000	10 ± 1.0	15–20
G-150	5000–300,000	15 ± 1.5	20–30
G-200	5000–600,000	20 ± 2.0	30–40
Polyacrylamide (Bio-Gels)[2]			
P-2	100–1800	1.5	3.0
P-4	800–4000	2.4	4.8
P-6	1000–6000	3.7	7.4
P-10	1500–20,000	4.5	9.0
P-30	2500–40,000	5.7	11.4
P-60	3000–60,000	7.2	14.4
P-100	5000–100,000	7.5	15.0
P-150	15,000–150,000	9.2	18.4
P-200	30,000–200,000	14.7	29.4
P-300	60,000–400,000	18.0	36.0
Dextran-polyacrylamide (Sephacryl)[1]			
S-100 HR	1000–100,000	—	—
S-200 HR	5000–250,000	—	—
S-300 HR	10,000–1,500,000	—	—
S-400 HR	20,000–8,000,000	—	—
Agarose			
Sepharose[1] 6B	10,000–4,000,000	—	—
Sepharose 4B	60,000–20,000,000	—	—
Sepharose 2B	70,000–40,000,000	—	—
Superose[1] 12 HR	1000–300,000	—	—
Superose 6 HR	5000–5,000,000	—	—
Bio-Gel[2] A-0.5	10,000–500,000	—	—
Bio-Gel A-1.5	10,000–1,500,000	—	—
Bio-Gel A-5	10,000–5,000,000	—	—
Bio-Gel A-15	40,000–15,000,000	—	—

TABLE 5.4 *Continued*

Name	Fractionation Range for Proteins (daltons)	Water Regain (mL/g dry gel)	Bed Volume (mL/g dry gel)
Bio-Gel A-50	100,000–50,000,000	—	—
Bio-Gel A-150	1,000,000–150,000,000	—	—
Vinyl (Fractogel TSK)[3]			
HW-40	100–10,000	—	—
HW-55	1000–700,000	—	—
HW-65	50,000–5,000,000	—	—
HW-75	500,000–50,000,000	—	—

[1] *Amersham – Pharmacia – LKB – Hoefer.*
[2] *Bio-Rad Laboratories.*
[3] *Pierce Biotechnology.*

various sizes of Sephadex. The number given each gel refers to the water regain multiplied by 10. Sephadex is available in various particle sizes labeled coarse, medium, fine, and superfine. Dextran-based gels cannot be manufactured with an exclusion limit greater than 600,000 daltons because the small extent of cross-linking is not sufficient to prevent collapse of the particles. If the dextran is cross-linked with N,N'-methylenebisacrylamide, gels for use in higher fractionation ranges are possible. Table 5.4 lists these gels, called Sephacryl.

Polyacrylamide gels are produced by the copolymerization of acrylamide and the cross-linking agent N,N'-methylenebisacrylamide. These are supplied by Bio-Rad Laboratories (Bio-Gel P). The Bio-Gel media are available in 10 sizes with exclusion limits ranging from 1800 to 400,000 daltons. Table 5.4 lists the acrylamide gels and their physical properties.

The agarose gels have the advantage of having very high exclusion limits. Agarose, the neutral polysaccharide component of agar, is composed of alternating galactose and anhydrogalactose units. The gel structure is stabilized by hydrogen bonds rather than by covalent cross-linking. Agarose gels, supplied by Bio-Rad Laboratories (Bio-Gel A) and by Pharmacia (Sepharose and Superose), are listed in Table 5.4.

The combined polyacrylamide-agarose gels are commercially available under the trade name Ultragel. These consist of cross-linked polyacrylamide with agarose trapped within the gel network. The polyacrylamide gel allows a high degree of separation and the agarose maintains gel rigidity, so high flow rates may be used.

Selecting a Gel

The selection of the proper gel is a critical stage in successful gel chromatography. Most gel chromatographic experiments can be classified as either **group separations** or **fractionations**. Group separations involve dividing the components of a sample into two groups, a fraction of relatively low-molecular-weight analytes and a fraction of relatively high-molecular-weight analytes. Specific examples of this are desalting a protein solution or removing small contaminating molecules from protein or nucleic acid extracts. For group separations, a gel should be chosen that allows complete exclusion of the high-molecular-weight molecules in the void volume. Sephadex G-25, Bio-Gel P-6, and Sephacryl S-100HR are recommended for most group separations. The particle size recommended is 100 to 200 mesh or 50 to 150 μm diameter.

Gel fractionation involves separation of groups of molecules of similar molecular weights in a multicomponent mixture. In this case, the gel should be chosen so that the fractionation range includes the molecular weights of the analytes. If the mixture contains macromolecules up to 120,000 in molecular weight, then Bio-Gel P-150, Sephacryl S-200HR, or Sephadex G-150 would be most appropriate. If P-100, G-100, or Sephacryl S-100HR were used, some of the higher-molecular-weight proteins in the sample would elute in the void volume. On the other hand, if P-200, P-300, or G-200 were used, there would be a decrease in both resolution and flow rate. If the molecular weight range of the mixture is unknown, empirical selection is necessary. The recommended gel grade for most fractionations is 100–200 or 200–400 mesh (20–80 μm or 10–40 μm). The finest grade that allows a suitable flow rate should be selected. For very critical separations, superfine grades offer the best resolution, but with very low flow rates.

Gel Preparation and Storage

The dextran and acrylamide gel products are sometimes supplied in dehydrated form and must be allowed to swell in water before use. The swelling time required differs for each gel, but the extremes are 3 to 4 hours at 20°C for highly cross-linked gels and up to 72 hours at 20°C for P-300 or G-200. The swelling time can be shortened if a boiling-water bath is used. Agarose gels and combined polyacrylamide-agarose gels are supplied in a hydrated state, so there is no need for swelling.

Before a gel slurry is packed into the column, it should be defined and deaerated. Defining is necessary to remove very fine particles, which would reduce flow rates. To define, pour the gel slurry into a graduated cylinder and add water equivalent to two times the gel volume. Invert the cylinder several times and allow the gel to settle. After 90 to 95% of the gel has settled, decant the supernatant, add water, and repeat the settling process. Two or three defining operations are usually sufficient to remove most small particles.

Deaerating (removing dissolved gases) should be done on the gel slurry and all eluting buffers. Gel particles that have not been deaerated tend to float and form bubbles in the column bed. Dissolved gases are removed by placing the gel slurry in a side-arm vacuum flask and applying a vacuum from a water aspirator. The degassing process is complete when no more small air bubbles are released from the gel (usually 1 to 2 hours).

Antimicrobial agents must be added to stored, hydrated gels. One of the best agents is sodium azide (0.02%).

Operation of a Gel Column

The procedure for gel-column chromatography is very similar to the general description given earlier. The same precautions must be considered in packing, loading, and eluting the column. A brief outline of important considerations follows.

Column Size

For fractionation purposes, it is usually not necessary to use columns greater than 100 cm in length. The ratio of bed length to width should be between 25 and 100. For group separations, columns less than 50 cm long are sufficient, and appropriate ratios of bed length to width are between 5 and 10.

Eluting Buffer

There are fewer restrictions on buffer choice in gel chromatography than in ion-exchange chromatography. Dextran and polyacrylamide gels are stable in the pH range 1 to 10, whereas agarose gels are limited to pH 4 to 10. Since there is such a wide range of stability of the gels, the buffer pH should be chosen on the basis of the range of stability of the macromolecules to be separated.

Sample Volume

The sample volume is a critical factor in planning a gel chromatography experiment. If too much sample is applied to a column, resolution is decreased; if the sample size is too small, the analytes are greatly diluted. For group separations, a sample volume of 10 to 25% of the column total volume is suitable. The sample volume for fractionation procedures should be between 1 and 5% of the total volume. Column total volume is determined by measuring the volume of water in the glass column that is equivalent to the height of the packed bed.

Column Flow Rate

The flow rate of a gel column depends on many factors, including length of column and type and size of the gel. It is generally safe to elute a gel column at a rate slightly less than free flow. A high flow rate reduces sample diffusion or zone broadening, but may not allow complete equilibration of analyte molecules with the gel matrix.

A specific flow rate cannot be recommended, since each type of gel requires a different range. The average flow rate given in literature references for small-pore-size gels is 8 to 12 $\mathrm{mL/cm^2}$ of cross-sectional bed area per hour (15 to 25 mL/hr). For large-pore-size gels, a value of 2 to 5 $\mathrm{mL/cm^2}$ of cross-sectional bed area per hour (5 to 10 mL/hr) is average.

Eluent can be made to flow through a column by either of two methods, gravity or pump elution. Gravity elution is most often used because no special equipment is required. It is quite acceptable for developing a column used for group separations and fractionations when small-pore-sized gels are used. However, if the flow rate must be maintained at a constant value throughout an experiment or if large-pore gels are used, pump elution is recommended.

One variation of gel chromatography is ascending eluent flow. Some investigators report more reproducible results, better resolution, and a more constant flow rate if the eluting buffer is pumped backward through the gel. This type of experiment requires special equipment, including a specialized column, and a peristaltic pump.

Applications of Gel-Exclusion Chromatography

Several experimental applications of gel chromatography have already been mentioned, but more detail will be given here.

Desalting

Inorganic salts, organic solvents, and other small molecules are used extensively for the purification of macromolecules. Gel chromatography provides an inexpensive, simple, and rapid method for removal of these small molecules. One especially attractive method for desalting very small samples (0.1 mL or less) of proteins or nucleic acid solutions is to use spin columns. These are prepacked columns of polyacrylamide exclusion gels. Spin columns are used in a similar fashion to microfiltration centrifuge tubes (Chapter 3, p. 75). The sample is placed on top of the gel column and spun in a centrifuge. Large molecules are eluted from the column and collected in a reservoir. The small molecules to be removed remain in the gel.

Purification of Biomolecules

This is probably the most popular use of gel chromatography. Because of a gel's ability to fractionate molecules on the basis of size, gel filtration complements other purification techniques that separate molecules on the basis of polarity and charge.

Estimation of Molecular Weight

The elution volume for a particular analyte is proportional to its molecular size. This indicates that it is possible to estimate the molecular weight of a molecule on the basis of its elution characteristics on a gel column. An elution curve for several standard proteins separated on Sephadex G-100 is shown in Figure 5.8. This curve, a plot of protein concentration A_{280} vs. volume collected, is representative of data obtained from a gel filtration experiment. The elution volume, V_e, for each protein can be estimated as shown in the figure. A plot of log molecular

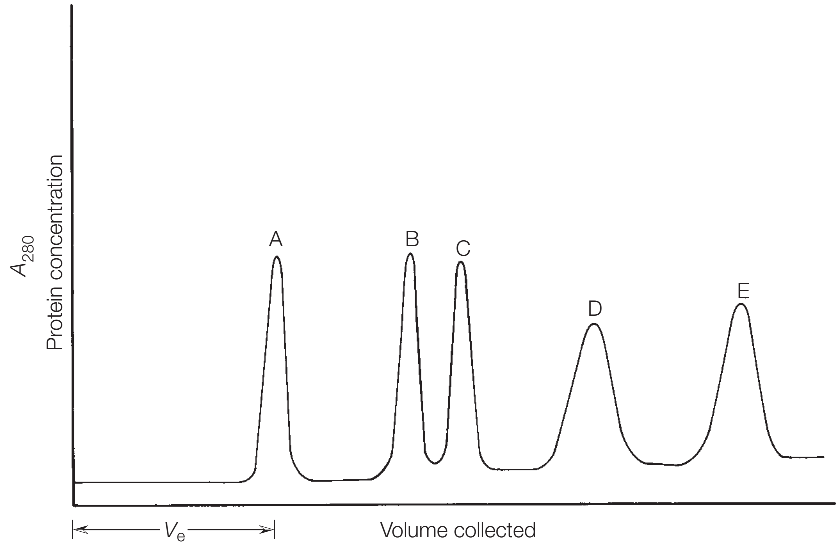

FIGURE 5.8 Elution curve for a mixture of several proteins using gel-filtration chromatography. **A** = hemoglobin; **B** = egg albumin; **C** = chymotrypsinogen; **D** = myoglobin; **E** = cytochrome *c*.

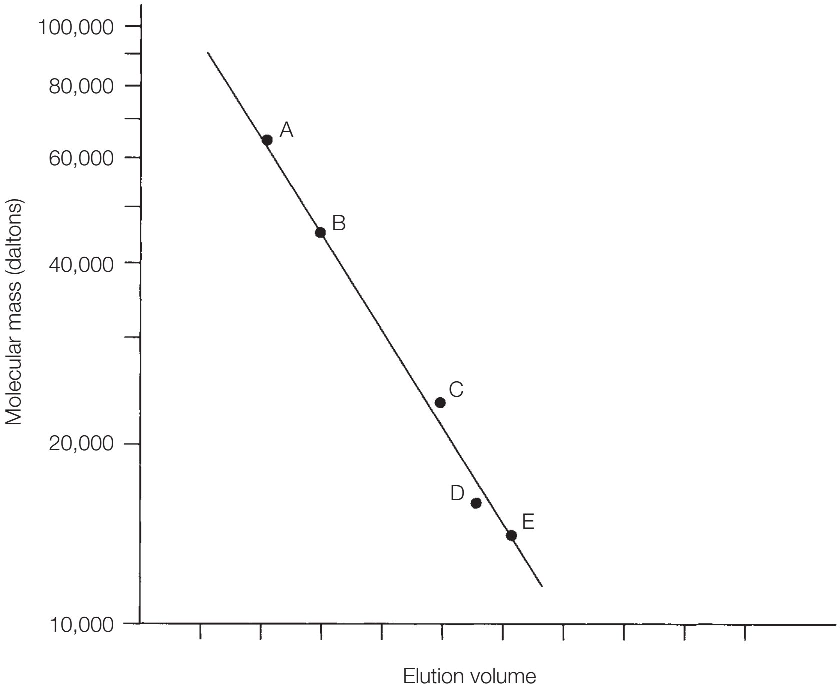

FIGURE 5.9 A plot of log molecular mass vs. elution volume for the proteins A, B, C, D, and E in Figure 5.8.

mass vs. elution volume for the proteins is shown in Figure 5.9. Note that the linear portion of the curve in Figure 5.9 covers the molecular mass range of 10,000 to 100,000 daltons. Molecules below 10,000 daltons are not eluted in an elution volume proportional to size. Since they readily diffuse into the gel particles, they are retarded. Molecules larger than 100,000 daltons are all excluded from the gel in the void volume. A solution of the unknown protein is chromatographed through the calibrated column under conditions identical to those for the standards, and the elution volume is measured. The unknown molecular size is then read directly from the graph. This method of molecular weight estimation is widely used because it is simple, inexpensive, and fast. It can be used with highly purified or impure samples. There are, of course, some limitations to consider. The gel must be chosen so that the molecular weight of the unknown is within the linear section of the curve. The method assumes that only steric and partition effects influence the elution of the standards and unknown. If a protein interacts with the gel by adsorptive or ionic processes, the estimate of molecular weight will be lower than the true value. The assumption is also made that the unknown molecules have a general spherical shape, not an elongated or rod shape. Although this is a convenient and rapid method for molecular weight determination, the final number is only an estimate.

Gel Chromatography in Organic Solvents

The gels discussed so far in this section are hydrophilic, and the inner matrix retains its integrity only in aqueous solvents. Because there is a need for gel chromatography of nonhydrophilic molecules, gels have been produced that can be used with organic solvents. The trade names of some of these products are Sepharose CL and Sephadex LH, Bio Beads S, and Styragel. Organic solvents that are of value in gel chromatography are ethanol, acetone, dimethylsulfoxide, dimethylformamide, tetrahydrofuran, chlorinated hydrocarbons, and acetonitrile.

STUDY EXERCISE 5.5 Gel-Exclusion Chromatography

Your first experiment in biochemistry lab is to analyze a mixture with the use of gel filtration. You are instructed to run a Sephadex G-50 column on the mixture containing the components listed below. All three components are colored, so you are able to observe the separation. The components are dissolved in water, and the column is eluted with water.

Blue dextran, a blue dye (MW = about 2,000,000)
Cytochrome *c*, a red protein (MW = 12,400)
Flavin adenine dinucleotide, a yellow coenzyme (FAD, MW = 830)

Predict the results of your experiment.

Solution: According to Table 5.4, the fractionation range for G-50 is 1500–30,000. The Blue dextran is above the fractionation range, so it will elute very rapidly in the void volume. The FAD is below the fractionation range, so it will take its time and elute much later than Blue dextran and cyt *c*. The red-colored cyt *c*, within the fractionation range, will elute somewhere between the Blue dextran and FAD.

STUDY EXERCISE 5.6 Separation of Proteins

Your biochemistry lab instructor gives you a mixture of three proteins and asks you to separate the proteins on a Sephadex G-100 gel. The three proteins are serum albumin, myoglobin, and chymotrypsinogen. The column is eluted with 0.1 *M* Tris buffer, pH 7.0. Predict the order of elution of the proteins from the Sephadex column. Use Appendix V to find the molecular weights of the proteins.

F. HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY (HPLC)

The previous discussions on the theory and practice of the various chromatographic methods should convince you of the tremendous influence chromatography has had on our biochemical understanding. It is tempting to make comparisons about the relative importance of the methods, but doing so would be unfair because each serves a specific purpose. There will, for example, always be a need for fast, inexpensive, and qualitative analyses as afforded by planar chromatography. Traditional column chromatography will probably always be preferred in large-scale protein purification.

However, during the past three decades, an analytical method has been developed that now surpasses the traditional liquid chromatographic techniques in importance for analytical separations. This technique, **high-performance liquid chromatography (HPLC)**, is ideally suited for the separation and identification of amino acids, carbohydrates, lipids, nucleic acids, proteins, pigments, steroids, pharmaceuticals, and many other biologically active molecules.

The future promise of HPLC is indicated by its classification as "modern liquid chromatography" when compared to other forms of column-liquid chromatography, now referred to as "classical" or "traditional." Compared to the classical forms of liquid chromatography (paper, TLC, column), HPLC has several advantages:

1. Resolution and speed of analysis far exceed the classical methods.
2. HPLC columns can be reused without repacking or regeneration.
3. Reproducibility is greatly improved because the parameters affecting the efficiency of the separation can be closely controlled.
4. Instrument operation and data analysis are easily automated.
5. HPLC is adaptable to very small sample sizes or large-scale, preparative procedures.

The advantages of HPLC are the result of two major advances: (1) the development of stationary supports with very small particle sizes and large surface areas, and (2) the improvement of elution rates by applying high pressure to the solvent flow.

The great versatility of HPLC is evidenced by the fact that all chromatographic modes, including partition, adsorption, ion exchange, chromatofocusing, and gel exclusion, are possible. In a sense, HPLC can be considered as automated liquid chromatography. The theory of each of these chromatographic modes has been discussed and needs no modification for application to HPLC. However, there are unique theoretical and practical characteristics of HPLC that should be introduced.

The **retention time** of an analyte in HPLC (t_R) is defined as the time necessary for maximum elution of the particular analyte. This is analogous to retention time measurements in GC. **Retention volume** (V_R) of an analyte is the solvent volume required to elute the analyte and is defined by Equation 5.2, where F is the flow rate of the solvent.

>> $$V_R = Ft_R \qquad \textbf{(Eq. 5.2)}$$

In all forms of chromatography, a measure of column efficiency is **resolution**, R. Resolution indicates how well analytes are separated; it is defined by Equation 5.3, where t_R and t'_R are the retention times of two analytes and w and w' are the base peak widths of the same two analytes.

>> $$R = 2\frac{t_R - t'_R}{w + w'} \qquad \textbf{(Eq. 5.3)}$$

Instrumentation

The increased resolution achieved in HPLC compared to classical column chromatography is primarily the result of adsorbents of very small particle sizes (less than 20 μm) and large surface areas. The smallest gel beads used in gel-exclusion chromatography are "superfine" grade with diameters of 20 to 50 μm. Recall that the smaller the particle size, the lower the flow rate; therefore, it is not feasible to use very small gel beads in liquid column chromatography because low flow rates lead to solute diffusion and the time necessary for completion of an analysis would be impractical. In HPLC, increased flow rates are obtained by applying a pressure differential across the column. A combination of high pressure and adsorbents of small particle size leads to the high resolving power and short analysis times characteristic of HPLC.

A schematic diagram of a typical high-pressure liquid chromatograph is shown in Figure 5.10. The basic components are a solvent reservoir, high-pressure pump, packed column, detector, and recorder. A computer is used to control the process and to collect and analyze data.

Solvent Reservoir

The solvent chamber should have a capacity of at least 500 mL for analytical applications, but larger reservoirs are required for preparative work. In order to avoid bubbles in the column and detector, the solvent must be degassed. Several

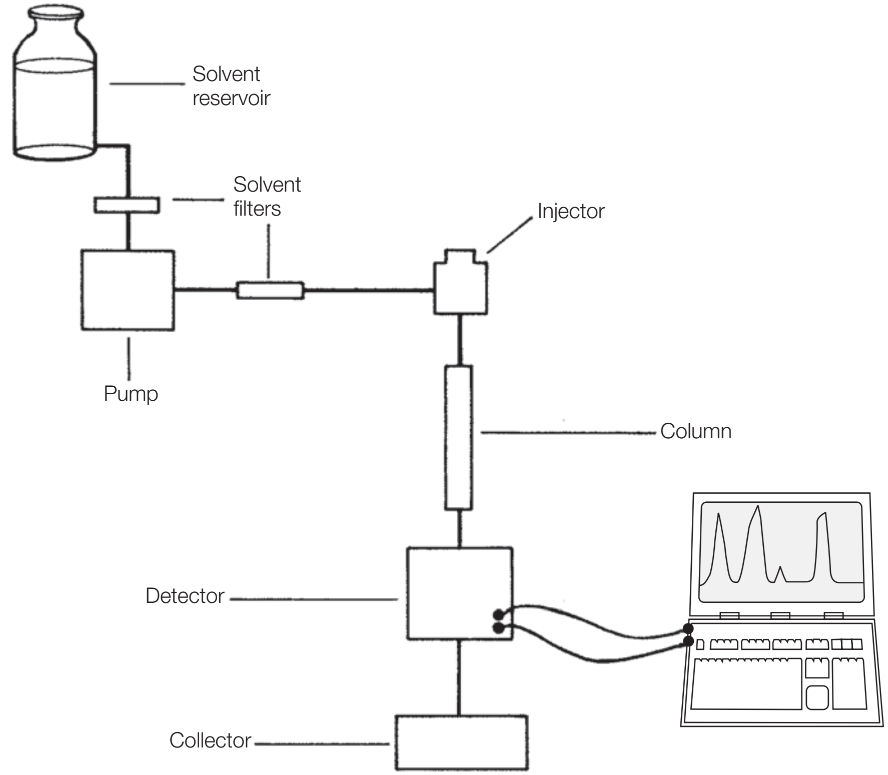

FIGURE 5.10 A schematic diagram of a high-performance liquid chromatograph.

methods may be used to remove unwanted gases, including refluxing, filtration through a vacuum filter, ultrasonic vibration, and purging with an inert gas. The solvent should also be filtered to remove particulate matter that would be drawn into the pump and column.

Pumping Systems

The purpose of the pump is to provide a constant, reproducible flow of solvent through the column. Two types of pumps are available—constant pressure and constant volume. Typical requirements for a pump are:

1. It must be capable of pressure outputs of at least 500 psi and preferably up to 5000 psi.
2. It should have a controlled, reproducible flow delivery of about 1 mL/min for analytical applications and up to 100 mL/min for preparative applications.
3. It should yield pulse-free solvent flow.
4. It should have a small holdup volume.

Although neither type of pump meets all these criteria, constant-volume pumps maintain a more accurate flow rate and allow for a more precise analysis.

Injection Port

A sample must be introduced onto the column in an efficient and reproducible manner. One of the most popular injectors is the syringe injector. The sample, in a microliter syringe, is injected through a neoprene/Teflon septum. This type of injection can be used at pressures up to 3000 psi.

Columns

HPLC columns are prepared from stainless steel or glass-Teflon tubing. Typical column inside diameters are 2.1, 3.2, or 4.5 mm for analytical separations and up to 30 mm for preparative applications. The length of the column can range from 5 to 100 cm, but 10- to 20-cm columns are common.

Detector

Liquid chromatographs are equipped with a means to continuously monitor the column effluent and recognize the presence of analyte. Only small sample sizes are used with most HPLC columns, so a detector must have high sensitivity. The type of detector that has the most universal application is the **differential refractometer**. This device continuously monitors the refractive index difference between the mobile phase (pure solvent) and the mobile phase containing sample (column effluent). The sensitivity of this detector is on the order of 0.1 μg, which, compared to other detectors, is only moderately sensitive. The major advantage of the refractometer detector is its versatility; its main limitation is that there must be at least 10^{-7} refractive index units between the mobile phase and sample.

The most widely used HPLC detectors are the **photometric detectors**. These detectors measure the extent of absorption of ultraviolet or visible radiation by a sample. Since few compounds are colored, visible detectors are of limited value. Ultraviolet detectors are the most widely used in HPLC. The typical UV detector functions by focusing radiation from a low-pressure mercury lamp on a flow cell that contains column effluent. The mercury lamp provides a primary radiation at 254 nm. The use of filters or other lamps provides radiation at 220, 280, 313, 334, and 365 nm. Many compounds absorb strongly in this wavelength range, and sensitivities on the order of 1 ng are possible. Most biochemicals are detected, including proteins, nucleic acids, pigments, vitamins, some steroids, and aromatic amino acids. Aliphatic amino acids, carbohydrates, lipids, and other biochemicals that do not absorb UV can be detected by chemical derivatization with UV-absorbing functional groups. UV detectors have many positive characteristics, including high sensitivity, small sample volumes, linearity over wide concentration ranges, nondestructiveness to sample, and suitability for gradient elution.

A third type of detector that has only limited use is the **fluorescence detector**. This type of detector is extremely sensitive: its use is limited to samples containing trace quantities of biological materials. Its response is not linear over a wide range of concentrations, but it may be up to 100 times more sensitive than the UV detector.

Collection of Eluent

All of the detectors described here are nondestructive to the samples, so column effluent can be collected for further chemical and physical analysis.

Analysis of HPLC Data

Most HPLC instruments are on-line, with an integrator and a computer for data handling. For quantitative analysis of HPLC data, operating parameters such as rate of solvent flow must be controlled. In modern instruments, the whole system (including the pump, injector, detector, and data system) is under the control of a computer.

Figure 5.11 illustrates the separation by HPLC of several phenylhydantoin derivatives of amino acids.

Stationary Phases in HPLC

The adsorbents in HPLC are typically small-diameter, porous materials. Two types of stationary phases are available. **Porous layer beads** (Figure 5.12A) have an inert solid core with a thin porous outer shell of silica, alumina, or ion-exchange resin. The average diameter of the beads ranges from 20 to 45 μm. They are especially useful for analytical applications, but, because of their short pores, their capacities are too low for preparative applications.

Microporous particles are available in two sizes: 20 to 40 μm diameter with longer pores and 5 to 10 μm with short pores (see Figure 5.12B). These are now more widely used than the porous layer beads because they offer greater resolution and faster separations with lower pressures. The microporous beads are prepared from alumina, silica, ion-exchanger resins, and chemically bonded phases (see next section).

The HPLC can function in several chromatographic modes. Each type of chromatography will be discussed in the following subsections, with information about stationary phases.

Column: Microsorb Cyano, 5 μm, 4.6 mm ID × 25 cm L
Mobile phase: 15.5% THF and 17.1% acetonitrile in 6 m*M* phosphate, pH 3.2
Flow: 1 mL/min
Temperature: 35°C
Detection: UV, 254 nm

Peak identification

1. Cys-A
2. Asn
3. Gln
4. Ser
5. Thr
6. Asp
7. Gly
8. His
9. Glu
10. Ala
11. Arg
12. Cys
13. Tyr
14. Pro
15. Val
16. Met
17. Ile
18. Phe
19. Leu
20. nor-Leu
21. Trp
22. Lys

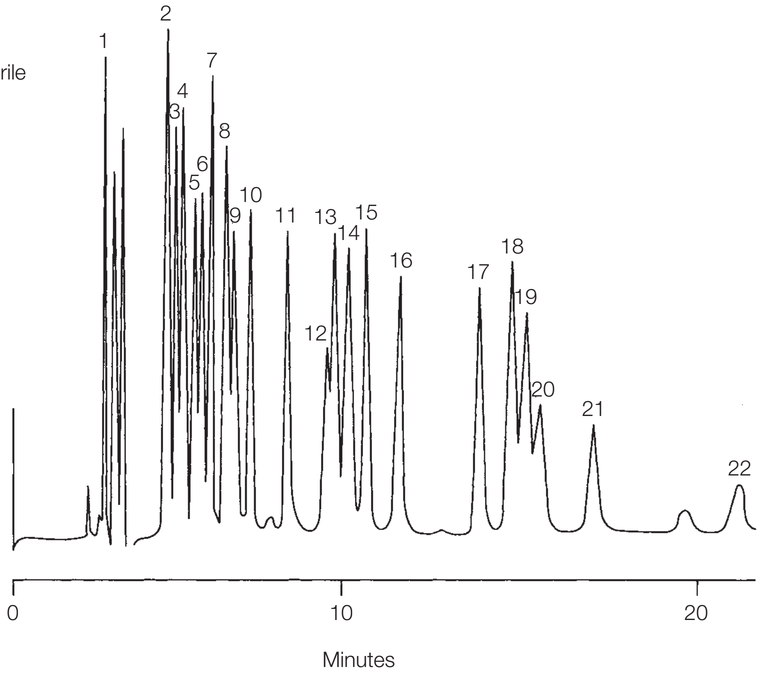

FIGURE 5.11 The separation of several amino acid phenylhydantoins by HPLC. *Courtesy of Rainin Instrument Co., Woburn, MA;* www.rainin.com.

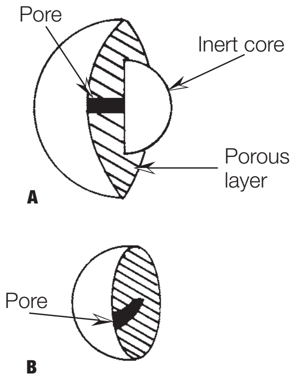

FIGURE 5.12 Adsorbents used in HPLC. **A** Porous layer with short pores. **B** Microporous particles with longer pores.

Liquid-Solid (Adsorption) Chromatography

HPLC in the adsorption mode can be carried out with silica or alumina porous-layer-bead columns. Small glass beads are often used for the inert core. Some of the more widely used packings are μ Porasil (Waters Associates), BioSilA (Bio-Rad Laboratories), LiChrosorb Si-100 Partisil, Vydac, ALOX 60D (several suppliers), and Supelcosil (Supelco).

In high-pressure adsorption chromatography, analytes adsorb with different affinities to binding sites in the solid stationary phase. Separation of analytes in a sample mixture occurs because polar molecules adsorb more strongly than nonpolar molecules. Therefore, the various components in a sample are eluted with different retention times from the column. This form of HPLC is usually called **normal phase** (a polar stationary phase and a nonpolar mobile phase).

Liquid-Liquid (Partition) Chromatography

In the early days of HPLC (1970–78), solid supports were coated with a liquid stationary phase. Columns with these packings had short lifetimes and a gradual decrease in resolution because there was continuous loss of the liquid stationary phase with use of the column.

This problem was remedied by the discovery of methods for chemically bonding the active stationary phase to the inert support. Most chemically bonded stationary phases are produced by covalent modification of the surface silica. Three modification processes are shown in Equations 5.4–5.6.

Silicate esters

>> $$—\mathrm{Si}—\mathrm{OH} + \mathrm{ROH} \longrightarrow —\mathrm{Si}—\mathrm{O}—\mathrm{R} \qquad \textbf{(Eq. 5.4)}$$

Silica − carbon

>> $$—\mathrm{Si}—\mathrm{Cl} \;+\; \begin{array}{c}\mathrm{RMgBr}\\ \text{or}\\ \mathrm{RLi}\end{array} \longrightarrow —\mathrm{Si}—\mathrm{R} \qquad \textbf{(Eq. 5.5)}$$

Siloxanes

>> $$—\mathrm{Si}—\mathrm{OH} \;+\; \begin{array}{c}\mathrm{ClSiR_3}\\ \text{or}\\ \mathrm{ROSiR_3}\end{array} \longrightarrow —\mathrm{Si}—\mathrm{O}—\mathrm{SiR_3} \qquad \textbf{(Eq. 5.6)}$$

The major advantage of a bonded stationary phase is stability. Since it is chemically bonded, there is very little loss of stationary phase with column use. The siloxanes are the most widely used silica supports. Functional groups that can be attached as siloxanes are alkylnitriles ($—\mathrm{Si}—\mathrm{CH_2CH_2}—\mathrm{CN}$), phenyl ($—\mathrm{Si}—\mathrm{C_6H_5}$), alkylamines ($—\mathrm{Si(CH_2)}_n\mathrm{NH_2})—$, and alkyl side chains ($—\mathrm{Si}—\mathrm{C_8H_{17}}$; $—\mathrm{Si}—\mathrm{C_{18}H_{37}}$).

The use of nonpolar chemically bonded stationary phases with a polar mobile phase is referred to as **reverse-phase HPLC**. This technique separates sample components according to hydrophobicity. It is widely used for the separation of all types of biomolecules, including peptides, nucleotides, carbohydrates, and derivatives of amino acids. Typical solvent systems are water-methanol, water-acetonitrile, and water-tetrahydrofuran mixtures. Figure 5.13 shows the results of protein separation on a silica-based reverse-phase column.

Conditions

column:	Hi-Pore RP-304 column, 250 × 4.6 mm
Sample:	Protein standard
Buffers:	A. H_2O, 0.1% B B. 95% Acetonitrile, 5% H_2O, 0.1% TFA
Gradient:	25%–100% B in 30 minutes
Flow rate:	1.5 mL/min
Temperature:	Ambient
Detection:	UV @ 280 nm; conductivity

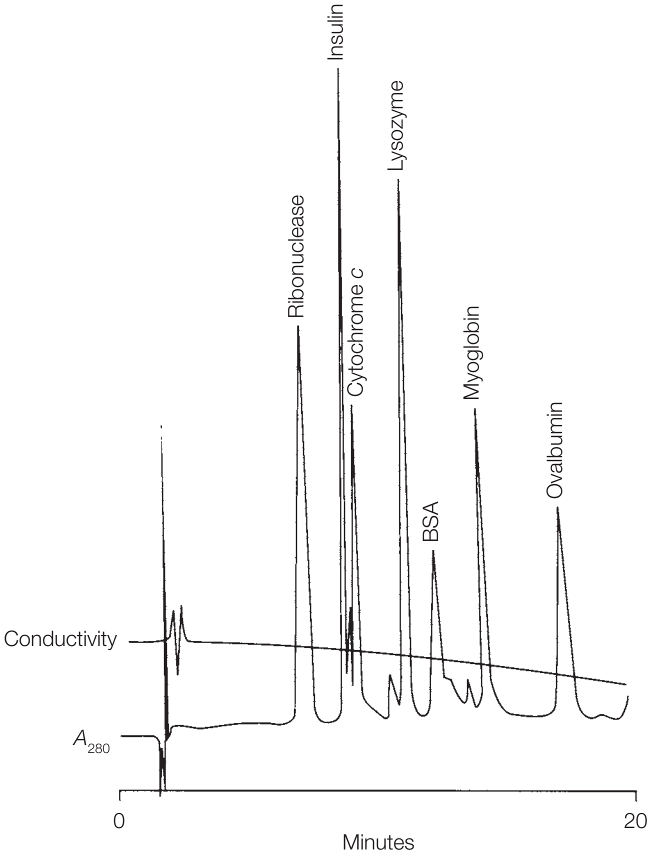

FIGURE 5.13 Reverse-phase chromatography of a mixture of standard proteins. *Separation courtesy of Bio-Rad Laboratories, Hercules, CA;* www.bio-rad.com.

Ion-Exchange Chromatography

Ion-exchange HPLC uses column packings with charged functional groups. Structures of typical ion exchangers are shown in Figure 5.14. They are prepared by chemically bonding the ionic groups to the support via silicon atoms or by using polystyrene-divinylbenzene resins. These stationary phases may be used for the separation of proteins, peptides, and other charged biomolecules.

Gel-Exclusion Chromatography

The combination of HPLC and gel-exclusion chromatography is used extensively for the separation of large biomolecules, especially proteins and nucleic acids.

The exclusion gels discussed in Section E are not appropriate for HPLC use because they are soft and, except for small-pore beads (G-25 and less), collapse under high-pressure conditions. Semirigid gels based on cross-linked styrene-divinylbenzene, polyacrylamide, and vinyl-acetate copolymer are available with various fractionation ranges useful for the separation of molecules up to 10,000,000 daltons.

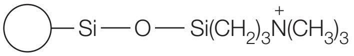

$-Si-O-Si(CH_2)_3\overset{+}{N}(CH_3)_3$

Anion exchanger

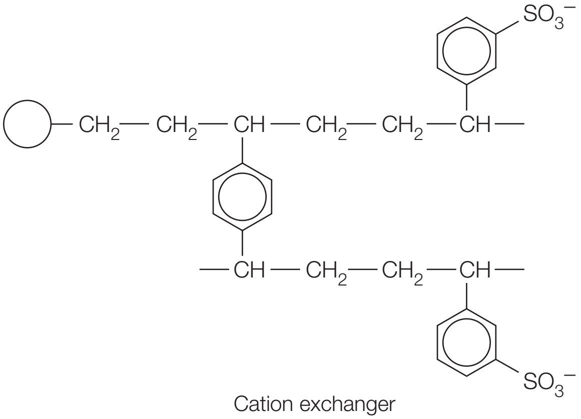

Cation exchanger

FIGURE 5.14 Ion exchangers used in HPLC.

Rigid packings for HPLC gel exclusion are prepared from porous glass or silica. They have multiple advantages over the semirigid gels, including several fractionation ranges, ease of packing, and compatibility with water and organic solvents.

Chiral Chromatography

HPLC has been extremely effective in separating and analyzing a broad range of biological molecules, including amino acids, proteins, lipids, nucleic acids, and carbohydrates, as described in the previous subsections. One class of biologically important compounds that has shown resistance to separation are the enantiomeric forms of biomolecules that exist because of the presence of a chiral center or stereocenter. For most molecules, a chiral center is a carbon atom that is surrounded by four different groups. Compounds that have a stereocenter exist in two molecular forms **(enantiomers)** that have the same physical and chemical properties. They differ only in the way they interact with plane-polarized light—one enantiomer rotates the light to the left (L), and one enantiomer rotates light to the right (D). We now know that enantiomers also differ in their biological actions or physiological effects. This has become increasingly important in medical treatment because many drugs are chiral and the enantiomers differ in how they interact with receptors. One enantiomeric form of a drug may interact to cause the desired effect, and one may lead to no effect or even to toxicity. For example, the drug thalidomide was widely used in Europe during the late 1950s and early 1960s as a sedative/tranquilizer, especially for pregnant women. The original drug form was the racemic mixture containing two thalidomide enantiomers. Thousands of women taking the drug later delivered malformed babies. We now know that only one enantiomer (D, *R*) causes the desired sedative effect. The other enantiomer (L, *S*) is teratogenic.

It has become essential to develop techniques for the separation and analysis of enantiomers by chromatography. Some current drugs whose active ingredient is a single enantiomer include Lipitor (Atorvastatin), Zocor (Simvastatin), Nexium (Esomeprazole), Plavix (Clopidogrel), Advair (Fluticasone), and Zoloft (Sertraline).

A logical approach to the separation of enantiomers is to use a chiral stationary phase, because it is known that chiral compounds are able to interact selectively, favoring one enantiomer over the other. Because of the different spatial arrangements around the stereocenters of the enantiomers, they interact differently with a chiral surface. The most effective chiral stationary phases contain proteins that are, of course, composed of amino acids, each of which has a stereocenter (except gly). Many protein sorbents have been tested, and the most successful results have come from alpha1-acid glycoproteins (AGP), human serum albumin (HSA), and cellobiohydrolase (CBH). For example, the proteins AGP and HSA on HPLC columns have been used to separate the enantiomers of the anti-inflammatory drug ibuprofen (see Figure 5.15).

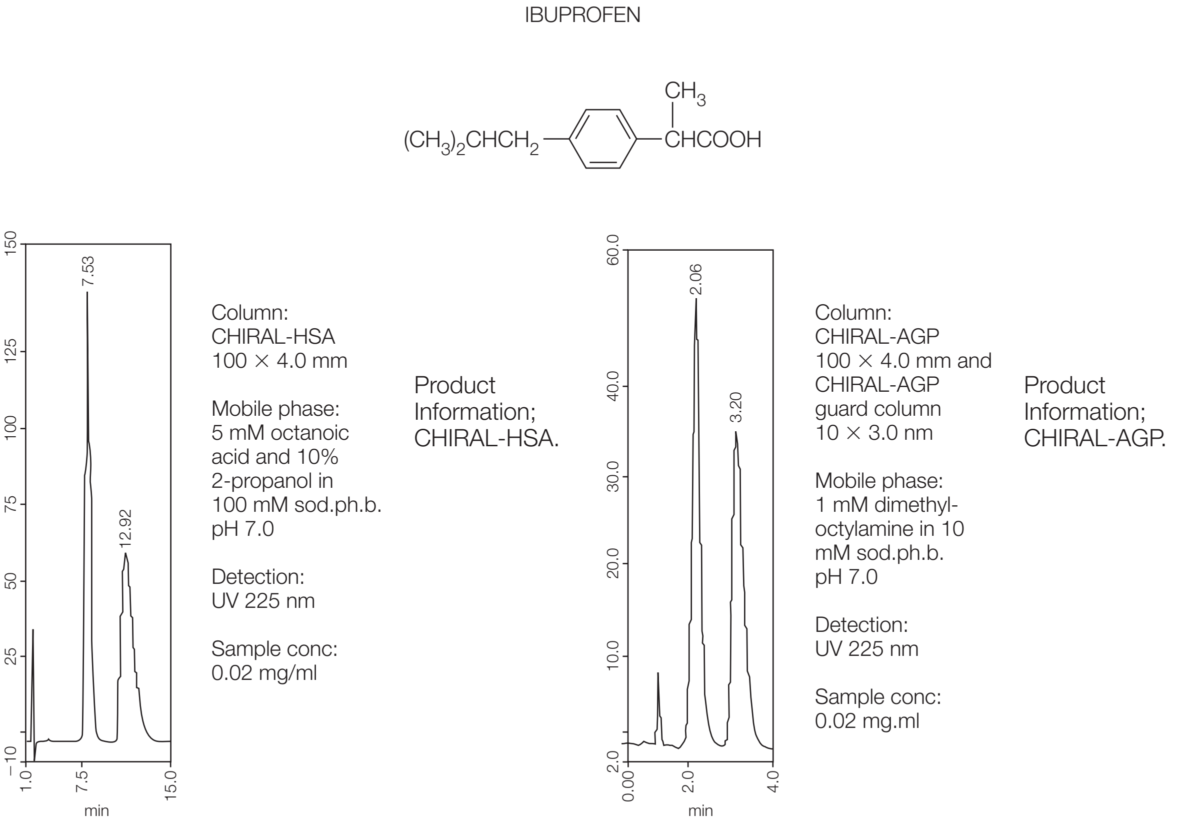

Column:
CHIRAL-HSA
100 × 4.0 mm

Mobile phase:
5 mM octanoic
acid and 10%
2-propanol in
100 mM sod.ph.b.
pH 7.0

Detection:
UV 225 nm

Sample conc:
0.02 mg/ml

Product
Information;
CHIRAL-HSA.

Column:
CHIRAL-AGP
100 × 4.0 mm and
CHIRAL-AGP
guard column
10 × 3.0 nm

Mobile phase:
1 mM dimethyl-
octylamine in 10
mM sod.ph.b.
pH 7.0

Detection:
UV 225 nm

Sample conc:
0.02 mg.ml

Product
Information;
CHIRAL-AGP.

FIGURE 5.15 Separation of the enantiomers of the painkiller ibuprofen on two different chiral supports. sod.ph.b = sodium phosphate buffer. *Courtesy of Chiral Technologies, Inc.* www.chiraltech.com.

The Mobile Phase

Selection of a column packing that is appropriate for a given analysis does not ensure a successful HPLC separation. A suitable solvent system must also be chosen. Several critical solvent properties will be considered here.

> ***Purity*** Very-high-purity solvents with no particulate matter are required. Many laboratory workers do not purchase expensive prepurified solvents, but, rather, purify lesser grade solvents by microfiltration through a Millipore system or distillation in glass.
>
> ***Reactivity*** The mobile phase must not react with the analytical sample or column packing. This does not present a major limitation since many relatively unreactive hydrocarbons, alkyl halides, and alcohols are suitable.
>
> ***Detector Compatibility*** A solvent must be carefully chosen to avoid interference with the detector. Most UV detectors monitor the column effluent at 254 nm. Any UV-absorbing solvent, such as benzene or olefins, would be unacceptable because of high background. Since refractometer detectors monitor the difference in refractive index between solvent and column effluent, a greater difference leads to greater ability to detect the solute.

Sample Preparation and Selection of HPLC Operating Conditions

During the initial stages of biochemical sample preparation, the sample is often quite crude; it may contain hundreds of components in addition to the desired biomolecules. Most samples must be pretreated before optimum HPLC results can be expected. The following procedures may be needed in order to convert a crude sample into a clean one: desalting, removal of anions and cations, removal of metal ions, concentration of the desired macromolecules, removal of detergent, and particulate removal. Sample preparation techniques used to achieve these results are gel-exclusion chromatography, ion-exchange chromatography, microfiltration, and metal-affinity chromatography. These procedures may be completed by commercially available prefilters or precolumns.

Each type of HPLC instrument has its own characteristics and operating directions. It is not feasible to describe those here. However, it is appropriate to outline the general approach taken when an HPLC analysis is desired. The following items must be considered:

1. Chemical nature and proper preparation of the sample.
2. Selection of type of chromatography (partition, adsorption, ion-exchange, gel-exclusion).
3. Choice of solvent system and mode of elution.
4. Selection of column packing.
5. Choice of equipment (type of detector).

FPLC—A Modification of HPLC

In 1982, Pharmacia introduced an innovative chromatographic method that provides a link between classical column chromatography and HPLC. This technique, called **fast protein liquid chromatography (FPLC)**, uses experimental conditions intermediate between those of column chromatography and HPLC. The typical FPLC system requires a pump that will deliver solvent to the column in the flow rate range 1–499 mL/hr with operating pressures of 0–40 bar. (HPLC pumps deliver solvent in a flow rate range of 0.010–10 mL/min with operating pressures of 1–400 bar. Classical chromatography columns are operated at atmospheric pressure.) Also required for FPLC are a controller, detector, and fraction collector. Since lower pressures are used in FPLC than in HPLC, a wider range of column supports is possible. Chromatographic techniques incorporated in an FPLC system are gel filtration, ion exchange, affinity (see Section G, p. 152), hydrophobic interaction, reversed phase, and chromatofocusing.

Perfusion Chromatography

A separation method that improves resolution and decreases the time required for analysis of biomolecules has recently been introduced. This method, called **perfusion chromatography**, relies on a type of particle support called POROS, which may be used in low-pressure and high-pressure liquid chromatography applications. In conventional chromatographic separations, some biomolecules in the sample move rapidly around and past the media particles while other molecules diffuse slowly through the particles (Figure 5.16A). The result is loss of resolution because some biomolecules exit the column before others. To improve resolution, the researcher with conventional media found it necessary to reduce the flow rate to allow for diffusion processes, increasing the time required for analysis. In other words, before the development of perfusion chromatography, the researcher had to choose between high speed–low resolution and low speed–high resolution. POROS particles have two types of pores—**through pores** (6000–8000 Å in diameter), which provide channels through the particles, and connected **diffusion pores** (800–1500 Å in diameter), which line the through pores and have very short diffusion path lengths (Figure 5.16B). This combination

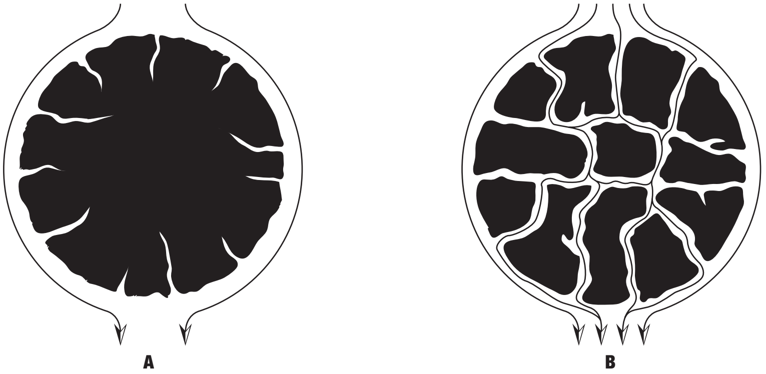

FIGURE 5.16 Transport of biomolecules through chromatographic media. **A** Conventional support particles. **B** POROS particles for perfusion chromatography.

pore system increases the porosity and the effective surface area of the particles and results in improved resolution and shorter analysis times (30 seconds to 3 minutes for POROS versus 30 minutes to several hours for conventional media).

POROS media, made by copolymerization of styrene and divinylbenzene, have high mechanical strength and are resistant to many solvents and chemicals. The functional surface chemistry of the particles can be modified to provide supports for many types of chromatography, including ion exchange, hydrophobic interaction, immobilized metal affinity, reversed phase, group-selective affinity, and conventional bioaffinity. Perfusion chromatography has been applied with success to the separation of peptides, proteins, and polynucleotides on both preparative and analytical scales.

In addition to high resolution and short analysis times, perfusion chromatography has the advantage of improved recovery of biological activity because active biomolecules spend less time on the column, where denaturing conditions may exist.

G. AFFINITY CHROMATOGRAPHY AND IMMUNOADSORPTION

The more conventional chromatographic procedures that we have studied up to this point rely on rather nonspecific physicochemical interactions between a sorbent and an analyte. The molecular characteristics of net charge, size, and polarity do not provide a basis for high selectivity in the separation and isolation of biomolecules. The desire for more specificity in chromatographic separations has led to the development of **affinity chromatography**. This technique offers the ultimate in specificity—separation on the basis of biological interactions. The biological function displayed by most macromolecules (antibodies, transport proteins, enzymes, nucleic acids, polysaccharides, receptor proteins, etc.) is a result of recognition of and interaction with specific molecules called **ligands**. This is illustrated by Equation 5.7, where M represents a macromolecule and L a smaller molecule or ligand. The two molecules interact in a specific manner to form a complex, L : M

>> $$\mathrm{L} + \mathrm{M} \rightleftharpoons \mathrm{L}:\mathrm{M} \longrightarrow \text{biological response} \qquad \textbf{(Eq. 5.7)}$$

In a biological system, the formation of the complex often triggers some response such as immunological action, control of a metabolic process, hormone action, catalytic breakdown of a substrate, or membrane transport. The biological response depends on proper molecular recognition and binding as shown in the reaction. The most common example of Equation 5.7 is the interaction that occurs between an enzyme molecule, E, and a substrate, S, with reversible formation of an ES complex. The biological event resulting from this interaction is the transformation of S to a metabolic product, P. Only the first step in Equation 5.7, formation of the complex, is of concern in affinity chromatography (see Chapter 8, Section B, p. 250).

In practice, affinity chromatography requires the preparation of an insoluble sorbent, to which appropriate ligand molecules (L) are covalently affixed. Thus, ligand molecules are immobilized on the stationary support. The affinity support is packed into a column through which a mixture containing

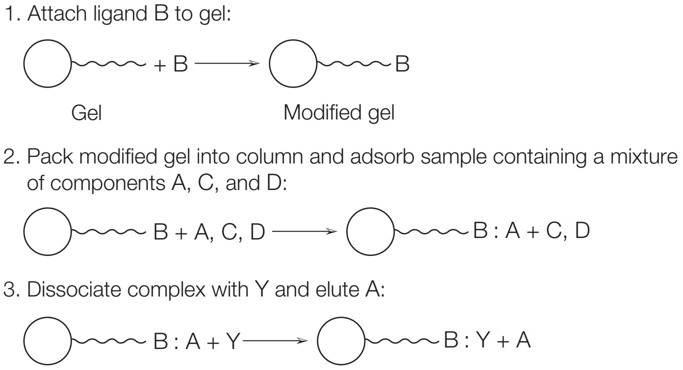
1. Attach ligand B to gel:
+ B
B
Gel
Modified gel
2. Pack modified gel into column and adsorb sample containing a mixture of components A, C, and D:
B + A, C, D
B : A + C, D
3. Dissociate complex with Y and elute A:
B : A + Y
B : Y + A

FIGURE 5.17 Purification of a macromolecule, **A,** by affinity chromatography. Ligand **B,** which has a specific affinity for **A,** is immobilized on the gel. **Y** represents an eluting agent that causes dissociation of **A**.

the desired macromolecule, M, is allowed to percolate. There are many types of molecules in the mixture, especially if it is a crude cell extract, but only macromolecules that recognize and bind to immobilized L are retarded in their movement through the column. After the nonbinding molecules have washed through the column, the desired macromolecules are eluted by gentle disruption of the L:M complex. Study Figure 5.17 for an illustration of affinity chromatography.

Affinity chromatography can be applied to the isolation and purification of virtually all biological macromolecules. It has been used to purify nucleic acids, enzymes, transport proteins, antibodies, hormone-receptor proteins, drug-binding proteins, neurotransmitter proteins, and many others.

Successful application of affinity chromatography requires careful design of experimental conditions. The essential components, which are outlined in the following subsections, are: (1) creation and preparation of a stationary matrix with immobilized ligand, and (2) design of column development and eluting conditions.

Chromatographic Media

Selection of the matrix used to immobilize a ligand requires consideration of several properties. The stationary supports used in gel-exclusion chromatography are found to be quite suitable for affinity chromatography because: (1) they are physically and chemically stable under most experimental conditions, (2) they are relatively free of nonspecific adsorption effects, (3) they have satisfactory flow characteristics, (4) they are available with very large pore sizes, and (5) they have reactive functional groups to which an appropriate ligand may be attached.

Four types of media possess most of these desirable characteristics: agarose, polyvinyl, polyacrylamide, and controlled-porosity glass (CPG) beads. Highly porous agarose beads such as Sepharose 4B (Amersham-Pharmacia-LKB-Hoefer) and Bio-Gel A-150 (Bio-Rad Laboratories) have virtually all of these characteristics and are the most widely used matrices. Polyacrylamide gels such as Bio-Gel P-300 (Bio-Rad) display many of the recommended features; however, their porosity is not especially high.

The Immobilized Ligand

The ligand, B, in Figure 5.17 can be selected only after the nature of the macromolecule to be isolated is known. When a hormone-receptor protein is to be purified by affinity chromatography, the hormone itself is an ideal candidate for the ligand. For antibody isolation, an antigen or hapten may be used as the ligand. If an enzyme is to be purified, a substrate analog, inhibitor, cofactor, or effector may be used as the immobilized ligand. The actual substrate molecule may be used as a ligand, but only if column conditions can be modified to avoid catalytic transformation of the bound substrate.

In addition to the foregoing requirements, the ligand must display a strong, specific, but reversible interaction with the desired macromolecule and it must have a reactive functional group for attachment to the matrix. It should be recognized that several types of ligand may be used for affinity purification of a particular macromolecule. Of course, some ligands will work better than others, and empirical binding studies can be performed to select an effective ligand.

Attachment of Ligand to Matrix

Several procedures have been developed for the covalent attachment of the ligand to the matrix. All procedures for gel modification proceed in two separate chemical steps: (1) activation of the functional groups on the matrix, and (2) joining of the ligand to the functional group on the matrix. The attachment method must leave the ligand in a form and position capable of binding to the desired macromolecule.

A wide variety of activated gels is now commercially available. The most widely used are described as follows.

Cyanogen Bromide-Activated Agarose

This gel is especially versatile because all ligands containing primary amino groups are easily attached to the agarose. It is available under the trade name CNBr-activated Sepahrose 4B. Since the gel is extremely reactive, very gentle conditions may be used to couple the ligand. One disadvantage of CNBr activation is that small ligands are coupled very closely to the matrix surface; macromolecules, because of steric repulsion, may not be able to interact fully with the ligand. The procedure for CNBr activation and ligand coupling is outlined in Figure 5.18A.

6-Aminohexanoic Acid (CH)-Agarose and 1,6-Diaminehexane (AH)-Agarose

These activated gels overcome the steric interference problems stated above by positioning a six-carbon spacer arm between the ligand and the matrix. Ligands with free primary amino groups can be covalently attached to CH-agarose, whereas ligands with free carboxyl groups can be coupled to AH-agarose. The attachment of ligands to AH and CH gels is outlined in Figure 5.18B,C.

Carbonyldiimidazole (CDI)-Activated Supports

Reaction with CDI produces gels that contain uncharged N-alkylcarbamate groups (see Figure 5.18D). CDI-activated agarose, dextran, and polyvinyl acetate are sold by Pierce Biotechnology Co. under the trade name Reacti-Gel.

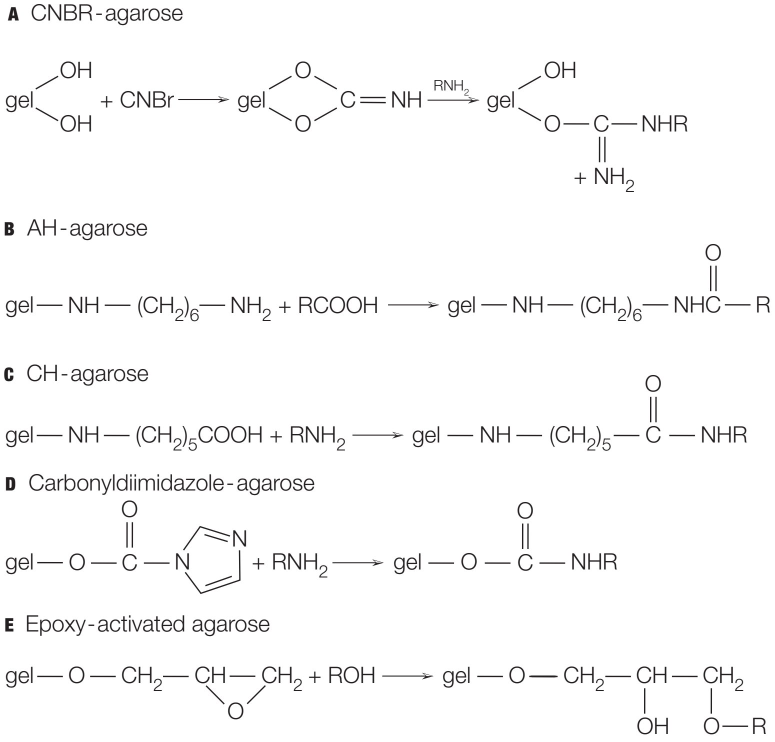
A CNBR-agarose
gel OH OH + CNBr → gel O O C=NH
RNH2
gel OH O—C—NHR
+NH2
B AH-agarose
gel—NH—(CH2)6—NH2 + RCOOH → gel—NH—(CH2)6—NHC—R
O
C CH-agarose
gel—NH—(CH2)5COOH + RNH2 → gel—NH—(CH2)5—C—NHR
O
D Carbonyldiimidazole-agarose
gel—O—C—N N + RNH2 → gel—O—C—NHR
O
E Epoxy-activated agarose
gel—O—CH2—CH—CH2 + ROH → gel—O—CH2—CH—CH2
O
OH O—R

FIGURE 5.18 Attachment of specific ligands to activated gels. R = ligands.

Epoxy-Activated Agarose

The structure of this gel is shown in Figure 5.18E. It provides for the attachment of ligands containing hydroxyl, thiol, or amino groups. The hydroxyl groups of mono-, oligo-, and polysaccharides can readily be attached to the gel.

Group-Specific Adsorbents

The affinity materials described up to this point are modified with a ligand having specificity for a particular macromolecule. Therefore, each time a biomolecule is to be isolated by affinity chromatography, a new adsorbent must be designed and prepared. Ligands of this type are called substance specific. In contrast, **group-specific adsorbents** contain ligands that have affinity for a class of biochemically related substances. Table 5.5 shows several commercially available group-specific adsorbents and their specificities. The principles behind binding of nucleic acids and proteins to group-specific adsorbents depend on the actual affinity adsorbent. In most cases, the immobilized ligand and macromolecule (protein or nucleic acid) interact through one or more of the following forces: hydrogen bonding, hydrophobic interactions, and/or covalent interactions. Some group-specific adsorbents deserve special attention. Phenyl- and octyl-Sepharose are gels used for **hydrophobic interaction chromatography**. These adsorbent materials separate proteins on the basis of their hydrophobic character. Because most proteins contain hydrophobic amino acid side chains, this method is widely used. Octyl-Sepharose is strongly hydrophobic; hence, it binds strongly to nonpolar

TABLE 5.5 Group-Specific Adsorbents Useful in Biochemical Applications

Group-Specific Adsorbent	Group Specificity
5′-AMP-agarose	Enzymes that have NAD^+ cofactor, ATP-dependent kinases
Benzamidine–Sepharose	Serine proteases
Boronic acid–agarose	Compounds with cis-diol groups, sugars: catecholamines, ribonucleotides, glycoproteins
Cibracron blue–agarose	Enzymes with nucleotide cofactors (dehydrogenases, kinases, DNA polymerases); serum albumin
Concanavalin A–agarose	Glycoproteins and glycolipids
Heparin–Sepharose	Nucleic acid–binding proteins, restriction endonucleases, lipoproteins
Iminodiacetate–agarose	Proteins with affinity for metal ions, serum proteins, interferons
Lentil lectin–Sepharose	Detergent-soluble membrane proteins
Lysine–Sepharose	Nucleic acids
Octyl-Sepharose	Weakly hydrophobic proteins, membrane proteins
Phenyl-Sepharose	Strongly hydrophobic proteins
Poly(A)-agarose	Nucleic acids containing poly(U) sequences, mRNA-binding proteins
Poly(U)-agarose	Nucleic acids containing poly(A) sequences, poly(U)-binding proteins
Protein A–agarose	IgG-type antibodies
Thiopropyl-Sepharose	—SH containing proteins

proteins. Phenyl-Sepharose is more weakly hydrophobic; therefore, it is more likely to reversibly bind strongly hydrophobic proteins.

The use of thiopropyl-Sepharose and boronic acid–agarose is an example of **covalent chromatography**, since relatively strong but reversible covalent bonds are formed between the affinity gel and specific macromolecules.

Metal affinity chromatography is a relatively new method that separates proteins on the basis of metal binding.

The availability of a great variety of group-specific adsorbents in prepacked columns makes possible the combination of FPLC and affinity chromatography for the separation and purification of proteins.

Immunoadsorption

One of the most effective modifications of affinity chromatography is **immunoaffinity,** also called **immunoadsorption**. The unique high specificity of antibodies for their antigens is exploited for the purification of antigens. The interaction between antigen and antibody is very selective and very strong, perhaps the most specific in affinity chromatography. In practice, the antibody is immobilized on a column support. The antibody may be obtained by immunizing a rabbit, but you must have relatively large amounts of the purified protein antigen. The better choice is to make monoclonal antibodies against the antigen. When a mixture containing several other proteins along with the protein antigen is passed through the column, only antigen binds; the other proteins wash off the column. Because antigen:antibody pairs form very strong complexes, it is not especially difficult to get the protein antigen to bind to the column. There is often difficulty, though, in eluting the bound protein without denaturing it. Effective eluting agents for disrupting the complexes and eluting the antigen are discussed in the next section.

Protein A-agarose in Table 5.5 is an example of immunoaffinity; however, this adsorbent does not recognize specific antibodies but, rather, the general family of immunoglobulin G antibodies.

Experimental Procedure for Affinity Chromatography

Although the procedure is different for each type of substance isolated, a general experimental plan is outlined here. Figure 5.19 provides a step-by-step plan in flowchart form. Many types of matrix-ligand systems are commercially available and the costs are reasonable, so it is not always necessary to spend valuable laboratory time for affinity gel preparation. Even if a specific gel is not available, time can be saved by purchasing preactivated gels for direct attachment of the desired ligand. Once the gel is prepared, the procedure is similar to that described earlier. The major difference is the use of shorter columns. Most affinity gels have high capacities and column beds less than 10 cm in length may be used. A second difference is the mode of elution. Ligand-macromolecule complexes immobilized on the column are held together by hydrogen bonding, ionic interactions, and hydrophobic effects. Any agent that diminishes these forces causes the release and elution of the macromolecule from the column. The elution step in affinity chromatography is represented in Figure 5.17, Step 3, where a general eluting agent, Y, dissociates the desired macromolecule, A. The common methods of elution are change of buffer pH, increase of buffer ionic strength, affinity elution, and chaotropic agents. The choice of elution method depends on many factors, including the types of forces responsible for complex formation and the stability of the ligand matrix and isolated macromolecule.

Buffer pH or Ionic Strength

If ionic interactions are important for complex formation, a change in pH or ionic strength weakens the interaction by altering the extent of ionization of ligand and macromolecule. In practice, either a decrease in pH or a gradual increase in ionic strength (continual or stepwise gradient) is used.

Affinity Elution

In this method of elution, a selective substance added to the eluting buffer competes for binding to the ligand or for binding to the adsorbed macromolecule.

Chaotropic Agents

If gentle and selective elution methods do not release the bound macromolecule, as may be the case in immunoadsorption, then mild denaturing agents can be added to the buffer. These substances deform protein and nucleic acid structure and decrease the stability of the complex formed on the affinity gel. The most useful agents are urea, guanidine $\cdot$ HCl, CNS^-, ClO_4^-, and CCl_3COO^-. These

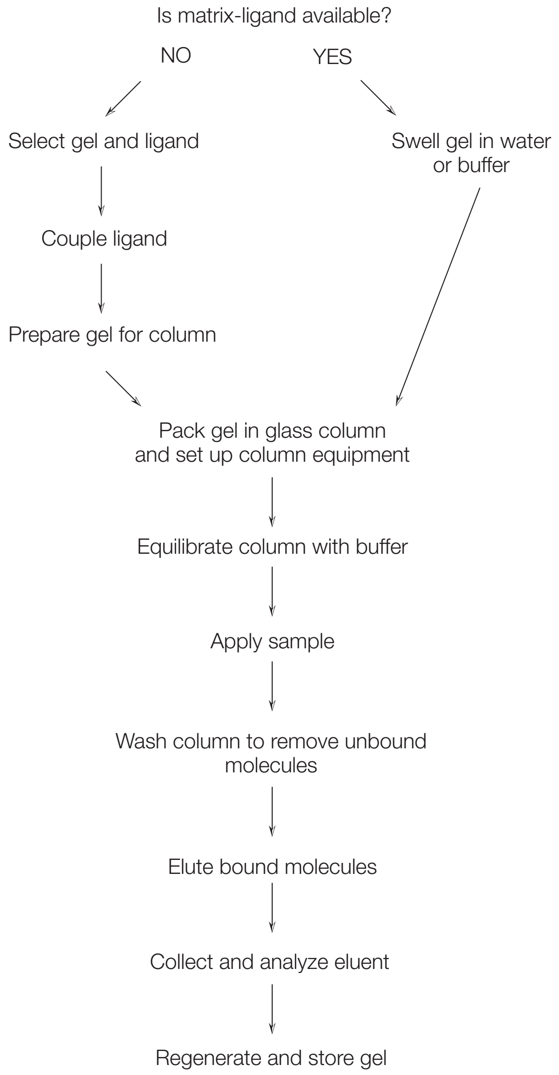
Is matrix-ligand available?
NO
YES
Select gel and ligand
Swell gel in water or buffer
Couple ligand
Prepare gel for column
Pack gel in glass column and set up column equipment
Equilibrate column with buffer
Apply sample
Wash column to remove unbound molecules
Elute bound molecules
Collect and analyze eluent
Regenerate and store gel

FIGURE 5.19 Experimental procedure for affinity chromatography.

substances should be used with care, as they may cause irreversible structural changes in the isolated macromolecule.

The application of affinity chromatography is limited only by the imagination of the investigator. Every year, literally hundreds of research papers appear with new and creative applications of affinity chromatography. Figure 5.20 illustrates the purification of α-chymotrypsin by affinity chromatography on immobilized D-tryptophan methyl ester. α-chymotrypsin can recognize and bind, but not chemically transform, D-tryptophan methyl ester. The enzyme catalyzes the hydrolysis of L-tryptophan methyl ester. The impure α-chymotrypsin mixture was applied to the gel, D-tryptophan methyl ester coupled to CH-Sepharose 4B, and the column washed with Tris buffer. At the point shown by the arrow, the eluent was changed to 0.1 *M* acetic acid. The decrease in pH caused release of α-chymotrypsin from the column.

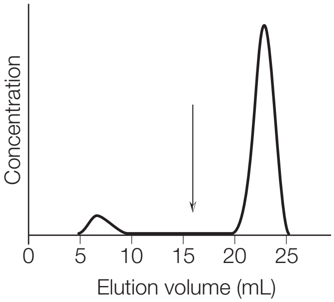

FIGURE 5.20 Purification of α-chymotrypsin by affinity chromatography on immobilized D-tryptophan methyl ester. From ***Affinity Chromatography: Principles and Methods***. Pharmacia (Uppsala, Sweden).

STUDY EXERCISE 5.7 Affinity Chromatography

You have discovered that leaves from the quaking aspen tree contain an enzyme that catalyzes the transfer of a phosphoryl group from ATP to various carbohydrates. You wish to purify the enzyme, a kinase, by affinity chromatography. Select two stationary phases from Table 5.5 that might be effective in this separation.

H. MEMBRANE-BASED CHROMATOGRAPHY

Most of the column chromatographic stationary supports discussed up to this point were based on particles, usually in the form of synthetic resin beads. Biotech labs and pharmaceutical companies involved in the manufacture of protein- and nucleic acid-based drugs require effective purification methods that offer large capacity and speed and are usually dependent on chromatography. There are several limitations to the use of bead-based stationary supports, and it is challenging to improve separating ability and speed especially in purifying large biomolecules ($>$ 200,000 daltons). Separation of biomolecules depends on their binding to functional groups (ion exchange, affinity, etc.) that are often present inside the beads and thus not readily available for binding of larger molecules. Increasing the number of functional groups by increasing the height of a column is not an option, as the resin beads are compressible and thus cause a reduction in the flow rate.

Many of the limitations of conventional bead columns may be overcome by the use of synthetic, microporous membranes as stationary supports. These membranes offer a three-dimensional structure with open pores, which are readily available to larger biomolecules (Figure 5.21). Research on chromatography membranes shows that movement of molecules takes place by convection rather than diffusion, which enhances the distribution of molecules throughout the support. Other advantages of membrane-based chromatography compared to conventional bead columns include higher flow rates (up to 100 times faster), higher capacity (10 times greater), and the fact that the membranes can be packed into layers, which greatly improves the effectiveness of biomolecule separation.

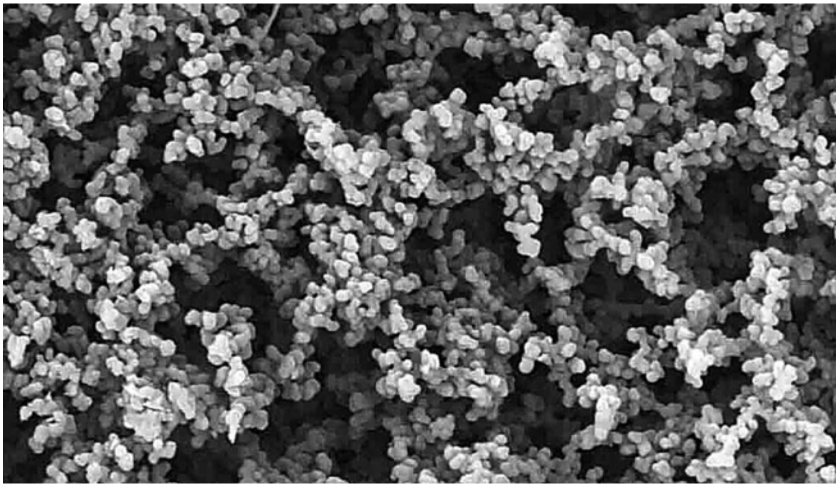

FIGURE 5.21 The inside of an ion-exchange chromatography membrane showing the readily accessible pores. *Photo courtesy of Pall Corporation,* www.pall.com.

Membranes are available for several chromatographic modes including ion exchange, affinity, hydrophobic interaction, and reverse phase. The ion-exchange membranes have a polyethersulfone base modified with sulfonic acid groups (cation exchanger) or quaternary amines (anion exchanger). The membranes have been found to be effective in separating and purifying proteins, plasmid DNA, and other biomolecules. Membranes are available in disposable plastic cartridges (5 mL, 140 mL, or 5000 mL volumes), which can be linked to HPLC or FPLC systems or used alone with a peristaltic pump or syringe (Figure 5.22). Membrane cartridges are designed so they can be set up in flow systems, which allows for purification of large sample sizes as used in industry (Figure 5.23).

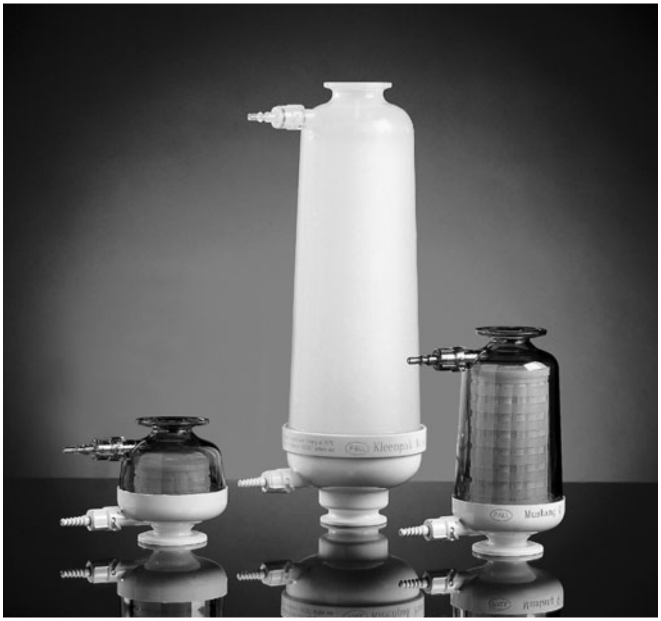

FIGURE 5.22 Disposable membrane cartridges. *Photo courtesy of Pall Corporation,* www.pall.com.

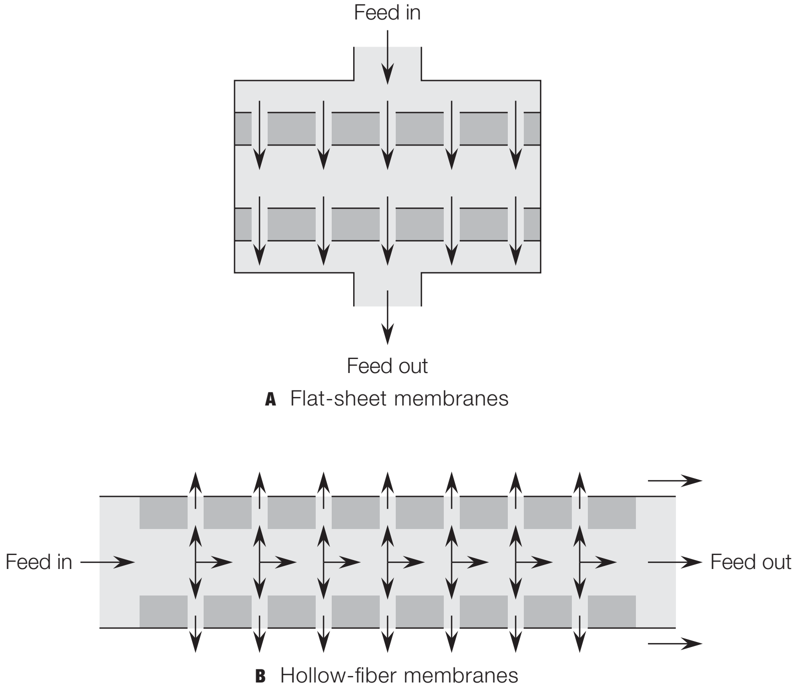

FIGURE 5.23 Flow options in membrane-based chromatography. **A** Flat-sheet and **B** Hollow-fiber membranes.

Study Problems

1. Amino acid analyzers are instruments that automatically separate amino acids by cation-exchange chromatography. Predict the order of elution (first to last) for each of the following sets of amino acids at pH = 4.
   (a) Gly, Asp, His
   (b) Arg, Glu, Ala
   (c) Phe, His, Glu
2. Predict the relative order of paper chromatography R_f values for the amino acids in the following mixture: Ser, Lys, Leu, Val, and Ala. Assume that the developing solvent is *n*-butanol, water, and acetic acid.
3. In what order would the following proteins be eluted from a DEAE-cellulose ion exchanger by an increasing salt gradient. The pH_I is listed for each protein.

Egg albumin, 4.6	Cytochrome *c*, 10.6
Pepsinogen, 1.0	Myoglobin, 6.8
Serum albumin, 4.9	Hemoglobin, 6.8

4. Describe the various detection methods that can be used in HPLC. What types of biomolecules are detected by each method?
5. Draw the elution curve (A_{280} vs. fraction number) obtained by passing a mixture of the following proteins through a column of Sephadex G-100. The molecular mass is given for each protein.

Myoglobin, 16,900	Myosin, 524,000
Catalase, 222,000	Serum albumin, 68,500
Cytochrome *c*, 13,370	Chymotrypsinogen, 23,240

6. Briefly describe how you would experimentally measure the exclusion limit for a Sephadex gel whose bottle has lost its label.
7. Name three enzymes that you predict will bind to the affinity support, 5′-AMP-agarose.
8. Explain the elution order of amino acids in Figure 5.11.
9. Describe how you would use affinity elution to remove the enzyme alcohol dehydrogenase bound to a Cibracron blue–agarose column.
10. You have read in your botany textbook that bean leaves contain an enzyme that catalyzes the hydrolysis of the methyl esters of aromatic amino acids (for example, the methyl ester of phenylalanine). You wish to design a plan for isolating and purifying the enzyme. Assume that your last step of purification is to be affinity chromatography. Describe an affinity sorbent that might be effective in purifying the protein.
11. In affinity chromatography and immunoadsorption, analytes are usually tightly bound to the sorbent. Which of the following types of bonding are important in forming these complexes?

    Hydrogen bonding
    Covalent bonds between carbon atoms
    Ionic bonds
    Hydrophobic interactions
    Van der Waals forces
12. You have isolated an unknown protein and wish to estimate its molecular weight using column gel filtration with several standard proteins. The gel material is Sephacryl S-100. The following elution volumes were measured.

Protein	**MW**	**Elution volume (mL)**
Hemoglobin	64,500	10.1
Egg albumin	45,000	14.5
Chymotrypsinogen	23,200	25.0
Myoglobin	16,900	28.2
Cytochrome ***c***	13,000	31.0
Unknown	?	7.5

    What is the approximate molecular weight of the unknown protein?
    Hint: Prepare a graph like in Figure 5.9.
13. Which gel filtration medium would be more effective in separating the six proteins in Study Problem 12, Sephadex G-50 or G-100? Why?
14. Which column chromatographic method is better to separate the proteins below, ion exchange or gel filtration?

Hemoglobin	MW = 64,500;	isoelectric pH = 6.8
Myoglobin	MW = 17,000;	isoelectric pH = 7.0

15. Use information from Chapter 2 to help you find the following:
    (a) a recent research article on the purification of egg white lysozyme by membrane-based chromatography
    (b) the availability of spin columns for membrane-based chromatography
    (c) chiral chromatography using membranes

Further Reading

H. Aboul-Enein and I. Ali, ***Chiral Separations by Liquid Chromatography: Theory and Applications*** (2003), Marcel Dekker (New York).

J. Berg, J. Tymoczko, and L. Stryer, ***Biochemistry, 6th ed.*** (2007), Freeman (New York). Introduction to chromatography.

R. Boyer, ***Concepts in Biochemistry, 3rd ed.*** (2006), John Wiley & Sons (Hoboken, NJ), pp. 84–86. An introduction to chromatography.

J. Coligan, (Editor), ***Short Protocols in Protein Science*** (2003), John Wiley & Sons (Hoboken, NJ).

R. Curtright, R. Emry, and J. Markwell, ***J. Chem. Educ. 76,*** 249 (1999). "Student Understanding of Chromatography: A Hands-On Approach."

R. Garrett and C. Grisham, ***Biochemistry, 4th ed.*** (2010), Brooks/Cole (Belmont, CA). An introduction to chromatography.

R. Ghosh, ***J. Chromatogr. 952***, 13–27 (2002), "Protein separation using membrane chromatography."

D. Gjerde, C. Hanna, and D. Hornby, ***DNA Chromatography*** (2002), John Wiley & Sons (New York).

F. Gorga, ***J. Chem. Educ.*** **77,** 264 (2000). "A Problem-Solving Approach to Chromatography in the Biochemistry Lab."

E. Grushka and P. Brown, ***Advances in Chromatography*** (2003), Marcel Dekker (New York).

M. Gupta, (Editor), ***Methods for Affinity-Based Separations for Enzymes and Proteins*** (2002), Birkhauser (Cambridge, MA).

D. Hage and J. Cazes (Editors), ***Handbook of Affinity Chromatography, 2nd ed.*** (2005), Marcel Dekker (New York).

R. Hatti-Kaul and B. Mattiasson, (Editors), ***Isolation and Purification of Proteins*** (2003), Marcel Dekker (New York).

J. Kemsley, ***Chem. Eng. News,*** May 18, 2009, pp. 11–18. "Modernizing TLC."

T. Kowalska and J. Sherma, ***Preparative Thin-Layer Chromatography***, (2006), CRC Press, Taylor & Francis Group (Boca Raton, FL).

J. Miller, ***Chromatography: Concepts and Contrasts,*** 2nd ed. (2009), Wiley-Interscience.

D. Nelson and M. Cox, ***Lehninger Principles of Biochemistry, 5th ed.*** (2009), Freeman (New York). Introduction to chromatography.

C. Poole, ***The Essence of Chromatography*** (2003), Elsevier (Amsterdam).

M. Pugh and E. Schultz, ***Biochem. Mol. Biol. Educ.*** **30**, 179–183 (2002). "Assessment of the Purification of a Protein by Ion Exchange and Gel Permeation Chromatography."

D. Sheehan, ***Physical Biochemistry: Principles and Applications, 2nd ed.*** (2009), Wiley-Blackwell (Hoboken, NJ).

J. Sherma and B. Fried (Editors), ***Handbook of Thin-Layer Chromatography*, 3rd ed.** (2003), Marcel Dekker (New York-Basel-Hong Kong).

L. Snyder, J. Kirkland, J. Dolan, ***Introduction to Modern Liquid Chromatography*, 3rd ed.** (2009), Wiley (Hoboken, NJ).

A. Striegel, W. Yau, J. Kirkland, D. Bly, ***Modern Size-Exclusion Liquid Chromatography*, 2nd ed.** (2009), Wiley.

M. Urh, D. Simpson, K. Zhao, ***Methods Enzymology,*** **463,** 417–438 (2009). "Affinity Chromatography: General Methods."

D. Voet and J. Voet, ***Biochemistry, 3rd ed.*** (2005), John Wiley & Sons (Hoboken, NJ). Use of chromatography in protein purification.

D. Voet, J. Voet, and C. Pratt, ***Fundamentals of Biochemistry, 3rd ed.*** (2008), John Wiley & Sons (Hoboken NJ). Chromatography in protein purification.

J. Watson et al., ***Molecular Biology of the Gene, 5th ed.*** (2004), Benjamin/Cummings (San Francisco), Chapter 20 (Methods).

K. Williams, ***J. Chem. Educ.* 79**, 922 (2002). "Colored Bands: History of Chromatography."

K. Wilson and J. Walker, (Editors), ***Principles and Techniques of Practical Biochemistry, 5th ed.*** (2000), Cambridge University Press (Cambridge), pp. 619–687. Chromatographic techniques.

M. Zachariou (Editor), ***Affinity Chromatography: Methods and Protocols*** (2007), Humana Press.

http://aesop.rutgers.edu/~dbm/affinity.html
Affinity and immunoadsorbent chromatography.

http://www.sigmaaldrich.com/life-science/proteomics/protein-chromatography/affinity-chromatography.html
Affinity chromatography.

http://www.bio.davidson.edu/Courses/genomics/method/Affinity.html
Methods of affinity chromatography.

http://andromeda.rutgers.edu/~mendelso/biochem581/BiochemToolsLecture.pdf
Methods of chromatography.

https://products.appliedbiosystems.com/ab/en/US/adirect/ab;jsessionid=y2fQLcRpjVn7gv5h1tmJ0yqWtPkF8WXpBKL1Qn1LxnsD8ZdpyfP1!1614902786?cmd=catNavigate2&catID=600500
Perfusion chromatography.

http://www.separationsnow.com/coi/cda/home.cda?chId=0
Resources for chromatography.

http://www.chiraltech.com
Supplier of chiral columns and chromatography materials.

http://www1.gelifesciences.com/aptrix/upp01077.nsf/content/na_homepage
On Home Page, select "Protein purification."

http://www.encyclopedia.com/doc/1P3-190345261.html
Membrane chromatography.

http://www.pall.com/biopharm_33053.asp
Membrane chromatography starter kits.

http://www.wpi.edu/Pubs/ETD/Available/etd-050405-174109/unrestricted/cao.pdf
A thesis on protein separation with ion-exchange membrane chromatography.

http://www.iptonline.com/articles/public/page50loresnonprint.pdf
Purifying large biomolecules using membrane chromatography.

http://www.camag.com
Products for planar chromatography.

http://www.piercenet.com/Products/Browse.cfm?fldID=1A66B494-33B2-4358-804A-6FDB69A7FFAB
Ion-exchange spin columns.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

CHAPTER 6

CHARACTERIZATION OF PROTEINS AND NUCLEIC ACIDS BY ELECTROPHORESIS

A. Theory of Electrophoresis 166
B. Methods of Electrophoresis 167
C. Practical Aspects of Electrophoresis 188

Study Problems 198
Further Reading 199

Electrophoresis is an analytical tool that allows biochemists to examine the differential movement of charged molecules in an electric field. A. Tiselius, a Swede who invented the technique in the 1930s, performed experiments in free solution that were severely limited by the effects of diffusion and convection currents. Modern electrophoretic techniques use a polymerized gel-like matrix, which is more stable as a support medium. The sample to be analyzed is applied to the medium as a spot or thin band; hence, the term **zonal** electrophoresis is often used. The migration of molecules is influenced by: (1) the size, shape, charge, and chemical composition of the molecules to be separated; (2) the rigid, mazelike matrix of the gel support; and (3) the applied electric field. Electrophoresis, which is a relatively rapid, inexpensive, and convenient technique, is capable of analyzing and purifying many different types of biomolecules, but is especially effective with proteins and nucleic acids. The newest version of the analytical technique, **capillary electrophoresis (CE),** provides extremely high resolution and is useful for analysis of both large and small molecules. CE has been found to be especially useful in the analysis of pharmaceuticals.

Proteomics, the discipline that attempts systematic, large-scale studies on the structure and function of gene products in an organism or cell, is expanding rapidly because of the availability of electrophoresis (especially two-dimensional techniques) to analyze proteins and peptides.

Even though electrophoresis has been studied for more than 80 years, it has been a challenge to provide an accurate, theoretical description of the electrophoretic movement of molecules in a gel support. However, the lack of theoretical understanding has not hampered growth in the use of the technique in separating and characterizing a wide variety of biomolecules.

A. THEORY OF ELECTROPHORESIS

Introduction

The movement of a charged molecule in a medium subjected to an electric field is represented by Equation 6.1.

>> $$v = \frac{Eq}{f} \qquad \textbf{(Eq. 6.1)}$$

where

E = the electric field in volts/cm

q = the net charge on the molecule

f = frictional coefficient, which depends on the mass and shape of the molecule

ν = the velocity of the molecule

The charged particle moves at a velocity that depends directly on the electric field (E) and charge (q), but inversely on a counteracting force generated by the viscous drag (f). The applied voltage represented by E in Equation 6.1 is usually held constant during electrophoresis, although some experiments are run under conditions of constant current (where the voltage changes with resistance) or constant power (the product of voltage and current). Under constant-voltage conditions, Equation 6.1 shows that the movement of a charged molecule depends only on the ratio q/f. For molecules of similar conformation (for example, a collection of linear DNA fragments or spherical proteins), f varies with size but not shape; therefore, the only remaining variables in Equation 6.1 are the charge (q) and mass dependence of f, meaning that under such conditions molecules migrate in an electric field at a rate proportional to their charge-to-mass ratio.

Theory and Practice

The movement of a charged particle in an electric field is often defined in terms of **mobility,** μ, the velocity per unit of electric field (Equation 6.2).

>> $$\mu = \frac{v}{E} \qquad \textbf{(Eq. 6.2)}$$

This equation can be modified using Equation 6.1.

>> $$\mu = \frac{Eq}{Ef} = \frac{q}{f} \qquad \textbf{(Eq. 6.3)}$$

In theory, if the net charge, q, on a molecule is known, it should be possible to measure f and obtain information about the hydrodynamic size and shape of that molecule by investigating its mobility in an electric field. Attempts to define f by electrophoresis have not been successful, primarily because Equation 6.3 does not adequately describe the electrophoretic process. Important factors that are not accounted for in the equation are interaction of migrating molecules with the support medium and shielding of the molecules by buffer ions. This means that electrophoresis is not useful for describing specific details about the shape of a molecule. Instead, it has been applied to the analysis of purity and size of macromolecules. Each molecule in a mixture is expected to have a unique charge and size, and its mobility in an electric field will therefore be unique. This expectation forms the basis for analysis and separation by all electrophoretic methods. The technique is especially useful for the analysis of amino acids, peptides, proteins, nucleotides, nucleic acids, and other charged molecules, including pharmaceuticals.

B. METHODS OF ELECTROPHORESIS

All modes of electrophoresis are based on the principles just outlined. The major difference among the various methods is the type of support medium. Cellulose and cellulose acetate are used as a support medium for low-molecular-weight biochemicals like amino acids and carbohydrates. Polyacrylamide and agarose gels are widely used as support media for larger molecules. In capillary electrophoresis, several different types of support media, including the natural, untreated surfaces inside a silica narrow bore capillary tube, are used. Geometries (vertical and horizontal), buffers, and electrophoretic conditions provide many different experimental arrangements for the variety of methods described here.

Polyacrylamide Gel Electrophoresis (PAGE)

Gels formed by polymerization of acrylamide have several positive features in electrophoresis: (1) high resolving power for small and moderately sized proteins and nucleic acids (up to approximately 1×10^6 daltons), (2) acceptance of relatively large sample sizes, (3) minimal interactions of the migrating molecules with the matrix, and (4) physical stability of the matrix. Recall from the earlier discussion of gel filtration (Chapter 5, p. 132) that gels can be prepared with different pore sizes by changing the concentration of cross-linking agents. Electrophoresis through polyacrylamide gels leads to enhanced resolution of sample components because the separation is based on both molecular sieving and electrophoretic mobility. The order of molecular movement in gel filtration and PAGE is very different, however. In gel filtration, large molecules migrate through the matrix faster than small molecules. The opposite is the case for gel electrophoresis, where there is no void volume in the matrix, but only a continuous network of pores throughout the gel. The electrophoresis gel is comparable to a single bead in gel filtration. Therefore, large molecules do not move easily through the medium, and the order of movement is small molecules followed by large molecules.

Preparation of Gels

Polyacrylamide gels are prepared by the free radical polymerization of acrylamide and the cross-linking agent *N*,*N*′-methylene-bis-acrylamide (Figure 6.1). Chemical polymerization is controlled by an initiator-catalyst system, ammonium persulfate–*N*,*N*,*N*′,*N*′-tetramethylethylenediamine (TEMED). Photochemical polymerization may be initiated by riboflavin in the presence of ultraviolet (UV) radiation. A standard gel for protein separation is 7.5% polyacrylamide. It can be used over the molecular size range of 10,000 to 1,000,000 daltons; however, the best resolution is obtained in the range of 30,000 to 300,000 daltons. The resolving power and molecular size range of a gel depend on the concentrations of acrylamide and bis-acrylamide (see Table 6.1 for effective ranges of protein separation). Lower concentrations give gels with larger pores, allowing analysis of higher-molecular-weight biomolecules. In contrast, higher concentrations of acrylamide give gels with smaller pores, allowing analysis of lower-molecular-weight biomolecules (see Table 6.2 for effective ranges of DNA separation).

Polyacrylamide gel electrophoresis has been done using either of two instrumental arrangements, column or slab. Column gels are rarely used today, but they do have historical significance as they are important forerunners of

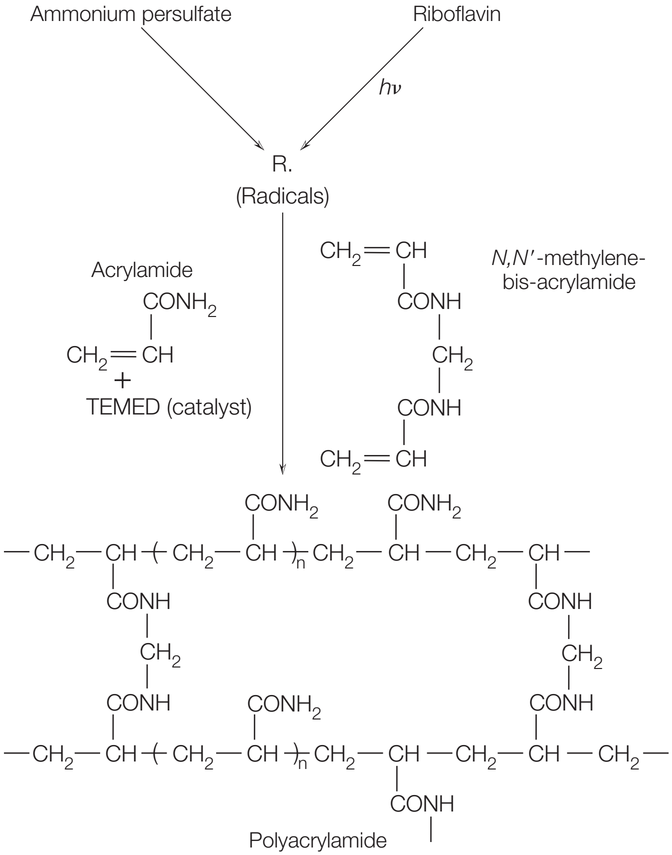

FIGURE 6.1 Chemical reactions illustrating the copolymerization of acrylamide and N,N′-methylene-bis-acrylamide. See text for details.

TABLE 6.1 Separation of Proteins by PAGE. The recommended concentrations of acrylamide (% w/v) are given that will best separate proteins of different molecular weight ranges

Protein Molecular Weight Range	Recommended % Acrylamide (w/v)[1]
<10,000	>15%
10,000–80,000	10–15%
20,000–150,000	5–10%
>100,000	3–5%

[1] Ratio of acrylamide to bis-acrylamide, 37.5:1 (g:g).

TABLE 6.2 Effective Range of Separation of DNA by PAGE

Acrylamide[1] (% w/v)	Range of Separation (bp)	Bromophenol Blue[2]	Xylene Cyanol[2]
3.5	1000–2000	100	450
5.0	80–500	65	250
8.0	60–400	50	150
12.0	40–200	20	75
20.0	5–100	10	50

[1] Ratio of acrylamide to bis-acrylamide, 20:1.
[2] The numbers (in bp) represent the size of DNA fragment with the same mobility as the dye.

more modern slab gels. Figure 6.2 shows the typical arrangement for a column gel. Glass tubes (10 cm × 6 mm i.d.) are filled with a mixture of acrylamide, N,N'-methylene-bis-acrylamide, buffer, and free radical initiator-catalyst. Polymerization occurs in 30 to 40 minutes. The gel column is inserted between two separate buffer reservoirs. The upper reservoir usually contains the cathode (−) and the lower the anode (+). Gel electrophoresis is usually carried out at basic pH, where most biological polymers are anionic; hence, they move down toward the anode. The sample to be analyzed is layered on top of the gel, and voltage is applied to the system. DNA and many proteins are colorless, so a reagent dye must be added to monitor the rate of electrophoresis. A “tracking dye” is applied, which moves more rapidly through the gel than the sample components. When the dye band has moved to the opposite end of the column, the voltage is turned off and the gel is removed from the column and stained with a dye. Chambers for column gel electrophoresis, instruments that hold several glass tubes, are commercially available or can be constructed from inexpensive materials.

Slab Gel Electrophoresis

Slab gels are now much more widely used than column gels. A slab gel on which several samples may be analyzed is more convenient to make and use than several individual column gels. Slab gels also offer the advantage that all samples are

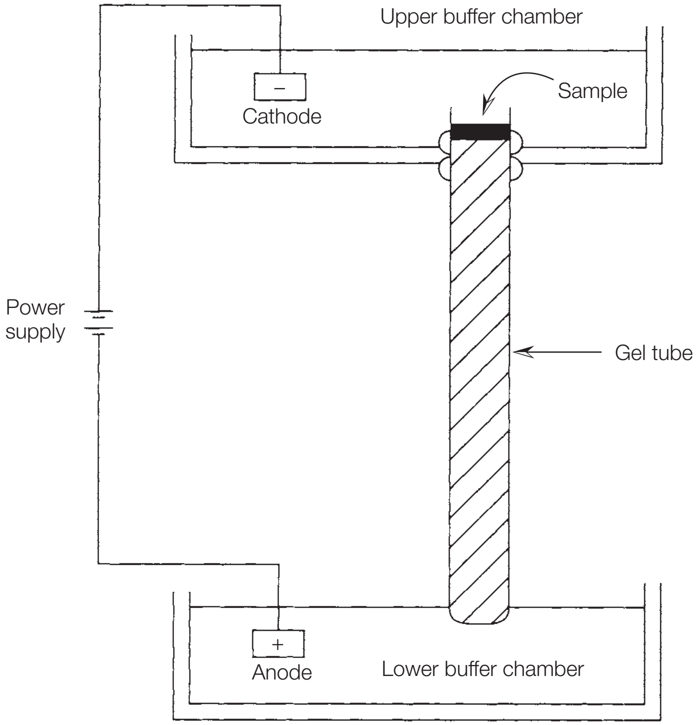

FIGURE 6.2 A column gel for polyacrylamide gel electrophoresis.

analyzed in a matrix environment that is identical in composition compared to glass tubes, which may have slightly different compositions. A typical vertical slab gel apparatus is shown in Figure 6.3. The polyacrylamide slab is prepared between two glass plates that are separated by spacers (Figure 6.4). The spacers allow a uniform slab thickness of 0.5 to 2.0 mm, which is appropriate for analytical and some preparative procedures. Slab gels are usually 8 × 10 cm or 10 × 10 cm, but for nucleotide sequencing, slab gels as large as 20 × 40 cm are often required.

A plastic "comb" inserted into the top of the slab gel during polymerization forms indentations in the gel that serve as sample wells. Up to 20 sample wells may be formed. After polymerization, the comb is carefully removed and the wells are rinsed thoroughly with buffer to remove salts and any unpolymerized acrylamide. The gel plate is clamped into place between two buffer reservoirs, a sample is loaded into each well, and voltage is applied. For visualization, the slab is removed and stained with an appropriate dye, or in some cases, the gel may be prestained with the dye before electrophoresis.

Perhaps the most difficult and inconvenient aspect of polyacrylamide gel electrophoresis is the preparation of gels. The monomer, acrylamide, is a neurotoxin and a carcinogenic agent; hence, special handling is required. Other necessary reagents including catalysts and initiators also require special handling and are unstable. In addition, it is difficult to make gels that have reproducible thicknesses and compositions. Many researchers are now turning to the use of precast polyacrylamide gels. Several manufacturers now offer gels precast in glass or plastic cassettes. Gels for all experimental operations are available, including

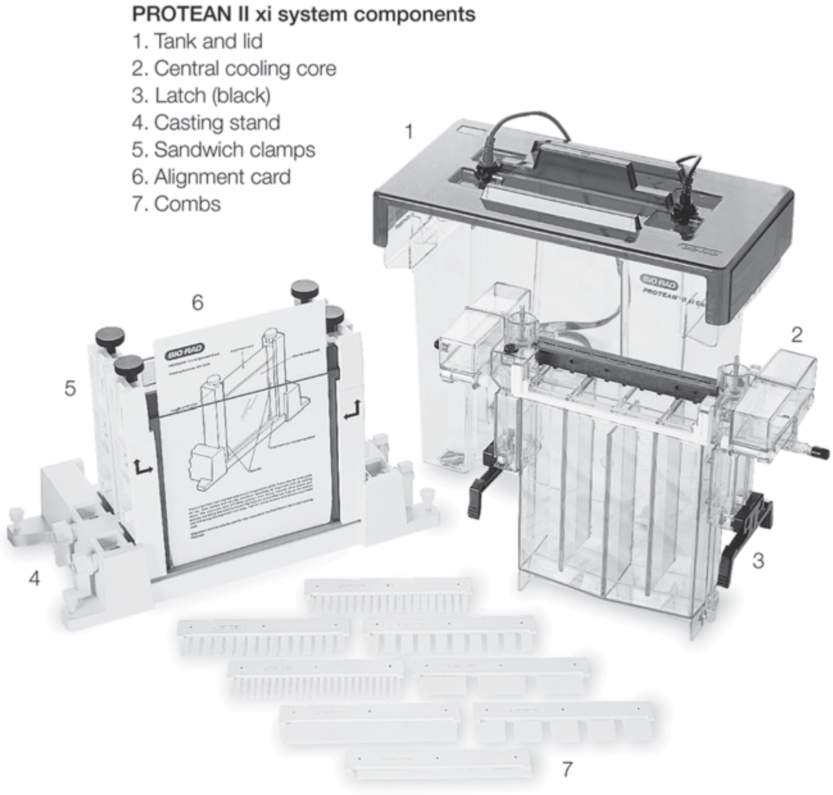
PROTEAN II xi system components
1. Tank and lid
2. Central cooling core
3. Latch (black)
4. Casting stand
5. Sandwich clamps
6. Alignment card
7. Combs
1
6
5
2
3
4
7

FIGURE 6.3 A vertical electrophoresis apparatus for slab gel analysis of proteins. *Permission to use the materials has been granted by Bio-Rad Laboratories, Inc.,* www.bio-rad.com.

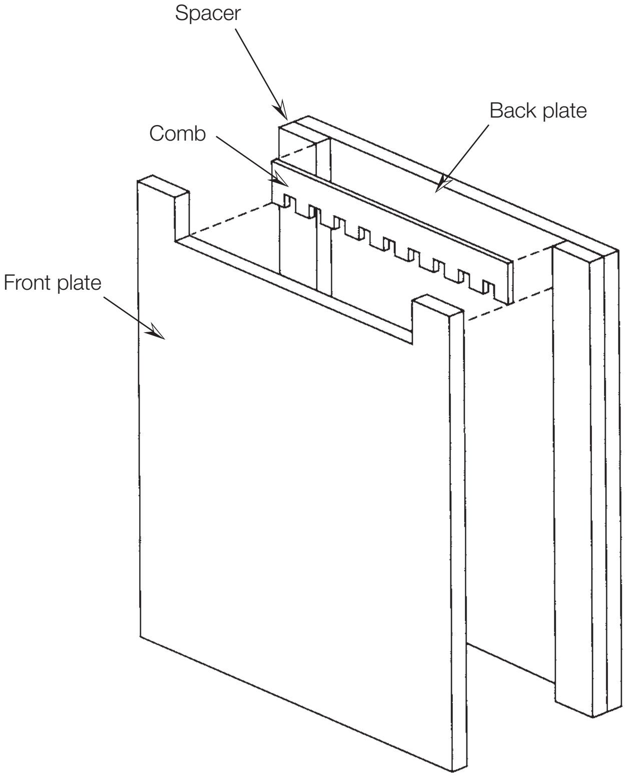
Spacer
Back plate
Comb
Front plate

FIGURE 6.4 Arrangement of two glass plates with spacers to form a slab gel. The comb is used to prepare wells for placement of samples.

single percentage (between 3% and 27%) or gradient gel concentrations and a variety of sample well configurations and buffer chemistries. More details on precast gels will be given in Section C, Practical Aspects of Electrophoresis.

Several modifications of PAGE have greatly increased its versatility and usefulness as an analytical tool.

Discontinuous Gel Electrophoresis

The experimental arrangement for "disc" gel electrophoresis is shown in Figure 6.5. Three significant characteristics of this method are that: (1) there are two gel layers, a lower or **resolving gel** and an upper or **stacking gel;** (2) the

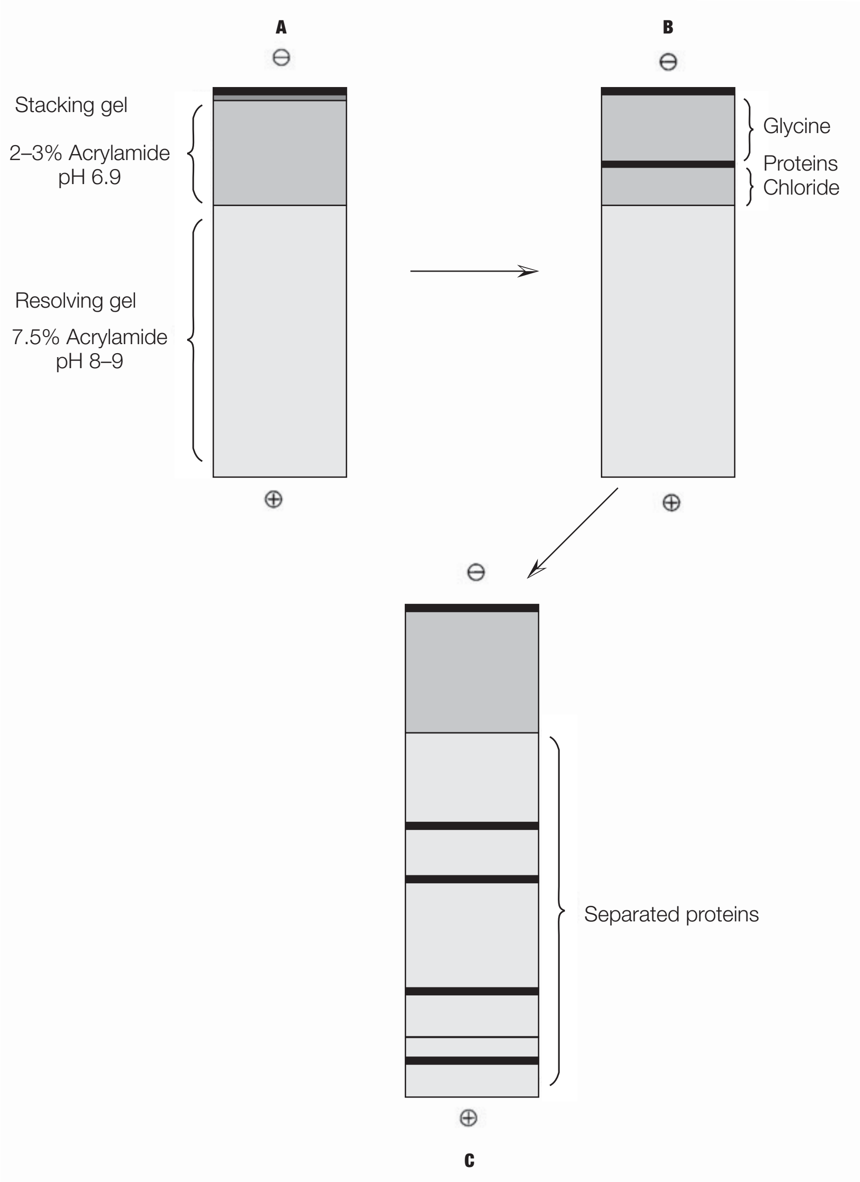
A
⊖
Stacking gel
2–3% Acrylamide
pH 6.9
Resolving gel
7.5% Acrylamide
pH 8–9
⊕
B
⊖
Glycine
Proteins
Chloride
⊕
⊖
Separated proteins
⊕
C

FIGURE 6.5 The process of disc gel electrophoresis. **A** Before electrophoresis. **B** Movement of chloride, glycinate, and protein through the stacking gel. **C** Separation of protein samples by the resolving gel.

buffers used to prepare the two gel layers are of different ionic strengths and pH; and (3) the stacking gel has a lower acrylamide concentration, so its pore sizes are larger. These three changes in the experimental conditions cause the formation of highly concentrated bands of sample in the stacking gel and greater resolution of the sample components in the lower gel. Sample concentration in the upper gel occurs in the following manner. The sample is usually dissolved in glycine-chloride buffer, pH 8 to 9, before loading on the gel. Glycine exists primarily in two forms at this pH, a zwitterion and an anion (Equation 6.4).

>> $$H_3\overset{+}{N}CH_2COO^- \rightleftharpoons H_2NCH_2COO^- + H^+ \qquad \textbf{(Eq. 6.4)}$$

The average charge on glycine anions at pH 8.5 is about −0.2. When the voltage is turned on, buffer ions (glycinate and chloride) and protein or nucleic acid sample move into the stacking gel, which has a pH of 6.9. Upon entry into the upper gel, the equilibrium of Equation 6.4 shifts toward the left, increasing the concentration of glycine zwitterion, which has no net charge and hence no electrophoretic mobility. In order to maintain a constant current in the electrophoresis system, a flow of anions must be maintained. Since most proteins and nucleic acid samples are still anionic at pH 6.9, they replace glycinate as mobile ions. Therefore, the relative ion mobilities in the stacking gel are chloride > protein or nucleic acid sample > glycinate. The sample will tend to accumulate and form a thin, concentrated band sandwiched between the chloride and glycinate as they move through the upper gel. Since the acrylamide concentration in the stacking gel is low (2% to 3%), there is little impediment to the mobility of the large sample molecules.

Now, when the ionic front reaches the lower gel with pH 8 to 9 buffer, the glycinate concentration increases and anionic glycine and chloride carry most of the current. The protein or nucleic acid sample molecules, now in a narrow band, encounter both an increase in pH and a decrease in pore size. The increase in pH would, of course, tend to increase electrophoretic mobility, but the smaller pores decrease mobility. The relative rate of movement of anions in the lower gel is chloride > glycinate > protein or nucleic acid sample. The separation of sample components in the resolving gel occurs as described in an earlier section on gel electrophoresis. Each component has a unique charge/mass ratio and a discrete size and shape, which directly influence its mobility. Disc gel electrophoresis yields excellent resolution and is the method of choice for analysis of proteins and nucleic acid fragments. Protein or nucleic acid bands containing as little as 1 or 2 μg can be detected by staining the gels after electrophoresis.

STUDY EXERCISE 6.1 PAGE

You are using the technique of nondenaturing PAGE to separate a mixture of two proteins, bovine serum albumin and bovine hemoglobin. Assume that you are using a buffer of pH 8.0 and a gel of 7.5% acrylamide, and that silver staining was used for detection. Show the results of your experiment by drawing a gel with dark bands for each protein.

Protein	pH_I	MW
Bovine serum albumin	4.9	65,000
Bovine hemoglobin	6.8	65,000

Sodium Dodecyl Sulfate–Polyacrylamide Gel Electrophoresis (SDS-PAGE)

If protein samples are treated so that they have a uniform charge, electrophoretic mobility then depends primarily on size (see Equation 6.3). The molecular weights of proteins may be estimated if they are subjected to electrophoresis in the presence of a detergent, sodium dodecyl sulfate (SDS), and a disulfide bond reducing agent, mercaptoethanol. This method is often called **denaturing electrophoresis**.

The electrophoretic techniques previously discussed are called **nondenaturing** or **"native" PAGE** and are used when an investigator requires that the protein analyzed still retains its biological activity. This would be the case when the protein is an enzyme or antibody, or contains a receptor binding site. The separation of proteins under these conditions in which they maintain their native conformation is influenced by both charge and size.

When protein molecules are treated with SDS, the detergent disrupts the secondary, tertiary, and quaternary structure to produce linear polypeptide chains coated with negatively charged SDS molecules. The presence of mercaptoethanol assists in protein denaturation by reducing all disulfide bonds. The detergent binds to hydrophobic regions of the denatured protein chain in a constant ratio of about 1.4 g of SDS per gram of protein. The bound detergent molecules carrying negative charges mask the native charge of the protein. In essence, polypeptide chains of a constant charge/mass ratio and uniform shape are produced. The electrophoretic mobility of the SDS-protein complexes is influenced primarily by molecular size: the larger molecules are retarded by the molecular sieving effect of the gel, and the smaller molecules have greater mobility. Empirical measurements have shown a linear relationship between the log molecular weight and the electrophoretic mobility (Figure 6.6).

In practice, a protein of unknown molecular weight and subunit structure is treated with 1% SDS and 0.1 *M* mercaptoethanol in electrophoresis buffer. A standard mixture of proteins with known molecular weights must also be subjected to electrophoresis under the same conditions. Two broad sets of standards are commercially available, one for low-molecular-weight proteins (molecular weight range 14,000 to 100,000) and one for high-molecular-weight proteins (45,000 to 200,000). Figure 6.7 shows a stained gel after electrophoresis of a standard protein mixture. (For details on stains and staining, see Section C.) After electrophoresis and dye staining, mobilities are measured and molecular weights determined graphically.

SDS-PAGE is valuable for estimating the molecular weight of protein subunits. This modification of gel electrophoresis finds its greatest use in characterizing the sizes and different types of subunits in oligomeric proteins. SDS-PAGE

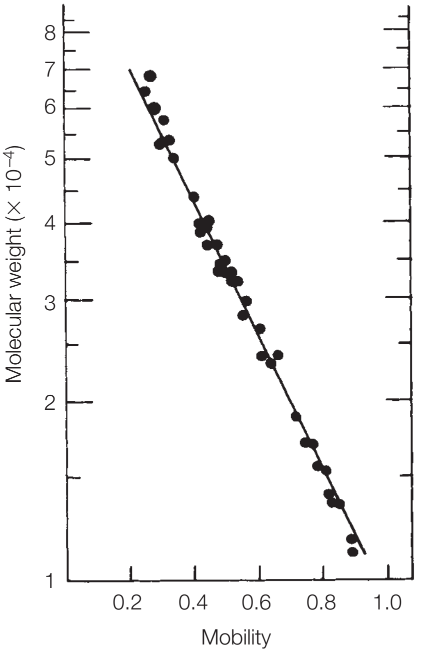

FIGURE 6.6 Graph illustrating the linear relationship between electrophoretic mobility of a protein and the log of its molecular weight. Thirty-seven different polypeptide chains with a molecular weight of 11,000 to 70,000 are shown. Gels were run in the presence of SDS. *From K. Weber and M. Osborn, J. Biol. Chem.* ***244,*** *4406 (1969). By permission of the copyright owner, the American Society for Biochemistry and Molecular Biology, Inc.*

is limited to a molecular weight range of 10,000 to 200,000. Gels of less than 2.5% acrylamide must be used for determining molecular weights above 200,000, but these gels do not set well and are very fragile because of minimal cross-linking. A modification using gels of agarose-acrylamide mixtures allows the measurement of molecular weights above 200,000.

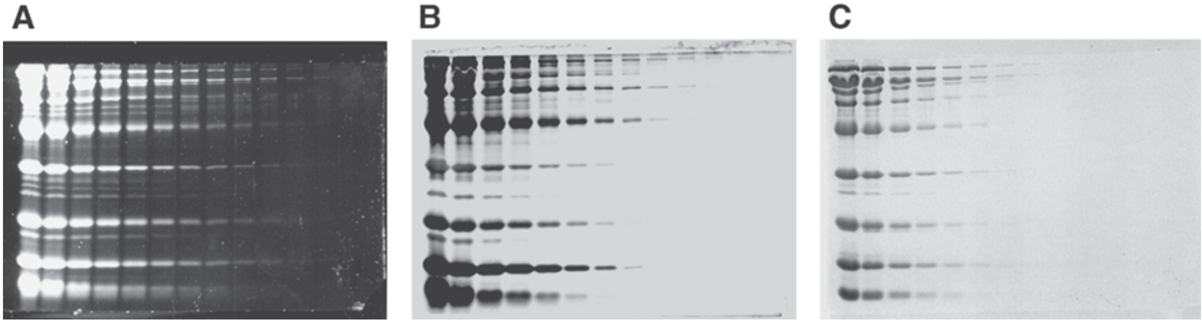

FIGURE 6.7 A comparison of the sensitivities achieved with three different protein stains. Identical SDS-polyacrylamide gels were stained with **A** SYPRO Red protein gel stain; **B** Silver stain; **C** Coomassie brilliant blue dye. *Courtesy of Molecular Probes; www.probes.com, a part of Invitrogen Corporation;* www.invitrogen.com.

STUDY EXERCISE 6.2 SDS-PAGE

You have just completed an experiment using SDS-PAGE to study the subunit structure of ferritin (iron storage protein). Ferritin is composed of two types of subunits, H and L, in about equal quantities. You have included in your electrophoresis several standard proteins of known molecular weight. Draw the final stained gel in the form of Figure 6.7.

Protein	MW
Ferritin	500,000
Subunit H	22,000
Subunit L	19,000
α-lactalbumin	14,200
Trypsinogen	24,000
Egg albumin	45,000
Bovine serum albumin	65,000

Nucleic Acid Sequencing Gels

Sequence analysis of nucleic acids is based on the generation of sets of DNA or RNA fragments with common ends and the separation of these oligonucleotide fragments by polyacrylamide electrophoresis. Two methods have been developed for sequencing nucleic acids: (1) the **partial chemical degradation method** of Maxam and Gilbert, which uses four specific chemical reactions to modify bases and cleave phosphodiester bonds, and (2) the **chain termination method** developed by Sanger, which requires a single-stranded DNA template and chain extension processes, followed by chain termination caused by the presence of dideoxynucleoside triphosphates. Both sequencing methods result in nested sets of DNA or RNA fragments that have one common end and chains varying in length. The smallest possible size difference of nucleic acid fragments is one nucleotide. Separation of the nucleic acid fragments by polyacrylamide electrophoresis allows one to "read" the sequence of nucleotides from the gel (see Chapter 9, p. 282).

Using Sequencing Gels

The experimental arrangement is the same as that previously described for PAGE; however, the gel is prepared with many sample wells to accommodate a large number of samples. Sequence gels of 6, 8, 12, and 20% polyacrylamide are routinely used. Gels of 20% may be used to sequence the first 50 to 100 nucleotides of a nucleic acid, and lower percentage gels allow sequencing out to 250 nucleotides. Sequencing gels are large (up to 40 cm), and power supplies must provide more power than for conventional methods. Precast sequencing gels are now commercially available from Stratagene and other suppliers. They have a gel concentration of 5.5%, have 32 sample wells, and will sequence up to 500 nucleotides. Denaturants such as urea and formamide are required to prevent renaturing of the nucleic acid fragments during electrophoresis. For detection, nucleic acid chains for sequencing must be end-labeled with ^{32}P, ^{35}S, or a fluorescent tag. ^{32}P- and ^{35}S-labeled nucleic acids on gels are detected by autoradiography (see later). Nucleic acids end-labeled with fluorescent molecules are detected by fluorimeter scanning of the gels. Many researchers working on the large and expensive Human Genome Project[1] have generated huge amounts of DNA sequence data. Much of this information is stored in computer data banks for use by researchers around the world.

Agarose Gel Electrophoresis

The electrophoretic techniques discussed up to this point are useful for analyzing proteins and small fragments of nucleic acids up to 350,000 daltons (500 bp) in molecular size; however, the small pore sizes in the gel are not appropriate for analysis of large nucleic acid fragments or intact DNA molecules. The standard method used to characterize RNA and DNA in the range 200 to 50,000 base pairs (50 kilobases) is electrophoresis with agarose as the support medium.

Agarose, a product extracted from seaweed, is a linear polymer of galactopyranose derivatives. Gels are prepared by dissolving agarose in warm electrophoresis buffer. After cooling the gel mixture to 50°C, the agarose solution is poured between glass plates as described for polyacrylamide. Gels with less than 0.5% agarose are rather fragile and must be used in a horizontal arrangement (Figure 6.8). The sample to be separated is placed in a sample well made with a comb, and voltage is applied until separation is complete. Precast agarose gels of all shapes, sizes, and percent composition are commercially available.

Nucleic acids can be visualized on the slab gel after separation by soaking in a solution of ethidium bromide, a dye that displays enhanced fluorescence when intercalated between stacked nucleic acid bases. Ethidium bromide may be added directly to the agarose solution before gel formation. This method allows monitoring of nucleic acids during electrophoresis. Irradiation of ethidium bromide–treated gels by UV light results in orange-red bands where nucleic acids are present. Nucleic acids may also be stained with the new, fluorescent SYBR dyes, which are less toxic

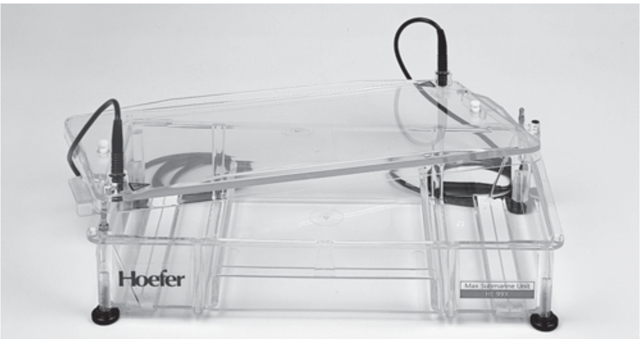

FIGURE 6.8 An apparatus for horizontal slab gel electrophoresis of nucleic acids. *Courtesy of Hoefer, Inc.* www.hoeferinc.com.

---

[1] The Human Genome Project was a federal, government-sponsored program to sequence all DNA in human chromosomes. The project was completed in 2001.

TABLE 6.3 Effective Range of Separation of DNA by Agarose

Agarose (% w/v)	Effective Range (kb)
0.3	5–50
0.5	2–25
0.7	0.8–10
1.2	0.4–5
1.5	0.2–3
2.0	0.1–2

than the mutagenic ethidium bromide. The newest, least toxic, and most sensitive stains for nucleic acids are GelRed and GelGreen from Biotium, Inc. (See Section C.)

The mobility of nucleic acids in agarose gels is influenced by the agarose concentration and the molecular size and molecular conformation of the nucleic acid. Agarose concentrations of 0.3 to 2.0% are most effective for nucleic acid separation (Table 6.3). Adding ethidium bromide to the gel can retard DNA mobility. Like proteins, nucleic acids migrate at a rate that is inversely proportional to the logarithm of their molecular weights; hence, molecular weights can be estimated from electrophoresis results using standard nucleic acids or DNA fragments of known molecular weight. The DNA conformations most frequently encountered are superhelical circular (form I), nicked circular (form II), and linear (form III). The small, compact, supercoiled form I molecules usually have the greatest mobility, followed by the rodlike, linear form III molecules. The extended, circular form II molecules migrate more slowly. The relative electrophoretic mobility of the three forms of DNA, however, depends on experimental conditions such as agarose concentration and ionic strength.

The versatility of agarose gels is obvious when one reviews their many applications in nucleic acid analysis. The rapid advances in our understanding of nucleic acid structure and function in recent years are due primarily to the development of agarose gel electrophoresis as an analytical tool. Two of the many applications of agarose gel electrophoresis will be described here.

Analysis of DNA Fragments after Digestion by Restriction Endonucleases

Restriction endonucleases are enzymes that recognize a specific base sequence in double-stranded DNA and catalyze cleavage (hydrolysis of phosphodiester bonds) in or near that specific region (see Chapter 10, Section B, p. 297). Many viral, bacterial, or animal DNA molecules are substrates for the enzymes. When each type of DNA is treated with a restriction endonuclease, a specific number of DNA fragments is produced. The base sequence recognized by the enzyme occurs only a few times in any particular DNA molecule; therefore, the smaller the DNA molecule, the fewer specific cleavage sites there are. Viral or phage DNA, for example, is cleaved into about 50 fragments depending on the enzyme used, whereas larger bacterial or animal DNA may be cleaved into hundreds or thousands of fragments. Smaller DNA molecules, upon cleavage with a particular enzyme, will produce a limited set of fragments. It is unlikely that this set of fragments will be the same for any two different DNA molecules, so the fragmentation pattern can be considered a "fingerprint" of the DNA substrate. The **restriction pattern** is produced by electrophoresis of the cleavage reaction mixture through agarose gels, followed by staining with ethidium bromide or SYBR dyes (Figure 6.9). The separation of the fragments is based on molecular size, with large fragments remaining near the origin and smaller fragments migrating farther down the gel. In addition to characterization of DNA

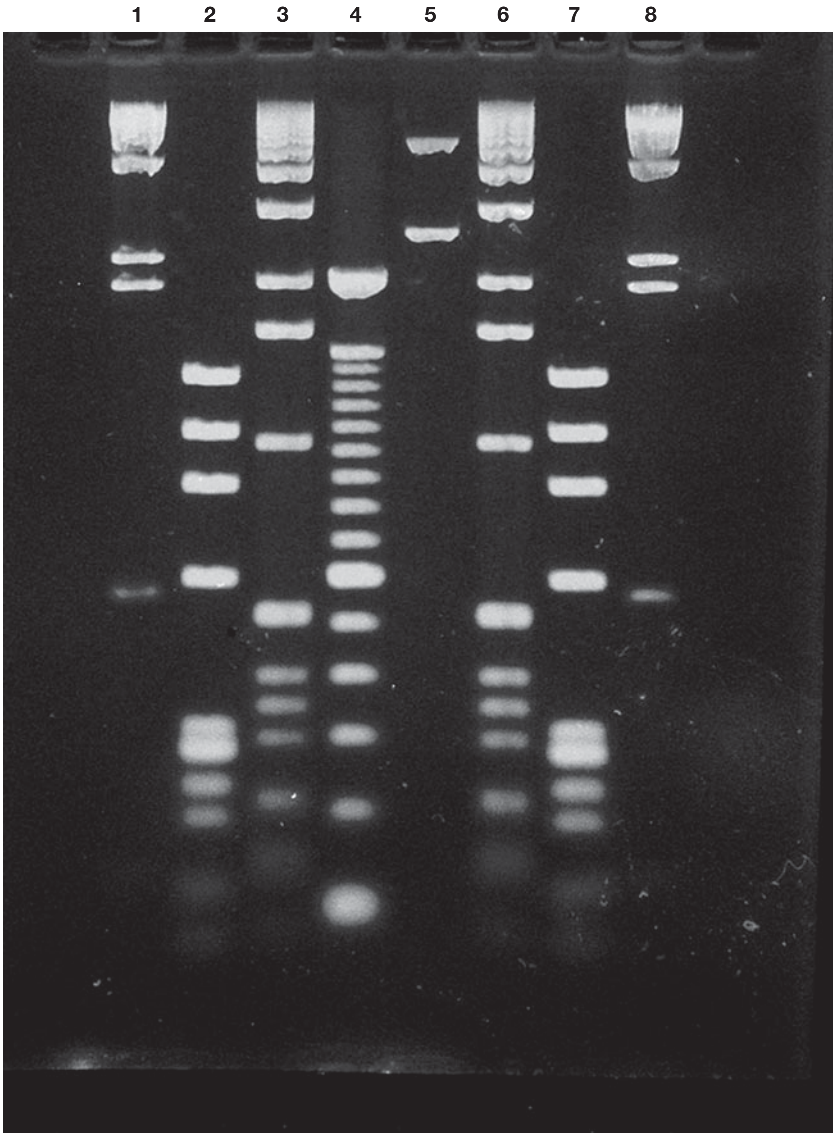

FIGURE 6.9 Restriction patterns produced by agarose electrophoresis of DNA fragments after endonuclease action. DNA molecular weight ladders have been electrophoresed on a 1% agarose gel and then stained with SYBR Green I nucleic acid gel stain. Lanes **1** and **8** contain *Hin*dIII-cut λ DNA; lanes **2** and **7,** *Hae*III-cut Φ X174 RF DNA; lanes **3** and **6,** 1 kilobase pair DNA ladder; lane **4,** 100 base pair DNA ladder; lane **5,** *Eco*R-I-cut pUC19 DNA mixed with *Pst*I-cut Φ X174 RF DNA.

structure, endonuclease digestion coupled with agarose gel electrophoresis is a valuable tool for plasmid mapping and DNA recombination experiments.

Characterization of Superhelical Structure of DNA

The structure of plasmid, viral, and bacterial DNA is often closed circular with negative superhelical turns. It is possible under various experimental conditions to induce reversible changes in the conformation of DNA. The intercalating dye ethidium bromide causes an unwinding of supercoiled DNA that affects its electrophoretic mobility. Electrophoresis of DNA on agarose in the presence of increasing concentrations of ethidium bromide provides an unambiguous method for distinguishing between closed circular and other DNA conformations.

Closed circular, negatively supercoiled DNA (form I) usually has the greatest electrophoretic mobility of all DNA forms because supercoiled DNA molecules tend to be compact. If ethidium bromide is added to form I DNA, the dye intercalates between the stacked DNA bases, causing unwinding of some of the negative supercoils. As the concentration of ethidium bromide is increased, more and more of the negative supercoils are removed until no more are present in the DNA. The conformational change of the DNA supercoil can be monitored by electrophoresis because the mobility decreases with each unwinding step. With increasing concentration of ethidium bromide, the negative supercoils are progressively unwound and the electrophoretic mobility decreases to a minimum. This minimum represents the free dye concentration necessary to remove all negative supercoils. (The free dye concentration at this minimum has been shown to be related to the superhelix density, which is a measure of the extent of supercoiling in a DNA molecule.) The circular DNA at this point is equivalent to the “relaxed” form. If more ethidium bromide is added to the relaxed DNA, positive superhelical turns are induced in the structure and the electrophoretic mobility increases. Forms II and III DNA, under the same conditions of increasing ethidium bromide concentration, show a gradual decrease in electrophoretic mobility throughout the entire concentration range.

Agarose gel electrophoresis is able to resolve topoisomers of native, covalently closed, circular DNA that differ only in their degree of supercoiling. This technique has proved useful in the analysis and characterization of enzymes that catalyze changes in the conformation or topology of native DNA. These enzymes, called **topoisomerases,** have been isolated from bacterial and mammalian cells. They change DNA conformations by catalyzing nicking and closing of phosphodiester bonds in circular duplex DNA. Agarose gel electrophoresis is an ideal method for identifying and assaying topoisomerases because the intermediate DNA molecules can be resolved on the basis of the extent of supercoiling. Topoisomerases may be assayed by incubating native DNA with an enzyme preparation, removing aliquots after various periods of time, and subjecting them to electrophoresis on an agarose gel with standard supercoiled and relaxed DNA.

Pulsed Field Gel Electrophoresis (PFGE)

Conventional agarose gel electrophoresis is limited in use for the separation of nucleic acid fragments smaller than 50,000 bp (50 kb). In practice, that limit is closer to 20,000 to 30,000 bp if high resolution is desired. Since chromosomal

DNA from most organisms contains thousands and even millions of base pairs, the DNA must be cleaved by restriction enzymes before analysis by standard electrophoresis. In the early 1980s, Schwartz and Cantor at Columbia University discovered that large molecules of DNA (yeast chromosomes, 200–3000 kb) could be separated by **pulsed field gel electrophoresis (PFGE).** There is one major distinction between standard gel electrophoresis and PFGE. In PFGE, the electric field is not constant as in the standard method, but is changed repeatedly (pulsed) in direction and strength during the separation (Figure 6.10). The physical mechanism for separation of the large DNA molecules as they move through the gel under these conditions is not yet well understood. An early explanation was that the electrical pulses abruptly perturbed the conformation of the DNA molecules. They would be oriented by the influence of the electric field coming from one direction and then reoriented as a new electric field at a different angle to the first was turned on. According to this explanation, it takes longer for larger molecules to reorient, so smaller fragments respond faster to the new pulse and move faster. More recent experiments on dyed DNA moving in gels have shown that conformational changes of the DNA are not abrupt, but more gradual, in response to the electrical pulse, and DNA molecules tend to "slither" through the gel matrix. In addition, it has been discovered that the gel becomes more fluid during electrical pulsing.

Applications of PFGE

Even though our theoretical understanding of PFGE is lacking, practical applications and experimental advances are expanding rapidly. The availability of PFGE has sparked changes in DNA research. New methods for isolating intact DNA molecules have been developed. Because of mechanical breakage, the average size of DNA isolated from cells in the presence of lysozyme, detergent, and EDTA is about 400–500 kb (see Chapter 9, Section B, p. 275). Intact chromosomal DNA can be isolated by embedding cells in an agarose matrix and disrupting the cells with detergents and enzymes. Slices or "plugs" of the agarose with intact

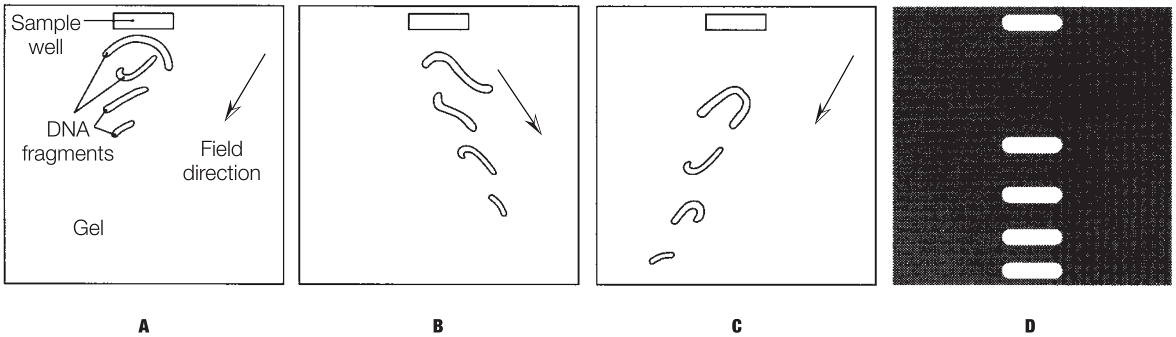

FIGURE 6.10 Pulsed field gel electrophoresis. Generalized PFGE separation of four DNA fragments of different sizes in one lane. **A** DNA molecules of various shapes and configurations move toward the positive field. **B** The new field orientation pulls the DNA in a different direction, realigning the molecules. **C** The field returns to the original configuration. **D** The bands show the final position of a large collection of the molecules. *Reprinted with permission from the Journal of NIH Research* ***1****, 115 (1999). Illustration by Terese Winslow.*

DNA are then placed on the gel for PFGE analysis. Newly discovered restriction endonucleases that cut DNA only rarely can now be used to subdivide chromosome-sized DNA. Two important endonucleases with eight-base recognition sites are *Not* I and *Sfi* I.

There are also many instrumental advances that allow changes in the experimental design of PFGE. Some of the variables that can be changed for each experiment are voltage, pulse length, number of electrodes, relative angle of electrodes, gel box design, temperature, agarose concentration, buffer pH, and time of electrophoresis.

Like all laboratory techniques, PFGE has its disadvantages and problems. Long periods of electrophoresis are often required for good resolution, and migration of fragments is extremely dependent on experimental conditions. Therefore, it is difficult to compare gels even when they are run under similar conditions. In spite of these shortcomings, PFGE will continue to advance as a significant tool for the characterization of very large molecules. The technique, which was widely used in the Human Genome Project, will greatly increase our understanding of chromosome structure and function.

Isoelectric Focusing of Proteins

Another important and effective use of electrophoresis for the analysis of proteins is **isoelectric focusing (IEF),** which examines electrophoretic mobility as a function of pH. The net charge on a protein is pH dependent. Proteins below their **isoelectric pH** (pH_I, or sometimes P_I; the pH at which they have zero net charge) are positively charged and migrate in a medium of fixed pH toward the negatively charged cathode. At a pH above its isoelectric point, a protein is deprotonated and negatively charged and migrates toward the anode. If the pH of the electrophoretic medium is identical to the pH_I of a protein, the protein has a net charge of zero and does not migrate toward either electrode. Theoretically, it should be possible to separate protein molecules and to estimate the pH_I of a protein by investigating the electrophoretic mobility in a series of separate experiments in which the pH of the medium is changed. The pH at which there is no protein migration should coincide with the pH_I of the protein. Because such a repetitive series of electrophoresis runs is a rather tedious and time-consuming way to determine the pH_I, IEF has evolved as an alternative method for performing a single electrophoresis run in a medium of gradually changing pH (i.e., a pH gradient).

Separating Proteins by IEF

Figure 6.11 illustrates the construction and operation of an IEF pH gradient. An acid, usually phosphoric, is placed at the cathode; a base, such as triethanolamine, is placed at the anode. Between the electrodes is a medium in which the pH gradually increases from 2 to 10. The pH gradient can be formed before electrophoresis is conducted or formed during the course of electrophoresis. The pH gradient can be either broad (pH 2–10) for separating several proteins of widely ranging pH_I values, or narrow (pH 7–8) for precise determination of the pH_I of a single protein. P in Figure 6.11 represents different molecules of

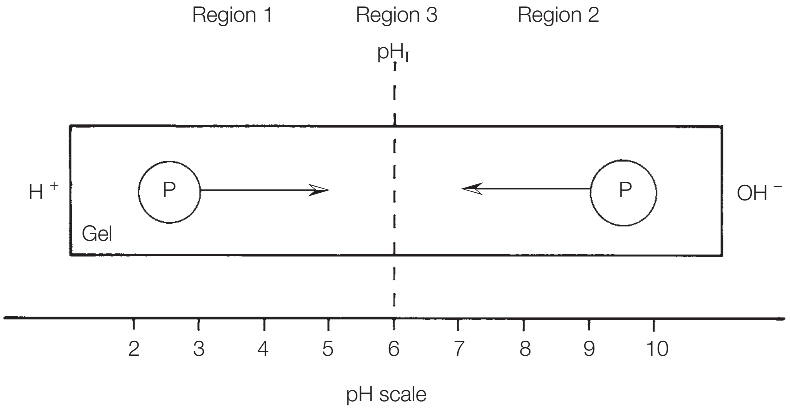

FIGURE 6.11 Illustration of isoelectric focusing. See text for details.

the same protein in two different regions of the pH gradient. Assuming that the pH in region 1 is less than the pH$_I$ of the protein and the pH in region 2 is greater than the pH$_I$ of the protein, molecules of P in region 1 will be positively charged and will migrate in an applied electric field toward the cathode. As P migrates, it will encounter an increasing pH, which will influence its net charge. As it migrates *up* the pH gradient, P will become increasingly deprotonated and its net charge will decrease toward zero. When P reaches a region where its net charge is zero (region 3), it will stop migrating. The pH in this region of the electrophoretic medium will coincide with the pH$_I$ of the protein and can be measured with a surface microelectrode, or the position of the protein can be compared to that of a calibration set of proteins of known pH$_I$ values. P molecules in region 2 will be negatively charged and will migrate toward the anode. In this case, the net charge on P molecules will gradually decrease to zero as P moves *down* the pH gradient, and P molecules originally in region 2 will approach region 3 and come to rest. The P molecules move in opposite directions, but the final outcome of IEF is that P molecules located anywhere in the gradient will migrate toward the region corresponding to their isoelectric point and will eventually come to rest in a sharp band; that is, they will "focus" at a point corresponding to their pH$_I$.

Since different protein molecules in mixtures have different pH$_I$ values, it is possible to use IEF to separate proteins. In addition, the pH$_I$ of each protein in the mixture can be determined by measuring the pH of the region where the protein is focused.

Practical Aspects of IEF

The pH gradient is prepared in a horizontal glass tube or slab. Special precautions must be taken so that the pH gradient remains stable and is not disrupted by diffusion or convective mixing during the electrophoresis experiment. The most common stabilizing technique is to form the gradient in a polyacrylamide, agarose, or dextran gel. The pH gradient is formed in the gel by electrophoresis of synthetic polyelectrolytes, called **ampholytes,** which migrate to the region of their pH_I values just as proteins do and establish a pH gradient that is stable for the duration of the IEF run. Ampholytes are low-molecular-weight polymers that have a wide range of isoelectric points because of their numerous amino and carboxyl or sulfonic acid groups. The polymer mixtures are available in specific pH ranges (pH 5–7, 6–8, 3.5–10, etc.). It is critical to select the appropriate pH range for the ampholyte so that the proteins to be studied have pH_I values in that range. The best resolution is, of course, achieved with an ampholyte mixture over a small pH range (about two units) encompassing the pH_I of the sample proteins. If the pH_I values for the proteins under study are unknown, an ampholyte of wide pH range (pH 3–10) should be used first and then a narrower pH range selected for use.

The gel medium is prepared as previously described except that the appropriate ampholyte is mixed prior to polymerization. The gel mixture is poured into the desired form (column tubes, horizontal slabs, etc.) and allowed to set. Immediately after casting of the gel, the pH is constant throughout the medium, but application of voltage will induce migration of ampholyte molecules to form the pH gradient. The standard gel for proteins with molecular sizes up to 100,000 daltons is 7.5% polyacrylamide; however, if larger proteins are of interest, gels with larger pore sizes must be prepared. Such gels can be prepared with a lower concentration of acrylamide (about 2%) and 0.5 to 1% agarose to add strength. Precast gels for isoelectric focusing are also commercially available.

The protein sample can be loaded on the gel in either of two ways. A concentrated, salt-free sample can be layered on top of the gel as previously described for ordinary gel electrophoresis. Alternatively, the protein can be added directly to the gel preparation, resulting in an even distribution of protein throughout the medium. The protein molecules move more slowly than the low-molecular-weight ampholyte molecules, so the pH gradient is established before significant migration of the proteins occurs. Very small protein samples can be separated by IEF. For analytical purposes, 10 to 50 μg is a typical sample size. Larger sample sizes (up to 20 mg) can be used for preparative purposes.

Two-Dimensional Electrophoresis (2-DE) of Proteins

The separation of proteins by IEF is based on charge, whereas SDS-PAGE separates molecules based on molecular size. A combination of the two methods leads to enhanced resolution of complex protein mixtures. Such an experiment was first reported by O'Farrell and the combined method has since become a routine and powerful separatory technique. Figure 6.12 shows the results of O'Farrell's analysis of total *Escherichia coli* protein. The sample was first separated in one dimension by IEF. The sample gel was then transferred to an SDS-PAGE slab and electrophoresis was continued in the second dimension. At least 1000 discrete protein spots are visible.

The technique of 2-DE is a powerful tool for researchers in the new field of **proteomics,** whose goal is the systematic study of the structure, interactions, and biological function of the proteins expressed by the genome (proteome) of an organism. We see that the 1000 or so proteins in the *E. coli* cell are well resolved by 2-DE (Figure 6.12), but what about the numerous proteins expressed in higher organisms like humans? It is currently estimated from data generated by the

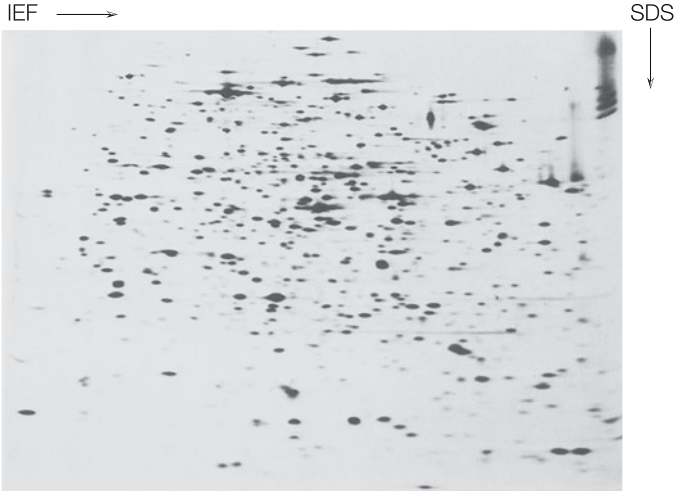

FIGURE 6.12 Two-dimensional electrophoresis of total *E. coli* proteins. *Photo courtesy of Dr. P. O'Farrell.*

Human Genome Project that humans have about 20,000–25,000 genes. This number does not relate directly to the number of proteins because many of the genes are alternately spliced and because of numerous post-translational modification processes. It is now estimated that there may be as many as one million different protein products in the human. This is a daunting number even for the high resolution displayed in 2-DE. It is not realistic to imagine that 2-DE techniques will ever be able to analyze this many proteins, not only because of the complexity, but also because the technique is tedious and time-consuming, and the gels are difficult to analyze. However, when 2-DE data are combined and analyzed by modern computer software, valuable on-line databases on protein expression may be constructed. In addition, when proteins separated by 2-DE are further analyzed by mass spectrometry (see Chapter 7, Section D, p. 230), partial protein sequence data can be obtained.

The 2-DE technique is also of value in developmental biochemistry, where the increase or decrease in intensity of a spot representing a specific protein may be monitored as a function of cell growth.

Capillary Electrophoresis (CE)

Capillary electrophoresis is a new technique that combines the high resolving power of electrophoresis with the speed, versatility, and automation of high-performance liquid chromatography (HPLC). It offers the ability to analyze very small samples (5–10 nL) utilizing up to 1 million theoretical plates to achieve high resolution and sensitivity to the attomole level (10^{-18} mole). It will become a widely used technique in the analysis of amino acids, peptides, proteins, nucleic acids, and pharmaceuticals.

CE in Practice

A general experimental design is diagrammed in Figure 6.13. The equipment consists of a power supply, two buffer reservoirs, a buffer-filled capillary tube, and an on-line detector. Platinum electrodes connected to the power supply are

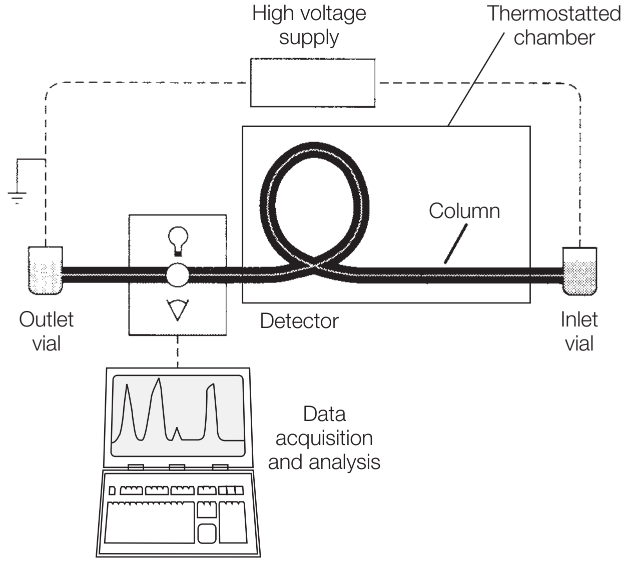

FIGURE 6.13 Experimental setup for capillary electrophoresis. *Courtesy of Bio-Rad Laboratories, Life Science Research Group, Hercules, CA.*

immersed in each buffer reservoir. A high voltage is applied along the capillary, and a small plug of sample solution is injected into one end of the capillary. Components in the solution migrate along the length of the capillary under the influence of the electric field. Molecules are detected as they exit from the opposite end of the capillary. The detection method used depends on the type of molecules separated, but the most common are UV-VIS fixed-wavelength detectors and diode-array detectors (see Chapter 7, pp. 210–211). The capillaries used are flexible, fused, silica tubes of $5-100\ \mu$m i.d. and 25–100 cm length that may or may not be filled with chromatographic matrix.

A major advantage of capillary electrophoresis is that many analytical experimental designs are possible, just as in the case of HPLC. In HPLC, a wide range of molecules can be separated by changing the column support (see Chapter 5, Section F, p. 140). In CE, the capillary tube may be coated or filled with a variety of materials. For separation of small, charged molecules, bare silica or polyimide-coated capillaries are often used. If separation by molecular sieving is desired, the tube is filled with polyacrylamide or SDS-polyacrylamide. If the capillary is filled with electrolyte and an ampholyte pH gradient, isoelectric focusing experiments on proteins may be done. We can expect to see numerous applications of CE in all aspects of biochemistry and molecular biology. New applications will include DNA sequencing, analysis of single cells, and separations of neutral molecules. (For example, see nucleotide separation in Figure 6.14.)

Immunoelectrophoresis (IE)

In immunoelectrophoresis, two sequential procedures are applied to the analysis of complex protein mixtures: (1) separation of the protein mixture by agarose gel electrophoresis, followed by (2) interaction with specific antibodies to examine the antigenic properties of the separated proteins.

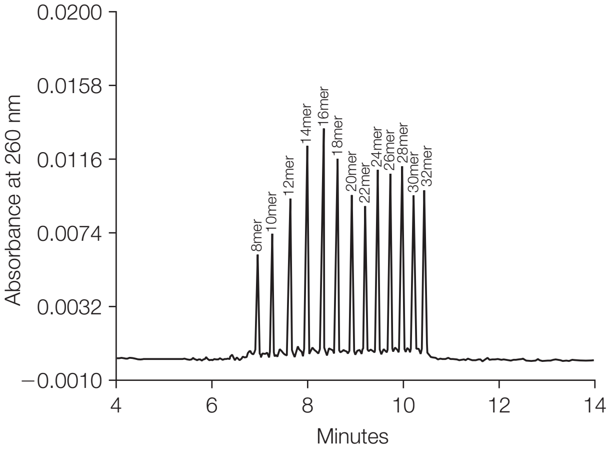

8 mer	**GACTGACT**
10 mer	**GACTGACTGT**
12 mer	**GACTGACTGACT**
14 mer	**GACTGACTGACTGT**
16 mer	**GACTGACTGACTGACT**
18 mer	**GACTGACTGACTGACTGT**
20 mer	**GACTGACTGACTGACTGACT**
22 mer	**GACTGACTGACTGACTGACTGT**
24 mer	**GACTGACTGACTGACTGACTGACT**
26 mer	**GACTGACTGACTGACTGACTGACTGT**
28 mer	**GACTGACTGACTGACTGACTGACTGACT**
30 mer	**GACTGACTGACTGACTGACTGACTGACTGT**
32 mer	**GACTGACTGACTGACTGACTGACTGACTGACT**

FIGURE 6.14 Separation of a series of oligonucleotides using capillary electrophoresis. The compounds are separated on a capillary coated on the inside with polyacryloylamino-ethoxyethanol. The structures of the standard oligonucleotides are shown in the lower diagram. Adapted from Bio-Rad Laboratories, Inc., www.bio-rad.com.

The technique of IE was first reported by Grabar and Williams in 1953 for the separation and immunoanalysis of serum proteins, but it can be applied to the analysis of any purified protein or complex mixture of proteins. In practice (Figure 6.15), a protein mixture is separated by standard electrophoresis in an agarose gel prepared on a small glass plate. This is followed by exposing the separated proteins to a specific antibody preparation. The antibody is added to a trough cut into the gel, as shown in Figure 6.15, and is allowed to diffuse through the gel toward the separated proteins. If the antibody has a specific affinity for one of the proteins, a visible precipitin arc forms. This is an insoluble complex formed at the boundary of antibody and antigen protein. The technique is most useful for the analysis of protein purity, composition, and antigenic properties. The basic IE technique described here allows only qualitative examination of antigenic proteins. If quantitative results in the form of protein antigen concentration are required, the advanced modifications rocket immunoelectrophoresis and two-dimensional (crossed) immunoelectrophoresis may be used.

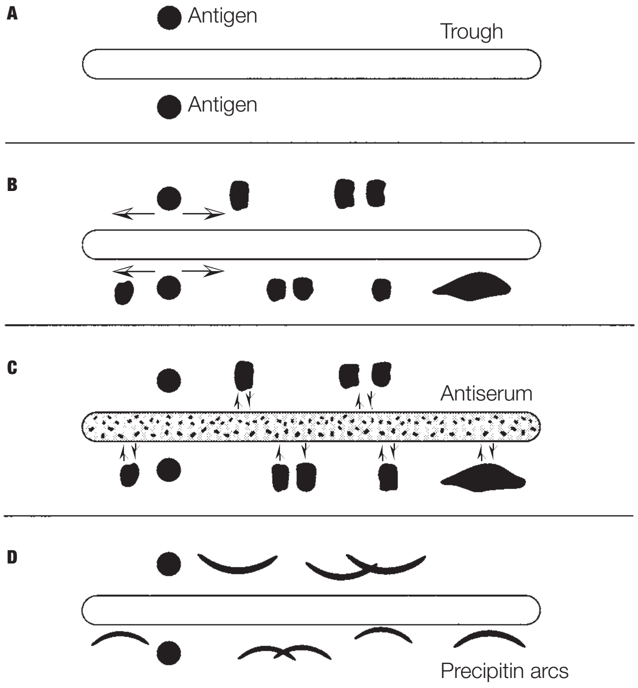

FIGURE 6.15 Immunoelectrophoresis. **A** Antigen is placed in sample wells. **B** Electrophoresis. **C** Antiserum containing antibody is placed in trough. **D** Insoluble antigen–antibody complexes form precipitin arcs.

C. PRACTICAL ASPECTS OF ELECTROPHORESIS

Instrumentation

The basic components required for electrophoresis are a power supply and an electrophoresis chamber (gel box). A power supply that provides a constant current is suitable for most conventional electrophoresis experiments. Power supplies that generate both constant voltage (up to 4 kV) and constant current (up to 200 mA) are commercially available. In order to implement the many modifications of electrophoresis, such versatile power supplies are essential in a research laboratory. If isoelectric focusing experiments are planned, a power supply that furnishes a constant voltage is necessary. DNA sequencing experiments require power supplies capable of generating 2–4 kV. Because of the high power requirements of these experiments, the glass plates sandwiching the gel must be covered with conductive aluminum plates to dissipate heat and prevent gel melting and glass plate breakage. CE also requires the use of high voltage, which generates heat as well. However, the heat is quickly dissipated through the thin walls of the capillary tubing. Pulsed field electrophoresis experiments have special requirements, including a high-voltage power supply, electrical switching devices for control of the field, equipment for temperature control, and a specially designed gel box with an array of electrodes. All types of electrophoresis chambers for both horizontal and vertical placements of gels are available from commercial suppliers.

Reagents

High-quality, electrophoresis grade chemicals must be used, since impurities may influence both the gel polymerization process and electrophoretic mobility. The reagents used for gel formation should be stored in a refrigerator.

Caution

Acrylamide, *N,N,N′N′*-tetramethylethylenediamine, *N,N′*-methylene-bis-acrylamide, and ammonium persulfate are toxic and must be used with care. Acrylamide is a neurotoxin, a carcinogenic agent, and a potent skin irritant, so gloves and a mask must be worn while handling it in the unpolymerized form.

Buffers appropriate for electrophoresis gels include Tris-glycine, Tris-acetate, Tris-phosphate, and Tris-borate at concentrations of about 0.05 *M* (Chapter 3, Section A, p. 61).

Several advances in gel electrophoresis have recently been made that have streamlined and improved many of the established techniques. Some of those improvements will be outlined here.

1. ***Precast Gels*** As previously mentioned, precast gels of polyacrylamide and agarose are now commercially available. A wide variety of gel sizes, types, configurations, and compositions may be purchased. Costs are reasonable, beginning at about $15 for an 8 × 10 cm single percentage polyacrylamide gel. Precast gels not only offer convenience, but also increase safety, save time, and provide more reproducible electrophoretic runs. Precast gels with dye/stain already applied are also available (prestaining).
2. ***Bufferless Precast Gels*** One of the major inconveniences of running gels is the necessity of having liquid buffer reservoirs to saturate gels during electrophoresis. There is always the chance that reservoirs will leak or that solutions perhaps containing acrylamide or ethidium bromide will be spilled. Precast gels (agarose and acrylamide) are now available in dry, plastic cassettes. They do not require liquid buffers because the gels contain ion-exchange matrices, which sustain the electric field.
3. ***Reusable Precast Gels*** Precast agarose gels are now available that may be recycled. After a run, the DNA samples on the gel are removed by reversing the direction of the electric field. The gels are then reloaded with new samples and reused.

Staining and Detecting Electrophoresis Bands

Tracking Dyes

During the electrophoretic process, it is important to know when to stop applying the voltage or current. If the process is run for too long, the desired components may pass entirely through the medium and into the buffer; if too short a period is used, the components may not be completely resolved. It is common practice to add a "tracking dye," usually bromphenol blue and/or xylene cyanol, to the sample mixture. These dyes, which are small and anionic, move rapidly through the gel ahead of most proteins or nucleic acids. After electrophoresis, bands have a tendency to widen by diffusion. Because this broadening may decrease resolution, gels should be analyzed as soon as possible after the power supply has been turned off. In the case of proteins, gels can be treated with an agent that "fixes" the proteins in their final positions, a process that is often combined with staining.

Staining Proteins with Coomassie Blue

Reagent dyes that are suitable for visualization of biomolecules after electrophoresis were mentioned earlier. The most commonly used general stain for proteins is **Coomassie Brilliant Blue.** This dye can be used as a 0.25% aqueous solution. It is followed by destaining (removing excess background dye) by repeated washing of the gel with 7% acetic acid. Alternatively, gels may be stained by soaking in 0.25% dye in H_2O, methanol, and acetic acid (5:5:1), followed by repeated washings with the same solvent. The most time-consuming procedure in the Coomassie Blue visualization process is destaining, which often requires days of washing. Rapid destaining of gels may be brought about electrophoretically. The gel, after soaking in the stain, is subjected to electrophoresis again, using a buffer of higher concentration to remove excess stain.

The search for more rapid and sensitive methods of protein detection after electrophoresis has led to the development of fluorescent staining techniques. Two commonly used fluorescent reagents are fluorescamine and anilinonaphthalene sulfonate. New dyes based on silver salts (silver diamine or silver–tungstosilicic acid complex) have been developed for protein staining. They are 10 to 100 times more sensitive than Coomassie Blue (Figure 6.7).

A visualization procedure that is specific for a certain biomolecule, such as an enzyme, is often needed. If the enzyme remains in an active form while in the gel, any substrate that produces a colored product could be used to locate the enzyme on the gel. Although it is less desirable for detection, the electrophoresis support medium may be cut into small segments and each part extracted with buffer and analyzed for the presence of the desired component.

Fluorescent SYPRO and CF Dyes

The new fluorescent SYPRO dyes, produced by Molecular Probes, Inc. (www.probes.com), promise to challenge the use of silver staining for protein gel staining. SYPRO Orange stain, SYPRO Ruby stain, and SYPRO Red stain may be used for 1-D and 2-D gels, and their sensitivities are equal to silver staining. However, the SYPRO dyes are much easier and faster to use (they require no fixing or destaining) than both Coomassie Blue or silver. Gels treated with the SYPRO dyes may be viewed using a standard UV light or with a laser scanner.

The new CF dyes produced by Biotium, Inc., have the advantage of excellent brightness, sensitivity, and photostability compared to the SYPRO dyes. They are designed primarily for labeling antibodies (see www.biotium.com).

Viewing Nucleic Acids on Gels

Nucleic acids may be visualized in agarose and polyacrylamide gels using the fluorescent dye ethidium bromide. The gel is soaked in a solution of the dye and washed to remove excess dye. Illumination of the rinsed slab with UV light reveals red-orange stains where nucleic acids are located.

Although ethidium bromide stains both single- and double-stranded nucleic acids, the fluorescence is much greater for double-stranded molecules. The electrophoresis may be performed with the dye incorporated in the gel and buffer, although this does influence mobility. This has the advantage that the gel can be illuminated with UV light during electrophoresis to view the extent of separation. Precast gels are made in plastic cassettes that are UV transparent. The mobility of double-stranded DNA may be reduced 10 to 15% in the presence of ethidium bromide. Destaining of the gel is not necessary because ethidium bromide–DNA complexes have a much greater fluorescent yield than free ethidium bromide, so relatively small amounts of DNA can be detected in the presence of free ethidium bromide. The detection limit for DNA is 10 ng.

Caution

Ethidium bromide must be used with great care, as it is a potent mutagen. Gloves should be worn at all times while using dye solutions or handling gels; however, even this does not ensure complete safety, as ethidium bromide has been shown to leak through some plastic gloves.

The new, fluorescent SYBR dyes, produced by Molecular Probes, Inc., offer several advantages over ethidium bromide. They are far less toxic than the mutagenic ethidium bromide, and some are as much as five times more sensitive. The most widely used SYBR dye for staining DNA gels is SYBR Green I, which is a cyanine dye and has a sensitivity similar to EtBr. The detailed mechanism for SYBR dye binding to DNA is unknown, but it is most likely not base intercalation as is the case for ethidium bromide. Other new dyes now being tested include Blueview, Carolina Blu, and 4′ 6-diamidino-2-phenylindole dihydrochloride hydrate (DAPI). Currently, the safest and most sensitive dyes now available for nucleic acids are the fluorescent GelRed and GelGreen, produced by Biotium. These dyes may be used for either prestaining or poststaining.

Because single-stranded nucleic acids do not stain deeply with ethidium bromide and some fluorescent dyes, alternate techniques must be used for detection, especially when only small amounts of biological material are available for analysis. One of the most sensitive techniques is to use radio-labeled molecules. For nucleic acids, this usually means labeling the 5′ or 3′ end with ^{32}P, a strong β emitter. Bands of labeled nucleic acids on an electrophoresis gel can easily be located by **autoradiography.** For this technique, the electrophoresed slab gel is transferred to heavy chromatography paper. After covering the gel and paper with plastic wrap, they are placed on X-ray film and wrapped in a cassette to avoid external light exposure. This procedure must be done in a darkroom. The gel-film combination is stored at −70°C for exposure. The low temperature maintains the gel in a rigid form and prevents diffusion of gel bands.

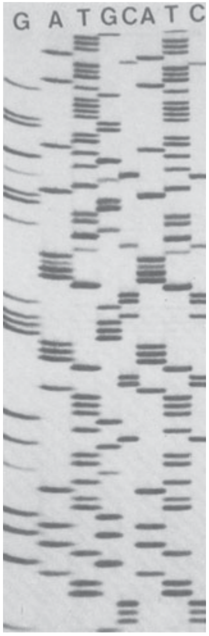

FIGURE 6.16 Autoradiogram of a DNA sequencing gel. *From Zyskind and Bernstein, DNA Laboratory Manual (1999), Academic Press (San Diego, CA).*

The exposure time depends on the amount of radioactivity, but can range from a few minutes to several days. Figure 6.16 shows the autoradiogram of a typical DNA sequencing gel. The autoradiogram also provides a permanent record of the gel for storage and future analysis. The actual gel is very fragile and difficult to store. Also, because of the short half-life of ^{32}P (14 days), an autoradiogram of the gel cannot be obtained in the distant future. Proteins labeled with ^{32}P or ^{125}I can be dealt with in a manner similar to nucleic acids. If an autoradiogram of a gel can be prepared, a permanent record of the experimental data is available.

Because of concerns about the safety of radioisotope use, researchers are developing fluorescent and chemiluminescent methods for detection of small amounts of biomolecules on gels. One attractive approach is to label biomolecules before analysis with the coenzyme biotin. Biotin forms a strong complex with enzyme-linked streptavidin. Some dynamic property of the enzyme is then measured to locate the biotin-labeled biomolecule on the gel. These new methods approach the sensitivity of methods involving radio-labeled molecules, and rapid advances are being made.

Protein and Nucleic Acid Blotting

Only a minute amount of protein or nucleic acid is present in bands on electropherograms. In spite of this, there is often a need to extract the desired biomolecule from the gel for further investigation. This sometimes involves the tedious and cumbersome process of crushing slices of the gel in a buffer to release the trapped proteins or nucleic acids. Techniques are now available for removing nucleic acids and proteins from gels and characterizing them using probes to detect certain structural features or functions. After electrophoresis, the desired biomolecules are transferred or "blotted" out of the gel onto a nitrocellulose filter or nylon membrane. The desired biomolecule is now accessible on the filter for further analysis. The first blotting technique was reported by E. Southern in 1975. Using labeled complementary DNA probes, he searched for certain nucleotide sequences among DNA molecules blotted from the gel. This technique of detecting DNA-DNA hybridization is called **Southern blotting.** The general blotting technique has now been extended to the transfer and detection of specific RNA with labeled complementary DNA probes **(Northern blotting)** and the transfer and detection of proteins that react with specific antibodies **(Western blotting).** In practice, the electropherogram is alkali treated, neutralized, and placed in contact with the filter or nylon membrane. A buffer is used to facilitate the transfer. Figure 6.17 shows the setup for a blotting experiment. The location

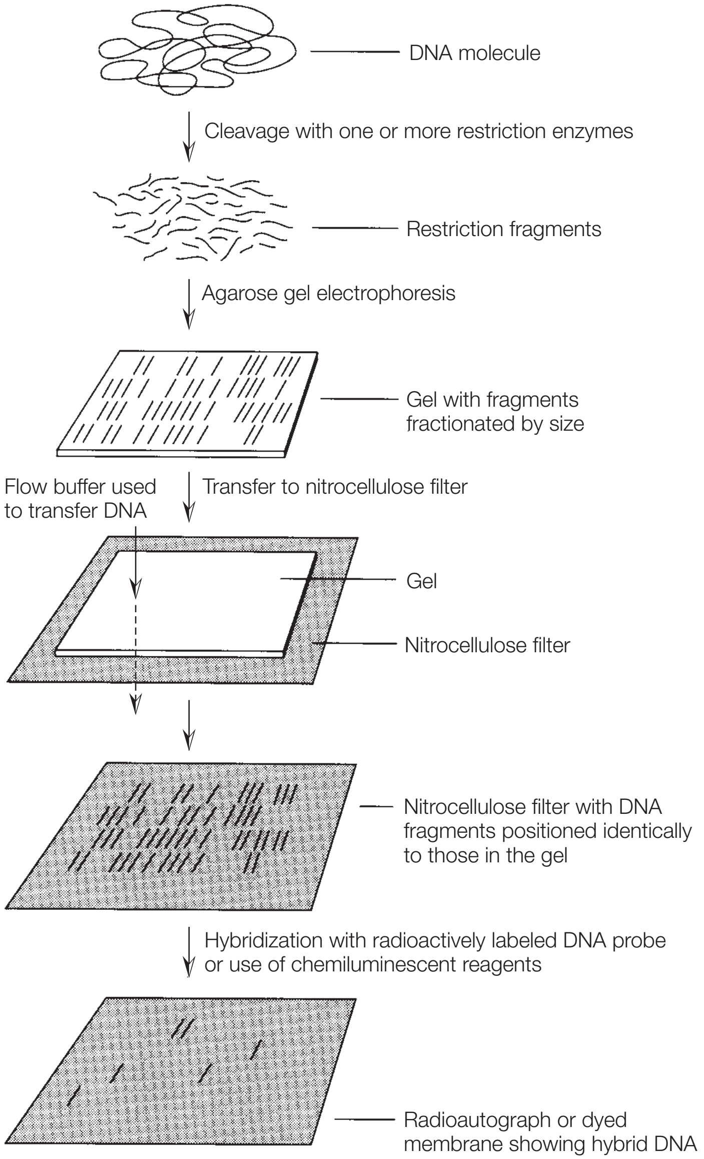
DNA molecule
Cleavage with one or more restriction enzymes
Restriction fragments
Agarose gel electrophoresis
Gel with fragments fractionated by size
Flow buffer used to transfer DNA
Transfer to nitrocellulose filter
Gel
Nitrocellulose filter
Nitrocellulose filter with DNA fragments positioned identically to those in the gel
Hybridization with radioactively labeled DNA probe or use of chemiluminescent reagents
Radioautograph or dyed membrane showing hybrid DNA

FIGURE 6.17
Diagram of a Southern blotting experiment.

of the desired nucleic acid or protein is then detected by incubation of the membrane with a radio-labeled probe and autoradiography, by use of a biotinylated probe, or by linkage to an enzyme-catalyzed reaction that generates a color.

Blotting techniques have many applications, including mapping the genes responsible for inherited diseases by using restriction fragment length polymorphisms (RFLPs), screening collections of cloned DNA fragments (DNA libraries), "DNA fingerprinting" for analysis of biological material remaining at the scene of a crime, and identification of specific proteins. More details on blotting are found in Chapter 10, Section C, p. 304.

The Western Blot

Western blotting has become an important, modern technique for analysis and characterization of proteins. The procedure consists of the electrophoretic transfer (blotting) of proteins from a polyacrylamide gel to a synthetic membrane. The transferred blots are then probed using various methods involving antibodies that bind to specific proteins.

PAGE is indeed a very effective analytical tool to achieve fractionation of protein mixtures, to analyze purity, and to estimate molecular weight, but it provides no experimental data to prove the identity of any of the protein bands. Stained spots simply indicate the presence and location of each and every protein on the gel. It is often possible to identify proteins by treating gel bands directly with chemical reagents that react with a specific protein; for example, enzymes on the gel could be treated with substrates that form a colored product. However, proteins are deeply embedded in the polyacrylamide gel matrix and are not readily accessible to most analytical reagents. This hinders specific analysis of the protein bands in order to identify individual proteins. Proteins separated by PAGE may be transferred (or blotted) from the gel to a thin support matrix, usually a nitrocellulose membrane, which strongly binds and immobilizes proteins. The protein blots on the membrane surface are more accessible to chemical or biochemical reagents for further analysis. When the transfer process is coupled with protein identification using highly specific and sensitive immunological detection techniques, the procedure is called **Western blotting.** Western blotting or immunoblotting assays of proteins have many advantages, including the need for only small reagent volumes, short processing times, relatively inexpensive equipment, and ease of performance (see immunoadsorption, Chapter 5, Section G, p. 152).

The Western Blotting Procedure

To begin the Western blot, a protein mixture for analysis and further characterization is fractionated by PAGE. Since denaturing SDS-PAGE results in better resolution than native PAGE, the SDS version is usually preferred; however, the detection method used at the conclusion of the blotting experiment must be able to recognize denatured protein subunits. The next step involves selection of the membrane matrix for transfer. Three types of support matrices are available for use: nitrocellulose, nylon, and polyvinyldifluoride (PVDF). Nitrocellulose membranes, currently the most widely used supports, have a satisfactory general protein-binding capacity (100 μg/cm^2), but they display weak binding of proteins of molecular weights smaller than 14,000 and they are subject to tearing. Binding of proteins to nitrocellulose membranes is noncovalent, most likely involving hydrophobic interactions. Nylon membranes are stronger than nitrocellulose, and some have a binding capacity up to 450 μg/cm^2. However, since they are cationic, they only weakly bind basic proteins. During detection procedures, nylon membranes often display high background colors, so it is difficult to visualize proteins of interest. PVDF membranes bind proteins strongly (125 μg/cm^2) and, because of their hydrophobic nature, give a light background color after analysis. For overall general use in protein transfer and immunoblotting, nitrocellulose membranes are the most common choice.

The actual blotting process may be accomplished by one of two methods: **passive (or capillary) transfer** and **electroblotting.** In passive transfer, the membrane is placed in direct contact with the polyacrylamide gel and organized in a sandwich-like arrangement consisting of (from bottom to top) filter paper soaked with transfer buffer, gel, membrane, and more filter paper (see Figure 6.17). The sandwich is compressed by a heavy weight. Buffer passes by capillary action from the bottom filter paper through the gel, transferring the protein molecules to the membrane, where the macromolecules are immobilized. Passive transfer is very time-consuming, sometimes requiring 1–2 days for complete protein transfer. Faster and more efficient transfer is afforded by the use of an electroblotter. Here a sandwich of filter paper, gel, membrane, and more filter paper is prepared in a cassette, which is placed between platinum electrodes. An electric current is passed through the gel, causing the proteins to electrophorese out of the gel and onto the membrane. The electroblotting process usually is complete in 1–4 hours.

Detection of Blotted Proteins

The Western blot procedure is concluded by probing the blotted protein bands and detecting a specific protein or group of proteins among the blots. In other words, visualization of specific protein blots is now possible. The most specific identification techniques are based on immunology (antigen–antibody interactions; see Chapter 5, Section G, p. 152). A general procedure for immunoblotting is outlined in Figure 6.18. Before the protein detection process can begin, it is necessary to block protein-binding sites on the membrane that are not occupied by blotted proteins. This is essential because antibodies used to detect blotted proteins are also proteins and will bind to the membrane and interfere with detection procedures. Protein-binding sites still remaining on blotted membranes may be blocked by treatment with solutions of casein (major protein in milk), gelatin, or bovine serum albumin.

The blotted membrane, with all protein-binding sites occupied, can now be treated with analytical reagents for detection of specific proteins. Typically, the blotted membrane is incubated with an antibody specific for the protein of interest. This is called the **primary antibody,** which is a protein of the immunoglobulin G (IgG) class (see Figure 6.18). The primary antibody binds to the desired protein, forming an antigen–antibody complex. The interaction between the protein and its antibody does not usually result in a visible signal. The blot is then incubated with a **secondary antibody,** which is directed against the general class

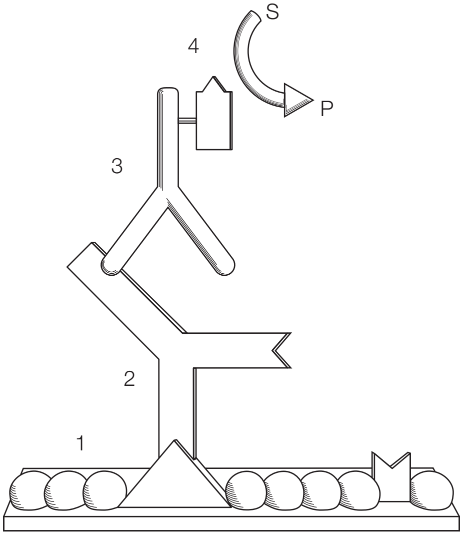

FIGURE 6.18 Specific detection of nitrocellulose membrane-bound proteins using a conjugated enzyme. **1** Proteins are transferred from electrophoresis gel to nitrocellulose membrane. Blocker proteins bind to unoccupied sites on the membrane. **2** The membrane is incubated with a primary antibody directed against the protein of interest. **3** A secondary antibody is directed against the primary antibody. **4** The secondary antibody is conjugated with an enzyme to provide a detection mechanism. Substrate solution is added to the blot. The conjugated enzyme (HRP or AP) catalyzes the conversion of substrate (S) to product (P) to form a colored precipitate at the site of the protein–antibody complex.

of primary antibody. For example, if the primary antibody was produced in rabbit serum, then the second antibody would be anti-rabbit IgG, usually from a goat or horse. The second antibody is labeled (conjugated) so that the interaction of the second antibody with the primary antibody produces some visual signal. For most detection procedures, the secondary antibody may be tagged with an enzyme, usually horseradish peroxidase (HRP) or alkaline phosphatase (AP). When the treated blot is incubated in a substrate solution, the conjugated enzyme catalyzes the conversion of the substrate into a visible product that precipitates at the blot site. The presence of a colored band indicates the position of the protein of interest (see Figure 6.19). This general procedure is also the basis for the widely used enzyme-linked immunosorptive assay (ELISA). The reactions catalyzed by conjugated enzymes are shown below:

>> $$\text{4-chloro-1-naphthol}_{\text{reduced}} + H_2O_2 \overset{\text{HRP}}{\rightleftharpoons} \text{4-chloro-1-naphthol}_{\text{oxidized}} + H_2O + \tfrac{1}{2}O_2 \qquad \textbf{(Eq. 6.5)}$$

>> $$\text{5-bromo-4-chloro-3-indolylphosphate} + \text{nitroblue tetrazolium}_{\text{oxidized}} \overset{\text{AP}}{\rightleftharpoons} \text{5-bromo-4-chloro-3-indole} + P_i + \text{nitroblue tetrazolium}_{\text{reduced}} \qquad \textbf{(Eq. 6.6)}$$

A modification of these coloring systems has recently been developed that leads to more sensitive detection. Chemiluminescent substrates have been

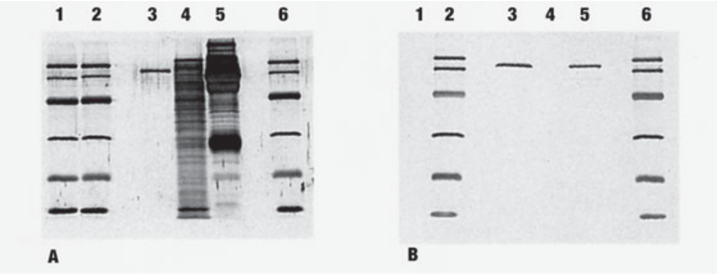

FIGURE 6.19 Results from a Western blot. **A** SDS-PAGE gels, 12%, were run and transferred to nitrocellulose. Lane **1,** MW standards; lane **2,** biotinylated standards; lane **3,** human transferrin; lane **4,** *E. coli* lysate; lane **5**, total human serum; lane **6**, biotinylated standards. Gel A was stained with a protein dye. Blot **B** was assayed using rabbit anti-human transferrin as the first antibody. The second antibody solution contained anti-rabbit HRP conjugates. Only the transferrin bands and the prestained biotinylated standards were detected by the antibodies and the avidin-HRP treatment.

designed that are converted by the enzymes to products that generate a light signal that can be captured on photographic film. This increases the level of sensitivity about 1000-fold over standard color detection methods.

Even though primary and secondary antibodies are widely used in Western blotting detection systems, they do have some disadvantages. For proteins to be detected, specific antibodies must be available. It is often very time-consuming and expensive for a research laboratory to generate the proper antibodies if they are not available commercially. Even if antibodies are commercially available, they are very expensive.

The Western blot has gained widespread use in biochemical and clinical investigations. It is one of the best methods for identifying the presence of specific proteins in complex biological mixtures. The Western blot procedure has been modified to develop a diagnostic assay that detects the presence in serum of antibodies to the AIDS virus. The presence of AIDS antibodies in a patient is an indication of viral infection.

Analysis of Electrophoresis Results

By separating biochemicals on the basis of charge, size, and conformation, electrophoresis can provide valuable information, such as purity, identity, and molecular weight. Purity is indicated by the number of stained bands in the electropherogram. One band usually means that only one detectable component is present; that is, the sample is homogeneous or "electrophoretically pure." Two or more bands usually indicate that the sample contains two or more components, contaminants, or impurities and is therefore heterogeneous or impure. There are, of course, exceptions. Other proteins or nucleic acids may be present in what appears to be a homogeneous sample, but they may be below the limit of detection of the staining method. Occasionally, a homogeneous sample may result in two or more bands because of degradation during the electrophoresis process. Information on purity may be obtained by all of the electrophoresis methods discussed.

The identity of unknown biomolecules can be confirmed by electrophoresis of the unknown along with possible known standards on the same gel. This is similar to the identification of unknowns by HPLC as discussed in Chapter 5, Section F, p. 140.

As previously discussed in this chapter, the molecular size of protein or nucleic acid samples may be determined by electrophoresis. This requires the preparation of standard curves of log molecular weight versus μ (mobility) using standard proteins or nucleic acids.

Study Problems

1. What physical characteristics of a biomolecule influence its rate of movement in an electrophoresis matrix?
2. Each of the proteins listed below is treated with sodium dodecyl sulfate and separated by electrophoresis on a polyacrylamide slab gel. Draw pictures of the final results.
   (a) Myoglobin
   (b) Hemoglobin (two α subunits, molecular weight = 15,500; two β subunits, molecular weight = 16,000)
3. Draw a slab gel to show the results of nondenaturing electrophoresis of the following mixture of proteins. The molecular weight is given for each.

Lysozyme (13,930)	Egg white albumin (45,000)
Chymotrypsin (21,600)	Serum albumin (65,400)

4. Explain the purpose of each of the chemical reagents that are used for PAGE.
   (a) acrylamide
   (b) N,N'-methylene-bis-acrylamide
   (c) TEMED
   (d) sodium dodecyl sulfate
   (e) Coomassie Blue dye
   (f) bromophenol blue
5. What is the main advantage of slab gels over column gels for PAGE?
6. Explain the purposes of protein and nucleic acid "blotting."
7. Is it possible to use polyacrylamide as a matrix for electrophoresis of nucleic acids? What are the limitations, if any?
8. Can polyacrylamide gels be used for the analysis of plasmid DNA with greater than 3000 base pairs? Why or why not?
9. Describe the toxic characteristics of acrylamide, and outline precautions necessary for its use.
10. The dye ethidium bromide is often used to detect the presence of nucleic acids on electrophoresis supports. Explain how it functions as an indicator.
11. The name "Western blot" is derived from other blotting procedures called the Southern blot (developed by Earl Southern) and the Northern blot. What are the differences among these three types of blotting techniques and what is the purpose of each?
12. An economical way to "block" a Western blot membrane is to incubate it in a 10% solution of nonfat milk powder. How does this solution function as a blocking reagent?
13. If a protein you wish to analyze by Western blotting is acidic (anionic under blotting conditions), what type of membrane would be best to ensure the tightest binding?
14. Design a diagnostic test based on the Western blot that would give an indication of infection by the AIDS virus. Assume that a blood serum sample is available from the patient.
15. A protein mixture can be fractionated either by native PAGE or by denaturing SDS-PAGE, before Western blotting. What factors would determine your choice of electrophoresis method?

16. Bromophenol blue dye is often used as a marker to tell you when to stop the electrophoresis process. What assumptions must you make about the relative movement of the dye versus sample proteins during electrophoresis?

17. In your biochemistry research project, you have isolated a new protein from spinach leaves. You wish to do a Western blot experiment with the protein to help determine its chemical structure and/or biological function. An amino acid analysis of the protein showed a great abundance of Phe, Leu, and Val. What would be your choice of membrane for the blotting experiment?

18. Define each of the following items in terms of their use in Western blotting.
   (a) SDS-PAGE
   (b) Nylon membrane
   (c) Electroblotting
   (d) Primary antibody
   (e) Secondary antibody
   (f) Protein molecular weight standard mixture
   (g) Conjugated enzyme

Further Reading

J. Berg, J. Tymoczko, and L. Stryer, ***Biochemistry,*** 6th ed. (2007), Freeman (New York). Analysis of proteins by electrophoresis.

K. Bourzac, L. LaVine, and M. Rice, ***J. Chem. Educ.*** **80,** 1292–1297 (2003). “Analysis of DAPI and SYBR Green I as Alternatives to Ethidium Bromide for Nucleic Acid Staining in Agarose Gel Electrophoresis.”

R. Boyer, ***Concepts in Biochemistry,*** 3rd ed. (2006), John Wiley & Sons (Hoboken, NJ), pp. 84–86. Introduction to electrophoresis.

R. Boyer, ***Modern Experimental Biochemistry,*** 3rd ed. (2000), Benjamin/Cummings (San Francisco), pp. 321–333. Identification of serum glycoproteins by SDS-PAGE and Western blotting.

E. Eberhardt et al., ***Biochem. Mol. Biol. Educ.*** **31,** 402–409 (2003). “Preparing Undergraduates to Participate in the Post-Genome Era: A Capstone Laboratory Experience in Proteomics.”

R. Garrett and C. Grisham, ***Biochemistry,*** 4th ed. (2010), Brooks Cole (Belmont, CA). An introduction to electrophoresis.

D. Nelson and M. Cox, ***Lehninger Principles of Biochemistry,*** 5th ed. (2009), Freeman (New York). Introduction to electrophoresis.

N. Price, R. Dwek, M. Wormald, and R. Ratcliffe, ***Principles and Problems in Physical Chemistry for Biochemists,*** 3rd ed. (2002), Oxford University Press (New York).

P. Schmitt-Kopplin, ***Capillary Electrophoresis: Methods and Protocols,*** (2008), Humana Press (Totowa, NJ).

D. Sheehan, ***Physical Biochemistry: Principles and Applications,*** 2nd ed. (2009), Wiley-Blackwell (Hoboken, NJ).

I. Tinoco, Jr., K. Sauer, J. Wang, and J. Puglisi, ***Physical Chemistry: Principles and Applications in Biological Sciences,*** 4th ed. (2002), Prentice Hall (Upper Saddle River, NJ). Chapter 6 covers electrophoresis.

A. Van Eeckhaut and Y. Michotte (Editors), ***Chiral Separations by Capillary Electrophoresis*** (2009), CRC Press.

D. Voet and J. Voet, ***Biochemistry,*** 3rd ed. (2005), John Wiley & Sons (Hoboken, NJ). Introduction to electrophoresis.

D. Voet, J. Voet, and C. Pratt, ***Fundamentals of Biochemistry,*** 3rd ed. (2008), John Wiley & Sons (Hoboken, NJ). An introduction to electrophoresis.

J. Walker, Editor, ***The Protein Protocols Handbook,*** 2nd ed. (2002), Humana Press (Totowa, NJ). Contains several chapters on electrophoresis of proteins and peptides.

R. Westermeier, ***Electrophoresis in Practice: A Guide to Methods and Applications of DNA and Protein Separation,*** (2005), Wiley-VCH.

J. Yan et al., ***Electrophoresis* 21,** 3657–3665 (2000). "Postelectrophoretic Staining of Proteins Separated by Two-Dimensional Gel Electrophoresis Using SYPRO Dyes."

http://biotech.biology.arizona.edu/labs/Electrophoresis_dyes_stude.html
The University of Arizona Biotech Project.

http://www.probes.com
This Web site for Molecular Probes Inc. has extensive information on gel staining for proteins and nucleic acids.

http://www.med.unc.edu/wrkunits/3ctrpgm/pmbb/mbt/WESTERN.htm
For introduction and procedures, read Unit 1: Western Blotting.

http://www.bio-rad.com
Information on all areas of electrophoresis and blotting.

http://proteomics.cancer.dk
Danish Center for Human Genome Research: Human 2-D PAGE Databases.

http://en.wikipedia.org/wiki/Gel_electrophoresis
Introduction to gel electrophoresis.

http://en.wikipedia.org/wiki/Agarose_gel_electrophoresis
Introduction to agarose gel electrophoresis.

http://harrisphq.com/eThermo/CMA/PDFs/Articles/articlesFile_21627.pdf
A basic primer to protein electrophoresis.

http://www.hoeferinc.com/static_products.asp?page=54&gclid=CLiytpynrp8CFRRcbQodHlDu1Q
Protein electrophoresis.

http://www.piercenet.com/Products/Browse.cfm?fldID=020510
Protein electrophoresis.

http://www.sigma-aldrich.com/life-science/proteomics/protein-electrophoresis.html
Protein electrophoresis.

http://www.invitrogen.com
Sypro dyes.

http://www.hoeferinc.com/static_products.asp?page=57
Nucleic acid electrophoresis.

http://www.interchim.com/interchim/bio/cat_bio/pdf/Genomics_DNA_analysis_Electro_Nucleic_acids.pdf
GelRed dye.

http://www.biotium.com/
View slide presentation on GelRed and GelGreen dye use and safety info.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

CHAPTER

7

SPECTROSCOPIC ANALYSIS OF BIOMOLECULES

A. Ultraviolet-Visible Absorption Spectrometry 202
B. Fluorescence Spectrometry 220
C. Nuclear Magnetic Resonance Spectroscopy 225
D. Mass Spectrometry 230
E. X-Ray Crystallography 233

Study Problems 234
Further Reading 235

Some of the earliest experimental measurements on biomolecules involved studies of their interactions with electromagnetic radiation of all wavelengths, including X-ray, ultraviolet–visible, and infrared. It was experimentally observed that when light impinges on solutions or crystals of molecules, at least two distinct processes occur: **light scattering** and **light absorption.** Both processes have led to the development of fundamental techniques for characterizing and analyzing biomolecules. We now use the term **spectroscopy** to label the discipline that studies the interaction of electromagnetic radiation with matter.

Absorption of ultraviolet–visible light by molecules is an especially valuable process for measuring concentration and for molecular structure elucidation. The absorption process is dependent upon two factors: (1) the properties of the radiation (wavelength, energy, etc.), and (2) the structural characteristics of the absorbing molecules (atoms, functional groups, etc.). The interaction of electromagnetic radiation with molecules is a quantum process and described mathematically by quantum mechanics; that is, the radiation is subdivided into discrete energy packets called **photons.** In addition, molecules have quantized excitation levels and can accept packets of only certain quantities of energy, thus allowing only certain electronic transitions.

With some molecules, the process of absorption is followed by emission of light of a longer wavelength. This process, called **fluorescence,** depends on molecular structure and environmental factors and assists in the characterization and analysis of biologically significant molecules and dynamic processes occurring between molecules.

Nuclear magnetic resonance spectroscopy and **mass spectrometry** techniques are also now being applied to the study of biological macromolecules and processes. NMR is especially versatile because, in addition to proton spectra, monitoring the presence of ^{13}C, ^{19}F, ^{15}N, and ^{31}P nuclei in biomolecules is possible. Multidimensional NMR is being applied to the study of protein secondary and tertiary structure. Accurate measurements of protein molecular weight and sequence analysis may now be carried out by mass spectrometry.

A. ULTRAVIOLET–VISIBLE ABSORPTION SPECTROMETRY

Wavelength and Energy

The electromagnetic spectrum, as shown in Figure 7.1, is composed of a continuum of waves with different properties defined by wavelength and energy. Several regions of the electromagnetic spectrum are of importance in biochemical studies, including X-ray (X-ray crystallography, up to 7 nm), the ultraviolet (UV, 180–340 nm), the visible (VIS, 340–800 nm), the infrared (IR, 1000–100,000 nm), and radio waves (NMR, 10^6-10^{10} nm). In this section, we will concentrate on the UV and VIS regions. Light in these regions has energy sufficient to excite the valence electrons of molecules and thus move the electrons from one energy level (or state) to a higher energy level (excited state).

Figure 7.2 shows that the propagation of light is due to an electrical field component E and a magnetic field component H that are perpendicular to each other. The **wavelength** of light, defined by Equation 7.1, is the distance between adjacent wave peaks, as shown in Figure 7.2.

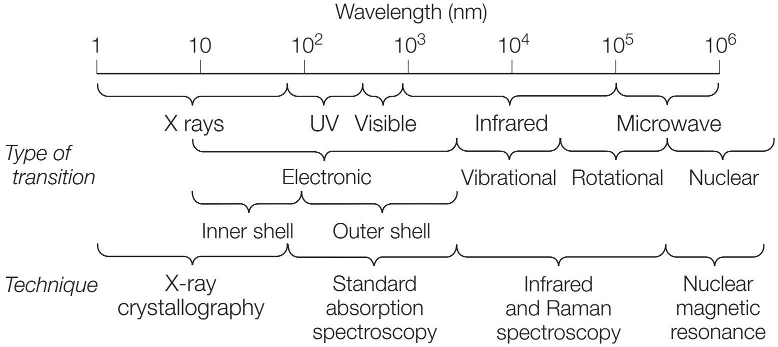
Wavelength (nm)
1
10
10^2
10^3
10^4
10^5
10^6
X rays
UV
Visible
Infrared
Microwave
Type of transition
Electronic
Vibrational
Rotational
Nuclear
Inner shell
Outer shell
Technique
X-ray crystallography
Standard absorption spectroscopy
Infrared and Raman spectroscopy
Nuclear magnetic resonance

FIGURE 7.1 The electromagnetic spectrum. The wavelengths of light associated with important spectroscopy techniques are shown in the figure.

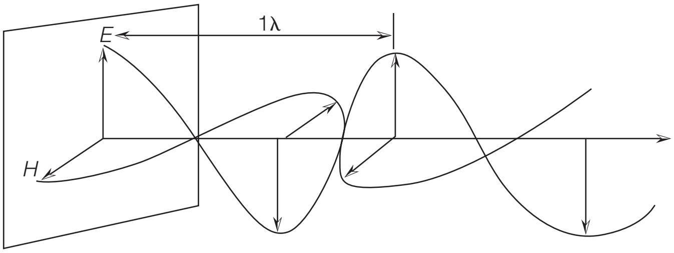

FIGURE 7.2 An electromagnetic wave, showing the *E* and *H* components.

>> $$\lambda = c/\nu \qquad \textbf{(Eq. 7.1)}$$

where

λ = wavelength
c = speed of light, 3×10^8 m/s
ν = frequency, the number of waves passing a certain point per unit time

Light also behaves as though it were composed of energetic particles. The amount of energy E associated with these particles (or **photons**) is given by Equation 7.2:

>> $$E = h\nu \qquad \textbf{(Eq. 7.2)}$$

where

h = Planck's constant, 6.63×10^{-34} Joules · s

Equations 7.1 and 7.2 may be combined to yield Equation 7.3:

>> $$E = hc/\lambda \qquad \textbf{(Eq. 7.3)}$$

Note the inverse relationship between wavelength and energy. In Figure 7.1, X-rays have the shortest wavelength, but the most energy, and microwaves have long wavelengths, but the least energy.

STUDY EXERCISE 7.1 UV-VIS Spectroscopy

What is the amount of energy (E) associated with light of wavelength 300 nm in the ultraviolet range?

Solution: Using Equation 7.3:

$$E = hc/\lambda$$

$$E = \frac{(6.63 \times 10^{-34}\ \mathrm{J \cdot s})(3 \times 10^{8}\ \mathrm{m/s})}{300 \times 10^{-9}\ \mathrm{m}}$$

$$= 6.63 \times 10^{-19}\ \mathrm{J/photon} = 398\ \mathrm{kJ/mole}$$

This is the approximate amount of energy associated with carbon-based covalent bonds.

Light Absorption

When a photon of specified energy interacts with a molecule, one of two processes may occur. The photon may be **scattered,** or it may transfer its energy to the molecule, producing an **excited state** of the molecule. The former process, called **Rayleigh scattering,** occurs when a photon collides with a molecule and is diffracted or scattered with unchanged frequency. Light scattering is the physical basis of several experimental methods used to characterize macromolecules. Before the development of electrophoresis, light scattering techniques were used to measure the molecular weights of macromolecules. The widely used techniques of X-ray diffraction (crystal and solution), electron microscopy, laser light scattering, and neutron scattering all rely in some way on the light scattering process.

The other process mentioned above, the transfer of energy from a photon to a molecule, is **absorption.** For a photon to be absorbed, its energy must match the energy difference between two energy levels of the molecule. Molecules possess a set of quantized energy levels, as shown in Figure 7.3. Although several states are possible, only two electronic states are shown, a **ground state,** *G*, and the **first excited state,** S_1. These two states differ in the distribution of valence electrons. When electrons are promoted from a ground state orbital in *G* to an orbital of higher energy in S_1, an **electronic transition** is said to occur. The energy associated with ultraviolet and visible light is sufficient to promote molecules from one electronic state to another, that is, to move electrons from one quantized level to another.

Within each electronic energy level is a set of **vibrational levels.** These represent changes in the stretching and bending of covalent bonds. The importance of these energy levels will not be discussed here, but transitions between the vibrational levels are the basis of infrared spectroscopy.

The electronic transition for a molecule from *G* to S_1, represented by the vertical arrow in Figure 7.3, has a high probability of occurring if the energy of the photon corresponds to the energy necessary to promote an electron from energy level E_1 to energy level E_2:

>> $$E_2 - E_1 = \Delta E = \frac{hc}{\lambda} \qquad \textbf{(Eq. 7.4)}$$

A transition may occur from any vibrational level in *G* to some other vibrational level in S_1, for example, $\nu = 3$; however, not all transitions have equal probability.

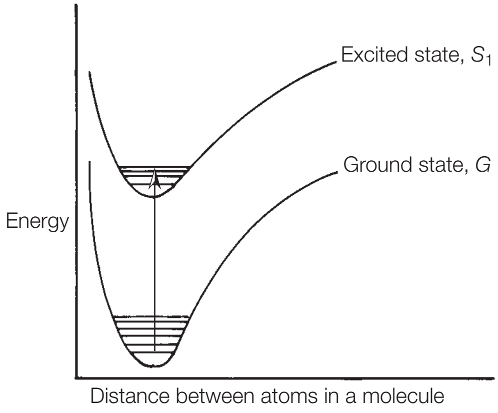

FIGURE 7.3
Energy-level diagram showing the ground state G and the first excited state S_1.

The probability of absorption is described by quantum mechanics and will not be discussed here.

Electronic Transitions in Biomolecules

Absorption of UV and VIS light occurs when the energy associated with the electromagnetic radiation matches the energy difference between molecular states in a molecule. But what are the actual electronic transitions that occur in a molecule? If this question can be answered, then it should be possible to predict which molecules will or will not absorb UV-VIS light. The reverse is also true—by observing the wavelengths of light absorbed by a molecule, it should be possible to speculate on molecular structure (what atoms and functional groups are present). Consider the electronic properties of a carbonyl group, 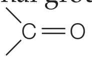 a functional group present in many types of biomolecules including amino acids, fatty acids, carbohydrates (open chain), nucleotides, proteins, and nucleic acids. A molecular orbital diagram showing the valence electrons in the higher-energy orbitals of the carbonyl group is displayed in Figure 7.4. There are three types of orbitals shown:

- n = nonbonding orbitals with lone pairs of electrons in atoms like O, N, and S
- π = bonding orbitals formed by overlap of p atomic orbitals
- π^* = antibonding orbitals

Two kinds of electronic transitions are possible:

- $\pi \rightarrow \pi^*$, which are "allowed" by quantum mechanics and have a high probability of occurring, and
- $n \rightarrow \pi^*$, which are "forbidden" and have a low probability of occurring.

These two types of electronic transitions are shown in Figure 7.4.

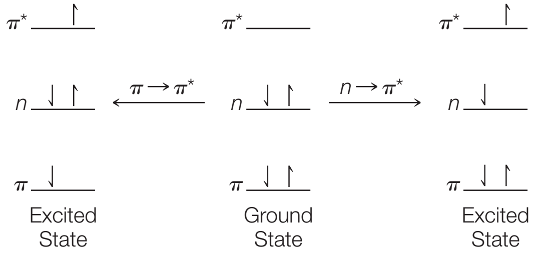

FIGURE 7.4 Molecular orbital diagram for the carbonyl group. Only valence electrons in higher-energy orbitals are shown. Two types of electronic transitions are illustrated: $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$.

Proteins

The following functional groups in proteins are responsible for absorption of UV-VIS light:

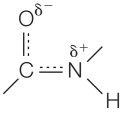

1. Peptide bonds:
    - $\pi \rightarrow \pi^*$, intense absorption at 190 nm
    - $n \rightarrow \pi^*$, very weak absorption at 210–220 nm
2. Amino acid residues containing aromatic phenyl groups: Phe, Tyr, Trp

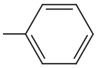

    - $\pi \rightarrow \pi^*$
      Phe: weak absorption at 250 nm
      Tyr: weak absorption at 274 nm
      Trp: strong absorption at 280 nm
      The absorbance at 280 nm is often used to measure protein concentration (See Chapter 3, Section B, p. 70, The Spectrophotometric Assay).
3. Prosthetic groups:
   Nucleotide cofactors: NAD^+, FAD, FMN, etc.
   Heme
   Chlorophyll
   Metal ions: Cu(II), Fe(II), Fe(III), Ni(II), etc.
   Highly conjugated compounds: retinal, etc.

Many of the cofactors and prosthetic groups above absorb in both the UV and VIS regions. Absorption in the VIS region causes many of the biomolecules in Group 3 to be colored.

Nucleic Acids

Absorption is due to the presence of pyrimidine and purine bases, which have several different types of electronic transitions (see structures, Chapter 9, p. 268):

1. Extensive aromatic character, ($\pi \rightarrow \pi^*$)
2. Carbonyl groups in bases, ($\pi \rightarrow \pi^*$; $n \rightarrow \pi^*$)
3. Several nitrogen and oxygen atoms, ($n \rightarrow \pi^*$).

Nucleic acids display intense absorption in the 200–300-nm range with a maximum at 260 nm. This is the basis for the spectrophotometric assay of DNA and RNA solutions (Chapter 3, Section C, p. 71).

STUDY EXERCISE 7.2 UV-VIS Spectroscopy

Predict the types of electronic transitions in each of the following biomolecules and suggest whether UV-VIS spectroscopy might be useful in their analysis and characterization. You may need to review their structures in your biochemistry textbook.

(a) Alanine
(b) Lysine
(c) Cysteine
(d) Glucose
(e) Lauric acid
(f) Adenine
(g) β-carotene
(h) Chlorophyll
(i) Bilirubin
(j) Ferritin, a protein containing Fe(III)

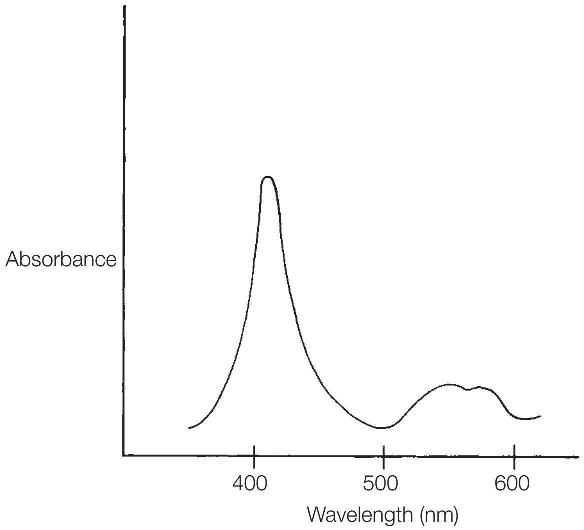

FIGURE 7.5 The visible absorbance spectrum of hemoglobin.

The Absorption Spectrum

A UV-VIS spectrum is obtained by measuring the light absorbed by a sample as a function of wavelength. Since only discrete packets of energy (specific wavelengths) are absorbed by molecules in the sample, the spectrum theoretically should consist of sharp discrete lines. However, the many vibrational levels of each electronic energy level increase the number of possible transitions. This results in several spectral lines, which together make up the familiar spectrum of broad peaks as shown in Figure 7.5.

An absorption spectrum can aid in the identification of a molecule because the wavelength of absorption depends on the functional groups or arrangement of atoms in the sample. The spectrum of oxyhemoglobin in Figure 7.5 is due to the presence of the iron porphyrin moiety and is useful for the characterization of heme derivatives or hemoproteins. Note that the spectrum consists of several peaks at wavelengths where absorption reaches a maximum (415, 542, and 577 nm). These points, called $\lambda_{\mathbf{max}}$, are of great significance in the identification of unknown molecules and will be discussed later in the chapter.

The Beer-Lambert Law

Quantitative measurements in spectrophotometry are evaluated using the Beer-Lambert law:

>> $$A = Elc \qquad \textbf{(Eq. 7.5)}$$

where

A = absorbance or $-\log(I/I_0)$
I_0 = intensity of light irradiating the sample
I = intensity of light transmitted through the sample
E = absorption coefficient or absorptivity

l = path length of light through the sample, or thickness of the cell
c = concentration of absorbing material in the sample

Since the absorbance, A, is derived from a ratio $(-\log I/I_0)$, it is unitless. The term E, which is a proportionality constant, defines the efficiency or extent of absorption. If this is defined for a particular chromophore at a specific wavelength, the term **absorption coefficient** or **absorptivity** is used. However, students should be aware that in the older biochemical literature, the term extinction coefficient is often used. The units of E depend on the units of l (usually cm) and c (usually molar) in Equation 7.5. For biomolecules, E is often used in the form **molar absorption coefficient,** ε, which is defined as the absorbance of a 1 M solution of pure absorbing material in a 1-cm cell under specified conditions of wavelength and solvent. The units of ε are $M^{-1}\,\text{cm}^{-1}$. To illustrate the use of Equation 7.5, consider the following calculation:

STUDY EXERCISE 7.3 Beer-Lambert Law

The absorbance, A, of a $5 \times 10^{-4}\ M$ solution of the amino acid tyrosine, at a wavelength of 280 nm, is 0.75. The path length of the cuvette is 1 cm. What is the molar absorption coefficient, ε?

Solution:

$$A = \varepsilon l c = 0.75$$

$$l = 1\ \text{cm}$$

$$c = 5 \times 10^{-4}\ M$$

$$\varepsilon = \frac{0.75}{(1\ \text{cm})(5 \times 10^{-4}\ \text{mole/liter})}$$

$$= 1500\ \frac{\text{liter}}{\text{mole} \times \text{cm}} = 1500\ M^{-1}\,\text{cm}^{-1}$$

Notice that the units of ε are defined by the concentration units of the tyrosine solution (M) and the dimension units of the cuvette (cm). Although E is most often expressed as a molar absorption coefficient, you may encounter other units such as $E_{\lambda,}^{1\%}$ which is the absorbance of a 1% (w/v) solution of pure absorbing material in a 1-cm cuvette at a specified wavelength, λ.

STUDY EXERCISE 7.4 Absorption Coefficients

In the older biochemical literature, absorption coefficients are often reported as **extinction coefficients ($E_{\lambda}^{1\%}$)**. The extinction coefficient of the enzyme yeast alcohol dehydrogenase is:

$$E_{\lambda}^{1\%} = 12.6$$

Calculate the molar absorption coefficient for yeast alcohol dehydrogenase. The molecular weight for the enzyme is 141,000.

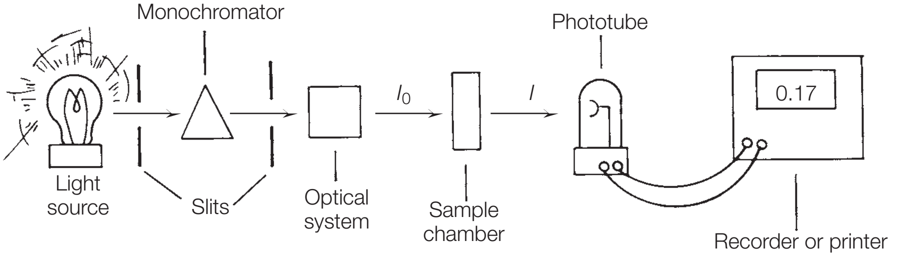

FIGURE 7.6 A diagram of a conventional, single-beam spectrophotometer.

Instrumentation

The **spectrophotometer** is used to measure absorbance experimentally. This instrument produces light of a preselected wavelength, directs it through the sample (usually dissolved in a solvent and placed in a cuvette), and measures the intensity of light transmitted by the sample. The major components are shown in Figure 7.6. These consist of a light source, a monochromator (including various filters, slits, and mirrors), a sample chamber, a detector, and a meter or recorder. All of these components are usually under the control of a computer.

Light Source

For absorption measurements in the ultraviolet region, a high-pressure hydrogen or deuterium lamp is used. These lamps produce radiation in the 200 to 340 nm range. The light source for the visible region is the tungsten-halogen lamp, with a wavelength range of 340 to 800 nm. Instruments with both lamps have greater flexibility and can be used for the study of most biologically significant molecules.

Monochromator

Both lamps discussed above produce continuous emissions of all wavelengths within their range. Therefore, a spectrophotometer must have an optical system to select monochromatic light (light of a specific wavelength). Modern instruments use a prism or, more often, a diffraction grating to produce the desired wavelengths. It should be noted that light emitted from the monochromator is not entirely of a single wavelength, but is enhanced in that wavelength. That is, most of the light is of a single wavelength, but shorter and longer wavelengths are present.

Before the monochromatic light impinges on the sample, it passes through a series of slits, lenses, filters, and mirrors. This optical system concentrates the light, increases the spectral purity, and focuses it toward the sample. The operator of a spectrophotometer has little control over the optical manipulation of the light beam, except for adjustment of slit width. Light passing from the monochromator to the sample encounters a "gate" or slit. The slit width, which is controlled by a computer, determines both the intensity of light impinging on the sample and the spectral purity of that light. Decreasing the slit width increases the spectral purity of the light, but the amount of light directed toward the sample decreases. The efficiency or sensitivity of the detector then becomes a limiting factor.

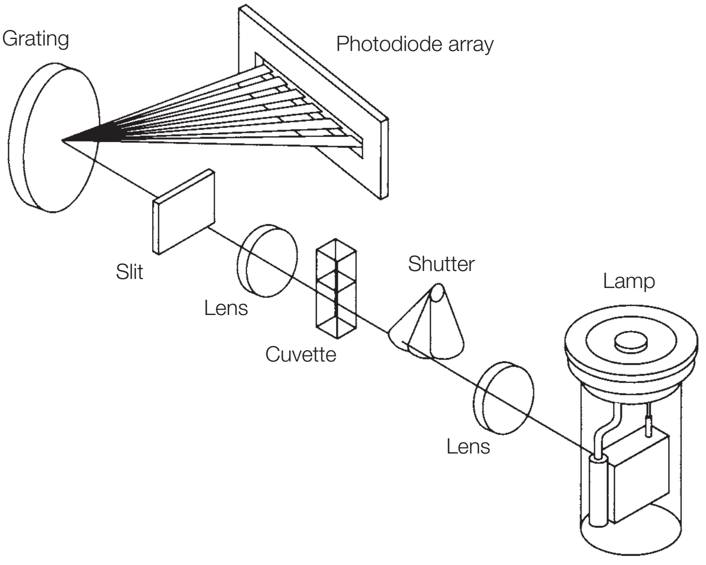

FIGURE 7.7 Optical diagram for a diode-array spectrophotometer.

In instruments equipped with photodiode array detectors (see Figure 7.7 and the following section on detectors), polychromatic light from a source passes through the sample and is focused on the entrance slit of a polychromator (a holographic grating that disperses the light into its wavelengths). The wavelength-resolved light beam is then focused onto the photodiode array detector. The relative positions of the sample and grating are reversed compared to conventional spectrometry; hence, the new configuration is often called reversed optics. (Compare Figures 7.6 and 7.7.)

Sample Chamber

The processed monochromatic light is then directed into a sample chamber, which can accommodate a wide variety of sample holders. Most UV-VIS measurements on biomolecules are taken on solutions of the molecules. The sample is placed in a tube or cuvette made of glass, quartz, or other transparent material. Figure 7.8 shows the transmission properties of several transparent materials used in cuvette construction.

Glass cuvettes are inexpensive, but, because they absorb UV light, they can be used only above 340 nm. Quartz or fused silica cuvettes may be used throughout the UV and visible regions ($\sim 200-800$ nm). Disposable plastic cuvettes are now commercially available in polymethacrylate (280–800 nm) and polystyrene (350–800 nm). The care and use of cuvettes were discussed in Chapter 1, p. 15.

Sample chambers for spectrometers come in two varieties—those holding only one cuvette at a time (single-beam) and those holding two cuvettes, one for a reference, usually solvent, and one for a sample (double-beam). In the past, single-beam instruments were usually less expensive but more cumbersome to use because reference and sample cuvettes required constant exchange. However, modern single-beam instruments with computer control and analysis

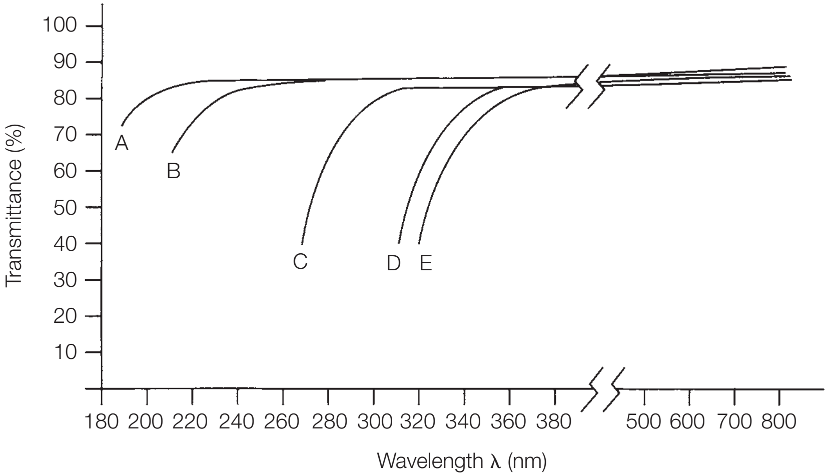

FIGURE 7.8 The transmission properties of several materials used in cuvettes.
A = silica (quartz)
B = NIR silica
C = polymethacrylate
D = polystyrene
E = glass

can be programmed to correct automatically for the reference spectrum, which may be stored in a memory file. In double-beam optics, the light beam is split into two paths by directing it through a monochromator. The two beams, which are of identical wavelength and intensity, pass through the sample cell (analyte plus solvent) and reference chamber (solvent only). This allows for the correction of sample absorbance by continuously subtracting the reference spectrum.

Detector

The intensity of the light that passes through the sample under study depends on the amount of light absorbed by the sample. Intensity is measured by a light-sensitive detector, usually a photomultiplier tube (PMT). The PMT detects a small amount of light energy, amplifies this by a cascade of electrons accelerated by dynodes, and converts it into an electrical signal that can be fed into a meter or recorder.

A new technology has been introduced during the past few years that greatly increases the speed of spectrophotometric measurements. New detectors called **photodiode arrays** are being used in modern spectrometers. Photodiodes are composed of silicon crystals that are sensitive to light in the wavelength range 170–1100 nm. Upon photon absorption by the diode, a current is generated in the photodiode that is proportional to the number of photons. Linear arrays of photodiodes are self-scanning and have response times on the order of 100 milliseconds; hence, an entire UV-VIS spectrum can be obtained with an extremely brief exposure of the sample to polychromatic light. New spectrometers designed by Hewlett-Packard, Perkin-Elmer, and others use this technology and can produce a full spectrum from 190 to 820 nm in one-tenth of a second or less.

Printers and Recorders

Less expensive instruments give a direct readout of absorbance and/or transmittance in analog or digital form. These instruments are suitable for single-wavelength measurements; however, if a scan of absorbance vs. wavelength (Figure 7.5) is desired, some type of device to display the spectrum must be available.

Modern, research-grade spectrometers are available that offer the latest in technology. All of the components discussed above are integrated into a single package and are completely under the control of a computer. By simply pushing a button, one can obtain the UV-VIS spectrum of a sample displayed on a computer screen in less than 1 second. In addition, these modern instruments with computers can be programmed to carry out several functions, such as subtraction of solvent spectrum, spectral overlay, storage, difference spectra, derivative spectra, and calculation of concentrations and rate constants.

Applications of UV–VIS Spectroscopy

Now that you are familiar with the theory and instrumentation of absorption spectrophotometry, you will more easily understand the actual operation and typical applications of a spectrophotometer. Since virtually all UV-VIS measurements are made on samples dissolved in solvents, only those applications will be described here. Although many different types of operations can be carried out on a spectrophotometer, all applications fall in one of two categories:

1. measurement of absorbance at a fixed wavelength;
2. measurement of absorbance as a function of wavelength.

Measurements at a fixed wavelength are most often used to obtain quantitative information, such as the concentration of a solute in solution or the absorption coefficient of a chromophore. Absorbance measurements as a function of wavelength provide qualitative information that assists in solving the identity and structure of a pure substance by detecting characteristic groupings of atoms in a molecule.

For fixed-wavelength measurements with a single-beam instrument, a cuvette containing solvent only is placed in the sample beam and the instrument is adjusted to read “zero” absorbance. A matched cuvette containing sample plus solvent is then placed in the sample chamber and the absorbance is read directly from the display. The adjustment to zero absorbance with only solvent in the sample chamber allows the operator to obtain a direct reading of absorbance for the sample.

Fixed-wavelength measurements using a double-beam spectrophotometer are made by first zeroing the instrument with no cuvette in either the sample or reference holder. Alternatively, the spectrophotometer can be balanced by placing matched cuvettes containing water or solvent in both sample chambers. Then, a cuvette containing pure solvent is placed in the reference position and a matched cuvette containing solvent plus sample is set in the sample position. The absorbance reading given by the instrument is that of the sample; that is, the absorbance due to solvent is subtracted by the instrument.

An **absorbance spectrum** of a compound is obtained by scanning a range of wavelengths and plotting the absorbance at each wavelength. Most double-beam spectrophotometers automatically scan the desired wavelength range and record the absorbance as a function of wavelength. If solvent is placed in the reference chamber and solvent plus sample in the sample position, the instrument will continuously and automatically subtract the solvent absorbance from the total absorbance (solvent plus sample) at each wavelength; hence, the recorder output is really a difference spectrum (absorbance of sample plus solvent, minus absorbance of solvent).

Both types of measurements (fixed wavelength and absorbance spectrum) are common in biochemistry, and you should be able to interpret results from each. The following four examples are typical of the kinds of problems readily solved by spectrophotometry.

Measurement of the Concentration of an Analyte in Solution

According to the Beer-Lambert law, the absorbance of a material in solution is directly dependent on the concentration of that material. Two methods are commonly used to measure concentration. If the absorption coefficient is known for the absorbing species, the concentration can be calculated after experimental measurement of the absorbance of the solution.

STUDY EXERCISE 7.5 Absorbance and Concentration

A solution of the nucleotide base uracil, in a 1-cm cuvette, has an absorbance at $\lambda_{\max}(260\,\text{nm})$ of 0.65. Pure solvent in a matched quartz cuvette has an absorbance of 0.07. What is the molar concentration of the uracil solution? Assume the molar absorption coefficient, ε, is $8.2 \times 10^3 M^{-1}\,\text{cm}^{-1}$.

Solution:

$$A = \varepsilon l c$$

$$A = (\text{absorbance of solvent} + \text{sample}) - (\text{absorbance of solvent})$$

$$A = 0.65 - 0.07 = 0.58$$

$$\varepsilon = 8.2 \times 10^3 M^{-1}\,\text{cm}^{-1}$$

$$l = 1\,\text{cm}$$

$$c = \frac{A}{\varepsilon l} = \frac{0.58}{(8.2 \times 10^3 M^{-1}\,\text{cm}^{-1})(1\,\text{cm})}$$

$$c = 7.1 \times 10^{-5} M$$

If the absorption coefficient for an absorbing species is known, the concentration of that species in solution can be calculated as outlined. However, there are limitations to this application. Most spectrophotometers are useful for measuring absorbances up to 1, although more sophisticated instruments can measure absorbances as high as 2. (If the absorbance is above 1, dilute sample with the same solvent.) Most absorbance readings below 0.1 are not accurate. Also, some substances do not obey the Beer-Lambert law; that is, absorbance may not increase in a linear fashion with concentration. Reasons for deviation from the Beer-Lambert law are many; however, the majority are instrumental, chemical, or physical. Spectrophotometers often display a nonlinear response at high absorption levels because of stray light. Physical reasons for nonlinearity include hydrogen bonding of the absorbing species with the solvent and intermolecular interactions at high concentrations. Chemical reasons may include reaction of the solvent with the absorbing species and the presence of impurities. Linearity is readily tested by preparing a series of concentrations of the absorbing species and measuring the absorbance of each. A plot of *A* vs. concentration should be linear if the Beer-Lambert law is valid. If the absorption coefficient for a species is unknown, its concentration in solution can be measured if the absorbance of a standard solution of the compound is known.

STUDY EXERCISE 7.6 Absorbance and Concentration

The absorbance of a 1% (w/v) solution of the enzyme tyrosinase, in a 1-cm cell at 280 nm, is 24.9. What is the concentration of a tyrosinase solution that has an A_{280} of 0.25?

Since the absorption coefficient, *E*%, is the same for both solutions, the concentration can be calculated by a direct ratio:

Solution:

$$\frac{A_{\text{std}}}{C_{\text{std}}} = \frac{A_x}{C_x}$$

A_{std} = absorbance of the 1% standard solution = 24.9

C_{std} = concentration of the standard solution = 1% (1 g/dL)

A_x = absorbance of the unknown solution = 0.25

C_x = concentration of the unknown solution in, %

$$\frac{24.9}{1\%} = \frac{0.25}{C_x}$$

$$C_x = 0.01\% = 0.01\ \text{g/dL} = 0.1\ \text{mg/mL}$$

Alternatively, the concentration of a species in solution can be determined by preparing a standard curve of absorbance vs. concentration.

STUDY EXERCISE 7.7 Bradford Protein Assay

The Bradford protein assay is one of the most used spectrophotometric assays in biochemistry. (For a discussion of the Bradford assay, see Chapter 3, Section B, p. 69.) Solutions of varying amounts of a standard protein are mixed with reagents that cause the development of a color. The amount of color produced depends on the amount of protein present. The absorbance at 595 nm of each reaction mixture is plotted against the known protein concentration. A protein sample of unknown concentration is treated with the Bradford reagents, and the color is allowed to develop.

The following absorbance measurements are typical for the standard curve of a protein:

Protein (μg per assay)	A_{595}
15	0.07
25	0.15
50	0.28
100	0.55
150	0.90
0.1 mL unknown protein solution	0.10
0.2 mL unknown protein solution	0.22

A standard curve for the data is plotted in Figure 3.6 on page 70. What standard protein was used in this case? Note the linearity, indicating that the Beer-Lambert law is obeyed over this concentration range of standard protein. Two different volumes of unknown protein were tested. This was to ensure that one volume would be in the concentration range of the standard curve. Since the accuracy of the assay is dependent on identical times for color development, the unknowns must be assayed at the same time as the standards. From graphical analysis, an A_{595} of 0.22 corresponds to 40 μg of protein per 0.20 mL. This indicates that the original protein solution concentration was approximately 200 μg/mL or 0.20 mg/mL. Using computer graphics, students can now quickly visualize experimental data and determine the need for further analysis. Most modern computer programs use the method of least squares to calculate automatically the slope, intercepts, and correlation coefficients.

Absorbance Measurements on Very Small Sample Volumes

The absorbance experiments described to this point have been applied to sample sizes of 1 and 3 mL measured in glass or quartz cuvettes. Biomolecules are often expensive, difficult to isolate, and not always available in these large sample sizes. Recent advances in the development of spectrophotometers now make available instruments for rapidly quantifying and analyzing micro-volume samples. For example, NanoDrop UV-VIS spectrophotometers from Thermo Scientific (www.thermo.com/nanodrop) can measure samples as small as 0.5μl in about 5 seconds. The wavelength range of the spectrometers is 220–750 nm, and they can make both individual absorbance measurements and absorption spectra (scans). Cuvettes and capillaries are not necessary, as the sample droplet is applied directly onto an optical pedestal for measurement. Measurements can be made on any biomolecules that absorb in the wavelength range. This includes proteins, DNA, RNA, nucleotides, cofactors, fluorescent substances, and many others.

Identification of Unknown Biomolecules by Spectrophotometry

The UV-VIS spectrum of a biomolecule reveals much about its molecular structure. Therefore, a spectral analysis is one of the first experimental measurements made on an unknown biomolecule. Natural molecules often contain chromophoric (color-producing) functional groups that have characteristic spectra. Figure 7.9 displays spectra of the well-characterized biomolecules DNA, FMN, $FMNH_2$, NAD^+, NADH, a nucleotide, and a pyrimidine base.

The procedure for obtaining a UV-VIS spectrum begins with the preparation of a solution of the species under study. A standard solution should be prepared in an appropriate solvent. An aliquot of the solution is transferred to a

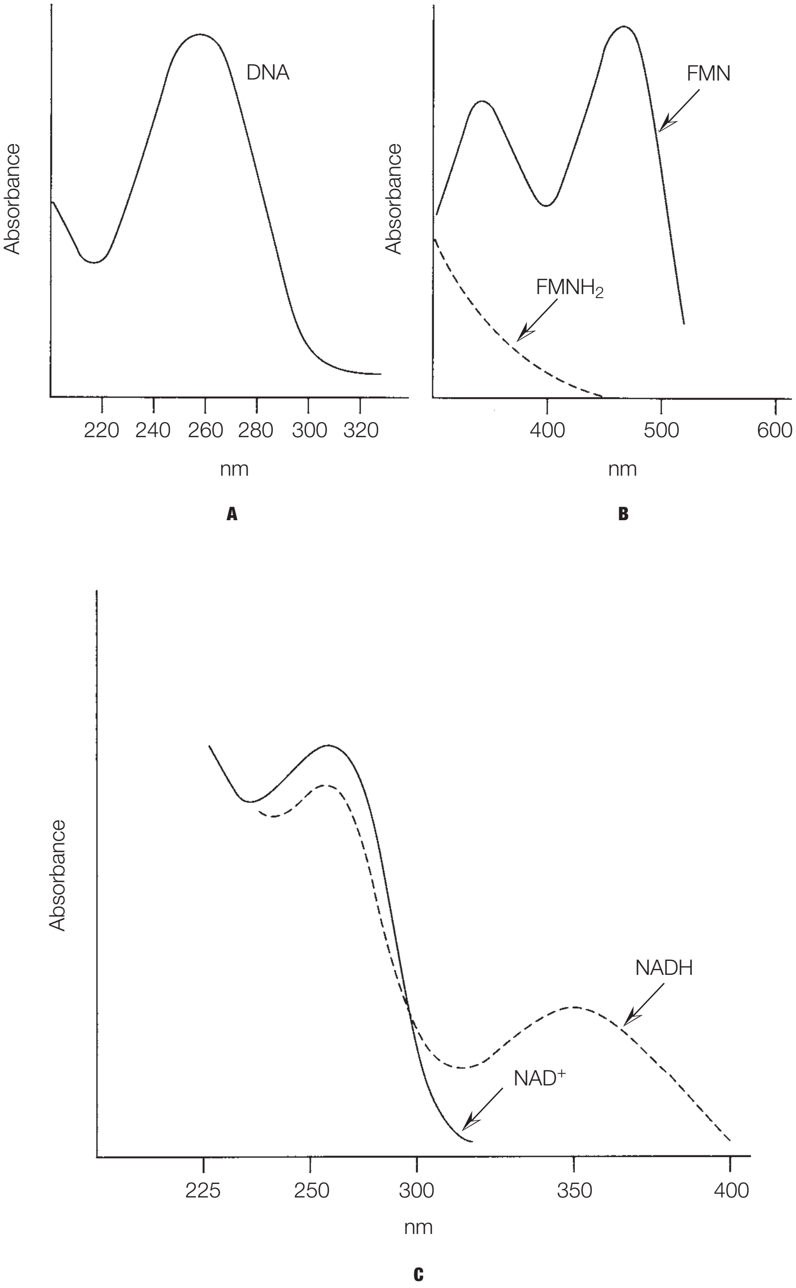

FIGURE 7.9
UV-VIS absorbance spectra of significant biomolecules.
A DNA, **B** FMN and $FMNH_2$, **C** NAD^+ and NADH,
D d-GMP,
E thymine.

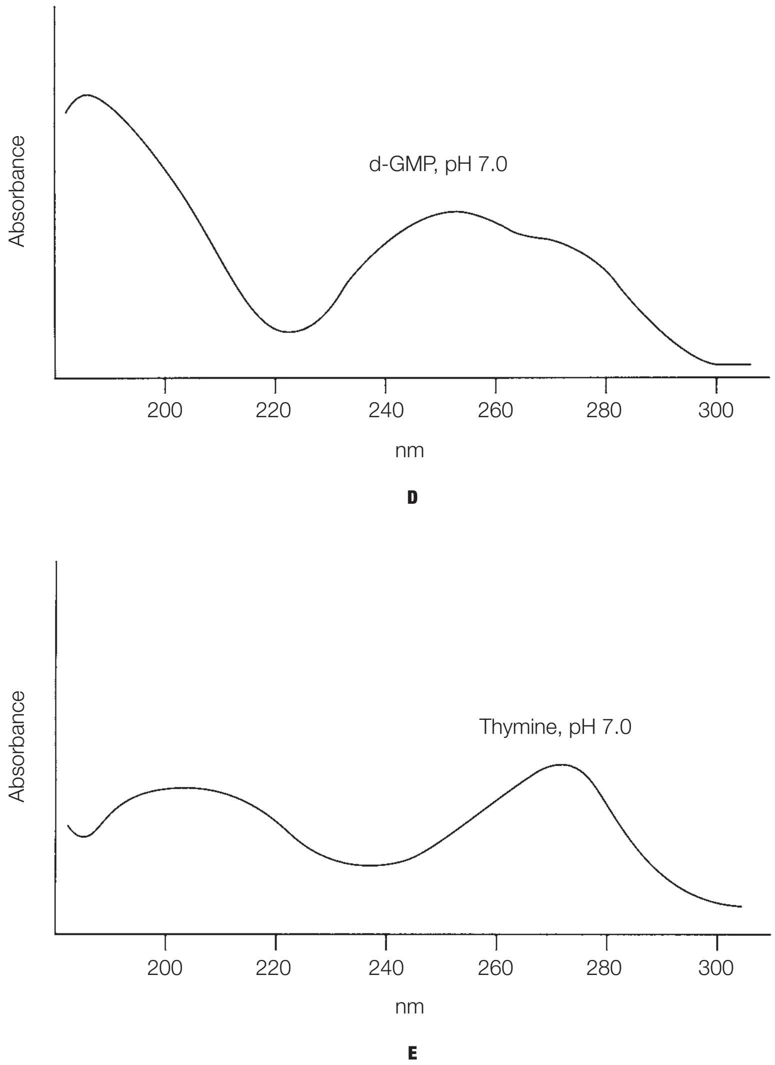

FIGURE 7.9
(*Continued*)

cuvette and placed in the sample chamber of a spectrophotometer. A cuvette containing solvent is placed in the reference holder. The spectrum is scanned over the desired wavelength range and an absorption coefficient is calculated for each major λ_{max}.

Kinetics of Biochemical Reactions

Spectrophotometry is one of the best methods available for measuring the rates of biochemical reactions. Consider a general reaction as shown in Equation 7.6.

>> $$A + B \rightleftharpoons C + D \qquad \textbf{(Eq. 7.6)}$$

If reactants A or B absorb in the UV-VIS region of the spectrum at some wavelength λ_1, the rate of the reaction can be measured by monitoring the decrease of absorbance at λ_1 due to loss of A or B. Alternatively, if products C or D absorb at a specific wavelength λ_2, the kinetics of the reaction can be evaluated by monitoring the absorbance increase at λ_2. According to the Beer-Lambert law, the absorbance change of a reactant or product is proportional to the concentration change of that species occurring during the reaction. This method is widely used to assay enzyme-catalyzed processes. Since the rates of chemical reactions vary with temperature, the sample cuvette containing the reaction mixture must be held in a thermostated chamber.

Characterization of Macromolecule-Ligand Interactions by Difference Spectroscopy

Many biological processes depend on a specific interaction between molecules. The interaction often involves a macromolecule (protein or nucleic acid) and a smaller molecule, a **ligand.** Specific examples include enzyme-substrate interactions and receptor protein–hormone interactions (see Chapter 8). One of the most effective and convenient methods for detecting and characterizing such interactions is **difference spectroscopy.** The interaction of small molecules with the transport protein hemoglobin is a classic example of the utility of difference spectroscopy. If a small molecule or ligand, such as inositol hexaphosphate, binds to hemoglobin, there is a change in the heme spectral properties. The spectral change is small and would be difficult to detect if the experimenter recorded the spectrum in the usual fashion. Normally, one would obtain a spectrum of a heme protein by placing a solution of the protein in a cuvette in the sample compartment of a spectrophotometer and the neat solvent in the reference compartment. Any absorption due to solvent is subtracted because the solvent is present in both light beams, so the spectrum is that due to the heme protein. Then the ligand to be tested would be added to the heme protein, and the spectrum would be obtained for this mixture vs. solvent. If the free or bound ligand molecule did not absorb light in the wavelength range studied, there would be no need to have ligand in the reference cell. The two spectra (heme protein in solvent vs. solvent and heme protein and ligand in solvent vs. solvent) can then be compared and differences noted.

A difference spectrum is faster than the preceding method because only one spectral recording is necessary. Two cuvettes are prepared in the following manner. The reference cuvette contains heme protein and solvent, whereas the sample cuvette contains heme protein, solvent, and ligand. There must be equal concentrations of the heme protein in the two cuvettes. (Why?) The two cuvettes are placed in a double-beam spectrophotometer and the spectrum is recorded. If the spectrum of the heme protein is not influenced by the ligand, the result would be a zero difference spectrum, that is, a straight line (Figure 7.10A). Both samples have identical spectral properties, indicating that there is probably little or no interaction between heme protein and ligand. However, such data should be treated with caution because it is possible that the heme group is not affected by ligand binding. A nonzero difference spectrum indicates that the ligand interacts with the heme protein and induces a change in the environment of the heme group (Figure 7.10B).

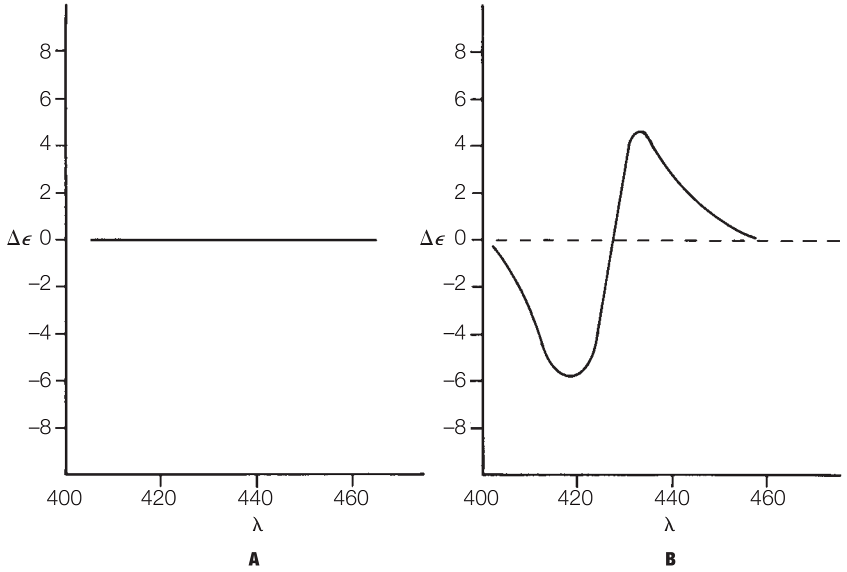

FIGURE 7.10 Difference spectroscopy. **A** Hemoglobin vs. hemoglobin. **B** Hemoglobin vs. hemoglobin + inositol hexaphosphate.

A difference spectrum can be analyzed and used in several ways. This is a useful technique for demonstrating qualitatively whether an interaction occurs between a macromolecule and a ligand. Quantitative analysis of difference spectra requires measurement of λ_{max} and ΔA at λ_{max}. This method can be used to quantify the strength of ligand-protein interaction. Note that every time you record a spectrum in a double-beam spectrometer, you are obtaining a difference spectrum between sample and reference. Although a heme protein was used in this example, this does not imply that only heme interactions can be characterized by difference spectroscopy. Any protein that contains a chromophoric group, whether it be an aromatic amino acid, cofactor, prosthetic group, or metal ion, can be studied by difference spectroscopy.

Limitations and Precautions in Spectrophotometry

The use of a spectrophotometer is relatively straightforward and can be mastered in a short period of time. There are, however, difficulties that must be considered. A common problem encountered with biochemical measurements is turbidity, or cloudiness, of biological samples. This can lead to great error in absorbance measurements because much of the light entering the cuvette is not absorbed but is scattered. This causes artificially high absorbance readings. Occasionally, absorbance readings on turbid solutions are desirable (as in measuring the rate of bacterial growth in a culture), but in most cases turbid solutions must be avoided or clarified by filtration or centrifugation.

A difficulty encountered in measuring the concentration of an unknown absorbing species in solution is deviation from the Beer-Lambert law. For reasons stated earlier in this chapter, some absorbing species do not demonstrate an increase in absorbance that is proportional to an increase in concentration. (In reality, most compounds follow the Beer-Lambert relationship over a relatively small concentration range.) When measuring solution concentration, adherence to the Beer-Lambert law must always be tested in the concentration range under study.

B. FLUORESCENCE SPECTROMETRY

Principles

In our discussion of absorption spectroscopy, we noted that the interaction of photons with molecules resulted in the promotion of valence electrons from ground state orbitals to higher energy level orbitals. The molecules were said to be in an excited state.

Molecules in the excited state do not remain there long, but spontaneously relax to the more stable ground state. With most molecules, the relaxation process is brought about by collisional energy transfer to solvent or other molecules in the solution. Some excited molecules, however, return to the ground state by emitting the excess energy as light. This process, called **fluorescence,** is illustrated in Figure 7.11. The solid vertical arrow in the figure indicates the photon absorption process in which the molecule is excited from *G* to some vibrational level in *S*. The excited molecule loses vibrational energy by collision with solvent and ground state molecules. This relaxation process, which is very rapid, leaves the molecule in the lowest vibrational level of *S*, as indicated by the wavy arrow. The molecule may release its energy in the form of light (fluorescence, dashed arrow) to return to some vibrational level of *G*.

Two important characteristics of the emitted light should be noted: (1) The emitted light is of longer wavelength (lower energy) than the excitation light. This is because part of the energy initially associated with the *S* state is lost as heat energy, and the energy lost by emission may be sufficient only to return the excited molecule to a higher vibrational level in *G*. (2) The emitted light is composed of many wavelengths, which results in a fluorescence spectrum as shown in Figure 7.12. This is due to the fact that fluorescence from any particular excited molecule may return the molecule to one of many vibrational levels in the ground state. Just as in the case of an absorption spectrum, a wavelength of

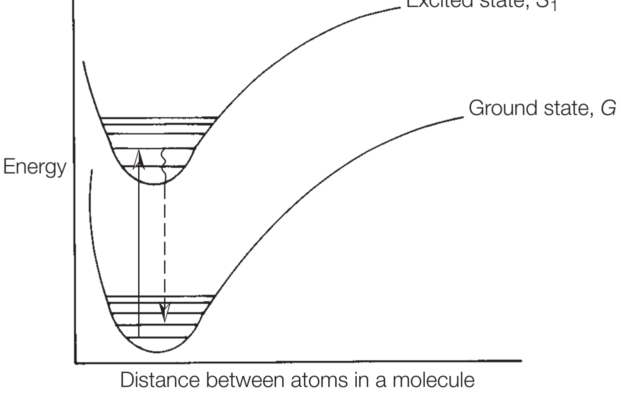

FIGURE 7.11 Energy-level diagram outlining the fluorescence process. See text for details.

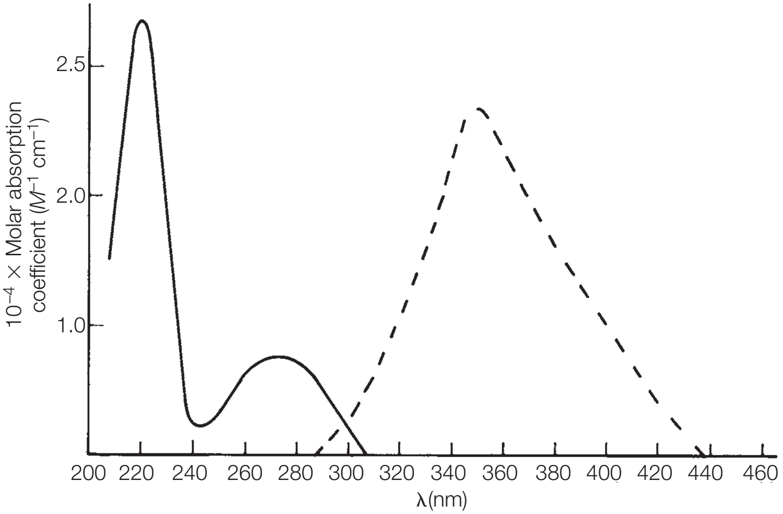

FIGURE 7.12 Absorption (—) and fluorescence (– – –) spectra of the amino acid tryptophan.

maximum fluorescence is observed, and the spectrum is composed of a wavelength distribution centered at this emission maximum.

Quantum Yield

In the foregoing discussion, it was pointed out that a molecule in the excited state can return to lower energy levels by collisional transfer or by light emission. Since these two processes are competitive, the **fluorescence intensity** of a fluorescing system depends on the relative importance of each process. The fluorescence intensity is often defined in terms of **quantum yield,** represented by Q. This describes the efficiency or probability of the fluorescence process. By definition, Q is the ratio of the number of photons emitted to the number of photons absorbed (Equation 7.7).

>> $$Q = \frac{\text{number of photons emitted}}{\text{number of photons absorbed}} \qquad \textbf{(Eq. 7.7)}$$

Measurement of quantum yield is often the goal in fluorescence spectroscopy experiments. Q is of interest because it may reveal important characteristics of the fluorescing system. Two types of factors affect the intensity of fluorescence, internal and external (environmental) influences. Internal factors, such as the number of vibrational levels available for transition and the rigidity of the molecules, are associated with properties of the fluorescent molecules themselves. Internal factors will not be discussed in detail here because they are of more interest in theoretical studies. The external factors that affect Q are of great interest to biochemists because information can be obtained about macromolecule conformation and molecular interactions between small molecules (ligands) and larger biomolecules (proteins, nucleic acids). Of special value is the study of experimental conditions that result in **quenching** or **enhancement** of the quantum yield. Quenching in biochemical systems can be caused by chemical reactions of the fluorescent species with added molecules, transfer of energy to other molecules by collision (actual contact between molecules), and transfer of energy over a distance (no contact, resonance energy transfer). The reverse of quenching, enhancement of fluorescent intensity, is also observed in some situations. Several fluorescent dye molecules are quenched in aqueous solution, but their fluorescence is greatly enhanced in a nonpolar or rigidly bound environment (the interior of a protein, for example). This is a convenient method for characterizing ligand binding. Both fluorescence quenching and fluorescence enhancement studies can yield important information about biomolecular structure and function.

Instrumentation

The basic instrument for measuring fluorescence is the **spectrofluorometer.** It contains a light source, two monochromators, a sample holder, and a detector. A typical experimental arrangement for fluorescence measurement is shown in Figure 7.13. The setup is similar to that for absorption measurements, with two significant exceptions. First, there are two monochromators, one for selection of the excitation wavelength and another for wavelength analysis of the emitted light. Second, the detector is at an angle (usually 90°) to the excitation beam. This is to eliminate interference by the light that is transmitted through the sample. Upon excitation of the sample molecules, the fluorescence is emitted in all directions and is detected by a photocell at right angles to the excitation light beam.

The lamp source used in most instruments is a xenon arc lamp that emits radiation in the ultraviolet, visible, and near-infrared regions (200 to 1400 nm). The light is directed by an optical system to the excitation monochromator, which allows either preselection of a wavelength or scanning of a certain wavelength range. The exciting light then passes into the sample chamber, which contains a fluorescence cuvette with dissolved sample. Because of the geometry of the optical system, a typical fused absorption cuvette with two opaque sides cannot be used; instead, special fluorescence cuvettes with four translucent

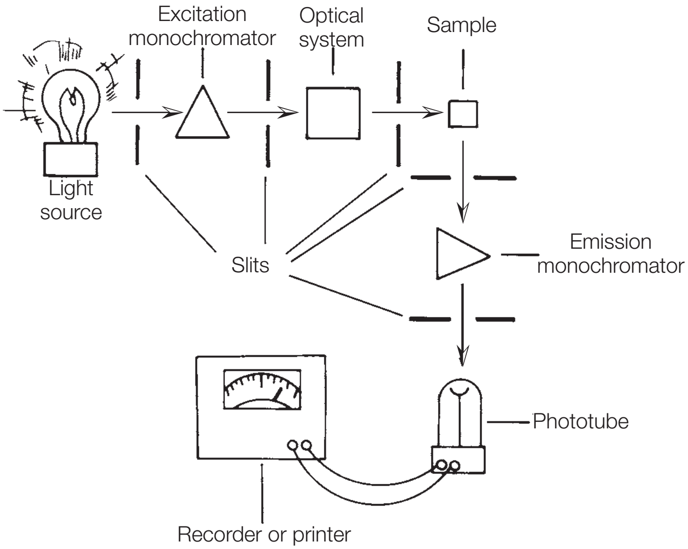

FIGURE 7.13 A diagram of a typical spectrofluorometer.

quartz or glass sides must be used. When the excitation light beam impinges on the sample cell, molecules in the solution are excited and some will emit light.

Light emitted at right angles to the incoming beam is analyzed by the emission monochromator. In most cases, the wavelength analysis of emitted light is carried out by measuring the intensity of fluorescence at a preselected wavelength (usually the wavelength of emission maximum). The analyzer monochromator directs emitted light of only the preselected wavelength toward the detector. A photomultiplier tube serves as a detector to measure the intensity of the light. The output current from the photomultiplier is fed to some measuring device that indicates the extent of fluorescence. The final readout is not in terms of Q, but in units of the photomultiplier tube current (microamperes) or in relative units of percent of full scale. Therefore, the scale must be standardized with a known. Some newer instruments provide, as output, the ratio of emitted light to incident light intensity.

Applications of Fluorescence Spectroscopy

Two types of measurements are most common in fluorescence experiments, measurements of **relative fluorescence intensities** and measurements of the **quantum yield.** Most experiments require only relative fluorescence intensity measurements, and they proceed as follows. The fluorometer is set to "zero" or "full scale" fluorescence intensity (microamps or %) with the desired biochemical system under standard conditions. Some perturbation is then made in the system (pH change, addition of a chemical agent in varying concentrations, change of ionic strength, etc.), and the fluorescence intensity is determined relative to the standard conditions. This is a straightforward type of experiment because it consists of replacing one solution with another in the fluorometer and reading the detector output for each. For these experiments, the excitation wavelength and the emission wavelength are preselected and set for each monochromator.

The measurement of quantum yield is a more complicated process. Before these measurements can be made, the instrument must be calibrated. A thermopile or chemical actinometer may be used to measure the absolute intensity of incident light on the sample. Alternatively, quantum yields may be measured relative to some accepted standard. Two commonly used fluorescence standards are quinine sulfate in $0.5\,M\ H_2SO_4$ ($Q = 0.70$) and fluorescein in $0.1\,M$ NaOH ($Q = 0.93$). The quantum yield of the unknown, Q_x, is then calculated by Equation 7.8.

>>

$$\frac{Q_x}{Q_{std}} = \frac{F_x}{F_{std}} \qquad \textbf{(Eq. 7.8)}$$

where

Q_x = quantum yield of unknown

Q_{std} = quantum yield of standard

F_x = experimental fluorescence intensity of unknown

F_{std} = experimental fluorescence intensity of standard

Some biomolecules are **intrinsic fluors;** that is, they are fluorescent themselves. The amino acids with aromatic groups (phenylalanine, tyrosine, and tryptophan) are fluorescent; hence, proteins containing these amino acids have intrinsic fluorescence. The purine and pyrimidine bases in nucleic acids (adenine, guanine, cytosine, uracil, thymine) and some coenzymes (NAD^+, FAD) are also intrinsic fluors. Intrinsic fluorescence is most often used to study protein conformational changes (protein folding) and to probe the location of active sites and coenzymes in enzymes.

Valuable information can also be obtained by the use of **extrinsic fluors.** These are fluorescent molecules that are added to the biochemical system under study. Many fluorescent dyes have enhanced fluorescence when they are in a nonpolar solution or bound in a rigid hydrophobic environment. Some of these dyes bind to specific sites on proteins or nucleic acid molecules, and the resulting fluorescence intensity depends on the environmental conditions at the binding site. Extrinsic fluorescence is of value in characterizing the binding of natural ligands to biochemically significant macromolecules. This is because many of the extrinsic fluors bind in the same sites as natural ligands. Extrinsic fluorescence has been used to study the binding of fatty acids to serum albumin, to characterize the binding sites for cofactors and substrates in enzyme molecules, to characterize the heme binding site in various hemoproteins, and to study the intercalation of small molecules into the DNA double helix.

Figure 7.14 shows the structures of extrinsic fluors that have been of value in studying biochemical systems. ANS, dansyl chloride, and fluorescein are used for protein studies, whereas ethidium, proflavine, and various acridines are useful for nucleic acid characterization. Ethidium bromide has the unique characteristic of enhanced fluorescence when bound to double-stranded DNA but not to single-stranded DNA. Aminomethyl coumarin (AMC) is of value as a fluorogenic leaving group in measuring peptidase activity.

Difficulties in Fluorescence Measurements

Fluorescence measurements have much greater sensitivity than absorption measurements. Therefore, the experimenter must take special precautions in making fluorescence measurements because any contaminant or impurity in the system can lead to inaccurate results. The following factors must be considered when preparing for a fluorescence experiment.

Preparation of Reagents and Solutions

Since fluorescence measurements are very sensitive, dilute solutions of biomolecules and other reagents are appropriate. Special precautions must be taken to maintain the integrity of these solutions. All solvents and reagents must be checked for the presence of fluorescent impurities, which can lead to large errors in measurement. "Blank" readings should be taken on all solvents and solutions, and any background fluorescence must be subtracted from the fluorescence of the complete system under study. Solutions should be stored in the dark, in clean glass-stoppered containers, in order to avoid photochemical breakdown of the reagents and contamination by corks and rubber stoppers. Some biomolecules, especially proteins, tend to adsorb to glass surfaces, which can lead to loss of

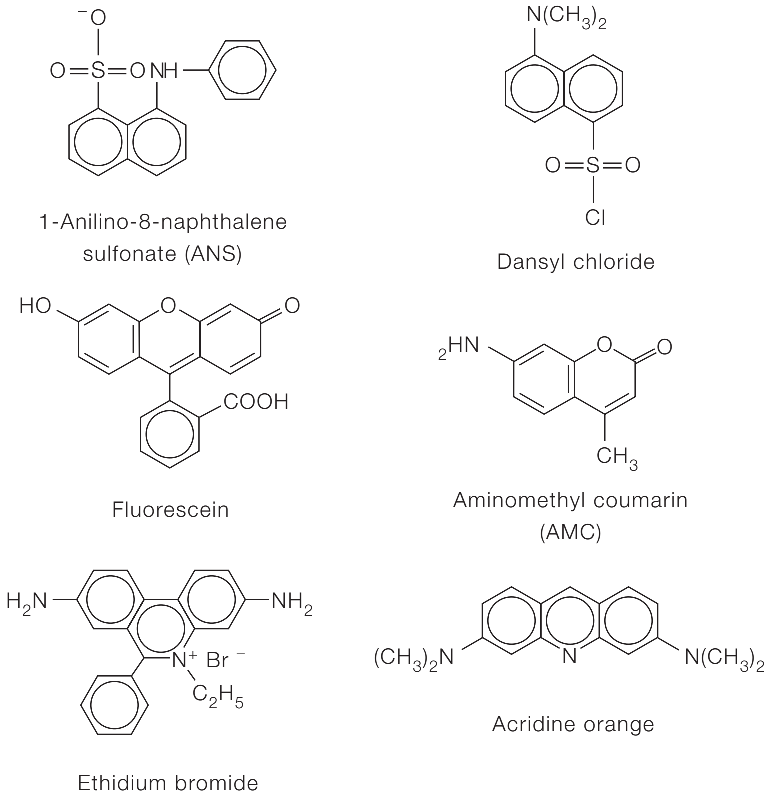

1-Anilino-8-naphthalene sulfonate (ANS)

Dansyl chloride

Fluorescein

Aminomethyl coumarin (AMC)

Ethidium bromide

Acridine orange

FIGURE 7.14 Molecules useful as extrinsic fluors in biochemical studies.

fluorescent material or to contamination of fluorescence cuvettes. All glassware must be scrupulously cleaned. Turbid solutions must be clarified by centrifugation or filtration.

Control of Temperature

Fluorescence measurements, unlike absorption, are temperature dependent. All solutions, especially if relative fluorescent measurements are taken, must be thermostated at the same temperature.

C. NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

Nuclear magnetic resonance (NMR) spectroscopy is a method of absorption spectroscopy that has some characteristics similar to ultraviolet and visible spectroscopy, but also some that are unique. In NMR, a molecular sample, usually dissolved in a liquid solvent, is placed in a magnetic field and the absorption of radio-frequency waves by certain nuclei (protons and others) is measured. NMR spectroscopy was originally developed as an analytical tool to determine molecular structure by monitoring the environment of individual protons. The relative positions and intensity of absorption signals provide detailed information from which chemical structures may be elucidated. Early proton NMR studies focused primarily on small organic molecules, some of which had biological origin and significance. Since most biomolecules have a very large number of protons, the observed spectra are extremely complex and difficult to interpret. With the advent of modern techniques—superconducting magnets, Fourier transform analysis, multidimensional spectra, and powerful computer control—NMR has become an important method for the study of biological macromolecules. The fundamental principles, instrumentation, and biochemical applications for NMR will be outlined here.

NMR Theory

All nuclei possess a positive charge. For some nuclei, this charge confers the property of spin, which causes the nuclei to act like tiny magnets. The angular momentum of the spin is described by the quantum spin number I. If I is an integral number ($I = 0, 1, 2$, etc.), then there is no net spin and no NMR signal for that nucleus. However, if I is half-integral ($I = {}^1/_2, {}^3/_2, {}^5/_2$, etc.), the nuclei have spins, and when placed in a magnetic field, the spins orient themselves with (parallel) or opposed (antiparallel) to the external magnetic field. The nuclei aligned with the magnetic field have lower energy (are more stable) than those opposed. Energy in the radio-frequency range is sufficient to flip the nuclei from the parallel to the antiparallel alignment. The NMR instrument is designed to measure the energy difference between the nuclear spin states. Absorption of energy may be detected by scanning the radio-frequency range and measuring the absorption that causes spin state transition (resonance). Modern NMR instruments instead maintain a constant radio-frequency and electrically induce small changes in the strength of the magnetic field until resonance is attained. The point of resonance for a nucleus is dependent upon the electronic environment of that nucleus, so an NMR spectrum provides information that helps elucidate biochemical structures.

NMR in Biochemistry

NMR has found a wide variety of applications in biochemistry. Proton NMR has a long and rich history in organic chemistry and biochemistry. The structures of many small but significant biomolecules were elucidated by proton NMR. Protons on different atoms and in different molecular environments absorb energy of different levels (measured by radio-frequency units or magnetic field units). Two experimentally measured characteristics, **chemical shifts** and **spin-spin coupling,** provide important structural information. The chemical shift (δ) of an absorbing nucleus, measured in parts per million (ppm), is the spectral position of resonance relative to a standard signal, usually tetramethylsilane. The NMR signal for a proton is "split" by interactions with neighboring protons. This characteristic, called spin-spin coupling, helps to determine positions and numbers of equivalent and nonequivalent protons.

TABLE 7.1 Nuclei Important in Biochemical NMR

Isotope[1]	Natural Abundance (%)[2]	Relevant Biomolecules
^{1}H	100	Most biomolecules
^{13}C	1	Most biomolecules
^{15}N	0.4	Amino acids, proteins, nucleotides
^{19}F	100	Substitute for H
^{31}P	100	Nucleotides, nucleic acids, and other phosphorylated compounds

[1] All have a spin of ½.

[2] The percentage of this isotope in naturally occurring molecules containing this element.

NMR experiments are not limited to the study of protons. Resonance signals from other atomic nuclei including ^{13}C, ^{15}N, ^{19}F, and ^{31}P can be detected and measured (see Table 7.1). ^{13}C NMR techniques have been especially valuable for the study of carbohydrate, amino acid, and fatty acid structures. Just as with protons, each distinct carbon atom in a molecule yields a signal that is split by neighboring interacting nuclei (see Figure 7.15). Because of the low natural abundance of the ^{13}C isotope, it is necessary either to enrich the ^{13}C content by chemical synthesis or to use powerful computers and magnets (500–600 MHz). The ^{31}P spectra of phosphorylated biomolecules display a peak for each type of phosphorus (Figure 7.16). NMR instruments now have the sensitivity necessary to measure the *in vivo* concentrations and reactions of biomolecules in cells. For example, the catabolism of glucose by glycolysis in erythrocytes has been monitored by ^{13}C NMR, and the involvement of ATP in phosphoryl-group transfer processes has been studied by ^{31}P NMR (see Figure 7.16).

NMR and Protein Structures

New computer techniques in data analysis and improvements in instrumentation have now made it possible to carry out detailed structural and conformational studies of all biopolymers, but especially proteins. NMR, which may be done on noncrystalline materials in solution, provides a technique complementary to X-ray diffraction, which requires crystals for analysis. An advantage of NMR is that proteins can be studied in solution in an environment that mimics that in the living cell. One-dimensional NMR, as described to this point, can be applied to structural analysis of smaller molecules. But proteins and other complex biopolymers with large numbers of protons will yield very crowded spectra with many overlapping lines. In multidimensional NMR (2-D, 3-D, 4-D), peaks are spread out through two or more axes to improve resolution. Protein structure determination depends on several homonuclear 2-D ^{1}H NMR experiments: **correlation spectroscopy (COSY), nuclear Overhauser effect spectroscopy (NOESY), 2-D total correlation spectroscopy (TOCSY), foldover-corrected spectroscopy (FOCSY), spin-echo correlation spectroscopy (SECSY),** and others. These computational techniques are based on

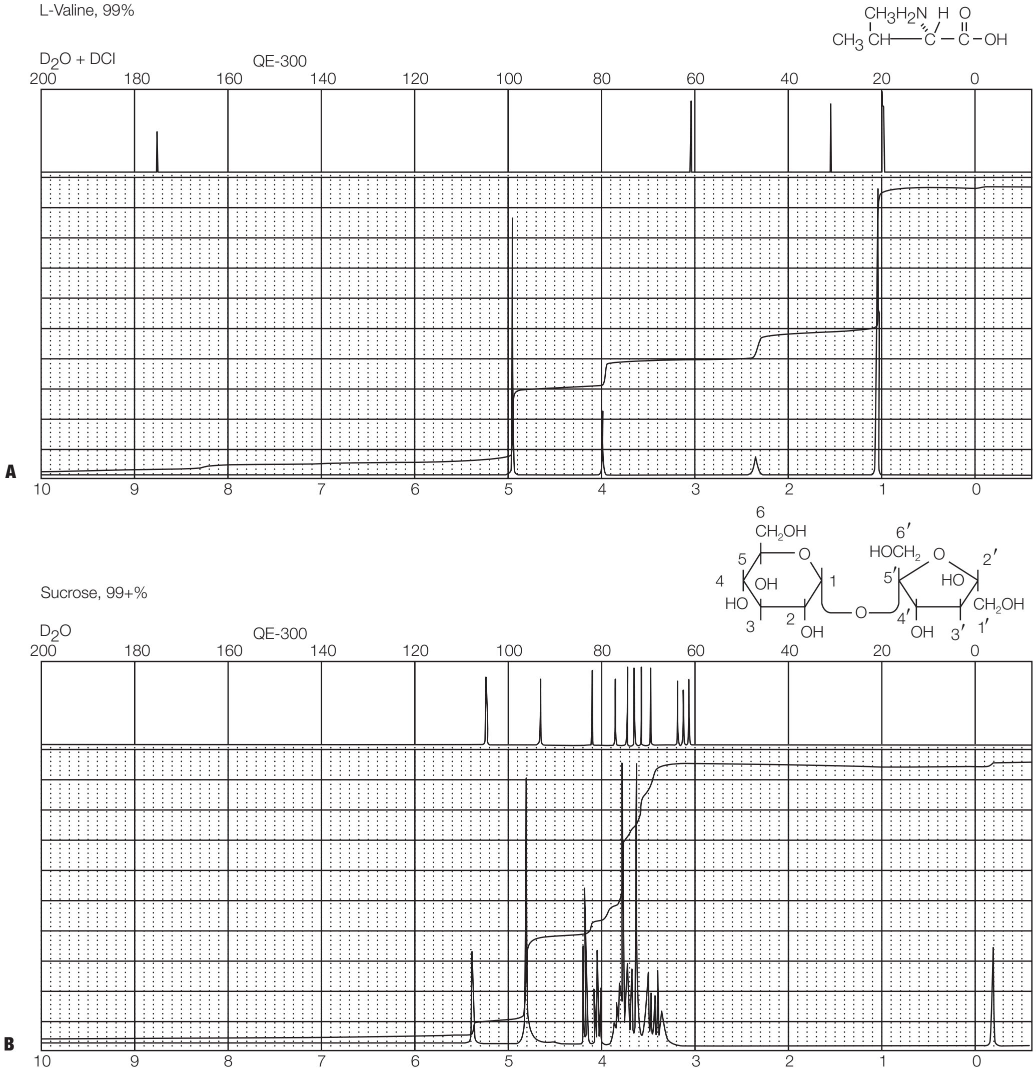

FIGURE 7.15 ^{1}H and ^{13}C FT-NMR spectra of biomolecules. The lower spectra are for ^{1}H showing a ppm scale of 0–10. TMS standard is at 0 ppm. The upper spectra are for ^{13}C showing a ppm scale of 0–200. **A.** L-valine in D_2O + DCl. Can you assign each peak to the correct protons and carbon atoms in the valine structure? Hint: The carboxyl carbon of valine has a peak at about 175 ppm. **B.** Sucrose in D_2O. Carbon numbers in the chemical structure correspond to the following peaks in order from 0 to 120 ppm: C-6; C-1'; C-6 '; C-4; C-2; C-5; C-3; C-4 '; C-3 '; C-5 '; C-1; C-2 '. *Reprinted with permission of Aldrich Chemical Co., Inc.*

the observation that nonequivalent protons interact with each other. By using multiple-pulse techniques, it is possible to perturb one nucleus and observe the effects on the spin states of other nuclei. The availability of powerful computers, strong magnets, and Fourier transform (FT) calculations makes it possible to elucidate structures of proteins up to one million daltons, and there is future promise for studies on larger proteins. In addition to structural studies, NMR

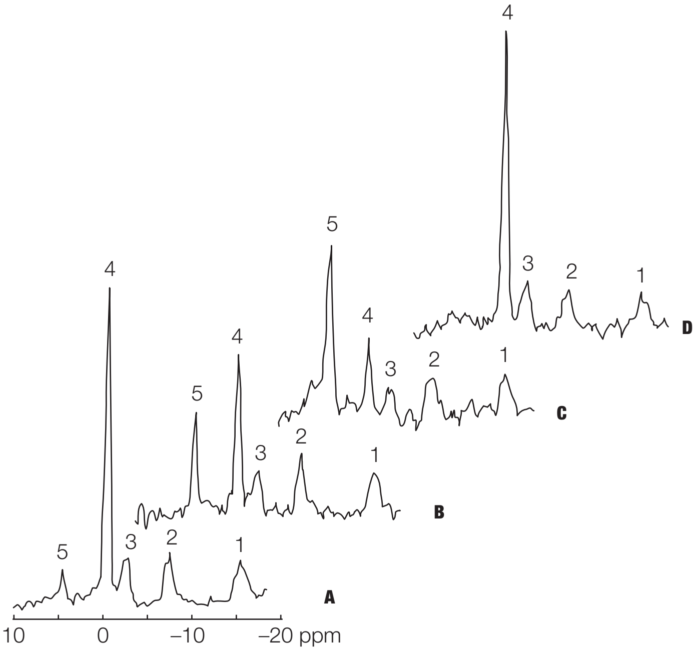

FIGURE 7.16 ^{31}P NMR spectra of human forearm muscle showing the effect of exercise. **A** Before exercise; **B** and **C** during 19 minutes of exercise; **D** 5–6 minutes after **C**. Peak assignments: 1, β-phosphorus of ATP; 2, α-phosphorus of ATP; 3, γ-phosphorus of ATP; 4, phosphocreatine; 5, Pi. Phosphocreatine is used as a major source of energy during exercise. It is hydrolyzed to creatine and Pi. Note that the level of ATP remains relatively constant during exercise because it is produced and used at about the same rate. *After G. Radda, Science* **233,** *641 (1986). Reprinted with permission from the American Association for the Advancement of Science.*

will also be applied to studies of conformational changes and interactions between ligands and biopolymers.

Many of the instrumental techniques and computational methods for NMR protein structure determination have been developed by Kurt Wüthrich of the Swiss Federal Institute of Technology (ETH) in Zurich. Beginning in the early 1980s, Wüthrich and his colleagues used NMR to determine the structure of several smaller proteins (below 10,000 daltons). By 1990, NMR structure determination could be applied to proteins of molecular masses up to about 20,000 daltons. One of the largest proteins studied so far by NMR is the GroEL-GroES complex, which is about 900,000 daltons. Some of the interesting proteins studied by Wüthrich include tendamistat (an α-amylase inhibitor), rabbit metallothionein, human cyclophilin A-cyclosporin A complex, and the murine prion protein, which may be involved in diseases related to bovine spongiform encephalopathy (BSE). Wüthrich was awarded the Nobel Prize in Chemistry in 2002 for this work, which occurred over a period of about 25 years. It is interesting to note that independent studies comparing protein three-dimensional structures derived from X-ray diffraction (on crystals) and NMR (in solution) have yielded identical results. The NMR methods have even been used to correct structural mistakes made in the X-ray diffraction method. Protein NMR is now a very active research area and more than 7000 NMR-derived protein structures had been submitted to the Protein Data Bank (PDB) (www.rcsb.org/pdb) by the year 2010. The current total of protein structures submitted to the PDB is over 66,000.

The complex principles and techniques behind Wüthrich's studies are best described in his own words (Wüthrich, 2001):

> *"Four principal elements are combined in the NMR method for protein structure determination: (i) the nuclear Overhauser effect (NOE) as an experimentally accessible NMR parameter in proteins that can yield the information needed for de novo global fold determination of a polymer chain; (ii) sequence-specific assignment of the many hundred to several thousand NMR peaks from a protein; (iii) computational tools for the structural interpretation of the NMR data and the evaluation of the resulting molecular structures; and, (iv) multidimensional NMR techniques for efficient data collection."*

The preparation of a protein sample for NMR studies is relatively straightforward; however, a rather large amount of protein is required compared to mass spectrometry and X-ray crystallography. For NMR analysis, the protein is usually dissolved in water at concentrations of 3–6 m*M*. The protein solution may be adjusted to the ionic strength, pH, and temperature found under physiological conditions. Samples as small as 0.5 mL are sufficient for analysis.

D. MASS SPECTROMETRY

Mass spectrometry (MS) is similar to NMR in that it has historically been of great value in the structure elucidation of relatively small (MW limit around 1000) organic and biomolecules. Only in the last 15 years has the tool of MS been applied to the analysis of biological macromolecules. The MS analysis of proteins and other biopolymers was initially hindered because analytes are usually measured in the gas phase and it was very difficult to vaporize these large molecules. The development of new methods for sample preparation (ionization), multi-quadrupole analysis, tandem MS instruments, and powerful computers now makes it possible to study large molecules, especially proteins.

Ionization and Analysis of Proteins

The mass spectrometer generally consists of three components that are linked together: an **ionization device,** a **mass analyzer,** and an **ion detector** (Figure 7.17). Neutral molecules are ionized, and their positively charged ion products are directed through an electric and/or magnetic field, where they are separated (analyzed) on the basis of their mass-to-charge ratio (m/z). A detector then records the ions after separation. The "spectrum" generated by MS displays ion intensity as a function of m/z.

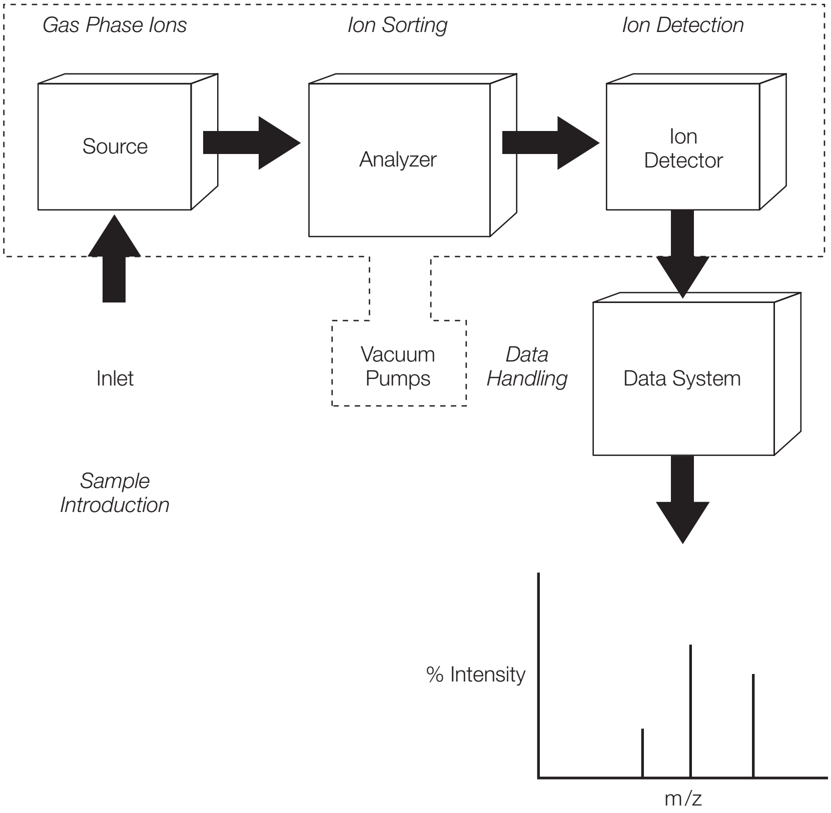

FIGURE 7.17 The components of a mass spectrometer. The final data output consists of a graph of ion intensity versus m/z. Mass spectrometry, unlike the other techniques described in this chapter, does not involve interaction of molecules with electromagnetic radiation; however, it has traditionally been combined with the spectrophotometric techniques.

Ionization of small organic and biological molecules prior to MS analysis is done by electron impact; however, ionization of nonvolatile, biological macromolecules like peptides, proteins, and nucleic acids requires special treatment. In the late 1980s, John Fenn, now at Virginia Commonwealth University, and Koichi Tanaka of the Shimadzu instrument company in Kyoto, Japan, introduced methods for the “soft ionization” of macromolecules. These methods are labeled “soft,” as they cause minimal degradation to the sample during ionization. Fenn introduced the **electrospray ionization (ESI)** method, in which a solution of the sample is sprayed from a metal needle or capillary tip held at a potential of about +5000 volts. This results in a spray of tiny droplets containing positively charged ions. The solvent evaporates, and the ions are directed into the analyzer for separation. Tanaka demonstrated that pulses from a nitrogen laser could be used to ionize proteins from a surface **(soft laser desorption, SLD).** Because ESI and SLD are liquid-phase processes, the samples may be those purified by HPLC or CE. Fenn and Tanaka were awarded the Chemistry Nobel Prize (shared with K. Wüthrich) for this work in 2002.

Laser desorption has now been modified by Franz Hillenkamp of the University of Münster, Germany, to a solid phase method, **matrix-assisted laser desorption ionization (MALDI),** which is the most widely used technique for protein analysis. In MALDI, the sample is placed in a matrix of small organic molecules and irradiated by a laser pulse. The matrices absorb the laser energy, causing ionization of the macromolecular sample. A common practice is the use of protein samples obtained from 1-D and 2-D PAGE gels.

Several types of mass analyzers are used for ion separation, including time-of-flight (TOF), quadrupoles, and ion traps. In TOF-MS, the most widely used technique, each ion produced has the same initial kinetic energy, but the speed varies with mass. Mass is determined by measurement of an ion's time-of-flight to the detector.

MS Applications in Biochemistry

MS is becoming a standard tool in the analysis of biological molecules and biological processes. Many of the applications of MS described below involve combining the technique with sample preparation and purification by HPLC, CE, and PAGE (especially 2-DE). (See section on proteomics, Chapter 6, Section B, p. 184.)

1. ***Identification of Peptides and Proteins*** Protein identification is perhaps the most widely used MS application at the present time. The most important ions resulting from the removal of an electron by ionization procedures are the positively charged **molecular ions** $(\mathbf{M}^{+})$ or **protonated molecular ions** $(\mathbf{M} + nH)^{\mathbf{n}-}$. Measurement of the molecular mass of these species provides the molecular weight of the original molecule. The accuracy of this method for molecular weight determination is about 0.01%. (Molecular mass measurements using gel filtration typically have accuracies no greater than 5–10%.) Some molecular ions are unstable and disintegrate to produce fragment ions. These fragmentation processes are useful in structural elucidation of smaller molecules. It should be expected that proteomic studies will be enhanced by coupling SDS-PAGE and 2-DE (for separation and purification) with further size analysis of the proteins by MALDI-TOF.

   One MS method for peptide identification is called **peptide mass fingerprinting.** The unknown sample is digested with the proteolytic enzyme trypsin to produce fragments that are analyzed by MALDI-MS. The resulting spectrum displays the masses of the peptide fragments. This may be used as a "fingerprint" of the sample and compared with known fragmented patterns in a sequence database.
2. ***Characterization of Post-Translational Modification Processes*** After translation, proteins in the cell are often modified by covalent attachment of specific functional groups. Such processes might include:

   (a) **phosphorylation** of hydroxyl groups in serine and tyrosine residues, or
   (b) addition of carbohydrate residues **(glycosylation)** to the hydroxyl groups of serine or threonine or to the amide nitrogen of asparagine residues.

   These chemical modifications can easily be detected by MS, as the addition of a phosphoryl group (—PO_3^{2-}) adds about 80 daltons to the molecular mass of a protein, whereas addition of a glucose residue adds about 180 daltons.

3. ***Peptide Sequencing*** Small peptides and proteins may be sequenced by MS procedures. The peptide sample is ionized and fragmented at the peptide bonds by collision-induced dissociation. Size analysis of the series of fragments produced leads to the sequence of amino acids.
4. ***Protein-protein Interactions*** One of the most important goals of proteomics is to study how proteins interact with each other in order to initiate biological processes. ESI-MS and MALDI-TOF are currently being used to investigate how proteins associate to form biologically active multisubunit complexes.

E. X-RAY CRYSTALLOGRAPHY

Historically, **X-ray crystallography** has been the most important tool for studying the three-dimensional structures of biomolecules. In fact, this technique was the only method available for determining macromolecular structures until the application of NMR in the 1980s. Protein crystals were first studied by X-ray crystallography in 1934 by J. D. Bernal and Dorothy Crowfoot Hodgkin. However, the diffraction patterns were extremely complex, and computers were not available for complete analysis of the proteins. John Kendrew, a biochemist working at Cambridge University, announced the first three-dimensional structure of a protein, sperm whale myoglobin, in 1958. Kendrew had determined the complex structure of the protein (153 amino acid residues) after several years of analysis by X-ray crystallography. Other important biomolecules that were studied by X-ray crystallography include DNA (Rosalind Franklin and Maurice Wilkins) and vitamin B_{12} (Dorothy Crowfoot Hodgkin). Analysis of protein structures is now routine. As of July 2010, the Protein Data Bank (www.rcsb.org/pdb) had registered approximately 66,000 proteins—about 56,000 were obtained by X-ray crystallography and about 10,000 by NMR. Although the number of structures derived by NMR is growing faster than the number derived by X-ray crystallography, crystallography is still an important source of structural data.

Methodology of X-ray Crystallography

An X-ray crystallography analysis requires three components (see Figure 7.18): (1) a protein crystal, (2) an X-ray source, and (3) a detector (i.e., radiation detector or photographic film).

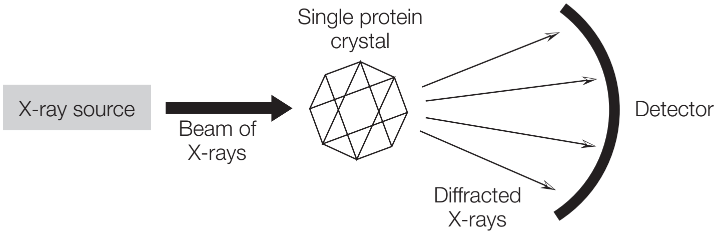

FIGURE 7.18 The components required for analysis of protein structure by X-ray crystallography.

Protein Crystals

For many proteins, this can be the most difficult step in three-dimensional analysis. The factors important in protein crystallization are not yet fully understood, so the methodology for growing crystals is still an art and involves much trial and error. Crystallization usually occurs best in a saturated or supersaturated solution of the protein. A common method often used involves changing solution conditions by adding a precipitant (like ammonium sulfate), changing the pH, or changing the ionic strength. The crystallization process may be hastened by dialysis of the protein against a solution containing the precipitating factor.

Data Collection and Analysis

The theory supporting X-ray crystallography is quite complex and beyond the scope of this text; however, some practical aspects can be described. The crystal is mounted in the diffractometer and bombarded by a collimated beam of X-rays. The crystal is rotated so that it is struck from many different directions. Many of the X-rays passing into the sample are diffracted (scattered) when they encounter electrons associated with atoms. The diffracted beams impinge upon the detector and are recorded. The data at this point are in the form of a regular array of spots called reflections. (Kendrew's pattern of myoglobin had nearly 25,000 reflections, which were analyzed by computer in order to construct the 3-D image of the protein.) The extent of scatter of the reflections depends on the size and position of each atom in the crystal. By extensive computer and mathematical analysis (by Fourier transform) of the angle of scatter and of the pattern collected by the detector, it is possible to construct an **electron density map** of the protein molecule showing the arrangement of the atoms. Using modern equipment and computer software, an analyst can collect data and determine a structure in a day or two.

Although X-ray analysis of proteins is slowly being replaced by NMR, there will continue to be interest in and a need for X-ray analysis of very large and unsymmetrical proteins and other biomolecules. NMR does have some advantages in structure determination—sample preparation is rapid and straightforward, and the sample may be studied under physiological-like conditions. Numerous comparative studies have confirmed that the analysis of an identical protein by X-ray and NMR leads to the same three-dimensional result, even though one is done in the solid state and one in the liquid state.

Study Problems

1. For each pair of wavelengths listed below, specify which one is higher in energy.
   (a) 1 nm (X-ray) or 10,000 nm (IR)
   (b) 280 nm (UV) or 360 nm (VIS)
   (c) 200,000 nm (microwave) or 800 nm (VIS)
2. In a laboratory experiment, you are asked to determine the molar concentration of a solution of an unknown compound, X. The solution diluted in half by water (1 mL of X and 1 mL of H_2O) has an absorbance at 425 nm of 0.8 and a molar extinction coefficient of $1.5 \times 10^3 M^{-1}\,\text{cm}^{-1}$. What is the molar concentration of the original solution of X?

3. Match the spectral region listed below with the appropriate molecular transition that occurs. The first problem is worked as an example.

Spectral region	**Transition**
_a_ 1. X-rays	a. inner-shell electrons
__2. UV	b. molecular rotations
__3. VIS	c. valence electrons
__4. IR	d. molecular vibrations
__5. microwave	

4. Explain the differences between a spectrophotometer that uses a phototube for a detector and one that uses a photodiode array detector.

5. Several spectroscopic techniques were studied in this chapter. Which experimental techniques involve an actual measurement of radiation absorbed?
(a) UV-VIS spectroscopy
(b) NMR spectroscopy
(c) MS
(d) Fluorescence spectroscopy

6. Why can you not use a glass cuvette for absorbance measurements in the UV spectral range?

7. What single structural characteristic do all of the fluorescent molecules in Figure 7.14 possess?

8. Study the ^{1}H spectrum of valine in Figure 7.15 and match each peak to the corresponding proton in the chemical structure. Explain any spin-spin coupling.

9. Why must a cuvette with four translucent sides be used for fluorescence measurements?

10. Study the ^{13}C spectrum of valine in Figure 7.15. Identify the carbon atom that produces each peak in the spectrum.

Further Reading

C. Arnaud, ***Chem. Eng. News,*** October 2, 17–25 (2006), “Mass Spec Tackles Proteins.”

J. Berg, J. Tymoczko, and L. Stryer, ***Biochemistry,*** 6th ed. (2007), Freeman and Company (New York). Application of NMR, X-ray, and MS to biochemistry.

D. Beussman et al., ***Biochem. Mol. Biol. Educ.* 37,** 49–55 (2008). “Identifying gel-separated proteins using in-gel digestion, mass spectrometry, and database searching.”

S. Borman, ***Chem. Eng. News,*** December 3, 11 (2007), “Probing Large Protein Systems.”

A. Burlingame, S. Carr, and M. Baldwin, ***Mass Spectrometry in Biology and Medicine*** (2000), Humana Press (Totowa, NJ).

J. Cavanagh et al., ***Protein NMR Spectroscopy,*** 2nd ed. (2006), Academic Press (Burlington, MA).

E. De Hoffmann, and V. Stroobant, ***Mass Spectrometry: Principles and Applications,*** 2nd ed. (2001), John Wiley & Sons (New York).

E. Eberhardt et al., ***Biochem. Mol. Biol. Educ.* 31,** 402–409 (2003). “Preparing Undergraduates to Participate in the Post-Genome Era: A Capstone Laboratory Experience in Proteomics.”

R. Ekman (Editor) et al. ***Mass Spectrometry: Instrumentation, Interpretation and Applications,*** (2008), John Wiley (Hoboken, NJ).

F. Finehout and K. Lee, ***Biochem. Mol. Biol. Educ* 32,** 93–100 (2004). "An Introduction to Mass Spectrometry Applications in Biological Research."

R. Garrett and C. Grisham, ***Biochemistry,*** 4th ed. (2010), Brooks/Cole (Belmont, CA), Introduction to UV-VIS, NMR, and MS.

B. Giles et al., ***J. Chem. Educ.* 76,** 1564–1566 (1999). "An In Vivo ^{13}C NMR Analysis of the Anaerobic Yeast Metabolism of 1-^{13}C-Glucose."

G. Hammes, ***Spectroscopy for the Biological Sciences,*** (2005), Wiley-Interscience (New York).

M. Kinter and N. Sherman, ***Protein Sequencing and Identification Using Tandem Mass Spectrometry*** (2000), Wiley-Interscience (New York).

N. Krishna and L. Berliner, ***Protein NMR for the Millennium*** (2003), Plenum (New York). Advances in protein NMR.

M. Mann and A. Pandley, ***Trends Biochem. Sci.* 26,** 54–61 (2001). "Use of Mass Spectrometry-Derived Data to Annotate Nucleotide and Protein Sequence Databases."

M. Mann et al., ***Ann. Rev. Biochem.* 70,** 437–473 (2001). "Analysis of Proteins and Proteomes by Mass Spectrometry."

D. Nelson, and M. Cox, ***Lehninger Principles of Biochemistry,*** 5th ed. (2009), Freeman (New York). Introduction to spectroscopy.

A. Pavia, G. Lampman, and G. Kriz, ***Introduction to Spectroscopy,*** 3rd ed. (2000), Brooks/Cole (Pacific Grove, CA). An introductory textbook.

K. Peterman, K. Lentz, and J. Duncan, ***J. Chem. Educ.* 75,** 1283–1284 (1998). "A ^{19}F NMR Study of Enzyme Activity."

N. Price, R. Dwek, M. Wormald, and R. Ratcliffe, ***Principles and Problems in Physical Chemistry for Biochemists,*** 3rd ed. (2002), Oxford University Press (Oxford). An introduction to spectroscopy with theory and applications.

G. Radda, ***Science,* 233,** 640–645 (1986). "The Use of NMR Spectroscopy for the Understanding of Disease."

D. Sheehan, ***Physical Biochemistry: Principles and Applications,*** 2nd ed. (2009), Wiley-Blackwell (Hoboken, NJ).

I. Tinoco, K. Sauer, J. Wang, and J. Puglisi, ***Physical Chemistry: Principles and Applications in Biological Sciences,*** 4th ed. (2002), Prentice-Hall (Upper Saddle River, NJ). Chapter 10 covers all areas of spectroscopy.

D. Voet and J. Voet, ***Biochemistry,*** 3rd ed. (2005), John Wiley & Sons (Hoboken, NJ), pp. 172–175, 240–244. NMR, X-ray, and MS in biochemistry.

D. Voet, J. Voet, and C. Pratt, ***Fundamentals of Biochemistry,*** 3rd ed. (2008), John Wiley & Sons (Hoboken, NJ). NMR, MS, and X-ray in biochemistry.

K. Wilson and J. Walker, Editors, ***Principles and Techniques of Practical Biochemistry,*** 5th ed. (2000), Cambridge University Press (Cambridge, UK). Chapter 6 covers X-ray diffraction. Chapter 9 covers atomic and molecular electronic spectroscopy. Chapters 10 and 11 cover NMR and MS.

K. Wüthrich, ***Nat. Struct. Biol* 8,** 923–925 (2001). "The Way to NMR Structures of Proteins."

K. Wüthrich, ***J. Biol. Chem.* 265,** 22059–22062 (1990). "Protein Structure Determination in Solution by NMR Spectroscopy."

http://www.scienceofspectroscopy.info/index.htm
The Science of Spectroscopy, funded by NASA.

http://www.cis.rit.edu/htbooks/nmr
The basics of NMR.

http://www.asms.org
Web site for the American Society of Mass Spectrometry.

http://www.rcsb.org/pdb
This Web site, the Protein Data Bank, is the single repository for the three-dimensional structures of proteins and nucleic acids.

http://www.spectroscopynow.com/coi/cda/home.cda?chId=0
Resource for spectroscopy.

http://biop.ox.ac.uk/www/mol_of_life/index.html
The Molecules of Life.

http://www.ebi.ac.uk/thornton-srv/software/PROCHECK
NMR for protein structures.

www.thermo.com/nanodrop
Information on nanodrop spectrometers.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

This page intentionally left blank

CHAPTER 8

BIOMOLECULAR INTERACTIONS: LIGAND BINDING AND ENZYME REACTIONS

A. Ligand-Macromolecule Interactions (Molecular Recognition) 239
B. Biological Catalysis (Enzymes) 250

Study Problems 262
Further Reading 264

Many of the dynamic processes occurring in biological cells and organisms are the result of interactions between molecules. Most often, these interactions involve the binding of a molecule or ion to a specific site (or sites) on a macromolecule, usually a protein or nucleic acid. The binding forces that hold the molecules together are noncovalent and they include hydrogen bonding, van der Waals forces, electrostatic bonds, and hydrophobic interactions. Usually the combination of the two molecules (formation of a complex) will lead to a specific biological action or response. For example, a hormonal response is the result of the hormone molecule interacting with its receptor protein; a chemical transformation in metabolism is the consequence of the initial binding of a substrate to an enzyme, forming an ES complex.

A. LIGAND-MACROMOLECULE INTERACTIONS (MOLECULAR RECOGNITION)

Throughout our study of biochemistry we will encounter many examples where noncovalent interactions bring together, in specific ways, two different molecules that form a complex. Such molecules display a phenomenon called **molecular recognition.** The importance of these interactions in biology is that the combination of two molecules will lead to some biological action. The action of hormones provides an excellent example. A hormonal response is the consequence of a weak, but specific, interaction between the molecule and a receptor protein in the membrane of the target cell. The biochemical response may often be an enhancement in production of an enzyme needed for metabolism **(signal transduction)**. Another example of molecular recognition is the interaction that brings together a substrate molecule with an enzyme. Before a metabolic reaction can occur, a substrate molecule must physically interact in a certain well-defined manner with the macromolecular catalyst, an enzyme. The biochemical action of drugs also depends on molecular recognition. A drug is first distributed throughout the body via the bloodstream. Drug molecules in the bloodstream are often bound to plasma proteins (e.g., serum albumin), which act as carriers. When the drug molecules are transported to their site of action, a second molecular interaction is likely to occur. Many drugs elicit their effects by interfering with biochemical processes. This may take the form of enzyme inhibition, where the drug molecule binds to a specific enzyme and prohibits catalytic action. Table 8.1 lists several other molecular recognition events that lead to some dynamic biochemical action. Details of these processes may be found in your biochemistry textbook.

Properties of Noncovalent Binding Interactions

The process of molecular recognition may be defined by a simple, reversible reaction that brings together a **ligand (L)** and a **macromolecule (M)** (R, for receptor, in some textbooks):

>> $$\mathrm{L} + \mathrm{M} \rightleftharpoons \mathrm{LM} \longrightarrow \text{Biological response} \qquad \textbf{(Eq. 8.1)}$$

LM represents a complex, held together by noncovalent interactions, that has a specialized biological function. The ligand is usually a relatively small molecule (hormone, substrate, inhibitor, drug, coenzyme, metal ion, etc.), but not always. Examples of large ligands include protein substrates that are cleaved by a protease, a protein molecule that binds to DNA as a transcription regulator, or large antigens that interact with antibodies.

TABLE 8.1 Examples of LM Interactions and Their Biological Responses

Type of Interaction (Ligand:Macromolecule)	Biological Significance or Response
Substrate:enzyme	Metabolic reactions
Inhibitor:enzyme	Metabolic regulation
Allosteric effector:enzyme	Metabolic regulation
Coenzyme:enzyme	Metabolic reactions
Hormone:receptor protein	Metabolic regulation; signal transduction
Antigen:antibody	Immune response
Ligand:carrier or transport protein	Storage or transport
Drug:protein or nucleic acid	Disease treatment
Regulatory protein:DNA	Transcription regulation

All molecular interactions that are the basis of molecular recognition have at least three common characteristics:

1. ***The binding forces that hold the complex together are noncovalent and relatively weak.*** Four types of noncovalent bonds are important:

   (a) **hydrogen bonds**
   (b) **van der Waals forces**
   (c) **hydrophobic interactions**
   (d) **ionic or electrostatic bonds**

   Properties and examples of these interactions are reviewed in Table 8.2. The strengths of noncovalent binding forces are in the range of 1–30 kJ/mol, compared to about 350 kJ/mol for a carbon–carbon single bond, a typical covalent bond. A single, noncovalent bond is usually insufficient to hold two molecules together. Typical biopolymers that may serve as macromolecules—DNA,

TABLE 8.2 Properties and Examples of Noncovalent Binding Forces

Type	Brief Description and Example	Stabilization Energy (kJ/mol)	Length (nm)
(i) *Hydrogen bonds*	Between a hydrogen atom bonded to an electronegative atom and a second electronegative atom Between neutral groups 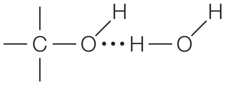 Between peptide bonds 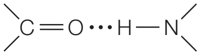	10–30	0.2–0.3
(ii) *van der Waals interactions*	Between molecules with temporary dipoles induced by fluctuating electrons. This may occur between any two atoms in close proximity.	1–5	0.1–0.2
(iii) *Hydrophobic interactions*	The presence of water forces nonpolar groups into ordered arrangements to avoid the water. 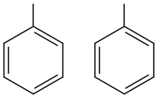 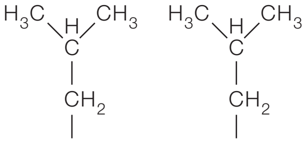	5–30	—
(iv) *Ionic bonds*	Interactions that occur between fully charged atoms or groups Na^+Cl^- $R—NH_3^+\,{}^-OOC—R$	20	0.25

RNA, polysaccharides, and proteins—have numerous functional groups that participate in noncovalent interactions. A collection of many of these interactions will lead to highly stabilized complexes.

2. ***Reactions that form noncovalent complexes are reversible.*** Noncovalent interactions are initiated when diffusing (wandering or moving) molecules come into close contact. Diffusion is brought about by thermal motions. An initial close encounter may not always result in the successful formation of a complex. A few weak bonds may form, but be disrupted by thermal motions, causing the molecules to dissociate. Therefore, bonds may constantly form and break until enough bonds have accumulated to result in an intermediate with transient, but significant, existence. The complex can then initiate a specific biological process. An intermediate rarely lasts longer than a few seconds. Eventually, thermal motions cause the complex to dissociate to the individual molecules. Reversibility is an important characteristic of these interactions so that a static, gridlock situation does not occur. The biological process initiated by the complex LM must have a finite lifetime—a starting and an ending time.
3. ***The binding between molecules is specific.*** Imagine that the interactions to form LM bring together two molecular surfaces. The two surfaces will be held together if noncovalent interactions are established. If on one surface there is a nonpolar molecular group (phenyl ring, hydrophobic alkyl chain), the adjacent region on the other surface must also be hydrophobic and nonpolar. If a positive charge exists on one surface, there may be a neutralizing negative charge on the other surface. A hydrogen-bond donor on one surface can interact favorably with a hydrogen-bond acceptor on the other surface. Simply stated, the two molecules must be compatible or complementary in a chemical sense so the development of stabilizing forces can hold molecules together. In molecular terms, the binding site (or sites) on M displays high specificity; therefore, only certain ligands can bind. L molecules, for a particular M, will be limited to a single, specific molecule or perhaps a group of structurally related molecules.

Quantitative Characterization of Ligand Binding

A thorough understanding of the biochemical significance of ligand binding to macromolecules comes only from a quantitative analysis of the strength of binding (affinity) between L and M. Binding affinity between two molecules is often expressed as an equilibrium constant, the **formation constant,** K_f, which is derived from the law of mass action. Consider the specific interaction between a

STUDY EXERCISE 8.1 Examples of Noncovalent Interactions

a. Show how the carboxylate group on the side chain of the amino acid aspartate can interact with a lysyl residue in a protein. Draw structures and define the possible types of noncovalent binding forces.

b. Show how a peptide carbonyl group can form a hydrogen bond with the free amino acid serine.

small molecule, L (for ligand), and a macromolecule, M (Equation 8.1). These two species combine to form a complex, LM.

K_f, the formation constant for the complex, is defined by Equation 8.2.

>> $$K_f = \frac{[LM]}{[L][M]} \qquad \textbf{(Eq. 8.2)}$$

Do not confuse K_f with K_d, the dissociation constant. The relationship between K_f and K_d is defined in Equation 8.3.

>> $$K_d = \frac{[L][M]}{[LM]} = \frac{1}{K_f} \qquad \textbf{(Eq. 8.3)}$$

The larger the value of K_f, the greater the strength of binding between L and M. (Large K_f implies a high concentration of LM relative to L and M.) Return to Equation 8.2 and note that in order to determine K_f, a method must be developed to measure equilibrium concentrations of L, M, and the complex LM. In a later section, we will describe experimental techniques that are applied to these measurements of binding constants, but first we must reorganize Equation 8.2 into a form that contains more readily measurable terms. We will begin with the assumption that the macromolecule, M, has several binding sites for L and that these sites do not interact with each other. That is, K_f is identical for all binding sites. The following definitions are necessary for the reorganization of Equation 8.2:

[L] = equilibrium concentration of free or unbound ligand
[M] = equilibrium concentration of macromolecule with no bound L, or the concentration of unoccupied binding sites
[LM] = equilibrium concentration of ligand–macromolecule complex, or the concentration of occupied sites
$[M]_0$ = total or initial concentration of macromolecule, or total concentration of available binding sites
$[L]_0$ = total concentration of bound and unbound ligand, or initial concentration of ligand
ν = fraction of available sites on M that are occupied, or the fraction of M that has L in binding site:

>> $$\nu = \frac{[LM]}{[M]_0} \qquad \textbf{(Eq. 8.4)}$$

The term ν is particularly significant because it can be considered a ratio of the number of occupied sites to the total number of potential binding sites on M. It can be measured experimentally, but first it must be redefined in the following manner: Since $[M]_0 = [LM] + [M]$, then

$$\nu = \frac{[LM]}{[LM] + [M]}$$

From Equation 8.2, $[LM] = K_f[L][M]$. Therefore,

$$\nu = \frac{K_f[L][M]}{K_f[L][M] + [M]}$$

Simplifying,

>> $$\nu = \frac{K_f[\mathrm{L}]}{K_f[\mathrm{L}] + 1} \qquad \textbf{(Eq. 8.5)}$$

You should recognize the similarity of Equation 8.5 to the Michaelis-Menten equation for enzyme catalysis. A graph of ν vs. [L] yields a hyperbolic curve (see Figure 8.1) that approaches a limiting value or saturation level. At this point, all binding sites on M are occupied. Because of the difficulty of measuring the exact point of saturation, this nonlinear curve is seldom used to determine K_f. Linear plots are more desirable, so Equation 8.5 is converted to an equation for a straight line. The equation will first be put into a more general form to account for any number of potential binding sites on M. The symbol $\bar{\nu}$ will be used to represent the average number of occupied sites per M, and n will represent the number of potential binding sites per M molecule. Assuming that all the binding sites on M are equivalent, Equation 8.5 becomes

>> $$\bar{\nu} = \frac{nK_f[\mathrm{L}]}{K_f[\mathrm{L}] + 1} \qquad \textbf{(Eq. 8.6)}$$

Scatchard's Equation

Equation 8.6 contains terms such as [L] that are difficult to determine, so it must be converted to a form amenable to experimental measurements.

If $\bar{\nu}$ is the average number of occupied sites per M molecule, then $n - \bar{\nu}$ is the average number of unoccupied sites per M molecule.

$$n - \bar{\nu} = n - \frac{nK_f[\mathrm{L}]}{K_f[\mathrm{L}] + 1}$$

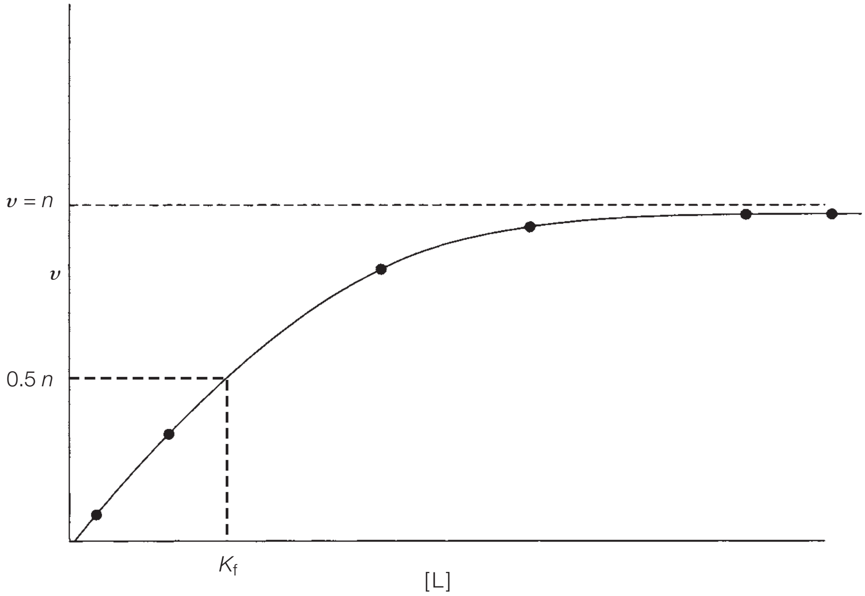

FIGURE 8.1 A plot illustrating saturation of binding sites with ligand. The n is estimated at the point when all binding sites are occupied by ligand. K_f is represented by the [L] when one-half of the sites are occupied by ligand.

Simplifying,

$$(n - \bar{\nu})(K_f[L] + 1) = n(K_f[L] + 1) - n(K_f[L])$$

>> $$(n - \bar{\nu}) = \frac{n}{K_f[L] + 1} \qquad \textbf{(Eq. 8.7)}$$

To further simplify Equation 8.7, let the term $\bar{\nu}/n - \bar{\nu}$ represent the ratio of occupied sites to nonoccupied sites on M; it can be mathematically represented as

$$\frac{\bar{\nu}}{n - \bar{\nu}} = \left(\frac{nK_f[L]}{K_f[L] + 1}\right)\left(\frac{K_f[L] + 1}{n}\right)$$

or

>> $$\frac{\bar{\nu}}{n - \bar{\nu}} = K_f[L] \qquad \textbf{(Eq. 8.8)}$$

A more desirable form for graphical use is known as **Scatchard's equation:**

>> $$\frac{\bar{\nu}}{[L]} = K_f(n - \bar{\nu}) \qquad \textbf{(Eq. 8.9)}$$

If a plot of $\bar{\nu}/[L]$ versus $\bar{\nu}$ yields a straight line, shown by the solid line in Figure 8.2, then all the binding sites on M are identical and independent, and K_f and n are estimated as shown in the figure.

Cooperative Binding of Ligands

The derivation of Equation 8.9 assumes that K_f is identical for all binding sites; that is, the binding of one molecule of L does not influence the binding of other L molecules to binding sites on M. However, it is common for ligand-macromolecule interactions to display such influences. The binding of one L molecule to M may encourage or inhibit the binding of a second L molecule to M. For example, the binding of oxygen to one of the four subunits of hemoglobin increases the affinity of the other subunits for oxygen. There is said to be **cooperativity** of sequential binding. If the sites do show cooperative binding, the plot is nonlinear, as shown by the dashed line in Figure 8.2. The shape of the nonlinear curve may be used to determine the number of types of binding sites. The dashed line in Figure 8.2 can be resolved into two lines, indicating that two types of binding sites are present on M. The n and K_f for each type of binding site may be estimated by resolving the smooth curve into straight lines as shown in Figure 8.2. K_f and n can be estimated by extrapolating the two straight lines to the axes.

Experimental Measurement of Ligand-Binding Interactions

To analyze ligand-macromolecule interactions quantitatively (to use Equation 8.9), one must be able to distinguish experimentally between bound ligand (LM) and free ligand (L). Many techniques have been developed for measuring the dynamics of LM interactions. Widely used techniques include equilibrium dialysis (Chapter 3, p. 73), ultrafiltration (Chapter 3, p. 74), and spectroscopic (especially

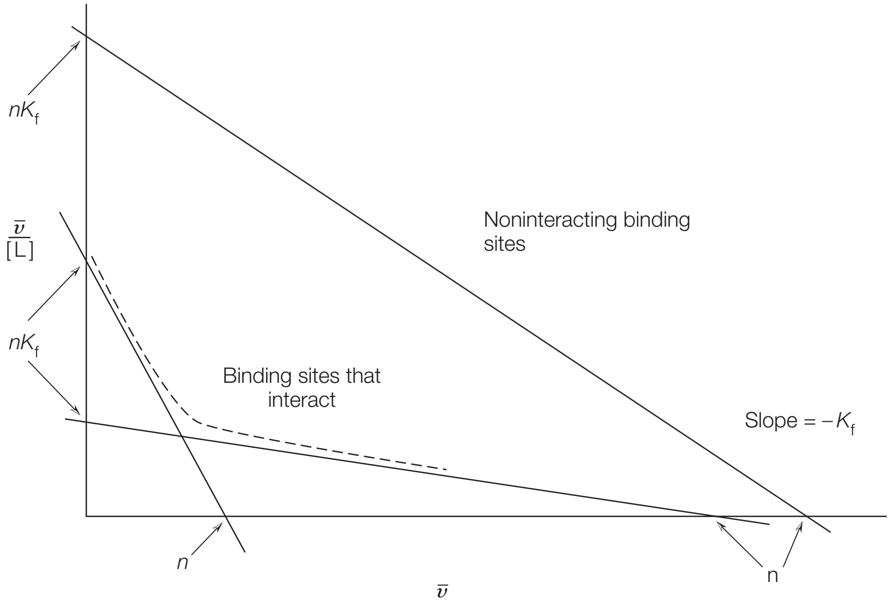

FIGURE 8.2 Scatchard plot. Two types of Scatchard curves are illustrated. The upper plot (solid line) represents binding to a macromolecule with noninteracting sites. The binding of a ligand molecule at one site is independent of the binding of a second ligand molecule at another site. The plot can be extrapolated to each axis and the n and K_f calculated. The lower, dashed line represents binding of ligand molecules to binding sites that interact. One ligand molecule bound to the macromolecule influences the rate of binding of other ligand molecules. The dashed line is evaluated as shown in the figure. The curved line has two distinct slopes. It can be resolved into two straight lines, each of which may be evaluated for n and K_f. This would indicate that there are two types of binding sites, each with a unique n and K_f.

fluorescence) measurements (Chapter 7, p. 220). Many of these methods require specialized and expensive equipment, sometimes cumbersome procedures, and the use of radio-labeled ligands.

STUDY EXERCISE 8.2 Using Equilibrium Dialysis

Ligand binding to a protein can be measured by the procedure of equilibrium dialysis (see Chapter 3, Section D, p. 73). A solution of the protein (M) is sealed in a dialysis bag, which is then placed in a large volume of solution containing the ligand (L). The pore size of the dialysis bag allows the passage of ligand molecules, but not protein molecules. After equilibrium is reached several hours later, the bag is opened and the concentrations of free ligand inside and outside are measured. The difference between these two values is the amount of ligand bound to the protein. The following experimental data are collected:

$$\text{Concentration of free ligand, [L]} = 1 \times 10^{-5}\ \text{mole/liter}$$
$$\text{Concentration of bound ligand, [LM]} = 5 \times 10^{-6}\ \text{mole/liter}$$
$$K_f \text{ for the reaction } L + M \rightleftharpoons LM = 1 \times 10^{5}\ \text{moles/liter}$$

Calculate the concentration of the protein, M, in moles per liter. Assume that M has only one binding site for the ligand.

Solution:

Use Equation 8.2 to calculate [M]. The concentration of M $= 5 \times 10^{-6}$ moles/liter.

One of the simplest and most convenient methods for monitoring ligand binding is the differential method, which detects and quantifies some measurable change in spectral absorption or fluorescence, in the UV-VIS regions, that accompanies ligand binding (see next section). A ligand may show enhanced fluorescence when bound to a macromolecule, or amino acid residues (i.e., Trp) in a protein may show enhanced fluorescence when a ligand molecule is bound.

The Bradford Protein Assay as an Example of Ligand Binding

When Coomassie Blue dye binds to proteins, the dye undergoes a significant spectral change in the visible region (Figure 8.3). The spectrum of free dye displays a minimum in the range 575–625 nm. When dye and a protein (ovalbumin) are mixed, a new absorption band appears in the region of 595 nm. The increase in absorption at 595 nm is directly related to the concentration of protein, which is the basis of the Bradford protein assay (Chapter 3, Section B, p. 69). The new absorption band is due to the presence of dye-ovalbumin complex. Note that the new band is rather broad and that a range of wavelengths may actually be used, but 595 nm is usually chosen. In this section, we will use the Bradford assay to illustrate an alternate method for using Scatchard's equation.

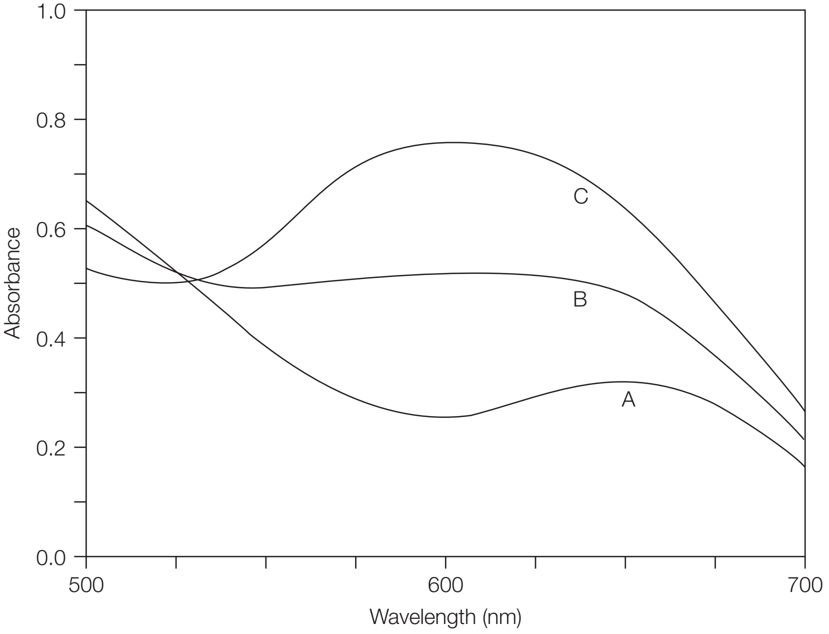

FIGURE 8.3 The spectral changes occurring during the binding of Coomassie Blue dye to ovalbumin. **A** free dye in solution, **B** dye plus 0.5 mg of ovalbumin, **C** dye plus 1.0 mg of ovalbumin. The dye concentration is identical in all three curves.

Equation 8.6 for evaluating protein-ligand interactions is easily adaptable to the spectroscopic differential method, but it must be converted into a linear form. Absorption data can be expressed in terms of [bound]/[free] ligands or [occupied]/[unoccupied] sites on the macromolecule. The data analysis described here assumes that L and LM, but not M, contribute to the measured absorbance. The measured absorbance change is proportional to the ratio of bound L to unbound L.

The absorbance observed from a mixture of L and M, A_{obs}, is

$$A_{\text{obs}} = A_{\text{L}} + A_{\text{LM}}$$

where

A_{L} = absorbance of free L
A_{LM} = absorbance of bound L

Absorbance data may be used directly to calculate the fraction of ligand bound:

>> $$\text{Fraction of ligand bound} = f = \frac{A_{\text{obs}} - A_{\text{L}}}{A_{\text{max}} - A_{\text{L}}} \qquad \textbf{(Eq. 8.10)}$$

where

A_{max} = total absorbance of LM when all the L molecules in solution are bound to M. This number is experimentally determined in this chapter, Section B.

The term $A_{\text{max}} - A_{\text{L}}$ is proportional to the total number of binding sites, and $A_{\text{obs}} - A_{\text{L}}$ represents the average number of occupied sites on M caused by a certain concentration of L.

For graphical analysis of the experimental data, $\bar{\nu}$ for Equation 8.6 can be defined in terms that can be measured experimentally:

>> $$\bar{\nu} = f\frac{[\text{L}]_0}{[\text{M}]_0} \qquad \textbf{(Eq. 8.11)}$$

If Equation 8.11 is combined with Equation 8.6, an equation for a straight line can be derived in terms that are readily measured.

$$f\frac{[\text{L}]_0}{[\text{M}]_0} = \frac{nK_{\text{f}}[\text{L}]}{K_{\text{f}}[\text{L}] + 1}$$

Take the reciprocal of each side and solve for $[\text{M}]_0/f$:

$$\frac{[\text{M}]_0}{f[\text{L}]_0} = \frac{K_{\text{f}}[\text{L}] + 1}{nK_{\text{f}}[\text{L}]}$$

$$\frac{[\text{M}]_0}{f} = \frac{[\text{L}]_0}{n} + \frac{[\text{L}]_0}{nK_{\text{f}}[\text{L}]}$$

The term $[\text{L}]/[\text{L}]_0 = 1 - f$; therefore:

$$\frac{[\text{M}]_0}{f} = \frac{[\text{L}]_0}{n} + \frac{1}{nK_{\text{f}}(1 - f)}$$

Rearrange to the form for a straight line, $y = mx + b$:

>> $$\frac{[\mathrm{M}]_0}{f} = \frac{1}{nK_\mathrm{f}(1 - f)} + \frac{[\mathrm{L}]_0}{n} \qquad \textbf{(Eq. 8.12)}$$

This is an alternate form of the Scatchard equation. The values n and K_f can be calculated from a plot of $1/(1 - f)$ vs. $[\mathrm{M}]_0/f$. The slope of the line is represented by $1/nK_\mathrm{f}$, whereas the y intercept is $[\mathrm{L}]_0/n$. If the data points fall on a straight line as in Figure 8.4, then all the dye binding sites on the protein M are identical and independent.

The derivation of Equation 8.12 assumes that K_f is identical for all binding sites; that is, the binding of one molecule of L does not influence the binding of other L molecules to binding sites on M. However, it is common for ligand-macromolecule interactions to display such influences. The binding of one L molecule to M may encourage or inhibit the binding of a second L molecule to M. For example, the binding of oxygen to one of the four subunits of hemoglobin increases the affinity of the other subunits for oxygen. There is said to be cooperativity of sequential binding. If the sites do show cooperative binding, the plot becomes nonlinear, as shown by the curved line in Figure 8.4. This line is resolved into two lines, indicating that two types of binding sites are present on M. The n and K_f for each type of binding may be estimated by resolving the smooth curve into two straight lines. K_f and n may be calculated from the intercepts and slopes.

Computer Software for Analysis of LM Binding

Numerous computer programs exist that may be used for graphical analysis of ligand binding data. Most programs perform a linear or nonlinear least-squares regression analysis of experimental data. Many are available as freeware on the

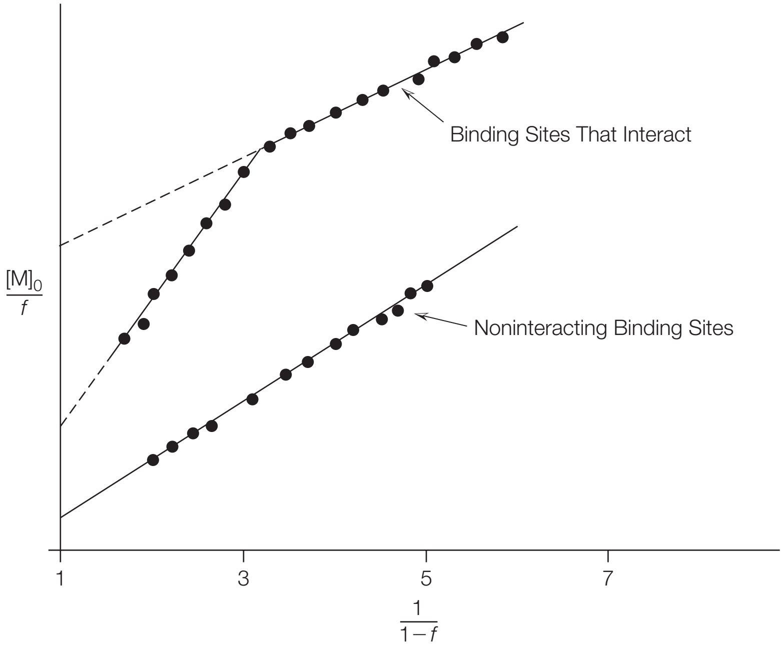

FIGURE 8.4 Scatchard plot for typical binding data. See text for details.

Internet. Some popular and effective software programs are listed and described in Table 8.6. Additional Web sites and references are available at the end of this chapter and in Appendix I. One very useful program is DynaFit (www.biokin.com), which may be applied to the analysis of ligand binding data or enzyme kinetic data. The program has been designed by Dr. Petr Kuzmic, President and CEO of Biokin Ltd. DynaFit is available free for academic users (students, professors, university researchers), but one must register for an academic license at the Web site.

B. BIOLOGICAL CATALYSIS (ENZYMES)

Enzyme-catalyzed reactions may also be described in terms of ligand-macromolecule interactions. Enzyme reactions are initiated by the interaction of a substrate molecule with an enzyme molecule:

>> $$E + S \rightleftharpoons ES$$ **(Eq. 8.13)**

Enzymes are biological catalysts. Without their presence in a cell, most biochemical reactions would not proceed at the required rate. The chemical and biological properties of enzymes have been investigated since the early 1800s. The unrelenting interest in enzymes is due to several factors—their dynamic and essential roles in the cell, their extraordinary catalytic power, and their selectivity.

Enzyme-catalyzed reactions proceed through an ES complex, as shown in Equation 8.14. Individual rate constants *k* are assigned to each reaction step.

>> $$E + S \underset{k_2}{\overset{k_1}{\rightleftharpoons}} ES \underset{k_4}{\overset{k_3}{\rightleftharpoons}} E + P$$ **(Eq. 8.14)**

E represents the enzyme, S the substrate or reactant, and P the product. For a specific enzyme, only one or a few different substrate molecules can bind in the proper manner and produce a functional ES complex. The substrate must have a size, shape, and polarity compatible with the active site of the enzyme. Some enzymes catalyze the transformation of many different molecules as long as there is a common type of chemical linkage in the substrate. Others have absolute specificity and can form reactive ES complexes with only one molecular structure. In fact, some enzymes are able to differentiate between D and L isomers of substrates.

Classes of Enzymes

Thousands of enzymes have now been isolated and studied; confusion would reign without some system for nomenclature and classification. Common names for enzymes are usually formed by adding the suffix **–ase** to the name of the substrate. The enzyme tyrosinase catalyzes the oxidation of tyrosine; cellulase catalyzes the hydrolysis of cellulose to produce glucose. Common names of this type define the substrate, but do not describe the chemistry of the reaction. Some very early names, such as catalase, trypsin, and pepsin, are even less descriptive and give no clue to their function or substrates. To avoid such confusion, enzymes now have official names that reflect the reactions they catalyze. All known enzymes can be classified into one of six categories (Table 8.3). Each enzyme has an official international name ending in *–ase* and a classification number, called the Enzyme Commision (E.C.) number. The number consists of four digits, each referring to a class and subclass of reaction. Table 8.3 shows an example from each class of enzyme.

TABLE 8.3 An Example of Each Class of Enzyme

1. Oxidoreductase

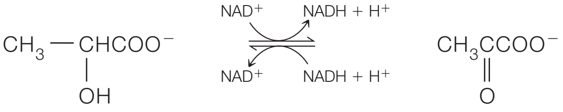

$$CH_3-\underset{\displaystyle OH}{\underset{|}{C}}HCOO^- \underset{NAD^+ \quad NADH + H^+}{\overset{NAD^+ \quad NADH + H^+}{\rightleftharpoons}} CH_3\underset{\displaystyle O}{\underset{\|}{C}}COO^-$$

Lactate Pyruvate

Common name: Lactate dehydrogenase

Official name: L-Lactate: NAD^+ oxidoreductase

Official E.C. number: 1.1.2.3

2. Transferase

$(d - NMP)_n + d - NTP \rightleftharpoons (d - NMP)_{n+1} + PP_i$

$(d - NMP)_n$ = DNA with n nucleotides

$(d - NMP)_{n+1}$ = DNA with $n + 1$ nucleotides

PP_i = Pyrophosphate

Common name: DNA polymerase

Official name: Deoxynucleoside triphosphate: DNA deoxynucleotidyltransferase (DNA-directed)

Official E.C. Number: 2.7.7.7

3. Hydrolase

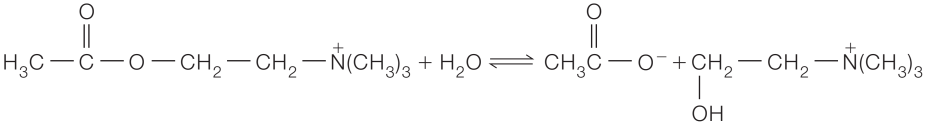

$$H_3C-\overset{\displaystyle O}{\overset{\|}{C}}-O-CH_2-CH_2-\overset{+}{N}(CH_3)_3 + H_2O \rightleftharpoons CH_3\overset{\displaystyle O}{\overset{\|}{C}}-O^- + \underset{\displaystyle OH}{\underset{|}{C}}H_2-CH_2-\overset{+}{N}(CH_3)_3$$

Acetylcholine Acetate Choline

Common name: Acetylcholinesterase

Official name: Acetylcholine acetylhydrolase

Official E.C. number: 3.1.1.7

4. Lyase

$CO_2 + H_2O \rightleftharpoons H_2CO_3$

Carbonic acid

Common name: Carbonic anhydrase

Official name: Carbonate hydrolyase

Official E.C. number: 4.2.1.1

(*Continued*)

TABLE 8.3 (*Continued*)

5. Isomerase

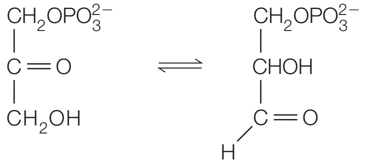

Dihydroxyacetone phosphate

Glyceraldehyde 3-phosphate

Common name: Triose phosphate isomerase

Official name: D-Glyceraldehyde-3-phosphate ketoisomerase

Official E.C. number: 5.3.1.1

6. Ligase

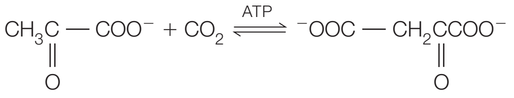

Pyruvate

Oxaloacetate

Common name: Pyruvate carboxylase

Official name: Pyruvate: CO_2 ligase (ADP-forming)

Official E.C. number: 6.4.1.1

Kinetic Properties of Enzymes

The **initial reaction velocity,** ν_0, of an enzyme-catalyzed reaction varies with the substrate concentration, [*S*], as shown in Figure 8.5. The Michaelis-Menten equation has been derived to account for the kinetic properties of enzymes. (Consult a biochemistry textbook for a derivation of this equation and for a discussion of the conditions under which the equation is valid.) The common form of the equation is

>>

$$\nu_0 = \frac{V_{\max}[S]}{K_M + [S]} \qquad \textbf{(Eq. 8.15)}$$

where

ν_0 = initial reaction velocity

$V_{\max}$ = maximal reaction velocity; attained when all enzyme active sites are filled with substrate molecules

[S] = substrate concentration

K_M = Michaelis constant $= \dfrac{k_2 + k_3}{k_1}$

The important kinetic constants $V_{\max}$ and K_M can be graphically determined as shown in Figure 8.5. Equation 8.15 and Figure 8.5 have all of the disadvantages of nonlinear kinetic analysis. $V_{\max}$ can only be estimated because of the asymptotic nature of the line. The value of K_M, the substrate concentration

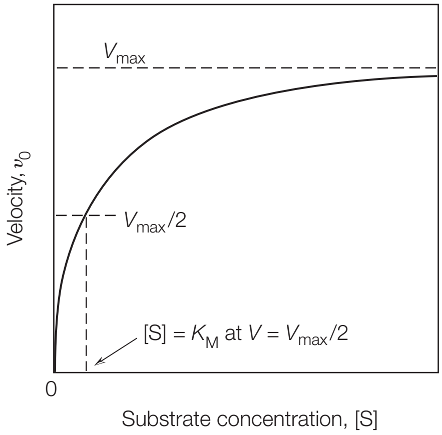

FIGURE 8.5 Michaelis-Menten plot for an enzyme-catalyzed reaction.

that results in a reaction velocity of $V_{max}/2$, depends on V_{max}, so both are just an estimate. By taking the reciprocal of both sides of the Michaelis-Menten equation, however, it is converted into the Lineweaver-Burk relationship (Equation 8.16).

>>

$$\frac{1}{v_0} = \frac{K_M}{V_{max}} \cdot \frac{1}{[S]} + \frac{1}{V_{max}} \quad \textbf{(Eq. 8.16)}$$

This equation, which is in the form $y = mx + b$, gives a straight line when $1/\nu_0$ is plotted against 1/[S] (Figure 8.6). The intercept on the $1/\nu_0$ axis is $1/V_{max}$, and the intercept on the 1/[S] axis is $-1/K_M$. A disadvantage of the Lineweaver-Burk plot is that the data points are compressed in the high substrate concentration region.

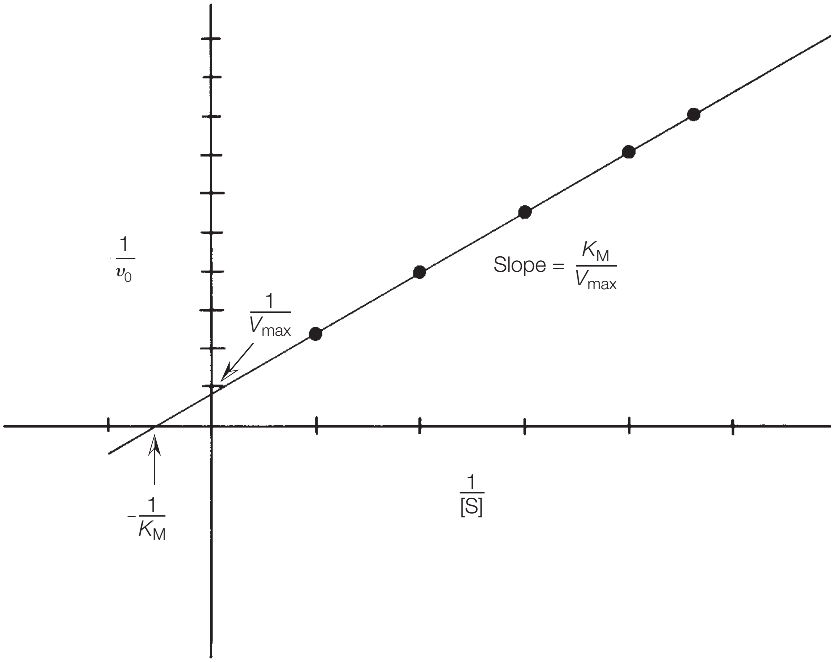

FIGURE 8.6 Lineweaver-Burk plot for an enzyme-catalyzed reaction.

Significance of Kinetic Constants

The **Michaelis constant,** K_M, for an enzyme-substrate interaction has two meanings: (1) K_M is the substrate concentration that leads to an initial reaction velocity of $V_{max}/2$ or, in other words, the substrate concentration that results in the filling of one-half of the enzyme active sites, and (2) $K_M = (k_2 + k_3)/k_1$. The second definition of K_M has special significance in certain cases. When $k_2 \gg k_3$, then $K_M = k_2/k_1$ so K_M is equivalent to the dissociation constant of the ES complex. When $k_2 \gg k_3$, a large K_M implies weak interaction between E and S, whereas a small K_M indicates strong binding between E and S.

V_{max} is important because it leads to the analysis of another kinetic constant, k_3, **turnover number.** The analysis of k_3 begins with the basic rate law for an enzyme-catalyzed process (Equation 8.17), which is derived from Equation 8.14.

>> $$\nu_0 = k_3[\mathrm{ES}] \qquad \textbf{(Eq. 8.17)}$$

If all of the enzyme active sites are saturated with substrate, then [ES] in Equation 8.17 is equivalent to $[E_T]$, the total concentration of enzyme, and ν_0 becomes V_{max}; hence

>> $$V_{max} = k_3[\mathrm{E_T}] \qquad \textbf{(Eq. 8.18)}$$

or

$$k_3 = \frac{V_{max}}{[\mathrm{E_T}]}$$

For an enzyme with one active site per molecule, the turnover number, k_3, is the number of substrate molecules transformed to product by one enzyme molecule per unit time, usually in minutes or seconds. The turnover number is a measure of the efficiency of an enzyme. K_M and k_3 values for some enzymes are listed in Table 8.4.

TABLE 8.4 K_M **and** k_3 **Values for Some Enzyme: Substrate Systems**

Enzyme	Substrate	K_M (mM)	$k_3(\mathrm{sec}^{-1})$
Catalase	H_2O_2	0.001	40,000,000
Carbonic anhydrase	HCO_3^-	9	400,000
Acetylcholinesterase	Acetylcholine	0.09	25,000
Penicillinase	Benzylpenicillin	0.050	2,000
Chymotrypsin	Glycyltyrosinylglycine	108	100
DNA polymerase	DNA	–	15
Ribulose-1,5-bisphosphate carboxylase	Ribulose-1,5-bisphosphate	0.028	3.3
	CO_2	0.009	–

STUDY EXERCISE 8.3 Enzyme Kinetics

An enzyme displays the following reaction kinetics:

$$V_{max} = 200\ \mu\text{mol/min}$$
$$\text{Initial rate } (\nu_0) = 75\ \mu\text{mol/min}$$
$$[S] = 10\ \mu\text{M}$$

Calculate K_M.

Inhibition of Enzyme Activity

Nonsubstrate molecules may interact with enzymes, leading to a decrease in enzymatic activity. Enzyme inhibition, the phenomenon in which molecules interfere with enzymes, is of interest because it often provides clues about the mechanism of enzyme action and also reveals information about metabolic control and regulation. In addition, many toxic substances, including drugs, express their action by enzyme inhibition. Inhibition of enzymes by reversible pathways will be reviewed here. For more details on enzyme inhibition, including irreversible processes, see your biochemistry course textbook.

The process of reversible inhibition is described by an equilibrium interaction between enzyme (E) and inhibitor molecule (I):

>>

$$E + I \rightleftharpoons EI \qquad \textbf{(Eq. 8.19)}$$

Most inhibition processes can be classified as **competitive, mixed,** or **uncompetitive,** depending on how the inhibitor interferes with enzyme action.

1. ***Competitive inhibition.*** A competitive inhibitor is usually similar in structure to the substrate and is capable of reversible binding to the enzyme active site. In contrast to the substrate molecule, the inhibitor molecule cannot undergo chemical transformation to a product; however, it does interfere with substrate binding. The binding of substrate and competitive inhibitor is a mutually exclusive process: When inhibitor is bound to enzyme, substrate is unable to bind and vice versa. The kinetic scheme for competitive inhibition is as follows:

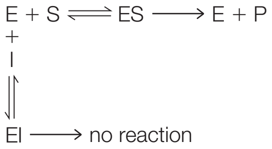

2. ***Mixed inhibition.*** A mixed inhibitor does not bind in the active site of the enzyme, but binds to some other region of the enzyme molecule. The presence of inhibitor may or may not affect substrate binding, but it does interfere with the catalytic functioning of the enzyme. [A noncompetitive inhibitor, a special type of mixed inhibitor, does not affect substrate binding

(no change in K_M), but does affect the catalytic efficiency of the enzyme (lowers V_{max}).] The enzyme with mixed inhibitor bound may be converted to a nonfunctional conformational state. The kinetic scheme for noncompetitive inhibition is as follows (note that the inhibitor may bind to the free enzyme, E, to EI, or to ES):

$$\begin{tikzcd}
\mathrm{E} + \mathrm{S} \arrow[r, shift left] & \mathrm{ES} \arrow[l, shift left] \arrow[r] & \mathrm{E} + \mathrm{P} \\
+\ \mathrm{I} \arrow[d, shift left] & +\ \mathrm{I} \arrow[d, shift left] & \\
\mathrm{EI} + \mathrm{S} \arrow[u, shift left] \arrow[r, shift left] & \mathrm{EIS} \arrow[l, shift left] \arrow[u, shift left] \arrow[r] & \text{no reaction}
\end{tikzcd}$$

3. ***Uncompetitive inhibition.*** An uncompetitive inhibitor is similar to a noncompetitive inhibitor—it allows substrate to bind to the active site. It differs, however, because an uncompetitive inhibitor binds only to the ES complex:

$$\begin{tikzcd}
\mathrm{E} + \mathrm{S} \arrow[r, shift left] & \mathrm{ES} \arrow[l, shift left] \arrow[r] & \mathrm{E} + \mathrm{P} \\
 & +\ \mathrm{I} \arrow[d, shift left] & \\
 & \mathrm{ESI} \arrow[u, shift left] \arrow[r] & \text{no reaction}
\end{tikzcd}$$

Because the inhibitor combines only with ES and not free enzyme, it will influence the activity of the enzyme only when substrate concentrations and, in turn, ES concentrations are high.

Any complete inhibition study requires the experimental differentiation among the three types of inhibition. The three inhibitory processes are kinetically distinguishable by application of the Lineweaver-Burk equation. For each inhibitor studied, at least two sets of rate experiments are completed. In all sets, the enzyme concentration is identical. In set 1, the substrate concentration is varied and no inhibitor is added. In set 2, the same variable substrate concentrations from set 1 are used, and a constant amount of inhibitor is added to each assay. If additional data are desired, more sets are prepared with variable substrate concentrations as in set 2; however, a constant, and different, concentration of inhibitor is present. These data, when plotted on a Lineweaver-Burk graph, lead to three lines as shown in Figure 8.7. In competitive inhibition, all three lines intersect at the same point on the $1/\nu_0$ axis; hence, V_{max} is not altered by the presence of the competitive inhibitor. If enough substrate is added, the competitive inhibition may be overcome. The apparent K_M value (measured on the $1/[S]$ axis) changes with each increase in inhibitor concentration. In mixed (noncompetitive) inhibition, the family of lines has a common intercept on the $1/[S]$ axis (unchanged K_M for the lines). For uncompetitive inhibition, parallel lines are obtained indicating that both V_{max} and K_M are changed. The kinetic characteristics of reversible inhibition are compared in Table 8.5.

Units of Enzyme Activity

The actual molar concentration of an enzyme in a cell-free extract or purified preparation is seldom known. Only if the enzyme is available in a pure crystalline form, carefully weighed, and dissolved in a solvent can the actual molar

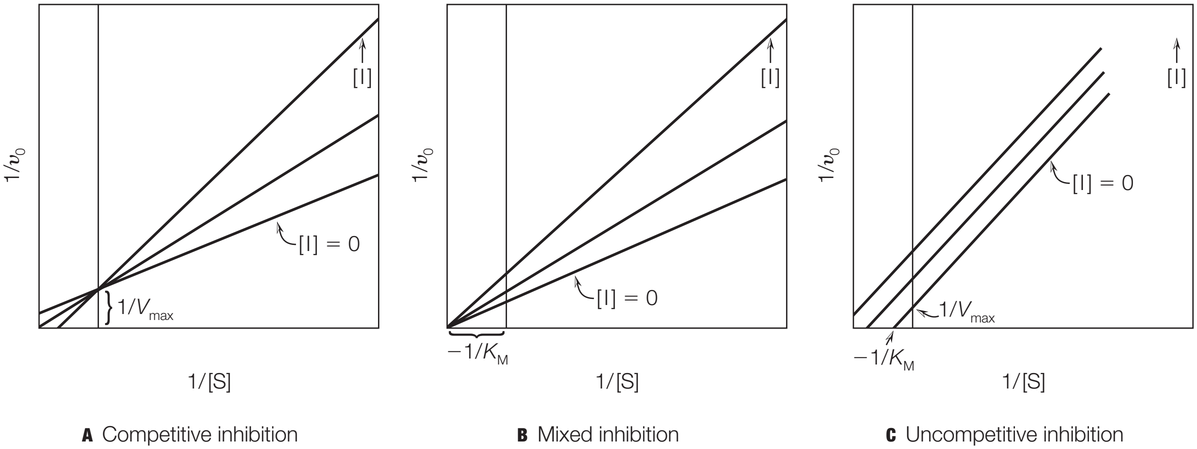

FIGURE 8.7 Lineweaver-Burk plots to determine type of inhibition. Each set of three lines includes a line representing no inhibitor present, and two lines showing two different concentrations of inhibitor. **A** competitive inhibition: The lines intersect on the 1/v_0 axis, where V_{max} can be calculated. **B** Mixed (noncompetitive) inhibition: The lines intersect on the 1/[S] axis, where K_M can be calculated. **C** Uncompetitive inhibition: The lines are parallel. The V_{max} and K_M can be calculated for each line as shown.

TABLE 8.5 Kinetic Characteristics of Reversible Inhibition

	Effect on Inhibited Reaction[1]		
Type of inhibition	K_m	V_{max}	K_m/V_{max} **(Slope)**
Competitive	Higher	Same	Increase
Mixed	Higher	Lower	Increase
Noncompetitive	Same	Lower	Increase
Uncompetitive	Lower	Lower	Same

[1] *Relative to uninhibited reaction.*

concentration be accurately determined. It is, however, possible to develop a precise and accurate assay for enzyme activity. Consequently, the amount of a specific enzyme present in solution is most often expressed in units of activity. Three units are in common use, the *international unit* (IU), the *katal*, and *specific activity*. The International Union of Biochemistry Commission on Enzymes has recommended the use of a standard unit, the **international unit** or simply **unit,** of enzyme activity. One IU of enzyme corresponds to the amount that catalyzes the transformation of 1 μmole of substrate to product per minute under specified conditions of pH, temperature, ionic strength, and substrate concentration. If a solution containing enzyme converts 10 μmoles of substrate to product in 5 minutes, the solution contains 2 enzyme units. A new unit of activity, the katal, has been recommended. One **katal** of enzyme activity represents the conversion of 1 mole of substrate to product in 1 second. One international unit is equivalent to 1/60 μkatal or 0.0167 μkatal. One katal, therefore, is equivalent to 6×10^7 international units. It is convenient to represent small amounts of enzyme in millikatals (mkatals), microkatals (μkatals), or nanokatals (nkatals). The enzyme activity in the above example was 2 units, which can be converted to katals as follows: Since one katal is 6×10^7 units, 2 units are equivalent to $2/6 \times 10^7$ or 33 nkatals (0.033 μkatals). If the enzyme is an impure preparation in solution, the activity is most often expressed as units/mL or katals/mL.

Specific Activity

Another useful quantitative definition of enzyme efficiency is **specific activity**. The specific activity of an enzyme is the number of enzyme units or katals per milligram of protein. This is a measure of the purity of an enzyme. If a solution contains 20 mg of protein that express 2 units of activity (33 nkatals), the specific activity of the enzyme is $2\,\text{units}/20\,\text{mg} = 0.1\,\text{units/mg}$, or $33\,\text{nkatals}/20\,\text{mg} = 1.65\,\text{nkatals}/20\,\text{mg}$. As an enzyme is purified, its specific activity increases. That is, during purification, the enzyme concentration increases relative to the total protein concentration until a limit is reached. The maximum specific activity is attained when the enzyme is homogeneous, or in a pure form.

The activity of an enzyme depends on several factors, including substrate concentration, cofactor concentration, pH, temperature, and ionic strength. The conditions under which enzyme activity is measured are critical and must be specified when activities are reported.

Design of an Enzyme Assay

Whether an enzyme is obtained commercially or prepared (purified) in a multistep procedure, an experimental method must be developed to detect and quantify the specific enzyme activity. During isolation and purification of an enzyme, the assay is necessary to determine the amount and purity of the enzyme and to evaluate its kinetic properties. An assay is also essential for a further study of the mechanism of the catalyzed reaction.

The design of an assay requires certain knowledge of the reaction:

1. The complete stoichiometry.
2. What substances are required (substrate, metal ions, cofactors, etc.) and their kinetic dependence.
3. Effect of pH, temperature, and ionic strength.

The most straightforward approach to the design of an enzyme assay is to measure the change in substrate or product concentration during the reaction.

If an enzyme assay involves continuous monitoring of substrate or product concentration, the assay is said to be **kinetic.** If a single measurement of substrate or product concentration is made after a specified reaction time, a **fixed-time assay** results. The kinetic assay is more desirable because the time course of the reaction is directly observed and any discrepancy from linearity can be immediately detected.

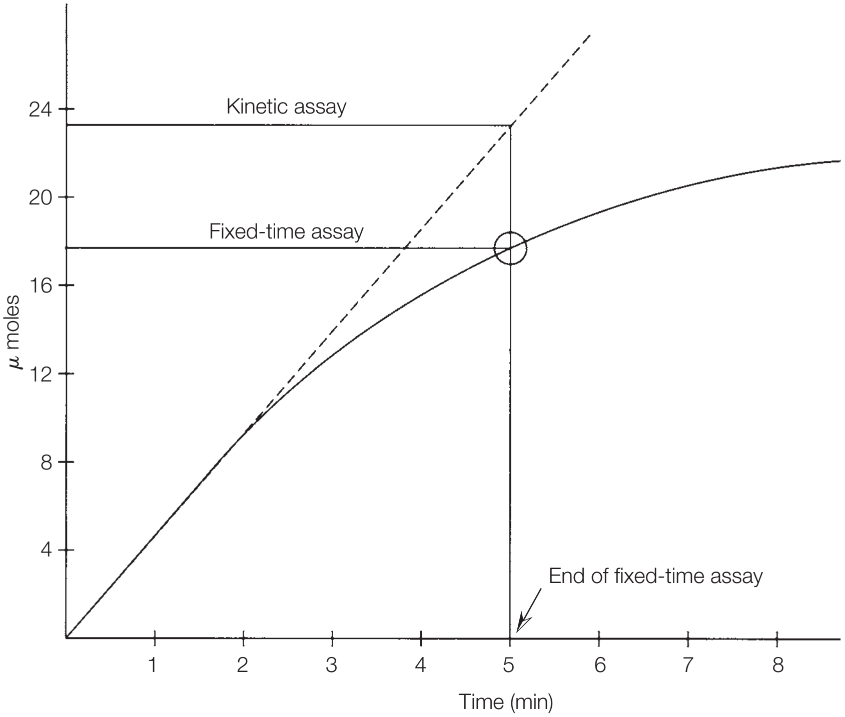

FIGURE 8.8 Kinetic progress of an enzyme-catalyzed reaction. See text for details.

Kinetic versus Fixed-time Assay

Figure 8.8 displays the kinetic progress curve of a typical enzyme-catalyzed reaction and illustrates the advantage of a kinetic assay. The rate of product formation decreases with time. This may be due to any combination of factors such as decrease in substrate concentration, denaturation of the enzyme, and product inhibition of the reaction. The solid line in Figure 8.8 represents the continuously measured time course of a reaction (kinetic assay). The true rate of the reaction is determined from the slope of the dashed line drawn tangent to the experimental result. From the data given, the rate is 5 μmoles of product formed per minute. Data from a fixed-time assay are also shown on Figure 8.8. If it is assumed that no product is present at the start of the reaction, then only a single measurement after a fixed period is necessary. This is shown by a circle on the experimental rate curve. The measured rate is now 16 μmoles of product formed every 5 minutes, or about 3 μmoles/minute, which is considerably lower than the rate derived from the continuous, kinetic assay. Which rate measurement is correct? Obviously, the kinetic assay gives the true rate because it corrects for the decline in rate with time. The fixed-time assay can be improved by changing the time of the measurement, in this example, to 2 minutes of reaction time, when the experimental rate is still linear. It is possible to obtain true rates with the fixed-time assay, but one must choose the time period very carefully. In the laboratory, this is done by removing aliquots of a reaction mixture at various times and measuring the concentration of product formed in each aliquot. Figure 8.8 reinforces an assumption used in the derivation of the Michaelis-Menten equation. Only measurements of initial velocities lead to true reaction rates. This avoids the complications of enzyme denaturation, decrease of [S], product inhibition, and reversion of the product to substrate.

Several factors must be considered when the experimental assay conditions are developed. The reaction rate depends on the concentrations of substrate, enzyme, and necessary cofactors. In addition, the reaction rate is under the influence of environmental factors such as pH, temperature, and ionic strength. Enzyme activity increases with increasing temperature until the enzyme becomes denatured. The enzyme activity then decreases until all enzyme molecules are inactivated by denaturation. During kinetic measurement, it is essential that the temperature of all reaction mixtures be maintained as constant.

Ionic strength and pH should also be monitored carefully. Although some enzymes show little or no change in activity over a broad pH range, most enzymes display maximum activity in a narrow pH range. Any assay developed to evaluate the kinetic characteristics of an enzyme must be performed in a buffered solution, preferably at the optimal pH.

Applications of an Enzyme Assay

The conditions used in an enzyme assay depend on what is to be accomplished by the assay. There are two primary applications of an enzyme assay procedure. First, it may be used to measure the concentration of active enzyme in a preparation. In this circumstance, the measured rate of the enzyme-catalyzed reaction must be proportional to the concentration of enzyme; stated in more kinetic terms, there must be a linear relationship between initial rate and enzyme concentration (the reaction is first-order in enzyme concentration). To achieve this, certain conditions must be met: (1) the concentrations of substrate(s), cofactors, and other required substances must be in excess; (2) the reaction mixture must not contain inhibitors of the enzyme; and (3) all environmental factors such as pH, temperature, and ionic strength should be controlled. Under these conditions, a plot of enzyme activity (μmole product formed/minute) vs. enzyme concentration is a straight line and can be used to estimate the concentration of active enzyme in solution.

Second, an enzyme assay may be used to measure the kinetic properties of an enzyme, such as K_M, V_{max}, and inhibition characteristics. In this situation, different experimental conditions must be used. If K_M for a substrate is desired, the assay conditions must be such that the measured initial rate is first-order in substrate. To determine K_M of a substrate, constant amounts of enzyme are incubated with varying amounts of substrate. A Lineweaver-Burk plot ($1/v_0$ vs. $1/[S]$) may be used to determine K_M and V_{max}. If a reaction involves two or more substrates, each must be evaluated separately. The concentration of only one substrate is varied, while the other is held constant at a saturating level. The same procedure holds for determining the kinetic dependence on cofactors. Substrate(s) and enzyme are held constant, and the concentration of the cofactor is varied.

Inhibition kinetics are included in the second category of assay applications. An earlier discussion outlined the kinetic differentiation among competitive, uncompetitive, and mixed inhibition. In summary, a study of enzyme kinetics is approached by measuring initial reaction velocities under conditions in which only one factor (substrate, enzyme, cofactor) is varied and all others are held constant.

STUDY EXERCISE 8.4 Converting Rate Data from ΔA/min Units to Concentration Units

Because most enzyme assays are completed with the use of spectroscopic methods, initial rate data are in the form of ΔA/min. The change of color intensity per time interval is not a useful unit for kinetics. A more valuable and widely used unit that actually defines the amount of product formed per time interval is mmol/min or μmol/min. The conversion of ΔA units to concentration units may be done using Beer's Law:

$$\mathrm{c} = A/El$$

where

A = the absorbance
E = the molar absorption coefficient of the substrate or product
l = path length of the cuvette, usually 1 cm
c = concentration of substrate or product

For our example, let us assume that we are measuring the appearance of product that has an E value of 3600 $M^{-1}\mathrm{cm}^{-1}$. According to Beer's Law, the absorbance change in a 3 mL cuvette during 1 minute for production of 1 molar product is 3600:

$$1\ \text{molar} = \frac{A}{3600\ M^{-1}\mathrm{cm}^{-1} \times 1\ \mathrm{cm}}$$

$$A = 3600$$

To convert the raw data to moles of product formed per minute per liter, each ΔA/min is divided by 3600. For example, a ΔA/min of 0.1 is converted to μmoles of product formed (Δc) in the following way:

$$\Delta c = \frac{\Delta A/\mathrm{min}}{El} = \frac{0.1/\mathrm{min}}{3600\ M^{-1}\mathrm{cm}^{-1}\mathrm{cm}} = 2.78 \times 10^{-5}\frac{M}{\mathrm{min}}$$

Now convert to moles/min:

$$2.78 \times 10^{-5}\frac{\text{moles}}{\text{liter} \times \text{min}} \times 0.003\ \text{liters} = 8.33 \times 10^{-8}\frac{\text{moles}}{\text{min}}$$

Now convert to μmoles/min:

$$\Delta c = 8.33 \times 10^{-2}\frac{\mu\text{moles}}{\text{min}}$$

TABLE 8.6 Computer software and databases for this chapter. Web addresses are at the end of this chapter and in Appendix I.

Name	Description
DynaFit	Graphical analysis of ligand-binding data and enzyme kinetic data
CurveFit	Analysis of ligand binding and enzyme kinetics from GraphPad
SigmaPlot	Enzyme kinetics module from Systat
Systat	Statistical and graphical software
ENZYME	General enzyme information from ExPASy
BRENDA	The comprehensive enzyme information system
PDB	Contains the known enzyme structures in Protein Data Bank
Enzyme Kinetics!Pro	Statistical evaluation of enzyme kinetic data
IUBMB	EC nomenclature for enzymes from the International Union of Biochemistry and Molecular Biology

Computer Software for Analysis of Enzyme Kinetic Data

Many of the computer programs listed for use in ligand binding analysis (this chapter, Section A and Table 8.6) may also be used to analyze enzyme kinetic data. A widely used program is DynaFit (www.biokin.com), which is available free to academic users. One purpose of DynaFit is to aid the graphical analysis of enzyme kinetic data. The experimental data may be in the form of initial reaction rates with dependence on the concentration of substrate or inhibitor. Reaction progress curves (time vs. absorbance) may also be analyzed.

In addition to kinetic analysis programs, many Internet databases also contain general information on enzyme names, EC numbers, catalytic mechanisms, cofactors, structures, reactions, kinetics, associated diseases, and other facts. The most popular, current ones are the ENZYME databases, which are part of the Expert Protein Analysis System (ExPASy), BRENDA, and the Enzyme Structure Database (Table 8.6 and Appendix I).

Performing enzyme kinetic/inhibition experiments in the laboratory and analyzing data may be too complex and not appropriate for some levels of students. In addition, equipment and reagents can be quite expensive and difficult to set up. Computer simulation of enzyme kinetic experiments may be an appropriate alternative (Gonzalez-Cruz et al., 2003).

Study Problems

1. A drug, X, was studied for its affinity to serum albumin. When X was bound to albumin, an increase in absorbance was noted. The $\bar{\nu}$ values were determined from these absorbance measurements. Use the data below to determine K_f and n for the interaction between X and albumin. Prepare two types of graphs and compare the results. In one graph plot $\bar{\nu}$ vs. [X], and in the second plot $\bar{\nu}/[X]$ vs. $\bar{\nu}$. Which is the better method? Why? You may also want to try Dynafit for analysis.

[X]	$\bar{v}$
0.36	0.43
0.60	0.68
1.2	1.08
2.4	1.63
3.6	1.83
4.8	1.95
6.0	1.98

2. One of the many straight-line modifications of the Michaelis-Menten equation is the Eadie-Hofstee equation:

$$.\nu_o = -K_M \frac{\nu_0}{[S]} + V_{max}$$

Beginning with the Lineweaver-Burk equation, derive the Eadie-Hofstee equation. Explain how it may be used to plot a straight line from experimental rate data. How are V_{max} and K_M calculated?

3. A research project you are working on involves the study of sugar binding to human albumin. The sugars to be tested are not fluorescent, and you do not wish to use a secondary probe such as a dye. Human albumin has only one tryptophan residue, and you know that this amino acid is fluorescent. You find that the tryptophan fluorescence spectrum of human albumin undergoes changes when various sugars are added. Can you explain the results of this experiment and discuss the significance of the finding?

4. Equation 8.6 in this chapter can be used to determine K_f values, but hyperbolic plots are obtained. Convert Equation 8.6 into an equation that will yield a linear plot without going through all the changes necessary for the Scatchard equation. Hint: See the Michaelis-Menten equation and the Lineweaver-Burk equation.

5. You are attempting to develop a colorimetric probe for use in binding studies. List several requirements that the probe must meet in order to be effective. For example, it must bind at specific locations of the macromolecule. Can you think of other requirements?

6. Using the data in Problems 7 and 8, calculate the specific activity of the enzyme in units/mg and katal/mg. Assume the enzyme has a molecular weight of 55,000 and the reaction mixture for each assay is contained in a total volume of 1.00 mL.

7. The table below gives initial rates of an enzyme-catalyzed reaction, along with the corresponding substrate concentration. Use any graphical method or computer program to determine k_M and V_{max}.

ν_0 (μM/min)	[S] (mole/liter)
130	65×10^{-5}
116	23×10^{-5}
87	7.9×10^{-5}
63	3.9×10^{-5}
30	1.3×10^{-5}
10	0.37×10^{-5}

8. The enzyme concentration used in each assay in Problem 7 is 5×10^{-7}mole/liter. Calculate k_3, the turnover number, in units of $\sec^{-1}$.
9. Compound X was tested as an inhibitor of the enzyme in Problem 6. Use the rate data in Problem 6 and the following inhibition data to evaluate compound X. Is it a competitive or noncompetitive inhibitor?

$[X] = 3.7 \times 10^{-4}$ *M* v_0 (μM/min)	[S] (mole/liter)	$[X] = 1.58 \times 10^{-3}$ *M* v_0 (μM/min)
125	6.5×10^{-4}	110
102	2.3×10^{-4}	80
70	7.9×10^{-5}	40
45	3.9×10^{-5}	20
20	1.3×10^{-5}	—
5	0.37×10^{-5}	—

10. Show the mathematical steps required to derive the Lineweaver-Burk equation beginning with the Michaelis-Menten equation.
11. Figure 8.8 displays the kinetic progress of an enzyme-catalyzed reaction. What time limit must be imposed on rate measurements taken with the use of the fixed-time assay? Why?
12. The following data were collected in a ligand binding experiment. The protein concentration was 7.5×10^{-6} mole/liter. Use a graphical analysis program to determine n and K_f.

Reaction No.	L added (μM)	L bound (μM)
1	20	11
2	50	26
3	100	44
4	150	55
5	200	60
6	400	70

13. In your biochemistry research project, you find that the binding of a ligand to a protein decreases if the ionic strength of the buffer solvent is increased. What type of noncovalent bonding might be involved in the ligand-protein complex?

Further Reading

A. Attie and R. Raines, ***J. Chem. Educ.* 72,** 119–124 (1995). "Analysis of Receptor-Ligand Interactions."

M. Benore-Parsons and K. Sufka, ***Biochem. Mol. Biol. Educ.* 31,** 85–92 (2003). "Teaching Receptor Theory to Biochemistry Undergraduates."

J. Berg, J. Tymoczko, and L. Stryer, ***Biochemistry,*** 6th ed. (2007), Freeman (New York).

A. Bordbar, A. Saboury, and A. Moosavi-Movahedi, ***Biochem. Educ.* 24,** 172–175 (1996). "The Shapes of Scatchard Plots for Systems with Two Sets of Binding Sites."

M. Bradford, ***Anal. Biochem.* 72,** 248–254 (1976). "A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein-Dye Binding."

B. Cantor and P. Schimmel, ***Biophysical Chemistry,*** Part III (1980), Freeman (San Francisco), pp. 849–886. "Ligand Interactions at Equilibrium."

P. Frey and A. Hegeman, ***Enzymatic Reaction Mechanisms*** (2007), Oxford University Press (New York).

R. Garrett and C. Grisham, ***Biochemistry,*** 4th ed. (2010), Brooks Cole-Cengage (Belmont, CA).

J. Gonzalez-Cruz, R. Rodriquez-Sotres, and M. Rodriquez-Penagos, ***Biochem. Mol. Biol. Educ.* 31**, 93–101 (2003). "On the Convenience of Using Computer Simulation to Teach Enzyme Kinetics to Undergraduate Students with Biological Chemistry-related Curricula."

J. Goodrich and J, Kugel, ***Binding and Kinetics for Molecular Biologists*** (2007), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY).

S. Harding and B. Chowdhry, ***Protein-Ligand Interactions: A Practical Approach,*** Vol. 2, (2001), Oxford University Press (New York).

P. Kuzmic, ***Anal. Biochem.* 237,** 260–273 (1996). "Program DYNAFIT for the Analysis of Enzyme Kinetic Data: Application to HIV Protease."

M. Moller and A. Denicola, ***Biochem. Mol. Biol. Educ.* 30,** 309–312 (2002). "Study of Protein-Ligand Binding by Fluorescence."

D. Nelson and M. Cox, ***Lehninger Principles of Biochemistry,*** 5th ed. (2009), Freeman (New York).

N. Price, R. Dwek, M. Wormald, and R. Ratcliffe, ***Principles and Problems in Physical Chemistry for Biochemists,*** 3rd ed. (2002), Oxford University Press (Oxford). Chapter 4 covers ligand binding, and Chapters 9–13 cover enzyme reactions.

N. Price and J. Nairn, ***Exploring Proteins: A Student's Guide to Experimental Skills and Methods,*** (2009), Oxford University Press (New York).

D. Purich, ***Contemporary Enzyme Kinetics and Mechanism,*** 3rd ed. (2009), Academic Press (Burlington, MA).

D. Purich, ***Enzyme Kinetics: Catalysis and Control,*** (2010), Elsevier (New York).

F. Ranaldi, P. Vanni, and E. Gischetti, ***Biochem. Educ.* 27**, 87–91 (1999). "What Students Must Know about the Determination of Enzyme Kinetic Parameters."

G. Rothenberg, ***Catalysis: Concepts and Green Applications*** (2008), Wiley-VCH.

I. Schomburg et al., ***Trends Biochem. Sci.* 27,** 54–56 (2002). "BRENDA: A Resource for Enzyme Data and Metabolic Information."

D. Sheehan, ***Physical Biochemistry: Principles and Applications*** 2nd ed. (2009), Wiley-Blackwell (Hoboken, NJ).

J. Sohl and A. Splittgerber, ***J. Chem. Educ.* 68,** 262–264 (1991). "The Binding of Coomassie Brilliant Blue to Bovine Serum Albumin—A Physical Biochemistry Experiment."

B. Tsai, ***Computational Biochemistry*** (2002), John Wiley & Sons (New York), pp. 107–146. Ligand interactions and enzyme kinetics.

D. Voet and J. Voet, ***Biochemistry,*** 3rd ed. (2005), John Wiley & Sons (Hoboken, NJ).

D. Voet, J. Voet, and C. Pratt, ***Fundamentals of Biochemistry,*** 3rd ed. (2008), John Wiley & Sons (Hoboken, NJ).

http://www.graphpad.com/www/radiolig/radiolig1.htm
GraphPad Software for analyzing binding data. Provides an introduction to analysis of ligand-protein interactions.

http://www.curvefit.com/index.htm
The CurveFit Web site by GraphPad.

http://www.chem.qmw.ac.uk/iubmb/enzyme
The Enzyme List maintained by the International Union of Biochemistry and Molecular Biology.

http://www.biokin.com
Contains the computer program DynaFit, which may be used for the graphical analysis of ligand-receptor binding and enzyme kinetic data.

http://www.systat.com
Computer programs for curve fitting.

http://www.chemsw.com/16029.htm
Enzyme Kinetics!Pro.

http://www.brenda-enzymes.org
The Comprehensive Enzyme Information System.

http://www.biochemweb.org/enzymes.shtml
The Virtual Library of Biochemistry, Molecular Biology and Cell Biology.

http://themedicalbiochemistrypage.org/enzyme-kinetics.html
Enzymes, Kinetics and Diagnostics.

http://www.sigmaplot.com/products/sigmaplot/enzyme-mod.php
Sigmaplot by Systat Software, Inc.

www.expasy.ch/enzyme
General enzyme information.

www.biochem.ucl.ac.uk/bsm/enzymes/index.html
or
http://www.ebi.ac.uk/thornton-srv/databases/enzymes
Enzymes in PDB.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

CHAPTER 9

MOLECULAR BIOLOGY I: STRUCTURES AND ANALYSIS OF NUCLEIC ACIDS

A. Introduction to the Nucleic Acids 268
B. Laboratory Methods for Investigation of DNA and RNA 275

Study Problems 285
Further Reading 286

In this chapter, we shall focus on the **nucleic acids,** biomolecules important for their cellular roles in the storage, transfer, and expression of genetic information. Two fundamental types of nucleic acids participate as genetic molecules: (1) **deoxyribonucleic acid (DNA)** and (2) **ribonucleic acid (RNA).** We begin this chapter with a review of the chemical and biological properties of the nucleic acids and then turn to a discussion of laboratory methods for their isolation and characterization. Several procedures are described here that provide information about the structure and function of DNA and RNA. These include the isolation of chromosomal and plasmid DNA, isolation of RNA, ultraviolet absorption of the nucleic acids, construction of thermal denaturation curves, ethidium bromide binding and fluorescence, agarose gel electrophoresis, and sequencing DNA molecules. Early studies on DNA and RNA delved into understanding the molecular details of DNA replication, RNA transcription, and translation to produce proteins. More recently, research on the nucleic acids has concentrated on the biotechnological development of **recombinant DNA** for molecular cloning procedures, which will be discussed in Chapter 10.

A. INTRODUCTION TO THE NUCLEIC ACIDS

Chemical Components of DNA and RNA

In chemical terms, the nucleic acids are linear polymers made up of monomeric units called **nucleotides.** A nucleotide is comprised of three chemical entities (see Figure 9.1):

1. A purine or pyrimidine nitrogen-containing, heterocyclic base (Figure 9.2).
2. A five-carbon carbohydrate (aldopentose), **β-D-ribose** or **β -D-2-deoxyribose** (Figure 9.3).
3. One, two, or three phosphate groups.

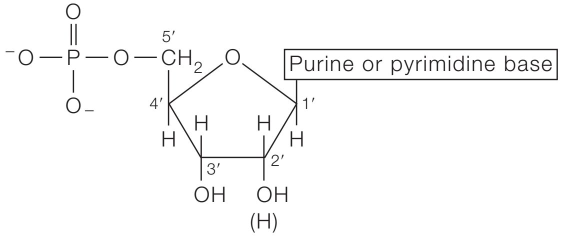

FIGURE 9.1 General structure of a nucleotide showing the three fundamental components: a purine or pyrimidine base, a ribose (or deoxyribose), and phosphate.

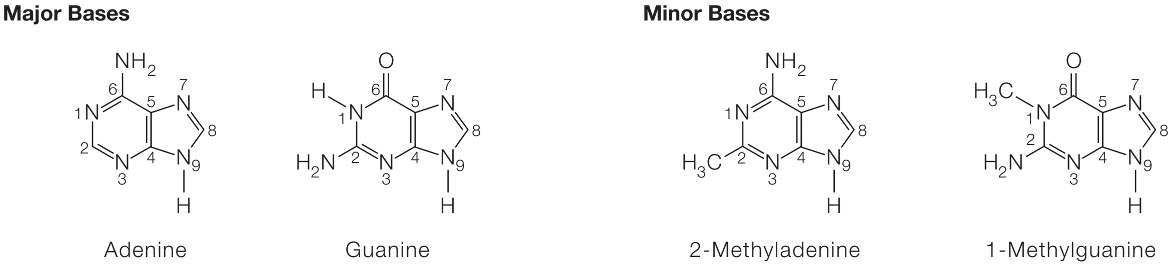

Adenine Guanine 2-Methyladenine 1-Methylguanine

Purines

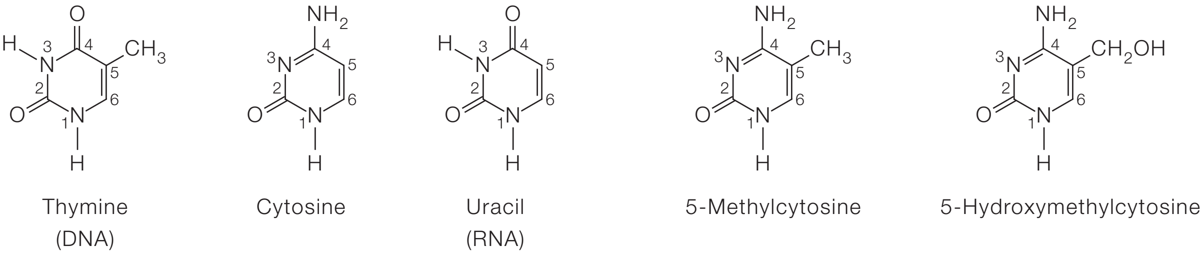

Thymine (DNA) Cytosine Uracil (RNA) 5-Methylcytosine 5-Hydroxymethylcytosine

Pyrimidines

FIGURE 9.2 The major and some minor heterocyclic bases in DNA and RNA. All are derived from purine or pyrimidine.

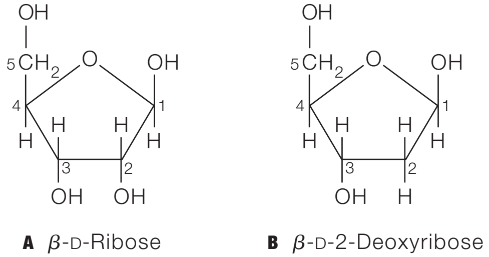

FIGURE 9.3 The aldopentoses in RNA and DNA: **A** β-D-ribose, **B** β-D-2-deoxyribose.

The predominant nitrogen bases in DNA include the purines, **adenine (A)** and **guanine (G),** and the pyrimidines, **thymine (T)** and **cytosine (C).** In RNA, the purines include **A** and **G** and the pyrimidines, **C** and **uracil (U).** The nucleic acids, especially RNA, contain small quantities of methylated nitrogen bases (Figure 9.2).

When a nitrogen base and aldopentose are combined via an *N*-glycosidic bond, the product is a **nucleoside** (Figure 9.4). Addition of a phosphoryl group ($-PO_3^{2-}$) to a hydroxyl group on the carbohydrate leads to a **nucleotide.** The most common site for phosphorylation is the 5′-hydroxyl group. The names of 5′ nucleotides derived from the common nitrogen bases are listed in Table 9.1.

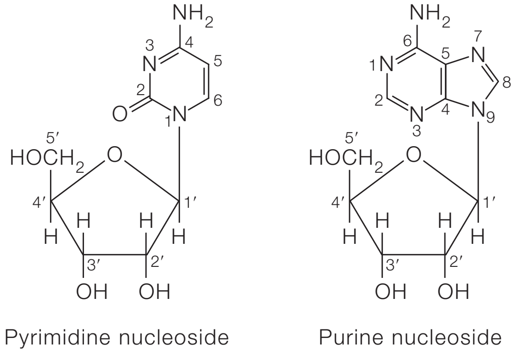

FIGURE 9.4 A nucleoside consists of a purine or pyrimidine base linked to a ribose or deoxyribose by an *N*-glycosidic bond. Two numbering systems (primed and unprimed) are necessary to distinguish the two rings.

TABLE 9.1 Nomenclature for Nucleosides and Nucleotides in DNA and RNA

Base	Nucleoside	Nucleotide (Abbreviation)	Nucleic Acid
Purine			
Adenine	Adenosine, deoxyadenosine	Adenosine 5′-monophosphate (5′-AMP) Deoxyadenosine 5′-monophosphate (5′-dAMP)	RNA DNA
Guanine	Guanosine, deoxyguanosine	Guanosine 5′-monophosphate (5′-GMP) Deoxyguanosine 5′-monophosphate (5′-dGMP)	RNA DNA
Pyrimidine			
Cytosine	Cytidine, deoxycytidine	Cytidine 5′-monophosphate (5′-CMP) Deoxycytidine 5′-monophosphate (5′-dCMP)	RNA DNA
Thymine	Deoxythymidine	Deoxythymidine 5′-monophosphate (5′-dTMP)	DNA
Uracil	Uridine	Uridine 5′-monophosphate (5′-UMP)	RNA

DNA Structure and Function

DNA in all forms of life is a polymer made up of nucleotides containing four major types of heterocyclic nitrogen bases. The nucleotides are held together by **3′,5′-phosphodiester bonds** (Figure 9.5). The quantitative ratio and sequence of bases vary with the source of the DNA. The covalent backbone of DNA (and RNA) consists of alternating deoxyriboses (or riboses) and phosphate groups. This feature, which is defined as a **common, invariant region,** is found in all

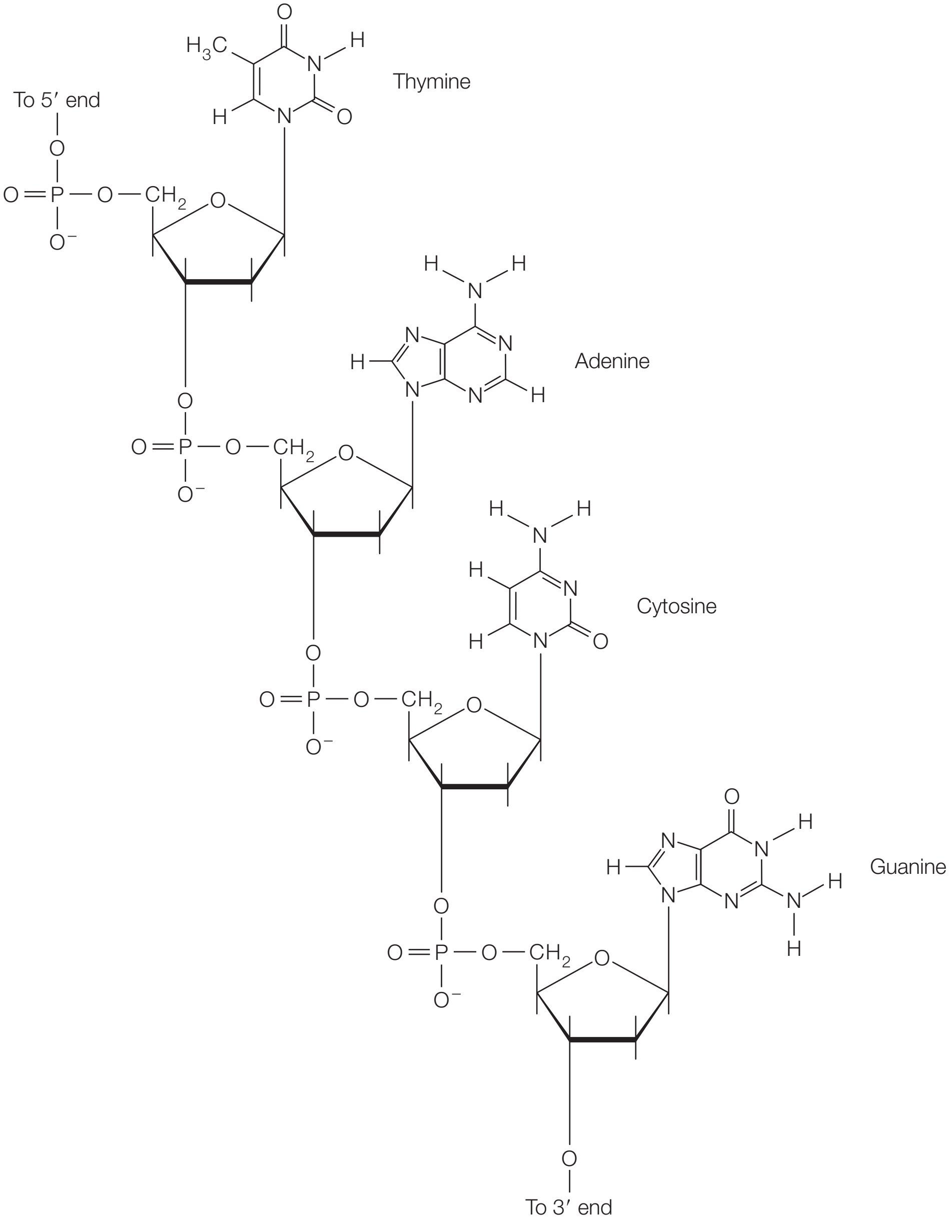

FIGURE 9.5 The covalent structure of DNA showing the phosphodiester backbone linking deoxyribose through the 3′ and 5′ hydroxyl groups.

TABLE 9.2 Comparison of DNA from Different Species

Organism	Number of Base Pairs	Length (μm)	Conformation
Viruses			
SV40	5100	1.7	Circular
Adenovirus	36,000	12	Linear
λ phage	48,600	17	Circular
Bacteria			
E. coli	4,700,000	1400	Circular
Eukaryotes			
Yeast	13,500,000	4600	Linear
Fruit fly	165,000,000	56,000	Linear
Human	3,200,000,000	$1-2 \times 10^6$	Linear

nucleic acids. The **variable region** of DNA and RNA displays the sequence of the four kinds of bases in the nucleic acid. The purine/pyrimidine nitrogen bases protrude from the backbone-like side chains. The sequence of bases carries the specific genetic message. The DNA or RNA chain also displays **directionality**—one end of the chain has a 3′-hydroxyl (or phosphate) group and the other end has a 5′-hydroxyl (or phosphate) group.

Native DNA exists as two complementary, antiparallel strands arranged in a **double helix** held together by noncovalent bonding. DNA in most prokaryotic cells (simple cells with no major organelles and a single chromosome) exists as a single molecule in a circular, double-stranded form with a molecular weight of at least 2×10^9 amu. Eukaryotic cells (cells with major organelles) contain several chromosomes and, thus, several very large DNA molecules (Table 9.2).

DNA was first isolated from biological material in 1869, but its participation in the transfer of genetic information was not recognized until the mid-1940s. Since that time, DNA has been the subject of thousands of physical, chemical, and biological investigations. A landmark discovery was Watson and Crick's elucidation of the three-dimensional structure of DNA by X-ray diffraction analysis in 1953. The double helix as envisioned by Watson and Crick is now recognized as a significant form of native DNA (Figure 9.6). The most important structural and functional feature of the double helix is its **complementary base pairing.** This not only holds the two strands of the double helix together, but also allows the DNA to function in the storage and transfer of genetic information.

The double helix is stabilized by two types of noncovalent forces (review Table 8.2):

1. ***Hydrogen bonding between pairs of complementary bases: A:T and G:C.*** This combination leads to the maximum number of hydrogen bonds for stability. Each base pair consists of a pyrimidine and a purine base. The

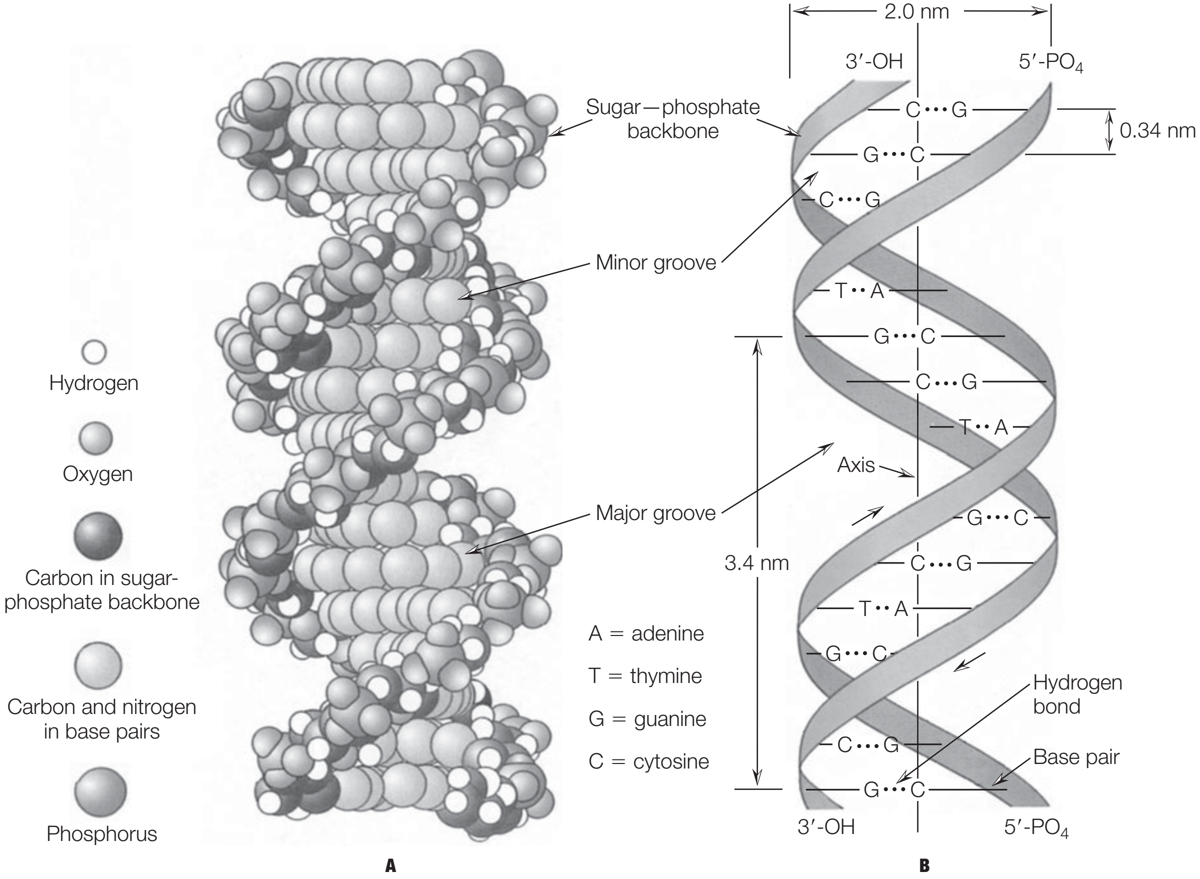
Hydrogen
Oxygen
Carbon in sugar-phosphate backbone
Carbon and nitrogen in base pairs
Phosphorus
Sugar—phosphate backbone
Minor groove
Major groove
A = adenine
T = thymine
G = guanine
C = cytosine
2.0 nm
3′-OH
5′-PO4
0.34 nm
C•••G
G•••C
C•••G
T••A
G•••C
C•••G
T••A
Axis
G•••C
3.4 nm
C•••G
T••A
G•••C
Hydrogen bond
C•••G
Base pair
G•••C
3′-OH
5′-PO4
A
B

FIGURE 9.6 **A** The Watson-Crick double helix. **B** Pairing of A-T and G-C bases in the DNA double helix. Note that the two strands are antiparallel; their 3′ and 5′ phosphodiester bonds run in opposite directions.

distance spanned across the double helix is just right for a purine:pyrimidine pair. Two pyrimidines are too small to form strong hydrogen bonds, and two purines are too sterically crowded.

2. ***Hydrophobic interactions and van der Waals forces between "stacked bases" (see Chapter 8, Section A, p.241).*** The planes of the nitrogen bases are nearly perpendicular to the common axis of the helix and take on a stacking arrangement, bringing the purine and pyrimidine rings close together. This allows for favorable hydrophobic interactions between nonpolar regions of the bases and interactions between π electrons in the sp^2 hybrid orbitals of the aromatic rings.

RNA Structure and Function

Unlike DNA, which is primarily a homogeneous molecule in the cell, RNA exists in three major forms (Table 9.3). **Ribosomal RNA (rRNA),** the most abundant form, is found associated with the ribosomes, the protein-synthesizing

TABLE 9.3 RNA Molecules in *E. coli*

Type	Relative Amount (%)	S^a Value	MW	Average Number of Nucleotides
Ribosomal RNA (rRNA)	80	23S 16S 5S	1.2×10^6 0.55×10^6 36,000	3700 1700 120
Transfer RNA (tRNA)	15	4S	25,000	74–93
Messenger RNA (mRNA)	5	4S	Heterogeneous mixture	

[a] The *S* value refers to the sedimentation coefficient (in Svedberg units) and is related to the size of the molecule.

organelles. The three sizes of prokaryotic rRNA, 5S, 16S, and 23S, can be separated by centrifugation (Chapter 4). **Messenger RNA (mRNA)** carries the transient message for protein synthesis from nuclear DNA to the ribosomes. The smallest RNA molecules, **transfer RNA (tRNA)**, select, bind, and activate amino acids for use in protein synthesis. Because RNA molecules are heterogeneous and short-lived in the cell, it has been a challenge to isolate them in an intact form in order to study their structures and properties. RNA is usually single-stranded, but it does have important structural features. tRNA molecules have been purified and crystallized for X-ray crystallography. The basic two-dimensional structure of all tRNAs is often shown in a cloverleaf pattern in order to display its single-stranded and double-stranded regions. mRNA and rRNA molecules recently isolated and crystallized display the structural elements of hairpin turns, right-handed double helixes, and internal loops. Complementary bases in RNA are A:U and G:C.

Now that some of the major players in the flow of biological information have been identified, it is possible to provide a schematic outline for the process of protein synthesis (see Figure 9.7). The general direction of information flow is:

$$\textit{DNA} \longrightarrow \textit{RNA} \longrightarrow \textit{Proteins} \longrightarrow \textit{Cell Structure/Function}$$

Note that DNA is the original source of genetic information. DNA is faithfully copied by the process of **replication.** In **transcription,** a smaller region of the DNA (gene) is converted into the language of mRNA. In **translation,** proteins are synthesized from amino acids. After synthesis, many proteins are not yet biologically active and must go through **posttranslational processing** before they are functional. Changes that are made in this process include protein folding into the native conformation (assisted by chaperones), protein shortening by proteases, amino acid residue modification (i.e., phosphorylation of seryl hydroxyl groups), attachment of carbohydrate residues, and addition of prosthetic groups. Finished and biologically active proteins play essential roles in the general maintenance, functioning, and structure of the cell.

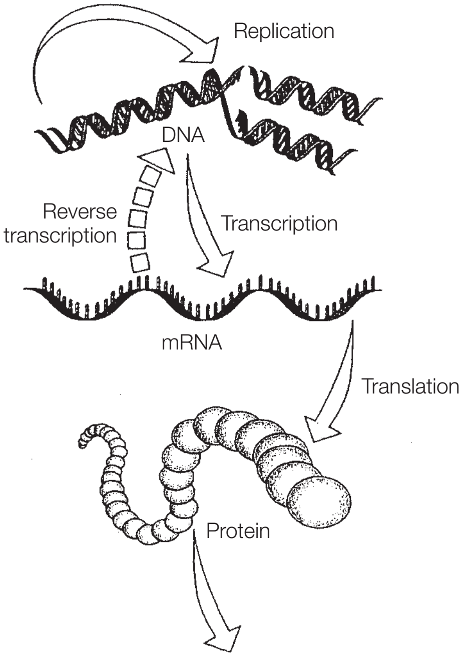

Cell structure and function
- Energy metabolism
- Synthesis and breakdown of biomolecules
- Storage and transport of biomolecules
- Muscle contraction
- Cellular communication (signal transduction)

FIGURE 9.7 The storage and replication of biological information in DNA and its transfer via RNA to synthesize proteins that direct and maintain cellular structure and function.

The nucleic acids are among the most complex molecules that you will encounter in your biochemical studies. When the dynamic roles that are played by DNA and RNA in the life of the cell are realized, the complexity is understandable. The discovery of the structures and functions of the nucleic acids will always be considered as some of the most significant breakthroughs in our understanding of the chemistry of life.

This section has provided a brief review of many important concepts in molecular biology. For more details, refer to one of the standard biochemistry textbooks and Web sites listed in the Further Reading section at the end of the chapter.

STUDY EXERCISE 9.1 Complementary Bases in DNA

A short polynucleotide strand of DNA has the following sequence of bases. Write the sequence of the complementary strand.

5′AGCTTACGTCC 3′

STUDY EXERCISE 9.2 Complementary Bases in RNA

A short polynucleotide strand of double-stranded RNA has the following sequence of bases. Write the sequence of the complementary strand of RNA.

5′UAGGUACUUCG 3′

B. LABORATORY METHODS FOR INVESTIGATION OF DNA AND RNA

Isolation of Chromosomal DNA

Because of the large size and the fragile nature of chromosomal DNA, it is difficult to isolate in a completely intact, undamaged form. Several isolation procedures have been developed that provide DNA in a biologically active form, but this does not mean it is completely undamaged. These preparations yield DNA that is stable, of high molecular weight, and relatively free of RNA and proteins. Here a general method will be described for the isolation of chromosomal DNA, in a relatively pure form, from microorganisms.

Designing an isolation procedure for DNA requires extensive knowledge of the chemical stability of DNA, as well as of its condition in the cellular environment. Figures 9.5 and 9.6 illustrate several chemical bonds in DNA that may be susceptible to cleavage during the extraction process. The experimental factors that must be considered and their effects on various structural aspects of intact DNA are outlined below.

1. **pH**
   (a) Hydrogen bonding between the complementary strands is stable between pH 4 and 10.
   (b) The phosphodiester linkages in the DNA backbone are stable between pH 3 and 12.
   (c) *N*-glycosidic bonds to purine bases (adenine and guanine) are hydrolyzed at pH values of 3 and less.
2. **Temperature**
   (a) There is considerable variation in the temperature stability of the hydrogen bonds in the double helix, but most DNA begins to unwind in the range of 80−90°C.
   (b) Phosphodiester linkages and *N*-glycosidic bonds are stable up to 100°C.
3. **Ionic strength**
   (a) DNA is most stable and soluble in salt solutions. Salt concentrations of less than 0.05 *M* weaken the hydrogen bonding between complementary strands.
4. **Cellular conditions**
   (a) Before the DNA can be released, the cell wall of the organism must be lysed. The ease with which the cell wall is disrupted varies from

organism to organism. In some cases (e.g., yeast), extensive grinding or sonic treatment is required, whereas in others (e.g., *Bacillus subtilis*), enzymatic hydrolysis of the cell wall is possible.

(b) Several enzymes are present in the cell that may act to degrade DNA, but the most serious damage is caused by the deoxyribonucleases. These enzymes catalyze the hydrolysis of phosphodiester linkages.

(c) Native DNA is present in the cell as DNA-protein complexes. Basic proteins called **histones** must be dissociated from the DNA during the extraction process.

5. **Mechanical stress on the DNA**

(a) Gentle manipulations may not always be possible during the isolation process. Grinding, shaking, stirring, and other disruptive procedures may cause cleavage (shearing or scission) of the DNA chains. This usually does not cause damage to the secondary structure of the DNA, but it does reduce the length of the molecules.

Now that these factors are understood, a general procedure of DNA extraction from bacteria will be outlined.

Step 1. Disruption of the cell wall and release of the DNA into a medium in which it is soluble and protected from degradation The isolation procedure described here calls for the use of an enzyme, lysozyme, to disrupt the cell wall. Lysozyme catalyzes the hydrolysis of glycosidic bonds in cell wall peptidoglycans, thus causing destruction of the cell wall and allowing the release of DNA and other cellular components. The medium for solution of DNA is a buffered saline solution containing EDTA. DNA, which is ionic, is more soluble and stable in salt solution than in distilled water. The EDTA serves at least two purposes. First, it binds divalent metal ions (Ca^{2+}, Mg^{2+}, Mn^{2+}) that could form salts with the anionic phosphate groups of the DNA. Second, it inhibits deoxyribonucleases that have a requirement for Mg^{2+} or Mn^{2+}. The mildly alkaline medium (pH 8) acts to reduce electrostatic interaction between DNA and the basic histones and the polycationic amines, spermine and spermidine. The relatively high pH also tends to diminish nuclease activity and denature other proteins.

Step 2. Dissociation of the protein-DNA complexes Detergents are used at this stage to solubilize the inner membrane and disrupt the ionic interactions between positively charged histones and the negatively charged backbone of DNA. Sodium dodecyl sulfate (SDS), an anionic detergent, binds to proteins and gives them extensive anionic character. A secondary action of SDS is to denature deoxyribonucleases and other proteins. Also favoring dissociation of protein-DNA complexes is the alkaline pH, which reduces the positive character of the histones. To ensure complete dissociation of the DNA-protein complex and to remove bound cationic amines, a high concentration of a salt (NaCl or sodium perchlorate) is added. The salt acts by diminishing the ionic interactions between DNA and cations.

Step 3. Separation of the DNA from other soluble cellular components Before DNA is precipitated, the solution must be deproteinized. This is brought about by treatment with chloroform–isoamyl alcohol followed by centrifugation. Upon centrifugation, three layers are produced: an upper aqueous phase, a lower organic layer, and a compact band of denatured protein at the interface between the aqueous and organic phases. Chloroform causes surface denaturation of proteins. Isoamyl alcohol reduces foaming and stabilizes the interface between the aqueous phase and the organic phases where the protein collects.

The upper aqueous phase containing nucleic acids is then separated and the DNA precipitated by addition of ethanol. Because of the ionic nature of DNA, it becomes insoluble if the aqueous medium is made less polar by addition of an organic solvent. The DNA forms a threadlike precipitate that can be collected by "spooling" onto a glass rod. The isolated DNA may still be contaminated with protein and RNA. Protein can be removed by dissolving the spooled DNA in saline medium and repeating the chloroform–isoamyl alcohol treatment until no more denatured protein collects at the interface.

RNA does not normally precipitate like DNA, but it could still be a minor contaminant. RNA may be degraded during the procedure by treatment with ribonuclease after the first or second deproteinization step. Removal of RNA sometimes makes it possible to denature more protein using chloroform–isoamyl alcohol. If DNA in a highly purified state is required, several deproteinization and alcohol precipitation steps may be carried out. It is estimated that up to 50% of the cellular DNA is isolated by this procedure. The average yield is 1 to 2 mg per gram of wet packed bacterial cells.

Isolation of Plasmid DNA

Many bacterial cells contain self-replicating, extrachromosomal DNA molecules called **plasmids.** This form of DNA is closed, circular, double-stranded, and much smaller than chromosomal DNA; its molecular weight ranges from 2×10^6 to 20×10^6, which corresponds to between 3000 and 30,000 base pairs. Plasmids are widely used as cloning vehicles in the preparation of recombinant DNA (see Chapter 10, p. 294). Bacterial plasmids normally contain genetic information for the translation of proteins that confer a specialized and sometimes protective characteristic (phenotype) on the organism. Examples of these characteristics include enzyme systems that degrade antibiotics, and enzymes necessary for the production of antibodies and toxins. Plasmids are replicated in the cell by one of two possible modes. **Stringent replicated plasmids** are present in only a few copies and **relaxed replicated plasmids** are present in many copies, sometimes up to 200. Some relaxed plasmids continue to be produced even after the antibiotic chloramphenicol is used to inhibit chromosomal DNA synthesis in the host cell. Under those conditions, many copies of the plasmid DNA may be produced (up to 2000 or 3000), and may compose up to 30 to 40% of the total cellular DNA.

The extensive use of plasmid DNA as a cloning vehicle often requires the isolation and characterization of plasmids. Characterization of plasmids might be necessary for any of the following reasons (see Chapter 10):

1. Construction of new recombinant DNA.
2. Analysis of molecular size by agarose gel electrophoresis.

3. Electrophoretic analysis of restriction enzyme digests and construction of a restriction enzyme map.
4. Sequence analysis of nucleotides by the Sanger or Maxam-Gilbert method.

Several methods for isolating plasmid DNA have been developed; some lead to a more highly purified product than others. All isolation methods have the same objective—separation of plasmid DNA from chromosomal DNA. Plasmid DNA has two major structural differences from chromosomal DNA.

1. Plasmid DNA is almost always extracted in a covalently closed circular form, whereas isolated chromosomal DNA usually consists of sheared linear fragments, and,
2. Plasmid DNA is much smaller than chromosomal DNA.

The structural differences cause physicochemical differences that can be exploited to separate the two types of DNA molecules. Methods for isolating plasmid DNA fall into three major categories:

1. Methods that rely on specific interaction between plasmid DNA and a solid support. Examples are adsorption to nitrocellulose microfilters and hydroxyapatite columns.
2. Methods that cause selective precipitation of chromosomal DNA by various agents. These methods exploit the relative resistance of covalently closed circular DNA to extremes of pH, temperature, or other denaturing agents.
3. Methods based on differences in sedimentation behavior between the two types of DNA. This is the approach of choice if highly purified plasmid DNA is required.

Two widely used isolation procedures based on Method 2 are described here. Both procedures yield plasmid DNA that is sufficiently pure for size analysis by agarose gel electrophoresis and for digestion by restriction enzymes.

Method A: Separation of Plasmid DNA by Boiling (Holmes and Quigley)

The total cellular DNA must first be released by lysis of the bacterial cells. This is brought about by incubation with the enzyme lysozyme in the presence of reagents that inhibit nucleases. Chromosomal DNA is then separated from the plasmid DNA by boiling the lysis mixture for a brief period, followed by centrifugation. In contrast to closed circular plasmid DNA, linear chromosomal DNA becomes irreversibly denatured by heating and forms an insoluble gel, which sediments during centrifugation. Even though plasmid DNA may become partially denatured during boiling, the closed circular helix reforms upon cooling. The boiling serves a second purpose, that of denaturing deoxyribonucleases and other proteins.

Method B: Microscale Isolation of Plasmids by Alkaline Lysis

Often it is necessary to detect and analyze plasmid DNA in a large number of small bacterial samples. One widely used microscale method is the alkaline lysis procedure. For this procedure, host bacterial cells harboring the plasmids are grown in small culture volumes (1–5 mL) or in single colonies on agar plates. The cells are lysed and their contents denatured by alkaline sodium dodecyl sulfate (SDS). Proteins and high-molecular-weight chromosomal DNA denatured under these conditions precipitate as a gel that can be centrifuged from the supernatant, which contains plasmid DNA and bacterial RNA.

Characterization of DNA

Ultraviolet Absorption

A complete understanding of the biochemical functions of DNA requires a clear picture of its structural and physical characteristics. DNA has significant absorption in the UV range because of the presence of the aromatic bases adenine, guanine, cytosine, and thymine. This provides a useful probe into DNA structure because structural changes such as helix unwinding affect the extent of absorption. In addition, absorption measurements are used as an indication of DNA purity. The major absorption band for purified DNA peaks at about 260 nm. Protein material, the primary contaminant in DNA, has a peak absorption at 280 nm. The ratio A_{260}/A_{280} is often used as a relative measure of the nucleic acid/protein content of a DNA sample. The typical A_{260}/A_{280} for isolated DNA is about 1.8. A smaller ratio indicates increased contamination by protein.

Thermal Denaturation

If DNA solutions are treated with denaturing agents (heat, alkali, organic solvents), their ultraviolet-absorbing properties are strikingly increased. Figure 9.8 shows the effect of temperature on the UV absorption of DNA. The curve is obtained by plotting $A_{260(T)}/A_{280(25^\circ)}$ vs. temperature (T). Heating through the

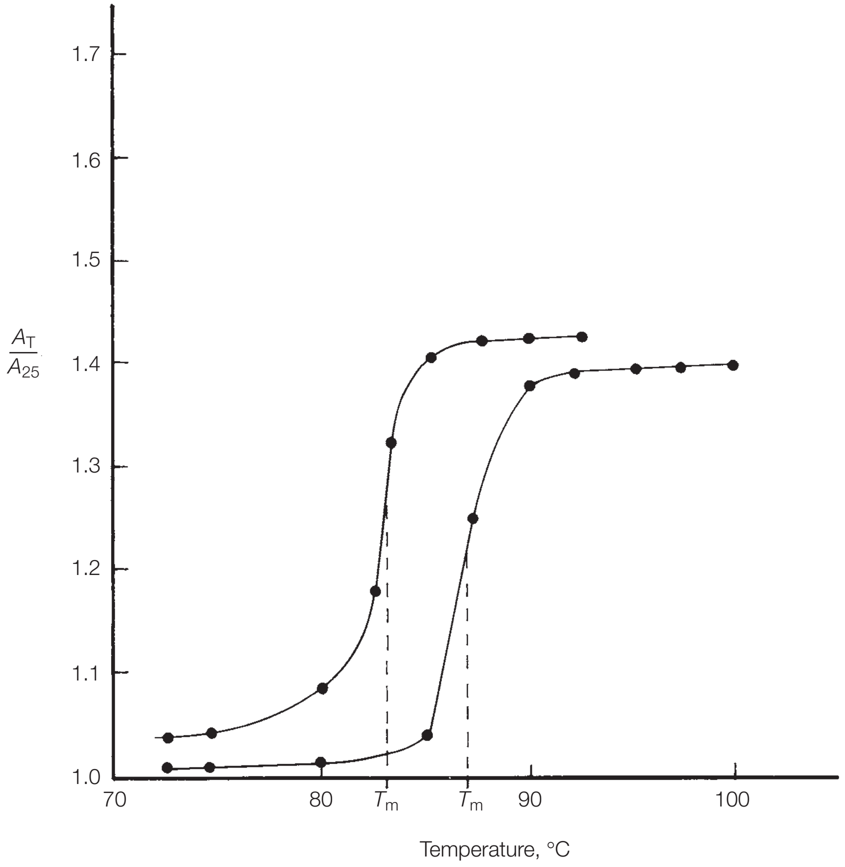

FIGURE 9.8 A typical thermal denaturation curve for DNA. T_m is measured as the midpoint of the absorbance increase.

temperature range of 25°C to about 80°C results in only minor increases in absorption. However, as the temperature is further increased, there is a sudden increase in UV absorption followed by a constant A_{260}. The total increase in absorption is usually on the order of 40% and occurs over a small temperature range. Figure 9.8 is called a **thermal denaturation curve,** a **temperature profile,** or a **melting curve.** The temperature corresponding to the midpoint of each absorption increase is defined as T_m, the **transition temperature** or **melting temperature.** (This should not be confused with melting point, the transformation of a substance from solid to liquid, as was routinely studied in the organic laboratory.) Each species of DNA has a characteristic T_m value that can be used for identification and characterization purposes.

The origin of the absorption increase, called a **hyperchromic effect,** is well understood. The absorption changes are those that result from the transition of an ordered double-helix DNA structure to a denatured state or random, unpaired DNA strands. Native DNA in solution exists in the double helix held together primarily by hydrogen bonding between complementary base pairs on each strand (see Figure 9.6). Hydrophobic and π-π interactions between stacked base pairs also strengthen the double helix. Agents that disrupt these forces (hydrogen bonding, hydrophobic and π-π interactions) cause dissociation or unwinding of the double helix. In a random coil arrangement, base-base interactions are at a minimum; this alters the resonance behavior of the aromatic rings, causing an increase in absorption. The process of DNA dissociation can, therefore, be characterized by monitoring the UV-absorbing properties of DNA under various conditions.

Ethidium Bromide Binding and Fluorescence

Fluorescence assays are considered to be among the most convenient, sensitive, and versatile of all laboratory techniques (Chapter 7, p. 220). However, the purine and pyrimidines of the nucleic acids yield only weak fluorescence spectra. The fluorescence of the dye, ethidium bromide, is enhanced about 25-fold when it interacts with DNA. Ethidium bromide, which is a relatively small planar molecule (Figure 9.9), binds to DNA by insertion between stacked base pairs (intercalation). The process of intercalation is especially significant for aromatic dyes, antibiotics, and other drugs. Some dyes, when intercalated into DNA, show enhanced fluorescence that can be used to detect DNA molecules after gel electrophoresis measurements (Chapter 6, Section C, p. 191).

The concentration of RNA and DNA solutions may be determined using ethidium bromide binding. The spectrophotometric assay for DNA/RNA (Chapter 3, p. 71) measures both double-stranded and single-stranded DNA because it measures the UV absorption by purine and pyrimidine bases. Ethidium

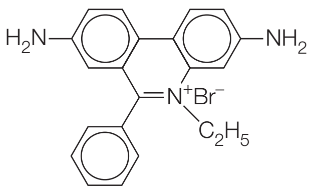

FIGURE 9.9 Structure of the fluorescent intercalation dye ethidium bromide.

bromide interaction with single-stranded DNA does not lead to increased fluorescence, so duplex DNA can be quantified in the presence of dissociated DNA. Solutions of purified DNA are commonly contaminated with RNA. Because single-stranded RNA can form hairpin loops with base pairing and duplex formation (as in tRNA), ethidium bromide also binds with enhanced fluorescence to these duplex regions of RNA. Addition of ribonuclease A results in digestion of RNA and loss of fluorescence due to ethidium binding of RNA. The amount of fluorescence lost is proportional to the concentration of RNA. The ethidium fluorescence remaining after ribonuclease treatment is directly proportional to the concentration of duplex DNA.

$$F(\text{total}) = F(\text{DNA}) + F(\text{RNA})$$

where

F = fluorescence yield due to each type of nucleic acid

When the F (RNA) term is reduced to zero, the total fluorescence is a direct measurement of concentration of double-stranded DNA. The actual concentration of DNA in solution can be calculated by using a standard solution of DNA or from a standard curve.

STUDY EXERCISE 9.3 Ethidium Bromide-Fluorescence Assay for DNA

Solution A of DNA (unknown concentration) is known to contain double-stranded DNA, damaged single-stranded DNA, and RNA. The ethidium bromide fluorescence assay was done on the solution, and the following data were collected. The fluorescence intensity values (F) were obtained on an instrument with a standardized scale of 0–100.

F of solution A = 87

F after treatment of A with ribonuclease = 80

F of a standard solution of double-stranded DNA (10 μg/mL) = 100

What is the concentration of double-stranded DNA in solution A in μg/mL?

Solution:

We begin by making several assumptions. The total fluorescence of solution A is due to double-stranded DNA and some contaminating RNA. Single-stranded DNA and RNA do not bind ethidium bromide. The concentration of double-stranded DNA is directly proportional to the fluorescence reading. The fluorescence intensity due to double-stranded DNA is 80. To calculate the concentration of double-stranded DNA in solution A, compare the fluorescence intensity (80) with that obtained for the standard DNA solution:

$$[\text{DNA-A}]/80 = [10\ \mu\text{g/mL}]/100$$

$$[\text{DNA-A}] = 8\ \mu\text{g/mL}$$

Caution

Ethidium bromide must be used with great care, as it is a potent mutagen. Gloves should be worn at all times while handling solutions of the dye. When finished with experiments, used and excess solutions of the dye must be disposed of as directed by your instructor or laboratory director.

Agarose Gel Electrophoresis

Several techniques for the characterization of nucleic acids have been introduced in this chapter, but the standard method for separation and analysis of plasmids and other smaller DNA molecules is agarose gel electrophoresis (Chapter 6, Section B, p.177). This method has several advantages, including ease of operation, sensitive staining procedures, high resolution, and the ability to analyze a wide range of molecular weights ($0.6 - 100 \times 10^6$). As discussed in Chapter 6, the mobility of nucleic acids in agarose gels is influenced by the agarose concentrations, the molecular size of the DNA, and the molecular shape of the DNA. In general, the lower the agarose concentration in the gels, the larger the DNA that can be analyzed. A practical lower limit of agarose concentration is reached at 0.3% agarose, below which gels become too fragile for ordinary use. This lower limit of agarose in the gels allows for analysis of linear double-stranded DNA within the range of 5 and 60 kilobase pairs (up to 150×10^6 in molecular weight). Gels with an agarose concentration of 0.8% can separate DNA in the range of 0.5–10 kilobase pairs, and 2% agarose gels are used to separate smaller DNA fragments (0.1–3 kilobase pairs).

Nucleic acids migrate in an agarose medium at a rate that is inversely proportional to their size (kilobase pairs or molecular weight). In fact, a near-linear relationship exists between mobility and the logarithm of kilobase pairs (or molecular weight) of a DNA fragment. A standard curve may be prepared by including on the gel a sample containing DNA fragments of known molecular weights (see Figure 9.10). Several types of molecular-weight ladders are commercially available to mark and help estimate the sizes and concentrations of unknown nucleic acids.

Agarose gel electrophoresis is an ideal technique for analysis of DNA fragments. In addition to the positive characteristics discussed previously, the technique is simple, rapid, and relatively inexpensive. Fragments that differ in molecular weight by as little as 1% can be resolved on agarose gels, and as little as 1 ng of DNA can be detected on a gel. Nucleic acids are visualized after electrophoresis by treatment with ethidium bromide or one of the less toxic SYBR and GelRed dyes mentioned in Chapter 6, p. 190.

Sequencing DNA Molecules

When it was determined that genetic information in DNA is coded in the form of nucleotide base sequence, the direction of research turned to the design of experimental procedures for determining the base order. Early sequence studies used the inefficient, insensitive, and labor-intensive methods of acid-, base-, and nuclease-catalyzed hydrolysis of DNA and analysis of fragments by electrophoresis.

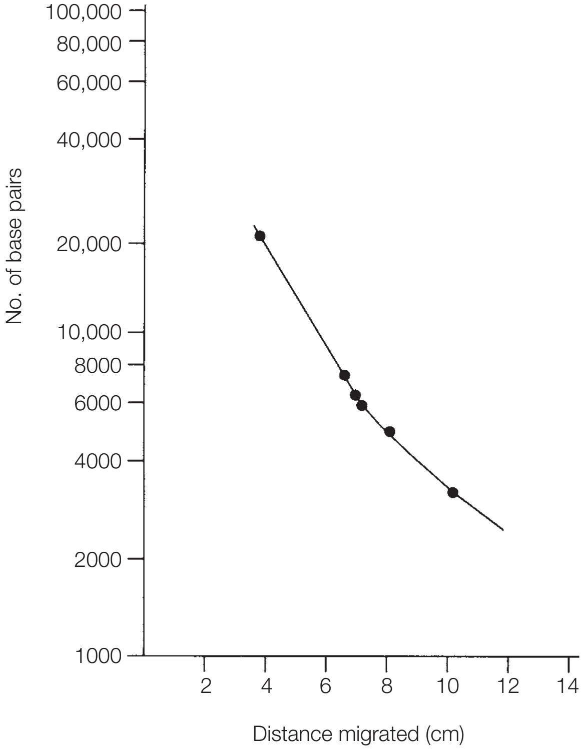

FIGURE 9.10 Standard curve obtained by electrophoresis of λ phage DNA fragments from EcoR1 cleavage.

Two sequencing methods that are both simple and inexpensive are now widely used:

1. The **Maxam-Gilbert chemical cleavage method** (also called the chemical degradation method) uses chemical modification of bases, which promotes cleavage at selected phosphodiester linkages. Cleavage products, which have one common end and vary in length, are analyzed by PAGE.
2. The **Sanger chain-terminating method** (also called the dideoxy method) uses the enzyme DNA polymerase, coupled with the presence of all four nucleoside triphosphates and one of the 2′, 3′-dideoxynucleoside triphosphates, to synthesize a DNA chain complementary to an added template. Replication stops when a dideoxynucleotide is in place on the new chain. Fragments of different lengths are produced and analyzed by PAGE.

In both sequencing methods, electrophoretic procedures are capable of separating nucleic acid fragments that differ in size by only one nucleotide. Detection of the fragments is done with chemiluminescent-labeling or ^{32}P-labeling of deoxynucleoside triphosphates (see Figure 6.16, p. 192).

Sequencing procedures that are automated and computer-controlled have now made possible the sequencing of entire genomes (Table 9.4). The first

TABLE 9.4 Sequenced Genomes

Organism	Genome Size (Kilobase Pairs)	Number of Chromosomes
Borrelia burgdorferi (carrier of Lyme disease)	1444	1
Haemophilus influenzae (human pathogenic bacterium)	1830	1
Mycobacterium tuberculosis (cause of tuberculosis)	4412	1
Escherichia coli (bacterium)	4639	1
Saccharomyces cerevisiae (yeast)	11,700	17*
Drosophila melanogaster (fruit fly)	137,000	4*
Oryza sativa (rice)	430,000	12
Homo sapiens (human)	3,200,000	23*

*in haploid chromosomes

complete genome of a free-living organism, the prokaryote *Haemophilus influenzae*, was published in 1995. This genome contains 1,830,137 bases and 1740 genes. Complete sequencing of the human genome was announced in 2001. The human genome contains over 3.2 billion base pairs and about 20,000–25,000 genes.

Genomic sequences are now made available in repositories on the Internet. Three primary resources, combined in the International Nucleotide Sequence Database Collaboration, are listed in Table 9.5. Database resources for genome projects are also listed in Table 9.5 and in Appendix I. Readers may practice using the databases by working Study Problem 12.

Isolation and Characterization of RNA

The methods used for the isolation of RNA are similar to those described for DNA at the beginning of this chapter, Section B. However, some procedural changes must be made in the purification of RNA because of the presence of

TABLE 9.5

Organization	Web site
The International Nucleotide Sequence Database Collaboration	
1. Genbank of the National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov/Genbank
2. EMBL (European Molecular Biology Laboratory)	http://www.ebi.ac.uk
3. DDBJ (DNA Data Bank of Japan)	http://www.ddbj.nig.ac.jp
Genome Project Databases	
1. dbEST (cDNA and partial sequences)	http://www.ncbi.nih.gov
2. Genethon (genetic maps based on repeat markers)	http://www.genethon.fr
3.The Broad Institute of MIT and Harvard University (Genomics)	http://www.broadinstitute.org

heterogeneous populations of molecules and the extra susceptibility to cleavage of the phosphodiester bonds in RNA. Here are some general comments on laboratory methods for the isolation and characterization of RNA:

- RNA molecules are naturally shorter than DNA. This makes RNA less susceptible to physical shearing, so more vigorous conditions may be used for cell lysis.
- The phosphodiester bonds in DNA are stable in the pH range of 3–12; however, the presence of the $2'$-OH groups in RNA makes it especially susceptible to base- and enzyme-catalyzed (RNase) cleavage. Special precautions must be taken to avoid chemical agents that degrade RNA, and chemicals must be added to isolation buffers that destroy endogenous proteins that act as nucleases. The addition of chelating agents like EDTA and citrate ties up metal ions that are essential for RNase activity. Strong detergents may be added to denature proteins, especially nucleases. The chemical reagent guanidinium thiocyanate is a useful addition, as it acts as an RNase inhibitor and protein denaturant.
- It is essential for lab workers to wear plastic gloves to avoid transfer of nucleases from bare skin.
- Gradient centrifugation procedures as described in Chapter 4 may be used to isolate, purify, and characterize all types of RNA.
- Isolated RNA samples may be analyzed for purity by agarose gel electrophoresis procedures similar to those for DNA.
- Specific types of RNA may be isolated and purified using affinity chromatography. For example, eukaryotic mRNAs, which make up only about 2–5% of cellular RNA, have a poly (A) segment at their $3'$ ends. This RNA may be purified using an affinity matrix consisting of poly (T) or poly (U) on agarose or cellulose gels (see Chapter 5, Section G, p. 152).

STUDY EXERCISE 9.4

Use PubMed or another computer search method to find new methods for isolation of RNA.

Study Problems

1. During the isolation of DNA, the solution becomes more viscous after treatment with lysozyme and SDS. Explain why.
2. How can RNA be removed from a DNA preparation?
3. Explain the action of SDS in disrupting the cell membrane in the extraction procedure of DNA.
4. A solution of purified DNA gave an absorbance of 0.55 at a wavelength of 260 nm. The absorbance was measured in a quartz cuvette with a path length of 1 cm. What is the concentration of DNA? (See Chapter 3, Section C, p. 71).
5. The concentration of a purified DNA solution is 35 μg/mL. Predict the absorbance of the solution at 260 nm.
6. Why is bromophenol blue dye added to a gel-loading buffer before electrophoresis?

7. What assumptions must be made about the relative mobility of bromophenol blue dye and DNA fragments during electrophoresis?

8. What is the purpose of the ethanol precipitation step in the preparation of plasmids?

9. Predict and explain the effect of each of the following conditions on the T_m of native DNA. Will the T_m be raised or lowered relative to the DNA in Tris-HCl buffer, pH 7.5?
   (a) Measure T_m in pH 12 buffer.
   (b) Measure T_m in distilled water.
   (c) Measure T_m in 50% methanol water.
   (d) Measure T_m in standard Tris-HCl, pH 7.5 buffer solution containing SDS.

10. The polyamines 1,4-diaminobutane, 1,5-diaminopentane, spermine, and spermidine are metabolic products found in many cells. Although the specific function of these compounds has not been determined, they are known to bind to the nucleic acids. Describe how you would combine the ethidium bromide fluorescence assay described in this chapter with ligand binding plots defined in Chapter 8, Section A, to characterize the binding of the polyamines to DNA or RNA. Hint: The polyamines are found to bind to some of the same sites on the nucleic acids as does ethidium bromide.

11. A solution of purified RNA gave an absorbance of 0.75 at 260 nm. The absorbance was measured in a quartz cuvette with path length of 1 cm. What is the concentration of RNA in the solution?

12. Use one of the DNA sequence databases in Table 9.5 to obtain the DNA sequence of:
   (a) Cytosolic *t*RNA-phenylalanine from yeast (Saccharomyces cerevisiae)
   (b) human insulin

Further Reading

B. Alberts, et al., ***Molecular Biology of the Cell,*** 5th ed. (2007), Garland Press (New York).

J. Berg, J. Tymoczko, and L. Stryer, ***Biochemistry,*** 6th ed. (2007), Freeman (New York).

H. Birnboim, ***Methods Enzymol.*** **100**, 243–255 (1983). "A Rapid Alkaline Extraction Method for the Isolation of Plasmid DNA."

R. Boyer, ***Concepts in Biochemistry,*** 3rd ed. (2006), John Wiley & Sons (Hoboken, NJ). Chapters 10–13 cover the concepts in molecular biology.

F. Collins, et al., ***Nature*** **409**, 745–708 (2001). The human genome.

R. Garrett and C. Grisham, ***Biochemistry,*** 4th ed. (2010), Brooks Cole-Cengage (Belmont, CA).

D. Holmes and M. Quigley, ***Anal. Biochem.*** **114**, 193–197 (1981). "A Rapid Boiling Method for the Preparation of Bacterial Plasmids."

C. Liu, ***Handbook of Nucleic Acid Purification,*** (2009), CRC Press (Boca Raton, FL).

J. Marmur, ***J. Mol. Biol.*** **3**, 208 (1961). "A Procedure for the Isolation of Deoxyribonucleic Acid from Microorganisms."

S. Neidle, ***Principles of Nucleic Acids,*** 2nd ed. (2007), Academic Press (Reston, VA).

D. Nelson, and M. Cox, ***Lehninger Principles of Biochemistry,*** 5th ed. (2009), Freeman (New York).

J. Sambrook and D. Russell, ***Molecular Cloning: A Laboratory Manual,*** 3rd ed. (2001), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY).

C. Tsai, ***Computational Biochemistry*** (2002), Wiley–Liss (New York). Chapter 9 covers nucleotide sequencing.

J. Venter, ***Science*** **291,** 1304 (2001). The sequence of the human genome.

D. Voet and J. Voet, ***Biochemistry,*** 3rd ed. (2005), John Wiley & Sons (Hoboken, NJ). Chapters 5–7 and 29–35 cover DNA and RNA.

D. Voet, J. Voet, and C. Pratt, ***Fundamentals of Biochemistry***, 3rd ed. (2008). John Wiley & Sons (Hoboken, NJ).

http://www.cgcsci.com/cminfo.shtml
Restriction map of pBR322 in circular and linear format. Designed by CGC Scientific, a provider of software tools for biologists.

http://www.fermentas.com
Click on "conventional restriction endonucleases" for a listing of enzymes.
Click on individual enzymes for specific data including cleavage site and reaction conditions.

http://biowww.net
Molecular Cell Biology Protocol Troubleshooting.

http://learn.genetics.utah.edu/content/labs/extraction
DNA Extraction Virtual Lab, "How to Extract DNA from Anything Living."

http://www.intute.ac.uk/biologicalsciences
Helping you find the best Web sites for study and research.

http://www.genome.ou.edu/protocol_book/protocol_index.html
Protocols for recombinant DNA isolation, cloning and sequencing.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

This page intentionally left blank

CHAPTER 10

MOLECULAR BIOLOGY II: RECOMBINANT DNA, MOLECULAR CLONING, AND ENZYMOLOGY

A. Recombinant DNA Biotechnology 290
B. Important Enzymes in Molecular Biology and Biotechnology 297
C. Nucleic Acid Blotting 304

Study Problems 304
Further Reading 305

Investigation of the nucleic acids during the first three quarters of the 20th century led to the significant discovery that all of the cell's attributes and activities have their origin in DNA. For example, we now believe that the information needed to make all the proteins in a cell and organism resides in DNA. This discovery prompted biochemists and molecular biologists to develop laboratory procedures for the manipulation of DNA, because it was suspected that changes to DNA could lead to molecular and genetic changes in cells and organisms. A culmination of this research activity was the recent announcement of the sequence of the human genome (Human Genome Project, 2001). Medical scientists are now applying the newly developed lab and clinical procedures to the area of gene therapy, where, dysfunctional genes that lead to disease are replaced with functional genes that produce viable proteins.

In this chapter, we will discuss some experimental procedures that have moved us to the point where the nucleic acids, especially DNA, are now among the easiest biomolecules to work with. Some of the concepts we will encounter include **molecular cloning, recombinant DNA, restriction enzymes, nucleic acid blotting,** and the **polymerase chain reaction.**

A. RECOMBINANT DNA BIOTECHNOLOGY

Students often perceive biochemistry as a very research-oriented discipline. It is easy to see how this idea can develop, because every day students listen to their instructors talk about research activities, attend research seminars, and watch their professors at work on basic research in the laboratory. Students may even participate in their own research projects. Biochemistry is very dependent on research activities that are basic and theoretical, but biochemistry has always had a very “practical” or “applied” side. Many of the early biochemical activities were encouraged and funded by companies involved in fermentation of beer and wine, baking, nutrition, and pharmaceuticals. We often refer to this practical side of biochemistry as **biotechnology,** which is defined as “the practical use of biological cells, biomolecules, and biological processes.” We see many examples of the daily use of applied biochemistry:

1. Using enzymes to catalyze reactions in the industrial production of specialty chemicals.
2. The use of gene-replacement therapy to treat individuals with diseases caused by dysfunctional genes.
3. Identification of biological specimens at crime scenes.
4. Using bacterial cells to produce large quantities of proteins for medical or industrial use.
5. Using bacteria to clean up chemical waste sites.

Today, one of the most active areas of biotechnology, and also the most controversial, is the preparation of **recombinant DNA** for use in molecular cloning. This activity has bioethical considerations because it has the power to change the genetic characteristics of fundamental life forms, including humans.

Molecular Cloning

Our knowledge of DNA structure and function has increased at a gradual pace over the past 50 years, but some discoveries have had a special impact on the progress and direction of DNA research. Although DNA was first discovered in cell nuclei in 1869, it was not confirmed as the carrier of genetic information until 1944. This major discovery was closely followed by the announcement of the double-helix structure of DNA in 1953. The more recent development of technology that allows for genetic manipulation by inserting “foreign” DNA fragments into an organism’s natural, replicating DNA may well have a greater impact on the direction of DNA research than the earlier discoveries. In the short time since the first construction and replication of plasmid recombinant DNA, several scientific and medical applications of the new technology have been announced. This new era of genetic engineering has captivated the general public and scientists alike. Some of the predicted achievements in recombinant DNA research have been slow in coming; however, future workers in biochemistry and related fields will continue to see advances in the use of recombinant DNA and molecular cloning (genetic engineering).

Recombinant DNA or **molecular cloning** consists of the covalent insertion of DNA fragments from one type of cell or organism into the replicating DNA of another type of cell. Many copies of the hybrid DNA may be produced by the progeny of the recipient cells; hence, the DNA molecule is **cloned.** If the inserted fragment contains a functional gene carrying the code for a specific protein, many copies of that gene and translated protein may be produced in the host cell. This process has become important for the large-scale production of proteins (insulin, somatostatin, bovine growth hormone, and other molecules) that are of value in medicine and basic science, but are difficult and expensive to obtain by other methods (see Chapter 11, Section B and Table 11.3, p. 317).

Our current knowledge of recombinant DNA is the result of several recent biotechnological advances. The first major breakthrough was the isolation of mutant strains of *E. coli* that are not able to degrade or restrict foreign DNA. These strains are now used as host organisms for the replication of recombinant DNA. The second advance was the development of bacterial extrachromosomal DNA **(plasmids)** and bacteriophage DNA as cloning vehicles to carry the DNA to be cloned into host cells. The final, necessary advance was the development of methods for inserting the foreign DNA into the natural vehicle. Manipulation of DNA was greatly aided by the discovery of **restriction endonucleases,** enzymes that catalyze the hydrolysis of phosphodiester bonds at selected sites in DNA (for details, see this chapter, Section B). The enzymes recognize specific base sequences (usually 4–8 bases) and catalyze hydrolytic cleavage of both DNA strands in or near the base sequence region (Reaction 10.1):

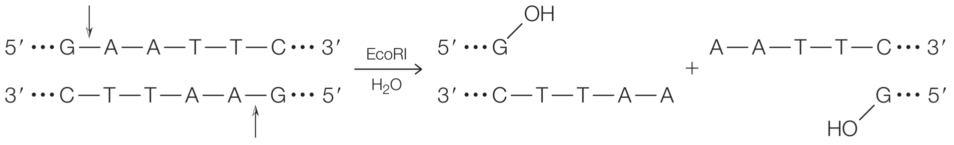

(Reaction 10.1)

The cleavage leads to one of two types of ends in the DNA (Figure 10.1):

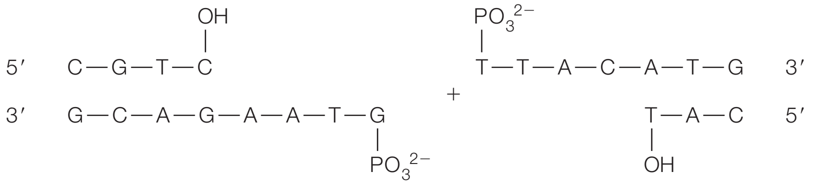

A Cohesive ends

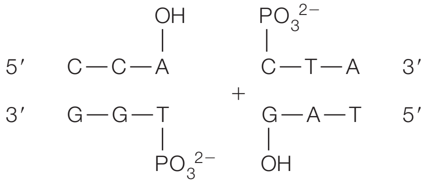

B Blunt ends

FIGURE 10.1 Results of the cleavage of DNA by a restriction endonuclease. Two types of ends after DNA cleavage: **A** Cohesive ends may remain weakly associated by hydrogen bonds between complementary bases. **B** Blunt ends do not overlap.

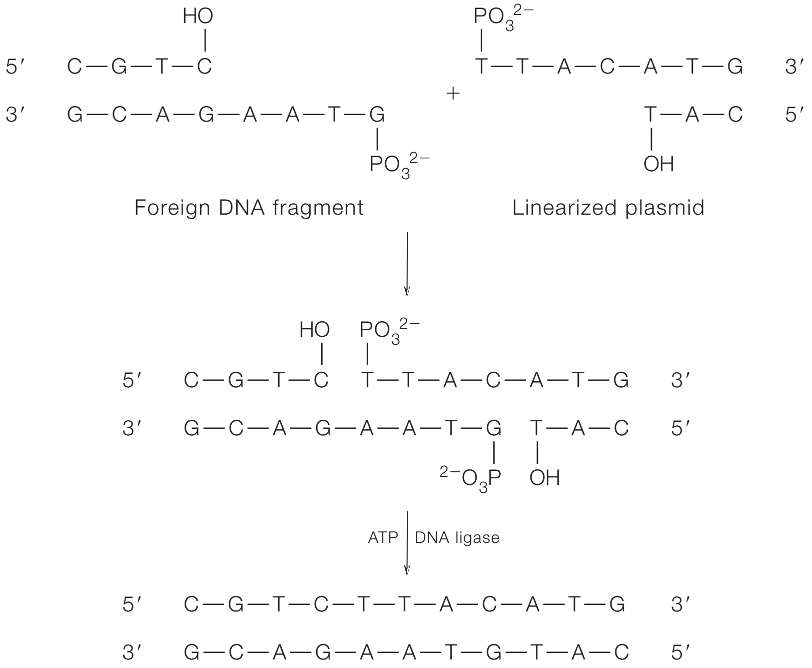

FIGURE 10.2 Insertion of a DNA fragment into a linearized plasmid using DNA ligase. DNA ligase catalyzes the formation of phosphodiester bonds.

1. ***Cohesive (or sticky) ends,*** where a few bases can remain weakly associated by hydrogen bonding, or
2. ***Blunt ends*** that do not overlap.

Restriction enzymes provide a gentle and specific method for opening (linearizing) circular vectors at predetermined sites, and for preparing fragments of DNA to be cloned.

Methods for covalent joining of the foreign DNA fragment to the vector ends and closure of the circular hybrid plasmid were then developed. The enzyme, DNA ligase, which can catalyze the ATP-dependent formation of phosphodiester linkages at the insertion sites, is used for final closure (Figure 10.2).

Steps for Preparing Recombinant DNA

The basic steps involved in performing a recombinant DNA experiment are listed below and are outlined in Figure 10.3:

1. ***Select and prepare a DNA fragment (X) that is to be incorporated into a host cell, where it will be replicated and translated into a protein product.*** Often the DNA fragment is a specific gene that carries the information for synthesis of a particular protein or proteins. As shown in Figure 10.3, the foreign DNA to be cloned may be prepared by chemical synthesis (usually by the polymerase chain reaction and sequence-specific primers; see Section B of this chapter), action of restriction endonucleases, or transcription of mRNA catalyzed by reverse transcriptase.
2. ***Choose a vector or vehicle to carry X, the DNA fragment, into the host cell.*** The vector may be plasmid DNA, DNA from a phage, or yeast artificial

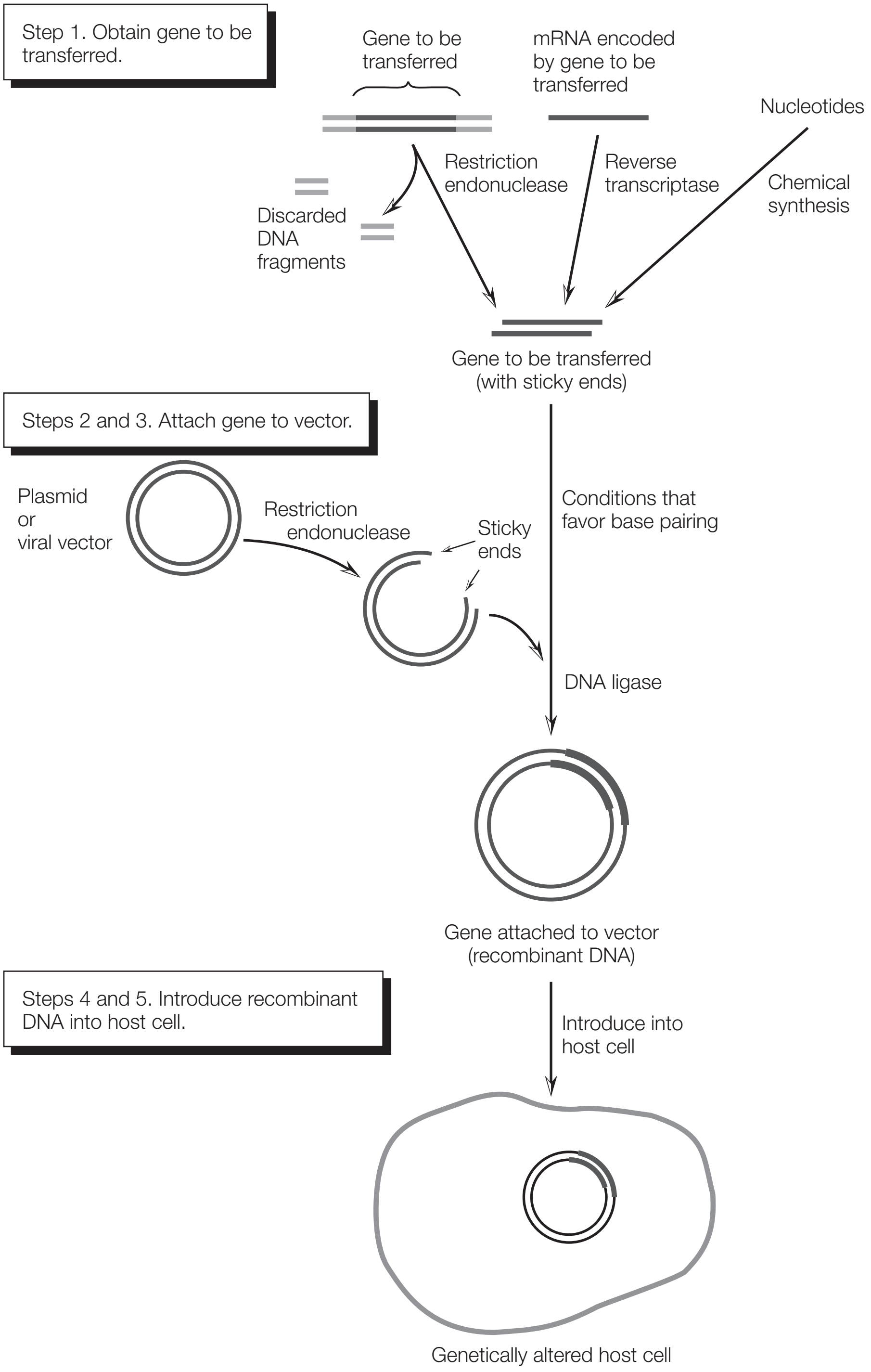
Step 1. Obtain gene to be transferred.
Gene to be transferred
mRNA encoded by gene to be transferred
Nucleotides
Restriction endonuclease
Reverse transcriptase
Chemical synthesis
Discarded DNA fragments
Gene to be transferred (with sticky ends)
Steps 2 and 3. Attach gene to vector.
Plasmid or viral vector
Restriction endonuclease
Sticky ends
Conditions that favor base pairing
DNA ligase
Gene attached to vector (recombinant DNA)
Steps 4 and 5. Introduce recombinant DNA into host cell.
Introduce into host cell
Genetically altered host cell

FIGURE 10.3
Outline of a typical experimental procedure for constructing recombinant DNA.

chromosomes. If a circular vector is used (i.e., a plasmid), it must be linearized (broken open) in order to accept X. Ideally, the same restriction enzyme should be used to prepare X and the vehicle so that there is the possibility of overlapping cohesive ends.

3. ***Insert the DNA fragment, X, into the vector by overlapping cohesive ends or by modifying blunt ends using homopolymer tails.*** The final covalent bonds to hold X into the vector are formed by the action of **DNA ligase.** This enzyme catalyzes the ATP-dependent formation of phosphodiester bonds (Figure 10.2). The final product represents a recombinant or hybrid DNA.
4. ***Introduce the hybrid DNA into a host organism (usually a bacterial cell), where it can be replicated.*** This process is called **transformation.** It has been demonstrated that incorporation of a hybrid plasmid or phage DNA into host *E. coli* cells is enhanced by the addition of calcium ions (Ca^{2+}).
5. ***Develop a method for identifying and screening for host cells that have accepted and are replicating the hybrid DNA.*** This is usually accomplished by screening for antibiotic resistance (Figure 10.4).

Cloning Vectors

Several types of cloning vectors are commonly used today. Some of the most effective ones include plasmids and bacteriophage DNA for prokaryotic cells, and yeast artificial chromosomes and baculovirus for eukaryotic cells.

Plasmids

Many bacterial cells contain self-replicating, extrachromosomal DNA molecules called **plasmids.** This form of DNA is closed circular, double-stranded, and much smaller than chromosomal DNA; their molecular weights range from 2×10^6 to 20×10^6, which corresponds to between 3000 and 30,000 base pairs. Bacterial plasmids normally contain genetic information for the translation of proteins that confer a specialized and sometimes protective characteristic (phenotype) on the organism. Examples of these characteristics are enzyme systems necessary for the production of antibiotics, enzymes that degrade antibiotics, and enzymes for production of toxins. Plasmids are replicated in the cell by one of two possible modes. **Stringent replicated** plasmids are present in only a few copies, and **relaxed replicated** plasmids are present in many copies, sometimes up to 200. In addition, some relaxed plasmids continue to be produced even after the antibiotic chloramphenicol is used to inhibit chromosomal DNA synthesis in the host cell. Under these conditions, many copies of the plasmid DNA may be produced (up to 2000 or 3000) and may accumulate to 30% to 40% of the total cellular DNA.

The ideal plasmid cloning vector has the following properties:

1. The plasmid should replicate in a relaxed fashion so that many copies are produced.
2. The plasmid should be small; then it is easier to separate from the larger chromosomal DNA, easier to handle without physical damage, and probably contains very few sites for attack by restriction endonucleases.
3. The plasmid should contain identifiable markers so that it is possible to screen progeny for the presence of the plasmid. At least two selective markers are

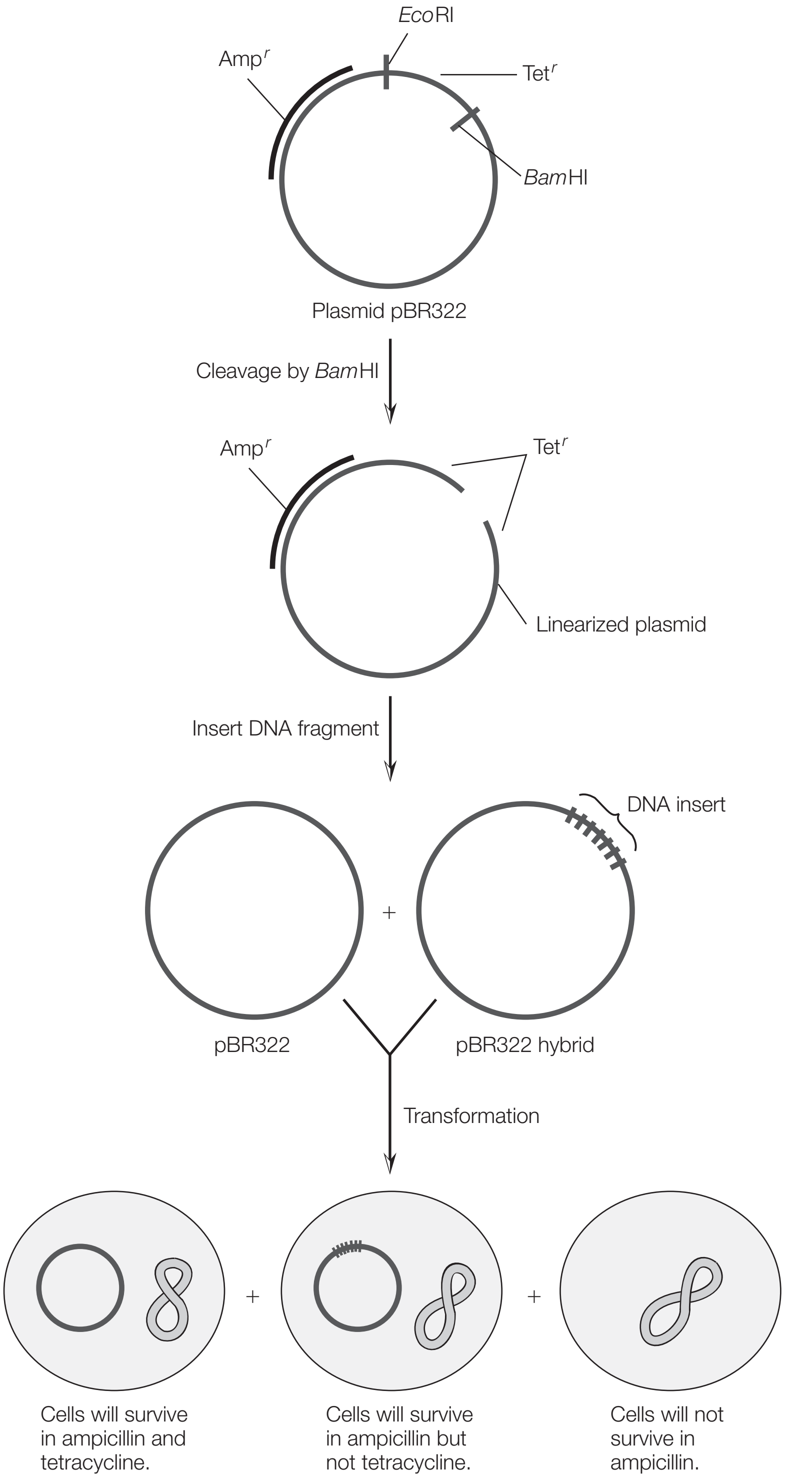
*Eco*RI
Amp^r
Tet^r
*Bam*HI
Plasmid pBR322
Cleavage by *Bam*HI
Amp^r
Tet^r
Linearized plasmid
Insert DNA fragment
DNA insert
+
pBR322
pBR322 hybrid
Transformation
+
+
Cells will survive in ampicillin and tetracycline.
Cells will survive in ampicillin but not tetracycline.
Cells will not survive in ampicillin.

FIGURE 10.4 Procedure for selecting bacterial cells that contain hybrid DNA.

desirable, a primary one to confirm the presence of the plasmid, and a secondary one to confirm the insertion of foreign DNA. Resistance to antibiotics is a convenient type of marker.

4. The plasmid should have only one cleavage site for a specific restriction endonuclease. This provides only two "ends" to which the foreign DNA can be attached. Ideally, the single restriction site should be within a gene, so that insertion of the foreign DNA will inactivate the gene (called insertional marker inactivation).

E. coli Plasmids

Among the most widely used *E. coli* plasmids are derivatives of the replicon **plasmid ColE1.** This plasmid carries a resistance gene against the antibiotic colicin E. The plasmid is under relaxed control, and up to 3000 copies may be produced when the proper *E. coli* strain is grown in the presence of chloramphenicol. One especially useful derivative plasmid of ColE1 is pBR322. It has all the properties previously outlined; in addition, its nucleotide sequence of 4363 base pairs is known, and it contains several different restriction endonuclease cleavage sites where foreign DNA can be inserted. For example, pBR322 has a single restriction site for the restriction enzyme *Eco*RI. *In vitro* insertion of a foreign DNA fragment into the *Eco*RI cleavage site and incorporation into a host cell (transformation) lead to immunity of the host cell to colicin E1, but the cell is unable to produce colicin.

Useful modifications of the pBR322 plasmid are the **pUC plasmids** (developed at The University of California). The pUC plasmids have a resistance gene for tetracycline, and sites for restriction endonuclease cleavage are concentrated in a region called the **multiple cloning site.**

Strains of *E. coli* are most often used as host cells because they are easy to grow and maintain. The rate of growth is exponential and can be monitored by measuring the absorbance of a culture sample at 600 nm. One absorbance unit corresponds to a cell density of approximately 8×10^8 cells/mL. Some *E. coli* strains that harbor the ColE1 plasmids are RR1, HB101, GM48, 294, SK1592, JC411Thy/ColE1, and CR34/ColE1. The typical procedure for growth and amplification of plasmids is, first, to establish the cells in normal medium for several hours. An aliquot of this culture is then used to inoculate medium containing the appropriate antibiotic. After overnight growth, a new portion of medium containing the antibiotic is inoculated with an aliquot of overnight culture. After the culture has been firmly established, a solution of chloramphenicol is added to inhibit chromosomal DNA synthesis. The ColE1 plasmids continue to replicate. The culture is then incubated for 12 to 18 hours and harvested by centrifugation.

Other Cloning Vectors

Another useful and widely used cloning vector is the DNA from **bacteriophage λ.** The DNA from λ phage is a double-stranded molecule with about 50,000 base pairs. It has many advantages as a cloning vector for prokaryotic proteins:

1. Many copies of recombinant phage DNA may be replicated in a host cell.
2. The recombinant phage DNA may be packaged as a viral particle for infecting the host bacteria.

3. Because λ phage DNA is larger than plasmid DNA, it is possible to insert larger fragments of DNA that include smaller fragments of eukaryotic DNA. The inserted DNA may be as large as 25,000 base pairs.

The insertion of larger DNA, including eukaryotic DNA that may contain over a million base pairs, is more efficiently done with **yeast artificial chromosomes (YACs)** or the **baculovirus expression vector system (BEVS).** YACs containing DNA inserts of 100,000 to 1,000,000 base pairs may be prepared for cloning. The baculoviruses are double-stranded DNA viruses that have as their natural hosts different insect species. They are not known to infect vertebrate hosts. The BEVS may be used to express complex proteins that require special processes like proteolytic cleavage, intron splicing, chemical modifications (phosphorylation or glycosylation of hydroxyl group), formation of disulfide bonds, and proper protein folding.

STUDY EXERCISE 10.1 Cloning Vectors

Use PubMed or a similar computer search to look for new cloning vectors that have been developed since publication of this book.

B. IMPORTANT ENZYMES IN MOLECULAR BIOLOGY AND BIOTECHNOLOGY

The Restriction Endonucleases

Bacterial cells produce many enzymes that act to degrade various forms of DNA. Of special interest are the **restriction endonucleases,** or restriction enzymes, that recognize specific base sequences in double-stranded DNA and catalyze hydrolytic cleavage of the two strands in or near that specific region. The biological function of these enzymes is to degrade or restrict foreign DNA molecules. Host DNA is protected from hydrolysis because some bases near the cleavage sites are methylated. The action of the restriction enzyme *Eco*RI is shown in Reaction 10.1.

The site of action of *Eco*RI is a specific hexanucleotide sequence. Two phosphodiester bonds are hydrolyzed (see arrows), resulting in fragmentation of both strands. Note the twofold rotational symmetry feature at the recognition site and the formation of cohesive ends. The weak base pairing between the cohesive ends is not sufficient to hold the two fragments together.

At the present time, more than 3500 restriction enzymes have been isolated and characterized. Nomenclature for the enzymes consists of a three-letter italicized abbreviation representing the source (*Eco* = *E. coli*), a letter representing the strain (R), and a Roman numeral designating the order of discovery. *Eco*RI is the first to be isolated from *E. coli* (strain R) and characterized. Table 10.1 lists several other restriction enzymes, their recognition sequence for cleavage, and optimum reaction conditions. A useful Web site for the restriction enzymes is REBASE: http://rebase.neb.com.

TABLE 10.1 Specificity and Optimal Conditions for Several Restriction Endonucleases

Name	Recognition Sequence 5′ … … … 3′	*T* (°C)	pH	Tris (m*M*)	NaCl (m*M*)	$MgCl_2$ (m*M*)	DTT[2] (m*M*)
*Aat*II	G — A — C — G — T ↓ — C	37	7.9	20	50	10	10
*Alu*I	A — G ↓ — C — T	37	7.5	10	50	10	10
*Bal*I	T — G — G ↓ — C — C — A	37	7.9	6	–	6	6
*Bam*HI	G ↓ — G — A — T — C — C	37	8.0	20	100	0.7	1
*Bcl*I	T ↓ — G — A — T — C — A	60	7.5	10	50	10	1
*Eco*RI	G ↓ — A — A — T — T — C	37	7.5	10	100	10	1
*Hae*II	Pu[1] — G — C — G — C ↓ — Py[1]	37	7.5	10	50	10	10
*Hind*III	A ↓ — A — G — C — T — T	37–55	7.5	10	60	10	1
*Hpa*I	G — T — T ↓ — A — A — C	37	7.5	10	50	10	1
*Mse*I	T ↓ — T — A — A	37	7.9	10	50	10	1
*Not*I	G — C ↓ — G — G — C — C — G — C	37	7.9	10	150	10	–
*Sal*I	G ↓ — T — C — G — A — C	37	8.0	10	150	10	1
*Sca*I	A — G — T ↓ — A — C — T	37	7.4	10	100	10	1
*Taq*I	T ↓ — C — G — A	65	8.4	10	100	10	10

[1] Pu = *a purine base;* Py = *a pyrimidine base.*

[2] DTT = *dithiothreitol.*

Applications of Restriction Enzymes

Restriction enzymes are used extensively in nucleic acid chemistry. They may be used to cleave large DNA molecules into smaller fragments that are more amenable to analysis. For example, λ phage DNA, a linear, double-stranded molecule of 48,502 base pairs (molecular weight 31×10^6), is cleaved into six fragments by *Eco*RI or into more than 50 fragments by *Hinf* I (*Haemophilus influenzae,* serotype f). The base sequence recognized by a restriction enzyme is likely to occur only a very few times in any particular DNA molecule; therefore, the smaller the DNA molecule, the fewer the number of specific sites. The λ phage DNA is cleaved into 0 to 50 or more fragments, depending on the restriction enzyme used, whereas larger bacterial or animal DNA will most likely have many recognition sites and be cleaved into hundreds of fragments. Smaller DNA molecules, therefore, have a much greater chance of producing a unique set of fragments with a particular restriction enzyme. It is unlikely that this set of fragments will be the same for any two different DNA molecules, so the fragmentation pattern is unique and can be considered a “fingerprint” of the DNA substrate. The fragments are readily separated and sized by agarose gel electrophoresis.

Restriction endonucleases are also valuable tools in the construction of hybrid DNA molecules. Several restriction enzymes act on bacterial plasmid vehicles that have only a single site of cleavage. This linearizes the circular plasmid and allows for the insertion of a foreign DNA fragment. For example, the popular plasmid vehicle pBR322 has a single restriction site for *Bam*HI (*Bacillus amyloliquefaciens,* H) that is within the tetracycline resistance gene. The enzyme not only opens the plasmid for insertion of a DNA fragment, but also destroys a phenotype; this fact aids in the selection of transformed bacteria.

Restriction enzymes can be used to produce physical maps of DNA molecules. Important information can be obtained from maps of DNA, whether the molecules are small and contain only a few genes, such as viral or plasmid DNA, or large and complex, such as bacterial or eukaryotic chromosomal DNA. An understanding of the genetics, metabolism, and regulation of an organism requires knowledge of the precise arrangement of its genetic material. The construction of a **restriction enzyme map** for a DNA molecule provides some of this information. Such a map displays the sites of cleavage by restriction endonucleases and the number of fragments obtained after digestion with each enzyme.

A restriction enzyme map is constructed by first digesting plasmid DNA with restriction endonucleases that yield only a few fragments. Each digest of DNA obtained with a single nuclease is analyzed by agarose gel electrophoresis, using standards for molecular weight determination. These are referred to as the primary digests. Second, each of the primary digests is treated with a series of additional restriction enzymes, and the digests are analyzed by agarose gel electrophoresis. The restriction map is then constructed by combining the various fragments by trial and error and logic, much as in sequencing a protein by proteolytic digestion by several enzymes and searching for overlap regions in the fragments. In the case of the restriction endonuclease digests, two characteristics of the DNA fragments are known: the approximate molecular weights (from electrophoresis) and the nature of the fragment ends (from the known selectivity of the individual restriction enzymes). The fragments may also be sequenced by the Sanger or Maxam-Gilbert method. A restriction enzyme map for the plasmid pBR322 is shown in Figure 10.5.

Practical Aspects of Restriction Enzyme Use

Restriction enzymes are heat-labile and expensive biochemical reagents. Their use requires considerable planning and care. Each restriction nuclease has been examined for optimal reaction conditions in regard to specific pH range, buffer composition, and incubation temperature. This information for each enzyme is readily available from the commercial supplier of the enzyme or from the literature. Table 10.1 gives important reaction information for several enzymes. The temperature range and pH optima for most restriction nucleases are similar (37°C, 7.5–8.0); however, optimal buffer composition is variable. Typical buffer components are Tris, NaCl, $MgCl_2$, and a sulfhydryl reagent (β-mercaptoethanol or dithiothreitol). Proper reaction conditions are crucial for optimal reaction rate, but, more importantly, changing reaction conditions have been shown to alter the specificity of some restriction enzymes.

Although restriction enzymes are very unstable reagents, they can be stored at −20°C in a buffer containing 50% glycerol. They are usually prepared in an appropriate buffer and shipped in packages containing dry ice.

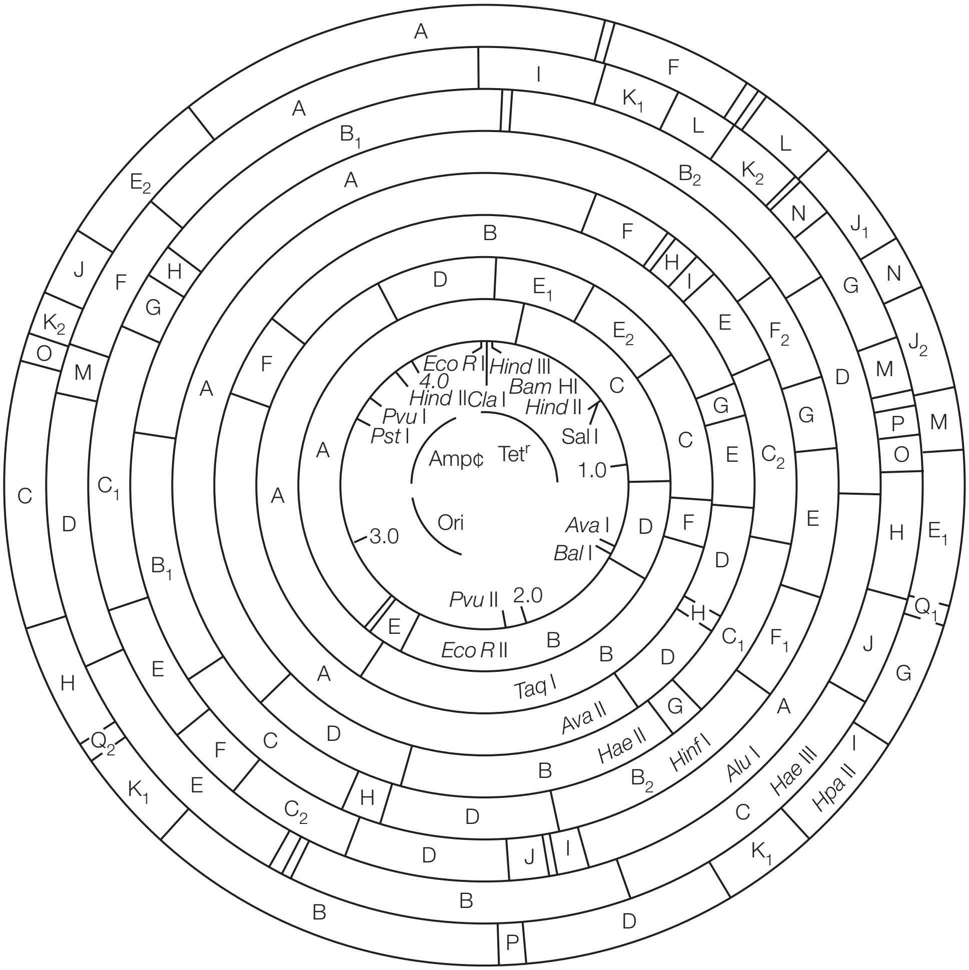

FIGURE 10.5 A restriction enzyme map for the plasmid pBR322. *From* Molecular Cloning: A Laboratory Manual, *by T. Maniatis, E. Fritsch, and J. Sambrook, Cold Spring Harbor Laboratory (Cold Spring Harbor, NY), 1982.*

Disposable gloves should be worn when you are handling the enzyme container. Remove the enzyme from the freezer just before you need it. Store the enzyme in an ice bucket at all times when it is outside the freezer. The enzyme should never be stored at room temperature. Because of high cost, digestion by restriction enzymes is carried out on a microscale level. A typical reaction mixture will contain about 1 μg or less of DNA and 1 unit of enzyme in the appropriate incubation buffer. One unit is the amount of enzyme that will degrade 1 μg of λ phage DNA in 1 hour at the optimal temperature and pH. The total reaction volume is usually between 20 and 50 μL. Incubation is most often carried out at the recommended temperature for about 1 hour. The reaction is stopped by adding EDTA solution, which complexes divalent metal ions essential for nuclease activity.

Reaction mixtures from restriction enzyme digestion may be analyzed directly by agarose gel electrophoresis. This technique combines high resolving power and sensitive detection to allow the analysis of minute amounts of DNA fragments.

STUDY EXERCISE 10.2 Restriction Enzymes

What is the class of the restriction enzyme, *Eco*RI? See Table 8.3 for classes, p. 251.

The Polymerase Chain Reaction

Producing multiple copies of a particular DNA fragment need not always require the tedious and time-consuming procedures of molecular cloning as described in Section A of this chapter. If at least part of the sequence of a DNA fragment is known, it is possible to make many copies using the **polymerase chain reaction (PCR).** The PCR method was conceived by Kary Mullis (1993 recipient of the Nobel Prize in chemistry) during a moonlit drive through the mountains of northern California.

Fundamentals of the PCR

The fundamentals of this relatively simple process are outlined here using a fragment of double-stranded DNA. The DNA template is divided into five regions, designated I to V in Figure 10.6. The complementary regions are I′ to V′. The region targeted for amplification is III (and III′), which can be reproduced by PCR if the nucleotide sequences of the flanking regions, II′ and IV′, are known. Experimental requirements for the PCR include:

1. Two synthetic oligonucleotide primers of about 20 base pairs each, which are complementary to the flanking sequences II and IV and also have similar binding ability.
2. A heat-stable DNA polymerase.
3. The four deoxyribonucleoside triphosphates, dATP, dGTP, dCTP, and dTTP.

The PCR is performed in cycles of three steps. Each cycle consists of:

1. Denaturation to achieve template DNA strand separation is done by heating a mixture of all components at 95°C for about 15 s.
2. Abrupt cooling of the mixture to the range of 37–55°C allows the primers to hybridize with the appropriate flanking regions. The primers are oriented on the template so their 3′ ends are directed toward each other. Synthesis of DNA extends across regions III and III′. Note the dual roles played by the complementary oligonucleotides: they locate the starting points for duplication of the desired DNA segment, and they serve as 3′-hydroxy primers to initiate DNA synthesis. A large excess of primers to DNA is added to the reaction mixture to favor hybridization and prevent reannealing of the DNA template strands.
3. Synthesis of the targeted DNA is catalyzed by *Taq* DNA polymerase. The temperature is raised to 72°C to enhance the rate of the polymerization reaction. The enzyme extends both primers, producing two new strands of DNA, II-III-IV-V and I′-II′-III′-IV′. The chosen polymerase is from a thermophilic bacterium, *Thermus aquaticus*, an organism originally discovered in a Yellowstone National Park hot spring. Other useful polymerases have

FIGURE 10.6 The polymerase chain reaction has three steps: **A** Strand separation by heating at 95°C. **B** Hybridization of the primers. **C** Extension of the primers by DNA synthesis. Segments are labeled I, II, III, IV, and V on the original DNA strand and I′, II′, III′, IV′, and V′ on the complementary strand. Primer II has diagonal lines and primer IV′ has dots. Newly synthesized DNA is shown in black.

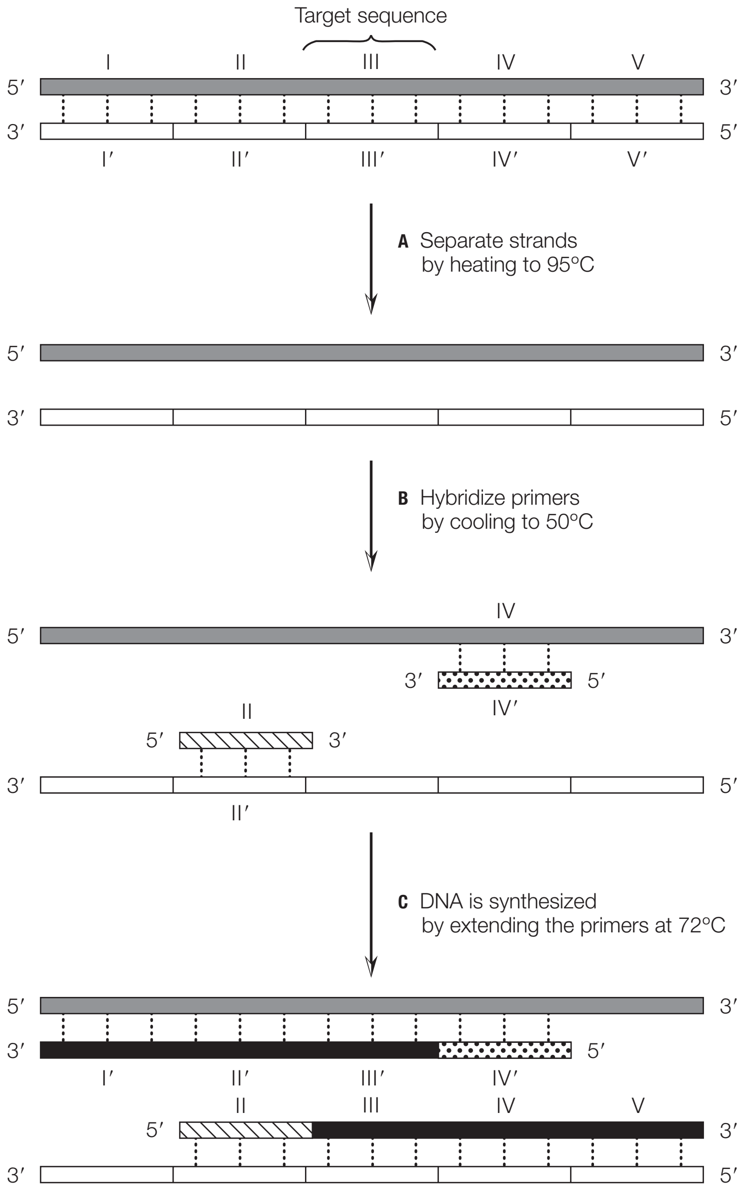
Target sequence
I
II
III
IV
V
5′
3′
3′
5′
I′
II′
III′
IV′
V′
A Separate strands by heating to 95°C
5′
3′
3′
5′
B Hybridize primers by cooling to 50°C
IV
5′
3′
3′
5′
IV′
II
5′
3′
3′
5′
II′
C DNA is synthesized by extending the primers at 72°C
5′
3′
3′
5′
I′
II′
III′
IV′
II
III
IV
V
5′
3′
3′
5′

been isolated from bacteria found in geothermal vents on the ocean floor. These enzymes are heat stable, so the reaction can be carried out at a high temperature, leading to a high rate of DNA synthesis. The DNA synthesis reaction is usually complete in about 30 s.

The usefulness of the PCR lies in its three steps—denaturation, hybridization, and DNA synthesis—that can be repeated many times simply by changing the temperature of the reaction mixture. Each newly synthesized strand of DNA can serve as a template, so the target DNA concentration increases at an exponential rate. In a process consisting of 20 cycles, the amplification for the DNA fragment is about a millionfold. Thirty cycles, which can be completed in 1–3 hours, provide a billionfold amplification. Theoretically, one could begin with a single molecule of target DNA and produce in 1 hour enough DNA for the Sanger dideoxy chain-terminating sequence procedure. Another benefit of the PCR is its simplicity. Commercially available instruments called **thermocyclers** allow the laboratory technician to mix the reagents and insert the reaction mixture in the cycler, which then automatically repeats the reaction steps by changing temperatures.

The PCR has become a routine tool in basic research carried out in universities, hospitals, and pharmaceutical companies. It is especially useful for amplifying small amounts of DNA for sequencing. It was used widely in the Human Genome Project, which in 2001 announced the sequence of the 3 billion base pairs in the human genome.

Applications of the PCR

Diagnostic Medicine

PCR may be used to detect the presence of infectious bacteria and viruses in an individual even before symptoms begin to appear. HIV infection is usually confirmed using tests that detect serum antibody proteins made against the viral proteins. However, there may be a period of at least six months between infection and when the anti-HIV antibodies reach a detectable concentration. PCR can be used to detect proviral DNA. An especially important application is the identification of HIV-infected infants. The antibody-based test cannot be done on infants because they have maternal antibodies for up to 15 months after birth. Other medical applications of PCR include early detection of tuberculosis and cancers, especially leukemias, and analysis of small samples of amniotic fluid to detect genetic abnormalities in fetuses.

Forensics

Recovery of fingerprints from a crime scene has been a long-standing and well-accepted method in forensics. A new procedure, **DNA fingerprinting,** is being developed and achieving widespread use. DNA fingerprinting is the biochemical analysis of DNA in biological samples (semen, blood, saliva) remaining at a crime scene. Since every individual possesses a unique hereditary composition, each individual has a characteristic phenotype that is reflected in his or her DNA sequence. DNA fingerprinting attempts to show the genetic variations that are found from person to person. These differences can be detected by analysis of an individual's DNA. Two experimental methods are currently used in DNA fingerprinting:

1. ***Restriction Fragment Length Polymorphisms (RFLPs)*** In this method, DNA samples are digested with restriction enzymes. The resulting fragments are separated by gel electrophoresis, blotted onto a membrane, hybridized to a DNA probe, and analyzed by autoradiography (see Section C on blotting in this chapter). Results vary from individual to individual (except in the case of identical twins, who have identical DNA) because of their differences in DNA sequence. Since each person's DNA has a unique sequence pattern, the restriction enzymes cut differently and lead to different-sized fragments.
2. ***PCR-Based Analysis*** The second method of DNA fingerprinting involves the use of PCR-amplified DNA. Allele-specific oligonucleotide (ASO)

primers are used for amplification of specific sequences. PCR-based DNA fingerprinting has many advantages over RFLP methods, including speed, simplicity, no requirements for radioactive probes, and greater sensitivity. DNA from a single hair or minute samples of blood, saliva, and semen can be analyzed.

C. NUCLEIC ACID BLOTTING

Many of the procedures for isolating and characterizing DNA and RNA molecules discussed in Chapters 9 and 10 depend on the separation of nucleic acids by an electrophoresis step, followed by some kind of analysis of the gel bands. Although electrophoresis provides excellent separation of molecules, the method gives no indication as to the identity of the molecules at each band. As described in Chapter 6, Section C, p. 192, **membrane blotting techniques** are now available for functional analysis of the molecules at each electrophoresis band.

The first blotting technique was reported by E. Southern in 1975. Using labeled complementary DNA probes, he searched for certain nucleotide sequences among DNA molecules blotted from the gel. This technique of detecting DNA-DNA hybridization is called **Southern blotting.** The general blotting technique has now been extended to the transfer and detection of specific RNA (for example, mRNA) using labeled complementary DNA probes **(Northern blotting)** and the transfer and detection of proteins that react with specific antibodies (**Western blotting**).

Blotting techniques have many applications, including mapping the genes responsible for inherited diseases by using restriction fragment length polymorphisms (RFLPs), screening collections of cloned DNA fragments (DNA libraries), and DNA fingerprinting for analysis of biological material remaining at the scene of a crime.

Study Problems

1. Explain the action of ampicillin as an inducer of plasmid replication.
2. How does chloramphenicol inhibit protein synthesis?
3. Why is polyacrylamide gel electrophoresis not suitable for analysis of most plasmid DNA?
4. Why is a wavelength of 600 nm used to measure growth of bacteria? Could other wavelengths be used?
5. How does the addition of an EDTA solution stop a restriction enzyme reaction?
6. What kind of chemical bonding holds cohesive ends together after action of a restriction enzyme?
7. Which of the following base sequences are probably not recognition sites for cleavage by restriction endonucleases? Why not?
   (a) 5′ GAATTC 3′
   (b) 5′ CATTAG 3′
   (c) 5′ CATATG 3′
   (d) 5′ CAATTG 3′

8. Use the REBASE Web site (http://rebase.neb.com) to search for information about the restriction enzymes listed below. For each enzyme, find the recognition sequence and the cleavage site, and determine if the DNA cut ends are blunt or cohesive.
   (a) *Apa*I
   (b) *Spm*I
   (c) *Alu*I
9. How many DNA fragments result from the action of restriction enzyme *Hae*II on the plasmid pBR322?
10. Assume the reaction digest from Problem 8 is analyzed by agarose gel electrophoresis. Draw an electrophoresis gel and show the approximate locations of each fragment.

Further Reading

W. Backman, ***Trends Biochem. Sci.*** **26,** 268–270 (2001). "The Advent of Genetic Engineering."

W. Bains, ***Biotechnology from A to Z,*** 3rd ed. (2003), Oxford University Press (New York).

J. Berg, J. Tymoczko, and L. Stryer, ***Biochemistry,*** 6th ed. (2007), Freeman (New York).

R. Boyer, ***Concepts in Biochemistry,*** 3rd ed. (2006), John Wiley & Sons (Hoboken, NJ). Chapter 13 covers biotechnology.

J. Brooks, ***Methods in Enzymol.*** **152,** 113–129 (1987). "Properties and Uses of Restriction Endonucleases."

T. Brown, ***Gene Cloning and DNA Analysis,*** 4th ed. (2001), Blackwell Science (Malden, MA).

F. Collins et al., ***Nature,*** **409** 708–745 (2001). The human genome.

C. Dieffenbach and G. Dveksier (Editors), ***PCR Primer: A Laboratory Manual,*** 2nd ed. (2003), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY).

R. Garrett and C. Grisham, ***Biochemistry,*** 4th ed. (2010), Brooks/Cole (Belmont, CA).

B. Glick et al., ***Molecular Biotechnology: Principles and Applications of Recombinant DNA,*** 4th ed. (2009), ASM Press (Washington, DC).

C. Hardin, et al. (Editors), ***Cloning, Gene Expression and Protein Purification: Experimental Procedures and Process Rationale*** (2001), Oxford University Press (New York).

E. Hilario and J. MacKay (Editors), ***Protocols for Nucleic Acid Analysis by Nonradioactive Probes,*** 2nd ed. (2007), Humana Press (New York).

S. Metzenberg, ***Working with DNA*** (2007), Taylor and Francis (New York).

K. Mullis, ***Sci. Am.*** **263(4),** 56–65 (1990). "The Unusual Origin of the Polymerase Chain Reaction."

D. Nelson and M. Cox, ***Lehninger Principles of Biochemistry,*** 5th ed. (2009), Freeman (New York).

D. Nicholl, ***An Introduction to Genetic Engineering,*** 2nd ed. (2003), Cambridge University Press (Cambridge, UK).

W. Ream and K. Field, ***Molecular Biology Techniques: An Intensive Laboratory Course*** (1998), Academic Press (San Diego).

J. Sambrook and D. Russell, ***Molecular Cloning: A Laboratory Manual,*** 3rd ed. (2001), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY).

C. Tsai, ***Computational Biochemistry*** (2002), Wiley-Liss (New York). Chapters 9 and 10 cover nucleotide sequencing and recombinant DNA.

J. Venter, ***Science*** **291,** 1304 (2001). "The Sequence of the Human Genome."

D. Voet and J. Voet, ***Biochemistry,*** 3rd ed. (2005), John Wiley & Sons (Hoboken, NJ).

J. Watson et al., ***Molecular Biology of the Gene,*** 5th ed. (2004), Benjamin/Cummings (San Francisco).

J. Watson et al. ***Recombinant DNA: Genes and Genomes—A Short Course,*** 3rd ed. (2006), W. H. Freeman (New York).

L. Yount (Editor), ***Cloning*** (2000), Greenhaven Press (San Diego).

B. Zimm, ***Trends Biochem. Sci.*** **24,** 121–123 (1999). "One Chromosome: One DNA Molecule."

http://www.fermentas.com
Click on "Products," then "Conventional Restriction Endonucleases" for a listing of enzymes.

http://rebase.neb.com
The Restriction Enzyme Database.

http://www.tutorvista.com/content/biology/biology-iii/biotechnology/biotechnology-tools.php
Enzymes of recombinant DNA.

http://www.rpi.edu/dept/chem-eng/Biotech-Environ/Projects00/rdna/rdna.html
Recombinant DNA.

http://biog-101-104.bio.cornell.edu/BIOG101_104/tutorials/recomb_DNA.html
Recombinant DNA.

http://www.biochem.arizona.edu/classes/bioc471/pages/materials.html
An excellent resource on recombinant DNA.

http://en.wikipedia.org/wiki/Restriction_enzyme
Restriction enzymes.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.)

CHAPTER 11

PROTEIN PRODUCTION, PURIFICATION, AND CHARACTERIZATION

A. Procedures for the Purification of Proteins 308
B. Production of Proteins by Expression of Foreign Genes 317
C. Protein Characterization 322
D. Determination of Primary Structure 323

Study Problems 328
Further Reading 328

Proteins are often considered the workhorses among biomolecules, as they are responsible for the general maintenance and daily functioning of cells and organisms. The proteins are extremely versatile functional molecules, as they serve roles in biological catalysis (enzymes), immune response (antibodies), structural integrity, muscle contraction, cell regulation, and storage/transport processes. Because of their roles as essential biomolecules, there has always been intense interest in the extraction of proteins from their natural sources and in their characterization in terms of structure and biological function. In this chapter, we will focus on the purification procedures that are used to obtain native protein molecules that may be further characterized. The new discipline of **proteomics**, the study of the thousands of proteins expressed in cells and organisms and how the proteins interact with each other, has renewed interest in the development of methods for their isolation and characterization. Because the normal concentration of proteins in cells is quite low, there is a strong interest in designing methods to enhance their production. Recombinant DNA procedures now play a major role in the production of relatively large quantities of scarce proteins for basic scientific study and for medical and industrial applications. In the medical technique of **gene therapy**, a hybrid DNA containing the genes for the desired protein is transferred into a host cell, which acts as a factory to produce the desired protein.

A. PROCEDURES FOR THE PURIFICATION OF PROTEINS

Protein purification is an activity that has occupied the time of biochemists and molecular life scientists throughout the last two centuries. In fact, a large percentage of the biochemical literature is a description of how specific proteins have been separated from thousands of other proteins and other biomolecules in tissues, cells, and biological fluids. Biochemical investigations of all biological processes require, at some time, the isolation, purification, and characterization of a protein. In contrast to the repetitive procedures for isolation of DNA and RNA (see Chapter 9, Section B, p. 275), there is no single technique or sequence of techniques that can be followed to isolate and purify all proteins. It is sometimes still necessary to proceed by trial and error. Fortunately, the experiences and discoveries of thousands of scientists have been combined so that, today, a general, systematic approach is available for protein isolation and purification.

Composition of Proteins

All proteins found in nature are constructed by amide linkages between α-amino acids. The amino acids are selected from a group of 20 molecules that have common structural characteristics. They each have at least one carboxyl group and at least one amino group (Figure 11.1). The distinctive physical, chemical, and biological properties associated with an amino acid are the result of the R group, a side chain that is unique for each amino acid. A list of the 20 common amino acids and their one-letter and three-letter abbreviations are given in Table 11.1. If you are not yet familiar with the general structures and properties of amino acids and proteins, refer to your general biochemistry text.

STUDY EXERCISE 11.1 Peptide Structure

Draw the structure of this peptide: Asp-Phe-Ala-Lys-Trp.

Amount of Protein Versus Purity of Protein Versus Expense

The procedure selected for protein purification is dependent primarily on how much of the protein is needed and how pure the protein must be; in other words, what is the purpose for isolating the protein? In terms of amount, one usually thinks on two scales: **preparative,** isolation of larger quantities (several milligrams or grams) to use for further characterization or structure and function studies; or **analytical,** isolation of microgram or milligram quantities in order to make a few precise measurements such as molecular weight, sequencing, or structure determination by NMR or X-ray diffraction.

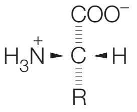

FIGURE 11.1 The general structure of the zwitterionic form of an L-amino acid. R represents the side chain.

TABLE 11.1 Abbreviations of the 20 Common Amino Acids Found in Proteins

	Abbreviation	
Name	**One-letter**	**Three-letter**
Glycine	G	Gly
Alanine	A	Ala
Valine	V	Val
Leucine	L	Leu
Isoleucine	I	Ile
Methionine	M	Met
Phenylalanine	F	Phe
Proline	P	Pro
Serine	S	Ser
Threonine	T	Thr
Cysteine	C	Cys
Asparagine	N	Asn
Glutamine	Q	Gln
Tyrosine	Y	Tyr
Tryptophan	W	Trp
Aspartate	D	Asp
Glutamate	E	Glu
Histidine	H	His
Lysine	K	Lys
Arginine	R	Arg

In terms of purity, a protein sample is considered pure when it contains a single type of protein. It is not practically possible to obtain a protein in 100% purity, but most studies can be done on samples that are 90–95% pure. It often is not worth the extra time and reagent expense to purify a protein greater than 90% unless absolutely necessary. In general, the best purification scheme is one that yields the maximum (or appropriate) amount of the protein, of desired purity, with a minimum amount of time and expense. The following discussion outlines the basic steps available to develop a protein purification scheme (Table 11.2).

Basic Steps in Protein Purification

Development of Protein Assay

First and foremost in any protein purification scheme is the development of an assay for the protein. This procedure, which may have a physical, chemical, or biological basis, is necessary in order to determine quantitatively and/or

TABLE 11.2 Typical Sequence for the Purification of a Protein

1. Develop an assay for the desired protein.
2. Select the biological source of the protein, including use of recombinant DNA techniques.
3. Release the protein from the source and solubilize it in an aqueous buffer system.
4. Fractionate the cell components by physical methods (centrifugation).
5. Separate the biomolecules present by differential solubility.
6. Apply chromatographic procedures.
    - (a) Gel-filtration chromatography
    - (b) Ion-exchange chromatography
    - (c) Affinity chromatography
7. Isoelectric focusing, if desired.
8. Determine purity by electrophoresis or HPLC.

qualitatively the presence of the specific protein. During the early stages of the purification, the particular protein desired must be distinguished from thousands of other proteins present in the crude cell homogenate. The desired protein may be less than 0.1% of the total protein. It should be obvious that the general protein assays (spectrophotometric, Bradford, Lowry, etc.; see Chapter 3, Section B, p. 67) are not useful as assays for a specific protein, because they simply indicate the total protein present. If the desired protein is an enzyme, the obvious assay should be based on biological function, that is, a measurement of the enzymatic activity after each isolation–purification step. If the protein to be isolated is not an enzyme or if the biological activity is unknown, physical or chemical methods must be used. One of the most useful analytical methods is electrophoresis. If an antibody has been prepared against the desired protein, it may be analyzed by electrophoresis and monitored during purification using Western blotting (see Chapter 6, Section C, p. 192).

Source of the Protein

We next consider the selection of the source from which the protein will be isolated. If the objective is simply to obtain a large quantity of a protein for further study, you would choose a source that contains large amounts of the protein. A possible choice is an organ from a large animal that can be obtained from a local slaughterhouse. Microorganisms are also good sources because they can be harvested in large quantities. If, however, you desire a specific protein from a specific type of organism, tissue, cell, cell organelle, or biological fluid, the source is limited. For example, if one desires to study the enzymes involved in glycolysis in the alfalfa plant, one would begin with leaves of the plant.

If the desired protein is known to be located in an organelle or subcompartment of the cell, partial purification of the protein is achieved by isolating the organelle by fractional centrifugation (Chapter 4, Section C, Figure 4.11, p. 107). If the protein is, instead, in the soluble cytoplasm of the cell, it will remain dissolved in the final supernatant obtained after centrifugation at 100,000 × g.

With the introduction of techniques in recombinant DNA research, it is now possible to synthesize relatively large quantities of specific proteins in selected host cells. One begins by preparing the recombinant DNA, which consists of inserting the specific gene that contains the code for the desired protein, along with appropriate promoters and markers, into a plasmid or other vector. The recombinant DNA may then be transferred to a host cell (*E. coli,* or other bacteria, or yeast), and the host cell is then allowed to grow and synthesize the desired protein (molecular cloning, Chapter 10, Section A, p. 290). This procedure has several advantages over the traditional method of isolating and purifying proteins as described in the previous paragraphs and in Table 11.2. Some of the benefits of using recombinant DNA include:

- Overexpression of the protein is possible, which produces relatively large amounts; this is especially helpful when dealing with scarce proteins, or with proteins for commercial or medical applications.
- Conditions can be controlled so there is less degradation of the desired protein by proteases and less contamination by other biomolecules. This makes it easier to isolate and purify the protein from the host cell. Some steps in Table 11.2 may be eliminated.
- With the use of genetic engineering techniques, including mutation, it is possible to produce modified proteins, i.e., proteins with one or more amino acid changes.

Because the techniques for purifying a protein using recombinant DNA procedures are somewhat different than those of traditional isolation and purification, a special section will be devoted to protein production by cloning (Section B of this chapter, p. 317).

Preparation of the Crude Extract

Once the protein source has been selected, the next step is to release the desired protein from its natural cellular environment and solubilize it in aqueous solution. This calls for disruption of the cell membrane without damage to the cell contents. Proteins are relatively fragile molecules, and only gentle procedures are allowed at this stage. The gentlest methods for cell breakage are osmotic lysis, gentle grinding in hand-operated glass homogenizers, and disruption by ultrasonic waves. These methods are useful for “soft tissue,” as found in green plants and animals. When dealing with bacterial cells, in which rigid cell walls are present, the most effective methods are grinding in a mortar with an inert abrasive such as sand or alumina, or treatment with lysozyme and/or a detergent. Lysozyme is an enzyme that catalyzes the hydrolysis of polysaccharide moieties present in cell walls. When very resistant cell walls are encountered (yeast, for example), the French press must be used. Here the cells are disrupted by passage, under high pressure, through a small hole.

Osmotic lysis consists of suspending cells in a solution of relatively high ionic strength. This causes water inside the cell to diffuse out through the membrane. The cells are then isolated by centrifugation and transferred to pure water. Water rapidly diffuses into the cell, bursting the membrane.

Another alternative, gentle **grinding,** is best accomplished with a glass or Teflon homogenizer. This consists of a glass tube with a close-fitting piston. Several varieties are shown in Figure 11.2A, B. The cells are forced against the glass walls under the pressure of the piston, and the cell components are released into an aqueous solution.

Ultrasonic waves, produced by a sonicator, are transmitted into a suspension of cells by a metal probe (Figure 11.2C). The vibration set up by the ultrasonic waves disrupts the cell membrane, releasing the cell components into the surrounding aqueous solution.

A less gentle, but widely used, solubilization device is the common electric blender. This method may be used for plant or animal tissue, but it is not effective for disruption of bacterial cell walls. A tool that is more scientifically designed is the **rotor stator homogenizer** (Figure 11.2D). The stator is a hollow tube and the rotor, attached to the stator, is a rapidly turning knife blade. Cells are torn apart by the turbulence and shear generated by the rotor.

The most recently developed homogenizer is the so-called **cell bomb,** which makes use of high pressure and decompression to disrupt cells. In this technique, a gas, usually nitrogen, helium, or air, is forced into cells under high pressure. When the pressure is released, expanding bubbles of the gas (nebulization) rupture the cell membrane. Cell bombs are available with pressure ranges of 250 to 25,000 psi. Cell bombs have an advantage over sonicators in that they cause no temperature increase that can denature proteins. A disadvantage of cell bombs is the potential for explosions because of the high-pressure conditions.

Since so many cell disruption methods are available, the experimenter must, by trial and error, find a convenient method that yields the maximum quantity of the protein with minimal damage to the molecules.

Stabilization of Proteins in a Crude Extract

Continued stabilization of the protein must always be considered in protein purification processes. While the protein is inside the cell, it is in a highly regulated environment. Cell components in these surroundings are protected against sudden changes in pH, temperature, or ionic strength and against oxidation and enzymatic degradation. Once the cell wall/membrane barrier is destroyed, the protective processes are no longer functional and degradation of the desired protein is likely to begin. An artificial environment that mimics the natural one must be maintained so that the protein retains its chemical integrity and biological function throughout the purification procedure. What factors are important in maintaining an environment in which proteins are stable? Although numerous factors must be considered, the following are the most critical:

1. ***Ionic strength and polarity*** The standard cellular environment is, of course, aqueous; because of the presence of inorganic salts, though, its ionic strength is relatively high. Addition of KCl, NaCl, or $MgCl_2$ to the cell extract may be necessary to maintain this condition. Proteins that are normally found in the hydrophobic regions of cells (i.e., membranes) are generally more stable in aqueous environments in which the polarity has been reduced by the addition of 1 to 10% glycerol or sucrose.

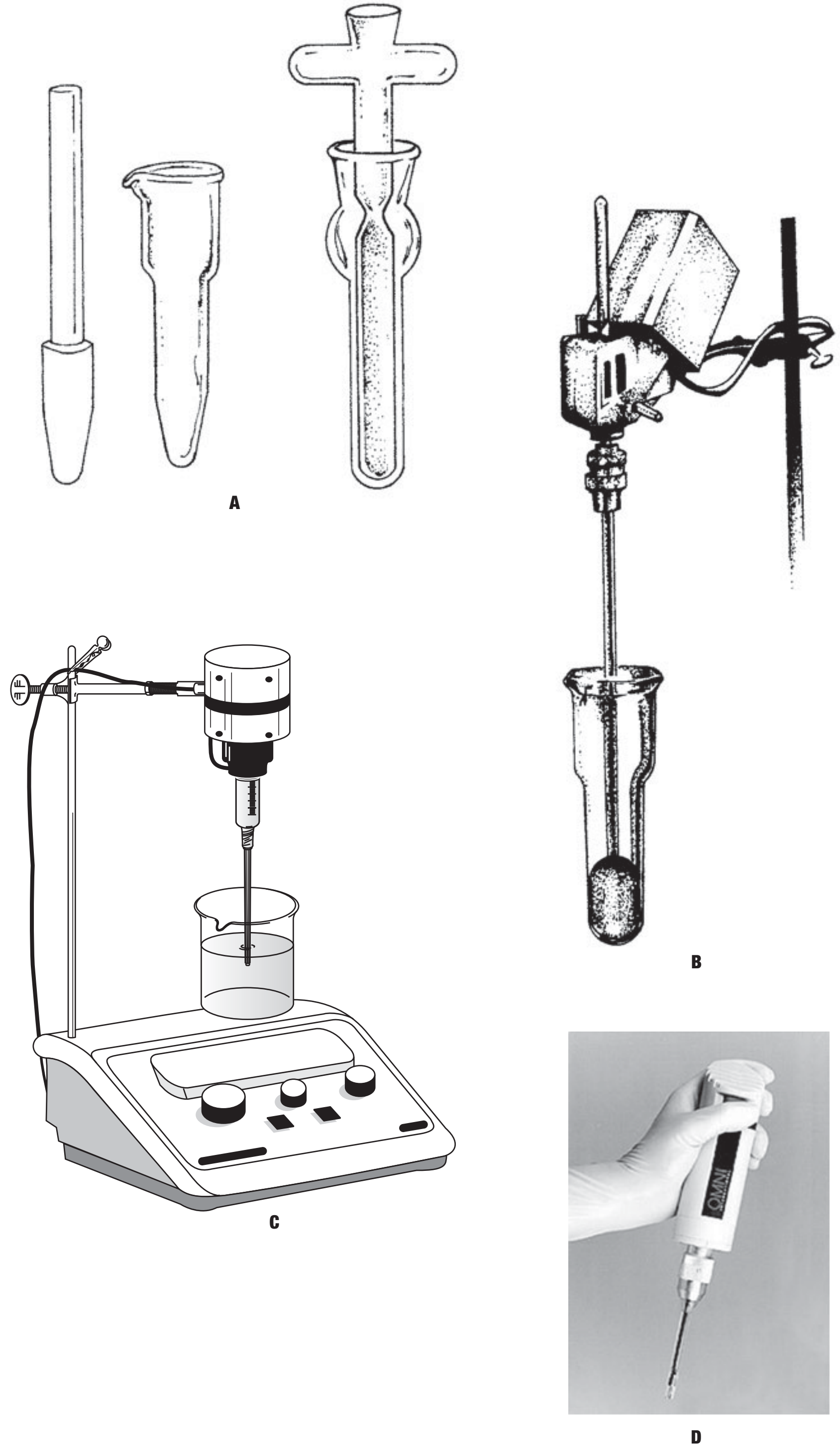

FIGURE 11.2 Tools for the preparation of a crude cell extract. **A** Hand-operated homogenizers, *courtesy of Ace Glass, Inc. Vineland, NJ.* **B** Homogenizer with electric motor, *courtesy of Sargent-Welch/VWR; www.sargentwelch.com.* **C** Ultrasonic homogenizer, *courtesy of BioLogics, Inc., www.biologics-inc.com.* **D** Rotor-stator, *courtesy of Omni International, Inc. www.omni-inc.com.*

2. ***pH*** The pH of a biological cell is controlled by the presence of natural buffers. Since protein structure is often irreversibly altered by extremes in pH, a buffer system must be maintained for protein stabilization. The importance of proper selection of a buffer system cannot be overemphasized. The criteria that must be considered in selecting a buffer are discussed in Chapter 3, Section A, p. 57. For most cell homogenates at physiological pH values, Tris and phosphate buffers are widely used.
3. ***Metal ions*** The presence of metal ions in mixtures of biomolecules can be both beneficial and harmful. Metal ions such as $Na^+, K^+, Ca^{2+}, Mg^{2+}$, and Fe^{3+} may actually increase the stability of dissolved proteins. Many enzymes require specific metal ions for activity. In contrast, heavy metal ions such as $Ag^{1+}, Cu^{2+}, Pb^{2+}$, and Hg^{2+} are deleterious, particularly to proteins that depend on sulfhydryl groups for structural and functional integrity. The main sources of contaminating metals are buffer salts, water used to make buffer solutions, and metal containers and equipment. To avoid heavy-metal contamination, you should use high-purity buffers, glass-distilled water, and glassware specifically cleaned to remove extraneous metal ions (see Chapter 1, Section C, p. 15). If metal contamination still persists, a chelating agent such as ethylenediaminetetraacetic acid (EDTA, $1 \times 10^{-4}\,M$) may be added to the buffer.
4. ***Oxidation*** Many proteins are susceptible to oxidation. This is especially a problem with proteins having free sulfhydryl groups, which are easily oxidized and converted to disulfide bonds. A reducing environment can be maintained by adding mercaptoethanol, cysteine, or dithiothreitol ($1 \times 10^{-3}\,M$) to the buffer system.
5. ***Proteases*** Many biological cells contain degradative enzymes (proteases) that catalyze the hydrolysis of peptide linkages. In the intact cell, functional proteins are protected from these destructive enzymes because the enzymes are stored in cell organelles (lysosomes, etc.) and released only when needed. The proteases are freed upon cell disruption and immediately begin to catalyze the degradation of protein material. This detrimental action can be slowed by the addition of specific protease inhibitors such as phenylmethylsulfonyl fluoride or certain bioactive peptides. These inhibitors are to be used with extreme caution because they are potentially toxic.
6. ***Temperature*** Many of the above conditions that affect the stability of proteins in solution are dependent on chemical reactions. In particular, metal ions, oxidative processes, and proteases bring about chemical changes in proteins. It is a well-accepted tenet in chemistry that lower temperatures slow down chemical processes. We generally assume that proteins are more stable at low temperatures. Although there are a few exceptions to this, it is fairly common practice to carry out all procedures of protein isolation under reduced-temperature conditions (usually in ice bath, 0–4°C).

After cell disruption, gross fractionation of the properly stabilized, crude cell homogenate may be achieved by physical methods, specifically centrifugation. Figure 4.11 outlines the stepwise procedure commonly used to separate subcellular organelles such as nuclei, mitochondria, lysosomes, and microsomes.

Separation of Proteins Based on Solubility Differences

Proteins are soluble in aqueous solutions primarily because their charged and polar amino acid residues are solvated by water. Any agent that disrupts these protein-water interactions decreases protein solubility because the protein-protein interactions become more important. Protein-protein aggregates are no longer sufficiently solvated, and they precipitate from solution. Because each specific type of protein has a unique amino acid composition and sequence, the degree and importance of water solvation vary from protein to protein. Therefore, different proteins precipitate at different concentrations of precipitating agent. The agents and factors most often used for protein precipitation are (1) inorganic salts, (2) organic solvents, (3) polyethylene glycol (PEG), (4) pH, and (5) temperature.

The most commonly used inorganic salt, ammonium sulfate, is highly solvated in water and actually reduces the water available for interaction with protein. As ammonium sulfate is added to a protein solution, a concentration of salt is reached at which there is no longer sufficient water present to maintain a particular type of protein in solution. The protein precipitates or is “salted out” of solution. The concentration of ammonium sulfate at which the desired protein precipitates from solution cannot be calculated, but must be established by trial and error. In practice, ammonium sulfate precipitation is carried out in stepwise intervals. For example, a crude cell extract is treated by slow addition of dry, solid, high-purity ammonium sulfate in order to achieve a change in salt concentration from 0 to 25%, is gently stirred for up to 60 minutes, and is subjected to centrifugation at 20,000 × *g*. The precipitate that separates upon centrifugation and the supernatant are analyzed for the desired protein. If the protein is still predominantly present in the supernatant, the salt concentration is increased from 25 to 35%. This process of ammonium sulfate addition and centrifugation is continued until the desired protein is salted out.

Organic solvents also decrease protein solubility, but they are not as widely used as ammonium sulfate because they sometimes denature proteins. They are thought to function as precipitating agents in two ways: (1) by dehydrating proteins, much as ammonium sulfate does, and (2) by decreasing the dielectric constant of the solution. The organic solvents used (which, of course, must be miscible with water) include methanol, ethanol, and acetone.

A relatively new method of selective protein precipitation involves the use of nonionic polymers. The most widely used agent in this category is polyethylene glycol. This polymer is available in a variety of molecular weights ranging from 400 to 7500. The biochemical literature reports successful use of different weights, but lower-molecular-weight polymer has been shown to be the most specific. The principles behind the action of PEG as a protein precipitating agent are not completely understood. Possible modes of action include (1) complex formation between protein and polymer, and (2) exclusion of the protein from part of the solvent (dehydration), followed by protein aggregation and precipitation. The most attractive advantage of PEG is its ability to fractionate proteins on the basis of size and shape, as in gel filtration.

Finally, changes in pH and temperature have been used effectively to promote selective protein precipitation. A change in the pH of the solution alters the ionic state of a protein and may even bring some proteins to a state of charge neutrality.

Charged protein molecules tend to repel each other and remain in solution; however, neutral protein molecules do not repel each other, so they tend to aggregate and precipitate from solution. A protein is least soluble in aqueous solution when it has no net charge, that is, when it is isoelectric. This characteristic can be used in protein purification, since different proteins usually have different isoelectric pH values (see Chapter 6, Section B, p. 182). An increase in temperature generally causes an increase in the solubility of solutes. This general rule is followed by most proteins up to about 40°C. Above this temperature, however, many proteins aggregate and precipitate from solution. If the protein of interest is heat stable and still water soluble above 40°C, a major step in protein purification can be achieved because most other proteins precipitate at these temperatures and can be removed by centrifugation.

Selective Techniques in Protein Purification

After gross fractionation of proteins, as discussed above and listed in Table 11.2 Step 5, more refined methods with greater resolution can be attempted. These methods, in order of increasing resolution, are gel-filtration chromatography, ion-exchange chromatography, affinity chromatography, and isoelectric focusing. Since the basis of protein separation is different for each of these techniques, it is often most effective and appropriate to use all of the techniques in the order given.

Chromatographic methods for protein purification are discussed in Chapter 5. However, we must consider another important topic, preparation of protein solutions for chromatography. Fractionation of heterogeneous protein mixtures by inorganic salts, organic solvents, or PEG usually precedes column chromatography. The presence of the precipitating agents will interfere with the later chromatographic steps. In particular, the presence of ammonium sulfate increases the ionic strength of the protein solution and damps the ionic protein–ion-exchange resin interactions. Procedures that are in current use to remove undesirable small molecules from protein solutions include ultrafiltration (see Chapter 3, Section D, p. 74), dialysis (see Chapter 3, Section D, p. 73), and gel filtration (see Chapter 5, Section E, p. 132).

Chromatography is now, and will continue to be for many years, the most effective method for selective protein purification. The more conventional methods (ion exchange and gel filtration) rely on rather nonspecific physicochemical interactions between a stationary support and protein molecule. These techniques, which separate proteins on the basis of net charge, size, and polarity, do not have a high degree of specificity. The highest level of selectivity in protein purification is offered by affinity chromatography—the separation of proteins on the basis of specific biological interactions (see Chapter 5, Section G. p. 152).

To some individuals, especially those who recall their experiences in the organic chemistry laboratory, the ultimate step in purification of a molecule is crystallization. The desire to obtain crystalline protein has long been strong, and many proteins have been crystallized. However, there is a common misconception that the ability to form crystals of a protein ensures that the protein is homogeneous. For many reasons (entrapment of contaminants within crystals, aggregation of protein molecules, etc.), the ability to crystallize a protein should not be used as a criterion of purity. The interest in protein crystals today has its origin in the demand for X-ray crystallographic analysis of protein structure (see Chapter 7, p. 233).

STUDY EXERCISE 11.2 Protein Purification

List at least three types of chromatography useful for protein purification, and identify the basis of separation for each type.

B. PRODUCTION OF PROTEINS BY EXPRESSION OF FOREIGN GENES

It is now possible to produce large quantities of selected proteins by using recombinant DNA technology. In general, a hybrid DNA is prepared by inserting the gene to be expressed (the gene carries the code for the desired protein) into a vector (plasmid, etc.; see Chapter 10, Section A, p. 294) and transferring it into a host cell where the desired protein is synthesized. These procedures have been used in the production of hundreds of proteins for scientific, medical, agricultural, and industrial purposes. Table 11.3 lists some of the important protein products now made by recombinant DNA methods.

Gene Expression in Prokaryotic Organisms

Bacterial cells, especially *E. coli*, have been widely used as host cells for the large-scale production of proteins. Special requirements must be considered while preparing the recombinant DNA containing the "foreign" gene. The cloned structural gene that contains the message for the desired protein is incorporated into an expression vector that has all the necessary transcriptional and translational

TABLE 11.3 Recombinant Proteins and Their Use

Protein	Use
Human insulin	Treatment of diabetes
Human somatotropin (growth hormone)	Treatment of dwarfism
Bovine somatotropin (BST)	Enhances milk production in dairy cattle
Porcine somatotropin (PST)	Enhances growth in pigs
Pulmozyme (DNase)	Treatment of cystic fibrosis
Tissue plasminogen activator (TPA)	Treatment of heart attack, stroke; dissolves blood clots
Erythropoietin	Treatment of anemia by stimulating erythrocyte production
Interferons	Treatment of cancers; antiviral agent
Atrial natriuretic factor	Reduces high blood pressure
Leptin	Treatment of obesity
Hepatitis B vaccine	Treatment of hepatitis
Herceptin	Treatment of metastatic breast cancer; monoclonal antibody
Amylase and cellulase	Removes carbohydrate precipitates from fruit juice

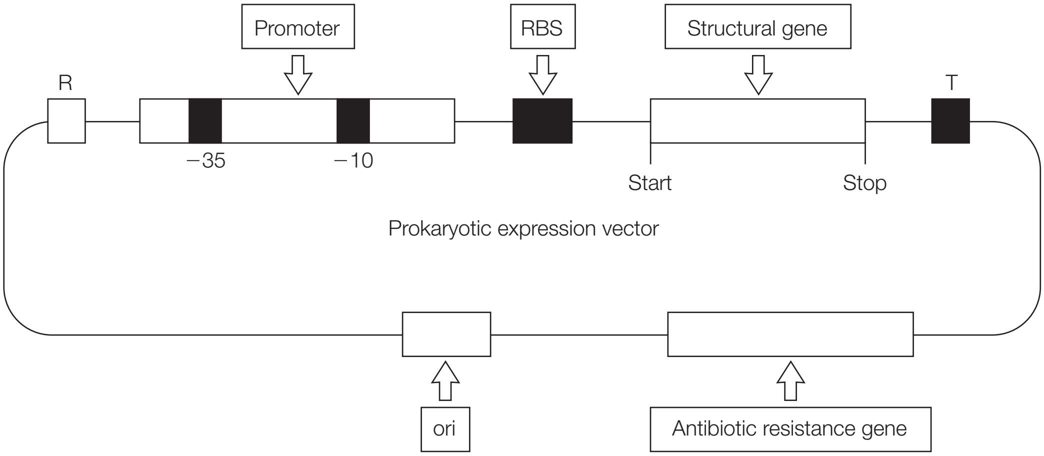

FIGURE 11.3 Components of a typical recombinant DNA prokaryotic expression vector. The structural gene contains the coding sequence for the protein. Essential components include the promoter (P), ribosome binding site (RBS), regulatory gene (R), and transcription terminator (T). "ori" refers to the origin of replication.

control sequences (Figure 11.3). If a relaxed plasmid is used as the vector (see Chapter 10, Section A, p. 294) and the proper signals for gene expression are present, it is possible to overexpress the desired protein so that its concentration may reach up to 40% of the host bacteria's total cellular protein. Although this provides a high yield of the expressed protein, the bacterial cells often concentrate the foreign protein into insoluble **inclusion bodies** in which the protein is present in a denatured, aggregated form. The protein may be renatured by isolating the inclusion bodies by centrifugation and incubating them with a denaturing agent (6 *M* guanidinium ion or urea) and then slowly removing the denaturing agent by ultrafiltration or dialysis (see Chapter 3, Section D, p. 73). Initially, the presence of the desired protein in the inclusion bodies seems to be a disadvantage in the production of the protein, but it turns out sometimes to be a benefit, as it may actually save time in further purification steps. Because the inclusion bodies contain the desired protein in the absence of most other cellular proteins, this serves as a major purification step. The renatured protein often is ready for further purification using the final steps of chromatography in Table 11.2. However, the renaturing process is very slow, and percent recovery of the protein may be as low as 50%.

Host-Cell Secretion of Protein

The formation of inclusion bodies and other technical problems of gene expression can be avoided if the coding gene is genetically engineered to induce the host cell to secrete the protein into the periplasmic space (the area between the cell wall and the plasma membrane). The protein can be released from this space using the procedure of osmotic lysis (see this chapter, p. 311). These procedures greatly facilitate purification of the protein, as they avoid the preparation of a crude cell homogenate in which the protein is in a mixture of all other cellular molecules. The host bacterial cell can be coaxed into secreting the desired protein by incorporating a signal sequence for protein targeting into the expression vector.

The yield of biologically active protein can also be increased by reducing its intracellular degradation. Because the host bacterial cell recognizes the cloned protein as "foreign" (and, indeed, some proteins are toxic to the bacterium), it attempts to remove the protein by protease-catalyzed hydrolysis. Degradation by protease action is lessened if the protein is secreted. Protein degradation can also be reduced if the expression vector is modified to add or label the desired protein with a protease-resistant sequence of amino acids.

Many of the above problems, including inclusion body formation, can also be minimized or even eliminated by tagging the desired protein with a protein of bacterial origin. This is accomplished by inserting into the vector a DNA fragment carrying the code for a bacterial protein. The bacteria can be tricked into recognizing the conjugated protein (bacterial protein covalently linked to the desired protein) as "native." The "label" is often added to the N-terminus of the cloned protein. Such modified proteins are called **hybrid** or **fusion proteins.** The label may be removed from the desired protein, after extraction, by using a proteolytic enzyme.

Histidine-Tagged Proteins

A unique approach to eliminating some of the problems outlined above and also facilitating protein purification is to program the expression vector to add a string of six or more histidine residues to the N-terminus of the desired protein. The histidine residues (along with the protein) will bind to an affinity chromatography column containing nickel ions immobilized to a nitrilotriacetic acid solid support. Protein purification is achieved by passing the cell lysate (prepared by treatment of the host cells with lysozyme and/or a detergent) directly through the affinity column (Figure 11.4; see immobilized metal-ion affinity chromatography, Chapter 5, Section G, p. 152). Equipment and reagents for these procedures are available in kit form from QIAGEN Inc. and Clontech Laboratories Inc. A modified version of this procedure, the MagneHis Protein Purification System, is available from Promega. The MagneHis system involves the use of paramagnetic precharged nickel particles to isolate His-tagged protein directly from a crude cell lysate. In all procedures described above, the His-tagged protein is recovered from the solid support by eluting with imidazole, which binds to the nickel ions and displaces the protein.

Other peptide fusion labels that facilitate protein purification include:

- ***Polyarginine***—A series of arginine residues is added to the C-terminus of the protein. This makes the protein highly basic and thus able to be purified using a cation-exchange resin.
- ***Flag***—A short amino acid sequence (Asp-Tyr-Lys-Asp-Asp-Asp-Asp-Lys–) is attached to the N-terminus of the protein. The protein may then be purified by immunoaffinity chromatography with the use of a commercially available monoclonal antibody that recognizes the hydrophilic peptide sequence.

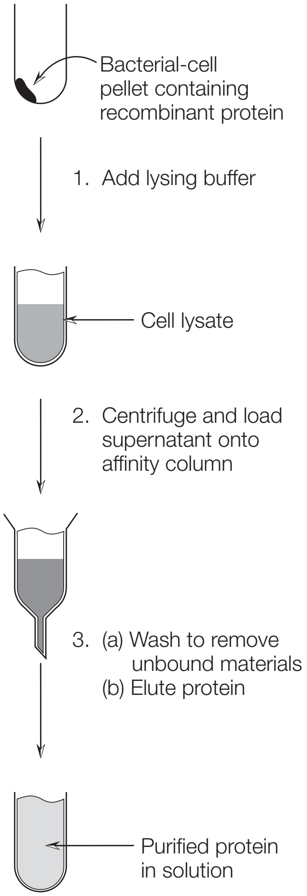
Bacterial-cell pellet containing recombinant protein
1. Add lysing buffer
Cell lysate
2. Centrifuge and load supernatant onto affinity column
3. (a) Wash to remove unbound materials
(b) Elute protein
Purified protein in solution

FIGURE 11.4 Purification of a His-tagged protein using immobilized metal-affinity chromatography. See text for details. www.clontech.com.

An extremely useful application of recombinant DNA technology is the ability to modify protein structures by changing the identity of amino acid residues. This provides a powerful tool for monitoring which amino acid residues are important in a protein's biological function. Amino acid changes may be made by a technique called **site-directed mutagenesis,** where an altered gene carrying the message for different amino acids is inserted into a recombinant DNA.

STUDY EXERCISE 11.3 Genetic Engineering and Chromatography

What kind of chromatography is used to separate His-tagged proteins?

Gene Expression in Eukaryotic Cells

Expression of a eukaryotic gene in a bacterial host cell often leads to many technical problems. Among these problems are:

1. Bacterial cells are unable to carry out gene splicing processes that remove introns from the eukaryotic mRNA.
2. Bacterial cells do not recognize and respond to eukaryotic transcriptional and translational regulatory elements.
3. Bacterial cells do not have the enzyme systems to direct post translational modifications of proteins, as eukaryotic cells are capable of doing. These protein modifications include chemical changes such as addition of carbohydrates (glycosylation), formation of correct disulfide bonds, proteolytic cleavage of inactive precursors, phosphorylation of amino acid hydroxyl groups (tyrosine, serine, etc.), and other changes that assist the protein to fold into its native conformation and become biologically active.

The challenges of incorporating all of these messages into an expression vector for bacterial cells are great, if not insurmountable. The best approach is to use eukaryotic cells such as yeast as hosts.

The many challenges presented above, such as preparing complex recombinant DNA with all the proper signals and messages, and maintaining and using bacterial and eukaryotic host-cell cultures, have encouraged scientists to search for more efficient systems for producing proteins.

cDNA

One procedure for making recombinant proteins that is growing in popularity is the use of **complementary DNA (cDNA).** All of the proteins in a cell are made from their specific mRNAs by translation, so it would seem logical to isolate the mRNA associated with a specific protein and then use it to make that protein. However, mRNA is too difficult to purify and too unstable to use directly for protein synthesis. Instead, the mRNA is converted into a DNA form using the RNA-viral enzyme system **reverse transcriptase (RT)**. RT is able to read an RNA template and catalyze the synthesis of a complementary strand of DNA (cDNA). (RT is the enzyme that retroviruses use to copy their genes, originally in the form of RNA, into DNA, in order to "take over" a cell or organism.) Researchers now can isolate and purify the mRNA associated with the desired protein, transform it into cDNA with RT, and insert the cDNA into a suitable vector for cloning.

Cell-Free Systems

Another promising alternative for protein production is to use cell-free transcription and translation systems for expression of the recombinant DNA. Two *in vitro* systems currently being used are wheat-germ extracts and rabbit reticulocyte lysates. The *in vitro* systems require the addition of the isolated mRNA fragment that contains the coded message for the protein, as well as amino acids, tRNAs, and ribosomes.

Protein Expression Using Synthesized Genes

Producing large quantities of a protein from naturally-occurring gene sequences as described above is not always efficient. It can be a challenging, time consuming, and costly process. Some of the most valuable and desired proteins needed by biochemists are often expressed in wild-type cells, but in very low amounts. Biochemists have begun to experiment with synthesized genes in which the three nucleotides in the codons are revised or rewritten with the hope of optimizing protein expression. How is this possible? Recall that the genetic code is degenerate. Almost all of the 20 amino acids have at least two codons (Leu has six); thus, different codons can specify the same amino acid. For example, a codon for Phe, U-U-U, could be changed to U-C-U, which would also specify Phe. Three of the codons signal termination, which leaves 61 of the 64 total codons to specify amino acids. Thus, it is possible to change about a third (20/61) of the nucleotides and still produce the same protein. The number of alternate gene sequences that will produce the same protein is enormous. According to Wikipedia, there are 10^{150} codon combinations for an average-sized protein of 300 amino acids. It has been shown experimentally that some combinations lead to better protein yields than others. Thus, modifying a synthetic gene can potentially enhance protein expression. Synthetic biology companies including DNA2.0, PolyQuant (Entelechon), GENEWIZ, GENEART, GenScript, Epoch Bioloabs, and others are developing computer software based on new design algorithms to predict the set of codons most effective for protein expression. Gene optimization can offer protein yields up to at least 10 times higher than the more conventional procedures. Finding just the right gene, however, requires extensive computer analysis. A major goal of these investigations is to determine the gene sequence characteristics that have a significant influence on protein expression.

C. PROTEIN CHARACTERIZATION

Now that the sought-after protein has been obtained in a highly purified state, it is usually suitable for further characterization and study. The characterization techniques applied to the protein depend on what is already known about the molecule and what kind of protein it is. One is, of course, interested in knowing the biological role the protein plays in the organism; however, such a determination must wait until more is known about its physical and chemical properties. Here is an outline of procedures one should apply to the protein for general characterization and for specific analysis. References are made to coverage of the characterization procedures in this book (see Appendix I for a table of useful Web sites, p. 331).

- Concentration of protein solutions—general assays like the Bradford, Lowry, spectrophotometric, etc. (see Chapter 3, Section B, p. 67).
- Purity of protein—PAGE and SDS-PAGE (see Chapter 6, Section B, p. 167), HPLC (see Chapter 5, Section F, p. 140), capillary electrophoresis (see Chapter 6, Section B, p. 185), isoelectric focusing (see Chapter 6, Section B, p. 182), and 2D electrophoresis (see Chapter 6, Section B, p. 184).
- Molecular weight, subunit structure—gel-filtration chromatography (see Chapter 5, Section E, p. 132), SDS-PAGE (see Chapter 6, p. 174), and MS (see Chapter 7, Section D, p. 230).
- Identity and structure determination (including cofactors, coenzymes, metal ions, carbohydrates, etc.)—UV/VIS spectroscopy (see Chapter 7, Section A, p. 202), NMR spectroscopy (see Chapter 7, Section C, p. 225), MS (see Chapter 7, Section D, p. 230), X-ray diffraction (see Chapter 7, Section E, p. 233), and protein databases (Table 11.4).
- Sequence of amino acids in the protein and of nucleotides in the coding gene—Web sites, databases (see Table 11.4 and Chapters 2, 9, and 10). For

TABLE 11.4 Protein Database Resources

Name	URL
ExPASy—UniProt Knowledgebase: Swiss-Prot and TrEMBL	http://ca.expasy.org/sprot
PIR—Protein Information Resource	http://pir.georgetown.edu
RCSB Protein Data Bank (PDB)	http://www.rcsb.org
The National Center for Biotechnology	http://www.ncbi.nlm.nih.gov
The European Bioinformatics Institute	http://www.ebi.ac.uk

analysis of the amino acid sequence (primary structure), see this chapter, Section D.

- Biological function:
  Enzyme–substrate specificity, kinetics, inhibition, regulation (see Chapter 8, Section B, p. 250).
  Transport protein–ligand binding (fluorescence spectroscopy, see Chapter 7, Section B, p. 220, and Chapter 8, Section A, p. 239).
  Other type of protein–procedures depend on function.
- Proteomics: identity of unknown proteins by amino acid composition and prediction of structure (see Web sites and databases in Table 11.4 and Chapter 2), analysis by 2D electrophoresis (see Chapter 6, Section B, p. 184), interactions with other proteins—MALDI-MS (see Chapter 7, Section D, p. 230).

D. DETERMINATION OF PRIMARY STRUCTURE

Structural elucidation of natural macromolecules is an important step in understanding the relationships between the chemical properties of a biomolecule and its biological function. The techniques used in organic structure determination (NMR, IR, UV, and MS; Chapter 7) are quite useful when applied to biomolecules, but the unique nature of natural molecules also requires the application of specialized chemical procedures. Proteins, polysaccharides, and nucleic acids are polymeric materials, each composed of hundreds or sometimes thousands of monomeric units (amino acids, monosaccharides, and nucleotides, respectively). But there is only a limited number of these types of units from which the biomolecules are synthesized. For example, only 20 different amino acids are found in proteins, but these different amino acids may appear several times in the same protein molecule. Therefore, the structure of a peptide or protein can be recognized only after the amino acid composition ***and*** sequence have been determined.

Amino Acid Composition

The amide bonds in peptides and proteins can be hydrolyzed in a strong acid or base. Treatment of a peptide or protein under either of these conditions yields a mixture of the constituent amino acids. Neither acid- nor base-catalyzed hydrolysis of a protein leads to ideal results because both tend to destroy some constituent amino acids. Acid-catalyzed hydrolysis destroys tryptophan and cysteine, causes some loss of serine and threonine, and converts asparagine and glutamine to aspartic acid and glutamic acid, respectively. Base-catalyzed hydrolysis leads to destruction of serine, threonine, cysteine, and cystine, and also results in racemization of the free amino acids. Because acid-catalyzed hydrolysis is less destructive, it is often the method of choice. The hydrolysis procedure consists of dissolving the protein sample in aqueous acid, usually 6 *M* HCl, and heating the solution in a sealed, evacuated vial at 100°C for 12 to 24 hours.

Now that the free amino acids present in a peptide or protein have been released by hydrolysis, they must be separated and identified. The most versatile, economical, and convenient techniques for separation are based on chromatographic methods (see Chapter 5 for a review, p. 118). Earlier workers relied on paper chromatography (PC) and thin-layer chromatography (TLC); however, more sensitive techniques are now available. Automated ion-exchange chromatography (amino acid analyzers), capillary electrophoresis (CE), and high-performance liquid chromatography (HPLC) are now powerful tools for the qualitative and quantitative analysis of amino acids and their derivatives. Because there is still some demand for rapid, qualitative, routine analysis of amino acids, thin-layer and paper chromatographic methods are still being developed and improved. Amino acids and their derivatives may be analyzed directly by PC and TLC without further derivatization. Several support materials are available, but most analyses are carried out on silica gel or cellulose. The free amino acids can be detected on the developed chromatographic plates by reaction with ninhydrin. A pink-purple color is obtained for all amino acids except proline. A yellow color develops with proline.

High-performance liquid chromatographic techniques have been successfully applied to the analysis of phenylthiohydantoin, 2,4-dinitrophenyl, and dansyl amino acid derivatives.

If only very small samples of amino acids are available for analysis, fluorescence is used for detection. One of the most sensitive methods of microanalysis is based on the reaction of amino acids with *o*-phthalaldehyde and β-mercaptoethanol (Equation 11.1). The isoindole derivative is fluorescent, and amounts as small as 10^{-12} mole may be measured.

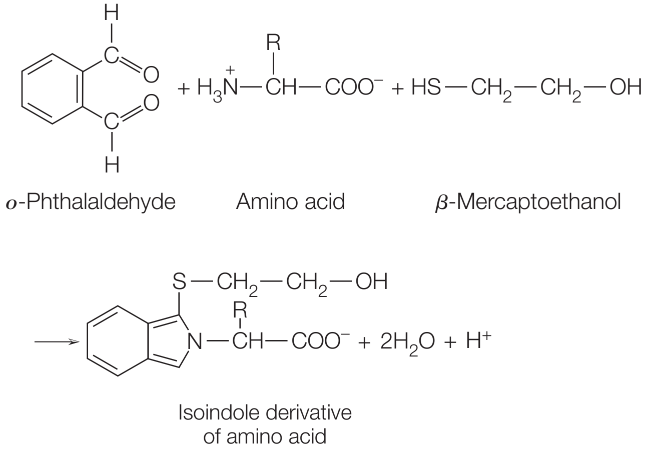

(Eq. 11.1)

Another important derivatizing reagent for amino acids is 9-fluorenylmethyl chloroformate (FMOC) (Equation 11.2). This reagent has the advantages that: (1) the reaction with amino acids is very fast, occurring in less than 1 min; (2) the products are stable for long periods of time; (3) the derivatives can be separated by reversed-phase column chromatography and capillary electrophoresis procedures; and (4) the products are fluorescent, allowing for easy detection.

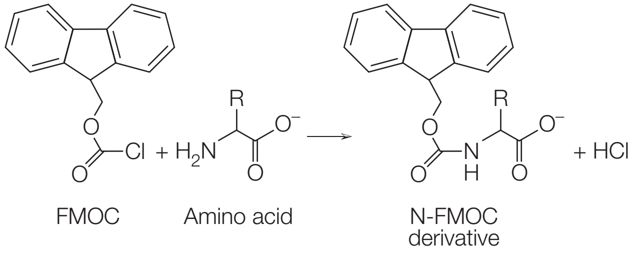
R
O
O⁻
Cl
+ H₂N
O
R
O⁻
N
H
O
+ HCl
FMOC
Amino acid
N-FMOC derivative

(Eq. 11.2)

Sequential Analysis

Sequential analysis of amino acids in purified peptides and proteins is best initiated by analysis of the terminal amino acids. A peptide has one amino acid with a free α amino group (NH_2-terminus) and one amino acid with a free α carboxyl group (COOH-terminus). Many chemical methods have been developed to selectively tag and identify these terminal amino acids.

N-Terminal Analysis

NH_2-terminal amino acid analysis is achieved by the use of: (1) 2,4-dinitrofluorobenzene (Sanger reagent), (2) 1-dimethylaminonaphthalene-5-sulfonyl chloride (dansyl chloride), or (3) phenylisothiocyanate (Edman reagent). Figure 11.5 shows the structures of these reagents. Although the chemistry is different for

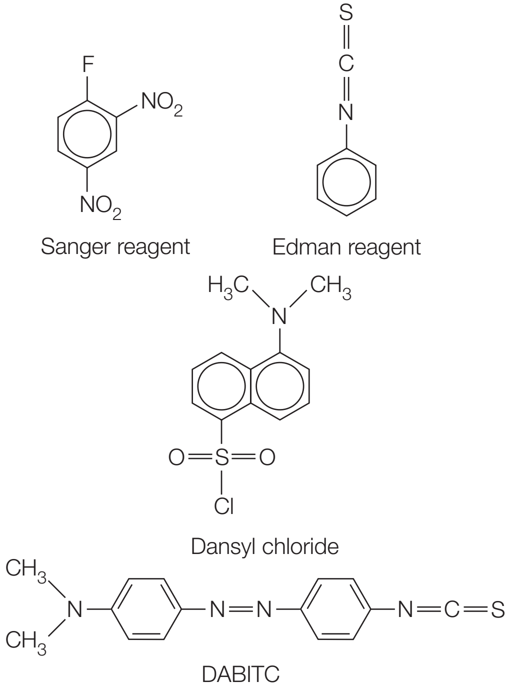
F
NO₂
NO₂
Sanger reagent
S
C
N
Edman reagent
H₃C
CH₃
N
O=S=O
Cl
Dansyl chloride
CH₃
N
CH₃
N=N
N=C=S
DABITC

FIGURE 11.5
Reagents useful in amino-terminal analysis of proteins.

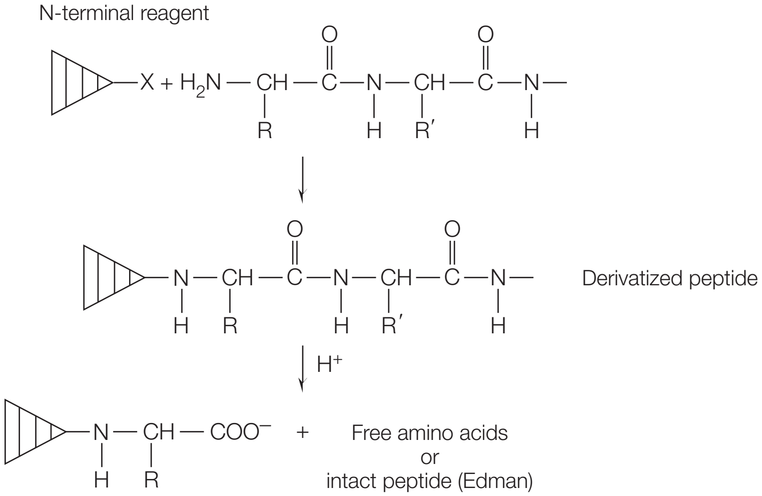

FIGURE 11.6 Scheme to show analysis of amino-terminal residue of protein.

each of these reagents, the same general concept applies. These compounds react with the NH_2-terminal amino acid of a peptide or protein to produce covalent derivatives that are stable to acid-catalyzed hydrolysis. The process is illustrated in Figure 11.6. The mixture of modified amino acid and free amino acids is separated, and the NH_2-terminal amino acid is identified by thin-layer, paper, or high-performance liquid chromatography; or capillary electrophoresis. The first NH_2-terminal method to be widely used was based on the Sanger reagent, which produces yellow-colored 2,4-dinitrophenyl (DNP) derivatives of NH_2-terminal amino acids. The Sanger method has several disadvantages, including poor yield of the DNP-peptide derivative, low sensitivity of analysis, and instability of some DNP-amino acids during acid hydrolysis. The dansyl chloride method has largely replaced the Sanger method because very sensitive fluorescence techniques may be used for detection and analysis of the dansyl amino acid derivatives, and the derivatives are more stable during acid hydrolysis.

Edman Sequence Method

Even more versatile than the dansyl method is the Edman method (Figure 11.7). The NH_2-terminal amino acid is removed as its phenylthiohydantoin (PTH) derivative under anhydrous acid conditions, while all other amide bonds in the peptide remain intact. The derivatized amino acid is then extracted from the reaction mixture and identified by paper, thin-layer, or high-performance liquid chromatography. The intact peptide (minus the original NH_2-terminal amino acid) may be isolated and recycled by reaction with phenylisothiocyanate. Since this method is nondestructive to the remaining peptide (aqueous acid hydrolysis is not required) and results in good yield, it can be used for stepwise sequential analysis of peptides.

Microsequencing

The development of new chemical reagents and instrumentation has now made it possible to achieve end-group analysis and sequencing on extremely small samples of protein (microsequencing). By using a chromophoric derivative of

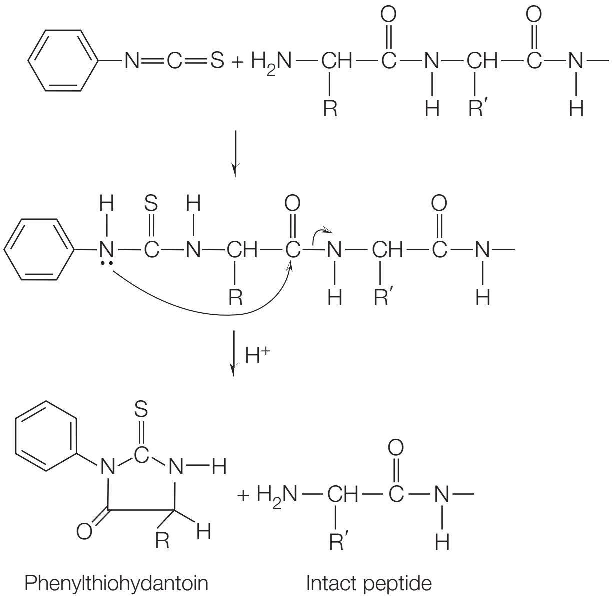

FIGURE 11.7 The Edman method for analysis of proteins.

phenylisothiocyanate, 4-*N*,*N*-dimethylaminoazobenzene-4′-isothiocyanate (DABITC; see Figure 11.5), and analyzing the derivatives by HPLC, it is possible to sequence peptides and proteins at the nanomolar level.

C-Terminal Analysis

The COOH-terminal amino acid of a peptide or protein may be analyzed by enzymatic methods. The method of choice is peptide hydrolysis catalyzed by carboxypeptidases A and B. These two enzymes catalyze the hydrolysis of amide bonds at the COOH-terminal end of a peptide (Equation 11.3), since carboxypeptidase action requires the presence of a free α carboxyl group in the substrate.

Carboxypeptidase A catalyzes the hydrolysis of carboxyl-terminal acidic or neutral amino acids; however, the rate of hydrolysis depends on the structure of the side chain R′. Amino acids with nonpolar aryl or alkyl side chains are cleaved more rapidly. Carboxypeptidase B is specific for the hydrolysis of basic COOH-terminal amino acids (lysine and arginine). Neither peptidase functions if proline occupies the COOH-terminal position or is the next to last amino acid.

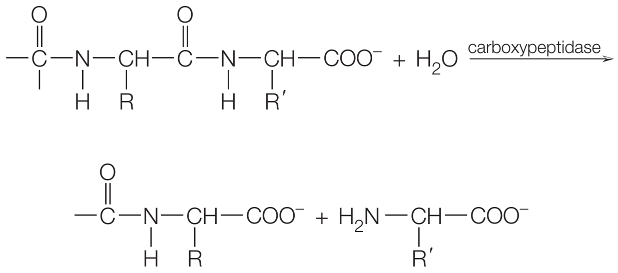

$$-\underset{|}{\overset{\overset{O}{\|}}{C}}-\underset{\underset{H}{|}}{N}-\underset{\underset{R}{|}}{CH}-\overset{\overset{O}{\|}}{C}-\underset{\underset{H}{|}}{N}-\underset{\underset{R'}{|}}{CH}-COO^- + H_2O \xrightarrow{\text{carboxypeptidase}}$$

(Eq. 11.3)

$$-\overset{\overset{O}{\|}}{C}-\underset{\underset{H}{|}}{N}-\underset{\underset{R}{|}}{CH}-COO^- + H_2N-\underset{\underset{R'}{|}}{CH}-COO^-$$

Sequencing DNA Instead of the Protein

New techniques for protein sequence analysis have emerged from recombinant DNA research. It is now possible to clone long stretches of DNA. Genes carrying the message for selected proteins can be isolated in quantities sufficient for nucleotide sequencing. By using the genetic code, the structure of the protein product from a specific gene can be deduced from the sequence of nucleotides in the DNA. Techniques for DNA sequencing are faster and more reliable than for protein sequencing; however, determining protein sequences indirectly from DNA sequences will never completely replace direct analysis of the protein product. An amino acid sequence determined from DNA will be that of the initial protein product synthesized. That initial form (which is often not biologically active) is changed by post translational modification processes. These changes may include removal of short stretches of amino acid residues, formation of disulfide bonds, and chemical changes on some amino acid residues.

Study Problems

1. A solution of a purified protein yielded an A_{280}/A_{260} of 1.3. Estimate the protein concentration of the solution.
2. Describe the action of the enzyme lysozyme in breaking bacterial cells for protein isolation and purification.
3. What amino acid residues in a protein absorb light of 280 nm?
4. Use a flowchart format to show how you would isolate a protein from an inclusion body.
5. A protein you wish to study is present in the mitochondria of the cell. Show how you would use fractional centrifugation to isolate mitochondria to begin purification of the protein.
6. When would you use the technique of isoelectric focusing in the isolation and characterization of a protein?
7. What type of electrophoresis would be better to study the subunit structure of a protein, PAGE or SDS-PAGE? Explain.
8. During the expression of a protein in a bacterial host cell, what are some of the advantages of fusing a natural, bacterial protein to the “foreign” protein?
9. When the histidine-tagged protein procedure is used for purification, the expressed protein is recovered from an affinity column by eluting with a solution of imidazole. What atoms of imidazole displace the protein by binding to nickel?
10. Use a protein database to find the primary structure of the active form of human insulin (51 amino acids).

Further Reading

J. Berg, J. Tymoczko, and L. Stryer, ***Biochemistry,*** 6th ed. (2007), Freeman (New York).

P. Bonner, ***Protein Purification: The Basics*** (2007), Taylor and Francis (New York).

J. Bornhorst and J. Falke, ***Met. Enzymol***. **326,** 245–254 (2000). “Purification of Proteins Using Polyhistidine Affinity Tags.”

R. Boyer, ***Concepts in Biochemistry,*** 3rd ed. (2006), John Wiley & Sons (Hoboken, NJ), pp. 84–90. Protein purification and analysis.

A. Campbell and L. Heyer, ***Discovering Genomics, Proteomics, and Bioinformatics*** (2003), Benjamin Cummings (San Francisco).

F. Carvalho and M. Gillespie, ***Biochem. Mol. Biol. Educ.*** **31,** 46–51 (2003). "Bringing Proteomics into the Lab."

E. Christodoulou and C. Vorgias, ***Biochem. Mol. Biol. Educ.*** **30,** 189–191 (2002). "Understanding Heterologous Protein Overproduction Under the T7 Promoter."

R. Garrett and C. Grisham, ***Biochemistry,*** 4th ed. (2010), Brooks/Cole-Cengage (Belmont, CA).

D. Nelson and M. Cox, ***Lehninger Principles of Biochemistry,*** 5th ed. (2009), Freeman (New York).

B. Nielsen and S. Echols, ***Biochem. Mol. Biol. Educ.*** **30,** 408–413 (2002), "Use of Chloroplast rRNA Gene to Introduce Basic Molecular Biology Techniques."

L. Plesniak and E. Bell, ***Biochem. Mol. Biol. Educ.*** **31,** 127–130 (2003). "Teaching Proteomics."

N. Price and J. Nairn, ***Exploring Proteins: A Student's Guide to Experimental Skills and Methods*** (2009), Oxford University Press (New York).

S. Roe, Editor, ***Protein Purification Techniques,*** 2nd ed. (2001), Oxford University Press (Oxford, UK).

R. Simpson (Editor), ***Purifying Proteins for Proteomics: A Laboratory Manual*** (2004), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY).

R. Simpson et al., ***Basic Methods in Protein Purification and Analysis: A Laboratory Manual*** (2008), Cold Spring Harbor Laboratory Press (Cold Spring Harbor, NY).

C. Sommer et al., ***Biochem. Mol. Biol. Educ.*** **32,** 7–10 (2004). "Teaching Molecular Biology to Undergraduate Biology Students."

C. Tsai, ***An Introduction to Computational Biochemistry*** (2002), John Wiley & Sons (New York).

D. Voet and J. Voet, ***Biochemistry,*** 3rd ed. (2005), John Wiley & Sons (Hoboken, NJ), pp. 116–160. Introduction to protein purification and molecular cloning.

D. Voet, J. Voet, and C. Pratt, ***Fundamentals of Biochemistry,*** 3rd ed. (2008), John Wiley & Sons (Hoboken, NJ).

http://www.piercenet.com
A discussion on protein purification.

http://www.promega.com
A discussion of the MagneHis Protein Purification System.

Additional Web sites for this chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

This page intentionally left blank

APPENDIX I

List of Software Programs and Web Sites Useful for Each Chapter

This table lists Web sites that are particularly relevant to specific chapters. The list is not all-inclusive, but it does include popular and especially helpful URLs. Individual chapters also contain detailed lists of associated Web sites. Additional Web sites for each chapter are available on the book's Companion Website, which can be found at http://www.chemplace.com.

Chapter No.	Name	Description	URL
1	Sigma-Aldrich	MSDS	www.sigma-aldrich.com
1	Swarthmore College	Scientific posters	http://www.swarthmore.edu/NatSci/courrin1/posteradvice.htm
1-3	Rainin	Pipetting	http://www.rainin.com
1-3	Graphpad	Statistics	http://www.graphpad.com/prism/Prism.htm
1-3	Journal of Chemical Education Resource Shelf	List of all current biochemistry textbooks	http://www.umsl.edu/divisions/artscience/chemistry/books/welcome.html
1-3	ChemDraw	Chemical structure drawing program	http://www.cambridgesoft.com
1-11	Intute	Best Web sites	http://www.intute.ac.uk/biologicalsciences
1-11	BioChemWeb	Biochemistry library	http://www.biochemweb.org
2-11	Protocol-Online	Your lab's reference book	http://www.protocol-online.org
2-11	Protein Data Bank (PDB)	Protein structures determined by X-ray and NMR	http://www.rcsb.org/pdb
2-11	ChemBioFinder	Life science solutions	http://www.chembiofinder.com
2-11	NCBI Molecules to Go	Structures of biomolecules	http://molbio.info.nih.gov.cgi-bin/pdb
2-11	PyMol Molecular Viewer	Structures of biomolecules	http://pymol.org
2-11	Chimera Molecular Modeling System (UCSF)	Advanced molecular modeling	http://www.cgl.ucsf.edu/chimera
2-11	Biomolecules Alive	Instructional material	http://www.biomoleculesalive.org
2-11	Entrez Browser of NCBI	Database searching including PubMed literature	http://www.ncbi.nlm.nih.gov/Entrez

(continued)

Chapter No.	Name	Description	URL
2-11	MEDLINE (PubMed)	U.S. National Library of Medicine	http://www.nlm.nih.gov
2-11	International Union of Biochemistry and Molecular Biology (IUBMB)	Nomenclature	http://www.chem.qmul.ac.uk/iubmb
7	Biological Magnetic Resonance Data Bank	NMR data of biomolecules	http://bmrb.wisc.edu
7	Science of Spectroscopy	Spectroscopy applications and techniques	http://www.scienceofspectroscopy.info/index.htm
8	DynaFit	Ligand binding	http://www.biokin.com
8	SigmaPlot	Enzyme kinetics module	http://www.sigmaplot.com/products/sigmaplot/enzyme-mod.php
8	Enzyme Structures Database	Enzyme structures in Protein Data Bank	http://www.biochem.ucl.ac.uk/bsm/enzymes/index.html
8-11	IUBMB Enzyme List	Catalog of enzymes	http://www.chem.qmul.ac.uk/iubmb/enzyme
8-11	Worthington Enzyme Manual	Properties of enzymes	http://www.worthington-biochem.com/index/manual/html
8-11	BRENDA	A comprehensive enzyme information system	http://www.brenda-enzymes.info
9-10	European Bioinformatics Institute (EBI)	DNA sequences	http://www.ebi.ac.uk
9-10	GenBank of the NCBI	Nucleotide sequence database	http://www.ncbi.nlm.nih.gov/Genbank
9-10	EMBL	Nucleotide sequence database	http://www.ebi.ac.uk
9-10	The Institute for Genomic Research	Collection of genomic databases	http://www.ebi.ac.uk/genomes/index.html
9-11	Enzyme Database of ExPASy	Enzyme names and numbers	http://www.expasy.ch/enzyme
9-11	The Broad Institute of MIT and Harvard	Genomics	http://www.broadinstitutes.org
10	REBASE-The Restriction Enzyme Database	Restriction enzyme directory and action	http://rebase.neb.com
10	Fermentas	Restriction enzymes	http://www.fermentas.com

(continued)

Chapter No.	Name	Description	URL
11	Munich Information Center for Protein Sequences (MIPS)	Protein sequences	http://www.helmholtz-muenchen.de/en/mips
11	Protein Information Resource (PIR)	Database searching for proteins	http://pir.georgetown.edu
11	Swiss-Protein	Protein sequences	http://ca.expasy.org/sprot

APPENDIX II

Properties of Common Acids and Bases

Compound	Formula	Molecular Weight	Specific Gravity	% by Weight	Molarity (*M*)
Acetic acid, glacial	CH_3COOH	60.1	1.05	99.5	17.4
Ammonium hydroxide	NH_4OH	35.0	0.89	28	14.8
Formic acid	$HCOOH$	46.0	1.20	90	23.4
Hydrochloric acid	HCl	36.5	1.18	36	11.6
Nitric acid	HNO_3	63.0	1.42	71	16.0
Perchloric acid	$HClO_4$	100.5	1.67	70	11.6
Phosphoric acid	H_3PO_4	98.0	1.70	85	18.1
Sulfuric acid	H_2SO_4	98.1	1.84	96	18.0

APPENDIX III

Properties of Common Buffer Compounds

			(20°C)			
Compound	**Abbreviation**	**Molecular Weight**	**pK_1**	**pK_2**	**pK_3**	**pK_4**
N-(2-Acetamido)-2-aminoethanesulfonic acid	ACES	182.2	6.9	–	–	–
N-(2-Acetamido)-2-iminodiacetic acid	ADA	212.2	6.60	–	–	–
Acetic acid	–	60.1	4.76	–	–	–
Arginine	Arg	174.2	2.17	9.04	12.48	–
Barbituric acid	–	128.1	3.79	–	–	–
N,N-bis (2-Hydroxyethyl)-2-aminoethane-sulfonic acid	BES	213.1	7.15	–	–	–
N,N-bis (2-Hydroxyethyl)glycine	Bicine	163.2	8.35	–	–	–
Boric acid	–	61.8	9.23	12.74	13.80	–
Citric acid	–	210.1	3.10	4.75	6.40	–
Ethylenediaminetetraacetic acid	EDTA	292.3	2.00	2.67	6.24	10.88
Formic acid	–	46.03	3.75	–	–	–
Fumaric acid	–	116.1	3.02	4.39	–	–
Glycine	Gly	75.1	2.45	9.60	–	–
Glycylglycine	–	132.1	3.15	8.13	–	–
N-2-Hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid	HEPES	238.3	7.55	–	–	–
N-2-Hydroxyethylpiperazine-*N*′-3-propanesulfonic acid	HEPPS	252.3	8.0	–	–	–
Histidine	His	209.7	1.82	6.00	9.17	–
Imidazole	–	68.1	6.95	–	–	–
2-(*N*-Morpholino) ethanesulfonic acid	MES	195	6.15	–	–	–
3-(*N*-Morpholino) propanesulfonic acid	MOPS	209.3	7.20	–	–	–
Phosphoric acid	–	98.0	2.12	7.21	12.32	–
Succinic acid	–	118.1	4.18	5.60	–	–
3-Tris (hydroxymethyl) aminopropanesulfonic acid	TAPS	243.2	8.40	–	–	–

(continued)

Compound	Abbreviation	Molecular Weight	(20°C)			
			pK_1	pK_2	pK_3	pK_4
N-Tris (hydroxymethyl)methyl-2-aminoethanesulfonic acid	TES	229.2	7.50	–	–	–
N-Tris (hydroxymethyl)methylglycine	Tricine	179	8.15	–	–	–
Tris (hydroxymethyl)aminomethane	Tris	121.1	8.30	–	–	–

APPENDIX IV

pK_a Values and pH_I Values of Amino Acids

Name	Abbreviations		pK_1 (α-carboxyl)	pK_2 (α-amino)	pK_R (side chain)	pH_I
Alanine	Ala	A	2.3	9.7	–	6.0
Arginine	Arg	R	2.2	9.0	12.5	10.8
Asparagine	Asn	N	2.0	8.8	–	5.4
Aspartate	Asp	D	2.1	9.8	3.9	3.0
Cysteine	Cys	C	1.7	10.8	8.3	5.0
Glutamate	Glu	E	2.2	9.7	4.3	3.2
Glutamine	Gln	Q	2.2	9.1	–	5.7
Glycine	Gly	G	2.3	9.6	–	6.0
Histidine	His	H	1.8	9.2	6.0	7.6
Isoleucine	Ile	I	2.4	9.7	–	6.1
Leucine	Leu	L	2.4	9.6	–	6.0
Lysine	Lys	K	2.2	9.0	10.5	9.8
Methionine	Met	M	2.3	9.2	–	5.8
Phenylalanine	Phe	F	1.8	9.1	–	5.5
Proline	Pro	P	2.0	10.6	–	6.3
Serine	Ser	S	2.2	9.2	–	5.7
Threonine	Thr	T	2.6	10.4	–	6.5
Tryptophan	Trp	W	2.4	9.4	–	5.9
Tyrosine	Tyr	Y	2.2	9.1	10.1	5.7
Valine	Val	V	2.3	9.6	–	6.0

APPENDIX V

Molecular Weight of Some Common Proteins

Name (Source)	**Molecular Weight**
Albumin (bovine serum)	65,400
Albumin (egg white)	45,000
Carboxypeptidase A (bovine pancreas)	35,268
Carboxypeptidase B (porcine)	34,300
Catalase (bovine liver)	250,000
Chymotrypsinogen (bovine pancreas)	23,200
Cytochrome *c*	13,000
Hemoglobin (bovine)	64,500
Insulin (bovine)	5,700
α-Lactalbumin (bovine milk)	14,200
Lysozyme (egg white)	14,600
Myoglobin (horse heart)	16,900
Pepsin (porcine)	35,000
Peroxidase (horseradish)	40,000
Ribonuclease I (bovine pancreas)	12,600
Trypsinogen (bovine pancreas)	24,000
Trypsin (bovine pancreas)	23,800
Tyrosinase (mushroom)	128,000
Uricase (pig liver)	125,000

APPENDIX VI

Common Abbreviations Used in This Text

A	adenine or absorbance
A	absorbance
AMP, ADP, ATP	adenosine mono-, di-, or triphosphate
AP	alkaline phosphatase
BCA	bicinchoninic acid
bp	base pairs
Bq	Becquerel
BSA	bovine serum albumin
C	cytosine
cDNA	complementary DNA
CE	capillary electrophoresis
Ci	Curie
CM cellulose	carboxymethyl cellulose
COSY	correlation spectroscopy
CPM	counts per minute
D	dalton
DANSYL	1,1-dimethylaminonaphthalene-5-sulfonyl chloride
2-DE	two-dimensional electrophoresis
DEAE cellulose	diethylaminoethyl cellulose
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DNP	dinitrophenyl
DOPA	dihydroxyphenylalanine
E	absorption coefficient
EDTA	ethylenediaminetetraacetic acid
ELISA	enzyme-linked immunosorbent assay
ELN	electronic lab rolebook
ESI	electrospray ionization
EtBr	ethidium bromide
FAB	fast-atom bombardment
FAD (H_2)	flavin adenine dinucleotide (reduced form)

(continued)

FAME	fatty-acid methyl ester
FMN (H_2)	flavin mononucleotide (reduced form)
FOCSY	foldover-corrected spectroscopy
FPLC	fast-protein liquid chromatography
FT	Fourier transform
g	gravitational force
G	guanine
GC	gas chromatography
GMP, GDP, GTP	guanosine mono-, di-, or triphosphate
Hb	hemoglobin
HMIS	Hazardous Materials Identification System
HPLC	high-performance liquid chromatography
HRP	horseradish peroxidase
HTML	hypertext markup language
IE	immunoelectrophoresis
IEF	isoelectric focusing
IHP	inositol hexaphosphate
IgG	immunoglobulin gamma
IMAC	immobilized metal-ion affinity chromatography
IR	infrared
ISP	Internet service provider
IVS	intervening sequence
kb	kilobase pairs
L	ligand
M	macromolecule
M	molarity
MALDI	matrix-assisted laser desorption ionization
MS	mass spectrometry
MSDS	material safety data sheet
NAD(H)	nicotinamide adenine dinucleotide (reduced form)
NADP(H)	nicotinamide adenine dinucleotide phosphate (reduced form)
NMR	nuclear magnetic resonance
NOESY	nuclear Overhauser effect spectroscopy
OSHA	Occupational Safety and Health Administration
PAGE	polyacrylamide gel electrophoresis
PC	paper chromatography

PCR	polymerase chain reaction
PEG	polyethylene glycol
PFGE	pulsed-field gel electrophoresis
PMT	photomultiplier tube
POPOP	1,4-bis[5-phenyl-2-oxazolyl]benzene
PPO	2,5-diphenyloxazole
PVDF	polyvinyldifluoride
RCF	relative centrifugal force
R_f	relative mobility in chromatography
RFLP	restriction fragment length polymorphism
RIA	radioimmunoassay
RNA	ribonucleic acid
RNase	ribonuclease
RT	reverse transcriptase
s	sedimentation coefficient
S	Svedberg unit (10^{-13} s)
SDS	sodium dodecyl sulfate
SECSY	spin-echo correlation spectroscopy
SMP	submitochondrial particles
T	thymine
TAG	triacylglycerol
TEMED	*N, N*, *N′*, *N′*-tetramethylethylenediamine
TLC	thin-layer chromatography
TOCSY	2-D total correlation spectroscopy
TOF	time of flight
Tris	tris(hydroxymethyl)aminomethane
TROSY	transverse relaxation-optimized spectroscopy
U	uracil
URL	uniform resource locator
UV	ultraviolet
VIS	visible
WWW	World Wide Web

APPENDIX VII

Units of Measurement

The International System of Units (SI)

Quantity	Unit	Abbreviation
length	meter	m
mass	kilogram	kg
time	second	s
temperature	kelvin	K
electric current	ampere	A
amount of substance	mole	mol
radioactivity	becquerel	Bq
volume	liter	L

Metric Prefixes

Name	Abbreviation	Multiplication Factor (relative to "1")
atto	a	10^{-18}
femto	f	10^{-15}
pico	p ($\mu\mu$)	10^{-12}
nano	n (mμ)	10^{-9}
micro	μ	10^{-6}
milli	m	10^{-3}
centi	c	10^{-2}
deci	d	10^{-1}
deca	da	10
kilo	k	10^{3}
mega	M	10^{6}
giga	G	10^{9}

Units of Length

Name	Abbreviation	Multiplication Factor (relative to meter)
Kilometer	km	10^3
meter	m	1
centimeter	cm	10^{-2}
millimeter	mm	10^{-3}
micrometer	μm	10^{-6}
nanometer	nm	10^{-9}
Angstrom	Å	10^{-10}

Units of Mass

Name	Abbreviation	Multiplication Factor (relative to gram)
kilogram	kg	10^3
gram	g	1
milligram	mg	10^{-3}
microgram	μg	10^{-6}
nanogram	ng	10^{-9}

Units of Volume

Name	Abbreviation	Multiplication Factor (relative to liter)
liter	L	1
deciliter	dL	10^{-1}
milliliter	mL	10^{-3}
microliter	μL	10^{-6}

APPENDIX VIII

Table of the Elements*

	Symbol	Atomic No.	Atomic Mass
Actinium	Ac	89	227.0278
Aluminum	Al	13	26.98154
Americium	Am	95	[243]†
Antimony	Sb	51	121.75
Argon	Ar	18	39.948
Arsenic	As	33	74.9216
Astatine	At	85	[210]
Barium	Ba	56	137.33
Berkelium	Bk	97	[247]
Beryllium	Be	4	9.01218
Bismuth	Bi	83	208.9804
Bohrium	Bh	107	[264]
Boron	B	5	10.81
Bromine	Br	35	79.904
Cadmium	Cd	48	112.41
Calcium	Ca	20	40.08
Californium	Cf	98	[251]
Carbon	C	6	12.011
Cerium	Ce	58	140.12
Cesium	Cs	55	132.9054
Chlorine	Cl	17	35.453
Chromium	Cr	24	51.996
Cobalt	Co	27	58.9332
Copernicium	Cn	112	[285]
Copper	Cu	29	63.546
Curium	Cm	96	[247]
Darmstadtium	Ds	110	[281]
Dubnium	Db	105	[262]
Dysprosium	Dy	66	162.50
Einsteinium	Es	99	[252]

	Symbol	Atomic No.	Atomic Mass
Erbium	Er	68	167.26
Europium	Eu	63	151.96
Fermium	Fm	100	[257]
Fluorine	F	9	18.998403
Francium	Fr	87	[223]
Gadolinium	Gd	64	157.25
Gallium	Ga	31	69.72
Germanium	Ge	32	72.59
Gold	Au	79	196.9665
Hafnium	Hf	72	178.49
Hassium	Hs	108	[270]
Helium	He	2	4.00260
Holmium	Ho	67	164.9304
Hydrogen	H	1	1.0079
Indium	In	49	114.82
Iodine	I	53	126.9045
Iridium	Ir	77	192.22
Iron	Fe	26	55.847
Krypton	Kr	36	83.80
Lanthanum	La	57	138.9055
Lawrencium	Lr	103	[262]
Lead	Pb	82	207.2
Lithium	Li	3	6.941
Lutetium	Lu	71	174.967
Magnesium	Mg	12	24.305
Manganese	Mn	25	54.9380
Meitnerium	Mt	109	[268]
Mendelevium	Md	101	[258]
Mercury	Hg	80	200.59
Molybdenum	Mo	42	95.94
Neodymium	Nd	60	144.24
Neon	Ne	10	20.179
Neptunium	Np	93	237.0482
Nickel	Ni	28	58.70
Niobium	Nb	41	92.9064

(continued)

	Symbol	Atomic No.	Atomic Mass
Nitrogen	N	7	14.0067
Nobelium	No	102	[259]
Osmium	Os	76	190.2
Oxygen	O	8	15.9994
Palladium	Pd	46	106.4
Phosphorus	P	15	30.97376
Platinum	Pt	78	195.09
Plutonium	Pu	94	[244]
Polonium	Po	84	[209]
Potassium	K	19	39.0983
Praseodymium	Pr	59	140.9077
Promethium	Pm	61	[145]
Protactinium	Pa	91	231.0359
Radium	Ra	88	226.0254
Radon	Rn	86	[222]
Rhenium	Re	75	186.207
Rhodium	Rh	45	102.9055
Rubidium	Rb	37	85.4678
Ruthenium	Ru	44	101.07
Rutherfordium	Rf	104	[261]
Samarium	Sm	62	150.4
Scandium	Sc	21	44.9559
Seaborgium	Sg	106	[266]
Selenium	Se	34	78.96
Silicon	Si	14	28.0855
Silver	Ag	47	107.868
Sodium	Na	11	22.98977
Strontium	Sr	38	87.62
Sulfur	S	16	32.06
Tantalum	Ta	73	180.9479
Technetium	Tc	43	[98]
Tellurium	Te	52	127.60
Terbium	Tb	65	158.9254
Thallium	Tl	81	204.37
Thorium	Th	90	232.0381

	Symbol	Atomic No.	Atomic Mass
Thulium	Tm	69	168.9342
Tin	Sn	50	118.69
Titanium	Ti	22	47.90
Tungsten	W	74	183.85
Uranium	U	92	238.029
Vanadium	V	23	50.9415
Xenon	Xe	54	131.30
Ytterbium	Yb	70	173.04
Yttrium	Y	39	88.9059
Zinc	Zn	30	65.38
Zirconium	Zr	40	91.22

[*] *Atomic masses are based on carbon-12.*

[†] *A value given in brackets denotes the mass number of the longest-lived or best-known isotope.*

APPENDIX IX

Answers to Odd-Numbered Study Problems

Chapter 1

1. Personal Protection Index for glacial acetic acid: splash goggles, gloves, synthetic apron, and vapor respirator.
3. Every biochemistry lab is different, so a general picture is not possible. You should know where safety features are present in your lab.
5. Water purified by ion exchange will contain organic molecules that are washed from the ion-exchange resin. These contaminants will increase the UV-absorbance properties of the water.
7. 100 m*M*.
9. 58.5 m*M*; 20 mg/mL; 2%.
11. **a.** Sample mean = +3.21°.
    b. Standard deviation = ±0.043°.
    c. 95% confidence limits = +3.21° ± 0.03° at a probability of 0.05.
13. Use the graduated cylinder to measure and pour 46.1 mL (1 mole) of 100% ethanol into the volumetric flask. Add water to the etched mark and mix well.

Chapter 2

1. **a.** The enzyme tyrosinase is found in a wide variety of plants, animals, and fungal species.
    b. Copper.
    c. Inhibitors: benzoic acid, cyanide, and other compounds that complex with copper. Search for references in PubMed.
    d. Another substrate: catechol.
    e. Use PDB or other site to search for a structure.
3. Proteins purified by IMAC: histidine-tagged proteins used in recombinant protein expression (see Chapter 11, Section B, p. 319); human and bovine α-lactalbumin.
5. -A-A-G-C-T-T-; between A and A.
7. α-lactalbumin is a regulatory protein in milk that when complexed with the enzyme galactosyl transferase stimulates the synthesis of milk sugar, lactose.
9. For a start, see Chapter 11, Section B, p. 321.

Chapter 3

1. NaH_2PO_4: 0.31 mole; 37.2 g.
    Na_2HPO_4: 0.19 mole; 26.9 g.
3. Use 0.03 mole of isoelectric glycine and 0.07 mole of sodium glycinate for 1 liter of solution.
5.

Tube No.	1	2	3	4	5	6
μg IgG	0	10	20	40	80	100
A_{595}	0.0	0.05	0.10	0.20	0.40	0.50

7. 17.5 μg/mL.
9. Pure DNA has a 260/280 ratio of about 1.8. The sample is probably contaminated with protein.
11. Primarily Tyr and Trp; no.
13. A at 260 nm = 0.7; A at 280 nm = 0.39.
15. The very best protein to use as a standard is a purified preparation of the protein being assayed; however, it is often not available.
17. **a.** Citrate.
    b. Imidazole.
    c. Glycine.
19. 0.49/day
21. Solution B: 0.2 M.
    Solution C: 0.02 M.
    Solution D: 0.002 M.
23. 77 days.

Chapter 4

1. (a), (b), (d).
3. Hemoglobin.
5. There are four subunits each with molecular weight 10,000 and two subunits each with molecular weight 30,000.
7. Magnesium ions bind to DNA in place of smaller protons, causing the DNA to spread out and thus become less dense.
9. 125,000 × g.
11. Most mitochondria sediment at 20,000 × g.

Chapter 5

1. **a.** Asp, Gly, His.
   b. Glu, Ala, Arg.
   c. Glu, Phe, His.
3. Cyt c, myoglobin = hemoglobin, serum albumin, egg albumin, pepsinogen.
5. Myosin, catalase, serum albumin, chymotrypsinogen, myoglobin, cyt c.
7. Malate dehydrogenase, alcohol dehydrogenase, glucokinase.
9. A dilute solution of NAD^+ should elute the enzyme from the affinity gel.
11. Hydrogen bonding, ionic bonds, hydrophobic interactions, van der Waals forces.
13. Sephadex G-100.
15. Use PubMed or other computer bibliographic searches.

Chapter 6

1. Charge, size.
3. From top to bottom: serum albumin, egg white albumin, chymotrypsin, lysozyme.
5. The gel matrix in slab gels is more uniform than the gel in column gels, which are made individually.
7. Polyacrylamide gels may be used for nucleic acids up to 2000 base pairs.
9. Acrylamide is a potent neurotoxin, suspected carcinogen, and skin irritant. Gloves and a mask must be worn while handling it in the unpolymerized form.

Weighing and making solutions of acrylamide should be carried out in a well-ventilated hood.

11. Western blotting: used to identify a specific protein or group of proteins by immunoblotting (detection by antibodies). Southern blotting: used to identify a specific base sequence in DNA. Northern blotting: used to identify a specific base sequence in RNA.
13. Nylon membranes, because they are cationic and strongly bind acidic proteins.
15. If SDS-PAGE is used for separation, the proteins to be blotted are denatured. The antibodies used for the detection process must be able to recognize and bind to denatured proteins.
17. The protein has an abundance of nonpolar amino acid residues; thus, it will bind best to polyvinyldifluoride (PVDF) membranes, which are hydrophobic.

Chapter 7

1. **a.** X-ray.
   b. Ultraviolet.
   c. Visible.
3. **(1)** a.
   (2) a.
   (3) c.
   (4) d.
   (5) b.
5. (a), (b), and (d).
7. All the molecules contain alternating double bonds (are highly conjugated).
9. Fluorescing light is measured at right angles to the light irradiating the sample.

Chapter 8

1. First plot: $n = 2$; $K_f = 1.1$ mM. Second plot: $n = 2.5$; $K_f = 0.6$ mM.
3. The tryptophan residue must be close to a site on the human albumin where sugars bind. Bound sugars interact with the tryptophan.
5. The binding of the molecular probe must be reversible; the bound probe must have a different absorbance spectrum from the unbound probe.
7. From Michaelis-Menten plot: $V_{\max} = 140\ \mu M/\text{min}$; $K_M = 40\text{–}45\ \mu M$.
9. Type of inhibition: competitive.
11. The rate is linear for only about 2 minutes.
13. Ionic bonding.

Chapter 9

1. DNA, which forms viscous solutions, is released from the cells into the medium.
3. SDS is a detergent that dissociates protein-lipid complexes in cell membranes.
5. The absorbance at 260 nm = 0.70.
7. It must be assumed that the blue dye marker moves faster in electrophoresis than any of the proteins or nucleic acids analyzed in the gel.
9. **a.** Lowered.
   b. Lowered.
   c. Lowered.
   d. Probably little or no effect.
11. 30 μg/mL.

Chapter 10

1. Many plasmids contain genes that carry messages for the synthesis of proteins that protect microorganisms against antibiotics. The presence of certain antibiotics is a signal to a microorganism that more of these proteins are necessary. The microorganism responds by increasing the rate of production of plasmids.
3. Natural DNA molecules and restriction fragments are too large to penetrate polyacrylamide gels. Even gels with low percentage cross linking are not useful.
5. Restriction endonucleases require the presence of magnesium ions for activity. The quench buffer contains EDTA which complexes transition metal ions in the solution. The metal ions are no longer available for binding by the nuclease molecules and enzyme activity is inhibited.
7. (b).
9. Ten fragments.

Chapter 11

1. The protein concentration is about 1.2 mg/mL.
3. Tyrosine and tryptophan.
5. See Figure 4.11.
7. SDS-PAGE.
9. Nitrogen atoms in imidazole.

Extraction Technologies for Medicinal and Aromatic Plants

INTERNATIONAL CENTRE
FOR SCIENCE AND HIGH TECHNOLOGY

Contributors

Chapter 1

An Overview of Extraction Techniques for Medicinal and Aromatic Plants

Sukhdev Swami Handa

Senior Specialist, Industrial Utilization of Medicinal and Aromatic Plants Earth, Environmental and Marine Sciences and Technologies, ICS-UNIDO, AREA Science Park, Bldg. L2, Padriciano 99, 34012 Trieste, Italy

Chapter 2

Role of Process Simulation to Extraction Technologies for Medicinal and Aromatic Plants

Maurizio Fermeglia

DICAMP-CASLAB, University of Trieste and Scientific Consultant for Process Simulation, ICS-UNIDO, AREA Science Park, Bldg. L2, Padriciano 99, 34012 Trieste, Italy

Chapter 3

Maceration, Percolation and Infusion Techniques for the Extraction of Medicinal and Aromatic Plants

Janardan Singh

Scientist E II, Botany and Pharmacognosy, Central Institute of Medicinal and Aromatic Plants, P. O. CIMAP, Lucknow, India

Chapter 4

Hydrolytic Maceration, Expression and Cold Fat Extraction

Anil Kumar Singh

Scientist F, Essential Oil Analysis Laboratory, Central Institute of Medicinal and Aromatic Plants, P. O. CIMAP, Lucknow, India

Chapter 5

Decoction and Hot Continuous Extraction Techniques

Sudeep Tandon and Shailendra Rane

Scientist EI, Chemical Engineer, Process and Product Development Division, Central Institute of Medicinal and Aromatic Plants, P. O. CIMAP, Lucknow, India

Chapter 6

Aqueous Alcoholic Extraction of Medicinal and Aromatic Plants by Fermentation

Chander Kant Katiyar

Director, Herbal Drug Research, Ranbaxy Research Labs, R&D-II, Plot 20, Sector 18, Udyog Vihar Industrial Area, Gurgaon, India

Chapter 7

Distillation Technology for Essential Oils

Sudeep Tandon

Scientist EI, Chemical Engineer, Process and Product Development Division, Central Institute of Medicinal and Aromatic Plants, P. O. CIMAP, Lucknow, India

Chapter 8

Microdistillation, Thermomicrodistillation and Molecular Distillation Techniques

Vishwas Govind Pangarkar

Professor, University Institute of Chemical Technology, Nathalal Parekh Marg Manunga (East) Mumbai 400 019, India

Chapter 9

Solid Phase Micro-extraction and Headspace Trapping Extraction

Rama Kant Harlalka

Director, Nishant Aromas 424, Milan Industrial Estate, Cotton Green Park, Mumbai 200 033, India

Chapter 10

Supercritical Fluid Extraction of Medicinal and Aromatic Plants: Fundamentals and Applications

***Alberto Bertucco*[1] *and Giada Franceschin*[2]**

[1]Professor, Dipartimento di Principi ed Impianti di Ingegneria Chimica "I. Sorgato", University of Padova, Via Marzolo 9, 35131 Padova, Italy

[2]DIPIC - Department of Chemical Engineering, University of Padova, via Marzolo 9, 35131 Padova, Italy

Chapter 11

Process-scale HPLC for Medicinal and Aromatic Plants

***Madan Mohan Gupta*[1] *and Karuna Shanker*[2]**

[1]Head, Analytical Chemistry Division, Central Institute of Medicinal and Aromatic Plants, P. O. CIMAP, Lucknow, India

[2]Scientist, Analytical Chemistry Division, Central Institute of Medicinal and Aromatic Plants, P. O. CIMAP, Lucknow, India

Chapter 12

Flash Chromatography and Low Pressure Chromatographic Techniques for Separation of Phytomolecules

Sunil Kumar Chattopadhyay

Scientist F, Process and Product Development Division, Central Institute of Medicinal and Aromatic Plants, P. O. CIMAP, Lucknow, India

Chapter 13

Counter-current Chromatography

Santosh Kumar Srivastava

Scientist E II, Phytochemistry, Central Institute of Medicinal and Aromatic Plants, P. O. CIMAP, Lucknow, India

Chapter 14

Quality Control of Medicinal and Aromatic Plants and their Extracted Products by HPLC and High Performance Thin Layer Chromatography

Karan Vasisht

Professor of Pharmacognosy, University Institute of Pharmaceutical Sciences, Panjab University, Chandigarh 160 014, India

Preface

Medicinal plants are the richest bioresource of drugs for traditional systems of medicine, modern medicines, nutraceuticals, food supplements, folk medicines, pharmaceutical intermediates and chemical entities for synthetic drugs. Aromatic plants are a source of fragrances, flavors, cosmeceuticals, health beverages and chemical terpenes. Medicinal and aromatic plants (MAPs) are traded as such in bulk from many developing countries for further value addition in developed countries. The first step in the value addition of MAP bioresources is the production of herbal drug preparations (i.e. extracts), using a variety of methods from simple traditional technologies to advanced extraction techniques.

Extraction (as the term is pharmaceutically used) is the separation of medicinally active portions of plant (and animal) tissues using selective solvents through standard procedures. Such extraction techniques separate the soluble plant metabolites and leave behind the insoluble cellular marc. The products so obtained from plants are relatively complex mixtures of metabolites, in liquid or semisolid state or (after removing the solvent) in dry powder form, and are intended for oral or external use. These include classes of preparations known as decoctions, infusions, fluid extracts, tinctures, pilular (semisolid) extracts or powdered extracts. Such preparations have been popularly called galenicals, named after Galen, the second century Greek physician. The purpose of standardized extraction procedures for crude drugs (medicinal plant parts) is to attain the therapeutically desired portions and to eliminate unwanted material by treatment with a selective solvent known as menstruum. The extract thus obtained, after standardization, may be used as medicinal agent as such in the form of tinctures or fluid extracts or further processed to be incorporated in any dosage form such as tablets and capsules. These products all contain complex mixture of many medicinal plant metabolites, such as alkaloids, glycosides, terpenoids, flavonoids and lignans. In order to be used as a modern drug, an extract may be further processed through various techniques of fractionation to isolate individual chemical entities such as vincristine, vinblastine, hyoscyamine, hyoscine, pilocarpine, forskolin and codeine.

The industrial processing of MAPs starts with the extraction of the active components using various technologies. The general techniques of medicinal plant extraction include maceration, infusion, percolation, digestion, decoction, hot continuous extraction (Soxhlet), aqueous-alcoholic extraction by fermentation, counter-current extraction, microwave-assisted extraction, ultrasound extraction (sonication), supercritical fluid extraction, and phytonic extraction (with hydrofluorocarbon solvents). For aromatic plants, hydrodistillation techniques (water distillation, steam distillation, water and steam distillation), hydrolytic maceration followed by distillation, expression and enfleurage (cold fat extraction) may be employed. Some of the latest extraction methods for aromatic plants include headspace trapping, solid phase micro-extraction, protoplast extraction, microdistillation, thermomicrodistillation and molecular distillation.

With the increasing demand for herbal medicinal products, nutraceuticals, and natural products for health care all over the world, medicinal plant extract manufacturers and essential oil producers have started using the most appropriate extraction technologies in order to produce extracts and essential oils of defined quality with the least variations from batch to batch. Such approach has to be adopted by MAP-rich developing countries in order to meet the increasing requirement of good quality extracts and essential oils for better revenue generation within the country, as well as for capturing this market in developed countries.

The basic parameters influencing the quality of an extract are the plant parts used as starting material, the solvent used for extraction, the manufacturing process (extraction technology) used with the type of equipment employed, and the crude-drug:extract ratio. The use of appropriate extraction technology, plant material, manufacturing equipment, extraction method and solvent and the adherence to good manufacturing practices certainly help to produce a good quality extract. From laboratory scale to pilot scale, all the conditions and parameters can be modelled using process simulation for successful industrial-scale production.

With the advances in extraction technologies and better knowledge for maintaining quality parameters, it has become absolutely necessary to disseminate such information to emerging and developing countries with a rich MAP biodiversity for the best industrial utilization of MAP resources.

The experts at the South-East Asian (SEA) Regional Workshop entitled ″Extraction Technologies for Medicinal and Aromatic Plants,″ held in 2006 in Lucknow, India, agreed to prepare a publication on extraction principles, technologies and analytical techniques for quality control of raw materials and processed products in the form of extracts and essential oils for medicinal and aromatic plants. This book focuses on different techniques of hydrodistillation, steam distillation, cohobation and fractional distillation of volatile oils from aromatic plants and on water-solvent extraction and supercritical fluid extraction for medicinal plant extracts. It also discusses general, specific and advanced technologies for preparing extracts of medicinal plants and the extraction of volatile oils and fragrances from aromatic plants. This book is intended to equip emerging and developing countries with techniques of extraction that can help them to produce economical and globally competitive quality extracts.

Gennaro Longo
Chief of Environment Area
Special Adviser on Technology Development

Contents

1 An Overview of Extraction Techniques for Medicinal and Aromatic Plants, *S. S. Handa* 21
- Abstract 21
- 1.1 Introduction 21
- 1.2 Medicinal Plant Extracts 22
  - 1.2.1 General Methods of Extraction of Medicinal Plants 22
    - 1.2.1.1 Maceration 22
    - 1.2.1.2 Infusion 22
    - 1.2.1.3 Digestion 22
    - 1.2.1.4 Decoction 23
    - 1.2.1.5 Percolation 23
    - 1.2.1.6 Hot Continuous Extraction (Soxhlet) 23
    - 1.2.1.7 Aqueous Alcoholic Extraction by Fermentation 24
    - 1.2.1.8 Counter-current Extraction 25
    - 1.2.1.9 Ultrasound Extraction (Sonication) 25
    - 1.2.1.10 Supercritical Fluid Extraction 25
    - 1.2.1.11 Phytonics Process 26
      - 1.2.1.11.1 Advantages of the Process 27
      - 1.2.1.11.2 Applications 28
    - 1.2.1.12 Parameters for Selecting an Appropriate Extraction Method 28
  - 1.2.2 Steps Involved in the Extraction of Medicinal Plants 29
    - 1.2.2.1 Size Reduction 29
    - 1.2.2.2 Extraction 30
      - 1.2.2.2.1 Cold Aqueous Percolation 30
      - 1.2.2.2.2 Hot Aqueous Extraction (Decoction) 30
      - 1.2.2.2.3 Filtration 30
      - 1.2.2.2.4 Spray Drying 30
    - 1.2.2.3 Solvent Extraction 31
      - 1.2.2.3.1 Cold Percolation 31
      - 1.2.2.3.2 Hot Percolation 32
      - 1.2.2.3.3 Concentration 33
- 1.3 Aromatic Plant Extracts 34
  - 1.3.1 Concrete 34
  - 1.3.2 Absolutes 35
  - 1.3.3 Resinoids 35
  - 1.3.4 Pomades 35
  - 1.3.5 Essential Oils 35
    - 1.3.5.1 Sources of Natural Essential Oils 36
    - 1.3.5.2 Essential Oil Constituents 37
    - 1.3.5.3 Methods of Producing Essential Oils 37

1.3.5.3.1 Hydrodistillation 40
1.3.5.3.1.1 Mechanism of Distillation 40
1.3.5.3.1.1.1 Hydrodiffusion 40
1.3.5.3.1.1.2 Hydrolysis 41
1.3.5.3.1.1.3 Effect of Heat 41
1.3.5.3.2 Three Types of Hydrodistillation 42
1.3.5.3.2.1 Water Distillation 42
1.3.5.3.2.1.1 Traditional Method of Producing Attar Using Hydrodistillation 43
1.3.5.3.2.1.2 Disadvantages of Water Distillation . . . 44
1.3.5.3.2.2 Water and Steam Distillation 44
1.3.5.3.2.2.1 Cohobation 45
1.3.5.3.2.2.2 Advantages of Water and Steam Distillation over Water Distillation . . 46
1.3.5.3.2.2.3 Disadvantages of Water and Steam Distillation 46
1.3.5.3.2.3 Direct Steam Distillation 46
1.3.5.3.2.3.1 Advantages of Direct Steam Distillation 47
1.3.5.3.2.3.2 Disadvantage of Direct Steam Distillation 47
1.3.5.3.3 Essential Oil Extraction by Hydrolytic Maceration Distillation 47
1.3.5.3.4 Essential Oil Extraction by Expression 47
1.3.5.3.4.1 Pelatrice Process 48
1.3.5.3.4.2 Sfumatrice Process 48
1.3.5.3.5 Essential Oil Extraction with Cold Fat (Enfleurage) . . . 48
1.3.5.3.5.1 Enfleurage and Defleurage 50
1.3.5.3.5.2 Hot Maceration Process 51
1.3.6 Modern (Non-traditional) Methods of Extraction of Essential Oils 51
1.4 Conclusions 52
Bibliography 52

2 Role of Process Simulation to Extraction Technologies for Medicinal and Aromatic Plants, *M. Fermeglia* 55
Abstract 55
2.1 Introduction 55
2.2 Process Simulation Goals and Definitions 56
2.3 Biotechnological and Phytochemical Processes Studied at ICS-UNIDO 59
2.3.1 Brief of Biotechnological Processes 59
2.3.1.1 Alcohol Production from Biomass 59
2.3.1.2 Soybean Oil Refinery Process and Treatment of the Waste 59
2.3.1.3 Production of Synthetic Hydrocarbon Fuels Starting from Biomass 60
2.3.1.4 Production of Bio-ethanol from Corn 60
2.3.2 Brief of Phytochemical Processes 60
2.3.2.1 Citral Recovery by Distillation of Lemon Peel Oil 60
2.3.2.2 Menthol Recovery by Crystallization of Mentha Oil 60
2.3.2.3 Carvone Recovery from Spearmint Oil 61
2.3.2.4 Peppermint Oil Extraction by Steam Distillation 61
2.3.2.5 Multiple-effect Evaporation of Milk Serum 61
2.3.3 Case Study: Turpentine Oil Batch Distillation 61
2.3.4 Case Study: Menthol Recovery by Crystallization of Mentha Oil 63
2.4 Conclusions 65
Bibliography 66
3 Maceration, Percolation and Infusion Techniques for the Extraction of Medicinal and Aromatic Plants, *J. Singh* 67
Abstract 67
3.1 Introduction 67
3.2 General Principles and Mechanism Involved for Crude Drug Extraction by Maceration, Percolation and Infusion 68
3.3 Factors Affecting Choice of Extraction Process 69
3.3.1 Nature of the Crude Drug 69
3.3.2 Stability of the Crude Drug 69
3.3.3 Cost of the Crude Drug 69
3.3.4 Solvent 69
3.3.5 Concentration of the Product 69
3.3.6 Recovery of Solvent 69
3.4 Quality Assurance of Extracts: The Extraction Process and Solvent 70
3.5 Maceration Processes (Steady-state Extraction) 70
3.5.1 General Procedure 70
3.5.2 Maceration Process for Organized and Unorganized Crude Drugs 71
3.5.3 Modifications to the General Processes of Maceration 72
3.5.4 Large-scale Extraction Procedures 72
3.5.4.1 Circulatory Extraction 73

3.5.4.2 Multistage Extraction 73
3.5.4.2.1 Advantages 74
3.5.4.2.2 Procedure 74
3.5.4.2.3 Extraction Battery 74
3.5.4.2.4 Equipment 74
3.6 Percolation (Exhaustive Extraction) 75
3.6.1 General Process of Percolation 75
3.6.2 Modifications to the General Process of Percolation 77
3.6.2.1 Reserved Percolation 77
3.6.2.2 Cover and Run Down Method 77
3.6.3 Percolators 78
3.6.3.1 Small-scale or Laboratory-scale Extraction 78
3.6.3.1.1 Soxhlet Apparatus 78
3.6.3.1.2 Official Extractor 79
3.6.3.2 Limitations of the Official Extractor 79
3.6.3.3 Large-scale Extractor 80
3.7 Infusion 81
3.7.1 General Considerations 81
3.7.2 General Method for Preparing Fresh Infusions 81
3.7.3 Preparation of Concentrated Infusions 81
3.8 Evaporation 81
3.9 Conclusions 82
Bibliography 82
4 Hydrolytic Maceration, Expression and Cold Fat Extraction, *A. K. Singh* 83
Abstract 83
4.1 Introduction 83
4.2 Hydrolytic Maceration 83
4.3 Expression Extraction of Essential Oils 84
4.3.1 Process of Expression 85
4.3.1.1 Hand Process 85
4.3.1.2 Ecuelle Process 86
4.3.1.3 Hand Machine 86
4.3.1.4 Sfumatrici and Pelatrici 86
4.3.1.4.1 Special Sfumatrice 86
4.3.1.5 Modern Machines 87
4.3.1.6 FMC Whole Fruit Extractor 87
4.4 Cold Fat Extraction (Enfleurage) 88
4.5 Conclusions 91
Bibliography 92
5 Decoction and Hot Continuous Extraction Techniques, *S. Tandon and S. Rane* 93
Abstract 93
5.1 Introduction 93

5.2 Solid Liquid Extraction Process 93
5.3 Process Parameters Affecting Solid-liquid Extraction 94
  5.3.1 Post-harvest Processing 94
  5.3.2 Matrix Characteristics 94
  5.3.3 Choice of Solvent 95
  5.3.4 Condition of Extraction 95
5.4 Method of Solid-liquid Extraction 95
5.5 Solid-liquid Extraction Equipment 97
  5.5.1 Continuous Extraction 98
    5.5.1.1 Percolation 98
    5.5.1.2 Immersion 98
  5.5.2 Continuous Extraction Equipment 98
    5.5.2.1 Continuous Horizontal Extractor 98
    5.5.2.2 Hildebrandt Extractor 99
    5.5.2.3 Bonotto Extractor 100
    5.5.2.4 Bollmann Extractor 100
    5.5.2.5 Kennedy Extractor 101
5.6 Conventional Solvent Extraction 102
  5.6.1 Principles and Mechanisms 102
  5.6.2 Advantages and Disadvantages of Soxhlet Extraction 103
5.7 Accelerated Solvent Extraction 104
  5.7.1 Principles and Mechanisms 104
  5.7.2 Advantages and Disadvantages of Accelerated Solvent Extraction 104
5.8 Important Factors for Designing a Solvent Extraction Plant for Medicinal Plants ... 104
5.9 Conclusions 105
Bibliography 106

6 **Aqueous Alcoholic Extraction of Medicinal and Aromatic Plants by Fermentation,** *C. K. Katiyar* . 107
Abstract 107
6.1 Introduction 107
6.2 Ayurvedic Dosage Forms 107
  6.2.1 Swarasa (Fresh Juice) 108
  6.2.2 Kalka (Wet Bolus) 108
  6.2.3 Kwatha (Decoction) 108
  6.2.4 Hima (Cold Infusion) 108
  6.2.5 Phanta (Hot Infusion) 108
  6.2.6 Solids 108
  6.2.7 Semisolids 109
  6.2.8 Liquids 109
  6.2.9 Fumes 109
6.3 Shelf Life of Dosage Forms 109
6.4 Asava and Arishta: Self-fermented Products 109
  6.4.1 Self-fermentation Process for Preparing Asava Arishta 110

6.4.2 Merits of the Process ... 111
6.5 Application of Asava Arishta Technology in New Drug Discovery ... 112
6.6 Conclusions ... 112
Bibliography ... 113
7 **Distillation Technology for Essential Oils,** *S. Tandon* ... 115
Abstract ... 115
7.1 Introduction ... 115
7.2 Principles of Distillation ... 115
7.3 Methods for Distillation ... 116
7.3.1 Hydrodistillation ... 116
7.3.2 Water and Steam Distillation ... 118
7.3.2.1 Improved Field Distillation Units ... 119
7.3.3 Direct Steam Distillation ... 120
7.3.3.1 Comparison of Boiler-operated Unit with Directly Fired Type ... 121
7.3.4 Distillation with Cohobation ... 121
7.4 Hydrodiffusion ... 123
7.5 Parameters Affecting Yield and Quality of Essential Oils ... 123
7.5.1 Mode of Distillation ... 123
7.5.2 Proper Design of Equipment ... 124
7.5.3 Material of Fabrication of Equipment ... 124
7.5.4 Condition of Raw Material ... 124
7.5.5 Time for Distillation ... 124
7.5.6 Loading of Raw Material and Steam Distribution ... 125
7.5.7 Operating Parameters ... 125
7.5.8 Condition of Tank and Equipment ... 125
7.6 Purification of Crude Essential Oils ... 125
7.6.1 Continuous Steam Distillation ... 126
7.7 Conclusions ... 126
Bibliography ... 127
8 **Microdistillation, Thermomicrodistillation and Molecular Distillation Techniques,** *V. G. Pangarkar* ... 129
Abstract ... 129
8.1 Introduction ... 129
8.2 Process Intensification ... 130
8.2.1 Multifunctional Equipment ... 130
8.2.2 Process-intensifying Equipment ... 130
8.3 Solvent Extractions of MAPs ... 131
8.3.1 Thermodynamics of Solvent Extraction and Choice of Solvent ... 131
8.3.2 Solid-liquid Mass Transfer ... 132
8.3.3 Microwave-assisted Extraction ... 133
8.3.3.1 Principle of Microwave Heating ... 133
8.3.3.2 Mechanism of MAE ... 133
8.3.3.3 Literature on MAE ... 135

8.3.3.4 Industrial-scale MAE 135
8.4 Microwave-assisted Hydrodistillation 136
8.5 Molecular Distillation or Short Path Distillation 137
8.5.1 Principle of MD 137
8.5.2 Advantages of MD 137
8.5.3 Separation Efficiency of MD 137
8.5.4 Parameters that Affect the MD Process 138
8.5.5 Typical Applications of MD 138
8.6 Recovery of Dissolved Essential Oils from Steam Distillation Condensates 138
8.6.1 Polymeric Adsorption Process 139
8.6.2 Pervaporation Process 140
8.7 Conclusions 141
Bibliography 141
9 **Solid Phase Micro-extraction and Headspace Trapping Extraction**, *R. Harlalka* 145
Abstract 145
9.1 Introduction 145
9.2 The SPME Device 146
9.3 Calibration, Optimization, Precision and Suitability of SPME 149
9.3.1 Selection of Fiber Coating 149
9.3.2 Selection of the Extraction Mode 149
9.3.3 Selection of the Agitation Technique 150
9.3.4 Selection of Separation or Detection Technique 150
9.3.5 Optimization of Desorption Conditions 150
9.3.6 Optimization of Sample Volume 151
9.3.7 Determination of the Extraction Time 151
9.3.8 Optimization of Extraction Conditions 151
9.3.9 Determination of the Linear Dynamic Range of the Method 152
9.3.10 Selection of the Calibration Method 152
9.3.11 Precision of the Method 152
9.3.12 Suitability 153
9.4 Headspace Trapping Extraction and GC-FID/MS Analysis 153
9.4.1 History of Headspace 154
9.4.2 The Aura 154
9.4.3 What is Diffusivity? 154
9.4.4 Application of Headspace Trapping 154
9.4.4.1 Jasmine 155
9.4.4.2 Yellow Tea Rose 155
9.4.4.3 Passion Flower 156
9.4.4.4 Lotus 156
9.4.4.5 Lavender 156
9.4.4.6 Chamomile 157
9.4.4.7 Shefali 157

9.4.4.8 Spearmint 157
9.4.4.9 Cinnamon Bark 158
9.4.4.10 Ginger 158
9.4.4.11 Peach 158
9.4.4.12 Pineapple 159
9.4.5 Classical Perfumes 159
9.4.6 Need for Headspace 160
9.5 Types of Headspace Trapping 160
9.5.1 Static Headspace Trapping 160
9.5.2 Dynamic Headspace Trapping 161
9.6 Principles of Static Headspace-GC Systems 162
9.6.1 Trace Analysis by HS-GC 163
9.7 Headspace Trapping Techniques 163
9.7.1 Static Headspace Trapping 163
9.7.2 Dynamic Headspace Trapping 164
9.7.3 Recovering the Adsorbed Volatiles by Thermal or Liquid Solvent Desorption 164
9.7.4 Some Practical Examples Using Headspace Technique Use 165
9.7.4.1 Tomato Juice 165
9.7.4.1.1 Preparation of Traps 166
9.7.4.1.2 Thermal Desorption 166
9.7.4.2 Headspace of Hedychium coronium 166
9.7.4.3 Volatiles of White Hyacinths Isolated by Dynamic Headspace Trapping 166
9.7.4.4 Medical Materials Testing by Headspace Trap-GC/MS 166
9.8 Conclusions 167
Bibliography 167

10 Supercritical Fluid Extraction of Medicinal and Aromatic Plants: Fundamentals and Applications, *A. Bertucco and G. Franceschin* 169
Abstract 169
10.1 Introduction 169
10.2 Supercritical Fluids 169
10.3 SFE Processes 172
10.4 The SFE Process and Equipment Development 175
10.5 SFE Applied to Medicinal and Aromatic Plants 177
10.6 Conclusions 179
Bibliography 180
11 Process-scale High Performance Liquid Chromatography for Medicinal and Aromatic Plants, *M. M. Gupta and K. Shanker* 181
Abstract 181
11.1 Introduction 181
11.2 Theoretical Aspects of HPLC 182

11.2.1 Chromatography Classification 182
11.2.2 Important Factors that Influence HPLC Separation 183
11.2.3 Main Components of HPLC 184
11.2.4 HPLC Classification 185
11.2.5 Advantages of HPLC 186
11.3 Preparative HPLC 186
11.3.1 Strategy for Preparative Separation 186
11.4 Practical Consideration in Preparative HPLC Scale-up 187
11.4.1 Sample Loading 187
11.4.2 Separation Time 188
11.4.3 Solvent Composition 188
11.4.4 Washing Steps 188
11.4.5 Recycling 188
11.5 Stepwise Operations in Process-scale HPLC 189
11.6 Problems Encounter in Preparative Scale-up 189
11.6.1 Purity of Crude Extract 189
11.6.2 Removal of Chromatographic Solvent 190
11.6.3 Temperature Variation from Laboratory to Pilot Scale 190
11.6.4 Increase in Pump Pressure Due to Accumulation of Impurities on the Column 190
11.7 Summary: Scale-up Strategy 190
11.8 Applications: Natural Products Isolation 191
11.9 Conclusions 193
Bibliography 193
12 Flash Chromatography and Low Pressure Chromatographic Techniques for Separation of Phytomolecules, *S. K. Chattopadhyay* 195
Abstract 195
12.1 Introduction 195
12.2 Flash Chromatography 196
12.2.1 Theory of Flash Chromatography 196
12.2.2 Converting TLC to Flash Chromatography 196
12.3 Isocratic versus Gradient Chromatography 199
12.4 Adsorbent Selection and Mode of Separation 200
12.4.1 Isolute Flash Columns 200
12.4.2 Method Development Using Isolute Flash Columns 200
12.4.2.1 Column Equilibration 200
12.4.2.2 Typical Equilibration Solvents 200
12.5 Sample Application 201
12.5.1 Wet Loading 201
12.5.1.1 Practical Tips for Wet Loading 201
12.5.2 Dry Loading 202
12.5.2.1 Practical Tips for Dry Loading 202

12.6 Elution 203
  12.6.1 Step Gradient Elution 203
  12.6.2 Linear Gradient Elution 203
  12.6.3 Method Development using Gradient Elution 204
  12.6.4 Practical Tips for Gradient Elution 204
  12.6.5 Optimizing Flow Rate 205
12.7 Fraction Collection 205
  12.7.1 Off-line Flash Chromatography 205
  12.7.2 On-line Flash Chromatography 206
12.8 Low Pressure Liquid Chromatography 206
  12.8.1 Gel Filtration Chromatography 206
  12.8.2 Ion Exchange Chromatography 206
  12.8.3 Affinity Chromatography 207
12.9 Conclusions 207
Bibliography 208
13 Counter-current Chromatography, *S. K. Srivastava* 209
Abstract 209
13.1 Introduction 209
13.2 Principles and Development of Counter-current Chromatography 210
  13.2.1 Liquid-liquid Extraction 210
  13.2.2 Partition Coefficient 211
  13.2.3 Droplet Counter-current Chromatography 212
  13.2.4 Applications of Droplet Counter-current Chromatography 214
  13.2.5 Limitations of DCCC 214
  13.2.6 Modern Counter-current Chromatography 214
13.3 HSCCC Instrument and Mechanism 215
13.4 CPC Instrument and Setup 218
  13.4.1 Various Kinds of CPC Instruments 218
13.5 How to Achieve Nice Separation of Various Kinds of Natural Products Using HSCCC and CPC 219
  13.5.1 Search for a Suitable Solvent System 219
    13.5.1.1 Solvent Systems for the Separation of a Large Variety of Natural Products 220
    13.5.1.2 Retention of the Stationary Phase 223
    13.5.1.3 Preparation of Sample Solution 224
    13.5.1.4 Separation Column 224
    13.5.1.5 Choice of the Mobile Phase 224
    13.5.1.6 Flow rate of the Mobile Phase 224
    13.5.1.7 Revolution Speed 225
    13.5.1.8 Filling the Column with the Stationary Phase 225
    13.5.1.9 Sample Loading 225
    13.5.1.10 On-line Monitoring of Effluent 226

13.5.1.11 Measurement of Stationary Phase Retention 226
13.5.2 Applications of HSCCC-CPC Technologies in Natural Products Isolation.... 227
13.5.2.1 Purification of Coenzyme Q_{10} from Fermentation Extract: HSCCC versus Silica Gel Column Chromatography 228
13.5.2.2 Preparative Separation of Gambogic Acid and its C-2 Epimer by HPCCC 228
13.5.2.3 Separation and Purification of 10-Deacetylbaccatin III by HSCCC 228
13.5.2.4 Large-scale Separation of Resveratrol and Anthraglycoside A and B from Polygonum cuspidatum by HSCCC 230
13.5.2.5 Separation of Andrographolide and Neoandrographolide from the Leaves of Andrographis paniculata using HSCCC 230
13.5.2.6 Separation of WAP-8294A Components, a Novel Anti-methicillin-resistant Staphylococcus aureus Antibiotic, using HSCCC....... 231
13.5.2.7 Other Examples of Separation of Phytoconstituents by CCC..... 232
13.6 Advantages of CCC.. 232
13.6.1 Advantages of HSCCC-CPC Technologies over HPLC 233
13.7 Manufacturers of CCC Instruments... 233
13.7.1 Manufacturers of HSCCC Machines 233
13.7.2 Manufacturers of CPC Machines.. 233
13.8 Selected Reviews on CCC.. 234
13.9 Conclusions 234
Bibliography 235
14 Quality Control of Medicinal and Aromatic Plants and their Extracted Products by HPLC and High Performance Thin Layer Chromatography, *K. Vasisht* 239
Abstract.. 239
14.1 Introduction.. 239
14.2 Quality Control of Medicinal Plants and their Products 240
14.3 Biological and Chemical Standardization of Drugs.................................. 242
14.3.1 Chemical Standardization and Markers.. 242
14.3.2 Analytical Techniques for Quantifying a Marker.............................. 243
14.3.3 Validation of Analytical Procedures.. 243
14.3.3.1 Specificity.. 244
14.3.3.2 Linearity 244
14.3.3.3 Range.. 244
14.3.3.4 Accuracy 245
14.3.3.5 Precision... 245
14.3.3.5.1 Repeatability 246
14.3.3.5.2 Intermediate Precision...................................... 246
14.3.3.5.3 Reproducibility 246
14.3.3.6 Detection Limit.. 246
14.3.3.7 Quantitation Limit.. 246

14.3.3.8 Robustness 246
14.4 Thin Layer Chromatography in Quality Control of Plant Products 247
  14.4.1 Sample Preparation 247
  14.4.2 Selection of Chromatographic Layer 248
  14.4.3 TLC versus HPTLC Layers 249
  14.4.4 Selection of the Mobile Phase 249
  14.4.5 Application of Sample 249
  14.4.6 Developing the Chromatogram 250
  14.4.7 Drying the Plate 251
  14.4.8 Derivatization 251
  14.4.9 Evaluation of the Chromatograms 251
  14.4.10 Improving the Efficiency of TLC 252
14.5 High Performance Liquid Chromatography 254
  14.5.1 Pumps 255
  14.5.2 Injector 255
  14.5.3 Columns 256
  14.5.4 Detectors 256
  14.5.5 Data Processing 256
  14.5.6 Factors Affecting the HPLC Analysis 256
  14.5.7 HPLC in Quality Control of Plant Products 257
14.6 TLC versus HPLC 258
14.7 Conclusions 259
Bibliography 259

1 An Overview of Extraction Techniques for Medicinal and Aromatic Plants

S. S. Handa

Abstract

A wide range of technologies is available for the extraction of active components and essential oils from medicinal and aromatic plants. The choice depends on the economic feasibility and suitability of the process to the particular situation. The various processes of production of medicinal plant extracts and essential oils are reviewed in this paper.

1.1 Introduction

Asia is the largest continent and has 60% of the world's population. It has abundant medicinal and aromatic plant species, well documented traditional knowledge, a long-standing practice of traditional medicine, and the potential for social and economic development of medicinal and aromatic plants (MAPs). Asia is one of the largest biodiversity regions in the world, containing some of the richest countries in plant resources. The continent has diverse plant flora but species richness is concentrated mainly in tropical and subtropical regions. Six of the world's 18 biodiversity hot spots, namely eastern Himalaya, North Borneo, Peninsular Malaysia, Sri Lanka, Philippines and the Western Ghats of South India, lie in Asia. The countries of the region have large flora: China has 30,000 species of higher plants; Indonesia, 20,000; India, 17,000; Myanmar, 14,000; Malaysia, 12,000; and Thailand, 12,000. The total numbers of plant species and the endemics in the region are given below:

Region	Species	Endemics
South East Asia	42-50,000	40,000
China and East Asia	45,000	18,650
Indian Subcontinent	25,000	12,000
South West Asia	23,000	7,100

Sustainable industrial exploitation of such a valuable bioresource, through use of appropriate technologies, can substantially contribute to the socio-economic growth of Asian countries. The International Centre for Science and High Technology (ICS-UNIDO) has thus organized this regional workshop on “extraction technologies for medicinal and aromatic plants” for South East Asian countries.

1.2 Medicinal Plant Extracts

Extraction, as the term is used pharmaceutically, involves the separation of medicinally active portions of plant or animal tissues from the inactive or inert components by using selective solvents in standard extraction procedures. The products so obtained from plants are relatively impure liquids, semisolids or powders intended only for oral or external use. These include classes of preparations known as decoctions, infusions, fluid extracts, tinctures, pilular (semisolid) extracts and powdered extracts. Such preparations popularly have been called galenicals, named after Galen, the second century Greek physician.

The purposes of standardized extraction procedures for crude drugs are to attain the therapeutically desired portion and to eliminate the inert material by treatment with a selective solvent known as *menstruum*. The extract thus obtained may be ready for use as a medicinal agent in the form of tinctures and fluid extracts, it may be further processed to be incorporated in any dosage form such as tablets or capsules, or it may be fractionated to isolate individual chemical entities such as ajmalicine, hyoscine and vincristine, which are modem drugs. Thus, standardization of extraction procedures contributes significantly to the final quality of the herbal drug.

1.2.1 General Methods of Extraction of Medicinal Plants

1.2.1.1 Maceration

In this process, the whole or coarsely powdered crude drug is placed in a stoppered container with the solvent and allowed to stand at room temperature for a period of at least 3 days with frequent agitation until the soluble matter has dissolved. The mixture then is strained, the marc (the damp solid material) is pressed, and the combined liquids are clarified by filtration or decantation after standing.

1.2.1.2 Infusion

Fresh infusions are prepared by macerating the crude drug for a short period of time with cold or boiling water. These are dilute solutions of the readily soluble constituents of crude drugs.

1.2.1.3 Digestion

This is a form of maceration in which gentle heat is used during the process of extraction. It is used when moderately elevated temperature is not objectionable. The solvent efficiency of the menstruum is thereby increased.

1.2.1.4 Decoction

In this process, the crude drug is boiled in a specified volume of water for a defined time; it is then cooled and strained or filtered. This procedure is suitable for extracting water-soluble, heat-stable constituents. This process is typically used in preparation of Ayurvedic extracts called "quath" or "kawath". The starting ratio of crude drug to water is fixed, e.g. 1:4 or 1:16; the volume is then brought down to one-fourth its original volume by boiling during the extraction procedure. Then, the concentrated extract is filtered and used as such or processed further.

1.2.1.5 Percolation

This is the procedure used most frequently to extract active ingredients in the preparation of tinctures and fluid extracts. A percolator (a narrow, cone-shaped vessel open at both ends) is generally used (Figure 1). The solid ingredients are moistened with an appropriate amount of the specified menstruum and allowed to stand for approximately 4 h in a well-closed container, after which the mass is packed and the top of the percolator is closed. Additional menstruum is added to form a shallow layer above the mass, and the mixture is allowed to macerate in the closed percolator for 24 h. The outlet of the percolator then is opened and the liquid contained therein is allowed to drip slowly. Additional menstruum is added as required, until the percolate measures about three-quarters of the required volume of the finished product. The marc is then pressed and the expressed liquid is added to the percolate. Sufficient menstruum is added to produce the required volume, and the mixed liquid is clarified by filtration or by standing followed by decanting.

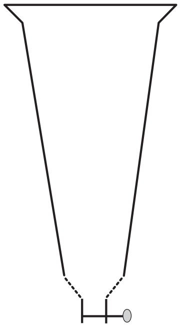

Figure 1: Percolator

1.2.1.6 Hot Continuous Extraction (Soxhlet)

In this method, the finely ground crude drug is placed in a porous bag or "thimble" made of strong filter paper, which is placed in chamber E of the Soxhlet apparatus (Figure 2). The extracting solvent in flask A is heated, and its vapors condense in condenser D. The condensed extractant drips into the thimble containing the crude drug, and extracts it by contact. When the level of liquid in chamber E rises to the top of siphon tube C, the liquid contents of chamber E siphon into flask A. This process is continuous and is carried out until a drop of solvent from the siphon tube does not leave residue when evaporated. The advantage of this method, compared to previously described methods, is that large amounts of drug can be extracted with a much smaller quantity of solvent. This effects tremendous economy in terms of time, energy and consequently financial inputs. At small scale, it is employed as a batch process only, but it becomes much more economical and viable when converted into a continuous extraction procedure on medium or large scale.

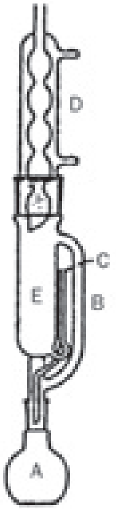

Figure 2: Soxhlet apparatus

1.2.1.7 Aqueous Alcoholic Extraction by Fermentation

Some medicinal preparations of Ayurveda (like *asava* and *arista*) adopt the technique of fermentation for extracting the active principles. The extraction procedure involves soaking the crude drug, in the form of either a powder or a decoction (*kasaya*), for a specified period of time, during which it undergoes fermentation and generates alcohol in situ; this facilitates the extraction of the active constituents contained in the plant material. The alcohol thus generated also serves as a preservative. If the fermentation is to be carried out in an earthen vessel, it should not be new: water should first be boiled in the vessel. In large-scale manufacture, wooden vats, porcelain jars or metal vessels are used in place of earthen vessels. Some examples of such preparations are *karpurasava, kanakasava, dasmularista*. In Ayurveda, this method is not yet standardized but, with the extraordinarily high degree of advancement in fermentation technology, it should not be difficult to standardize this technique of extraction for the production of herbal drug extracts.

1.2.1.8 Counter-current Extraction

In counter-current extraction (CCE), wet raw material is pulverized using toothed disc disintegrators to produce a fine slurry. In this process, the material to be extracted is moved in one direction (generally in the form of a fine slurry) within a cylindrical extractor where it comes in contact with extraction solvent. The further the starting material moves, the more concentrated the extract becomes. Complete extraction is thus possible when the quantities of solvent and material and their flow rates are optimized. The process is highly efficient, requiring little time and posing no risk from high temperature. Finally, sufficiently concentrated extract comes out at one end of the extractor while the marc (practically free of visible solvent) falls out from the other end.

This extraction process has significant advantages:

i) A unit quantity of the plant material can be extracted with much smaller volume of solvent as compared to other methods like maceration, decoction, percolation.
ii) CCE is commonly done at room temperature, which spares the thermolabile constituents from exposure to heat which is employed in most other techniques.
iii) As the pulverization of the drug is done under wet conditions, the heat generated during comminution is neutralized by water. This again spares the thermolabile constituents from exposure to heat.
iv) The extraction procedure has been rated to be more efficient and effective than continuous hot extraction.

1.2.1.9 Ultrasound Extraction (Sonication)

The procedure involves the use of ultrasound with frequencies ranging from 20 kHz to 2000 kHz; this increases the permeability of cell walls and produces cavitation. Although the process is useful in some cases, like extraction of rauwolfia root, its large-scale application is limited due to the higher costs. One disadvantage of the procedure is the occasional but known deleterious effect of ultrasound energy (more than 20 kHz) on the active constituents of medicinal plants through formation of free radicals and consequently undesirable changes in the drug molecules.

1.2.1.10 Supercritical Fluid Extraction

Supercritical fluid extraction (SFE) is an alternative sample preparation method with general goals of reduced use of organic solvents and increased sample throughput. The factors to consider include temperature, pressure, sample volume, analyte collection, modifier (cosolvent) addition, flow and pressure control, and restrictors. Generally, cylindrical extraction vessels are used for SFE and their performance is good beyond any doubt. The collection of the extracted analyte following SFE is another important step: significant analyte loss can occur during this step, leading the analyst to believe that the actual efficiency was poor.

There are many advantages to the use of CO_2 as the extracting fluid. In addition to its favorable physical properties, carbon dioxide is inexpensive, safe and abundant. But while carbon dioxide is the preferred fluid for SFE, it possesses several polarity limitations. Solvent polarity is important when extracting polar solutes and when strong analyte-matrix interactions are present. Organic solvents are frequently added to the carbon dioxide extracting fluid to alleviate the polarity limitations. Of late, instead of carbon dioxide, argon is being used because it is inexpensive and more inert. The component recovery rates generally increase with increasing pressure or temperature: the highest recovery rates in case of argon are obtained at 500 atm and 150° C.

The extraction procedure possesses distinct advantages:

i) The extraction of constituents at low temperature, which strictly avoids damage from heat and some organic solvents.
ii) No solvent residues.
iii) Environmentally friendly extraction procedure.

The largest area of growth in the development of SFE has been the rapid expansion of its applications. SFE finds extensive application in the extraction of pesticides, environmental samples, foods and fragrances, essential oils, polymers and natural products. The major deterrent in the commercial application of the extraction process is its prohibitive capital investment.

1.2.1.11 Phytonics Process

A new solvent based on hydrofluorocarbon-134a and a new technology to optimize its remarkable properties in the extraction of plant materials offer significant environmental advantages and health and safety benefits over traditional processes for the production of high quality natural fragrant oils, flavors and biological extracts.

Advanced Phytonics Limited (Manchester, UK) has developed this patented technology termed “phytonics process”. The products mostly extracted by this process are fragrant components of essential oils and biological or phytopharmacological extracts which can be used directly without further physical or chemical treatment.

The properties of the new generation of fluorocarbon solvents have been applied to the extraction of plant materials. The core of the solvent is 1,1,2,2-tetrafluoroethane, better known as hydrofluorocarbon-134a (HFC-134a). This product was developed as a replacement for chlorofluorocarbons. The boiling point of this solvent is -25° C. It is not flammable or toxic. Unlike chlorofluorocarbons, it does not deplete the ozone layer. It has a vapor pressure of 5.6 bar at ambient temperature. By most standards this is a poor solvent. For example, it does not mix with mineral oils or triglycerides and it does not dissolve plant wastes.

The process is advantageous in that the solvents can be customized: by using modified solvents with HFC-134a, the process can be made highly selective in extracting a specific class of phytoconstituents. Similarly, other modified solvents can be used to extract a broader spectrum of components. The biological products made by this process have extremely low residual solvent. The residuals are invariably less than 20 parts per billion and are frequently below levels of detection. These solvents are neither acidic nor alkaline and, therefore, have only minimal potential reaction effects on the botanical materials. The processing plant is totally sealed so that the solvents are continually recycled and fully recovered at the end of each production cycle. The only utility needed to operate these systems is electricity and, even then, they do no consume much energy. There is no scope for the escape of the solvents. Even if some solvents do escape, they contain no chlorine and therefore pose no threat to the ozone layer. The waste biomass from these plants is dry and “ecofriendly” to handle.

1.2.1.11.1 Advantages of the Process

- Unlike other processes that employ high temperatures, the phytonics process is cool and gentle and its products are never damaged by exposure to temperatures in excess of ambient.
- No vacuum stripping is needed which, in other processes, leads to the loss of precious volatiles.
- The process is carried out entirely at neutral pH and, in the absence of oxygen, the products never suffer acid hydrolysis damage or oxidation.
- The technique is highly selective, offering a choice of operating conditions and hence a choice of end products.
- It is less threatening to the environment.
- It requires a minimum amount of electrical energy.
- It releases no harmful emissions into the atmosphere and the resultant waste products (spent biomass) are innocuous and pose no effluent disposal problems.
- The solvents used in the technique are not flammable, toxic or ozone depleting.
- The solvents are completely recycled within the system.

1.2.1.11.2 Applications

The phytonics process can be used for extraction in biotechnology (e.g for the production of antibiotics), in the herbal drug industry, in the food, essential oil and flavor industries, and in the production of other pharmacologically active products. In particular, it is used in the production of top-quality pharmaceutical-grade extracts, pharmacologically active intermediates, antibiotic extracts and phytopharmaceuticals. However, the fact that it is used in all these areas in no way prevents its use in other areas. The technique is being used in the extraction of high-quality essential oils, oleoresins, natural food colors, flavors and aromatic oils from all manner of plant materials. The technique is also used in refining crude products obtained from other extraction processes. It provides extraction without waxes or other contaminants. It helps remove many biocides from contaminated biomass.

1.2.1.12 Parameters for Selecting an Appropriate Extraction Method

i) Authentication of plant material should be done before performing extraction. Any foreign matter should be completely eliminated.
ii) Use the right plant part and, for quality control purposes, record the age of plant and the time, season and place of collection.
iii) Conditions used for drying the plant material largely depend on the nature of its chemical constituents. Hot or cold blowing air flow for drying is generally preferred. If a crude drug with high moisture content is to be used for extraction, suitable weight corrections should be incorporated.
iv) Grinding methods should be specified and techniques that generate heat should be avoided as much as possible.
v) Powdered plant material should be passed through suitable sieves to get the required particles of uniform size.
vi) Nature of constituents:
  a) If the therapeutic value lies in non-polar constituents, a non-polar solvent may be used. For example, lupeol is the active constituent of *Crataeva nurvala* and, for its extraction, hexane is generally used. Likewise, for plants like *Bacopa monnieri* and *Centella asiatica*, the active constituents are glycosides and hence a polar solvent like aqueous methanol may be used.
  b) If the constituents are thermolabile, extraction methods like cold maceration, percolation and CCE are preferred. For thermostable constituents, Soxhlet extraction (if non-aqueous solvents are used) and decoction (if water is the menstruum) are useful.
  c) Suitable precautions should be taken when dealing with constituents that degrade while being kept in organic solvents, e.g. flavonoids and phenyl propanoids.

d) In case of hot extraction, higher than required temperature should be avoided. Some glycosides are likely to break upon continuous exposure to higher temperature.
e) Standardization of time of extraction is important, as:
  - Insufficient time means incomplete extraction.
  - If the extraction time is longer, unwanted constituents may also be extracted. For example, if tea is boiled for too long, tannins are extracted which impart astringency to the final preparation.
f) The number of extractions required for complete extraction is as important as the duration of each extraction.

vii) The quality of water or menstruum used should be specified and controlled.
viii) Concentration and drying procedures should ensure the safety and stability of the active constituents. Drying under reduced pressure (e.g. using a Rotavapor) is widely used. Lyophilization, although expensive, is increasingly employed.
ix) The design and material of fabrication of the extractor are also to be taken into consideration.
x) Analytical parameters of the final extract, such as TLC and HPLC fingerprints, should be documented to monitor the quality of different batches of the extracts.

1.2.2 Steps Involved in the Extraction of Medicinal Plants

In order to extract medicinal ingredients from plant material, the following sequential steps are involved:

1. Size reduction
2. Extraction
3. Filtration
4. Concentration
5. Drying

1.2.2.1 Size Reduction

The dried plant material is disintegrated by feeding it into a hammer mill or a disc pulverizer which has built-in sieves. The particle size is controlled by varying the speed of the rotor clearance between the hammers and the lining of the grinder and also by varying the opening of the discharge of the mill. Usually, the plant material is reduced to a size between 30 and 40 mesh, but this can be changed if the need arises. The objective for powdering the plant material is to rupture its organ, tissue and cell structures so that its medicinal ingredients are exposed to the extraction solvent. Furthermore, size reduction maximizes the surface area, which in turn enhances the mass transfer of active principle from plant material to the solvent. The 30-40 mesh size is optimal, while smaller particles may become slimy during extraction and create difficulty during filtration.

1.2.2.2 Extraction

Extraction of the plant material is carried out in three ways:

i) Cold aqueous percolation
ii) Hot aqueous extraction (decoction)
iii) Solvent extraction (cold or hot)

1.2.2.2.1 Cold Aqueous Percolation

The powdered material is macerated with water and then poured into a tall column. Cold water is added until the powdered material is completely immersed. It is allowed to stand for 24 h so that water-soluble ingredients attain equilibrium in the water. The enriched aqueous extract is concentrated in multiple-effect evaporators to a particular concentration. Some diluents and excipients are added to this concentrated extract, which is then ready for medicinal use.

1.2.2.2.2 Hot Aqueous Extraction (Decoction)

This is done in an open-type extractor. The extractor is a cylindrical vessel made from type 316 stainless steel and has a diameter (D) greater than the height (H), i.e. the H/D ratio is approximately 0.5. The bottom of the vessel is welded to the dished end and is provided with an inside false bottom with a filter cloth. The outside vessel has a steam jacket and a discharge valve at the bottom.

One part powdered plant material and sixteen parts demineralized water are fed into the extractor. Heating is done by injecting steam into the jacket. The material is allowed to boil until the volume of water is reduced to one-fourth its original volume. By this time the medicinal ingredients present in the plant material have been extracted out.

1.2.2.2.3 Filtration

The extract so obtained is separated out from the marc (exhausted plant material) by allowing it to trickle into a holding tank through the built-in false bottom of the extractor, which is covered with a filter cloth. The marc is retained at the false bottom, and the extract is received in the holding tank. From the holding tank, the extract is pumped into a sparkler filter to remove fine or colloidal particles from the extract.

1.2.2.2.4 Spray Drying

The filtered extract is subjected to spray drying with a high pressure pump at a controlled feed rate and temperature, to get dry powder. The desired particle size of the product is obtained by controlling the inside temperature of the chamber and by varying the pressure of the pump. The dry powder is mixed with suitable diluents or excipients and blended in a double cone mixer to obtain a homogeneous powder that can be straightaway used, for example, for filling in capsules or making tablets.

1.2.2.3 Solvent Extraction

The principle of solid-liquid extraction is that when a solid material comes in contact with a solvent, the soluble components in the solid material move to the solvent. Thus, solvent extraction of plant material results in the mass transfer of soluble active principle (medicinal ingredient) to the solvent, and this takes place in a concentration gradient. The rate of mass transfer decreases as the concentration of active principle in the solvent increases, until equilibrium is reached, i.e. the concentrations of active principle in the solid material and the solvent are the same. Thereafter, there will no longer be a mass transfer of the active principle from plant material to the solvent.

Since mass transfer of the active principle also depends on its solubility in the solvent, heating the solvent can enhances the mass transfer. Moreover, if the solvent in equilibrium with the plant material is replaced with fresh solvent, the concentration gradient is changed. This gives rise to different types of extractions: cold percolation, hot percolation and concentration.

1.2.2.3.1 Cold Percolation

The extraction of plant material is carried out in a percolator which is a tall cylindrical vessel with a conical bottom and a built-in false bottom with a filter cloth. The percolator is connected to a condenser and a receiver for stripping solvent from the marc.

The powdered material is fed into the percolator along with a suitable solvent (ethyl alcohol or another non-polar solvent). The material is left in contact with the solvent until equilibrium of the active principle is achieved. The solvent extract, known as *miscella*, is taken out from the bottom discharge valve of the percolator. Fresh solvent is added into the percolator and the miscella is drained out after acquiring equilibrium. Overall, the plant material is washed four to five times until it gets exhausted. All washes from the percolator are pooled and concentrated.

The solvent in the marc is stripped out by passing steam from the bottom of the percolator. The solvent and steam vapors rise and are condensed in a tubular condenser. The condensate, which is a mixture of alcohol and water, is collected in a receiver and then subjected to fractional distillation to get 95% pure ethyl alcohol which is again used as a fresh solvent.

This type of percolation is not efficient as it takes a long time to reach equilibrium due to the slow mass transfer rate. The mass transfer rate can be enhanced if some sort of movement is created between the particles and the solvent.

This can be achieved either by providing inside agitation with a mechanical stirrer or by repeated circulation of the extract back to the percolator. The first method is cumbersome and power intensive whereas the latter has been successful. A circulation pump that continuously circulates the miscella back to the top of the percolator gives a better mass transfer rate and reduces the equilibrium time considerably. Still, this type of percolation is energy-consuming as large amounts of miscella from multiple washes must be concentrated to remove the solvent.

To overcome this problem, a battery of percolators can be connected in series. If three washes are required for completion of the extraction, four percolators are connected in series with their respective miscella storage tanks.

At a particular time, one percolator is out of circuit, for charging and discharging the material and also for stripping solvent from the marc, whereas the other three percolators are in operation. Material is fed into all the percolators and the solvent is fed into the first percolator. When the equilibrium in the first percolator is reached, the extract from the first percolator is sent to the second percolator. The first percolator is again filled with fresh solvent. The extract of second percolator is transferred to the third, the extract of first is transferred to second, and fresh solvent is added to the first. The extract of the third percolator is transfered to the fourth percolator. After attaining equilibrium, the extract from the fourth percolator is drained off. The extract of the third percolator goes to fourth, the extract of second goes to third, and the extract of first goes to second percolator. The material of the first percolator, which has received three washes, is completely exhausted. This percolator is taken out of the system for stripping the solvent and discharging the extracted marc. This is again filled with fresh plant material and the sequence is repeated with other percolators. In this way, solvent of each percolator comes in contact three times with solid material and gets fully enriched with active principle. The enriched extract is sent for solvent recovery and concentration. Thus, instead of concentrating three volumes of solvent, only one volume has to be concentrated; this saves energy and the process is efficient.

1.2.2.3.2 Hot Percolation

Increasing the temperature of the solvent increases the solubility of the active principle, which increases the concentration gradient and therefore enhances the mass transfer of active principle from solid material to the solvent, provided the active principle is not heat sensitive. This is achieved by incorporating a heat exchanger between the circulation pump and the feed inlet of the percolator. The extract is continuously pumped into a tubular heat exchanger which is heated by steam. The temperature of the extract in the percolator is controlled by a steam solenoid valve through a temperature indicator controller. This sort of arrangement can be incorporated in single percolators or in a battery of percolators as needed.

The percolators that are tall cylindrical towers must be housed in sheds of relatively great height. Tall towers are difficult to operate, especially when charging material and discharging the marc from the top and bottom manholes, which are time-consuming and labor-intensive procedures. Tall towers have been replaced by extractors of smaller height for which the H/D ratio is not more than 1.5.

These extractors have perforated baskets in which the material to be extracted is charged. These perforated baskets, when loaded outside, can be inserted into the extractor with a chain pulley block and, after the extraction, they can be lifted out from the extractor for discharging the marc. Some extractors have an electrical hoist for the charging the material and discharging the marc, which makes the operation less labor-intensive, quick and efficient.

The other type of instrument for extraction of medicinal ingredients from plant material is the Soxhlet apparatus, which consists of an extractor, a distillation still, a tubular condenser for the distillation still, a tubular condenser for the recovery of solvent from the marc, a receiver for collecting the condensate from the condenser, and a solvent storage tank. The plant material is fed into the extractor, and solvent is added until it reaches the siphon point of the extractor. Then, the extract is siphoned out into the distillation still, which is heated with steam. The solvent vapors go to the distillation condenser, get condensed and return to the extractor. The level of the solvent in the extractor again rises to the siphon point and the extract is siphoned out into the distillation still. In this way, fresh solvent comes in contact with the plant material a number of times, until the plant material is completely extracted. The final extract in the distillation still, which is rich in active principle, is concentrated and the solvent is recovered.

1.2.2.3.3 Concentration

The enriched extract from percolators or extractors, known as miscella, is fed into a wiped film evaporator where it is concentrated under vacuum to produce a thick concentrated extract. The concentrated extract is further fed into a vacuum chamber dryer to produce a solid mass free from solvent. The solvent recovered from the wiped film evaporator and vacuum chamber dryer is recycled back to the percolator or extractor for the next batch of plant material. The solid mass thus obtained is pulverized and used directly for the desired pharmaceutical formulations or further processed for isolation of its phytoconstituents.

1.3 Aromatic Plant Extracts

The types of volatile isolates that are obtained commercially from aromatic plants are essential oils, concretes, absolutes, pomades and resinoids. Essential oils are isolated from plant material by distillation whereas other volatile isolates are obtained by solvent extraction.

1.3.1 Concrete

This is an extract of fresh flowers, herbs, leaves and the flowering tops of plants obtained by the use of a hydrocarbon solvent such as butane, pentane, hexane and petroleum ether. Concrete is rich in hydrocarbon-soluble material and devoid of water-soluble components. It is generally a waxy, semisolid, dark-colored material free from the original solvent.

In practice, concretes are produced in static extractors. These extractors are fitted with numerous perforated trays so that the flowers do not get compressed by their own weight. Each perforated tray has a spacer so the number and distance between them are predetermined.

The set of perforated trays can be within a removable cylindrical basket. In the centre of the lower tray, there is a rod on which the spacers and the perforated trays are fitted while at the top there is a ring or a hook so that the entire contents of the extractor can be readily removed by a chain pulley block.

While stacking the flowers on these trays, care should be taken to minimize bruising and damage of the flowers, because such damage can result in the release of enzymes in the flower juice which deteriorates the quality of concrete. The basket stacked with flowers is inserted into the extractor and the solvent of choice is introduced from the bottom into the extractor until the material on the perforated disc assembly is completely immersed. Four to five such washes are given until the material is exhausted.

The enriched solvent from the extractor is pumped into an evaporator for solvent recovery and the solvent content is reduced to about one-tenth the original volume. The recovered solvent is pumped to the solvent tanks to be used again. The concentrated material from the evaporator is pumped into a vacuum evaporator where the solvent is removed more carefully under high vacuum and the recovered solvent is returned to the solvent tanks for repeated use. The resultant concrete has an odor similar to but stronger than the material from which it was extracted.

In concrete manufacturing, it is a normal practice to circulate fresh solvent through a battery of extractors. At each cycle, the solvent becomes more enriched with the flower volatiles until extraction is complete.

The number of extractors has to be synchronized with the number of solvent washes.

1.3.2 Absolutes

Concretes are not widely used in perfumery in their native form but are generally converted into an alcohol-soluble volatile concentrate known as an absolute, i.e. they have to be extracted with alcohol.

To make an absolute, the concrete is mixed with absolute alcohol and agitated thoroughly in a vessel with an agitator. During agitation, the temperature is kept at 40°-60° C and the concrete is immersed in the solution. The solution is cooled down to -5° to -10° C to precipitate out the wax, since waxes are normally insoluble in alcohol below -1° C. The precipitated wax is removed by passing the solution through a rotary filter. The filtrate from the rotary filter is pumped into a primary evaporator, where it is concentrated to about 10% alcohol content. Finally, the concentrated extract is pumped into an agitating-type evaporator, where the alcohol is carefully removed under high vacuum.

1.3.3 Resinoids

Resinoid is an extract of naturally resinous material, made with a hydrocarbon solvent. Resinoids are usually obtained from dry materials. The extraction process is same as that of concrete production, except that perforated discs are not used for stacking the material; instead powder from dry plant material is fed into the extractor.

1.3.4 Pomades

Pomades are obtained by a process known as enfleurage, which is a cold fat extraction method. The fat is spread out on glass plates contained in wooden frames, leaving a clear margin near the edges. The absorptive surface of the fat is increased by surface grooves made with a wooden spatula.

Fresh flowers are spread out on the surface of the fat and the frames are stacked in piles. After the perfume oils have been absorbed from the flowers, the spent flowers are removed by hand. Fresh flowers are again spread on the fat surface. This is repeated until the fat surface is completely enriched with perfume oils. The pomade so obtained is ready for cold alcoholic extraction.

1.3.5 Essential Oils

Essential oils are used in a wide variety of consumer goods such as detergents, soaps, toilet products, cosmetics, pharmaceuticals, perfumes, confectionery food products, soft drinks, distilled alcoholic beverages (hard drinks) and insecticides. The world production and consumption of essential oils and perfumes are increasing very fast. Production technology is an essential element to improve the overall yield and quality of essential oil. The traditional technologies pertaining to essential oil processing are of great significance and are still being used in many parts of the globe. Water distillation, water and steam distillation, steam distillation, cohobation, maceration and enfleurage are the most traditional and commonly used methods. Maceration is adaptable when oil yield from distillation is poor. Distillation methods are good for powdered almonds, rose petals and rose blossoms, whereas solvent extraction is suitable for expensive, delicate and thermally unstable materials like jasmine, tuberose, and hyacinth. Water distillation is the most favored method of production of citronella oil from plant material.

1.3.5.1 Sources of Natural Essential Oils

Plant organs containing natural essential oils are illustrated in Figure 3. Essential oils are generally derived from one or more plant parts, such as flowers (e.g. rose, jasmine, carnation, clove, mimosa, rosemary, lavander), leaves (e.g. mint, *Ocimum* spp., lemongrass, jamrosa), leaves and stems (e.g. geranium, patchouli, petitgrain, verbena, cinnamon), bark (e.g. cinnamon, cassia, canella), wood (e.g. cedar, sandal, pine), roots (e.g. angelica, sassafras, vetiver, saussurea, valerian), seeds (e.g fennel, coriander, caraway, dill, nutmeg), fruits (bergamot, orange, lemon, juniper), rhizomes (e.g. ginger, calamus, curcuma, orris) and gums or oleoresin exudations (e.g. balsam of Peru, balsam of Tolu, storax, myrrh, benzoin).

Specialized plant structures that produce and store essential oils are shown in Figure 4. Depending upon the plant family, essential oils may occur in specialized secretary structures such as glandular hairs (Labiatae, Verbenaceace, Geraniaceae), modified parenchymal cells (Piperaceae), resin canals (conifers), oil tubes called vittae (Umbelliferae), lysigenous cavities (Rutaceae), schizogenous passages (Myrtaceae, Graminae, Compositae) or gum canals (Cistacae, Burseraceae). It is well known that when a geranium leaf is lightly touched, an odor is emitted because the long stalked oil glands are fragile. Similarly, the application of slight pressure on a peppermint leaf will rupture the oil gland and release oil. In contrast, pine needles and eucalyptus leaves do not release their oils until the epidermis of the leaf is broken. Hence, the types of structures in which oil is contained differ depending on the plant type and are plant-family specific. Unfortunately, not enough is known even today about these oil secretary structures to carefully categorize them. From the practical standpoint, they can be categorized into superficial and subcutaneous oils. Based on the currently available information, it may be inferred that oils of the Labiatae, Verbenaceae and Geraniaceae families are the only superficial oils known; consequently, the others are considered subcutaneous oils.

During handling, some flowers continue to produce aroma while other quickly loose their odor. Flowers collected at different times may also give different perfumery values. Regarding the rose, half-open flowers with plump anthers give higher oil yield than fully opened flowers with shrivelled anthers. Humidity, wind, rain and surface temperature also affect the oil yield considerably. Harvesting schedule affects both quantity and quality of the oil.

1.3.5.2 Essential Oil Constituents

Major constituents of essential oils are shown in Figure 5, from which it is clear that most essential oils consist of hydrocarbons, esters, terpenes, lactones, phenols, aldehydes, acids, alcohols, ketones, and esters. Among these, the oxygenated compounds (alcohols, esters, aldehydes, ketones, lactones, phenols) are the principal odor source. They are more stable against oxidizing and resinifying influences than other constituents. On the other hand, unsaturated constituents like monoterpenes and sesquiterpenes have the tendency to oxidize or resinify in the presence of air and light. The knowledge of individual constituents and their physical characteristics, such as boiling point, thermal stability and vapor-pressure-temperature relationship, is of paramount importance in technology development of oxygenated compounds.

1.3.5.3 Methods of Producing Essential Oils

Methods for producing essential oils from plant materials are summarized in Figure 6. Regarding hydrodistillation, the essential oils industry has developed terminology to distinguish three types: water distillation; water and steam distillation; and direct steam distillation.

Originally introduced by Von Rechenberg, these terms have become established in the essential oil industry. All three methods are subject to the same theoretical considerations which deal with distillation of two-phase systems. The differences lie mainly in the methods of handling the material.

Some volatile oils cannot be distilled without decomposition and thus are usually obtained by expression (lemon oil, orange oil) or by other mechanical means. In certain countries, the general method for obtaining citrus oil involves puncturing the oil glands by rolling the fruit over a trough lined with sharp projections that are long enough to penetrate the epidermis and pierce the oil glands located within outer portion of the peel (*ecuelle* method). A pressing action on the fruit removes the oil from the glands, and a fine spray of water washes the oil from the mashed peel while the juice is extracted through a central tube that cores the fruit. The resulting oil-water emulsion is separated by centrifugation. A variation of this process is to remove the peel from the fruit before the oil is extracted.

Often, the volatile oil content of fresh plant parts (flower petals) is so small that oil removal is not commercially feasible by the aforementioned methods. In such instances, an odorless, bland, fixed oil or fat is spread in a thin layer on glass plates. The flower petals are placed on the fat for a few hours; then repeatedly, the oil petals are removed, and a new layer of petals is introduced. After the fat has absorbed as much fragrance as possible, the oil may be removed by extraction with alcohol. This process, known as enfleurage, was formerly used extensively in the production of perfumes and pomades.

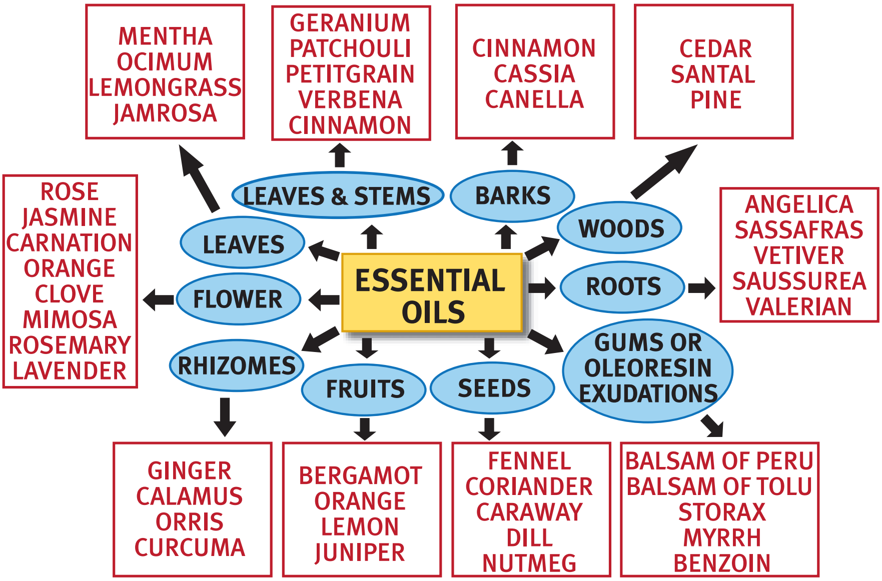
MENTHA OCIMUM LEMONGRASS JAMROSA
GERANIUM PATCHOULI PETITGRAIN VERBENA CINNAMON
CINNAMON CASSIA CANELLA
CEDAR SANTAL PINE
ROSE JASMINE CARNATION ORANGE CLOVE MIMOSA ROSEMARY LAVENDER
LEAVES & STEMS
BARKS
WOODS
LEAVES
ANGELICA SASSAFRAS VETIVER SAUSSUREA VALERIAN
FLOWER
ESSENTIAL OILS
ROOTS
RHIZOMES
FRUITS
SEEDS
GUMS OR OLEORESIN EXUDATIONS
GINGER CALAMUS ORRIS CURCUMA
BERGAMOT ORANGE LEMON JUNIPER
FENNEL CORIANDER CARAWAY DILL NUTMEG
BALSAM OF PERU BALSAM OF TOLU STORAX MYRRH BENZOIN

Figure 3: Plant organs containing essential oils

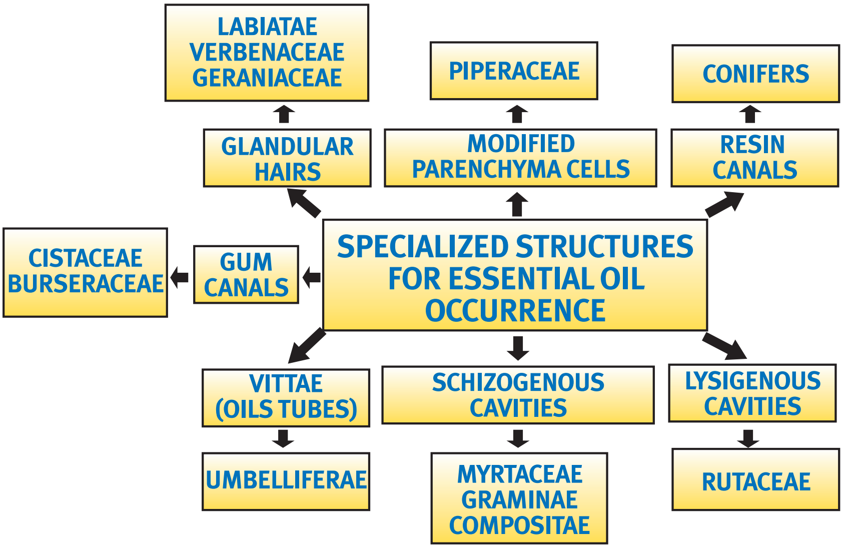
LABIATAE VERBENACEAE GERANIACEAE
PIPERACEAE
CONIFERS
GLANDULAR HAIRS
MODIFIED PARENCHYMA CELLS
RESIN CANALS
CISTACEAE BURSERACEAE
GUM CANALS
SPECIALIZED STRUCTURES FOR ESSENTIAL OIL OCCURRENCE
VITTAE (OILS TUBES)
SCHIZOGENOUS CAVITIES
LYSIGENOUS CAVITIES
UMBELLIFERAE
MYRTACEAE GRAMINAE COMPOSITAE
RUTACEAE

Figure 4: Family-specific plant tissues responsible for producing or storing essential oil

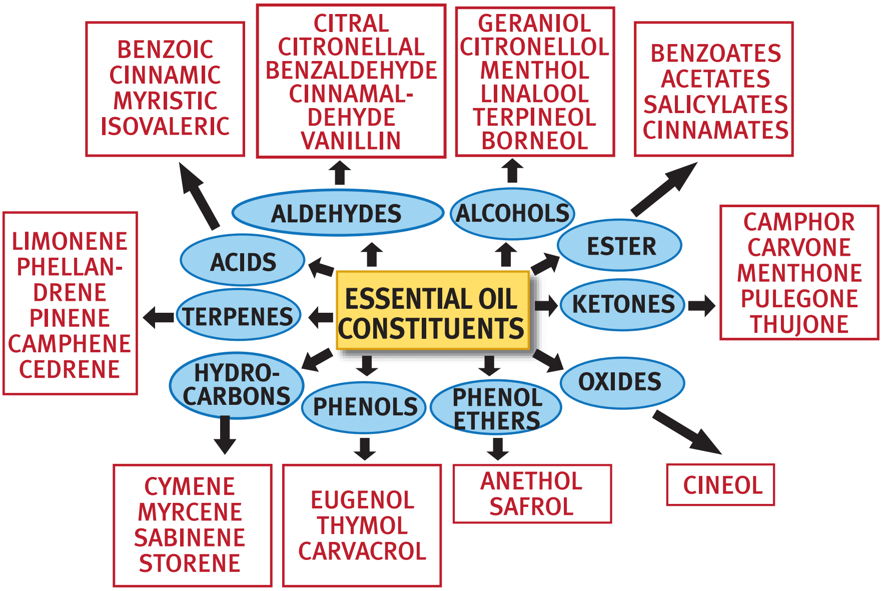
BENZOIC
CINNAMIC
MYRISTIC
ISOVALERIC
CITRAL
CITRONELLAL
BENZALDEHYDE
CINNAMALDEHYDE
VANILLIN
GERANIOL
CITRONELLOL
MENTHOL
LINALOOL
TERPINEOL
BORNEOL
BENZOATES
ACETATES
SALICYLATES
CINNAMATES
ALDEHYDES
ALCOHOLS
ACIDS
ESTER
LIMONENE
PHELLANDRENE
PINENE
CAMPHENE
CEDRENE
TERPENES
ESSENTIAL OIL CONSTITUENTS
KETONES
CAMPHOR
CARVONE
MENTHONE
PULEGONE
THUJONE
HYDROCARBONS
PHENOLS
PHENOL ETHERS
OXIDES
CYMENE
MYRCENE
SABINENE
STORENE
EUGENOL
THYMOL
CARVACROL
ANETHOL
SAFROL
CINEOL

Figure 5: Heterogeneous chemical groups present in essential oil

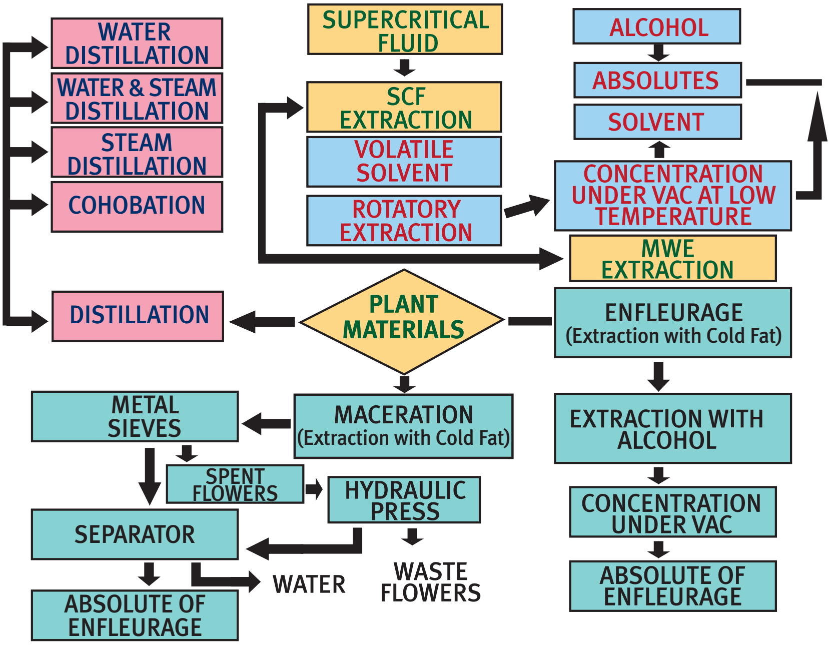
WATER DISTILLATION
WATER & STEAM DISTILLATION
STEAM DISTILLATION
COHOBATION
DISTILLATION
SUPERCRITICAL FLUID
SCF EXTRACTION
VOLATILE SOLVENT
ROTATORY EXTRACTION
ALCOHOL
ABSOLUTES
SOLVENT
CONCENTRATION UNDER VAC AT LOW TEMPERATURE
MWE EXTRACTION
PLANT MATERIALS
ENFLEURAGE
(Extraction with Cold Fat)
METAL SIEVES
MACERATION
(Extraction with Cold Fat)
EXTRACTION WITH ALCOHOL
SPENT FLOWERS
HYDRAULIC PRESS
CONCENTRATION UNDER VAC
SEPARATOR
ABSOLUTE OF ENFLEURAGE
WATER
WASTE FLOWERS
ABSOLUTE OF ENFLEURAGE

Figure 6: Methods of producing essential oils from plant materials

In the perfume industry, most modern essential oil production is accomplished by extraction, using volatile solvents such as petroleum ether and hexane. The chief advantages of extraction over distillation is that uniform temperature (usually 50° C) can be maintained during the process, As a result, extracted oils have a more natural odor that is unmatched by distilled oils, which may have undergone chemical alteration by the high temperature. This feature is of considerable importance to the perfume industry; however, the established distillation method is of lower cost than the extraction process.

Destructive distillation means distilling volatile oil in the absence of air. When wood or resin of members of the Pinaceae or Cupressaceae is heated without air, decomposition takes place and a number of volatile compounds are driven off. The residual mass is charcoal. The condensed volatile matter usually separates into 2 layers: an aqueous layer containing wood naptha (methyl alcohol) and pyroligneous acid (crude acetic), and a tarry liquid in the form of pine tar, juniper tar, or other tars, depending on the wood used. This dry distillation is usually conducted in retorts and, if the wood is chipped or coarsely ground and the heat is applied rapidly, the yield often represents about 10% of the wood weight used.

1.3.5.3.1 Hydrodistillation

In order to isolate essential oils by hydrodistillation, the aromatic plant material is packed in a still and a sufficient quantity of water is added and brought to a boil; alternatively, live steam is injected into the plant charge. Due to the influence of hot water and steam, the essential oil is freed from the oil glands in the plant tissue. The vapor mixture of water and oil is condensed by indirect cooling with water. From the condenser, distillate flows into a separator, where oil separates automatically from the distillate water.

1.3.5.3.1.1 Mechanism of Distillation

Hydrodistillation of plant material involves the following main physicochemical processes:

i) Hydrodiffusion
ii) Hydrolysis
iii) Decomposition by heat

1.3.5.3.1.1.1 Hydrodiffusion

Diffusion of essential oils and hot water through plant membranes is known as hydrodiffusion. In steam distillation, the steam does not actually penetrate the dry cell membranes. Therefore, dry plant material can be exhausted with dry steam only when all the volatile oil has been freed from the oil-bearing cells by first thorough comminution of the plant material. But, when the plant material is soaked with water, exchange of vapors within the tissue is based on their permeability while in swollen condition. Membranes of plant cells are almost impermeable to volatile oils. Therefore, in the actual process, at the temperature of boiling water, a part of volatile oil dissolves in the water present within the glands, and this oil-water solution permeates, by osmosis, the swollen membranes and finally reaches the outer surface, where the oil is vaporized by passing steam.

Another aspect of hydrodiffusion is that the speed of oil vaporization is not influenced by the volatility of the oil components, but by their degree of solubility in water. Therefore, the high-boiling but more water-soluble constituents of oil in plant tissue distill before the low-boiling but less water-soluble constituents. Since hydrodiffusion rates are slow, distillation of uncomminuted material takes longer time than comminuted material.

1.3.5.3.1.1.2 Hydrolysis

Hydrolysis in the present context is defined as a chemical reaction between water and certain constituents of essential oils. Esters are constituents of essential oils and, in the presence of water, especially at high temperatures, they tend to react with water to form acids and alcohols. However, the reactions are not complete in either direction and the relationship between the molal concentrations of various constituents at equilibrium is written as:

$$K = \frac{(\text{alcohol}) \times (\text{acid})}{(\text{ester}) \times (\text{water})}$$

where K is the equilibrium constant.

Therefore, if the amount of water is large, the amounts of alcohol and acid will also be large, resulting in a decreased yield of essential oil. Furthermore, since this is a time-dependent reaction, the extent to which hydrolysis proceeds depends on the time of contact between oil and water. This is one of the disadvantages of water distillation.

1.3.5.3.1.1.3 Effect of Heat

Almost all constituents of essential oils are unstable at high temperature. To obtain the best quality oil, distillation must be done at low temperatures. The temperature in steam distillation is determined entirely by the operating pressure, whereas in water distillation and in water and steam distillation the operating pressure is usually atmospheric.

All the previously described three effects, i.e. hydrodiffusion, hydrolysis and thermal decomposition, occur simultaneously and affect one another. The rate of diffusion usually increases with temperatures as does the solubility of essential oils in water. The same is true for the rate and extent of hydrolysis. However, it is possible to obtain better yield and quality of oils by: (1) maintaining the temperature as low as possible, (2) using as little water as possible, in the case of steam distillation, and (3) thoroughly comminuting the plant material and packing it uniformly before distillation.

1.3.5.3.2 Three Types of Hydrodistillation

Three are three types of hydrodistillation for isolating essential oils from plant materials:

1. Water distillation
2. Water and steam distillation
3. Direct steam distillation

1.3.5.3.2.1 Water Distillation

In this method, the material is completely immersed in water, which is boiled by applying heat by direct fire, steam jacket, closed steam jacket, closed steam coil or open steam coil. The main characteristic of this process is that there is direct contact between boiling water and plant material.

When the still is heated by direct fire, adequate precautions are necessary to prevent the charge from overheating. When a steam jacket or closed steam coil is used, there is less danger of overheating; with open steam coils this danger is avoided. But with open steam, care must be taken to prevent accumulation of condensed water within the still. Therefore, the still should be well insulated. The plant material in the still must be agitated as the water boils, otherwise agglomerations of dense material will settle on the bottom and become thermally degraded. Certain plant materials like cinnamon bark, which are rich in mucilage, must be powdered so that the charge can readily disperse in the water; as the temperature of the water increases, the mucilage will be leached from the ground cinnamon. This greatly increases the viscosity of the water-charge mixture, thereby allowing it to char. Consequently, before any field distillation is done, a small-scale water distillation in glassware should be performed to observe whether any changes take place during the distillation process. From this laboratory trial, the yield of oil from a known weight of the plant material can be determined. The laboratory apparatus recommended for trial distillations is the Clevenger system (Figure 7).

During water distillation, all parts of the plant charge must be kept in motion by boiling water; this is possible when the distillation material is charged loosely and remains loose in the boiling water. For this reason only, water distillation possesses one distinct advantage, i.e. that it permits processing of finely powdered material or plant parts that, by contact with live steam, would otherwise form lumps through which the steam cannot penetrate. Other practical advantages of water distillation are that the stills are inexpensive, easy to construct and suitable for field operation. These are still widely used with portable equipment in many countries.

The main disadvantage of water distillation is that complete extraction is not possible. Besides, certain esters are partly hydrolyzed and sensitive substances like aldehydes tend to polymerize. Water distillation requires a greater number of stills, more space and more fuel. It demands considerable experience and familiarity with the method. The high-boiling and somewhat water-soluble oil constituents cannot be completely vaporized or they require large quantities of steam. Thus, the process becomes uneconomical. For these reasons, water distillation is used only in cases in which the plant material by its very nature cannot be processed by water and steam distillation or by direct steam distillation.

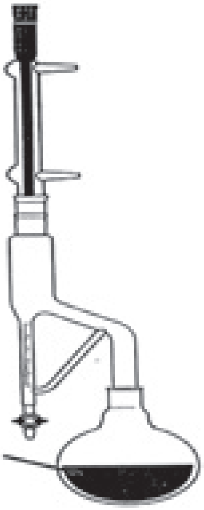

Figure 7: Clevenger-type laboratory-scale hydrodistillation apparatus

1.3.5.3.2.1.1 Traditional Method of Producing Attar Using Hydrodistillation

Floral attars are defined as the distillates obtained by hydrodistillation of flowers (such as saffron, marigold, rose, jasmine, pandanus) in sandal wood oil or other base materials like paraffin.

Attar manufacturing takes place in remote places because the flowers must be processed quickly after collection. The apparatus and equipment used to manufacture attar are light, flexible, easy to repair, and have a fair degree of efficiency. Keeping in view these facts, the traditional “deg and bhapka” process has been used for centuries and is used even now with the following traditional equipment (Figure 8).

- Deg (still)
- Bhapka (receiver)
- Chonga (bamboo condenser)

- Traditional bhatti (furnace)
- Gachchi (cooling water tank)
- Kuppi (leather bottle)

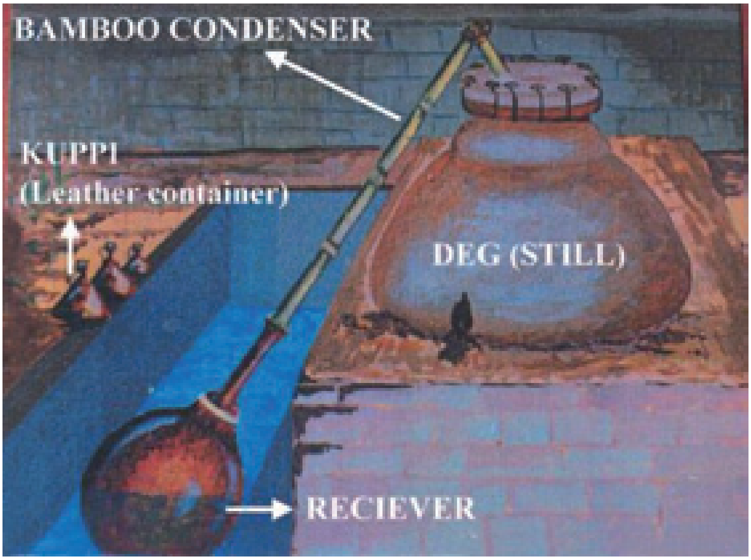

Figure 8: Traditional process of water distillation for making attar

1.3.5.3.2.1.2 Disadvantages of Water Distillation

- Oil components like esters are sensitive to hydrolysis while others like acyclic monoterpene hydrocarbons and aldehydes are susceptible to polymerization (since the pH of water is often reduced during distillation, hydrolytic reactions are facilitated).
- Oxygenated components such as phenols have a tendency to dissolve in the still water, so their complete removal by distillation is not possible.
- As water distillation tends to be a small operation (operated by one or two persons), it takes a long time to accumulate much oil, so good quality oil is often mixed with bad quality oil.
- The distillation process is treated as an art by local distillers, who rarely try to optimize both oil yield or quality.
- Water distillation is a slower process than either water and steam distillation or direct steam distillation.

1.3.5.3.2.2 Water and Steam Distillation

In water and steam distillation, the steam can be generated either in a satellite boiler or within the still, although separated from the plant material. Like water distillation, water and steam distillation is widely used in rural areas. Moreover, it does not require a great deal more capital expenditure than water distillation. Also, the equipment used is generally similar to that used in water distillation, but the plant material is supported above the boiling water on a perforated grid. In fact, it is common that persons performing water distillation eventually progress to water and steam distillation.

It follows that once rural distillers have produced a few batches of oil by water distillation, they realize that the quality of oil is not very good because of its still notes (subdued aroma). As a result, some modifications are made. Using the same still, a perforated grid or plate is fashioned so that the plant material is raised above the water. This reduces the capacity of the still but affords a better quality of oil. If the amount of water is not sufficient to allow the completion of distillation, a cohobation tube is attached and condensate water is added back to the still manually, thereby ensuring that the water, which is being used as the steam source, will never run out. It is also believed that this will, to some extent, control the loss of dissolved oxygenated constituents in the condensate water because the re-used condensate water will allow it to become saturated with dissolved constituents, after which more oil will dissolve in it.

1.3.5.3.2.2.1 Cohobation

Cohobation is a procedure that can only be used during water distillation or water and steam distillation. It uses the practice of returning the distillate water to the still after the oil has been separated from it so that it can be re-boiled. The principal behind it is to minimize the losses of oxygenated components, particularly phenols which dissolve to some extent in the distillate water. For most oils, this level of oil loss through solution in water is less than 0.2%, whereas for phenol-rich oils the amount of oil dissolved in the distillate water is 0.2%-0.7%. As this material is being constantly re-vaporized, condensed and re-vaporized again, any dissolved oxygenated constituents will promote hydrolysis and degradation of themselves or other oil constituents. Similarly, if an oxygenated component is constantly brought in contact with a direct heat source or side of a still, which is considerably hotter than 100° C, then the chances of degradation are enhanced. As a result, the practice of cohobation is not recommended unless the temperature to which oxygenated constituents in the distillate are exposed is no higher than 100° C.

In steam and water distillation, the plant material cannot be in direct contact with the fire source beneath the still; however, the walls of the still are good conductors of heat so that still notes can also be obtained from the thermal degradation reactions of plant material that is touching the sides of the still. As the steam in the steam and water distillation process is wet, a major drawback of this type of distillation is that it will make the plant material quite wet. This slows down distillation as the steam has to vaporize the water to allow it to condense further up the still. One way to prevent the lower plant material resting on the grid from becoming waterlogged is to use a baffle to prevent the water from boiling too vigorously and coming in direct contact with the plant material.

1.3.5.3.2.2.2 Advantages of Water and Steam Distillation over Water Distillation

- Higher oil yield.
- Components of the volatile oil are less susceptible to hydrolysis and polymerization (the control of wetness on the bottom of the still affects hydrolysis, whereas the thermal conductivity of the still walls affects polymerization).
- If refluxing is controlled, then the loss of polar compounds is minimized.
- Oil quality produced by steam and water distillation is more reproducible.
- Steam and water distillation is faster than water distillation, so it is more energy efficient.

Many oils are currently produced by steam and water distillation, for example lemongrass is produced in Bhutan with a rural steam and water distillation system.

1.3.5.3.2.2.3 Disadvantages of Water and Steam Distillation

- Due to the low pressure of rising steam, oils of high-boiling range require a greater quantity of steam for vaporization - hence longer hours of distillation.
- The plant material becomes wet, which slows down distillation as the steam has to vaporize the water to allow it to condense further up the still.
- To avoid that the lower plant material resting on the grid becomes waterlogged, a baffle is used to prevent the water from boiling too vigorously and coming in direct contact with the plant material.

1.3.5.3.2.3 Direct Steam Distillation

As the name suggests, direct steam distillation is the process of distilling plant material with steam generated outside the still in a satellite steam generator generally referred to as a boiler. As in water and steam distillation, the plant material is supported on a perforated grid above the steam inlet. A real advantage of satellite steam generation is that the amount of steam can be readily controlled. Because steam is generated in a satellite boiler, the plant material is heated no higher than 100° C and, consequently, it should not undergo thermal degradation. Steam distillation is the most widely accepted process for the production of essential oils on large scale. Throughout the flavor and fragrance supply business, it is a standard practice.

An obvious drawback to steam distillation is the much higher capital expenditure needed to build such a facility. In some situations, such as the large-scale production of low-cost oils (e.g. rosemary, Chinese cedarwood, lemongrass, litsea cubeba, spike lavender, eucalyptus, citronella, cornmint), the world market prices of the oils are barely high enough to justify their production by steam distillation without amortizing the capital expenditure required to build the facility over a period of 10 years or more.

1.3.5.3.2.3.1 Advantages of Direct Steam Distillation

- Amount of steam can be readily controlled.
- No thermal decomposition of oil constituents.
- Most widely accepted process for large-scale oil production, superior to the other two processes.

1.3.5.3.2.3.2 Disadvantage of Direct Steam Distillation

- Much higher capital expenditure needed to establish this activity than for the other two processes.

1.3.5.3.3 Essential Oil Extraction by Hydrolytic Maceration Distillation

Certain plant materials require maceration in warm water before they release their essential oils, as their volatile components are glycosidically bound. For example, leaves of wintergreen (*Gaultheria procumbens*) contain the precursor gaultherin and the enzyme primeverosidase; when the leaves are macerated in warm water, the enzyme acts on the gaultherin and liberates free methyl salicylate and primeverose. Other similar examples include brown mustard (sinigrin), bitter almonds (amygdalin) and garlic (alliin).

1.3.5.3.4 Essential Oil Extraction by Expression

Expression or cold pressing, as it is also known, is only used in the production of citrus oils. The term expression refers to any physical process in which the essential oil glands in the peel are crushed or broken to release the oil. One method that was practiced many years ago, particularly in Sicily (*spugna* method), commenced with halving the citrus fruit followed by pulp removal with the aid of sharpened spoon-knife (known as a *rastrello*). The oil was removed from the peel either by pressing the peel against a hard object of baked clay (*concolina*) which was placed under a large natural sponge or by bending the peel into the sponge. The oil emulsion absorbed by the sponge was removed by squeezing it into the *concolina* or some other container. It is reported that oil produced this way contains more of the fruit odor character than oil produced by any other method.

A second method known as equaling (or the *scodella* method), uses a shallow bowl of copper (or sometimes brass) with a hollow central tube; the equaling tool is similar in shape to a shallow funnel. The bowl is equipped with brass points with blunt ends across which the whole citrus fruit is rolled by hand with some pressure until all of the oil glands have burst. The oil and aqueous cell contents are allowed to dribble down the hollow tube into a container from which the oil is separated by decantation. Obviously, hand pressing is impractical because it is an extremely slow process, e.g. on average only 2-4 lbs oil per day can be produced by a single person using one of these hand methods. As a result, over the years a number of machines have been designed to either crush the peel of a citrus fruit or crush the whole fruit and then separate the oil from the juice.

1.3.5.3.4.1 Pelatrice Process

In the *pelatrice* process, citrus fruits are fed from a hopper into the abrasive shell of the machine. The fruits are rotated against the abrasive shell by a slow-moving Archimedian screw whose surface rasps the fruit surfaces causing some of the essential oil cavities on the peel to burst and release their oil-water emulsion. This screw further transports the fruit into a hopper in which rollers covered with abrasive spikes burst the remaining oil cavities. The oil and water emulsion is washed away from the fruit by a fine spray of water. The emulsion next passes through a separator where any solids are removed, after which it passes through two centrifugal separators working in series to yield the pure oil. Most bergamot oil and some lemon oil are produced this way in Italy.

1.3.5.3.4.2 Sfumatrice Process

The *sfumatrice* equipment consists of a metallic chain that is drawn by two horizontal ribbed rollers. The peels are conveyed through these rollers during which time they are pressed and bent to release their oil. As in *pelatrice*, the oil is washed away from the *sfumatrice* rollers by fine sprays of water. Again, the oil is initially passed through a separator prior to being sent to two centrifuges in series, so that purified oil can be produced. At one time, *sfumatrice* was the most popular process for citrus oil isolation in Italy; however, today the *pelatrice* method appears more popular.

1.3.5.3.5 Essential Oil Extraction with Cold Fat (Enfleurage)

Despite the introduction of the modern process of extraction with volatile solvents, the old fashioned method of enfleurage, as passed on from father to son and perfected in the course of generations, still plays an important role. Enfleurage on a large scale is today carried out only in the Grasse region of France, with the possible exception of isolated instances in India where the process has remained primitive.

The principles of enfleurage are simple. Certain flowers (e.g. tuberose and jasmine) continue the physiological activities of developing and giving off perfume even after picking. Every jasmine and tuberose flower resembles, so to speak, a tiny factory continually emitting minute quantities of perfume. Fat possesses a high power of absorption and, when brought in contact with fragrant flowers, readily absorbs the perfume emitted. This principle, methodically applied on a large scale, constitutes enfleurage. During the entire period of harvest, which lasts for eight to ten weeks, batches of freshly picked flowers are strewn over the surface of a specially prepared fat base (corps), let there (for 24 h in the case of jasmine and longer in the case of tuberose), and then replaced by fresh flowers. At the end of the harvest, the fat, which is not renewed during the process, is saturated with flower oil. Thereafter, the oil is extracted from the fat with alcohol and then isolated.

The success of enfleurage depends to a great extent upon the quality of the fat base employed. Utmost care must be exercised when preparing the corps. It must be practically odorless and of proper consistency. If the corps is too hard, the blossoms will not have sufficient contact with the fat, curtailing its power of absorption and resulting in a subnormal yield of flower oil. On the other, if it is too soft, it will tend to engulf the flowers and the exhausted ones will adhere; when removed, the flowers will retain adhering fat, resulting in considerable shrinkage and loss of corps. The consistency of the corps must, therefore, be such that it offers a semihard surface from which the exhausted flowers can easily be removed. The process of enfleurage is carried out in cool cellars, and every manufacturer must prepare the corps according to the prevailing temperature in the cellars during the months of the flower harvest.

Many years of experience have proved that a mixture of one part of highly purified tallow and two parts of lard is eminently suitable for enfleurage. This mixture assures a suitable consistency of the corps in conjunction with high power of absorption. The fat corps thus prepared is white, smooth, absolutely of uniform consistency, free of water and practically odorless. Some manufacturers also add small quantities of orange flower or rose water when preparing the corps. This seems to be done for the sake of convention. Such additions somewhat shade the odor of the finished product by imparting a slight orange blossom or rose note.

1.3.5.3.5.1 Enfleurage and Defleurage

Every enfleurage building is equipped with thousands of so-called chassis, which serve as vehicles for holding the fat corps during the process. A chassis consists of a rectangular wooden frame. The frame holds a glass plate upon both sides of which the fat corps is applied with a spatula at the beginning of the enfleurage process. When piled one above the other, the chassis form airtight compartments, with a layer of fat on the upper and lower side of each glass plate.

Every morning during the harvest the freshly picked flowers arrive, and after being cleaned of impurities, such as leaves and stalks, are strewn by hand on top of the fat layer of each glass plate. Blossoms wet from dew or rain must never be employed, as any trace of moisture will turn the corps rancid. The chassis are then piled up and left in the cellars for 24 h or longer, depending upon the type of flowers. The latter rest in direct contact with one fat layer (the lower one), which acts as a direct solvent whereas the other fat layer (beneath the glass plate of the chassis above) absorbs only the volatile perfume given off by the flowers.

After 24 h, the flowers have emitted most of their oil and start to wither, developing an objectionable odor. They must then be removed from the corps, which process, despite all efforts to introduce labor-saving devices, is still done by hand. Careful removal of the flower (defleurage) is almost more important than charging the corps on the chassis with fresh flowers (enfleurage) and, therefore, the persons doing this work must be experienced and skilled. Most of the exhausted flowers will fall from the fat layer on the chassis glass plate when the chassis is struck lightly against the working table, but since it is necessary to remove every single flower and every particle of the flower, tweezers are used for this delicate operation. Immediately following defleurage, that is, every 24 h, the chassis are recharged with fresh flowers. For this purpose the chassis are turned over and the fat layer, which in the previous operation formed the top (ceiling) of the small chamber, is now directly charged with flowers. In the case of jasmine, the entire enfleurage process lasts about 70 days: daily the exhausted flowers are removed and the chassis are recharged with fresh ones. At the beginning of, and several times during, the harvest, the fat on the chassis is scratched over with metal combs and tiny furrows are drawn in order change and increase the surface of absorption.

At the end of the harvest, the fat is relatively saturated with flower oil and possesses the typical fragrance. The perfumed fat must then be removed from the glass plates between the chassis. For this purpose, it is scraped off with a spatula and then carefully melted and bulked in closed containers. The final product is called pomade (pomade de jasmine, pomade de tuberous, pomade de violet, etc.). The most highly saturated pomade is pomade no. 36, because the corps on the chassis have been treated with fresh flowers 36 times during the whole process of enfleurage. At the beginning of the harvest, every chassis is charged with about 360 g fat corps on each side of the glass plate, in other words, with 720 g per chassis. Every kilogram of fat corps should be in contact with about 2.5 kg (preferably with 3.0 kg) of jasmine flowers for the entire period of enfleurage, which lasts from 8 to 10 weeks. The quantities differ somewhat for different flowers. At the end of enfleurage, the fat corps has lost about 10% of its weight because of the various manipulations.

1.3.5.3.5.2 Hot Maceration Process

In this process, the long enfleurage time is reduced by the immersion of petals in molten fat heated at 45°-60° C for 1 to 2 h, depending upon the plant species. After each immersion, the fat is filtered and separated from the petals. After 10 to 20 immersions, the fat is separated from waste flowers and water. Absolute of maceration is then produced from fat containing oil through the process of extraction and concentration under reduced pressure. It is mainly used for highly delicate flowers whose physiological activities are lost rapidly after their harvest, such as lily of valley.

1.3.6 Modern (Non-traditional) Methods of Extraction of Essential Oils

Traditional methods of extraction of essential oils have been discussed and these are the methods most widely used on commercial scale. However, with technological advancement, new techniques have been developed which may not necessarily be widely used for commercial production of essential oils but are considered valuable in certain situations, such as the production of costly essential oils in a natural state without any alteration of their thermosensitive components or the extraction of essential oils for micro-analysis. These techniques are as follows:

- Headspace trapping techniques
  - Static headspace technique
  - Vacuum headspace technique
  - Dynamic headspace technique
- Solid phase micro-extraction (SPME)
- Supercritical fluid extraction (SFE)
- Phytosol (phytol) extraction
- Protoplast technique
- Simultaneous distillation extraction (SDE)
- Microwave distillation
- Controlled instantaneous decomposition (CID)
- Thermomicrodistillation
- Microdistillation
- Molecular spinning band distillation
- Membrane extraction

Some of these techniques are discussed in other chapters. Here, a few important, relevant references are provided.

1.4 Conclusions

Some of the major constraints in sustainable industrial exploitation of medicinal and aromatic plants (MAPs) are due to the fact that the countries of South East Asia have poor agricultural practices for MAPs, unscientific and indiscriminate gathering practices from the wild, poor post-harvest and post-gathering practices leading to poor quality raw material, lack of research for the development of high-yielding varieties of MAPs, poor propagation methods, inefficient processing techniques, poor quality control procedures, lack of research on process and product development, difficulty in marketing, non-availability of trained personnel, lack of facilities and tools to fabricate equipment locally, and finally lack of access to the latest technologies and market information. This calls for co-operation and co-ordination among various institutes and organizations of the region, in order to develop MAPs for sustainable commercial exploitation.

The process of extracting MAPs determines how efficiently we add value to MAP bioresources. In the case of essential oils, the extraction process affects the physical as well as internal composition. External appearance, at times, can result in rejection of the batch even if the analytical results are within acceptable limits. Furthermore, essential oils are evaluated internationally for their olfactory properties by experienced perfumers and these olfactory qualities supersede analytical results. Variations in the chemical constituents of the extracts of medicinal plants may result by using non-standardized procedures of extraction. Efforts should be made to produce batches with quality as consistent as possible (within the narrowest possible range).

Bibliography

Abraham, M. A., Sunol, A. K., 1997, Supercritical Fluids: Extraction and Pollution Prevention, *ACS Symposium Series, 670*, Washington

Banoti, A., 1980, Problems relating to the preparation and use of extracts from medicinal plants, *Fitoterapia*, 51: 5-11

Baser, K. H. C. and Kurkcuoglu, M., 1999, SPME and Headspace Assay Development: Application to Rose Products, In: *Proceedings of the IFEAT Conference*, 8-12 November 1998, London, p. 298-305

Baser, K. H. C., 1999, Essential oil extraction from natural products – nontraditional methods. ICS-UNIDO training course on Process simulation and essential oil extraction from Aromatic plants, 18-22 Oct 1999, Trieste, Italy

Buttery, R. G., Ling, L. C., 1996, Methods for Isolating Food and Plant volatiles. In: Takeoka, G. R., Teranishi, R, Williams, P. J. and Kobayashi (Eds), Biotechnology for Improved Foods and Flavours, *ACS Symposium Series*, 637: 240-248

Clery, R., 1999, Advances in Headspace Analysis of Flowers. In: *Proceedings of the IFEAT Conference*, 8-12 November 1998, London, p. 294-297

Craveiro, A. A., Matos, F. J. A., Alenkar, J. W. and Plumel, M. M., 1989, Microwave for extraction of essential oil, *Flavour and Fragrance Journal*, 4: 43-44

Crippa, F., 1980, Problems involved in pharmaceutical and cosmetic formulations containing extracts, *Fitoterapia*, 51: 59-66

DeSilva, T. K., 1995, Development of Essential Oil Industry. In: A Manual on the Essential oil Industry, Ed. T. K. DeSilva, UNIDO, Vienna p. 1-11

Fleisher, A., 1999, The poroplast extraction technique in the flavour and fragrance industry, *Perfumer and Flavor*, 15(5): 27-36

Forni, G. P., 1980, Thin layer and high performance liquid chromatography in the analysis of extracts, *Fitoterapia*, 51: 13-33

Guenther, E., 1952, The Essential Oils, vol. 6D Von Nostrand, Princeton, N. J.

Guenthur, E., 1949, The Essential Oils, vol. 3D, Van Nostrand, Princeton, N. J.

Handa S. S. and Kaul, M. K., 1997, Cultivation and Utilization of Aromatic Plants. Regional Research Laboratory, Jammu, India

Handa, S. S. and Kaul, M. K., 1996, Cultivation and Utilization of Medicinal Plants. Regional Research Laboratory, Jammu, India

Handa, S. S., 1999, Essential oil extraction from natural products, the traditional methods. ICS-UNIDO training course on Process simulation and essential oil extraction from Aromatic Plants, 18-22 Oct. 1999, Trieste, Italy

Handa, S. S., 2005, Traditional and Modern methods of extraction of essential oils from aromatic plants. Presentation at the training course on cultivation, post-harvesting and processing technologies of medicinal and aromatic plants in developing countries. ICS-UNIDO organized at Bomako, Mali (West Africa), 25-29 July 2005

Handa, S. S., Rakesh, D. D. and Vasisht, K., 2006, Compendium of Medicinal and Aromatic Plants, Asia, Vol. II, ICS-UNIDO, Trieste, Italy

Humphrey, J. L., Keller, I. I. G. E., 1997, Separation Process Technology, McGraw Hill, New York

Kaiser, R., 1991, Trapping, Investigation and Reconstruction of floral Scents. In: Muller, P. M. and Lamparsky, D. (Eds.), Perfumes, Art, Science and Technology, Elsevier, p. 213

Koedam, A., 1987, Some Aspects of Essential Oil Preparation. In: Sandra,P. and Bicchi, C. (Eds.), Capillary Gas Chromatography in Essential Oil Analysis, A. Huethig Verlag, Heidelberg, p. 13-15

Lang, F., and Stumpf, H., 1999, Considerations on future pharmacopoeial monographs for plant extracts, *Pharmeuropa*, 11: 2

Lawrence, B. M., 1995, The Isolation of Aromatic Materials from Natural Plant Products. In: De Silva, K. T. (Ed.), A Manual on the Essential Oil Industry, UNIDO, Vienna, p. 57-154

Lowrence, B. M., 1979, Commercial production of non-citrus oils in North America, *Perfumer and Flavor*, 3(6): 21-23

Magistretti, M. J., 1980, Remarks on pharmacological examination of plant extracts, *Fitoterapia*, 51: 67-78

Martinelli, E. M., 1980, Gas chromatography in the control of extracts, *Fitoterapia*, 51: 35-57

Rezzong, S. A., Baghdadi, M. W., Louka, N., Boutekedijiret, C. and Allaf, K., 1998, Study of a new extraction process: controlled instantaneous decompression. Application to the extraction of essential oil from rosemary leaves, *Flavour and Fragrance Journal*, 13: 251-258

Scheffer, J. J. C., 1997, Various methods for the isolation of essential oils, *Phytotherapy Research*, 10: S6-S7

Seader, J. D. and Henley, E. J., 1996, Separation Process Principles, Wiley, New York

Taylor, L. T, 1996, Supercritical Fluid Extraction, Wiley, New York

Verrall, M., 1996, Downstream Procession of Natural Products. A Practical Handbook. Wiley, Chichester

Wenclawiak, B., 1992, Analysis with Supercritical fluids: Extraction and Chromatography, Springer, Berlin

Werkhoff, P., Brennecke, S. and Bretschneider, W., 1998, Modern methods and extraction techniques for isolating volatile flavour compounds, *Contact*, 2: 16-23

2 Role of Process Simulation in Extraction Technologies for Medicinal and Aromatic Plants

M. Fermeglia

Abstract

This paper illustrates the role of process simulation in the field of extraction technologies for medicinal and aromatic plants. The paper starts with a brief introduction to process simulation fundamentals and the role of process simulation in the industry today. It describes procedures to follow in simulating a process and the benefits of process simulation. In the second part of the paper, phytochemical processes that have been simulated at ICS-UNIDO are listed, followed by two case studies to illustrate the applicability of the methodology proposed: (i) turpentine oil batch distillation and (ii) menthol recovery by crystallization of mentha oil. At the end, recommendations are given advocating the importance of process simulation for developing countries.

2.1 Introduction

Developing countries are rich in medicinal and aromatic plants (MAPs) but, due to difficulty in accessing efficient extraction technologies, value addition to this rich bioresource is difficult. In most cases, and particularly in very poor countries, the technologies used are inappropriate and not economical. The crucial problem is related to the quality of the product: primitive extraction technologies do not guarantee a stable and high-quality product and, in some cases, inappropriate technologies and procedures result in producing contaminated product which has low market value.

In order to assist developing countries to achieve the objective of using rich MAP resource for producing value-added products, dissemination of knowledge of existing extraction technologies and of the latest developments in these technologies is essential.

Commercial process simulation software can be used to predict, on a computer, the real plant and consequently is a useful tool for optimizing the process conditions and enhancing the capacity of managing the phytochemical processes. In particular, process simulation can assist developing and emerging countries in optimizing an advanced process rather than managing a primitive process, which should be substituted by more efficient and standardized procedures. The focus in this case is more related to practical problems such as the quality of the materials and of the water to be used for the extraction. In most cases, developing countries face problems in the type of vessel, quality of water and stability of the product during the processing.

This paper describes the use of process simulation software in the extraction and purification of essential oils at both pilot and industrial scales. Such processes have been developed and are in operation in developing and emerging countries. The goal of this paper is to illustrate a procedure for obtaining better knowledge of the extraction process and, therefore, for optimizing the process in terms of energy use, raw material consumption and environmental impact.

2.2 Process Simulation Goals and Definitions

Simulation is the act of representing some aspects of the real world by numbers or symbols which may be manipulated to facilitate their study. A process simulator is an engineering tool that performs several tasks, including automated calculations, material and energy balances, physical property estimations, design or rating calculations, and process optimization. A process simulator is not a process engineer, and a process engineer is always needed to analyze the problem and the output of a process simulator. A process simulator solves material and energy balances by means of computer code.

In principle, a process simulator for the study of a chemical process goes through the procedure outlined in Figure 1. One starts from the definition of the problem (problem analysis) and then develops the process model, i.e. the system of equations (algebraic or differential). Furthermore, one collects the necessary additional data and solves the model with a suitable method, depending on the system of equations. Finally, the process engineer analyzes the results and perhaps starts over again to develop a more realistic model.

The same picture applies to steady-state simulation, dynamic simulation and optimization problems; only the process model and the method of solution change. Solution of the system of material and energy balance equations is not an easy task because it must be solved considering many components, complex thermophysical models for phase equilibrium calculations, a large number of subsystems (equipment), rather complex equipment (e.g. distillation columns), recycle streams and control loops.

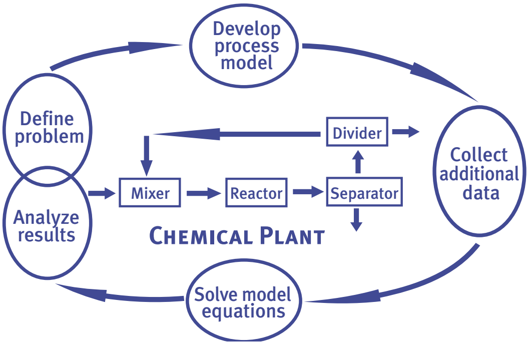

Figure 1: Fundamental steps in running process simulation software

A typical process simulation scheme, with the most important elements and their connections, is shown in Figure 2. Clearly, a process simulator includes cost estimation as well as economic evaluation. The importance of the database is shown in the figure as a necessary source of information for different objects in the structure.

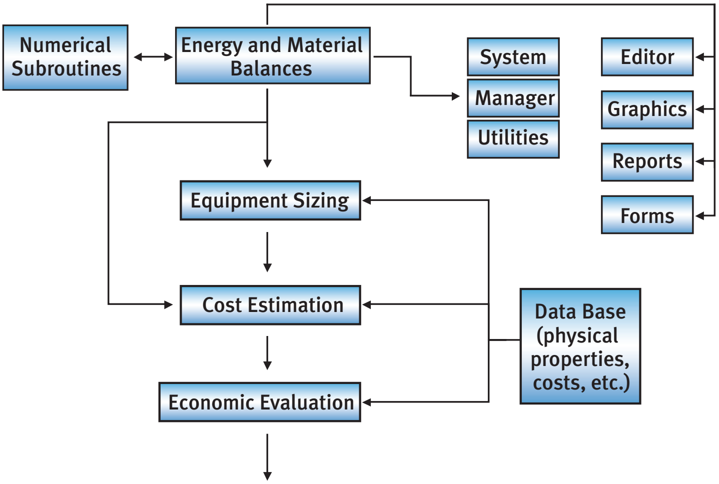
Numerical Subroutines
Energy and Material Balances
System
Manager
Utilities
Editor
Graphics
Reports
Forms
Equipment Sizing
Cost Estimation
Economic Evaluation
Data Base (physical properties, costs, etc.)

Figure 2: General scheme of a steady-state process simulator

The following approaches are available in process simulation:

- Steady state simulation, which considers a snapshot in time of the process.
- Dynamic simulation, which considers the evolution in the time domain of the equations describing the process.
- Integrated steady-state–dynamic simulation, which combines the previous two approaches.

These three approaches may be used in different ways when dealing with process simulation. One possibility is to perform *process analysis*, in which an existing process is studied and alternative conditions as well as dynamic behavior are investigated in the appraisal of effectiveness of the design. The second is *process synthesis*, in which different process configurations are compared in order to identify the best choice of units and the connections between them. The third possibility is *process design and simulation*, which aims at establishing the optimal operating conditions of a given process.

In all these possibilities, impact on industry is pervasive rather than restricted to a single moment in the development of the process. Process simulation has strongly affected the way engineering knowledge is used in processes. The traditional way of using process simulation was mainly focused on designing flowsheets and on defining critical equipment parameters, such as distillation column stages and column diameter. Today, engineers are oriented to a more comprehensive use of process simulation in the entire “life” of the plant, as in designing control strategies, optimizing process parameters, studying process time evolution for understanding startup and shutdown procedures, performing risk analysis, training operators, and defining procedures to reduce non-steady-state operations.

The main benefits gained from such a comprehensive use of process simulation are the partial or total replacement of pilot plants (reduction of the number of runs and planning), the reduction of time to market for the development of new processes, and the fast screening of process alternatives to select the best solution in terms of economics, environmental aspects, energy consumption and flexibility. Due to the high complexity of chemical processes, to get these benefits one must critically simplify the process and apply process simulation techniques in the entire life cycle of a process.

Steady-state simulators are considered the core products of process simulation and are used for designing processes, evaluating process changes and analyzing what-if scenarios. Steady-state simulation is normally performed before all other kinds of simulation: dynamic simulation, process synthesis with pinch technology, detailed equipment design, off-line and on-line equation-based optimization, and application technologies for vertical markets (e.g. polymers).

The problems involved in a process simulation run are the definition of an accurate thermodynamic model (equations of state or excess Gibbs energy model), the necessity of defining dummy operations (not always easy to identify), and the tear streams identification to achieve rapid convergence.

The logical procedure for performing a simulation is as follows. First, one defines all the components to be used in the simulation, including conventional and non-conventional components. Next, the most important step in the definition of the simulation is the selection of physicochemical properties to be used in the calculations. Having done this, one proceeds to flowsheet connectivity and to the definition of the feed conditions. The next step is the unit operation internal definitions. At this stage, it is possible to run a base case and check that the system is converging.

Process specification definition, control parameters, and equipment hold-up definition are added later to refine the simulation and to obtain results similar to the reality. Various different results are obtained from a simulation run. The most important are the validation of phase equilibria models for the real system to be used in similar conditions, the verification of the process operating conditions, information on intermediate streams and enthalpy balance, verification of plant specifications, and influence of the operative parameters on the process specifications.

All this information is useful for de-bottlenecking the entire process, or part of it, for identifying process control strategies, and for tuning the instrumentation. This is important since it allows one to verify the behavior of security systems when process conditions are varied.

As far as dynamic simulation is concerned, applications can be found in continuous processes, concurrent process and control design, evaluation of alternative control strategies, troubleshooting process operability, and verification of process safety.

The most important benefits of dynamic modeling are: the capital avoidance and lower operating costs through better engineering decisions; the throughput, product quality, safety and environmental improvements through improved process understanding; and the increased productivity through enhanced integration of engineering work processes.

2.3 Biotechnological and Phytochemical Processes Studied at ICS-UNIDO

This section summarizes ongoing work involving the simulation of important biotechnological and phytochemical processes.

2.3.1 Brief of Biotechnological Processes

2.3.1.1 Alcohol Production from Biomass

The goal of the process is the steady-state simulation of the production of ethanol from biomass fermentation. The process is divided into two parts: (i) biomass fermentation that produces a mixture of ethanol, water and other components, and (ii) separation using a distillation column that concentrates the ethanol. The modelled reactors are continuous stirred tank reactors in series or parallel.

2.3.1.2 Soybean Oil Refining and Treatment of the Waste

The goal of this work is to simulate the soybean oil refining process. This is a complex biotechnological process that involves many reactions and the treatment of solids. The process is divided into three sections: (i) degumming and neutralization, (ii) bleaching, and (iii) deodorization. The main objective of the simulation is to reduce the consumption of steam by applying the pinch technology concept. Steam is consumed in the heat exchangers to heat the oil, in the bleacher equipment, and in the deodorizer. The difficulty of this simulation lies in the large number of undefined components that must be characterized in order to obtain a reliable simulation. In addition, an alternative way of reducing the free acids, by using extraction with ethanol, is being examined. Key features are the achievements of the required product quality by minimizing the capital and operating costs.

2.3.1.3 Production of Synthetic Hydrocarbon Fuels from Biomass

Starting from natural gas, coal, or wood, a syngas of suitable composition can be produced by gasification. Then, water-gas shift reaction, Fischer-tropsch synthesis and hydrocracking can be applied to the syngas to obtain a mixture of liquid hydrocarbons that can be used as synthetic fuel. This is a complex process that can be simplified by neglecting the kinetics of the chemical reactions involved. The purpose of this project is to quantify the mass and energy consumption and the emission of carbon dioxide. Key features are: the selection of the thermodynamic model to achieve a realistic simulation; the heat integration among different process sections to minimize the environmental impact during recovery and recycling of the entrainer; and the reduction of energy duties.

2.3.1.4 Production of Bio-ethanol from Corn

By fermentation of sugar cane, corn or wheat, ethanol (bio-ethanol) can be easily produced. This process includes the steps of liquefaction, cooking, fermentation, distillation, dehydration, evaporation and drying of the solid by-product. By a careful simulation of the operations involved, the needs for water and energy can be minimized, and the use of fossil fuels can be avoided. Key features are the energy balance starting from the content in the feedstock biomass, and the water saving.

A second problem can be addressed by accurately simulating the distillation and dehydration aspects, which have the highest energy demand of this process. Key feature is the use of pressure as an operating variable.

2.3.2 Brief of Phytochemical Processes

2.3.2.1 Citral Recovery by Distillation of Lemon Peel Oil

The goal of this process is the production of citral. Lemon peel oil is fractioned by a traditional method of separation to get an oxygenated substance (citral). A simulation model of the distillation helps identify the optimal operating conditions. The objective of this simulation is to separate the oxygenated compounds from terpenes.

2.3.2.2 Menthol Recovery by Crystallization of Mentha Oil

Mentha oil contains menthol, a commercially important product. Menthol is separated from the other components on the basis of differences in melting temperatures. Crystallization from solution is an industrially important unit operation due to its ability to provide high purity separations.

The crystal growth and nucleation kinetic parameters must be determined experimentally before systematically designing a crystallizer and computing optimal operations and control procedures.

2.3.2.3 Carvone Recovery from Spearmint Oil

Spearmint oil contains the major component carvone that must be separated from other components. The separation of carvone is done by continuous distillation and the process is optimized after identification of the relevant parameters worked out by sensitivity analysis. The objective of this simulation is to obtain carvone at 95% purity or more.

2.3.2.4 Peppermint Oil Extraction by Steam Distillation

For the steam distillation of peppermint, the peppermint leaves are placed at the bottom of a distillation flask and steam is percolated through. The peppermint oil evaporates, and the emerging mixture of vaporized water and oil moves through a coil, usually cooled with running water, where the steam condenses. The mixture of condensed water and essential oil is collected and separated by decantation or, in rare cases, by centrifugation.

2.3.2.5 Multiple-effect Evaporation of Milk Serum

Evaporation is a widely used operation for the recovery of valuable products from dilute aqueous mixtures, such as milk serum. In this case, a four-effect process helps minimize the energy consumption and makes this process economically attractive. Key features are the effects of both pressure and heat transfer coefficients on the overall performance of the apparatus.

2.3.3 Case Study: Turpentine Oil Batch Distillation

Turpentine oil obtained from species of *Pinus* (family Pinaceae) is mainly used in paint and soap manufacturing industries, and in the pharmaceutical industry its use is limited to balms and oil bases. Semi-fluid mixtures of resins remain dissolved in the volatile oil, thus it is produced by fractional distillation. The objectives of the process simulation are to: (i) develop the process simulation base case, (ii) understand how to obtain complete fractionation of the oil, (iii) optimize the composition of pinene, carene and longifolene in the product streams, and (iii) optimize the time and energy consumption of the process.

Figure 3a shows the fractional composition profiles of pinene, carene and longifolene versus time obtained in the top of the distillation column. Figure 3b shows the instantaneous energy consumption of a constant reflux operation.

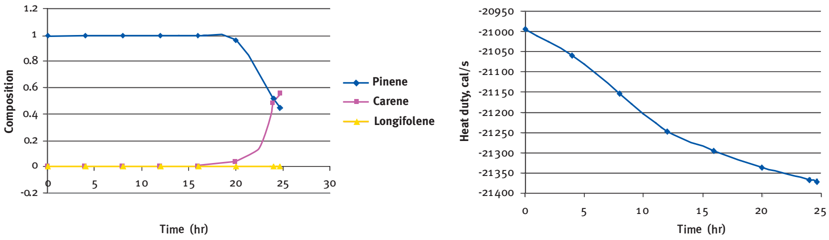

Figure 3: a) Fractional composition versus time, and
b) Energy consumption versus time for the base case

Figure 4 shows the same process in which the reflux ratio is varied in order to obtain a constant composition at the top of the column. It is interesting to note that the distillation time as well as energy consumption are greatly reduced.

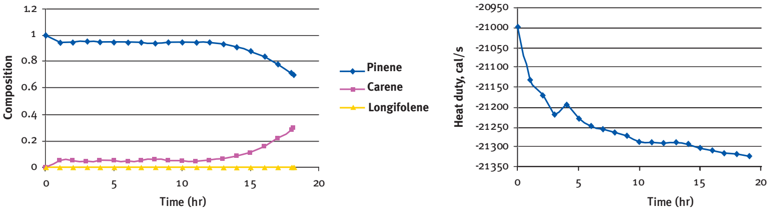

Figure 4: a) Fractional composition versus time, and
b) Energy consumption versus time for the constant composition case

Figure 5 shows that a total separation of the oil constituents is achievable.

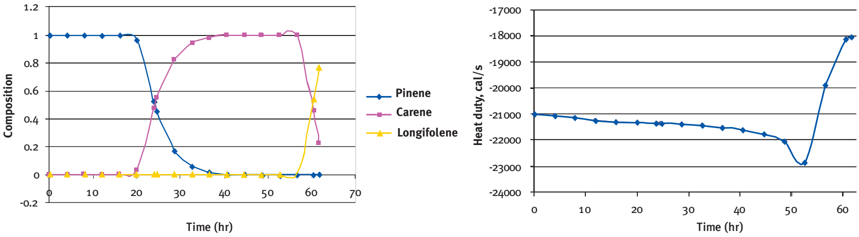

Figure 5: a) Fractional composition versus time, and
b) Energy consumption versus time for the total fractionation case

These simulations show that, in the base case with a reflux ratio of 15 and a high consumption of energy (Figure 4), the distillate accumulator collects a high percentage of pinene (93%). If a PID controller is introduced to maintain the concentration constant, the composition of pinene is around 90% and it takes only 12 h and a reflux ratio of 5 to achieve the desired value, thus saving time and energy. In the case of complete fractionation, we can collect 93% of pinene, 88% of carene and 5.7% of dump products at the end of the process.

2.3.4 Case Study: Menthol Recovery by Crystallization of Mentha Oil

Crystallization from solution is an industrially important operation due to its ability to provide high-purity separations. The menthol crystallization process using menthe oil is rather simple, and consists of a cascade crystallization as shown in Figure 6.

The objective of the simulation is to optimize the menthol crystallization process. The oil, composed of 75% menthol and also containing menthyl acetate, limonene and menthone (Table 1), is fed into the first crystallizer where the temperature is 35° F. The menthol crystals produced here are separated by decantation. The decanted liquid is passed to the second crystallizer and the crystals obtained in this second stage are also separated from the liquid by decantation. Thus, the assumptions made are: (i) limonene is present in all the fractions, and (ii) the separation of solid material from the liquid portion is complete. Furthermore, the thermophysical properties of menthone and menthyl acetate are included in the software's database. The feed stream conditions are: temperature, 80° F; pressure, 1 atm; and flow rate, 50 lb · mol/h.

Table 1: Major constituents of mentha oil

Constituent	Concentration
Menthol	75%
Menthyl acetate	11%
Limonene	8%
Menthone	6%

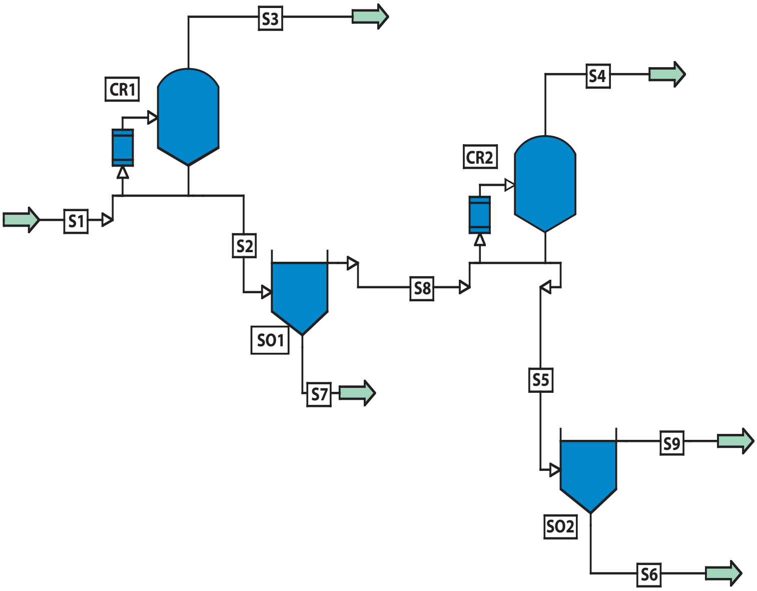

Figure 6: Flowsheet for the menthol crystallization process

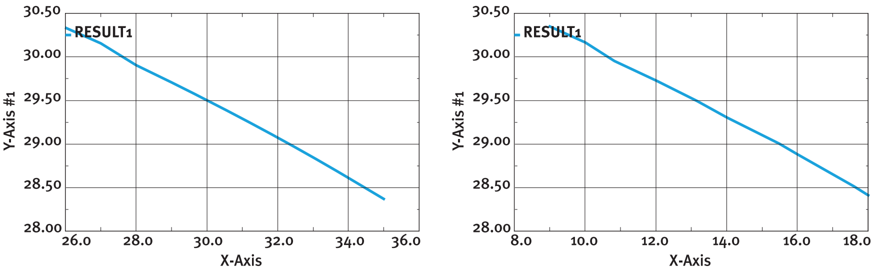

a *b*

Figure 7: Variation of menthol flow rate (Y axis, lb mol/hr) as product of both crystallizers (S6 + S7) versus T (X axis, °F). a) Crystallizer 1. b) Crystallizer 2

Figures 7a and b show on the Y axes the total flow rate (lb mol/hr) of menthol crystal produced in both crystallizers (the combination of S6+S7 of Figure 6). It is evident from the sensitivity analysis that the temperature of the crystallizers has an effect on the total amount of pure product obtained, and consequently on the product yield. In fact, the base case (Table 2) reports a figure of 28.469 lb mol/hr and this amount can be raised linearly if the temperature of the crystallizers is lowered. The sensitivity analysis helps the engineer to select the right temperatures of the crystallizers for a given production.

The material and energy balance of the menthol extraction plant for the base case is listed in Table 2. This is an example of the simulator's output and these values may change if the process conditions are changed. In the first crystallizer, the menthol produced is equal to 5.633 lb mol/hr (purity one) and in the second crystallizer the amount is 22.836 lb mol/hr (purity one). If we compare the total amount of menthol produced by the two crystallizers (22.836 + 5.633 = 28.469 lb mol/hr) with the total amount of menthol fed to the process (37.5 lb mol/hr), we obtain a recovery (amount of pure menthol produced/total amount of menthol fed to the process) of 75.91%, which is rather satisfactory. Table 2 also reports the temperature, pressure, total flow and composition of each single stream considered in the process.

Table 2: Material and energy balance for the menthol crystallization process. The stream names refer to Figure 6

Stream name		**S1**	**S2**	**S5**	**S6**	**S7**	**S8**	**S9**
Phase		**Liquid**	**S/L**	**S/L**	**Solid**	**Solid**	**Liquid**	**Liquid**
Temperature	F	80	35	18	18	35	35	18
Pressure	PSIA	14.696	14.696	14.696	14.696	14.696	14.696	14.696
Molecular weight		159.170	159.170	161.608	156.270	156.270	161.608	163.004
Total	lb mol/hr	50.000	50.000	27.164	5.633	22.836	27.164	21.531
Component mole fractions								
D-limene		0.0800	0.0800	0.1472			0.1473	0.1858
Menthol		0.7500	0.7500	0.5399	1.000	1.0000	0. 5398	0.4195
Menthyl acetate		0.1100	0.1100	0.2025			0.2025	0.2554
Menthone		0.0600	0.0600	0.1104			0.1104	0.1393

2.4 Conclusions

There are two important benefits of the application of process simulation to phytochemical processes of industrial interest.

The first is to improve process knowledge. This is achieved by verifying “in silico” the operating conditions and the estimates of data for intermediate streams, which are difficult to measure. Process knowledge also includes: (i) enthalpy balance information, (ii) verification of plant specifications, (iii) influence of operative parameters on process specifications, (iv) validation of phase equilibrium models for the real system to be used in similar conditions, and finally (v) process de-bottlenecking for each section.

The second important benefit is process optimization, in terms of: (i) consumption of energy and raw materials, (ii) identification of process control strategies, and (iii) clarification of security system behavior when process conditions are varied.

Running process simulation software requires: (i) availability of thermodynamic properties for all components involved, (ii) definition of an accurate thermodynamic model (equations of state or excess Gibbs energy model) for binary and multi-component mixtures, (iii) availability of all necessary interaction parameters, (iv) availability of all necessary unit operation modules, and (v) identification of tear streams to achieve rapid convergence in case of recycles. Furthermore, sometimes it is necessary to define and develop user models and user thermodynamic models.

It is necessary to stress some important principles. First, the program is an aid in making calculations and decisions: the process engineer must ensure that it is “fit for purpose” and is responsible for the results generated and for any use which is made of the results. Second, it is the professional, ethical and legal responsibility of the process engineer to take care and to exercise good judgment. Process simulation is, in essence, a program.

Nonetheless, process simulation is important since it: (i) has high accuracy, (ii) allows one to focus on the interpretation of the results rather than on the methods for obtaining the results, (iii) allows a global vision of the process by assembling theories and models, (iv) is essential in the design of new and existing processes, (v) is essential in the analysis of existing plants in terms of environmental impact, and (vi) is a simple tool for treating real cases.

Process simulation is a well established tool in the chemical industry and has been used for a decade in the petrochemical industry. Process simulation is now applicable in many different fields besides the petrochemical and fine chemical industries, and is particularly interesting for biotechnological and phytochemical processes.

In summary, process simulation may play an important role in the optimization of phytochemical processes and, therefore, application of process simulation can assist in the development of advanced processes. This paper showed that it is possible to achieve energy reduction and maximization of product yield.

Process simulation perhaps is not a good tool for countries that are using primitive techniques, since they should aim at reaching a stable quality of the product rather than an optimization of energy consumption and environmental issues. Moreover, they should focus on practical problems such as the quality of the materials and the water to be used for extraction.

Bibliography

Hazra, P. and Kahol, A. P., 1990, Improved process for production of liquid menthol by catalytic reduction of menthone, *Research and Industry*, 35: 174-176

Prausnitz, J. M., Lichtenthaler, R. N. and Gomes de Azevedo, E., 1998, Molecular Thermodynamics of Fluid Phase Equilibria, 3rd Edition, Prentice Hall

Turton, R., Bailie, R., Whiting, W. and Schweitz, J., 2003, Analysis, Synthesis and Design of Chemical Processes, 2nd Edition, Prentice Hall

3 Maceration, Percolation and Infusion Techniques for the Extraction of Medicinal and Aromatic Plants

J. Singh

Abstract

Techniques of maceration, percolation and infusion have been traditionally used for making galenicals and tinctures from medicinal and aromatic plants (MAPs). This article describes the underlying principals and mechanisms of these extraction techniques, and discusses the various modifications made for the small- and large-scale extraction of MAPs, the factors affecting the extraction process, and the quality of the extracts produced.

3.1 Introduction

Before the nineteenth century, there was no real progress in methods of extraction of plant materials for industrial use. Nonetheless, the various classes of preparations involving simple expression and extraction techniques were in vogue for a long time for the preparation of medicines used in traditional medicine and in complimentary and alternative medicine, practiced throughout the world. The techniques available were limited to expression, aqueous extraction and evaporation; later on, the use of extraction processes was extended by using alcohol as a solvent. Such techniques were highly successful in the phytochemical field and, consequently, single pure molecules were isolated for industrial and medicinal uses. After the nineteenth century, rapid progress was made in extraction processes which led to the isolation and characterization of many groups of plant metabolites of therapeutic importance, including both single chemical constituents as well as standardized extracts of crude drugs. In manufacturing various classes of medicinal plant preparations, such as decoctions, infusions, fluid extracts, tinctures, semisolid extracts (pilular) and powdered extracts, popularly known as galenicals, both simple traditional methods and advanced technologies are used, conforming to the official procedures and specifications as laid down in various pharmacopoeias and codices of the world.

Maceration, percolation and infusion are the general techniques used for the extraction of medicinal plants and are mostly applied for galenical preparations. The sole purpose of such basic extraction procedures is to obtain the therapeutically desirable portion and eliminate the inert material by treatment with a selective solvent known as menstruum. These techniques also play a decisive role in the qualitative and quantitative evaluation of the extracts. The standardized extracts thus obtained are further processed for inclusion in other solid and semisolid herbal dosage forms. These extracts are also used as sources of therapeutically active chemical constituents for various dosage forms of modern medicines.

Historically, galenical preparations were much more extensively used than they are at the present time. Nonetheless, due to resurgence in interest of herbal drugs throughout the world, these extraction procedures are still relevant and are mentioned in official and unofficial monographs about drug preparations. The preparations involving these procedures are primarily intended for extemporaneous dispensing and must be freshly prepared, due to the fact that they rapidly produce a deposit because of coagulation of inert colloidal material and readily support microbial growth due to absence of preservatives.

This article describes the principal methods of extraction by maceration, percolation and infusion as well as the modifications in these procedures for small-scale, official and large-scale extraction. In addition, the paper discusses the choice of extraction method, quality assurance, and factors affecting the extraction procedures.

3.2 General Principles and Mechanisms Involved in Crude Drug Extraction by Maceration, Percolation and Infusion

The general principles and mechanisms involved in maceration, percolation and infusion for the extraction of the crude drugs are same as to those for the extraction of soluble constituents from solid materials using solvent, which is generally referred to as leaching. The processes of leaching may involve simple physical solution or dissolution. The extraction procedures are affected by various factors, namely the rate of transport of solvent into the mass, the rate of solubilization of the soluble constituents by the solvent, the rate of transport of solution out of the insoluble material. The extraction of crude drugs is mostly favored by increasing the surface area of the material to be extracted and decreasing the radial distances traversed between the solids (crude drug particle). Mass transfer theory states that the maximum surface area is obtained by size reductions which entail reduction of material into individual cells. However, this is not possible or desirable in many cases of vegetable material. It has been demonstrated that even 200 mesh particles contain hundred of unbroken cells with intact cell wall. Therefore, it is pertinent to carry out extraction with unbroken cells to obtain an extract with a high degree of purity and to allow enough time for the diffusion of solvent through the cell wall for dissolution of the desired solute (groups of constituents) and for diffusion of the solution (extract) to the surface of the cell wall.

3.3 Factors Affecting the Choice of Extraction Process

The choice of the process to be used for the extraction of a drug depends on a number of factors.

3.3.1 Nature of the Crude Drug

The choice to use maceration or percolation primarily depends upon the nature and characteristics of the crude drugs to be extracted. Therefore, knowledge of the type of organs and tissues of the plant matter is essential for achieving the best result.

3.3.2 Stability of the Crude Drug

Continuous hot extraction procedures should be avoided when constituents of the drug are thermolabile.

3.3.3 Cost of the Crude Drug

When the crude drug is expensive (e.g. ginger), it is desirable to obtain complete extraction. Therefore, from the economic point of view, percolation should be used. For inexpensive drugs, maceration, despite its lower efficiency, is acceptable in view of its lower cost.

3.3.4 Solvent

Selection of the solvent depends on the solubility of the desired components of the material. If the constituents demand a solvent other than a pure boiling solvent or an azeotrope, continuous extraction should be used.

3.3.5 Concentration of the Product

Dilute products such as tinctures can be made by maceration or percolation. For semi-concentrated preparations, the more efficient percolation process is used. Concentrated preparations, such as liquid or dry extracts, are made by percolation.

3.3.6 Recovery of Solvent

Solvent is preferably recovered under reduced pressure to save thermolabile constituents.

3.4 Quality Assurance: the Extraction Process and Solvent

The type of extraction procedure also plays a decisive role in determining the qualitative and quantitative composition of the extract. Some important points regarding the quality of the extracts need to be considered:

i) The more exhaustive the extraction, the better is the yield of the constituents from the herbal drugs.
ii) If maceration is facilitated by stirring and by use of comminuted material, the additional stirring and shearing forces may lead to better extraction.
iii) Other factors determining the quality of the extracts are extraction time, temperature and solvent volume.
iv) Some drugs (e.g. Hypericum spp.) are extracted very slowly so that exhaustive extraction can only be achieved by percolation or multistage motion extraction. In many cases, the transfer of quality-relevant constituents from the herbal drugs to the extract (i.e. extraction rate) can be considerably improved by raising the temperature. Hypericin, pseudohypericin and biapigenin are extracted better at higher temperature and with longer extraction times.
v) The quality of the extracts and the spectrum of constituents obtained by maceration or digestion (i.e. maceration at higher temperature) are also influenced by the ratio of herbal drug to solvent. The quantity of extracted matter increases with the volume of extraction solvent. For example, maceration of Salvia officinalis flowers achieves almost exhaustive extraction and thus the full spectrum of constituents obtained with percolation can be achieved with a drug:solvent ratio of 1:20.
vi) The composition of an herbal extract depends on the type, concentration and elution strength of the solvent. The spectrum of constituents may vary considerably depending on the hydrophilic or lipophilic nature of the solvent.

3.5 Maceration Processes (Steady-state Extraction)

3.5.1 General Procedure

The general process of maceration on a small scale consists of placing the suitably crushed plant material, or a moderately coarse powder made from it, in a closed vessel and adding the selected solvent called menstruum. The system is allowed to stand for seven days, with occasional shaking. The liquid is then strained off and the solid residue, called marc, is pressed to recover as much occluded solution as possible. The strained and expressed liquid thus obtained is mixed and clarified by filtration. Plant material in fine powder form is never used, as it makes subsequent clarification of the extract difficult. In the case of vegetable drugs, sufficient time is allowed for the menstruum to diffuse through the cell wall to solubilize the constituents within the cells and for the resulting solution to diffuse out. As the system is static, except for occasional shaking, the process of extraction works by molecular diffusion, which is very slow. Occasional shaking assists diffusion and also ensures dispersal of the concentrated solution accumulating around the surface of the particles, thereby bringing fresh menstruum to the particle surface for further extraction. A closed vessel is used to prevent evaporation of the menstruum during the extraction period and thus avoids batch to batch variation.

At the end of the maceration process, when equilibrium has been reached, the solution is filtered through a cloth; the marc may be strained through a special press. The concentrations of active constituents in the strained and expressed liquids, sometimes called *miscella*, are the same and hence they can be combined. The expressed liquid may be cloudy with colloidal and small particles, and sufficient time (perhaps several weeks) is necessary for coagulation and settling. The settled matter is filtered through a filter press or any other suitable equipment.

3.5.2 Maceration Process for Organized and Unorganized Crude Drugs

Organized drugs have a defined cellular structure whereas unorganized drugs are non-cellular. Bark and roots are examples of organized crude drugs, while gum and resin are unorganized crude drugs. The processes of maceration for organized and unorganized drugs are slightly different, as shown in Table 1.

Table 1: Four differentiating steps of the maceration process, for organized and unorganized crude drugs

Organized drugs	**Unorganized drugs**
(i) Drug + entire volume of menstruum	(i) Drug + four-fifths of menstruum (in most cases)
(ii) Shake occasionally for 7 days	(ii) Shake occasionally on days 2 to 7, as specified
(iii) Strain liquid, press the marc	(iii) Decant the liquid. Marc is not pressed
(iv) Mix the liquids, clarify by subsidence for filtration. Filtrate is not adjusted for volume	(iv) Filter the liquid and add remaining menstruum through the filter

During maceration of organized drugs, the marc is pressed because a considerable proportion of liquid adheres to it and cannot otherwise be separated. Moreover, the volume is not adjusted because a variable amount of liquid containing soluble matter is left in the marc. If the volume is adjusted, a weak product will result. Omitting adjustment, the volume of liquid expressed influences the product yield and the percentage of soluble matter, regardless of the efficiency with which the marc is pressed in a hand press, screw press or hydraulic press; the strength of the product is not affected. Preparations made by this processes include vinegar of squill (British Pharmaceutical Codex, BPC), oxymel of squill (BPC), tincture of orange (Indian Pharmacopoeia, IP), tincture of capsicum (BPC), compound tincture of gentian, tincture of lemon, and tincture of squill (BPC).

In maceration of unorganized drugs, the marc is not pressed because the desirable material is mostly dissolved and the remaining marc is gummy and slimy. Thus, it is neither practicable nor necessary to press it. Moreover, the volume is adjusted because the clear upper layer is easily separated by filtration from the lower layer. The solution contains practically all the soluble matter of the drug; the small amount adherent to the gummy matter is recovered when the marc is washed by menstruum in the filter. Therefore, adjustment of volume leads to uniformity. Preparations made by this process include compound tincture of benzoin, tincture of myrrh (BPC), and tincture of tolu (BPC).

3.5.3 Modifications to the General Processes of Maceration

Repeated maceration may be more efficient than a single maceration process, as described earlier, because an appreciable amount of active principle may be left behind in the first pressing of the marc. Double maceration is used when the active constituents are particularly valuable and also when the concentrated infusions contain volatile oil. Where the marc cannot be pressed, a process of triple maceration is sometimes employed. The total volume of solvent used is, however, large and the second and third macerates are usually mixed and evaporated before being added to the first macerate.

3.5.4 Large-scale Extraction Procedures

For large-scale, industrial extraction, certain modifications are warranted. When the extraction vessel contains a small amount of solvent (500-1000 ml), occasional shaking is no problem. But, for industrial work where a large amount of solvent and huge vessels are involved, shaking the vessels is difficult. Obviously, there are alternative methods of agitation that are just as effective and much simpler to put into practice. In addition, economics become increasingly important and one of the most important objectives is to improve the efficiency of extraction so that less solvent is needed and evaporation requirements for concentrated products are reduced. Reducing the cost of evaporation has the further advantage of minimizing the heat damage to thermolabile constituents. Some of the modified maceration procedures used for large-scale extraction are described in the next paragraphs.

3.5.4.1 Circulatory Extraction

The efficiency of extraction in a maceration process can be improved by arranging the solvent to be continuously circulated through the drug, as indicated in the Figure 1. Solvent is pumped from the bottom of the vessel (through an outlet) and is distributed by spray nozzles over the surface of the drug. The movement of the solvent reduces boundary layers and the uniform distribution minimizes local concentration in a shorter time.

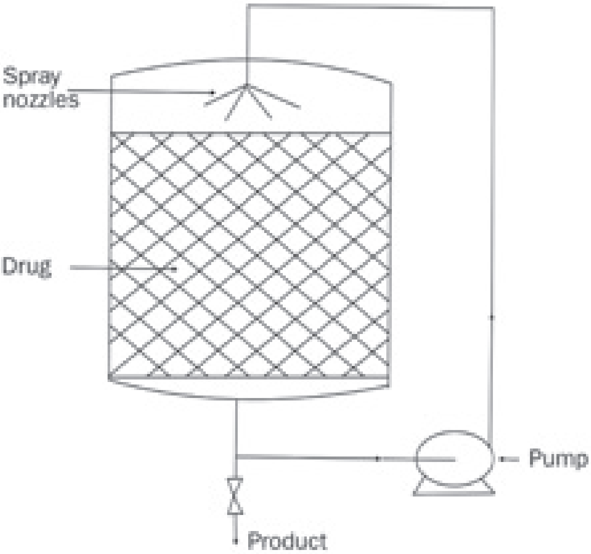

Figure 1: Circulatory extraction

3.5.4.2 Multistage Extraction

In the normal maceration process, extraction is incomplete, since mass transfer ceases when equilibrium is reached. This problem can be overcome using a multistage process. The equipment needed for this method is a vessel for the crude drug, a circulating pump, spray distributors and a number of tanks to receive the extracted solution. The extractor and tanks are connected with piping and valves as shown in Figure 2, so that any of the tanks may be connected to the extractor for transfer of the solution. Each batch of drug is treated several times with solvent and, once a cycle is in process, the receivers contain solution with the strongest in receiver 1 and the weakest in receiver 3.

3.5.4.2.1 Advantages

The crude drug is extracted as many times as there are receivers (in Figure 2 there are three receivers). If more extraction stages are required, it is only necessary to have more receivers. The last treatment of the drug – before it is discharged – is with fresh solvent, giving maximum extraction. The solution is in contact with fresh drug before removal for evaporation, giving the highest possible concentration.

3.5.4.2.2 Procedure

Fill the extractor with crude drug, add solvent and circulate. Run off to receiver 1. Refill the extractor with solvent and circulate. Run off to receiver 2. Refill the extractor with solvent and circulate. Run off to receiver 3. Remove drug from the extractor and recharge. Return solution from receiver 1 to the extractor. Remove for evaporation. Return solution from receiver 2 to the extractor and circulate. Run off to receiver 1. Return solution from receiver 3 to the extractor and circulate. Run off to receiver 2. Add fresh solvent to the extractor and circulate. Run off to receiver 3. Remove drug from the extractor and recharge. Repeat cycle.

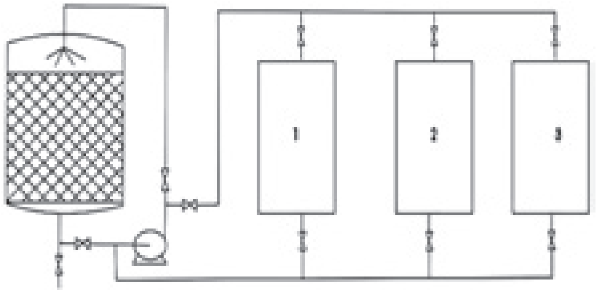

Figure 2: Multistage extraction

3.5.4.2.3 Extraction Battery

In the normal percolation process, the percolate is a very dilute solution, while the ideal situation is to obtain the maximum concentration possible. Continuous extraction devices of battery type are used when large amounts of a single material are handled. Such devices can be achieved by treating percolation as a multistage process. In an extraction battery process, a series of vessels is used and extraction is semicontinuous.

3.5.4.2.4 Equipment

An extraction battery consists of a number of vessels with interconnecting piping. Vessels are so arranged that solvent can be added to and the product taken from any vessel. These vessels can, therefore, be made into a series with any of vessels as the first of the series. The use of an extraction battery is illustrated in Figure 3, which shows the simplest arrangement of three vessels.

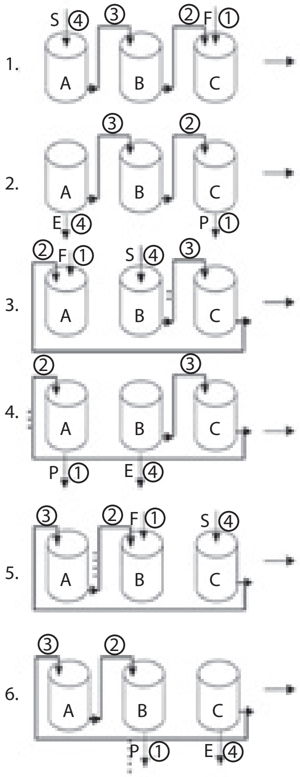

1. Charge fresh drug to C.
2. Transfer solution from B to C.
3. Transfer solution from A to B.
4. Add fresh solvent to A.

1. Remove product solution from C.
2. Transfer solution from B to C.
3. Transfer solution from A to B.
4. Dump exhausted drug from A.

1. Charge fresh drug to A.
2. Transfer solution from C to A.
3. Transfer solution from B to C.
4. Add fresh solvent to B.

1. Remove product solution from A.
2. Transfer solution from C to A.
3. Transfer solution from B to C.
4. Dump exhausted drug from B.

1. Charge fresh drug to B.
2. Transfer solution from A to B.
3. Transfer solution from C to A.
4. Add fresh solvent to C.

1. Remove product solution from B.
2. Transfer solution from A to B.
3. Transfer solution from C to A.
4. Dump exhausted drug from C.

Arrows indicate transfer of material to or from the vessels. An asterisk(.) is added after each contact stage. Key: F = fresh drug. S = solvent. P = product. E = exhausted drug. A, B, C, extraction vessels. Start with C empty and units A and B containing drug, with A more nearly exhausted.

Figure 3: Extraction battery

3.6 Percolation (Exhaustive Extraction)

3.6.1 General Process of Percolation

In this process, an organized vegetable drug, in a suitably powdered form, is packed in a percolator and the solvent is allowed to percolate through it. Although some materials (e.g. ginger) may be packed directly into the percolator in a dry state, this may cause difficulties with other drugs. With the addition of solvent, the dry material swells and this swelling increases with increasing aqueous nature of the solvent. This swelling reduces or blocks the flow of the solvent, thus seriously affecting the extraction process. Furthermore, if the dry powder is packed, fine particles may be washed down the column and settle at the lower levels, reducing the porosity drastically, blocking the column and making the column nonuniform. The finer particles may even be washed out of the column. These difficulties can be prevented by a preliminary uniform moistening of the raw material with the menstruum for a period of 4 h in a separate closed vessel; this process is called imbibition. During this period, the crude drug is allowed to swell to the maximum extent. Hence, when aqueous solvents are used for extraction, more menstruum is needed during imbibition. Also, the occluded air in the drug powder is replaced by the vapor of the solvent, thereby enabling the material to be more evenly packed and allowing the menstruum to flow more uniformly. Uneven packing permits more solvent to pass through channels offering less resistance to the flow of the solvent, thus resulting in inefficient extraction. After imbibition, the drug is packed evenly into the percolator. The imbibed drug is packed over a loose plug of tow or other suitable material previously moistened with the solvent. Even packing can be achieved by introducing the material layer by layer and pressing it with a suitable implement to give even compression; the pressure exerted on the material depends on the nature of the material and its permeability. After packing is over, a piece of filter paper is placed on the surface followed by a layer of clean sand such that the top layer of the drug is not disturbed when solvent is added for extraction. Sufficient menstruum is now poured over the drug slowly and evenly to saturate it, keeping the tap at the bottom open to allow the occluded gases between particles to pass out. Menstruum should never be poured with the tap closed since the occluded air will escape from the top, disturbing the bed. When the menstruum begins to drip through the tap, the tap is closed; sufficient menstruum is added to maintain a small layer above the drug and allowed to stand for 24 h. The layer of menstruum above the surface of the bed prevents drying of the top layer, which may result in the development of cracks on the top surface of the bed. The 24-h maceration period allows the solvent to diffuse through the drug, solubilize the constituents and leach out the soluble material. In this way, the extraction is more efficient than carrying out percolation without the maceration period.

After the maceration, the outlet is opened and the solvent is percolated at a controlled rate with continuous addition of fresh solvent. The volume of percolate collected depends on the nature of the final product. In general, about 75% of the volume of the finished product is collected, the marc is pressed and the expressed liquid is added to the percolate, giving about 80%-90% of the final volume. After assay, the volume is adjusted with calculated quantities of fresh menstruum. If no assay is available, the volume is adjusted after adding the other constituents, if any. In percolation, the expressed liquid is devoid of active constituents as they are already extracted during the percolation period; pressing the marc is only to recover the valuable solvent. This is in contrast with maceration in which the marc is pressed.

3.6.2 Modifications to the General Process of Percolation

In the general process of percolation, particularly in the manufacture of concentrated preparations like liquid extracts, the following problems may arise:

- If the active substances are thermolabile, evaporation of large volumes of dilute percolate may result in partial loss of the active constituents.
- In the case of alcohol-water mixtures, evaporation results in preferential vaporization of alcohol, leaving behind an almost aqueous concentrate. This may not be able to retain the extracted matter in solution and hence the substances may precipitate.

In such cases, the general process of percolation is modified, as described in the next paragraphs.

3.6.2.1 Reserved Percolation

In this case, extraction is done through the general percolation procedure. At the end, the evaporation is done under reduced pressure in equipment like a climbing evaporator to the consistency of a soft extract (semisolid) such that all the water is removed. This is then dissolved in the reserved portion, which is strongly alcoholic and easily dissolves the evaporated portion with any risk of precipitation.

3.6.2.2 Cover and Run Down Method

This is a process that combines the maceration and percolation techniques. This process cannot be used for materials that contain volatile principles or for those which undergo change during the evaporation stage. This procedure is advantageous because industrial methylated spirit may be used for extraction instead of the costly rectified spirit.

The detailed procedure is as follows. After the imbibition stage, the material is packed in a percolator and macerated for a few hours with a suitable diluted industrial methylated spirit. Then, the liquid is run off and the bed is covered with more menstruum. Maceration is done as before and the second volume of the extract is collected. This process is repeated several times with the later weaker extracts used for extraction of a fresh batch of the drug. More concentrated fractions are evaporated under reduced pressure to eliminate the toxic methanol. After the concentrate is assayed for the active principle or for total solids content, it is diluted with water and ethanol to obtain the correct concentration of alcohol and active principle.

3.6.3 Percolators

Different types of percolators are used for small- and large-scale extraction.

3.6.3.1 Small-scale or Laboratory-scale Extraction

The processes for the manufacture of concentrated preparations, maceration and percolations, are involved in extraction followed by the evaporation of solvents. The two operations are combined in a continuous extraction process.

The general procedures and apparatus used for small- or laboratory-scale extraction are described in the following paragraphs.

3.6.3.1.1 Soxhlet Apparatus

On the laboratory scale, the Soxhlet apparatus is used. It consists of a flask, a Soxhlet extractor and a reflux condenser. The raw material is usually placed in a thimble made of filter paper and inserted into the wide central tube of the extractor. Alternatively, the drug, after imbibition with menstruum, may be packed in the extractor taking care that the bottom outlet for the extract is not blocked. Solvent is placed in the flask and brought to its boiling point. Its vapors pass through the larger right hand tube into the upper part of the extractor and then to the condenser where they condense and drop back onto the drug. During this period, the soluble constituents are extracted. When the level of the extract reaches the top of the syphon tube, the entire volume of extract syphons over into the flask. The process is continued until the drug is completely extracted. The extract in the flask is then processed. This procedure is thus a series of short macerations.

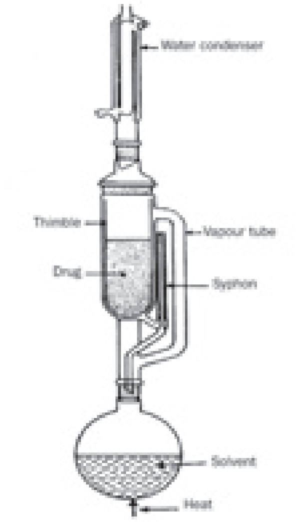

Figure 4: Soxhlet apparatus for hot extraction

3.6.3.1.2 Official Extractor

An official extractor, as described in the official monographs (e.g. IP, British Pharmacopoeia), is illustrated in Figure 5. In this case, the extraction is a continuous percolation procedure. In this apparatus, vapors rise through the extraction chamber passing the drug container; the vapor condenses in the reflux condenser and returns through the drug, taking the soluble constituents to the flask.

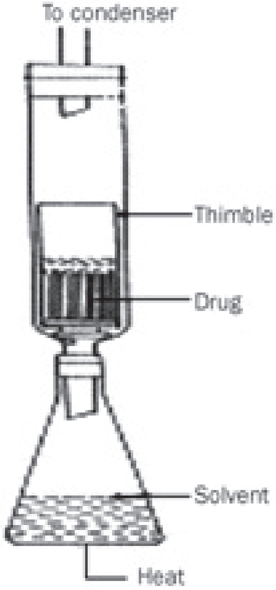

Figure 5: Apparatus for the continuous extraction of drugs

3.6.3.2 Limitations of the Official Extractor

It is not useful when the raw material contains thermolabile active constituents, because the extraction is carried out at an elevated temperature and the extract in the flask is also maintained in the hot condition until the process is complete. It can be used only with pure solvents or with solvent mixtures forming azeotropes. If an ordinary binary mixture is used as the menstruum, the composition of the vapor will be different from the liquid composition.

Similar methods can be used in large-scale production.

A typical industrial set-up for continuous extraction is shown in Figure 6. The principle of operation resembles that of the laboratory equipment.

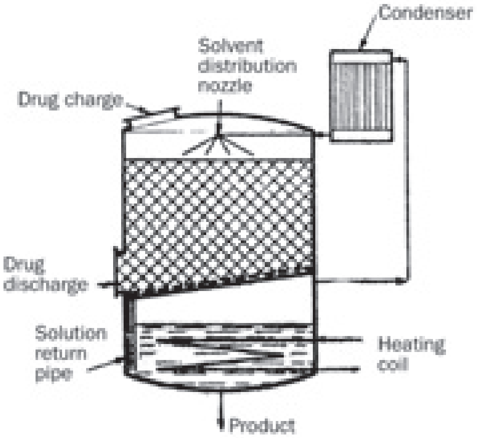
Condenser
Solvent distribution nozzle
Drug charge
Drug discharge
Solution return pipe
Heating coil
Product

Figure 6: Continuous extraction: large-scale plant

3.6.3.3 Large-scale Extractor

The drug is supported on a perforated metal plate covered with a layer of sacking or straw. The percolator has a removable lid which contains portholes for running the solvent in and for observing the flow of solvent. The outlet from the percolator is fitted with a tap and pipeline. This outlet permits removal of the percolate for subsequent processing or for use as a menstruum in a second percolator in series, resulting in more efficient use of the menstruum by carrying out the extraction in a counter-current manner. On the small scale, copper percolators were originally used but these are now largely replaced with percolators made of glass or stainless steel. A slightly conical percolator is better than a cylindrical one since the sloping sides permit the eventual expansion of the bed and also allow the solvent to permeate the material present near the sides at the bottom, which is a problem with a cylindrical percolator.

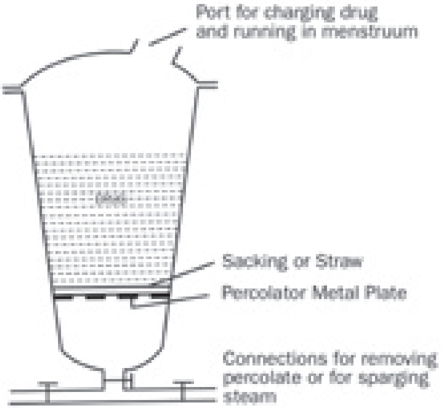
Port for charging drug and running in menstruum
Sacking or Straw
Percolator Metal Plate
Connections for removing percolate or for sparging steam

Figure 7: Commercial scale percolator

3.7 Infusion

3.7.1 General Considerations

Infusions are dilute solutions containing the readily soluble constituents of crude drugs. Formerly, fresh infusions, prepared by macerating the drug for a short period in cold or boiling water, were used and diluted to eight volumes but, now, infusions are usually prepared by diluting one volume of a concentrated infusion to ten volumes with water. Concentrated infusions are prepared by a modified percolation or maceration process. After dilution with water, concentrated infusions resemble in potency and aroma the corresponding fresh infusion. Infusions are liable to fungal and bacterial growth, and it is necessary to dispense them within 12 h of their preparation.

3.7.2 General Method for Preparing Fresh Infusions

The coarsely powdered crude drug (50 g) is moistened, in a suitable vessel with a cover, with 50 ml cold water and is allowed to stand for 15 min. Then, 900 ml boiling water is added, and the vessel is covered tightly and allowed to stand for 30 min. The mixture is strained and enough water is passed to make the infusion measure 1000 ml. Some drugs are supplied (accurately weighed) in muslin bags for preparing specific amounts of infusion and as such are used for infusion preparation. If the activity of the infusion is affected by the heat of the boiling water, cold water should be used. As fresh infusions do not keep well, they should be made extemporaneously and in small quantities at the time of use.

3.7.3 Preparation of Concentrated Infusions

The official monographs also recognize certain “concentrated infusions” in which 25% alcohol is added during or after the infusion process and then diluted as per pharmacopoeial (official) requirement. Concentrated infusions are especially prepared in cases in which the active and desirable principles of drug are equally soluble in water and in the menstruum used for both concentrates and infusions.

3.8 Evaporation

One quality-relevant parameter is the evaporation of the eluate from the soft extract. The state of art is cautious vacuum evaporation, in which evaporation temperatures do not exceed 55° C. The temperature in relation to the evaporation time is of special importance for quality of this step, especially if the extract contains volatile or thermolabile constituents.

3.9 Conclusions

The spectrum of constituents obtained by steady-state extraction (simple maceration) differs from that obtained by exhaustive extraction (percolation). With maceration, one can achieve a spectrum of constituents similar to that of percolation. Different extraction procedures may be considered to be equivalent if they respect critical quality parameters and if the analysis of numerous production batches confirms their compliance with standards.

Bibliography

Anonymous, 1955, Indian Pharmacopoeia, the Manager of Publication, Delhi, p. 273

Anonymous, 1973, British Pharmaceutical Codex, the Pharmaceutical Press, London, p. 703-704

Anonymous, 1980, British Pharmacopoeia, VOL. II, University Press, Cambridge, London, p. 576

Anonymous, 2002, Bentley's Text book of Pharmaceutics EA Rawlins (Ed.). Reprint, Bailliere Tindall, London/All India Traveller Book Selter, New Delhi

Cooper, J. W. and Gunn, C., 1975, Tutorial Pharmacy, S. J., Carter Reprint, CBS Publication, Delhi, p. 251- 261

Cooper, J. W. and Gunn, C., 1985, General Pharmacy, CBS Publishers and Distributors Delhi, p. 308-333

Evans, W. C., 1998, Trease and Evan's Pharmacognosy (14th Edition), W. B. Saunders Company Limited, London, p. 119

Sambamurthy, K., 2002, Pharmaceutical Engineering Reprint, New Age International (P.) Ltd., New Delhi, p. 173-194

Singh, J., Bagchi, G. D., and Khanuja, S. P .S., 2003, Manufacturing and quality control of Ayurvedic and herbal preparations, In: Verpoorte, R. and Mukherjee, P. K.(Eds), GMP for Botanicals, Regulatory and Quality Issues on Phytomedicine (1st Edition), Business Horizons, New Delhi, p. 201-230

Waldesch, F. G., Konigswinter, B. S., and Blasius, H., 2003, Herbal Medicinal Products, Medpharm, Stuttgart, Germany and CRS Press, London, p. 48-54

4 Hydrolytic Maceration, Expression and Cold Fat Extraction

A. K. Singh

Abstract

The incorporation of bioactive ingredients without loss of activity into foods, flavors, pharmaceuticals, pesticides and cosmaceutical products is very important. Extraction of active constituents from raw materials is an important and critical step in maintaining bioactivity. A number of methods are available for extraction, and these are selected in such a way that the activity of the phytoconstituents is retained. This paper discusses the processes of hydrolytic maceration, expression and cold fat extraction.

4.1 Introduction

The extraction of active constituents from plants is one of the most critical steps in the development of natural products for commercial use. The simplest example of extraction may be brewing a cup of coffee, wherein caffeine and tannins are extracted from coffee beans in hot water. All living organisms contain complex mixtures of chemicals, usually held within cellular structural material (protein, lipid, polysaccharides etc.) of which some are desired while others are not. Thus, taking out the desired part from the whole crude drug is referred to as extraction and it is done in solvents where ingredients move from one phase to another.

A number of methods are available for extraction and the choice among them is dictated by the physicochemical properties and stability of the phytoconstituents to be obtained. For the extraction of essential oils, the simplest methods are hydrodistillation and steam distillation while other methods also employed are cold fat extraction, expression, maceration and solvent extraction. Nowadays, more advanced technologies are used, such as supercritical fluid extraction, solid phase micro-extraction and phytonic extraction. The present article deals with extraction by hydrolytic maceration, expression and cold fat extraction.

4.2 Hydrolytic Maceration

The word maceration is derived from the Latin word *maceratus,* which means to soften. In reference to medicinal and aromatic plants, maceration refers to the preparation of a solution by soaking plant material in vegetable oil or water. Maceration methods are based on the immersion of crude drug in bulk solvent, while percolation methods depend on the flow of solvent through the powdered drug. The rate of extraction depends upon the following:

- The rate of transport of solvent into the mass to be leached.
- The rate of solubilization of soluble constituents into solvent.
- The rate of transport of solute out of the insoluble material and from the surface of the insoluble material into the solution.

In the process of maceration, the powered solid material is placed in a closed vessel (to prevent evaporation) and the chosen solvent (menstruum) is added. It is allowed to stand for a long time (varying from hours to days) with occasional shaking. Sufficient time is allowed for the menstruum to diffuse through the cell wall to solubilize the constituent present in the cells and for the resulting solution to diffuse out. The process takes place only by molecular diffusion. After the desired time, the liquid is strained off; the solid residue (marc) is pressed to recover as much solvent as possible. When the menstruum is water and the period of maceration is long, a small quantity of alcohol may be added to prevent microbial growth.

Cold maceration of crushed grapes is done at room temperature before the onset of fermentation. In this process, the skin and seeds are permitted to soak for one to two days prior to the initiation of fermentation to get more aqueous extraction without the effect of ethanol on grape cells. In certain cases, hydrolysis is done with a suitable agent (enzyme) prior to maceration, e.g. maceration of wine. The quality of wine is judged by its appearance, color, aroma, taste (mouth feel) and flavor. Grape-derived aroma and flavor precursors exist partially as non-volatile, sugar-bound glycosides. Hydrolysis modifies sensory attributes and potentially enhances wine quality. Flavored aglycones potentially affect the wine quality after hydrolysis. Cell wall-degrading enzymes help in the release of grape aroma, when cold maceration prior to fermentation has been carried out.

4.3 Expression Extraction of Essential Oils

Most essential oils are isolated from the respective plant parts by the process of hydrodistillation or steam distillation. A few essential oils such as those present in the citrus fruit peel can be, and in large part are, obtained by pressing, which yields a product of superior quality. The long action of heat with boiling water affects some thermolabile constituents which may decompose due to hydrolysis, polymerization and resinification. Therefore, the essential oil obtained by distillation does not represent the natural oil as it originally occurred in the plant material. In such plant materials, essential oil is extracted by the process of expression or solvent extraction.

Before dealing with the process of expression in citrus oil, it is important to review the structure of a citrus peel. The citrus peel contains numerous oval, balloon-shaped oil sacs and glands or vesicles (0.4-0.6 mm diameter). These are ductless glands which are irregularly distributed in the outer colored (flavedo) portion of the peel of maturing and matured fruit, in the outer mesocarp beneath the epicarp and hypoderm and above the inner mesocarp (albedo), respectively. Albedo is made of cellulose, hemicellulose, lignin, pectin, sugars, glycosides, etc. On maturation, cells of albedo become elongated and branched, with large intracellular spaces which give the ripe peel its spongy texture. The spongy layer plays an important role in the expression of oil, but it easily absorbs the oil ejected from the sacs and causes some mechanical difficulties in oil extraction. The freshness and stage of maturity affect the oil ejection from the peel. The total weight of the peel is about half that of the fruit and the oil content is 0.5%-0.7%. The cells surrounding the oil sac contain salt (colloids) in aqueous solution. In contact with water, the higher osmotic pressure of the cell results in diffusion of water into them, increasing turgor pressure and causing the oil sac to stress from every side. If the spongy tissue is not filled with water, it will absorb the oil as the sacs are broken and hold it with great tenacity. So the pressure exerted yields first aqueous fluid and later oil.

In Sicily, peels are immersed in water for several hours before being subjected to a “sponge” process. A dilute aqueous solution of salt acts as a carrier to prevent loss by spurting. The product of expression is not a simple mixture of oil and water but a thin emulsion which is left to stand; gradually a supernatant layer forms. Filtration through a sponge absorbs the colloidal material and leaves the mixture of oil and water.

4.3.1 Process of Expression

4.3.1.1 Hand Process

In this process, the freshly harvested fruits are cut transversely into two halves. The pulp is removed with a sharp-edged spoon called *rastrello*. The peel is then immersed in water for several hours and finally pressed by hand. The worker holds, with the left hand, either one large or two smaller flat sponges on top of a wooden crossbar and, with the right hand, presses the peel against the upper sponge. Thinner peels can be pressed from inside. The emulsion ejected from the oil sacs is soaked by the sponge, which retains solid matter and absorbs colloidal substances. Squeezing the content of the sponge from time to time, oil is finally decanted and drawn off. This process requires much labor and the yield of this method is 50%-70% of the total oil present in the peel. The quality of oil obtained from hand pressing is near to the quality present in fruit peel. A large number of small units in Sicily and Calabria were formerly responsible for the entire Italian production of lemon and orange oil, but the process is not in use now.

4.3.1.2 Ecuelle Process

This process was common in the south of France. Ecuelle consists of a shallow bowl of copper with a hollow central tube with which it forms a sort of funnel. The bowl is equipped with large brass nails with blunt ends, on which the fruit is rolled by hand with some pressure until the entire surface of fruit yields its oil. The oil and aqueous cell contents drain into the central tube where they are separated by decantation. The yield is only 20% of the total oil present in the peel.

These methods were quite laborious and, with the advance of technology, machines have been invented to do them. Nowadays, rinds are extracted for oil using centrifugal force.

4.3.1.3 Hand Machine

For expression, the peel is placed in a hollow sponge attached with other sponges to a plate actuated by the lever and fixed below with additional sponges to the base. The sponge is fitted with a funnel through which oil and aqueous material pass to the receiving vessel. The part that comes in contact with oil is made of brass or bronze.

4.3.1.4 Sfumatrici and Pelatrici

The machines that treat only the peel after removal of juice and pulp are called *sfumatrici*, while those that process the whole fruit are known as *pelatrici*.

4.3.1.4.1 Special Sfumatrice

This is a specially designed roller type machine in which each peel is bent to expel the maximum quantity of oil. Not much pressure is exerted to expel other contents of the cells. The emulsion is collected and filtered through wool or sponges to yield oil and water. A number of *sfumatrici* have been developed and were in use in many countries.

In expression using *sfumatrici*, two approaches are used. In the first approach (used by Ramino Sfumatrice), only the peel is treated, so halving the fruit, removing the pulp and expressing the peel are the steps involved. In the second approach, the oil is extracted by either puncturing the peel glands or cutting a superficial layer of peel to expose the oil glands; this is followed by washing away the oil with a spray of water.

In whole fruit extraction with a rasping machine, the whole fruit is crushed and oil is subsequently separated from the aqueous phase (juice and cell liquid). As citrus peel oil is a byproduct of the citrus juice industry, both the oil and juice are extracted subsequently, e.g. in a rotatory juice extractor or Pipkin peel oil press. Here, the fruit is washed automatically, sorted according to size and then cut in two halves. The halved fruit is passed between two cylinders which remove only the juice by gentle pressing without affecting the oil glands on the surface. The residual halved peels are extracted for essential oil. In the Pipkin peel oil press, for example, oil is expressed by two close stainless steel cylinders with capillary grooves running around the circumference. The expressed oil automatically comes out of the grooves and there is no need to spray water..

4.3.1.5 Modern Machines

Nowadays, complete mechanization has been achieved and whole fruit processing machines have been developed. These machines either crush whole fruit and then separate the oil from the aqueous phases by distillation or centrifugation, or they express the oil in such a way that it does not come in contact with the juice during the process. The oil extractor developed by Brown International, California, liberates the essential oil from whole citrus fruit. Oil removal is achieved by lightly puncturing the entire surface of the fruit with over three million sharp stainless steel points configured in the form of rotating rolls. An adjustable speed differential between adjacent rolls controls the amount of pressure applied to the fruit. Whole fruits roll across the brown oil extractor (BOE, Figure 1), which is made of toothed rollers partially set in a flowing bed of water which propels the fruit across the machine and simultaneously punctures the peel to release the oil from the glands. The fruit proceeds to the extractor and the oil-water mixture goes to centrifugation and to the oil recovery chamber.

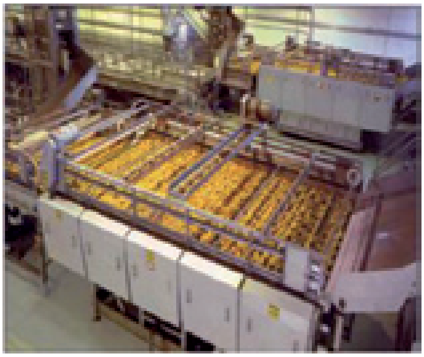

Figure 1: Brown oil extractor

4.3.1.6 FMC Whole Fruit Extractor

Food Machinery Corporation (FMC) of San José, California, has developed many designs and improved extractors in which both the juice and the volatile oil are extracted without getting mixed with each other. Further details are available at: http://files.asme.org/ASMEORG/Communities/History/Landmarks/5549.pdf.

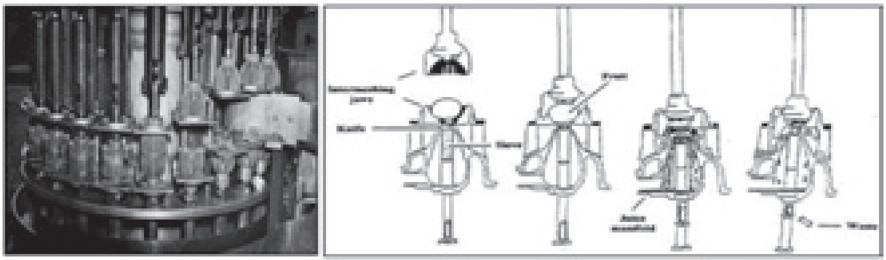

Figure 2: FMC whole fruit extractor

The FMC juice extractor has two inter-meshing jaws which encompass the fruit, crushing it between them. The juice exits through a mesh screen which penetrates the center of the fruit and the juice is separated from the peel, pith and seeds. This crushing action is sufficient to force the oil out of the glands on the surface of the flavado. The FMC machine sprays water onto the surface of the fruit, and the oil-water emulsion is subjected to centrifugation. Traces of water and waxy material are separated by keeping low temperature. About 75% of citrus oil production utilizes this technology. The odor of cold pressed oil is true in nature and similar to that of the oil present in fruit.

4.4 Cold Fat Extraction (Enfleurage)

Certain high-quality odor-producing flowers such as jasmine, tuberose and gardenia yield small quantities of oil and cannot be distilled by hydrodistillation. Furthermore, the oil components are thermolabile and such flowers, even after plucking, continue to emit small quantities of perfume. The oil from these types of flowers is extracted by cold fat extraction, i.e. enfleurage.

Figure 3: Chassis for holding the fat crops

Fat possesses a high power of absorption of volatile oil and, if brought in contact with fragrant flowers, it absorbs the perfume. This principle methodically applied on large scale constitutes enfleurage. The quality of the fat base is important for the quality of the flower oil. It must be odorless and of proper consistency. If the crops (fat) is too hard, the blossoms will not have sufficient contact with it, the absorption will be insufficient, and the quality and yield will be poor. If the crops is loose, it may engulf the flowers so that exhausted ones are difficult to remove and retain adhering fat when removed. The crops must have a consistency suitable to produce a semihard surface from which exhausted flowers can be removed easily. Saturated fragrant fat extract is known as pomade. Enfleurage process of cold extraction is carried out in Bulgaria, Egypt, Algeria, Sicily (Italy), and Grasse (France). France remains the main centre of production of highly prized so-called natural flower oil. Natural flower oil does not include the distilled essential oil but applies only to flower oils obtained by enfleurage, maceration or solvent extraction. The whole process is carried out in cold atmosphere cellars (cold rooms). The mixture of one part highly purified tallow and two parts lard is best suited for enfleurage.

Different bases have been used in the preparation of crops. For example, vegetable fat, mineral oil, esters of polyhydric aliphatic alcohol (ester of glycol, glycerol, mannitol, hexitol), and fatty acids of high molecular weight have been tried, but the best results were from the old-fashioned mixture of lard and tallow. To purify tallow, it is melted after cleaning, washing and removing blood and muscles; the skin is beaten and cleaned, and benzoin (0.6%) and alum (0.15%-0.30%) are added. Benzoin is a preservative while alum causes impurities to coagulate during heating. Warm melted fat is filtered through cloth and left to cool.

The vehicle for holding the fat crops during the process is a specially designed “chassis”. These are rectangular wooden frames (2 in. high, 20 in. long and 16 in. wide) that hold a glass plate (Figure 3). The fat crops is applied to both sides of the glass with a spatula at the beginning of enfleurage: 360 g crops on each side is required. When piled one above the other, the chassis form an air-tight compartment with a layer of fat on the upper and lower sides of each glass (Figure 4). Every morning, freshly picked flowers are cleaned by removing the leaves and stalks and eliminating the moisture from dew or rain. The flowers are strewn (Figure 3) by hand on top of the fat layer of each glass plate. Traces of moisture will cause the crops to go rancid, so precaution must be taken.

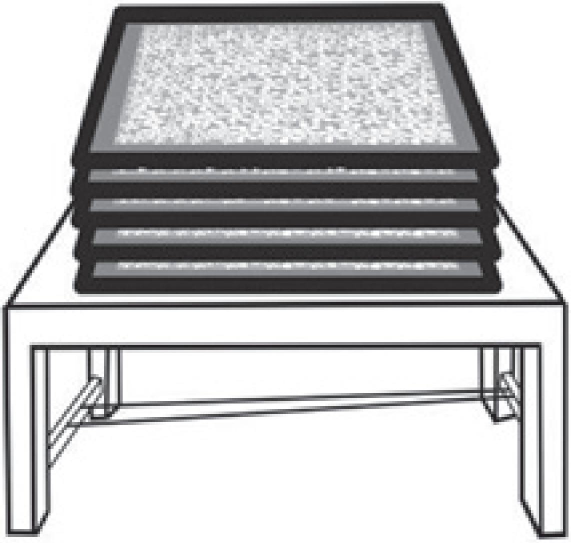

Figure 4: Pile of chassis

When chassis are piled one above the other, flowers remain in contact with the lower fat crops which acts as direct solvent, whereas the upper fat layer (beneath the glass plate of the chassis) absorbs only the volatile perfume given off by the flowers. After 24 h, the flowers are removed from the chassis. The time of removal, however, depends on type of the flowers. Removal of the flowers from the crops is known as defleurage. Immediately following defleurage, the chassis are recharged with fresh flowers. For this purpose, the chassis are turned over, thus the fat layer which in the previous operation formed on top (ceiling) of the small chamber is now directly charged with fresh flowers. When the fat crops becomes saturated with fragrance (pomade), the crops is removed from the chassis with a spatula. The pomades are extracted with absolute alcohol in a process called *extrait*. At the beginning and several times during the harvest, the fat on the chassis is scratched over with a metal comb and tiny furrows are drawn in order to change and increase the surface area available for maximum absorption. The most highly saturated pomade is pomade no. 36, meaning the crops on the chassis has been treated with fresh flowers 36 times during the whole process of enfleurage. In the case of jasmine, every kilogram of fat is charged with 3 kg flowers during the entire period. The alcohol washing of pomade no. 36 are called extrait no. 36, which reproduces the natural flower oil to a remarkable degree. Sometimes a fatty note is present in the extrait, which can be removed by freezing and filtering the alcoholic washes.

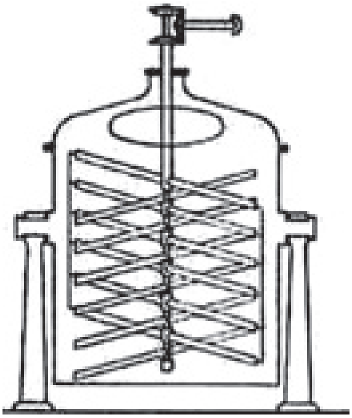

Figure 5: Batteuses

Pomades are processed during winter in cool cellars called batteuses (Figure 5) made of copper and equipped with stirrers. Several batteuses are arranged serially. The alcohol driven from one batteuse is poured in second, third and first washings of successive batteuses. For the last washing, fresh alcohol is used. Extrait no. 36 is cooled in a refrigerator to separate the fat. The exhausted fat is odorless and used for the manufacture of soap. Complete concentration of extrait is done under vacuum at low temperature to remove alcohol, producing absolute of enfleurage which is semisolid in nature. The flowers removed from the chassis are extracted with solvents (petroleum ether) and the concentrated residue is dissolved in absolute alcohol. The fat is removed by freezing. This preparation is known as absolute of chassis. Absolute of enfleurage and absolute of chassis logically supplement one another because each represents only part of the total oil present in the flowers but, due to cost difference, they are kept separately.

4.5 Conclusions

In spite of recent technological development in the field of extraction, hydrolytic maceration, expression and cold fat extraction techniques are inevitable for certain types of raw material. For these materials, no substitutes are available at present, although hydrodistillation or solvent extraction can be used for essential oils of citrus and flower oils or other perfumery materials. Nonetheless, the real replication of essential oil or the true natural fragrances present in these materials can only be produced by expression and cold fat extraction. Only technological improvements for easy extraction can be made without affecting the basic principles of these processes.

Bibliography

Anonymous, 2007, Brown International Corporation, Brown Oil Extractor Model 6100, 633 N. Barranca Avenue Covina, California. 91723-1297 USA. Available at: http://www.brown-intl.com/download/BOE.PDF.

Anonymous, 1983, Food Machinery Corporation. Available at: http://files.asme.org/ASMEORG/Communities/History/Landmarks/5549.pdf

Cooper, J. W. and Gunn, C., 1975, Tutorial Pharmacy, Carter S. J. (Ed.), Reprint CBS Publication, Dehli, p. 255

Ganga, A., Pinaga, A., Quero, A., Valles, S. and Ramon, D., 2001, Cell wall degrading enzyme in release of grape aroma precursors, *Food Science and Technology International*, 7: 83-87

Ghosh, D. and Laddha, K. S., 2005, Herbal drug extraction. Chemical Weekly, Raghavan R. (Ed.), Sevak Publications. 602-B, K. J. Somaiya Hospital Road, Sion (Ed), Mumbai India, p. 185

Goodrich, R. M. and Braddock, R. J. 2004, Major By-products of the Florida Citrus Processing Industry. University of Florida, Institute of Food and Agricultural Sciences (UF/IFAS). Available at: http://edis.ifas.ufl.edu/BODY_FS107#beginning

Guenther, E., 1949, The Essential Oils, D. Van Nostrand Company, Inc. 257 Fourth Avenue, New York 10, a. Vol 3, p. 6-75, b. Vol 1, p. 189-200

Kim, D. H., Kim, J. H., Bae, S. E., Seo, J. H., Oh, T. K. and Lee, O. H., 2005, Enhancement of natural pigment extraction using *Bacillus* species xylanase, *Journal of Agricultural and Food Chemistry*, 53(7): 2541-2545

Mc Mahon, H. M., Zoecklein, B. W. and Jasinski, Y. W., 1999, The effects of pre fermentation maceration temperature and percent alcohol (V/V) at press on the concentration of cabernet sauvignon grape glucosides and glycoside fractions, *American Journal of Enology and Viticulture*, 50(4): 385-390

Rotter, B., 2006, Prefermentation cold maceration. Available at: www.brsquared.org/wine/articles/coldsoak.htm

Sambamurthy, K., 2002, Pharmaceutical Engineering (Reprint) New Age International (P.) Ltd. New Delhi, p. 174-175

Cross, S.D., 1986, Citrus juice extractor. FMC Corporation, Chicago III U.S. 4,700,620, Oct. 20, 1987

Tsuchiya, T. and Nakamura, C., 1979, Acetocarmine squash method for observing sugar beet chromosomes, *Euphytica*, 28(2): 249-256

5 Decoction and Hot Continuous Extraction Techniques

S. Tandon and S. Rane

Abstract

The chapter describes the techniques, parameters and equipment used for the extraction of plants by decoction and hot continuous extraction. Principles, mechanisms, merits and demerits of conventional solvent extraction and accelerated solvent extraction are also discussed.

5.1 Introduction

Of the traditional methods of extraction of medicinal plant material for making an aqueous extract, decoction is one of the most described. Decoction is a water-based preparation to extract active compounds from medicinal plant materials. In this process, the liquid preparation is made by boiling the plant material with water. Decoction differs from infusion in that the latter is not actively boiled. Decoction is the method of choice when working with tough and fibrous plants, barks and roots and with plants that have water-soluble chemicals. The plant material is generally broken into small pieces or powdered. Different methods have been described for the preparation of decoctions. In the Ayurvedic method, traditionally known as *kwatha*, the crude drug in form of *yavakuta* (small pieces) is placed in earthen pots or tinned copper vessels with clay on the outside. Water is added and the pot is heated on a fire. If the material is soft, four times water is used per 1 part drug; if the drug is moderately hard, eight times water is used and if the drug is very hard, sixteen times water is recommended. The mixture is then boiled on low flame until it is reduced to one-fourth starting volume, in case of soft drugs, and one-eighth in case of moderately or very hard drugs. The extract is then cooled and strained, and the filtrate is collected in clean vessels.

5.2 Solid-liquid Extraction Process

Solid-liquid extraction is one of the most widely used unit operations in the medicinal and aromatic plant industry. One example of solid-liquid extraction is the solvent extraction of herbs. This process, also referred to as leaching, is a separation technique that is often employed to remove a solute from a solid mixture with the help of a solvent. The insoluble solid may be colossal and permeable; more often it is particulate and the particles may be openly porous, cellular with selectively permeable cell walls, or surface-activated. The stream of solids being leached and the accompanying liquid are known as the underflow. The solid content of the stream is called marc. The stream of liquid containing the leached solute is known as the overflow.

5.3 Process Parameters Affecting Solid-liquid Extraction

The following parameters generally affect the rate of solid-liquid extraction (leaching):

- Post-harvest processing
- Matrix characteristics
- Choice of solvent
- Method of contact
- Temperature of extraction
- Number of washes
- Condition for extraction (e.g. agitation)

5.3.1 Post-harvest Processing

After harvesting, most herbs have a moisture content of 60%-80% and cannot be stored without drying. Otherwise, important compounds can breakdown or micro-organisms may contaminate the material. Drying of the herbs in shade in a thin layer is generally preferred. Some medicinal plants, like pyrethrum, lose their active constituents if exposed to direct sun light for a long period. For drying large quantities of plant material, a hot air drying oven is used where material can be placed on a large number of trays stacked over each other. Oven temperature must be kept at a safe level so as not to damage the active constituents of the medicinal plant.

5.3.2 Matrix Characteristics

Knowledge of the matrix characteristics of the carrier solid is important to determine whether it needs prior treatment to make the solute more associable to the solvent. Grinding of plant material means mechanically breaking down leaves, roots, seeds, or other parts of a plant into small units, ranging from large course fragments to fine powder.

Solute may exist in the inert solids in a variety of ways:

1. On the surface of the solid,
2. Surrounded by a matrix of inert material, Chemically combined, or inside the cells.

Solute adhering to the solid surface is readily removable by solvent. When the solute exists in pores surrounded by a matrix of inert material, the solvent has to diffuse to the interior of the solid to capture the solute. In the medicinal plant industry, generally ball mills and fluid mills are employed for powdering, and the optimum particle size is established prior to the large-scale extraction.

5.3.3 Choice of Solvent

The following factors should be considered when selecting a solvent for commercial use:

- *Solvent power (selectivity)*. Only the active, desired constituents should be extracted from the plant material, which means that a high selectivity is required.
- *Boiling temperature*. The boiling point of the solvent is as low as possible in order to facilitate removal of the solvent from the product.
- *Reactivity*. The solvent should not react chemically with the extract, nor should it readily decompose.
- *Viscosity*. A low viscosity of the solvent leads to low pressure drop and good heat and mass transfer.
- *Safety*. The solvent should be non-flammable and non-corrosive, and should not present a toxic hazard; its disposal should not imperil the environment.
- *Cost*. The solvent should be readily available at low cost.
- *Vapor pressure*. To prevent loss of solvent by evaporation, a low vapor pressure at operating temperature is required.
- *Recovery*. The solvent has to be separated easily from the extract to produce a solvent-free extract.

5.3.4 Conditions for Extraction

Too fine particle size may result in problems with packing of solids for extraction, preventing free flow of solvent through the solid bed. In such a case, extraction is more difficult, especially when finely divided solids are treated in an un-agitated state. Dispersion of the particles in liquid solvent by agitation facilitates contact of the solid with the solvent. Agitation, while giving good extraction, may cause suspension of fine particles in overflowing solution.

5.4 Method of Solid-liquid Extraction

The three principle types of flow used in leaching systems are:

1. Single-stage system
2. Multistage counter-current system
3. Multistage co-current system

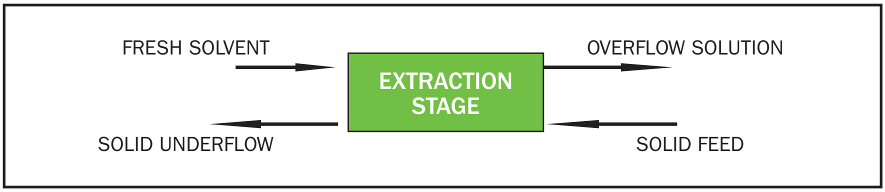
FRESH SOLVENT
EXTRACTION STAGE
OVERFLOW SOLUTION
SOLID UNDERFLOW
SOLID FEED

Figure 1: Single-stage solid-liquid extraction

The single-stage system represents the complete operation of contacting the solid feed and fresh solvent. This is rarely encountered in industrial practice because of the low recovery of solute obtained and relatively dilute solution produced. Efficiency of extraction is somewhat improved by dividing the solvent into a number of smaller portions and then carrying out multiple successive extractions instead of only one contact of the entire amount of solvent with the solid.

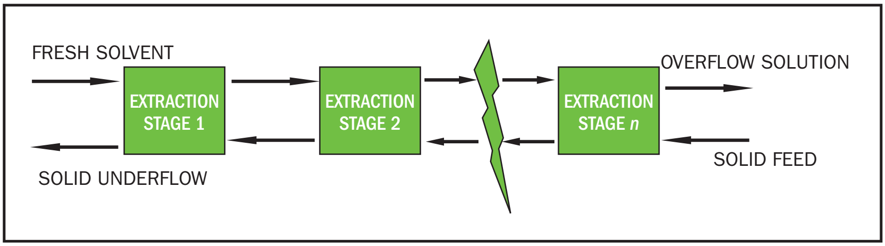
FRESH SOLVENT
EXTRACTION STAGE 1
EXTRACTION STAGE 2
EXTRACTION STAGE n
OVERFLOW SOLUTION
SOLID FEED
SOLID UNDERFLOW

Figure 2: Multistage counter-current solid-liquid extraction

In the continuous counter-current multistage system shown in Figure 2, the underflow and overflow streams flow counter-current to each other. This system allows high recovery of solute with a highly concentrated product because the concentrated solution leaves the system after contact with fresh solid.

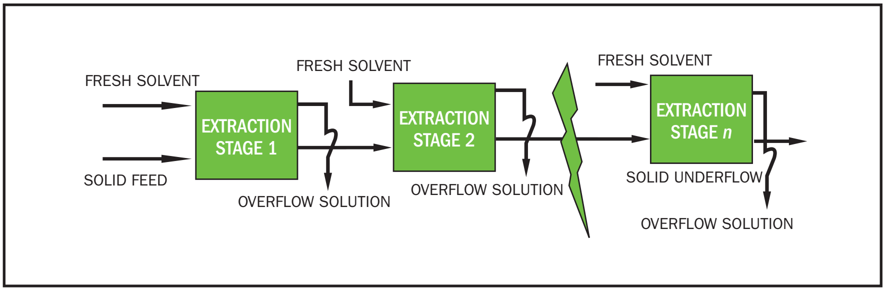
FRESH SOLVENT
SOLID FEED
EXTRACTION STAGE 1
OVERFLOW SOLUTION
FRESH SOLVENT
EXTRACTION STAGE 2
OVERFLOW SOLUTION
FRESH SOLVENT
EXTRACTION STAGE n
SOLID UNDERFLOW
OVERFLOW SOLUTION

Figure 3: Multistage co-current solid-liquid extraction

In the multistage co-current (parallel) system shown in Figure 3, fresh solvent and solid feeds are contacted in the first stage. Underflow from the first stage is sent to the second stage, where it comes in contact with more fresh solvent. This scheme is repeated in all succeeding stages.

5.5 Solid-liquid Extraction Equipment

Equipment for solid-liquid extraction is of two types:

a) Batch solid-liquid extractor
b) Continuous solid liquid extractor

The most common batch extractors employed for solid-liquid extraction of medicinal plants are:

- *Pot extractor*. The extractor has a volume of 2-10 m^3 and a mixer is necessary to guarantee good mixing for treatment of fine materials. For structured materials, the mixer is only used for evaporation of the solvent and for emptying the extractor.
- *Rotating extractor*. The extractor is filled with extraction material and solvent and starts then to rotate. The installation of heating coils and the use of a double jacket make it possible to evaporate the solvent at the end of the extraction cycle. A special form of heating coil can act as mixer during the extraction period.

The advantage of batch extractors is that they are simple to operate and are robustly constructed. Disadvantages of batch extractors are the limited capacity and the discontinuous output of the product.

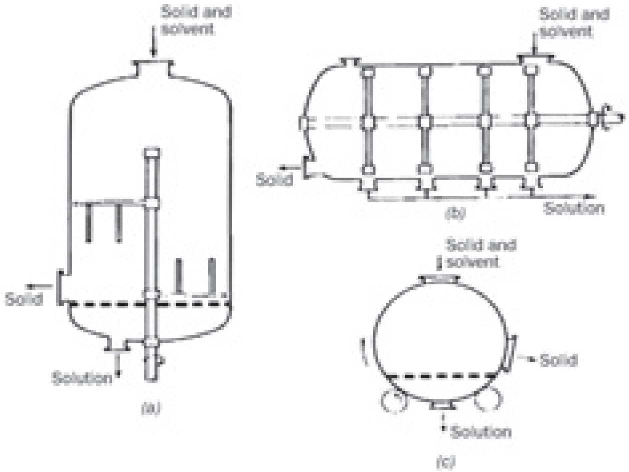

Figure 4: Pot extractor and rotating extractor

5.5.1 Continuous Extraction

For continuously operating extraction, percolation and immersion are used.

5.5.1.1 Percolation

The solvent passes through the non-moving solid material and extracts the soluble active constituents. One advantage of this method is that the solid material requires little mechanical treatment because it does not need to move in the percolator while the product passes in the solution. Moreover, since self-filtration takes place, there is minimum content of fine solid particles in the extract.

5.5.1.2 Immersion

In this process, the solid material dips completely into the solvent and is mixed with it. Therefore, no special percolation properties of the solid material are necessary. The disadvantage is that no self-filtration of the extract solution takes place. Therefore, a filtration step has to be added.

5.5.2 Continuous Extraction Equipment

5.5.2.1 Continuous Horizontal Extractor

The solid material is placed in baskets and comes in contact with the solvent by percolation. The flow of solvent through the extractor is counter-current to the flow of solid material.

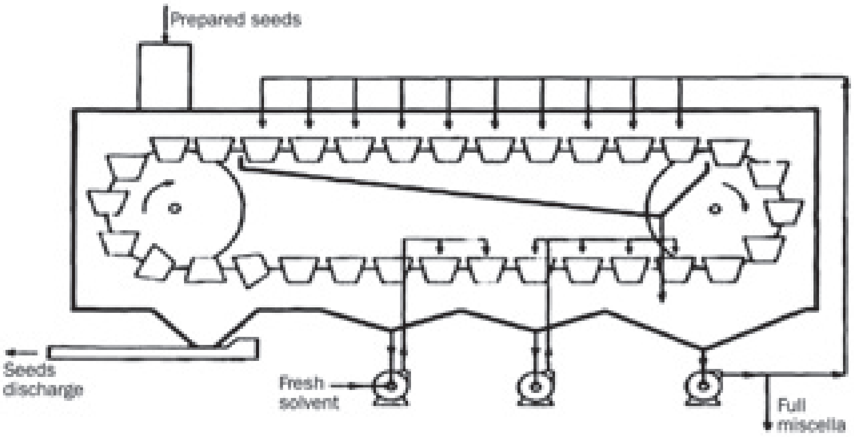

Figure 5: Continuous horizontal extractor

5.5.2.2 Hildebrandt Extractor

The solid material is extracted according to the immersion method. Screw conveyors are installed in the extractor for transporting the solid material. Again, the solvent flows counter-current to the solid materials through the extractor.

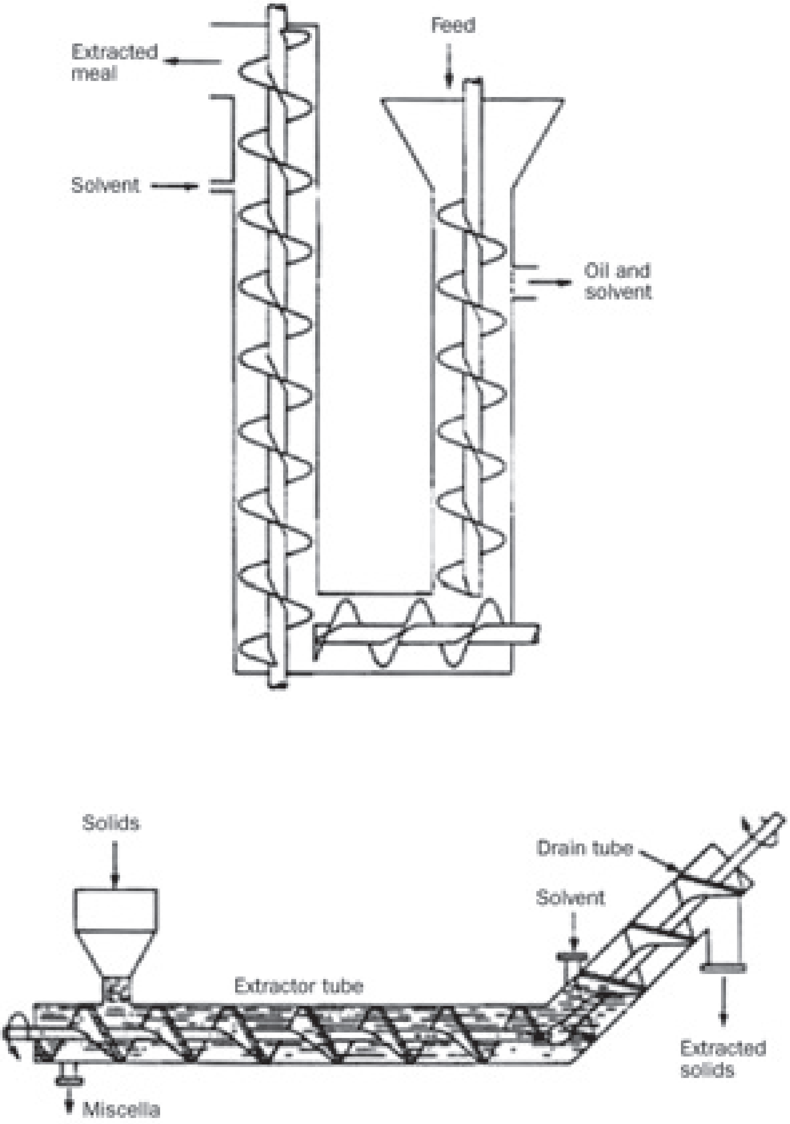
Feed
Extracted meal
Solvent
Oil and solvent
Solids
Drain tube
Solvent
Extractor tube
Extracted solids
Miscella

Figure 6: Hildebrandt extractor

5.5.2.3 Bonotto Extractor

The Bonotto extractor is used for counter-current extraction according to the immersion method. The solid material is transported by the mixer on a tray until it reaches the open sector where it falls onto the next tray. The screw conveyor at the outlet withdraws the extracted solid material (underflow) and prevents the solution from flowing out of the extractor.

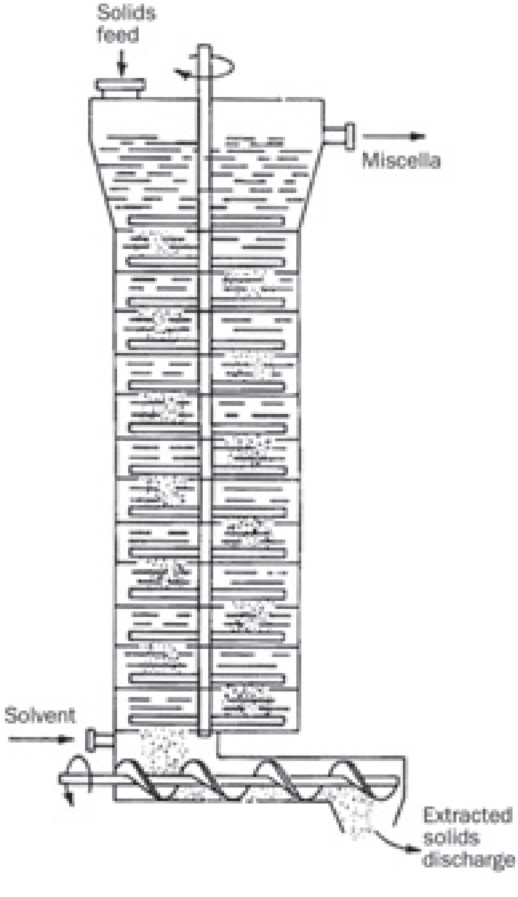

Figure 7: Bonotto extractor

5.5.2.4 Bollmann Extractor

The fresh solvent is added during the upward movement of the baskets so that this part operates in counter current. The already preloaded solution is withdrawn from the bottom of the extractor and enters the downward-moving baskets so that this part of the extractor operates in a co-current way. The full *miscella* is withdrawn at the bottom of the extractor.

In the baskets, self-filtration takes place, so that no further treatment of the *miscella* before distillation is necessary.

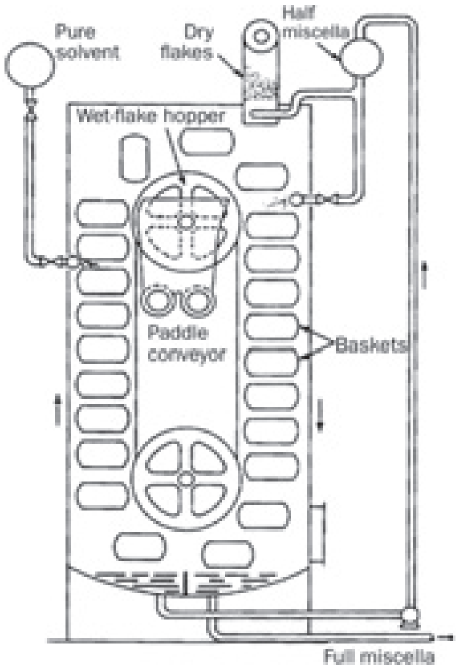
Pure solvent
Dry flakes
Half miscella
Wet-flake hopper
Paddle conveyor
Baskets
Full miscella

Figure 8: Bollmann extractor

5.5.2.5 Kennedy Extractor

The solid material is transported by paddles from one chamber to the next, in counter-current way to the solvent. The chamber where the *miscella* is withdrawn is used as a filtration unit where fine particles are separated from the extract solution.

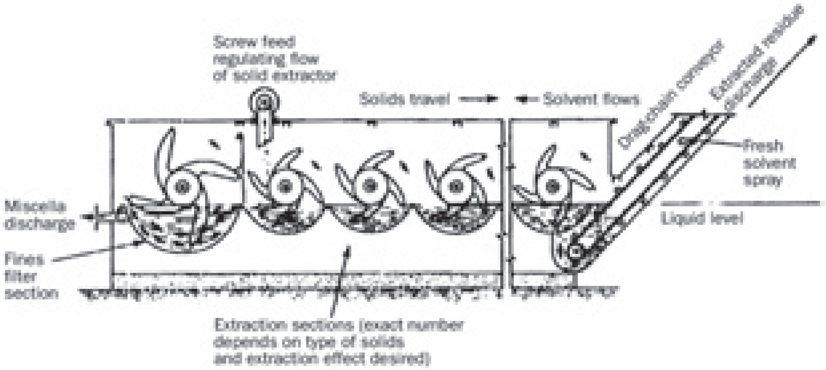
Screw feed regulating flow of solid extractor
Solids travel
Solvent flows
Dragchain conveyor
Extracted residue discharge
Fresh solvent spray
Liquid level
Miscella discharge
Fines filter section
Extraction sections (exact number depends on type of solids and extraction effect desired)

Figure 9: Kennedy extractor

5.6 Conventional Solvent Extraction

5.6.1 Principles and Mechanisms

Classic techniques for the solvent extraction of active constituents from medicinal plant matrices are based on the choice of solvent coupled with the use of heat or agitation. Existing classic techniques used to obtain active constituents from plants include: Soxhlet, hydrodistillation and maceration with an alcohol-water mixture or other organic solvents. Soxhlet extraction is a general and well-established technique, which surpasses in performance other conventional extraction techniques except for, in limited fields of application, the extraction of thermolabile compounds.

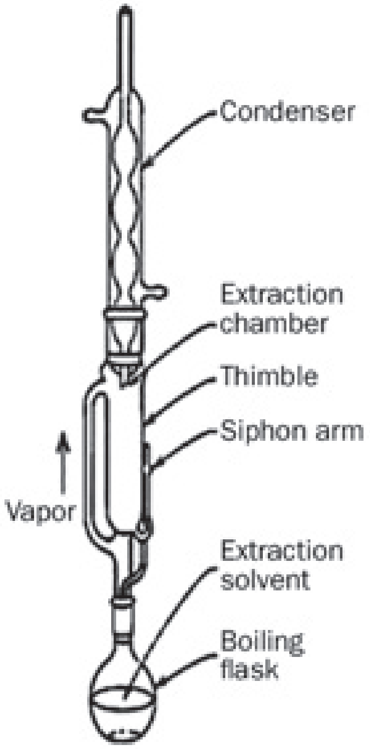

Figure 10: Soxhlet extractor

In a conventional Soxhlet system, as shown in Figure 10, plant material is placed in a thimble-holder, which is filled with condensed fresh solvent from a distillation flask. When the liquid reaches the overflow level, a siphon aspirates the solution of the thimble-holder and unloads it back into the distillation flask, carrying extracted solutes into the bulk liquid. Solute is left in the flask and fresh solvent passes back into the plant solid bed. The operation is repeated until complete extraction is achieved.

5.6.2 Advantages and Disadvantages of Soxhlet Extraction

Advantages:

1. The displacement of transfer equilibrium by repeatedly bringing fresh solvent into contact with the solid matrix.
2. Maintaining a relatively high extraction temperature with heat from the distillation flask.
3. No filtration of the extract is required.

Disadvantages:

1. Agitation is not possible in the Soxhlet device.
2. The possibility of thermal decomposition of the target compounds cannot be ignored as the extraction usually occurs at the boiling point of the solvent for a long time.

Worldwide, most of the solvent extraction units are based on the Soxhlet principle with recycling of solvents. Basic equipment for a solvent extraction unit consists of a drug holder-extractor, a solvent storage vessel, a reboiler kettle, a condenser, a breather system (to minimize solvent loss) and supporting structures like a boiler, a refrigerated chilling unit and a vacuum unit.

Table 1: Some common solvents used for the extraction of medicinal and aromatic plants

Solvent	**Boiling point, °C**	**Miscibility with H_2O**	**Threshold limit values, ppm**
Acetone	56	∞	1000
Acetic acid	116-117	∞	10
Ethyl acetate	77	80%	400
Benzene	80	<0.01%	25
2-Butanol	79.5	19%	2200
Cyclohexane	80.7	<0.01%	300
Dichloromethane	39.7	1.3%	2200
Chloroform	61	8%	50
Carbon tetrachloride	76.77	0.8%	10
Hexane	69	<0.01%	-
Ethanol	78	∞	1000
Ethyl ether	34.6	1.2%	400
Petrol ether	30-50	-	500
Propanetriole	290*	∞	-
Methanol	64.7	∞	200
1-Propanol	91	M	400
2-Propanol	82.4	M ?	400
Toluene	110.6	0.06	100

t = < 0.01%; * with decomposition; M miscible; ∞ completely miscible

5.7 Accelerated Solvent Extraction

5.7.1 Principles and Mechanisms

Accelerated solvent extraction (ASE) is a solid-liquid extraction process performed at elevated temperatures, usually between 50° and 200° C, and at pressures between 10 and 15 MPa. Therefore, accelerated solvent extraction is a form of pressurized solvent extraction. Increased temperature accelerates the extraction kinetics and elevated pressure keeps the solvent in the liquid state, thus achieving safe and rapid extraction. Also, high pressure allows the extraction cell to be filled faster and helps to force liquid into the solid matrix. A typical accelerated solvent extraction system is illustrated in Figure 11. Although the solvent used in ASE is usually organic, pressurized hot water can also be used. In these cases, one refers to pressurized hot water extraction or sub-critical water extraction.

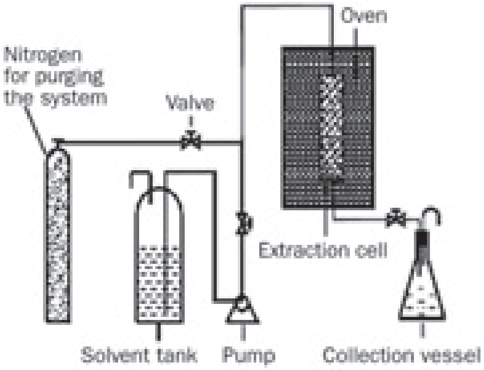

Figure 11: Accelerated solvent extraction

5.7.2 Advantages and Disadvantages of Accelerated Solvent Extraction

Compared with traditional Soxhlet extraction, ASE presents a dramatic reduction in the amount of solvent and extraction time. Particular attention should be paid to ASE performed at high temperature, which may lead to degradation of thermolabile compounds.

5.8 Important Factors for Designing a Solvent Extraction Plant for Medicinal Plants

- High efficiency of extraction
- Minimal solvent loss
- Facilities for cold and hot extractions
- Extraction with agitation

- Multiple solvent extraction systems
- Multiple fraction collection systems
- On-line filtration unit
- Solvent recycling and condensing unit
- Vent lines with breather for minimizing solvent loss and maximizing safety
- Brine circulation unit
- Fractionating column for separation of solvent mixtures
- Efficient evaporating systems like falling film, wiped film or rotary evaporators to work under low pressures
- Vacuum manifold system with cold traps
- GMP compatible
- Automation

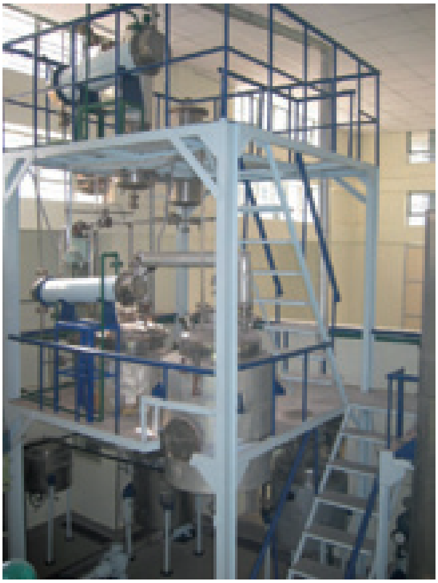

Figure 12: Solvent extraction plant at CIMAP, India

5.9 Conclusions

Decoction, a water-based preparation, is one of the most used traditional methodologies for the extraction of active constituents of a medicinal plant. It is generally carried out by boiling the plant part for a fixed period. Hot continuous extraction or solvent extraction technique is one of the most widely used extraction techniques for the processing of medicinal plants. The solvent extraction method is simple, well established and economical. Important factors that can affect the efficiency of extraction, such as post-harvest processing, solid characteristics, choice of solvent, method of contact, and temperature, should be optimized for best yield. The choice of solvent especially for commercial plants and high efficiency usually depends on many factors such as selectivity, polarity, boiling point, chemical and thermal stability, safety, flammability, and costs. Despite the economic advantages of solvent extraction, the use of volatile organic solvents such as hexane, acetone and methanol for processing medicinal plants has been limited due to environmental considerations. Hot continuous extraction technology shall always remain the method of choice for high efficiency, economical extraction and with less capital investment.

Bibliography

Dung N. X. and Dinh, T., 2005, Extraction and Distillation of essential Oils, Processing, Analysis and Application of Essential Oils, 1st Edition, Har Krishan Bhalla & Sons Book Company, p. 59

Gamse, T., 2002, Lecture on Liquid-liquid Extraction and Solid-liquid Extraction. Available at: www.iq.uva.es/separacion/archivos/SkriptumExtraction.pdf

Hazra, P. et al., 1989, Solvent extraction of *Artemisia annua* L. on pilot scale, *Research and Industry*, 36: 14-16

Kahol, A. P., Tandon, S. and Singh, K. L., 1998, Developments in separation technologies for the processing of medicinal and aromatic plants, *PAFAI Journal*, 20(3): 17-28

Pangarkar, V. G., 1990, Liquid-liquid contact operation: review of equipment, *PAFAI Journal,* 12(3): 13-17

Perry, R. H., 2003, Perry's Chemical Engineer's Handbook, McGraw-Hill Company

Sambamurthy, K., 2002, Pharmaceutical Engineering (Reprint), New Age International (Pvt.) Ltd., New Delhi, p. 173-195

Schweitzer P. A., 1979, Handbook of Separation Technique for Chemical Engineers, 1st Edition, McGraw-Hill Book Company

Tandon, S. et. al., 2003, Pilot plant processing data for the isolation of artemisinin from the new variety of *Artemisia annua* 'Jeevanrakha', *Journal of Scientific and Industrial Research*, 62: 457-461

Treybal, R. E., 1988, Mass Transfer Operations, McGraw-Hill Book Company, 3rd Edition, p. 717

Wang, L. and Weller, C. L., 2006, Recent advances in extraction of nutraceuticals from plants, *Trends in Food Science & Technology*, 17: 300–312

Warren, L. M., 1987, Unit Operations of Chemical Engineering, McGraw-Hill Book Company, p. 531

6 Aqueous Alcoholic Extraction of Medicinal and Aromatic Plants by Fermentation

C. K. Katiyar

Abstract

The history of the development of pharmaceutical dosage forms can be traced back to the Vedic era. It provides insight into the innovations made during ancient times when the crude herbs were initially used in powder form and later on they were used as decoctions, self-fermented products, paste, pills and other advanced dosage forms. New drug development involves the search for novel and new pharmacophores. Combinatorial chemistry cannot solve the problem in a satisfactory manner nor can the reductionist approach be applied to medicinal plants. Asava arishta, a fermented Ayurvedic product, is a good source of novel pharmacophores for new drug discovery.

6.1 Introduction

The history of development of pharmaceutical dosage forms can be traced back to *Charak Samhita*, the first systematic documentation of Ayurveda. Ayurveda has recommended a comprehensive *Materia Medica* including medicinal plants, minerals, metals, and products of marine and animal origin. However, the use of herbs has been given priority.

Medicinal plants have been used for therapeutic purposes for centuries. Initially, these were used in fresh or dried powder form, which caused the problems of high dose, high volume and low shelf life. This led to the development of extraction processes. Extracts were found to be more useful as the necessary dose was less, the volume was low and shelf life was higher. Initially the solvents used for extraction were either water or alcohol, or their mixture. This evolutionary phase is continuing even today where solvents of all kinds of polarity are tried and extraction technologies have evolved from simple water decoction centuries ago to supercritical extraction. While this is true for some streams of products, Ayurveda did not exactly follow the same route but adopted a *sui generis* system of innovation.

6.2 Ayurvedic Dosage Forms

During ancient times, various dosage forms were developed. The number of dosage forms that developed over time is given below:

Charak Samhita (12th century BC)	128 dosage forms
Sushruta Samhita (10th century BC)	129 dosage forms
Ashtanga Hridaya (6th century AD)	90 dosage forms

Chakradutta (9th century AD)	90 dosage forms
Sarngadhara (14th century AD)	75 dosage forms
Bhaishajya Ratnavali (18th century AD)	98 dosage forms

The evolutionary phases of dosage forms followed by Ayurveda starting from crude plant material are mentioned in the following paragraphs.

6.2.1 *Swarasa* (Fresh Juice)

The evolution of liquid orals started from the administration of freshly obtained juices of plant material. To obtain fresh juices, green herbs are crushed and the juice is expressed by squeezing the crushed material. The product is referred to as *swarasa*.

6.2.2 *Kalka* (Wet Bolus)

In this method, the crushed fresh plant material is administered as such, without expressing the juices.

6.2.3 *Kwatha* (Decoction)

One part of coarsely powdered herb is boiled with 16 times its weight of water in an earthen pot over a mild fire until the liquid is reduced to one-fourth or one-eighth of the original quantity, depending upon the nature of the plant material.

6.2.4 *Hima* (Cold Infusion)

The plant material is dried and coarsely powdered. As and when required, the powder is soaked in plain water for a defined period. Then it is filtered, the marc is squeezed, and the combined filtrate is used.

6.2.5 *Phanta* (Hot Infusion)

As a further advancement of *hima*, the *phanta* method was adopted. This method uses boiled water for obtaining a hot infusion.

6.2.6 Solids

Anjana, Churna, Mansa potli, Utkarika, Kshara, Gutika, Guda, Dhumravarti, Puplika, Prithuka, Mandura, Modaka, Rasakriya, Vati, Varti, Shashkuli, Saktu, Bhasmas, Rasaushadhis.

6.2.7 Semisolids

Oral: o*dana* (rice preparation), *kalka, krishara, avaleha.*

Topical: *lepa, upnaha* (poultice), *tilapishta, patrasveda, madhucchishta* (beeswax).

6.2.8 Liquids

Oral: *taila, ghrita, asava/arishta, arka, kwatha, kshirapaka, takra, phanta, him, swarasa, peya, phanita, manda, manasa rasa, yusha, vesavara, vilepi, madya*

Topical: *ashchyotana, karna purana.*

6.2.9 Fumes

Dhumrapana, dhupana.

6.3 Shelf Life of Dosage Forms

Sarngadhara Samhita gave the shelf life of various dosage forms. In ancient times, Ayurvedic physicians themselves prepared the recipes for patients. In ancient times, Ayurvedic physicians themselves prepared the recipes for patients; during the fourteenth century AD, they became aware of the problem of poor shelf life of the botanicals in some dosage forms, such as powder and decoction. This led to the discovery of novel dosage forms termed *asava* and *arishtas,* which are self-fermented preparations having approximately 10%-12% alcohol. These are similar to medicated wines. In the preparation of *asava* cold infusion of unprocessed plant material is used, whereas for the preparation of *arishta* decoction of the plant material is used for fermentation.

6.4 *Asava* and *Arishta*: Self-fermented Products

This unique dosage form discovered by Ayurveda is supposed to have indefinite shelf life and it was said that the “older the better it is”. In terms of current understanding, this phrase assumes more importance because this dosage form has an inherent attribute of continuous hydro-alcoholic extraction and probably formation of natural analogues of the chemical compounds present in the medicinal plants. Some of the major self-fermented Ayurvedic preparations are given in Table 1.

Table 1: Major self-fermented Ayurvedic preparations

Product	**No. of ingredients**	**Indication**
Aravindasava	27	Pediatric tonic
Arjunarishta	5	Cardiotonic
Ashokarishta	14	Menstrual cycle regulator, especially to control excessive bleeding for prolonged periods during menstrual cycle
Ashwagandharishta	24	General tonic
Dashamularishta	67	Normalization of physiological processes after childbirth in women; also anti-inflammatory
Drakshasava	17	General tonic
Jeerakadyarishta	13	Galactogogue
Kumaryasava	46	Liver disorders
Kutajarishta	6	Diarrhea and dysentery
Lohasava	14	Anemia

6.4.1 Self-fermentation Process for Preparing *Asava Arishta*

Preparation of *asava arishtas* involves complicated processes where the following factors play important roles:

- Ingredients and their ratios
- Process
- Pot
- Season
- Place

The method of preparing *asava arishtas* is known as *sandhana kalpana* in Ayurveda. Briefly, a decoction or cold infusion of several herbs is taken and a defined amount of jaggery (dried juice of sugarcane) is added along with flowers of *Woodfordia fruticosa* as inoculum to initiate fermentation. It is kept for about four weeks for fermentation by anaerobic method to obtain a particular level of self-generated alcohol. The product is then kept for some time for maturation. Spices like cardamom and cinnamon are added as flavoring agents. A brief outline of the process of preparing *asava arishta* is given in Figure 1.

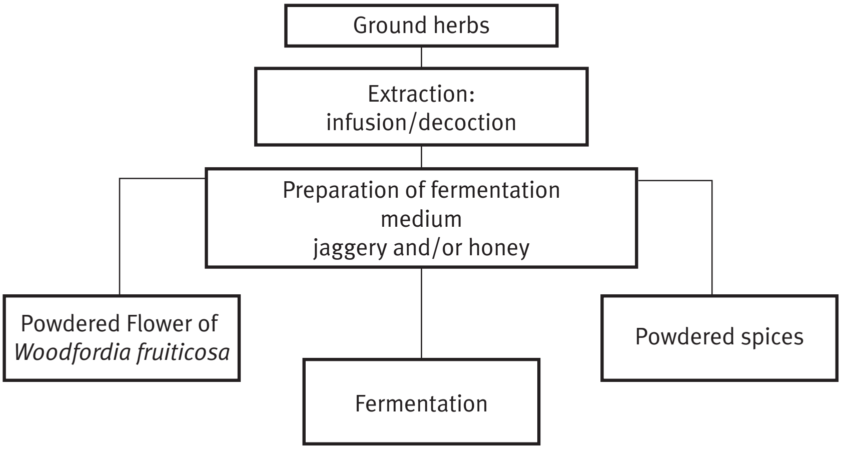
Ground herbs
Extraction: infusion/decoction
Preparation of fermentation medium jaggery and/or honey
Powdered Flower of *Woodfordia fruiticosa*
Powdered spices
Fermentation

Figure 1: The traditional process of preparing asava-arishtas

A crude match-box method is applied to check whether fermentation has occurred. This method depends upon the release of carbon dioxide during the process. The major role in this dosage form is played by *Woodfordia fruticosa,* which is used as inoculum for fermentation but appears to play a role beyond that.

6.4.2 Merits of the Process

Prahst has mentioned some of the benefits of fermented herbal products which are reproduced below:

- Fermentation removes most of the undesirable sugars from plant material, makes the product more bio-available and eliminates side effects such as gas and bloating.
- Fermentation extracts a wider range of active ingredients from the herb than any extraction method since the menstruum undergoes a gradient of rising alcohol levels.
- Yeast cell walls naturally bind heavy metals and pesticide residues and, therefore, act as a natural cleansing system.
- Not only does fermentation remove contaminants, it can also lower the toxicity of some of the toxic components in plants.
- Fermentation actively ruptures the cells of the herb, exposing it openly to the menstruum and bacteria have enzymes that break down cell walls to further assist in the leaching process. Fermentation also creates an active transport system that moves the dissolved constituents from the herbal material to the menstruum.

6.5 Application of *Asava Arishta* Technology in New Drug Discovery

The Ayurvedic dictum with regard to *asava arishtas* that “older is better” needs to be scientifically evaluated. The process of preparing *asava arishtas* appears to involve:

1. Slow hydro-alcoholic extraction at room temperature of crude plant material particles floating in the liquid. Since the particle size of the plant material floating in the liquid is small, the effectiveness of extraction may be higher because of the larger surface area.
2. During the process, if the product is kept for a prolonged period, the probability of development of analogues of some of the pure chemical compounds of the plant material is high.

With a view to enhance the success rate of isolation of pure “druggable” compounds from medicinal plants, it is advised to start from 2- to 3-year-old self-fermented preparations than from solvent extracts. The chances of successful isolation of effective therapeutic compounds using this approach may be high and need to be evaluated.

6.6 Conclusions

Fermentation was applied hundreds of years ago in Ayurveda to develop *asava arishtas*, a multiherbal product, with a view to increase the shelf life and also to enhance the efficacy profile. The race for discovery of new molecules is getting increasingly competitive; at the same time, lack of new and novel pharmacophores is a big impediment which slows down drug discovery. Nature continues to be a source of pharmacophores, although the compounds isolated may not be “druggable” as such. They need to be derived, mostly to enhance their potency. In *asava arishtas*, the self-fermented products can undergo continuous chemical transformation which goes on beyond hydro-alcoholic extraction of the suspended material. This may result in novel natural molecules with enhanced therapeutic activity.

Bibliography

Charak Samhita, 2001, Sharma & Dash (Eds.), Vol. I-VI. Pub. Chaukhambha Sanskrit Series, Varanasi

Clardy, J. and Walsh, C. 2004, Lessons from natural molecules, Nature, 432(16): 829-837

Das, G., 1961, Bhaishajya Ratnavali. Ambikadatta Shastri (Ed.), Pub. Chowkhamba Sanskrit Series, Varanasi

Newman, D. J. and Cragg, G. M., 2007, Natural products as sources of new drug over the last 25 years, Journal of Natural Products, 70: 461-477

Prahst, A., 2007, Liquid Fermented Tinctures - A Natural choice. Available at: www.vistamagonline.com

Sarngadhara, 2006, Sarngadhara Samhita. Srikantha Murthy K. R. (Ed.), Pub. Chaukhamba Orientalia., Varanasi

7 Distillation Technology for Essential Oils

S. Tandon

Abstract

In many developing countries, the technology employed for extraction of essential oils from aromatic plants is primitive and obsolete. This results in low yield and poor quality of essential oils. Thus, there is great need for attaining adequate technological capability in the area of processing essential oil plants. This article deals with the different techniques of distillation, the principles of distillation, and important processing and design aspects which affect the yield and quality of the essential oils.

7.1 Introduction

Distillation is the most popular, widely used and cost-effective method for producing essential oils throughout the world.

Distillation of aromatic plants simply implies vaporizing or liberating the oils from the plant cellular membranes in the presence of moisture, by applying high temperature and then cooling the vapor mixture to separate the oil from the water on the basis of the immiscibility and density of the essential oil with respect to water.

7.2 Principles of Distillation

The choice of a particular process for the extraction of essential oil is generally dictated by the following considerations:

a) Sensitivity of the essential oil to the action of heat and water
b) Volatility of the essential oil
c) Water solubility of the essential oil

Essential oils with high solubility in water and those that are susceptible to damage by heat cannot be steam distilled. Also, the oil must be steam volatile for steam distillation to be feasible. Most of the essential oils in commerce are steam volatile, reasonably stable to heat and practically insoluble in water; hence they are suitable for processing by steam distillation.

Essential oils are a mixture of various aroma chemicals, basically monoterpenes, sesquiterpenes and their oxygenated derivatives, having a boiling point ranging from 150° to 300° C. When the plant material is subjected to heat in the presence of moisture from the steam, these oils are liberated from the plant. For the oil to change from the liquid to the vapor phase, it must receive latent heat that, within the tank, can only come from condensing steam. Consequently, the temperature of the steam within the still must be higher than the temperature at which the oil boils in the presence of water on the surface of the plant material, otherwise there would not be a temperature gradient to take the latent heat from the condensing steam to vaporize the oil droplet. Thus, the energy from the steam in form of heat as latent heat of vaporization converts the oil into a vapor. But, as the boiling point of the oil is higher than that of water, the vaporization takes place with steam on the basis of their relative vapor pressures.

It is imperative to note that a liquid always boils at the temperature at which its vapor pressure equals the atmospheric or surrounding pressure. For any two immiscible liquids, the total vapor pressure of the mixture is always equal to the sum of their partial pressures. The composition of the mixture in the vapor phase (in this case, oil and water) is determined by the concentration of the individual components multiplied by their respective partial pressures. For example, if a sample of an essential oil comprised of component A (boiling point, 190° C) and water (boiling point, 100° C) is boiled, after some time, once their vapors reach saturation, the temperature will immediately drop to 99.5° C, which is the temperature at which the sum of the two vapor pressures equals 760 mmHg. In other words, the oil forms an azeotropic mixture with water. Thus, any essential oil having high boiling point can be evaporated with steam in a ratio such that their combined vapor pressures equal the atmospheric pressure; the essential oil can be recovered from the plant by the wet distillation process.

7.3 Methods for Distillation

The following four techniques for the distillation of essential oils from aromatic plants are employed:

1. Water distillation (or hydrodistillation)
2. Water and steam distillation
3. Direct steam distillation
4. Distillation with cohobation

7.3.1 Hydrodistillation

Hydrodistillation is the simplest and oldest process available for obtaining essential oils from plants. Hydrodistillation differs from steam distillation mainly in that the plant material is almost entirely covered with water in the still which is placed on a furnace. An important factor to consider in water distillation is that the water present in the tank must always be enough to last throughout the distillation process, otherwise the plant material may overheat and char. In this method, water is made to boil and the essential oil is carried over to condenser with the steam which is formed. Water-distilled oil is slightly darker in color and has much stronger still notes than oils produced by other methods. The stills based on this principle are simple in design and are extensively used by small-scale producers of essential oils. Care should be taken during distillation of powdered herbs, as they tend to settle on the bottom of the still and get thermally degraded. Also, for plant material that tends to form mucilage and increase the viscosity of the water, the chances of charring are greater. For plant material that has a tendency to agglomerate or to agglutinate into an impenetrable mass when steam is passed through (like rose petals), water distillation is the preferred method of oil isolation.

The primitive, traditional Indian system of essential oil distillation, *bhapka* method, is also based on water distillation (Figure 1). In this process, the plant material is entirely covered with water in a distillation still, which is made of copper and is known as *deg*. This *deg* is placed in a brick furnace. Another copper vessel with a long neck is placed in a water tank or natural pond to serve as a condenser. A bamboo pipe is used as the vapor connection and mud is used to seal the various joints. The water is boiled, the oil vapors along with steam are condensed in the copper vessel, and oil is separated. The capacity of one *deg* is around 40 kg/batch. These types of units are still being used in Kannauj in Uttar Pradesh and in the Ganjam district of Orissa, India for the preparation of *rooh* and attars of *gulab, kewda, khus, rajnigandha*, and *bela*. These units can easily be transported from one place to another, but are not suitable for large-scale distillation of aromatic crops like grasses and mints.

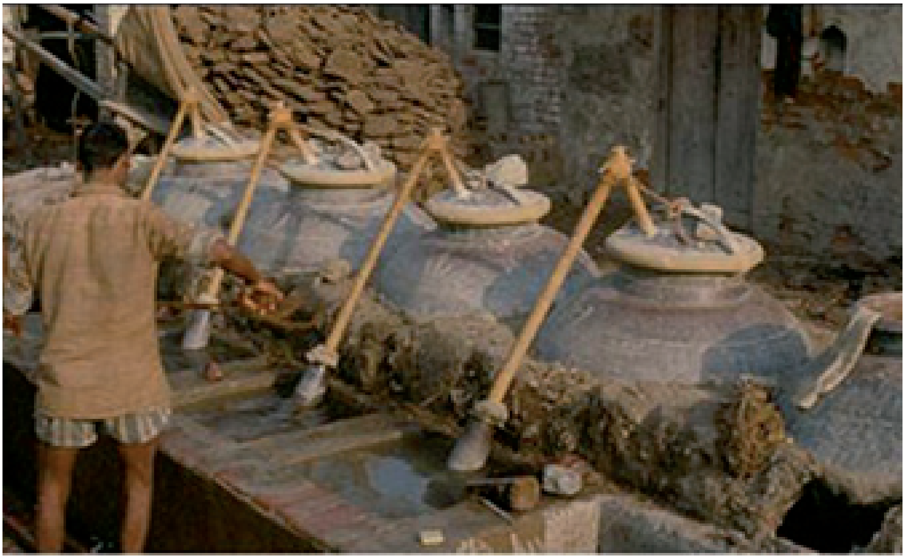

Figure 1: Traditional Indian deg bhapka method

Although hydrodistillation (water distillation) is still being used, the process suffers from the following serious drawbacks:

a) As the plant material near the bottom of the still comes in direct contact with the fire from the furnace, it may char and thus impart an objectionable odor to the essential oil.
b) The prolonged action of hot water can cause hydrolysis of some constituents of the essential oil, such as esters.
c) Heat control is difficult, which may lead to variable rates of distillation.
d) The process is slow and distillation times are much longer than those of steam distillation.

7.3.2 Water and Steam Distillation

To eliminate some of the drawbacks of water distillation, some modifications were made to the distillation units. A perforated grid was introduced in the still, to support the plant material and to avoid its direct contact with the hot furnace bottom. When the water level is kept below the grid, the essential oil is distilled by the rising steam from the boiling water. This mode of distillation is generally termed water and steam distillation.

The field distillation unit (FDU), also known as a directly fired-type distillation unit, is designed according to the principle of water and steam distillation. The FDU consists of a still or tank made of mild stainless steel with a perforated grid and is fitted directly to a brick furnace. A chimney is connected to the furnace to minimize the pollution at the workplace and also to induce proper firing and draft. The plant material is loaded on the perforated grid of the tank and water is filled below it. The tank is connected to the condenser through a vapor line. The water is boiled and the steam vapors pass through the herb, vaporize the oil and get condensed, mostly in a coil condenser by cooling water. The condensate (oil-vapor mixture) is then separated in the oil separator.

These units are simple to fabricate and can be installed in the farmer's field. Due to their simple construction, low cost and easy operation, FDUs are extremely popular with essential oil producers in developing countries. The furnace is always fueled by locally available firewood or straw. This makes the unit suited for use in remote areas where the raw material is available. This also helps in reducing transportation costs in the production of essential oils. FDUs are currently finding application in distillation of patchouli oil in Indonesia, aromatic grass and mint oil in India, citronella oil in Taiwan and many more all over the world. A local FDU currently being used by rural farmers in India for the distillation of mint oils is shown in Figure 2. Such field units generally can hold 100-2000 kg plant material. Total time for distillation with these units is about 6-8 h.

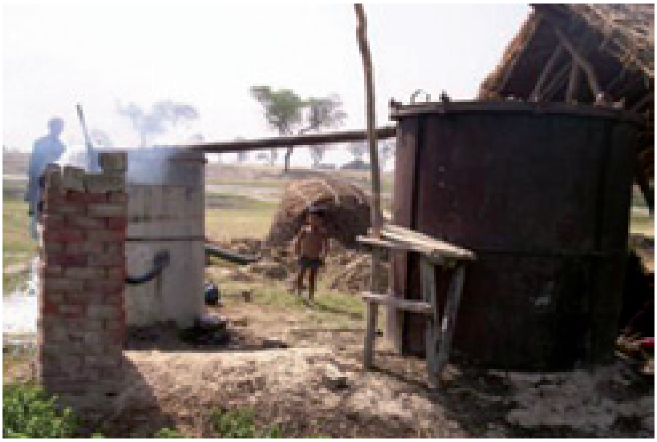

Figure 2: Local type field distillation unit in India

7.3.2.1 Improved Field Distillation Units

Due to the limited heating surface available, the rate of steam production in the FDU is always insufficient. This results in prolonged distillation periods and sometimes lower oil yields. Refluxing of oil back into the still due to inadequate steam rate may lead to decomposition reactions and poorer oil quality. Experimental measurements made at the Central Institute of Medicinal and Aromatic Plants (CIMAP), India, have shown that firewood consumption in a conventional field still may be up to 2.5-times greater than that of a modern steam distillation unit operated by an external boiler. This factor may not be critical where fuel supplies are cheap and abundant but, in many developing countries, fuel supplies are getting scarce and costly and low thermal efficiency can directly affect the cost of production.

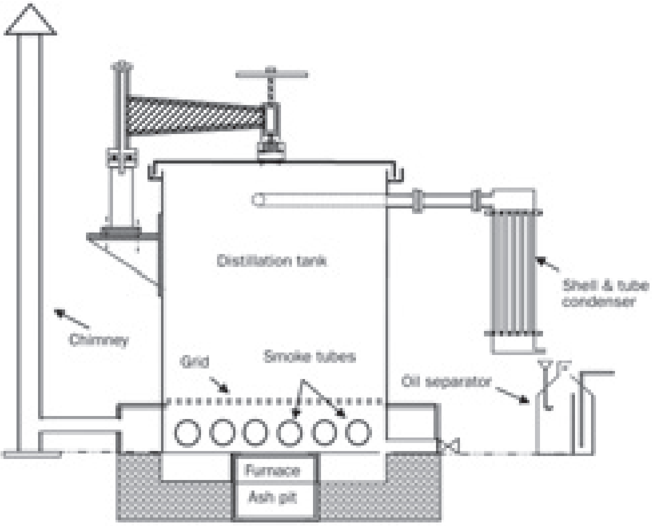

Figure 3: CIMAP's improved field distillation unit

Considering the previously mentioned demerits of FDUs, designs of economical and improved units with capacities 500-2000 kg per batch are now being preferred (Figure 3). The units are fabricated with high quality mild stainless steel, keeping in view the plant materials to be distilled. The improved distillation unit consists of a cylindrical distillation tank fitted on a square inbuilt boiler (calandria) having smoke pipes which reduces the heating time of the water, resulting in a high rate of steam generation and lower fuel consumption (20%-30%). Hot flue gasses of the furnace are led through the smoke tubes where they impart heat to the water, thus raising additional steam. The tank is fitted on a specially designed furnace having fire grate, flue ducts and fire door for proper controlling of the firing and draft. The furnace is connected to a chimney of optimum height to maximize the air draft and control the pollution by smoke in the workplace. A similarly designed stainless steel shell and tube-type condenser having higher condensation capacity are used for cooling the vapors. It prevents loss of oil due to improper condensation. The condensed oil-water mixture is then allowed to pass through a specially designed stainless steel oil separator. The separator has an inbuilt baffle to maximize the retention time of the mixture, thereby resulting in no loss of oil with the outgoing water from the separator. The unit also has a chain pulley hoist system with a support structure that makes work easier and saves time during discharge of the distillation waste material from the tank. CIMAP has designed, fabricated and supplied these improved units to entrepreneurs and farmers in different parts of India.

7.3.3 Direct Steam Distillation

In direct steam distillation, plant material is distilled with steam generated outside the tank in a steam generator or boiler. As in water and steam distillation, the plant material is supported on a perforated grid above the steam inlet. As already noted, the steam in an FDU is at atmospheric pressure and hence its maximum temperature is 100° C. But, steam in a modern pressure boiler operating at, for example, 50 psi pressure will have a temperature correspondingly higher. Moreover, there is no limitation to the steam generation when an external boiler is used as a source of steam. The use of high-pressure steam in modern steam distillation units permits much more rapid and complete distillation of essential oils.

Steam distillation is preferred when a lot of area is under cultivation and more than one unit is to be installed. Also, for distillation of high boiling oils and hardy materials such as roots and woods like sandalwood, cedar wood and nagarmotha, steam distillation is more efficient. Steam distillation also reduces the time required for the extraction of oils. A charge of Java citronella, which takes up to 5 h in an FDU, is processed within two to 3 h in a steam distillation still. In this method of distillation, steam is generated separately in a steam boiler and is passed through the distillation tank through a steam coil (Figure 4). The plant material is tightly packed above the perforated grid. Steam, containing the oil vapor, is condensed in a tube condenser and is separated in the oil receiver. Fuel costs are generally lower in modern steam distillation units due to higher thermal efficiency at which most of the boilers operate. Capital cost is higher, thus only bigger producers can afford to own such units. Still capacities range from 1 to 3 tonne plant material per batch.

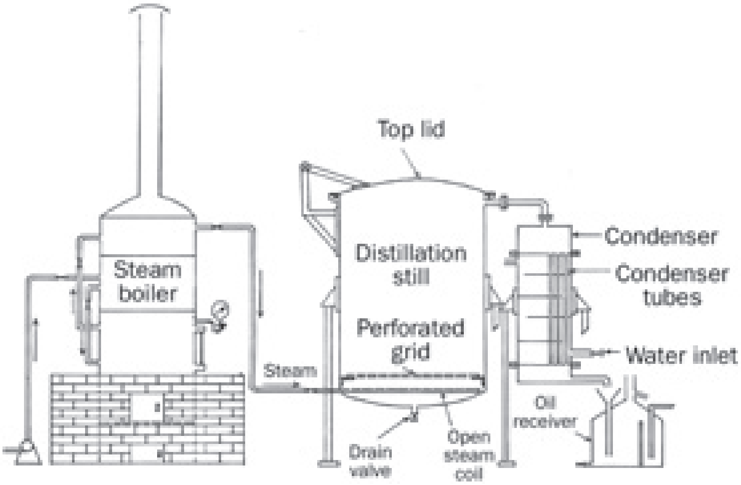

Figure 4: Boiler-operated steam distillation unit

7.3.3.1 Comparison of Boiler-operated Unit with Directly Fired Type

a) It is possible to run a large number of distillation units by a single steam boiler. Hence a boiler system is ideal for large-scale production of essential oils. An FDU is more suited for small and medium-sized farmers.
b) The efficiency of extraction of essential oil in a well-designed FDU can equal that obtained by a boiler unit. But in a poorly designed FDU, oil recovery may be low and fuel wastage may be heavy accompanied by smoke pollution.
c) Steam injection rate in a boiler-operated unit can be adjusted with ease but steam generation rate in an FDU is limited by the heat transfer area provided in the unit. Insufficient steam generation in an FDU can result in low oil yield.
d) Boiler-operated units require a skilled boiler man for operation but an FDU can be operated by relatively low-skilled workers.

7.3.4 Distillation with Cohobation

Cohobation is a technique that can be used for water distillation or for water and steam distillation. It uses the process of returning the distillate water to the still after the oil has been separated from it so that it can be re-boiled. This is basically an improvised methodology of the directly fired type steam and water distillation units for oils which have partial solubility in water. Although most of the essential oils have finite solubility in water, some oils like those of rose, lavender and geranium have comparatively higher solubility. In such extractions, the loss of oil with the outgoing water of distillation can become alarmingly high. This problem can be solved by returning the condensate water from the separator back to the still; this is known as cohobation. It is evident that this cannot be done with steam distillation as the water level in the still will keep building up due to continuous steam injection.

In a further improved version, a packed column is placed on top of the column for providing mass transfer to the oil-water vapors, so as to increase the concentration of the outgoing condensate and to coalesce the oil droplets in the oil separator (Figure 5). The condenser is placed above the column so that the condensate water from the separator can be recycled back to the still by means of gravity. Additional heat, if required, can be provided by a closed steam coil immersed in the tank bottom. The condenser is moved above the distillation still so that condensed water from the separator can flow by means of gravity to the still. By limiting the total quantity of water in this closed cycle operation, it is possible to obtain increased yields of essential oils that are more water soluble. It is relevant to point out here that prolonged recirculation of the distillation water allows the various impurities and plant decomposition products to build up in the system. This may sometimes affect the quality of the oil. One must always keep this in mind when considering a cohobation distillation system for any application.

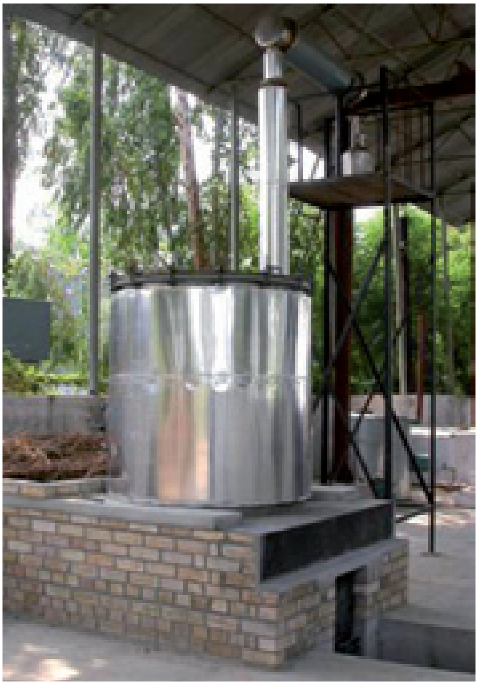

Figure 5: Distillation unit with cohobation

7.4 Hydrodiffusion

This system was first described in 1983. Unlike traditional steam distillation, hydrodiffusion works on the diffusion principle of allowing steam to enter the top of the plant charge and diffuse through the charge by gravity. The process uses the principle of osmotic pressure to diffuse oil from the oil glands. The system is connected to a steam source, and low pressure steam is passed into the plant material from a boiler. The condenser, which is directly under the basket within the still, is of the tube type. The oil and water are collected below the condenser in a typical oil separator. Hydrodiffusion is an efficient process that is easy to use, especially regarding the processes of loading and unloading the plant material. The yield of oil is higher and the process is advantageous because of the reduced steam consumption, shorter distillation time and absence of hydrolysis, as the raw material does not come in contact with boiling water. However, because of the downward flow of steam and condensate, co-extraction of other non-volatile compounds (such as lipids, chlorophyll and fatty acids) and polar components makes the process complicated. Although it may seem that hydrodiffusion is a better alternative to conventional distillation processes, the fact remains that commercial ventures based on hydrodiffusion have not been able to take off successfully.

7.5 Parameters Affecting Yield and Quality of Essential Oils

The yield and quality of essential oil from steam distillation is affected by the various process parameters. It is advisable to keep them in mind while designing such systems. Some of the important parameters are being listed below.

7.5.1 Mode of Distillation

The technique for distillation should be chosen considering the boiling point of the essential oil and the nature of the herb, as the heat content and temperature of steam can alter the distillation characteristics. For high boiling oils such as woody oils (e.g. sandalwood, cedar wood) and roots (e.g. *Cyperus*), the oil should be extracted using boiler-operated steam distillation. Since the heat content and temperature of steam depend upon its pressure, a change in steam pressure can alter the distillation characteristics. High-boiling constituents of essential oils normally require high-pressure steam to distill over. For oil of rose and other florals, the material is generally immersed in water, i.e. hydrodistillation, as flowers tend to aggregate and form lumps which cannot be distilled using water and steam distillation or direct steam distillation.

7.5.2 Proper Design of Equipment

Improper designing of tank, condenser or separators can lead to loss of oil and high capital investments. The design of the furnace and chimney affects the firing and heat control of the distillation rates. Tank height:diameter ratio is important. Similarly the use of a condenser with an improper design and without calculating the heat transfer areas based on the steam generation areas will lead to improper condensation and loss of oil.

7.5.3 Material of Fabrication of Equipment

Essential oils which are corrosive in nature should be preferably distilled in stills made of resistant materials like aluminum, copper or stainless steel. The tank still can be made from a cheaper metal like mild steel or galvanized iron, and the condenser and separator can be made from a resistant material like stainless steel. As only vapor is present in the tank still, the rust and other products of corrosion may not be carried over into the oil. This can result in considerable savings in the capital cost of the equipment. Expensive, high-value essential oils like rose, agarwood, *kewda*, sandalwood and lavender should be distilled in stainless steel systems. Although copper was the most common material of fabrication of distillation stills since ancient times, its availability is getting reduced and with the arrival of superior alloys like stainless steel, it is slowly disappearing from the scene.

7.5.4 Condition of Raw Material

The condition of the raw material is important because some materials like roots and seeds will not yield essential oil easily if distilled in their natural state. These materials have to be crushed, powdered or soaked in water to expose their oil cells. Chopping of plants will also change the packing density of the material when placed in the distillation still. One can pack up to 50% more plant material in the same still after chopping of some aromatic herbs like mint. Air drying and wilting the herb prior to distillation also has considerable effect on distillation. If required, drying of the herbs prior to distillation should be done in shaded areas and the dried material should not be kept in heaps.

7.5.5 Time for Distillation

Different constituents of the essential oil get distilled in the order of their boiling points. Thus, the highest boiling fractions will be last to come over when, generally, very little oil is distilling. If the distillation is terminated too soon, the high-boiling constituents will be lost. In many aromatic plants, like vetiver, patchouli, chamomile, sandalwood and agarwood, these high-boiling fractions are valuable due to the quality of their aromas. Thus, the time of distillation must be chosen with due care.

7.5.6 Loading of Raw Material and Steam Distribution

Improper loading of the herb may result in steam channeling, causing incomplete distillation. The herb should be evenly and uniformly loaded in the tank without leaving any voids. Excessive filling of plant material may also lead to formation of “rat holes” which may allow steam to escape without vaporizing the oil. For powdered herbs, a proper stainless steel wire mesh or muslin cloth should be put at the false bottom to prevent plant material from falling into the tank base.

7.5.7 Operating Parameters

Proper control of injection rates and pressure in boiler-operated units is necessary to optimize the temperature of extraction for maximal yield. Generally, high-pressure steam is not advisable for the distillation of essential oils. The temperature of the condensate should not be high, as it can result in oil loss due to evaporation. In directly fired-type FDUs, the firing of the furnace should be well controlled as it can result in high flow rates and high condensate temperatures.

7.5.8 Condition of Tank and Equipment

The tank and other equipment should not be rusted. If rusted, the tank should be cleaned with dilute caustic solutions. The perforated grids should not be corroded or have large gaps permitting the plant material to settle to the bottom of the tank and emit a burnt odor. The distillation tanks should be well steamed prior to distillation for multiple crop distillation.

7.6 Purification of Crude Essential Oils

Essential oil as obtained from the oil separator is in crude form. It may have suspended impurities and appreciable moisture content. It might even contain some objectionable constituents which degrade its flavor quality. The presence of moisture and impurities adversely affects the keeping quality of oil and accelerates polymerization and other undesirable reactions. Addition of a drying agent like anhydrous sodium sulphate to the oil, standing overnight followed by filtration will remove the moisture and free the oil of suspended impurities. Use of high-speed centrifugation to clarify the essential oils is common.

Essential oils are frequently rectified or re-distilled to remove objectionable constituents. In order to keep the temperature of re-distillation within permissible limits, the process is carried out under vacuum or with the help of steam distillation.

7.6.1 Continuous Steam Distillation

Steam distillation units involve manual charging and discharging of plant material from the tank still. These operations are labor intensive and time consuming. To overcome these problems, continuous steam distillation plants have been developed in the Soviet Union and have been in operation since the last couple of decades. These units are being used for distillation of lavender and require negligible manual handing. Capacities of 2 tonnes per hour are quite common. Incoming plant material is first chopped with special ensilage cutters and then conveyed to the top of a tall distillation column by means of a belt conveyor. The movement of material inside the column is by gravity or by special helical screw conveyors. Sometimes two columns in series are used for complete removal of oil. Steam is injected at multiple points in the column. Spent material is continuously ejected out of the bottom of the distillation column by special screw conveyors with a vapor lock which does not allow steam to escape.

Fabrication and operation of continuous distillation columns is rather complicated and these have not yet gained acceptance and popularity outside the former Soviet Union. In another development, containerized distillation is also being used for the distillation of *Mentha piperita* and lavender in some parts of the United States. In this method, large capacity containers mounted on wheels are attached to a harvester which directly loads the plant material into the containers from the fields (these containers have inbuilt steam coils); these are then taken to the distillation area where steam is directly connected to the coils and the top is closed and connected through a vapor line to the condenser and subsequently to the oil separator.

7.7 Conclusions

Distillation is the most widely used method for the extraction of essential oils. Proper selection of the distillation technique, design and material of fabrication of the equipment, and processing parameters all play vital roles in determining the quality and yield of an essential oil.

Bibliography

Ames, G. R. and Matthews, W. S. A., 1969, The distillation of essential oils, *Perfumery & Essential Oil Records*, 9-18

Boucard, G. R. and Serth, W. R., 1991, A continuous steam stripping process for the distillation of essential oils, *Perfumer and Flavorist*, 16(2): 1-8

Denny, E. F. K., 1969, Hydro distillation of oils from aromatic herbs, *Perfumer and Flavours*, 14(4): 57-63

Denny, E. F. K., 1989, The new approach to oil distillation from the herb, *Indian Perfumer*, 33(1): 70-75

Denny, E. F. K., 1991, Field Distillation for Herbaceous Oils (2nd Edition.), Denny-McKenzie, Assoc., Lilydale, Tasmania, p. 265

Guenther, E., 1965, The Essential Oils, Vol 4, D. Van Nostrand Co., New Jersey, p. 682

Kahol, A. P., 1984, Distillation Technology. In: Practical Manual on: The Essential Oils Industry, Wijesekera, R. O. B (Ed.), UNIDO, Vienna

Lawrence, B. M., 1995, The Isolation of Aromatic Materials from Natural Plant Products. In: A Manual on the Essential Oil Industry, Tuley De Silva (Ed.), United Nations Industrial Development Organization (UNIDO), Vienna, Austria

8 Microdistillation,Thermomicrodistillation and Molecular Distillation Techniques

V. G. Pangarkar

Abstract

Medicinal and aromatic plants (MAPs) have assumed considerable significance in view of their special attributes. There are many compounds of great therapeutic value which can be obtained only from the plant kingdom. Most of the ingredients in such extracts and oils are large bulky molecules highly sensitive to processing conditions. The processes for such extractions have been mostly based on "recipes". In the recent past, significant advances have been made in the unit operations which are part of the recipe-based processes. It is imperative that these advances, which are essentially aimed at achieving better yields at lower costs and are termed "process intensification", are incorporated into the processing of MAPs. This paper introduces the theme of process intensification as applied to the processing of MAPs and presents an overview of several new technologies which allow rapid, cost-effective extraction. The application of such innovative technologies can yield significant benefits in terms of the quality of the product and its yield per unit weight of the plant material processed. Recovery of dissolved essential oil components from steam distillation condensates is also addressed and the two available techniques are discussed in detail.

8.1 Introduction

Phytochemicals derived from medicinal and aromatic plants (MAPs) have been important to humans for centuries. Before the advent of modern synthetic chemistry, many aroma and flavor chemicals were derived from sources of natural origin such as flowers, roots and stems. The contemporary system of allopathic medicine, which has gained tremendous importance in the treatment of various diseases, is mainly based on active pharmaceutical compounds made synthetically. However, in recent years increasing attention has been paid to the traditional systems of treatments followed in Asia and Africa. The variety of medicinal plants and their constituents are being discovered only recently. There are many compounds of great therapeutic value which can be obtained only from the plant kingdom. Thus, vincristine, perhaps better known as the chemotherapy agent Oncovin, is only synthesized in the periwinkle plant *Catharanthus roseus*, and the sole source of the compound is this plant species. There are many other compounds which are equally valuable in other sectors such as food flavors, fragrances and cosmetics.

The processing of MAPs for obtaining the required extracts and oils has been based on traditionally established "recipes". Most of the ingredients in such extracts and oils are large bulky molecules highly sensitive to processing conditions. Generally, relatively mild conditions are used in such processes to protect the integrity of the valuable components. The recipe-based methods are time-tested but arguably not the most efficient in terms of yield, energy consumption per unit of product, etc. With the advent of modern processing techniques, there is an urgent need to revisit the recipe-based processing, understand the science underlying them and develop modern, cost-effective processes. This article deals with some important advances made in the extraction of MAPs and the post-extraction treatment of the products and their byproducts.

8.2 Process Intensification

The modern chemical industry is undergoing drastic changes driven mainly by economic considerations. There is an upsurge of interest in clean, energy-efficient and material-conserving processes. An entirely new discipline, “process intensification” (PI), has become the focus of a large and sustained effort all over the world. Stankiewicz and Moulijn have given a precise definition of PI as “Any chemical engineering development that leads to a substantially smaller, cleaner and more energy efficient technology”. India has not been lagging behind in developing innovative PI concepts.

PI can be broadly divided into two categories, with specific reference to processing of MAPs as per the definition of Stankiewicz and Moulijn. These are processes that employ multifunctional equipment (MF) and those that use process-intensifying equipment.

8.2.1 Multifunctional Equipment

This category of PI employs equipment that can perform multiple functions simultaneously. Thus, earlier process plants that required a number of different instruments devoted to individual tasks are being replaced by such MF equipment. A brilliant example of the use of MF equipment is the conversion of a slow and polluting process for the enzymatic hydrolysis of penicillin G to 6-amino penicillanic acid (important intermediate for semisynthetic antibiotics) into an intensified and sustainable process. Since MF equipment-based plants are smaller and consume less energy, they have become popular for globally competitive and sustainable processes.

8.2.2 Process-intensifying Equipment

This category of PI employs equipment that specifically focuses on intensifying the rates of the various steps. In the case of MAP processing, the main resistance in the overall extraction process is the diffusion of the active molecules through the plant cell membrane to the surface before extraction by the fluid. Microwave-assisted extraction (MAE) is highly useful in obtaining rapid and complete extraction without significant damage to the active molecules; this technique is discussed in some detail later. Ultrasound-assisted extraction is also an alternative. However, considering that ultrasound waves can produce free radicals and that many active molecules are susceptible to such highly reactive species, this approach does not seem feasible.

8.3 Solvent Extraction of MAPs

This process entails the extraction of solid MAPs by liquid solvents. This is a typical solid-liquid extraction process. Two factors that affect the extent and rate of extraction are the thermodynamics and kinetics (rate of mass transfer) of the process.

8.3.1 Thermodynamics of Solvent Extraction and Choice of Solvent

Relative sorption of solutes in the solvent depends on the interactions between the solutes and the solvent. Solubility or miscibility of a component with the solvent depends on their relative solubility parameters. For mutual solubility of two components, their free energy of mixing, ΔG_m should be negative. ΔG_m is defined as

$$\Delta G_m = \Delta H_m - T\,\Delta S_m \tag{3.1}$$

Enthalpy of mixing, ΔH_m can be correlated to cohesive energy density, i.e. solubility parameter (δ) as:

$$\Delta H_m = n_1\, n_2\, V_1\, (\delta_1 - \delta_2)^2 \tag{3.2}$$

In equation (3.2), the solubility parameter is that due to only dispersive forces between structural units of the concerned solute and solvent, since the original regular solution theory of Scatchard and Hildebrand was restricted to non-polar, non-hydrogen bonding solute-solvent systems. However, for many liquids and solutes, contributions from polar and hydrogen bonding forces need to be considered. Accordingly, equation (3.2) becomes:

$$\Delta H_m = n_1\, n_2\, V_1\, [(\Delta\delta_d)^2 + (\Delta\delta_p)^2 + (\Delta\delta_h)^2] \tag{3.3}$$

From equations (3.2) and (3.3), it is clear that to make ΔG_m negative, the difference between δ_i (solvent) and δ_i (solute), i.e. ($\Delta\delta$) for all the three forces of interactions, should be as small as possible. It implies that the solvent and the desired solute to be extracted should have comparable polarity and hydrogen bonding capabilities to achieve similar solubility parameter values. Grulke has given an exhaustive tabulation of solubility parameters for the most common chemical compounds. The individual δ_h, δ_d and δ_p values for compounds not listed in the tabulation can be obtained using the group contributions due to various different groups given by Grulke. When the individual δ_i (solvent) and δ_i (solute) values are very close, a high solubility of the solute in the solvent is obtained. For instance, it is well known that non-polar solvents dissolve terpene fractions more than oxygenated compounds because both are non-polar. On the other hand, mixed solvents of polar and non-polar compounds can yield better results for oxygenated compounds. Bio-ethanol is a good solvent for such oxygenated compounds on two accounts: (i) it is natural, and (ii) it is “green” (renewable). However, most MAPs contain water and the complete miscibility of ethanol with water implies dilution of the solvent after each use. This is further complicated by the fact that ethanol forms an azeotrope at high concentration (~95 wt%). As a result, ingress of small quantities of water is sufficient to reach the azeotropic composition. Implementation of the Montreal Protocol, the Clean Air Act, and the Pollution Prevention Act of 1990 has resulted in increased awareness of organic solvent use in chemical processing.

8.3.2 Solid-liquid Mass Transfer

The MAPs to be processed are in solid form. Solid-liquid extraction is a typical heterogeneous mass transfer process. In such processes, the rate of extraction depends upon: (i) the interface area, and (ii) the mass transfer coefficient. Both should be high. High effective interface area can be obtained by comminuting the solid material to be processed. During comminution, the ensuing friction can increase the temperature of the solid and thereby possibly lead to degradation of thermally labile components. To avoid this, special water-cooled roll crushers are used.

The mass transfer coefficient depends on the diffusivity of the solute in the solid matrix (main resistance) and the level of turbulence in the extractor. Traditional extraction has relied upon percolation or extraction in stirred vessels. In the case of percolation, the solid is packed in a vessel which is filled with solvent. The latter is allowed to percolate in the solid matrix under stagnant conditions. In the case of extraction in stirred vessels, different types of agitators are used to suspend the solid in the solvent and accelerate the mass transfer process. In both percolation and extraction in stirred vessels, the solvent is first sorbed by the matrix of the solid. This sorption, which causes swelling of the matrix, is a relatively slow process. However, once the matrix is swollen, the diffusion coefficient increases several fold or even by an order of magnitude as compared to the dry matrix. Evidently the controlling step is the diffusion of the solute through the solid matrix to the surface of the solid. Once the solute is available at the surface, the solvent can dissolve it depending upon the rate of transport from the solid surface into the bulk of the solvent. In percolation vessels, this latter transport is predominantly by molecular diffusion and hence is slow, although not as slow as the transport through the solid matrix. The stirred vessels, on the other hand, provide a high level of turbulence and hence facilitate transport into the bulk solvent phase. In both percolation and stirred vessels, the dominant resistance is diffusion through the solid matrix. It is then clear that even stirred vessels with high power inputs may not intensify the mass transfer process. Therefore, instead of focusing on the transport at the solid surface, it is desirable to increase the rate of transport through the solid matrix by rupturing the cells which contain the solute or oil and consequently bring the same in direct contact with the solvent.

8.3.3 Microwave-assisted Extraction

8.3.3.1 Principle of Microwave Heating

Microwave radiation interacts with dipoles of polar and polarizable materials. The coupled forces of electric and magnetic components change direction rapidly (2450 MHz). Polar molecules try to orient in the changing field direction and hence get heated. In non-polar solvents without polarizable groups, the heating is poor (dielectric absorption only because of atomic and electronic polarizations). This thermal effect is practically instantaneous at the molecular level but limited to a small area and depth near the surface of the material. The rest of the material is heated by conduction. Thus, large particles or agglomerates of small particles cannot be heated uniformly, which is a major drawback of microwave heating. It may be possible to use high power sources to increase the depth of penetration. However, microwave radiation exhibits an exponential decay once inside a microwave-absorbing solid.

The various industrial techniques used for heating are listed in Table 1, which shows that microwaves have the highest efficiency when compared with the other competitive techniques.

8.3.3.2 Mechanism of MAE

In microwave-assisted extraction (MAE): 1) the heat of the microwave irradiation is directly transferred to the solid without absorption by the microwave-transparent solvent; 2) the intense heating of step 1 causes instantaneous heating of the residual microwave-absorbing moisture in the solid; 3) the heated moisture evaporates, creating a high vapor pressure; 4) the vapor pressure generated by the moisture breaks the cell; and 5) breakage of cell walls releases the oil trapped within it (Figure 1).

Table 1: Relative efficiencies of common heating devices

Appliance	Temperature, °C	Rating, W	Time	Energy used, kWh	Energy cost, US$
Electric oven	177	2000	1 h	2	0.17
Convection oven	163	1853	45 min	1.39	0.12
Gas oven	177	36	1 h	3.57	0.07
Microwave oven	High	1440	15 min	0.36	0.03

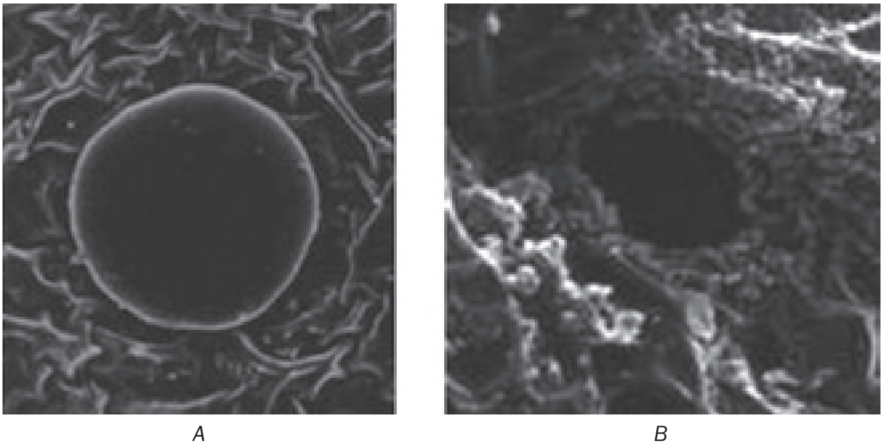

Figure 1: Mint gland: (A) before and (B) after microwave irradiation (Microphotographs courtesy of Radient Technologies Inc.)

It is evident then that the main resistance to solid-liquid mass transfer, the transport of the solute through cell membrane, is eliminated because of the rupture of the cells. Besides cell breakage, the other advantages of microwave heating are:

1. Improved “existing” products
2. Increased marker recovery
3. Increased purity of the extract
4. Reduced heat degradation
5. Reduced processing costs
6. Significantly faster extraction
7. Much lower energy usage
8. Much lower (order of magnitude) solvent usage
9. Potential for “new” products

8.3.3.3 Literature on MAE

Some interesting results on MAE have recently been published. For example, the extraction of vanillin from *V. planifolia* pods using MAE and ultrasound-assisted extraction has been described. Using absolute ethanol as the solvent at room temperature, the yield of vanillin was 1.25 wt% at each of 3 conventional extractions performed over 24 h. Using ultrasound-assisted extraction, the yield was 0.99 wt%, while it was 1.86 wt% using MAE. These investigations clearly showed that vanillin extraction by MAE is superior to other techniques in terms of yield, purity of vanillin, and the time taken to extract the same percentage of the vanillin from the pods. The extraction of vanillin and *p*-hydroxy benzaldehyde (PHB) from vanilla beans using MAE has also been studied: MAE was superior to the conventional, official method of extraction in Mexico, which involves maceration of the beans with ethanol for 12 h. Specifically, extraction time decreased 62-fold and vanillin and PHB concentrations increased between 40% and 50% with respect to the Mexican extraction method. This study also showed that extraction of commercial samples was superior to extraction of dried and lyophilized beans. This observation illustrates the role played by moisture in aiding extraction, as discussed in Section 8.3.3.2. Several other investigations have shown that MAE has gained acceptance as a mild and controllable processing tool. MAE is a simple, rapid and low-solvent-consuming process.

8.3.3.4 Industrial-scale MAE

As mentioned earlier, microwave radiation decays exponentially inside a solid matrix. This aspect must be carefully weighed while designing industrial-scale MAE. The major requirements that must be met are:

1. Free distribution of particles allows uniform heating of all the particles in the solid bed. This criterion also enhances the extent and probability of proximity of the substrate to the wall of the sample holder where the microwave exposure is highest. Most comminuted samples of MAPs which are used for commercial extraction are not of the same shape and size. Therefore, there is a strong tendency to “segregate”, which must be curbed by regular renewal of the layer.
2. Thin and uniform spreading of the substrate layers. This permits complete and uniform penetration of microwave radiation even at large water contents.
3. Low depth of the layers. Since microwaves have low penetration depth (~1.5 cm in H_2O at 2.45 GHz), the layers should be <1.5 cm thick.

Large-scale commercial (3 tonne/hour) MAE is available for industrial use (www.radientinc.com). In view of the advantages of MAE and the development of equipment for large-scale commercial operation, MAE has a bright future. Figure 2 shows a flowsheet for industrial-scale MAE.

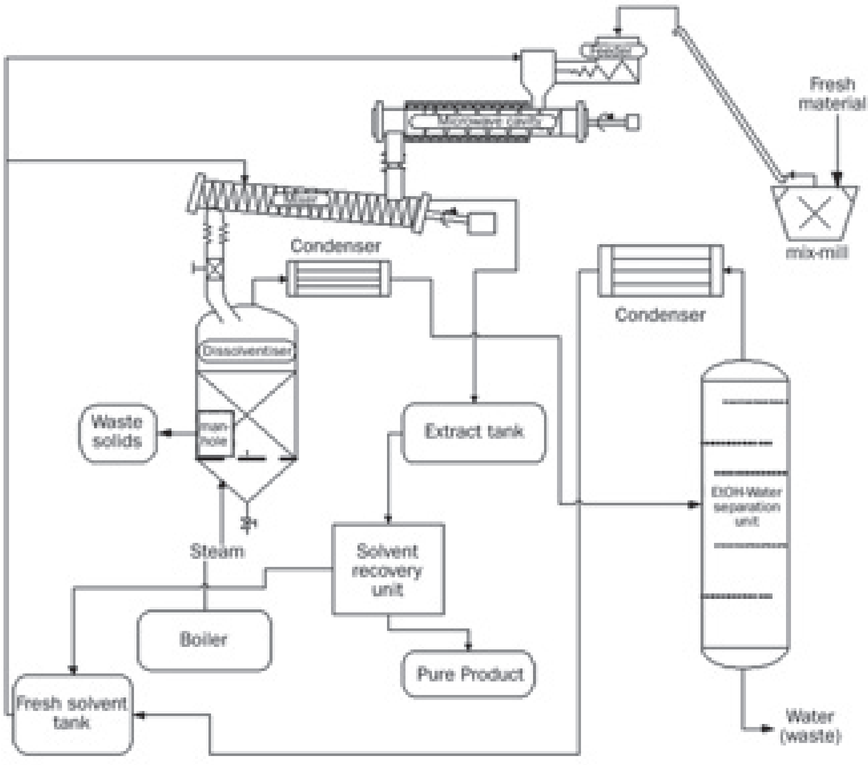
Fresh material
Microwave cavity
Mixer
mix-mill
Condenser
Condenser
Dissolventiser
Waste solids
man-hole
Extract tank
EtOH-Water separation unit
Steam
Solvent recovery unit
Boiler
Pure Product
Fresh solvent tank
Water (waste)

Figure 2: Flowsheet of microwave-assisted extraction (courtesy, Radient Technologies Inc).

8.4 Microwave-assisted Hydrodistillation

The ability of microwave radiation to heat solid material effectively can also be used for obtaining essential oils. Thus, the herb is placed in a microwave cavity and irradiated with microwaves. This process yields essential oils consisting of relatively low volatile fractions as compared to hydrodistillation. For instance, in coriander oil, the percentage of tetradecanoic and hexadecanoic acid increased whereas that of linalool decreased. This is possibly due to the poor stability of linalool, a tertiary alcohol.

Dill seed oil obtained by microwave-assisted hydrodistillation (MWAHD) contained greater quantities of compounds with higher boiling points and lesser quantities of compounds of lower stability. These and other findings indicate that MWAHD is better for extracting stable, high-boiling point components, whereas it is not suitable for recovering chemically unstable compounds.

8.5 Molecular Distillation or Short Path Distillation

Molecular distillation (MD), also known as short path distillation, is a fairly well established technique. In view of this, the discussion of MD is restricted to its principle, advantages and applications in the processing of MAPs.

8.5.1 Principle of MD

The term MD refers to a non-equilibrium process. The still used has an evaporating surface very close to a condensing surface. Under very low pressures, this results in a situation where the distance traveled by the evaporating molecules is comparable to the mean free path of the molecules. The nomenclature MD is derived from this particular condition under which the so-called distillation is carried out.

8.5.2 Advantages of MD

1. Operating pressures as low as 0.001 mbar can yield relatively low processing temperatures, thereby reducing thermal degradation.
2. Agitated film MD units can process high viscosity feeds with very good turndown.
3. Combination of low pressures and high temperatures (up to 300° C) allows processing of extremely high-boiling materials without degradation.
4. Short exposure to high temperature (low residence time) prevents degradation.
5. Very low liquid hold-up allowing use in applications involving low volume, high value materials.
6. Available in low (laboratory scale) to high heat transfer areas to suit the requirements.

8.5.3 Separation Efficiency of MD

For high viscosity liquid films falling under gravity, agitated film MD units perform far better than those without agitation of the film. This is due to the fact that, particularly for high viscosity liquids, the agitation of the film renews the surface more frequently than when there is no agitation. The surface renewal model is useful for predicting the efficiency η of MD stills without mechanical control.

8.5.4 Parameters that Affect the MD Process

The amount of low-boiling volatiles as well as dissolved air, moisture, or other gasses in the feed material has a deleterious effect on the efficiency of MD. This is due to the fact that the non-condensable components cover the condensing surface.

Higher temperature difference between the condensing and evaporating surfaces yields higher efficiency. High viscosity liquids (without mechanical agitation) yield high liquid film thicknesses and hence lower efficiency. As a rule, the relative volatility of organics increases with decreasing pressure, particularly in the very low pressure range common to MD. Therefore, low operating pressure generally yields higher efficiency.

8.5.5 Typical Applications of MD

1. Concretes obtained by solid-liquid extraction are conventionally converted to the absolutes by dissolving in aqueous alcohol solvents and then precipitating the waxes by chilling to sub-zero temperatures. This process is highly energy intensive due to the electrical energy required for refrigeration.
2. Red palm oil (high vitamin E content).
3. Separation of tocopherols from vegetable oil deodorization residues.
4. Natural vitamins A, E, K-1 and K-2 (replacing synthetics in the pharmaceutical industry).
5. Purification and separation of natural extracts into crude fractions.
6. Recovery of lanolin from wool grease, the soft wax from hair of sheep (cosmetic industry).
7. Fragrances derived from fatty acids.

8.6 Recovery of Dissolved Essential Oils from Steam Distillation Condensates

The major prerequisite of the process used for production of essential oils is that the product obtained must resemble the natural aroma and flavor of the original source, which is a combination of different compounds of varying organoleptic characteristics. Oxygenated organic compounds like aldehydes, ketones, alcohols and esters are the dominant contributors to the overall aroma and flavor. The essential oil produced should ideally have all these components in the same proportions as in the original natural product in order to match the natural aroma and flavor. For example, steam-distilled rose oil contains less than 1 wt% phenyl ethyl alcohol (PEA) whereas the solvent-extracted rose oil contains greater than 60 wt% PEA. It is a common experience that the steam distillation condensate has an odor similar to that of the oil. Thus, this condensate has some value. The sale of rose water (otto of rose) for use in weddings in the sub-continent is probably the only way by which the distiller realizes the value of the condensate. In practically all other cases, the condensate is wasted. Table 2 gives estimates of the values of wasted oil in the condensate in India for some essential oils. The estimate is conservative because it does not include oil physically carried with the condensate. Even this conservative estimate is a mind-boggling number and is particularly important in the social context of developing countries like India where marginal farmers are main contributors to the overall produce. The value of the recovered oil from central distillation facilities and pro rata distribution of the value will be a big bonus to the small farmer and can certainly stop the downslide. Table 2 shows a major contribution from *Mentha arvensis*. If the relatively high value flavor sector is included, for India alone the value of wasted oil can easily reach US$ 100 million. On a rough estimate, the combined number for the South East Asian countries can be upwards of US$ 160 million. It must be noted that these numbers are simple statistics. They do not reflect in any way the high value that can be gained when the recovered oil is blended with the main distilled oil fraction to obtain a premium grade of the respective essential oil.

Table 2: Loss of essential oils in distillation condensate water. Production figures are for India only

Essential oil	**Production, 2006, million tonnes**	**Unit price, 2006**	**Volume, 2007**	**Oil lost in water, kg***	**Value of oil lost in water, $/yr**
Arvensis	28,000	$ 14/kg	35,000 MT	5.6-7x10^6	46-58x10^6
Basil	100	$ 8/kg	100 MT	1x10^4	8.2x10^4
Citrodora	100	$ 8/kg	100 MT	1x10^4	7.4x10^4
Citronella	300	$ 8/kg	300 MT	3x10^4	2.52x10^5
Peppermint	450	$ 23/kg	450 MT	4.5x10^4	1.3x10^5
Spearmint	250	$ 23/kg	300 MT	2.5x10^4	7.2x10^4
Total	29,200		36,200 MT	5.72-7.72x10^6	47-59x10^6

** Data refer to 100 kg water per kilogram oil. Solubility of oil in water is 1000 ppm.*

8.6.1 Polymeric Adsorption Process

Various techniques, such as cohobation, poroplast extraction and adsorption, which can be used to recover the dissolved substances, have been discussed in the literature. Polymeric adsorbents can be advantageously used to recover dissolved essential oil components. Several investigators have established the utility of adsorption in this context beyond doubt. One study showed that although cis-rose oxide could not be detected in the condensate, this valuable component was found in the recovered oil in significant proportions. These investigations show that more than 95% of the oil in the condensate can be recovered. The polymeric adsorbents used are hard cross-linked macroreticular beads which can be used in adsorption-regeneration cycles practically indefinitely. The technique is simple to use and does not require sophisticated instrumentation as the breakthrough can be judged from the smell of the water coming out of the adsorbent bed. The regeneration of the spent bed can be done using low-boiling alcohols or ketones, and the eluate can be distilled in a relatively short distillation column to obtain a relatively high boiling oil fraction.

8.6.2 Pervaporation Process

Membrane separation processes have been receiving increasing attention particularly for situations involving recovery from relatively dilute (~1000 ppm) aqueous solutions. Pervaporation is one such process which yields very high (~1000 ppm or more) selectivity in the very dilute solution range. Essential oil components which have high affinity for organophilic polymers can be recovered at very high selectivities. One study showed that silicone rubber membranes yielded bold menthol crystals when the *Mentha* condensate water was studied under the pervaporation mode. Similar results were also obtained for basil water. Subsequent studies showed that the high selectivity of properly selected membranes results in a permeate concentration far exceeding the solubility limit of the organics, resulting in phase separation.

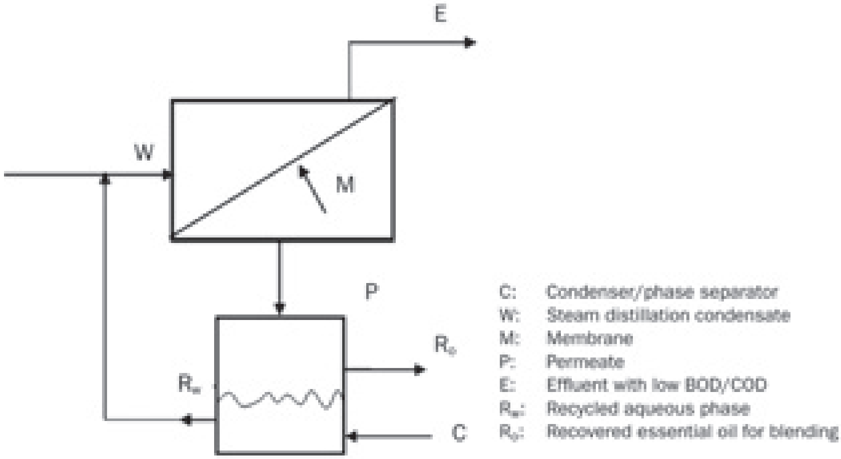
E
W
M
P
R_o
R_w
C
C: Condenser/phase separator
W: Steam distillation condensate
M: Membrane
P: Permeate
E: Effluent with low BOD/COD
R_w: Recycled aqueous phase
R_o: Recovered essential oil for blending

Figure 3: Recovery of dissolved organics using pervaporation

The separate oil layer can be directly recovered to blend with the main oil fraction to obtain premium grade oil. Figure 3 shows a schematic of pervaporation-based recovery of dissolved essential oils in the condensate. It is evident that this technique consists of a closed-loop operation with only treated water going out of the battery limits. This treated water has very low biochemical oxygen demand (BOD) and chemical oxygen demand (COD), which is another bonus for the processor.

8.7 Conclusions

Various new technologies for efficient and cost-effective extraction of medicinal and aromatic plants have been discussed. Microwave-assisted extraction (MAE) is highly efficient for obtaining extracts under mild conditions. MAE is particularly important since the active components which are thermally labile can be recovered without any damage. The loss of valuable aroma components in steam distillation condensates is estimated to be of the order of US$ 50 million per year from aroma oils for India alone. Two types of separation processes – adsorptive and membrane-based pervaporation – are useful in recovering practically all the oil that is lost with the condensate water. The recovered oil can be sold as such or blended with the main oil fraction to yield a much more natural aroma and hence a high value. This recovered oil will be a big bonus even for the marginal farmer and hence this approach needs to be seriously considered.

Bibliography

Amin, L. P., Pangarkar, V. G. and Beenackers, A. A. C. M., 2001, Recovery of valuable perfumery compounds from geranium steam distillation condensates by polymeric adsorbents, *Separation Science and Technology*, 36(16): 3639-3655

Bohra, P. M., Vaze, A. S. and Pangarkar, V. G., 1994, Adsorptive recovery of water soluble essential oil components, *Journal of Chemical Technology & Biotechnology*, 60: 97-102

Gaidhani, H. K., Tolani, V. L., Pangarkar, K. V. and Pangarkar, V. G., 2002a, Intensification of enzymatic hydrolysis of penicillin G: 2. Model for enzymatic reaction with reactive extraction, *Chemical Engineering Science*, 57(11): 1985-1992

Gaidhani, H. K., Wasewar, K. L. and Pangarkar, V. G., 2002b, Intensification of enzymatic hydrolysis of penicillin G: 1. Equilibria and kinetics of extraction of phenyl acetic acid by alamine 336, *Chemical Engineering Science*, 57(11): 1979-1984

Grulke, E. A., 1975, Solubility parameter values, In: Immergut, J. and Grulke, E. A. (Eds.), Polymer Handbook, Wiley-interscience Publications, New York, Vol. VII, p. 675

Guenther, E., 1952, The Essential Oils, Robert E. Krueger Pub. Co., Florida, Vol. I, p. 10-12

Hansen, C. M., 1967, The three dimensional solubility parameters, Key to paint component affinities, *Journal of Paint Technology*, 505: 104-112

Jadhav, S. V. and Pangarkar, V. G., 1989, Gas-liquid and solid-liquid mass transfer in three phase sparged reactors with and without ultrasound, *Journal of the American Oil Chemists' Society* (JAOCS), 66(3): 362-364

Jogdeo, D. A., Niranjan, K. and Pangarkar, V. G., 2000, Recovery of alkyl isothiocyanate from steam distillation condensate using adsorption, *Journal of Chemical Technology & Biotechnology*, l75(8): 673-680

Kanani, D. M., Nikhade, B. P., Balakrishnan, P., Singh, G. and Pangarkar V. G., 2003, Recovery of valuable tea aroma components from steam distillation condensate by pervaporation, *Industrial & Engineering Chemistry Research*, 42(26): 6924-6932

Kaufmann, B. and Christen, P., 2002, Recent extraction techniques for natural products: microwave-assisted extraction and pressurized solvent extraction, *Phytochemical Analysis*, 13(2): 105-113

Kosar, M., Özek, T., Göger, F., Kürkcüoglu, M. and Baser, K. H. C., 2005, Comparison of microwave-assisted hydrodistillation and hydrodistillation methods for the analysis of volatile secondary metabolites, *Pharmaceutical Biology*, 43(6): 491-495

Longares-Patron, A. and Canizares-Macias, M. P., 2006, Focused microwaves-assisted extraction and simultaneous spectrophotometric determination of vanillin and *p*-hydroxybenzaldehyde, *Talanta*, 69: 882-887

Machale, K. W., Niranjan, K. and Pangarkar, V. G., 1997, Recovery of dissolved essential oils from condensate water of basil and *Mentha arvensis* distillation, *Journal of Chemical Technology & Biotechnology*, 69(3): 362-366

Nelida, E. G. and Witte, L., 1999, International Society of Ecology 99–Marseille, Microwave assisted solvent extraction from macerated species, Poster presentation. Available at: http://www.chemecol.org

Netke, S. A., Sawant, S. B., Joshi, J. B. and Pangarkar, V. G., 1995, Comparative study of membranes for pervaporation of trace organics from aqueous solutions, In: Bowen, W. R., Field, R. W. and Howell, J. A. (Eds.), Proceedings of Euro-membrane-95 Conference, p. 116-121

Pangarkar, V. G., 2002, Process Intensification in Chemical Industry, Guest Editorial: *Chemical Industry News*, Indian Chemical Manufacturers Association (ICMA), XLVII (5): 5-6

Pangarkar, V. G., Yawalkar, A. A., Sharma, M. M. and Beenackers, A. A. C. M., 2002, Particle–liquid mass transfer coefficient in two/three-phase stirred tank reactors, *Industrial & Engineering Chemistry Research*, 41(17): 4141-4167

Pare, J. R. J., 1995, Microwave-assisted extraction from materials containing organic matter, United States Patent 5458897

Save, S. V. and Pangarkar, V. G. 1994, Liquid-liquid extraction using aphrons, *Separation Technology*, 4(2): 104-111

Save, S. V., and Pangarkar, V.G., 1995, Harvesting of *Saccharomyces cerevisiae* using colloidal gas aphrons, *Journal of Chemical Technology & Biotechnology*, 62(2): 192-199

Save, S. V., Pangarkar, V. G. and Kumar, S. V., 1993, Intensification of mass transfer in aqueous two-phase systems, *Biotechnology & Bioengineering*, 41(1): 72-78

Shanks, J. V, 2005, Phytochemical engineering: combining chemical reaction engineering with plant science, *AIChE Journal*, 51(1): 2-7

Sharma, A., Verma, S. C., Saxena, N., Chadda, N., Singh, N. P. and Sinha, A. K., 2006, Microwave- and ultrasound-assisted extraction of vanillin and its quantification by high-performance liquid chromatography in *Vanilla planifolia*, *Journal of Separation Science*, 29: 613-619

Sherman, C. B., Huibers, P. D. T. Garcia-Valls, R. and Hatton, T. A., 1998, Solvent replacement for green processing, *Environmental Health Perspectives Supplements*, 106(S1): 202-209

Stankiewicz A. I. and Moulijn J. A., 2000, Process intensification: transforming chemical engineering, *Chemical Engineering Progress*, 96: 22-34

Wasewar, K. L., Heesink, A. B. M., Versteeg, G. F. and Pangarkar, V. G., 2003, Intensification of enzymatic conversion of glucose to lactic acid by reactive extraction, *Chemical Engineering Science*, 58(15): 3385-3393(9)

Wilson, A., Thorne, J. and Morrill, J., 2003, Consumer Guide to Home Energy Savings, 8th Edition, Washington DC: American Council for Energy Efficient Economy

Zanwar, S. S. and Pangarkar, V. G., 1988, Solid liquid mass transfer in packed beds: Enhancement due to ultrasound, *Chemical Engineering Communications*, 68: 133-140

9 Solid Phase Micro-extraction and Headspace Trapping Extraction

R. Harlalka

Abstract

Solid phase micro-extraction (SPME) is a technique used in the quantitative analysis of analytes in aqueous and gaseous phases. This novel technology captures aroma molecules surrounding flower petals without touching the flower or other part of the plant. SPME has gained widespread acceptance as the technique of choice in many fields of application, including forensics, toxicology, and the analysis of flavors, fragrances, and environmental and biological matrices. SPME is ideal for field monitoring. SPME sampling can be performed in three basic modes: direct extraction, headspace trapping and extraction with membrane protection. Headspace trapping is essentially a gas extraction technique permitting the direct analysis of volatile compounds present in a non-volatile matrix. This technique is needed because the aromas of living plant materials are different from those of the extracted oil. Headspace trapping permits getting closer to the natural aroma of the living plant, and gives a clearer view of the differences in volatile constituents between the living plant and the extracted phase. There are two types of headspace trapping: static and dynamic, which is also called the purge-and-trap method. A few examples of headspace trapping of well known aromatic flowers, fruits and leaves, in comparison to the analyses of the extracted oil, are presented in this paper. Some classical perfumes are also discussed.

9.1 Introduction

Solid phase micro-extraction (SPME) was developed in the 1990s by Professor J. Pawliszyn to provide a quick and solventless technique for the isolation of analytes from a sample matrix. The traditional methods by which the analytes of interest were isolated are typically time- and labor-intensive and involve multistep procedures, which could reduce sensitivity. Also, the use of solvents can be hazardous to the operators' health and can damage the environment.

SPME was developed from the technique of solid phase extraction, but the sorbing material is permanently attached to the fiber, allowing reuse of the extracting phase. SPME uses a small volume of sorbent, typically dispersed on the surface of small fibers, to isolate and concentrate analytes from the sample matrix. After contact with the sample, analytes are absorbed or adsorbed by the fiber phase (depending on the nature of the coating). After the extraction step, the fibers are transferred, with a syringe-like handling device, to the analytical instrument, for separation and quantification of the analytes. This technique integrates sampling, extraction and sample introduction, and is a simple way of performing on-site monitoring. Applications of this technique include environmental monitoring, fragrance drug analysis, and in-laboratory and on-site analyses.

SPME was introduced in 1990 as a solvent-free sample preparation technique. The basic principal of this approach is to use a small amount of the extracting phase, usually less than 1 microliter. Sample volume can be large when the investigated material is sampled directly, e.g. the air in a room. The extracting phase can be either a high molecular weight polymeric liquid, similar in nature to stationary phases in chromatography, or a solid sorbent, typically of a high porosity, to increase the surface area available for adsorption.

The configuration of SPME is a small, fused silica fiber, usually coated with a polymeric phase. The fiber is mounted for protection in syringe-like device. The analytes are absorbed or adsorbed by the fiber phase until equilibrium is reached in the system. The amount of an analyte extracted by the coating at equilibrium is determined by the magnitude of the partition coefficient of the analyte between the sample matrix and the coating material.

In SPME, analytes typically are not extracted quantitatively from the matrix. Equilibrium methods are more selective because they take full advantage of the difference between extracting phase and matrix distribution constants to separate target analytes from interferences. Exhaustive extraction can be achieved in SPME, and this can be accomplished for most compounds by the application of an internally cooled fiber. In exhaustive extraction, selectivity is sacrificed to obtain quantitative transfer of target analytes to the extracting phase.

SPME is ideal for field monitoring. It is unnecessary to measure the volume of the extracted sample, and therefore the SPME device can be exposed directly to the investigated material for quantification of analytes of interest. In addition, extracted analytes are introduced into the instrument by simply placing the fiber in the desorbtion unit. This convenient, solvent-free process results in sharp injection bands and rapid separations.

9.2 The SPME Device

The commercial SPME device manufactured by Supelco (Bellefonte, USA) is presented in Figure 1. The fiber glued into a piece of stainless steel tubing is mounted on a special holder. The holder is equipped with an adjustable depth gauge, which makes it possible to control repeatedly, how far the needle of the device penetrates the sample container or the injector. This is important, as the fiber can break if it hits an obstacle. The movement of the plunger is limited by a small screw that moves in the z-shaped slot of the device. For protection during storage or septum piercing, the fiber is withdrawn into the needle of the device, with the screw in the uppermost position. During extraction or desorption, the fiber is exposed by depressing the plunger. The plunger is moved to its lowermost position only for replacement of the fiber assembly. Each type of fiber has a hub of a different color.

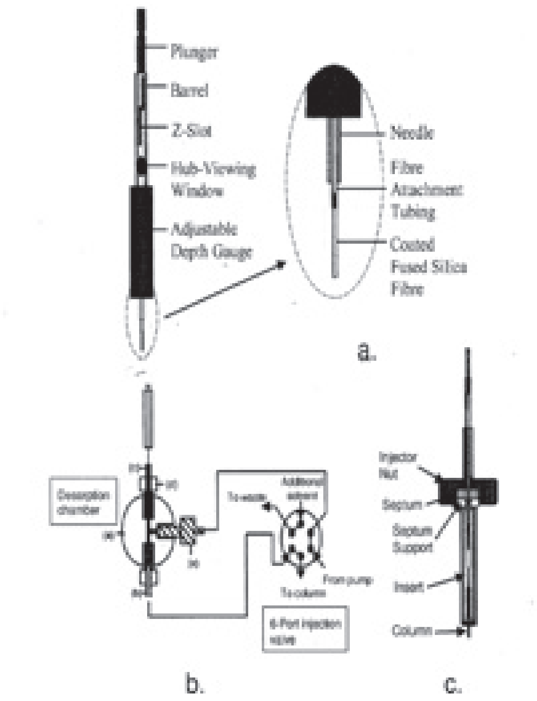
Plunger
Barrel
Z-Slot
Hub-Viewing Window
Adjustable Depth Gauge
Needle
Fibre Attachment Tubing
Coated Fused Silica Fibre
a.
Desorption chamber
To waste
Additional solvent
From pump
To column
6-Port injection valve
b.
Injector Nut
Septum
Septum Support
Insert
Column
c.

Figure 1: The SPME device

If the sample is in a vial, the septum of the vial is first pierced with the needle (with the fiber in the retracted position), and the plunger is lowered, which exposes the fiber to sample. The analytes are allowed to partition into the coating for a pre-determined time, and the fiber is then retracted back to the needle. The device is then transferred to the SPME instrument. When gas chromatography is used for analyte separation and quantification, the fiber is inserted into a hot injector, where thermal desorption of the trapped analyte takes place.

For spot sampling, the fiber is exposed to a sample matrix until partitioning equilibrium is reached between sample matrix and the coating material. In the time average approach, on the other hand, the fiber remains in the needle during exposure of the SPME device to the sample. The coating works as a trap for the analytes that diffuse into the needle, resulting in integral concentration over time measurements.

SPME sampling can be performed in three basic modes: direct extraction, headspace trapping, and extraction with membrane protection.

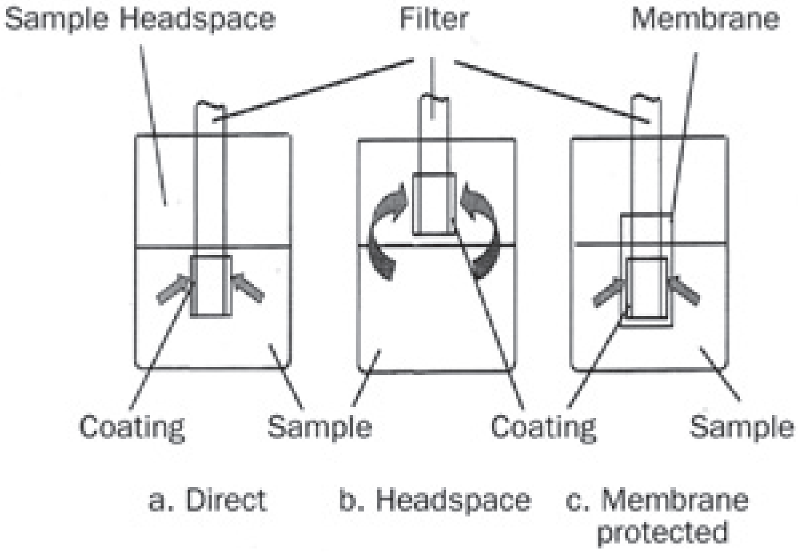

Figure 2: Modes of SPME operation: direct extraction (a), headspace trapping (b) and membrane-protected SPME (c)

In direct extraction, the coated fiber is inserted into the sample and the analytes are transported directly from the sample matrix to the extracting phase. To facilitate rapid extraction, some agitation is required to transport the analytes from the bulk of the sample to the vicinity of the fiber. For gaseous samples, natural flow of air (e.g. convection) is usually sufficient to facilitate rapid equilibrium for volatile analytes.

In headspace mode, the analytes are extracted from the gas phase equilibrated with the sample. The primary reason for this modification is to protect the fiber from the adverse effects caused by non-volatile, high molecular weight substances present in the sample matrix (e.g. human acids or proteins). Here, the amount of an analyte extracted by the fiber coating does not depend on the location of the fiber, in the liquid or gas phase; therefore, the sensitivity of headspace trapping is the same as that of direct sampling as long as the volumes of the two phases are the same in both sampling modes. When no headspace is used in direct extraction, a significant sensitivity difference between direct and headspace trapping can occur only for very volatile analytes. However, the choice of sampling mode has a significant impact on the extraction kinetics. When the fiber is in the headspace, the analytes are removed from the headspace first, followed by indirect extraction from the matrix.

In general, the equilibration times for volatile compounds are shorter for headspace SPME than for direct extraction under similar agitation conditions, because of the following reasons: a substantial portion of the analytes is present in the headspace prior to the beginning of the extraction process; there is typically a large interface between sample matrix and headspace; and the diffusion coefficients in the gas phase are typically higher by four orders of magnitude than in liquids. The concentration of semivolatile compounds in the gaseous phase at room temperature is small, and headspace extraction rates for these compounds are substantially lower. They can be improved by using efficient agitation or by increasing the extraction temperature.

In the third mode (SPME extraction with membrane protection), the fiber is separated from the sample with a selective membrane, which lets the analytes through while blocking the interferences. The main purpose for the use of the membrane barrier is to protect the fiber against adverse effects caused by high molecular weight compounds when dirty samples are analyzed. While headspace trapping serves the same purpose, membrane protection enables the analysis of less volatile compounds. Use of thin membranes and an increase in extraction temperature result in shorter extraction times.

9.3 Calibration, Optimization, Precision and Suitability of SPME

9.3.1 Selection of Fiber Coating

The chemical nature of the analyte of interest determines the type of coating used. A simple general rule, “like dissolves like”, applies very well for liquid coatings. Selection of the coating is based primarily on the polarity and volatility of the analyte. Poly(dimethylsiloxane) (PDMS) is the most useful coating and should be considered first. It is rugged and able to withstand high injector temperatures, up to about 300° C. PDMS is a non-polar liquid, thus it extracts non-polar analytes very well with a wide linear dynamic range. However, it can also be applied successfully to more polar compounds, particularly after optimizing extraction conditions.

Both the coating thickness and the distribution constant determine the sensitivity of the method and the extraction time. Thick coatings offer increased sensitivity, but require much longer equilibration times. As a general rule, to speed up the sampling process, the thinnest coating offering the sensitivity required should be used.

9.3.2 Selection of the Extraction Mode

Extraction mode selection is based on the sample matrix composition, analyte volatility, and its affinity to the matrix. For dirty samples, the headspace or fiber-protection mode should be selected. For clean matrices, both direct and headspace trapping can be used. The latter is applicable for analytes of medium to high volatility. Headspace trapping is always preferential for volatile analytes because the equilibration times are shorter in this mode than in direct extraction. Fiber protection should be used only for dirty samples in cases where neither of the first two modes can be applied.

9.3.3 Selection of the Agitation Technique

Equilibration times in the gaseous samples are short and frequently limited only by the rate of diffusion of the analytes in the coating. When the aqueous and gaseous phases are at equilibrium prior to the beginning of the sampling process, most of the analytes are in the headspace. As a result, the extraction times are short even when no agitation is used. However, for aqueous samples, agitation is required in most cases to facilitate mass transport between the bulk of the aqueous sample and the fiber.

Magnetic stirring is most commonly used in manual SPME experiments. Care must be taken when using this technique to ensure that the rotational speed of the stirring bar is constant and that the base plate does not change temperature during stirring. This usually implies the use of high quality digital stirrers. Alternatively, with cheaper stirrers, the base plate should be thermally insulated from the vial containing the sample to eliminate variations in sample temperature during extraction. Magnetic stirring is efficient when fast rotational speeds are applied.

9.3.4 Selection of Separation or Detection Technique

Most SPME applications have been developed for gas chromatography (GC), but other separation techniques, including high performance liquid chromatography, capillary electrophoresis (CE) and supercritical fluid chromatography, can be used in conjunction with this technique. The complexity of the extraction mixture determines the proper quantitative device. Regular chromatographic and CE detectors can normally be used for all but the most complex samples, for which mass spectrometry (MS) should be applied.

9.3.5 Optimization of Desorption Conditions

Standard gas chromatographic injectors, such as the popular split-splitless types, are equipped with large volume inserts to accommodate the vapors of the solvent introduced during liquid injections. As a result, the linear flow rates of the carrier gas in those injectors are very low in splitless mode, and the transfer of the volatilized analytes onto the front of the GC column is also slow. No solvent is introduced during SPME injection; therefore, the large insert volume is unnecessary. Opening the split line during SPME injection is not practical, since it results in reduced sensitivity. Efficient desorption and rapid transfer of the analytes from the injector to the column require high linear flow rates of the carrier gas around the coating. This can be accomplished by reducing the internal diameter of the injector insert, matching it as closely as possible to the outside diameter of the coated fiber.

9.3.6 Optimization of Sample Volume

The volume of the sample should be selected based on the estimated distribution constant. The distribution constant can be estimated by using published values for the analyte or a related compound, with the coating selected. The distribution constant can also be calculated or determined experimentally by equilibrating the sample with the fiber and measuring the amount of analyte extracted by the coating.

9.3.7 Determination of the Extraction Time

The equilibration time is defined as the time after which the amount of analyte extracted remains constant and corresponds within the limits of experimental error to the amount extracted after infinite time. Care should be taken when determining the equilibration time, since in some cases a substantial reduction of the slope of the curve might be wrongly taken as the point at which equilibrium is reached. Determination of the amount extracted at equilibrium allows calculation of the distribution constants.

When equilibrium times are excessively long, shorter extraction times can be used. However, in such cases the extraction time and mass transfer conditions have to be strictly controlled to assure good precision. At equilibrium, small variations in the extraction time do not affect the amount of the analyte extracted by the fiber.

On the other hand, at the steep part of the curve, even small variations in extraction time may result in significant variations of the amount extracted. Shorter is the extraction time, larger is the relative error. Autosamplers can measure the time precisely, and the precision of analyte determination can be good, even when equilibrium is not reached in the system. However, this requires that the mass transfer conditions and the temperature remain constant during all experiments.

9.3.8 Optimization of Extraction Conditions

An increase in extraction temperature increases the extraction rate but simultaneously decreases the distribution constants. In general, if the extraction rate is of major concern, the highest temperature that still provides satisfactory sensitivity should be used.

Adjustment of the pH of the sample can improve the sensitivity of the method for basic and acidic analytes. This is related to the fact that unless ion exchange coatings are used, SPME can extract only neutral (non-ionic) species from water. By properly adjusting the pH, weak acids and bases can be converted to their neutral forms, so that they can be extracted by the SPME fiber.

9.3.9 Determination of the Linear Dynamic Range of the Method

Modification of the extraction conditions affects both the sensitivity and the equilibration time. It is advisable, therefore, to check the previously determined extraction time before proceeding to the determination of the linear dynamic range. This step is required if substantial changes in sensitivity occur during the optimization process.

SPME coating includes polymeric liquids, such as PDMS, which by definition have a broad linear range. For solid sorbents, such as Carbowax/DVB or PDMS/DVB, the linear range is narrower because of the limited number of sorption sites on the surface, but it still can span over several orders of magnitude for typical analytes in pure matrices. In some rare cases when the analyte has extremely high affinity to the surface, saturation can occur at low analyte concentrations. In such cases, the linear range can be expanded by shortening the extraction time.

9.3.10 Selection of the Calibration Method

Standard calibration procedures such as external calibration can be used with SPME. The fiber blank should first be checked to ensure that neither the fiber nor the instrument causes interference with the determination. The fiber should be conditioned prior to the first use by desorption in a GC injector or in a specially designed conditioning device. This process ensures that the fiber coating itself does not introduce interference. Fiber conditioning may have to be repeated after analysis of samples containing large amounts of high molecular weight compounds, since such compounds may require longer desorption times than the analytes of interest.

A special calibration procedure, such as isotopic dilution or standard addition, should be used for more complex samples. In these methods, it is assumed that the target analytes behave similarly to spikes during the extraction. This is usually a valid assumption when analyzing homogeneous samples.

9.3.11 Precision of the Method

The most important factors affecting precision in SPME are:

- Agitation conditions
- Sampling time (if non-equilibrium conditions are used)
- Temperature
- Condition of the fiber coating (cracks, adsorption of high molecular weight species)
- Geometry of the fiber (thickness and length of the coating)
- Sample matrix components (salt, organic material, humidity, etc.)

- Time between extraction and analysis
- Analyte loss (adsorption on the walls, permeation of Teflon, absorption by septa)

9.3.12 Suitability

SPME is well suited to the analysis of flavor and fragrance compounds. The typically small, volatile compounds are easily extracted by the fibers, and the simplicity of the method allows easy coupling to analytical instruments. Headspace trapping can reduce the potential for interference peaks and prevent contamination of both the needle and the instrument. Loss of these volatile compounds during sample preparation steps is minimized or eliminated compared to conventional methods, and the method is amenable to field sampling and analysis.

SPME has been shown to be useful for semivolatile compounds, even though these appeared more challenging in the early years. With appropriate matrix modification, one can take advantage of headspace trapping for these as well. SPME provides significant convenience for field and air analysis. Quantification is relatively straightforward, even in the presence of varying air temperature. Finally, the use of SPME for time-weighted average sampling provides simplicity in monitoring flavor and fragrance concentrations over time.

9.4 Headspace Trapping Extraction and GC-FID/MS Analysis

Orange juice volatiles were extracted from the juice headspace using a syringe-like SPME device equipped with a 75 µm Carboxen-PDMS fiber (Supelco). Aliquots (25 ml) of juice were placed in 40-ml glass vials with plastic screw caps and Teflon-coated septa, warmed to 40° C, and gently swirled to coat the walls of the vial. Juices were allowed to equilibrate for at least 15 min prior to fiber insertion and were maintained at 40° C throughout the 35-min extraction period. The fiber was then removed from the headspace and inserted into the heated GC injector, where the volatile compounds were thermally desorbed. Flavor extract was separated using an HP 5890 GC instrument equipped with a 30 m x 0.32 mm i.d. DB5 capillary column. Column temperature was initially 32° C, with a 3-min hold, and was then increased at 6° C/min to 200° C. Helium carrier gas linear velocity was 29 cm/s. A special narrow boar (0.75 mm) injector liner was used to improve peak shape and chromatographic efficiency; the entire separation was conducted in the splitless mode.

9.4.1 History of Headspace

In November 1986, at the 10th International Congress on Essential Oils, in Washington DC, USA, Dr. B. D. Mookherjee presented a paper on the impact of “live vs. dead” on headspace trapping extraction, using as example jasmine flowers.

The SPME needle which is 2- to 3-mm solid glass fiber coated with a high-boiling liquid adsorbent, is placed in close proximity to a flower without touching it and is kept there for a period of 30-60 min depending on the odor strength of the blossom. The aroma molecules around the petals are absorbed onto the fiber. Then with GC/MS, the fiber is analyzed to determine the aroma profile of that particular flower. The aroma of the living flower was brought into space by NASA in 1998.

9.4.2 The Aura

When the Sun is totally eclipsed by the moon, the surrounding glow is called an aura. Similarly, if we consider a drop of fragrance, the molecules surrounding the drop form an aura of that particular fragrance.

It is a common belief that one smells a fragrance, layer by layer, from the top note of the volatile components, to the middle note of components with boiling points in the middle range, and finally to the bottom note of components with the highest boiling point.

In reality, when a drop of fragrance is placed on the skin, several different molecules, from the lowest to the highest boiling types, irrespective of their molecular weights, boiling points and vapor pressures, form an aura, which eventually reaches our nose and gives us our first impression of the particular fragrance. The composition of this aura depends on a characteristic property of each fragrance molecule, knows as its diffusivity.

9.4.3 What is Diffusivity?

Diffusivity is the inherent property of a compound to emit its molecules into the air. One compound is said to be more diffusive than another if its molecules tend to pass into the air to a greater extent than those of other compounds. Diffusivity is independent of boiling point, molecular weight, odor threshold or odor value.

9.4.4 Application of Headspace Trapping

Some examples of headspace trapping are discussed here.

9.4.4.1 Jasmine

The headspace constituents of living and picked *Jasminum grandiflorum* flowers are:

Compound	**Living flowers, %**	**Picked flowers, %**
Benzyl acetate	60.0	40.0
Linalool	3.0	30.0
Indole	11.0	2.0
Cis-jasmone	3.0	–
3,5-Dimethyl-2-ethyl pyrazine	–	0.5
Epi-methyl jasmonate	0.5	–
Methyl jasmonate	0.3	–

Differences in the volatile compounds of living flowers from *Jasminum grandiflorum* and *Jasminum sambac* are:

Compound	***J. grandiflorum*, %**	***J. sambac*, %**
Methyl benzoate	–	5.0
Benzyl acetate	60.0	37.0
Indole	11.0	5.0
Linalool	3.0	9.0
Epi-methyl jasmonate	0.5	–
Methyl jasmonate	0.3	–

9.4.4.2 Yellow Tea Rose

The differences in headspace constituents between living and picked yellow tea rose flowers are:

Compound	**Living flowers, %**	**Picked flowers, %**
Cis-3-hexenyl acetate	20.67	5.39
Hexyl acetate	8.40	4.26
Phenylethyl alcohol	5.73	3.30
3,5-Dimethoxy toluene	9.96	18.58
Alpha-elemene	–	4.07
Geranyl acetone	2.17	–
Dihydro-beta-ionol	–	2.62
Isocaryophyllene	0.30	2.12
Alpha-farnesene	5.83	2.96

9.4.4.3 Passion Flower

Volatile constituents of living passion flower (*Passiflora* spp.) are:

Compound	Living flowers, %
Methyl benzoate	90.3
Methyl salicylate	1.1
Methyl cinnamate	1.6

9.4.4.4 Lotus

The major differences in headspace constituents of living and picked lotus (*Nelumbo nucifera*) are:

Compound	Living flowers, %	Picked flowers, %
Sabinene	6.0	12.0
p-Dimethoxy benzene	18.0	8.0
4-Terpineol	3.0	1.5
Alpha terpineol	9.0	1.0
Cis-jasmone	0.1	–
C15 hydrocarbons	20.0	30.0

9.4.4.5 Lavender

The volatile constituents of living French lavender (*Lavandula dentata)* and English lavender (*Lavandula angustifolia*) are:

Compound	French lavender, %	English lavender, %
Limonene	18	6
1-Octen-3-ol	–	7
Hexyl acetate	11	2
Eucalyptol	9	3
Linalool	7	–
Cis-3-hexenyl acetate	17	13
Borneol	–	2
Cryptone	–	6

9.4.4.6 Chamomile

The volatile constituents of living Roman and German chamomile are:

Compound	**Roman chamomile, %**	**German chamomile, %**
Ethyl 2-methyl butyrate	–	12
Cis-3-hexenyl acetate	3	22
Isobutyl methacrylate	7	–
Isobutyl angelate	18	–
Ocimene	–	1
Iso-amyl angelate	10	–
Isohexyl angelate	10	–

9.4.4.7 Shefali

The volatile constituents of living shefali (*Nycanthus arbortristis*) are:

Compound	**Living flower, %**
Benzyl alcohol	11.2
Phenyl acetaldehyde	9.4
Phenyl ethyl alcohol	6.3
Methyl anthranilate	10.7

9.4.4.8 Spearmint

The major differences between living and picked spearmint are:

Compound	Living plant, %	Picked plant, %
Hexanal	0.5	–
Hexanol	–	–
Beta-pinene	0.8	2.0
Sabinene	0.5	–
Myrcene	8.8	4.0
Alpha-phellandrene	0.7	–
Limonene	18.0	2.0
Cis- and trans-ocimene	1.0	–
Dihydrocarvone	0.6	2.6
Carvone	24.0	70.0
Alpha- and beta-elemene	5.0	0.1
Beta-caryophyllene	4.0	0.1

9.4.4.9 Cinnamon Bark

Comparative analysis of fresh cinnamon bark headspace and commercial oil has revealed:

Compound	Fresh cinnamon bark, %	Commercial oil, %
Cis- and trans-cinnamic aldehyde	80.3	71.7
Eugenol	–	12.7
Ortho-methoxy cinnamaldehyde	0.3	–
Eugenyl acetate	–	0.5

9.4.4.10 Ginger

Comparative analysis of the headspace of fresh ginger root and commercial oil has revealed:

Compound	Fresh ginger root, %	Commercial oil, %
Citral	15.3	1.2
Beta-bisabolene	3.3	6.2
Alpha zingiberene	15.2	34.4
Cis- and trans-alpha-farnesene	13.7	6.0
ar-Curcumene	11.3	4.8
Beta-sesquiphellandrene	8.0	11.8

9.4.4.11 Peach

Volatile constituents of living and picked peach (*Prunus persica*) are:

Compound	Living peach, %	Picked peach, %
Ethyl acetate	6.2	–
Dimethyl disulfide	0.6	–
Cis-3-hexenyl acetate	9.7	–
Methyl octanoate	34.2	7.1
Ethyl octanoate	7.4	11.0
6-Pentyl alpha pyrone	Trace	10.6
Gamma decalactone	2.5	39.2

9.4.4.12 Pineapple

Volatile constituents of the interior and exterior of living pineapple are:

Compound	Exterior, %	Interior, %
Methyl hexanoate	13.3	24.6
Ethyl hexanoate	25.6	4.7
Methyl 3-methylthiopropionate	0.9	0.5
Ethyl 3-methylthiopropionate	1.1	–

9.4.5 Classical Perfumes

Almost all successful classical perfumes are based on floral aromas. Perfumers created them using natural flower oil such as rose and jasmine. Few persons are aware of the fact that fruit and flower oils that are made by extraction of picked material exhibit different aromas from those of the living entities. Examples of classical perfumes based on floral aromas are Amarige (Givenchy), Joy (Jean Patou), White Linen (Estée Lauder), Aura (Hugo Boss), Anais Anais (Cacharel) and Beautiful (Estee Lauder). The difference in composition between the oil and the aura of Amarige is as follows:

Compound	Oil, %	Aura on skin after 60 min, %
Linalool	1.7	17.9
Benzyl acetate	4.9	22.7
Styrallyl acetate	1.2	9.7
Cashmeran	–	0.5
Bacdanol	0.2	0.5
Hedione	29.9	4.9
Cedramber	1.5	4.9
Iso E super	7.1	12.1
Ambrox	0.2	0.1
Benzyl salicylate	32.5	1.1
Muskalactone	0.9	0.4

The examples include Joy by Jean Patou; White Linen by Estée Lauder; Aura of Hugo BOSS; Anais Anais by Cacharel; and Beautiful by Estee Lauder.

9.4.6 Need for Headspace

When we go to a rose-field full of bloomed roses, we detect a pleasant smell in the atmosphere and expect the same fragrance when we use the bottled perfume or 100% genuine essential oil extracted from the same roses. However, this is not true. The fragrance we detect in the field is completely different from the bottled perfume or essential oil, for the following reasons:

a) When a flower or herb is processed to obtain the essential oils, the low volatile compounds cannot always be recovered and often evaporate. These low volatiles are important for aroma.
b) During the processing of an herb, many chemical reactions take place, such as saponification, trans-esterification, polymerization and condensation. These reactions actually change the character of the oil, so that its aroma no longer resembles that of the actual plant and the product is different in composition. Many stereoterpenes, which are highly volatile in nature, cannot be extracted and remain in the herbs. These stereoterpenes do not contribute directly to the odor but, in combination with other ingredients, impart a synergic effect to the overall odor quality.

9.5 Types of Headspace Trapping

Headspace trapping can be static or dynamic, which is generally called the purge-and-trap method. In static headspace trapping, gas extraction is carried out in a single step or in a limited number of steps. On the other hand, the purge-and-trap technique consists of two or three separate steps, the first of which is continuous gas extraction.

9.5.1 Static Headspace Trapping

This is a single-step gas extraction procedure (Figure 3). By thermosetting the sample for a certain time at a preselected temperature, equilibrium is reached between the sample phase and the gas phase of the sample vial. Subsequently, a single aliquot of the headspace is introduced into the carrier gas flow, which then carries it to the column where the volatile compounds are separated in the usual way.

The equilibrium of the two phases in the sample vial is characterized by a partition coefficient (Ki) representing the ratio of the analyte’s concentration in the sample phase and in the gas phase.

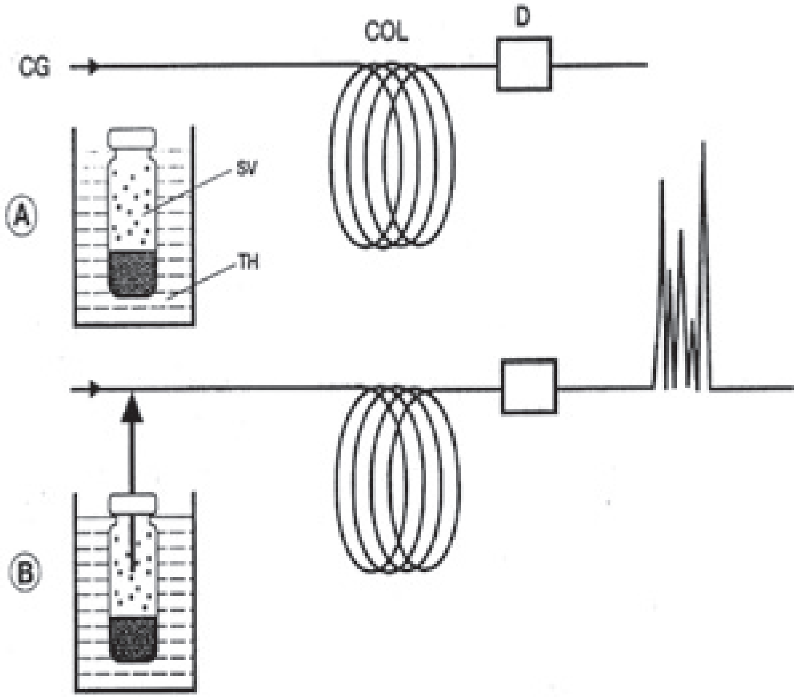

Figure 3: Static headspace trapping technique

9.5.2 Dynamic Headspace Trapping

In this technique (Figure 4), the sample is continuously purged with an inert gas (the purge gas), until all volatile compounds are removed. During this step, the gas effluent leaving the sample vessel is conducted through a trap, either cooled to low temperature or containing an adsorbent. This trap retards the volatile analytes purged from the sample. When gas extraction is complete, the condensed or adsorbed analytes by rapid heating of the trap now get purged with the carrier gas.

In Figure 4, the desorbed analytes are conducted directly into the gas chromatograph. Thermal desorption from an adsorbent is not instantaneous: thus, the sample "slug" might be too long, creating broad peaks, with tailing. This is particularly the case when a capillary column is used in the gas chromatograph. For this reason, usually a second, small trap, cooled to low temperature, is placed in the carrier gas line between the primary trap and the column. When desorption is finished, this small trap is then heated very rapidly: in this way, a sharp band of the analytes enters the column.

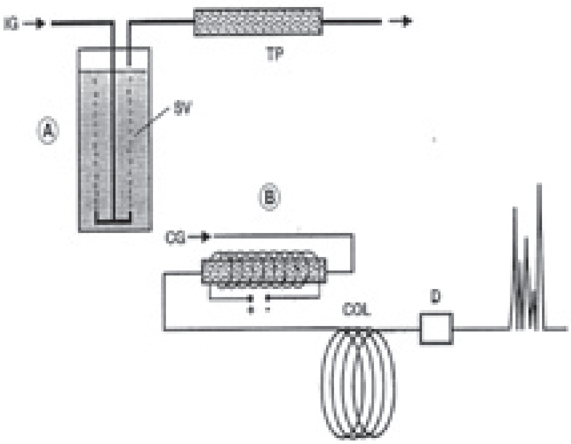

Figure 4: Dynamic headspace trapping technique

This technique is being used more generally after the introduction of Tenax (poly(2,6-diphenyl-*p*-phenylene oxide)) as a universal adsorbent for dynamic headspace GC by Zlatkis and his group, at the University of Houston, in 1973. They used the technique for the investigation of biological fluids and demonstrated the reproducibility of the purge-and-trap method.

9.6 Principles of Static Headspace-GC Systems

Gas from the headspace of a closed vessel can be sampled simply with a gas-tight syringe. However, with such a manual method, it is difficult to reproduce all the conditions necessary for reliable quantitative analysis. Therefore, today, headspace-gas chromatography (HS-GC) is carried out almost exclusively with automated instruments, in which thermosetting, aliquoting the headspace and introducing it into the gas chromatograph are fully automated. In this way and using the proper calibration methods, the required precision, accuracy and reliability are assured.

Present-day HS-GC instruments are of two types. In the first, the headspace aliquot is taken by an automated syringe which then is moved above the injection port of the gas chromatograph and the sample is injected. In essence, such systems are similar to the autosamplers used in GC. In the second case, the aliquot from the vial's headspace is not withdrawn by suction as in the case of a syringe: instead, after equilibrium is reached, the vial is pressurized by the carrier gas. After pressurization there are two possibilities. The carrier gas flow can be temporary interrupted while the pressurized gas in the vial is allowed to expand onto the column; the transferred volume of headspace can be accurately controlled by controlling the time of transfer and the pressure. The second possibility is to have a gas introduced between the sample vial and the column, and fill the sample loop of the valve by the pressurized headspace gas. Today, automated instruments based on these principles are commercially available.

9.6.1 Trace Analysis by HS-GC

HS-GC in both its dynamic and static versions permits the determination of analytes at low concentrations. Usually the dynamic technique is considered to be more sensitive; however, this is not necessarily true. For example, trace impurities in a water sample, at the parts-per-billion level, can be determined relatively easily by static HS-GC.

9.7 Headspace Trapping Techniques

9.7.1 Static Headspace Trapping

Using the static method (Figure 5), a food sample is normally placed in a heated vessel, which is sealed gas-tight by a septum. The food sample stays inside the vessel for a certain period of time, so that the volatile compounds evaporate to a certain concentration in the air or to certain equilibrium. In order to determine the best conditions for the experiment, the odor of the headspace can be checked by sniffing the vessel. Subsequently, a distinct volume is taken out of the vessel by a gas-tight syringe and directly injected into a gas chromatographic column, with or sometimes without prior concentration (e.g. cryofocussing). The advantage of this method is that it accurately assesses the composition of the odorants. An application of this technique, called GC olfactometry of static headspace samples, has been widely used to identify the highly volatile compounds causing the first odor impression of foods.

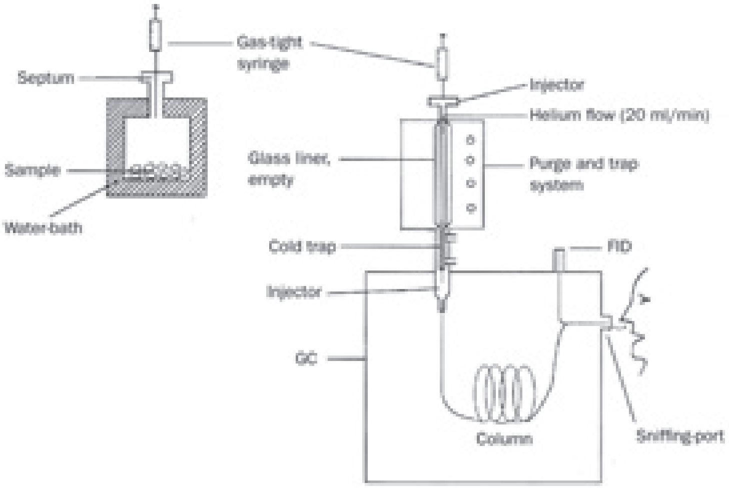

Figure 5: Static headspace trapping technique

However, static headspace samples are normally too small to quantify odorants that are present only at low concentrations in the vapor phase. In other words, one can smell them, but in many cases it is not possible to obtain a signal in a mass spectrometer.

9.7.2 Dynamic Headspace Trapping

To overcome the disadvantages of headspace trapping method, dynamic headspace trapping can be used (Figure 6). Again, the food sample is placed in a heated vessel but the evaporating compounds are continuously swept by a stream of inert gas into a trap containing a porous polymer, which adsorbs more or less the organic constituents. This method yields a much higher amount of trapped volatiles so that, after desorption, it is no longer problematic to obtain an MS signal.

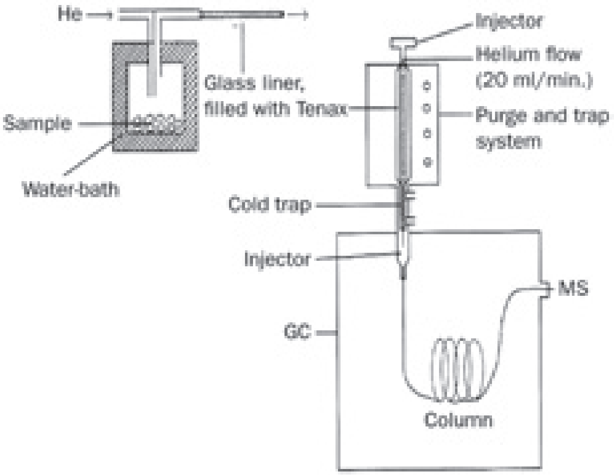

Figure 6: Dynamic headspace trapping technique

However, the disadvantage of this procedure is the strong dependence on the yield of the odorants, on the velocity of the carrier gas and on the selectivity of the adsorption and desorption process for different compounds. It is very difficult to control these parameters precisely and therefore, the results of such quantitative measurements might be inaccurate.

9.7.3 Recovering the Adsorbed Volatiles by Thermal or Liquid Solvent Desorption

Several studies have reported methods of desorption using organic solvents. Drawbacks of the use of solvent desorption include the loss of volatile compounds during removal of excess solvent before GC analysis, solvent selectivity and solvent impurities. We recently developed a sensitive and highly reproducible dynamic headspace (DHS) protocol with thermal desorption (using injector glass liners packed with Tenax-TA as adsorbent traps for aroma collection at ambient room temperature) and desorption at the interior of a GC injector. This DHS-type protocol was used to characterize fresh tomato flavor compounds; the results were compared with published data from a static headspace method (Table 1).

Table 1: Concentration of selected tomato aromas from heat-processed tomato juice by static headspace trapping (SHT) and dynamic headspace trapping (DHT), expressed in parts per billion (ppb)

Compound	SHT, ppb	DHT, ppb
(E)-2-hexanal	5	340
1-Penten-3-one	61	100
2-Isobutylthiazole	2	450
2-Methylfuran	97	1,060
2-Pentylfuran	26	700
3-Methybutanal	17	750
3-Methylfuran	717	3,200
6-Methyl-5-hepten-2-one	21	1,330
Acetone	325	-
Benzaldehyde	3	30
Dimethyl disulfide	16	630
Dimethyl sulfide	5,205	2,974
Ethanol	311	-
Geranial	2	130
Hexanal	188	6,210
Pentanal	48	470

In the present study, this DHT-type protocol was used to characterize fresh tomato flavour compounds for comparison with related literature methods.

9.7.4 Some Practical Examples of Headspace Technique Use

9.7.4.1 Tomato Juice

Fresh tomato juice was made from vine-ripe fruit by Campbell Soup Company's R&D centre in Davis, USA. Chemicals were reagent grade, supplied from reliable sources.

9.7.4.1.1 Preparation of Traps

Traps were prepared using silane-treated glass tubing (79 mm x 6 mm) packed with 13 mg 60/80 mesh Tenax-TA (2,6-diphenyl-*p*-phenylene oxide) polymers held in place by silanized glass wool. The traps were initially conditioned at 330° C for 2 h under nitrogen gas at a 20 ml/min flow rate. The traps were regenerated at 250° C for 1 h immediately before each purge-and-trap experiment.

9.7.4.1.2 Thermal Desorption

Adsorbed volatile compounds were recovered from the trap directly inside the GC injector. The desorption time and temperature were previously determined. The injector temperature was 200° C and a loop of the analytical column at the injector end was immersed in a liquid nitrogen-filled Dewar flask to cryogenically trap the desorbed volatiles. Subsequently, the injector glass liner (insert) was replaced with the trap to desorb volatiles. Thermal desorption was carried out for 5 min with the split vent and septum purge closed.

9.7.4.2 Headspace of *Hedychium coronium*

Hedychium coronium flower has a delicate, pleasant fragrance, but the essential oil and concrete extracted by traditional methods usually lose this fragrance. Thus, the headspace of the *H. coronium* flower was analyzed. The essential oil of *H. coronium* flowers, which was absorbed by XAD-4 resin, eluted by organic solvent and concentrated, had a fragrance similar to the natural fragrance of *H. coronarium* flowers.

9.7.4.3 Volatiles of White Hyacinths Isolated by Dynamic Headspace Trapping

More than 70 constituents of white hyacinths can be identified by GC and GC-MS. The principal constituents are benzyl acetate and (E,E)-α-farnesene. Beside these, sensorily important substances like indole, oct-1-en-3-ol and phenylacetaldehyde were identified. Minor traces of three substituted pyrazines were detected by GC-sniffing. The advantages of the simultaneous closed-loop stripping technique using various adsorbing agents at the same time were demonstrated. By this method, artifact formation and discrimination of individual components can be determined

9.7.4.4 Medical Materials Testing by Headspace Trapping-GC-MS

The new technology provided by the HS-40 Trap coupled with a sensitive detection method such as GC-MS allows volatile organic compounds in medical sutures to be analyzed easily at trace levels. Individual compounds present in the sutures can be analyzed by GC-MS and identified by a NIST library search of the acquired mass spectral data. The innovative, patent-pending, headspace trapping technology used in this application provides sensitivity beyond the capability of traditional static headspace. This presents a new level of detection capability for the evaluation of materials used in medical applications, as well as in other types of material testing, including pharmaceutical formulations and food-packaging film.

9.8 Conclusions

Advanced technologies such as SPME and headspace trapping extraction are well suited for the analysis of flavor and fragrance compounds. The typically small, volatile compounds are easily extracted. The simplicity of the method allows easy coupling to analytical instruments. Loss of volatile compounds during sample preparation steps is minimized or eliminated, compared to conventional methods.

These techniques are useful for semivolatile compounds, even though these were more challenging in the early years. With appropriate matrix modification, one can take advantage of this analytical method, which provides significant convenience for field and air analyses.

Bibliography

Anonymous, 2006, On sampling, conditioning and gas handling. Available at: http://www2.nose-network.org/members/3_Sampling.doc

Anonymous, 2006, Welcome to the SPME. Available at: http://www.Sigmaaldrich.com/ Brands/Supelco_Home/Spotlights/SPME_central.html

Mookherjee, B. D., 1998, A novel technology to study the emission of fragrances from skin, *Perfumer & Flavorist*, 23(11): 1-11

Rouseff, R. L. and Cadwallader, K. R. (eds.), 2001, Headspace Analysis of Foods and Flavors: Theory and Practice, Kleur Publishing: New York, p. 212

Snow, M. and Grecsek, H. 2006, Medical Materials Testing by Headspace Trap-GC/MS. Available at: http://las.perkinelmer.com/Content/ApplicationNotes/FAR_GCMSMedical Materials.pdf

10 Supercritical Fluid Extraction of Medicinal and Aromatic Plants: Fundamentals and Applications

A. Bertucco and G. Franceschin

Abstract

The main issues related to supercritical fluid extraction of medicinal and aromatic plants are discussed in view of the development of this separation technique at industrial scale. After an introduction to supercritical fluid extraction, the roles of thermodynamics and mass transfer properties are emphasized, and the effects of the main operating variables on product recovery are briefly examined. Fundamental concepts about the equipment needed and basic technology are presented, including economical evaluation. Finally, a short literature survey of successful supercritical extraction processes of medicinal and aromatic plants is reported and a future outlook is given.

10.1 Introduction

In the second half of last century, an increasing interest has been paid to supercritical fluids as alternate solvents for the extraction of natural bioactive molecules from plants. The main reason for the interest in supercritical fluid extraction (SFE) was the possibility of carrying out extractions at temperature near to ambient, thus preventing the substance of interest from incurring in thermal denaturation.

A thorough review of the results achieved up to the early 1980s is presented in a book by Stahl *et al.*, published in 1986. Clearly, by that time, the fundamentals of this new extraction process were already understood, even though the technical-economical assessment and the design criteria for large-scale application of SFE were still missing. After twenty years of research and development, it is now possible to say that such achievements are at hand, so SFE is currently a well-established unit operation for extraction and separation. Moreover, its design and operating criteria are fully understood, so that it can profitably be applied in the extraction of medicinal and aromatic plants (MAPs).

10.2 Supercritical Fluids

A fluid at supercritical condition, also referred to as a dense gas, is a fluid above its critical temperature (T_C) and critical pressure (P_C) to a certain extent: to be supercritical, the reduced temperature T_r (i.e. T/T_C) must not exceed 1.2 or 1.3, whereas the reduced pressure P_r (i.e. P/P_C) may be as high as allowed by technological limits.

At suitable conditions, any fluid can reach its supercritical state. However, only those having a critical temperature not far from ambient temperature can be used as alternative solvents for the extraction of MAPs. Carbon dioxide (CO_2), with T_C=31.06° C and P_C=73.81 bar, is the most attractive solvent, because of its proprieties regarding toxicity, flammability and cost.

The possibility of using supercritical fluids (SFs) as extraction solvents is directly linked to their density. In fact, according to an empirical correlation proposed by Chrastil in 1982,

$$s = \rho^{a} \exp\left(\frac{b}{T} + c\right) \qquad (1)$$

where *s* is the solute solubility, ρ is the solvent density and *T* is the absolute temperature; *a*, *b* and *c* are correlation parameters to be adjusted to experimental solubility data in supercritical CO_2.

When a fluid approaches the critical conditions, its density gets closer and closer to that of the liquid state. This can be seen, for CO_2, in Figure 1, where density isotherms are plotted against the reduced pressure. For example, at T = 35° C and P = 200 bar, ρ = 866 kg/m$^{3.}$

It is also clear from Figure 1 that, close to the critical point, both the compressibility and expansion coefficient of the fluid are high, so slight changes in the operating conditions can significantly modify the density, i.e. the supercritical fluid solvent power. The importance of the Chrastil equation (Eq. 1) lies in the fact that solvent density is identified as the key factor in a successful SFE process.

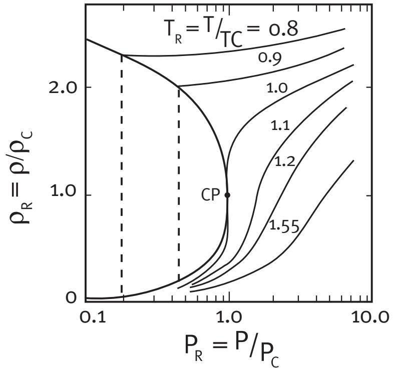

Figure 1: Density vs. pressure diagram for carbon dioxide

When plotted against solvent density, solubility data for supercritical CO_2 always display a regular trend such as that in Figure 2a, whereas a more complex behavior is seen when pressure is improperly used as the independent process variable (Figure 2b).

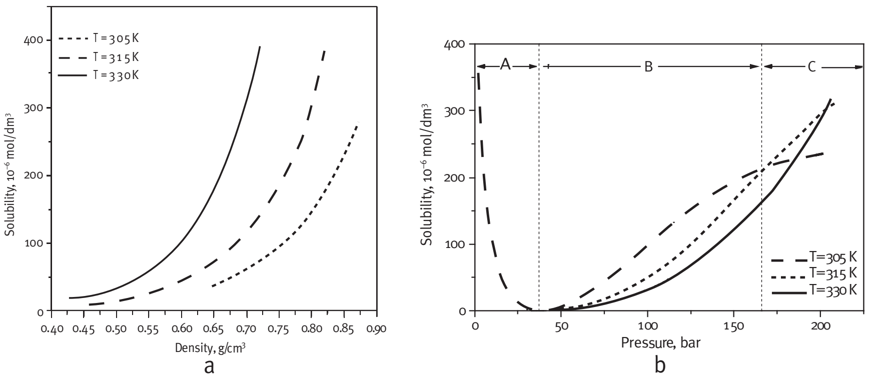

a b

Figure 2: Solubility of 1,4-bis-(n-propylamino)-9,10-anthraquinone in supercritical CO_2: as a function of solvent density (2a) and of pressure (2b)

Coming back to Eq. 1, is important to point out that it is not theoretically correct and must be applied within restricted temperature ranges only.

More importantly, the exponential form of Eq. 1 does not guarantee that the solubility of a solute in SF is high. The solubility depends on parameters *a*, *b* and *c*; in fact, in the case of CO_2, the solubility of a solute of interest for MAP applications is at best in the range of 1 to 1000 by weight or often 1 to 10,000 (Figure 2a illustrates this). This is because CO_2 is a poor solvent, even at supercritical conditions. In addition, this holds for non-polar substances only, as supercritical CO_2 does not dissolve polar molecules at all. Actually, CO_2 is a good solvent only for low molecular weight solutes.

The limit on solubility is dictated by thermodynamics. According to the iso-fugacity criterion applied to the substance to be extracted, between the two phases at equilibrium (the condensed one – either solid or liquid – and the supercritical one), we have:

$$y_i = \frac{P^{sat}}{P} E_i \tag{2}$$

$$E_i = \frac{\varphi_i^{o,v}}{\varphi_i^{v}} \exp\left(v^{S/L} \frac{P - P^{sat}}{RT}\right) \tag{3}$$

where y_i is the mole fraction of *i* in the supercritical phase, $\varphi_i^{o,v}$ and φ_i^{v} are the fugacity coefficients of *i* in the standard state and in the mixture, respectively, at the process conditions, P^{sat} is the solute saturation (or sublimation) partial pressure (i.e. the component volatility), and $v^{S/L}$ is the molar volume of the condensed phase (either solid or liquid). *T* is the absolute temperature and *R* is the universal gas constant. E_i is the so-called enhancement factor, which accounts for the increasing solubility due to system nonidealities with respect of the ideal behavior (given by Eq. 2).

According to Eq. 2, the solubility of *i* in the SF can be calculated at the process condition, provided that the fugacity coefficients φ_i can be evaluated accurately by means of an equation of state. However, the substance vapor pressure directly influences the solubility when P^{sat} is as usual for MAPs, very low. Only an equally low value of the fugacity coefficient, i.e. a high system nonideality, can partially counteract the lack of volatility of the pure component.

Regardless of the way its value has been obtained, i.e. from Eq. 1 or Eqs. 2-3, the solubility is only one of two fundamental pieces of information that must be known in order to assess the feasibility of an SFE process for MAPs. The second one is selectivity, which is defined as the ratio of the solubility of the substance *i* of interest with respect to a reference substance *j*:

$$\alpha_{ij} = \frac{s_i}{s_j} \tag{4}$$

If on one hand high solubility is desirable, to reduce the solvent consumption per unit product extracted, on the other hand selectivity must be as far as possible from 1, to ensure that the substance of interest is extracted as pure rather than in mixture with other components. In summary, to develop a successful SFE process for MAPs, both solubility and selectivity issues must be fulfilled properly.

Coming back to CO_2, it must be kept in mind that this solvent is rather non-selective: when it is able to dissolve a group of similar substances (for example, in terms of carbon atoms), all of them are extracted to a similar extent, provided they have similar polarities.

Therefore, it can be stated that CO_2 alone is not as selective as a good and pure solvent. It is also noteworthy that CO_2 capacity and selectivity may be improved by using an organic solvent as the entrainer, also called the co-solvent, with the function of modifying chemical interactions between CO_2 and the substance to be dissolved in it.

But by doing this, the SFE process becomes more complicated, as an extra chemical component needs to be introduced into the process. However, the co-solvent can be easily separated from the product downstream, due to the high selectivity displayed by supercritical CO_2 in this respect.

10.3 SFE Processes

An SFE process for extracting MAPs is basically composed of two main sections (Figure 3a). The feed, containing the substance of interest, indicated by A, comes in contact with supercritical CO_2, at suitable temperature and pressure, in an extraction device. In this simple scheme, component A is selectively extracted and must be recovered from the supercritical solution, which is usually a dilute one for the reason explained in the previous section. Product recovery occurs in the separation section, whose temperature and pressure can be adjusted in order to optimize the amount of A produced.

Note that, due to the low solubility in supercritical CO_2, after recovery of the product of interest the solvent must be recycled and pumped back to the extractor, in order to minimize operating costs. It is also noteworthy that the separator can be operated either at the same temperature or at the same pressure of the extractor, the best condition resulting from an economical analysis of the overall production costs.

If the temperature is kept constant, product separation is achieved by depressurization (Figure 3a), and mechanical energy has to be provided to the system to raise the CO_2 pressure from the separator to the extractor conditions. On the other hand, extracted products can be separated from CO_2 by increasing the temperature, and thermal energy must be supplied in this case (Figure 3b), where the circulation of the solvent can be done at nearly isobaric conditions. Of course, the way the separation of products from CO_2 is achieved can be more complex: for instance, both temperature and pressure can be varied when passing from the extractor to the separator sectors, or a solid can be used to promote separation by adsorption.

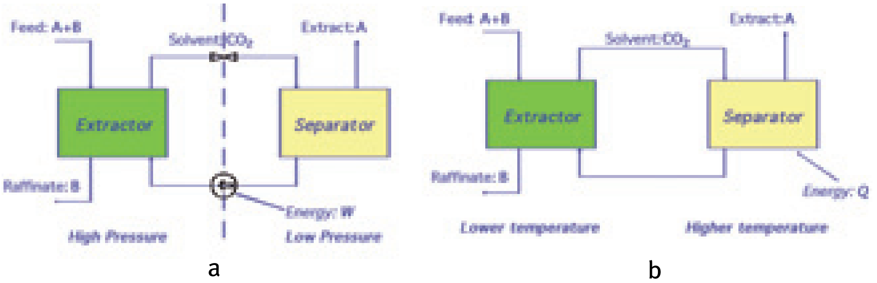

Figure 3: Block flow diagrams of simple SFE processes: with separation obtained by pressure change (a) and by temperature change (b)

If, as it often occurs, many substances are extracted by CO_2 at the extraction conditions because of lower CO_2 selectivity, their fractionation can also be achieved in the separation section, by simply using more than one separator, operated at different conditions. As shown in Figure 4, multiple fractions with different properties can be recovered from the same extraction.

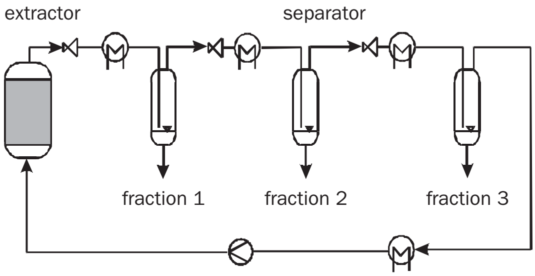

Figure 4: Single extractor, multiple separator scheme

Finally, a multiple extractor scheme can also be envisaged, as represented in Figure 5 with only one separation step, for sake of clarity. This configuration is particularly useful when, as in the case of SFE of MAPs, the substances to be extracted are embedded in a solid matrix, which is initially loaded in the extraction vessel as a fixed bed. In this case, the extractors can be connected either in parallel or in series, depending on specific requirements.

More details on the development of SFE processes are provided in a book by Brunner listed in the bibliography. SFE of solids is a semibatch operation, which can also be operated in a simulated moving bed configuration to obtain a continuous production.

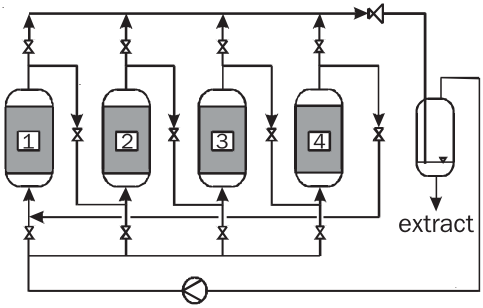

Figure 5: Multiple extractor, single separator scheme

Typical extraction profiles from solid materials (single vessel) are shown in Figure 6 where the extractor yield, i.e. the amount of substance of interest extracted with respect to the total amount initially contained in the solid, is plotted against extraction time. The profiles, which are steeper if the temperature is higher, have two parts: a straight line corresponding to the extraction of the substance “readily available” to supercritical CO_2, and an asymptotic curve representing the extraction of the part attached to the solid matrix. In the first case, the extraction is limited by solubility; in the second, mass transport (diffusion) properties are important and can be limiting and crucial for the success of SFE.

The effects of operating temperature are also clear in Figure 6. Other important operating variables are pressure, CO_2 flow rate and humidity of the material to be extracted.

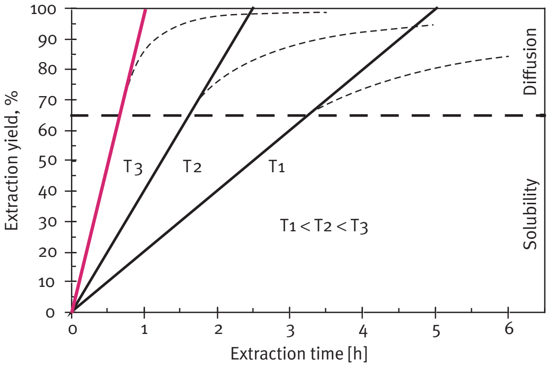

Figure 6: Extraction yield versus time at different temperatures

10.4 The SFE Process and Equipment Development

In order to design and develop an SFE process for MAPs with CO_2 (possibly assisted by ethanol or water as entrainers), we need to know and optimize:

1. The solubility of the substance of interest
2. The selectivity of this substance with respect to others that are extracted simultaneously
3. The extraction profiles (such as those in Figure 6)
4. The way to separate the substance of interest from the total extract

All this information can be obtained by simple measurements performed in a laboratory-scale apparatus of minimum volume such as that illustrated in Figure 7.

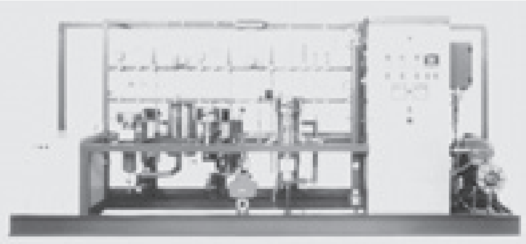

Figure 7: Laboratory-scale apparatus for SFE process design

To measure extraction profiles, a small pilot-scale apparatus can be used. Extractor and separator volumes do not need to exceed 1 liter each. The analytical system must be suitable to measure the concentration and purity of the products of interest. Basic requirements in terms of equipment are:

1. A liquid CO_2 storage tank
2. A pump for liquid CO_2
3. A cooler to prevent CO_2 from evaporating in the pump
4. A heat exchanger to control the temperature of CO_2 entering the extractor
5. An extraction vessel
6. A heat exchanger to control the CO_2 plus solute mixture entering the separator
7. A separation vessel

Note that condensing and recycling of CO_2 after separation is not needed at the laboratory-scale developmental level, whereas these are essential requirements at the industrial production level.

All parts of the SFE laboratory-scale plant must be designed in order to resist the maximum operating pressure. If this does not exceed 300-350 bar, the entire equipment (e.g. vessels, valves, fittings) is pretty much of standard type and relatively inexpensive. If, as usual, stainless steel is used, the thickness of any part of the plant can be easily calculated by applying the Von-Mises equation:

$$P_i \frac{k^2}{k^2 - 1} \sqrt{3} < \sigma_{am} \tag{5}$$

with:

$$\sigma_{am} = \frac{\sigma_s}{S_f} \tag{6}$$

where P_i is the internal pressure, k is the external to internal diameter ratio, σ_s is the yield stress, and S_f is a suitable safety factor (usually S_f=1.5).

From Eq. 5, it can be seen that the thickness of a cylindrical vessel depends on its diameter. Examples are given in Table 1 for a vessel of 0.2 m internal diameter, with both stainless steel and carbon steel construction materials.

Table 1: Thickness (s) of a thick-wall cylindrical vessel of 200 mm internal diameter as a function of pressure (SS=stainless steel, CS=carbon steel)

P [atm]		**50**	**100**	**150**	**200**	**250**	**300**
s [mm]	SS	3.8	8.1	12.9	18.5	25.0	32.7
	CS	2.2	4.6	7.2	10.0	13.0	16.2

Special care must be paid to closures and seals. SFE of MAPs is mostly an extraction operation from solid materials, which is carried out in batch or semibatch mode. Therefore, extraction vessels need to be pressurized, depressurized, opened, filled, and closed again several times per day. In order to ensure fast and safe operation procedures and reliable seals, gaskets like O-rings are useful and closure devices have been specifically designed. Again, the technology needed is already fully developed. We refer to chapter 4 of the book by Bertucco and Vetter for details. The book also describes the machinery for moving fluids under pressure, i.e. pumps and compressors. We conclude that setting up a laboratory-scale apparatus with which to perform feasibility studies concerning the possibility of applying SFE to MAPs is not really an issue, and can be done with a relatively small capital cost.

However, this does not mean that SFE of MAPs is in itself an economically convenient operation. An accurate evaluation of production costs, including both capital and utility costs, must be performed before scaling up a process whose technical feasibility has been demonstrated at the laboratory level. Costs are also discussed in the book by Bertucco and Vetter (chapter 8), but are only indicative. The reader should remember that capital costs have been steadily decreasing in the last years must be taken into account.

10.5 SFE Applied to Medicinal and Aromatic Plants

A large number of MAPs has been considered for possible extraction by supercritical CO_2. The most recent developments suitable to have industrial relevance are listed in Table 2. These examples illustrate the great potential of SFE in this field.

Table 2: Medicinal and aromatic plants extracted by SFE

Plant name (part used)	**Product(s) extracted**	**Reference**
Calendula officinalis (flowers)	Oleoresin	Campos *et al.,* 2005, Experimental data and modeling the supercritical fluid extraction of marigold (*Calendula officinalis*) oleoresin, *J Supercritical Fluids*, 34: 163-170 Danielski *et al.,* 2007, Marigold (*Calendula officinalis* L.) oleoresin: solubility in SC-CO_2 and composition profile, *Chem Eng Proc*, 46: 99–106

Plant name (part used)	**Product(s) extracted**	**Reference**
Echinacea purpurea (whole herb)	Alkamides, polyphenolics including chichoric acid, carbohydrates	Catchpole *et al.,* 2002*,* Supercritical extraction of herbs I: saw palmetto, St John's wort, kava root, and Echinacea, *J Supercritical Fluids*, 22: 129-138
Eucalyptus spp. (leaves)	Essential oil	Della Porta, *et al.,* 1999, Isolation of eucalyptus oil by supercritical fluid extraction, *Flavour Fragr J,* 14: 214–218
Ginkgo biloba (leaves)	Flavonol glycosides (flavonoids) and terpenoids	Chun Yang *et al.,* 2002, Extraction of pharmaceutical components from *Ginkgo biloba* leaves using supercritical carbon dioxide, *J Agric Food Chem,* 50: 846-849
Hypericum perforatum (herb)	Naphthodianthones, hypericin and pseudohypericin	Catchpole *et al.,* 2002
Levisticum officinale (dry rhyzomes, roots)	Essential oil	Daukšas *et al.,* 1999, Supercritical CO_2 extraction of the main constituents of lovage (*Levisticum officinale* Koch.) essential oil in model systems and overground botanical parts of the plant, *J Supercritical Fluids,* 15: 51–62
Matricaria chamomilla (flowers)	Oleoresin	Kotnik *et al.,* 2007, Supercritical fluid extraction of chamomile flower heads: comparison with conventional extraction, kinetics and scale-up, *J Supercritical Fluids*, Available online 13 February 2007 (in print)
Mentha spp. (leaves)	Essential oil	Marongiu *et al.,* 2001, Extraction and isolation of *Salvia desoleana* and *Mentha spicata* subsp. insularis essential oils by supercritical CO_2, *Flavour Fragr J,* 16: 384–388
Origanum spp. (herb)	Essential oil	Leeke *et al.,* 2002, *Eng Chem Res,* 41: 2033–2039
Piper methysticum (roots, rhizomes)	Kava lactones	Catchpole *et al.,* 2002
Piper nigrum (fruit)	Oleoresin	Ferreira *et al.,* 1999, Supercritical fluid extraction of black pepper (*Piper nigrum* L.) essential oil, *J Supercritical Fluids*, 14: 235–245.

Plant name (part used)	**Product(s) extracted**	**Reference**
Saccharum spp. (crude wax)	Long chain n-alcohols	De Lucas *et al*., 2005, Supercritical extraction of long chain n-alcohols from sugar cane crude wax, *J Supercritical Fluids*, 34: 163-170
Salvia desoleana (leaves)	Essential oil	Marongiu *et al.,* 2001
Serenoa repens (fruit)	Free fatty acids, phytosterols (low concentrations), fatty alcohols and triglycerides	Catchpole *et al.,* 2002
Solanum lycopersicum (fruit)	Carotenoids, tocopherols and sitosterols	Vagi *et al*., 2007, Supercritical carbon dioxide extraction of carotenoids, tocopherols and sitosterols from industrial tomato by-products, *J Supercritical Fluids*, 40: 218–226
Taxus brevifolia (bark)	Taxol	Jennings *et al.,* 1992, Supercritical extraction of taxol from the bark of Taxus, *J Supercritical Fluids*, 5: 1-6
*Taxus cuspidate (*needles)	Paclitaxel and baccatin III	Moon-Kyoon Chun *et al.,* 1996, Supercritical fluid extraction of paclitaxel and baccatin III from needles of *Taxus cuspidate*, *J Supercritical Fluids*, 9: 192-198
Vitis vinifera (seeds)	Procyanidins	Cao *et al.,* 2003, Supercritical fluid extraction of grape seed oil and subsequent separation of free fatty acids by high-speed counter-current chromatography, *J Chromatogr* A 1021: 117–124
Zingiber officinale (rhizome)	Oleoresin	Badalyan *et al.,* 1998, Extraction of Australian ginger root with carbon dioxide and ethanol entrainer, *J Supercritical Fluids*, 13: 319-324

10.6 Conclusions

SFE with CO_2 is a technically and possibly economically valid technique to extract bioactive components from MAPs. Organic solvent-free products can be obtained and the low operating temperature makes it possible to preserve all their natural properties. The feasibility study on specific products can be performed rather easily at laboratory scale. However, accurate evaluation of production costs, including both capital and operating ones, must be done in order to exploit SFE at the industrial level.

Bibliography

Bertucco, A. and Vetter, G. (Eds.), 2001, High Pressure Process Technology: Fundamentals and Applications, Elsevier Science, Amsterdam

Brunner, G., 1994, Gas Extraction: An Introduction to Fundamentals of Supercritical Fluid and Application to Separation Processes, Steinkopff Darmstadt Springer, New York

Chrastil, J., 1982, Solubility of solids and liquids in supercritical gases, *Journal of Physical Chemistry,* 86: 3016-3021

Shi, J. (Ed.), 2006, Functional Food Ingredients and Nutraceuticals, Processing Technologies, Taylor and Francis, Boca Raton, USA

Stahl, E. K., Quirin, W. and Gerard, D., 1986. Dense Gases for Extraction and Refining, Springer Verlag, Berlin

11 Process-scale High Performance Liquid Chromatography for Medicinal and Aromatic Plants

M. M. Gupta and K. Shanker

Abstract

High performance liquid chromatography (HPLC) is widely used by chromatographers and by the pharmaceutical industry for the accurate and precise analysis of chemicals and drugs of diverse nature. The systematic scale-up from analytical to preparative and process scale and further scale-up to industrial scale can be used in the medicinal and aromatic plant industry for the isolation and purification of phytomolecules of therapeutic and commercial interest. Due to the gradual increase in the demand for phytomolecules, the importance of process-scale HPLC as a purification tool has been increasing. In this article, we discuss the practical aspects of process-scale HPLC and focus on terminology, operational problems, advantages and applications of this technology to medicinal and aromatic plants.

11.1 Introduction

The term liquid chromatography (LC) refers to a range of chromatographic systems, indicating liquid-solid, liquid-liquid, ion-exchange and size exclusion chromatography. Glass column chromatography is an example of classic liquid column chromatography in which the mobile phase percolates under gravity through a glass column filled with a finely divided stationary phase. Liquid chromatography has overtaken gas chromatography, as high performance liquid chromatography (HPLC) systems now provide features such as:

i) High resolving power
ii) Fast separation
iii) Continuous monitoring of column effluent
iv) Qualitative and quantitative measurements and isolation
v) Automation of analytical procedures and data handling

There has been tremendous growth in this technique since 1964 when the first HPLC instrument was constructed by Csaba Horvath at Yale University. For the isolation of compounds, preparative mode HPLC (prep-HPLC) can be used in pharmaceutical development for trouble-shooting purposes or as part of a systematic scale-up process. The importance of prep-HPLC in pharmaceutical production as a purification tool has been increasing. Chromatographic separation can remove impurities of different polarity and can reduce the content of an enantiomer in a racemic mixture. In both of these instances, crystallization may be used to prepare the pure product. Bench to pilot scale production of natural products needs some form of automation: thus, developing well-automated preparative chromatographic methods is a necessary but demanding task.

Innovations in micro-analytical to preparative HPLC played an important role in the progress of natural product chemistry. HPLC is used routinely in phytochemistry to pilot the preparative-scale isolation of natural products and to control the final purity of the isolated compounds. The development of hyphenated techniques related to this efficient separation technique in the past 20 years has provided powerful new tools such as LC/UV-photodiode array detection, LC/mass spectrometry (LC/MS) and LC/NMR. The combination of high separation efficiency of HPLC with these different detectors has made possible the acquisition of data on an LC peak of interest within a complex mixture.

11.2 Theoretical Aspects of HPLC

Separation of chemical compounds is carried out by passing the mobile phase, containing the mixture of the components, through the stationary phase, which consists of a column packed with solid particles. The cause for retention is physical and chemical forces acting between the solute and the two phases, on the chromatographic column. The reason for retention is the difference in the magnitude of forces; this results in the resolution and hence separation of the individual solutes. The separation of compounds occurs by distribution of solutes between the two phases.

11.2.1 Chromatography Classification

Chromatography can be classified according to mechanism of separation as: adsorption chromatography, partition chromatography, ion exchange chromatography, size exclusion chromatography and affinity chromatography. In HPLC, separation is mainly governed by adsorption and partition chromatography. In adsorption chromatography, separation is based on the difference between the adsorption affinities of the sample components on the surface of an active site, whereas in partition chromatography separation is mainly based on the difference between the solubility of sample components in the stationary phase and the mobile phase.

There are two modes of analysis depending on the operation techniques viz. isocratic and gradient. Isocratic analysis is the procedure in which the composition of the mobile phase remains constant during the elution process. In gradient elution, the composition of the mobile phase changes continuously or stepwise during the elution process. HPLC can also be classified according to special techniques, such as reverse phase (RP) and normal phase chromatography. Reverse phase is an elution procedure used in liquid chromatography where the mobile phase is significantly more polar than the stationary phase. On the other hand, in the normal phase procedure, the stationary phase is more polar than the mobile phase. Lipophilic substances like oils, fats and lipids are separated by normal phase chromatography. Commonly used mobile solvents are n-hexane, heptane, chloroform, and alcohols. Most biomedical substances are separated by reverse phase chromatography using aqueous mixture with methanol, acetonitrile and additives (buffers, ion-pairs).

11.2.2 Important Factors that Influence HPLC Separation

HPLC separation is influenced by dead volume, capacity factor, theoretical plate count and selectivity:

- ***Dead volume*** (V_0) is the volume at which an un-retained component elutes.
- ***Capacity factor*** (K') is a measurement of the retention time of a sample molecule, relative to column dead volume. It changes with variations in mobile phase composition, column surface chemistry or operating temperature. Capacity factor is calculated as follows:

  $$K' = \frac{V_1 - V_0}{V_0} \quad E_i$$

  V_1 = Retention volume of peak 1
- ***Theoretical plate count*** (N) is a measure of column efficiency in terms of band-spreading of a peak. The smaller the band-spread, the higher the number of theoretical plates, which indicates good column and system performances.
- ***Resolution*** (R_s) is the distance between the peak centres of two component peaks divided by the average base of the peaks, as follows:

  $$R_S = \frac{V_2 - V_1}{\sqrt{W_1 + W_2}}$$

  W_1 = width of peak 1
  W_2 = width of peak 2
- ***Selectivity*** (α) is the relative retention of two peaks in a chromatogram.

  $$\alpha = \frac{K'_2}{K'_1} = \frac{V_2 - V_0}{V_1 - V_0}$$

  K_1 and K_2 = capacity factors for retention volume of peak 1 and peak 2 respectively.

Capacity factor (K'), selectivity (α) and column efficiency (N) are three fundamental parameters that influence the resolution of a chromatographic separation, as follows:

$$R_s = \frac{1}{4} \left(\frac{\alpha - 1}{\alpha} \right) \sqrt{N} \left(\frac{K'}{1 + K'} \right)$$

11.2.3 Main Components of HPLC

An HPLC system contains the following components:

a) *Reservoir.* This is meant for the mobile solvents. Acetonitrile, methanol, heptane, isopropanol and cyclohexane are the organic modifiers most commonly used. Trifluoroacetic acid, heptafluorobutyric acid, phosphoric acid and triethylamine phosphate are ion-pairing reagents for better chromatographic results. All tubing and fittings should be chemically inert. Solvent must be filtered through a 0.45-µm filter unit.

b) *Degasser.* In analytical operations, the mobile phase should be free of air bubbles. For this purpose, a degasser is used.

c) *Pumps.* These are devices that deliver the mobile solvent at a controlled flow rate to the separation system. HPLC uses reciprocating pumps: a pump with a single or multiple chambers, from which the mobile phase is displaced by reciprocating pistons or diaphragms. Binary gradients are created by the selected mixing of two solvents, on a single-headed two-pump system. Accurate gradient is maintained by microprocessor control.

d) *Injector/autosampler.* This device introduces a liquid sample into the mobile phase or onto the chromatographic bed. An autosampler can perform repeated functions without operator attendance, and thus is a labor-saving device.

e) *Column.* Silica and modified silica columns are available for various applications. Examples are octyl (C_8), octadecyl (C_{18}), phenyl (C_6H_5), and cyno (CN) columns.

f) *Guard column.* This is used to protect the main column.

g) *Detectors.* No universal detector is available for all molecules. However, according to the characteristic of the molecules investigated, various detectors are used (Table 1).

h) *Fraction collector.* This device collects the fractions containing the molecules of interest during the chromatographic run.

i) *Records.* A computer is used for chromatographic data acquisition.

Table 1: Characteristics of various HPLC detectors

Detector	**Application**	**Advantages and limitations**
Electrochemical	Responds to substances that are oxidizable or reducible	Commercially available. Non-specific. High LOD

Detector	**Application**	**Advantages and limitations**
Fluorescence	Detects trace-level analytes such as aflatoxins, carbamates and polycyclic aromatic hydrocarbons	Very specific. Low LOD. Not everything fluoresces
Infrared	Works for all molecules	Many solvents are infrared-active
Mass spectrometry	Analyte identification	Ability to ionize analyte. Low LOD
Photodiode array	Works for wavelengths 190-800 nm	High LOD
Refractive index	Works for nearly all molecules	Temperature sensitive. High LOD
Scattering	Uniform response	Non-specific. LOD, 5 ng per 25 mL. Interference from solvent
Ultraviolet and visible	Works for molecules with chromophores and for complex samples	Non-specific. All molecules that absorb UV and visible light can be detected.

LOD, level of detection

11.2.4 HPLC Classification

HPLC may be characterized depending on column diameter, which is the governing factor for flow rate from microscale to industrial scale chromatography. Column internal diameter (i.d.) defines the sample load and flow rates (Figure 1).

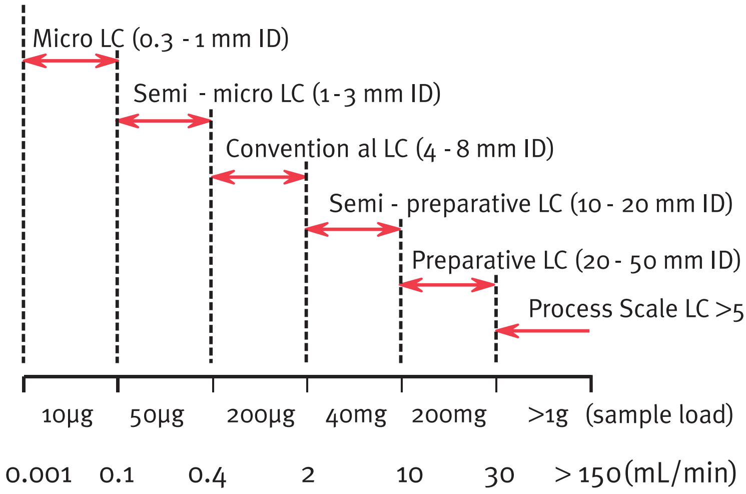
Micro LC (0.3 - 1 mm ID)
Semi - micro LC (1 - 3 mm ID)
Convention al LC (4 - 8 mm ID)
Semi - preparative LC (10 - 20 mm ID)
Preparative LC (20 - 50 mm ID)
Process Scale LC >50 mm ID)
10μg
50μg
200μg
40mg
200mg
>1g
(sample load)
0.001
0.1
0.4
2
10
30
> 150(mL/min)

Figure 1: Classification of HPLC according to column diameter

11.2.5 Advantages of HPLC

The use of HPLC in the isolation and purification of complex compounds is increasing tremendously due to its flexibility and efficiency. It has several advantages over traditional methods of isolation and purification:

i) Variety of separating techniques.
ii) Variety of column packings for different techniques.
iii) Separation optimized by alteration of the mobile phase.
iv) Mobile phase easily manipulated in gradient systems.
v) RP technique separates very similar and very different compounds simultaneously.
vi) HPLC can be used as a preparative method.
vii) HPCL can be used as a purification technique. More than one detector can be connected in series (e.g. UV and evaporative light scattering detector).
viii) Most sample analysis is carried out at room temperature.
ix) Short analysis runs. More than 70% of HPLC separations are performed on UV detectors and 15% rely on fluorescence without any derivatization.

11.3 Preparative HPLC

Preparative chromatography is the most powerful and versatile method for isolation as well as purification of complex compounds used in drug development studies. Prior to performing preparative HPLC, the following points must be taken into account to optimize the separation and maximize the sample load on a small column:

- Prior to pilot-plant scale, a systematic method for development is required
- Validated robust analytical methods are required
- Scale-up of parameters from analytical method to prep-HPLC
- In prep-HPLC, buffer is not used
- Stationary phase with large particle sizes to decrease costs for prep-HPLC

11.3.1 Strategy for Preparative Separation

Selection of the appropriate mode of chromatography is followed by the optimization of the separation, i.e. stationary phase, mobile phase, temperature, additives. The next step is optimization of the throughput, i.e. sample amount and column overloading. In the final step, stepwise scale-up of separation is performed to obtain the desired compound.

It is always important to optimize the small-scale separation (which will significantly impact on throughput), the size of packing material and the column needed to obtain the desired throughput, the mobile solvent and the instrument capability. Normal-phase methods are the first choice because: direct transfer from normal-phase thin-layer LC or HPLC to prep-LC is possible; costs of RP packing materials are still high; cleaning normal-phase silica is easier because the material is more robust; removing organic solvents typically used in normal-phase chromatography from the final product solution is easier than removing water from an RP chromatographic fraction and can be achieved at lower temperatures and provides higher product quality and lower energy costs. Particle size of the stationary phase material also plays an important role in the isolation of the desired compounds. The choice of 5-μm particle size in a preparative column is not practical because it not only increases the column pressure but also is extremely expensive. Moreover, when the sample amount is increased, resolution performances of 5-μm and 15-μm particles are not different.

11.4 Practical Consideration in Preparative HPLC Scale-up

11.4.1 Sample Loading

If the tests on analytical columns with analytical loadings show good separations, a scale-up to a larger column diameter can be performed on prep-HPLC. Instead of jumping directly to the largest column diameter, stepwise scale-up should be done. The first step in the scale-up process is the transfer of the analytical separation procedure to a 5 cm i.d. preparative LC column. Optimization in a preparative column is required. The sample injected onto the column usually starts at 1 g and increases to as much as 20 g, depending upon the quality (resolution) of the separation achieved, the quality of the initial material, and the specifications for the pure product. Start with the 1-g injection, collect fractions and re-analyze them for purity using the analytical method, because with an increase in sample loading there is a decrease in resolution.

$$\text{(A) Scale up factor for column size} = \frac{(\text{Diameter}_{\text{prep}})^2 \text{ x Length}_{\text{prep}}}{(\text{Diameter}_{\text{anal}})^2 \text{ x Length}_{\text{anal}}}$$

$$\text{(B) Flow rate (prep)} = \text{Flow rate (analytical) x } \frac{(\text{Diameter}_{\text{prep}})^2}{(\text{Diameter}_{\text{anal}})^2}$$

$$\text{(C) Gradient duration (prep)} = \text{Gradient duration (anal) x } = \frac{\text{Length}_{\text{anal}}}{\text{Length}_{\text{prep}}}$$

11.4.2 Separation Time

In preparative separations, the stationary phase is usually recovered and used again to purify the next batch of the same substance. Often the major operating cost in preparative LC is the solvent rather than the packing material. Therefore, the choice of solvent is important in method development and scale-up. As per the need for separation, isocratic mixture or gradient elution of water with organic solvent (methanol or acetonitrile) is used. Gradient elution has a shorter run time than isocratic elution, but sometimes purity of the isolates is compromised.

11.4.3 Solvent Composition

Methanol is often used in preparative separations. It is an inexpensive and strongly polar solvent commonly used in combination with water as a mobile phase in RP separations. Methanol can be used for flushing normal-phase silica columns to remove adsorbed polar contaminants. It can also be recovered easily from many mixtures. In RP applications, acetonitrile yields better peaks but is too expensive in most situations for process-scale separations.

An initial goal of the scale-up process is to find an acceptable separation. If analysts find more than one set of valid conditions, then the cost of solvents becomes a major criterion. Solvent selection is usually determined during the initial method development with the 4-mm i.d. analytical-scale columns. Sometimes, when the overall costs of the goods is important to a final product, one can perform a systematic solvent selection even at later stages of development.

11.4.4 Washing Steps

The accumulation of impurities on the column can decrease the resolution of the subsequent separation, and late-eluted impurities can spoil the collected fractions of the subsequent separation. Therefore, washing steps are often implemented between chromatographic runs. Solvent gradients and recycling steps are sometimes necessary to increase the resolution for difficult separation problems. Sometimes temperature programming is used to remove strongly held impurities.

11.4.5 Recycling

Sometimes gradients and recycling steps are required for better preparative separation of complex mixture of compounds.

11.5 Stepwise Operations in Process-scale HPLC

Various aspects should be taken into consideration for operating process-scale HPLC and stepwise scale-up, as follows.

- Develop a robust analytical method and scale up the method for process scale using the same stationary phase.
- The main objective of method development is a simple, well-automated and robust separation process able to run 24 h per day.
- Optimize sample loading, flow rate and column pressure.
- Select the best solvent for both sample preparation and elution.
- The injection solvent should be optimized because sharp peaks and high loadability are important goals.
- In process-scale separations, control the first two runs manually and observe the process. If no technical problems occur, subsequent runs can be performed automatically.
- Collect fractions and re-analyze them for purity using the analytical method.
- Peak purity at three points (i.e. up slope, apex and down slope) should be confirmed.

11.6 Problems Encountered in Preparative Scale-up

11.6.1 Purity of Crude Extract

A typical problem encountered in process scale-up is that the plant material or enriched fraction used during method development had been produced in analytical scale and differs in solubility and impurity from the material that is being processed in pilot scale. The process-scale plant material can be either of a different quality or show larger amounts of the same impurities, and, in some instances, even new impurities can arise. If an impurity profile shows larger amounts of the same impurities or new impurities, the chromatographer must retest the separation method at analytical scale before starting the process-scale separation.

If during a process scale up, a compound shows higher purity, the solubility in the weak solvent chosen during optimization may not be good enough. In this instance, productivity can be lower than expected because the amount separated in each run will be less. Because scale-up is linear, the chromatographic run takes the same time in preparative scale as in analytical scale, and the substance is not stressed longer in the separation equipment.

11.6.2 Removal of Chromatographic Solvent

The final work-up after the separation step is removal of the chromatographic solvent. The desired fraction collected is a solution which contains the substance of interest in the range of a small percent by weight and, therefore, large amounts of solvent must be removed. The evaporation of solvents, especially water, takes time, so the purified drug substance can be changed or even destroyed during the concentration process. This step should be performed with care considering the thermal stability of the compound of interest.

11.6.3 Temperature Variation from Laboratory to Pilot Scale

It is useful to test the temperature stability of the substance during analytical method development. Temperature also influences the separation performance. For example, the mixing of organic solvents before they enter the column can result in a strong increase or decrease in solvent temperature and can influence the operating temperatures of the mixing unit and column. Temperature effects at the centre of the column caused by heat dissipation can also influence the separation and ultimately the purity of isolates.

11.6.4 Increase in Pump Pressure Due to Accumulation of Impurities on the Column

Another problem that often occurs during the first separation in process scale is that some impurities accumulate on the column during a series of sequential runs. The quality of the separation deteriorates during the sequence. Because the fraction collection is commonly controlled by peak height, a UV detector does not detect this problem and the purity of the fractions decreases. When impurities accumulate on the column, the peak shapes or the retention times of the components of interest might change, so the chromatographer can see quality problems. Unfortunately, sometimes peak shapes and retention times show no changes. An additional indicator that impurities have accumulated on the column is a pressure increase; therefore, it is helpful to monitor column pressure. The increase in pressure is commonly related to instrument failure.

11.7 Summary: Scale-up Strategy

- Define the problem
- Find the chromatographic mode
- Develop and optimize the separation
- Maximize throughput
- Increase sample mass and volume to the maximum while meeting purity objectives
- Determine recovery

- Scale up to desired column size to meet throughput and load objectives
- Pool fractions of comparable purity and rerun if necessary
- Check fraction purity using an analytical column

11.8 Applications: Natural Products Isolation

A few examples of prep-HPLC for the isolation of natural products are summarized herein. First, tannins from *Guiera seregalensis* can be isolated using the following conditions:

Column	RP-18e (250 x 10 mm i.d.)
Mobile phase	Water:methanol:THF (90:10:0.25) and (80:20:0.25)
Flow rate	2.5 ml/min
Detection	280 nm
Compounds isolated	Galloylquinic acids, namely, 3-*O*-, 5-*O*-, 1,3-di-*O*-, 3,4-di-*O*-, 3,5-di-*O*-, 4,5-di-*O*-, 1,3,4-tri-*O*-, 3,4,5-tri-*O*- and 1,3,4,5-tetra-*O*-galloylquinic acid

Flavonoids from *Lychnophora ericoides* require the following conditions:

Columns	Silica (250 x 10 mm i.d.) and ODS (250 x 10 mm i.d.)
Mobile phase	Water:methanol:THF
Flow rate	2.5 ml/min
Detection	280/225 nm
Compounds isolated	7,4’-dihydroxy-flavonol; 5,7-dihydroxy-3-methoxy-flavonol; galangine, 7,4’-dihydroxy-dihydroflavono l,5,7,4’-trihydroxy-dihydroflavonol,7-hydroxy-4’-methoxy-dihydroflavonol; pinobanksin, 5,7’-dihydroxy-4’-hydroxy-flavanone; 7-hydroxy-4’-methoxy-flavanone; 5,7-dihydroxy-flavone; acacetin; 7-hydroxy-3’,4’-dihydroxy-isoflavone; 15-desoxigoiazensolide, 2’,3’-dihydro-15-desoxygoyazensolide; eremantholides A and C; 4,5-dihydroeremantholide A and lychnopholide

For the isolation of peptide components of bacitracin, use:

Column	C_8 (250 x 16 mm i.d.)
Mobile phase	Gradient acetonitrile, methanol and phosphate buffer
Flow rate	9.0 ml/min
Detection	254 nm
Compounds isolated	Peptides

The isolation and purification of β-carotene from carrot involves:

Column	Shim-pack PREP-ODS (H) kit (250 x 20 mm i.d.)
Mobile phase	Ethanol (99.5%)
Flow rate	10 ml /min
Detection	480 nm monitor with photodiode array (PDA)
Compound isolated	β-carotene

The isolation of anti-HIV compounds from *Gleditsia japonica* and *Gymnocladus chinensis* requires the following conditions:

Column	C18-μBondpak (300 x 24.4 mm i.d.)
Mobile phase	Methanol-water (varying percentages)
Detection	Refractive index detector
Compound isolated	Saponins

Isolation of procyanidins from *Vicia faba* requires these conditions:

Column	Sephadex LH-20 column (580 × 25 mm i.d.)
Mobile phase	Sequential elution with ethanol, ethanol:methanol, methanol and finally methanol:acetone
Flow rate	Varying
Detection	280 nm
Compounds isolated	(+)-gallocatechin-4-phloroglucinol; (–)-epigallocatechin-4-phloroglucinol; (+)-gallocatechin; (–)-epicatechin-4-phloroglucinol; (+)-catechin-4-phloroglucinol; (+)-catechin and (–)-epicatechin

Isolation of the anticancer compound taxol from *Taxus yunnanensis* is done with:

Column	D1 (4000 x 200 mm i.d.) packed with 956 polymeric resin
Mobile phase	Acetone:water (58:42)
Flow rate	79 ml/min
Detection	228 nm
Compound isolated	Taxol

11.9 Conclusions

Stepwise scale-up starting with analytical scale to process scale is an important issue that needs to be considered. Optimization of operating conditions is always useful for getting high purity phytomolecules. Thus, process-scale HPLC is the choice for isolating valuable molecules with desired purity for commercialization. Its significance will continue to grow because of the increasing requirements for high-purity molecules.

Bibliography

Anonymous, 2007, Shimadzu application note SC-AP-LC-0186. Available at: http://www2.shimadzu.com/apps/appnotes/app113.pdf

Bouchet, N., Levesque, J. and Pousset, J. L., 2000, HPLC isolation, identification and quantification of tannins from *Guiera senegalensis, Phytochemistry Annals,* 11: 52–56

Brandt, A. and Kueppers, S., 2002, Practical aspects of preparative HPLC in pharmaceutical and development production, In: LC-GC-Europe*,* p. 2-5

Konoshima, T., Yasuda, I., Kashiwada, Y., Cosentino, L. M. and Lee, K. H., 1995, Anti-AIDS agents. XXI: Triterpenoid saponins as anti-HIV principles from fruits of *Gleditsia japonica* and *Gymnocladus chinensis*, and a structure-activity correlation, *Journal of Natural Products,* 58: 1372-1377

Lorenz, H., Sheehan, P. and Morgenstern, A. S., 2001, Coupling of simulated moving bed chromatography and fractional crystallisation for efficient enantioseparation, *Journal of Chromatography A,* 908: 201-214

Merghem, R., Jay, M., Brun, N. and Voirin, B., 2004, Qualitative analysis and HPLC isolation and identification of procyanidins from *Vicia faba, Phytochemistry Annals*, 15: 95-99

Pavli, V., Kmetec, V. and Tanja, T., 2004, Isolation of peptide components of bacitracin by preparative HPLC and solid phase extraction (SPE), *Journal of Liquid Chromatography & Related Technology,* 27: 2381-2396

Sargenti, S. R. and Vichnewski, W. 2000, Sonication and liquid chromatography as a rapid technique for extraction and fractionation of plant material, *Phytochemistry Annals,* 11: 69-73

Venn, R. F. (Ed.), 2000, Practicals and Practice of Bioanalysis, Taylor & Francis, London, p. 44-130

Yang, X., Liu, K. and Xie, M. 1998, Purification of taxol by industrial preparative liquid chromatography, *Journal of Chromatography A,* 813: 201-204

12 Flash Chromatography and Low Pressure Chromatographic Techniques for Separation of Phytomolecules

S. K. Chattopadhyay

Abstract

Flash chromatography is a rapid form of preparative column chromatography that employs prepacked columns through which a solvent is pumped at high flow rate. Two types of solvent systems are used in flash chromatography: isocratic and gradient. In the isocratic system, a single-strength mobile phase brings about the desired separation. The gradient system, in which the solvent composition changes during the course of elution, is suited for complex samples containing compounds that differ greatly in column retention times. The optimum flow rate for a flash separation is related to the particle size and dimensions of the column. Typical sorbents for normal phase flash chromatography are polar (e.g. silica, NH_2*) and elution solvents are non-polar. In reverse phase chromatography, the stationary phase is non-polar (such as C18) and the mobile phase is polar. Compounds are retained by the interaction of their non-polar functional groups with the non-polar groups on the packing surface. Therefore, the most polar compounds elute first followed by other compounds in decreasing order of polarity. To achieve a desired separation, one must select a sorbent that effectively retains the compounds of interest under solvent conditions that are appropriate for the sample's solubility. Sample loading onto a flash column can be done with wet loading (the liquid sample is loaded directly and allowed to percolate into the sorbent bed) or dry loading (when samples are pre-absorbed to a small amount of sorbent which is then loaded onto the column).*

12.1 Introduction

Mikhail Semyonovich Tsvet of Russia invented the first chromatographic technique in 1901 during his research on chlorophyll. He used a liquid adsorption column containing calcium carbonate to separate plant pigments. The method was described on 30 December 1901 at the XIth Congress of Naturalists and Doctors in St. Petersburg. The first printed description was published in 1903 in the *Proceedings of the Warsaw Society of Naturalists,* section of biology. He first used the term chromatography in print in 1906 in his two papers about chlorophyll in the German botanical journal, *Berichte der Deutschen Botanischen Gesellschaft.*

In 1952, Archer John Portor Martin and Richard Laurence Millington Synge were awarded the Nobel Prize in Chemistry for their invention of partition chromatography. Since then, the technique has advanced rapidly. Researchers have found that the principles underlying Tsvet's chromatography can be applied in many ways, giving rise to the different varieties of chromatography and allowing increasing similar molecules to be resolved.

12.2 Flash Chromatography

Flash chromatography, also known as medium pressure chromatography, is a rapid form of preparative column chromatography that uses optimized, prepacked columns through which a solvent is pumped at a high flow rate. Initially developed in 1978 by W. C. Stills of Columbia University, New York, USA, flash chromatography is now a method of purification and separation using normal phases. Use of reverse phase packing materials is opening up the technique to a wider range of preparative separations. Currently, it is considered to be a simple and economical approach to preparative liquid chromatography (LC).

Flash chromatography differs from conventional techniques in two ways. First, slightly smaller silica gel particles (250-400 mesh) are used. Second, due to the limited flow of solvents caused by the small gel particles, pressurized gas (10-15 psi) is used to drive the solvent through the column of stationary phase. The net result is rapid ("over in a flash") and high resolution chromatography.

12.2.1 Theory of Flash Chromatography

Chromatography is a separation method that exploits the differences in partitioning behavior between a mobile phase and a stationary phase to separate the components in a mixture. Compounds of a mixture may interact with the stationary phase based on charge, relative solubility or adsorption. Retention is a measure of the speed at which a substance moves in a chromatographic system. In continuous development systems like high performance LC (HPLC) and gas chromatography (GC), where the compounds are eluted with the eluents, retention is usually measured as the retention time (Rt or t_R), i.e. the time between injection and detection. In uninterrupted development systems like thin layer chromatography (TLC), retention is measured as the retention factor (R_f), i.e. the run length of the compound divided by the run length of the eluent front:

$$R_f = \frac{\text{Distance travelled by the analyte}}{\text{Distance travelled by the solvent front}}$$

12.2.2 Converting TLC to Flash Chromatography

TLC separations can be used to help determine effective solvent compositions for flash chromatography. R_f is a common TLC unit and ΔR_f is the distance between the compounds:

$$\Delta R_f = R_{f1} - R_{f2}$$

The ideal solvent system for TLC is one that moves the compound of interest in the mixture to an R_f of 0.15-0.35 and that separates this component from the others nearest to it by a ΔR_f value of at least 0.15. In contrast to TLC, flash chromatography separations are governed by column volumes.

A column volume (CV) is defined as the volume of solvent required to fill all the adsorbent pores and interstitial spaces between adsorbent particles in a given column. The volume required to elute a compound of interest from a column is expressed in terms of the number of CV. The volume that separates the elution of two substances from the same volume is called column volume difference (ΔCV). The ideal flash chromatography solvent system is one that elutes the desired compound of interest in 3-6 CV and that separates this component from others nearest to it by a ΔCV greater than 1.

The relationship between numbers of CV to R_f for a given compound is $1/R_f$; therefore, for two compounds $\Delta CV = 1/R_{f1} - 1/R_{f2.}$ For a particular set of separation conditions, a weakly retained, fast-eluting component with an R_f=0.9 can be eluted in just over 1 CV, whereas a strongly retained, slow-eluting component with an R_f=0.1 requires 10 CV for complete elution (Table 1).

Table 1: Relationship between R_f and CV

R_f	CV
0.90	1.10
0.70	1.40
0.50	2.00
0.30	3.33
0.10	10.0

Due to factors such as change in the TLC solvent flow rate with respect to time and interference from adhesives used to bind TLC sorbents, solvent conditions that provide an acceptable TLC separation will not necessarily work effectively for flash chromatography without modification.

Although some empirical experimentation may be required, the steps below help streamline the process of converting a TLC solvent system into a flash chromatography mobile phase:

1. Use matching sorbent chemistries on the TLC plate and in the flash chromatography column. Stationary phase sorbent chemistries (including silica) can differ from one manufacturer to another. It is important to match these sorbent chemistries if the solvent systems are expected to provide equivalent results.

2. Optimize the TLC solvent mixture so that the compound of interest has an $R_f \approx 0.15$-0.35 and $\Delta R_f > 0.20$. These conditions will provide the most reliable starting point for a successful flash chromatography separation.

a) *Adjust the solvent selectivity to provide an $\Delta R_f > 0.20$.* Solvent selectivity is defined as the ability to affect the retention of one compound in the mixture relative to the others, therefore affecting ΔR_f and number of CV. Experimenting with different solvent combinations to obtain the desired TLC separation usually reveals appropriate conditions for effective flash chromatography separation. Different solvent mixtures such as hexane:ethyl acetate (1:1) and hexane:dichloromethane (1:2) may provide different solvent selectivities while providing similar solvent strengths. Different solvent mixtures can even reverse the elution order of some of the components in the sample. It is interesting to note that ΔR_f and ΔCV may vary greatly relative to one another for a given separation. ΔCV predicts column capacity, i.e. the amount of material that can be effectively separated in a single column loading (Table 2). Greater the ΔCV, the better the effective capacity of the column.

b) *Adjust the solvent strength to obtain an R_f between 0.15 and 0.35 (CV, 3-6).* Solvent strength refers to the solvent's simultaneous effect on the retention of all compounds in the mixture; therefore, solvent strength affects R_f and CV. Once the optimum separation has been established by modifying solvent selectivity, it may be useful to move some or all of the compounds off the flash column as quickly as possible by increasing solvent strength. Often, slight changes in solvent strength can make large differences in retention. In some cases, a lower-strength mobile phase provides improved separations. It is important to remember that the sample loading solvent should have equal or lower elution strength than the starting strength of the mobile phase. Additional adjustments to the selectivity and strength of the flash solvent system may be necessary to optimize the separation and to achieve a CV$\approx$3-6 and a ΔCV>1. This can often be achieved by using a less polar solvent system or by decreasing the proportion of polar modifier.

Table 2: Approximate capacity of a 20 g/70 ml ISOLUTE Flash Si column (Biotage)

ΔCV	**Sample load, g**
6	1.0
2	0.5
1	0.25

12.3 Isocratic versus Gradient Chromatography

Two types of solvent systems are used in flash chromatography: isocratic and gradient. The most common is an isocratic (meaning “same solvent strength”) system where a single-strength mobile phase mixture brings about the desired separation.

If the mixture is complex and contains compounds that differ greatly in column retention times, chemists may use a gradient solvent system that changes solvent composition during the course of elution. For example, in a normal phase system equipped with a silica column, a non-polar solvent such as hexane is applied to elute non-polar compounds. Then, a more polar solvent such as ethyl acetate is added to the hexane to elute the more polar compounds. The percentage of the polar solvent in the mixture is increased until all components of the mixture have eluted.

In a step-gradient system, the various solvent concentrations are typically changed in large increments (or steps). Alternatively, a linear gradient can be employed whereby a continuous linear change in the concentrations of the solvent (and thus mobile phase strength) is achieved. Chemists can often achieve effective separations more rapidly by using gradient solvent systems. Chemists must select miscible solvents for use in gradient solvent systems. A common solvent system for flash separations using polar sorbents such as silica is hexane and ethyl acetate, where ethyl acetate is the more polar solvent.

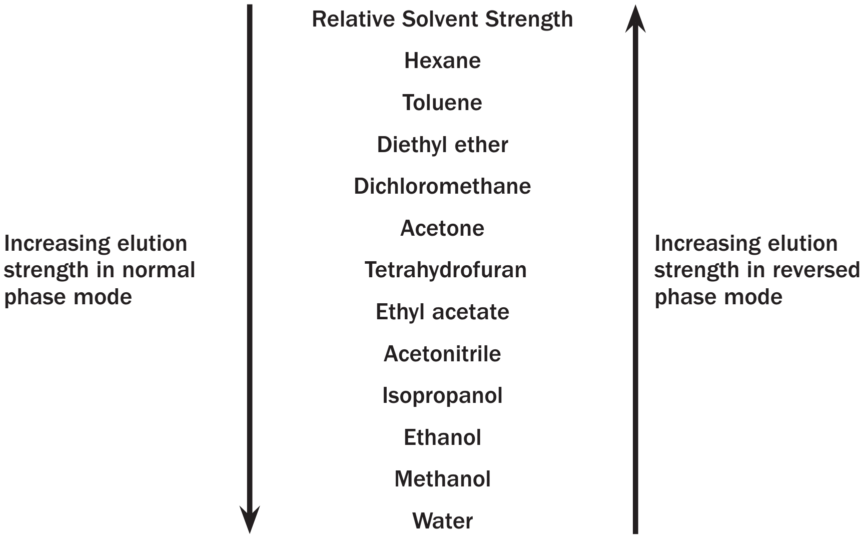
Relative Solvent Strength
Hexane
Toluene
Diethyl ether
Dichloromethane
Acetone
Tetrahydrofuran
Ethyl acetate
Acetonitrile
Isopropanol
Ethanol
Methanol
Water
Increasing elution strength in normal phase mode
Increasing elution strength in reversed phase mode

Figure 1: Step-gradient system

12.4 Adsorbent Selection and Mode of Separation

Typical sorbents for normal phase flash chromatography are polar (e.g. silica, NH_2) and elution solvents are non-polar (e.g. hexane, heptane, dichloromethane, sometimes modified with small amounts of more polar solvents such as isopropanol). The sample is usually applied in a weak (very non-polar) solvent, and separation occurs when the elution solvent is applied. In normal phase flash chromatography, the most non-polar sample component elutes first, followed by successively more polar compounds.

12.4.1 Isolute Flash Columns

Isolute flash columns (Biotage) are polypropylene columns prepacked with Isolute flash sorbents. These columns are appropriate for use in both off-line and on-line flash chromatography. In off-line flash chromatography, chemists apply the sample and elution solvent volumes to the top of an Isolute column fixed in a FlashVac Sample Processing Station. In off-line flash chromatography, chemists perform isocratic or step gradient separations. In on-line flash chromatography, an Isolute column is mounted on a system that connects the column to an external liquid pump system to produce a continuous flow of solvent through the column. Depending on the capability of the pump, the solvent composition can be isocratic (a single solvent or solvent mixture) or a gradient with an increasing proportion of stronger solvent (either in a step gradient or a linear gradient).

12.4.2 Method Development Using Isolute Flash Columns

12.4.2.1 Column Equilibration

Prior to sample loading, the column should be prepared for the separation by equilibration (prewetting) with a suitable solvent:

- *Off-line*. Apply the equilibration solvent to the top of the column and allow it to flow through the column under gravity.
- *On-line.* Equilibrate the flash column for on-line mode separation in the off-line mode using a vacuum manifold such as the FlashVac system or by mounting the column on the on-line apparatus and pumping a suitable volume of equilibration solvent through the column.

12.4.2.2 Typical Equilibration Solvents

Suitable solvents are non-polar, e.g. hexane or pentane. For best results, prewet the silica and NH_2 columns prior to use. A suitable volume for column equilibration is approximately two bed volumes (Table 3).

Table 3: Typical column solvent volumes for column equilibration

Column size, g	Approximate column volume, ml	Typical solvent volume, ml
2	2.5	5
5	6.5	13
10	12.5	25
20	25	50
25	31.3	62.5
50	62.5	125
70	88	176
100	125	250
150	188	376

Normal phase flash columns can be used without prewetting, but some column-to-column variation may be experienced.

12.5 Sample Application

There are two popular approaches to flash chromatography sample loading: wet loading and dry loading.

12.5.1 Wet Loading

Load the liquid sample directly onto the top frit of the column, and allow it to percolate through the top of the sorbent bed. For best results, load the sample onto a prewetted column in a non-polar solvent.

12.5.1.1 Practical Tips for Wet Loading

- Dissolve the sample in as non-polar as solvent as possible.
- If compounds are not easily soluble in a non-polar solvent, either dissolve them in a small volume of polar solvent and then dilute with a non-polar solvent to reduce the elution strength, or consider dry loading of the sample (discussed in next section).
- Position column on a FlashVac vacuum manifold equipped with PTFE stopcock needles.
- Apply the sample evenly over the entire area of the top frit. To do so, seal the column by closing the stopcock.
- Alternatively, load samples onto the column in on-line mode using a Flash Master system equipped with a 3-way injection valve.

12.5.2 Dry Loading

Dry loading is the method of choice for loading reaction mixtures consisting of polar solvents onto silica or other normal phase columns. Pre-absorb the reaction mixture onto a small amount of bulk material of the chosen sorbent. Evaporate off the majority of the solvent leaving the compounds bound to the surface of the sorbent.

Add this blend to the top of the prepacked (and prewetted) flash column; allow settling and add a top frit to secure the blend in place. The top frit can be placed using a suitably sized frit inserter.

A popular alternative sorbent for dry loading using the flash sorbent is a diatomaceous earth such as Isolute HM-N. This can be used in the same way as the flash sorbent, but has several advantages including more efficient desorption of the compounds into the mobile phase.

12.5.2.1 Practical Tips for Dry Loading

- Dissolve the sample initially in a suitable solvent, ensuring complete dissolution if possible. Use the smallest volume possible.
- Add the bulk material of choice. The ideal proportion of sample to bulk material ranges from 1:1 to 1:3 by volume.
- Evaporate off the residual solvent using a rotary evaporator to ensure even adsorption of the sample on the bulk material.
- Pack the dry blend on top of the flash column (above the top frit) and add another frit. Push down the new surface to prevent movement of the new blend. When loading with Isolute HM-N, ensure that the material is not crushed at this stage.
- When dry loading using bulk silica, ensure that it is identical to the material in the flash column. If this is not possible, use a lower surface area material, ensuring that the surface pH and moisture content are as similar as possible to the column packing. All Isolute sorbents are available as bulk material.

Table 4: Capacity guidelines

Reaction scale, g	Column size, g
0.1	1-2
0.25	2-5
0.50	5-10
1.0	10-20
2.5	25-20
5.0	50-70
10.0	70-100

As a general guideline, the amount of sample loaded onto the normal phase flash column should be 5%-10% of the column size. Factors affecting the capacity of the column include compounds, sample matrix, concentration of reaction products and elution solvent used.

12.6 Elution

12.6.1 Step Gradient Elution

Step gradients provide controlled elution with discrete changes in eluent strength. Each step is optimized to elute only those components that are soluble in that eluent. This technique can be applied to both off-line and on-line flash chromatography.

1. Use TLC to determine a suitable solvent strength to elute the components at discrete intervals, choosing different solvent mixtures that elute each component separately with an R_f of 0.2-0.5.
2. Calculate the volume of solvent required to elute each component using $CV=1/R_f$. Values of CV for different column sizes are listed in Table 3.
3. Apply between 2 and 5 CV of solvent for each step, starting with the solvent with weakest strength.
4. Collect the eluent from each step in a separate vessel.

12.6.2 Linear Gradient Elution

Liner gradients are a quick way of separating complex mixture, reducing the complexity of the subsequent purification of the fractions collected. This technique is suitable for on-line flash chromatography.

1. Use TLC to find both the weakest and strongest elution solvent. The weak solvent (solvent A) should give retention of the majority of components ($R_f<0.1$). The strong solvent (solvent B) should allow elution of all of the components of interest ($R_f>0.5$).
2. Run a gradient starting with 100% solvent A and ending with 100% solvent B.
3. Collect the eluent at regular intervals.

Automated flash chromatography instrumentation such as the FlashMaster II and Solo systems can monitor the signal from a UV detector, and collect only the fractions of eluent carrying the compounds of interest.

12.6.3 Method Development Using Gradient Elution

Method development can be performed without TLC. Gradient elution analysis offers a useful approach to method development, particularly for non-silica-based flash chromatography (e.g. reverse phase) where suitable TLC plates are not available.

1. Load the sample onto a prewetted (equilibrated) flash column in as weak a solvent as possible (or use dry loading); for example, for normal phase work use hexane for loading.
2. Elute the column with aliquots (2 CV each) of successively increasing solvent strength. A typical scheme for mixing the two solvents is given in Table 5.
3. Collect each fraction and analyze for the presence of the components of interest.
4. Using these data, identify the solvent mixture that elutes the components of interest separately, and set up a step or continuous gradient as described.

Table 5: Typical scheme of mixing solvents for method development. Solvent A is a weak solvent (e.g. hexane) while solvent B is a strong solvent (polar modifier, e.g. isopropanol)

Aliquot number	**Solvent A, %**	**Solvent B, %**
1	100	0
2	99	1
3	98	2
4	97	3
5	06	4
6	95	5
7	94	6
8	93	7
9	92	8
10	91	9

12.6.4 Practical Tips for Gradient Elution

1. The use of a gradient does not improve the selectivity of a separation if isocratic elution using the same solvent system does not effectively separate the sample components. However, a gradient can be used to decrease the time required to achieve a separation.

2. The gradient starting conditions must not cause chromatography separation. Start the gradient with a weak solvent that matches the sample loading conditions.
3. The flow rate at the beginning of a gradient can be high but, for best results, it should be reduced to the optimum flow rate at the separation area.
4. To speed up a gradient separation, use either a higher flow rate (most suited to samples with many components, high sample load) or a steeper gradient (most suited to samples with few components, low sample load).

12.6.5 Optimizing Flow Rate

The optimum flow rate for a flash separation is related to the particle size and dimensions of the column. Theoretical optimum flow rate for flash columns of different dimensions can be predicted. However, in practice, increasing the flow rate has not had a significant effect on separation and offers important productivity advantages (Table 6). Other factors, such as mobile phase composition and back pressure, also affect the range of effective flow rates. For recommended flow rates, see Table 6.

Table 6: Recommended flow rates

Column diameter (configuration)	**Flow rate range ml/min**
16 mm (D)	5-25
20 mm (E)	5-25
27 mm (F)	10-30
37 mm (J)	20-50
40 mm (V, W, X)	20-50

12.7 Fraction Collection

12.7.1 Off-line Flash Chromatography

When performing flash chromatography on a vacuum manifold such as the FlashVac system, successive fractions can be collected as follows:

- Load collection rack with vials of a suitable volume in each position.
- Place a single flash column in position 1 of the collection rack and apply the first solvent aliquot.
- Collect the aliquot in the vial in position 1 of the collection rack.
- Move the column to position 2 and apply the second solvent aliquot.

- Collect the aliquot in the vial in position 2 of the collection rack.
- Continue until all the components of interest have been collected.

Alternatively, multiple columns can be processed by replacing collection vials at each elution step. A typical volume for each fraction is two CV.

12.7.2 On-line Flash Chromatography

Using an automated system equipped with a fraction collector, fractions can be collected in a variety of ways, for example:

- Fixed volume fraction collection.
- Individual peak fraction collection under microprocessor control with UV detector input.

12.8 Low Pressure Liquid Chromatography

In low pressure column chromatography, a column of particulate material such as silica or alumina has a solvent passed through it at atmospheric or low pressure. There are different kinds of low pressure chromatographic techniques:

i) Gel filtration chromatography (separation on the basis of size)
ii) Ion exchange chromatography (separation of the basis of charge)
iii) Affinity chromatography (separation on the basis of specific binding sites on the proteins)

12.8.1 Gel Filtration Chromatography

Proteins of different sizes are separated on a column in which the stationary phase consists of polymerized agarose or acrylamide beads with pores of particular sizes. A small protein in the mobile phase (aqueous buffered solution) can enter the pores in the beads while a large protein cannot due to size restriction. The result is that a smaller fraction of the overall volume of the column is available to the large protein than to the small protein, which thus spends a longer time on the column and is eluted by the mobile solvent after the large protein.

12.8.2 Ion Exchange Chromatography

The material used for this type of chromatography consists of an agarose, acryl amide or cellulose resin or bead which is derivatized to contain covalently linked positively or negatively charged groups. Proteins in the mobile phase bind through electrostatic interactions to the charged groups on the column.

In a mixture of proteins, positively charged proteins bind to a resin containing negatively charged groups like carboxymethyl ($-OCH_2COO^-$) or sulfopropyl ($-OCH_2CH_2CH_2SO_3^-$), while the negatively charged proteins pass through the column. The positively charged proteins are eluted from the column with a mobile phase containing either a gradient of increasing salt concentration or a single higher salt concentration. The most positively charged proteins are eluted last, at the highest salt concentration.

Likewise, negatively charged proteins bind to a resin containing positively charged groups, like diethylaminoethyl ($-CH_2CH_2NHEt_2^+$) or a quaternary ethyl amino group. These proteins are separated in an analogous way.

12.8.3 Affinity Chromatography

In this technique, the chromatography resin is derivatized with a group that binds to a specific site on a protein of interest. It may be a group that binds to the active site of an enzyme (such as benzamidine-agarose used for the purification of trypsin) or an antibody that recognizes a specific amino acid sequence on a protein.

This method exploits the specific binding of antibody to antigen held on a solid matrix. Antigen is bound covalently to small, chemically reactive beads which are loaded in the column and the antiserum is allowed to pass over the beads. The specific antibodies bind while all other proteins in the serum including antibodies to other substances are washed away.

Affinity chromatography can also be used to purify antigens from complex mixtures by using beads coated with the specific antibody.

12.9 Conclusions

Flash chromatographic systems have been developed for the separation and purification of organic molecules from natural sources and from reaction mixtures. Method development for separations can be optimized with proper selection of adsorbents, solvent systems and flow rate of solvents used for the separation. Scale-up of flash chromatographic methods can easily be achieved with minimum optimization. Therefore, they are considered to be efficient and cost-effective methods to purify compounds in little time.

Bibliography

Glasl, S., Gunbilig, D., Narantuya, S., Werner, I., and Jurenitsch, J., 2001, Combination of chromatographic and spectroscopic methods for the isolation and characterization of polar guaianolides from *Achillea asiatica*, *Journal of Chromatography A,* 936: 193-200

Still, C. W., Kahn, M., and Mitra, A., 1978, Rapid chromatographic technique for preparative separations with moderate resolution, *Journal of Organic Chemistry,* 43: 2923

13 Counter-current Chromatography

S. K. Srivastava

Abstract

Counter-current Chromatography (CCC) is based on liquid-liquid partitioning and is an excellent alternative to circumvent the problems associated with solid phase adsorbents and to preserve the chemical integrity of mixtures subjected to fractionation. The uniqueness of this technique is that it does not use a stationary phase and separation is achieved between two immiscible liquid phases. One phase is used as a stationary phase while the other one is used as mobile phase. The separation takes place on the basis of partition of a sample between the two phases. In contrast to other chromatographic techniques, since there is no irreversible adsorption of sample on the stationary phase, recovery is up to 100%. Centrifugal Partition Chromatography (CPC) and High Speed Counter-current Chromatography (HSCCC) are the two variants of CCC being used today. Separation of compounds with a wide range of polarities is possible with the use of aqueous and non-aqueous solvent systems. Separation of crude plant extracts, semipurified fractions and synthetic mixtures can be carried out, with sample loads ranging from milligrams to grams.

13.1 Introduction

Natural products have been a key source for the discovery of new drugs. Isolation of the active components from a natural product has always been associated with complex separation problems due to the enormous chemical complexity of the extracts, but recent advances in separation sciences have facilitated the isolation of these active components from natural products.

Crude extracts of natural products which show desirable biological activity are subjected to activity-guided fractionation until an active component is isolated and identified. This exploratory process of fractionation typically involves suboptimal chromatographic conditions; hence, in order to avoid destruction of potentially labile components, utmost care must be taken throughout the entire process of isolation.

At present, most of the chromatographic separations of natural products are being carried out on solid supports. However, SiO_2, Al_2O_3 and reverse phase adsorbents are not chemically inert. Separation of a natural product on alumina or silica gel sometimes results in recovery of only 70%–90%. Sometimes severe losses of valuable materials result because of irreversible adsorption on a solid support. In addition, isolation of artifacts has also been reported due to chemical reactions of the substrates with solid phase adsorbents.

Counter-current Chromatography (CCC) is a unique form of liquid partition chromatography which utilizes a separation column free of solid support matrix. Because of this support-free system, the method provides an important advantage over other chromatographic methods by eliminating various complications such as adsorptive loss and deactivation of samples and contamination. During the 1970s, the method was steadily improved by accelerating the separation speed and efficiency. In the early 1980s, an epoch-making advance was achieved by the advent of high-speed CCC (HSCCC) which can yield highly efficient separation in a short period of time. Because of its high performance, the recent research and development of the CCC technology have been almost entirely focused on HSCCC, high performance centrifugal partition chromatography (HPCPC), and fast centrifugal partition chromatography (FCPC).

Recent developments in CCC instruments, explaining the use of CCC for better separation and their applications in the separation of bioactive natural products from plants, are discussed in this paper.

13.2 Principles and Development of Counter-current Chromatography

Anyone conversant with the technique of liquid-liquid extraction (using a separatory funnel) can readily understand the principles of Counter-current Chromatography (CCC), where separation is based on the partition of solutes between two immiscible liquid phases. In CCC, one of the two phases is retained in the column and is called the stationary phase. The second phase, which is called the mobile phase, percolates through the stationary one.

13.2.1 Liquid-liquid Extraction

Liquid-liquid extraction is a simple means of separating large quantities of materials, using a minimum of solvent. After dissolving the sample in a two-phase solvent system (in a separatory funnel), the steps in performing liquid-liquid extractions are as follows:

- Shake vigorously to thoroughly mix the two phases.
- Allow the mixture to settle into two phases.
- Separate the phases from each other.

These steps are crucial to achieve the separation of sample components.

The chief disadvantage of liquid-liquid extraction is that it provides only one plate of separation in the original sense.

Accordingly, either this single-step separation must be designed to suit one's needs, or multiple liquid-liquid extractions must be used to increase the separation.

CCC has its origin in the work of Archer John Porter Martin and Richard Laurence Millington Synge (Martin and Synge, 1941; Synge, 1946) carried out in Britain during World War II. For their pioneering work, Martin and Synge shared the 1952 Nobel Prize in Chemistry. Soon after their work appeared, Lyman Creighton Craig and Otto Post developed an apparatus that essentially consisted of a series of separatory funnels ("tubes") (Craig and Post, 1949). The sample was "automatically" transferred through the Craig-post apparatus (Figure 1). Over 1000 mixing and separation steps could be accomplished in one day.

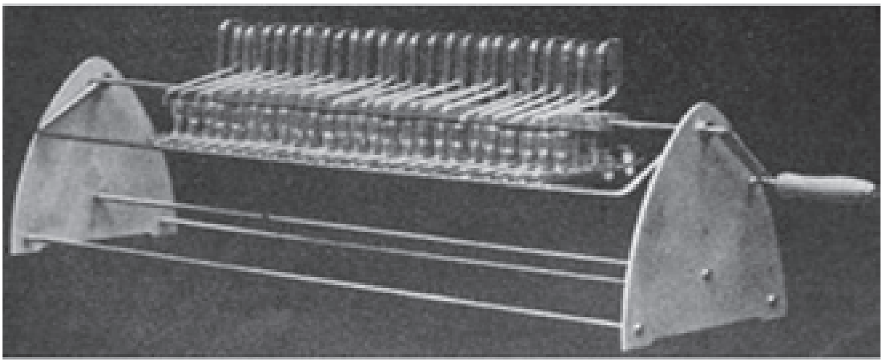

Figure 1: Craig-post apparatus

Individual components were separated based on their partitioning behavior. Craig and Post continuously improved their apparatus and were commercially quite successful. Over 1000 publications on "counter-current distribution" appeared during the period 1950-1970 citing the use of the Craig-post apparatus.

13.2.2 Partition Coefficient

For a given substance A, the partition coefficient KA is defined the concentration of A in the upper phase divided by that in the lower phase (Figure 2).

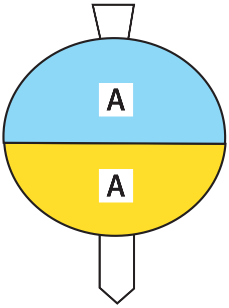

Figure 2: KA= [A] upper phase/[A] lower phase

KA is a constant at any given temperature, and it is unaffected by the presence of other substances or by the concentration of the solute. Usually it is expressed as the amount of solute in the stationary phase divided by that of the mobile phase as in conventional liquid chromatography. For a better CCC separation, the K values of the target compounds should be in the range of $0.5 \leq K \leq 1.0$. A smaller K value elutes the solute closer to the solvent front with lower resolution while a larger K value tends to give better resolution but broader, more dilute peaks due to a longer elution time. In CCC, one can choose either the upper or lower phase as the stationary phase, but before deciding which phase is to be used as the stationary phase, one should temporarily express the partition coefficient as $K_{U/L} = C_U/C_L$, where C_U is the solute concentration in the upper phase and C_L is that of the lower phase. If $K_{U/L} = 2$, the lower phase should be used as the stationary phase, which gives $K = 0.5$. It is important that this preliminary $K_{U/L}$ be clearly distinguished from K using the subscripts to avoid confusion. The measurement of $K_{U/L}$ values may be performed by the shake-flask method.

Add a small amount (a few milligrams) of each target compound (one at a time) to the two mutually equilibrated solvent phases (1–2 ml each) in a stoppered test tube (13 mm × 100 mm). Thoroughly mix to equilibrate the contents. After settling, pipette and deliver an equal volume of the upper and the lower phases (100–200 µl) each into a separate test tube, dilute each with an equal volume (2 ml) of a suitable solvent such as methanol, and measure the absorbance with a spectrophotometer at a suitable wavelength to obtain the $K_{U/L}$ value. If the pure sample is not available, one can subject each phase to HPLC analysis to compare the peak heights (or areas under the peaks). If the sample does not absorb in the UV or visible wavelengths, the $K_{U/L}$ value may be determined by other ways such as thin layer chromatography (color reaction) or evaporative light scattering detection (ELSD). Thin layer chromatography combined with densitometry is effectively used for simultaneous determination of $K_{U/L}$ of multiple components from a sample mixture.

13.2.3 Droplet Counter-current Chromatography

In the early 1970s, Ito and colleagues at the US National Institutes of Health introduced Droplet Counter-current Chromatography (DCCC). In DCCC, as with the Craig-post apparatus, the stationary phase is held in place only with unit gravity.

The apparatus consists of a set of vertical straight tubes serially connected with narrow transfer tubing (Figure 3). The original CCC apparatus was equipped with 300 glass tubes, each 60 cm × 1.8 mm i.d.

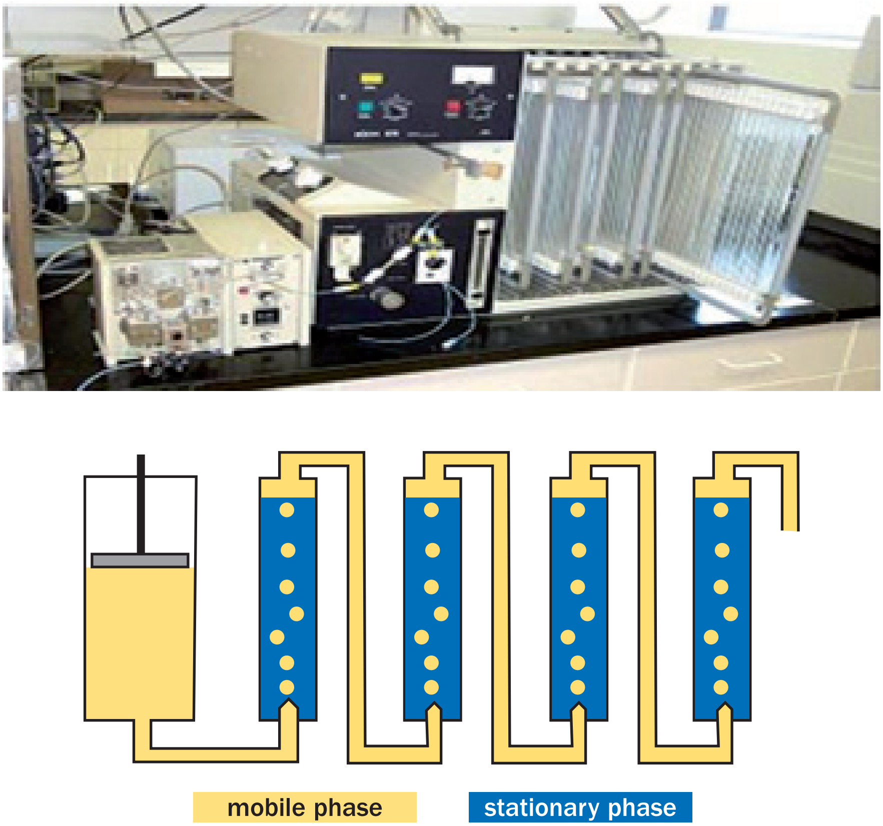

Figure 3: Droplet counter-current chromatography

The total capacity is about 600 ml including the volume in the transfer tubing (about 15% of total capacity). The operation of DCCC is initiated by filling the entire column with the stationary phase of an equilibrated two-phase solvent system followed by injection of sample solution. Then, the other phase is introduced into the first unit in such a way that the mobile phase travels through the column of the stationary phase by the effect of gravity, i.e. the mobile phase is introduced from the bottom if it is the lighter phase, and from the top if it is heavier. Consequently, the solutes are separated according to their partition coefficients. As with all types of chromatography, compounds more soluble in the mobile phase move more quickly, while those more soluble in the stationary phase lag behind. Under optimum flow rate, the mobile phase forms multiple droplets in the stationary phase to divide the column space into numerous partition units; this process is repeated in each partition unit.

DCCC necessitates a proper choice of solvent systems for producing a droplet flow of the mobile phase in the column. The most popularly used solvent system is composed of chloroform, methanol and water at various volume ratios, although a sizable separation usually takes a few days.

13.2.4 Applications of Droplet Counter-current Chromatography

This simple DCCC instrument has been used to perform efficient preparative separations such as:

- Separation of non-polar compounds.
- Isolation of virginiamycin-M1 and parthenolide.
- Isolation of vitamin B12.
- Pharmacognostical studies of *Tabernaemontana* species: ion-pair DCCC of indole alkaloids from suspension cultures.
- Chiral resolution of a carboxylic acid.
- Separation of natural polar substances by reverse phase HPLC, centrifugal thin layer chromatography and DCCC.
- Efficient isolation of ecdysteroids from the silkworm, *Bombyx mori*.
- Analytical DCCC isolation of 20-hydroxyecdysone from *Vitex thyrsiflora* (Verbenaceae).
- Increasing the speed of DCCC separations.
- Complete resolution of isoleucine.
- DCCC of anthocyanins.
- Efficient isolation of phytoecdysones from *Ajuga* plants by high-performance liquid chromatography and DCCC.
- DCCC with non-aqueous solvent systems.
- Use of DCCC in log *P* determinations.
- Purification of *Stevia rebaudiana* sweet constituents (potential sweetening agent of plant origin).
- Water-free solvent system for DCCC and its suitability for the separation of non-polar substances.
- Isolation of phorbol, 4α-phorbol and croton oil.
- Purification of antibiotics such as gramicidins, tyrocidines and tetracyclines.

13.2.5 Limitations of DCCC

- Extremely low flow rates (sometimes solute retention is measured in days).
- Only biphasic solvents systems that form stable droplets can be used.
- Poor mixing of phases, which results in relatively low efficiency.

13.2.6 Modern Counter-current Chromatography

Modern CCC has split into two basic directions. The first, which is called High-speed Counter-current Chromatography (HSCCC), uses an apparatus with a variable gravity field produced by a double axis gyratory motion (Figure 4). The second, termed Centrifugal Partition Chromatography

(CPC), employs a constant gravity field produced by a single axis rotation, together with rotatory seals for supply of solvent. Separation takes place in cartridges or disks. CPC with cartridges or disks is a hydrostatic equilibrium system. If the coil is filled with the stationary phase of a biphasic solvent system and then the other phase is pumped through the coil at a suitable speed, a point is reached at which no further displacement of the stationary phase occurs and the apparatus contains approximately 50% of each of the two phases. Steady pumping-in of mobile phase results in elution of mobile phase alone. This basic system uses only 50% of the efficient column space for actual mixing of the two phases.

A more effective way of using the column space is to rotate the coil around its central axis while eluting the mobile phase. A hydrodynamic equilibrium is rapidly established between the two phases and almost 100% of the column space can be used for their mixing. CCC with rotating coil instruments is an example of this latter mechanism.

- The planetary motion is produced by engaging a planetary gear mounted on the column holder axis to an identical stationary sun gear rigidly fixed to the centrifuge framework.
- This 1:1 gear coupling produces a particular type of planetary motion of the column holder, i.e. the holder rotates about its own axis while revolving around the centrifuge axis at the same angular velocity (synchronous) in the same direction.
- This planetary motion provides two major functions for performing CCC separation. The first is a rotary-seal-free elution system so that the mobile phase is continuously eluted through the rotating separation column.
- The second and more important function is that it produces a unique hydrodynamic motion of two solvent phases within the rotating multilayer coiled column mainly due to the Archimedean screw effect.

13.3 HSCCC Instrument and Mechanism

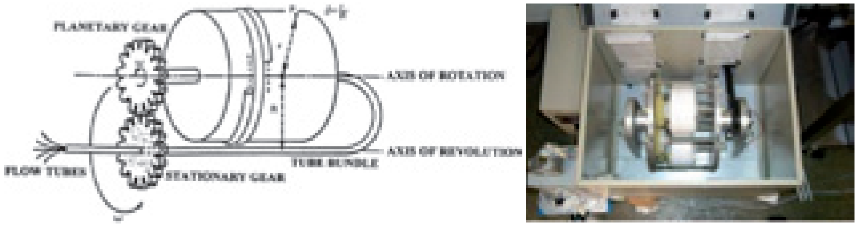

Figure 4: High speed counter-current chromatography instrument

When two immiscible solvent phases are introduced in an end-closed coiled column, the rotation separates the two phases completely along the length of the tube where the lighter phase occupies one end (called the head) and the heavier phase occupies the other end (called the tail). Although the cause of this bilateral hydrodynamic phase distribution of two immiscible solvents is still unknown, it can be efficiently used for performing CCC. In Figure 5A the coil at the top shows bilateral hydrodynamic distribution of the two phases in the coil where the white phase (head phase) occupies the head half and the black phase (tail phase) the tail half.

This condition clearly indicates that the white phase introduced at the tail end will move toward the head and similarly the black phase introduced at the head will move toward the tail. This hydrodynamic trend is effectively used for performing CCC (Figure 5B). The coil is first entirely filled with the white phase followed by pumping the black phase from the head end (Figure 5B, top). Similarly, the coil is filled with the black phase followed by pumping the white phase from the tail (Figure 5B, bottom). In either case, the mobile phase quickly moves through the coil, leaving a large volume of the other phase stationary in the coil.

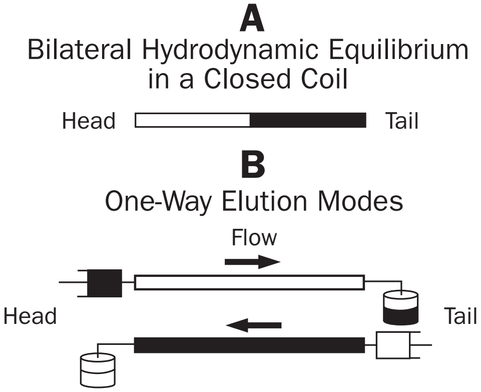

Figure 5: Mechanism of HSCCC. (A) Bilateral hydrodynamic distribution of two phases in the coiled column. (B) Elution mode of both lighter and heavier phases through the rotating coiled column.

- The motion and distribution of the two phases in the rotating coil were observed under stroboscopic illumination, and are illustrated in Figure 6A.
- A spiral column undergoes type-J planetary motion. The area in the spiral column is divided into two zones: the mixing zone occupying about one-quarter of the area near the center of revolution and the settling zone in the rest of the area.
- In Figure 6B, the spiral column is stretched and arranged according to the positions I–IV to visualize the motion of the mixing zones along the tubing.
- Each mixing zone travels through the spiral column at a rate of one round per revolution.

- The solute in the spiral column is subjected to the repetitive partition process of mixing and settling at an enormously high rate of over 13 times per second (at 800 rpm).
- HSCC is highly efficient.

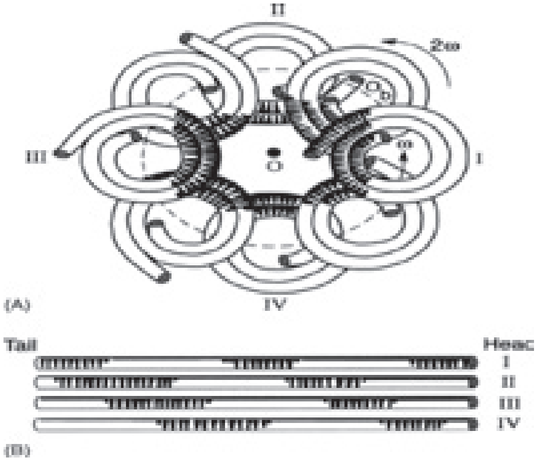
II
2ω
III
I
O
IV
(A)
Tail
Head
I
II
III
IV
(B)

Figure 6: (A) Motion and distribution of two phases in the rotating coil under stroboscopic illumination. (B) Column in stretched position to visualize the motion of the mixing zones along the tubing.

There are numerous potential variants of this instrument design. The most significant of these is represented by the third instrument in the Pharma-tech product line: TCC-1000-toroidal CCC (Figure 7). In some respects it is like CPC, but retains the advantage of not needing rotary seals. In addition, it employs a capillary tube, instead of the larger-diameter tubes employed in the helices of the other CCC models. This capillary passage makes the mixing of two phases very thorough, despite lack of any shaking or other mixing forces.

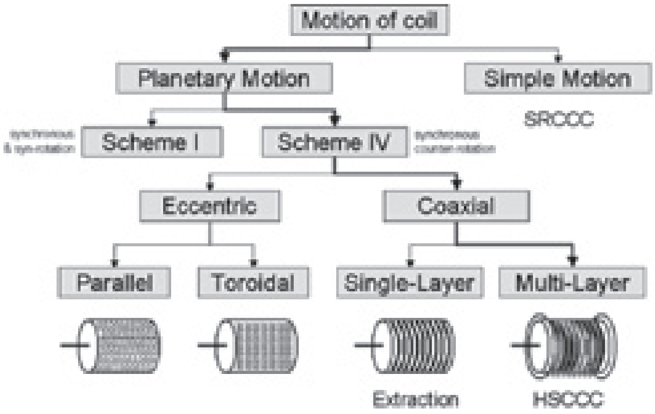
Motion of coil
Planetary Motion
Simple Motion
SRCCC
synchronous & syn-rotation
Scheme I
Scheme IV
synchronous counter-rotation
Eccentric
Coaxial
Parallel
Toroidal
Single-Layer
Multi-Layer
Extraction
HSCCC

Figure 7: The TCC-1000-toroidal CCC

13.4 CPC Instrument and Setup

Figure 8: Centrifugal partition chromatography instrument

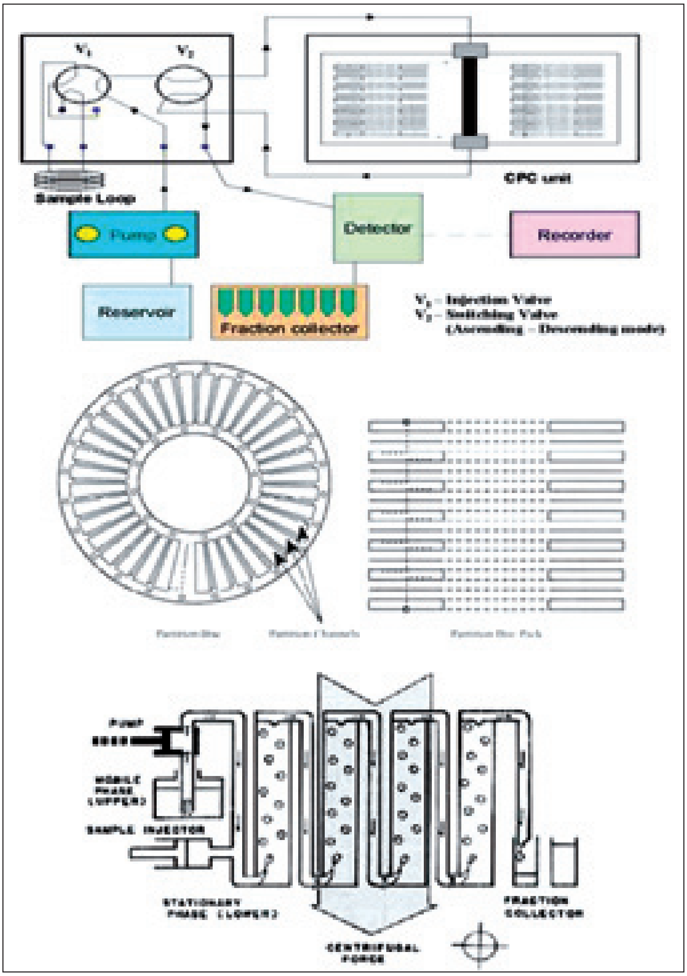
V1
V2
CPC unit
Sample Loop
Pump
Detector
Recorder
Reservoir
Fraction collector
V1 – Injection Valve
V2 – Switching Valve
(Ascending – Descending mode)
Partition Disc
Partition Channels
Partition Disc Pack
PUMP
MOBILE PHASE (UPPER)
SAMPLE INJECTOR
STATIONARY PHASE (LOWER)
FRACTION COLLECTOR
CENTRIFUGAL FORCE

Figure 9: Centrifugal partition chromatography setup

13.4.1 Various Kinds of CPC Instruments

CPC instruments are available in various sizes depending upon the rotor volume. As per the sample size and the purpose, one can choose on the basis of the following:

Rotor volume	Scale	Sample amount
100 ml	Analytical screening	Milligrams
200 ml	Analytical preparative	<5 g
1000 ml	Pilot scale	<40 g
5000 ml	Production scale	100–200 g
10000 ml	Production scale	200–400 g
Other volumes	On request	Up to one kilogram

13.5 How to Achieve Good Separation of Various Kinds of Natural Products Using HSCCC and CPC

In order to achieve the best separation of a desired natural product, one must systematically follow the guidelines presented in this section on choice of a solvent system.

13.5.1 Search for a Suitable Solvent System

In Table 1, sets of two-phase solvent systems are arranged from top to bottom in decreasing order of hydrophobicity in the organic phase.

Table 1: Solvent systems arranged in decreasing order of hydrophobicity in the organic phase

	n-Hexane	EtOAc	MeOH	n-BuOH	Water	
	10	0	5	0	5	↑
	9	1	5	0	5	
	8	2	5	0	5	
	7	3	5	0	5	
	6	4	5	0	5	
	5	5	5	0	5	HYDROPHOBIC
	4	5	4	0	5	
START HERE →	3	5	3	0	5	
	2	5	2	0	5	
	1	5	1	0	5	
	0	5	0	0	5	
	0	4	0	1	5	POLAR
	0	3	0	2	5	
	0	2	0	3	5	
	0	1	0	4	5	
	0	0	0	5	5	↓

When the polarity of the target compounds is unknown, the search may start with the two-phase solvent system composed of hexane–ethyl acetate–methanol–water at a volume ratio of 3:5:3:5, which has a moderate degree of polarity. If the partition coefficient is slightly off from the proper range, it can be adjusted by modifying the volume ratio. For example, if $K_{U/L}$ is slightly over 2, the volume ratio may be modified toward more hydrophobic such as 3.2:5:3.2:5, and if $K_{U/L}$ is slightly less than 0.5, the volume ratio may be modified in the opposite direction to 2.8:5:2.8:5.

If the target compound is mostly distributed in the upper organic phase, the search is directed upward along the arrow. If it is mostly distributed in the lower aqueous phase, the search is directed downward along the arrow. If the sample is an extract of plant material, the search may start at any point according to the polarity of the solvent used for the extraction.

If the sample is an ethyl acetate extract (relatively hydrophobic solvent), the search may start at hexane–ethyl acetate–methanol–water (1:1:1:1), whereas if the sample is a methanol extract (polar solvent), the search may start at 1-butanol–water. The search should be continued until a suitable range of $K_{U/L}$ values for all of the compounds of interest is obtained.

13.5.1.1 Solvent Systems for the Separation of a Large Variety of Natural Products

Table 2 presents different kinds of solvent systems that have been used for the separation of a large variety of natural products, which may be of immense importance for the researcher to find a suitable solvent system for the desired separation of natural products.

Table 2: Solvent systems used for the separation of natural products

Sample	Solvent system		Mobile phase
	Solvents	Ratio	
Aryl ketones			
Phenylethanone derivatives	*n*-Heptane–EtOAc–MeOH–H_2O	4:2:2:1	Lower
Styrylpyrones			
Kava lactones	*n*-Hexane–acetone–MeOH–H_2O	4:1:3:1	Upper
Benzofurans			
Dibenzofurans	*n*-Heptane–CH_2Cl_2–CH_3CN	10:3:7	Lower, upper

Sample	**Solvent system**		**Mobile phase**
	Solvents	**Ratio**	
Phenolic glycosides			
Gastrodin	EtOAc–1-BuOH–H_2O	3:2:5	Lower
Phenylpropanoids			
Honokiol, magnolol	*n*-Hexane–EtOAc–MeOH–H_2O	5:2:5:2	Lower
Phenylpropanoid glycosides			
Verbascoside	EtOAc–CH_3CN–H_2O	13:13:24	Lower
Acteoside	EtOAc–H_2O	1:1	Lower
Hydrolyzable tannins			
Ellagic acid and corilagin	1-BuOH–HOAc–H_2O	4:1:5	Lower
Catechins			
Epigallocatechin	*n*-Hexane–EtOAc–H_2O	1:9:10	Lower
Tea catechins	*n*-Hexane–EtOAc–MeOH–H_2O	3:10:3:10	Lower
Proanthocyanidins			
Tea proanthocyanidins	*n*-Hexane–EtOAc–MeOH–H_2O	1:5:1:5	Lower
Anthocyanins	MTBE–1-BuOH–CH_3CN–H_2O	2:2:1:5	Lower
Bilberry anthocyanins	MTBE–1-BuOH–CH_3CN–H_2O–TFA	1:4:1:5:0.01	Lower
Flavonoids			
Flavones and chalcones	$CHCl_3$–CH_2Cl_2–MeOH–H_2O	2:2:3:2	Lower
	CH_2Cl_2–MeOH–CH_3CN–H_2O	40:11:25:20	Lower
	EtOAc–1-PrOH–H_2O	140:8:80	Upper
Flavonoid glycosides	*n*-Hexane–EtOAc–MeOH–H_2O	1:8:1:8	Lower
	n-Hexane–EtOAc–MeOH–H_2O	2:12:3:15	Lower
	$CHCl_3$–MeOH–H_2O	8:7:4	Lower
C-Glycosyl flavones	$CHCl_3$–MeOH–H_2O	6:25:29	Lower
	EtOAc–1-BuOH–MeOH–H_2O	35:10:11:44	Lower
Icariin	*n*-Hexane–1-BuOH–MeOH–H_2O	1:4:2:6	Lower
Isoflavone glycosides	EtOAc–1-BuOH–EtOH–H_2O	30:10:6:50	Lower
	EtOAc–EtOH–HOAc–H_2O	16:4:1:20	Lower
Flavonol glycosides	1-BuOH–H_2O	1:1	Lower
Flavonolignans			

Sample	Solvent system		Mobile phase
	Solvents	Ratio	
Silymarin	*n*-Hexane–EtOAc–MeOH–H_2O	1:4:3:4	Lower
Lignans			
Arctiin	EtOAc–1-BuOH–EtOH–H_2O	10:1:2:10	Lower
Lignan diglucoside	MTBE–1-BuOH–CH_3CN–H_2O	1:3:1:5	Upper
Coumarins			
Psoralen and isopsoralen	*n*-Hexane–EtOAc–MeOH–H_2O	10:7:10:8	Lower
Inflacoumarin A	*n*-Hexane–$CHCl_3$–MeOH–H_2O	5:6:3:2	Lower
Chalcones			
Licochalcone A	*n*-Hexane–$CHCl_3$–MeOH–H_2O	3:12:6:4	Lower
Quinones			
Tanshinones	*n*-Hexane–EtOH–H_2O	(a) 20:11:9 (b) 10:7:3	Lower
Shikonin	*n*-Hexane–EtOAc–EtOH–H_2O	16:14:14:5	Lower
Diterpene quinone	*n*-Hexane–CCl_4–MeOH–H_2O	1:3:3:2	Lower
Monoterpenes			
Monoterpenes and monoterpene glycosides	$CHCl_3$–MeOH–H_2O	7:13:8	Upper
Iridoid glycosides	$CHCl_3$–MeOH–1-PrOH–H_2O	5:6:1:4	Lower
Sesquiterpenes			
Artemisinine, artemisitene arteannuin B	*n*-Hexane–EtOAc–EtOH–H_2O	6:4:5:4	Lower
Diterpenes			
10-Deacetylbaccatin III	*n*-Hexane–EtOAc–EtOH–H_2O	2:5:2:5	Lower
	n-hexane–$CHCl_3$–MeOH–H_2O	5:25:34:20	Lower
Trachylobanes and isopimaranes	*n*-Heptane–CH_3CN–CH_2Cl_2	10:7:3	Lower
Triterpenes			
Celastrol	*n*-Hexane–EtOAc–CCl_4–MeOH–H_2O	1:1:8:6:1	Lower
Saponins			
Ginsenosides	$CHCl_3$–MeOH–2-BuOH–H_2O	5:6:1:4	Lower
	EtOAc–1-BuOH–H_2O	1:1:2	Upper

Sample	**Solvent system**		**Mobile phase**
	Solvents	**Ratio**	
	n-Hexane–1-BuOH–H_2O	3:4:7	Lower
Phytolacca saponins	$CHCl_3$–MeOH–2-PrOH–H_2O	5:6:1:4	Lower
Glycyrrhizin	EtOAc–MeOH–H_2O	5:2:5	Lower
Carotenoids			
Zeaxanthin	*n*-Hexane–EtOAc–EtOH–H_2O	8:2:7:3	Lower
Lycopene	*n*-Hexane–CH_3CN–CH_2Cl_2	20:13:7	Upper
Alkaloids			
Aporphine alkaloids	CH_2Cl_2–MeOH–5% HOAc	5:5:3	Lower
Naphthylisoquinoline alkaloids	$CHCl_3$–EtOAc–MeOH–0.1 M HCl	5:3:5:3	Lower
Diterpene alkaloids	*n*-Hexane–CH_2Cl_2–MeOH–H_2O	15:15:24:8	Lower
	C_6H_6–$CHCl_3$–MeOH–H_2O	5:5:7:2	Lower
Tetramethylpyrazines			
Chuanxiongzine	*n*-Hexane–EtOAc–EtOH–H_2O	5:5:3:7	Lower
Glucosinolates	PrOH–CH_3CN–$(NH_4)_2SO_4$–H_2O	10:5:12:10	Upper
Cyclodepsipeptides	*n*-Heptane–EtOAc–MeOH–H_2O	2:8:2:8	Lower

13.5.1.2 Retention of the Stationary Phase

- Successful separation in HSCCC-CPC largely depends on the amount of the stationary phase retained in the column. In general, higher the retention of stationary phase, better the peak resolution.
- The amount of stationary phase retained in the column is highly correlated with the settling time of the two phases in a test tube.
- Measure the settling time of the two-phase solvent system to be used for the separation.
- The procedure is as follows: the two phases are first equilibrated in a separatory funnel. Deliver 2 ml of each phase, a total volume of 4 ml, into a test tube or graduated cylinder, which is then capped. Gently invert the container for several times and then immediately place it in an upright position to measure the time required for the two phases to form clear layers with a distinct interface.
- If the settling time is less than 20 seconds, the solvent system will provide satisfactory retention of the stationary phase, usually over 50% of the total column capacity, in a proper range of flow rates.

13.5.1.3 Preparation of Sample Solution

- The sample for HSCCC-CPC may be dissolved directly in the stationary phase or in a mixture of the two phases.
- The recommended sample volume is less than 5% of the total column capacity.
- Introduction of a larger sample volume into the column will reduce peak resolution of the analytes, especially for those having small *K* values.
- Ideally the analyte is injected in a small volume of the stationary phase to preserve the sharpness of the early elution peak.

13.5.1.4 Separation Column

- We must understand the head–tail orientation of the separation column.
- A lower (heavier) mobile phase should be introduced through the head toward the tail, and an upper (lighter) mobile phase in the opposite direction.
- This is extremely important because the elution of either phase in the wrong direction results in an almost complete loss of the stationary phase from the column.

13.5.1.5 Choice of the Mobile Phase

- In HSCCC-CPC, either phase can be used as the mobile phase provided that the *K* value of the analyte is in a proper range.
- If one has a choice, the lower phase may be used as the mobile phase, because the system provides more stable retention of the stationary phase and one can avoid trapping air bubbles in the flow cell of the detector by introducing the effluent from the lower end of the cell.
- On the other hand, when the upper organic mobile phase (excluding the chloroform system) is used as the mobile phase, it will facilitate the evaporation of solvent from the collected fractions.

13.5.1.6 Flow Rate of the Mobile Phase

- The flow rate of the mobile phase determines the separation time, the amount of stationary phase retained in the column, and therefore the *peak resolution*.
- A lower flow rate usually gives higher retention level of the stationary phase and improves the peak resolution, although it requires a longer separation time.
- The typical flow rates for the commercial multilayer coil are as follows:
  i) 5–6 ml/min for a preparative column with 2.6 mm i.d. PTFE tubing (600–800 rpm) (up to 1 g sample load);

  - ii) 2–3 ml/min for a semi preparative column with 1.6 mm i.d. PTFE tubing (800–1000 rpm) (up to 500 mg sample load);
  - iii) 1 ml/min for an analytical column with 0.85–1.0 mm i.d. PTFE tubing (1000–1200 rpm) (up to 50 mg sample load).
- These flow rates should be modified according to the settling time of the two-phase solvent system as well as other factors. When the settling time is around 20 s and the *K* value of the analyte is small, a lower flow rate is recommended.

13.5.1.7 Revolution Speed

The optimum revolution speed (revolution and planetary rotation speeds are always same) for the commercial HSCCC-CPC instrument for preparative separation ranges between 600 and 1400 rpm.

- Use of a lower speed will reduce the volume of the stationary phase retained in the column, leading to lower peak resolution.
- On the other hand, the higher speeds may produce excessive sample band broadening by violent pulsation of the column because of elevated pressure.

13.5.1.8 Filling the Column with the Stationary Phase

- In each separation, the column is first entirely filled with the stationary phase.
- Before introducing the stationary phase, the column may be flushed with a column volume of a solvent miscible with the two phases used in the previous run (e.g. ethanol or methanol) to wash out materials remaining in the column.
- This will also ensure a stable, clean baseline before the solvent front emerges.
- Avoid trapping the air in the column, especially in a preparative column. This can be easily tested as:
  - i) If no air is present in the column, the flow from the column outlet is ceased shortly after stopping the pumping.
  - ii) If the solvent keeps flowing from the outlet for more than several seconds, the air trapped in the column should be eliminated by resuming the pumping of the stationary phase under low speed column rotation (100–200 rpm) in a tail to head elution mode to accelerate air movement toward the outlet of the column.

13.5.1.9 Sample Loading

There are two ways to load samples; both are satisfactory for HSCCC-CPC separations:

- In the first method, the column is entirely filled with stationary phase and this is immediately followed by sample injection.

- The mobile phase is then eluted through the column while the column is rotated at the optimum rate.
- In the second method, after the column is filled with stationary phase, the mobile phase is eluted through the column at a desired rate until the solvent front emerges and hydrodynamic equilibrium is established throughout the column as evidenced by diminished carryover of the stationary phase.
- The sample is then injected into the column through the sample port.

Each method has advantages. The second method gives a clear tracing of the elution curve because of the minimum carryover of the stationary phase from the column.

The first method produces a distinct solvent front and saves separation time by eliminating the waiting period to reach hydrodynamic equilibrium.

One can conveniently use an injection valve with a sample loop as in HPLC.

13.5.1.10 On-line Monitoring of Effluent

- The effluent from the outlet of the HSCCC-CPC columns may be continuously monitored by a UV–VIS detector as in conventional liquid chromatography.
- An important difference between these two methods is that HSCCC-CPC uses the liquid stationary phase which, if carried over from the column, tends to disturb the tracing of the elution curve.
- Avoid trapping the stationary phase in the vertical flow cell by eluting the lower mobile phase upward from the bottom; do the reverse if the upper is used as the mobile phase.
- When the upper mobile phase is eluted from the top of the flow cell downward, it is important to prevent the formation of air bubbles which may become trapped in the flow cell and disturb the tracing of the elution curve.
- Bubble formation can be largely avoided by degassing the two phases in the separatory funnel before use, and also by connecting fine PTFE tubing (typically 30 cm × 0.4–0.5 mm i.d.) to the outlet of the monitor so that the pressure within the flow cell is substantially increased.

13.5.1.11 Measurement of Stationary Phase Retention

- When the separation is completed, rotation is stopped and the column contents are collected into a graduated cylinder by connecting the column inlet to a nitrogen cylinder (ca. 50 psi; 1 psi = 6894.76 Pa).

- After nitrogen appears at the outlet, the column is slowly rotated (100 rpm) in the tail to head elution mode so that solvent remaining inside the column is pumped out by an Archimedean screw force assisted by the nitrogen flow. Measure the retained stationary phase in a graduated cylinder.
- Measurement of the retained stationary phase is useful in efforts to improve the separation: when the peak resolution is unsatisfactory, a measure of stationary phase retention will serve as a guide for the next trial.
- If it is less than 30%, the separation may be improved by increasing retention, by applying a lower flow rate of the mobile phase, increasing the revolution speed, or modifying the solvent system to shorten the settling time.
- If instead the stationary phase retention is over 50%, efforts should be directed to search for a new two-phase solvent system, which provides an improved separation factor (α) between the analytes.

13.5.2 Applications of HSCCC-CPC Technologies in Natural Products Isolation

HSCCC-CPC technologies have applications in the following industries: nutraceuticals, fine chemicals, pharmaceuticals, biomedical, biotechnology, fats and oils, and fermentation.

Compounds that can be isolated in high purity by HSCCC-CPC technologies include: saponins, alkaloids, chlorophylls, tannins, carotenoids, phospholipids, fat-soluble vitamins, mono- and oligosaccharides, anthocyanins, lignans, phenolic compounds, synthetic compounds, other active compounds present in medicinal and aromatic plants (e.g. herbs and spices) and much more. The numerous applications of CCC have resulted in a growth in the annual number of publications in which this separation technology has been cited (Figure 10).

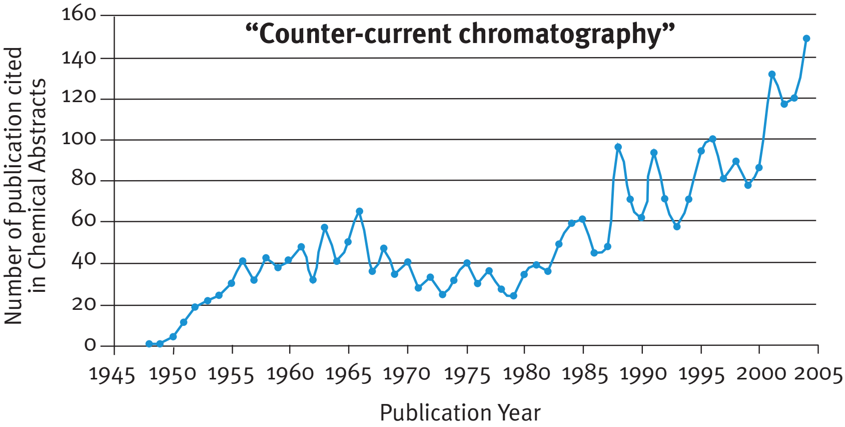

Figure 10: Increasing number of publications on separation by CCC

Examples of some important natural products that have been separated using HSCCC-CPC in the recent past are summarized herein.

13.5.2.1 Purification of Coenzyme Q_{10} from Fermentation Extract: HSCCC versus Silica Gel Column Chromatography

HSCCC was applied to the purification of coenzyme Q_{10} (CoQ_{10}) for the first time. CoQ_{10} was obtained from a fermentation broth extract. A non-aqueous two-phase solvent system composed of heptane–acetonitrile–dichloromethane (12:7:3.5, v/v/v) was selected by analytical HSCCC and used for purification of CoQ_{10} from 500 mg crude extract. The separation yielded 130 mg CoQ_{10} at an HPLC purity of over 99%. The results showed the advantages of HSCCC over an alternative of silica gel chromatography followed by recrystallization. These advantages regard purity, recovery and yield (Table 3).

Table 3: Purification of coenzyme Q_{10} from fermentation extract: HSCCC vs. silica gel column chromatography with subsequent crystallization

	Crude extract	**CoQ_{10} purified by silica gel chromatography**	**CoQ_{10} purified by HSCCC**
HPLC purity, %	89.2	96.0	99.2
Absolute purity, %	29.4	93.3	97.8
Recovery, %*	–	74.3	88.0
Yield, %*	–	23.4	26.4

** Recovery, amount of CoQ_{10} in purified product/amount of CoQ_{10} in crude extract. Yield,amount of purified product/amount of crude extract (Journal of Chromatography A 1127 (1-2), 15 September 2006, 92-96)*

13.5.2.2 Preparative Separation of Gambogic Acid and its C-2 Epimer by HPCCC

For the preparative separation of epimers, gambogic acid and epigambogic acid, from *Garcinia hanburyi,* a two-phase solvent system composed of *n*-hexane–methanol–water (5:4:1, v/v/v) was used. From 50 mg mixture, 28.2 mg gambogic acid and 18.4 mg epigambogic acid were separated. The purities of both were above 97% as determined by HPLC. The chemical structures were then identified by their ^1H NMR and ^{13}C NMR spectra.

13.5.2.3 Separation and Purification of 10-deacetylbaccatin III by HSCCC

At present, the most promising approach is the semisynthesis of paclitaxel or its analogs from 10-deacetylbaccatin III, a compound available in a relatively high quantity from the foliage of several yew species. HSCCC was used for the separation and purification of 10-deacetylbaccatin III (Figure 11). A crude needle extract (500 mg/5 ml) from Chinese yew (*Taxus chinensis*) was first separated with a two-phase solvent system composed of *n*-hexane–ethyl acetate–ethanol–water (2:5:2:5, v/v). The partially purified fraction was again purified with a different solvent system composed of *n*-hexane–chloroform–methanol–water (5:25:34:20, v/v). HPLC analysis of the final fraction showed that the purity of 10-deacetylbaccatin (20 mg) was over 98%.

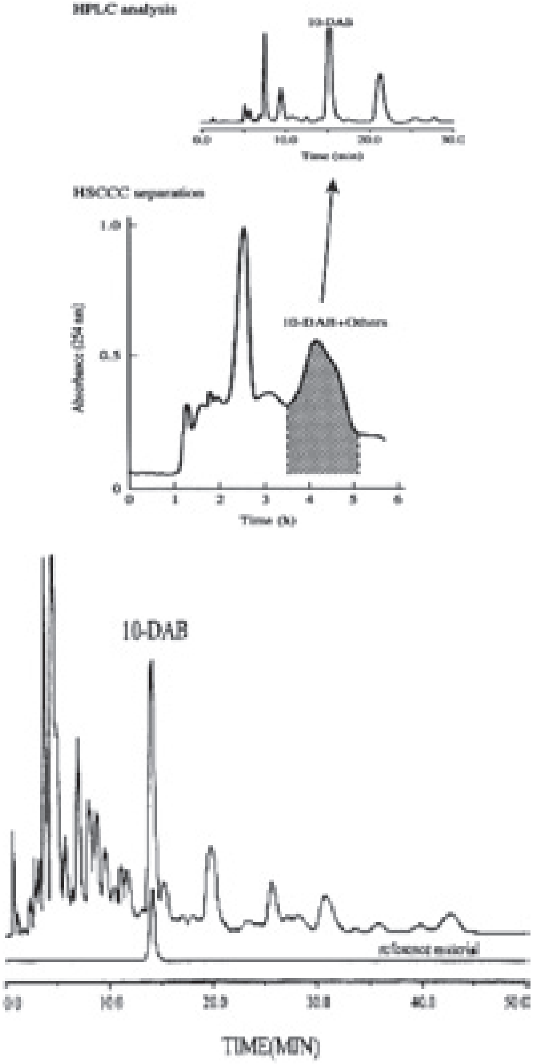

Figure 11: Separation and purification of 10-deacetylbaccatin III (DAB) by HSCCC

13.5.2.4 Large-scale Separation of Resveratrol and Anthraglycoside A and B from *Polygonum cuspidatum* by HSCCC

HSCCC was successfully applied to the large-scale (5 g) separation of resveratrol, anthraglycoside A and anthraglycoside B from a crude extract of *Polygonum cuspidatum* Sieb. et Zucc (Figure 12). A two-phase solvent system composed of chloroform, methanol and water (4:3:2, v/v) was used. The separation yielded 200 mg to 1 g of these three compounds, each at over 98% purity as determined by HPLC. Resveratrol is important as it has a cancer chemopreventive activity.

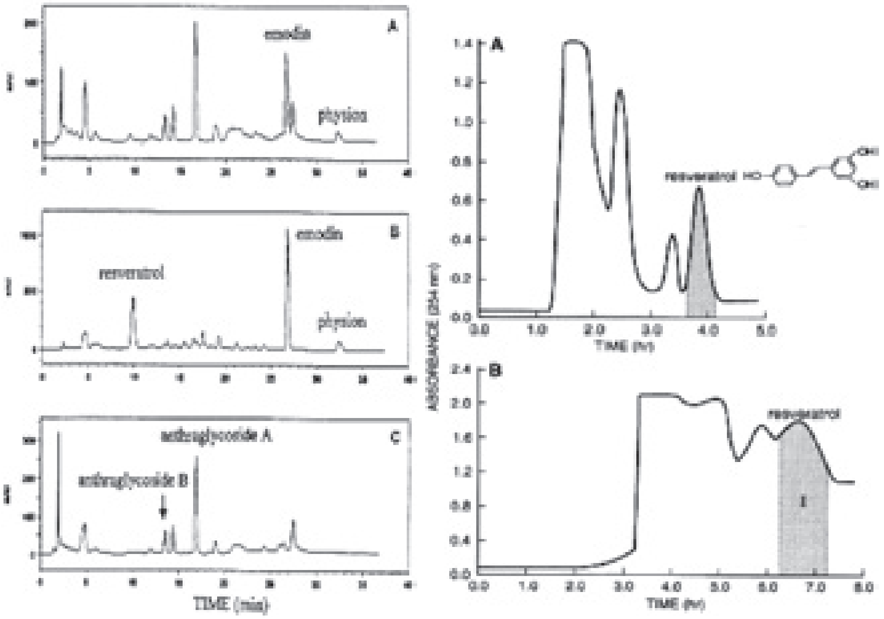

Figure 12: Separation of resveratrol, anthraglycoside A and anthraglycoside B from a crude extract of Polygonum cuspidatum Sieb. et Zucc

13.5.2.5 Separation of Andrographolide and Neoandrographolide from the Leaves of *Andrographis paniculata* using HSCCC

The bioactive diterpenes andrographolide and neoandrographolide from the leaves of *Andrographis paniculata* NEES (Acanthaceae) were successfully separated by CCC (Figure 13). A single 280-min separation yielded 189 mg of 99.9% andrographolide and 9.5 mg of 98.5% neoandrographolide. Water–methanol–ethyl acetate–*n*-hexane (2.5:2.5:4:1) solvent system was used.

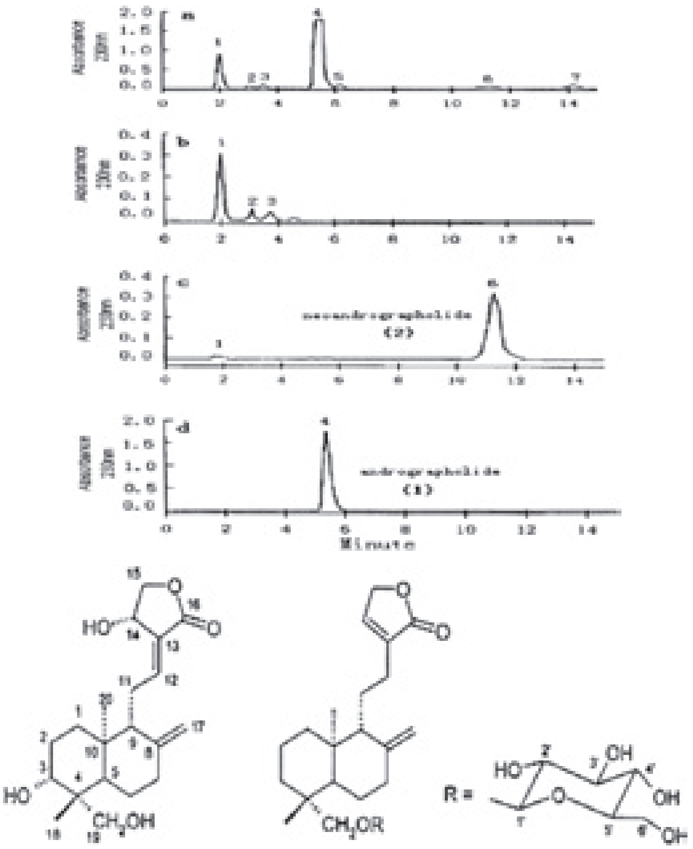

Figure 13: Separation of Andrographolide and Neoandrographolide from the leaves of Andrographis paniculata using HSCCC

13.5.2.6 Separation of WAP-8294A Components, a Novel Anti-methicillin-resistant *Staphylococcus aureus* Antibiotic, using HSCCC

The WAP-8294A complex was isolated from the fermentation broth of *Lysobacter* sp. WAP-8294 (Figure 14). The major component, WAP-8294A2, shows strong activity against gram-positive bacteria including methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococci* in vitro, and also exhibited a potent activity against MRSA in vivo.

The previous separations were unsatisfactory. Hence HSCCC was applied. *n*-Butanol–ethyl acetate–aqueous 0.005 *M* trifluoroacetic acid (1.25:3.75:5, v/v/v) was used as biphasic solvent. A sample size of 25 mg yielded pure fractions of three components (1–6 mg). The method will contribute to the clinical development of WAP-8294A2 as an anti-MRSA agent.

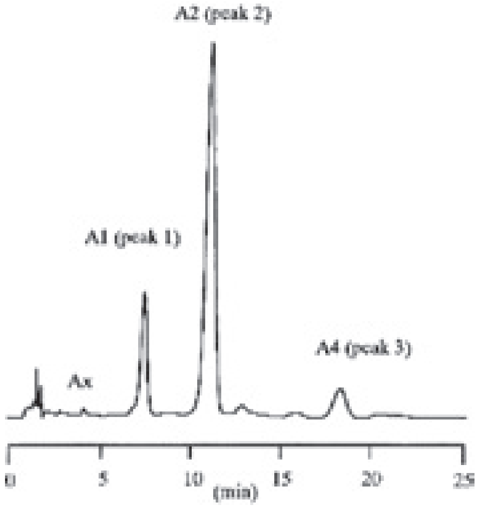

Figure 14: Separation of WAP-8294A components using HSCCC

13.5.2.7 Other Examples of Separation of Phytoconstituents by CCC

Apart from the previously mentioned examples, isolation and purification of polymethoxylated flavones from tangerine peel, catechin constituents from five tea cultivars, rupestonic acid from the Chinese medicinal plant *Artemisia rupestris* L., lycopene from tomato paste, spiramycin, gallic acid from *Cornus officinalis*, lutein from the microalga *Chlorella vulgaris*, naphthopyranone glycosides, salvianolic acid B from the Chinese medicinal plant *Salvia miltiorrhiza*, dammarane saponins from *Panax notoginseng*, isoflavan and pterocarpan glycosides from *Astragalus membranaceus* Bge. var. *mongholicus* (Bge.), glycyrrhizin from the root of liquorice and active principles from the roots of *Sophora flavescens* have been carried out successfully using HSCCC.

13.6 Advantages of CCC

- Quick (high throughput in preparative separation).
- Inexpensive (only solvent costs, which are 5 times less than for other LC techniques).
- Gentle and versatile, for separation of varied compounds, with less chance of decomposition.
- Able to resolve from milligrams to tens of grams on the same instrument.
- Able to switch between normal and reverse phase at will.
- A CCC machine, which is a chromatographic column with a liquid stationary phase, can be used as a liquid–liquid reactor for chemical reactions involving a liquid catalyst.

- No irreversible adsorption to a solid support (100% recovery of sample).
- Increased capacity for the same volume of stationary phase; a CPC column gives a higher capacity than the HPLC one.
- Quantity of sample depends on two factors: solubility of the sample and properties of the solvent system.
- No sample loss as a simple rinsing of the instrument allows a full recovery of the noneluted fractions.
- Easy maintenance, no costly solid phase to change.
- No degradation or denaturation of compounds and no interaction with silica.
- No polarity restriction; all biphasic mixtures can be used.
- Dual mode (off-line and on-line) exchange of stationary and mobile phase (CPC).

13.6.1 Advantages of HSCCC-CPC Technologies over HPLC

HSCCC-CPC	**HPLC**
No column	Expensive columns
High recovery	Irreversible adsorption
High throughput	Poor loadability
Retention of fragile compounds (molecular integrity)	Loss of biological activity (denaturation)
Volume ratio of stationary/ mobile very high (better resolution)	Ratio is low

13.7 Manufacturers of CCC Instruments

13.7.1 Manufacturers of HSCCC Machines

- AECS (http://www.ccc4labprep.com/)
- Conway Centri Chrom (http://www.centrichrom.com/)
- Dynamic Extractions (http://www.dynamicextractions.com/)
- Pharma-tech Research Corporation (http://www.pharma-tech.com/)

13.7.2 Manufacturers of CPC Machines

- EverSeiko Corporation (http://www.everseiko.co.jp)
- Kromaton Technologies (http://www.kromaton.com/)
- Partus Technologies (http://www.partus-technologies.com)

13.8 Selected Reviews on CCC

- Marston and Hostettmann, K., 1994, Counter-current Chromatography as a Preparative Tool, and Perspectives, *J. Chromatog. A; 658:* 315-341
- Jarvis, 1992, Macrocyclic Trichothecenes from Brazilian Baccarin Species: From Microanalysis to Large-scale Isolation, *Phytochemical Analysis;* 3: 241-249
- Foucault, 1991, Counter-current Chromatography, *Analytical Chemistry;* 63(10): 569-579
- Marston, I., Slacanin and Hostettmann, K., 1990, Centrifugal Partition Chromatography in the Separation of Natural Products, *Phytochemical Analysis;* 1: 3-17
- Sutherland, 1987, Counter-current Chromatography, *Laboratory Practice*
- Martin, D. G., Biles, C. and Peltonen, R. E., 1986, Counter-current Chromatography in the Fractionation of Natural Products, *American Laboratory*
- Ito, Y., 1985, High-speed Counter-current Chromatography, *Critical Reviews in Analytical Chemistry;* 17(1): 65-143
- Conway, W. D. and Ito, Y., 1984, Recent Applications of Counter-current Chromatography, *LC*; 2(5)

13.9 Conclusions

Counter-current Chromatography is an excellent alternative to circumvent the problems associated with solid-phase adsorbents and to preserve the chemical integrity of mixtures subjected to fractionation. It provides efficient resolution of samples by a mechanism which relies solely on partition. Separation of compounds with a wide range of polarities can be achieved with the use of aqueous and non-aqueous solvent systems. CCC is very flexible: solvent gradients are possible; flow rates can be varied during a chromatographic run; lower and upper phases can be interchanged as mobile phases during a separation, provided that the flow direction is also changed accordingly; and instruments can be stopped during chromatography and re-started hours later without affecting separation efficiency. A wide range of pH is tolerated in CCC, with implications in the separation of acidic and basic samples, notably in the technique of pH-zone-refining. Separation of crude plant extracts, semipurified fractions or synthetic mixtures can be carried out, with samples of any quantity ranging from 100 mg to 1500 g (for the largest model). With these advantages, CCC is gaining popularity as a separation method for natural products, and especially in the bioassay-guided fractionation of natural products.

Bibliography

Becker, H., Reichling, J. and Hsieh, W., 1982, Water-free solvent system for droplet counter-current chromatography and its suitability for the separation of non-polar substances, *Journal of Chromatography A,* 237(2): 307-310

Campaner dos Santos, L. and Vilegas, W., 2001, Preparative separation of the naphthopyranone glycosides by high-speed counter-current chromatography, *Journal of Chromatography A,* 915(1-2): 259-263

Cao, X. L., Xu, Y. T., Zhang, G. M., Xie, S. M., Dong, Y. M. and Ito, Y., 2006, Purification of coenzyme Q_{10} from fermentation extract: High-speed counter-current chromatography versus silica gel column chromatography, *Journal of Chromatography A,* 1127(1-2): 92-96

Cao, X., Tian, Y., Zhang, T. Y. and Ito, Y., 1998, Separation and purification of 10-deacetylbaccatin III by high-speed counter-current chromatography, *Journal of Chromatography A,* 813(2): 397-401

Chen, X., Yi, C., Yang, X. and Wang, X., 2004, Liquid chromatography of active principles in *Sophora flavescens* root, *Journal of Chromatography B,* 812(1-2): 149-163

Desai, H. K., Joshi, B. S., Ross, S. A. and Pelletier, S., 1989, Methanolysis of the C-8 acetoxyl group in aconitine-type alkaloids: a partial synthesis of hokbusine A, *Journal of Natural Products*, 52: 720-725

Domon, B., Hostettmann, M. and Hostettmann, K., 1982, Droplet counter-current chromatography with non-aqueous solvent systems, *Journal of Chromatography A,* 246(1): 133-135

Du, Q., Jerz, G., Waibel, R. and Winterhalter, P., 2003, Isolation of dammarane saponins from *Panax notoginseng* by high-speed counter-current chromatography, *Journal of Chromatography A,* 1008(2): 173-180

Francis, G. W. and Andersen, Ø. M., 1984, Droplet counter-current chromatography of anthocyanins, *Journal of Chromatography A,* 283: 445-448

Gago, F., Alvarez-Builla, J., Elguero, J., 1986, Use of droplet counter-current chromatography in log *P* determinations, *Journal of Chromatography A,* 360: 247-251

Glinski, J. A. and Caviness, G. O., 1995, Partition chromatography in assay-guided isolation of natural products: a case study of immunosuppressive components of *Tripterygium wilfordii*. In: Foucault, A. P. (Ed.), Centrifugal Partition Chromatography, Marcel Dekker, New York, p. 133-136

Han, Q. B., Song, J. Z., Qiao, C. F., Wong, L. and Xu, H. X., 2006, Preparative separation of gambogic acid and its C-2 epimer using recycling high-speed counter-current chromatography, *Journal of Chromatography A,* 1127(1-2): 298-301

Hanke, F. J. and Kubo, I., 1985, Increasing the speed of droplet counter-current chromatography separations, *Journal of Chromatography A,* 329: 395-398

Harad, K. I., Suzuki, M., Kato, A., Fujii, K., Oka, H. and Ito, Y., 2001, Separation of WAP-8294A components, a novel anti-methicillin-resistant *Staphylococcus aureus* antibiotic, using high-speed counter-current chromatography, *Journal of Chromatography A,* 932(1-2): 75-81

Heijden, R., Van Der, Hermans-Lokkerbol, A., Verpoorte, R. and Svendsen A. B., 1987, Pharmacognostical studies of *Tabernaemontana* species: XX. Ion-pair droplet counter-current chromatography of indole alkaloids from suspension cultures, *Journal of Chromatography A,* 396: 410-415

Jiang, Y., Lu, H. T. and Chen, F., 2004, Preparative purification of glycyrrhizin extracted from the root of liquorice using high-speed counter-current chromatography, *Journal of Chromatography,* 1033(1): 183-186

Kéry, Á. Turiák, Gy. and Tétényi, P., 1988, Isolation of parthenolide by droplet counter-current chromatography, *Journal of Chromatography A,* 446: 157-161

Kinghorn, A. D., Nanayakkara, N. P. D., Soejarto, D. D., Medon, S. and Kamath, P. J., 1982, Potential sweetening agents of plant origin: I. Purification of *Stevia rebaudiana* sweet constituents by droplet counter-current chromatography, *Journal of Chromatography A,* 237(3): 478-483

Kubo, I., Klocke, J. A., Ganjian, I., Ichikawa, N. and Matsumoto, T., 1983, Efficient isolation of phytoecdysones from *Ajuga* plants by high-performance liquid chromatography and droplet counter-current chromatography, *Journal of Chromatography A,* 257: 157-161

Kubo, I., Matsumoto, A. and Asano, S. 1985, Efficient isolation of ecdysteroids from the silkworm, *Bombyx mori* by droplet counter-current chromatography, *Insect Biochemistry,* 15(1): 45-47

Kubo, I., Matsumoto, A., Hanke, F. J. and Ayafor, J. F., 1985, Analytical droplet counter-current chromatography isolation of 20-hydroxyecdysone from V*itex thyrsiflora* (Verbenaceae), *Journal of Chromatography A,* 321: 246-248

Kumar, N., Savitri, and Maheshinie, R., 2005, Separation of catechin constituents from five tea cultivars using high-speed counter-current chromatography, *Journal of Chromatography A,* 1083(1-2): 223-228

Kurumaya, K., Sakamoto, T., Okada, Y., Kajiwara, M., 1988, Application of droplet counter-current chromatography to the isolation of vitamin B_{12}, *Journal of Chromatography A,* 435: 235-240

Li, H. B., Chen, F., Zhang, T. Y., Yang, F. Q. and Xu, G. Q., 2001, Preparative isolation and purification of lutein from the microalga *Chlorella vulgaris* by high-speed counter-current chromatography, *Journal of Chromatography A,* 905(1-2): 151-155

Li, H. B., Lai, J. P., Jiang, Y. and Chen, F. 2002, Preparative isolation and purification of salvianolic acid B from the Chinese medicinal plant *Salvia miltiorrhiza* by high-speed counter-current chromatography, *Journal of Chromatography A,* 943(2): 235-239

Ma, X., Tu, P., Chen, Y., Zhang, T., Wei, Y. and Ito, Y., 2004, Preparative isolation and purification of isoflavan and pterocarpan glycosides from *Astragalus membranaceus* Bge. var. *mongholicus* (Bge.) Hsiao by high-speed counter-current chromatography, *Journal of Chromatography A,* 1023(2): 311-315

Ma, Y., Aisha, H. A., Liao, M. L., Aibai, S., Zhang, T., Ito, Y., 2005, Preparative isolation and purification of rupestonic acid from the Chinese medicinal plant *Artemisia rupestris* L. by high-speed counter-current chromatography, *Journal of Chromatography A,* 1076(1-2): 198-201

Marshall, T. and Kinghorn, A. D., 1981, Isolation of phorbol and 4α-phorbol and croton oil by droplet counter-current chromatography, *Journal of Chromatography A,* 206(2): 421-424

Miething, H. and Seger, V., 1989, Separation of non-polar compounds by droplet counter-current chromatography, *Journal of Chromatography,* 478: 433-437

Oka, H., Harada, K., Ito, Y. and Ito, Y., 1998, Separation of antibiotics by counter-current chromatography, *Journal of Chromatography A, 812(1-2)*: 35-52

Oka, H., Harada, K., Suzuki, M. and Ito Y., 2000, Separation of spiramycin components using high-speed counter-current chromatography, *Journal of Chromatography A,* 903(1-2): 93-98

Orsini, F. and Verotta, L. 1985, Separation of natural polar substances by reversed-phase high-performance liquid chromatography, centrifugal thin-layer chromatography and droplet counter-current chromatography, *Journal of Chromatography A,* 349(1): 69-75

Oya, S. and Snyder, J. K. 1986, Chiral resolution of a carboxylic acid using droplet counter-current chromatography, *Journal of Chromatography A,* 370: 333-338

Pelletier, S. W., Desai, H. K. Jiang, Q. and Ross, S. A., 1990, An unusual epimerization of the α-OH-1 group of the norditerpenoid alkaloid delphisine, *Phytochemistry,* 29: 3649-3652

Pelletier, S. W., Venkov, A. P., Finer-Moore, J. and Mody, N. V., 1980, An alumina catalyzed addition of secondary amines to exocyclic α,β-unsaturated ketones, *Tetrahedron Letters,* 21: 809-812

Qizhen, D., Jerz, G. and Winterhalter, P., 2003, Separation of andrographolide and neoandrographolide from the leaves of *Andrographis paniculata* using high-speed counter-current chromatography, *Journal of Chromatography A,* 984(1): 147-151

Saito, K., Horie, M., Hoshino, Y., Nose, N., Shida, Y., Nakazawa, H., Fujita, M., 1988, Isolation of virginiamycin-M_1 by droplet counter-current chromatography, *Journal of Chromatography A,* 45: 387-391

Takeuchi, T., Horikawa, R. and Tanimura, T., 1984, Complete resolution of DL-isoleucine by droplet counter-current chromatography, *Journal of Chromatography A,* 284: 285-288

Tian, G., Zhang, T., Yang, F. and Ito, Y., 2000, Separation of gallic acid from *Cornus officinalis* Sieb. et Zucc by high-speed counter-current chromatography, *Journal of Chromatography A,* 886(1-2): 309-312

Wang, X., Li, F., Zhang, H., Geng, Y., Yuan, J. and Jiang, T., 2005, Preparative isolation and purification of polymethoxylated flavones from tangerine peel using high-speed counter-current chromatography, *Journal of Chromatography A,* 1090(1-2): 188-192

Wei, Y., Zhang, T., Xu, G. and Ito, Y., 2001, Application of analytical and preparative high-speed counter-current chromatography for separation of lycopene from crude extract of tomato paste, *Journal of Chromatography A,* 929(1-2): 169-173

Yang, F., Zhang, T. and Ito, Y., 2001, Large-scale separation of resveratrol, anthraglycoside A and anthraglycoside B from *Polygonum cuspidatum* Sieb. et Zucc by high-speed counter-current chromatography, *Journal of Chromatography A,* 919(2): 443-448

14 Quality Control of Medicinal and Aromatic Plants and their Extracted Products by HPLC and High Performance Thin Layer Chromatography

K. Vasisht

Abstract

The interest in medicinal plants and their products has increased manifold in recent years. The increasing public demand for natural medicines has resulted in increased commercial activity and production of these medicines. There is also increasing concern for ensuring quality and safety of plant medicines. Plant drugs, unlike active pharmaceutical ingredients, possess some inherent limitations which deter the process of laying standards for these drugs. This aspect has received considerable attention from different quarters including policy planners, scientists and manufacturers. This paper describes briefly the roles of high performance thin layer chromatography and high performance liquid chromatography in quality assurance of plant products. Some practical aspects of these techniques are also discussed.

14.1 Introduction

The use of medicinal plants products has increased several fold during the last decades. Individual countries are also giving increasing emphasis to promote their use under the direction of the World Health Organization (WHO). Besides this, one finds enormous interest in natural products from the public, which is attributable to several factors. These medicines are affordable, safer and better tolerated by the biological system. This has led to an increased consumption and cross-country movement of raw materials of medicinal plants.

In some parts of the world, e.g. several places in Africa and Asia, traditional medicines are the only affordable option. On the other hand, the same medicines are the option of choice in developed nations like Japan and the United States and in the European States. Despite being the more common medical option in Africa, use of traditional medicines has not matured to the expected level. But, some countries in Asia, especially India and China, have developed them to a level that has benefited all countries of the world.

Europe did not inherit a well-developed traditional system of medicine but it has put in place the strongest evidence-based and scientifically supported system of plant medicines. North America is the most prolific and flamboyant market for plant products, but government regulations are such that except for select plant drugs most are used as dietary supplements or nutraceuticals. South America, which has given the world some of the all-time great plant medicines like quinine and pilocarpine, follows the medical system of North America. Australia has a strong impact from Chinese medicines on the continent which is otherwise dominated by the modern system of medicine.

14.2 Quality Control of Medicinal Plants and their Products

The quality control of consumer products has become more challenging and demanding. The quality considerations of drugs are the most stringent among all consumer products. The purity of active pharmaceutical ingredients has been stretched to an all-time high with more and more restrictions on the level of the impurities. The situation is opposingly different in the case of plant-derived medicine, where we are still striving to define specifications to ensure consistency and safety. Therefore, the standards of plant drugs are more relaxed and are in the process of development. The inherent problems of plant drugs are obvious; unlike single chemical entities of modern drugs, they are combinations of infinite chemical molecules, known and unknown; the knowledge of the active components is incomplete; the natural variations in content and quantity of the chemical constituents are large and exercising a precise control is impractical; and the complete chemical profiling of plant drugs is beyond scope. Therefore, laying standards for such drugs is not an easy task and a comprehensive system of standards cannot be laid down for such drugs. As our knowledge of plant drugs will advance, the standards for them will become more meaningful and complete.

The quality issue of plant drugs was irrelevant in ancient times when these medicines were dispensed by the medical men for their patients. However, the issue has taken front seat with the commercialization of plant drugs. The matter has been further complicated by the vested interest of manufacturers who are out to exploit the loopholes in the standards and laws governing the production and distribution of plant drugs.

Several national and international agencies have prioritized the issue of assuring the quality of plant drugs. The effort of the World Health Organization is outstanding: over 20 years ago it first published *Quality control methods for medicinal plant materials*, which has been regularly updated and followed by a series of monographs on globally important medicinal plants.

The quality of a plant product cannot be assured without assuring the quality of the raw material. Also required to ensure quality products are in-process control, quality control of the finished product, good manufacturing practice (GMP) controls and process validation. In this regard, it is imperative to define specifications of raw materials to minimize variations in the quality of finished products and to achieve consistency. The specifications of plant materials include macro- and microscopic descriptions, tests of identity, and analytical and physicochemical determinations. The expected results of these tests and measurements are presented as numerical limits or as a range or discretely observable result.

While fixing the limits of specifications, naturally met variations in plant drugs need to be accommodated. A plant material conforming to the prescribed specifications should be considered acceptable for intended use. Several factors contribute to variation in the content and composition of raw materials. These factors can broadly be grouped in four categories of climatic, nutritional, collection and post-harvest factors. Climatic factors include prevailing temperature, rainfall, humidity, daylight and altitude of the growing region. The nutritional factors are those which affect the health of a growing plant and are reflected in the production of biomass and its composition; several soil factors such as availability of micro- and macronutrients, pH and cation exchange capacity are important for optimal growth of plants. Collection factors control the content of active components by giving due attention to the age, season, collection time and part of the plant collected. Post-harvest factors are important as the collected material is still live and carries out metabolic processes and respires; the enzymatic processes continue after collection until they are deactivated by drying or other suitable treatment; the crushing and cutting of material leads to de-compartmentalization of reactive chemical constituents of plants which were naturally located in intact cells; and the collected material faces direct impact of oxidation by air and light besides physical loss of some components.

The World Health Organization, in its volume *Quality control methods for medicinal plant materials,* has listed several parameters which are valuable in assuring quality of plant drugs. These include identification, visual inspection, sensory characters, macro- and microscopic characteristics, moisture content, foreign matter, fingerprint by thin layer chromatography (TLC), ash values, extractive values, volatile matter, microbial load, heavy metals and pesticide residues, radioactive contaminants and, according to the nature of the drug, one or more determinations for bitter value, tanning test, foaming, hemolytic and swelling indices. The European Medicines Agency's "Guideline on quality of herbal medicinal products" and "Guideline on specifications" more precisely describe the quality-related issues of medicinal plants. These documents define and differentiate herbal substances (equivalent to herbal drugs), herbal preparations (equivalent to herbal drug preparations) and herbal medicinal products (equivalent to traditional herbal medicinal products). The guidelines emphasize the quantification of active or analytical markers and describe procedures to ensure quality of raw material, semifinished and finished products.

14.3 Biological and Chemical Standardization of Drugs

All test procedures ultimately aim to determine the intrinsic potency of a drug, which is attributable to the chemical constituents present. Evaluating the biological potency of a drug provides a direct assessment of its quality. But, the complexity of the procedures and methods forbids implying this assessment. Moreover, it is not the practical option when one handles large numbers of samples. The other option, besides biological testing, is chemical testing, which uses assay procedures to determine the quantity of chemical compounds, preferably the active ones, to assess the quality of a product. This is complementary to recording other specifications like macro- and microscopic characteristics.

14.3.1 Chemical Standardization and Markers

Chemical standardization requires first to identify and select a chemical constituent of a drug and then to elaborate the assay procedure for quantification of the chosen compound. Selection of the constituent, called marker, is a difficult task based on several considerations: the chemical profile of the drug, biological activity of the chemical constituents, successful development of the assay procedure, ease of procuring or isolating the marker, and stability of the marker.

The European Union Guidelines define markers as “Chemically defined constituents or groups of constituents of a herbal substance, a herbal preparation or a herbal medicinal product which are of interest for control purposes independent of whether they have any therapeutic activity.” The WHO defines markers as “constituents of a medicinal plant material which are chemically defined and of interest for control purposes.”

Markers serve to calculate the quantity of herbal substances or herbal preparations in the herbal medicinal product if the marker has been quantitatively determined in the herbal substance or herbal preparations. The European Medicines Agency differentiates an active marker from an analytical marker: an active marker contributes to the therapeutic activity of the drug, whereas an analytical marker serves only the analytical purposes.

The choice of a marker, in first place, should be for a chemical constituent of the drug responsible for the activity. More than one marker can be employed for control purposes, separately or in combination. A second choice of marker is for a defined chemical constituent that is uniquely associated with the drug. This is true when active constituents of the drug are not known or active constituents are unobtainable or unstable for assay purposes. As a last choice, a marker can be selected among more commonly or ubiquitously present phytoconstituents. In rare cases, a chemical compound not associated with the plant or its activity can also be used as a marker. Such chemical compounds allow the estimation of an active constituent of the plant, which itself is unstable and unfit for the purpose of analysis. For example, dantron is recommended by the European Pharmacopoeia for the estimation of valerenic and acetoxyvalerenic acids in valerian root. Valerenic and acetoxyvalerenic acids, which are active constituents of drugs, are highly unstable and difficult to isolate and to use as markers. The test response measured for these acids in the drug is interpreted from the standard plot of dantron to calculate the concentration of valerenic and acetoxyvalerenic acids in the test solutions. It has been experimentally shown that the standard curves of dantron and these two acids, in the range of estimation, are linear and parallel.

Identification of a marker requires knowledge of the chemical and active constituents of the plant drug. Markers are usually generated in-house and a limited number of markers are commercially available from different sources. Sigma Chemicals, Chromadex, Regional Research Laboratory (Jammu) and Natural Remedies (Bangalore) are some commercial sources of markers.

14.3.2 Analytical Techniques for Quantifying a Marker

After a marker has been identified, it needs to be quantified or assayed in the test material for the purpose of quality control. Any of the major analytical techniques, including high performance liquid chromatography (HPLC), high performance thin layer chromatography (HPTLC), gas chromatography (GC), radioimmunoassay, ultraviolet or infrared spectrometry, and mass spectrometry, can be used to determine the quantities of marker in the test samples. These techniques possess some advantages and disadvantages. Some are more commonly used while the others have limited applications. The assay method requires the procedure and technique to be simple, quick, specific, economical and robust. The ideal method can be used in different laboratories across the world without compromising on the accuracy and precision, which is possible if the procedure has a minimum number of critical variables. Whenever possible, the assay procedure should use simple, inexpensive equipment which is affordable in most places. The assay methods are developed through experimentation and from the existing knowledge of the drug and of techniques.

14.3.3 Validation of Analytical Procedures

Development of the assay method is followed by its validation. The validation procedure considers issues of specificity, linearity, range, accuracy, precision, detection limit, quantitation limit, robustness and system suitability. The International Conference on Harmonisation (ICH) has issued guidelines on validating analytical procedures, which are widely accepted and commonly used. Besides validation of a procedure, the instruments used are also validated from time to time. Practically all manufacturers of modern equipment supply detailed instructions and methods of equipment validation. The validation record is maintained and, in case a deviation is observed, a service engineer is called.

The ICH guidelines on validation are available from the internet for comprehensive understanding of the subject. A brief explanation to the validation procedure is provided here.

14.3.3.1 Specificity

Specificity indicates the extent to which an assay procedure specifically measures the analyte of interest. It is not always possible to demonstrate that an analytical procedure is specific for a particular analyte (complete discrimination). It may require two or more assay procedures to demonstrate the necessary level of discrimination. Spiking the sample with the analyte or related compounds and observing the effect on the estimations demonstrates assay specificity. If spiking with compounds related to the analyte produces no effect on the result, the procedure is considered specific. Specificity is particularly valuable when analyzing an analyte among several similar compounds present in the sample.

The issue of specificity with respect to herbal materials has lesser relevance as in several instances; we tend to determine the content of total active compounds rather than one active constituent, e.g. total sennosides in senna and not a particular sennoside. In such cases, a procedure that is able to assay all sennosides together is preferred than a specific procedure which discriminates and estimates only one sennoside.

14.3.3.2 Linearity

The linearity of an analytical procedure is its ability (within a given range) to obtain test results which are directly proportional to the concentration (amount) of the analyte in the sample. A linear relationship (between amount of analyte and response) should be demonstrated across the range used in the assay procedure. A minimum of 5 concentrations is recommended. A regression line, using appropriate statistical methods, should be drawn to calculate correlation coefficient, y-intercept and slope. A plot of the data should be included in the test report. Another important feature, which estimates the degree of linearity, is calculation of the deviation of the actual data points from the regression line. In some cases, mathematical transformation of data, prior to the regression analysis, may be required.

14.3.3.3 Range

The range is the interval between the lower and upper analyte concentrations for which it has been demonstrated that the analytical procedure has a suitable level of accuracy, precision and linearity. It is normally derived from linearity studies.

14.3.3.4 Accuracy

The accuracy of an analytical procedure is the closeness of agreement between the conventional true value (or an accepted reference value) and the value calculated. It tells how closely the analyte amount is determined to its true amount present in the test sample. Accuracy should be specified across the range of analytical procedure and inferred from 9 measurements (triplicates of three concentrations in the range).

Accuracy can be demonstrated by application of the proposed procedure to an analyte of known purity or by comparing the results of the proposed analytical procedure with those of a second well-characterized procedure, whose accuracy is already known. Application of the procedure to test samples after spiking at three different levels of 50%, 100% and 150% of expected analyte concentration helps to determine the accuracy of the procedure.

A sample containing 1.0 mg analyte may, in different analyses, be found to contain 1.2, 0.9 and 0.8 mg. The assay procedure determining 0.9 mg is more accurate than the two other procedures.

14.3.3.5 Precision

The precision of an analytical method is closeness of results for a series of measurements of multiple samples from the same homogeneous material. The precision may suffer upon varying the experimental conditions, which are therefore assumed to be kept constant. System precision demonstrates error in recording the response and can be determined by repeatedly analyzing a sample within a short period of time.

It is possible that the results are precise but not accurate or vice versa. Triplicate measurements of 1.0 mg true quantity as 0.6, 1.0 and 1.4 result in an average value of 1.0 mg which is accurate but the three measurements themselves are not precise. Similarly, triplicate measurements of the same quantity as 0.6, 0.7 and 0.6 are precise but not accurate. Triplicate measurements of 1.0, 0.9 and 1.0 are precise as well as accurate.

Precision is expressed at three levels of short, medium and long intervals, which are respectively referred to as repeatability, intermediate precision and reproducibility.

14.3.3.5.1 Repeatability

Repeatability is precision under the same operating conditions over a short interval of time (intra-day precision). It is demonstrated by a minimum of 9 determinations covering the specified range for the procedure (e.g. 3 concentrations/3 replicates each) or a minimum of 6 determinations at 100% of the test concentration.

14.3.3.5.2 Intermediate Precision

Intermediate precision expresses intra-laboratory variations: different days, different analysts, different equipment, etc.

14.3.3.5.3 Reproducibility

Reproducibility is precision at the inter-laboratory level. It is especially important if the analytical procedure is to be used in different laboratories, for instance, a pharmacopoeial procedure.

14.3.3.6 Detection Limit

Detection limit (DL) is the lowest amount of analyte in a sample which can be detected, but not necessarily quantitated as an exact value. It is determined by analyzing samples containing known concentrations of analyte and by establishing the minimum level at which the analyte can be reliably detected. Other methods of determining DL are based on the signal-to-noise ratio and on the standard deviation of the response and the slope. A signal-to-noise ratio of 3:1 or 2:1 is a fair estimate of DL. Using the standard deviation of the response (σ) and slope of the calibration curve (S), $DL=3.3\sigma/S$.

14.3.3.7 Quantitation Limit

Quantitation limit (QL) is the lowest concentration of the analyte that can be determined with acceptable precision and accuracy under the stated experimental conditions. It is generally determined by the analysis of samples with known concentrations of analyte and by establishing the minimum level at which the analyte can be quantified with acceptable accuracy and precision. QL can also be defined from the signal-to-noise ratio (a typical signal-to-noise ratio is 10:1), or from the standard deviation of the response (σ) and the slope (S) of the calibration curve ($QL=10\sigma/S$).

14.3.3.8 Robustness

Robustness of an analytical method ensures that it performs well and has few variables affecting its performance. For example, a robust HPLC analytical procedure does not show variation when columns of different lots or from different manufacturers are used, when there is a slight variation in pH or composition of the mobile phase, or when temperature and flow rate vary. Steps of extraction, purification or enrichment of an analyte in herbal material should be simple and time effective. Solutions used in the analysis should be stable over a reasonable period of time. Lastly, the procedure should be feasible in most laboratories.

If measurements are prone to variations with small changes in the test conditions then such conditions should be suitably controlled or a precautionary statement should be included in the procedure.

14.4 Thin Layer Chromatography in Quality Control of Plant Products

Thin layer chromatography (TLC), also called planar chromatography, is a widely accepted and extensively used separation technique that is over 65 years old. The technique is simple, cost effective, versatile, and useable in all laboratories around the globe. It can be easily adapted to any given situation of qualitative, quantitative or preparative separation. Despite the great variety and complete automation of the technique, it still lags behind other chromatographic techniques when it comes to its use as analytical technique. However, there is no substitute for this technique for situations requiring qualitative analyses of plant extracts. TLC has nearly become indispensable for the standardization of plant materials, be it the fingerprint profiling or analysis of a marker. The advantages of the technique over other analytical techniques are many when handling plant materials. The samples can be applied without undertaking tedious, time-consuming processes of sample preparation. The loss in sensitivity is far compensated by the gain on several fronts, including ease of assays, multiple sample analyses and low cost per sample.

The two prominent uses of TLC in the standardization of plant materials include fingerprint profiling for the assessment of chemical constituents of a drug and quantitative analysis of markers in plant drugs. A typical TLC procedure involves sample preparation, selection of the chromatographic layer and the mobile phase, sample application, development and drying of the plate, derivatization (if required) and chromatogram evaluation.

14.4.1 Sample Preparation

Methods of sample preparation for fingerprinting and estimation of marker differ significantly. Whereas in fingerprinting, only proportionate quantities of components must be extracted, in assaying the marker complete and exhaustive extraction has to be ensured. Correspondingly, particle size of the crude drug, the solute-solvent ratio, extraction period and number of extractions assume greater significance in marker estimations. Normally 1-2 g of moderately fine powder (unless specified) of plant material is extracted with 25-50 ml solvent at room temperature, in a Soxhlet apparatus or under reflux on a water-bath. The extraction is repeated a number of times to ensure complete and exhaustive extraction of the marker from the drug matrix. The extract is filtered and solvent is removed from the combined filtrate. The residue is dissolved in the solvent, filtered again, and the volume is adjusted. The concentration of the marker is determined in the solution.

On the other hand, in comparing fingerprint profiles, the procedure requires a shorter extraction scheme. A powdered specimen of pharmacopoeial quality may be required as the reference material for comparison of the fingerprint profiles. The test and sample solutions are prepared under identical conditions of extraction and concentration. Usually 0.1-1.0 g material is extracted with 1-10 ml solvent for 5-30 min, by shaking at room temperature or heating to boiling. The extract is filtered, concentrated and used. Sometimes the solvent is completely evaporated and the residue is dissolved in a small volume of solvent (typically less than 1 ml) and filtered to separate the insoluble particles. The solution of a marker, of preferably known strength, is required if marker presence is to be ascertained. Using known strength of marker additionally provides semiquantitative information.

In certain cases, the extracts require further purification using extraction of the residue with another solvent at different pH or using distillation, sublimation or other appropriate method.

14.4.2 Selection of Chromatographic Layer

A wide variety of options is available for the adsorbent layer. Laboratory-made plates have given way to precoated plates marketed by several manufacturers. The precoated plates are machine-made of glass, aluminium or plastic base coated with different adsorbents. The different adsorbents include normal phase silica gel (most commonly used), reverse phase silica gel (RP_2, RP_8, RP_{18}, cyano, diol and amino plates), aluminium oxide, cellulose, kieselguhr, hybrid (capable of being used as normal and reverse phase) and derivatized adsorbent layers. They come in different sizes, from small strips to continuous rolls (20 x 20 cm^2 is most common).

The nature of the compounds defines the choice of adsorbent layer; a stronger adsorbent (aluminium oxide) is used for weakly adsorbed compounds and a weak adsorbent (cellulose) is used for strongly adsorbed compounds. Normal phase silica gel is more suited for non-polar components and reverse phase silica gel is more suited for polar constituents, which are eluted first on reverse phase TLC. The silica gel plates containing fluorescent dye (F_{254}) of aluminium base are most widely used; about 80% of the analyses are done using these plates as they are optimally efficient and cost-effective.

14.4.3 TLC versus HPTLC Layers

High performance TLC (HPTLC) plates use thin layers of adsorbent (100 µm instead of 200-250 µm) and smaller particles (5-6 µm versus 10-12 µm) of more homogeneous size (4-8 µm versus 5-20 µm). Moreover, they give better resolution (5- to 10-fold more) over shorter runs (3-6 cm versus 8-15 cm), reduce separation time (3-20 min versus 20-200 min), accommodate more samples per plate (more than double), use smaller sample volumes (0.1-0.5 µl versus 1-5 µl) with improved detection limits (100-500 pg), and significantly improve the precision, accuracy and sensitivity. HPTLC plates are substantially more expensive (4- to 6-times more) than normal plates but are an efficient alternative when high sensitivity, accuracy and precision are required in situations demanding high performance. More improvements in adsorbent layers include use of spherical particles of narrow size distribution (reducing resolution time and size of spots while improving the detection limit) and ultrathin layers (10 µm) that improve the resolution and sensitivity and drastically reduce the development time.

14.4.4 Selection of the Mobile Phase

Infinite combinations and a wide choice of solvents are available for TLC developments. Unlike HPLC, where choice is limited, TLC provides no or few restrictions. A mobile phase with 1-3 components is preferred over a multicomponent mobile phase. The polarity of the compounds of interest is the key to selection of a mobile phase. Personal experience applied to existing knowledge and a trial and error method is used to select the composition of the mobile phase. The mobile phase is freshly prepared for each run and the constituting solvents are mixed outside before transferring to the developing chambers. It is advised to allow the developing chamber to saturate unless otherwise specified. Saturation of the chamber is quickened by lining half or more of the total area of the inside walls with filter paper and pouring the mobile phase over it. Closing the chamber and allowing it to stand at room temperature saturates the chambers. It is possible to use another solvent alongside the mobile phase for chamber saturation in twin troughs, e.g. ammonia placed in one trough and mobile phase in another.

The TLC results are sensitive to temperature and humidity variations. All operations during which the plate is exposed to the air should be carried out at a relative humidity of 50%-60% under controlled temperature of 20°-30° C.

14.4.5 Application of Sample

Three typical options of delivering the sample solution onto the plate are manual, semi-automatic and automatic application. Manual application is done using a capillary, which can have a specific volume of 1, 2 or 5 µl for quantitative purposes. The solution is applied by the technique of touch and deliver. The precision and accuracy, as known to the author from personal experience, is fairly high after a short experience. The semi-automatic application uses devices such as Linomat 5 from Camag and Applicator AS 30 from Desaga, which use a syringe that has to be manually cleaned and filled. The remaining part of the application is automated through computer commands. The solution is applied as a spot or band of predetermined size at predetermined points by touch and delivery or spray-on technique. The needle touches the surface of the adsorbent layer and delivers, whereas in spray-on technique the predetermined volumes are sprayed onto the plate. In the fully automated application, all steps are controlled through a computer including washing of the delivery line.

The typical concentration of the applied samples ranges from 0.1 to 1 mg/ml for qualitative analysis but is usually much lower for quantitative purposes, which further depends on the molar absorption of the marker. The typical volume for spot application is 1-5 μl, and 10 μl for band application. These volumes are drastically reduced in HPTLC plates or ultrathin TLC plates. Bands are known to give better resolution and results than spots, as a narrow band is better suited to the optics of the TLC scanner.

14.4.6 Developing the Chromatogram

Development of plates is carried out in chambers which are special purpose jars or simple containers good enough to hold the solvent in an air-tight environment. There is no doubt that special purpose chambers produce better chromatograms. Twin-trough chambers allow use of another mobile phase in the chamber for the purpose of saturation, besides consuming smaller quantities of solvent. The cost of the chamber, which seems high in the beginning, is recovered by way of savings on the quantity of expensive solvents. Presaturation of the chambers decreases R_f values and corrects side distortions of the solvent front. The plate is placed as nearly vertical as possible in the chamber, ensuring that the points of application are above the surface of the mobile phase and the sides of the plate do not touch the container walls.

The developing chamber should always be kept out of direct sunlight. It should be protected from light during development, if the components being investigated are suspected to be unstable. If sun rays fall directly on the developing chamber, they may be refracted to different degrees through the glass walls, producing areas of high temperature on the plate and resulting in erratic flow of the mobile phase.

The technique of development has been largely improved in horizontal developing chambers and completely automated in automated development chambers or automated multiple development chambers. However, the cost of this equipment (except for the horizontal development chamber) is excessively high.

14.4.7 Drying the Plate

After development, the plate is dried. This is an automatic procedure in automated development chambers, but it has to be accomplished in air at room temperature, in a vacuum desiccator or by heating or blowing hot-air over the surface of the plate. In all instances, the mobile phase should be as completely removed as possible before proceeding to derivatization or scanning the plate.

14.4.8 Derivatization

Derivatization involves treatment of developed chromatograms with suitable spray reagents for locating the position of the constituents for qualitative evaluation and for quantifying ultraviolet-insensitive markers. Two methods are employed for derivatization of plates: spraying with a fine mist of a reagent (a traditional method) and dip-in technique, which of late has become more popular. The spray method does not allow the uniform wetting of the plate, producing areas of high wetting and deficient spray. This affects the precision and accuracy in case of quantitative determinations. The dip-in technique produces more uniform wetting; special equipment is available for this purpose. In most cases of derivatization, heating is required after spraying the plate. Heating the plate uniformly in the open air produces better results than heating in an oven. The fumes from heating are strongly reactive and damage the inner walls of the oven. The plate is heated at about 110° C for about 10 min or until the spots are best seen. Special purpose heating plates are available from the manufacturers of TLC equipment.

14.4.9 Evaluation of the Chromatograms

The TLC plate is observed in daylight, under short-wave and long-wave ultraviolet light, for comparing the chromatograms of standard and test samples or for observing the presence of a marker or compounds of interest in the test chromatogram. The centre of each spot is marked with a needle. The distance from the centre of each spot to the point of application is measured to record the R_f value (the ratio of the distance travelled by a given compound to that travelled by the solvent front) or the R_r value (the ratio of the distances moved by a compound and a stated reference substance). R_f values may vary depending on the temperature, degree of saturation, the activity of the adsorbent layer and the composition of the mobile phase.

Quantitative evaluation is done by scanning the plate in a TLC densitometer or scanner. The densitometer uses two modes of transmittance and reflectance depending upon the available optics. It uses fluorescence mode, ultraviolet absorption or visible light for quantitation of the marker depending upon the option exercised. Ultraviolet and visible light absorption modes come as a standard option on a scanner and the fluorescence mode is optional. Data acquisition and analysis is through standard PC-based software. Multi-wavelength scanning, recording and comparing ultraviolet spectra, and generating and acquiring spectra libraries are among several options available on the software provided with the TLC scanner.

The determination of analyte concentration is through a standard plot or single or double point calibrations.

14.4.10 Improving the Efficiency of TLC

Several precautions can be taken to improve the efficiency of TLC analysis. These include carefully selecting the range of concentrations for analysis; using correct instrument parameters like slit dimensions, wavelength selection, scanning speed, base line correction; using HPTLC plates for high sensitivity and resolution; use of appropriate sorbent from a wide range of sorption properties to optimize selectivity; use of automated sample application, development and detection; use of precise in situ recording and quantitation of chromatograms; and avoiding derivatization in assay procedures and, if necessary, using dip-in technique of derivatization.

The following example, of one of the Ayurvedic drugs, illustrates the use of TLC in quality control of plant material. The drug was analyzed for one of the active compounds and the TLC fingerprint profile was used for the purpose of positively identifying the plant material.

To prepare the fingerprint profile, about 5 g plant material was extracted with 50 ml methanol for 30 min at 50° C in a conical flask. The extract was filtered and the filtrate was concentrated to about 5 ml under vacuum. One of the active constituents isolated from this plant (code name DPH-1) was used as a reference. The solution of the reference substance was prepared by dissolving about 5 mg in 1 ml chloroform. About 10 µl of each test and reference solution was manually applied in band form on aluminium base, silica gel 60 F_{254}, 0.2-mm thick TLC plates (Merck). The plate was developed using mobile phase containing 95 volumes toluene and 5 volumes ethyl acetate. The plate after development was dried and visualized under 254 nm ultraviolet light (Figure 1A). The same plate was sprayed with anisaldehyde-sulphuric acid reagent and heated for about 10 min to visualize the spots (Figure 1B). These profiles can be used to confirm the identity of the plant material and to obtain semiquantitative information on the amount of DPH-1 present in the drug.

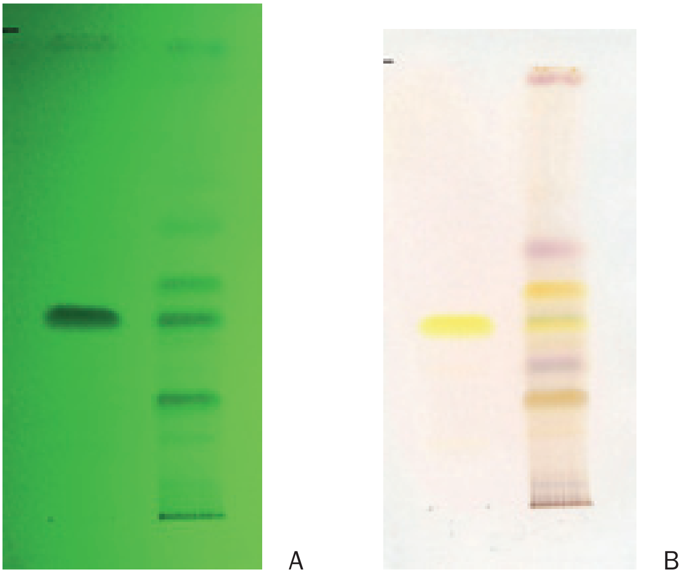

Figure 1: TLC chromatograms as visualized (A) under 254 nm UV and (B) after being sprayed with anisaldehyde-sulphuric acid reagent

Besides developing the fingerprint profile of the drug, the quantity of DPH-1 was also estimated in different samples of the plant material. For the purposes of analysis, 5 g moderately fine powder of drug material was extracted with methanol in a Soxhlet apparatus for 4 h. The extract was filtered and the volume was adjusted with methanol to 50 ml in a volumetric flask. One milliliter of this solution was diluted to 10 ml in a volumetric flask and used for analysis. A standard solution of DPH-1 was prepared by dissolving 4.95 mg DPH-1 in 10 ml methanol and diluting 0.5 ml of this solution to 10 ml in a volumetric flask. Six different concentrations of this solution were applied in triplicate on a precoated TLC plate, which was developed using mobile phase containing 90 volumes toluene and 10 volumes ethyl acetate. The developed and dried plate was scanned at 305 nm in a TLC scanner and the standard plot was constructed (Figure 2). One microliter of test solution was similarly analyzed using the same conditions as used for DPH-1, and the amount of DPH-1 in the test sample was calculated from the response obtained in a TLC scanner. The analyzed drug samples showed large variations in the content of DPH-1, ranging from below 0.3% to over 1.4%. The method was validated according to ICH guidelines.

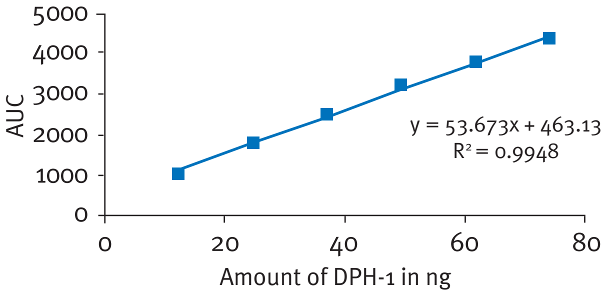

Figure 2: Calibration curve of DPH-1

The TLC-based method of analysis and fingerprint development is quick and reliable, and can be used conveniently in different laboratories. Similarly, it is possible to apply this technique to other plant drugs to develop fingerprint profiles and also to estimate the percentage of marker substances in the crude drugs or in finished products.

14.5 High Performance Liquid Chromatography

In a period of less than 50 years, HPLC has become the most widely used analytical tool in most laboratories of the world. The technique has received great attention for innovations leading to its overall development, regarding both consumables and equipment. HPLC separations are achieved using any of the five basic chromatographic modes: liquid-solid (adsorption), liquid-liquid (partition), bonded-phase (partition), ion exchange, and size exclusion chromatography. The selected mode depends on the nature and properties of the analyte. Bonded-phase chromatography, in which a stationary phase of organosilanes of varying carbon lengths is chemically bonded to silanol groups, is the most commonly used mode of separation. In liquid-liquid chromatography, the solid support (usually silica or kieselguhr) is mechanically coated with a film of high boiling point organic liquid, unlike bonded-phase chromatography where non-polar hydrocarbon chains are chemically bonded to hydroxyls of the silica support. Liquid-liquid chromatography, by virtue of its mechanism, is more susceptible to changes by interaction with mobile phase than bonded-phase chromatography.

A typical HPLC operation includes pumping of mobile phase at moderately high pressure through a narrow-bore column containing adsorbent. The separation of the mixture takes place in the column and separated components are detected by employing a suitable detector. As the mobile phase is being pumped at high pressures, a system is required to inject the mixture into the system without dropping the pressure and disturbing the flow characteristics, i.e. rate and pressure. To accomplish these requirements, an HPLC system requires a pump to push the mobile phase against high pressure, an injector to insert a solution of standard substance or test mixture, a column to effect separation, a detector to reveal the presence of analyte in the eluate, and a suitable data processing unit.

14.5.1 Pumps

The pump is considered a heart of the HPLC system, as all depends on the composition of the mobile phase and its flow rate accuracy. The pump gives a pulse-free flow of mobile phase; the expected variations in flow rate are less than 1.0%. Online mixing of solvents is preferred to manual mixing. However, compositions containing less than 10% of a particular solvent are better prepared by manual mixing. The composition of the mobile phase is either constant during the analysis (isocratic mode) or it is changed (gradient mode). The type and design of modern pumps allow low pressure mixing of up to four solvents; else, different pumps, one for each solvent, are required for gradient operation and the solvents are mixed at high pressure. A typical analytical procedure uses a flow rate of about 1 ml/min and operating pressure between 1000 and 2000 psi. Higher flow rates generating higher pressure should always be used with justification, as they decrease column life besides requiring frequent servicing of the pump.

The flow accuracy of the pumps is critical for analysis. The constancy of retention time of the last eluted peak is a measure of long-term flow accuracy of the pumps, whereas short-term flow accuracy is checked from the average peak areas of each component and their standard deviations. The mobile phase must be free of dissolved gases to ensure an accurate flow and to minimize noise due to bubbles. Vacuum filtration, sonication and helium gas purging are methods for degassing.

14.5.2 Injector

The injector allows a predetermined volume of test solution to be introduced into the flow channel of the system, without disturbing the flow kinetics. Typically, fixed volume injections are preferred over variable volume injections. When using fixed volume loops, it is advisable to flush higher volumes of the sample through the loop to ensure complete filling of the loop with the sample solution. The mobile phase close to the inner walls of the loop can only be assured to have pushed out after injecting volumes larger than the loop volume, e.g. injecting 20 μl test solution into a 20-μl loop cannot assure accurate injection; if the quantity of the test material is not a problem, flush the loop with over 100 μl test solution. Only the appropriate needle (compatible with the injection port) should be employed for the purpose of making an injection. It is important to select a syringe of appropriate size when giving variable injections; a thumb rule for any analytical technique is not to use the volumetric apparatus if less than 20% of its total volume is being used. Thus, a syringe of 25 μl should not be used to measure or inject volumes less than 5 μl.

14.5.3 Columns

Columns come in varied sizes, structural architecture and chemistry. The chromatographic material is usually packed in stainless steel casing. The material is composed of porous particles which vary in nature (inorganic ceramic, organic polymer, or hybrid), shape (irregular or spherical), size (ranging from 2 to 20 µm; normally around 5 µm) and surface modifications (silanes of different carbon lengths, aminopropyl, diol, cyano, sulphonic acid and ammonium ions). The choice of column is based on the type of analysis. Comprehensive information is available on the websites of the leading manufacturers of HPLC columns, which serves as good guides in choosing columns for analysis. Most analyses are reported on reverse phase columns, usually C_{18}, with increasing emphasis on reducing the column length, diameter and analysis time. Most HPLC separations are successful on columns maintained at ambient temperature, but thermostatted column maintained to ±0.2° C is necessary for reproducible results. This is because all mechanisms of separation are temperature-dependent and any shift in temperature has remarkable bearings on the result.

14.5.4 Detectors

A wide variety of detectors is available to cater to diverse needs of the analysts. Ultraviolet detectors of fixed wavelength, dual wavelength or variable wavelength (photodiode array detector) are most frequently used. Other options are refractive index detector, fluorescent detector, electrochemical detector, evaporative light scattering detector and chemiluminescence detector.

14.5.5 Data Processing

The electrical response from the detector is digitalized and fed to a data processing module, which in present days is invariably a computer, and computations are made using special software. Several software programs are available for data processing, from both manufacturers and third parties. Besides computing the data, they also control the entire operation of the machine.

14.5.6 Factors Affecting HPLC Analysis

Numerous variables affect an HPLC analysis. This topic is beyond the scope of this paper, but some critical variables are discussed briefly. Increased emphasis is now paid to control the temperature of the column within a narrow range to ensure precision of the result. This is desirable, as factors such as solubility, solute diffusion, viscosity of the mobile phase, and column plate number all are affected by temperature. Mobile phase composition is another vital parameter that affects the resolution, retention time and peak area. Pumps contribute the most towards variations of results, as precise composition of mobile phase and flow rate can only be assured by accurate pumps. Gases dissolved in the mobile phase are a source of flow-rate inaccuracies and errors in detector response.

Retention time variations are often discussed to know the tolerable limits. The retention time is affected by flow rate, column temperature, mobile phase composition and integration. An error in flow rate leads to changes in the retention time to the same extent. Small variations in column temperature have more significant effects on retention time. Ideally, a column is thermostatted to ±0.2° C. However, a high precision of 0.1% in retention time requires the column to be thermostatted to ±0.04° C. Changes in mobile phase composition leave a stronger impact on the retention time. It is estimated that in a typical isocratic elution, a variation of ±1% in mobile phase composition occurs, which introduces an error of 0.4%-0.7% in retention times. The observed variations in composition of the mobile phase are more in the gradient elution. Recording devices also introduce variations in the retention time through faulty recording, but the effect is much smaller (in the range of 0.1% to 0.04%).

Peak area is affected by all the factors that affect retention time. Additionally, the recorder response in marking the start and end of the peak is crucial; this has been seen to be the main source of error in recording peak areas.

Several more factors, like injection volume, connecting tubing, end fittings and detector volume, also have bearings on the final results. Large injection volume and quantity of analyte result in broadening of the peak. Preparing the sample in the mobile phase produces the best result and should be taken as the first choice.

14.5.7 HPLC in Quality Control of Plant Products

HPLC is the most popular technique among all the analytical techniques used today. It is therefore understandable that most happenings are taking place in the modernization of this technique. As discussed in the section on TLC, HPLC can be used for similar purposes. There are two applications of HPLC: one to generate the profile, for which TLC is preferred, and one to estimate the quantity of markers, where HPLC is preferred. The initial steps of sample preparation are similar to those for TLC with the exception that the samples for HPLC are filtered through a filter of 0.45 μm or less. Furthermore, it is assured that the test sample does not contain substances which are permanently retained on the HPLC column, which means in most cases purification procedures are applied to extracts before injecting them onto the column.

After the sample has been prepared, it is injected onto the column and the response is recorded preferably using a variable wavelength ultraviolet detector. As the nature of all the compounds of the extract cannot be known beforehand, the photodiode array detector is useful, especially when constructing profiles of plant extracts. The fingerprint profile of plant extracts can be used for identification purposes and also for obtaining semi-quantitative information if the sample preparation was not done for quantitative analysis. Similarly, the profile can be generated for the finished product and used to record batch to batch variations. The fingerprint profile can be used to study changes in the composition of the finished product or, in other terms, to indicate the stability of the product.

The most important use of HPLC is in estimation of markers in plant drugs. The steps in HPLC analysis are fundamentally the same as used for any other analytical technique. The response of the test sample is compared to that of a known quantity of the marker to quantify the marker in the test substance. The HPLC method is developed from knowledge of the technique and chemistry of the marker. In chemical analysis, HPLC has no parallel and can be customized to produce the most precise and accurate results. The HPLC analysis is vital in the analysis of a finished product and the expected results are superior to those from TLC, as the separations in HPLC are better. However, run time of HPLC analyses usually varies from 15 to 30 min, which restricts its use if large numbers of samples are to be analyzed.

14.6 TLC versus HPLC

TLC has emerged as a major tool in standardization of plant materials. The advancement and automation of the technique has made it a first choice for plant drugs. Its use has become more popular in developing countries where advancements of HPLC are not cost efficient. TLC offers several advantages over HPLC. Sample and mobile phase preparation do not require elaborated steps of purification, degassing, and filtration, which are essential to protect expensive columns from deterioration. Several samples (up to 18) can be accommodated on a single 20 x 20 cm^2 plate. The test samples and standards are analyzed simultaneously under the same conditions. Several analysts can work simultaneously as each step in analysis is carried out independently using separate equipment. The choice of solvent systems is unlimited, unlike for HPLC where column chemistry disallows the use of extremes of pH in mobile phase. The technique allows enormous flexibility of derivatization with chromogenic spray reagents, making possible the detection of an analyte that is transparent to ultraviolet light. It also allows multiple evaluations of the developed chromatogram, which is not possible in HPLC. There is no leftover from the previous analysis to interfere in the next, as each time a fresh plate is employed. Lastly, it saves tremendously on the time and cost of the analysis.

TLC offers many advantages but also has some disadvantages. It fails to match the sensitivity of HPLC and has not kept with the pace of developments and advancements happening in the area of HPLC. TLC is

Chemistry

Like the author's other companion books, **The Chemistry Companion** provides high quality information in unique one-page-per-topic presentations that do not overburden and distract readers with excessive details. The book offers concise summaries of general chemistry concepts, easily accessible in a convenient, reader-friendly format.

Suitable as an introduction or study guide, this companion presents the minimum of what readers need to know to understand the subject. It emphasizes the physics underlying chemistry. By looking at chemistry processes from a physics point of view, readers better appreciate what is happening from the chemical perspective that is usually found in traditional chemistry books.

The author focuses on the structure of matter, chemical components and bonds, the periodic table, states of matter, thermodynamics, reaction rates, carbon chemistry, biochemistry, and chemical, ionic, and electronic equilibria. Each topic is covered in a single-page outline format with just enough detail to enable a good understanding of the subject.

Features

- Provides a physical understanding of chemical concepts
- Presents clear explanations of difficult material, working through any inconsistencies in understanding
- Uses a convenient format for checking formulas and definitions
- Includes self-contained information on each page, assuming little prior knowledge

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K11517

ISBN: 978-1-4398-3088-8
90000
9 781439 830888

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press

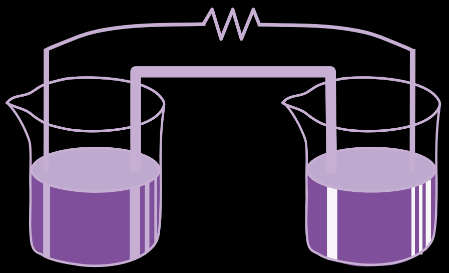

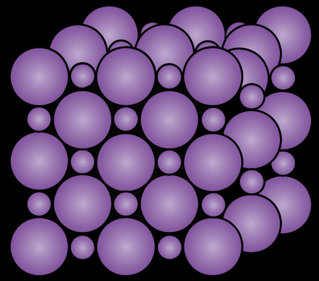

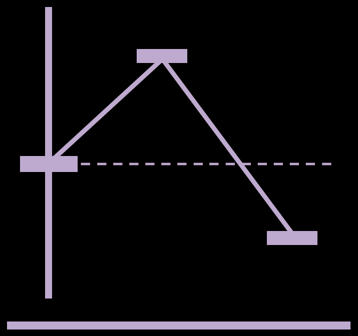

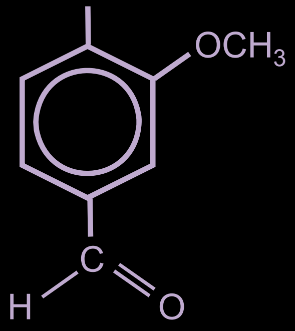

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press
Taylor & Francis Group

Contents

Preface xiii

1. Structure of Matter 1
- 1.1 Atoms 2
- 1.2 Bohr Atom 3
- 1.3 Energy Levels 4
- 1.4 Schrödinger Equation 5
- 1.5 The Infinite Square Well 6
- 1.6 The Coulomb Potential 7
- 1.7 Covalent Bond 8
- 1.8 Ionisation Energy 9
- 1.9 Electron Affinity 10
- 1.10 Ionic Bond 11
- 1.11 Electronegativity 12
- 1.12 Metallic Bond 13
- 1.13 Electronic Structure of Solids 14

2. Chemical Components 15
- 2.1 Matter 16
- 2.2 Atomic Weight 17
- 2.3 Ions 18
- 2.4 Molecules 19
- 2.5 Mole 20
- 2.6 Compounds 21
- 2.7 Cations 22
- 2.8 Anions 23
- 2.9 Chemical Equation 24
- 2.10 Stoichiometry 25
- 2.11 Example 26

3. The Periodic Table 27
- 3.1 Electron Configuration 28
- 3.2 Periodic Law 29
- 3.3 Periodic Table 30
- 3.4 Groups 31
- 3.5 Energy Levels 32
- 3.6 Noble Gases 33
- 3.7 Atomic Size 34
- 3.8 Covalent Radii 35
- 3.9 Ionic Radii..36
- 3.10 Ionisation Energy 37
- 3.11 Electronegativity 38

4. Chemical Bonds 39
4.1 Chemical Bond 40
4.2 Lewis (Electron Dot) Formulae 41
4.3 Multiple Bonds 42
4.4 Lewis Single-Bonded Structures 43
4.5 Lewis Multiple-Bonded Structures 44
4.6 Lewis Exceptions to the Octet Rule 45
4.7 Oxidation Number 46
4.8 Oxidation Number Examples 47
4.9 Polar Bonds 48
4.10 Hybrid Orbitals 49
4.11 Polarisation 50
4.12 van der Waals Forces 51
4.13 Hydrogen Bond 52

5. States of Matter 53
5.1 Changes of State 54
5.2 Changes of State of Liquids/Gases 55
5.3 Phases of Matter 56
5.4 Gases 57
5.5 Solutions 58
5.6 Aqueous Solutions 59
5.7 Solubility of Solids 60
5.8 Solubility Equilibrium 61
5.9 Electrolytes 62
5.10 Osmosis 63
5.11 Solids 64
5.12 Crystalline Lattice Structures 65
5.13 Metallic Solids 66

6. Chemical Thermodynamics 67
6.1 Molecular Energy 68
6.2 Specific Heat Capacity 69
6.3 Enthalpy 70
6.4 Heat of Reaction 71
6.5 Heat of Reaction 72
6.6 Enthalpy of Formation 73
6.7 Entropy 74
6.8 Entropy Calculations 75
6.9 Gibbs Energy 76
6.10 Spontaneous Processes 77
6.11 Melting of Ice 78
6.12 Freezing of Water 79

6.13 Ice/Water Equilibrium 80
6.14 Chemical Equilibrium 81
6.15 Statistical Entropy 82

7. Rates of Reaction 83
7.1 Rates of Reaction 84
7.2 Collision Theory 85
7.3 Reaction Mechanism 86
7.4 Activation Energy 87
7.5 Nature of Reactants 88
7.6 Concentration 89
7.7 Rate Law 90
7.8 Rates of Reactions 91
7.9 Determination of Order 92
7.10 Determination of 1st Order Rate Law 93
7.11 Half-Life Method 94
7.12 Temperature 95
7.13 Catalysts 96

8. Chemical Equilibrium 97
8.1 Chemical Equilibrium 98
8.2 Law of Chemical Equilibrium 99
8.3 Equilibrium Constant 100
8.4 Le Chatelier’s Principle 101
8.5 Summary of Le Chatelier’s Principle 102
8.6 Equilibrium in Gaseous Systems 103
8.7 Solubility of Solids 104
8.8 Factors Affecting Equilibrium 105
8.9 Common Ion Effect 106
8.10 Precipitation 107
8.11 Complex Ions 108

9. Ionic Equilibrium 109
9.1 Electrolytes 110
9.2 Ionisation of Water 111
9.3 H^+ and OH^- Concentrations 112
9.4 Acids and Bases 113
9.5 Bronsted–Lowry 114
9.6 Strength of Acids 115
9.7 Acid–Base Reactions 116
9.8 Buffer Solutions 117
9.9 Indicators 118
9.10 Neutralisation 119

9.11 Titration 120
9.12 Choice of Indicator 121
9.13 Hydrolysis 122
9.14 Simultaneous Equilibria 123
9.15 Complex Ions 124

10. Electronic Equilibrium 125
10.1 Oxidation and Reduction 126
10.2 Redox Reaction 127
10.3 Single-Cell Redox Reaction 128
10.4 Standard Hydrogen Electrode 129
10.5 Standard Electrode Potentials 130
10.6 Spontaneous Redox Reactions 131
10.7 Oxidation Numbers 132
10.8 Balancing Redox Half-Reactions 133
10.9 Balancing Redox Reactions with Oxidation Numbers 134
10.10 Electrochemical Cell 135
10.11 EMF vs. Concentration 136
10.12 Electronic Equilibrium 137
10.13 Equilibrium Constant 138
10.14 Lead/Acid Battery 139
10.15 Dry Cell 140
10.16 Corrosion 141
10.17 Electrolysis 142

11. Carbon Chemistry..143
11.1 Carbon 144
11.2 Carbon Compounds 145
11.3 Functional Groups 146
11.4 Alkanes 147
11.5 Alkenes 148
11.6 Alkynes 149
11.7 Benzene 150
11.8 Alcohols 151
11.9 Ethers 152
11.10 Aldehydes 153
11.11 Ketones 154
11.12 Carboxylic Acids 155
11.13 Esters 156
11.14 Amides 157
11.15 Amines 158
11.16 Polymers 159
11.17 Physical and Chemical Properties 160

12. Biochemistry 161
12.1 Sugars 162
12.2 Polysaccharides 163
12.3 Lipids 164
12.4 Proteins 165
12.5 Nucleic Acids 166
12.6 DNA 167
12.7 Enzymes 168
12.8 ATP 169
12.9 Anaerobic Metabolism 170
12.10 Aerobic Metabolism 171
12.11 Cyclic Photophosphorylation 172
12.12 Non-Cyclic Photophosphorylation 173
12.13 Metabolism 174

1. Structure of Matter

Summary

Energy levels Bohr H-like atom:

$$E_n = -\frac{m_e Z^2 q_e{}^4}{8\varepsilon_o{}^2 h^2 n^2}$$

Energy levels Schrödinger equation,
Coulomb potential:

$$E_n = -\frac{Z^2 q_e{}^4 m}{(4\pi\varepsilon_o)^2 2\hbar^2 n^2}$$

Covalent bond: the co-sharing of valence electrons.

Ionic bond: the electrostatic attraction between ions formed after electrons are transferred from one atom to another.

Metallic bond: free electrons in the valence band reduce the energy of the system and so act to hold atoms together.

The electronegativity describes the relative ability of an atom, when it combines with another atom, to become "more negative" by more strongly capturing a shared electron or electron pair.

1.1 Atoms

Sixth century BC Thales of Miletos proposes that the basic element from which all things are made is water. 450 BC Empedocles teaches that all matter is composed of earth, air, water and fire. Around 400 BC, Greek philosophers (Leucippus, **Democritus**, Epicurus) proposed that if one could divide a piece of matter again and again, eventually a limit would be reached where no further subdivision could take place, this limiting amount of matter was called the **atom**. About 300 BC, **Aristotle** rejects the atomistic view and argues that matter is based upon the four basic elements of Empedocles but adds the qualities of coldness, hotness, dryness and moistness. Aristotle's considerable reputation ensured that his ideas became embodied in religious teaching for many hundreds of years.

In the fifteenth century, new advances in physics suggested that matter was made from particles, in agreement with the ancient Greek atomists. Robert **Boyle** taught that matter consists of different types of **elements** that were composed of atoms of the same type. Different elements could join together in fixed proportions to form **compounds**. Later, in 1803, John **Dalton** proposed the **atomic theory of matter** that was based on quantitative experimental evidence from the weighing of different elements in combination. He created a scale atomic mass for the different elements that were then known. Dalton's reference atom was the lightest element known, hydrogen, which was assigned an atomic mass of one. Other elements were given atomic masses according to how heavy they were compared to a hydrogen atom.

In 1807, Humphry Davy decomposed potash into sodium and potassium metals using electrolysis. In 1832, Michael **Faraday** discovered a quantitative connection between electricity and the separation of compounds into elements in electrolysis. These observations suggested that atoms themselves contain electric charge. Experiments by William Crookes demonstrated visible "cathode rays" that emanated from a negatively charged electrode (cathode) and travelled towards the positive electrode (anode) in an evacuated tube.

In 1872, **Mendeleev** arranged elements in increasing order of atomic mass and discovered that the properties of certain elements were repeated at regular intervals. When elements were ordered in columns with the atomic mass going across from left to right, and similar chemical properties going down, a **periodic table** was formed whereby, using the known elements at the time, Mendeleev was able to predict the properties of some as-yet undiscovered elements from gaps in the table.

1.2 Bohr Atom

In 1897, **Thomson** demonstrated that the rays observed to be emitted from the cathodes of vacuum tubes were in fact charged particles which he called **electrons**. Thomson proposed that atoms consisted of a positively charged sphere within which were embedded negatively charged electrons.

Rutherford found in 1911 that the electrons were actually located at some distance from a central positively charged **nucleus**. He proposed that electrons orbited the nucleus and the electrostatic attraction between the nucleus and the electron was balanced by the centrifugal force arising from the orbital motion. However, a major problem with this was that if this were the case, then the electrons would continuously radiate all their energy as **electromagnetic waves** and very quickly fall into the nucleus.

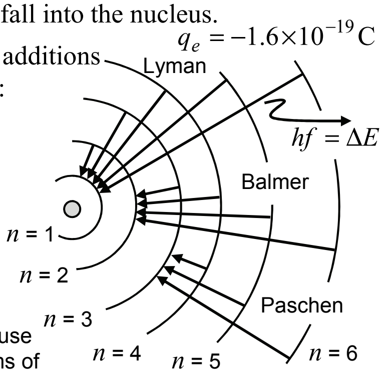

In 1913, **Bohr** postulated two important additions to Rutherford's theory of atomic structure:

1. Electrons of mass m_e can orbit the nucleus at radius r with velocity v in what are called **stationary states** in which no emission of radiation occurs and in which the **angular momentum** L is constrained to have values:

   $$L = m_evr = \frac{nh}{2\pi}$$

   The 2π appears because L is expressed in terms of ω rather than f.

2. Electrons can make transitions from one state to another accompanied by the emission or absorption of a single **photon** of energy $E = hf$, this being the absorption and emission spectra observed experimentally.

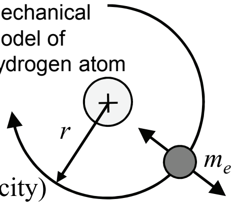

Mechanical model of hydrogen atom

As in the Rutherford atom, the centrifugal force is balanced by Coulomb attraction:

$$\frac{1}{4\pi\varepsilon_o}\frac{q_e^2}{r^2} = \frac{m_ev^2}{r}$$

with the addition that: $$m_evr = \frac{nh}{2\pi}$$

By summing the **kinetic energy** (from the orbital velocity) and the **potential energy** from the electrostatic force, the **total energy** of an electron at a given energy level n is given by:

$$E_n = -\frac{m_eZ^2q_e^4}{8\varepsilon_o^2h^2n^2}$$

Note: **atomic number** Z = 1 for the hydrogen atom where the energy of the ground state is –13.6 eV. The energy levels for each state n rise as Z^2. Thus, according to the Bohr model, the energy level of the innermost shell for multi-electron atoms can in principle be several thousand eV.

1.3 Energy Levels

The stationary states or energy levels allowed by the Bohr model are called **electron shells** or orbitals, and are labelled K, L, M, N, etc. with K corresponding to n = 1. The number n is called the **principal quantum number**. According to the Bohr model, the electron energy only depends on n, but experiments show that in multi-electron atoms, electron shells consist of sub-levels (evidenced by fine splitting of spectral lines). For example, the L shell n =2 has two **sub-shells**, 2*s* and 2*p*.

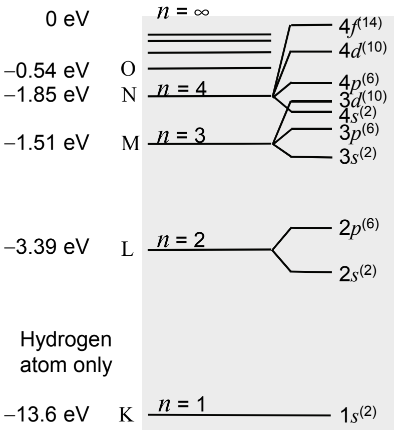

It is convenient to assign the energy at infinity as being 0 since as an electron moves closer to the nucleus, which is positively charged, its potential to do work is less and thus the energy levels for each shell shown are negative. In hydrogen, a single-electron atom, the energies for each shell are given by:

$$E = -\frac{13.6}{n^2} \quad \text{for hydrogen}$$

The electron-volt is a unit of energy. 1 eV = 1.602 × 10^{-19} J.

Sometimes the splitting of principal shells into sub-shells results in some overlap (e.g. 4*s* is lower in energy than 3*d*).

At each value of n the **angular momentum** can take on several distinct values. The number of values is described by the second quantum number l. The allowed values of l are 0, 1, … (n−1). Each value of l is indicated by a letter that indicates the sub-shell:

l = 0	*s*
l = 1	*p*
l = 2	*d*
l = 3	*f*
l = 4	*g*
l = 5	*h*

A third quantum number m describes the allowable changes in angle of the **angular momentum** vector in the presence of an electric field. It takes the values −l to 0 to +l.

A fourth quantum number describes the **spin** of an electron where the spin can be either −1/2 or +1/2.

According to the **Pauli exclusion principle**, no electron in any one atom can have the same combination of quantum numbers. When all the electrons in an atom are in the lowest possible energy levels, the atom is said to be in its **ground state**. The outermost electrons in an atom are called the **valence electrons**.

For example, the 3*d* sub-shell can hold up to 10 electrons:

$n = 3$
thus: l = 0, 1, 2 (*s*, *p*, *d*)
and: $m = \underbrace{-2, -1, 0, 1, 2}$

5 values of m times 2 for spin, thus 10 possible electrons

1.4 Schrödinger Equation

The total energy E of an electron in an atom is the sum of the potential and kinetic energies. Expressed in terms of **momentum**, p, and mass of electron m, this is stated:

$$E = \frac{p^2}{2m} + V$$

The value of the potential function may depend on both position and time. The form of $V(x,t)$ is different for different arrangements of atoms (e.g. a single isolated atom, an atom in a regular array of a crystal).

Thus: $$hf = \frac{p^2}{2m} + V(x,t)$$

since $E = hf$

Let $$p = -i\hbar\frac{\delta\Psi}{\delta x}$$

$$E = i\hbar\frac{\delta\Psi}{\delta t}$$

Ψ is a variable, the form and value of which provide information about the motion of a wave/particle.

Thus: $$\boxed{-\frac{\hbar^2}{2m}\frac{\delta^2\Psi}{\delta x^2} + V(x,t)\Psi = i\hbar\frac{\delta\Psi}{\delta t}}$$

Schrödinger equation

The solution to the Schrödinger wave equation is the **wave function** Ψ. If V is a function of x only, then the wave equation can be separated into time-independent and time-dependent equations that can be readily solved.

$$-\frac{\hbar^2}{2m}\frac{\delta^2\psi}{\delta x^2} + V(\psi) = E\psi \qquad \phi(t) = e^{i\frac{E}{\hbar}t}$$

The resulting solutions of these equations, when multiplied together, give the **wave function**: $$\Psi(x,t) = \psi(x)\phi(t)$$

The wave function gives all the information about the motion of an electron in an atom. Ψ is a complex quantity, the magnitude of which $|\Psi|$ is interpreted as a **probability density function** which in turn can be used to determine the probability of an electron being at some position between x and Δx.

Quantum mechanics is concerned with determining the wave function (i.e. solving the Schrödinger equation) for particular potential energy functions such as those inside atoms. It is found that valid solutions to the time-independent wave equation occur only when the total energy is quantised. The solutions correspond to **stationary states**.

Solutions to the Schrödinger equation can be found for potential functions which are a function of both x and t. This enables time-dependent phenomena (e.g. the probability of transitions of electrons between energy levels in an atom) to be calculated and hence the intensity of spectral lines.

1.5 The Infinite Square Well

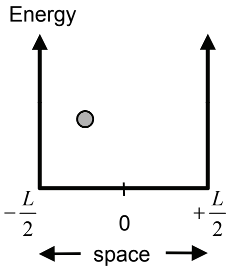

Consider an electron that is confined to be located in one-dimensional space $-L/2$ and $+L/2$ and where the **potential energy** of the electron is a constant (which can be conveniently set to zero). This is an example of an **infinite square well potential**. This potential is not usually found around electrons in an isolated atom, but does often represent that experienced by electrons in a solid and also in a chemical bond.

In the infinite square well potential, the electron cannot move more than a distance $x = \pm L/2$ from the centre position because it is constrained or bound by the infinite potential at the walls. In classical physics, the electron can have any value of total energy as it moves within this space.

In quantum physics, the allowed **stationary states** indicate the possible total energy of the electron. That is, the electron is moving as a particle whose probability of being in a particular position x is described by the standing wave patterns from the solutions to the **Schrödinger equation**. For the simple case of ***zero* potential energy** within the well, the energy levels are given by:

$$E = \frac{n^2\pi^2\hbar^2}{2L^2m} \qquad n = 1,2,3,4\ldots,\ V(x) = 0$$

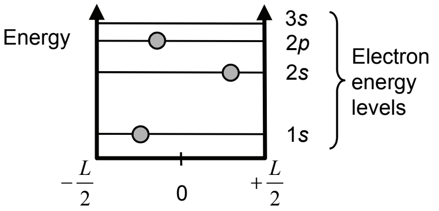

The electron can move anywhere within the confines ± L, but can only have kinetic energies allowable by the stationary states.

It can be seen that the energy increases as n increases. That is, the electron has a greater **kinetic energy** if it exists in a **stationary state** with a larger n. The minimum allowable energy is greater than the minimum potential energy (in this case, 0). This is the **zero point energy**.

Note also that the kinetic energy of the electron, for a given value of n, decreases as the length of the well L increases. That is, if the electron is given more room to move, then its total energy is lowered (in this case, the total energy is kinetic, since the potential energy was set to be 0). "Available space" is effectively a mechanism for providing a reduction in total energy of the system. This is important when the **valence electrons** of different atoms come near each other during the formation of **chemical bonds**.

1.6 The Coulomb Potential

In the **infinite square well potential**, the potential energy of the electron inside the well was constant, and independent of position. In an isolated **single-electron atom**, a more realistic scenario is the case where electrons are bound by the **Coulomb potential**.

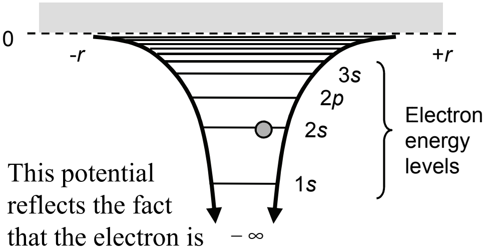

This potential reflects the fact that the electron is attracted more strongly by the nucleus at shorter distances r and so the potential energy is no longer constant – i.e. the potential varies with distance from the nucleus. Mathematically, the Coulomb potential is:

$$V(r) = -\frac{Zq_e^{\,2}}{4\pi\varepsilon_o r}$$

where Z is the **atomic number**, the number of protons in the nucleus.

By convention, an electron is assigned a zero total energy when it is at rest, at an infinite distance from the nucleus. As the negatively charged electron moves towards the positively charged nucleus (opposite to the direction of the electric field), it can acquire kinetic energy and/or do work and so its electrical potential is reduced. That is, the potential energy becomes negative. This is the same convention in electric field theory, where it is usual to consider a *positive* test charge moving in the *same* direction of the electric field – whereupon potential energy is reduced.

Note that in this potential, the energy (the potential energy) becomes more negative as r decreases. That is, electrons further away from the nucleus have a higher potential energy.

The total energy levels for an electron for this potential are given by the solution to the **Schrödinger equation** with $V(r)$ as above and are expressed:

$$E_n = -\frac{Z^2 q_e^{\,4} m}{(4\pi\varepsilon_o)^2 2\hbar^2 n^2} \qquad n = 1,2,3,4\ldots$$

This applies to an isolated **single electron atom** (e.g. hydrogen). Compare with Bohr model 1.2.

Here, the energy E becomes more negative as n decreases, and for a single electron ($Z = 1$) , and at $n = 1$, we obtain the total energy $E = -13.6$ eV. The Coulomb potential, as drawn above, also reflects the greater range of movement available to the electron as r increases.

When an electron is bound by a potential, it is forced into having discrete allowable energies. That is, the allowable states $E < 0$ are **bound states**. Above this level, the electron is free and can have any energy. This is indicated by the grey band in the figure above. Note that in this potential, the energy depends on the principle quantum number n. As n increases, the energy becomes less negative and is therefore at a higher potential.

1.7 Covalent Bond

Consider a **hydrogen atom**. Here we have one electron orbiting the nucleus. In the ground state, the electron is in one of the available 1*s* energy states.

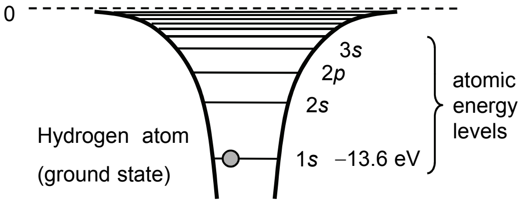
0
3s
2p
2s
Hydrogen atom
(ground state)
1s −13.6 eV
atomic energy levels

When a hydrogen atom comes into proximity with another hydrogen atom, we might be tempted to draw the energy levels like this:

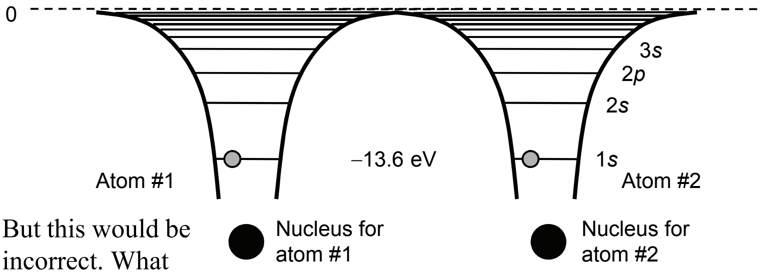
0
3s
2p
2s
−13.6 eV
1s
Atom #1
Atom #2
Nucleus for atom #1
Nucleus for atom #2

But this would be incorrect. What happens is that the electron in atom #2 is attracted by the positively charged nucleus of atom #1 and vice versa. Overall, the nucleus from atom #1 is attracted to both electrons. The nucleus from atom #2 is attracted to the same two electrons. Therefore, the two nuclei behave as if they were bonded together. The ***co***-sharing of these ***valence*** electrons and the resulting attraction of the two atoms is called a ***covalent*** **bond**.

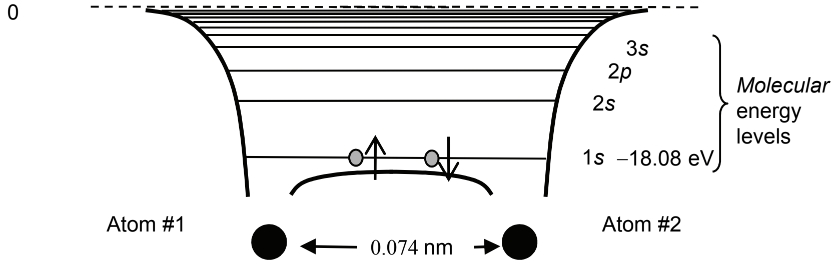
0
3s
2p
2s
1s −18.08 eV
Molecular energy levels
Atom #1
Atom #2
0.074 nm

When this attraction occurs, the sides of the "well" are reduced so the extent that the electrons are completely shared between the two atoms. The electrons (by the **Pauli exclusion principle**) have to have different spins ↑↓. Since this is energetically favourable (electrons have more space to move), then hydrogen naturally forms the **molecule** H_2. When a covalent bond is formed, energy is released (heat). To break the bond and separate the atoms, energy –(18.08–13.6) = –4.48 eV has to be supplied.

1.8 Ionisation Energy

In order for an electron to make a transition from a lower state to a higher state in an atom, it must absorb energy (via collision with a photon of sufficient energy, by heat, etc). The energy required to move the outermost (highest potential energy) electron in an isolated atom from its ground state to infinity is called the **first ionisation energy** and is usually expressed in units of eV. The resulting positively charged atom is called an **ion**.

The energy required to move the next outermost bound electron from its **ground state** to infinity is called the second ionisation energy, and so on.

Sometimes ionisation energies are expressed as kJ per mole. Since 1 eV is 1.602×10^{-19}J, then 1 eV per atom is 96.49 kJ per mole.

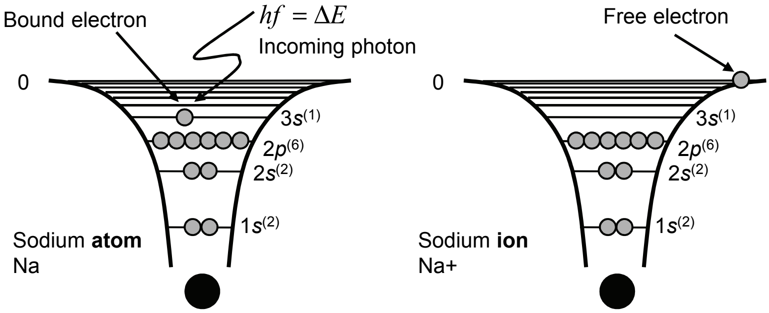

The incoming photon must have sufficient energy (in this example, > +5.1 eV for Na) to lift the electron to infinity distance from the nucleus for the atom to be ionised. If it does not, and if the incoming photon has sufficient energy to lift the electron to a higher energy level, then the atom may become "**excited**" rather than "**ionised**". If the incoming photon has insufficient energy to excite or ionise an atom, then it will just pass through or be scattered by the atom.

Important: We cannot just assign energies to these levels by calculation like we did with the hydrogen atom. The energies associated with each level in a multi-electron atom depend on the size of the atom, the atomic number, the degree of screening of outer electrons by the inner electrons, whether or not the atom is bonded with another, or exists in the gaseous phase, and most importantly, whether the level is filled or not. The ionisation energies are usually measured experimentally. The formula worked OK for hydrogen because we were dealing with just a single electron.

1.9 Electron Affinity

Consider a neutral chlorine atom. The first ionisation energy for Cl is 13 eV.

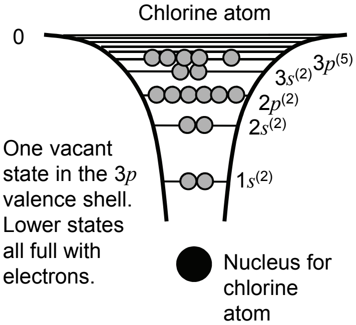
Chlorine atom
0
3s(2) 3p(5)
2p(2)
2s(2)
1s(2)
One vacant state in the 3p valence shell. Lower states all full with electrons.
Nucleus for chlorine atom

What happens when a free electron comes into contact with the neutral atom? The electron may occupy a vacancy in the 3*p* level, thus endowing the atom with a full outer energy shell. This condition is energetically desirable and is called a **noble gas configuration**. When the electron is bought from infinity into the 3*p* shell, its potential energy is lowered and experiments show that −3.7 eV is released (perhaps as heat). This energy that is released is called the **electron affinity** of the atom.

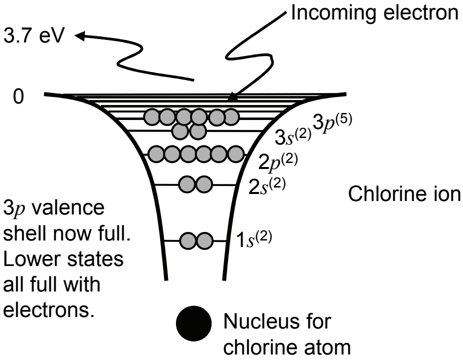
Incoming electron
3.7 eV
0
3s(2) 3p(5)
2p(2)
2s(2)
Chlorine ion
3p valence shell now full. Lower states all full with electrons.
1s(2)
Nucleus for chlorine atom

The **electron affinity** is the energy released when an electron is bought from infinity into a neutral atom. Like **ionisation energy**, it is measured experimentally (although usually indirectly). The incoming electron has to overcome the repulsion of the electrons already there, but, in the case of elements like Cl, the attainment of a **noble gas configuration** is a sufficient payoff.

1.10 Ionic Bond

Ionic bonds usually form between elements that have unpaired valence electrons. Consider the reaction between sodium and chlorine.

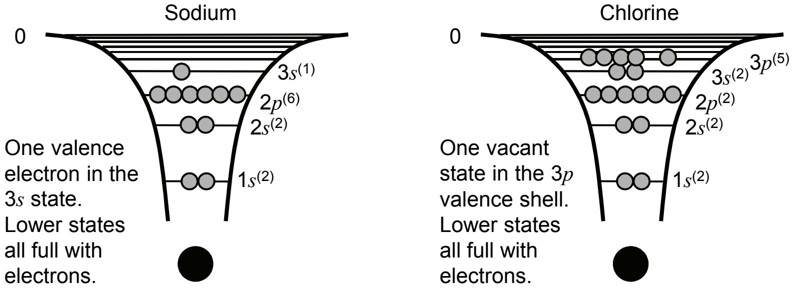

One valence electron in the 3*s* state. Lower states all full with electrons.

One vacant state in the 3*p* valence shell. Lower states all full with electrons.

The ionisation energy for Na is +5.1 eV. The electron affinity for Cl is –3.7 eV. Thus, if the lone electron in the 3*s* band in Na can climb over the energy barrier (–5.1 – –3.7 = –1.4 eV), then this will create an Na^+ sodium **ion** and a Cl^- chlorine ion, leaving each with a net charge q_e. The resulting electrostatic (Coulomb) attraction is called an **ionic bond**.

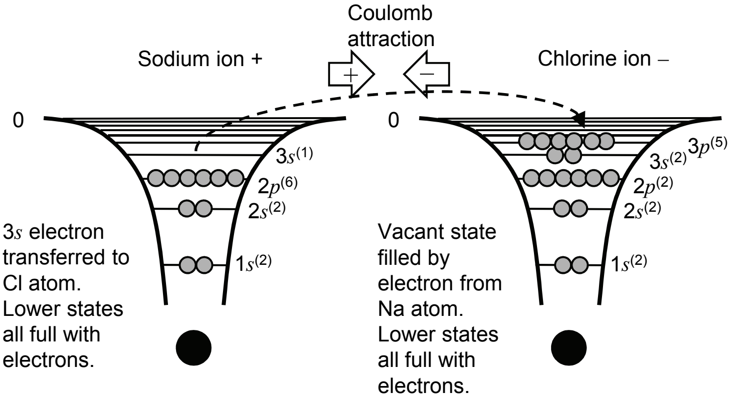

3*s* electron transferred to Cl atom. Lower states all full with electrons.

Vacant state filled by electron from Na atom. Lower states all full with electrons.

When the ions are formed, the attraction between them causes them to move towards each other and the electrical potential between them drops. The ions reach an **equilibrium distance** determined by the electrostatic attraction of their overall charge and the repulsion offered by their positively charged nuclei. Experiments show that the **bond energy** is –5.5 eV. This energy is lower than the –1.4 eV barrier and so an ionic bond between these two atoms is energetically favourable (since atomic systems tend settle to a state of minimum energy).

1.11 Electronegativity

It is energetically favourable for an atom to have completely filled outer energy shells (**noble gas configuration**). Metals tend to lose electrons to achieve this configuration because their few loosely bound outer shell valence electrons are easily removed (low **ionisation energy**). Non-metals, on the other hand, may have only a few vacancies in their outer shells and so prefer to gain electrons to attain a noble gas configuration.

When two atoms come together to form a **chemical bond**, the ability for one atom to gain, or pull, an electron from the other atom depends on both the ionisation energies and the electron affinities of the two atoms. A combined property that includes these two characteristics is called the **electronegativity** of the element.

Here's how it works in simplified terms. Hydrogen and chlorine can combine to form HCl by the formation of a covalent bond. An electron pair is shared between the H atom and the Cl atom. But is this sharing equal? The ionisation energy for H is +13.6 eV, and for Cl is +13 eV. The electron affinity of H is −0.75 eV and for Cl is −3.7 eV. Thus, for an electron to be transferred from Cl to H, a net energy of $13 - 0.75 = +12.25$ eV is required. For an electron to be transferred from H to Cl, a net energy of $13.6 - 3.7 = +9.9$ eV is required. Thus, in this covalent bond, the shared electron is biased towards being over near the Cl atom because less energy is required to transfer the electron from H to Cl compared to Cl to H. This unequal sharing makes the bond **polar** (since one end, the Cl end, has a net negative charge). Although we call the bond "covalent" it does have an ionic character as well. In general, there is a gradation of bond types from ionic to covalent depending on the nature of the atoms.

The **electronegativity** describes the relative ability for an atom, when it combines with another atom, to become "more negative" by more strongly capturing a shared electron or electron pair. It is measured by a variety of means, but at its simplest, depends on both the ionisation energy and the electron affinity of the atom. Electronegativities (no units) range from 4 (for Fl on the right-hand side of the periodic table) down to <1 (for elements on the left-hand side of the periodic table).

Two atoms with very different electronegativities are expected to form **ionic bonds**. Atoms with much the same electronegativities are expected to form **covalent bonds**.

An example is when an Na atom (with electronegativity 0.9) meets a Cl atom (electronegativity 3.0). The bond between them is expected to be predominantly ionic, with the chlorine atom becoming more negative. When H (with electronegativity 2.1) meets a Cl, the bond is expected to be predominantly covalent.

1.12 Metallic Bond

Consider what happens when we have two Li atoms close together. We might be first tempted to draw the potentials as:

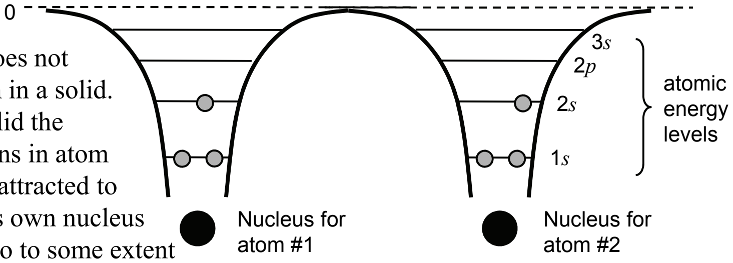
0
3*s*
2*p*
2*s*
1*s*
atomic energy levels
Nucleus for atom #1
Nucleus for atom #2

This does not happen in a solid. In a solid the electrons in atom #1 are attracted to both its own nucleus and also to some extent by the nucleus of atom #2. For example, the 1*s* electron orbiting the nucleus of atom #1 is also captured to some extent by the nucleus of atom #2. But, due to the **Pauli exclusion principle**, for each atom, we cannot have more than two electrons in the same energy level (i.e. the two 1*s* electrons for atom #1 as seen by atom #2 are no longer permitted to have energies at the 1*s* level because there are already two electrons from its own atom at that level).

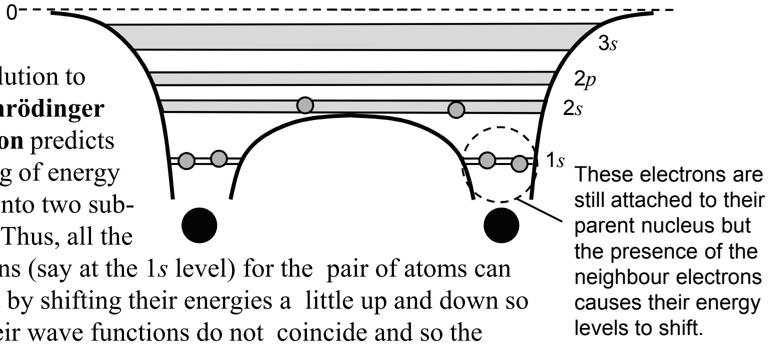
0
3*s*
2*p*
2*s*
1*s*
These electrons are still attached to their parent nucleus but the presence of the neighbour electrons causes their energy levels to shift.

The solution to the **Schrödinger equation** predicts splitting of energy levels into two sub-levels. Thus, all the electrons (say at the 1*s* level) for the pair of atoms can coexist by shifting their energies a little up and down so that their wave functions do not coincide and so the exclusion principle is satisfied. When there are many atoms present, the splitting of many levels into fine gradations creates a **band** of energies and if the energy barrier between the atoms is low enough for the upper bands to carry across to meet those of neighbouring atoms, the electrons in these bands become free to migrate from atom to atom. The electrons behave as if they were in a **potential well**. Because of the large range of movement of these electrons (compared to if they were still attached to their parent nuclei), their kinetic energy is lower (see Section 1.5) and a lowering of energy equates to stability – that is, the free electrons in the valence band reduce the energy of the system and so act to hold the two atoms together. This is called a **metallic bond.**

1.13 Electronic Structure of Solids

In a solid, the interacting potentials of many millions of relatively closely spaced atoms causes atomic energy levels to split into a very large number of sub-levels. The energy difference between each sub-level is so fine that each molecular level is considered to be virtually a continuous **band** of energies.

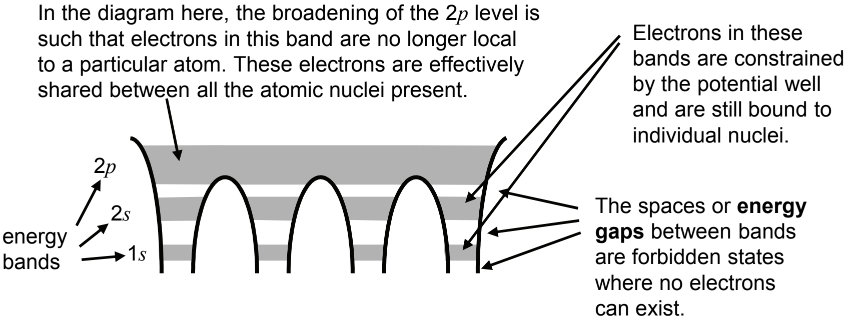

In the diagram here, the broadening of the 2*p* level is such that electrons in this band are no longer local to a particular atom. These electrons are effectively shared between all the atomic nuclei present.

Electrons in these bands are constrained by the potential well and are still bound to individual nuclei.

The spaces or **energy gaps** between bands are forbidden states where no electrons can exist.

If the highest energy band that contains electrons in the ground state (the **valence band**) in a solid is not completely full, then electrons within that band can easily move around from state to state *within the band*. Such movement can be readily obtained by applying an electric field to the solid. Such solids are thermal and electrical **conductors**.

If the valence band in a solid is full, and the next highest available band is positioned some distance away in terms of its energy levels, then the electrons within the topmost band cannot easily move from place to place or to the next highest band. Such materials are thermal and electrical **insulators**.

If the next highest available band is positioned fairly closely to the valence band, then even at room temperature, there may be sufficient thermal energy given to some electrons to be promoted to this higher level. The material becomes conducting and is a **semiconductor**. The band containing the conducting electrons is called the **conduction band**. In a conductor, the valence band is the conduction band. In a semiconductor, the conduction band (at 0K) is separated from the valence band (defined at 0K) by an **energy gap**.

Atoms in a solid generally form molecules which either arrange themselves in a regular pattern (**crystalline solids**) or the molecules do not repeat themselves in an orderly way (**amorphous solids**). When a solid forms, atoms are pushed and pulled around and settle into place when the interaction electron potentials reach a minimum level.

2. Chemical Components

Summary

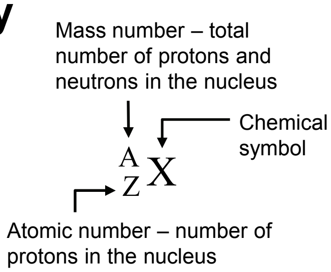
Mass number – total number of protons and neutrons in the nucleus
Chemical symbol
A
Z
X
Atomic number – number of protons in the nucleus

One atomic mass unit (amu) is 1/12th the mass of a carbon 12 atom. 1 amu = 1.6602×10^{-27} kg.

Avogadro's number: 6.022×10^{23} = 1 mole.

Atoms that lose or gain electrons are called ions:

- Cations (+) (electrons lost)
- Anions (−) (electrons gained)

A molecule is the smallest collection of atoms that is electrically neutral and can exist as a separate identifiable unit.

The molecular weight of a substance is the sum of the atomic weights of its constituents.

Stoichiometry is the process of accounting for the masses of atoms, molecules and compounds in chemical reactions.

2.1 Matter

Elements: cannot be reduced to simpler units by physical processes or chemical reactions. (Examples are oxygen, iron, carbon.) Generally, elements are substances which have only one kind of atom (as having a certain number of protons on the nucleus). There are just over 100 elements known, with about 95% of these occurring naturally, the remainder being created synthetically by nuclear processes.

Elements can be broadly divided into non-metals, semi-metals and metals.

Mixtures: can be separated into constituent parts by physical means. Mixtures are a product of mechanical or physical processes, such as metal alloys, suspensions, dispersions and colloids. **Homogenous** mixtures have uniform composition and appearance (examples are air, salty water) over a molecular length scale while **heterogeneous** mixtures have physically distinct regions. (Examples are fruit cake, concrete, fruit juice with pulp.)

Compounds: combinations of two or more elements in definite proportions where the composition of the compound is uniform throughout. Compounds have a chemical structure, the atoms being held in place by chemical bonds. Compounds keep their chemical identity when altered physically, and can only be reduced to their constitutive elements by chemical reactions. (Examples are water, salt, sulphuric acid.)

2.2 Atomic Weight

Except in the case of nuclear transformations, we can usually treat atoms as being much like the Bohr atom. Atoms are characterised by an **atomic number**, a **chemical symbol**, and the **mass number**.

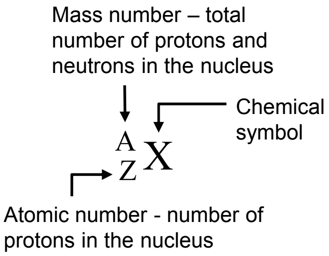
Mass number – total number of protons and neutrons in the nucleus
Chemical symbol
A
Z
X
Atomic number - number of protons in the nucleus

$m_e = 9.1096 \times 10^{-31}$ kg

$m_p = 1.6726 \times 10^{-27}$ kg

$m_N = 1.6749 \times 10^{-27}$ kg

The mass number is approximately equal to the atomic weight of the element. The **atomic weight** of an element is the mass (in relative terms) of an "average" atom of an element. The more precise term **atomic mass** is used to denote the absolute mass of a particular atom (in kg).

Since the **mass of an electron** m_e is very small nearly all of the mass of an atom is contributed by the **protons** and **neutrons** in the nucleus. Protons and neutrons have very nearly the same mass, so the number of neutrons in an atom is given by Z – A. However, it is found in nature that many atoms of the same element with atomic number Z have different mass numbers A because of having a different number of neutrons in the nucleus. Each type of atom A of an element is called a **nuclide**, or **isotope**.

By international agreement, one **atomic mass unit** (**amu**) is 1/12th the mass of a single carbon 12 atom. The atomic weight (or **relative atomic mass**) of an element is the ratio of the average mass per atom of the element to 1/12 of the mass of an atom of ^{12}C.

Note: The atomic weight of an element is not found from adding the masses of protons, neutrons and electrons. When an atom is formed, some of the mass is used as nuclear binding energy via the **Einstein** relationship $E = mc^2$.

The **carbon 12** isotope has 6 protons and 6 neutrons in the nucleus. The **carbon 13** atom has 6 protons and 7 neutrons in the nucleus. Both isotopes have 6 electrons. Experiments show that carbon in nature consists of 1.11% ^{13}C and 98.89% ^{12}C. The atomic weight of carbon is determined to be 12.011.

In chemistry, much of the arithmetic of chemical reactions involves the atomic weights for the atoms that take part. However, the actual nature of the chemical reactions is due to the number and arrangement of the outer-shell **valence electrons** in the atoms rather than the heavy nuclei. "Valence" is a word that means "power" and in some sense, it is the number and arrangement of valence electrons that give an atom "combining power" to form **chemical bonds** with other atoms.

2.3 Ions

Atoms that lose or gain electrons are called **ions**. When an atom loses one or more electrons, it is called a **cation**. When an atom gains one or more electrons, it is called an **anion**.

Examples of ions:

Atom	Z (atomic number)	No. electrons gained or lost	No. electrons remaining	Formula of ion
H	1	-1	0	H^+
Na	11	-1	10	Na^+
Ca	20	-2	18	Ca^{2+}
F	9	+1	10	F^-
S	16	+2	18	S^{2-}

When an ion is formed, the atom is no longer electrically neutral. The formula of the ion signifies the net electronic charge. Atoms may lose electrons, especially unpaired outer shell or valence electrons, by physical removal, absorption of **ionising radiation**, the electron being taken by another type of atom which has a stronger attraction for it, and so on.

Metallic elements tend to lose one or more electrons when they combine with other elements and so form cations. Non-metallic elements tend to gain electrons to form anions. When naming compounds, it is usual to list the more metallic element first. In **binary ionic compounds**, we list the metal cation first followed by the non-metal anion with an "ide" suffix. For covalent compounds, the more metallic element is listed first.

Sodium chloride
Carbon monoxide }
Carbon dioxide }
Nitrogen oxide

Where there is more than one possible compound, prefixes are used to distinguish them.

2.4 Molecules

Most elements bond with others to form **molecules**. A molecule is the smallest collection of atoms, that is electrically neutral, and can exist as a separate identifiable unit. The concept was first proposed by **Avogadro** in 1811, who, on the basis of observations made by Dalton and Gay-Lussac proposed that atoms combine to form molecules, *and* at the same temperature and pressure, equal volumes of *all* gases contain the same number of molecules. **Avogadro's law** provides the justification for a method of determining relative molecular weights

because there was now a way to obtain equal numbers of different molecules by using different gases all at the same temperature, pressure and volume.

A molecule is most conveniently described in terms of a **molecular formula**.

H_2O

A water molecule consists of two H atoms and one O atom.

In some cases, such as in a crystalline solid, the concept of a molecule is not appropriate because the atoms which make up the substance are arranged in a regular repeating pattern. In this case, we speak of the **empirical formula**.

NaCl

A sodium chloride crystal consists of equal numbers of sodium and chlorine atoms arranged in a regular array or crystal lattice.

The **molecular weight** (or **relative molar mass**, or **relative molecular mass**) M_r of a substance is the sum of the atomic weights of its constituents.

If we had a certain mass of an element (or a molecule), say 12 grams of ^{12}C, how many atoms would this be? This number is called **Avogadro's number** N_A and has the value 6.0220943×10^{23}. N_A is found by experimental methods.

By international agreement, one **atomic mass unit** (**amu**) is 1/12th the mass of a carbon 12 atom.

12 g of ^{12}C = 6.0220943×10^{23} atoms; therefore the mass of one ^{12}C atom is:

"But!", you might say, "How can you have 6.022... atoms if atoms are indivisible? Shouldn't this be a whole number?" Look at the 10^{23}. If you write Avogadro's number out in full, it *is* a whole number.

$$\begin{aligned}\text{mass} &= \frac{0.012}{6.02\times10^{23}}\,\text{kg}\\ &= 1.99\times10^{-23}\,\text{kg}\end{aligned}$$

It is convenient when working with quantities usually involved in chemical reactions to round down Avogadro's number to 6.02×10^{23}.

2.5 Mole

Chemical reactions usually involve large numbers of atoms and molecules. It is convenient to have a unit of measure that relates weight (i.e. mass in grams or kilograms, etc) of bulk chemicals with atomic or molecular weight. By definition, this unit of measure is called a **mole**, and it is the Avogadro number of atoms or molecules.

12 grams of ^{12}C contains 6.02×10^{23} atoms and is called 1 mole of ^{12}C

18 grams of H_2O also contains 6.02×10^{23} molecules and is called 1 mole of H_2O

The word "mole" is a shorthand way of saying "**Avogadro's number**". That is, it is easier to say "Consider one mole of sodium atoms" rather than "Consider one Avogadro's number of sodium atoms" or "Consider 6.02×10^{23} of sodium atoms". To find out the mass in grams of a mole of some atoms, we simply look up the **atomic weight**. To find out the mass in grams of a mole of some molecules, we simply use the molecular weight.

→ because the atomic weight (in amu) is expressed relative to the ^{12}C atom, and thanks to Avogadro, we know that 12 g of ^{12}C has N_A atoms.

Moles, why bother? It's convenient. One mole of *anything* is 6.02×10^{23} of the "anythings". When we are talking about atoms or molecules, we know that one mole has a mass equal to the atomic weight of the element, or the molecule because that's the way Avogadro's number was determined. It's the link between the *relative* atomic mass and the *actual* mass in grams.

Units of chemical accounting:

- One **atomic mass unit** (amu) is 1/12th the mass of a carbon 12 atom by international agreement. 1 amu is $1.66020943 \times 10^{-27}$ kg.
- The **atomic weight** (or **relative atomic mass**) of an element is the ratio of the average mass per atom of the element to 1/12 of the mass of an atom of ^{12}C.
- The total sum of the atomic weights for a molecule of substance is called the **molecular weight** or **relative molar mass** M_r.
- The molecular/atomic weight expressed in grams contains one mole of molecules/atoms. This is Avogadro's number 6.0220943×10^{23}.
- The **molar mass** M of an element is the atomic weight of the element expressed in g/mol. The molar mass of a compound is the sum of the atomic weights of the constituent elements expressed in g/mol.

2.6 Compounds

Compounds consist of two or more different elements which are chemically combined in a definite fixed ratio (as distinct from mixtures, which can have any proportions of ingredients). There are generally two types of compounds:

Molecular compounds are composed of atoms held together by chemical bonds in fixed proportion and are electrically neutral overall. For example, the water molecule H_2O is a molecular compound and the **molecular formula** shows the proportions of the atomic species within it.

Name	Molecular formula
Methane	CH_4
Carbon dioxide	CO_2
Ammonia	NH_3
Acetylene	C_2H_2
Ethylene	C_2H_4
Ethane	C_2H_6
Water	H_2O
Benzene	C_6H_6
Ethanol	C_2H_5OH
Naphthalene	$C_{10}H_8$
Aspirin	$C_9H_8O_4$

There is a distinction to be made between the **molecular formula** and the **empirical formula** of a compound. The empirical formula gives the ratio of component atoms in the lowest possible numerical terms. The empirical formula shows the ratio of component atoms as found to exist in practice. In many cases, the formulae are the same, but this cannot be always assumed. For example, the empirical formula for benzene is CH but the molecular formula is C_6H_6.

Ionic compounds are composed of charged atoms, or a charged group of atoms, of opposite sign that are held together by **ionic bonds**. When an ionic compound is formed from the joining of a charged group of atoms, it is said to be a **complex ion** (or a polyatomic ion, or a molecular ion). The ion with the positive charge in the compound is the **cation**. The ion with the negative charge is the **anion**. The ionic formula gives the ratio of anions and cations. The ionic formula is electrically neutral overall.

Name	Ions formed	Formula
Magnesium chloride	Mg^{2+} Cl^-	$MgCl_2$
Silver sulphate	Ag^+ SO_4^{2-}	Ag_2SO_4
Ammonium sulphate	NH_4^+ SO_4^{2-}	$(NH_4)_2SO_4$
Chromium (III) hydroxide	Cr^{3+} OH^-	$Cr(OH)_3$
Sodium chloride	Na^+ Cl^-	$NaCl$

2.7 Cations

+1

Name	Formula
Hydrogen	H^+
Lithium	Li^+
Sodium	Na^+
Potassium	K^+
Rubidium	Rb^+
Caesium	Cs^+
Silver	Ag^+
Copper	Cu^+
Mercury (I)	Hg_2^{2+}
Ammonium	NH_4^+

+2

Name	Formula
Beryllium	Be^{2+}
Magnesium	Mg^{2+}
Calcium	Ca^{2+}
Strontium	Sr^{2+}
Barium	Ba^{2+}
Lead	Pb^{2+}
Zinc	Zn^{2+}
Cadmium	Cd^{2+}
Nickel	Ni^{2+}
Manganese	Mn^{2+}
Tin (II)	Sn^{2+}
Iron (II)	Fe^{2+}
Mercury (II)	Hg^{2+}
Cobalt (II)	Co^{2+}
Chromium (II)	Cr^{2+}
Copper (II)	Cu^{2+}

+3 , +4

Name	Formula
Aluminium	Al^{3+}
Bismuth	Bi^{3+}
Iron (III)	Fe^{3+}
Cobalt (III)	Co^{3+}
Chromium (III)	Cr^{3+}
Tin (IV)	Sn^{4+}

In the naming of chemical compounds, it is customary to write the name of the most metallic element first, followed by the more non-metallic one. The second element or ion is given the *ide* suffix (e.g. sodium chloride).

Prefixes

1 – mono
2 – di
3 – tri
4 – tetra
5 – penta
6 – hexa

2.8 Anions

−1

Name	Formula
Fluoride	F^-
Chloride	Cl^-
Bromide	Br^-
Iodide	I^-
Hydroxide	OH^-
Nitrite	NO_2^-
Nitrate	NO_3^-
Chlorate	ClO_3^-
Perchlorate	ClO_4^-
Cyanide	CN^-
Permanganate	MnO_4^-
Thiocyanate	NCS^-
Bicarbonate	HCO_3^-
Bisulphate	HSO_4^-
Dihydrogen phosphate	$H_2PO_4^-$
Acetate	$CH_3CO_2^-$

−2

Name	Formula
Oxide	O^{2-}
Peroxide	O_2^{2-}
Sulphide	S^{2-}
Sulphate	SO_4^{2-}
Sulphite	SO_3^{2-}
Thiosulphate	$S_2O_3^{2-}$
Carbonate	CO_3^{2-}
Oxalate	$C_2O_4^{2-}$
Chromate	CrO_4^{2-}
Dichromate	$Cr_2O_7^{2-}$
Monohydrogen phosphate	HPO_4^{2-}

−3 , −4

Name	Formula
Nitride	N^{3-}
Phosphate	PO_4^{3-}
Hexacyanoferrate (III)	$Fe(CN)_6^{3-}$
Hexacyanoferrate (II)	$Fe(CN)_6^{4-}$

For polyatomic anions, the assignment of a suffix depends on the **oxidation number** (see Section 4.8) of the central non-metal atom.

Suffixes	
ide	
ite	ous
ate	ic

In some cases, the oxidation number of the metal is stated in roman numerals to distinguish compounds which would otherwise have the same name (e.g. iron (II) chloride and iron (III) chloride).

2.9 Chemical Equation

When atoms or molecules (the **reactants**) combine with other atoms or molecules to form compounds (the **products**), a **chemical reaction** is said to take place. Chemical reactions involve the breaking and formation of **chemical bonds**.

$$\text{Reactants} \rightarrow \text{Products}$$

This is the forward reaction

$$\text{Products} \rightarrow \text{Reactants}$$

This is the reverse reaction

Often one can tell if a chemical reaction has occurred by observation of products that are considerably different in appearance than the reactants. Examples are the formation of a **precipitate**, evolution of gas, change of colour, production of heat and so on. Often, chemical reactions proceed spontaneously, other times they have to be forced to proceed by the input of energy or the lowering of energy barriers.

Chemical equations are written using chemical formulae where the number of atoms or molecules is written before the formula:

$$2CO + O_2 \rightarrow 2CO_2$$

This equation says that two molecules of carbon monoxide react with one molecule of oxygen to product two molecules of carbon dioxide. This is a **balanced equation**, because the number of atoms in each reactant equals the number of atoms in each product. That is, we have two carbon atoms on the left-hand side and two carbon atoms on the right-hand side, and two oxygen atoms on the left side and two on the right side. We can also say that two **moles** of carbon monoxide combine with one mole of oxygen to form two moles of carbon dioxide. Since the mass of a mole of an atom or a molecule is the atomic or molecular weight expressed in grams, a balanced equation is balanced in terms of both number of atoms and mass on both sides of the equation. The net electric charge must also be the same on both sides of the equation.

Usually, reactions proceed until there is **chemical equilibrium**. Dynamic chemical equilibrium occurs when the rate of the forward reaction becomes equal to the rate of the reverse reaction:

$$\text{Reactants} \leftrightarrow \text{Products}$$

Dynamic equilibrium

By rate, we mean the rate of formation of products in units of mol/s.

2.10 Stoichiometry

Stoichiometry is the process of accounting for the masses of atoms, molecules and compounds in chemical reactions. The basic component of such accounting is the arrangement of chemical formulae in chemical equations. Stoichiometric calculations are performed in a logical sequence:

1. Write a balanced equation for the reaction.
2. Convert known masses of reactants and/or products into moles.
3. Calculate the number of moles of the reactants and/or products whose masses are unknown.
4. Calculate the masses of the unknown reactants and/or products.

Balancing an equation involves both a balance of atoms and also electric charge. This is particularly important in the case of ionic equations where spectator ions may be present and do not take part in the reaction. Some examples of unbalanced equations (with spectator ions) and the corresponding **net ionic equations** are:

$$\mathrm{Pb(NO_3)_2(aq) + KCl(aq) \rightarrow PbCl_2(s) + K^+ + NO^{3-}}$$

$$\mathrm{Pb^{2+}(aq) + 2Cl^-(aq) \rightarrow PbCl_2(s)}$$

$$\mathrm{Ni(OH)_2(s) + HNO_3(aq) \rightarrow NiNO_3(aq) + H_2O}$$

$$\mathrm{Ni(OH)_2(s) + 2H^+ \rightarrow Ni^{2+} + 2H_2O}$$

$$\mathrm{NaCO_3(aq) + HNO_3(aq) \rightarrow NaNO_3(aq) + H_2O + CO_2}$$

$$\mathrm{CO_3^{2-} + 2H^+ \rightarrow 2H_2O + CO_2}$$

Practical difficulties can arise when attempting to determine masses of reactants and products. For example, some reactants may be present in excess (more present than combine with one or other reactants). More than one reaction may take place at the same time, or a reaction may proceed in a series of intermediate steps. Products (such as gases) may escape and alter the equilibrium state of the reaction. The nature of the reactants may prevent mixing of the reactants and so there is incomplete yield of product.

2.11 Example

Acetylene gas is used together with oxygen in an “oxy” welding set. It is prepared from a reaction of calcium carbide and water:

$$\mathrm{CaC_2 + H_2O \rightarrow Ca(OH)_2 + C_2H_2} \quad \text{Unbalanced eqn.}$$

Balance this equation and determine the mass of acetylene that is produced from the reaction of 130 g CaC_2 and 100 g water.

$$\mathrm{CaC_2 + 2H_2O \rightarrow Ca(OH)_2 + C_2H_2} \quad \text{Balanced eqn.}$$

$$\begin{aligned}
\mathrm{CaC_2} &= 130\ \mathrm{g} \\
&= \frac{130}{40.1 + 2(12)} \\
&= 2.03\ \text{moles} \\
\mathrm{H_2O} &= 100\ \mathrm{g} \\
&= \frac{100}{2 + 16} \\
&= 5.56\ \text{moles}
\end{aligned}$$

But according to the balanced equation, for each mole of CaC_2, two moles of H_2O are required. Therefore, 2(2.03) = 4.06 moles of H_2O required, and so the mass of C_2H_2 formed would be:

$$\begin{aligned}
\mathrm{C_2H_2} &= 2.03\ \text{moles} \\
2.03 &= \frac{m}{2(12) + 2(1)} \\
&= 52.8\ \mathrm{g}
\end{aligned}$$

3. The Periodic Table

Summary

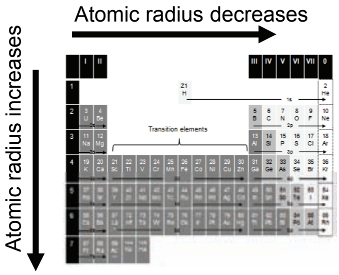

Atomic radius decreases

Atomic radius increases

Bohr radius

$$r = n^2\left(\frac{\varepsilon_o h^2}{\pi m_e Z q_e^2}\right)$$

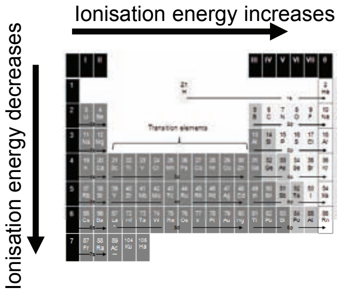

Ionisation energy increases

Ionisation energy decreases

Noble gases

He	$1s^2$
Ne	$He2s^22p^6$
Ar	$Ne3s^23p^6$
Kr	$Ar3d^{10}4s^24p^6$
Xe	$Kr4d^{10}5s^25p^6$
Rn	$Xe5d^{10}6s^26p^6$

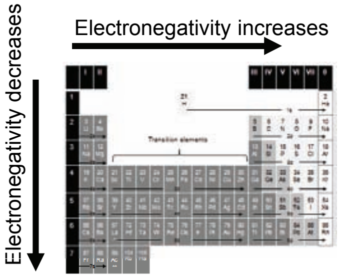

Electronegativity increases

Electronegativity decreases

3.1 Electron Configuration

In a **multi-electron atom**, the electrons do not find themselves in a **Coulomb potential** of the form $1/r$ as in the **hydrogen atom** due to the **screening effect** of inner-shell electrons on the outer-shell electrons. In a multi-electron atom, the potential in which an electron finds itself (and hence its energy) depends on not only n (as in the single-electron atom), but also the second quantum number l. The arrangement of electrons in atoms in the **ground state** can be complicated, but follows certain rules:

1. Electrons occupy the lower energy levels first, before occupying higher levels.
2. Electrons at a particular energy level have opposite spin (**Pauli exclusion principle**).
3. Electrons tend to occupy an energy level as single electrons before pairing up (**Hund's rule**).

Examples of the ground-state electron configuration of the first few elements arranged in order of increasing **atomic number** Z

Hydrogen	$1s^1$	
Helium	$1s^2$	
Lithium	$1s^2 2s^1$	
Beryllium	$1s^2 2s^2$	
Boron	$1s^2 2s^2 2p^1$	
Carbon	$1s^2 2s^2 2p^2$	
Nitrogen	$1s^2 2s^2 2p^3$	
Oxygen	$1s^2 2s^2 2p^4$	
Fluorine	$1s^2 2s^2 2p^5$	
Neon	$1s^2 2s^2 2p^6$	
Sodium	$1s^2 2s^2 2p^6 3s^1$	
Magnesium	$1s^2 2s^2 2p^6 3s^2$	
Aluminium	$1s^2 2s^2 2p^6 3s^2 3p^1$	
Silicon	$1s^2 2s^2 2p^6 3s^2 3p^2$	
Phosphorus	$1s^2 2s^2 2p^6 3s^2 3p^3$	
Sulphur	$1s^2 2s^2 2p^6 3s^2 3p^4$	
Chlorine	$1s^2 2s^2 2p^6 3s^2 3p^5$	
Argon	$1s^2 2s^2 2p^6 3s^2 3p^6$	
Potassium	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^1$	Note $4s$ has lower energy than $3d$ in K.
Calcium	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^2$	Note $4s$ has lower energy than $3d$ in Ca.
Scandium	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^1 4s^2$	Keep sub-shells together even if no longer in increasing energy order.

Electrons spread out to occupy unfilled sub-shells before pairing up. For example, in oxygen, the arrangement of electrons is:

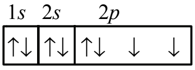

When we write the electron configuration of elements, we keep the principal quantum numbers together even if they are not arranged in energy order. For example, in potassium, the $4s$ sub-shell actually has a lower energy than the $3d$ sub-shell, but we usually write the $3d$ along with the $3s$ and $3p$ sub-shells.

3.2 Periodic Law

In the nineteenth century, it was noticed that if elements were arranged in order of increasing **atomic weight**, certain properties of the elements tended to occur at periodic intervals. This is a striking observation. It is now known that **periodicity** occurs when the elements are arranged in increasing **atomic number** Z. Consider the first few elements and the property of reactivity:

1 Hydrogen	$1s^1$
2 Helium	$1s^2$
3 Lithium	$1s^2 2s^1$
4 Beryllium	$1s^2 2s^2$
5 Boron	$1s^2 2s^2 2p^1$
6 Carbon	$1s^2 2s^2 2p^2$
7 Nitrogen	$1s^2 2s^2 2p^3$
8 Oxygen	$1s^2 2s^2 2p^4$
9 Fluorine	$1s^2 2s^2 2p^5$
10 Neon	$1s^2 2s^2 2p^6$
11 Sodium	$1s^2 2s^2 2p^6 3s^1$
12 Magnesium	$1s^2 2s^2 2p^6 3s^2$
13 Aluminium	$1s^2 2s^2 2p^6 3s^2 3p^1$
14 Silicon	$1s^2 2s^2 2p^6 3s^2 3p^2$
15 Phosphorus	$1s^2 2s^2 2p^6 3s^2 3p^3$
16 Sulphur	$1s^2 2s^2 2p^6 3s^2 3p^4$
17 Chlorine	$1s^2 2s^2 2p^6 3s^2 3p^5$
18 Argon	$1s^2 2s^2 2p^6 3s^2 3p^6$
19 Potassium	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^1$
20 Calcium	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^2$
21 Scandium	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^1 4s^2$

Another example of periodicity. Lithium is a highly reactive metal and so are sodium and potassium.

An example of periodicity. Helium is not very reactive (just about inert) and so are neon, and also argon.

Chemical reactivity is not the only property that tends to occur periodically. Other properties such as whether an element is a metal or a non-metal, whether they form certain compounds with other elements and so on can also occur periodically.

Initially it seems that the chemical reactivity of elements repeats in intervals of eight in atomic number Z. However, when the elements are arranged more completely we see that the periodicity actually depends upon the arrangement of outer-shell **valence electrons** – a not surprising observation in hindsight because it is the valence electrons that are involved in the formation of **chemical bonds**. For example, neon (10), argon (18), krypton (36), xenon (54) and radon (86) all have eight electrons in the outer shell and are inert. Na and K have one valence electron and easily form positive ions.

3.3 Periodic Table

Periodicity in chemical (and sometimes physical) properties can more readily be appreciated when elements are arranged in a table. Vertical columns in the table (**groups**) show elements with similar properties. Each horizontal row is called a **period**. The table also indicates (approximately) the way in which electron shells are occupied.

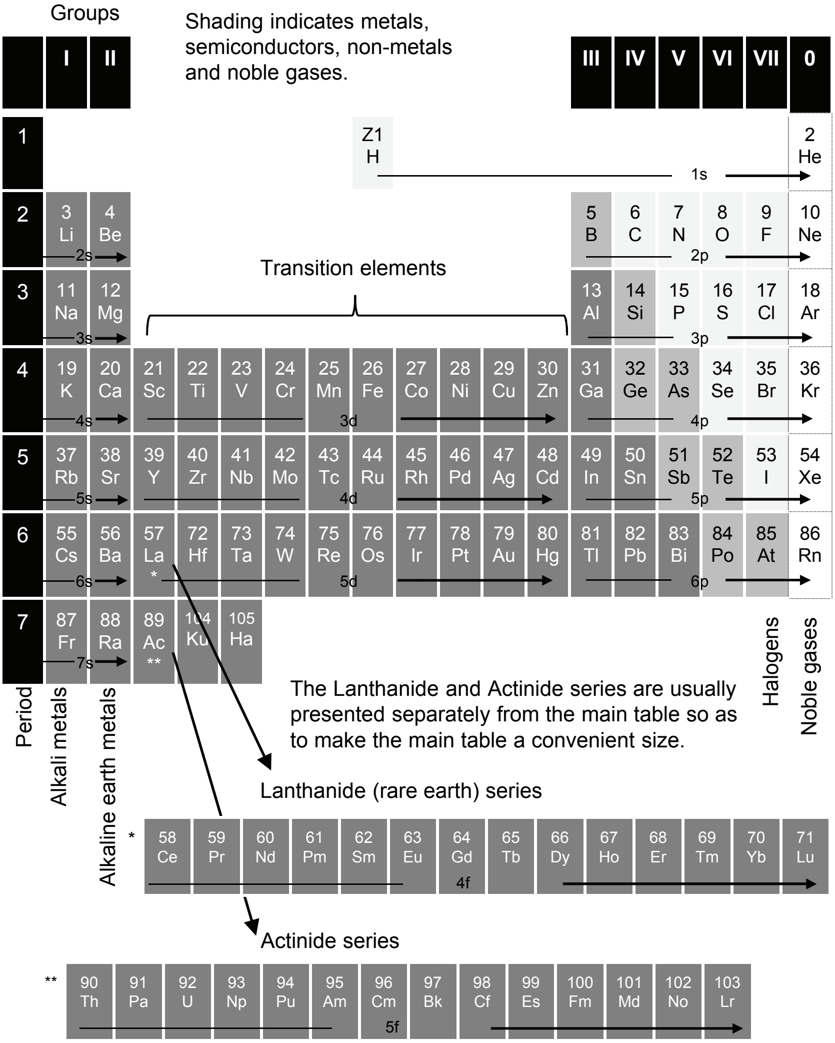
Groups
Period
Alkali metals
Alkaline earth metals
Transition elements
Halogens
Noble gases
1s
2s
2p
3s
3p
4s
3d
4p
5s
4d
5p
6s
5d
6p
7s
4f
5f
Lanthanide (rare earth) series
Actinide series

Shading indicates metals, semiconductors, non-metals and noble gases.

	I	II											III	IV	V	VI	VII	0
1									Z1 H									2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57 La *	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89 Ac **	104 Ku	105 Ha													

The Lanthanide and Actinide series are usually presented separately from the main table so as to make the main table a convenient size.

*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
**	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

3.4 Groups

Hydrogen is placed in a separate position in the periodic table.

Hydrogen has one electron, which in the ground state is in the 1*s* orbital. This unpaired electron makes hydrogen a reactive element that is able to form compounds with many other elements, including itself, to form a gas H_2.

${}^{1}_{1}\mathrm{H}$	Protium 99.9% abundance
${}^{2}_{1}\mathrm{H}$	Deuterium (heavy hydrogen)
${}^{3}_{1}\mathrm{H}$	Tritium (radioactive)

Group I elements are the **alkali metals**. Having one weakly bound outer valence electron, they are highly reactive, and metallic in character. Elements in this group usually form M^+ ions which are water soluble.

Group II elements are the **alkaline earth metals**. Elements in this group usually form M^{2+} ions and are generally insoluble, occurring naturally as silicates, carbonates, sulphates and phosphates.

The **transition elements** generally occupy the positions between Groups II and III. All these elements are metals since their outermost shells contain only a few electrons. However, unlike alkali metals, transition metals are hard, brittle and have a high melting point (with the exception of mercury).

Group III elements are also considered metals (although boron has only semi-metallic properties). With the exception of boron, they form M^{3+} ions and are relatively soft.

Group IV elements range from non-metal, semi-metals (or semiconductors) to metals down the group with increasing size of atom (and hence screening effect of electrons). All have four electrons in their outer shell. **Carbon** is responsible for the formation of hydrocarbons and derivatives, the basis of life on Earth. **Silicon**, unlike carbon, tends to form bonds with oxygen and is the basis for most of the minerals of the Earth.

Group V elements range from non-metallic to metallic down the group with increasing atomic size and five valence electrons.

Group VI elements show little metallic character due to the increasing ionisation potentials as we go across the periodic table. Increasing atomic size going down the group confers some metallic properties to Se, Te and Po.

Group VII elements are referred to as the halogens. They all have a high tendency to complete their electron shells by forming salts. These elements are largely non-metallic, and mostly reactive.

Group 0 elements are the **noble gases**, so called because of their inertness, although they are able to form compounds with oxygen and fluorine under certain conditions. Argon was discovered in 1894.

3.5 Energy Levels

The ordering of the outer electron shells in terms of energy is the basis for interpretation of the **periodic table**. Calculations show that the ordering of the energy shells proceeds in the following sequence from lowest (more negative) to highest (less negative potential):

1*s*, 2*s*, 2*p*, 3*s*, 3*p*, 4*s*, 3*d*, 4*p*, 5*s*, 4*d*, 5*p*, 6*s*, 4*f*, 5*d*, 6*p*, 7*s*, 5*f*, 6*d*

From 1*s* to 3*p*, it is easy to understand because the electron shells are at an increasing radial distance from the nucleus (*n* quantum number dominates the energy distribution just as in the one-electron atom). The *n* quantum number determines the distance the electrons are from the nucleus. The *l* quantum number determines the shape of the electron shell. When $l = 0$, we have an *s* sub-shell which is spherical. When $l = 1$, we have a p sub-shell which is lobed. $l = 2$ gives *d* sub-shells which are four-lobed geometries. The significance of this is that in the fourth period, in K and Ca, the 4*s* sub-shell has a lower energy than the 3*d* sub shell. In these elements, electrons prefer to be in 4*s* rather than 3*d* in the ground state even though the radius of the 4*s* shell is larger than the 3*d* shell.

$l = 0$	*s*
$l = 1$	*p*
$l = 2$	*d*
$l = 3$	*f*
$l = 4$	*g*
$l = 5$	*h*

↓

This is a consequence of the increasing dominance of the *l* quantum number in the *d* sub-shell in determining the energy for a sub-shell. The *l* quantum number is connected with the angular momentum of the electrons and so, much like the case where the angular momentum for a rotating wheel is greater if the mass is concentrated at the outer edge compared to the case when the mass is evenly distributed, the concentration of mass of electrons in *d*-shaped shells results in a greater kinetic energy component to the total energy compared to spherical *s* shells.

For the first row of the **transition elements**, the 3*d* levels become occupied only after the 4*s* levels are filled (except for Cr). In these elements, the electrons in the 4*s* shell, being further from the nucleus (but having lower energy than the 3*d* electrons) shield those in the inner 3*d* shell and so the chemical properties for these elements are all very similar since in each case, the valence electrons are the 4*s* *outer* electrons while it is the number of inner 3*d* electrons that is changing. It is the outer **valence electrons** (with the highest principal quantum number) that interact with other atoms to form chemical bonds.

The ordering of shells given above is not the ordering of *all* the energy levels *within* an atom, only the *energy* ordering of the *outer* shells. For example, in K, the 4*s* shell is at a lower energy than the 3*d* shell ($3s^23p^64s^1$). By the time we get to Zn, the 3*d* shell is at a lower energy than the 4*s* shell ($3d^{10}4s^2$). There is no one sequence of energy shells that applies to all elements.

3.6 Noble Gases

The **noble gas** elements are extremely stable and generally do not form compounds with any other elements. The notable feature of noble gas elements is the interesting property of having full outer electron energy shells. Consider the second period elements. As we go from left to right, the number of electrons in the energy level with principal quantum number $n = 2$ is increasing. At neon, we have eight valance electrons: $2s^2$ $2p^6$. A group of eight (**octet**) valence electrons that completely fills an energy shell is very energetically stable and is called the **noble gas configuration**.

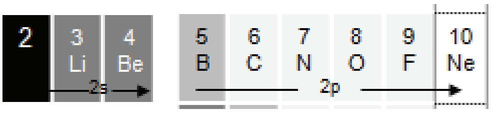

An atom with one more valence electron makes the element sodium, a highly reactive metal. Sodium is highly reactive because its single $3s^1$ valence electron is easily removed so that the sodium **cation** has a noble gas configuration in its outer shell. An atom with one electron less than sodium is the highly reactive gas fluorine. Fluorine readily attracts an electron from another atom to form an **anion** to achieve a **noble gas configuration**.

He	$\boldsymbol{1s^2}$
Ne	$1s^2\boldsymbol{2s^2 2p^6}$
Ar	$1s^2 2s^2 2p^6\boldsymbol{3s^2 3p^6}$
Kr	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10}\boldsymbol{4s^2 4p^6}$
Xe	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10}\boldsymbol{5s^2 5p^6}$
Rn	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^6 5d^{10}\boldsymbol{6s^2 6p^6}$

The noble gas configuration is where there is a completely filled energy shell and the next available higher energy level is an *s* shell. This is true for He where the $1s$ shell is filled and the next energy level is the $2s$ shell. For the other noble gases, we have completely filled *p* shells and the next highest energy level is the *s* sub-shell for the next quantum number *n*. There is a large energy gap between a *p* sub-shell and the next highest *s* shell and this gives the noble gas elements a high **ionisation energy**. Because all the occupied energy shells are filled, the electric charge distribution within noble gas elements is symmetric, resulting in no external electric field. The total angular momenta within these atoms adds up to zero, resulting in no external magnetic fields. These elements find it very difficult to form bonds with other atoms because they have little opportunity to interact electrostatically magnetically, or energetically.

3.7 Atomic Size

The size of an atom cannot be precisely stated since the location of the electrons can only be described statistically. **Atomic size**, for practical purposes, can be defined as the most probable distance from the nucleus to the most outer-shell electron but depends on whether the atom is free or combined with another.

Electrons are attracted to the nucleus by **electrostatic forces**, but are constrained to occupy defined orbits, or energy shells, by the principles of **quantum mechanics**. The electrostatic attraction between the nucleus and a near-shell electron is much larger than between the nucleus and an outer-shell electron, not only because of the increased distance, but also because of the **screening effect** that the inner electrons have on the outer electrons.

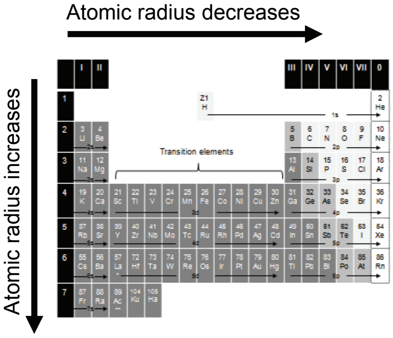

As we go across a **period** in the periodic table, the atomic number increases and so the magnitude of the positive nuclear charge also increases, and therefore so does the magnitude of the **electric field** within which the electrons find themselves. However, across any one period, the outer electrons exist in energy levels for one particular quantum number *n*. For example, in the 3rd period, the 3*s* and 3*p* levels are occupied. Because the electric field is also increasing as we go from left to right, so the electrons are acted upon by a larger Coulomb force and the **atomic radius** decreases across a period. However, because of the **screening effect** offered by the inner-shell electrons, the decrease in size is not as large as expected on the basis of the nuclear charge alone.

For a given vertical **group** of elements in the periodic table, the size of atoms increases as the atomic number Z increases (as we go down the column). We may therefore expect that the atomic radii would decrease but as we go down a group, electrons are filling states with a higher quantum number *n* and this more than offsets the effect of increasing Z.

Therefore **atomic radius** increases as we go down a group in the periodic table.

$$r = n^2\left(\frac{\varepsilon_o h^2}{\pi m_e Z q_e^2}\right)$$

The Bohr model of the atom shows how the atomic radius increases in proportion to n^2 and decreases (inversely) with Z.

3.8 Covalent Radii

When atoms form a **covalent bond**, the nuclei of the two atoms move close together (see Section 1.7) compared to two isolated atoms adjacent to each other.

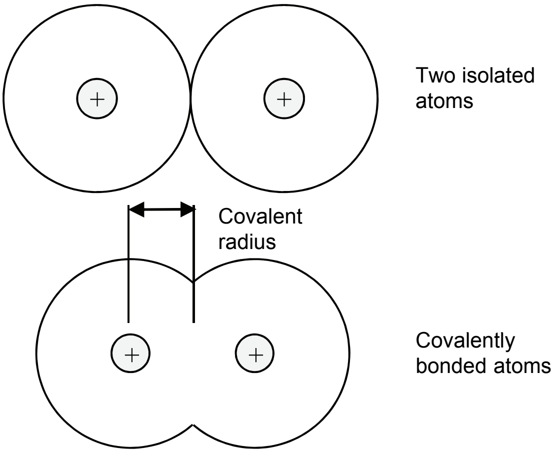

The **covalent radius** of an atom is defined as one half the distance between the nuclei of two identical *covalently bonded* atoms. The actual covalent radius of an atom in a molecule where the atoms are not identical depends on the atoms involved. The general term "**atomic radius**" usually means the covalent radius (where the two atoms are identical) and this serves as a useful benchmark for comparison with other atomic sizes.

For metallic atoms, the atomic or **metallic radius** is usually defined as half the distance between two nuclei of the atoms in the solid metallic state.

The shapes of the orbitals in which the valence electrons participate in bonding is important. The *s* orbital has a somewhat spherical shape centred on the nucleus, while the *p* orbitals have a dumbbell shape along the three coordinate axes. The shapes of the atomic orbitals involved in the formation of a particular bond determine the shape of the molecule.

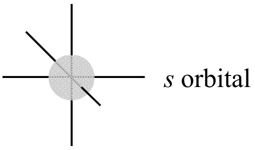

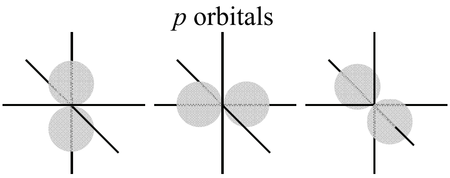

3.9 Ionic Radii

When a metal atom loses an electron to become a positively charged **ion**, a reduction in **atomic size** occurs. The electron that is lost is usually an outer-shell electron and so this electron shell becomes vacant and so the outermost electron shell is the next one closer in towards the nucleus. However, the reduction in size is not just due to an outer electron shell becoming vacant. When a positive ion is formed, the remaining electrons can bunch together a little more closely because of the reduced degree of mutual repulsion (since there are now fewer electrons surrounding the nucleus).

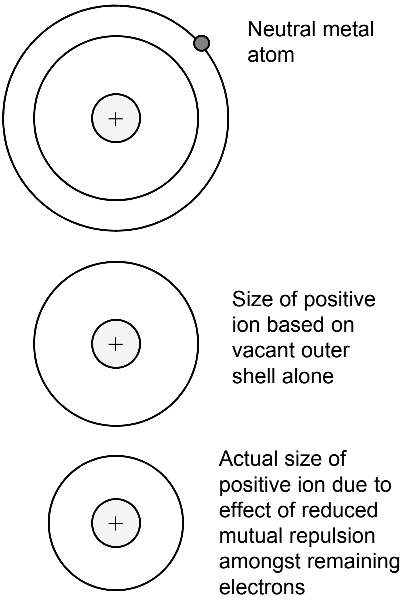
Neutral metal atom
Size of positive ion based on vacant outer shell alone
Actual size of positive ion due to effect of reduced mutual repulsion amongst remaining electrons
+
+
+

When a non-metal atom gains an electron to become a negative ion, this usually results in an outer-level electron shell accepting an additional electron – which would not ordinarily cause an appreciable increase in size, but the addition of this electron means that there is now an excess of negative charge in the outer shell and mutual repulsion causes these electrons to move apart and the atom "expands".

In a sodium chloride ionic crystal, therefore, the positive sodium ions shrink and the negative chlorine ions expand such that the sodium ions are almost packed in between the spaces of the chlorine ions.

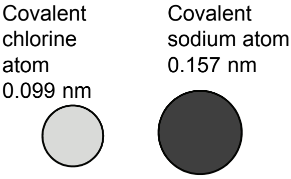
Covalent chlorine atom 0.099 nm
Covalent sodium atom 0.157 nm

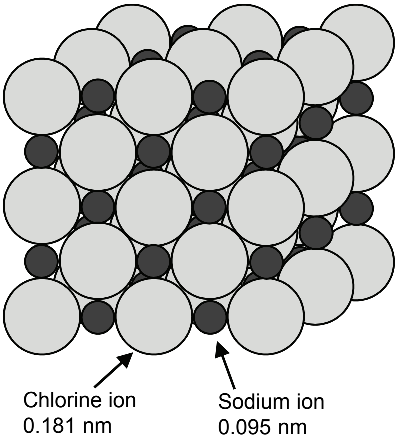
Chlorine ion 0.181 nm
Sodium ion 0.095 nm

3.10 Ionisation Energy

The **ionisation energy**, or **ionisation potential**, is the amount of energy needed to remove an electron from an isolated atom. The first ionisation energy is that needed to remove one outermost electron from an atom. Successive ionisation energies are those needed to remove further electrons from the atom.

Successive ionisation energies always increase at each electron removal. As each electron is removed, the size of the atom decreases markedly and so the remaining electrons are closer to the positively charged nucleus, and so more energy is needed to remove them. Further, a particularly large increase in energy is needed to remove an electron from a filled energy level because a noble gas configuration of electrons is extremely stable.

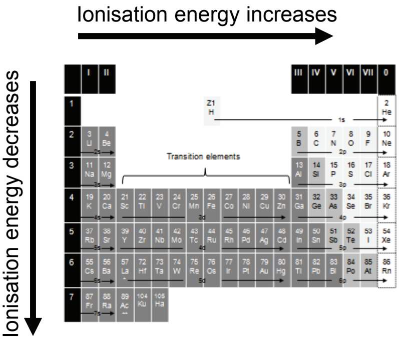

Within a horizontal **period**, there is an increase in nuclear charge. As the **nuclear charge** increases (number of protons on the nucleus), the magnitude of the electric field within which the electrons find themselves also increases, and so generally speaking, the ionisation energy increases within a period. Also, the **atomic size** decreases from left to right and so the electrons for atoms towards the right are closer to the nucleus compared to those on the left, and so this also results in a general increase of **ionisation energy** from left to right.

Within a vertical **group**, there is generally a decrease of ionisation energy from top to bottom which is predicted on the basis of size of the atom alone. Any potential increase in ionisation energy on the basis of increased nuclear charge, going from top to bottom, is essentially cancelled out by the increased **screening effect** of the inner electrons on the outer electrons.

Generally speaking, elements with a low ionisation energy are on the left-hand side of the periodic table and those with a high ionisation energy are on the right. Noble gas configurations (eight valence electrons) can cause exceptions to these general rules.

3.11 Electronegativity

The **electronegativity** describes the relative ability for an atom, when it combines with another atom, to become “more negative” by more strongly capturing a shared electron or electron pair.

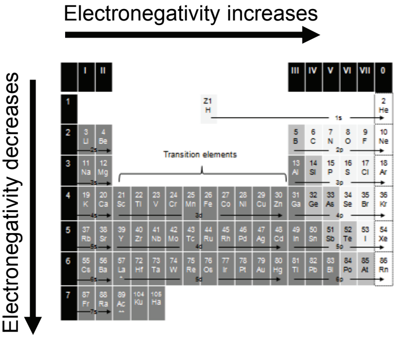

It is found that the more difficult it becomes for an atom to lose an electron, the easier it is for the atom to add an extra electron. That is, a higher **ionisation energy** also means a higher electronegativity.

Some electronegativities within a **group** are:

Be	1.5
Mg	1.3
Ca	1.0
Sr	0.95
Ba	0.9
Ra	0.9

Some electronegativities within a **period** are:

Li	Be	B	C	N	O	F	Ne
1.0	1.5	2.0	2.5	3.0	3.5	4.0	-

The electronegativity of fluorine is so strong that in some circumstances, it is able to attract an electron from helium and so produce a compound that involves a noble gas.

The electronegativity of an atom influences the nature of the bonding between different types of atoms. Even when a bond is covalent, the shared electron spends more time nearer to the atom which has a higher electronegativity. The resulting molecule becomes polar.

4. Chemical Bonds

Summary

Lewis electron dot structures

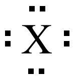

Van der Waals

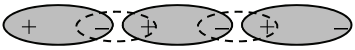

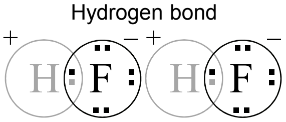

Oxidation numbers:

a) In free, uncombined, elements, the oxidation number of each atom is set to 0.

b) In compounds involving hydrogen, the oxidation number of hydrogen is 1+ .

c) In compounds involving oxygen, the oxidation number of O is usually 2–.

d) The sum of all the oxidation numbers of all atoms in an ion is equal (in both magnitude and sign) to the charge on the ion.

e) The sum of all the oxidation numbers of all atoms in a neutral molecule is 0.

Oxidation → Increase in Oxidation No. → Loss of Electrons

Reduction → Decrease in Oxidation No. → Gain of Electrons

4.1 Chemical Bond

Consider two hydrogen atoms that are initially a long way apart from each other. The two atoms are both electrically neutral, and have no ionic character. Why should they then be attracted to each other and form a stable H_2 molecule? Atoms tend to arrange themselves in the lowest possible energy state. In a **covalent bond** (like in H_2), the sharing of electrons provides the necessary reduction in total energy since the shared electrons, now having more space to move (over two atoms instead of one) have a lower **energy** than a single electron orbiting a single nucleus (see Section 1.5). This reduction in energy appears to us like a force of attraction acting between the atoms.

When two atoms form a **chemical bond**, they take up equilibrium positions according to the balance between long-range attractive and short-range repulsive forces $F(r)$ between them. At the **equilibrium position**, the potential energy $V(r)$ of the bond is a minimum.

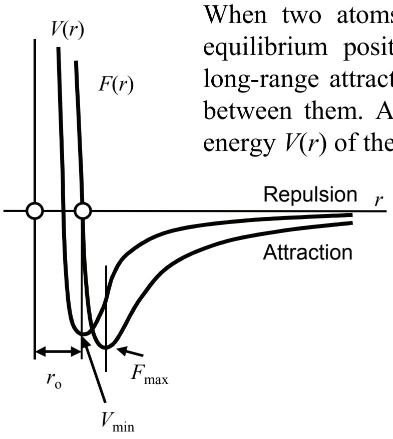

The potential energy function (V(r)) acting between two atoms can be very complex. Simple models are usually used, a very popular one being the **Lennard-Jones potential**:

$$V(r) = -\frac{A}{r^6} + \frac{B}{r^{12}}$$

In this potential, the negative term is the energy associated with the attractive forces and the positive term is the energy associated with the repulsive forces.

It is convenient to assign a potential energy of zero to widely spaced molecules or atoms so that when they approach, the potential energy becomes more negative. Since we generally assign a positive number to **work** done on a system (energy entering a system) and a negative number to work done by a system (energy leaving a system), work has to be done on the molecules or atoms to separate them. Thus, when the atoms are at the equilibrium position, energy is required to be done on the system to move one atom from the equilibrium distance to infinity against the force of attraction. This energy is called the **bond energy** (usually in kJ/mol).

To be consistent with the assignment of energy, we must therefore treat **attractive forces** as negative and repulsive forces as positive. The force F, with this sign convention, when multiplied by the distance of movement r, then results in an energy of the correct sign.

4.2 Lewis (Electron Dot) Formulae

Molecules and compounds form when bonds exist between atoms involving the transfer or sharing of valence electrons. The **Lewis symbol** for an atom is the element symbol surrounded by dots that represent the possible **valence electrons** for the atom.

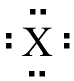

Lewis recognised that bonding or sharing of electrons occurred so that an element tended to acquire eight valence electrons (noble gas configuration): the **octet rule**. To represent a bond between two atoms, a **Lewis structure** is written so as to indicate the position of the valence electrons:

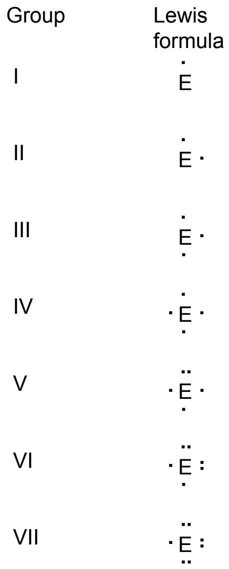

“E” denotes an element.

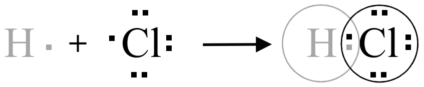

This **Lewis structure** shows how each element acquires a noble gas configuration when a chemical (in this case, ionic) bond is formed.

When determining a Lewis structure, the total number of dots that appear (the total number of **valence electrons**) is the sum of the number of valence electrons for each participating atom.

In some cases, it is not easily determined what the Lewis or **electron dot** formula for an atom should be. For example, in carbon, the electronic configuration is: $1s^2$ $2s^2$ $2p^2$ and we might be tempted to write the **electron dot** formula as:

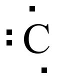

Although this arrangement can occur, when carbon usually combines with other atoms, the $2s$ orbitals combine with the $2p$ orbitals and the electrons are spread out over the resulting combined “*sp*” energy level giving four valence electrons. Carbon is **tetravalent**.

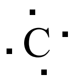

4.3 Multiple Bonds

Covalent bonds between atoms often involve sharing of more than one electron at a time. For example, carbon has a bonding capacity, or valence, of four, and so the carbon compound **ethane** is a stable molecule and is fairly unreactive.

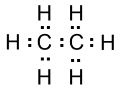

$$\mathrm{H}:\underset{\mathrm{H}}{\overset{\mathrm{H}}{\ddot{\underset{\cdot\cdot}{\mathrm{C}}}}}:\underset{\mathrm{H}}{\overset{\mathrm{H}}{\ddot{\underset{\cdot\cdot}{\mathrm{C}}}}}:\mathrm{H}$$

Ethylene, with molecular formula C_2H_4 should therefore leave two lone valence electrons:

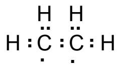

$$\mathrm{H}:\overset{\mathrm{H}}{\ddot{\underset{\cdot}{\mathrm{C}}}}:\overset{\mathrm{H}}{\ddot{\underset{\cdot}{\mathrm{C}}}}:\mathrm{H}$$

We would therefore expect ethylene to be a very reactive compound, in much the same way as Group I elements. However, this is not the case. Instead, the two carbon atoms arrange for the orbits of the two lone electrons to overlap and be shared. That is, two pairs of electrons are shared between the carbon atoms to form a **double bond**.

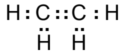

$$\mathrm{H}:\underset{\mathrm{H}}{\underset{\cdot\cdot}{\mathrm{C}}}::\underset{\mathrm{H}}{\underset{\cdot\cdot}{\mathrm{C}}}:\mathrm{H}$$

Triple bonds are also possible whereby three pairs of electrons are shared. The double and triple bonds offer even a greater range of movement of valence electrons (since more electrons are shared and so more electrons have more overall room to move) compared to a single bond. The strength of the bond is higher, and the C nuclei are closer together (shorter bond length).

Although the bond energy associated with a double bond is higher, the double bond results in a more reactive compound than a comparable single-bonded structure. This is because the double bond offers more opportunities for other more reactive elements than carbon to attach to these valence electrons – as if other elements can open up the bond and find two potential bonding sites rather than one. It is energetically more favourable to have two single bonds than one double bond.

Molecule	Structural formula	C – C length (Å)	C – C energy (kJ/mol)
C_2H_6 Ethane	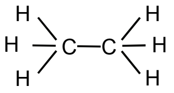	1.54	334
C_2H_4 Ethylene	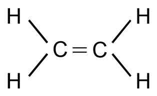	1.33	606
C_2H_2 Acetylene	$H-C\equiv C-H$	1.20	828

4.4 Lewis Single-Bonded Structures

Some examples of single-bonded structures are:

Molecular formula	Total No. valence electrons	Lewis formula	N	Shape	Note
F_2 Fluorine gas	14	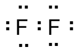	-	F — F	Linear
CH_4 Methane	8	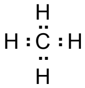	4	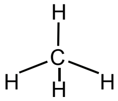	Tetrahedral
$SiCl_4$	32	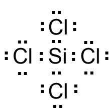	4	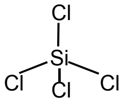	Tetrahedral
NH_3 Ammonia	8	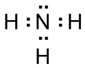	4	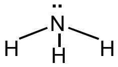	Pyramidal
H_2O Water	8	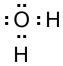	4	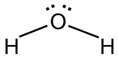	Bent
NH_4^+	8	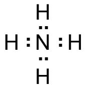	4	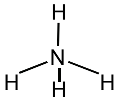	Tetrahedral
C_2H_6 Ethane	14	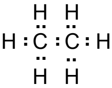	4,4	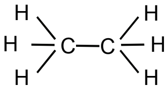	2 tetrahedral
H_2O_2	14	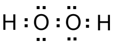	4,4	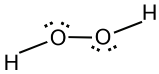	

Note that **lone pairs**, being closer to the central atom, have a greater repulsive effect than bonded pairs and so influence the shape of the molecule. Note that each atom is surrounded by eight electrons (except for H which requires two electrons for noble gas configuration).

4.5 Lewis Multiple-Bonded Structures

Examples of multiple-bonded structures:

Molecular formula	No. valence electrons	Lewis formula	N	Shape	Note
N_2 Nitrogen gas	10	$\ddot{N}:::\ddot{N}$	-	$N \equiv N$	Linear
C_2H_4 Ethylene	12	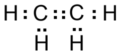 $H:\underset{\underset{H}{..}}{C}::\underset{\underset{H}{..}}{C}:H$	3,3	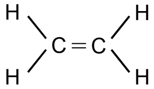	2 Tetrahedral
C_2H_2 Acetylene	10	$H:C:::C:H$	2,2	$H—C \equiv C—H$	Linear
CO_2 Carbon dioxide	16	$:\ddot{O}::C::\ddot{O}:$	2	$O = C = O$	Linear
HCN Hydrogen cyanide	10	$H:\underset{..}{\ddot{C}}:::\ddot{N}$	2	$H—C \equiv N$	Linear
$COCl_2$ Phosgene	24	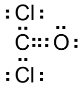 $\begin{array}{c} :Cl: \\ \ddot{C}:::\ddot{O}: \\ :\ddot{Cl}: \end{array}$	3	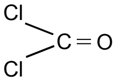	Planar
H_2CO Formaldehyde	12	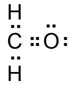 $\begin{array}{c} H \\ \ddot{C}::\ddot{O}: \\ H \end{array}$	3	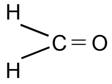	Planar

Note that each atom is surrounded by eight electrons (except for H which requires two electrons for noble gas configuration).

4.6 Lewis Exceptions to the Octet Rule

Examples of structures that do not obey the **octet rule** include:

(1) Atoms having more than eight electrons involved in bonding:

Molecular formula	No. valence electrons	Lewis formula	N	Shape	Note
PCl_5	40	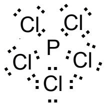	5	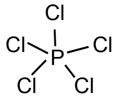	More than an octet
SF_4	43	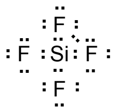	5	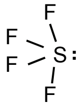	More than an octet with lone pair

When there are more than eight electrons involved, such as in PCl_5, the bonding usually involves electrons from the *d* sub-shell.

(2) Atoms having less than eight electrons involved in bonding:

BeH_2	4	H:Be:H	2	H — Be — H	Less than an octet
NO	11	$:\dot{N}::\ddot{O}:$		N = O	Less than an octet

When more than one valid structure can be written and the true structure cannot be written, the actual or true structure is called a **resonance hybrid** of the different structures. An example is NO. The unpaired electron (as shown above) can be written as attached to either the N or the O atom.

Example of **resonance structure**: CO_3^{2-}

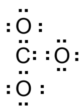

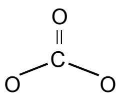

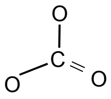

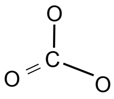

4.7 Oxidation Number

The **oxidation number**, or **oxidation state**, is a quantity that describes the number of electrons that appear to be gained or lost by an atom when a chemical bond is formed. That is, it is essentially the number of electrons that participate in the formation of the chemical bond. If we imagine all bonds to involve a transfer of electrons (even if in reality the bond may be covalent), the oxidation number is essentially the electronic charge that an atom appears to have after the bond has formed. The oxidation number is determined by a set of agreed rules. It is not strictly a physical quantity.

In a molecule or a compound, electrons that are shared between atoms are counted as belonging to the more **electronegative** atom. Electrons that are shared between two identical atoms are equally divided between the two atoms.

(a) In free, uncombined elements the oxidation number of each atom is set to zero. For example, the oxidation number of H in H_2 is zero.

(b) In compounds involving hydrogen, the oxidation number of hydrogen is 1+ except in the case of metal hydrides (where the hydrogen is bonded to an atom which is less electronegative) where it is 1−.

(c) In compounds involving oxygen, the oxidation number of O is usually 2−. The exceptions are the oxygen atom in peroxides and when oxygen is bonded with fluorine.

(d) The sum of all the oxidation numbers of all atoms in an ion is equal (in both magnitude and sign) to the charge on the ion.

(e) The sum of all the oxidation numbers of all atoms in a neutral molecule is 0.

Consider the compound H_2O. The oxygen atom, being more **electronegative** than the hydrogen, is seen to gain two electrons and so is given an oxidation number of 2− (since it seems to have acquired a net “negative” charge). Each hydrogen atom is seen to lose an electron and so the oxidation number of hydrogen in this compound is assigned as 1+.

When the oxidation number of an atom has increased, the atom is said to have been **oxidised** (as in the hydrogen above). This usually involves the loss of electrons. When the oxidation number of an atom has decreased, the atom is said to have been **reduced**, that is, its oxidation number has decreased. This usually involves a gain of electrons (as in the oxygen above).

The oxidation number is useful when naming some compounds. Prefixes like *hypo* and *per*, and suffixes like *ous*, *ic*, *ite*, are assigned according to the oxidation number of the central metal or non-metal atom in a compound.

4.8 Oxidation Number Examples

Example	Element	Oxidation state	Rule
S	S	0	a
H_2S	S	-2	b, e
Fe^{2+}	Fe	+2	d
Fe^{3+}	Fe	+3	d
MnO_4^-	Mn	+7	c, e
Mn^{2+}	Mn	+2	d
$Cr_2O_7^{2-}$	Cr	+6	c, e
Cr^{3+}	Cr	+3	d
NO_3^-	N	+5	c, e
NO	N	+2	c, e
NO_2	N	+4	c, e
SO_4^{2-}	S	+6	c, e
CO_2	C	+4	c, e
H_2O_2	O	-1	special case

When a substance is **oxidised**, it can be done so by either adding oxygen or removing hydrogen. Similarly, when a substance is **reduced**, it can be done so by either removing oxygen or adding hydrogen. For example, consider the production of methane CH_4 from carbon and hydrogen:

$$\begin{array}{rl} \mathrm{C} + 2\mathrm{H}_2 \rightarrow \mathrm{CH}_4 & \text{Oxidation No.} \\ 2\mathrm{H}_2 \rightarrow 4\mathrm{H}^+ + 4\mathrm{e}^- & 0 \text{ to } 4+ \\ \mathrm{C} + 4\mathrm{e}^- \rightarrow \mathrm{C}^{4-} & 0 \text{ to } 4- \end{array}$$

From the point of view of the hydrogen, it has lost electrons and so is oxidised. From the point of view of the carbon, it has gained electrons and so is reduced. Of course in this compound, CH_4, the electrons are shared between the C and H atoms as covalent bonds, but from the perspective of oxidation numbers, the electrons are effectively transferred from H to C.

In some biochemical reactions, electrons (in the company of protons H^+) are transferred during oxidation and reduction reactions as a means of passing energy from one molecule to another. The addition of an electron via the addition of hydrogen stores energy in the compound being reduced. The energy is released when the compound is oxidised.

When CH_4 is combined with oxygen (e.g. when it is burnt), the C is oxidised to CO_2 and water: $\mathrm{CH}_4 + \mathrm{O}_2 \rightarrow \mathrm{CO}_2 + 2\mathrm{H}_2\mathrm{O}$

4.9 Polar Bonds

Although bonds between atoms may be classified as being **covalent** or **ionic**, in many different substances, there is a gradation from one type to the other. Unequal sharing of electrons in a covalent bond gives the bond an ionic character. Further, such a covalent molecule may acquire regions of net positive and negative charge and become **polarised**.

Separation of charges within a covalent molecule usually results in electrons spending more time on the atom which has the greatest **electronegativity**.

:C :::O:

In CO, the shared electrons spend more time near O than C and so the centre of positive charge is offset from the centre of negative charge.

For example, in the molecule carbon monoxide, there is a difference in electronegativity between carbon (2.5) and oxygen (3.5) and so the shared electrons spend more time near the oxygen atom than the carbon atom. This results in a net separation of centres of **electric charge** within the molecule. The molecule is thus polarised and is a **dipole**. The molecule as a whole remains electrically neutral, but the distribution of charge within the molecule shifts so that the geometrical centre of positive charge is different than the centre of negative charge.

Although it is straightforward to determine if a diatomic molecule is polar or non-polar, by a consideration of the electronegativities of their atoms, the situation is more complicated for molecules with more than two atoms. For example, in CO_2, the individual bonds between the carbon and the oxygen molecules may be polar, but the molecule as a whole is non-polar because of the symmetry of the shape of the molecule.

:Ö::C::Ö: O = C = O μ = 0 debye

When CO_2 is placed in an electric field there is no net **dipole moment**.

Experiments show that **water** has a permanent dipole moment due to the presence of the electron pairs on the oxygen.

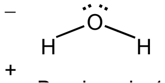

μ = 1.84 debye

Bond angle 104.5°

Polarization is responsible for **physical bonds** forming between molecules which in turn result in molecules forming liquids and solids.

4.10 Hybrid Orbitals

In a covalent bond, when valence electrons are shared between atoms, they do so by overlapping orbitals. Consider an oxygen atom: $1s^2, 2s^2, 2p^4$

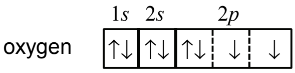

The shapes of molecules are difficult to predict on the basis of the orbitals of isolated atoms because the nature of these orbitals change when there are more than one electron in an atom and when atoms come together to form molecules.

Hybrid	Shape
sp	linear
sp^2	plane triangular
sp^3	tetrahedral

When a molecule of water forms, the 2*s* and the three 2*p* orbitals in the oxygen combine to form a new molecular orbital called the sp^3 **hybrid**. The superscript 3 indicates that three of the original *p* orbitals are involved in the hybrid orbital. It is this hybrid orbital $2sp^3$ that combines with the *s* orbitals of the H atoms.

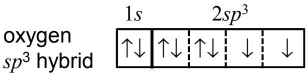

Each of the $2sp^3$ hybrid orbitals is at the same energy level.

When hybridization occurs, the rearrangement of electrons often results in there being more unpaired valence electrons than there were originally. For example, in carbon compounds, we have:

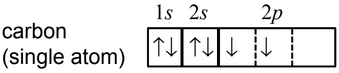

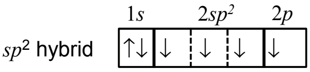

In CH_4, it is the hybridization of the C atomic energy levels into four identical $2sp^3$ orbitals that results in a regular tetrahedral geometry with bond angle 109.5°. In PCl_5 (more than eight electrons bonded), the hybrid orbitals are sp^3d.

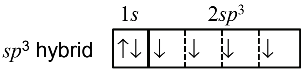

In the double C=C bond of C_2H_4 one of the bonds is the end-to-end overlap of one of the sp^2 hybrid orbitals (this is called the σ **sigma bond**) along the line of the two C atoms (the other two sp^2 orbitals overlap with the *s* orbitals of the H atoms), while the other C bond is formed by the side-by-side overlap of the unhybridised (dumbbell shaped) 2*p* orbitals parallel to the plane of the C and H atoms (this is called the π **pi bond**).

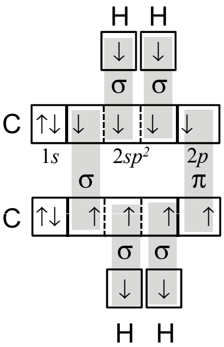

4.11 Polarisation

When an electric field *E* is applied to a molecule in a non-polar insulating material, the centre of charge of electrons moves left, the centre of charge of protons move right and the two centres of charge are then separated by distance *d*. The atom or molecule becomes **polarised** in the presence of *E*.

Polarisation in a material may occur due to several mechanisms, all of which may occur to some extent depending on the atom. Most often, polarisation occurs when an external electric field *E* is applied.

1. **Electronic polarisation**
   Small induced dipole moment arising from difference in the net centres of nucleus and electrons in an atom.
   $\varepsilon_r \approx 2-4$

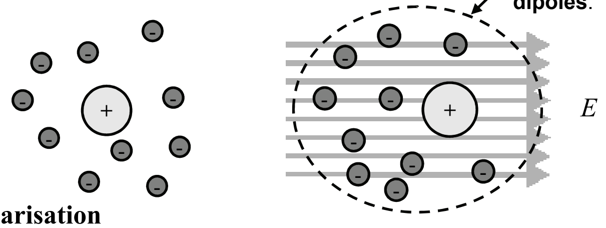

The entities with circles around them are **dipoles**.

2. **Ionic polarisation**
   Dipole moment created by shift of positive ions with respect to negative ions in unit cell.
   $\varepsilon_r \approx 6-10$

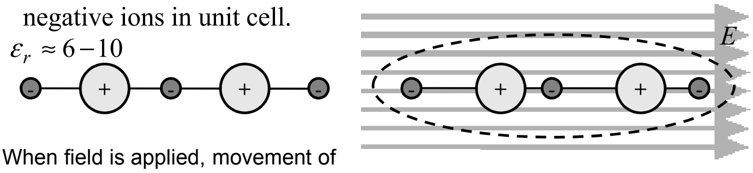

When field is applied, movement of ions produces a net dipole moment in a unit cell and hence a net polarisation.

3. **Dipolar polarisation**
   Net dipole moment created by alignment of molecule with external field due to presence of internal permanent dipoles from geometrical structure of molecule.
   $\varepsilon_r \approx 20-100$

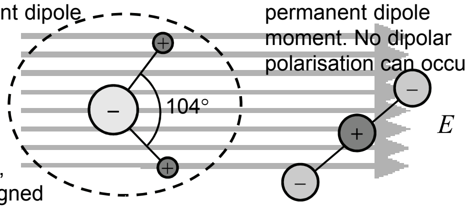

Water molecule has a permanent dipole moment.

CO_2 molecule has no permanent dipole moment. No dipolar polarisation can occur.

Note: Even though there is a permanent dipole moment, in the absence of an electric field, these dipole moments are not aligned (due to thermal agitation) and so the net polarisation over all molecules in the material is zero. When a field *E* is applied, molecular dipoles tend to align themselves with the field and there exists a net polarisation.

4.12 van der Waals Forces

The arrangement of electrons around atoms is not constant in time. Electrons move within their energy shells or levels with kinetic energy. An instantaneous movement of an electron within an atom may result in the atom becoming **polarised** for a short period of time. During this time, the polarisation in one atom may induce polarisation in a neighbour atom, resulting in a very brief electrostatic attraction between them. These instantaneous forces of attraction and repulsion are termed **London forces** after the scientist who proposed the explanation for them and are found in all covalent materials.

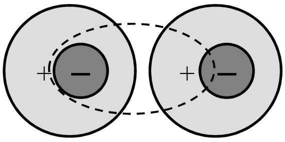

Instantaneous polarisation of one molecule results in induced dipole in neighbour molecule and a net instantaneous attraction between the two molecules.

In materials containing molecules that have permanent dipole moments, the molecules tend to align themselves (against the tendency of thermal agitation to keep them in random orientation). The electrostatic attraction between the polarised ends of the molecules draws them together. **Dipole-dipole interaction** may persist over many molecules and appears to us like a force holding the molecules together.

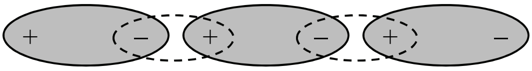

When considered over millions of atoms, the London and permanent dipole forces can be substantial and are called physical, as distinct from chemical, bonds between molecules. The intermolecular forces of this type are called **van der Waals forces**. In most cases, it is the van der Waals forces that bind atoms and molecules into liquids.

Intermolecular forces on this scale are responsible for deviations from ideal behaviour in liquids and gases. Van der Waals forces operate over a small scale in comparison to distances between molecules in a gas. It is only when gas molecules become closer together (such as under increased pressure) that these forces become strong enough to cause the molecules to then clump together and **condense** into a **liquid**. The forces become less effective when the temperature increases due to the tendency for an increase in random order associated with thermal agitation of molecules.

4.13 Hydrogen Bond

Hydrogen has one valence electron and so can readily form a single covalent bond. However, there are some compounds involving hydrogen in which the hydrogen atom can be said to be bonded to two atoms at the same time. Such compounds involve hydrogen bonding with small electronegative atoms.

Consider the **hydrogen fluoride** molecule:

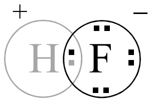

Because fluorine has a large **electronegativity**, the shared electron spends much of its time near the fluorine atom. The molecule, as shown above, is **polar**. Now, if there are other HF molecules nearby, then the positively charged H side of the molecule will be drawn into contact with the negatively charged F end of the other molecule. Two molecules thus join up so:

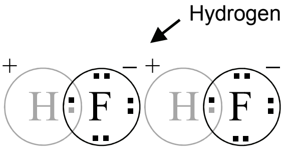

It looks to us like the H atom in the middle is holding the two outer F atoms in place like a tie bar. This electrostatic attraction is called a **hydrogen bond**. The hydrogen bond can only form between hydrogen and small electronegative atoms because the H "ion" is very small and only has space for one other atom to be located near enough to its polar end to form an **electrostatic** bond.

Hydrogen bonds also form between **water molecules**. Hydrogen bonding between water molecules gives rise to a characteristic crystal structure in **ice**, which in turn is reflected in the shape of **snowflakes**.

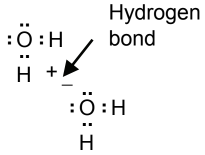

Hydrogen bonding is very important in biological processes. The bonds, having an energy of about 20 kJ, are stronger than those due to van der Waals forces (about 0.2 kJ) but are significantly weaker than those in chemical bonds (about 200 kJ). Certain biological processes rely on these bonds being of just the right strength for events to occur with ease (such as in the forming and breaking of the two halves of the molecular spiral in **DNA** – these hydrogen bonds involve H, O and N atoms).

5. States of Matter

Summary

Gases: $$\frac{p_1V_1}{T_1} = \frac{p_2V_2}{T_2}$$ Combined gas law

$$p_1V_1 = p_2V_2$$ Boyle's law

$$\frac{V_1}{T_1} = \frac{V_2}{T_2}$$ Charles' law

$$\underset{\text{Absolute pressure}}{p}\,\overset{\text{Volume}}{V} = \underset{\text{no. moles}}{n}\,\underset{\text{Universal gas constant 8.3145 J/mol/K}}{R}\,\overset{\text{Absolute temperature}}{T}$$

Liquids: $$\text{Molarity} = \frac{\text{No. moles of solute}}{\text{Volume of solution in litres}}$$

Solids:

- Molecular solids - in which the binding forces between discrete atoms or molecules are due to van der Waals interactions between instantaneous electric dipoles.
- Ionic solids - in which the binding energy is due to Coulomb electrostatic forces between positively and negatively charged ions in the crystal.
- Covalent solids - in which the binding energy is due to shared valence electrons between atoms in the solid.
- Metallic solids - in which valence electrons are effectively shared amongst all the atoms in the solid.
- Amorphous solids – no long-range structure.

5.1 Changes of State

When a solid is heated, atoms/molecules within the solid absorb heat energy and this is stored as **internal energy** of vibration. As the temperature increases, the magnitude or amplitude of the vibration increases until the bonds holding the molecules together become temporarily ruptured and molecules flow over one another. This is called **melting**.

At the **melting point**, further heating results in no increase in temperature. The energy input in this case goes to breakage of bonds between the molecules and the material changes phase from solid to liquid.

Upon further heating, molecules of liquid increase their translational energy until they acquire sufficient kinetic energy to escape the liquid into the gaseous phase.

The presence of impurities lowers the melting point. A measurement of the melting point can be used as an index of the purity of a particular substance.

When a liquid cools, molecules begin to lose their internal energy of motion until eventually, intermolecular forces are able to hold them in place as a solid. The temperature at which this occurs is called the **freezing point** – which is usually equal to the melting point.

If the molecules in the cooling solid do not immediately lock into the pattern of the solid, then the temperature may fall below the official freezing point and the material remains in the liquid phase. This is called **supercooling**. Introduction of a seed crystal may cause the liquid to avoid the supercooling condition.

Forces between molecules:

- Long-range attractive
- Short-range repulsive

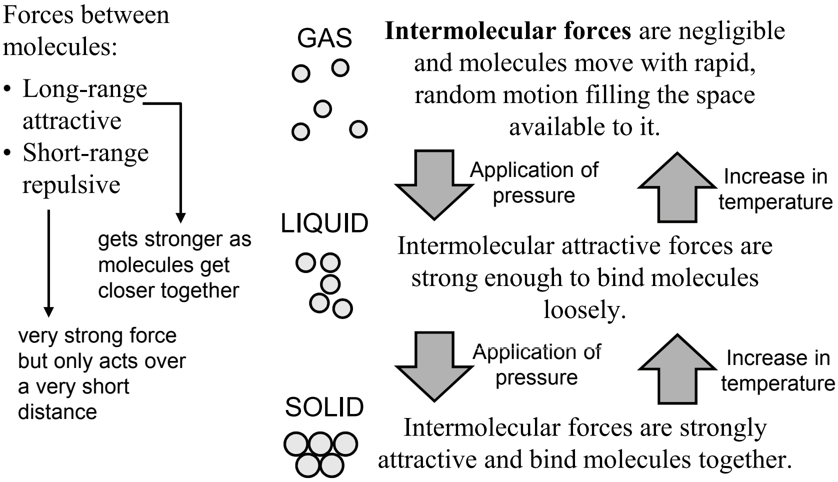
gets stronger as molecules get closer together
very strong force but only acts over a very short distance
GAS
Application of pressure
Increase in temperature
LIQUID
Application of pressure
Increase in temperature
SOLID

Intermolecular forces are negligible and molecules move with rapid, random motion filling the space available to it.

Intermolecular attractive forces are strong enough to bind molecules loosely.

Intermolecular forces are strongly attractive and bind molecules together.

5.2 Changes of State of Liquids/Gases

Closed system

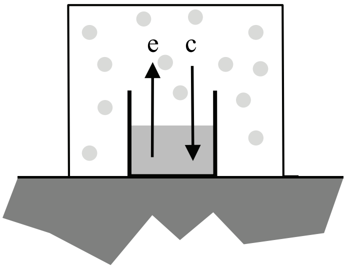

A dynamic equilibrium is set up where the rate of evaporation = the rate of condensation.

The **equilibrium vapour pressure** (**EVP**) or **saturated vapour pressure** is the pressure exerted by the vapour when in equilibrium with its liquid. The magnitude of the EVP depends on the nature of the liquid and the temperature.

- Nature of liquid. The stronger the **van der Waals forces** between molecules, the lower is their ability to escape the liquid and the lower the EVP.
- Temperature. Increase in temperature increases kinetic energy of molecules in the liquid and hence their ability to escape from the surface and hence the higher the EVP.

Open system

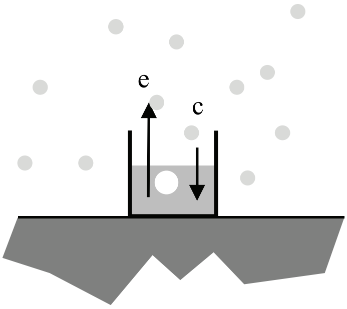

A liquid is said to boil at a temperature called the **boiling point**, at which the EVP of the liquid becomes equal to the prevailing atmospheric pressure. Vapour bubbles form in the liquid and rise to the surface.

The **normal boiling point** is the boiling point at 1 atm of pressure (101.3 kPa).

Increasing the prevailing atmospheric pressure increases the boiling point.

5.3 Phases of Matter

In a ***p–V* diagram**, the temperature is kept constant, volume decreased and pressure recorded.

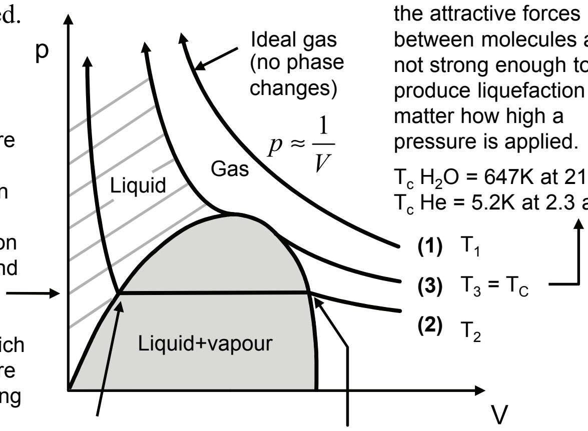
p
V
Liquid
Gas
Liquid+vapour
Ideal gas (no phase changes)
$p \approx \frac{1}{V}$
(1) T_1
(3) $T_3 = T_C$
(2) T_2

Vapour pressure
The partial pressure exerted by the vapour when it is in equilibrium with its liquid. It depends on the temperature and nature of the substance. The temperature at which the vapour pressure equals the prevailing atmospheric pressure is called the **boiling point**.

Critical Temperature T_c
There is, for each gas, a temperature above which the attractive forces between molecules are not strong enough to produce liquefaction no matter how high a pressure is applied.

T_c H_2O = 647K at 218 atm
T_c He = 5.2K at 2.3 atm

All gas condensed into liquid, attempts to further reduce volume produce large increase in pressure as liquid is compressed.

Gas starts condensing into liquid, no change in pressure as volume decreases.

In a **phase diagram**, we keep the volume *V* constant and plot pressure versus temperature.

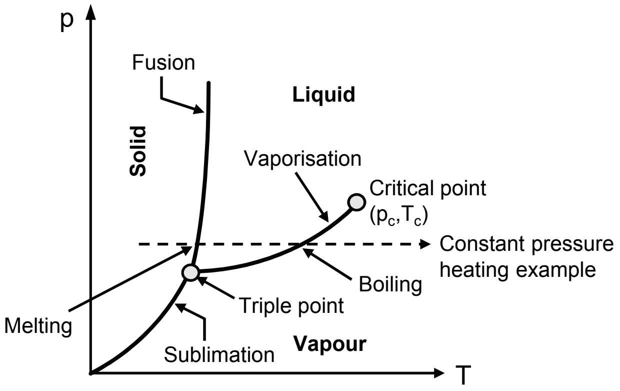
p
T
Fusion
Liquid
Solid
Vaporisation
Critical point (p_c,T_c)
Constant pressure heating example
Boiling
Triple point
Melting
Sublimation
Vapour

At each point (*p*,*T*) only a single phase can exist except on the lines where there is phase equilibrium. At the **triple point**, solid, liquid and vapour exist together in **equilibrium**.

5.4 Gases

The properties of a perfect or ideal gas are the most convenient to study. The properties of an ideal gas are:

- The molecules of the gas occupy a very small volume compared to the volume of the container.
- The molecules are very distant from one another and only interact elastically during collisions.

Real gases often behave like ideal gases at conditions in which no phase changes occur.

Macroscopic properties of a gas:

- Pressure
- Temperature
- Volume
- Mass

These quantities specify the **state** of a gas

Consider a mass (m) of gas:

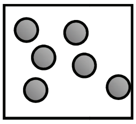

If p, V and T all vary, then:

$$\frac{p_1 V_1}{T_1} = \frac{p_2 V_2}{T_2}$$ **Combined gas law**

If temperature T is a constant:

$$p_1 V_1 = p_2 V_2$$ **Boyle's law**

If pressure p is a constant:

$$\frac{V_1}{T_1} = \frac{V_2}{T_2}$$ **Charles' law**

Let us express the mass of a gas indirectly by specifying the number of moles. Experiment shows that Boyle's law and Charles' law lead to:

Volume
Absolute temperature

$$pV = nRT$$

Absolute pressure
No. moles
Universal gas constant 8.3145 J/mol/K

This equation links all the macroscopic quantities needed to describe the (steady) state of an *ideal* gas and is thus called an **equation of state**.

Note: These laws cannot be applied when the mass of gas changes during the process. Pressures and temperatures are absolute.

Mass in kg

No. moles $\longrightarrow n = \dfrac{m}{M_m} \longleftarrow$ Molar mass

By using moles, we get the ideal gas equation with the universal gas constant R (units J/mol/K). Otherwise, value of R depends on the nature of the gas (i.e. no longer universal) and has units J/kg/K.

Example: Calculate the volume occupied by one mole of an ideal gas at 273 K at atmospheric pressure.

$$pV = nRT$$
$$101.3(V) = 1(8.314)(273)$$
$$V = 22.406\,\mathrm{L}$$

5.5 Solutions

Homogenous mixtures of two or more gases, liquids, or solids are called **solutions**.

- Gases usually mix in any proportion. Air is a solution of predominantly nitrogen, oxygen and carbon dioxide.
- Liquid solutions can be made by dissolving a solid, liquid or gas (the **solute**) in a liquid (the **solvent**). When the solvent is water, the solution is called an **aqueous solution**.
- Solid solutions occur where atoms of one solid are randomly dispersed throughout the other solid. **Alloys** are common examples of solid solutions although some alloys are **compounds** (combine in definite proportions). Some alloys are also heterogeneous **mixtures**.

Intermediate between heterogeneous mixtures and homogenous solutions is the case where the dispersed solid particles are not broken up into molecular units and yet are not in collections large enough to be called a separate phase. In this case, the dispersion is called a **colloid**.

The strength, or **concentration**, of a solution can be described as:

- The ratio of the number of moles of one of the components (the solute or the solvent) relative to the total number of moles present is called the **mole fraction** of the component.
- The number of moles of solute per litre of final *solution* is called the **molarity** and given the symbol M (this is the most common method but depends on volume, which in turn is temperature dependent).
- The number of moles per kilogram of *solvent* is called the **molality** and is given the symbol m (has the advantage of not being dependent on temperature).
- The percent solute by mass or volume of the final solution.
- % by weight (**w/w**) is the number of grams of solute per 100 g of solution.
- % by volume (**w/v**) is the number of grams of solute per 100 mL of solution.

The limiting concentration of the solution to which a solute can be dissolved in a solvent is called the **solubility**. At this concentration, the solution is said to be **saturated**. Addition of excess solute results in no increase in concentration and a portion of the solute remains undissolved.

5.6 Aqueous Solutions

Many chemical reactions involve chemicals that exist as water, or **aqueous solutions**.

$$\text{Molarity} = \frac{\text{No. moles of solute}}{\text{Volume of solution in litres}}$$

Examples:

(1) Calculate the mass of solute needed to make up 1 litre solution of 0.5M sodium hydroxide:

$$\begin{aligned} 1\text{ mole NaOH} &= 23 + 16 + 1 \\ &= 40\text{ g} \\ \text{mass required for 1 L solution at 0.5M} &= 40(0.5) \\ &= 20\text{ g} \end{aligned}$$

(2) Calculate the volume of 10M acid required to make up 250 mL of 0.4M solution of HCl:

$$\begin{aligned} 1\text{ L} &= 0.4\text{M} \\ &= (0.4)(36.5)\text{ g} \\ &= 14.6\text{ g} \\ 0.25\text{ L} &= 3.65\text{ g HCl required} \\ 10\text{M} &= 365\text{ g/L HCl} \\ 3.65\text{ g HCl} &= 10\text{ mL} \end{aligned}$$

$$\begin{aligned} \text{Cl} &= 35.5 \\ \text{H} &= 1 \end{aligned}$$

It is often observed that for dilute aqueous solutions, the addition of the solute lowers the **freezing point** and raises the **boiling point** (reduction in vapour pressure) compared to pure water.

When an aqueous solution evaporates, a **hydrated ionic compound** may be formed. The hydrated compound contains water of crystallization:

$$\text{Co}^{2+}(\text{aq}) + 2\text{Cl}^-(\text{aq}) \xrightarrow{\text{evap}} \text{CoCl}_2 + 2\text{H}_2\text{O(s)}$$

5.7 Solubility of Solids

When a solute dissolves in a solvent, three distinct processes can be identified:

- Solute-solute interactions – whereby the solute atoms or molecules are separated from each other (such as the dissociation of Na^+ and Cl^- into ions). This is an **endothermic** process (heat is required to break the ionic bonds).
- Solvent-solvent interactions – whereby the solvent atoms or molecules may be separated from each other (such as when H_2O molecules are separated from one another to accommodate the incoming solute molecules). This is an **endothermic** process.
- Solute-solvent interactions – whereby processes occur between the solute and the solvent. These are usually **exothermic** (reduction in total energy).

Some ionic compounds are soluble in water, but insoluble in non-polar solvents (like benzene and carbon tetrachloride). When water is the solvent, a **hydration** process can occur which serves to insulate the positive and negative ions from each other and prevent them reforming into a solid. This is a solute-solvent interaction.

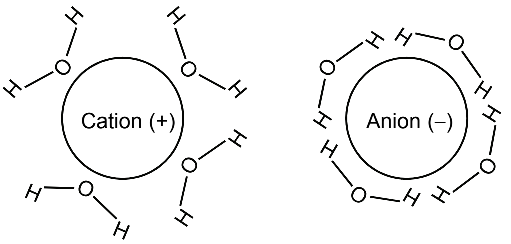

Hydration involves the formation of bonds and so is exothermic (heat is released as the total energy is lowered) and so the solute-solvent interactions dominate the solute-solute interactions. Non-polar solvents cannot hydrate the ions.

Some non-polar molecular solids are soluble in non-polar solvents but are insoluble in water (a polar solvent). Non-polar solute molecules are not hydrated by water, and so solute-solvent interactions are not significant here. If more energy is required to break the hydrogen bonds linking water molecules together than that required to break the bonds holding the components of the solute together, then the solute components remain insoluble. When placed in a non-polar solvent, solvent-solvent interactions may occur and the solvent molecules separate easily to accommodate solute molecules and the solute dissolves.

5.8 Solubility Equilibrium

If a solute is added to the solvent in increasing quantity, a point is reached where the atomic or molecular species being dissolved can no longer be accommodated in the solute. As more solute is added, the reverse process begins to occur so that some of the previously dissolved solute returns to the solute phase. This is called **precipitation**. This is most vividly illustrated when a solid solute is added to a liquid solvent. When the rate of precipitation becomes equal to the rate of **dissolution**, dynamic equilibrium is achieved and the solution is said to be **saturated**.

The **concentration** of a solute in a given solvent at **saturation** is called the **solubility** of the solute in the solvent.

Solubility depends on **temperature**. When a solute is dissolved in a solvent and heat is generated, the solubility generally decreases with increasing temperature. If heat is absorbed, the solubility generally increases with increasing temperature. Often, whether or not heat is absorbed or generated depends on the physical processes involved in any chemical reactions that may occur in the solution.

Solubility depends on **pressure**. For a non-reactive gas being dissolved in a liquid, the solubility C increases as the pressure P increases according to **Henry's law**: $C \propto P$

For example, dissolved CO_2 in a soft drink readily precipitates (observed as bubbles) when the pressure above the liquid is released (such as when the bottle is opened). An easy way to remember Henry's law is to think of it as the soft drink law.

For example, air dissolves in water to some extent and when it does so, heat is produced (since the gas "condenses" into a liquid form). The solubility of air in water thus decreases as the temperature increases. When water is heated, the dissolved air can be seen to "precipitate" as bubbles before the boiling point of water is reached.

One of the most common solvents is water, but not all compounds are soluble in water.

Compound	General rule	Exceptions
Nitrates	Soluble	
Chlorides	Soluble	AgCl, $PbCl_2$, Hg_2Cl_2
Sulphates	Soluble	$PbSO_4$, $BaSO_4$
Carbonates	Insoluble	Na_2CO_3, K_2CO_3, $(NH_4)2CO_3$
Hydroxides	Insoluble	NaOH, KOH, $Ba(OH)_2$, (NH_4OH)
Oxides	Insoluble	Na_2O, K_2O, BaO

5.9 Electrolytes

Some molecular and ionic compounds dissolve in water to give a solution that is able to conduct electric current. The compounds, when dissolved in water, dissociate into **ions**. The ions, being electrically charged, are thus able to move under the influence of an electric field and the movement of these ions is therefore an electric current.

- **Strong electrolytes** dissociate completely into ions. Many ionic compounds are strong electrolytes. Molecular compounds that are strong electrolytes are usually acids.
- **Weak electrolytes** only partially dissociate into ions. Usually, molecular compounds that are weak electrolytes are either an acid or a base.
- **Non-electrolytes** do not dissociate into ions when placed in solution. Molecular compounds that are neither acids or bases are usually non-electrolytes.

Molecular compounds that are **strong electrolytes**:

	Molecular formula	Ions formed in solution	
Hydrochloric acid	HCl	H^+, Cl^-	Strong acid
Nitric acid	HNO_3	H^+, NO_3^-	Strong acid
Sulphuric acid	H_2SO_4	H^+, SO_4^{2-}	Strong acid
Perchloric acid	$HClO_4$	H^+, ClO_4^-	Strong acid

Molecular compounds that are **weak electrolytes**:

	Molecular formula	Ions formed in solution	
Acetic acid	CH_3COOH	H^+, $CH_3CO_2^-$	Weak acid
Benzoic acid	C_6H_5COOH	H^+, $C_6H_5CO_2^-$	Weak acid
Ammonia	NH_3	NH_4^+, OH^-	Weak base

Molecular compounds that are **non-electrolytes**:

	Molecular formula	Ions formed in solution
Ethyl alcohol	C_2H_5OH	
Sucrose	$C_{12}H_{22}O_{11}$	

5.10 Osmosis

Osmosis is a movement of molecules through a **semi-permeable membrane**. A **semi-permeable membrane** is a sheet of substance that can allow some molecules through it but not others. Semi-permeable membranes can be things like **cell walls** in living tissue that contain openings that allow small molecules to pass through but not large molecules.

The concept of **osmosis** depends on the phenomenon of **diffusion**. Diffusion occurs when there is a **concentration gradient**. The concentration gradient drives the diffusion process. It is a fundamental consequence of the **second law of thermodynamics**. Diffusion happens in a wide variety of physical processes (e.g. the formation of the barrier potential at the p-n junction of a **semiconductor**).

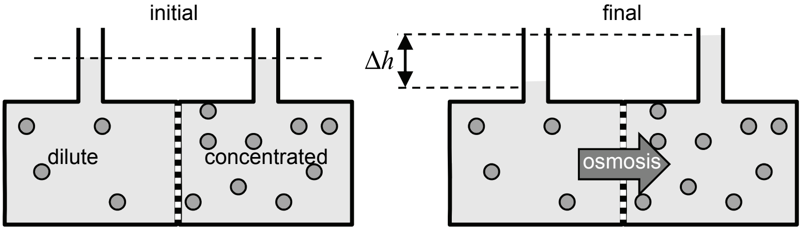

Molecules of solvent pass through to try and make concentration of solute the same in each.

Consider a **dilute solution** on one side of a semi-permeable membrane and a **concentrated solution** on the other side. In this example, the semi-permeable membrane allows molecules of **solvent** to pass through it, but not the **solute**. Be careful. In this case, the **concentration gradient** has to be looked at from the point of view of the solvent. In this case, molecules of solvent tend to pass from the dilute solution into the concentrated solution so as to attempt to achieve uniformity of concentration throughout. Molecules of solute from the concentrated solution would very much like to pass into the dilute side to achieve the same aim, but cannot pass through the membrane. The movement of solvent from one side to the other results in an increase in pressure on the concentrated side. The increase in pressure tends to inhibit the further flow of solvent and an equilibrium condition is reached where no more net movement occurs across the membrane. This pressure increase is called **osmotic pressure**.

For dilute solutions, the osmotic pressure π in atm (atmospheres) of non-electrolytes is given by $\pi = MRT$ where M is the molar concentration of the non-electrolyte, T is in K, and $R = 0.0821$ L atm/mol/K.

5.11 Solids

Generally speaking, there are two kinds of solids. **Crystalline solids** possess a repeating order of atoms, ions, complex ions, or molecules. The repeating pattern is called the **crystal lattice**. The repeating unit in the crystal lattice is called a **unit cell**. **Amorphous solids** have no repeating structure of atoms, ions or molecules.

Crystalline solids can be broadly classified as follows:

Molecular solids are those in which the binding forces between discrete molecules are due to instantaneous van der Waals interactions in addition to dipole-dipole attractions in polar molecules. Covalent bonds exist within the molecules between the atoms.

- Generally soft
- Gas, liquid or solid at room temperature
- Low melting point (about 20 ± 200°C)
- Not usually soluble in water
- Poor conductors of electricity
- Heat insulator

Examples: molecular crystals of noble gases, halogens. Compounds such as CO_2, CH_4, H_2O, C_6H_6, $C_{12}H_{22}O_{11}$ (sucrose) and most organic compounds.

Ionic solids are those in which the binding energy is due to Coulomb electrostatic forces between positively and negatively charged ions in the crystal. That is, ionic solids are composed of **cations** and **anions**.

- Generally hard
- Solid at room temperature
- High melting point (about 800°C)
- Usually soluble in water
- Poor conductors of electricity in solid form
- Heat insulator

Examples: cations from Group I (e.g. Na^+) or Group II (e.g. Mg^{2+}) with anions O^{2-}, S^{2-}, NO_3^-, CO_3^{2-}, ClO_4^-, SO_4^{2-}, CrO_4^{2-}, PO_4^{3-}, H^+, OH^-, MnO_4^-, CN^-.

Covalent solids are those in which the binding energy is due to shared valence electrons (**covalent bonds**) between atoms in the solid.

- Generally hard
- Solid at room temperature
- Very high melting point (about 1200°C)
- Insoluble in water
- Poor conductors of electricity
- Heat insulator

Examples : Si, C (**diamond**), Ge, SiO_2 (**quartz**), BN, SiC.

Metallic solids in which valence electrons are effectively shared amongst all the atoms in the solid.

Amorphous solids have no long-range regular repeating pattern of atoms or molecules. Examples are glass and most plastics. In these materials, there is an orderly structure in the neighbourhood of any one atom, but this is not regularly repeated throughout the material.

5.12 Crystalline Lattice Structures

For geometrical reasons, there are only fourteen types of lattices that satisfy symmetry operations such as translation, rotation, reflection and inversion. Each of these fourteen lattices is called a **Bravais lattice**. There are seven convenient **crystal systems** in the set of Bravais lattices: cubic, tetragonal, orthorhombic, trigonal, monoclinic, hexagonal and triclinic.

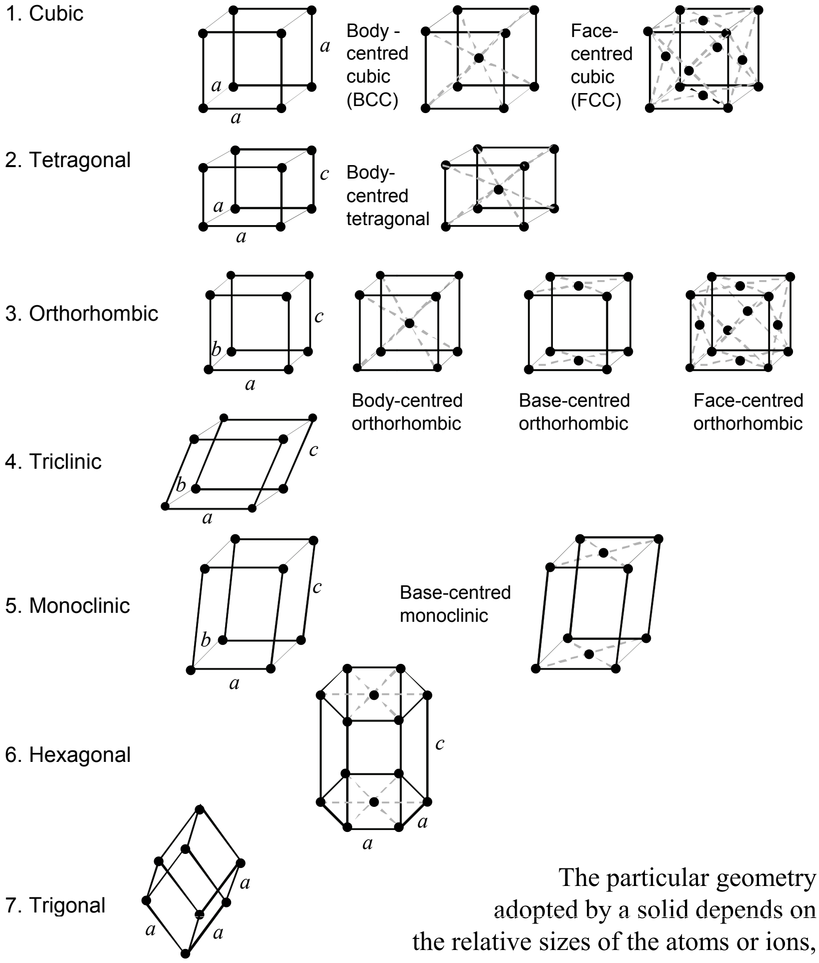
1. Cubic
a
a
a
Body - centred cubic (BCC)
Face- centred cubic (FCC)
2. Tetragonal
c
a
a
Body- centred tetragonal
3. Orthorhombic
c
b
a
Body-centred orthorhombic
Base-centred orthorhombic
Face-centred orthorhombic
4. Triclinic
c
b
a
5. Monoclinic
c
b
a
Base-centred monoclinic
6. Hexagonal
c
a
a
7. Trigonal
a
a
a

The particular geometry adopted by a solid depends on the relative sizes of the atoms or ions, their charge (for ions). Note: The lines shown in these images do not actually exist in a real material, but just serve to illustrate the overall shape of the lattice.

5.13 Metallic Solids

Metallic solids are those in which valence electrons are effectively shared amongst all the atoms in the solid. Atoms are held in place by the **metallic bond**.

- Can be either hard or soft
- Solid at room temperature (except Hg)
- Low to high melting point (room temperature and above)
- Insoluble in water
- Good conductors of electricity
- Heat conductor

Most metals crystallise into either body-centred cubic (BCC), hexagonal close-packed, and face-centred cubic (FCC). Common metals with **BCC** structure are Fe (**iron**), Cr (**chromium**), Mo (**molybdenum**), W (**tungsten**), V (**vanadium**). Common metals with **FCC** structure are Al (**aluminium**), Cu (**copper**), Au (**gold**), Pb (**lead**), Ni (**nickel**), Pt (**platinum**), Ag (**silver**). Common metals with close-packed hexagonal form are Cd (**cadmium**), Co (**cobalt**), Mg (**magnesium**), Ti (**titanium**), Zn (**zinc**). Metals with BCC structure have a higher yield strength than those with FCC structure.

At room temperature, iron has a BCC structure, but above 910°C, iron rearranges into an FCC structure (at which time it becomes non-magnetic). This change in structure takes place in the solid state and is called an **allotropic change**.

Solid solutions are very important in **metallurgy**, where they are called **alloys**. In a true solid solution, one component is randomly dispersed throughout the other component. Other types of alloys are molecular compounds (in fixed proportions) and heterogeneous mixtures. A particularly important alloy is that of iron and carbon, which makes **steel**.

Carbon Content:

0.05–0.3%	Low-carbon steel	Mild steel, tough, ductile, easily forged and welded.
0.3–0.45%	Medium-carbon steel	Strong, hard, not so easily forged or welded.
0.45–0.75%	High-carbon steel	Very strong (high-strength steel) and hard. May be annealed for machining or heat treated for different degrees of hardness.
0.75–1.5%	Very high-carbon steel	

6. Chemical Thermodynamics

Summary

Internal energy: $\Delta U = nC_p(\Delta T)$

Work: $W = p\Delta V$

First Law: $Q_p = \Delta U + p\Delta V$

Enthalpy:
$$\begin{aligned}\Delta H &= \Delta U + p\Delta V \\ &= Q_p\end{aligned}$$

Entropy:
$$\Delta S = \frac{Q}{T}$$

$\Delta S_{\text{total}} > 0$ irreversible process
$\Delta S_{\text{total}} = 0$ reversible process
$\Delta S_{\text{total}} < 0$ impossible process

Gibbs energy: $\Delta G = \Delta H - T\Delta S$

Spontaneous reactions:

$$\Delta G_{\text{reaction}} = \Delta_{\text{f}} G^0{}_{\text{products}} - \Delta_{\text{f}} G^0{}_{\text{reactants}}$$

6.1 Molecular Energy

Molecules in a gas are capable of **independent motion**. Consider a diatomic gas molecule:

(a) the molecule itself can travel as a whole from one place to another

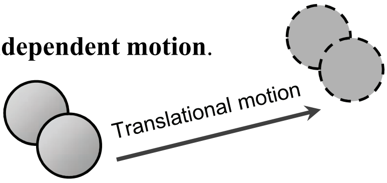

(b) the molecule can spin around on its axis

(c) the atoms within the molecule can vibrate backwards and forwards

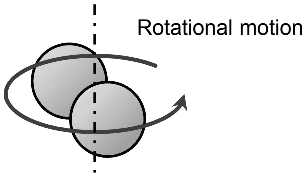

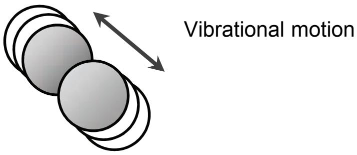

These **kinetic energy** terms represent the **internal energy** and have significance for the interpretation of the **specific heat** capacity of a substance. Changes in internal energy of a molecule manifest themselves physically as changes in **temperature** of the molecule.

The total energy of the molecule also contains some **potential energy** components. Firstly, there are **intermolecular forces** (such as **van der Waals forces**) that lead to the formation of liquids, and hydrogen bonds.

Secondly, there is the potential energy associated with the **chemical bond** between the atoms of the molecule, the **bond energy**.

These potential energy terms are associated with the *position* of the atoms within a molecule, and the position of molecules in relation to other molecules, as distinct from the *motion* of the atoms or molecules.

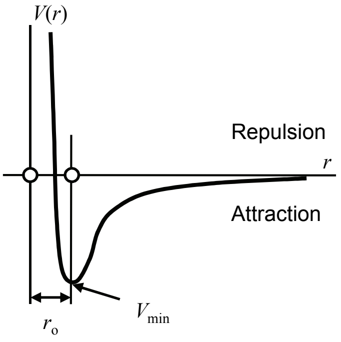

6.2 Specific Heat Capacity

The amount of energy (ΔQ) required to change the temperature of a mass of material is found to be dependent on the:

• Mass of the body (*m*)	kg
• Temperature increase (Δ*T*)	^oC, K
• Nature of the material (*c*)	J/kg$^{/}$/K

↓

$$\Delta Q = mc(\Delta T)$$

Specific heat or **heat capacity**

Material	**c (kJ/kg/K)**
water	4.186
steel	0.45
cast iron	0.54
aluminium	0.92

The **specific heat** is the amount of heat required to change the temperature of 1 kg of material by 1^oC

Molar specific heat, or **molar heat capacity** (*C*), is the amount of heat required to raise the temperature of 1 mole of the substance by 1^oC.

$$\Delta Q = nC(\Delta T)$$

Experiments show that when a gas is heated at constant volume, the **molar heat capacity** C_v is always less than that if the gas is heated at constant pressure C_p. This is because for a constant volume process, there is no opportunity for there to be mechanical work done on or by the system.

For a constant pressure process, for a given temperature rise Δ*T*, there will always be a volume change ΔV. Therefore the energy into the system has to both raise the temperature *and* do work, thus C_p is always greater than C_v.

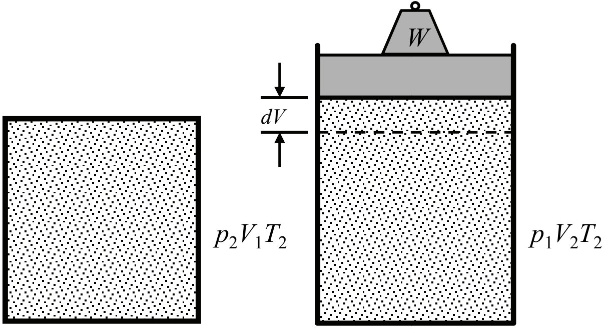

For a constant pressure process, a volume change involves a force acting through a distance and hence work is done on or by the system. Volume change in solids and liquids very small, and hence distinction between C_p and C_v not usually made.

$$Q_v = nC_v(T_2 - T_1)$$

$$Q_p = nC_v(T_2 - T_1) + p(V_2 - V_1)$$
$$= nC_p(T_2 - T_1)$$

↓

Q_p is a very important quantity in industrial processes and is given the name **enthalpy**.

6.3 Enthalpy

Consider a beaker of solution which is heated from an external source. That is, heat energy is transferred into the system across the boundary.

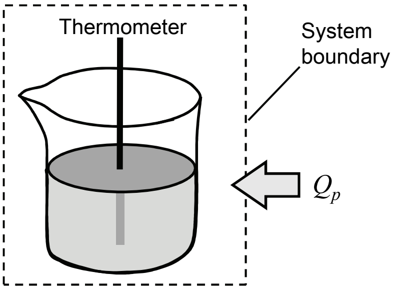

The temperature of the solution increases by an amount dependent on the **molar heat capacity** C_V of the solution. The change in **internal energy** of the system will be:

$$\Delta U = nC_v(\Delta T)$$

The solution may also expand by an amount ΔV and perform mechanical **work** (e.g. by a piston that fits into the top of the beaker and floats on the surface of the solution):

$$W = p\Delta V$$

The relationship between the heat entering the system, the work done by the system and the rise in internal energy is expressed by the **first law of thermodynamics**:

$$Q_p = \Delta U + p\Delta V$$

Q_p is positive when heat enters the system. W is positive when work is done by the system on the surroundings. ΔU is positive when ΔT is positive.

If heat energy Q_p is added to the system, then some of this heat goes into mechanical work and some into internal energy. The sum of the change in internal energy and the work done on or by the system is called the change in **enthalpy** ΔH of the system:

$$\begin{aligned}\Delta H &= \Delta U + p\Delta V \\ &= Q_p\end{aligned}$$

Note: $p\Delta V$ is usually small in comparison to ΔU in experiments in the laboratory whereupon ΔH is approximately equal to ΔU.

That is, the change in enthalpy of the system is the same as the amount of heat energy that has been put into it.

If the heating of the solution involves a **phase change**, then some of the heat energy input would be consumed by latent heat which would have to be accounted for.

If the heating of the solution initiates a **chemical reaction**, then heat energy may be released (e.g. as happens during combustion) or absorbed by the solution as the potential energy of the **chemical bonds** (chemical energy) changes. This is called the **heat of reaction**. The heat of reaction is a release, or absorption of potential energy of the chemical bonds.

6.4 Heat of Reaction

Consider the reaction between two chemical species in an aqueous solution in a beaker which is placed in a temperature controlled water bath so that the beaker is maintained at a constant temperature. That is, heat flow can occur across the system boundary but the solution is held **isothermal**. If, as a result of the reaction, heat energy is released, then this heat must pass out through the system boundary. The surrounding water bath ensures that the heat energy is taken away fast enough to prevent any temperature rise of the solution.

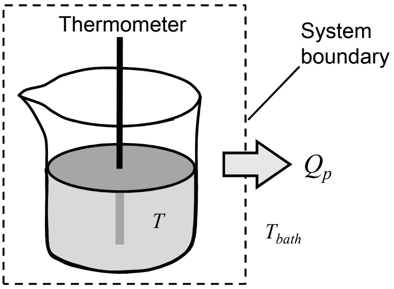

- The **internal energy** of the system remains the same after the reaction as indicated by no rise in temperature.
- **Energy** has come from the **heat of reaction** and been transferred into the water bath.
- If there has been a volume change, then mechanical **work** also would have been performed on (+ve) or by (-ve) the system.

In this example, there must have been a net release of bond energy as a result of the chemical reaction. Some of this energy has been carried away across the boundary by the water bath, and some (a small amount in this case) has gone into mechanical work against the pressure of the atmosphere. The change in **enthalpy** of the system is:

$$\Delta H = \Delta U + p\Delta V$$

We can assume that the mechanical work involved in this example is small compared to the heat lost by the system to the surroundings. Thus, $\Delta H = Q_p$. In this example, heat Q_p has passed out of the system and so the change in enthalpy is a decrease. When heat is passed out to the surroundings, the reaction is said to be **exothermic**. If the reaction involved an absorption of heat from the surroundings, then the change in enthalpy would be positive, and the reaction is called **endothermic**.

The heat of reaction represents the change in enthalpy in this isothermal system. The question now is, if there is a chemical reaction, *and* the temperature changes, how is the change in enthalpy calculated?

6.5 Heat of Reaction

Consider the reaction between two chemical species in an aqueous solution in an insulated beaker. A **thermometer** measures the temperature change during the reaction.

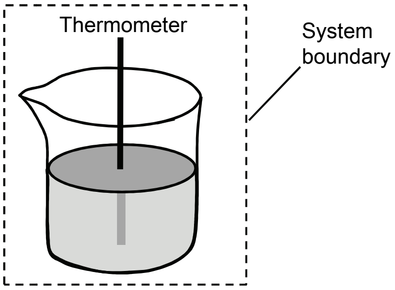

The reaction takes place inside the boundary of a system. For the present example, we will assume that the system boundary is **adiabatic**. That is, no heat flow will occur across the boundary, as if the beaker were fully insulated.

If, as a result of the reaction, the temperature of the solution rises, then:

- The **internal energy** of the system is higher after the reaction, as indicated by a rise in temperature.
- **Energy** has come from the **heat of reaction** and been transferred into internal energy of motion of the molecules in the solution.
- If there has been a volume change, then mechanical **work** also would have been performed on (+ve) or by (-ve) the system.

Since no external heat energy was involved (no heat flow across the system boundary), the heat energy needed to raise the temperature of the solution must have come from the net sum of energies from the breakage and formation of the **chemical bonds** involved in the reaction. In this example, there must have been a net release of bond energy as a result of the reaction. Some of this energy has gone into internal energy of motion of the molecules, while some (a small amount in this case) has gone into mechanical work against the pressure of the atmosphere. In this case, $\Delta H = \Delta U + p\Delta V$. If the work done is negligible, then $\Delta H \approx \Delta U$. In this case, the change in enthalpy is equal to the **heat of reaction**. Although no heat has crossed the system boundary, heat has appeared as a result of the chemical reaction – just as if it had come from outside the system. That is, in this *adiabatic* example, heat Q_p has entered the system from the chemical reaction and $\Delta H \approx \Delta U$.

6.6 Enthalpy of Formation

A chemical reaction is the process of breakage and formation of chemical bonds. Energy is absorbed when a bond is broken, and energy is released when a bond is formed. When a chemical reaction occurs, heat may be either given off, or absorbed to or from the surroundings. That is, the change in enthalpy may be either positive (**endothermic**) or negative (**exothermic**).

Liquid water molecules are made by the reaction between H_2 and O_2 gas molecules:

$$H_2(g) + \frac{1}{2}O_2(g) \to H_2O(aq)$$

The equation is written in terms of the formation of 1 mole of product. Under standard conditions (1 atm pressure and 25ºC) it is found that 286 kJ of heat energy is produced. That is, the change in enthalpy of the system is $\Delta H = -286$ kJ. This is the **heat of reaction**.

When the reactants are at standard conditions, and the products are also kept at standard conditions (e.g. by use of a water bath to maintain temperature), the heat of reaction is referred to as the **heat of formation**, or **enthalpy of formation**. Standard heats of formation are determined by experiment.

Name	Molecular Formula	Heat of Formation ΔH, kJ/mol
Methane	CH_4	-75
Carbon dioxide	CO_2	-394
Ammonia	NH_3	-46
Acetylene	C_2H_2	+227
Ethylene	C_2H_4	+227
Ethane	C_2H_6	-85
Water	H_2O	-286
Hydrochloric acid	HCl	-92
Sodium hydroxide	NaOH	-427
Aluminium oxide	Al_2O_3	-1670
Carbon monoxide	CO	-110

When the reaction proceeds in reverse, the sign of the heat of formation is reversed.

The reference point for heats of formation are elements at standard conditions. For example the heat of formation of $O_{2(g)}$ is set to zero.

Heats of any reaction can be obtained by addition of heats of formations of known compounds, or **Hess' law**. The change in enthalpy for the combustion of ethane is:

$$C_2H_6 + \frac{7}{2}O_2 \to 2CO_2 + 3H_2O$$

$$(C_2H_6(g) \to 2C(s) + 3H_2(g)) \quad \Delta H = +85$$

$$2(C + O_2 \to CO_2) \quad \Delta H = 2(-394)$$

$$3\left(H_2 + \frac{1}{2}O_2 \to H_2O\right) \quad \Delta H = 3(-286)$$

$$\Delta H = 85 - 2(394) - 3(286) = -1561 \text{ kJ/mol}$$

Note that the sign of ΔH for C_2H_6 has been reversed in this calculation since the reaction splits this compound rather than forms it. ΔH for the combustion of ethane is called the **calorific value**.

6.7 Entropy

Consider a beaker of ice which is heated from an external source. The ice remains at 0°C (273K) and melts. The amount of energy Q_p needed is determined by the **latent heat of fusion** of ice.

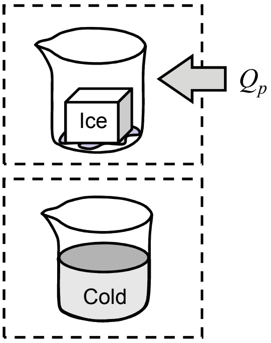

The water molecules have a greater amount of disorder after melting compared to that before heating due to their increased velocity. A quantitative measure of disorder is **entropy**.

Absolute values of entropy S can be calculated from thermodynamic theory or measured in the laboratory. Of more interest to us in chemistry are changes in entropy as a result of physical changes of state, and chemical reactions. When a system undergoes a process at temperature T involving heat flow, the *change* in entropy of the **system** is:

$$\Delta S = \frac{Q}{T}$$

When a system absorbs heat, Q is positive and so the entropy increases, $\Delta S>0$. When a system rejects heat, Q is negative and the entropy decreases, $\Delta S<0$.

If the process occurs at a low temperature, then the change in entropy is greater than if the same process occurs at a higher temperature. At high temperatures, the entropy is already high and so the proportional change in entropy for a given heat flow Q is less compared to that which would occur at low temperatures.

For 1 mole ice:

$$\begin{aligned}\Delta S &= \frac{mL}{T}\\ &= \frac{0.018(335000)}{273}\\ &= 22.1 \text{ J/K}\end{aligned}$$

When a system undergoes a process and interacts with its surroundings (i.e. a non-isolated system), the total **entropy change** (of the universe) is:

$$\Delta S_{total} = \Delta S + \Delta S_{\text{surroundings}}.$$

The sign of ΔS_{total} signifies the presence of the following types of processes:

$\Delta S_{\text{total}} > 0$ irreversible process
$\Delta S_{\text{total}} = 0$ reversible process
$\Delta S_{\text{total}} < 0$ impossible process

Entropy is not energy. There is no "law of conservation of entropy". Irreversible processes *create* entropy.

When a phase change or a chemical reaction involves heat flow to or from the surroundings, entropy must be considered. The **total entropy** of the system and the surroundings cannot decrease. If the total entropy change is zero, then the process is reversible and so cannot be **spontaneous** in any one direction. For a reaction to proceed spontaneously from reactants to products, the total entropy ΔS_{total} must increase.

6.8 Entropy Calculations

Changes in entropy of a system at constant temperature are expressed as:

$$\Delta S = \frac{Q}{T}$$

When the temperature is not constant, we add up incremental changes in entropy for incremental transfers of heat:

$$dQ = mcdT$$

$$\begin{aligned}\Delta S &= \int_{Q_1}^{Q_2} \frac{1}{T} dQ \\ &= \int_{T_1}^{T_2} \frac{1}{T} mc\, dT \\ &= mc \ln \frac{T_2}{T_1}\end{aligned}$$

The **standard molar entropy** S^{o} is the entropy content of one mole of substance under standard conditions of 1 atm pressure and 25°C. Standard entropies are absolute values and represent the change in entropy to bring the substance from 0K to 273K in its standard state. Standard entropies are always positive. The zero point for entropy is 0K. This statement of zero point for entropy is called the **third law of thermodynamics**.

For a chemical reaction or a physical process, the change in entropy can be calculated from:

$$\Delta S_{\text{reaction}} = S^{\text{o}}{}_{\text{products}} - S^{\text{o}}{}_{\text{reactants}}$$

For example:

$$\text{H}_2(\text{g}) + \frac{1}{2}\text{O}_2(\text{g}) \rightarrow \text{H}_2\text{O(g)}$$

$$\begin{aligned}\Delta S^{\text{o}} &= 188 - (131 + 0.5 \times 205) \\ &= -45.5 \text{ J/K/mol}\end{aligned}$$

Element	S^{o}. (J/K/mol)
$H_2(g)$	131
$O_2(g)$	205
$N_2(g)$	192
Si(s)	19
C(s)	6
Fe(s)	27

Compound	S^{o} (J/K/mol)
$NH_3(g)$	192.8
$H_2O(l)$	70.0
$H_2O(g)$	188.8
$CO_2(g)$	213.8
CO(g)	197.7
$CH_4(g)$	186.2

Note that in this example, the change in entropy for this system is < 0 at the expense of an increase in entropy elsewhere in the universe. Hydrogen and oxygen will, under certain conditions, form water since it is energetically favourable to do so.

6.9 Gibbs Energy

It is a natural tendency of systems to attain the lowest possible potential energy, such as a ball rolling down a hill, electrons moving in an electric field, and so on. In an **exothermic** process, heat leaves a system ($\Delta H < 0$) and we might expect that such processes are always spontaneous and that **endothermic** processes are not. However, this is not observed. For example, when water evaporates spontaneously at room temperature, $\Delta H > 0$ and the water vapour has a higher energy than the liquid.

$$H_2(g) + \frac{1}{2}O_2(g) \rightarrow H_2O(aq) \quad \Delta H = -286\text{ kJ/mol}$$

$$H_2(g) + \frac{1}{2}O_2(g) \rightarrow H_2O(g) \quad \Delta H = -242\text{ kJ/mol}$$

$$H_2O(aq) \rightarrow H_2O(g) \qquad \Delta H = 286 + -242 = +44\text{ kJ/mol}$$

+ΔH signifies energy of products is greater than reactants (endothermic)

Whether or not a phase change or a chemical reaction proceeds spontaneously cannot be judged on the basis of energy alone. The change in entropy ΔS_{total} of both the system (the reaction) and the surroundings must also be considered. Whether or not the change in entropy or enthalpy is the more important factor in any particular process depends on the absolute temperature T.

Now, the total energy of a system is given by the **enthalpy** and this consists of the internal energy and the external work. Changes in total energy are expressed as: $\Delta H = \Delta U + p\Delta V$.

Part of the total energy H is unavailable for conversion into work. This part is connected with changes in **entropy**. Since the units of entropy are J/K, and the determining factor that quantifies the contribution of entropy in a particular process is T, then the unavailable energy can be calculated from the product TS. Thus, the total energy of a *system* can be written as:

$$H = G + TS$$

where G represents that part of the total energy that is freely available for conversion into work. This freely available energy is called the **Gibbs energy**. As the temperature increases, the amount of Gibbs energy decreases and the unavailable term TS increases. Changes in the amount of Gibbs energy in a constant temperature process are thus: $\Delta G = \Delta H - T\Delta S$.

The concept of Gibbs energy allows us to predict the direction of spontaneous chemical reactions by considering both energy and entropy. A spontaneous chemical reaction will proceed only when there is a decrease in available Gibbs energy: $\Delta G < 0$.

6.10 Spontaneous Processes

The concept of Gibbs free energy embodies both energy and entropy considerations in determining in which direction a chemical reaction proceeds. A decrease in Gibbs energy can arise due to a decrease in **enthalpy** or an increase in the product TS. In a particular chemical reaction, ΔH and ΔTS may have the same or opposite signs. For a constant pressure process, at constant temperature:

$$\Delta G = \Delta H - T\Delta S.$$

- If a reaction is exothermic $\Delta H < 0$ and the entropy change $\Delta S > 0$, then the free energy change $\Delta G < 0$ and the process is **spontaneous**.
- If a reaction is endothermic $\Delta H > 0$ and the entropy change $\Delta S < 0$, then the free energy change $\Delta G > 0$ and the process is not spontaneous.
- If the enthalpy change $\Delta H > 0$ and the entropy change $\Delta S > 0$, then the process is spontaneous at high temperature. When $\Delta H < 0$ and $\Delta S < 0$, the process is spontaneous at low temperature.

It is important to know that ΔS in the above equation is the entropy change of the *system*, not the total entropy change (of the universe). This enables us to conveniently predict and quantify the direction of a process based upon attributes of the system only.

In chemical reactions, the change in Gibbs energy depends on the physical states of the reactants and products, the temperature and pressure, and the concentration of the reactants and products. When reactants and products are at standard conditions (1 atm and 25°C), the free energy change is referred to as the **standard Gibbs energy change** ΔG^o.

The standard Gibbs energy change of the formation of a compound from its elements is called the **standard Gibbs energy of formation** $\Delta_f G^o$. Elements at their standard states have zero standard Gibbs energy of formation.

For a process or a chemical reaction, the change in Gibbs energy can be calculated from:

$$\Delta G_{\text{reaction}} = \Delta_f G^o{}_{\text{products}} - \Delta_f G^o{}_{\text{reactants}}$$

Compound	$\Delta_f G^o$. (kJ/mol)
NH_3(g)	−16.6
H_2O(l)	−237.2
H_2O(g)	−228.6
CO_2(g)	−394.3
CO(g)	−137.3
CH_4(g)	−50.8

6.11 Melting of Ice

Consider a mole of ice initially at 0°C. Now, let the temperature of the surroundings be 283K (+10°C).

For the system:
$$\begin{aligned}\Delta H &= mL + mc\Delta T\\ &= 0.018(335000) + 0.018(4186)(10)\\ &= 6780\text{ J} \qquad \text{(endothermic)}\end{aligned}$$

For the surroundings:
$$\Delta S_{\text{surroundings}} = \frac{-6780}{283} = -23.9\text{ J/K}$$

For the system: $dQ = mL + mcdT$

$$\begin{aligned}\Delta S &= \frac{mL}{T_1} + \int_{Q_1}^{Q_2} \frac{1}{T}\,dQ\\ &= \frac{mL}{T_1} + \int_{T_1}^{T_2} \frac{1}{T}\,mc\,dT\\ &= \frac{0.018(335000)}{273} + 0.018(4186)\ln\frac{283}{273} = 24.8\text{ J/K}\end{aligned}$$

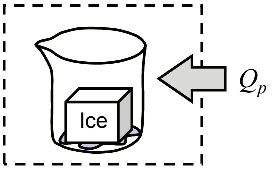

Total change in entropy:

$$\begin{aligned}\Delta S_{\text{total}} &= \Delta S + \Delta S_{\text{surroundings}}\\ &= 24.8 - 23.9\\ &= +0.9\text{ J/K}\end{aligned}$$

This is an **irreversible process**. Even though the change in entropy is < 0 for the *surroundings*, the change in entropy of the system is greater, and so the change in **total entropy** is > 0.

$$\begin{aligned}\Delta G &= \Delta H - T\Delta S\\ &= 6780 - 283(24.8)\\ &= 6780 - 7018\\ &= -238\text{ J}\end{aligned}$$

The **Gibbs free energy** change is < 0 so this process is spontaneous.

Data:
c_{water} = 4186 J/kg/K
c_{Al} = 920 J/kg/K
$L_v = 22.57 \times 10^5$ J/kg
$L_f = 3.35 \times 10^5$ J/kg
ρ_{water} = 1000 kg/m^3
c_{ice} = 2110 J/kg/K

6.12 Freezing of Water

Consider a mole of water, initially at 0°C. Now, let the temperature of the surroundings be 263K (−10°C).

For the system:
$$\begin{aligned}\Delta H &= -mL + -mc\Delta T\\ &= -0.018(335000) + -0.018(2110)(10)\\ &= -6.03 - 0.38\\ &= -6410\ \text{J} \qquad \text{(exothermic)}\end{aligned}$$

For the surroundings:
$$\Delta S_{\text{surroundings}} = \frac{6410}{263} = 24.4\ \text{J/K}$$

For the system: $dQ = -mL + mcdT$

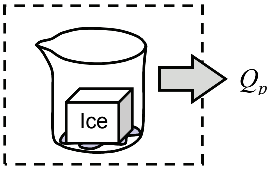

$$\begin{aligned}\Delta S &= \frac{-mL}{T_1} + \int_{Q_1}^{Q_2} \frac{1}{T}\,dQ\\ &= \frac{-mL}{T_1} + \int_{T_1}^{T_2} \frac{1}{T}\,mc\,dT\\ &= \frac{-0.018(335000)}{273} + 0.018(2110)\ln\frac{263}{273} = -23.52\ \text{J/K}\end{aligned}$$

Total change in entropy:
$$\begin{aligned}\Delta S_{\text{total}} &= \Delta S + \Delta S_{\text{surroundings}}\\ &= -23.52 + 24.4\\ &= +0.88\ \text{J/K}\\ &> 0\end{aligned}$$

This is an **irreversible process**. Even though the change in entropy is < 0 for the *system*, the change in entropy of the surroundings is greater, and so the change in **total entropy** is > 0.

$$\begin{aligned}\Delta G &= \Delta H - T\Delta S\\ &= -6410 - 263(-23.5)\\ &= -6410 + 6180\\ &= -229\ \text{J}\end{aligned}$$

This is a spontaneous process since the freezing of ice results in a decrease in **Gibbs free energy**.

Data:
c_{water} = 4186 J/kg/K
c_{Al} = 920 J/kg/K
L_v = 22.57 × 10^5 J/kg
L_f = 3.35 × 10^5 J/kg
ρ_{water} = 1000 kg/m^3
c_{ice} = 2110 J/kg/K

6.13 Ice/Water Equilibrium

Consider the conversion of one mole of water to ice at 273K:

For the system: $\Delta H = -mL$

$$= -0.018(335000)$$

$$= -6030\text{ J} \quad \text{(endothermic)}$$

For the surroundings:

$$\Delta S_{\text{surroundings}} = \frac{6030}{273} = 22.08\text{ J/K}$$

For the system: $dQ = -mL$

$$\Delta S = -\frac{mL}{T}$$

$$= \frac{-0.018(335000)}{273}$$

$$= -22.08\text{ J/K}$$

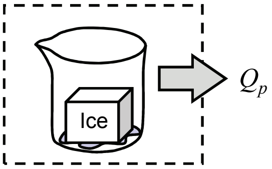

Total change in entropy:

$$\Delta S_{\text{total}} = \Delta S + \Delta S_{\text{surroundings}}$$

$$= -22.08 + 22.08$$

$$= 0\text{ J/K}$$

The total change of entropy is zero. This is a **reversible process**. Extraction of Q_p from the melted ice would cause it to re-freeze.

$$\Delta G = \Delta H - T\Delta S$$

$$= -6030 - 273(-22.08)$$

$$= 0\text{ J}$$

This process is at **phase equilibrium**.

Data:
c_{water} = 4186 J/kg/K
c_{Al} = 920 J/kg/K
L_v = 22.57 × 10^5 J/kg
L_f = 3.34 × 10^5 J/kg
ρ_{water} = 1000 kg/m^3
c_{ice} = 2110 J/kg/K

6.14 Chemical Equilibrium

A chemical reaction involves reactants and products. Whether or not a chemical reaction proceeds spontaneously depends on the sign of the change in free energy.

$$\Delta G_{\text{reaction}} = \Delta_{\text{f}} G^0{}_{\text{products}} - \Delta_{\text{f}} G^0{}_{\text{reactants}}$$

$$\text{Reactants} \rightarrow \text{Products}$$

If a chemical reaction proceeds from reactants to products spontaneously, we may at first expect the process to continue until the reactants are all converted to product. This is not observed. Usually, reactions proceed until there is **chemical equilibrium**:

$$\text{Reactants} \leftrightarrow \text{Products}$$

At equilibrium, the rate of forward reaction = rate of reverse reaction.

In chemical reactions, the **free energy** change depends on the temperature, physical state, pressure, and concentrations of the reactants and the products. That is, as the reaction proceeds, the reactants are consumed and the products are created – and the free energy change changes as the reaction proceeds in time. When the free energy change becomes equal to zero, the rate of forward reaction has become equal to the rate of the reverse reaction. This usually happens (for spontaneous reactions) when the concentration of the products is somewhat greater than those of the reactants – but it is important to note that this does not necessarily mean that all the products are used up. At **chemical equilibrium**, the forward and reverse reactions are both happening at the same rate.

Precisely what is meant by **rate of reaction** depends upon the context. Rates of reaction may be expressed as the change in concentration of one of the reactants per unit time where concentration is expressed in terms of moles per unit volume.

State	Concentration units
Gases	Partial pressure units
Liquids	Molarity (moles per litre)
Solids	Moles or mass units

The appropriate time period may be expressed in milliseconds, seconds, days or years depending on the reaction.

6.15 Statistical entropy

The change in entropy of a system can be calculated from macroscopic quantities, Q and T. Molecules of gas in a volume V at temperature T have a distribution of velocities, and hence, kinetic energies. The distribution of velocities over the total number of molecules was determined by Boltzmann:

The number of molecules N with an energy E was computed by Boltzmann using the **Maxwell velocity distribution** function:

$$N = Ce^{\frac{-E}{kT}}$$ Maxwell-Boltzmann *energy* distribution

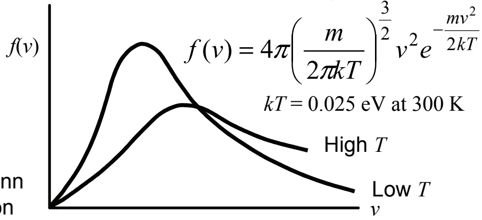

Consider a partitioned volume V containing N molecules. There is a discrete number of energies possible, and there is a finite number of ways that the molecules present can have a distribution of energies that matches the macroscopic properties of P, T and V. Each of these possibilities is called a **microstate**.

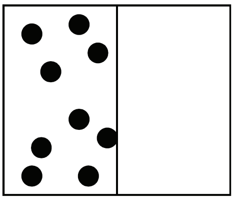

Now, one such possible state is if all the energy of the system were concentrated in one molecule and the others had zero energy. We know that this is very unlikely to ever happen, just as we know that the likelihood of throwing zero heads and 100 tails in 100 tosses of a coin is unlikely. The most likely scenario for our N molecules is that the energies will follow the Maxwell-Boltzmann distribution.

In mathematical symbols, the **statistical weight function** of the system $\Omega(E,V,N)$ is the number of possible microstates that are consistent with the observed **macrostate**. The lower the number of possible microstates, the more accurately we know information about the molecules. The entropy of the system is:

$$S = k \ln \Omega.$$

If there were only one possible microstate, then S would be zero. As the number of possible microstates increases, the entropy becomes larger. When the partition is removed, the molecules rush to fill the entire volume. It is exceedingly unlikely that at any one instant, all the molecules will be located over on one half of the now larger volume. The number of possible microstates has increased. However, one might ask what if the container had only two molecules in it? In this case, it is likely that at any one time, the two molecules *will* find themselves over to one side of the container. The number of possible microstates, for the two-molecule scenario, is very small, and so is the entropy.

7. Rates of Reaction

Summary

Rates of reactions:

- The nature of the reactants
- The concentration of the reactants
- The temperature
- The presence of a catalyst

Rate law (concentrations):

$$aA + bB \rightarrow P$$

$$\frac{d[P]}{dt} = k[A]^n[B]^m$$

Rate law (temperature):

$$k = Ae^{\frac{-E}{RT}}$$ Arrhenius equation

Catalyst:

- Has no effect on the stoichiometry of the overall reaction.
- Does not affect the equilibrium position of the reaction.
- Affects the reaction mechanism so as to lower the activation energy, thus increasing the number of molecules that can make a productive collision.

7.1 Rates of Reaction

Consider the reaction between hydrogen gas and oxygen gas to produce liquid water:

$$H_2(g)+\frac{1}{2}O_2(g)\rightarrow H_2O(aq) \quad \Delta H=-286\text{ kJ/mol}$$

$$H_2(g)+\frac{1}{2}O_2(g)\rightarrow H_2O(aq)$$

$$\begin{aligned}\Delta S^{o} &= 70-(131+0.5*205)\\ &= -163.3\text{ J/K/mol}\end{aligned}$$

$$\begin{aligned}\Delta G &= \Delta H - T\Delta S\\ &= -286000-273(-163.3)\\ &= -286000+44580.9\\ &\approx -240\text{ kJ/mol}\end{aligned}$$

The indications are that this reaction should proceed spontaneously when these two gases are mixed. While this might be so, experience indicates that this reaction does not proceed at all unless there is a spark or source of ignition present – in which case the reaction proceeds with explosive violence.

Clearly then, the **rates of reaction** are an important part of chemical reactions. While thermodynamics might provide information about the direction in which a reaction might proceed on its way to chemical equilibrium, more information is needed to determine at what rate the products are produced.

Associated with the rate of reaction is the **reaction mechanism**. The production of products and the consumption of reactants, while observable, only represents the net reaction in general. Chemical reactions may proceed in several steps. Intermediate species may form and be consumed, and these intermediate steps often influence the overall rate of the reaction as a whole.

The most important factors that govern the rate of a reaction are

- The nature of the reactants
- The concentration of the reactants
- The temperature
- The presence of a catalyst

The rate of reaction is the rate of change of **concentration** of either the reactants or the products.

7.2 Collision Theory

Chemical reactions involve the breaking and formation of bonds between atoms, ions or molecules. Elements or compounds that comprise the reactants form new compounds as products. The relationship between energy and entropy – the **Gibbs free energy** – determines which direction a chemical reaction will proceed. The question now is to ask just how a chemical reaction takes place.

Central to the answer of this question is the fact that atoms, ions and molecules are in constant motion – especially when they are in liquid and gaseous form. Because they are in motion, they undergo collisions with other atoms and molecules.

When atoms, ions or molecules collide, a chemical reaction may occur. That is, chemical reactions occur as a result of collisions between molecules. This is known as **collision theory**.

The notion of chemical reactions only taking place between atoms or molecules when they contact each other in a certain way is not so obvious as it might seem and was only proposed in the early 1900s.

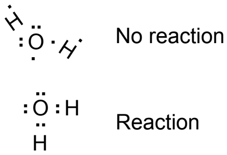

During the collision, bonds are broken and formed as atoms and electrons are rearranged into new compounds. Several variables serve to determine how fast reactants are converted into products:

- Particles (either atoms, ions or molecules) have to collide in a particular way. That is, the orientation of the colliding particles has to be a certain way in order that the collision be effective at breaking bonds and forming new ones.
- The particles have to possess enough kinetic energy to allow them to interact to the extent that bonds may be broken and allow new ones to be formed.
- The number of collisions per unit time determines how fast the reaction proceeds and this is related to the concentration of the reactants.

The nature of the reactants, the temperature, and concentration and action of **catalysts** can be all explained by collision theory. Not all collisions result in a reaction. Only a small proportion are **productive collisions**.

7.3 Reaction Mechanism

We have agreed that for a chemical reaction to proceed, atoms, ions, or molecules must collide. Further, they must not just touch or brush past one another. They must collide with a certain amount of kinetic energy, and be in the right orientation with respect to one another in order for bonds to be broken and allow new ones to form. That is, there has to be a **productive collision**.

Consider the rusting of iron. Simply put, the reaction is:

$$4\mathrm{Fe(s)} + 3\mathrm{O_2(g)} \rightarrow 2\mathrm{Fe_2O_3(s)}.$$

This is a reasonably slow reaction. However, if one were to just take collision theory at face value, it would be luck indeed if four atoms of iron were to collide with three molecules of oxygen at the same time to form two molecules of iron oxide. In reality, reactions usually proceed as a series of smaller steps. The series of smaller steps is called the **reaction mechanism**.

$$\mathrm{Fe^{2+}} + 2\mathrm{OH^-} \rightarrow \mathrm{Fe(OH)_2}$$

$$4\mathrm{Fe(OH)_2} + \mathrm{O_2} + 2\mathrm{H_2O} \rightarrow \mathrm{Fe(OH)_3(s)}$$

$$\mathrm{Fe(OH)_3(s)} \rightarrow \mathrm{Fe_2O_3, H_2O} + 2\mathrm{H_2O}$$

The chances of a collision occurring between two atoms, ions or molecules is significantly greater than the chances of a collision between three or four of the required chemical species at the same time. Most reaction steps in chemical reactions involve two atoms or molecules. The slowest step in the reaction mechanism is the **rate-determining step**. The nature of the reaction steps cannot be deduced from the net reaction. The reaction steps can only be found by experiment.

Identification of the rate-determining step is important for establishing a quantitative rate law for the reaction as a whole. Often, the rates of reaction depend on the concentration of the reactants, but it is the concentration of the reactants within the rate-determining step that is they key issue. During the **elementary reactions**, or the **reaction steps**, intermediate chemical species may be formed which are used up in a later step and do not appear in the net equation.

Some reaction mechanisms are noted for their explosive properties. In these cases, the steps become a self-sustaining **chain reaction** (such as in **combustion**).

7.4 Activation Energy

A reaction step usually involves a two-particle collision (either atoms, ions or molecules). As two molecules approach each other, the outer electrons of one molecule will begin to feel the electrostatic repulsion of the other molecule. The kinetic energy of motion will be converted into electrical potential energy. The potential energy rises as the molecules come closer and closer together.

When all the kinetic energy has been expended, the molecules are in very close proximity and are in a somewhat unstable state. The molecules at this point form what is called an **activated complex**. The activated complex is a conglomeration of the participating atoms. At this point, the activated complex may split into the product molecules (the forward reaction) or bounce backwards and reform the original molecules (the reverse reaction).

The colliding molecules require sufficient **kinetic energy** to overcome mutual repulsion and allow the collision and formation of the activated complex. This is called the **activation energy** E_{act}. The activation energy depends on the nature of the reactants.

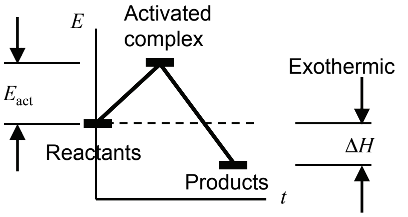

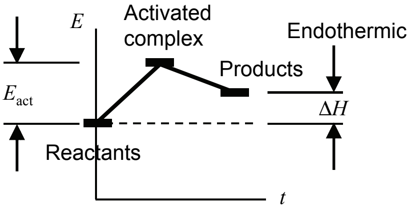

When a spark is introduced into a mixture of H_2 and O_2, the energy of the spark overcomes the activation energy and allows the reaction to proceed. The heat of the reaction furnishes the required activation energy for further reactions. That is, the H_2 gas undergoes **ignition**.

The activation energy is the difference in energy between the activated complex and the average energy of the reactants.

If a particular reaction has a large activation energy, then only a small proportion of collisions may result in enough kinetic energy to overcome it, and the reaction will be slow. If a particular reaction has a lower activation energy, then a greater proportion of collisions will have enough kinetic energy to overcome it and the reaction proceeds at a faster rate.

7.5 Nature of Reactants

The rate of a chemical reaction, that is, the rate of change of concentration of either the reactants or the products, depends on the nature of the reactants. When the permanganate ion is reduced by an Fe^{2+} ion in an acidic solution, the reaction is very fast and is limited by the time it takes to mix the solutions:

$$5Fe^{2+}(aq) + MnO_4^-(aq) + 8H^+(aq) \rightarrow 5Fe^{3+}(aq) + Mn^{2+}(aq) + 4H_2O \qquad \text{fast}$$

But when the permanganate ion is reduced by oxalic acidic $H_2C_2O_4$, the reaction is fairly slow, taking several minutes.

$$5C_2O_4^{2-}(aq) + 2MnO_4^-(aq) + 16H^+(aq) \rightarrow 10CO_2(g) + 2Mn^{2+}(aq) + 8H_2O \qquad \text{slow}$$

The main difference between the two reactions is the reducing agent. It could be argued that in the case of $H_2C_2O_4$, many more bonds must be broken compared to the case of the Fe^{2+} solution and this might influence the rate of the reaction. That is, the reaction mechanism for the second reaction may be quite different in character than that of the first reaction.

Experiments indicate that:

- Reactions that involve the breaking and formation of **covalent bonds** are usually slow at room temperature.
- Reactions that involve electron transfer, such as **ionic reactions** in aqueous solutions, are generally fast. Examples are precipitation reactions, **acid** on **base** reactions, most oxidation/reduction reactions.

Since the **activation energy** depends on the chemical species involved, then so does the rate of reaction. There is more chance of a successful or productive collision in the case of a low activation energy because a greater number of particles will have sufficient kinetic energy to overcome it.

It is not only the chemical nature of the reactants, but the physical nature as well. Since collision theory requires atoms or molecules to make contact to form the activated complex, any mechanism that affects the degree of contact between atoms or molecules will affect the rate of reaction. In homogenous systems, such as aqueous solutions, there is intimate contact, whereas in inhomogeneous systems, such as solid/liquids, or solid/gas, reactions only take place at the surface of contact of the atoms or molecules. Increasing the surface area by mixing or grinding increases the rate of reaction. A common example is that sawdust is almost explosive in air, while a block of wood burns relatively slowly in comparison.

7.6 Concentration

One of the most important effects on the rate of reaction is the **concentration** of one or more of the reactants. Generally speaking, increasing the concentration of a reactant increases the rate of the reaction (but not always). It is easy to accept that the chances of a successful, or productive, collision between two reacting particles is increased if there are more of them, per unit volume, involved in the process. Consider a reaction:

$$A + B \rightarrow AB.$$

If we only had one molecule of A and one of B, then we may have to wait some time before they encounter each other and collide in a way so as to produce the reaction. The collision rate, and hence the reaction rate, would be low.

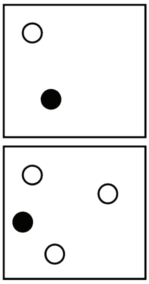

If we now increase the concentration of A, but putting more molecules of A in the reaction chamber, then there would be an increase in the chances of a successful collision – the reaction rate would increase. It is possible that doubling the concentration of one or other of the reactants would double the reaction rate.

The concentration of a species in moles per litre can be conveniently written using [ ] notation. Thus, the reaction rate is indicated by the rate of change of appearance of the product, or the rate of change of concentration of AB.

Reaction rate is expressed as moles/litre/second

$$\frac{d[\mathrm{AB}]}{dt} = k[\mathrm{A}][\mathrm{B}]$$

This equation is called a **rate law**. *k* is the constant of proportionality, or the **rate constant**. It should be noted that the rate law actually observed for a reaction cannot be necessarily deduced from the balanced net equation. This is because one of the intermediate reaction steps may be the rate-determining reaction, in which case the rate law for that step would be required. Unless more of the reactants are continually added, the rate of reaction will generally decrease with time as the reactants are consumed (and their concentrations decrease).

7.7 Rate Law

Consider the reaction: $A + A \rightarrow A_2$

$$2A \rightarrow A_2 \quad \text{Balanced equation}$$

The **rate law** would be written:

$$\frac{d[A_2]}{dt} = k[A][A]$$

$$= k[A]^2 \quad \text{Rate law}$$

Note that the exponent of [A] in the rate law is the same as the coefficient of A in the reactants in the balanced equation. This is a second-order reaction.

In general, for an equation of the form: $aA + bB \rightarrow P$ the rate law is written as:

$$\frac{d[P]}{dt} = k[A]^n[B]^m.$$

For a single-step reaction, the exponents *n* and *m* are equivalent to *a* and *b*. However, a chemical reaction nearly always proceeds via a **reaction mechanism**, that is, a series of reaction steps. Thus, in most cases, there is no connection between the exponents *n*, *m* and the coefficients *a*, *b* for the net reaction. In general, the rate law for the net reaction has to be determined by experiment.

The exponents *n* and *m* indicate the **order of the reaction** with respect to that reactant. For example, when $n = 2$ and $m = 1$, we say that the reaction is second order with respect to A and first order with respect to B. The overall order of the reaction is the sum $m+n$. When *n* or *m* equals zero, then the rate of reaction is independent of the concentration of the reactant and is a **zero-order reaction** with respect to that reactant.

The **rate constant**, *k*, is specific to the particular reaction and depends on the experimental conditions. The units of *k* depend on the order of the reaction. A high value of *k* indicates fast reaction. For first-order reactions, if $k > 10\ \text{sec}^{-1}$, then the reaction is considered instantaneous.

The **rate law** shows how the instantaneous rate of a reaction changes with the concentration of the reactants, as long as other factors such as temperature, the nature of the reactants and the presence of catalysts remain constant.

The rate of a reaction with respect to either reactants or products is expressed in terms of the **rate constant** *k* and the **order** of the reaction.

7.8 Rates of Reactions

There are several methods of expressing the rates of a reaction:

Average rate of disappearance of a reactant R: $-\frac{[\mathrm{R}]_2-[R]_1}{t_2-t_1}$

Average rate of appearance of a product P: $\frac{[\mathrm{P}]_2-[P]_1}{t_2-t_1}$

Instantaneous rate of disappearance of a reactant: $-\frac{d[\mathrm{R}]}{dt}$

Instantaneous rate of appearance of a product: $\frac{d[\mathrm{P}]}{dt}$

Half-life (the time taken for half of a reactant to be consumed or product to be formed): $t_{1/2}$

The rate of formation of products P as a whole is equal to the negative of the rate of disappearance of the reactants R:

$$\frac{d[\mathrm{P}]}{dt}=-\frac{d[\mathrm{R}]}{dt}$$

But we must be careful in expressing the rate in terms of a single product or reactant. For example, consider the following reaction:

$$\mathrm{H}_2(\mathrm{g})+\mathrm{I}_2(\mathrm{g})\rightarrow 2\mathrm{HI}(\mathrm{s}).$$

Two moles of HI are produced at the rate of disappearance of 1 mole of H_2 per second. Thus, the rate of change of concentration, from **stoichiometry**, is

$$\frac{d[\mathrm{HI}]}{dt}=-2\frac{d[\mathrm{H}_2]}{dt}.$$

The rate at which H_2 disappears is the same as the rate at which I_2 disappears, and is half the rate at which HI appears. We cannot determine the order of the reaction from this net equation; we can only determine the relative magnitudes of the rates of products and reactants, no matter what the order for each participant might be.

7.9 Determination of Order

One of the most common ways of determining the rate law for a reaction is to measure the **initial rate of reaction** (e.g. formation of products) while varying the concentration of one of the reactants and holding the concentration of the other reactants constant.

$$\frac{d[\mathrm{P}]}{dt} = k[\mathrm{A}]^n[\mathrm{B}]^m$$

Let us hold the concentration of B constant and express the rate as the rate of change of concentration of A:

$$-\frac{d[\mathrm{A}]}{dt} = k'[\mathrm{A}]^n$$

$$\log\frac{d[\mathrm{A}]}{dt} = -n\log[A] - \log k'$$

By plotting the log of the initial reaction rate (d[A]/d*t*) versus log of the initial concentration [A], then the slope of the resulting straight line will be the order *n*.

Similarly, we may hold the concentration of A constant and vary the concentration of B to obtain a value for *m*.

$$-\frac{d[\mathrm{B}]}{dt} = k''[\mathrm{B}]^m$$

$$\log\frac{d[\mathrm{B}]}{dt} = -m\log[B] - \log k''$$

It is important to know that the concentrations, and the rates, are determined at the initial conditions, that is, the rate of the reaction is measured as the slope of the tangent of the rate curve at $t = 0$.

The procedure can also be performed numerically, where we substitute values of reaction rate and concentration into the general rate law and solve simultaneous equations for the two unknowns *n* and *m*.

This method (either graphical or arithmetic) is called the **initial rate method**.

7.10 Determination of 1st Order Rate Law

An example for a simple **first-order reaction** is the decomposition of hydrogen peroxide to water and oxygen:

$$2H_2O_2 \rightarrow 2H_2O + O_2(g)$$

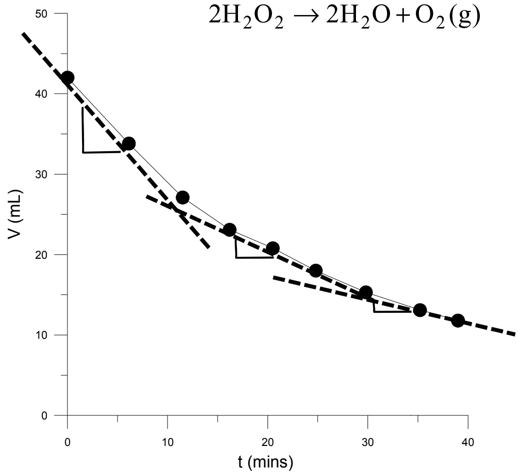

This is ordinarily a very slow reaction, but may be sped up by the addition of a catalyst. The concentration of H_2O_2 at any point is found by inhibiting the reaction by addition of dilute sulphuric acid, and the titration with $KMnO_4$. The volume of $KMnO_4$ used in the titration is an indication of the concentration of H_2O_2 left in the solution.

The slope of the tangent to the curve decreases as *t* increases, thus indicating that the rate of the reaction decreases with time. Since there is only one reactant, the decrease in reaction rate is therefore due to the decrease in concentration of H_2O_2. In general, the rate law would be written:

$$\frac{d[H_2O_2]}{dt} = k[H_2O_2]^n$$

For a first-order reaction, $n = 1$ and thus:

$$-\frac{d[H_2O_2]}{dt} = k[H_2O_2]$$

$$\frac{1}{[H_2O_2]}d[H_2O_2] = -kdt$$

$$\int_{[H_2O_2]_o}^{[H_2O_2]} \frac{1}{[H_2O_2]}d[H_2O_2] = -\int_0^t kdt$$

$$\ln[H_2O_2] - \ln[H_2O_2]_o = -kt$$

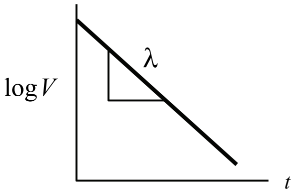

$[H_2O_2]$ is proportional to the titration volume in this experiment, and so a plot of log *V* versus *t* is a straight line if the reaction is first order, and the value of *k* can be found from the slope of a plot of log *V* versus *t*.

$$\ln[H_2O_2] = -kt + \ln[H_2O_2]_o$$

$$\log[H_2O_2] = -\frac{k}{2.303}t + \log[H_2O_2]_o \quad \text{(change of base to 10)}$$

7.11 Half-Life Method

The **half-life** $t_{1/2}$ of a reaction is the time for half of a reactant to be consumed, or half of a product to be formed. Consider previous case of a first-order reaction with one reactant:

$$2H_2O_2 \rightarrow 2H_2O + O_2(g)$$

Proceeding as before, we obtain

$$-\frac{d[H_2O_2]}{dt} = k[H_2O_2]$$
$$\ln[H_2O_2] = -kt + \ln[H_2O_2]_o$$
$$[H_2O_2]_t = [H_2O_2]_o e^{-kt}$$

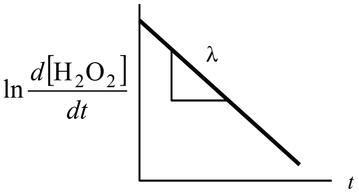

The half-life $t_{1/2}$ is the time for one half of the reactant to be consumed:

$$\frac{[H_2O_2]_t}{[H_2O_2]_o} = e^{-kt_{1/2}} = \frac{1}{2}$$
$$\ln\frac{1}{2} = -kt_{1/2}$$
$$t_{1/2} = \frac{0.69}{k}$$

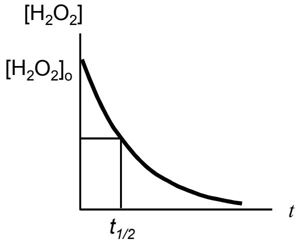

Either the rate constant *k* or the half-life (time measured for half the reactant to be consumed) can be determined from experimental data to give the other quantity. In a first-order reaction, the half-life is a constant and depends only on *k*.

7.12 Temperature

The rate of a chemical reaction depends on the frequency of collisions between atoms, ions or molecules, and also the fraction of which these particles have an energy greater than the activation energy. Increasing the concentration of reactants generally increases the frequency of collisions. Increasing the temperature also increases the collision frequency because the particles are moving faster and make more collisions per unit time. However, the most profound effect of an increase in temperature is to increase the fraction of particles with an energy above the activation energy.

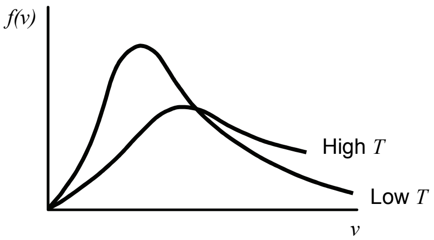

In gases, the distribution of velocities of molecules can be given by the Maxwell velocity distribution:

$$f(v) = 4\pi\left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} v^2 e^{-\frac{mv^2}{2kT}}.$$

At low temperature, a greater proportion of the molecules have a more narrow range of relatively low velocities compared to high temperature, where the range of velocities is more spread out, but shifted to higher values. Increasing the temperature therefore

- Shifts the molecular distribution of energies to a higher value
- Increases the proportion of molecules having higher energies

thus increasing the frequency and number of productive collisions per unit time and hence increasing the reaction rate.

The number of particles N with a total energy E was computed by Boltzmann using the **Maxwell velocity distribution** function:

$$N = Ce^{\frac{-E}{kT}}.$$

The rate constant for a specific reaction increases as the temperature increases, roughly in accordance with the **Arrhenius equation**:

$$k = Ae^{\frac{-E}{RT}},$$

where it can be seen that k is an exponential function of $1/T$. At room temperature, an increase of about 10°C results in an increase in reaction rate in the order of 2 to 3.

If E represents the activation energy, then this may be determined by experiment by plotting the logarithm of k versus $1/T$ over a range of temperatures.

7.13 Catalysts

The decomposition of hydrogen peroxide is a spontaneous, but slow reaction:

$$2H_2O_2 \rightarrow 2H_2O + O_2(g)$$

If powdered MnO_2 is added, the reaction proceeds much faster. Further, the MnO_2 molecules do not get used up during the reaction nor do they appear in the net equation. Substances which increase the rate of a reaction without appearing in the net reaction are called **catalysts** and the reaction is said to be **catalysed**.

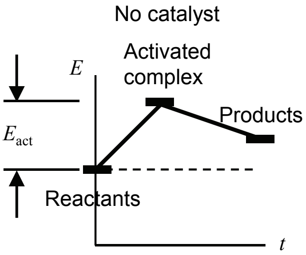

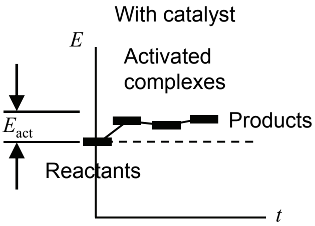

The role of a catalyst is to alter the reaction mechanism by insertion of additional steps so that the activation energy is reduced, allowing the reaction rate to increase.

While the catalyst may be consumed in one reaction step, and regenerated in another, it does not appear in the net reaction.

Catalysts may be solids, liquids or gases. When the reaction being catalysed is all of the same phase (e.g. all liquid), it is called **homogenous catalysis**. When the reaction occurs at the surface of a catalyst, such as over the platinum bed of a **catalytic converter** in an automobile exhaust system, it is called **heterogeneous catalysis**.

- A catalyst has no effect on the stoichiometry of the overall reaction.
- A catalyst does not affect the equilibrium position of the reaction. That is, the reaction proceeds to the same point of chemical equilibrium as it would without the catalyst. It just gets to equilibrium faster.
- A catalyst affects the reaction mechanism so as to lower the activation energy, thus increasing the number of molecules that can make a productive collision.

A negative catalyst decreases the rate of reaction and is called an **inhibitor**. The means by which this does so may not necessarily be to increase the activation energy, but is sometimes due to the inhibitor preventing another chemical present from acting as a catalyst. Auto catalysis, or **self-catalysis**, occurs when one of the products of a reaction acts as a catalyst for the reaction.

Enzymes act as catalysts in biological processes. Indeed, nearly all biological chemical reactions occur due to the action of catalysts.

8. Chemical Equilibrium

Summary

Law of chemical equilibrium:

$$K_c = \frac{[\mathrm{C}]^c[\mathrm{D}]^d}{[\mathrm{A}]^a[\mathrm{B}]^b}$$

Le Chatelier's principle:

> When a chemical system is at equilibrium, and that equilibrium is disturbed in some way, then the system will change in a way so as to counteract the disturbance until equilibrium is re-established.

Gaseous equilibrium:

$$K_p = \frac{K_c}{RT} = \frac{P_c}{P_a P_b}$$

Solid/solubility equilibrium:

$$K_{sp} = \left[\mathrm{A}^+(\mathrm{aq})\right]\left[\mathrm{B}^-(\mathrm{aq})\right]$$

Common ion effect:

> The presence of a common ion has reduced the solubility of a compound by several orders of magnitude.

8.1 Chemical Equilibrium

If a chemical reaction proceeds from reactants to products spontaneously, we may at first expect the process to continue until the reactants are all converted to product. This is not observed. Usually, reactions proceed until there is **chemical equilibrium**.

$$\text{Reactants} \leftrightarrow \text{Products}$$

At equilibrium, the rate of forward reaction = rate of reverse reaction.

As the reaction proceeds, the reactants are consumed and the products are created. The concentration of the reactants decreases (unless they are continuously replenished) and the concentration of the products increases. The rate of a chemical reaction depends on the concentration of the reactants. Therefore if the concentration of reactants is steadily decreasing, then the rate of conversion of reactants to products (in moles/litre/second) is also decreasing. At the same time, the rate of the conversion of products back to reactants is increasing because the concentration of the products is increasing. At chemical equilibrium, the rates of both forward and reverse reactions are equal.

$$A + B \leftrightarrow C + D$$

At equilibrium, we cannot tell from this equation how many moles of A and B exist, and how many moles of the product C or D exist.

The concentrations of A and B start off at some particular values, then decrease and level off with time. The concentrations of C and D start from zero and increase and level off with time. At equilibrium, the concentrations of A, B, C and D are all constant.

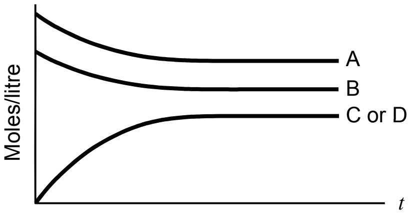

Although the balanced equation does not indicate the concentrations of reactants or products at any one time, they may be measured experimentally.

At equilibrium:

- The rate of formation of product C or D in moles per litre per second is equal to the rate of conversion of C and D back to molecules A and B.
- The concentrations of A, B and C, D are constant.

Chemical equilibrium is a **dynamic equilibrium**. Chemical activity still occurs at the equilibrium condition, but externally, we see no change in concentration of reactants or products.

It is important to know that **catalysts** do not alter the equilibrium state; they only serve to increase the rate of both forward and reverse reactions.

8.2 Law of Chemical Equilibrium

Consider the following single-step forward reaction:

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} + \mathrm{D}$$

The rate law for this reaction is written:

$$\frac{d[\mathrm{C}+\mathrm{D}]}{dt} = k[\mathrm{A}][\mathrm{B}]$$

Now, consider the reverse reaction and its **rate law**:

$$\mathrm{C} + \mathrm{D} \rightarrow \mathrm{A} + \mathrm{B}$$

$$\frac{d[\mathrm{A}+\mathrm{B}]}{dt} = k'[\mathrm{C}][\mathrm{D}]$$

At equilibrium, the rate of forward reaction must equal the rate of the reverse reaction, and so:

$$k[\mathrm{A}][\mathrm{B}] = k'[\mathrm{C}][\mathrm{D}]$$

$$\frac{[\mathrm{C}][\mathrm{D}]}{[\mathrm{A}][\mathrm{B}]} = \frac{k}{k'} = K_c$$

That is, the ratio of the concentrations of the products over the reactants is a constant *at equilibrium*. A similar treatment can be written for more complex reaction steps.

More generally, it is found from experiment that for the general chemical equation $a\mathrm{A} + b\mathrm{B} \rightarrow c\mathrm{C} + d\mathrm{D}$ the **equilibrium constant** K_c is given by

$$K_c = \frac{[\mathrm{C}]^c[\mathrm{D}]^d}{[\mathrm{A}]^a[\mathrm{B}]^b}$$ **law of chemical equilibrium**

The magnitude of K_c indicates the extent of the reaction.

$K_c \leq 10^{-5}$	Forward reaction is negligible
$10^{-5} \leq K_c \leq 10^{5}$	Appreciable concentration of products and reactants
$K_c \geq 10^{5}$	Reverse reaction is negligible and almost complete conversion of reactants to products

The value of the equilibrium constant depends on the temperature and the chemical species involved. However, it does not depend on the starting concentrations of either reactants or products. Whatever concentrations of A, B, C and D there are, when equilibrium is reached, the ratio of the concentrations expressed in this way will have the same value at that temperature. The value of *K* also depends on how the balanced equation is written (i.e. the choice of coefficients). The equilibrium constant must be associated with a particular balanced equation.

Note that the equilibrium constant does not provide any information on how long it takes for equilibrium to be established. It only provides information on the extent to which equilibrium favours the products over the reactants.

8.3 Equilibrium Constant

For any chemical reaction, the ratio of concentrations raised to the power of the coefficients can be written as:

$$\frac{[\mathrm{C}]^c[\mathrm{D}]^d}{[\mathrm{A}]^a[\mathrm{B}]^b}$$

This is more formally called the **mass-action expression**. It is only at chemical equilibrium that this expression becomes equal to a constant value K_c (the subscript c indicates concentrations in moles/litre).

K_c depends on the temperature. The units of K_c depend on the powers to which the concentrations are raised.

$$\mathrm{N_2O_4} \rightarrow 2\mathrm{NO_2(g)}$$

$$K_c = \frac{[\mathrm{NO_2}]^2}{[\mathrm{N_2O_4}]} = \mathrm{mol/L}$$

$$2\mathrm{CO}_{(g)} + \mathrm{O_2} \rightarrow 2\mathrm{CO_2}$$

$$K_c = \frac{[\mathrm{CO_2}]^2}{[\mathrm{CO}]^2[\mathrm{O_2}]} = \mathrm{L/mol}$$

A particular value of K_c depends on the form of the balanced equation.

compare

$$\frac{1}{2}\mathrm{N_2O_4} \rightarrow \mathrm{NO_2(g)}$$

$$K_c' = \frac{[\mathrm{NO_2}]}{[\mathrm{N_2O_4}]^{1/2}} = \mathrm{mol}^{1/2}\ \mathrm{L}^{1/2}$$

$$= (K_c)^{1/2}$$

Homogenous equilibrium occurs when all reactants and products are in the same physical state. **Heterogeneous equilibrium** occurs when reactants and products are in different physical states.

The concentrations of pure **solids** and **liquids** are constant. This is because if more *pure* solid or liquid is added, then the volume of that solid or liquid is also increased in the same proportion and so the number of moles per litre remains the same as before. The concentrations of these substances are removed from the equilibrium expression and are effectively incorporated into the value of K_c.

Gases cannot be removed from the equilibrium expression because their concentration changes during the reaction.

Writing a chemical equation in reverse inverts the value of K_c.

8.4 Le Chatelier's Principle

When a chemical system is at equilibrium, and that equilibrium is disturbed in some way, then the system will change in a way so as to counteract the disturbance until equilibrium is re-established. This is known as **Le Chatelier's principle**.

One way of disturbing the equilibrium condition is to add more of a reactant or a product. When more of a reactant is added, the forward reaction increases in rate (since reaction rate depends on the concentration) and more product is formed until the equilibrium condition, and the value of K_c, is restored. When more of product is added, then the concentration of the products increases, thus increasing the rate of the reverse reaction until the equilibrium condition is re-established. Similarly, if either a reactant or product is removed, then the forward or reverse reaction proceeds at a faster rate so as to restore the concentration of the removed species until the equilibrium condition is re-established.

Consider the effect of an increase in concentration of one reactant in the equilibrium expression:

$$N_2O_4 \rightarrow 2NO_2(g)$$

$$K_c = \frac{[NO_2]^2}{[N_2O_4]} = 0.87 \text{ at } 55^\circ C$$

Increasing the amount of N_2O_4 makes the denominator of the expression larger, and makes the computed value of K_c smaller. The concentration of NO_2 must increase until the original value of K_c is re-established. Note that at the new equilibrium, the concentrations of both products and reactants will have increased but their ratio, as expressed by the equilibrium expression, remains the same. Looked at another way, introducing more reactant tends to increase the rate of the forward reaction until the equilibrium constant is re-established back to its former level.

Another way of disturbing the equilibrium condition is to alter the temperature. The value of K_c at equilibrium depends very much on the temperature. In an exothermic reaction, increasing the temperature tends to favour the reverse reaction. A new equilibrium is reached with a new value of K_c.

8.5 Summary of Le Chatelier's Principle

Summary of Le Chatelier's Principle:

Change	**Effect on K_c**	**Shift in reaction**
Increase in concentration of one reactant	Constant	Forward reaction
Decrease in concentration of one reactant	Constant	Reverse reaction
Increase in concentration of one of the products	Constant	Reverse reaction
Decrease in concentration of one of the products	Constant	Forward reaction
Pressure increase with volume decrease	Constant	Decrease in number of gaseous molecules
Pressure decrease with volume increase	Constant	Increase in number of gaseous molecules
Temperature increase (exothermic reaction)	Decrease	Reverse reaction
Temperature increase (endothermic reaction)	Increase	Forward reaction
Temperature decrease (exothermic)	Increase	Forward reaction
Temperature decrease (endothermic)	Decrease	Reverse reaction
Catalyst added	Constant	No shift, but rate of both forward and reverse reactions increased

8.6 Equilibrium in Gaseous Systems

For reactions involving gases, the law of chemical equilibrium can be written in terms of **partial pressures**. The amount of gas in a mixture can be measured in terms of its partial pressure. The total of all the partial pressures is the total pressure in the system. Consider the reaction

$$\mathrm{A(g)} + \mathrm{B(g)} \to \mathrm{C(g)}$$

General gas equation: $PV = nRT$

Therefore, for A(g) $P_aV = nRT$ where P_a is the partial pressure of gas A in volume V

$$P_a = \frac{n}{V}RT$$

$$= [\mathrm{A}]RT$$

and similarly for B(g) and C(g).

Therefore,

$$K_c = \frac{[\mathrm{C}]}{[\mathrm{A}][\mathrm{B}]}$$

$$= RT\frac{P_c}{P_aP_b}$$

$$\frac{K_c}{RT} = \frac{P_c}{P_aP_b}$$

a constant.

$$= K_p$$

That is, a **gaseous equilibrium constant** K_p is the ratio of the partial pressures of the products over the partial pressures of the reactants.

Example: $N_2(g) + 3H_2(g) \to 2NH_3(g)$

$$K_p = \frac{P_{NH_3}^2}{(P_{N_2})(P_{H_2})^3}$$

$$K_c = \frac{[NH_3]^2}{[N_2][H_2]^3}$$

$$= \frac{(P_{NH_3}/RT)^2}{(P_{N_2}/RT)(H_2/RT)^3}$$

$$= \frac{P_{NH_3}^2}{(P_{N_2})(P_{H_2})^3}(RT)^2$$

$$= K_p(RT)^2$$

8.7 Solubility of Solids

Many chemical species dissolve in water. They can do so by a simple dissolution where the molecules just become separated from each other and are surrounded by water molecules, by dissociation into ions, and by chemical reaction with the water molecules.

Silver chloride, AgCl, has a low solubility in water. Those molecules that do dissolve, do so by forming ions. The **equilibrium constant** is therefore written as:

$$\mathrm{AgCl(s)} \leftrightarrow \mathrm{Ag^+(aq)} + \mathrm{Cl^-(aq)}$$

$$K_c = \frac{\left[\mathrm{Ag^+(aq)}\right]\left[\mathrm{Cl^-(aq)}\right]}{\left[\mathrm{AgCl(s)}\right]}$$

Note that the chemical equation is written with the ions on the product side.

The concentration of a solid is a constant, and so is incorporated into the equilibrium constant to form a new constant, the **solubility product constant**, K_{sp}:

$$K_{sp} = \left[\mathrm{Ag^+(aq)}\right]\left[\mathrm{Cl^-(aq)}\right] = 1.7\times10^{-10} \text{ at } 25^\circ\mathrm{C}$$

The **solubility product constant**, K_{sp}, applies to saturated aqueous solutions. The product of the ion concentrations is called the ion product. At equilibrium, the ion product equals K_{sp}. A low value of K_{sp} indicates that the concentration of ions in solution is low at equilibrium – in other words, the solubility of the compound is low.

The **solubility** of a compound in water is the number of moles of that compound that can be dissolved in 1 litre of water (at a specific temperature). The solubility of a compound can be obtained from the solubility product constant.

$$\begin{aligned} K_{sp} &= \left[\mathrm{Ag^+(aq)}\right]\left[\mathrm{Cl^-(aq)}\right] \\ &= 1.7\times10^{-10} \text{ at } 25^\circ\mathrm{C} \\ s &= \left[\mathrm{Ag^+(aq)}\right] = \left[\mathrm{Cl^-(aq)}\right] \\ 1.7\times10^{-10} &= s^2 \\ s &= 1.3\times10^{-5}\,\mathrm{mol/L} \end{aligned}$$

From the balanced equation, 1 mole of AgCl will produce one mole of Ag^+ and 1 mole of Cl^- ions. Therefore in a 1 litre solution, there will be *s* moles of Ag^+ and *s* moles of Cl^- ions.

At equilibrium, the rate of **dissolution** is equal to the rate of **precipitation**.

In general, the solubility product constant is

$$\mathrm{M}_n\mathrm{X}_m \leftrightarrow n\mathrm{M}^{n+}\mathrm{(aq)} + m\mathrm{X}^{n-}\mathrm{(aq)}$$

$$K_{sp} = \left[\mathrm{M}^{n+}\mathrm{(aq)}\right]^n\left[\mathrm{X}^{n-}\mathrm{(aq)}\right]^m \quad \text{at solubility equilibrium}$$

Ionic solids (salts) and some molecular solids may dissociate into ions in aqueous solution.

8.8 Factors Affecting Equilibrium

A system tends towards a state of minimum **energy** and maximum **entropy**. The two tendencies can oppose or reinforce one another.

- If the forward reaction is exothermic $\Delta H < 0$ and the entropy change $\Delta S > 0$, then the forward reaction is favoured.
- If the forward reaction is endothermic $\Delta H > 0$ and the entropy change $\Delta S < 0$, then the reverse reaction is favoured.
- If the enthalpy change $\Delta H > 0$ and the entropy change $\Delta S > 0$, then the forward reaction is favoured at high temperature. When $\Delta H < 0$ and $\Delta S < 0$, the forward reaction is favoured at low temperature.

For a **solid** dissolving in a **liquid**, in saturated solution, the position of minimum energy can be on either side of the equation. The position of maximum entropy is on the dissolved side:

Solid ↔ Saturated solution
entropy gain →
← energy minimum →

For the dissolving of a **gas** in a **liquid**, entropy decreases as the gas dissolves and so the tendency is to oppose the dissolving process, however, when a gas dissolves, the gas molecules, being attracted to the liquid molecules, enter a state of lower potential energy and so this favours the dissolving process.

Gas ↔ Saturated solution
← entropy gain
energy minimum →

It is important to realise that the **equilibrium constant** is not the same thing as the **solubility product constant**. The solubility product constant is an experimentally derived value that applies to the solubility of a compound in **aqueous solution**. The equilibrium constant applies to any chemical reaction, not just those in aqueous (water) solution.

Many chemical reactions occur in dilute aqueous solutions. In these cases, for reactions that consume, or produce, water molecules, the concentration of the water is considered constant and is incorporated into the equilibrium constant. This is because the amount of water consumed or produced is usually a small fraction of the overall total amount of water present.

8.9 Common Ion Effect

The solubility of an ionic solid in a solution containing a common ion is greatly reduced. This is because the concentration of the common ion appears in the expression for the solubility product constant.

K_{sp} = 1.5×10^{-9} for $BaSO_4$ in water . The solubility of this salt in water can be calculated and compared with the solubility in 0.1M Na_2SO_4.

Solubility in water:

$$\begin{aligned} K_{sp} &= \left[\mathrm{Ba}^{2+}(\mathrm{aq})\right]\left[\mathrm{SO}_4^{2-}(\mathrm{aq})\right] \\ &= (s)(s) \\ s^2 &= 1.5\times10^{-9} \\ s &= 3.9\times10^{-5}\,\mathrm{mol/L} \end{aligned}$$

Solubility in 0.1M Na_2SO_4:

$$\begin{aligned} K_{sp} &= \left[\mathrm{Ba}^{2+}(\mathrm{aq})\right]\left[\mathrm{SO}_4^{2-}(\mathrm{aq})\right] \\ &= (s)(s+0.1) \\ &= 1.5\times10^{-9} \\ (s+0.1) &\approx 0.1 \quad \text{since } s \text{ will be} \ll 0.1 \\ 1.5\times10^{-9} &= s(0.1) \\ s &= 1.5\times10^{-8}\,\mathrm{mol/L} \end{aligned}$$

The presence of the **common ion** has reduced the solubility of this compound by several orders of magnitude.

8.10 Precipitation

The **solubility product constant** K_{sp} can be used to predict whether or not a solid precipitate will form as a result of a chemical reaction between solutions.

The product of the concentrations of the ions in a solution is called the **ion product**. When the ion product is greater than the solubility product constant, **precipitation** will occur until the concentration of the ion reduces to the extent that the ion product becomes equal to K_{sp}.

Consider the mixing of 1 litre of 0.5M $Ca(NO_3)_2$ and 1 litre of 0.2M Na_2SO_4. We begin by writing the ionic equation of interest:

$$CaSO_4 \leftrightarrow Ca^{2+} + SO_4^{2-}$$

Note that the equation has been written with the ions on the product side of the equation so that the ion product can be related to K_{sp} directly, and not the inverse.

The final volume of the mixed solution is 2 litres, and therefore the concentrations of the original solutions are:

$$\left[Ca^{2+}\right] = 0.25$$
$$\left[SO_4^{2-}\right] = 0.1$$
$$\begin{aligned} IP &= \left[Ca^{2+}(aq)\right]\left[SO_4^{2-}(aq)\right] \\ &= (0.25)(0.1) \\ &= 0.025 \end{aligned}$$

ion product

K_{sp} for $CaSO_4$ = 2.6×10^{-4}. Since the ion product is greater than K_{sp}, then $CaSO_4$ will form as a precipitate until the concentration of the Ca^{2+} and SO_4^{2-} ions fall until the ion product becomes equal to K_{sp}. That is, precipitation is a method of forcing out ions from the solution until the ion product becomes equal to K_{sp}, at which time the solid and the solution are in equilibrium. Precipitation has important economic consequences whereby the extraction of the desired product must be achieved as economically as possible. Precipitates are sometimes washed with a solution containing a **common ion** to further reduce their solubility.

8.11 Complex Ions

Some molecules or anions have a lone pair of electrons. This enables them to combine, or coordinate, with some transition metal ions to form a **complex ion**. For example, the molecule ammonia, NH_3, has a lone pair of electrons:

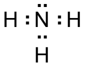

$$\begin{array}{c} \cdot\cdot \\ H : \mathrm{N} : H \\ \cdot\cdot \\ H \end{array}$$

NH_3 can coordinate to Ag^+, Co^{3+}, Zn^{2+}, Ni^{2+} to form complex ions:

$$\mathrm{Ag(NH_3)_2^+, Co(NH_3)_4^{2+}, Zn(NH_3)_4^{2+}, Ni(NH_3)_6^{2+}}$$

The OH- ion also has a lone pair of electrons and can coordinate with Pb^{2+}, Zn^{2+}, Sn^{2+}, Al^{3+} :

$$\mathrm{Pb(OH)_4^{2-}, Zn(OH)_4^{2-}, Sn(OH)_4^{2-}, Al(OH)_4^-}$$

The **ligand** (the species with the lone pair) arranges itself around the metal ion and the electron pairs fill up vacant energy levels to create a noble gas configuration for the metal ion. The electron orbitals that are associated with isolated atoms are rearranged to give **hybrid orbitals** to accommodate the electron pairs. The overall charge on the complex ion remains unchanged.

The equilibrium condition between the reactants and products for complex ion formation is called the **stability constant** K_{stab}.

$$\mathrm{Ni^{2+} + 4CN^- \rightarrow Ni(CN)_4^{2-}}$$

Tetracyanonickelate

$$K_{stab} = \frac{\left[\mathrm{Ni(CN)_4^{2-}}\right]}{\left[\mathrm{Ni^{2+}}\right]\left[\mathrm{CN^-}\right]^4}$$

$$\mathrm{Pt^{2+} + 6NH_3 \rightarrow Pt(NH_3)_6^{2+}}$$

Hexaamineplatinum (II)

$$K_{stab} = \frac{\left[\mathrm{Pt(NH_3)_6^{2+}}\right]}{\left[\mathrm{Pt^{2+}}\right]\ \left[\mathrm{NH_3}\right]^6}$$

$$\mathrm{Zn^{2+} + 4OH^- \rightarrow Zn(OH)_4^{2-}}$$

Tetrahydroxozincate

$$K_{stab} = \frac{\left[\mathrm{Zn(OH)_4^{2-}}\right]}{\left[\mathrm{Zn^{2+}}\right]\left[\mathrm{OH^-}\right]^4}$$

9. Ionic Equilibrium

Summary

- Strong electrolytes dissociate completely into ions.
- Weak electrolytes only partially dissociate into ions.
- Non-electrolytes do not dissociate into ions when placed in solution.

Ion product of water:

$$K = \frac{[H^+][OH^-]}{[H_2O]}$$

$$K_w = [H^+][OH^-]$$

[H+] mol/L	**pH**	**pOH**
1	0	14
1×10^{-7}	7	7
1×10^{-14}	14	0

$$pH = -\log_{10}[H^+]$$

$$pOH = -\log_{10}[OH^-]$$

Acids – proton (H^+) donors

Bases – proton (H^+) acceptors

Acid equilibrium constant:

$$K_A = \frac{[H_3O^+][A^-]}{[HA]}$$

Base equilibrium constant:

$$K_B = \frac{[BH^+][OH^-]}{[B]}$$

Neutralisation: An acid on a base gives salt and water

Titration: The progressive addition of an acid to a base, or a base to an acid.

Hydrolysis: The chemical reaction of ions with water.

9.1 Electrolytes

When a substance dissolves, it may do so by simply separating from its neighbour molecules and become surrounded by water molecules, or the constituent molecules may break apart, or dissociate, and form negatively charged **anions** and positively charged **cations**. Such solutions conduct electricity by virtue of the movement of the ions under the influence of an applied electric field. Substances which dissolve to produce conducting solutions are called **electrolytes**. Substances which do not are called **non-electrolytes**.

- **Strong electrolytes** dissociate completely into ions.
- **Weak electrolytes** only partially dissociate into ions.
- **Non-electrolytes** do not dissociate into ions when placed in solution.

When an ionic substance dissociates into ions in an aqueous solution, the ions themselves are attracted to the polar water molecules and become **hydrated**. That is, a negative anion is surrounded by the positive (or H) end of water molecules and this hydrated ion can move around in solution under the influence of an electric field because as a whole, it is still negatively charged. Similarly, a positive cation is surrounded by the oxygen end of nearby water molecules and becomes hydrated, but the hydrated ion remains positively charged.

Both ionic and molecular compounds may be strong electrolytes. In the case of HCl, a molecular compound, we have

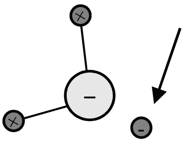

H^+ proton with positive charge showing one water molecule being attracted to it. Other water molecules may also be attracted to this proton.

$$\mathrm{HCl} + \mathrm{H_2O} \rightarrow \mathrm{H_3O^+} + \mathrm{Cl^-}$$

The **hydronium ion**, H_3O^+, is a hydrated H^+ ion. For convenience, we omit the water of hydration and just write

$$\mathrm{HCl} \rightarrow \mathrm{H^+} + \mathrm{Cl^-}$$

Examples of ionic compounds which are strong electrolytes are NaCl, $MgCl_2$, KBr, $CuSO_4$, $LiNO_3$. Examples of molecular covalent compounds which are strong electrolytes are HCl, HBr, HNO_3. Examples of weak electrolytes are $HC_2H_3O_2$, HNO_2, NH_3, HCN.

The degree to which dissociation occurs depends on energy changes (in particular, the energy of hydration) and entropy considerations as well as the nature of the solvent, the concentration of the solute, and the temperature. The dissociation of HCl is an example of **ionic equilibrium**.

9.2 Ionisation of Water

Experiments indicate that pure **water** is a very poor conductor of electricity, but it does conduct, indicating the presence of ions. Water is a weak **electrolyte**. Water dissociates into ions according to

$$H_2O \leftrightarrow H^+ + OH^-$$

The H^+ ions attach themselves to other water molecules to form the **hydronium** ion H_3O^+, but here we will simply write the H^+ ions as if they exist on their own.

The degree of dissociation is very small, and so the concentration of the H^+ and OH^- ions is also very small.

The concentration of water, expressed as moles/litre, can be readily established from its molecular weight and its density:

$$\begin{aligned} \text{m.w.}_{H_2O} &= 16+2 \\ &= 18 \end{aligned}$$

One litre of water has a mass of 1000 g, therefore the number of moles *n* in a litre of water is:

$$\begin{aligned} n &= \frac{1000}{18} \\ &= 55.6\ \text{moles} \end{aligned}$$

That is, water has a concentration of 55.6M.

Since the concentration of the H^+ and OH^- ions in a litre of water is exceedingly small, the concentration of H_2O is virtually a constant, and so the **ion product** K_w is formed:

$$\begin{aligned} K &= \frac{[H^+][OH^-]}{[H_2O]} \\ K_w &= [H^+][OH^-] \end{aligned}$$

The concentration of H_2O (55.6M) is taken to be a constant and incorporated into K_w.

Experiments show that the **ion product of pure water** at 25°C is 1.0×10^{-14} mole^2/litre^2. The concentrations of H^+ and OH^- ions are 1.0×10^{-7} moles/litre respectively.

A solution containing equal concentrations of H^+ and OH^- ions at these concentrations are said to be **neutral**.

$$\begin{aligned} [H^+] &= 1\times10^{-7}\ \text{mol/litre} \\ [OH^-] &= 1\times10^{-7}\ \text{mol/litre} \end{aligned}$$

9.3 H^+ and OH^- Concentrations

because HCl dissociates completely to H^+ and Cl^- ions.

Consider a 0.1M solution of HCl:

$$\mathrm{HCl} \rightarrow \mathrm{H}^+ + \mathrm{Cl}^-$$

The concentration of H^+ ions will be 0.1M and Cl^- ions 0.1M. We can ignore the small amount of H^+ ions from the dissociation of the water solvent. But, the equilibrium between the H^+ and the OH^- ions in the solution must still be established such that

$$\begin{aligned} 1\times10^{-14} &= \left[\mathrm{H}^+\right]\left[\mathrm{OH}^-\right] \\ &= 0.1\left[\mathrm{OH}^-\right] \\ \left[\mathrm{OH}^-\right] &= 1\times10^{-13}\,\mathrm{mol/litre} \end{aligned}$$

That is, compared to neutral water, the concentration of the OH^- has been reduced from 1×10^{-7} to 1×10^{-13} due to a **common ion effect**.

Consider now an aqueous solution of 0.1M NaOH. Using the same calculation as above, we find that

$$\mathrm{NaOH} \rightarrow \mathrm{Na}^+ + \mathrm{OH}^-$$
$$\left[\mathrm{H}^+\right] = 1\times10^{-13}\,\mathrm{mol/litre}$$

The added H^+ or OH^- ions suppress or inhibit the dissociation of the water molecules (in accordance with Le Chatelier).

Because the changes in concentration of these ions can cover such a large range of magnitudes, it is convenient to express them on a logarithmic scale – the "p" scale. The **pH** is the concentration of H^+ ions found from:

$$\mathrm{pH} = -\log_{10}\left[\mathrm{H}^+\right]$$

The **pOH** is the concentration of OH^- ions found from

$$\mathrm{pOH} = -\log_{10}\left[\mathrm{OH}^-\right]$$

The pH is a measure of the potential to combine with H^+ ions. A high pH means a high potential for H^+ combination.

Since the logarithm is the index to which the base must be raised to equal the number, the pH is the index of the H^+ concentration. For example, for a neutral solution,

$$\begin{aligned} \mathrm{pH} &= -\log^{10}\left[\mathrm{H}^+\right] = -\log_{10}\left(10^{-7}\right) \\ &= 7 \end{aligned}$$

It is easy to show that for a solution at equilibrium,

$$\mathrm{pH} + \mathrm{pOH} = 14$$

[H+] (mol/L)	pH	pOH
1	0	14
1×10^{-7}	7	7
1×10^{-14}	14	0

Say we have a 5M solution of HCl. What is the pH? $\mathrm{pH} = -\log_{10} 5 = -0.7$

A negative pH is possible, but at high concentrations, the assumptions we make about the derivation of these equilibrium formulae breaks down.

9.4 Acids and Bases

A wide variety of substances can be generally classified into being either **acids**, **bases** or **salts**. Each of these has several distinctive characteristics when dissolved in water.

Acids	**Bases**
Electrical conductor	Electrical conductor
Sharp sensation on skin	Slippery feel
Sour taste	Bitter taste
Change the colour of litmus from blue to red	Change colour of litmus from red to blue.
React with some metals to form H_2 gas	
Dissolve carbonates to form CO_2 gas	

It was noticed that substances such as HCl, HNO_3, $HC_2H_3O_2$, H_2SO_4 and so on had the characteristics of acids. Substances such as NaOH, KOH, $Ca(OH)_2$, $Mg(OH)_2$ had the characteristics of bases. It was therefore postulated by **Arrhenius** (1859-1927) that acids are substances which dissolve in water to form H^+ ions, and bases are substances that dissolve in water to form OH^- ions:

$$\begin{array}{cc} \text{Acid} & \text{Base} \\ \mathrm{HA} \rightarrow \mathrm{H}^+ + \mathrm{A}^- & \mathrm{BOH} \rightarrow \mathrm{B}^+ + \mathrm{OH}^- \end{array}$$

According to Arrhenius, all acids have the general formula HA and all bases have the general formula BOH.

A slightly more general definition of Arrhenius' scheme is that acids dissociate so as to increase the H^+ ions in water while bases dissociate and result in an increase in OH^- ions.

It was also known that when an acid was mixed with a base, the resulting solution lost its properties of either (except for being an electrical **conductor**) and formed a **salt** with the general form BA.

The degree of dissociation in water indicates the strength of the acid or the base:

$\mathrm{HCl} \rightarrow \mathrm{H}^+ + \mathrm{Cl}^-$	Strong acid (equilibrium favours products)
$\mathrm{HC_2H_3O_2} \rightarrow \mathrm{H}^+ + \mathrm{C_2H_3O_2^-}$	Weak acid (equilibrium favours reactants)
$\mathrm{NaOH} \rightarrow \mathrm{Na}^+ + \mathrm{OH}^-$	Strong base (equilibrium favours products)
$\mathrm{NH_3} + \mathrm{H_2O} \leftrightarrow \mathrm{NH_4^+} + \mathrm{OH}^-$	Weak base (equilibrium favours reactants)
$\mathrm{HCl} + \mathrm{NaOH} \rightarrow \mathrm{NaCl} + \mathrm{H_2O}$	acid + base = salt + water

The Arrhenius definition of acids and bases remains the simplest and most popular despite it having some limitations in only dealing with H^+ and OH^- ions as the cause of acidic or basic properties of substances.

9.5 Bronsted–Lowry

Reactions can take place in other solvents besides water. H^+ and OH^- ions may not be involved at all, yet the reactions still have the characteristics of acids and bases. The **Bronsted–Lowr**y definitions of acids and bases are

Acids – proton (H^+) donors
Bases – proton (H^+) acceptors

The Bronsted–Lowry classification is consistent with the Arrhenius definition of acids and bases but includes certain substances that act as bases which are not covered by the Arrhenius definition.

Consider the dissociation of HCl in water, and this time, we write the equation in full with the **hydronium ion**:

$$\mathrm{HCl} + \mathrm{H_2O} \rightarrow \mathrm{H_3O^+} + \mathrm{Cl^-}$$

Now, let's focus our attention on the water molecule. The water molecule accepts a H^+ ion, a proton, to become a hydronium ion. That is, the water molecule, according to Bronsted and Lowry, acts as a base. At equilibrium, the reverse reaction also occurs:

$$\mathrm{HCl} + \mathrm{H_2O} \leftarrow \mathrm{H_3O^+} + \mathrm{Cl^-}$$

Here, the hydronium ion donates a proton to the proceedings and the result is a water molecule. For all intents and purposes, the hydronium ion acts as an acid. In the Bronsted–Lowry scheme, the water molecule is the **conjugate base** of the hydronium ion. Or, the hydronium ion is the **conjugate acid** of the water molecule.

In the forward reaction, HCl donates a proton and becomes a Cl^- ion. HCl acts as an acid. In the reverse reaction, the Cl^- ion becomes HCl. That is, the Cl^- ion acts like a base, accepting a proton and becoming HCl. HCl is the **conjugate acid** of Cl^- and Cl^- is the **conjugate base** of HCl.

Every Bronsted–Lowry acid is accompanied by a Bronsted–Lowry base. The Bronsted–Lowry scheme has this extra dimension of creation of a **conjugate pair** compared to the Arrhenius theory:

$$\underset{\text{acid}}{\mathrm{HA}} + \underset{\text{base}}{\mathrm{H_2O}} \rightarrow \underset{\text{acid}}{\mathrm{H_3O^+}} + \underset{\text{base}}{\mathrm{A^-}}$$

In the forward reaction, HA is the acid and H_2O is the base. For the reverse reaction, H_3O^+ is the acid and A^- is the base.

The degree of dissociation indicates the strength of the acid or the base. A strong Bronsted–Lowry acid implies a weak conjugate base. One of the greatest strengths of this system is that acid-base reactions that do not involve the OH^- ion can be accommodated.

9.6 Strength of Acids

The strength of an acid is described by its tendency to dissociate into ions and produce H^+. The **pH** scale does not necessarily indicate whether a solution contains a strong acid (or base); it only indicates the concentration of H^+ (and OH^-) ions which depends on both the intrinsic strength of the acid and the molar concentration used to make the solution.

HCl is considered an intrinsically strong acid because it almost completely dissociates into H^+ and Cl^- ions:

$$\mathrm{HCl} \rightarrow \mathrm{H}^+ + \mathrm{Cl}^-$$

The **equilibrium constant** K_{HCl} for this reaction is very large (≈10^7), favouring he products. Acetic acid is a weak acid:

$$\mathrm{HC_2H_3O_2} \leftrightarrow \mathrm{H}^+ + \mathrm{C_2H_3O_2}^-$$

The equilibrium constant for this reaction is very small (about 1.8×10^{-5}), favouring the reactants.

where we always know that H^+ is actually hydrated

The general expression for the **acid equilibrium constant** is given as

$$\mathrm{HA} + \mathrm{H_2O} \leftrightarrow \mathrm{H_3O}^+ + \mathrm{A}^-$$

$$K = \frac{[\mathrm{H_3O}^+][\mathrm{A}^-]}{[\mathrm{HA}][\mathrm{H_2O}]}$$

$$K_{\mathrm{A}} = \frac{[\mathrm{H_3O}^+][\mathrm{A}^-]}{[\mathrm{HA}]}$$

since the concentration of H_2O is considered constant.

Acid	K_A	**Conjugate base**
$HClO_4$	≈10^{10}	ClO_4^-
H_2SO_4	≈10^9	HSO_4^-
HCl	≈10^7	Cl^-
HSO_4	1.3×10^{-2}	SO_4^-
H_3PO_4	8×10^{-3}	$H_2PO_4^-$
$HC_2H_3O_2$	1.74×10^{-5}	$CH_3CO_2^-$
NH_4^+	6.3×10^{-10}	NH_3
NH_3	≈10^{-30}	NH_2^-

K_A provides a quantitative measure of the strength of an acid. That is, its tendency for dissociation to provide H^+ ions. pH is an indication of the concentration of H^+ ions in a particular solution. A highly concentrated solution of a weak acid may have a similar pH to a dilute solution of a strong acid.

The strength of a base can be expressed in a similar manner by the **base equilibrium constant** K_B.

$$\mathrm{B} + \mathrm{H_2O} \leftrightarrow \mathrm{BH}^+ + \mathrm{OH}^-$$

$$K_{\mathrm{B}} = \frac{[\mathrm{BH}^+][\mathrm{OH}^-]}{[\mathrm{B}]}$$

These equilibrium constants can be expressed on a logarithmic scale:

$$\mathrm{p}K_{\mathrm{A}} = -\log_{10}[K_A]$$

$$\mathrm{p}K_{\mathrm{B}} = -\log_{10}[K_B]$$

For an acid and its conjugate base,

$$\mathrm{p}K_{\mathrm{A}} + \mathrm{p}K_{\mathrm{B}} = 14$$

9.7 Acid–Base Reactions

A few acids and bases are strong, but most are weak. Consider the dissociation of acetic acid $HC_2H_3O_2$, a weak acid with $K_A = 1.74 \times 10^{-5}$.

$$HC_2H_3O_2 + H_2O \leftrightarrow H_3O^+ + C_2H_3O_2^-$$

The concentration of H_3O^+ is found from:

$$K = \frac{\left[H_3O^+\right]\left[C_2H_3O_2^-\right]}{\left[HC_2H_3O_2\right]\left[H_2O\right]}$$

$$\left[H_3O^+\right] = 1.74\times10^{-5}\,\frac{\left[HC_2H_3O_2\right]}{\left[C_2H_3O_2^-\right]}$$

Strong acids:
HCl, HBr, HI, HNO_3,
H_2SO_4, $HClO_4$

If a solution is made with 0.1 moles of $HC_2H_3O_2$, then the concentration $[HC_2H_3O_2]$ is 0.1M. What then is the concentration of $[C_2H_3O_2^-]$? It must be small since the value of K_A implies a low incidence of dissociation. Now, let us add 0.1 moles of solid $NaC_2H_3O_2$. Since the salt, $NaC_2H_3O_2$ dissociates almost completely, the concentration of the acetate ion is now approximately 0.1M. Thus

$$\left[H_3O^+\right] = 1.74\times10^{-5}\,\frac{0.1}{0.1}$$

$$pH = 4.75$$

If to this solution we now add a small amount of a strong base, say 0.001 moles (a drop) of 2M NaOH, then the NaOH reacts with the acetic acid molecules: $HC_2H_3O_2 + NaOH \leftrightarrow H_2O + NaC_2H_3O_2$

The quantity $[HC_2H_3O_2]$ is reduced from 0.1M to 0.099M. The quantity $[C_2H_3O_2^-]$ increases to 0.11M (since the $NaC_2H_3O_2$ dissociates). The pH of this new solution is

$$pH = \log\left(1.74\times10^{-5}\,\frac{0.099}{0.11}\right) = 4.8$$

The $HC_2H_3O_2$ / $C_2H_3O_2^-$ solution is called a **buffer solution**, because of its ability to resist changes in pH. If a small amount of acid is added to the solution, this reacts with the acetate ion and the concentrations shift slightly the other way, and the pH remains almost unchanged.

If the concentration of the introduced acid or base becomes too large, then the **buffering capacity** of the solution is exhausted and the pH changes appreciably.

Note, without the common ion, addition of small amounts of acid or base can significantly alter the pH by several units.

9.8 Buffer Solutions

Buffer solutions have the important property of being able to maintain pH when disturbed by the addition of other acids or bases. Buffer solutions can be made in two ways:

1. Weak acid/salt of conjugate base (e.g. acetic acid and sodium acetate)

$$\mathrm{NaA_{(s)}} \rightarrow \mathrm{Na^+} + \mathrm{A^-}$$

$$\mathrm{HA} + \mathrm{H_2O} \leftrightarrow \mathrm{H_3O^+} + \mathrm{A^-}$$

$$\left[\mathrm{H_3O^+}\right] = K_A \frac{[\mathrm{HA}]}{\left[\mathrm{A^-}\right]}$$

$$\mathrm{pH} = -\log\left(K_A \frac{[\mathrm{HA}]}{\left[\mathrm{A^-}\right]}\right)$$

There are two equilibrium reactions going on here. One is the almost complete dissociation of the salt into its ions, and the other is the partial dissociation of the acid. The [A$^-$] is dominated by the molarity of the salt solution. The other reaction is the acid equilibrium. Because of the presence of the common ion A$^-$, the equilibrium for this reaction lies on the reactants side of the equation and [HA] is therefore practically equal to the molarity of the acid.

Since [HA] is high (equilibrium to the left), any added OH$^-$ from an introduced chemical will react with these ions:

$$\mathrm{OH^-} + \mathrm{HA} \rightarrow \mathrm{A^-} + \mathrm{H_2O}$$

added in buffer

$$\mathrm{H_3O^+} + \mathrm{A^-} \rightarrow \mathrm{HA} + \mathrm{H_2O}$$

added in buffer

The added $\mathrm{H_3O^+}$ or OH$^-$ ions will be prevented from affecting the pH by combing with either the HA or A$^-$ ions already in the solution.

2. Weak base/salt of conjugate acid (e.g. ammonia/ammonium chloride)

$$\mathrm{BH^+X^-} \rightarrow \mathrm{BH^+} + \mathrm{X^-}$$

$$\mathrm{B} + \mathrm{H_2O} \leftrightarrow \mathrm{BH} + \mathrm{OH^-}$$

$$\left[\mathrm{OH^-}\right] = K_B \frac{[\mathrm{B}]}{\left[\mathrm{BH^+}\right]}$$

$$\mathrm{pOH} = -\log\left(K_B \frac{[\mathrm{B}]}{\left[\mathrm{BH^+}\right]}\right)$$

$$\mathrm{pH} = 14 - \mathrm{pOH}$$

Complete dissociation of the salt into its ions results in the [BH$^+$] being dominated by the molarity of the salt solution. Because of the presence of the common ion BH$^+$, the equilibrium for the base reaction lies on the reactants side of the equation and [B] is therefore practically equal to the molarity of the base.

Any added $\mathrm{H_3O^+}$ or OH$^-$ ions will be prevented from affecting the pH by combing with either the B or BH$^+$ ions already in the solution.

9.9 Indicators

Indicators are either a weak acid or a weak base with the property of exhibiting a change in colour depending on the pH of the solution. The molecular form of the acid HA has one colour, and the ionic form, A^-, another colour.

$$\underset{\text{Colour 1}}{HA + H_2O} \leftrightarrow \underset{\text{Colour 2}}{H_3O^+ + A^-}$$

If a small amount of indicator is added to a solution containing a large concentration of H_3O^+, then the equilibrium for the indicator solution shifts towards the reactants and Colour 1 dominates. If the indicator is added to a solution of large concentration of OH^-, then the equilibrium shifts towards the products and Colour 2 dominates.

Typically, a few drops of indicator are added to a solution whose pH is to be determined, and the resulting colour of the solution indicates the pH when read off against a colour chart.

For the indicator:

$$HA + H_2O \leftrightarrow H_3O^+ + A^-$$

$$K_A = \frac{[H_3O^+][A^-]}{[HA]}$$

$$[H_3O^+] = K_A \frac{[HA]}{[A^-]}$$

Indicator	K_A	pH of colour change
Methyl orange	2×10^{-4}	3.7
Litmus	1×10^{-7}	7
Phenolphthalein	7×10^{-10}	9.1

An indicator changes colour when $[HA] = [A^-]$,

$$\therefore [H_3O^+] = K_A$$

Indicators are useful when they exist in a solution whose pH is slowly changing. If the pH, or the $[H_3O^+]$ of a solution changes so that [HA] becomes equal to $[A^-]$, and the $[H_3O^+]$ of the solution becomes equal to the K_A of the indicator and the colour changes.

It is important to know that an indicator does not necessarily change colour when $[H_3O^+] = [OH^-]$. Some indicators change colour on the basic side, while others change on the acidic side, and some at the neutral position, or **equivalence point**.

The **end point** occurs when the indicator changes colour. The reason why this is called the end point becomes clear when indicators are used in the process of **titration**.

9.10 Neutralisation

When an acid reacts with a base, the H^+ ions combine with the OH^- ions to form water, and the acidic and basic properties of the solution disappear along with the disappearing ions. This is **neutralisation**. That is, an acid on a base gives salt and water.

Consider 0.1M HCl added to 0.1M NaOH:

$$\mathrm{HCl + NaOH \rightarrow NaCl + H_2O}$$

More specifically, in 0.1M solutions, we do not have HCl molecules and NaOH molecules. They are already dissociated into their constituent ions. The reaction is therefore actually

$$\mathrm{H_3O^+ + Cl^- + Na^+ + OH^- \rightarrow Na^+ + Cl^- + 2H_2O}$$

The net reaction is therefore

$$\mathrm{H_3O^+ + OH^- \rightarrow 2H_2O}$$

We know from experience that the products of this reaction are salt and water, or more specifically, Na^+ ions and Cl^- ions and water. That is, salty water, and not a mixture of acidic and basic solutions. But, how do we know this? Because in any aqueous solution at equilibrium, $K_w = 1 \times 10^{-14}$.

Equilibrium favours the products because the reverse reaction, the dissociation of water, has such a low **equilibrium constant** (or **ion product** in this case). The H_3O^+ ions combine with the OH^- ions to form water until such time as the product of the concentrations of H_3O^+ and OH^- ions fall to 1×10^{-14} moles/litre.

For strict neutralisation of a solution to occur, the initial concentrations of the H_3O^+ and OH^- ions must be similar – that is, the acid and the base must be of about the same strength. That is, if equal moles of acid and base are added, the final solution $[H_3O^+] = [OH^-] = 1 \times 10^{-7}$.

In a solution of acetic acid, very little dissociation occurs and so the reaction with a strong base in aqueous solution is:

$$\mathrm{HC_2H_3O_2 + Na^+ + OH^- \rightarrow H_2O + Na + C_2H_3O_2^-}$$

$$\mathrm{HC_2H_3O_2 + OH^- \rightarrow H_2O + C_2H_3O_2^-}$$

The resulting solution is not neutral, but slightly basic in the sense that one of the products, $C_2H_3O_2^-$, is a Bronsted–Lowry base. In general,

$$\mathrm{HA + OH^- \rightarrow H_2O + A^-}$$

An interesting application of neutralisation is the use of antacids for relief of heartburn. In order to raise the pH in the stomach, sodium bicarbonate can be used to react with HCl in the gastric juices to give H_2O and $CO_2(g)$.

9.11 Titration

The progressive addition of an acid to a base, or a base to an acid, is called **titration**. Typically, when a base of known concentration is added to an acid of unknown pH, we say that the acid is titrated with the base. A known number of moles of base is added to the acid, either in solid form, or more commonly, as a solution via a burette.

As the base is added, the concentration of $[H_3O^+]$ changes due to neutralisation. If the pH is monitored during the process, a titration curve can be plotted showing pH against moles of added base.

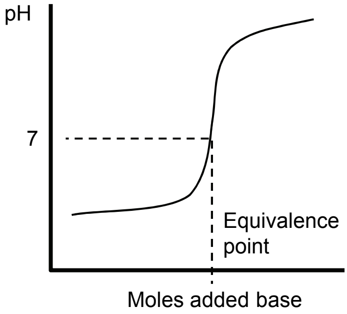

The moles of added base are measured usually by use of a burette. For the case of a strong acid and a strong base, there is a sharp change in pH at the equivalence point, where $[H_3O^+] = [OH^-]$. At or near this point, a very small addition of base results in a large increase in pH.

At the start of the titration, $[H_3O^+]$ is the molarity of the acid (for a strong acid). Before the equivalence point is reached, $[H_3O^+]$ is found from the initial moles of $[H_3O^+]$ less the moles of added OH^- (from the burette). At the equivalence point, $[H_3O^+] = [OH^-] = 1 \times 10^{-7}$. After the equivalence point, there is an excess of OH^-. Since we know how much OH^- is used to neutralise the acid, we know how much excess $[OH^-]$ exists after the equivalence point (from the burette reading) and the pH can be calculated from 14 – pOH.

The change in pH of the solution as the base is added can be measured using a pH meter, or more conventionally, using an **indicator**. When the indicator changes colour, the **end point**, the titration is finished.

9.12 Choice of Indicator

During a titration, the indicator to be used should have an **end point** which corresponds to the **equivalence point**. At first sight, this might sound like that the only useful indicator would be one that has an end point at pH = 7. However, the choice of indicator depends on the nature of the reactants.

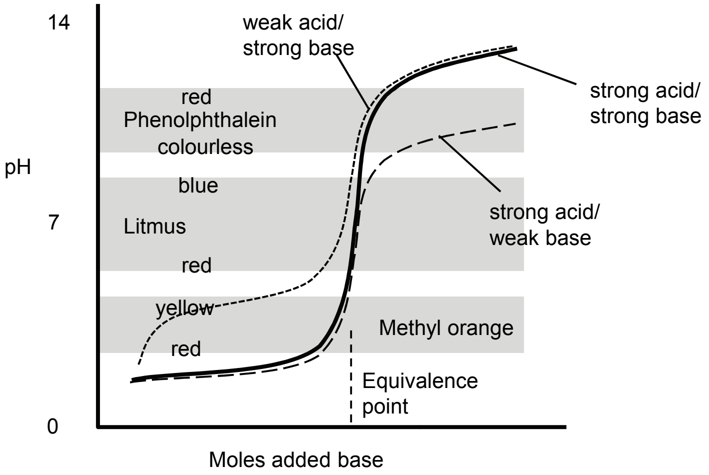

In the case of a strong acid being titrated with a strong acid, any of the three indicators above are suitable. However, for the case of a weak acid and a strong base, **Methyl orange** would not be suitable because it would change colour too early. **Phenolphthalein** would be a better choice because its end point (pH at which colour changes) more closely coincides with the equivalence point. For the case of a strong acid and a weak base, Methyl orange would be suitable, but not Phenolphthalein. **Litmus** is usually considered a poor choice because it changes colour over too wide a range of pH compared to other indicators. Weak acids are not usually used with weak bases in titration experiments.

A **universal indicator** is a mixture of other indicators designed so that different colours indicate different pH values.

In swimming pool maintenance, a few drops of phenol red (colour change from yellow to red over a pH range of 6.8 to 8.4) are used to measure pH. When the water is basic, the acid demand test involves putting a few drops of acid into the test mixture to determine how much acid is required to bring the solution back to the desired range of pH (usually 7.2 to 7.6).

9.13 Hydrolysis

Hydrolysis is the chemical reaction of ions with water. **Anions** may function as Bronsted-Lowry bases and receive a proton from the water molecule:

$$B^{n-} + H_2O \leftrightarrow OH^- + BH^{n-1}$$

Halide ions do not **hydrolyse** to any appreciable extent:

$$Cl^- + H_2O \leftarrow HCl + OH^- \qquad pH \approx 7$$

Oxide and sulphide ions hydrolyse:

$$O_2^{2-} + H_2O \rightarrow 2OH^- \qquad pH < 7$$

$$S^{2-} + H_2O \rightarrow HS^- + 2OH^-$$

Weak conjugate bases of strong acids do not hydrolyse:

$$ClO_4^- + H_2O \leftarrow HClO_4 + OH^- \qquad pH \approx 7$$

Strong conjugate bases of weak acids do hydrolyse:

$$C_2H_3O_2{}^- + H_2O \leftrightarrow HC_2H_3O_2 + OH^- \qquad pH < 7$$

Cations may function as Bronsted-Lowry acids and donate proton to the water molecule:

$$A^{n+} + H_2O \leftrightarrow H_3O^+ + A^{n-1}$$

Many elements occur in compounds as oxides and hydroxides. When hydrolysed, water molecules attach themselves to anions or cations formed by their dissociation in solution and these hydrolysed ions may act as acids or bases. The extent to which this occurs depends on the size of the atom and the charge.

Hydrated cations of groups I and II do not hydrolyse. Hydrated transition metal ions (Al_3^+, NH_4^+) hydrolyse to give acid solutions.

$$[Fe(H_2O)_6]^{3+} + H_2O \rightarrow H_3O^+ + [Fe(H_2O)_5OH]^{2+}$$

The **salts** of strong acids to not hydrolyse. For example, NaCl dissolved in water produces

$$NaCl \rightarrow Na^+ + Cl^-$$

$$Na^+ + H_2O \rightarrow \text{no reaction}$$

$$pH = 7$$

$$Cl^- + H_2O \rightarrow \text{no reaction}$$

The salts of weak acids and strong bases do hydrolyse. For example, $NaC_2H_3O_2$ (sodium acetate) when dissolved in water produces

$$NaC_2H_3O_2 \rightarrow Na^+ + C_2H_3O_2{}^-$$

$$Na^+ + H_2O \rightarrow \text{no reaction}$$

$$C_2H_3O_2{}^- + H_2O \leftrightarrow HC_2H_3O_2 + OH^- \qquad pH > 7$$

The salts of strong acids and weak bases do hydrolyse:

$$NH_4Cl \rightarrow NH^{4+} + Cl^-$$

$$Cl^- + H_2O \rightarrow \text{no reaction}$$

$$NH^{4+} + H_2O \rightarrow NH_3 + H_3O^+ \qquad pH < 7$$

The salts of weak acids and weak bases do hydrolyse but the resulting pH depends on the solubility constants K_A and K_B.

9.14 Simultaneous Equilibria

In the ideal case, only one equilibrium constant need be satisfied in a chemical reaction, but in practice, several equilibria are operating at the same time. For example, consider the hydrolysis reaction involving the dissociation of $(NH_4)_2S$ in solution:

$$NH_4^+ + H_2O \leftrightarrow NH_3 + H_3O^+$$
$$S^{2-} + H_2O \leftrightarrow HS^- + OH^-$$

There are two equilibria to be satisfied:

$$\begin{aligned} pK_{A\left(NH_4^+\right)} &= 9.2 \\ pK_{B\left(S^{2-}\right)} &= 0.2 \\ pK_{B\left(S^{2-}\right)} &> pK_{B\left(NH_4^+\right)} \\ &\therefore pH > 7 \end{aligned}$$

The quantitative analysis of hydrolysis involves the simultaneous equilibria of two reactions. For example consider the hydrolysis of the acetate ion:

$$C_2H_3O_2{}^- + H_2O \leftrightarrow HC_2H_3O_2 + OH^-$$

$$\begin{aligned} K_w &= \left[H_3O^+\right]\left[OH^-\right] \\ &= 1\times10^{-14} \end{aligned}$$

$$K_h = \frac{\left[OH^-\right]\left[HC_2H_3O_2\right]}{\left[C_2H_3O_2^-\right]}$$

the **hydrolysis equilibrium constant**

$$\left[OH^-\right] = \frac{K_w}{\left[H_3O^+\right]}$$

$$\begin{aligned} K_h &= \frac{K_w\left[HC_2H_3O_2\right]}{\left[C_2H_3O_2^-\right]} \\ &= \frac{K_w}{K_A} \end{aligned}$$

The hydrolysis equilibrium constant is expressed in terms of the associated acid (or base) equilibrium constant.

9.15 Complex Ions

Ligands are molecules or anions that have a lone pair of electrons. NH_3, H_2O and CN^- are examples. Some ligands combine with transition elemental metal ions – they become coordinated with the ion to form a **complex ion**.

$$[\mathrm{Cu(NH_3)_4}]^{2+}$$
$$[\mathrm{Al(H_2O)_6}]^{3+}$$
$$[\mathrm{Ag(CN)_2}]^{-}$$

$$[\underset{\text{metal ion}}{\mathrm{M}}(\overset{\text{ligand}}{\mathrm{L}})_{\underset{\text{coordination number}}{\mathrm{X}}}]^{n+/-}$$

The **coordination number** usually equals twice the charge on the metal ion. The overall complex ion charge equals the sum of the cation and ligand charges. Complex ions may be anionic, cationic or neutral.

The properties of complex ions depend very much on the geometrical arrangement of molecules around the central atom. In general terms, the ligand provides electron pairs which find a place to attach themselves at available electron *d* orbitals around the central metal ion. However, this can happen in several ways. For example, consider the complex ion formed by Co and the ligands Cl^- and H_2O:

$[\mathrm{Co(Cl)_4}]^{2-}$ coordination number 4

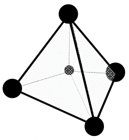

$[\mathrm{Co(H_2O)_6}]^{2+}$ coordination number 6

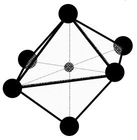

The lone pairs on the ligands face inwards towards and occupy *d* orbitals on the central metal ion. The reaction in water is

$$\underset{\text{blue}}{[\mathrm{Co(Cl)_4}]^{2-}} + 6\mathrm{H_2O} \leftrightarrow \underset{\text{pink}}{[\mathrm{Co(H_2O)_6}]^{2+}} + 4\mathrm{Cl}^-$$

Colours are often prominent in these compounds because the energy levels associated within the *d* orbitals are in the **visible light** range. Certain wavelength components of white light passing through solutions containing these ions are absorbed and the transmitted light that we see appears coloured.

10. Electronic Equilibrium

Summary

- **Oxidation** occurs at the **anode**. Neutral atoms being oxidised **lose** their electrons and become positively charged **cations** which tend to move towards the **cathode**.

$$\begin{aligned}4\mathrm{Fe}(\mathrm{s}) &\rightarrow 2\mathrm{Fe}_2^{2+}(\mathrm{aq})+4\mathrm{e}^-\\ \mathrm{Cu}(\mathrm{s}) &\rightarrow \mathrm{Cu}^{2+}(\mathrm{aq})+2\mathrm{e}^-\\ \mathrm{Zn}(\mathrm{s}) &\rightarrow \mathrm{Zn}^{2+}(\mathrm{aq})+2\mathrm{e}^-\\ \mathrm{H}_2^{2+}(\mathrm{g}) &\rightarrow 2\mathrm{H}^+ + 2\mathrm{e}^-\end{aligned}$$

Oxidation occurs when the oxidation number of an atom is increased (loss of electrons).

- **Reduction** occurs at the **cathode**. Positively charged ions in solution take up the electrons coming from the oxidation reaction and become neutral atoms. The remaining negatively charged **anions** tend to move towards the **anode**.

$$\begin{aligned}3\mathrm{O}_2 + 4\mathrm{e}^- &\rightarrow 2\mathrm{O}_3^{2-}\\ 2\mathrm{Ag}^{2+} + 2\mathrm{e}^- &\rightarrow 2\mathrm{Ag}(\mathrm{s})\\ 2\mathrm{H}^+ + 2\mathrm{e}^- &\rightarrow 2\mathrm{H}_2(\mathrm{g})\\ \mathrm{Cu}^{2+} + 2\mathrm{e}^- &\rightarrow \mathrm{Cu}(\mathrm{s})\end{aligned}$$

Reduction occurs when the oxidation number of an atom is decreased (gain of electrons).

10.1 Oxidation and Reduction

Oxidation and reduction reactions involve the **equilibrium** of a competition for electrons between atoms and ions. It is **electronic equilibrium**. To see how this comes about, we need to look at some numbers that quantify what drives this type of reaction.

The **oxidation number**, or **oxidation state**, is a quantity that describes the number of electrons that seem to be gained or lost by an atom when a chemical bond is formed. During a chemical reaction, **oxidation** occurs when the oxidation number of an atom is increased (loss of electrons). **Reduction** occurs when the oxidation number of an atom is decreased (gain of electrons). That is, the term reduction indicates a reduction in oxidation number.

Consider a simplified version of the oxidation of iron:

$$4\mathrm{Fe(s)} + 3\mathrm{O_2(g)} \rightarrow 2\mathrm{Fe_2O_3(s)}$$

The **oxidation half-reaction** is

$$4\mathrm{Fe(s)} \rightarrow 2\mathrm{Fe_2^{2+}} + 4e^-$$

In this half-reaction, the oxidation number of the Fe atoms increases from 0 to 2+ (loses electrons)

The **reduction half-reaction** is

$$3\mathrm{O_2} + 4e^- \rightarrow 2\mathrm{O_3^{2-}}$$

In this half-reaction, the oxidation number of the O atoms decreases from 0 to 2− (gains electrons).

The terms oxidation and reduction are historical in origin. They arose from the common observation that some reactants combine with oxygen (and so are said to be **oxidised**) while others lose oxygen atoms (and so are said to be **reduced**). Now it is recognised that it is the loss and gain of electrons that is the important feature of such reactions.

The **oxidising agent**, or **oxidant**, is the substance that is reduced. In the example above, oxygen is the oxidising agent since oxygen oxidises the iron. The **reducing agent**, or the **reductant**, is the substance that is oxidised. A **redox couple** is an oxidant and its complementary reductant. In the above example, the redox couples are written

$$\mathrm{Fe/Fe_2^{2+}} \quad \text{and} \quad \mathrm{O_2/O_3^{2-}}$$

Oxidant/Reductant
(or Reduction/Oxidation)

The processes of oxidation and reduction do not occur in all chemical reactions. In many reactions, the oxidation numbers of the participating atoms do not change. Examples of **non-redox reactions** are some precipitation reactions, acids reacting with bases, acids and oxides, hydroxides, sulphides and carbonates. In **acid-base reactions**, it is more an exchange of protons that is important, not electrons.

10.2 Redox Reaction

Consider the redox reaction: $\mathrm{Cu(s)+2Ag^{2+}(aq)\rightarrow Cu^{2+}(aq)+2Ag(s)}$

We can separate out the oxidation and reduction processes physically using separate beakers of solution, a wire for electrons to flow along, and a salt bridge for ions to move across and electrodes where reactions take place.

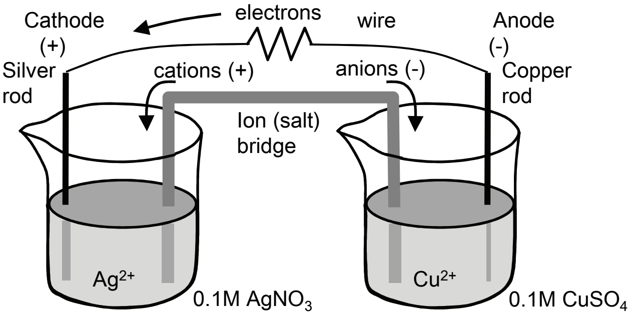

$\mathrm{2Ag^{2+}(aq)+2e^{-}\rightarrow 2Ag(s)}$ $\mathrm{Cu(s)\rightarrow Cu^{2+}(aq)+2e^{-}}$

Reduction half-reaction Oxidation half-reaction

The salt bridge is a tube containing a salt solution (e.g. $NaNO_3$) and allows negatively charged anions to move into the anode solution and positively charged cations to move into the cathode solution while at the same time preventing bulk mixing of the solutions.

The electrode where reduction occurs is called the **cathode**. Silver has a greater affinity for **electrons** than copper. So, the Ag^{2+} ions in the solution combine with the free electrons in the cathode to form solid silver. This loss of free electrons at the cathode represents a net positive charge. The Ag^{2+} ions are pulling the electrons off the silver rod and forming solid silver at the cathode. But the consumption of Ag^{2+} ions in the solution results in a build-up of negative charge (the NO_3^- ions), and if this is not balanced in some way, the reaction will come to a stop as the ions in the solution repel any more electrons arriving at the cathode wanting to take part in the reaction.

The electrode where oxidation occurs is called the **anode.** The copper rod at the anode dissolves away to form Cu^{2+} ions and free **electrons**. These free electrons travel along the wire and are consumed at the cathode. The build-up of positively charge Cu^{2+} ions would eventually overcome the affinity for electrons by the silver and the reaction would come to a stop.

If the NO_3^- **anions** are allowed to flow from the cathode solution to the anode solution, and the Cu^{2+} **cations** flow from the anode solution to the cathode solution, then build-up of excess charge would be prevented and the reactions at each electrode would proceed spontaneously.

The redox reactions, arranged in this way, constitute an **electrochemical cell** (or **galvanic cell**) since a **voltage** is produced between the cathode and the anode. The sign of the anode and cathode is determined by the flow of electrons in the external wire, not the ion flow in the solutions.

10.3 Single-Cell Redox Reaction

Consider the redox reaction

$$\mathrm{Cu(s)} + 2\mathrm{Ag}^{2+}(\mathrm{aq}) \rightarrow \mathrm{Cu}^{2+}(\mathrm{aq}) + 2\mathrm{Ag(s)}$$

We have seen that this is an oxidation-reduction reaction whereby each half-reaction can be arranged to occur in separate solutions. This reaction can also occur in a single solution.

$$2\mathrm{Ag}^{2+}(\mathrm{aq}) + 2\mathrm{e}^{-} \rightarrow 2\mathrm{Ag(s)}$$

Reduction half-reaction

$$\mathrm{Cu(s)} \rightarrow \mathrm{Cu}^{2+}(\mathrm{aq}) + 2\mathrm{e}^{-}$$

Oxidation half-reaction

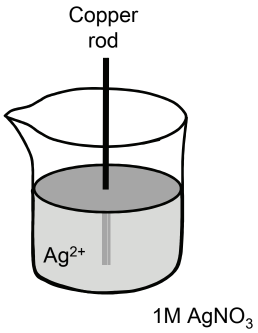

Redox reactions of this kind find application in the silvering of copper conductors in electrical switch gear whereby the oxidation (dulling) of copper due to atmospheric contaminants is to be avoided.

When a copper rod is dipped into a silver nitrate solution, silver metal is spontaneously deposited on the rod. The solution turns blue. Electron transfer occurs from the copper to the silver ions in the vicinity of the rod.

Another good example of a single-cell reaction is when a zinc rod is dipped into hydrochloric acid.

$$2\mathrm{H}^{+} + 2\mathrm{e}^{-} \rightarrow 2\mathrm{H}_2(\mathrm{g})$$

Reduction half-reaction

$$\mathrm{Zn(s)} \rightarrow \mathrm{Zn}^{2+}(\mathrm{aq}) + 2\mathrm{e}^{-}$$

Oxidation half-reaction

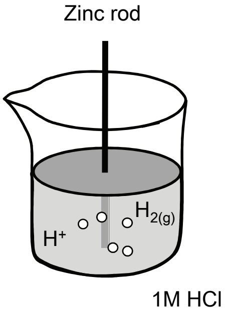

The metal dissolves in the acid, producing H_2 gas. The same thing happens with Mg, Al, Fe and Ni, but not with Cu, Ag and Au. The reason for some metals being able to be dissolved in HCl and others not can be found by a consideration of **standard electrode potentials**.

10.4 Standard Hydrogen Electrode

Consider the redox reaction

$$Cu^{2+}(aq) + H_2(g) \rightarrow Cu(s) + 2H^+(aq)$$

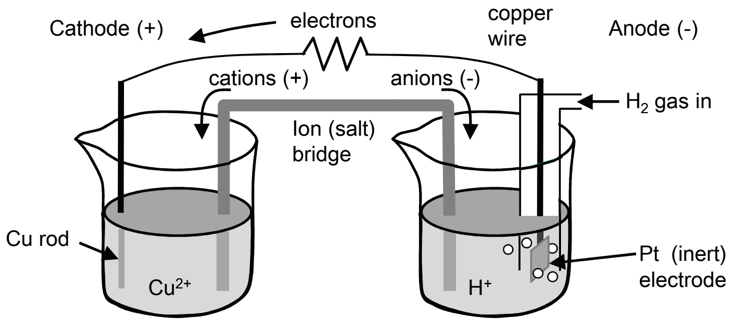

$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$ $H_2^{2+}(g) \rightarrow 2H^+ + 2e^-$

Reduction half-reaction Oxidation half-reaction

Cu^{2+}/Cu Redox couple $H^+/H_2(g)$ Redox couple

When the H_2 gas is at 1 atm and the H^+ solution is 1 M, the H^+/H_2 redox couple is referred to as a **standard hydrogen electrode** and is assigned a 0V potential. The potential difference between the hydrogen standard cell and the other electrode is called the **standard electrode potential** E^o for the half-reaction occurring at that other electrode.

Standard electrode potentials are provided from the point of view of **oxidation** at the hydrogen standard electrode. That is, positive values of standard electrode potentials for a redox reaction indicate that the reaction will be one of **reduction** when the other half-cell is the standard cell. When the standard electrode potential for a redox reaction is listed as negative, it indicates that this reaction will be an **oxidation** reaction if the other electrode is the standard hydrogen electrode.

The standard electrode potential E^o for the Cu^{2+}/Cu half-reaction above is +0.34V and so the copper is reduced and the hydrogen is oxidised. The value of E^o for Zn is –0.76V and so if a zinc electrode and solution were used in place of the copper, the zinc would be oxidised and the hydrogen would be reduced.

10.5 Standard Electrode Potentials

Standard electrode potentials E° for some **reduction half-reactions** where gases are at 1 atm and solutions are 1M:

$F_2(g) + 2e^- \rightarrow 2F^-$	+2.87
$O_3(g) + 2H^+(aq) + 2e^- \rightarrow 2H_2O$	+2.07
$MnO_4^{2+}(aq) + 8H^+(aq) + 5e^- \rightarrow Mn^{2-}(aq) + 4H_2O$	+1.51
$Cl_2(g) + 2e^- \rightarrow 2Cl^-(aq)$	+1.36
$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O$	+1.229
$Ag^+(aq) + e^- \rightarrow Ag(s)$	+0.799
$Fe^{3+}(aq) + e^- \rightarrow Fe^{2+}(aq)$	+0.771
$Cu^+(aq) + e^- \rightarrow Cu(s)$	+0.52
$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$	+0.337
$Sn^{4+}(aq) + 2e^- \rightarrow Sn^{2+}(aq)$	+0.15
$S(aq) + 2H^+ + 2e^- \rightarrow H_2S(g)$	+0.141
$2H^+(aq) + 2e^- \rightarrow H_2^{2+}(g)$	0
$Pb^{2+}(aq) + 2e^- \rightarrow Pb(s)$	-0.126
$Sn^{2+}(aq) + 2e^- \rightarrow Sn(s)$	-0.136
$Fe^{2+}(aq) + 2e^- \rightarrow Fe(s)$	-0.440
$Zn^{2+}(aq) + 2e^- \rightarrow Zn(s)$	-0.763
$Al^{3+}(aq) + 3e^- \rightarrow Al(s)$	-1.66
$Mg^{2+}(aq) + 2e^- \rightarrow Mg(s)$	-2.37
$Na^+(aq) + e^- \rightarrow Na(s)$	-2.714
$Ca^{2+}(aq) + 2e^- \rightarrow Ca(s)$	-2.89
$K^+(aq) + e^- \rightarrow K(s)$	-2.925
$Li^+(aq) + e^- \rightarrow Li(s)$	-3.045

10.6 Spontaneous Redox Reactions

A very useful application of standard electrode potentials is the determination of whether a redox reaction will proceed spontaneously or not. Consider the copper and zinc redox couples:

$$\mathrm{Cu}^{2+}/\mathrm{Cu} \qquad \mathrm{Zn}^{2+}/\mathrm{Zn}$$

$$E^{\mathrm{o}} = +0.34\ \mathrm{V} \qquad E^{\mathrm{o}} = -0.76\ \mathrm{V}$$

There are four possible reactants and two possible reactions:

1 $\mathrm{Cu}^{2+}_{(ox)} + \mathrm{Zn}^{2+}_{(ox)}$ No reaction (simultaneous oxidation)

2 $\mathrm{Cu}^{2+}_{(ox)} + \mathrm{Zn}_{(red)}$ $\mathrm{Cu}^{2+}(\mathrm{aq}) + \mathrm{Zn(s)} \rightarrow \mathrm{Cu(s)} + \mathrm{Zn}^{2+}(\mathrm{aq})$

3 $\mathrm{Cu}_{(red)} + \mathrm{Zn}^{2+}_{(ox)}$ $\mathrm{Cu}_{(red)} + \mathrm{Zn}^{2+}_{(ox)} \rightarrow \mathrm{Cu}^{2+}(\mathrm{aq}) + \mathrm{Zn(s)}$

4 $\mathrm{Cu}_{(red)} + \mathrm{Zn}_{(red)}$ No reaction (simultaneous reduction)

The first and fourth possibilities cannot happen since we must have complementary oxidation and reduction processes happening. To determine which of the second and third possibilities will happen spontaneously when these chemicals are in an **electrochemical cell**, we add the **standard electrode potentials**. However, the standard electrode potentials are expressed in terms of reduction reactions, and so we must reverse the sign for a corresponding oxidation reaction.

For the second possibility, the electrode potential for the oxidation of copper in solution would be –0.34V. This is added to –0.76 for the reduction of zinc to obtain –1.1V. Since this total is a negative voltage for the cell as a whole, we say that this reaction will not proceed of its own accord.

For the third possibility, the electrode potential for the reduction of copper in solution is +0.34V. This is added to +0.76 for the oxidation of zinc to obtain +1.1V. Since this total is a positive value, we say that this reaction will proceed of its own accord and what's more, we would measure a **potential difference** of +1.1V across the terminals of the cell.

Standard electrode potentials for a redox reaction indicate the **reduction** potential of the half-reaction. Generally speaking, a redox reaction will proceed spontaneously if the **oxidant** (the one that is reduced) from the redox couple with the higher value of E^{o} reacts with the **reductant** of the other.

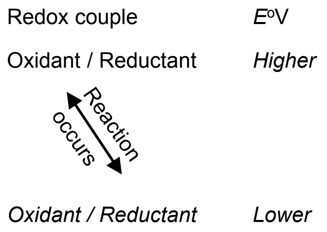

10.7 Oxidation Numbers

Although we can predict which redox reactions will occur spontaneously from a consideration of **standard electrode potentials**, we have yet to consider why the standard potentials are what they are. Consider the redox reaction

$$\mathrm{Cu(s) + 2Ag^{2+}(aq) \rightarrow Cu^{2+}(aq) + 2Ag(s)} \quad \text{Redox reaction}$$

$$\mathrm{2Ag^{2+}(aq) + 2e^{-} \rightarrow 2Ag(s)} \quad \text{Reduction half-reaction}$$

$$\mathrm{Cu(s) \rightarrow Cu^{2+}(aq) + 2e^{-}} \quad \text{Oxidation half-reaction}$$

From the above, we can conclude that silver ions have a stronger affinity for attracting **electrons** than do copper ions, and so the silver ions are reduced to a solid by taking up electrons and the solid copper is oxidised by releasing electrons, rather than the reverse. It is the relative affinity of ions to attract electrons that places a particular element in the list of standard potentials – which as will be remembered is presented from the point of view of reduction half-reactions.

This affinity for electrons by atoms is something we have met before: **electronegativity**. It was found that a practical measure of electronegativity is the method of electron counting by **oxidation numbers**. In a chemical reaction, when the oxidation number of an atom (i.e. its apparent charge) has increased, the atom is said to have been **oxidised**. When the oxidation number of an atom has decreased, the atom is said to have been **reduced**.

In the above reactions, the oxidation number of copper is initially 0. After reaction, the copper atoms have apparently lost two electrons and so its oxidation number is now 2+ (its apparent charge), and is thus an increase. The oxidation number of silver in the $\mathrm{Ag^{2+}}$ ions is initially 2+, and after reaction is 0. The oxidation number has decreased and the silver is therefore reduced. All this comes down to a competition for the attraction of electrons.

- **Oxidation** represents a loss of electrons or increase in oxidation number.
- **Reduction** represents a gain of electrons or decrease in oxidation number.

Oxidation numbers allow us to recognise oxidation-reduction reactions and are also useful for balancing the equations for these reactions. Changes in oxidation number do not necessarily indicate that an atom has gained or lost an electron, or has gained or lost net electric charge, even though we use them in this way for the purpose of establishing the numbers. Rather, they indicate how many electrons are involved for the atom in a chemical reaction.

10.8 Balancing Redox Half-Reactions

In redox reactions, both atoms and charges must be conserved. A convenient method of balancing redox equations is to use the equations for the half-reactions.

Consider the reaction between copper and nitric acid, HNO_3. The redox couples involved are

$Cu^{2+}/Cu(s)$	$H^+, NO_3^- / NO(g)$
$E^o = +0.34V$	$E^o = +0.96V$
Since E^o for the nitric acid is more positive, then the copper must undergo oxidation:	and the NO_3 must undergo reduction:
$Cu(s) \rightarrow Cu^{2+}(aq) + 2e^-$	$4H^+ + NO_3^- + 3e^- \rightarrow NO(g) + H_2O$

Both these half-reactions are balanced, but they now need to be combined to form the net reaction, which is required also to be balanced. The procedure is to balance the charge. The lowest common factor in this example is 6, and so

$3Cu(s) \rightarrow 3Cu^{2+}(aq) + 6e^-$ $\quad 8H^+ + 2NO_3^- + 6e^- \rightarrow 2NO(g) + 4H_2O$

These half-reactions can now be combined:

$$3Cu + 8H^+ + 2NO_3^- \rightarrow 3Cu^{2+} + 2NO(g) + 4H_2O$$

Not all reactions can be balanced so easily. A general method is to

1. Balance for atoms other than O and H
2. Balance for O by adding water
3. Balance for H by adding H^+
4. Add OH^- ions to both sides for basic solution to form water with the H^+ added from (3)
5. Balance for charge by adding electrons

It should be noted that balancing a half-reaction by adjusting coefficients does not alter the electrode potential for that half-reaction.

10.9 Balancing Redox Reactions with Oxidation Numbers

Another method of balancing redox reactions is by the use of oxidation numbers. Consider the reaction between copper and nitric acid HNO_3. The redox couples involved are:

$$\begin{array}{ll} Cu^{2+}/Cu(s) & NO_3^-/NO(g) \\ E^o = +0.34\text{ V} & E^o = +0.96\text{ V} \end{array}$$

Since E^o for the nitric acid is more positive, then the copper must undergo oxidation. The unbalanced equation is:

$$Cu + NO_3^- \rightarrow Cu^{2+} + NO(g)$$

Next, we write the oxidation numbers to each element in the reaction:

$$\begin{array}{ccccccc} Cu & + & NO_3^- & \rightarrow & Cu^{2+} & + & NO(g) \\ 0 & & 5+,2- & & 2+ & & 2+,2- \end{array}$$

Cu has undergone a change in oxidation number from 0 to 2+. An increase in oxidation number indicates oxidation. The change in oxidation number in this element is +2.

Rules for oxidation numbers

a) In free, uncombined, elements, the oxidation number of each atom is set to be 0.
b) In compounds involving hydrogen, the oxidation number of hydrogen is usually 1+.
c) In compounds involving oxygen, the oxidation number of O is usually 2–.
d) The sum of all the oxidation numbers of all atoms in an ion is equal (in both magnitude and sign) to the charge on the ion.
e) The sum of all the oxidation numbers of all atoms in a neutral molecule is 0.

N has undergone a change in oxidation number of 5+ to 2+. The change in oxidation number is –3. A decrease in oxidation number indicates reduction. There is no net change in oxidation number for the oxygen atoms.

We now introduce a further rule for oxidation numbers, and that is, the net *change* in oxidation number for the overall reaction is to be zero. That is, the overall *change* in oxidation number for the left side of the reaction has to be equal to the negative of the overall *change* oxidation number for the right side. The only way to do this here is to multiply the Cu atoms by 3 and the NO_3^- ions by 2. That is, operating on the changes in oxidation numbers, we have: $3(+2)+2(-3)=0$ Thus,

$$3Cu + 2NO_3^- \rightarrow 3Cu^{2+} + 2NO(g).$$

We need to add H_2O to get a balance for oxygen atoms:

$$3Cu + 2NO_3^- \rightarrow 3Cu^{2+} + 2NO(g) + 4H_2O$$

And finally, add H^+ ions to balance for H atoms:

$$3Cu + 8H^+ + 2NO_3^- \rightarrow 3Cu^{2+} + 2NO(g) + 4H_2O$$

10.10 Electrochemical Cell

An **electrochemical cell**, **or galvanic cell**, is a redox reaction that proceeds spontaneously and has an **electrical potential difference** at its electrodes. This potential difference, when measured at a negligible current draw, is called the **emf** of the cell E_{cell}.

$$E_{\text{cell}} = \underset{\substack{\big\downarrow\\ \text{reduction}}}{E_{\text{cathode}}} - \underset{\substack{\big\downarrow\\ \text{oxidation}}}{E_{\text{anode}}}$$

When the half-cells are standard electrodes, the standard emf of the cell is given by

$$E^{\mathrm{o}}_{\text{cell}} = E^{\mathrm{o}}_{\text{cathode}} - E^{\mathrm{o}}_{\text{anode}}$$

Example of electrochemical cell:

$$\begin{array}{ll} \mathrm{Cu}^{2+}/\mathrm{Cu} & \mathrm{Fe}^{3+}(\mathrm{aq})/\mathrm{Fe}^{2+}(\mathrm{aq}),\mathrm{Pt} \\ E^{\mathrm{o}} = +0.34\mathrm{V} & E^{\mathrm{o}} = +0.77\mathrm{V} \end{array}$$

All at 1M concentrations

Since E^{o} for the iron reaction is higher than that for copper, then the Fe^{3+} is undergoing reduction (it is the oxidant) and so is the **cathode** for the cell. Copper is being oxidised and so is the **anode**.

$$\begin{aligned} E^{\mathrm{o}}_{\text{cell}} &= E^{\mathrm{o}}_{\text{cathode}} - E^{\mathrm{o}}_{\text{anode}} \\ &= 0.77 - 0.34 \\ &= +0.43\mathrm{V} \end{aligned}$$

Note: we do not have to "balance" the half-equations to determine the overall cell potential. For example, the **standard electrode potential** of the copper half cell is still +0.34 if we write

$$2\mathrm{Cu}^{2+} + 4\mathrm{e}^{-} \rightarrow 2\mathrm{Cu}(\mathrm{s})$$

The potentials do depend on the concentration of the species, but not the number of atoms that participate in the overall reaction.

In an electrochemical cell that does not have a connection between the anode and cathode, the redox reaction proceeds until the build-up of positive charge on the cathode prevents any more cations from moving from the anode and vice versa and the reaction stops. The system has acquired **electrical potential energy** as a result of the reaction proceeds. The reduction of energy of the system as a whole (i.e. the chemical reaction tending to send the system towards a lower energy state) appears as mechanical work or heat by virtue of the current flow.

10.11 EMF vs. Concentration

The potential of an electrochemical cell depends upon the concentration of the reactants. Standard potentials are given for 1M solutions. As the cell operates, the concentrations of the reactants change. The dependence of the cell potential on concentration is given by the **Nernst equation**.

$$p\mathrm{Ox} + ne^{-} \rightarrow q\mathrm{Red}$$

$$E = E^{\mathrm{o}} + \frac{RT}{nF}\ln\frac{[\mathrm{Ox}]^{p}}{[\mathrm{Red}]^{q}}$$

$$= E^{\mathrm{o}} + \frac{0.059}{n}\log_{10}\frac{[\mathrm{Ox}]^{p}}{[\mathrm{Red}]^{q}} \quad \text{at } 25°\mathrm{C}$$

R - universal gas constant

T - temperature (K)

F - Faraday's constant 96487 C/mol

Consider the electrochemical cell at 25°C:

$$\mathrm{Pt}\,|\,\mathrm{Sn}^{4+}(0.5M),\mathrm{Sn}^{2+}(1.5M)\,||\,\mathrm{Fe}^{3+}(1.0M),\mathrm{Fe}^{2+}(0.5M)\,|\,\mathrm{Pt}$$

Anode Cathode

$$\mathrm{Sn}^{4+} + 2\mathrm{e}^{-} \leftrightarrow \mathrm{Sn}^{2+}$$

$$E_{\mathrm{Sn}} = E^{\mathrm{o}}_{\mathrm{Sn}} + \frac{0.059}{2}\log_{10}\frac{[\mathrm{Sn}^{4+}]}{[\mathrm{Sn}^{2+}]}$$

$$= 0.15 + \frac{0.059}{2}\log_{10}\frac{[0.5]}{[1.5]}$$

$$= 9.14\mathrm{V}$$

$$\mathrm{Fe}^{3+} + \mathrm{e}^{-} \leftrightarrow \mathrm{Fe}^{2+}$$

$$E_{\mathrm{Fe}} = E^{\mathrm{o}}_{\mathrm{Fe}} + \frac{0.059}{1}\log_{10}\frac{[\mathrm{Fe}^{4+}]}{[\mathrm{Fe}^{2+}]}$$

$$= 0.77 + \frac{0.059}{1}\log_{10}\frac{[1.0]}{[0.5]}$$

$$= 0.79\ \mathrm{V}$$

Since $E_{\mathrm{Fe}} > E_{\mathrm{Sn}}$, $\mathrm{Fe}^{3+}|\mathrm{Fe}^{2+}$ is the cathode and the cell reaction is

$$2\mathrm{Fe}^{3+} + \mathrm{Sn}^{2+} \rightarrow 2\mathrm{Fe}^{2+} + \mathrm{Sn}^{4+}$$

$$E_{\mathrm{cell}} = 0.79 + -0.14$$

$$= 0.65\mathrm{V}$$

When the electrode is a metal in contact with its cation (e.g. $\mathrm{Cu}^{2+}|\mathrm{Cu}$), then

$$\mathrm{Cu}^{2+} + 2\mathrm{e}^{-} \leftrightarrow \mathrm{Cu}_{(\mathrm{s})}$$

$$E = E^{\mathrm{o}} + \frac{0.059}{2}\log_{10}[\mathrm{Cu}^{2+}]$$

When the electrode is a gas in contact with its ion (e.g. $\mathrm{H}^{+}|\mathrm{H}_2(\mathrm{g})$), then

$$2\mathrm{H}^{+} + 2\mathrm{e}^{-} \leftrightarrow \mathrm{H}_2(\mathrm{g})$$

$$E = E^{\mathrm{o}} + \frac{0.059}{2}\log_{10}\frac{[\mathrm{H}^{+}]^{2}}{(P_{H_2})} \longrightarrow \text{Pressure in atm}$$

10.12 Electronic Equilibrium

Thus far we have considered the conditions under which an oxidation-reduction reaction will proceed. Consider an electrochemical cell:

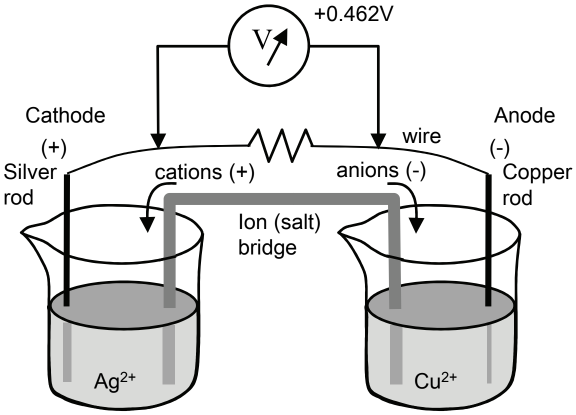

$2Ag^{2+} + 2e^- \rightarrow 2Ag(s)$

Reduction half-reaction

$Cu(s) \rightarrow Cu^{2+} + 2e^-$

Oxidation half-reaction

How long does the spontaneous reaction proceed for? Does this cell "run down"? Does the voltmeter reading start to drop and then go to zero?

Initially, we know from our standard electrode potentials that the attainment of equilibrium, the objective of all chemical reactions, favours the forward reaction as written:

$$Cu(s) + 2Ag^{2+}(aq) \rightarrow Cu^{2+}(aq) + 2Ag(s)$$

As the reaction proceeds, the Ag^{2+} ions are being used up in the cathode solution, and there is a build-up of Cu^{2+} ions in the anode solution. Electrons lose potential energy as they pass through the wire and this energy is converted to heat in the resistor. When the electrons combine with the silver ions at the cathode, they are in a lower energy state compared to when they were bound to the copper. Attainment of a lower energy state is the driving force behind most spontaneous chemical reactions.

When the silver ions are in short supply, then not all the electrons coming from the wire will find a home to go to and there will be a slight increase in negative charge at the cathode. When there are no more silver ions to accept these electrons, the build-up of negative electric charge will be great enough to prevent any more from arriving. At this point, the electrical potential difference between the anode and cathode will be zero and the cell will be exhausted.

10.13 Equilibrium Constant

The potential of an electrochemical cell depends upon the concentration of the reactants. When the cell is exhausted, we reach a state of equilibrium and the **equilibrium constant** for the reaction is satisfied. Consider the redox reaction:

$$5\mathrm{Fe}^{2+} + \mathrm{MnO}_4^- + 8\mathrm{H}^+ \rightarrow 5\mathrm{Fe}^{3+} + \mathrm{Mn}^{2+} + 4\mathrm{H}_2\mathrm{O}$$

$$K_\mathrm{c} = \frac{\left[\mathrm{Fe}^{3+}\right]^5\left[\mathrm{Mn}^{2+}\right]}{\left[\mathrm{Fe}^{2+}\right]^5\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}$$

At equilibrium, $E = 0$, and so

$$E^\mathrm{o}_\mathrm{Fe} + 0.059\log_{10}\frac{\left[\mathrm{Fe}^{3+}\right]}{\left[\mathrm{Fe}^{2+}\right]} = E^\mathrm{o}_{\mathrm{MnO}_4^-} + \frac{0.059}{5}\log_{10}\frac{\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}{\left[\mathrm{Mn}^{2+}\right]}$$

$$E^\mathrm{o}_{\mathrm{MnO}_4{}^-} - E^\mathrm{o}_\mathrm{Fe} = 0.059\log_{10}\frac{\left[\mathrm{Fe}^{3+}\right]}{\left[\mathrm{Fe}^{2+}\right]} - \frac{0.059}{5}\log_{10}\frac{\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}{\left[\mathrm{Mn}^{2+}\right]}$$

$$\Delta E^\mathrm{o} = \frac{0.059}{5}\left(\log_{10}\frac{\left[\mathrm{Fe}^{3+}\right]^5}{\left[\mathrm{Fe}^{2+}\right]^5} - og_{10}\frac{\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}{\left[\mathrm{Mn}^{2+}\right]}\right)$$

$$= \frac{0.059}{5}\log 10\frac{\left[\mathrm{Fe}^{3+}\right]^5\left[\mathrm{Mn}^{2+}\right]}{\left[\mathrm{Fe}^{2+}\right]^5\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}$$

$$= \frac{0.059}{5}\log_{10} K_c$$

$$K_c = \log 10\frac{5\Delta E^\mathrm{o}}{0.059}$$

$$= 5.1\times 10^{62}$$

$$E^\mathrm{o}{}_{\mathrm{MnO}_4^-} = +1.51\mathrm{V}$$

$$E^\mathrm{o}{}_\mathrm{Fe} = +0.77\mathrm{V}$$

10.14 Lead/Acid Battery

A **battery** is a collection of cells arranged in series. In a motor vehicle **lead acid battery**, there are six cells, each producing about 2V, arranged in series to provide 12V for the vehicle. The redox couples involved are:

$$\begin{array}{ll} PbSO_4(s)/Pb(s),SO_4^{2-} & PbO_2(s),H^+,SO_4^{2-}/PbSO_4(s) \\ E^o = -0.36V & E^o = +1.69V \end{array}$$

Since the standard potential for the PbO_2 reaction is higher, then this species undergoes reduction and is the cathode.

Reduction (cathode) $$PbO_2(s) + 4H^+ + SO_4^{2-} + 2e^- \rightarrow PbSO_4(s) + 2H_2O$$

Oxidation (anode) $$Pb(s) + SO_4^{2-} \rightarrow PbSO_4(s) + 2e^-$$

$$PbO_2(s) + 4H^+ + 2SO_4^{2-} \rightarrow 2PbSO_4(s) + 2H_2O$$ **Discharge** reaction

$$\begin{aligned} E^o_{cell} &= E^o_{cathode} - E^o_{anode} \\ &= 1.69 - -0.36 \\ &= 2.05V \end{aligned}$$

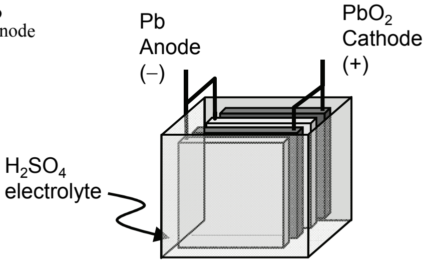

The concentration of H_2SO_4 in a motor vehicle battery is usually about 5M and so the actual cell potential is about 2.15V.

The **anode** (the negative plates) is solid lead, Pb, while the **cathode** (the positive plates) is **lead peroxide**, PbO_2, in a solution of **sulphuric acid**. The reaction product, **lead sulphate** $PbSO_4$, sometimes appears as a white precipitate on the battery terminals.

As the cell discharges, the concentration of H_2SO_4 decreases and the density of the solution also decreases. The density of the solution is thus a measure of the state of **charge** of the battery and is measured with a **hydrometer**.

When the battery is charged, the above reaction proceeds in reverse. Overcharging results in the evolution of gaseous H_2 at the anode and O_2 at the cathode, and the water in the electrolyte has to be replenished.

10.15 Dry Cell

In a conventional **dry cell**, or **Leclanché cell**, oxidation occurs at a zinc anode and reduction occurs at an inert carbon cathode. The electrolyte is not a liquid, but a paste of magnesium dioxide MnO_2, zinc chloride $ZnCl_2$, ammonium chloride NH_4Cl, and carbon black particles.

$$\mathrm{Zn(s)} \rightarrow \mathrm{Zn^{2+}(aq)} + 2\mathrm{e^-}$$

Oxidation (anode −)

$$2\mathrm{MnO_2(s)} + \mathrm{H_2O} + 2e^- \rightarrow \mathrm{Mn_2O_3(s)} + 2\mathrm{OH^-}$$

Reduction (cathode +)

$$\mathrm{NH_4^+(aq)} + \mathrm{OH^-(aq)} \rightarrow \mathrm{NH_3(g)} + \mathrm{H_2O}$$

$$\mathrm{Zn^{2+}(aq)} + 4\mathrm{NH_3(aq)} \rightarrow \mathrm{[Zn(NH_3)_4]^{2+}(aq)}$$

Acid/base absorbs the OH^- ions
Complex ion formation absorbs NH_3 gas

$$\mathrm{Zn(s)} + 2\mathrm{MnO_2(s)} \rightarrow \mathrm{ZnMn_2O_4(s)}$$

Overall cell reaction

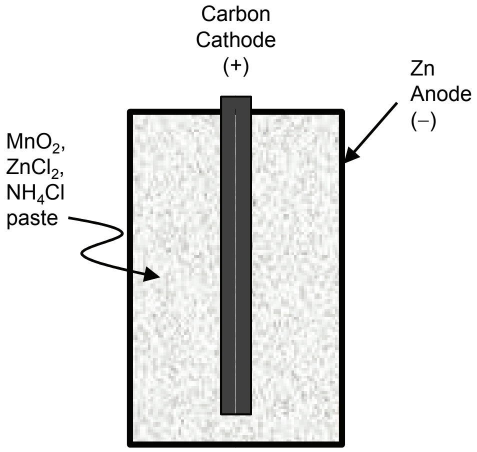

Modern **alkaline cells** use an alkaline electrolyte of KOH instead of the MnO_2, $ZnCl_2$, NH_4Cl paste of the zinc-carbon type of dry cell. In an alkaline cell, the half-reactions are

$$\mathrm{Zn(s)} + 2\mathrm{OH^-(aq)} \rightarrow \mathrm{ZnO(s)} + \mathrm{H_2O(l)} + 2\mathrm{e^-}$$

$$2\mathrm{MnO_2(s)} + \mathrm{H_2O(l)} + 2\mathrm{e^-} \rightarrow \mathrm{Mn_2O_3(s)} + 2\mathrm{OH^-(aq)}$$

10.16 Corrosion

Corrosion is the undesirable oxidation of a metal. The most well-known example is the **rusting of iron**. In simple terms, the complex redox reaction for the rusting of iron is written as

$$4\mathrm{Fe}(s)+3\mathrm{O}_2(g)\rightarrow 2\mathrm{Fe}_2\mathrm{O}_3(s)$$

In practice, the rusting of iron requires both air (O_2) and water, and the reaction proceeds as a series of steps.

$\mathrm{Fe}(s)\rightarrow \mathrm{Fe}_2^{2+}+2e^-$ $\quad$ $\mathrm{O}_2+2\mathrm{H}_2\mathrm{O}+4e^-\rightarrow 4\mathrm{OH}^-$

Oxidation half-reaction $\quad$ Reduction half-reaction

The OH^- ions react with the Fe^{2+} ions:

$$\mathrm{Fe}^{2+}+2\mathrm{OH}^-\rightarrow \mathrm{Fe}(\mathrm{OH})_2$$

then $\quad 4\mathrm{Fe}(\mathrm{OH})_2+\mathrm{O}_2+2\mathrm{H}_2\mathrm{O}\rightarrow \mathrm{Fe}(\mathrm{OH})_3(s)$

and finally $\quad \mathrm{Fe}(\mathrm{OH})_3(s)\rightarrow \mathrm{Fe}_2\mathrm{O}_3,\mathrm{H}_2\mathrm{O}+2\mathrm{H}_2\mathrm{O}$

The brown coloured metal we usually associate with rust is the compound $\mathrm{Fe}_2\mathrm{O}_3,\mathrm{H}_2\mathrm{O}$, an Fe(III) hydrated oxide

Iron and steel often rust badly because the rust has a lower density than the parent metal and tends to flake off, thus exposing more of the material to water and oxygen and promoting further rusting.

Water and oxygen are typically excluded by painting the iron article. However, a very effective way to reduce rusting of iron is to attach a more reactive metal (i.e. one with a lower E^o so that the iron undergoes reduction and the **sacrificial anode** (usually zinc) undergoes oxidation. This process is often called **cathodic protection**.

$\mathrm{Fe}/\mathrm{Fe}_2^{2+}$ $\quad$ $\mathrm{Zn}^{2+}/\mathrm{Zn}$

$E^o=-0.44\mathrm{V}$ $\quad$ $E^o=-0.76\mathrm{V}$

Iron or steel coated with zinc is said to be **galvanised** (since it acts like a galvanic cell).

Water and oxygen are typically excluded by painting the iron article. However, a very effective way to reduce rusting of iron is to attach a more reactive metal (i.e. one with a lower E^o so that the iron undergoes reduction and the sacrificial metal (usually zinc) undergoes oxidation.

Aluminium undergoes corrosion in the presence of atmospheric oxygen. A very thin (nm) tough layer of Al_2O_3 (**alumina**) appears within nanoseconds on the metallic surface which shields the parent metal from further corrosion. Alumina is a very hard ceramic material and is a non-conductor of electricity.

10.17 Electrolysis

An **electrolytic cell** is not a spontaneous redox reaction, and requires an electrical energy input to proceed.

Consider the redox reaction:

$$\mathrm{Cu^{2+}(aq) + 2Ag(s) \rightarrow Cu(s) + 2Ag^{2+}(aq)}$$

Arranged in this way, the half-reactions are:

$\mathrm{2Ag(s) \rightarrow 2Ag^{2+}(aq) + 2e^-}$	$\mathrm{Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)}$
Oxidation half-reaction	Reduction half-reaction
-0.799V	+0.337V

The total potential is −0.462V and so this is not a spontaneous reaction. If a cell were constructed and connected with a wire between the two electrodes, the reaction would proceed in the reverse to that as shown above. However, if we apply an external **potential difference** to the electrodes, say a 1V potential, then

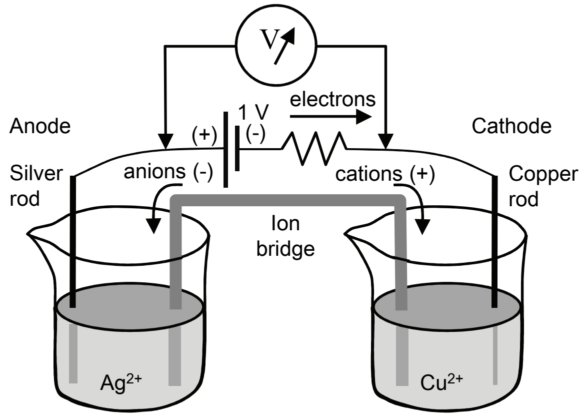

$\mathrm{2Ag(s) \rightarrow 2Ag^{2+}(aq) + 2e^-}$	$\mathrm{Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)}$
Oxidation half-reaction	Reduction half-reaction

Arranged in this way, we have an **electrolytic cell** in which the copper rod is the cathode (+) and the silver rod is the anode (-). The reaction proceeds as written.

Note: **oxidation** occurs at the **anode** which in this case is the silver. **Reduction** occurs at the **cathode**, in this case the copper. The silver rod dissolves away and copper particles coat the copper rod.

The applied external potential has to overcome +0.462V in order for the reaction to proceed, and so the potential difference across the electrodes would be 1 – 0.462 = 0.538V.

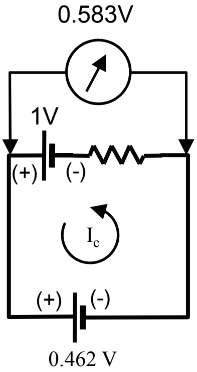

I_c indicates conventional current flow

11. Carbon Chemistry

Summary

Functional groups

H	Alkanes	Single C bond
C=C	Alkenes	Double C bond
C≡C	Alkynes	Triple C bond
OH	Alcohols	ROH
O	Ethers	ROR'
CHO	Aldehydes	RCHO
CO	Ketones	R(C=)OR
COOH	Acids	RCOOH
COO	Esters	RCOOR'
NH_2	Amines	RNH_2 R_2NH, R_3N

Alkyl groups

Methyl
Ethyl
Propyl
Butyl
Pentyl
Hexyl
Heptyl
Octyl
Nonyl
Decyl

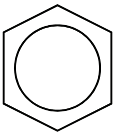

Benzene ring

11.1 Carbon

The element **carbon** has four outer valence electrons and is the first element of Group IV in the periodic table. The other elements in this group become more metallic in nature with increasing atomic number. At room temperature, carbon is relatively inert. At high temperatures, it can form covalent bonds with itself and many other elements. Carbon compounds are the basis for all life on Earth. It is interesting to note that **silicon**, the next element down in the group and also with four valence electrons, is the basis of the composition for nearly all of the Earth's minerals.

With four valence electrons, carbon has the choice of either losing or gaining four electrons to achieve a noble gas configuration but in practice, it tends to share its four electrons via covalent bonds. On the other hand, tin and lead, the heaviest elements in Group IV, with more loosely bound valence electrons, find it energetically favourable to share these electrons in the form of **metallic bonds**.

Carbon reacts with metals to form carbides: e.g. silicon carbide SiC; with non-metals to form molecular compounds: e.g. methane CH_4; with oxygen to form carbon monoxide CO and carbon dioxide CO_2; and with nitrogen to form cyanides (e.g. HCN).

Pure carbon forms three crystalline structures found naturally: **diamond**, **graphite,** and the **fullerenes**. These are **allotropes**. These have significantly different mechanical, thermal and electrical properties and differ only in the geometrical arrangement of their atoms in the crystal structures.

Diamond has a three-dimensional tetrahedral structure. Each carbon atom has four neighbours, and all the valence electrons are equally shared with no vacancies or lone electrons. Thus, diamond is a poor electrical conductor, has a high **hardness** and **elastic modulus**. Diamond is a good heat conductor due to the free passage of **phonons**.

Graphite has a two-dimensional hexagonal sheet structure with the sheets bonded together as layers by **van der Waals forces**. Each carbon atom has three neighbours. Thus, three valence electrons are involved in covalent bonding while the fourth is left in a relatively free state and is responsible for graphite being an electrical conductor.

The **fullerene** form of carbon consists of a three-dimensional arrangement of C atoms in a pentagon which together form a sphere or tube. The most famous is the C_{60} molecule, often referred to as a **buckyball**.

11.2 Carbon Compounds

Carbon forms a great number of compounds with hydrogen. These are called **hydrocarbons**. In addition, **hydrocarbon derivatives** are formed when other atoms in addition to hydrogen become involved in the bonding.

Central to this remarkable ability to form such a wide variety of compounds is the formation of chains of atoms. For example, consider the simple hydrocarbon **methane** CH_4:

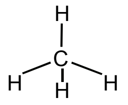

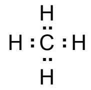

This molecule can be easily expanded to form **ethane** C_2H_6:

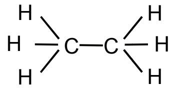

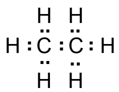

And so on for **propane**, C_3H_8, and **butane**, C_4H_{10}.

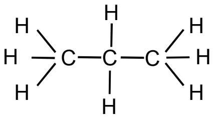

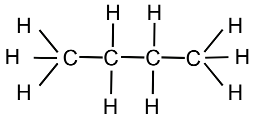

Other possibilities abound, including compounds involving a double carbon bond such as in **ethylene** and the triple bond in **acetylene**:

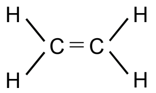

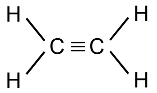

The reason for this is that saturated compounds are saturated with H atoms. When there is a double bond, the maximum possible number of H atoms that can attached to a C atom is decreased and so the C atoms is unsaturated.

Compounds with single bonds between carbon atoms are **saturated hydrocarbons**. Compounds with double or triple bonds are called **unsaturated hydrocarbons**.

Hydrocarbons with the same chemical formula but with different structures are possible. For example, **butene** C_4H_8, can exist in four different structural arrangements, each of which is called an **isomer**.

The substitution of other elements in place of one or more hydrogen atoms in carbon compounds is responsible for the existence of a large number of hydrocarbon derivatives. Alcohols are hydrocarbons with an attached OH **functional group**.

The ability of carbon to form chains, double bonds, isomers and contain functional groups is responsible for the great variety of carbon compounds which form the chemistry of life on Earth – or **organic chemistry**.

11.3 Functional Groups

Hydrocarbon derivatives are formed when other atoms in addition to hydrogen become involved in the bonding. In these cases, the core hydrocarbon is written with an attached **functional group**. The core or hydrocarbon compound (also called the **remainder** or **residue)** is written with the symbol R. A common functional group is the alcohol, or OH group.

methanol CH_3OH:

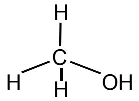

Alcohols can be written as ROH where R is the hydrocarbon residue, or the core hydrocarbon component, and OH is the alcohol functional group.

ethanol C_2H_5OH:

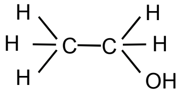

Functional groups

H	Alkanes
C=C	Alkenes
C≡C	Alkynes
OH	Alcohols
O	Ethers
CHO	Aldehydes
CO	Ketones
COOH	Acids
COO	Esters
NH_2	Amines

Functional groups determine the nature of the reactivity of the hydrocarbon involved. The chemistry of the compound is usually determined by the character of the functional group while the hydrocarbon residue is relatively inert.

Consider some compounds based upon the CH_3CH_2- residue:

Ethane: $CH_3CH_2 - H$

Ethanol: $CH_3CH_2 - OH$

Ethyl bromide: $CH_3CH_2 - Br$

The CH_3CH_2– residue is called the **ethyl group**. Similarly, the CH_3– is the **methyl group** and the $CH_3CH_2CH_2$– is the **propyl group**.

These named **alkyl groups** form bonds with various different functional groups and the resulting compounds R–X are thus **derivatives** of the parent hydrocarbon (e.g. methane, ethane, etc).

11.4 Alkanes

Alkanes are **saturated hydrocarbons** since the bonds with the carbon atoms are single bonds and they contain only carbon and hydrogen. The names of the compounds are derived from the number of carbon atoms.

The first 10 alkanes

Methane	CH_4
Ethane	C_2H_6
Propane	C_3H_8
Butane	C_4H_{10}
Pentane	C_5H_{12}
Hexane	C_6H_{14}
Heptane	C_7H_{16}
Octane	C_8H_{18}
Nonane	C_9H_{20}
Decane	$C_{10}H_{22}$

Alkanes can have chain or branched chain (**aliphatic**), or cyclic (**alicyclic**) structures. As the complexity of the molecules increases, there are increased ways in which molecules can form structures with the same molecular formula.

Compounds having the same molecular formula but different structure are called **isomers**.

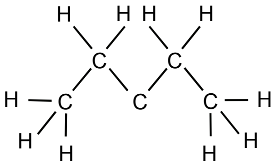

n-pentane (or normal pentane):

Alkanes are the principal components of **petroleum**. Low molecular weight compounds are generally gases at room temperature and pressure. Gasoline (petrol) and turpentine typically contain light fractions (C_7H_{16} to $C_{10}H_{22}$). Kerosene, diesel oil, lubricating oils, pitch, and so on consist of heavier compounds. Alkanes are relatively inert, but can undergo exothermic oxidation (**combustion**) which make them useful sources of energy.

Fuel	**Calorific value** (MJ/kg)
Petrol	45
Diesel	44.4
Ethanol	27.2

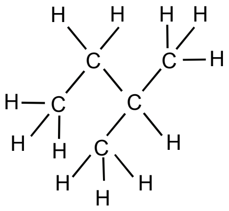

Alkanes with a chain structure have the molecular formula C_nH_{2n+2}.

iso-pentane (single branch at the end)

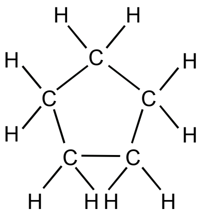

Alkanes having a cyclic or a ring structure have the formula C_nH_{2n}.

cyclo-pentane (ring structure)

Alkanes are sometimes called **paraffins** because they have little (parum) affinity for forming bonds with other atoms or molecules. When they do participate in reactions, this usually involves the **substitution** of a hydrogen atom with another atomic species.

11.5 Alkenes

Alkenes (also known as **olefins**) are unsaturated hydrocarbons because they contain a double bond between two of the available carbon atoms instead of single bonds. As a consequence, there are fewer than four atoms attached to each of the carbon atoms.

Alkenes in a chain configuration have the general formula C_nH_{2n}. Ethene or **ethylene** (C_2H_4) is the simplest alkene.

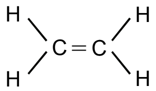

Bond length: 1.33 Å
Energy: 606 kJ/mol
ΔH = +52.5 kJ/mol

Note that the carbon double bond can occur in different places in the structure, leading to formation of isomers. In **propylene** (C_3H_6):

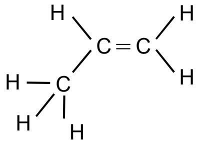

Ethene, propene, and butene are gases. Higher alkenes are generally liquids and those with greater than sixteen carbon atoms are generally waxy solids.

While alkenes are relatively inert, they are more reactive than alkanes because of the double bond. The carbon double bond has a higher bond strength compared to the single bond, but the double bond allows a greater potential for attachment for other atoms and this leads to greater reactivity compared to single-bonded structures. Chemical reactions involving alkenes are usually associated with the breaking of the double bond and the addition of other atoms.

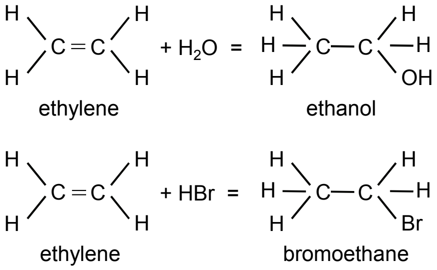

An important reaction involving alkenes is that of **polymerisation**, where a chain reaction occurs in which the C=C double bonds are successively broken and employed to form larger and larger single-bonded structures.

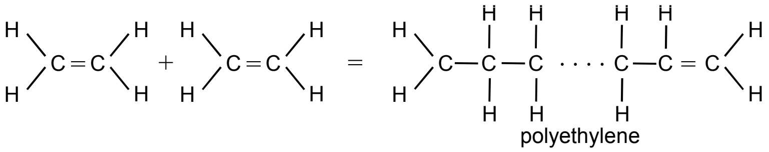

11.6 Alkynes

Alkynes are similar to alkenes except that they feature a carbon-carbon triple bond. Alkynes are unsaturated hydrocarbons and have the general formula C_nH_{2n-2}. Alkynes are also known as **acetylenes**, but the common use of this term refers to **ethyne**: C_2H_2.

C_2H_2
acetylene

$H—C\equiv C—H$

Bond length: 1.2 Å
Bond energy: 828 kJ/mol
ΔH = +226.9 kJ/mol

The triple bond can occur anywhere in the chain in higher alhynes. Consider the compound 1-butyne. The 1- indicates that the triple bond is at the end of the chain:

$CH_3CH_2C\equiv CH$

The CH_3CH_2 is the ethyl group, and so this compound can also be named **ethylacetylene**.

Ethyne	$HC\equiv CH$
Propyne	$CH_3C\equiv CH$
1-Butyne	$CH_3CH_2C\equiv CH$
1-Pentyne	$CH_3(CH_2)_2C\equiv CH$
1-Hexyne	$CH_3(CH_2)_3C\equiv CH$
1-Heptyne	$CH_3(CH_2)_4C\equiv CH$
1-Octyne	$CH_3(CH_2)_5C\equiv CH$
1-Nonyne	$CH_3(CH_2)_6C\equiv CH$
1-Decyne	$CH_3(CH_2)_7C\equiv CH$

Alkynes have a similar chemistry to alkenes. The C≡CH unit is a linear structure and in acetylene, the molecules are therefore rod-like in shape.

The triple bond is capable of opening up and accommodating additional atoms during reactions. Additions may occur with halides, halogens, water, and ammonia.

The C≡CH unit, located at the end of the structure, imparts an acidic nature to alkynes and so acid-base reactions are also possible. Alkynes can also undergo **polymerisation** to give **polyacetylenes**.

One of the most common uses of acetylene is in gas welding. The combustion of acetylene in oxygen provides a flame temperature of about 3300°C making it one of the hottest flames available in industry.

$$2C_2H_2 + 5O_2 \rightarrow 4CO_2 + 2H_2O$$

Acetylene is a gas, but an unstable one, especially at moderate to high pressures. In welding gas bottle sets, acetylene is usually dissolved in liquid **acetone** or dimethylformamide and stored under pressure in a cylinder containing a porous medium, **agamassam**.

11.7 Benzene

An important class of carbon compounds arises when carbon atoms are arranged in a special ring formation.

The **benzene ring** C_6H_6 has all the atoms in a plane, with a 120° angle between the carbon atoms. The initial difficulty with this structure is that each carbon atoms has three neighbours sharing one electron, but there are four valence electrons in carbon. Evidently, there either has to be three double bonds present or some way of accommodating the fourth valence electron. If each second carbon atom had a double bond, then all available valence electrons would be shared, but experimental evidence indicates that the bond strength for the C-C bonds are equal around the ring, and that it is stronger than a single bond, but not as strong as a double bond.

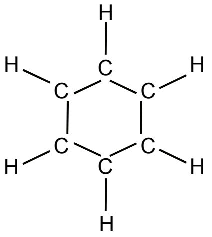

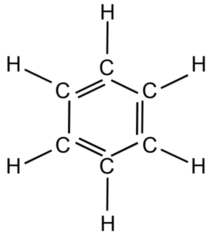

The bonding in a benzene ring is explained by the allowable movement of the **valence electrons** in the ring structure. Rather than forming single or double bonds, electrons in the ring form a resonance hybrid, or a superposition of single and double bonds, whereby they may have movement throughout the ring, much like valence electrons in a metal have movement throughout the **conduction band**.

To signify this, the benzene ring is often written in shorthand form using one of the following diagrams:

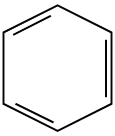

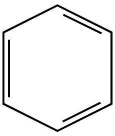

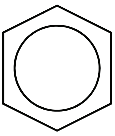

It should be always remembered that there are not single or double bonds, but a special bond somewhere between the two based upon electron movement around the ring.

Ring compounds of this type are called **aromatic compounds**. When they consist of C and H only, they are called **aromatic hydrocarbons**. Note: **aliphatic** compounds in a ring structure are not aromatic; they are simple ring structures (**alicyclic**). Aromatic compounds have mobile electrons within the ring and it is these that determine the chemistry of these compounds. In contrast to the addition reactions for alkenes with a chain structure, benzene undergoes substitution reactions in which H atoms are replaced by other atoms.

11.8 Alcohols

Hydrocarbon derivatives arise when functional groups combine with a residual hydrocarbon core. **Alcohols** are those compounds which have one or more OH (**hydroxyl**) groups attached, ROH.

Three types of alcohol are to be found depending on the number of C atoms attached to the C atom with the OH group. Primary (n) alcohols are written RCH_2OH.

$CH_3CH_2CH_2 — OH$ 1-propyl alcohol
n-propanol

Secondary (s) alcohols are written RR'CHOH:

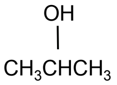

$CH_3\overset{\overset{OH}{|}}{C}HCH_3$

isopropyl alcohol
2-propanol
s-propanol

Tertiary (t) alcohols are RR'R"COH.

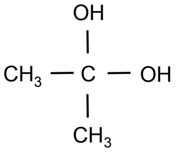

$CH_3-\underset{\underset{CH_3}{|}}{\overset{\overset{OH}{|}}{C}}-OH$

t-butyl alcohol

Lower alcohols are soluble in water due to the presence of the polar OH group. In higher alcohols, the longer the carbon chain, the less water-like the compound is and behaves more like a hydrocarbon – volatile and relatively inert. **Pentanol** and higher compounds are insoluble in water.

Alcohols can react with other atoms and can act as weak **bases** (due to the unshared electron pairs on the O atom) or weak **acids** – where the H atom on the OH group can act as a proton donor.

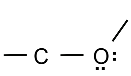

Phenol (hydroxybenzene) has the OH group attached to a benzene ring and forms the basis of synthesis for many other compounds.

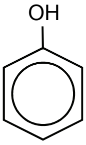

Ethanol (C_2H_5OH) is "alcohol" to the common person (e.g. in alcoholic drinks). **Denatured ethanol** is the usual form used for cleaning and other uses. Additives render it unpalatable and, in some cases poisonous (such as when methanol is added to the ethanol to give **methylated spirits**).

Methanol (CH_3OH) is highly toxic and may cause blindness or death when only a few millilitres are ingested.

A common cleaning alcohol is **isopropyl alcohol**, or 2-propanol $(CH_3)_2CHOH$.

Ethylene glycol $C_2H_4(OH)_2$ is commonly used as an anti-freeze/ anti-boil additive in motor vehicle cooling systems. In its pure form it has a boiling point of 197.3 °C but has a lower heat capacity than water and so is used diluted. It does this by depressing the hydrogen bonding between water molecules so that, upon freezing, the temperature must be significantly lowered before ice crystals can form.

An interesting compound prepared from phenol, sodium hydroxide, and carbon dioxide is **aspirin**.

11.9 Ethers

Ethers have the general formula ROR' where R is an alkyl group. They can occur as aliphatic (chained), alicyclic, or aromatic (**benzene ring**) formations, or a combination. They are relatively unreactive hydrocarbon derivatives.

The lone pairs of electrons on the O atom allow ethers to form **hydrogen bonds** with water molecules and so they are generally partially soluble in water.

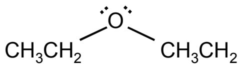

diethyl ether

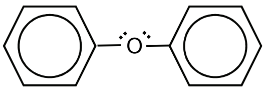

diphenyl ether

However, they cannot form hydrogen bonds between themselves and so generally have a lower boiling point than comparable alcohols.

In some respects, ethers are similar to alcohols in that they can be considered derivatives of water. In an alcohol, the R takes the place of one of the H atoms in the water molecule. In an ether, both of the H atoms are replaced by alkyl groups R and R'. Ethers are generally isomers of the corresponding alcohols. For example, ethanol, C_2H_5OH, has the same formula as dimethylether, but of course has quite different chemical and physical properties.

Perhaps one of the most well known ethers is **diethyl ether**, $CH_3CH_2OCH_2CH_3$, a low boiling point colourless liquid used as an anaesthetic. Diethyl ether can be formed from the removal of water from ethanol.

Dimethyl ether, CH_3OCH_3, (or **DME**) is a gas at room temperature is often used as an aerosol propellant since it can be stored as a liquid under pressure in equilibrium with the gas phase, evaporates readily and leaves the desired product as an aerosol mist for the application.

Methyl phenyl ether $CH_3OC_6H_5$ is found in aniseed oil.

Ethers are generally unreactive substances and can withstand attack from either acids or bases, but are more reactive than comparable alkanes.

Simple ethers have the alkyl groups on each end of the oxygen atom and if the groups are the same, are thus named dialkyl ether (such as in dimethylether). If one of the groups is a more complex group, then the structure of the name is alk-oxy-alkane. For example, $CH_3CH_2CH_2OCH_3$ is **methoxypropane**.

11.10 Aldehydes

Alcohols can be thought of as a step towards complete oxidation of the corresponding alkane. Primary alcohols can also undergo oxidation to form **aldehydes** and **carboxylic acids**, and secondary alcohols can oxidise to form **ketones**.

Aldehydes have a carbon-to-oxygen double bond (which itself is called the **carbonyl group**) with the general formula RCHO. They are usually prepared by **oxidation** of the corresponding alcohol.

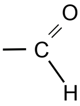
aldehyde group

Perhaps one of the most well known aldehyde is that made from **formaldehyde** dissolved in water to give **formalin**, a preservative fluid used for embalming biological tissues.

Consider the reaction between **methanol** and potassium dichromate:

$$3CH_3OH + Cr_2O_7^{2-} + 8H^+ \rightarrow 3CH_2O + 2Cr^{3+} + 7H_2O$$

In this equation, the Cr_2 ion has been reduced to the Cr^{3+} ion, and the methanol has lost two electrons (and two protons H^+) and hence has been oxidised to give CH_2O, formaldehyde.

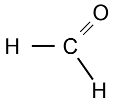
methanal (formaldehyde)

Methanal	HCHO	Formaldehyde
Ethanal	CH_3CHO	Acetaldehyde
Propanal	CH_3CH_2CHO	Propionaldehyde
Butanal	$CH_3CH_2CH_2CHO$	Butyraldehyde

Aldehydes often have a pleasant odour or flavour, such as in the benzene derivative vanillin:

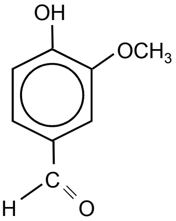

and cinnamaldehyde, the active ingredient in cinnamon.

To avoid confusion with alcohols (OH), the carbonyl group with an attached H is written CHO and not COH.

Cyclic forms also exist, such as **benzaldehyde**, an aromatic aldehyde:

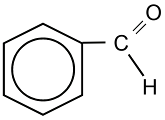

Aldehydes are closely related to **ketones**, but have the **carbonyl group** C=O at the end of a carbon chain rather than in between two C atoms. Lower aldehydes are soluble in water. Aldehydes can be further oxidised to form **carboxylic acids**.

11.11 Ketones

Ketones have a **carbonyl group** C=O attached to two other C atoms that form part of the same or different **alkyl groups**. (In an aldehyde, one of the other atoms is an H atom.) The general formula is RC(=O)R'.

The oxygen atom is more **electronegative** than attached carbon (i.e. the electron pairs that constitute the double bond are closer to the oxygen atom than the carbon atom), making this group polar. Unlike aldehydes, ketones do not have an H atom attached to the carbonyl group.

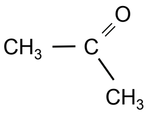

2-propanone
(acetone)

Acetone can be formed by the oxidation of 2-propanol according to

$$3CH_3CHOHCH_3 + Cr_2O_7^{2-} + 8H^+ \rightarrow$$

$$3CH_3COCH_3 + 2Cr^{3+} + 7H_2O$$

Aldehydes and **ketones** have many similarities and both are formed from the oxidation of an alcohol.

Ketones cannot be further oxidised to an acid since there is no H atom available on the functional group (as in the case of **aldehydes**). The absence of an H atom bonded to an oxygen atom means that both aldehydes and ketones do not form **hydrogen bonds** with each other and so generally have a lower boiling point than comparable alcohols. However, the lone pairs of electrons on the oxygen atom can form a hydrogen bond with an H atom in a water molecule and so lower ketones (and also aldehydes) are generally soluble in water.

The names of ketones (and aldehydes) are formed by the number of C atoms in the longest chain, and include the C atom in the carbonyl group.

Acetone CH_3COCH_3 is widely used as a relatively non-toxic solvent for various carbon compounds (e.g. paint, plastics, nail varnish, superglue, etc). It also has the useful property of being soluble in water and is commonly used as a drying agent for glassware in chemistry laboratories. It is very volatile and dries without leaving a residue.

Another common ketone used as a solvent for paints, varnish and gums is 2-butanone, or **methyl ethyl ketone** (**MEK**), or methyl acetone, $CH_3COCH_2CH_3$, a flammable clear liquid that is soluble in water. It has a characteristic sweet odour. As well as industrial uses as a solvent, it is also used as a polystyrene cement in plastic model kits.

Another significant ketone, with a cyclic (but not aromatic) structure, is **cyclohexanone** $(CH_2)_5CO$ which is used in the manufacture of **nylon**.

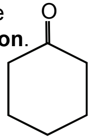

cyclohexanone

Ketones have significant biochemical functions and are produced when body fat is used as a source of energy when there is a shortage of **glucose** (such as during fasting).

11.12 Carboxylic Acids

Hydrocarbon derivatives that contain both the carbonyl (C=O) and hydroxyl (OH) functional groups are called **carboxylic acids** with the general formula RCOOH. When R is an aliphatic residue, the compound is called a **fatty acid**. When R is an aromatic ring, the acid is a **benzene ring derivative**.

The COOH group is called the **carboxyl group**. Carboxylic acids can be prepared from the oxidation of the corresponding alcohol.

$$5CH_3OH + 4MnO_4^- + 12H^+ \rightarrow$$
$$5HCOOH + 4Mn^{2+} + 11H_2O$$

In this reaction CH_3OH and H_2O produces HCOOH and four protons and four electrons. Since there is a loss of four electrons (compared to two for the oxidation of methanol to methanal, or formaldehyde) , formic acid is a more highly oxidised compound than **formaldehyde**.

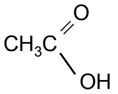

ethanoic acid (acetic acid)

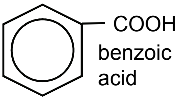

benzoic acid

A very common aliphatic acid is **acetic acid**, written CH_3COOH and is found in vinegar. Formic acid (from the latin formica for ant) in the found in venom of ants. Stearic acid is found in soaps, waxes and plant oils. Other carboxylic acids can be found in coconuts, chocolate,

Some carboxylic acids:

Formic	Methanoic	HCOOH
Acetic	Ethanoic	CH_3COOH
Propionic	Propionic	CH_3CH_2COOH
Butyric	Butanoic	CH_32CH_2COOH
Stearic	Octa decanoic	CH_316CH_2COOH

Similarly, ethanol can be oxidised to form **acetic acid**. Acetic acid can also be formed from the oxidation of acetaldehyde. Further oxidation of these compounds can be obtained via the rather energetic process of **combustion**, in which case the product is CO_2 and water.

Carboxylic acids are weak acids and dissociate only slightly, giving H^+ ions in water, and act as proton donors in reactions. The H^+ ions come from the COOH functional group. Lower carboxylic acids are soluble in water.

A common chemical demonstration reaction involves that between vinegar and **baking soda** to give sodium ethanoate, carbon dioxide and water:

$$CH_3COOH + NaHCO_3 \rightarrow 3CH_3COONa + CO_2 + 7H_2O$$

There is an ionic bond between the CH_3COO^- and Na^+ ions. By convention, the ionic compound is written with the metal ion last.

Carboxylic acids react with alcohols to form fats and oils (esters). **Fatty acids** are those needed by the body and are obtained from digestion of animal and vegetable oils and fats.

11.13 Esters

Carboxylic acids contain the functional group COOH. The H atom can be easily given up (proton donor): hence the acid classification.

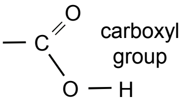

Derivatives of carboxylic acids are formed when the OH group is replaced by something else. When the OH group is replaced by OR, then an **ester** is formed. Esters are generally colourless liquids insoluble in water and have the general formula RCOOR'. For example, a reaction between methanol and acetic acid forms the ester **methyl acetate**:

$$CH_3OH + CH_3COOH \rightarrow CH_3COOCH_3 + H_2O$$

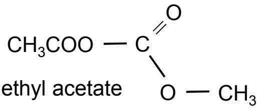

methyl acetate

Note that the methyl group is the CH_3 at the end of the molecule. The "acetate" is from the carboxylic acid. Esters are written with the functional group first.

In this reaction, water is formed from the H atom in the incoming alcohol, and the OH group from the carboxylic acid, leaving behind an ester.

Esters are commonly used in food flavourings and perfumes due to their pleasant odour.

Methyl butyrate	Apple	$CH_3CH_2CH_2COOCH_3$
Ethyl butyrate	Strawberry	$CH_3CH_2CH_2COOCH_2CH_3$
Pentyl butyrate	Apricot	$CH_3CH_2CH_2COO(CH_2)_4CH_3$
Pentyl acetate	Banana	$CH_3COO(CH_2)_4CH_3$
Isoamyl acetate	Pear	$CH_3COOCH_2CH_2\,CH(CH_3)_2$
Octyl acetate	Orange	$CH_3COO(CH_2)_7CH_3$

Most naturally occurring fats and oils are very large esters made from organic acids and alcohols (usually **glycerol**). When there are no C=C bonds, the acid is said to be **saturated**. When the acid contains C=C bonds, it is **unsaturated**. When an acid has one C=C bond it is **monounsaturated**, and when there is more than one C=C bond, it is referred to as being **polyunsaturated**. These **fatty acids** form saturated, unsaturated and polyunsaturated fats and oils (oils are esters that are liquid at room temperature whereas waxes are solid).

Esters also undergo polymerisation to form common products such as **polyesters** used for clothing and recyclable polyethylene terephthalate (**PET**) drink bottles. An interesting example of an ester is nitroglycerine (which is somewhat incorrectly named).

11.14 Amides

Amides are derivatives of carboxylic acids in a similar manner to esters. When the OH group is replaced by NH_2, then an **amide** is formed. A special class of amides occurs when the OH group is replaced by NHR. Amides can be produced by the reaction between ammonia and an ester.

The reaction between methyl acetate and ammonia produces **ethanamide** (acetamide) and methanol:

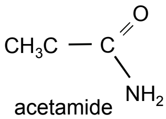

acetamide

$$CH_3COOCH_3 + NH_3 \rightarrow$$
$$CH_3CONH_2 + CH_3OH$$

In the NH_2 group, the electrons are more strongly drawn towards the more electronegative N atoms, leaving the H end with a positive charge and so amides are able to form **hydrogen bonds** with each other as they are attracted to the negatively charged lone pairs on neighbouring O atoms. This confers a relatively high melting point for these compounds. Amides have much the same solubility as comparable esters, with lower amides being soluble in water.

The most notable characteristic of amides are their mechanical strength and are often used, both in nature and industrially, as structural items.

In nature, amides for the basis of links between amino acids in **proteins**.

In industrial use, amides are joined together in long chains to produce polyamides, the most well known of which are **nylon** and **kevlar**.

Lysergic acid diethylamide (**LSD**) is also an amide.

Methanamide	$HCONH_2$
Ethanamide	CH_3CONH_2
Propanamide	$CH_3CH_2CONH_2$

The **amide link** joins amino acids together in biological processes to form **proteins**. **Amino acids** are carboxylic acids with an amine group of the form $H_2NCHRCOOH$. During polymerisation, water is eliminated between two acid groups to be joined together.

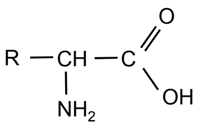

amino acid

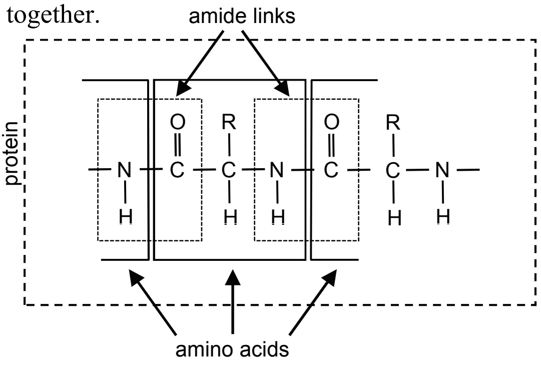

11.15 Amines

In an alcohol, the R takes the place of one of the H atoms in the water molecule to give the general formula ROH. Now consider the molecule **ammonia** NH_3.

If one or more H atoms are replaced by an alkyl group, then the resulting compound is called an **amine**. Primary amines RNH_2 have one R substitution, secondary amines R_2NH have two, and tertiary amines R_3N have all three H atoms replaced with alkyl groups. Aliphatic and aromatic structures are found.

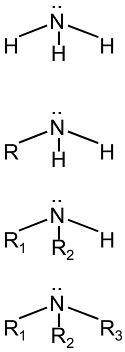

Ammonia	NH_3
Methylamine	CH_3NH_2
Ethylamine	$CH_3CH_2NH_2$
Propylamine	$CH_3CH_2CH_2NH_2$
Butylamine	$CH_3CH_2CH_2CH_2NH_2$
Aniline	$C_6H_5NH_2$

Just as in the case of ammonia, amines are bases due to the lone electron pair on the N atom. Primary and secondary amines can form hydrogen bonds with each other and all can form **hydrogen bonds** with water, thus making them soluble, the lower gaseous amines being very soluble.

Phenylamine, or **aniline**, is a primary amine where an H atom on the ammonia is replaced with a benzene ring. This has the interesting consequence of the lone pair of electrons on the N atom interacting with, and to some extent being absorbed into the ring, which in turn decreases its solubility and raises its boiling point compared to comparable aromatic compounds.

Ammonia is a colourless gas with a very sharp odour. Ammonium hydroxide in household cleaning products is a solution of ammonia gas in water where it slightly dissociates into the ammonium NH^{4+} and hydroxide OH^- ions.

Many primary aromatic amines are used as dyes.

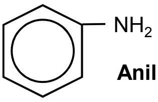

Aniline

Amines are important in biological functions and are produced from the breakdown of amino acids in the body. **Amino acids** are carboxylic acids with an amine group of the form $H_2NCHRCOOH$. Amines serve as neurotranmitters. A common natural amine is **histamine** which triggers an immune response. Anti-histamines suppress this activity in severe allergic reactions.

Methylphenethylamine (or **amphetamine**) is a powerful drug that modifies the action of natural neurotransmitters in the brain.

11.16 Polymers

The chemical and physical properties of hydrocarbon compounds and derivatives depend very much on the size of the molecule. Smaller groups such as methyl and ethyl compounds generally have a high solubility, a low boiling point. When these small molecular units, collectively called **monomers**, and joined together into long chains, the molecular size can increase greatly with the production of a variety of useful substances which are called **polymers**, or **plastics**.

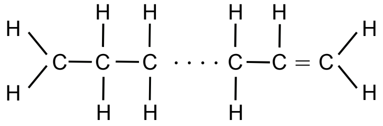
polyethylene

For example, in ethylene, the C=C bonds may be successively broken and joined to form larger and larger molecules of polyethylene. This process is usually performed with the help of a catalyst. The resulting compounds are called **addition polymers**. When other atoms or functional groups are involved, polymers of a great variety of properties are possible.

Another class of polymers are **condensation polymers**. In these, chains are formed by the expulsion of a water molecule from between the monomers to be joined. In **Nylon**, amide links are used to join monomers into large chain molecules.

One of the monomers used to make **Nylon** 6,6 is adipic (hexanedioic) acid:

$$HOOC—CH_2—CH_2—CH_2—CH_2—COOH$$

This is used with another monomer, 1,6 diaminehexane:

$$H_2N—CH_2—CH_2—CH_2—CH_2—CH_2—CH_2—NH_2$$

When these two monomers are chained together, water drops out as shown:

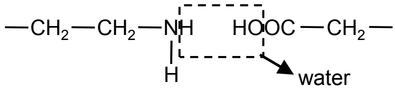

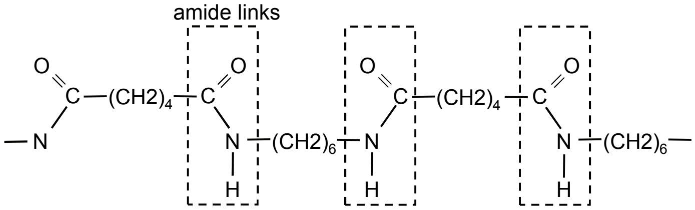

During polymerisation, these molecules continue to chain together, expelling a water molecule each time and forming amide linkages. **Kevlar** is a similar polymer but the amide links join benzene rings together.

11.17 Physical and Chemical Properties

Group	Boiling point	Solubility	Notes
Alkanes	Lower compounds are gases at room temp. Boiling point increases as size of molecule increases.	Do not form H bonds, insoluble in water.	Mainly unreactive with other compounds.
Alkenes	Similar to alkenes.	Similar to alkenes.	Slightly more reactive due to presence of C=C.
Alkynes			More reactive than alkenes due to C≡C.
Alcohols	H bonding causes boiling point to be higher than comparable alkanes.	Lower alcohols are soluble in water. Compounds become less water-like with increasing size.	OH group makes the molecules polar, and able to form H bonds.
Ethers	Cannot form H bonds between themselves so lower boiling point than comparable alcohol.	Can form H bonds with water molecules. Lower ethers soluble in water.	More reactive than comparable alkanes.
Aldehydes	Cannot form H bonds with each other and so generally have a lower boiling point than comparable alcohols.	Lower aldehydes soluble in water due to presence of lone pair electrons on O atom.	
Ketones	Similar to aldehydes.	Similar to aldehydes.	
Carboxylic acids	High boiling point due to H bonding.	Lower acids soluble.	Typically weak acids.
Esters	Lower boiling point due to lack of self-H bonds.	H bond with water confer solubility on lower esters.	
Amines	Reasonably high boiling point but lower than alcohols.	Lower amines soluble.	

12. Biochemistry

Summary

Glucose

$C_6H_{12}O_6$

Glycerol

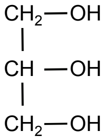

$$\begin{array}{l} CH_2-OH \\ \quad | \\ CH-OH \\ \quad | \\ CH_2-OH \end{array}$$

Amino acid

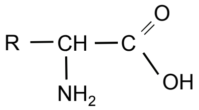

Base-pairing rules

A-T

G-C

Nucleotide

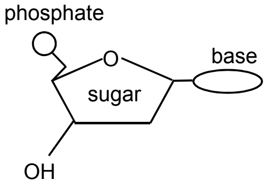

Anaerobic metabolism

$$C_6H_{12}O_6 + 2NAD^+ + 4ADP \rightarrow 2C_3H_4O_3 + 2NADH + 2H^+ + 4ATP$$

Aerobic metabolism

$$C_6H_{12}O_6 + 6O_2 + 36ADP + 36P \rightarrow 6CO_2 + 6H_2O + 36ATP$$

Photosynthesis

$$6CO_2 + 12H_2O + \text{energy} \rightarrow C_6H_{12}O_6 + 6O_2 + 6H_2O$$

$NAD^+ + 2H^+ + 2e^- \rightarrow NADH + H^+$ reduction

$NADH + H^+ \rightarrow NAD^+ + 2H^+ + 2e^-$ oxidation

12.1 Sugars

Sugars are part of a larger group of compounds called **carbohydrates**. Carbohydrates have the general formula $C_m(H_2O)_n$ and can be thought of as hydrated carbon compounds.

Glucose, $C_6H_{12}O_6$, an important sugar in biochemistry is an aldehyde .

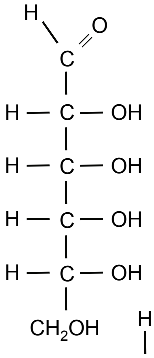

Fructose, an isomer of glucose, $C_6H_{12}O_6$, and is a ketone.

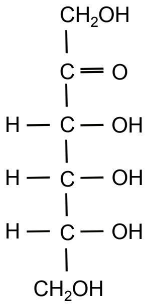

These compounds can readily form cyclic structures, which are often more common than the chain structures.

Glucose and fructose are examples of **monosaccharides**. **Sucrose**, or table sugar, is a **disaccharide** and is formed by the condensation (with the elimination of one water molecule) of one molecule of fructose and one of glucose. The reverse reaction, hydrolysis of sucrose, splits sucrose into glucose and fructose.

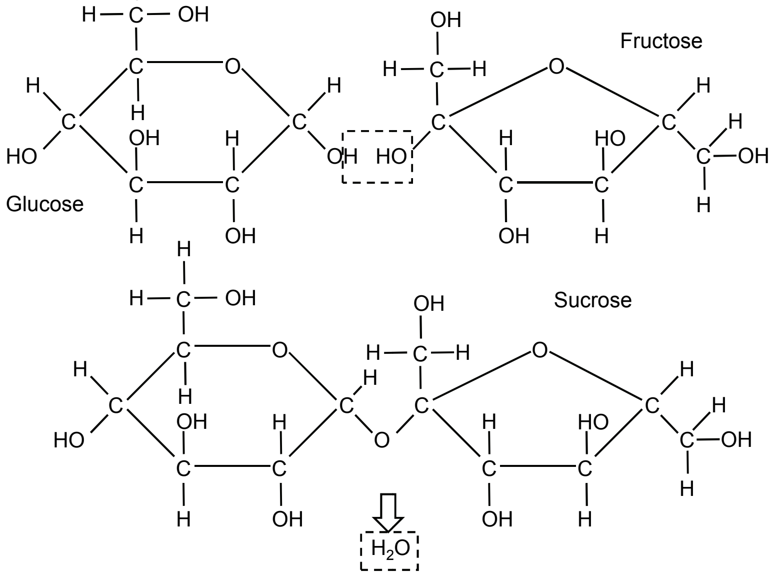

Glucose is the basic fuel for the process of **metabolism**. Glucose is recovered from fats and sugars in the diet and stored as **glycogen** in the body until converted back to glucose for use directly by cells.

12.2 Polysaccharides

Cellulose is a straight-chain polysaccharide $(C_6H_{10}O_5)_n$ consisting of thousands of glucose units. It is insoluble in water. The nature of particular forms of cellulose in organisms depends mainly upon the chain length of the molecule.

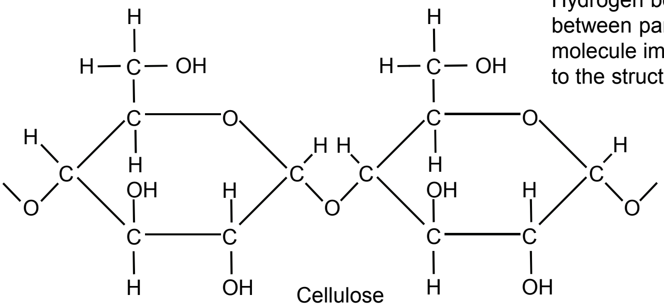

Cellulose

Hydrogen bonding between parts of the molecule impart strength to the structure.

Starch is similar to cellulose in that it is constructed from glucose monosaccharide units, but has a different orientation of linkage between the glucose elements, with more branching than cellulose. It is generally a mixture of glucose polymers of varying lengths, the shorter constituents (of molecular weight ≈4000) being water soluble.

Starch typically contains a mixture of linear (soluble) amylose and branched (relatively insoluble) amylopectin molecules. Complete hydrolysis gives glucose. Partial hydrolysis results in various starch sugars called dextrins.

Starch is broken down into glucose in the body by enzymes, (chiefly amylase, which is a constituent of saliva and pancreatic juice).

Glycogen is similar to the amylopectin form of starch in structure and is found in the liver and muscles of animals as an energy store. It is synthesised from glucose from digestion by enzymes in the liver where it is stored until released as **glucose** into the bloodstream as needed.

Cellulose acts as the main structural element of plants and forms the cell wall that gives plants their rigidity. Wood is about 50% cellulose. Mammals generally are incapable of breaking down cellulose to glucose although cows and sheep, and other ruminants, have the necessary bacteria in their digestive system that perform this function. Insects (e.g. termites) can also digest cellulose.

Starch is used as an energy store in plants. The glucose produced by photosynthesis is stored in plant tissue as starch until it is needed by the plant. Starch in plants, when eaten as food, is an important source of glucose for animals.

12.3 Lipids

Fats and oils are **esters**. Esters are derivatives of **carboxylic acids** where the H on the OH group is replaced by R so that the general formula is RCOOR'. Esters are usually derived from the reaction between acids and alcohols with the elimination of a water molecule.

Animal and plant fats are triple esters, or **triglycerides**, of **glycerol** (**glycerin**) $C_3H_8O_3$. Each of the three available OH groups in glycerol is available for combination with three molecules of **fatty acid**. Fatty acids typically have from 12 to 20 carbon atoms.

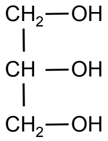

$$\begin{array}{l} CH_2-OH \\ | \\ CH\ -OH \\ | \\ CH_2-OH \end{array}$$

Glycerol
(propane 1,2,3 triol)

Fats, which are solid at room temperature, are made from **saturated fatty acids**, while **oils**, which are liquid at room temperature, are formed mainly from **unsaturated fatty acids**. Fats and oils from plants and animals typically contain a mixture of different types of esters, two or three fatty acids attached to a glycerol molecule.

Linolenic acid has the general formula

$$CH_3(CH_2)_4\underset{\uparrow}{C}H=CHCH_2CH=CH(CH_2)_7COOH$$

and is called an omega-6 fatty acid because the first C=C double bond is on the sixth carbon atom from the (left-hand) end. **Linoleic acid** is am omega-3 acid. (Note: normally numbering of atom positions is from the right.)

$$CH_3CH_2\underset{\uparrow}{C}H=CHCH_2CH=CHCH_2CH_2=CH(CH_2)_7COOH$$

Linoleic acid

Fats, along with oils and waxes, are one example of a broader group of **lipids**. Compared to carbohydrates, lipids generally contain a smaller proportion of O atoms. Lipids are relatively insoluble in water.

Fat in the body acts as a storage place for glucose. When fat reserves are called upon, the fats are converted back into glycerol and **fatty acids**. The glycerol in turn is converted into glucose by the liver.

Fats in the diet are also a source of **linolenic** $C_{17}H_{29}CO_2H$ and **linoleic** $C_{17}H_{31}CO_2H$ unsaturated fatty acids which are essential in animal dietary intake because they are unable to be synthesised by the body directly and are required for metabolism into a variety of other acids required by various bodily functions.

Fats also serve a structural role in the body by providing heat insulation and a barrier against mechanical shock for organs.

When fats hydrolyse, or react with water, in an alkaline solution, glycerol is formed along with the metal salt of the carboxylic acid. Such a reaction is called **saponification**, from which soaps (e.g. sodium stearate) is formed.

12.4 Proteins

Proteins are amino acids joined together by **amide** (or **peptide**) links and are classified as **polyamides**. The arrangement, or sequence, of amino acids within the protein structure determines the function of the protein. A typical protein may contain 100 or more amino acids.

The mechanical structure of proteins (i.e. the sequence of amino acids and hydrogen bonds) has great significance for living organisms. Proteins form the basis of skin, hair, muscle, tendons, and other tissues in the body.

Some proteins are enzymes and act as catalysts for chemical reactions which would otherwise proceed too slowly for use in organisms.

The secondary structure of proteins is stable within narrow temperature and pH ranges. When the secondary structure of a protein is disrupted, either by heat (such as by cooking), or immersion in acids or alkalis, the protein is said to be denatured and the physical properties change markedly.

Amino acids with the amine group attached to the same C atom as the COOH group are called **alpha-amino acids** and are the most important in biochemistry.

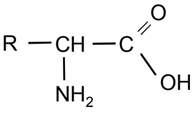

amino acid

The twenty or so amino acids found in proteins differ in the makeup of the R residual. In **glycine**, R is just an H atom, while in phenylaniline, R is a ring structure.

On one end of the chain, there is a free NH_2 group; this is the N terminal. At the other end of the chain, there is a COOH group, the C terminal. In between, water molecules have been eliminated to leave amino acid units, or residues.

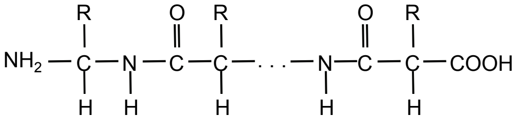

$$NH_2-\underset{H}{\overset{R}{C}}-\underset{H}{N}-\overset{O}{\overset{\|}{C}}-\underset{H}{\overset{R}{C}}-\cdots-\underset{H}{N}-\overset{O}{\overset{\|}{C}}-\underset{H}{\overset{R}{C}}-COOH$$

A typical **polypeptide** protein consists of between 100 and 500 or more amino acid residues. The sequence of amino acids is the **primary structure** of the molecule. The long chains of proteins themselves have a **secondary structure**, usually in the form of a helix or spiral. **Hydrogen bonds** between different amino acid groups are responsible for this structure. Often, there is more than one polypeptide chain present and they exist as intertwined helixes held together by hydrogen or ionic bonds leading to a **tertiary structure**.

When a protein undergoes complete hydrolysis, the amino acids are recovered.

12.5 Nucleic Acids

Carbohydrates, lipids and proteins make up the bulk of living organisms. A fourth major class of compounds comprise the nucleic acids **DNA** and **RNA**. Nucleic acids are long-chain polymers that consist of smaller units, **nucleotides**, joined together.

In DNA, there are four types of nucleotide. Each nucleotide consists of a 5-carbon sugar (**deoxyribose**), with an attached phosphate, and a nitrogen base. Each type is distinguished by the identity of the nitrogenous bases.

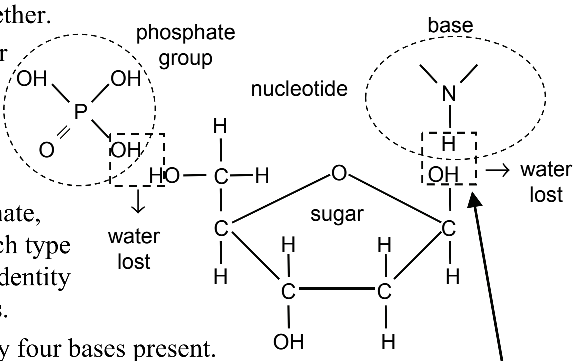
phosphate group
OH
OH
P
O
OH
HO
water lost
nucleotide
H
C
H
C
H
O
sugar
H
H
C
C
OH
H
base
N
H
OH
C
H
water lost

The NH group on the base combines with the OH group on the sugar and a H_2O is lost.

In DNA, there are only four bases present.

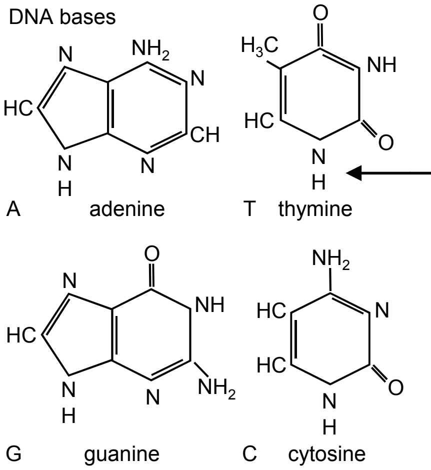
DNA bases
N
NH2
N
HC
CH
N
N
H
A adenine
O
H3C
NH
HC
O
N
H
T thymine
O
N
NH
HC
N
N
NH2
H
G guanine
NH2
HC
N
HC
O
N
H
C cytosine

The primary structure of DNA is a sequence of nucleotides bonded together as long chains. The phosphate group of one nucleotide bonds with the sugar of another (losing a molecule of water in the process).

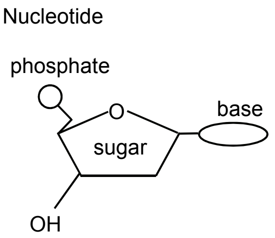
Nucleotide
phosphate
O
base
sugar
OH

A shorthand way of writing these groups is to just show the covalent bonds between the groups, and leave out the C and H atoms.

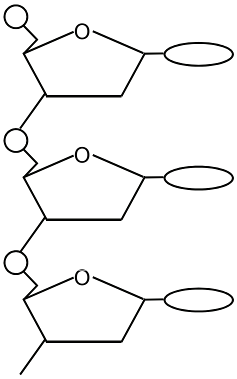

12.6 DNA

The primary structure of DNA is essentially the sequence of bases in the nucleotides in the chain. However, DNA does not ordinarily exist as a single chain, but is paired with an opposite chain so that the bases are paired and held together by **hydrogen bonds**.

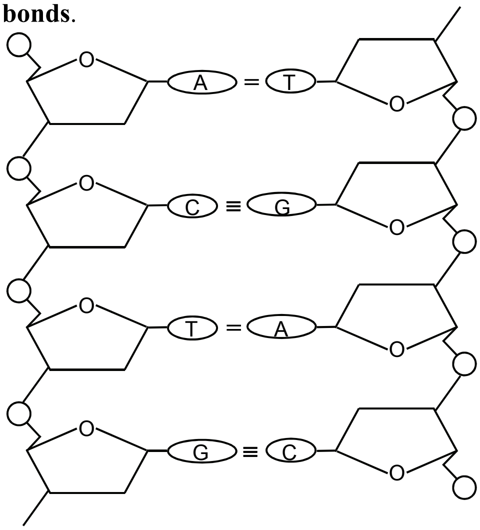

The bases can only pair according to certain rules: A is always paired with T by a double hydrogen bond. G is always paired with C by a triple hydrogen bond.

The ladder-like structure is further characterised by being coiled up in a spiral, each chain with bases pointing inwards and bonding with the corresponding base on the other – the so-called **double-helix**. The nucleotides, identified by their bases, can appear in any particular order, but the sequence on one chain has to be reflected in the other by virtue of the **base-pairing rules** A-T and G-C.

Sequences of nucleotides fall into functional groups called **genes**, which in turn lie along a single large DNA molecule called a **chromosome**. In most cells, chromosomes occur in duplicated pairs. When cell division by **mitosis** occurs, the DNA molecules are replicated by rupturing the hydrogen bonds between the two chains and then each chain forming bonds with new nucleotides according to base-pairing rules. When cell division is complete, the two new cells each contain paired chromosomes of the same genetic sequence as the parent cell.

Although there are only four bases in human DNA, the molecule typically consists of about 1500 nucleotides in the chain, allowing billions of combinations of base sequences possible in a single molecule.

RNA is similar to DNA but consists of ribose as the sugar and a single chain structure with the base **uracil** rather than thymine. There are different types of RNA; some act as messengers, carrying the DNA template to the site of protein synthesis (**ribosomes**) in the cell, while others transport amino acids to the ribosomes.

12.7 Enzymes

The many chemical reactions which take place in an organism to sustain life are collectively called **metabolism**. Typical reactions are **condensation** and **hydrolysis** reactions:

sugars	Condensation	polysaccharides (cellulose, starch)
fatty acids + glycerol	→	fats
	Hydrolysis	
amino acids	←	proteins

These reactions occur at temperatures of about 37°C under atmospheric pressure. The reaction rates would be too slow unless reaction pathways were altered by use of **catalysts**. Nearly every biological chemical reaction involves the use of a specific biological catalyst called an **enzyme**.

Enzymes are very large protein molecules of a specific shape. The enzyme molecule is typically much larger than the molecules actually involved in the chemical reaction. The reaction molecule (or molecules in the case of a condensation reaction), called the substrate, attaches itself to an **active site** on the enzyme molecule.

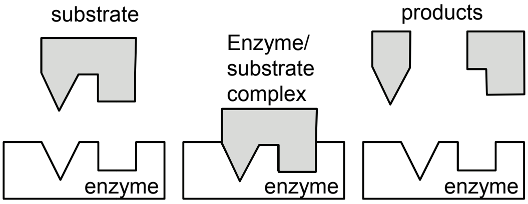

An **activated complex** is formed, and the reaction proceeds along a different pathway than would normally occur in the laboratory.

When the reaction is complete, the products detach from the enzyme and the enzyme is available to catalyse another reaction.

Enzymes lower the **activation energy** by increasing the collision frequency of molecules by virtue of bringing the molecules together spatially in an optimum manner, or altering the bond energies involved in the reaction by forming temporary bonds on the reacting molecules, and in some cases, applying mechanical stress to the reacting molecules to facilitate contact.

Enzyme inhibitors are used within the organism to control the activity of enzymes by masking or blocking the active sites on the enzyme to regulate reaction rates as a whole.

The importance of enzymes cannot be over-stated. In a living cell, the reagents of all the chemical reactions are not stored in isolated bottles and mixed when needed as in a laboratory, but exist all together in solution. The process of chemical reactions between these mixed reagents is orchestrated by the action of enzymes.

12.8 ATP

Organisms need nutrients to stay alive. Nutrients comprise carbohydrates, proteins and fats and, in animals, are usually ingested as food. Nutrients provide a source of energy, essential amino acids and essential fatty acids. Vitamins and other trace elements are also required for the proper functioning of enzymes.

Metabolism is the oxidation of glucose to carbon dioxide, water and energy:

$$C_6H_{12}O_6 + 6O_2 + 36ADP + 36P \rightarrow 6CO_2 + 6H_2O + 36ATP$$

Energy stored as phosphate bonds via the conversion of ADP (adenosinediphosphate) to ATP (adenosinetriphosphate)

The energy released by the oxidation process is stored as chemical potential energy in the formation of **ATP** molecules from **ADP** molecules. **Adenosinediphosphate** consists of a molecule of adenine, ribose, and two phosphate groups:

$$ADP + P + \text{energy} \leftrightarrow ATP + H_2O$$

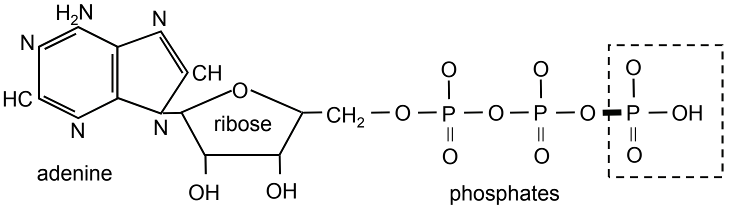

When energy is used to bind a third phosphate group to the end of the molecule, **adenosinetriphosphate** is formed. This is an endothermic process and is called **phosphorylation**. The energy used to create ATP is stored in the phosphate bond. Energy-rich ATP molecules are then hydrolysed to release their energy when used directly in other biological processes such as muscle action, protein synthesis and nerve impulse transmission.

Another important molecule involved in metabolism is the co-enzyme NAD (**nicotinamide adenine dinucleotide phosphate**). NAD takes part in oxidation-reduction reactions in metabolism by losing and accepting electrons ultimately necessary for conversion of ADP to ATP. The redox reactions are

$NAD^+ + 2H^+ + 2e^- \rightarrow NADH + H^+$ reduction

$NADH + H^+ \rightarrow NAD^+ + 2H^+ + 2e^-$ oxidation

Reduction can occur by adding hydrogen.
Oxidation can occur by removing hydrogen.

When NAD^+ is reduced, it stores energy by storing an excited electron in the form of NADH and a H^+ ion. When $NADH+H^+$ is oxidised, it releases energy, usually to form ATP from ADP.

12.9 Anaerobic Metabolism

Anabolic processes build up low-energy reactants into high-energy products. Photosynthesis in plants is an anabolic process. **Catabolic** processes break down high-energy reactants to low-energy products, releasing energy in doing so. This energy is transferred to **ATP**.

Anaerobic metabolism involves two steps: (i) the breakdown of glucose from a **carbohydrate** into **pyruvic acid**:

$$C_6H_{12}O_6 + 2NAD^+ + 4ADP \rightarrow 2C_3H_4O_3 + 2NADH + 2H^+ + 4ATP$$

and (ii) the fermentation of pyruvic acid into either CO_2 and ethanol, or lactic acid. The ultimate aim is to produce ATP from ADP.

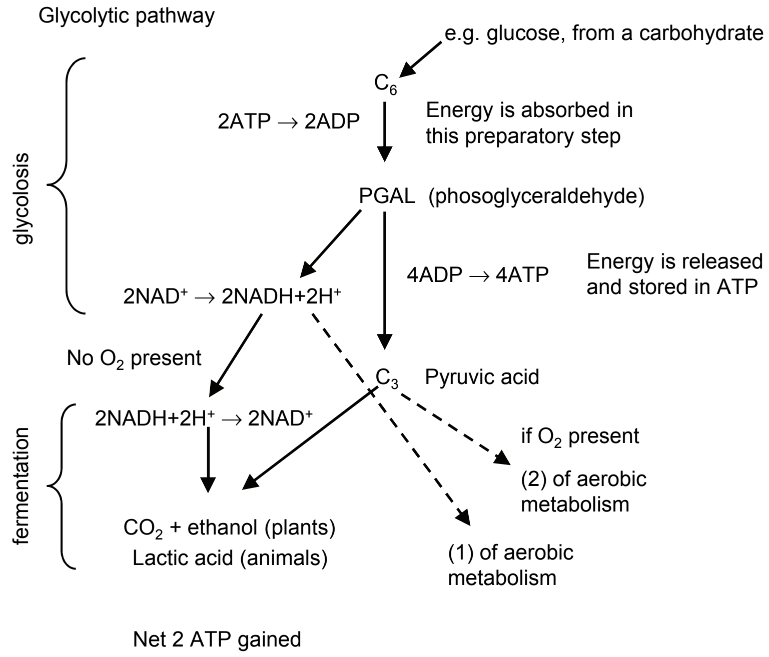
Glycolytic pathway
e.g. glucose, from a carbohydrate
C6
2ATP → 2ADP
Energy is absorbed in this preparatory step
glycolosis
PGAL (phosoglyceraldehyde)
2NAD+ → 2NADH+2H+
4ADP → 4ATP
Energy is released and stored in ATP
No O2 present
C3 Pyruvic acid
2NADH+2H+ → 2NAD+
if O2 present
(2) of aerobic metabolism
fermentation
CO2 + ethanol (plants)
Lactic acid (animals)
(1) of aerobic metabolism
Net 2 ATP gained

Anaerobic metabolism occurs in all cells. It does result in a particularly high yield of ATP since the products, ethanol, lactic acid and pyruvic acid, still contain a substantial portion of the original energy from the glucose. While anaerobic metabolism can provide ATP for energy use (in muscles) when O_2 is in short supply, or none at all, it also provides a source of **pyruvic acid** for use in **aerobic respiration**.

12.10 Aerobic Metabolism

Aerobic metabolism, or **respiration**, is the process by which glycerol is oxidised with the production of ATP molecules from ADP molecules in the presence of molecular oxygen:

from glycolysis

$$(1)\quad 2NADH + 2H^+ + O_2 \rightarrow 2NAD^+ + 2H_2O \quad + \rightarrow 6\text{ ATP}$$

$$(2)\quad \underset{\text{Pyruvic acid}}{2C_3} + \overset{\text{Coenzyme A}}{2CoA} + 2NAD^+ \rightarrow \overset{C_2}{2AcetylCoA} + 2NADH + 2H^+ + 2CO_2$$

$$(3)\quad 2NADH + 2H^+ + O_2 \rightarrow 2NAD^+ + 2H_2O \quad + \rightarrow 6\text{ ATP}$$

$$(4)\quad 2AcetylCoA + \text{Krebs cycle} \rightarrow CO_2 + H_2O \quad + \rightarrow 2\text{ ATP directly} + NADH{+}H^+$$

(1) The NADH from **glycolysis** is oxidised, by the addition of O_2, to produce six ATP.

(2) **Pyruvic acid** C_3 is oxidised to an activated form of acetic acid C_2. NAD^+ is reduced to $NADH{+}H^+$.

(3) $NADH{+}H^+$ from (2) is oxidised by O_2, to form six ATP.

(4) **AcetylCoA** enters the **Krebs cycle** (or **citric acid cycle**). This consists of multiple reactions, the net effect of which is to form two molecules of ATP plus energy-rich NADH.

During the Krebs cycle, six molecules of $NADH{+}H^+$ and two molecules of a related compound, FADH, are formed. These electron-carrier molecules are energy-rich and can be further oxidised as in (1) and (3) which require O_2. The net result of (4) is the production of twenty two ATP molecules. **Anaerobic metabolism** produces two molecules of ATP, which added to the thirty four ATP molecules from aerobic metabolism makes thirty six molecules of ATP from the complete oxidation of one molecule of **glucose**.

Fats and **proteins** are also sources of ATP, by virtue of their breakdown into components which can be inserted into glycolytic and aerobic pathways.

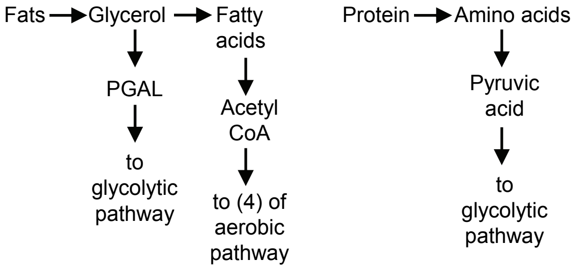

12.11 Cyclic Photophosphorylation

Ultimately, the energy contained within glucose used in animal metabolism comes from **photosynthesis** in plants. The overall reaction in plants, for the production of glucose, is

$$6CO_2 + 12H_2O + \overset{\overset{\text{light}}{\downarrow}}{\text{Energy}} \rightarrow C_6H_{12}O_6 + 6O_2 + 6H_2O$$

Note that O_2 is a product of this reaction.

Chlorophyll is the unique ingredient of plants that captures photons to promote an electron of low energy into an excited state. The excited electron is passed from one transfer molecule to another, losing energy at each step, until it returns to a chlorophyll molecule in an unexcited state. At each step, the energy lost is used to convert ADP to ATP:

$$ADP + P + \text{Energy} \leftrightarrow ATP + H_2O$$

As the excited electron is passed from molecule to molecule, *some* of the energy is used to create ATP. The process is not 100% efficient.

Because electrons are donated by chlorophyll from the high-energy state and then to acceptance in a low-energy state, this synthesis of ATP is called **cyclic photophosphorylation**. The ATP, essentially created from ADP by the energy of sunlight, is ultimately used to create high-energy glucose (which is stored in the plant as starch) from low-energy CO_2. Animals eat the plants and convert starch into glucose, which is then used to create ATP for their own metabolic processes.

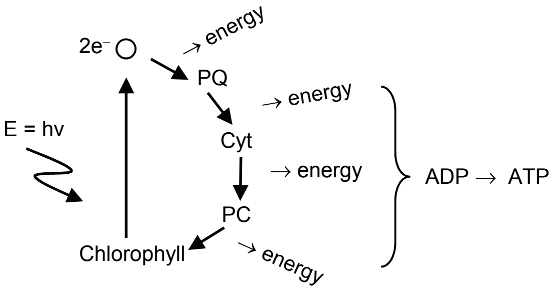

PQ, Cyt and PC are electron carrier molecules that undergo reduction (when they accept the electron) and oxidation (when they donate the electron).

In this mode of photophosphorylation, electrons are transported around in a cycle; no oxygen or NADH+H^+ is produced.

12.12 Non-Cyclic Photophosphorylation

Another way of producing ATP used by plants has a beginning similar to cyclic photophosphorylation, but employs a more complicated path for extraction of energy from excited electrons. This non-cyclic photophosphorylation is used by modern green plants :

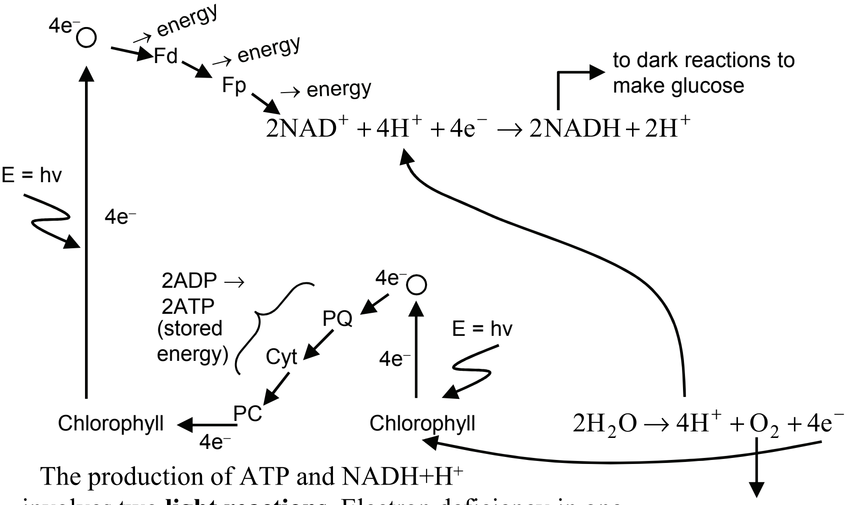
$4e^-$
→ energy
Fd
→ energy
Fp
→ energy
$2NAD^+ + 4H^+ + 4e^- \rightarrow 2NADH + 2H^+$
to dark reactions to make glucose
E = hv
$4e^-$
2ADP → 2ATP (stored energy)
$4e^-$
PQ
Cyt
PC
E = hv
$4e^-$
Chlorophyll
$4e^-$
Chlorophyll
$2H_2O \rightarrow 4H^+ + O_2 + 4e^-$
Note that O_2 is a product of this reaction

The production of ATP and NADH+H^+ involves two **light reactions**. Electron deficiency in one chlorophyll is balanced by different electrons from those from the water dissociation.

Additional reactions take place in which the energy obtained from the light reactions is converted into carbohydrates for storage. That is, CO_2 is reduced, and (if the energy is coming from NADH+H^+) the NDAH+H^+ is oxidised with the net result being glucose. These are known as **dark reactions** since no light is required.

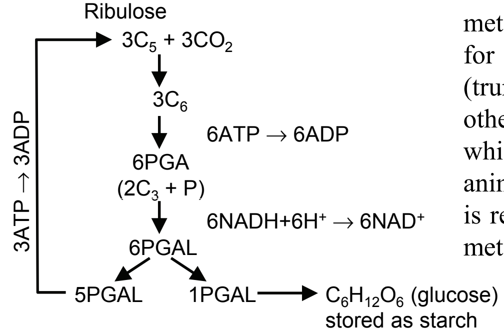
Ribulose
$3C_5 + 3CO_2$
$3C_6$
6ATP → 6ADP
6PGA
$(2C_3 + P)$
$6NADH+6H^+ \rightarrow 6NAD^+$
6PGAL
5PGAL
1PGAL
$C_6H_{12}O_6$ (glucose) stored as starch
3ATP → 3ADP

The resulting glucose is used in metabolism by the plant cells for making cellulose/starch (trunks and branches, etc) and other sugars (such as in fruit), which may in turn be eaten by animals from which the glucose is recovered and used in animal metabolism.

12.13 Metabolism

Summary:

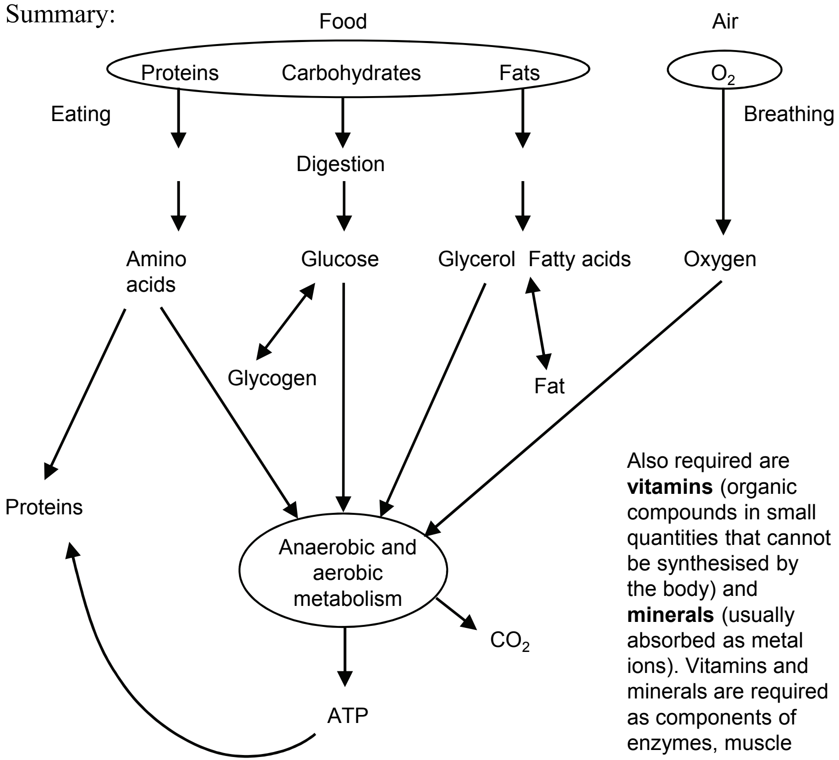
Food
Air
Proteins
Carbohydrates
Fats
O_2
Eating
Breathing
Digestion
Amino acids
Glucose
Glycerol
Fatty acids
Oxygen
Glycogen
Fat
Proteins
Anaerobic and aerobic metabolism
CO_2
ATP

Also required are **vitamins** (organic compounds in small quantities that cannot be synthesised by the body) and **minerals** (usually absorbed as metal ions). Vitamins and minerals are required as components of enzymes, muscle action, bones, teeth, haemoglobin, etc.

ATP is used to provide the energy source for conversion of amino acids into proteins required by the body for the production of skin, hair, cells, enzymes, muscles, nerves, signalling, anaerobic respiration, as well as for energising DNA replication. It also provides energy for muscle action, vision, brain activity, and nearly all biological energy transformations. Glycogen and fats can be stored in the body and sent back into the metabolism pathway as glucose and glycerol for oxidation into CO_2 and energy when there is no food being digested.

Chemistry

Like the author's other companion books, **The Chemistry Companion** provides high quality information in unique one-page-per-topic presentations that do not overburden and distract readers with excessive details. The book offers concise summaries of general chemistry concepts, easily accessible in a convenient, reader-friendly format.

Suitable as an introduction or study guide, this companion presents the minimum of what readers need to know to understand the subject. It emphasizes the physics underlying chemistry. By looking at chemistry processes from a physics point of view, readers better appreciate what is happening from the chemical perspective that is usually found in traditional chemistry books.

The author focuses on the structure of matter, chemical components and bonds, the periodic table, states of matter, thermodynamics, reaction rates, carbon chemistry, biochemistry, and chemical, ionic, and electronic equilibria. Each topic is covered in a single-page outline format with just enough detail to enable a good understanding of the subject.

Features

- Provides a physical understanding of chemical concepts
- Presents clear explanations of difficult material, working through any inconsistencies in understanding
- Uses a convenient format for checking formulas and definitions
- Includes self-contained information on each page, assuming little prior knowledge

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K11517

ISBN: 978-1-4398-3088-8
90000
9 781439 830888

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press

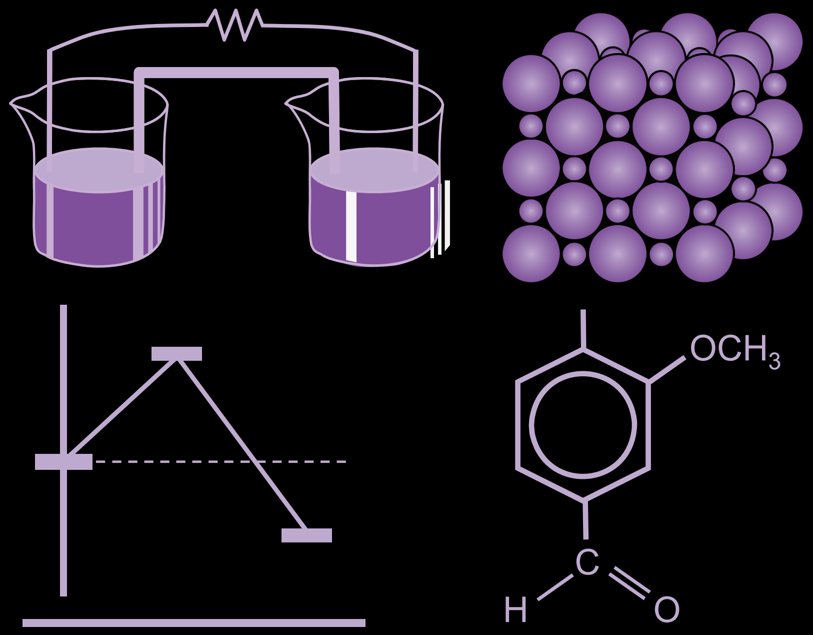

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press
Taylor & Francis Group

SCIENCE VISUAL RESOURCES

CHEMISTRY

An Illustrated Guide to Science

SCIENCE VISUAL RESOURCES

CHEMISTRY

An Illustrated Guide to Science

The Diagram Group

CHELSEA HOUSE
PUBLISHERS
An imprint of Infobase Publishing

Chemistry: An Illustrated Guide to Science

Copyright © 2006 The Diagram Group

Author:	Derek McMonagle BSc PhD CSci CChem FRSC
Editors:	Eleanora von Dehsen, Jamie Stokes, Judith Bazler
Design:	Anthony Atherton, Richard Hummerstone, Lee Lawrence, Phil Richardson
Illustration:	Peter Wilkinson
Picture research:	Neil McKenna
Indexer:	Martin Hargreaves

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage or retrieval systems, without permission in writing from the publisher. For information contact:

Chelsea House
An imprint of Infobase Publishing
132 West 31st Street
New York NY 10001

For Library of Congress Cataloging-in-Publication data, please contact the publisher.

ISBN 0-8160-6163-7

Chelsea House books are available at special discounts when purchased in bulk quantities for businesses, associations, institutions, or sales promotions. Please call our Special Sales Department in New York at 212/967-8800 or 800/322-8755.

You can find Chelsea House on the World Wide Web at
http://www.chelseahouse.com

Printed in China

CP Diagram 10 9 8 7 6 5 4 3 2 1

This book is printed on acid-free paper.

Introduction

Chemistry is one of eight volumes of the **Science Visual Resources** set. It contains eight sections, a comprehensive glossary, a Web site guide, and an index.

Chemistry is a learning tool for students and teachers. Full-color diagrams, graphs, charts, and maps on every page illustrate the essential elements of the subject, while parallel text provides key definitions and step-by-step explanations.

Atomic Structure provides an overview of the very basic structure of physical matter. It looks at the origins of the elements and explains the nature of atoms and molecules.

Elements and Compounds examines the characteristics of the elements and their compounds in detail. Tables give the boiling points, ionization energies, melting points, atomic volumes, atomic numbers, and atomic masses key elements. Plates also describe crystal structures and covalent bonding.

Changes in Matter is an overview of basic chemical processes and methods. It looks at mixtures and solutions, solubility, chromatography, and the pH scale.

Patterns—Non-Metals and **Patterns—Metals** focus on the properties of these two distinct groups of elements. These sections also include descriptions of the industrial processes used when isolating important elements of both types.

Chemical Reactions looks at the essential factors that influence reactions. It includes information on proton transfer, electrolysis, redox reactions, catalysts, and the effects of concentration and temperature.

Chemistry of Carbon details the chemical reactions involving carbon that are vital to modern industry—from the distillation of crude oil to the synthesis of polymers and the manufacture of soaps and detergents. This section also includes an overview of the chemistry of life.

Radioactivity is concerned with ionizing radiation, nuclear fusion, nuclear fission, and radioactive decay, as well as the properties of radiation. Tables describe all known isotopes, both radioactive and non-radioactive.

Contents

1 ATOMIC STRUCTURE

8 Formation of stars
9 Fate of stars
10 The solar system
11 Planet composition
12 Planetary density, size, and atmosphere
13 Atomic structure
14 Geiger and Marsden's apparatus
15 Investigating the electron 1
16 Investigating the electron 2
17 Cathode ray oscilloscope
18 Measuring the charge on the electron
19 Size and motion of molecules
20 Determination of Avogadro's constant
21 The mole
22 Atomic emission spectrum: hydrogen
23 Energy levels: hydrogen atom
24 Luminescence

2 ELEMENTS AND COMPOUNDS

25 Organizing the elements
26 The periodic table
27 First ionization energies of the elements
28 Variation of first ionization energy
29 Melting points of the elements °C
30 Variation of melting points
31 Boiling points of the elements °C
32 Variation of boiling points
33 Atomic volumes of the elements
34 Variation of atomic volumes
35 Atomic mass
36 Periodic table with masses and numbers
37 Calculating the molecular mass of compounds
38 Structure of some ionic crystals
39 Crystal structure of metals: lattice structure
40 Crystal structure of metals: efficient packing
41 Chemical combination: ionic bonding
42 Chemical combination: ionic radicals
43 Chemical combination: covalent bonding
44 Chemical combination: coordinate bonding

3 CHANGES IN MATTER

45 Mixtures and solutions
46 Colloids
47 Simple and fractional distillation
48 Separating solutions
49 Paper chromatography
50 Gas-liquid chromatography and mass spectrometry
51 The pH scale
52 Indicators
53 Titration of strong acids
54 Titration of weak acids

55 pH and soil
56 The water cycle
57 Treatment of water and sewage
58 The water molecule
59 Water as a solvent of ionic salts
60 Ionic solutions
61 Solubility
62 Solubility curves
63 Solubility of copper(II) sulfate

4 PATTERNS—NON-METALS

64 Hydrogen: preparation
65 Hydrogen: comparative density
66 Hydrogen: reaction with other gases
67 Hydrogen: anomalies in ammonia and water
68 Basic reactions of hydrogen
69 The gases in air
70 Nitrogen
71 Other methods of preparing nitrogen
72 The nitrogen cycle
73 Preparation and properties of ammonia
74 Industrial preparation of ammonia (the Haber process): theory
75 Industrial preparation of ammonia (the Haber process): schematic
76 Industrial preparation of nitric acid
77 Nitrogen: reactions in ammonia and nitric acid
78 Basic reactions of nitrogen
79 Nitrate fertilizers
80 Oxygen and sulfur
81 Extraction of sulfur—the Frasch process
82 Oxygen and sulfur: allotropes
83 Oxygen and sulfur: compound formation
84 The oxides of sulfur
85 Industrial preparation of sulfuric acid (the contact process): theory
86 Industrial preparation of sulfuric acid (the contact process): schematic
87 Affinity of concentrated sulfuric acid for water
88 Oxygen and sulfur: oxidation and reduction
89 Basic reactions of oxygen
90 Basic reactions of sulfur
91 The halogens: group 7
92 Laboratory preparation of the halogens
93 Compounds of chlorine
94 Hydrogen chloride in solution
95 Acid/base chemistry of the halogens
96 Redox reactions of the halogens
97 Reactivity of the halogens

5 PATTERNS—METALS

98 World distribution of metals
99 Main ores of metals
100 The group 1 metals
101 The group 1 metals: sodium
102 The group 2 metals

103 The group 2 metals: general reactions
104 The transition metals: electron structure
105 The transition metals: ionization energies and physical properties
106 Aluminum
107 Iron: smelting
108 The manufacture of steel
109 Rusting
110 Copper smelting and converting
111 Reactions of copper
112 Reaction summary: aluminum, iron, and copper
113 The extraction of metals from their ores
114 Reactivity summary: metals
115 Tests on metals: flame test
116 Tests on metals: metal hydroxides
117 Tests on metals: metal ions
118 Uses of metals

6 CHEMICAL REACTIONS

119 Reactivity of metals 1
120 Reactivity of metals 2
121 Electrolysis
122 Electrolysis: electrode activity and concentration
123 Acids: reactions
124 Preparation of acids
125 Bases: reactions
126 Bases: forming pure salts
127 Proton transfer: neutralization of alkalis
128 Proton transfer: neutralization of bases
129 Proton transfer: metallic carbonates
130 Proton transfer: neutralization of acids
131 Collision theory
132 Rates of reaction: surface area and mixing
133 Rates of reaction: temperature and concentration
134 Rates of reaction: concentration over time
135 Rate of reaction vs. concentration
136 Variation of reaction rate
137 Rates of reaction: effect of temperature 1
138 Rates of reaction: effect of temperature 2
139 Exothermic and endothermic reactions
140 Average bond dissociation energies
141 Catalysts: characteristics
142 Catalysts: transition metals
143 Oxidation and reduction
144 Redox reactions 1
145 Redox reactions 2
146 Demonstrating redox reactions
147 Assigning oxidation state

7 CHEMISTRY OF CARBON

148 The allotropes of carbon: diamond and graphite
149 The allotropes of carbon: fullerenes
150 The carbon cycle
151 Laboratory preparation of carbon oxides
152 The fractional distillation of crude oil
153 Other refining processes

154 Carbon chains
155 Naming hydrocarbons
156 Table of the first six alkanes
157 Table of the first five alkenes
158 Ethene
159 Polymers
160 Polymers: formation
161 Polymers: table of properties and structure
162 Functional groups and homologous series
163 Alcohols
164 Carboxylic acids
165 Esters
166 Soaps and detergents
167 Organic compounds: states
168 Functional groups and properties
169 Reaction summary: alkanes and alkenes
170 Reaction summary: alcohols and acids
171 Optical isomerism
172 Amino acids and proteins
173 Monosaccharides
174 Disaccharides and polysaccharides

8 RADIOACTIVITY

175 Ionizing radiation
176 Radiation detectors
177 Properties of radiations: penetration and range
178 Properties of radiations: in fields
179 Stable and unstable isotopes
180 Half-life
181 Measuring half-life
182 Radioactive isotopes
183 Nuclear fusion
184 Nuclear fission
185 Nuclear reactor
186 The uranium series
187 The actinium series
188 The thorium series
189 The neptunium series
190 Radioactivity of decay sequences
191 Table of naturally occurring isotopes 1
192 Table of naturally occurring isotopes 2
193 Table of naturally occurring isotopes 3
194 Table of naturally occurring isotopes 4
195 Table of naturally occurring isotopes 5
196 Table of naturally occurring isotopes 6
197 Table of naturally occurring isotopes 7

APPENDIXES

198 Key words
205 Internet resources
207 Index

Key words

Big Bang
black hole
brown dwarf
neutron star
protostar
supernova
white dwarf

Formation of stars

Beginnings

- According to the *Big Bang* theory, the universe resulted from a massive explosion that created matter, space, and time.
- During the first thee minutes following the Big Bang, hydrogen and helium were formed as the universe began to cool.

Initial formation

- Stars were formed when gravity caused clouds of interstellar gas and dust to contract. These clouds became denser and hotter, with their centers boiling at about a million kelvins.
- These heaps became round, glowing blobs called *protostars*.
- Under the pressure of gravity, contraction continued, and a protostar gradually became a genuine star.
- A star exists when all solid particles have evaporated and when light atoms such as hydrogen have begun building heavier atoms through nuclear reactions.
- Some cloud fragments do not have the mass to ignite nuclear reactions. These become *brown dwarfs*.
- The further evolution of stars depends on their size (See page 9).
- Stars the size of our Sun will eventually shed large amounts of matter and contract into a very dense remnant—a *white dwarf*, composed of carbon and oxygen atoms.
- More massive stars collapse quickly shedding much of their mass in dramatic explosions called *supernovae*. After the explosion, the remaining material contracts into an extremely dense *neutron star*.
- The most massive stars eventually collapse from their own gravity to *black holes*, whose density is infinite.

© Diagram Visual Information Ltd.

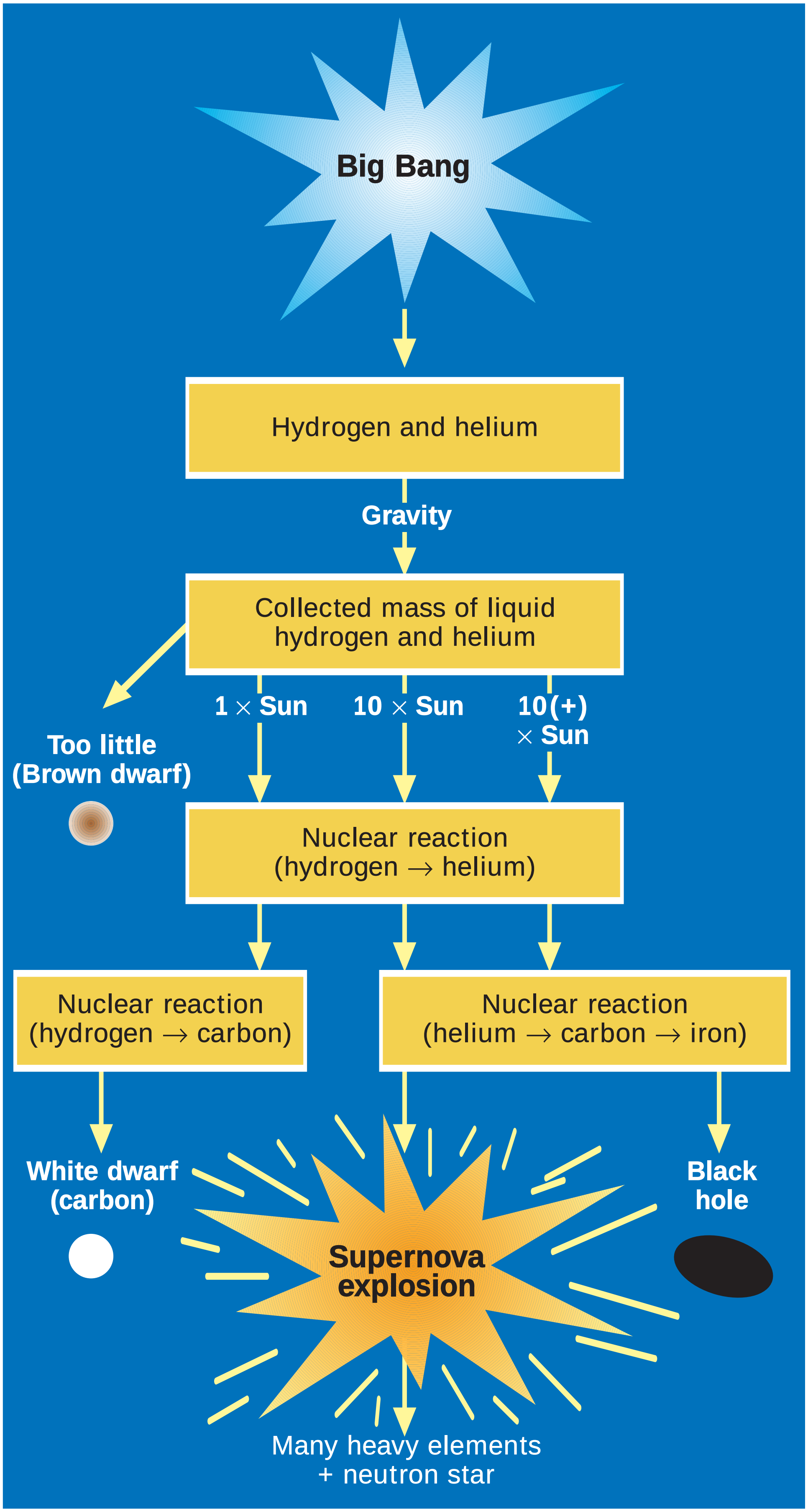
Big Bang
Hydrogen and helium
Gravity
Collected mass of liquid hydrogen and helium
Too little (Brown dwarf)
1 × Sun
10 × Sun
10(+) × Sun
Nuclear reaction (hydrogen → helium)
Nuclear reaction (hydrogen → carbon)
Nuclear reaction (helium → carbon → iron)
White dwarf (carbon)
Supernova explosion
Black hole
Many heavy elements + neutron star

Fate of stars

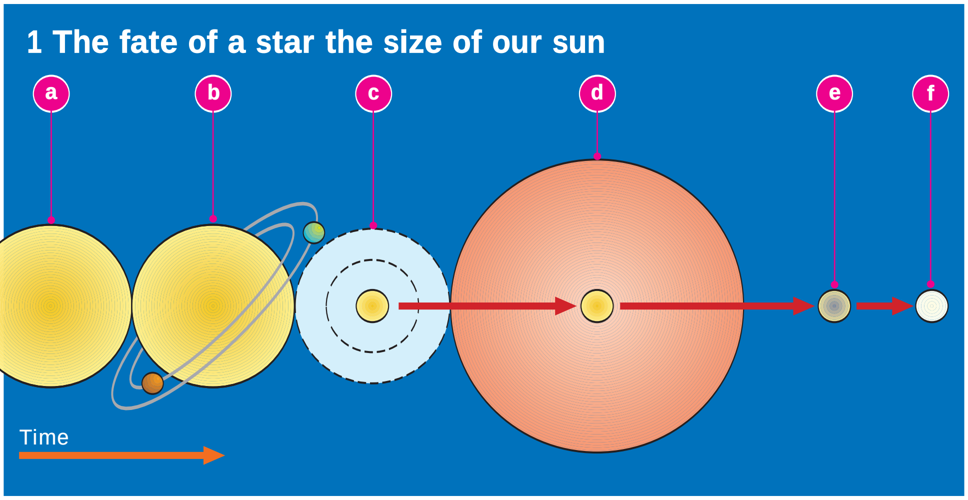

1 The fate of a star the size of our sun

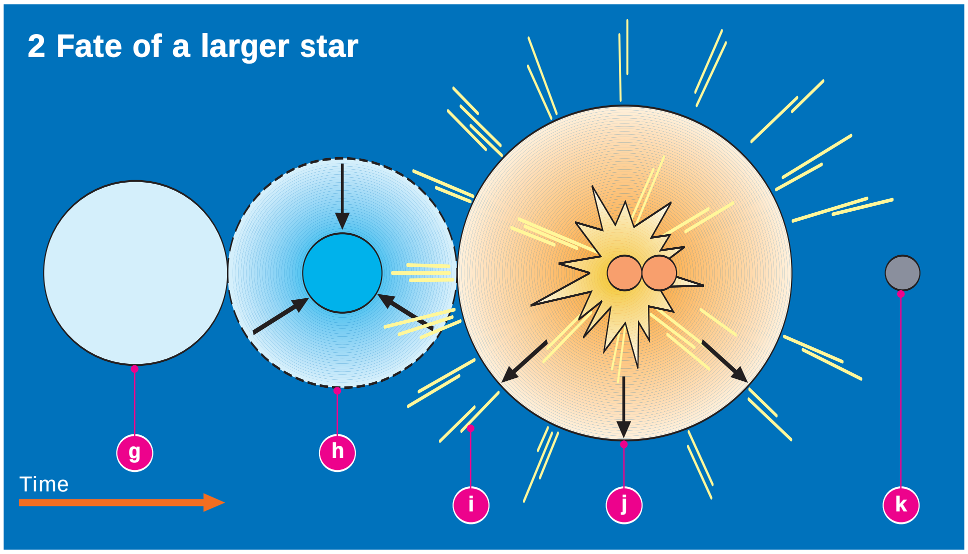

2 Fate of a larger star

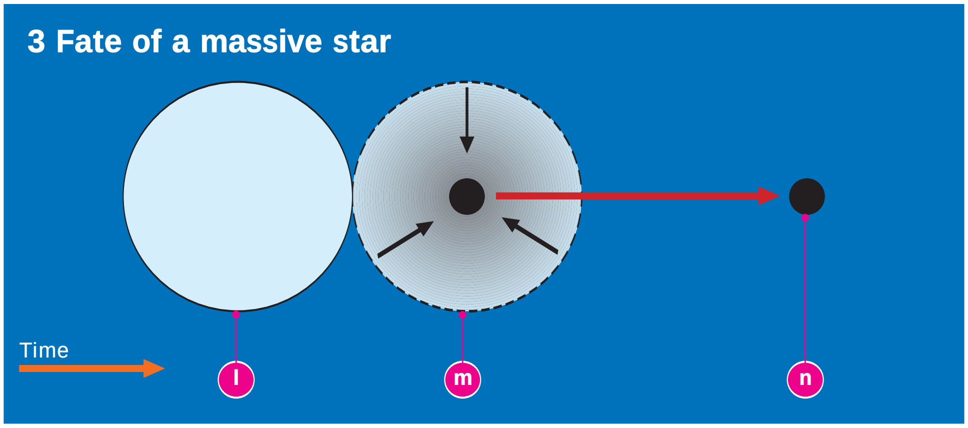

3 Fate of a massive star

a hydrogen is converted to helium
b planetary system evolves
c hydrogen runs out and helium is converted to carbon
d star cools to form a red giant
e carbon
f star evolves to form a white dwarf
g hydrogen is converted to helium and carbon, and eventually iron
h hydrogen runs out, and star undergoes gravitational collapse
i The collapsed star suddenly expands rapidly, creating a supernova explosion
j creates many different elements
k the core of the dead star becomes a neutron star
l hydrogen converted to many different elements
m hydrogen runs out, and the star collapses to form a black hole
n black hole

Key words

black hole
fusion
neutron star
red giant
supernova
white dwarf

Fate of stars

- During most of a star's life, the outward pressure from nuclear fusion balances the pull of gravity, but as nuclear fuel is exhausted, gravity compresses the star inward and the core collapses. How and how far it collapses depends on the size of the star.

1 The fate of a star the size of our sun

- A star the size of our Sun burns hydrogen into helium until the hydrogen is exhausted and the core begins to collapse. This results in nuclear *fusion* reactions in a shell around the core. The outer shell heats up and expands to produce a *red giant*.
- Ultimately, as its nuclear reactions subside, a red giant cools and contracts. Its core becomes a very small, dense hot remnant, a *white dwarf*.

2 Fate of a larger star

- Stars with an initial mass 10 times that of our Sun go further in the nuclear fusion process until the core is mostly carbon. The fusion of carbon into larger nuclei releases a massive amount of energy. The result is a huge explosion in which the outer layers of the star are blasted out into space. This is called a *supernova*.
- After the explosion, the remaining material contracts, and the core collapses into an extraordinary dense object composed only of neutrons—a neutron star.

3 Fate of a massive star

- Stars with an initial mass of 30 times our Sun undergo a different fate altogether. The gravitational field of such stars is so powerful that material cannot escape from them. As nuclear reactions subside, all matter is pulled into the core, forming a *black hole*.

© Diagram Visual Information Ltd.

The solar system

Key words

ammonia
fission
helium
hydrogen
methane

1 Birth of the solar system

- The solar system is thought to have formed about 4.6 billion years ago as a result of nuclear *fission* in the Sun.
- A nebula (cloud) of gases and dust that resulted from the explosion. flattened into a disk with a high concentration of matter at the center.

2 Formation of the inner and outer planets

- Near the Sun, where the temperature was high, volatile substances could not condense, so the inner planets (Mercury, Venus, Earth, and Mars) are dominated by rock and metal. They are smaller and more dense than those farther from the Sun.
- In the colder, outer areas of the disk, substances like *ammonia* and *methane* condensed, while *hydrogen* and *helium* remained gaseous. In this region, the planets formed (Jupiter, Saturn, Uranus, and Neptune) were gas giants.

3 Inner planets

- Inner planets consist of a light shell surrounding a dense core of metallic elements.
- Mercury, the planet closest to the Sun, has a proportionately larger core than Mars, the inner planet farthest from the Sun.

4 Outer planets

- The outer planets have low densities and are composed primarily of hydrogen and helium.
- The outer planets are huge in comparison to the inner planets.
- Jupiter and Saturn, the largest of the gas giants, contain the greatest percentages of hydrogen and helium; the smaller Uranus and Neptune contain larger fractions of water, ammonia, and methane.

© Diagram Visual Information Ltd.

1 Birth of the solar system

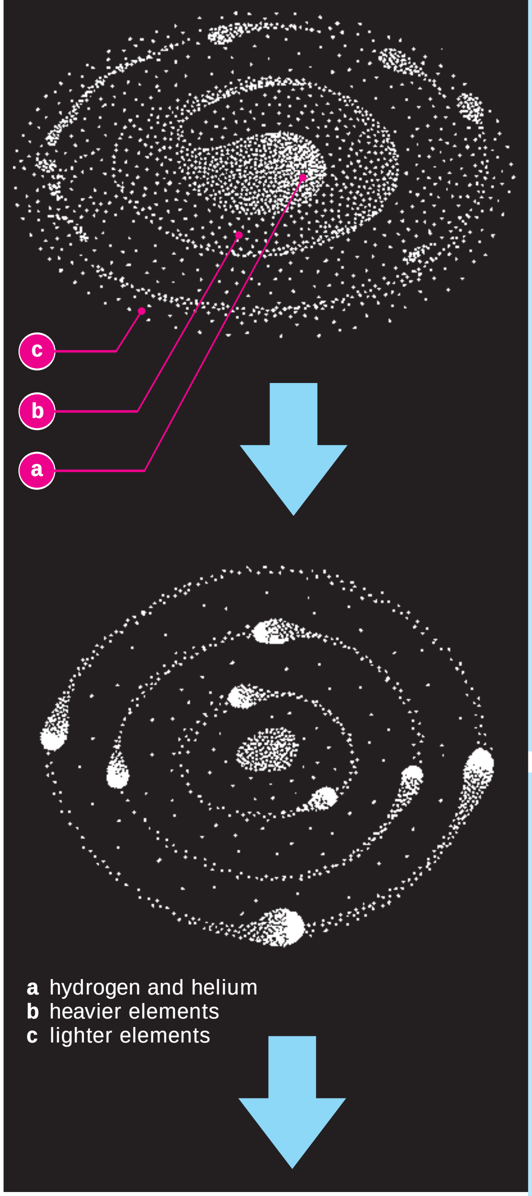

a hydrogen and helium
b heavier elements
c lighter elements

2 Formation of the inner and outer planets

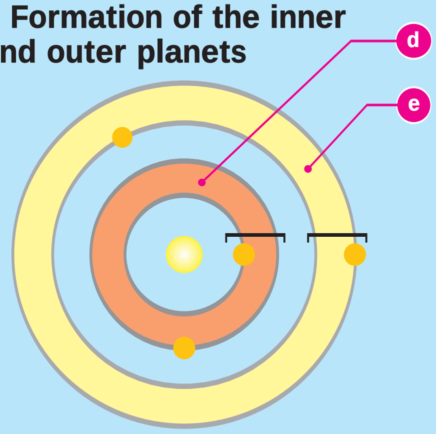

d denser inner planets
e less dense outer planets

3 Inner planets

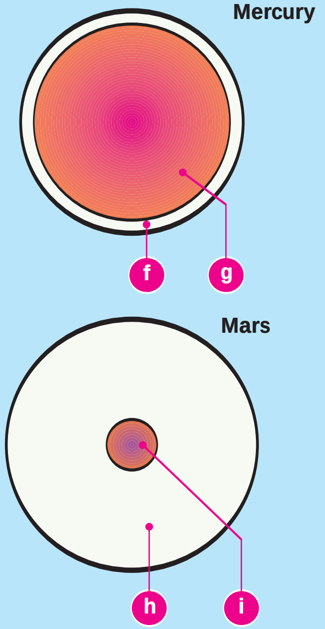

f light shell
g dense core
h light shell
i dense core

4 Outer planets

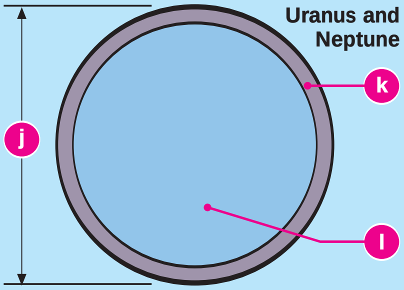

j diameter = 2 or 3 that of the Earth
k solid water, methane, and ammonia
l liquid water, methane, and ammonia

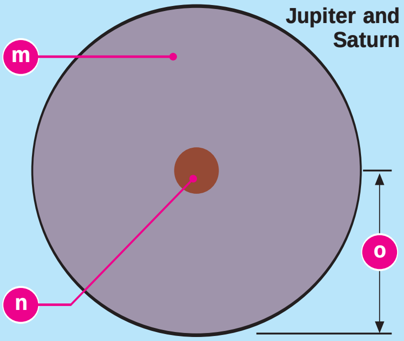

m liquid hydrogen and helium
n small rocky center
o radii:
Jupiter = 11 × radius of Earth
Saturn = 9 × radius of Earth

Planet composition

1 Basic composition of the planets

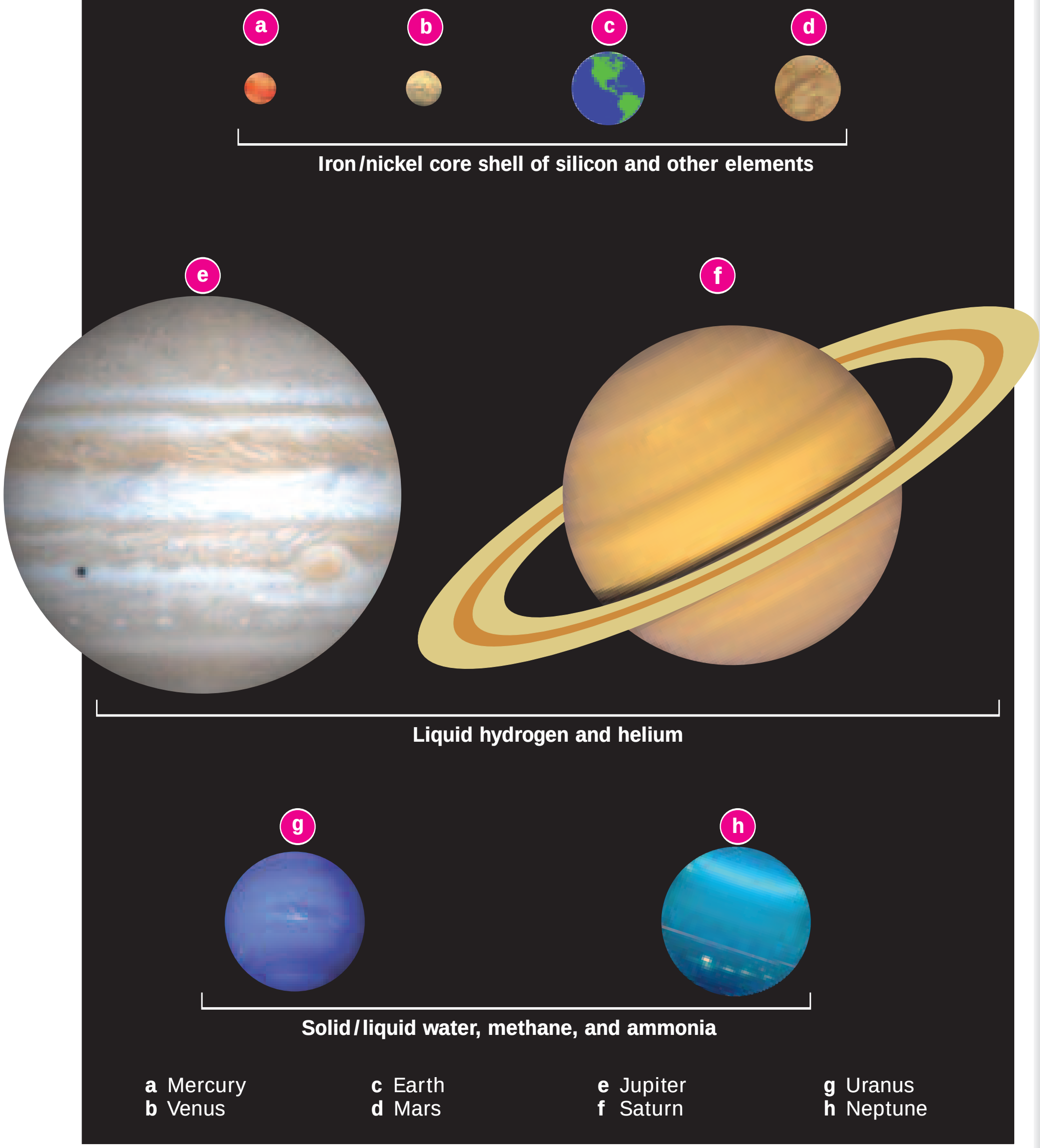
a
b
c
d
Iron/nickel core shell of silicon and other elements
e
f
Liquid hydrogen and helium
g
h
Solid/liquid water, methane, and ammonia
a Mercury
b Venus
c Earth
d Mars
e Jupiter
f Saturn
g Uranus
h Neptune

2 Composition of Earth

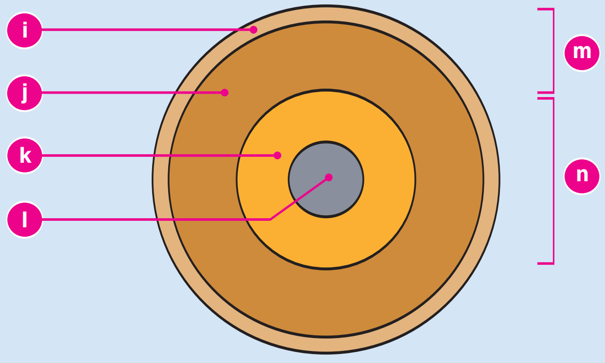

Composition of Earth	%
oxygen	46
silicon	28
aluminum	8
iron	5
calcium	4
sodium	3
potassium	3
magnesium	2

i crust
j mantle (oxygen, silicon, aluminum, iron)
k outer core (liquid – nickel and iron)
l inner core (solid – nickel and iron)
m crust, mantle, and oceans = $^{2}/_{3}$ of mass)
n core = $^{1}/_{3}$ of mass

Key words

atmosphere
carbonate
crust
mantle
nitrate
oxide
sulfate

1 Basic composition of the planets

- The inner planets—Mercury, Venus, Earth, and Mars—consist of an iron–nickel core surrounded by a shell of silicon and other elements.
- The outer planets—Jupiter, Saturn, Uranus, and Neptune—consist largely of solid or liquid methane, ammonia, liquid hydrogen, and helium.
- Pluto is not included in this comparison because it is atypical of the other outer planets, and its origins are uncertain.

2 Composition of Earth

- Earth consists of a dense, solid inner core and a liquid outer core of nickel and iron. The core is surrounded by the *mantle* (a zone of dense, hot rock), and finally by the *crust*, which is the surface of Earth.
- Since most of the materials of Earth are inaccessible (the deepest drilled holes only penetrate a small distance into the crust), we can only estimate the composition of Earth by looking at the composition of the materials from which Earth formed. Meteorites provide this information.
- Oxygen is the most common element on Earth, and about one fifth of Earth's *atmosphere* is gaseous oxygen.
- Oxygen is also present in many compounds, including water (H_2O), carbon dioxide (CO_2), and silica (SiO_2), and metal salts such as *oxides*, *carbonates*, *nitrates*, and *sulfates*.

© Diagram Visual Information Ltd.

Planetary density, size, and atmosphere

Key words

atmosphere
carbon dioxide
chlorophyll
photosynthesis

1 Densities and radii of the planets

- The inner planets—Mercury, Venus, Earth, and Mars—are relatively small but have a higher density than the outer planets.
- The outer planets—Jupiter, Saturn, Uranus, and Neptune—are relatively large but have a lower density than the inner planets.

2 Atmospheric composition of the inner planets

- Earth's *atmosphere* was probably similar to that of Venus and Mars when the planets formed. However, the particular conditions on Earth allowed life to start and flourish. With this came drastic changes to the composition of the atmosphere. Of particular importance is the evolution of green plants.
- Green plants contain a pigment called *chlorophyll*. Plants use this pigment to trap energy from sunlight and make carbohydrates. The process is called *photosynthesis*.
- As Earth became greener, the proportion of *carbon dioxide* in the atmosphere fell until it reached the present level of about 0.04 percent.
- The green plants provided a means of turning the Sun's energy into food, which in turn, provided animals with the energy they needed to survive. Thus, animals could evolve alongside plants.
- Conditions on the two planets adjacent to Earth—Venus and Mars—were not suitable for life as we know it, and the atmospheres on these planets have remained unchanged.

© Diagram Visual Information Ltd.

1 Densities and radii of the planets

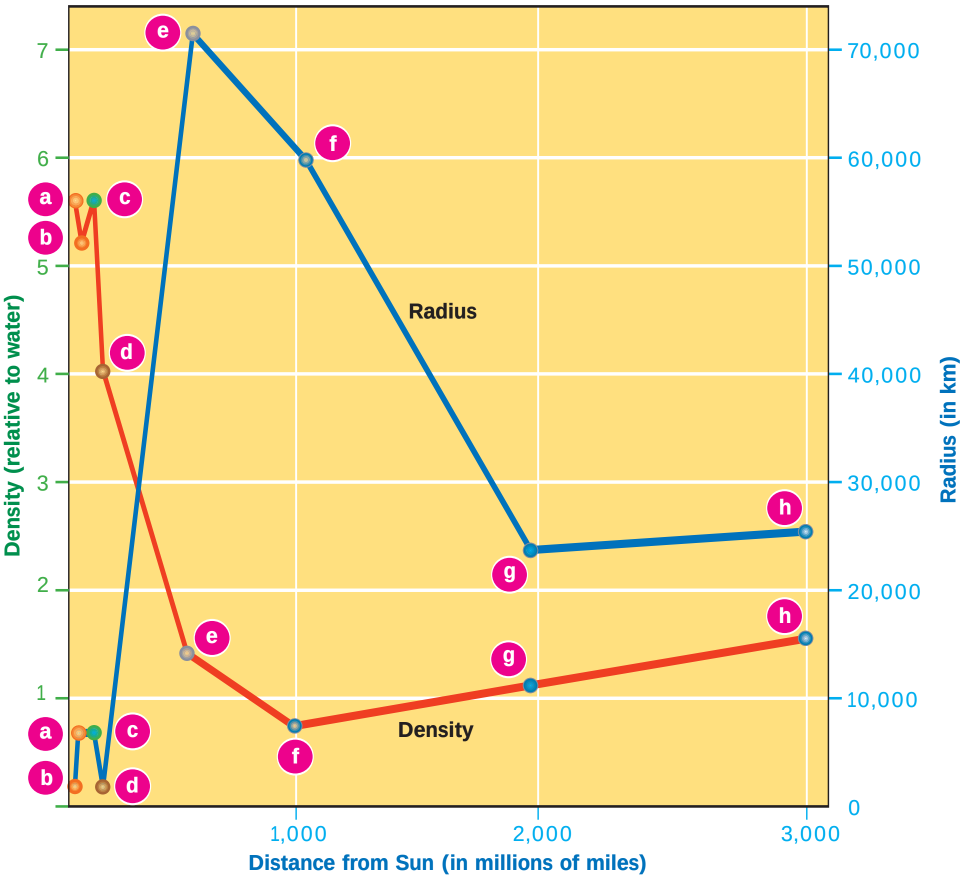

a Mercury
b Venus
c Earth
d Mars
e Jupiter
f Saturn
g Uranus
h Neptune

2 Atmospheric composition of the inner planets

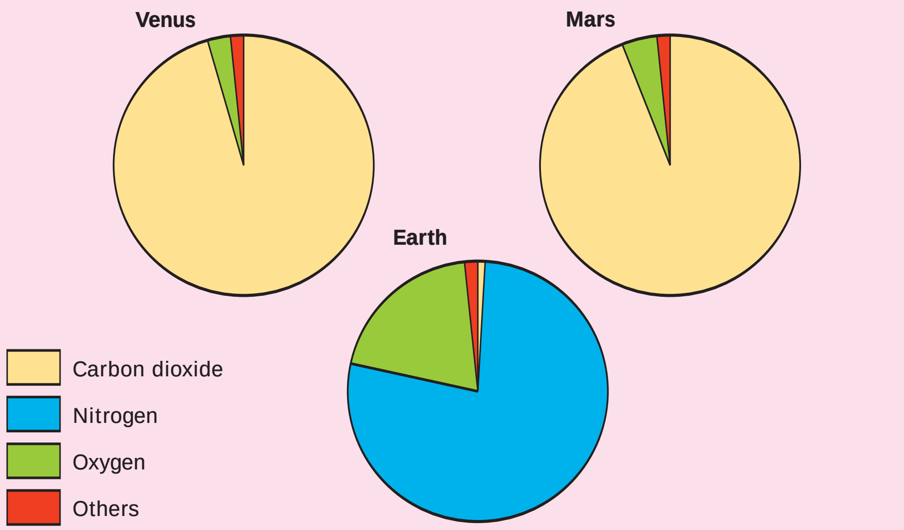

Atomic structure

1 Principle subatomic particles

Particle	Relative atomic mass	Relative charge
Electron	$\frac{1}{1836}$	–1
Neutron	1	0
Proton	1	1

2 The atom

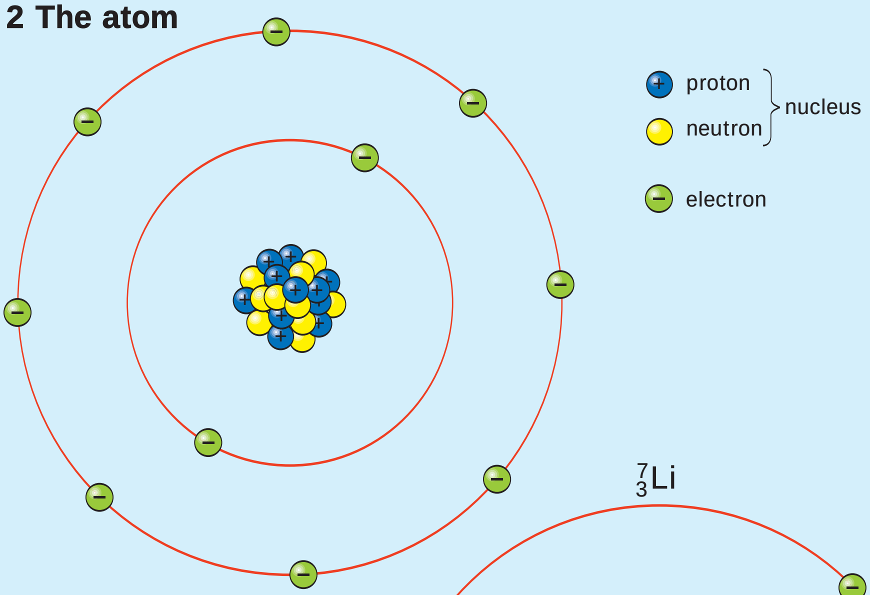

3 Representing an element

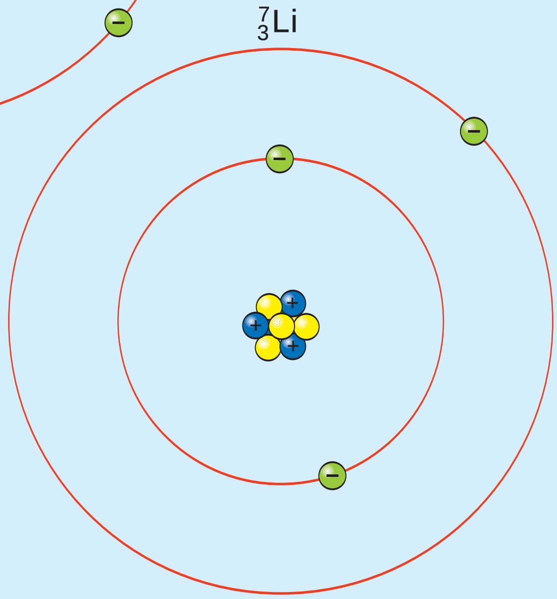

4 Isotopes

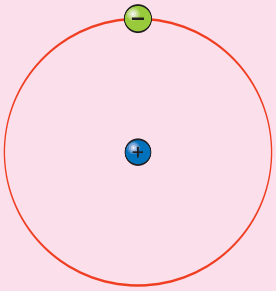
Hydrogen 1

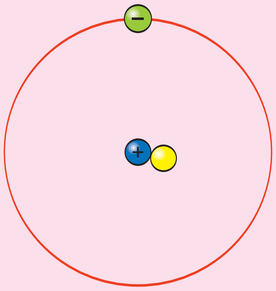
Hydrogen 2

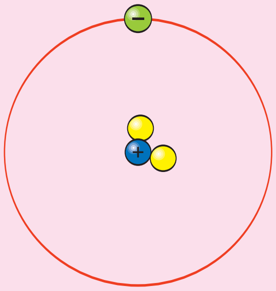
Hydrogen 3

Key words

atom
atomic number
electron
isotope
mass number
neutron
nucleus
proton
subatomic particle

1 Principle subatomic particles

- An *atom* is the smallest particle of an element. It is made up of even smaller *subatomic particles*: negatively charged *electrons*, positively charged *protons*, and *neutrons*, which have no charge.

2 The atom

- An atom consists of a *nucleus* of protons and neutrons surrounded by a number of electrons.
- Most of the mass of an atom is contained in its nucleus.
- The number of protons in the nucleus is always equal to the number of electrons around the nucleus. Atoms have no overall charge.

3 Representing an element

- Elements can be represented using their *mass number*, *atomic number*, and atomic symbol:

$^{\text{mass number}}_{\text{atomic number}}\text{Symbol}$

- The atomic number of an atom is the number of protons in its nucleus.
- The mass number is the total number of protons and neutrons in its nucleus. Thus, an atom of one form of lithium (Li), which contains three protons and four neutrons, can be represented as: $^{7}_{3}\text{Li}$

4 Isotopes

- All atoms of the same element have the same atomic number; however, they may not have the same mass number because the number of neutrons may not always be the same. Atoms of an element that have different mass numbers are called isotopes. The diagram at left illustrates isotopes of hydrogen.

© Diagram Visual Information Ltd.

Key words

alpha particle
atom
atomic mass

Geiger and Marsden's apparatus

Developing the atomic model

- At end of the 19th century, scientists thought that the *atom* was a positively charged blob with negatively charged electrons scattered throughout it. At the suggestion of British physicist Ernest Rutherford, Johannes Geiger and Earnest Marsden conducted an experiment that changed this view of the atomic model.
- Scientists had recently discovered that some elements were radioactive—they emitted particles from their nuclei as a result of nuclear instability. One type of particle, alpha radiation, is positively charged. Geiger and Marsden investigated how *alpha particles* scattered by bombarding them against thin sheets of gold, a metal with a high *atomic mass*.
- They used a tube of radon, a radioactive element, in a metal block (a) as the source of a narrow beam of alpha particles and placed a sheet of gold foil in the center of their apparatus (b). After they bombarded the sheet, they detected the pattern of alpha particle scattering by using a fluorescent screen (c) placed at the focal length of a microscope (d).
- If the existing model had been correct, all of the particles would have been found within a fraction of a degree of the beam. But Geiger and Marsden found that alpha particles were scattered at angles as large as 140°.
- From this experiment, Rutherford deduced that the positively charged alpha particles had come into the repulsive field of a highly concentrated positive charge at the center of the atom. He, therefore, concluded that an atom has a small dense nucleus in which all of the positive charge and most of the mass is concentrated. Negatively charged electrons surround the nucleus—similar to the way the planets orbit the Sun.

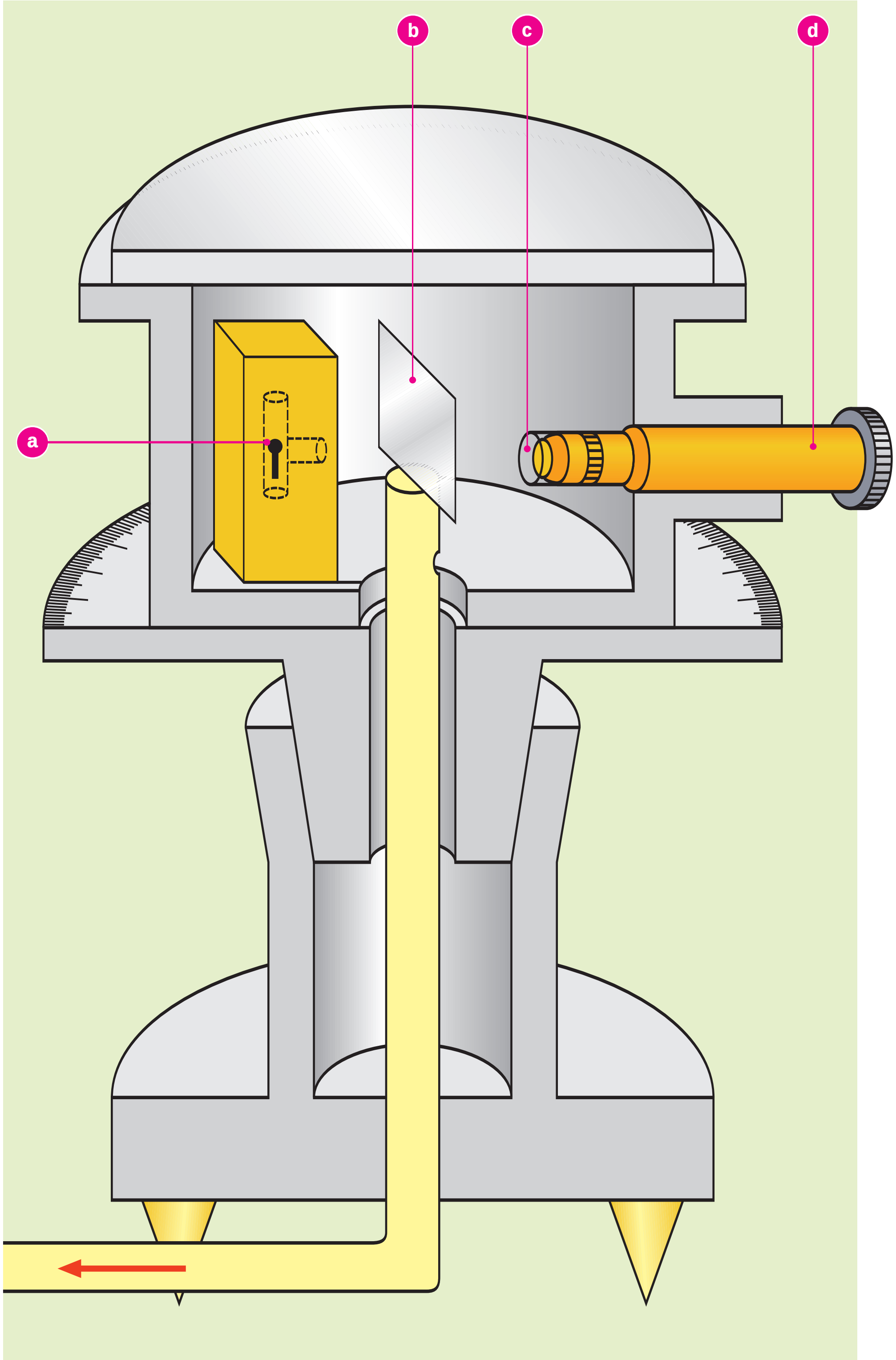

a source of alpha particles (radon tube)
b gold foil
c screen
d microscope

© Diagram Visual Information Ltd.

Investigating the electron 1

1 Maltese-Cross tube

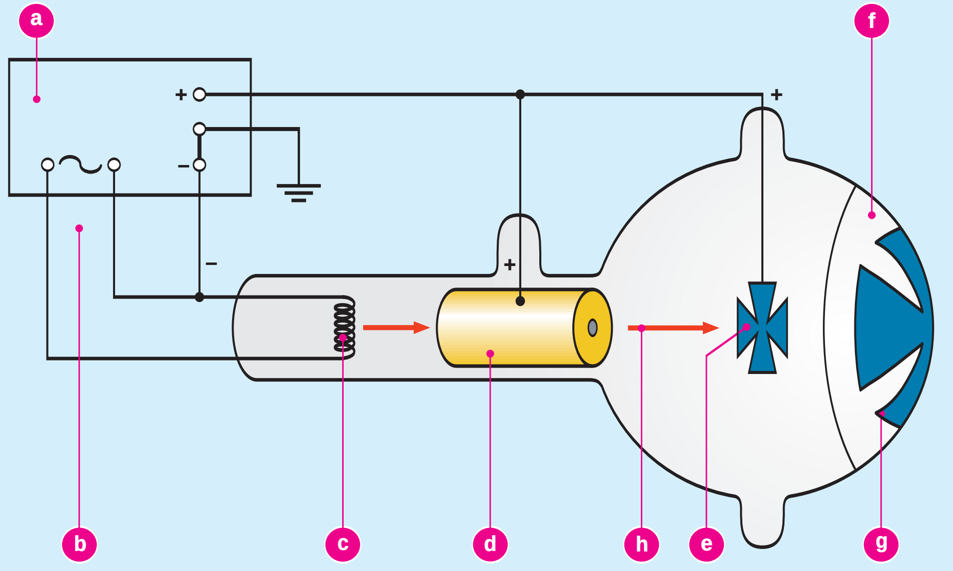

a E.h.t. supply
b low voltage
c heated filament and cathode
d anode
e Maltese-Cross (connected to anode)
g shadow
h invisible cathode rays
f fluorescent screen

2 The Perrin tube

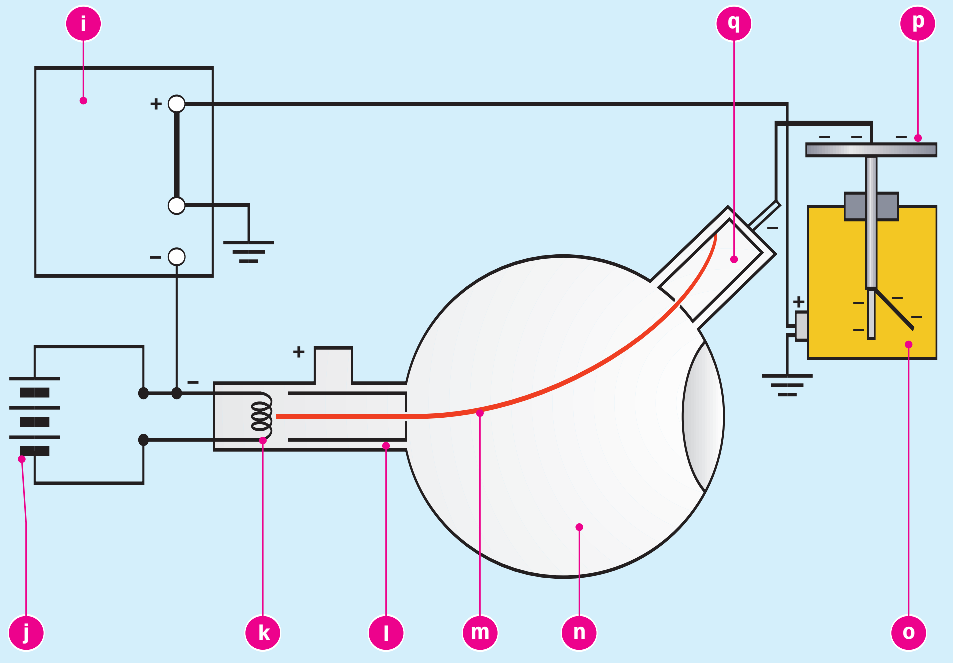

i E.h.t. supply
j 6 V supply
k cathode
l anode
m track of electron beam in magnetic field
n vacuum
o gold-leaf electroscope
p electrons are collected
q insulated metal cylinder

Key words

anode
cathode
cathode rays
electron
fluorescence

Investigating the electron

- During the last half of the nineteenth century, scientists observed that when an electric current passes through a glass tube containing a small amount of air, the air glowed. As air was removed, a patch of *fluorescence* appeared on the tube, which they called *cathode rays*. Scientists then began investigated these streams of *electrons* traveling at high speed.

1 Maltese cross tube

- In the 1880s, William Crookes experimented on cathode rays using a Maltese cross tube.
- The stream of electrons emitted by the hot *cathode* is accelerated toward the *anode*. Some are absorbed, but the majority passes through and travels along the tube. Those electrons that hit the Maltese cross are absorbed. Those electrons that miss the cross strike the screen, causing it to fluoresce with a green light.
- The result of this experiment is that a shadow of the cross is cast on the screen. This provides evidence that cathode rays travel in straight lines.

2 The Perrin tube

- In 1895 Jean Perrin devised an experiment to demonstrate that cathode rays convey negative charge.
- He constructed a cathode ray tube in which the cathode rays were accelerated through the anode, in the form of a cylinder open at both ends, into an insulated metal cylinder called a Faraday cylinder.
- This cylinder has a small opening at one end. Cathode rays enter the cylinder and build up charge, which is indicated by the electroscope. Perrin found that the electroscope had become negatively charged.
- Perrin's experiments helped to prepare the way for English physicist J. J. Thompson's work on electrons a few years later.

© Diagram Visual Information Ltd.

Investigating the electron 2

Key words

anode
cathode
cathode rays
electron
photoelectric effect
radiation

1 J.J. Thomson's cathode ray tube

- In 1897 J.J. Thomson devised an experiment with *cathode rays* that resulted in the discovery of the *electron*.
- Up to this time, it was thought that the hydrogen atom was the smallest particle in existence. Thomson demonstrated that electrons (which he called corpuscles) comprising cathode rays were nearly 2,000 times smaller in mass than the then lightest-known particle, the hydrogen ion.
- When a high voltage is placed across a pair of plates, they become charged relative to each other. The positively charged plate is the *anode*, and the negatively charged plate the *cathode*.
- Electrons pass from the surface of the cathode and accelerate toward the oppositely charged anode. The anode absorbs many electrons, but if the anode has slits, some electrons will pass through.
- The electrons travel into an evacuated tube, where they move in a straight line until striking a fluorescent screen. This screen is coated with a chemical that glows when electrons strike it.

2 Evidence of the photoelectric effect

- The *photoelectric effect* is the emission of electrons from metals upon the absorption of electromagnetic *radiation*.
- Scientists observed the effect in the nineteenth century, but they could not explain it until the development of quantum physics.
- To observe the effect, a clean zinc plate is placed in a negatively charged electroscope. The gold leaf and brass plate carry the same negative charge and repel each other.
- When ultraviolet radiation strikes the zinc plate, electrons are emitted. The negative charge on the electroscope is reduced, and the gold leaf falls.

1 J.J. Thomson's cathode ray tube

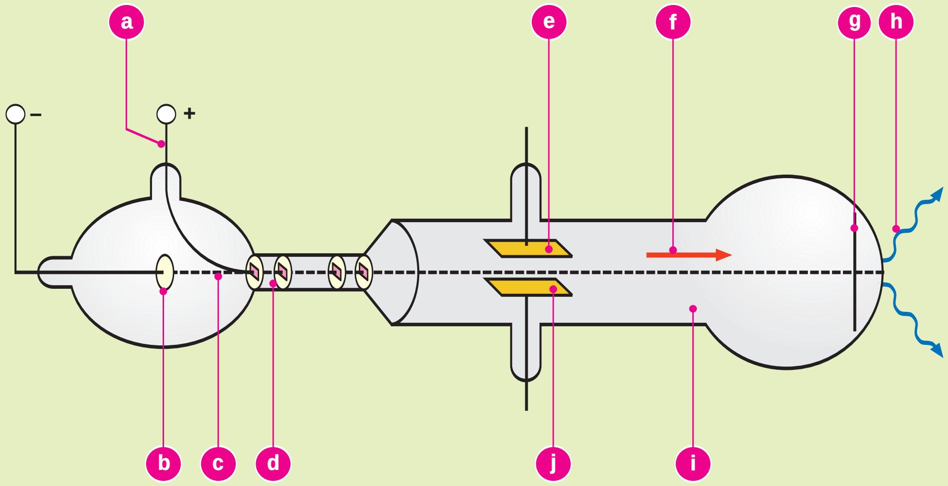

a high voltage
b cathode
c gas discharge provides free electrons
d anode with slit
e y-deflecting plate
f direction of travel of the cathode rays
g flourescent screen
h light
i evacuated tube
j x-deflecting plate

2 Evidence of the photoelectric effect

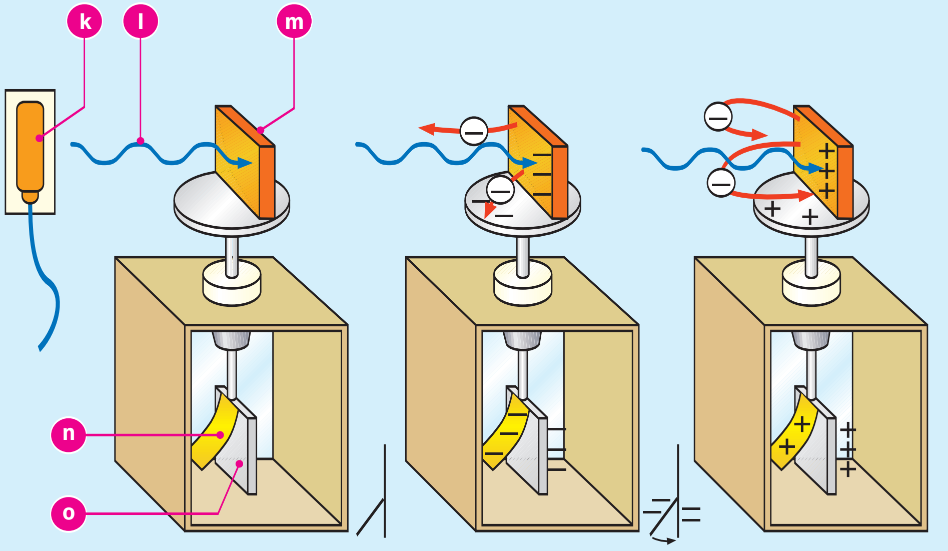

Negatively charged electroscope with zinc plate attached

The leaf falls as electrons are ejected from the zinc plate

If positively charged the electroscope remains charged

k mercury vapor lamp
l ultraviolet light
m brass plate
n gold leaf
o zinc plate

© Diagram Visual Information Ltd.

Cathode ray oscilloscope

1 The cathode ray oscilloscope

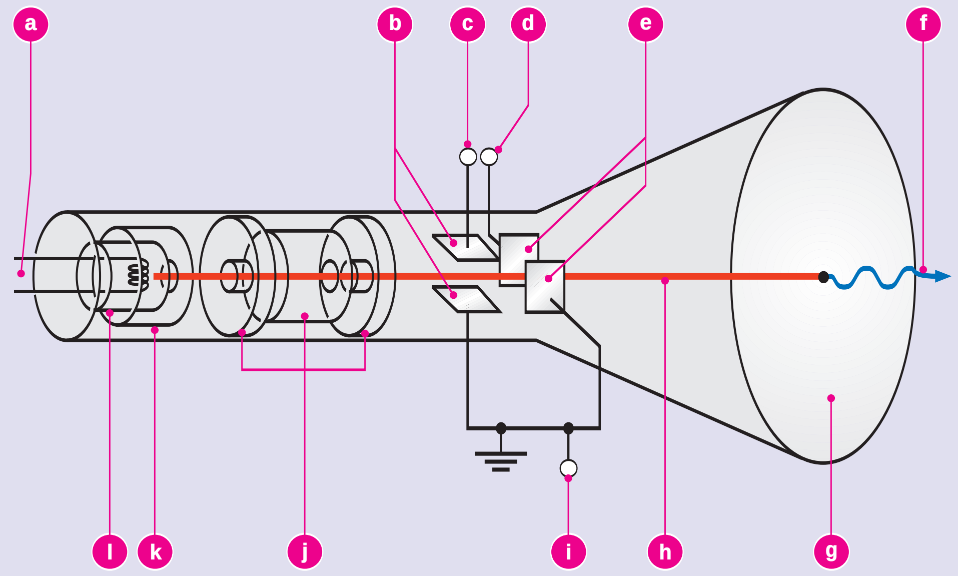

a heater
b y-deflection plates
c y-input terminal
d x-input terminal
e x-deflection plates
f light
g phosphor coating
h electron beam
i common-input terminal
j accelerating and focusing anodes
k grid
l cathode

2 Electron gun

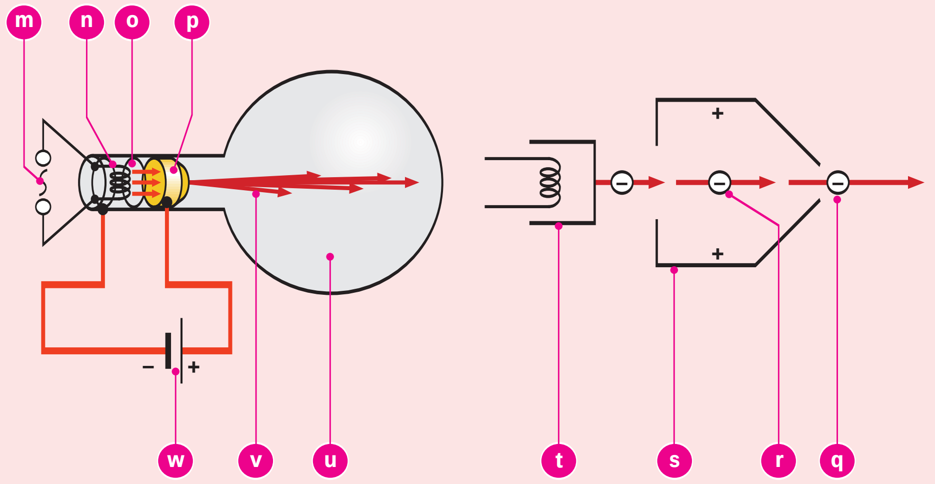

m low voltage
n heater
o cathode
p cyclindrical anode
q high speed electrons
r accelerated electrons
s anode
t cathode
u evacuated tube
v electron beam
w high voltage

Key words

anode
cathode
cathode rays

1 Cathode ray oscilloscope

- The *cathode ray* oscilloscope (CRO) is one of the most important scientific instruments ever to be developed. It is often used as a graph plotter to display a waveform showing how potential difference changes with time. The CRO has three essential parts: the electron gun, the deflecting system, and the fluorescent gun.
- The electron gun consists of a heater and *cathode*, a grid, and several *anodes*. Together these provide a stream of cathode rays. The grid is at negative potential with respect to the cathode and controls the number of electrons passing through its central hole. It is the brightness control.
- The deflecting system consists of a pair of deflecting plates across which potential differences can be applied. The Y-plates are horizontal but deflect the beam vertically. The X-plates are vertical and deflect the bean horizontally.
- A bright spot appears on the fluorescent screen where the beam hits it.

2 Electron gun

- When a current passes through the heater, electrons are emitted from the surface of the cathode and attracted towards an oppositely charged anode. Some will be absorbed by the anode, while others pass through and are accelerated, forming a stream of high-speed electrons.

© Diagram Visual Information Ltd.

Key words

electron
radiation

Measuring the charge on the electron

Measuring the charge on the electron

- In the early part of the 20th century, American physicist Robert Millikan constructed an experiment to accurately determine the electric charge carried by a single *electron*.
- Millikan's apparatus consisted of two horizontal plates about 20 cm in diameter and 1.5 cm apart, with a small hole in the center of the upper plate.
- At the beginning of the experiment, an atomizer sprayed a fine mist of oil on to the upper plate.
- As a result of gravity, a droplet would pass through the hole in the plate into a chamber that was ionized by *radiation*. Electrons from the air attached themselves to the droplet, causing it to acquire a negative charge. A light source illuminated the droplet, making it appear as a pinpoint of light. Millikan then measured its downward velocity by timing its fall through a known distance.
- Millikan measured hundreds of droplets and found that the charge on them was always a simple multiple of a basic unit, 1.6 x 10^{-19} coulomb. From this he concluded that the charge on an electron was numerically 1.6 x 10^{-19} coulomb.

© Diagram Visual Information Ltd.

Millikan's apparatus

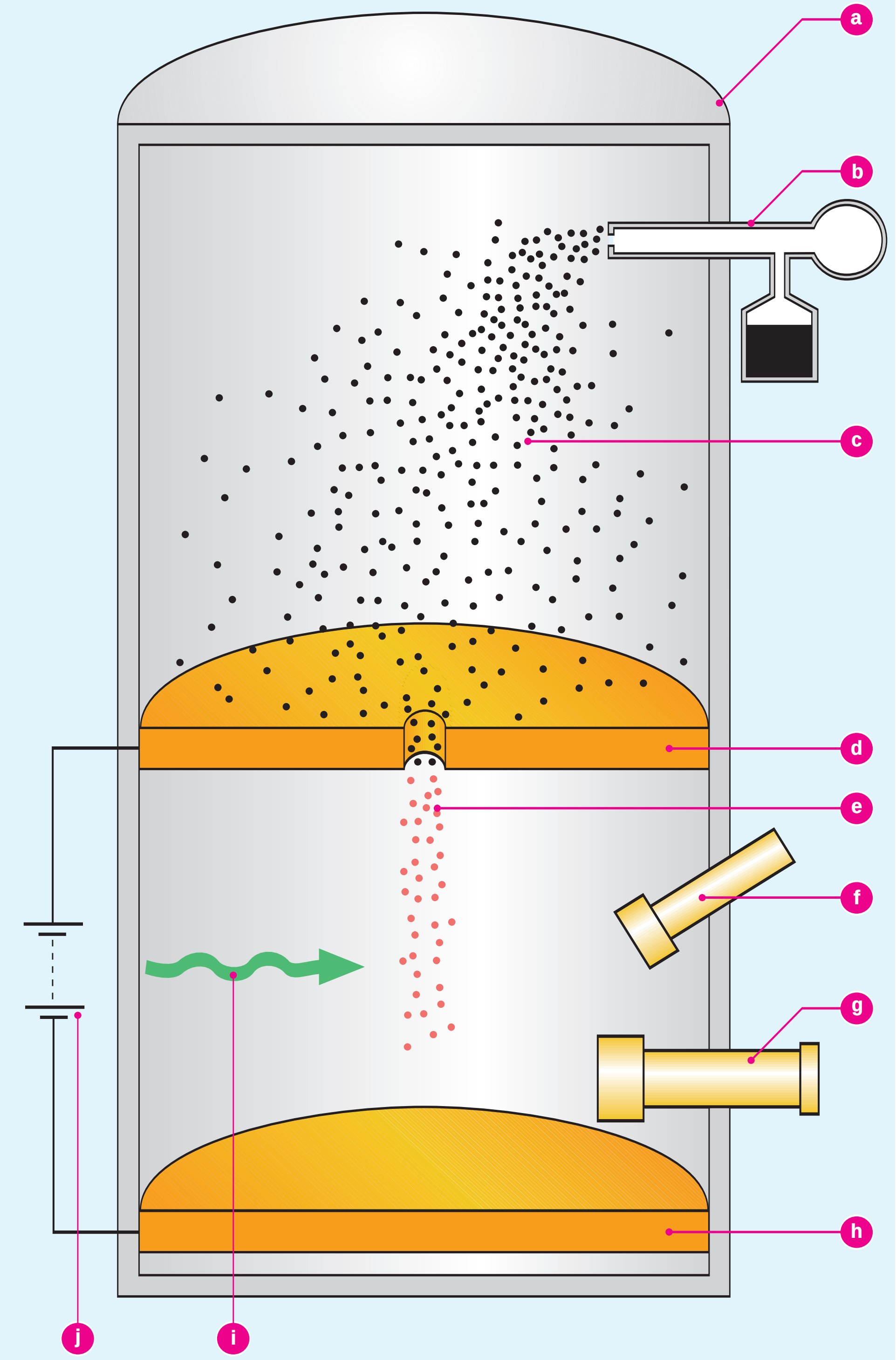

a sealed container
b atomizer
c oil droplets
d charged metal plate (+)
e charged oil droplets
f light source
g viewing microscope
h charged metal plate (–)
i ionizing radiation
j power source

Size and motion of molecules

1 Estimating the size of a molecule

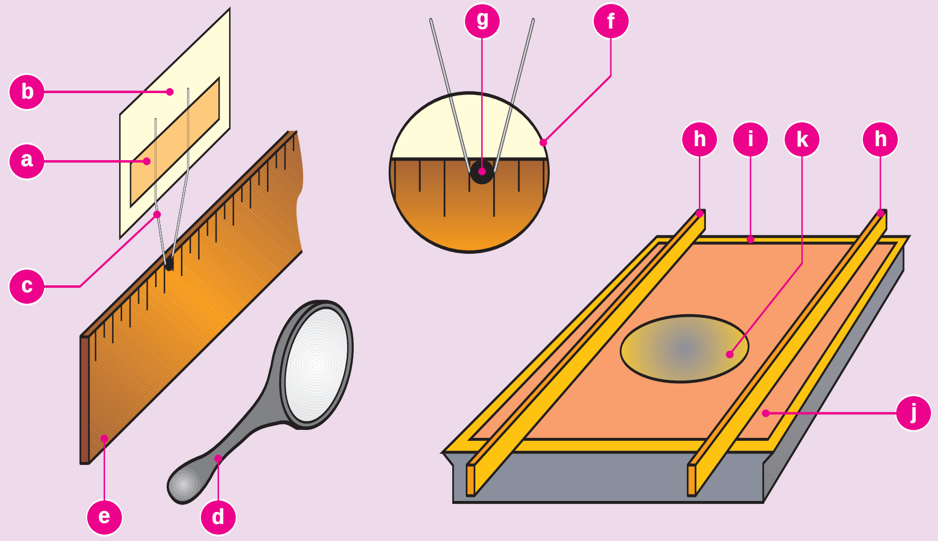

Determining the radius of an oil drop

Determining the radius of an oil drop spread

2 Brownian motion in air

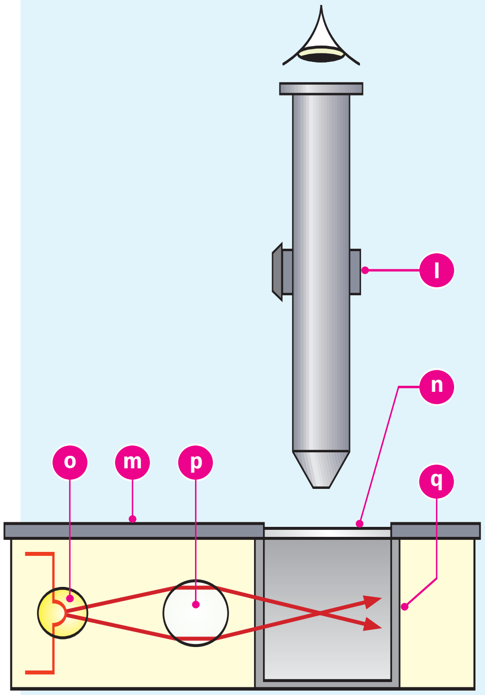

3 Diffusion

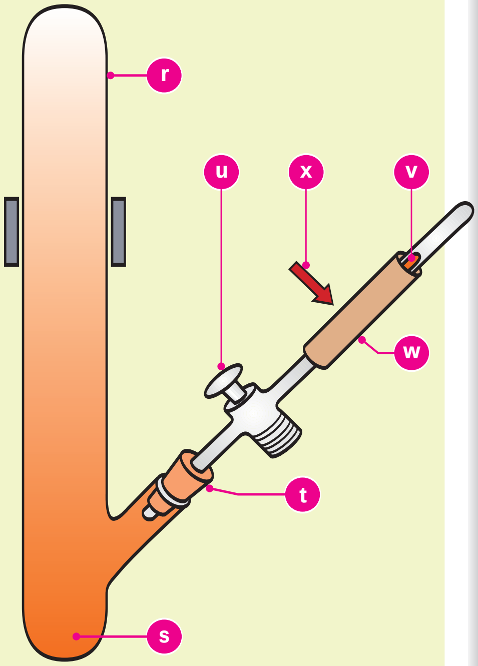

a tape
b cardboard
c fine stainless steel wire
d magnifying glass
e 1/2 mm scale
f view through magnifying glass
g oil drop
h waxed sticks
i wax-coated tray
j lycopodium powder
k oil patch
l microscope
m removable lid
n window
o lamp
p glass rod for converging light
q glass smoke chamber
r glass diffusion tube
s liquid bromine capsule
t rubber stopper
u tap
v bromine capsule
w rubber tube
x point at which pressure is applied to break capsule

Key words

Brownian motion
diffusion
molecule

1 Estimating the size of a molecule

- Scientists can estimate the size of a *molecule* by dividing the volume of a sphere by the volume of a cylinder.
- In the example in the diagram, the volume of a spherical oil drop of radius, r_s, is given by:

$$\frac{4 \times \pi \times r_s^3}{3}$$

- When the oil drop spreads across the surface of water, it takes the shape of a cylinder of radius, r_c, and thickness, h. The volume of such a cylinder is:

$$\pi \times r_c^2 \times h$$

- If we assume that the layer of oil is one molecule thick, then h gives the size of an oil molecule.
- When spread on water the drop of oil will have the same volume therefore:

$$h = \frac{4 \times \pi \times r_s^3}{3} \times \frac{1}{\pi \times r_c^2}$$

$$h = \frac{4\ r_s^3}{3\ r_c^2}$$

2 Brownian motion in air

- *Brownian motion* is the random motion of particles through a liquid or gas. Scientists can observe this by using a glass smoke chamber.
- Smoke consists of large particles that can be seen using a microscope.
- In the smoke chamber, the smoke particles move around randomly due to collisions with air particles.

3 Diffusion

- *Diffusion* is the spreading out of one substance through another due to the random motion of particles.
- The diagram illustrates how scientists use a diffusion tube to observe this. Initially the color of the substance is strongest at the bottom of the tube.
- After a period of time, as a result of diffusion, the particles of the substance mix with air particles, and the color becomes uniform down the length of the tube.

© Diagram Visual Information Ltd.

Determination of Avogadro’s constant

Key words

anode
Avogadro’s constant
electrolysis
Faraday constant
mole

Defining Avogadro’s constant

- *Avogadro’s constant* is the number of particles in a *mole* of a substance. It equals 6.023 x 10^{23} mol^{-1}.
- It is F, the *Faraday constant*—96,500 coulombs per mole, the amount of electric charge of one mole of electrons—divided by 1.60 x 10^{-19} coulomb—the charge on one electron (expressed as e).
- Thus, the Avogadro constant, N, is given by: $N = \frac{F}{e}$

or:

$$\frac{96{,}500}{1.60 \times 10^{-19}} = 6.023 \times 10^{23}\ \mathrm{mol}^{-1}$$

Determining the Constant

- The number of molecules in one mole of substance can be determined by using electrochemistry.
- During *electrolysis*, current (electron flow) over time is measured in an electrolytic cell (see diagram). The number of atoms in a weighed sample is then related to the current to calculate Avogadro’s constant.

Illustrating the Procedure

- The diagram illustrates the electrolysis of copper sulfate. To calculate Avogadro’s constant, the researcher weighs the rod to be used as the *anode* before submerging the two copper rods in copper sulfate. She then connects the rods to a power supply and an ammeter (an instrument used to measure electric current). She measures and records the current at regular intervals and calculates the average amperage (the unit of electric current). Once she turns off the current, she weighs the anode to see how much mass was lost. Using the figures for anode mass lost, average current, and duration of the electrolysis, she calculates Avogadro’s constant.

© Diagram Visual Information Ltd.

Determination of Avogadro’s constant

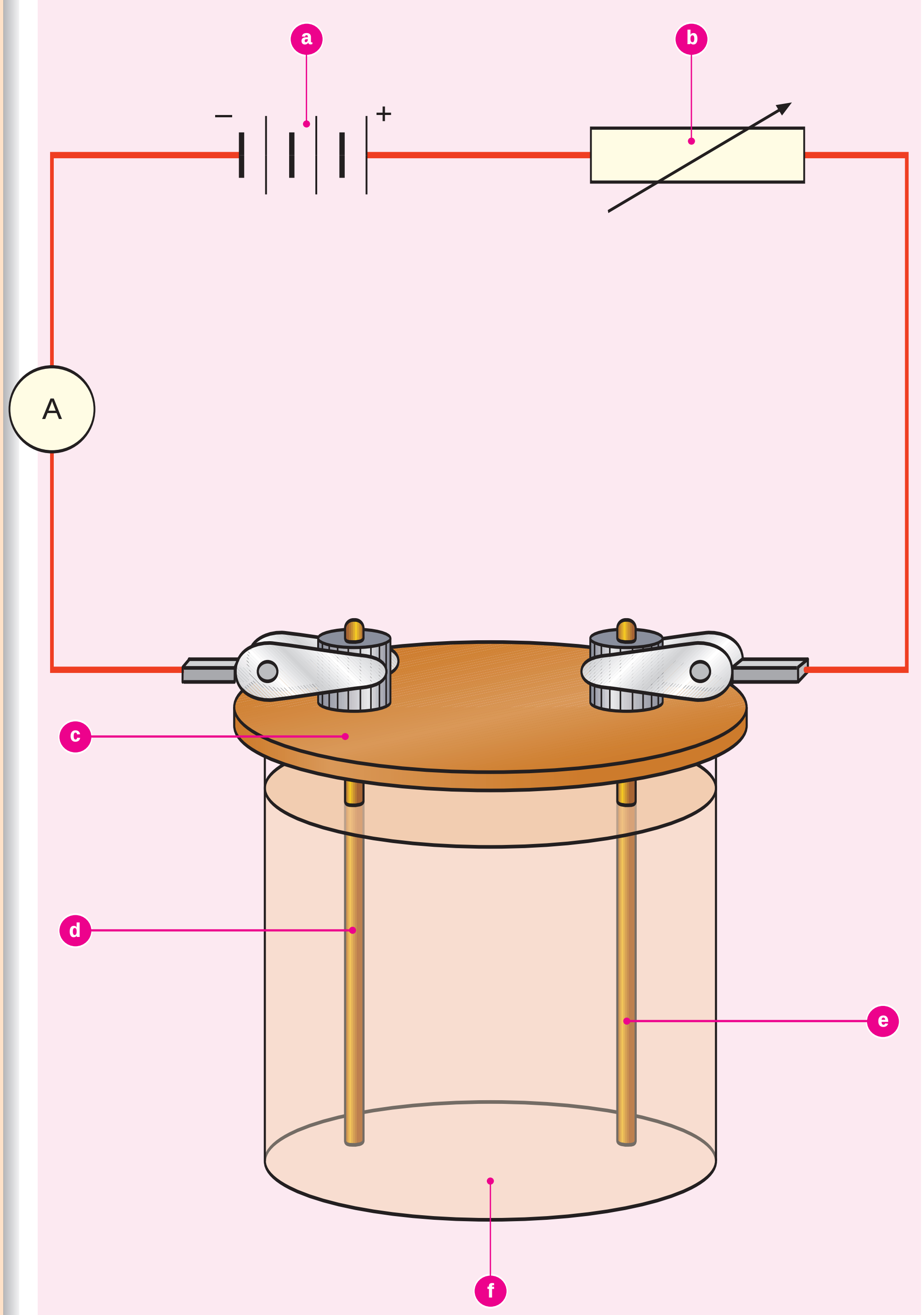

a power supply with ammeter
b rheostat
c hardboard or wooden electrode holder
d copper rod cathode
e copper rod anode
f copper sulfate solution

The mole

1 Defining a mole

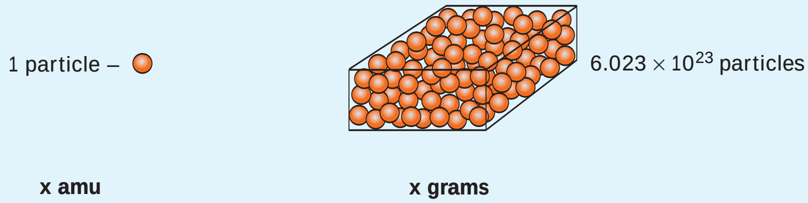

2 Moles of gas

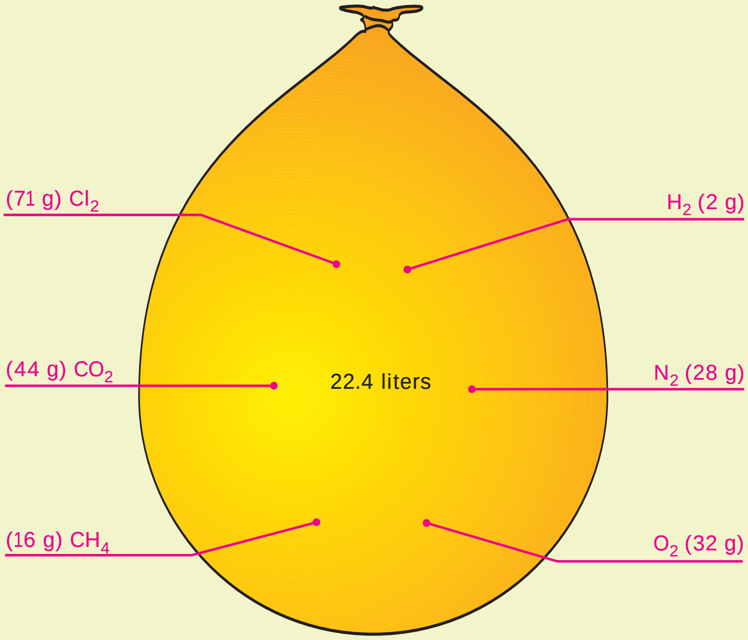

3 Molarity

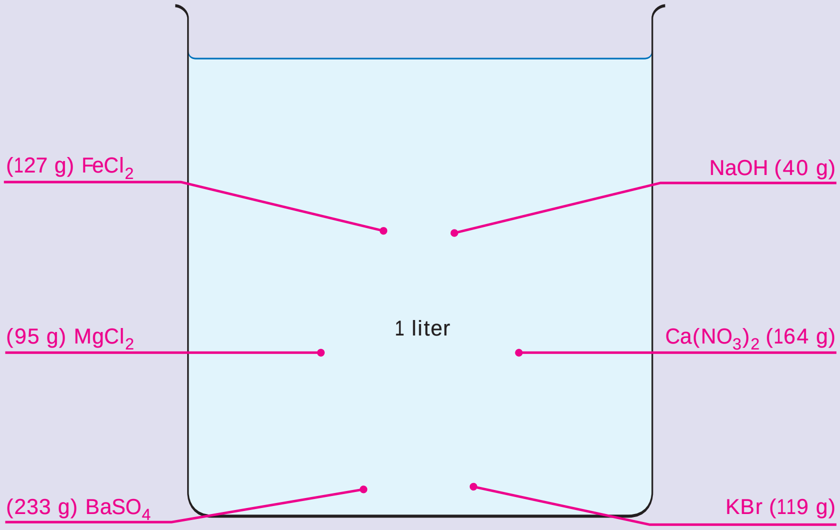

Key words

atom
ion
molarity
mole
molecule

1 Defining a mole

- Because *atoms*, *ions*, and *molecules* have very small masses, it is impossible to count or weigh them individually. As a result, scientists use *moles* in a chemical reaction.
- A mole is the amount of substance that contains as many elementary entities (atoms, molecules, ions, any group of particles) as there are atoms in exactly 0.012 kilogram of carbon-12. This quantity is Avogadro's constant (6.023 x 10^{23} mol^{-1}).
- The significance of this number is that it scales the mass of a particle in atomic mass units (amu) exactly into grams (g).
- Chemical equations usually imply that the quantities are in moles.

2 Moles of gas

- One mole of any gas occupies 22.4 liters at standard temperature and pressure, (which is 0 °C and atmospheric pressure).
- The diagram shows the mass in grams of one mole of the following gases: chlorine (Cl_2), carbon dioxide (CO_2), methane (CH_4), hydrogen (H_2), nitrogen (N_2), and oxygen (O_2).

3 Molarity

- *Molarity* is concerned with the concentration of a solution. It indicates the number of particles in 1 liter of solution.
- A 1 molar solution contains 1 mole of a substance dissolved in water or some other solvent to make 1 liter of solution.
- The diagram shows the mass in grams of one mole of the following substances: iron(II) chloride ($FeCl_2$), magnesium chloride ($MgCl_2$), barium sulfate ($BaSO_4$), sodium hydroxide (NaOH), calcium nitrate ($Ca(NO_3)_2$), and potassium bromide (KBr).

© Diagram Visual Information Ltd.

Key words

atomic emission spectrum
infrared
spectrum
ultraviolet
wavelength

Atomic spectrum

- The *atomic emission spectrum* of an element is the amount of electromagnetic radiation it emits when excited. This pattern of wavelengths is a discrete line *spectrum*, not a continuous spectrum. It is unique to each element.

Investigating hydrogen

- Toward the end of the nineteenth century, scientists discovered that when excited in its gaseous state, an element produces a unique spectral pattern of brightly colored lines. Hydrogen is the simplest element and, therefore, was the most studied. Hydrogen has three distinctively observable lines in the visible spectrum—red, blue/cyan, and violet.

Series

- In 1885 Swiss mathematician and physicist Johannes Jakob Balmer proposed a mathematical relationship for lines in the visible part of the hydrogen emission spectrum that is now known as the Balmer series.
- The series in the *ultraviolet* region at a shorter *wavelength* than the Balmer series is known as the Lyman series.
- The series in the *infrared* region at the longer wavelength than the Balmer series is known as the Paschen series.
- The Brackett series and the Pfund series are at the far infrared end of the hydrogen emission series.

Atomic emission spectrum: hydrogen

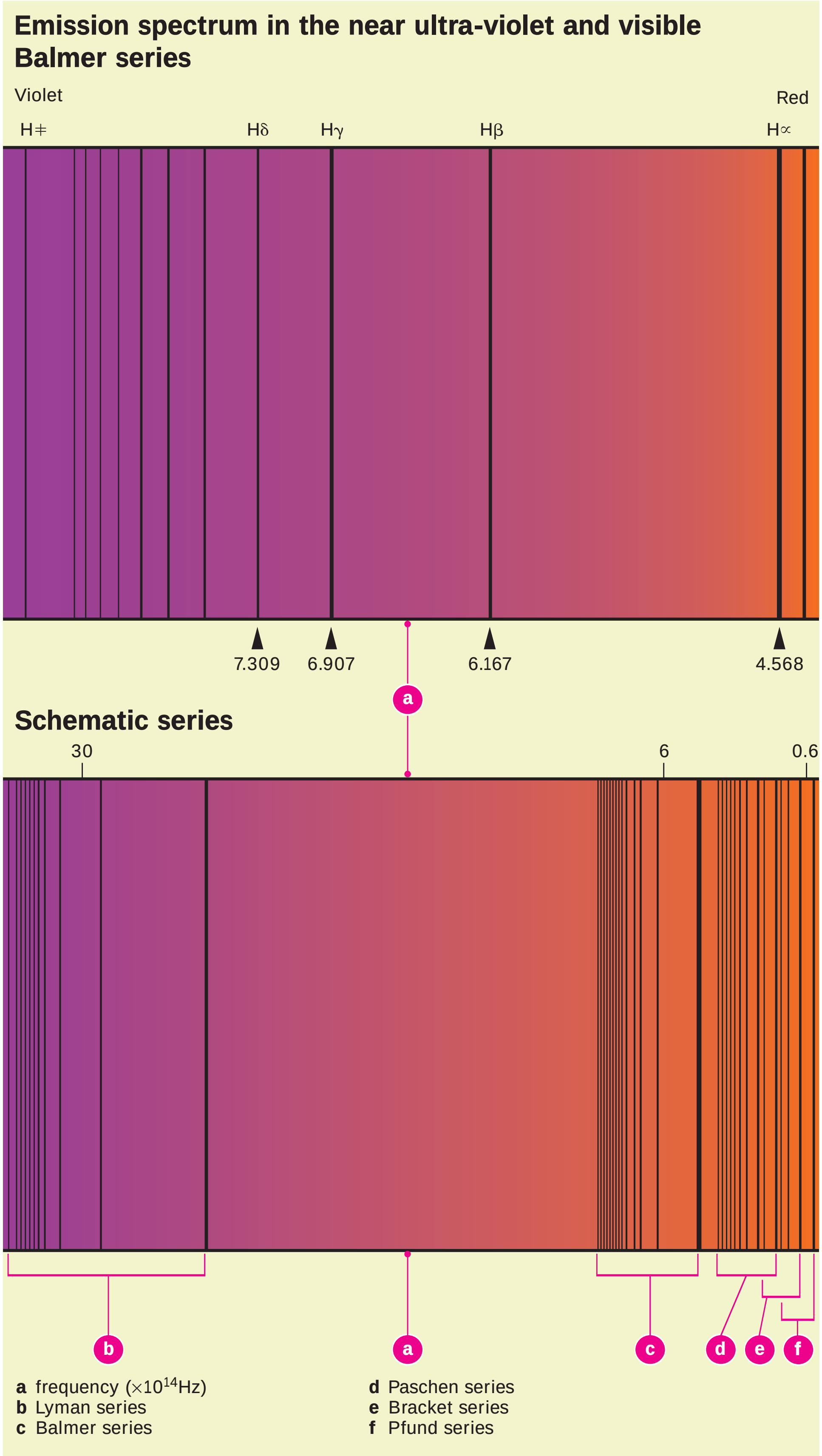
Emission spectrum in the near ultra-violet and visible Balmer series

Schematic series

a frequency ($\times 10^{14}$Hz)
b Lyman series
c Balmer series
d Paschen series
e Bracket series
f Pfund series

© Diagram Visual Information Ltd.

Energy levels: hydrogen atom

Energy-level schematic

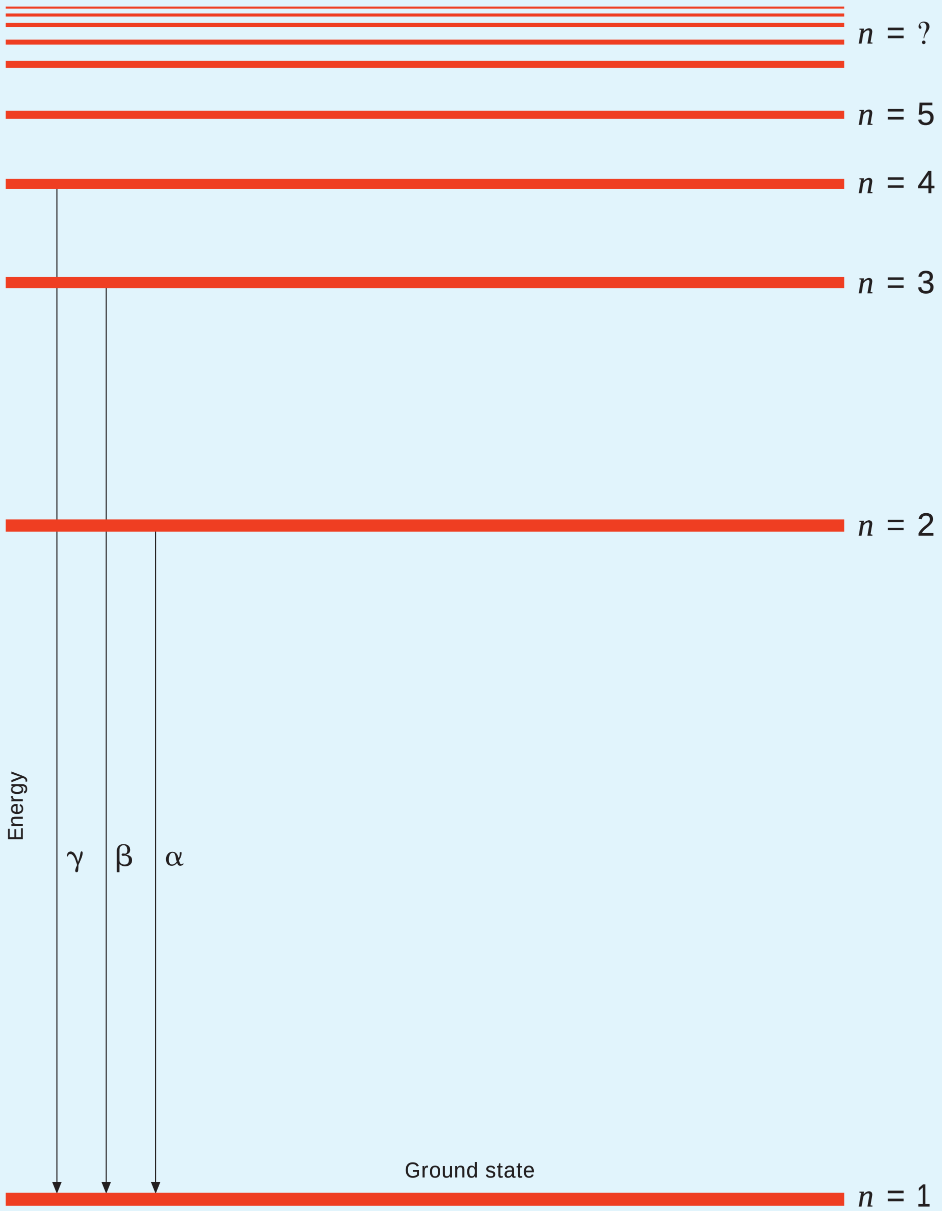

Line spectrum

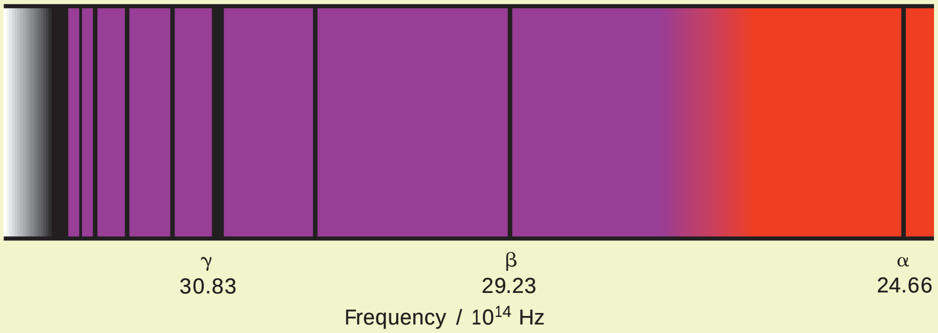

Key words

ground state
orbital
quantum number
shell
ultraviolet

Energy levels

- Electrons are arranged in definite energy levels (also called *shells* or *orbitals*), at a considerable distance from the nucleus.
- Electrons jump between the orbits by emitting or absorbing energy.
- The energy emitted or absorbed is equal to the difference in energy between the orbits.

Energy levels of hydrogen

- The graph shows the energy levels for the hydrogen atom. Each level is described by a *quantum number* (labeled by the integer n).
- The shell closest to the nucleus has the lowest energy level. It is generally termed the *ground state*. The states farther from the nucleus have successively more energy.

Transition from n level to ground state

- Transition from n=2 to the ground state, n=1:
  Frequency =24.66 x 10^{14} Hz
- Transition from n=3 to the ground state, n=1:
  Frequency =29.23 x 10^{14} Hz
- Transition from n=4 to the ground state, n=1:
  Frequency =30.83 x 10^{14} Hz

Line spectrum

- This radiation is in the *ultraviolet* region of the electromagnetic spectrum and cannot be seen by the human eye.

© Diagram Visual Information Ltd.

Luminescence

Key words

fluorescence
luminescence
phosphorescence

1 Luminescence

- *Luminescence* is the emission of light caused by an effect other than heat.
- Luminescence occurs when a substance is stimulated by radiation and subsequently emits visible light.
- The incident radiation excites electrons, and as the electrons return to their ground state, they emit visible light.
- If the electrons remain in their excited state and emit light over a period of time, the phenomenon is called *phosphorescence*.
- If the electrons in a substance return to the ground state immediately after excitation, the phenomenon is called *fluorescence*.

2 Fluorescence

- In this diagram, a fluorescent light tube contains mercury vapor at low pressure. Electrons are released from hot filaments at each end of the tube and collide with the mercury atoms, exciting the electrons in the mercury atoms to higher energy levels. As the electrons fall back to lower energy states, photons of ultraviolet light are emitted.
- The ultraviolet photons collide with atoms of a fluorescent coating on the inside of the tube. The electrons in these atoms are excited and then return to lower energy levels, emitting visible light.

1 Luminescence

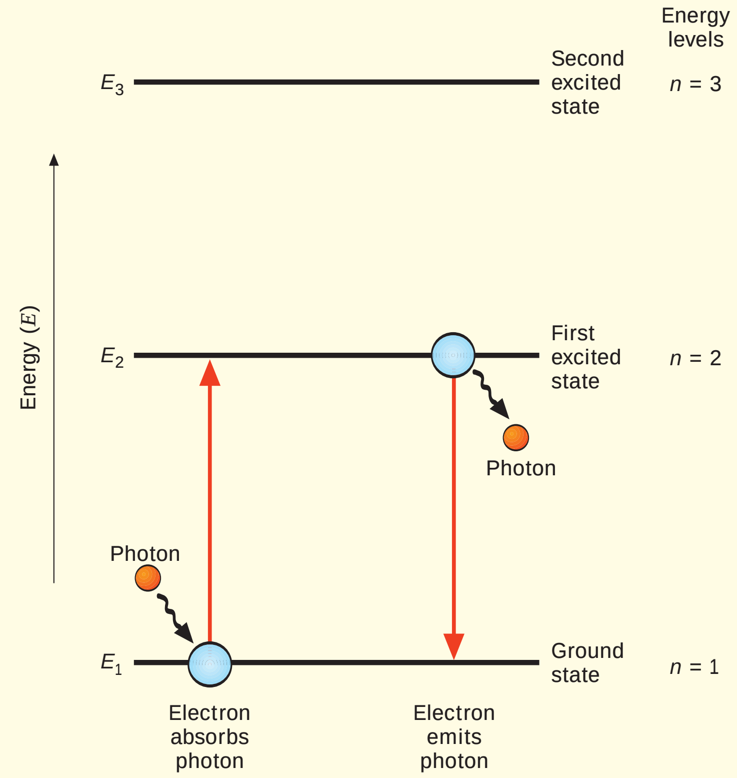
Energy levels
E_3
Second excited state
$n = 3$
Energy (E)
E_2
First excited state
$n = 2$
Photon
Photon
E_1
Ground state
$n = 1$
Electron absorbs photon
Electron emits photon

2 Fluorescence

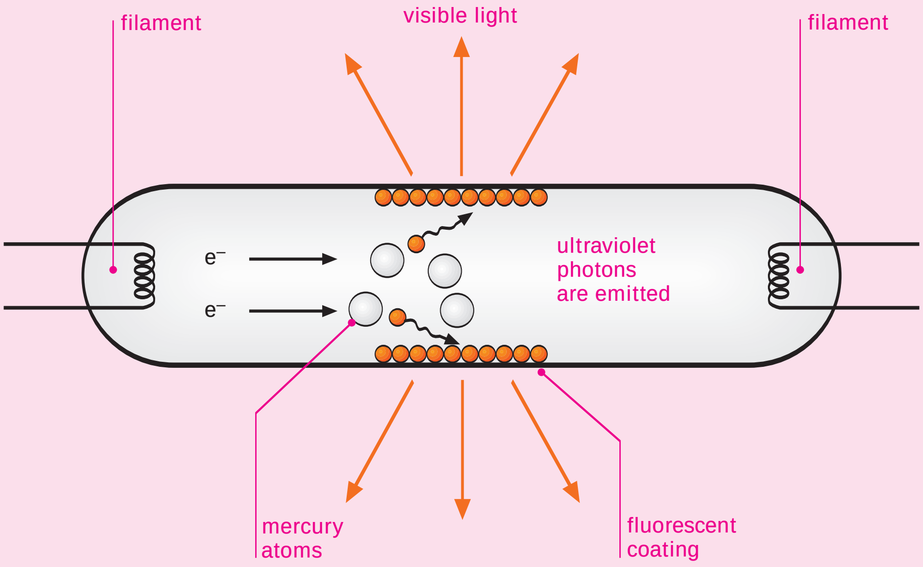
filament
visible light
filament
e⁻
e⁻
ultraviolet photons are emitted
mercury atoms
fluorescent coating

© Diagram Visual Information Ltd.

Organizing the elements

1 Antoine Lavoisier

Group 1	Group 2	Group 3	Group 4
heat light oxygen nitrogen	sulfur phosphorus	copper tin lead zinc	lime baryta magnesia alumina silica

2 Johann Dobereiner

Triad			Relative atomic mass		
Li	Na	K	7	23	39
S	Se	Te	32	79	128
Cl	Br	I	35.5	80	127
Ca	Sr	Ba	40	88	137

3 John Newlands

H	Li	Be	B	C	N	O
1	2	3	4	5	6	7
F	Na	Mg	Al	Si	P	S
8	9	10	11	12	13	14
Cl	K	Ca	Cr	Ti	Mn	Fe
15	16	17	18	19	20	21

4 Lothar Meyer

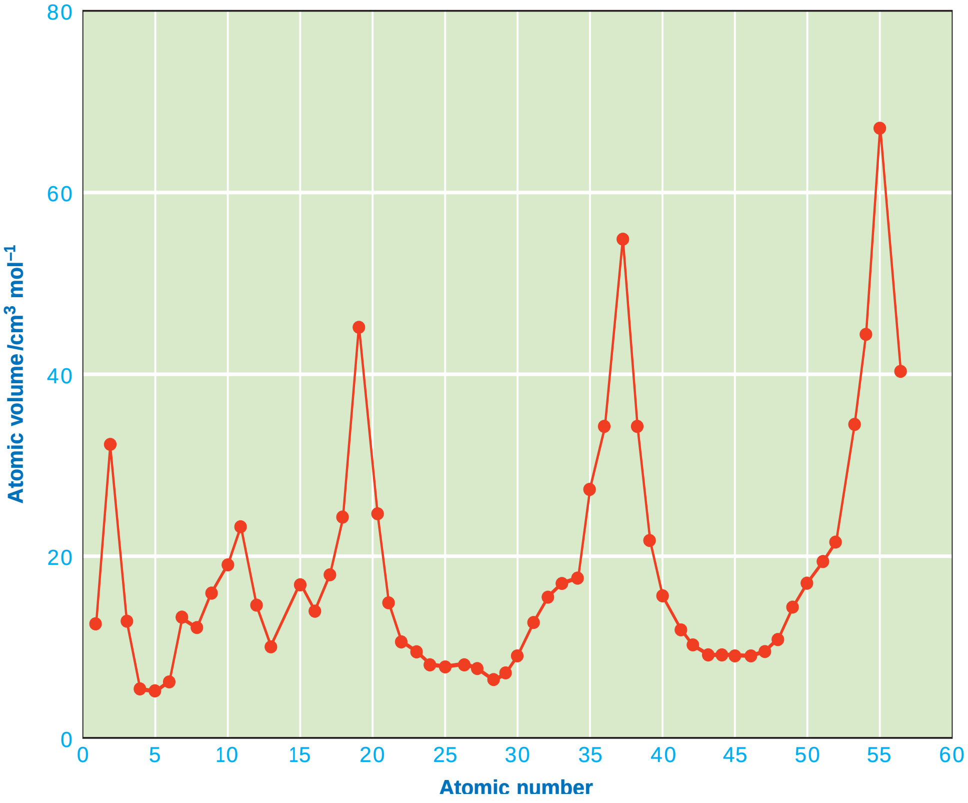

Key words

atomic mass
atomic volume
element

1 Antoine Lavoisier

- In 1789, French chemist Antoine Lavoisier organized what he believed were the *elements* into four groups: Group 1 gases, Group 2 non-metals, Group 3 metals, and Group 4 earths.

2 Johann Dobereiner

- In 1817, German chemist Johann Dobereiner noticed that the *atomic mass* of strontium was about half way between that of calcium and barium. After further study, he found that he could organize other elements into similar groups based on the same relationship to each other. He called these groups triads. Subsequently, scientists attempted to arrange all of the known elements into triads.

3 John Newlands

- In 1864, English chemist John Newlands noticed that if the elements were arranged in increasing order of atomic mass, the eighth element after any given one had similar properties. He likened this to an octave of music and called the regularity the “law of octaves.”
- Newlands’s arrangement worked well for the first 17 elements but broke down thereafter. Consequently, it was not well received by other scientists.

4 Lothar Meyer

- In 1870, German chemist Lothar Meyer plotted a graph of *atomic volume* against atomic mass.
- He found a pattern in which elements of similar properties appeared in similar positions on the graph.

© Diagram Visual Information Ltd.

The periodic table

Key words

atomic mass	*noble gases*
atomic number	*period*
element	*periodic table*
group	
group 1	

1 Mendeleyev's periodic table

- The modern *periodic table* is based on that developed by Russian chemist Dmitry Mendeleyev in the 1860s.
- He arranged the *elements* in order of increasing *atomic mass*. He called the horizontal rows *periods* and the vertical columns *groups*. He grouped the elements on the basis of their properties.
- Mendeleyev made a separate group for those elements that did not appear to fit the pattern. He also left spaces where there was no known element that fit the pattern and made predictions about the missing elements.
- There were some problems with Mendeleyev's table. For example, iodine was placed after tellurium on the basis of its chemistry, even though its atomic mass was lower than tellurium. Also, there was no obvious place for the *noble gases*. These problems were subsequently resolved when, in 1914, English physicist Henry Moseley showed that the elements could be arranged in a pattern on the basis of their *atomic number*.

2 The modern periodic table

- Metals occupy positions to the left and center, while non-metals are found to the right. Hydrogen is the exception to this pattern. The atomic structure of hydrogen would indicate that it belongs at the top left of the table; however, it is a non-metal and has very different properties from the *group 1* elements.

© Diagram Visual Information Ltd.

1 Part of Mendeleyev's periodic table

	Group							
Period	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
1	H							
2	Li	Be	B	C	N	O	F	
3	Na	Mg	Al	Si	P	S	Cl	
	K	Ca	*	Ti	V	Cr	Mn	Fe Co Ni
4	Cu	Zn	*	*	As	Se	Br	
	Rb	Sr	Y	Zr	Nb	Mo	*	Ru Rh Pd
5	Ag	Cd	In	Sn	Sb	Te	I	

Spaces were left for elements that had not been discovered. They were candium, gallium, germanium, and technetium.

2 Modern Periodic Table

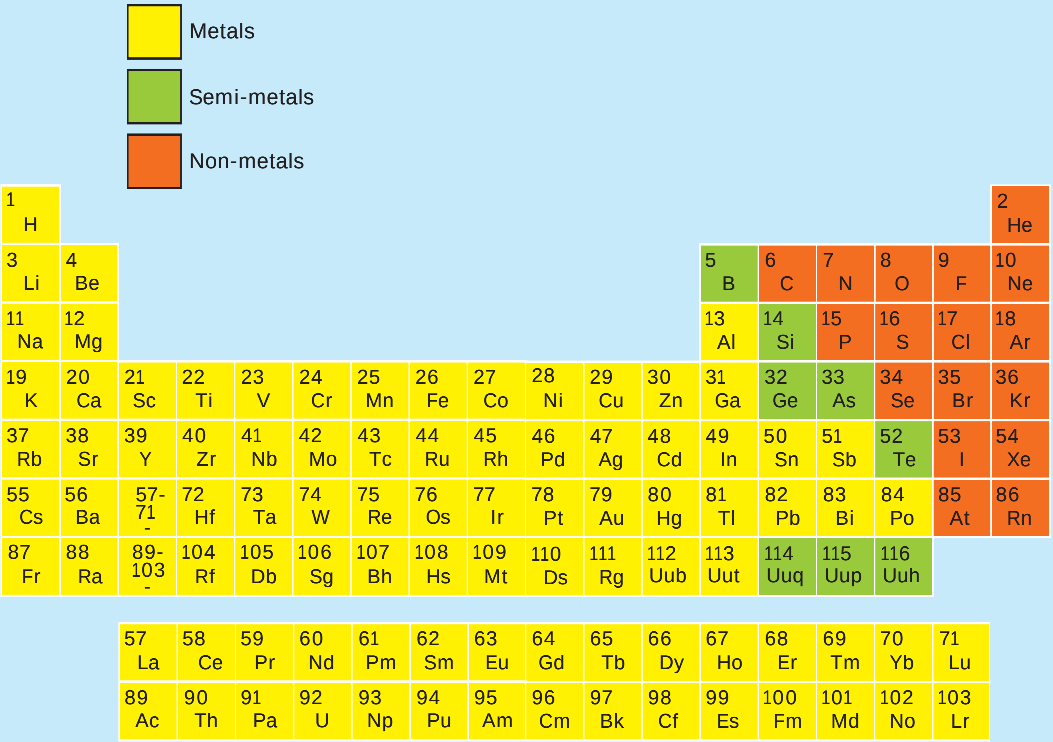

🟨 Metals

🟩 Semi-metals

🟧 Non-metals

1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	57-71	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89-103	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Uub	113 Uut	114 Uuq	115 Uup	116 Uuh		
		57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu	
		89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr	

First ionization energies of the elements

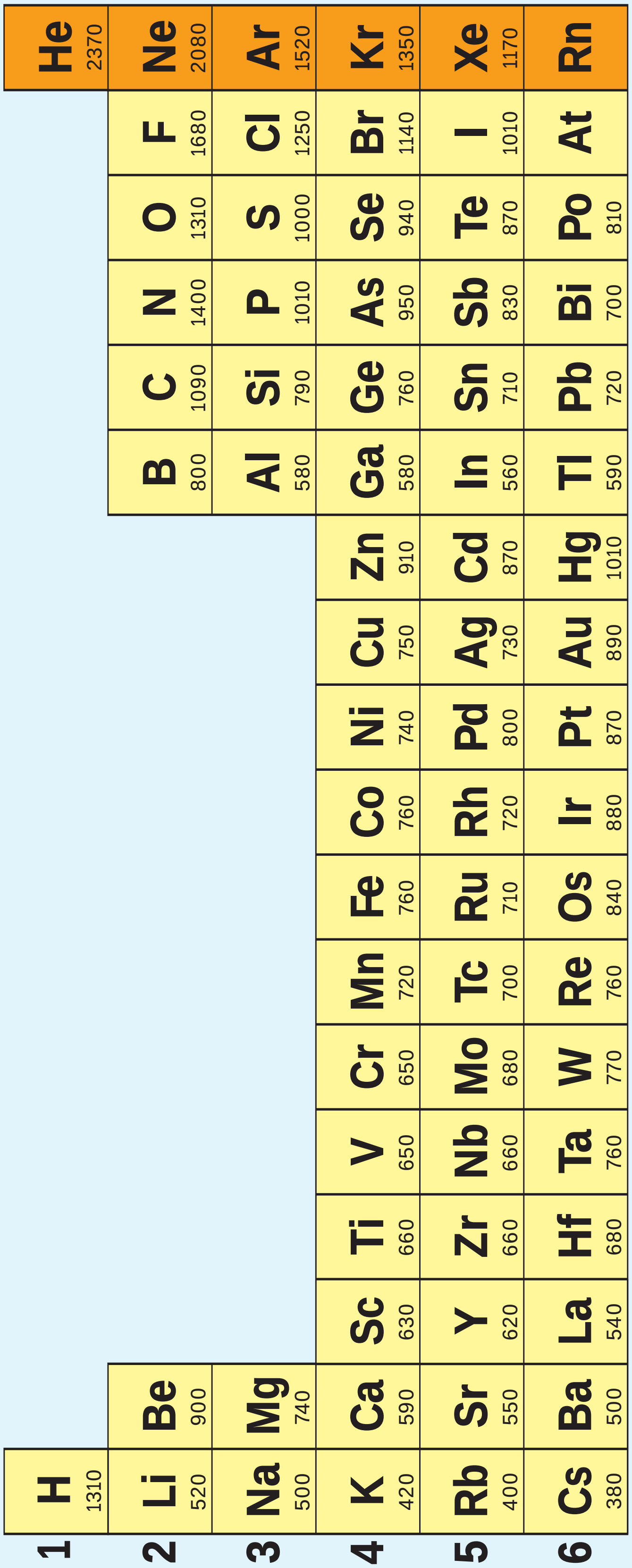

1	**H** 1310																	**He** 2370
2	**Li** 520	**Be** 900											**B** 800	**C** 1090	**N** 1400	**O** 1310	**F** 1680	**Ne** 2080
3	**Na** 500	**Mg** 740											**Al** 580	**Si** 790	**P** 1010	**S** 1000	**Cl** 1250	**Ar** 1520
4	**K** 420	**Ca** 590	**Sc** 630	**Ti** 660	**V** 650	**Cr** 650	**Mn** 720	**Fe** 760	**Co** 760	**Ni** 740	**Cu** 750	**Zn** 910	**Ga** 580	**Ge** 760	**As** 950	**Se** 940	**Br** 1140	**Kr** 1350
5	**Rb** 400	**Sr** 550	**Y** 620	**Zr** 660	**Nb** 660	**Mo** 680	**Tc** 700	**Ru** 710	**Rh** 720	**Pd** 800	**Ag** 730	**Cd** 870	**In** 560	**Sn** 710	**Sb** 830	**Te** 870	**I** 1010	**Xe** 1170
6	**Cs** 380	**Ba** 500	**La** 540	**Hf** 680	**Ta** 760	**W** 770	**Re** 760	**Os** 840	**Ir** 880	**Pt** 870	**Au** 890	**Hg** 1010	**Tl** 590	**Pb** 720	**Bi** 700	**Po** 810	**At**	**Rn**

Key words

electron
element
group 1
ion
ionization energy
lanthanide series
mole
nucleus
period
shell

First ionization energy

- The first *ionization energy* of an element is the energy needed to remove a single *electron* from 1 *mole* of atoms of the *element* in the gaseous state, in order to form 1 mole of positively charged *ions*.
- Reading down *group 1*, there is a decrease in the first ionization energies. This can be explained by considering the electronic configuration of the elements in the group. Reading down, the outer electron is further from the positively charged *nucleus*, and there is an increasing number of complete *shells* of inner electrons, which to some extent, shield the outer electron from the nucleus. The result is that less energy is needed to remove the outer electron. A similar situation exists in other groups.

Increase in ionization energy

- There is a general increase in the first ionization energies across a *period*. This increase is due to electrons at the same main energy level being attracted by an increasing nuclear charge. This charge is caused by the increasing number of protons in the nucleus. The increase makes it progressively more difficult to remove an electron; thus more energy is needed.
- The diagram illustrates this principle using the first six periods minus the *lanthanide series*.

Elements whose ionization energies are the greatest in their period

He	Helium
Ne	Neon
Ar	Argon
Kr	Krypton
Xe	Xenon
Rn	Radon

© Diagram Visual Information Ltd.

Variation of first ionization energy

Key words

ionization energy
noble gases
period
periodicity
shell

First ionization energies

- The graph shows a repeating pattern, or *periodicity*, corresponding to reading down the *periods* of the periodic table.
- Within a period, it becomes increasingly more difficult to remove an electron due to the increasing nuclear charge. The graph peaks at the last element in each period, which is a *noble gas* (labeled on the graph).
- The noble gases have complete outer *shells* of electrons. This electron configuration provides great stability, and consequently, the noble gases are very unreactive. Some are totally unreactive. The first *ionization energies* of the noble gases are very high.

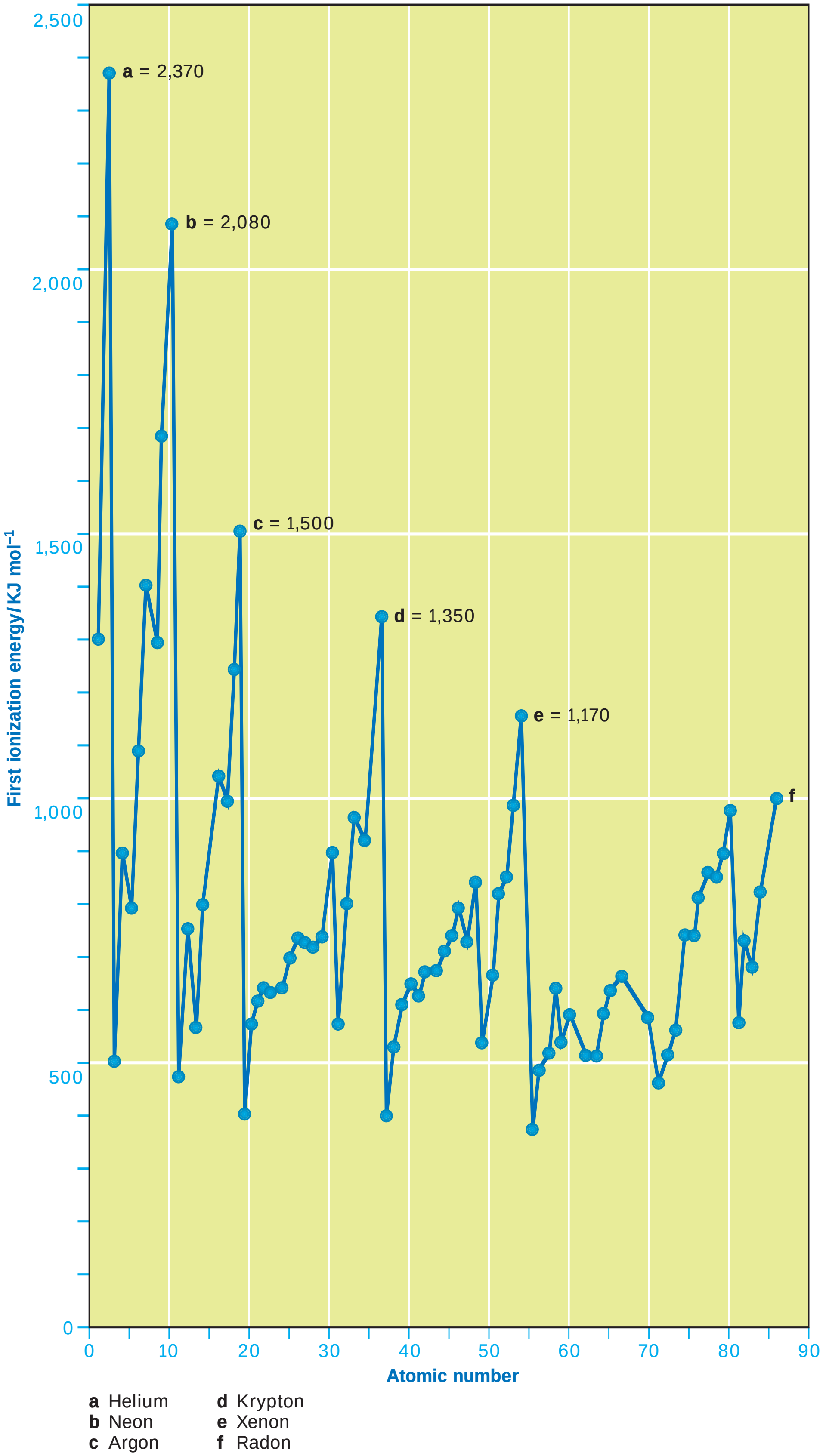

© Diagram Visual Information Ltd.

Melting points of the elements °C

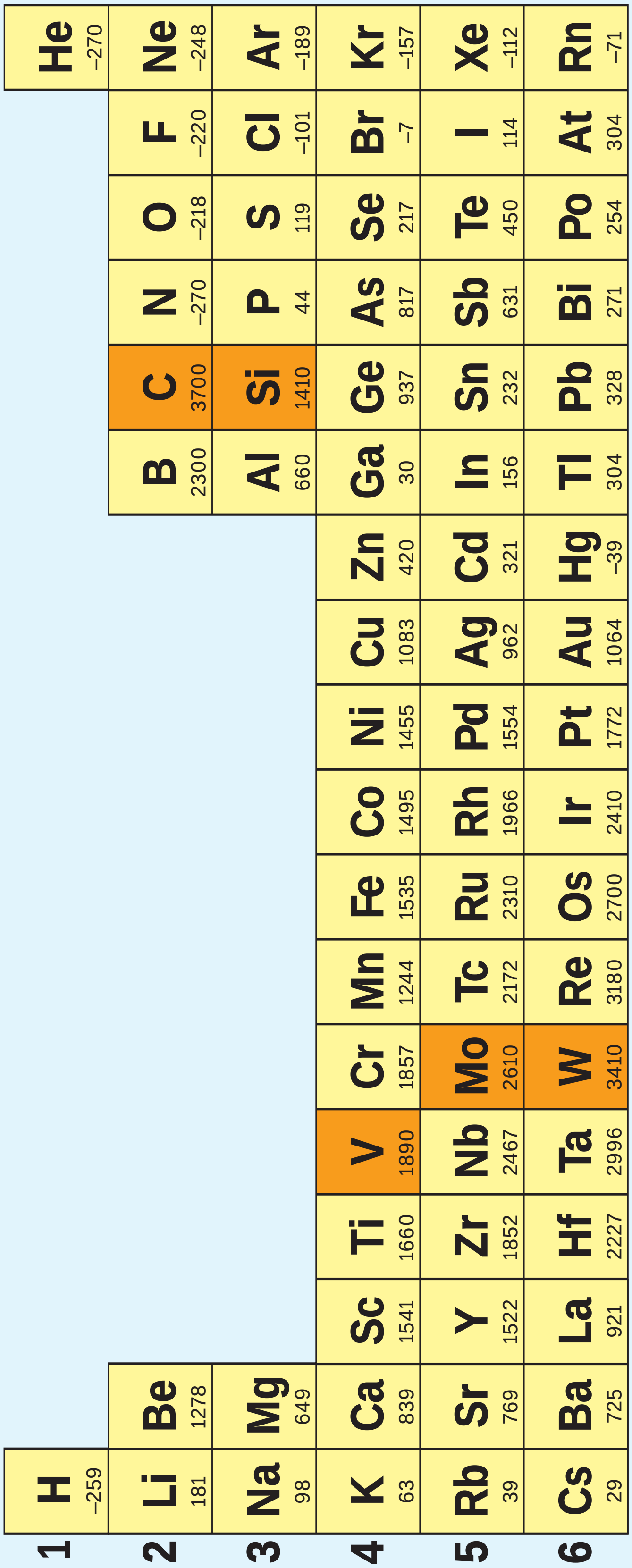

1	H −259																	He −270
2	Li 181	Be 1278											B 2300	**C 3700**	N −270	O −218	F −220	Ne −248
3	Na 98	Mg 649											Al 660	**Si 1410**	P 44	S 119	Cl −101	Ar −189
4	K 63	Ca 839	Sc 1541	Ti 1660	**V 1890**	Cr 1857	Mn 1244	Fe 1535	Co 1495	Ni 1455	Cu 1083	Zn 420	Ga 30	Ge 937	As 817	Se 217	Br −7	Kr −157
5	Rb 39	Sr 769	Y 1522	Zr 1852	Nb 2467	**Mo 2610**	Tc 2172	Ru 2310	Rh 1966	Pd 1554	Ag 962	Cd 321	In 156	Sn 232	Sb 631	Te 450	I 114	Xe −112
6	Cs 29	Ba 725	La 921	Hf 2227	Ta 2996	**W 3410**	Re 3180	Os 2700	Ir 2410	Pt 1772	Au 1064	Hg −39	Tl 304	Pb 328	Bi 271	Po 254	At 304	Rn −71

Key words

group
group 1
lanthanide series
liquid
melting point
noble gases
period
solid
transition metals

Melting points

- The *melting point* is the point at which the solid and *liquid* phase of a substance is in equilibrium at a given pressure.
- In a *solid*, the particles are held in a rigid structure by the strong forces of attraction that exist between them. They vibrate but cannot move position. When a solid is heated to its melting point, the particles gain sufficient energy to overcome these forces of attraction, and the particles are able to move position.
- Within *groups* of metallic elements, the melting point decreases down the group. The converse is true for non-metals, where the melting point increases down the group.
- Reading across *periods* 2 and 3, the elements follow a pattern of metallic structure, giant covalent structure, and simple covalent structure. The melting point increases until a maximum is reached with the element that exists as a giant covalent structure.
- The more reactive metals in *group 1* are soft and have low melting points. *Transition metals* (elements that have an incomplete inner electron structure) are generally harder and have higher melting points.
- The *noble gases* exist as single atoms with only weak forces of attraction between them. Consequently, their melting points are very low.
- Using the first six periods minus the *lanthanide series*, the diagram highlights the element with the highest melting point in a period.

Elements whose melting points are the greatest in their period

C	Carbon
Si	Silicon
V	Vanadium
Mo	Molybdenum
W	Tungsten

© Diagram Visual Information Ltd.

Key words

element
melting point
period
periodicity
periodic table

Melting points

- The graph shows a repeating pattern, or *periodicity*, corresponding to reading down the *periods* of the *periodic table*.
- The structure of periods 2 and 3 with regard to the nature of the *elements*, is:
  Elements having a metallic structure: *melting point* increasing
  Elements having a giant covalent structure: melting point maximum
  Elements having a simple covalent structure: melting point decreasing
- In general, the melting point increases at the start of these periods, corresponding to elements that have metallic structure. The melting point is at maximum for elements that have a giant covalent structure (labeled on the graph). After this, the melting point rapidly falls to low values, corresponding to those elements that have a simple covalent structure.

© Diagram Visual Information Ltd.

Variation of melting points

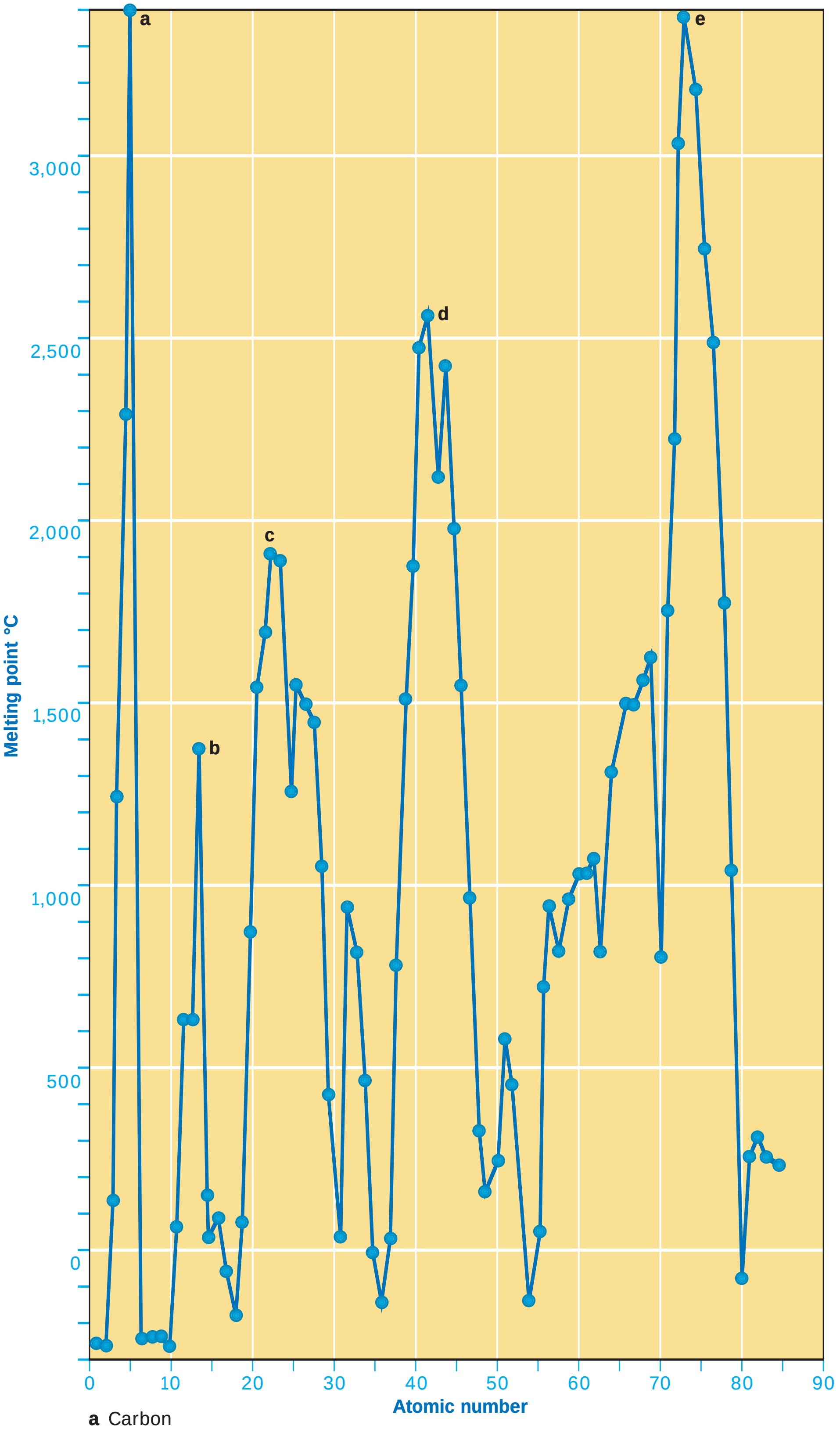

a Carbon
b Silicon
c Vanadium
d Molybdenum
e Tungsten

Boiling points of the elements °C

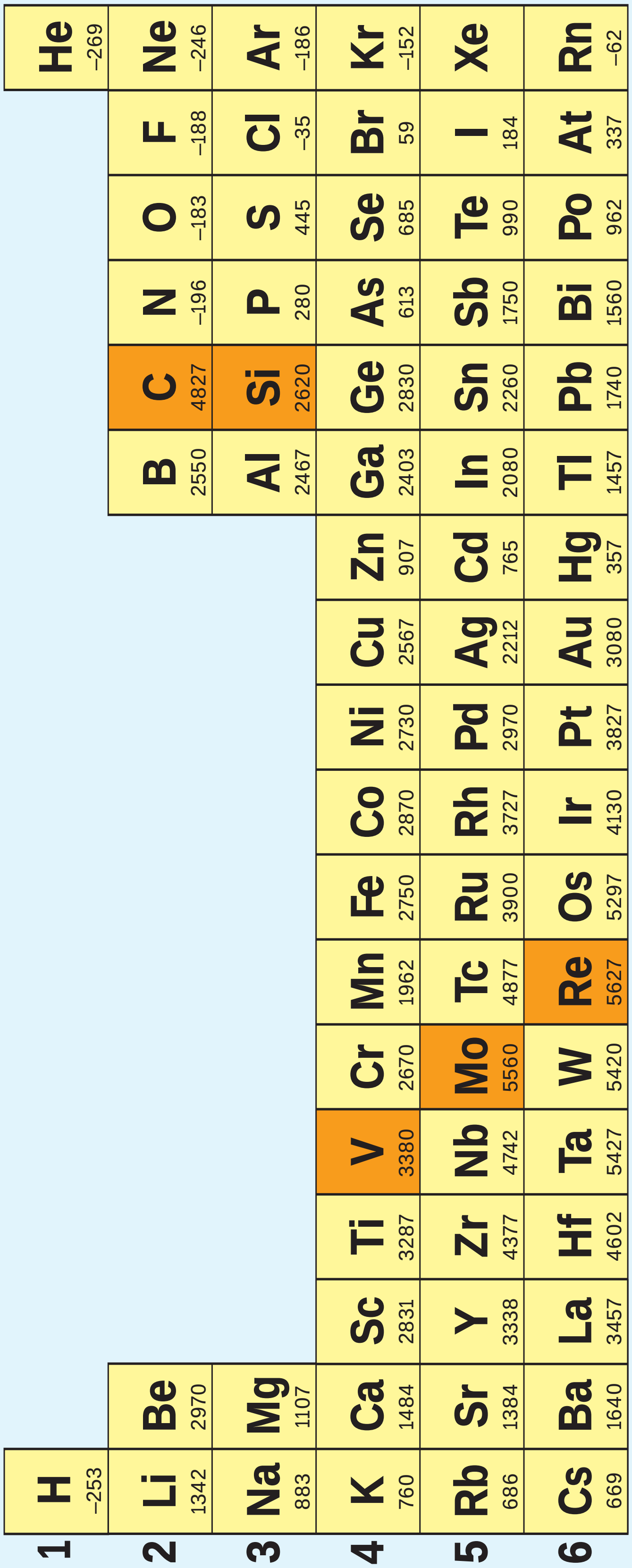

1	H −253																He −269
2	Li 1342	Be 2970											B 2550	C 4827	N −196	O −183	F −188	Ne −246
3	Na 883	Mg 1107											Al 2467	Si 2620	P 280	S 445	Cl −35	Ar −186
4	K 760	Ca 1484	Sc 2831	Ti 3287	V 3380	Cr 2670	Mn 1962	Fe 2750	Co 2870	Ni 2730	Cu 2567	Zn 907	Ga 2403	Ge 2830	As 613	Se 685	Br 59	Kr −152
5	Rb 686	Sr 1384	Y 3338	Zr 4377	Nb 4742	Mo 5560	Tc 4877	Ru 3900	Rh 3727	Pd 2970	Ag 2212	Cd 765	In 2080	Sn 2260	Sb 1750	Te 990	I 184	Xe
6	Cs 669	Ba 1640	La 3457	Hf 4602	Ta 5427	W 5420	Re 5627	Os 5297	Ir 4130	Pt 3827	Au 3080	Hg 357	Tl 1457	Pb 1740	Bi 1560	Po 962	At 337	Rn −62

Key words

boiling point
gas
group
group 1
kinetic energy
lanthanide series
liquid
noble gases
transition metals

Boiling points

- The *boiling point* is the temperature at which a *liquid* becomes a *gas*.
- The particles in a liquid are held together by the strong forces of attraction that exist between them. The particles vibrate and are able to move around, but they are held closely together. When a liquid is heated to its boiling point, the particles gain *kinetic energy*, moving faster and faster. Eventually, they gain sufficient energy to break away from each other and exist separately. There is a large increase in the volume of any substance going from a liquid to a gas.
- Within *groups* of metallic elements, the boiling point decreases down the group. The converse is true for non-metals: the melting point increases down the group.
- The more reactive metals in *group 1* have relatively low boiling points. *Transition metals* generally have very high boiling points.
- The *noble gases* exist as single atoms with only weak forces of attraction between them. Consequently, their boiling points are very low because it takes relatively little energy to overcome these forces.
- Using the first six periods minus the *lanthanide series*, the diagram highlights the element with the highest boiling point in a period.

Elements whose boiling points are the greatest in their period

C	Carbon
Si	Silicon
V	Vanadium
Mo	Molybdenum
Re	Rhenium

© Diagram Visual Information Ltd.

Key words

boiling point
gas
group 1
group 2
transition metals

Variation of boiling point

- The majority of non-metallic elements are *gases* at room temperature and atmospheric pressure. Most non-metallic elements have simple covalent structures and have very low *boiling points*.
- Elements with metallic and giant covalent structures have very high boiling points (see diagram). The boiling points of *transition metals* are generally much higher than those of the *group 1* and *group 2* metals.

Variation of boiling points

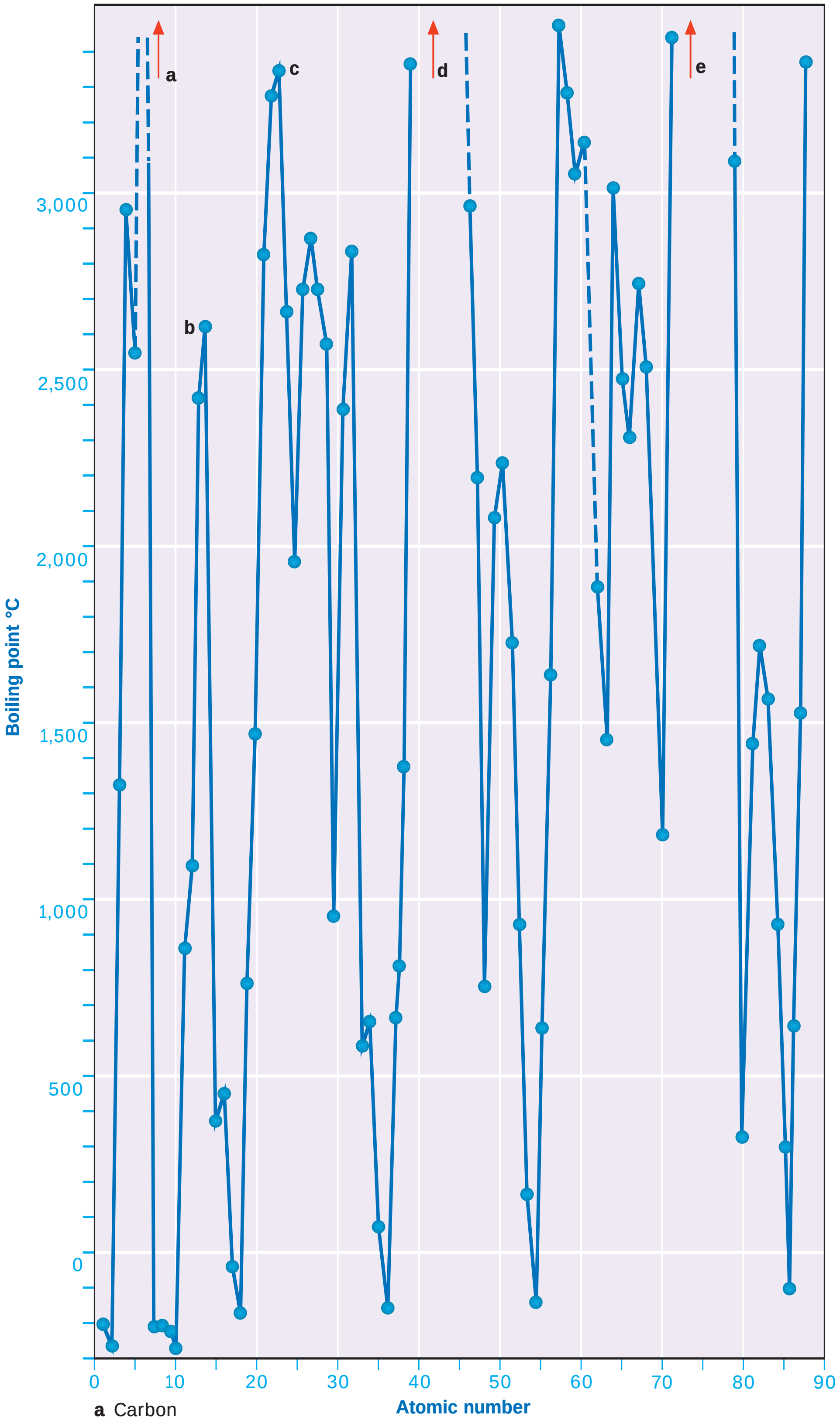

a Carbon
b Silicon
c Vanadium
d Molybdenum
e Rhenium

© Diagram Visual Information Ltd.

Atomic volumes of the elements

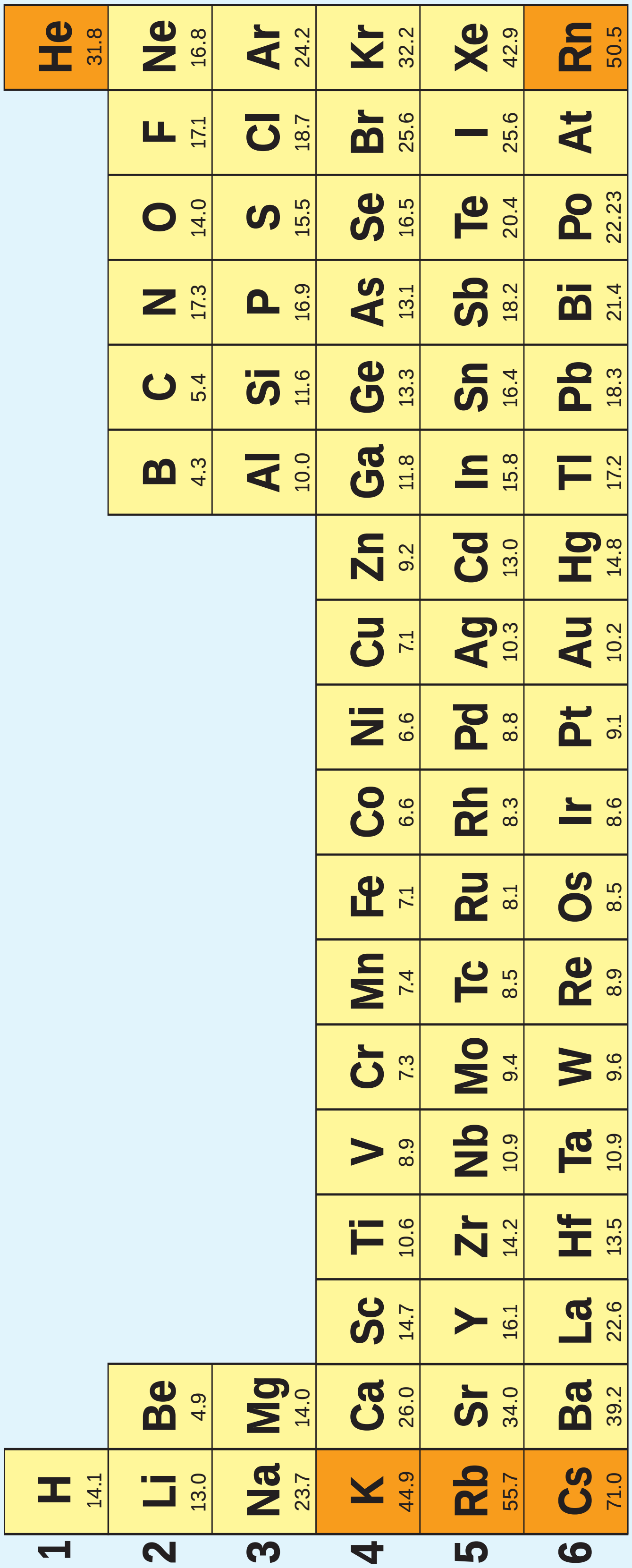

1	H 14.1																	He 31.8
2	Li 13.0	Be 4.9											B 4.3	C 5.4	N 17.3	O 14.0	F 17.1	Ne 16.8
3	Na 23.7	Mg 14.0											Al 10.0	Si 11.6	P 16.9	S 15.5	Cl 18.7	Ar 24.2
4	K 44.9	Ca 26.0	Sc 14.7	Ti 10.6	V 8.9	Cr 7.3	Mn 7.4	Fe 7.1	Co 6.6	Ni 6.6	Cu 7.1	Zn 9.2	Ga 11.8	Ge 13.3	As 13.1	Se 16.5	Br 25.6	Kr 32.2
5	Rb 55.7	Sr 34.0	Y 16.1	Zr 14.2	Nb 10.9	Mo 9.4	Tc 8.5	Ru 8.1	Rh 8.3	Pd 8.8	Ag 10.3	Cd 13.0	In 15.8	Sn 16.4	Sb 18.2	Te 20.4	I 25.6	Xe 42.9
6	Cs 71.0	Ba 39.2	La 22.6	Hf 13.5	Ta 10.9	W 9.6	Re 8.9	Os 8.5	Ir 8.6	Pt 9.1	Au 10.2	Hg 14.8	Tl 17.2	Pb 18.3	Bi 21.4	Po 22.23	At	Rn 50.5

Key words

atomic mass
atomic volume
density
element
group
group 8
lanthanide series
mole
noble gases
period

Atomic volume

- The *atomic volume* is the volume of one *mole* of the atoms of an *element*. It can be found by dividing the *atomic mass* of one mole of atoms by the density of the element:

  $$\text{Atomic volume} = \frac{\text{Atomic mass}}{\text{Density}}$$

- Since there are 6.023 x 10^{23} atoms per mole of atoms, it would seem possible to use the atomic volume to calculate the volume of a single atom, and thus its radius. However there are two problems with doing this. First, the state of an element, and therefore its density, changes with temperature and pressure. Second, using the atomic volume to calculate the volume of a single atom assumes that an element consists of atoms that are not bonded to each other. This is true only of the *group 8* elements (*noble gases*). For these reasons, it is not possible to consider the volume of an atom in isolation, but only as part of the structure of an element.
- In general, atomic volume increases down a *group*. Across a *period*, it decreases and then increases.
- The diagram highlights the element with the highest atomic volume in the first six periods (minus the *lanthanide series*).

Elements with peak atomic volumes

He	Helium
K	Potassium
Rb	Rubidium
Cs	Cesium
Rn	Radon

© Diagram Visual Information Ltd.

Key words

atomic mass
atomic number
atomic volume
periodicity

Periodicity

- As early as the Middle Ages, scientists recognized that elements could be differentiated by their properties and that these physical and chemical properties were periodic.
- The German chemist Lothar Meyer demonstrated *periodicity* by plotting *atomic volumes* against atomic weights (the term *atomic mass* is now used).
- This periodicity is better shown by plotting atomic volumes against *atomic number*.
- You can see periodicity most clearly by the pattern between potassium (b) and rubidium (c), and between rubidium (c) and cesium (d) in the diagram. These correspond to the changing values across period 4 and period 5, respectively.

© Diagram Visual Information Ltd.

Variation of atomic volumes

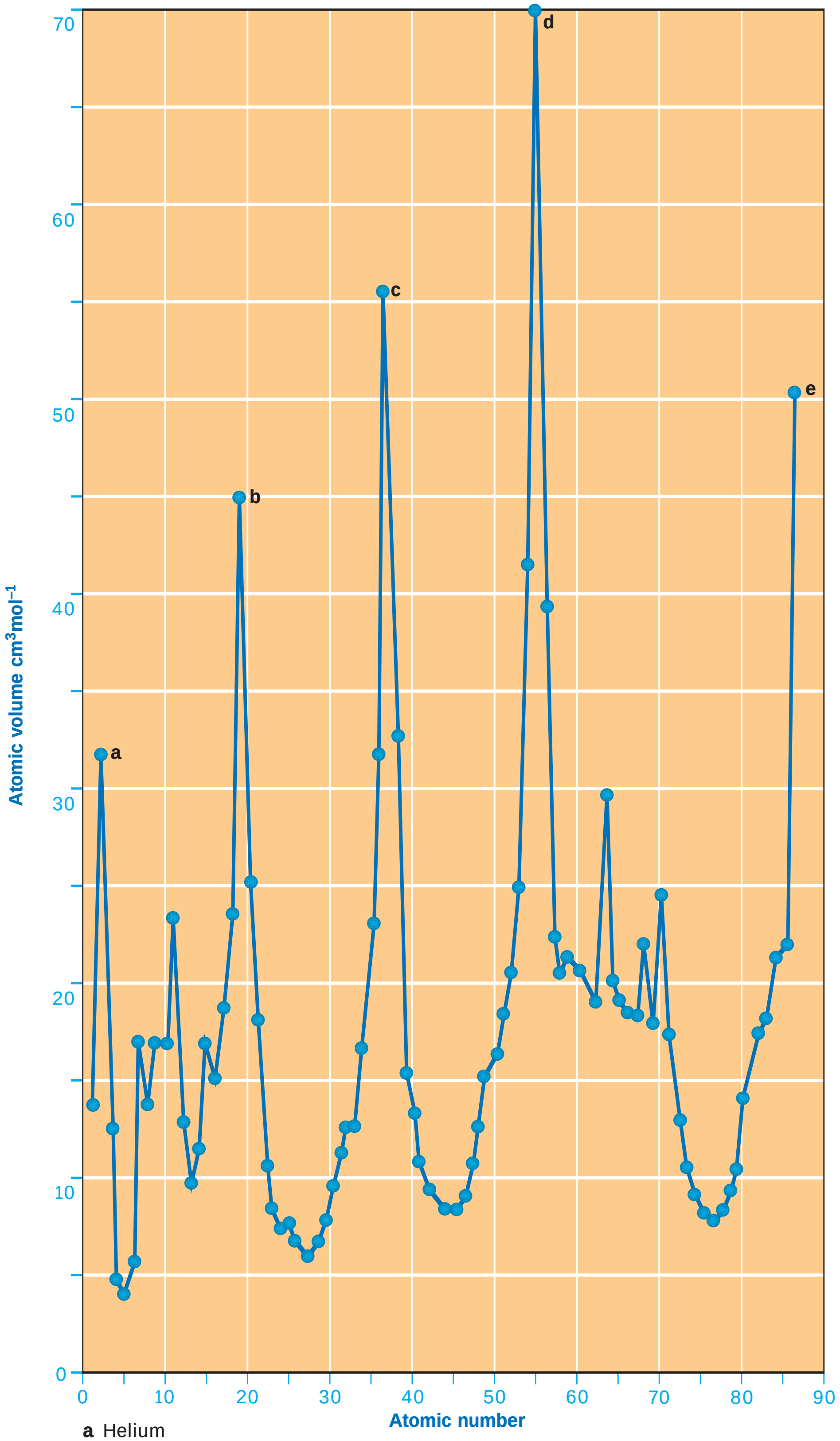

a Helium
b Potassium
c Rubidium
d Cesium
e Radon

Atomic mass

1 Carbon-12

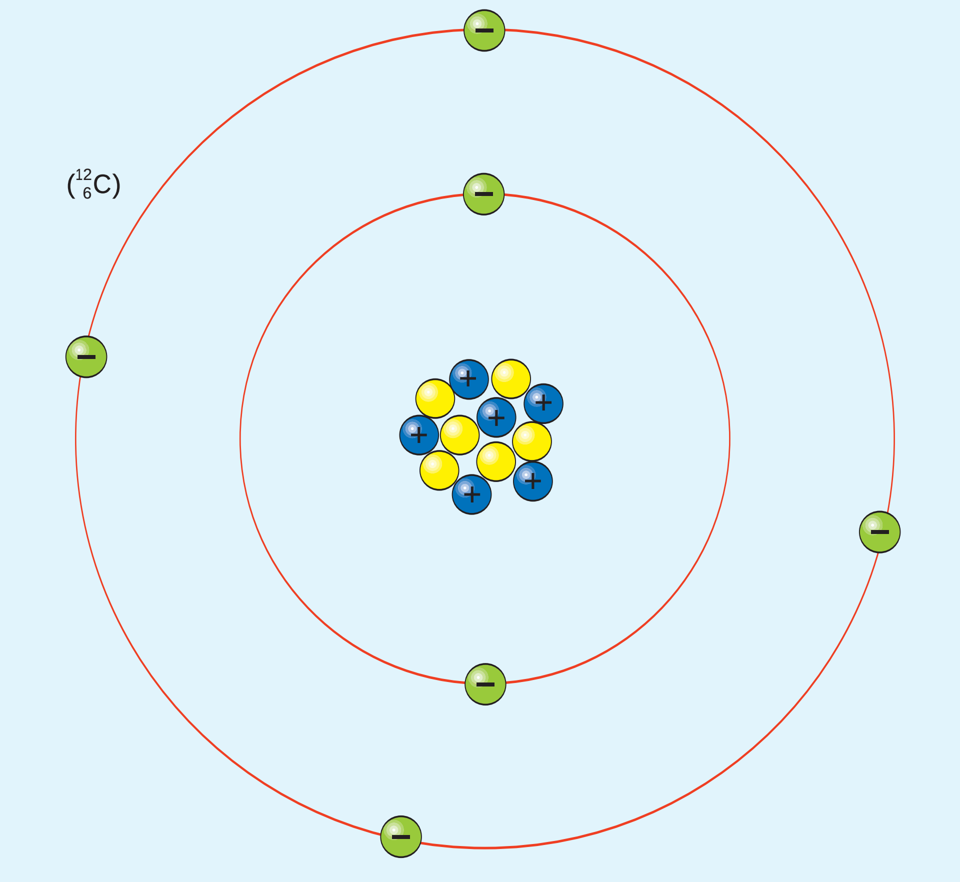

2 Lithium

Isotope	Natural abundance
Lithium-6 ($^{6}_{3}$Li)	7.5%
Lithium-7 ($^{7}_{3}$Li)	92.5%

The relative atomic mass of lithium is given by:

$$\frac{(6 \times 7.5) + (7 \times 92.5)}{100} = 6.925$$

3 Chlorine

Isotope	Natural abundance
Chlorine-35 ($^{35}_{17}$Cl)	75.77%
Chlorine-37 ($^{37}_{17}$Cl)	24.23%

The relative atomic mass of lithiumis given by:

$$\frac{(35 \times 75.77) + (37 \times 24.23)}{100} = 35.4846$$

Key words

atomic mass
isotope

1 Carbon-12

- To compare the masses of different atoms accurately, scientists need a standard mass against which all other masses can be calculated. Masses are given relative to this standard.
- The *isotope* carbon-12 is used as the standard. On this scale, atoms of carbon-12 are given a mass of exactly 12. The *atomic masses* of all other atoms are given relative to this standard.
- If an element contained only one isotope, its atomic mass would be the relative mass of that isotope. However, most elements contain a mixture of several isotopes in varying proportions.
- Natural abundance gives the proportion of each isotope in a sample of the element.
- If more than one isotope of an element is present, the atomic mass is calculated by taking an average that takes into account the relative proportion of each isotope. Diagrams 2 and 3 illustrate how the atomic mass of common isotopes of lithium and chlorine would be calculated.

2 Lithium

- There are two common isotopes of lithium: lithium-6 and lithium-7.
- The atomic mass of lithium is 6.925, but for most calculates a value of 7 is sufficiently accurate.

3 Chlorine

- There are also two common isotopes of chlorine: chlorine-35 and chlorine-37.
- The atomic mass of chlorine is 35.4846, but for most calculations a value of 35.5 is sufficiently accurate.
- Rounding the atomic mass of chlorine to the nearest whole number would lead to significant errors in calculations.

© Diagram Visual Information Ltd.

Periodic table with masses and numbers

Key words

atomic mass
atomic number
element
isotope

Atomic mass

- The *atomic mass* of an *element* is the average of the relative masses of its *isotopes*. It provides the relative mass of an "average" atom of the element, which is useful for calculations.
- The atomic mass is represented by the symbol A(r)
- The atomic mass of an isotope is its mass relative to the isotope carbon-12.
- The atomic mass of an isotope is the sum of the protons and neutrons in its nucleus.
- The atomic masses of the elements are presented below the element on the periodic table at right.

Atomic number

- The *atomic number* of an element is the number of protons in its nucleus.
- The atomic number is usually represented by Z.
- The number of neutrons in the nucleus of an isotope is:
  A(r) – Z
- The atomic numbers of the elements are presented above the element in the periodic table at right.

1 **H** 1.008																	2 **He** 4.00
3 **Li** 6.94	4 **Be** 9.01											5 **B** 10.81	6 **C** 12.01	7 **N** 14.01	8 **O** 16.00	9 **F** 19.00	10 **Ne** 20.18
11 **Na** 23.00	12 **Mg** 24.31											13 **Al** 27.00	14 **Si** 28.08	15 **P** 30.97	16 **S** 32.06	17 **Cl** 35.45	18 **Ar** 39.95
19 **K** 39.10	20 **Ca** 40.08	21 **Sc** 44.96	22 **Ti** 47.87	23 **V** 50.94	24 **Cr** 52.00	25 **Mn** 54.94	26 **Fe** 55.84	27 **Co** 58.93	28 **Ni** 58.69	29 **Cu** 63.55	30 **Zn** 65.41	31 **Ga** 69.72	32 **Ge** 72.64	33 **As** 74.92	34 **Se** 78.96	35 **Br** 79.90	36 **Kr** 83.80
37 **Rb** 85.47	38 **Sr** 87.62	39 **Y** 88.90	40 **Zr** 91.22	41 **Nb** 92.91	42 **Mo** 95.94	43 **Tc** 98.00	44 **Ru** 101.07	45 **Rh** 102.90	46 **Pd** 106.42	47 **Ag** 106.90	48 **Cd** 112.41	49 **In** 114.82	50 **Sn** 118.71	51 **Sb** 121.76	52 **Te** 127.60	53 **I** 126.90	54 **Xe** 131.29
55 **Cs** 132.90	56 **Ba** 137.33	57–71	72 **Hf** 178.49	73 **Ta** 180.95	74 **W** 183.84	75 **Re** 186.21	76 **Os** 190.23	77 **Ir** 192.22	78 **Pt** 195.08	79 **Au** 196.97	80 **Hg** 200.59	81 **Tl** 204.38	82 **Pb** 207.20	83 **Bi** 208.98	84 **Po** 210	85 **At** 210	86 **Rn** 220
87 **Fr** 223	88 **Ra** 226	89–103	104 **Rf** 261	105 **Db** 262	106 **Sg** 266	107 **Bh** 264	108 **Hs** 277	109 **Mt** 268	110 **Ds** 271	111 **Rg** 272	112 **Uub** 285	113 **Uut** 284	114 **Uuq** 289	115 **Uup** 288	116 **Uuh** 292		

57 **La** 138.90	58 **Ce** 140.12	59 **Pr** 140.91	60 **Nd** 144.24	61 **Pm** 145.00	62 **Sm** 150.36	63 **Eu** 151.96	64 **Gd** 157.25	65 **Tb** 158.92	66 **Dy** 162.50	67 **Ho** 164.93	68 **Er** 167.26	69 **Tm** 168.93	70 **Yb** 173.04	71 **Lu** 174.97
89 **Ac** 227	90 **Th** 232.04	91 **Pa** 231.03	92 **U** 238.03	93 **Np** 237	94 **Pu** 244	95 **Am** 243	96 **Cm** 247	97 **Bk** 247	98 **Cf** 251	99 **Es** 252	100 **Fm** 257	101 **Md** 258	102 **No** 259	103 **Lr** 262

© Diagram Visual Information Ltd.

Calculating the molecular mass of compounds

1 Diatomic molecule (chlorine)

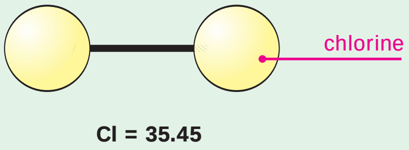

2 Covalent compound (ethanol)

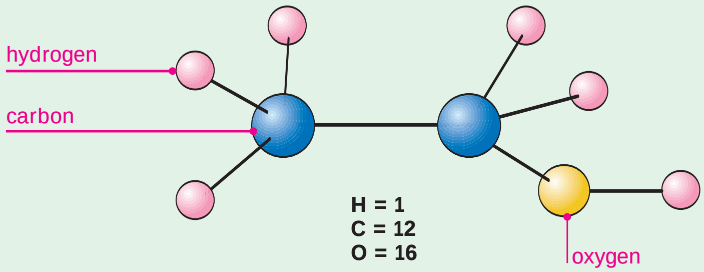

3 Ionic compound (sodium chloride)

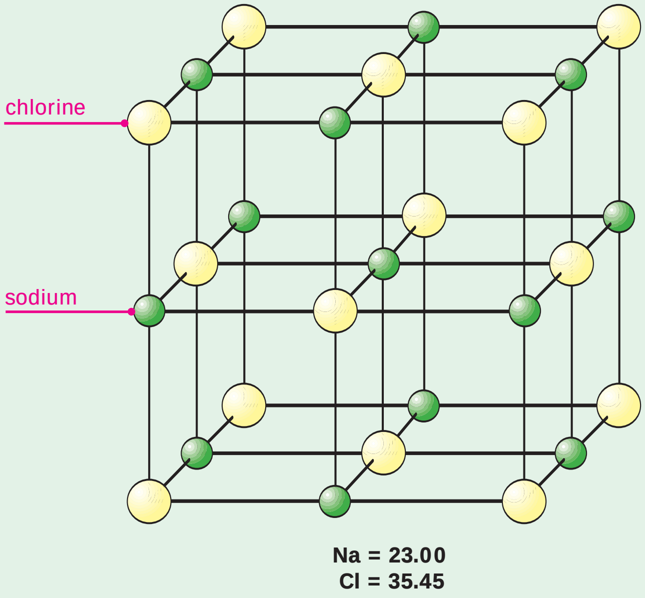

Key words

atomic mass
covalent compound
diatomic molecule
ionic compound
lattice
molecular mass

Calculating molecular mass

You calculate the *molecular mass* of a compound the same way regardless of structure:

1. Multiply the number of atoms in an element by its *atomic mass*.
2. Repeat this process for each element in the compound, then
3. Add the numbers.

1 Diatomic molecule (chlorine)

- The element chlorine exists as a *diatomic molecule* Cl_2.
  Atomic mass of chlorine = 35.5
  Molecular mass of chlorine
  = 2 x 35.5 = 71

2 Covalent compound (ethanol)

- Ethanol is a simple *covalent compound* that has the formula C_2H_5OH.
  Atomic mass of carbon = 12; hydrogen = 1; oxygen = 16.
  Molecular mass of ethanol
  = (2 x 12) + (6 x 1) + (1 x 16) = 46

3 Ionic compound (sodium chloride)

- *Ionic compounds* do not exist as molecules but as a giant *lattice* composed of ions in a fixed ratio. The formula mass of an ionic compound is the sum of the atomic masses of the ions in their simplest ratio.
- Sodium chloride consists of an ionic lattice in which the ions are present in the ratio 1:1. Therefore, the formula of sodium chloride is taken to be NaCl.
  Atomic mass of sodium = 23; chlorine = 35.5.
  Formula mass of sodium chloride
  = 23 + 35.5 = 58.5

© Diagram Visual Information Ltd.

Structure of some ionic crystals

Key words

anion
bond
cation
coordination number
ionic crystal
ion
lattice

Ionic crystals

- In an *ionic crystal*, each *ion* is surrounded by a number of oppositely charged ions in a *lattice* structure.
- There are several types of ionic structures.
  Simple: The atoms form grids.
  Body centered: One atom sits in the center of each cube.
  Face centered: One atom sits in each “face” of the cube.
- The lattice structure is determined by two factors:
  1. the ratio of the number of *cations* (positively charged ions) to *anions* (negatively charged ions)
  2. the ratio of the radii of the ions.
- In general, the higher the value of the radius ratio the higher the *coordination number* of the lattice. The coordination number is the number of atoms, ions, or molecules to which *bonds* can be formed.

1 Simple cubic structure (CsCl)

- In cesium chloride, the radius ratio is 0.94 (due to the large cesium ion). The coordination is 8:8. Each ion is surrounded by 8 oppositely charged ions.

2 Face-centered cubic structure (NaCl)

- The radius ratio in the sodium chloride lattice is 0.57. The coordination is 6:6. Each ion is surrounded by 6 oppositely charged ions.

3 Body-centered cubic structure (CaF_2)

- In calcium fluoride the radius ratio is 0.75. The coordination is 8:4. Each calcium ion is surrounded by 8 fluoride ions, while each fluoride ion is surrounded by 4 calcium ions.

© Diagram Visual Information Ltd.

1 Simple cubic structure (CsCl)

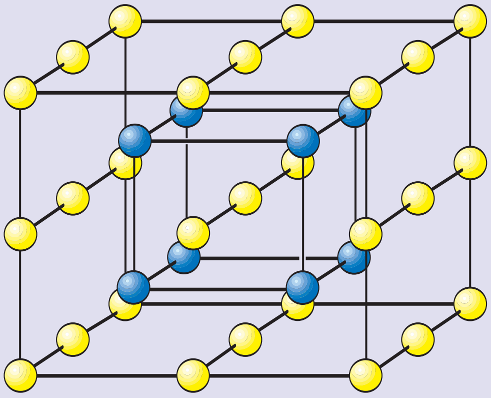

2 Face-centered cubic structure (NaCl)

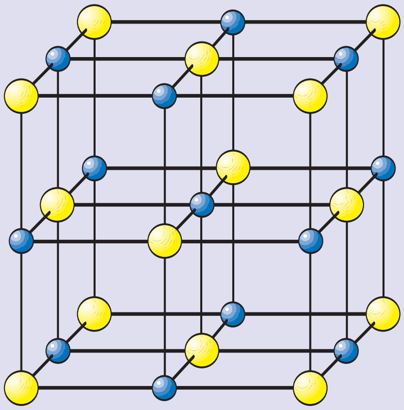

3 Body-centered cubic structure (CaF_2)

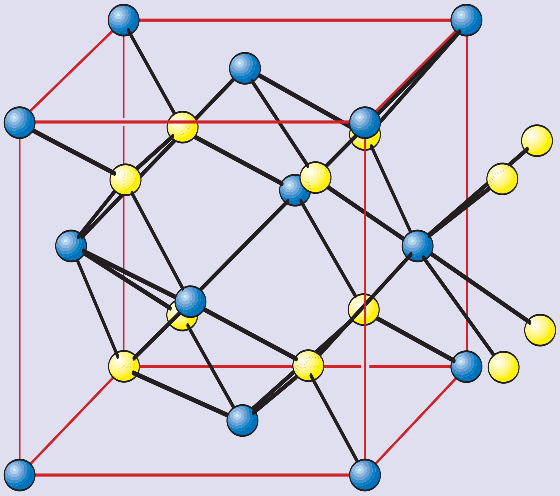

Crystal structure of metals: lattice structure

1 Hexagonal close packing

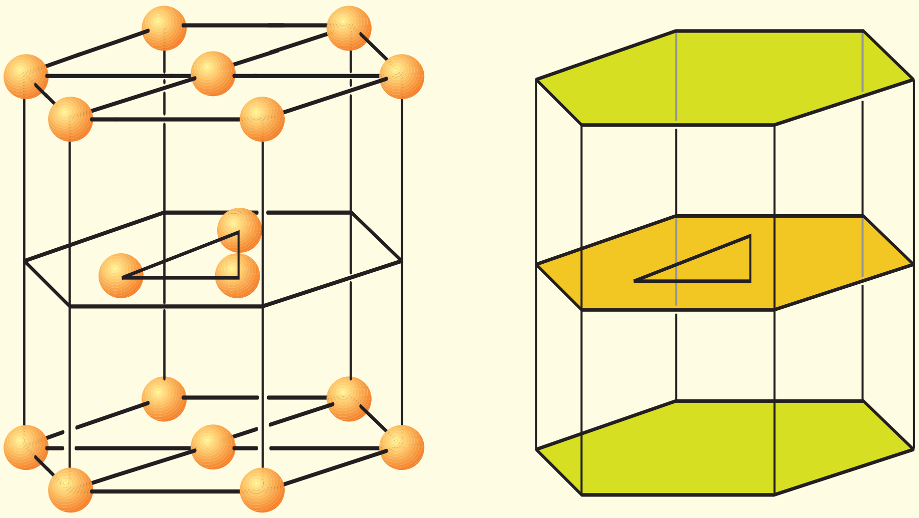

2 Face-centered cubic close packing

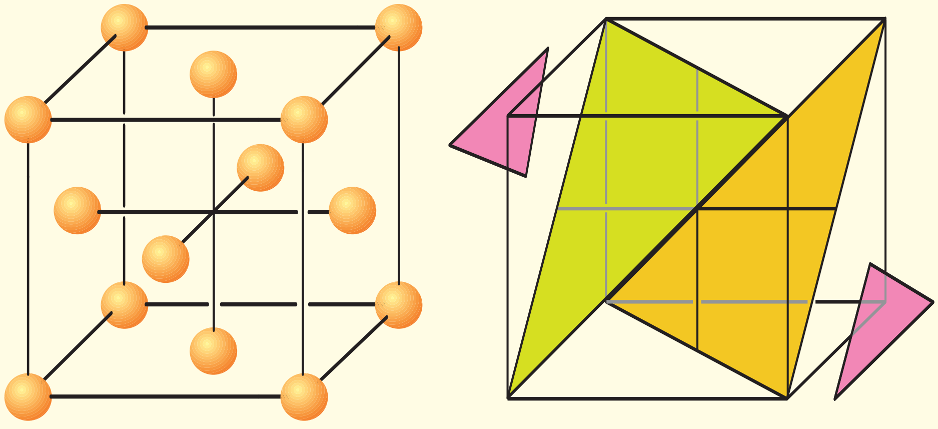

3 Body-centered cubic packing

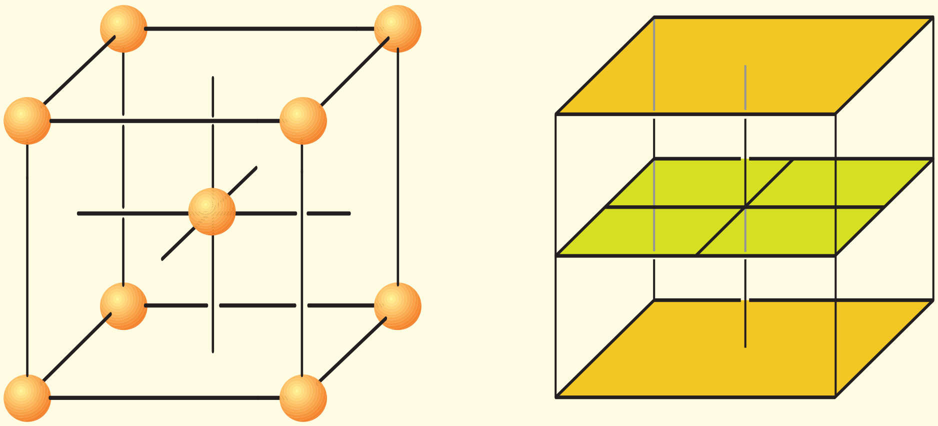

Key words

body centered cubic packing	*hexagonal close packing*
crystal	*lattice*
face-centered cubic close packing	*unit cell*

Metallic crystals

- Like all other *crystals*, metallic crystals are composed of *unit cells*, sets of atoms, ions, or molecules in orderly three dimensional arrangements called *lattices*.

1 Hexagonal close packing

- When arranged in a single layer, the most efficient method of packing the ions is in the form of a hexagon in which each ion is surrounded by six other ions.
- In *hexagonal close packing*, a second layer is positioned so that each ion in the second layer is in contact with three ions in the first layer. The third layer is placed directly above the first, and the fourth layer directly above the second, etc. This arrangement is sometimes represented as ABABAB.

2 Face-centered cubic close packing

- Here the third layer does not sit directly above either the first or second layers. The pattern is repeated after three layers, giving rise to an ABCABCABC arrangement.

3 Body-centered cubic packing

- Here the layers are formed from ions arranged in squares. The second layer is positioned so that each sphere in the second layer is in contact with four spheres in the first layer. The third layer sits directly above the first layer, giving rise to an ABABAB arrangement.

© Diagram Visual Information Ltd.

Crystal structure of metals: efficient packing

Key words

body-centered cubic packing
coordination number
face-centered cubic close packing
hexagonal close packing

1 Efficient packing

- Both *hexagonal close packing* and *face-centered cubic close packing* may be considered as efficient packing since the spheres occupy 74 percent of the available space. In both arrangements, each sphere is in contact with 12 others, and is said to have a *coordination number* of 12.

2 Less efficient packing

- *Body-centered cubic packing* is less efficient than hexagonal and face-centered cubic close packing. Spheres occupy only 68 percent of the available space. Each sphere is in contact with eight others (four in the layer above and four in the layer below) and, therefore, has a coordination number of eight.

Metals showing hexagonal close packing

- Cobalt
- Magnesium
- Titanium
- Zinc

Metals showing face-centered cubic close packing

- Aluminum
- Calcium
- Copper
- Lead
- Nickel

Metals showing body-centered cubic packing

- Group 1 metals
- Barium
- Chromium
- Iron
- Vanadium

© Diagram Visual Information Ltd.

1 Efficient packing

Hexagonal close packing

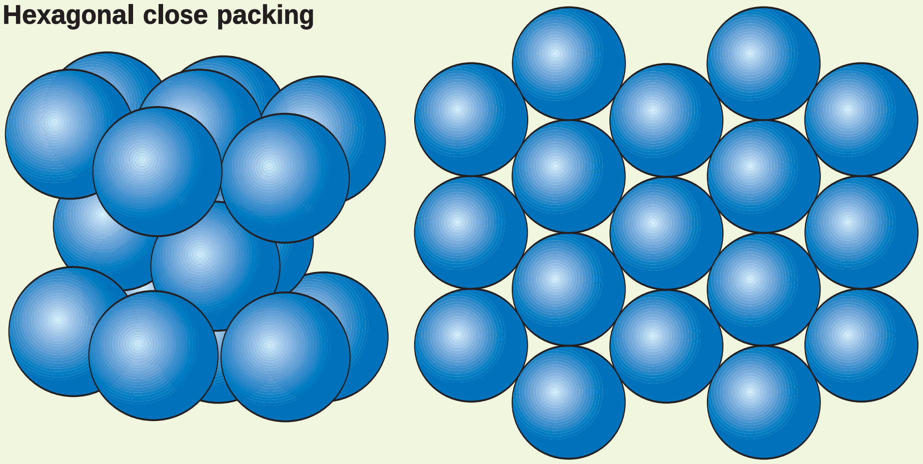

Face-centered cubic close packing

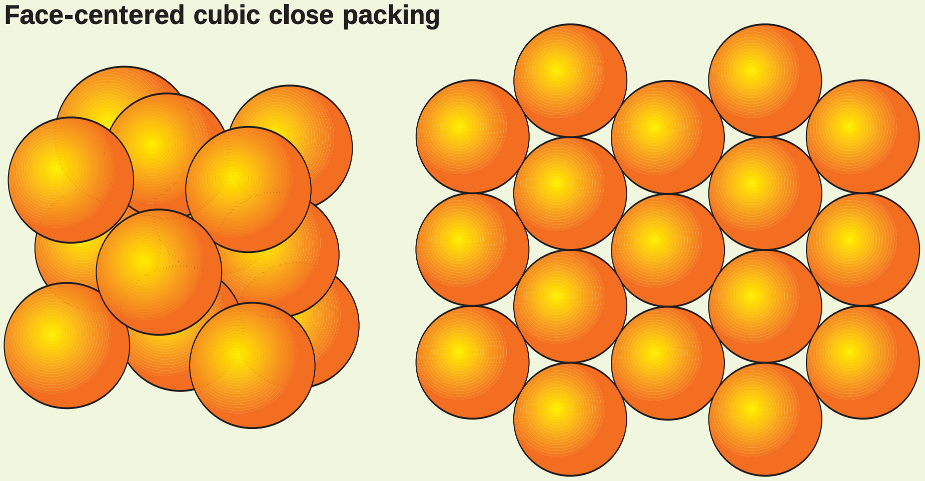

2 Less efficient packing

Body-centered cubic packing

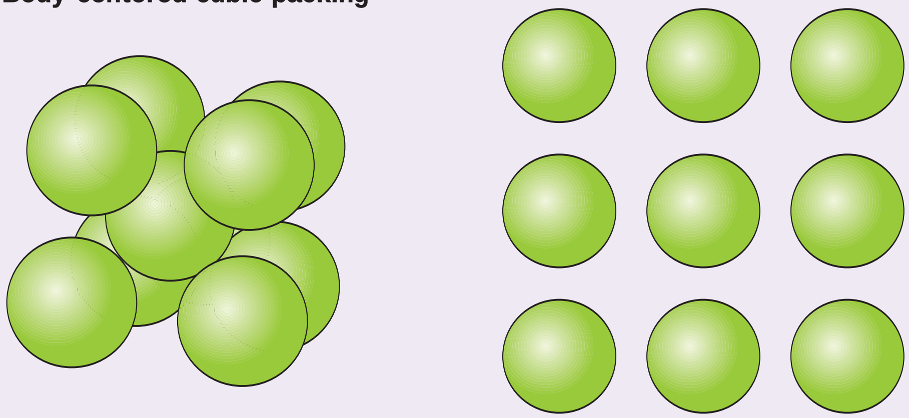

Chemical combination: ionic bonding

1 Formation of sodium chloride (NaCl)

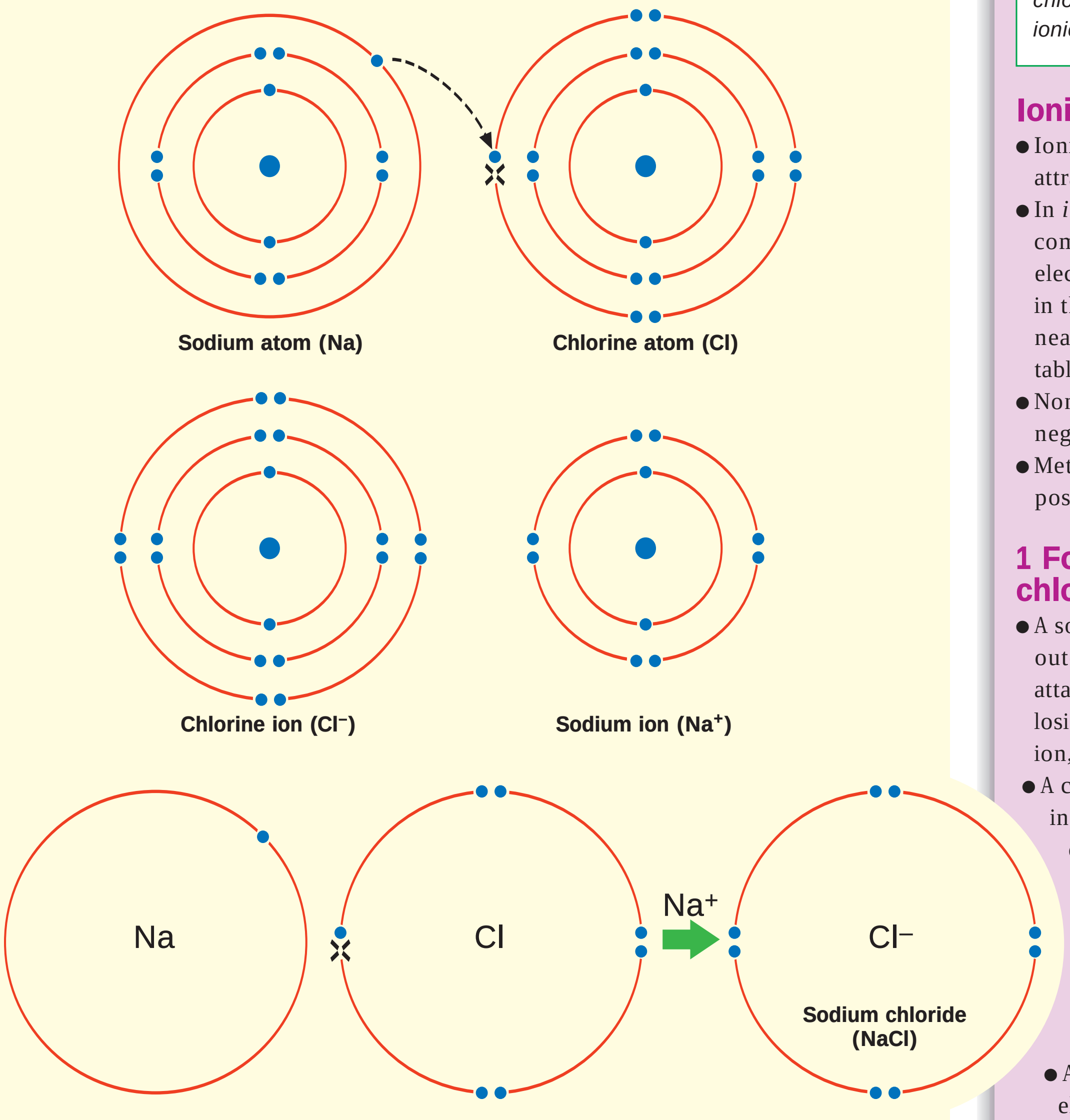

2 Formation of magnesium oxide (MgO)

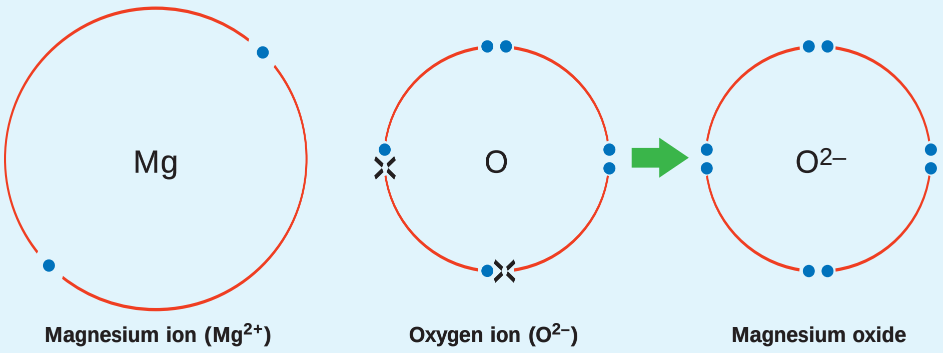

Key words

anion
bond
cation
chloride
ionic bonding
noble gases
oxide
shell

Ionic bonding

- Ionic *bonds* are formed by the attraction of opposite charges.
- In *ionic bonding*, the atoms in a compound gain, lose, or share electrons so the number of electrons in their outer *shell* is the same as the nearest *noble gas* on the periodic table.
- Non-metals gain electrons to give negatively charged ions (*anions*).
- Metal atoms loose electrons to give positively charged ions (*cations*).

1 Formation of sodium chloride (NaCl)

- A sodium atom has one electron in its outer shell. The easiest way it can attain a complete outer shell is by losing this electron to form a sodium ion, Na^+.
- A chlorine atom has seven electrons in its outer shell. The easiest way it can attain a complete outer shell is by gaining one more electron to form a *chloride* ion, Cl^-.

2 Formation of magnesium oxide (MgO)

- A magnesium atom has two electrons in its outer shell. It loses these electrons to form a magnesium ion, Mg^{2+}.
- An oxygen atom has six electrons in its outer shell. It gains two electrons to form an *oxide* ion, O^{2-}.

Electronic configuration

Na (sodium atom)	2.8.1
Na^+ (sodium ion)	2.8
Cl (chlorine atom)	2.8.7
Cl^- (chloride ion)	2.8.8
Mg (magnesium atom)	2.8.2
Mg^{2+} (magnesium ion)	2.8
O (oxygen atom)	2.6
O^{2-} (oxide ion)	2.8

© Diagram Visual Information Ltd.

Chemical combination: ionic radicals

Key words

carbonate
ion
limiting form
nitrate
radical
resonance structure
sulfate

Radicals

- A *radical* is a group of atoms that cannot be represented by one structural formula. It can pass unchanged through a series of chemical reactions. Radicals include the *carbonate* ion, $CO_3{}^{2-}$, the *nitrate* ion, $NO_3{}^{-}$, and the *sulfate* ion, $SO_4{}^{2-}$.

1 Carbonate ion

- The carbon atom is bonded to three oxygen atoms. By transferring electrons, it is possible to write three *limiting forms* for this *ion*. (Limiting forms are the possibilities for the distribution of electrons in a molecule or ion.)
- Electrons are continually being transferred in the ion. Thus its exact form is constantly changing. The ion is best represented as a *resonance structure* (the average of the limiting forms) in which dotted lines indicate that the charge on the ion, 2-, is spread over all three of the carbon–oxygen bonds.

2 Nitrate ion

- The nitrogen atom is bonded to three oxygen atoms. This ion has three limiting forms.
- It is best represented as a resonance structure in which dotted lines indicate that the charge on the ion, 1-, is spread over all three of the nitrogen–oxygen bonds.

3 Sulfate ion

- The sulfur atom is bonded to four oxygen atoms. This ion has three limiting forms.
- The ion's exact form is constantly changing. It is best represented as a resonance structure in which dotted lines indicate that the charge on the ion, 2-, is spread over all four of the sulfur–oxygen bonds.

© Diagram Visual Information Ltd.

1 Carbonate ion

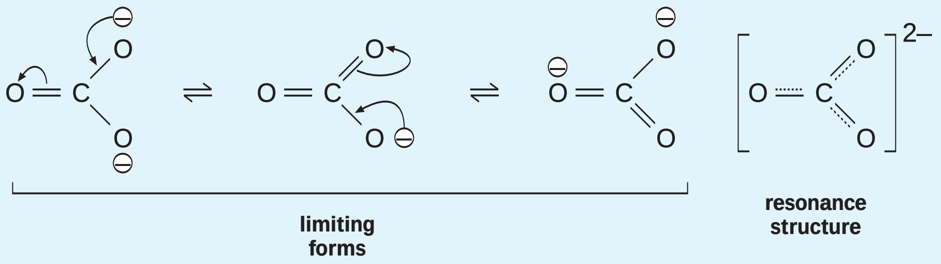

2 Nitrate ion

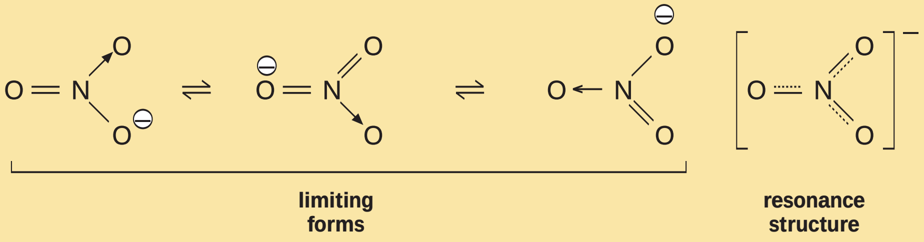

3 Sulfate ion

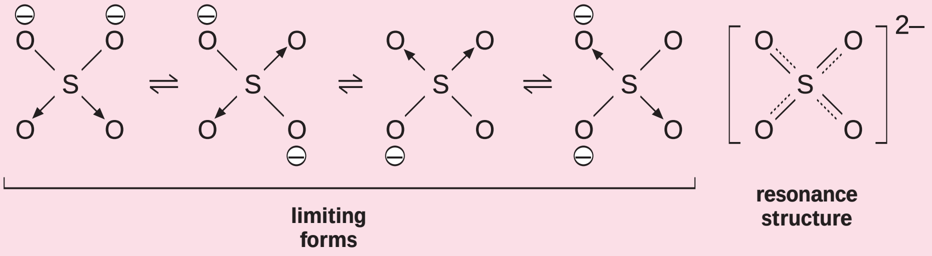

Chemical combination: covalent bonding

1 The hydrogen molecule

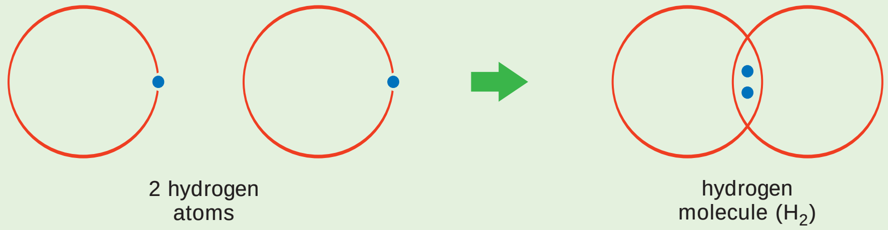

3 The ammonia molecule

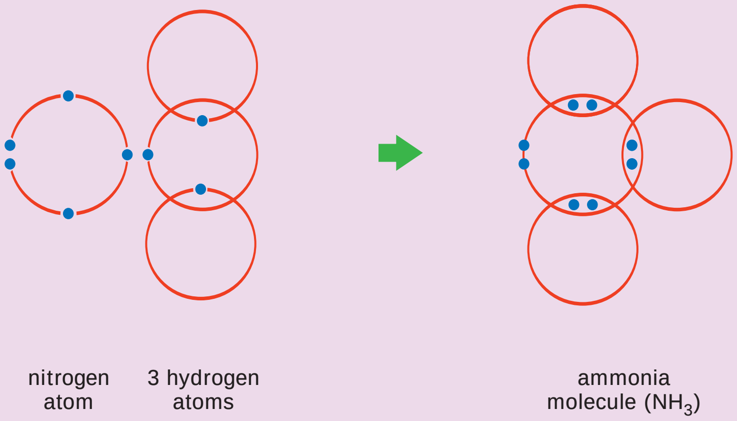

4 The methane molecule

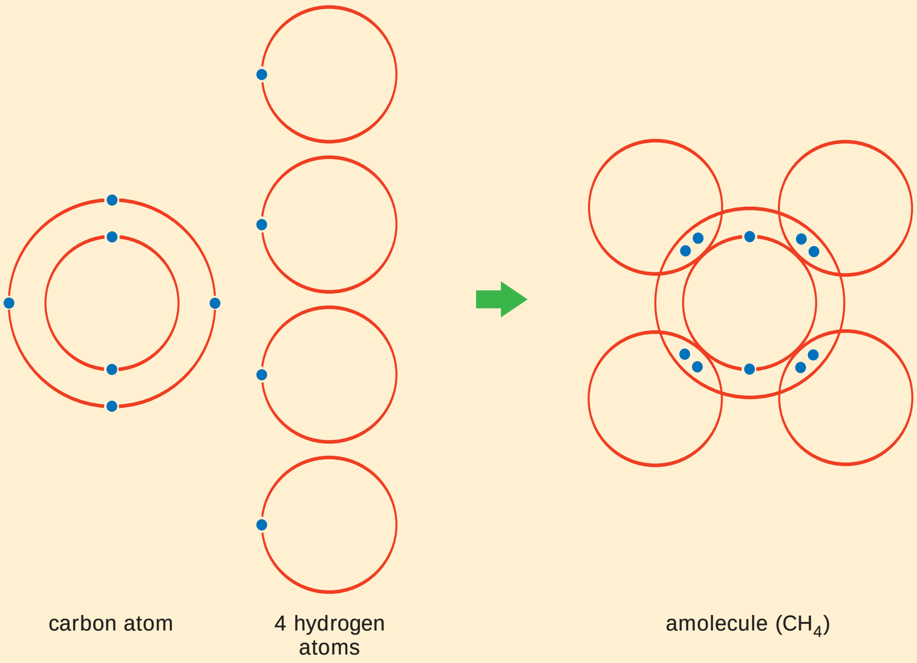

Key words

ammonia
bond
carbon
hydrogen
ionic compound
methane
nitrogen
shell

Covalent bonding

- Atoms gain stability by having a complete outer *shell* of electrons. In *ionic compounds*, this is achieved by the transfer of electrons. In covalent bonding, atoms share electrons.

1 The hydrogen molecule

- A *hydrogen* atom has one electron in its outer shell. In a hydrogen molecule, two hydrogen atoms each donate this electron to form a *bond*. Each hydrogen atom can be thought of as having control of the pair of electrons in the bond. Thus, each can be thought of as having a full outer shell of electrons. The single bond is shown as H-H.

2 The ammonia molecule

- A *nitrogen* atom has five electrons in its outer shell and needs another three electrons to complete the shell. In *ammonia*, three hydrogen atoms each donate one electron to form three N-H bonds. The nitrogen atom now has control of eight electrons and has a complete outer shell, while each hydrogen atom has control of two electrons and also has a complete outer shell.

3 The methane molecule

- A *carbon* atom has four electrons in its outer shell and needs another four electrons to complete the shell. In *methane*, four hydrogen atoms each donate one electron to form four C-H bonds. The carbon atom now has control of eight electrons and has a complete outer shell, while each hydrogen atom has control of two electrons and also has a complete outer shell.

© Diagram Visual Information Ltd.

Chemical combination: coordinate bonding

Key words

ammonium ion
coordinate bonding
covalent bond
covalent compound
hydronium ion
lone pair

Coordinate bonding

- *Coordinate bonding* is a particular form of covalent bonding in which one atom provides both electrons that the two atoms share.

1 Ammonium ion

- There is a non-bonding or *lone pair* of electrons on the nitrogen atom of an ammonia molecule. Nitrogen uses this lone pair to form a coordinate bond with a hydrogen ion, forming the *ammonium ion*, NH_4^+.

2 Hydronium ion

- The *hydronium ion*, H_3O^+, forms in a similar way.

3 Aluminum chloride

- The Al^{3+} ion is very small and carries a high charge. It attracts electrons so strongly that aluminum chloride is a *covalent compound*. It exists as Al_2Cl_6 molecules in which two $AlCl_3$ molecules are linked by coordinate bonds formed by the donation of lone pairs of electrons from two chlorine atoms.

4 Ionic compounds with covalent character

- The ions in a sodium chloride lattice are perfectly spherical. Thus the bonds in this compound are said to be perfectly ionic.
- But in an ionic compound consisting of a small, highly charged positive ion and a large negative ion such as lithium iodide, the positive ion attracts electron charge away from the negative ion. The result is that the negative ion becomes distorted, and electron density becomes concentrated between the ions, creating a bond similar to a *covalent bond*. Compounds like lithium iodide are said to be ionic with covalent character.

© Diagram Visual Information Ltd.

1 Ammonium ion

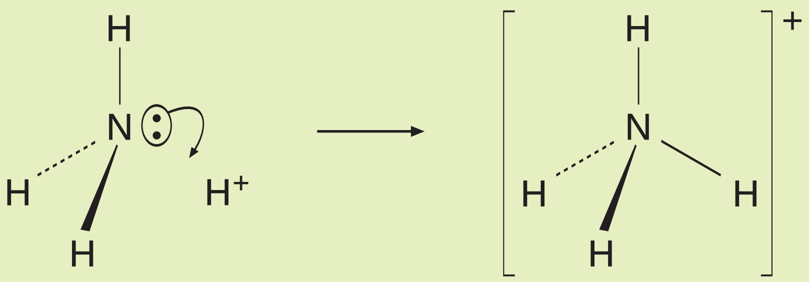

2 Hydronium ion

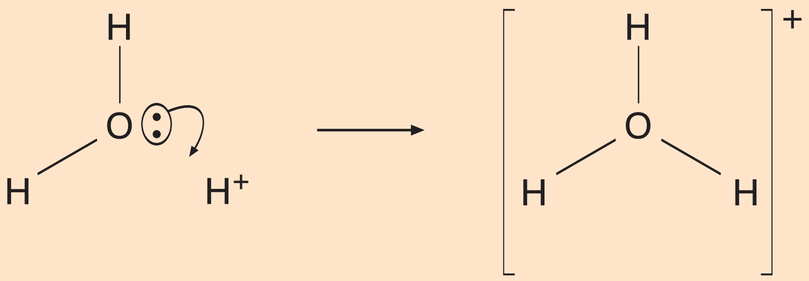

3 Aluminum chloride

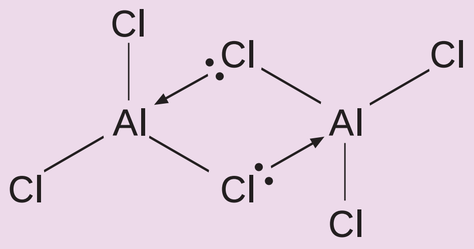

4 Ionic compounds with covalent character

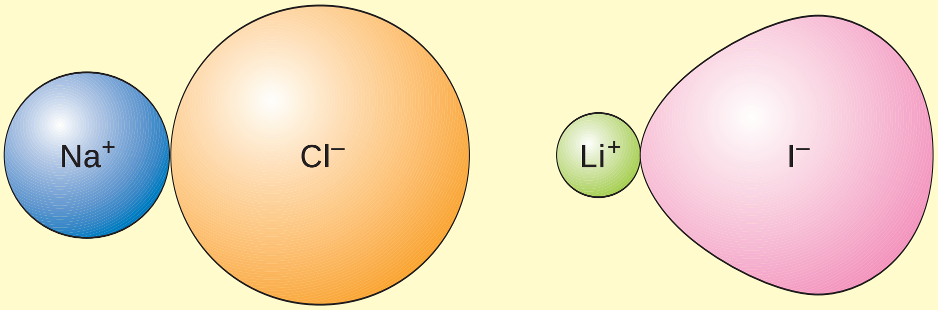

Mixtures and solutions

Producing a solution

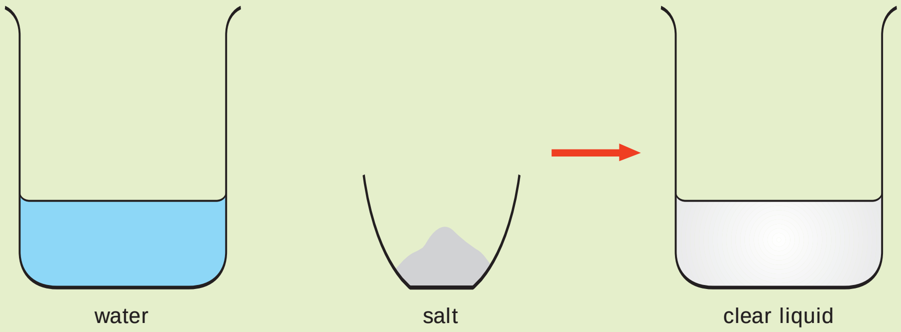
water
salt
clear liquid

Producing a suspension

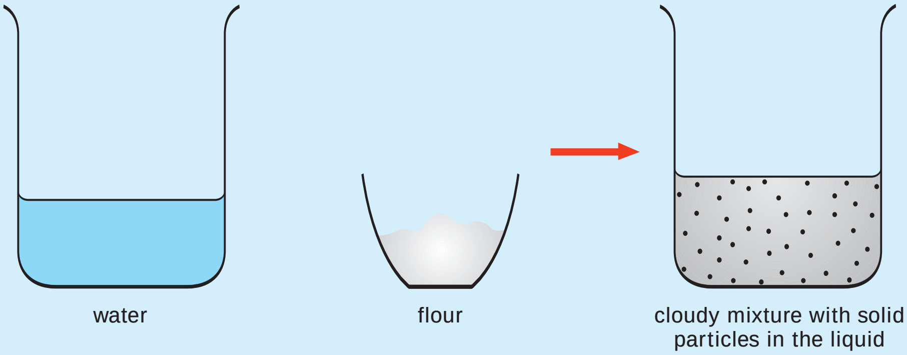
water
flour
cloudy mixture with solid particles in the liquid

Producing an emulsion

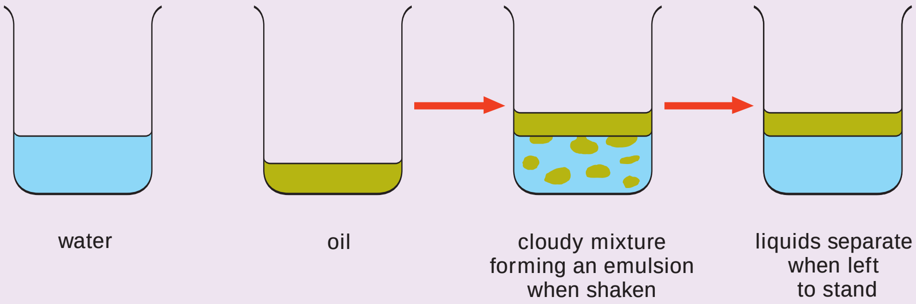
water
oil
cloudy mixture forming an emulsion when shaken
liquids separate when left to stand

Key words

emulsion
ionic compound
mixture
soluble
solution
suspension

1 Solutions

- A *solution* is a homogeneous *mixture* of substances. Particles in solutions are very small and cannot be seen. The particles may be atoms, ions, or molecules, and their diameters are typically less than 5 nm. Salt dissolves in water to form a clear colorless solution.
- Many, but not, all *ionic compounds* are *soluble* in water.
- A small proportion of organic compounds are soluble in water. However, organic compounds are generally more soluble in organic solvents such as hexane and ethanol.

2 Suspensions

- A *suspension* is a heterogeneous mixture of two components. The particles will settle out over a period of time. Suspended particles have diameters that are typically 1,000 nm or more.
- When flour is mixed with water, it forms a white suspension. Tiny particles of flour are suspended in the water. The flour particles can be filtered off from the suspension.

3 Emulsions

- An *emulsion* is a colloidal dispersion of small droplets of one liquid in another (See page 46).
- When oil and water are mixed, they form an emulsion. Oil is less dense than water and forms the upper layer.
- Tiny oil droplets are suspended in the water. After a while, the oil droplets join together, and two layers are formed.
- The mixture of oil and water can be separated using a separating funnel.

© Diagram Visual Information Ltd.

Key words

aerosol	*gel*
colloid	*sol*
emulsion	*solution*
filtration	*suspension*
foam	

Colloids

- A *colloid* is a substance made of particles whose size is intermediate between those in *solutions* and *suspensions*.
- The particles in a suspension have a diameter of typically 1,000 nm or more. The particles in a suspension will settle over a period of time.
- The particles in a colloid are approximately 500 nm or less in diameter and do not settle on standing.
- The particles in a colloid cannot be separated from the dispersion medium by ordinary techniques like *filtration* and centrifugation.
- A colloid consists of a dispersing medium and dispersed substance. These terms are analogous to the terms solute and solvent.
- Colloids are classified according to the original phases of their constituents. The main types are: *aerosols*, *foams*, *emulsions*, *sols*, and *gels*.
- Aerosols are extremely small solid or liquid particles suspended in air or another gas.
- Foams form when a gas is suspended in a liquid or a solid.
- Emulsions form when small particles of a liquid are suspended in another liquid.
- Sols form when solid particles are suspended in a liquid.
- Gels are solid particles arranged as a fine network in a liquid to form a jelly.

Colloids

1 In Air

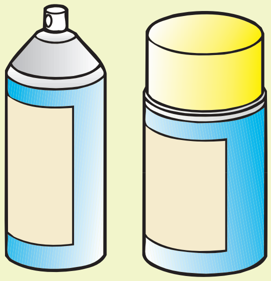

1 Colloid type	2 Phase of dispersing medium	3 Phase of dispersed substance	Examples
aerosol	gas	liquid	clouds, fog, insecticide spray
aerosol	gas	solid	dust, smoke

2 In liquids

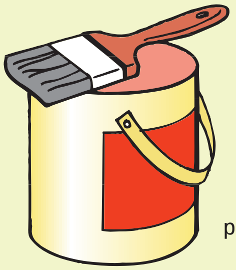

1 Colloid type	2 Phase of dispersing medium	3 Phase of dispersed substance	Examples
foam	liquid	gas	froth, whipped cream
emulsion	liquid	liquid	milk, salad dressing
sol	liquid	solid	milk of magnesia, paint

3 In solids

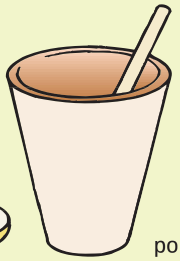

1 Colloid type	2 Phase of dispersing medium	3 Phase of dispersed substance	Examples
solid foam	solid	gas	cork, polyurethane
gel	solid	liquid	agar, geletine, jelly
solid sol	solid	solid	alloys

© Diagram Visual Information Ltd.

Simple and fractional distillation

Simple distillation of sea water

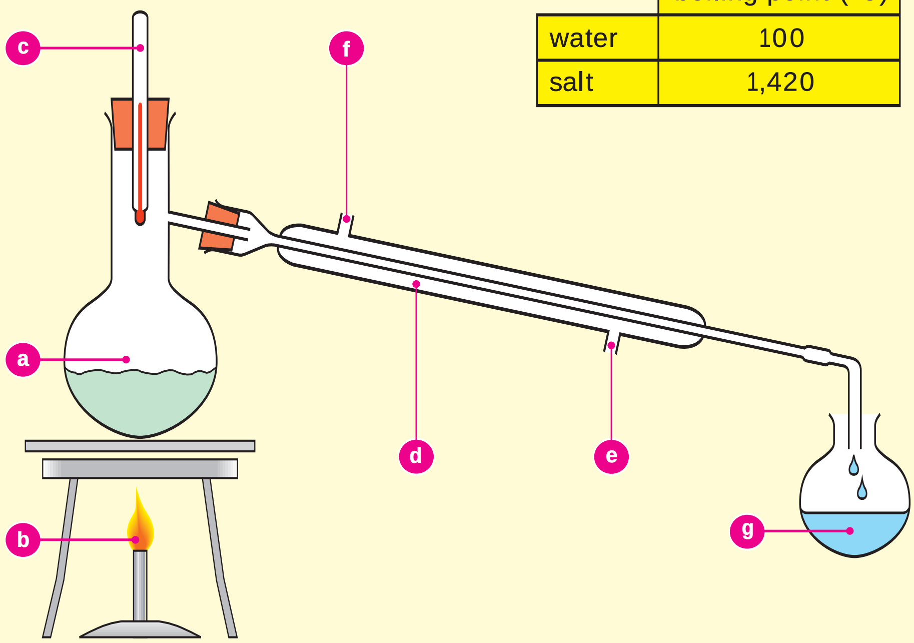

	boiling point (°C)
water	100
salt	1,420

Fractional distillation of ethanol

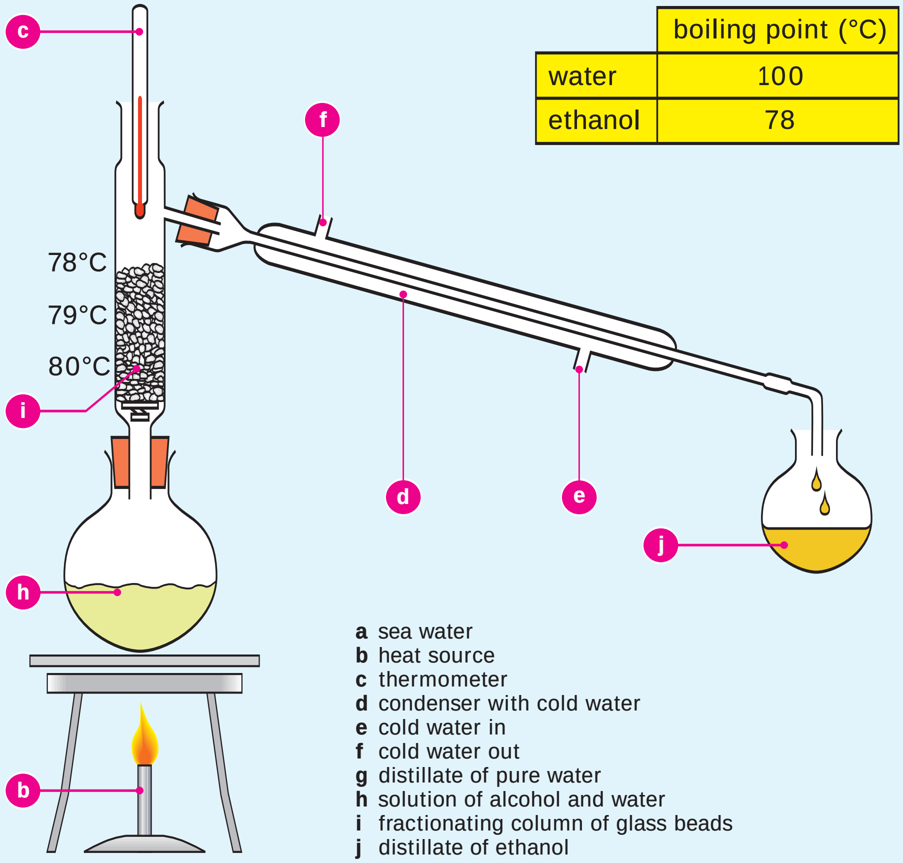

	boiling point (°C)
water	100
ethanol	78

a	sea water
b	heat source
c	thermometer
d	condenser with cold water
e	cold water in
f	cold water out
g	distillate of pure water
h	solution of alcohol and water
i	fractionating column of glass beads
j	distillate of ethanol

Key words

boiling point
distillation
fractional
distillation
mixture

1 Simple distillation

- *Distillation* is a process in which a *mixture* of materials is heated to separate the components.
- Simple distillation is used when the *boiling points* of the components are widely separated.
- In the diagram, salt water is placed in a round-bottom flask. Water boils at 100°C and becomes water vapor.
- A condenser consists of an inner tube surrounded by a jacket of cold water. This jacket ensures that the inner tube remains cool.
- The vapor passes into the condenser, where it is cooled and changes back into liquid.
- The water runs out of the condenser and is collected in a second flask.
- The salt remains in the round-bottomed flask.

2 Fractional distillation

- *Fractional distillation* is used to separate components whose boiling points are similar.
- Ethanol boils at 78°C and turns to vapor. Because the boiling point of water is only 100°C, a significant amount of water also becomes vapor as a result of evaporation.
- The fractionating column contains glass beads, which provide a large surface area for vapor to condense and the resulting liquid to subsequently boil.
- As the vapor mixture moves up the fractionating column, it condenses and then boils again to become vapor. Each time, the proportion of ethanol in the mixture increases.

© Diagram Visual Information Ltd.

Separating solutions

Key words

- *filtrate*
- *immiscible*
- *insoluble*
- *mixture*
- *residue*
- *soluble*
- *solvent*

1 Separating a mixture of two solids

- Sugar is *soluble* in ethanol, while salt is *insoluble*. When a *mixture* of sugar and salt is mixed with ethanol, the sugar dissolves while the salt does not.
- When the mixture is filtered, the undissolved salt remains as the *residue* in the filter. The *filtrate*, sugar solution, passes through the filter.
- If the filtrate is left open to the air, the ethanol evaporates, and solid sugar remains.

2 Separating two solutes in solution

- Salt dissolves in water but not in carbon tetrachloride.
- Iodine is slightly soluble in water but is far more soluble in carbon tetrachloride.
- When a mixture of salt and iodine is shaken in a mixture of water and carbon tetrachloride, the salt dissolves in the water and the iodine in carbon tetrachloride.
- Water and carbon tetrachloride are *immiscible*, they do not mix, and form two layers in a separating funnel. Carbon tetrachloride is more dense than water and forms the lower layer.
- When the layers are run into separate evaporating basins and left, the *solvents*—carbon tetrachloride and water—evaporate, leaving salt and iodine respectively.

1 Separating a mixture of two solids

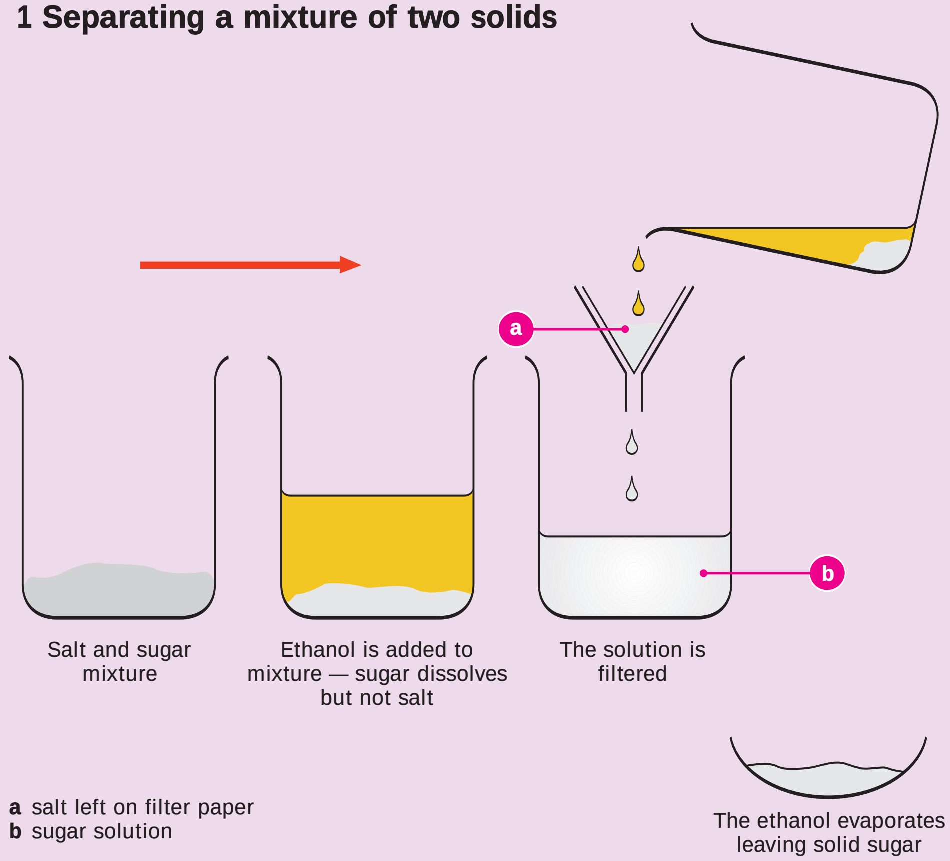
a
b
Salt and sugar mixture
Ethanol is added to mixture — sugar dissolves but not salt
The solution is filtered
The ethanol evaporates leaving solid sugar

a salt left on filter paper
b sugar solution

2 Separating two solutes in solution

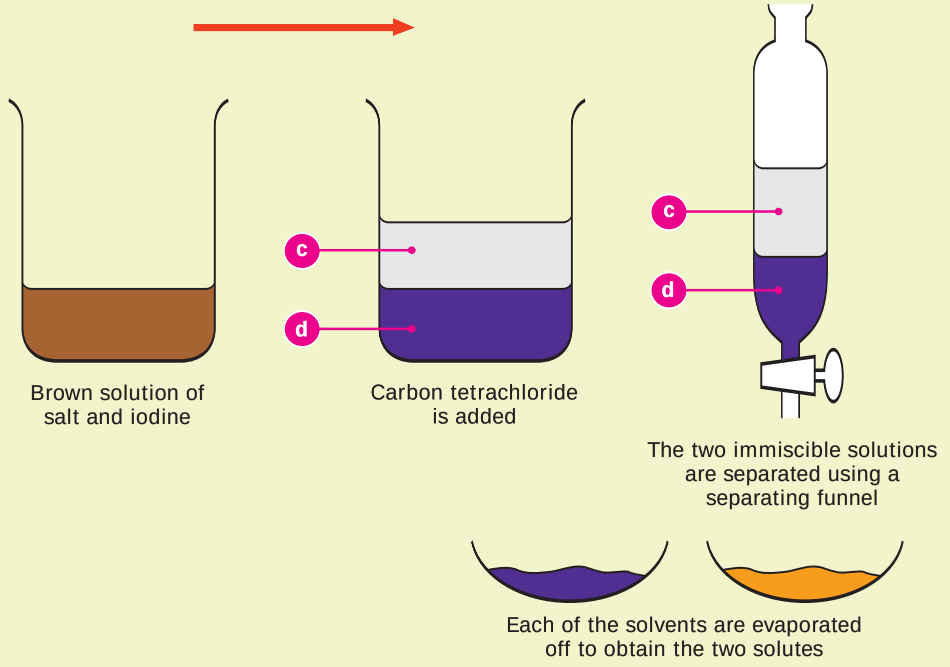
c
d
c
d
Brown solution of salt and iodine
Carbon tetrachloride is added
The two immiscible solutions are separated using a separating funnel
Each of the solvents are evaporated off to obtain the two solutes

c salt solution in water
d purple solution of iodine in carbon tetrachloride

© Diagram Visual Information Ltd.

Paper chromatography

1 Paper chromatography

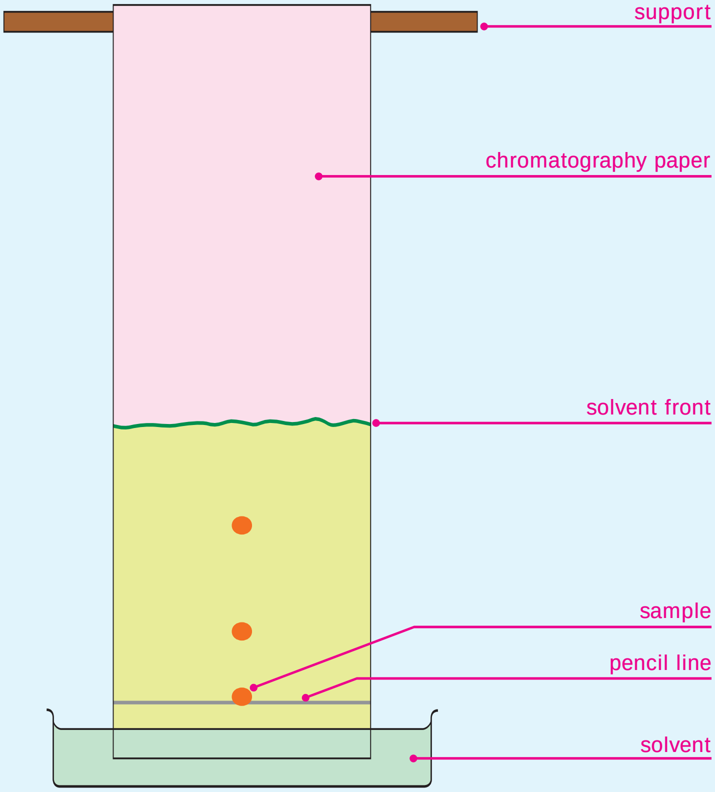

2 R_f value

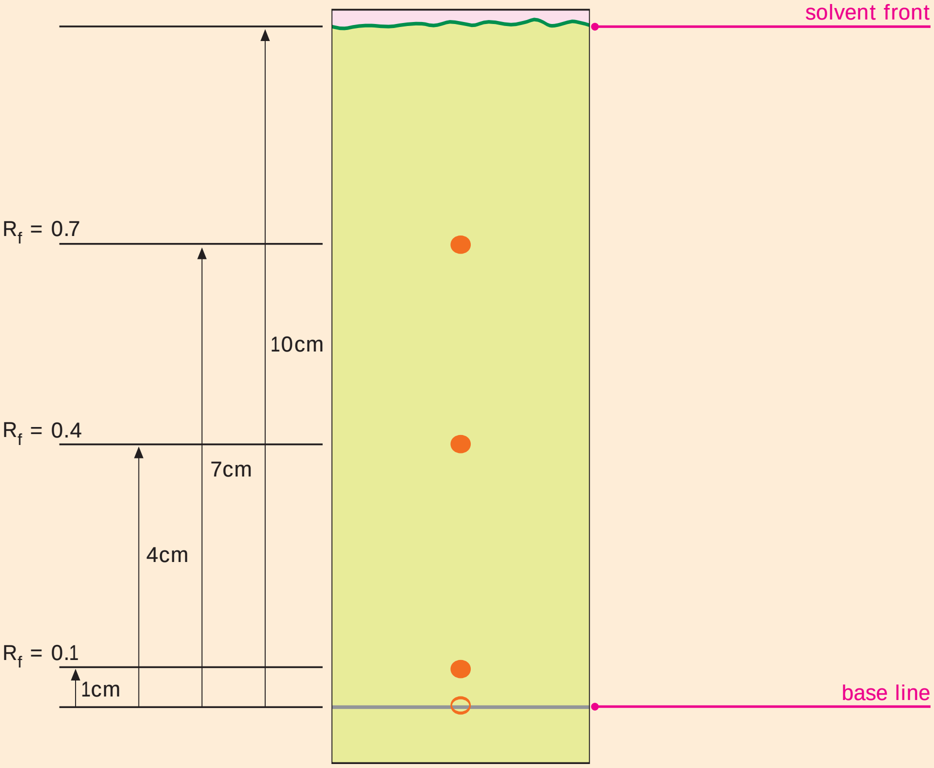

Key words

chromatography
R_f value
mixture
solute
solution
solvent

1 Paper chromatography

- *Chromatography* is a technique for separating and identifying *mixtures* of *solutes* in *solutions*.
- In paper chromatography, absorbent paper is suspended on a support so that only the bottom rests in the *solvent*.
- A base line is drawn in pencil above the level of the solvent. (If ink were used, the dyes in the ink would separate during the process and mix with the sample.)
- A concentrated solution of the sample mixture is made by dissolving as much as possible in a very small volume of solvent.
- A small amount of the concentrated solution is spotted onto the base line. The chromatography paper is suspended over the solvent.
- The solvent rises up the chromatography paper.

2 R_f value

- The *R_f value* is the ratio of the distance moved by a substance in a chromatographic separation to the distance moved by the solvent. The greater the attraction between a substance and the solvent molecules, the greater the R_f value.
- The molecules of each substance in a mixture are attracted both to the chromatography paper and to the solvent molecules.
- The greater the attraction between a substance and the solvent molecules, the quicker it will be carried up the chromatography paper. Dyes that are very soluble in the solvent are carried up to the top of the paper, while those that are less soluble remain lower down.
- The R_f value is independent of the height of the solvent front but is dependent on the solvent used.

© Diagram Visual Information Ltd.

Gas-liquid chromatography and mass spectrometry

Key words

- *alkane*
- *chromatography*
- *gas-liquid chromatography*
- *mass spectrometry*
- *mobile phase*
- *stationary phase*

1 Gas-liquid chromatography

- In chromatography, substances are partitioned between a *stationary phase* and a *mobile phase*. The stationary phase is the substance that retards the components of the sample. The mobile phase is the components of the sample.
- In *gas-liquid chromatography*, the stationary phase, packed into the column, consists of a high-boiling point liquid, such as a long-chain *alkane*, supported by a porous inert solid, such as charcoal or silica.
- The mobile phase consists of a carrier gas—usually nitrogen, hydrogen, helium, or argon.
- A sample mixture is injected into the chamber where it vaporizes and is carried through the column by the carrier gas. Various compounds in the sample pass through the column at different rates due to their attraction to the stationary phase.
- The separated compounds pass to a detector or directly into a mass spectrometer.

2 Mass spectrometry

- *Mass spectrometry* is a technique used to identify the chemical constitution of a substance by means of analyzing its ions.
- The sample passes into the ionization chamber, where it is bombarded by electrons and forms a series of positive ions.
- The ions are accelerated by an electric field and deflected along a circular path by a magnetic field. The lighter the ions, the greater the deflection.
- The intensity of the ion beam is detected electrically, amplified, and finally recorded.
- Each compound gives a characteristic spectrum from which it can be identified.

© Diagram Visual Information Ltd.

1 Gas-liquid chromatography

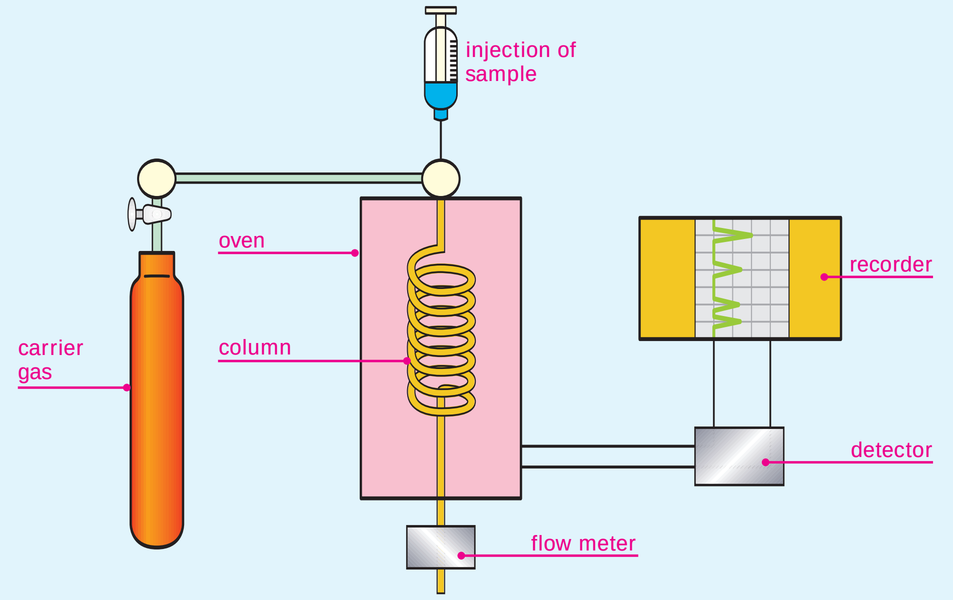
injection of sample
oven
recorder
carrier gas
column
detector
flow meter

2 Mass spectrometry

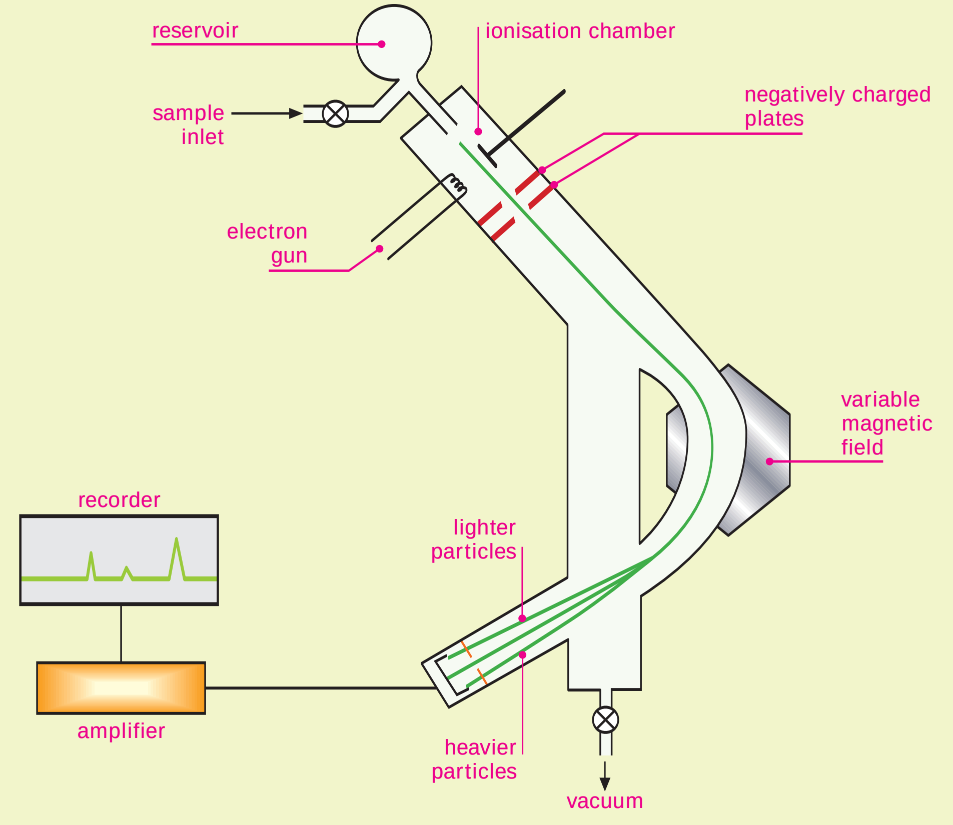
reservoir
ionisation chamber
sample inlet
negatively charged plates
electron gun
variable magnetic field
recorder
lighter particles
amplifier
heavier particles
vacuum

The pH scale

1 pH scale

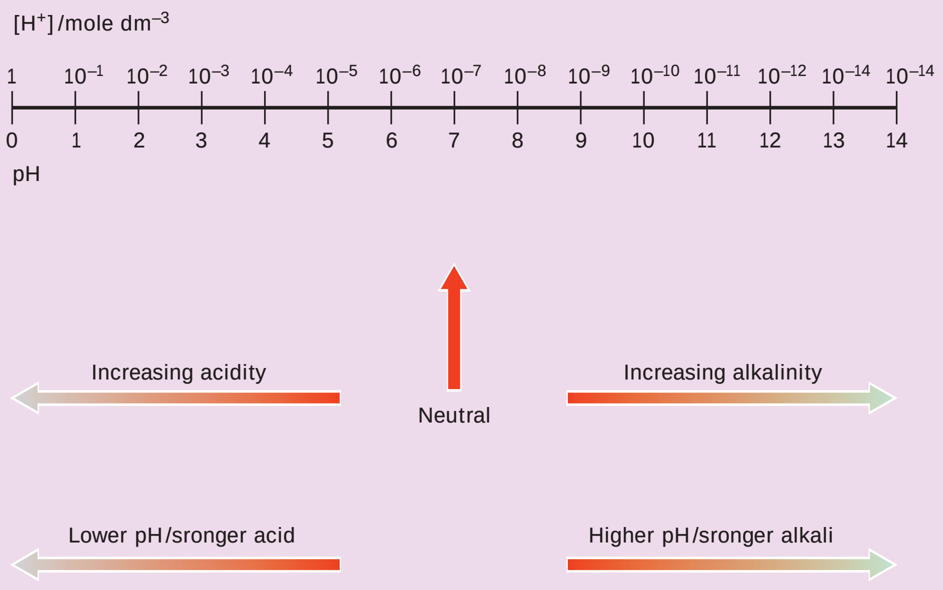
$[H^+]$/mole dm^{-3}
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
10^{-12}
10^{-14}
10^{-14}
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
pH
Increasing acidity
Neutral
Increasing alkalinity
Lower pH/sronger acid
Higher pH/sronger alkali

2 Schematic of a pH meter

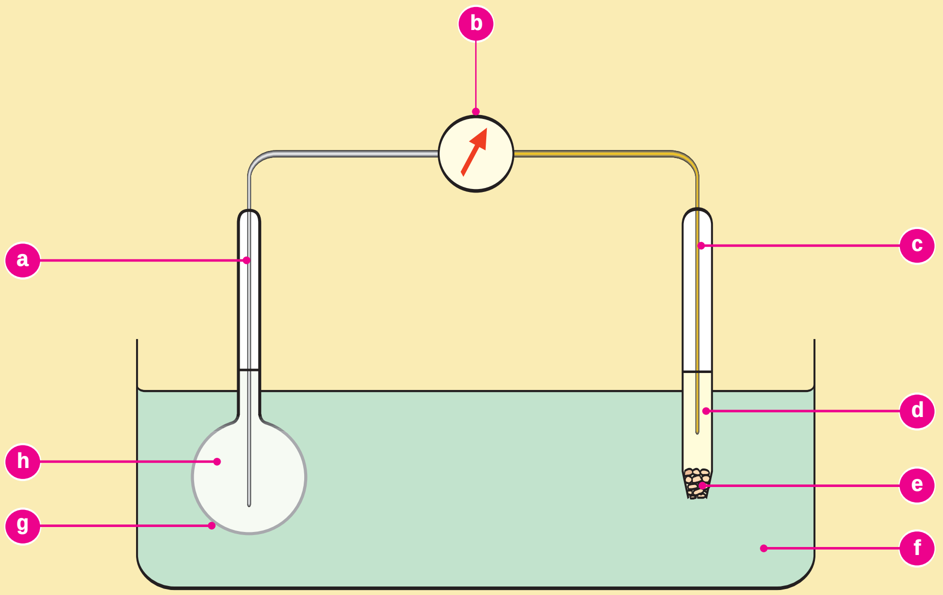

a platinum wire
b sensitive voltmeter
c silver wire coated with silver chloride (AgCl)
d saturated potassium chloride (KCl)
e capillary opening with porous plug
f solution of unknown pH
g Thin glass membrane through which H^+ ions can pass
h solution of fixed acid pH

Key words

acidity
alkalinity
pH
pH meter

1 pH scale

- *pH* is a measure of the *acidity* or *alkalinity* of a solution. The term pH was originally introduced by the Danish biochemist Søren Sørensen in 1909 while working on methods of improving the quality control of beer. The letters pH stand for "potential of hydrogen."
- Acidic solutions always have a pH of less than 7, and alkaline solutions always have a pH of more than 7. The lower the pH value, the more acidic the solution; conversely, the higher the pH value, the more alkaline the solution.
- The pH of a solution is the logarithm to base 10 of the reciprocal of the numerical value of the hydrogen ion concentration:
  $pH = lg(1/[H^+]0 = -lg [H^+]$
- The pH of a neutral solution can be calculated directly from the ionic product (K_w) of water:
  $K_w = [H^+][OH^-] = 10^{-14}\ mol^2\ dm^{-6}$
  For a neutral solution:
  $[H^+] = [OH^-] = 10^{-7}\ mol\ dm^{-3}$
  therefore the pH of a neutral solution = 7.
- The pH scale is logarithmic, so hydrogen ion concentration increases or decreases by a power of 10 for each step down or up the scale.

2 pH meter

- A *pH meter* is an electrochemical cell consisting of an electrode, such as a glass electrode, which is sensitive to hydrogen ion concentration, and a reference electrode.
- The emf (electromotive force) of the cell can be measured using a high-resistance voltmeter. A pH meter is a high-resistance voltmeter calibrated with the pH scale.

© Diagram Visual Information Ltd.

Indicators

Key words

acid
acid-base indicator
alkali
end point
equilibrium
titration
universal indicator

1 Common indicators

- *Acid-base indicators* are substances that are different colors in *acids* and *alkalis* so they “indicate” whether a solution is an acid or alkali.

2 Changing equilibrium

- Acid-base indicators are usually weak acids that disassociate to give an ion that is a different color than the acid. A change in pH causes a change in the position of the equilibrium of the reaction and, therefore, the color of the solution.
- Phenolphthalein is such an indicator. It is a colorless, weak acid that dissociates in water, forming pink anions. Under acidic conditions, the *equilibrium* of the reaction is to the left, and the concentration of the anions is too low for the color to be visible. Under alkaline conditions, the equilibrium is to the right, and the concentration of anions is high enough for the pink to be seen.

3 Universal indicator

- In contrast to an indicator such as phenolphthalein, which is able to show whether a substance is an acid or base only in the broadest terms, a *universal indicator* has a range of colors that indicate how acidic or how alkaline a solution is.

4 pH range of indicators

- Most indicators do not change color when the pH of a solution is exactly 7. This means that the *end point* of the *titration*, the point at which the indicator undergoes the maximum color change, occurs at a different time to the equivalence point of the titration, the point at which there are equivalent amounts of acid and alkali.
- The suitability of an indicator for use in a titration depends on what combination of strong and weak acid and alkali is to be used.

© Diagram Visual Information Ltd.

1 Table of common indicators

Indicator	Color in acid	Color in alkali
litmus	red	blue
methyl orange	red	yellow
phenolphthalein	colorless	pink

2 Changing equilibrium (phenol phthalein)

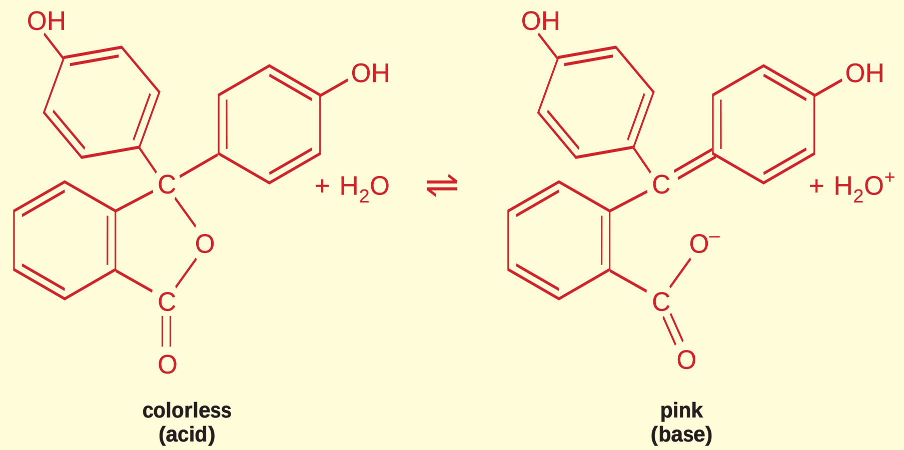

3 Universal indicator

pH														
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
red					orange	yellow	green	blue			purple			
Color														

4 Table of pH range over which acid–base indicators change color

Indicator	Color		pH range over which color change occurs
	Acid	Alkali	
bromocresol green	yellow	blue	3.8–5.4
bromothymol blue	yellow	blue	6.0–7.6
methyl orange	red	yellow	3.2–4.4
methyl red	yellow	red	4.8–6.0
phenolphthalein	colorless	pink	8.2–10.0
phenol red	yellow	red	6.8–8.4

Titration of strong acids

1 Titration of strong acid against strong alkali

(pH changes during the titration of 50 cm³ of 0.1M HCl with 0.1M NaOH)

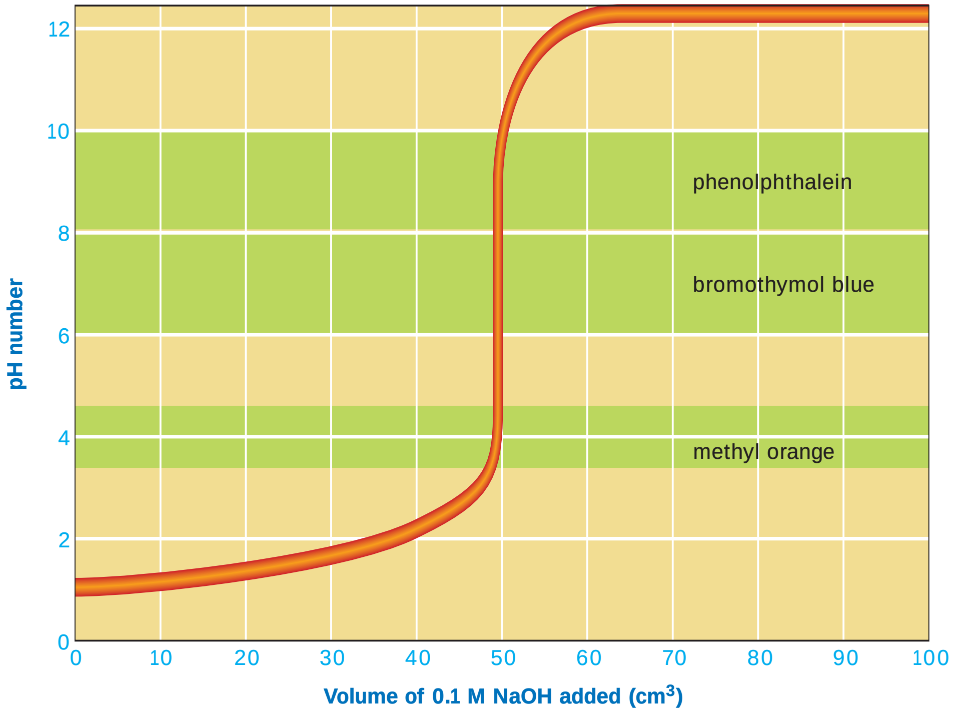

2 Titration of strong acid against weak alkali

(pH changes during the titration of 50 cm³ of 0.1 M HCl with 0.1 M NH_3)

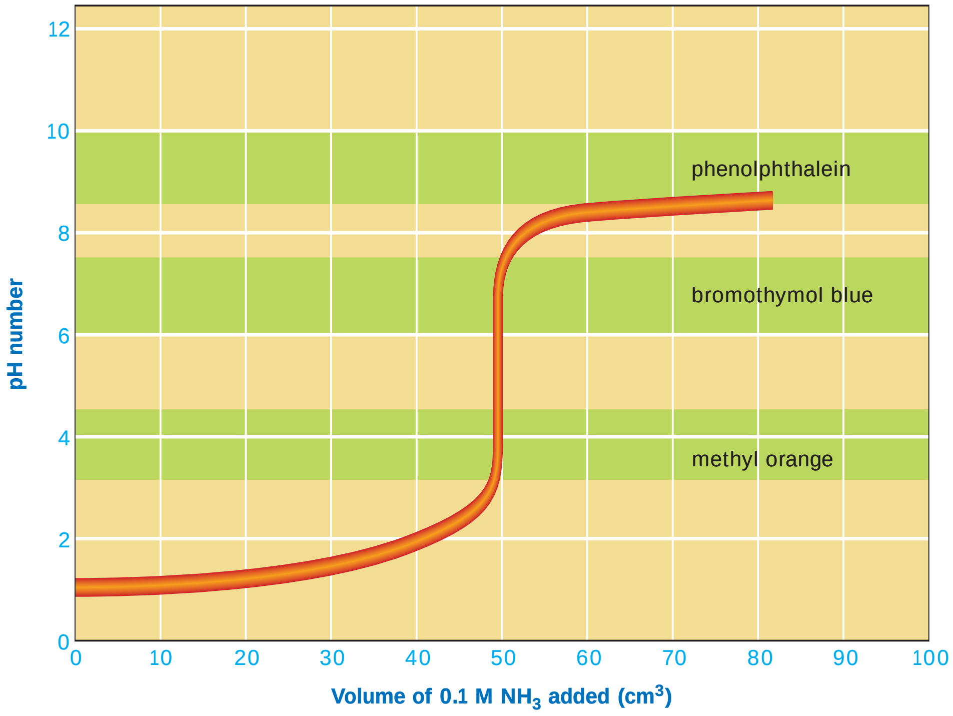

Key words

acid	*titration*
alkali	
base	
end point	
pH	

1 Titration of strong acid against strong alkali

- At the *end point* of strong *acid*–strong *base alkali titration*, the *pH* changes by 5 or 6 pH units when only 1 drop of acid or alkali is added.
- Methyl orange, bromothymol blue, and phenolphthalein are all suitable indicators for this titration because they all change color within a very small change in volume of sodium hydroxide solution.

2 Titration of strong acid against weak alkali

- At the end point of a strong acid–weak alkali titration, the pH change for the addition of one drop of acid or alkali is significant. However, the pH at equivalence (when there are equivalent amounts of acid and alkali) is less than 7. A suitable indicator should change color below or around pH 7. Thus both methyl orange and bromothymol blue would be suitable indicators because they change color between pH 3.2 and 7.6. Within this range, the pH of the titration mixture changes significantly for a very small change in volume of sodium hydroxide solution.
- Phenolphthalein would not be a good choice of indicator because it changes color between pH 8.2 and 10.0. In order to change the pH of the titration mixture over this pH range, a significant volume of sodium hydroxide solution must be added. The result would be an overestimate of the volume of sodium hydroxide solution needed to neutralize the acid.

© Diagram Visual Information Ltd.

Titration of weak acids

Key words

acid	*titration*
alkali	
end point	
equivalence point	
pH	

1 Titration of weak acid against strong alkali

- At the *end point* of a weak *acid*–strong *alkali titration*, the *pH* change for the addition of one drop of acid or alkali is significant. However, the pH at equivalence (when there are equivalent amounts of acid and alkali) is greater than 7. A suitable indicator should change color above pH 7. Thus phenolphthalein would be a good choice because it changes color between pH 8.2 and 10.0. Within this range, the pH of the titration mixture changes significantly for a very small change in volume of sodium hydroxide solution.
- Conversely, bromothymol blue and methyl orange would not be good choices because they change color between pH 3.2 and 7.6, which is before the *equivalence point* of the titration is reached. The result would be an underestimate of the volume of sodium hydroxide solution needed to neutralize the acid.

2 Titration of weak acid against weak alkali

- The pH changes too slowly around the equivalence point to give a color change with the addition of one drop of acid or alkali. The use of methyl orange, bromothymol blue, or phenolphthalein would not give accurate results.
- It is not usual to titrate weak acids with weak alkali, but if it must be done, a pH meter is necessary to find the equivalence point accurately. There is no suitable indicator for this type of titration.

1 Titration of weak acid against strong alkali

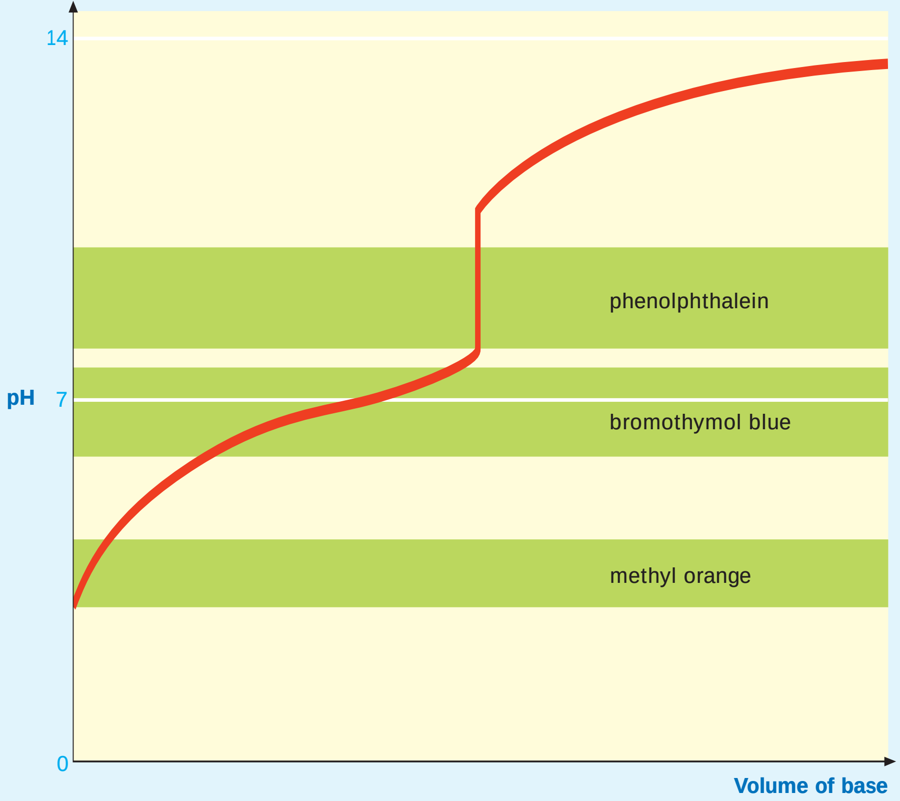

2 Titration of weak acid against weak alkali

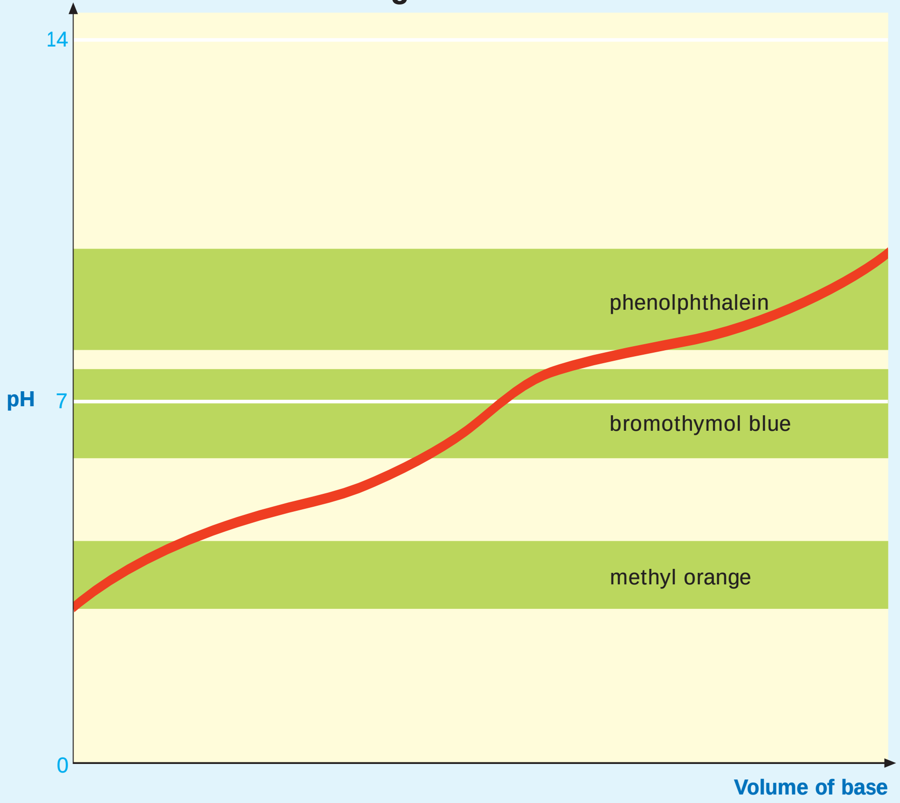

© Diagram Visual Information Ltd.

pH and soil

1 Soil classification

pH	Description
< 5.5	strongly acid
5.5 – 5.9	medium acid
6.0 – 6.4	slightly acid
6.4 – 6.9	very slightly acid
7.0	neutral
7.1 – 7.5	very slightly alkaline
7.6 – 8.0	slightly alkaline
8.1 – 8.5	medium alkaline
> 8.5	strongly alkaline

2 pH range of common fruit and vegetables

Fruit or vegetable	Soil pH range
cabbage	6.0 – 7.5
cauliflower	6.5 – 7.5
celery	6.5 – 7.5
cucumber	5.5 – 7.0
potato	5.0 – 6.0
peas	6.0 – 7.5
strawberry	5.0 – 6.0
tomato	5.5 – 7.0

3 Elements needed by plants

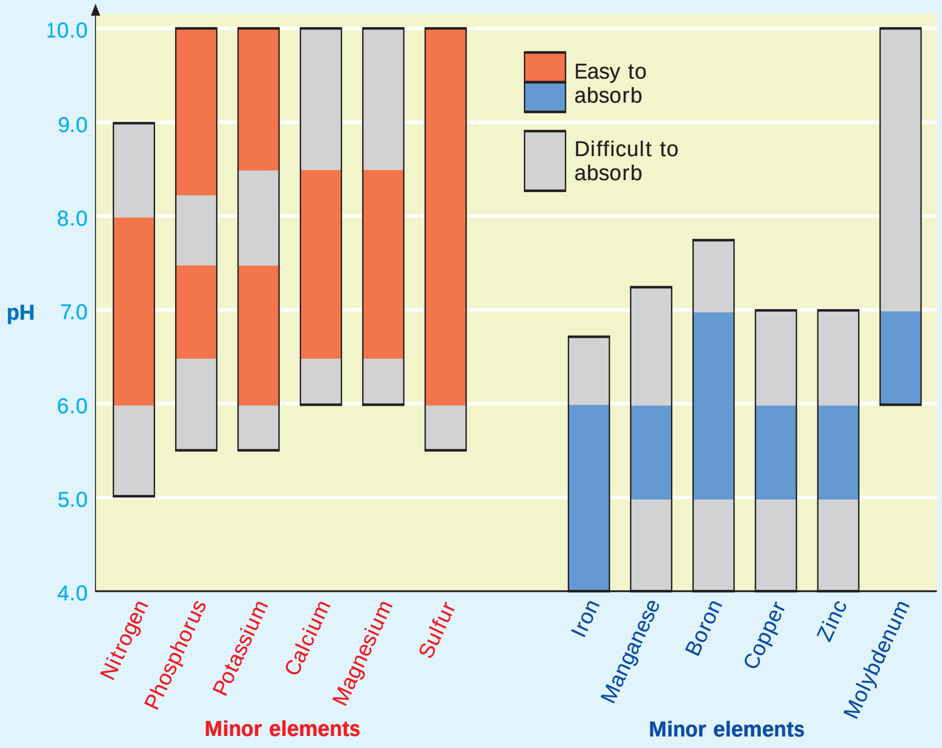

Key words

pH

1 Soil classification

- Soil can be classified according to its *pH*.
- Soils naturally tend to become more acidic due to organic acids being released into the soil as a result of the decay of organic material.
- The acidity of soil can be reduced by spreading slaked lime (calcium hydroxide) or lime (calcium carbonate).

2 pH range of common fruits and vegetables

- Most plants grow best in soil that is slightly acidic, with a pH value between 6.3 and 7.2. Plants will grow outside this range but not as well. This has serious implications for food crops.
- The soil pH is an important consideration in preparing soil to grow crops.

3 Elements needed by plants

- Plants need a number of major and minor elements in order to grow well, and they obtain these from the soil. The minerals dissolve in soil water and are absorbed into the plant through the roots.
- The pH of the soil determines how easily minerals containing these elements can be absorbed. At soil pH values between 6.0 and 7.0, all major elements and minor elements can be absorbed, although some are absorbed more easily than others. In very acidic or very alkaline soils, relatively few plants prosper because they cannot absorb all of the minerals needed for healthy growth.

© Diagram Visual Information Ltd.

Key words

atmosphere
convection current

The water cycle

- Earth’s water is always moving in a cycle called the hydrologic or water cycle.
- The Sun provides the energy driving the cycle.

1 Evaporation

- Heat energy causes water to evaporate from the surface of the oceans, leaving all dissolved substances behind. The rate of evaporation is greater in areas of Earth where the seas are warmer.
- Water vapor rises into the *atmosphere*, where it eventually condenses to form clouds. These are dispersed by winds, which carry them to the colder regions of Earth.

2 Transportation

- When clouds reach landmasses, they are carried up on *convection currents*. As they rise, the temperature decreases, and eventually the water vapor condenses, forming precipitation, which falls to Earth.
- Rainwater contains dissolved gases, which makes it slightly acidic.

3 Deposition

- The fresh water flows over rocks and through soils before gathering in streams and rivers. As the water flows through the ground, solids dissolve in it.
- Water is removed from rivers for both industrial and domestic use. Much of this water is ultimately returned to the rivers. Finally, the water flows out to sea, thus completing the cycle. Any dissolved solids are carried in it and eventually deposited in the oceans.

© Diagram Visual Information Ltd.

The water cycle

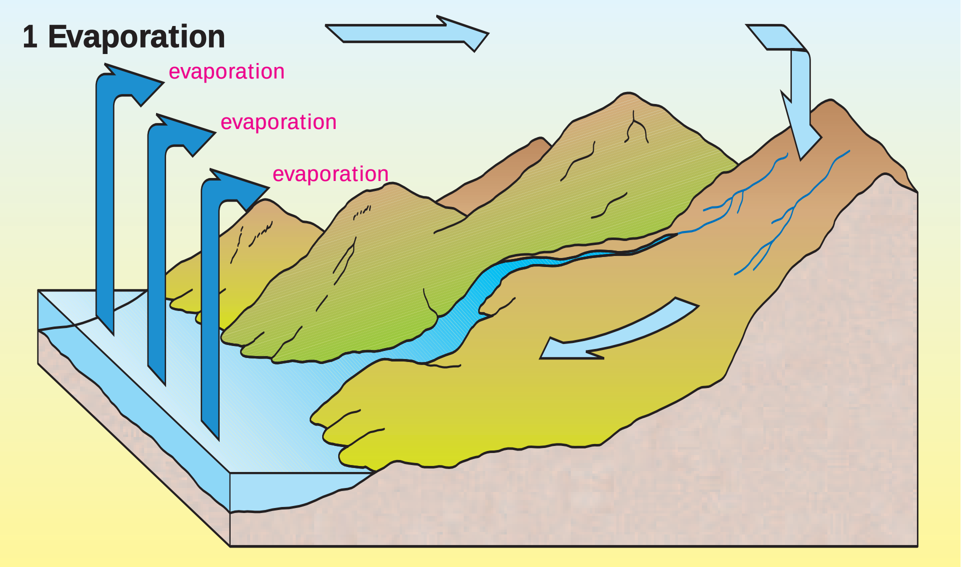
1 Evaporation

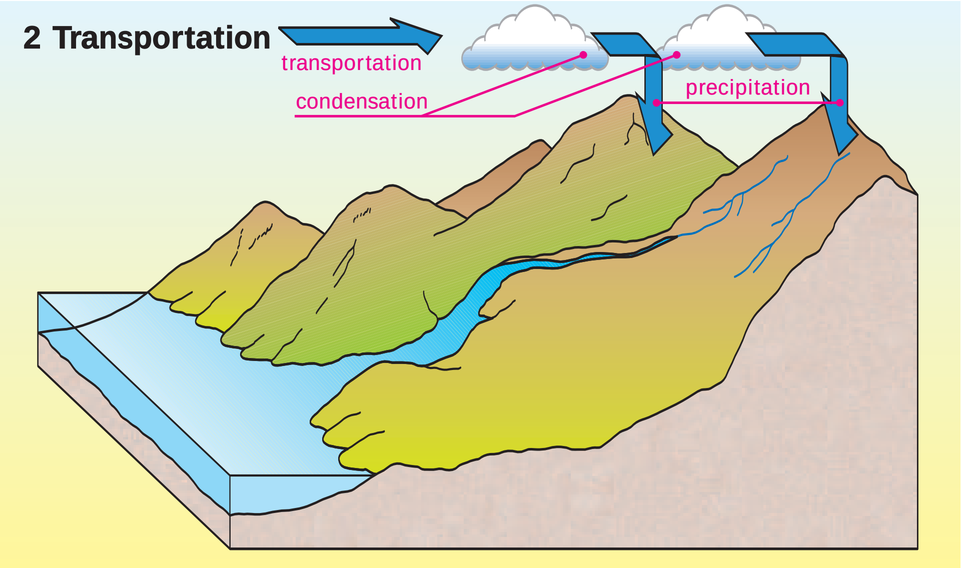
2 Transportation

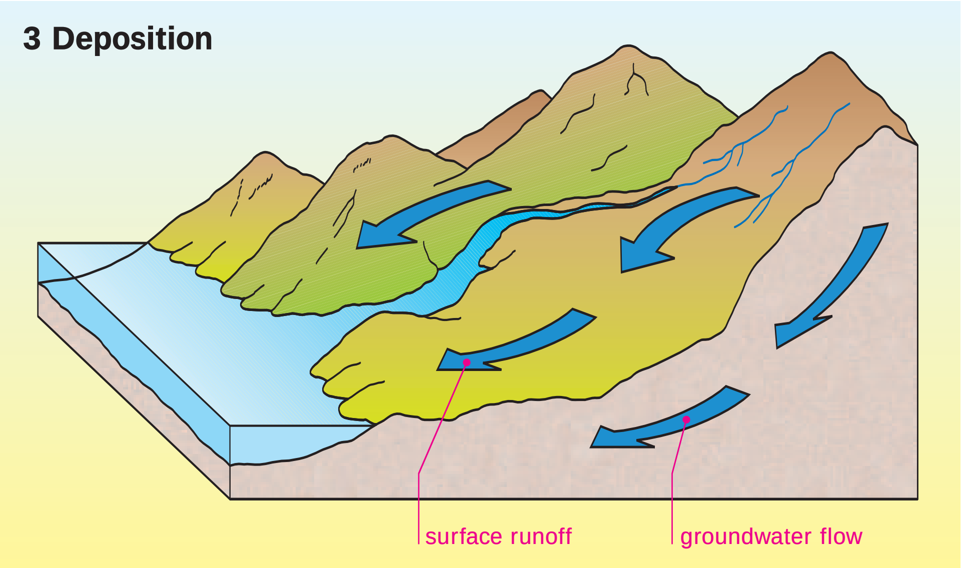
3 Deposition

Treatment of water and sewage

1 water treatment

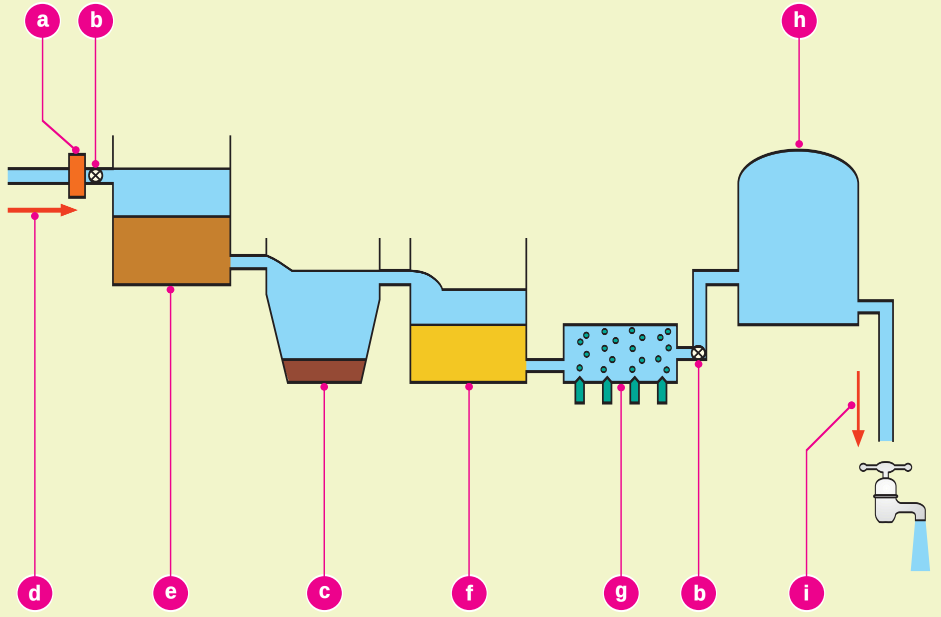

2 Sewage treatment

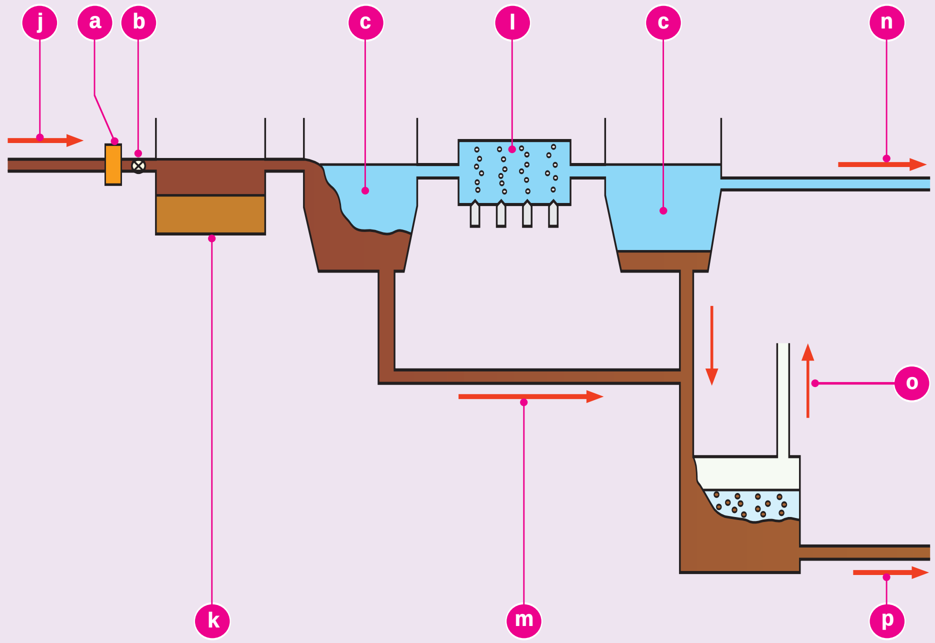

a screen
b pump
c sedimentation tank
d water in
e coarse sand filter
f fine sand filter
g chlorine added
h covered storage tank
i to homes and factories
j sewage in
k settling tank
l digester aeration tank
m sludge collected
n clean water to river
o methane out
p digested sludge out

Key words

methane
oxidizing agent
sewage

1 Water treatment

- Particles are removed from water by passing it through a series of sand filter beds and sedimentation tanks. The filter beds also contain bacteria, which break down and destroy micro-organisms in the water.
- Chlorine is a powerful *oxidizing agent* that is used to kill any remaining microorganisms in the water before is stored ready for distribution. Storage tanks are covered to prevent the entry of foreign bodies.

2 Sewage treatment

- Raw *sewage* cannot be released into rivers because of the threat to health and the effects on the environment. The waste materials it contains must first be broken down by the action of decomposing bacteria.
- Solids are removed from the sewage by a series of screens and settling tanks. The remaining liquid passes into a digester, where bacteria break down the waste products. Streams of air are blown into the tank in order to provide the bacteria with the oxygen needed to survive and to keep the mixture circulating.
- After settling, the clean water is allowed to pass into the river, while the sludge undergoes further digestion during which *methane* is released. The digested sludge contains nitrogenous compounds and is often used as a fertilizer.

© Diagram Visual Information Ltd.

Key words

bond	*oxygen*
covalent compound	
hydrogen	
lone pair	

1 A covalent compound

- Water is essentially a *covalent compound* formed by two atoms of *hydrogen* and one atom of *oxygen*.

2 The water molecule

- The oxygen atom in a water molecule has two pairs of bonding electrons (sometimes called shared pairs) and two pairs of non-bonding electrons (sometimes called *lone pairs*).
- These four pairs of electrons are directed toward the corners of a tetrahedron. However, the tetrahedral shape is distorted. The non-bonding pairs of electrons repel each other more strongly than the bonding pairs of electrons. Repulsion between these and the bonding pairs of electrons reduces the angle between the oxygen–hydrogen bonds to 104.5 °.

3 The polar nature of the molecule

- Oxygen is more electronegative than hydrogen and, therefore, has a stronger attraction for the electrons in the oxygen–hydrogen *bond*. The result is that the electrons in the bond reside closer to the oxygen atom. Since electrons are negatively charged, this leaves the oxygen atom slightly negative and the hydrogen atom slightly positive. This is shown using δ notation; oxygen is δ- and hydrogen is δ+.

© Diagram Visual Information Ltd.

The water molecule

1 A covalent compound

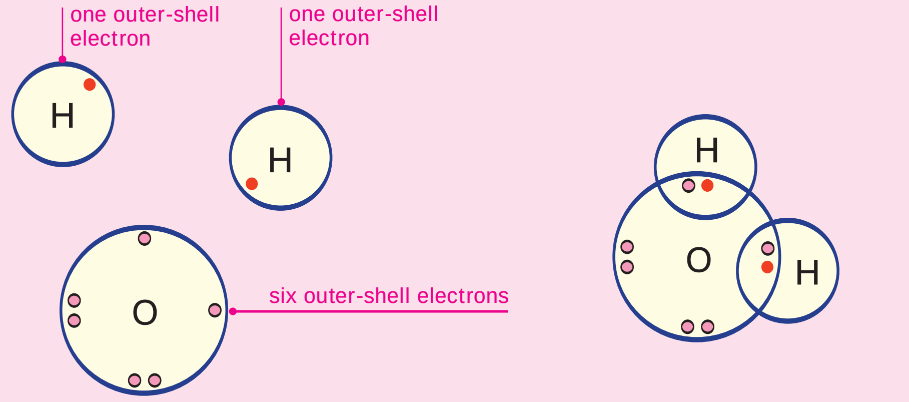

Atoms present

Full shells of a water molecule

2 The water molecule

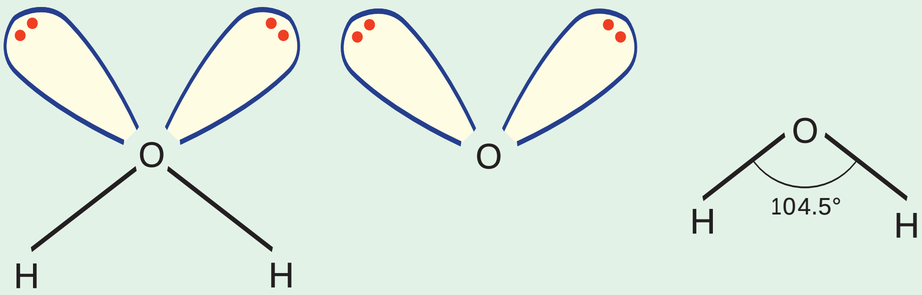

Water molecule

Non-bonding pairs (lone pairs)

Bonding pairs

3 The polar nature of the molecule

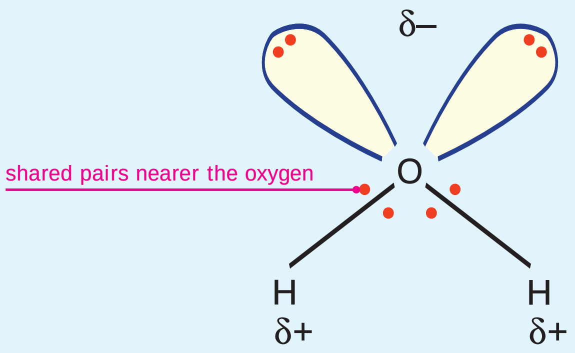

Water as a solvent of ionic salts

1 Ionic lattices

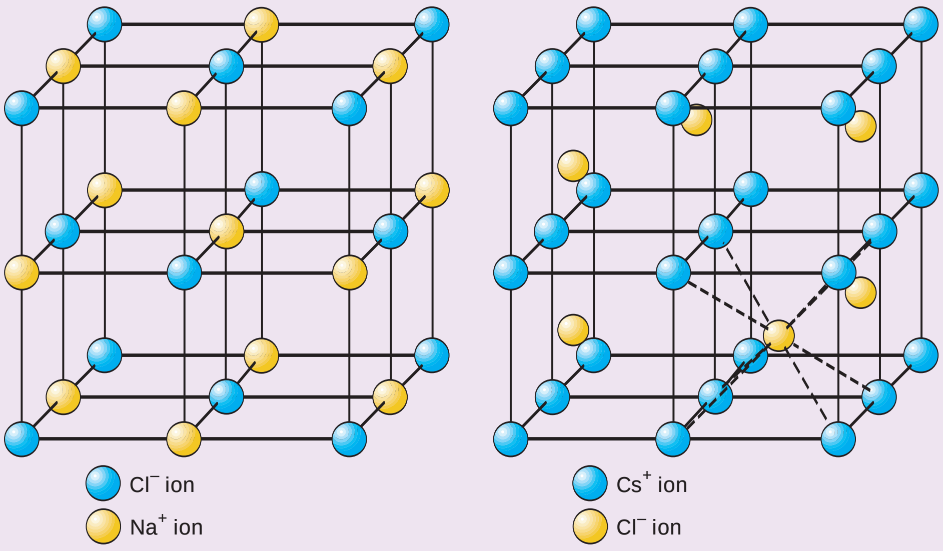

2 The effect of water on an ionic lattice

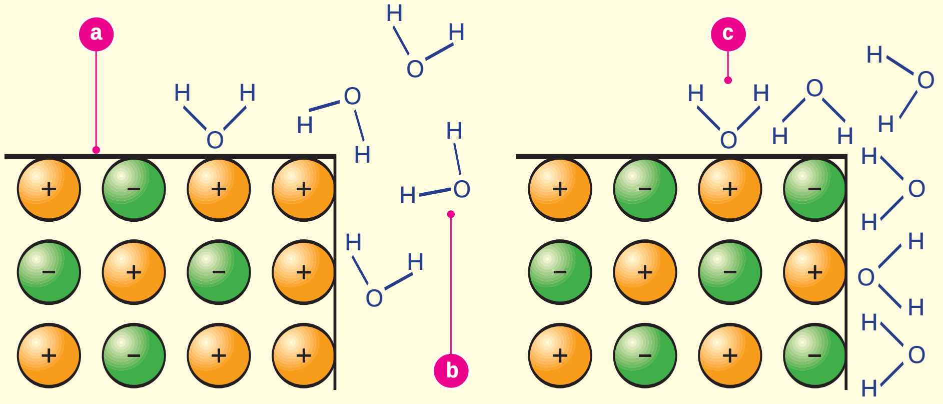

Ionic lattice put into water

The ions attract the water molecules

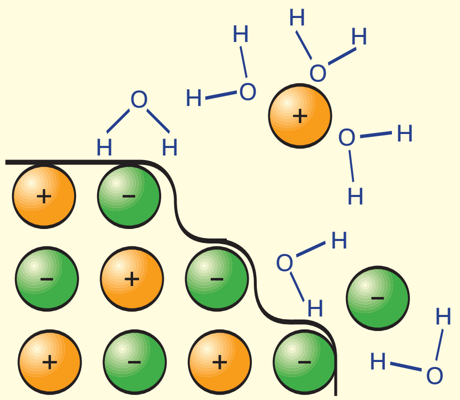

The lattice starts to split up

a lattice
b water molecules

c charged ends of molecules attracted to ions of opposite charge

Key words

anion
body-centered cubic
cation
face-centered cubic
hydration
ionic crystal lattice

1 Ionic lattices

- The ions in an *ionic crystal* are arranged in a *lattice*. Each ion is surrounded by a number of oppositely charged ions. The lattice structure is determined by:
  - the ratio of the number of positively charged ions (*cations*) to negatively charged ions (*anions*)
  - the ratio of the radii of the ions (r_{cation} / r_{anion})
- The radius ratio in sodium chloride is 0.57. The ions are arranged in a *face-centered cubic* structure in which each sodium ion is surrounded by six chloride ions, and each chloride ion is surrounded by six sodium ions.
- The radius ratio of cesium chloride is 0.94 (due to the larger cesium ion). The ions are arranged in a *body-centered cubic* structure in which each cesium ion is surrounded by eight chloride ions, and each chloride ion is surrounded by eight cesium ions.

2 The effect of water on an ionic lattice

- When an ionic compound is placed in water, the water molecules collide with the lattice. If the water molecules collide with sufficient energy to overcome the forces of attraction between the oppositely charged ions, *hydration* occurs, and the compound will dissolve, forming a solution.

© Diagram Visual Information Ltd.

Ionic solutions

Key words

anion
bond
cation
ion
ionic compound
lattice
transition metals

Ionic solutions

- Ionic solutions are *ionic compounds* dissolved in water.

1 Stabilizing free ions

- Due to the uneven sharing of electrons in the oxygen–hydrogen *bonds* of a water molecule, the oxygen atom is slightly negatively charged and the hydrogen atoms are slightly positively charged.
- Water molecules surround and stabilize the *ions* in the compound: positively charged *cations* are stabilized by the negative oxygen atoms, and negatively charged *anions* are stabilized by the positive hydrogen atoms.

2 Transition metals in solution

- *Transition metal* ions form complexes with water.
- Transition metals have an incomplete outer shell and can fill this with the electric charge on the water molecule. Non-bonding pairs of electrons from the water molecules are donated to form coordinate bonds.
- The bonds between the metal ions and water are so strong that they remain when solids are obtained from their solutions.

3 Production of silver chloride

- When silver nitrate solution and sodium chloride solution are mixed, insoluble silver chloride forms a white precipitate.

  $Ag^+(aq) + Cl^-(aq) \rightarrow AgCl(s)$
- The silver ions and chloride ions are more stable when bonded together in an ionic *lattice* than existing apart surrounded by water molecules.

1 Stabilizing free ions

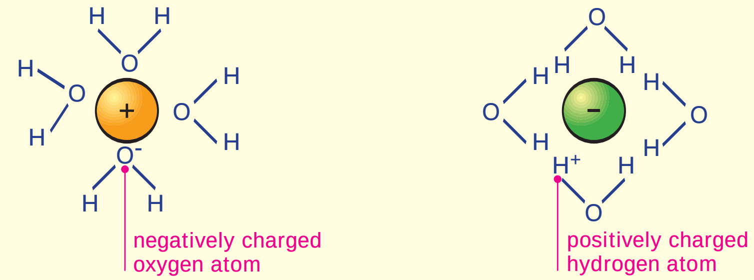

negatively charged oxygen atom

positively charged hydrogen atom

2 Transition metals in solution

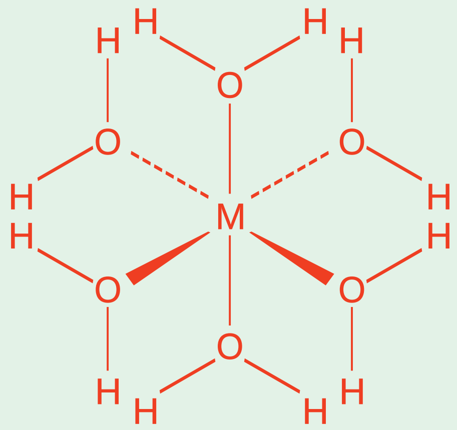

3 Production of silver chloride when a metal chloride is added to silver nitrate solution

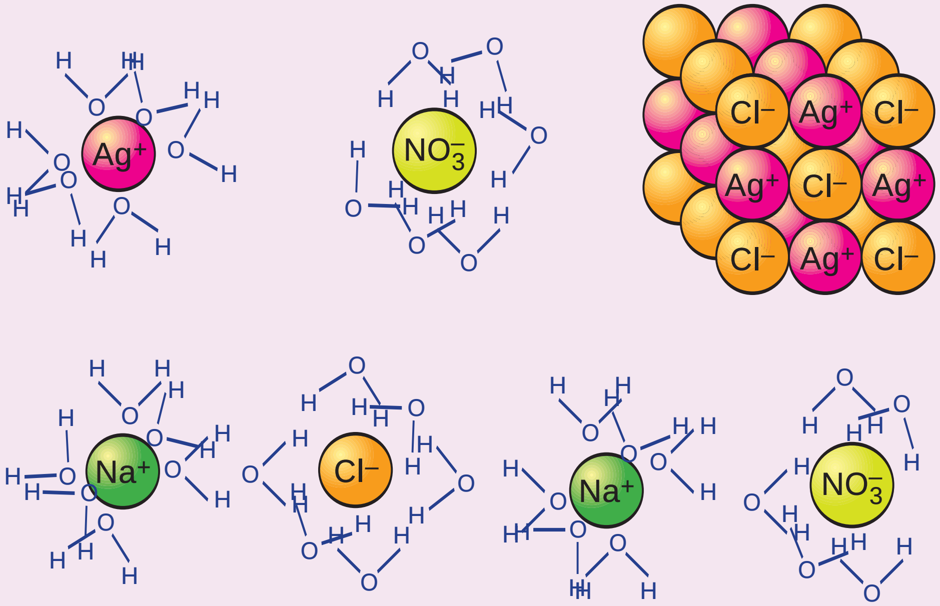

Silver nitrate solution and sodium chloride solution

After reaction, clusters of water molecules have been forced away from aqueous silver and chloride ions to leave solid silver chloride

© Diagram Visual Information Ltd.

Solubility

1 Table of solubility of ionic compounds

Soluble	Insoluble
all salts of ammonium, potassium, and sodium	
all nitrates	
most bromides, chlorides, and iodides	lead and silver bromides, chlorides, and iodides
most sulphates (calcium sulfate is slightly soluble)	barium and lead sulfate
ammonium, potassium, and sodium carbonate	most other carbonates
ammonium, potassium, and sodium hydroxides (calcium hydroxide is slightly soluble)	most other hydroxides

2 Table of the most abundant compounds in seawater

Name of compound	Formula of compound	Percentage of solids in seawater
sodium chloride	$NaCl$	78
magnesium chloride	$MgCl_2$	9
magnesium sulphate	$MgSO_4$	7
calcium sulphate	$CaSO_4$	4
potassium chloride	KCl	2
calcium carbonate	$CaCO_3$	less than 1
magnesium bromide	$MgBr_2$	less than 1

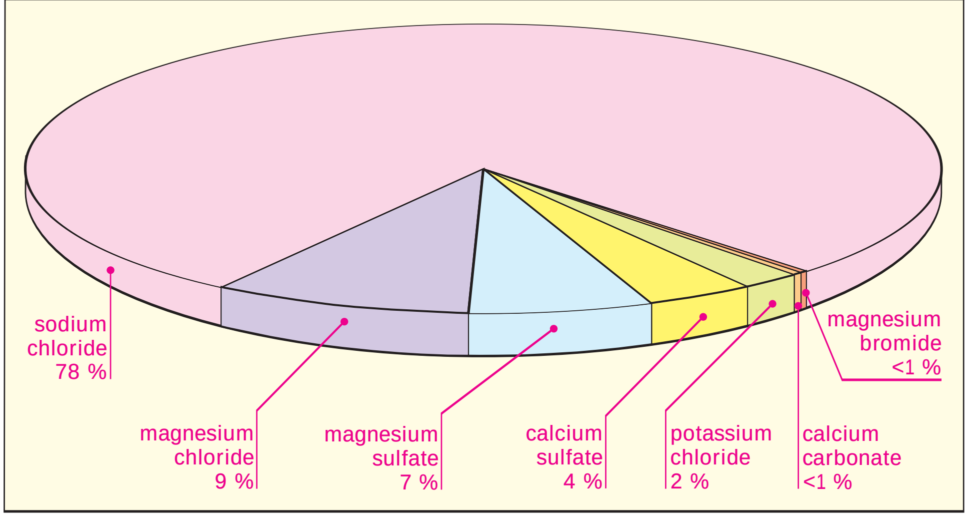

Key words

ionic compound
insoluble
soluble

1 Solubility of ionic compounds

- All *ionic compounds* are *soluble* in water to some extent. However the solubility of some is so low that they are best regarded as *insoluble*. Solubility generally follows the following rules.
- All ammonium, potassium, and sodium salts are soluble.
- All nitrates are soluble.
- With the exception of lead and sliver bromides, chlorides, and iodides, bromides, chlorides, and iodides are soluble.
- Most sulfates are soluble, with the exception of barium and lead sulfate.
- Most carbonates are insoluble, with the exception of ammonium, potassium, and sodium carbonate.
- Most hydroxides are insoluble, with the exception of ammonium, potassium, and sodium hydroxides.
- Calcium hydroxide is only slightly soluble.

2 Most abundant compounds in seawater

- Seawater is a solution of many different salts. The main salt present in seawater is sodium chloride.
- The concentration of solids in seawater depends on the location. The saltiest water occurs in the Red Sea, where there is 40 g of dissolved solids per 1,000 g of water. The North Atlantic is the saltiest of the major oceans, with an average of 37.9 g of dissolved solids per 1,000 g of water. The least salty waters are found in polar seas and the Baltic Sea, which contains only 5–15 g of dissolved solids per 1,000g of water.

© Diagram Visual Information Ltd.

Solubility curves

Key words

solubility curve

Expressing solubility

- Solubility is normally expressed in g / 100 g of water. The solubility of a compound varies (normally increases) with temperature, so when quoting solubility, it is necessary to state the temperature for which it is given.
- A *solubility curve* is a graph that shows how the solubility of a salt varies between 0°C (the freezing point of water) and 100°C (the boiling point of water).
- The solubility curve of a compound is plotted using data about the solubility of the compound over the whole temperature range. Solubility curves are generally not straight lines.
- Over the temperature range 0–100°C, the solubility of some salts remains nearly constant (sodium chloride), while the solubility of others either increases gradually (potassium sulfate) or increases very rapidly (potassium nitrate).

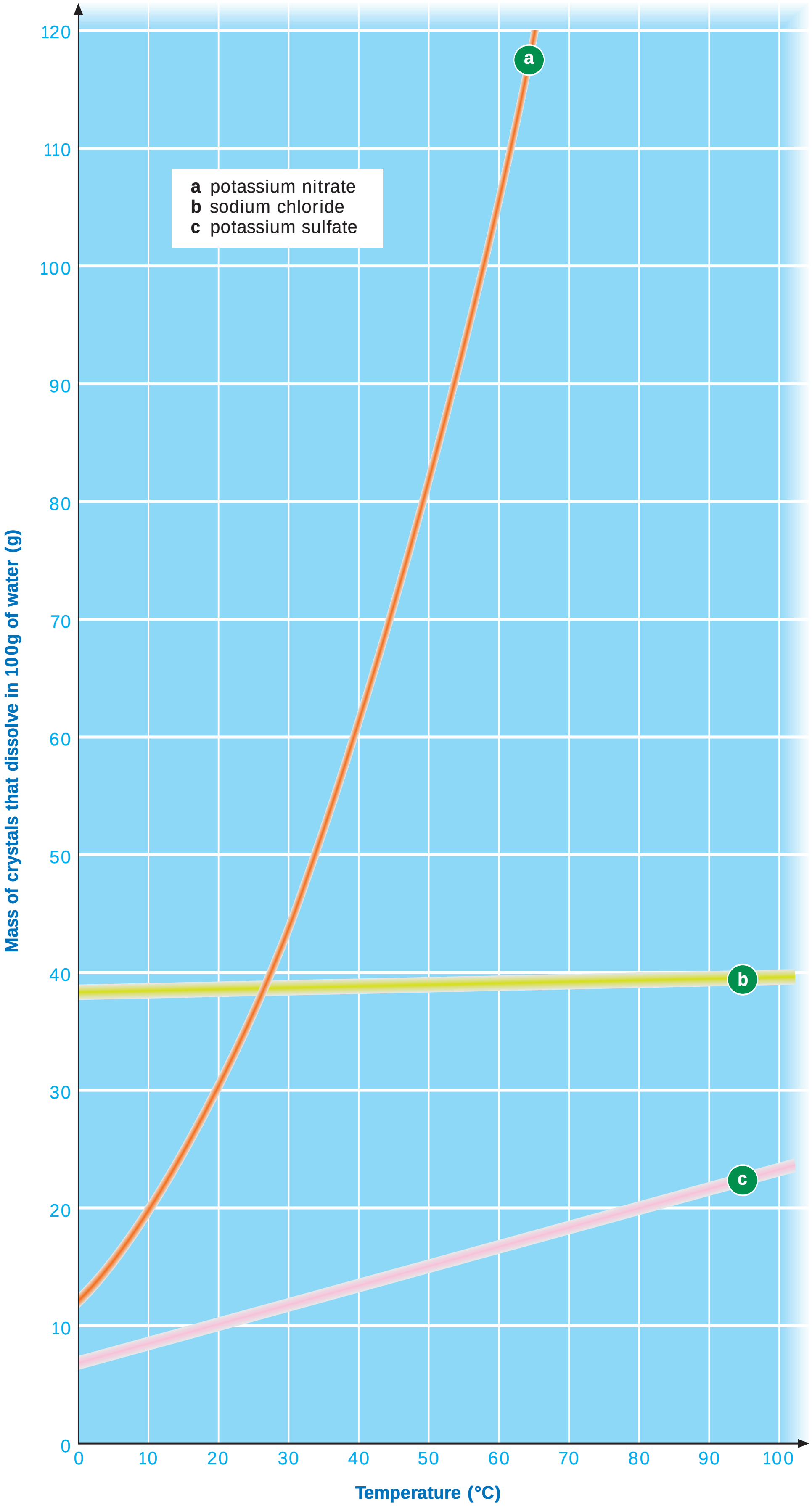

© Diagram Visual Information Ltd.

Solubility of copper(II) sulfate

1 Table of solubility of copper(II) sulfate at different temperatures

Temperature °C	Solubility g/100g water	Temperature °C	Solubility g/100g water
0	14.3	60	40.0
10	17.4	70	47.1
20	20.7	80	55.0
30	24.3	90	64.8
40	28.5	100	75.4
50	34.0		

2 Solubility curve for copper(II) sulfate

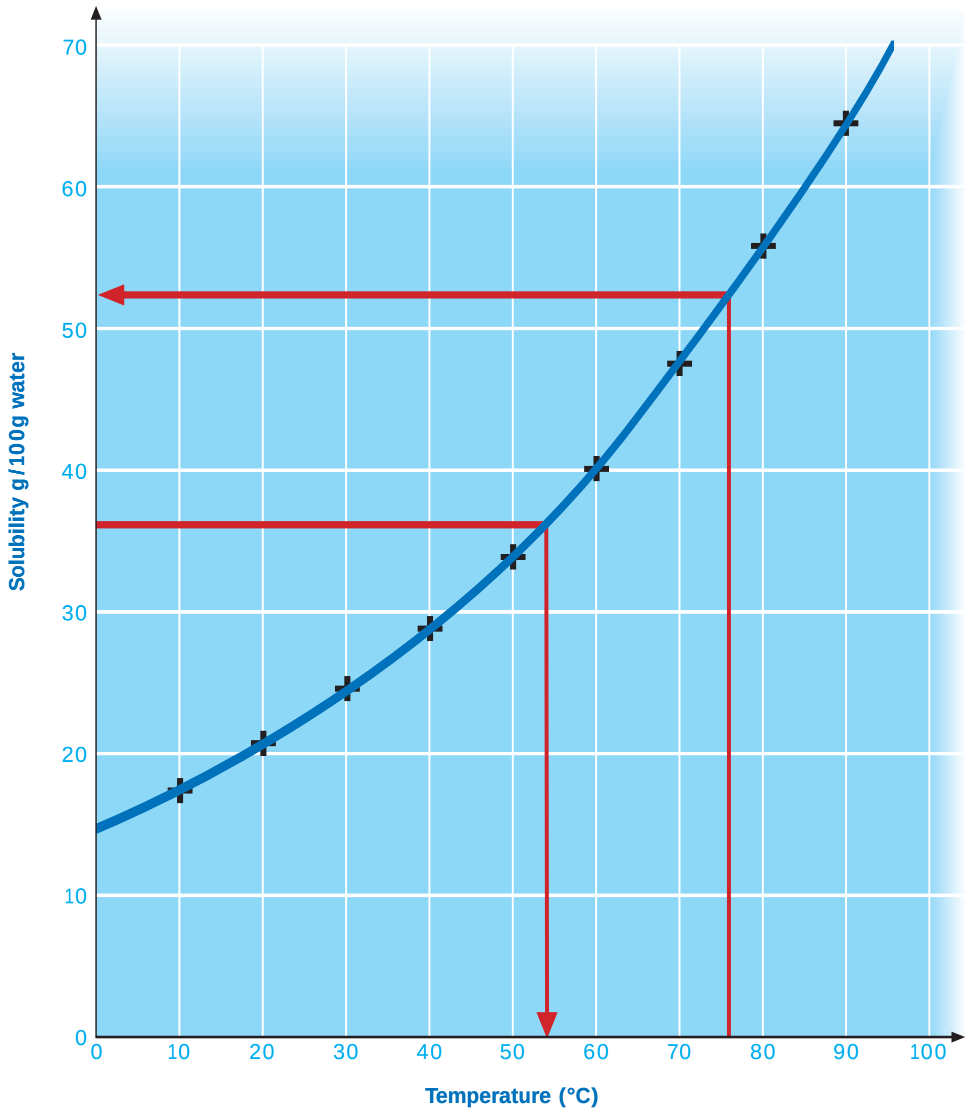

Key words

saturated
solute
solution
solubility curve

1 Solubility of copper(II) sulfate at different temperatures

- When no more solid will dissolve in a *solution* at a given temperature, the solution is said to be *saturated*.
- Normally when a solution is cooled below the saturation temperature for the quantity of *solute* present, some solute crystallizes out. Under certain conditions a solution may be cooled below this temperature without crystallization occurring. Such a solution is said to be supersaturated.
- A *solubility curve* is plotted by finding the amount of solid needed to make a saturated solution at a number of different temperatures over the range 0–100°C.

2 Solubility curve for copper(II) sulfate

- The solubility of copper(II) sulfate at 76°C is found by drawing a vertical line from 76°C to the solubility curve and then a horizontal line to the solubility axis. The value is 52 g of copper(II) sulfate per 100 g of water.
- The temperature at which the solubility of copper(II) sulfate is exactly 36 g per 100 g of water is found by drawing a horizontal line from 36 g / 100 g water to the solubility curve and then a vertical line to the temperature axis. The temperature is 54°C.
- The solubility of copper(II) sulfate at 90°C is 64.8 g / 100 g water and at 20°C is 20.7 g / 100 g water. If 100 g of saturated copper(II) sulfate solution was allowed to cool from 90°C to 20°C the mass of copper(II) sulfate crystals formed would be 64.8 – 20.7 = 44.1 g.
- The size of crystals formed is related to how quickly the solution is cooled. If cooling is rapid many small crystals are formed but if cooling is slow a smaller number of large crystals are formed.

© Diagram Visual Information Ltd.

Hydrogen: preparation

Key words

anhydrous
dry gas
hydrochloric acid
hydrogen

1 Obtaining hydrogen using Kipps apparatus

- Traditionally, *hydrogen* is generated using a Kipps apparatus. When the tap is opened, dilute *hydrochloric acid* floods the bottom compartment and the level rises until it reacts with granulated zinc in the middle compartment. Zinc reacts with dilute hydrochloric acid to produce hydrogen:
  $Zn(s) + 2HCl(aq) \rightarrow ZnCl_2(aq) + H_2(g)$
- If the tap is closed, hydrogen continues to be produced for a short time, and the pressure of the gas in the middle compartment gradually increases. Eventually the pressure is sufficient to force the dilute hydrochloric acid back down and out of the middle compartment, and the reaction stops.
- In a modern laboratory, hydrogen is often obtained directly from a cylinder of the gas.

2 Collecting dry hydrogen

- A *dry gas* is a natural gas from which all water vapor has been reduced.
- Hydrogen is dried by passing through *anhydrous* calcium chloride. The gas is collected by upward delivery (downward displacement of air) because it is less dense than air.

1 Obtaining hydrogen using Kipps apparatus

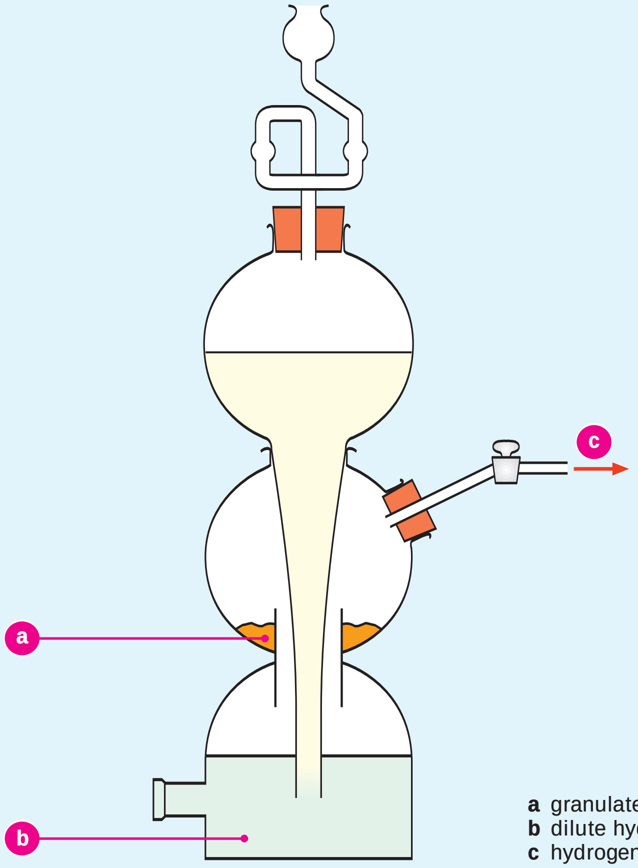

a granulated zinc
b dilute hydrochloric acid
c hydrogen gas

2 Collecting dry hydrogen

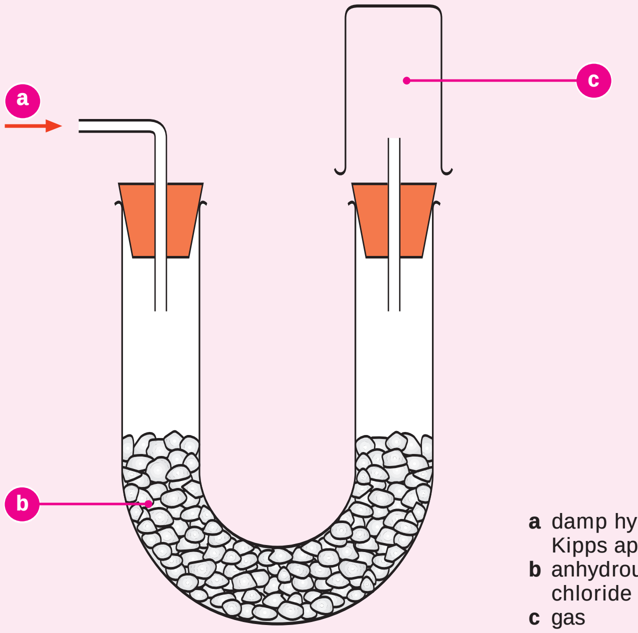

a damp hydrogen from Kipps apparatus
b anhydrous calcium chloride
c gas

© Diagram Visual Information Ltd.

Hydrogen: comparative density

1 Gaseous diffusion of hydrogen in air

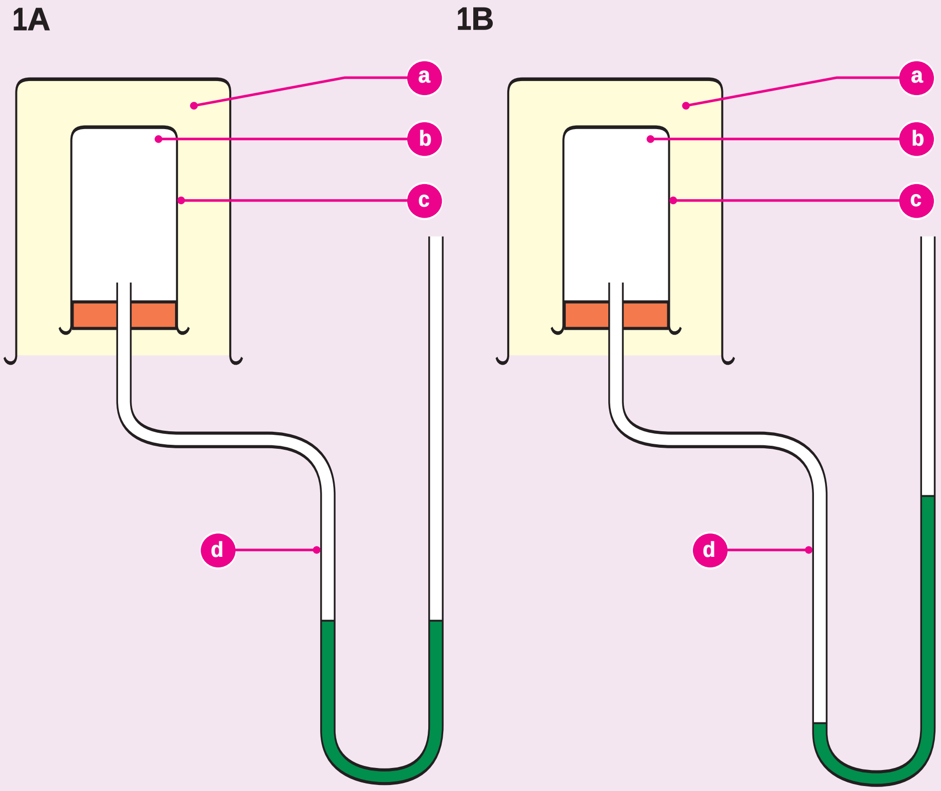

2 Gaseous diffusion of carbon dioxide in air

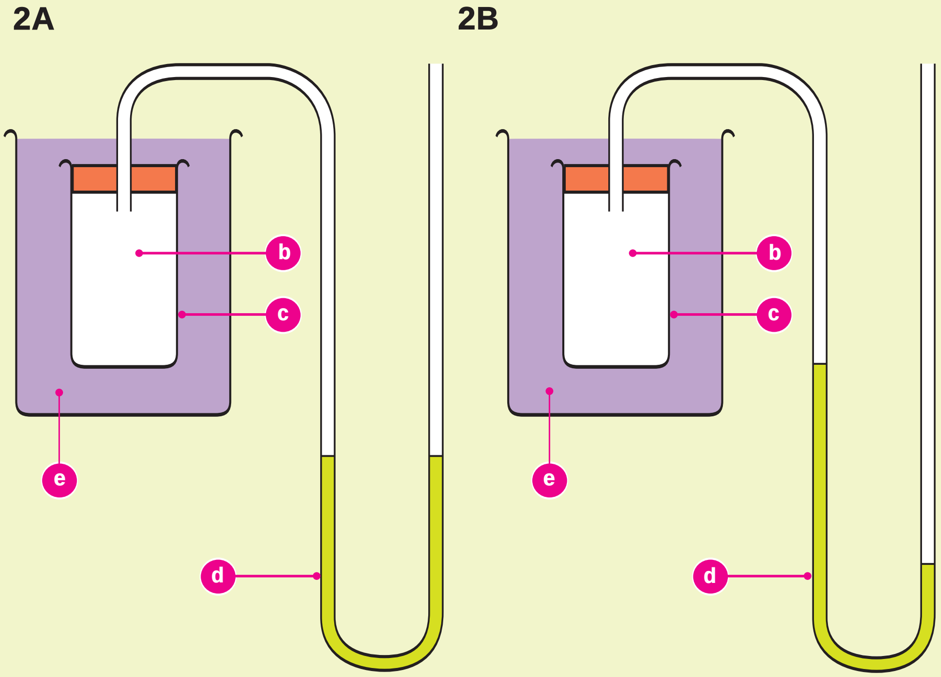

a hydrogen
b air
c porous vessel
d manometer
e carbon dioxide

Key words

carbon dioxide
diffusion
Graham's law
hydrogen

1 Gaseous diffusion of hydrogen

- Gases are able to pass into and out of a porous vessel.
- *Hydrogen* is less dense than air, so it is contained in an inverted beaker (1A).
- Hydrogen diffuses into the porous vessel more quickly than air diffuses out of it.
- The gas pressure inside the porous vessel increases and becomes greater than atmospheric pressure (1B). Liquid is forced up the right side of the manometer (an instrument used to measure the pressure of a fluid).

2 Gaseous diffusion of carbon dioxide

- *Carbon dioxide* is more dense than air, so it is contained in an upright beaker (2A).
- Carbon dioxide diffuses into the porous vessel more slowly than air diffuses out of it.
- The gas pressure inside the porous vessel decreases and becomes less than atmospheric pressure. Liquid is forced up the left side of the manometer (2B).

Comparative density

- *Graham's law* of *diffusion* states that the rate at which a gas diffuses (r) is proportional to the square root of 1 over its density (d):
  $r \propto \sqrt{1/d}$
- The density of hydrogen is lower that air. Therefore, it diffuses more quickly.
- The density of carbon dioxide is higher that air. Therefore it diffuses more slowly.

© Diagram Visual Information Ltd.

Hydrogen: reaction with other gases

Key words

anhydrous
anode
calcium carbonate
cathode
electrolysis
hydrogen sulfide
lead sulfide
photochemical reaction
sulfuric acid

1 Hydrogen and oxygen

- *Electrolysis* of dilute *sulfuric acid* produces oxygen at the *anode* (positive electrode) and hydrogen at the *cathode* (negative electrode).
- At the anode:
  $4OH^{-}(aq) \rightarrow O_2(g) + 2H_2O(l) + 4e^{-}$
- At the cathode:
  $4H^{+}(aq) + 4e^{-} \rightarrow 2H_2(g)$
- A mixture of hydrogen and oxygen explodes when ignited:
  $2H_2(g) + O_2(g) \rightarrow 2H_2O(g)$

2 Hydrogen and air

- Hydrogen burns in air to produce water.
- The hydrogen gas is dried by passing it through *anhydrous calcium carbonate* so any water produced must be the result of combustion.
- Water vapor condenses on the outer surface of the beaker of cold water and collects in the water glass.
- The liquid turns anhydrous blue cobalt chloride paper pink, showing it is water.

3 Hydrogen and sulfur

- Hydrogen reacts with sulfur to produce *hydrogen sulfide*:
  $H_2(g) + S(l) \rightarrow H_2S$
- Hydrogen sulfide turns damp lead acetate paper black due to the formation of *lead sulfide*:
  $Pb^{2+}(aq) + H_2S(g) \rightarrow PbS(s) + 2H^{+}(aq)$

4 hydrogen and chlorine

- A mixture of hydrogen and chlorine react in bright light.
- This is an example of a *photochemical reaction*. Light provides the energy needed for the reaction to start:
  $H_2(g) + Cl_2(g) \rightarrow 2HCl(g)$

© Diagram Visual Information Ltd.

1 Explosion of hydrogen/oxygen mixture

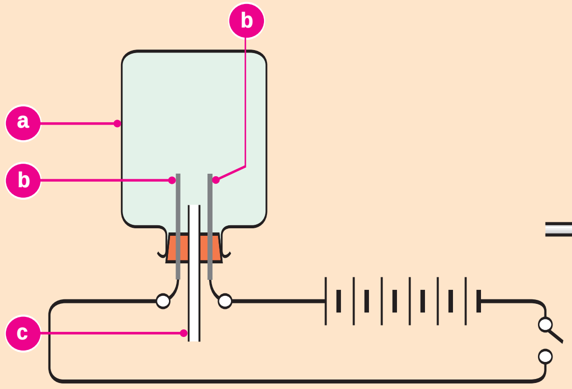

Preparation of hydrogen and oxygen from electrolysis of sulfuric acid

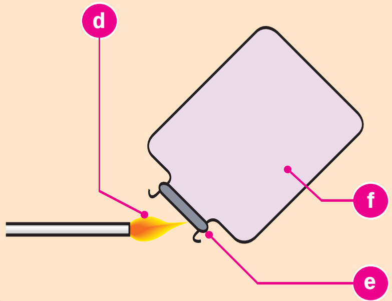

Explosion technique

2 Hydrogen and air

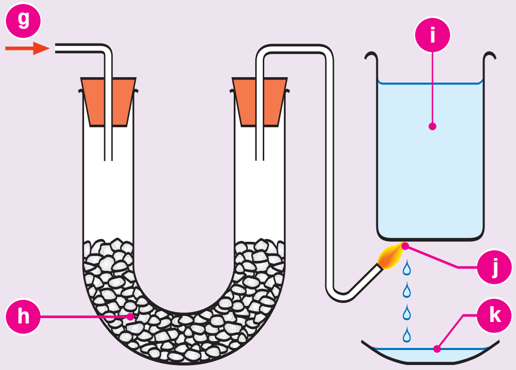

3 Reaction of hydrogen with sulfur

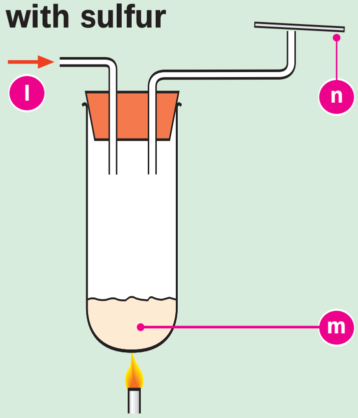

4 Photocatalytic explosion of hydrogen with chlorine

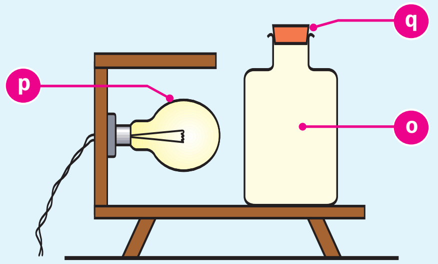

Before light is switched on

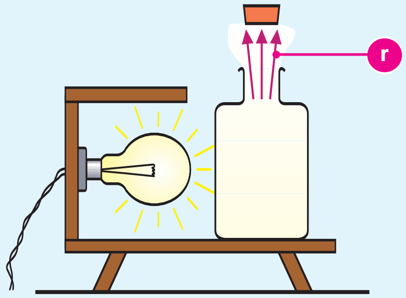

Light is switched on

a polyethylene bottle containing dilute H_2SO_4
b platinum electrodes
c polyethelene tubing
d bunsen flame
e loose support
f hydrogen/oxygen mixture
g hydrogen from Kipps apparatus
h anhydrous calcium chloride
i beaker containg cold water
j hydrogen burning
k water collected
l dry hydrogen
m boiling sulfur
n filter paper soaked in lead acetate
o polyethylene bottle containing mixture of chlorine and hydrogen
p powerful light source
q rubber bung
r white fumes of hydrogen chloride

Hydrogen: anomalies in ammonia and water

1 Anomalous boiling points of ammonia and water

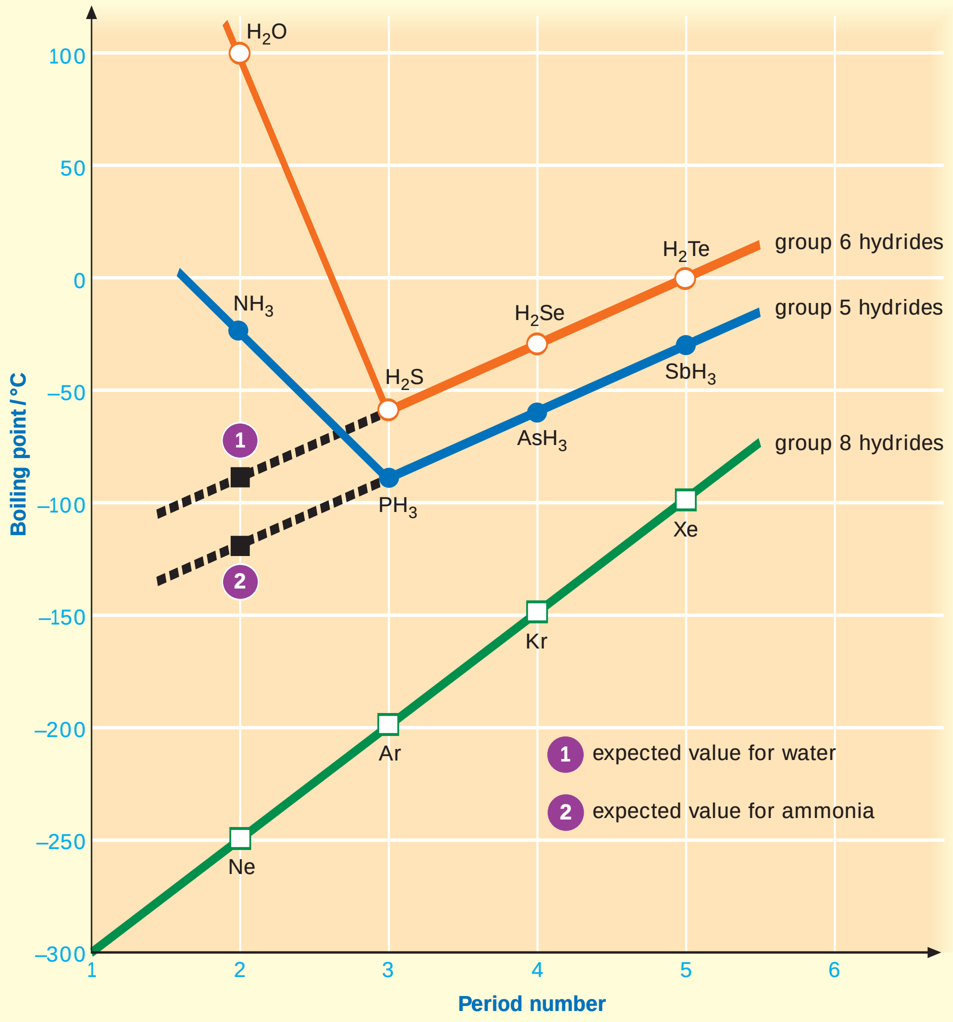

2 Strong bonding between water molecules

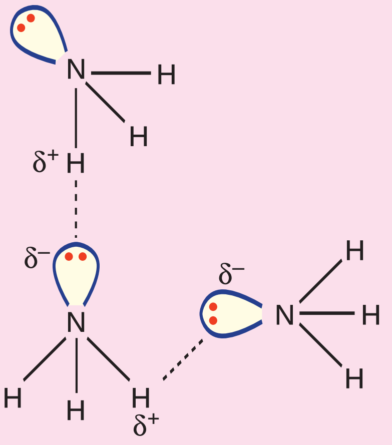

Key words

ammonia
boiling point
dipole
group 5
group 6
group 8
hydride
lone pair
orbital

1 Anomalous boiling points of ammonia and water

- Atoms of the *group 8* elements have a full outer *orbital* of electrons. They exist as single atoms, and there are very weak forces of attraction between them. The *boiling point* of these elements increases in proportion to the size of the atom.
- The boiling points of the *group 5* and *group 6 hydrides* also increases with molecular size. However, the boiling point of water and *ammonia* are significantly higher than might be expected when compared to the other hydrides in their groups. This is the result of strong forces of attraction, called hydrogen bonding, between water molecules and between ammonia molecules.

2 Strong bonding between water molecules

- There are five electrons in the outer orbital of a nitrogen atom. In ammonia, three of the electrons are used in nitrogen–hydrogen bonds, while the remaining two form a non-bonding or *lone pair*.
- The nitrogen atom in ammonia is more electronegative than the hydrogen atoms. The result is that a pair of bonding electrons lies nearer to the nitrogen atom than the hydrogen atom in each nitrogen–hydrogen bond.
- This forms a *dipole*, a chemical compound with an unequally distributed electric charge. The nitrogen–hydrogen bond is polarized, leaving the nitrogen slightly negative (sometimes shown as δ-) and the hydrogen slightly positive (sometimes shown as δ+).
- The slightly positive hydrogen atom is attracted to the slightly negative nitrogen atom and the non-bonding pair of electrons on neighboring ammonia molecules.

© Diagram Visual Information Ltd.

Basic reactions of hydrogen

Key words

alkali
base
catalyst
hydride
hydrogen
hydrocarbon
hydroxide
oxide
peroxide

1 Preparation

- Water, in the form of steam, can be reduced by either carbon or *hydrocarbons* to give *hydrogen*.
- This reaction is used to provide hydrogen for the Haber process in the industrial manufacture of ammonia (See page 74):
  $CH_4(g) + H_2O(g) \rightleftharpoons CO(g) + 3H_2(g)$

2 Reactions of hydrogen

- Hydrogen forms hydrides with both metals and non-metals.
- Hydrogen can be used to reduce the *oxides* of metals that are low in the reactivity series, such as copper(II) oxide.

3 The oxides

- Hydrogen reacts with oxygen to form both an oxide (water) and a *peroxide* (hydrogen peroxide).
- Hydrogen peroxide contains an unstable peroxide –O-O- bond.
- A variety of substances, including manganese(IV) oxide, act as *catalysts* to break this bond, forming water and oxygen.

4 The hydroxy compounds

- All metal *hydroxides* are *bases*.
- Metal hydroxides that dissolve in water are *alkalis*.
- Sodium hydroxide solution can be used to form precipitates of insoluble metal hydroxides like copper(II) hydroxide.

5 Acids

- All acids produce solutions containing hydrogen ions, H^+.

© Diagram Visual Information Ltd.

1 Preparation

A $C + H_2O \longrightarrow CO + H_2$

B $C_7H_{16} + 7H_2O \longrightarrow 7CO + 15H_2$

C $2H_2O + 2e^- \longrightarrow 2OH^- + H_2$

D $Zn + 2HCl \longrightarrow ZnCl_2 + H_2$

2 Reactions of hydrogen

A $2Na + H_2 \longrightarrow 2NaH$

$N_2 + 3H_2 \longrightarrow 2NH_3$

$O_2 + 2H_2 \longrightarrow 2H_2O$

$Cl_2 + H_2 \longrightarrow 2HCl$

B $CuO + H_2 \longrightarrow Cu + H_2O$

3 The oxides

A $2Na + 2H_2O \longrightarrow 2NaOH + H_2$

B $2H_2O_2 \xrightarrow{MnO_2} 2H_2O + O_2$

4 The hydroxy compounds

$NaOH + HCl \longrightarrow NaCl + H_2O$

$2NaOH + CuCl_2 \longrightarrow Cu(OH)_2\downarrow + 2NaCl$

5 Acids

$HCl \longrightarrow H^+ + Cl^-$

$HNO_3 \longrightarrow H^+ + NO_3^-$

$H_2SO_4 \longrightarrow 2H^+ + SO_4^{2-}$

1A from coal
1B from oil
1C electrolysis of brine
1D laboratory preparation
2A it reduces other elements
2B it reduces compound
3A water
3B hydrogen peroxide
4 metallic hydroxides are bases

The gases in air

1 Composition of air

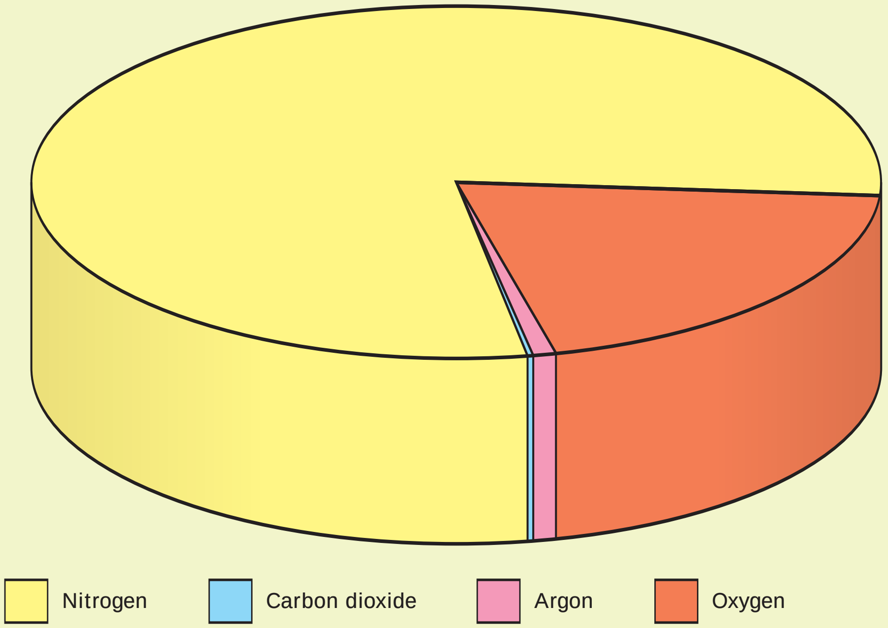

2 Gases present in clean dry air

Gases	Symbol or formula	Percentage of volume
Nitrogen	N	78.08
Oxygen	O	20.95
Argon	Ar	0.93
Carbon dioxide	CO_2	0.04
Neon	Ne	traces
Helium	He	
Krpton	Kr	
Xenon	Xe	

3 Gases that may be found in polluted air

Gases	Symbol or formula	Example of source
Ammonia	NH_3	Industrial processes
Carbon monoxide	CO	Motor vehicle exhaust
Methane	CH_4	Decay of organic material, passed in wind by herbivorous animals
Nitrogen oxides	NO_x	Motor vehicle and furnace exhaust
Ozone	O_3	Motor vehicle exhaust
Sulphur dioxide	SO_2	Coal-fired power stations

Key words

argon
carbon dioxide
carbon monoxide
compound
element
methane
ozone
pollutant
sulfur dioxide

1 Composition of air

- Air is not a chemical *compound* but a mixture of *elements* and compounds. The exact composition of air varies slightly from place to place depending on conditions. For example, the proportion of *carbon dioxide* in the air above a forest will be less than in the air above a city. Trees remove carbon dioxide for photosynthesis, while burning fossil fuels releases carbon dioxide.

2 Gases present in clean dry air

- Clean dry air is approximately 80 percent nitrogen and 20 percent oxygen.
- *Argon* makes up the largest portion of the remaining 10 percent.
- Air usually contains water vapor, with the amount depending on local conditions.

3 Gases in polluted air

- Air may contain *pollutants* released from a variety of sources.
- *Sulfur dioxide* and nitrogen oxides dissolve in water in the atmosphere and increase the acidity of rainwater.
- *Carbon monoxide* is a poisonous gas. If it is inhaled, it bonds onto the hemoglobin in red blood cells and prevents them from transporting oxygen.
- *Ozone* is essential in the upper atmosphere (ozone layer), where it prevents harmful ultraviolet radiation from reaching Earth. At ground level, however, it is responsible for the formation of smog.
- *Methane* and carbon dioxide are greenhouse gases. In the upper atmosphere, they prevent heat radiation from passing out into space, thus causing the global temperature to rise.

© Diagram Visual Information Ltd.

Nitrogen

Key words

group 5
nitrogen
solubility
triple bond

1 Nitrogen atom and molecule

- *Nitrogen* is in *group 5* of the periodic table. Nitrogen atoms have five electrons in the outer orbital and require an additional three electrons in order to fill the shell.
- Nitrogen exists as nitrogen molecules, N_2, in which each nitrogen atom provides three electrons. The nitrogen atoms are held together by a *triple bond*. A large amount of energy is needed to break the N≡N triple bond (945.4 kJ mol^{-1}), which accounts for the unreactive nature of nitrogen.

2 Physical properties of nitrogen

- The melting point and boiling point of nitrogen are both very low, and nitrogen is only a liquid over a small range of temperature. This indicates that the intermolecular forces in nitrogen are weak.
- The *solubility* of gases decreases with increasing temperature. The solubility of nitrogen falls from 2.3 cm^3 per 100 g water at 0°C to 1.0 cm^3 per 100 g water at 100°C.

3 Laboratory preparation of nitrogen

- Nitrogen is formed by the thermal decomposition of ammonium nitrite:
  $NH_4NO_2(s) \rightarrow 2H_2O(g) + N_2(g)$
- Ammonium nitrite is difficult to store because it decomposes over time. It is best made when it is needed by mixing ammonium chloride and sodium nitrite:
  $NH_4Cl(s) + NaNO_2(s) \rightarrow$
  $NH_4NO_2(s) + NaCl(s)$

© Diagram Visual Information Ltd.

1 Nitrogen atom and molecule

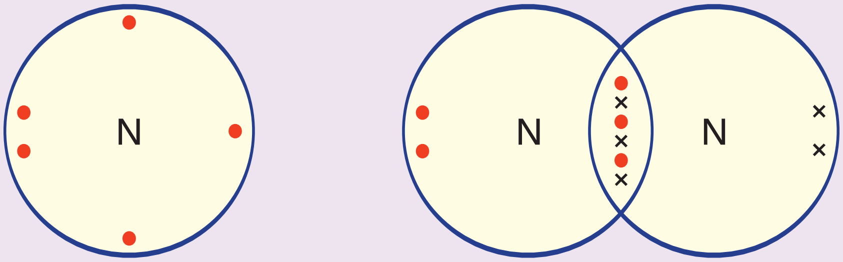

Nitrogen atom, $^{14}_{7}N$

Nitrogen molecule, N_2

2 Physical properties of nitrogen

Physical properties	m.p. / °C	b.p. / °C	density (relative to air)	color	smell	solubility at STP
Nitrogen	–120	–196	0.97	none	none	1.52 cm^3 in 100g of water

3 Laboratory preparation of nitrogen from ammonium chloride and sodium nitrate

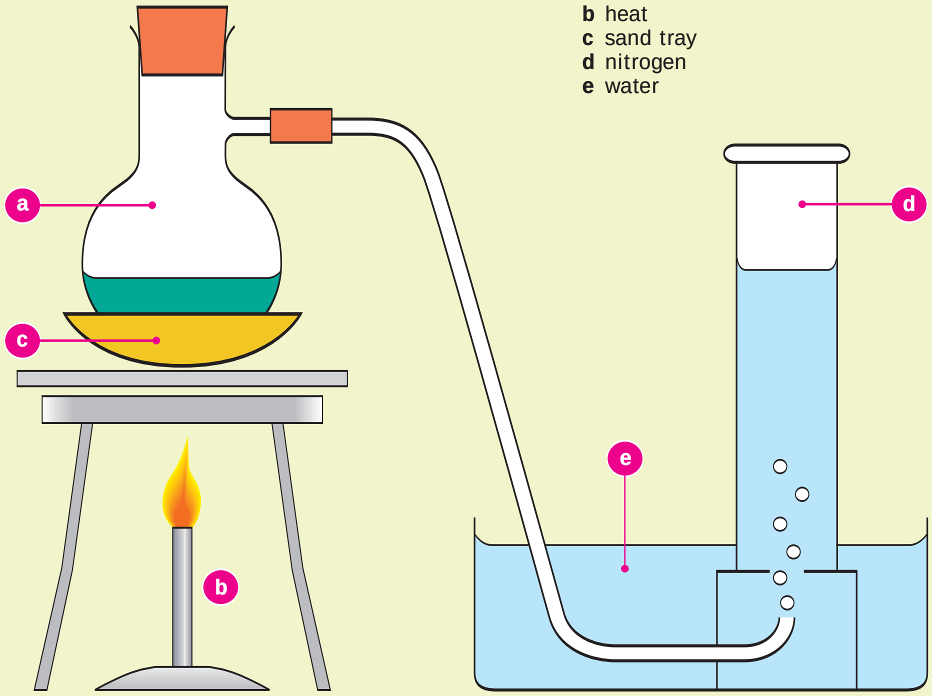

a ammonium chloride and sodium nitrate in solution
b heat
c sand tray
d nitrogen
e water

Other methods of preparing nitrogen

1 Extraction of nitrogen from the atmosphere

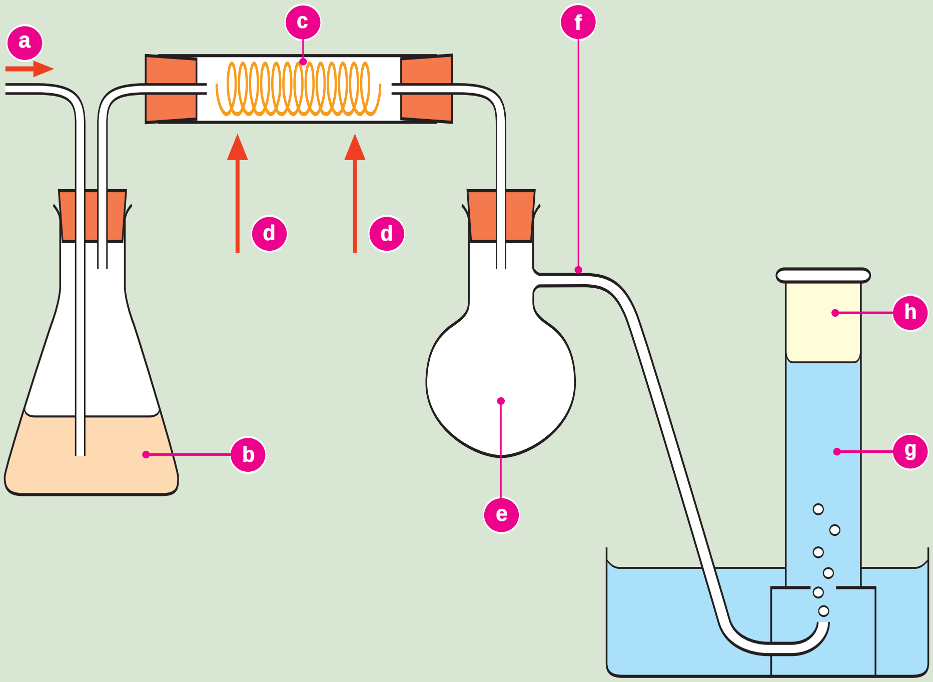

2 Preparation of nitrogen from ammonia

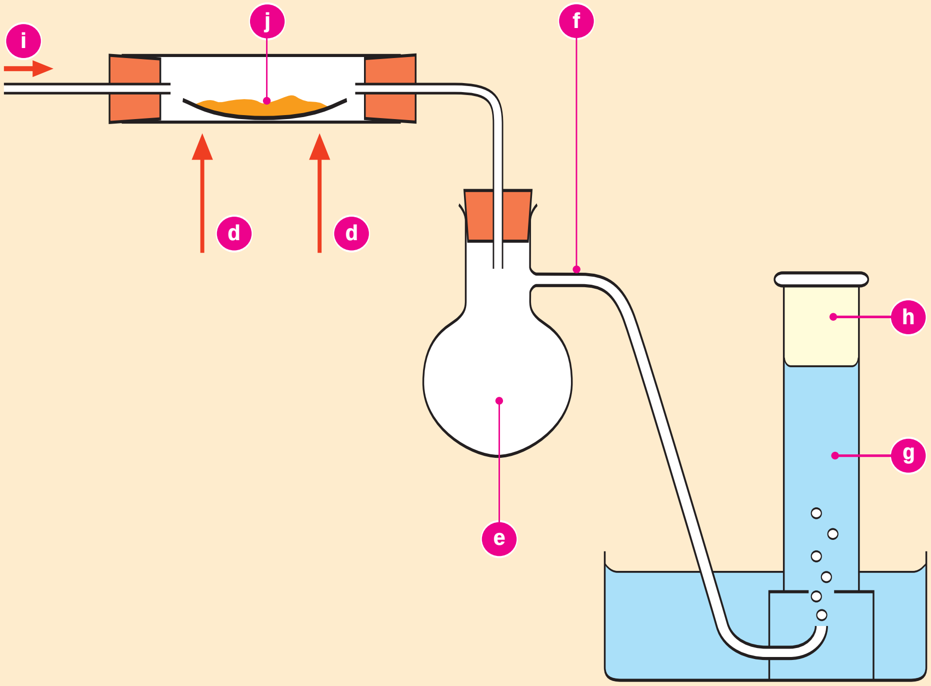

a air intake
b potassium hydroxide solution to absorb carbon
c hot copper coil to remove oxygen
d heat
e safety trap
f delivery tube
g water
h nitrogen collected over water
i dry ammonia gas
j dry copper(II) oxide

Key words

ammonia
group 8
inert
nitrogen
soluble

Other methods

- In addition to preparing *nitrogen* from ammonium chloride and sodium nitrate, it can be extracted from the air or from the reduction of copper(II) oxide by *ammonia*.

1 Extraction of nitrogen from the atmosphere

- Air is approximately 80 percent nitrogen and 20 percent oxygen. Impure nitrogen can be obtained by removing carbon dioxide and oxygen from air.
- Potassium hydroxide solution reacts with carbon dioxide:
  $KOH(aq) + CO_2(g) \rightarrow KHCO_3(aq)$
- A hot copper pile reacts with oxygen:
  $2Cu(s) + O_2(g) \rightarrow 2CuO(s)$
- Nitrogen prepared in this way contains argon and other *group 8* gases. However, because these are chemically *inert*, they would not interfere with any subsequent reactions.

2 Preparation of nitrogen from ammonia

- Ammonia can be used to reduce copper(II) oxide to copper. Nitrogen is one of the other products of this reaction:
  $3CuO(s) + 2NH_3(g) \rightarrow$
  $3Cu(s) + 3H_2O(g) + N_2(g)$
- Nitrogen is not very *soluble* and is readily collected over water.
- Water vapor condenses in the water trough. Ammonia is very soluble, and unreacted ammonia dissolves in the water.

© Diagram Visual Information Ltd.

Key words

ammonia
nitrate
nitrite
nitrogen
nitrogen cycle

Nitrogen cycle

- *Nitrogen* is continually being recycled between the soil and the air by natural processes.
- Lightning provides the energy for atmospheric nitrogen and oxygen to react, forming nitrogen oxides. For example:
  $N_2(g) + O_2(g) \rightarrow 2NO(g)$
  $N_2(g) + 2O_2(g) \rightarrow 2NO_2(g)$
- These oxides dissolve in water vapor in the atmosphere and eventually fall to the ground as rain.
- Atmospheric nitrogen is also converted into nitrogen compounds by nitrogen-fixing bacteria found in the root nodules of some plants.
- Nitrogen compounds are released into the soil as a result of the decay of animal waste products and by the decay of dead plants and animals. In the soil, the *ammonia* formed during decay processes is converted into *nitrates* by the action of nitrifying bacteria:
  Ammonia → Ammonium compounds → *Nitrites* → Nitrates
- Nitrogen compounds may also be added to soil as artificial fertilizers. Ammonium nitrate is widely used by farmers and gardeners to provide growing plants with essential nitrogen.
- Nitrogen compounds are taken out of the soil by plants, which use them to make proteins and other essential chemicals. Animals subsequently eat the plants.
- Nitrogen compounds in the soil are converted to nitrogen gas by denitrifying bacteria.

© Diagram Visual Information Ltd.

The nitrogen cycle

Nitrogen cycle schematic

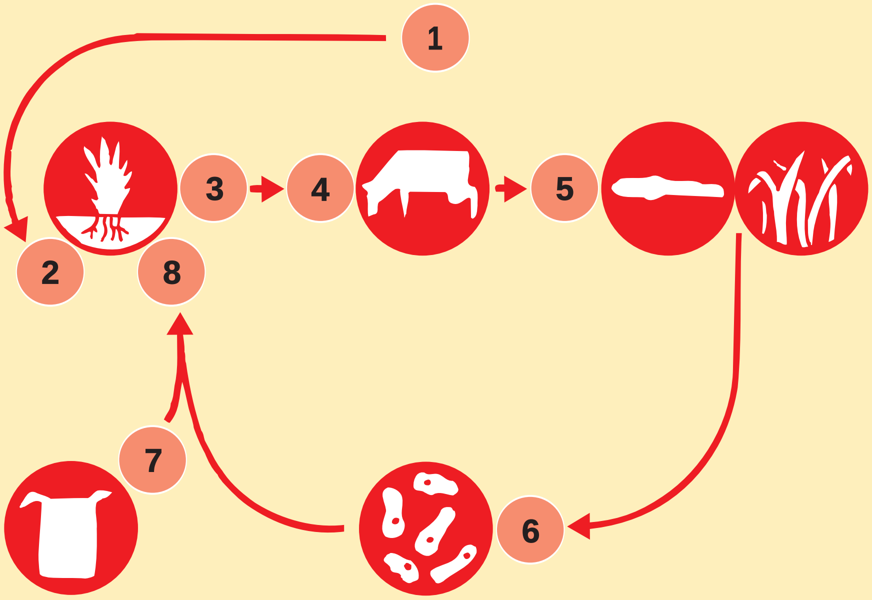

All plants and animals need nitrogen, present in proteins and nucleic acids. Most living things, however, cannot use nitrogen directly from the atmosphere.

1 Nitrogen in the air.
2 Nitrogen in the atmosphere is trapped by some plant roots.
3 Plants use nitrogen for making proteins.
4 Animals eat plant proteins.
5 The proteins in dead organisms and in body wastes are converted to ammonia by bacteria and fungi.
6 Other bacteria convert the ammonia to nitrates.
7 Artificial nitrates are added to the soil as fertilizer.
8 Plants absorb the nitrates.

Nitrogen cycle

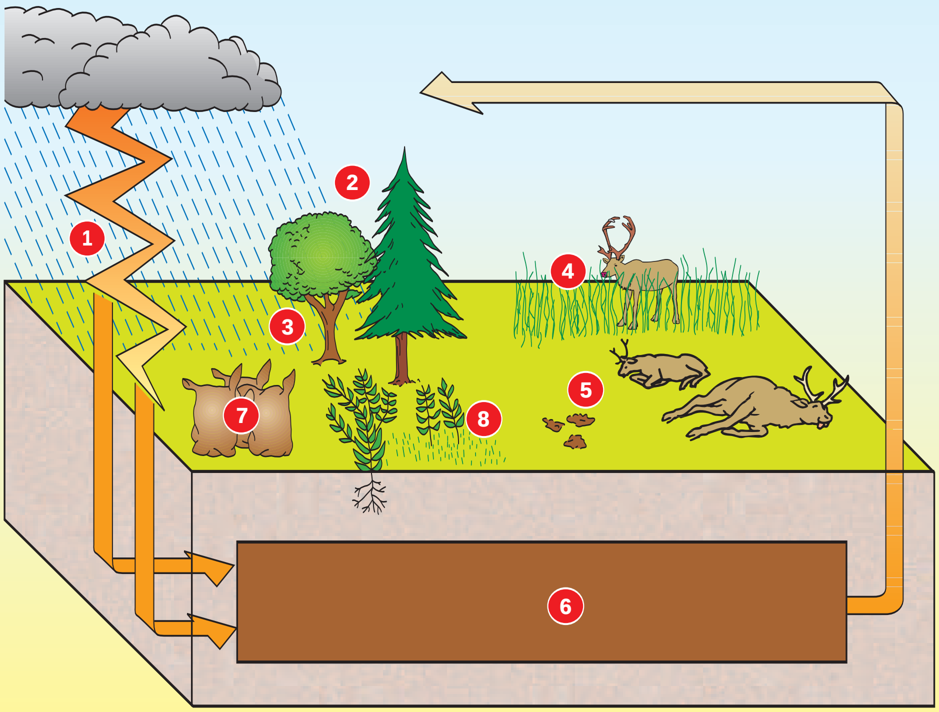

Preparation and properties of ammonia

1 Laboratory preparation of ammonia

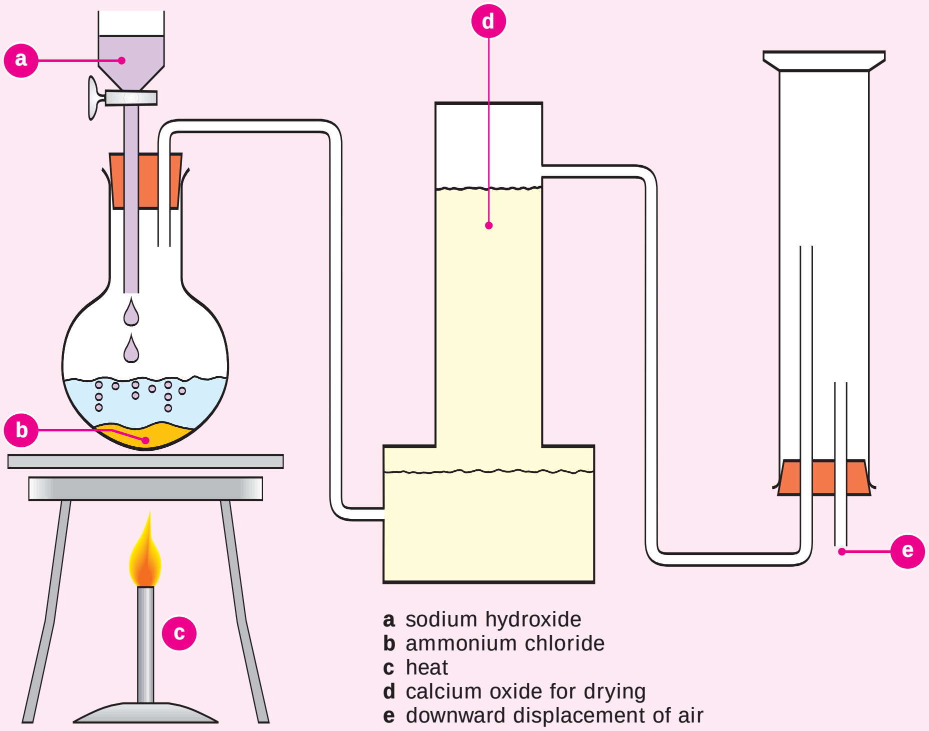

a sodium hydroxide
b ammonium chloride
c heat
d calcium oxide for drying
e downward displacement of air

2 Bonding angle

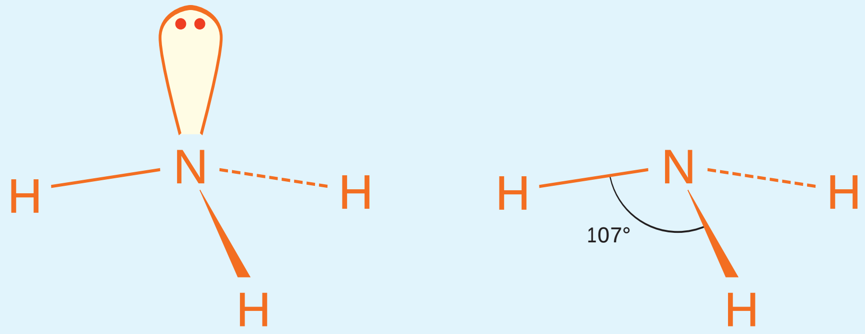

3 Physical properties of ammonia

Physical property	Ammonia
Ammonia	–77.7
Carbon monoxide	–33.5
Methane	0.59
Nitrogen oxides	colorless
Ozone	characteristic unpleasant odour
Sulphur dioxide	68 000 cm^3 in 100 g of water

Key words

alkali
ammonia
ammonium hydroxide
bond angle
lone pair
shell

1 Laboratory preparation of ammonia

- *Ammonia* is formed by the reaction of an ammonium compound, such as ammonium chloride, with an *alkali*, such as sodium hydroxide:
  $NH_4Cl(s) + NaOH(aq) \rightarrow$
  $NH_3(g) + NaCl(aq) + H_2O(l)$
- Ammonia gas is dried by passing it through a column of calcium oxide.

2 Bonding angle

- The five electrons in the outer electron *shell* of nitrogen plus three shared electrons from the hydrogen atoms form three pairs of bonding electrons and one pair of non-bonding electrons (*lone pair*).
- The four pairs of electrons are directed toward the corners of a tetrahedron. However, repulsion between the non-bonding pair of electrons and the bonding pairs of electrons forces the nitrogen–hydrogen bonds slightly closer to each other, resulting in a *bond angle* of 107°. By comparison, the bond angle in methane is 109.5°.

3 Physical properties of ammonia

- Ammonia is less dense than air and is collected by upward delivery (downward displacement of air).
- Ammonia is exceptionally soluble and cannot be collected over water. 680 cm^3 of ammonia will dissolve in 1 g of water at 20°C.
- Ammonia dissolves in water to form a solution that is a weak alkali:
  $NH_3(g) + H_2O(l) \rightleftharpoons$
  $NH_4^+(aq) + OH^-(aq)$
- Ammonia solution is sometimes referred to as *ammonium hydroxide*.

© Diagram Visual Information Ltd.

Industrial preparation of ammonia (the Haber process): theory

Key words

ammonia
equilibrium
Le Chatelier's principle
product
reactant

1 Reversible reaction

- *Ammonia* is made industrially by the reaction of nitrogen and hydrogen.
- The reaction is reversible.
  Forward reaction:
  $N_2(g) + 3H_2(g) \rightarrow 2NH_3(g)$
  Backward reaction:
  $N_2(g) + 3H_2(g) \leftarrow 2NH_3(g)$
- The concentration of ammonia in the equilibrium mixture depends on the pressure and temperature at which the reaction is carried out.
- According to *Le Chatelier's principle*, if any change is made to the external conditions (such as temperature, concentration and pressure) of a system at *equilibrium*, the equilibrium position will alter so as to oppose the change.

2 Variation of percent ammonia with pressure

- In the forward reaction, four moles of *reactants* are converted into two moles of *product* so there is a drop in pressure.
- According to Le Chatelier's principle, an increase in pressure will favor the forward reaction. The equilibrium position will move to the right in order to oppose the increase in pressure, so the equilibrium mixture will contain more ammonia.

3 Variation of percent ammonia with temperature

- The reaction is exothermic; heat is given out.
- According to Le Chatelier's principle, a decrease in temperature will favor the forward reaction. The equilibrium position will move to the right in order to oppose the decrease in temperature, so the equilibrium mixture will contain more ammonia.

© Diagram Visual Information Ltd.

1 Reversible reaction

$$N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g) \qquad \Delta H = -92\ \text{kJ mol}^{-1}$$

4 moles — 2 moles

2 Variation of percent ammonia with pressure

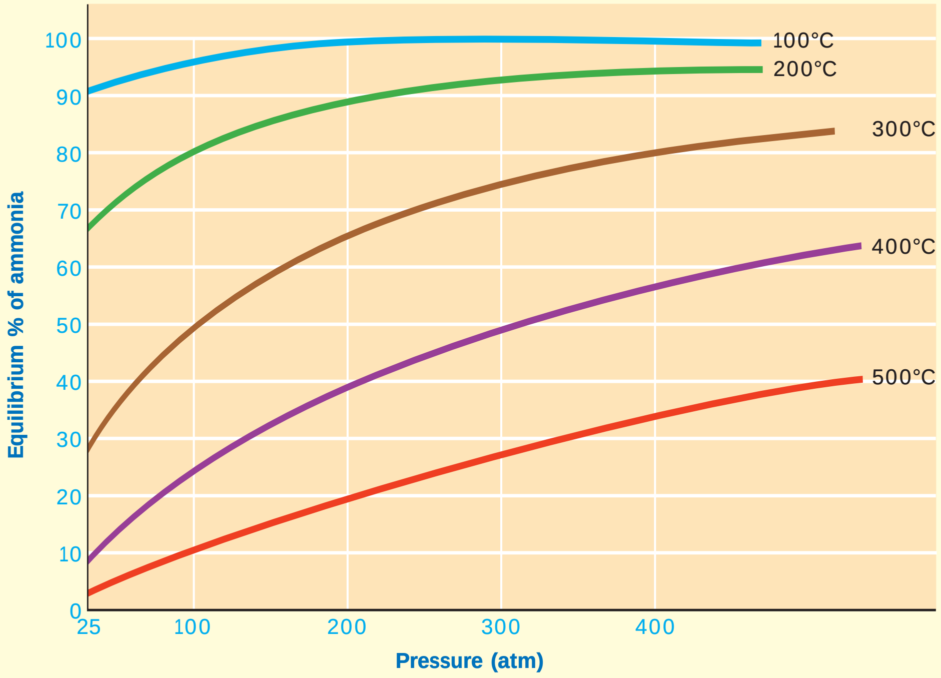

3 Variation of percent ammonia with temperature

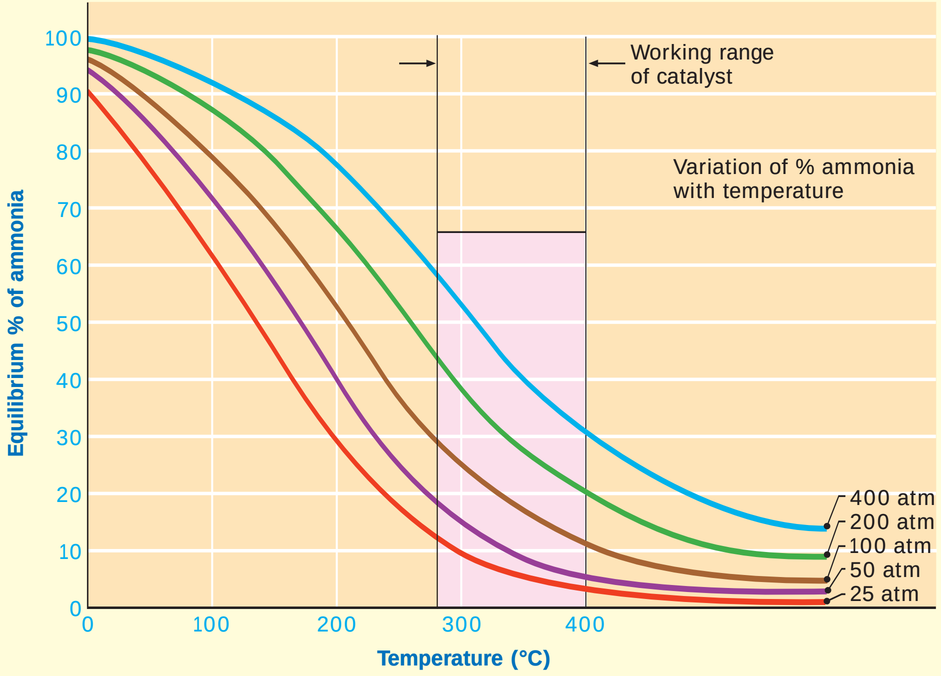

Industrial preparation of ammonia (the Haber process): schematic

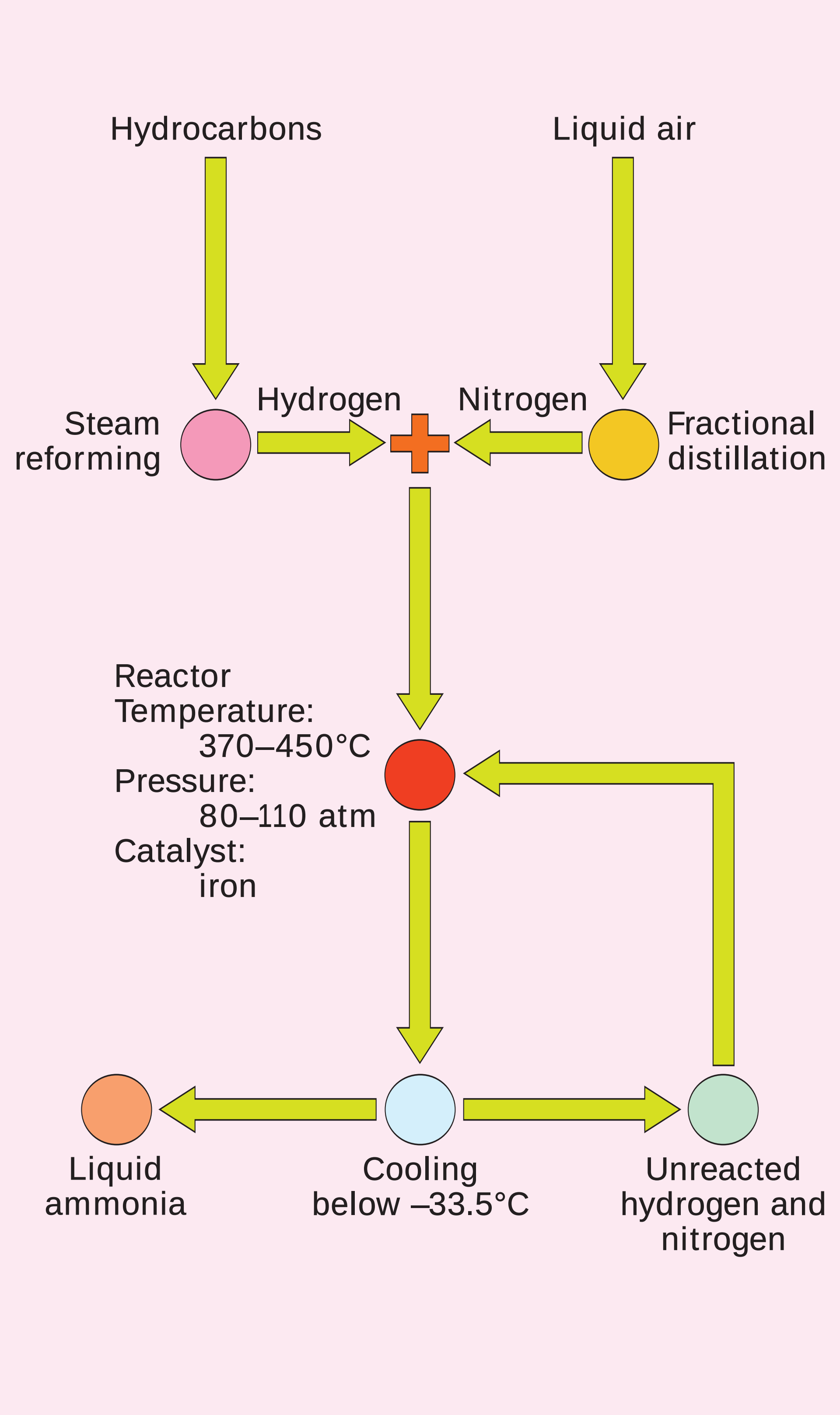
Hydrocarbons
Liquid air
Steam reforming
Hydrogen
Nitrogen
Fractional distillation
Reactor
Temperature: 370–450°C
Pressure: 80–110 atm
Catalyst: iron
Liquid ammonia
Cooling below –33.5°C
Unreacted hydrogen and nitrogen

Key words

ammonia	*hydrocarbon*
catalyst	*hydrogen*
equilibrium	*nitrogen*
fractional distillation	

The process

- The raw materials for the Haber process are *hydrogen* and *nitrogen*.
- Hydrogen is obtained by the steam reforming of *hydrocarbons*, such as methane, or the reaction of steam with coke.

  $CH_4(g) + H_2O(g) \rightleftharpoons CO(g) + 3H_2(g)$

  $C(s) + H_2O(g) \rightleftharpoons CO(g) + H_2(g)$
- Nitrogen is obtained from the *fractional distillation* of liquid air.
- The formation of *ammonia* is favored by high pressure, and the reaction is normally carried out at 80–110 atm. It is also favored by low temperature, but this also lowers the rate of reaction, so a *catalyst* is used. The reaction is normally carried out at 370–450°C in the presence of a finely divided iron catalyst.
- In reality, the reaction is not allowed to reach *equilibrium*. A single pass through the reactor results in about 15 percent conversion to ammonia.
- The reaction mixture is cooled to below the boiling point of ammonia, at which point liquid ammonia condenses out and is removed. The mixture of unreacted hydrogen and nitrogen is recycled back into the reactor.
- Around 80 percent of the ammonia produced each year is used to make fertilizers, including ammonia solution, ammonium nitrate, ammonium sulfate, and urea.

© Diagram Visual Information Ltd.

Industrial preparation of nitric acid

Key words

ammonia
azeotropic mixture
exothermic
nitric acid
oxidation

1 Preparation

- The industrial production of *nitric acid* involves two stages: the *oxidation* of *ammonia* and the absorption of the resulting nitrogen oxides.
- In the converter, a mixture of ammonia and air is passed through a platinum–rhodium gauze and the ammonia is oxidized:
  $4NH_3(g) + 5O_2(g) \rightleftharpoons 4NO(g) + 6H_2O(g)$
  $\Delta H = -909$ kj mol^{-1}
- The reaction is *exothermic*, and a large quantity of heat is produced.
- Conditions are carefully controlled to minimize a competing reaction in which ammonia is oxidized to nitrogen:
  $4NH_3(g) + 3O_2(g) \rightleftharpoons 2N_2(g) + 6H_2O(g)$
- Air is added to the nitrogen oxides in order to make nitrogen dioxide and, subsequently, dinitrogen tetroxide.
  $2NO(g) + O_2(g) \rightleftharpoons 2NO_2(g)$
  $2NO_2(g) \rightleftharpoons N_2O_4(g)$
- Dinitrogen tetroxide reacts with water to produce nitric acid.
  $3N_2O_4(g) + 2H_2O(l) \rightleftharpoons$
  $4HNO_3(aq) + 2NO(g)$
- The acid from the absorption towers typically contains 55–60 percent nitric acid by mass. Most of the modern demand is for acid of this concentration.
- Nitric acid and water form an *azeotropic mixture* (a mixture that boils without a change in composition) containing 68.5 percent nitric acid by mass. Thus, concentrated nitric cannot be obtained by distillation. Concentrated sulfuric acid is used to reduce the water content and give concentrated nitric acid.

2 Uses of nitric acid

- Over two thirds of nitric acid production is directed to making ammonium nitrate, which is used as a fertilizer and in explosives.

© Diagram Visual Information Ltd.

1 Industrial manufacture of nitric acid

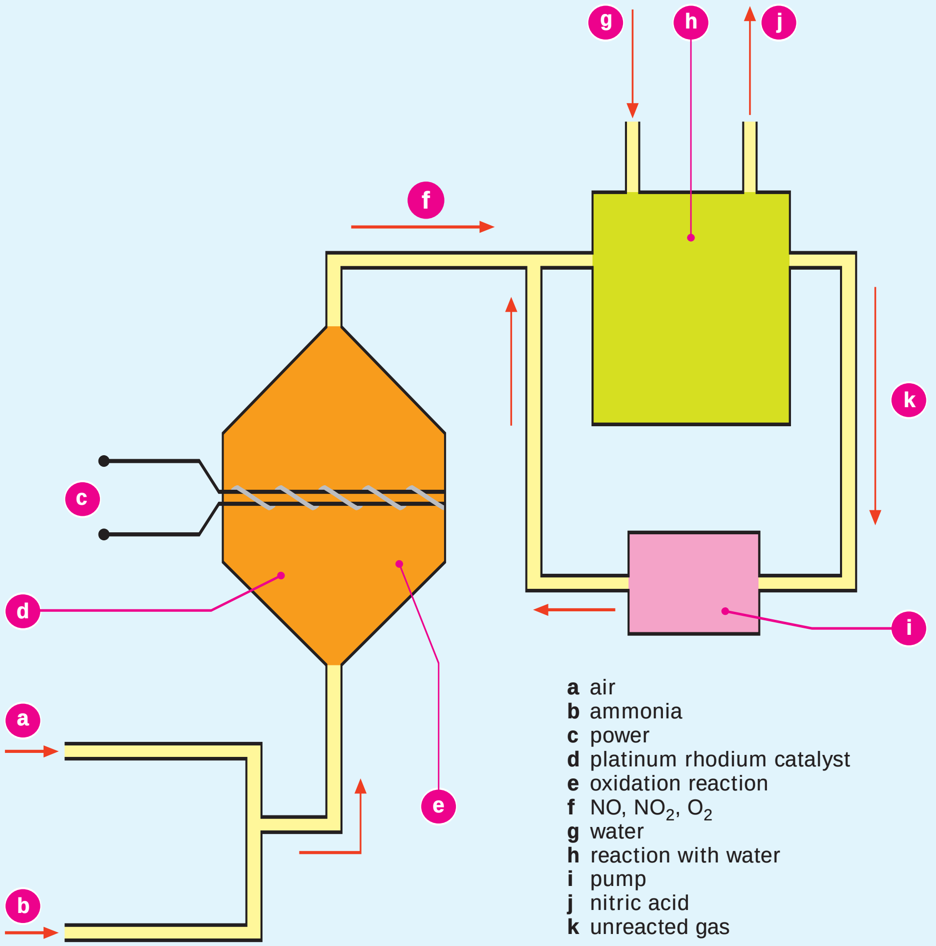

a air
b ammonia
c power
d platinum rhodium catalyst
e oxidation reaction
f NO, NO_2, O_2
g water
h reaction with water
i pump
j nitric acid
k unreacted gas

2 Summary of uses

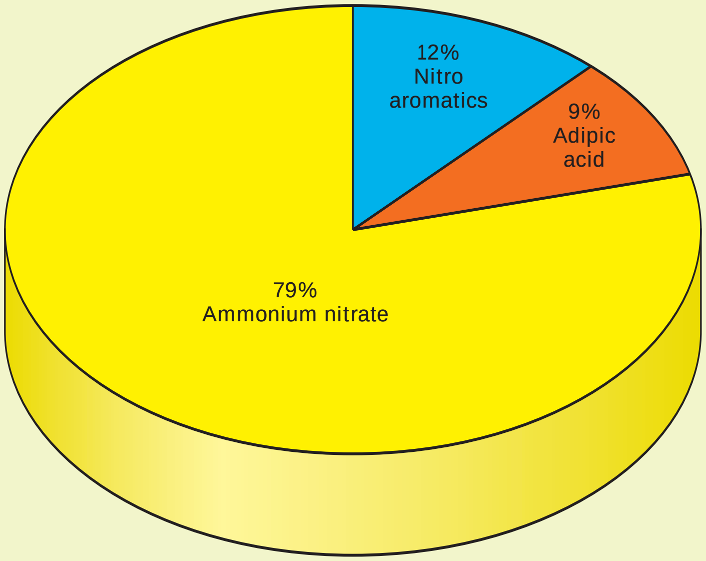

Nitrogen: reactions in ammonia and nitric acid

1 Redox chemistry (ammonia)

$$2NH_3 + 3Cl_2 \longrightarrow N_2 + 6HCl$$

Ammonia reduces chlorine to hydrogen chloride and nitrogen

$$2NH_3 + 3CuO \longrightarrow 3Cu + 3H_2O + N_2$$

Ammonia reduces copper oxide to copper, water, and nitrogen

2 Redox chemistry (nitric acid)

$$C + 4HNO_3 \longrightarrow CO_2 + 4NO_2 + 2H_2O$$

Concentrated nitric acid oxidizes carbon to carbon dioxide

$$3Cu + 8HNO_3(\text{dilute}) \longrightarrow 3Cu(NO_3)_2 + 4H_2O + 2NO\uparrow$$

Dilute nitric acid oxidizes copper to produce copper nitrate, water, and nitrogen oxide

$$Cu + 4HNO_3(\text{conc}) \longrightarrow Cu(NO_3)_2 + 2H_2O + 2NO_2\uparrow$$

Concentrayed nitric acid oxidizes copper to produce copper nitrate, water, and nitrogen dioxide

3 Complex ammonia salts

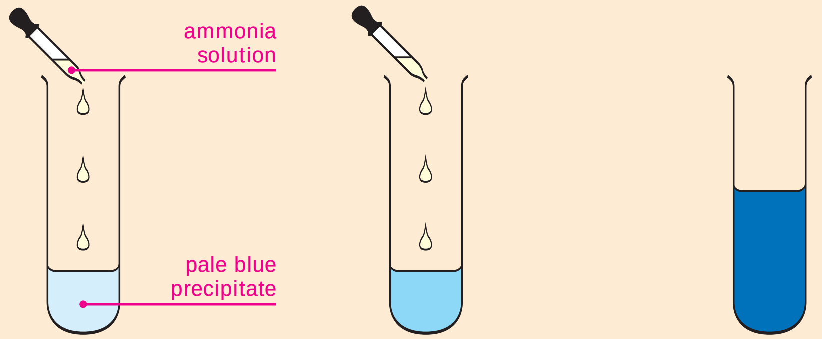

Ammonia solutionis added to $CuSo_4$ solution

More ammonia solution is added

The precipitate diasppears, leaving a deep royal blue solution (a complex salt)

4 Tetraamminecopper(II) ion

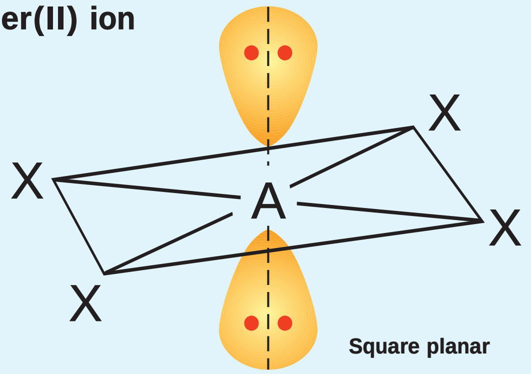

Square planar

Key words

ammonia
nitric acid
oxide
oxidizing agent
reducing agent

1 Redox chemistry (ammonia)

- *Ammonia* is a *reducing agent* and will reduce chlorine and heated metal *oxides* such as copper oxide. During the reactions, the ammonia is oxidized to nitrogen.

2 Redox chemistry (nitric acid)

- In addition to its properties as an acid, *nitric acid* is also a powerful *oxidizing agent*. It is able to oxidize both non-metals and metals.

3 Complex ammonia salts

- When ammonia solution is added drop by drop to copper(II) sulfate solution, a pale blue precipitate of copper(II) hydroxide is formed:
  $Cu^{2+}(aq) + 2OH^-(aq) \rightarrow Cu(OH)_2(s)$
- When an excess of ammonia solution is added, the pale blue precipitate redissolves, forming a deep blue solution:
  $Cu(OH)_2(s) + 4NH_3(aq) \rightarrow [Cu(NH_3)_4]^{2+}$
- The deep blue solution contains the complex ion tetraamminecopper(II), $[Cu(NH_3)_4]^{2+}$.
- Ammonia also forms complex ions with other metals, such as diamminesilver(I), $[Ag(NH_3)_2]^+$ and hexaamminenickel(II), $[Ni(NH_3)_6]^{2+}$.

4 Tetraamminecopper(II)

- In the tetraamminecopper(II) ion, a copper ion is surrounded by four ammonia molecules in a square planar arrangement. The non-bonding pair of electrons on each nitrogen atom is attracted to the central positively charged copper ion.

© Diagram Visual Information Ltd.

Basic reactions of nitrogen

Key words

acid
ammonia
nitrate
nitric acid
nitrite
nitrogen
oxide
oxidizing agent
salt

1 With metals and non-metals

- *Nitrogen* combines directly with both metals, such as magnesium, and non-metals, such as sodium and hydrogen.

2 Basic reactions of ammonia

- *Ammonia* reacts with water to produce a soluble alkaline gas (2A).
- Ammonia reacts with an *acid* to produce a *salt* (2B).
- Ammonia reacts with an *oxide* to produce a metal and nitrogen (2C).
- Ammonia reacts with oxygen to produce *nitric acid* (2D).

3 Nitric acid

- Nitric acid is both a strong acid (3A) and a powerful *oxidizing agent* (3B).
- Cold, dilute nitric acid produces nitrogen oxide when reacting with a metal (3C).
- Hot, concentrated nitric acid produces nitrogen dioxide when reacting with a metal (3D).

4 Nitrates

- All *nitrates* are very soluble in water.
- Group 1 metal nitrates (apart form lithium nitrate) decompose on heating to form metal *nitrites* and oxygen.
- Other metal nitrates decompose on heating to form metal oxides, nitrogen dioxide, and oxygen.
- Ammonium nitrate decomposes on heating, forming water and dinitrogen oxide.

1 With metals and non-metals

$$3Mg + N_2 \longrightarrow Mg_3N_2$$

$$6Na + N_2 \longrightarrow 2Na_3N$$

$$N_2 + 3H_2 \overset{\text{Fe catalyst}}{\rightleftharpoons} 2NH_3$$

2 Basic reactions of ammonia

A $NH_3 + H_2O \rightleftharpoons NH_4^+ + OH_2^-$

$NH_3 + HCl \rightleftharpoons NH_4Cl$

B $2NH_3 + 3CuO \longrightarrow 3Cu + 3H_2O + N_2$

C $4NH_3 + 5O_2 \longrightarrow 4NO + 6H_2O$

then $2NO + O_2 \longrightarrow 2NO_2$

and $4NO_2 + O_2 + 2H_2O \longrightarrow 4HNO_3$

D $Cu(OH)_2 + 4NH_3 \rightleftharpoons Cu(NH_3)_4^{2+} + 2OH^-$

3 Nitric acid

A $HNO_3 + H_2O \longrightarrow H_3O^+ + NO_3^-$

B $C + 2HNO_3 \longrightarrow CO_2 + 2NO_2 + H_2O$

$Cu + 4HNO_3 \longrightarrow Cu(NO_3)_2 + 2H_2O + 2NO_2$

4 Nitrates

A $2NaNO_3 \longrightarrow 2NaNO_2 + O_2$

B $2Pb(NO_3)_2 \longrightarrow 2PbO + 4NO_2 + O_2$

C $NH_4NO_3 \longrightarrow 2H_2O + N_2O$

2A soluble alkaline gas
2B reducing agent
2C ammonia with oxygen to make nitric acid
2D complexing agent
3A strong acid
3B oxidizing agent

4A group 1 (excluding $LiNO_3$)
4B others
4C ammonium nitrate

© Diagram Visual Information Ltd.

Nitrate fertilizers

1 Atmospheric nitrogen

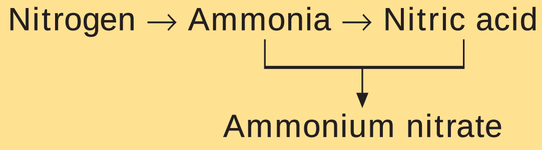

$$NH_3(aq) + HNO_3(aq) \rightarrow NH_4NO_3(aq)$$

2 Ammonia nitrate

Formula mass of ammonium nitrate = $14 + 4 + 14 + 3 \times 16 = 80$

Percentage of nitrogen in ammonium nitrate = $\frac{2 \times 14 \times 100}{80} = 35\%$

3 Nitrogen in plants

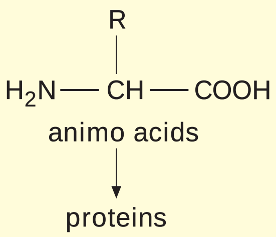

$$H_2N - \overset{\overset{\displaystyle R}{|}}{C}H - COOH$$

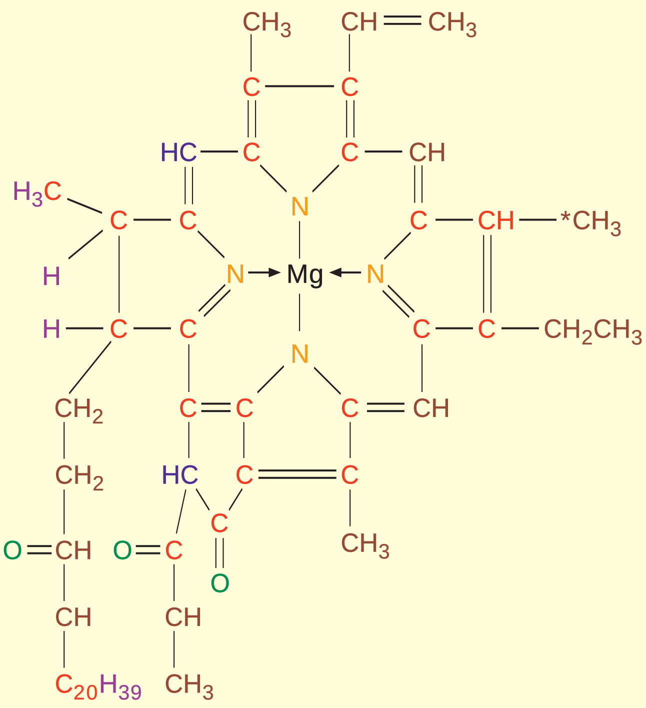

Chlorophyll a
(In chlorophyll b, the methyl group marked by an asterisk is replaced by a —CHO group)

4 Table of nitrogen fertilizers

Compound	Formula	Percentage of nitrogen
Ammonium nitrate	NH_4NO_3	35.00
Ammonium sulphate	$(NH_4)_2SO_4$	21.21
Urea	H_2NCONH_2	46.67

Key words

alkali
amino acid
ammonia
chlorophyll
neutralization
nitric acid
protein

1 Atmospheric nitrogen

- Atmospheric nitrogen is an important raw material in the manufacture of *ammonia* and *nitric acid*.
- Ammonia solution is *alkali*, while nitric acid is acidic.
- The two solutions undergo a *neutralization* reaction to form the salt ammonium nitrate.

2 Ammonia nitrate

- The percentage of nitrogen in a nitrogenous fertilizer is an important factor in determining how much fertilizer should be used on an area of crops.
- Ammonium nitrate is very soluble in water. Any excess that is applied to soil is readily washed out into streams and rivers, where it causes environmental problems.

3 Nitrogen in plants

- Plants use nitrogen to make *amino acids* and, from these, to make *proteins*.
- Plants also use nitrogen to make other important chemicals, such as *chlorophyll*.

4 Nitrogen fertilizers

- Ammonium sulfate is formed by the neutralization reaction between ammonia solution and sulfuric acid:
  $2NH_3(aq) + H_2SO_4(aq) \rightarrow (NH_4)_2SO_4(aq)$
- Urea is a waste product of animal metabolism and is excreted from the body in sweat and urine. It is made industrially by the reaction of ammonia with carbon dioxide. This reaction is carried out at 200 °C and 200 atmospheres:
  $CO_2(g) + 2NH_3(g) \rightarrow H_2NCONH_2(l) + H_2O(g)$

© Diagram Visual Information Ltd.

Oxygen and sulfur

Key words

catalyst
double bond
enzyme
group 6
oxygen
sulfur

1 Atoms and molecules

- *Oxygen* and *sulfur* are both in *group 6* of the periodic table. Atoms of each element have six electrons in the outer electron shell and require two electrons to fill the shell.
- Oxygen exists as oxygen molecules, O_2, in which each oxygen atom provides two electrons. The oxygen atoms are held together by a *double bond*, O=O.
- At room temperature, sulfur exists as a molecule composed of eight sulfur atoms, S_8, arranged in the shape of a crown.

2 Laboratory preparation of oxygen

- Oxygen is prepared in the laboratory by the decomposition of hydrogen peroxide using a suitable *catalyst*, such as manganese dioxide:

$2H_2O_2(l) \rightarrow O_2(g) + 2H_2O(l)$

- Hydrogen peroxide is rapidly decomposed by a variety of catalysts, including the *enzyme* catalase. In one second, one molecule of catalase can decompose up to 50,000 molecules of hydrogen peroxide.
- Oxygen is only slightly soluble and can be collected over water.
- Hydrogen peroxide is usually supplied in solutions designated by volume strength. For example, 20-volume hydrogen peroxide yields 20 volumes of oxygen gas per volume of solution.

3 Physical properties of oxygen and sulfur

- At room temperature, oxygen is a colorless, odorless gas, while sulfur is a yellow solid.

© Diagram Visual Information Ltd.

1 Atoms and molecules

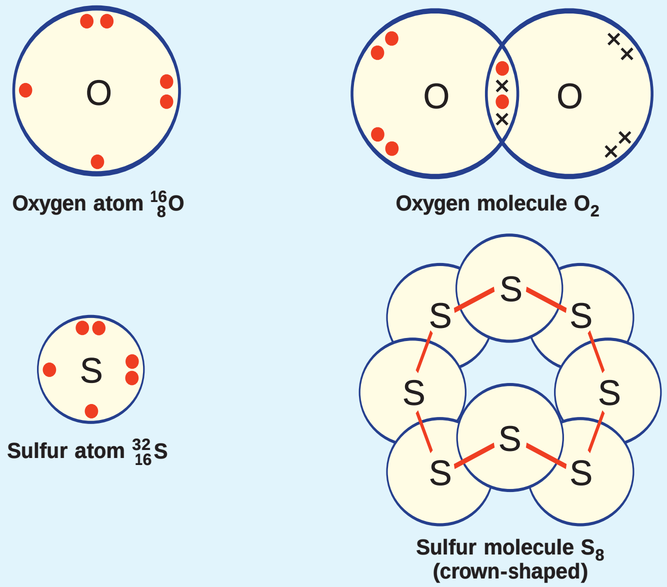

Oxygen atom $^{16}_{8}O$

Oxygen molecule O_2

Sulfur atom $^{32}_{16}S$

Sulfur molecule S_8 (crown-shaped)

2 Laboratory preparation of oxygen

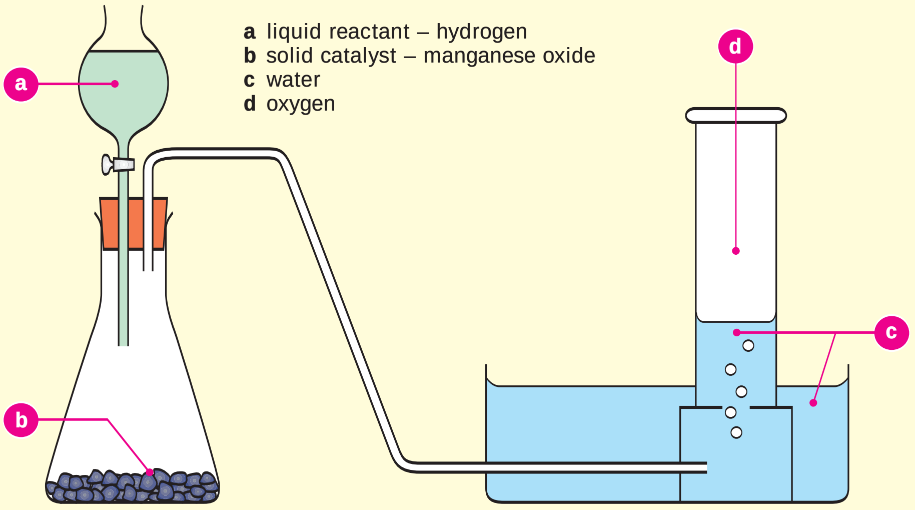

a liquid reactant – hydrogen
b solid catalyst – manganese oxide
c water
d oxygen

3 Physical properties of oxygen and sulfur

Physical properties	Oxygen	Sulfur
m.p./°C	–218	119
b.p./°C	–183	444
Density g/dm^3	1.31	2070
Color	none	yellow
Smell	none	slight
Solubility	0.007g per 100g/H_2O	almost insoluble

Extraction of sulfur—the Frasch process

The Frasch process

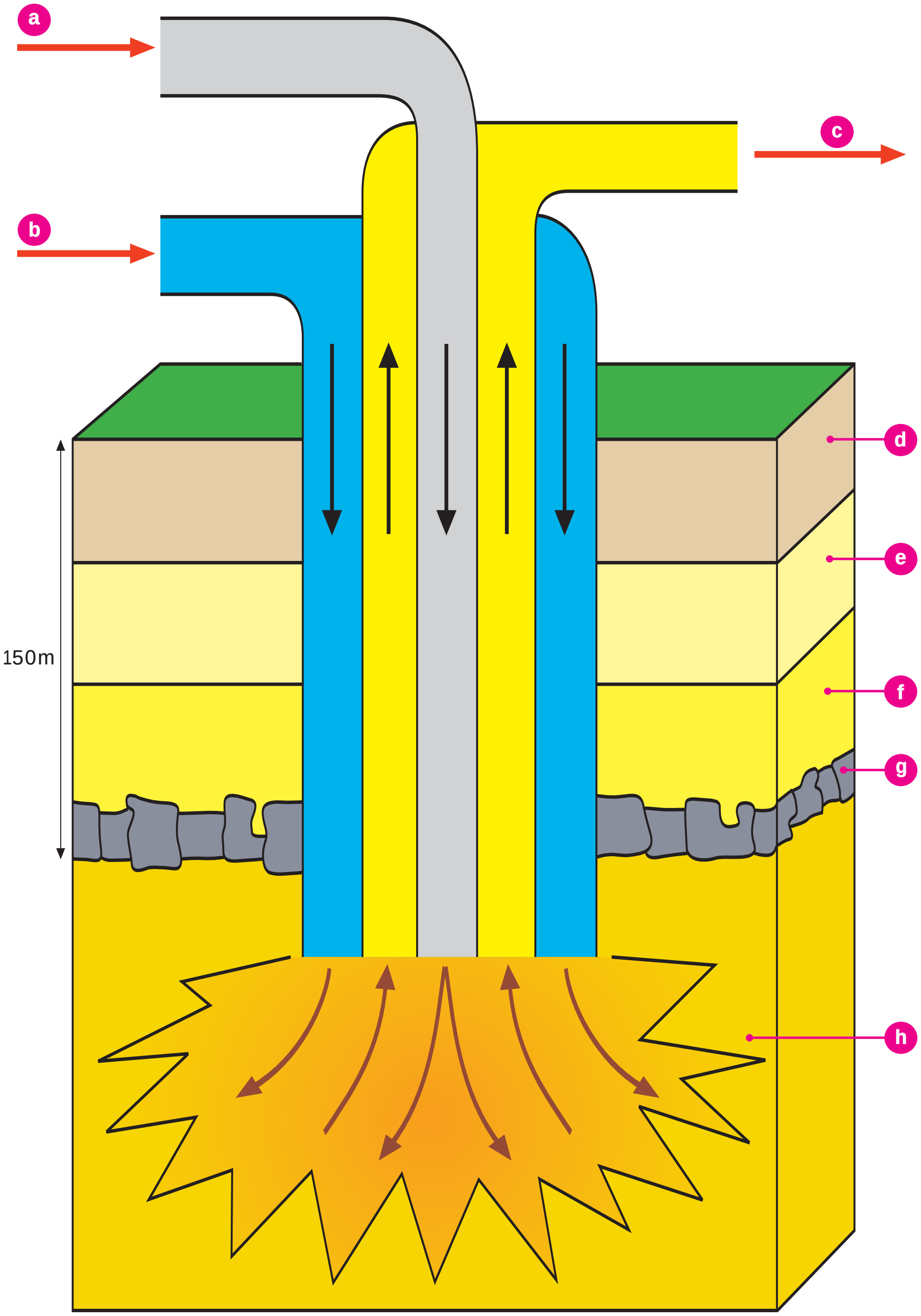

a hot compressed air
b superheated water (at 170°C)
c molten sulfur and water
d clay
e quicksand
f sand
g limestone
h sulfur

Key words

ore
sulfide
sulfur
sulfur dioxide

Processing and uses

- Hot compressed air and superheated steam are piped underground. This forces water and molten *sulfur* to the surface.
- The sulfur obtained is about 99.5 percent pure, and may be stored and transported molten or allowed to cool and solidify.
- A significant proportion of the elemental sulfur used in industry is obtained as a by-product of other industrial processes such as the refining of metal *sulfide ores* and petroleum refining.
- In petroleum refining, sulfur compounds like thiols (R-SH) and disulphides (R-S-S-R) are removed from some of the petroleum fractions because they would damage the catalysts used in refining processes and also because of their potential to cause environmental problems. For example, if they were not removed from fuels like gasoline, they would burn to form *sulfur dioxide*. This gas dissolves in water in the atmosphere, forming an acid, and would significantly increase the acidity of rain water.
- Sulfur compounds are converted to hydrogen sulfide by catalytic hydrogenation:

$$\text{R-SH(g)} + H_2\text{(g)} \rightarrow \text{R-H(g)} + H_2S\text{(g)}$$

$$\text{R-S-S-R(g)} + 3H_2\text{(g)} \rightarrow 2\text{R-H(g)} + 2H_2S\text{(g)}$$

- Hydrogen sulfide can be converted to sulfur using the Claus process:

$$6H_2S\text{(g)} + 5O_2\text{(g)} \rightarrow 2SO_2\text{(g)} + 2S_2\text{(g)} + 6H_2O\text{(g)}$$

$$4H_2S\text{(g)} + 2SO_2\text{(g)} \rightarrow 3S_2\text{(g)} + 4H_2O\text{(g)}$$

© Diagram Visual Information Ltd.

Oxygen and sulfur: allotropes

Key words

- *allotrope*
- *bond*
- *diatomic molecule*
- *oxygen*
- *ozone*
- *sulfur*
- *viscosity*

Allotropes

- *Allotropes* are different forms of the same element in the same physical state. Many elements, including *oxygen* and *sulfur*, exist as more than one allotrope.

1 Oxygen

- The most common form of oxygen is a *diatomic molecule*, O_2. The gas also exists as a triatomic molecule, O_3, which is called *ozone*.
- Oxygen has two *bonds* between the atoms. Each atom donates one electron to each bond.
- In ozone, the central oxygen atom donates a pair of electrons to form a bond with the other two atoms. One of the other atoms also donates a pair of electrons, while the other does not.

2 Sulfur

- Sulfur has several allotropes in the solid form, including rhombic sulfur and monocline sulfur.
- Rhombic sulfur crystals have a lemon-yellow appearance.
- Monocline sulfur crystals are needle-like and have a deeper yellow color.
- Each allotrope is composed of S8 puckered molecular rings, but arranged in different ways.

3 Heating sulfur

- Sulfur melts when gently heated, and the sulfur molecules are able to move around, forming a low-*viscosity* liquid.
- On stronger heating, the sulfur rings break open, yielding sulfur molecules. These molecules join by cross-linking, causing a sharp increase in viscosity.
- On even stronger heating, the cross-linked structure breaks, yielding small sulfur molecules, which are free-moving, and the viscosity falls.

© Diagram Visual Information Ltd.

1 Allotrope of oxygen

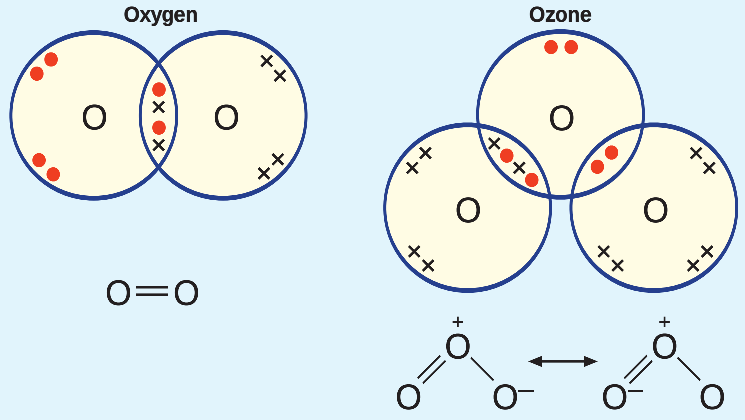

2 Allotropes of sulfur

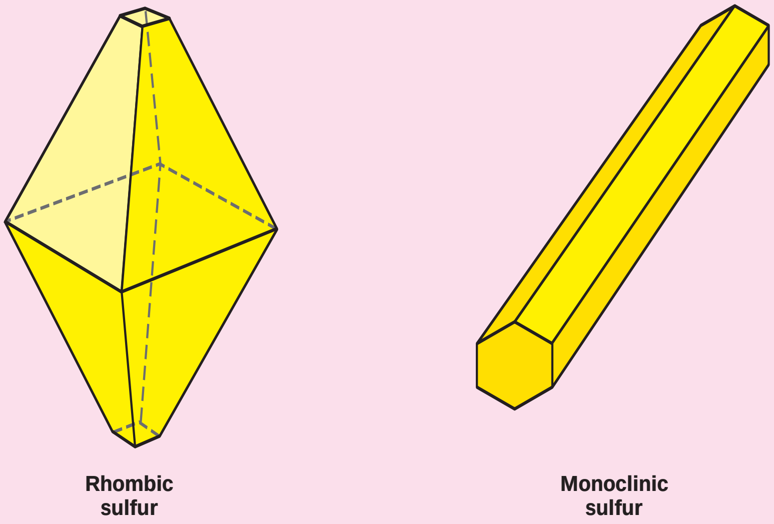

3

Structure	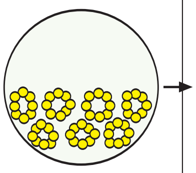	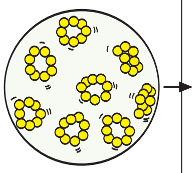	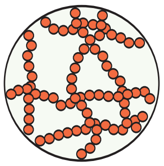	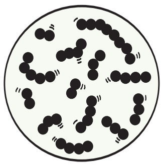
Color	yellow	yellow	red	black
State of matter	solid	liquid	liquid	liquid
Viscosity	–	low	high	low

Oxygen and sulfur: compound formation

1 Oxygen and magnesium

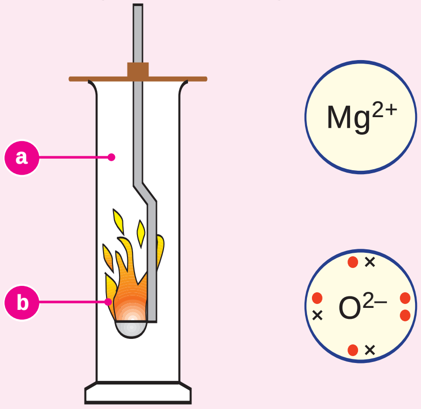

Oxygen reacts with magnesiun

Magnesium oxide is ionic

2 Reaction of sulfur and iron

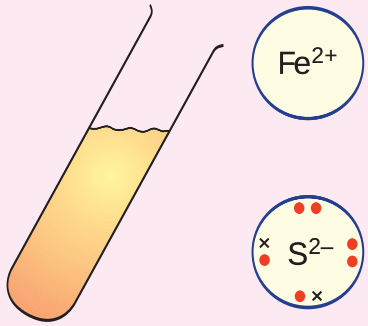

Sulfur and iron glow

Iron sulfide is also ionic

3 Phosphorous reacts with atmospheric oxygen

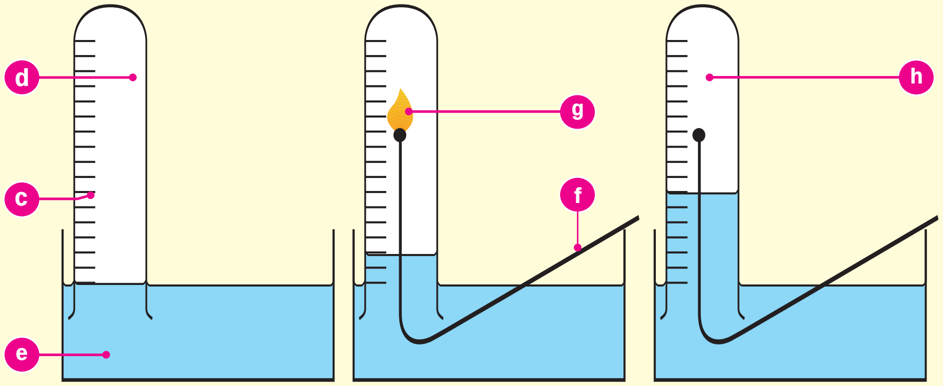

A tube full of air

Phosphorus is added

Unreacted phosphorus is left behind

a oxgen
b magnesium burning
c volume scale
d 100 cm^3 of air
e water
f stiff wire
g phosphorus reacting
h 79 cm^3 of air is left behind

4 Water, hydrogen peroxide, and hydrogen sulfide

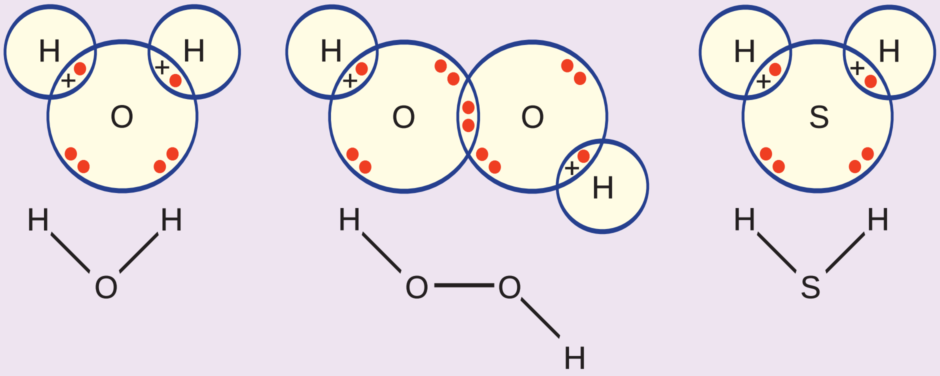

Water is formed when oxygen and hydrogen are exploded together

Hydrogen peroxide is made by reacting a metal peroxide with acid

Hydrogen sulfide is made by the reaction of a metal sulfide with dilute acid

Key words

hydrogen
hydrogen peroxide
hydrogen sulfide
magnesium
oxide
oxygen
peroxide
sulfide
sulfur

1 Oxygen and magnesium

- *Oxygen* reacts with most metals to form metal *oxides*. Magnesium burns in air with a bright flame, producing a white smoke of magnesium oxide. The reaction is even more vigorous in pure oxygen:
  $2Mg(s) + O_2(g) \rightarrow 2MgO$
- Soluble metal oxides dissolve in water to form alkaline solutions:
  $MgO(s) + H_2O(l) \rightarrow Mg(OH)_2(aq)$

2 Sulfur and iron

- *Sulfur* forms *sulfides* with many metals. When iron and sulfur are heated together, iron sulfide is formed:
  $8Fe + S_8 = 8FeS$

3 Oxygen and phosphorous

- Oxygen also reacts with non-metals to form oxides. Phosphorus burns in air to form phosphorus(V) oxide. Approximately 20 percent of the air is used:
  $P_4(s) + 5O_2(g) \rightarrow P_4O_{10}(s)$
- Non-metal oxides dissolve in water to form acids:
  $P_4O_{10}(s) + H_2O(l) \rightarrow 4H_3PO_4(aq)$

4 Other common reactions

- *Hydrogen* burns in oxygen to form water:
  $2H_2(g) + O_2(g) \rightarrow 2H_2O(l)$
- *Hydrogen peroxide* is formed by the reaction of metal *peroxides*, such as barium peroxide, with dilute acids:
  $BaO_2(s) + H_2SO_4(aq) \rightarrow$
  $H_2O_2(aq) + BaSO_4(s)$
- *Hydrogen sulfide* is formed by the reaction of a metal sulfide with a dilute acid:
  $FeS(s) + 2HCl(aq) \rightarrow$
  $FeCl_2(aq) + H_2S(g)$

© Diagram Visual Information Ltd.

The oxides of sulfur

Key words

catalyst
lone pair
oxide
sulfite
sulfur
sulfur dioxide
sulfurous acid
sulfur trioxide

Sulfur oxides

- *Sulfur* combines with oxygen to form two *oxides*: *sulfur dioxide* (sulfur(IV) oxide) and *sulfur trioxide* (sulfur(VI) oxide.

1 Laboratory preparation of sulfur dioxide

- Metal *sulfites*, such as sodium sulfite, react with dilute acids to from sulfur dioxide:
  $Na_2SO_3(s) + 2HCl(aq) \rightarrow$
  $2NaCl(aq) + H_2O(l) + SO_2(g)$
- Sulfur dioxide is dried by passing it through anhydrous calcium chloride and collected by downward delivery.
- Sulfur dioxide dissolves in water to form *sulfurous acid*:
  $H_2O(l) + SO_2(g) \rightarrow H_2SO_3(aq)$

2 Laboratory preparation of sulfur trioxide

- Sulfur trioxide is formed when dry sulfur dioxide and oxygen are heated in the presence of a platinized asbestos *catalyst*:
  $2SO_2(g) + O_2(g) \rightleftharpoons 2SO_3(g)$
- Sulfur trioxide forms needle-like crystals when cooled.
- Sulfur trioxide dissolves in water to form sulfuric acid:
  $H_2O(l) + SO_3(g) \rightarrow H_2SO_4(aq)$

3 SO_2 and SO_3 molecules

- In sulfur dioxide, there are four pairs of bonding electrons and a non-bonding or *lone pair* of electrons around the sulfur atom. The pairs of electrons are kept as far from each other as possible by adopting a bent shape in which the double bonds between the sulfur and oxygen atoms are at an angle of 120°.
- In sulfur trioxide, the three double bonds form a trigonal planar structure around the sulfur atom in which the bond angle is also 120°.

© Diagram Visual Information Ltd.

1 Laboratory preparation of sulfur dioxide

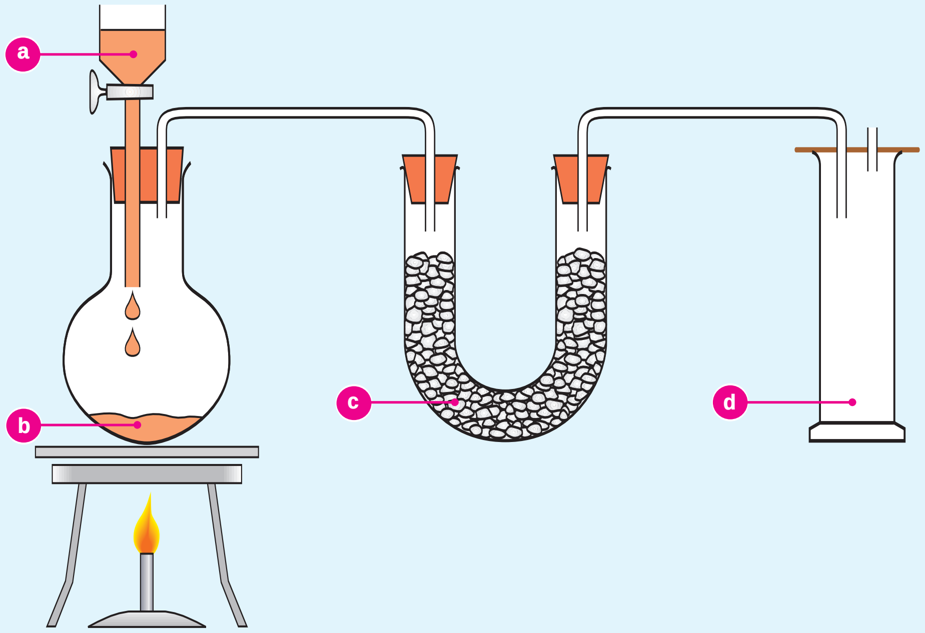

2 Laboratory preparation of sulfur trioxide

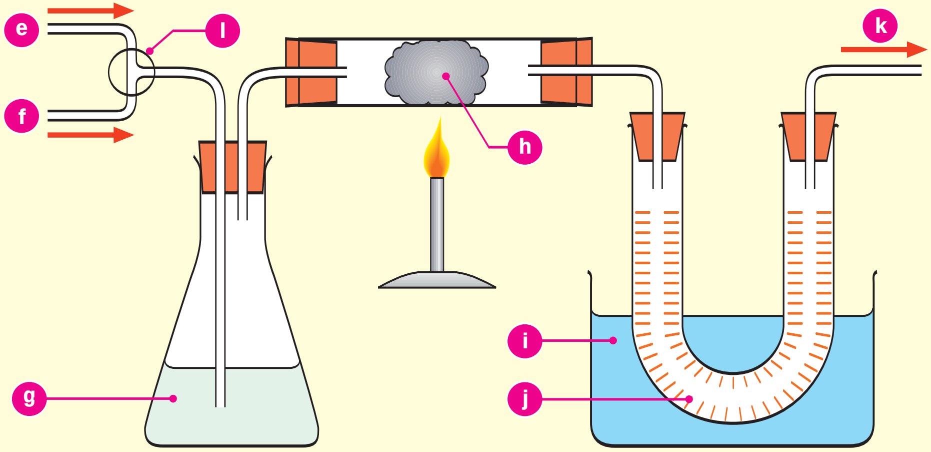

a dilute
b sodium sulfite
c anhydrous $CaCl_2$
d upward displacement
e oxygen
f sulfur dioxide
g concentrated sulfuric acid for drying
h plantinized asbestos
i freezing mixture of ice and salt
j needles of sulfur oxide
k to the fume cupboard
l 3-way tap

3 SO_2 and SO_3 molecules

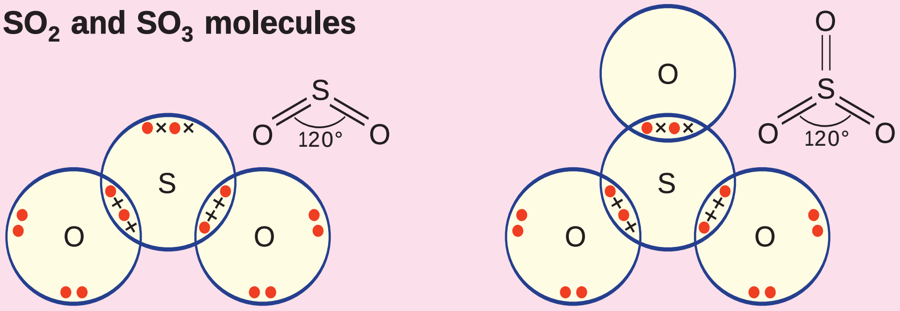

Molecule of sulfur dioxide

Molecule of sulfur trioxide

Industrial preparation of sulfuric acid (the contact process): theory

1 Sulfur burning

$$S(1) + O_2(g) \rightarrow SO_2(g) \qquad \Delta H = -297\ \text{kJ mol}^{-1}$$

2 Conversion

$$\underset{\text{4 moles}}{2SO_2(g) + O_2(g)} \rightleftharpoons \underset{\text{2 moles}}{2SO_3(g)} \qquad \Delta H = -192\ \text{kJ mol}^{-1}$$

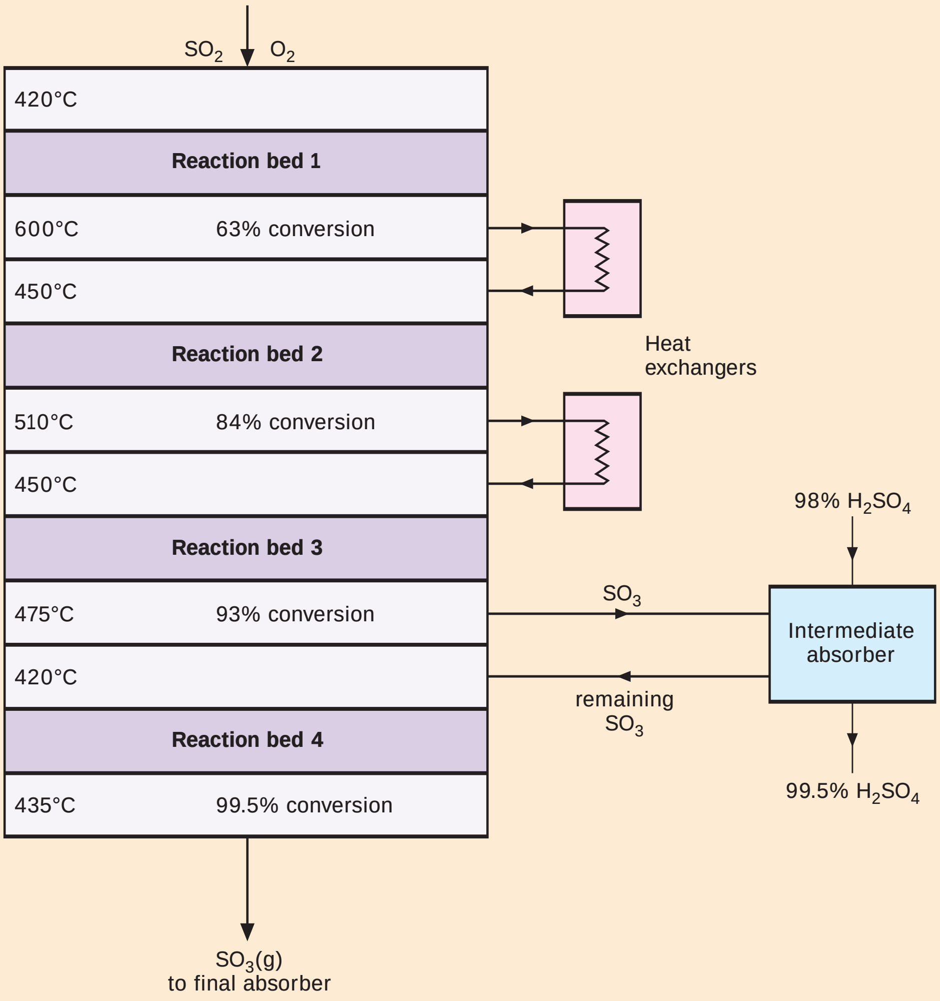
SO_2
O_2
420°C
Reaction bed 1
600°C
63% conversion
450°C
Reaction bed 2
510°C
84% conversion
450°C
Reaction bed 3
475°C
93% conversion
420°C
Reaction bed 4
435°C
99.5% conversion
Heat exchangers
98% H_2SO_4
SO_3
Intermediate absorber
remaining SO_3
99.5% H_2SO_4
$SO_3(g)$ to final absorber

3 Absorption

$$H_2SO_4(1) + SO_3(g) \rightleftharpoons H_2S_2O_7(1)$$

$$H_2S_2O_7(1) + H_2O(1) \rightarrow 2H_2SO_4(1)$$

Key words

catalyst
exothermic
equilibrium
sulfur
sulfur dioxide
sulfuric acid
sulfur trioxide

Preparation of sulfuric acid

- The industrial preparation of *sulfuric acid* is a three stage process:
  1. *Sulfur* burning
  2. Conversion of *sulfur dioxide* to *sulfur trioxide*
  3. Absorption of sulfur trioxide to form sulfuric acid.

1 Sulfur burning

- Molten sulfur is sprayed into a furnace and burned in a blast of dry air. The reaction is very *exothermic*, and the reaction temperature rises to over 1,000°C. The mixture of gases, containing sulfur dioxide and oxygen, is cooled before conversion.

2 Conversion

- The conversion of sulfur dioxide into sulfur trioxide is an exothermic reaction. The forward reaction is favored by a low temperature. However, this would also reduce the rate of reaction, so it would take longer for the reaction mixture to reach *equilibrium*.
- The reaction is carried out at temperatures between 420–620°C in the presence of a vanadium(V) oxide *catalyst*.
- Modern converters consist of four reaction beds. The reaction mixture is cooled after passing through each of the first two beds in order to maximize conversion in subsequent beds.

3 Absorption

- Sulfur trioxide is removed and absorbed after the reaction mixture has passed through both the third and fourth beds.
- Sulfur trioxide is absorbed into 98 percent sulfuric acid to form 99.5 percent sulfuric acid, which is sometimes referred to as oleum and given the chemical formula $H_2S_2O_7$.
- Oleum is then diluted to give 98 percent sulfuric acid.

© Diagram Visual Information Ltd.

Industrial preparation of sulfuric acid (the contact process): schematic

Key words

catalyst
ore
sulfide
sulfur
sulfur dioxide
sulfuric acid
sulfur trioxide

Preparing sulfuric acid

- The *sulfur* needed for the manufacture of *sulfuric acid* is obtained either directly from the ground or as a by-product of other industrial processes, such as the refining of metal *sulfide ores* and the refining of crude oil.
- An excess of dry air is used to ensure there is sufficient oxygen remaining in the reaction mixture for the conversion of *sulfur dioxide* to *sulfur trioxide*.
- The vanadium(V) oxide (vanadium pentoxide) *catalyst* is activated by potassium sulfate on a silica support. It is generally in the form of small cylindrical pellets that ensure a large surface area for reaction. The catalyst is inactive below about 380°C and has an optimum working temperature between 420–620°C.
- The catalyst pellets are packed onto perforated plate supports to form reaction beds. In a modern converter, there four reaction beds, each consisting of a layer of catalyst pellets about 0.6 m deep.
- A conversion of 99.5 percent of sulfur dioxide to sulfur trioxide is essential for both economic and environmental reasons. This can only be attained by removing heat between reaction beds and by removing the sulfur trioxide produced between the third and fourth beds.
- The mixture of gases that remain after absorption consists mostly of nitrogen, together with a small proportion of oxygen and traces of sulfur dioxide. The gases are filtered to remove any sulfuric acid before being released into the atmosphere at high level via a stack.
- Sulfuric acid is used in a wide variety of industries. Important uses include the manufacture of general chemicals, paints and pigments, detergents and soaps, and phosphatic fertilizers.

© Diagram Visual Information Ltd.

Industrial preparation of sulfuric acid

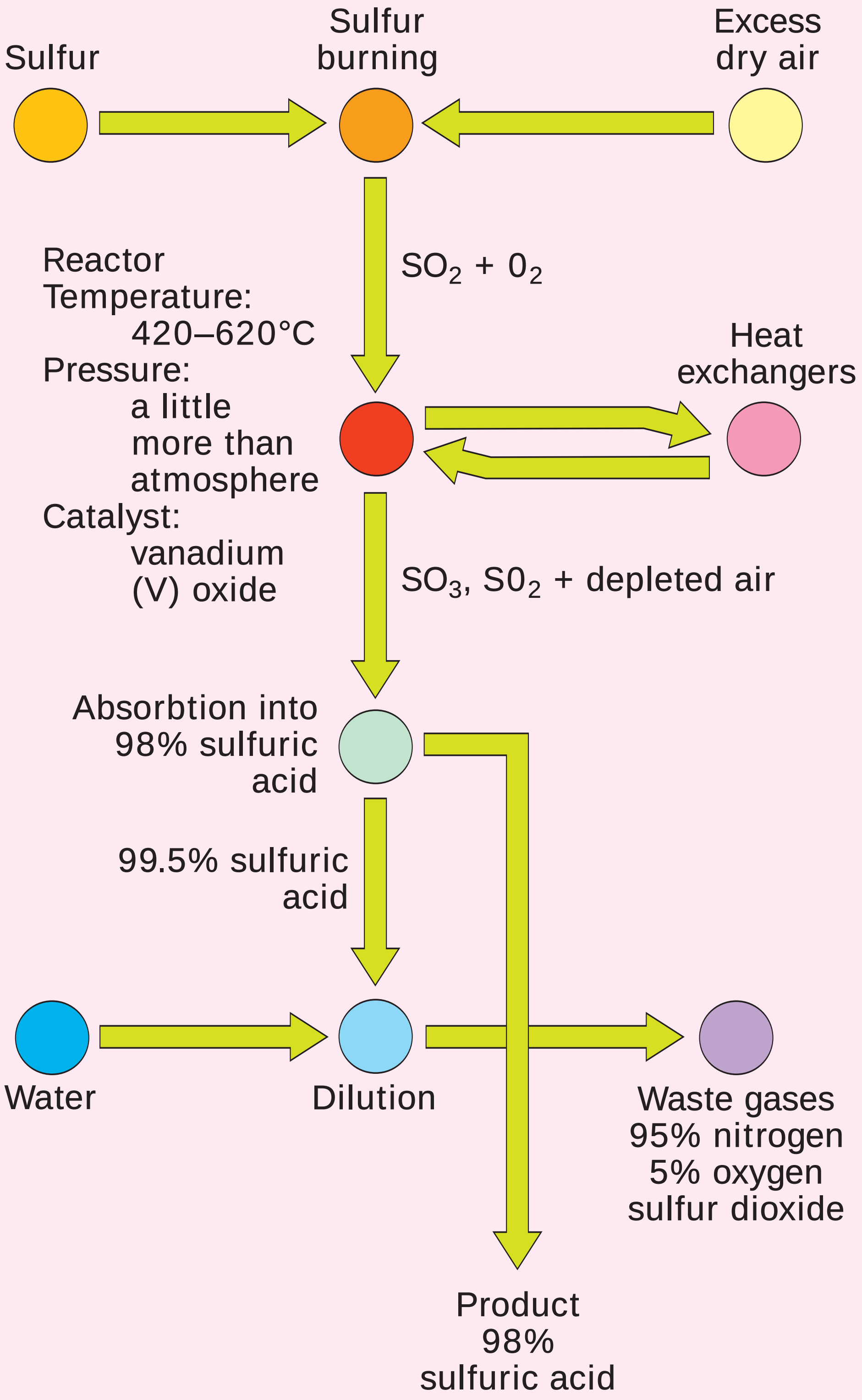
Sulfur
Sulfur burning
Excess dry air
$SO_2 + 0_2$
Reactor
Temperature: 420–620°C
Pressure: a little more than atmosphere
Catalyst: vanadium (V) oxide
Heat exchangers
SO_3, $S0_2$ + depleted air
Absorbtion into 98% sulfuric acid
99.5% sulfuric acid
Water
Dilution
Waste gases 95% nitrogen 5% oxygen sulfur dioxide
Product 98% sulfuric acid

Affinity of concentrated sulfuric acid for water

1 Sulfuric acid and atmospheric water

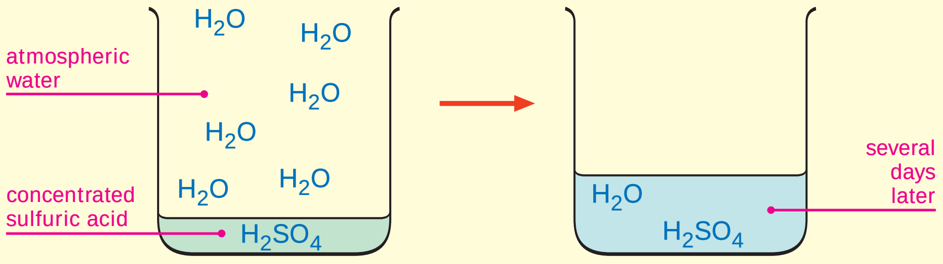
atmospheric water
concentrated sulfuric acid
H_2O
H_2SO_4
several days later

2 Illustrating sulfuric acid's affinity for water

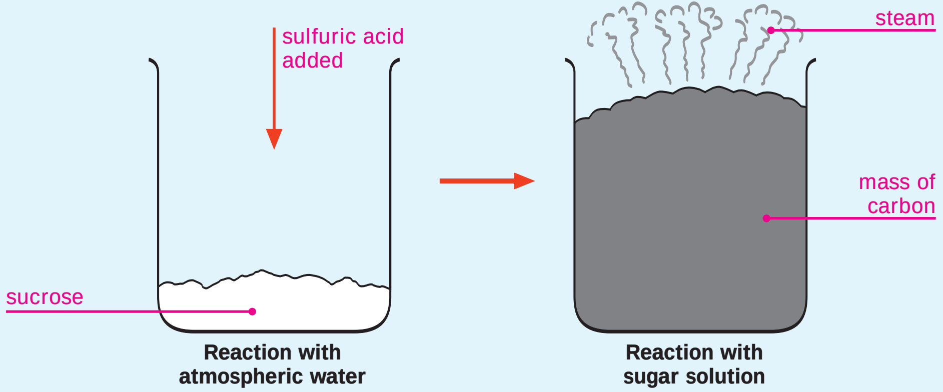
sulfuric acid added
sucrose
steam
mass of carbon
Reaction with atmospheric water
Reaction with sugar solution

3 Drying gases

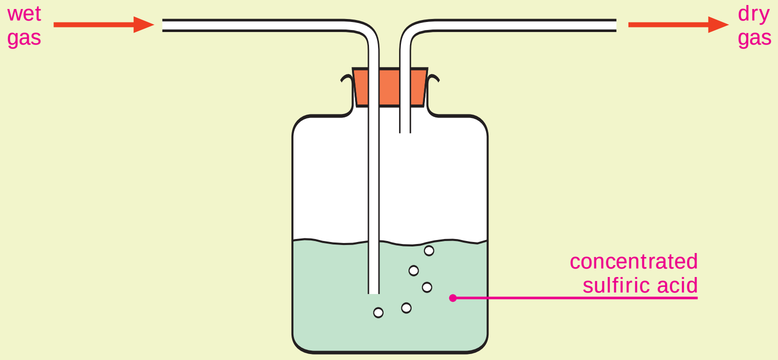
wet gas
dry gas
concentrated sulfiric acid

4 Forming esters and alkenes

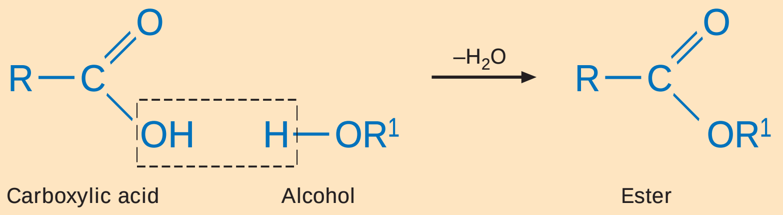
R
C
O
OH
H
OR1
–H2O
Carboxylic acid
Alcohol
Ester

Key words

alcohol
alkene
carboxylic acid
catalyst
ester
sulfuric acid

1 Sulfuric acid and atmospheric water

- Concentrated *sulfuric acid* will absorb water from the air.
- The level of liquid in a beaker containing concentrated sulfuric acid rises as water is absorbed and the acid becomes more dilute.

2 Sulfuric acid's affinity for water

- Concentrated sulfuric acid will remove the elements of water from organic chemicals such as sucrose ($C_{12}H_{22}O_{11}$).
- When concentrated sulfuric acid is poured onto sucrose, the crystalline white sucrose turns into a black amorphous mass of carbon.

3 Drying gases

- Concentrated sulfuric acid can be used to dry gases, such as hydrogen. The gas is bubbled through the acid in a suitable container.
- Some gases, such as ammonia, react with concentrated sulfuric acid and cannot be dried in this way.

4 Forming esters and alkenes

- Concentrated sulfuric acid is used as a *catalyst* in the formation of *esters* from *carboxylic acids* and *alcohols*. Concentrated sulfuric acid removes the elements of water from alcohols to form *alkenes*:

$$R\text{-}CH_2\text{-}CH_2OH \xrightarrow{-H_2O} R\text{-}CH{=}CH_2$$

© Diagram Visual Information Ltd.

Oxygen and sulfur: oxidation and reduction

Key words

hydrogen peroxide
hydrogen sulfide
oxidizing agent
oxygen
redox reaction
reduction
sulfur
sulfur dioxide

1 Common redox reactions

- *Oxidation* reactions and *reduction* reactions are more accurately described as *redox reactions* because they cannot occur in isolation. One substance is reduced and another is oxidized at the same time.
- *Hydrogen sulfide* reduces chlorine gas to chloride ions. The sulfide is oxidized to elemental *sulfur*:
  $Cl_2 + 2e^- \rightarrow 2Cl^-$ reduction
  $S^{2-} \rightarrow S + 2e^-$ oxidation
- *Hydrogen peroxide* is an oxidizing agent but is itself oxidized by a more powerful oxidizing agent such as potassium manganate(VII).
- *Sulfur dioxide* is a reducing agent both in the gaseous form and in aqueous solution. It reduces iron(III) to iron(II) while being itself oxidized to a sulfate:
  $Fe^{3+} + e^- \rightarrow Fe^{2+}$ reduction
- Sulfur dioxide can be oxidized to sulfur trioxide by reaction with *oxygen*.
- Copper reacts with concentrated sulfuric acid. In this reaction, the concentrated sulfuric acid acts as an oxidizing agent, oxidizing copper metal to copper(II). The sulfuric acid is itself reduced to sulfur dioxide:
  $Cu \rightarrow Cu^{2+} + 2e^-$ oxidation

2 Sulfuric acid and sulfate reactions

- Sulfuric acid reacts with metal oxides to form sulfates and water.
- Sulfuric acid reacts with metal hydroxides to form sulfates and water
- Sulfuric acid reacts with metal carbonates to form salts, carbon dioxide, and water
- In all of these reactions, no oxidation or reduction takes place. The charge on the metal ion in the oxide, hydroxide, and carbonate is the same as it is in the sulfate.

© Diagram Visual Information Ltd.

1 Common redox reactions

$$8H_2S + 8Cl_2 \longrightarrow S_8 + 16HCl$$

Hydrogen sulfide reduces chlorine gas

$$H_2O_2 + 3KMnO_4 \longrightarrow 2KOH + 2MnO_2 + 2O_2$$

Hydrogen peroxide reduces potassium manganate(VII) solution

$$SO_2 + 2FeCl_3 + 2H_2O \longrightarrow 2FeCl_2 + 2HCl + H_2SO_4$$

Sulfur oxide reduces iron chloride solution

$$O_2 + 2SO_2 \longrightarrow 2SO_3$$

Oxygen oxidizes sulfur dioxide

$$Cu + 2H_2SO_4 \longrightarrow CUSO_4 + H_2O + SO_2\uparrow$$

Concentrated sulfuric acid oxidizes copper

2 Sulfuric acid and sulfate reactions

Acid + metal oxide = Salt + water

$$H_2SO_4 + MgO \longrightarrow MgSO_4 + H_2O$$

Acid + metal hydroxide = Salt + water

$$H_2SO_4 + Cu(OH)_2 \longrightarrow CuSO_4 + 2H_2O$$

Acid + metal carbonate = Salt + carbon dioxide + water

$$H_2SO_4 + ZnCO_3 \longrightarrow ZnSO_4 + CO_2 + H_2O$$

Basic reactions of oxygen

1 Reactions

Oxidizes other elements

$2H_2 + O_2 \longrightarrow 2H_2O$

$4Na + O_2 \longrightarrow 2Na_2O$

$2Na + O_2 \longrightarrow Na_2O_2$

$S_8 + 8O_2 \longrightarrow 8SO_2$

$P_4 + 5O_2 \longrightarrow P_4O_{10}$

$2Cu + O_2 \longrightarrow 2CuO$

Oxidizes other elements

$2CO + O_2 \longrightarrow 2CO_2$

$2SO_2 + O_2 \rightleftharpoons 2SO_3$

$4NH_3 + 5O_2 \longrightarrow 4NO + 6H_2O$

2 The hydrides

Water

ⓐ $2Na + 2H_2O \longrightarrow 2NaOH + H_2\uparrow$

ⓑ $H_2O + HCl \longrightarrow H_3O^+ + Cl^-$

ⓒ $H_2O + NH_3 \rightleftharpoons NH_4^+ + OH^-$

ⓓ $H_2O + CO_2 \rightleftharpoons H_2CO_3$

$H_2O + SO_2 \rightleftharpoons H_2SO_3$

Oxidizes other elements

ⓔ $BaO_2 + H_2SO_4 \longrightarrow H_2O_2 + BaO_4\downarrow$

ⓕ $H_2O_4 + 2FeCl_2 + 2HCl \longrightarrow 2FeCl_3 + 2H_2O$

ⓖ $H_2O_2 + 2KMnO_4 \longrightarrow 2KOH + 2MnO_2 + 2O_2$

3 The oxides

Oxidizes other elements

ⓗ $CuO + H_2SO_4 \longrightarrow CuSO_4 + H_2O$

ⓘ $ZnO + H_2SO_4 \longrightarrow ZnSO_4 + H_2O$

$ZnO + 2NaOH \longrightarrow Na_2Zn(CH)_4 + H_2O$

Oxidizes other elements

ⓙ SO_2, CO_2

ⓚ H_2O, CO

a oxidizes reactive metals
b accepts protons from acids
c gives protons to bases
d reacts with non-metal oxides
e preparation
f an oxiding agent
g a reducing agent
h basic
i amphoteric
j acidic
k neutral

Key words

alkali
amphoteric
cation
hydroxide
oxide
oxygen
transition metals
valency

1 Reactions

- *Oxygen* reacts with both metals and non-metals to form *oxides*.
- Group 1 and group 2 metals often form more than one oxide:

Na_2O	sodium monoxide
Na_2O_2	sodium peroxide
NaO_2	sodium dioxide

- *Transition metals* form oxides in which the metal exhibits different *valency* states:

Cu_2O	copper(I) oxide
CuO	copper(II) oxide

- Non-metallic elements also form more than one oxide:

P_4O_6	phosphorus(III) oxide
P_4O_{10}	phosphorus(V) oxide

2 The hydrides

- Water reacts with group 1 and group 2 metals to give the metal *hydroxide* and hydrogen. In these reactions, the metal is oxidized to a metal *cation* and the water is reduced:
  $M \rightarrow M^+$ and $M \rightarrow M^{2+}$
- In the Brønsted-Lowry theory of acids and bases, an acid is defined as a substance that donates protons and a base as a substance that accepts protons. Water acts as both an acid and a base.
- Non-metallic oxides dissolve in water to form acids:
  $H_2CO_3 \rightleftharpoons H^+ + HCO_3^-$ carbonic acid
  $H_2SO_3 \rightleftharpoons H^+ + HSO_3$ sulfurous acid

3 The oxides

- Metal oxides are basic and react with acids to form salts and water.
- Non-metallic oxides are acidic or neutral.
- Some metal oxides are *amphoteric*: they react with both acids and *alkalis*.

© Diagram Visual Information Ltd.

Basic reactions of sulfur

Key words

dehydrating agent
hydrogen sulfide
oxidizing agent
redox reaction
sulfide
sulfur
sulfur dioxide
sulfuric acid

1 Reactions of sulfur

- *Sulfur* reacts with most metals and hydrogen to form *sulfides*. Metal sulfides react with acids to give *hydrogen sulfide*:
  $FeS(s) + 2HCl(aq) \rightarrow FeCl_2(aq) + H_2S(g)$
- Sulfur combines with non-metals such as oxygen and chlorine.
- *Sulfur dioxide* is formed by the reaction between concentrated *sulfuric acid* and sulfur.

2 The hydrides

- Hydrogen sulfide dissolves in water to form a weak acid.
- Group 1 metal sulfides are soluble in water, however, other metal sulfides are only sparingly soluble and form characteristically colored precipitates. Moist lead(II) ethanoate paper turns black in the presence of hydrogen sulfide due to the formation of lead sulfide:
  $Pb^{2+} + S^{2-} \rightarrow PbS$
- Hydrogen sulfide and sulfur dioxide undergo a *redox reaction* to form elemental sulfur.

3 The oxides SO_2 and SO_3

- Sulfur dioxide dissolves in water to form sulfurous acid, a weak acid.
- Sulfur dioxide is oxidized to sulfur trioxide.
- Sulfur trioxide dissolves in water to form sulfuric acid, a strong acid.

4 The hydroxy compound

- Concentrated sulfuric acid is a strong *oxidizing agent* and will oxidize both metals and non-metals.
- Concentrated sulfuric acid is a strong *dehyrating agent* and will remove water or the elements of water.

© Diagram Visual Information Ltd.

1 Reactions of sulfur

Oxidizes some elements

$$Fe + S \longrightarrow FeS$$
$$2Cu + S \longrightarrow Cu_2S$$
$$H_2 + S \longrightarrow H_2S$$

Reduces other elements elements

$$S + O_2 \longrightarrow SO_2$$
$$2S + Cl_2 \longrightarrow S_2Cl_2$$

Reduces some compounds

$$3S + 2H_2SO_4 \longrightarrow 3SO_2\uparrow + 2H_2O$$

2 The hydrides

Is a sparingly soluble acidic gas

$$H_2S + H_2O \rightleftharpoons H_3O^+ + HS^-$$

Is a reducing agent

$$2H_2S + SO_2 \longrightarrow 2H_2O + 3S\downarrow$$

Causes precipitation of insoluble metel sulfides

$$CuSO_4 + H_2S \longrightarrow CuS\downarrow + H_2SO_4$$

3 The oxides SO_2 and SO_3

Is a sparingly soluble acidic gas

$$SO_2 + H_2O \rightleftharpoons H_2SO_3$$

Is reducing agent

$$2SO_2 + O_2 \rightleftharpoons 2SO_3$$
$$SO_2 + 2H_2S \longrightarrow 3S\downarrow + 2H_2O$$

SO_3 is very acidic

$$SO_3 + H_2O \longrightarrow H_2SO_4$$

4 The hydroxy compound

As a strong acid

$$H_2SO_4 + H_2O \longrightarrow H_3O^+ + HSO_4^-$$

As an oxidizing agent

$$2H_2SO_4 + C \longrightarrow CO_2 + 2SO_2\uparrow + 2H_2O$$
$$Cu + 2H_2SO_4 \longrightarrow CuSO_4 + SO_2\uparrow + 2H_2O$$

As a dehydrating agent

$$CH_3CH_2OH \xrightarrow[170°C]{H_2SO_4} CH_2 = CH_2$$
$$C_{12}H_{22}O_{11} \xrightarrow[-11H_2O]{H_2SO_4} 12C$$
$$CuSO_4.5H_2O \xrightarrow[-5H_2O]{H_2SO_4} CuSO_4$$

The halogens: group 7

1 Electron structure

F			2	7
Cl			8	7
Br		2	18	7
I	2	8	18	7

inner electrons **outer shell**

2 Halogen atom and molecule

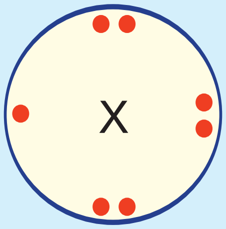

Halogen atom

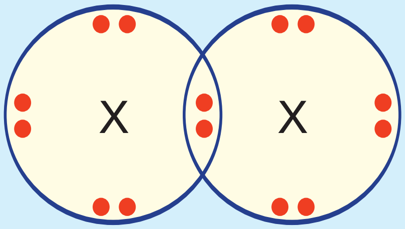

Halogen molecule

3 Physical properties

Element	Fluorine	Chlorine	Bromine	Iodine
Atomic number	9	17	35	53
Relative atomic mass	19.0	35.5	79.9	126.9
State at 20°C	gas	gas	liquid	solid
Color	pale yellow	pale green	red-brown	black
m.p./°C	–220	–101	–7	113
b.p./°C	–188	–35	59	183
Solubility/g per 100g of water at 20°C	reacts readily with water	0.59 (reacts slightly	3.6	0.018

Key words

astatine
bromine
chlorine
covalent bond
covalent compound
fluorine
halogens
iodine
ionic compound

Halogens

- The elements of group 7 are sometimes referred to as the *halogens*. They are *flourine*, *chlorine*, *bromine*, *iodine*, and *astatine*. The symbol 'X' is often used to denote a halogen atom and 'X^-' a halogen ion.

1 Electron structure

- All halogen atoms have seven electrons in their outer shell. A halogen atom needs one more electron to fill the outer shell, and it can obtain this either by forming a single *covalent bond* or by forming an ion, X^-. Halogens form both *covalent compounds* and *ionic compounds*.

2 Halogen atom and molecule

- Halogens exist as diatomic molecules. Each atom in the molecule provides a single electron to form a covalent bond. The result is that each atom has control over eight electrons.

3 Physical properties

- There is a gradation of physical properties going down group 7.
  1. State changes from solid to liquid to gas. Bromine is one of only two elements that exist as liquids at room temperature.
  2. The color darkens from pale yellow to black.
  3. Melting point and boiling point increase.
- There is a gradual decrease in chemical reactivity going down group 7.
- Fluorine oxidizes water to give oxygen: $2F_2 + 2H_2O \rightarrow 4HF + O_2$
- Chlorine reacts less vigorously with water, forming an acidic solution: $C_2 + H_2O \rightleftharpoons HCl + HOCl$
- Bromine and iodine form solutions in water, although the latter is not very soluble.

© Diagram Visual Information Ltd.

Laboratory preparation of the halogens

Key words

bromine
chlorine
distillation
iodine
oxidation
oxidizing agent

1 Laboratory preparation of chlorine

- *Chlorine* is made in the laboratory by the *oxidation* of concentrated hydrochloric acid using a suitable *oxidizing agent* such as manganese dioxide (manganese(IV) oxide):
  $MnO_2(s) + 4HCl(aq) \rightarrow$
  $MnCl_2(aq) + 2H_2O(l) + Cl_2(g)$
- The gas is first passed through water to remove any hydrogen chloride gas, and then through concentrated sulfuric acid to dry the gas. Chlorine is more dense than air and is collected by downward delivery.
- Chlorine can also be conveniently made in the laboratory from bleaching powder, using dilute hydrochloric acid:
  $Ca(OCl)_2(s) + 4HCl(aq) \rightarrow$
  $CaCl_2(aq) + 2H_2O(l) + 2Cl_2(g)$

2 Laboratory preparation of bromine

- *Bromine* is made in a similar way to chlorine:
  $MnO_2(s) + 2NaBr(aq) + 2H_2SO_4(aq) \rightarrow$
  $MnSO_4(aq) + Na_2SO_4(aq)$
  $+ 2H_2O(l) + Br_2(g)$
- Because it boils at 59°C, bromine is removed from the reaction mixture by *distillation*.

3 Laboratory preparation of iodine

- *Iodine* is made in a similar way to bromine. Hydrogen iodide is made *in situ* by reacting sodium iodide with concentrated sulfuric acid:
  $MnO_2(s) + 2KI(aq) + 2H_2SO_4(aq) \rightarrow$
  $MnSO_4(aq) + K_2SO_4(aq) +$
  $2H_2O(l) + I_2(g)$
- Iodine is removed from the reaction mixture by sublimation. On heating, it changes directly from solid to vapor and then back to solid on cooling.

© Diagram Visual Information Ltd.

1 Laboratory preparation of chlorine

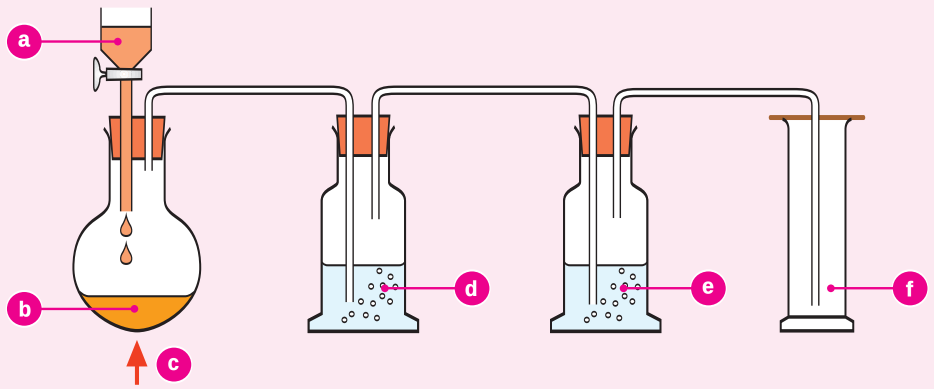

2 Laboratory preparation of bromine

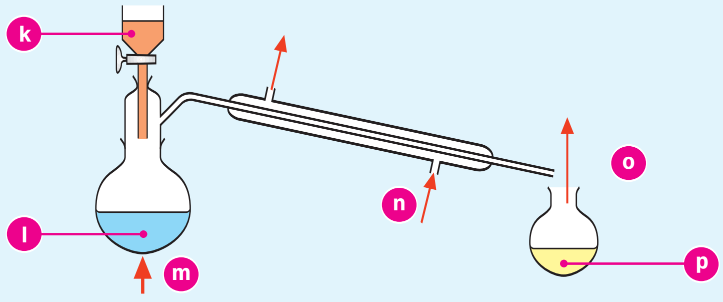

3 Laboratory preparation of iodine

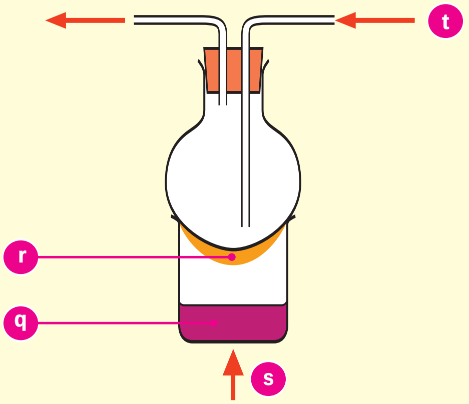

a	concentrated hydrochloric acid
b	manganese dioxide
c	warm gently
d	water to remove HCl fumes
e	concentrated H_2SO_4 to dry Cl_2
f	chlorine gas
g	concentrated brine
h	chlorine gas
i	hydrogen gas
j	water
k	concentrated sulfuric acid
l	manganese oxide + sodium bromide
m	warm gently
p	cold water
o	fumes of HBr
p	bromine
q	manganese oxide + potassium iodine + concentrated H_2SO_4
s	warm gently
t	cold water

Compounds of chlorine

1 Chlorine and metals

Calcium burns in chlorine

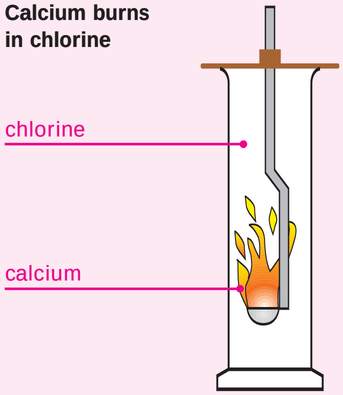

Aluminum reacts when warmed in a stream of chloride

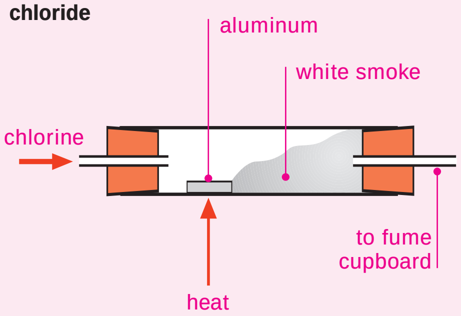

2 Laboratory preparation of hydrogen chloride

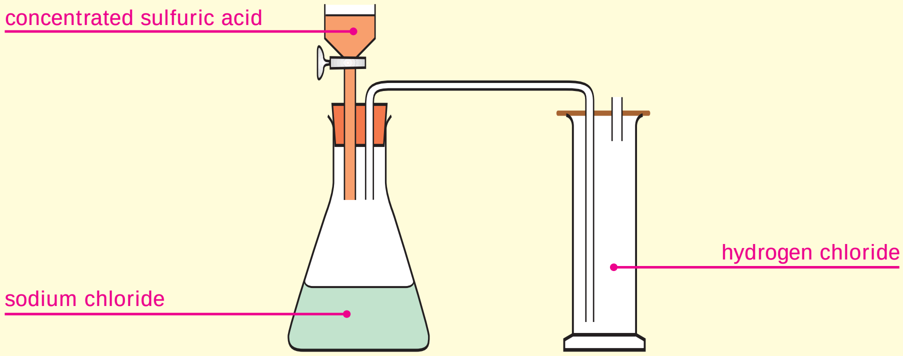

3 Compounds with non-metals

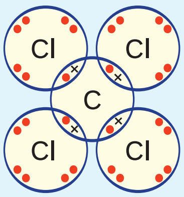

Carbon tetrachloride (CCl_4)

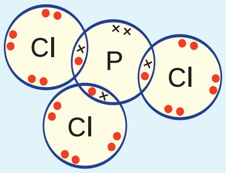

Phosphorus trichloride (PCl_3)

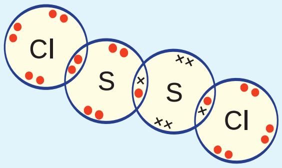

Sulfur monochloride (S_2Cl_2)

4 The structure of some chlorinated pesticides

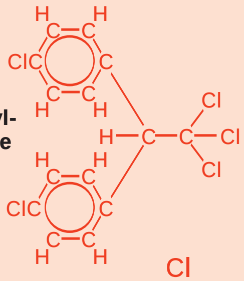

Dichlorophenyl-trichloroethane (DDT)

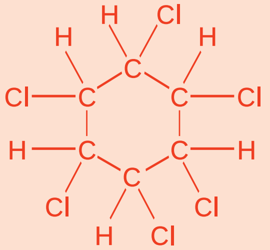

Benzene hexachloride (BHC)

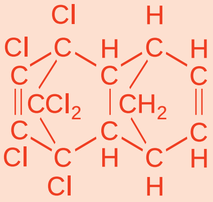

Aldrin

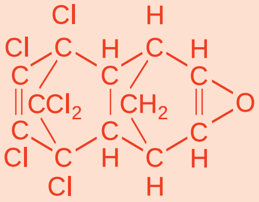

Dieldrin

Key words

chloride
chlorine
covalent compound
halide
hydrogen chloride
sodium chloride
sulfuric acid

1 Chlorine and metals

- Metals, such as calcium, burn in *chlorine* to produce the corresponding metal *chloride*:
  $Ca(s) + Cl_2(g) \rightarrow CaCl_2(s)$
- Aluminum reacts with chlorine to form aluminum chloride:
  $2Al(s) + 3Cl_2(g) \rightarrow 2AlCl_3(s)$
- Unlike many metal chlorides, aluminum chloride is hydrolyzed by water, giving off *hydrogen chloride* gas:
  $AlCl_3(s) + 3H_2O(l) \rightarrow Al(OH)_3(s) + 3HCl(g)$
  It is for this reason that aluminum halides fume when they come into contact with moist air.

2 Laboratory preparation of hydrogen chloride

- Hydrogen chloride is made by the reaction of *sodium chloride* with concentrated *sulfuric acid*:
  $NaCl(s) + H_2SO_4(aq) \rightarrow NaHSO_4(s) + HCl(g)$

3 Compounds with non-metals

- Chlorine forms *covalent compounds* with non-metals such as carbon, phosphorus, and sulfur.

4 Pesticides

- Chlorinated compounds provide a range of pesticides.
- DDT, BHC, Aldrin, and Dieldrin have been the source of environmental concern, and their use is now prohibited or severely restricted.

© Diagram Visual Information Ltd.

Hydrogen chloride in solution

Key words

carbonate
covalent bond
covalent compound
hydrogen chloride
ion
ionic compound

Hydrogen chloride

- *Hydrogen chloride* gas is a *covalent compound*. In solutions in organic solvents, it remains a covalent compound. It becomes an *ionic compound* in aqueous solutions.

1 In organic solvents

- In solution in organic solvents such as methylbenzene, hydrogen chloride remains a covalent compound. The hydrogen atom and the chlorine atom each donate one electron to form the *covalent bond*.
- The solution contains no ions and does not conduct electricity.
- The solution has no effect on blue litmus paper or on *carbonates*, thus showing that it is not an acid.

2 In aqueous solution

- In aqueous solution, hydrogen chloride becomes an ionic compound. The hydrogen atom loses an electron to become a hydrogen *ion*, H^+, and the chlorine atom gains an electron to become a chloride ion, Cl^-.
- The solution contains ions and conducts electricity. The ions are able to carry a charge through the solution.
- The solution turns blue litmus paper red and reacts with carbonates, showing that it is an acid:

$$Na_2CO_3(aq) + 2HCl(aq) \rightarrow 2NaCl(aq) + CO_2(g) + H_2O(l)$$

1 In organic solvents

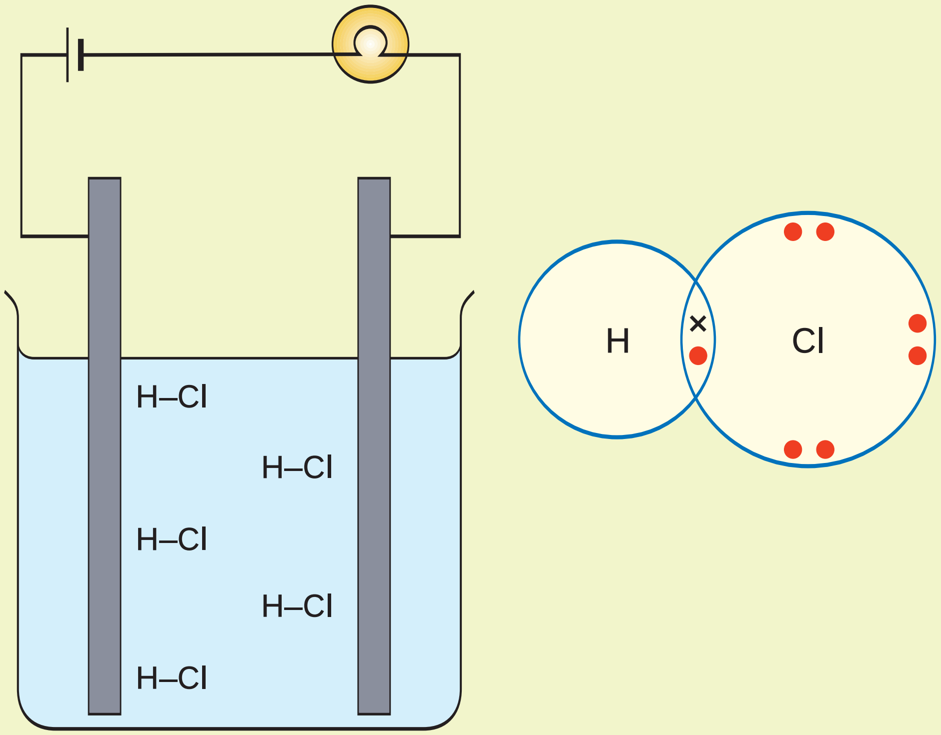

2 In aqueous solution

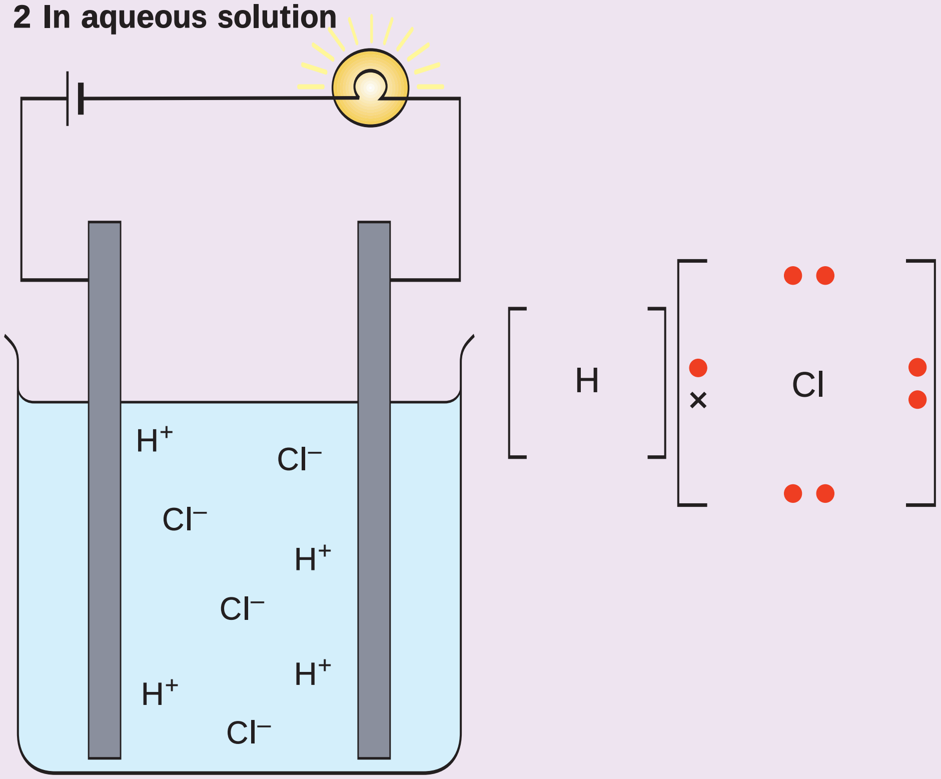

© Diagram Visual Information Ltd.

Acid/base chemistry of the halogens

1 Laboratory preparation of hydrochloric acid

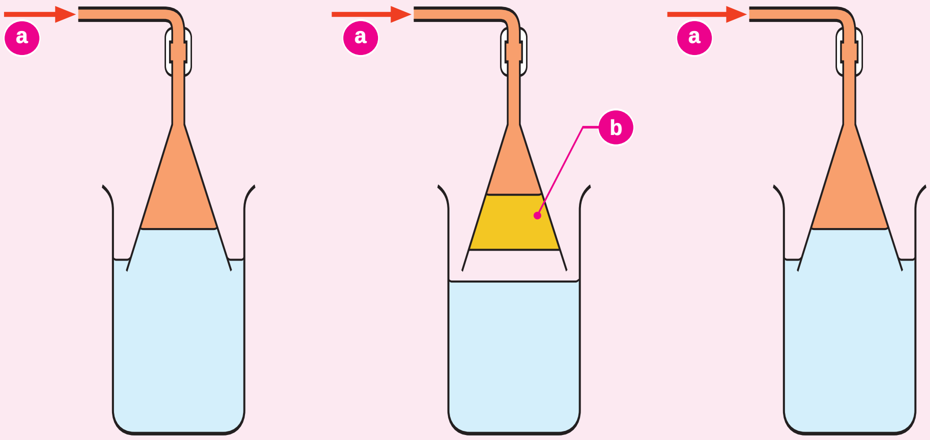

Hydrogen chloride enters water via a filter funnel

The level in the beaker drops

The cycle starts again

a hydrochloric acid (HCl)
b plug of liquid in funnel

2 Solubility of the halogens

$$2F_2 + 2H_2O \longrightarrow 4HF + O_2$$

Flouorine is so reactive it decomposes water producing hydrofluoric acid and oxygen

$$Cl_2 + H_2O \rightleftharpoons HCl + HOCl$$

Chlorine is the next most reactive halogen after fluorine

3 Chloride test

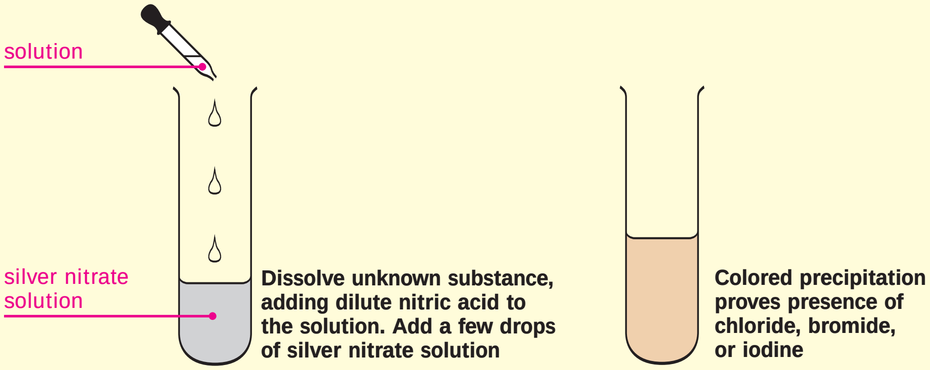

Dissolve unknown substance, adding dilute nitric acid to the solution. Add a few drops of silver nitrate solution

Colored precipitation proves presence of chloride, bromide, or iodine

Key words

halide
hydrochloric acid
hydrogen chloride
nitric acid
oxidizing agent
reducing agent
silver nitrate

1 Laboratory preparation of hydrochloric acid

- *Hydrochloric acid* is made in the laboratory by dissolving *hydrogen chloride* gas in water.
- Hydrogen chloride is very soluble in water. It is dissolved by passing through an inverted filter funnel, the rim of which sits just below the water level. When water is sucked into the funnel, the water level drops, and the funnel rim is no longer submerged. This prevents water being sucked back into the apparatus.

2 Solubility of the halogens

- All halogens are *oxidizing agents*. However, oxidizing power decreases down the group:
  fluorine > chlorine > bromine > iodine
- *Halide* ions are *reducing agents*. The reducing power increases down the group:
  fluorine < chlorine < bromine < iodine

3 Chloride test

- The presence of halide ions in solution can be detected by adding a few drops of dilute *nitric acid* followed by several drops of *silver nitrate* solution.
  1. Chloride ions form a white precipitate of insoluble silver chloride:
  $Ag^+(aq) + Cl^-(aq) \rightarrow AgCl(s)$
  2. Bromide ions form a cream precipitate of insoluble silver bromide:
  $Ag^+(aq) + Br^-(aq) \rightarrow AgBr(s)$
  3. Iodide ions form a yellow precipitate of insoluble silver iodide:
  $Ag^+(aq) + I^-(aq) \rightarrow AgI(s)$

© Diagram Visual Information Ltd.

Key words

chloride
chlorine
noble gases
oxidizing agent
reactivity series
redox reaction
sulfur

1 Calcium and chlorine

- When hydrogen sulfide and chlorine are mixed, elemental *sulfur* is formed. *Chlorine* acts as an *oxidizing agent* and oxidizes the hydrogen sulfide by removing hydrogen. In turn, the chlorine gains hydrogen and is reduced to hydrogen chloride:
$8H_2S(g) + 8Cl_2(g) \rightarrow S_8(s) + 16HCl(g)$

2 Chlorine and ferrous chloride

- Chlorine can also be used to oxidize iron(II) to iron(III). When chlorine is bubbled into iron(II) chloride solution, the color changes from green to yellow-brown, showing the formation of iron(III). The chlorine atoms are reduced to *chloride* ions:
$2FeCl_2(aq) + Cl_2(g) \rightarrow 2FeCl_3(aq)$

3 Halogens and metals

- Halogens readily oxidize metals. Fluorine oxidizes all metals, including gold and silver, easily.
- Chlorine oxidizes all but the least reactive metals. When iron is heated in a stream of dry chlorine, iron(III) chloride is produced:
$2Fe(s) + 3Cl_2(g) \rightarrow 2FeCl_3(s)$
- The ease with which halogens oxidize metals decreases down the group, but even iodine will slowly oxidize metals low in the *reactivity series*.

4 Halogens and non-metals

- Fluorine oxidizes most non-metals except nitrogen and most of the *noble gases*.
- Chlorine reacts directly with phosphorus and sulfur, but carbon, nitrogen, and oxygen do not react directly with chlorine, bromine, or iodine.
- The relative reactivities of the halogens in *redox reactions* with non-metals is illustrated at right by their reaction with hydrogen.

© Diagram Visual Information Ltd.

Redox reactions of the halogens

1 Calcium burns in chlorine

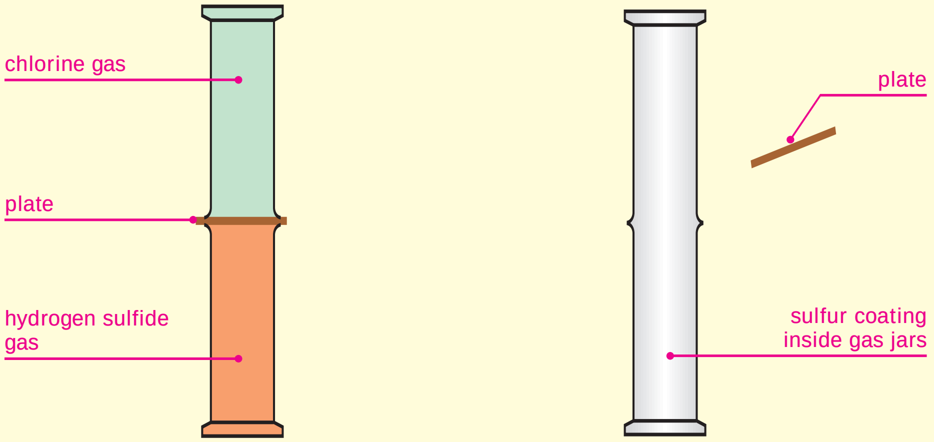

Chlorine gas and hydrogen sulfide gas are separated by plate

Plate is removed

2 Reaction of chlorine with ferrous chloride

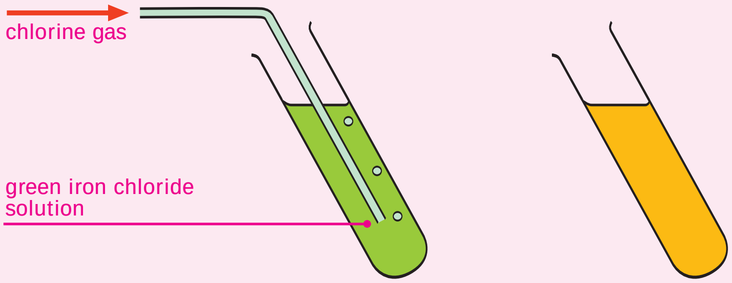

Chlorine gas is passed into ferrous chloride solution

Yellow-brown iron chloride solution

3 Halogens and metals

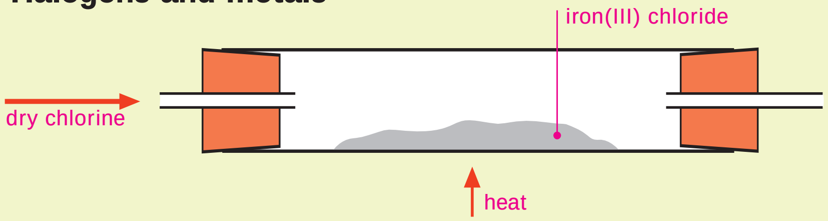

4 Halogens and non-metals

Reaction	Observations
$H_2(g) + F_2(g) \rightarrow 2HF(g)$	explosive
$H_2(g) + Cl_2(g) \rightarrow 2HCl(g)$	explosive in sunlight but slow in the dark
$H_2(g) + Br_2(g) \rightarrow 2HBr(g)$	needs heat and a catalyst
$H_2(g) + I_2(g) \rightarrow 2HI(g)$	slow even when heated

Reactivity of the halogens

1 Chemical reactivity of halogens with each other

	Chlorine	Bromine	Iodine
Chloride		✗	✗
Bromide	✓		✗
Iodide	✓	✓	

2 Reaction of chlorine and bromine

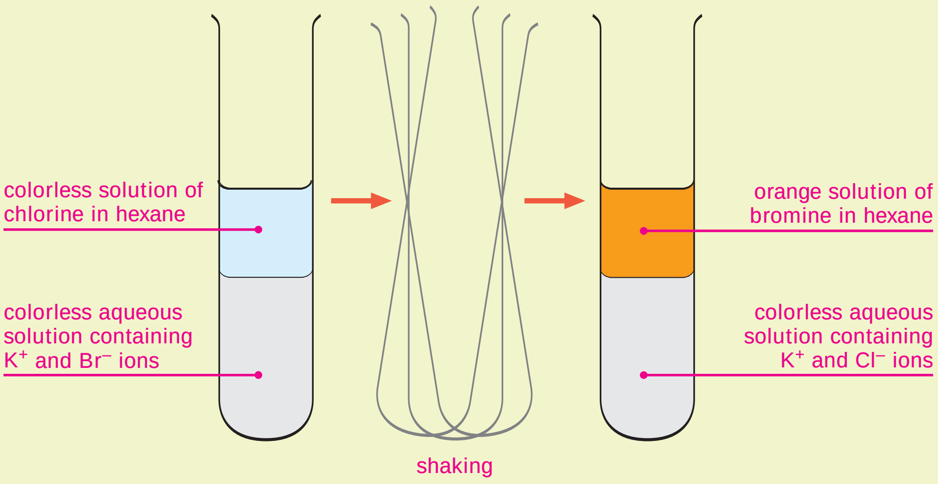
colorless solution of chlorine in hexane
colorless aqueous solution containing K⁺ and Br⁻ ions
shaking
orange solution of bromine in hexane
colorless aqueous solution containing K⁺ and Cl⁻ ions

3 Reaction of chlorine and iodine

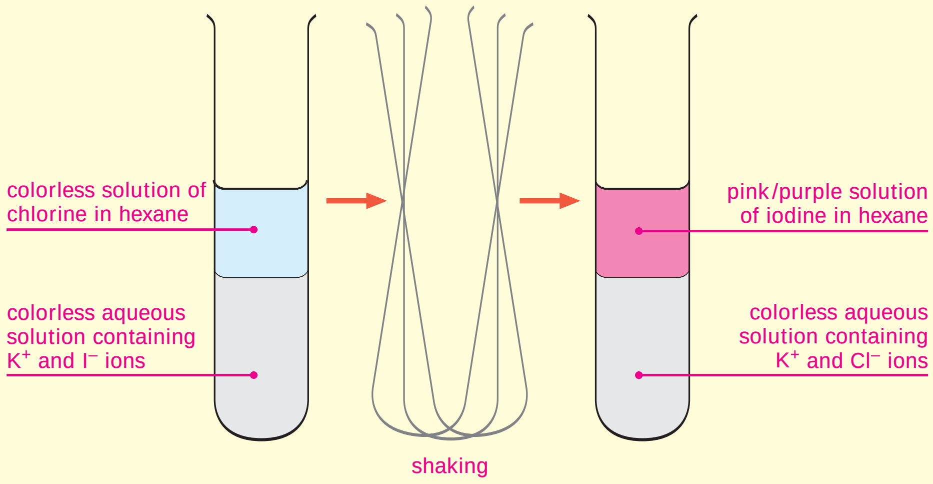
colorless solution of chlorine in hexane
colorless aqueous solution containing K⁺ and I⁻ ions
shaking
pink/purple solution of iodine in hexane
colorless aqueous solution containing K⁺ and Cl⁻ ions

Key words

displacement reaction
halide
halogens
immiscible

1 Reactivity of halogens

- The chemical reactivity of the *halogens* decreases down group 7: fluorine > chlorine > bromine > iodine
- A more reactive halogen will displace the ions of a less reactive halogen from a metal *halide* solution. This is called a *displacement reaction*.
- Chlorine will displace bromide ions and iodide ions from solution.
- Bromine will displace iodide ions from solution.
- In a displacement reaction, the halogen acts as an oxidizing agent and is reduced while the halide ion is oxidized.

2 Chlorine and bromine

- Chlorine dissolves in the organic solvent hexane to give a colorless or slightly green solution, depending on concentration.
- Hexane is *immiscible* with water. When the two liquids are mixed, hexane forms a layer above water.
- When solutions of chlorine in hexane and potassium bromide in water are shaken together, chlorine displaces bromide ions from the aqueous solution. Bromine is more soluble in hexane than in water, and an orange layer of bromine in hexane forms:

$2KBr + Cl_2 \rightarrow 2KCl + Br_2$

$2Br^- + Cl_2 \rightarrow 2Cl^- + Br_2$

3 Chlorine and iodine

- When solutions of chlorine in hexane and potassium iodide in water are shaken together, chlorine displaces iodide ions from the aqueous solution. Iodine is more soluble in hexane than in water, and a pink-purple layer of iodine in hexane forms:

$2KI + Cl_2 \rightarrow 2KCl + I_2$

$2I^- + Cl_2 \rightarrow 2Cl^- + I_2$

© Diagram Visual Information Ltd.

World distribution of metals

Key words

- aluminum
- copper
- gold
- iron
- lead
- platinum
- silver
- uranium
- zinc

1 Gold, sliver, platinum, and uranium

- *Gold*: South Africa, USA, Canada, Russia
- *Silver*: USA, South America
- *Platinum*: South Africa, USA, South America
- *Uranium*: North America, Europe, Central and South Africa, Australia

2 Aluminum and copper

- *Aluminum*: South America, Jamaica, West Africa, Russia, India, Australia
- *Copper*: North America, Central and South Africa, Europe, Russia

3 Iron, zinc, and lead

- *Iron*: North and South America, Russia, Europe, Angola, Australia
- *Zinc* and *lead*: USA, Europe, Australia, Russia

1 Gold, sliver, platinum, and uranium

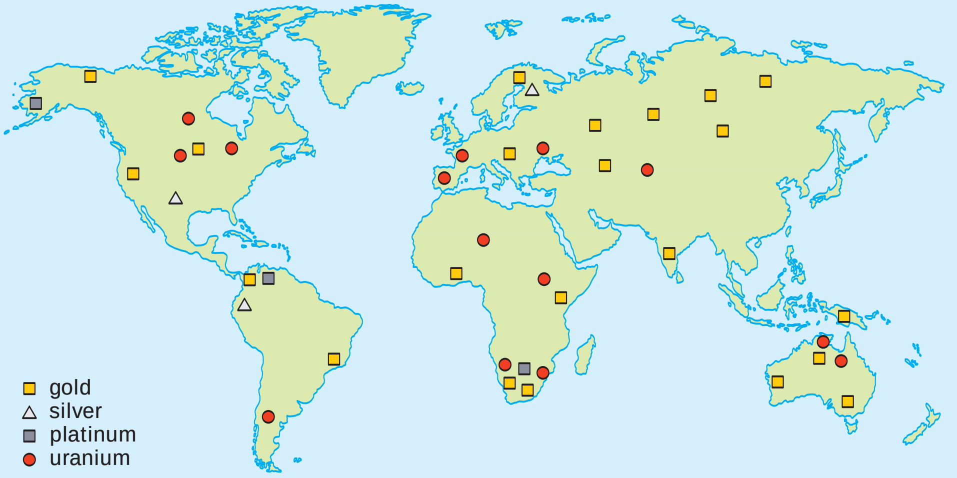

2 Aluminum and copper

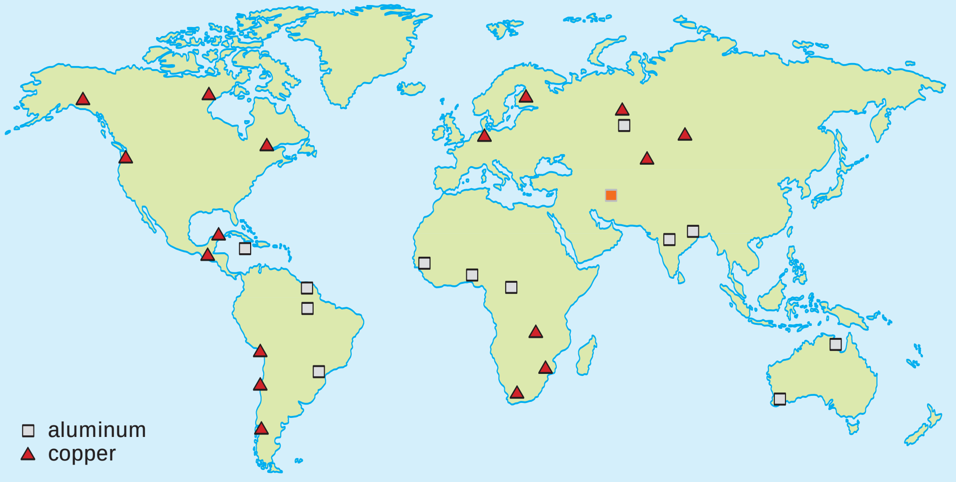

3 Iron, zinc, and lead

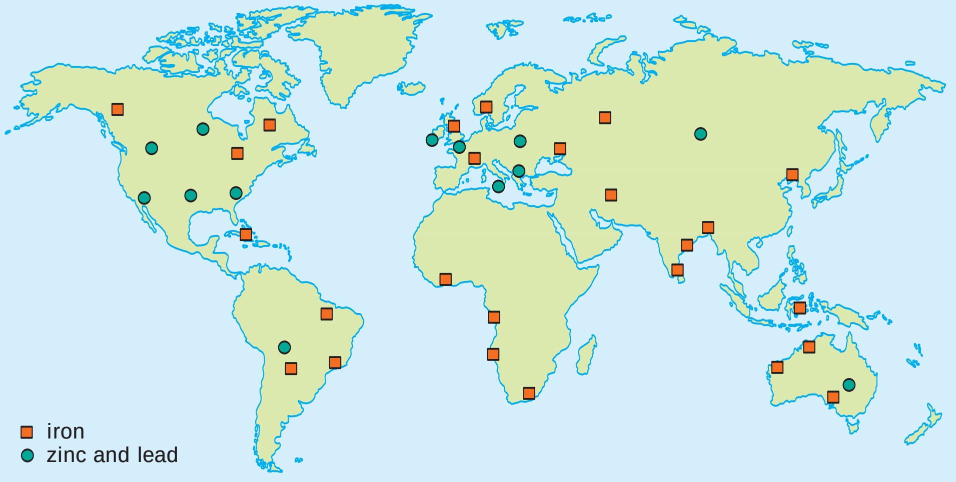

© Diagram Visual Information Ltd.

Main ores of metals

Metal **Common name for ore(s)**	**Chemical name for ore(s)**	**Formula of ore(s)**
aluminum bauxite	aluminum oxide	$Al_2O_3.2H_2O$
chromium chromite	iron chromium oxide	$FeCr_2O_4$
copper chalcopyrite bornite chalcocite	copper iron sulfide copper iron sulfide copper(I) sulfide	$CuFeS_2$ Cu_5FeS_4 Cu_2S
iron haematite magnetite	iron(III) oxide iron(II)iron(III) oxide	Fe_2O_3 Fe_3O_4
lead galena cerussite anglesite	lead(II) sulfide lead(II) carbonate lead(II) sulfate	PbS $PbCO_3$ $PbSO_4$
magnesium magnesite	magnesium carbonate	$MgCO_3$
mercury cinnabar	mercury sulfide	HgS
silver argentite	silver sulfide	Ag_2S
sodium salt	sodium chloride	$NaCl$
tin cassiterite	tin oxide	SnO_2
titanium rutile ilmenite	titanium oxide iron titanium oxide	TiO_2 $FeTiO_3$
uranium uraninite	uranium oxide	UO_2
zinc zinc blende calamine	zinc sulfide zinc carbonate	ZnS $ZnCO_3$

Key words

grade
mineral
ore
oxide
sulfide

Ores

- An *ore* is a *mineral* from which a metal (or non-metal) can be extracted.
- Metal ores are often metal *oxides* or metal *sulfides*.
- A metal may be present in a range of different minerals, but not all minerals will be suitable sources of that metal.

Recovering ores

- To be appropriate for mining, an ore must contain minerals that are valuable and that are concentrated enough to be mined profitably. It must also be economically viable to extract the ore from waste rock.
- Mineral deposits that are economically recoverable are called ore deposits. Not all mineral deposits are suitable for recovery. Some may be too low in *grade* (the concentration of the ore in the rock) or technically impossible to extract.

Formation

- The process of ore formation is called ore genesis.
- Ore genesis involves a variety of geological, internal, hydrothermal, metamorphic, and surficial processes.

© Diagram Visual Information Ltd.

The group 1 metals

Key words

alkali metals	*radioactive*
boiling point	*shell*
group 1	
melting point	
orbital	

1 Position in periodic table

- The *group 1* metals occupy the first column of the periodic table. Historically, they were known as the *alkali metals* because they all react with water to give alkaline solutions.
- The elements include lithium, sodium, potassium, rubidium, and cesium. Francium lies below cesium in the periodic table. However, it is not considered when discussing the group because it is *radioactive*, and little is known of its chemistry.

2 Electron-shell structure

- The electrons surrounding the nucleus of an atom are arranged in a series of *orbitals*, areas around the atom where there is a high probability of finding an electron. Orbitals are grouped in a series of *shells* (energy levels) at a gradually increasing distance from the nucleus. Different orbitals have different shapes: s orbitals are spherically symmetric; p orbitals point in a particular direction; and d orbitals have complicated shapes. Scientists describe an atom by describing the orbital structure. Thus, as the table indicates, sodium has 2s orbitals in the first shell, 2s and 6p orbitals in the second shell, and 1s orbital in the third shell.

3 Physical properties

- Reading down the group, the *melting point* decreases, the *boiling point* increases, the density increases, and the hardness decreases.

© Diagram Visual Information Ltd.

1 Position in the periodic table

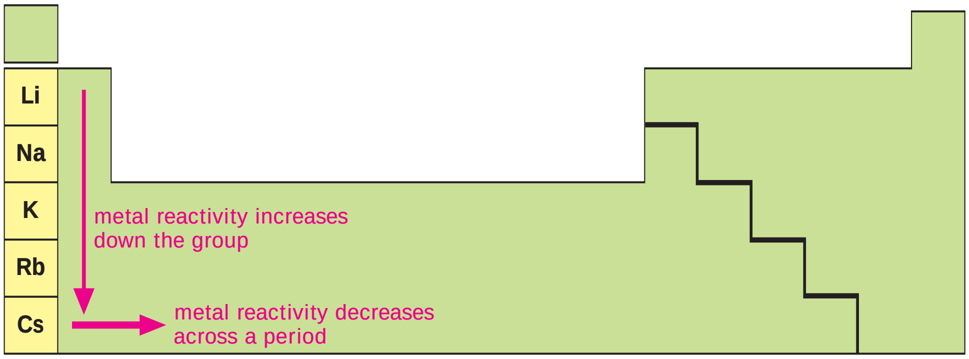
Li
Na
K
Rb
Cs
metal reactivity increases down the group
metal reactivity decreases across a period

2 Electron-shell structure

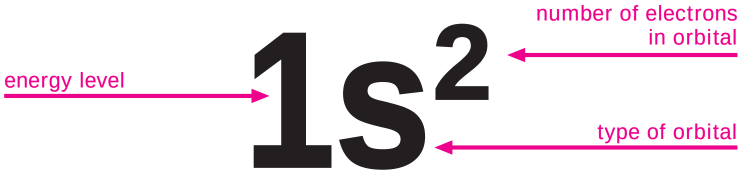

Li	$1s^2 2s^1$
Na	$1s^2\ 2s^2\ 2p^6\ 3s^1$
K	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^1$
Rb	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2\ 3d^{10}\ 4p^6\ 5s^1$
Cs	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2\ 3d^{10}\ 4p^6\ 5s^2\ 4d^{10}\ 5p^6\ 6s^1$

3 Physical properties of group I elements compared with a typical metal

Group 1 element		**m.p. / °C**	**b.p. / °C**	**Density /g cm^{-3}**	**Hardness / Moh**	**Conductivity Ω^{-1}cm^{-1}**
Lithium	Li	180	1336	0.53	0.6	11700
Sodium	Na	98	883	0.97	0.4	23800
Potassium	K	64	759	0.86	0.5	16400
Rubidium	Rb	39	700	1.53	0.3	9100
Cesium	Cs	29	669	1.88	0.2	2000
Typical metal						
Iron	Fe	1530	3000	7860	4–5	11300

The group 1 metals: sodium

1 Reaction of sodium hydroxide (NaOH) with hydrochloric acid (HCl)

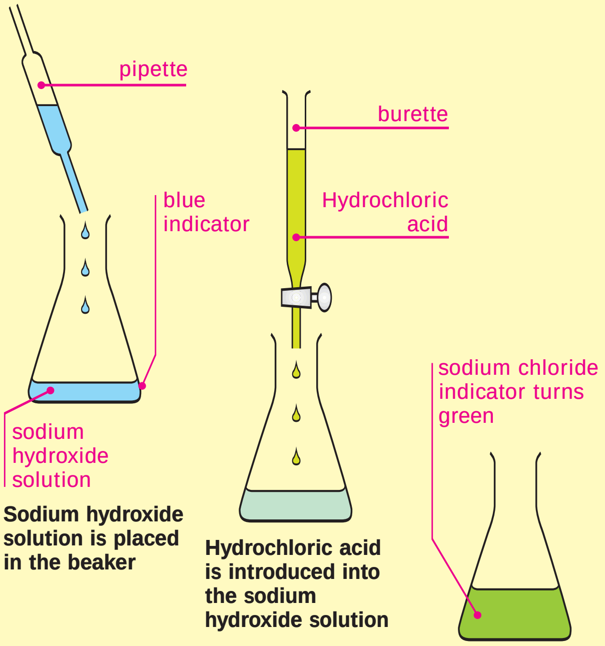
pipette
blue indicator
sodium hydroxide solution
burette
Hydrochloric acid
sodium chloride indicator turns green

Sodium hydroxide solution is placed in the beaker

Hydrochloric acid is introduced into the sodium hydroxide solution

Results of the reaction

2 Sodium burns readily in chlorine or oxygen

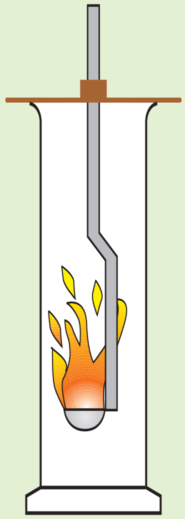

3 Sodium reacts with nitrogen gas

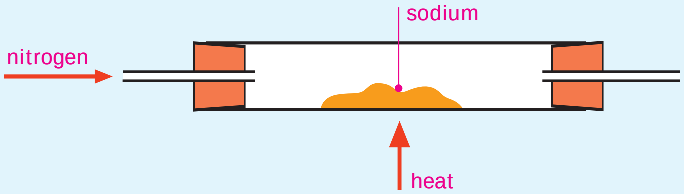
sodium
nitrogen
heat

4 Commercial preparation of sodium

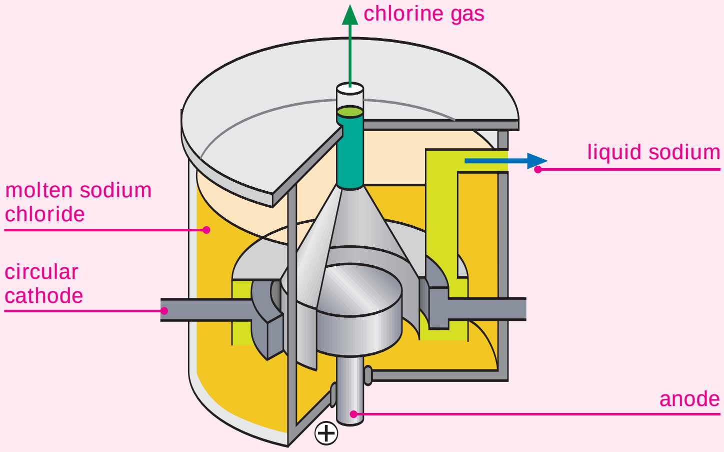
chlorine gas
liquid sodium
molten sodium chloride
circular cathode
anode

Key words

- *acid*
- *acid-base indicator*
- *alkali*
- *anode*
- *cathode*
- *electrolysis*
- *salt*
- *titration*

1 Reaction of NaOH with HCl

- Group 1 hydroxides are *alkalis*. They can be used to form *salts* by neutralizing them with *acids* in a process call *titration*. Sodium chloride is made by neutralizing sodium hydroxide with hydrochloric acid.
  $NaOH + HCl \rightarrow NaCl + H_2O$
- A given volume of sodium hydroxide solution is put into a conical flask.
- A few drops of an *acid–base indicator* are added to the sodium hydroxide solution. The indicator is a different color in acids and alkalis.
- Hydrochloric acid is run into the sodium hydroxide solution burette until the indicator just changes color.

2 Burning sodium

- Sodium burns vigorously in chlorine to form sodium chloride:
  $2Na + Cl_2 \rightarrow 2NaCl$
- Sodium also burns vigorously in oxygen to form sodium oxide:
  $4Na + O_2 \rightarrow 2Na_2O$

3 Sodium and nitrogen

- When heated in a stream of nitrogen, sodium reacts to form sodium nitride:
  $6Na + N_2 \rightarrow 2Na_3N$

4 Preparation of sodium

- Sodium is obtained by the *electrolysis* of molten sodium chloride in a Downs cell. Sodium is discharged at the negative electrode (*cathode*):
  $Na^+ + e^- \rightarrow Na$
  The product at the positive electrode (*anode*) is chlorine gas:
  $2Cl^- \rightarrow Cl_2 + 2e^-$

© Diagram Visual Information Ltd.

The group 2 metals

Key words

alkaline earth metals
group 2
ion
ionization energy
orbital
radioactive
reactivity

1 Position in periodic table

- The *group 2* metals occupy the second column of the periodic table. They include beryllium, magnesium, calcium, strontium, and barium. Radium, which lies below barium, is not usually considered when discussing the group because it is *radioactive*. Historically, group 2 metals were known as the *alkaline earth metals* because all but beryllium react with water to give alkaline solutions.
- The atomic radius and, therefore, the size of the atoms increases going down the group.

2 Electron-shell structure

- Scientists can describe an atom by describing its electron-shell structure (see page 150).
- All group 2 elements form *ions* by losing two outer electrons. The energy needed to do this is the sum of the first and second *ionization energies*, i.e., the energy needed to remove the first electron and the second electron.
- Going down the group, there is an increase in the number of *orbitals* of electrons. This affects the value of the ionization energy in two ways: 1) the two outer electrons are further from the positively charged nucleus, and 2) there are more layers of electrons between the nucleus and the outer electrons, which partially shields the outer electrons from the nucleus. Consequently, going down the group, less energy is needed to remove the outer two electrons, and the metals become progressively more reactive.

3 Reactivity

- As with the group 1 metals, the *reactivity* of the group 2 metals increases going down the group.
- The group 2 metals have similar chemical properties as group 1 metals; however, the reactivity of group 2 metals in the same period is less.

© Diagram Visual Information Ltd.

1 Position in the periodic table

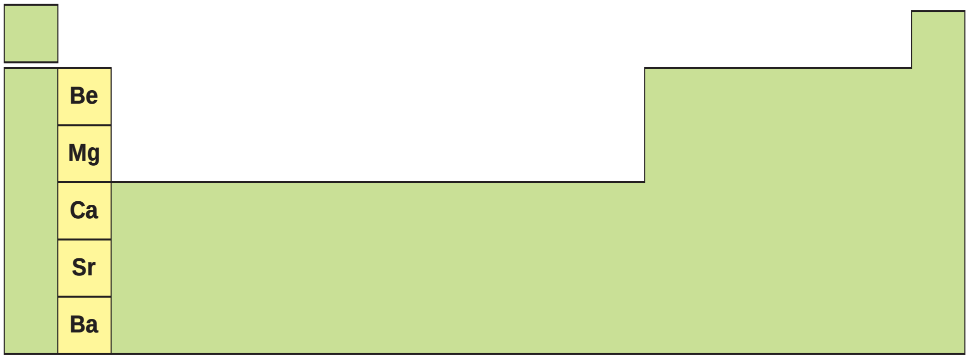

2 Electron-shell structure

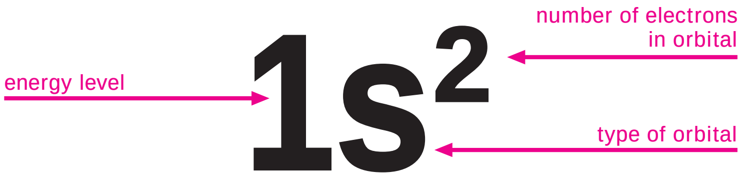

Be	$1s^2 2s^2$
Mg	$1s^2\ 2s^2\ 2p^6\ 3s^2$
Ca	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2$
Sr	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2\ 3d^{10}\ 4p^6\ 5s^2$
Ba	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2\ 3d^{10}\ 4p^6\ 5s^2\ 4d^{10}\ 5p^6\ 6s^2$

3 Reactivity comparison with group I

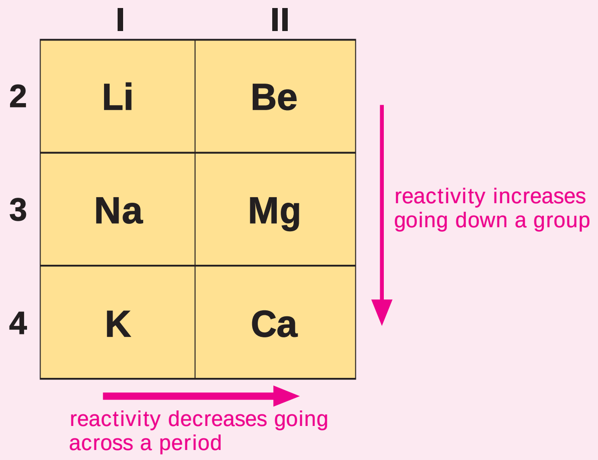
I
II
2
Li
Be
3
Na
Mg
4
K
Ca
reactivity increases going down a group
reactivity decreases going across a period

The group 2 metals: general reactions

1 Sodium and water

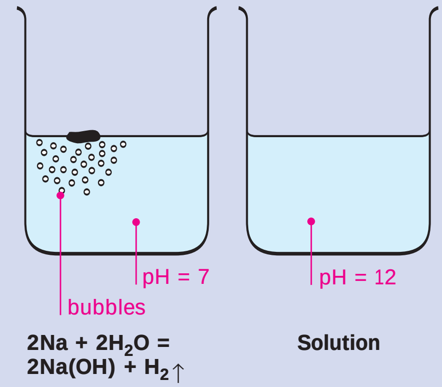
bubbles
pH = 7
pH = 12
$2Na + 2H_2O = 2Na(OH) + H_2\uparrow$
Solution

2 Calcium and water

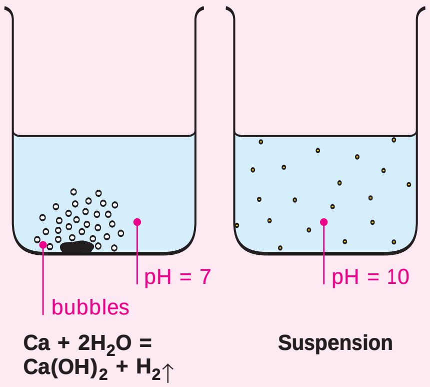
bubbles
pH = 7
pH = 10
$Ca + 2H_2O = Ca(OH)_2 + H_2\uparrow$
Suspension

3 Production of “rainwater”

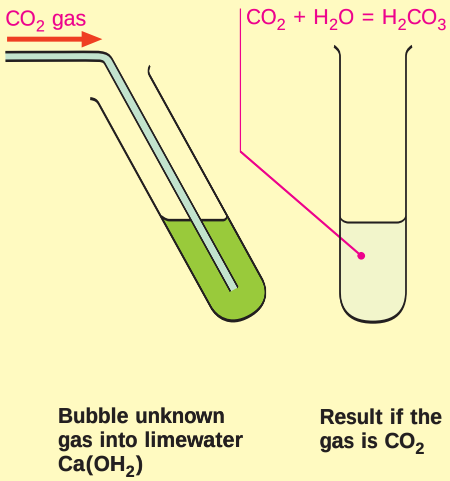
CO_2 gas
$CO_2 + H_2O = H_2CO_3$
Bubble unknown gas into limewater Ca(OH$_2$)
Result if the gas is CO_2

4 Effect of rainfall on fresh water

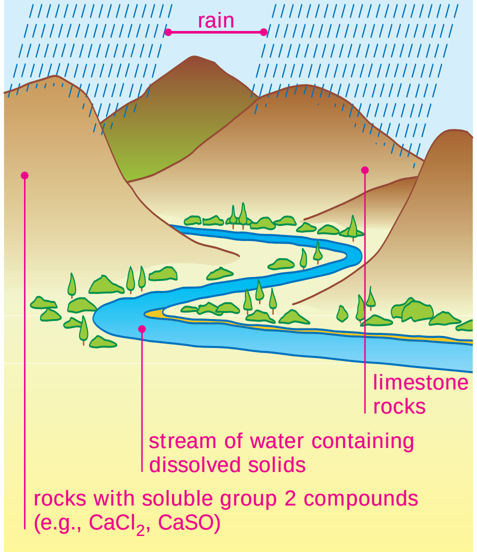
rain
limestone rocks
stream of water containing dissolved solids
rocks with soluble group 2 compounds (e.g., $CaCl_2$, CaSO)

5 Calcium hydroxide [Ca(OH)$_2$] test for carbon dioxide gas

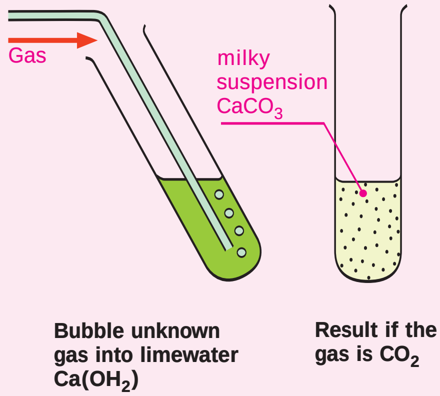
Gas
milky suspension $CaCO_3$
Bubble unknown gas into limewater Ca(OH$_2$)
Result if the gas is CO_2

6 Barium chloride ($BaCl_2$) test for metal sulfates

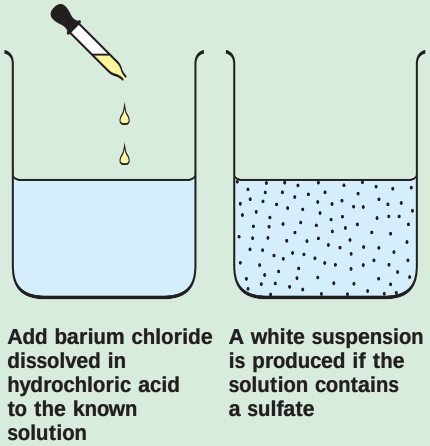
Add barium chloride dissolved in hydrochloric acid to the known solution
A white suspension is produced if the solution contains a sulfate

Key words

alkali
carbonic acid
group 2
hydrochloric acid
insoluble
limewater
soluble

1 Sodium and water

- Sodium reacts vigorously with water to form sodium hydroxide solution (a strong alkali) and hydrogen gas.

2 Calcium and water

- Calcium reacts less vigorously with water than sodium to form calcium hydroxide solution and hydrogen gas.
- Calcium hydroxide is less *soluble* than sodium hydroxide and forms a weak *alkali* solution containing suspended particles of undissolved solid.

3 Rainwater

- Naturally occurring rainwater is always weakly acidic because carbon dioxide from the air dissolves in it, forming weak *carbonic acid*, H_2CO_3.

4 Effect of rainfall

- What rain flows over rocks, *group 2* metal compounds dissolve in it, resulting in water that contains dissolved solids.
- Magnesium and calcium carbonates are effectively *insoluble* in water, but they react with rainwater, because it is acidic, to form soluble hydrogencarbonates.

5 Ca(OH)$_2$ test

- *Limewater*, an aqueous solution of calcium hydroxide, is used to test for carbon dioxide.
- When carbon dioxide is bubbled into limewater, it turns milky due to the formation of insoluble calcium carbonate, CaCo$_3$.

6 BaCl$_2$ test

- The test solution is first acidified with dilute *hydrochloric acid*, and a few drops of barium chloride solution are then added. If the solution contains sulfate ions, a white precipitate of barium sulfate is formed.

© Diagram Visual Information Ltd.

The transition metals: electron structure

Key words

actinides	*tensile strength*
catalyst	*transition metals*
orbital	
oxidation state	
shell	

Characteristics of transition metals

- The *transition metals* are any of the metallic elements with an incomplete inner electronic structure. While the outermost *shell* contains at most two electrons, their next-to-outermost shells have incompletely filled *orbitals*, which fill up going across a period. The filling is not always regular.
- The 40 transitional metals are organized into four series: The first series, shown in the table, runs from element 21 (scandium) to element 30 (zinc) and is in period 4. The second series, elements 39 (yttrium) to 48 (cadmium), is in period 5. The third, elements 71 (lanthanum) to 80 (mercury), is in period 6. The fourth series, from 103 (lawrencium) to 112 (ununbium), is the *actinides* and transactinides.
- Moving away from the nucleus, successive electron shells become progressively closer in energy. The energy levels of the third and fourth orbitals are close in the first series of transition metals.
- The electronic structure of all of the elements in period 4 can be written as that of the element argon together with additional electrons filling the 3d and 4s orbitals (see table).
- Transition metals often have colored compounds because their ions contain electrons in the 3d orbitals that can move between energy levels, giving out light.
- Transition metals tend to have high *tensile strength* (the maximum stress a material can withstand without breaking), density, and melting and boiling points. They have a variety of different *oxidation states* and are often good *catalysts*.

Table to show the electron structures of atoms and ions of elements from scandium to zinc

Element	Symbol	Electronic structure of atom	Common ion	Electronic structure of ion
Scandium	Sc	(Ar)$3d^1 4s^2$	Sc^{3+}	(Ar)
Titanium	Ti	(Ar)$3d^2 4s^2$	Ti^{4+}	(Ar)
Vanadium	V	(Ar)$3d^3 4s^2$	V^{3+}	(Ar)$3d^2$
Chromium	Cr	(Ar)$3d^5 4s^1$	Cr^{3+}	(Ar)$3d^3$
Manganese	Mn	(Ar)$3d^5 4s^2$	Mn^{2+}	(Ar)$3d^5$
Iron	Fe	(Ar)$3d^6 4s^2$	Fe^{2+}	(Ar)$3d^6$
			Fe^{3+}	(Ar)$3d^5$
Cobalt	Co	(Ar)$3d^7 4s^2$	Co^{2+}	(Ar)$3d^7$
Nickel	Ni	(Ar)$3d^8 4s^2$	Ni^{2+}	(Ar)$3d^8$
Copper	Cu	(Ar)$3d^{10} 4s^1$	Cu^+	(Ar)$3d^{10}$
			Cu^{2+}	(Ar)$3d^9$
Zinc	Zn	(Ar)$3d^{10} 4s^2$	Zn^{2+}	(Ar)$3d^{10}$

(Ar) = electron structure of argon

Note: As the shells of electrons get further and further from the nucleus successive shells become closer in energy

© Diagram Visual Information Ltd.

The transition metals: ionization energies and physical properties

1 Graphs showing the second and third ionization energies of the elements from scandium to zinc

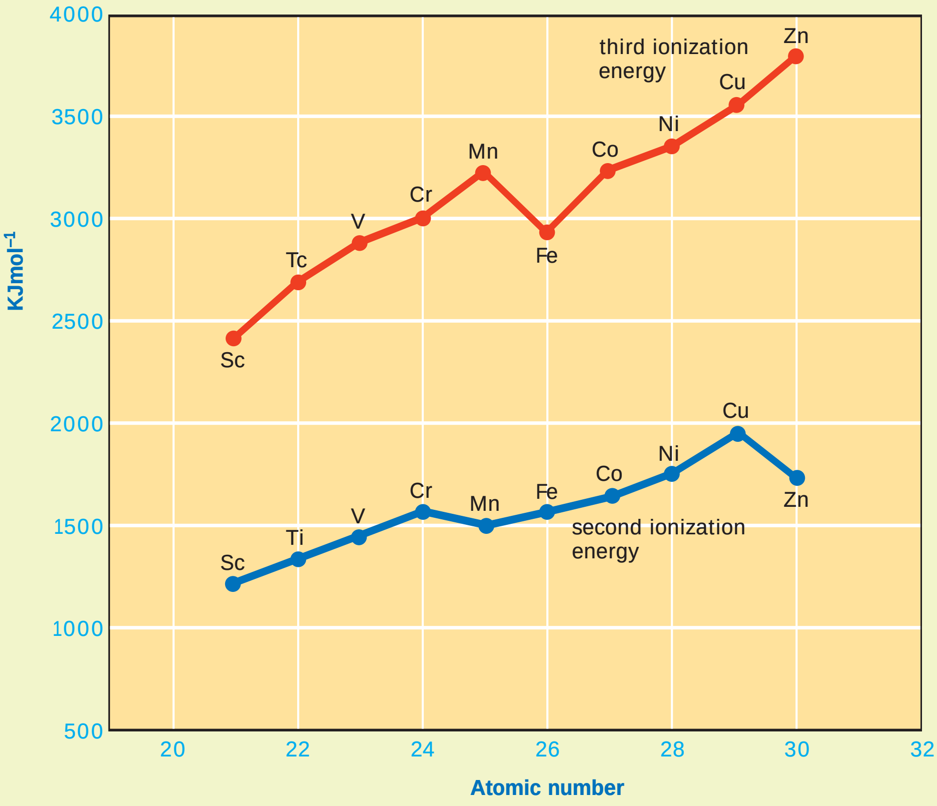

2 Physical properties of the elements from scandium to zinc

Element		Atomic radius/nm	m.p./°C	b.p./°C	Density/ gcm–3	Ionic radius/nm	
						M^{2+}	M^{3+}
Scandium	Sc	0.16	1540	2730	3.0		0.081
Titanium	Ti	0.15	1680	3260	4.5	0.090	0.076
Vanadium	V	0.14	1900	3400	6.1	0.088	0.074
Chromium	Cr	0.13	1890	2480	7.2	0.084	0.069
Manganese	Mn	0.14	1240	2100	7.4	0.080	0.066
Iron	Fe	0.13	1540	3000	7.9	0.076	0.064
Cobalt	Co	0.13	1500	2900	8.9	0.074	0.063
Nickel	Ni	0.13	1450	2730	8.9	0.072	0.062
Copper	Cu	0.13	1080	2600	8.9	0.070	
Zinc	Zn	0.13	420	910	7.1	0.074	

Key words

boiling point
conductor
ionization energy
melting point
transition metals

1 Ionization energies

- *Ionization energy* is the energy needed to remove an electron from a neutral gaseous atom or ion against the attraction of the nucleus.
- The second ionization energy is the energy needed to go from M^+ to M^{2+}, (where M = metal), and the third ionization energy is the energy needed to go from M^{2+} to M^{3+}.
- The second ionization energy increases across period 4 because there is an increasing positive charge on the nucleus of the ion, making it increasingly more difficult to remove the second electron.
- The third ionization energy for all elements is significantly higher than the second. Removal of the second electron results in a greater net difference between the positive charge on the nucleus of the ion and the negative charge surrounding it, so it requires more energy to remove a third electron.

2 Physical properties

- Like other metals, *transition metals* are good *conductors* of both heat and electricity.
- The transition metals in general have higher *melting points* and *boiling points* than groups 1 and 2 metals.
- The atomic radii and ionic radii for the M^{2+} ion decrease across period 4 because the increasing positive charge on the nucleus of the atom and of the ion provides a greater attraction for the surrounding electrons.

© Diagram Visual Information Ltd.

Aluminum

Key words

acid
alkali
aluminum
amphoteric
cryolite
electrolysis
electrolyte
filtrate
ore
precipitate

1 Extraction

- Bauxite, the *ore* from which *aluminum* is obtained, contains impurities, principally iron(III) oxide (Fe_2O_3), that must be removed before the ore can be processed to obtain aluminum.
- Aluminum oxide is an *amphoteric* oxide (it reacts with both *acids* and *alkalis*). After grinding, the ore is mixed with an excess of sodium hydroxide solution, forming sodium tetrahydroxoaluminate(III) solution.
- Iron(III) oxide and the other impurities remain undissolved in the sodium hydroxide solution and are filtered off.
- The *filtrate*, containing sodium tetrahydroxoaluminate(III), is transferred into a precipitation tank, where the solution decomposes, giving a *precipitate* of pure solid aluminum oxide.

2 Manufacture

- Aluminum oxide is reduced by *electrolysis* in a Hall-Hérault cell.
- For electrolysis to occur, the *electrolyte* must be molten so that the ions are mobile and able carry electric charge. The electrolyte consists of a solution of aluminum oxide and molten *cryolite* (a compound of aluminum fluoride and sodium fluoride).
- Aluminum oxide dissociates in the cryolite solution, giving aluminum ions, Al^{3+}, and oxide ions, O^{2-}.
- Aluminum ions are reduced to aluminum metal, which is tapped off molten from the bottom of the cell. Oxide ions are oxidized to oxygen.
- The graphite anode readily reacts with the oxygen produced to give carbon dioxide. The graphite anode is gradually eaten away and must be replaced at regular intervals.

© Diagram Visual Information Ltd.

1 Extraction of pure aluminum oxide (Al_2O_3)

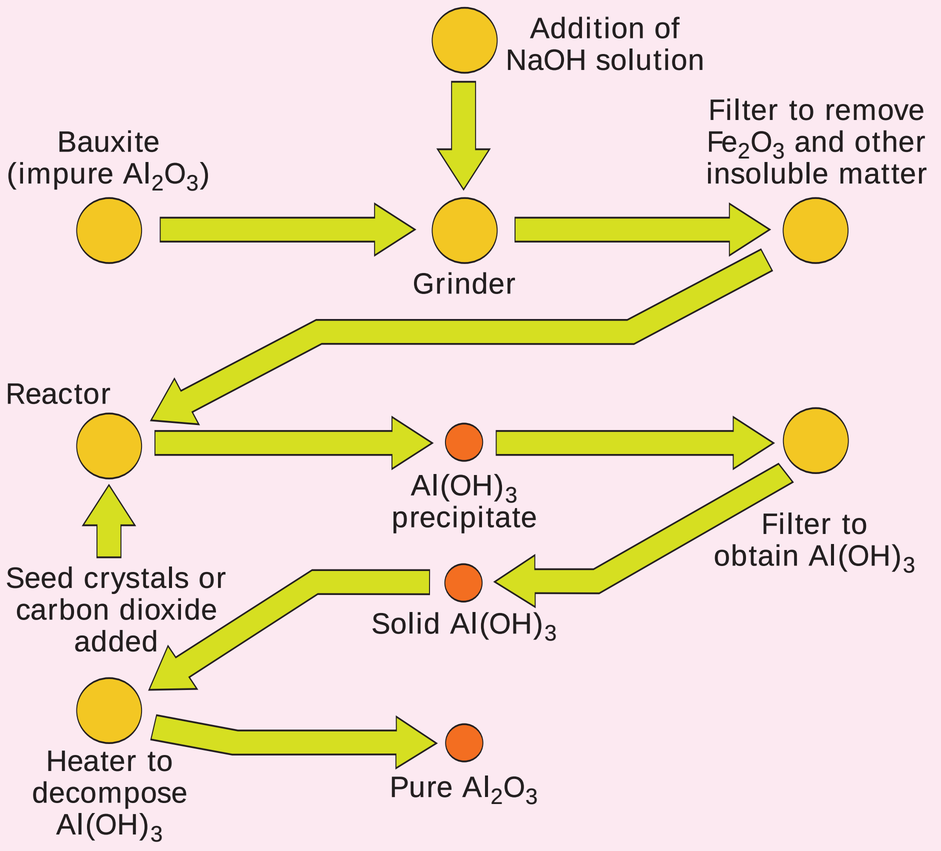
Addition of NaOH solution
Bauxite (impure Al_2O_3)
Filter to remove Fe_2O_3 and other insoluble matter
Grinder
Reactor
$Al(OH)_3$ precipitate
Filter to obtain $Al(OH)_3$
Seed crystals or carbon dioxide added
Solid $Al(OH)_3$
Heater to decompose $Al(OH)_3$
Pure Al_2O_3

2 The electrolytic manufacture of aluminum

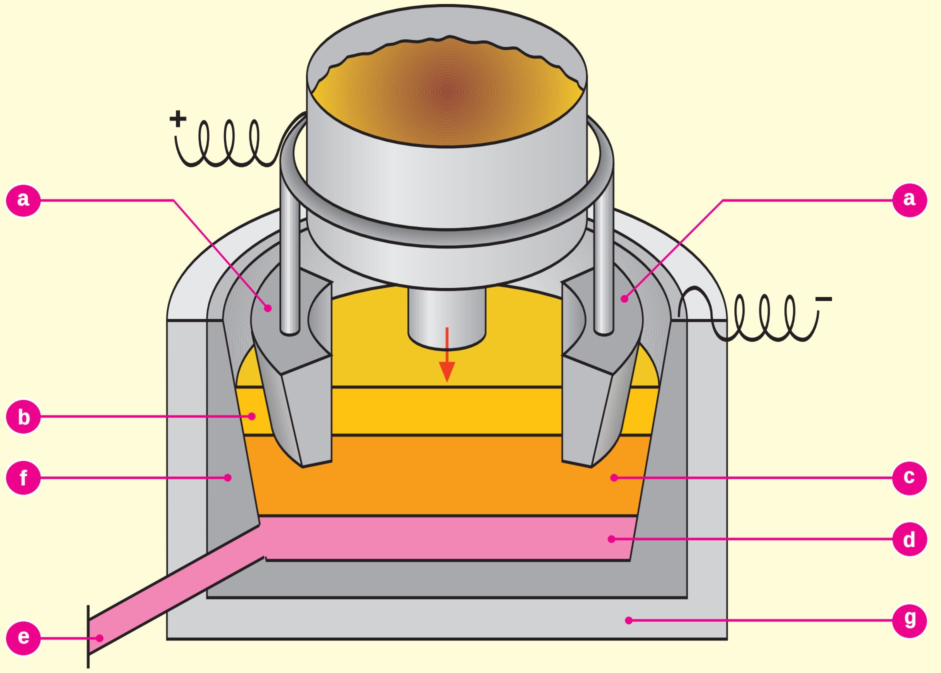

a graphite anodes
b solid crust of electrolyte
c molten electrolyte (aluminum oxide dissolved in cryolite)
d molten aluminum oxide
e tapping hole
f graphite lining to cell (cathode)
g insulation

Iron: smelting

1 The blast furnace

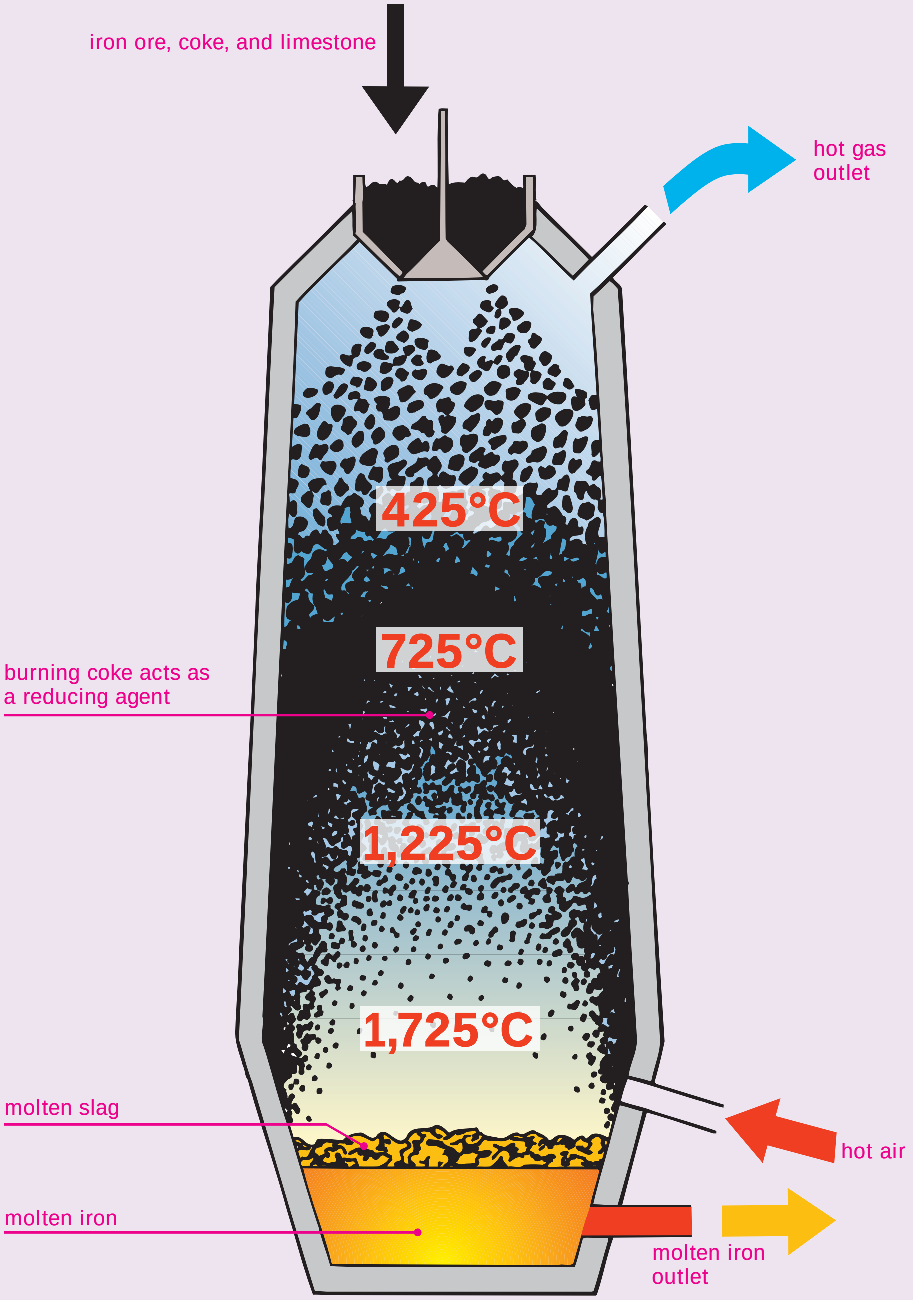

2 Table of impurities of pig iron

Impurity	% impurity in pig iron
Carbon	3 to 5
Silicon	1 to 2
Sulfur	0.05 to 0.10
Phosphorus	0.05 to 1.5
Manganese	0.5 to 1.0

Key words

flux	*smelting*
ore	
reducing agent	
reduction	
slag	

1 The blast furnace

- Iron *ores* such as hematite and magnetite contain oxygen. To create pure iron, the ores are smelted in a blast furnace to remove the oxygen.
- A charge of iron ore, limestone, and coke is fed into the top of the furnace, and hot air is blown in toward the bottom through pipes called tuyeres.
- The coke is used as a fuel, as a *reducing agent*, and also to supply carbon, which dissolves in the molten iron formed.
- The limestone acts as a *flux* (cleaning agent), combining with acidic impurities in the iron ore to form a liquid *slag* (the waste produce of *smelting*).
- Molten iron falls to the bottom of the furnace, where it is tapped.
- Molten slag floats on the molten iron and is drawn off.
- Hot gases (carbon monoxide, carbon dioxide, sulfur dioxide, nitrogen, and unreacted oxygen) are removed at the top of the furnace.
- The conversion of iron oxide to iron is a *reduction*. The main reducing agent is carbon monoxide.
- Iron oxide is reduced to iron by carbon monoxide, which itself is oxidized to carbon dioxide.
- The temperature inside the blast furnace is sufficient to decompose limestone into calcium oxide and carbon dioxide. Calcium oxide then combines with impurities such as silicon dioxide to form slag.

2 Impurities

- The iron that leaves the blast furnace (called pig iron) contains a variable amount of impurities, including carbon, silicon, sulfur, phosphorus, and manganese.

© Diagram Visual Information Ltd.

The manufacture of steel

Key words

alloy
slag

1 Basic oxygen process

- Steel is an *alloy* of iron, carbon, and other metals and non-metals.
- In the basic oxygen process, the furnace is charged with controlled amounts of steel scrap and molten iron from a blast furnace. An oxygen lance, cooled by circulating water, is lowered into the furnace, and high purity oxygen is injected into the vessel at twice the speed of sound. Impurities are readily oxidized. Molten iron is also oxidized.
- With the exception of carbon monoxide, the remaining oxides all react with calcium oxide, which is added during the oxygen blow, to form a *slag*.
- The resulting steel is highly oxidized and not suitable for casting. It is deoxidized by adding controlled amounts of aluminum and silicon in a separate reaction vessel. Additional metals and non-metals are added at this point to make different types of steel.

2 Electric arc furnace

- The electric arc furnace process uses only cold scrap metal. The furnace is a circular bath with a moveable roof through which carbon electrodes can be raised or lowered as required.
- Scrap steel is placed in the furnace, the roof closed, and the electrodes lowered into position. When a current is passed, an arc forms between the scrap steel and the electrodes, and the heat generated melts the scrap steel.
- Lime, fluorspar, and iron ore are added, and these combine with impurities forming a slag. When the steel has reached the correct composition, the slag is poured off, and the steel is tapped from the furnace.

© Diagram Visual Information Ltd.

1 Basic oxygen process

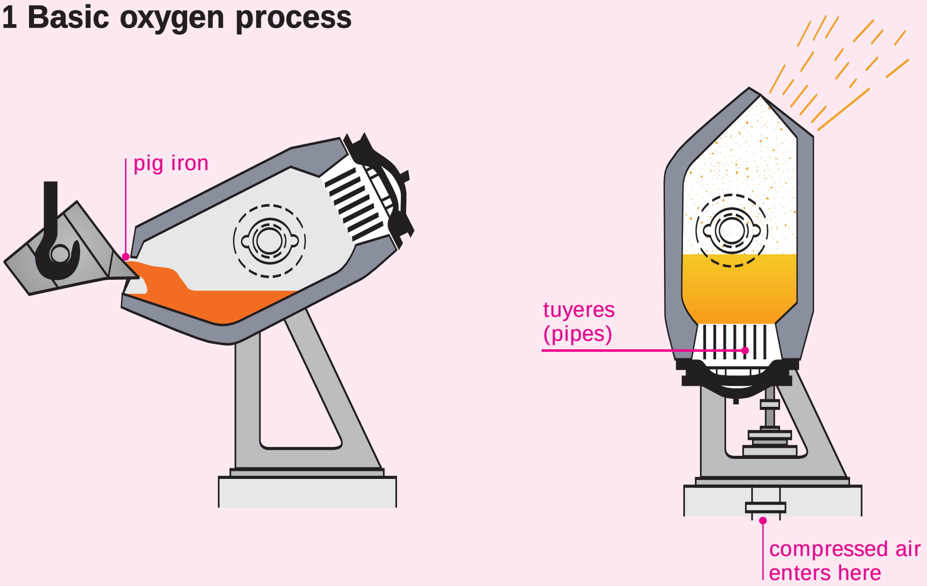
pig iron
tuyeres (pipes)
compressed air enters here

Charging the converter **Position during blowing**

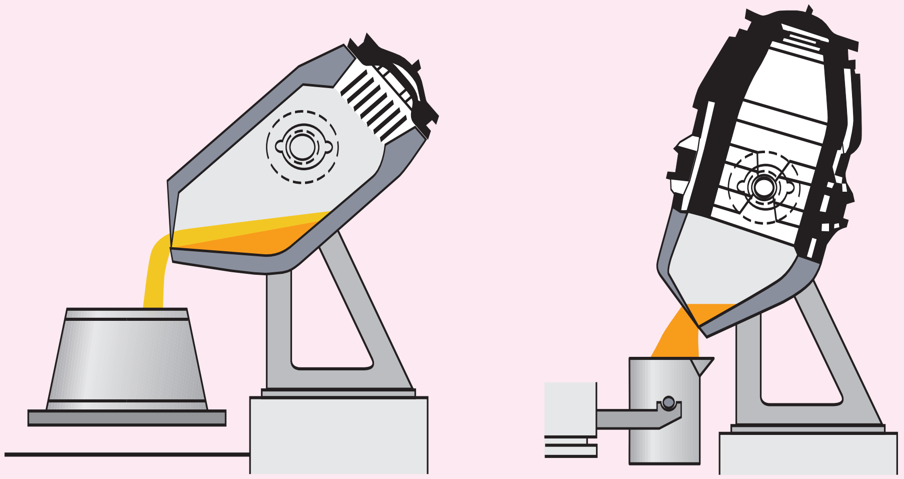

Discharging the slag **Discharging the steel**

2 Electric arc furnace

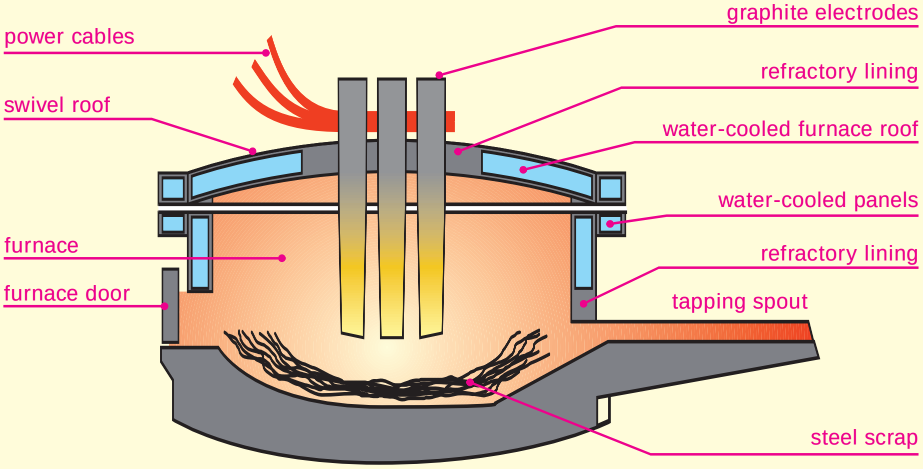
power cables
graphite electrodes
refractory lining
swivel roof
water-cooled furnace roof
water-cooled panels
furnace
refractory lining
furnace door
tapping spout
steel scrap

Rusting

1 Rust experiment

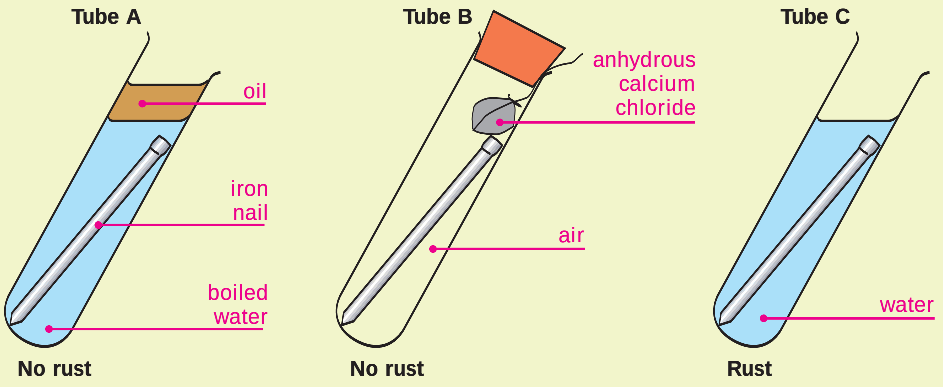
Tube A
oil
iron nail
boiled water
No rust
Tube B
anhydrous calcium chloride
air
No rust
Tube C
water
Rust

2 Chemical process

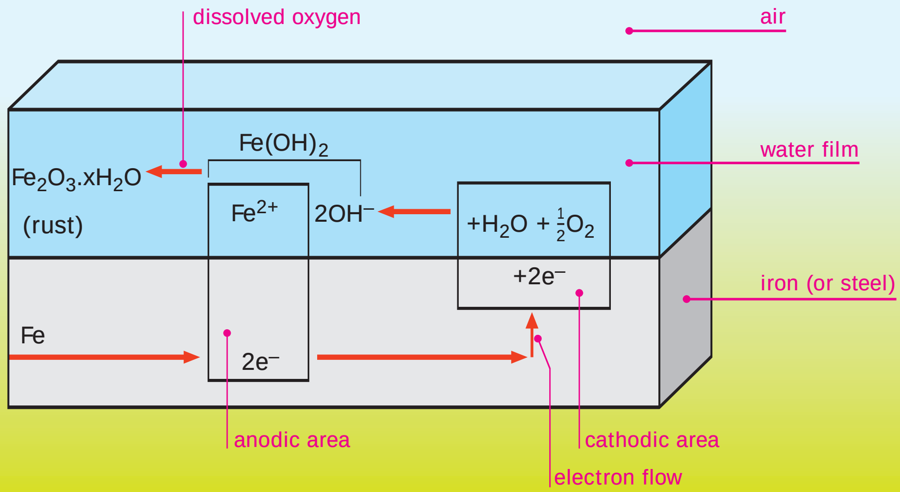
dissolved oxygen
air
$Fe(OH)_2$
$Fe_2O_3.xH_2O$
(rust)
Fe^{2+}
$2OH^-$
$+H_2O + \frac{1}{2}O_2$
$+2e^-$
water film
iron (or steel)
Fe
$2e^-$
anodic area
cathodic area
electron flow

3 Rust prevention

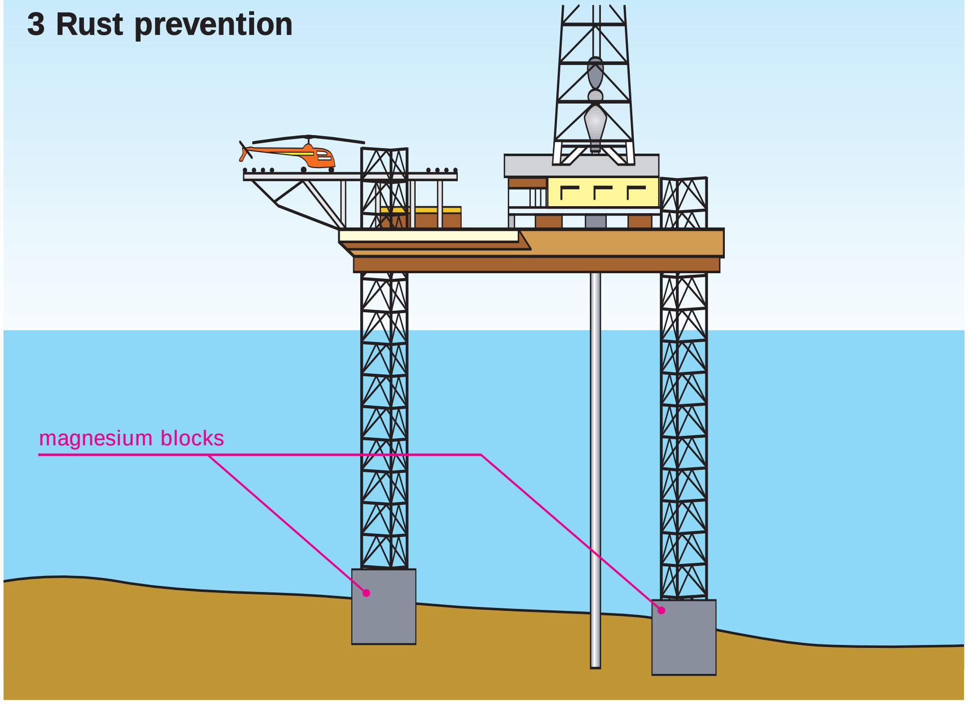
magnesium blocks

Key words

galvanizing
hydroxide ion
iron
magnesium
rust

Rusting

- Rusting is the result of a chemical cell being formed on the surface of *iron* when it is in contact with water and oxygen from the air.

1 Rust experiment

- The experiment at left proves that both water and oxygen are needed for rusting.
- Tube A: When water is boiled, the air it contains is expelled, and oil prevents any air redissolving in the water. The nail is exposed to water but not oxygen and does not *rust*.
- Tube B: Anhydrous calcium chloride removes moisture from the air. The nail is exposed to oxygen but not water and does not rust.
- Tube C: The nail is exposed to both water and oxygen, and rust forms on it.

2 Chemical process

- Iron atoms are oxidized to form first iron(II) ions, Fe^{2+}, and then iron(III) ions, Fe^{3+}, present in rust, $Fe_2O_3.xH_2O$.
- Oxygen is reduced and combined with water to form *hydroxide ions*, OH^-.

3 Rust prevention

- Most methods of rust prevention involve stopping iron or, more commonly, steel from coming into contact with water and/or oxygen in air. These methods include painting, greasing, coating in plastic, coating in zinc (*galvanizing*), and coating in tin.
- Sacrificial protection involves bolting blocks of a more reactive metal, such as *magnesium*, to a steel structure. The magnesium will oxidize more readily than the iron and will thus “sacrifice itself” in order to prevent iron from rusting.

© Diagram Visual Information Ltd.

Copper smelting and converting

Key words

anode
cathode
electrolyte
slag

1 Matte smelting

- Matte smelting is used to produce a liquid sulfide phase (matte) containing as much copper as possible, and an immiscible liquid *slag*, which contains virtually no copper.
- Copper sulfide ores, such as chalcopyrite ($CuFeS_2$) are mixed with sand and blown into the flash furnace:

$$4CuFeS_2(s) + 5O_2(g) + 2SiO_2(s) \rightarrow \underbrace{2Cu_2S.FeS(l)}_{\text{matte}} + \underbrace{2FeSiO_3(l)}_{\text{slag}} + 4SO_2(g)$$

- As the iron content of the matte falls to about 1 percent, copper starts to form. This product is called "blister copper" and is 98–99.5 percent pure. It is porous and brittle and requires further refining to be commercially useful.
- Blister copper is melted to drive off sulfur dioxide, and air is blown through it to remove any sulfur. The impure copper is cast into *anodes* for electro-refining.

2 Electro-refining

- In electro-refining, a large impure copper anode and a small pure copper *cathode* are suspended in an *electrolyte* consisting of copper(II) sulfate solution and sulfuric acid.
- At the anode, copper atoms are oxidized to copper ions and pass into solution. The anode gradually becomes smaller:

$$Cu(s) \rightarrow Cu^{2+}(aq) + 2e^-$$

- At the cathode, copper ions are removed from solution as they are reduced to copper atoms. The cathode gradually becomes larger:

$$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$$

- Impurities that are insoluble in the electrolyte fall to the bottom of the cell. These may include gold, silver, platinum, and tin, and in some circumstances may be more valuable than the copper produced.

© Diagram Visual Information Ltd.

1 Matte smelting

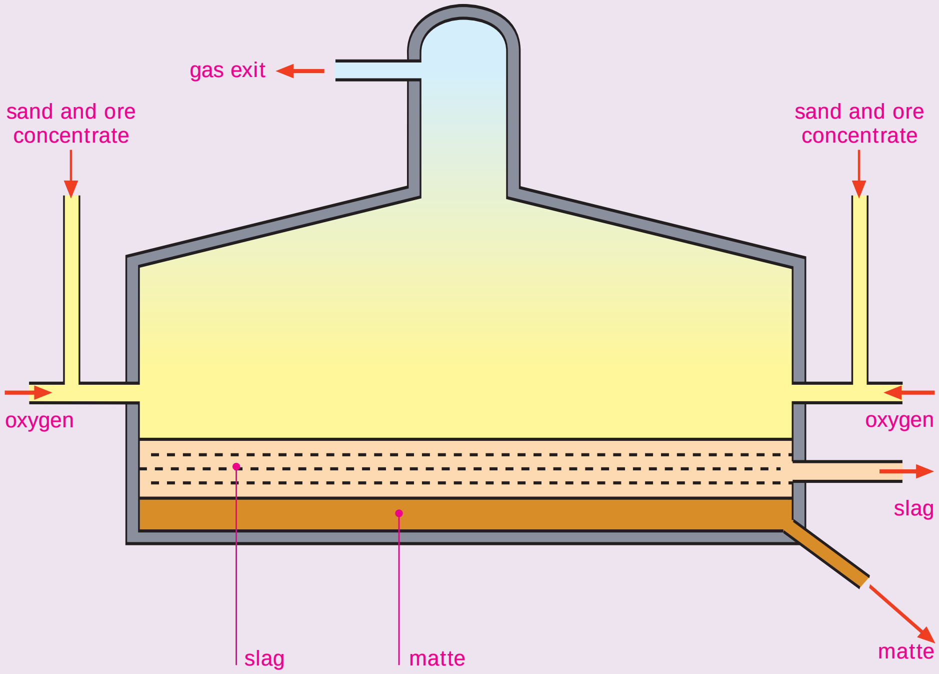
gas exit
sand and ore concentrate
sand and ore concentrate
oxygen
oxygen
slag
slag
matte
matte

2 Electro-refining

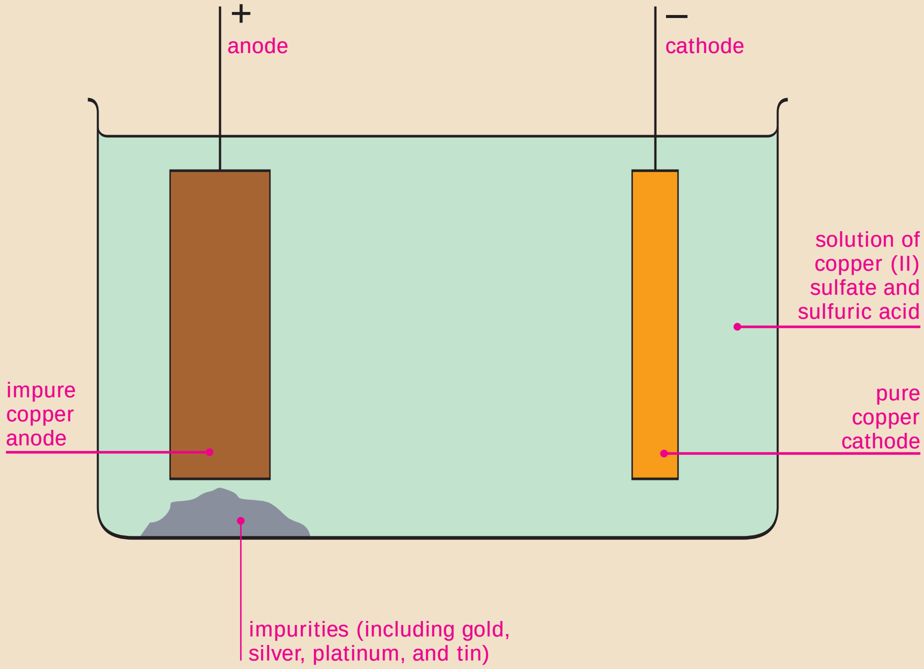
+
anode
–
cathode
solution of copper (II) sulfate and sulfuric acid
impure copper anode
pure copper cathode
impurities (including gold, silver, platinum, and tin)

Reactions of copper

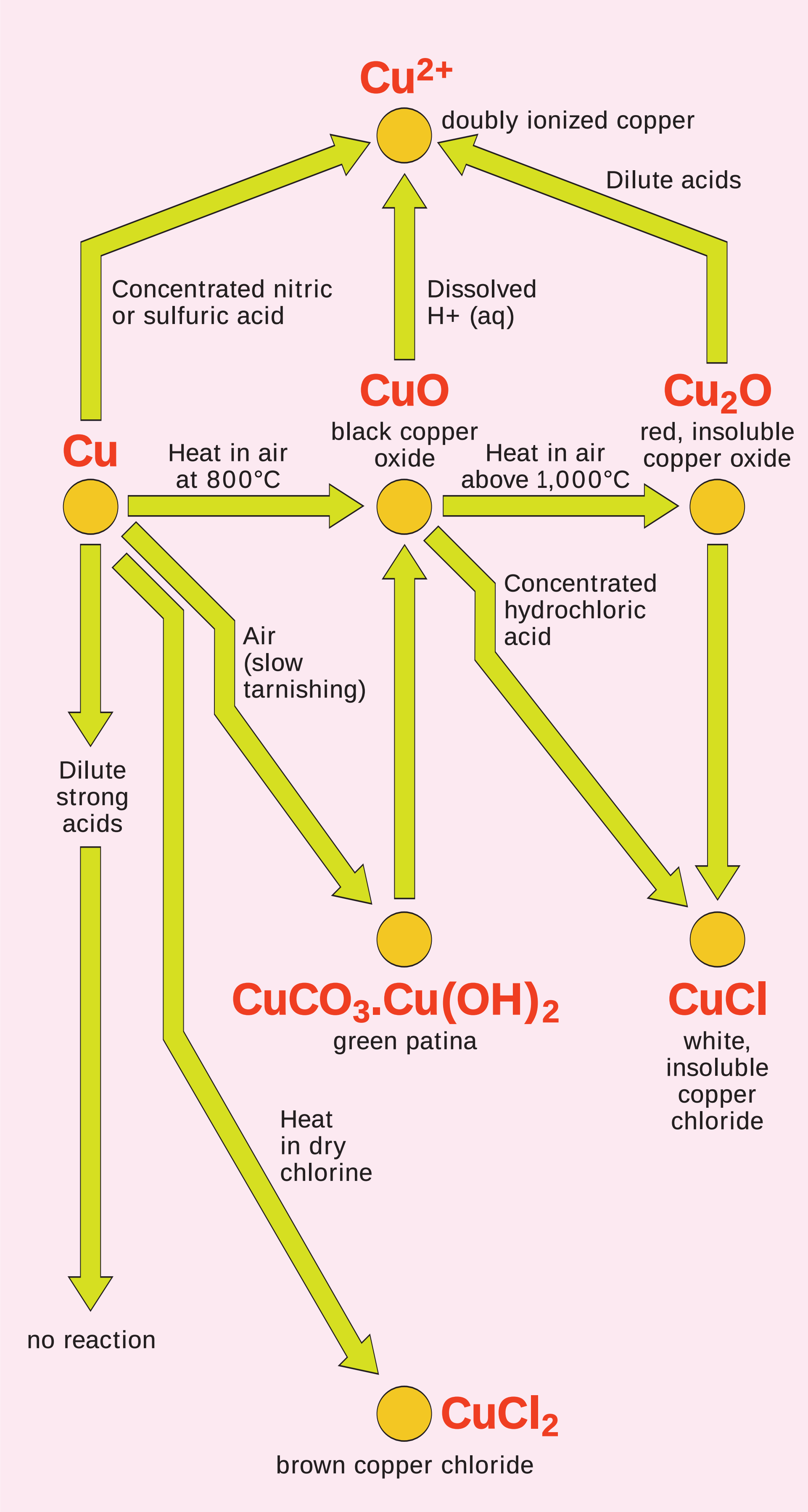
Cu2+
doubly ionized copper
Dilute acids
Concentrated nitric or sulfuric acid
Dissolved H+ (aq)
CuO
black copper oxide
Cu2O
red, insoluble copper oxide
Cu
Heat in air at 800°C
Heat in air above 1,000°C
Concentrated hydrochloric acid
Air (slow tarnishing)
Dilute strong acids
no reaction
CuCO3.Cu(OH)2
green patina
CuCl
white, insoluble copper chloride
Heat in dry chlorine
CuCl2
brown copper chloride

Key words

hydrochloric acid
nitric acid
oxidation state
oxidizing agent
sulfuric acid
transition metals

Reactions of copper

- Copper is a *transition metal*. Its normal *oxidation state* is copper(II), Cu^{2+}, but it also forms some copper(I), Cu^{+}, compounds. Copper is a relatively unreactive metal. It does not react with dilute strong acids, water, or steam.
- When heated in air at 800°C, copper is oxidized to black copper(II) oxide:
  $2Cu(s) + O_2(g) \rightarrow 2CuO(s)$
  At temperatures over 1,000°C, red copper(I) oxide is formed:
  $4Cu(s) + O_2(g) \rightarrow 2Cu_2O(s)$
  Both oxides react with dilute acids to form copper(II) salts.
- When heated in chlorine, copper forms brown copper(II) chloride.
  $Cu(s) + Cl_2(g) \rightarrow CuCl_2(s)$
- White copper(I) chloride also exists and can be made by strongly heating copper(II) chloride:
  $2CuCl_2(s) \rightarrow 2CuCl(s) + Cl_2(g)$
  It is also formed by the reaction of copper(II) oxide with concentrated *hydrochloric acid* via an complex ion, $[CuCl_2]^-$. When a solution containing this ion is poured into water, copper(I) chloride is precipitated.
- Copper tarnishes slowly in air, forming basic copper(II) carbonate, a compound of copper(II) carbonate, and copper(II) hydroxide, $CuCO_3.Cu(OH)_2$. It is this compound that produces the green coloration, referred to as patina, on weathered copper.
- Copper reacts with both concentrated nitric and concentrated *sulfuric acid*. Both of these concentrated acids are powerful *oxidizing agents* and react with copper in a different way than a dilute acid reacts with a metal. Copper does not react with dilute acids.
  With concentrated sulfuric acid:
  $Cu(s) + 2H_2SO_4(l) \rightarrow$
  $CuSO_4(aq) + 2H_2O(l) + SO_2(g)$
  With concentrated *nitric acid*:
  $Cu(s) + 4HNO_3 \rightarrow$
  $Cu(NO_3)_2(aq) + 2H_2O(l) + 2NO_2(g)$

© Diagram Visual Information Ltd.

Reaction summary: aluminum, iron, and copper

Key words

- *aluminum*
- *amphoteric*
- *carbonate*
- *copper*
- *hydroxide*
- *iron*
- *oxidation state*
- *oxide*
- *transition metals*
- *valency*

Reactivity

- *Aluminum* is the most reactive and *copper* is least reactive of the three metals.
- All three metals react directly with non-metals.

Oxides

- Aluminum has one *oxide*, Al_2O_3, which is *amphoteric* and thus reacts with both acids and alkalis. *Iron* has three oxides: FeO, Fe_2O_3, and Fe_3O_4. Copper has two: Cu_2O and CuO. All metal oxides react with dilute acids to form salts and water.

Hydroxides

- Aluminum *hydroxide*, like aluminum oxide, is amphoteric. Iron forms two hydroxides by the addition of sodium hydroxide solution to solutions of its salts. Iron(II) salts produce a dirty green precipitate of iron(II) hydroxide, while iron(III) salts produce a red-brown precipitate of iron(III) hydroxide. Copper(II) hydroxide forms as a blue precipitate when sodium hydroxide is added to a solution of a copper salt.
- All metal hydroxides react with alkalis to give metal salts and water.

Carbonates

- Aluminum and iron(III) do not form *carbonates*. Iron(II) carbonate and copper(II) carbonate decompose on heating to the corresponding metal oxide with the loss of carbon dioxide gas. The carbonates also react with dilute acids to forms metal salts, carbon dioxide, and water.

Valency

- Aluminum is in group 3 of the periodic table and exhibits only one *oxidation state*, +3, in its compounds. Iron and copper are both *transition metals* and exhibit two oxidation states in their compounds.

© Diagram Visual Information Ltd.

Aluminum	
Preparation	Electrolysis of aluminum oxide $Al^{3+} + 3e^- \rightarrow Al$ at cathode
Reaction of elements	$4Al + 3O_2 \rightarrow 2Al_2O_3$ oxide layer formed $2Al + 3Cl_2 \rightarrow Al_2Cl_6$ $2Al + 3H_2SO_4 \rightarrow Al_2(SO_4)_3 + 3H_2$
Oxide	$Al_2O_3 + 3H_2SO_4 \rightarrow Al_2(SO_4)_3 + 3H_2O$ $Al_2O_3(s) + 2NaOH(aq) \rightarrow Na[Al(OH)_4](aq)$
Hydroxide	$AlCl_3 + 3NaOH \rightarrow Al(OH)_3 + 3NaCl$ $Al(OH)_3 + 3HCl \rightarrow AlCl_3 + 3H_2O$ $Al(OH)_3(s) + NaOH(aq) \rightarrow 2Na[Al(OH)_4](aq)$ (amphoteric)
Carbonate	Not formed
Change of valency	Only on oxidation

Iron	
Preparation	Chemical reduction in blast furnace $Fe_2O_3 + 3CO \rightarrow 2Fe + 3CO_2$
Reaction of elements	$2Fe + 2H_2O + O_2 \rightarrow 2Fe(OH)_2$ rust $Fe + 2HCl \rightarrow FeCl_2 + H_2$ $2Fe + 3Cl_2 \rightarrow 2FeCl_3$ $Fe + S \rightarrow FeS$ $Fe + H_2SO_4 \rightarrow FeSO_4 + H_2$
Oxide	$FeO + H_2SO_4 \rightarrow FeSO_4 + H_2O$ $Fe_2O_3 + 3H_2SO_4 \rightarrow Fe_2(SO_4)_3 + 3H_2O$
Hydroxide	$FeCl_2 + 2NaOH \rightarrow Fe(OH)_2 + 2NaCl$ $FeCl_3 + 3NaOH \rightarrow Fe(OH)_3 + 3NaCl$ $Fe(OH)_2 + 2HCl \rightarrow FeCl_2 + 2H_2O$ $Fe(OH)_3 + 3HCl \rightarrow FeCl_3 + 3H_2O$
Carbonate	Unstable to heat. $FeCO_3 \rightarrow FeO + CO_2$ $FeCO_3 + H_2SO_4 \rightarrow FeSO_4 + CO_2 + H_2O$
Change of valency	$2Fe^{2+}_{(aq)} + Cl_{2(g)} \rightarrow 2Fe^{3+}_{(aq)} + 2Cl^-_{(aq)}$

Copper	
Preparation	Thermal decomposition in furnace $Cu_2S \xrightarrow{air} 2Cu + SO_2$
Reaction of elements	$2Cu + O_2 \rightarrow 2CuO$ $Cu + Cl_2 \rightarrow CuCl_2$ $2Cu + S \rightarrow Cu_2S$ $CuCl_2 + H_2SO_4 \rightarrow$ no reaction with dilute acid $Cu + 2H_2SO_4 \rightarrow CuSO_4 + SO_2 + 2H_2O$ with conc. acid
Oxide	$CuO + H_2SO_4 \rightarrow CuSO_4 + H_2O$
Hydroxide	$CuCl_2 + 2NaOH \rightarrow Cu(OH)_2 + 2NaCl$ $Cu(OH)_2 + 2HCl \rightarrow CuCl_2 + 2H_2O$
Carbonate	Unstable to heat. $CuCO_3 \rightarrow CuO + CO_2$ $CuCO_3 + H_2SO_4 \rightarrow CuSO_4 + CO_2 + H_2O$
Change of valency	$2Cu^{2+}_{(aq)} + 2I^-_{(aq)} \rightarrow 2Cu^+_{(aq)} + I_{2(s)}$ then $Cu^+_{(aq)} + I^-_{(aq)} \rightarrow CuI_{(s)}$

The extraction of metals from their ores

Metal (Date of discovery) Ranked from highest to lowest in reactivity series	Main ore from which it is obtained	Main method of extraction
Sodium (1807) Group 1	Rock salt NaCl	Electrolysis of molten NaCl
Magnesium (1808) Group 2	Magnesite $MgCO_3$ and Mg^{2+} ions in seawater	Electrolysis of molten $MgCl_2$
Aluminum (1827) Group 3	Bauxite $Al_2O_3.2H_2O$	Electrolysis of Al_2O_3 in molten cryolite (Na_3AlF_6)
Zinc (1746) Transition metal	Zinc blende ZnS	Heat sulfide in air → oxide. Dissolve oxide in H_2SO_4, electrolyze
Iron (ancient) Transition metal	Hematite Fe_2O_3	Reduce Fe_2O_3 with carbon monoxide
Tin (ancient) Group 4	Tinstone SnO_2	Reduce SnO_2 with carbon
Lead (ancient) Group 4	Galena PbS	Heat sulfide in air → oxide. Reduce oxide with carbon
Copper (ancient) Transition metal	Copper pyrites $CuFeS_2$ (CuS + FeS)	Controlled heating with correct amount of air → Cu + SO_2
Mercury (ancient) Transition metal	Cinnabar HgS	Heat in air → Hg + SO_2

Key words

electrolysis
ore
reactivity series
reduction

Extraction of metals

- The ease with which a metal is obtained from its *ore* is directly related to its position in the *reactivity series* of metals.

Electrolytic reduction

- All of the group 1 and group 2 metals and aluminum from group 3 are reactive metals and in the upper half of the reactivity series. They cannot be obtained from their ores by chemical *reduction*, i.e., by heating the ore with a reducing agent such as carbon monoxide or carbon. These metals can only be obtained by electrolytic reduction or *electrolysis*.
- Consequently, it was impossible to obtain these metals before the discovery and development of electricity at the end of the eighteenth century. All of these metals were first made in the early years of the nineteenth century, several by English chemist Sir Humphrey Davy.

Heating

- Zinc oxide and iron oxide are reduced by heating with carbon monoxide. Although zinc can be obtained by chemical reduction, approximately 80 percent of the world's annual production is, in fact, obtained by electrolysis.
- All of the metals from iron and below in the reactivity series are relatively easy to obtain from their ores by heating.
- Iron is obtained by reduction with carbon monoxide
- Tin is obtained by reduction with carbon
- Lead is obtained by heating lead sulfide in air to produce an oxide, which is then reduced with carbon.
- Copper is obtained by controlled heating with the correct amount of air
- Mercury is obtained by heating in air.

© Diagram Visual Information Ltd.

Reactivity summary: metals

Key words

oxide
reactivity
reactivity series

Reactivity summary

- Metals can be arranged in order of their *reactivity*, starting with the most reactive. This is called the *reactivity series*. The relative reactivity of metals is reflected through all of their chemistry.

Reaction with oxygen

- Metals at the top of the reactivity series readily burn in oxygen. Less reactive metals do not burn but form a surface layer of *oxide*. Metals at the bottom of the reactivity series are not oxidized by atmospheric oxygen.

Reaction with cold water

- Metals at the top of the reactivity series react readily with cold water but with decreasing vigor down to magnesium. The metals below magnesium do not react with cold water.

Reaction with steam

- Metals react more vigorously with steam than with cold water. All of the metals down to iron react with steam with decreasing vigor. The metals below iron do not react with steam.

Reaction with dilute acid

- All of the metals down to lead react with dilute acids, with decreasing vigor. The metals below lead do not react with dilute acids.

	K Na	Ca Mg Al Zn Fe	Pb Cu Hg	Ag Pt Au
Reaction with $O_{2(g)}$ on heating	form oxides (e.g.,Na_2O) in limited supplies of O_2, but peroxides (e.g., Na_2O_2) with excess O_2	burn with decreasing vigor to form oxides	do *not* burn, but only form a surface layer of oxide	do *not* burn or oxidize on surface
Heat evolved when metal reacts with 1 mole of O_2 to form oxide shown / kJ	**K** K_2O 723 **Na** Na_2O 832	**Ca** CaO 1272 **Mg** MgO 1204 **Al** Al_2O_3 1114 **Zn** ZnO 697 **Fe** Fe_2O_3 548	**Pb** PbO 436 **Cu** CuO 311 **Hg** HgO 182	**Ag** Ag_2O 61 **Pt** — — Au Au_2O_3 54
Reaction with cold water	displace $H_{2(g)}$ from cold water with decreasing reactivity **K**, violently	displace $H_{2(g)}$ from cold water with decreasing reactivity **Mg**, very slowly	do *not* displace $H_{2(g)}$ from cold water	do *not* displace $H_{2(g)}$ from cold water
Reaction with steam	displace $H_{2(g)}$ from steam with decreasing vigor **K**, very violently	displace $H_{2(g)}$ from steam with decreasing vigor **Fe**, very slowly)	do *not* displace $H_{2(g)}$ from steam	do *not* displace $H_{2(g)}$ from steam
Reaction with dilute acid	displace $H_{2(g)}$ from dilute acid with decreasing vigor **K**, explosively	**Mg**,very vigorous **Fe**,steadily	do *not* displace $H_{2(g)}$ from dilute acid **Pb**, very slowly	do *not* displace $H_{2(g)}$ from dilute acid

© Diagram Visual Information Ltd.

Tests on metals: flame test

1 Flame test

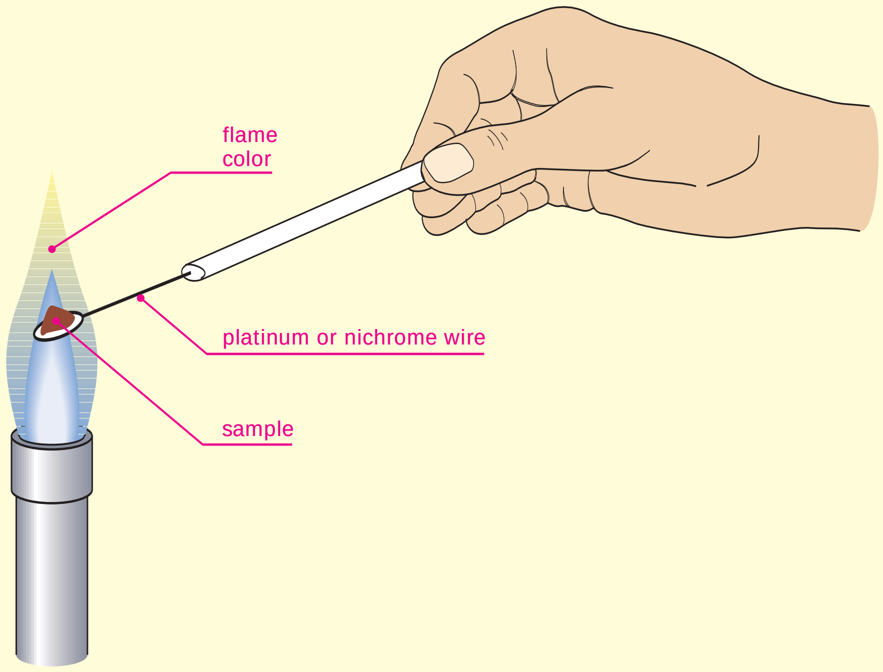

2 Table of flame coloration

Color of flame	Likely ion present	Metal
Apple green	Ba^{2+}	barium
Blue-green	Cu^{2+}	copper
Brick red	Ca^{2+}	calcium
Crimson	Sr^{2+}	strontium
Lilac	K^{+}	potassium
Orange-yellow	Na^{+}	sodium
Red	Li^{+}	lithium

Key words

ion
salt
solution

1 Flame test

- Several metal *ions* produce characteristic colors when introduced to a bunsen flame either as a solid or as a *solution* of a *salt*.
- A clean platinum or nichrome wire is dipped in concentrated hydrochloric acid and then into the solid or solution.
- The sample is introduced to the middle of a non-luminous bunsen flame.

2 Flame coloration

- The following metals produce the following colors in the flame test:
  barium: apple green
  calcium: brick red
  copper: blue-green
  lithium: red
  potassium: lilac
  sodium: orange-yellow
  strontium: crimson
- The lilac color of potassium is sometimes difficult to see and is better observed through blue glass that makes the flame appear purple.
- The orange-yellow color of sodium is very intense and may mask the color of other metal ions present.

© Diagram Visual Information Ltd.

Tests on metals: metal hydroxides

Key words

amphoteric
hydroxide
precipitate
salt
sodium hydroxide
transition metals

1 Producing the hydroxide

- Group 1 metal *hydroxides* are very soluble and form strong alkaline solutions. Group 2 metal hydroxides are less soluble but still dissolve sufficiently to form weak alkaline solutions. All other metals form insoluble hydroxides. If several drops of sodium hydroxide ($NaOH$) solution are added to a solution of a metal *salt*, a *precipitate*, often gelatinous, is formed. Care must be taken when carrying out this reaction because some metals form precipitates that redissolve in excess *sodium hydroxide* solution. If sodium hydroxide solution is added too quickly, the initial precipitate may not be seen.

2 The reactions

- The reactions of metal salt solutions with sodium hydroxide solution can be used to identify the metal.
- Aluminum, zinc, and lead hydroxides are all *amphoteric*. When sodium hydroxide solution is added to solutions of salts of these metals, an initial white precipitate is formed. However, if excess sodium hydroxide solution is added, the precipitate dissolves, forming a solution of a soluble complex compound.

  $Al(OH)_3(s) + NaOH(aq) \rightarrow Na[Al(OH)_4](aq)$
  sodium tetrahydroxoaluminate(III)

  $Zn(OH)_2(s) + 2NaOH(aq) \rightarrow Na_2[Zn(OH)_4](aq)$
  sodium tetrahydroxozincate(II)

  $Pb(OH)_2(s) + 2NaOH(aq) \rightarrow Na_2[Pb(OH)_4](aq)$
  sodium tetrahydroxoplumbate(II).
- Iron and copper are *transition metals* and form characteristic colored precipitates with sodium hydroxide solution.

© Diagram Visual Information Ltd.

1 Producing the hydroxide from the metallic salt

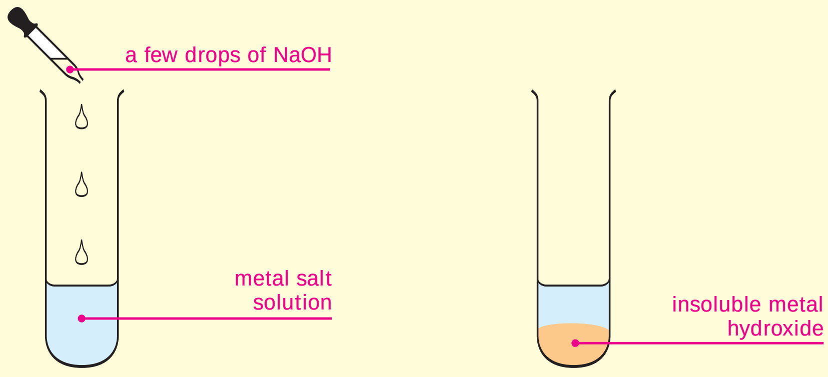

Add a small amount of NaOH to metal salt solution

A jelly-like solid forms

2 The reactions

Aluminum nitrate → white precipitate of aluminum hydroxide
$Al(NO_3)_3 + 3NaOH \rightarrow 3NaNO_3 + Al(OH)_3 \downarrow$

Zinc nitrate → white precipitate of zinc hydroxide
$Zn(NO_3)_2 + 2NaOH \rightarrow 2NaNO_3 + Zn(OH)_2 \downarrow$

Lead nitrate → white precipitate of lead hydroxide
$Pb(NO_3)_2 + 2NaOH \rightarrow 2NaNO_3 + Pb(OH)_2 \downarrow$

Iron(II) nitrate → green precipitate of iron(II) hydroxide
$Fe(NO_3)_2 + 2NaOH \rightarrow 2NaNO_3 + Fe(OH)_2 \downarrow$

Iron(III) nitrate → rust-brown precipitate of iron(III) hydroxide
$Fe(NO_3)_3 + 3NaOH \rightarrow 3NaNO_3 + Fe(OH)_3 \downarrow$

Copper nitrate → royal blue precipitate of copper hydroxide
$Cu(NO_3)_2 + 2NaOH \rightarrow 2NaNO_3 + Cu(OH)_2 \downarrow$

Tests on metals: metal ions

Key words

hydroxide
reagent

Reacations with reagents

- In addition to flame tests and the properties of their hydroxides, the presence of some metal ions in solution can be demonstrated by their reactions with particular reagents.

Metal ion in solution	To the test solution	Positive result
Alumnum Al^{3+}	Add 1 or 2 drops of litmus solution followed by dilute hydrochloric acid until the mixture is just acidic. Then add ammonia solution until just alkaline.	Blue lake – a gelatinous precipitate of aluminum hydroxide – is formed, and this absorbs the litmus, leaving the solution almost colorless.
Barium Ba^{2+}	Add several drops of potassium chromate solution.	A yellow precipitate of barium chromate. Lead ions also give a yellow precipitate, but lead chromate is deeper yellow and turns orange on heating.
Copper Cu^{2+}	Add ammonia solution drop by drop until it is in excess.	An initial blue precipitate of copper(II) hydroxide that dissolves in excess ammonia solution to give a deep blue solution containing the complex ion $[Cu(NH_3)_4]^{2+}$.
Iron(II) Fe^{2+}	Add several drops of potassium hexacyanoferrate(III) solution.	A deep blue solution is formed.
Iron(III) Fe^{3+}	Add several drops of ammonium thiocyanate solution.	Deep blood-red coloration.
Lead Pb^{2+}	Add several drops of potassium iodide solution.	A yellow precipitate of lead(II) iodide.
Silver Ag^{+}	Add several drops of potassium chromate solution.	A brick-red precipitate of silver chromate.
Zinc Zn^{2+}	Add ammonium chloride and ammonia solution, then pass hydrogen sulfide through the mixture.	A white, or more often dirty white, precipitate of zinc sulfide.

© Diagram Visual Information Ltd.

Key words

alloy
aluminum
copper
ductile
galvanizing
iron
lead
reactivity series
zinc

Uses of metals

- The uses of metals are related to both their physical and chemical properties. The physical properties of a metal are sometimes altered by mixing it with other metals or non-metals to form *alloys*.

Aluminum

- *Aluminum* has a low density but is too soft for many applications. It is frequently used as duralumin (an alloy of aluminum and *copper*) as a structural material in the manufacture of airplanes.

Zinc

- *Zinc* is above *iron* in the *reactivity series*. During *galvanizing*, iron is dipped in molten zinc, and the layer of zinc formed on the iron protects it from rusting. If the galvanized iron is scratched, exposing the iron, an electrolytic cell forms between the iron and zinc, and the zinc corrodes in preference to the iron.

Iron

- Iron is used for all sorts of structures, most often as steel (an alloy of iron and carbon). The one serious problem with iron and steel is that they rust on exposure to water and oxygen in the air.

Lead

- *Lead* has a high density and is impervious to water, so it used as flashing on roofs. It is also used as in the manufacture of car batteries. In the past, before its toxic nature was understood, lead was also used for water pipes and in paints. Solder (an alloy of lead and tin) is widely used to join copper wires and copper pipes.

© Diagram Visual Information Ltd.

Uses of metals

Metal	Use	Reason
Aluminum 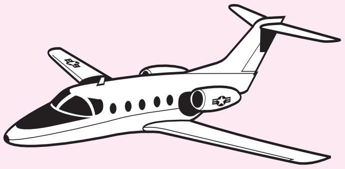	Structural material for ships, planes, cars, cookware Electric cables	Strong but light; oxide layer prevents corrosion. Light, but good conductor.
Zinc 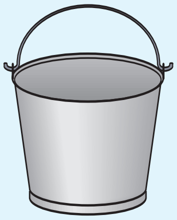	Coating (galvanizing) steel Alloys: brass (Zn/Cu) bronze (Zn/Sn/Cu)	Reactive — gives sacrificial protection to iron; does not corrode easily. Modifies the properties of the other elements.
Iron 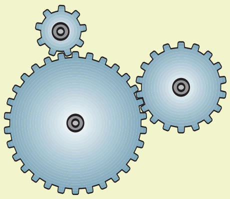	Structural material for all industries (in the form of steel)	Strong and cheap; properties can be made suitable by alloying.
Lead 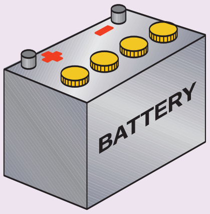	Roofing Car batteries Solder (Pb/Sn alloy)	Very malleable and does not corrode. Design of battery makes recharging possible. Low melting point.
Copper 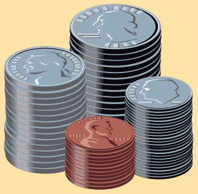	Electric cables Pipes Alloys (see above) Coins (alloyed with nickel)	Very good conductor. Very ductile, does not corrode easily.

Copper

- Copper is very *ductile* and can be easily drawn into wires. It is a good conductor of electricity and is used for the conducting parts of electric cables. Copper does not react with water and is a good conductor of heat. It is used for water pipes and radiators.

Reactivity of metals 1

1 Forming oxides and chlorides

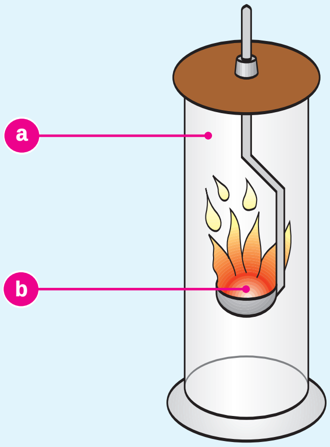

a oxygen or chlorine
b burning piece of reactive metal

2 Forming hydroxides

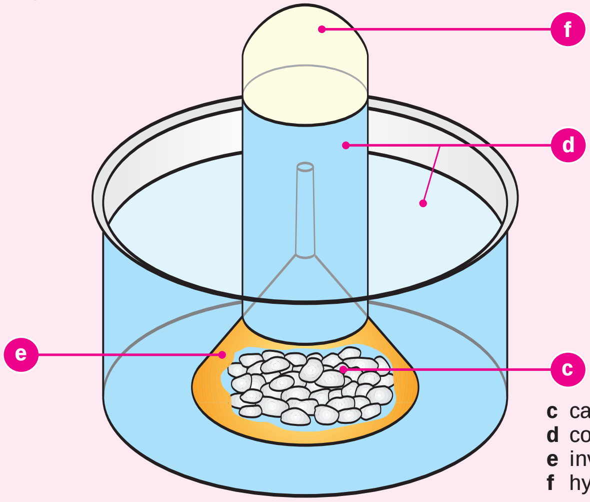

c calcium
d cold water
e inverted filter funnel
f hydrogen

3 Less reactive metals

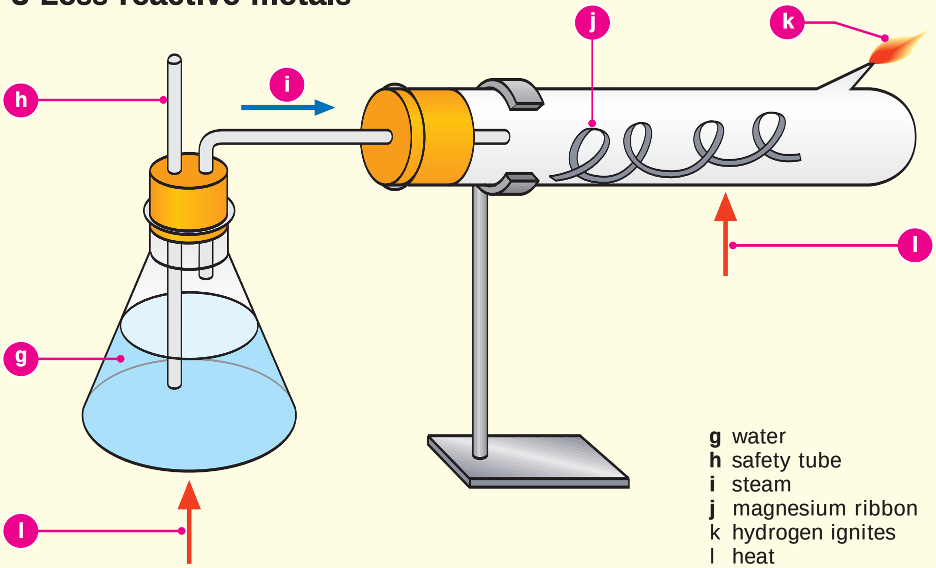

g water
h safety tube
i steam
j magnesium ribbon
k hydrogen ignites
l heat

Key words

alkali
calcium
chloride
copper
iron
hydroxide
magnesium
oxide
sodium

1 Forming oxides and chlorides

- Most metals react with air to form metal *oxides*. Reactive metals like *magnesium* burn, producing light and heat. Less reactive metals like *copper* simply change color on heating:

  $2Mg(s) + O_2(g) \rightarrow 2MgO(s)$

  $2Cu(s) + O_2 \rightarrow 2CuO(s)$

- Metals will also form *chlorides* when heated in chlorine:

  $Mg(s) + Cl_2(g) \rightarrow MgCl_2(s)$

  $Cu(s) + Cl_2 \rightarrow CuCl_2(s)$

2 Forming hydroxides

- Very reactive metals like *calcium* and *sodium* react with water to form solutions of metal *hydroxides* and hydrogen gas:

  $Ca(s) + 2H_2O(l) \rightarrow Ca(OH)_2(aq) + H_2(g)$

  $2Na(s) + 2H_2O(l) \rightarrow 2NaOH(aq) + H_2(g)$

- Calcium hydroxide is less soluble in water and forms a weak *alkali*.
- Sodium hydroxide is very soluble in water and forms a strong alkali.

3 Less reactive metals

- Less reactive metals, which react with water very slowly or not at all, react with steam to form metal oxides and hydrogen gas.
- Magnesium reacts very slowly with water but readily with steam:

  $Mg(s) + H_2O(g) \rightarrow MgO(s) + H_2(g)$

- Iron does not react with water but reacts with steam to form iron(II) diiron(III) oxide:

  $3Fe(s) + 4H_2O(g) \rightarrow Fe_3O_4(s) + 4H_2(g)$

- The least reactive metals, such as copper, do not react with water or steam.

© Diagram Visual Information Ltd.

Reactivity of metals 2

Key words

anode
carbonate
cathode
limewater
oxide
reactivity series
sulfuric acid

1 Metal compounds

- The *oxides* of metals that are low in the *reactivity series*, like copper, can be reduced by heating them in a stream of hydrogen gas.
- All group 1 metal *carbonates*, with the exception of lithium carbonate, are not decomposed on heating. All other metal carbonates decompose on heating, forming the metal oxide and carbon dioxide gas:

$Li_2CO_3(s) \rightarrow Li_2O(s) + CO_2(g)$

$MgCO_3(s) \rightarrow MgO(s) + CO_2(g)$

$CuCO_3(s) \rightarrow CuO(s) + CO_2(g)$

- Carbon dioxide gas is more dense than air and can be poured from one test tube into another. Carbon dioxide turns *limewater* milky.

2 Generating electric current

- When rods of zinc and copper are placed in dilute *sulfuric acid*, a simple electrical cell is formed, and there is a potential voltage difference between the two metals. If the two metals are connected externally, electric current flows.
- The zinc rod becomes the positive electrode (*anode*) of the cell. Zinc atoms are oxidized to form zinc ions:

$Zn(s) \rightarrow Zn^{2+}(aq) + 2e^-$

- The copper rod becomes the negative electrode (*cathode*) of the cell. Hydrogen ions are reduced to hydrogen gas:

$2H^+(aq) + 2e^- \rightarrow H_2(g)$

- If the copper rod is surrounded by a porous vessel containing copper(II) sulfate solution, a different reaction occurs at the cathode:

$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$

- Zinc atoms are oxidized to zinc ions, while copper ions are reduced to copper atoms.

© Diagram Visual Information Ltd.

1 Reactions of metal compounds

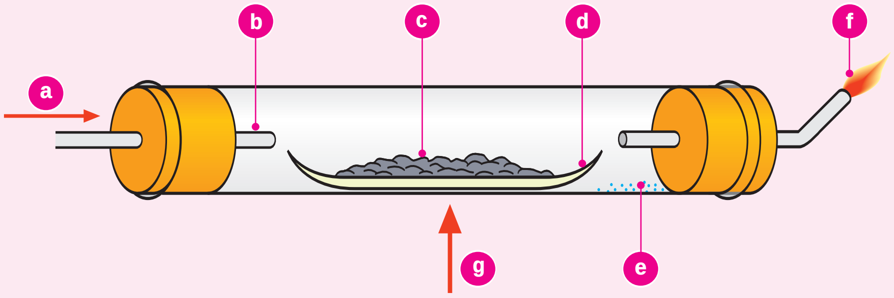

Reduction of oxides

a hydrogen
b combustion tube clamped to slope downward
c metallic oxide
d porcelain vessel
e moisture collects here
f hydrogen ignited
g heat

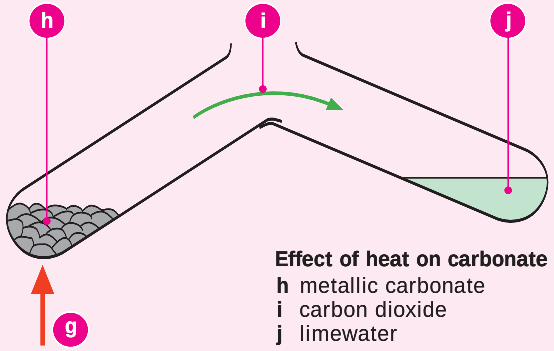

Effect of heat on carbonate

h metallic carbonate
i carbon dioxide
j limewater

2 Generation of electric current by mechanical reaction

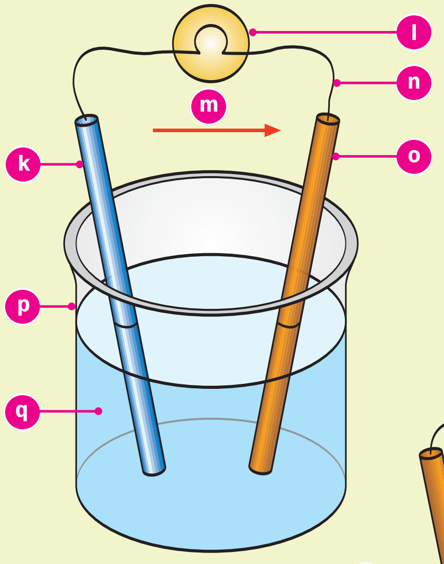

Method 1: simple cell

k zinc rod
l electric bulb
m electron transfer
n connecting wire
o copper rod
p beaker
q dilute sulfuric acid

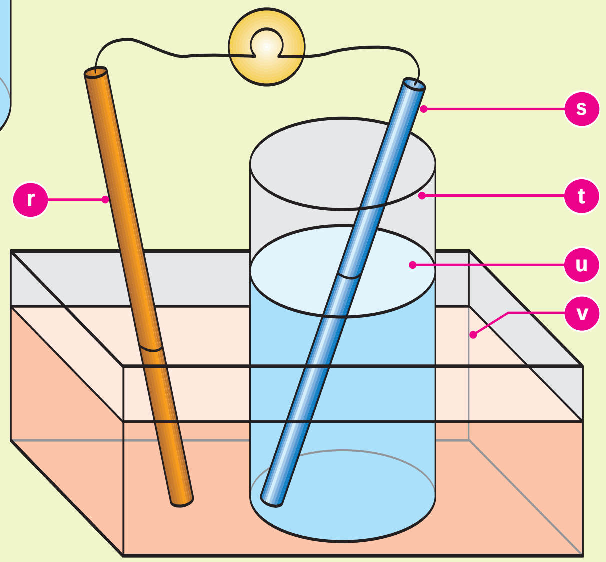

Method 2

r copper rod
s zinc rod
t porous vessel
u dilute sulfuric acid
v copper sulfate solution

Electrolysis

1 Electrolysis: schematic

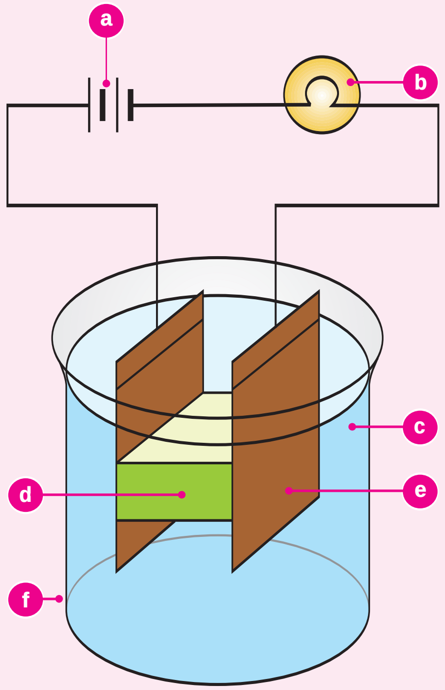

2 Electrolysis of salt solutions

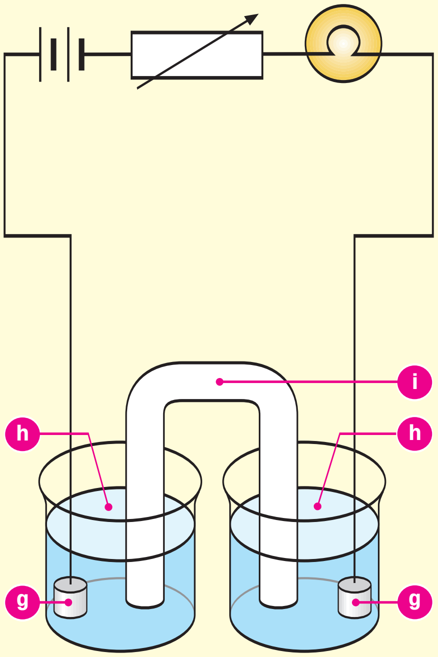

3 Electrolysis of water

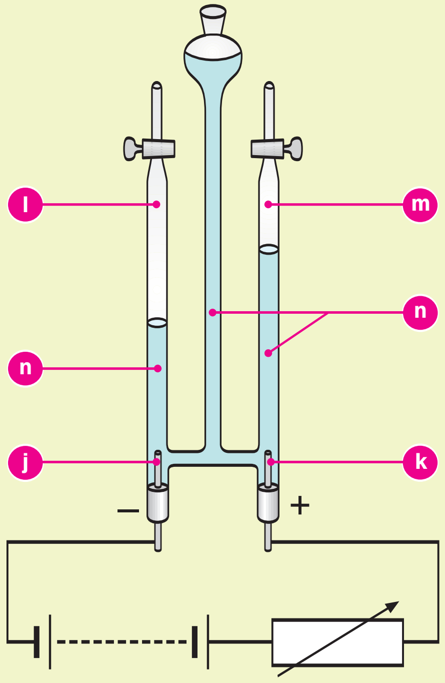

4 U tube

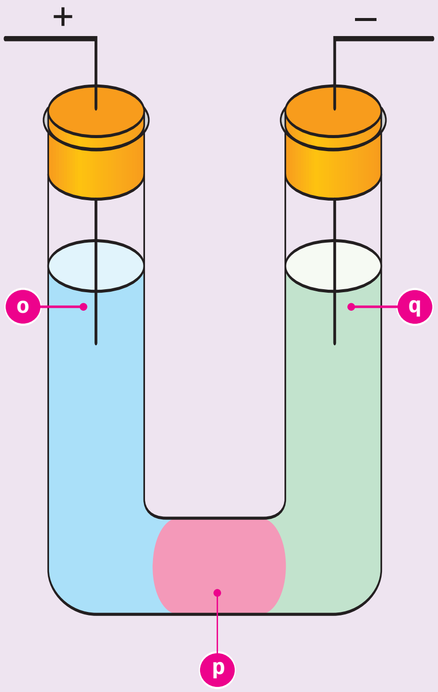

a battery
b electric bulb
c liquid under test
d poly(ethene) support
e copper plates
f glass vessel
g platinum electrodes
h electrolyte solution in beakers
i salt bridge
j platinum cathode
k platinum anode
l hydrogen
m oxygen
n water acidified with dilute sulfuric acid
o dilute sulfuric acid
p agar jelly colored pink by phenolphthalein and alkali
q sodium sulfate solution

Key words

anion	*electrolysis*
anode	*electrolyte*
cathode	
cation	
electrode	

1 Electrolysis

- *Electrolysis* is the process by which an *electrolyte* (a substance that conducts electricity) is decomposed when a direct current is passed through it between *electrodes*. Positive *cations* move to the *cathode* to gain electrons; negative *anions* move to the *anode* to lose electrons.
- Substances are either deposited or liberated at the electrodes depending on the nature of the electrodes and electrolyte.

2 Salt solutions

- Two electrolytes undergo electrolysis at the same time when they are connected in a circuit by a salt bridge.
- The platinum electrode in the left-hand beaker is the anode and attracts negative ions, which are oxidized.
- The platinum electrode in the right-hand beaker is the cathode and attracts positive ions, which are reduced.

3 Water

- The electrolysis of water yields hydrogen at the cathode and oxygen at the anode. Hydrogen and oxygen are formed in the ratio of 2:1.

4 U tube

- The ions present in dilute sulfuric acid are H^+, OH^-, and SO_4^{2-}. Hydroxide ions are discharged at the anode, leaving a surplus of hydrogen ions, so the electrolyte in the left side of the U tube becomes increasingly acidic.
- The ions present in sodium sulfate solution are H^+, Na^+, OH^-, and SO_4^{2-}. Hydrogen ions are discharged at the cathode, leaving a surplus of hydroxide ions, so the electrolyte in the right side of the U tube becomes increasingly alkaline.

© Diagram Visual Information Ltd.

Electrolysis: electrode activity and concentration

Key words

anode
cathode
electrode
electrolysis
inert

Electrode activity and concentration

- The results of *electrolysis* differ depending on the concentration of the solution and type of *electrodes* used.
- *Inert* electrodes take no part in the reaction; active electrodes take part in the reaction.

1 Dilute solution

- Reaction at the *anode*: oxygen produced
- Reaction at the *cathode*: hydrogen produced

2 Concentrated solution

- Reaction at the anode: chlorine produced
- Reaction at the cathode: hydrogen produced

3 Inert electrodes

- The following reactions occur at the electrodes when copper(II) sulfate undergoes electrolysis using carbon (inert) electrodes.
- Reaction at the anode: oxygen is produced
- Reaction at the cathode: copper metal is deposited on the cathode.

4 Active electrodes

- The following reactions occur at the electrodes when copper(II) sulfate undergoes electrolysis using copper (active) electrodes.
- Reaction at the anode: copper goes into solution as copper ions, and the anode grows smaller.
- Reaction at the cathode: copper metal is deposited, and the cathode grows bigger.

1 Dilute solution sodium chloride

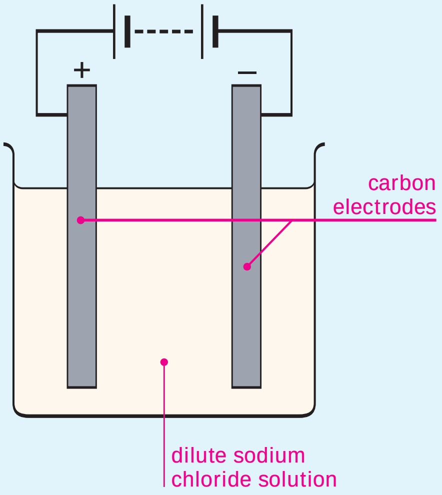
+
–
carbon electrodes
dilute sodium chloride solution

2 Concentrated solution sodium chloride

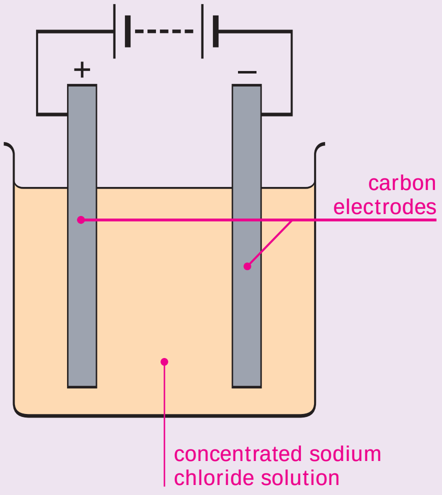
+
–
carbon electrodes
concentrated sodium chloride solution

3 Inert electrodes

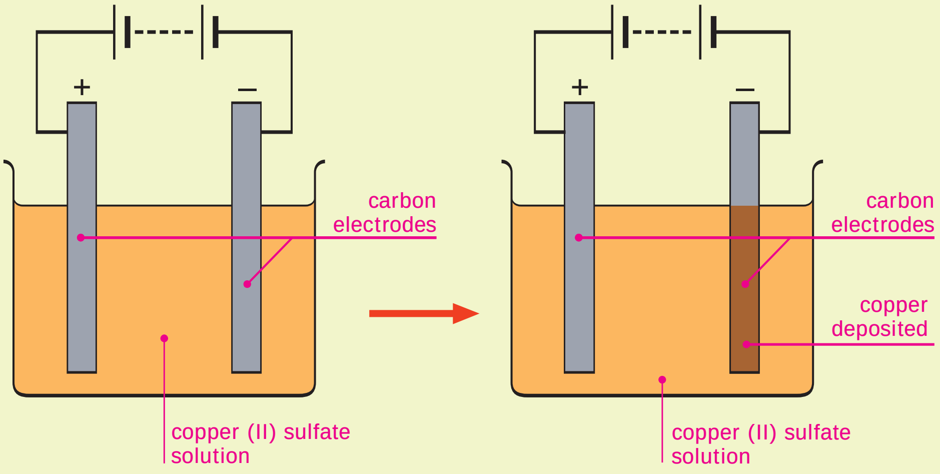
+
–
carbon electrodes
copper (II) sulfate solution
+
–
carbon electrodes
copper deposited
copper (II) sulfate solution

4 Active electrodes

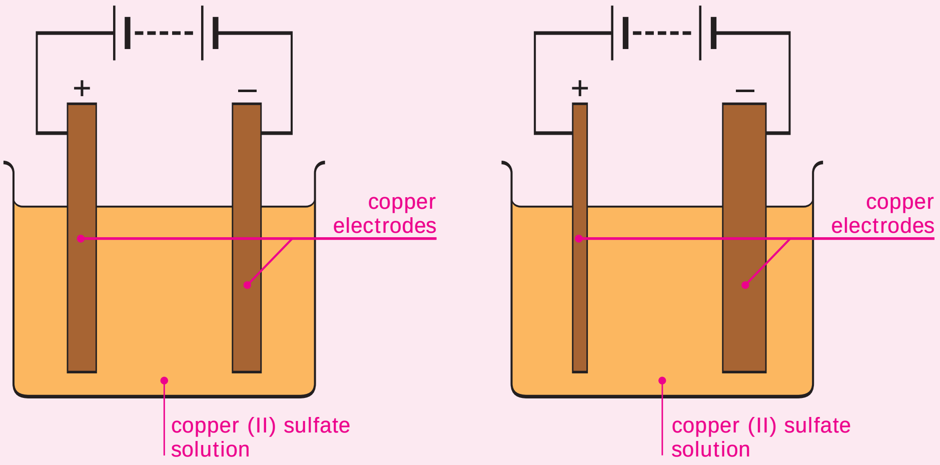
+
–
copper electrodes
copper (II) sulfate solution
+
–
copper electrodes
copper (II) sulfate solution

© Diagram Visual Information Ltd.

Acids: reactions

1 Main reactions of an acid

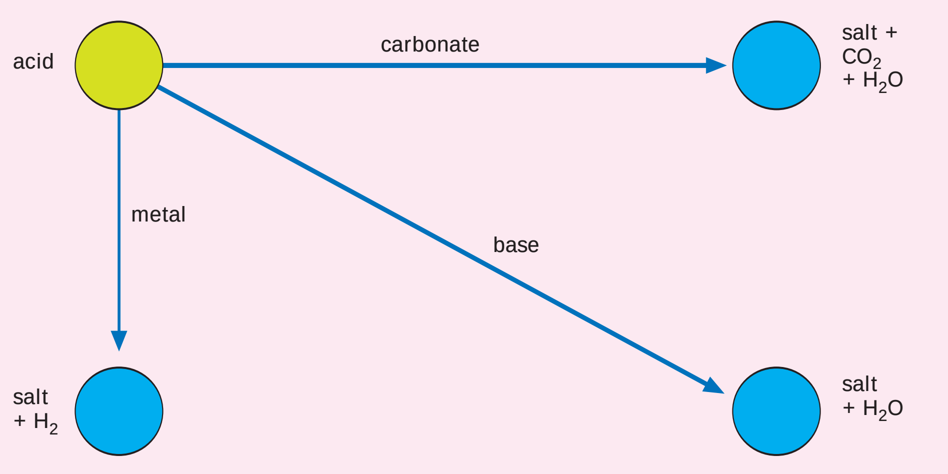

2 Examples of reaction type

Acid with carbonate	$Na_2CO_3(s) + 2HNO_3(aq) \rightarrow 2NaNO_3(aq) + CO_2(g) + H_2O(l)$
Acid with base	$HCl(aq) + NaOH(s) \rightarrow NaCl(s) + H_2O(l)$
Acid with metal	$Zn(s) + 2HCl(aq) \rightarrow ZnCl_2(aq) + H_2(g)$
Acid neutralized by oxide	$CuO(s) + H_2SO_4(aq) \rightarrow CuSO_4(aq) + H_2O(l)$

3 Laboratory preparation of sulfur trioxide

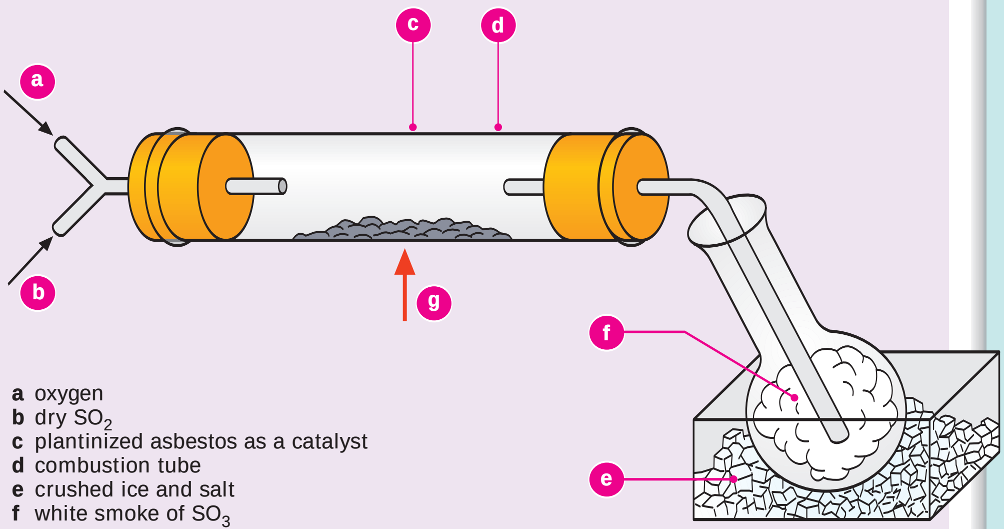

a oxygen
b dry SO_2
c plantinized asbestos as a catalyst
d combustion tube
e crushed ice and salt
f white smoke of SO_3
g heat

Key words

acid
base
carbonate
catalyst
hydroxide
oxidation
oxide
salt

1 Main reactions of an acid

- Dilute *acids* react with all metal *carbonates* to give a metal *salt*, carbon dioxide, and water.
- Dilute acids react with *bases* to give salts plus water.
- Dilute acids react with most metals to give a metal salt and hydrogen.
- Dilute acids are neutralized by metal *oxides* and metal *hydroxides* to form a metal salt and water.

2 Example of reaction type

- Sodium carbonate reacts with dilute nitric acid to give sodium nitrate, carbon dioxide, and water.
- Hydrochloric acid reacts with sodium hydroxide to form a salt and water.
- Zinc reacts with dilute hydrochloric acid to give zinc chloride and hydrogen.
- Copper(II) oxide reacts with dilute sulfuric acid to give copper(II) sulfate and water.

3 Sulfur trioxide

- Sulfur trioxide is a white crystalline solid obtained by *oxidation* of sulfur dioxide. It dissolves in water with a hissing noise and the production of heat, forming sulfuric acid. Sulfur trioxide is employed as a dehydrating agent.
- Sulfur trioxide is made in the laboratory by passing a mixture of dry sulfur dioxide and dry oxygen over a heated platinum *catalyst*. Sulfur trioxide melts at 17°C and condenses as a solid in a suitably cooled beaker.
- Industrially it is made using the contact process (see pages 75 & 76).

© Diagram Visual Information Ltd.

Preparation of acids

Key words

hydrochloric acid
hydrogen chloride
nitric acid
soluble
sulfuric acid

1 Preparing HCl gas

- *Hydrogen chloride* gas is made by the reaction of sodium chloride and concentrated sulfuric acid:

$2NaCl(s) + H_2SO_4(aq) \rightarrow$
$Na_2SO_4(aq) + 2HCl(g)$

The gas is more dense than air and is collected by downward delivery.

2 Preparing HCl acid

- Hydrogen chloride is extremely *soluble* in water, forming *hydrochloric acid*. It cannot be dissolved simply by placing a delivery tube carrying the gas directly into water because the water would be sucked back into the reaction vessel.
- The gas is dissolved in water by passing it into an inverted funnel positioned so the lip is just under the surface of the water. The funnel prevents suck back.

3 Preparing nitric acid

- *Nitric acid* can be made by the reaction of solid sodium or potassium nitrate with concentrated *sulfuric acid*:

$KNO_3(s) + H_2SO_4(aq) \rightarrow$
$KHSO_4(aq) + HNO_3$

- The product of this reaction is normally yellow due to the presence of nitrogen dioxide, formed by the thermal decomposition of the acid:

$4HNO_3(l) \rightarrow$
$4NO_2(g) + 2H_2O(g) + O_2(g)$

4 Industrial preparation of HNO_3

- Nitric acid is made industrially by the oxidation of ammonia in a process involving three stages (see page 76): production of nitrogen oxide gas, oxidation of nitrogen oxide to nitrogen dioxide gas, reaction of nitrogen dioxide and water.
- This process can be modeled in the laboratory by passing ammonia vapor over a heated platinum catalyst.

© Diagram Visual Information Ltd.

1 Preparation of hydrogen chloride (gas)

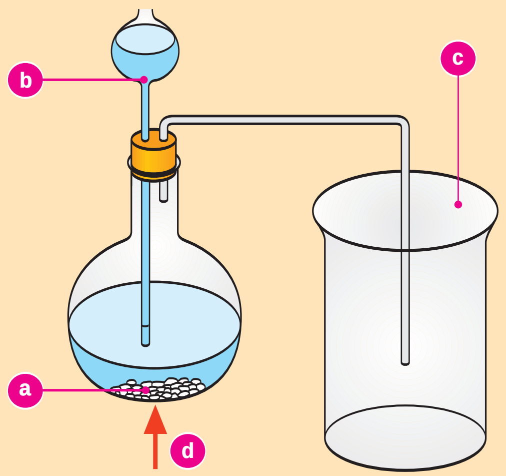

a rock salt
b concentrated sulfuric acid
c HCl gas collected
d heat

2 Preparation of hydrochloric acid

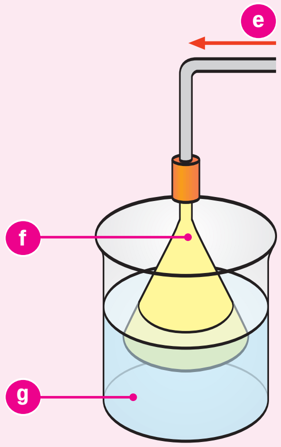

e HCl filter
f filter funnel
g water (to become dilute HCl acid)

3 Laboratory preparation of nitric acid

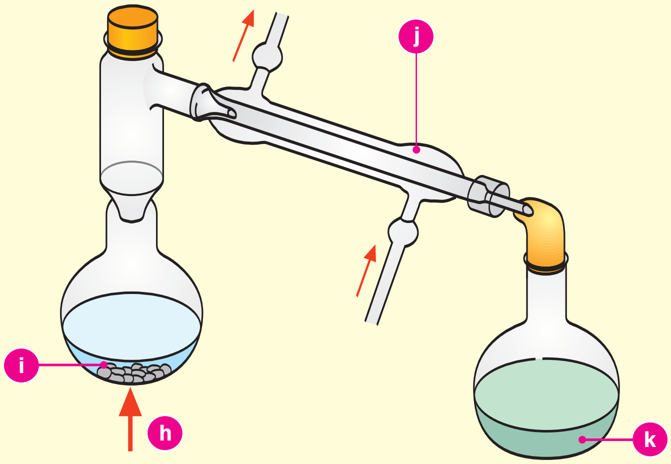

h heat
i solid sodium nitrate plus concentrated sulfuric acid
j water jacket
k pure nitric acid

4 Industrial preparation of nitric acid

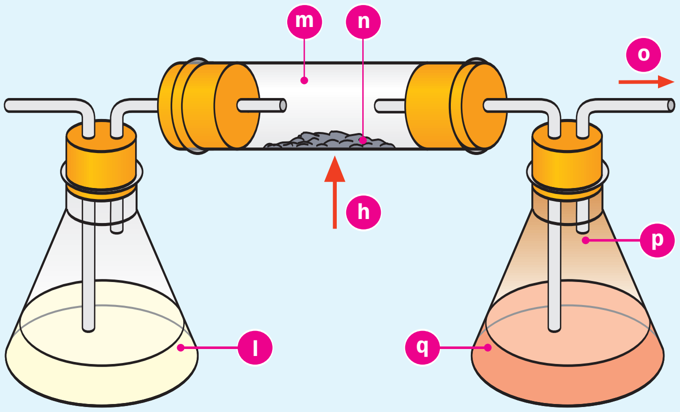

l concentrated ammonia diluted with water (50%)
m combustion tube
n platinized asbestos
o pump sucks gases through apparatus
p brown gas
q litmus goes red

Bases: reactions

1 General reactions of a base with an acid

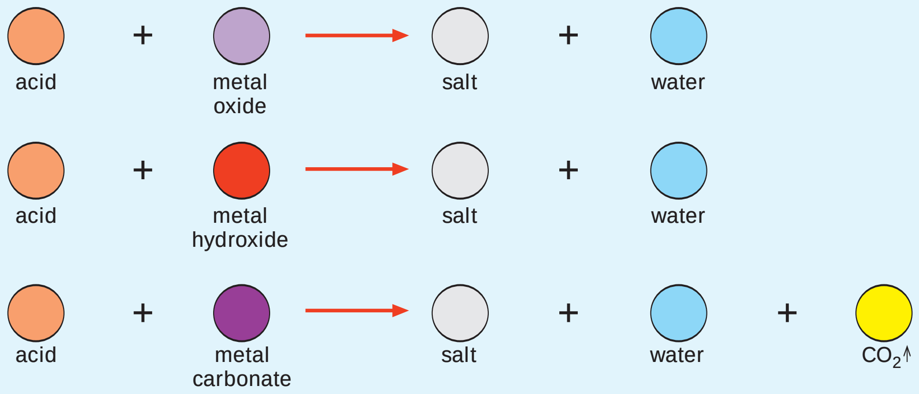

2 Metal oxide and acid

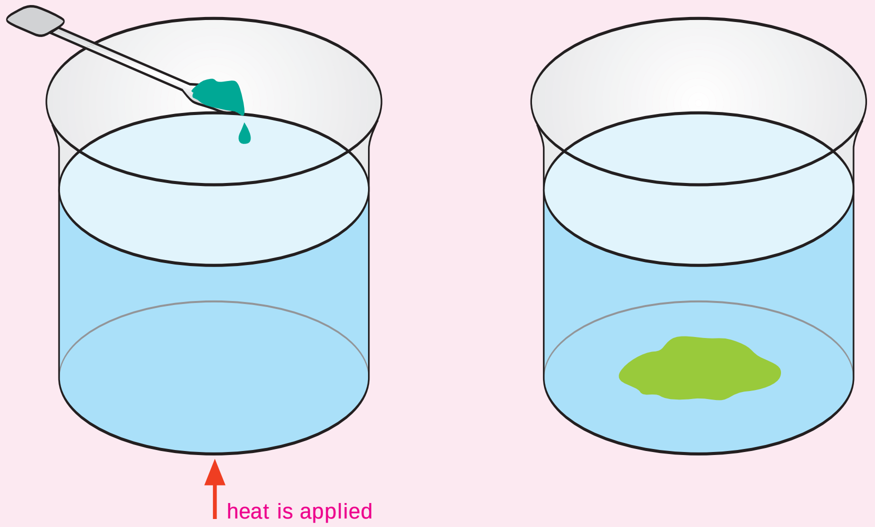

Magnesium oxide is added to hydrochloric acid and indicator

Neutral solution, indicator is green

3 Carbonate and acid

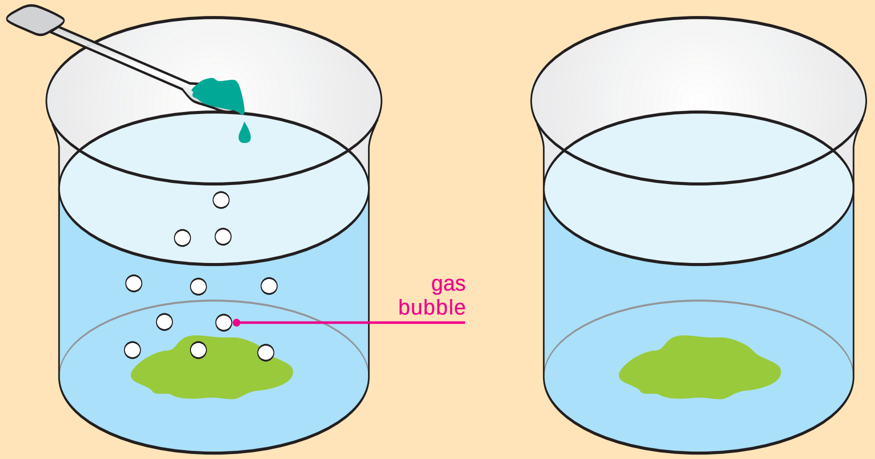

Cabonate is added to hydrochloric acid and indicator

Neutral solution, indicator is green

Key words

acid
base
carbonate
hydroxide
oxide
salt
universal indicator

Bases

- A *base* is a compound that reacts with an *acid* to form a *salt*. Common bases are metal *oxides*, metal *hydroxides*, and metal *carbonates*.

1 General reactions with acids

- Metal oxides react with acids to form salts and water.
- Metal hydroxides also react with acids to form salts and water.
- Metal carbonates react with acids to form salts, water, and carbon dioxide.

2 Metal oxide and acid

- The reaction of magnesium oxide (MgO) with hydrochloric acid (HCl) can be followed by adding a few drops of *universal indicator* to the acid.
- Initially the indicator is red. When magnesium oxide is added to the reaction, the following reaction occurs:
  $MgO(s) + 2HCl(aq) \rightarrow$
  $MgCl_2(aq) + H_2O(l)$
- When there are equivalent amounts of magnesium oxide and hydrochloric acid, the indicator turns green, signifying all of the acid has reacted and the mixture is neutral.

3 Carbonate and acid

- The reaction of magnesium carbonate ($MgCO_3$) with hydrochloric acid (HCl) can be followed by observing the carbon dioxide gas evolved.
- Initially bubbles of gas are evolved as the following reaction occurs:
  $MgCO_3(s) + 2HCl(aq) \rightarrow$
  $MgCl_2(aq) + H_2O(l) + CO_2(g)$
- When all of the hydrochloric acid has reacted, no gas is produced, and excess insoluble magnesium carbonate remains in the beaker.

© Diagram Visual Information Ltd.

Bases: forming pure salts

Key words

acid
base
indicator
insoluble
neutral
salt
soluble
titration

1 From a soluble base

- *Titration* is used to make *salts* from *acids* and *soluble bases*, e.g., sodium chloride from hydrochloric acid and sodium hydroxide.
- The burette is filled with hydrochloric acid, and a known volume of sodium hydroxide solution is placed in a conical flask. A few drops of a suitable *indicator* are added to the sodium hydroxide solution. Hydrochloric acid is run into the flask until the color of the indicator changes, showing that the reaction mixture is *neutral*. The volume of hydrochloric acid in the burette is noted before and after addition so the volume of acid needed can be calculated.
- The flask contains a solution of sodium chloride, which is impure due to the presence of the indicator. The procedure must be repeated using exactly the same volumes of hydrochloric acid and sodium hydroxide solution but no indicator.
- Sodium chloride crystals are obtained by boiling off some of the water from the sodium chloride solution and allowing the remaining solution to cool.

2 From an insoluble base

- Salts are made from *insoluble* bases by adding an excess of the base to an acid. For example, copper(II) sulfate is formed by the reaction of copper(II) oxide and sulfuric acid.
- An excess of copper(II) oxide is used to ensure that all of the sulfuric acid has reacted and no acid residue remains. The excess is filtered off, leaving a blue solution of copper(II) sulfate.
- Copper(II) sulfate crystals are obtained by boiling off some of the water from the copper(II) sulfate solution and allowing the remaining solution to cool.

© Diagram Visual Information Ltd.

1 From a soluble base (alkali)

Example: sodium chloride from sodium hydroxide and hydrochloric acid

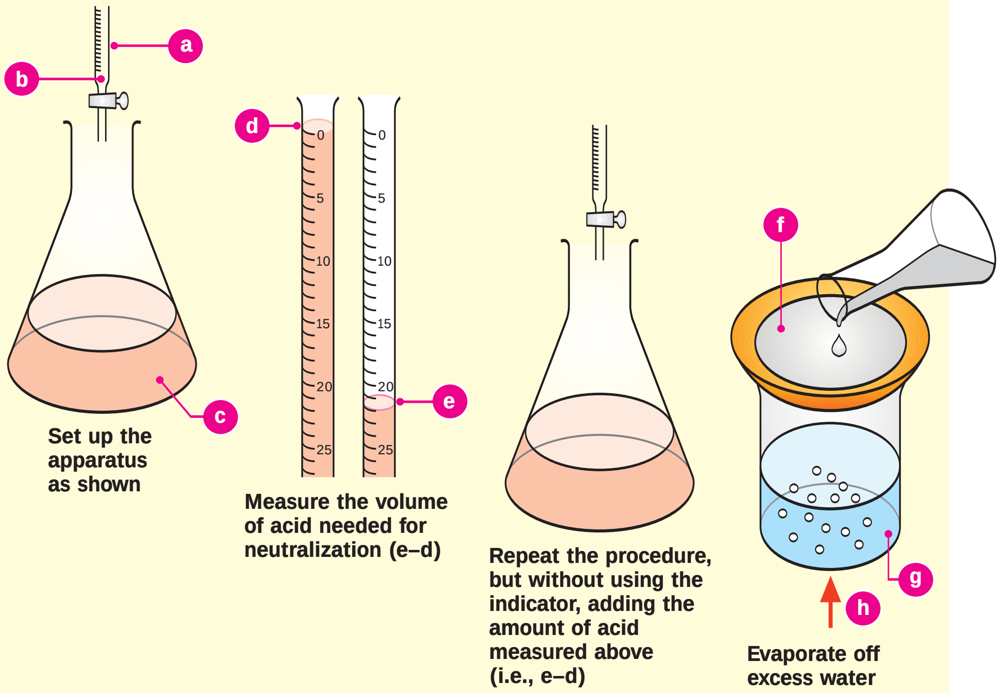
a
b
d
0
5
10
15
20
25
0
5
10
15
20
25
e
c
f
g
h
Set up the apparatus as shown
Measure the volume of acid needed for neutralization (e–d)
Repeat the procedure, but without using the indicator, adding the amount of acid measured above (i.e., e–d)
Evaporate off excess water

2 From an insoluble base

Example: copper oxide and sulfuric acid

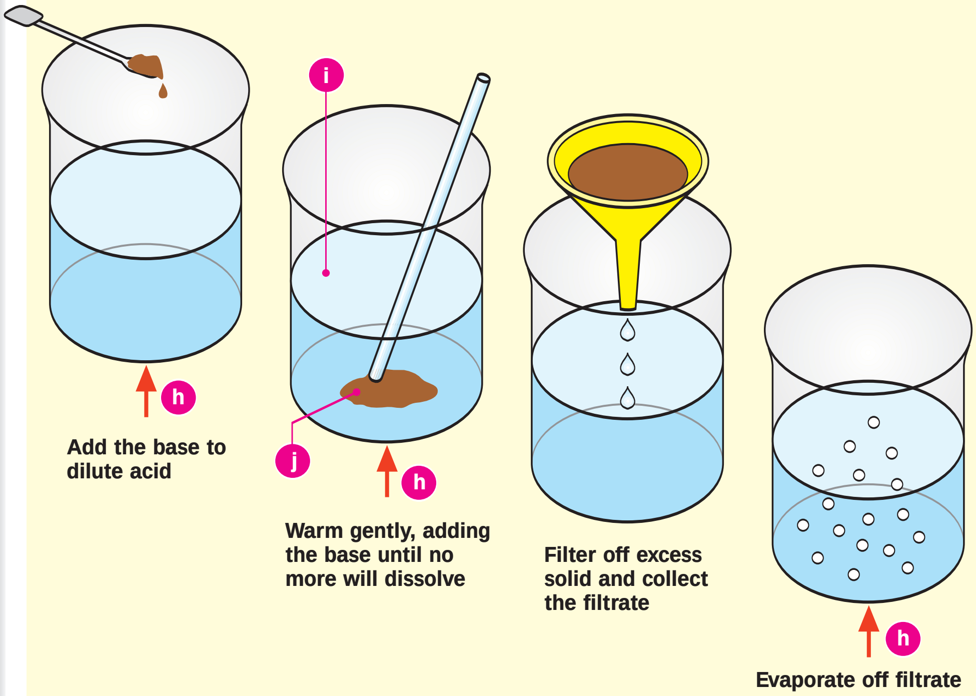
h
Add the base to dilute acid
i
j
h
Warm gently, adding the base until no more will dissolve
Filter off excess solid and collect the filtrate
h
Evaporate off filtrate

a burette
b acid
c alkali and phenolphthalein indicator Add the acid until the solution just turns colorless.
d volume of acid in the burette before carrying out the procedure
e volume of acid remaining when the indicator has turned colorless
f salt solution
g boiling water
h heat
i neutralized acid
j excess solid

Proton transfer: neutralization of alkalis

1 Water particles

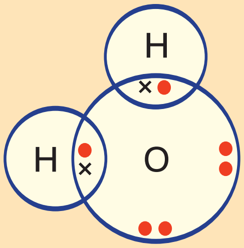

Neutral water molecule

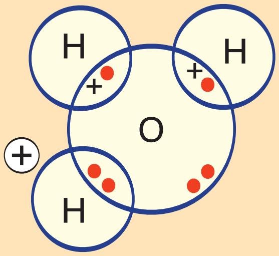

Hydronium ion: protonated water molecule

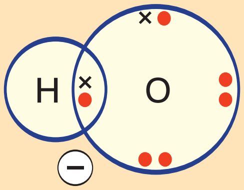

Hydroxide ion: deprotonated water molecule

2 Ammonia solution turns universal indicator blue

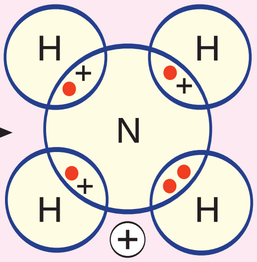

Ammonia molecule has an extra proton

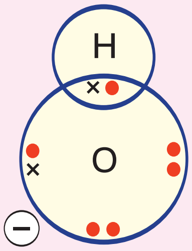

To show the extra-electron in the hydroxide ion

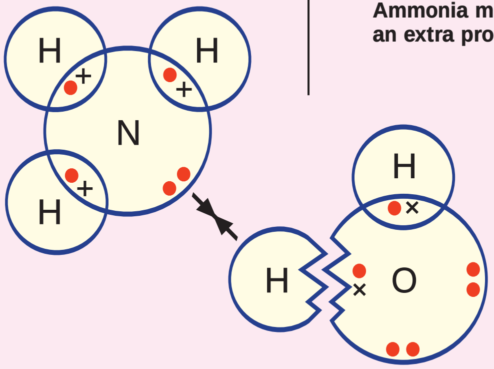

To show the attraction between the molecules and the breaking of the bond in the molecule

3 Schematic of proton transfer in diagram 2

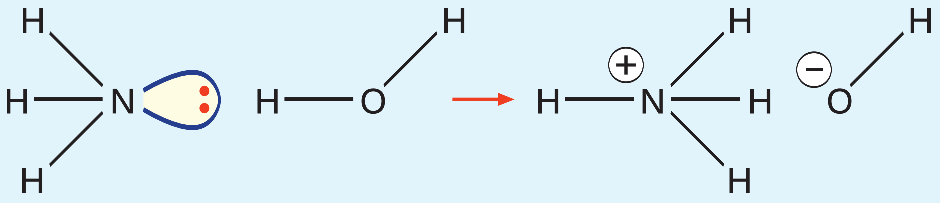

Key words

ammonium hydroxide
ammonium ion
hydronium ion
hydroxide ion
proton
species

1 Water particles

- In a water molecule, the oxygen atom forms bonds with two hydrogen atoms. The oxygen atom and the hydrogen atom each donate one electron to the bond. The oxygen atom also has two pairs of non-bonding electrons, which can be donated to form bonds with other *species*.
- In acidic solutions, each proton reacts with a water molecule to form a *hydronium ion*. A pair of non-bonding electrons forms the new H-O bond: $H^+ + H_2O \rightarrow H_3O^+$
- An *hydroxide ion* is formed by the loss of a *proton* from a water molecule: $H_2O \rightleftharpoons H^+ + OH^-$

2 Ammonium ions

- The ammonia molecule, NH_3, has a similar structure to the water molecule, H_2O, in the sense that the nitrogen atom has a pair of non-bonding electrons that it can donate to form a bond with another species.
- Ammonia reacts with the protons in an acid to form the *ammonium ions*: $NH_3 + H^+ \rightarrow NH_4^+$
- The four N-H bonds in the ammonium ion are directed toward the corners of a tetrahedron, giving a similar structure to methane. This keeps the bonding pairs of electrons as far away from each other as possible.

3 Schematic of proton transfer

- Ammonia is very soluble in water and dissolves to form a weak alkaline solution that is sometimes referred to as *ammonium hydroxide*: $NH_3 + H_2O \rightleftharpoons NH_4OH \rightleftharpoons NH_4^+ + OH^-$
- Ammonia solution contains ammonium ions, NH_4^+, and hydroxide ions, OH^-, and has similar reactions to solutions of soluble metal hydroxides, such as sodium hydroxide.

© Diagram Visual Information Ltd.

Proton transfer: neutralization of bases

Key words

acid
base
hydronium ion
lattice
magnesium oxide
oxide
proton
salt

Neutralizing bases

- Metallic *bases* neutralize *acids* to form a *salt* plus water.

1 MgO in acid

- *Magnesium oxide* consists of a regular *lattice* of magnesium ions, Mg^{2+}, and *oxide* ions, O^{2-}.
- An acid contains *hydronium ions*, H_3O^+.

2 Attractions

- Hydronium ions carry a positive charge, while oxide ions carry a negative charge. When solid magnesium oxide is added to an acid, these oppositely charged ions are attracted to each other.

3 Transfer

- In an oxide ion, there are eight electrons in the outer orbital of the oxygen atom. Two pairs of electrons are donated to form bonds with oppositely charged hydronium ions: $2H_3O^+ + O^{2-} \rightarrow 3H_2O$

4 Neutral solution

- Each hydronium ion transfers a *proton* to the oxide ion, forming a molecule of water.
- The magnesium oxide lattice breaks down, releasing magnesium ions into solution.
- The acid is neutralized, and a solution of a magnesium salt is formed. The nature of the salt depends on the acid used.

$MgO(s) + 2HCl(aq) \rightarrow$
$MgCl_2(aq) + H_2O(l)$
Hydrochloric acid →
magnesium chloride

$MgO(s) + 2HNO_3(aq) \rightarrow$
$Mg(NO_3)_2(aq) + H_2O(l)$
Nitric acid → magnesium nitrate

$MgO(s) + H_2SO_4(aq) \rightarrow$
$MgSO_4(aq) + H_2O(l)$
Sulfuric acid → magnesium sulfate

© Diagram Visual Information Ltd.

1 Magnesium oxide (MgO) solid and dilute acid

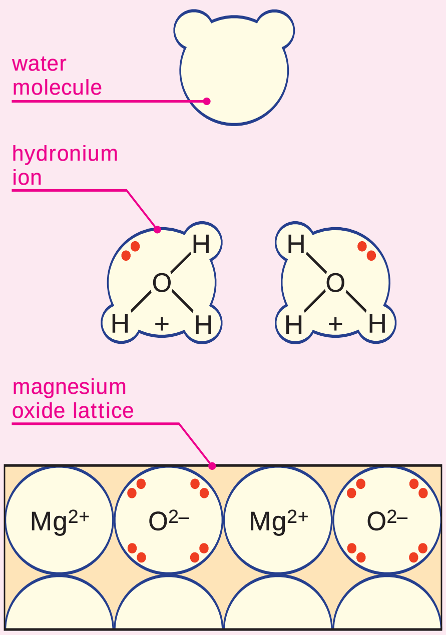

2 The oxide ions attract the hydronium ions

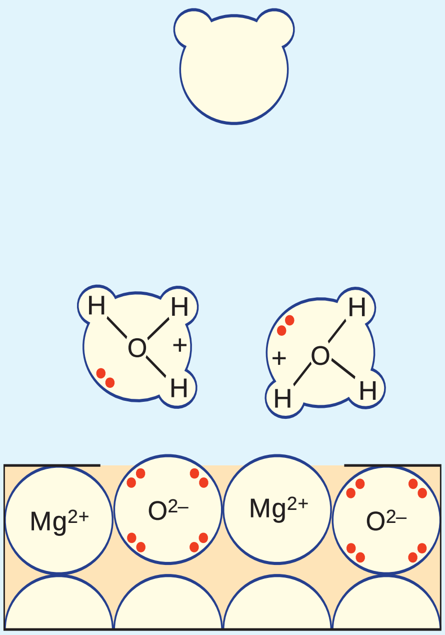

3 Proton transfer takes place

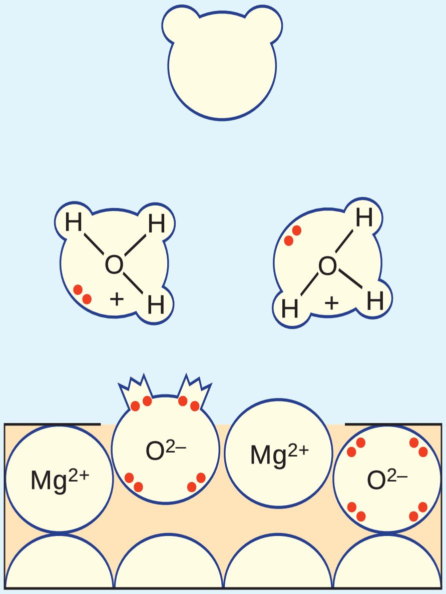

4 A neutral solution is produced and part of the oxide lattice has dissolved

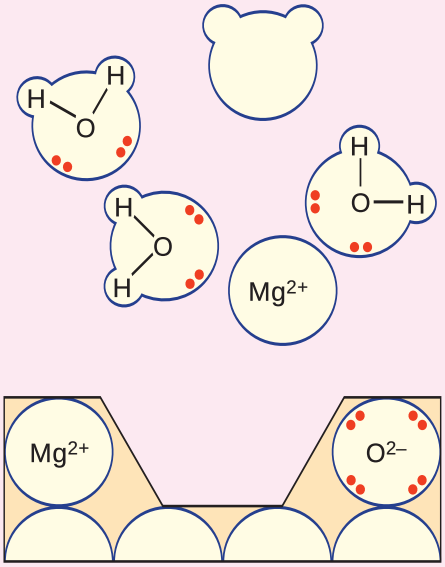

Proton transfer: metallic carbonates

1 Carbonate ions attract hydronium ions

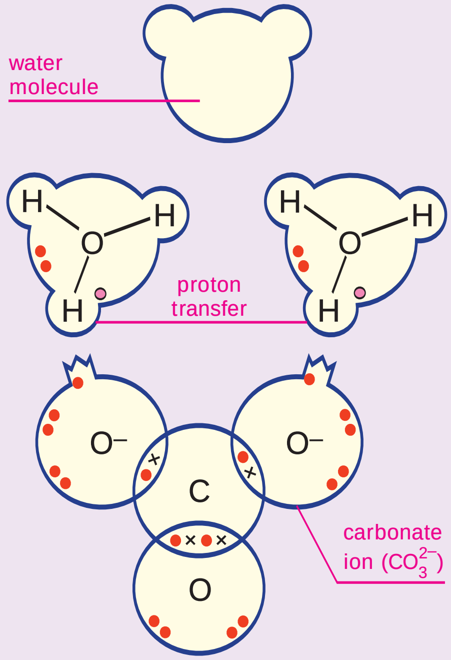

2 Hydrogen carbonate molecules and water molecules are produced

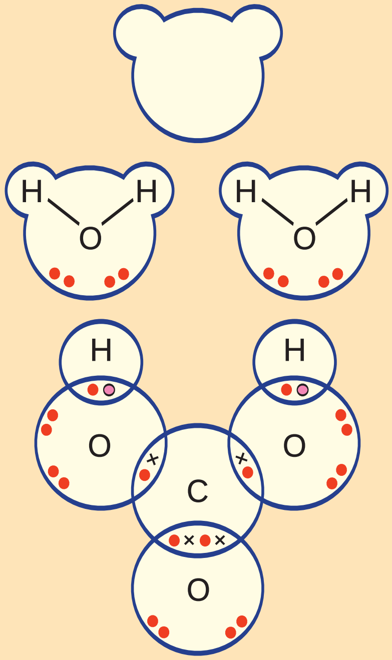

3 A hydrogen carbonate molecule splits

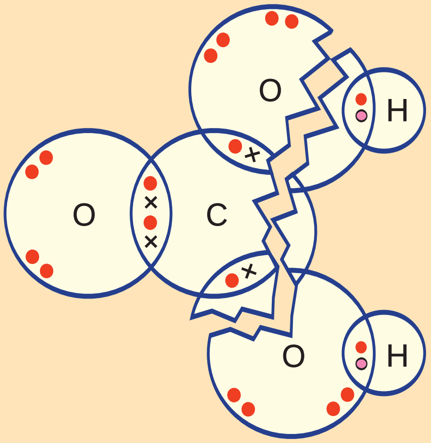

4 A carbon dioxide molecule and water molecule are produced

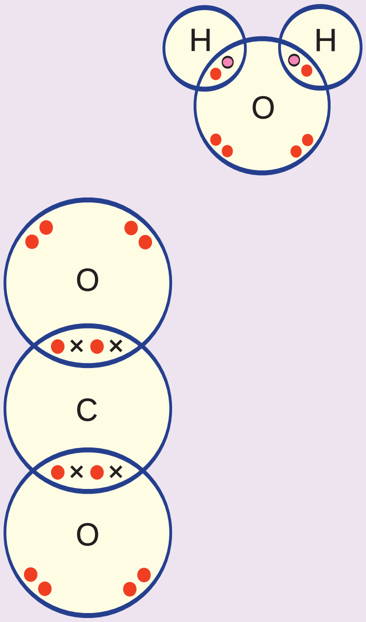

Key words

carbonate
carbonic acid
hydronium ion
orbital
salt

Metallic carbonates

- Metallic *carbonates* neutralize acids to form a metal *salt*, carbon dioxide, and water.

1 Attraction

- Group 1 metal carbonates are soluble in water and can be used as solids or in solution. Other metal carbonates are insoluble in water and are used as solids.
- All metal carbonates contain the carbonate ion, CO_3^{2-}. All acids contain the *hydronium ion*, H_3O^+.
- Hydronium ions carry a positive charge, while carbonate ions carry a negative charge. When a carbonate is added to an acid, these oppositely charged ions are attracted to each other.
- In a carbonate ion, each of the three oxygen atoms has eight electrons in its outer *orbital*. A pair of electrons is donated from two of the oxygen atoms to form bonds with oppositely charged hydronium ions.

2 H_2CO_3 and water

- The result is the formation of *carbonic acid*, H_2CO_3, and water:
$2H_3O^+ + CO_3^{2-} \rightarrow H_2CO_3 + 2H_2O$

3 H_2CO_3 splits

- Carbonic acid is a weak acid that only exists in solution. It readily breaks down to carbon dioxide and water:
$H_2CO_3 \rightleftharpoons H_2O + CO_2$

4 CO_2 and water

- In an acid–carbonate reaction, some of the carbon dioxide will remain in solution, but most will be given off as bubbles of gas.
- The gas can be identified by bubbling it into limewater. Carbon dioxide turns limewater milky due to the formation of insoluble calcium carbonate.
- The acid is neutralized by the carbonate, and a salt is formed. The nature of the salt depends on the metal carbonate and the acid used.

© Diagram Visual Information Ltd.

Proton transfer: neutralization of acids

Key words

ammonia
covalent bond
hydrogen chloride

Neutralizing acids

- Bases react with acids to produce a salt and water.

1 Molecules

- A chlorine atom has seven electrons in its outer orbit. In *hydrogen chloride*, the chlorine atom forms a *covalent bond* with one hydrogen atom, forming the molecule HCl.
- An oxygen atom has six electrons in its outer orbit. In hydrogen oxide (water), the oxygen atom forms covalent bonds with two hydrogen atoms, forming the molecule H_2O.
- A nitrogen atom has five electrons in its outer orbit. In *ammonia*, the nitrogen atom forms covalent bonds with three hydrogen atoms, forming the molecule NH_3.

2 Schematic

- In a hydrogen chloride molecule, each chlorine atom is surrounded by eight electrons: one pair of bonding electrons and three pairs of non-bonding electrons (lone pairs).
- In a water molecule, each oxygen atom is surrounded by eight electrons: two pairs of bonding electrons and two lone pairs of electrons.
- In an ammonia molecule, each nitrogen atom is surrounded by eight electrons: three pairs of bonding electrons and one pair lone pair of electrons.

3 & 4 Proton transfer and schematic

- Hydrogen chloride gas is a covalent compound and exists as diatomic molecules.
- When hydrogen chloride dissolves in water, an acidic solution is formed:
  $H_2O + HCl \rightarrow H_3O^+ + Cl^-$
- A lone pair of electrons from an oxygen atom is donated to create a covalent bond between the oxygen atom and a hydrogen atom, forming a hydronium ion and a chloride ion.
- When hydrogen chloride is dissolved in water, it forms an ionic compound.

© Diagram Visual Information Ltd.

1 Examples of molecules

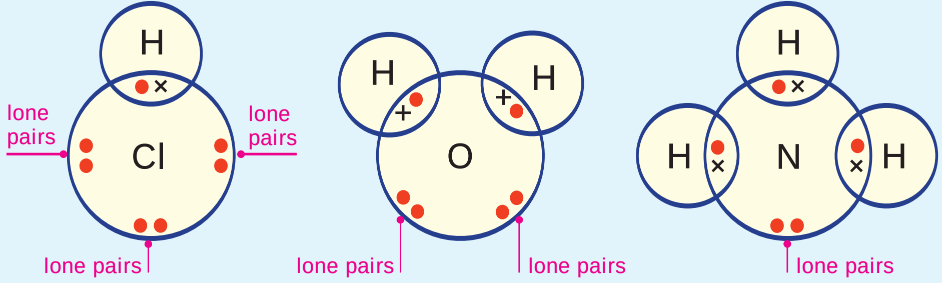

Hydrogen chloride molecule

Water molecule

Ammonia molecule

2 Schematic of the molecules shown in diagram 1

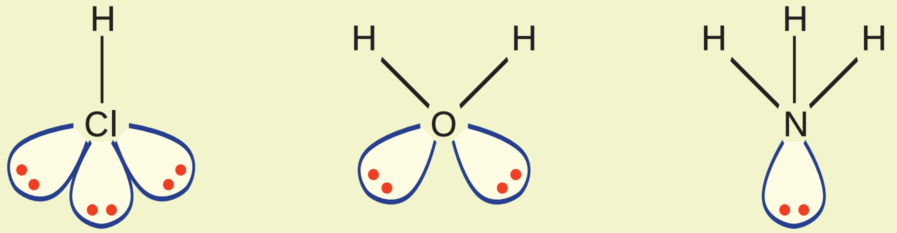

Hydrogen chloride molecule

Water molecule

Ammonia molecule

3 Proton transfer

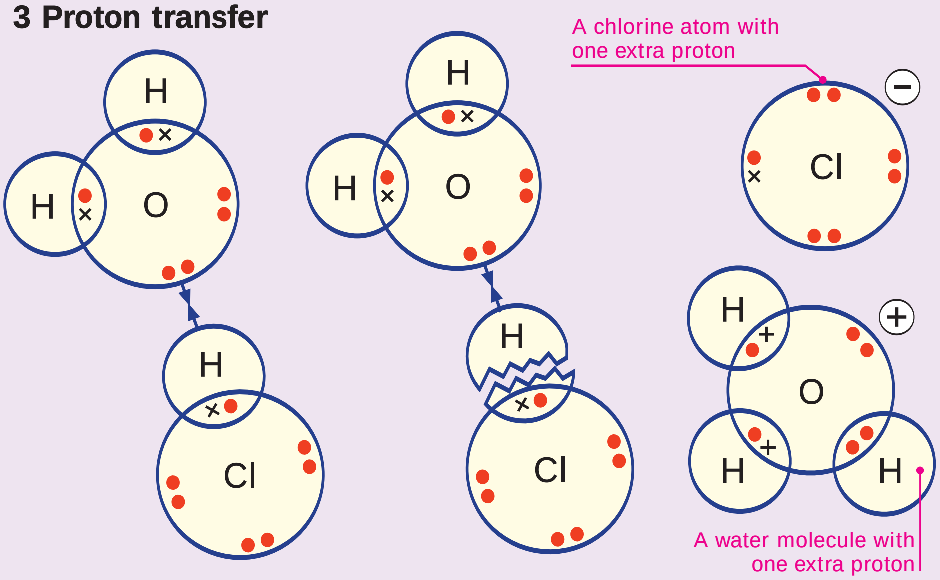
H
H
O
H
Cl
H
H
O
H
Cl
A chlorine atom with one extra proton
Cl
H
O
H
H
A water molecule with one extra proton

Attraction begins

The bond breaks

Transfer of the proton is complete

4 Schematic of proton transfer

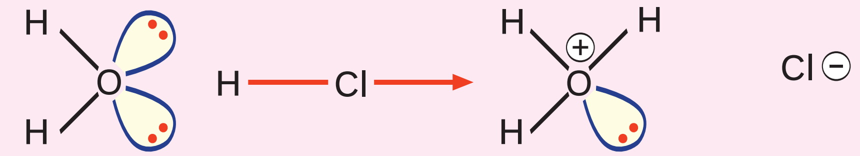

Collision theory

1 Collision theory

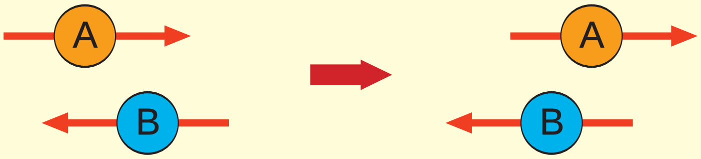

No collision between the particles of the reactants: no reaction

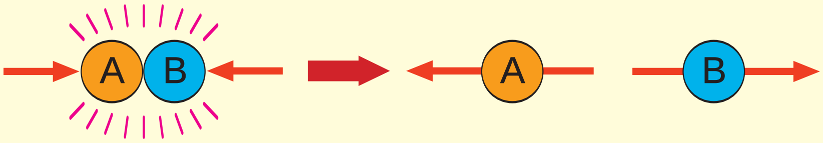

Weak collision: no reaction

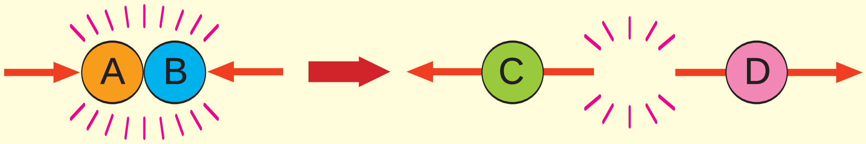

Effective collision: reaction

2 Maxwell-Boltzman distribution

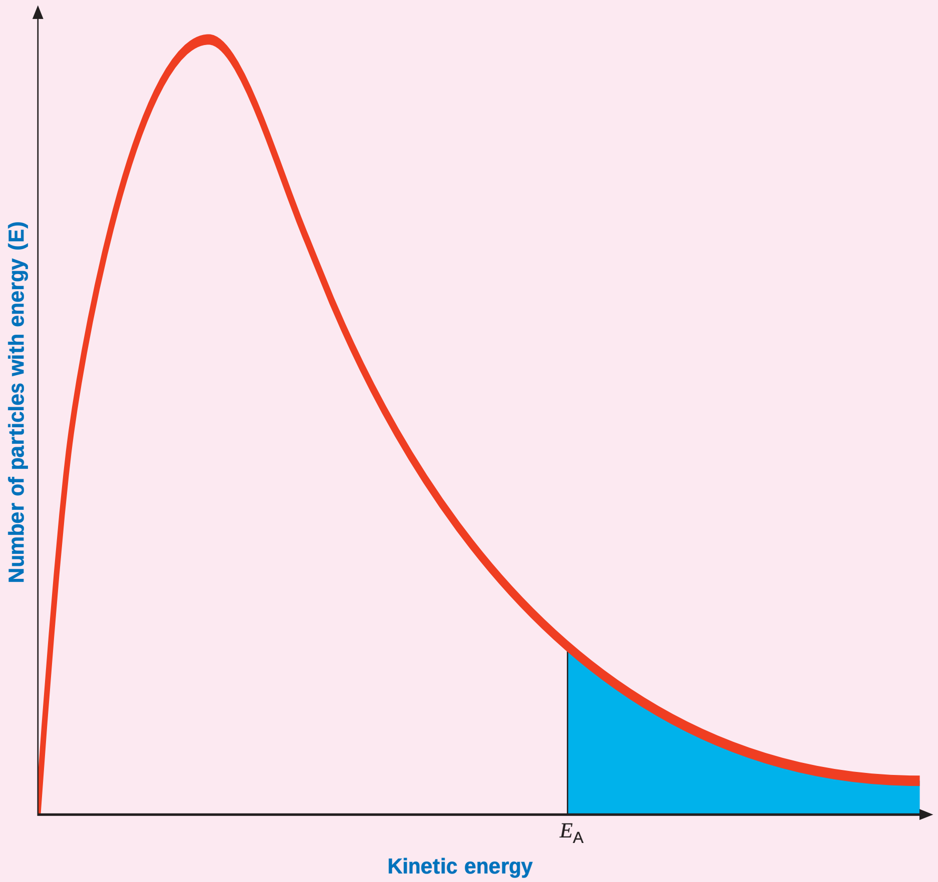

Key words

activation energy
effective collision
product
reactant

1 Collision theory

- Reactions occur when particles collide with sufficient force to provide the energy needed to start a reaction.
- If particles collide with insufficient force to start a reaction, they simply bounce off each other.
- A collision that brings about a reaction is called an *effective collision*. Particles of *reactant* collide, and particles of *product* are formed:

$$\underbrace{A + B}_{\text{reactants}} \rightarrow \underbrace{C + D}_{\text{products}}$$

- Not every collision between particles gives rise to a reaction, but every set of particles that do react have to collide.

2 Maxwell-Boltzman distribution

- Because all the particles of a particular chemical, element, or compound have the same mass, the energy of the particles is directly related to their speed.
- In any mixture of moving particles, the energy at which an individual particle is moving will vary.
- The Maxwell-Boltzman distribution shows how the number of particles in a sample is distributed at different energies at a particular temperature.
- There are no particles at zero energy. There are relatively few particles at very high energy, but there is no maximum energy value.
- In order to react, particles need to have a minimum amount of energy, called *activation energy*. The activation energy is marked on the graph by a line, parallel to the Y axis, at a point on the X axis that symbolizes the activation energy (E_A).

© Diagram Visual Information Ltd.

Rates of reaction: surface area and mixing

Key words

diffusion
immiscible
reactant
surface area

Surface area

- In order for a reaction to take place, the *reactants* must come into contact with each other. Thus, for a given mass of reactant, the smaller the objects, the greater the *surface area* on which the chemical reaction can occur. If all of the reactants are gases or liquids, it is easy for them to mix, giving the maximum opportunity for the particles to collide.

1 Total surface area

- The reaction can only take place on the surface of the solid.
- A cube with sides 2 cm has a total surface area of 2 x 2 x 6 = 24 cm^2. If the same cube is divided into 8 cubes with sides 1 cm, the total surface area now becomes 1 x 1 x 6 x 8 = 48 cm^2.

2 Reduced surface area

- Zinc reacts with dilute hydrochloric acid to form zinc chloride and hydrogen gas:

  $Zn + 2HCl \rightarrow ZnCl_2 + H_2$
- This reaction proceeds much more quickly if zinc dust (fine powder) is used rather than granulated zinc (large lumps).

3 Mixing

- When reactant particles are added together, they will eventually mix by *diffusion*, and a reaction will take place.
- Stirring reactants speeds the process of mixing so the reaction takes place more quickly.

4 Interface surface area

- If one of the reactants is a liquid and one a gas, or if the two reactants are *immiscible* liquids, then the reaction can only take place at the interface. The larger the surface area of the interface, the faster the reaction will take place.

© Diagram Visual Information Ltd.

1 Total suface area

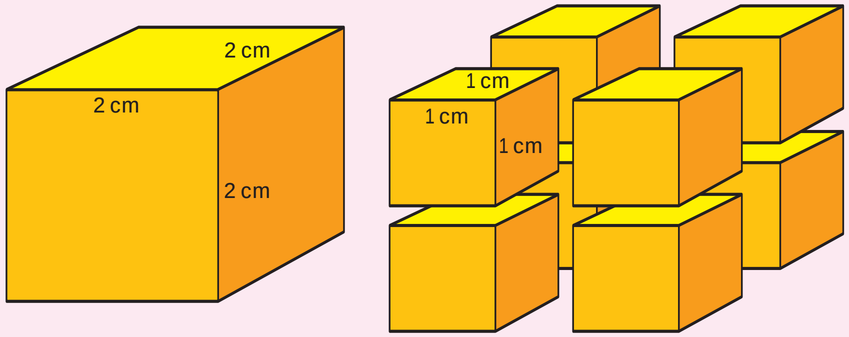

2 Reduced surface area reaction

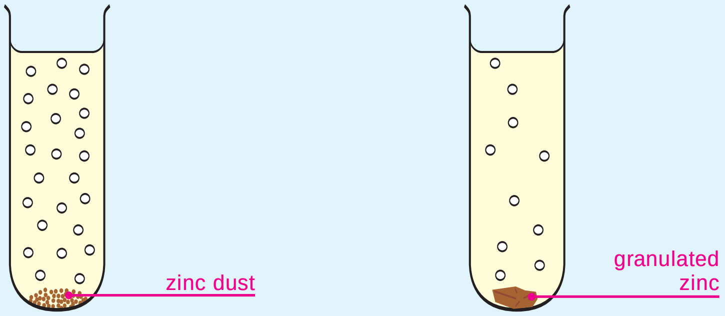

3 Mixing

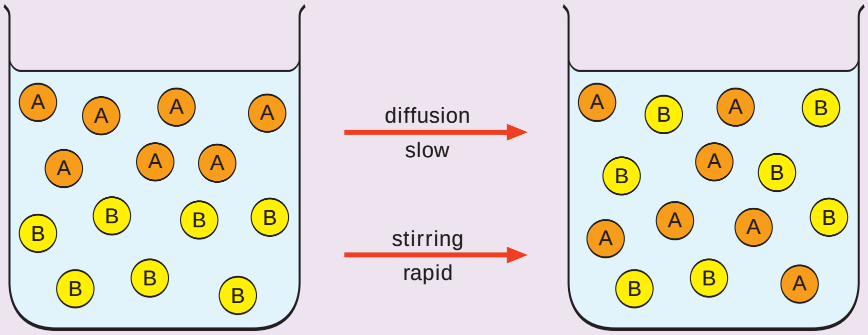

4 Interface surface area

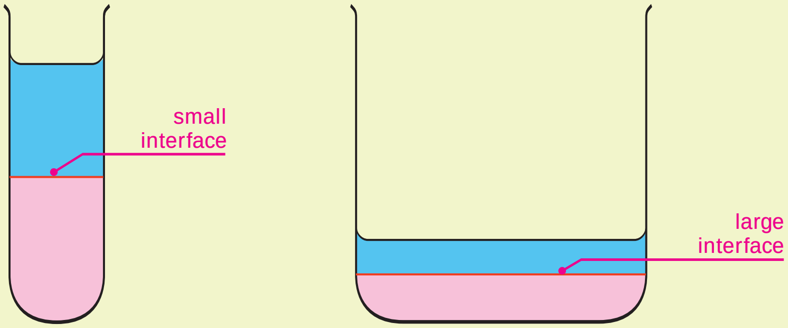

Rates of reaction: temperature and concentration

1 Temperature (distribution of molecular energies at T_1 and T_2)

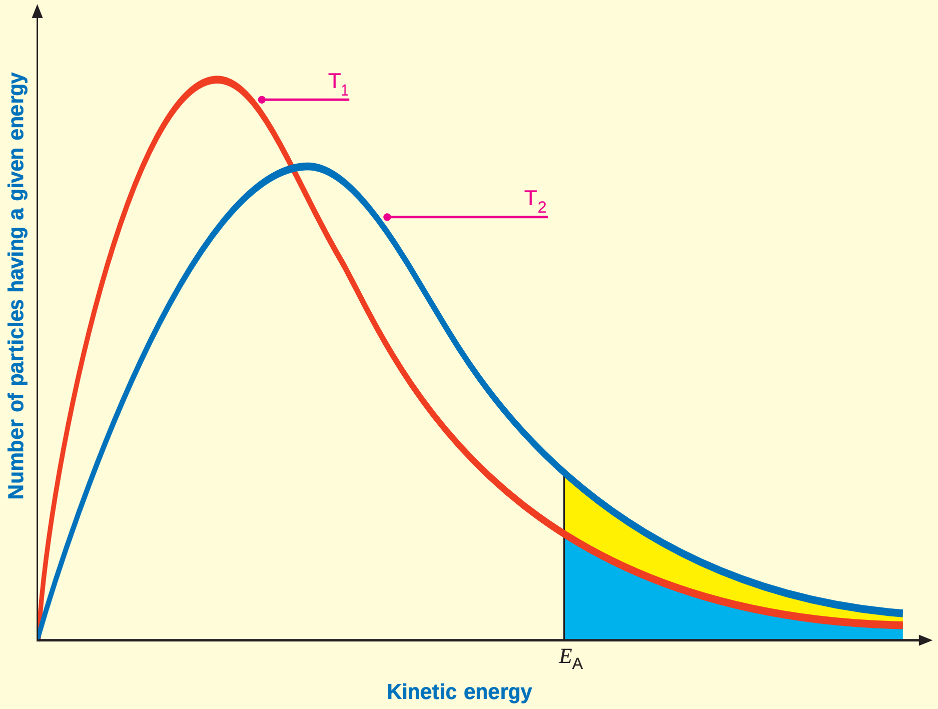

2 Concentration

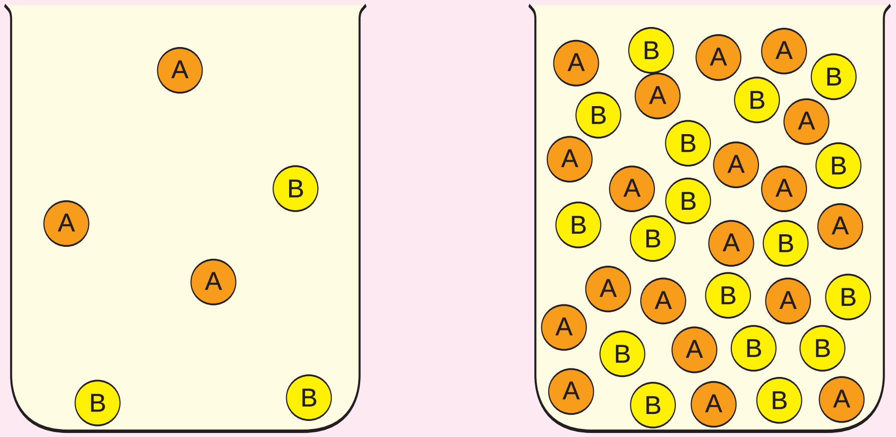

Low concentration

High concentration

3 Rate of reaction

$$\text{Rate of reaction} = \frac{\text{Change in concentration}}{\text{Time}}$$

Key words

activation energy
product
reactant

1 Temperature

- Temperature is an important factor in determining rate of reaction.
- When temperature increases, the average speed of the particles in a substance increases. The graph shows the Maxwell-Boltzman distribution at two temperatures, T_2 is greater than T_1.
- The number of particles is constant, so the area under the two curves is the same. However, the average energy of the particles at T_2 is greater. The area of the curve to the right of the *activation energy* line (E_A) is greater for T_2. Therefore, at this temperature a higher proportion of particles have sufficient energy to react.

2 Concentration

- An increase in the concentration of a chemical, or the pressure of a gas, means that there will be more particles within a given space, so particles will collide more often.

3 Rate of reaction

- The rate of any reaction is the speed at which the *reactants* are converted to *products*. This can be qualified as the change of concentration of reactants or products.
- Changes in concentration can be measured by:
  1. appearance or disappearance of color in reactants or products
  2. volume of gas evolved
  3. changes in pH
  4. heat produced
  5. changes in pressure.

© Diagram Visual Information Ltd.

Key words

concentration
rate of reaction

Concentration over time

- Bromine reacts with an excess of methanoic acid in aqueous solution according to the following equation. The reaction is catalyzed by acid:

$$Br_2(aq) + HCOOH(aq) \xrightarrow{H^+} 2Br^-(aq) + 2H^+(aq) + CO_2(g)$$

- The reaction can be followed by measuring the intensity of the red-brown at different time intervals and relating this to the *concentration* of bromine.
- The concentration of bromine, $[Br_2]$, falls during the reaction, so the rate of the reaction can be expressed in terms of the rate at which the bromine concentration changes.
- The *rate of reaction* = - rate of change of bromine concentration = $-\frac{d[Br_2]}{dt}$
- The rate of change of bromine concentration is negative because the bromine is being used up. The negative sign in the expression is necessary to give the rate of reaction a positive value.
- In order to obtain the rate of reaction at any given time, a tangent to the curve must be drawn at that particular time and the gradient measured. The concentration of bromine after 300 seconds (s) is 0.0035 mol dm^{-3}. The rate of reaction at this time is 1.2×10^{-5} mol dm^{-3} s^{-1}.

© Diagram Visual Information Ltd.

Rates of reaction: concentration over time

Graph to show the variation of bromine concentration with time in the reaction between methanoic acid and bromine

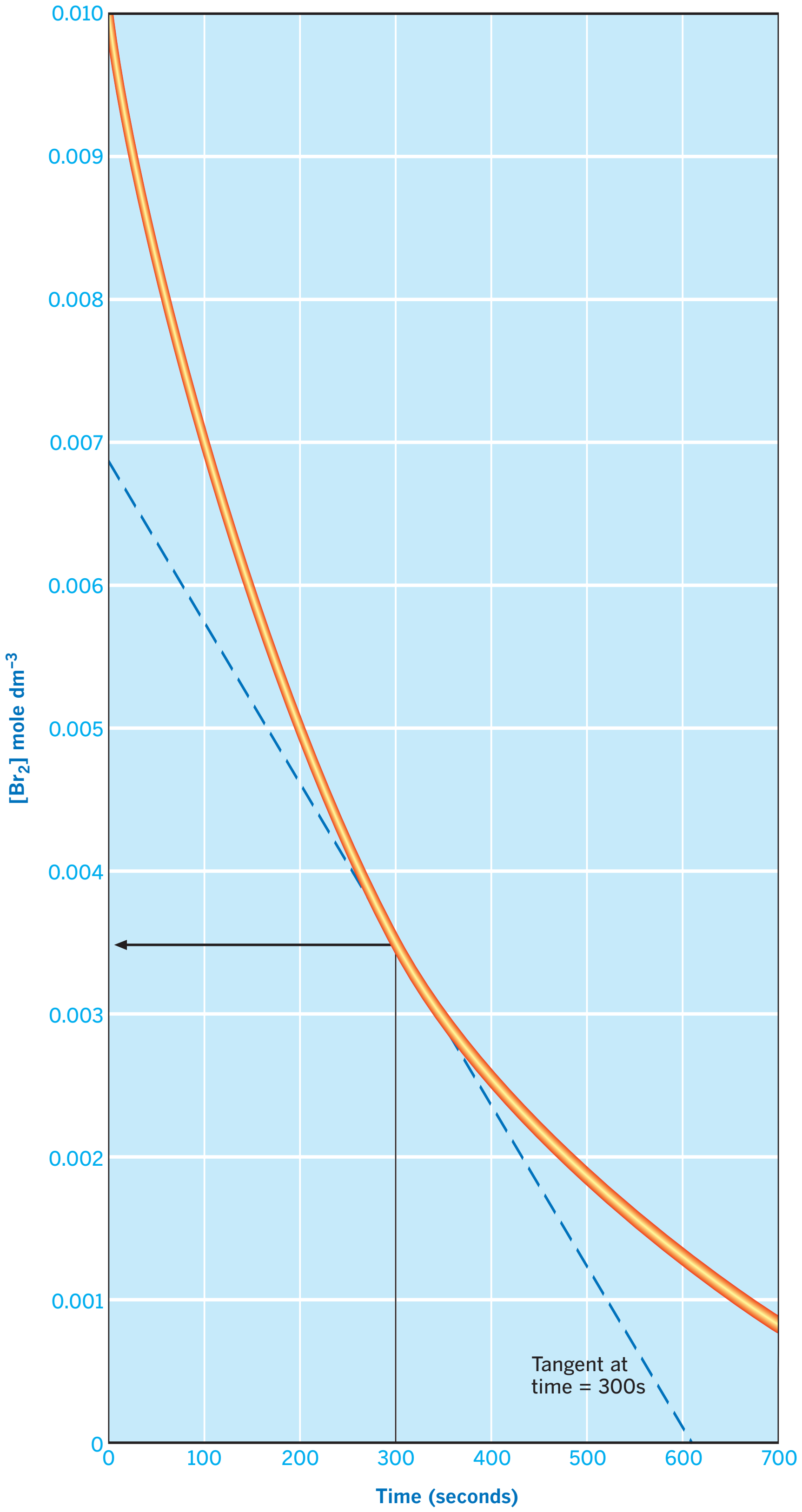

Rate of reaction vs. concentration

Graph to show the variation of reaction rate with bromine concentration

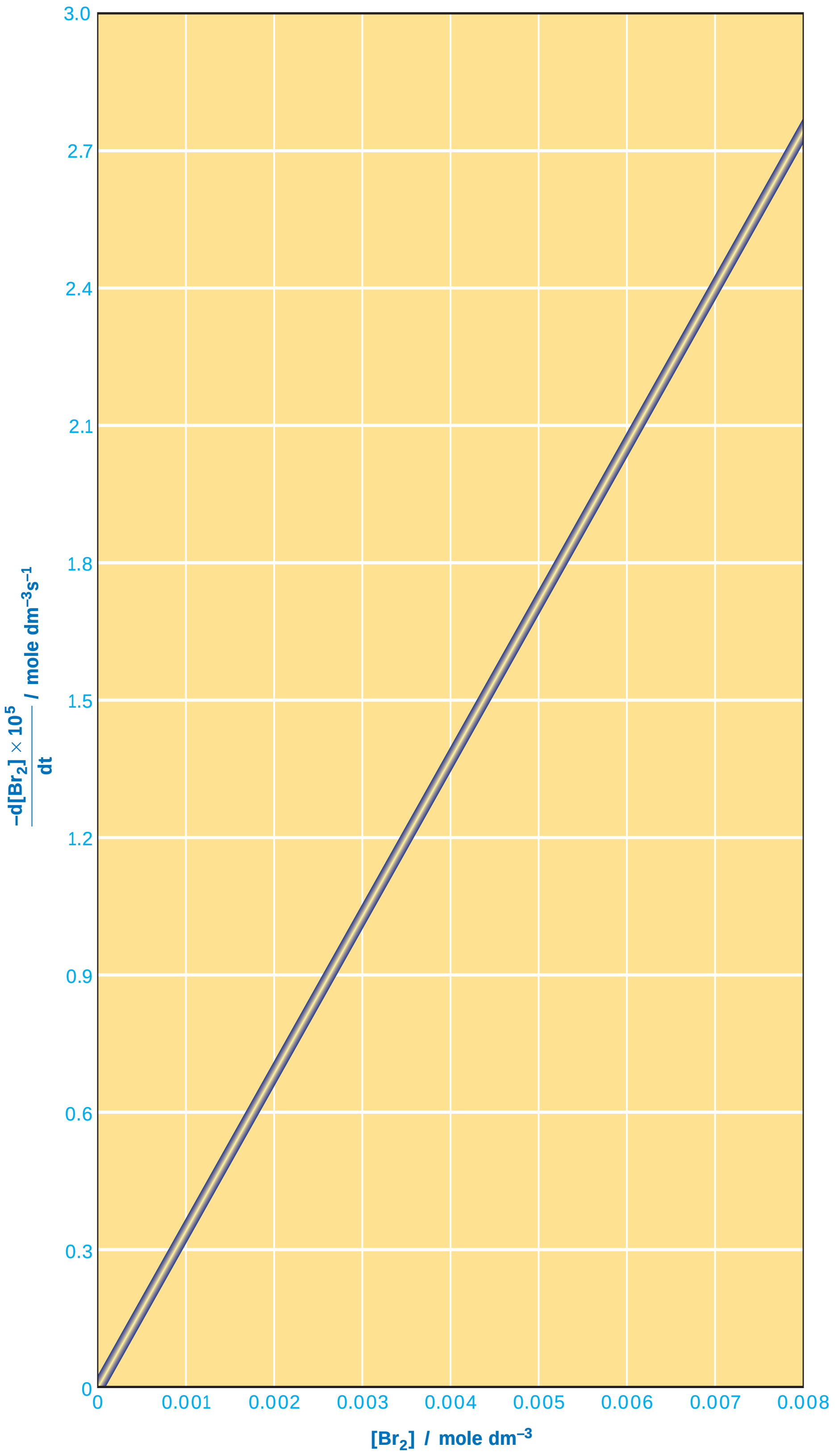

Key words

concentration
rate of reaction

Rate vs. concentration

- In order to draw a graph showing how the *rate of reaction* varies with bromine *concentration*, it is necessary to find the rate of reaction at different times and, therefore, different bromine concentrations.
- The graph shows that the rate of reaction is directly proportional to the bromine concentration.

  Reaction rate $\propto [Br_2]$, therefore,
  Rate of reaction = $k[Br_2]$ where k is a constant, known as the rate constant or the velocity constant for the reaction.
- This reaction is said to be first rate with respect to bromine since doubling the concentration of bromine doubles the rate of the reaction.
- Since rate of reaction = $k[Br_2]$, then to find the units of k:

  $$k = \frac{\text{rate of reaction}}{[Br_2]} = \frac{\text{mol dm}^{-3}\ \text{s}^{-1}}{\text{mol dm}^{-3}} = \text{s}^{-1}$$

  The unit of the rate constant, k, for first order reactions is s^{-1}.

© Diagram Visual Information Ltd.

Variation of reaction rate

Key words

concentration
product
rate of reaction
reactant

1 Clock technique

- *Rate of reaction* = $\frac{\text{change in concentration of a substance}}{\text{time}}$
- In order to monitor the progress of a reaction, we could measure the *concentration* of a *reactant* or a *product* at regular time intervals, say every 10 seconds.
- Strictly speaking, this would give us the average reaction rate during the 10 second period. By measuring the change in concentration over shorter and shorter time periods, we would obtain an increasingly more accurate estimate of the rate of reaction at any particular moment.
- Using a clock technique, the rate is obtained as the inverse of the time for a certain proportion of the reaction to occur. Provided the reaction has only gone a small way toward completion, the error is very small, but the error increases as the reaction moves further to completion.

2 Increasing concentration

- If doubling the concentration of a reactant has no effect on the rate of a reaction, then the reaction is said to be zero order with respect to the reactant. The rate equation is:
  rate = $k[\text{reactant}]^0 = k$
- If doubling the concentration of a reactant doubles the rate of a reaction, then the reaction is said to be first order with respect to the reactant. The rate equation is:
  rate = $k[\text{reactant}]$
- If doubling the concentration of a reactant quadruples the rate of a reaction, then the reaction is said to be second order with respect to the reactant. The rate equation is:
  rate = $k[\text{reactant}]^2$

© Diagram Visual Information Ltd.

1 Clock technique for measuring reaction rates

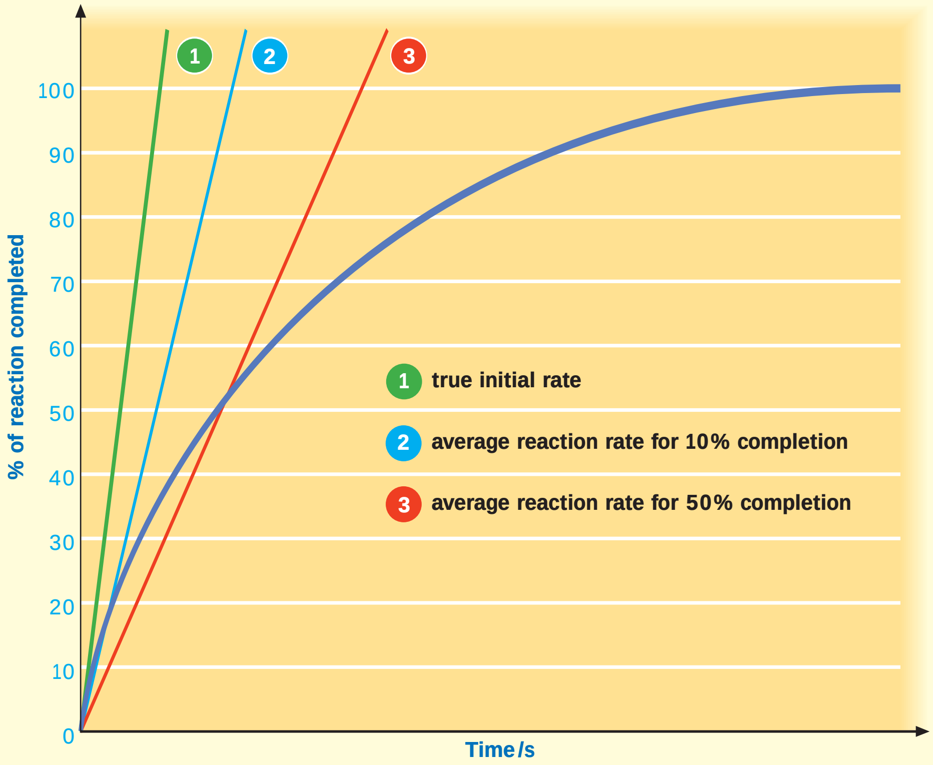

2 Increasing concentration

The variation of reaction rate with concentration for reactions which are zero, first, and second order

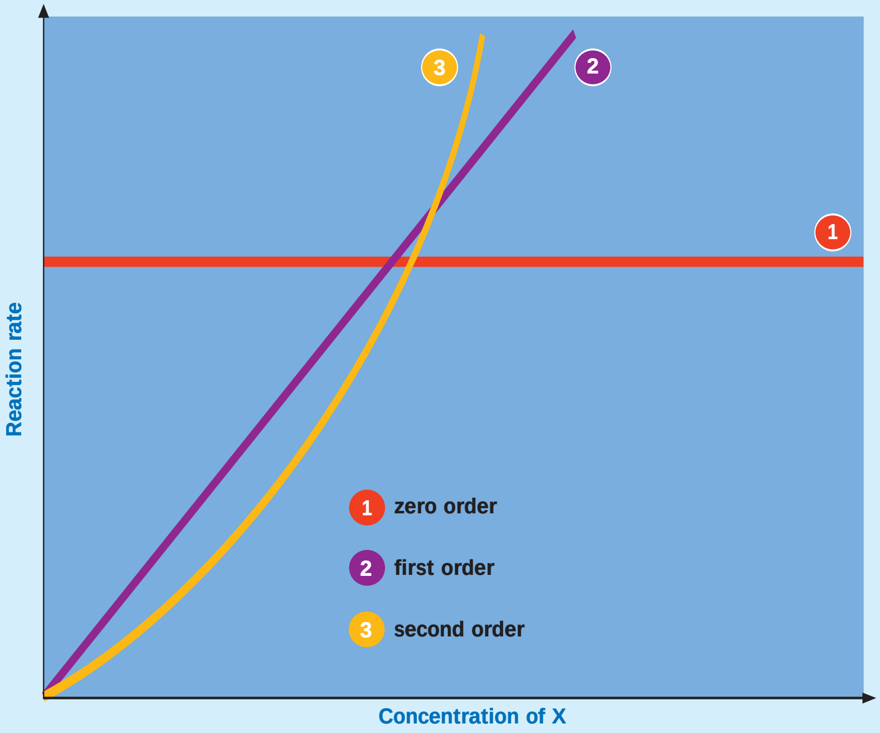

Rates of reaction: effect of temperature 1

The effect of temperature on different reactions

1 Most reactions

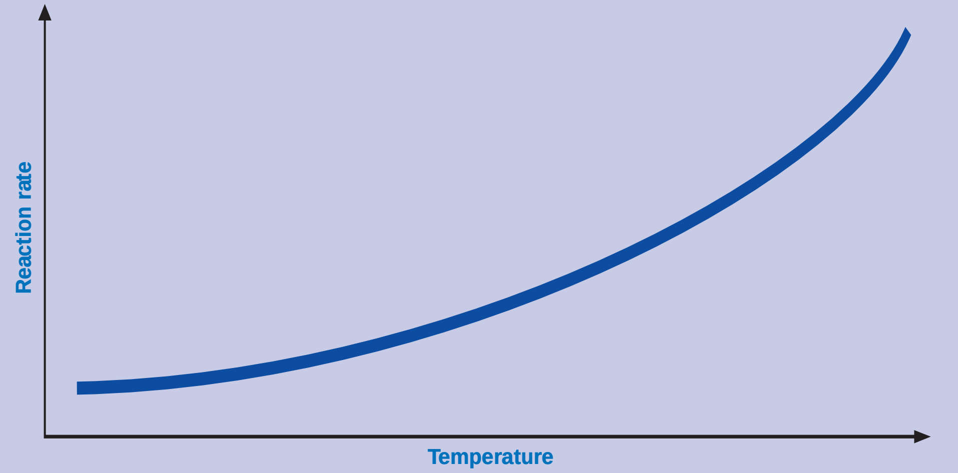

2 Enzyme-catalyzed reactions

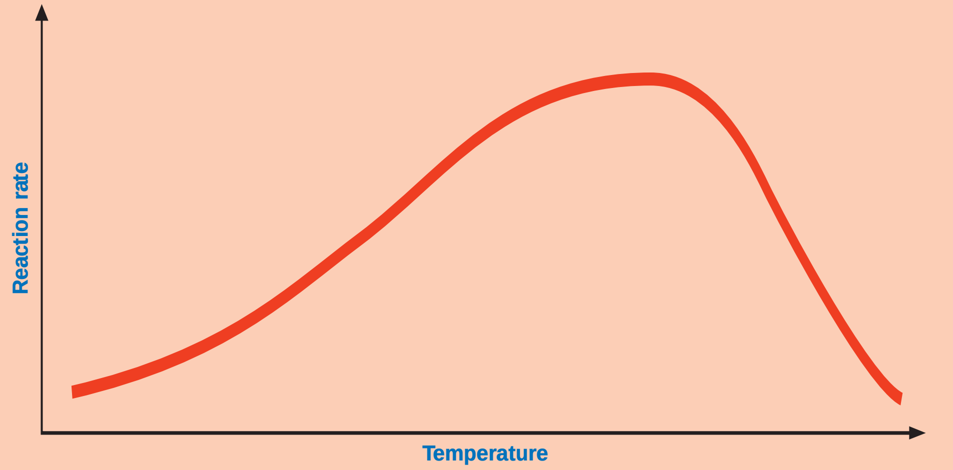

3 Explosive reactions

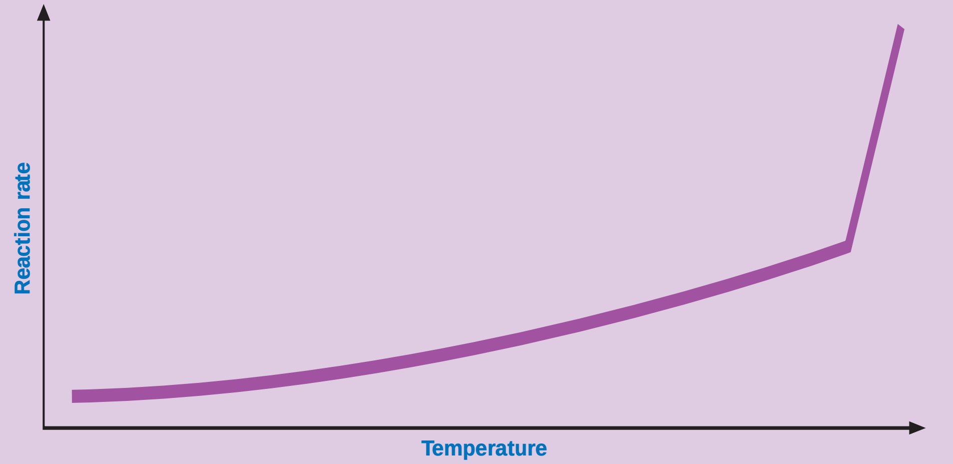

Key words

effective collision
enzyme
kinetic energy
polymer
rate of reaction

Effect of temperature

- When a substance is heated, its particles gain *kinetic energy* and move around more quickly. The frequency of collisions increases, and because the particles have a greater momentum, the frequency of *effective collisions* also increases. The result is an increase in the *rate of reaction*.

1 Most reactions

- In most chemical reactions, the rate of reaction increases steadily with rising temperature. It is for this reason that chemical reactions are often heated.

2 Enzyme-catalyzed reactions

- *Enzymes* catalyze chemical reactions with a high degree of specificity and efficiency. An enzyme molecule is a *polymer* composed of a long chain of amino acids that folds over on itself, giving it a particular shape. Reacting molecules, called the substrate, fit into this shape rather like a key in a lock.
- Up to a point, the rate of an enzyme-catalyzed reaction increases with rising temperature in the same way as most other reactions. However, after reaching an optimum temperature at which the activity of the enzyme is greatest, the reaction rate rapidly falls.
- Heating an enzyme causes its shape to change, and thus the enzyme ceases to be able to catalyze the reaction. It is said to be denatured.

3 Explosive reactions

- In an explosive reaction, the reaction rate increases with rising temperature up to some point where the reaction rate suddenly rises sharply.

© Diagram Visual Information Ltd.

Rates of reaction: effect of temperature 2

Key words

activation energy

1 Rate constant for reaction

- As a general rule of thumb, the rate of a reaction doubles for every 10 K rise in temperature. This would seem to suggest that there is an exponential relationship between rate and temperature.
- The exact relationship was proposed by the Swedish chemist Svante Arrhenius in 1889. The Arrhenius equation relates the rate constant (not the rate of reaction) to temperature.
- The equation can be expressed in a logarithmic form and in terms of log to the base 10. The latter form of the equation is the most useful for calculation purposes.

2 Plotting the Arrhenius constant

- The constants A and E_a for a given reaction can be obtained by plotting log k against 1/T: the temperature, T, must be expressed in kelvin.
  The slope of the graph is equal to E_a / 2.303R.
- The Arrhenius constant, A, can be obtained by substituting values for the slope (E_a / 2.303R), log k and T in the Arrhenius equation.
- The *activation energy*, E_a, can also be found from the slope of the graph.
  Slope = - E_a / 2.303 R
  The slope of the graph is negative, and its unit is K therefore
  E_a = 2.303 x R x slope
  The gas constant, R, = 0.008314 kJ K^{-1} mol^{-1} therefore
  E_a = 2.303 x 0.008314 x slope kJ mol^{-1}

© Diagram Visual Information Ltd.

1 Rate constant for reaction

$$k = Ae^{-Ea/RT}$$

$$\ln k = \ln A - E_a/RT$$

$$\log k = \log A - E_a/2.303RT$$

k = rate constant for the reaction

A = constant for the reaction (Arrhenius constant)

E_a = activation energy

R = gas constant

T = absolute temperature

2 Plotting the Arrhenius constant

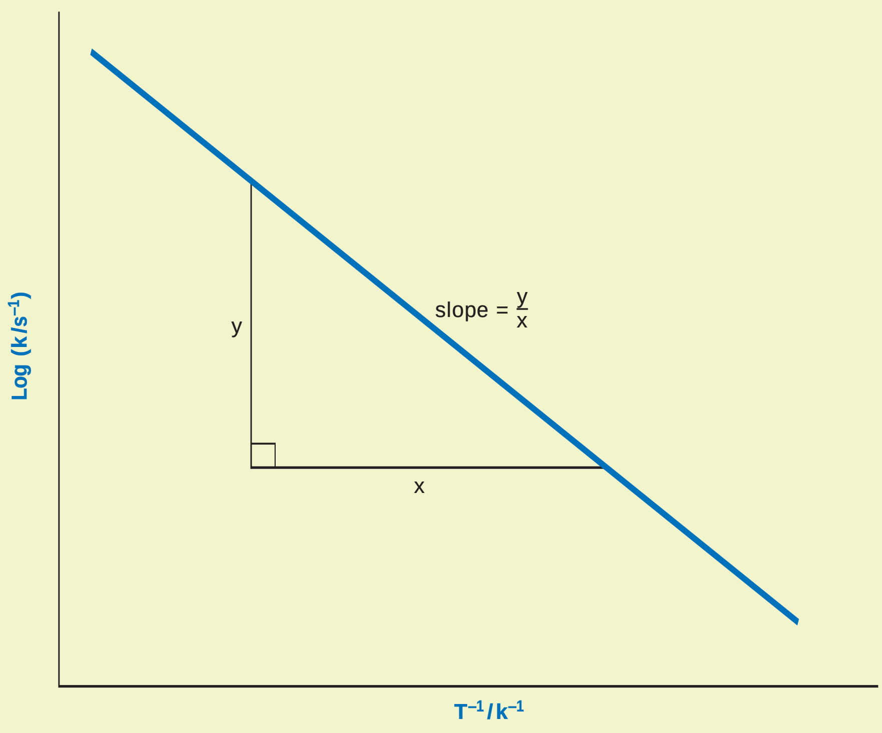

Exothermic and endothermic reactions

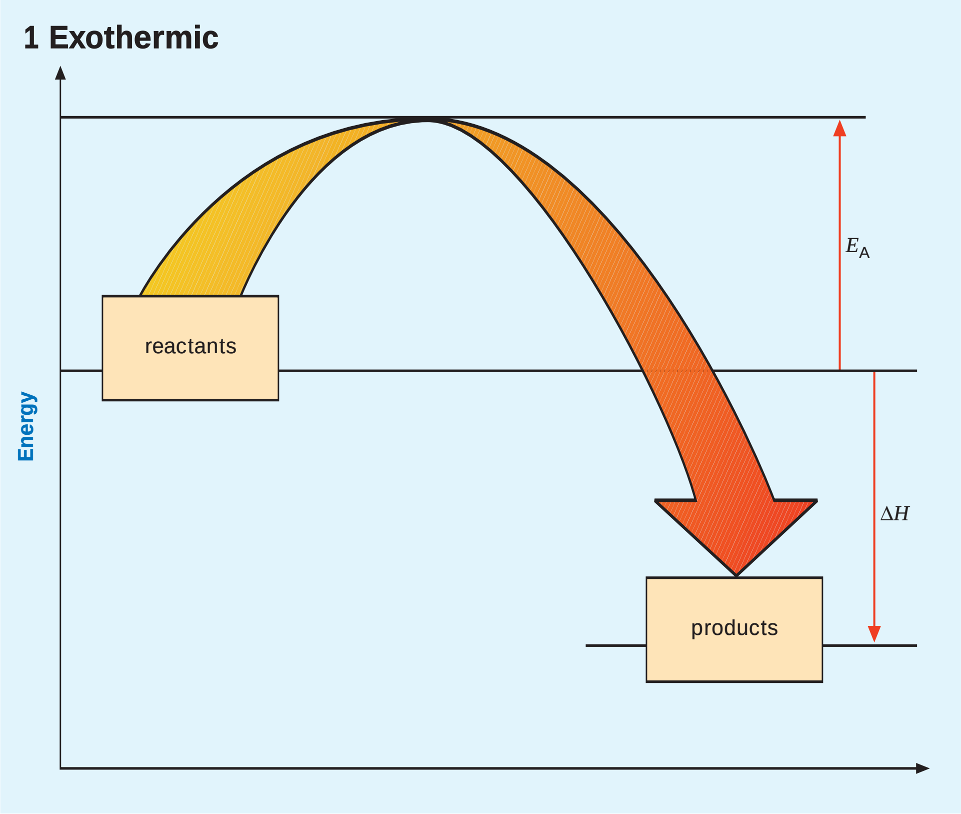
1 Exothermic

E_A = activation energy ΔH = heat of reaction

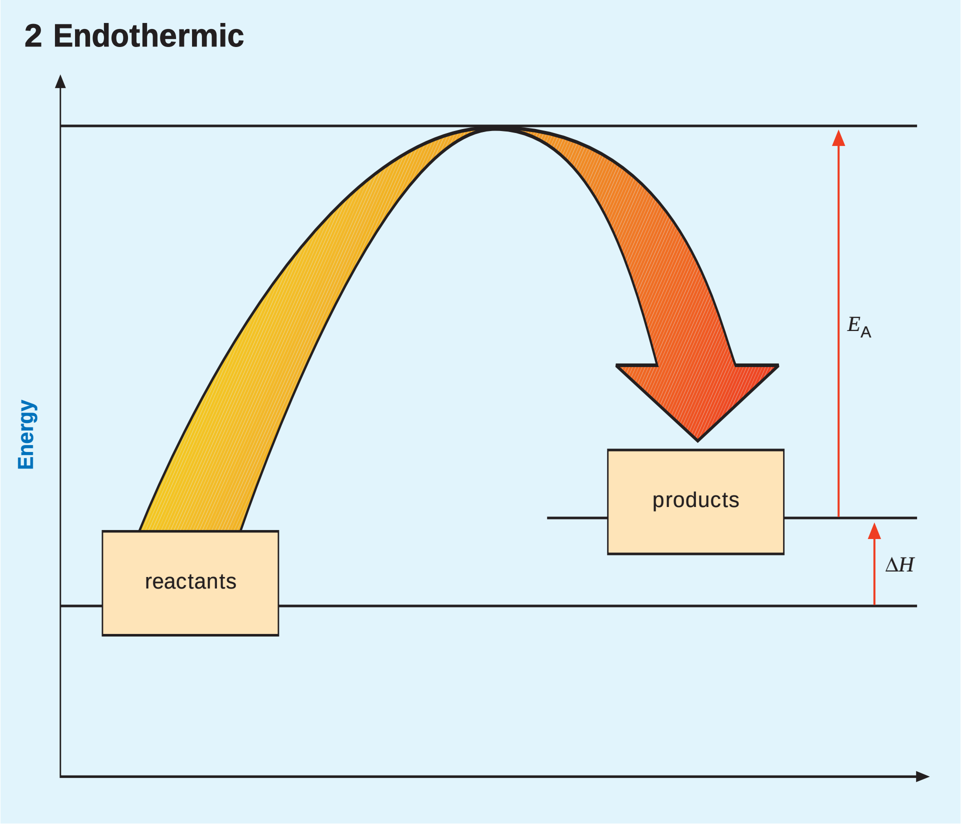
2 Endothermic

Key words

endothermic
enthalpy
exothermic
product
reactant

1 Exothermic

- In an *exothermic* reaction, energy is given out, and the temperature of the reaction mixture increases as the reaction proceeds. The *products* are at a lower energy than the *reactants*.
- The energy released is due to a decrease in the *enthalpy*, ΔH, of the system. Enthalpy is a measure of the stored heat energy of a substance. Therefore, ΔH is negative for an exothermic reaction.
- The following equation represents the combustion of methane in a good supply of air:

  $CH_4(g) + 2O_2(g) \rightarrow CO_2(g) + 2H_2O(g)$

  $\Delta H = -890 \text{ kJ mol}^{-1}$

  This is an exothermic reaction. 890 kJ of energy are released per mole of methane combusted.

2 Endothermic

- In an *endothermic* reaction, energy is taken in, and the temperature of the reaction mixture decreases as the reaction proceeds. The products are at a higher energy than the reactants.
- The energy taken in is due to an increase in the enthalpy, ΔH, of the system. Therefore, ΔH is positive for an endothermic reaction.
- The following equation represents the steam reforming of methane:

  $CH_4(g) + H_2O(g) \rightarrow 3H_2(g) + CO(g)$

  $\Delta H = +206 \text{ kJ mol}^{-1}$

  This reaction is an endothermic reaction. 206 kJ of energy are taken in per mole of methane reformed.

© Diagram Visual Information Ltd.

Average bond dissociation energies

Key words

dissociation
enthalpy

1 Average bond enthalpy

- Bond *dissociation* energy is the energy change when one mole of bonds is broken. It refers to a specific bond in a molecule. However, the exact value depends on the local environment of the bond. For example, if the C-H bonds in methane are broken one after another, each will have a different bond dissociation *enthalpy*:

$CH_4(g) \rightarrow CH_3(g) + H(g)$
$\Delta H = +425\ kJ\ mol^{-1}$
$CH_3(g) \rightarrow CH_2(g) + H(g)$
$\Delta H = +470\ kJ\ mol^{-1}$
$CH_2(g) \rightarrow CH(g) + H(g)$
$\Delta H = +416\ kJ\ mol^{-1}$
$CH(g) \rightarrow C(g) + H(g)$
$\Delta H = +335\ kJ\ mol^{-1}$

- For this reason, in a molecule composed of more than one atom, it is more useful to know the average amount of energy needed to break a particular bond.

2 Estimating enthalpy change

- The table at right utilizes the complete combustion of propane to illustrate how bond enthalpies can be used to estimate the enthalpy change in a reaction.
- 6,488 kJ mol^{-1} of total energy is taken in to break the bonds.
- 8,542 kJ mol^{-1} of total energy is given out when the bonds are formed.
- The enthalpy change when 1 mole of propane is completely combusted is 6,488 – 8,542 = 2,054 kJ mol^{-1}.

1 Average bond enthalpy

Bond	Average bond enthalpy / kJ mol^{-1}	Bond	Average bond enthalpy / kJ mol^{-1}
C–C	347	C=O	805
C=C	612	H–Cl	432
C–Cl	346	H–H	436
C–H	413	N–H	391
C–N	286	O–H	464
C–O	336	O=O	498

2 Estimating the enthalpy change in a reaction.

Complete combustion of propane.

$C_3H_8(g) + 5O_2(g) \rightarrow 3CO_2(g) + 4H_2O(g)$

Energy is taken in to break bonds:

Bond	Average bond dissociation energy / kJ mol^{-1}	Number of bonds	Energy taken in / kJ mol^{-1}
C–H	413	8	3,304
C–C	347	2	694
O=O	498	5	2,490
Total energy taken in			6,488

Energy is given out when bonds are formed:

Bond	Average bond dissociation energy / kJ mol^{-1}	Number of bonds	Energy taken in / kJ mol^{-1}
C=O	805	6	4,830
O–H	464	8	3,712
Total energy taken in			8,542

The enthalpy change when 1 mole of propane is completely combusted is 6,488 – 8,542 = 2,054 kJ mol^{-1}.

© Diagram Visual Information Ltd.

Catalysts: characteristics

1 Characteristics of catalysts

Stoichiometry	The overall stoichiometry of a reaction is unaltered.
Specificity	Catalysts may alter the rate of one reaction but have no effect on others.
Reaction mechanism	A catalyst provides an alternative reaction pathway for a reaction to take place.
Chemical involvement	A catalyst is chemically involved in a reaction. It is consumed during one step and regenerated in another. A catalyst does not undergo a net chemical change, but it may change its physical form.
Equilibrium	A catalyst speeds up the rates of both forward and backward reactions, and this speeds up the rate at which equilibrium is attained.
Yield	A catalyst does not alter the yield of a reaction.

2 Increasing reaction rate

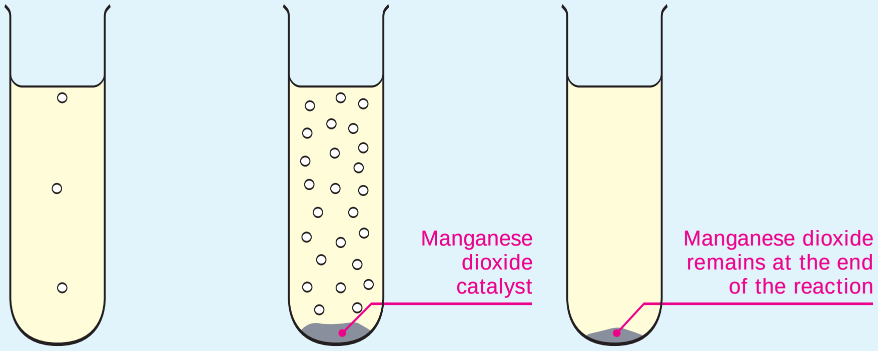
Manganese dioxide catalyst
Manganese dioxide remains at the end of the reaction

3 Distribution of the kinectic energies of reacting particles and the activation energies for catalyzed and uncatalyzed reactions

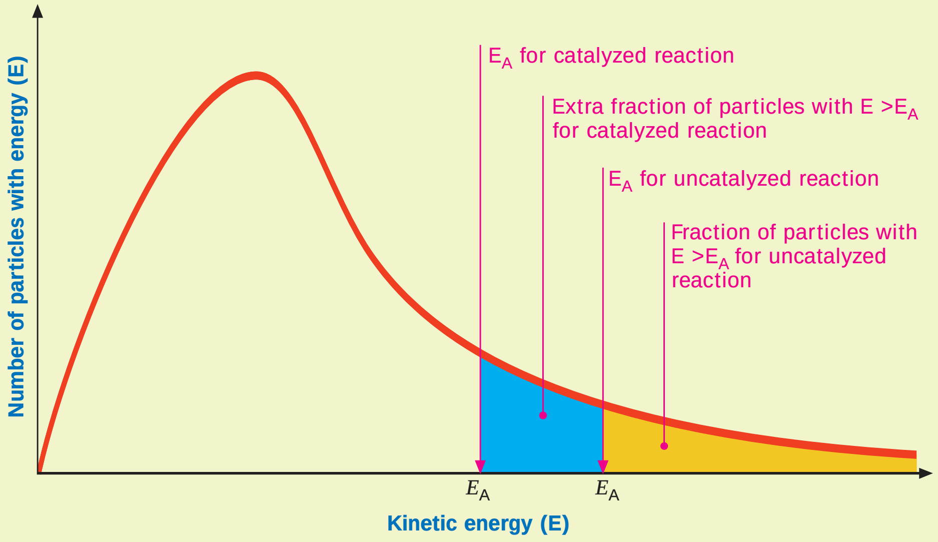

Key words

activation energy
active site
catalyst
effective collision

Catalysts

- A *catalyst* is a substance that alters the rate of a chemical reaction but remains chemically unchanged by it.

1 Characteristics of catalysts

- Catalysts may be classified as homogenous or heterogeneous. Homogenous catalysts are in the same phase (solid, liquid, or gas) as the reactants; heterogeneous catalysts are in a different phase.
- A large number of reactions are catalyzed on the surface of solid catalysts. The surface provides *active sites* where reactions can occur. Thus, an increase in the surface area will increase the effect of the catalyst.

2 Increasing reaction rate

- Hydrogen peroxide decomposes very slowly on its own to form water and oxygen gas:
  $2H_2O_2(aq) \rightarrow 2H_2O(l) + O_2(g)$
  The rate of this reaction is greatly increased by adding manganese dioxide, MnO_2.
- Manganese dioxide acts as a catalyst and remains unchanged after all of the hydrogen peroxide has decomposed.

3 Activation energies

- A catalyst lowers the minimum energy, or *activation energy* (E_A), required for a reaction to occur. The frequency of *effective collisions* is, therefore, increased, resulting in an increase in the rate of a reaction.

© Diagram Visual Information Ltd.

Catalysts: transition metals

Key words

catalyst
equilibrium
exothermic
Le Chatelier's principle
oxidation
oxidation state
reduction
transition metals
vanadium

1 Reaction catalyzed

- *Catalysts* are often *transition metals* or transition metal compounds. Transitional metals are useful as catalysts because or their ability to exist in different *oxidation states*.

2 V_20_5 as catalyst

- The contact process is an important step in the manufacture of sulfuric acid (see pages 85 and 86). Sulfur dioxide is oxidized to sulfur trioxide in the presence of a *vanadium* pentoxide, V_2O_5, catalyst.
- This reaction involves the *reduction* and subsequent *oxidation* of the catalyst. In the reduction reaction, the oxidation state of vanadium changes from +5 to +4. In the oxidation reaction, it changes back from +4 to +5.

3 Iron as catalyst

- The Haber process for the manufacture of ammonia uses finely divided iron as the catalyst (see pages 74 and 75):

$$N_2 + 3H_2 \overset{Fe(s)}{\rightleftharpoons} 2NH_3$$

$$\Delta H = -92\ kJ\ mol^{-1}$$

- This reaction is *exothermic*. According to *Le Chatelier's principle*, a low temperature would produce more ammonia in the *equilibrium* mixture, but it would take longer to reach equilibrium.
- The catalyst does not alter the yield of ammonia in the equilibrium mixture, but it does increase the speed with which equilibrium is attained. Using a catalyst, a reasonable rate of reaction is achieved at a lower temperature than would otherwise be the case.

© Diagram Visual Information Ltd.

1 Transition metals and reaction catalyzed

Transition metal /compound	Reaction catalyzed
$TiCl_3$	polymerization of ethene to poly(ethene)
V_2O_5	contact process in production of sulfuric acid
Fe	Haber process on production of ammonia
Ni	hydrogenation of alkenes in hardening of vegetable oils
Cu	oxidation of ethanol to ethanal
Pt	oxidation of ammonia in manufacture of nitric acid

2 Vanadium oxide as catalyst in contact process

$$2SO_2(g) + O_2(g) \overset{V_2O_5(s)}{\rightleftharpoons} 2SO_3(g)$$

$$SO_2 + V_2O_5 \longrightarrow SO_3 + V_2O_4$$

$$2V_2O_4 + O_2 \longrightarrow 2V_2O_5$$

3 Iron as catalyst in Haber process

(Energy profiles for the reaction $N_2 + 3H_2 \overset{Fe(s)}{\rightleftharpoons} 2NH_3$)

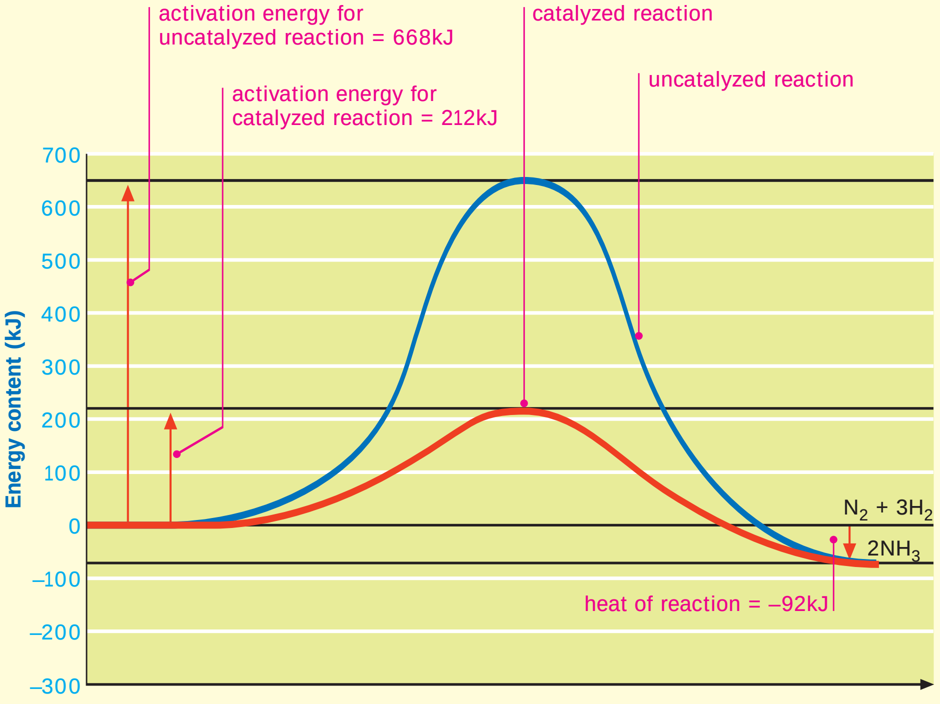

Oxidation and reduction

1 Oxygen

Oxidation is the addition of oxygen to a substance

Reduction is the removal of oxygen from a substance

2 Hydrogen

Oxidation is the removal of hydrogen from a substance

Reduction is the addition of hydrogen to a substance

3 Modern definition

Oxidation is the loss of electrons from a substance

Reduction is the gain of electrons by a substance

4 Redox reaction

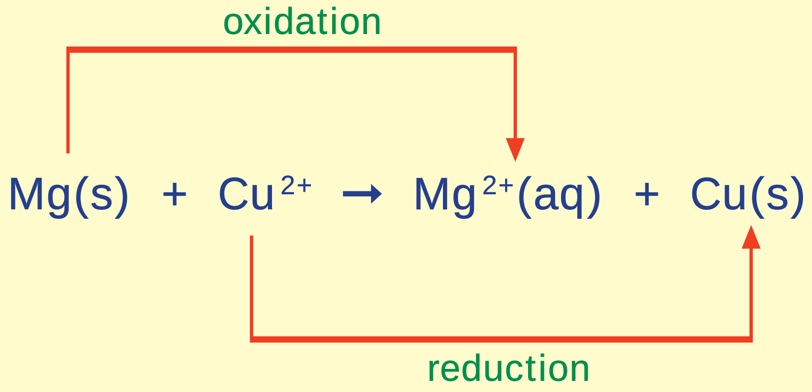

Key words

oxidation
oxidation state
redox reaction
reduction

Evolving definition

- Over time, scientists have extended the definitions of *oxidation* and *reduction*.

1 Oxygen

- Historically the terms oxidation and reduction were applied to reactions involving either the addition or the removal of oxygen. For example:

$2Cu(s) + O_2 \rightarrow 2CuO(s)$
copper is oxidized
$Fe_2O_3(s) + 3CO(g) \rightarrow 2Fe(s) + 3CO_2(g)$
iron is reduced

2 Hydrogen

- The terms were extended to include the removal or addition of hydrogen:

$CH_3\text{-}CH_3(g) \rightarrow CH_2{=}CH_2(g) + H_2(g)$
ethane is oxidized
$CH_3COH(l) + H_2(g) \rightarrow CH_3CH_2OH(l)$
ethanal is reduced

3 Modern definition

- The terms oxidation and reduction are now used more widely to describe changes in *oxidation state*:

$Cu(s) \rightarrow Cu^{2+}(aq) + 2e^-$
copper is oxidized to copper(II)
$Fe^{3+}(aq) + e^- \rightarrow Fe^{2+}(aq)$
iron(III) is reduced to iron(II)

- This definition covers all of those reactions involving the gain or loss of oxygen and other reactions that do not involve oxygen.

4 Redox reaction

- Reactions that involve a reduction must also involve an oxidation. If one reactant is reduced, then another must be oxidized. Such reactions are described as *redox reactions*.

© Diagram Visual Information Ltd.

Redox reactions 1

Key words

displacement reaction
oxidation
redox reaction
reduction

Redox reactions

- *Reduction* and *oxidation* reactions always occur together and are collectively referred to as *redox reactions*.

1 Oxidation and reduction

- When magnesium is heated in air, it forms magnesium oxide:
  $2Mg(s) + O_2(g) \rightarrow 2MgO(s)$
- Magnesium atoms are oxidized to magnesium ions by losing two electrons.
- Oxygen atoms are reduced to oxide ions by gaining two electrons.
- This is true of all metals when they are converted to metal oxides.

2 Electron transfer

- A more reactive metal displaces the ions of a less reactive metal from a solution of its salts. This type of reaction is called a *displacement reaction*:
  $Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)$
  Zinc atoms are oxidized to zinc ions by losing two electrons.
- Copper ions are reduced to copper atoms by gaining two electrons.

3 Balancing redox reactions

- In balancing redox reactions, the electrons lost must equal the electrons gained.
- In the example at right, bromine (a) is gaining two electrons and iron (b) is losing 1 electron.
- In order to balance the equation, the entire reaction has to be multiplied by 2 (c).
- The result is a balanced equation (d).

© Diagram Visual Information Ltd.

1 Redox reactions: oxidation and reduction

$$2Mg + O_2 \rightarrow 2Mg^{2+} + 2O^{2-}$$

$$4Na + O_2 \rightarrow 2(Na^2)_2O^{2-}$$

When metals react with oxygen they form oxides

$$\begin{tikzcd} 2Mg \arrow[r] & 2Mg^{2+} + \textcircled{4e^-} \arrow[dl] \\ O_2 + \textcircled{4e^-} \arrow[r] & 2O^{2-} \end{tikzcd}$$

The metal is oxidized and the metal is reduced. The oxygen takes the electrons given up by the metal

2 Electron transfer in redox reactions

$$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)$$

When powered zinc is added to copper sulfate (II) solution, an exothermic reaction occurs

$$\begin{tikzcd} Zn(s) \arrow[r] & Zn^{2+}(aq) + \textcircled{2e^-} \arrow[dl] \\ Cu^{2+}(aq) + \textcircled{2e^-} \arrow[r] & Cu(s) \end{tikzcd}$$

Redox equations for the reaction

3 Balancing redox reactions

$$Br_2 + 2e^- \rightarrow 2Br^-$$

(a) The oxidizing agent is Br^2

$$Fe^{2+} \rightarrow Fe^{3+} + e^-$$

(b) The reducing agent is Fe^{2+}

$$\begin{tikzcd} 2Fe^{2+} \arrow[r] & 2Fe^{3+} + \textcircled{2e^-} \arrow[dl] \\ Br_2 + \textcircled{2e^-} \arrow[r] & 2Br^- \end{tikzcd}$$

(c) Redox reaction

$$2Fe^{2+} + Br_2 \rightarrow 2Fe^{3+} + 2Br^-$$

(d) Balance equation

Redox reactions 2

1 The reaction of metals with non-metals

$Fe + S \rightarrow Fe^{2+}S^{2-}$

Iron and sulfur

$Fe \rightarrow Fe^{2+} + \boxed{2e^-}$

$S + \boxed{2e^-} \rightarrow S^{2-}$

Redox equation

$2Fe + 3Cl_2 \rightarrow 2FeCl_3$

Iron and chlorine

$2Fe \rightarrow 2Fe^{3+} + \boxed{6e^-}$

$3Cl_2 + \boxed{6e^-} \rightarrow 6Cl^-$

Redox equation

2 The reaction of metals with water

$Ca + 2H_2O \rightarrow Ca^{2+}(OH^-)_2 + H_2$

Calcium and water

$Ca \rightarrow Ca^{2+} + \boxed{2e^-}$

$2H_2O + \boxed{2e^-} \rightarrow 2OH^- + H_2$

Redox equation

3 The reaction of metals with acids

$Zn(s) + 2H^+(aq) \rightarrow Zn^{2+}(aq) + H_2\uparrow$

The reaction of zinc

$Zn \rightarrow Zn^{2+} + \boxed{2e^-}$

$2H^+ + \boxed{2e^-} \rightarrow H^2$

Redox equation

Key words

redox reaction

1 Metals with non-metals

- Iron undergoes *redox reactions* with non-metals.
  Iron and sulfur:
  $Fe(s) + S(s) \rightarrow Fe^{2+}S^{2-}(s)$
  Iron atoms are oxidized to iron(II) ions by losing two electrons.
  Sulfur is reduced to sulfide ions by gaining two electrons.
  Iron and chlorine:
  $2Fe(s) + 3Cl_2(s) \rightarrow 2Fe^{3+}Cl^-_3(s)$
  Iron atoms are oxidized to iron(III) ions by losing three electrons.
  Chlorine is reduced to chloride ions by gaining one electron.

2 Metals with water

- Metals are oxidized when they react with water.
  Metal + water → metal hydroxide + hydrogen
  $Ca(s) + 2H_2O(l) \rightarrow Ca(OH)_2(aq) + H_2(g)$
  $H_2O(l) \rightleftharpoons H^+(aq) + OH^-(aq)$
  $Ca(s) + 2H^+(aq) \rightarrow Ca^{2+}(aq) + H_2(g)$
  In the example at left, calcium atoms are oxidized to calcium ions by the loss of two electrons. Hydrogen ions are reduced to hydrogen atoms by gaining one electron.

3 Metals with acids

- Metals are oxidized when they react with acids.
  Metal + acid → metal salt + hydrogen
  $Zn(s) + 2HCl(aq) \rightarrow ZnCl_2(aq) + H_2(g)$
  $Zn(s) + 2H^+(aq) \rightarrow Zn^{2+}(aq) + H_2(g)$
  In the example at left, zinc atoms are oxidized to zinc ions by the loss of two electrons. Hydrogen ions are reduced to hydrogen atoms by gaining one electron.

© Diagram Visual Information Ltd.

Key words

reactivity series
redox reaction

1 Electron transfer in redox reactions

- The movement of electrons during *redox reactions* can be demonstrated using a simple cell consisting of two metals rods, suspended in solutions of their salts, connected by a wire and a salt bridge.
- At the zinc rod, zinc atoms lose two electrons to become zinc ions. The electrons pass along the wire and through the bulb to the copper rod. The zinc ions pass into solution.
- At the copper rod, copper ions gain two electrons to become copper atoms. The copper ions come out of solution and are deposited as copper metal on the copper rod.
- Electric current passes through the wire in the external circuit as a flow of negatively charged electrons.
- Ions flow through the salt bridge: positive ions from the zinc sulfate solution to the copper sulfate, and negative ions in the opposite direction from the copper sulfate solution to the zinc sulfate solution.
- Electric current passes through ionic solutions as a flow of positively charged and negatively charged ions.

2 Reaction equations

- Zinc is higher than copper in the *reactivity series*. Zinc atoms are oxidized to zinc ions, while copper ions are reduced to copper atoms.
- When any two metals are placed in a cell, the direction of electrons in the external circuit depends on their reactivities. The metal that is higher in the reactivity series will be oxidized, while the ions of the metal that is lower in the reactivity series will be reduced.

© Diagram Visual Information Ltd.

Demonstrating redox reactions

1 Electron transfer in redox reactions

Experimental set-up

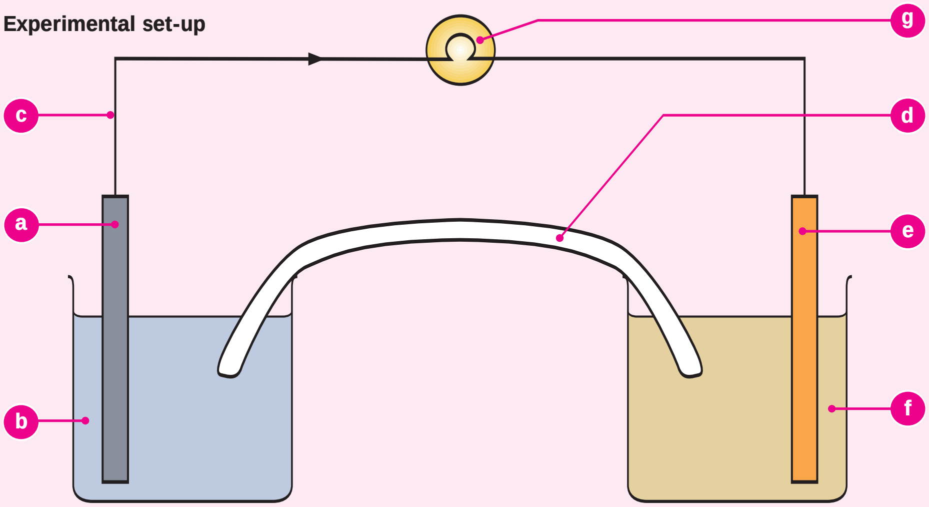

Movement of charge around the circuit

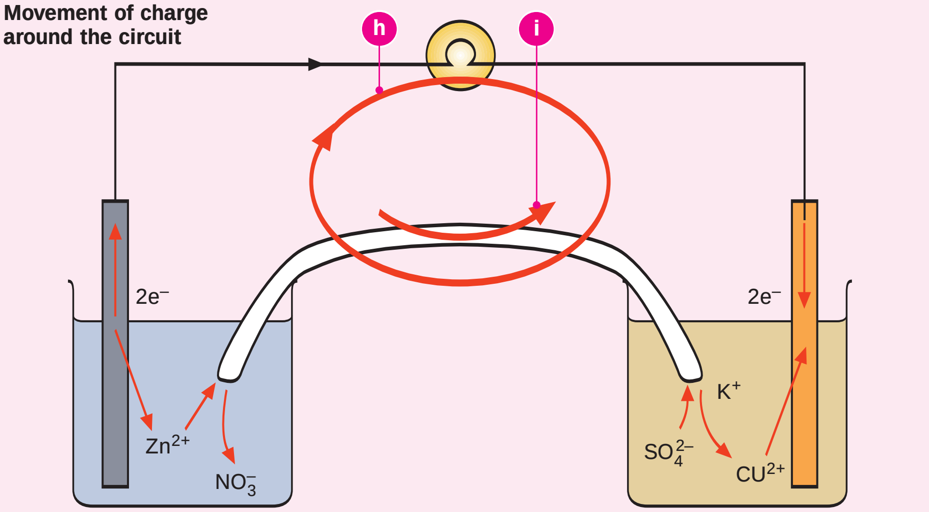

a zinc rod
b zinc sulfate solution
c electron flow
d filter paper soaked in potassium nitrate as a salt bridge
e copper rod
f copper sulfate solution
g small light bulb
h movement of negative charge (electrons and anions)
i movement of positive charge (cations)

2 Reaction equations

$$Zn(s) + Cu^{2+} \rightarrow Zn^{2+}_{(aq)} + Cu(s)$$

$$Zn(s) \rightarrow Zn^{2+}_{(aq)} + 2e^-$$

$$Cu^{2+}_{(aq)} + 2e^- \rightarrow Cu(s)$$

Copper ions are reduced to a deposit of red-brown

Assigning oxidation state

1 Oxidation state

Component	Oxidation state
uncombined elements	0
Group 2 metals in compounds	+2
Group 1 metals in compounds	+1
combined hydrogen except in metal hydrides	+1
combined hydrogen in metal hydrides	–1
combined halogens	–1
combined oxygen	–2

2 Tetrachlorocuprate ion

$$CuCl_4^{2-}$$

the tetrachlorocuprate ion contains the transition metal copper

3 Manganate ion

$$MnO_4^{-}$$

the manganate ion contains the transition metal manganese

4 Dichromate ion

$$Cr_2O_7^{2-}$$

the dichromate ion contains the transition metal chromium

Key words

oxidation state
transition metals

1 Oxidation state

- The ability of *transition metals* to exhibit different *oxidation states* in different compounds is central to the behavior of these elements.
- The oxidation state of simple ions is given by the charge they carry, e.g.:
  Na^+ has an oxidation state of +1
  O^{2-} has an oxidation state of -2
- The situation is more complicated in a complex ion. The oxidation state of the central atom in a complex ion is the charge that the ion would have if it were a simple ion. This is found by adding the oxidation states of the various components in the complex ion.

2 Tetrachlorocuprate ion

- Total oxidation number due to chlorine = 4 x -1 = -4.
- Overall charge on the ion = -2.
- Oxidation state of the central copper atom = -2 – (-4) = +2.
- This complex ion is more correctly called the tetrachlorocuprate(II) ion.

3 Manganate ion

- Total oxidation number due to oxygen = 4 x -2 = -8.
- Overall charge on the ion = -1.
- Oxidation state of the central manganese atom = -1 – (-8) = +7.
- This complex ion is more correctly called the manganate(VII) ion.

4 Dichromate ion

- Total oxidation number due to oxygen = 7 x -2 = -14.
- Overall charge on the ion = -2.
- Total oxidation state of the two central chromium atoms = -2 – (-14) = +12.
- Oxidation state of each chromium atom = +12 / 2 = +6.
- This complex ion is more correctly called the dichromate(VI) ion.

© Diagram Visual Information Ltd.

The allotropes of carbon: diamond and graphite

Key words

allotrope
carbon
diamond
fullerenes
graphite

Carbon allotropes

- *Carbon* exists in three *allotropes: diamond, graphite,* and *fullerenes*.

1 Diamond

- In diamond, each carbon atom is covalently bonded to four other carbon atoms.
- The four bonds are directed toward the corners of a pyramid or tetrahedron, and all bonds are the same length, 0.154 nm. The angle between any two bonds is 109.5°.
- All four of the outer electrons on the carbon atom form bonds with other carbon atoms so there are no mobile electrons. Diamond does not, therefore, conduct electricity.
- Diamond has a rigid structure and is very hard.

2 Graphite

- The carbon atoms in graphite are arranged in layers consisting of interlocking hexagons in which each carbon atom is covalently bonded to three other carbon atoms. The length of the bond is 0.141 nm, and the angle between bonds is 120°.
- The fourth outer electron on each carbon atom forms bonds with adjacent layers. The bond length is much greater than between carbon atoms within a layer.
- The electrons between the layers are mobile; therefore, graphite conducts electricity. Also, the layers are able to slide over each other relatively easily.
- Graphite is soft.

© Diagram Visual Information Ltd.

Diamond

Bond angle 109.5°
Bond length 0.154 nm

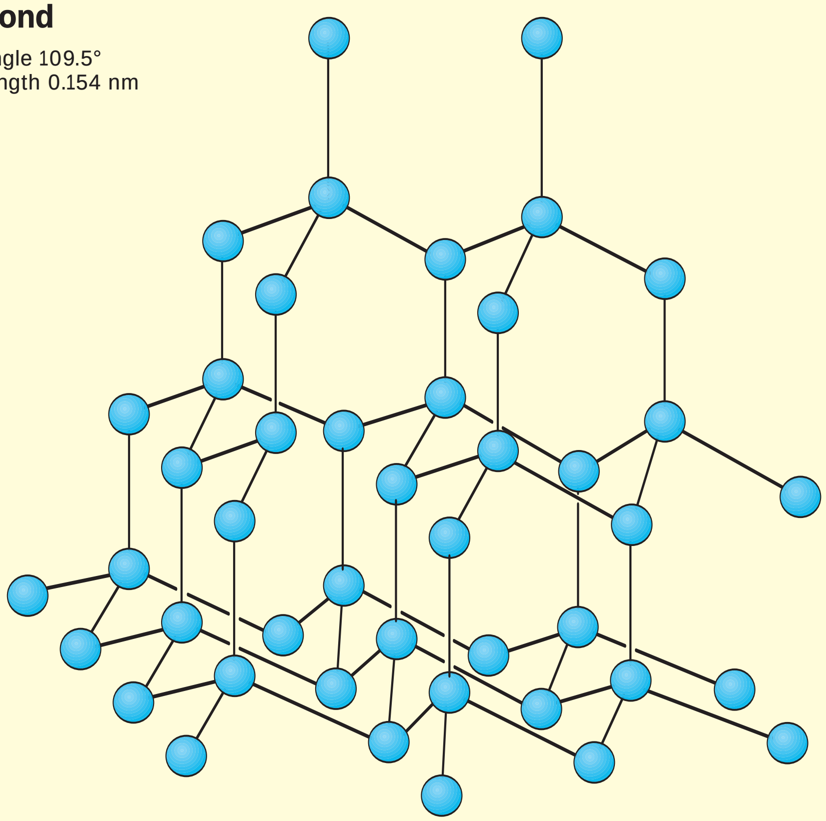

Graphite

Bond angle 120°
Bond lengths
– in layers 0.141 nm
– between layers
0.335 nm

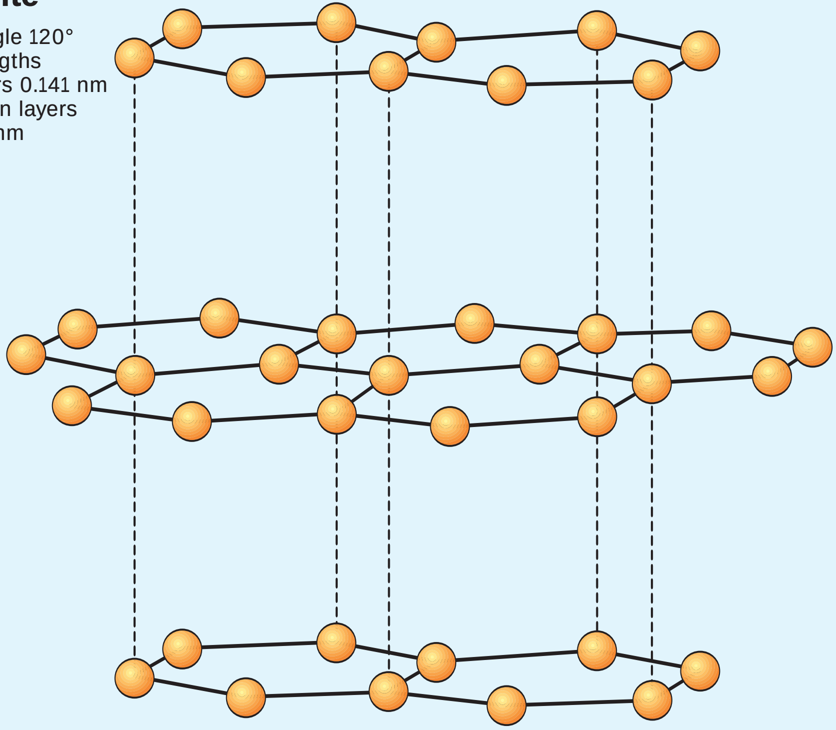

The allotropes of carbon: fullerenes

Buckyball

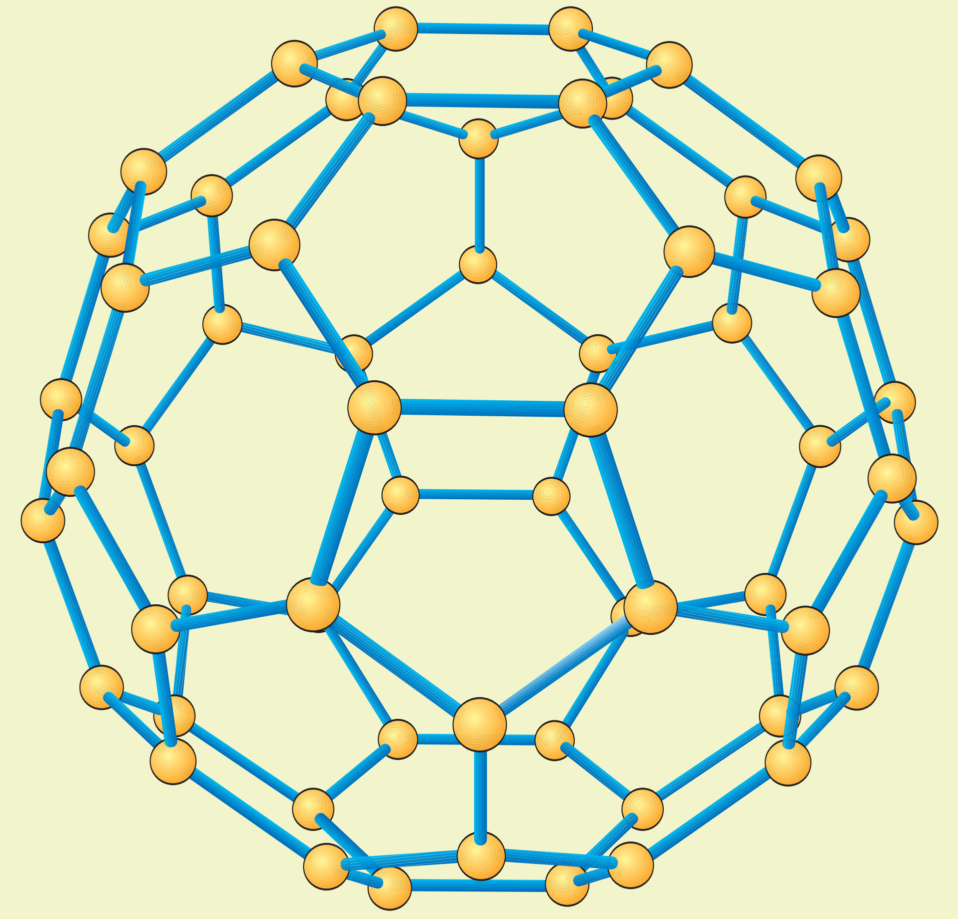

Nanotube

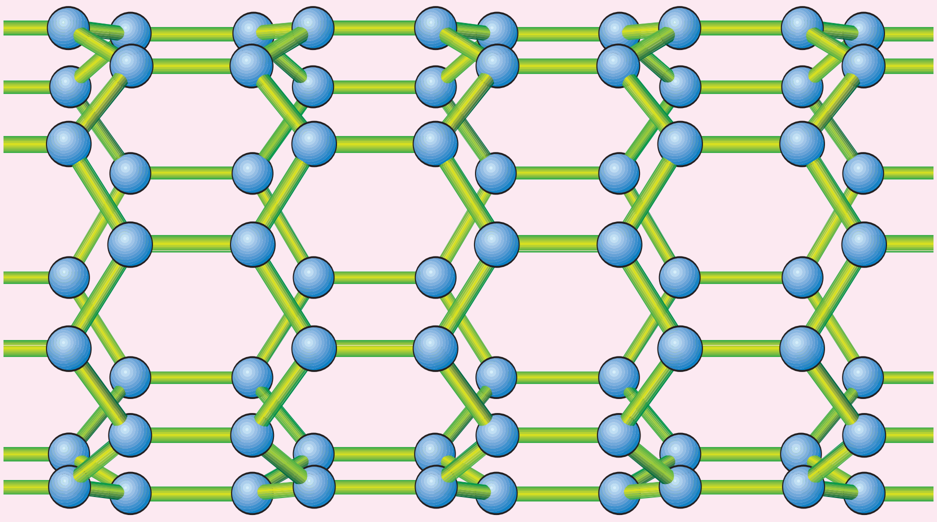

Key words

allotrope	*fullerenes*
buckyball	*graphite*
carbon	*nanotube*

Fullerenes

- *Fullerenes* are *allotropes* of *carbon* in the form of a hollow sphere or tube. Spherical fullerenes are sometimes called *buckyballs*, and cylindrical fullerenes are called *nanotubes*.
- Because the allotrope was only discovered in the late twentieth century, its physical and chemical properties are still being studied.
- Fullerenes are not very reactive and are only slightly soluble in many solvents. They are the only known allotrope of carbon that can be dissolved.

Buckminsterfullerene

- This form of carbon is composed of 60 carbon atoms bonded together in a polyhedral structure composed of pentagons and hexagons. The molecules are made when an electric arc is struck between *graphite* electrodes in an inert atmosphere. This method also produces small amounts of other fullerenes that have less symmetrical molecular structures, such as C_{70}.
- Buckminsterfullerene was first identified in 1985 and named after the architect Richard Buckminster Fuller because of the resemblance of its structure to the geodesic dome.
- The substance is a yellow crystalline solid that is soluble in benzene, an organic solvent.
- It is possible to trap metal ions within the C_{60} sphere. Some of these structures are semiconductors.

Nanotubes

- Nanotubes, first identified in 1991, are long thin cylinders of carbon closed at either end with caps containing pentagonal rings.
- Nanotubes have a very broad range of electronic, thermal, and structural properties that change depending on the kind of nanotube (defined by its diameter, length, and twist).

© Diagram Visual Information Ltd.

Key words

atmosphere
carbon
carbon cycle
chlorophyll
glucose
photosynthesis
respiration

The carbon cycle

- *Carbon* is the fourth most abundant element in the Universe.
- The total amount of carbon on planet Earth is fixed. The same carbon atoms have been used in countless other molecules since Earth began. The *carbon cycle* is the complex set of processes through which all carbon atoms rotate.
- Carbon exists in Earth's *atmosphere* primarily as carbon dioxide.
- All green plants contain *chlorophyll*, a pigment that gives them their characteristic color. During *photosynthesis*, chlorophyll traps energy from sunlight and uses it to convert carbon dioxide and water into *glucose* and oxygen.
- Carbon is transferred from green plants to animals when animals eat plants or other animals.
- All animals and plants need energy to drive their various metabolic processes. This energy is provided by *respiration*. During this process, glucose reacts with oxygen to form carbon dioxide and water. These waste products are subsequently released into the atmosphere. In essence, respiration is the opposite process to photosynthesis.
- When plants and animals die, their bodies decompose. In the presence of air, the carbon they contain becomes carbon dioxide, which is released into the atmosphere.
- When plants and animals decay in the absence of air, carbon cannot be converted into carbon dioxide. Instead, it remains and forms fossil fuels such as coal, crude oil, and natural gas.
- When fossil fuels are burned, the carbon they contain becomes carbon dioxide and is released into the atmosphere.

© Diagram Visual Information Ltd.

The carbon cycle

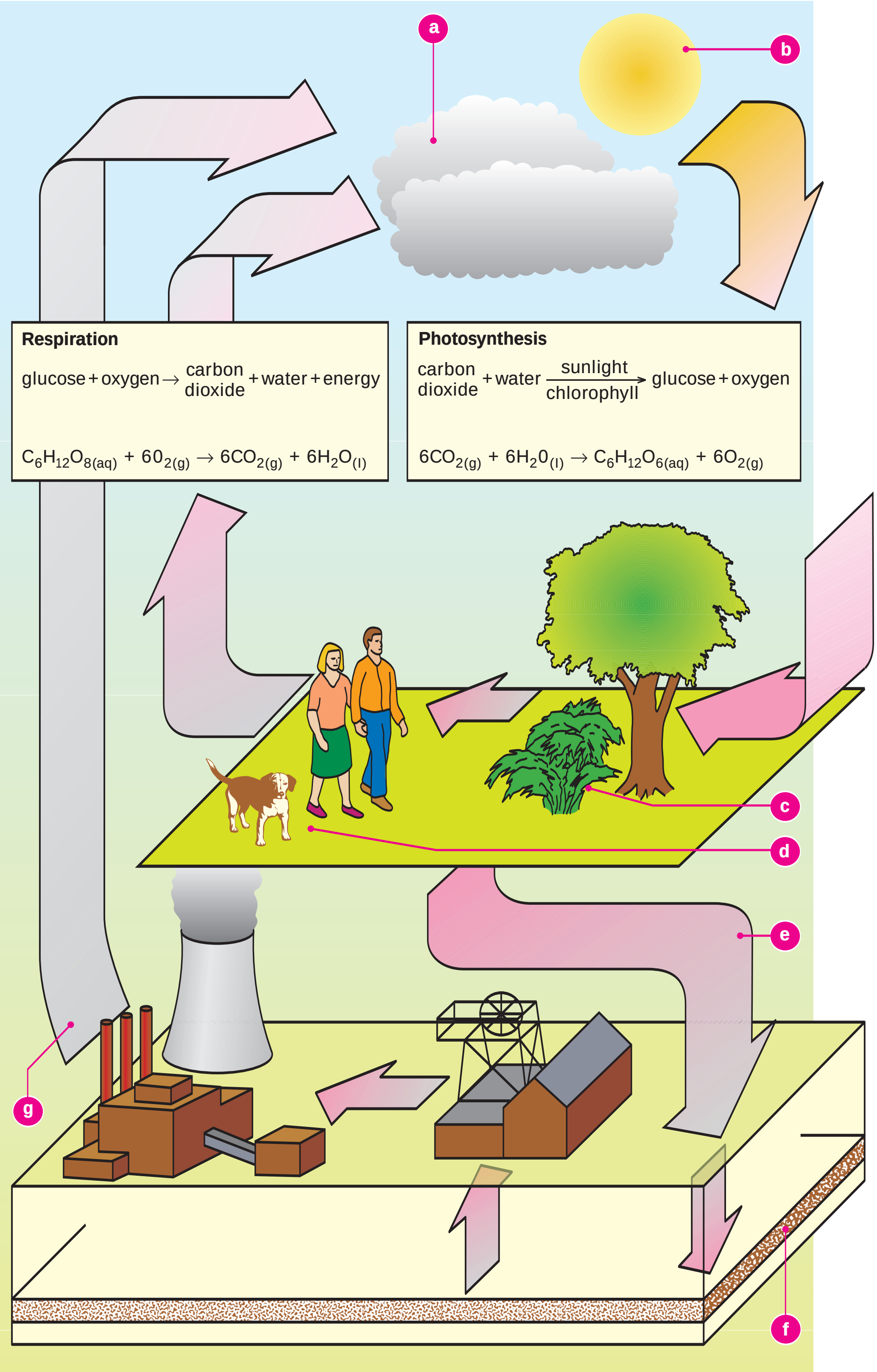

Respiration

glucose + oxygen → carbon dioxide + water + energy

$C_6H_{12}O_{8(aq)} + 60_{2(g)} \rightarrow 6CO_{2(g)} + 6H_2O_{(l)}$

Photosynthesis

$\text{carbon dioxide} + \text{water} \xrightarrow[\text{chlorophyll}]{\text{sunlight}} \text{glucose} + \text{oxygen}$

$6CO_{2(g)} + 6H_20_{(l)} \rightarrow C_6H_{12}O_{6(aq)} + 6O_{2(g)}$

a carbon dioxide in the air
b sunlight
c plants take in carbon dioxide and give out oxygen
d animals take in oxygen, eat plants and vegetables, and breath out CO_2
e death and decay
f carbon compounds (e.g., in oil and coal)
g burning fuel produces CO_2

Laboratory preparation of carbon oxides

1 Preparation of carbon dioxide

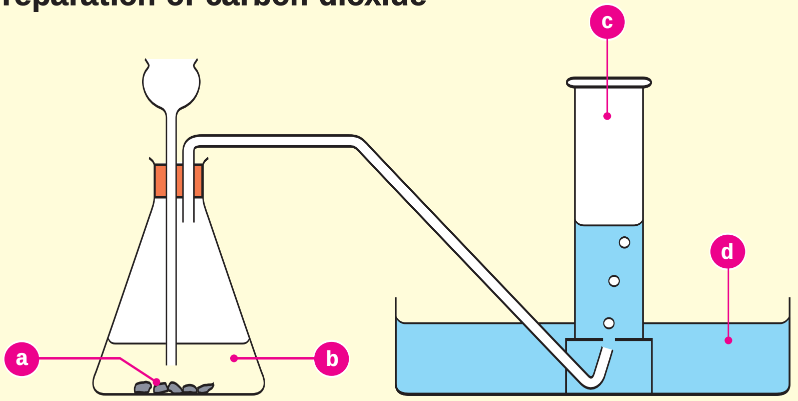

2 Preparation of dry carbon dioxide

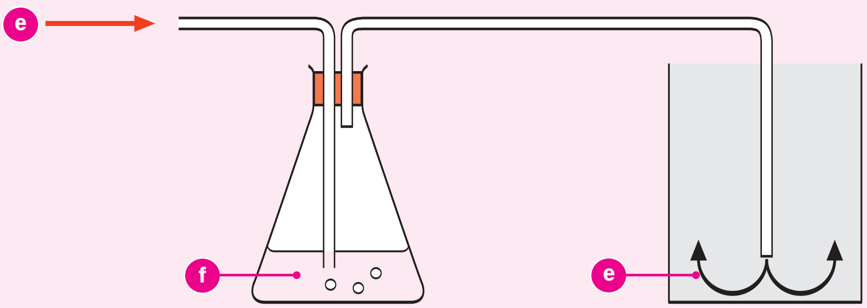

3 Preparation of carbon monoxide

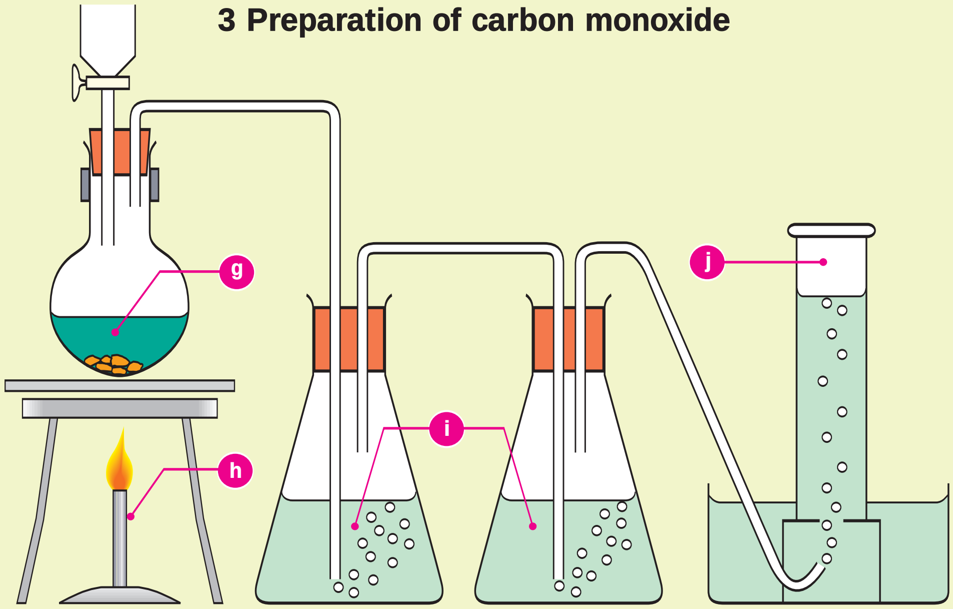

a marble chips
b dilute hydrochloric acid
c carbon dioxide
d water
e carbon dioxide
f concentrated sulfuric acid
g ethanedioic (oxalic) acid crystals and concentrated sulfuric acid
h heat
i concentrated potassium hydroxide solution
j carbon monoxide

Key words

carbonate
carbon dioxide
carbon monoxide

Carbon oxides

- The most common forms of carbon oxides are *carbon dioxide*, which is instrumental in the carbon cycle, and *carbon monoxide*, a colorless, odorless gas that is the result of the incomplete combustion of fuels. They can be prepared in the laboratory using the following techniques.

1 Carbon dioxide

- Carbon dioxide is formed when a metal *carbonate* reacts with a dilute acid:

metal carbonate + dilute acid →
metal salt + carbon dioxide + water

- When calcium carbonate (marble chips) reacts with dilute hydrochloric acid:

$$CaCO_3(s) + 2HCl(aq) \rightarrow CaCl_2(aq) + CO_2(g) + H_2O(l)$$

- Carbon dioxide is not very soluble in water, so it can be conveniently collected over water.

2 Dry carbon dioxide

- Carbon dioxide can be dried by passing it through concentrated sulfuric acid and collected by downward delivery (upward displacement) because it is denser than air.

3 Carbon monoxide

- Carbon monoxide is formed by the dehydration of ethanedioic (oxalic) acid using concentrated sulfuric acid:

$$HOOC\text{-}COOH(l) \xrightarrow{\text{conc. sulfuric acid}} CO_2(g) + CO(g) + H_2O(l)$$

- Acid residues and carbon dioxide are removed by passing the gas through a potassium hydroxide solution. Carbon monoxide can be collected over water because it is only slightly soluble.

© Diagram Visual Information Ltd.

The fractional distillation of crude oil

Key words

fractional distillation
hydrocarbon

Fractional distillation

- *Fractional distillation* is one of several processes used to refine crude oil. Refining converts crude oil into a range of useful products.
- Crude oil is a complex mixture of *hydrocarbons*. During fractional distillation, this mixture is separated into a series of fractions (components) on the basis of boiling point.
- The crude oil is passed through a furnace, where it is heated to 400°C and turns mostly into vapor. The gases pass into a distillation column within which there is a gradation of temperature. The column is hottest at the bottom and coolest at the top.
- Hydrocarbons with the highest boiling points are the first to condense at the bottom of the column, along with any remaining liquid residue from the crude oil. This fraction provides bitumen for use in road building.
- Rising up the column, other fractions condense out: first diesel oil, then kerosene, and finally gasoline. All of these fractions are used as fuels.
- The hydrocarbons with the lowest boiling points remain as gases and rise to the top of the column. This fraction is used as a fuel in the refinery.
- The hydrocarbon vapor moves up the column through a series of bubble caps. At each level, the hydrocarbon vapor passes through condensed hydrocarbon liquid. This helps to ensure a good separation into the various fractions.

Crude oil composition

- Crude oil varies in composition, depending on where it was obtained. Fractional distillation of different crude oils provides different proportions of the various fractions.

© Diagram Visual Information Ltd.

1 Fractional distillation of crude oil

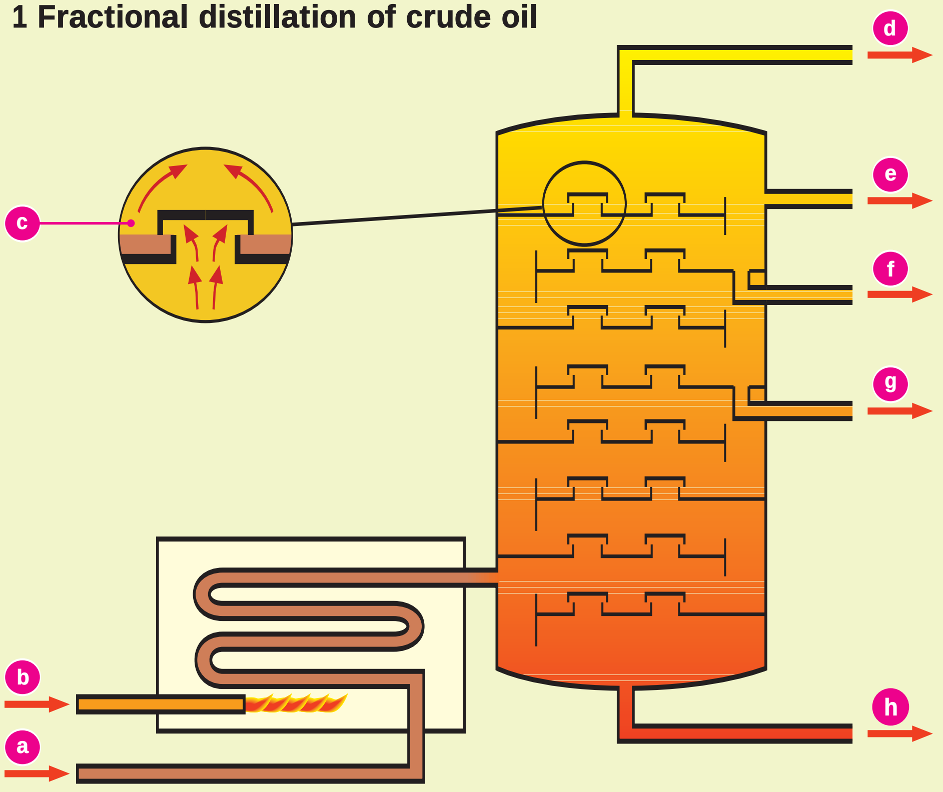

2 Crude oil composition (percent yields and uses of oil)

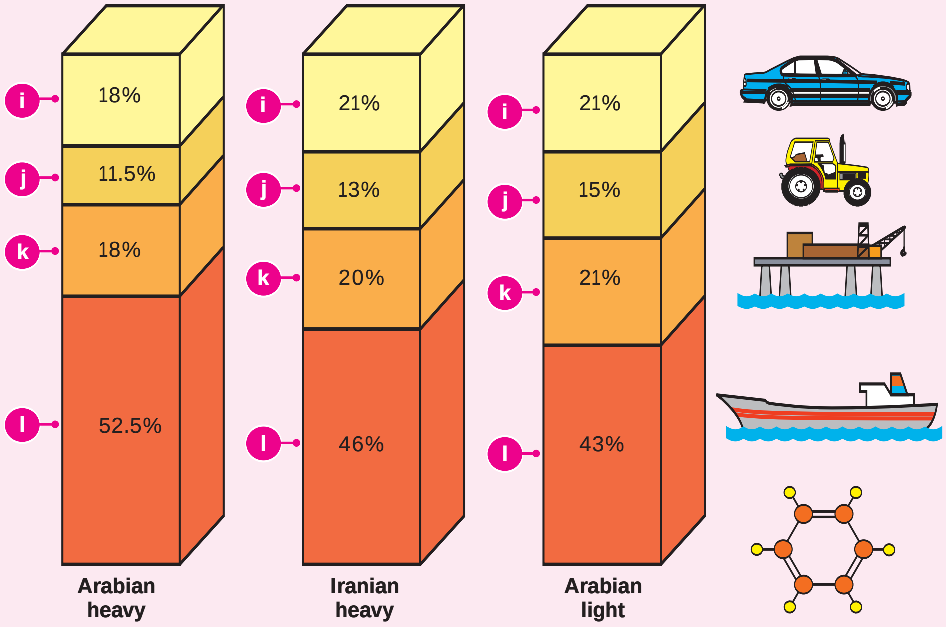

a crude oil
b heater
c bubble cap
d refinery gas
e gasoline (110°C)
f kerosine (180°C)
g diesel oil (260°C)
h residue — bitumen tar (400°C)
i gasoline and chemical feedstock
j kerosine
k gas oil
l fuel oil

Other refining processes

1 Other refining processes

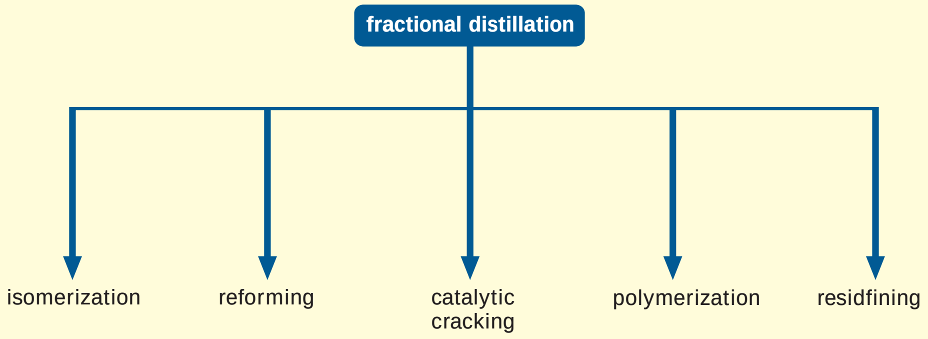

2 Isomerization

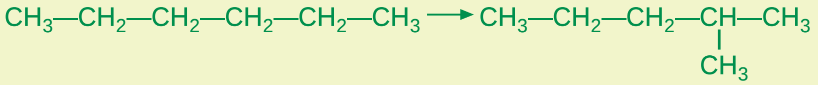

$$CH_3-CH_2-CH_2-CH_2-CH_2-CH_3 \longrightarrow CH_3-CH_2-CH_2-\underset{\displaystyle CH_3}{\underset{|}{CH}}-CH_3$$

pentane 2-methylbutane

3 Reforming

Dehydration

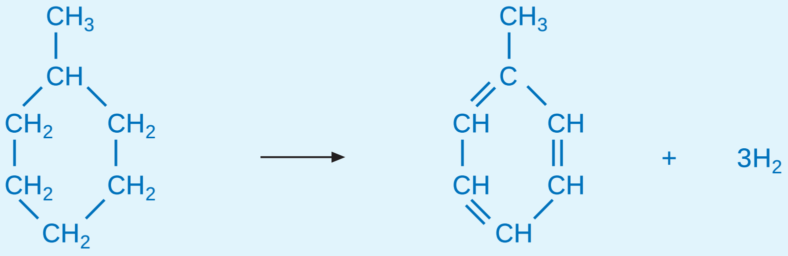

methylcyclohexane methylbenzene

Cyclization

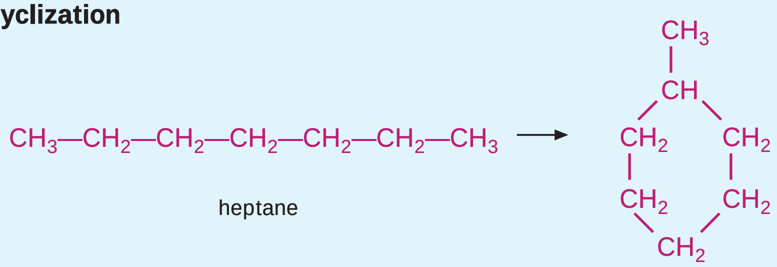

heptane

methylcyclohexane

4 Catalytic cracking

Octane → propane and pent-1-ene

$CH_3CH_2CH_2CH_2CH_2CH_2CH_2CH_3 \rightarrow CH_3CH_2CH_3 + CH_3CH_2CH_2CH{=}CH_2$

octane propane pent-1-ene

5 Polymerization

Two propene molecules combine to form hexene

$CH_3CH{=}CH_2 + CH_3CH{=}CH_2 \rightarrow CH_3CH_2CH_2CH_2CH{=}CH_2$

propene propene hex-1-ene

Key words

alkane
alkene
catalytic cracking
fractional distillation
isomerization
polymerization
reforming
residfining

1 Other processes

- Other refining processes are used to modify the products of *fractional distillation*. These include *isomerization*, *reforming*, *catalytic cracking*, *polymerization*, and *residfining*.

2 Isomerization

- Isomerization changes the shape of hydrocarbon molecules. For example, pentane is converted into 2-methlybutane.

3 Reforming

- Reforming converts straight chain molecules into branched molecules in order to improve the efficiency of gasoline. One type of reaction involves the dehydration of saturated compounds to unsaturated compounds. Another involves the cyclization of hydrocarbons.

4 Catalytic cracking

- In general, smaller hydrocarbon molecules, such as those in gasoline, are in greater demand than larger ones. Catalytic cracking redresses this balance by breaking (cracking) large alkane molecules into smaller *alkane* and *alkene* molecules.

5 Polymerization

- Polymerization combines small molecules to form larger molecules that can be used to make various products.

Residfining

- Resifining is the process used on the residue fraction to convert it into usable products. It also removes impurities that would damage the catalyst used in catalytic cracking.

© Diagram Visual Information Ltd.

Carbon chains

Key words

alkane
alkene
alkyne
bond
carbon
catenation
van der Waals forces

1 Catenation

- *Carbon* has the ability to form long chains of carbon atoms in its compounds. This is called *catenation*.

2 Melting and boiling points

- Forces of attraction, called *van der Waals forces*, exist between molecules. As molecular size increases, there is more overlap between the molecules, and the intermolecular forces of attraction increase.
- In order to melt and to boil, the forces of attraction between molecules must be overcome. The greater these forces, the more energy is needed. This is reflected in a steady increase in melting point and boiling point as molecules increase in size.

3 Types of bonds

- A carbon atom may form one, two, or three *bonds* with another carbon atom in its compounds. These bonds are described as single bonds (C–C), double bonds (C=C), and triple bonds (C≡C).
- *Alkanes* contain only carbon–carbon single bonds.
- *Alkenes* contain a carbon–carbon double bond.
- *Alkynes* contain a carbon–carbon triple bond.
- Alkanes, alkenes, and alkynes are all hydrocarbons since they consist only of hydrogen and carbon atoms.

1 Catenation

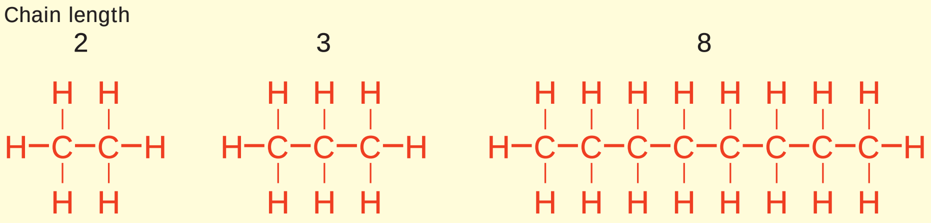

2 Melting and boiling points of some chains

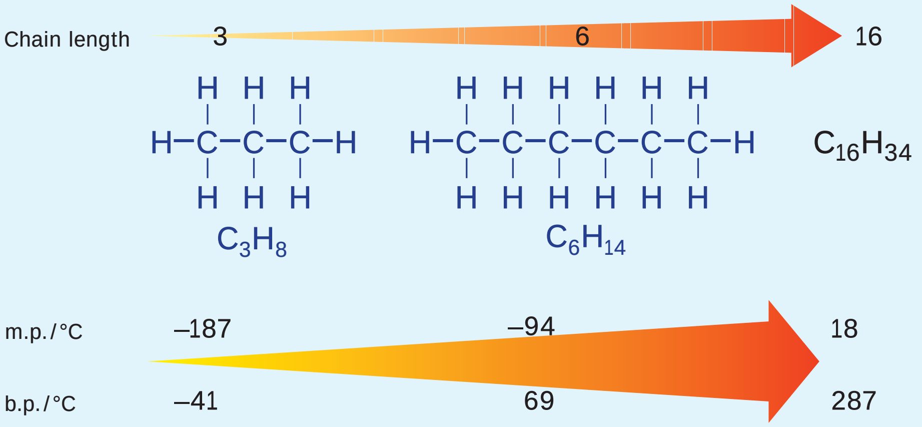

3 Types of bonds

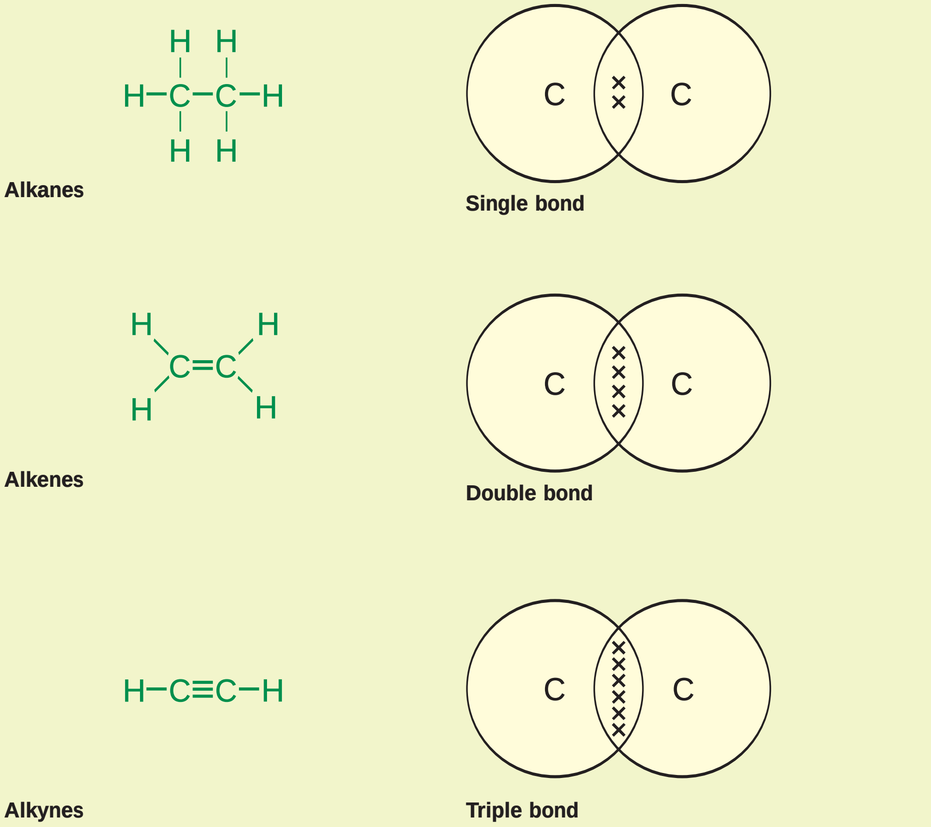

© Diagram Visual Information Ltd.

Naming hydrocarbons

1 Chain length gives first part of name

Chain length	C_1	C_2	C_3	C_4	C_5	C_6
First part of name	meth-	eth-	prop-	but-	pent-	hex-

2 Functional group gives second part of name

Functional group	Alkane 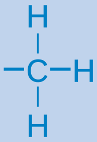	Alkene $C=C$	Alkyne $C\equiv C$
Second part of name	-ane	-ene	-yne

3 Examples of organic compound names

Molecule	**Chain length**	**Functional group**	**Name**
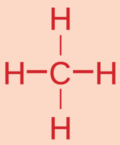	1 → meth-	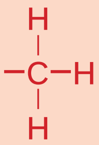 → -ane	methane
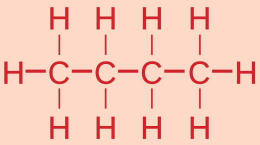	4 → but-	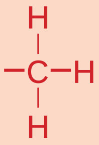 → -ane	butane
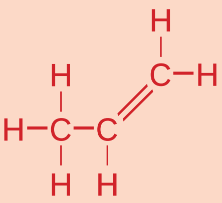	3 → prop-	$C=C$ → -ene	propene
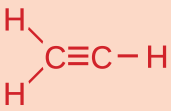	2 → eth-	$C\equiv C$ → -yne	ethyne

Key words

alkane
alkene
alkyne
functional group
hydrocarbon

Naming hydrocarbons

- The name of a *hydrocarbon* indicates the number of carbon atoms in the molecule and what sort of carbon–carbon bonds is present.

1 Chain length

- The first part of name is determined by the number of carbon atoms in the molecule. The same prefixes are used for all groups of organic compounds.

2 Functional group

- The second part of the name is determined by the type of carbon–carbon bonds present. Each *functional group* has a unique suffix.
- The position of the functional group in a carbon chain is identified by numbering the carbon atoms in the carbon chain.

3 Examples of compound names

- The first two examples in the diagram are *alkanes*. If there is one carbon atom in the molecule it is:
  “meth” (1 carbon atom in the chain) + “ane” (for alkane): methane.
  If there are four carbon atoms in the molecule it is:
  “but” (4 carbon atoms in the chain) + “ane” (for alkane): butane.
- The third example is propane, an *alkene* with three carbon atoms:
  “pro” (3 carbon atoms in the chain) + “ene” (for alkene).
- The fourth example is ethyne, an *alkyne* with a two carbon chain:
  “eth” (2 carbon atoms in the chain) + “yne” (for alkyne).

© Diagram Visual Information Ltd.

Key words

alkane
homologous series
hydrocarbon
van der Waals forces

The first six alkanes

- The *alkanes* form an *homologous series* of compounds that have the general formula C_nH_{2n+2}, where n is a positive integer. Each alkane molecule differs from the previous one in the series by -CH_2-.
- They have similar chemical properties and show a gradation of physical properties, such as melting point and boiling point, as the molecular size increases.
- Alkane molecules are attracted to each other by *van der Waals forces*. As molecular size increases, there is more overlap between the molecules, and the intermolecular forces of attraction increase.
- Alkane molecules are frequently shown as having a flat two-dimensional structure because this is easy to draw, but in reality, the four bonds around each carbon atom are directed toward the corners of a tetrahedron. The angle between any two bonds is 109.5°.
- Alkanes are relatively unreactive substances when compared with other groups of *hydrocarbons*. Their most important reaction is combustion, and they are the main constituent of a range of fuels. Natural gas is largely composed of methane:
  $CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$
- In a good supply of air, hydrocarbons burn to give carbon dioxide and water. In a restricted supply of air, carbon monoxide and/or carbon may be formed:
  $C_2H_6 + 2O_2 \rightarrow CO + C + 3H_2O$

© Diagram Visual Information Ltd.

Table of the first six alkanes

Alkane	Methane	Ethane	Propane
Formula	CH_4	C_2H_6	C_3H_8
Structural formula	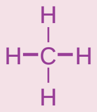 $H-\overset{H}{\underset{H}{C}}-H$	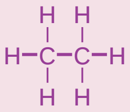 $H-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-H$	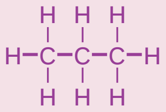 $H-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-H$
Boiling point (°C)	–164	–87	–42
Physical state at room temperature	Gas	Gas	Gas
Molecular model	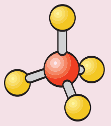	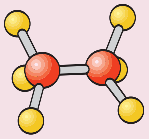	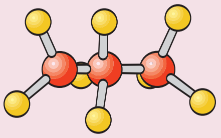

Alkane	Butane	Pentane	Hexane
Formula	C_4H_{10}	C_5H_{12}	C_6H_{14}
Structural formula	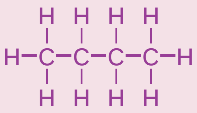 $H-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-H$	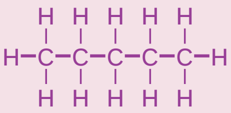 $H-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-H$	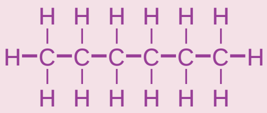 $H-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-\overset{H}{\underset{H}{C}}-H$
Boiling point (°C)	0	36	69
Physical state at room temperature	Gas	Liquid	Liquid
Molecular model	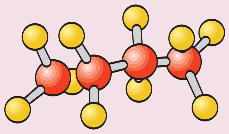	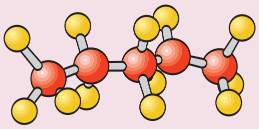	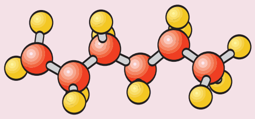

Table of the first five alkenes

Alkene	Formula	Number of carbon atoms per molecule	Structural formula	Boiling point (°C)	Physical state at room temperature	Molecular model
Ethene	C_2H_4	2	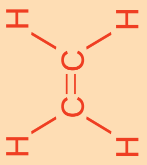	–104	Gas	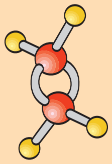
Propene	C_3H_6	3	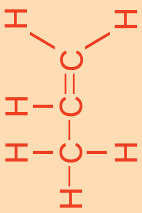	–47	Gas	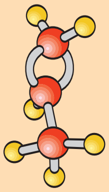
Butene	C_4H_8	4	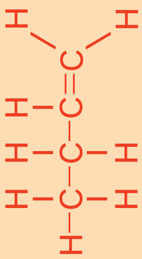	–6	Gas	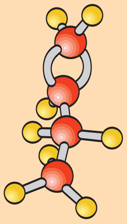
Pentene	C_5H_{10}	5	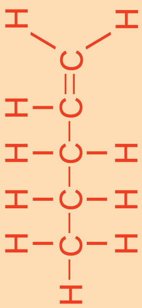	30	Liquid	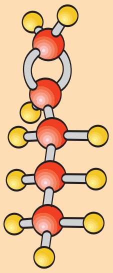
Hexene	C_6H_{12}	6	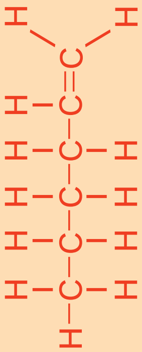	64	Liquid	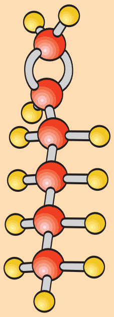

Key words

addition reaction
alkene
functional group
homologous series
van der Waals forces

The first five alkenes

- The *alkenes* form an *homologous series* of compounds with the general formula C_nH_{2n}, where n is a positive integer. Each alkene molecule differs from the previous one in the series by $-CH_2-$.
- Alkene molecules are attracted to each other by *van der Waals forces*. As molecular size increases, there is more overlap between the molecules, and the intermolecular forces of attraction increase. The series thus shows a gradation of physical properties, such as melting point and boiling point.
- Alkenes all contain the same *functional group*, a carbon–carbon double bond, represented by C=C.
- The bonds around each of the carbon atoms in a carbon–carbon double bond are in the same plane and directed toward the corners of an equilateral triangle. The angle between any two bonds is 120°.
- Alkenes undergo combustion in the same way as alkanes. However, they have other chemistry resulting from the reactive carbon–carbon double bond.
- Alkenes undergo addition reactions in which a molecule is added across the carbon–carbon double bond. For example, ethene undergoes the following *addition reactions*:

$CH_2{=}CH_2 + H\text{-}H \rightarrow CH_3\text{-}CH_3$
ethene + hydrogen → ethane

$CH_2{=}CH_2 + H\text{-}OH \rightarrow CH_3\text{-}CH_2\text{-}OH$
ethene + steam → ethanol

$CH_2{=}CH_2 + H\text{-}Br \rightarrow CH_3\text{-}CH_2Br$
ethene + hydrogen bromide → bromoethane

$CH_2{=}CH_2 + Br\text{-}Br \rightarrow CH_2Br\text{-}CH_2Br$
ethene + bromine → 1,2-dibromoethane

© Diagram Visual Information Ltd.

Ethene

Key words

alkene
ethane
ethene
ethanol
geometric isomerism
halogens
isomer

Ethene

- *Ethene* is the first member of the *alkene* series. It is a colorless, flammable gas.

1 Preparation

- In the laboratory, ethene can be made by the dehydration of *ethanol* using concentrated sulfuric acid.

2 Structure

- Ethene, like all alkenes, contains a carbon–carbon double bond about which rotation is impossible.

3 Isomerism

- *Isomers* are compounds having the same molecular formula and relative molecular mass but different three-dimensional structures.
- The existence of two compounds with the same molecular formula but where groups are distributed differently around a carbon–carbon double bond is described as *geometric isomerism* or cis / trans isomerism.
- The prefix "cis" is used when the substituent groups (an atom or group of atoms substituted in place of a hydrogen atom or chain) of a hydrocarbon are or the same side of a plane through the carbon–carbon double bond. The prefix "trans" is used when the substituent groups are on the opposite side.
- In trans-1,2-dibromoethene the bromine atoms are on opposite sides of a plane through the carbon–carbon double bond.
- In cis-1,2-dibromoethene the bromine atoms are on the same side.

4 Reactivity

- The carbon–carbon double bond in ethene is very reactive and will undergo various addition reactions. Ethene reacts with: *halogens* (such as chlorine) to form 1,2-dihaloethane, hydrogen to form *ethane*, and hydrogen halides to form haloethane.

© Diagram Visual Information Ltd.

1 Dehydration of ethanol to produce ethene

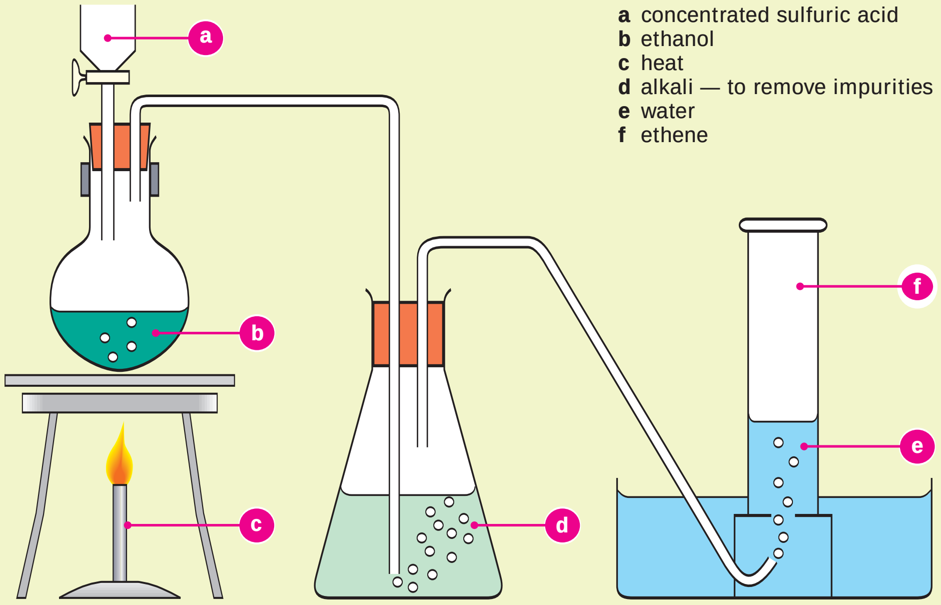

a concentrated sulfuric acid
b ethanol
c heat
d alkali — to remove impurities
e water
f ethene

$$\underset{\text{ethanol}}{CH_3CH_2OH} \xrightarrow{-H_2O} \underset{\text{ethene}}{CH_2{=}CH_2}$$

2 Structure

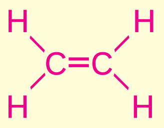

Ethene

3 Isomerism

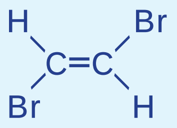

Trans-1, 2-dibromoethene

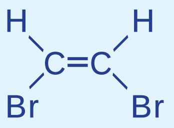

Cis-1, 2-dibromoethene

4 Reactivity

$$CH_2 = CH_2 + Cl_2 \longrightarrow CH_2Cl{-}CH_2Cl$$

Reaction with chlorine to form 1, 2-dichloroethane

$$CH_2 = CH_2 + H_2 \xrightarrow{Ni} CH_3CH_3$$

Reaction with hydrogen to form ethane

$$CH_2 = CH_2 + HX \longrightarrow CH_3CH_2X$$

Reaction with hydrogen halides to form haloethane

Polymers

1 Types of branching

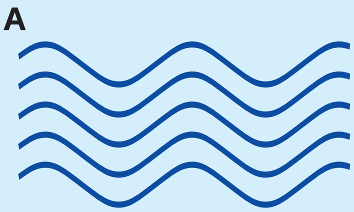

Polymer with few branched chains, e.g., high-density polyethene

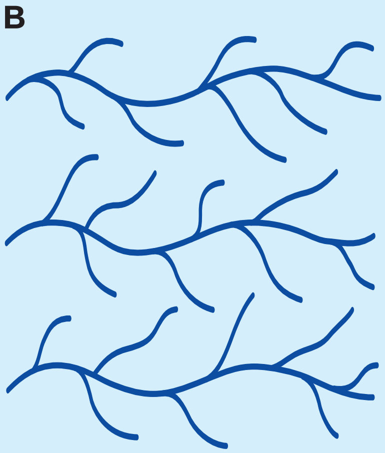

Polymer with many branched chains, e.g., low-density polyethene

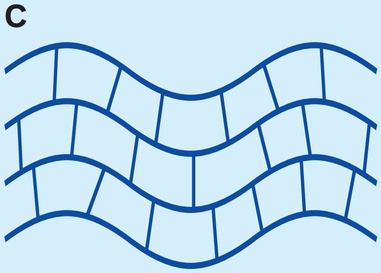

Polymer with much cross-linking, e.g., bakelite

2 Additional polymerization (illustrating how ethene can be restructured to form poly(ethylene) i.e., polyethene)

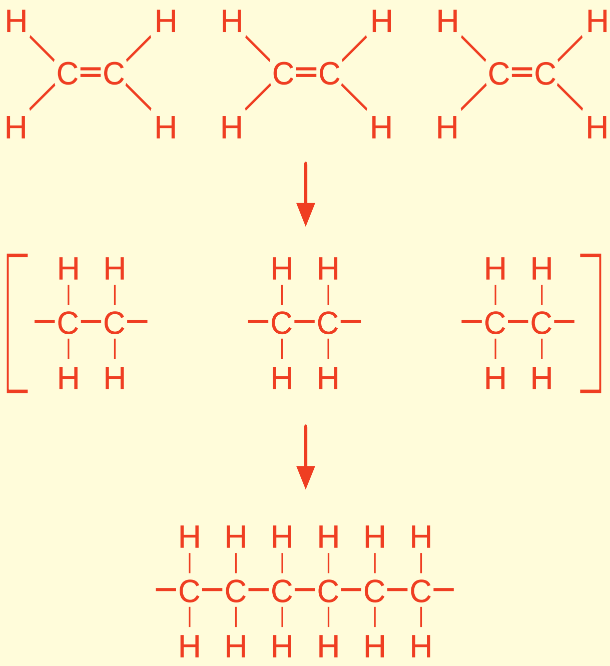

Key words

addition polymerization
bakelite
ethene
polyethene
polymer
polymerization

Polymers

- A *polymer* is a large organic molecule composed of repeating carbon chains. The physical properties of a polymer depend on the nature of these carbon chains and how they are arranged.

1 Types of branching

- A certain amount of side branching occurs during *polymerization*, depending on the reaction conditions.
- Low pressure and low temperature results in a high-density polymer.
- Very high pressure and moderate temperatures produce a low-density polymer.
- In high-density polymers, the carbon chains are unbranched, and they can be packed closely together forming a dense substance, e.g., high-density *polyethene* (1A).
- In low-density polymers, the carbon chains are branched, and it is not possible to pack them as closely together, e.g., low-density polyethene (1B).
- In polymers like *bakelite*, there are cross links between the carbon chains, producing a hard, rigid structure (1C).

2 Addition polymerization

- *Ethene* forms a polymer by a process called *addition polymerization*.
- In this process, one of the bonds from the carbon–carbon double bond is used to form a bond with an adjacent molecule. This process is repeated many times, resulting in long chains containing thousands of carbon atoms.

© Diagram Visual Information Ltd.

Polymers: formation

Key words

alkene
monomer
polymer

Monomers

- *Monomers* are the basic units from which a polymer is made.
- The systematic name for a *polymer* is derived from the name of the monomer. For example, polypropene is “poly” (for polymer) + the *alkene* propene.
- The diagrams at right illustrate the formation of some alkene polymers.

1 Forming polypropene

- Propene molecules combine to form polypropene.
- Most polypropene is produced as a monopolymer (a polymer formed from propene only).

2 Forming polychloroethene

- Chloroethene molecules combine to form polychloroethene.
- 1,2-dichloroethane is made by chlorinating ethene. This product is then cracked to form chloroethene.

3 Forming polyphenylethene

- Phenylethene molecules combine to form polyphenylethene.
- Phenylethane is made from ethene and benzene by a Friedel-Crafts reaction using aluminum(III) chloride/hydrochloric acid catalyst. This is dehydrogenated to give the phenylethene monomer.

4 Forming polytetrafluoroethene

- Tetrafluoroethene molecules combine to form polytetrafluoroethene.
- Trichloromethane is produced by the reaction of methane with controlled amounts of chlorine/hydrochloric acid. This is reacted with anhydrous hydrogen fluoride in the presence of antimony(III) chloride to give chlorodifluoromethane, which is subsequently cracked to produce tetrafluoroethene.

1 Restructuring of propene to make poly(propene)

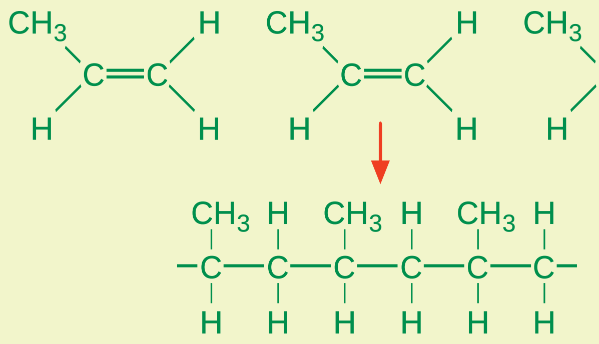

2 Restructuring of chloroethene to make poly(chloroethene) i.e., polyvinylchloride

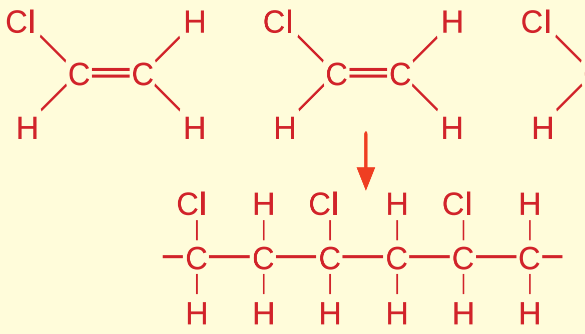

3 Poly(phenylethene)

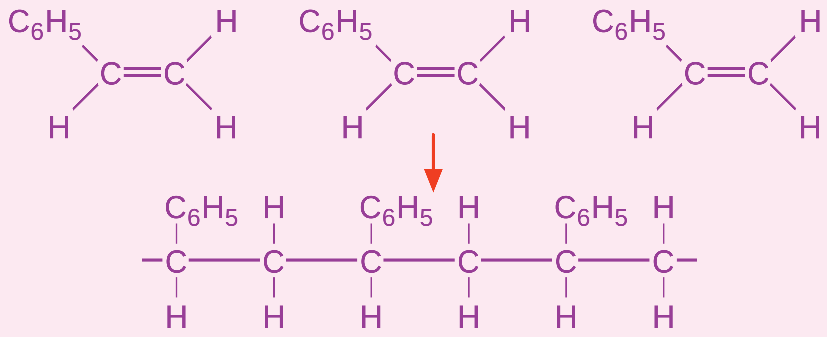

4 Poly(tetrafluoroethene)

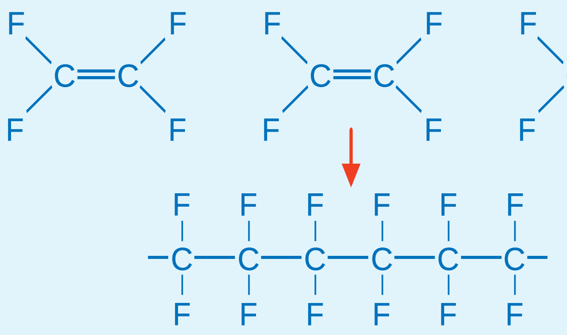

© Diagram Visual Information Ltd.

Polymers: table of properties and structure

Polymer systematic name	Polymer common name	Properties	Uses	Structure of monomer
Poly(ethene)	Polyethene	low density; high density	film and bags; molding rigid articles	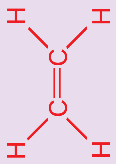
Poly(propane)	Polypropylene	high density	molding rigid articles, film and fibers	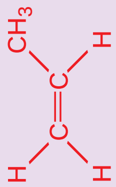
Poly(chloroethene)	PVC (polyvinylchloride)	flexible	coating fabrics and insulation on wires and cables	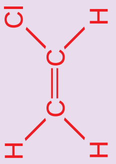
Poly(phenylethene)	Polystyrene	brittle but cheap	plastic toys, expanded with air and used for insulation	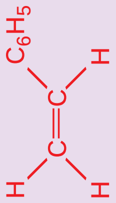
Poly(ethene)	PTFE (polytetrafluoroethene)	low friction and stable to heat	non-stick coating on pans	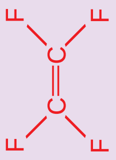
Poly(methyl-2-methyl-propenoate)	Perspex	transparent	substitute for glass	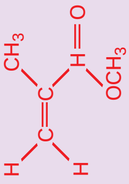
Poly(propenenitrile)	Acrilan	strong fibers	wool substitute in textiles	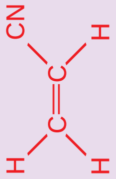

Key words

monomer
polymer
polymerization

Polymers

- Most *polymers* have common names that are used in everyday language.
- The uses of polymers depend on their properties.

Classification

- There are several ways in which polymers can be classified.
- Heat. Thermoplastics soften when heated and harden on cooling, so they can be reshaped many times without changing their chemical structure. Thermosets are chemically altered on heating and produce a permanently hard material that cannot be softened by heating.
- Method of polymerization. Addition polymers are usually formed from *monomers* containing a –CH=CH- unit to which different atoms or groups are attached. On *polymerization*, one of the carbon–carbon bonds becomes a bond to another unit. Condensation polymers are formed from condensation reactions in which a small molecule, sometimes but not always water, is lost.
- Formula. Homopolymers are formed from one monomer unit. Co-polymers are formed from two or more monomers.
- Chemical structure. Linear chains may have straight, zigzag, coiled, or random spatial arrangements. Branched chains have side branch chains attached to the main chains. Cross-linked chains have two or three dimensional cross-linkage between chains.
- Steric structure. Isotactic: in which all side groups are on the same side of the main chain. Syndiotactic: in which each alternative side group has the same orientation. Atactic: in which there is no specific pattern to the distribution of side groups.

© Diagram Visual Information Ltd.

Functional groups and homologous series

Key words

alcohol
alkene
carboxylic acid
ester
functional group
homologous series
oxidation

Functional groups

- A *functional group* is the atom or group of atoms present in a molecule that determines the characteristic properties of the molecule.
- A *homologous series* is a group of compounds that contain the same functional group. The physical properties of a homologous series show a gradation as molecular size increases. The chemical properties of a homologous series are similar because they are determined by the functional group.

1 Alkenes

- *Alkenes* contain the functional group $C=C$.
- Their general formula is C_nH_{2n}.
- Alkenes are reactive and undergo additional reactions.

2 Alcohols

- *Alcohols* contain the functional group C-OH.
- Their general formula is $C_nH_{2n+1}OH$.
- Alcohols can also undergo *oxidation* to give carboxylic acids, or they can be dehydrated to alkenes. They can also react to form ester compounds

3 Carboxylic acids

- *Carboxylic acids* contain the functional group -COOH.
- Carboxylic acids are typically weak acids that partially dissociate into H^+ cations and $RCOO^-$ anions in aqueous solution.
- Carboxylic acids are widespread in nature.

4 Esters

- *Esters* contain the functional group –COOC-.
- Esters are formed by a reaction between a carboxylic acid and an alcohol.
- Esters are used in flavorings and perfumes.

© Diagram Visual Information Ltd.

Functional group

Example

1 Alkenes

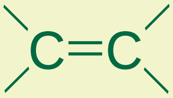

Propene

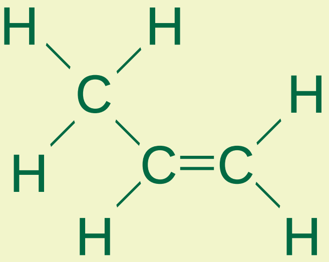

2 Alcohols

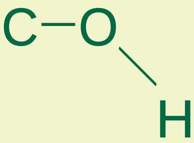

Ethanol

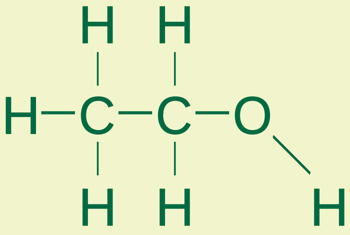

3 Carboxylic acids

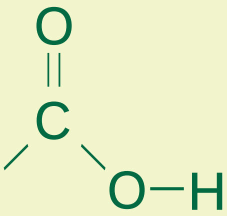

Ethanoic acid

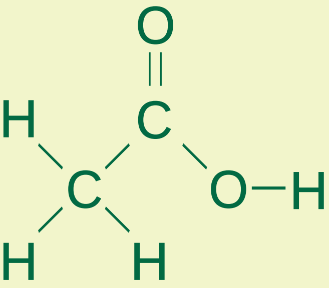

4 Esters

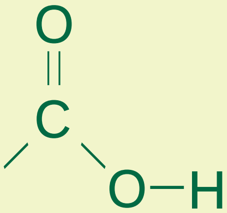

Methylethanoate

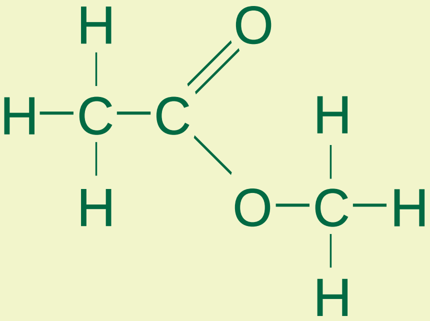

Alcohols

1 The first six alcohols

Structure	Name
CH_3-OH	methanol
CH_3CH_2-OH	ethanol
$CH_3CH_2CH_2$-OH	propan-1-ol
$CH_3CH_2CH_2CH_2$-OH	butan-1-ol
$CH_3CH_2CH_2CH_2CH_2$-OH	pentan-1-ol
$CH_3CH_2CH_2CH_2CH_2CH_2$-OH	hexan-1-ol

2 Classification

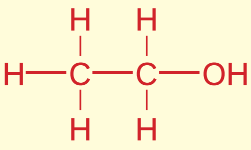
Primary alcohol

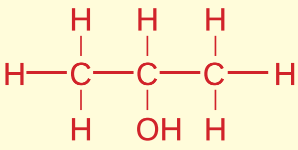
Secondary alcohol

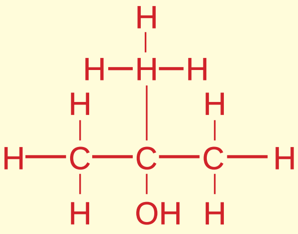
Tertiary alcohol

3 Sharing of electrons

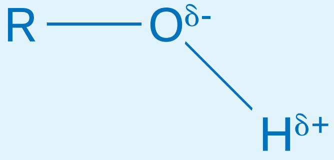

4 Hydrogen bonding

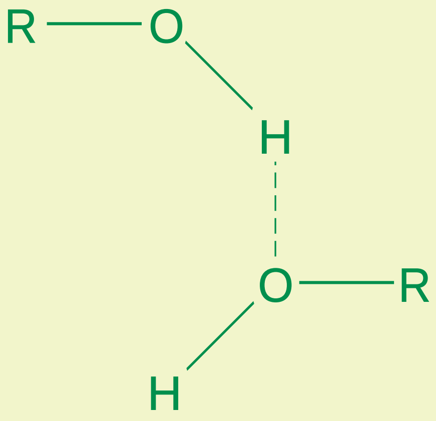

Key words

alcohol
alkane
functional group
hydrogen bond

1 Naming

- *Alcohols* are named by dropping the terminal “e” from the *alkane* chain and adding “ol.” For example, methane is the alkane; methanol is the alkanol, or alcohol. When necessary, the position of the hydroxyl (-OH) group is indicated by a number between the alkane name and the “ol,” e.g., propan-1-ol, or in front of the name, e.g., 2-propanol.

2 Classification

- Alcohols may be classified as primary, secondary, or tertiary on the basis of the number of carbon atoms bonded to the carbon carrying the *functional group* (-OH).

3 Sharing of electrons

- An oxygen atom is more electronegative than a hydrogen atom, and this leads to an unequal sharing of the electrons in the O-H bond. The bonding electrons are drawn more toward the oxygen atom and, because the electrons carry a negative charge, the oxygen atom becomes slightly negative. This is described as delta minus and is denoted by δ-. Conversely, the hydrogen atom becomes slightly positive—delta plus, denoted by δ+. (R represents the carbon group attached to the oxygen.)

4 Hydrogen bonding

- The -OH functional group generally makes the alcohol molecule polar. It has a positive charge at one end and a negative at the other. Molecules can form *hydrogen bonds* with one another and other compounds when the oppositely charged parts are attracted to each other, forming hydrogen bonds.

© Diagram Visual Information Ltd.

Carboxylic acids

Key words

alkali	*pH*
carbonate	*salt*
carboxylic acid	
dissociation	
homologous series	

1 Naming

- *Carboxylic acids* are named by adding the suffix "anoic acid" to the prefixes used for all *homologous series* of organic compounds. For example, the carboxylic acid containing three carbon atoms is "prop" + "anoic acid" = "propanoic acid."

2 Hydrogen bonding

- Hydrogen bonding is present between carboxylic acid molecules, resulting in higher boiling points than might otherwise be expected and miscibility with water.

3 Ionization

- Carboxylic acids ionize to give hydrogen ions, H^+; however, they are weak acids because they are only partially ionized.
- The *dissociation* constant for ethanoic acid, for example, is 1.75 x 10^{-5} mol^3dm^{-6}. This means that only about 4 molecules in every 1,000 are ionized at any one time.

Characteristics

- Carboxylic acids have a *pH* value of approximately 3–5.
- Carboxylic acids react with *carbonates* and hydrogencarbonates to produce carbon dioxide:

  $2H^+(aq) + CO_3^{2-}(aq) \rightarrow$
  $H_2O(l) + CO_2(g)$
  $H^+(aq) + HCO_3^-(aq) \rightarrow$
  $H_2O(l) + CO_2(g)$
- Carboxylic acids form *salts* with *alkalis*:

  $CH_3COOH(aq) + NaOH(aq) \rightarrow$
  ethanoic acid sodium hydroxide
  $CH_3COO^-Na^+(aq) + H_2O(l)$
  sodium ethanoate water

© Diagram Visual Information Ltd.

1 The first six carboxylic acids

Structure	Name
CHOOH	methanoic acid
CH_3COOH	ethanoic acid
CH_3CH_2COOH	propanoic acid
$CH_3CH_2CH_2COOH$	butanoic acid
$CH_3CH_2CH_2CH_2COOH$	pentanoic acid
$CH_3CH_2CH_2CH_2CH_2COOH$	hexanoic acid

2 Hydrogen bonding

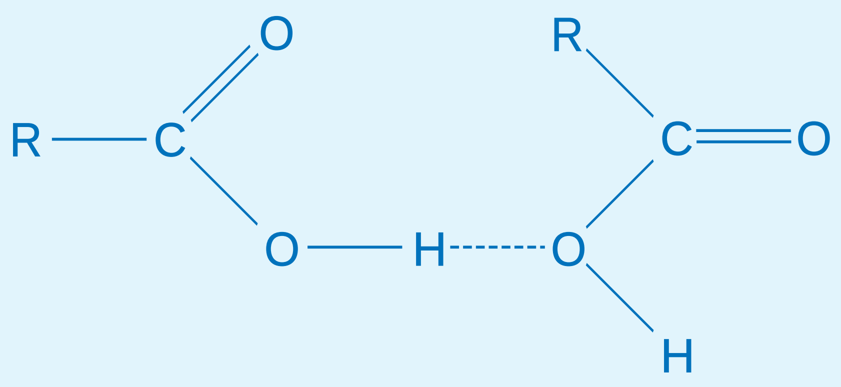

3 Ionization

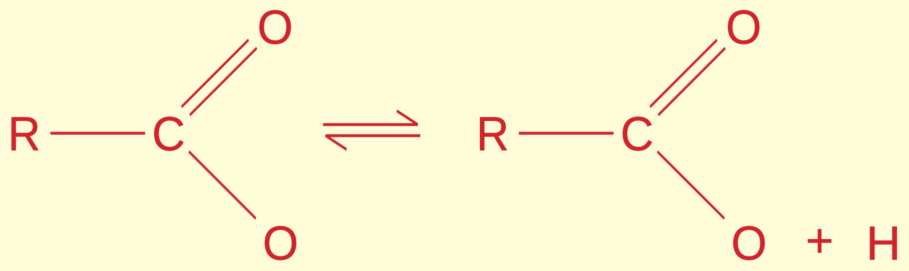

Esters

1 Forming esters

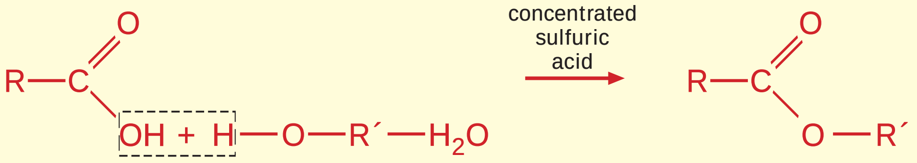

2 Naming

Structure of ester	Name of ester
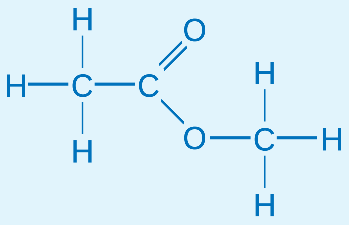	methyl ethanoate
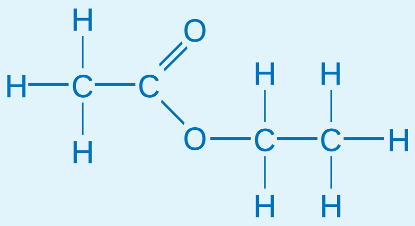	ethyl ethanoate
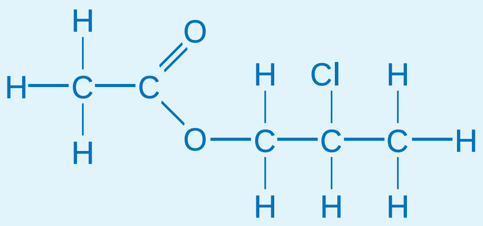	propyl ethanoate
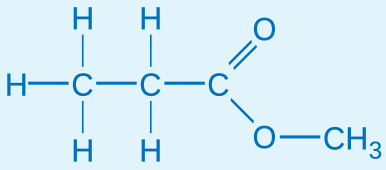	methyl propanoate

3 Saponification

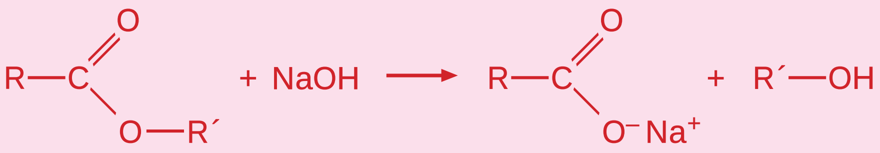

Key words

alcohol
alkyl
aryl
carbon
carboxylic acid
ester
functional group
saponification

Esters

- *Esters* contain the *functional group* –COOR, where R is an *alkyl* or an *aryl* group.

1 Forming esters

- Esters are formed by the reaction of *carboxylic acids* with *alcohols* in the presence of a strong acid catalyst, such as concentrated sulfuric acid. The reaction involves the loss of water.
- Esters generally have a fruity smell that can be used to identify their presence. They are used for food flavorings and in cosmetics.
- Esters have no –OH group, so they cannot form hydrogen bond like carboxylic acids and alcohols. Consequently, they are more volatile and are insoluble in water.

2 Naming

- The name of an ester is derived from the carboxylic acid and the alcohol from which it is formed.
- The alcohol part of an ester is written at the beginning of the ester name; from methanol we get methyl, from ethanol we get ethyl, etc.
- The acid part of an ester is written at the end of the ester name. It is written as if it was an ionic carboxylate group in a salt; from ethanoic acid we get ethanoate, from propanoic acid we get propanoate, etc.

3 Saponification

- When esters are heated with an alkali, such as sodium hydroxide, they are readily hydrolyzed to form an alcohol and a carboxylic acid salt.
- This may be described as a *saponification* reaction. It is important in the production of soaps from fats and oils.

© Diagram Visual Information Ltd.

Soaps and detergents

Key words

carboxylic acid
detergent
ester
fatty acid
hydrophilic
hydrophobic
soap

Soaps and Detergents

- *Soaps* are cleansing agents made from *fatty acids* derived from natural oils and fats. *Detergents* are made from synthetic chemical compounds.

1 Fatty acids

- *Carboxylic acids* occur in animal and plant fats and oils. They may contain from 7 to 21 carbon atoms and are often referred to as fatty acids.

2 Making soap

- Most naturally occurring fats and oils are *esters* of propane-1,2,3-triol (glycerine). When the fats are boiled with sodium hydroxide, propane1,2,3,-triol and a mixture of sodium salts of the three carboxylic acids are formed. These salts are what we call soaps.

3 Soap molecule

- One end of a soap molecule is ionic, while the other end is covalent. The ionic end is described as *hydrophilic* because it dissolves in water. Conversely, the covalent end is described as *hydrophobic* because it does not dissolve in water, but it will dissolve in organic substances like oils.

4 Cleaning action

- The cleaning action of soap is the result of the different affinities of the two ends of the soap molecule.
- The hydrophobic end of the molecule dissolves in oils and fats on the fabric, while the hydrophilic end of the molecule remains in the water.
- The oil and fat particles are lifted off the fabric and held in the water by soap molecules.

5 Detergent molecule

- Alkylbenzene sulfonates are common examples of detergents.

© Diagram Visual Information Ltd.

1 Common fatty acids

Name	Formula	Found in
palmitic acid	$CH_3(CH_2)_{14}COOH$	animal and vegetable fats
stearic acid	$CH_3(CH_2)_{16}COOH$	animal and vegetable fats
oleic acid	$CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH$	most fats and oils
linoleic acid	$CH_3(CH_2)_4CH{=}CHCH_2CH{=}$ $CH(CH_2)_7COOH$	soya-bean oil and nut oil

2 Making soap

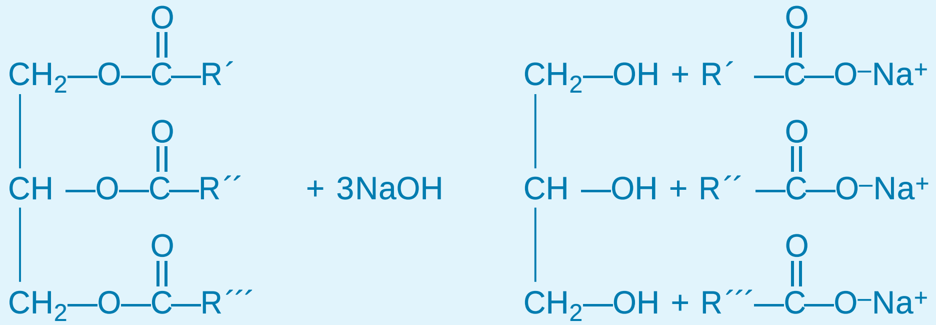

3 Soap molecule

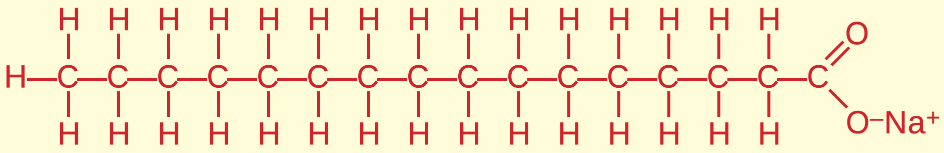

4 Cleaning action

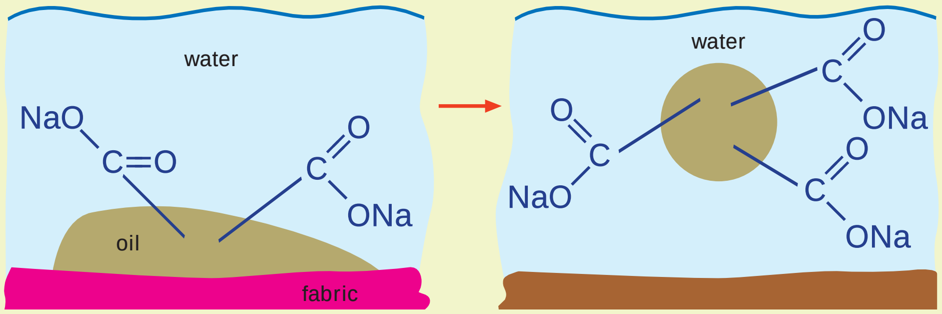

5 Detergent molecule

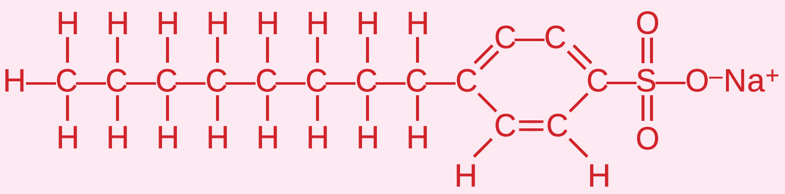

Organic compounds: states

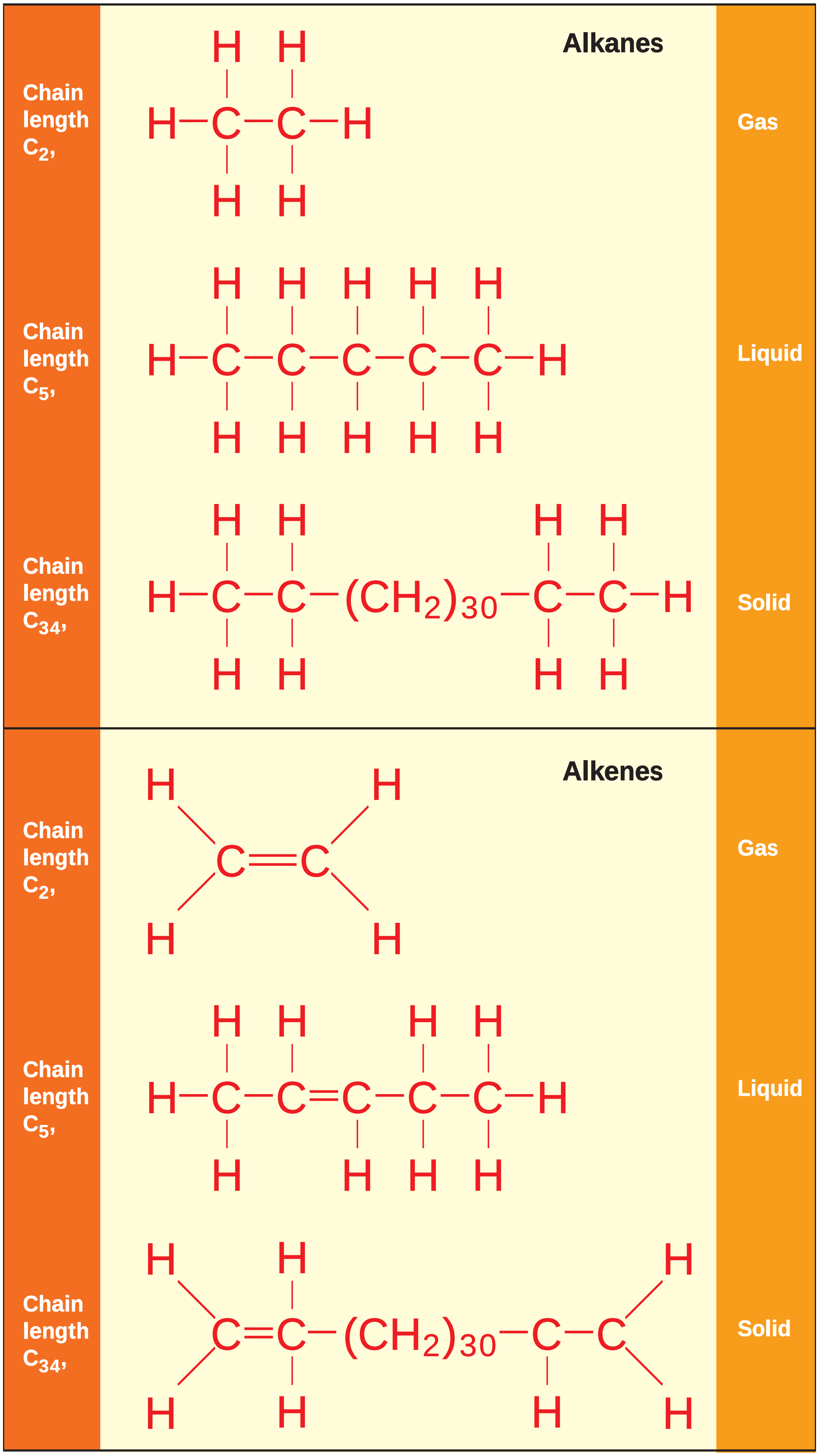
Alkanes
Chain length C2,
Gas
Chain length C5,
Liquid
Chain length C34,
Solid
Alkenes
Chain length C2,
Gas
Chain length C5,
Liquid
Chain length C34,
Solid

Key words

alkane
alkene
homologous series

Physical properties

- All *homologous series* of compounds show a gradation of physical properties as the carbon chain length increases.

Alkanes

- The simplest *alkane* is CH_4, methane. The next simplest alkane is the two carbon alkane, ethane (C_2H_6). Both of these are gases.
- The five carbon alkane, pentane (C_5H_{12}), is a liquid.
- The 34 carbon compound, butadecane is a solid.

Alkenes

- The simplest *alkene* is the two carbon alkene, ethene (C_2H_4), which is a gas.
- 2-pentene (C_5H_{10}), which is a five carbon alkene, is a liquid.
- 2-butedecane, which is a 34 carbon alkene, is a solid.

© Diagram Visual Information Ltd.

Key words

alcohol
aldehyde
alkene
carboxylic acid
ester
functional group
homologous series
ketone
polymer

Functional groups and properties

- All members of an *homologous series* of compounds has the same *functional group*. Because the functional group determines most of the chemistry of a compound, members of a particular homologous series will have similar chemical reactions.
- *Alkenes* are unsaturated compounds because they all contain a carbon–carbon double bond that makes them very reactive. Typically, they will undergo addition reactions with hydrogen, halogens, and water. They also form a variety of *polymers*.
- *Alcohols* with a small relative molecular mass are flammable liquids and readily dissolve in water. Primary alcohols are readily oxidized: first to *aldehydes* and then to *carboxylic acids*. Secondary alcohols are oxidized to *ketones*:

$$R\text{-}CH_2\text{-}OH \xrightarrow{[O]} R\text{-}CHO \xrightarrow{[O]} R\text{-}COOH$$

primary alcohol → aldehyde → carboxylic acid

$$R\text{-}CHOH\text{-}R \xrightarrow{[O]} R\text{-}CO\text{-}R$$

secondary alcohol → ketone

- Carboxylic acids are weak acids since they only partially ionize. They have similar reactions to fully ionized mineral acids but they react with less vigor. Sodium salts of carboxylic acids are ionic compounds. Those with short carbon chains are readily soluble in water.
- *Esters* are volatile liquids or low-melting solids. They are usually insoluble in water but soluble in ethanol and diethyl ether. Esters have sweet fruity smells and are used in perfumes, flavorings, and essences.

© Diagram Visual Information Ltd.

Functional groups and properties

Class of compound	Example	Functional group
Alkene	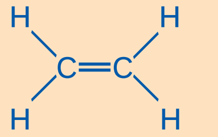 **Ethene**	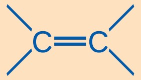
Alcohol	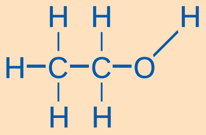 **Ethanole**	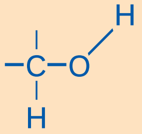
Carboxylic acid	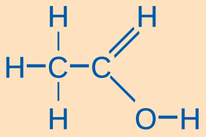 **Ethanoic acid**	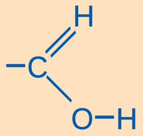
Ester	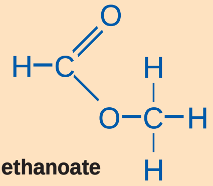 **Methyl methanoate**	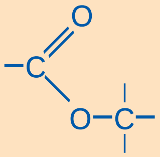

Class of compound	Typical chemical property
Alkene	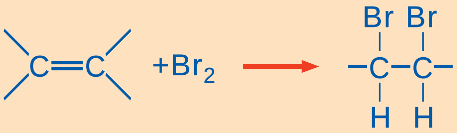 **Decolorizes bromine water**
Alcohol	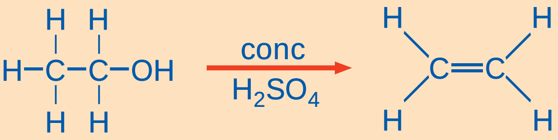 **Decolorizes bromine water**
Carboxylic acid	$2CH_3COOH + Na_2CO_3 \rightarrow 2CH_3COO^-Na^+ + CO_2\uparrow + H_2O$ **Reacts with sodium carbonate solution**
Ester	$HCOOH_3 + NaOH \rightarrow HCOO^-Na^+ + CO_3OH$ **Can be hydrolized by alkali**

Reaction summary: alkanes and alkenes

Alkanes

Combustion

$C_3H_8 + 5O_2 \longrightarrow 3CO_2 + 4H_2O$

Substitution

$CH_4 + Cl_2 \longrightarrow CH_3Cl + HCl$

Cracking

$C_8H_{18} \longrightarrow H_2C = CH_2 + C_6H_{14}$

Alkenes

Hydrogenation

$H_2C = CH_2 + H_2 \longrightarrow H_3C - CH_3$

Substitution

$H_2C = CH_2 + Br_2 \longrightarrow$

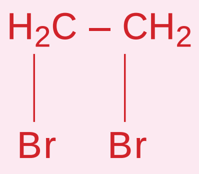

$H_2C(Br) - CH_2(Br)$

General reaction alkene to alkane

$nCH_2 = CH_2 \longrightarrow (CH_2 - CH_2)_n-$

Key words

addition polymerization
alkane
alkene
solvent

Reaction of alkanes and alkenes

- Both *alkanes* and *alkenes* burn readily in a good supply of air to produce carbon dioxide and water.
- Crude oil is a complex mixture of alkanes, which are separated into fractions (components) on the basis of boiling point during the refining process. Some of these fractions provide gasoline, diesel, aviation fuel, and fuel oil.
- The quality of gasoline (how smoothly it burns) in indicated by its octane number, which ranges from 0–100: the higher the octane number the smoother burning the gasoline. The octane number is the percentage by volume of 2,2,4-trimethylpentane (also known as iso-octane) in a mixture of 2,2,4-trimethylpentane and heptane, which has the same knocking characteristics as the gasoline being tested.
- Historically, tetraethyllead(IV) $Pb(C_2H_5)_4$ was added to gasoline as an anti-knock additive to make it burn more smoothly. A growing knowledge of the poisonous nature of lead has resulted in the development of lead-free fuels in which other anti-knock additives, such as MTBE (methyltert-butyl ether), are used.
- Crude oil contains no alkenes, but they are produced in cracking and other refining processes. Alkenes are important feedstock for *addition polymerization* but are also used in gasoline blending, making plasticizers, and as *solvents*.
- Much of the chemistry of the alkenes is the result of the reactive nature of the carbon–carbon double bond. Alkenes undergo addition reactions with a variety of substances.

© Diagram Visual Information Ltd.

Key words

alcohol
aldehyde
carboxylic acid
ester
ethanol
ethene
oxidizing agent

1 Alcohols

- The majority of the world's annual production of *ethanol* is made by the catalytic hydration of *ethene*. A mixture of ethene and steam at 300°C and 70 atmospheres is passed over a phosphoric acid catalyst.
- Ethanol is also made industrially by the fermentation of carbohydrates.
- It can also be prepared in the laboratory using concentrated sulfuric acid and heat.
- Ethanol burns readily in air. In some countries it is used as a blending agent in motor fuels.
- *Alcohols* can be oxidized to *carboxylic acids* by heating with a suitable *oxidizing agent* such as acidified potassium dichromate. The oxidation involves two stages and goes via a group of compounds called *aldehydes*. Under suitable conditions, the ethanal can be removed from the reaction mixture before it is further oxidized to ethanoic acid.

2 Acids

- Salts of short-chain carboxylic acids, like sodium ethanoate, are ionic compounds and are soluble in water.
- Ethanoic acid and ethanol react in the presence of a concentrated sulfuric acid catalyst to form the *ester* ethyl ethanoate. This reaction is reversed by heating ethyl ethanoate with an alkali such as sodium hydroxide solution. The sodium salt formed, sodium ethanoate, can be neutralized by dilute mineral acid to regenerate ethanoic acid.

ethanoic acid + ethanol $\xrightarrow{\text{esterification}}$
ethyl ethanoate

ethyl ethanoate $\xrightarrow{\text{hydrolysis}}$
ethanoic acid + ethanol

© Diagram Visual Information Ltd.

Reaction summary: alcohols and acids

Alcohols

Preparation in industry

$$CH_2 = CH_2 + H_2O \xrightarrow[\text{+70 Atmospheres}]{H_3PO_4 \text{ at } 300°C} CH_3CH_2OH$$

Fermentation

$$C_{12}H_{22}O_{11} + H_2O \xrightarrow[\text{catalyst}]{\text{enzyme}} 2C_6H_{12}O_6 \xrightarrow[\text{catalyst}]{\text{enzyme}} 4CH_3CH_2OH + 4CO_2$$

Preparation in the laboratory

$$CH_2 = CH_2 + H_2O \xrightarrow[\text{+ heat}]{\text{conc } H_2SO_4} CH_3CH_2OH$$

Oxidation by burning

$$CH_3CH_2OH + 3O_2 \longrightarrow 2CO_2 + 3H_2O$$

Oxidation by oxidizing agent

$$CH_3CH_2OH \xrightarrow{K_2CR_2O_7 + \text{dil } H_2SO_4} CH_3 + COOH$$

Reaction to produce an ester

$$CH_3CH_2OH + CH_3CO_2H \xrightarrow[H_2SO_4]{\text{conc}} H_2O + CH_3CO_2CH_2CH_3$$

Organic acids

Reaction giving ionic salt

$$CH_3CO_2 + NaOH \longrightarrow CH_3COO^-Na^+ + H_2O$$

Reaction giving covalent ester

$$CH_3CO_2 + CH_3CH_2OH \longrightarrow CH_3COOCH_2CH_3 + H_2O$$

Reaction giving hydrolysis of an ester

$$CH_3COOCH_2CH_3 + NaOH \longrightarrow CH_3COO^-Na^+ + CH_3CH_2OH$$

Optical isomerism

1 Chiral molecule

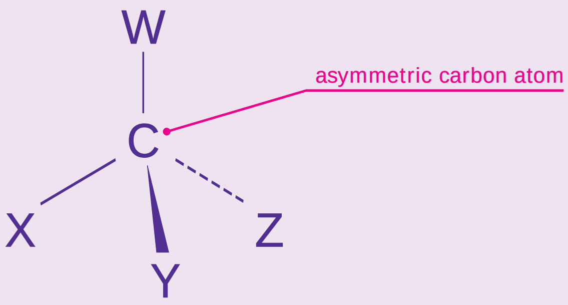

2 Enantioners

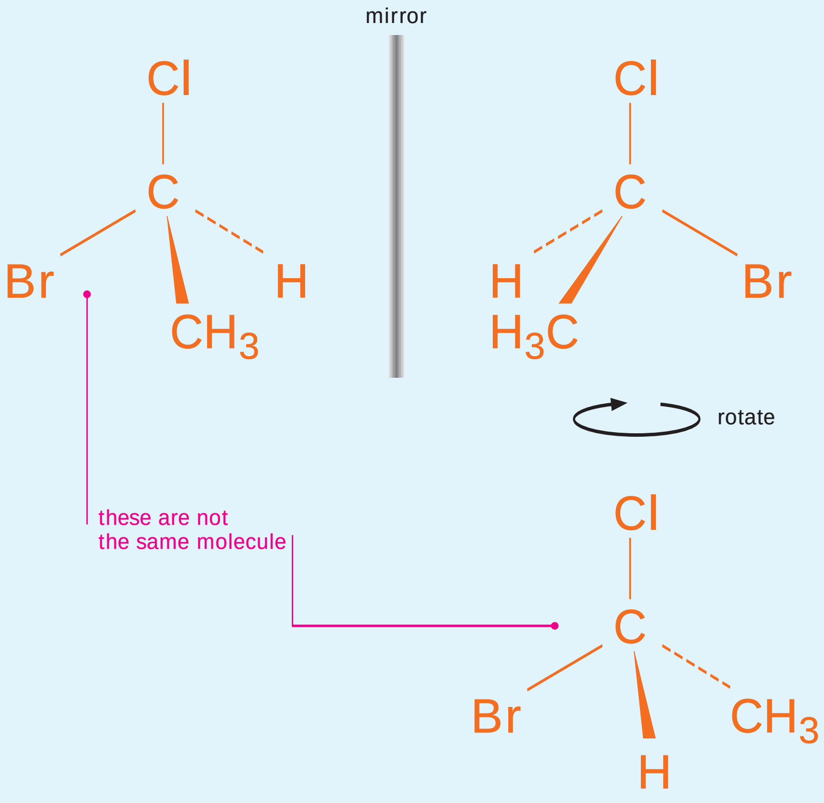

3 Optical activity

Polarized light	→	dextro-rotatory or (+) rotatory
	←	laevo-rotatory or (–) rotatory

Key words

chiral
enantiomer
optical isomerism
racemate

Optical isomerism

- *Optical isomerism* is a form of isomerism in which two isomers are the same in every way except that they are mirror images that cannot be superimposed on each other.

1 Chiral molecule

- When four different groups are attached to a carbon atom, the resulting molecule has no symmetry. The molecule is said to be *chiral*, and the carbon atom at the center is described as asymmetric.

2 Enantiomers

- 1-bromo-1-chloroethane is a chiral molecule. It exists in two forms, called *enantiomers*, that differ only in the way that the bonds are arranged in space.
- The enatiomers of a chiral molecule are mirror images of each other and cannot be superimposed on each other.

3 Optical activity

- Chiral molecules are said to be optically active since they rotate the plane of polarized light. If polarized light is passed through a solution containing only one of the enantiomers, the plane of the light will be rotated either to the right (dextro-rotatory) or to the left (laevo-rotatory). A similar solution containing only the other enantiomer will rotate the plane of the light by the same amount in the opposite direction.
- A solution containing equal amounts of the enantiomers is called a racemic mixture or *racemate*. It is optically inactive since the two effects cancel each other out.

© Diagram Visual Information Ltd.

Amino acids and proteins

Key words

amine	*protein*
amino acid	*zwitterion*
carboxylic acid	
functional group	
optical isomerism	

1 Amino acids

- *Amino acids* are compounds that contain both *amine* (-NH_2) and a *carboxylic acid* (-COOH) *functional groups*.
- Amino acids are generally crystalline solids that decompose on melting. They are soluble in water and insoluble in organic solvents such as ethanol.

2 Alanine

- Like most α-amino acids, alanine contains an asymmetric carbon atom and exhibits *optical isomerism*. There are two forms of alanine; L-alanine and D-alanine (L = laevo-[left] rotatory; D = dextro-[right] rotatory.)

3 Zwitterions

- In aqueous solution, amino acids are able to form ions that carry both positive and negative charge. Such ions are called *zwitterions*. They form by the loss of a proton from the carboxylic acid group and the gain of a proton on the amine group.

4 Proteins

- *Proteins* are polymers consisting of long chains of amino acids. The amino acids join together forming peptide bonds by the loss of water:

$$H_2N\text{-}CHR\text{-}COOH + H_2N\text{-}CHR\text{-}COOH \xrightarrow{-H_2O} H_2N\text{-}CHR\text{-}CONH\text{-}CHR\text{-}COOH$$

- All of the amino acids in proteins are the L-isomers.

1 Amino acids

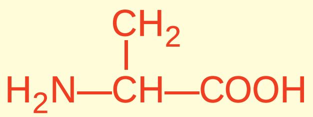

$H_2N-\overset{\overset{\displaystyle CH_2}{|}}{CH}-COOH$

alanine

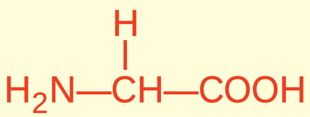

$H_2N-\overset{\overset{\displaystyle H}{|}}{CH}-COOH$

glycine

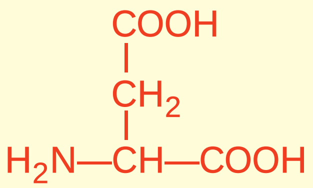

$H_2N-\overset{\overset{\overset{\overset{\displaystyle COOH}{|}}{CH_2}}{|}}{CH}-COOH$

aspartic acid

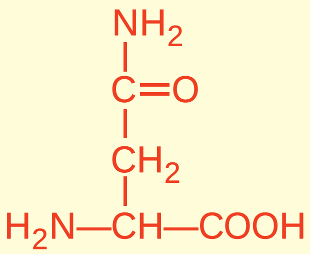

$H_2N-\overset{\overset{\overset{\overset{\overset{\overset{\displaystyle NH_2}{|}}{C=O}}{|}}{CH_2}}{|}}{CH}-COOH$

asparagine

2 Alanine

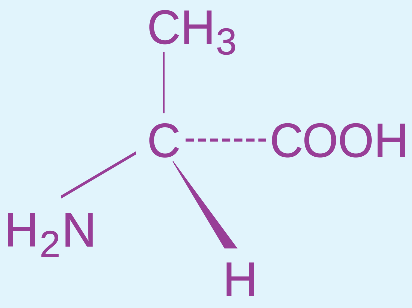

L-alanine

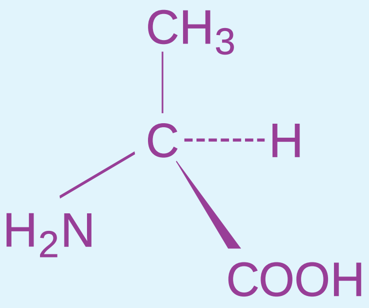

D-alanine

3 Zwitterions

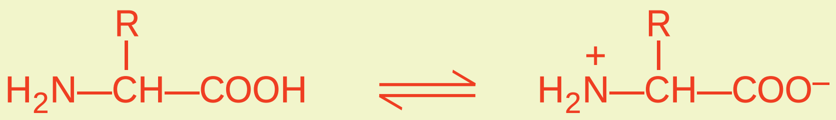

$H_2N-\overset{\overset{\displaystyle R}{|}}{CH}-COOH \rightleftharpoons \overset{+}{H_2N}-\overset{\overset{\displaystyle R}{|}}{CH}-COO^-$

4 Proteins

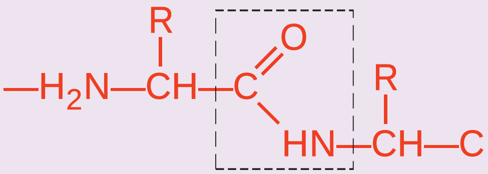

$-H_2N-\overset{\overset{\displaystyle R}{|}}{CH}-\overset{\overset{\displaystyle O}{\|}}{C}-HN-\overset{\overset{\displaystyle R}{|}}{CH}-C$

peptide link

© Diagram Visual Information Ltd.

Monosaccharides

1 Chain structure

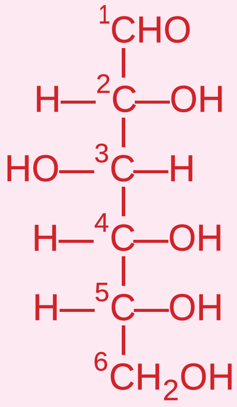

D-glucose

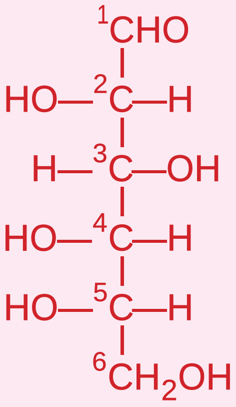

L-glucose

2 Ring structure

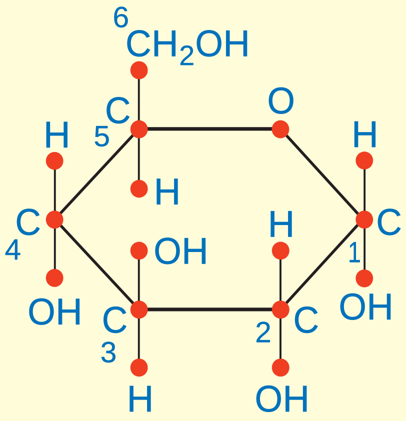

α-D-glucose

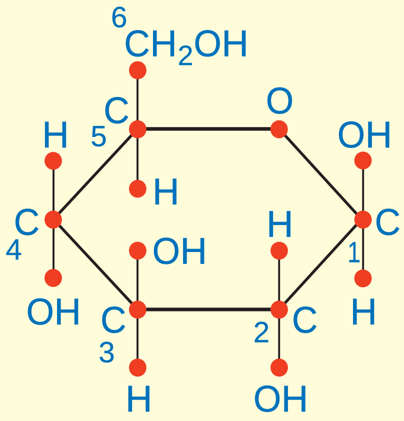

β-D-glucose

3 Hexagonal ring

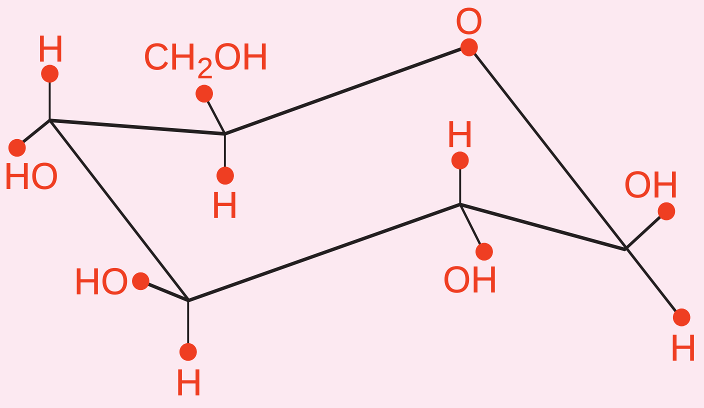

Key words

aldehyde
aldohexose
aldose
anomer
glucose
hexose
monosaccharide

Monosaccharides

- *Monosaccharides* are simple sugars that have between three and six carbon atoms. Those with six carbon atoms are known as the *hexoses* and have the general formula $C_6H_{12}O_6$.
- Monosaccharides with an *aldehyde* group (-CHO) are called *aldoses*.
- *Glucose* has both an aldehyde group and six carbon atoms and is therefore an *aldohexose*.

1 Chain structure

- For simplicity, monosaccharides are sometimes displayed as vertical open chain structures to which the -H and –OH groups are attached.
- Aldohexoses contain four asymmetrical carbon atoms: C-2, C-3, C-4, and C-5. There are 8 different possible ways of arranging the –H and –OH groups on these carbon atoms, and each of these has two optical isomers, making a total of 16.
- The most important of these are the two optical isomers of glucose.
- For glucose the D- and L- indicate the configuration of the –H and –OH groups on C-5.

2 Ring structure

- In reality, solid monsaccharides do not exist as open chain structures but as ring structures.
- In Howarth projections of monosaccharides, groups are shown on vertical bonds above and below a flat hexagonal ring.
- D-glucose can exist in two separate crystalline forms known and α-D-glucose and β-D-glucose. These forms are known as *anomers*.

3 Hexagonal ring

- The hexagonal ring in a monosaccharide is not flat but in the form of a chair.

© Diagram Visual Information Ltd.

Key words

cellulose
disaccharide
glycogen
monosaccharide
polysaccharide
starch
sucrose

Disaccharides and polysaccharides

Di-and polysaccharides

- A *disaccharide* is formed when two *monosaccharides* join together. A molecule of water is lost and a glycosidic link is formed.
- A *polysaccharide* is a polymer formed by the joining of many monosaccharide units.

1 Sucrose

- *Sucrose*, the sugar widely used on foods, is a disaccharide.

2 Cellulose

- *Cellulose*, a polysaccharide, provides plant cells with a rigid structure.

3 Starch

- *Glycogen* is the storage polysaccharide of animals.
- *Starch* is the storage polysaccharide of plants.

1 Sucrose

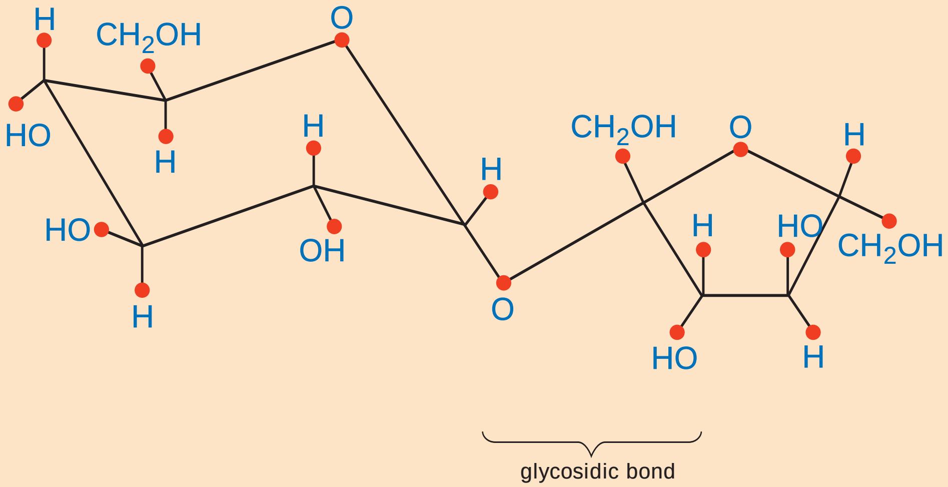
H
CH2OH
O
HO
H
H
HO
OH
H
H
CH2OH
O
H
H
HO
CH2OH
O
HO
H
glycosidic bond

2 Cellulose

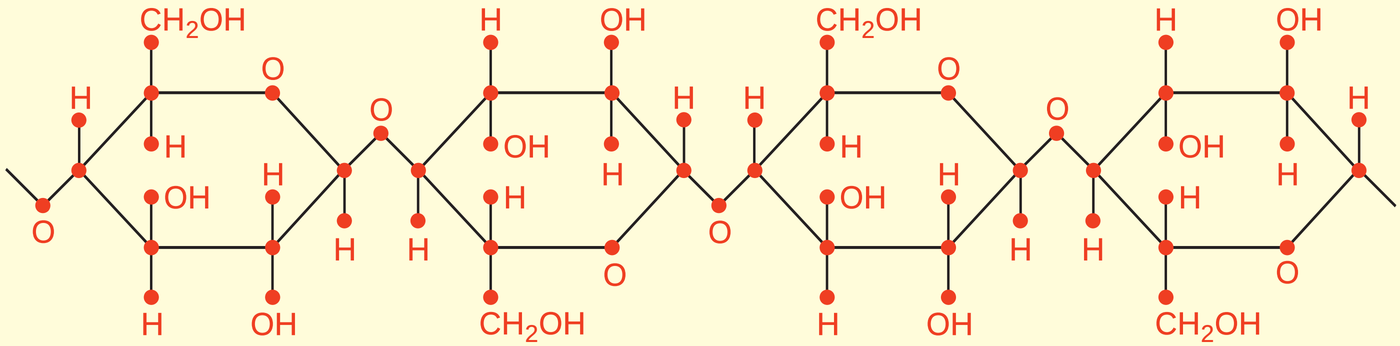
CH2OH
O
H
H
OH
H
O
H
OH
H
OH
H
H
O
CH2OH
O
H
H
OH

3 Starch

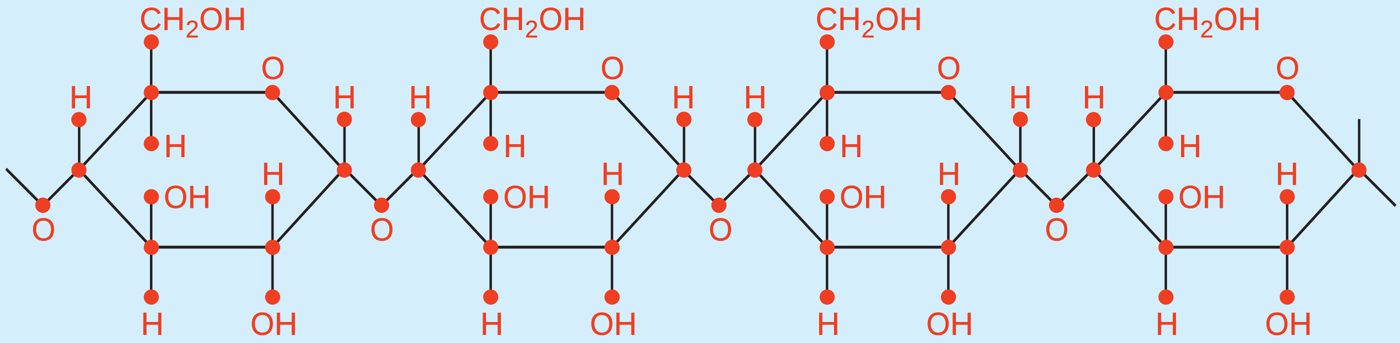
CH2OH
O
H
H
OH
H
O
H
OH
H
OH

© Diagram Visual Information Ltd.

Ionizing radiation

1 Alpha particles

α-radiation consists of a stream of α particles

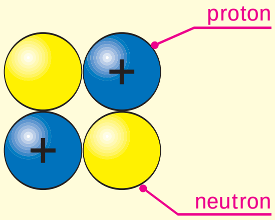

Produces intense ionization in a gas

2 Beta particles

β-radiation consists of a stream of β particles

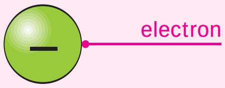

Produces less intense ionization in a gas than a particles

3 Gamma radiation

γ-radiation is a form of electromagnetic radiation

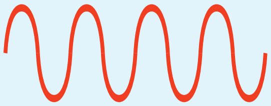

Wavelength $<10^{-12}$ m

Frequency $>10^{21}$ Hz

Only weakly ionizes a gas

4 Radiation in laboratories

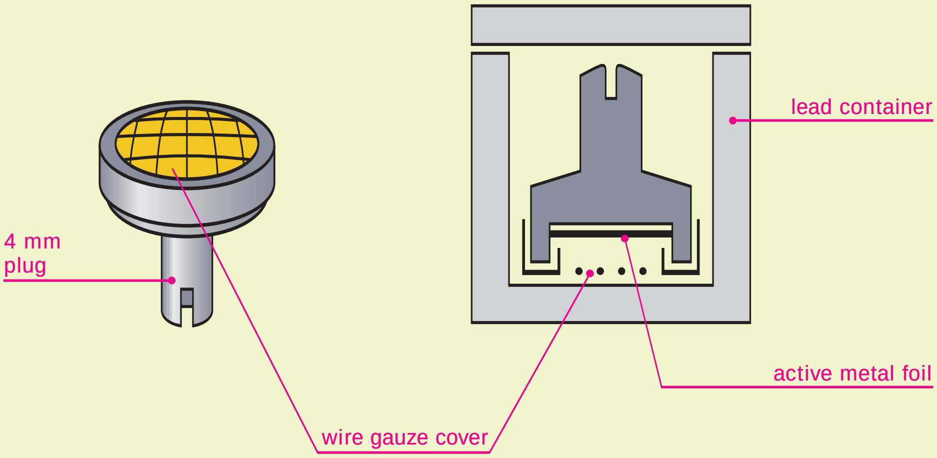

Key words

alpha particle
beta particle
gamma radiation
ionizing radiation

Ionizing radiation

- *Ionizing radiation* is any radiation capable of displacing electrons from atoms or molecules and so producing ions. Examples include *alpha particles*, *beta particles*, and *gamma radiation*.

1 Alpha particles

- An alpha (α) particle has the same structure as a helium nucleus (two protons and two neutrons).
- Alpha particles are relatively heavy, high-energy particles with a positive charge.
- Alpha particles produce intense ionization in a gas.
- Emission speeds are typically of the order of 5–7 percent of the speed of light.

2 Beta particles

- A beta (β) particle is a fast-moving electron with a negative charge.
- Beta particles produce less ionization in a gas than alpha particles and on average produce only 1/1000th as many ions per unit length.
- Emission speeds can be as high as 99 percent of the speed of light.

3 Gamma radiation

- Gamma (γ) rays ionize gas only weakly and on average produce only 1/1000th as many ions per unit length as beta particles.

4 Radiation in laboratories

- Sources of radiation used for laboratory experiments are usually supplied mounted in a holder. The active material is sealed in metal foil, which is protected by a wire gauze cover. When not in use, the material is stored in a small lead container.

© Diagram Visual Information Ltd.

Radiation detectors

Key words

ionizing radiation
radiation
radioactivity

Detectors

- *Radioactivity* is invisible, but because it affects the atoms that it passes, scientists can easily detect it using a variety of methods.

1 Spark counter

- High voltage is applied between the stiff wire (anode) and the gauze (cathode) and reduced until it just stops sparking.
- When a radium source is brought near the gauze, the air between the wire and the gauze is ionized, and sparks are seen and heard at irregular intervals.

2 Cloud chamber

- When air containing ethanol vapor is cooled, it becomes saturated. If *ionizing radiation* passes through this air, further cooling causes the vapor to condense on the ions created in the air. The result is a white line of tiny liquid droplets that shows up as a track when illuminated.

3 GM tube

- When *radiation* enters the metal tube, either through the mica window or through the tube wall, it creates argon ions and electrons. These are accelerated toward the electrodes and collide with other argon atoms. On reaching the electrodes, the ions produce a current pulse, which is amplified before being fed to a pulse counter.

4 Testing absorption

- The ability of materials to absorb alpha, beta, and gamma radiation can be tested by placing the material between a radioactive source and a GM tube and comparing the count per minute with the count over the same period when the material is removed.

1 Spark counter

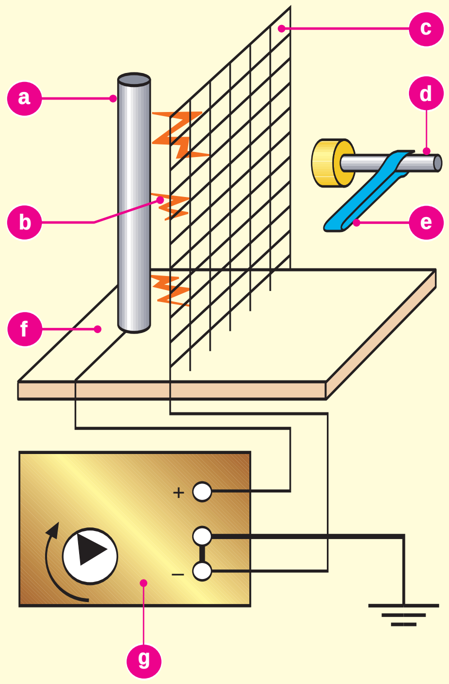

2 Cloud chamber

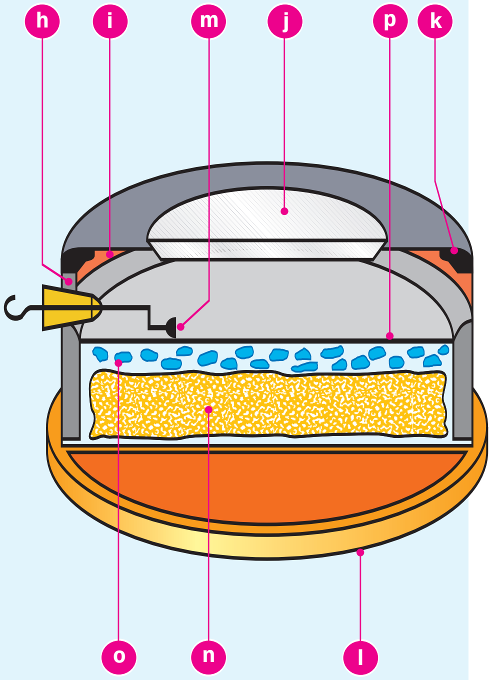

3 Geiger-Muller tube (GM tube)

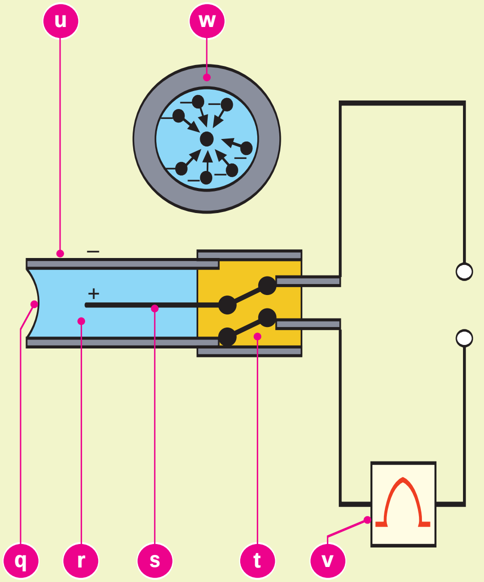

4 Testing absorbtion of alpha, beta, and gamma radiation

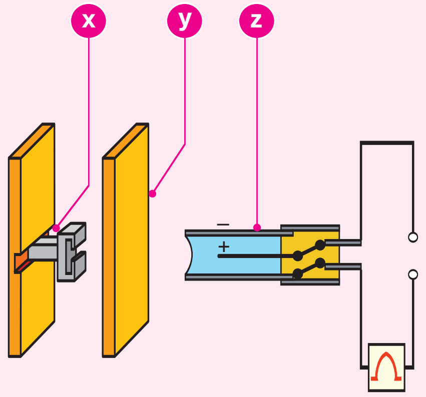

a stiff wire (anode)
b sparks
c wire gauze (cathode)
d radium source
e forceps
f insulating base
g E.h.t. supply
h circular transparent plastic chamber
i super-cooled vapor
j transparent lid
k felt strip soaked with alcohol and water
l base
m radioactive source
n foam sponge
o crushed dry ice
p black metal base plate
q mica window
r argon gas at low pressure
s anode wire
t insulator
u cathode metal tube
v pulse counter
w electrons are pulled toward the anode wire in an avalanche
x source
y absorbing material
z GM tube

Properties of radiations: penetration and range

1 Penetration of radiation

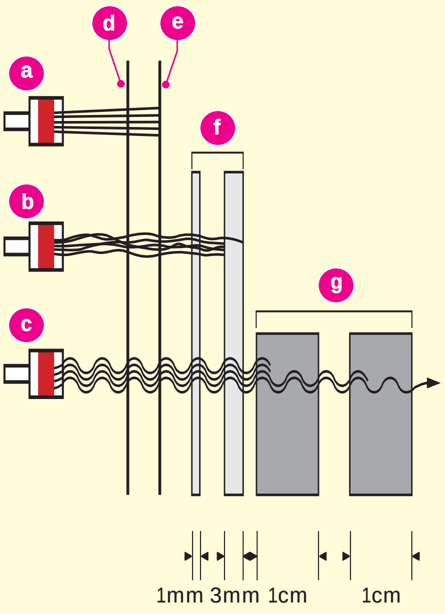

a α – source
b β – source
c γ – source
d metal foil
e paper
f aluminum
g lead

2 Range of radiation in air

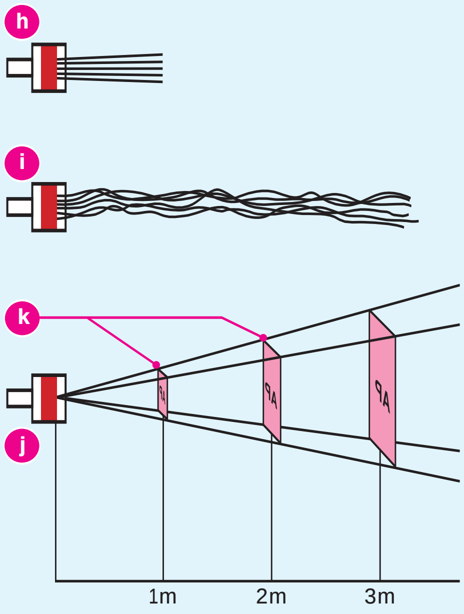

h α – a few centimeters
i β – a few meters
j γ – many meters
k area covered by γ radiation at 1m distances

3 The inverse square law for gamma radiation penetration

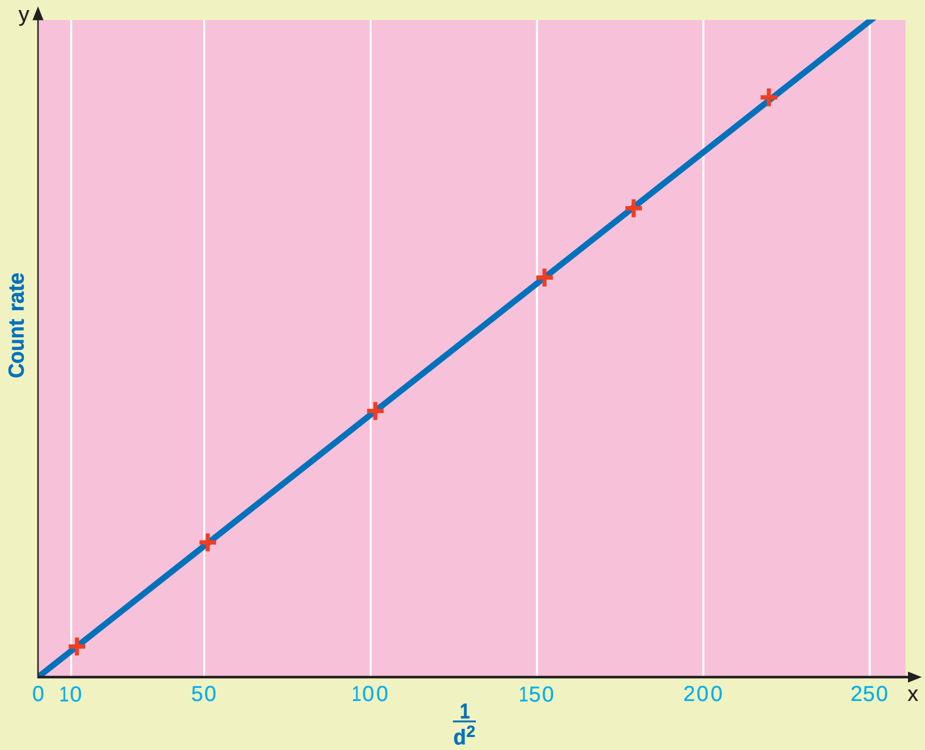

Key words

alpha particle
beta particle
gamma radiation
radiation

1 Penetration

- Alpha, beta, and *gamma radiation* penetrate by different amounts.
- Alpha radiation is the least penetrating and is stopped by a sheet of paper or very thin metal foil.
- Beta radiation is stopped by aluminum a few millimeters thick.
- Gamma radiation is most penetrating, and is only stopped by a thick block of lead.

2 Range

- The penetrating power of alpha, beta, and gamma radiation is reflected in the distance that they can travel through air. *Alpha particles* can only travel a few centimeters before colliding with air particles. *Beta particles* travels a few meters, while gamma radiation can travel many meters.

3 Gamma penetration

- Gamma rays are highly penetrating because they have relatively little interaction with matter. There is very little absorption or scattering as they pass through air.
- The intensity falls off with distance according to the inverse square law:

  $$I = \frac{k}{d^2}$$

  where I is intensity, d is the distance from the source, and k is a constant. At a distance x, the intensity of the gamma radiation:

  $$I_x = \frac{k}{x^2}$$

  At a distance 2x, the intensity of the gamma radiation:

  $$I_{2x} = \frac{k}{(2x)^2} = \frac{k}{4x^2}$$

  As the distance increases by a factor of 2, the intensity of the gamma radiation decreases by a factor of 4.

© Diagram Visual Information Ltd.

Properties of radiations: in fields

Key words

alpha particle
beta particle
electric field
gamma radiation

Electric and magnetic fields

- An *electric field* is a field extending outward in all directions from a charged particle.
- A magnetic field is an area of force that exists around a magnetic body or a current-carrying conductor. Alpha, beta, and gamma radiation behave differently in both.

1 Electric field

- Alpha radiation is composed of positively charged particles. A stream of *alpha particles* is deflected when passing through the electric field between two oppositely charged plates. The particles are repelled from the positively charged plate and attracted toward the negatively charged plate.
- Beta radiation is composed of negatively charged particles. A stream of *beta particles* is deflected by an electric field in the opposite direction to alpha particles. The deflection is greater because the beta particles have a much smaller mass.
- *Gamma radiation* is not deflected by an electric field. This is evidence that gamma radiation carries no charge.

2 Magnetic field

- Alpha radiation is deflected by a strong magnetic field. Weak magnetic fields have no noticeable effect due to the greater mass of alpha particles compared to beta particles.
- Beta radiation is deflected by a relatively weak magnetic field. Beta radiation is deflected in the opposite direction to alpha radiation, indicating its particles carry an opposite charge.
- Gamma radiation is not deflected by a magnetic field, indicating that gamma radiation carries no charge.

Radiation	1 Electric field	2 Magnetic field
Alpha	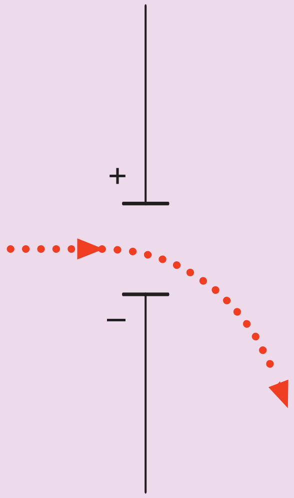	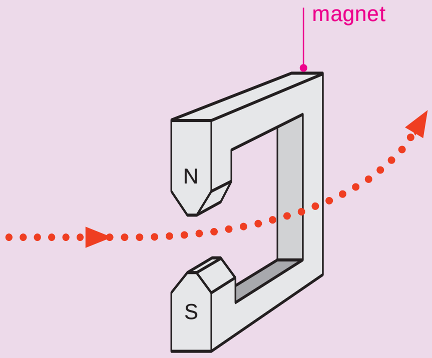
Beta	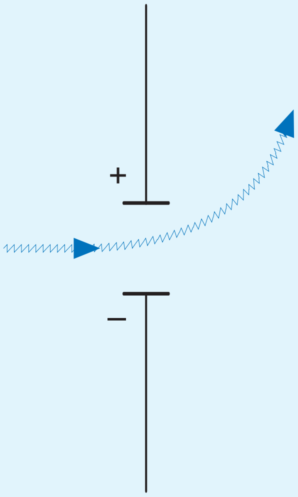	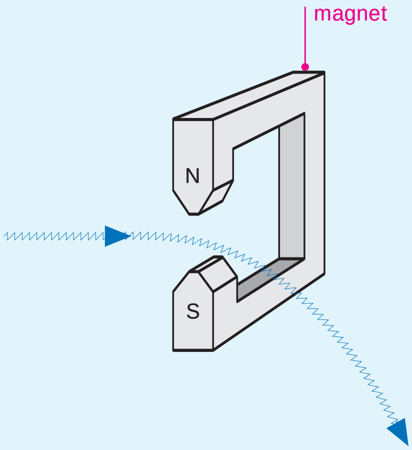
Gamma	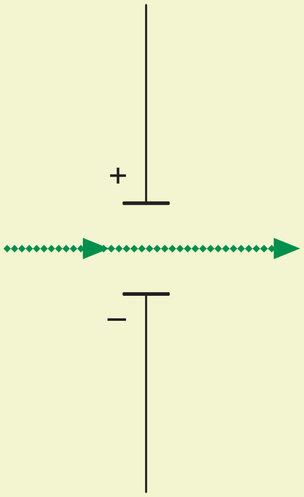	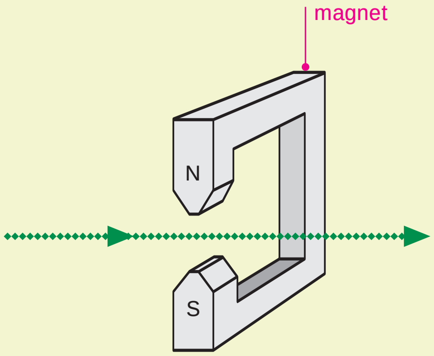

Stable and unstable isotopes

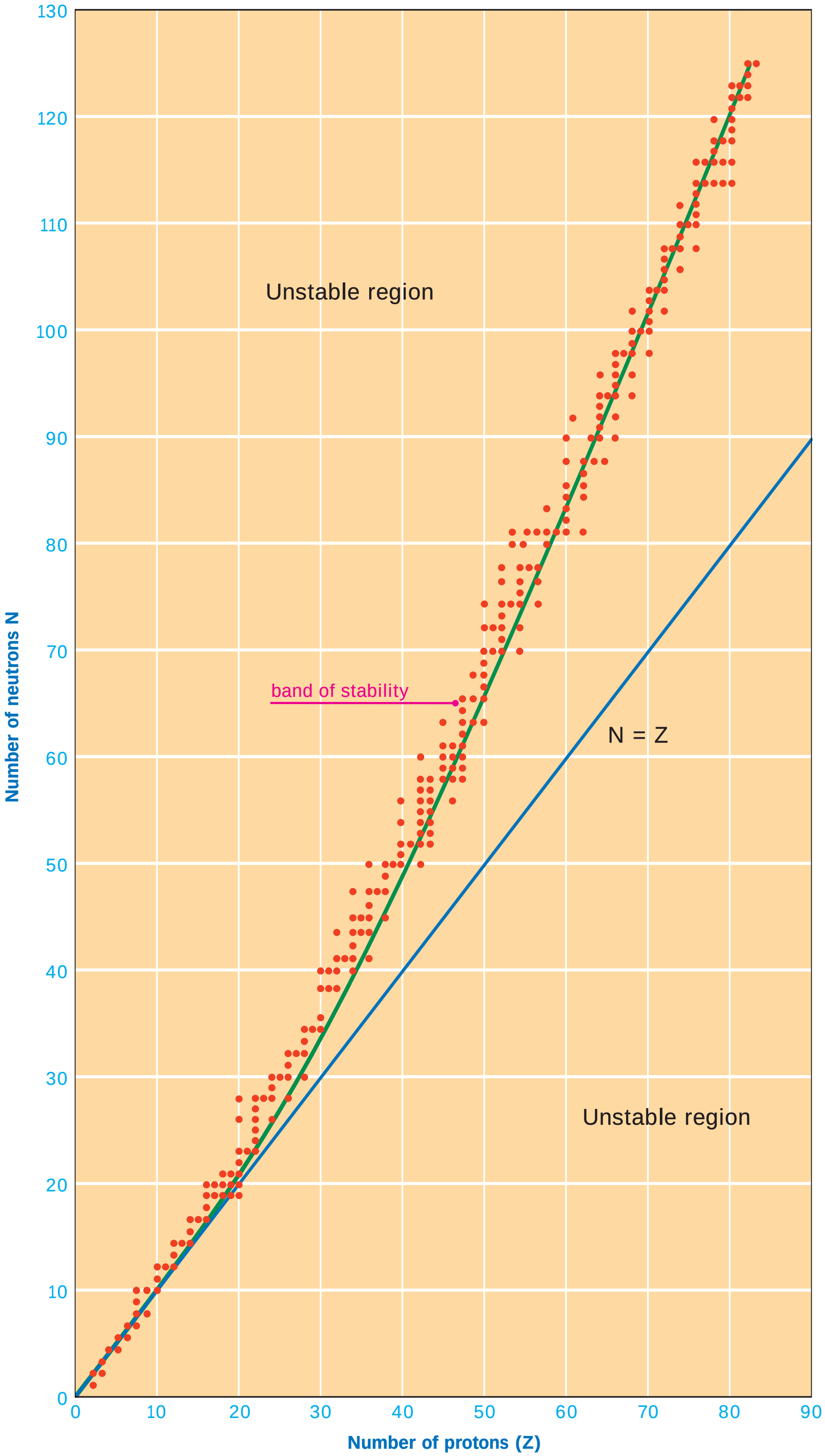

Key words

isotope
nucleon
nuclide

Stability

- The stability of *isotopes* is based on the ratio of neutrons and protons in their nucleus. Although most nuclei are stable, some are not and spontaneously decay, emitting radiation.
- The lightest stable *nuclides* (particular isotopes of an element) have almost equal numbers of protons and neutrons. The heavier stable nuclides require more neutrons than protons. The heaviest stable nuclides have approximately 50 percent more neutrons than protons.

Odd-even rule

- Isotopes tend to be more stable when they have even numbers of protons and neutrons than when they have odd. This is the result of the spins of the *nucleons* (the constituents of the atomic nucleus). When two protons or neutrons have paired spins (spins in opposite directions), their combined energy is less than when they are unpaired.

Decay

- When unstable nuclides disintegrate, they tend to produce new nuclides that are nearer to the stability line. This will continue until a stable nuclide is formed.
- An unstable nuclide above the band of stability decays by beta emission. This increases the proton number and decreases the neutron number. Thus, the neutron to proton ratio is decreased.
- An unstable nuclide below the band of stability disintegrates so as to decrease the proton number and increase the neutron to proton ratio. In heavy nuclides this can occur by alpha emission.

© Diagram Visual Information Ltd.

Half-life

Key words

alpha particle
half-life
isotope
nuclide

1 Half-life

- *Half-life* is the time required for half the nuclei in a sample of an *isotope* to undergo *radioactive decay*.
- Radioactive decay is a completely random process in which nuclei disintegrate independently of each other or external factors such as temperature and pressure.

2 Rate of decay

- There are always very large numbers of active *nuclides* even in small amounts of radioactive material, so statistical methods can be employed to predict the fraction that will have decayed, on average, over a given period of time.
- The rate of decay of a nuclide at any time is directly proportional to the number of nuclei, N, of the nuclide:

  $$\frac{-dN}{dt} \propto N \text{ or } \frac{dN}{dt} = -\lambda N$$

  where N is the number of undecayed nuclei and λ is the decay constant. The minus sign indicates that the number of undecayed nuclei falls with time. Integrating this gives the exponential law equation:

  $$N_t = N_0 e^{-\lambda t}$$

  where N_0 is the number of undecayed atoms at time t = 0 and N_t the number of undecayed atoms after time t.
- After one half life ($t_{1/2}$) has passed, the number of undecayed atoms remaining in the sample will be $N_0/2$. Substituting this into the exponential law equation for N_t and taking natural logs of both sides provides a mathematical relationship between the decay constant and the half life of a radioactive atom:

  $$t_{(1/2)} = \frac{0.693}{\lambda}$$

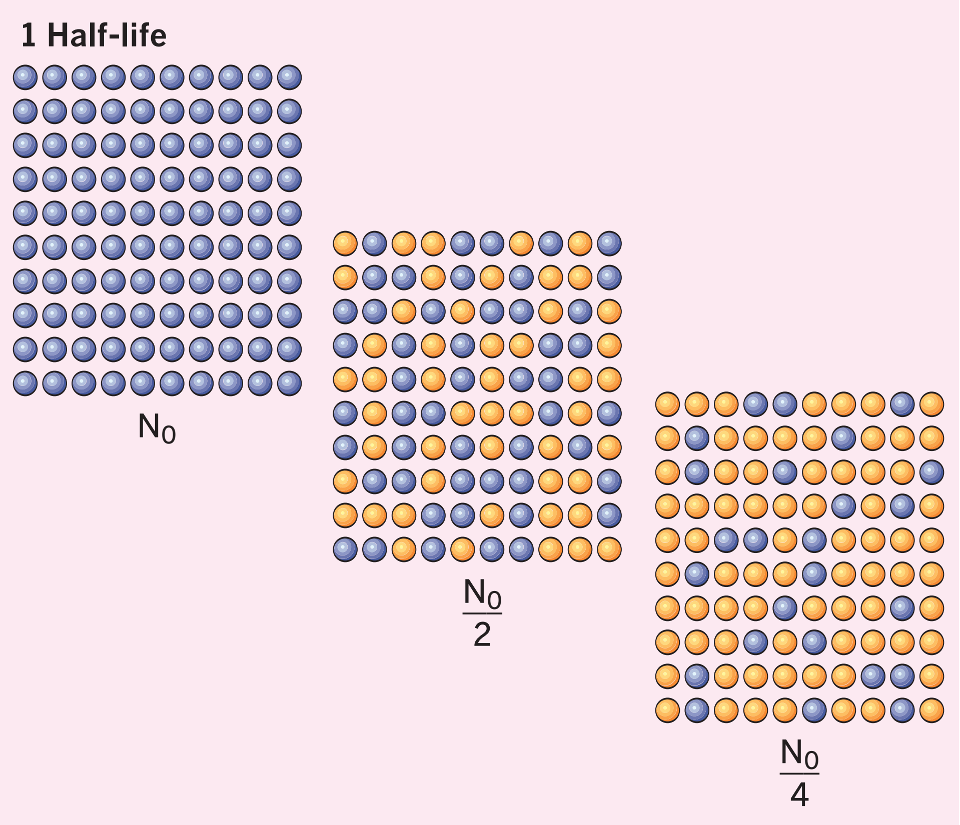

1 Half-life

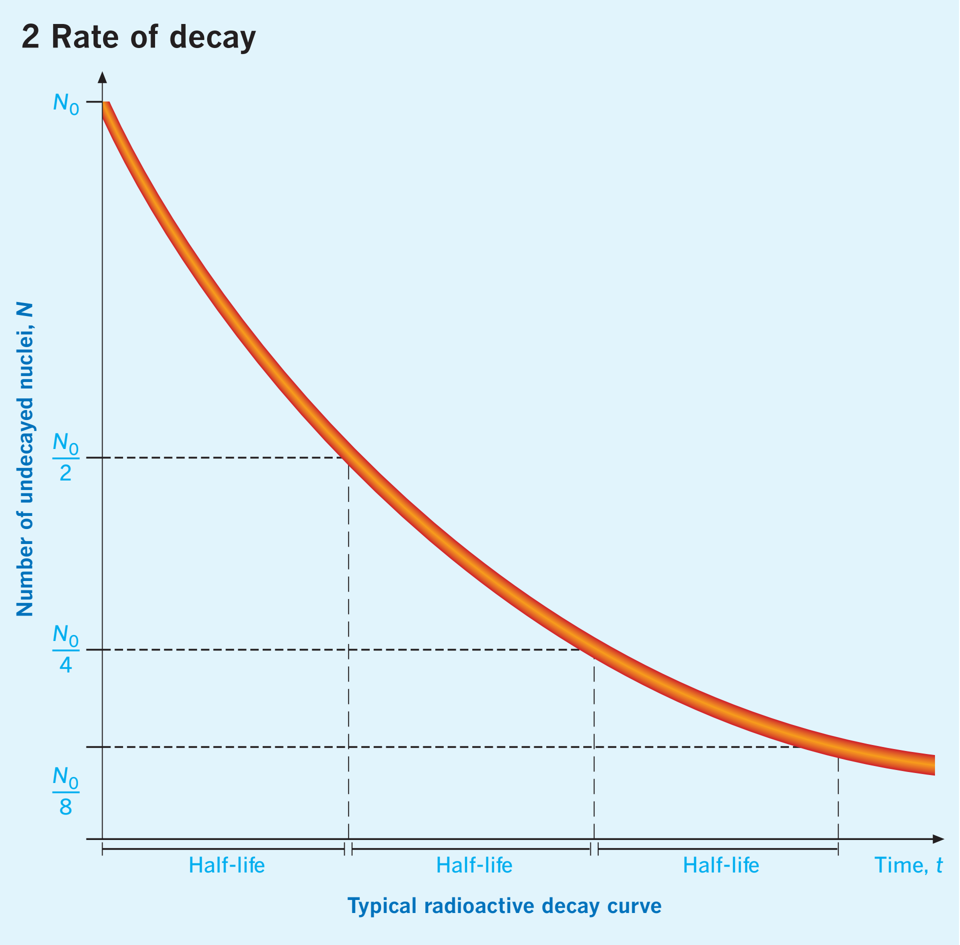

2 Rate of decay

Typical radioactive decay curve

Measuring half-life

1 Half-life of radon

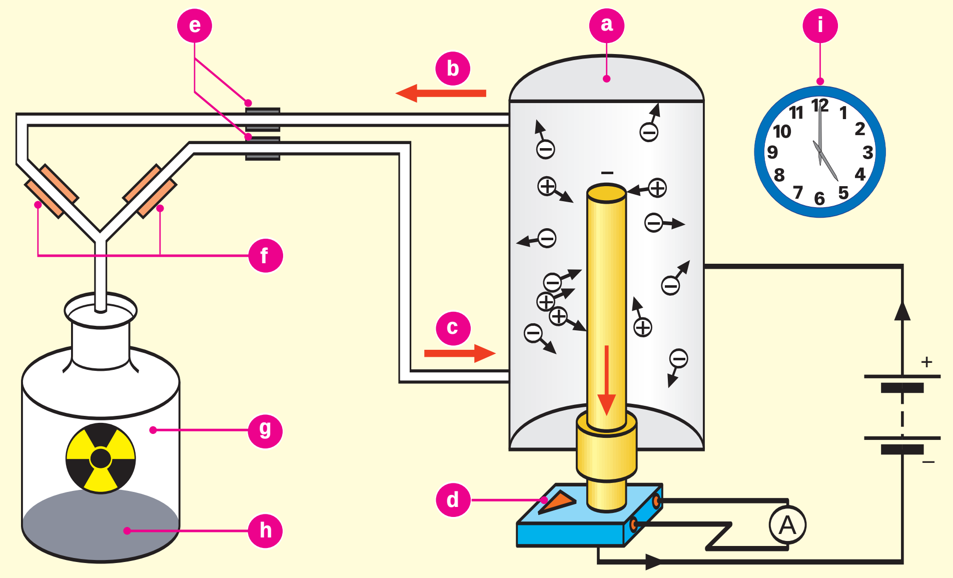

a ionization chamber
b air
c radon
d d.c. amplifier
e clips
f valves
g squeezable polyethylene bottle
h thorium hydroxide powder
i clock

2 Exponential decay: decay curve for radon gas

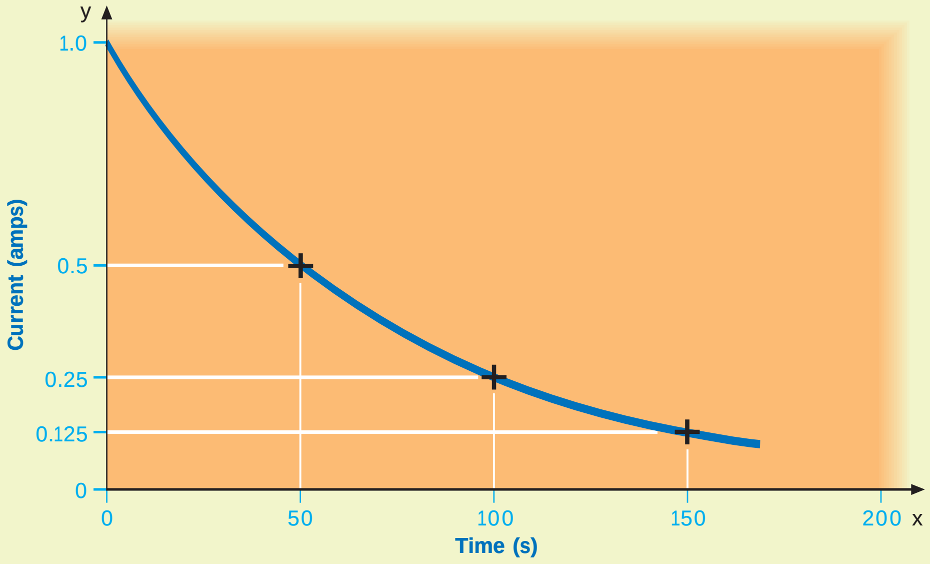

3 Radon decay

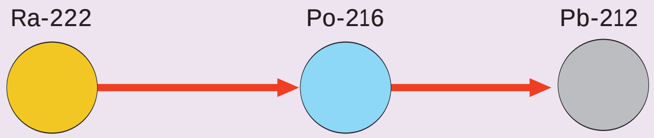

Key words

alpha particle
half-life
isotope

1 Half-life of radon

- Thorium decays to produce the radioactive *isotope* radon-220. This isotope is sometimes referred to as thoron.
- The bottle containing thorium hydroxide powder is squeezed a few times to transfer some radon-220 to the flask. The clips are then closed.
- As the radon decays, the ionization current decreases. It is always a measure of the number of *alpha particles* present and, therefore, the proportion of radon-220 remaining.
- The current is noted every 15 seconds for 2 minutes and then every 60 seconds for several minutes.

2 Exponential decay

- A graph of current against time is plotted.
- In this experiment, the *half-life* is indicated by the amount of time taken for the current to fall to half of its original value.
- The half-life of radon-220 is approximately 55 seconds.

3 Radon decay

- Radon-220 decays with the loss of an alpha particle to form polonium-216, which decays to form lead-212. The half life of polonium-216 is 0.145 seconds, and the half life of lead-212 is 10.64 hours.

© Diagram Visual Information Ltd.

Radioactive isotopes

Key words

alpha particle
gamma radiation
irradiation
isotope

1 Tracers

- Radioactive *isotopes* are used as tracers to monitor the movement of substances in plants and animals. A solution containing radioactive phosphorus-32 is introduced into the stem of a plant. A Geiger counter is used to detect the movement of the isotope through the plant.

2 Thyroid monitor

- A solution containing iodine-131 is introduced to the bloodstream of a patient with a defective thyroid. A Geiger counter is used to detect the isotope and monitor thyroid activity.

3 Food preservation

- Food is irradiated by exposing it to *gamma radiation*. *Irradiation* destroys disease-causing bacteria as well as those that spoil food, so the shelf life of food is extended.

4 Sterilization

- Gamma radiation is used to sterilize medical equipment.

5 Smoke detectors

- Americium-241, a source of alpha radiation, is widely used in smoke detectors. The *alpha particles* ionize the air in the sensing circuit. Any smoke particles interfere with this and cause a change in the current, which triggers an alarm.

6 Duration of death

- All organisms contain a specific ratio of radioactive carbon-14 to carbon-12. When an organism dies, no carbon-14 is added. After death, carbon-14 decays at a predictable rate: the half-life is 5,700 years. By comparing the ratio of carbon-14 to carbon-12, it is possible to say when an organism died.

1 Tracers

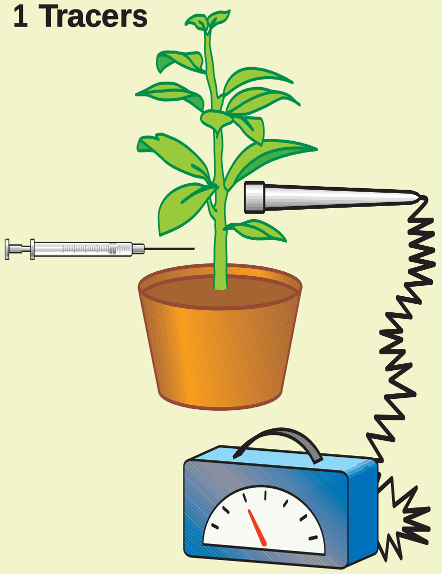

2 Thyroid monitor

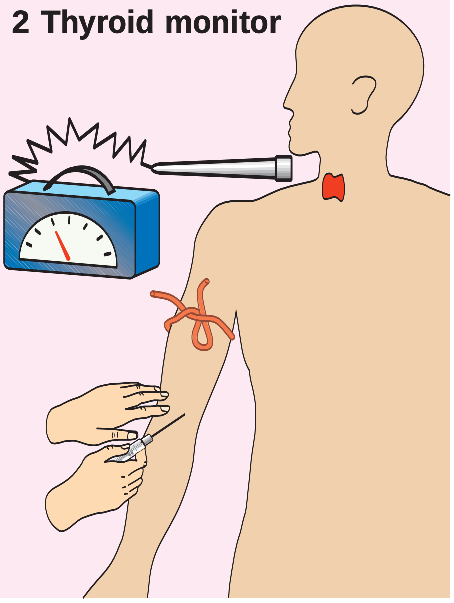

3 Food preservation

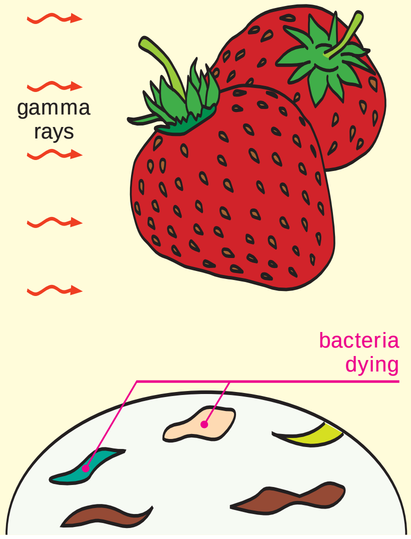
gamma
rays
bacteria
dying

4 Sterilization

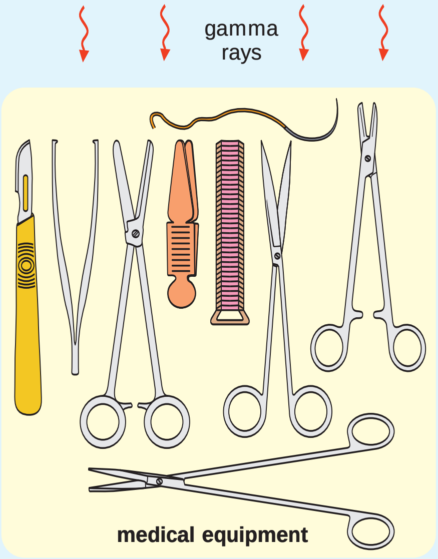
gamma
rays
medical equipment

5 Smoke detectors

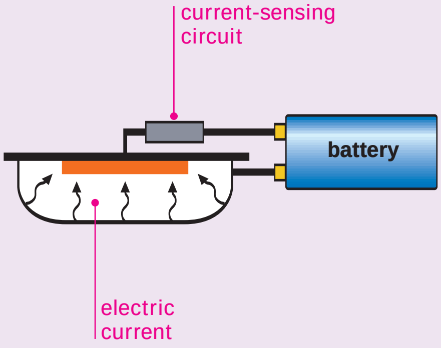
current-sensing
circuit
battery
electric
current

6 Duration of death

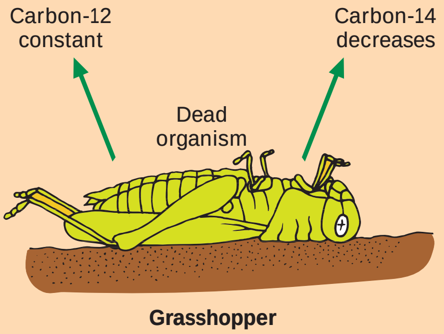
Carbon-12
constant
Carbon-14
decreases
Dead
organism
Grasshopper

Nuclear fusion

1 Nuclear fusion

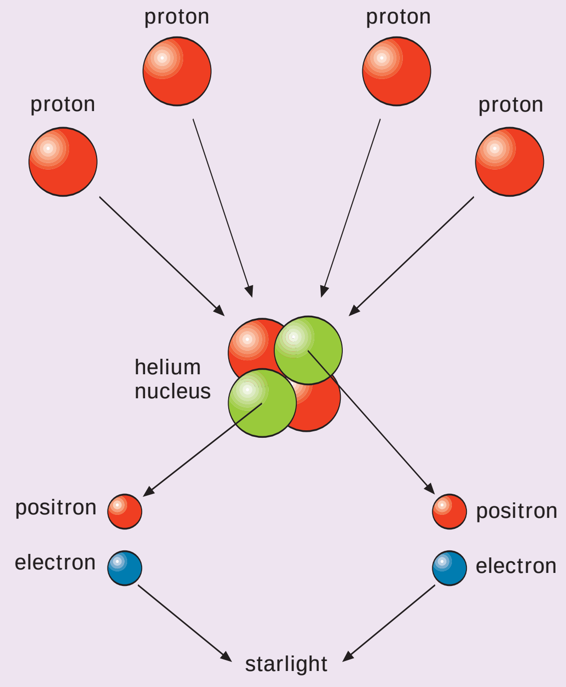

2 Fusion of deuterium

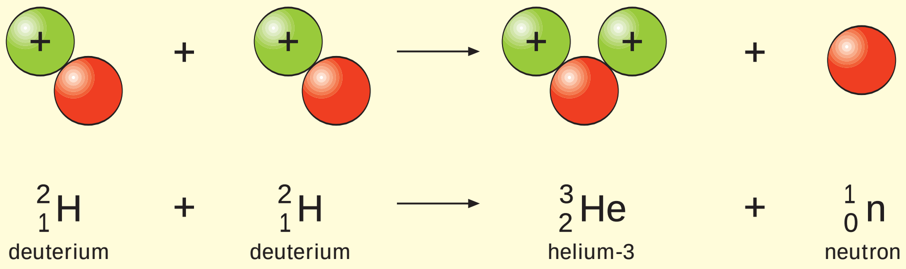

$^{2}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} \longrightarrow {}^{3}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}$

3 Fusion of deuterium and tritium

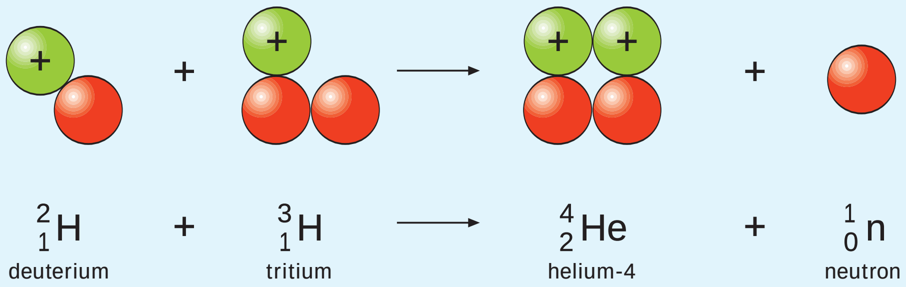

$^{2}_{1}\mathrm{H} + {}^{3}_{1}\mathrm{H} \longrightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}$

Key words

fusion
isotope

1 Nuclear fusion

- In nuclear *fusion*, two or more light atomic nuclei join to make a more massive one. During the process, some of the mass of the nuclei is converted into energy. Nuclear fusion, which first occurred during the Big Bang, powers stars. It also occurs in hydrogen bombs. Currently scientists are working to control fusion so it can be used in nuclear reactors.

2 Deuterium

- Deuterium is an *isotope* of hydrogen known as heavy hydrogen. The nucleus of a deuterium atom consists of one neutron and one proton.
- The fusion of two deuterium nuclei results in the formation of a helium-3 nucleus. A small amount of mass is converted into energy:

  Mass of two deuterium nuclei =
  2 x 2.014 = 4.028 u
  Mass of helium-3 nucleus plus a neutron =
  3.016 + 1.009 = 4.025 u
  Mass converted to energy by fusion =
  4.028 – 4.025 = 0.003 u
  Energy released by the fusion reaction = 4.5×10^{-13} J
  Energy released per kilogram of deuterium is approximately 9×10^{13} J.

3 Tritium

- Tritium is another isotope of hydrogen. The nucleus of a tritium atom consists of two neutrons and one proton.
- The fusion of a deuterium nucleus and a tritium nucleus results in the formation of a helium-4 nucleus and the release of energy. The energy released per kilogram of deuterium and tritium is approximately 30×10^{13} J.
- This reaction produces more energy, and the fusion takes place at a lower temperature.

© Diagram Visual Information Ltd.

Nuclear fission

Key words

chain reaction
fission

Nuclear fission

- In nuclear *fission*, a heavy atomic nucleus divides to make two smaller ones. Some of the mass of the nuclei is converted into energy during the process.

1 Reaction with uranium

- In a nuclear reaction with uranium and slow-moving neutrons, the nucleus of the uranium-235 atom undergoes fission and forms two smaller nuclei (lanthanum-148 and bromine-85) plus three neutrons. A small amount of mass is converted to energy.

2 Chain reaction

- A nuclear *chain reaction* is a series of self-sustaining reactions in which the particles released by one nucleus trigger the fission of at least as many other nuclei.
- Under normal circumstances, only a very small proportion of fission neutrons act in this way. However, if there is a sufficient amount of a radioactive isotope, a chain reaction can start.
- In an atomic bomb, an increasing uncontrolled chain reaction occurs in a very short time when two pieces of uranium-235 (or plutonium-239) are rapidly brought together.
- In a nuclear power station, the chain reaction is steady and controlled, so only a limited number of fission neutrons bring about further fission reactions.

1 Reaction with uranium

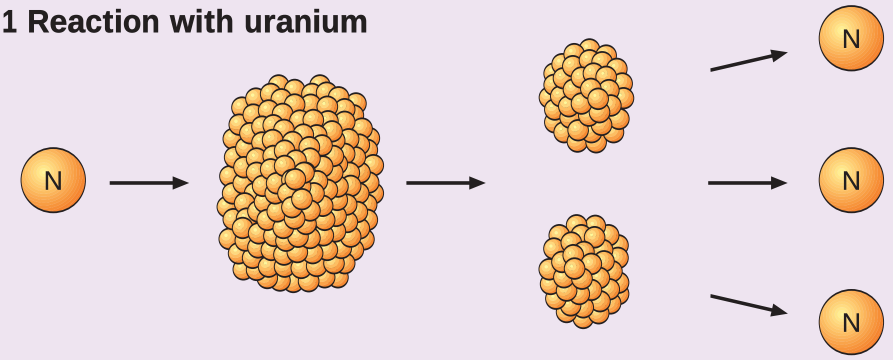

$$^{1}_{0}n + ^{235}_{92}U \longrightarrow ^{148}_{57}La + ^{85}_{35}Br + 3\,^{1}_{0}n$$

2 Chain reaction

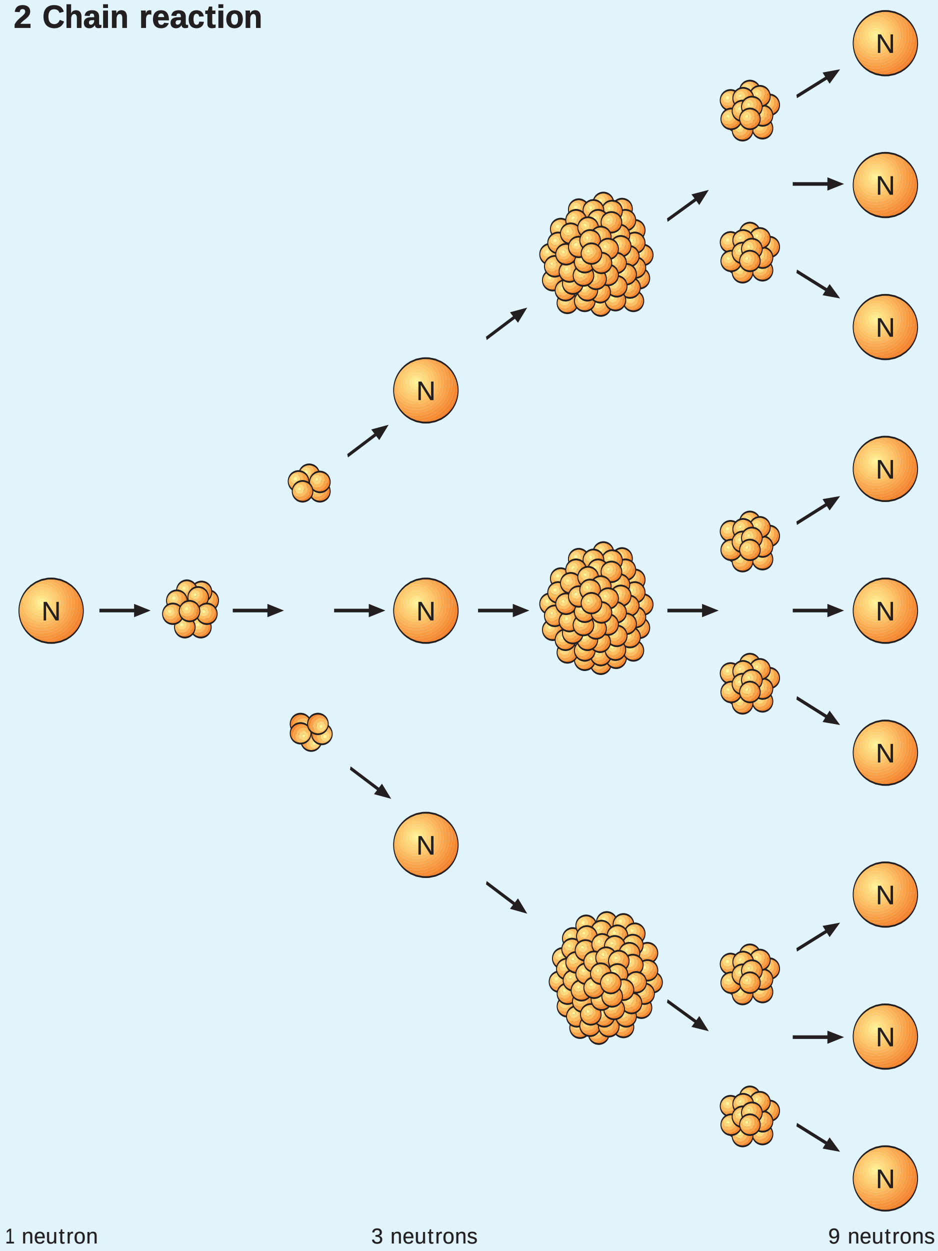

Nuclear reactor

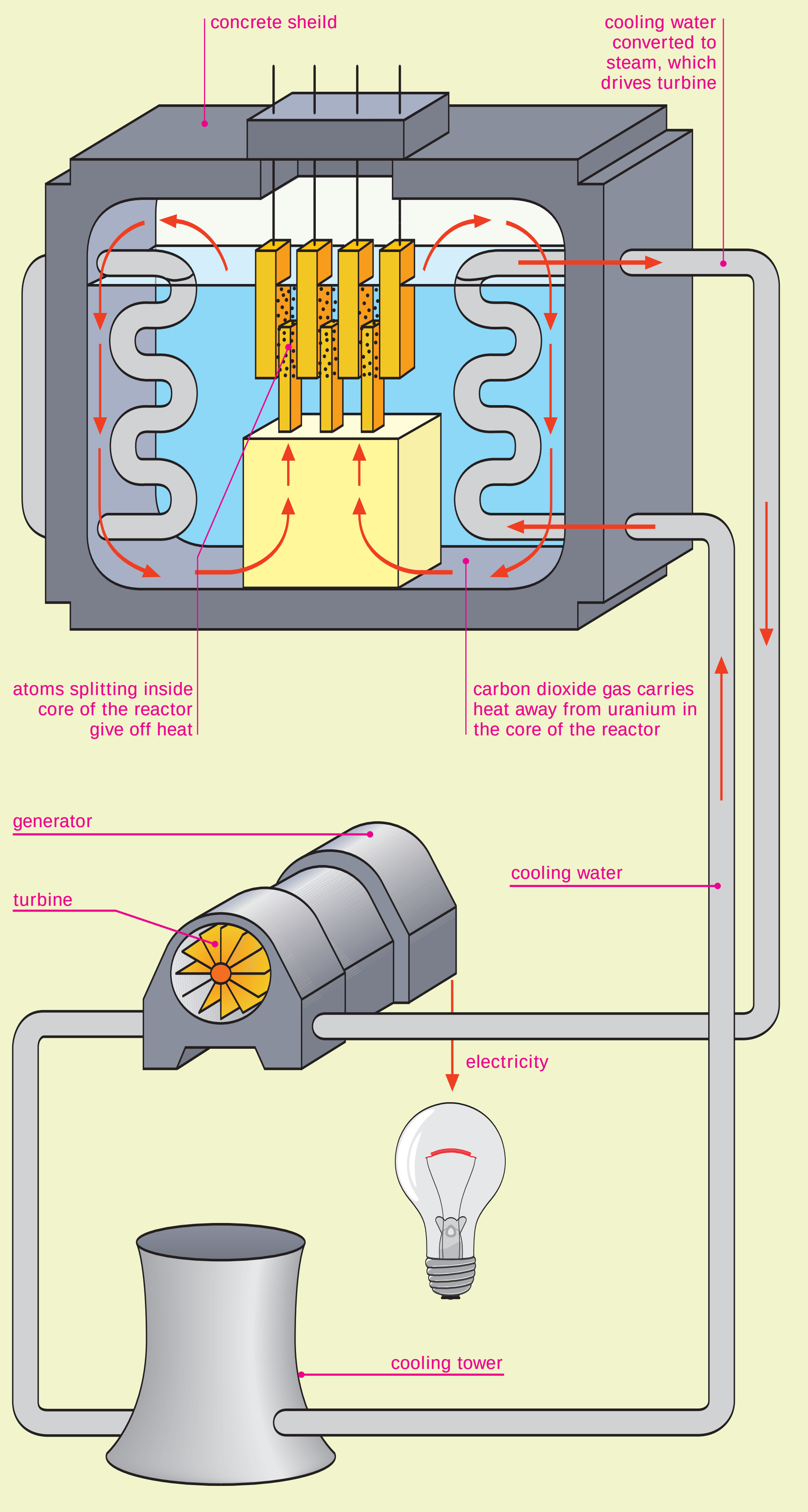
concrete sheild
cooling water converted to steam, which drives turbine
atoms splitting inside core of the reactor give off heat
carbon dioxide gas carries heat away from uranium in the core of the reactor
generator
turbine
cooling water
electricity
cooling tower

Key words

fission
isotope
uranium

Nuclear reactor

- *Uranium*, either the metal or the metal oxide, is used as fuel in nuclear reactors. The fuel is in the form of fuel rods, which are suspended in the reactor.
- Naturally occurring uranium contains 99.3 percent uranium-238 and only 0.7 percent of the radioactive *isotope* uranium-235. The uranium-235 content must be increased to approximately 3 percent before the uranium can be use as a fuel.
- Uranium-235 undergoes spontaneous *fission*. However, in a nuclear power station, the fission is brought about by bombarding the uranium nuclei with neutrons.
- The fission of one atom of uranium-235 absorbs one neutron and releases three others. In order to increase the chances that these neutrons will strike other uranium-235 atoms, they are slowed down by a moderator.
- Control rods are suspended between the fuel rods. These can be raised or lowered as needed to control the nuclear reaction. The control rods are made of alloys that absorb neutrons. When they are lowered, more neutrons are absorbed.
- The heat produced by the fission reaction is removed through a heat exchanger. The loop between the nuclear reactor and the heat exchanger is sealed so there is no danger of radioactive material escaping into the environment.
- The heat is used to convert water into pressurized steam. The high pressure steam drives a turbine connected to a generator, which produces electricity.

© Diagram Visual Information Ltd.

The uranium series

Key words

alpha particle
atomic number
beta particle
daughter nucleus
half-life
mass number
nuclide
radioactive decay
uranium
uranium series

Radioactive decay

- Radioactive nuclei break down by a process known as *radioactive decay* in order to become more stable. In a radioactive decay series, each member of the series is formed by the decay of the *nuclide* before it until a stable nuclide is produced. As the nuclei disintegrate, they emit *alpha* (α) or *beta* (β) *particles*.
- There are three naturally occurring radioactive decay series: the *uranium series*, the actinium series, and the thorium series. Each ends with a stable isotope of lead.

The uranium series

- The uranium series involves the radioactive decay of U-238 to stable Pb-206. It is also known as the 4n+2 series (where n is an integer), because each member of the series has a mass equivalent to 4n+2.
- The graph indicates how the decay occurs. *Atomic numbers* are plotted on the x-axis. The *mass numbers* are on the y-axis. The symbol for the element is at the top of the graph. Each diagonal line represents an alpha (α) decay; each horizontal line a beta (β) decay. A circle indicates the *daughter nucleus* (the nucleus produced by the decay of the previous nucleus). *Half-life* is indicated in years (a), days (d), hours (h), minutes (m), and seconds (s).

Decay chain

U-238 → Th-234 → Pa-234 → U-234 → Th-230 → Ra-226 → Rn-222 → Po-218 → At-218 → Pb-214 → Bi-214 → Po-214 → Ti-210 → Pb-210 → Bi-210 → Po-210 → Pb-206 (stable)

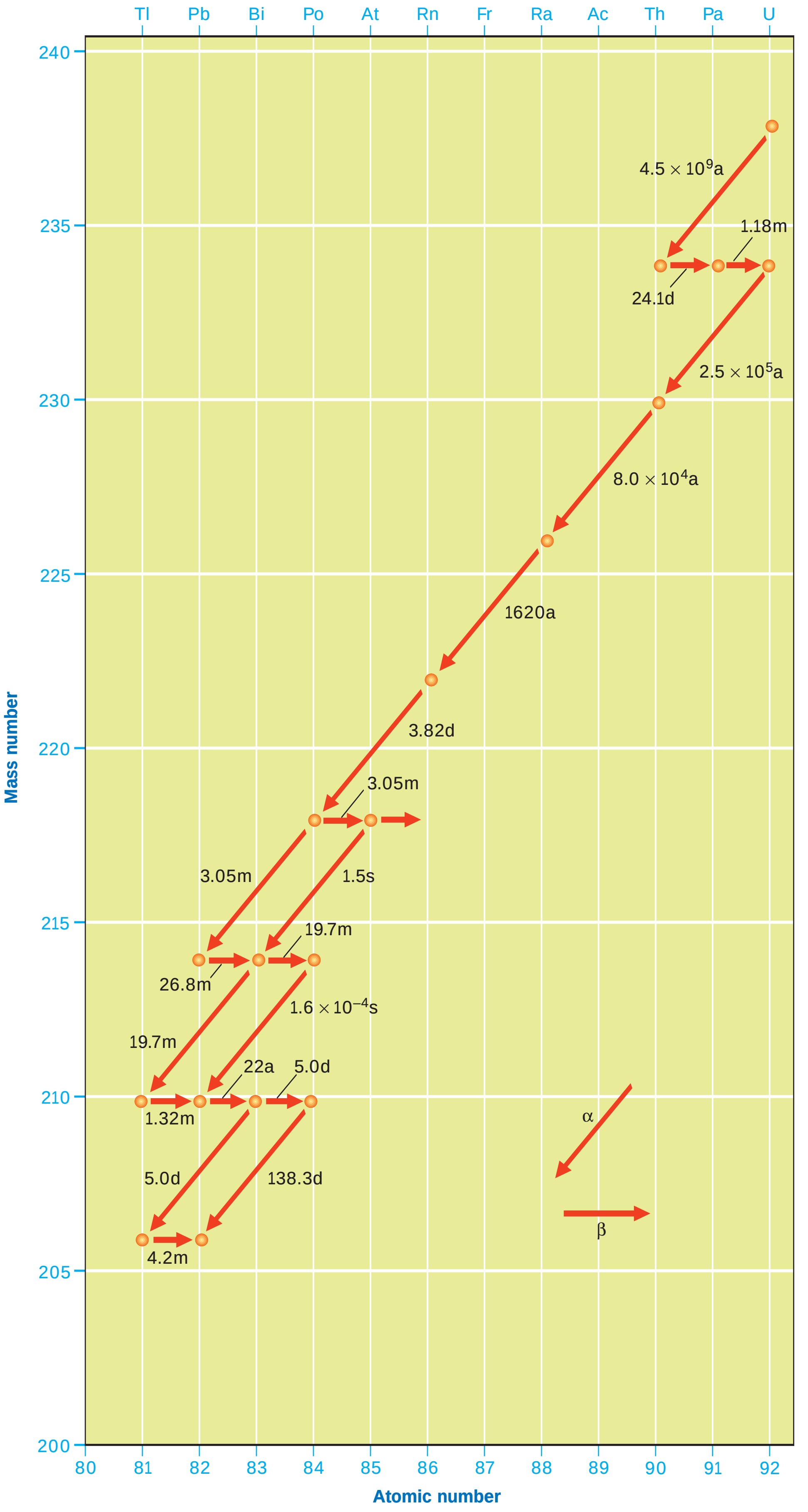

The actinium series

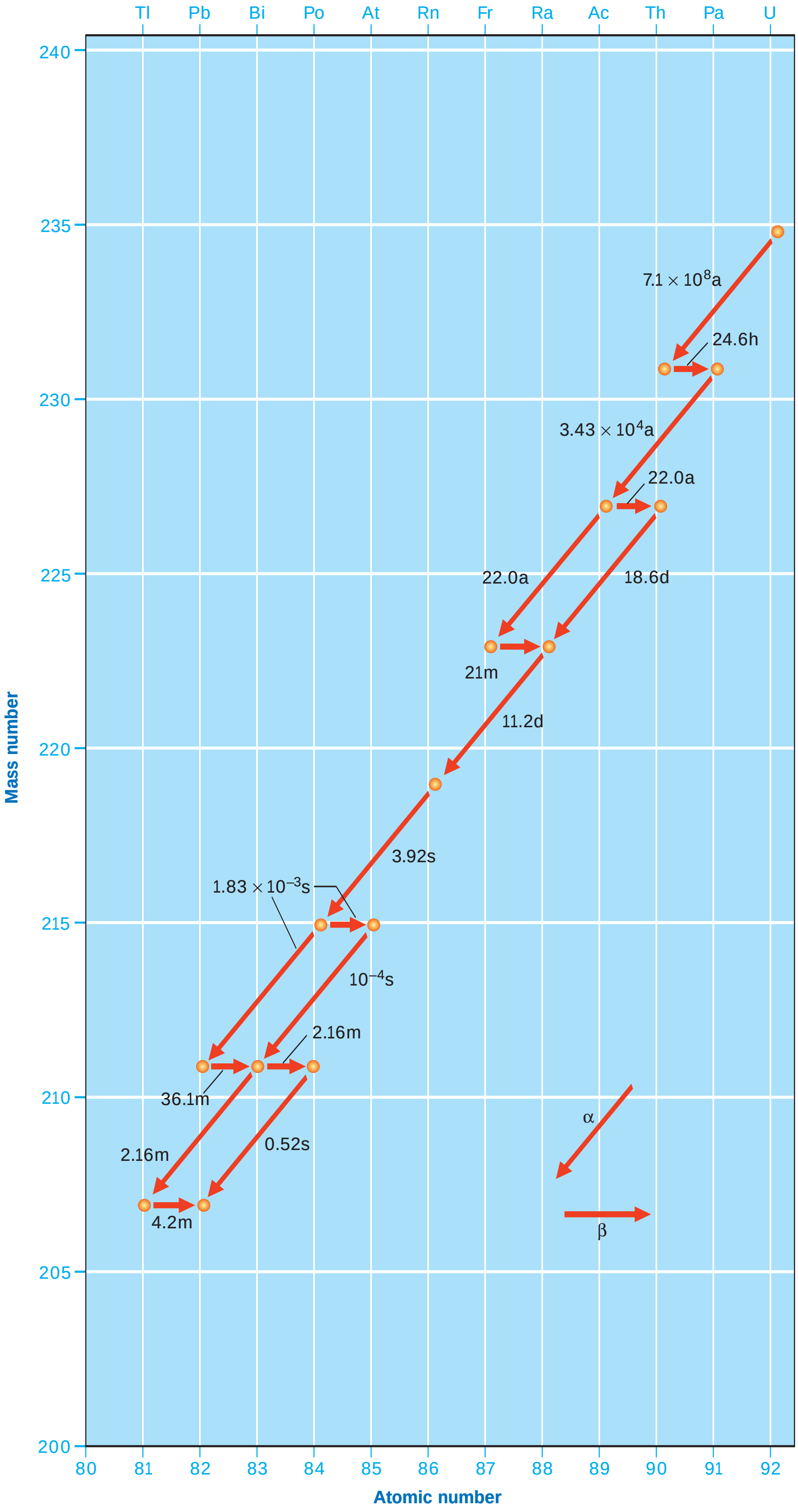

Key words

actinium
actinium series
alpha particle
atomic number
beta particle
daughter nucleus
half-life
mass number
nuclide
radioactive decay

Radioactive decay

- Radioactive nuclei break down by a process known as *radioactive decay* in order to become more stable. In a radioactive decay series, each member of the series is formed by the decay of the *nuclide* before it until a stable nuclide is produced. As the nuclei disintegrate, they emit *alpha* (α) or *beta* (β) *particles*.
- There are three naturally occurring radioactive decay series: the uranium series, the *actinium series*, and the thorium series. Each ends with a stable isotope of lead.

The actinium series

- The actinium series involves the radioactive decay of U-235 to stable Pb-207. It is also known as the 4n+3 series (where n is an integer), because each member of the series has a mass equivalent to 4n+3.
- The graph indicates how the decay occurs. *Atomic numbers* are plotted on the x-axis. The *mass numbers* are on the y-axis. The symbol for the element is at the top of the graph. Each diagonal line represents an alpha (α) decay; each horizontal line a beta (β) decay. A circle indicates the *daughter nucleus* (the nucleus produced by the decay of the previous nucleus). *Half-life* is indicated in years (a), days (d), hours (h), minutes (m), and seconds (s).

Decay chain

U-235 → Th-231 → Pa-231 → Ac-227 → Th-227 → Fr-223 → Ra-223 → Rn-219 → Po-215 → At-215 → Pb-211 → Bi-211 → Po-211 → Tl-207 → Pb-207 (stable)

© Diagram Visual Information Ltd.

The thorium series

Key words

alpha particle
atomic number
beta particle
daughter nucleus
half-life
mass number
nuclide
radioactive decay
thorium
thorium series

Radioactive decay

- Radioactive nuclei break down by a process known as *radioactive decay* in order to become more stable. In a radioactive decay series, each member of the series is formed by the decay of the *nuclide* before it until a stable nuclide is produced. As the nuclei disintegrate, they emit *alpha* (α) or *beta* (β) *particles.*
- There are three naturally occurring radioactive decay series: the uranium series, the actinium series, and the *thorium series*. Each ends with a stable isotope of lead.

The thorium series

- The thorium series involves the radioactive decay of Th-232 to stable Pb-208. It is also known as the (4n) series (where n is an integer) because each member of the series has a mass equivalent to 4n.
- The graph indicates how the decay occurs. *Atomic numbers* are plotted on the x-axis. The *mass numbers* are on the y-axis. The symbol for the element is at the top of the graph. Each diagonal line represents an alpha (α) decay; each horizontal line a beta (β) decay. A circle indicates the *daughter nucleus* (the nucleus produced by the decay of the previous nucleus). *Half-life* is indicated in years (a), days (d), hours (h), minutes (m), and seconds (s).

Decay chain

Th-232 → Ra-228 → Ac-228 → Th-228 → Ra-224 → Rn-220 → Po-216 → Pb-212 → Bi-212 → Po-212 → Tl-208 → Pb-208 (stable)

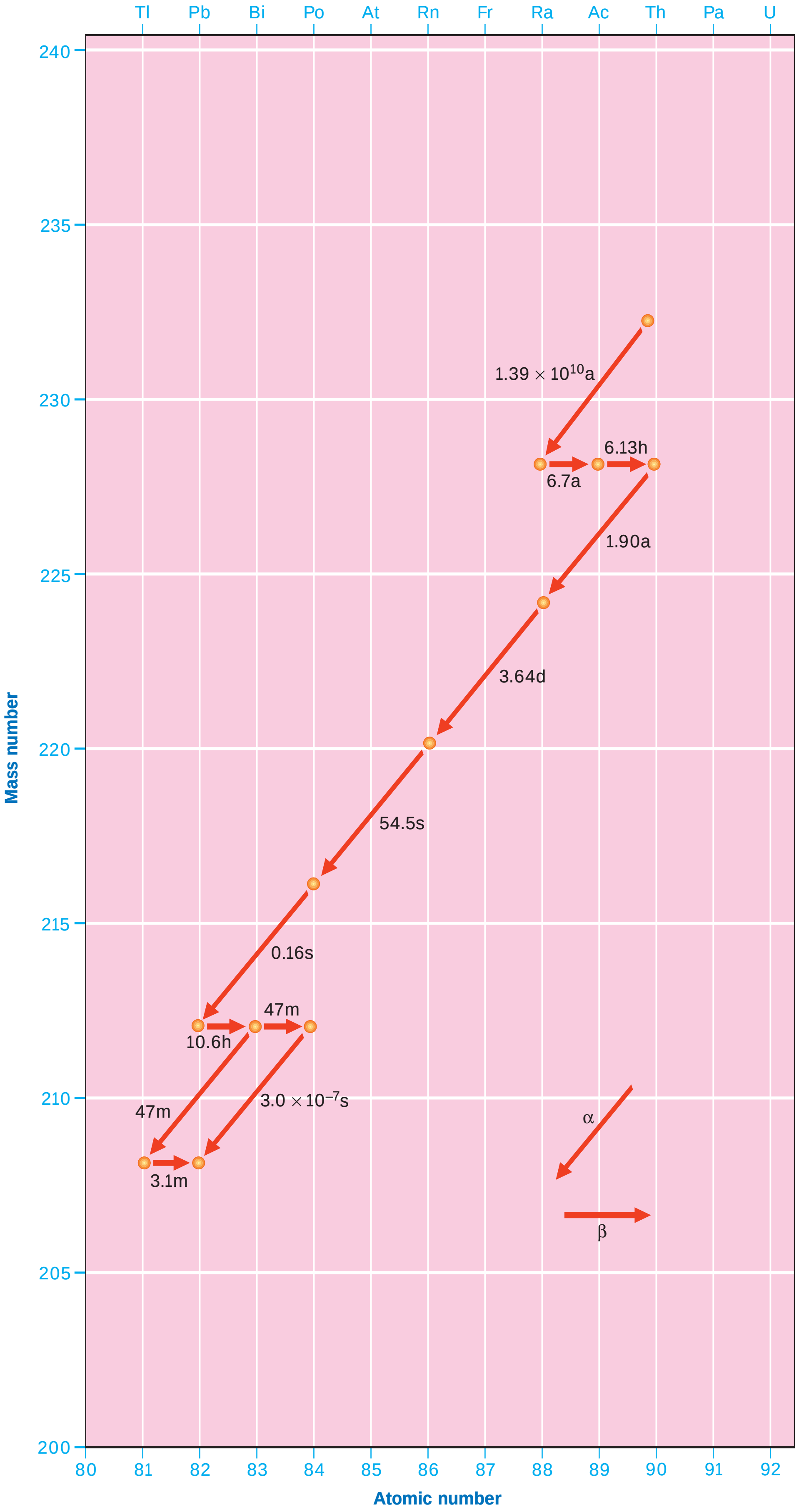

The neptunium series

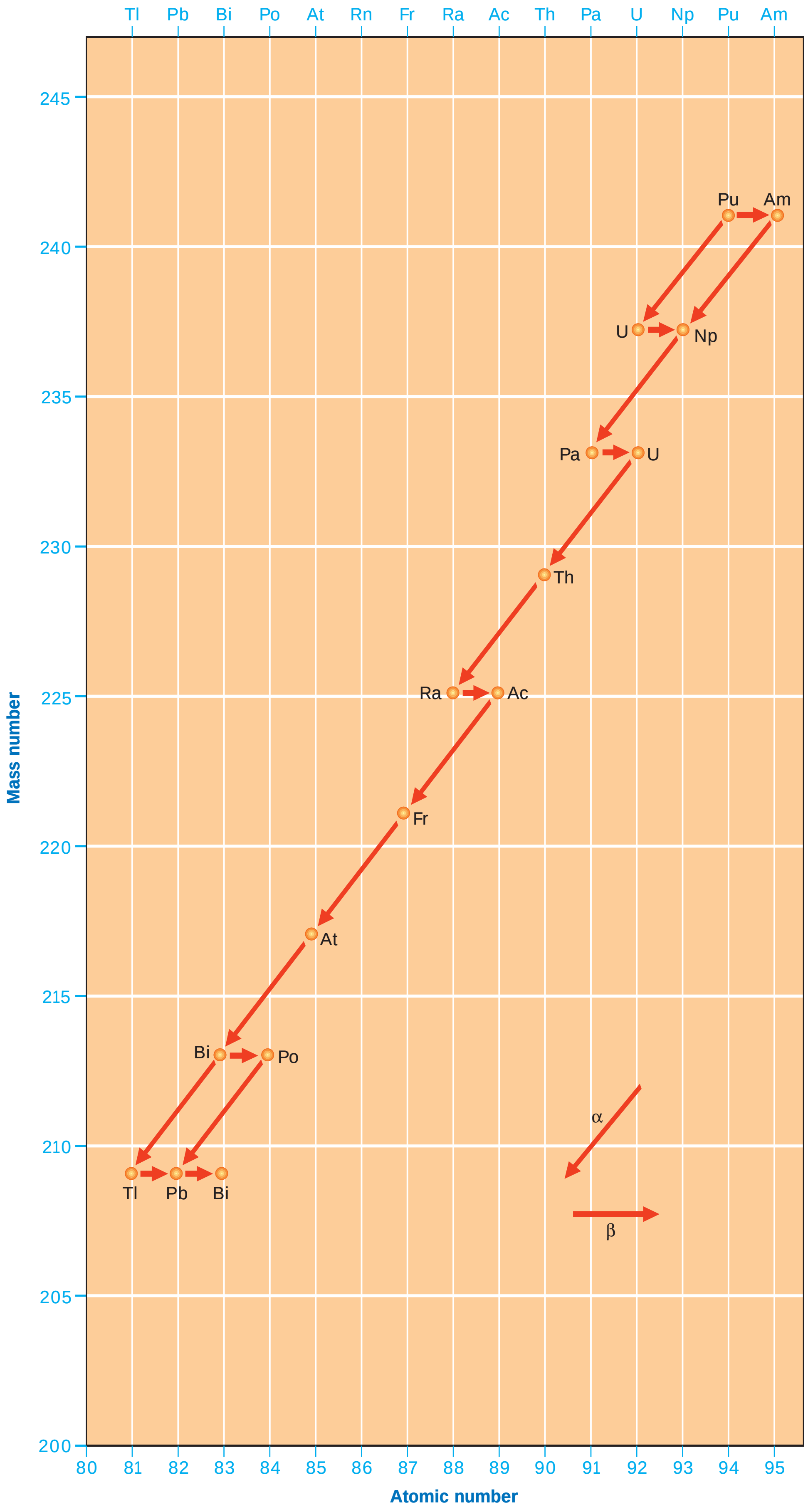

Key words

alpha particle
atomic number
beta particle
daughter nucleus
half-life
mass number
neptunium
neptunium series
nuclide
radioactive decay

Radioactive decay

- Radioactive nuclei break down by a process known as *radioactive decay* in order to become more stable. In a radioactive decay series, each member of the series is formed by the decay of the *nuclide* before it until a stable nuclide is produced. As the nuclei disintegrate, they emit *alpha* (α) or *beta* (β) *particles*.
- The *neptunium* series is composed of isotopes that do not occur in nature.

The neptunium series

- The neptunium series starts with the artificial isotope plutonium-241 and ends with bismuth-209. Each member of the series has a mass equivalent to 4n+1 (where n is an integer).
- The graph indicates how the decay occurs. *Atomic numbers* are plotted on the x-axis. The *mass numbers* are on the y-axis. The symbol of the element is at the top of the graph. Each diagonal line represents an alpha (α) decay; each horizontal line a beta (β) decay. A circle indicates the *daughter nucleus* (the nucleus produced by the decay of the previous nucleus).

Decay chain

Pu-241 → Am-241 → Np-237 → Pa-233 → U-233 → Th-229 → Ra-225 → Ac-225 → Fr-221 → At-217 → Bi-213 → Po-213 → Pb-209 → Bi-209 (stable)

© Diagram Visual Information Ltd.

Radioactivity of decay sequences

Key words

alpha decay
alpha particle
atomic number
beta decay
beta particle
gamma radiation
ground state
kinetic energy
mass number
nuclide

1 Alpha decay

- *Alpha decay* is the process in which the nucleus of an atom emits an *alpha particle* (which has the same structure as the helium-4 nucleus: ${}^{4}_{2}He$).
- The new atom's atomic *mass number* (A) is reduced by 4 and its *atomic number* (Z) is decreased by 2.
- Uranium-238 decays to thorium-234 by the loss of an alpha particle.
- Energy is also released as *gamma* (γ) *radiation*.

2 Alpha particle spectrum

- The *ground state* of the uranium nucleus (the natural state of the lowest energy of the nucleus) is at a higher energy than the ground state of the thorium nucleus.
- Some energy is released in the form of *kinetic energy*, which is carried by the alpha particle.
- The remaining energy is released as gamma radiation.

3 Beta decay

- *Beta decay* is the process in which the nucleus of an atom emits a *beta particle* (an electron).
- The new atom's atomic number (Z) is increased by 1, while the atomic mass number (A) remains unchanged.
- Thorium-234 decays to protactinium-234 by the loss of a beta particle. The half-life for this decay is 6.75 hours.

4 Beta particle spectrum

- The ground state of the thorium *nuclide* is at a higher energy than the ground state of the protractinium nucleus.
- Some energy is released in the form of kinetic energy, which is carried by the beta particle.
- The remaining energy is released as gamma radiation.

1 Alpha decay

$${}^{A}_{Z}X \xrightarrow{\alpha} {}^{A-4}_{Z-2}Y + {}^{4}_{2}He + \gamma$$

General sequence of alpha decay

$${}^{238}_{92}U \xrightarrow{\alpha} {}^{234}_{90}Th + {}^{4}_{2}He + \gamma$$

Example of alpha decay: uranium decay to thorium

2 Alpha particle spectrum

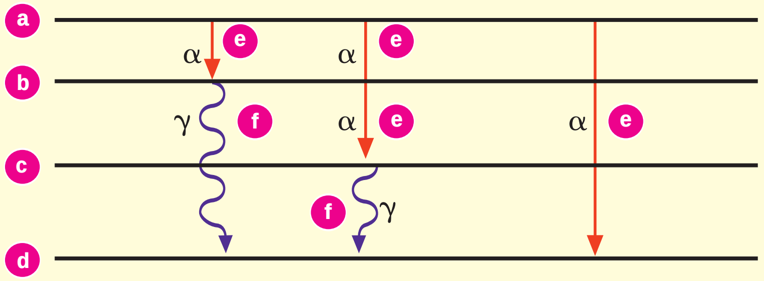

3 Beta decay

$${}^{A}_{Z}X \xrightarrow{\beta} {}^{A}_{Z+1}Y + e^{-} + \bar{\nu}$$

General sequence of beta decay

$${}^{234}_{90}Th \xrightarrow{\beta} {}^{A}_{91}Pa + e^{-} + \bar{\nu}$$

Example of beta decay: thorium decay to protactinium

4 Beta particle spectrum

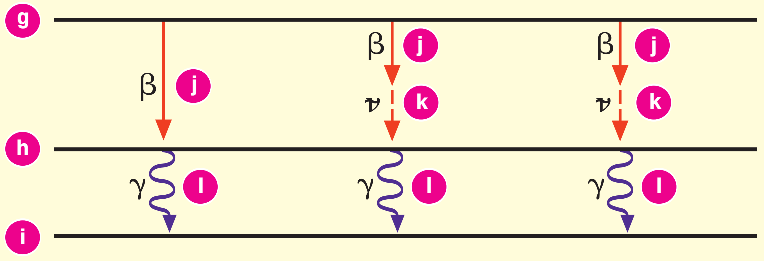

a nuclide *z*
b excited states of *Z* – 2
c excited states of *Z* – 2
d ground state of *Z* – 2
e alpha particle energy
f gamma radiation
g nuclide *Z*
h excited state of *Z* + 1
i ground state of *Z* + 1
j beta particle energy
k neutrino energy
l gamma radiation

Table of naturally occurring isotopes 1

Table of masses and abundance of naturally occurring isotopes

Atomic number (Z)	Element	Symbol	Mass number (A)	Percentage	Atomic mass
0	Neutron	n	1	—	1.008665
1	Hydrogen	H	1	99.99	1.007825
			2	0.01	2.014102
2	Helium	He	3	1.3×10^{-1}	3.016030
			4	100	4.002604
3	Lithium	Li	6	7.4	6.015126
			7	92.6	7.016005
4	Beryllium	Be	9	100	9.012186
5	Boron	B	10	19.6	10.012939
			11	80.4	11.009305
6	Carbon	C	12	98.9	12.000000
			13	1.1	13.003354
7	Nitrogen	N	14	99.6	14.003074
			15	0.4	15.000108
8	Oxygen	O	16	99.76	15.994915
			17	0.04	16.999133
			18	0.20	17.999160
9	Fluorine	F	19	100	18.998405
10	Neon	Ne	20	90.9	19.992440
			21	0.3	20.993849
			22	8.8	21.991384
11	Sodium	Na	23	100	22.989773
12	Magnesium	Mg	24	78.8	23.985045
			25	10.2	24.985840
			26	11.1	25.982591
13	Aluminum	Al	27	199	26.981535
14	Silicon	Si	28	92.2	27.976927
			29	4.7	28.976491
			30	3.1	29.973761
15	Phosphorus	P	31	100	30.973763
16	Sulfur	S	32	95.0	31.972074
			33	0.8	32.971460
			34	4.2	33.967864
			36	0.01	35.967091
17	Chlorine	Cl	35	75.5	34.968854
			37	24.5	36.965895
18	Argon	Ar	36	0.34	35.967548
			38	0.06	37.962724
			40*	99.6	30.962384
19	Potassium	K	39	93.1	38.963714
			40	0.012	39.964008
			41	6.9	49.961835
20	Calcium	Ca	40	97.0	39.962589
			42	0.6	41.958628
			43	0.1	42.958780
			44	2.1	43.955490
			46	0.003	45.953689
			48	0.2	47.952519
21	Scandium	Sc	45	100	44.955919

*denotes radioactive isotope

Key words

atomic mass
atomic number
isotope
mass number

Atomic number

- The *atomic number* (Z) of an element is the number of protons in the nucleus of one atom of that element. All atoms of the same element have the same atomic number.

Element

- "Element" refers to the common name of the element. This list is restricted to the 89 naturally occurring elements.

Symbol

- "Symbol" refers to the shorthand form of the element's name used in chemical equations.

Mass number

- The *mass number* (A) represents the number of protons or neutrons in the nucleus of one atom of that element. Not all atoms of the same element have the same mass number. Atoms of an element that have different mass numbers are called *isotopes*.

Percentage

- "Percentage" refers to isotopic abundance. For example, 99.99 percent of naturally-occurring hydrogen has the mass number 1. Only 0.01 percent has the mass number 2.

Atomic mass

- "*Atomic mass*" refers to the average atomic mass of that element's isotope weighted by isotopic abundance.

© Diagram Visual Information Ltd.

Table of naturally occurring isotopes 2

Key words

atomic mass
atomic number
isotope
mass number

Atomic number

- The *atomic number* (Z) of an element is the number of protons in the nucleus of one atom of that element. All atoms of the same element have the same atomic number.

Element

- "Element" refers to the common name of the element. This list is restricted to the 89 naturally occurring elements.

Symbol

- "Symbol" refers to the shorthand form of the element's name used in chemical equations.

Mass number

- The *mass number* (A) represents the number of protons or neutrons in the nucleus of one atom of that element. Not all atoms of the same element have the same mass number. Atoms of an element that have different mass numbers are called *isotopes*.

Percentage

- "Percentage" refers to isotopic abundance. For example, 99.99 percent of naturally-occurring hydrogen has the mass number 1. Only 0.01 percent has the mass number 2.

Atomic mass

- "*Atomic mass*" refers to the average atomic mass of that element's isotope weighted by isotopic abundance.

Table of masses and abundance of naturally occurring isotopes

Atomic number (Z)	Element	Symbol	Mass number (A)	Percentage	Atomic mass
22	Titanium	Ti	46	8.0	45.952633
			47	7.3	46.95176
			48	74.0	47.947948
			49	5.5	48.947867
			50*	5.2	49.944789
23	Vanadium	V	50	0.25	49.947165
			51	99.75	50.943978
24	Chromium	Cr	50	4.3	49.946051
			52	83.8	51.940514
			53	9.5	52.940651
			54	2.4	53.938879
25	Manganese	M	55	100	54.938054
26	Iron	Fe	54	5.8	53.93962
			56	91.7	55.93493
			57	2.2	56.93539
			58	0.3	57.93327
27	Cobalt	Co	59	100	58.933189
28	Nickel	Ni	58	67.8	57.93534
			60	26.2	59.93078
			61	1.2	60.93105
			62	3.7	61.92834
			64	1.1	63.92796
29	Copper	Cu	63	69.1	62.92959
			65	30.9	64.92779
30	Zinc	Zn	64	48.9	63.929145
			66	27.8	65.92605
			67	4.1	66.92715
			68	18.6	67.92486
			70	0.6	69.92535
31	Gallium	Ga	69	60.5	68.92568
			71	39.5	70.92484
32	Germanium	Ge	70	20.5	69.92428
			72	27.4	71.92174
			73	7.7	72.9234
			74	36.7	73.9211
			76	7.7	75.9214
33	Arsenic	As	75	100	74.92158
34	Selenium	Se	74	0.9	73.9224
			76	9.0	75.91923
			77	7.6	76.91993
			78	23.5	77.91735
			80	49.8	79.91651
			82	9.2	81.9167
35	Bromine	Br	79	50.6	78.91835
			81	49.4	80.91634
36	Krypton	Kr	78	0.3	77.920368
			80	2.3	79.91639
			82	11.6	81.91348
			83	11.5	82.91413
			84	56.9	83.911504
			86	17.4	85.91062

*denotes radioactive isotope

Table of naturally occurring isotopes 3

Table of masses and abundance of naturally occurring isotopes

Atomic number (Z)	Element	Symbol	Mass number (A)	Percentage	Atomic mass
37	Rubidium	Rb	85	72.1	84.9117
			87*	27.9	86.9092
38	Strontium	Sr	84	0.6	83.91338
			86	9.9	85.9093
			87	7.0	86.9089
			88	82.5	87.9056
39	Yttrium	Y	89*	100	88.9054
40	Zirconium	Zr	90	51.5	89.9043
			91	11.2	90.9052
			92	17.1	91.9046
			94	17.4	93.9061
			96	2.8	95.9082
41	Niobium	Nb	93	100	92.9060
42	Molybdenum	Mo	92	15.9	91.9063
			94	9.1	93.9047
			95	15.7	94.9057
			96	16.5	95.9045
			97	9.4	96.9057
			98	23.8	97.9055
			100	9.6	99.9076
43	Technetium	Tc	has no stable or naturally-occuring isotopes		
44	Ruthenium	Ru	96	5.6	95.9076
			98	1.9	97.905
			99	12.7	98.9061
			100	12.6	
			101	17.1	
			102	31.6	101.9037
			104	18.5	103.9055
45	Rhodium	Rh	103	100	102.9048
46	Palladium	Pd	102	1.0	101.9049
			104	11.0	103.9036
			105	22.2	104.9046
			106	27.3	105.9032
			108	26.7	107.9039
			110	11.8	109.9045
47	Silver	Ag	107	51.4	106.9050
			109	48.6	108.9047
48	Cadmium	Cd	106	1.2	105.9059
			108	0.9	107.9040
			110	12.4	109.9030
			111	12.7	110.9041
			112	24.1	111.9028
			113	12.3	112.9046
			114	28.8	113.9036
			116	7.6	115.9050
49	Indium	In	113	4.3	112.9043
			115*	95.7	114.9041

*denotes radioactive isotope

Key words

atomic mass
atomic number
isotope
mass number

Atomic number

- The *atomic number* (Z) of an element is the number of protons in the nucleus of one atom of that element. All atoms of the same element have the same atomic number.

Element

- “Element” refers to the common name of the element. This list is restricted to the 89 naturally occurring elements.

Symbol

- “Symbol” refers to the shorthand form of the element’s name used in chemical equations.

Mass number

- The *mass number* (A) represents the number of protons or neutrons in the nucleus of one atom of that element. Not all atoms of the same element have the same mass number. Atoms of an element that have different mass numbers are called *isotopes*.

Percentage

- “Percentage” refers to isotopic abundance. For example, 99.99 percent of naturally-occurring hydrogen has the mass number 1. Only 0.01 percent has the mass number 2.

Atomic mass

- “*Atomic mass*” refers to the average atomic mass of that element's isotope weighted by isotopic abundance.

© Diagram Visual Information Ltd.

Key words

atomic mass
atomic number
isotope
mass number

Atomic number

- The *atomic number* (Z) of an element is the number of protons in the nucleus of one atom of that element. All atoms of the same element have the same atomic number.

Element

- "Element" refers to the common name of the element. This list is restricted to the 89 naturally occurring elements.

Symbol

- "Symbol" refers to the shorthand form of the element's name used in chemical equations.

Mass number

- The *mass number* (A) represents the number of protons or neutrons in the nucleus of one atom of that element. Not all atoms of the same element have the same mass number. Atoms of an element that have different mass numbers are called *isotopes*.

Percentage

- "Percentage" refers to isotopic abundance. For example, 99.99 percent of naturally-occurring hydrogen has the mass number 1. Only 0.01 percent has the mass number 2.

Atomic mass

- "*Atomic mass*" refers to the average atomic mass of that element's isotope weighted by isotopic abundance.

Table of naturally occurring isotopes 4

Table of masses and abundance of naturally occurring isotopes

Atomic number (Z)	Element	Symbol	Mass number (A)	Percentage	Atomic mass
50	Tin	Sn	112	1.0	111.9049
			114	0.6	113.9030
			115	0.3	114.9035
			116	14.2	115.9021
			117	7.6	116.9031
			118	24.0	117.9018
			119	8.8	118.9034
			120	33.0	119.9021
			122	4.7	121.9034
			124	6.0	123.9052
51	Antimony	Sb	121	57.3	120.9037
			123	42.7	122.9041
52	Tellurium	Te	120	0.1	119.9045
			122	2.4	121.9030
			123	0.9	122.9042
			124	4.6	123.9028
			125	7.0	124.9044
			126	18.7	125.90324
			128	31.8	127.9047
			130	34.5	129.9067
53	Iodine	I	127	100	126.90435
54	Xenon	Xe	124	0.1	123.9061
			126	0.1	125.90417
			128	1.9	127.90354
			129	26.4	128.90478
			130	4.1	129.90351
			131	21.2	130.90509
			132	26.9	131.90416
			134	10.4	133.90540
			136	8.9	135.90722
55	Cesium	Ca	133	100	132.9051
56	Barium	Ba	130	0.1	129.90625
			132	0.2	131.9051
			134	2.6	133.9043
			135	6.7	134.9056
			136	8.1	135.9044
			137	11.9	136.9056
			138	70.4	137.9050
57	Lanthanum	La	138*	0.1	137.9068
			139	99.9	138.9061
58	Cerium	Ce	136	0.2	135.9071
			138	0.2	137.9057
			140	88.5	139.90528
			142*	11.1	141.9090
59	Praseodymium	Pr	141	100	140.90739

*denotes radioactive isotope

Table of naturally occurring isotopes 5

Table of masses and abundance of naturally occurring isotopes

Atomic number (Z)	Element	Symbol	Mass number (A)	Percentage	Atomic mass
60	Neodymium	Nd	142	27.3	141.90748
			143	12.3	142.90962
			144*	23.8	143.90990
			145	8.3	144.9122
			146	17.1	145.9127
			148	5.7	147.9165
			150	5.5	149.9207
61	Promethium	Pm	has no naturally occuring isotope		
62	Samarium	Sm	144	3.1	143.9116
			147*	15.1	146.91462
			148	11.3	146.9146
			149	14.0	148.9169
			150	7.5	149.9170
			152	26.6	151.9193
			154	22.4	153.9217
63	Europium	Eu	151	47.8	150.9196
			153	52.2	152.9207
64	Gadolinium	Gd	152	0.2	151.9194
			154	2.2	153.9202
			155	15.1	154.9220
			156	20.6	155.9222
			157	15.7	156.9240
			158	24.5	157.9242
			160	21.7	159.9273
65	Terbium	Tb	159	100	158.924
66	Dysprosium	Dy	156	0.1	
			158	0.1	
			160	2.3	159.924
			161	19.0	160.926
			162	25.5	161.926
			163	24.9	162.928
			164	28.1	163.928
67	Holmium	Ho	165	100	164.930
68	Erbium	Er	162	0.1	
			164	1.6	163.929
			166	33.4	165.929
			167	22.9	166.931
			168	27.1	167.931
			170	14.9	169.935
69	Thulium	Tm	169	100	
70	Ytterbium	Yb	168	0.1	
			170	3.1	
			171	14.4	
			172	21.9	171.929
			173	16.2	
			174	31.7	173.926
			176	12.6	
71	Lutetium	Lu	175	97.4	
			176*	2.6	175.9414

*denotes radioactive isotope

Key words

atomic mass
atomic number
isotope
mass number

Atomic number

- The *atomic number* (Z) of an element is the number of protons in the nucleus of one atom of that element. All atoms of the same element have the same atomic number.

Element

- "Element" refers to the common name of the element. This list is restricted to the 89 naturally occurring elements.

Symbol

- "Symbol" refers to the shorthand form of the element's name used in chemical equations.

Mass number

- The *mass number* (A) represents the number of protons or neutrons in the nucleus of one atom of that element. Not all atoms of the same element have the same mass number. Atoms of an element that have different mass numbers are called *isotopes*.

Percentage

- "Percentage" refers to isotopic abundance. For example, 99.99 percent of naturally-occurring hydrogen has the mass number 1. Only 0.01 percent has the mass number 2.

Atomic mass

- "*Atomic mass*" refers to the average atomic mass of that element's isotope weighted by isotopic abundance.

© Diagram Visual Information Ltd.

Table of naturally occurring isotopes 6

Key words

atomic mass
atomic number
isotope
mass number

Atomic number

- The *atomic number* (Z) of an element is the number of protons in the nucleus of one atom of that element. All atoms of the same element have the same atomic number.

Element

- “Element” refers to the common name of the element. This list is restricted to the 89 naturally occurring elements.

Symbol

- “Symbol” refers to the shorthand form of the element’s name used in chemical equations.

Mass number

- The *mass number* (A) represents the number of protons or neutrons in the nucleus of one atom of that element. Not all atoms of the same element have the same mass number. Atoms of an element that have different mass numbers are called *isotopes*.

Percentage

- “Percentage” refers to isotopic abundance. For example, 99.99 percent of naturally-occurring hydrogen has the mass number 1. Only 0.01 percent has the mass number 2.

Atomic mass

- “*Atomic mass*” refers to the average atomic mass of that element's isotope weighted by isotopic abundance.

Table of masses and abundance of naturally occurring isotopes

Atomic number (Z)	Element	Symbol	Mass number (A)	Percentage	Atomic mass
72	Hafnium	Hf	174	0.2	
			176	5.2	175.9403
			177	18.6	176.9419
			178	27.1	177.9425
			179	13.7	178.9444
			180	35.2	179.9451
73	Tantalum	Ta	180	0.01	179.9457
			181	99.99	180.9462
74	Tungsten	W	180	0.2	179.9450
			182	26.4	181.9465
			183	14.4	182.9485
			184	30.6	183.9491
			186	28.4	185.951
75	Rhenium	Re	185	37.1	184.950
			187*	62.9	186.9550
76	Osmium	Os	184	0.02	
			186	1.6	185.9529
			187	1.6	186.9550
			188	13.3	187.9550
			189	16.1	188.9572
			190*		189.9574
			192*		191.9605
77	Iridium	Ir	191	38.5	190.9599
			193	61.5	192.9623
78	Platinum	Pt	190		189.9592
			192	0.8	191.9605
			194	32.9	193.9624
			195		194.9645
			196		195.9646
			198	7.2	197.9675
79	Gold	Au	197	100	196.96655
80	Mercury	Hg	196	0.1	195.96582
			198	10.0	197.96677
			199	16.9	198.96826
			200	23.1	199.96834
			201	13.2	200.97031
			202	29.8	201.97063
			204	6.9	203.97348
81	Thallium	Tl	203	29.5	202.97233
			205	70.5	204.97446
			206*	—	205.97608
			207*	—	206.97745
			208*	—	207.98201
			210*	—	209.99000

*denotes radioactive isotope

Table of naturally occurring isotopes 7

Table of masses and abundance of naturally occurring isotopes

Atomic number (Z)	Element	Symbol	Mass number (A)	Percentage	Atomic mass
82	Lead	Pb	204	1.4	203.97307
			206	25.2	205.97446
			207	21.7	206.97590
			208	51.7	207.97664
			210*	—	209.98418
			211*	—	210.98880
			212*	—	211.99190
			214*	—	213.99976
83	Bismuth	Bi	209	100	208.98042
			210*	—	209.98411
			211*	—	210.98729
			212*	—	211.99127
			214*	—	213.99863
84	Polonium	Po	210*	—	209.98287
			211*	—	210.98665
			212*	—	211.98886
			214*	—	213.99519
			215*	—	214.99947
			216*	—	216.00192
			218*	—	218.0089
85	Astatine	At	215*	—	214.99866
			218*	—	218.00855
86	Emanation	Em	219*	—	219.00952
			220*	—	220.01140
			222*	—	222.0175
87	Francium	Fr	223*	—	223.01980
88	Radium	Ra	223*	—	223.01857
			224*	—	224.02022
			226*	—	226.0254
			228*	—	228.03123
89	Actinium	Ac	227*	—	227.02781
			228*	—	228.03117
			230*	—	230.0331
			231*	—	231.03635
			232*	100	232.03821
			234*	—	234.0436
91	Protactinium	Pa	231*	—	231.03594
			234*	—	234.0434
92	Uranium	U	234*	0.006	234.04090
			235*	0.718	235.04393
			238*	99.276	238.0508

*denotes radioactive isotope

Key words

atomic mass
atomic number
isotope
mass number

Atomic number

- The *atomic number* (Z) of an element is the number of protons in the nucleus of one atom of that element. All atoms of the same element have the same atomic number.

Element

- "Element" refers to the common name of the element. This list is restricted to the 89 naturally occurring elements.

Symbol

- "Symbol" refers to the shorthand form of the element's name used in chemical equations.

Mass number

- The *mass number* (A) represents the number of protons or neutrons in the nucleus of one atom of that element. Not all atoms of the same element have the same mass number. Atoms of an element that have different mass numbers are called *isotopes*.

Percentage

- "Percentage" refers to isotopic abundance. For example, 99.99 percent of naturally-occurring hydrogen has the mass number 1. Only 0.01 percent has the mass number 2.

Atomic mass

- "*Atomic mass*" refers to the average atomic mass of that element's isotope weighted by isotopic abundance.

© Diagram Visual Information Ltd.

Key words

accelerator A chemical that increases the rate of a chemical reaction.

acid Any substance that releases hydrogen ions when added to water. It has a pH of less than 7.

acid-base indicator A chemical compound that changes color when going from acidic to basic solutions. An example is Methyl orange.

acidity The level of hydrogen ion concentration in a solution.

actinides The name of the radioactive group of elements with atomic numbers from 89 (actinium) to 103 (lawrencium).

actinium (Ac) A silvery radioactive metallic element that occurs naturally in pitchblende and can be synthesized by bombarding radium with neutrons.

actinium series One of the naturally occurring radioactive series.

activation energy The energy barrier to be overcome in order for a reaction to occur.

active site The part of an enzyme where the chemical reaction occurs.

addition polymerization A chemical reaction in which simple molecules are added to each other to form long-chain molecules without by-products.

addition reaction A reaction in which a molecule of a substance reacts with another molecule to form a single compound.

adsorption The process by which molecules of gases or liquids become attached to the surface of another substance.

aerosol Extremely small liquid or solid particles suspended in air or another gas.

alcohol A member of a family of organic compounds whose structure contains the –OH functional group.

aldehyde One of a group of organic compounds containing the aldehyde group (–CHO). Names have the suffix -al.

aldohexose A monosaccharide having six carbon atoms and an aldehyde group.

aldose A sugar containing one aldehyde group per molecule

alkali A solution of a substance in water that has a pH of more than 7 and has an excess of hydroxide ions in the solution.

alkali metals Metallic elements found in group 1 of the periodic table. They are very reactive, electropositive, and react with water to form alkaline solutions.

alkaline earth metals Metallic elements found in group 2 of the periodic table.

alkalinity Having a pH greater than 7.

alkane A member of the hydrocarbon group whose general formula is C_nH_{2n+2}. They have single bonds between the carbon atoms and are not very reactive.

alkanol See alcohol

alkene A member of the hydrocarbon group whose general formula is C_nH_{2n}. They have a double bond between a pair of carbon atoms and are thus reactive.

alkyl A member of the hydrocarbon group whose general formula is CnH_{2n+1}.

alkyne A member of the hydrocarbon group with the general formula CnH_{2n-1}. They have a triple bond between a pair of carbon atoms in each molecule and are thus reactive.

allotrope An element that can exist in more than one physical form while in the same state.

alloy A metallic material made of two or more metals or of a metal and non-metal.

alpha decay The process of radioactive decay in which the nucleus of an atom emits an alpha particle.

alpha particle A particle released during radioactive decay that consists of two neutrons and two protons.

aluminum (Al) A silvery-white. metallic element that is non-magnetic and oxidizes easily.

amine A member of a group of organic compounds containing the amino functional group $–NH_2$.

amino acid An organic compound containing both the carboxyl group (–COOH) and the amino group ($–NH_2$).

ammonia (NH_3) A colorless, strong-smelling poisonous gas that is very soluble in water.

ammonium hydroxide (NH_4OH) An aqueous solution of ammonia. It is a corrosive chemical with a strong odor.

ammonium ion (NH_4^+) An ion found in ammonia solution and in ammonium compounds.

amphoteric Exhibiting properties of both an acid and a base.

anhydride The substance remaining when one or more molecules of water have been removed from an acid or a base.

anhydrous Containing no water. Term applied to salts without water of crystallization.

anion An ion having a negative charge.

anode The electrode carrying the positive charge in a solution undergoing electrolysis.

anomer A stereoisometric form of a sugar, involving different arrangements of atoms or molecules around a central atom,

aqueous solution A solution in which water is the solvent.

argon Ar. A colorless, odorless. gaseous element. One of the noble gases.

aryl A member of an aromatic hydrocarbon group formed by the removal of a hydrogen atom from an aromatic hydrocarbon.

association The process by which molecules of a substance combine to form a larger structure.

astatine At. A non-metallic radioactive element that is highly unstable and rare in nature.

atmosphere The layer of gases surrounding Earth.

atom The smallest particle of an element that can exhibit that element’s properties.

atomic emission spectrum The amount of electromagnetic radiation an element emits when excited.

atomic mass The ratio of the mass of an average atom of an element to 1/12th of the mass of an atom of the carbon-12 isotope.

atomic number The number of protons in the nucleus of an atom.

atomic volume The volume of one mole of the atoms of an element.

Avogadro’s constant The number of particles present in a mole of substance.

© Diagram Visual Information Ltd.

azeotropic mixture A mixture of liquids that boils without a change in composition.

bakelite A phenol/methanal resin that has good electrical and heat insulation properties.

base A substance existing as molecules or ions that can take up hydrogen ions.

beta decay The process of radioactive decay in which the nucleus of an atom emits a beta particle.

beta particle A high-speed electron emitted by the nucleus of certain radioactive elements during beta decay.

Big Bang The primeval explosion that most astronomers think gave rise to the Universe.

black hole An object with infinite density.

body-centered cubic packing A crystalline structure in which one atom sits in the center of each cube.

boiling point The point at which a substance changes from the liquid state to the gas state.

bond The chemical connection between atoms within a molecule. Bonds are forces and are caused by electrons.

bond angle In a molecule, the angle between the two straight lines joining the centers of the atoms concerned.

bromine (Br) A non-metallic element that is isolated as a dark red liquid. It is a very reactive oxidizing agent.

brown dwarf A ball of gas like a star but whose mass is too small to have nuclear fusion occur at its core.

Brownian motion The random movement of particles through a liquid or gas.

buckminsterfullerene See buckyball.

buckyball The nickname for buckminsterfullerene. An allotropic form of carbon. It has a cage-like structure and has the formula C_{50}, C_{60}, or C_{70}.

burette A long, graduated glass tube with a tap at the lower end. It is used to measure a volume of liquid accurately.

calcium (Ca) A soft, slivery-white metal.

calcium carbonate A white solid, occurring naturally in marble and limestone, that dissolves in dilute acids.

carbide A compound that contains carbon and an element with lower electronegativity.

carbon (C) A non-metallic element whose compounds occur widely in nature.

carbonate A salt of carbonic acid (containing the ion CO_3^{2-}).

carbon cycle The circulation of carbon through the biosphere.

carbon dioxide (CO_2) A dense, colorless, odorless gas that does not support combustion. It exists in the atmosphere and is instrumental in the carbon cycle.

carbonic acid (H_2CO_3) A very weak acid formed by dissolving carbon dioxide in water.

carbon monoxide(CO) A colorless, odorless, very poisonous gas. It is sparingly soluble in water.

carboxyl group The organic radical –CO.OH.

carboxylic acid An organic acid that contains one or more carboxyl groups.

catalyst A substance that alters the rate of a chemical reaction but remains chemically unchanged by it.

catalytic cracking The process used in the petroleum industry to convert large-chain hydrocarbon molecules to smaller ones.

catenation The formation of chains of bonded atoms.

cathode The electrode carrying the negative charge in a solution undergoing electrolysis.

cathode rays A stream of electrons emitted from the cathode in a vacuum tube.

cation An ion having a positive charge.

cellulose A complex carbohydrate that is the main component of the cell walls of plants.

centrifuge A machine that rotates an object at high speed.

chain reaction A self-sustaining nuclear reaction yielding energy and electrons emitted by the fission of an atomic nucleus, which proceeds to cause further fissions.

chemical compound A substance composed of two or more elements linked by chemical bonds that may be ionic or covalent.

chemical energy The energy stored in the bonds between atoms and molecules that is released during a chemical reaction.

chemical reaction The process in which one or more substances reacts to form new substances.

chiral An object or a system that differs from its mirror image.

chloride A compound containing chlorine and another element.

chlorine (Cl) A poisonous, greenish, gaseous element that is a powerful oxidizing agent.

chlorophyll A green pigment found in most plants. It absorbs light energy during photosynthesis.

chromatography A technique for separating and identifying mixtures of solutes in a solution.

chromium (Cr) A hard, brittle, gray-white metallic element that is very resistant to corrosion and takes a high polish.

cobalt (Co) A hard, lustrous, silvery-white metallic element found in ores,

colloid A substance made of very small particles whose size (1–100 nm) is between those of a suspension and those in solution.

compound See chemical compound

concentration A measure of the quantity of solute dissolved in a solution at a given temperature.

conductor A material that is able to conduct heat and electricity.

convection current A circular current in a fluid such as air.

coordinate bonding A type of covalent bond in which one of the atoms supplies both electrons.

coordination number The number of atoms, ions, or molecules to which bonds can be formed.

copper (Cu) A pinkish metallic element used widely in alloys and electrical wires.

covalent bond A bond formed when two electrons are shared between two atoms (usually between two non-metallic atoms), one contributed by each atom.

covalent compound A compound in which the atoms in the molecules are held together by covalent bonds.

crust The outer layer of Earth.

cryolite A compound of aluminum fluoride and sodium fluoride.

crystal A substance with an orderly arrangement of atoms, ions, or molecules in a regular geometrical shape.

daughter nucleus In radio active decay, the nucleus produced by the decay of the previous nucleus.

© Diagram Visual Information Ltd.

dehydrating agent A substance that has an attraction for water and is therefore used as a drying agent.

dehydrogenation The chemical process of removal of hydrogen atoms from a molecule (a form of oxidation), increasing its degree of unsaturation.

density The mass per unit volume of a given substance.

detergent The term for a synthetic soap substitute.

diamond A transparent crystalline allotrope of carbon. It is the hardest naturally occurring substance.

diatomic molecule A molecule that consists of two atoms.

diffusion The process of rapid random movement of the particles of a liquid or gas that eventually form a uniform mixture.

dipole A chemical compound with an unequally distributed electric charge.

disaccharide A sugar molecule formed by a condensation reaction between two monosaccharide molecules.

displacement reaction A reaction in which a more reactive substance displaces the ions of a less reactive substance.

dissociation The breaking down of a molecule into smaller molecules, atoms, or ions.

dissolve To add a solute to a solvent to form a uniform solution.

distillation A process in which a solution is boiled and its vapor then condensed.

double bond A covalent bond formed between two atoms in which two pairs of electrons contribute to the bond..

dry gas A gas from which all water has been removed

ductile Capable of being drawn out, shaped, or bent.

effective collision A collision that brings about a reaction.

electric field A field of force around a charged particle.

electrode A conductor that allows current to flow through an electrolyte, gas, vacuum, or semiconductor.

electrolysis The process by which an electrolyte is decomposed when a direct current is passed through it between electrodes.

electrolyte A substance that forms ions when molten or dissolved in a solvent and that carries an electric current during electrolysis.

electron One of the three basic subatomic particles. Very light and carrying a negative charge, it orbits around the nucleus of an atom.

element A substance that cannot be split into simpler substances using chemical methods.

emulsion A colloidal dispersion of small droplets of one liquid dispersed within another, such as oil in water or water in oil.

enantiomer One of two "mirror images" of a chiral molecule.

end point The point at which a reaction is complete.

endothermic a chemical change during which heat is absorbed.

enthalpy A measure of the stored heat energy of a substance.

enzyme An organic catalyst, made of proteins, that increases the rate of a specific biochemical reaction.

equilibrium The state of a reversible chemical reaction where the forward and backward reactions take place at the same rate.

equivalence point The point at which there are equivalent amounts of acid and alkali.

ester A member of a hydrocarbon group that is formed by a reaction between a carboxylic acid and an alcohol.

ethane (C_2H_6) A colorless, flammable alkane that occurs in natural gas.

ethanol (C_2H_5OH) A volatile, colorless liquid alcohol used in beverages and as a gasoline octane enhancer.

ethene (C_2H_4) A colorless, flammable unsaturated gas, manufactured by cracking petroleum gas, used in ethanol and polyethene production.

evaporation The change in state from liquid to vapor.

exothermic A chemical change resulting in the liberation of heat.

face-centered cubic close packing A crystal structure in which one atom sits in each "face" of the cube.

Faraday constant The amount of electricity needed to liberate one mole of a monovalent ion during electrolysis (9.648 670 x 10^{-4} C mol^{-1}).

fatty acid A hydrocarbon chain with a carboxyl group at one end.

filtrate A clear liquid that has passed through a filter.

filtration The process of removing particulate matter from a liquid by passing the liquid through a porous substance.

fission A process during which a heavy atomic nucleus disintegrates into two lighter atoms and the lost mass is converted to energy.

fluorescence The emission of light from an object that has been irradiated by light or other radiations.

fluorine (F) A gaseous non-metallic element that is poisonous and very reactive gas.

flux A substance that combines with another substance (usually an oxide), forming a compound with a lower melting point than the oxide.

foam A dispersion of gas in a liquid or solid. Small bubbles of gas are separated by thin films of the liquid or solid.

formula mass The relative molecular mass of a compound calculated using its molecular formula. The mass of a mole of the substance.

forward reaction A reaction in which reactants are converted to products.

fractional distillation The separation of a mixture or liquids that have differing but similar boiling points.

fullerenes Allotropes of carbon in the form of a hollow sphere (buckyball) or tube (nanotube).

functional group The atom (or group of atoms) present in a molecule that determines the characteristic properties of that molecule.

fusion The process by which two or more light atomic nuclei join, forming a single heavier nucleus. The products of fusion are lighter than the components. The mass lost is liberated as energy.

galvanizing The coating of iron or steel plates with a layer of zinc to protect against rusting.

gamma radiation Very short-wave electromagnetic radiation emitted as a result of radioactive decay.

gas One of the states of matter. In a gas, the particles can move freely throughout the space in which it is contained. Gas is the least dense of the states of matter.

© Diagram Visual Information Ltd.

gas-liquid chromatography A type of chromatography in which the mobile phase is a carrier gas and the stationary phase is a microscopic layer of liquid on an inert solid support.

gel A colloidal solution that has formed a jelly. The solid particles are arranged as a fine network in the liquid phase.

geometric isomerism A form of isomerism that describes the orientation of functional groups at the ends of a bond where no rotation is possible.

glucose In animals and plants, the most widely distributed hexose sugar and the most common energy source in respiration.

glycogen A polysaccharide composed of branched chains of glucose, used to store energy in animals and some fungi..

gold (Au) A shiny, yellow metallic element used in coins, jewelry, and electrical contacts.

grade The concentration of ore in rock.

Graham's law The velocity with which a gas will diffuse is inversely proportional to the square root of its density.

graphite A soft, grayish-black, solid allotrope of carbon.

ground state The lowest allowed energy state of an atom, molecule, or ion.

group The vertical columns of elements in the periodic table. Elements in a group react in a similar way and have similar physical properties.

group 1 elements The alkali metals. The elements lithium, sodium, potassium, rubidium, cesium, and francium. These elements have one electron in their outer shell.

group 2 elements The alkaline earth metals. The elements beryllium, magnesium, calcium, strontium, barium, and radium. These elements have two electrons in their outer shell.

group 3 elements The elements boron, aluminum, gallium, indium, and thallium. These elements have a full s orbital and one electron in a p orbital in their outer shell.

group 4 elements The elements carbon, silicon, germanium, tin, and lead. These elements have a full s orbital and two electrons in two p orbitals in their outer shell.

group 5 elements The elements nitrogen, phosphorus, arsenic, antimony, and bismuth. These elements have a full s orbital and three electrons in three p orbitals in their outer shell.

group 6 elements The chalcogens. The elements oxygen, sulfur, selenium, tellurium, and polonium. These elements have a full s orbital, one full p orbital, and two half-full p orbitals in their outer shell.

group 7 elements The halogens. The elements fluorine, chlorine, bromine, iodine, and astatine. These elements have a full s orbital, two full p orbitals, and one half-full p orbital in their outer shell.

group 8 elements. The noble or inert gases. The elements helium, neon, argon, krypton, xenon, and radon. The outer shell of the atoms in these elements is complete, rendering these elements unreactive.

half-life The time required for half the atoms of a radioactive substance to disintegrate.

halide A compound that a halogen makes with another element. Metal halides are ionic; non-metal halides are formed by covalent bonding.

halogens See Group 7 elements.

helium (He) A colorless, odorless gaseous element that is the second most abundant element on Earth.

hexagonal close packing In crystalline structures, a way of packing atoms so that alternating layers overlie one another in an ABABAB pattern.

hexose A monosaccharide with six carbon atoms.

homologous series A series of related organic compounds. The formula of each member differs from the preceding member by the addition of a $-CH_2-$ group.

hydration The combination of water and another substance to produce a single product.

hydride A compound formed between hydrogen and another element.

hydrocarbon An organic molecule consisting only of carbon and hydrogen.

hydrochloric acid (HCl) A colorless fuming solution of hydrogen chloride.

hydrogen (H) An odorless, easily flammable gaseous element that is the most abundant on Earth.

hydrogen bond A weak bond between hydrogen and another element with partial but opposite electrical charges.

hydrogen chloride (HCl) A colorless gas with a pungent smell that fumes in moist air. It is very soluble in water.

hydrogen peroxide (H_2O_2) A colorless or pale blue viscous liquid. It is a strong oxidizing agent, but it can also act as a reducing agent.

hydrogen sulfide (H_2S) A colorless, poisonous gas smelling of bad eggs that is moderately soluble in water. It is a reducing agent.

hydronium ion The positive ion $(H_3O)^+$. It is the hydrated form of the hydrogen ion (H^+) or proton.

hydrophilic Water-loving. In solution, it refers to a chemical or part of a chemical that is highly attracted to water.

hydrophobic Water-hating. It refers to a chemical or part of a chemical that repels water.

hydroxide A compound containing the hydroxide ion or the hydroxyl group bonded to a metal atom.

hydroxide ion The negative ion (OH^-) present in alkalis.

immiscible Incapable of mixing.

indicator A substance that indicates by a change in its color the degree of acidity or alkalinity of a solution or the presence of a given substance.

inert A substance that is either very or completely unreactive.

inert gases See noble gases.

infrared Electromagnetic radiation with a greater wavelength than the red end of the visible spectrum.

insoluble A substance that does not dissolve in a particular solvent under certain conditions of temperature and pressure.

iodine (I) A grayish-black non-metallic element that is essential in the diet and is used in disinfectants and photography.

ion An electrically charged atom or group of atoms.

ionic bonding A type of bonding that occurs when atoms form ions and electrons are transferred from one atom to another.

ionic compound Compounds consisting of ions held together by strong ionic bonds. Ionic compounds are electrolytes.

© Diagram Visual Information Ltd.

ionic crystal A type of crystal where ions of two of more elements form a regular three-dimensional arrangement (crystal structure).

ionization energy The energy needed to remove completely an electron from a neutral gaseous atom or ion against the attraction of the nucleus.

ionizing radiation Any radiation capable of displacing electrons from atoms or molecules and so producing ions

iron (Fe) A silvery, malleable and ductile metallic element used in construction.

irradiation The use of radiation to destroy microorganisms in foods.

isomer One of two or more (usually organic) compounds having the same molecular formula and relative molecular mass but different three-dimensional structures.

isomerism The rearrangement atoms in a molecule to make it more efficient.

isomerization The transformation of a molecule into a different isomer.

isotope Atoms of the same element (all chemically identical) having the same atomic number but containing different numbers of neutrons, giving a different mass number.

ketone An organic compound that contain two organic radicals connected to a carbonyl group.

kinetic energy The energy a body has by virtue of its motion.

lanthanide series A series of metallic elements with the atomic numbers 57 to 71. The metals are shiny and are attacked by water and acids.

lattice The orderly three-dimensional arrangements of atoms, molecules, or ions seen in crystals.

lead (Pb) A silvery-white metallic element used in batteries and in water, noise, and radiation shielding.

lead sulfide (PbS) A brownish-black insoluble crystal. It occurs naturally as the mineral galena.

Le Chatelier's principle If a chemical reaction is at equilibrium and a change is made to any of the conditions, further reaction will take place to counteract the changes in order to re-establish equilibrium.

limewater A solution of calcium hydroxide that is used to test for the presence of carbon dioxide.

limiting form The possibilities for the distribution of electrons in a molecule or ion.

liquid A state of matter between solid and gas. Particles are loosely bonded, so can move relatively freely.

lone pair A pair of electrons in the outermost shell of an atom that are not involved in the formation of covalent bonds.

luminescence Light emission from a substance caused by an effect other than heat.

magnesium (Mg) A silvery-white metallic element used in alloys and castings.

magnesium oxide (MgO) A white solid used for reflective coatings and as a component of semiconductors.

manganese (Mn) A soft, gray metallic element used in making steel alloys.

mantle The layer of Earth between the crust and the core.

mass The measure of a body's resistance to acceleration.

mass number The total number of protons and neutrons in the nucleus of an atom.

mass spectrometry A technique for determining the composition of molecules by using the mass of their basic constituents

melting point The point at which a substance changes state from solid to liquid.

methane (CH_4) The simplest alkane. A colorless, tasteless, odorless flammable gas used as a fuel.

mineral A natural inorganic substance with distinct chemical composition and internal structure.

mixture A system consisting of two or more substances that are not chemically combined.

mobile phase The phase that moves along the stationary phase. It is the solvent in paper chromatography.

molarity The concentration of solution giving the number of moles of solute dissolved in 1 kg of solvent.

mole The amount of a substance that contains the same number of entities (atoms, molecules, ions, etc.) as there are atoms in 12 g of the carbon-12 isotope.

molecular mass The sum of the atomic masses of all atoms in a molecule. The mass of a mole of the substance.

molecule The smallest part of an element or chemical compound that can exist independently with all the properties of the element or compound.

monomer A basic unit from which a polymer is made.

monosaccharide A simple sugar such as glucose.

nanotube An isotope of carbon consisting of long thin cylinders closed at either end with caps containing pentagonal rings.

neptunium (Np) A radioactive metallic element that can be synthesized by bombarding U-238 with neutrons.

neptunium series A radioactive series composed of artificial isotopes.

neutral A solution whose pH is 7.

neutralization The reaction of an acid and a base forming a salt and water..

neutron One of the two major components of the atomic nucleus. It has no electric charge.

neutron star The smallest but densest kind of star, apparently resulting from a supernova explosion.

nickel (Ni) A hard, malleable and ductile, silvery-white metallic element that is a component of Earth's core.

nitrate A salt of nitric acid.

nitric acid (HNO_3) A colorless, corrosive, poisonous, fuming liquid that is a strong oxidizing agent.

nitrite A salt of nitrous acid.

nitrogen (N) A colorless gaseous element essential for the growth of plants and animals.

nitrogen cycle The process by which nitrogen is recycled in the ecosystem.

noble gases Group 8 elements: helium, neon, argon, krypton, xenon, and radon. These gases do not combine chemically with other materials.

nucleon A proton or neutron.

nucleus The positively charged core of an atom that contains almost all its mass.

nuclide A particular isotope of an element, identified by the number of protons and neutrons in the nucleus.

optical isomerism A form of isomerism in which two isomers are the same in every way except that they are mirror images that cannot be superimposed on each other.

© Diagram Visual Information Ltd.

orbital An area around an atom or molecule where there is a high probability of finding an electron.

ore A mineral from which a metal or non-metal may be profitably extracted.

oxidation The process by which a substance gains oxygen, loses hydrogen, or loses electrons.

oxidation state The sum of negative and positive charges in an atom.

oxide A compound consisting only of oxygen and another element. Oxides can be either ionic or covalent.

oxidizing agent A substance that can cause the oxidation of another substance by being reduced itself.

oxygen (O) A colorless, odorless gaseous element. It the most common element in Earth's crust and is the basis for respiration in plants and animals..

ozone (O_3) One of the two allotropes of oxygen. A bluish gas with a penetrating smell, it is a strong oxidizing agent.

period The horizontal rows of elements in the periodic table.

periodic table A table of elements, arranged in ascending order of atomic number, that summarizes the major properties of the elements.

periodicity Recurring at regular intervals.

peroxide A compound that contains the peroxide ion O_2^{2-}. Peroxides are strong oxidizing agents.

pH A scale from 0 to 14 that measures the acidity or alkalinity of a solution. A neutral solution has a pH of 7, while an acidic solution has a lower value and an alkaline solution a higher value.

pH meter A device that uses an electrochemical cell to measure pH.

phosphorescence The emission of light by an object, and the persistence of this emission over long periods, following irradiation by light or other forms of radiation.

photochemical reaction A chemical reaction that is initiated by a particular wavelength of light.

photoelectric effect The emission of electrons from metals upon the absorption of electromagnetic radiation.

photosynthesis The photochemical reaction by which green plants make carbohydrates using carbon dioxide and water.

platinum (Pt) A soft, shiny, silver metallic transition element that is malleable and ductile.

pollutant A substance that harms the environment when it mixes with air, soil, or water.

polyethene A thermoplastic polymer made by addition polymerization of ethene.

polymer A material containing very large molecules built up from a series of repeated small basic units (monomers).

polymerization The building up of long chain hydrocarbons from smaller ones.

polysaccharide A organic polymer composed of many simple sugars (monosaccharides).

precipitate An insoluble substance formed by a chemical reaction.

product A substance produced during a chemical reaction.

protein A large, complex molecule composed of a long chain of amino acids.

proton The positively charged particle found in the nucleus of the atom.

protostar The early stage in a star's formation before the onset of nuclear burning.

quantum number The number used when describing the energy levels available to atoms and molecules.

racemate A mixture of equal amounts of left- and right-handed stereoisomers of a chiral molecule.

radiation Energy that is transmitted in the form of particles, rays, or waves.

radical A group of atoms forming part of many molecules.

radioactive decay The process by which unstable radioactive atoms are transformed into stable, non-radioactive atoms.

radioactivity The spontaneous disintegration of certain isotopes accompanied by the emission of radiation.

rate of reaction The speed at which a chemical reaction proceeds.

reactant A substance present at the start of a chemical reaction that takes part in the reaction.

reaction A process in which substances react to form new substances.

reactivity The ability of substances to react to form new substances.

reactivity series of metals Metallic elements arranged in order of their decreasing chemical reactivity.

reagent A substance that takes part in a chemical reaction, one that is usually used to bring about a chemical change.

red giant A very large, cool star in the final stages of its life.

redox reaction A process in which one substance is reduced and another is oxidized at the same time.

reducing agent A chemical that can reduce another while being oxidized itself.

reduction A chemical reaction in which a substance gains electrons, looses oxygen, or gains hydrogen. It is the reverse of oxidation.

reforming The conversion of straight chain molecules into those that are branched in order to improve their efficiencies.

residfining The process used on the residue fraction of crude oil to convert it into a usable product.

residue The solid remaining after the completion of a chemical process.

resonance structure In organic chemistry, a diagrammatic tool to symbolize bonds between atoms in molecules.

respiration The chemical reaction by which an organism derives energy from food.

reverse reaction A reaction in which the products are converted into reactants.

reversible reaction A chemical reaction that can proceed in either direction. It does not reach completion but achieves dynamic equilibrium.

R_f value The ratio of the distance moved by a substance in a chromatographic separation to the distance moved by the solvent.

rust A reddish-brown oxide coating on iron or steel caused by the action of oxygen and water.

salt A compound formed from an acid in which all or part of the hydrogen atoms are replaced by a metal or metal-like group. Salts are generally crystalline.

saponification The treatment of an ester (hydrolysis) with a strong alkaline solution to form a salt of a carboxylic acid and an alcohol.

© Diagram Visual Information Ltd.

saturated A solution where there is an equilibrium between the solution and its solute.

scandium (Sc) Silvery-white metallic element in the lanthanide series found in nature only in minute quantities.

sewage Wastewater from domestic and industrial sources.

shell A group of orbitals at a similar distance from an atomic nucleus.

silver (Ag) A white, shiny, ductile metallic element.

silver nitrate ($AgNO_3$) A very soluble white salt that decomposes to form silver, oxygen, and nitrogen dioxide on heating.

slag Waste material that collects on the surface of a molten metal during the process of either extraction or refining.

smelting The process of extracting a metal from its ores.

soap A cleansing agent made from fatty acids derived from natural oils and fats.

sodium (Na) A soft, silver-white metallic element.

sodium chloride (NaCl) A nonvolatile ionic compound that is soluble in water.

sodium hydroxide (NaOH) A white, translucent, crystalline solid that forms a strongly alkaline solution in water.

sol A liquid solution or suspension of a colloid.

solid A state of matter in which the particles are not free to move but in which they can vibrate about fixed positions.

solubility A measure of the quantity of a solute that will dissolve in a certain amount of solvent to form a saturated solution under certain conditions of temperature and pressure.

solubility curve A graphic representation of the changing solubility of a solute in a solvent at different temperatures.

soluble A relative term that describes a substance that can dissolve in a particular solvent.

solute A substance that dissolves in a solvent and thus forms a solution.

solution A uniform mixture of one or more solutes in a solvent.

solvent A substance, usually a liquid, in which a solute dissolves to form a solution.

species The common name for entities (atoms, molecules, molecular fragments, and ions) being subjected to investigation.

spectrum The arrangement of electromagnetic radiation into its constituent wavelengths.

starch A polysaccharide with the formula ($C_6H_{10}O_5$). It is composed of many molecules of glucose.

stationary phase That which the mobile phase moves on. In paper chromatography it is the paper.

stoichiometry The calculation of the quantities of reactants and products involved in a chemical reaction.

subatomic particles The particles from which atoms are made. Neutrons and protons are found in the nucleus of the atom. Electrons form a cloud around the nucleus.

sucrose A disaccharide sugar that occurs naturally in most plants.

sulfate A salt or ester of sulfuric acid.

sulfide A compound of sulfur and a more electropositive element.

sulfur (S) A yellow, non-metallic element that is found abundantly in nature.

sulfuric acid (H_2SO_4) An oily, colorless, odorless liquid that is extremely corrosive.

sulfur dioxide (SO_2) A colorless gas with a pungent odor of burning sulfur. It is very soluble in water.

sulfur trioxide (SO_3) A white, soluble solid that fumes in moist air. It reacts violently with water to form sulfuric acid.

supernova The explosion caused when a massive star dies and collapses.

surface area The sum of the area of the faces of a solid.

suspension A type of dispersion. Small solid particles are dispersed in a liquid or gas.

tensile strength The amount of stress a material can stand without breaking.

thorium Th. A gray, radioactive metallic element used as fuel in nuclear reactors.

thorium series One of the naturally occurring radioactive series.

titanium (Ti) A lightweight, gray metallic element that is very strong and resistant to corrosion.

titration In analytical chemistry, A technique used to determine the concentration of a solute in a solution.

transition metals Metallic elements that have an incomplete inner electron structure and exhibit variable valencies.

triple bond A covalent bond formed between two atoms in which three pairs of electrons contribute to the bond.

ultraviolet Electromagnetic radiation of shorter wavelengths than visible light, but of longer wavelength than X rays.

unit cell The smallest repeating array of atoms, ions, or molecules in a crystal.

universal indicator A mixture of substances that shows a gradual color change over a wide range of pH values.

uranium (U) A hard, white, radioactive metallic element used in nuclear reactors and nuclear weapons.

uranium series One of the naturally occurring radioactive series.

valency The measure of an element's ability to combine with other elements.

vanadium (V) A silvery-white or gray metallic element used as a steel additive and in catalysts.

van der Waals forces Weak intermolecular or interatomic forces between neutral molecules or atoms. They are much weaker than chemical bonds.

viscosity A measure of the resistance of a fluid to flow.

wavelength The distance between two corresponding points on a wave.

white dwarf The small, dense remnant of a star near the end of its period of nuclear fusion.

zinc (Zn) A hard, brittle, bluish-white metallic element used in alloys and in galvanizing.

zwitterion An ion that carries both a positive and negative charge.

© Diagram Visual Information Ltd.

Internet resources

There is a lot of useful information on the internet. Information on a particular topic may be available through a search engine such as Google (**http://www.google.com**). Some of the Web sites that are found in this way may be very useful, others not. Below is a selection of Web sites related to the material covered by this book.

The publisher takes no responsibility for the information contained within these Web sites. All the sites were accessible on March 1, 2006.

About Chemistry

Includes links to a glossary, encyclopedia, experiments, periodic table, chemical structure archive, chemistry problems, and articles.

http://chemistry.about.com

Allchemicals.info

Hundreds of definitions and descriptions from absolute zero to zinc.

http://www.allchemicals.info

Chem4Kids

Accessible information on matter, atoms, elements, reactions, biochemistry, and much more, for grades 5–9.

http://www.chem4kids.com

Chemistry Carousel: A Trip Around the Carbon Cycle

Site explaining the carbon cycle.

http://library.thinkquest.org/11226

Chemistry Central

Offers basic atomic information, information on the periodic table, chemical bonding, and organic chemistry as well as extensive links to a wide variety of other resources.

http://users.senet.com.au/~rowanb/chem

Chemistry.org

Offers publications, career advice, information, and curriculum materials for K–12.

http://www.acs.org/

The Chemistry Research Center

Offers high school students links to useful sites for help with homework.

http://library.thinkquest.org/21192

Chemistry Tutor

Help for high school students with chemistry homework. Includes an introduction to chemistry, equations, calculations, types of reactions, information on lab safety, and links to other sources.

http://library.thinkquest.org/2923

ChemSpy.com

Links to chemistry and chemical engineering terms, definitions, synonyms, acronyms, and abbreviations.

http://www.chemspy.com

Chemtutor

A guide to the basics of chemistry for high school and college students.

http://www.chemtutor.com

CHEMystery

A virtual chemistry textbook, providing an interactive guide for high school chemistry students and links to other resources.

http://library.thinkquest.org/3659

Common Molecules

Information and 3-D presentation on molecules studied in chemistry classes or of interest for their structural properties.

http://www.reciprocalnet.org/edumodules/commonmolecules

Delights of Chemistry

Presents more than 40 chemistry demonstrations and 500 photographs/animations of experiments and chemical reactions.

http://www.chem.leeds.ac.uk/delights

EnvironmentalChemistry.com

Includes a chemical and environmental dictionary; a detailed periodic table of elements; articles on environmental and hazardous materials issues; a geologic timeline.

http://environmentalchemistry.com

Eric Weisstein's World of CHEMISTRY

Online encyclopedia, still under construction, with excellent graphics; good source for chemical reactions.

http://scienceworld.wolfram.com/chemistry

General Chemistry Online

Contains searchable glossary, frequently asked questions, database of compounds, tutorials, simulations, and toolbox of periodic table and calculators.

http://antoine.frostburg.edu/chem/senese/101

© Diagram Visual Information Ltd.

IUPAC Nomenclature Home Page

Definitions of terms used in chemistry provided by the International Union of Pure and Applied Chemistry. The "Gold book" is particularly good for basic terms.

http://www.chem.qmul.ac.uk/iupac

The Learning Matters of Chemistry

Offers visualizations of molecules and atomic orbits, interactive chemistry exercises, and links to other resources.

http://www.knowledgebydesign.com/tlmc

The Macrogalleria: A Cyberwonderland of Polymer Fun

An Internet "mall" for learning about polymers and polymer science.

http://www.pslc.ws/macrog

Nuclear Chemistry and the Community

Introduction to nuclear chemistry and its impact on society.

http://www.chemcases.com/nuclear

Open Directory Project: Biochemistry and Molecular Biology

A comprehensive listing of internet resources in the field of biochemistry.

http://dmoz.org/Science/Biology/Biochemistry_and_Molecular_Biology

Open Directory Project: Chemistry

A comprehensive listing of internet resources in the field of chemistry.

http://dmoz.org/science/chemistry

The pH Factor

Introduction to acids and bases for middle school students.

http://www.miamisci.org/ph

PSIgate: Chemistry

Offers interactive tutorials, timeline, and links, in many areas.

http://www.psigate.ac.uk/newsite/chemistry-gateway

Reactive Reports

Web chemistry magazine offering news stories and links to sites.

http://www.reactivereports.com

ScienceMaster

News, information, links, columns, and homework help in all major areas of science.

http://www.sciencemaster.com

Science News for Kids

Science Service Suggestions for hands-on activities, books, articles, Web resources, and other useful materials for students ages 9–13.

http://www.sciencenewsforkids.org

The Science of Spectroscopy

Introduction to spectroscopy with descriptions of common spectroscopic analysis techniques, as well as applications of spectroscopy in consumer products, medicine, and space science.

http://www.scienceofspectroscopy.info

Virtual Chemistry

3-D simulated laboratory for teaching chemistry, with links to an online encyclopedia, tutorials, and close-ups of molecules.

http://neon.chem.ox.ac.uk/vrchemistry

A Visual Interpretation of the Table of Elements

Striking visual representations of 110 elements. Site includes detailed information on the elements and on the history of the periodic table.

http://www.chemsoc.org/viselements

Web Elements™ Periodic Table Scholar Edition

High quality source of information about the periodic table for students. There is also a professional edition.

http://www.webelements.com/webelements/scholar

What's that Stuff?

Explores the chemistry of everyday objects.

http://pubs.acs.org/cen/whatstuff/stuff.html

© Diagram Visual Information Ltd.

Index

Index of subject headings.

a

acids
  neutralization of 130
  preparation of 124
  reactions of 123, 170
  titration of 53–4
actinium series 187
air, gases in 69
alcohols 163, 170
alkalis, neutralization of 127
alkanes 156
  reaction summary 169
alkenes 157
  reaction summary 169
allotropes
  of carbon 148–9
  of oxygen 82
  of sulfur 82
aluminum 106
  reactions of 112
amino acids 172
ammonia 67, 73–7
  Haber process 74–5
  properties of 73
atomic emission spectrum 22
atomic mass 35
atomic structure 13
atomic volumes 33–4
Avogadro’s constant 20

b

bases 125–6
  neutralization of 128
boiling points 31–2
bond dissociation energies 140
bonding of chemicals 41–4

c

carbon, chemistry of 148–74
carbon chains 154
carbon cycle 150
carbon oxides 151
carboxylic acids 164
catalysts 141–2
cathode ray oscilloscopes 17
changes in matter 45–63
chemical combination 41–4
chemical reactions 119–47
chlorine, compounds of 93
chromatography 49–50
collision theory 131
colloids 46
compounds, molecular, masses of 37
contact process 85–6
coordinate bonding 44
copper
  reactions of 111–12
  smelting and conversion of 110
copper sulfate, solubility of 63
covalent bonding 43
crude oil, fractional distillation of 152
crystal structure of metals 39–40
crystals, ionic 38

d

decay sequences
  actinium series 187
  neptunium series 189
  radioactivity of 190
  thorium series 188
  uranium series 186
detergents 166
diamonds 148
disaccharides 174
distillation, simple and fractional 47

e

efficient packing 40
electrode activity and concentration 122
electrolysis 121–2
electron structure of metals 104
electrons 15–18
elements
  atomic volumes of 33
  boiling points of 31
  first ionization energies of 27
  melting points of 29
  organization of 25
  periodic table 26, 36
endothermic reactions 139
energy levels 23
esters 165
ethene 158
exothermic reactions 139

f

fertilizers, nitrate 79
first ionization energy 27–8
fission, nuclear 184
flame tests on metals 115
fractional distillation 47
  of crude oil 152
fullerenes 149
functional groups 162

g

gas-liquid chromatography 50
gases in air 69
Geiger and Marsden’s apparatus 14
graphite 148

h

Haber process 74–5
half-life 180–1
halogens 91–2, 95–7
homologous series 162
hydrocarbons, naming of 155
hydrogen 22–3, 64–8
hydrogen chloride 94

i

ionic bonding 41
ionic crystals, structure of 38
ionic radicals 42
ionic salts 59
ionic solutions 60
ionization energies of transition metals 105
ionizing radiation 175
iron
  smelting of 107
  reactions of 112
isotopes
  naturally occurring 191–7
  radioactive 182
  stable and unstable 179

© Diagram Visual Information Ltd.

l

lattice structure of metals 39
luminescence 24

m

mass
  atomic 35
  molecular 37
mass spectrometry 50
melting points 29–30
metal hydroxides 116
metal ions 117
metallic carbonates 129
metals
  crystal structure of 39–40
  electron structure of 104
  extraction of 113
  group 1 and group 2 100–3
  ores of 99
  reactivity of 114, 119–20
  tests on 115–17
  uses of 118
  world distribution of 98
mixtures 45
molecular mass 37
mole, the 21
molecules, size and motion of 19
monosaccharides 173

n

naturally occurring isotopes 191–7
neptunium series 189
neutralization of alkalis, bases and acids 127–30
nitrate fertilizers 79
nitric acid 76–7
nitrogen 70–1, 77–8
nitrogen cycle 72
nuclear power 183–5

o

optical isomerism 171
ores of metals 99, 113
organic compounds 167
oscilloscopes 17
oxidation and oxidation states 88, 143, 147
oxides of sulfur 84
oxygen 80–3, 88–9

p

paper chromatography 49
periodic table 26, 36
pH scale 51, 55
planets
  composition of 11
  density, size, and atmosphere of 12
polymers 159–61
polysaccharides 174
proteins 172
proton transfer 127–30

r

radiations
  detection of 176
  properties of 177–8
radioactivity 175–97
  of decay sequences 190
reaction rates 132–8
reactivity of metals 119–20
redox reactions 144–6
  of halogens 96
reduction of oxygen and sulfur 88
refining processes 152–3
rusting 109

s

salts
  formation of 126
  ionic 59
separation of solutions 48
sewage 57
simple distillation 47
smelting
  of copper 110
  of iron 107
soaps 166
sodium 101
soil 55
solar system 10
solubility, solutions and solvents 45, 48, 59, 61–3
stars 8–9
steel manufacture 108
sulfur 80–3, 88, 90
  oxides of 84
sulfuric acid 85–7

t

temperature, effect on reaction rates of 137–8
thorium series 188
titration of acids 53–4
transition metals 104–5
  as catalysts 142

u

uranium series 186

w

water 56–9, 67, 87

© Diagram Visual Information Ltd.

Flexibook

Color Atlas of Biochemistry

J. Koolman
K. H. Roehm

Second edition, revised and enlarged

basic sciences

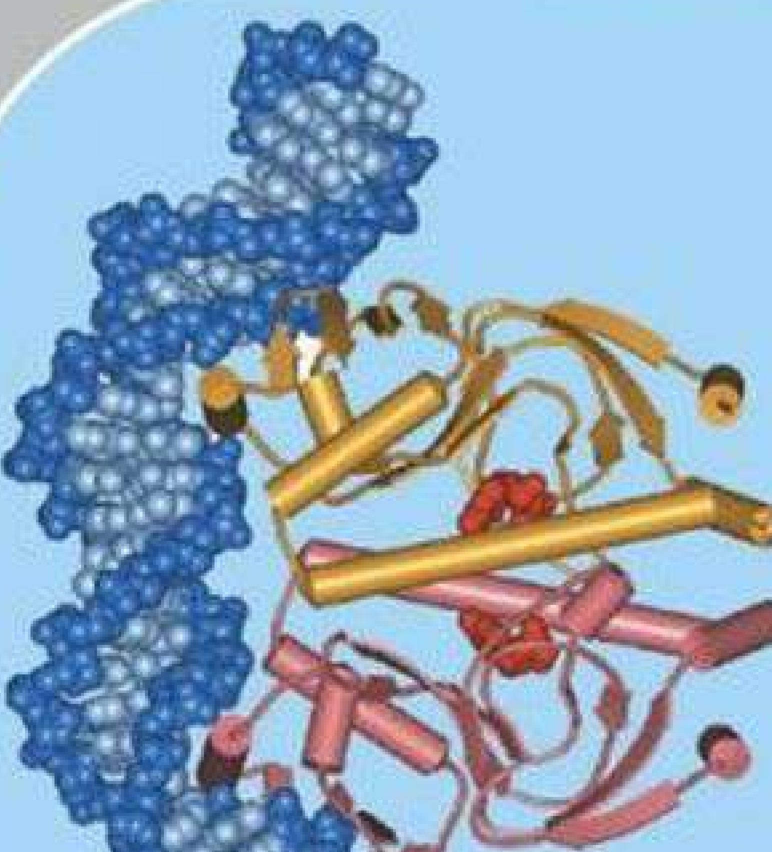

Thieme

Contents

Introduction 1

Basics

Chemistry

Periodic table. 2
Bonds 4
Molecular structure 6
Isomerism 8
Biomolecules I 10
Biomolecules II 12
Chemical reactions. 14

Physical Chemistry

Energetics 16
Equilibriums 18
Enthalpy and entropy. 20
Reaction kinetics 22
Catalysis 24
Water as a solvent 26
Hydrophobic interactions. 28
Acids and bases 30
Redox processes. 32

Biomolecules

Carbohydrates

Overview. 34
Chemistry of sugars 36
Monosaccharides and disaccharides . . . 38
Polysaccharides: overview 40
Plant polysaccharides. 42
Glycosaminoglycans and glycoproteins. 44

Lipids

Overview. 46
Fatty acids and fats 48
Phospholipids and glycolipids 50
Isoprenoids 52
Steroid structure 54
Steroids: overview. 56

Amino Acids

Chemistry and properties. 58
Proteinogenic amino acids 60
Non-proteinogenic amino acids 62

Peptides and Proteins

Overview. 64
Peptide bonds 66
Secondary structures 68
Structural proteins 70
Globular proteins 72
Protein folding 74
Molecular models: insulin. 76
Isolation and analysis of proteins 78

Nucleotides and Nucleic Acids

Bases and nucleotides. 80
RNA 82
DNA. 84
Molecular models: DNA and RNA 86

Metabolism

Enzymes

Basics. 88
Enzyme catalysis 90
Enzyme kinetics I 92
Enzyme kinetics II 94
Inhibitors 96
Lactate dehydrogenase: structure 98
Lactate dehydrogenase: mechanism . . . 100
Enzymatic analysis 102
Coenzymes 1 104
Coenzymes 2 106
Coenzymes 3 108
Activated metabolites 110

Metabolic Regulation

Intermediary metabolism 112
Regulatory mechanisms 114
Allosteric regulation 116
Transcription control 118
Hormonal control. 120

Energy Metabolism

ATP 122
Energetic coupling 124
Energy conservation at membranes. . . . 126
Photosynthesis: light reactions 128
Photosynthesis: dark reactions 130
Molecular models: membrane proteins. 132
Oxoacid dehydrogenases. 134
Tricarboxylic acid cycle: reactions 136
Tricarboxylic acid cycle: functions. 138
Respiratory chain 140
ATP synthesis 142
Regulation 144
Respiration and fermentation 146
Fermentations 148

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Carbohydrate Metabolism

Glycolysis 150
Pentose phosphate pathway 152
Gluconeogenesis. 154
Glycogen metabolism 156
Regulation 158
Diabetes mellitus 160

Lipid Metabolism

Overview 162
Fatty acid degradation 164
Minor pathways of fatty acid degradation 166
Fatty acid synthesis 168
Biosynthesis of complex lipids 170
Biosynthesis of cholesterol 172

Protein Metabolism

Protein metabolism: overview 174
Proteolysis 176
Transamination and deamination 178
Amino acid degradation 180
Urea cycle 182
Amino acid biosynthesis 184

Nucleotide Metabolism

Nucleotide degradation. 186
Purine and pyrimidine biosynthesis . . . 188
Nucleotide biosynthesis 190

Porphyrin Metabolism

Heme biosynthesis 192
Heme degradation 194

Organelles

Basics

Structure of cells 196
Cell fractionation 198
Centrifugation 200
Cell components and cytoplasm 202

Cytoskeleton

Components. 204
Structure and functions 206

Nucleus 208

Mitochondria

Structure and functions 210
Transport systems 212

Biological Membranes

Structure and components 214
Functions and composition 216
Transport processes 218
Transport proteins 220
Ion channels. 222
Membrane receptors 224

Endoplasmic Reticulum and Golgi Apparatus

ER: structure and function. 226
Protein sorting 228
Protein synthesis and maturation 230
Protein maturation 232

Lysosomes. 234

Molecular Genetics

Overview 236
Genome 238
Replication 240
Transcription. 242
Transcriptional control 244
RNA maturation 246
Amino acid activation 248
Translation I: initiation 250
Translation II: elongation and termination. 252
Antibiotics 254
Mutation and repair 256

Genetic engineering

DNA cloning 258
DNA sequencing 260
PCR and protein expression. 262
Genetic engineering in medicine 264

Tissues and organs

Digestion

Overview 266
Digestive secretions. 268
Digestive processes 270
Resorption 272

Blood

Composition and functions 274
Plasma proteins. 276
Lipoproteins 278
Hemoglobin 280
Gas transport 282
Erythrocyte metabolism 284
Iron metabolism 286
Acid–base balance. 288
Blood clotting 290
Fibrinolysis, blood groups 292

Immune system

Immune response 294
T-cell activation. 296
Complement system 298
Antibodies 300
Antibody biosynthesis. 302
Monoclonal antibodies, immunoassay . 304

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Liver

Functions 306
Buffer function in organ metabolism 308
Carbohydrate metabolism 310
Lipid metabolism 312
Bile acids 314
Biotransformations 316
Cytochrome P450 systems 318
Ethanol metabolism 320

Kidney

Functions 322
Urine 324
Functions in the acid–base balance 326
Electrolyte and water recycling 328
Renal hormones 330

Muscle

Muscle contraction 332
Control of muscle contraction 334
Muscle metabolism I 336
Muscle metabolism II 338

Connective tissue

Bone and teeth 340
Calcium metabolism 342
Collagens 344
Extracellular matrix 346

Brain and Sensory Organs

Signal transmission in the CNS 348
Resting potential and action potential 350
Neurotransmitters 352
Receptors for neurotransmitters 354
Metabolism 356
Sight 358

Nutrition

Nutrients

Organic substances 360
Minerals and trace elements 362

Vitamins

Lipid-soluble vitamins 364
Water-soluble vitamins I 366
Water-soluble vitamins II 368

Hormones

Hormonal system

Basics 370
Plasma levels and hormone hierarchy 372

Lipophilic hormones 374
Metabolism of steroid hormones 376
Mechanism of action 378

Hydrophilic hormones 380
Metabolism of peptide hormones 382
Mechanisms of action 384
Second messengers 386
Signal cascades 388

Other signaling substances

Eicosanoids 390
Cytokines 392

Growth and development

Cell proliferation

Cell cycle 394
Apoptosis 396
Oncogenes 398
Tumors 400
Cytostatic drugs 402

Viruses 404

Metabolic charts 406
Calvin cycle 407
Carbohydrate metabolism 408
Biosynthesis of fats and membrane liquids 409
Synthesis of ketone bodies and steroids 410
Degradation of fats and phospholipids 411
Biosynthesis of the essential amino acids 412
Biosynthesis of the non-essential amino acids 413
Amino acid degradation I 414
Amino acid degradation II 415
Ammonia metabolism 416
Biosynthesis of purine nucleotides 417
Biosynthesis of the pyrimidine nucleotides and C_1 metabolism 418
Nucleotide degradation 419

Annotated enzyme list 420

Abbreviations 431

Quantities and units 433

Further reading 434

Source credits 435

Index 437

Key to color-coding:
see front and rear inside covers

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Introduction

This paperback atlas is intended for students of medicine and the biological sciences. It provides an introduction to biochemistry, but with its modular structure it can also be used as a reference book for more detailed information. The 216 color plates provide knowledge in the field of biochemistry, accompanied by detailed information in the text on the facing page. The degree of difficulty of the subject-matter is indicated by symbols in the text:

● stands for "basic biochemical knowledge"
◑ indicates "standard biochemical knowledge"
○ means "specialist biochemical knowledge."

Some general rules used in the structure of the illustrations are summed up in two *explanatory plates* inside the front and back covers. Keywords, definitions, explanations of unfamiliar concepts and chemical formulas can be found using the *index*. The book starts with a few **basics** in biochemistry (pp. 2–33). There is a brief explanation of the concepts and principles of chemistry (pp. 2–15). These include the periodic table of the elements, chemical bonds, the general rules governing molecular structure, and the structures of important classes of compounds. Several basic concepts of *physical chemistry* are also essential for an understanding of biochemical processes. Pages 16–33 therefore discuss the various forms of energy and their interconversion, reaction kinetics and catalysis, the properties of water, acids and bases, and redox processes.

These basic concepts are followed by a section on the structure of the important biomolecules (pp. 34–87). This part of the book is arranged according to the different classes of metabolites. It discusses carbohydrates, lipids, amino acids, peptides and proteins, nucleotides, and nucleic acids.

The next part presents the reactions involved in the interconversion of these compounds—the part of biochemistry that is commonly referred to as **metabolism** (pp. 88–195). The section starts with a discussion of the enzymes and coenzymes, and discusses the mechanisms of metabolic regulation and the so-called *energy metabolism*. After this, the central metabolic pathways are presented, once again arranged according to the class of metabolite (pp. 150–195).

The second half of the book begins with a discussion of the functional compartments within the cell, the **cellular organelles** (pp. 196–235). This is followed on pp. 236–265 by the current field of **molecular genetics** (*molecular biology*). A further extensive section is devoted to the biochemistry of individual **tissues and organs** (pp. 266–359). Here, it has only been possible to focus on the most important organs and organ systems—the digestive system, blood, liver, kidneys, muscles, connective and supportive tissues, and the brain.

Other topics include the biochemistry of **nutrition** (pp. 360–369), the structure and function of important **hormones** (pp. 370–393), and **growth and development** (pp. 394–405).

The paperback atlas concludes with a series of schematic **metabolic "charts"** (pp. 407–419). These plates, which are not accompanied by explanatory text apart from a brief introduction on p. 406, show simplified versions of the most important synthetic and degradative pathways. The charts are mainly intended for reference, but they can also be used to review previously learned material. The enzymes catalyzing the various reactions are only indicated by their EC numbers. Their names can be found in the systematically arranged and annotated enzyme list (pp. 420–430).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Periodic table

A. Biologically important elements ◑

There are 81 stable elements in nature. Fifteen of these are present in all living things, and a further 8–10 are only found in particular organisms. The illustration shows the first half of the **periodic table**, containing all of the biologically important elements. In addition to physical and chemical data, it also provides information about the distribution of the elements in the living world and their abundance in the human body. The laws of atomic structure underlying the periodic table are discussed in chemistry textbooks.

More than 99% of the atoms in animals' bodies are accounted for by just four elements—hydrogen (H), oxygen (O), carbon (C) and nitrogen (N). Hydrogen and oxygen are the constituents of water, which alone makes up 60–70% of cell mass (see p. 196). Together with carbon and nitrogen, hydrogen and oxygen are also the major constituents of the **organic compounds** on which most living processes depend. Many biomolecules also contain sulfur (S) or phosphorus (P). The above **macroelements** are essential for all organisms.

A second biologically important group of elements, which together represent only about 0.5% of the body mass, are present almost exclusively in the form of **inorganic ions**. This group includes the *alkali metals* sodium (Na) and potassium (K), and the *alkaline earth metals* magnesium (Mg) and calcium (Ca). The halogen *chlorine* (Cl) is also always ionized in the cell. All other elements important for life are present in such small quantities that they are referred to as **trace elements**. These include transition metals such as iron (Fe), zinc (Zn), copper (Cu), cobalt (Co) and manganese (Mn). A few *nonmetals,* such as iodine (I) and selenium (Se), can also be classed as essential trace elements.

B. Electron configurations: examples ○

The chemical properties of atoms and the types of bond they form with each other are determined by their electron shells. The **electron configurations** of the elements are therefore also shown in Fig. **A**. Fig. **B** explains the symbols and abbreviations used. More detailed discussions of the subject are available in chemistry textbooks.

The possible states of electrons are called **orbitals**. These are indicated by what is known as the principal quantum number and by a letter—s, p, or d. The orbitals are filled one by one as the number of electrons increases. Each orbital can hold a maximum of two electrons, which must have oppositely directed "spins." Fig. **A** shows the distribution of the electrons among the orbitals for each of the elements. For example, the six electrons of carbon (**B1**) occupy the 1s orbital, the 2s orbital, and two 2p orbitals. A filled 1s orbital has the same electron configuration as the noble gas helium (He). This region of the electron shell of carbon is therefore abbreviated as "He" in Fig. **A**. Below this, the numbers of electrons in each of the other filled orbitals (2s and 2p in the case of carbon) are shown on the right margin. For example, the electron shell of chlorine (**B2**) consists of that of neon (Ne) and seven additional electrons in 3s and 3p orbitals. In iron (**B3**), a transition metal of the first series, electrons occupy the 4s orbital even though the 3d orbitals are still partly empty. Many reactions of the transition metals involve empty d orbitals—e. g., redox reactions or the formation of complexes with bases.

Particularly stable electron arrangements arise when the outermost shell is fully occupied with eight electrons (the "**octet rule**"). This applies, for example, to the noble gases, as well as to ions such as Cl^- ($3s^2 3p^6$) and Na^+ ($2s^2 2p^6$). It is only in the cases of hydrogen and helium that two electrons are already sufficient to fill the outermost 1s orbital.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biologically important elements

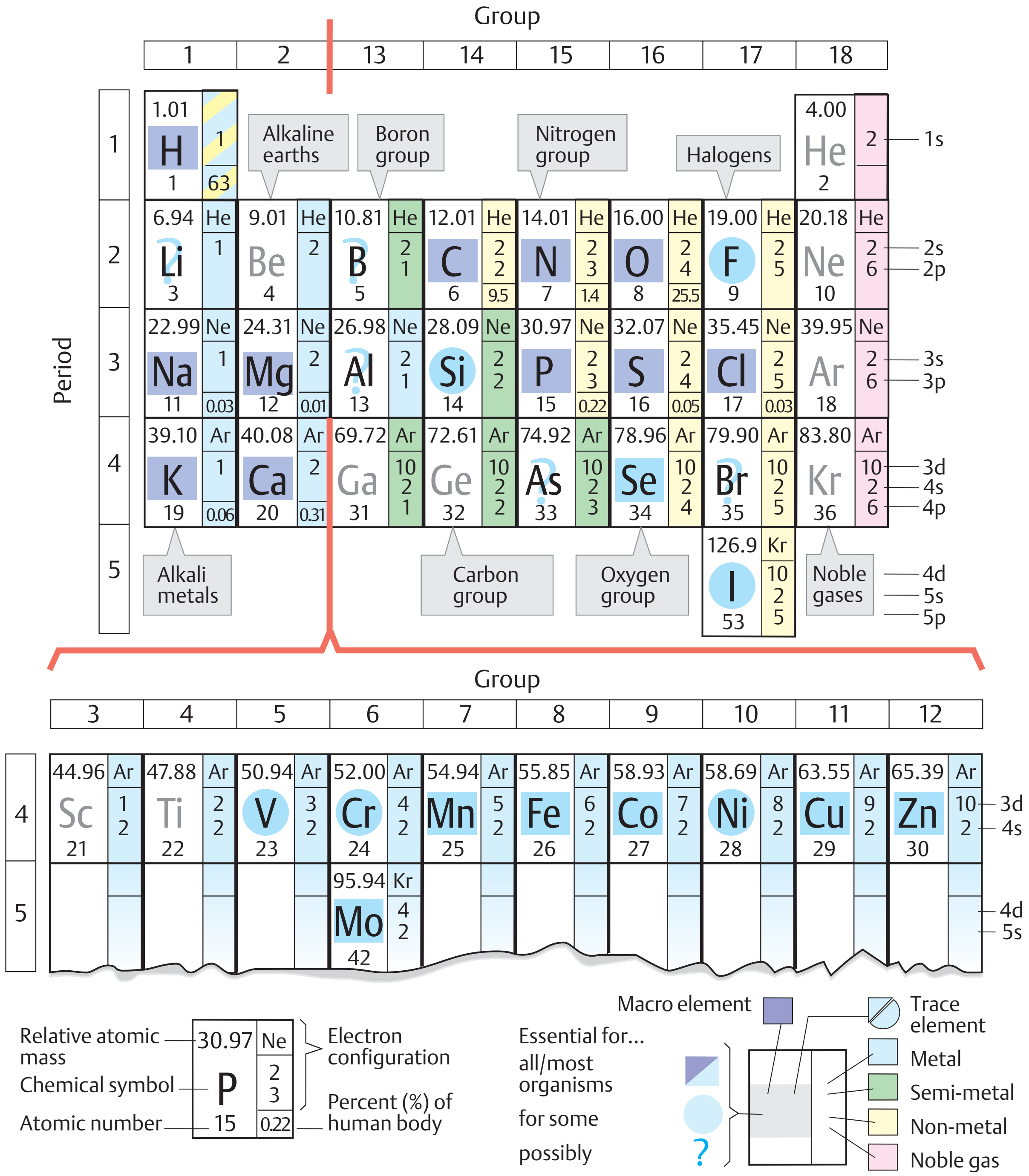
Group
1
2
13
14
15
16
17
18
Period
1
2
3
4
5
Alkaline earths
Boron group
Nitrogen group
Halogens
Alkali metals
Carbon group
Oxygen group
Noble gases
1.01 H 1 1 63
4.00 He 2 2
6.94 Li 3 He 1
9.01 Be 4 He 2
10.81 B 5 He 2 1
12.01 C 6 He 2 2 9.5
14.01 N 7 He 2 3 1.4
16.00 O 8 He 2 4 25.5
19.00 F 9 He 2 5
20.18 Ne 10 He 2 6
22.99 Na 11 Ne 1 0.03
24.31 Mg 12 Ne 2 0.01
26.98 Al 13 Ne 2 1
28.09 Si 14 Ne 2 2
30.97 P 15 Ne 2 3 0.22
32.07 S 16 Ne 2 4 0.05
35.45 Cl 17 Ne 2 5 0.03
39.95 Ar 18 Ne 2 6
39.10 K 19 Ar 1 0.06
40.08 Ca 20 Ar 2 0.31
69.72 Ga 31 Ar 10 2 1
72.61 Ge 32 Ar 10 2 2
74.92 As 33 Ar 10 2 3
78.96 Se 34 Ar 10 2 4
79.90 Br 35 Ar 10 2 5
83.80 Kr 36 Ar 10 2 6
126.9 I 53 Kr 10 2 5
1s
2s
2p
3s
3p
3d
4s
4p
4d
5s
5p
Group
3
4
5
6
7
8
9
10
11
12
44.96 Sc 21 Ar 1 2
47.88 Ti 22 Ar 2 2
50.94 V 23 Ar 3 2
52.00 Cr 24 Ar 4 2
54.94 Mn 25 Ar 5 2
55.85 Fe 26 Ar 6 2
58.93 Co 27 Ar 7 2
58.69 Ni 28 Ar 8 2
63.55 Cu 29 Ar 9 2
65.39 Zn 30 Ar 10 2
95.94 Mo 42 Kr 4 2
3d
4s
4d
5s
Relative atomic mass
Chemical symbol
Atomic number
30.97 P 15 Ne 2 3 0.22
Electron configuration
Percent (%) of human body
Macro element
Essential for...
all/most organisms
for some
possibly
?
Trace element
Metal
Semi-metal
Non-metal
Noble gas

B. Electron configurations: examples

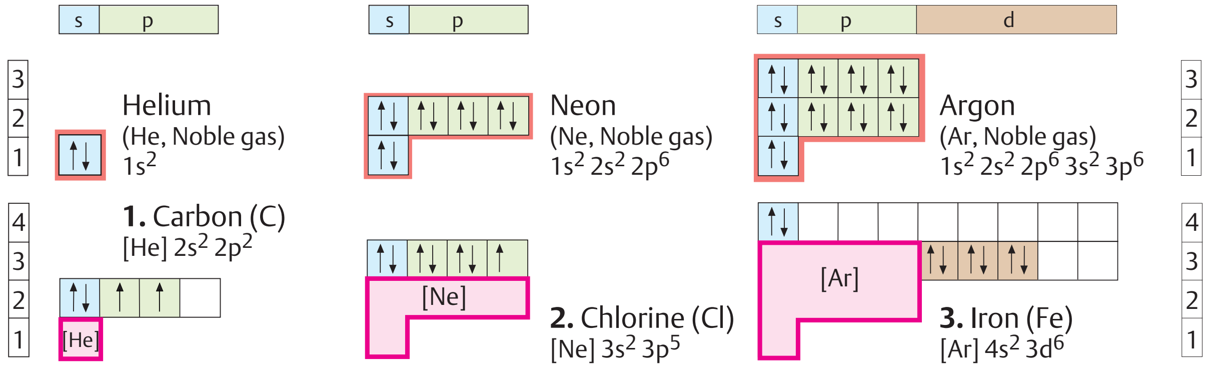
s
p
d
3
2
1
4
Helium
(He, Noble gas)
1s²
Neon
(Ne, Noble gas)
1s² 2s² 2p⁶
Argon
(Ar, Noble gas)
1s² 2s² 2p⁶ 3s² 3p⁶
1. Carbon (C)
[He] 2s² 2p²
[He]
[Ne]
2. Chlorine (Cl)
[Ne] 3s² 3p⁵
[Ar]
3. Iron (Fe)
[Ar] 4s² 3d⁶

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Bonds

A. Orbital hybridization and chemical bonding ○

Stable, covalent bonds between nonmetal atoms are produced when orbitals (see p. 2) of the two atoms form **molecular orbitals** that are occupied by one electron from each of the atoms. Thus, the four bonding electrons of the carbon atom occupy 2s and 2p atomic orbitals (**1a**). The 2s orbital is spherical in shape, while the three 2p orbitals are shaped like dumbbells arranged along the x, y, and z axes. It might therefore be assumed that carbon atoms should form at least *two different* types of molecular orbital. However, this is not normally the case. The reason is an effect known as **orbital hybridization**. Combination of the s orbital and the three p orbitals of carbon gives rise to four equivalent, tetrahedrally arranged sp^3 atomic orbitals (**sp^3 hybridization**). When these overlap with the 1s orbitals of H atoms, four equivalent σ-molecular orbitals (**1b**) are formed. For this reason, carbon is capable of forming four bonds—i. e., it has a valency of four. Single bonds between nonmetal atoms arise in the same way as the four σ or **single bonds** in methane (CH_4). For example, the hydrogen phosphate ion (HPO_4^{2-}) and the ammonium ion (NH_4^+) are also tetrahedral in structure (**1c**).

A second common type of orbital hybridization involves the 2s orbital and only *two* of the three 2p orbitals (2a). This process is therefore referred to as **sp^2 hybridization**. The result is three equivalent sp^2 hybrid orbitals lying in one plane at an angle of 120° to one another. The remaining $2p_x$ orbital is oriented perpendicular to this plane. In contrast to their sp^3 counterparts, sp^2-hybridized atoms form two *different* types of bond when they combine into molecular orbitals (**2b**). The three sp^2 orbitals enter into σ bonds, as described above. In addition, the electrons in the two $2p_x$ orbitals, known as **π electrons**, combine to give an additional, elongated π molecular orbital, which is located above and below the plane of the σ bonds. Bonds of this type are called **double bonds**. They consist of a σ bond and a π bond, and arise only when both of the atoms involved are capable of sp^2 hybridization. In contrast to single bonds, double bonds are not freely rotatable, since rotation would distort the π-molecular orbital. This is why all of the atoms lie in one plane (**2c**); in addition, *cis–trans* isomerism arises in such cases (see p. 8). Double bonds that are common in biomolecules are C=C and C=O. C=N double bonds are found in aldimines (Schiff bases, see p. 178).

B. Resonance ◑

Many molecules that have several double bonds are much less reactive than might be expected. The reason for this is that the double bonds in these structures cannot be localized unequivocally. Their π orbitals are not confined to the space between the double-bonded atoms, but form a shared, extended **π-molecular orbital**. Structures with this property are referred to as **resonance hybrids**, because it is impossible to describe their actual bonding structure using standard formulas. One can either use what are known as **resonance structures**—i. e., idealized configurations in which π electrons are assigned to specific atoms (cf. pp. 32 and 66, for example)—or one can use dashed lines as in Fig. **B** to suggest the extent of the delocalized orbitals. (Details are discussed in chemistry textbooks.)

Resonance-stabilized systems include carboxylate groups, as in *formate*; aliphatic hydrocarbons with conjugated double bonds, such as *1,3-butadiene*; and the systems known as **aromatic ring systems**. The best-known aromatic compound is *benzene,* which has six delocalized π electrons in its ring. Extended resonance systems with 10 or more π electrons absorb light within the visible spectrum and are therefore *colored.* This group includes the aliphatic carotenoids (see p. 132), for example, as well as the heme group, in which 18 π electrons occupy an extended molecular orbital (see p. 106).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Orbital hybridization and chemical bonding

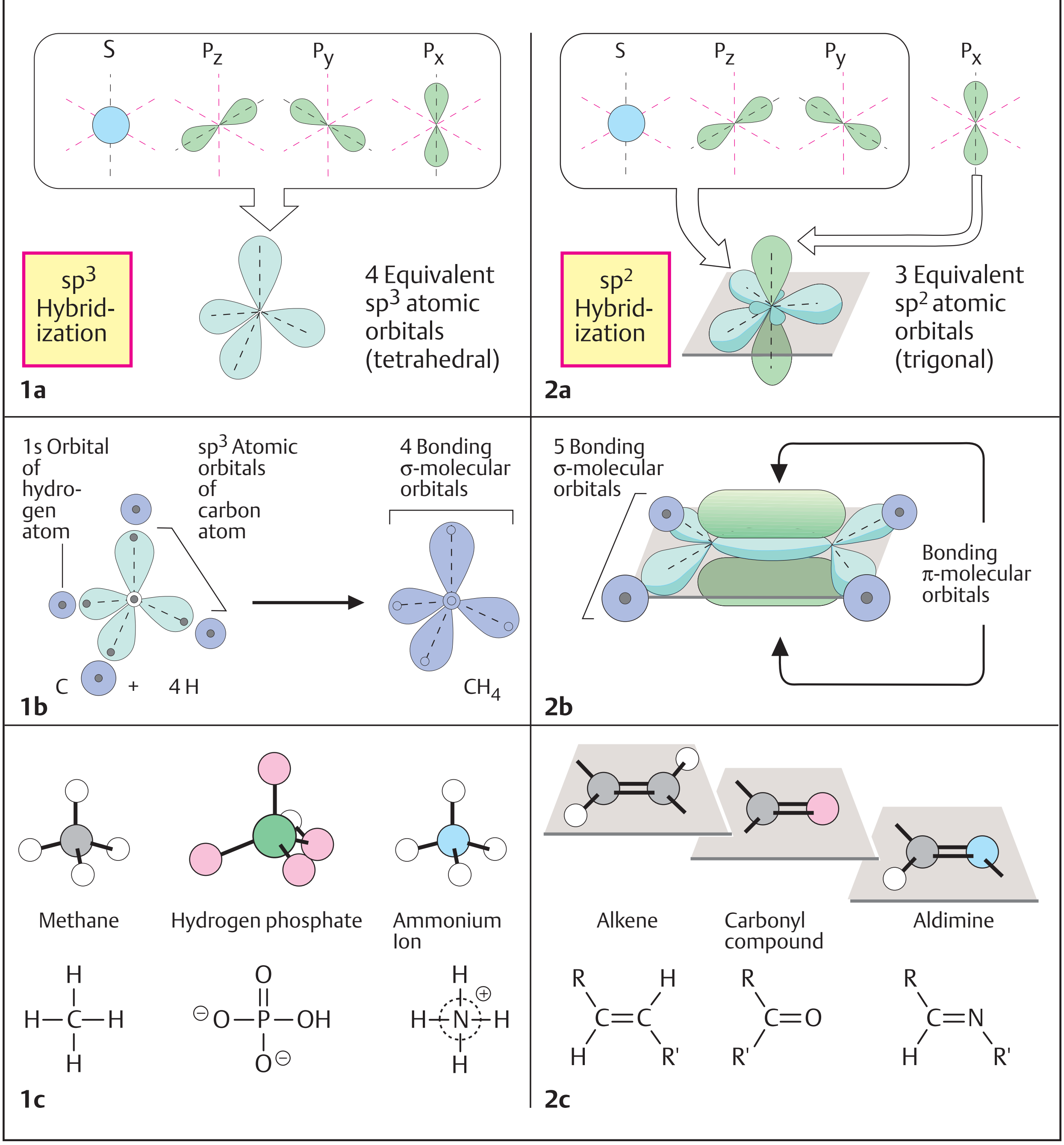
S
p_z
p_y
p_x
sp^3 Hybrid-ization
4 Equivalent sp^3 atomic orbitals (tetrahedral)
1a
S
p_z
p_y
p_x
sp^2 Hybrid-ization
3 Equivalent sp^2 atomic orbitals (trigonal)
2a
1s Orbital of hydro-gen atom
sp^3 Atomic orbitals of carbon atom
4 Bonding σ-molecular orbitals
C
+
4 H
CH_4
1b
5 Bonding σ-molecular orbitals
Bonding π-molecular orbitals
2b
Methane
Hydrogen phosphate
Ammonium Ion
H
H—C—H
H
O
⊖O—P—OH
O⊖
H
H—N—H
⊕
H
1c
Alkene
Carbonyl compound
Aldimine
R
H
C=C
H
R'
R
C=O
R'
R
C=N
H
R'
2c

B. Resonance

	Formate	1,3-Butadiene	Benzene
π-Molecular orbitals	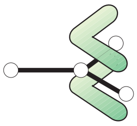	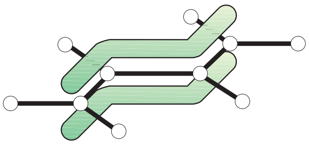	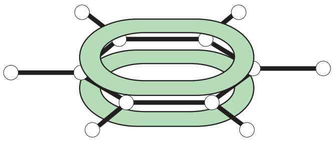
Formula	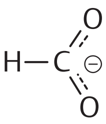	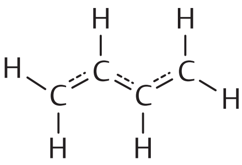 H H H C C C C H H H	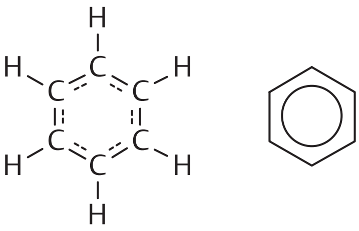 H H C C H C C C C H C H H

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Molecular structure

The physical and chemical behavior of molecules is largely determined by their **constitution** (the type and number of the atoms they contain and their bonding). Structural formulas can therefore be used to predict not only the chemical reactivity of a molecule, but also its size and shape, and to some extent its conformation (the spatial arrangement of the atoms). Some data providing the basis for such predictions are summarized here and on the facing page. In addition, L-dihydroxyphenylalanine (L-dopa; see p. 352), is used as an example to show the way in which molecules are illustrated in this book.

A. Molecule illustrations ◑

In traditional two-dimensional **structural formulas** (**A1**), atoms are represented as letter symbols and electron *pairs* are shown as lines. Lines between two atomic symbols symbolize two **bonding electrons** (see p. 4), and all of the other lines represent **free electron pairs**, such as those that occur in O and N atoms. Free electrons are usually not represented explicitly (and this is the convention used in this book as well). Dashed or continuous circles or arcs are used to emphasize delocalized electrons.

Ball-and-stick models (**A2**) are used to illustrate the spatial structure of molecules. Atoms are represented as colored balls (for the color coding, see the inside front cover) and bonds (including multiple bonds) as gray cylinders. Although the relative bond lengths and angles correspond to actual conditions, the size at which the atoms are represented is too small to make the model more comprehensible.

Space-filling **van der Waals models** (**A3**) are useful for illustrating the actual shape and size of molecules. These models represent atoms as truncated balls. Their effective extent is determined by what is known as the van der Waals radius. This is calculated from the energetically most favorable distance between atoms that are not chemically bonded to one another.

B. Bond lengths and angles ○

Atomic radii and distances are now usually expressed in picometers (pm; 1 pm = 10^{-12} m). The old angstrom unit (Å, Å = 100 pm) is now obsolete. The length of single bonds approximately corresponds to the sum of what are known as the **covalent radii** of the atoms involved (see inside front cover). Double bonds are around 10–20% shorter than single bonds. In sp^3-hybridized atoms, the angle between the individual bonds is approx. 110°; in sp^2-hybridized atoms it is approx. 120°.

C. Bond polarity ○

Depending on the position of the element in the periodic table (see p. 2), atoms have different **electronegativity**—i. e., a different tendency to take up extra electrons. The values given in **C2** are on a scale between 2 and 4. The higher the value, the more electronegative the atom. When two atoms with very different electronegativities are bound to one another, the bonding electrons are drawn toward the more electronegative atom, and the **bond** is **polarized**. The atoms involved then carry positive or negative partial charges. In **C1**, the van der Waals surface is colored according to the different charge conditions (red = negative, blue = positive). Oxygen is the most strongly electronegative of the biochemically important elements, with C=O double bonds being especially highly polar.

D. Hydrogen bonds ◑

The **hydrogen bond**, a special type of noncovalent bond, is extremely important in biochemistry. In this type of bond, hydrogen atoms of OH, NH, or SH groups (known as hydrogen bond **donors**) interact with free electrons of **acceptor** atoms (for example, O, N, or S). The bonding energies of hydrogen bonds (10–40 kJ mol^{-1}) are much lower than those of covalent bonds (approx. 400 kJ mol^{-1}). However, as hydrogen bonds can be very numerous in proteins and DNA, they play a key role in the stabilization of these molecules (see pp. 68, 84). The importance of hydrogen bonds for the properties of water is discussed on p. 26.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Molecule illustrations

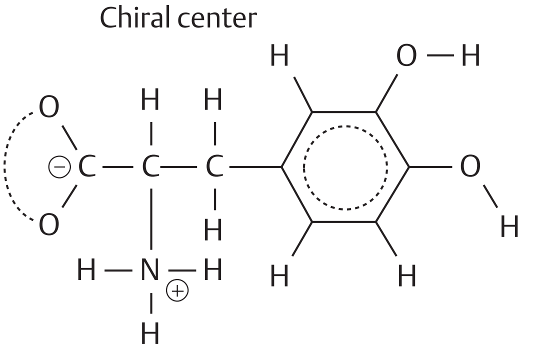
Chiral center
H
O — H
O
H H
⊖C — C — C
O
H
O
H
H — N — H
⊕
H
H
H

1. Formula illustration

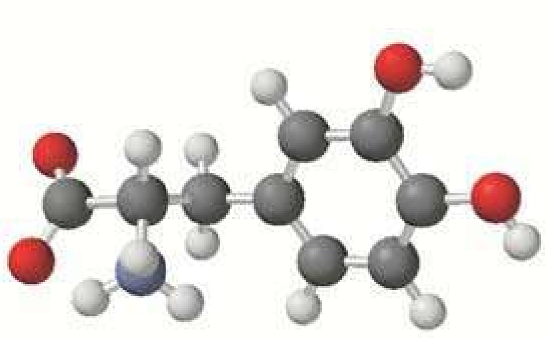

2. Ball- and-stick model

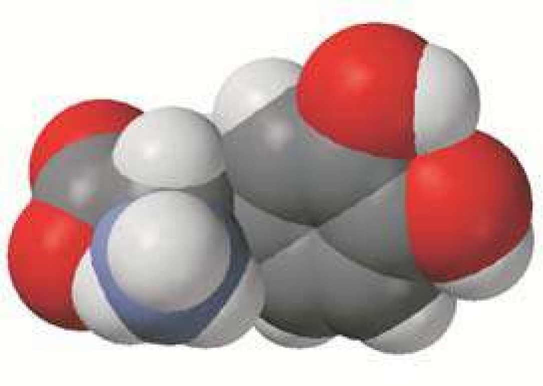

3. Van der Waals model

B. Bond lengths and angles

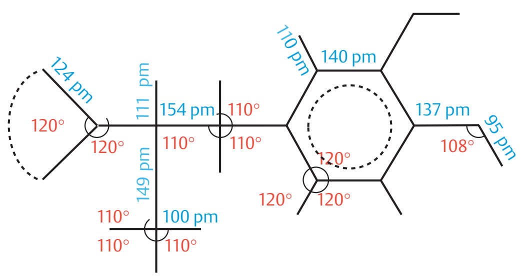
124 pm
120°
120°
111 pm
154 pm
110°
110°
110°
110°
149 pm
110°
110°
100 pm
110°
110°
110 pm
140 pm
137 pm
95 pm
108°
120°
120°
120°

C. Bond polarity

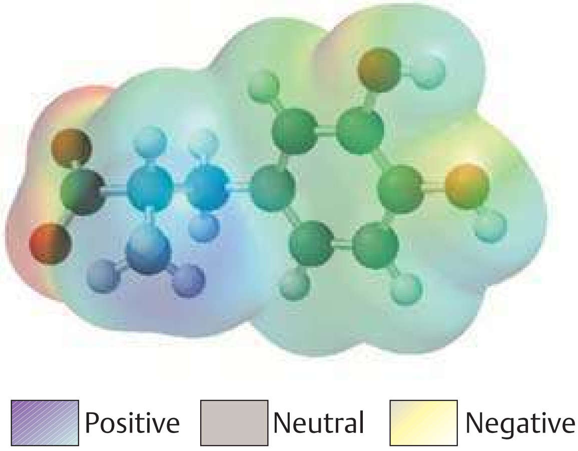
Positive
Neutral
Negative

1. Partial charges in L-dopa

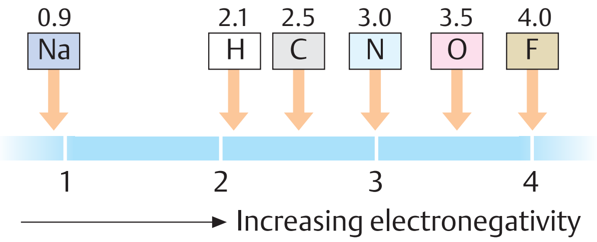
0.9
Na
2.1
H
2.5
C
3.0
N
3.5
O
4.0
F
1
2
3
4
Increasing electronegativity

2. Electronegativities

D. Hydrogen bonds

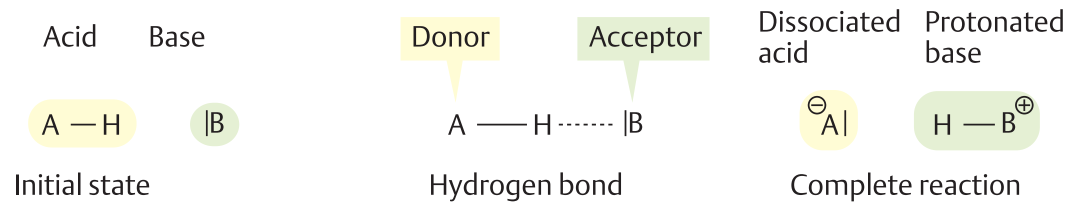
Acid
Base
A — H
|B
Initial state
Donor
Acceptor
A — H ······· |B
Hydrogen bond
Dissociated acid
Protonated base
⊖A|
H — B⊕
Complete reaction

1. Principle

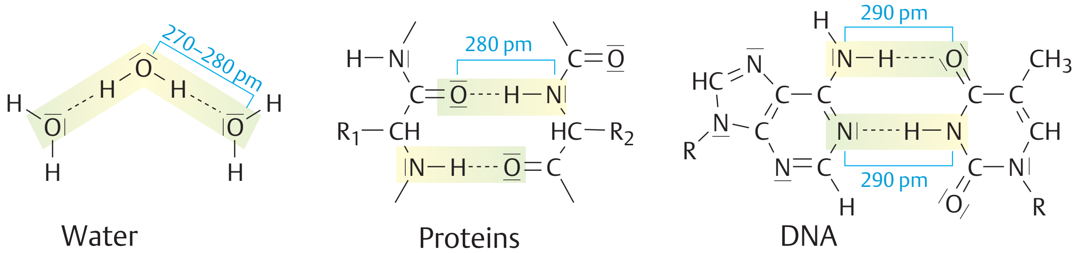
270–280 pm
H
O
H
Water
280 pm
H—N
C=O
R₁—CH
HC—R₂
Proteins
290 pm
290 pm
CH₃
R
DNA

2. Examples

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Isomerism

Isomers are molecules with the same composition (i. e. the same molecular formula), but with different chemical and physical properties. If isomers differ in the way in which their atoms are bonded in the molecule, they are described as **structural isomers** (cf. citric acid and isocitric acid, **D**). Other forms of isomerism are based on different arrangements of the substituents of bonds (**A**, **B**) or on the presence of chiral centers in the molecule (**C**).

A. *cis–trans* isomers ◑

Double bonds *are not freely rotatable* (see p. 4). If double-bonded atoms have different substituents, there are two possible orientations for these groups. In **fumaric acid**, an intermediate of the tricarboxylic acid cycle (see p. 136), the carboxy groups lie on *different* sides of the double bond (*trans* or *E* position). In its isomer **maleic acid**, which is not produced in metabolic processes, the carboxy groups lie on the *same* side of the bond (*cis* or *Z* position). *Cis–trans* isomers (**geometric isomers**) have different chemical and physical properties—e. g., their melting points (Fp.) and pK_a values. They can only be interconverted by chemical reactions.

In lipid metabolism, *cis–trans* isomerism is particularly important. For example, double bonds in natural fatty acids (see p. 48) usually have a *cis* configuration. By contrast, unsaturated intermediates of β oxidation have a *trans* configuration. This makes the breakdown of unsaturated fatty acids more complicated (see p. 166). Light-induced *cis–trans* isomerization of retinal is of central importance in the visual cycle (see p. 358).

B. Conformation ◑

Molecular forms that arise as a result of rotation around freely rotatable bonds are known as **conformers**. Even small molecules can have different conformations in solution. In the two conformations of **succinic acid** illustrated opposite, the atoms are arranged in a similar way to fumaric acid and maleic acid. Both forms are possible, although conformation 1 is more favorable due to the greater distance between the COOH groups and therefore occurs more frequently. Biologically active macromolecules such as proteins or nucleic acids usually have well-defined (“native”) conformations, which are stabilized by interactions in the molecule (see p. 74).

C. Optical isomers ◑

Another type of isomerism arises when a molecule contains a **chiral center** or is chiral as a whole. Chirality (from the Greek *cheir,* hand) leads to the appearance of structures that behave like image and mirror-image and that cannot be superimposed (“mirror” isomers). The most frequent cause of chiral behavior is the presence of an asymmetric C atom—i. e., an atom with four *different* substituents. Then there are two forms **(enantiomers)** with different **configurations.** Usually, the two enantiomers of a molecule are designated as L and D forms. Clear classification of the configuration is made possible by the *R/S system* (see chemistry textbooks).

Enantiomers have very similar chemical properties, but they rotate polarized light in opposite directions (**optical activity**, see pp. 36, 58). The same applies to the enantiomers of **lactic acid**. The dextrorotatory L-lactic acid occurs in animal muscle and blood, while the D form produced by microorganisms is found in milk products, for example (see p. 148). The Fischer projection is often used to represent the formulas for chiral centers (cf. p. 58).

D. The aconitase reaction ○

Enzymes usually function *stereospecifically.* In chiral substrates, they only accept one of the enantiomers, and the reaction products are usually also sterically uniform. *Aconitate hydratase* (aconitase) catalyzes the conversion of citric acid into the constitution isomer isocitric acid (see p. 136). Although citric acid is not chiral, aconitase only forms one of the four possible isomeric forms of isocitric acid (*2R,3S*-isocitric acid). The intermediate of the reaction, the unsaturated tricarboxylic acid *aconitate,* only occurs in the *cis* form in the reaction. The *trans* form of aconitate is found as a constituent of certain plants.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. *cis–trans* isomers

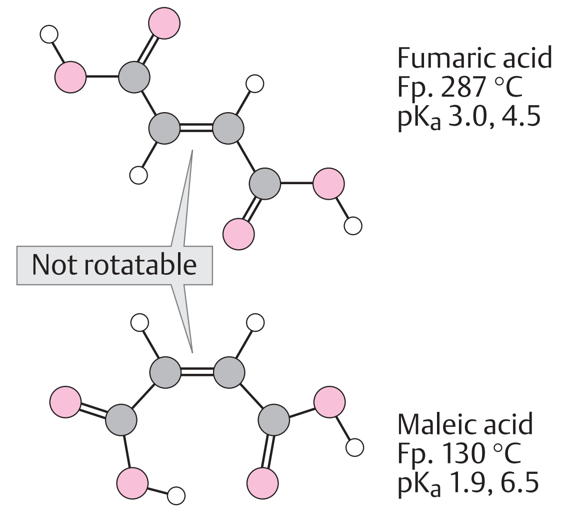
Fumaric acid
Fp. 287 °C
pKa 3.0, 4.5
Not rotatable
Maleic acid
Fp. 130 °C
pKa 1.9, 6.5

B. Conformers

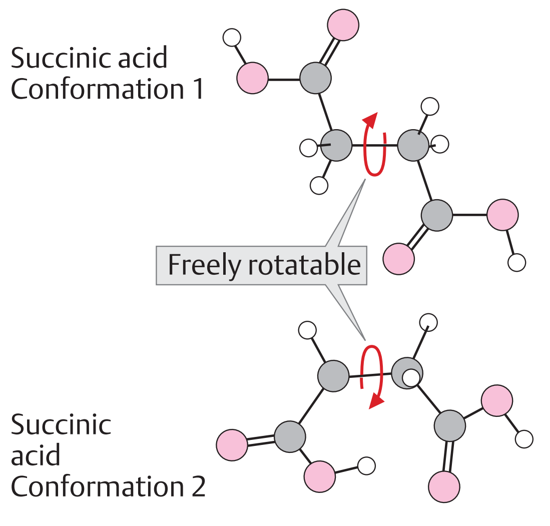
Succinic acid
Conformation 1
Freely rotatable
Succinic
acid
Conformation 2

C. Optical isomers

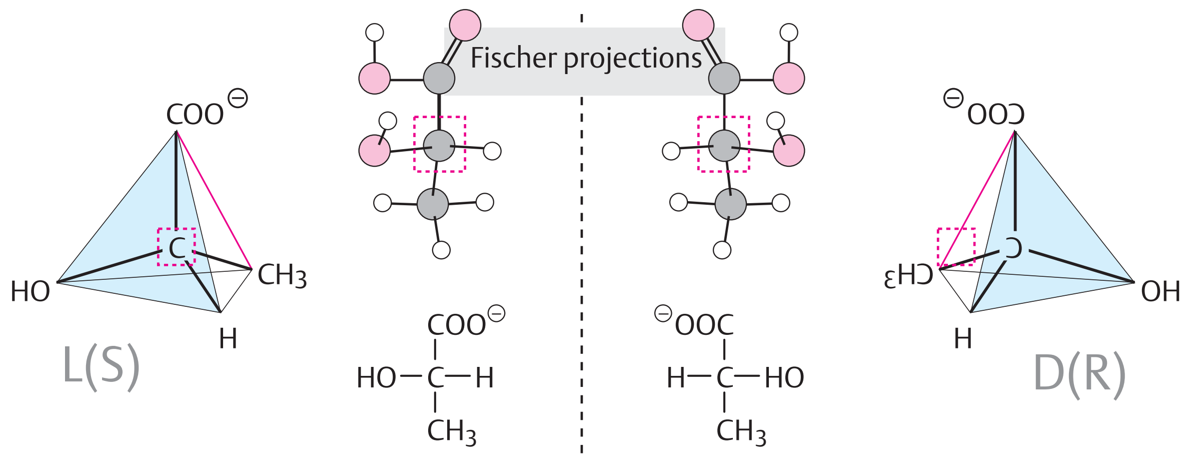
Fischer projections
COO⊖
C
HO
CH3
H
L(S)
COO⊖
HO—C—H
CH3
⊖OOC
H—C—HO
CH3
COO⊖
C
CH3
OH
H
D(R)

	L-lactic acid
Fp.	53 °C
pKa value	3.7
Specific rotation	+ 2.5˚

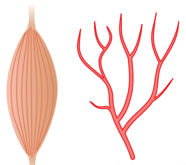
In muscle, blood

In milk products

D-lactic acid	
53 °C	Fp.
3.7	pKa value
-2.5˚	Specific rotation

D. The aconitase reaction

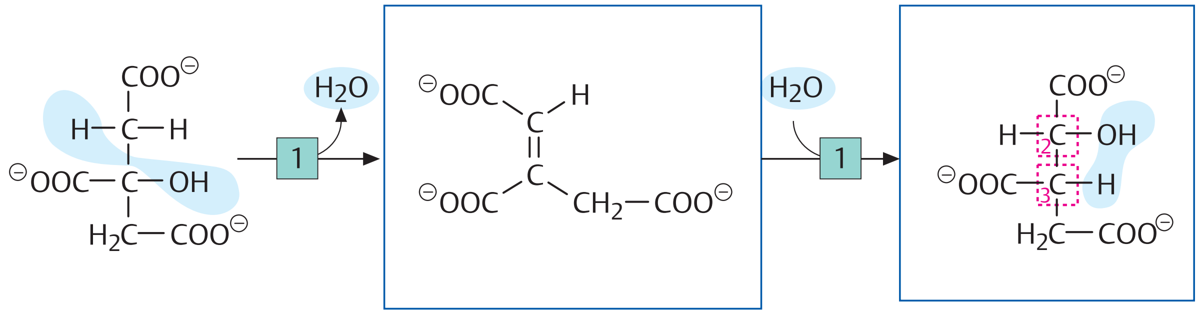
COO⊖
H—C—H
⊖OOC—C—OH
H2C—COO⊖
H2O
1
⊖OOC
H
C
C
⊖OOC
CH2—COO⊖
H2O
1
COO⊖
H—C—OH
2
⊖OOC—C—H
3
H2C—COO⊖

Citrate (prochiral)

cis-Aconitate (intermediate product)

(*2R,3S*)-Isocitrate

trans-Aconitate occurs in plants

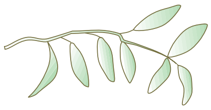

1 Aconitase *4.2.1.3*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Biomolecules I

A. Important classes of compounds ●

Most biomolecules are derivatives of simple compounds of the non-metals oxygen (O), hydrogen (H), nitrogen (N), sulfur (S), and phosphorus (P). The biochemically important oxygen, nitrogen, and sulfur compounds can be formally derived from their compounds with hydrogen (i. e., H_2O, NH_3, and H_2S). In biological systems, phosphorus is found almost exclusively in derivatives of phosphoric acid, H_3PO_4.

If one or more of the hydrogen atoms of a non-metal hydride are replaced formally with another group, R—e. g., alkyl residues—then derived compounds of the type $R\text{-}XH_{n-1}$, $R\text{-}XH_{n-2}\text{-}R$, etc., are obtained. In this way, **alcohols** (R-OH) and **ethers** (R-O-R) are derived from water (H_2O); primary **amines** ($R\text{-}NH_2$), secondary amines (R-NH-R) and tertiary amines (R-N-R′R″) amines are obtained from ammonia (NH_3); and **thiols** (R-SH) and **thioethers** (R-S-R′) arise from hydrogen sulfide (H_2S). Polar groups such as -OH and $\text{-}NH_2$ are found as substituents in many organic compounds. As such groups are much more reactive than the hydrocarbon structures to which they are attached, they are referred to as **functional groups**.

New functional groups can arise as a result of **oxidation** of the compounds mentioned above. For example, the oxidation of a thiol yields a **disulfide** (R-S-S-R). Double oxidation of a primary alcohol ($R\text{-}CH_2\text{-}OH$) gives rise initially to an **aldehyde** (R-C(O)-H), and then to a **carboxylic acid** (R-C(O)-OH). In contrast, the oxidation of a secondary alcohol yields a **ketone** (R-C(O)-R). The carbonyl group (C=O) is characteristic of aldehydes and ketones.

The addition of an amine to the carbonyl group of an aldehyde yields—after removal of water—an **aldimine** (not shown; see p. 178). Aldimines are intermediates in amino acid metabolism (see p. 178) and serve to bond aldehydes to amino groups in proteins (see p. 62, for example). The addition of an alcohol to the carbonyl group of an aldehyde yields a **hemiacetal** (R-O-C(H)OH-R). The cyclic forms of sugars are well-known examples of hemiacetals (see p. 36). The oxidation of hemiacetals produces carboxylic acid esters.

Very important compounds are the **carboxylic acids** and their derivatives, which can be formally obtained by exchanging the OH group for another group. In fact, derivatives of this type are formed by nucleophilic substitutions of activated intermediate compounds and the release of water (see p. 14). **Carboxylic acid esters** (R-O-CO-R′) arise from carboxylic acids and alcohols. This group includes the fats, for example (see p. 48). Similarly, a carboxylic acid and a thiol yield a **thioester** (R-S-CO-R′). Thioesters play an extremely important role in carboxylic acid metabolism. The best-known compound of this type is acetyl-coenzyme A (see p. 12).

Carboxylic acids and primary amines react to form **carboxylic acid amides** (R-NH-CO-R′). The amino acid constituents of peptides and proteins are linked by carboxylic acid amide bonds, which are therefore also known as peptide bonds (see p. 66).

Phosphoric acid, H_3PO_4, is a tribasic (three-protic) acid—i. e., it contains three hydroxyl groups able to donate H^+ ions. At least one of these three groups is fully dissociated under normal physiological conditions, while the other two can react with alcohols. The resulting products are phosphoric acid monoesters (R-O-P(O)O-OH) and diesters (R-O-P(O)O-O-R′). **Phosphoric acid monoesters** are found in carbohydrate metabolism, for example (see p. 36), whereas **phosphoric acid diester** bonds occur in phospholipids (see p. 50) and nucleic acids (see p. 82).

Compounds of one acid with another are referred to as **acid anhydrides**. A particularly large amount of energy is required for the formation of an acid—anhydride bond. Phosphoric anhydride bonds therefore play a central role in the storage and release of chemical energy in the cell (see p. 122). Mixed anhydrides between carboxylic acids and phosphoric acid are also very important “energy-rich metabolites” in cellular metabolism.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

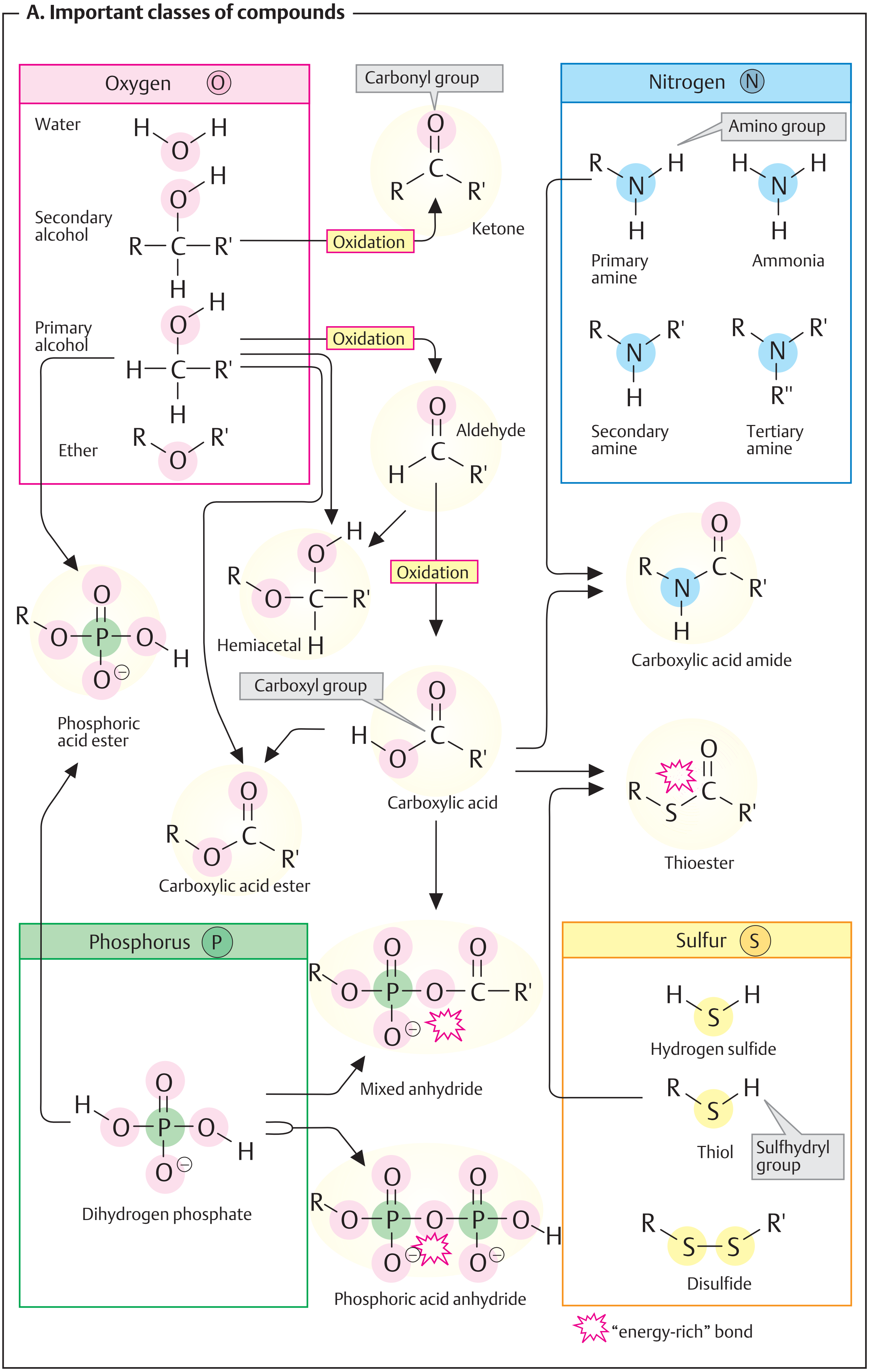
A. Important classes of compounds
Oxygen O
Water
Secondary alcohol
Primary alcohol
Ether
Carbonyl group
Ketone
Oxidation
Oxidation
Aldehyde
Oxidation
Hemiacetal
Carboxyl group
Carboxylic acid
Carboxylic acid ester
Phosphoric acid ester
Nitrogen N
Amino group
Primary amine
Ammonia
Secondary amine
Tertiary amine
Carboxylic acid amide
Thioester
Phosphorus P
Dihydrogen phosphate
Mixed anhydride
Phosphoric acid anhydride
Sulfur S
Hydrogen sulfide
Thiol
Sulfhydryl group
Disulfide
"energy-rich" bond

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Biomolecules II

Many biomolecules are made up of smaller units in a modular fashion, and they can be broken down into these units again. The construction of these molecules usually takes place through condensation reactions involving the removal of water. Conversely, their breakdown functions in a hydrolytic fashion—i. e., as a result of water uptake. The page opposite illustrates this modular principle using the example of an important coenzyme.

A. Acetyl CoA ◑

Coenzyme A (see also p. 106) is a nucleotide with a complex structure (see p. 80). It serves to activate residues of carboxylic acids (acyl residues). Bonding of the carboxy group of the carboxylic acid with the thiol group of the coenzyme creates a **thioester bond** (-S-CO-R; see p. 10) in which the **acyl residue** has a **high chemical potential**. It can therefore be transferred to other molecules in exergonic reactions. This fact plays an important role in lipid metabolism in particular (see pp. 162 ff.), as well as in two reactions of the tricarboxylic acid cycle (see p. 136).

As discussed on p. 16, the **group transfer potential** can be expressed quantitatively as the change in free enthalpy (ΔG) during hydrolysis of the compound concerned. This is an arbitrary determination, but it provides important indications of the chemical energy stored in such a group. In the case of acetyl-CoA, the reaction to be considered is:

$$\text{Acetyl CoA} + H_2O \rightarrow \text{acetate} + \text{CoA}$$

In standard conditions and at pH 7, the change in the chemical potential G (ΔG^0, see p. 18) in this reaction amounts to -32 kJ mol^{-1} and it is therefore as high as the ΔG^0 of ATP hydrolysis (see p. 18). In addition to the "energy-rich" **thioester bond**, acetyl-CoA also has seven other hydrolyzable bonds with different degrees of stability. These bonds, and the fragments that arise when they are hydrolyzed, will be discussed here in sequence.

(1) The reactive thiol group of coenzyme A is located in the part of the molecule that is derived from **cysteamine**. Cysteamine is a *biogenic amine* (see p. 62) formed by decarboxylation of the amino acid cysteine.

(2) The amino group of cysteamine is bound to the carboxy group of another biogenic amine via an **acid amide bond** (-CO-NH-). β-**Alanine** arises through decarboxylation of the amino acid aspartate, but it can also be formed by breakdown of pyrimidine bases (see p. 186).

(3) Another **acid amide bond** (-CO-NH-) creates the compound for the next constituent, **pantoinate**. This compound contains a *chiral center* and can therefore appear in two enantiomeric forms (see p. 8). In natural coenzyme A, only one of the two forms is found, the (*R*)-pantoinate. Human metabolism is not capable of producing pantoinate itself, and it therefore has to take up a compound of β-alanine and pantoinate—**pantothenate** ("pantothenic acid")—in the form of a vitamin in food (see p. 366).

(4) The hydroxy group at C-4 of pantoinate is bound to a **phosphate** residue by an **ester bond**.

The section of the molecule discussed so far represents a functional unit. In the cell, it is produced from pantothenate. The molecule also occurs in a protein-bound form as **4′-phosphopantetheine** in the enzyme *fatty acid synthase* (see p. 168). In coenzyme A, however, it is bound to 3′,5′-adenosine diphosphate.

(5) When two phosphate residues bond, they do not form an ester, but an "energy-rich" **phosphoric acid anhydride bond**, as also occurs in other nucleoside phosphates. By contrast, (6) and (7) are ester bonds again.

(8) The base **adenine** is bound to C-1 of **ribose** by an **N-glycosidic** bond (see p. 36). In addition to C-2 to C-4, C-1 of ribose also represents a *chiral* center. The *β-configuration* is usually found in nucleotides.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Acetyl CoA

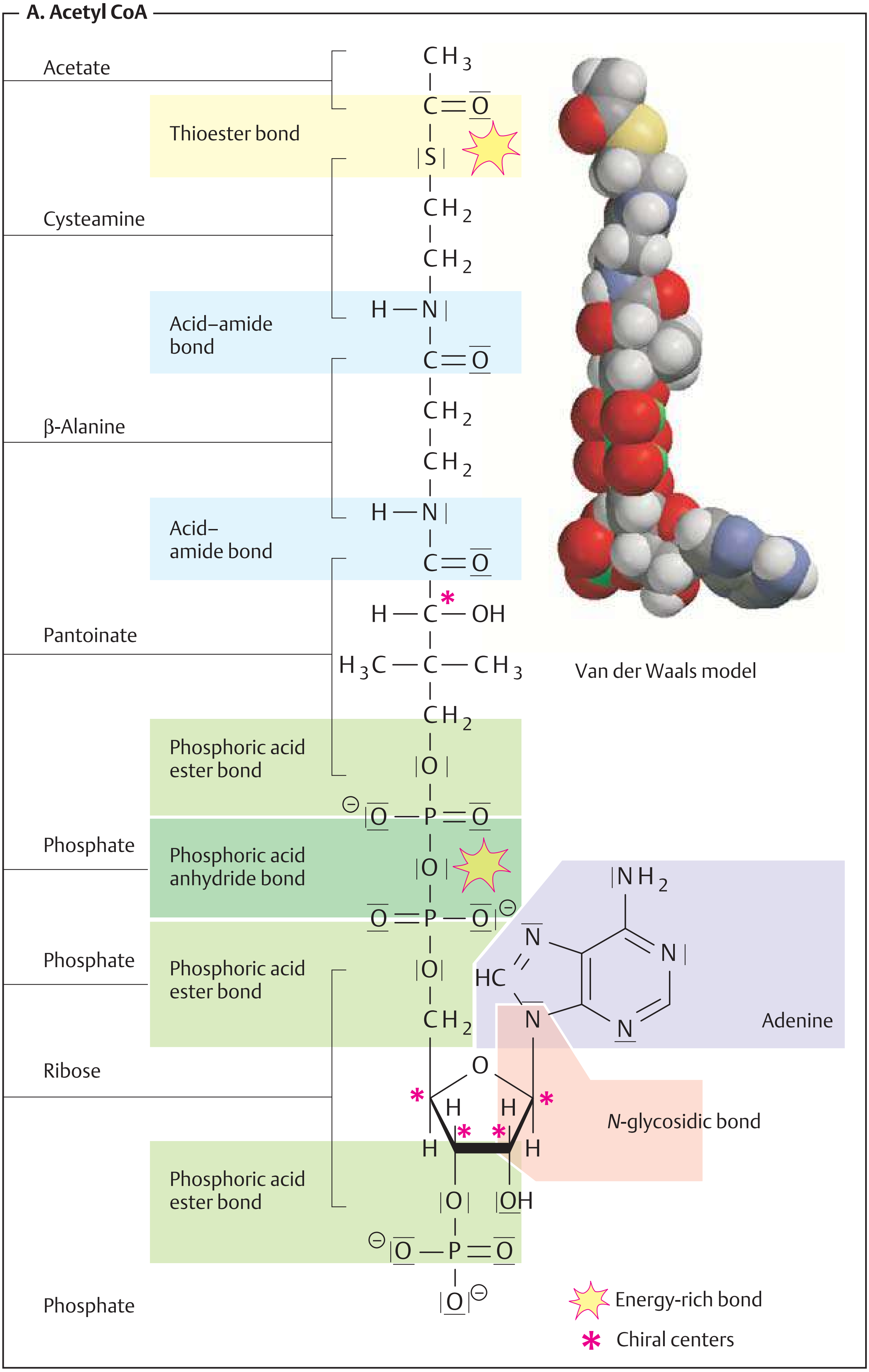
Acetate
Thioester bond
Cysteamine
Acid–amide bond
β-Alanine
Acid–amide bond
Pantoinate
Phosphoric acid ester bond
Phosphate
Phosphoric acid anhydride bond
Phosphate
Phosphoric acid ester bond
Ribose
Phosphoric acid ester bond
Phosphate
Van der Waals model
Adenine
N-glycosidic bond
Energy-rich bond
Chiral centers
CH3
C=O
S
CH2
CH2
H—N
C=O
CH2
CH2
H—N
C=O
H—C—OH
H3C—C—CH3
CH2
O
O—P=O
O
O=P—O
O
CH2
O
H H H H
O OH
O—P=O
O
NH2
N
N
HC
N
N

Chemical reactions

Chemical reactions are processes in which electrons or groups of atoms are taken up into molecules, exchanged between molecules, or shifted within molecules. Illustrated here are the most important types of reaction in organic chemistry, using simple examples. Electron shifts are indicated by red arrows.

A. Redox reactions ◑

In redox reactions (see also p. 32), **electrons** are **transferred** from one molecule (the reducing agent) to another (the oxidizing agent). One or two protons are often also transferred in the process, but the decisive criterion for the presence of a redox reaction is the electron transfer. The reducing agent is oxidized during the reaction, and the oxidizing agent is reduced.

Fig. **A** shows the oxidation of an alcohol into an aldehyde (**1**) and the reduction of the aldehyde to alcohol (**2**). In the process, one *hydride ion* is transferred (two electrons and one proton; see p. 32), which moves to the oxidizing agent A in reaction **1**. The superfluous proton is bound by the catalytic effect of a base B. In the reduction of the aldehyde (**2**), A-H serves as the reducing agent and the acid H-B is involved as the catalyst.

B. Acid–base reactions ◑

In contrast to redox reactions, only **proton transfer** takes place in acid–base reactions (see also p. 30). When an acid dissociates (**1**), water serves as a proton acceptor (i. e., as a base). Conversely, water has the function of an acid in the protonation of a carboxylate anion (**2**).

C. Additions/eliminations ◑

A reaction in which atoms or molecules are taken up by a multiple bond is described as **addition**. The converse of addition—i. e., the removal of groups with the formation of a double bond, is termed **elimination**. When water is added to an alkene (**1a**), a proton is first transferred to the alkene. The unstable carbenium cation that occurs as an intermediate initially takes up water (not shown), before the separation of a proton produces alcohol (**1b**). The elimination of water from the alcohol (**2**, dehydration) is also catalyzed by an acid and passes via the same intermediate as the addition reaction.

D. Nucleophilic substitutions ◑

A reaction in which one functional group (see p. 10) is replaced by another is termed **substitution**. Depending on the process involved, a distinction is made between nucleophilic and electrophilic substitution reactions (see chemistry textbooks). Nucleophilic substitutions start with the addition of one molecule to another, followed by elimination of the so-called *leaving group*.

The hydrolysis of an ester to alcohol and acid (**1**) and the esterification of a carboxylic acid with an alcohol (**2**) are shown here as an example of the S_N2 mechanism. Both reactions are made easier by the marked polarity of the C=O double bond. In the form of ester hydrolysis shown here, a proton is removed from a water molecule by the catalytic effect of the base B. The resulting strongly nucleophilic OH^- ion attacks the positively charged carbonyl C of the ester (**1a**), and an unstable sp^3-hybridized transition state is produced. From this, either water is eliminated (**2b**) and the ester re-forms, or the alcohol ROH is eliminated (**1b**) and the free acid results. In esterification (**2**), the same steps take place in reverse.

Further information

In **rearrangements** (isomerizations, not shown), groups are shifted within one and the same molecule. Examples of this in biochemistry include the isomerization of sugar phosphates (see p. 36) and of methylmalonyl-CoA to succinyl CoA (see p. 166).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Redox reactions

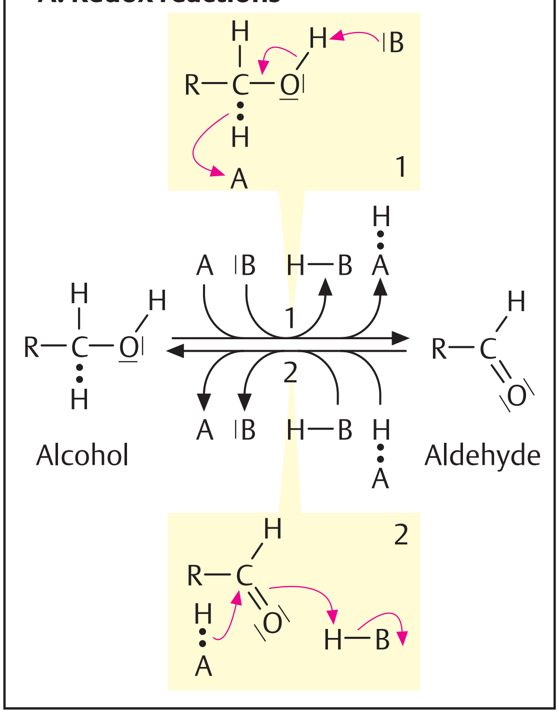
1
2
A
B
H—B
Alcohol
Aldehyde

B. Acid–base reactions

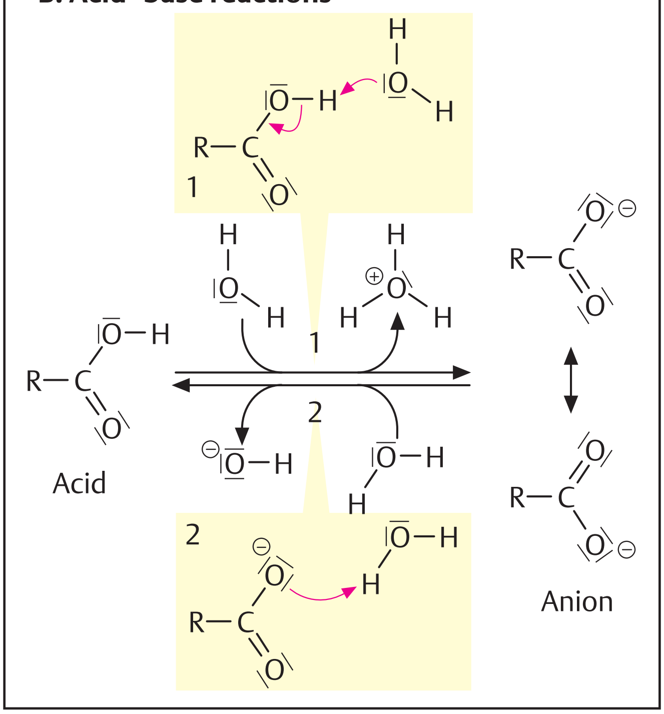
1
2
Acid
Anion

C. Additions/eliminations

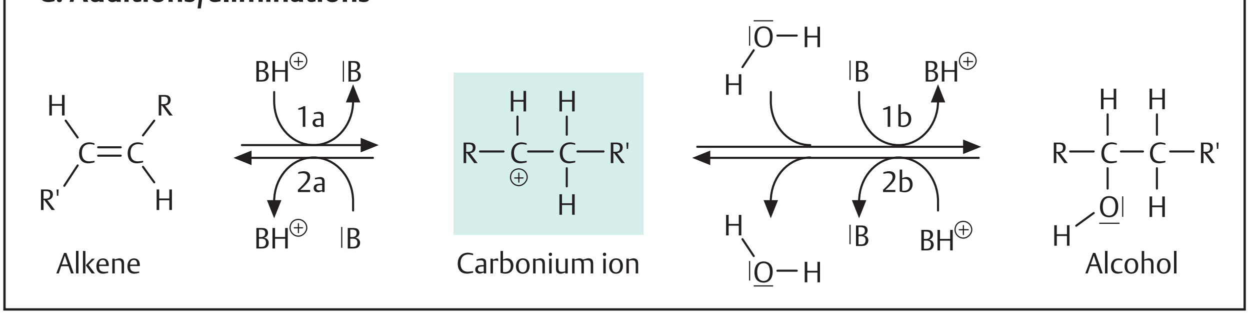
Alkene
1a
2a
BH⊕
B
Carbonium ion
1b
2b
Alcohol

D. Nucleophilic substitutions

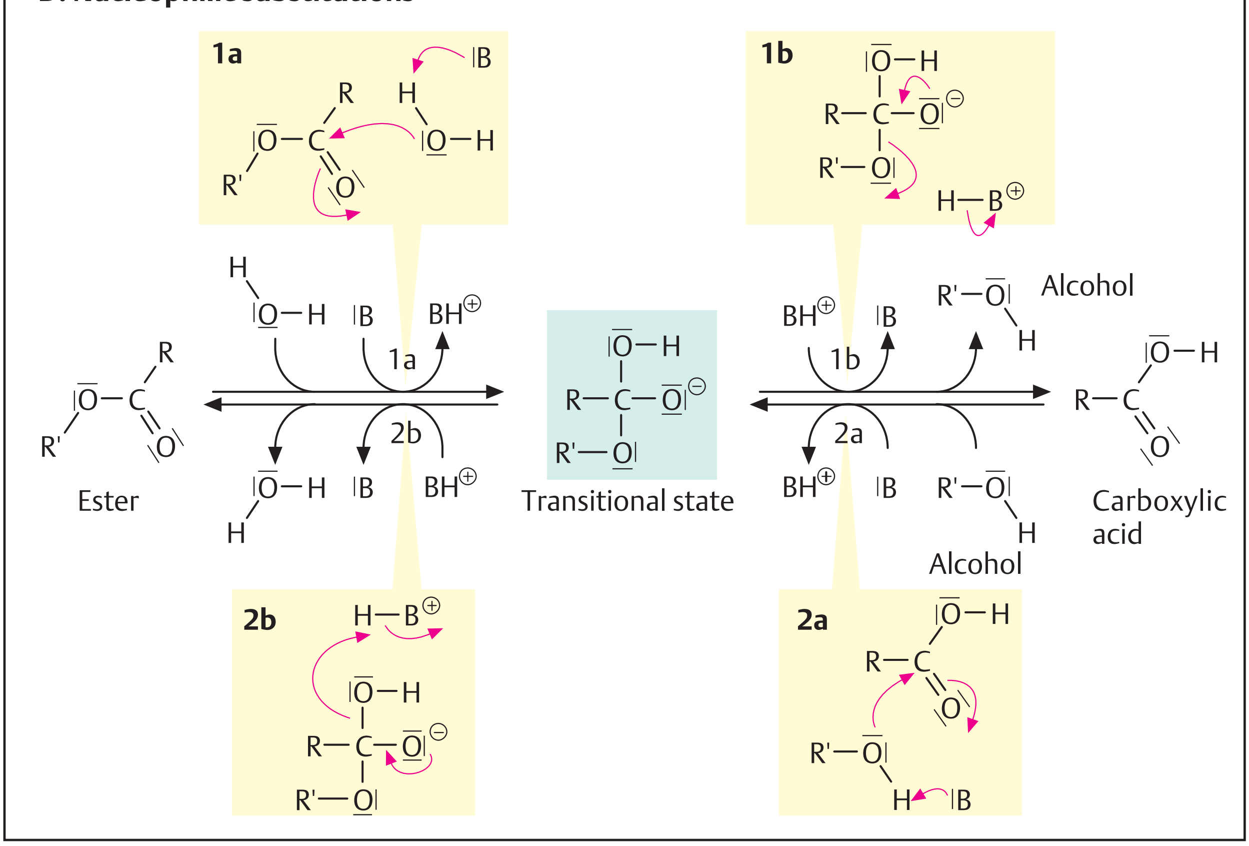
1a
1b
2a
2b
Ester
Transitional state
Alcohol
Carboxylic acid
BH⊕
B

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Energetics

To obtain a better understanding of the processes involved in energy storage and conversion in living cells, it may be useful first to recall the physical basis for these processes.

A. Forms of work ●

There is essentially no difference between work and energy. Both are measured in **joule** (J = 1 N m). An outdated unit is the **calorie** (1 cal = 4.187 J). **Energy is defined as the ability of a system to perform work.** There are many different forms of energy—e. g., mechanical, chemical, and radiation energy.

A system is capable of performing work when matter is moving along a potential gradient. This abstract definition is best understood by an example involving mechanical work (**A1**). Due to the earth's gravitational pull, the mechanical potential energy of an object is the greater the further the object is away from the center of the earth. A **potential difference** (ΔP) therefore exists between a higher location and a lower one. In a waterfall, the water spontaneously follows this potential gradient and, in doing so, is able to perform work—e. g., turning a mill.

Work and energy consist of two quantities: an **intensity** factor, which is a measure of the potential difference—i. e., the "driving force" of the process—(here it is the height difference) and a **capacity factor**, which is a measure of the quantity of the substance being transported (here it is the weight of the water). In the case of electrical work (**A2**), the intensity factor is the voltage—i. e., the electrical potential difference between the source of the electrical current and the "ground," while the capacity factor is the amount of charge that is flowing.

Chemical work and chemical energy are defined in an analogous way. The intensity factor here is the **chemical potential** of a molecule or combination of molecules. This is stated as **free enthalpy** *G* (also known as "Gibbs free energy"). When molecules spontaneously react with one another, the result is products at lower potential. The difference in the chemical potentials of the educts and products (the **change in free enthalpy**, **ΔG**) is a measure of the "driving force" of the reaction. The capacity factor in chemical work is the amount of matter reacting (in mol). Although absolute values for free enthalpy G cannot be determined, ΔG can be calculated from the equilibrium constant of the reaction (see p. 18).

B. Energetics and the course of processes ●

Everyday experience shows that water never flows uphill *spontaneously*. Whether a particular process can occur spontaneously or not depends on whether the potential difference between the final and the initial state, $\Delta P = P_2 - P_1$, is positive or negative. If P_2 is smaller than P_1, then ΔP will be negative, and the process will take place and perform work. Processes of this type are called **exergonic** (**B1**). If there is no potential difference, then the system is in **equilibrium** (**B2**). In the case of **endergonic** processes, ΔP is positive (**B3**). Processes of this type do *not* proceed spontaneously.

Forcing endergonic processes to take place requires the use of the principle of **energetic coupling**. This effect can be illustrated by a mechanical analogy (**B4**). When two masses M_1 and M_2 are connected by a rope, M_1 will move upward even though this part of the process is endergonic. The *sum* of the two potential differences ($\Delta P_{eff} = \Delta P_1 + \Delta P_2$) is the determining factor in coupled processes. When ΔP_{eff} is negative, the entire process can proceed.

Energetic coupling makes it possible to convert different forms of work and energy into one another. For example, in a flashlight, an exergonic chemical reaction provides an electrical voltage that can then be used for the endergonic generation of light energy. In the luminescent organs of various animals, it is a chemical reaction that produces the light. In the musculature (see p. 336), chemical energy is converted into mechanical work and heat energy. A form of storage for chemical energy that is used in all forms of life is **adenosine triphosphate** (ATP; see p. 122). Endergonic processes are usually driven by coupling to the strongly exergonic breakdown of ATP (see p. 122).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Forms of work

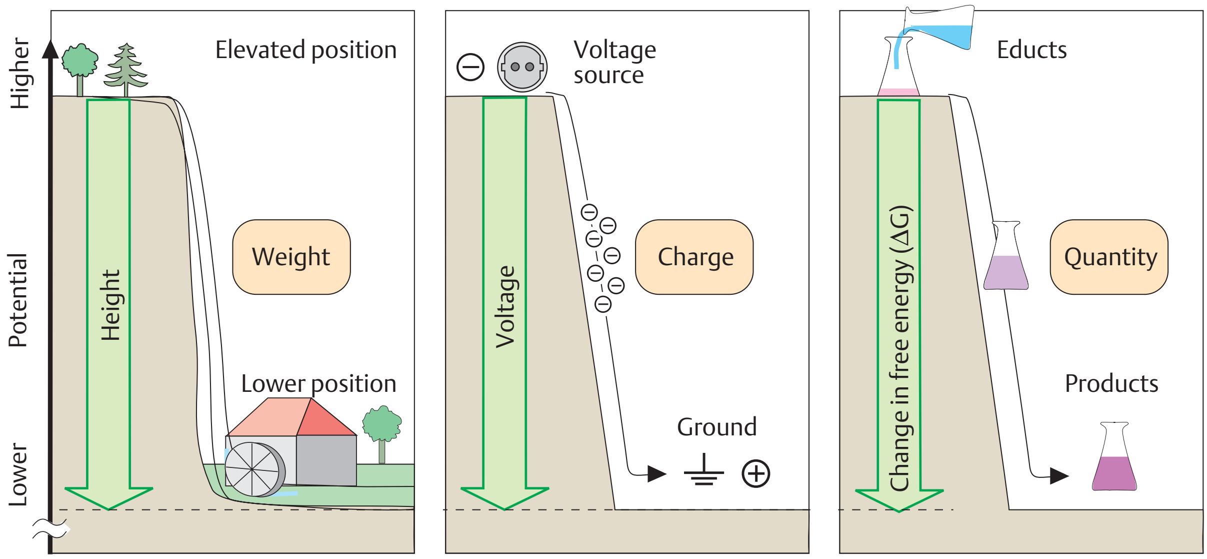
Higher
Potential
Lower
Elevated position
Height
Weight
Lower position
Voltage source
Voltage
Charge
Ground
Educts
Change in free energy (ΔG)
Quantity
Products

1. Mechanical work
2. Electrical work
3. Chemical work

$J = \text{Joule} = N \cdot m = 1\ kg \cdot m^2 \cdot s^{-2}$, 1 cal = 4.187 J

Form of work	Intensity factor	Unit	Capacity factor	Unit	Work = ⇨ · 🟧	Unit
Mechanical	Height	m	Weight	$J \cdot m^{-1}$	Height · Weight	J
Electrical	Voltage	$V = J \cdot C^{-1}$	Charge	C	Voltage · Charge	J
Chemical	Free-enthalpy change ΔG	$J \cdot mol^{-1}$	Quantity	mol	ΔG · Quantity	J

B. Energetics and the course of processes

$\Delta P < 0$	$\Delta P = 0$	$\Delta P > 0$	$\Delta P_{eff} < 0$
Process occurs spontaneously		Process cannot occur	Coupled processes can occur spontaneously

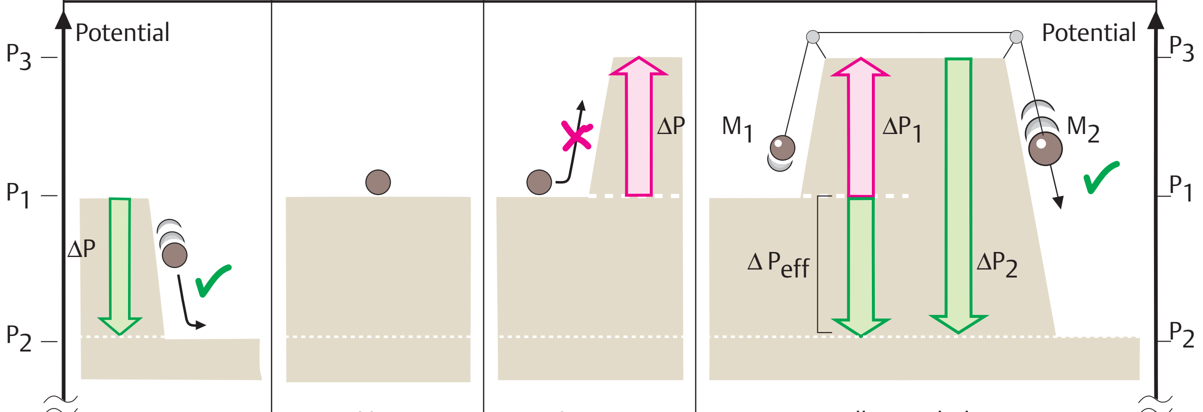
Potential
P_3
P_1
P_2
ΔP
ΔP
M_1
ΔP_1
M_2
ΔP_{eff}
ΔP_2
Potential
P_3
P_1
P_2

1. Exergonic
2. Equilibrium
3. Endergonic
4. Energetically coupled

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Equilibriums

A. Group transfer reactions ◑

Every chemical reaction reaches after a time a **state of equilibrium** in which the forward and back reactions proceed at the same speed. The **law of mass action** describes the concentrations of the educts (A, B) and products (C, D) *in equilibrium.* The **equilibrium constant K** is directly related to ΔG^0, the change in free enthalpy *G* involved in the reaction (see p. 16) under standard conditions ($\Delta G^0 = - R \quad T \quad \ln K$). For any given concentrations, the lower equation applies. At $\Delta G < 0$, the reaction proceeds spontaneously for as long as it takes for equilibrium to be reached (i. e., until $\Delta G = 0$). At $\Delta G > 0$, a *spontaneous* reaction is no longer possible (endergonic case; see p. 16). In biochemistry, ΔG is usually related to pH 7, and this is indicated by the “prime” symbol ($\Delta G^{0\prime}$ or $\Delta G^{\prime}$).

As examples, we can look at two group transfer reactions (on the right). In ATP (see p. 122), the terminal phosphate residue is at a high chemical potential. Its transfer to water (reaction **a**, below) is therefore strongly **exergonic**. The equilibrium of the reaction ($\Delta G = 0$; see p. 122) is only reached when more than 99.9% of the originally available ATP has been hydrolyzed. ATP and similar compounds have a high **group transfer potential** for phosphate residues. Quantitatively, this is expressed as the **ΔG of hydrolysis** ($\Delta G^{0\prime} = -32$ kJ mol^{-1}; see p. 122).

In contrast, the **endergonic** transfer of ammonia (NH_3) to glutamate (Glu, reaction **b**, $\Delta G^{0\prime} = +14$ kJ mol^{-1}) reaches equilibrium so quickly that only minimal amounts of the product glutamine (Gln) can be formed in this way. The synthesis of glutamine from these preliminary stages is only possible through **energetic coupling** (see pp. 16, 124).

B. Redox reactions ◑

The course of electron transfer reactions (redox reactions, see p. 14) also follows the law of mass action. For a single redox system (see p. 32), the Nernst equation applies (top). The **electron transfer potential** of a redox system (i. e., its tendency to give off or take up electrons) is given by its **redox potential E** (in standard conditions, E^0 or $E^{0\prime}$). The *lower* the redox potential of a system is, the *higher* the chemical potential of the transferred electrons. To describe reactions between two redox systems, ΔE—the difference between the two systems’ redox potentials—is usually used instead of ΔG. ΔG and ΔE have a simple relationship, but opposite signs (below). A redox reaction proceeds spontaneously when $\Delta E > 0$, i. e. $\Delta G < 0$.

The right side of the illustration shows the way in which the redox potential E is dependent on the composition (the proportion of the reduced form as a %) in two biochemically important redox systems (pyruvate/lactate and $NAD^+/NADH+H^+$; see pp. 98, 104). In the standard state (both systems reduced to 50%), electron transfer from lactate to NAD^+ is *not* possible, because ΔE is negative ($\Delta E = -0.13$ V, red arrow). By contrast, transfer can proceed successfully if the pyruvate/lactate system is reduced to 98% and $NAD^+/NADH$ is 98% oxidized (green arrow, $\Delta E = +0.08$ V).

C. Acid–base reactions ◑

Pairs of *conjugated* acids and bases are always involved in proton exchange reactions (see p. 30). The dissociation state of an acid–base pair depends on the H^+ concentration. Usually, it is not this concentration itself that is expressed, but its negative decadic logarithm, the **pH value**. The connection between the pH value and the dissociation state is described by the *Henderson–Hasselbalch equation* (below). As a measure of the **proton transfer potential** of an acid–base pair, its **pK_a value** is used—the negative logarithm of the acid constant K_a (where “a” stands for acid).

The *stronger* an acid is, the *lower* its pK_a value. The acid of the pair with the lower pK_a value (the stronger acid—in this case acetic acid, CH_3COOH) can protonate (green arrow) the base of the pair with the higher pK_a (in this case NH_3), while ammonium acetate (NH_4^+ and CH_3COO^-) only forms very little CH_3COOH and NH_3.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Group transfer reactions

Reaction $A + B \rightleftarrows C + D$

Law of mass action $K = \dfrac{[C] \cdot [D]}{[A] \cdot [B]}$

Only applies in chemical equilibrium

Equilibrium constant

Relationship between ΔG^0 and K

$\Delta G^\circ = -R \cdot T \cdot \ln K$

$R = 8.314\ J \cdot mol^{-1} \cdot K^{-1}$

In any conditions

$\Delta G = \Delta G^\circ + R \cdot T \cdot \ln \dfrac{[C] \cdot [D]}{[A] \cdot [B]}$

Measure of group transfer potential

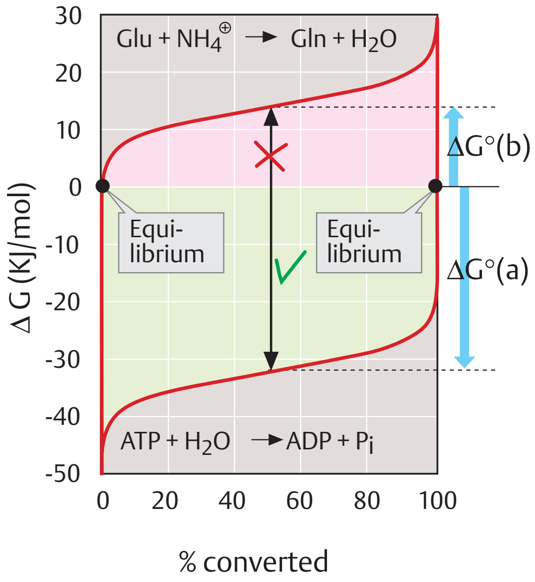

B. Redox reactions

$A_{red} \rightleftarrows A_{ox}$

For a redox system $E = E^\circ + \dfrac{R \cdot T}{n \cdot F} \cdot \ln \dfrac{[A_{ox}]}{[A_{red}]}$

Measure of electron transfer potential

For any redox reaction $\Delta E = \Delta E^\circ + \dfrac{R \cdot T}{n \cdot F} \cdot \ln \dfrac{[B_{ox}] \cdot [A_{red}]}{[B_{red}] \cdot [A_{ox}]}$

Definition and sizes

$\Delta E = E_{Acceptor} - E_{Donor}$

$\Delta G = -n \cdot F \cdot \Delta E$

n = No. of electrons transferred
F = Faraday constant

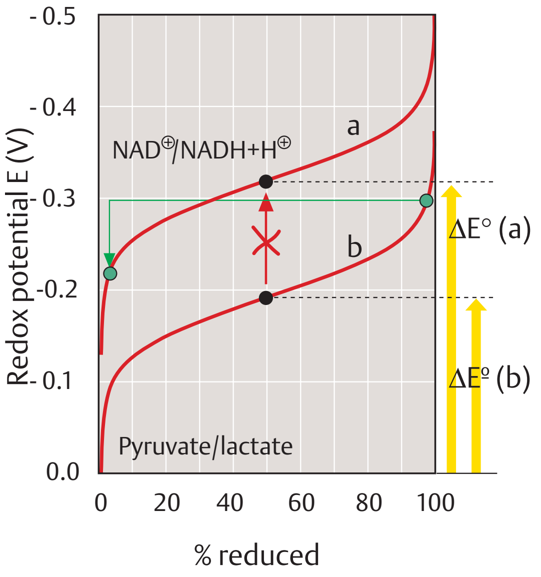

C. Acid–base reactions

Standard reaction $HA + H_2O \rightleftarrows A^\ominus + H_3O^\oplus$

Law of mass action $K = \dfrac{[A^\ominus] \cdot [H_3O^\oplus]}{[HA] \cdot [H_2O]}$

Simplified $K_a = \dfrac{[A^\ominus] \cdot [H^\oplus]}{[HA]}$

Henderson–Hasselbalch equation $pH = pK_a + \log \dfrac{[A^\ominus]}{[HA]}$

Measure of proton transfer potential

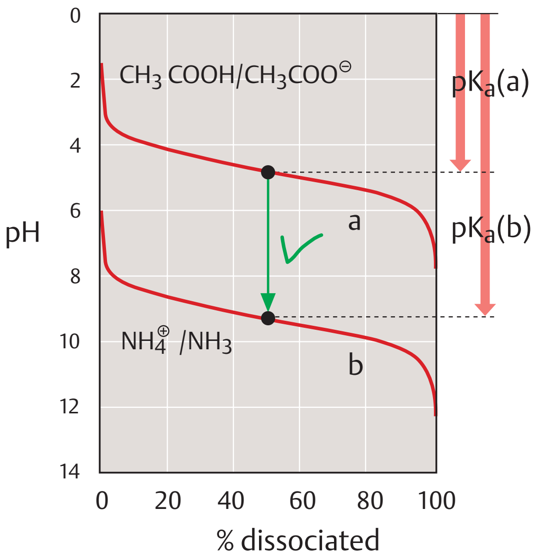

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Enthalpy and entropy

The change in the free enthalpy of a chemical reaction (i. e., its ΔG) depends on a number of factors—e. g., the concentrations of the reactants and the temperature (see p. 18). Two further factors associated with molecular changes occurring during the reaction are discussed here.

A. Heat of reaction and calorimetry ◑

All chemical reactions involve heat exchange. Reactions that release heat are called **exothermic**, and those that consume heat are called **endothermic**. Heat exchange is measured as the enthalpy change ΔH (the heat of reaction). This corresponds to the heat exchange at constant pressure. In exothermic reactions, the system *loses* heat, and ΔH is negative. When the reaction is endothermic, the system gains heat, and ΔH becomes positive.

In many reactions, ΔH and ΔG are similar in magnitude (see **B1**, for example). This fact is used to estimate the caloric content of foods. In living organisms, nutrients are usually oxidized by oxygen to CO_2 and H_2O (see p. 112). The maximum amount of chemical work supplied by a particular foodstuff (i. e., the ΔG for the oxidation of the utilizable constituents) can be estimated by burning a weighed amount in a **calorimeter** in an oxygen atmosphere. The heat of the reaction increases the water temperature in the calorimeter. The reaction heat can then be calculated from the temperature difference ΔT.

B. Enthalpy and entropy ◑

The reaction enthalpy ΔH and the change in free enthalpy ΔG are not always of the same magnitude. There are even reactions that occur spontaneously ($\Delta G < 0$) even though they are endothermic ($\Delta H > 0$). The reason for this is that changes in the degree of order of the system also strongly affect the progress of a reaction. This change is measured as the **entropy change (ΔS)**.

Entropy is a physical value that describes the **degree of order of a system**. The *lower* the degree of order, the larger the entropy. Thus, when a process leads to increase in disorder—and everyday experience shows that this is the normal state of affairs—ΔS is positive for this process. An increase in the order in a system ($\Delta S < 0$) always requires an input of energy. Both of these statements are consequences of an important natural law, the Second Law of Thermodynamics. The connection between changes in enthalpy and entropy is described quantitatively by the **Gibbs–Helmholtz equation** ($\Delta G = \Delta H - T \ \Delta S$). The following examples will help explain these relationships.

In the *knall-gas* **(oxyhydrogen) reaction** (**1**), gaseous oxygen and gaseous hydrogen react to form liquid water. Like many redox reactions, this reaction is strongly exothermic (i. e., $\Delta H < 0$). However, during the reaction, the degree of order increases. The total number of molecules is reduced by one-third, and a more highly ordered liquid is formed from freely moving gas molecules. As a result of the increase in the degree of order ($\Delta S < 0$), the term $-T \ \Delta S$ becomes positive. However, this is more than compensated for by the decrease in enthalpy, and the reaction is still strongly exergonic ($\Delta G < 0$).

The **dissolution of salt in water** (**2**) is endothermic ($\Delta H > 0$)—i. e., the liquid cools. Nevertheless, the process still occurs spontaneously, since the degree of order in the system *decreases*. The Na^+ and Cl^- ions are initially rigidly fixed in a crystal lattice. In solution, they move about independently and in random directions through the fluid. The decrease in order ($\Delta S > 0$) leads to a negative $-T \ \Delta S$ term, which compensates for the positive ΔH term and results in a negative ΔG term overall. Processes of this type are described as being **entropy-driven**. The folding of proteins (see p. 74) and the formation of ordered lipid structures in water (see p. 28) are also mainly entropy-driven.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

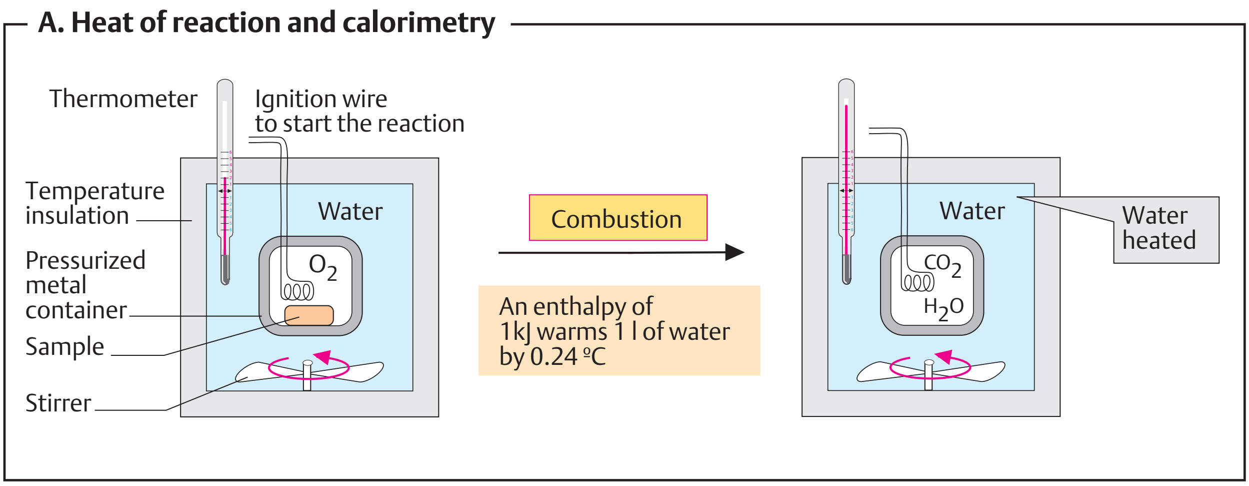
A. Heat of reaction and calorimetry
Thermometer
Ignition wire to start the reaction
Temperature insulation
Pressurized metal container
Sample
Stirrer
Water
O_2
Combustion
An enthalpy of 1 kJ warms 1 l of water by 0.24 ºC
Water
CO_2
H_2O
Water heated

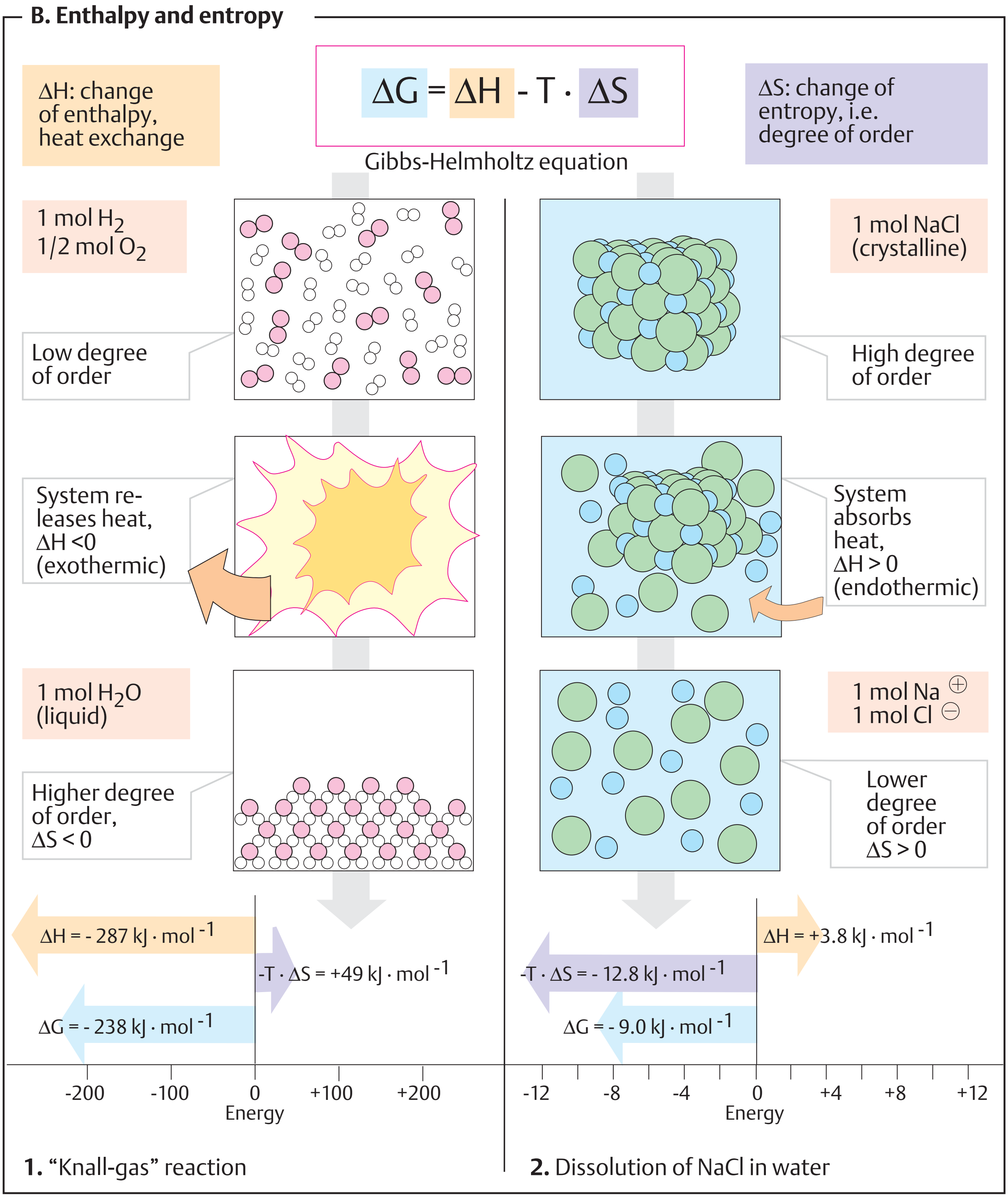
B. Enthalpy and entropy
ΔH: change of enthalpy, heat exchange
$\Delta G = \Delta H - T \cdot \Delta S$
Gibbs-Helmholtz equation
ΔS: change of entropy, i.e. degree of order
1 mol H_2
1/2 mol O_2
Low degree of order
System releases heat, ΔH <0 (exothermic)
1 mol H_2O (liquid)
Higher degree of order, ΔS < 0
$\Delta H = -287\ kJ \cdot mol^{-1}$
$-T \cdot \Delta S = +49\ kJ \cdot mol^{-1}$
$\Delta G = -238\ kJ \cdot mol^{-1}$
-200
-100
0
+100
+200
Energy
1. “Knall-gas” reaction
1 mol NaCl (crystalline)
High degree of order
System absorbs heat, ΔH > 0 (endothermic)
1 mol $Na^{\oplus}$
1 mol $Cl^{\ominus}$
Lower degree of order ΔS > 0
$\Delta H = +3.8\ kJ \cdot mol^{-1}$
$-T \cdot \Delta S = -12.8\ kJ \cdot mol^{-1}$
$\Delta G = -9.0\ kJ \cdot mol^{-1}$
-12
-8
-4
0
+4
+8
+12
Energy
2. Dissolution of NaCl in water

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Reaction kinetics

The change in free enthalpy ΔG in a reaction indicates whether or not the reaction can take place spontaneously in given conditions and how much work it can perform (see p. 18). However, it does not tell us anything about the *rate* of the reaction—i. e., its **kinetics**.

A. Activation energy ◑

Most organic chemical reactions (with the exception of acid–base reactions) proceed only very slowly, regardless of the value of ΔG. The reason for the slow reaction rate is that the molecules that react—the educts—have to have a certain minimum energy before they can enter the reaction. This is best understood with the help of an energy diagram (**1**) of the simplest possible reaction $A \rightarrow B$. The educt A and the product B are each at a specific **chemical potential** (G_e and G_p, respectively). The change in the free enthalpy of the reaction, ΔG, corresponds to the difference between these two potentials. To be converted into B, A first has to overcome a potential energy barrier, the peak of which, G_a, lies well above G_e. The potential difference $G_a - G_e$ is the **activation energy E_a** of the reaction (in kJ mol^{-1}).

The fact that A can be converted into B at all is because the potential G_e only represents the average potential of all the molecules. Individual molecules may occasionally reach much higher potentials—e. g., due to collisions with other molecules. When the increase in energy thus gained is greater than E_a, these molecules can overcome the barrier and be converted into B. The energy distribution for a group of molecules of this type, as calculated from a simple model, is shown in (**2**) and (**3**). $\Delta n/n$ is the fraction of molecules that have reached or exceeded energy E (in kJ per mol). At 27 °C, for example, approximately 10% of the molecules have energies > 6 kJ mol^{-1}. The typical activation energies of chemical reactions are much higher. The course of the energy function at energies of around 50 kJ mol^{-1} is shown in (**3**). Statistically, at 27 °C only two out of 10^9 molecules reach this energy. At 37 °C, the figure is already four. This is the basis for the long-familiar "Q_{10} law"—a rule of thumb that states that the speed of biological processes approximately doubles with an increase in temperature of 10 °C.

B. Reaction rate ◑

The velocity v of a chemical reaction is determined experimentally by observing the change in the concentration of an educt or product over time. In the example shown (again a reaction of the $A \rightarrow B$ type), 3 mmol of the educt A is converted per second and 3 mmol of the product B is formed per second in one liter of the solution. This corresponds to a rate of

$$v = 3\ \text{mM}\quad \text{s}^{-1} = 3\quad 10^{-3}\ \text{mol}\quad \text{L}^{-1}\quad \text{s}^{-1}$$

C. Reaction order ◑

Reaction rates are influenced not only by the activation energy and the temperature, but also by the concentrations of the reactants. When there is only one educt, A (**1**), v is proportional to the concentration [A] of this substance, and a **first-order reaction** is involved. When *two* educts, A and B, react with one another (**2**), it is a **second order reaction** (shown on the right). In this case, the rate v is proportional to the *product* of the educt concentrations (12 mM^2 at the top, 24 mM^2 in the middle, and 36 mM^2 at the bottom). The proportionality factors k and k′ are the **rate constants** of the reaction. They are *not* dependent on the reaction concentrations, but depend on the external conditions for the reaction, such as temperature.

In **B**, only the kinetics of simple irreversible reactions is shown. More complicated cases, such as reaction with three or more reversible steps, can usually be broken down into first-order or second-order partial reactions and described using the corresponding equations (for an example, see the Michaelis–Menten reaction, p. 92).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Activation energy

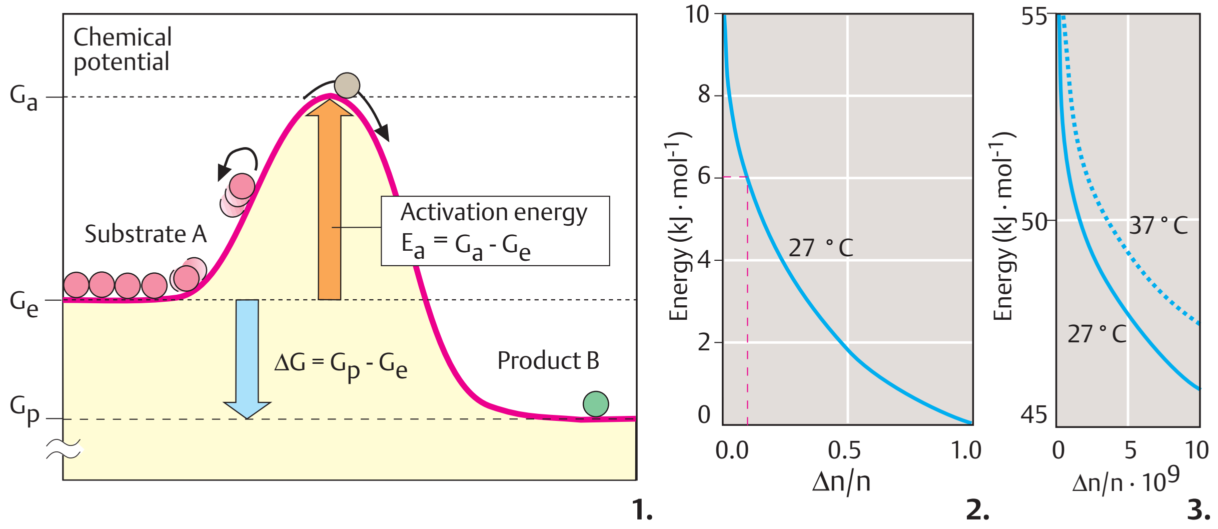

B. Reaction rate

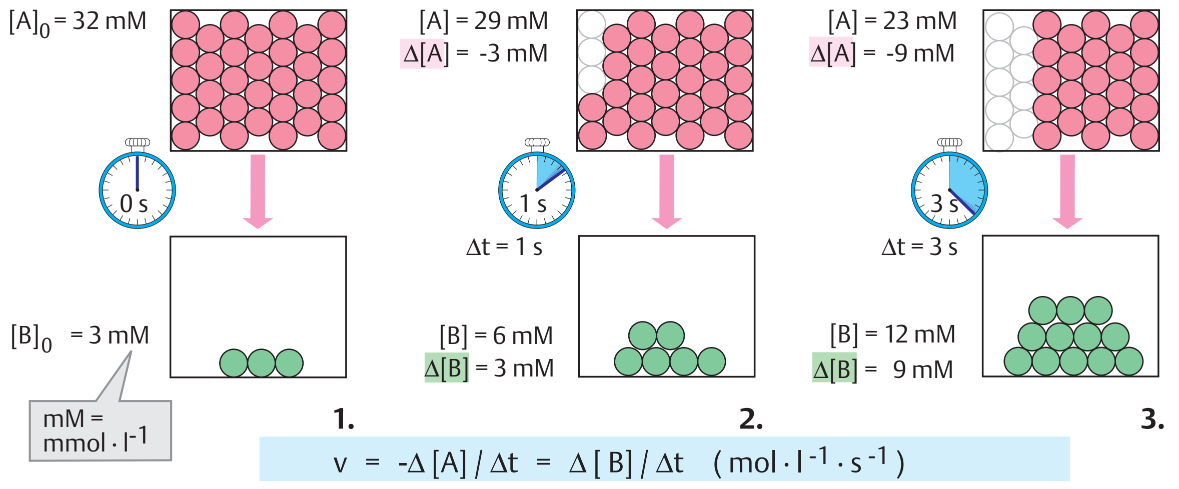
$[A]_0$ = 32 mM
0 s
$[B]_0$ = 3 mM
mM = mmol · l^{-1}
1.
[A] = 29 mM
Δ[A] = -3 mM
1 s
Δt = 1 s
[B] = 6 mM
Δ[B] = 3 mM
2.
[A] = 23 mM
Δ[A] = -9 mM
3 s
Δt = 3 s
[B] = 12 mM
Δ[B] = 9 mM
3.
$v = -\Delta[A]/\Delta t = \Delta[B]/\Delta t \quad (mol \cdot l^{-1} \cdot s^{-1})$

C. Reaction order

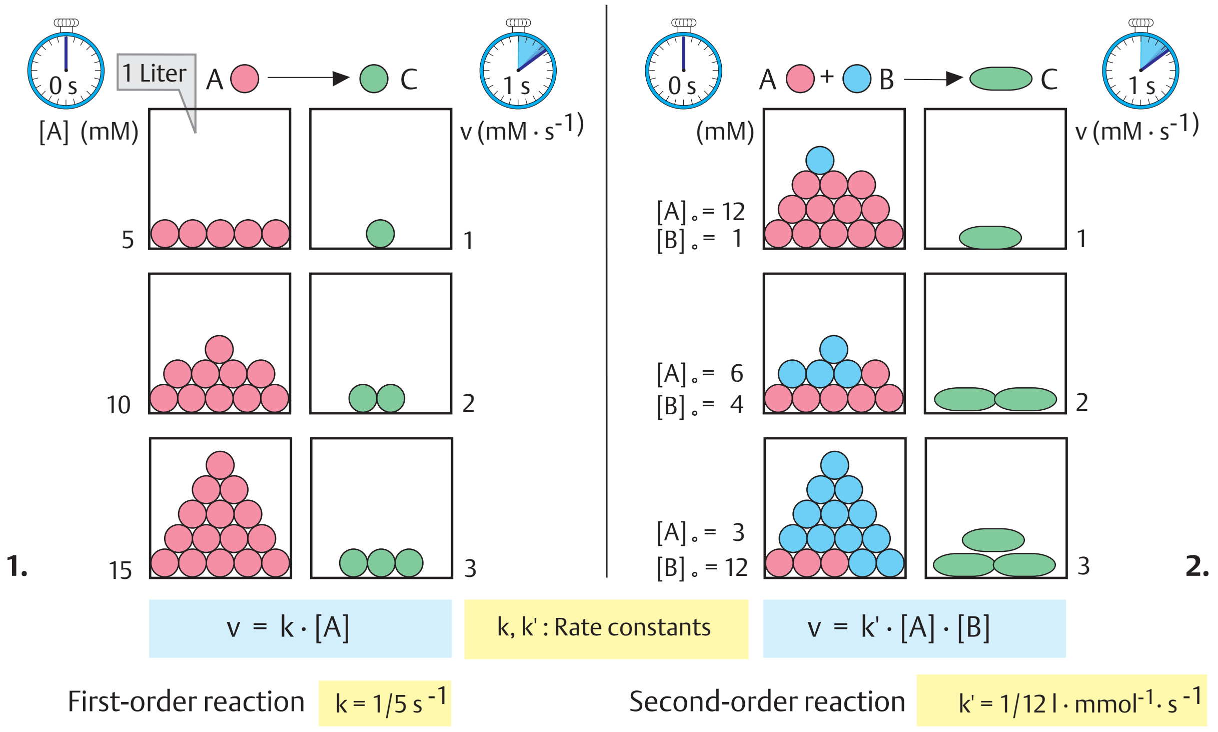
0 s
1 Liter
A → C
1 s
[A] (mM)
v (mM · s^{-1})
5
1
10
2
15
3
1.
0 s
A + B → C
1 s
(mM)
v (mM · s^{-1})
$[A]_0$ = 12
$[B]_0$ = 1
1
$[A]_0$ = 6
$[B]_0$ = 4
2
$[A]_0$ = 3
$[B]_0$ = 12
3
2.
$v = k \cdot [A]$
k, k' : Rate constants
$v = k' \cdot [A] \cdot [B]$
First-order reaction
$k = 1/5\ s^{-1}$
Second-order reaction
$k' = 1/12\ l \cdot mmol^{-1} \cdot s^{-1}$

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Catalysis

Catalysts are substances that accelerate chemical reactions without themselves being consumed in the process. Since catalysts emerge from the catalyzed reaction without being changed, even small amounts are usually sufficient to cause a powerful acceleration of the reaction. In the cell, **enzymes** (see p. 88) generally serve as catalysts. A few chemical changes are catalyzed by special RNA molecules, known as *ribozymes* (see p. 246).

A. Catalysis: principle ●

The reason for the slow rates of most reactions involving organic substances is the high **activation energy** (see p. 22) that the reacting molecules have to reach before they can react. In aqueous solution, a large proportion of the activation energy is required to remove the hydration shells surrounding the educts. During the course of a reaction, resonance-stabilized structures (see p. 4) are often temporarily suspended; this also requires energy. The highest point on the reaction coordinates corresponds to an energetically unfavorable **transition state** of this type (**1**).

A catalyst creates a new pathway for the reaction (**2**). When all of the transition states arising have a lower activation energy than that of the uncatalyzed reaction, the reaction will proceed more rapidly along the alternative pathway, even when the number of intermediates is greater. Since the starting points and end points are the same in both routes, the change in the enthalpy ΔG of the reaction is not influenced by the catalyst. Catalysts—including enzymes—are in principle *not* capable of altering the equilibrium state of the catalyzed reaction.

The often-heard statement that "a catalyst reduces the activation energy of a reaction" is not strictly correct, since a *completely different* reaction takes place in the presence of a catalyst than in uncatalyzed conditions. However, its activation energy is lower than in the uncatalyzed reaction.

B. Catalysis of H_2O_2 – breakdown by iodide ○

As a simple example of a catalyzed reaction, we can look at the disproportionation of hydrogen peroxide (H_2O_2) into oxygen and water. In the uncatalyzed reaction (at the top), an H_2O_2 molecule initially decays into H_2O and atomic oxygen (O), which then reacts with a second H_2O_2 molecule to form water and molecular oxygen (O_2). The activation energy E_a required for this reaction is relatively high, at 75 kJ mol^{-1}. In the presence of **iodide** (I^-) as a catalyst, the reaction takes a different course (bottom). The intermediate arising in this case is hypoiodide (OI^-), which also forms H_2O and O_2 with another H_2O_2 molecule. In this step, the I^- ion is released and can once again take part in the reaction. The lower activation energy of the reaction catalyzed by iodide (E_a = 56 kJ mol^{-1}) causes acceleration of the reaction by a factor of 2000, as the reaction rate depends exponentially on E_a ($v \sim e^{-Ea/RT}$).

Free metal ions such as iron (Fe) and platinum (Pt) are also effective catalysts for the breakdown of H_2O_2. **Catalase** (see p. 284), an enzyme that protects cells against the toxic effects of hydrogen peroxide (see p. 284), is much more catalytically effective still. In the enzyme-catalyzed disproportionation, H_2O_2 is bound to the enzyme's heme group, where it is quickly converted to atomic oxygen and water, supported by amino acid residues of the enzyme protein. The oxygen atom is temporarily bound to the central iron atom of the heme group, and then transferred from there to the second H_2O_2 molecule. The activation energy of the enzyme-catalyzed reaction is only 23 kJ mol^{-1}, which in comparison with the uncatalyzed reaction leads to acceleration by a factor of 1.3 10^9.

Catalase is one of the most efficient enzymes there are. A single molecule can convert up to 10^8 (a hundred million) H_2O_2 molecules per second.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Catalysis: principle

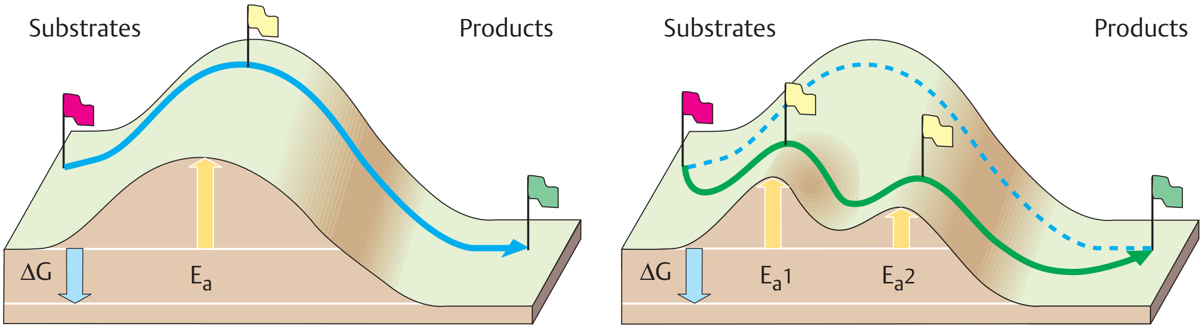
Substrates
Products
ΔG
E_a
Substrates
Products
ΔG
E_a1
E_a2

1. Energy profile without catalyst **2.** Energy profile with catalyst

B. Catalysis of H_2O_2 – breakdown by iodide

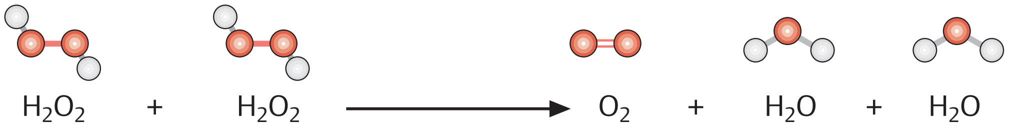
H_2O_2
+
H_2O_2
O_2
+
H_2O
+
H_2O

1. Breakdown of hydrogen peroxide

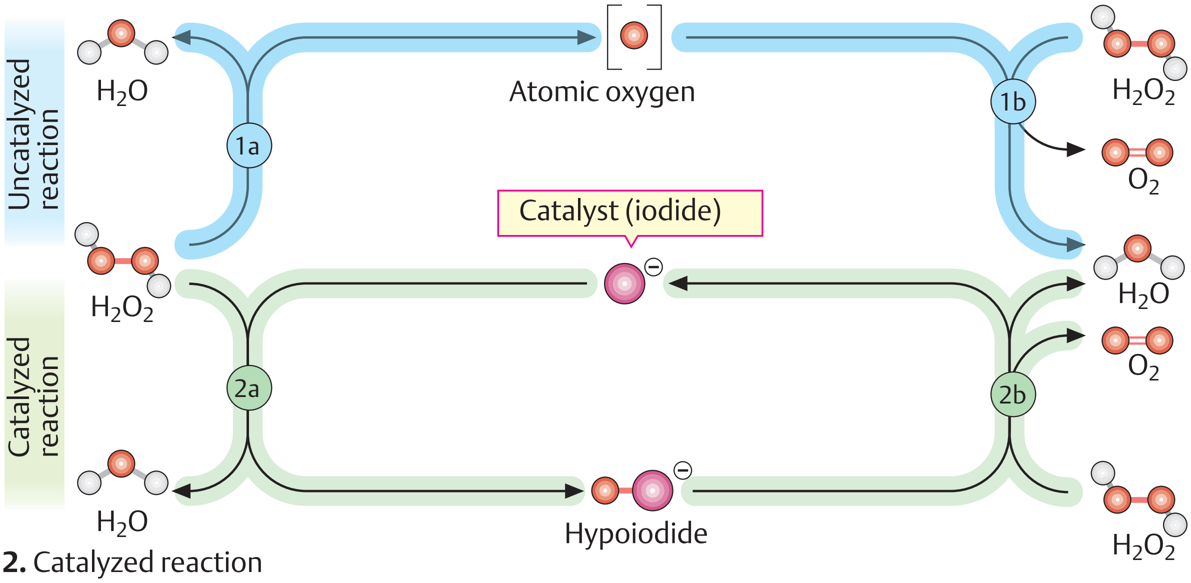
Uncatalyzed reaction
H_2O
1a
H_2O_2
Atomic oxygen
1b
H_2O_2
O_2
Catalyst (iodide)
H_2O
Catalyzed reaction
2a
2b
H_2O
O_2
H_2O
Hypoiodide
H_2O_2

2. Catalyzed reaction

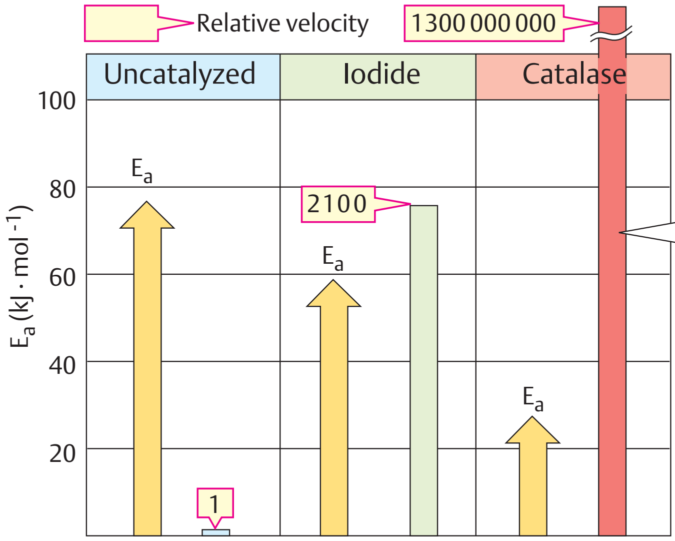

3. Activation energies

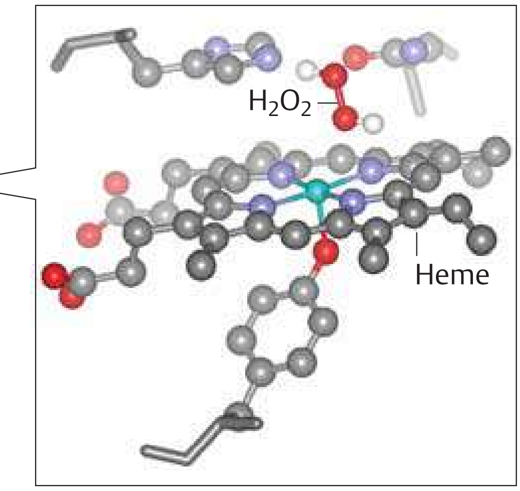

Active center of catalase

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Water as a solvent

Life as we know it evolved in water and is still absolutely dependent on it. The properties of water are therefore of fundamental importance to all living things.

A. Water and methane ◑

The special properties of **water (H_2O)** become apparent when it is compared with **methane (CH_4)**. The two molecules have a similar mass and size. Nevertheless, the boiling point of water is more than 250 °C above that of methane. At temperatures on the earth's surface, water is liquid, whereas methane is gaseous. The high boiling point of water results from its high vaporization enthalpy, which in turn is due to the fact that the density of the electrons within the molecule is unevenly distributed. Two corners of the tetrahedrally-shaped water molecule are occupied by unshared electrons (green), and the other two by hydrogen atoms. As a result, the H–O–H bond has an angled shape. In addition, the O–H bonds are polarized due to the high electronegativity of oxygen (see p. 6). One side of the molecule carries a partial charge (δ) of about –0.6 units, whereas the other is correspondingly positively charged. The spatial separation of the positive and negative charges gives the molecule the properties of an **electrical dipole**. Water molecules are therefore attracted to one another like tiny magnets, and are also connected by hydrogen bonds (**B**) (see p. 6). When liquid water vaporizes, a large amount of energy has to be expended to disrupt these interactions. By contrast, methane molecules are not dipolar, and therefore interact with one another only weakly. This is why liquid methane vaporizes at very low temperatures.

B. Structure of water and ice ◑

The dipolar nature of water molecules favors the formation of **hydrogen bonds** (see p. 6). Each molecule can act either as a donor or an acceptor of H bonds, and many molecules in liquid water are therefore connected by H bonds (**1**). The bonds are in a state of constant fluctuation. Tetrahedral networks of molecules, known as water "clusters," often arise. As the temperature decreases, the proportion of water clusters increases until the water begins to crystallize. Under normal atmospheric pressure, this occurs at 0 °C. In **ice**, most of the water molecules are fixed in a **hexagonal lattice** (**3**). Since the distance between the individual molecules in the frozen state is on average greater than in the liquid state, the density of ice is lower than that of liquid water. This fact is of immense biological importance—it means, for example, that in winter, ice forms on the surface of open stretches of water first, and the water rarely freezes to the bottom.

C. Hydration ◑

In contrast to most other liquids, water is an excellent **solvent for ions**. In the electrical field of cations and anions, the dipolar water molecules arrange themselves in a regular fashion corresponding to the charge of the ion. They form **hydration shells** and shield the central ion from oppositely charged ions. Metal ions are therefore often present as hexahydrates ($[Me(H_2O)_6{}^{2+}]$, on the right). In the inner hydration sphere of this type of ion, the water molecules are practically immobilized and follow the central ion. Water has a high dielectric constant of 78—i. e., the electrostatic attraction force between ions is reduced to 1/78 by the solvent. Electrically charged groups in organic molecules (e. g., carboxylate, phosphate, and ammonium groups) are also well hydrated and contribute to water solubility. Neutral molecules with several hydroxy groups, such as glycerol (on the left) or sugars, are also easily soluble, because they can form H bonds with water molecules. The higher the proportion of polar functional groups there is in a molecule, the more water-soluble (**hydrophilic**) it is. By contrast, molecules that consist exclusively or mainly of hydrocarbons are poorly soluble or insoluble in water. These compounds are called **hydrophobic** (see p. 28).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Water and methane

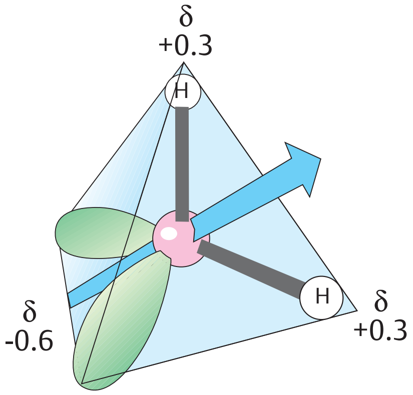
δ
+0.3
H
H
δ
-0.6
δ
+0.3

Water (H_2O)

H_2O		CH_4
18 Da	Molecular mass	16 Da
+100 °C	Boiling point	-162 °C
41	Heat of vaporization ($kJ \cdot mol^{-1}$)	8
6.2	Dipole moment (10^{-30} C · m)	0

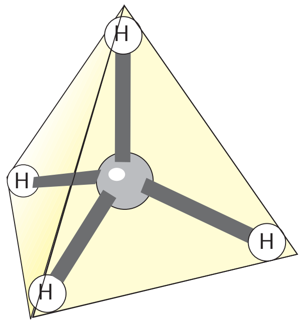

Methane (CH_4)

B. Structure of water and ice

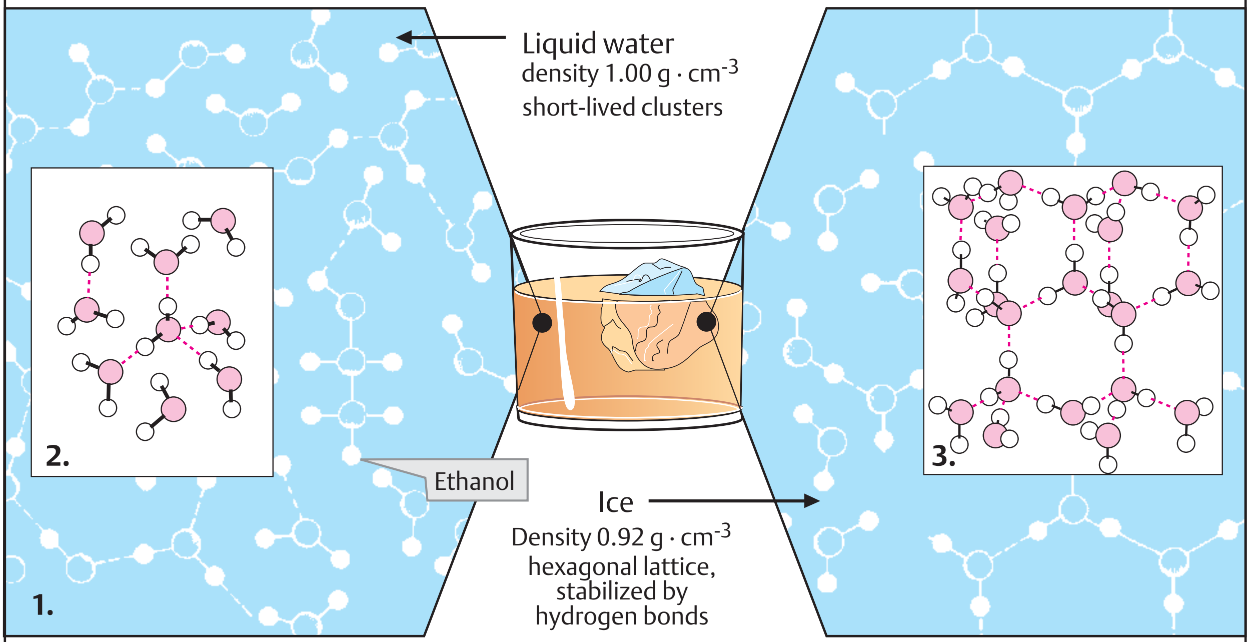
Liquid water
density 1.00 g · cm-3
short-lived clusters
2.
3.
Ethanol
Ice
Density 0.92 g · cm-3
hexagonal lattice,
stabilized by
hydrogen bonds
1.

C. Hydration

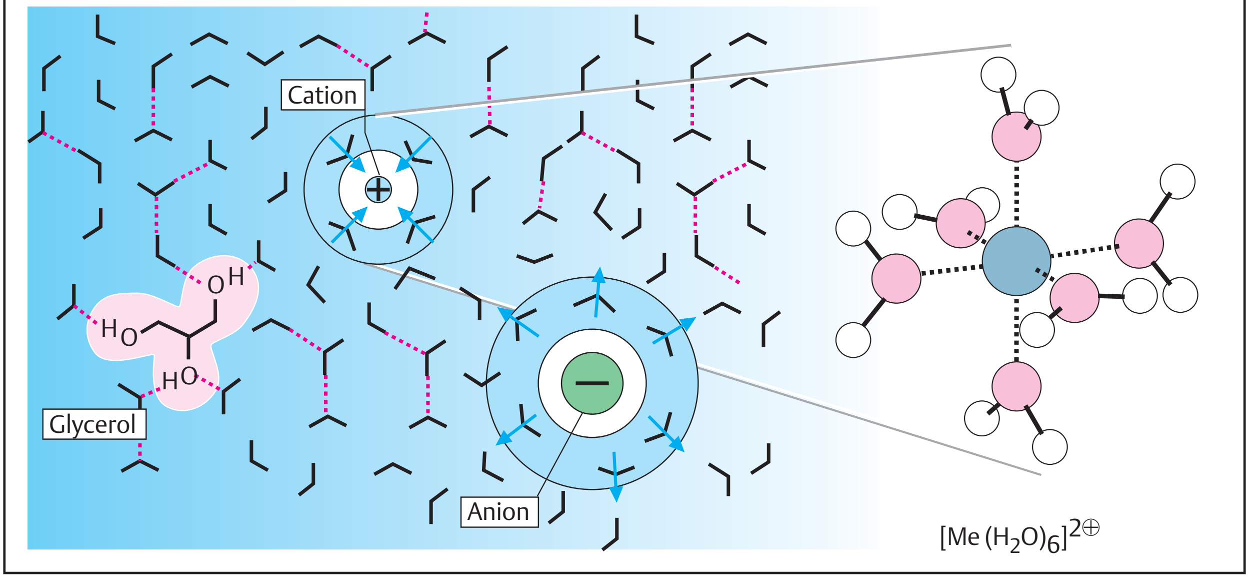
Cation
+
O H
H O
H O
Glycerol
−
Anion

$[Me(H_2O)_6]^{2\oplus}$

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Hydrophobic interactions

Water is an excellent solvent for ions and for substances that contain polarized bonds (see p. 20). Substances of this type are referred to as **polar** or **hydrophilic** ("water-loving"). In contrast, substances that consist mainly of hydrocarbon structures dissolve only poorly in water. Such substances are said to be **apolar** or **hydrophobic**.

A. Solubility of methane ○

To understand the reasons for the poor water solubility of hydrocarbons, it is useful first to examine the energetics (see p. 16) of the processes involved. In (**1**), the individual terms of the Gibbs–Helmholtz equation (see p. 20) for the simplest compound of this type, **methane**, are shown (see p. 4). As can be seen, the transition from gaseous methane to water is actually exothermic ($\Delta H^0 < 0$). Nevertheless, the change in the free enthalpy ΔG^0 is positive (the process is endergonic), because the entropy term $T \cdot \Delta S^0$ has a strongly positive value. The entropy change in the process (ΔS^0) is evidently negative—i. e., a solution of methane in water has a *higher* degree of order than either water or gaseous methane. One reason for this is that the methane molecules are less mobile when surrounded by water. More importantly, however, the water around the apolar molecules forms cage-like **"clathrate" structures**, which—as in ice—are stabilized by H bonds. This strongly increases the degree of order in the water—and the more so the larger the area of surface contact between the water and the apolar phase.

B. The "oil drop effect" ◑

The spontaneous separation of oil and water, a familiar observation in everyday life, is due to the energetically unfavorable formation of clathrate structures. When a mixture of water and oil is firmly shaken, lots of tiny oil drops form to begin with, but these quickly coalesce spontaneously to form larger drops—the two phases separate. A larger drop has a smaller surface area than several small drops with the same volume. Separation therefore reduces the area of surface contact between the water and the oil, and consequently also the extent of clathrate formation. The ΔS for this process is therefore positive (the *dis*order in the water increases), and the negative term $-T \cdot \Delta S$ makes the separation process exergonic ($\Delta G < 0$), so that it proceeds spontaneously.

C. Arrangements of amphipathic substances in water ◑

Molecules that contain both polar *and* apolar groups are called **amphipathic** or amphiphilic. This group includes soaps (see p. 48), phospholipids (see p. 50), and bile acids (see p. 56).

As a result of the "oil drop effect" amphipathic substances in water tend to arrange themselves in such a way as to minimize the area of surface contact between the apolar regions of the molecule and water. On water surfaces, they usually form single-layer **films** (top) in which the polar "head groups" face toward the water. **Soap bubbles** (right) consist of double films, with a thin layer of water enclosed between them. In water, depending on their concentration, amphipathic compounds form **micelles**—i. e., spherical aggregates with their head groups facing toward the outside, or extended bilayered **double membranes**. Most biological membranes are assembled according to this principle (see p. 214). Closed hollow membrane sacs are known as **vesicles**. This type of structure serves to transport substances within cells and in the blood (see p. 278).

The separation of oil and water (**B**) can be prevented by adding a strongly amphipathic substance. During shaking, a more or less stable **emulsion** then forms, in which the surface of the oil drops is occupied by amphipathic molecules that provide it with polar properties externally. The emulsification of fats in food by bile acids and phospholipids is a vital precondition for the digestion of fats (see p. 314).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Solubility of methane

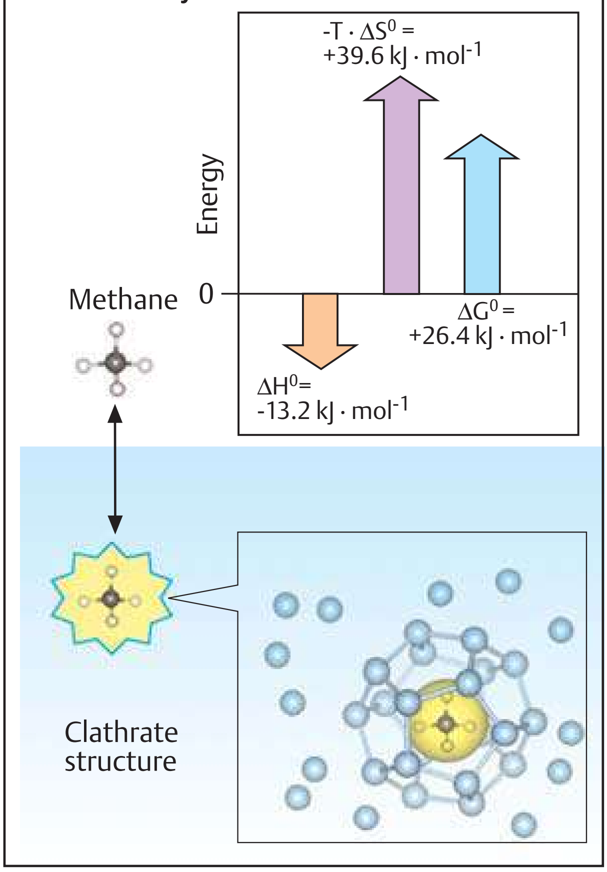
-T · ΔS⁰ =
+39.6 kJ · mol⁻¹
Energy
Methane
0
ΔG⁰ =
+26.4 kJ · mol⁻¹
ΔH⁰=
-13.2 kJ · mol⁻¹
Clathrate
structure

B. The "oil drop effect"

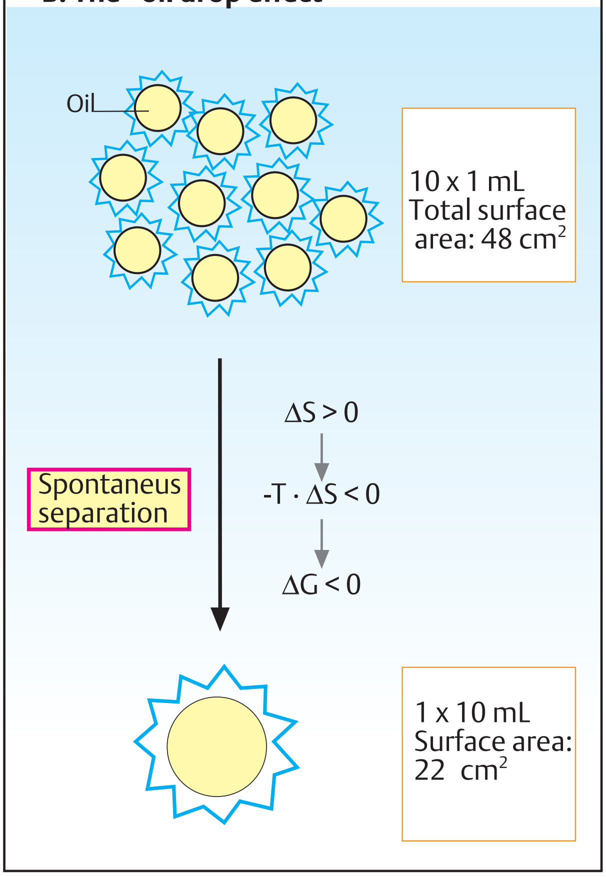
Oil
10 x 1 mL
Total surface
area: 48 cm²
ΔS > 0
-T · ΔS < 0
ΔG < 0
Spontaneus
separation
1 x 10 mL
Surface area:
22 cm²

C. Arrangements of amphipathic substances in water

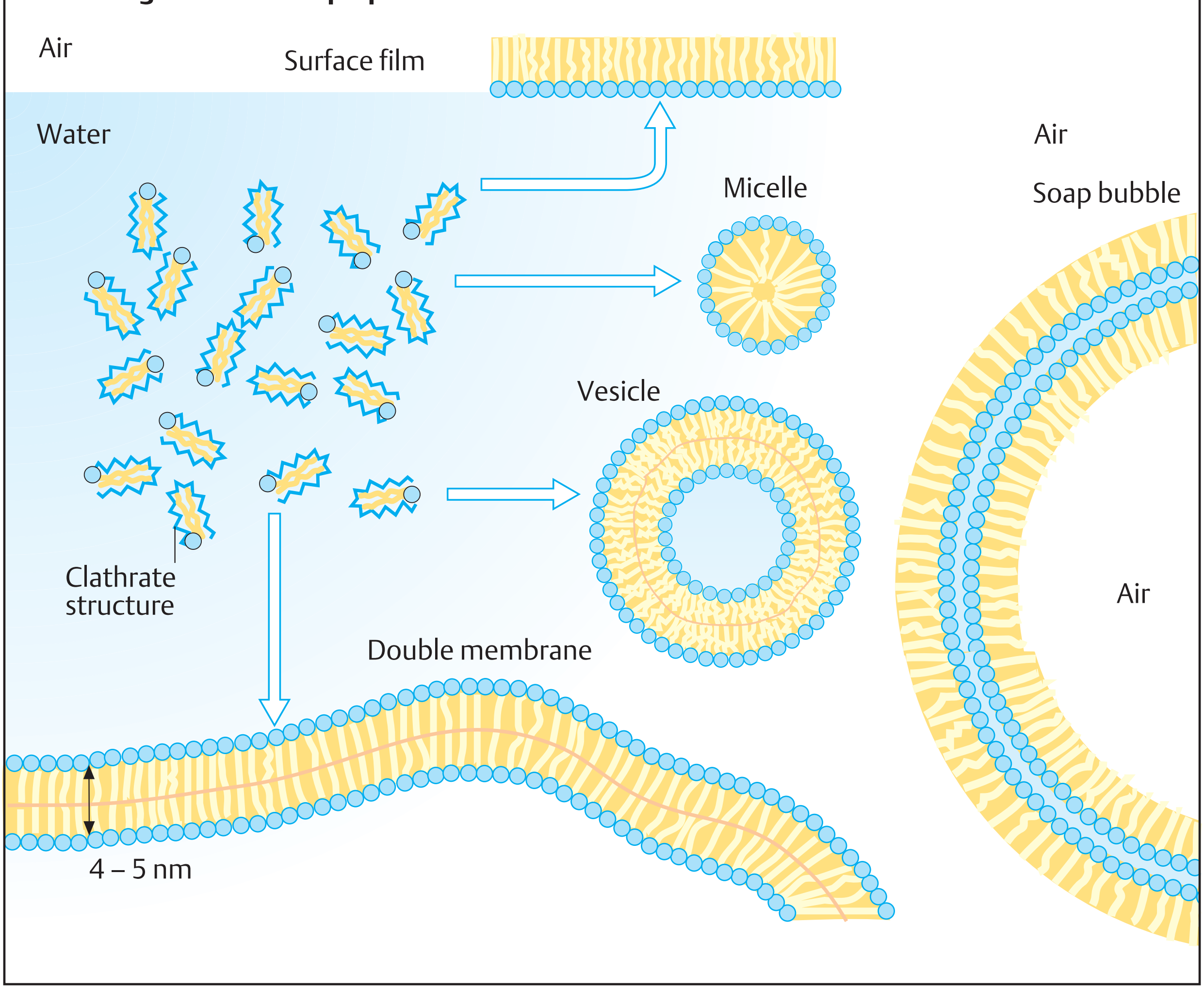
Air
Surface film
Water
Air
Micelle
Soap bubble
Vesicle
Clathrate
structure
Double membrane
Air
4 – 5 nm

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Acids and bases

A. Acids and bases ●

In general, **acids** are defined as substances that can donate hydrogen ions (protons), while **bases** are compounds that accept protons.

Water enhances the acidic or basic properties of dissolved substances, as water itself can act as either an acid or a base. For example, when **hydrogen chloride** (HCl) is in aqueous solution, it donates protons to the solvent (**1**). This results in the formation of chloride ions (Cl^-) and protonated water molecules (**hydronium ions, H_3O+**, usually simply referred to as H^+). The proton exchange between HCl and water is virtually quantitative: in water, HCl behaves as a *very strong acid* with a negative pK_a value (see p. 18).

Bases such as **ammonia** (NH_3) take over protons from water molecules. As a result of this, **hydroxyl ions** (OH^-) and positively charged ammonium ions (NH_4^+, **3**) form. Hydronium and hydroxyl ions, like other ions, exist in water in hydrated rather than free form (see p. 26).

Acid–base reactions always involve *pairs* of **acids** and the associated **conjugated bases** (see p. 18). The stronger the acid or base, the *weaker* the conjugate base or acid, respectively. For example, the very strongly acidic hydrogen chloride belongs to the very weakly basic chloride ion (**1**). The weakly acidic ammonium ion is conjugated with the moderately strong base ammonia (**3**).

The equilibrium constant K for the acid—base reaction between H_2O molecules (**2**) is very small. At 25 °C,

$$K = [H^+]\quad [OH^-] / [H_2O] = 2 \quad 10^{-16}\ \text{mol} \quad L^{-1}$$

In pure water, the concentration $[H_2O]$ is practically constant at 55 mol L^{-1}. Substituting this value into the equation, it gives:

$$K_w = [H^+]\quad [OH^-] = 1 \quad 10^{-14}\ \text{mol} \quad L^{-1}$$

The product $[H^+]$ $[OH^-]$—the **ion product** of water—is constant even when additional acid–base pairs are dissolved in the water. At 25 °C, pure water contains H^+ and OH^- at concentrations of 1 10^{-7} mol L^{-1} each; it is **neutral** and has a pH value of exactly 7.

B. pH values in the organism ◑

pH values in the cell and in the extracellular fluid are kept constant within narrow limits. In the blood, the pH value normally ranges only between 7.35 and 7.45 (see p. 288). This corresponds to a maximum change in the H^+ concentration of ca. 30%. The pH value of cytoplasm is slightly lower than that of blood, at 7.0–7.3. In lysosomes (see p. 234; pH 4.5–5.5), the H^+ concentration is several hundred times higher than in the cytoplasm. In the lumen of the gastrointestinal tract, which forms part of the outside world relative to the organism, and in the body's excretion products, the pH values are more variable. Extreme values are found in the stomach (ca. 2) and in the small bowel (> 8). Since the kidney can excrete either acids or bases, depending on the state of the metabolism, the pH of urine has a particularly wide range of variation (4.8–7.5).

C. Buffers ●

Short-term pH changes in the organism are cushioned by **buffer systems**. These are mixtures of a weak acid, HB, with its conjugate base, B^-, or of a weak base with its conjugate acid. This type of system can neutralize both hydronium ions and hydroxyl ions.

In the first case (left), the base (B^-) binds a large proportion of the added protons (H^+) and HB and water are formed. If hydroxyl ions (OH^-) are added, they react with HB to give B^- and water (right). In both cases, it is primarily the $[HB]/[B^-]$ ratio that shifts, while the pH value only changes slightly. The **titration curve** (top) shows that buffer systems are most effective at the pH values that correspond to the pK_a value of the acid. This is where the curve is at its steepest, so that the pH change, ΔpH, is at its smallest with a given increase Δc in $[H^+]$ or $[OH^-]$. In other words, the **buffer capacity** $\Delta c / \Delta pH$ is highest at the pK_a value.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Acids and bases

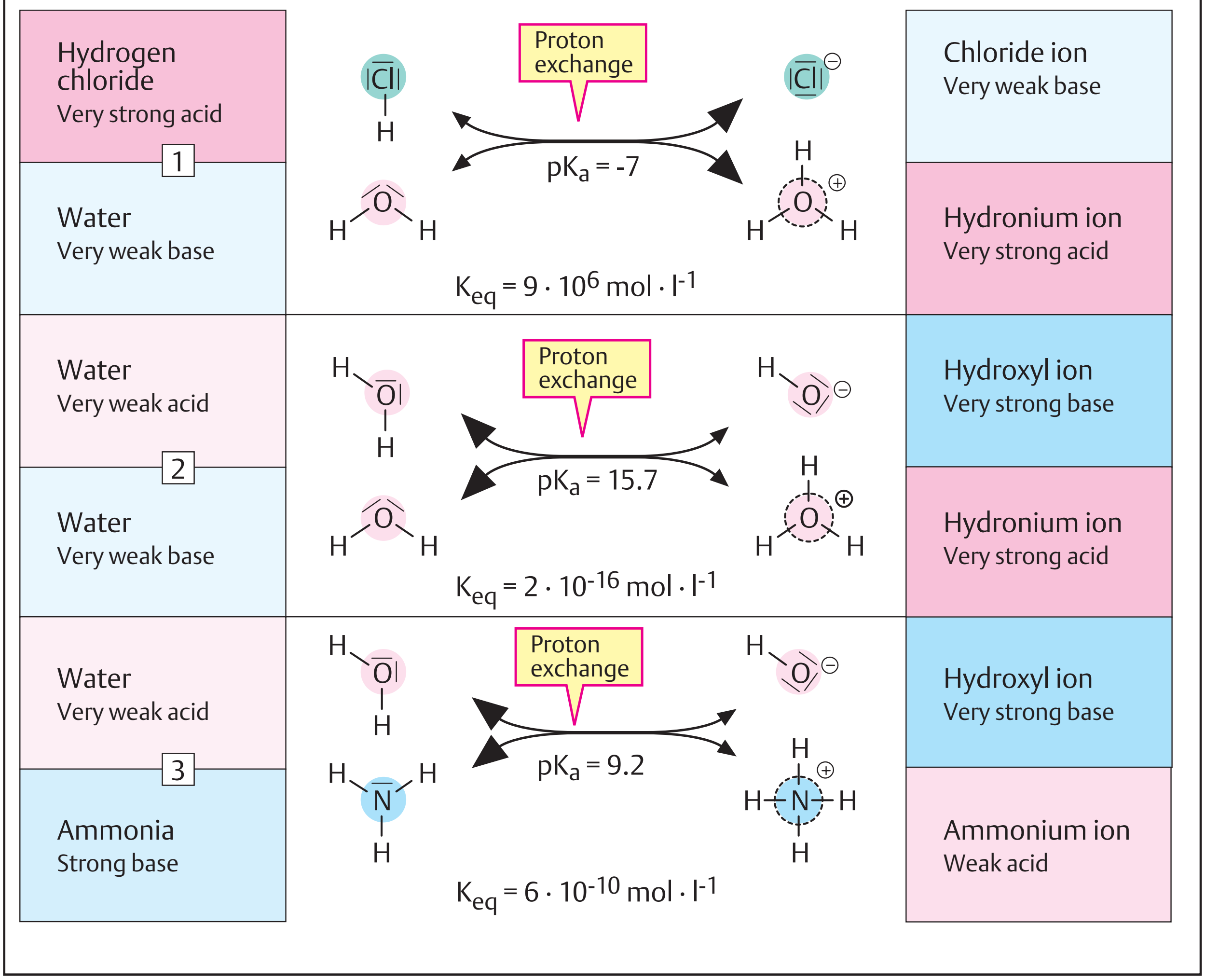
Hydrogen chloride
Very strong acid
1
Water
Very weak base
Proton exchange
pKa = -7
$K_{eq} = 9 \cdot 10^6$ mol · l^{-1}
Chloride ion
Very weak base
Hydronium ion
Very strong acid
Water
Very weak acid
2
Water
Very weak base
Proton exchange
pKa = 15.7
$K_{eq} = 2 \cdot 10^{-16}$ mol · l^{-1}
Hydroxyl ion
Very strong base
Hydronium ion
Very strong acid
Water
Very weak acid
3
Ammonia
Strong base
Proton exchange
pKa = 9.2
$K_{eq} = 6 \cdot 10^{-10}$ mol · l^{-1}
Hydroxyl ion
Very strong base
Ammonium ion
Weak acid

B. pH values in the body

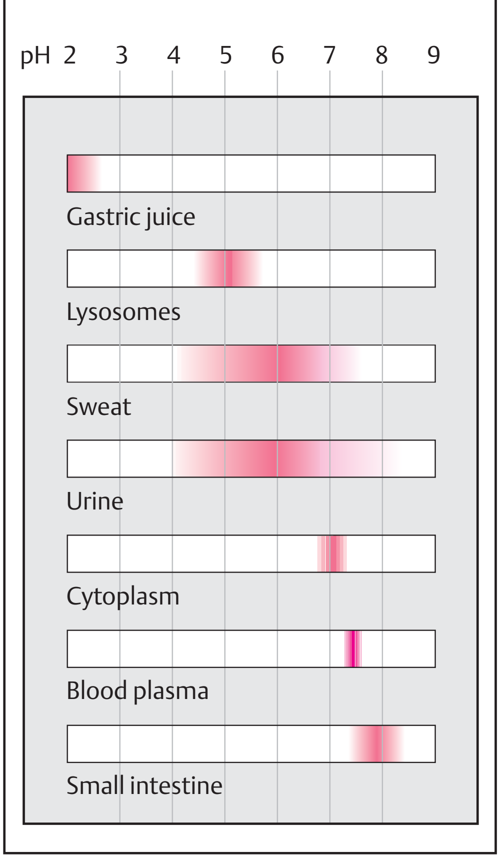

C. Buffers

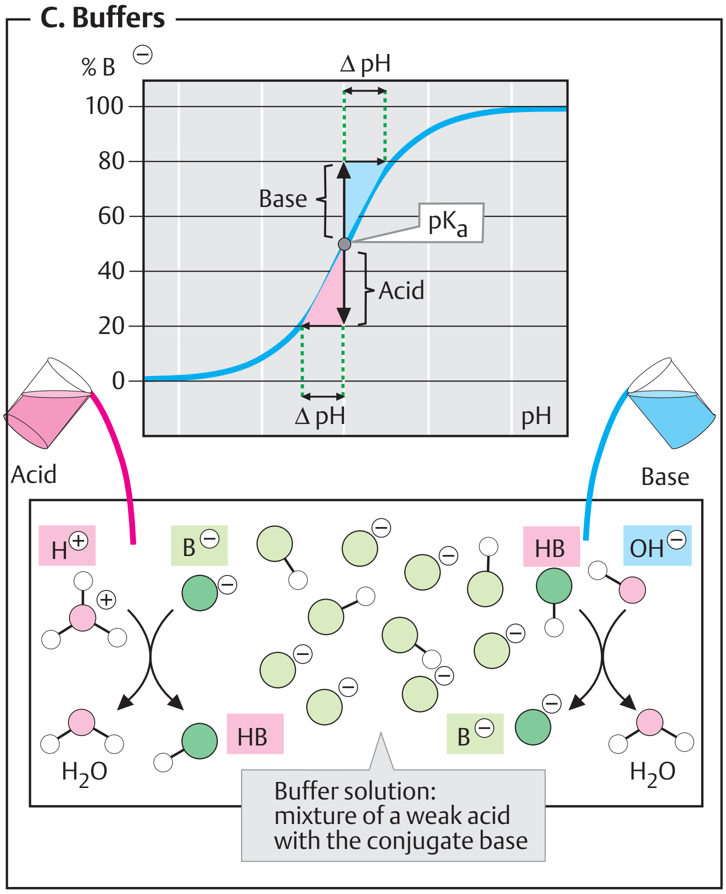
% B⊖
100
80
60
40
20
0
Δ pH
Base
pKa
Acid
Δ pH
pH
Acid
Base
H⊕
B⊖
HB
OH⊖
H₂O
HB
B⊖
H₂O
Buffer solution:
mixture of a weak acid
with the conjugate base

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Redox processes

A. Redox reactions ●

Redox reactions are chemical changes in which electrons are transferred from one reaction partner to another (**1**; see also p. 18). Like acid–base reactions (see p. 30), redox reactions always involve *pairs* of compounds. A pair of this type is referred to as a **redox system** (**2**). The essential difference between the two components of a redox system is the number of electrons they contain. The more electronrich component is called the **reduced form** of the compound concerned, while the other one is referred to as the **oxidized form**. The reduced form of one system (the **reducing agent**) donates electrons to the oxidized form of another one (the **oxidizing agent**). In the process, the reducing agent becomes oxidized and the oxidizing agent is reduced (**3**). Any given reducing agent can reduce only certain other redox systems. On the basis of this type of observation, redox systems can be arranged to form what are known as **redox series** (**4**).

The position of a system within one of these series is established by its **redox potential E** (see p. 18). The redox potential has a sign; it can be more negative or more positive than a reference potential arbitrarily set at zero (the normal potential of the system [$2\ H^+/H_2$]). In addition, E depends on the concentrations of the reactants and on the reaction conditions (see p. 18). In redox series (**4**), the systems are arranged according to their increasing redox potentials. Spontaneous electron transfers are only possible if the redox potential of the donor is *more negative* than that of the acceptor (see p. 18).

B. Reduction equivalents ◑

In redox reactions, protons (H^+) are often transferred along with electrons (e^-), or protons may be released. The combinations of electrons and protons that occur in redox processes are summed up in the term **reduction equivalents**. For example, the combination $1\ e^-/1\ H^+$ corresponds to a hydrogen *atom*, while $2\ e^-$ and $2\ H^+$ together produce a hydrogen *molecule*. However, this does not mean that atomic or molecular hydrogen is actually transferred from one molecule to the other (see below). Only the combination $2\ e^-/1\ H^+$, the **hydride ion**, is transferred as a unit.

C. Biological redox systems ◑

In the cell, redox reactions are catalyzed by enzymes, which work together with soluble or bound redox cofactors.

Some of these factors contain **metal ions** as redox-active components. In these cases, it is usually single electrons that are transferred, with the metal ion changing its valency. Unpaired electrons often occur in this process, but these are located in d orbitals (see p. 2) and are therefore less dangerous than single electrons in non-metal atoms ("free radicals"; see below).

We can only show here a few examples from the many organic redox systems that are found. In the complete reduction of the **flavin coenzymes** FMN and FAD (see p. 104), $2\ e^-$ and $2\ H^+$ are transferred. This occurs in two separate steps, with a *semiquinone radical* appearing as an intermediate. Since organic radicals of this type can cause damage to biomolecules, flavin coenzymes never occur freely in solution, but remain firmly bound in the interior of proteins.

In the reduction or oxidation of **quinone/quinol systems**, free radicals also appear as intermediate steps, but these are less reactive than flavin radicals. Vitamin E, another quinone-type redox system (see p. 104), even functions as a radical scavenger, by delocalizing unpaired electrons so effectively that they can no longer react with other molecules.

The **pyridine nucleotides** NAD^+ and $NADP^+$ always function in unbound form. The oxidized forms contain an aromatic nicotinamide ring in which the positive charge is delocalized. The right-hand example of the two *resonance structures* shown contains an electron-poor, positively charged C atom at the *para* position to nitrogen. If a **hydride ion** is added at this point (see above), the reduced forms NADH or NADPH arise. No radical intermediate steps occur. Because a proton is released at the same time, the reduced pyridine nucleotide coenzymes are correctly expressed as $NAD(P)H+H^+$.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Redox reactions

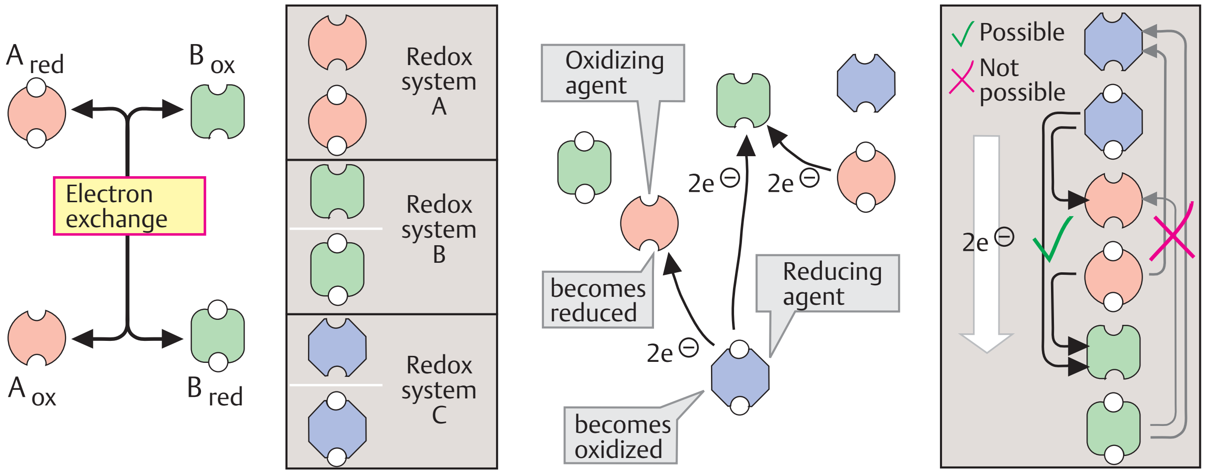
A red
B ox
Electron exchange
A ox
B red
Redox system A
Redox system B
Redox system C
Oxidizing agent
becomes reduced
2e⊖
2e⊖
2e⊖
Reducing agent
becomes oxidized
✓ Possible
✗ Not possible
2e⊖

1. Principle **2.** Redox systems **3.** Possible electron transfers **4.** Redox series

B. Reducing equivalents

Transferred components	$1e^{\ominus}$	$1e^{\ominus}$ + $1H^{\oplus}$	$2e^{\ominus}$ + $1H^{\oplus}$	$2e^{\ominus}$ + $2H^{\oplus}$
	$e^{\ominus}$	[H]	$H^{\ominus}$	$[H_2]$
Equivalent	Electron	Hydrogen atom	Hydride ion	Hydrogen molecule

C. Biological redox systems

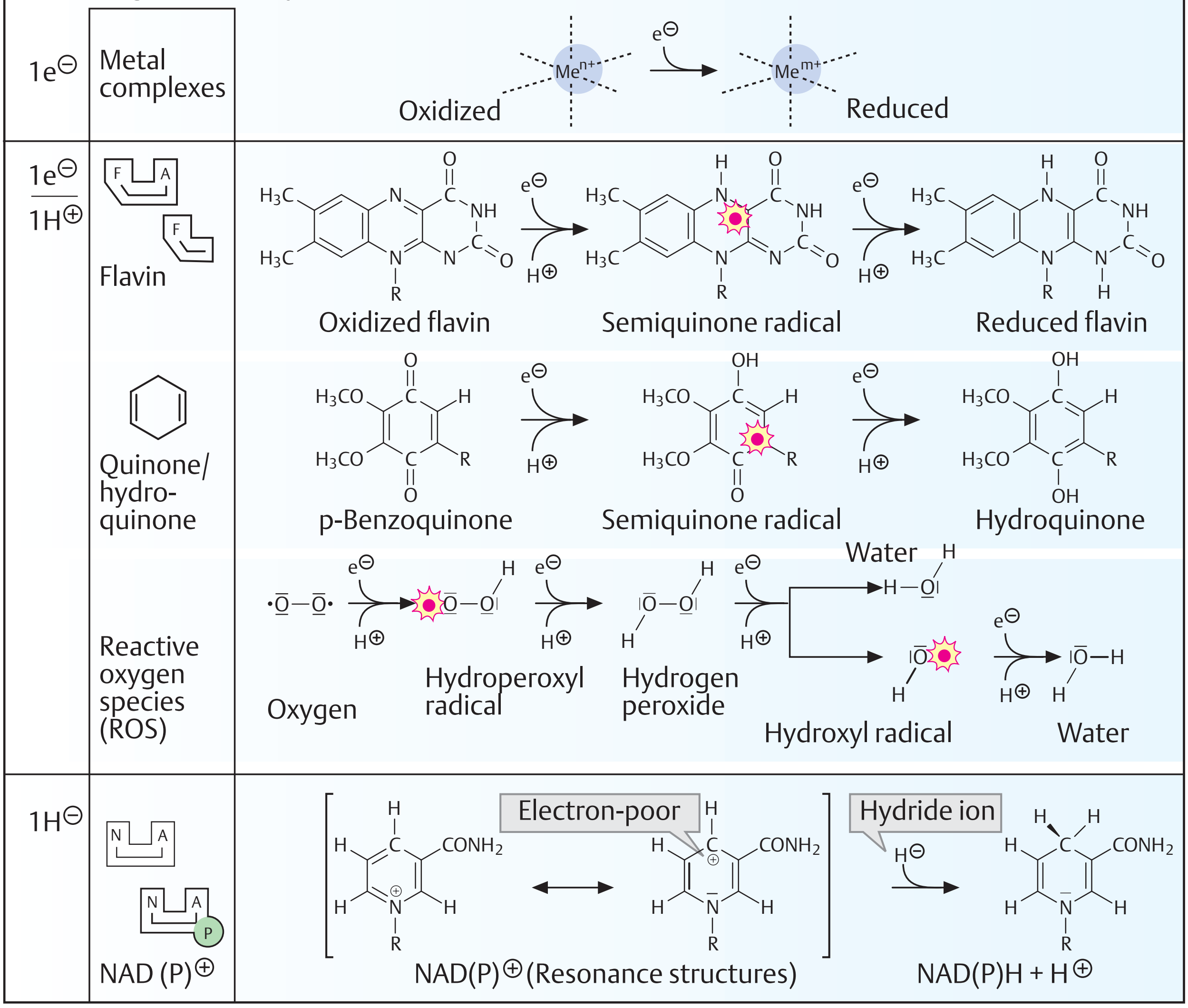
1e⊖
Metal complexes
Oxidized
Me n+
e⊖
Me m+
Reduced
1e⊖ / 1H⊕
F A
F
Flavin
Oxidized flavin
Semiquinone radical
Reduced flavin
e⊖
H⊕
Quinone/ hydro- quinone
p-Benzoquinone
Semiquinone radical
Hydroquinone
Reactive oxygen species (ROS)
Oxygen
Hydroperoxyl radical
Hydrogen peroxide
Water
Hydroxyl radical
Water
1H⊖
N A
N A P
NAD (P)⊕
Electron-poor
Hydride ion
H⊖
NAD(P)⊕ (Resonance structures)
NAD(P)H + H⊕

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Overview

The **carbohydrates** are a group of naturally occurring carbonyl compounds (aldehydes or ketones) that also contain several hydroxyl groups. The carbohydrates include **single sugars (monosaccharides)** and their polymers, the **oligosaccharides** and **polysaccharides**.

A. Carbohydrates: overview ●

Polymeric carbohydrates–above all starch, as well as some disaccharides–are important (but not essential) **components of food** (see p. 360). In the gut, they are broken down into monosaccharides and resorbed in this form (see p. 272). The form in which carbohydrates are distributed by the blood of vertebrates is *glucose* ("blood sugar"). This is taken up by the cells and either broken down to obtain energy (glycolysis) or converted into other metabolites (see pp. 150–159). Several organs (particularly the liver and muscles) store *glycogen* as a polymeric **reserve carbohydrate** (right; see p. 156). The glycogen molecules are covalently bound to a protein, *glycogenin.* Polysaccharides are used by many organisms as **building materials**. For example, the cell walls of bacteria contain *murein* as a stabilizing component (see p. 40), while in plants *cellulose* and other polysaccharides fulfill this role (see p. 42). Oligomeric or polymeric carbohydrates are often covalently bound to lipids or proteins. The **glycolipids** and **glycoproteins** formed in this way are found, for example, in cell membranes (center). Glycoproteins also occur in the blood in solute form (plasma proteins; see p. 276) and, as components of *proteoglycans,* form important constituents of the intercellular substance (see p. 346).

B. Monosaccharides: structure ◑

The most important natural monosaccharide, **D-glucose**, is an aliphatic aldehyde with six C atoms, five of which carry a hydroxyl group (**1**). Since C atoms 2 to 5 represent chiral centers (see p. 8), there are 15 further isomeric *aldohexoses* in addition to D-glucose, although only a few of these are important in nature (see p. 38). Most natural monosaccharides have the same configuration at C-5 as D-glyceraldehyde–they belong to the D **series**.

The open-chained form of glucose shown in (**1**) is found in neutral solution in less than 0.1% of the molecules. The reason for this is an intramolecular reaction in which one of the OH groups of the sugar is added to the aldehyde group of the *same* molecule (**2**). This gives rise to a cyclic **hemiacetal** (see p. 10). In aldohexoses, the hydroxy group at C-5 reacts preferentially, and a six-membered pyran ring is formed. Sugars that contain this ring are called **pyranoses**. By contrast, if the OH group at C-4 reacts, a five-part furan ring is formed. In solution, *pyranose* forms and *furanose* forms are present in equilibrium with each other and with the open-chained form, while in glucose polymers only the pyranose form occurs.

The **Haworth projection** (**2**) is usually used to depict sugars in the cyclic form, with the ring being shown in perspective as viewed from above. Depending on the configuration, the substituents of the chiral C atoms are then found above or below the ring. OH groups that lie on the *right* in the Fischer projection (**1**) appear *under* the ring level in the Haworth projection, while those on the *left* appear *above* it.

As a result of hemiacetal formation, an additional chiral center arises at C-1, which can be present in both possible configurations (anomers) (see p. 8). To emphasize this, the corresponding bonds are shown here using wavy lines.

The Haworth formula does not take account of the fact that the pyran ring is not plain, but usually has a *chair conformation.* In **B3**, two frequent conformations of D-glucopyranose are shown as ball-and-stick models. In the $^1C_4$ conformation (bottom), most of the OH groups appear vertical to the ring level, as in the Haworth projection (**axial** or **a** position). In the slightly more stable $^4C_1$ conformation (top), the OH groups take the **equatorial** or **e** position. At room temperature, each form can change into the other, as well as into other conformations.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Carbohydrates: overview

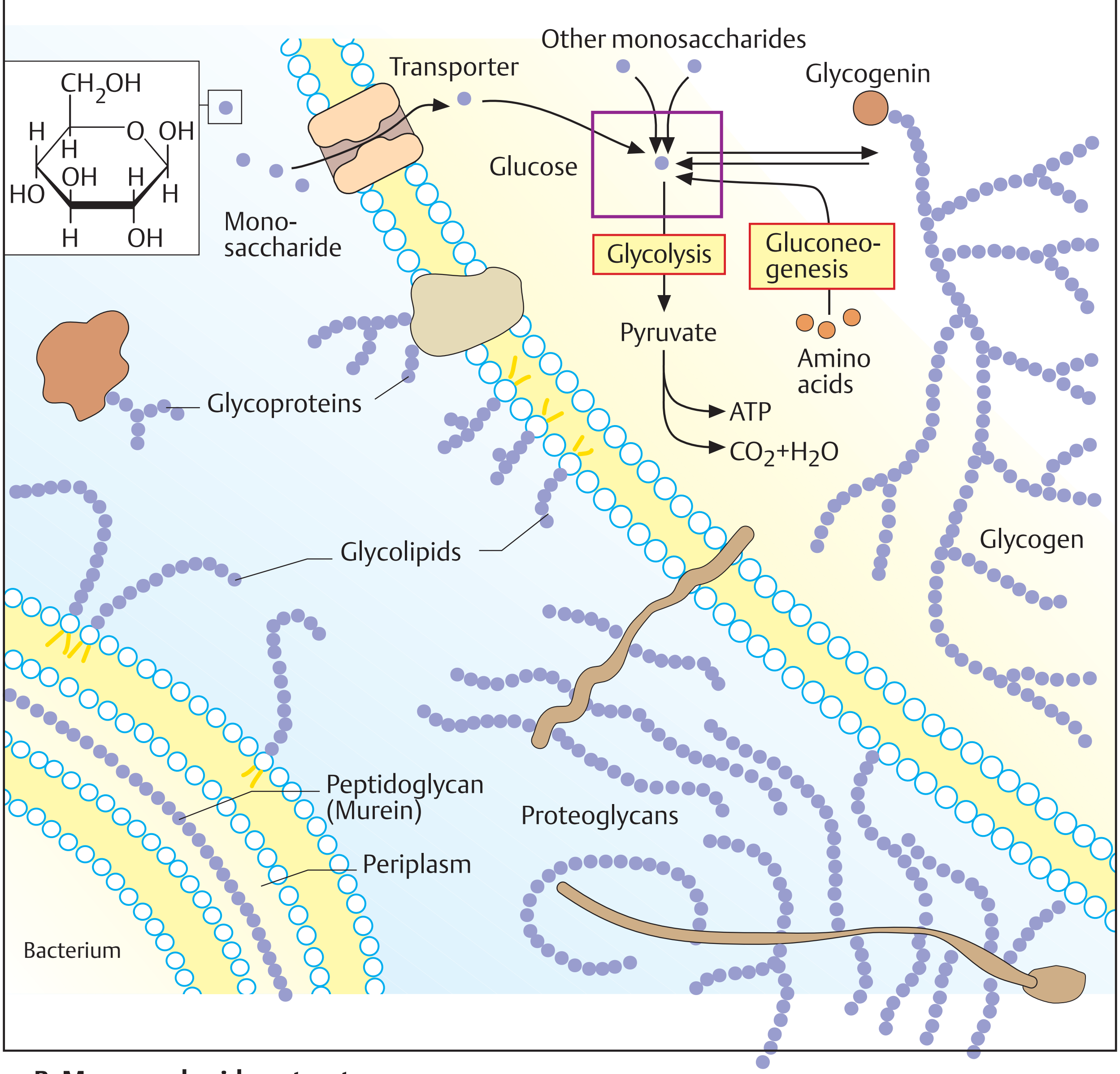
CH2OH
H
O
OH
H
OH
H
HO
H
H
OH
Mono-
saccharide
Transporter
Other monosaccharides
Glucose
Glycogenin
Glycolysis
Gluconeo-
genesis
Pyruvate
Amino
acids
ATP
CO2+H2O
Glycoproteins
Glycolipids
Glycogen
Peptidoglycan
(Murein)
Periplasm
Proteoglycans
Bacterium

B. Monosaccharides: structure

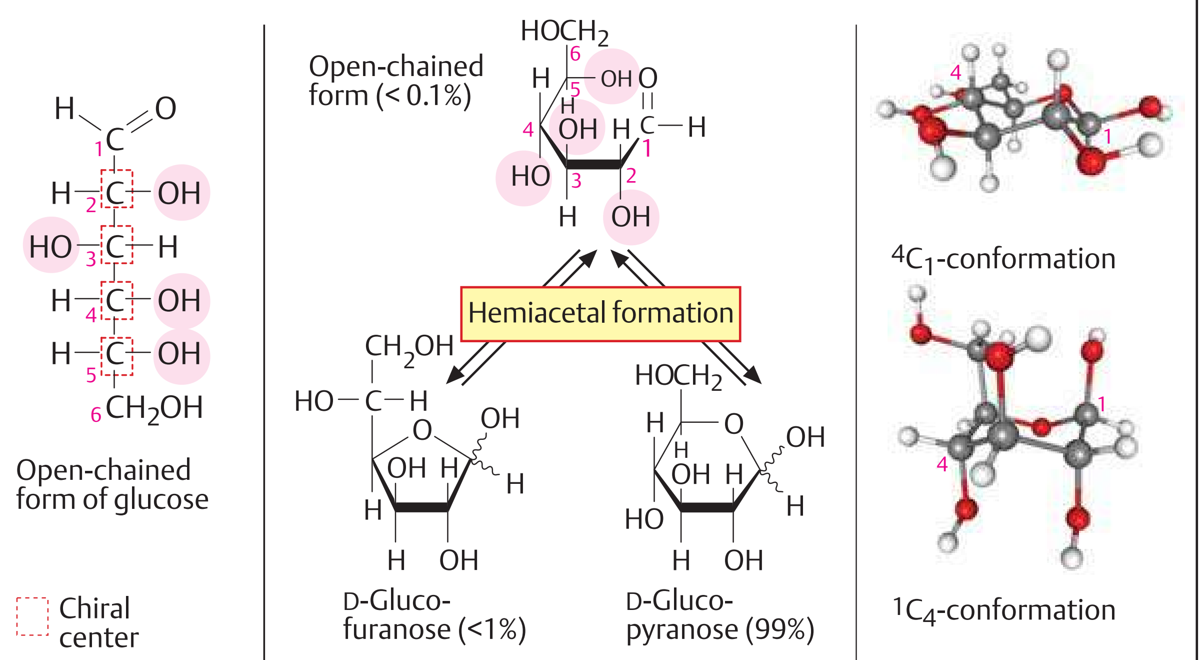
H
O
1
C
H
2
C
OH
HO
3
C
H
H
4
C
OH
H
5
C
OH
6
CH2OH
Open-chained
form of glucose
Chiral
center
Open-chained
form (< 0.1%)
HOCH2
6
H
5
OH
O
H
4
OH
H
C
H
1
HO
3
2
H
OH
Hemiacetal formation
CH2OH
HO
C
H
O
OH
OH
H
H
H
H
OH
D-Gluco-
furanose (<1%)
HOCH2
H
O
OH
H
OH
H
HO
H
H
OH
D-Gluco-
pyranose (99%)
4
1
4C1-conformation
1
4
1C4-conformation

1. Fischer projection **2.** Ring forms (Haworth projection) **3.** Conformations

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Chemistry of sugars

A. Reactions of the monosaccharides ◑

The sugars (monosaccharides) occur in the metabolism in many forms (derivatives). Only a few important conversion reactions are discussed here, using **D**-glucose as an example.

1. Mutarotation. In the cyclic form, as opposed to the open-chain form, aldoses have a chiral center at C-1 (see p. 34). The corresponding isomeric forms are called **anomers**. In the β-anomer (center left), the OH group at C-1 (the anomeric OH group) and the CH_2OH group lie on the *same* side of the ring. In the α-anomer (right), they are on different sides. The reaction that interconverts anomers into each other is known as *mutarotation* (**B**).

2. Glycoside formation. When the anomeric OH group of a sugar reacts with an alcohol, with elimination of water, it yields an *O–glycoside* (in the case shown, α –methylglucoside). The glycosidic bond is not a normal ether bond, because the OH group at C-1 has a hemiacetal quality. Oligosaccharides and polysaccharides also contain *O*-glycosidic bonds. Reaction of the anomeric OH group with an NH_2 or NH group yields an *N-glycoside* (not shown). *N*-glycosidic bonds occur in nucleotides (see p. 80) and in glycoproteins (see p. 44), for example.

3. Reduction and oxidation. Reduction of the anomeric center at C-1 of glucose (2) produces the sugar alcohol *sorbitol*. Oxidation of the aldehyde group at C-1 gives the intramolecular ester (lactone) of *gluconic acid* (a glyconic acid). Phosphorylated gluconolactone is an intermediate of the pentose phosphate pathway (see p. 152). When glucose is oxidized at C-6, *glucuronic acid* (a glycuronic acid) is formed. The strongly polar glucuronic acid plays an important role in biotransformations in the liver (see pp. 194, 316).

4. Epimerization. In weakly alkaline solutions, glucose is in equilibrium with the ketohexose D-*fructose* and the aldohexose D-*mannose*, via an enediol intermediate (not shown). The only difference between glucose and mannose is the configuration at C-2. Pairs of sugars of this type are referred to as *epimers*, and their interconversion is called *epimerization*.

5. Esterification. The hydroxyl groups of monosaccharides can form *esters* with acids. In metabolism, phosphoric acid esters such as *glucose 6-phosphate* and *glucose 1-phosphate* (6) are particularly important.

B. Polarimetry, mutarotation ○

Sugar solutions can be analyzed by **polarimetry**, a method based on the interaction between chiral centers and linearly polarized light—i. e., light that oscillates in only *one* plane. It can be produced by passing normal light through a special filter (a **polarizer**). A second polarizing filter of the same type (the **analyzer**), placed behind the first, only lets the polarized light pass through when the polarizer and the analyzer are in alignment. In this case, the field of view appears bright when one looks through the analyzer (**1**). Solutions of chiral substances rotate the plane of polarized light by an angle α either to the left or to the right. When a solution of this type is placed between the polarizer and the analyzer, the field of view appears darker (**2**). The angle of rotation, α, is determined by turning the analyzer until the field of view becomes bright again (**3**). A solution's **optical rotation** depends on the type of chiral compound, its concentration, and the thickness of the layer of the solution. This method makes it possible to determine the sugar content of wines, for example.

Certain procedures make it possible to obtain the α and β anomers of glucose in pure form. A 1-molar solution of α-D-glucose has a rotation value $[\alpha]_D$ of +112°, while a corresponding solution of β-D-glucose has a value of +19°. These values change spontaneously, however, and after a certain time reach the same end point of +52°. The reason for this is that, in solution, **mutarotation** leads to an equilibrium between the α and β forms in which, independently of the starting conditions, 62% of the molecules are present in the β form and 38% in the α form.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Reactions of the monosaccharides

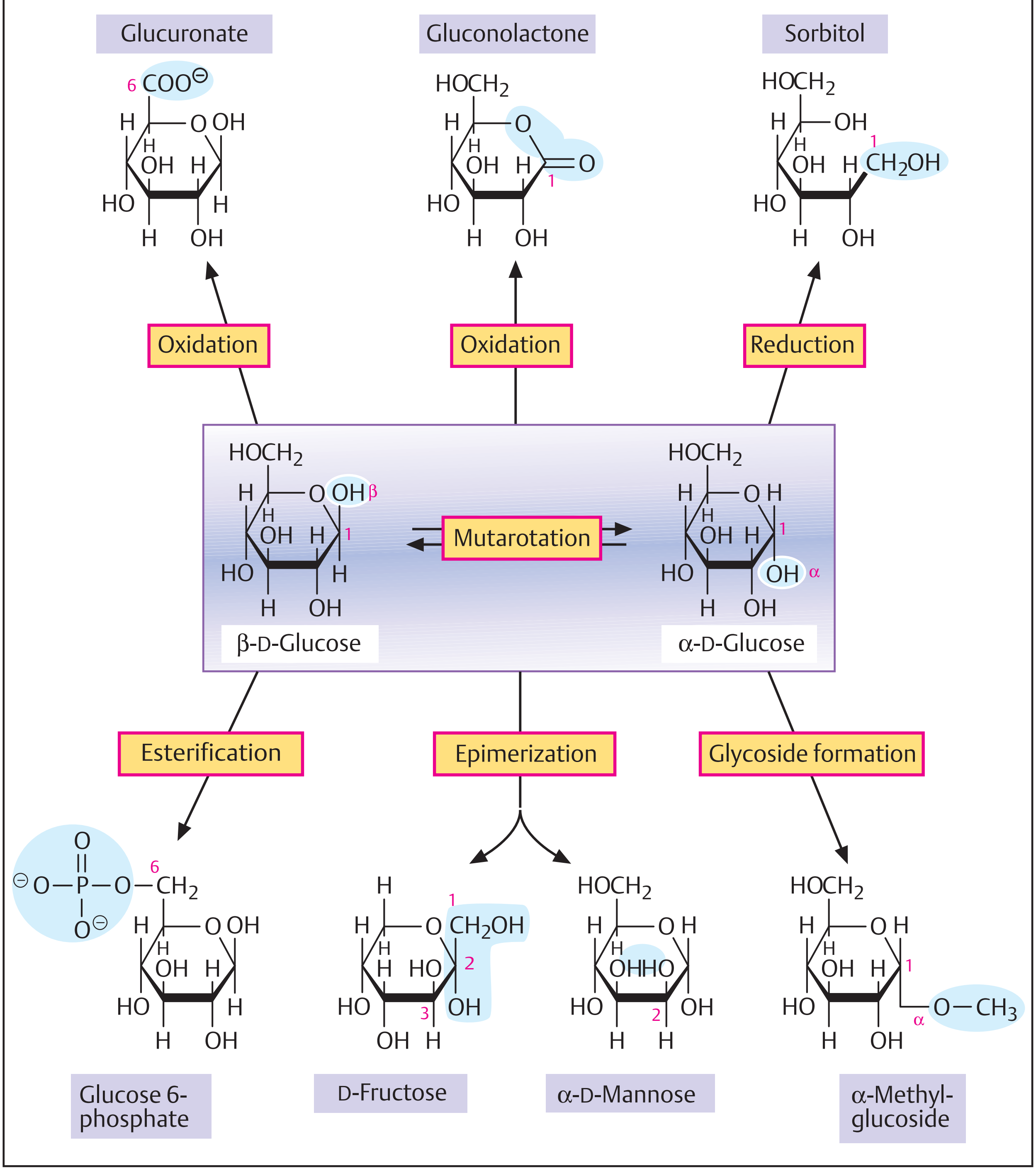
Glucuronate
6 COO⊖
H O OH H OH H HO H H OH
Gluconolactone
HOCH2
H O H OH H O 1 HO H OH
Sorbitol
HOCH2
H OH H OH H 1 CH2OH HO H OH
Oxidation
Oxidation
Reduction
HOCH2
H O OH β H OH H 1 HO H H OH
β-D-Glucose
Mutarotation
HOCH2
H O H H OH H 1 HO OH α H OH
α-D-Glucose
Esterification
Epimerization
Glycoside formation
O
⊖ O P O 6 CH2
O⊖
H O OH H OH H HO H H OH
Glucose 6-phosphate
H
1
H O CH2OH H H HO 2 HO 3 OH OH H
D-Fructose
HOCH2
H O H H OHHO HO 2 OH H H
α-D-Mannose
HOCH2
H O H H OH H 1 HO α O CH3 H OH
α-Methyl-glucoside

B. Polarimetry, mutarotation

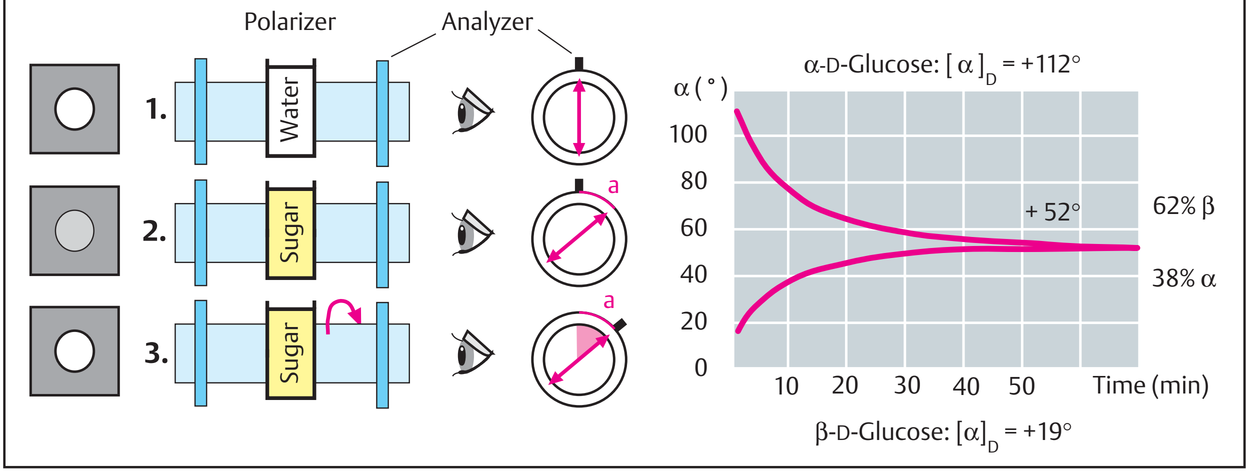

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Monosaccharides and disaccharides

A. Important monosaccharides ◑

Only the most important of the large number of naturally occurring **monosaccharides** are mentioned here. They are classified according to the number of C atoms (into pentoses, hexoses, etc.) and according to the chemical nature of the carbonyl function into aldoses and ketoses.

The best-known **aldopentose** (**1**), D-*ribose*, is a component of RNA and of nucleotide coenzymes and is widely distributed. In these compounds, ribose always exists in the furanose form (see p. 34). Like ribose, D-*xylose* and L-*arabinose* are rarely found in free form. However, large amounts of both sugars are found as constituents of polysaccharides in the walls of plant cells (see p. 42).

The most important of the **aldohexoses** (**1**) is D-glucose. A substantial proportion of the biomass is accounted for by glucose polymers, above all cellulose and starch. Free D-glucose is found in plant juices ("grape sugar") and as "blood sugar" in the blood of higher animals. As a constituent of lactose (milk sugar), D-galactose is part of the human diet. Together with D-mannose, galactose is also found in glycolipids and glycoproteins (see p. 44).

Phosphoric acid esters of the **ketopentose** D-*ribulose* (**2**) are intermediates in the pentose phosphate pathway (see p. 152) and in photosynthesis (see p. 128). The most widely distributed of the **ketohexoses** is D-*fructose*. In free form, it is present in fruit juices and in honey. Bound fructose is found in sucrose (**B**) and plant polysaccharides (e. g., inulin).

In the **deoxyaldoses** (**3**), an OH group is replaced by a hydrogen atom. In addition to 2-*deoxy-D-ribose,* a component of DNA (see p. 84) that is reduced at C-2, L-fucose is shown as another example of these. Fucose, a sugar in the λ series (see p. 34) is reduced at C-6.

The **acetylated amino sugars** *N-acetyl-D-glucosamine* and *N-acetyl-D-Galactosamine* (**4**) are often encountered as components of glycoproteins.

N-acetylneuraminic acid (sialic acid, **5**), is a characteristic component of glycoproteins. Other **acidic monosaccharides** such as D-*glucuronic acid*, D-*galacturonic acid*, and *liduronic acid*, are typical constituents of the glycosaminoglycans found in connective tissue.

Sugar alcohols (**6**) such as *sorbitol* and *mannitol* do not play an important role in animal metabolism.

B. Disaccharides ◑

When the anomeric hydroxyl group of one monosaccharide is bound glycosidically with one of the OH groups of another, a **disaccharide** is formed. As in all glycosides, the glycosidic bond does *not* allow mutarotation. Since this type of bond is formed stereospecifically by enzymes in natural disaccharides, they are only found in *one* of the possible configurations (α or β).

Maltose (**1**) occurs as a breakdown product of the starches contained in malt ("malt sugar"; see p. 148) and as an intermediate in intestinal digestion. In maltose, the anomeric OH group of one glucose molecule has an α-glycosidic bond with C-4 in a second glucose residue.

Lactose ("milk sugar," **2**) is the most important carbohydrate in the milk of mammals. Cow's milk contains 4.5% lactose, while human milk contains up to 7.5%. In lactose, the anomeric OH group of galactose forms a β-glycosidic bond with C-4 of a glucose. The lactose molecule is consequently elongated, and both of its pyran rings lie in the same plane.

Sucrose (**3**) serves in plants as the form in which carbohydrates are transported, and as a soluble carbohydrate reserve. Humans value it because of its intensely sweet taste. Sources used for sucrose are plants that contain particularly high amounts of it, such as sugar cane and sugar beet (*cane sugar, beet sugar*). Enzymatic hydrolysis of sucrose-containing flower nectar in the digestive tract of bees—catalyzed by the enzyme *invertase*—produces **honey**, a mixture of glucose and fructose. In sucrose, the two anomeric OH groups of glucose and fructose have a glycosidic bond; sucrose is therefore one of the non-reducing sugars.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Important monosaccharides

① Aldoses

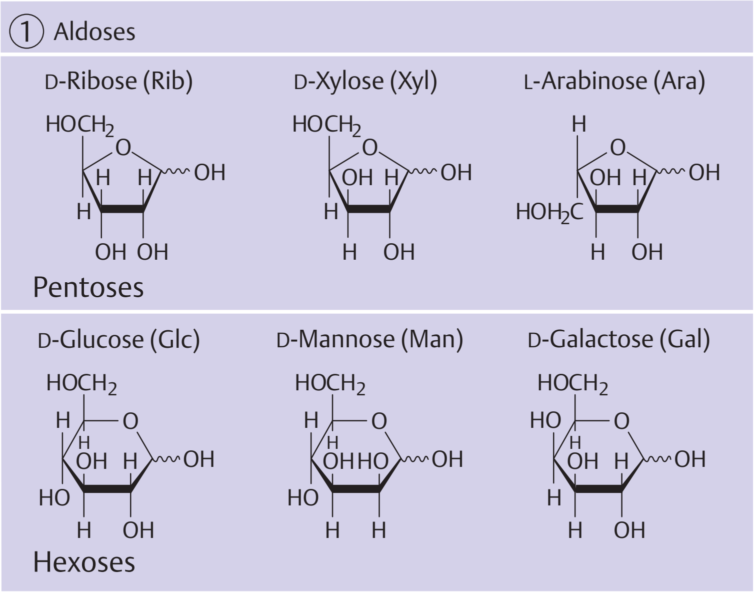
D-Ribose (Rib)
D-Xylose (Xyl)
L-Arabinose (Ara)
Pentoses
D-Glucose (Glc)
D-Mannose (Man)
D-Galactose (Gal)
Hexoses

② Ketoses

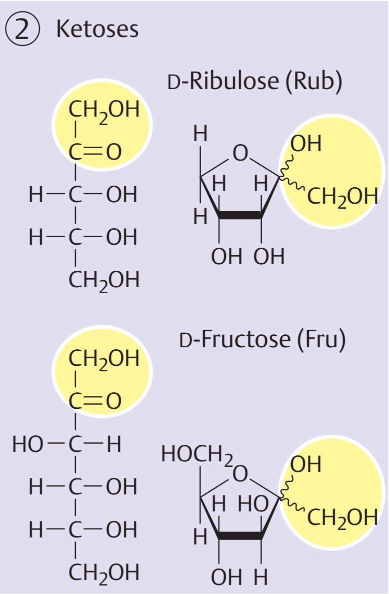
D-Ribulose (Rub)
D-Fructose (Fru)

③ Deoxyaldoses

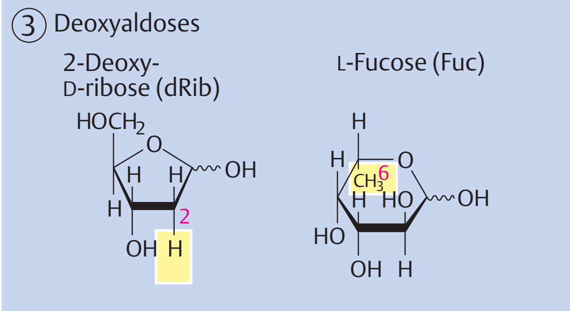
2-Deoxy-D-ribose (dRib)
L-Fucose (Fuc)

④ Acetylated amino sugars

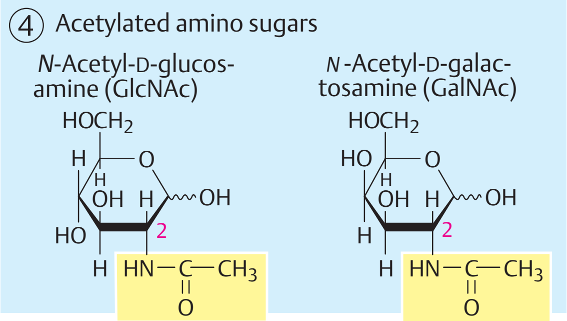
N-Acetyl-D-glucosamine (GlcNAc)
N-Acetyl-D-galactosamine (GalNAc)

⑤ Acidic monosaccharides

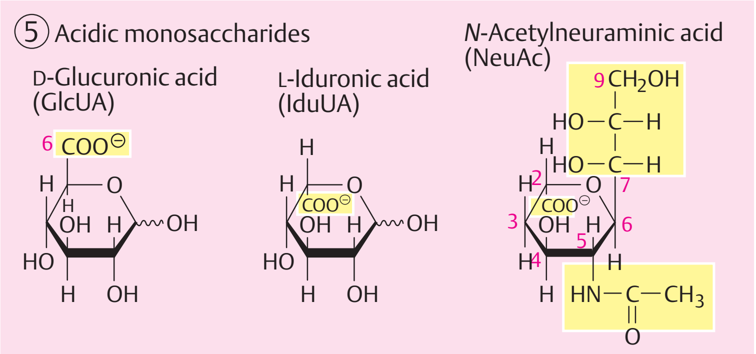
D-Glucuronic acid (GlcUA)
L-Iduronic acid (IduUA)
N-Acetylneuraminic acid (NeuAc)

⑥ Sugar alcohols (alditoles)

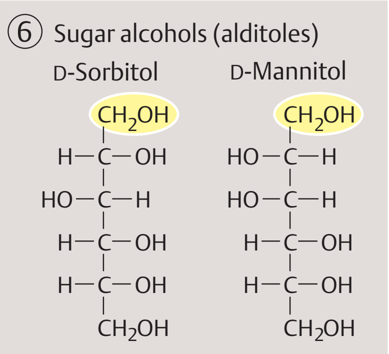
D-Sorbitol
D-Mannitol

B. Disaccharides

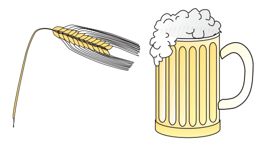

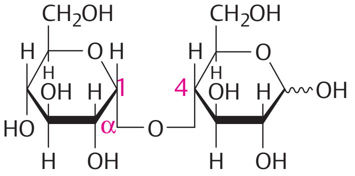

1. Maltose
α-D-Glucopyranosyl-
(1→4)-D-glucopyranose

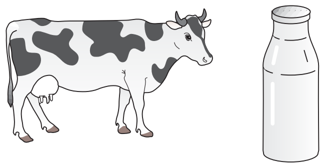

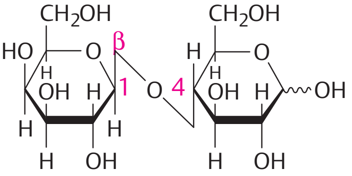

2. Lactose
β-D-Galactopyranosyl-
(1→4)-D-glucopyranose

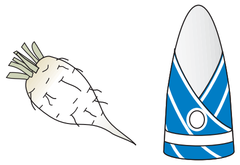

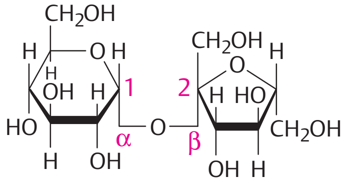

3. Sucrose
α-D-Glucopyranosyl-
(1↔2)-β-D-fructofuranoside

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Polysaccharides: overview

Polysaccharides are ubiquitous in nature. They can be classified into three separate groups, based on their different *functions.* **Structural polysaccharides** provide mechanical stability to cells, organs, and organisms. **Waterbinding polysaccharides** are strongly hydrated and prevent cells and tissues from drying out. Finally, **reserve polysaccharides** serve as carbohydrate stores that release monosaccharides as required. Due to their polymeric nature, reserve carbohydrates are osmotically less active, and they can therefore be stored in large quantities within the cell.

A. Polysaccharides: structure ◑

Polysaccharides that are formed from only *one* type of monosaccharide are called **homoglycans**, while those formed from different sugar constituents are called **heteroglycans**. Both forms can exist as either linear or branched chains.

A section of a **glycogen** molecule is shown here as an example of a branched homoglycan. Amylopectin, the branched component of vegetable starch (see p. 42), has a very similar structure. Both molecules mainly consist of α1→4-linked glucose residues. In glycogen, on average every 8th to 10th residue carries —via an α1→6 bond—another 1,4-linked chain of glucose residues. This gives rise to branched, tree-like structures, which in animal glycogen are covalently bound to a protein, *glycogenin* (see p. 156).

The linear heteroglycan **murein**, a structural polysaccharide that stabilizes the cell walls of bacteria, has a more complex structure. Only a short segment of this thread-like molecule is shown here. In murein, two different components, both β1→4-linked, alternate: *N-acetylglucosamine* (GlcNAc) and *N-acetylmuraminic acid* (MurNAc), a lactic acid ether of *N*-acetylglucosamine. *Peptides* are bound to the carboxyl group of the lactyl groups, and attach the individual strands of murein to each other to form a three-dimensional network (not shown). Synthesis of the network-forming peptides in murein is inhibited by *penicillin* (see p. 254).

B. Important polysaccharides ◑

The table gives an overview of the composition and make-up both of the glycans mentioned above and of several more.

In addition to murein, bacterial polysaccharides include **dextrans**—glucose polymers that are mostly α1→6-linked and α1→3-branched. In water, dextrans form viscous slimes or gels that are used for chromatographic separation of macromolecules after chemical treatment (see p. 78). Dextrans are also used as components of blood plasma substitutes (plasma expanders) and foodstuffs.

Carbohydrates from algae (e. g., **agarose** and **carrageenan**) can also be used to produce gels. Agarose has been used in microbiology for more than 100 years to reinforce culture media ("agar-agar"). Algal polysaccharides are also added to cosmetics and ready-made foods to modify the consistency of these products.

The **starches**, the most important vegetable reserve carbohydrate and polysaccharides from plant cell walls, are discussed in greater detail on the following page. **Inulin**, a fructose polymer, is used as a starch substitute in diabetics' dietary products (see p. 160). In addition, it serves as a test substance for measuring renal clearance (see p. 322).

Chitin, a homopolymer from β1→4-linked *N*-acetylglucosamine, is the most important structural substance in insect and crustacean shells, and is thus the most common animal polysaccharide. It also occurs in the cell wall of fungi.

Glycogen, the reserve carbohydrate of higher animals, is stored in the liver and musculature in particular (**A**, see pp. 156, 336). The formation and breakdown of glycogen are subject to complex regulation by hormones and other factors (see p. 120).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Polysaccharides: structure

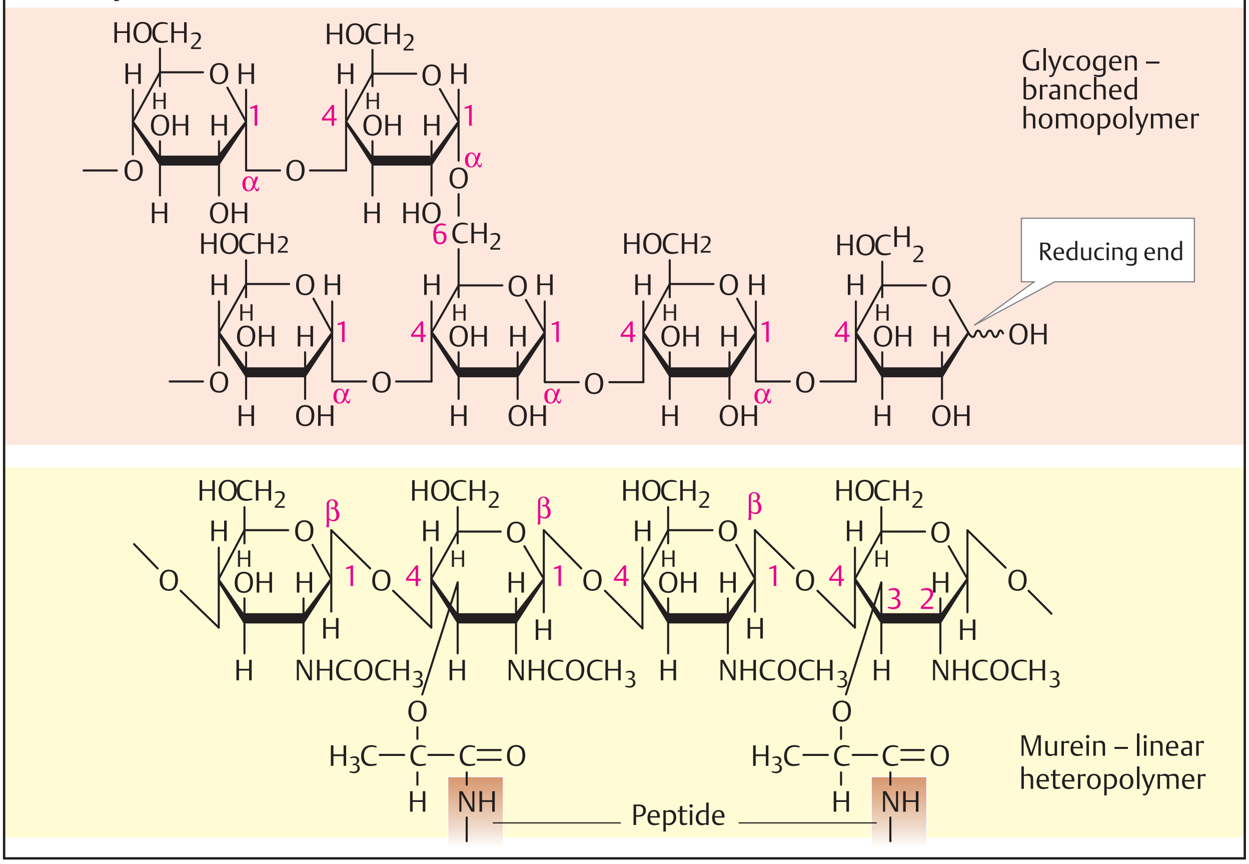
Glycogen – branched homopolymer
Reducing end
Murein – linear heteropolymer
Peptide

B. Important polysaccharides

Poly-saccharide	Mono-saccharide 1	Mono-saccharide 2	Linkage	Branch-ing	Occurrence	Function
Bacteria						
Murein	D-GlcNAc	D-MurNAc[1]	β1→4	—	Cell wall	SC
Dextran	D-Glc	—	α1→6	α1→3	Slime	WB
Plants						
Agarose	D-Gal	L-aGal[2]	β1→4	β1→3	Red algae (agar)	WB
Carrageenan	D-Gal	—	β1→3	α1→4	Red algae	WB
Cellulose	D-Glc	—	β1→4	—	Cell wall	SC
Xyloglucan	D-Glc	D-Xyl (D-Gal, L-Fuc)	β1→4	β1→6 (β1→2)	Cell wall (Hemicellulose)	SC
Arabinan	L-Ara	—	α1→5	α1→3	Cell wall (pectin)	SC
Amylose	D-Glc	—	α1→4	—	Amyloplasts	RC
Amylopectin	D-Glc	—	α1→4	α 1→6	Amyloplasts	RC
Inulin	D-Fru		β2→1	—	Storage cells	RC
Animals						
Chitin	D-GlcNAc	—	β1→4	—	Insects, crabs	SK
Glycogen	D-Glc	—	α1→4	α1→6	Liver, muscle	RK
Hyaluronic acid	D-GlcUA	D-GlcNAc	β1→4 β1→3	—	Connective tissue	SK,WB

SC= structural carbohydrate, RC= reserve carbohydrate, WB = water-binding carbohydrate; [1] *N*-acetylmuramic acid, [2] 3,6-anhydrogalactose

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Plant polysaccharides

Two glucose polymers of plant origin are of special importance among the polysaccharides: β1→4-linked polymer **cellulose** and **starch**, which is mostly α1→4-linked.

A. Cellulose ◑

Cellulose, a linear homoglycan of β1→4-linked glucose residues, is the *most abundant organic substance* in nature. Almost half of the total biomass consists of cellulose. Some 40–50% of plant *cell walls* are formed by cellulose. The proportion of cellulose in *cotton fibers,* an important raw material, is 98%. Cellulose molecules can contain more than 10^4 glucose residues (mass 1–2 10^6 Da) and can reach lengths of 6–8 μm.

Naturally occurring cellulose is *extremely mechanically stable* and is highly *resistant* to chemical and enzymatic hydrolysis. These properties are due to the conformation of the molecules and their supramolecular organization. The unbranched β1→4 linkage results in linear chains that are stabilized by hydrogen bonds within the chain and between neighboring chains (**1**). Already during biosynthesis, 50–100 cellulose molecules associate to form an **elementary fibril** with a diameter of 4 nm. About 20 such elementary fibrils then form a **microfibril** (**2**), which is readily visible with the electron microscope.

Cellulose microfibrils make up the basic framework of the **primary wall** of young plant cells (**3**), where they form a complex network with other polysaccharides. The linking polysaccharides include **hemicellulose**, which is a mixture of predominantly neutral heteroglycans (xylans, xyloglucans, arabinogalactans, etc.). Hemicellulose associates with the cellulose fibrils via noncovalent interactions. These complexes are connected by neutral and acidic **pectins**, which typically contain galacturonic acid. Finally, a collagen-related protein, **extensin**, is also involved in the formation of primary walls.

In the higher animals, including humans, cellulose is **indigestible**, but important as **roughage** (see p. 273). Many herbivores (e. g., the ruminants) have symbiotic unicellular organisms in their digestive tracts that break down cellulose and make it digestible by the host.

B. Starch ◑

Starch, a **reserve polysaccharide** widely distributed in plants, is the *most important carbohydrate in the human diet.* In plants, starch is present in the chloroplasts in leaves, as well as in fruits, seeds, and tubers. The starch content is especially high in cereal grains (up to 75% of the dry weight), potato tubers (approximately 65%), and in other plant storage organs.

In these plant organs, starch is present in the form of microscopically small granules in special organelles known as **amyloplasts**. *Starch granules* are virtually insoluble in cold water, but swell dramatically when the water is heated. Some 15–25% of the starch goes into solution in colloidal form when the mixture is subjected to prolonged boiling. This proportion is called amylose ("soluble starch").

Amylose consists of *unbranched* α1→4-linked chains of 200–300 glucose residues. Due the α configuration at C-1, these chains form a *helix* with 6–8 residues per turn (**1**). The blue coloring that soluble starch takes on when iodine is added (the "iodine–starch reaction") is caused by the presence of these helices—the iodine atoms form chains inside the amylose helix, and in this largely nonaqueous environment take on a deep blue color. Highly branched polysaccharides turn brown or reddishbrown in the presence of iodine.

Unlike amylose, **amylopectin**, which is practically insoluble, is *branched.* On average, one in 20–25 glucose residues is linked to another chain via an α1→6 bond. This leads to an extended tree-like structure, which—like amylose—contains only *one* anomeric OH group (a "reducing end"). Amylopectin molecules can contain hundreds of thousands of glucose residues; their mass can be more than 10^8 Da.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

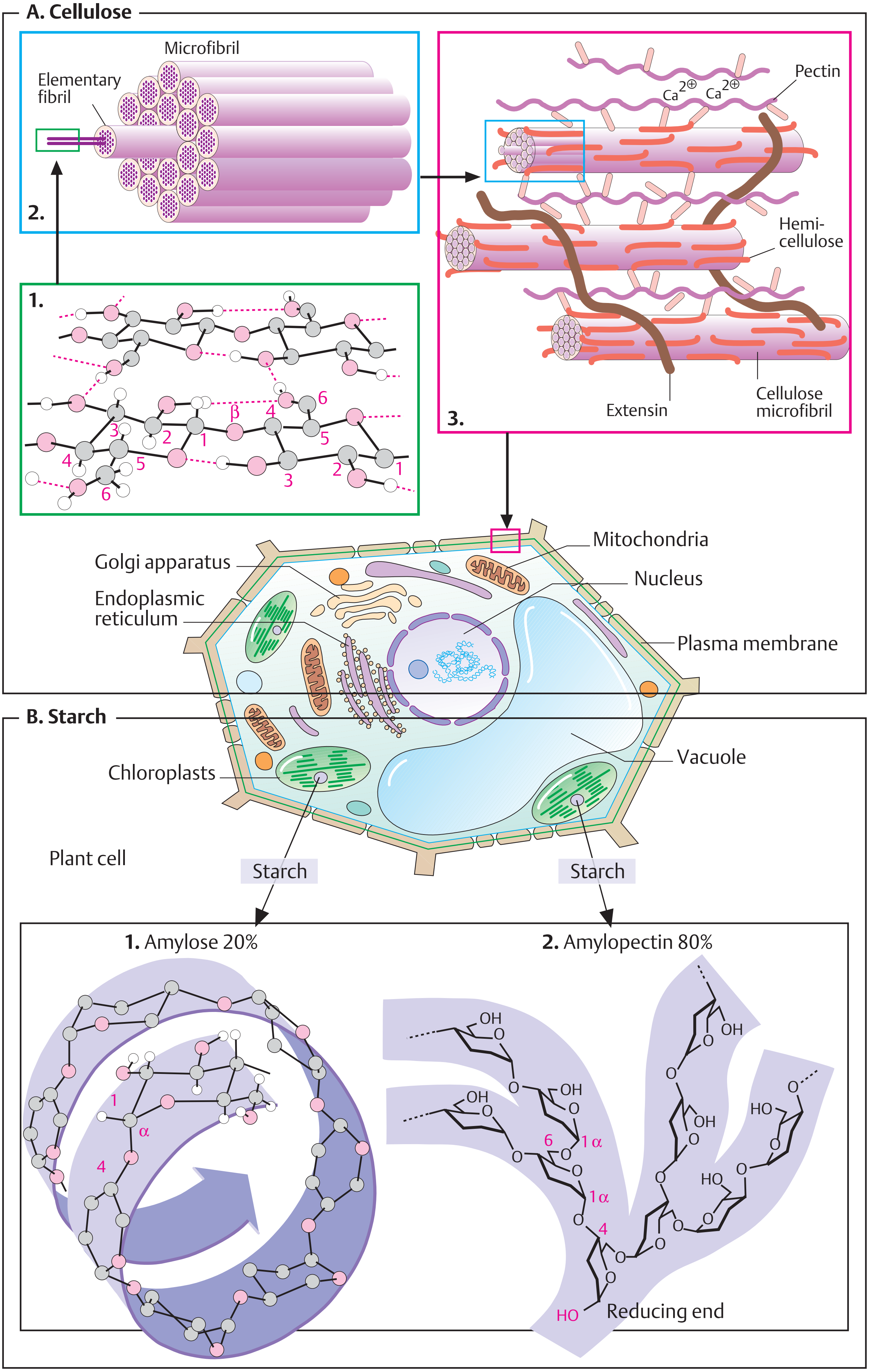
A. Cellulose
Microfibril
Elementary fibril
2.
1.
β
1
2
3
4
5
6
Pectin
$Ca^{2⊕}$
$Ca^{2⊕}$
Hemi-cellulose
Extensin
Cellulose microfibril
3.
Golgi apparatus
Endoplasmic reticulum
Mitochondria
Nucleus
Plasma membrane
B. Starch
Chloroplasts
Vacuole
Plant cell
Starch
Starch
1. Amylose 20%
α
1
4
2. Amylopectin 80%
OH
O
6
1 α
4
HO
Reducing end

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Glycosaminoglycans and glycoproteins

A. Hyaluronic acid ○

As constituents of proteoglycans (see p. 346), the glycosaminoglycans—a group of acidic heteropolysaccharides—are important structural elements of the extracellular matrix.

Glycosaminoglycans contain *amino sugars* as well as *glucuronic acid* and *iduronic acid* as characteristic components (see p. 38). In addition, most polysaccharides in this group are esterified to varying extents by sulfuric acid, increasing their acidic quality. Glycosaminoglycans can be found in free form, or as components of proteoglycans throughout the organism.

Hyaluronic acid, an unesterified glycosaminoglycan with a relatively simple structure, consists of disaccharide units in which *N-acetylglucosamine* and *glucuronic acid* are alternately β1→4-linked and β1→3-linked. Due to the unusual β1→3 linkage, hyaluronic acid molecules–which may contain several thousand monosaccharide residues—are coiled like a helix. Three disaccharide units form each turn of the helix. The outwardfacing hydrophilic carboxylate groups of the glucuronic acid residues are able to bind Ca^{2+} ions. The **strong hydration** of these groups enables hyaluronic acid and other glycosaminoglycans to bind water up to 10 000 times their own volume in gel form. This is the function which hyaluronic acid has in the *vitreous body* of the eye, which contains approximately 1% hyaluronic acid and 98% water.

B. Oligosaccharide in immunoglobulin G (IgG) ○

Many proteins on the surface of the plasma membrane, and the majority of secreted proteins, contain oligosaccharide residues that are post-translationally added to the endoplasmic reticulum and in the Golgi apparatus (see p. 230). By contrast, cytoplasmic proteins are rarely glycosylated. **Glycoproteins** can contain more than 50% carbohydrate; however, the proportion of protein is generally much greater.

As an example of the carbohydrate component of a glycoprotein, the structure of one of the oligosaccharide chains of immunoglobulin G (IgG; see p. 300) is shown here. The oligosaccharide has an *N-glycosidic* link to the amide group of an asparagine residue in the F_c part of the protein. Its function is not known.

Like all *N*-linked carbohydrates, the oligosaccharide in IgG contains a T-shaped **core structure** consisting of two *N-acetylglucosamines* and three *mannose* residues (shown in violet). In addition, in this case the structure contains two further *N-acetylglucosamine* residues, as well as a *fucose* residue and a *galactose* residue. Glycoproteins show many different types of branching. In this case, we not only have β1→4 linkage, but also β1→2, α1→3, and α1→6 bonds.

C. Glycoproteins: forms ◑

On the cell surface of certain glycoproteins, **O-glycosidic** links are found between the carbohydrate part and a serine or threonine residue, instead of *N*-glycosidic links to asparagine residues. This type of link is less common than the ***N*-glycosidic** one.

There are two types of oligosaccharide structure with *N*-glycosidic links, which arise through two different biosynthetic pathways. During glycosylation in the ER, the protein is initially linked to an oligosaccharide, which in addition to the core structure contains six further mannose residues and three terminal glucose residues (see p. 230). The simpler from of oligosaccharide (the **mannose-rich** type) is produced when only the glucose residues are cleaved from the primary product, and no additional residues are added. In other cases, the mannose residues that are located outside the core structure are also removed and replaced by other sugars. This produces oligosaccharides such as those shown on the right (the **complex type**). At the external end of the structure, glycoproteins of the complex type often contain *N*-acetylneuraminic acid residues, which give the oligosaccharide components negative charges.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Hyaluronic acid

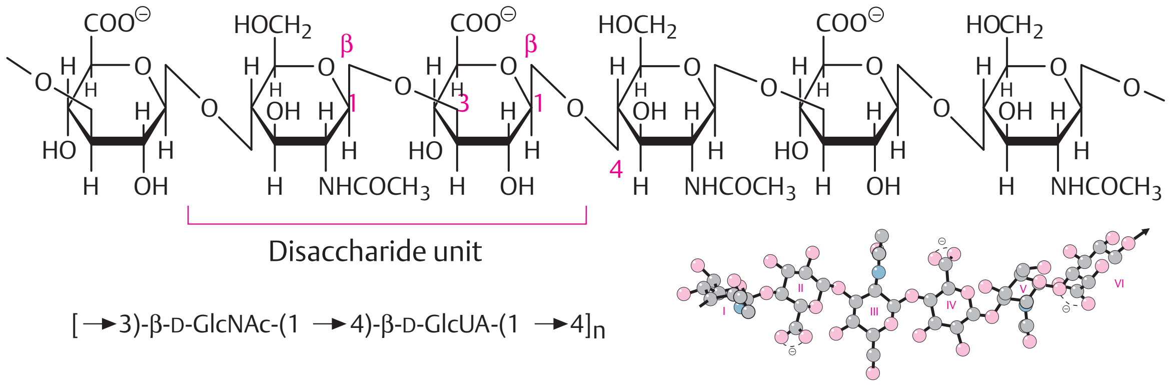
COO⊖
HOCH2
β
1
3
4
O
H
OH
HO
NHCOCH3
Disaccharide unit
[→3)-β-D-GlcNAc-(1 →4)-β-D-GlcUA-(1 →4]n
I
II
III
IV
V
VI

B. Oligosaccharide in immunoglobulin G (IgG)

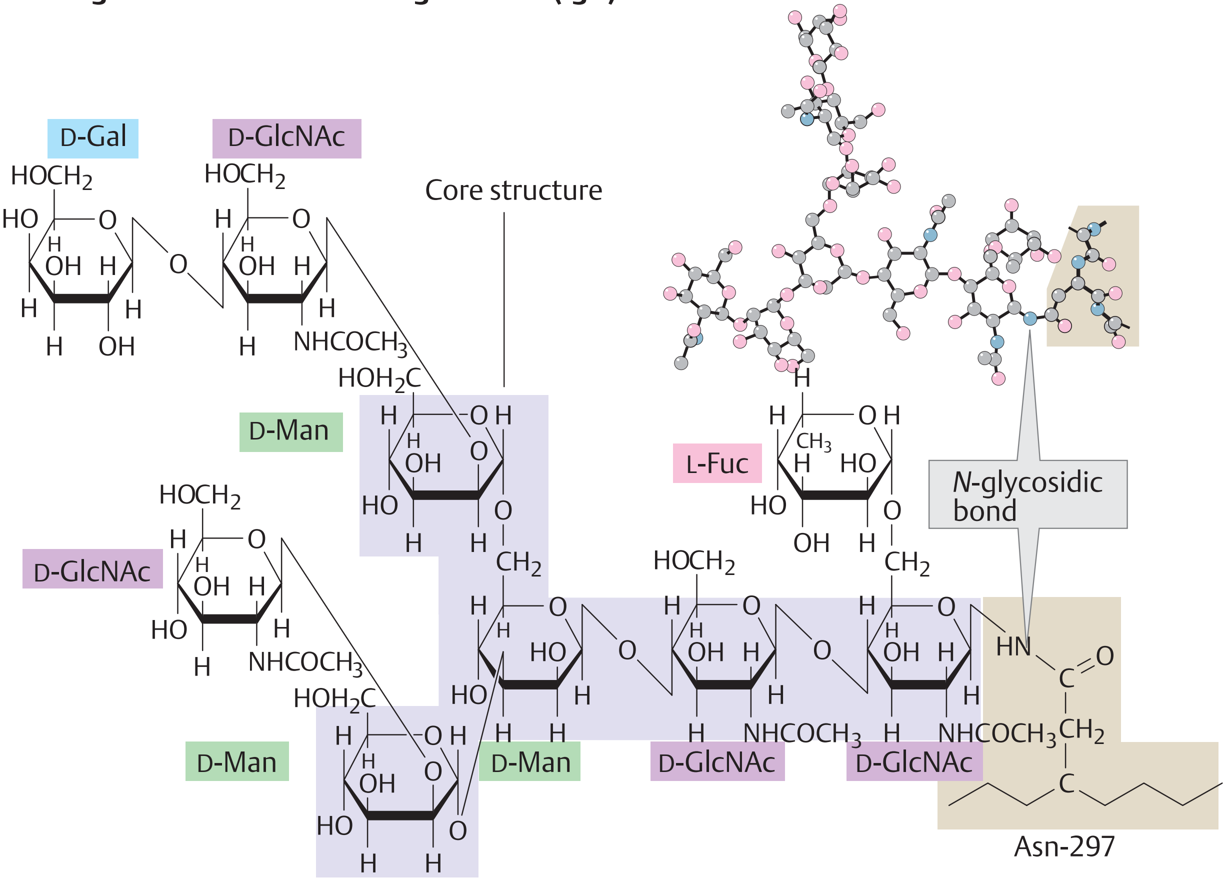
D-Gal
D-GlcNAc
Core structure
HOCH2
HOH2C
D-Man
D-GlcNAc
D-Man
D-Man
D-GlcNAc
D-GlcNAc
L-Fuc
CH3
CH2
NHCOCH3
N-glycosidic bond
HN
C=O
C
Asn-297

C. Glycoproteins: forms

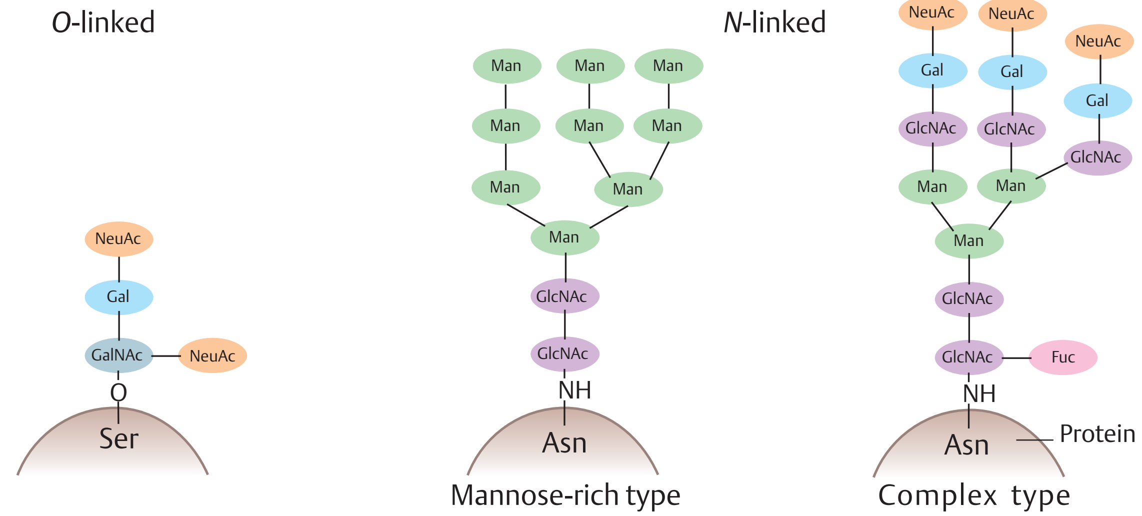
O-linked
N-linked
NeuAc
Gal
GalNAc
NeuAc
O
Ser
Man
GlcNAc
NH
Asn
Fuc
Protein
Mannose-rich type
Complex type

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Overview

A. Classification ●

The **lipids** are a large and heterogeneous group of substances of biological origin that are easily dissolved in organic solvents such as methanol, acetone, chloroform, and benzene. By contrast, they are either insoluble or only poorly soluble in water. Their low water solubility is due to a lack of polarizing atoms such as O, N, S, and P (see p. 6).

Lipids can be classified into substances that are either *hydrolyzable*— i. e., able to undergo hydrolytic cleavage—or *nonhydrolyzable*. Only a few examples of the many lipids known can be mentioned here. The individual classes of lipids are discussed in more detail in the following pages.

Hydrolyzable lipids (components shown in brackets). The simple **esters** include the *fats* (triacylglycerol; one glycerol + three acyl residues); the *waxes* (one fatty alcohol + one acyl residue); and the *sterol esters* (one sterol + one acyl residue). The **phospholipids** are esters with more complex structures. Their characteristic component is a phosphate residue. The phospholipids include the *phosphatidic acids* (one glycerol + two acyl residues + one phosphate) and the *phosphatides* (one glycerol + two acyl residues + one phosphate + one amino alcohol). In the **sphingolipids**, glycerol and one acyl residue are replaced by sphingosine. Particularly important in this group are the sugar-containing **glycolipids** (one sphingosine + one fatty acid + sugar). The *cerebrosides* (one sphingosine + one fatty acid + one sugar) and *gangliosides* (one sphingosine + one fatty acid + several different sugars, including neuraminic acid) are representatives of this group.

The components of the hydrolyzable lipids are linked to one another by **ester bonds**. They are easily broken down either enzymatically or chemically.

Non-hydrolyzable lipids. The **hydrocarbons** include the *alkanes* and *carotenoids*. The **lipid alcohols** are also not hydrolyzable. They include long-chained *alkanols* and cyclic *sterols* such as cholesterol, and *steroids* such as estradiol and testosterone. The most important **acids** among the lipids are *fatty acids*. The *eicosanoids* also belong to this group; these are derivatives of the polyunsaturated fatty acid arachidonic acid (see p. 390).

B. Biological roles ●

1. Fuel. Lipids are an important source of energy in the diet. In quantitative terms, they represent the principal energy reserve in animals. Neutral fats in particular are stored in specialized cells, known as *adipocytes*. Fatty acids are released from these again as needed, and these are then oxidized in the mitochondria to form water and carbon dioxide, with oxygen being consumed. This process also gives rise to reduced coenzymes, which are used for ATP production in the respiratory chain (see p. 140).

2. Nutrients. Amphipathic lipids are used by cells to build membranes (see p. 214). Typical membrane lipids include phospholipids, glycolipids, and cholesterol. Fats are only weakly amphiphilic and are therefore not suitable as membrane components.

3. Insulation. Lipids are excellent insulators. In the higher animals, neutral fats are found in the subcutaneous tissue and around various organs, where they serve as mechanical and thermal insulators. As the principal constituent of cell membranes, lipids also insulate cells from their environment mechanically and electrically. The impermeability of lipid membranes to ions allows the formation of the membrane potential (see p. 126).

4. Special tasks. Some lipids have adopted special roles in the body. Steroids, eicosanoids, and some metabolites of phospholipids have *signaling functions*. They serve as hormones, mediators, and second messengers (see p. 370). Other lipids form *anchors* to attach proteins to membranes (see p. 214). The lipids also produce *cofactors for enzymatic reactions*—e. g., vitamin K (see p. 52) and ubiquinone (see p. 104). The carotenoid retinal, a light-sensitive lipid, is of central importance in the *process of vision* (see p. 358).

Several lipids are not formed independently in the human body. These substances, as **essential fatty acids** and **fat-soluble vitamins**, are indispensable components of nutrition (see pp. 364ff.)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Classification

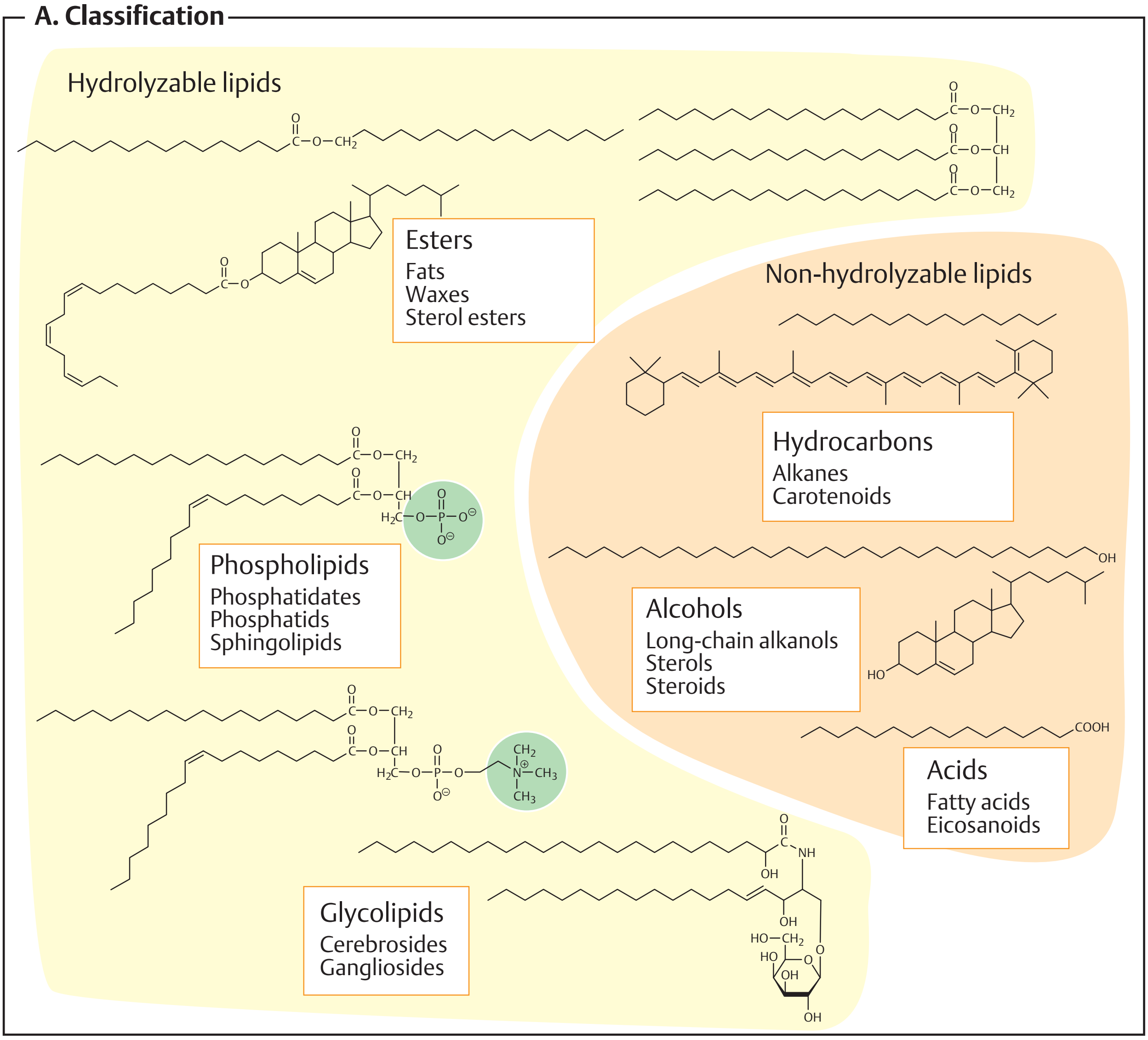
Hydrolyzable lipids
Esters
Fats
Waxes
Sterol esters
Phospholipids
Phosphatidates
Phosphatids
Sphingolipids
Glycolipids
Cerebrosides
Gangliosides
Non-hydrolyzable lipids
Hydrocarbons
Alkanes
Carotenoids
Alcohols
Long-chain alkanols
Sterols
Steroids
Acids
Fatty acids
Eicosanoids
C–O–CH2
C–O–CH
C–O–CH2
H2C–O–P–O⊖
O⊖
N–CH3
CH3
NH
OH
HO–CH2
HO
COOH

B. Biological roles

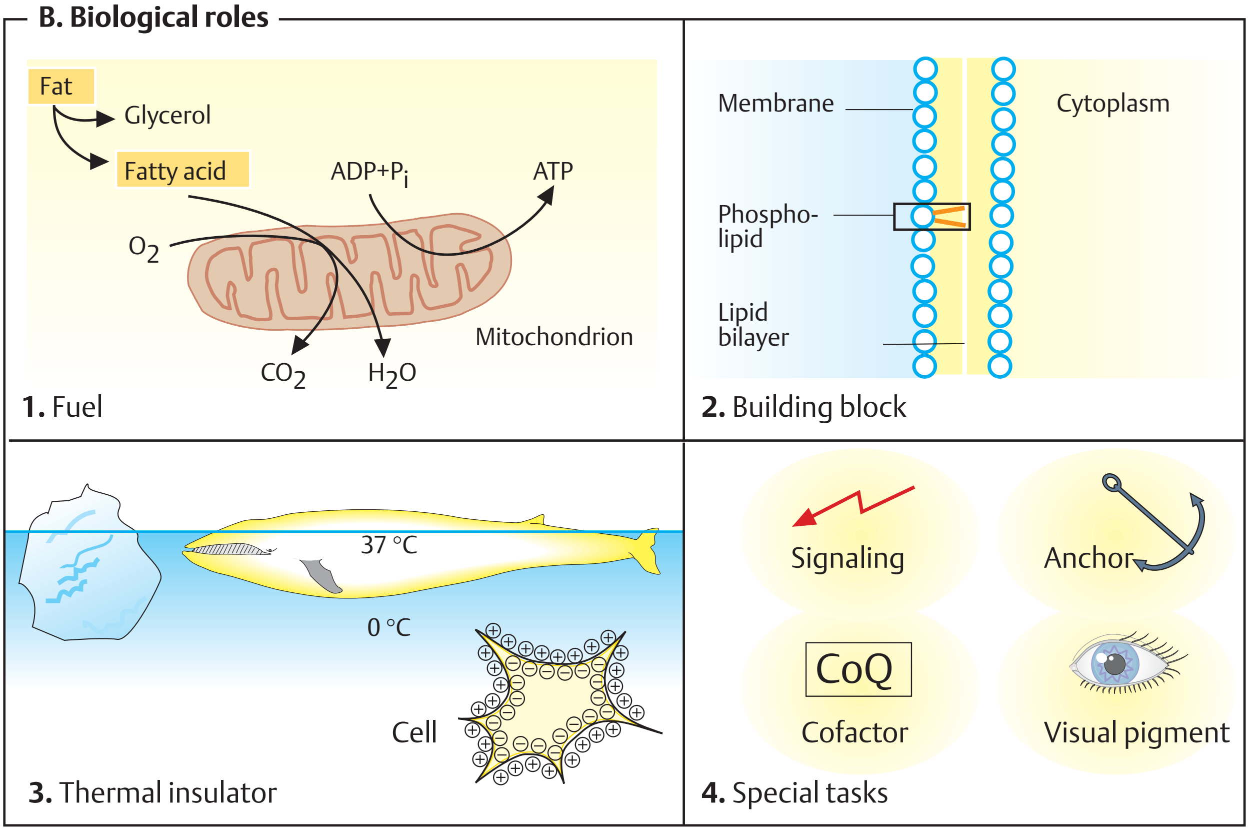
Fat
Glycerol
Fatty acid
ADP+Pi
ATP
O2
Mitochondrion
CO2
H2O
1. Fuel
Membrane
Cytoplasm
Phospho-lipid
Lipid bilayer
2. Building block
37 °C
0 °C
Cell
3. Thermal insulator
Signaling
Anchor
CoQ
Cofactor
Visual pigment
4. Special tasks

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Fatty acids and fats

A. Carboxylic acids ◑

The naturally occurring **fatty acids** are carboxylic acids with unbranched hydrocarbon chains of 4–24 carbon atoms. They are present in all organisms as components of fats and membrane lipids. In these compounds, they are esterified with alcohols (glycerol, sphingosine, or cholesterol). However, fatty acids are also found in small amounts in unesterified form. In this case, they are known as *free fatty acids* (FFAs). As free fatty acids have strongly amphipathic properties (see p. 28), they are usually present in protein-bound forms.

The table lists the full series of aliphatic carboxylic acids that are found in plants and animals. In higher plants and animals, unbranched, longchain fatty acids with either 16 or 18 carbon atoms are the most common—e. g., palmitic and stearic acid. The number of carbon atoms in the longer, natural fatty acids is always even. This is because they are biosynthesized from C_2 building blocks (see p. 168).

Some fatty acids contain one or more isolated *double bonds*, and are therefore *“unsaturated.”* Common **unsaturated fatty acids** include oleic acid and linoleic acid. Of the two possible *cis–trans* isomers (see p. 8), usually only the *cis* forms are found in natural lipids. Branched fatty acids only occur in bacteria. A shorthand notation with several numbers is used for precise characterization of the structure of fatty acids—e g., 18:2;9,12 for linoleic acid. The first figure stands for the number of C atoms, while the second gives the number of double bonds. The positions of the double bonds follow after the semicolon. As usual, numbering starts at the carbon with the highest oxidation state (i. e., the carboxyl group corresponds to C-1). Greek letters are also commonly used (α = C-2; β = C-3; ω = the last carbon, ω-3 = the third last carbon).

Essential fatty acids are fatty acids that have to be supplied in the diet. Without exception, these are all polyunsaturated fatty acids: the C_{20} fatty acid *arachidonic acid* (20:4;5,8,11,14) and the two C_{18} acids *linoleic acid* (18:2;9,12) and *linolenic acid* (18:3;9,12,15). The animal organism requires arachidonic acid to synthesize eicosanoids (see p. 390). As the organism is capable of elongating fatty acids by adding C_2 units, but is not able to introduce double bonds into the end sections of fatty acids (after C-9), arachidonic acid has to be supplied with the diet. Linoleic and linolenic acid can be converted into arachidonic acid by elongation, and they can therefore replace arachidonic acid in the diet.

B. Structure of fats ◑

Fats are esters of the trivalent alcohol *glycerol* with three fatty acids. When a single fatty acid is esterified with glycerol, the product is referred to as a *monoacylglycerol* (fatty acid residue = acyl residue).

Formally, esterification with additional fatty acids leads to *diacylglycerol* and ultimately to *triacylglycerol,* the actual fat (formerly termed “triglyceride”). As triacylglycerols are uncharged, they are also referred to as *neutral fats*. The carbon atoms of glycerol are not usually equivalent in fats. They are distinguished by their *“sn”* number, where sn stands for “stereospecific numbering.”

The three acyl residues of a fat molecule may differ in terms of their chain length and the number of double bonds they contain. This results in a large number of possible combinations of individual fat molecules. When extracted from biological materials, fats always represent mixtures of very similar compounds, which differ in their fatty acid residues. A chiral center can arise at the middle C atom (*sn* -C-2) of a triacylglycerol if the two external fatty acids are different. The monoacylglycerols and diacylglycerols shown here are also chiral compounds. Nutritional fats contain palmitic, stearic, oleic acid, and linoleic acid particularly often. Unsaturated fatty acids are usually found at the central C atom of glycerol.

The length of the fatty acid residues and the number of their double bonds affect the melting point of the fats. The shorter the fatty acid residues and the more double bonds they contain, the lower their melting points.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Carboxylic acids

Name	Number of carbons	Number of double bonds	Position of double bonds
Formic acid	1	0	
Acetic acid	2	0	
Propionic acid	3	0	
Butyric acid	4	0	
Valerianic acid	5	0	
Caproic acid	6	0	
Caprylic acid	8	0	
Capric acid	10	0	
Lauric acid	12	0	
Myristic acid	14	0	
Palmitic acid	16	0	
Stearic acid	18	0	
Oleic acid	18	1	9
✷ Linoleic acid	18	2	9,12
✷ Linolenic acid	18	3	9,12,15
Arachidic acid	20	0	
✷ Arachidonic acid	20	4	5,8,11,14
Behenic acid	22	0	
Erucic acid	22	1	13
Lignoceric acid	24	0	
Nervonic acid	24	1	15

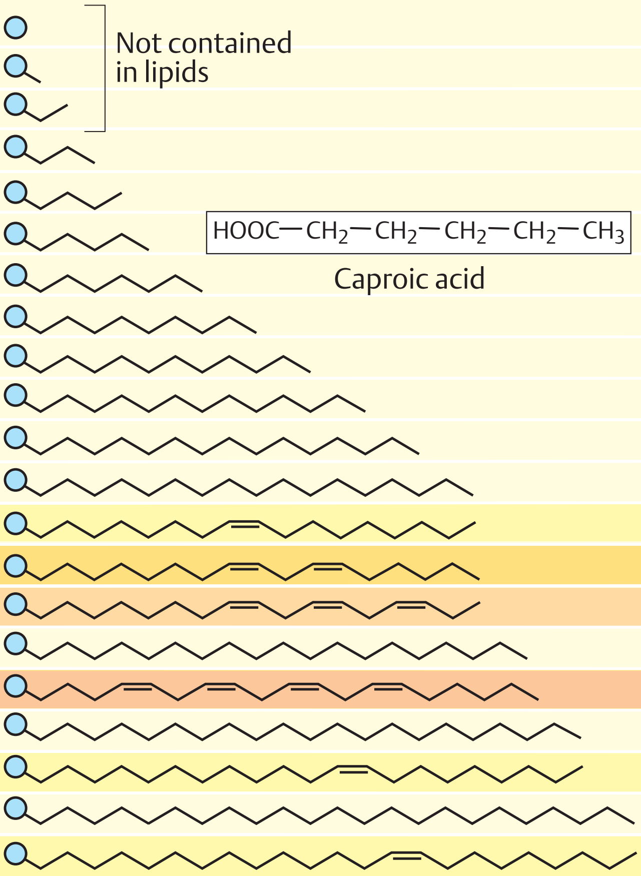
Not contained
in lipids

$HOOC{-}CH_2{-}CH_2{-}CH_2{-}CH_2{-}CH_3$

Caproic acid

✷ Essential in human nutrition

B. Structure of fats

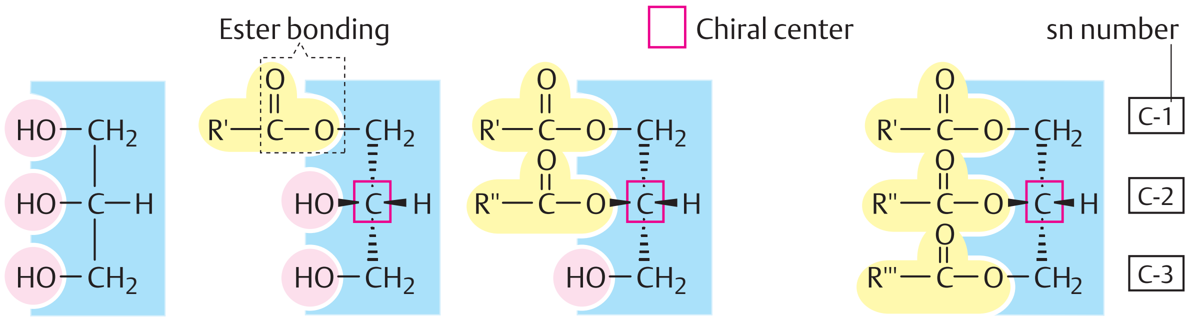
Ester bonding
Chiral center
sn number
HO
CH2
C
H
R'
O
R''
R'''
C-1
C-2
C-3

Glycerol Monoacylglycerol Diacylglycerol Triacylglycerol = Fat

Van der Waals model
of tristearylglycerol

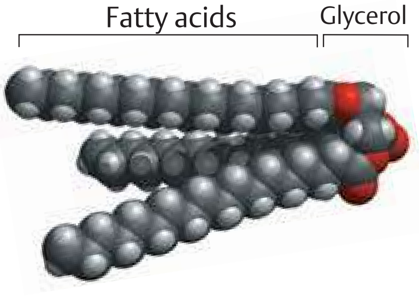
Fatty acids
Glycerol

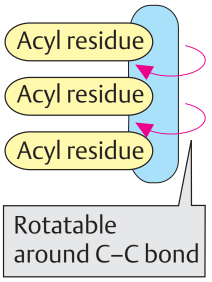
Acyl residue
Acyl residue
Acyl residue
Rotatable
around C–C bond

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Phospholipids and glycolipids

A. Structure of phospholipids and glycolipids ◑

Fats (triacylglycerol, **1**) are esters of glycerol with three fatty acids (see p. 48). Within the cell, they mainly occur as fat droplets. In the blood, they are transported in the hydrophobic interior of lipoproteins (see p. 278).

Phospholipids (**2**) are the main constituents of biological membranes (see pp. 214–217). Their common feature is a phosphate residue that is esterified with the hydroxyl group at C-3 of glycerol. Due to this residue, phospholipids have at least one negative charge at a neutral pH.

Phosphatidates (anions of the phosphatidic acids), the simplest phospholipids, are phosphate esters of diacylglycerol. They are important intermediates in the biosynthesis of fats and phospholipids (see p. 170). Phosphatidates can also be released from phospholipids by phospholipases.

The other phospholipids can be derived from phosphatidates (residue = phosphatidyl). Their phosphate residues are esterified with the hydroxyl group of an amino alcohol (*choline*, *ethanolamine*, or *serine*) or with the cyclohexane derivative *myo-inositol*. *Phosphatidylcholine* is shown here as an example of this type of compound. When two phosphatidyl residues are linked with one glycerol, the result is *cardiolipin* (not shown), a phospholipid that is characteristic of the inner mitochondrial membrane. *Lysophospholipids* arise from phospholipids by enzymatic cleavage of an acyl residue. The hemolytic effect of bee and snake venoms is due in part to this reaction.

Phosphatidylcholine (lecithin) is the most abundant phospholipid in membranes. *Phosphatidylethanolamine* (cephalin) has an ethanolamine residue instead of choline, and *phosphatidylserine* has a serine residue. In *phosphatidylinositol,* phosphatidate is esterified with the sugarlike cyclic polyalcohol *myo*-inositol. A doubly phosphorylated derivative of this phospholipid, phosphatidylinositol 4,5-bisphosphate, is a special component of membranes, which, by enzymatic cleavage, can give rise to two *second messengers*, diacylglycerol (DAG) and inositol 1,4,5trisphosphate ($InsP_3$; see p. 386).

Some phospholipids carry additional charges, in addition to the negative charge at the phosphate residue. In phosphatidylcholine and phosphatidylethanolamine, the *N*-atom of the amino alcohol is positively charged. As a whole, these two phosphatides therefore appear to be neutral. In contrast, phosphatidylserine—with one additional positive charge and one additional negative charge in the serine residue—and phosphatidylinositol (with no additional charge) have a negative net charge, due to the phosphate residue.

Sphingolipids (**3**), which are found in large quantities in the membranes of nerve cells in the brain and in neural tissues, have a slightly different structure from the other membrane lipids discussed so far. In sphingolipids, *sphingosine*, an amino alcohol with an unsaturated alkyl side chain, replaces glycerol and one of the acyl residues. When sphingosine forms an amide bond to a fatty acid, the compound is called *ceramide* (**3**). This is the precursor of the sphingolipids. *Sphingomyelin* (2)—the most important sphingolipid—has an additional phosphate residue with a choline group attached to it on the sphingosine, in addition to the fatty acid.

Glycolipids (**3**) are present in all tissues on the outer surface of the plasma membrane. They consist of sphingosine, a fatty acid, and an oligosaccharide residue, which can sometimes be quite large. The phosphate residue typical of phospholipids is absent. *Galactosylceramide* and *glucosylceramide* (known as cerebroside) are simple representatives of this group. Cerebrosides in which the sugar is esterified with sulfuric acid are known as *sulfatides*. *Gangliosides* are the most complex glycolipids. They constitute a large family of membrane lipids with receptor functions that are as yet largely unknown. A characteristic component of many gangliosides is *N*-acetylneuraminic acid (sialic acid; see p. 38).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Structure of fats, phospholipids, and glycolipids

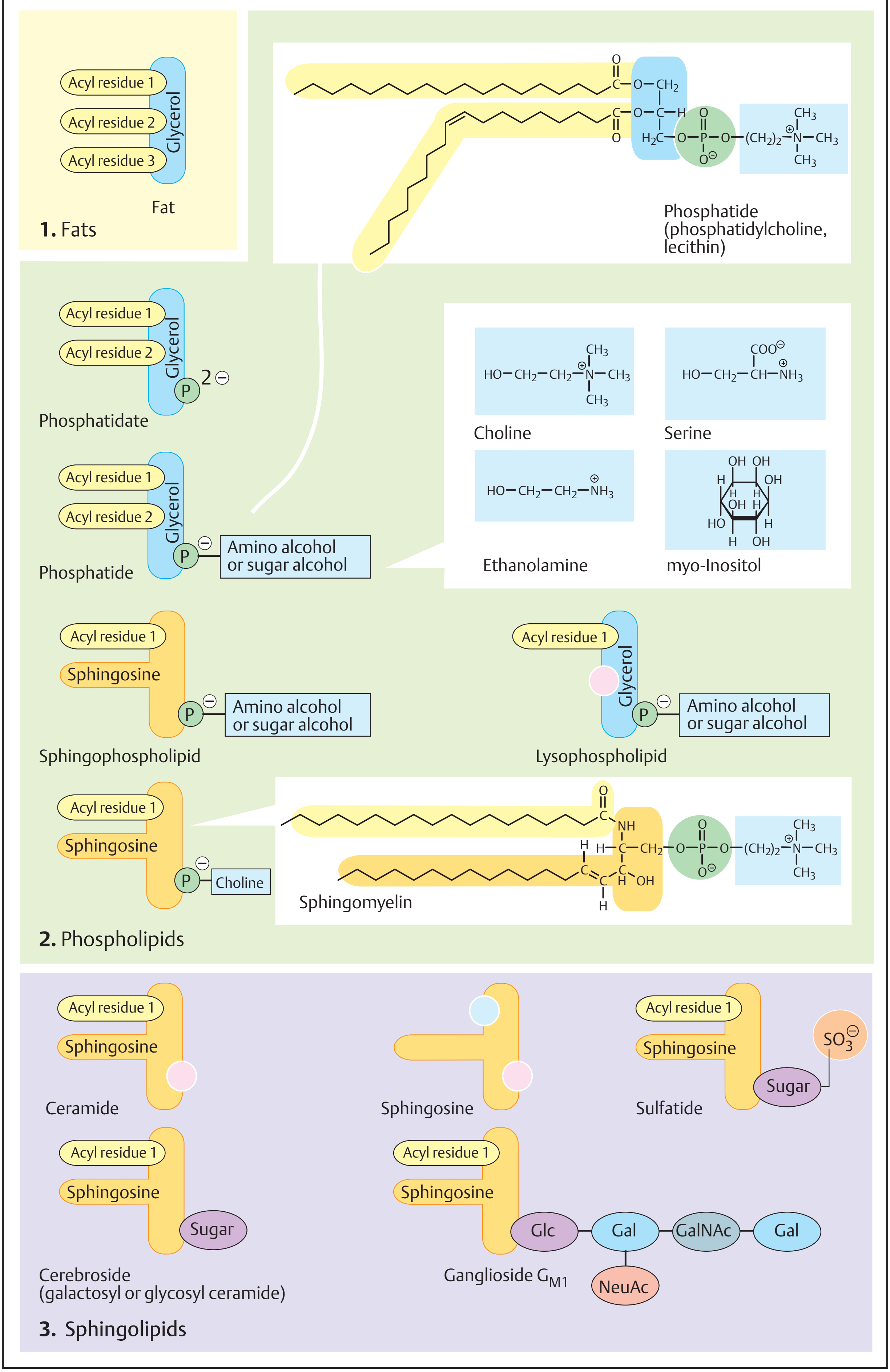
Acyl residue 1
Acyl residue 2
Acyl residue 3
Glycerol
Fat
1. Fats
Phosphatide (phosphatidylcholine, lecithin)
Acyl residue 1
Acyl residue 2
Glycerol
P
2 ⊖
Phosphatidate
Choline
Serine
Ethanolamine
myo-Inositol
Acyl residue 1
Acyl residue 2
Glycerol
P
Amino alcohol or sugar alcohol
Phosphatide
Acyl residue 1
Sphingosine
P
Amino alcohol or sugar alcohol
Sphingophospholipid
Acyl residue 1
Glycerol
P
Amino alcohol or sugar alcohol
Lysophospholipid
Acyl residue 1
Sphingosine
P
Choline
Sphingomyelin
2. Phospholipids
Acyl residue 1
Sphingosine
Ceramide
Sphingosine
Acyl residue 1
Sphingosine
Sugar
$SO_3^⊖$
Sulfatide
Acyl residue 1
Sphingosine
Sugar
Cerebroside (galactosyl or glycosyl ceramide)
Acyl residue 1
Sphingosine
Glc
Gal
GalNAc
Gal
NeuAc
Ganglioside G_{M1}
3. Sphingolipids

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Isoprenoids

A. Activated acetic acid as a component of lipids ◑

Although the lipids found in plant and animal organisms occur in many different forms, they are all closely related biogenetically; they are all derived from **acetyl-CoA**, the “activated acetic acid” (see pp. 12, 110).

1. One major pathway leads from acetyl-CoA to the activated fatty acids (**acyl-CoA**; for details, see p. 168). *Fats, phospholipids,* and *glycolipids* are synthesized from these, and fatty acid derivatives in particular are formed. Quantitatively, this is the most important pathway in animals and most plants.

2. The second pathway leads from acetyl-CoA to isopentenyl diphosphate (*“active isoprene”*), the basic component for the **isoprenoids**. Its biosynthesis is discussed in connection with biosynthesis of the isoprenoid, cholesterol (see p. 172).

B. Isoprenoids ◑

Formally, isoprenoids are derived from a single common building block, isoprene (2-methyl-1,3-butadiene), a methyl-branched compound with five C atoms. Activated isoprene, *isopentenyl diphosphate,* is used by plants and animals to biosynthesize linear and cyclic oligomers and polymers. For the isoprenoids listed here—which only represent a small selection—the number of isoprene units (I) is shown.

From activated isoprene, the metabolic pathway leads via dimerization to activated *geraniol* (I = 2) and then to activated *farnesol* (I = 3). At this point, the pathway divides into two. Further extension of farnesol leads to chains with increasing numbers of isoprene units—e. g., *phytol* (I = 4), *dolichol* (I = 14–24), and *rubber* (I = 700–5000). The other pathway involves a “head-to-head” linkage between two farnesol residues, giving rise to *squalene* (I = 6), which, in turn, is converted to *cholesterol* (I = 6) and the other *steroids.*

The ability to synthesize particular isoprenoids is limited to a few species of plants and animals. For example, rubber is only formed by a few plant species, including the rubber tree (*Hevea brasiliensis*). Several isoprenoids that are required by animals for metabolism, but cannot be produced by them independently, are vitamins; this group includes *vitamins A, D, E,* and *K.* Due to its structure and function, vitamin D is now usually classified as a steroid hormone (see pp. 56, 330).

Isoprene metabolism in plants is very complex. Plants can synthesize many types of aromatic substances and volatile oils from isoprenoids. Examples include *menthol* (I= 2), *camphor* (I = 2), and *citronellal* (I = 2). These C_{10} compounds are also called *monoterpenes.* Similarly, compounds consisting of three isoprene units (I = 3) are termed *sesquiterpenes*, and the steroids (I = 6) are called *triterpenes.*

Isoprenoids that have hormonal and signaling functions form an important group. These include *steroid hormones* (I = 6) and *retinoate* (the anion of retinoic acid; I = 3) in vertebrates, and *juvenile hormone* (I = 3) in arthropods. Some plant hormones also belong to the isoprenoids—e. g., the cytokinins, abscisic acid, and brassinosteroids.

Isoprene chains are sometimes used as lipid anchors to fix molecules to membranes (see p. 214). Chlorophyll has a *phytyl* residue (I = 4) as a lipid anchor. Coenzymes with isoprenoid anchors of various lengths include *ubiquinone* (coenzyme Q; I = 6–10), *plastoquinone* (I = 9), and *menaquinone* (vitamin K; I = 4–6). Proteins can also be anchored to membranes by *isoprenylation.*

In some cases, an isoprene residue is used as an element to modify molecules chemically. One example of this is *N*′-isopentenyl-AMP, which occurs as a modified component in tRNA.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Activated acetic acid as a component of lipids

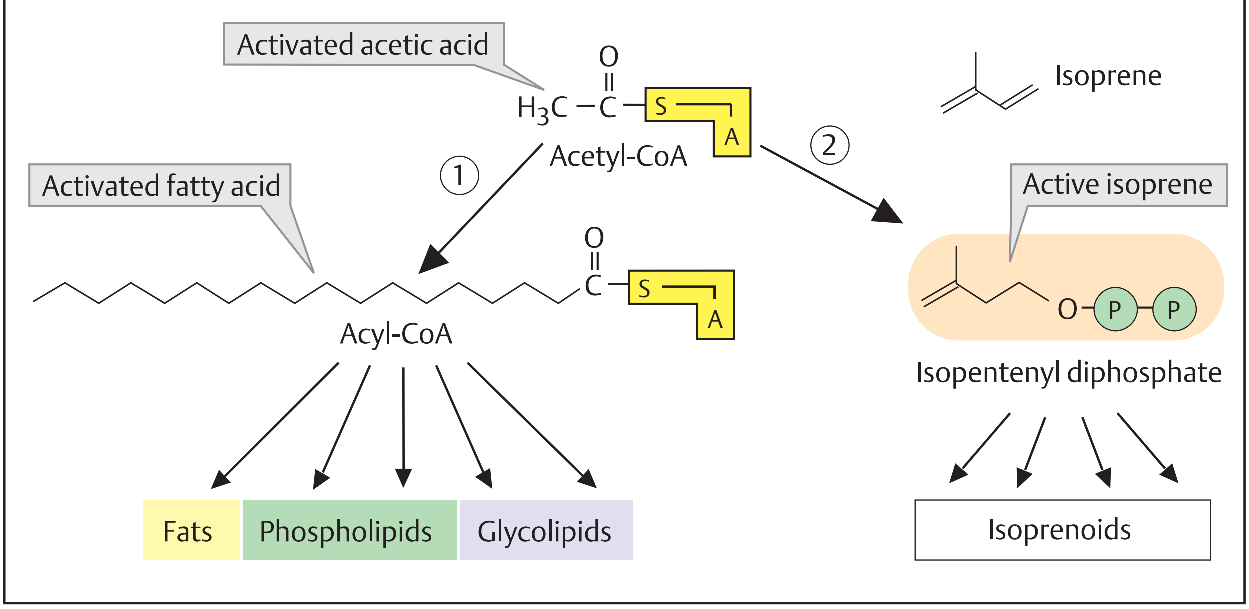
Activated acetic acid
O
H3C – C – S A
Acetyl-CoA
Isoprene
1
2
Activated fatty acid
Active isoprene
O
C – S A
O – P – P
Acyl-CoA
Isopentenyl diphosphate
Fats
Phospholipids
Glycolipids
Isoprenoids

B. Isoprenoids

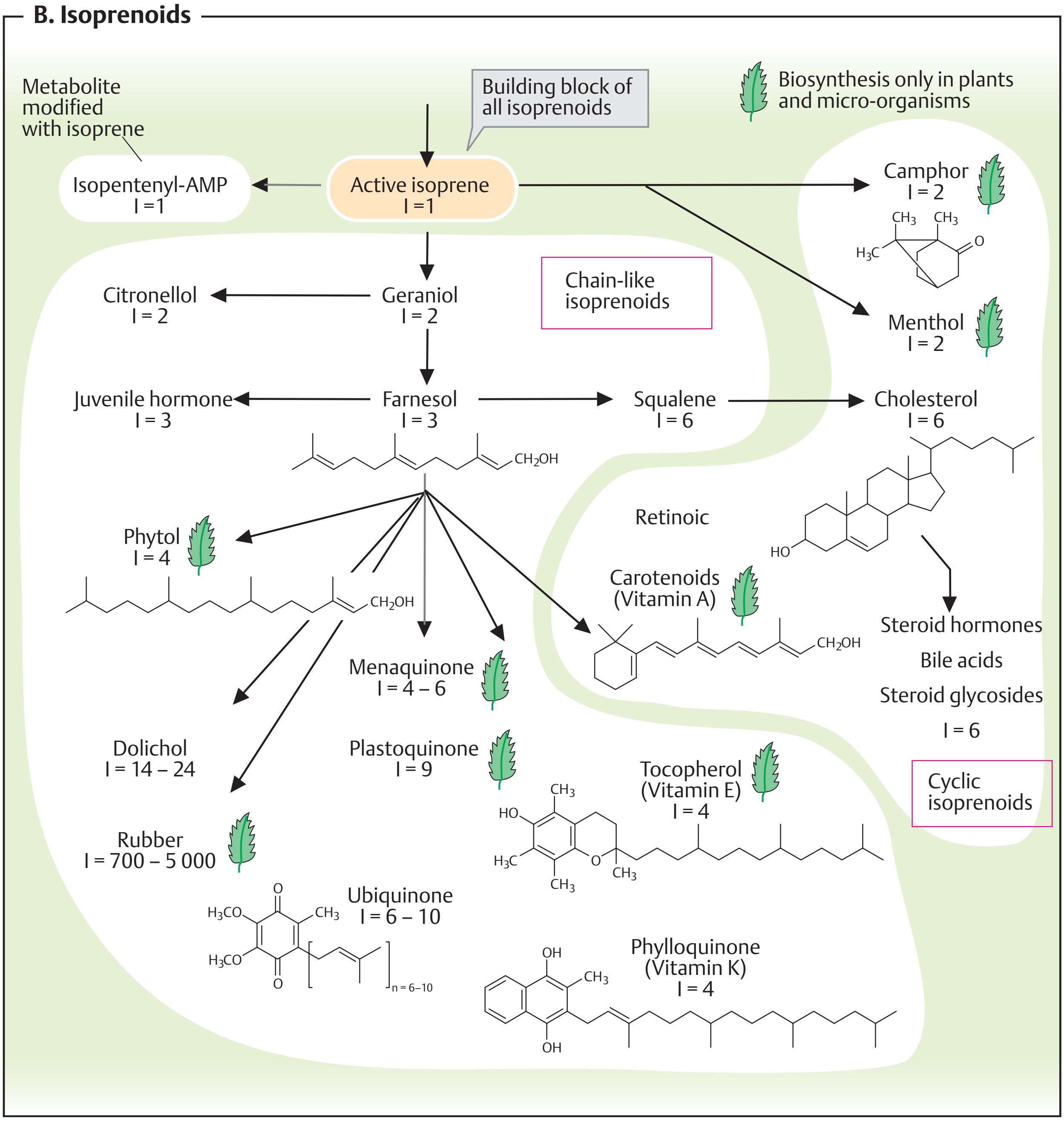
Metabolite modified with isoprene
Building block of all isoprenoids
Biosynthesis only in plants and micro-organisms
Isopentenyl-AMP
I = 1
Active isoprene
I = 1
Camphor
I = 2
CH3
CH3
H3C
O
Menthol
I = 2
Chain-like isoprenoids
Citronellol
I = 2
Geraniol
I = 2
Juvenile hormone
I = 3
Farnesol
I = 3
CH2OH
Squalene
I = 6
Cholesterol
I = 6
HO
Retinoic
Carotenoids (Vitamin A)
CH2OH
Phytol
I = 4
CH2OH
Steroid hormones
Bile acids
Steroid glycosides
I = 6
Menaquinone
I = 4 – 6
Dolichol
I = 14 – 24
Plastoquinone
I = 9
Cyclic isoprenoids
Tocopherol (Vitamin E)
I = 4
CH3
HO
H3C
O
CH3
CH3
Rubber
I = 700 – 5 000
Ubiquinone
I = 6 – 10
O
H3CO
CH3
H3CO
O
n = 6–10
Phylloquinone (Vitamin K)
I = 4
OH
CH3
OH

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Steroid structure

A. Steroid building blocks ◑

Common to all of the steroids is a molecular core structure consisting of four saturated rings, known as *gonane*. At the end of the steroid core, many steroids also carry a side chain, as seen in *cholestane*, the basic component of the *sterols* (steroid alcohols).

B. Spatial structure ○

The four rings of the steroids are distinguished using the letters A, B, C, and D. Due to the tetrahedral arrangement of the single carbon bonds, the rings are not flat, but puckered. Various *ring conformations* are known by the terms "chair," "boat," and "twisted" (not shown). The *chair* and *boat* conformations are common. Fivemembered rings frequently adopt a conformation referred to as an "envelope". Some rings can be converted from one conformation to another at room temperature, but with steroids this is difficult.

Substituents of the steroid core lie either approximately in the same plane as the ring (e = *equatorial*) or nearly perpendicular to it (a = *axial*). In threedimensional representations, substituents pointing toward the observer are indicated by an unbroken line (β position), while bonds pointing into the plane of the page are indicated by a dashed line (α position). The so-called *angular* methyl groups at C-10 and C-13 of the steroids always adopt the β position.

Neighboring rings can lie in the same plane (*trans*; **2**) or at an angle to one another (*cis*; **1**). This depends on the positions of the substituents of the shared ring carbons, which can be arranged either *cis* or *trans* to the angular methyl group at C-10. The substituents of steroid that lie at the points of intersection of the individual rings are usually in *trans* position. As a whole, the core of most steroids is more or less planar, and looks like a flat disk. The only exceptions to this are the ecdysteroids, bile acids (in which A:B is *cis*), cardiac glycosides, and toad toxins.

A more realistic impression of the three-dimensional structure of steroids is provided by the space-filling model of *cholesterol* (**3**). The four rings form a fairly rigid scaffolding, onto which the much more mobile side chain is attached.

Steroids are relatively apolar (hydrophobic). Some polar groups—e. g., hydroxyl and oxo groups—give them amphipathic properties. This characteristic is especially pronounced with the bile acids (see p. 314).

C. Thin-layer chromatography ○

Thin-layer chromatography (TLC) is a powerful, mainly analytic, technique for rapidly separating lipids and other small molecules such as amino acids, nucleotides, vitamins, and drugs. The *sample* being analyzed is applied to a *plate* made of glass, aluminum, or plastic, which is covered with a thin layer of silica gel or other material (**1**). The plate is then placed in a chromatography chamber that contains some *solvent*. Drawn by capillary forces, the solvent moves up the plate (**2**). The substances in the sample move with the solvent. The speed at which they move is determined by their distribution between the *stationary phase* (the hydrophilic silica), and the *mobile phase* (the hydrophobic solvent). When the solvent reaches the top edge of the plate, the chromatography is stopped. After evaporation of the solvent, the separated substances can be made visible using appropriate staining methods or with physical processes (e. g., ultraviolet light) (**3**). The movement of a substance in a given TLC system is expressed as its R_f value. In this way, compounds that are not known can be identified by comparison with reference substances.

A process in which the polarity of the stationary and mobile phases is reversed—i. e., the stationary phase is apolar and the solvent is polar—is known as "reversed-phase thin-layer chromatography" (RP-TLC).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Steroid building blocks

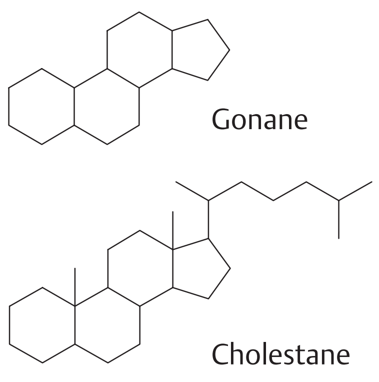
Gonane
Cholestane

B. 3D structure

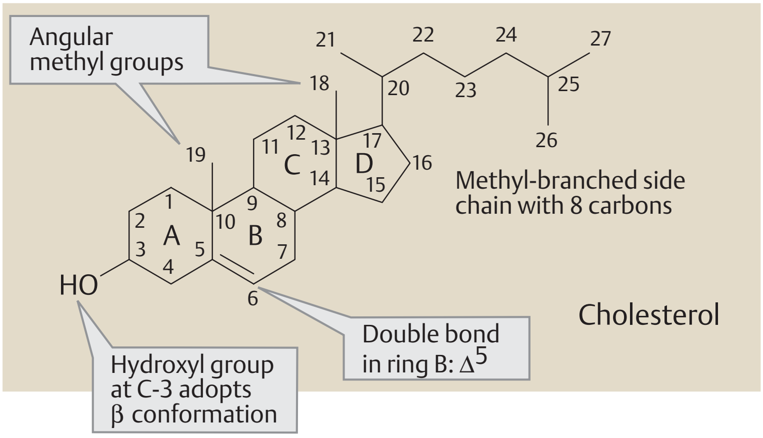
Angular methyl groups
21
22
24
27
18
20
23
25
26
19
11
12
13
17
C
D
16
14
15
1
9
2
A
10
B
8
3
5
7
4
HO
6
Methyl-branched side chain with 8 carbons
Cholesterol
Double bond in ring B: Δ5
Hydroxyl group at C-3 adopts β conformation

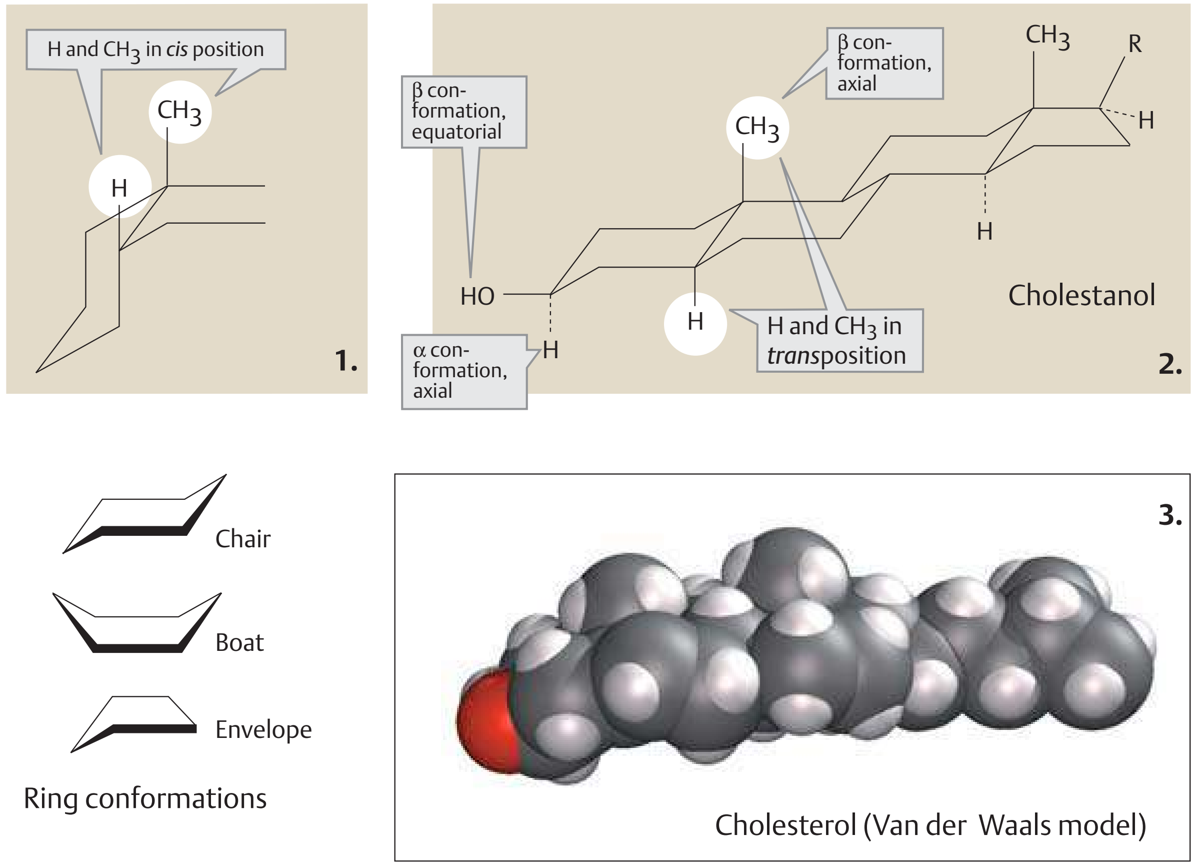
H and CH3 in cis position
CH3
H
1.
β con-formation, equatorial
β con-formation, axial
CH3
CH3
R
H
H
HO
H
H
α con-formation, axial
H and CH3 in transposition
Cholestanol
2.
Chair
Boat
Envelope
Ring conformations
3.
Cholesterol (Van der Waals model)

C. Thin-layer chromatography

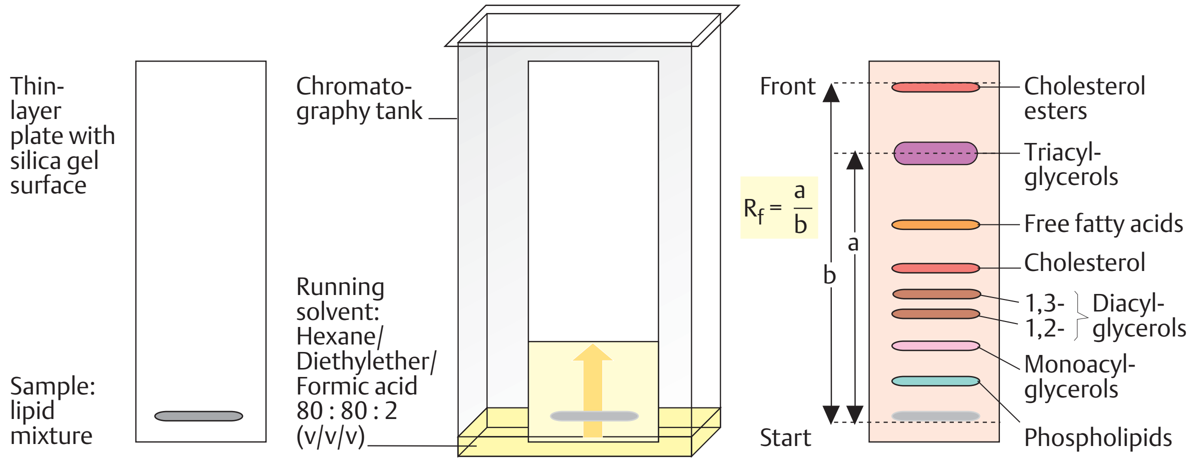
Thin-layer plate with silica gel surface
Sample: lipid mixture
Chromato-graphy tank
Running solvent: Hexane/ Diethylether/ Formic acid 80 : 80 : 2 (v/v/v)
Front
Start
$R_f = \frac{a}{b}$
a
b
Cholesterol esters
Triacyl-glycerols
Free fatty acids
Cholesterol
1,3- 1,2- Diacyl-glycerols
Monoacyl-glycerols
Phospholipids

1. Load **2.** Develop **3.** Make visible

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Steroids: overview

The three most important groups of steroids are the *sterols*, *bile acids*, and *steroid hormones*. Particularly in plants, compounds with steroid structures are also found that are notable for their pharmacological effects—steroid alkaloids, digitalis glycosides, and saponins.

A. Sterols ◑

Sterols are *steroid alcohols*. They have a β-positioned hydroxyl group at C-3 and one or more double bonds in ring B and in the side chain. There are no further oxygen functions, as in the carbonyl and carboxyl groups.

The most important sterol in animals is **cholesterol**. Plants and microorganisms have a wide variety of closely related sterols instead of cholesterol—e. g., **ergosterol**, **β-sitosterol**, and **stigmasterol**.

Cholesterol is present in all animal tissues, and particularly in neural tissue. It is a major constituent of cellular membranes, in which it regulates fluidity (see p. 216). The storage and transport forms of cholesterol are its esters with fatty acids. In lipoproteins, cholesterol and its fatty acid esters are associated with other lipids (see p. 278). Cholesterol is a constituent of the bile and is therefore found in many gallstones. Its biosynthesis, metabolism, and transport are discussed elsewhere (see pp. 172, 312).

Cholesterol-rich lipoproteins of the LDL type are particularly important in the development of arteriosclerosis, in which the arterial walls are altered in connection with an excess plasma cholesterol level. In terms of dietary physiology, it is important that plant foodstuffs are low in cholesterol. By contrast, animal foods can contain large amounts of cholesterol—particularly butter, egg yolk, meat, liver, and brain.

B. Bile acids ◑

Bile acids are synthesized from cholesterol in the liver (see p. 314). Their structures can therefore be derived from that of cholesterol. Characteristic for the bile acids is a side chain shortened by three C atoms in which the last carbon atom is oxidized to a carboxyl group. The double bond in ring B is reduced and rings A and B are in *cis* position relative to each other (see p. 54). One to three hydroxyl groups (in α position) are found in the steroid core at positions 3, 7, and 12. Bile acids keep bile cholesterol in a soluble state as micelles and promote the digestion of lipids in the intestine (see p. 270). **Cholic acid** and **chenodeoxycholic acid** are *primary bile acids* that are formed by the liver. Their dehydroxylation at C-7 by microorganisms from the intestinal flora gives rise to the *secondary bile acids* **lithocholic acid** and **deoxycholic acid**.

C. Steroid hormones ◑

The conversion of cholesterol to *steroid hormones* (see p. 376) is of minor importance quantitatively, but of major importance in terms of physiology. The steroid hormones are a group of lipophilic signal substances that regulate metabolism, growth, and reproduction (see p. 374).

Humans have six steroid hormones: **progesterone**, **cortisol**, **aldosterone**, **testosterone**, **estradiol**, and **calcitriol**. With the exception of calcitriol, these steroids have either no side chain or only a short side one consisting of two carbons. Characteristic for most of them is an oxo group at C-3, conjugated with a double bond between C-4 and C-5 of ring A. Differences occur in rings C and D. Estradiol is aromatic in ring A, and its hydroxyl group at C-3 is therefore phenolic. Calcitriol differs from other vertebrate steroid hormones; it still contains the complete carbon framework of cholesterol, but lightdependent opening of ring B turns it into what is termed a "secosteroid" (a steroid with an open ring).

Ecdysone is the steroid hormone of the arthropods. It can be regarded as an early form of the steroid hormones. Steroid hormones with signaling functions also occur in plants.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Sterols

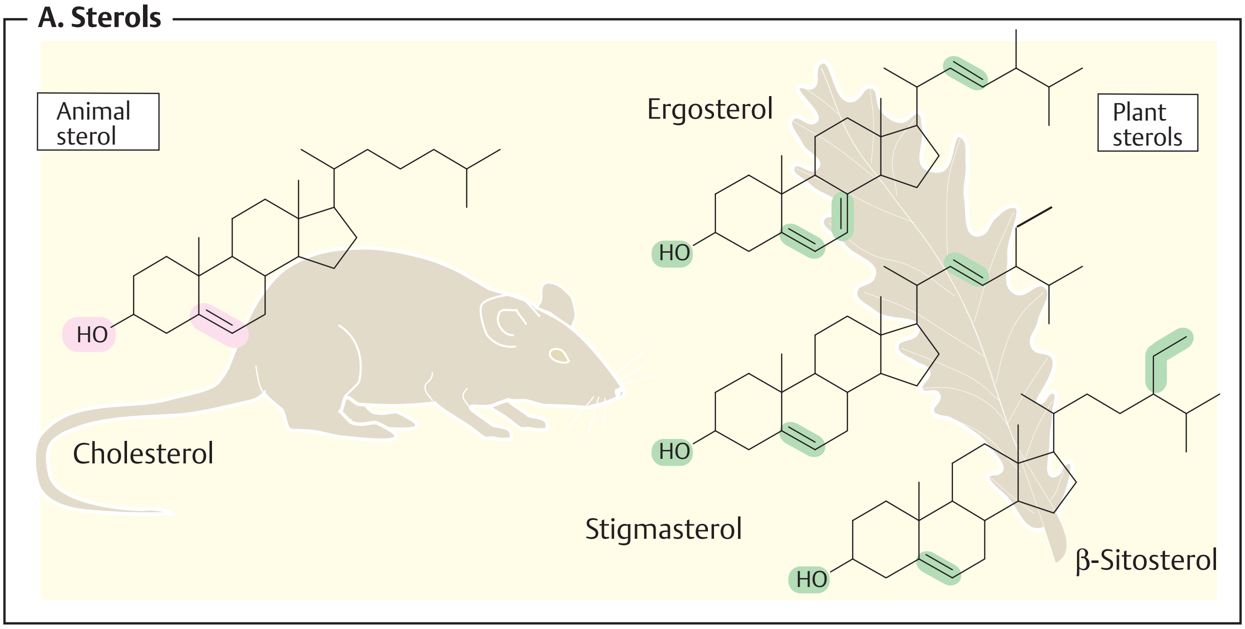
Animal sterol
HO
Cholesterol
Ergosterol
Plant sterols
HO
HO
Stigmasterol
HO
β-Sitosterol

B. Bile acids

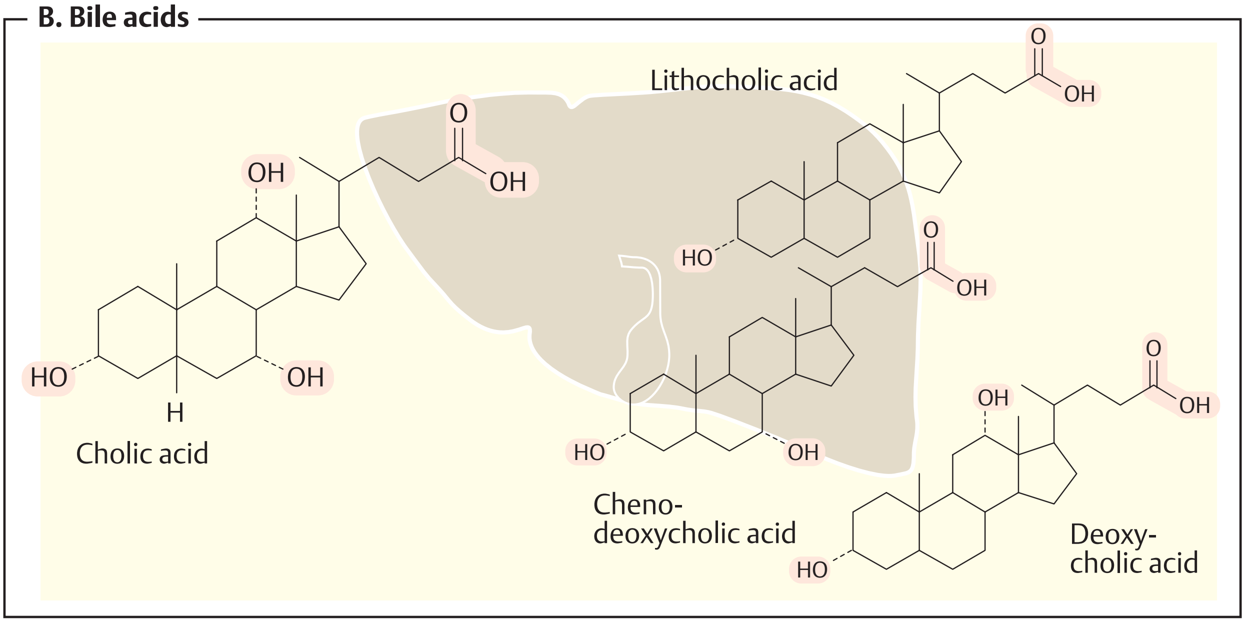
OH
O
OH
HO
OH
H
Cholic acid
Lithocholic acid
O
OH
HO
O
OH
HO
OH
Cheno-deoxycholic acid
OH
O
OH
HO
Deoxy-cholic acid

C. Steroid hormones

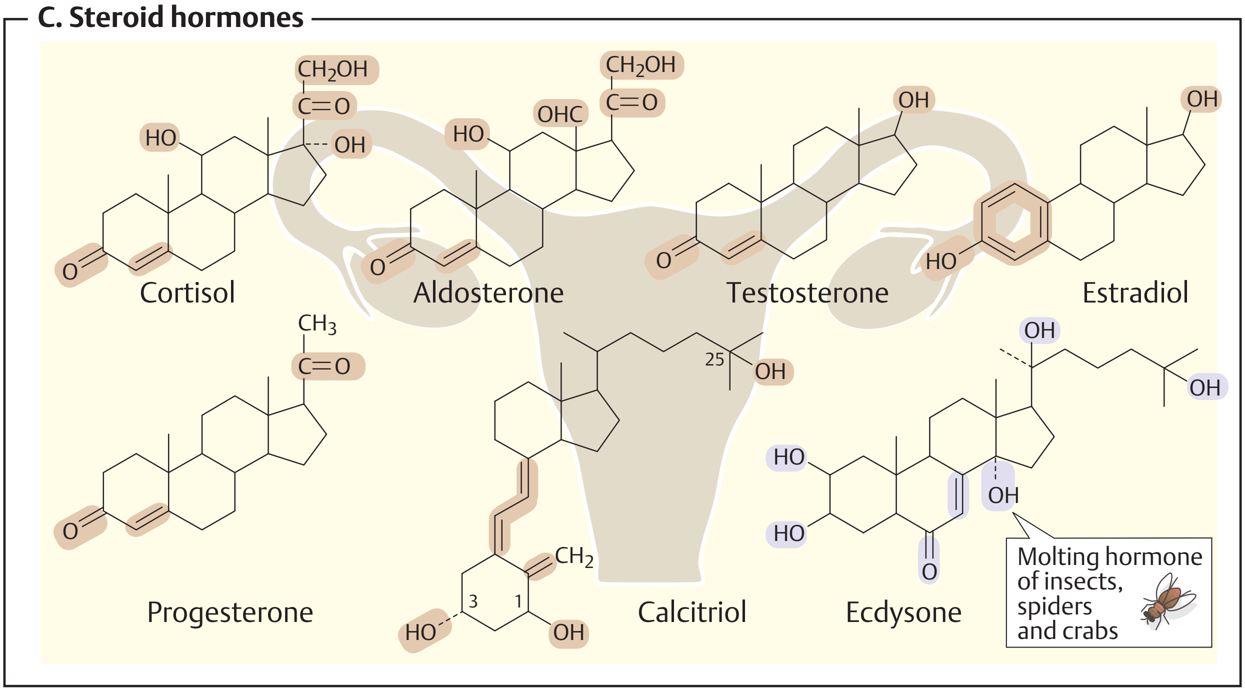
CH₂OH
C=O
HO
OH
O
Cortisol
CH₂OH
C=O
OHC
HO
O
Aldosterone
OH
O
Testosterone
OH
HO
Estradiol
CH₃
C=O
O
Progesterone
25
OH
CH₂
3
1
HO
OH
Calcitriol
OH
OH
HO
HO
OH
O
Ecdysone
Molting hormone of insects, spiders and crabs

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Amino acids: chemistry and properties

A. Amino acids: functions ●

The amino acids (2-aminocarboxylic acids) fulfill various functions in the organism. Above all, they serve as the **components of peptides and proteins**. Only the 20 *proteinogenic amino acids* (see p. 60) are included in the genetic code and therefore regularly found in proteins. Some of these amino acids undergo further (post-translational) change following their incorporation into proteins (see p. 62). Amino acids or their derivatives are also form components of **lipids**—e. g., serine in phospholipids and glycine in bile salts. Several amino acids function as **neurotransmitters** themselves (see p. 352), while others are precursors of neurotransmitters, mediators, or hormones (see p. 380). Amino acids are important (and sometimes essential) components of food (see p. 360). Specific amino acids form **precursors** for other metabolites—e. g., for glucose in gluconeogenesis, for purine and pyrimidine bases, for heme, and for other molecules. Several non-proteinogenic amino acids function as intermediates in the synthesis and breakdown of proteinogenic amino acids (see p. 412) and in the urea cycle (see p. 182).

B. Optical activity ◑

The natural amino acids are mainly α-amino acids, in contrast to β-amino acids such as β-alanine and taurine. Most α-amino acids have four different substituents at C-2 (Cα). The α atom therefore represents a *chiral center*—i. e., there are two different **enantiomers** (L- and D-amino acids; see p. 8). Among the proteinogenic amino acids, only glycine is *not* chiral (R = H). In nature, it is almost exclusively **L-amino acids** that are found. **D**-Amino acids occur in bacteria—e. g., in murein (see p. 40)—and in peptide antibiotics. In animal metabolism, D-Amino acids would disturb the enzymatic reactions of L-amino acids and they are therefore broken down in the liver by the enzyme *D-amino acid oxidase.*

The **Fischer projection** (center) is used to present the formulas for chiral centers in biomolecules. It is derived from their three-dimensional structure as follows: firstly, the tetrahedron is rotated in such a way that the most oxidized group (the carboxylate group) is at the top. Rotation is then continued until the line connecting line COO^- and R (red) is level with the page. In L-amino acids, the NH_3^+ group is then on the left, while in **D**-amino acids it is on the right.

C. Dissociation curve of histidine ◑

All amino acids have at least two ionizable groups, and their net charge therefore depends on the pH value. The COOH groups at the α-C atom have pK_a values of between 1.8 and 2.8 and are therefore more acidic than simple monocarboxylic acids. The basicity of the α-amino function also varies, with pK_a values of between 8.8 and 10.6, depending on the amino acid. Acidic and basic amino acids have additional ionizable groups in their side chain. The pK_a values of these side chains are listed on p. 60. The electrical charges of peptides and proteins are mainly determined by groups in the side chains, as most α-carboxyl and α-amino functions are linked to peptide bonds (see p. 66).

Histidine can be used here as an example of the pH-dependence of the net charge of an amino acid. In addition to the carboxyl group and the amino group at the α-C atom with pK_a values of 1.8 and 9.2, respectively, histidine also has an imidazole residue in its side chain with a pK_a value of 6.0. As the pH increases, the net charge (the sum of the positive and negative charges) therefore changes from +2 to –1. At pH 7.6, the net charge is zero, even though the molecule contains two almost completely ionized groups in these conditions. This pH value is called the **isoelectric point**.

At its isoelectric point, histidine is said to be **zwitterionic**, as it has both anionic *and* cationic properties. Most other amino acids are also zwitterionic at neutral pH. Peptides and proteins also have isoelectric points, which can vary widely depending on the composition of the amino acids.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Amino acids: functions

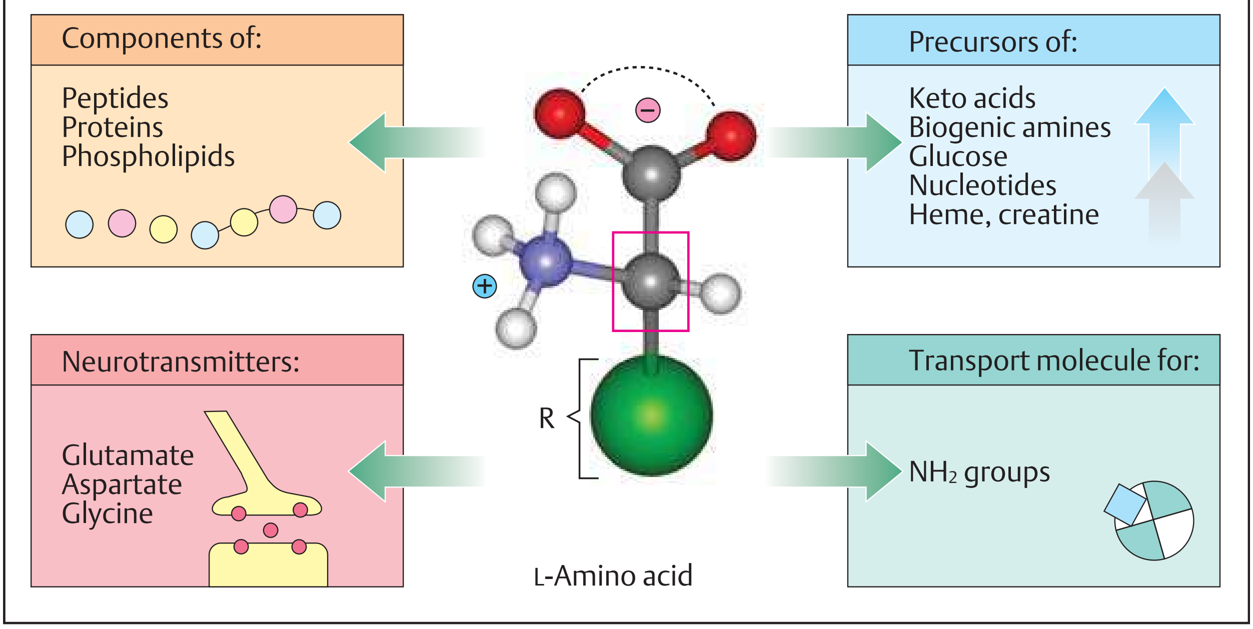
Components of:
Peptides
Proteins
Phospholipids
Precursors of:
Keto acids
Biogenic amines
Glucose
Nucleotides
Heme, creatine
Neurotransmitters:
Glutamate
Aspartate
Glycine
Transport molecule for:
NH₂ groups
R
⊖
⊕
L-Amino acid

B. Optical activity

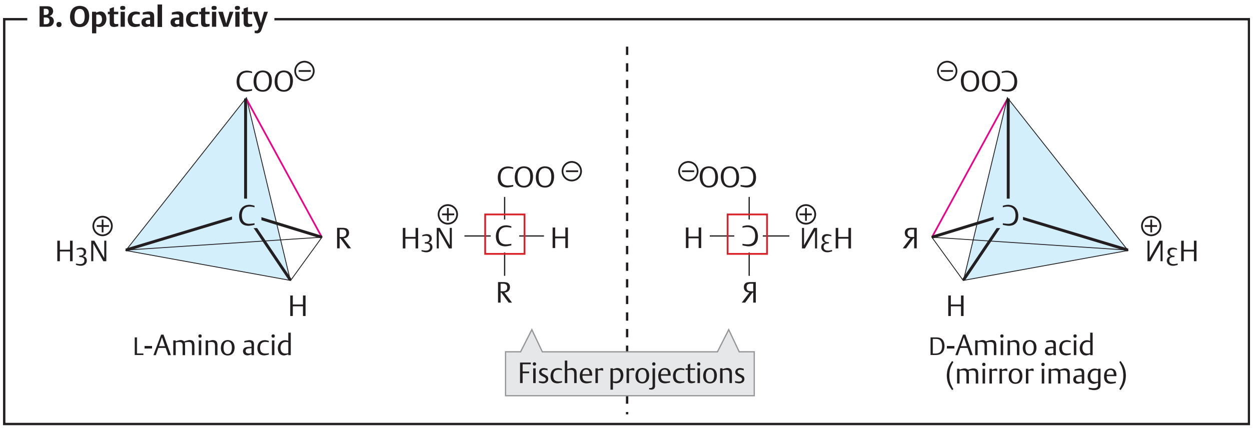
COO⊖
H₃N⊕
C
R
H
L-Amino acid
COO⊖
H₃N⊕
C
H
R
Fischer projections
COO⊖
H
C
H₃N⊕
R
COO⊖
C
R
H₃N⊕
H
D-Amino acid
(mirror image)

C. Dissociation curve of histidine

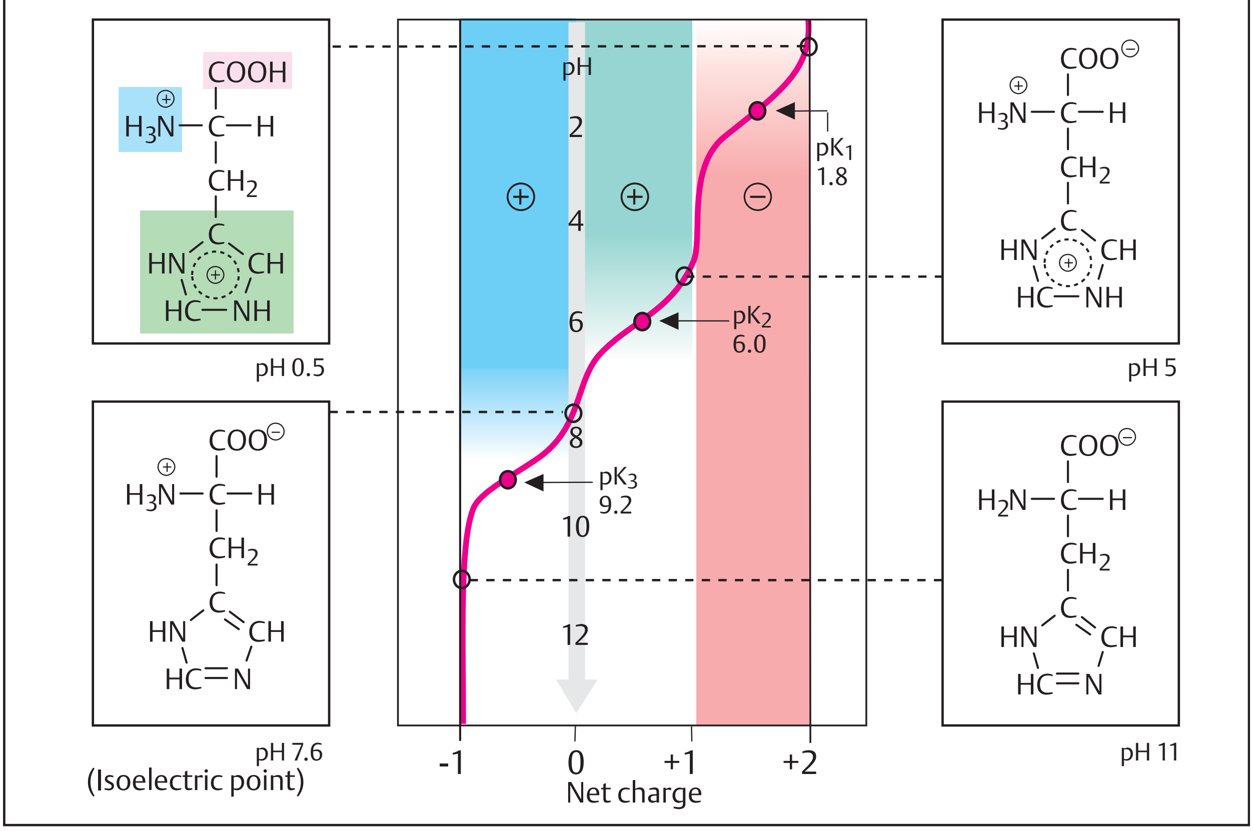

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Proteinogenic amino acids

A. The proteinogenic amino acids ◑

The amino acids that are included in the genetic code (see p. 248) are described as "proteinogenic." With a few exceptions (see p. 58), only these amino acids can be incorporated into proteins through *translation*. Only the side chains of the 20 proteinogenic amino acids are shown here. Their classification is based on the chemical structure of the side chains, on the one hand, and on their polarity on the other (see p. 6). The literature includes several slightly different systems for classifying amino acids, and details may differ from those in the system used here.

For each amino acid, the illustration names:

- *Membership of structural classes* I–VII (see below; e. g., III and VI for histidine)
- Name and abbreviation, formed from the first three letters of the name (e. g., histidine, His)
- The *one-letter symbol* introduced to save space in the electronic processing of sequence data (H for histidine)
- A quantitative *value for the polarity* of the side chain (bottom left; 10.3 for histidine). The more positive this value is, the *more polar* the amino acid is.

In addition, the polarity of the side chains is indicated by color. It increases from yellow, through light and dark green, to bluish green. For ionizing side chains, the corresponding pK_a values are also given (red numbers).

The **aliphatic** amino acids (class I) include *glycine, alanine, valine, leucine,* and *isoleucine*. These amino acids do not contain heteroatoms (N, O, or S) in their side chains and do not contain a ring system. Their side chains are markedly apolar. Together with threonine (see below), valine, leucine, and isoleucine form the group of *branched-chain amino acids*. The **sulfurcontaining amino acids** *cysteine* and *methionine* (class II), are also apolar. However, in the case of cysteine, this only applies to the undissociated state. Due to its ability to form disulfide bonds, cysteine plays an important role in the stabilization of proteins (see p. 72). Two cysteine residues linked by a disulfide bridge are referred to as *cystine* (not shown).

The **aromatic amino acids** (class III) contain resonancestabilized rings. In this group, only *phenylalanine* has strongly apolar properties. *Tyrosine* and *tryptophan* are moderately polar, and *histidine* is even strongly polar. The imidazole ring of histidine is already protonated at weakly acidic pH values. Histidine, which is only aromatic in protonated form (see p. 58), can therefore also be classified as a basic amino acid. Tyrosine and tryptophan show strong light absorption at wavelengths of 250–300 nm.

The **neutral** amino acids (class IV) have hydroxyl groups (*serine, threonine*) or amide groups (*asparagine, glutamine*). Despite their nonionic nature, the amide groups of asparagine and glutamine are markedly polar.

The carboxyl groups in the side chains of the **acidic** amino acids *aspartic acid* and *glutamic acid* (class V) are almost completely ionized at physiological pH values. The side chains of the **basic** amino acids *lysine* and *arginine* are also fully ionized—i. e., positively charged—at neutral pH. Arginine, with its positively charge guanidinium group, is particularly strongly basic, and therefore extremely polar.

Proline (VII) is a special case. Together with the α-C atom and the α-NH_2 group, its side chain forms a fivemembered ring. Its nitrogen atom is only weakly basic and is not protonated at physiological pH. Due to its ring structure, proline causes *bending of the peptide chain* in proteins (this is important in collagen, for example; see p. 70).

Several proteinogenic amino acids cannot be synthesized by the human organism, and therefore have to be supplied from the diet. These **essential amino acids** (see p. 360) are marked with a star in the illustration. Histidine and possibly also arginine are essential for infants and small children.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. The proteinogenic amino acids

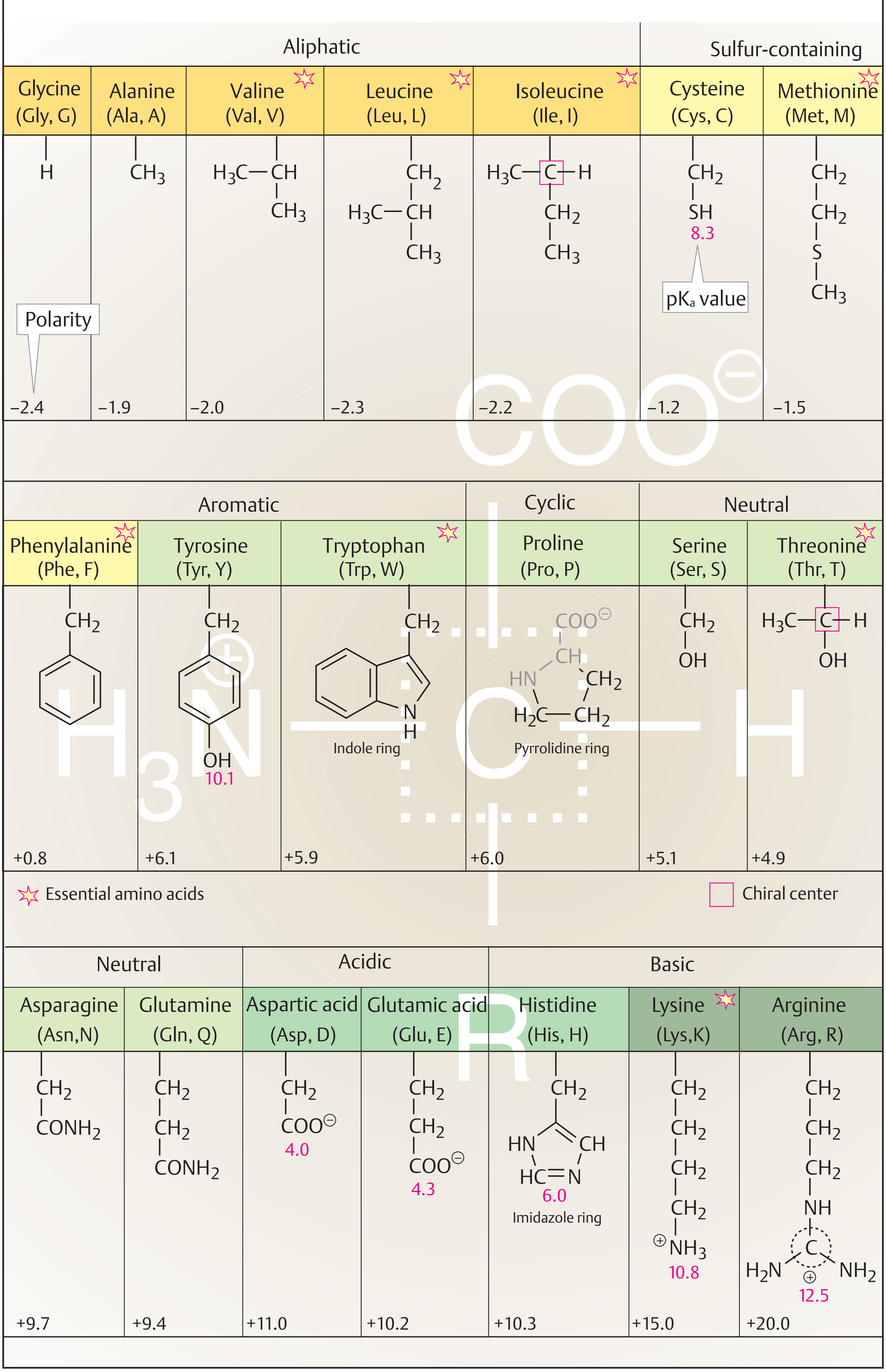
Aliphatic
Sulfur-containing
Glycine (Gly, G)
H
–2.4
Polarity
Alanine (Ala, A)
CH3
–1.9
Valine (Val, V)
H3C—CH
CH3
–2.0
Leucine (Leu, L)
CH2
H3C—CH
CH3
–2.3
Isoleucine (Ile, I)
H3C—C—H
CH2
CH3
–2.2
Cysteine (Cys, C)
CH2
SH
8.3
pKa value
–1.2
Methionine (Met, M)
CH2
CH2
S
CH3
–1.5
COO⊖
Aromatic
Cyclic
Neutral
Phenylalanine (Phe, F)
CH2
+0.8
Tyrosine (Tyr, Y)
CH2
OH
10.1
+6.1
Tryptophan (Trp, W)
CH2
N
H
Indole ring
+5.9
Proline (Pro, P)
COO⊖
CH
HN
CH2
H2C—CH2
Pyrrolidine ring
+6.0
Serine (Ser, S)
CH2
OH
+5.1
Threonine (Thr, T)
H3C—C—H
OH
+4.9
H3N⊕
C
H
R
Essential amino acids
Chiral center
Neutral
Acidic
Basic
Asparagine (Asn,N)
CH2
CONH2
+9.7
Glutamine (Gln, Q)
CH2
CH2
CONH2
+9.4
Aspartic acid (Asp, D)
CH2
COO⊖
4.0
+11.0
Glutamic acid (Glu, E)
CH2
CH2
COO⊖
4.3
+10.2
Histidine (His, H)
CH2
HN
CH
HC=N
6.0
Imidazole ring
+10.3
Lysine (Lys,K)
CH2
CH2
CH2
CH2
⊕NH3
10.8
+15.0
Arginine (Arg, R)
CH2
CH2
CH2
NH
C
H2N
NH2
⊕
12.5
+20.0

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Non-proteinogenic amino acids

In addition to the 20 proteinogenic amino acids (see p. 60), there are also many more compounds of the same type in nature. These arise during metabolic reactions (**A**) or as a result of enzymatic modifications of amino acid residues in peptides or proteins (**B**). The "biogenic amines" (**C**) are synthesized from α-amino acids by decarboxylation.

A. Rare amino acids ○

Only a few important representatives of the non-proteinogenic amino acids are mentioned here. The basic amino acid **ornithine** is an analogue of lysine with a shortened side chain. Transfer of a carbamoyl residue to ornithine yields **citrulline**. Both of these amino acids are intermediates in the urea cycle (see p. 182). **Dopa** (an acronym of 3,4-**d**ihydr**o**xy**p**henyl**a**lanine) is synthesized by hydroxylation of tyrosine. It is an intermediate in the biosynthesis of catecholamines (see p. 352) and of melanin. It is in clinical use in the treatment of *Parkinson's disease*. **Selenocysteine**, a cysteine analogue, occurs as a component of a few proteins—e. g., in the enzyme glutathione peroxidase (see p. 284).

B. Post-translational protein modification ◑

Subsequent alteration of amino acid residues in finished peptides and proteins is referred to as *post-translational modification*. These reactions usually only involve polar amino acid residues, and they serve various purposes.

The free α-amino group at the *N*-terminus is blocked in many proteins by an acetyl residue or a longer acyl residue (**acylation**). *N*-terminal glutamate can cyclize into a pyroglutamate residue, while the *C*-terminal carboxylate group can be present in an amidated form (see TSH, p. 380). The side chains of serine and asparagine residues are often linked to oligosaccharides (**glycosylation**, see p. 230). **Phosphorylation** of proteins mainly affects serine and tyrosine residues. These reactions have mainly regulatory functions (see p. 114). Aspartate and histidine residues of enzymes are sometimes phosphorylated, too. A special modification of glutamate residues, **γ-carboxylation**, is found in coagulation factors. It is essential for blood coagulation (see p. 290). The ε-amino group of lysine residues is subject to a particularly large number of modifications. Its **acetylation** (or deacetylation) is an important mechanism for controlling genetic activity (see p. 244). Many coenzymes and cofactors are covalently linked to lysine residues. These include biotin (see p. 108), lipoic acid (see p. 106), and pyridoxal phosphate (see p. 108), as well as retinal (see p. 358). Covalent modification with **ubiquitin** marks proteins for breakdown (see p. 176). In collagen, lysine and proline residues are modified by **hydroxylation** to prepare for the formation of stable fibrils (see p. 70). Cysteine residues form **disulfide bonds** with one another (see p. 72). Cysteine **prenylation** serves to anchor proteins in membranes (see p. 214). Covalent bonding of a cysteine residue with heme occurs in cytochrome *c*. Flavins are sometimes covalently bound to cysteine or histidine residues of enzymes. Among the modifications of tyrosine residues, conversion into iodinated **thyroxine** (see p. 374) is particularly interesting.

C. Biogenic amines ◑

Several amino acids are broken down by *decarboxylation*. This reaction gives rise to what are known as biogenic amines, which have various functions. Some of them are **components of biomolecules**, such as *ethanolamine* in phospholipids (see p. 50). *Cysteamine* and *γ-alanine* are components of coenzyme A (see p. 12) and of pantetheine (see pp. 108, 168). Other amines function as signaling substances. An important **neurotransmitter** derived from glutamate is γ-aminobutyrate (GABA, see p. 356). The transmitter *dopamine* is also a precursor for the catecholamines epinephrine and norepinephrine (see p. 352). The biogenic amine *serotonin*, a substance that has many effects, is synthesized from tryptophan via the intermediate 5-hydroxytryptophan.

Monamines are inactivated into aldehydes by *amine oxidase* (monoamine oxidase, "MAO") with deamination and simultaneous oxidation. MAO inhibitors therefore play an important role in pharmacological interventions in neurotransmitter metabolism.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Rare amino acids

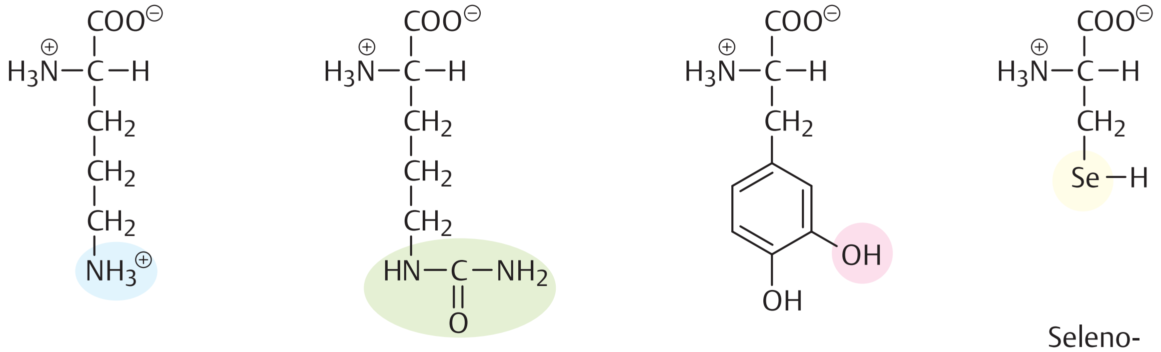
Ornithine

Citrulline

L-Dopa

Seleno-cysteine

B. Post-translational protein modification

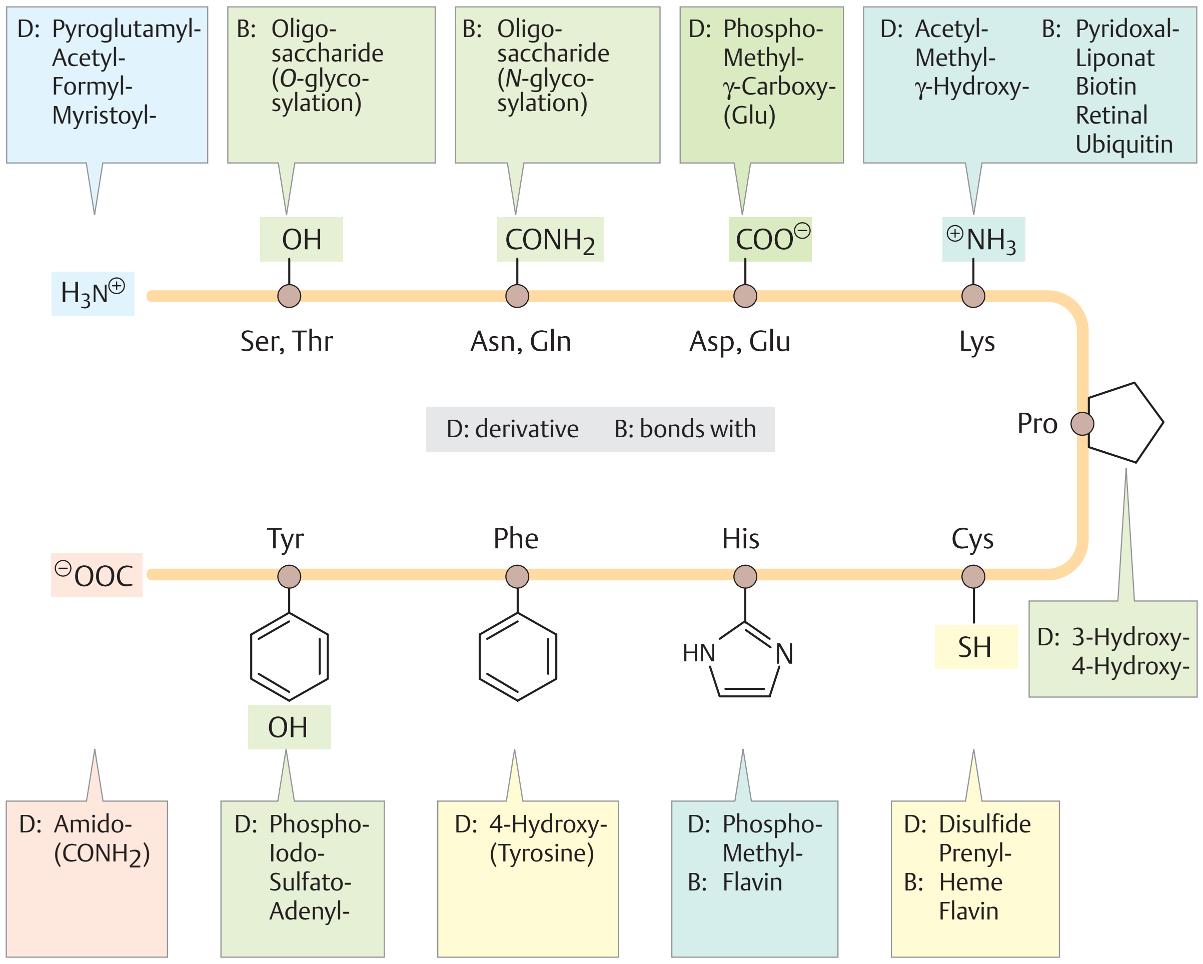
D: Pyroglutamyl- Acetyl- Formyl- Myristoyl-
B: Oligo-saccharide (O-glycosylation)
B: Oligo-saccharide (N-glycosylation)
D: Phospho- Methyl- γ-Carboxy- (Glu)
D: Acetyl- Methyl- γ-Hydroxy-
B: Pyridoxal- Liponat Biotin Retinal Ubiquitin
OH
CONH2
COO⊖
⊕NH3
H3N⊕
Ser, Thr
Asn, Gln
Asp, Glu
Lys
D: derivative B: bonds with
Pro
Tyr
Phe
His
Cys
⊖OOC
HN
N
SH
OH
D: 3-Hydroxy- 4-Hydroxy-
D: Amido- (CONH2)
D: Phospho- Iodo- Sulfato- Adenyl-
D: 4-Hydroxy- (Tyrosine)
D: Phospho- Methyl- B: Flavin
D: Disulfide Prenyl- B: Heme Flavin

C. Biogenic amines

Amino acid	Amine	Function
Serine	Ethanol-amine	Glutamate
Cysteine	Cysteamine	Component of coenzyme A
Threonine	Amino-propanol	Component of vitamin B_{12}
Aspartate	β-Alanine	Component of coenzyme A

Amino acid	Amine	Function
Glutamate	γ-Amino-butyrate	Neurotrans-mitter (GABA)
Histidine	Histamine	Mediator, neuro-transmitter
Dopa	Dopamine	Neurotransmitter
5-Hydroxy-tryptophan	Serotonin	Mediator, neuro-transmitter

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Peptides and proteins: overview

A. Proteins ◑

When amino acids are linked together by acid–amide bonds, linear macromolecules (peptides) are produced. Those containing more than ca. 100 amino acid residues are described as **proteins** (polypeptides). Every organism contains thousands of different proteins, which have a variety of functions. At a magnification of ca. 1.5 million, the semischematic illustration shows the structures of a few intra and extracellular proteins, giving an impression of their variety. The functions of proteins can be classified as follows.

Establishment and maintenance of structure. Structural proteins are responsible for the *shape and stability* of cells and tissues. A small part of a **collagen** molecule is shown as an example (right; see p. 70). The complete molecule is 1.5 300 nm in size, and at the magnification used here it would be as long as three pages of the book. **Histones** are also structural proteins. They organize the arrangement of DNA in chromatin. The basic components of chromatin, the *nucleosomes* (top right; see p. 218) consist of an octameric complex of histones, around which the DNA is coiled.

Transport. A wellknown transport protein is **hemoglobin** in the erythrocytes (bottom left). It is responsible for the transport of oxygen and carbon dioxide between the lungs and tissues (see p. 282). The blood plasma also contains many other proteins with transport functions. **Prealbumin** (transthyretin; middle), for example, transports the thyroid hormones thyroxin and triiodothyronine. **Ion channels** and other integral membrane proteins (see p. 220) facilitate the transport of ions and metabolites across biological membranes.

Protection and defense. The immune system protects the body from pathogens and foreign substances. An important component of this system is **immunoglobulin G** (bottom left; see p. 300). The molecule shown here is bound to an erythrocyte by complex formation with surface glycolipids (see p. 292).

Control and regulation. In biochemical signal chains, proteins function as signaling substances (hormones) and as hormone receptors. The complex between the growth hormone **somatotropin** and its **receptor** is shown here as an example (middle). Here, the extracellular domains of two receptor molecules here bind one molecule of the hormone. This binding activates the cytoplasmic domains of the complex, leading to further conduction of the signal to the interior of the cell (see p. 384). The small peptide hormone **insulin** is discussed in detail elsewhere (see pp. 76, 160). DNA-binding proteins (*transcription factors;* see p. 118) are decisively involved in regulating the metabolism and in differentiation processes. The structure and function of the **catabolite activator protein** (top left) and similar bacterial transcription factors have been particularly well investigated.

Catalysis. *Enzymes,* with more than 2000 known representatives, are the largest group of proteins in terms of numbers (see p. 88). The smallest enzymes have molecular masses of 10–15 kDa. Intermediatesized enzymes, such as **alcohol dehydrogenase** (top left) are around 100–200 kDa, and the largest—including **glutamine synthetase** with its 12 monomers (top right)—can reach more than 500 kDa.

Movement. The interaction between actin and myosin is responsible for muscle contraction and cell movement (see p. 332). **Myosin** (right), with a length of over 150 nm, is among the largest proteins there are. Actin filaments (**F-actin**) arise due to the polymerization of relatively small protein subunits (G-actin). Along with other proteins, **tropomyosin,** which is associated with F-actin, controls contraction.

Storage. Plants contain special **storage proteins**, which are also important for human nutrition (not shown). In animals, *muscle proteins* constitute a nutrient reserve that can be mobilized in emergencies.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Proteins

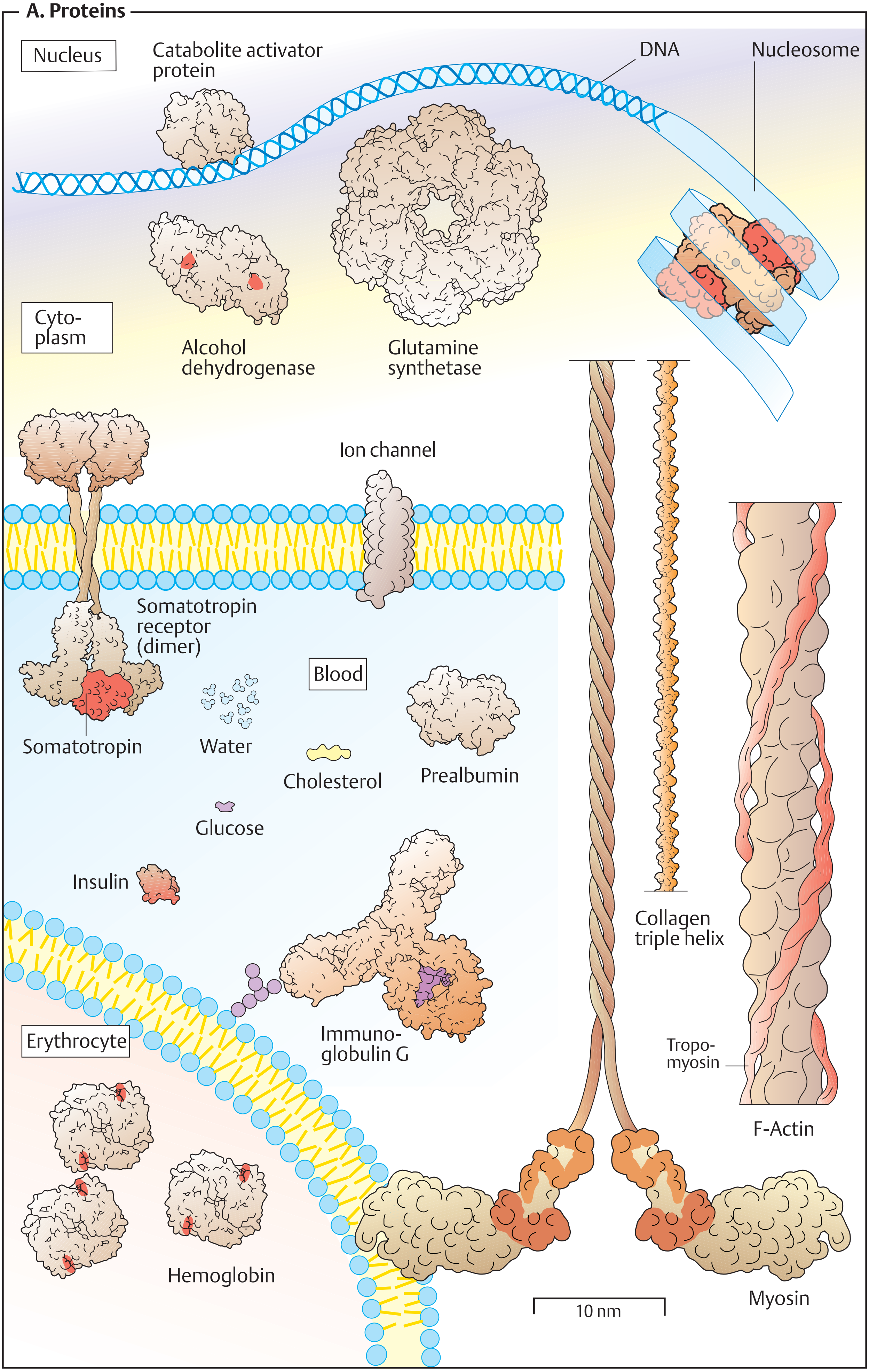
Nucleus
Catabolite activator protein
DNA
Nucleosome
Cyto-plasm
Alcohol dehydrogenase
Glutamine synthetase
Ion channel
Somatotropin receptor (dimer)
Somatotropin
Water
Blood
Cholesterol
Prealbumin
Glucose
Insulin
Collagen triple helix
Erythrocyte
Immuno-globulin G
Tropo-myosin
F-Actin
Hemoglobin
10 nm
Myosin

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Peptide bonds

A. Peptide bond ◑

The amino acid components of peptides and proteins are linked together by *amide* bonds (see p. 60) between α-carboxyl and α-amino groups. This type of bonding is therefore also known as **peptide bonding**. In the **dipeptide** shown here, the serine residue has a free ammonium group, while the carboxylate group in alanine is free. Since the amino acid with the free NH_3^+ group is named first, the peptide is known as **seryl alanine**, or in abbreviated form Ser-Ala or SA.

B. Resonance ◑

Like all acid–amide bonds, the peptide bond is **stabilized by resonance** (see p. 4). In the conventional notation (top right) it is represented as a combination of a C=O double bond with a C–N single bond. However, a C=N double bond with charges at O and N could also be written (middle). Both of these are only extreme cases of electron distribution, known as *resonance structures.* In reality, the π electrons are *delocalized* throughout all the atoms (bottom). As a mesomeric system, the peptide bond is *planar.* Rotation around the C–N bond would only be possible at the expense of large amounts of energy, and the bond is therefore *not freely rotatable.* Rotations are only possible around the single bonds marked with arrows. The state of these is expressed using the angles ϕ and ψ (see **D**). The plane in which the atoms of the peptide bond lie is highlighted in light blue here and on the following pages.

C. Peptide nomenclature ◑

Peptide chains have a *direction* and therefore two different ends. The amino terminus (**N terminus**) of a peptide has a free ammonium group, while the carboxy terminus (**C terminus**) is formed by the carboxylate group of the last amino acid. In peptides and proteins, the amino acid components are usually linked in linear fashion. To express the **sequence** of a peptide, it is therefore sufficient to combine the three-letter or single-letter abbreviations for the amino acid residues (see p. 60). This sequence always starts at the N terminus. For example, the peptide hormone *angiotensin II* (see p. 330) has the sequence Asp-Arg-Val-Tyr-Ile-His-Pro-Phe, or DRVYIHPF.

D. Conformational space of the peptide chain ○

With the exception of the terminal residues, every amino acid in a peptide is involved in *two* peptide bonds (one with the preceding residue and one with the following one). Due to the restricted rotation around the C–N bond, rotations are only possible around the N–C_α and C_α–C bonds (**2**). As mentioned above, these rotations are described by the dihedral angles ϕ (phi) and ψ (psi). The angle describes rotation around the N–C_α bond; ψ describes rotation around C_α–C—i. e., the position of the subsequent bond.

For steric reasons, only specific combinations of the dihedral angles are possible. These relationships can be illustrated clearly by a so-called ϕ/ψ *diagram* (**1**). Most combinations of ϕ and ψ are sterically “forbidden” (red areas). For example, the combination $\phi = 0°$ and $\psi = 180°$ (**4**) would place the two carbonyl oxygen atoms less than 115 pm apart—i. e., at a distance much smaller than the sum of their van der Waals radii (see p. 6). Similarly, in the case of $\phi = 180°$ and $\psi = 0°$ (**5**), the two NH hydrogen atoms would collide. The combinations located within the green areas are the only ones that are sterically feasible (e. g., **2** and **3**). The important secondary structures that are discussed in the following pages are also located in these areas. The conformations located in the yellow areas are energetically less favorable, but still possible.

The ϕ/ψ diagram (also known as a **Ramachandran plot**) was developed from modeling studies of small peptides. However, the conformations of most of the amino acids in proteins are also located in the permitted areas. The corresponding data for the small protein, insulin (see p. 76), are represented by black dots in **1**.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Peptide bonds

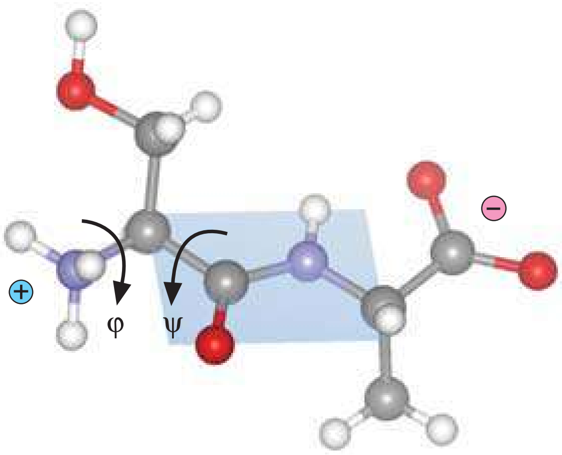

Seryl alanine
(Ser-Ala, $^{\oplus}H_3N$-Ser-Ala-$COO^{\ominus}$, SA)

B. Resonance

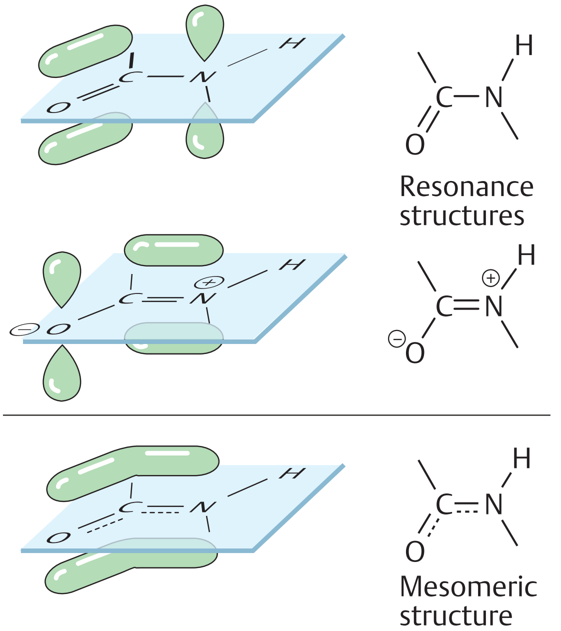
C — N
O
H
Resonance structures
C = N
⊕
⊖ O
Mesomeric structure

C. Peptide nomenclature

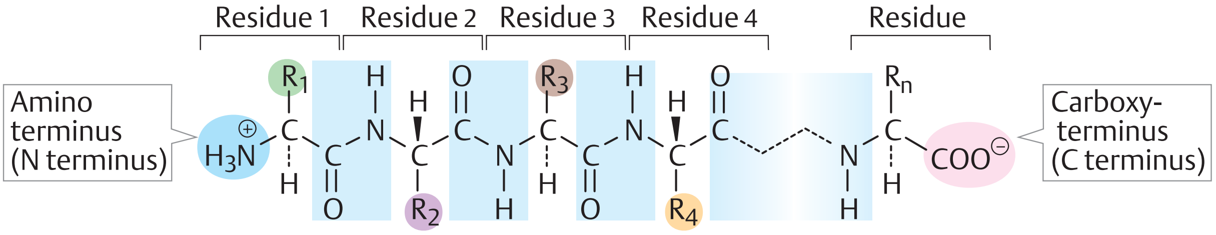
Residue 1
Residue 2
Residue 3
Residue 4
Residue
Amino terminus (N terminus)
H3N ⊕
R1
R2
R3
R4
Rn
COO⊖
Carboxy-terminus (C terminus)

D. Conformation space of the peptide chain

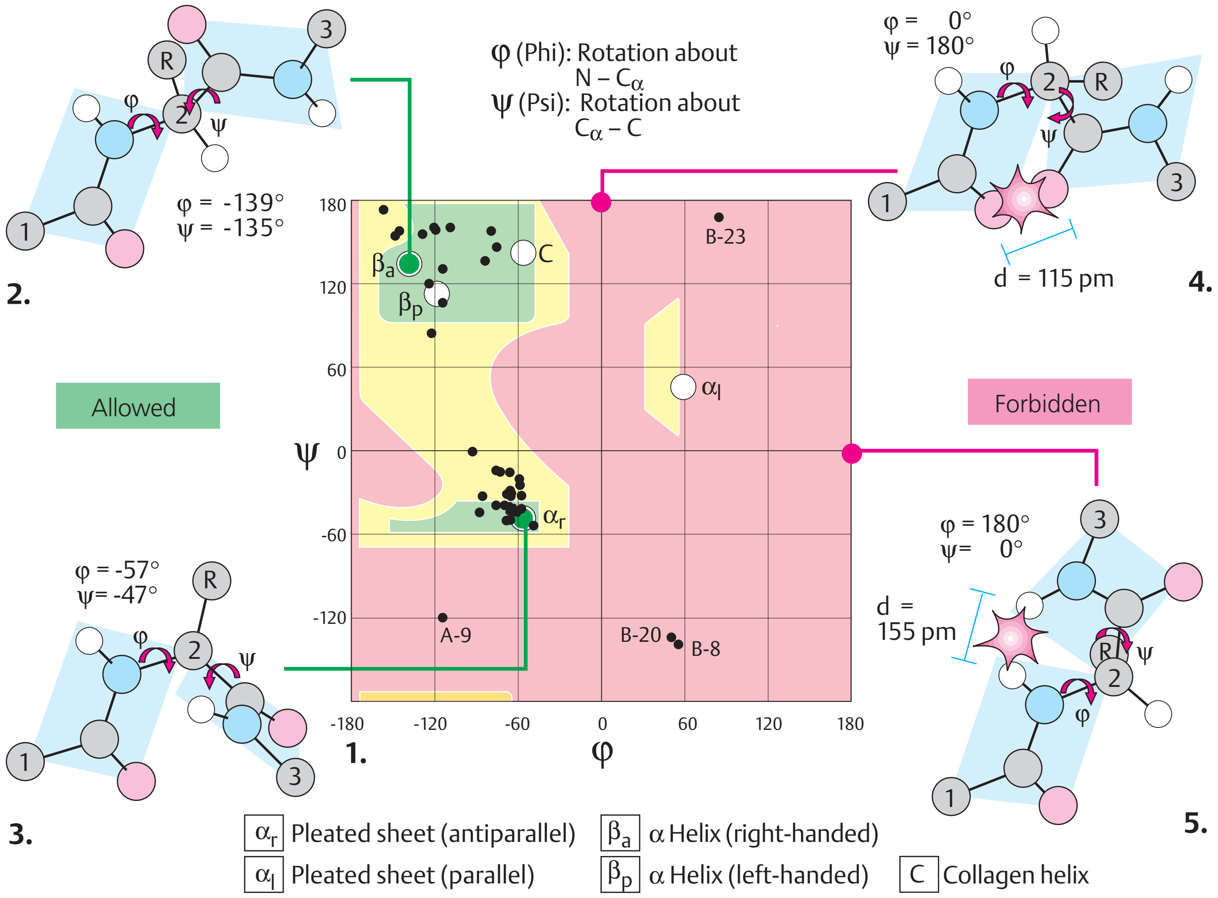

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Secondary structures

In proteins, specific combinations of the dihedral angles ϕ and ψ (see p. 66) are much more common than others. When several successive residues adopt one of these conformations, defined **secondary structures** arise, which are stabilized by hydrogen bonds either within the peptide chain or between neighboring chains. When a large part of a protein takes on a defined secondary structure, the protein often forms mechanically stable filaments or fibers. **Structural proteins** of this type (see p. 70) usually have characteristic amino acid compositions.

The most important secondary structural elements of proteins are discussed here first. The illustrations only show the course of the peptide chain; the side chains are omitted. To make the course of the chains clearer, the levels of the peptide bonds are shown as blue planes. The dihedral angles of the structures shown here are also marked in diagram D1 on p. 67.

A. α-Helix ◑

The **right-handed** α-helix (α_R) is one of the most common secondary structures. In this conformation, the peptide chain is wound like a screw. Each turn of the screw (the screw axis in shown in orange) covers approximately 3.6 amino acid residues. The *pitch* of the screw (i. e., the smallest distance between two equivalent points) is 0.54 nm. α-Helices are stabilized by almost linear *hydrogen bonds* between the NH and CO groups of residues, which are four positions apart from each another in the sequence (indicated by red dots; see p. 6). In longer helices, most amino acid residues thus enter into *two* H bonds. Apolar or amphipathic α-helices with five to seven turns often serve to anchor proteins in biological membranes (*transmembrane helices;* see p. 214).

The mirror image of the α_R helix, the **left-handed α-helix** (α_L), is rarely found in nature, although it would be energetically “permissible.”

B. Collagen helix ◑

Another type of helix occurs in the collagens, which are important constituents of the connectivetissue matrix (see pp. 70, 344). The **collagen helix** is **left-handed**, and with a pitch of 0.96 nm and 3.3 residues per turn, it is steeper than the α-helix. In contrast to the α-helix, H bonds are not possible *within* the collagen helix. However, the conformation is stabilized by the association of three helices to form a righthanded **collagen triple helix** (see p. 70).

C. Pleated-sheet structures ◑

Two additional, almost stretched, conformations of the peptide chain are known as β **pleated sheets**, as the peptide planes are arranged like a regularly folded sheet of paper. Again, H bonds can only form between *neighboring chains* (“strands”) in pleated sheets. When the two strands run in opposite directions (**1**), the structure is referred to as an **antiparallel pleated sheet** (β_a). When they run in the same direction (**2**), it is a **parallel pleated sheet** (β_p). In both cases, the α-C atoms occupy the highest and lowest points in the structure, and the side chains point alternately straight up or straight down (see p. 71 C). The β_a structure, with its almost linear H bonds, is energetically more favorable. In extended pleated sheets, the individual strands of the sheet are usually not parallel, but twisted relative to one another (see p. 74).

D. β Turns ◑

β **Turns** are often found at sites where the peptide chain changes direction. These are sections in which four amino acid residues are arranged in such a way that the course of the chain reverses by about 180° into the opposite direction. The two turns shown (types I and II) are particularly frequent. Both are stabilized by hydrogen bonds between residues 1 and 4. β Turns are often located between the individual strands of antiparallel pleated sheets, or between strands of pleated sheets and α helices.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. α Helix

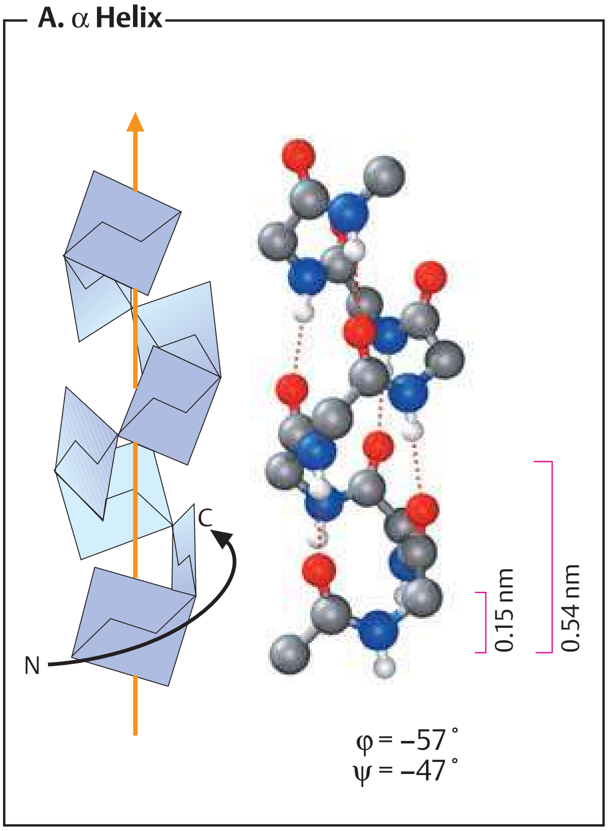
C
N
0.15 nm
0.54 nm
φ = −57°
ψ = −47°

B. Collagen helix

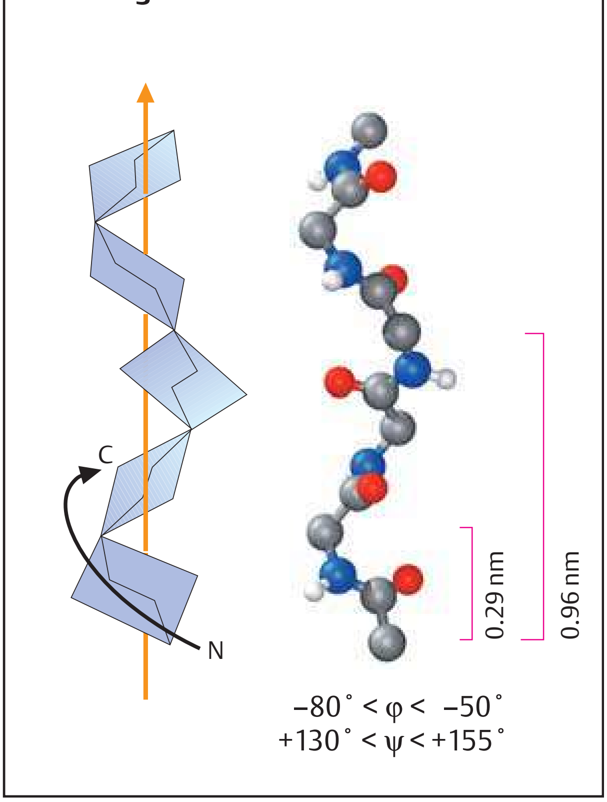
C
N
0.29 nm
0.96 nm
−80° < φ < −50°
+130° < ψ < +155°

C. Pleated-sheet structures

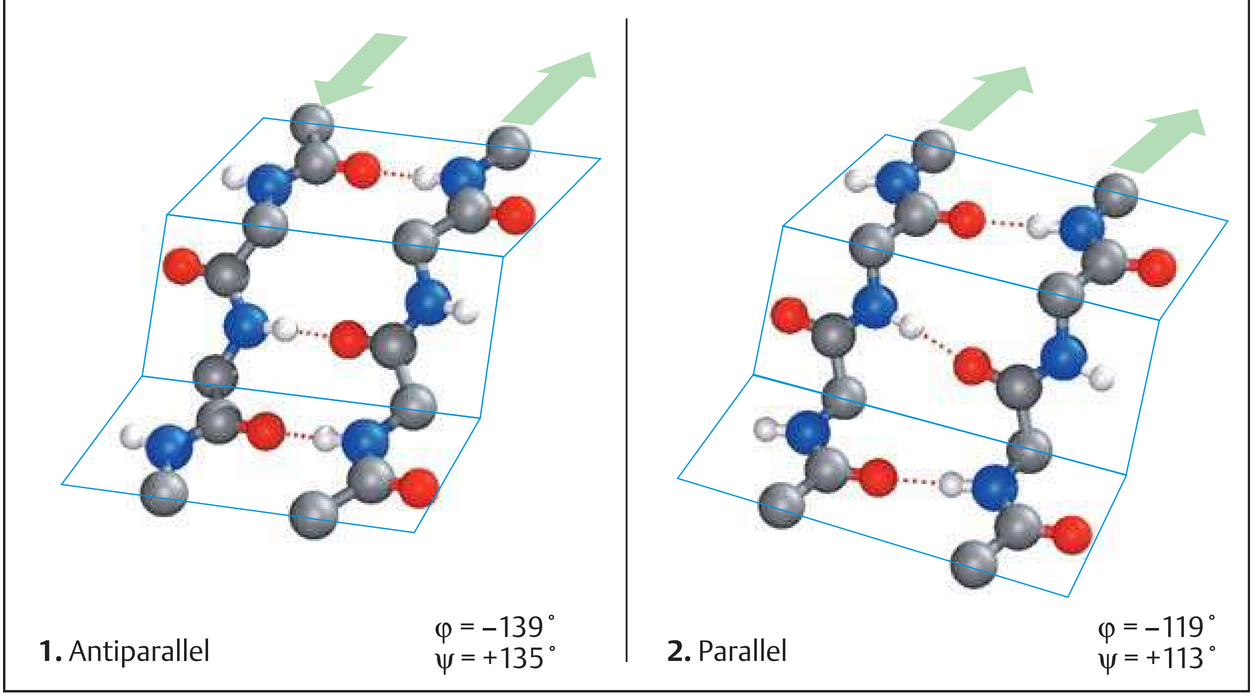
φ = −139°
ψ = +135°
φ = −119°
ψ = +113°

1. Antiparallel

2. Parallel

D. β Turns

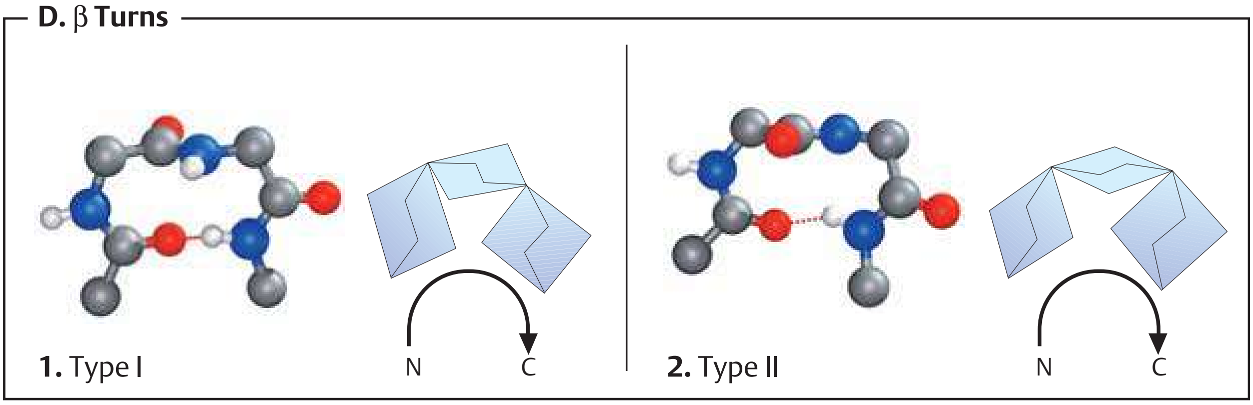

1. Type I

2. Type II

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Structural proteins

The **structural proteins** give extracellular structures mechanical stability, and are involved in the structure of the cytoskeleton (see p. 204). Most of these proteins contain a high percentage of specific secondary structures (see p. 68). For this reason, the amino acid composition of many structural proteins is also characteristic (see below).

A. α Keratin ○

α-Keratin is a structural protein that predominantly consists of α helices. Hair (wool), feathers, nails, claws and the hooves of animals consist largely of keratin. It is also an important component of the cytoskeleton (cytokeratin), where it appears in intermediate filaments (see p. 204).

In the keratins, large parts of the peptide chain show right-handed α-helical coiling. Two chains each form a left-handed **superhelix**, as is also seen in myosin (see p. 65). The superhelical keratin dimers join to form tetramers, and these aggregate further to form **protofilaments**, with a diameter of 3 nm. Finally, eight protofilaments then form an **intermediate filament**, with a diameter of 10 nm (see p. 204).

Similar keratin filaments are found in **hair**. In a single wool fiber with a diameter of about 20 μm, millions of filaments are bundled together within dead cells. The individual keratin helices are cross-linked and stabilized by numerous disulfide bonds (see p. 72). This fact is exploited in the *perming* of hair. Initially, the disulfide bonds of hair keratin are disrupted by reduction with thiol compounds (see p. 8). The hair is then styled in the desired shape and heat-dried. In the process, new disulfide bonds are formed by oxidation, which maintain the hairstyle for some time.

B. Collagen ◑

Collagen is the quantitatively most important protein in mammals, making up about 25% of the total protein. There are many different types of collagen, particularly in connective tissue. Collagen has an unusual amino acid composition. Approximately one-third of the amino acids are *glycine* (Gly), about 10% *proline* (Pro), and 10% *hydroxyproline* (Hyp). The two latter amino acids are only formed during collagen biosynthesis as a result of *posttranslational modification* (see p. 344).

The triplet Gly-X-Y (**2**) is constantly repeated in the sequence of collagen, with the X position often being occupied by Pro and the Y position by Hyp. The reason for this is that collagen is largely present as a **triple helix** made up of three individual collagen helices (**1).** In triple helices, every third residue lies on the inside of the molecule, where for steric reasons there is only room for glycine residues (**3**; the glycine residues are shown in yellow). Only a small section of a triple helix is illustrated here. The complete collagen molecule is approximately 300 nm long.

C. Silk fibroin ○

Silk is produced from the spun threads from silkworms (the larvae of the moth *Bombyx mori* and related species). The main protein in silk, **fibroin**, consists of antiparallel *pleated sheet structures* arranged one on top of the other in numerous layers (**1**). Since the amino acid side chains in pleated sheets point either straight up or straight down (see p. 68), only compact side chains fit between the layers. In fact, more than 80% of fibroin consists of glycine, alanine, and serine, the three amino acids with the shortest side chains. A typical repetitive amino acid sequence is *(Gly-Ala-Gly-Ala-Gly-Ser)*. The individual pleated sheet layers in fibroin are found to lie alternately 0.35 nm and 0.57 nm apart. In the first case, only glycine residues (R = H) are opposed to one another. The slightly greater distance of 0.57 nm results from repulsion forces between the side chains of alanine and serine residues (**2**).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. α-Keratin

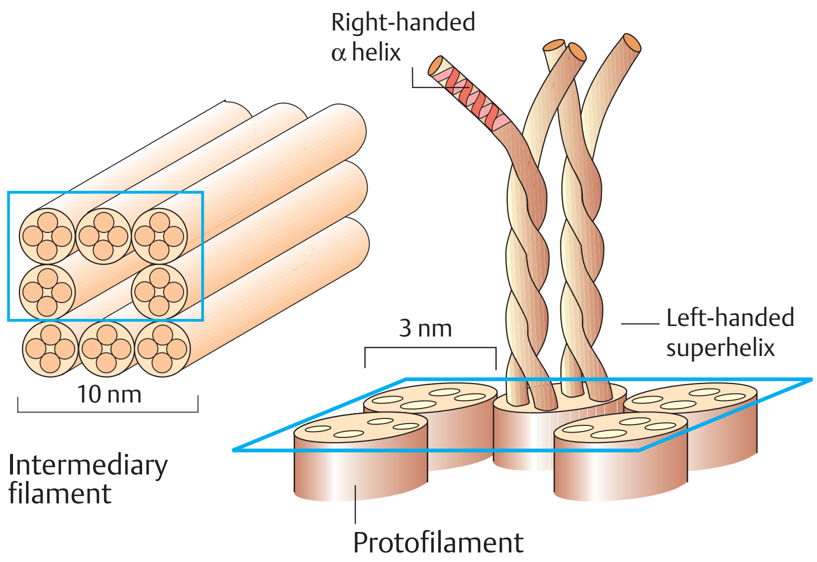
Right-handed
α helix
3 nm
Left-handed
superhelix
10 nm
Intermediary
filament
Protofilament

B. Collagen

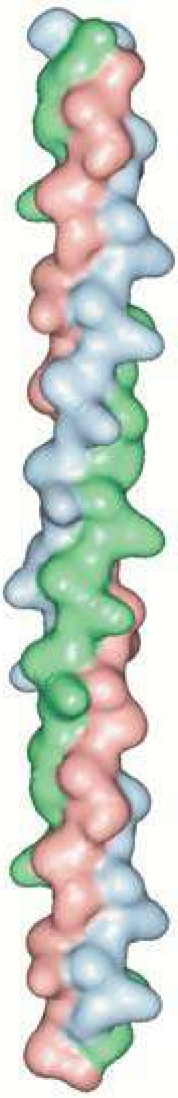

1. Triple helix (section)

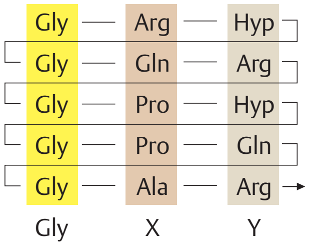

Gly	X	Y
Gly	Arg	Hyp
Gly	Gln	Arg
Gly	Pro	Hyp
Gly	Pro	Gln
Gly	Ala	Arg

2. Typical sequence

3. Triple helix
(view from above)

C. Silk fibroin

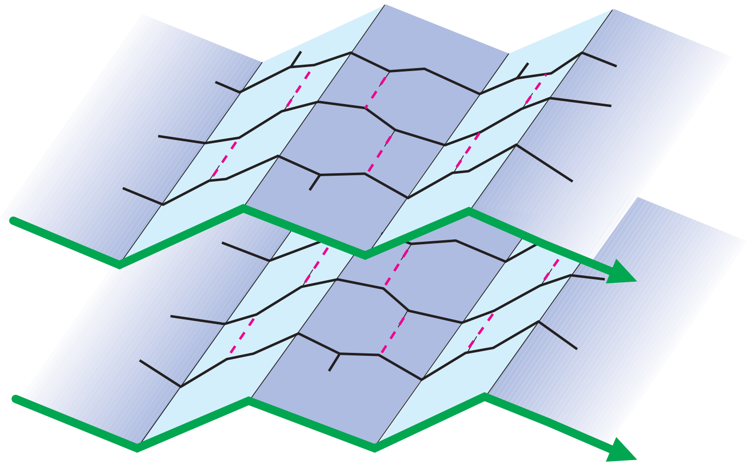

1. Spatial illustration

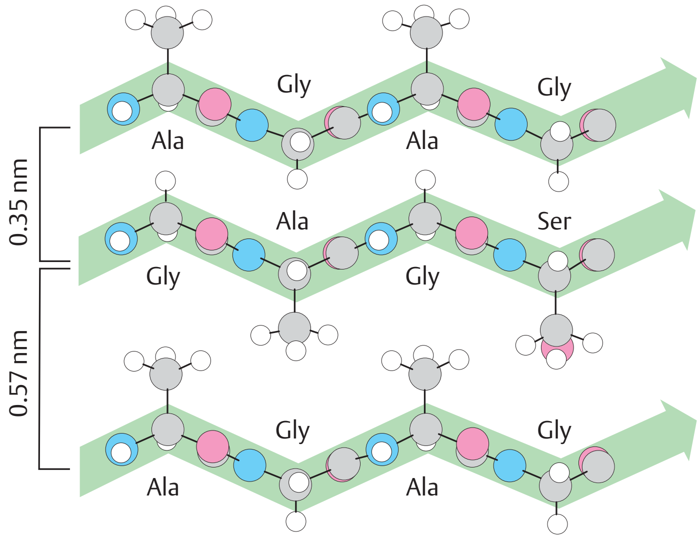
Gly
Ala
Gly
Ala
0.35 nm
Ala
Gly
Ser
Gly
0.57 nm
Gly
Ala
Gly
Ala

2. Front view

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Globular proteins

Soluble proteins have a more complex structure than the fibrous, completely insoluble structural proteins. The shape of soluble proteins is more or less spherical (globular). In their biologically active form, **globular proteins** have a defined spatial structure (the **native conformation**). If this structure is destroyed (**denaturation**; see p. 74), not only does the biological effect disappear, but the protein also usually precipitates in insoluble form. This happens, for example, when eggs are boiled; the proteins dissolved in the egg white are denatured by the heat and produce the solid egg white.

To illustrate protein conformations in a clear (but extremely simplified) way, *Richardson diagrams* are often used. In these diagrams, α-helices are symbolized by red cylinders or spirals and strands of pleated sheets by green arrows. Less structured areas of the chain, including the β-turns, are shown as sections of gray tubing.

A. Conformation-stabilizing interactions ◑

The native conformation of proteins is stabilized by a number of different interactions. Among these, only the **disulfide bonds** (**B**) represent covalent bonds. **Hydrogen bonds**, which can form inside secondary structures, as well as between more distant residues, are involved in all proteins (see p. 6). Many proteins are also stabilized by **complex formation** with metal ions (see pp. 76, 342, and 378, for example). The **hydrophobic effect** is particularly important for protein stability. In globular proteins, most hydrophobic amino acid residues are arranged in the interior of the structure in the native conformation, while the polar amino acids are mainly found on the surface (see pp. 28, 76).

B. Disulfide bonds ◑

Disulfide bonds arise when the SH groups of two cysteine residues are covalently linked as a dithiol by oxidation. Bonds of this type are only found (with a few exceptions) in extracellular proteins, because in the interior of the cell *glutathione* (see p. 284) and other reducing compounds are present in such high concentrations that disulfides would be reductively cleaved again. The small plant protein *crambin* (46 amino acids) contains three disulfide bonds and is therefore very stable. The high degree of stability of insulin (see p. 76) has a similar reason.

C. Protein dynamics ◑

The conformations of globular proteins are not rigid, but can change dramatically on binding of ligands or in contact with other proteins. For example, the enzyme *adenylate kinase* (see p. 336) has a mobile domain (domain = independently folded partial structure), which folds shut after binding of the substrate (yellow). The larger domain (bottom) also markedly alters its conformation. There are large numbers of **allosteric proteins** of this type. This group includes, for example, *hemoglobin* (see p. 280), *calmodulin* (see p. 386), and many allosteric enzymes such as *aspartate carbamoyltransferase* (see p. 116).

D. Folding patterns ○

The globular proteins show a high degree of variability in folding of their peptide chains. Only a few examples are shown here. Purely helically folded proteins such as *myoglobin* (**1**; see p. 74, heme yellow) are rare. In general, pleated sheet and helical elements exist alongside each other. In the hormone-binding domain of the *estrogen receptor* (**2**; see p. 378), a small, two-stranded pleated sheet functions as a "cover" for the hormone binding site (estradiol yellow). In *flavodoxin*, a small flavoprotein with a redox function (**3**; FMN yellow), a fan-shaped, pleated sheet made up of five parallel strands forms the core of the molecule. The conformation of the β subunit of the G-protein *transducin* (**4**; see pp. 224, 358) is very unusual. Seven pleated sheets form a large, symmetrical "β propeller." The *N*-terminal section of the protein contains one long and one short helix.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Conformation-stabilizing interactions

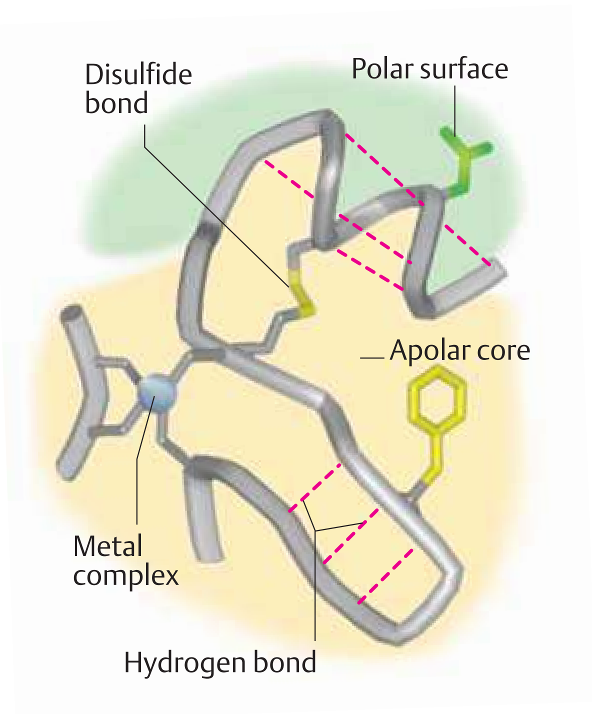
Disulfide bond
Polar surface
Apolar core
Metal complex
Hydrogen bond

B. Disulfide bonds

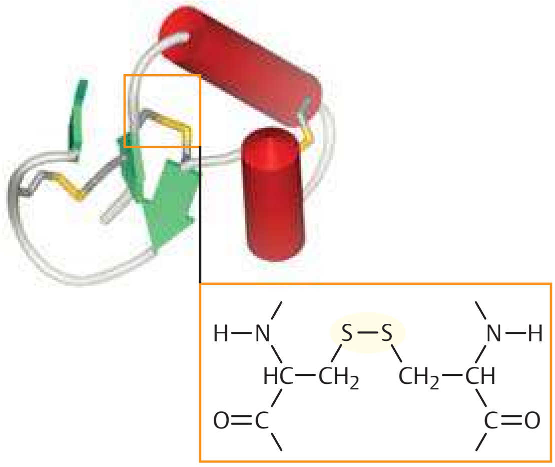
H—N
S—S
N—H
HC—CH2
CH2—CH
O=C
C=O

C. Protein dynamics

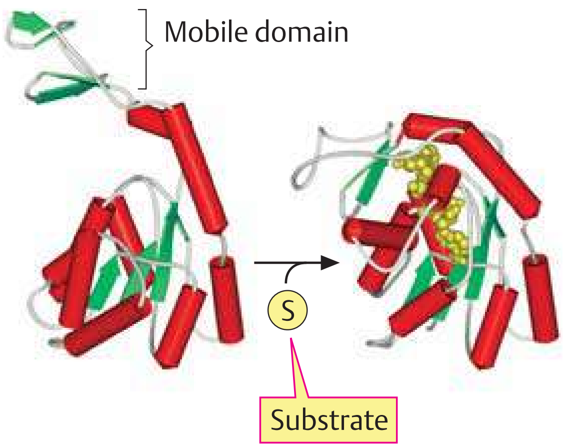
Mobile domain
S
Substrate

Adenylate kinase

D. Folding patterns

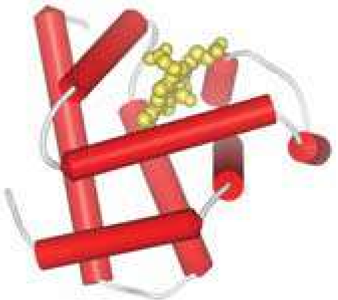

1. Myoglobin

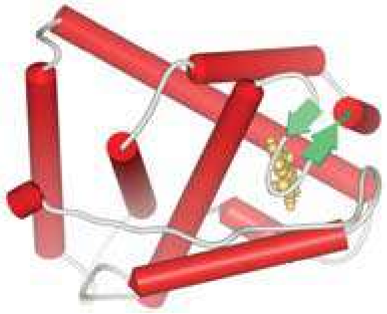

2. Estrogen receptor (domain)

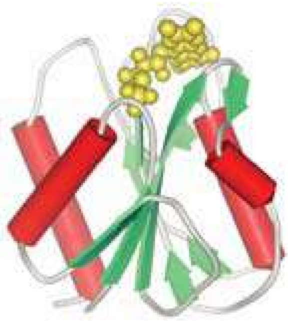

3. Flavodoxin

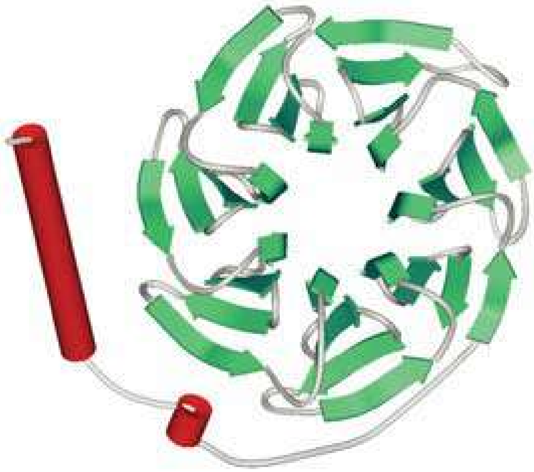

4. Transducin (β subunit)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Protein folding

Information about the biologically active (*native*) conformation of proteins is already encoded in their amino acid sequences. The native forms of many proteins arise spontaneously in the test tube and within a few minutes. Nevertheless, there are special auxiliary proteins (chaperonines) that support the folding of other proteins in the conditions present within the cell (see p. 232). An important goal of biochemistry is to understand the laws governing **protein folding**. This would make it possible to predict the conformation of a protein from the easily accessible DNA sequence (see p. 260).

A. Folding and denaturation of ribonuclease A ◑

The **folding** of proteins to the native form is favored under physiological conditions. The native conformation is lost, as the result of **denaturation**, at extreme pH values, at high temperatures, and in the presence of organic solvents, detergents, and other denaturing substances, such as urea.

The fact that a denatured protein can spontaneously return to its native conformation was demonstrated for the first time with **ribonuclease**, a digestive enzyme (see p. 266) consisting of 124 amino acids. In the native form (top right), there are extensive pleated sheet structures and three α helices. The eight cysteine residues of the protein are forming four disulfide bonds. Residues His-12, Lys-41 and His-119 (pink) are particularly important for catalysis. Together with additional amino acids, they form the enzyme's *active center.*

The disulfide bonds can be reductively cleaved by *thiols* (e. g., mercaptoethanol, $HO\text{-}CH_2\text{-}CH_2\text{-}SH$). If *urea* at a high concentration is also added, the protein unfolds completely. In this form (left), it is up to 35 nm long. Polar (green) and apolar (yellow) side chains are distributed randomly. The denatured enzyme is completely inactive, because the catalytically important amino acids (pink) are too far away from each other to be able to interact with each other and with the substrate.

When the urea and thiol are removed by dialysis (see p. 78), secondary and tertiary structures develop again spontaneously. The cysteine residues thus return to a sufficiently close spatial vicinity that disulfide bonds can once again form under the oxidative effect of atmospheric oxygen. The active center also reestablishes itself. In comparison with the denatured protein, the native form is astonishingly compact, at 4.5 2.5 nm. In this state, the apolar side chains (yellow) predominate in the interior of the protein, while the polar residues are mainly found on the surface. This distribution is due to the "hydrophobic effect" (see p. 28), and it makes a vital contribution to the stability of the native conformation (**B**).

B. Energetics of protein folding ○

The **energetics** of protein folding are not at present satisfactorily understood. Only a simplified model is discussed here. The conformation of a molecule is stable in any given conditions if the change in its free enthalpy during folding (ΔG_{fold}) is negative (see p. 16). The magnitude of the folding enthalpy is affected by several factors. The main factor working *against* folding is the strong increase in the ordering of the molecule involved. As discussed on p. 20, this leads to a negative change in entropy of ΔS_{conf} and therefore to a strongly positive entropy term $-T \quad \Delta S$ (violet arrow). By contrast, the covalent and noncovalent bonds in the interior of the protein have a *stabilizing* influence. For this reason, the change in folding enthalpy ΔH_{fold} is negative (red arrow). A third factor is the change in the system's entropy due to the hydrophobic effect. During folding, the degree of order in the *surrounding water* decreases—i. e., ΔS_{water} is positive and therefore $-T \quad \Delta S$ is negative (blue arrow). When the sum of these effects is negative (green arrow), the protein folds spontaneously into its native conformation.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

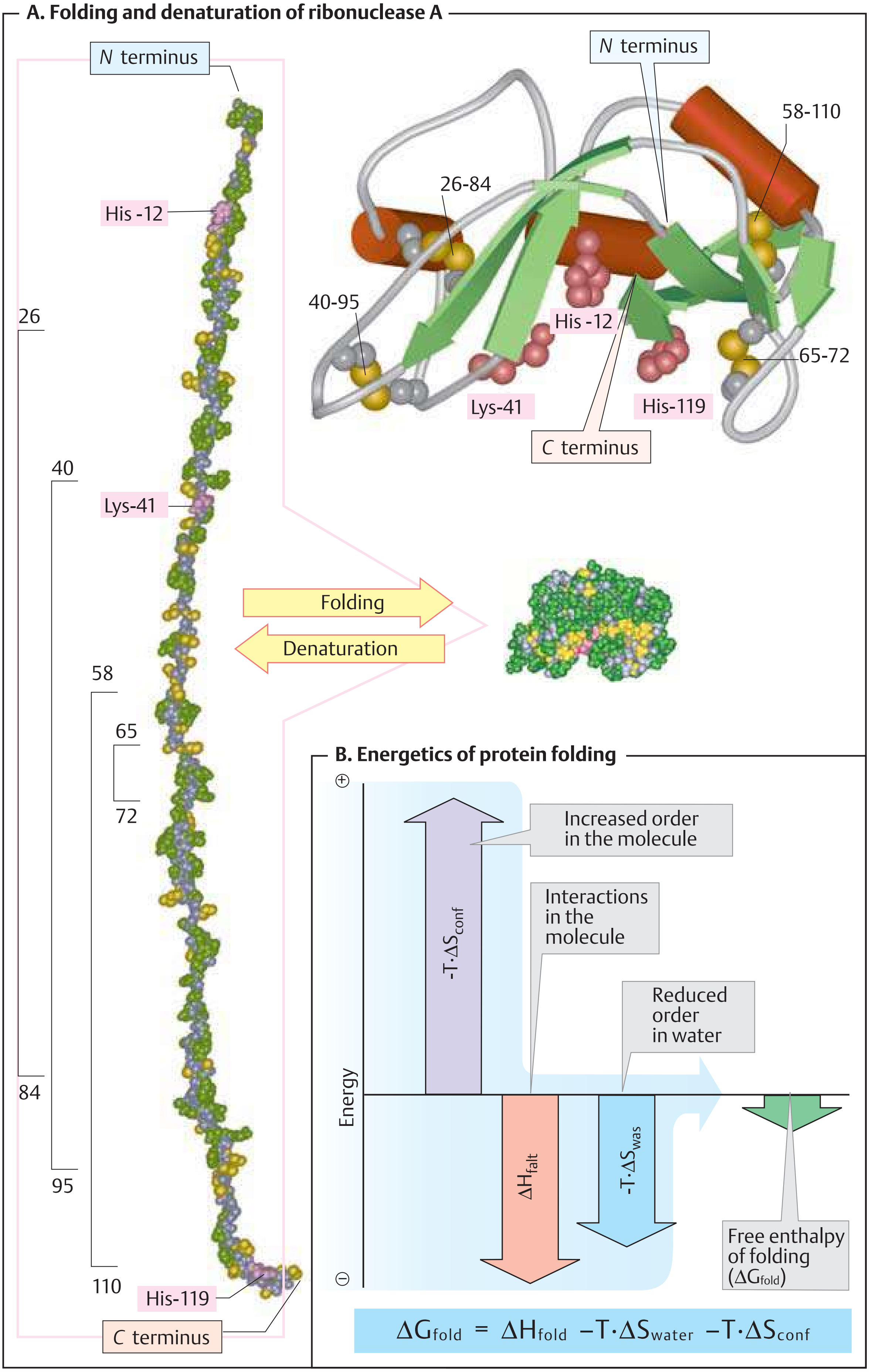

A. Folding and denaturation of ribonuclease A

N terminus
His -12
26
40
Lys-41
58
65
72
84
95
110
His-119
C terminus
N terminus
58-110
26-84
40-95
His -12
65-72
Lys-41
His-119
C terminus
Folding
Denaturation

B. Energetics of protein folding

⊕
⊖
Energy
Increased order in the molecule
Interactions in the molecule
Reduced order in water
Free enthalpy of folding (ΔG_{fold})
$-T \cdot \Delta S_{conf}$
ΔH_{falt}
$-T \cdot \Delta S_{was}$

$$\Delta G_{fold} = \Delta H_{fold} - T \cdot \Delta S_{water} - T \cdot \Delta S_{conf}$$

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Molecular models: insulin

The opposite page presents models of insulin, a small protein. The biosynthesis and function of this important hormone are discussed elsewhere in this book (pp. 160, 388). Monomeric insulin consists of 51 amino acids, and with a molecular mass of 5.5 kDa it is only half the size of the smallest enzymes. Nevertheless, it has the typical properties of a globular protein.

Large quantities of pure insulin are required for the treatment of *diabetes mellitus* (see p. 160). The annual requirement for insulin is over 500 kg in a country the size of Germany. Formerly, the hormone had to be obtained from the pancreas of slaughtered animals in a complicated and expensive procedure. **Human insulin**, which is produced by *overexpression* in genetically engineered bacteria, is now mainly used (see p. 262).

A. Structure of insulin ○

There are various different structural levels in proteins, and these can be briefly discussed again here using the example of insulin.

The **primary structure** of a protein is its amino acid sequence. During the biosynthesis of insulin in the pancreas, a continuous peptide chain with 84 residues is first synthesized—*proinsulin* (see p. 160). After folding of the molecule, the three disulfide bonds are first formed, and residues 31 to 63 are then proteolytically cleaved releasing the so-called *C peptide*. The molecule that is left over (**1**) now consists of two peptide chains, the *A chain* (21 residues, shown in yellow) and the *B chain* (30 residues, orange). One of the disulfide bonds is located inside the A chain, and the two others link the two chains together.

Secondary structures are regions of the peptide chain with a defined conformation (see p. 68) that are stabilized by H-bonds. In insulin (**2**), the α-helical areas are predominant, making up 57% of the molecule; 6% consists of β-pleated-sheet structures, and 10% of β-turns, while the remainder (27%) cannot be assigned to any of the secondary structures.

The three-dimensional conformation of a protein, made up of secondary structural elements and unordered sections, is referred to as the **tertiary structure**. In insulin, it is compact and wedge-shaped (**B**). The tip of the wedge is formed by the B chain, which changes its direction at this point.

Quaternary structure. Due to non-covalent interactions, many proteins assemble to form symmetrical complexes (oligomers). The individual components of oligomeric proteins (usually 2–12) are termed *subunits* or *monomers.* Insulin also forms quaternary structures. In the blood, it is partly present as a dimer. In addition, there are also hexamers stabilized by Zn^{2+} ions (light blue) (**3**), which represent the form in which insulin is stored in the pancreas (see p. 160).

B. Insulin (monomer) ○

The van der Waals model of monomeric insulin (**1**) once again shows the wedge-shaped tertiary structure formed by the two chains together. In the second model (**3**, bottom), the side chains of polar amino acids are shown in blue, while apolar residues are yellow or pink. This model emphasizes the importance of the "hydrophobic effect" for protein folding (see p. 74). In insulin as well, most hydrophobic side chains are located on the inside of the molecule, while the hydrophilic residues are located on the surface. Apparently in contradiction to this rule, several apolar side chains (pink) are found on the surface. However, all of these residues are involved in hydrophobic interactions that stabilize the dimeric and hexameric forms of insulin.

In the third model (**2**, right), the colored residues are those that are located on the surface and occur *invariably* (red) or *almost invariably* (orange) in all known insulins. It is assumed that amino acid residues that are not replaced by other residues during the course of evolution are essential for the protein's function. In the case of insulin, almost all of these residues are located on one side of the molecule. They are probably involved in the binding of the hormone to its receptor (see p. 224).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Structure of insulin

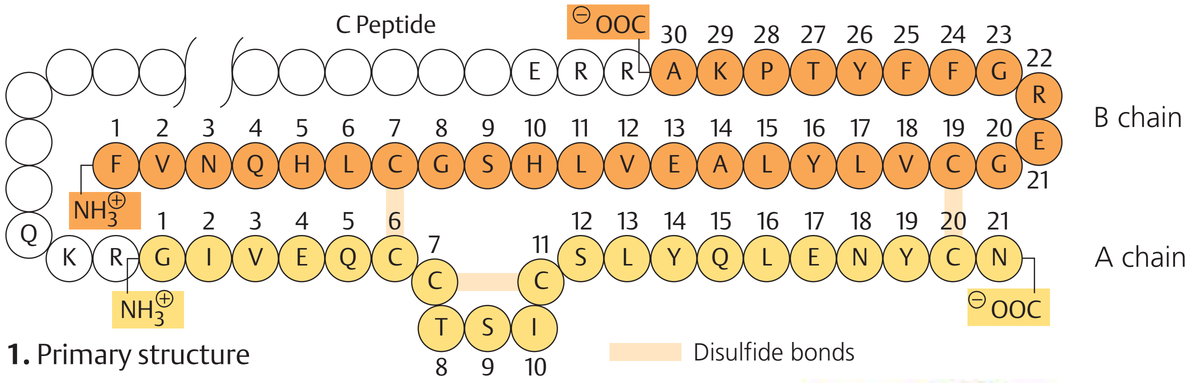
C Peptide
⊖OOC
30 29 28 27 26 25 24 23 22
A K P T Y F F G R
E R R
B chain
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
F V N Q H L C G S H L V E A L Y L V C G E
NH3⊕
Q
K R
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
G I V E Q C C T S I C S L Y Q L E N Y C N
NH3⊕
⊖OOC
A chain
Disulfide bonds

1. Primary structure

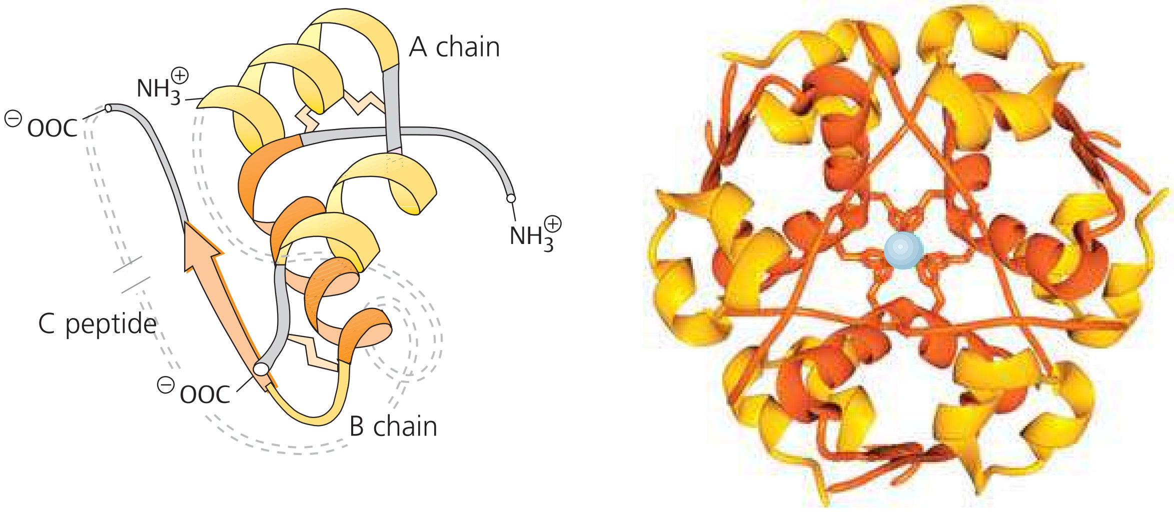
A chain
NH3⊕
⊖OOC
NH3⊕
C peptide
⊖OOC
B chain

2. Secondary and tertiary structure

3. Quaternary structure

B. Insulin (monomer)

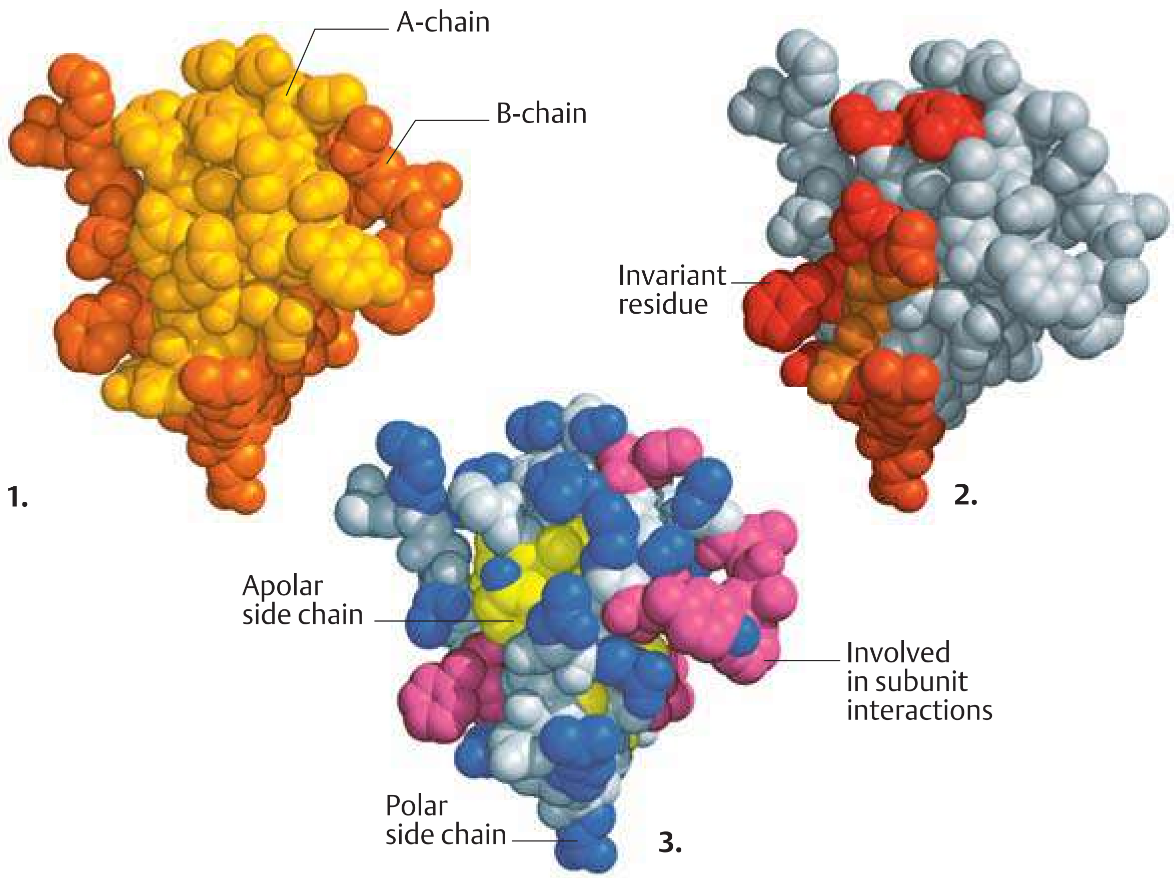
A-chain
B-chain
1.
Invariant residue
2.
Apolar side chain
Involved in subunit interactions
Polar side chain
3.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Isolation and analysis of proteins

Purified proteins are nowadays required for a wide variety of applications in research, medicine, and biotechnology. Since the globular proteins in particular are very unstable (see p. 72), purification is carried out at low temperatures (0–5 °C) and particularly gentle separation processes are used. A few of the methods of purifying and characterizing proteins are discussed on this page.

A. Salt precipitation ○

The solubility of proteins is strongly dependent on the salt concentration (*ionic strength*) of the medium. Proteins are usually poorly soluble in pure water. Their solubility increases as the ionic strength increases, because more and more of the well-hydrated anorganic ions (blue circles) are bound to the protein's surface, preventing aggregation of the molecules (**salting in**). At very high ionic strengths, the salt withdraws the hydrate water from the proteins and thus leads to aggregation and precipitation of the molecules (**salting out**). For this reason, adding salts such as ammonium sulfate $(NH_4)_2SO_4$ makes it possible to separate proteins from a mixture according to their degree of solubility (fractionation).

B. Dialysis ○

Dialysis is used to remove lower-molecular components from protein solutions, or to exchange the medium. Dialysis is based on the fact that due to their size, protein molecules are unable to pass through the pores of a **semipermeable membrane**, while lower-molecular substances distribute themselves evenly between the inner and outer spaces over time. After repeated exchanging of the external solution, the conditions inside the *dialysis tube* (salt concentration, pH, etc.) will be the same as in the surrounding solution.

C. Gel filtration ○

Gel permeation chromatography ("gel filtration") separates proteins according to their size and shape. This is done using a *chromatography column,* which is filled with spherical *gel particles* (diameter 10–500 µm) of polymeric material (shown schematically in **1a**). The insides of the particles are traversed by channels that have defined diameters. A protein mixture is then introduced at the upper end of the column (**1b**) and *elution* is carried out by passing a buffer solution through the column. Large protein molecules (red) are unable to penetrate the particles, and therefore pass through the column quickly. Medium-sized (green) and small particles (blue) are delayed for longer or shorter periods (**1c**). The proteins can be collected separately from the effluent (*eluate*) (**2**). Their elution volume V_e depends mainly on their molecular mass (**3**).

D. SDS gel electrophoresis ○

The most commonly used procedure for checking the purity of proteins is sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). In electrophoresis, molecules move in an electrical field (see p. 276). Normally, the speed of their movement depends on three factors—their size, their shape, and their electrical charge.

In SDS-PAGE, the protein mixture is treated in such a way that only the molecules' mass affects their movement. This is achieved by adding *sodium dodecyl sulfate* ($C_{12}H_{25}$-OSO_3Na), the sulfuric acid ester of lauryl alcohol (dodecyl alcohol). SDS is a *detergent* with strongly amphipathic properties (see p. 28). It separates oligomeric proteins into their subunits and denatures them. SDS molecules bind to the unfolded peptide chains (ca. 0.4 g SDS / g protein) and give them a strongly negative charge. To achieve complete denaturation, thiols are also added in order to cleave the disulfide bonds (**1**).

Following electrophoresis, which is carried out in a vertically arranged gel of polymeric acrylamide (**2**), the separated proteins are made visible by staining. In example (**3**), the following were separated: **a**) a cell extract with hundreds of different proteins, **b**) a protein purified from this, and **c**) a mixture of proteins with known masses.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

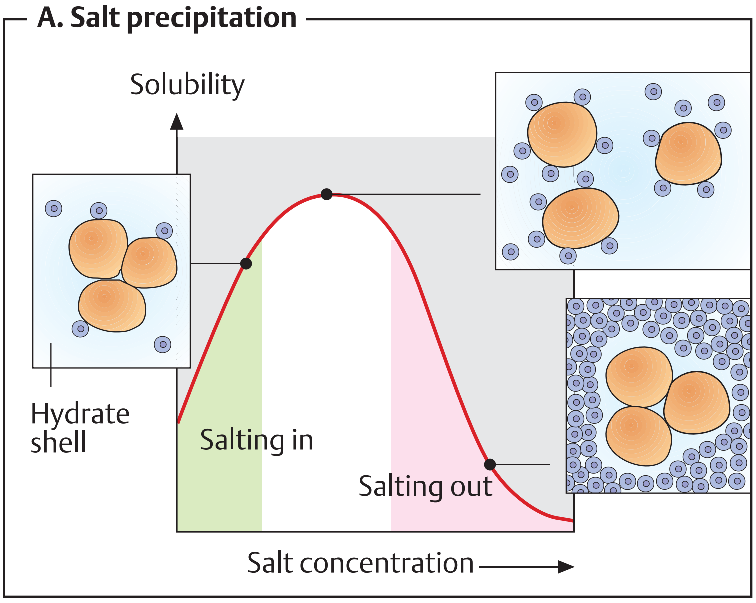

A. Salt precipitation

Solubility
Hydrate shell
Salting in
Salting out
Salt concentration

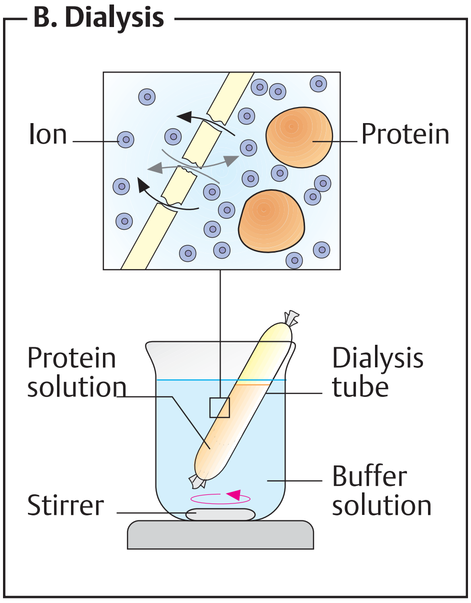

B. Dialysis

Ion
Protein
Protein solution
Dialysis tube
Buffer solution
Stirrer

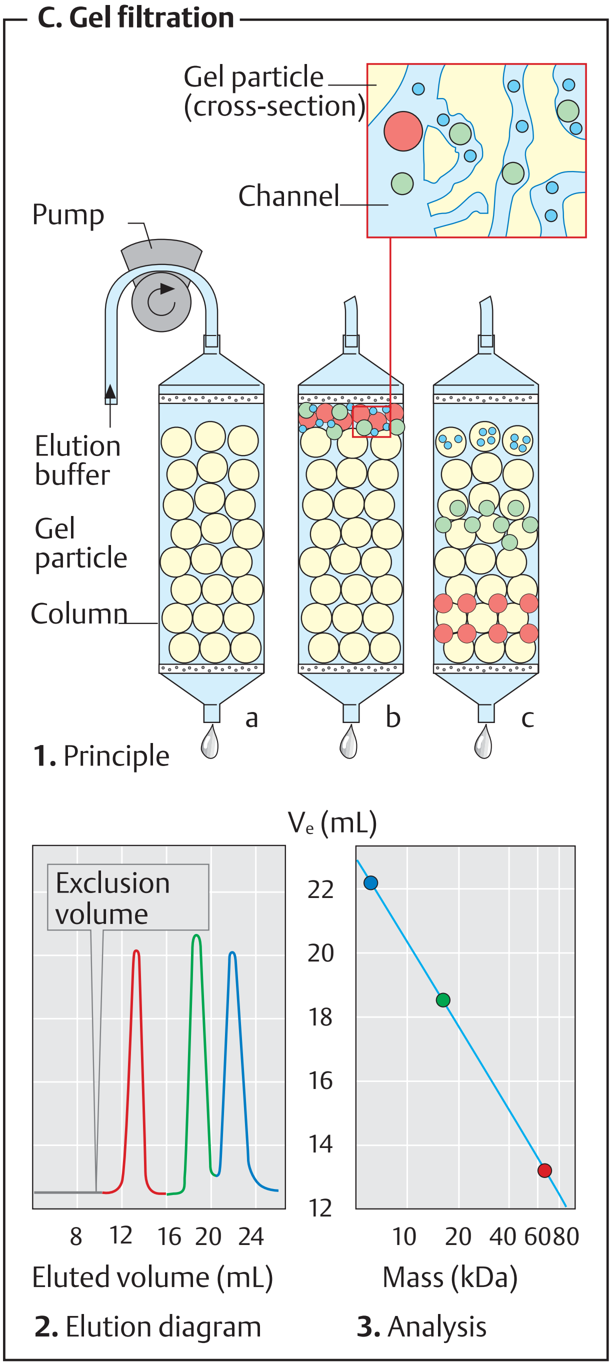

C. Gel filtration

Gel particle (cross-section)
Channel
Pump
Elution buffer
Gel particle
Column
a
b
c
1. Principle
Exclusion volume
8
12
16
20
24
Eluted volume (mL)
2. Elution diagram
Vₑ (mL)
22
20
18
16
14
12
10
20
40
60
80
Mass (kDa)
3. Analysis

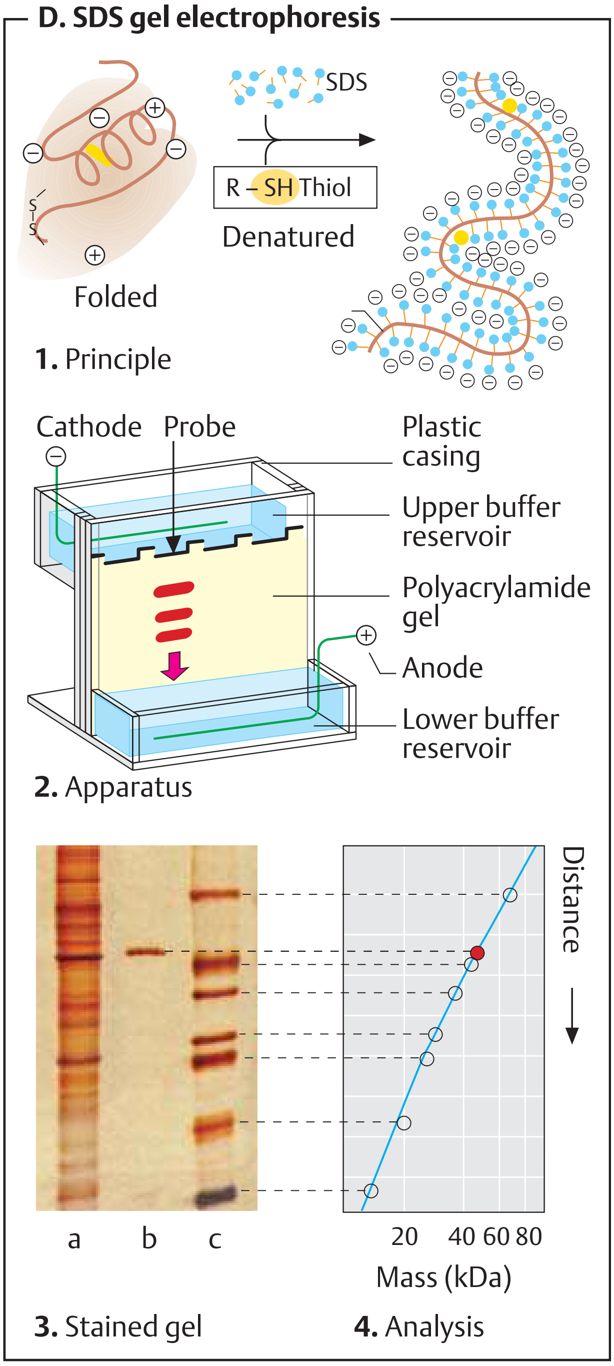

D. SDS gel electrophoresis

SDS
R – SH Thiol
Denatured
Folded
S
S
1. Principle
Cathode
Probe
Plastic casing
Upper buffer reservoir
Polyacrylamide gel
Anode
Lower buffer reservoir
2. Apparatus
a
b
c
3. Stained gel
Distance
20
40
60
80
Mass (kDa)
4. Analysis

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Bases and nucleotides

The nucleic acids play a central role in the storage and expression of genetic information (see p. 236). They are divided into two major classes: **deoxyribonucleic acid (DNA)** functions solely in information storage, while **ribonucleic acids (RNAs)** are involved in most steps of gene expression and protein biosynthesis. All nucleic acids are made up from **nucleotide components**, which in turn consist of a *base*, a *sugar*, and a *phosphate residue*. DNA and RNA differ from one another in the type of the sugar and in one of the bases that they contain.

A. Nucleic acid bases ◑

The bases that occur in nucleic acids are *aromatic* heterocyclic compounds derived from either **pyrimidine** or **purine**. Five of these bases are the main components of nucleic acids in all living creatures. The purine bases **adenine** (abbreviation Ade, *not* "A"!) and **guanine** (Gua) and the pyrimidine base **cytosine** (Cyt) are present in both RNA *and* DNA. In contrast, **uracil** (Ura) is only found in RNA. In DNA, **uracil** is replaced by thymine (Thy), the 5-methyl derivative of uracil. 5-methylcytosine also occurs in small amounts in the DNA of the higher animals. A large number of other modified bases occur in tRNA (see p. 82) and in other types of RNA.

B. Nucleosides, nucleotides ◑

When a nucleic acid base is N-glycosidically linked to ribose or 2-deoxyribose (see p. 38), it yields a **nucleoside**. The nucleoside **adenosine** (abbreviation: A) is formed in this way from adenine and ribose, for example. The corresponding derivatives of the other bases are called *guanosine* (G), *uridine* (U), *thymidine* (T) and *cytidine* (C). When the sugar component is 2-deoxyribose, the product is a **deoxyribonucleoside**—e. g., 2′-deoxyadenosine (dA, not shown). In the cell, the 5′OH group of the sugar component of the nucleoside is usually esterified with phosphoric acid. 2′-Deoxythymidine (dT) therefore gives rise to **2′-deoxythymidine-5′-monophosphate (dTMP)**, one of the components of DNA (**2**). If the 5′phosphate residue is linked via an acid–anhydride bond to additional phosphate residues, it yields nucleoside diphosphates and triphosphates—e. g., ADP and ATP, which are important coenzymes in energy metabolism (see p. 106). All of these nucleoside phosphates are classified as **nucleotides**.

In nucleosides and nucleotides, the pentose residues are present in the furanose form (see p. 34). The sugars and bases are linked by an *N*-**glycosidic bond** between the C-1 of the sugar and either the N-9 of the purine ring or N-1 of the pyrimidine ring. This bond always adopts the β-configuration.

C. Oligonucleotides, polynucleotides ◑

Phosphoric acid molecules can form acid–anhydride bonds with each other. It is therefore possible for two nucleotides to be linked via the phosphate residues. This gives rise to *dinucleotides with a phosphoric acid–anhydride* structure. This group includes the coenzymes $NAD(P)^+$ and CoA, as well as the flavin derivative **FAD** (**1**; see p. 104).

If the phosphate residue of a nucleotide reacts with the 3′-OH group of a second nucleotide, the result is a *dinucleotide with a phosphoric acid diester structure*. Dinucleotides of this type have a free phosphate residue at the *5′end* and a free OH group at the *3′ end*. They can therefore be extended with additional mononucleotides by adding further phosphoric acid diester bonds. This is the way in which **oligonucleotides**, and ultimately **polynucleotides**, are synthesized.

Polynucleotides consisting of ribonucleotide components are called **ribonucleic acid (RNA)**, while those consisting of deoxyribonucleotide monomers are called **deoxyribonucleic acid (DNA**; see p. 84). To describe the structure of polynucleotides, the abbreviations for the *nucleoside* components are written from left to right *in the 5′→3′* direction. The position of the phosphate residue is also sometimes indicated by a "p". In this way, the structure of the RNA segment shown Fig. **2** can be abbreviated as ..pUpG.. or simply as ..^{UG}.. .

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Nucleic acid bases

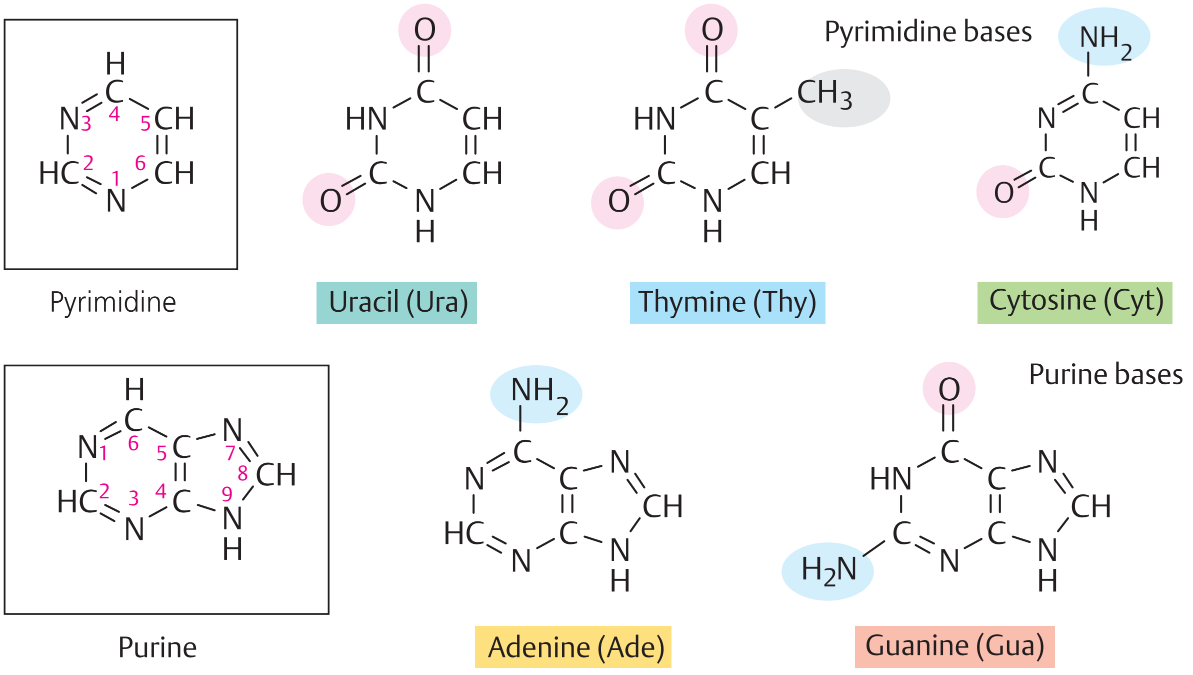
Pyrimidine bases
Pyrimidine
Uracil (Ura)
Thymine (Thy)
Cytosine (Cyt)
Purine bases
Purine
Adenine (Ade)
Guanine (Gua)

B. Nucleosides, nucleotides

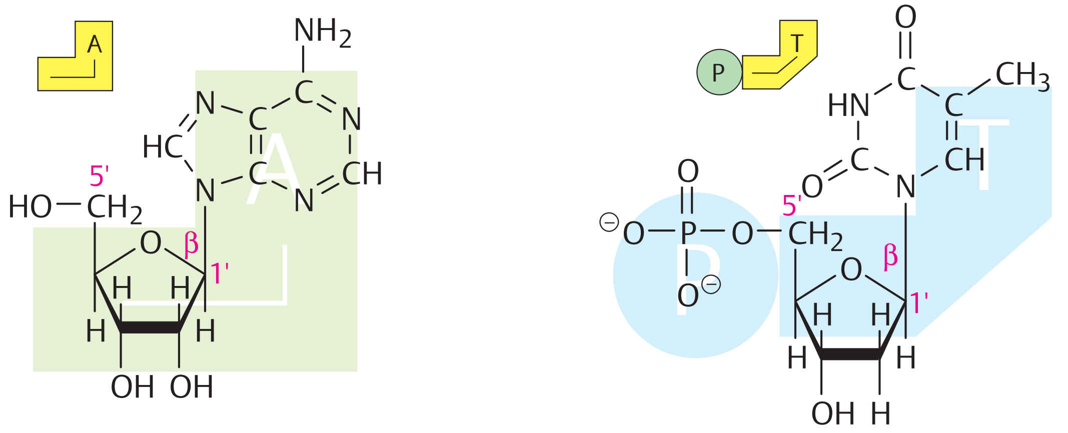
A
NH2
5'
β
1'
OH OH
P
T
5'
β
1'
OH H

1. Adenosine (Ado)

2. 2'-Deoxythymidine 5'-monophosphate (dtMP)

C. Oligonucleotides, polynucleotides

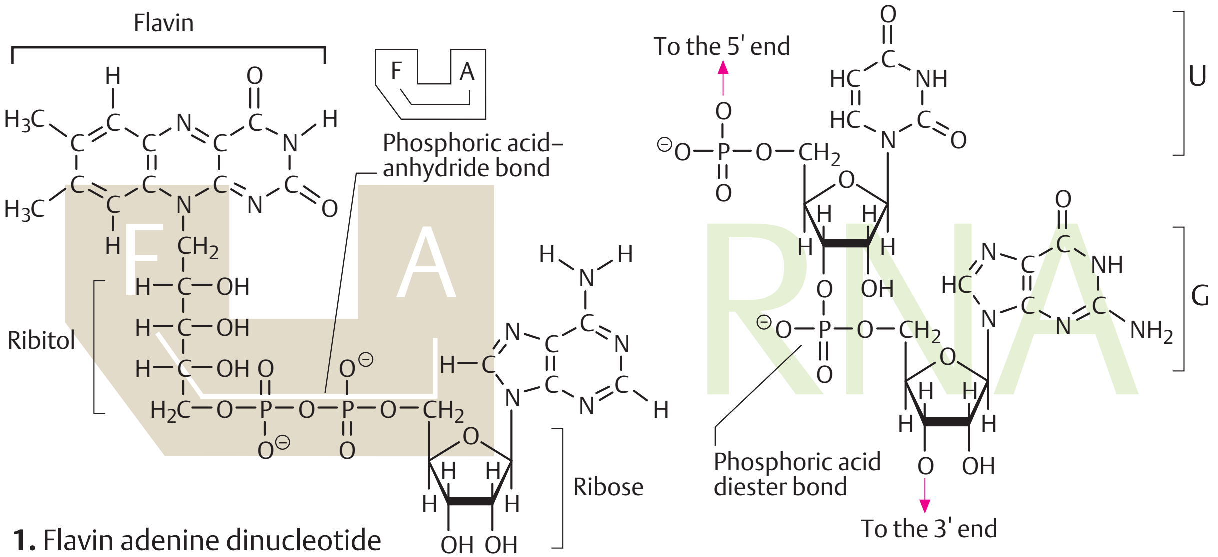
Flavin
F
A
Phosphoric acid–anhydride bond
Ribitol
Ribose
To the 5' end
U
G
RNA
Phosphoric acid diester bond
To the 3' end

1. Flavin adenine dinucleotide (FAD)

2. RNA (section)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

RNA

Ribonucleic acids (RNAs) are polymers consisting of nucleoside phosphate components that are linked by phosphoric acid diester bonds (see p. 80). The bases the contain are mainly uracil, cytosine, adenine, and guanine, but many unusual and modified bases are also found in RNAs (**B**).

A. Ribonucleic acids (RNAs) ◑

RNAs are involved in all the individual steps of gene expression and protein biosynthesis (see pp. 242–253). The properties of the most important forms of RNA are summarized in the table. The schematic diagram also gives an idea of the secondary structure of these molecules.

In contrast to DNA, RNAs do not form extended double helices. In RNAs, the base pairs (see p. 84) usually only extend over a few residues. For this reason, substructures often arise that have a finger shape or clover-leaf shape in two-dimensional representations. In these, the paired stem regions are linked by loops. Large RNAs such as ribosomal 16S-rRNA (center) contain numerous "stem and loop" regions of this type. These sections are again folded three-dimensionally—i. e., like proteins, RNAs have a tertiary structure (see p. 86). However, tertiary structures are only known of small RNAs, mainly tRNAs. The diagrams in Fig. **B** and on p. 86 show that the "clover-leaf" structure is not recognizable in a three-dimensional representation.

Cellular RNAs vary widely in their size, structure, and lifespan. The great majority of them are ribosomal RNA **(rRNA)**, which in several forms is a structural and functional component of *ribosomes* (see p. 250). Ribosomal RNA is produced from DNA by transcription in the nucleolus, and it is processed there and assembled with proteins to form ribosome subunits (see pp. 208, 242). The bacterial 16S-rRNA shown in Fig. **A**, with 1542 nucleotides (nt), is a component of the small ribosomae subunit, while the much smaller 5S-rRNA (118 nt) is located in the large subunit.

Messenger RNAs **(mRNAs)** transfer genetic information from the cell nucleus to the cytoplasm. The primary transcripts are substantially modified while still in the nucleus (mRNA maturation; see p. 246). Since mRNAs have to be read codon by codon in the ribosome, they must not form a stable tertiary structure. This is ensured in part by the attachment of *RNA-binding proteins,* which prevent base pairing. Due to the varying amounts of information that they carry, the lengths of mRNAs also vary widely. Their lifespan is usually short, as they are quickly broken down after translation.

Small nuclear RNAs **(snRNAs)** are involved in the splicing of mRNA precursors (see p. 246). They associate with numerous proteins to form "spliceosomes."

B. Transfer RNA (tRNA^{Phe}) ◑

The transfer RNAs **(tRNAs)** function during translation (see p. 250) as links between the nucleic acids and proteins. They are small RNA molecules consisting of 70–90 nucleotides, which "recognize" specific mRNA codons with their *anticodons* through base pairing. At the same time, at their 3′ end (sequence .. CCA-3′) they carry the amino acid that is assigned to the relevant mRNA codon according to the genetic code (see p. 248).

The base sequence and the tertiary structure of the yeast tRNA specific for phenylalanine (tRNA^{Phe}) is typical of all tRNAs. The molecule (see also p. 86) contains a high proportion of unusual and modified components (shaded in dark green in Fig. **1**). These include *pseudouridine* (Ψ), *dihydrouridine* (D), *thymidine* (T), which otherwise only occurs in DNA, and many methylated nucleotides such as *7-methylguanidine* (m^{7}G) and—in the anticodon—*2′-O-methylguanidine* (m^{2}G). Numerous base pairs, sometimes deviating from the usual pattern, stabilize the molecule's conformation (**2**).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Ribonucleic acids (RNAs)

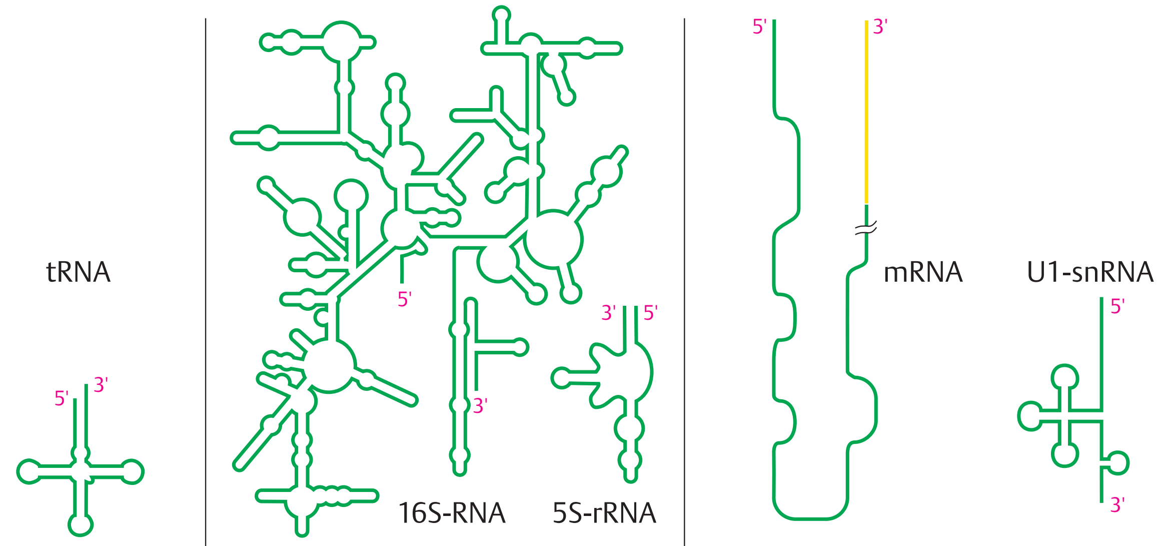
tRNA
5'
3'
5'
3'
3'
5'
3'
16S-RNA
5S-rRNA
5'
3'
mRNA
U1-snRNA
5'
3'

tRNA	rRNA	Type	mRNA	snRNA
>50	4	Species per cell	> 1000	~ 10
74 - 95	120 - 5000	Length (b)	400 - 6000	100 - 300
10-20%	80%	Proportion	5%	< 1%
Long	Long	Lifespan	Short	Long
Translation	Translation	Function	Translation	Splicing

B. Transfer RNA ($tRNA^{Phe}$)

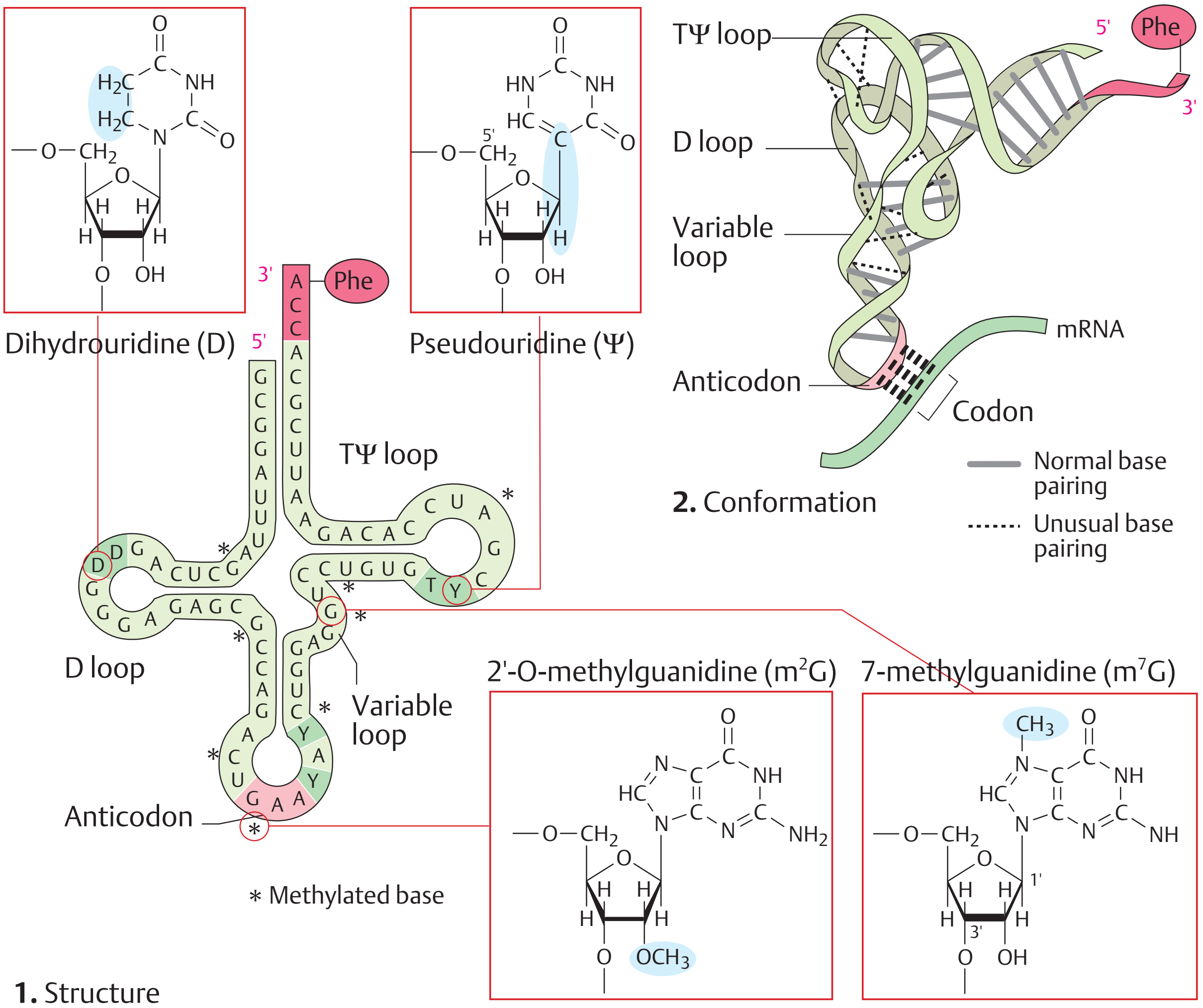
Dihydrouridine (D)
Pseudouridine (Ψ)
3'
5'
Phe
A
C
C
TΨ loop
D loop
Variable loop
Anticodon
D
Y
G
* Methylated base
2'-O-methylguanidine (m²G)
7-methylguanidine (m⁷G)
CH3
OCH3
TΨ loop
D loop
Variable loop
Anticodon
mRNA
Codon
5'
Phe
3'
Normal base pairing
Unusual base pairing

2. Conformation

1. Structure

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

DNA

A. DNA: structure ●

Like RNAs (see p. 82), deoxyribonucleic acids (DNAs) are polymeric molecules consisting of nucleotide building blocks. Instead of ribose, however, DNA contains 2′-deoxyribose, and the *uracil* base in RNA is replaced by *thymine.* The spatial structure of the two molecules also differs (see p. 86).

The first evidence of the special structure of DNA was the observation that the amounts of adenine and thymine are almost equal in every type of DNA. The same applies to guanine and cytosine. The model of DNA structure formulated in 1953 explains these *constant base ratios:* intact DNA consists of *two* polydeoxynucleotide molecules ("strands"). Each base in one strand is linked to a *complementary* base in the other strand by H-bonds. Adenine is complementary to thymine, and guanine is complementary to cytosine. One purine base and one pyrimidine base are thus involved in each **base pair**.

The complementarity of A with T and of G with C can be understood by considering the H bonds that are possible between the different bases. Potential donors (see p. 6) are amino groups (Ade, Cyt, Gua) and ring NH groups. Possible acceptors are carbonyl oxygen atoms (Thy, Cyt, Gua) and ring nitrogen atoms. *Two* linear and therefore highly stable bonds can thus be formed in A–T pairs, and *three* in G–C pairs.

Base pairings of this type are only possible, however, when the *polarity* of the two strands differs—i. e., when they run in opposite directions (see p. 80). In addition, the two strands have to be intertwined to form a **double helix**. Due to steric hindrance by the 2′-OH groups of the ribose residues, RNA is unable to form a double helix. The structure of RNA is therefore less regular than that of DNA (see p. 82).

The conformation of DNA that predominates within the cell (known as **B-DNA**) is shown schematically in Fig. **A2** and as a van der Waals model in Fig. **B1**. In the schematic diagram (**A2**), the deoxyribose–phosphate "backbone" is shown as a ribbon. The bases (indicated by lines) are located on the inside of the **double helix**. This area of DNA is therefore apolar. By contrast, the molecule's surface is polar and negatively charged, due to the sugar and phosphate residues in the backbone. Along the whole length of the DNA molecule, there are two depressions—referred to as the "minor groove" and the "major groove"—that lie between the strands.

B. Coding of genetic information ●

In all living cells, DNA serves to **store genetic information**. Specific segments of DNA ("genes") are transcribed as needed into RNAs, which either carry out structural or catalytic tasks themselves or provide the basis for synthesizing proteins (see p. 82). In the latter case, the DNA codes for the primary structure of proteins. The "language" used in this process has four letters (A, G, C, and T). All of the words ("codons") contain three letters ("triplets"), and each triplet stands for one of the 20 proteinogenic amino acids.

The two strands of DNA are not functionally equivalent. The **template strand** (the (–) strand or "codogenic strand," shown in light gray in Fig. **1**) is the one that is read during the synthesis of RNA (transcription; see p. 242). Its sequence is complementary to the RNA formed. The **sense strand** (the (+) strand or "coding strand," shown in color in Figs. **1** and **2**) has the *same sequence as the RNA,* except that T is exchanged for U. By convention, it is agreed that gene sequences are expressed by reading the sequence of the sense strand in the 5′→3′ direction. Using the genetic code (see p. 248), in this case the protein sequence (**3**) is obtained directly in the reading direction usual for proteins—i. e., from the *N* terminus to the *C* terminus.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. DNA: structure

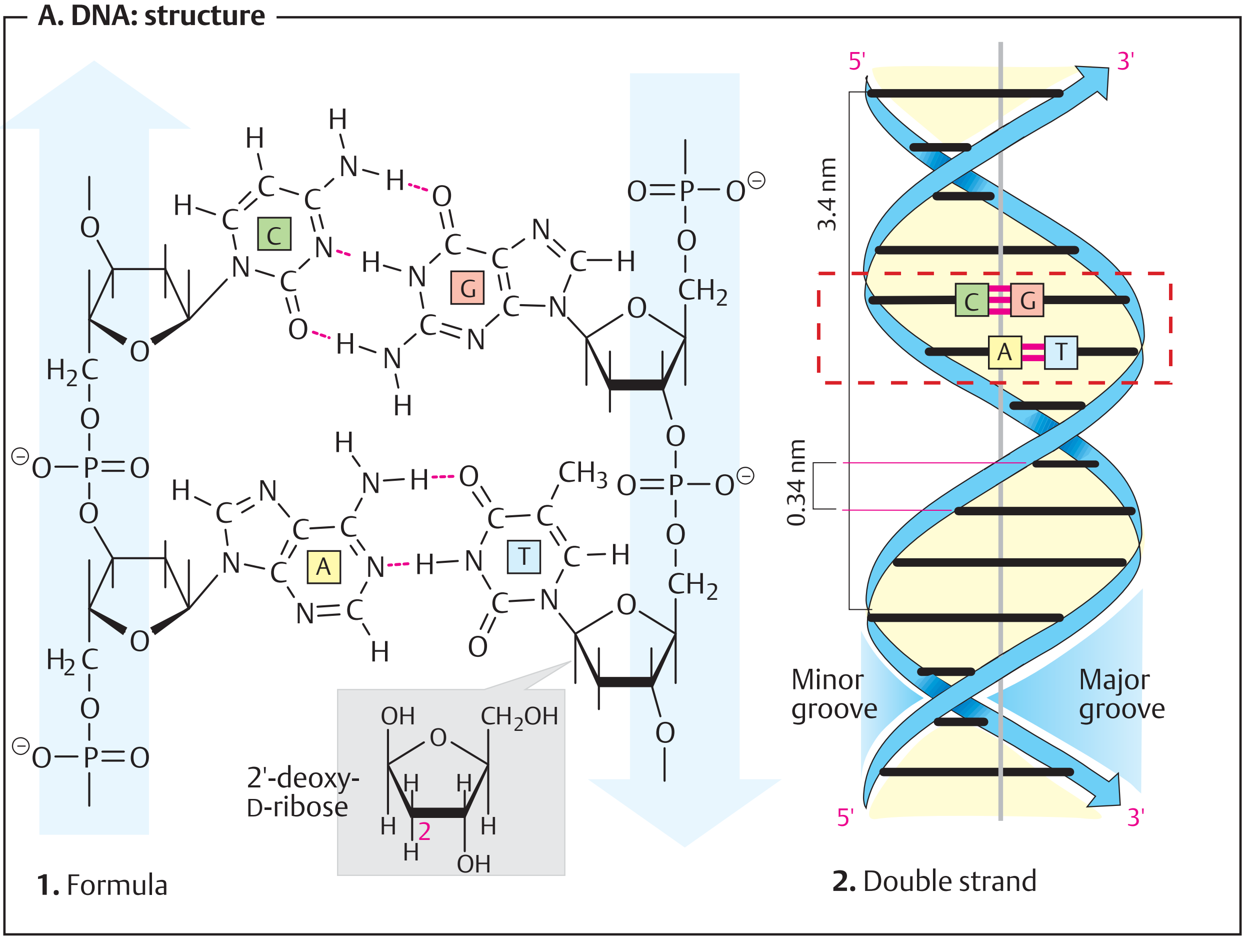
C
G
A
T
$O{=}P{-}O^{\ominus}$
CH_2
H_2C
CH_3
2'-deoxy-D-ribose
OH
CH_2OH
5'
3'
3.4 nm
0.34 nm
Minor groove
Major groove
1. Formula
2. Double strand

B. Coding of genetic information

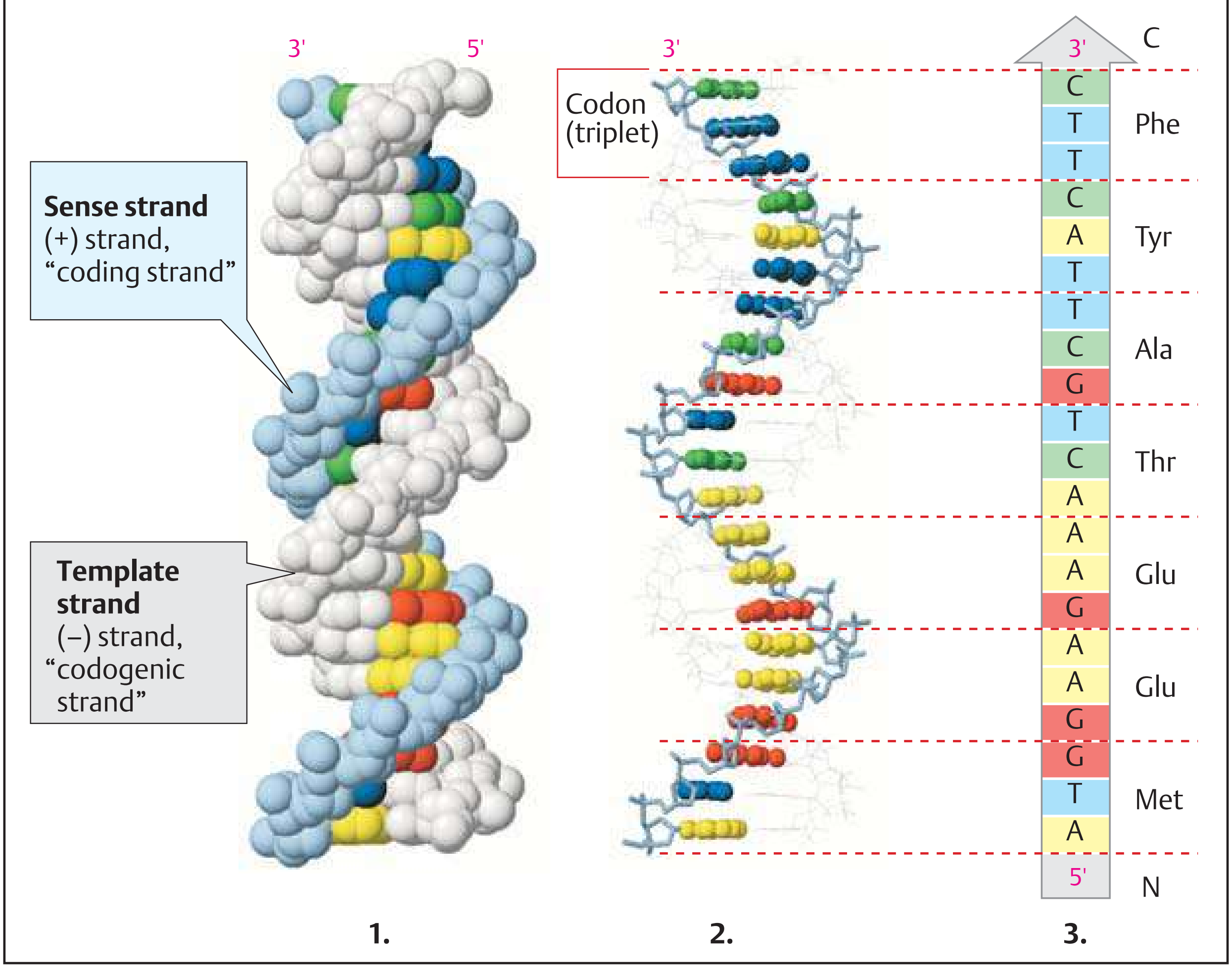
3'
5'
Sense strand
(+) strand,
“coding strand”
Template strand
(–) strand,
“codogenic strand”
Codon
(triplet)
C
C T T Phe
C A T Tyr
T C G Ala
T C A Thr
A A G Glu
A A G Glu
G T A Met
N
1.
2.
3.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Molecular models: DNA and RNA

The illustration opposite shows selected nucleic acid molecules. Fig. **A** shows various conformations of DNA, and Fig. **B** shows the spatial structures of two small RNA molecules. In both, the van der Waals models (see p. 6) are accompanied by ribbon diagrams that make the course of the chains clear. In all of the models, the polynucleotide "backbone" of the molecule is shown in a darker color, while the bases are lighter.

A. DNA: conformation ○

Investigations of synthetic DNA molecules have shown that DNA can adopt several different conformations. All of the DNA segments shown consist of 21 base pairs (bp) and have the same sequence.

By far the most common form is **B-DNA** (**2**). As discussed on p. 84, this consists of two antiparallel polydeoxynucleotide strands intertwined with one another to form a **right-handed double helix**. The "backbone" of these strands is formed by deoxyribose and phosphate residues linked by phosphoric acid diester bonds.

In the B conformation, the aromatic rings of the nucleobases are stacked at a distance of 0.34 nm almost at right angles to the axis of the helix. Each base is rotated relative to the preceding one by an angle of 35°. A complete turn of the double helix (360°) therefore contains around 10 base pairs (abbreviation: bp), i. e., the *pitch* of the helix is 3.4 nm. Between the backbones of the two individual strands there are two grooves with different widths. The *major groove* is visible at the top and bottom, while the narrower *minor groove* is seen in the middle. DNA-binding proteins and transcription factors (see pp. 118, 244) usually enter into interactions in the area of the major groove, with its more easily accessible bases.

In certain conditions, DNA can adopt the **A conformation** (**1**). In this arrangement, the double helix is still right-handed, but the bases are no longer arranged at right angles to the axis of the helix, as in the B form. As can be seen, the A conformation is more compact than the other two conformations. The minor groove almost completely disappears, and the major groove is narrower than in the B form. A-DNA arises when B-DNA is dehydrated. It probably does not occur in the cell.

In the **Z-conformation** (**3**), which can occur within GC-rich regions of B-DNA, the organization of the nucleotides is completely different. In this case, the helix is *left-handed,* and the backbone adopts a characteristic *zig-zag* conformation (hence "Z-DNA"). The Z double helix has a smaller pitch than B-DNA. DNA segments in the Z conformation probably have physiological significance, but details are not yet known.

B. RNA ○

RNA molecules are unable to form extended double helices, and are therefore less highly ordered than DNA molecules. Nevertheless, they have defined secondary and tertiary structures, and a large proportion of the nucleotide components enter into base pairings with other nucleotides. The examples shown here are **5S-rRNA** (see p. 242), which occurs as a structural component in ribosomes, and a **tRNA** molecule from yeast (see p. 82) that is specific for phenylalanine.

Both molecules are folded in such a way that the 3′ end and the 5′ end are close together. As in DNA, most of the bases are located in the inside of the structures, while the much more polar "backbone" is turned outwards. An exception to this is seen in the three bases of the *anticodon* of the tRNA (pink), which have to interact with mRNA and therefore lie on the surface of the molecule. The bases of the conserved CCA triplet at the 3′ end (red) also jut outward. During amino acid activation (see p. 248), they are recognized and bound by the ligases.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

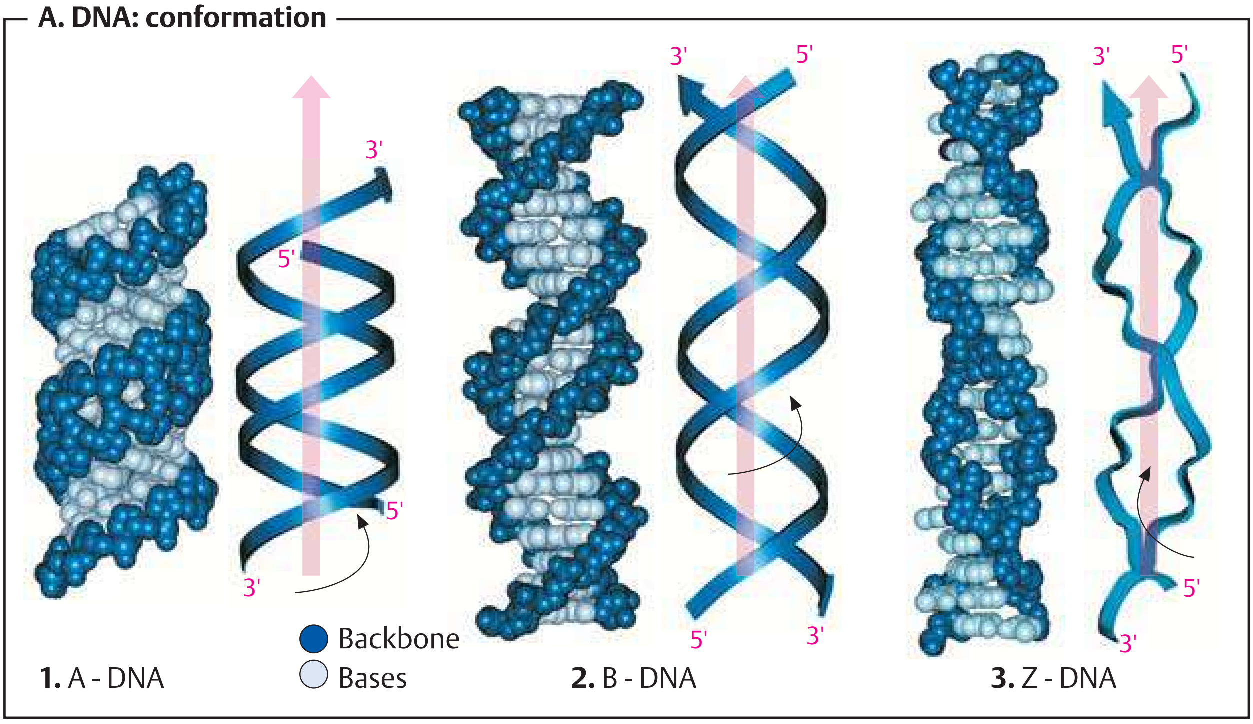

A. DNA: conformation

1. A - DNA

2. B - DNA

3. Z - DNA

3'
5'
5'
3'
● Backbone
○ Bases
3'
5'
5'
3'
3'
5'
5'
3'

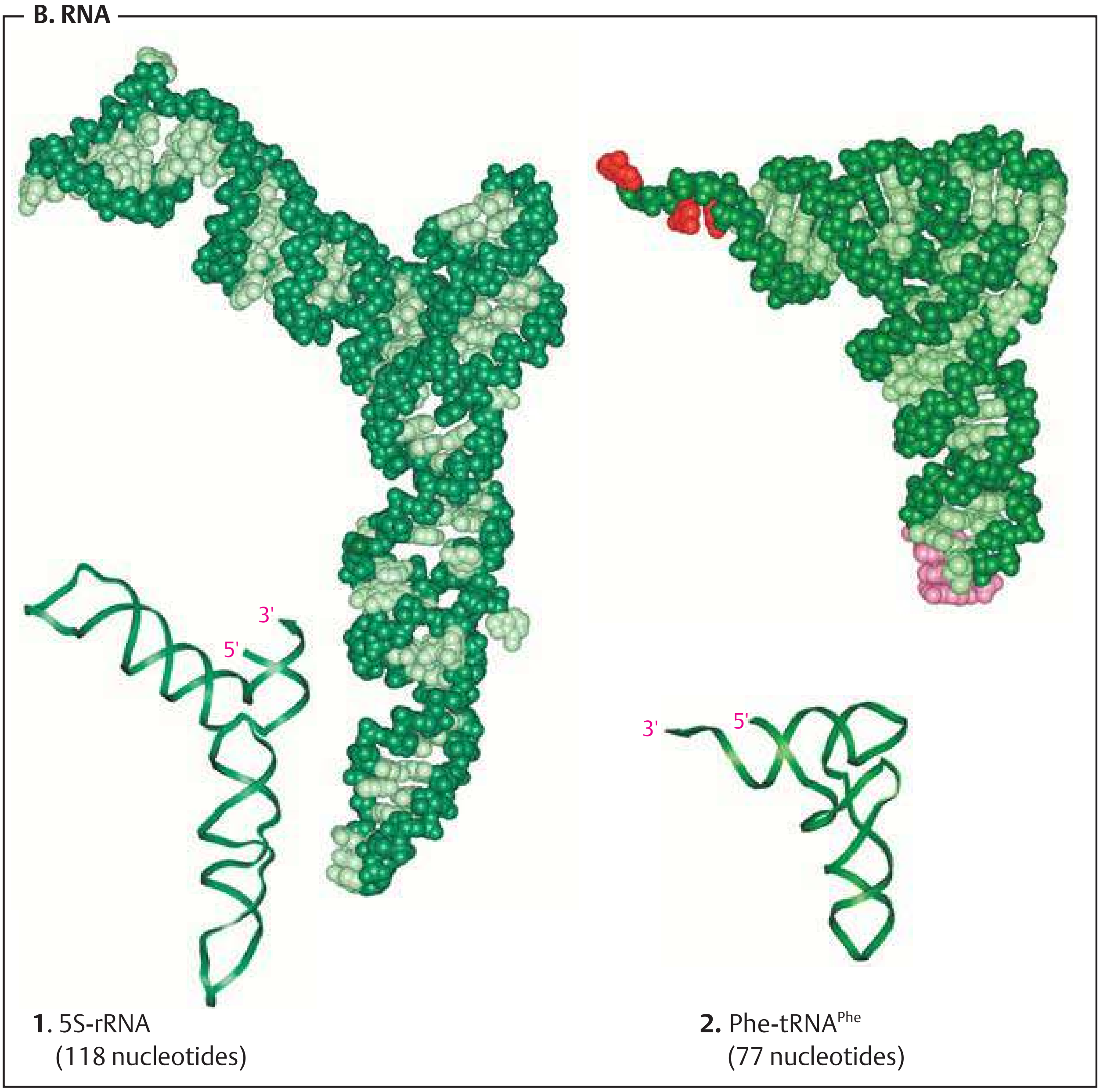

B. RNA

1. 5S-rRNA
(118 nucleotides)

2. Phe-tRNA^{Phe}
(77 nucleotides)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Enzymes: basics

Enzymes are **biological catalysts**—i. e., substances of biological origin that accelerate chemical reactions (see p. 24). The orderly course of metabolic processes is only possible because each cell is equipped with its own genetically determined set of enzymes. It is only this that allows coordinated sequences of reactions (**metabolic pathways**; see p. 112). Enzymes are also involved in many regulatory mechanisms that allow the metabolism to adapt to changing conditions (see p. 114). Almost all enzymes are **proteins**. However, there are also catalytically active ribonucleic acids, the *"ribozymes"* (see pp. 246, 252).

A. Enzymatic activity ●

The catalytic action of an enzyme, its **activity**, is measured by determining the **increase in the reaction rate** under precisely defined conditions—i. e., the difference between the turnover (violet) of the catalyzed reaction (orange) and uncatalyzed reaction (yellow) in a specific time interval. Normally, reaction rates are expressed as the *change in concentration per unit of time* (mol l^{-1} s^{-1}; see p. 22). Since the catalytic activity of an enzyme is independent of the volume, the unit used for enzymes is usually *turnover per unit time*, expressed in **katal** (kat, mol s^{-1}). However, the **international unit U** is still more commonly used (µmol turnover min^{-1}; 1 U = 16.7 nkat).

B. Reaction and substrate specificity ●

The action of enzymes is usually very *specific*. This applies not only to the type of reaction being catalyzed (**reaction specificity**), but also to the nature of the reactants ("substrates") that are involved (**substrate specificity**; see p. 94). In Fig. **B**, this is illustrated schematically using a bond-breaking enzyme as an example. Highly specific enzymes (type A, top) catalyze the cleavage of only *one* type of bond, and only when the structure of the substrate is the correct one. Other enzymes (type B, middle) have narrow reaction specificity, but broad substrate specificity. Type C enzymes (with low reaction specificity *and* low substrate specificity, bottom) are very rare.

C. Enzyme classes ◑

More than 2000 different enzymes are currently known. A system of *classification* has been developed that takes into account both their *reaction specificity* and their *substrate specificity*. Each enzyme is entered in the *Enzyme Catalogue* with a four-digit Enzyme Commission number (**EC number**). The first digit indicates membership of one of the six **major classes**. The next two indicate subclasses and subsubclasses. The last digit indicates where the enzyme belongs in the subsubclass. For example, lactate dehydrogenase (see pp. 98–101) has the EC number *1.1.1.27* (class 1, oxidoreductases; subclass 1.1, CH–OH group as electron *donor*; sub-subclass 1.1.1, $NAD(P)^+$ as electron *acceptor*).

Enzymes with similar reaction specificities are grouped into each of the six major classes:

The **oxidoreductases** (*class 1*) catalyze the transfer of reducing equivalents from one redox system to another.

The **transferases** (*class 2*) catalyze the transfer of other groups from one molecule to another. Oxidoreductases and transferases generally require coenzymes (see pp. 104 ff.).

The **hydrolases** (*class* 3) are also involved in group transfer, but the acceptor is always a *water molecule.*

Lyases (*class 4*, often also referred to as "synthases") catalyze reactions involving either the cleavage or formation of chemical bonds, with double bonds either arising or disappearing.

The **isomerases** (*class 5*) move groups within a molecule, without changing the gross composition of the substrate.

The ligation reactions catalyzed by **ligases** ("synthetases," *class* 6) are energy-dependent and are therefore always coupled to the hydrolysis of nucleoside triphosphates.

In addition to the enzyme name, we also usually give its EC number. The *annotated enzyme list* (pp. 420 ff.) includes all of the enzymes mentioned in this book, classified according to the Enzyme Catalog system.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Enzymatic activity

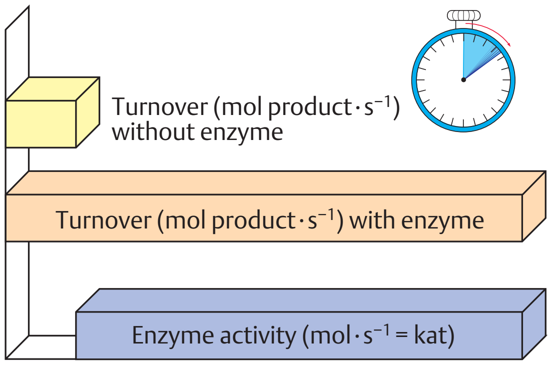
Turnover (mol product · s^{-1}) without enzyme
Turnover (mol product · s^{-1}) with enzyme
Enzyme activity (mol · s^{-1} = kat)

1 Katal (kat): Amount of enzyme which increases turnover by 1 mol · s^{-1}

B. Reaction and substrate specificity

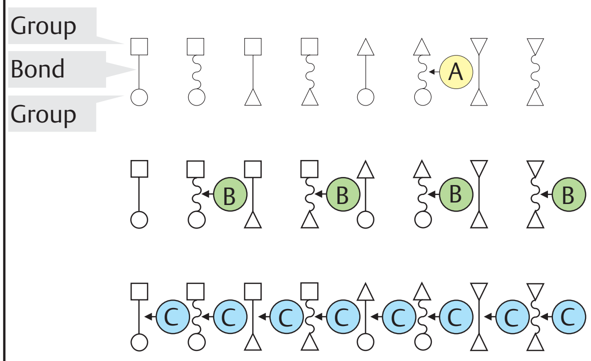
Group
Bond
Group
A
B
B
B
B
C
C
C
C
C
C
C
C

	Reaction specificity	Substrate specificity
A	High	High
B	High	Low
C	Low	Low

C. The enzyme classes

Class	Reaction type	Important subclasses
1 Oxidoreductases	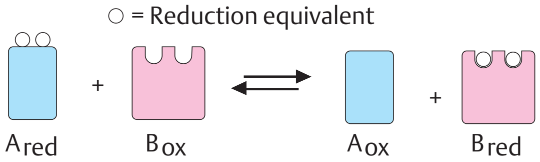 ○ = Reduction equivalent A_{red} + B_{ox} A_{ox} + B_{red}	Dehydrogenases Oxidases, peroxidases Reductases Monooxygenases Dioxygenases
2 Transferases	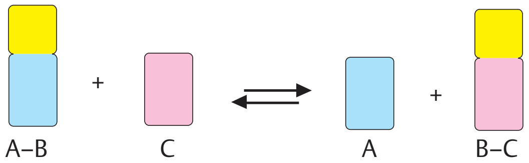	C_1-Transferases Glycosyltransferases Aminotransferases Phosphotransferases
3 Hydrolases	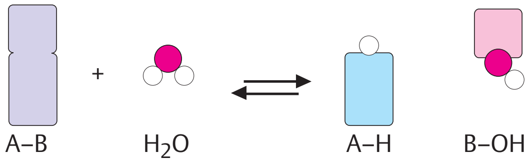 A–B + H_2O A–H B–OH	Esterases Glycosidases Peptidases Amidases
4 Lyases ("synthases")	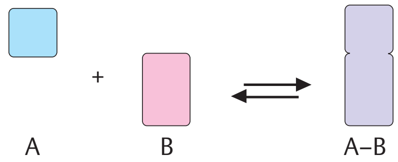	C-C-Lyases C-O-Lyases C-N-Lyases C-S-Lyases
5 Isomerases	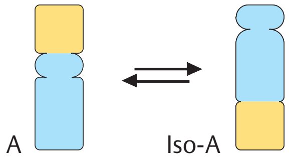	Epimerases *cis trans* Isomerases Intramolecular transferases
6 Ligases ("synthetases")	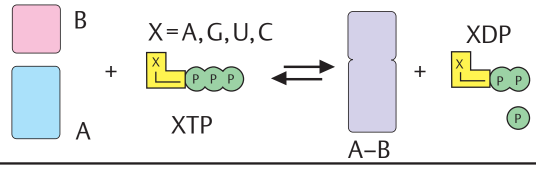 B A + X = A, G, U, C X P P P XTP A–B + XDP X P P P	C-C-Ligases C-O-Ligases C-N-Ligases C-S-Ligases

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Enzyme catalysis

Enzymes are extremely effective **catalysts**. They can increase the rate of a catalyzed reaction by a factor of 10^{12} or more. To grasp the mechanisms involved in enzyme catalysis, we can start by looking at the course of an uncatalyzed reaction more closely.

A. Uncatalyzed reaction ○

The reaction A + B → C + D is used as an example. In solution, **reactants A and B** are surrounded by a shell of water molecules (the *hydration shell*), and they move in random directions due to thermal agitation. They can only react with each other if they collide in a favorable orientation. This is not very probable, and therefore only occurs rarely. Before conversion into the products C + D, the **collision complex A-B** has to pass through a **transition state**, the formation of which usually requires a large amount of **activation energy, E_a** (see p. 22). Since only a few A–B complexes can produce this amount of energy, a productive transition state arises even less often than a collision complex. In solution, a large proportion of the activation energy is required for the *removal of the hydration shells* between A and B. However, charge displacements and other *chemical processes* within the reactants also play a role. As a result of these limitations, conversion only happens occasionally in the absence of a catalyst, and the reaction rate v is low, even when the reaction is thermodynamically possible—i. e., when $\Delta G < 0$ (see p. 18).

B. Enzyme-catalyzed reaction ◑

Shown here is a *sequential mechanism* in which substrates A and B are bound and products C and D are released, in that order. Another possible reaction sequence, known as the *"ping-pong mechanism,"* is discussed on p. 94.

Enzymes are able to bind the reactants (their *substrates*) specifically at the **active center**. In the process, the substrates are oriented in relation to each other in such a way that they take on the *optimal orientation* for the formation of the transition state (**1**–**3**). The **proximity and orientation of the substrates** therefore strongly increase the likelihood that *productive* A–B complexes will arise. In addition, binding of the substrates results in removal of their hydration shells. As a result of the **exclusion of water**, very different conditions apply in the active center of the enzyme during catalysis than in solution (**3**–**5**). A third important factor is the **stabilization of the transition state** as a result of interactions between the amino acid residues of the protein and the substrate (**4**). This further reduces the activation energy needed to create the transition state. Many enzymes also take up groups from the substrates or transfer them to the substrates during catalysis.

Proton transfers are particularly common. This **acid–base catalysis** by enzymes is much more effective than the exchange of protons between acids and bases in solution. In many cases, chemical groups are temporarily bound covalently to the amino acid residues of the enzyme or to coenzymes during the catalytic cycle. This effect is referred to as **covalent catalysis** (see the transaminases, for example; p. 178). The principles of enzyme catalysis sketched out here are discussed in greater detail on p. 100 using the example of lactate dehydrogenase.

C. Principles of enzyme catalysis ◑

Although it is difficult to provide quantitative estimates of the contributions made by individual catalytic effects, it is now thought that the enzyme's **stabilization of the transition state** is the most important factor. It is not tight binding of the *substrate* that is important, therefore—this would increase the activation energy required by the reaction, rather than reducing it—but rather the binding of the transition state. This conclusion is supported by the very high affinity of many enzymes for analogues of the transition state (see p. 96). A simple mechanical analogy may help clarify this (right). To transfer the metal balls (the reactants) from location EA (the substrate state) via the higher-energy transition state to EP (the product state), the magnet (the catalyst) has to be orientated in such a way that its attractive force acts on the transition state (bottom) rather than on EA (top).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Uncatalyzed reaction

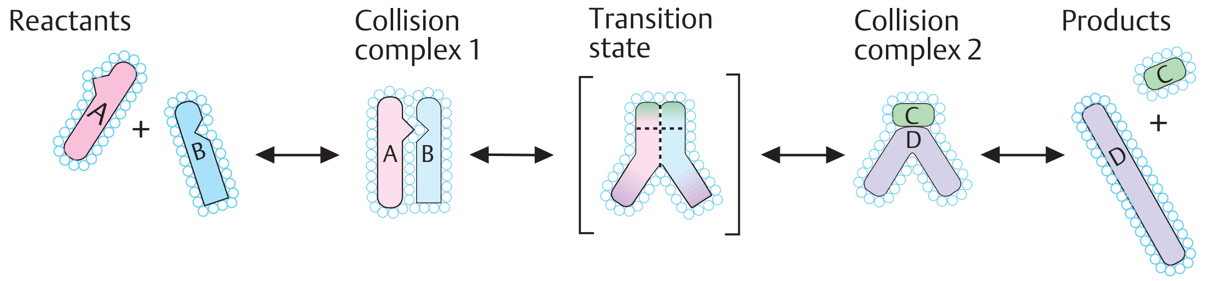
Reactants
A
+
B
Collision complex 1
A
B
Transition state
Collision complex 2
C
D
Products
C
+
D

B. Enzyme-catalyzed reaction

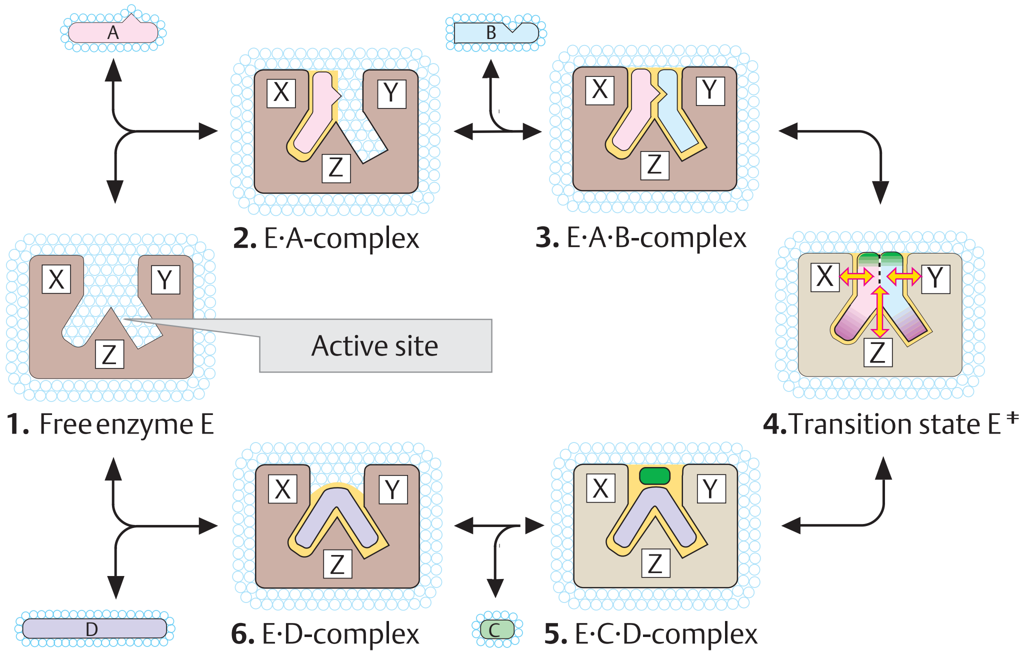
A
B
X
Y
Z
2. E·A-complex
3. E·A·B-complex
Active site
1. Free enzyme E
4. Transition state E‡
6. E·D-complex
5. E·C·D-complex
D
C

C. Principles of enzyme catalysis

a Approximation and orientation of the substrates

b Exclusion of water

c Stabilization of the transition state

d Group transfer

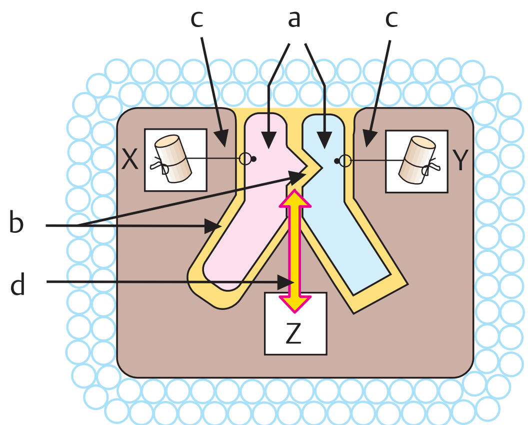

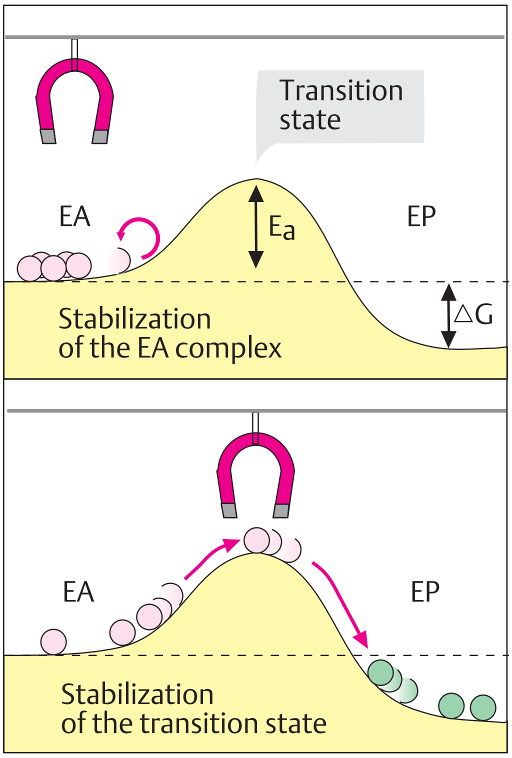
Transition state
EA
Ea
EP
ΔG
Stabilization of the EA complex
EA
EP
Stabilization of the transition state

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Enzyme kinetics I

The **kinetics** of enzyme-catalyzed reactions (i. e., the dependence of the reaction rate on the reaction conditions) is mainly determined by the *properties of the catalyst.* It is therefore more complex than the kinetics of an uncatalyzed reaction (see p. 22). Here we discuss these issues using the example of a simple first-order reaction (see p. 22)

A. Michaelis–Menten kinetics ◑

In the absence of an enzyme, the reaction rate v is proportional to the concentration of substance A (top). The constant k is the *rate constant* of the uncatalyzed reaction. Like all catalysts, the enzyme E (total concentration $[E]_t$) creates a new reaction pathway. Initially, A is bound to E (partial reaction 1, left). If this reaction is in chemical equilibrium, then with the help of the law of mass action—and taking into account the fact that $[E]_t = [E] + [EA]$—one can express the concentration [EA] of the *enzyme–substrate* complex as a function of [A] (left). The **Michaelis constant K_m** thus describes the state of equilibrium of the reaction. In addition, we know that $k_{cat} > k$—in other words, enzyme-bound substrate reacts to B much faster than A alone (partial reaction 2, right). k_{cat}, the enzyme's **turnover number**, corresponds to the number of substrate molecules converted by one enzyme molecule per second. Like the conversion A → B, the formation of B from EA is a first-order reaction—i. e., $v = k\ [EA]$ applies. When this equation is combined with the expression already derived for EA, the result is the **Michaelis–Menten equation.**

In addition to the *variables* v and [A], the equation also contains two *parameters* that do not depend on the substrate concentration [A], but describe properties of the enzyme itself: the product $k_{cat}\ [E]_g$ is the limiting value for the reaction rate at a very high [A], the **maximum velocity V_{max}** of the reaction (recommended abbreviation: V). The **Michaelis constant** K_m characterizes the *affinity* of the enzyme for a substrate. It corresponds to the substrate concentration at which v reaches half of V_{max} (if $v = V_{max}/2$, then $[A]/(K_m + [A]) = 1/2$, i. e. [A] is then = K_m). A *high* affinity of the enzyme for a substrate therefore leads to a *low* K_m value, and vice versa. Of the two enzymes whose *substrate saturation curves* are shown in diagram 1, enzyme 2 has the higher affinity for A [$K_m = 1\ \text{mmol}\ l^{-1}$]; V_{max}, by contrast, is much lower than with enzyme 1.

Since v approaches V *asymptotically* with increasing values of [A], it is difficult to obtain reliable values for V_{max}—and thus for K_m as well—from diagrams plotting v against [A]. To get around this, the Michaelis–Menten equation can be arranged in such a way that the measured points lie on a *straight line.* In the **Lineweaver–Burk plot (2)**, 1/v is plotted against 1/[A]. The intersections of the line of best fit with the axes then produce $1/V_{max}$ and $-1/K_m$. This type of diagram is very clear, but for practical purposes it is less suitable for determining V_{max} and K_m. Calculation methods using personal computers are faster and more objective.

B. Isosteric and allosteric enzymes ◑

Many enzymes can occur in various *conformations* (see p. 72), which have different catalytic properties and whose proportion of the total number of enzyme molecules is influenced by substrates and other ligands (see pp. 116 and 280, for example). **Allosteric enzymes** of this type, which are usually present in oligomeric form, can be recognized by their S-shaped (*sigmoidal*) saturation curves, which cannot be described using the Michaelis model. In the case of isosteric enzymes (with only *one* enzyme conformation, **1**), the efficiency of substrate binding (dashed curve) declines constantly with increasing [A], because the number of free binding sites is constantly decreasing. In most allosteric enzymes (**2**), the binding efficiency initially rises with increasing [A], because the free enzyme is present in a low-affinity conformation (square symbols), which is gradually converted into a higher-affinity form (round symbols) as a result of binding with A. It is only at high [A] values that a lack of free binding sites becomes noticeable and the binding strength decreases again. In other words, the affinity of allosteric enzymes is not constant, but depends on the type and concentration of the ligand.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Michaelis Menten kinetics

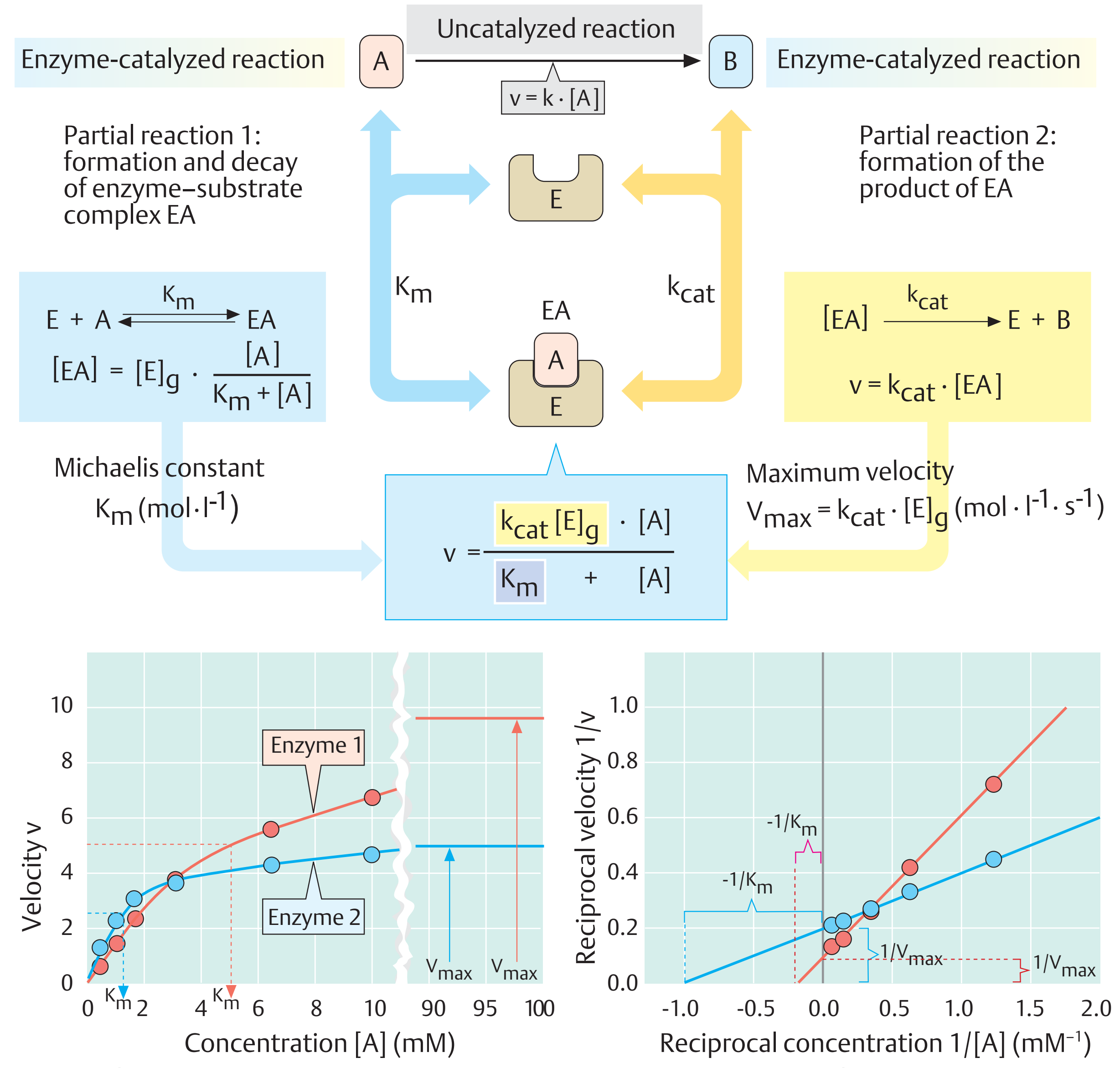
Enzyme-catalyzed reaction
A
Uncatalyzed reaction
v = k · [A]
B
Enzyme-catalyzed reaction
Partial reaction 1: formation and decay of enzyme–substrate complex EA
Partial reaction 2: formation of the product of EA
E
Km
kcat
EA
A
E
E + A ⇄ EA (Km)
[EA] = [E]g · [A] / (Km + [A])
[EA] → E + B (kcat)
v = kcat · [EA]
Michaelis constant Km (mol·l-1)
Maximum velocity Vmax = kcat · [E]g (mol · l-1 · s-1)
v = kcat [E]g · [A] / (Km + [A])
Velocity v
0
2
4
6
8
10
Enzyme 1
Enzyme 2
Km
Km
Vmax
Vmax
0 2 4 6 8 10 90 95 100
Concentration [A] (mM)
Reciprocal velocity 1/v
0.0
0.2
0.4
0.6
0.8
1.0
-1/Km
-1/Km
1/Vmax
1/Vmax
-1.0 -0.5 0.0 0.5 1.0 1.5 2.0
Reciprocal concentration 1/[A] (mM-1)

1. Hyperbolic plot **2.** Lineweaver–Burk plot

B. Isosteric and allosteric enzymes

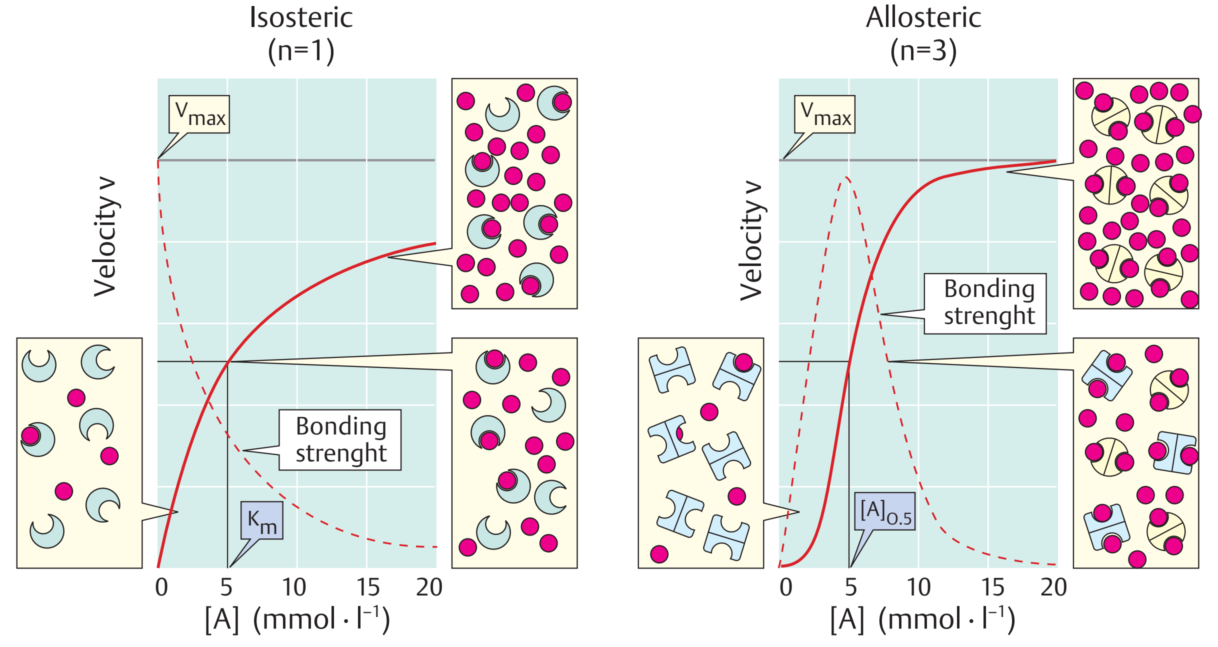
Isosteric (n=1)
Allosteric (n=3)
Vmax
Vmax
Velocity v
Velocity v
Bonding strenght
Bonding strenght
Km
[A]0.5
0 5 10 15 20
[A] (mmol · l-1)
0 5 10 15 20
[A] (mmol · l-1)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Enzyme kinetics II

The catalytic properties of enzymes, and consequently their activity (see p. 90), are influenced by numerous factors, which all have to be optimized and controlled if activity measurements are to be carried out in a useful and reproducible fashion. These factors include physical quantities (temperature, pressure), the chemical properties of the solution (pH value, ionic strength), and the concentrations of the relevant substrates, cofactors, and inhibitors.

A. pH and temperature dependency of enzyme activity ◑

The effect of enzymes is strongly dependent on the pH value (see p. 30). When the activity is plotted against pH, a *bell-shaped curve* is usually obtained (**1**). With animal enzymes, the **pH optimum**—i. e., the pH value at which enzyme activity is at its maximum—is often close to the pH value of the cells (i. e., pH 7). However, there are also exceptions to this. For example, the proteinase *pepsin* (see p. 270), which is active in the acidic gastric lumen, has a pH optimum of 2, while other enzymes (at least in the test tube) are at their most active at pH values higher than 9. The bell shape of the activity–pH profile results from the fact that amino acid residues with ionizable groups in the side chain are essential for catalysis. In example (**1**), these are a basic group B ($pK_a = 8$), which has to be protonated in order to become active, and a second acidic amino acid AH ($pK_a = 6$), which is only active in a dissociated state. At the optimum pH of 7, around 90% of both groups are present in the active form; at higher and lower values, one or the other of the groups increasingly passes into the inactive state.

The **temperature dependency** of enzymatic activity is usually asymmetric. With increasing temperature, the increased thermal movement of the molecules initially leads to a rate acceleration (see p. 22). At a certain temperature, the enzyme then becomes unstable, and its activity is lost within a narrow temperature difference as a result of denaturation (see p. 74). The optimal temperatures of the enzymes in higher organisms rarely exceed 50 °C, while enzymes from thermophilic bacteria found in hot springs, for instance, may still be active at 100 °C.

B. Substrate specificity ◑

Enzymes "recognize" their substrates in a highly specific way (see p. 88). It is only the marked **substrate specificity** of the enzymes that makes a regulated metabolism possible. This principle can be illustrated using the example of the two closely related proteinases *trypsin* and *chymotrypsin*. Both belong to the group of serine proteinases and contain the same "triad" of catalytically active residues (Asp–His–Ser, shown here in green; see p. 176). Trypsin selectively cleaves peptide bonds on the C-terminal side of basic amino acids (lysine and arginine), while chymotrypsin is specific for hydrophobic residues. The substrate binding "pockets" of both enzymes have a similar structure, but their amino acid sequences differ slightly. In trypsin, a negatively charged aspartate residue (Asp-189, red) is arranged in such a way that it can bind and fix the basic group in the side chain of the substrate. In chymotrypsin, the "binding pocket" is slightly narrower, and it is lined with neutral and hydrophobic residues that stabilize the side chains of apolar substrate amino acids through hydrophobic interactions (see p. 28).

C. Bisubstrate kinetics ○

Almost all enzymes—in contrast to the simplified description given on p. 92—have more than one substrate or product. On the other hand, it is rare for more than two substrates to be bound *simultaneously*. In bisubstrate reactions of the type $A + B \rightarrow C + D$, a number of reaction sequences are possible. In addition to the *sequential mechanisms* (see p. 90), in which all substrates are bound in a specific sequence before the product is released, there are also mechanisms in which the first substrate A is bound and immediately cleaved. A part of this substrate remains bound to the enzyme, and is then transferred to the second substrate B after the first product C has been released. This is known as the **ping-pong mechanism**, and it is used by *transaminases*, for example (see p. 178). In the Lineweaver—Burk plot (right; see p. 92), it can be recognized in the parallel shifting of the lines when [B] is varied.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. pH and temperature dependency of enzyme activity

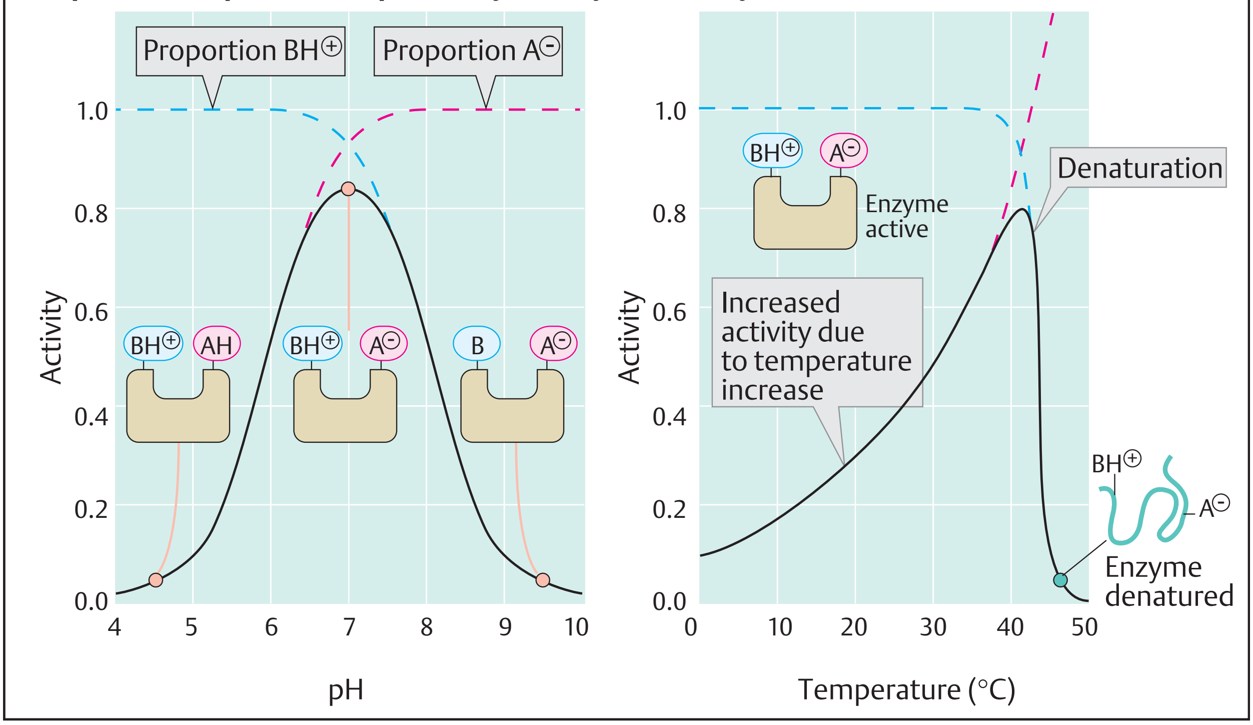

B. Substrate specificity

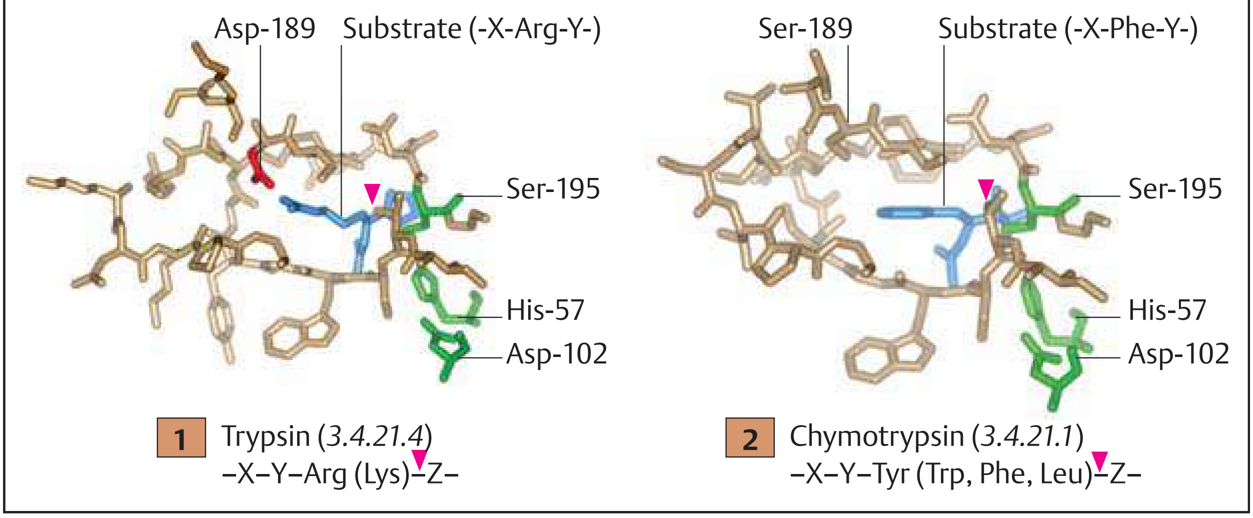
Asp-189
Substrate (-X-Arg-Y-)
Ser-195
His-57
Asp-102
1
Trypsin (3.4.21.4)
–X–Y–Arg (Lys)–Z–
Ser-189
Substrate (-X-Phe-Y-)
Ser-195
His-57
Asp-102
2
Chymotrypsin (3.4.21.1)
–X–Y–Tyr (Trp, Phe, Leu)–Z–

C. Bisubstrate kinetics

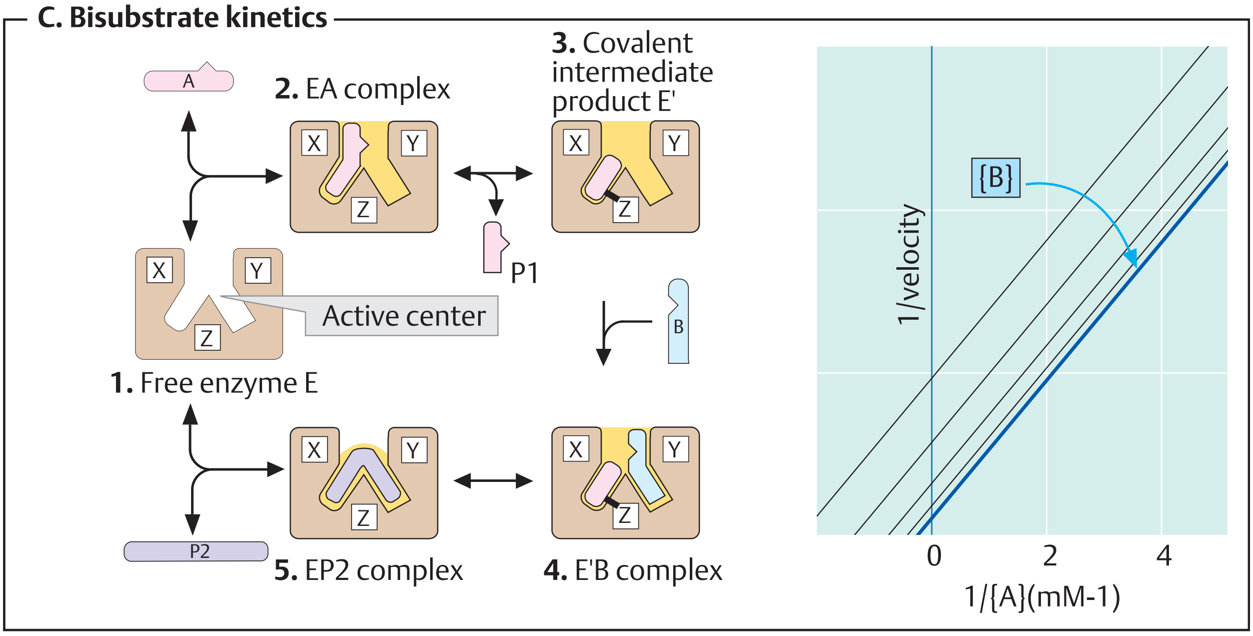
A
2. EA complex
3. Covalent intermediate product E'
X
Y
Z
P1
Active center
1. Free enzyme E
B
5. EP2 complex
4. E'B complex
P2
[B]
1/velocity
0
2
4
1/[A](mM-1)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Inhibitors

Many substances can affect metabolic processes by influencing the activity of enzymes. **Enzyme inhibitors** are particularly important here. A large proportion of **medicines** act as enzyme inhibitors. Enzyme-kinetic experiments are therefore an important aspect of drug development and testing procedures. Natural *metabolites* are also involved in regulatory processes as inhibitors (see p. 114).

A. Types of inhibitor ◑

Most enzyme inhibitors act **reversibly**—i. e., they do not cause any permanent changes in the enzyme. However, there are also **irreversible** inhibitors that permanently modify the target enzyme. The mechanism of action of an inhibitor—its **inhibition type**—can be determined by comparing the kinetics (see p. 92) of the inhibited and uninhibited reactions (**B**). This makes it possible to distinguish *competitive inhibitors* (left) from *noncompetitive inhibitors* (right), for example. *Allosteric inhibition* is particularly important for metabolic regulation (see below).

Substrate analogs (**2**) have properties similar to those of one of the substrates of the target enzyme. They are bound by the enzyme, but cannot be converted further and therefore *reversibly* block some of the enzyme molecules present. A *higher* substrate concentration is therefore needed to achieve a half-maximum rate; the Michaelis constant K_m increases (**B**). High concentrations of the substrate displace the inhibitor again. The maximum rate V_{max} is therefore not influenced by this type of inhibition. Because the substrate and the inhibitor compete with one another for the *same* binding site on the enzyme, this type of inhibition is referred to as **competitive**. **Analogs of the transition state** (**3**) usually also act competitively.

When an inhibitor interacts with a group that is important for enzyme activity, but does not affect binding of the substrate, the inhibition is **non-competitive** (right). In this case, K_m remains unchanged, but the concentration of functional enzyme $[E]_t$, and thus V_{max}, decrease. Non-competitive inhibitors generally act irreversibly, by modifying functional groups of the target enzyme (**4**).

"Suicide substrates" (**5**) are substrate analogs that also contain a reactive group. Initially, they bind reversibly, and then they form a covalent bond with the active center of the enzyme. Their effect is therefore also non-competitive. A well-known example of this is the antibiotic *penicillin* (see p. 254).

Allosteric inhibitors bind to a separate binding site outside the active center (**6**). This results in a *conformational change* in the enzyme protein that indirectly reduces its activity (see p. 116). Allosteric effects practically only occur in *oligomeric enzymes.* The kinetics of this type of system can no longer be described using the simple Michaelis–Menten model.

B. Inhibition kinetics ◑

In addition to the Lineweaver–Burk plot (see p. 92), the *Eadie–Hofstee plot* is also commonly used. In this case, the velocity *v* is plotted against *v* /[A]. In this type of plot, V_{max} corresponds to the intersection of the approximation lines with the v axis, while K_m is derived from the gradient of the lines. Competitive and non-competitive inhibitors are also easily distinguishable in the Eadie—Hofstee plot. As mentioned earlier, **competitive** inhibitors only influence K_m, and not V_{max}. The lines obtained in the absence and presence of an inhibitor therefore intersect on the ordinate. **Non-competitive inhibitors** produce lines that have the same slope (K_m unchanged) but intersect with the ordinate at a lower level. Another type of inhibitor, not shown here, in which V_{max} and K_m are reduced by the same factor, is referred to as **uncompetitive**. Inhibitors with purely uncompetitive effects are rare. A possible explanation for this type of inhibition is selective binding of the inhibitor to the EA complex.

Allosteric enzymes shift the target enzyme's saturation curve to the left (see p. 92). In Eadie–Hofstee and Lineweaver–Burk plots (see p. 92), allosteric enzymes are recognizable because they produce curved lines (not shown).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Types of inhibitor

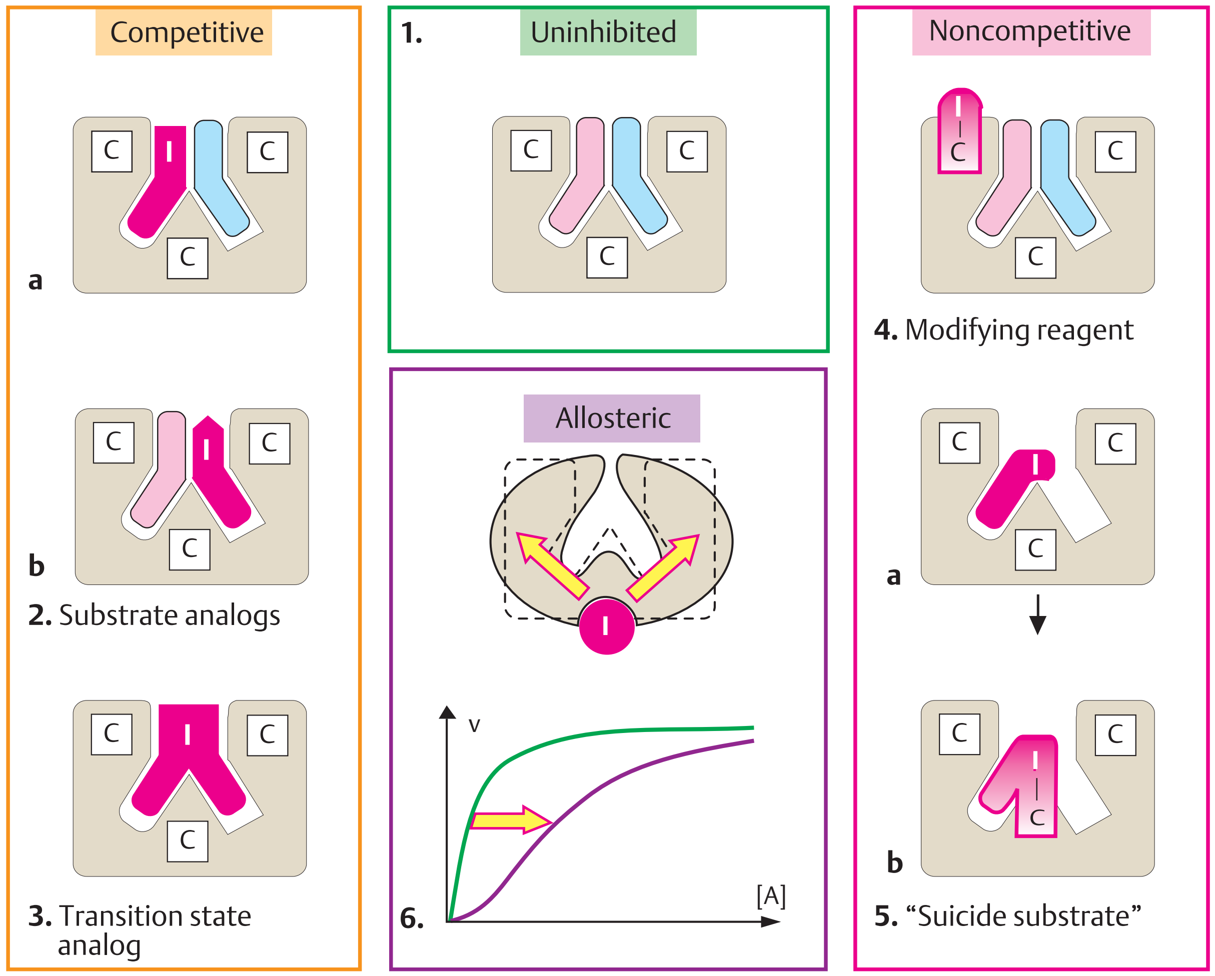
Competitive
a
b
2. Substrate analogs
3. Transition state analog
1.
Uninhibited
Allosteric
v
[A]
6.
Noncompetitive
4. Modifying reagent
a
b
5. "Suicide substrate"
C
I

B. Kinetics of inhibition

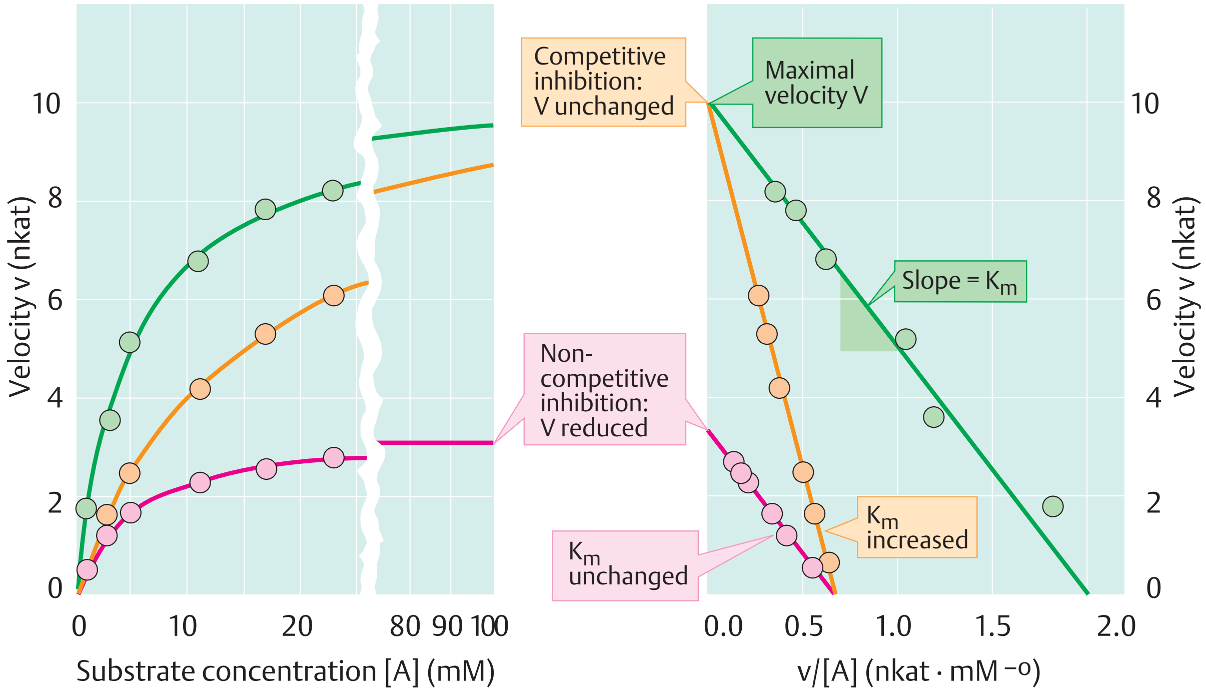

1. Hyperbolic plot

2. Eadie Hofstee plot

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Lactate dehydrogenase: structure

Lactate dehydrogenase (LDH, *EC 1.1.1.27*) is discussed in some detail here and on the next page as an example of the structure and function of an enzyme.

A. Lactate dehydrogenase: structure ○

The active form of lactate dehydrogenase (mass 144 kDa) is a **tetramer** consisting of four subunits (**1**). Each monomer is formed by a peptide chain of 334 amino acids (36 kDa). In the tetramer, the subunits occupy *equivalent positions* (**1**); each monomer has an active center. Depending on metabolic conditions, LDH catalyzes NADH-dependent reduction of pyruvate to lactate, or NAD^+-dependent oxidation of lactate to pyruvate (see p. 18).

The active center of an LDH subunit is shown schematically in Fig. **2**. The peptide backbone is shown as a light blue tube. Also shown are the substrate *lactate* (red), the coenzyme *NAD^+* (yellow), and three amino acid side chains (Arg-109, Arg-171, and His-195; green), which are directly involved in the catalysis. A *peptide loop* (pink) formed by amino acid residues 98–111 is also shown. In the absence of substrate and coenzyme, this partial structure is open and allows access to the substrate binding site (not shown). In the enzyme lactate NAD^+ complex shown, the peptide loop closes the active center. The catalytic cycle of lactate dehydrogenase is discussed on the next page.

B. Isoenzymes ◑

There are two *different* LDH subunits in the organism—M and H—which have a slightly different amino acid sequence and consequently different catalytic properties. As these two subunits can associate to form tetramers randomly, a total of five different **isoenzymes** of LDH are found in the body.

Fig. **1** shows sections from the amino acid sequences of the two subunits, using the single-letter notation (see p. 60). A common precursor gene was probably duplicated at some point in evolution. The two genes then continued to develop further independently of each other through mutation and selection. The differences in sequence between the M and H subunits are mainly *conservative*—i. e., both residues are of the same type, e. g. glycine (G) and alanine (A), or arginine (R) and lysine (K). Non-conservative exchanges are less frequent—e. g., lysine (K) for glutamine (Q), or threonine (T) for glutamic acid (E). Overall, the H subunit contains more acidic and fewer basic residues than the M form, and it therefore has a more strongly negative charge. This fact is exploited to separate the isoenzymes using electrophoresis (**2**; see pp. 78, 276). The isoenzyme LDH-1, consisting of four H subunits, migrates fastest, and the M_4 isoenzyme is slowest.

The separation and analysis of isoenzymes in blood samples is important in the diagnosis of certain diseases. Normally, only small amounts of enzyme activity are found in serum. When an organ is damaged, intracellular enzymes enter the blood and can be demonstrated in it (**serum enzyme diagnosis**). The total activity of an enzyme reflects the severity of the damage, while the type of isoenzyme found in the blood provides evidence of the site of cellular injury, since each of the genes is expressed in the various organs at different levels. For example, the liver and skeletal muscles mainly produce M subunits of lactate dehydrogenase (M for muscle), while the brain and cardiac muscle mainly express H subunits (H for heart). In consequence, each organ has a characteristic *isoenzyme pattern* (**3**). Following cardiac infarction, for example, there is a strong increase in the amount of LDH-1 in the blood, while the concentration of LDH-5 hardly changes. The isoenzymes of *creatine kinase* (see p. 336) are also of diagnostic importance.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Lactate dehydrogenase: structure

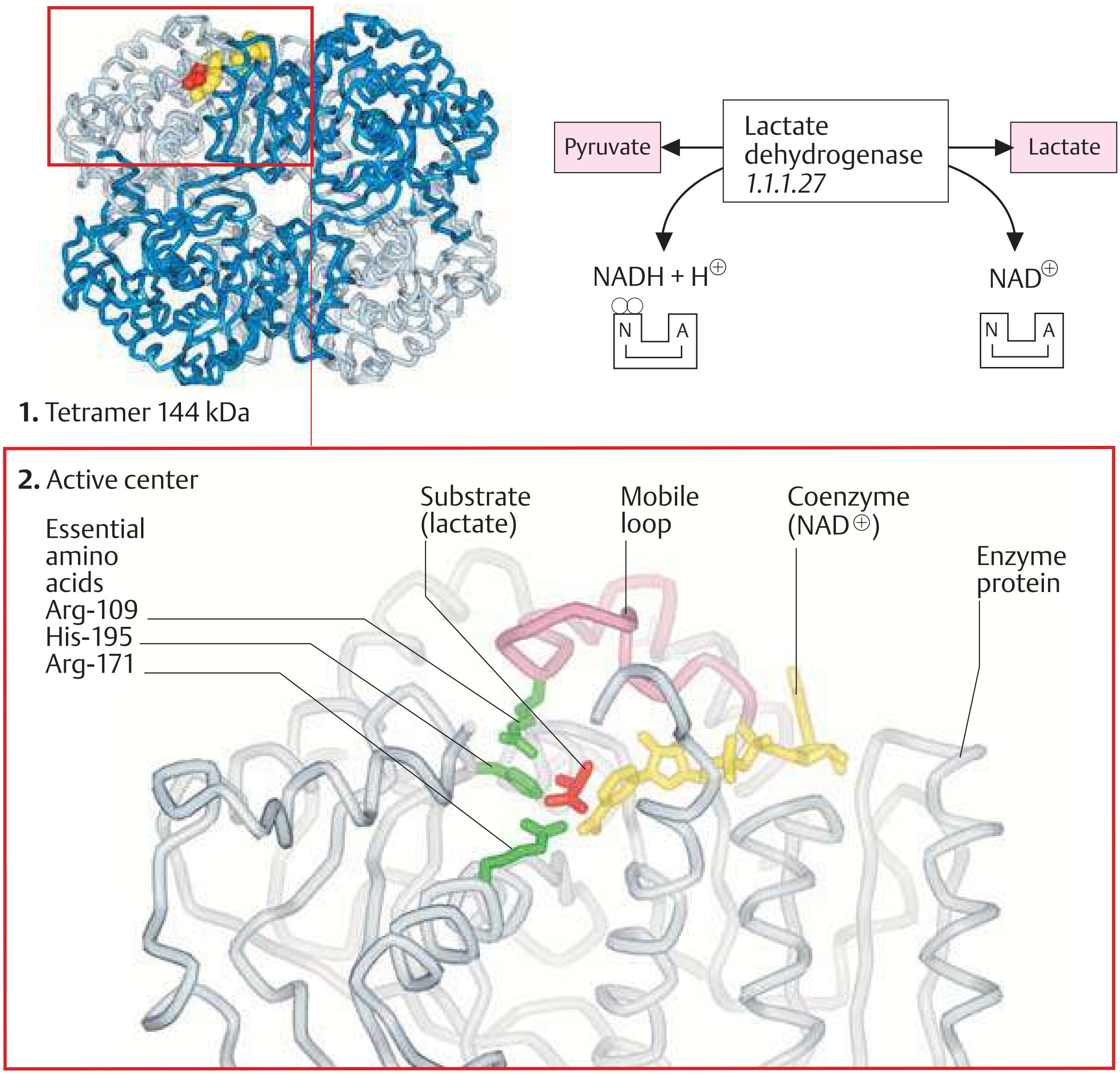
Pyruvate
Lactate dehydrogenase
1.1.1.27
Lactate
NADH + H⊕
N
A
NAD⊕
N
A
1. Tetramer 144 kDa
2. Active center
Essential amino acids
Arg-109
His-195
Arg-171
Substrate (lactate)
Mobile loop
Coenzyme (NAD⊕)
Enzyme protein

B. Isoenzymes

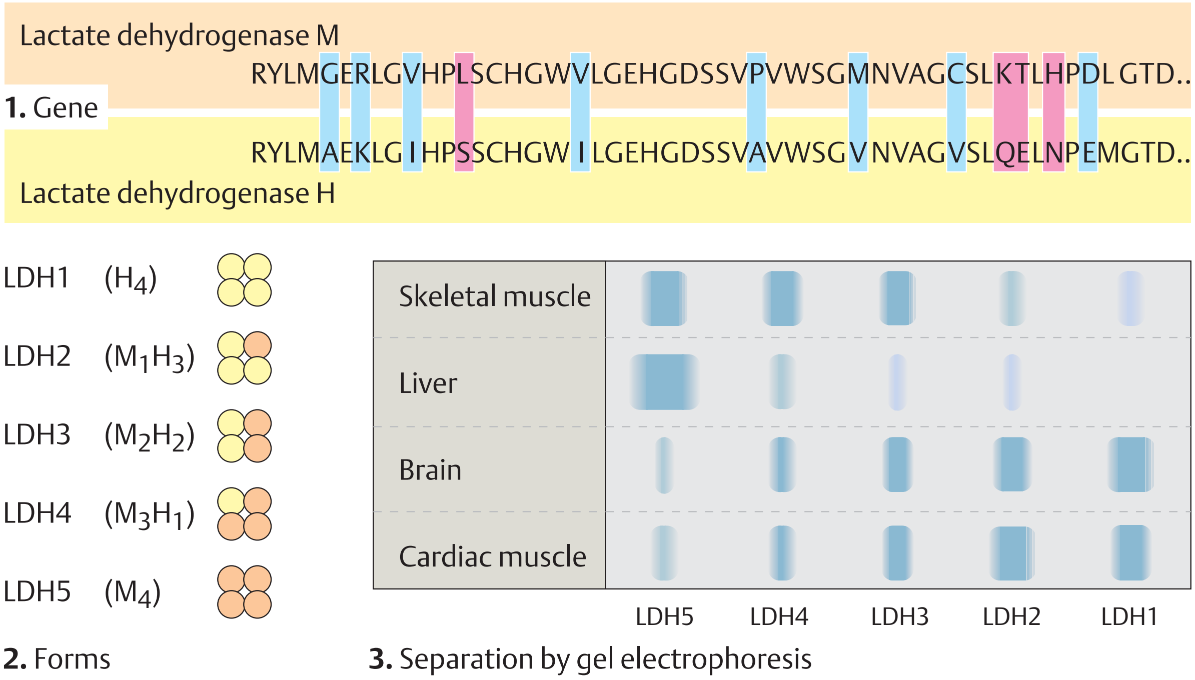
Lactate dehydrogenase M
RYLMGERLGVHPLSCHGWVLGEHGDSSVPVWSGMNVAGCSLKTLHPDLGTD..
1. Gene
RYLMAEKLGIHPSSCHGWILGEHGDSSVAVWSGVNVAGVSLQELNPEMGTD..
Lactate dehydrogenase H
LDH1 (H4)
LDH2 (M1H3)
LDH3 (M2H2)
LDH4 (M3H1)
LDH5 (M4)
2. Forms
Skeletal muscle
Liver
Brain
Cardiac muscle
LDH5
LDH4
LDH3
LDH2
LDH1
3. Separation by gel electrophoresis

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Lactate dehydrogenase: mechanism

The principles of enzyme catalysis discussed on p. 90 can be illustrated using the reaction mechanism of lactate dehydrogenase (LDH) as an example.

A. Lactate dehydrogenase: catalytic cycle ○

LDH catalyzes the transfer of hydride ions (see p. 32) from lactate to NAD$^+$ or from NADH to pyruvate.

$$\text{L-lactate} + \text{NAD}^+ \rightleftharpoons \text{pyruvate} + \text{NADH} + \text{H}^+$$

The equilibrium of the reaction strongly favors lactate *formation*. At high concentrations of lactate and NAD$^+$, however, oxidation of lactate to pyruvate is also possible (see p. 18). LDH catalyzes the reaction in *both* directions, but—like all enzymes—it has *no* effect on chemical equilibrium.

As the reaction is reversible, the catalytic process can be represented as a closed loop. The **catalytic cycle** of LDH is reduced to six “snapshots” here. Intermediate steps in catalysis such as those shown here are extremely short-lived and therefore difficult to detect. Their existence was deduced indirectly from a large number of experimental findings—e. g., kinetic and binding measurements.

Many amino acid residues play a role in the **active center** of LDH. They can mediate the binding of the substrate and coenzyme, or take part in one of the steps in the catalytic cycle directly. Only the side chains of three particularly important residues are shown here. The positively charged guanidinium group of **arginine-171** binds the carboxylate group of the substrate by electrostatic interaction. The imidazole group of **histidine-195** is involved in acid–base catalysis, and the side chain of **arginine-109** is important for the stabilization of the transition state. In contrast to His-195, which changes its charge during catalysis, the two essential arginine residues are constantly protonated. In addition to these three residues, the **peptide loop 98–111** mentioned on p. 98 is also shown here schematically (red). Its function consists of closing the active center after binding of the substrate and coenzyme, so that water molecules are largely excluded during the electron transfer.

We can now look at the **partial reactions** involved in LDH-catalyzed pyruvate reduction.

In the free enzyme, His195 is protonated (**1**). This form of the enzyme is therefore described as E H^+. The coenzyme NADH is bound first (**2**), followed by pyruvate (**3**). It is important that the carbonyl group of the pyruvate in the enzyme and the active site in the nicotinamide ring of the coenzyme should have a fairly optimal position in relation to each other, and that this orientation should become fixed (*proximity and orientation of the substrates*). The 98–111 loop now closes over the active center. This produces a marked decrease in polarity, which makes it easier to achieve the **transition state** (**4**; *water exclusion*). In the transition state, a hydride ion, H$^-$ (see p. 32), is transferred from the coenzyme to the carbonyl carbon (*group transfer*). The transient—and energetically unfavorable—negative charge on the oxygen that occurs here is stabilized by electrostatic interaction with Arg-109 (*stabilization of the transition state*). At the same time, a proton from His-195 is transferred to this oxygen atom (*group transfer*), giving rise to the enzyme-bound products lactate and NAD$^+$ (**5**). After the loop opens, lactate dissociates from the enzyme, and the temporarily uncharged imidazole group in His-195 again binds a proton from the surrounding water (**6**). Finally, the oxidized coenzyme NAD$^+$ is released, and the initial state (**1**) is restored. As the diagram shows, the proton that appears in the reaction equation (NADH + H$^+$) is not bound together with NADH, but after release of the lactate—i. e., between steps (**5**) and (**6**) of the *previous* cycle.

Exactly the same steps occur during the oxidation of lactate to pyruvate, but in the opposite direction. As mentioned earlier, the direction which the reaction takes depends not on the enzyme, but on the equilibrium state—i. e., on the concentrations of all the reactants and the pH value (see p. 18).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Lactate dehydrogenase: catalytic cycle

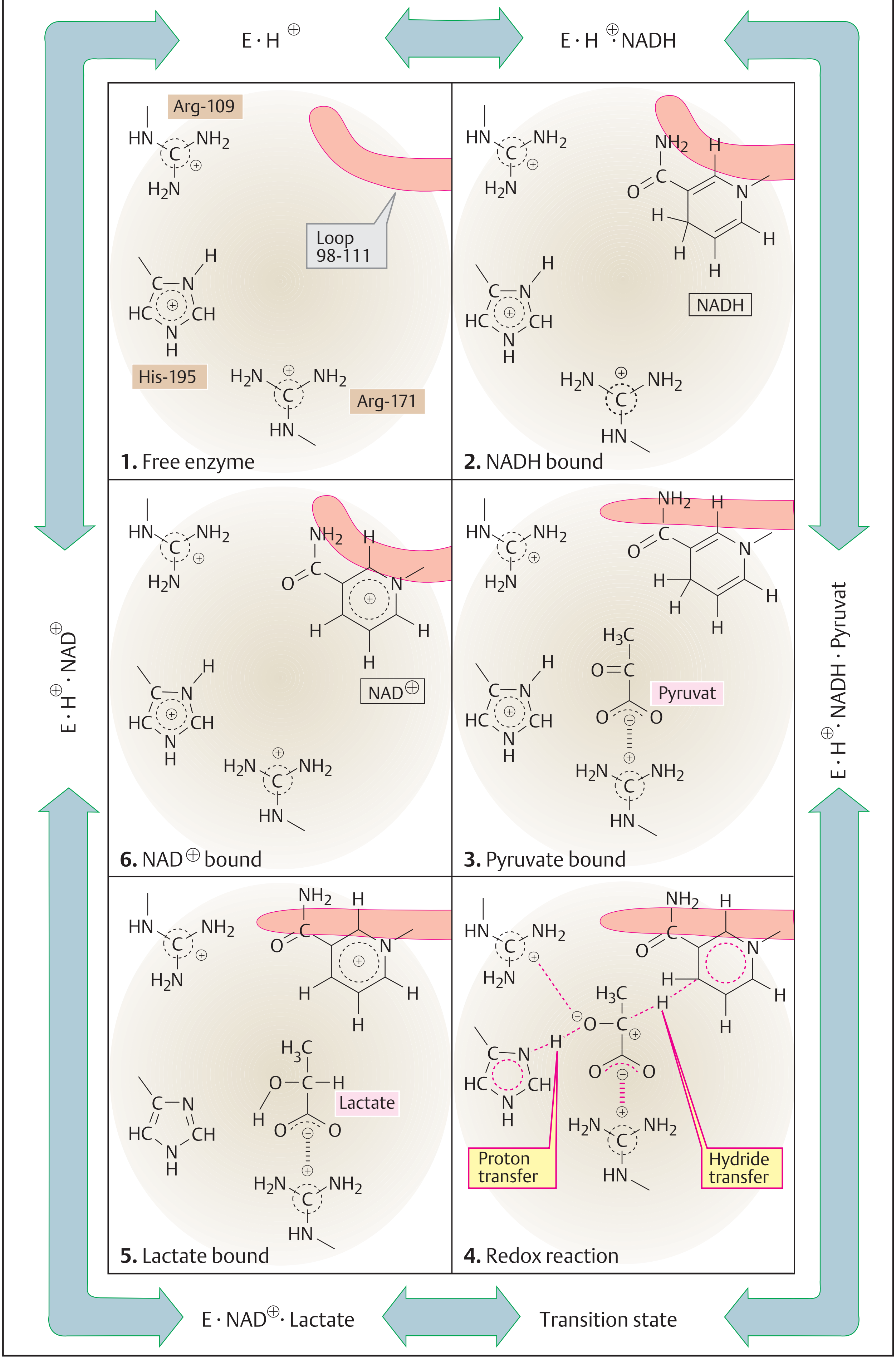
E · H⊕
E · H⊕ · NADH
Arg-109
Loop 98-111
His-195
Arg-171
1. Free enzyme
NADH
2. NADH bound
NAD⊕
6. NAD⊕ bound
Pyruvat
3. Pyruvate bound
Lactate
5. Lactate bound
Proton transfer
Hydride transfer
4. Redox reaction
E · H⊕ · NAD⊕
E · H⊕ · NADH · Pyruvat
E · NAD⊕ · Lactate
Transition state

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Enzymatic analysis

Enzymes play an important role in *biochemical analysis*. In biological material—e. g., in body fluids—even tiny quantities of an enzyme can be detected by measuring its catalytic activity. However, enzymes are also used as *reagents* to determine the concentrations of metabolites—e. g., the blood glucose level (**C**). Most enzymatic analysis procedures use the method of spectrophotometry (**A**).

A. Principle of spectrophotometry ○

Many substances *absorb* light in the visible or ultraviolet region of the spectrum. This property can be used to determine the concentration of such a substance. The extent of light absorption depends on the type and concentration of the substance and on the wavelength of the light used. **Monochromatic light**—i. e., light with a defined wavelength isolated from white light using a monochromator—is therefore used. Monochromatic light with an intensity of I_0 is passed through a rectangular vessel made of glass or quartz (a *cuvet*), which contains a solution of the absorbing substance. The **absorption A** of the solution (often also referred to as its *extinction*) is defined as the *negative decadic logarithm of the quotient* I/I_0. The **Beer–Lambert law** states that A is proportional to the concentration c of the absorbing substance and the thickness d of the solution it passes through. As mentioned earlier, the **absorption coefficient** ε depends on the type of substance and the wavelength.

B. Measurement of lactate dehydrogenase activity ○

Measurement of lactate dehydrogenase (LDH) activity takes advantage of the fact that while the reduced coenzyme NADH + H^+ absorbs light at 340 nm, oxidized NAD^+ does not. *Absorption spectra* (i. e., plots of A against the wavelength) for the substrates and the coenzymes of the LDH reaction are shown in Fig. **1**. Differences in absorption behavior between NAD^+ and NADH between 300 and 400 nm result from changes in the nicotinamide ring during oxidation or reduction (see p. 32). To measure the activity, a solution containing lactate and NAD^+ is placed in a cuvet, and absorption is recorded at a *constant wavelength* of 340 nm. The uncatalyzed LDH reaction is very slow. It is only after addition of the enzyme that measurable quantities of NADH are formed and absorption increases. Since according to the Beer–Lambert law the rate of the increase in absorption $\Delta A/\Delta t$ is proportional to the reaction rate $\Delta c/\Delta t$. The absorption coefficient ε at 340 nm or comparison with a standard solution can be used to calculate LDH activity.

C. Enzymatic determination of glucose ○

Most biomolecules do not show any absorption in the visible or ultraviolet spectrum. In addition, they are usually present in the form of mixtures with other—similar—compounds that would also react to a chemical test procedure. These two problems can be avoided by using an appropriate enzyme to produce a colored dye selectively from the metabolite that is being analyzed. The absorption of the dye can then be measured.

A procedure (**1**) that is often used to measure glucose when monitoring blood glucose levels (see p. 160) involves two successive reactions. The glucose-specific enzyme *glucose oxidase* (obtained from fungi) first produces hydrogen peroxide, H_2O_2, which in the second step—catalyzed by a *peroxidase*—oxidizes a colorless precursor into a green dye (**2**). When all of the glucose in the sample has been used up, the amount of dye formed—which can be measured on the basis of its light absorption—is equivalent to the quantity of glucose originally present.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Principle of spectrophotometry

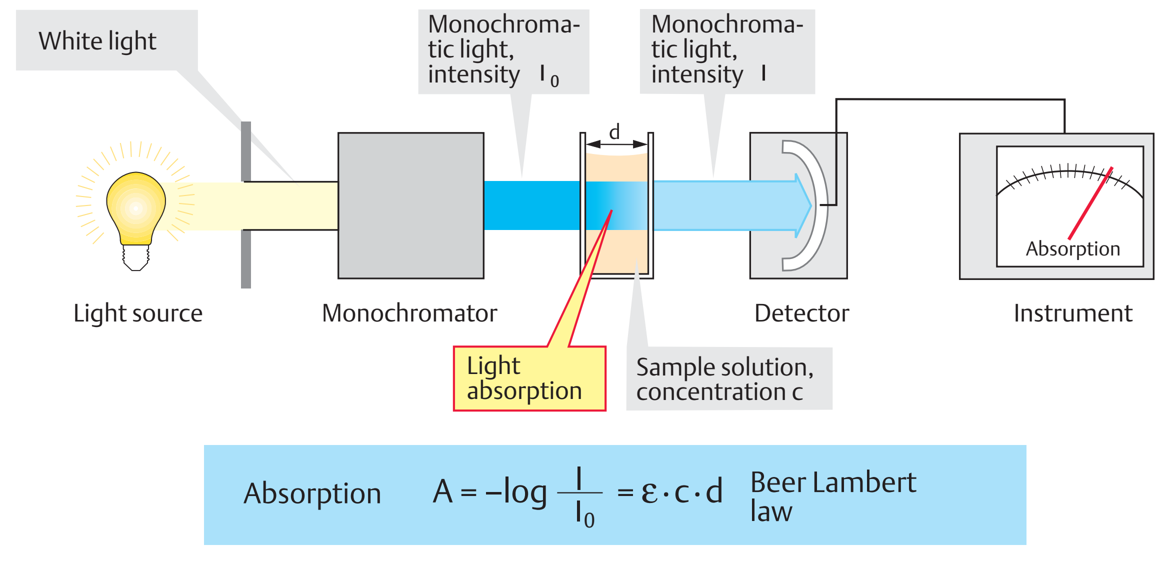
White light
Monochromatic light, intensity I_0
Monochromatic light, intensity I
d
Light source
Monochromator
Detector
Instrument
Absorption
Light absorption
Sample solution, concentration c
Absorption $A = -\log \frac{I}{I_0} = \varepsilon \cdot c \cdot d$ Beer Lambert law

B. Assay of lactate dehydrogenase activity

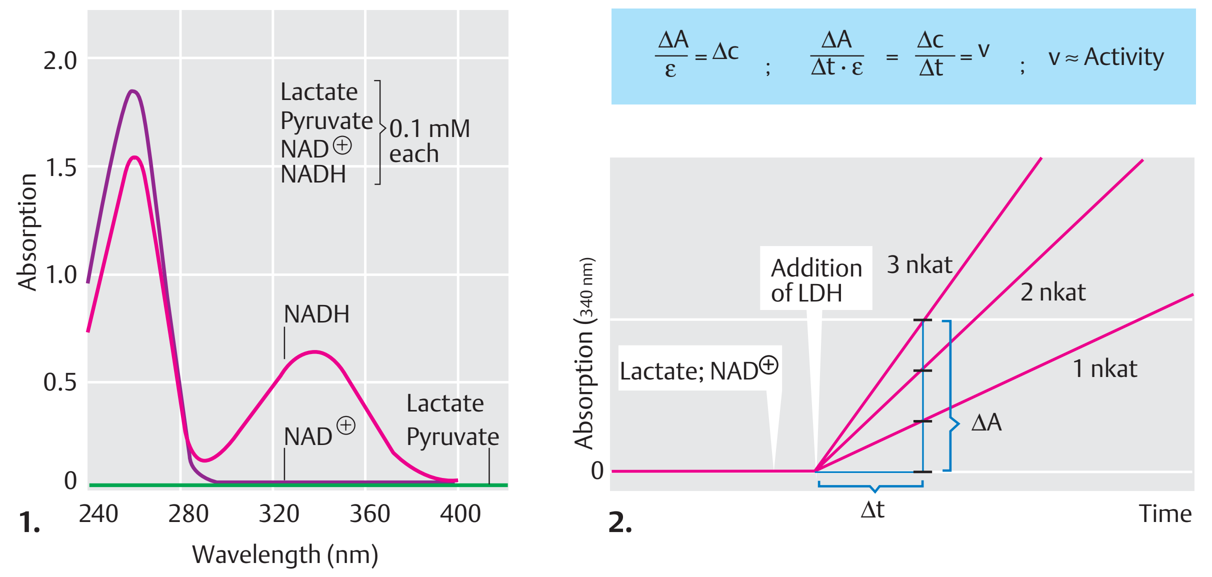

C. Enzymatic determination of glucose

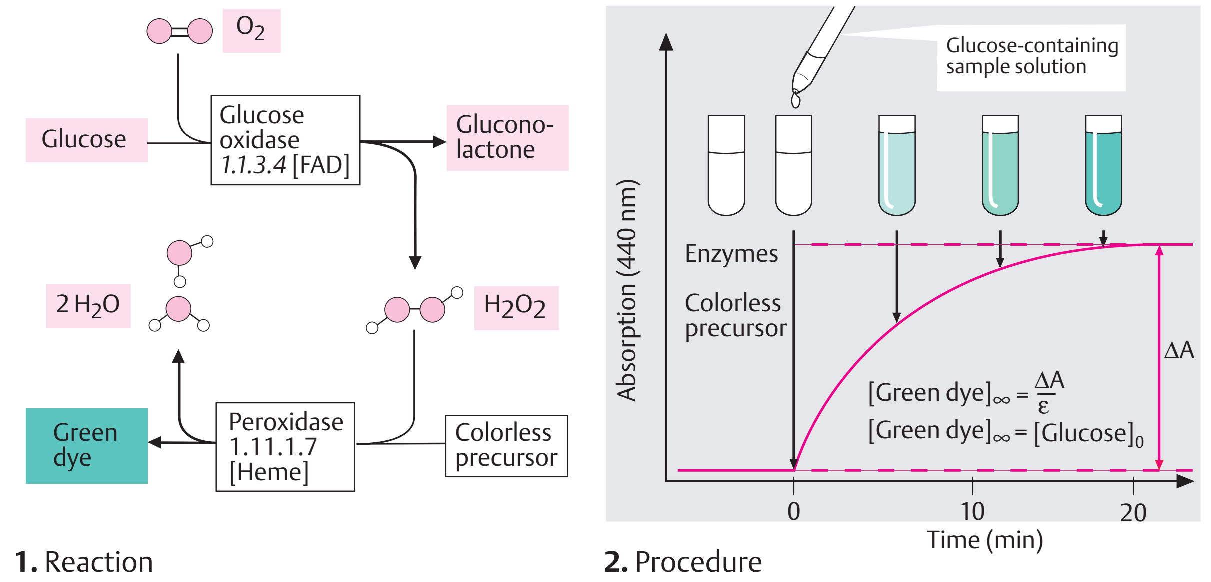
O_2
Glucose
Glucose oxidase 1.1.3.4 [FAD]
Glucono-lactone
$2\,H_2O$
H_2O_2
Green dye
Peroxidase 1.11.1.7 [Heme]
Colorless precursor
1. Reaction
Glucose-containing sample solution
Enzymes
Colorless precursor
Absorption (440 nm)
ΔA
$[\text{Green dye}]_\infty = \frac{\Delta A}{\varepsilon}$
$[\text{Green dye}]_\infty = [\text{Glucose}]_0$
0
10
20
Time (min)
2. Procedure

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Coenzymes 1

A. Coenzymes: definitions ●

In many enzyme-catalyzed reactions, electrons or groups of atoms are transferred from one substrate to another. This type of reaction always also involves additional molecules, which temporarily accept the group being transferred. Helper molecules of this type are called **coenzymes**. As they are not catalytically active themselves, the less frequently used term *"cosubstrate"* would be more appropriate. In contrast to substrates for which a given enzyme is usually specific (see p. 88), coenzymes cooperate with many enzymes of varying substrate specificity. We have rather arbitrarily divided the coenzymes here into group-transferring and redox coenzymes. Strictly speaking, redox coenzymes also transfer groups—namely, reducing equivalents (see p. 32).

Depending on the type of interaction with the enzyme, a distinction is made between soluble coenzymes and prosthetic groups. **Soluble coenzymes** (**1**) are *bound like substrates* during a reaction, undergo a chemical change, and are then *released again*. The original form of the coenzyme is regenerated by a second, independent reaction. **Prosthetic groups** (**2**), on the other hand, are coenzymes that are *tightly bound to the enzyme* and remain associated with it during the reaction. The part of the substrate bound by the coenzyme is later transferred to another substrate or coenzyme of the *same* enzyme (not shown in Fig. **2**).

B. Redox coenzymes 1 ◑

All oxidoreductases (see p. 88) require coenzymes. The most important of these redox coenzymes are shown here. They can act in soluble form (S) or prosthetically (P). Their normal potentials $E^{0'}$ are shown in addition to the type of reducing equivalent that they transfer (see p. 18).

The pyridine nucleotides **NAD⁺** and **NADP⁺** (**1**) are widely distributed as coenzymes of dehydrogenases. They transport *hydride ions* ($2e^-$ and 1 H^+; see p. 32) and *always* act in soluble form. NAD^+ transfers reducing equivalents from catabolic pathways to the respiratory chain and thus contributes to energy metabolism. In contrast, reduced $NADP^+$ is the most important *reductant* involved in biosynthesis (see p. 112).

The flavin coenzymes **FMN** and **FAD** (**2**, **3**) contain *flavin* (isoalloxazine) as a redox-active group. This is a three-membered, *N*-containing ring system that can accept a maximum of two electrons and two protons during reduction. FMN carries the phosphorylated sugar alcohol *ribitol* at the flavin ring. FAD arises from FMN through bonding with AMP. The two coenzymes are functionally similar. They are found in *dehydrogenases, oxidases,* and *monooxygenases*. In contrast to the pyridine nucleotides, flavin reactions give rise to *radical intermediates* (see p. 32). To prevent damage to cell components, the flavins always remain bound as prosthetic groups in the enzyme protein.

The role of **ubiquinone** (coenzyme Q, **4**) in transferring reducing equivalents in the respiratory chain is discussed on p. 140. During reduction, the *quinone* is converted into the *hydroquinone* (ubiquinol). The isoprenoid side chain of ubiquinone can have various lengths. It holds the molecule in the membrane, where it is freely mobile. Similar coenzymes are also found in photosynthesis (plastoquinone; see p. 132). *Vitamins E and K* (see p. 52) also belong to the quinone/hydroquinone systems.

L-Ascorbic acid (vitamin C, **5**) is a powerful reducing agent. As an **antioxidant**, it provides nonspecific protection against oxidative damage (see p. 284), but it is also an essential **cofactor** for various monooxygenases and dioxygenases. Ascorbic acid is involved in the hydroxylation of proline and lysine residues during the biosynthesis of collagen (see p. 344), in the synthesis of catecholamines (see p. 352) and bile acids (see p. 314), as well as in the breakdown of tyrosine (see p. 415). The reduced form of the coenzyme is a relatively strong acid and forms salts, the **ascorbates**. The oxidized form is known as **dehydroascorbic acid**. The stimulation of the immune system caused by ascorbic acid has not yet been fully explained.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Coenzymes: definitions

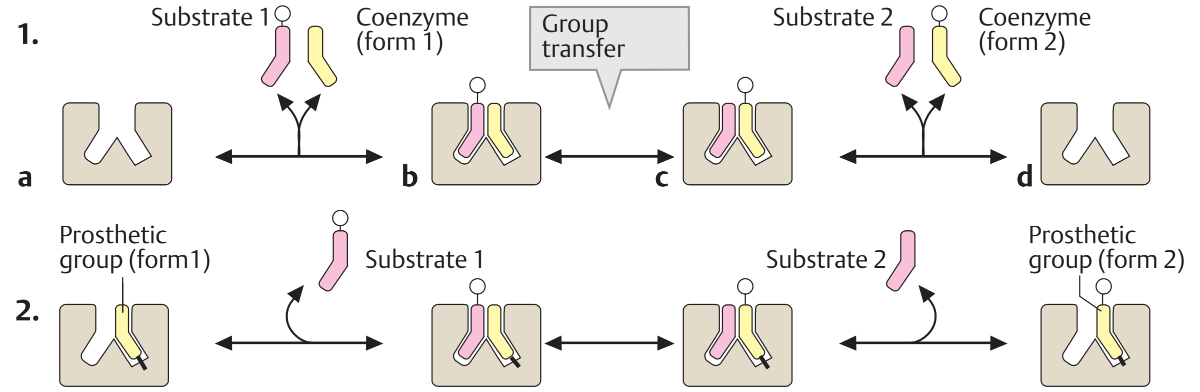
1.
Substrate 1
Coenzyme (form 1)
Group transfer
Substrate 2
Coenzyme (form 2)
a
b
c
d
2.
Prosthetic group (form1)
Substrate 1
Substrate 2
Prosthetic group (form 2)

B. Redox coenzymes

Coenzyme	Oxidized form Reduced form	Type	Trans-ferred	$E^{0\prime}$ (V)
1. NAD(P)$^{\oplus}$ 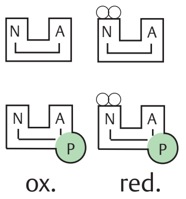	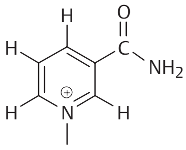 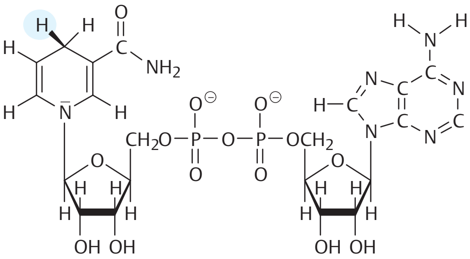	L	$H^{\ominus}$	−0.32
2. Flavin mononucleotide (FMN) 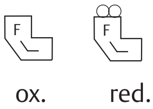	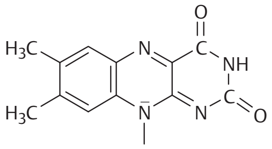 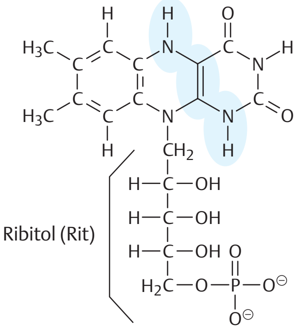 Ribitol (Rit)	P	2[H]	−0.3 to +0.2
3. Flavin adenine dinucleotide (FAD) 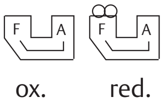	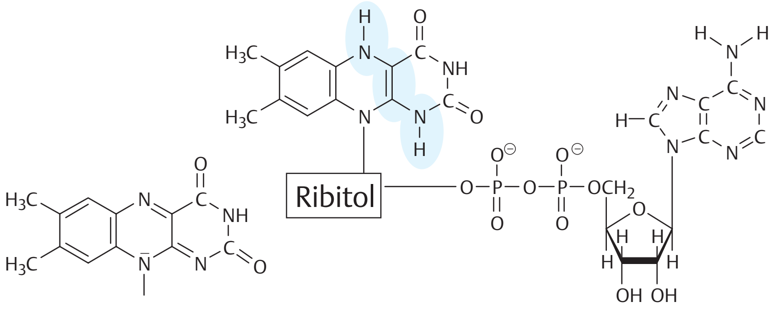	P	2[H]	−0.3 to +0.2
4. Ubiquinone (coenzym Q) 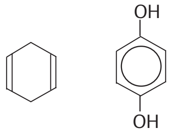	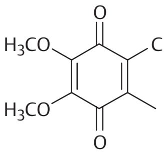 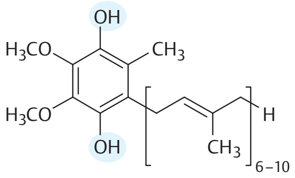	L	2[H]	−0 to +0.2
5. Ascorbic acid	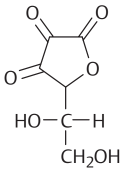 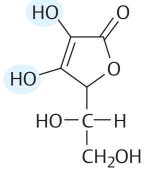	L	2[H]	+0.1

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Coenzymes 2

A. Redox coenzymes 2 ◑

In **lipoic acid** (**6**), an intramolecular *disulfide bond* functions as a redox-active structure. As a result of reduction, it is converted into the corresponding *dithiol*. As a prosthetic group, lipoic acid is usually covalently bound to a lysine residue (R) of the enzyme, and it is then referred to as **lipoamide**. Lipoamide is mainly involved in oxidative decarboxylation of 2-oxo acids (see p. 134). The peptide coenzyme **glutathione** is a similar disulfide/dithiol system (not shown; see p. 284).

Iron–sulfur clusters (**7**) occur as prosthetic groups in oxidoreductases, but they are also found in lyases—e. g., *aconitase* (see p. 136) and other enzymes. Iron–sulfur clusters consist of 2–4 iron ions that are coordinated with cysteine residues of the protein (–SR) and with anorganic sulfide ions (S). Structures of this type are only stable in the interior of proteins. Depending on the number of iron and sulfide ions, distinctions are made between [Fe_2S_2], [Fe_3S_4], and [Fe_4S_4] clusters. These structures are particularly numerous in the respiratory chain (see p. 140), and they are found in all complexes except complex IV.

Heme coenzymes (**8**) with redox functions exist in the *respiratory chain* (see p. 140), in *photosynthesis* (see p. 128), and in *monooxygenases* and *peroxidases* (see p. 24). Heme-containing proteins with redox functions are also referred to as **cytochromes**. In cytochromes, in contrast to hemoglobin and myoglobin, the iron changes its valence (usually between +2 and +3). There are several classes of heme (a, b, and c), which have different types of substituent $-R_1$ to $-R_3$. Hemoglobin, myoglobin, and the heme enzymes contain heme b. Two types of heme a are found in cytochrome *c* oxidase (see p. 132), while heme c mainly occurs in cytochrome *c*, where it is covalently bound with cysteine residues of the protein part via thioester bonds.

B. Group-transferring coenzymes 1 ◑

The **nucleoside phosphates** (**1**) are not only *precursors* for nucleic acid biosynthesis; many of them also have coenzyme functions. They serve for *energy conservation*, and as a result of *energetic coupling* (see p. 124) also allow endergonic processes to proceed. Metabolites are often made more reactive ("activated") as a result of the transfer of phosphate residues (*phosphorylation*). Bonding with nucleoside diphosphate residues (mainly UDP and CDP) provides activated precursors for polysaccharides and lipids (see p. 110). Endergonic formation of bonds by *ligases* (enzyme class 6) also depends on nucleoside triphosphates.

Acyl residues are usually activated by transfer to **coenzyme A** (**2**). In coenzyme A (see p. 12), *pantetheine* is linked to 3′-phospho-ADP by a phosphoric acid anhydride bond. Pantetheine consists of three components connected by amide bonds—*pantoic acid*, *β-alanine*, and *cysteamine*. The latter two components are biogenic amines formed by the decarboxylation of aspartate and cysteine, respectively. The compound formed from pantoic acid and β–alanine (*pantothenic acid*) has vitamin-like characteristics for humans (see p. 368). Reactions between the thiol group of the cysteamine residue and carboxylic acids give rise to **thioesters**, such as acetyl CoA. This reaction is strongly endergonic, and it is therefore coupled to exergonic processes. Thioesters represent the *activated form of carboxylic acids,* because acyl residues of this type have a high chemical potential and are easily transferred to other molecules. This property is often exploited in metabolism.

Thiamine diphosphate (TPP, **3**), in cooperation with enzymes, is able to activate aldehydes or ketones as *hydroxyalkyl groups* and then to pass them on to other molecules. This type of transfer is important in the transketolase reaction, for example (see p. 152). Hydroxyalkyl residues also arise in the decarboxylation of oxo acids. In this case, they are released as aldehydes or transferred to lipoamide residues of 2-oxoacid dehydrogenases (see p. 134). The functional component of TPP is the sulfur- and nitrogen-containing *thiazole ring.*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Redox coenzymes 2

Coenzyme	Oxidized form	Reduced form	Type	Transferred	E°'
6. Lipoamide S S ox. HS SH red.	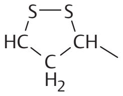	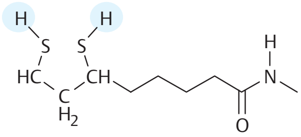	P	2[H]	−0.29
7. Iron–sulfur cluster	$[Fe_2S_2]^{n+}$ 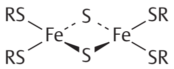	$[Fe_4S_4]^{m+}$ 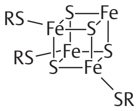	P	$1e^{\ominus}$	−0.6 to +0.5
8. Heme 3+ ox. 2+ red.	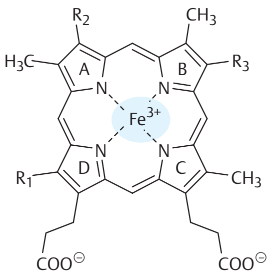	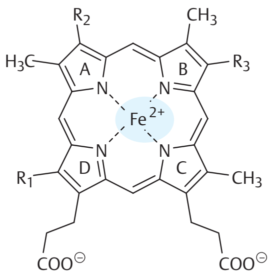	P	$1e^{\ominus}$	0 to +0.5

B. Group-transferring coenzymes 1

Coenzyme (symbol)	Free form	Charged form	Group(s) transferred	Important enzymes
1. Nucleoside phosphates N Ⓟ Ⓟ Ⓟ	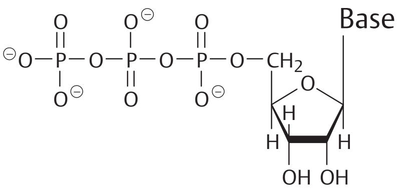		Ⓟ B-Rib B-Rib-Ⓟ B-Rib-ⓅⓅ	Phosphotransferases Nucleotidyltransferases (2.7.n.n) Ligases (6.n.n.n)
2. Coenzyme A A S	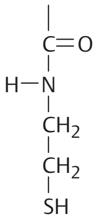	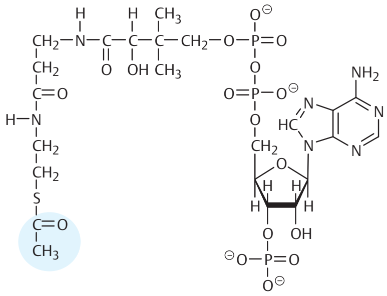	Acyl residues	Acyltransferases (2.3.n.n) CoA transferases (2.8.3.n)
3. Thiamine diphosphate TPP	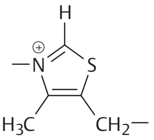	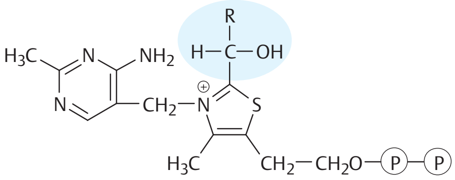	Hydroxyalkyl residues	Decarboxylases (4.1.1.n) Oxoacid dehydrogenases (1.2.4. n) Transketolase (2.2.1.1)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Coenzymes 3

A. Group-transferring coenzymes 2 ◑

Pyridoxal phosphate (**4**) is the most important coenzyme in amino acid metabolism. Its role in *transamination* reactions is discussed in detail on p. 178. Pyridoxal phosphate is also involved in other reactions involving amino acids, such as *decarboxylations* and *dehydrations*. The aldehyde form of pyridoxal phosphate shown here (left) is not generally found in free form. In the absence of substrates, the aldehyde group is covalently bound to the ε-amino group of a lysine residue as *aldimine* ("Schiff's base"). **Pyridoxamine phosphate** (right) is an intermediate of transamination reactions. It reverts to the aldehyde form by reacting with 2-oxoacids (see p. 178).

Biotin (**5**) is the coenzyme of the *carboxylases.* Like pyridoxal phosphate, it has an amide-type bond via the carboxyl group with a lysine residue of the carboxylase. This bond is catalyzed by a specific enzyme. Using ATP, biotin reacts with hydrogen carbonate (HCO_3^-) to form *N*-**carboxybiotin** . From this activated form, *carbon dioxide* (CO_2) is then transferred to other molecules, into which a carboxyl group is introduced in this way. Examples of biotindependent reactions of this type include the formation of oxaloacetic acid from pyruvate (see p. 154) and the synthesis of malonyl-CoA from acetyl-CoA (see p. 162).

Tetrahydrofolate (THF, **6**) is a coenzyme that can transfer *C_1 residues* in different oxidation states. THF arises from the vitamin *folic acid* (see p. 366) by double hydrogenation of the heterocyclic pterin ring. The C_1 units being transferred are bound to N-5, N-10, or both nitrogen atoms. The most important derivatives are:

a) N^5**-formyl-THF** and N^{10}**-formyl-THF**, in which the formyl residue has the oxidation state of a carboxylic acid;

b) N^5**-methylene-THF**, with a C_1 residue in the oxidation state of an aldehyde; and

c) N^5**-methyl-THF**, in which the methyl group has the oxidation state of an alcohol.

C_1 units transferred by THF play a role in the synthesis of methionine (see p. 412), purine nucleotides (see p. 188), and dTMP (see p. 190), for example. Due to the central role of THF derivatives in the biosynthesis of DNA precursors, the enzymes involved in THF metabolism are primary targets for cytostatic drugs (see p. 402).

The **cobalamins** (**7**) are the chemically most complex form of coenzyme. They also represent the only natural substances that contain the transition metal *cobalt* (Co) as an essential component. Higher organisms are unable to synthesize cobalamins themselves, and are therefore dependent on a supply of **vitamin B_{12}** synthesized by bacteria (see p. 368).

The central component of the cobalamins is the **corrin** ring, a member of the tetrapyrroles, at the center of which the cobalt ion is located. The end of one of the side chains of the ring carries a nucleotide with the unusual base *dimethylbenzimidazole.* The ligands for the metal ion are the four N atoms of the pyrrole ring, a nitrogen from dimethylbenzimidazole, and a **group X**, which is organometallically bound—i. e., *mainly covalently.*

In **methylcobalamin**, X is a methyl group. This compound functions as a coenzyme for several *methyltransferases,* and among other things is involved in the synthesis of methionine from homocysteine (see p. 418). However, in human metabolism, in which methionine is an essential amino acid, this reaction does not occur.

Adenosylcobalamin (coenzyme B_{12}) carries a covalently bound adenosyl residue at the metal atom. This is a coenzyme of various *isomerases,* which catalyze rearrangements following a radical mechanism. The radical arises here through *homolytic cleavage* of the bond between the metal and the adenosyl group. The most important reaction of this type in animal metabolism is the rearrangement of methylmalonyl-CoA to form succinyl-CoA, which completes the breakdown of odd-numbered fatty acids and of the branched amino acids valine and isoleucine (see pp. 166 and 414).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Group-transferring coenzymes 2

Coenzyme	Free form	Charged form	Group(s) transferred	Important enzymes
4. Pyridoxal phosphate PLP	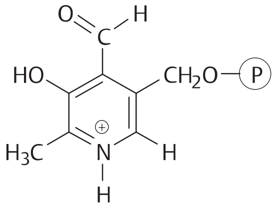	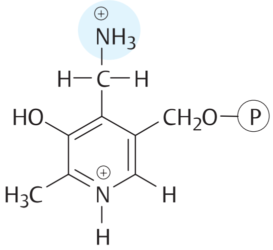	Amino group Amino acid residues	Transaminases (2.6.1.n) Many lyases (4.n.n.n)
5. Biotin B	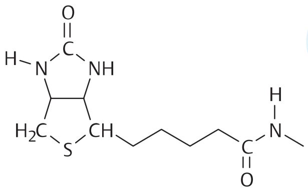	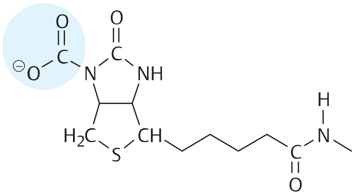	[CO_2]	Carboxylases (6.4.1.n)
4. Pyridoxal phosphate THF	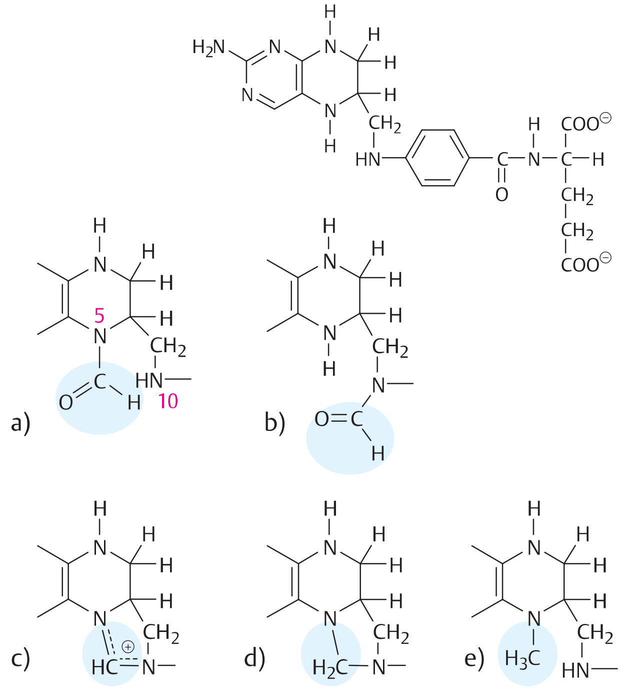		C_1 groups a) N^5-Formyl b) N^{10}-Formyl c) N^5N^{10}-Methenyl d) N^5N^{10}-Methylene e) N^5N^{10}-Methyl	C_1 transferases (2.1.n.n)
7. Cobalamin coenzymes	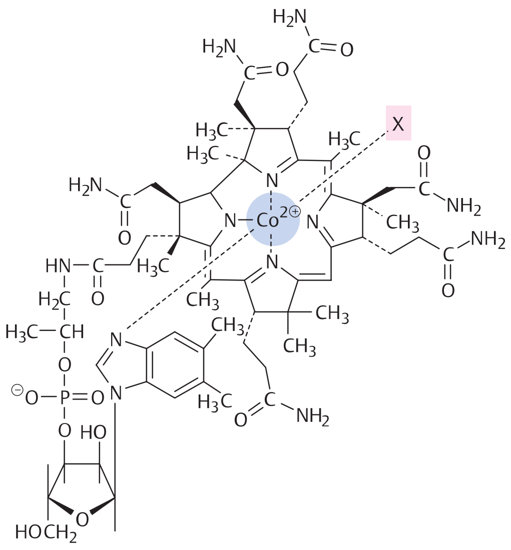		X = Adenosyl- X = Methyl-	Mutases *(5.4.n.n)* Methyltransferases *(2.1.1.n.)*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Activated metabolites

Many coenzymes (see pp. 104ff.) serve to *activate* molecules or groups that are poorly reactive. Activation consists of the formation of reactive intermediate compounds in which the group concerned is located at a higher chemical potential and can therefore be transferred to other molecules in an exergonic reaction (see p. 124). Acetyl-CoA is an example of this type of compound (see p. 12).

ATP and the other **nucleoside triphosphate coenzymes** not only transfer phosphate residues, but also provide the nucleotide components for this type of activation reaction. On this page, we discuss metabolites or groups that are activated in the metabolism by bonding with nucleosides or nucleotides. Intermediates of this type are mainly found in the metabolism of complex carbohydrates and lipids.

A. Activated metabolites ◑

1. Uridine diphosphate glucose (UDPglucose)

The inclusion of glucose residues into polymers such as glycogen or starches is an endergonic process. The activation of the **glucose** building blocks that is required for this takes places in several steps, in which two ATPs are used per glucose. After the phosphorylation of free glucose, glucose 6-phosphate is isomerized to glucose 1-phosphate (**a**), reaction with UTP (**b**) then gives rise to UDPglucose, in which the anomeric OH group at C-1 of the sugar is bound with phosphate. This “energy-rich” compound (an acetal phosphate) allows exergonic transfer of glucose residues to glycogen (**c**; see pp. 156, 408) or other acceptors.

2. Cytidine diphosphate choline (CDPcholine)

The amino alcohol **choline** is activated for inclusion in phospholipids following a similar principle (see p. 170). Choline is first phosphorylated by ATP to form choline phosphate (**a**), which by reaction with CTP and cleavage of diphosphate, then becomes CDPcholine. In contrast to (**1**), it is not choline that is transferred from CDPcholine, but rather choline phosphate, which with diacylglycerol yields phosphatidylcholine (lecithin).

3. Phosphoadenosine phosphosulfate (PAPS)

Sulfate residues occur as strongly polar groups in various biomolecules—e. g., in *glycosaminoglycans* (see p. 346) and *conjugates* of steroid hormones and xenobiotics (see p. 316). In the synthesis of the “activated sulfate” PAPS, ATP first reacts with anorganic sulfate to form adenosine phosphosulfate (APS, **a**). This intermediate already contains the “energy-rich” mixed anhydride bond between phosphoric acid and sulfuric acid. In the second step, the 3′-OH group of APS is phosphorylated, with ATP being used again. After transfer of the sulfate residue to OH groups (**c**), adenosine-3′,5′-bisphosphate remains.

4. *S*-adenosyl methionine (SAM)

The coenzyme *tetrahydrofolate* (THF) is the main agent by which C_1 fragments are transferred in the metabolism. THF can bind this type of group in various oxidation states and pass it on (see p. 108). In addition, there is “activated methyl,” in the form of *S*-adenosyl methionine (SAM). SAM is involved in many **methylation reactions**—e. g., in creatine synthesis (see p. 336), the conversion of norepinephrine into epinephrine (see p. 352), the inactivation of norepinephrine by methylation of a phenolic OH group (see p. 316), and in the formation of the active form of the cytostatic drug 6-mercaptopurine (see p. 402).

SAM is derived from degradation of the proteinogenic amino acid **methionine**, to which the adenosyl residue of an ATP molecule is transferred. After release of the activated methyl group, *S*-adenosyl homocysteine (SAH) is left over. This can be converted back into methionine in two further steps. Firstly, cleavage of the adenosine residue gives rise to the non-proteinogenic amino acid **homocysteine**, to which a methyl group is transferred once again with the help of N^5-methyl-THF (see p. 418). Alternatively, homocysteine can also be broken down into propionyl-CoA.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Activated metabolites

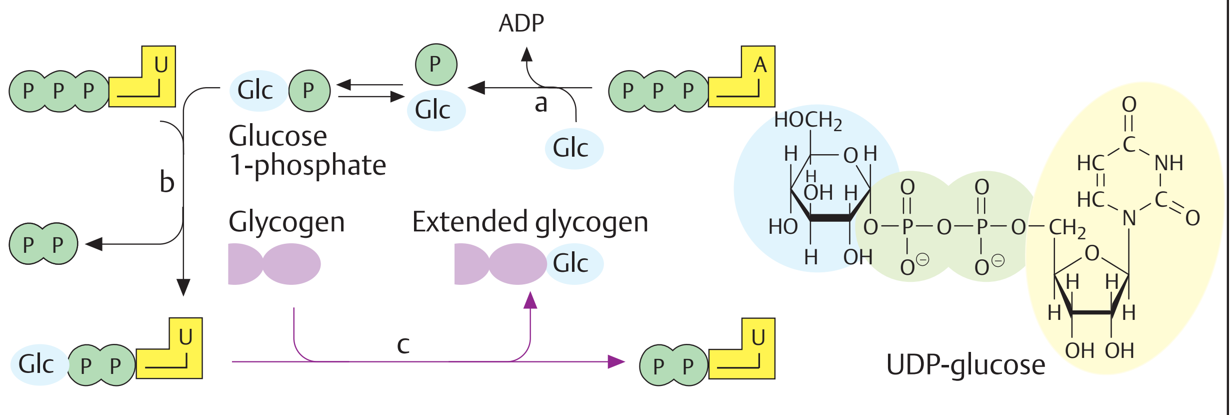
U
Glc
P
ADP
P
Glc
a
A
Glucose 1-phosphate
b
Glc
HOCH$_2$
Glycogen
Extended glycogen
Glc
UDP-glucose
c
P P P
P P
U

1. Uridine diphosphate glucose (UDP-glucose)

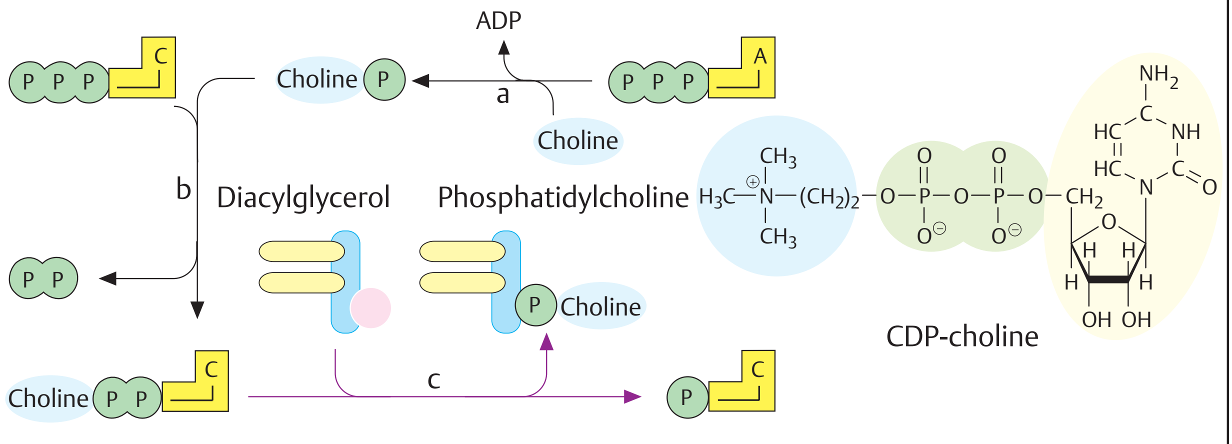
C
ADP
Choline
P
a
A
Choline
b
Diacylglycerol
Phosphatidylcholine
P P
Choline
CDP-choline
c
Choline P P
C
P

2. Cytidine diphosphate choline (CDPcholine)

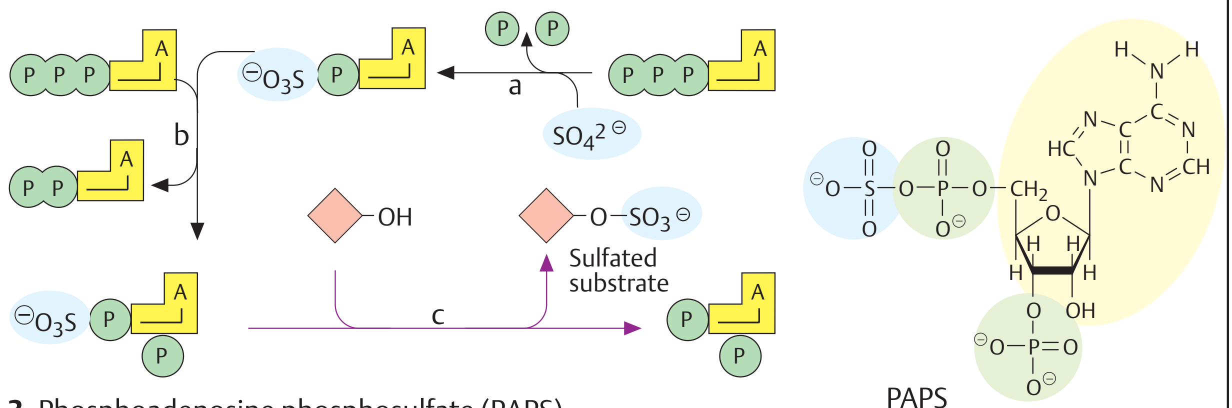
P P
A
$^{\ominus}O_3S$
a
$SO_4^{2\ominus}$
b
OH
O—$SO_3^{\ominus}$
Sulfated substrate
c
PAPS

3. Phosphoadenosine phosphosulfate (PAPS)

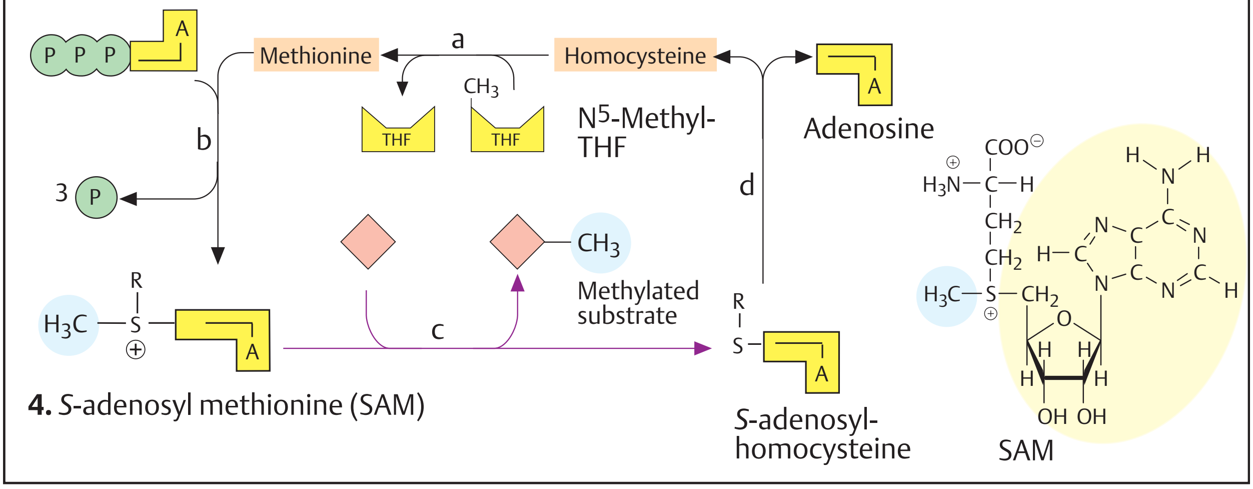
A
Methionine
a
Homocysteine
CH_3
THF
THF
N5-Methyl-THF
Adenosine
b
d
3 P
CH_3
Methylated substrate
R
H_3C
S
c
S-adenosyl-homocysteine
SAM

4. *S*-adenosyl methionine (SAM)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Intermediary metabolism

Hundreds of chemical reactions are constantly taking place in every cell, and taken together these are referred to as the **metabolism**. The chemical compounds involved in this are known as **metabolites**. Outside of the cell, almost all of the chemical changes in metabolites would only take place very slowly and without any specific direction. By contrast, organized sequences of chemical reactions with a high rate of throughput, known as **metabolic pathways**, become possible through the existence of specific **enzymes** (see p. 88).

A. Intermediary metabolism: overview ●

A number of central metabolic pathways are common to most cells and organisms. These pathways, which serve for synthesis, degradation, and interconversion of important metabolites, and also for energy conservation, are referred to as the **intermediary metabolism**.

In order to survive, all cells constantly require organic and inorganic *nutrients*, as well as *chemical energy,* which is mainly derived from ATP (see below). Depending on the way in which these needs are satisfied, organisms can be classified into autotrophic and heterotrophic groups. The **autotrophs**, which include plants and many microorganisms, can synthesize organic molecules from inorganic precursors (CO_2). An autotrophic lifestyle is possible through **photosynthesis**, for example (see p. 128). The **heterotrophs**—e. g., animals and fungi—depend on organic substances supplied in their diet. The schema shown on this page provides an overview of animal metabolism.

The polymeric substances contained in the diet (proteins, carbohydrates, and nucleic acids—top) cannot be used by the organism directly. Digestive processes first have to degrade them to monomers (amino acids, sugars, nucleotides). These are then mostly broken down by **catabolic pathways** (pink arrows) into smaller fragments. The metabolites produced in this way (generally referred to as the “metabolite pool”) are then either used to obtain energy through further catabolic conversion, or are built up again into more complex molecules by **anabolic pathways** (blue arrows). Of the numerous metabolites in the pool, only three particularly important representatives—pyruvate, acetyl-CoA, and glycerol—are shown here. These molecules represent connecting links between the metabolism of proteins, carbohydrates, and lipids. The metabolite pool also includes the intermediates of the tricarboxylic acid cycle (**6**). This cyclic pathway has both catabolic and anabolic functions—i. e., it is **amphibolic** (violet; see p. 138).

Waste products from the degradation of organic substances in animal metabolism include *carbon dioxide* (CO_2), *water* (H_2O), and *ammonia* (NH_3). In mammals, the toxic substance ammonia is incorporated into *urea* and excreted in this form (see p. 182).

The most important form of storage for chemical energy in all cells is **adenosine triphosphate** (ATP, see p. 122). ATP *synthesis* requires energy—i. e., the reaction is *endergonic*. Conversely, cleavage of ATP into ADP and phosphate releases energy. *Exergonic* hydrolysis of ATP, as a result of **energetic coupling** (see p. 16), makes energy-dependent (*endergonic*) processes possible. For example, most anabolic pathways, as well as movement and transport processes, are energy-dependent.

The most important pathway for the synthesis of ATP is **oxidative phosphorylation** (see p. 122). In this process, catabolic pathways first form reduced cofactors ($NADH+H^+$, QH_2, $ETFH_2$). Electrons are then transferred from these compounds to oxygen. This strongly exergonic process is catalyzed by the **respiratory chain** and used indirectly for the ATP synthesis (see p. 140). In *anaerobic conditions*—i. e., in the absence of oxygen—most organisms can fall back on ATP that arises in glycolysis (**3**). This less efficient type of ATP synthesis is referred to as **fermentation** (see p. 146).

While NADH exclusively supplies oxidative phosphorylation, $NADPH+H^+$—a very similar coenzyme—is the reducing agent for anabolic pathways. $NADPH + H^+$ is mainly formed in the pentose phosphate pathway (PPP, **1**; see p. 152).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Intermediary metabolism: overview

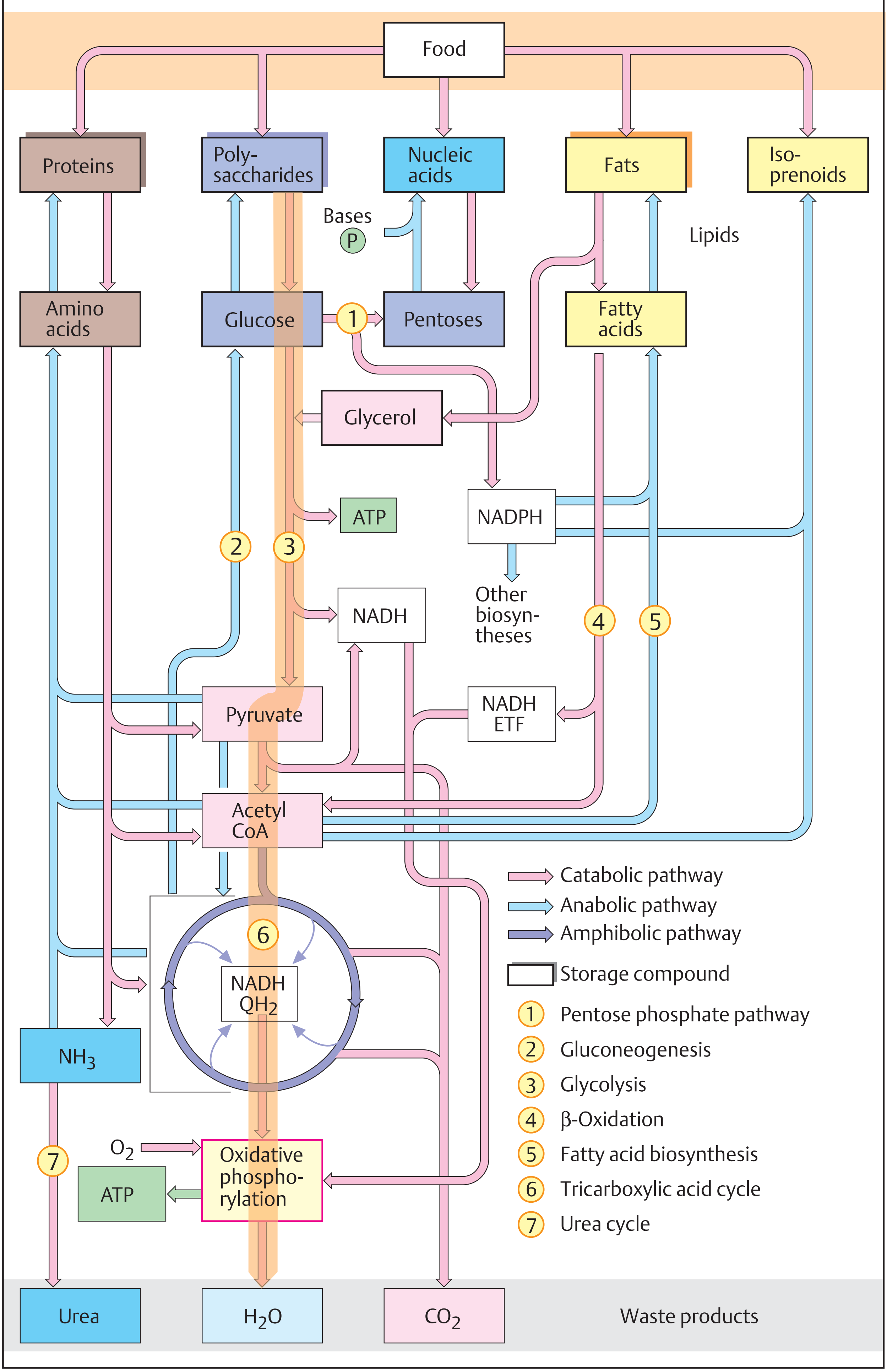
Food
Proteins
Poly-saccharides
Nucleic acids
Fats
Iso-prenoids
Bases
P
Lipids
Amino acids
Glucose
1
Pentoses
Fatty acids
Glycerol
ATP
NADPH
2
3
Other biosyn-theses
NADH
4
5
NADH ETF
Pyruvate
Acetyl CoA
6
NADH QH$_2$
NH$_3$
7
O$_2$
ATP
Oxidative phospho-rylation
Urea
H$_2$O
CO$_2$
Waste products
Catabolic pathway
Anabolic pathway
Amphibolic pathway
Storage compound
1 Pentose phosphate pathway
2 Gluconeogenesis
3 Glycolysis
4 β-Oxidation
5 Fatty acid biosynthesis
6 Tricarboxylic acid cycle
7 Urea cycle

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Regulatory mechanisms

A. Fundamental mechanisms of metabolic regulation ◑

The activities of all metabolic pathways are subject to precise regulation in order to adjust the synthesis and degradation of metabolites to physiological requirements. An overview of the regulatory mechanisms is presented here. Further details are shown on pp. 116ff.

Metabolite flow along a metabolic pathway is mainly determined by the activities of the **enzymes** involved (see p. 88). To regulate the pathway, it is sufficient to change the activity of the enzyme that catalyzes the *slowest* step in the reaction chain. Most metabolic pathways have **key enzymes** of this type on which the regulatory mechanisms operate. The activity of key enzymes is regulated at three independent levels:

Transcriptional control. Here, Biosynthesis of the enzyme protein is influenced at the genetic level (**1**). Interventions in enzyme synthesis mainly affect synthesis of the corresponding mRNA—i. e., *transcription* of the gene coding for the enzyme. The term “transcriptional control” is therefore used (see pp. 118, 244). This mechanism is mediated by *regulatory proteins* (transcription factors) that act directly on DNA. The genes have a special regulatory segment for this purpose, known as the *promoter* region, which contains binding sites (control elements) for regulatory proteins. The activity of these proteins is, in turn, affected by metabolites or hormones. When synthesis of a protein is increased by transcriptional control, the process is referred to as **induction**; when it is reduced or suppressed, it is referred to as **repression**. Induction and repression processes take some time and are therefore not immediately effective.

Interconversion of key enzymes (**2**) takes effect considerably faster than transcriptional control. In this case, the enzyme is already present at its site of effect, but it is initially still inactive. It is only when needed that it is converted into the catalytically active form, after signaling and mediation from second messengers (see p. 120) through an *activating enzyme* (E_1). If the metabolic pathway is no longer required, an *inactivating enzyme* (E_2) returns the key enzyme to its inactive resting state.

Interconversion processes in most cases involve **ATP-dependent phosphorylation** of the enzyme protein by a *protein* kinase or **dephosphorylation** of it by a *protein phosphatase* (see p. 120). The phosphorylated form of the key enzyme is usually the more active one, but the reverse may also occur.

Modulation by ligands. An important variable that regulates flow through a metabolic pathway is **precursor availability** (metabolite A in the case shown here). The availability of precursor A increases along with the activity of the metabolic pathways that form A (**3**) and it decreases with increasing activity of other pathways that also consume A (**4**). Transport from one cell compartment to another can also restrict the availability of A.

Coenzyme availability can also often have a limiting effect (**5**). If the coenzyme is regenerated by a second, independent metabolic pathway, the speed of the second pathway can limit that of the first one. For example, glycolysis and the tricarboxylic acid cycle are mainly regulated by the availability of NAD^+ (see p. 146). Since NAD^+ is regenerated by the respiratory chain, the latter indirectly controls the breakdown of glucose and fatty acids (respiratory control, see p. 144).

Finally, the activity of key enzymes can be regulated by *ligands* (substrates, products, coenzymes, or other effectors), which as *allosteric effectors* do not bind at the active center itself, but at another site in the enzyme, thereby modulating enzyme activity (**6**; see p. 116). Key enzymes are often inhibited by immediate reaction products, by end products of the reaction chain concerned (*“feedback” inhibition*), or by metabolites from completely different metabolic pathways. The precursors for a reaction chain can stimulate their own utilization through enzyme activation.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Fundamental mechanisms of metabolic regulation

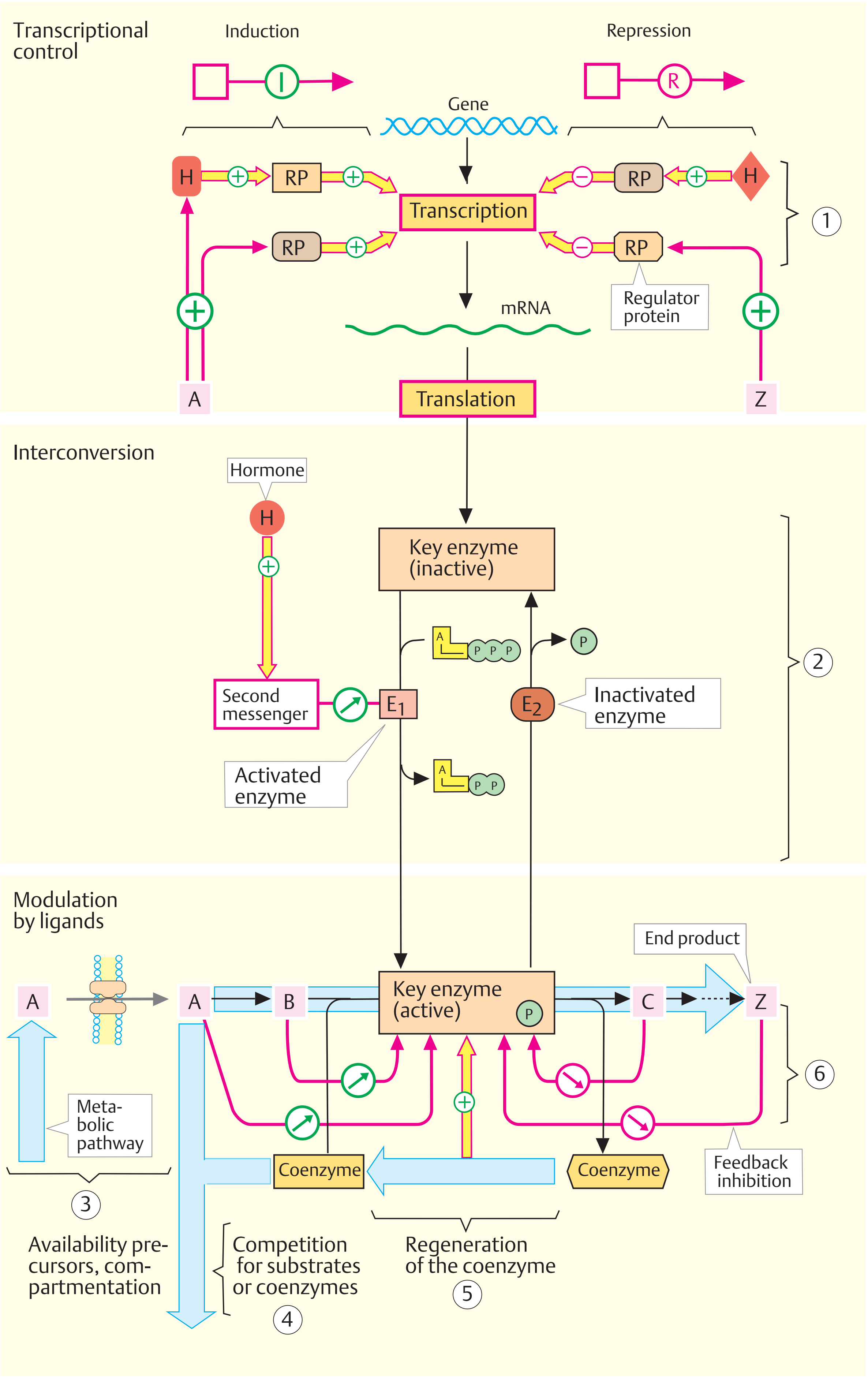
Transcriptional control
Induction
Repression
I
R
Gene
H
RP
+
Transcription
–
H
①
RP
RP
mRNA
Regulator protein
A
Translation
Z
Interconversion
Hormone
H
Key enzyme (inactive)
A
P P P
P
②
Second messenger
E1
E2
Inactivated enzyme
Activated enzyme
A
P P
Modulation by ligands
End product
A
A
B
Key enzyme (active)
P
C
Z
⑥
Meta-bolic pathway
Coenzyme
Coenzyme
Feedback inhibition
③
Availability pre-cursors, com-partmentation
Competition for substrates or coenzymes
④
Regeneration of the coenzyme
⑤

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Allosteric regulation

The regulation of **aspartate carbamoyltransferase** (ACTase), a key enzyme of pyrimidine biosynthesis (see p. 188) is discussed here as an example of allosteric regulation of enzyme activity. Allosteric effects are mediated by the substrate itself or by inhibitors and activators (*allosteric effectors,* see p. 114). The latter bind at special sites outside the active center, producing a conformational change in the enzyme protein and thus indirectly lead to an alteration in its activity.

A. Aspartate carbamoyltransferase: reaction ○

ACTase catalyzes the transfer of a carbamoyl residue from carbamoyl phosphate to the amino group of L-aspartate. The *N*-carbamoyl L-aspartate formed in this way already contains all of the atoms of the later pyrimidine ring (see p. 188). The ACTase of the bacterium *Escherichia coli* is inhibited by cytidine triphosphate (CTP), an end product of the anabolic metabolism of pyrimidines, and is activated by the precursor ATP.

B. Kinetics ◑

In contrast to the kinetics of isosteric (normal) enzymes, allosteric enzymes such as ACTase have **sigmoidal** (S-shaped) **substrate saturation curves** (see p. 92). In allosteric systems, the enzyme's affinity to the substrate is not constant, but depends on the substrate concentration [A]. Instead of the Michaelis constant K_m (see p. 92), the *substrate concentration at half-maximal rate* ($[A]_{0.5}$) is given. The sigmoidal character of the curve is described by the **Hill coefficient h**. In isosteric systems, h = 1, and h increases with increasing sigmoidicity.

Depending on the enzyme, *allosteric effectors* can influence the maximum rate V_{max}, the semi-saturation concentration $[A]_{0.5}$, and the Hill coefficient h. If it is mainly V_{max} that is changed, the term "**V system**" is used. Much more common are "**K systems**", in which allosteric effects only influence $[A]_{0.5}$ and h.

The K type also includes ACTase. The inhibitor CTP in this case leads to *right-shifting* of the curve, with an increase in $[A]_{0.5}$ and h (curve II). By contrast, the activator ATP causes a *left shift;* it reduces both $[A]_{0.5}$ and h (curve III). This type of allosteric effect was first observed in *hemoglobin* (see p. 280), which can be regarded as an "honorary" enzyme.

C. R and T states ○

Allosteric enzymes are almost always *oligomers* with 2–12 subunits. ACTase consists of six catalytic subunits (blue) and six regulatory subunits (yellow). The latter bind the allosteric effectors CTP and ATP. Like hemoglobin, ACTase can also be present in two conformations—the less active **T state** (for "tense") and the more active **R state** (for "relaxed"). Substrates and effectors influence the equilibrium between the two states, and thereby give rise to sigmoidal saturation behavior. With increasing aspartate concentration, the equilibrium is shifted more and more toward the R form. ATP also stabilizes the R conformation by binding to the regulatory subunits. By contrast, binding of CTP to the same sites promotes a transition to the T state. In the case of ACTase, the structural differences between the R and T conformations are particularly dramatic. In T → R conversion, the catalytic subunits separate from one another by 1.2 nm, and the subunits also rotate around the axis of symmetry. The conformations of the subunits themselves change only slightly, however.

D. Structure of a dimer ○

The subunits of ACTase each consist of two *domains*—i. e., independently folded partial structures. The *N*-terminal domain of the regulatory subunit (right) mediates interaction with CTP or ATP (green). A second, Zn^{2+}-containing domain (Zn^{2+} shown in light blue) establishes contact with the neighboring catalytic subunit. Between the two domains of the catalytic subunit lies the active center, which is occupied here by two substrate analogs (red).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Aspartate carbamoyltransferase: reaction

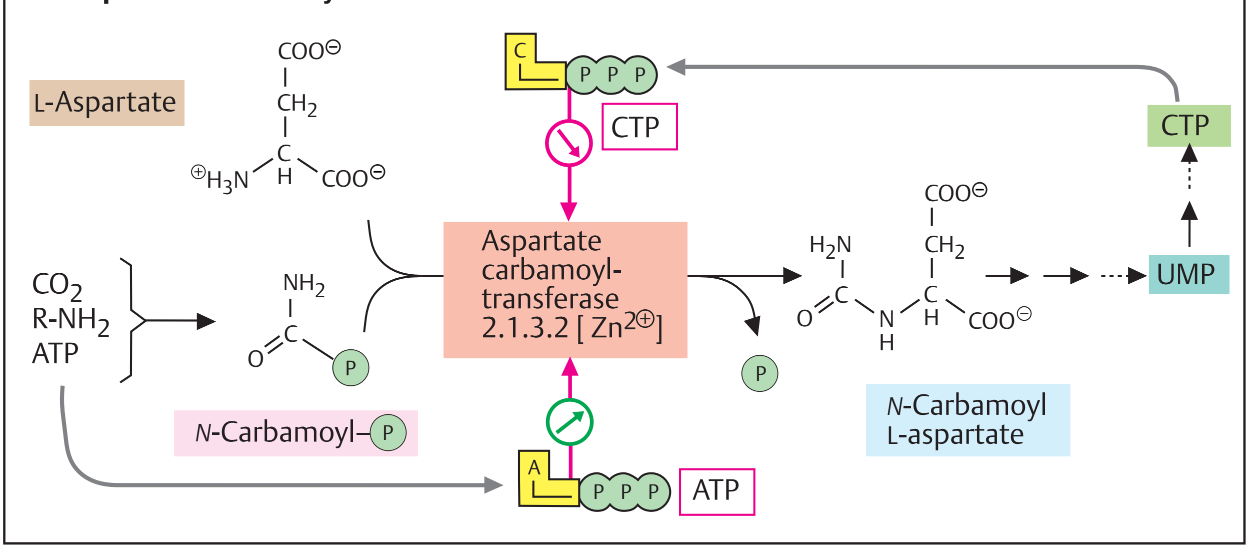
L-Aspartate
COO⊖
CH2
C
⊕H3N
H
COO⊖
C
P P P
CTP
CTP
Aspartate carbamoyl-transferase 2.1.3.2 [Zn2⊕]
COO⊖
H2N
CH2
C
C
O
N
H
H
COO⊖
UMP
CO2
R-NH2
ATP
NH2
C
O
P
P
N-Carbamoyl–P
N-Carbamoyl L-aspartate
A
P P P
ATP

B. Kinetics

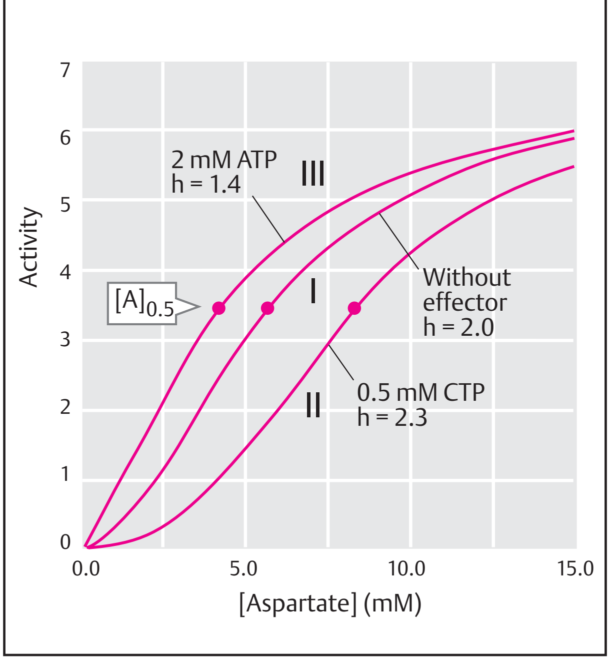

C. R and T conformation

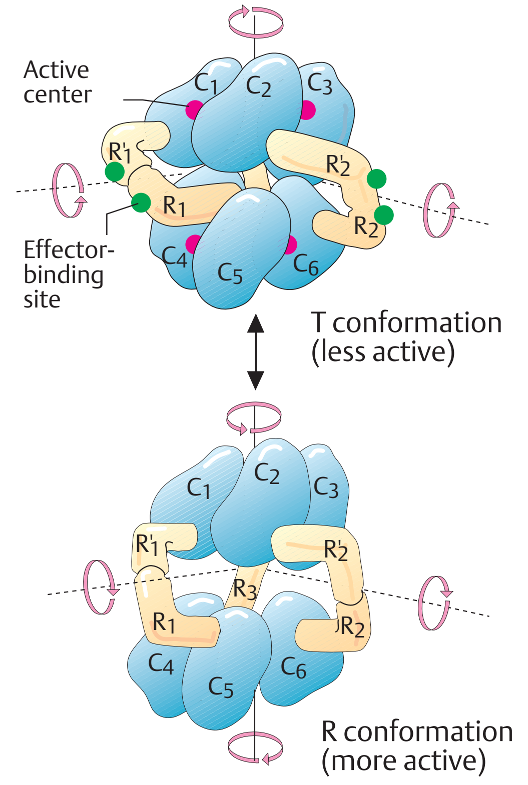
Active center
C1
C2
C3
R'1
R'2
R1
R2
Effector-binding site
C4
C5
C6
T conformation (less active)
C1
C2
C3
R'1
R'2
R3
R1
R2
C4
C5
C6
R conformation (more active)

D. Structure of a dimer

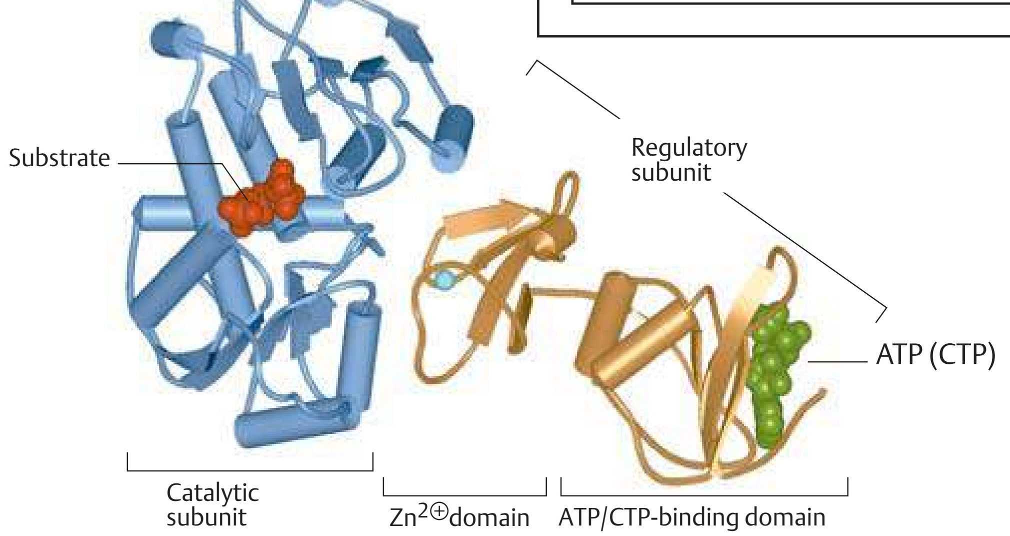
Substrate
Regulatory subunit
ATP (CTP)
Catalytic subunit
Zn2⊕domain
ATP/CTP-binding domain

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Transcription control

A. Functioning of regulatory proteins ◑

Regulatory proteins (transcription factors) are involved in controlling gene expression in all cells. These regulatory proteins bind to specific DNA sequences and thereby activate or inhibit the transcription of genes **(Transcription control)**. The effects of transcription factors are usually reversible and are often controlled by *ligands* or by *interconversion.*

The nomenclature for transcription factors is confusing. Depending on their mode of action, various terms are in use both for the proteins themselves and for the DNA sequences to which they bind. If a factor blocks transcription, it is referred to as a **repressor**; otherwise, it is called an **inducer.** DNA sequences to which regulatory proteins bind are referred to as **control elements.** In prokaryotes, control elements that serve as binding sites for RNA polymerases are called **promoters,** whereas repressor-binding sequences are usually called **operators**. Control elements that bind activating factors are termed **enhancers**, while elements that bind inhibiting factors are known as **silencers**.

The numerous regulatory proteins that are known can be classified into four different groups (**1**–**4**), based on their mechanisms of action. **Negative gene regulation**—i. e., switching off of the gene concerned—is carried out by **repressors**. Some repressors only bind to DNA (**1a**) in the absence of specific ligands (L). In this case, the complex between the repressor and the ligand loses its ability to bind to the DNA, and the promoter region becomes accesible for binding of RNA polymerase (**1b**). It is often the free repressor that does not bind to the DNA, so that transcription is only blocked in the presence of the ligand (**2a**, **2b**). A distinction between two different types of **positive gene regulation** can be made in the same way. If it is only the free inducer that binds, then transcription is inhibited by the appropriate ligand (**3**). Conversely, many **inducers** only become active when they have bound a ligand (**4**). This group includes the receptors for steroid hormones, for example (see p. 378).

B. Lactose operon ○

The well-investigated **lactose operon** of the bacterium *Escherichial coli* can be used here as an example of transcriptional control. The *lac* operon is a DNA sequence that is simultaneously subject to negative and positive control. The operon contains the *structural genes* for three proteins that are required for the utilization of lactose (one transporter and two enzymes), as well as *control elements* that serve to regulate the operon.

Since lactose is converted to glucose in the cell, there is no point in expressing the genes if glucose is already available. And indeed, the genes are in fact only transcribed when *glucose is absent* and *lactose is present* (**3**). This is achieved by interaction between two regulatory proteins. In the absence of lactose, the ***lac* repressor** blocks the promoter region (**2**). When lactose is available, it is converted into *allolactose*, which binds to the repressor and thereby detaches it from the operator (**3**). However, this is still not sufficient for the transcription of the structural genes. For binding of the RNA polymerase to take place, an *inducer*—the **catabolite activator protein (CAP)**—is required, which only binds to the DNA when it is present as a complex with 3,5′-cyclo-AMP (cAMP; see p. 386). cAMP, a signal for nutrient deficiency, is only formed by *E. coli* in the *absence* of glucose.

The interaction between the CAP–cAMP complex and DNA is shown in Fig. **4**. Each subunit of the dimeric inducer (yellow or orange) binds one molecule of cAMP (red). Contact with the DNA (blue) is mediated by two “recognition helices” that interact with the major groove of the DNA. The bending of the DNA strand caused by CAP has functional significance.

Transcription control is much more complex in eukaryotes (see p. 244). The number of transcription factors involved is larger, and in addition the gene activity is influenced by the state of the chromatin (see p. 238).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Functions of regulatory proteins

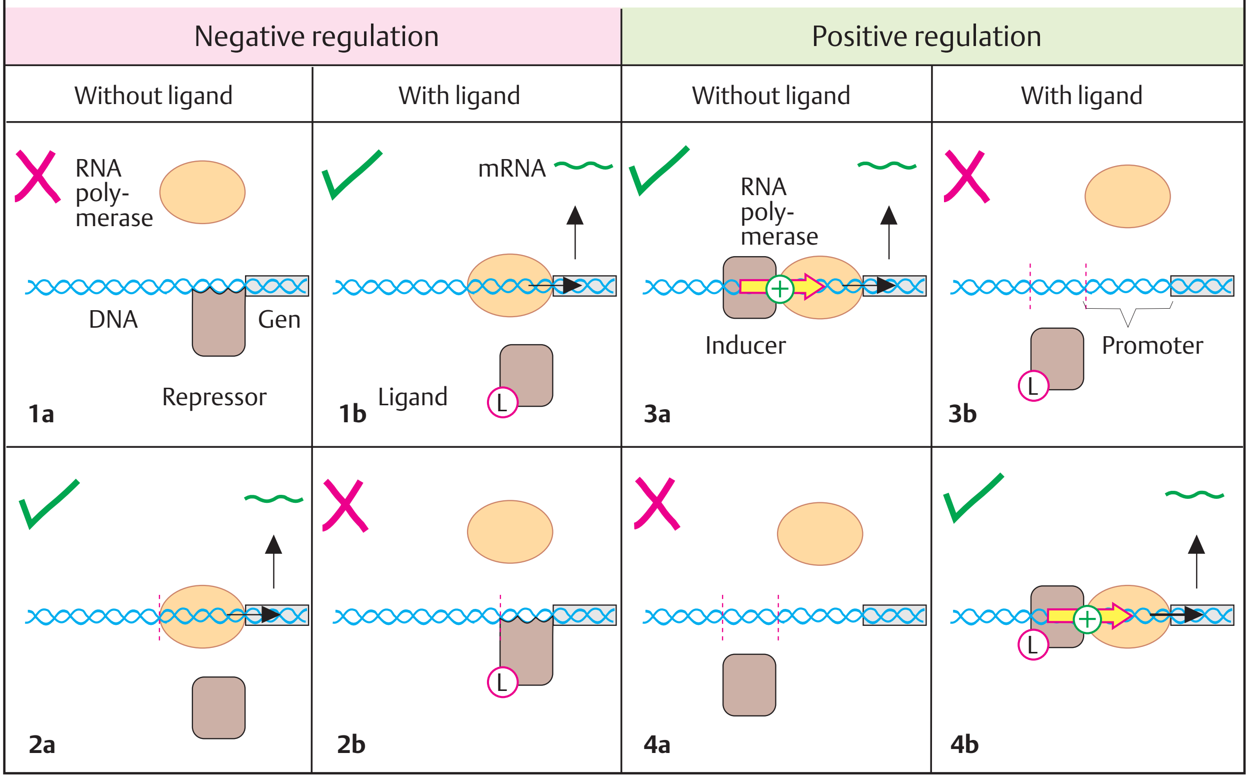
Negative regulation
Positive regulation
Without ligand
With ligand
Without ligand
With ligand
✗
RNA poly-merase
DNA
Gen
Repressor
1a
✓
mRNA
Ligand
L
1b
✓
RNA poly-merase
Inducer
3a
✗
Promoter
L
3b
✓
2a
✗
L
2b
✗
4a
✓
L
4b

B. Lactose operon

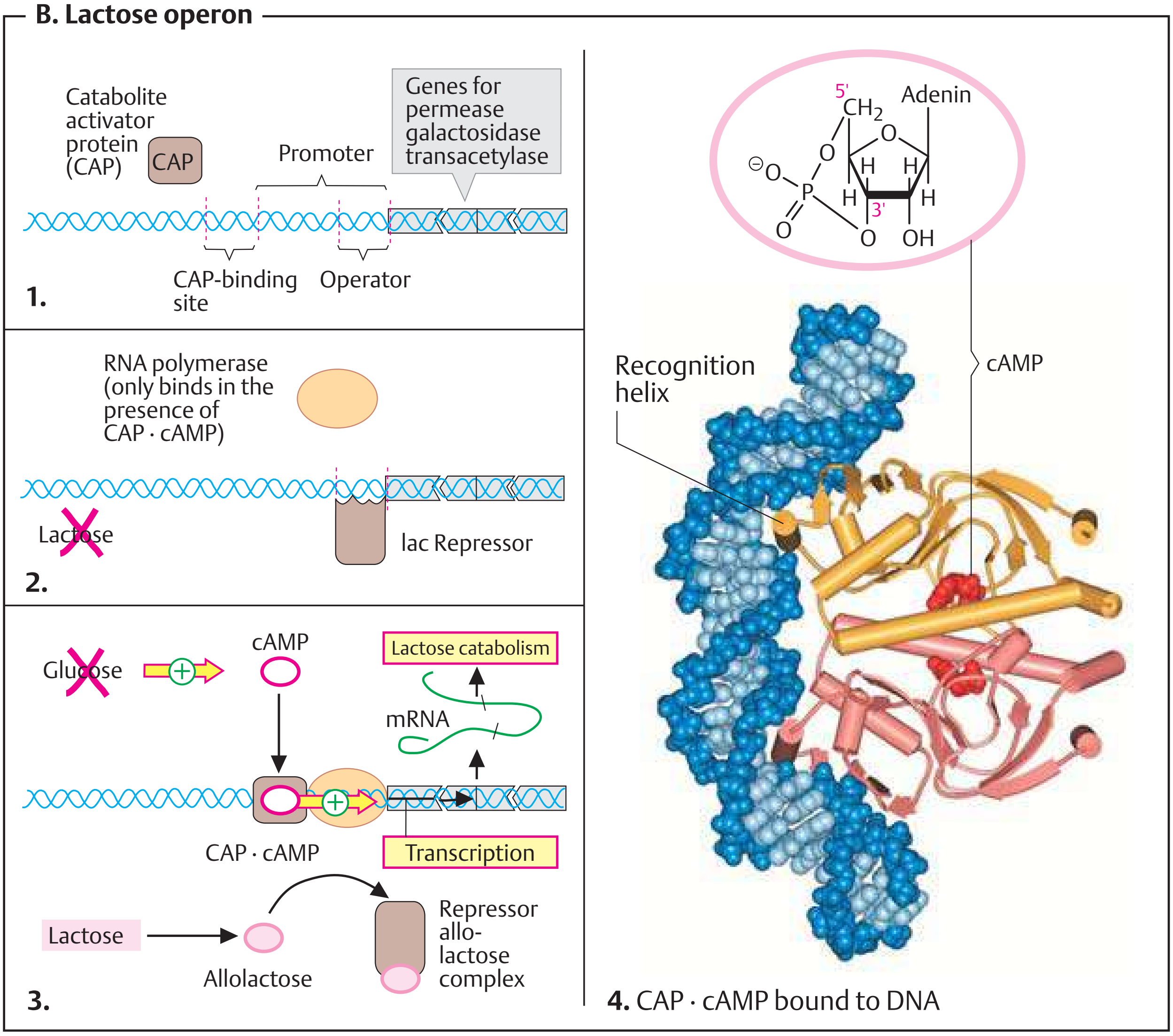
Catabolite activator protein (CAP)
CAP
Promoter
Genes for permease galactosidase transacetylase
CAP-binding site
Operator
1.
RNA polymerase (only binds in the presence of CAP · cAMP)
Lactose
lac Repressor
2.
Glucose
cAMP
Lactose catabolism
mRNA
CAP · cAMP
Transcription
Lactose
Allolactose
Repressor allo-lactose complex
3.
5'
CH2
Adenin
O
⊖O
P
H
3'
OH
cAMP
Recognition helix
4. CAP · cAMP bound to DNA

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Hormonal control

In higher organisms, metabolic and other processes (growth, differentiation, control of the internal environment) are controlled by **hormones** (see pp. 370ff.)

A. Principles of hormone action ●

Depending on the type of hormone, hormone signals are transmitted to the target cells in different ways. Apolar (lipophilic) hormones penetrate the cell and act in the cell nucleus, while polar (hydrophilic) hormones act on the external cell membrane.

Lipophilic hormones, which include the steroid hormones, thyroxine, and retinoic acid, bind to a specific *receptor protein* inside their target cells. The complex formed by the hormone and the receptor then influences *transcription* of specific genes in the cell nucleus (see pp. 118, 244). The group of **hydrophilic hormones** (see p. 380) consists of hormones derived from amino acids, as well as peptide hormones and proteohormones. Their *receptors* are located in the plasma membrane. Binding of the hormone to this type of receptor triggers a signal that is transmitted to the interior of the cell, where it controls the processes that allow the hormone signal to take effect (**signal transduction**; see pp. 384ff.)

B. Hormonal regulation of glucose metabolism in the liver ◑

The liver plays a major role in glucose homeostasis in the organism (see p. 310). If glucose deficiency arises, the liver releases glucose into the blood, and when blood sugar levels are high, it takes glucose up from the blood and converts it into different metabolites. Several hormones from both groups are involved in controlling these processes. A very simplified version of the way in which they work is presented here. **Glycogen** is the form in which glucose is stored in the liver and muscles. The rate of glycogen synthesis is determined by *glycogen synthase* (bottom right), while its breakdown is catalyzed by *glycogen phosphorylase* (bottom left).

Regulation by interconversion (bottom). If the blood glucose level falls, the peptide hormone **glucagon** is released. This activates glycogen breakdown, releasing glucose, and at the same time inhibits glycogen synthesis. Glucagon binds to receptors in the plasma membrane (bottom left) and, with mediation by a G-protein (see p. 386), activates the enzyme *adenylate cyclase*, which forms the *second messenger* 3,5′-cyclo-AMP (**cAMP**) from ATP. cAMP binds to another enzyme, *protein kinase A* (PK-A), and activates it. PK-A has several points of attack. Through *phosphorylation,* it converts the active form of *glycogen synthase* into the inactive form, thereby terminating the synthesis of glycogen. Secondly, it activates another protein kinase (not shown), which ultimately converts the inactive form of *glycogen phosphorylase* into the active form through phosphorylation. The active phosphorylase releases glucose 1-phosphate from glycogen, which after conversion into glucose 6-phosphate supplies free glucose. In addition, via an inhibitor (I) of protein phosphatase (PP), active PK-A inhibits inactivation of glycogen phosphorylase. When the cAMP level falls again, *phosphoprotein phosphatases* become active, which dephosphorylate the various phosphoproteins in the cascade described, and thereby arrest glycogen breakdown and re-start glycogen synthesis. Activation and inactivation of proteins through phosphorylation or dephosphorylation is referred to as **interconversion**.

In contrast to glucagon, the peptide hormone **insulin** (see p. 76) increases glycogen synthesis and inhibits glycogen breakdown. Via several intermediates, it inhibits protein kinase GSK-3 (bottom right; for details, see p. 388) and thereby prevents inactivation of glycogen synthase. In addition, insulin reduces the cAMP level by activating *cAMP phosphodiesterase* (PDE).

Regulation by transcriptional control (top). If the liver's glycogen reserves have been exhausted, the steroid hormone **cortisol** maintains glucose release by initiating the conversion of amino acids into glucose (*gluconeogenesis;* see p. 154). In the cell nucleus, the complex of cortisol and its receptor (see p. 378) binds to the promoter regions of various key enzymes of gluconeogenesis and leads to their transcription. The active enzymes are produced through translation of the mRNA formed. Control of the transcription of the gluconeogenesis enzyme *PEP carboxykinase* is discussed on p. 244.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Principles of hormone action

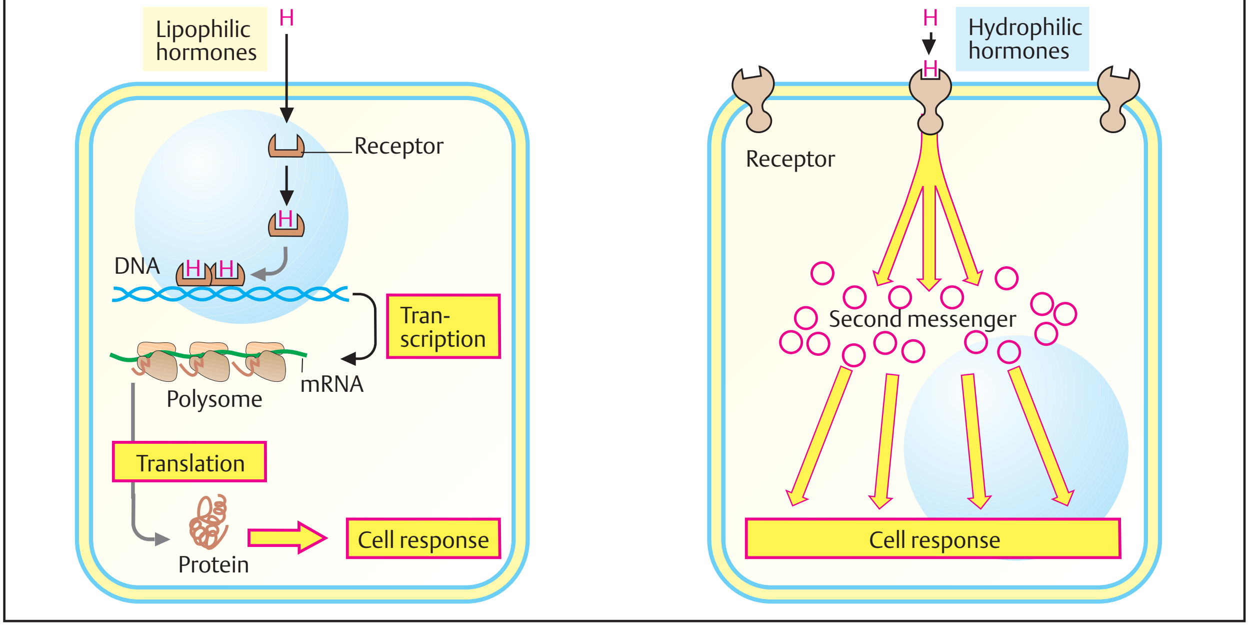
Lipophilic hormones
H
Receptor
H
DNA
H H
Tran-scription
mRNA
Polysome
Translation
Protein
Cell response
H
Hydrophilic hormones
H
Receptor
Second messenger
Cell response

B. Hormonal regulation of glucose metabolism in the liver

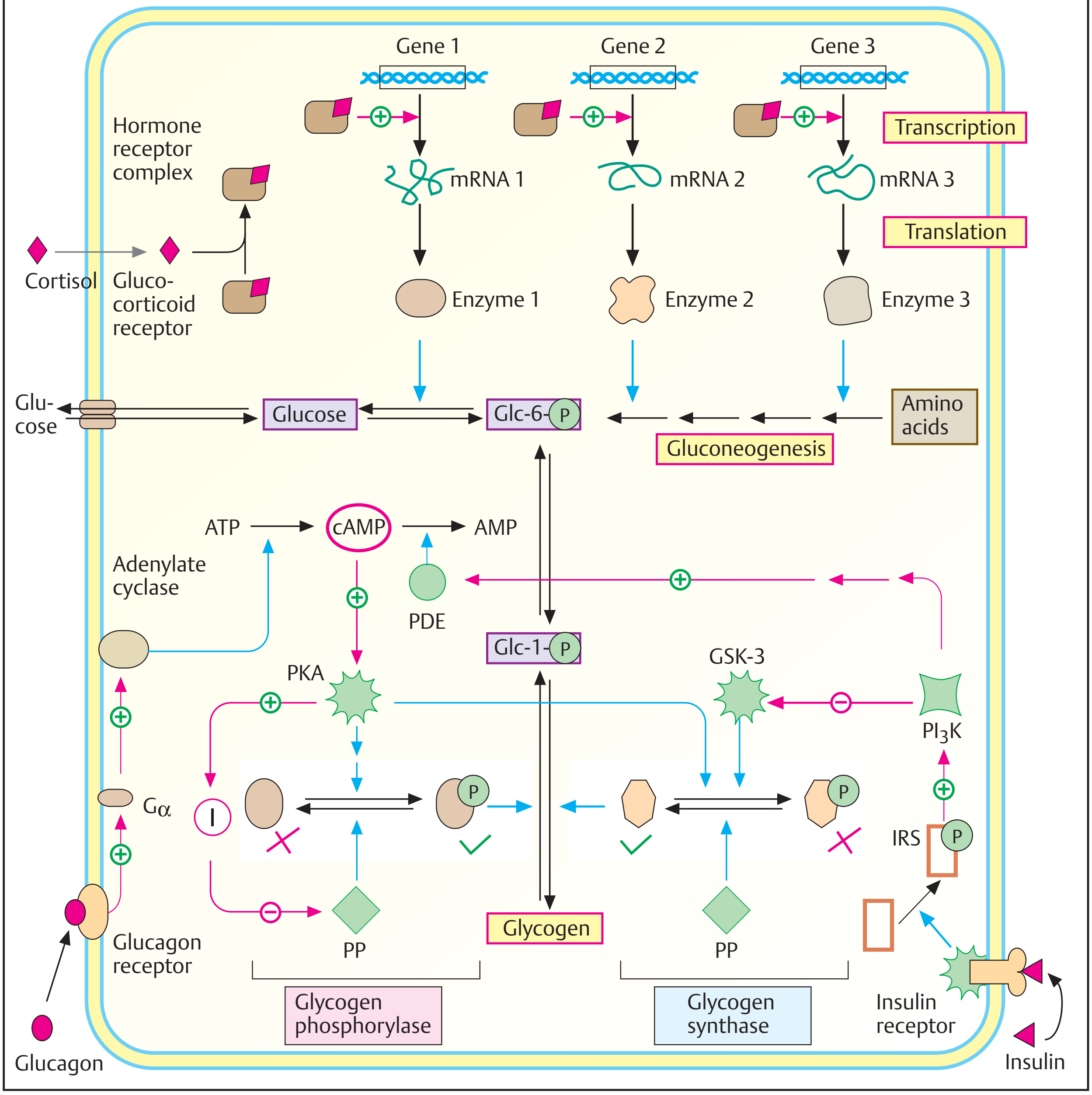
Gene 1
Gene 2
Gene 3
Transcription
Hormone receptor complex
mRNA 1
mRNA 2
mRNA 3
Translation
Cortisol
Gluco-corticoid receptor
Enzyme 1
Enzyme 2
Enzyme 3
Glu-cose
Glucose
Glc-6-P
Amino acids
Gluconeogenesis
ATP
cAMP
AMP
Adenylate cyclase
PDE
Glc-1-P
PKA
GSK-3
PI3K
Gα
I
P
P
IRS
P
PP
Glycogen
PP
Glucagon receptor
Glycogen phosphorylase
Glycogen synthase
Insulin receptor
Glucagon
Insulin

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

ATP

The nucleotide coenzyme **adenosine triphosphate** (ATP) is the most important **form of chemical energy** in all cells. Cleavage of ATP is strongly exergonic. The energy this provides (ΔG; see p. 16) is used to drive endergonic processes (such as biosynthesis and movement and transport processes) through *energetic coupling* (see p. 124). The other *nucleoside triphosphate coenzymes* (GTP, CTP, and UTP) have similar chemical properties to ATP, but they are used for different tasks in metabolism (see p. 110).

A. ATP: structure ◑

In ATP, a chain of three phosphate residues is linked to the 5′-OH group of the nucleoside adenosine (see p. 80). These phosphate residues are termed α, β, and γ. The α phosphate is bound to ribose by a *phosphoric acid ester bond*. The linkages *between* the three phosphate residues, on the other hand, involve much more unstable *phosphoric acid anhydride bonds*. The active coenzyme is in fact generally a complex of ATP with an Mg^{2+} ion, which is coordinatively bound to the α and β phosphates (Mg^{2+} ATP^{4-}). However, the term "ATP" is usually used for the sake of simplicity.

B. Hydrolysis energies ◑

The formula for phosphate residues shown in Fig. **A**, with single and double bonds, is not an accurate representation of the actual charge distribution. In ATP, the oxygen atoms of all three phosphate residues have similarly strong negative charges (orange), while the phosphorus atoms represent centers of positive charge. One of the reasons for the instability of phosphoric anhydride bonds is the *repulsion between these negatively charged oxygen atoms*, which is partly relieved by cleavage of a phosphate residue. In addition, the free phosphate anion formed by hydrolysis of ATP is *better hydrated* and *more strongly resonance-stabilized* than the corresponding residue in ATP. This also contributes to the strongly exergonic character of ATP hydrolysis.

In *standard conditions,* the change in free enthalpy $\Delta G^{0\prime}$ (see p. 18) that occurs in the hydrolysis of phosphoric acid anhydride bonds amounts to –30 to –35 kJ mol^{-1} at pH 7. The particular anhydride bond of ATP that is cleaved only has a minor influence on $\Delta G^{0\prime}$ (**1**–**2**). Even the hydrolysis of diphosphate (also known as pyrophosphate; **4**) still yields more than –30 kJ mol^{-1}. By contrast, cleavage of the ester bond between ribose and phosphate only provides –9 kJ mol^{-1} (**3**).

In the cell, the ΔG of ATP hydrolysis is substantially larger, because the concentrations of ATP, ADP and P_i are much lower than in standard conditions and there is an excess of ATP over ADP (see p. 18). The pH value and Mg^{2+} concentration also affect the value of ΔG. The *physiological energy yield* of ATP hydrolysis to ADP and anorganic phosphate (P_i) is probably around –50 kJ mol^{-1}.

C. Types of ATP formation ●

Only a few compounds contain phosphate residues with a group transfer potential (see p. 18) that is high enough to transfer them to ADP and thus allow **ATP synthesis**. Processes that raise anorganic phosphate to this type of high potential are called **substrate level phosphorylations** (see p. 124). Reactions of this type take place in glycolysis (see p. 150) and in the tricarboxylic acid cycle (see p. 136). Another "energy-rich" phosphate compound is *creatine phosphate*, which is formed from ATP in muscle and can regenerate ATP as needed (see p. 336).

Most cellular ATP does not arise in the way described above (i. e., by transfer of phosphate residues from organic molecules to ADP), but rather by **oxidative phosphorylation**. This process takes place in mitochondria (or as light-driven phosphorylation in chloroplasts) and is energetically coupled to a proton gradient over a membrane. These H^+ gradients are established by electron transport chains and are used by the enzyme *ATP synthase* as a source of energy for direct linking of anorganic phosphate to ADP. In contrast to substrate level phosphorylation, oxidative phosphorylation requires the presence of oxygen (i. e., *aerobic* conditions).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. ATP: structure

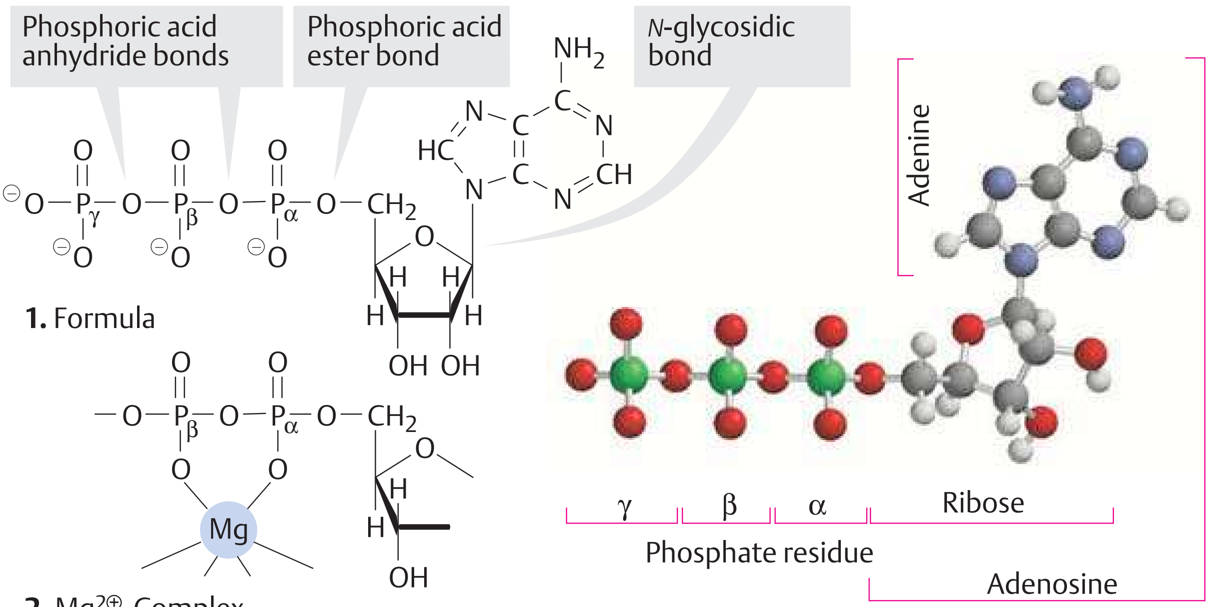
Phosphoric acid anhydride bonds
Phosphoric acid ester bond
N-glycosidic bond
Pγ
Pβ
Pα
CH₂
NH₂
OH OH
Mg
Adenine
γ
β
α
Ribose
Phosphate residue
Adenosine

1. Formula

2. $Mg^{2\oplus}$-Complex

B. Hydrolysis energies

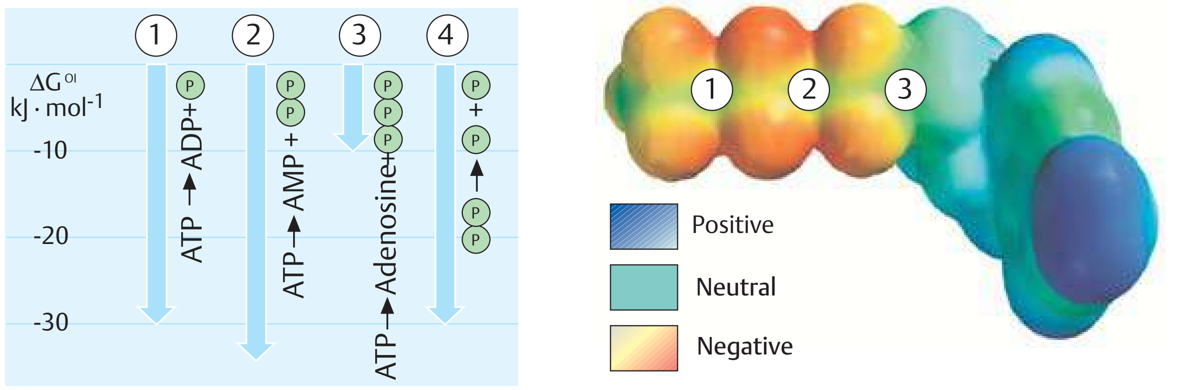

1. Hydrolysis energies

2. ATP: charge density

C. Types of ATP formation

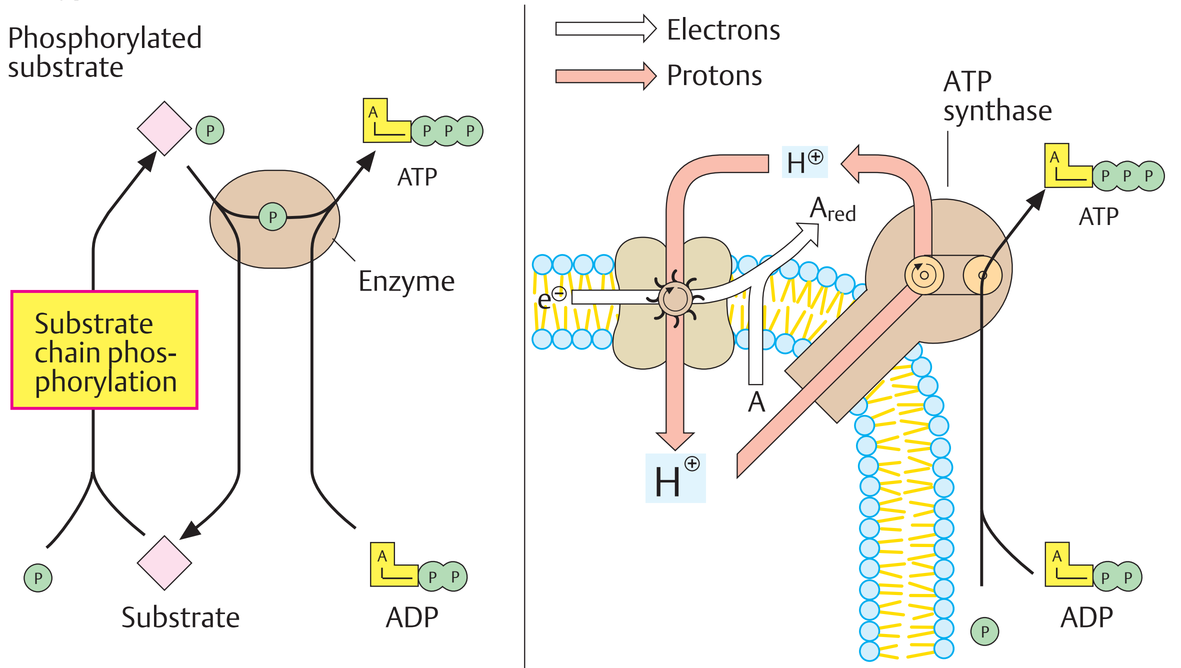
Phosphorylated substrate
P
ATP
Enzyme
Substrate chain phosphorylation
Substrate
ADP
Electrons
Protons
ATP synthase
H⊕
A_red
e⊖
A
H⊕
ATP
ADP
P

1. Phosphate transfer

2. Oxidative phosphorylation

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Energetic coupling

The cell stores chemical energy in the form of "energy-rich" metabolites. The most important metabolite of this type is adenosine triphosphate (**ATP**), which drives a large number of energy-dependent reactions via **energetic coupling** (see p. 16).

A. Energetic coupling ◑

The change in free enthalpy ΔG^0 during hydrolysis (see p. 18) has been arbitrarily selected as a measure of the group transfer potential of "energy-rich" compounds. However, this does not mean that ATP is in fact hydrolyzed in energetically coupled reactions. If ATP hydrolysis and an endergonic process were simply allowed to run alongside each other, the hydrolysis would only produce heat, without influencing the endergonic process. For coupling, the two reactions have to be linked in such a way that a *common intermediate* arises. This connection is illustrated here using the example of the **glutamine synthetase reaction**.

Direct transfer of NH_3 to glutamate is endergonic ($\Delta G^{0\prime} = +14\ \text{kJ mol}^{-1}$; see p. 18), and can therefore not take place. In the cell, the reaction is divided into two exergonic steps. First, the γ-phophate residue is transferred from ATP to glutamate. This gives rise to an "energy-rich" *mixed acid anhydride*. In the second step, the phosphate residue from the intermediate is substituted by NH_3, and glutamine and free phosphate are produced. The energy balance of the reaction as a whole ($\Delta G^{0\prime} = -17\ \text{kJ mol}^{-1}$) is the sum of the changes in free enthalpy of direct glutamine synthesis ($\Delta G^{0\prime} = 14\ \text{kJ mol}^{-1}$) plus ATP hydrolysis ($\Delta G^{0\prime} = -31\ \text{kJ mol}^{-1}$), although ATP has not been hydrolyzed at all.

B. Substrate-level phosphorylation ◑

As mentioned earlier (see p. 122), there are a few metabolites that transfer phosphate to ADP in an exergonic reaction and can therefore form ATP. In ATP synthesis, anorganic phosphate or phosphate bound in an ester-like fashion is transferred to bonds with a high phosphate transfer potential. Reactions of this type are termed *"substrate-level phosphorylations,"* as they represent individual steps within metabolic pathways.

In the **glyceraldehyde 3-phosphate dehydrogenation** reaction, a step involved in glycolysis (**1**; see also **C**), the aldehyde group in glyceraldehyde 3-phosphate is oxidized into a carboxyl group. During the reaction, an anorganic phosphate is also introduced into the product, producing a mixed acid anhydride—1,3-bisphosphoglycerate. **Phosphopyruvate hydratase** ("enolase", **2**) catalyzes the elimination of water from 2-phosphoglycerate. In the *enol phosphate* formed (phospho*enol* pyruvate), the phosphate residue—in contrast to 2-phosphoglycerate—is at an extremely high potential ($\Delta G^{0\prime}$ of hydrolysis: $-62\ \text{kJ mol}^{-1}$). A third reaction of this type is the formation of succinyl phosphate, which occurs in the tricarboxylic acid cycle as an individual step in the **succinyl CoA ligase** reaction. Here again, anorganic phosphate is introduced into a mixed acid anhydride bond to be transferred from there to GDP. Succinyl phosphate is only an intermediate here, and is not released by the enzyme.

In the literature, the term "substrate level phosphorylation" is used inconsistently. Some authors use it to refer to reactions in which *anorganic* phosphate is raised to a high potential, while others use it for the subsequent reactions, in which ATP or GTP is formed from the energy-rich intermediates.

C. Glyceraldehyde-3-phosphate dehydrogenase ○

The reaction catalyzed during glycolysis by *glyceraldehyde-3-phosphate dehydrogenase* (GADPH) is shown here in detail. Initially, the SH group of a cysteine residue of the enzyme is added to the carbonyl group of glyceraldehyde 3-phosphate (**a**). This intermediate is oxidized by NAD^+ into an "energy-rich" thioester (**b**). In the third step (**c**), anorganic phosphate displaces the thiol, and the mixed anhydride *1,3-bisphosphoglycerate* arises. In this bond, the phosphate residue is at a high enough potential for it to be transferred to ADP in the next step (not shown; see p. 150).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Energetic coupling

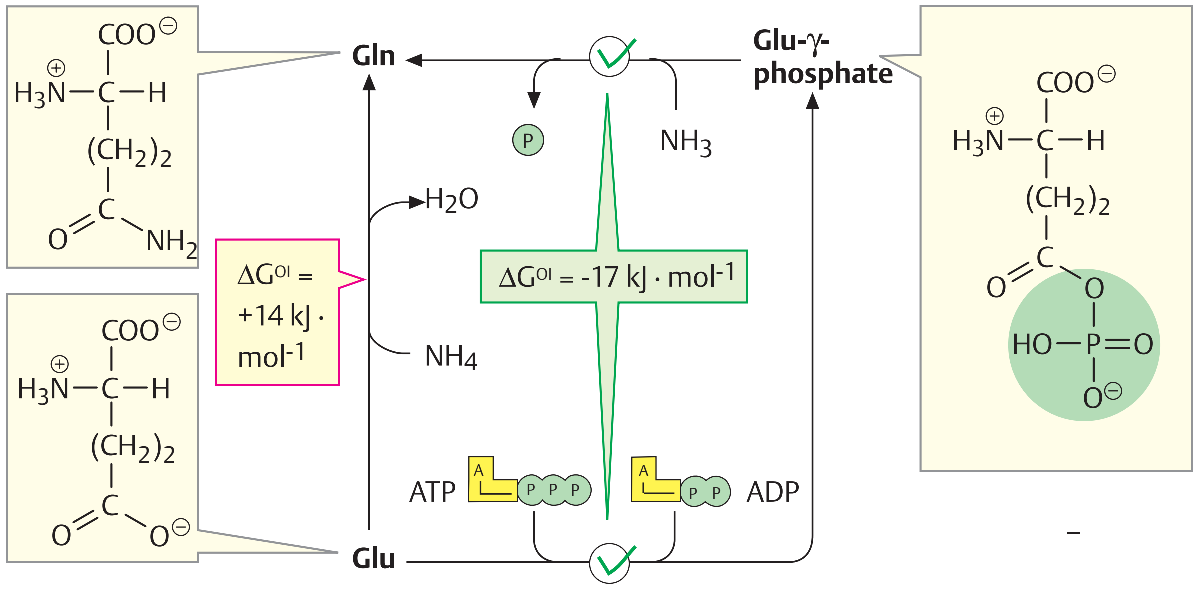
COO⊖
⊕
H_3N—C—H
$(CH_2)_2$
C
O
NH_2
Gln
P
NH_3
Glu-γ-phosphate
H_2O
ΔG°' = +14 kJ · mol-1
ΔG°' = -17 kJ · mol-1
NH_4
COO⊖
⊕
H_3N—C—H
$(CH_2)_2$
C
O
O⊖
ATP
A
P P P
A
P P
ADP
Glu
COO⊖
⊕
H_3N—C—H
$(CH_2)_2$
C
O
O
HO—P=O
O⊖

1. Glutamine synthetase reaction

Reaction 1:	Glutamate	+	NH_3	+14 kJ · mol−1	Glutamine	+	H_2O		
Reaction 2:	ATP	+	H_2O	− 31 kJ · mol−1	ADP	+	Ⓟ		
Total:	Glutamate	+	NH_3 + ATP	− 17 kJ · mol−1	Glutamine	+	ADP	+	Ⓟ

2. Energy balance

B. Substrate level phosphorylation

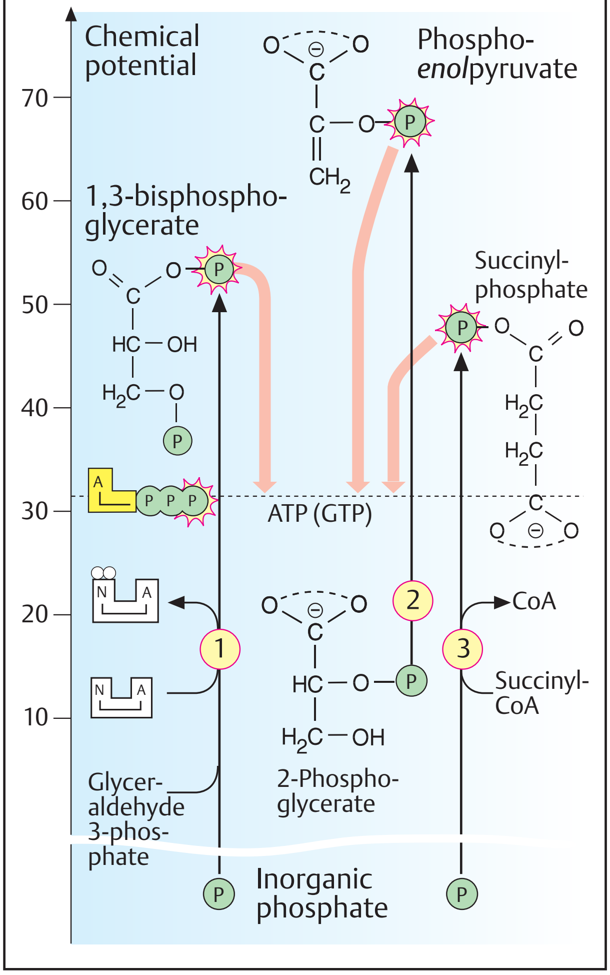
Chemical potential
70
60
50
40
30
20
10
Phospho-*enol*pyruvate
O ⊖ O
C
C—O—P
CH_2
1,3-bisphospho-glycerate
O
C
O—P
HC—OH
H_2C—O
P
Succinyl-phosphate
P
O
C
O
H_2C
H_2C
C
O ⊖ O
A
P P P
ATP (GTP)
N A
N A
1
2
3
CoA
O ⊖ O
C
HC—O—P
H_2C—OH
2-Phospho-glycerate
Succinyl-CoA
Glycer-aldehyde 3-phos-phate
P
Inorganic phosphate
P

C. Glyceraldehyde 3-phosphate dehydrogenase

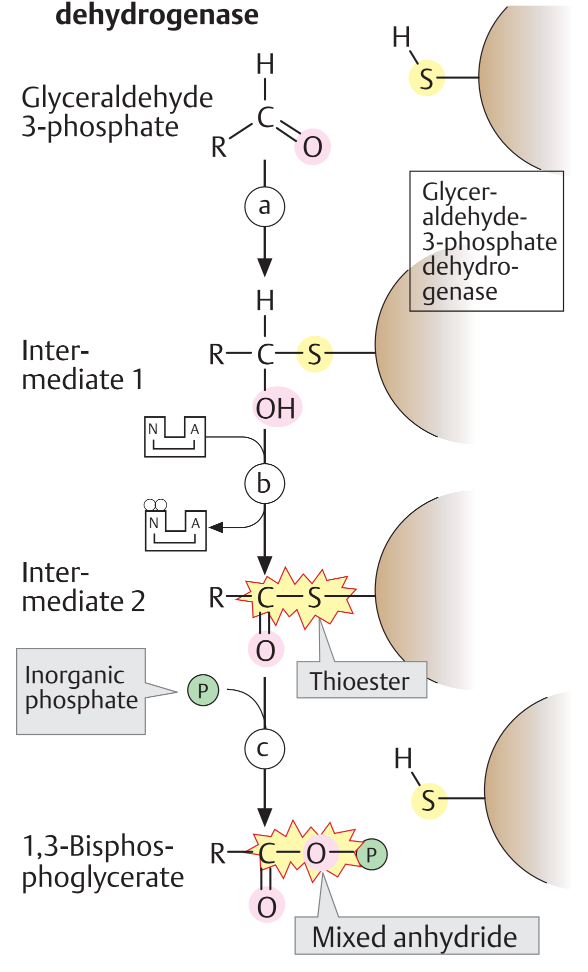
H
S
Glyceraldehyde 3-phosphate
H
C
R
O
a
Glycer-aldehyde-3-phosphate dehydro-genase
H
Inter-mediate 1
R—C—S
OH
N A
b
N A
Inter-mediate 2
R—C—S
O
Thioester
Inorganic phosphate
P
c
H
S
1,3-Bisphos-phoglycerate
R—C—O—P
O
Mixed anhydride

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Energy conservation at membranes

Metabolic energy can be stored not only in the form of “energy-rich” bonds (see p. 122), but also by separating electric charges from each other using an insulating layer to prevent them from redistributing. In the field of technology, this type of system would be called a *condenser.* Using the same principle, energy is also stored (“conserved”) at cell membranes. The membrane functions as an insulator; electrically charged atoms and molecules (*ions*) function as charges.

A. Electrochemical gradient ◑

Although artificial lipid membranes are almost impermeable to ions, biological membranes contain **ion channels** that selectively allow individual ion types to pass through (see p. 222). Whether an ion can cross this type of membrane, and if so in which direction, depends on the **electrochemical gradient**—i. e., on the concentrations of the ion on each side of the membrane (the *concentration gradient*) and on the *difference* in the electrical potential between the interior and exterior, the **membrane potential**.

The membrane potential of resting cells (**resting potential**; see p. 350) is –0.05 to –0.09 V—i. e., there is an excess negative charge on the inner side of the plasma membrane. The main contributors to the resting potential are the two cations Na$^+$ and K$^+$, as well as Cl$^-$ and organic anions (**1**). Data on the concentrations of these ions outside and inside animal cells, and permeability coefficients, are shown in the table (**2**).

The behavior of an ion type is described quantitatively by the **Nernst equation** (**3**). $\Delta\psi_G$ is the membrane potential (in volts, V) at which there is *no* net transport of the ion concerned across the membrane (**equilibrium potential**). The factor R T/F n has a value of 0.026 V for monovalent ions at 25 °C. Thus, for K$^+$, the table (**2**) gives an equilibrium potential of ca. –0.09 V—i. e., a value more or less the same as that of the resting potential. By contrast, for Na$^+$ ions, $\Delta\psi_G$ is much higher than the resting potential, at +0.07 V. Na$^+$ ions therefore immediately flow into the cell when Na$^+$ channels open (see p. 350). The disequilibrium between Na$^+$ and K$^+$ ions is constantly maintained by the enzyme *Na$^+$/K$^+$-ATPase,* which consumes ATP.

B. Proton motive force ◑

Hydronium ions (“H$^+$ ions”) can also develop electrochemical gradients. Such a **proton gradient** plays a decisive part in cellular ATP synthesis (see p. 142). As usual, the energy content of the gradient depends on the concentration gradients—i. e., on the **pH difference** эpH between the two sides of the membrane. In addition, the **membrane potential** $\Delta\psi$ also makes a contribution. Together, these two values give the **proton motive force** Δp, a measure for the work that the H$^+$ gradient can do. The proton gradient across the inner mitochondrial membrane thus delivers approximately 24 kJ per mol H$^+$.

C. Energy conservation in proton gradients ◑

Proton gradients can be built up in various ways. A very unusual type is represented by **bacteriorhodopsin** (**1**), a *light-driven proton pump* that various bacteria use to produce energy. As with rhodopsin in the eye, the light-sensitive component used here is covalently bound retinal (see p. 358). In photosynthesis (see p. 130), reduced *plastoquinone* (QH$_2$) transports protons, as well as electrons, through the membrane (**Q cycle**, **2**). The formation of the proton gradient by the **respiratory chain** is also coupled to redox processes (see p. 140). In complex III, a Q cycle is responsible for proton translocation (not shown). In *cytochrome c oxidase* (complex IV, **3**), H$^+$ transport is coupled to electron flow from cytochrome *c* to O$_2$.

In each of these cases, the H$^+$ gradient is utilized by an **ATP synthase** (**4**) to form ATP. ATP synthases consist of two components—a proton channel (F$_0$) and an inwardly directed protein complex (F$_1$), which conserves the energy of back-flowing protons through ATP synthesis (see p. 142).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Electrochemical gradient

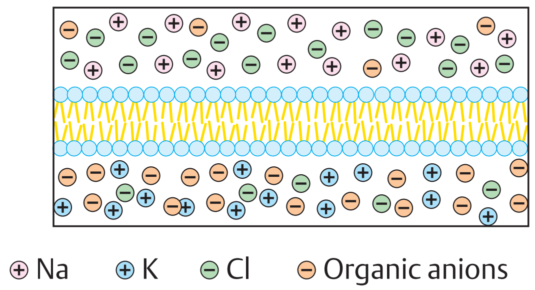
⊕ Na
⊕ K
⊖ Cl
⊖ Organic anions

1. Cause

Ion	Concentrations Cytoplasm (mM)	Extracelluar space (mM)	Permeability coefficient (cm · s^{-1} · 10^9)
K^{+}	100	5	500
Na^{+}	15	150	5
Ca^{2+}	0.0002	2	
Cl^{-}	13	150	10
Organic anions	138	34	0

2. Concentrations

$$\Delta\Psi_G = \frac{R \cdot T}{F \cdot n} \cdot \ln \frac{C\ \text{outside}}{C\ \text{inside}}$$

R = gas constant

T = temperature (K)

n = Ion charge

F = Faraday constant

3. Nernst equation

B. Proton motive force

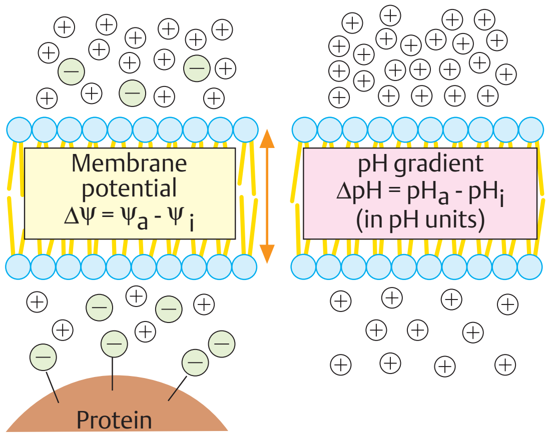
Membrane potential
$\Delta\psi = \psi_a - \psi_i$
pH gradient
$\Delta pH = pH_a - pH_i$
(in pH units)
Protein

Proton motive force

🟩 ≈ 0.06 V

$$\Delta p = \Delta\psi - \frac{2.3 \cdot R \cdot T}{F} \cdot \Delta pH$$

$$\Delta G_{H^{+}_{a} \leftrightarrow H^{+}_{i}} = -F \cdot \Delta p$$

C. Energy conservation in proton gradients

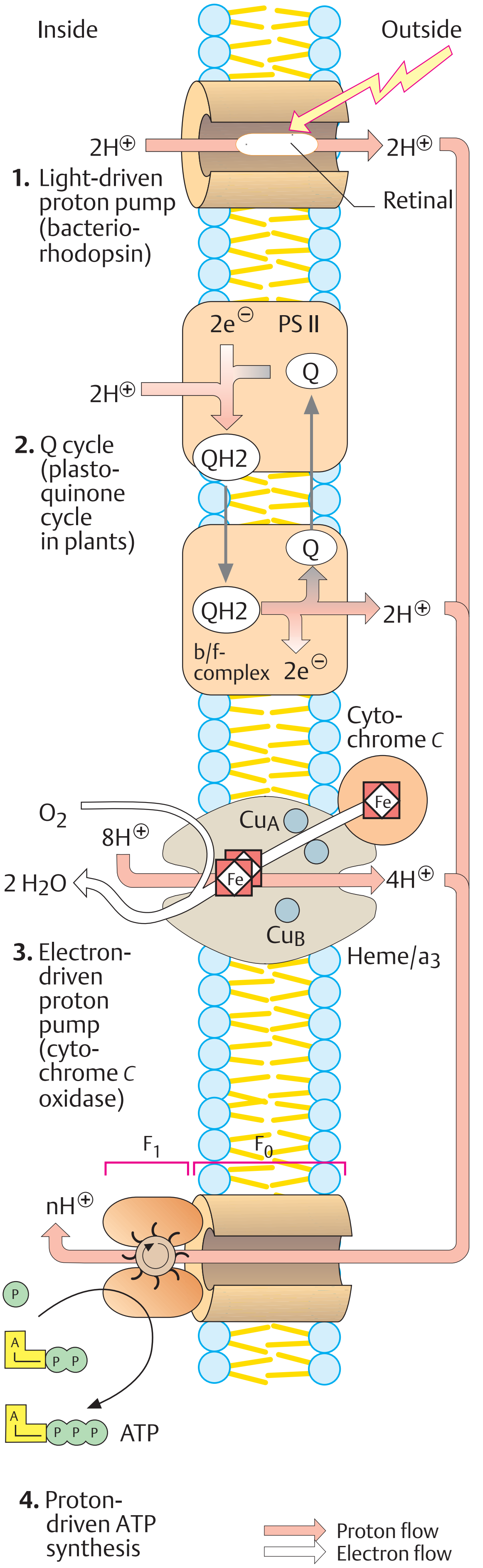
Inside
Outside
$2H^{+}$
$2H^{+}$
Retinal
1. Light-driven proton pump (bacterio-rhodopsin)
$2e^{-}$
PS II
Q
$2H^{+}$
QH2
2. Q cycle (plasto-quinone cycle in plants)
Q
QH2
$2H^{+}$
b/f-complex
$2e^{-}$
Cyto-chrome c
O_2
$8H^{+}$
CuA
Fe
Fe
$2\ H_2O$
$4H^{+}$
CuB
Heme/a3
3. Electron-driven proton pump (cyto-chrome c oxidase)
F_1
F_0
nH^{+}
P
A P P
A P P P
ATP
4. Proton-driven ATP synthesis
Proton flow
Electron flow

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Photosynthesis: light reactions

Sunlight is the most important source of energy for nearly all living organisms. With the help of **photosynthesis,** light energy is used to produce organic substances from CO_2 and water. This property of *phototrophic organisms* (plants, algae, and some bacteria) is exploited by *heterotrophic* organisms (e. g., animals), which are dependent on a supply of organic substances in their diet (see p. 112). The atmospheric *oxygen* that is vital to higher organisms is also derived from photosynthesis.

A. Photosynthesis: overview ◑

The chemical balance of photosynthesis is simple. Six molecules of CO_2 are used to form one hexose molecule (right). The hydrogen required for this reduction process is taken from water, and molecular oxygen is formed as a by-product (left). Light energy is required, since water is a very poor reducing agent and is therefore not capable of reducing CO_2.

In the light-dependent part of photosynthesis—the "**light reactions** "—H_2O molecules are split into protons, electrons, and oxygen atoms. The electrons undergo *excitation* by light energy and are raised to an energy level that is high enough to reduce $NADP^+$. The $NADPH+H^+$ formed in this way, in contrast to H_2O, is capable of "fixing" CO_2 reductively—i. e., of incorporating it into organic bonds. Another product of the light reactions is ATP, which is also required for CO_2 fixation. If $NADPH+H^+$, ATP, and the appropriate enzymes are available, CO_2 fixation can also take place in darkness. This process is therefore known as the "**dark reaction.**"

The excitation of electrons to form NADPH is a complex photochemical process that involves **chlorophyll**, a tetrapyrrole dye containing Mg^{2+} that bears an extra phytol *residue* (see p. 132).

B. Light reactions ○

In green algae and higher plants, photosynthesis occurs in **chloroplasts.** These are organelles, which—like mitochondria—are surrounded by two membranes and contain their own DNA. In their interior, the *stroma*, *thylakoids* or flattened membrane sacs are stacked on top of each other to form *grana*. The inside of the thylakoid is referred to as the *lumen*. The light reactions are catalyzed by enzymes located in the thylakoid membrane, whereas the dark reactions take place in the stroma.

As in the respiratory chain (see p. 140), the light reactions cause electrons to pass from one redox system to the next in an **electron transport chain.** However, the *direction of transport* is opposite to that found in the respiratory chain. In the respiratory chain, electrons flow from $NADH+H^+$ to O_2, with the production of water and energy.

In photosynthesis, electrons are taken up from water and transferred to $NADP^+$, with an *expenditure of energy.* Photosynthetic electron transport is therefore energetically "uphill work." To make this possible, the transport is stimulated at two points by the *absorption of light energy.* This occurs through two **photosystems**—protein complexes that contain large numbers of chlorophyll molecules and other pigments (see p. 132). Another component of the transport chain is the **cytochrome b/f complex,** an aggregate of integral membrane proteins that includes two cytochromes (b_{563} and f). **Plastoquinone**, which is comparable to ubiquinone, and two soluble proteins, the coppercontaining **plastocyanin** and **ferredoxin**, function as mobile electron carriers. At the end of the chain, there is an enzyme that transfers the electrons to $NADP^+$.

Because photosystem II and the cytochrome b/f complex release protons from reduced plastoquinone into the lumen (via a Q cycle), photosynthetic electron transport establishes an **electrochemical gradient** across the thylakoid membrane (see p. 126), which is used for ATP synthesis by an *ATP synthase.* ATP and $NADPH+H^+$, which are both needed for the dark reactions, are formed in the stroma.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Photosynthesis: overview

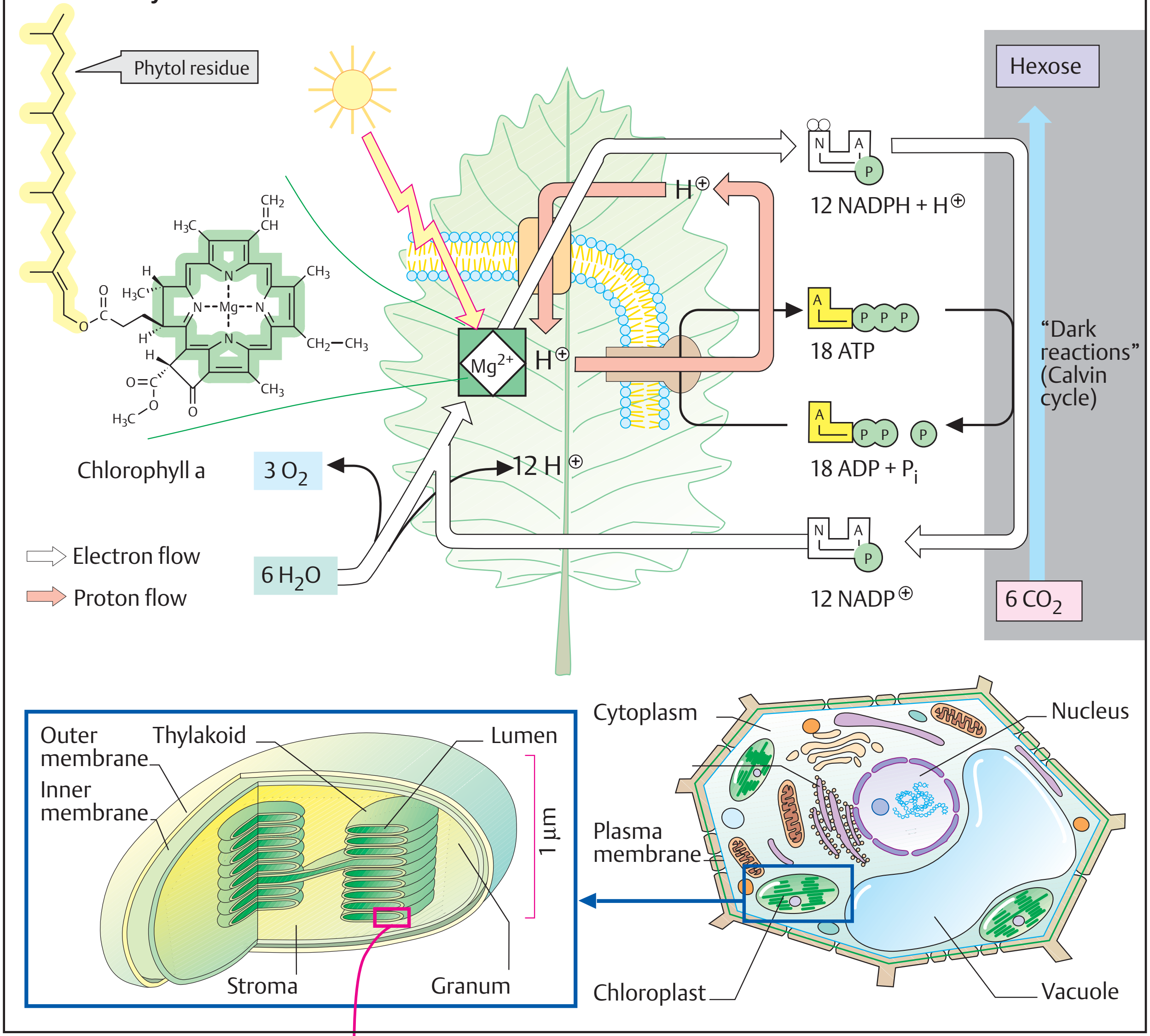
Phytol residue
Hexose
$H^{\oplus}$
N A P
12 NADPH + $H^{\oplus}$
CH_2
CH
H_3C
CH_3
N--Mg--N
CH_2-CH_3
O=C
O
H_3C
A P P P
18 ATP
"Dark reactions" (Calvin cycle)
Mg^{2+}
$H^{\oplus}$
Chlorophyll a
$3\ O_2$
12 $H^{\oplus}$
A P P P
18 ADP + P_i
Electron flow
Proton flow
$6\ H_2O$
12 $NADP^{\oplus}$
$6\ CO_2$
Outer membrane
Thylakoid
Lumen
Inner membrane
1 µm
Stroma
Granum
Cytoplasm
Nucleus
Plasma membrane
Chloroplast
Vacuole

B. Light reactions

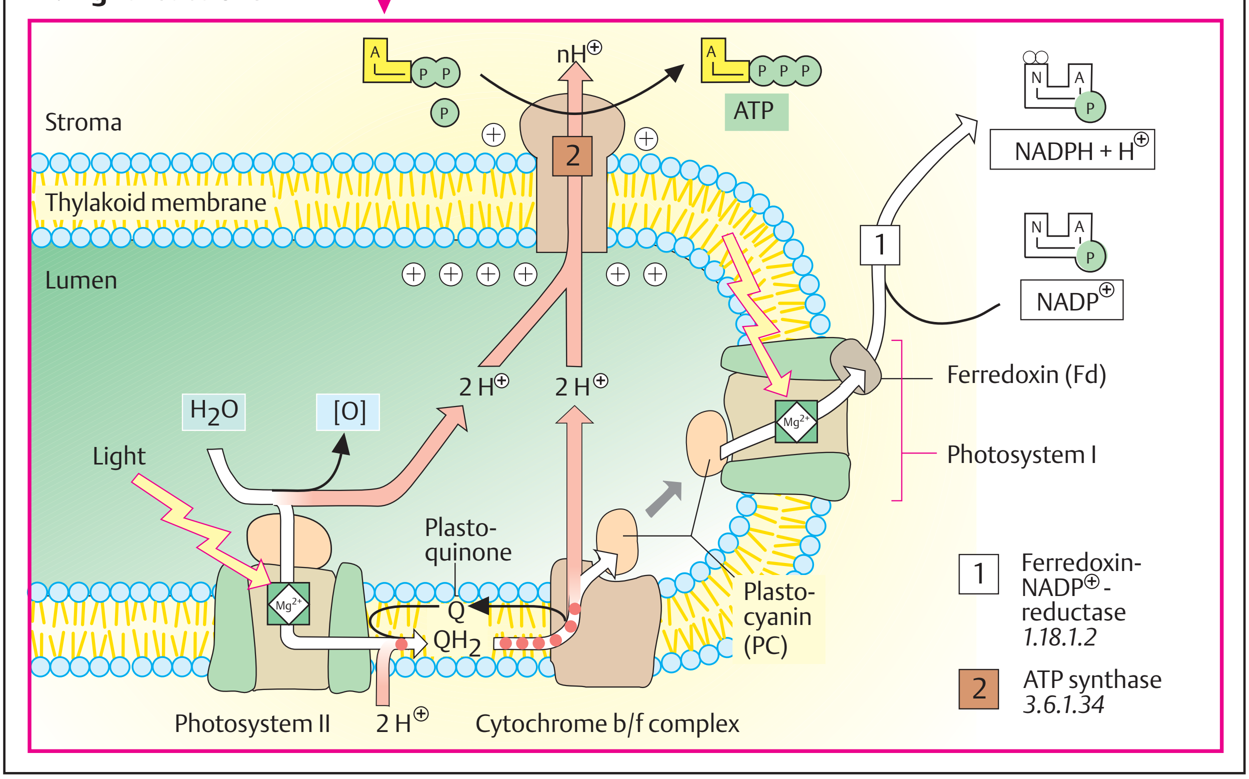
Stroma
A P P
P
n$H^{\oplus}$
A P P P
ATP
2
Thylakoid membrane
Lumen
N A P
NADPH + $H^{\oplus}$
$NADP^{\oplus}$
1
Ferredoxin (Fd)
Photosystem I
H_2O
[O]
2 $H^{\oplus}$
2 $H^{\oplus}$
Light
Mg^{2+}
Plasto-quinone
Q
QH_2
Plasto-cyanin (PC)
Photosystem II
2 $H^{\oplus}$
Cytochrome b/f complex
1 Ferredoxin-$NADP^{\oplus}$-reductase 1.18.1.2
2 ATP synthase 3.6.1.34

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Photosynthesis: dark reactions

The “light reactions” in photosynthesis bring about two strongly endergonic reactions—the reduction of $NADP^+$ to $NADPH+H^+$ and ATP synthesis (see p. 122). The chemical energy needed for this is produced from radiant energy by two photosystems.

A. Photosystem II ○

The photosynthetic electron transport chain in plants starts in **photosystem II** (PS II; see p. 128). PS II consists of numerous protein subunits (brown) that contain bound **pigments**—i. e., dye molecules that are involved in the absorption and transfer of light energy.

The schematic overview of PS II presented here (**1**) only shows the important pigments. These include a special chlorophyll molecule, the *reaction center* P_{680}; a neighboring Mg^{2+} free chlorophyll (*pheophytin*); and two bound *plastoquinones* (Q_A and Q_B). A third quinone (Q_P) is not linked to PS II, but belongs to the plastoquinone pool. The white arrows indicate the direction of electron flow from water to Q_P. Only about 1% of the chlorophyll molecules in PS II are *directly* involved in photochemical excitation (see p. 128). Most of them are found, along with other pigments, in what are known as light-harvesting or antenna complexes (green). The energy of light quanta striking these can be passed on to the reaction center, where it can be utilized.

In Fig. **2**, photosynthetic electron transport in PS II is separated into the individual steps involved. Light energy from the light-harvesting complexes (**a**) raises an electron of the chlorophyll in the reaction center to an excited *“singlet state.”* The excited electron is immediately passed on to the neighboring pheophytin. This leaves behind an “electron gap” in the reaction center—i. e., a positively charged P_{680} radical (**b**). This gap is now filled by an electron removed from an H_2O molecule by the *water-splitting enzyme* (b). The excited electron passes on from the pheophytin via Q_A to Q_B, converting the latter into a *semiquinone radical* (**c**). Q_B is then reduced to *hydroquinone* by a second excited electron, and is then exchanged for an oxidized quinone (Q_P) from the plastoquinone pool. Further transport of electrons from the plastoquinone pool takes place as described on the preceding page and shown in **B**.

B. Redox series ○

It can be seen from the *normal potentials* E^0 (see p. 18) of the most important redox systems involved in the light reactions why two excitation processes are needed in order to transfer electrons from H_2O to $NADP^+$. After excitation in PS II, E^0 *rises* from around –1 V back to positive values in plastocyanin (PC)—i. e., the energy of the electrons has to be increased again in PS I. If there is no $NADP^+$ available, photosynthetic electron transport can still be used for ATP synthesis. During *cyclic photophosphorylation,* electrons return from ferredoxin (Fd) via the plastoquinone pool to the b/f complex. This type of electron transport does not produce any NADPH, but does lead to the formation of an H^+ gradient and thus to ATP synthesis.

C. Calvin cycle ○

The synthesis of hexoses from CO_2 is only shown in a very simplified form here; a complete reaction scheme is given on p. 407. The actual **CO_2 fixation**—i. e., the incorporation of CO_2 into an organic compound—is catalyzed by *ribulose bisphosphate carboxylase/oxygenase* (“rubisco”). Rubisco, the most abundant enzyme on Earth, converts ribulose 1,5-bisphosphate, CO_2 and water into *two molecules* of 3-phosphoglycerate. These are then converted, via 1,3-bisphosphoglycerate and 3-phosphoglycerate, into glyceraldehyde 3-phosphate (glyceral 3-phosphate). In this way, 12 glyceraldehyde 3-phosphates are synthesized from six CO_2. Two molecules of this intermediate are used by gluconeogenesis reactions to synthesize *glucose 6-phosphate* (bottom right). From the remaining 10 molecules, six molecules of *ribulose 1,5-bisphosphate* are regenerated, and the cycle then starts over again. In the Calvin cycle, ATP is required for phosphorylation of 3-phosphoglycerate and ribulose 5-phosphate. $NADPH+H^+$, the second product of the light reaction, is consumed in the reduction of 1,3-bisphosphoglycerate to glyceraldehyde 3-phosphate.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Photosystem II

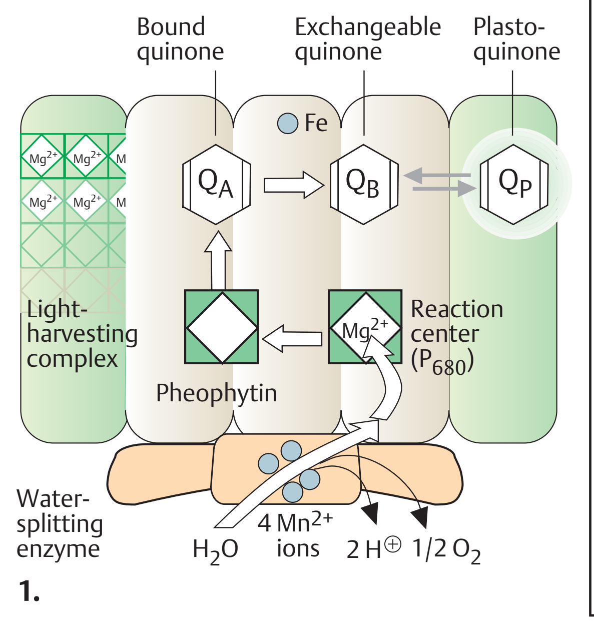
Bound quinone
Exchangeable quinone
Plasto-quinone
Fe
Q_A
Q_B
Q_P
Mg^{2+}
Light-harvesting complex
Pheophytin
Reaction center (P_{680})
Water-splitting enzyme
H_2O
4 Mn^{2+} ions
2 H⊕
1/2 O_2
1.

B. Redox series

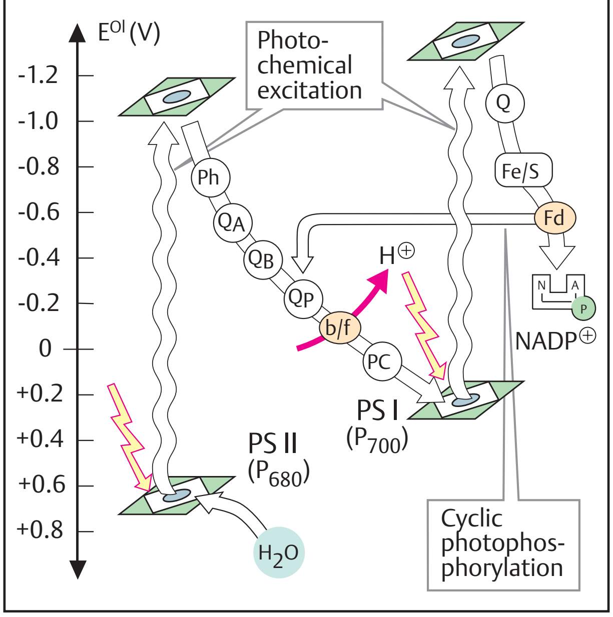
$E^{0\prime}$ (V)
-1.2
-1.0
-0.8
-0.6
-0.4
-0.2
0
+0.2
+0.4
+0.6
+0.8
Photo-chemical excitation
Ph
QA
QB
QP
b/f
PC
H⊕
Q
Fe/S
Fd
NADP⊕
PS II (P_{680})
PS I (P_{700})
H_2O
Cyclic photophos-phorylation

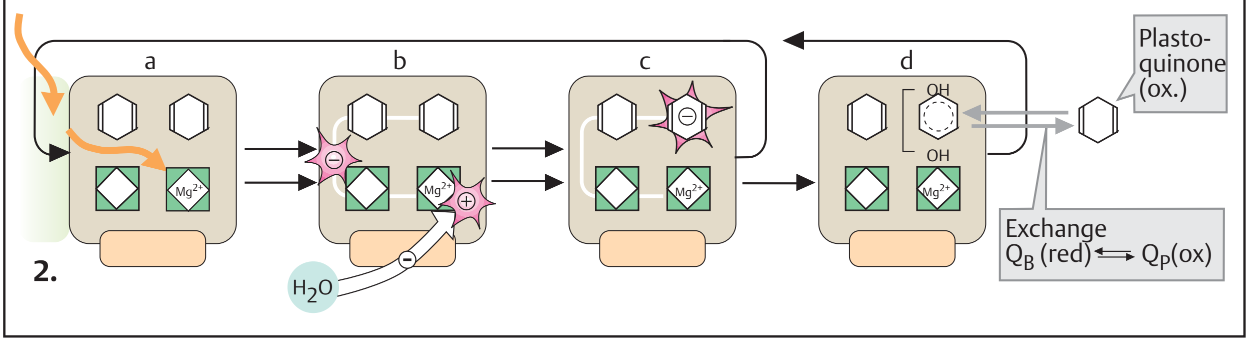
a
b
c
d
Mg^{2+}
⊖
⊕
H_2O
OH
Plasto-quinone (ox.)
Exchange Q_B (red) ⟷ Q_P(ox)
2.

C. Calvin cycle

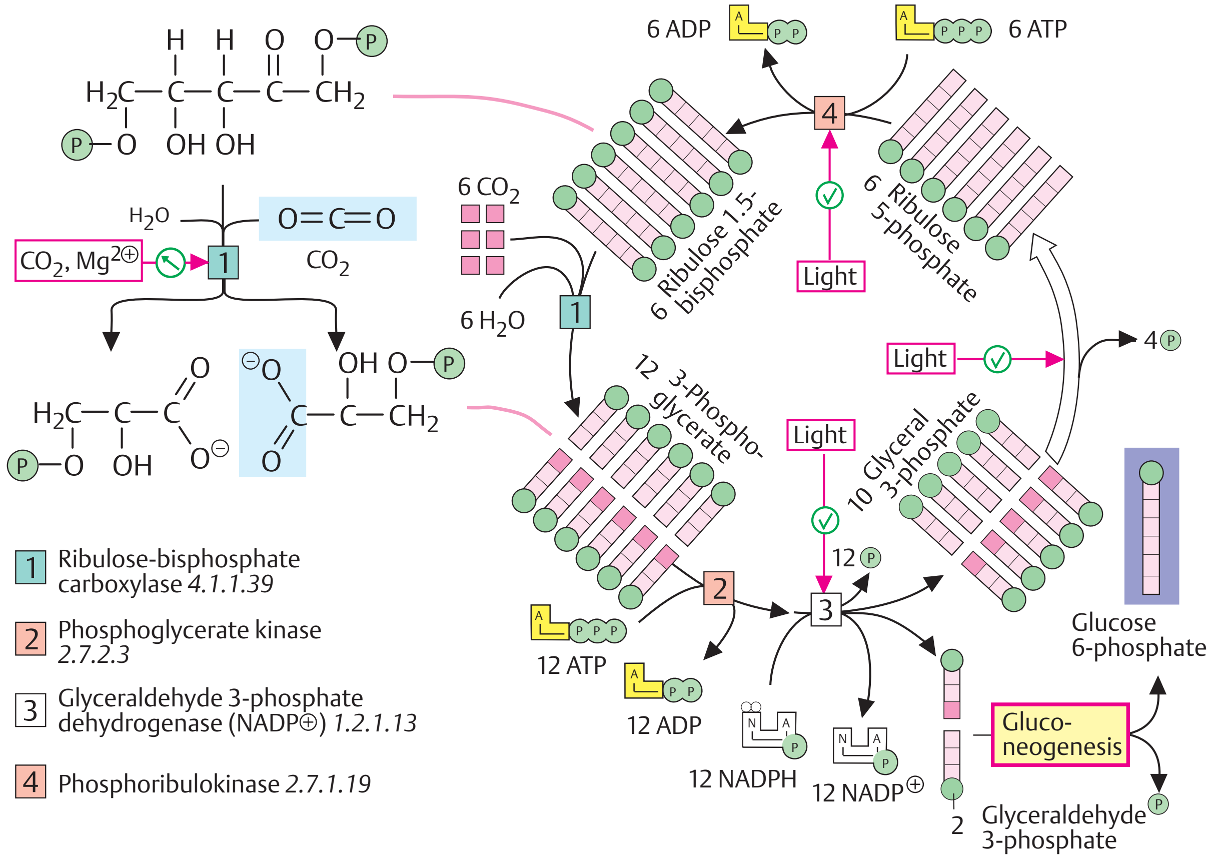
H H O O—P
H_2C—C—C—C—CH_2
P—O OH OH
H_2O
O=C=O
CO_2
CO_2, $Mg^{2⊕}$
1
H_2C—C—C
P—O OH O⊖
⊖O
C—C—CH_2
OH O—P
6 ADP
6 ATP
4
6 Ribulose 1.5-bisphosphate
6 Ribulose 5-phosphate
6 CO_2
6 H_2O
Light
Light
Light
4 P
12 3-Phospho-glycerate
10 Glyceral 3-phosphate
2
3
12 P
12 ATP
12 ADP
12 NADPH
12 NADP⊕
Glucose 6-phosphate
Gluco-neogenesis
2
Glyceraldehyde 3-phosphate
1 Ribulose-bisphosphate carboxylase 4.1.1.39
2 Phosphoglycerate kinase 2.7.2.3
3 Glyceraldehyde 3-phosphate dehydrogenase (NADP⊕) 1.2.1.13
4 Phosphoribulokinase 2.7.1.19

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Molecular models: membrane proteins

The plates show, in simplified form, the structures of **cytochrome *c* oxidase** (**A**; complex IV of the respiratory chain) and of **photosystem I** of a cyanobacterium (**B**). These two molecules are among the few integral membrane proteins for which the structure is known in detail. Both structures were determined by X-ray crystallography.

A. Cytochrome *c* oxidase ○

The enzyme cytochrome *c* oxidase ("COX," *EC 1.9.3.1*) catalyzes the final step of the respiratory chain. It receives electrons from the small heme protein cytochrome *c* and transfers them to molecular oxygen, which is thereby reduced to water (see p. 140). At the same time, 2–4 protons per water molecule formed are pumped from the matrix into the intermembrane space.

Mammalian COX (the illustration shows the enzyme from bovine heart) is a dimer that has two identical subunits with masses of 204 kDa each. Only one subunit is shown in detail here; the other is indicated by gray lines. Each subunit consists of 13 different polypeptides, which all span the inner mitochondrial membrane. Only polypeptides I (light blue) and II (dark blue) and the linked cofactors are involved in electron transport. The other chains, which are differently expressed in the different organs, probably have regulatory functions. The two heme groups, heme a (orange) and heme a_1 (red) are bound in polypeptide 1. The copper center Cu_A consists of two copper ions (green), which are coordinated by amino acid residues in polypeptide II. The second copper (Cu_B) is located in polypeptide I near heme a_3.

To reduce an O_2 molecule to two molecules of H_2O, a total of four electrons are needed, which are supplied by cytochrome *c* (pink, top left) and initially given off to Cu_A. From there, they are passed on via heme a and heme a_3 to the enzyme's reaction center, which is located between heme a_3 and Cu_B. The reduction of the oxygen takes place in several steps, without any intermediate being released. The four protons needed to produce water and the H^+ ions pumped into the intermembrane space are taken up by two channels (D and K, not shown). The mechanism that links proton transport to electron transfer is still being investigated.

B. Reaction center of *Synechococcus elongatus* ○

Photosystem I (PS I) in the cyanobacterium *Synechococcus elongatus* is the first system of this type for which the structure has been solved in atomic detail. Although the bacterial photosystem differs slightly from the systems in higher plants, the structure provides valuable hints about the course of the light reactions in photosynthesis (see p. 128). The functioning of the photosystem is discussed in greater detail on p. 130.

The functional form of PS I in *S. elongatus* consists of a trimer with a mass of more than 10^6 Da that is integrated into the membrane. Only one of the three subunits is shown here. This consists of 12 different polypeptides (gray-blue), 96 chlorophyll molecules (green), 22 carotenoids (orange), several phylloquinones (yellow), and other components. Most of the chlorophyll molecules are so-called **antenna pigments**. These collect light energy and conduct it to the **reaction center**, which is located in the center of the structure and therefore not visible. In the reaction center, an electron is excited and transferred via various intermediate steps to a ferredoxin molecule (see p. 128). The **chlorophylls** (see formula) are heme-like pigments with a highly modified tetrapyrrole ring, a central Mg^{2+} ion, and an apolar phytol side chain. Shown here is chlorophyll a, which is also found in the reaction center of the *S. elongatus* photosystem.

The yellow and orange-colored **carotenoids**—e. g., *β-carotene* (see formula)—are auxiliary pigments that serve to protect the chloroplasts from oxidative damage. Dangerous radicals can be produced during the light reaction—particularly *singlet oxygen*. Carotenoids prevent compounds of this type from arising, or render them inactive. Carotenoids are also responsible for the coloring of leaves seen during fall. They are left behind when plants break down chlorophyll in order to recover the nitrogen it contains.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Cytochrome *C* oxidase

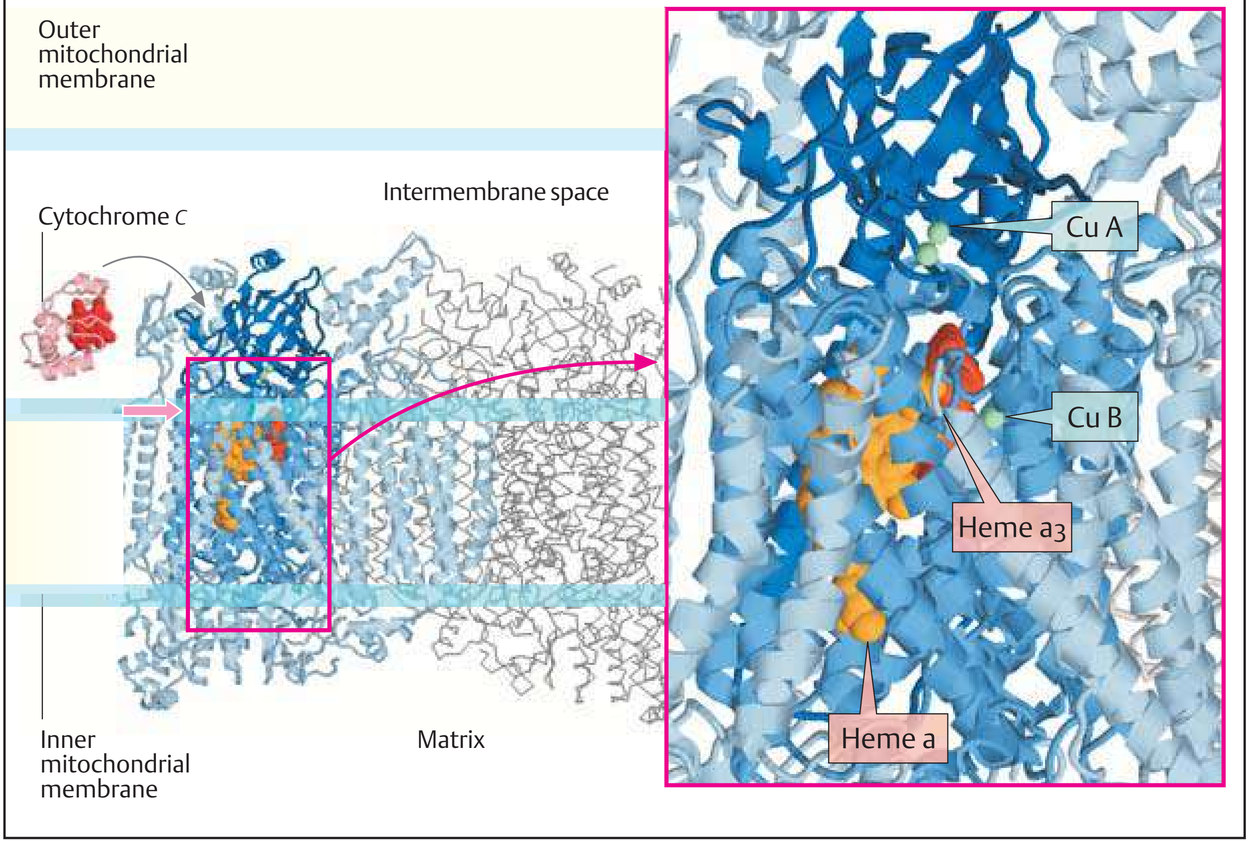
Outer mitochondrial membrane
Cytochrome *c*
Intermembrane space
Cu A
Cu B
Heme a3
Heme a
Inner mitochondrial membrane
Matrix

B. Photosystem I

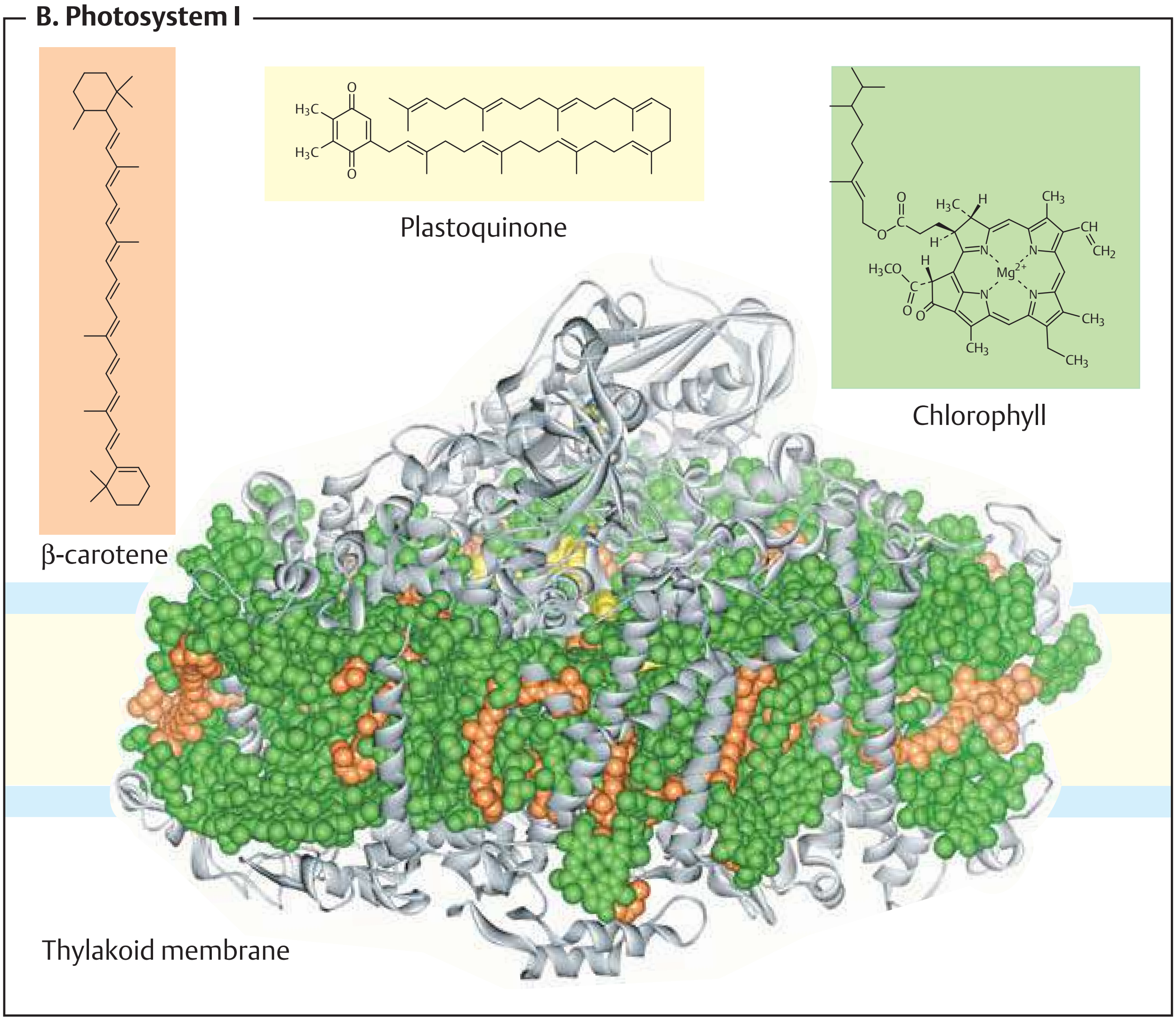
β-carotene
Plastoquinone
Chlorophyll
Thylakoid membrane

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Oxoacid dehydrogenases

The intermediary metabolism has *multienzyme complexes* which, in a complex reaction, catalyze the **oxidative decarboxylation** of 2-oxoacids and the transfer to coenzyme A of the acyl residue produced. NAD^+ acts as the electron acceptor. In addition, thiamine diphosphate, lipoamide, and FAD are also involved in the reaction. The *oxoacid dehydrogenases* include a) the *pyruvate dehydrogenase complex* (PDH, pyruvate → acetyl CoA), b) the *2-oxoglutarate dehydrogenase complex* of the tricarboxylic acid cycle (ODH, 2-oxoglutarate → succinyl CoA), and c) the *branched chain dehydrogenase complex,* which is involved in the catabolism of valine, leucine, and isoleucine (see p. 414).

A. Pyruvate dehydrogenase: reactions ◑

The pyruvate dehydrogenase reaction takes place in the mitochondrial matrix (see p. 210). Three different enzymes [E1–E3] form the PDH multienzyme complex (see **B**).

[1] Initially, *pyruvate dehydrogenase* [E1] catalyzes the decarboxylation of pyruvate and the transfer of the resulting hydroxyethyl residue to **thiamine diphosphate** (TPP, **1a**). The same enzyme then catalyzes oxidation of the TPP-bound hydroxyethyl group to yield an acetyl residue. This residue and the reducing equivalents obtained are then transferred to **lipoamide** (**1b**).

[2] The second enzyme, *dihydrolipoamide acetyltransferase* [E2], shifts the acetyl residue from lipoamide to **coenzyme A** (**2**), with dihydrolipoamide being left over.

[3] The third enzyme, *dihydrolipoamide dehydrogenase* [E3], reoxidizes dihydrolipoamide, with **$NADH+H^+$** being formed. The electrons are first taken over by enzyme-bound **FAD** (**3a**) and then transferred via a catalytically active disulfide bond in the E3 subunit (not shown) to soluble NAD^+ (**3b**).

The five different **coenzymes** involved are associated with the enzyme components in different ways. Thiamine diphosphate is non-covalently bound to E1, whereas lipoamide is covalently bound to a lysine residue of E2 and FAD is bound as a *prosthetic group* to E3. NAD^+ and coenzyme A, being soluble coenzymes, are only temporarily associated with the complex.

An important aspect of PDH catalysis is the spatial relationship between the components of the complex. The covalently bound lipoamide coenzyme is part of a mobile domain of E2, and is therefore highly mobile. This structure is known as the *lipoamide arm,* and swings back and forth between E1 and E3 during catalysis. In this way, lipoamide can interact with the TPP bound at E1, with solute coenzyme A, and also with the FAD that serves as the electron acceptor in E3.

B. PDH complex of *Escherichia coli* ○

The PDH complex of the bacterium *Escherichia coli* has been particularly well studied. It has a molecular mass of 5.3 10^6, and with a diameter of more than 30 nm it is larger than a ribosome. The complex consists of a total of 60 polypeptides (**1**, **2**): 24 molecules of E2 (eight trimers) form the almost cube-shaped core of the complex. Each of the six surfaces of the cube is occupied by a dimer of E3 components, while each of the twelve edges of the cube is occupied by dimers of E1 molecules. Animal oxoacid dehydrogenases have similar structures, but differ in the numbers of subunits and their molecular masses.

Further information

The PDH reaction, which is practically irreversible, occupies a strategic position at the interface between carbohydrate and fatty acid metabolism, and also supplies acetyl residues to the tricarboxylic acid cycle. PDH activity is therefore strictly regulated (see p. 144). **Interconversion** is particularly important in animal cells (see p. 120). Several PDH-specific *protein kinases* inactivate the E1 components through phosphorylation, while equally specific *protein phosphatases* reactivate it again. The binding of the kinases and phosphatases to the complex is in turn regulated by metabolites. For example, high concentrations of acetyl CoA promote binding of kinases and thereby inhibit the reaction, while Ca^{2+} increases the activity of the phosphatase. Insulin activates PDH via inhibition of phosphorylation.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Pyruvate dehydrogenase: reactions

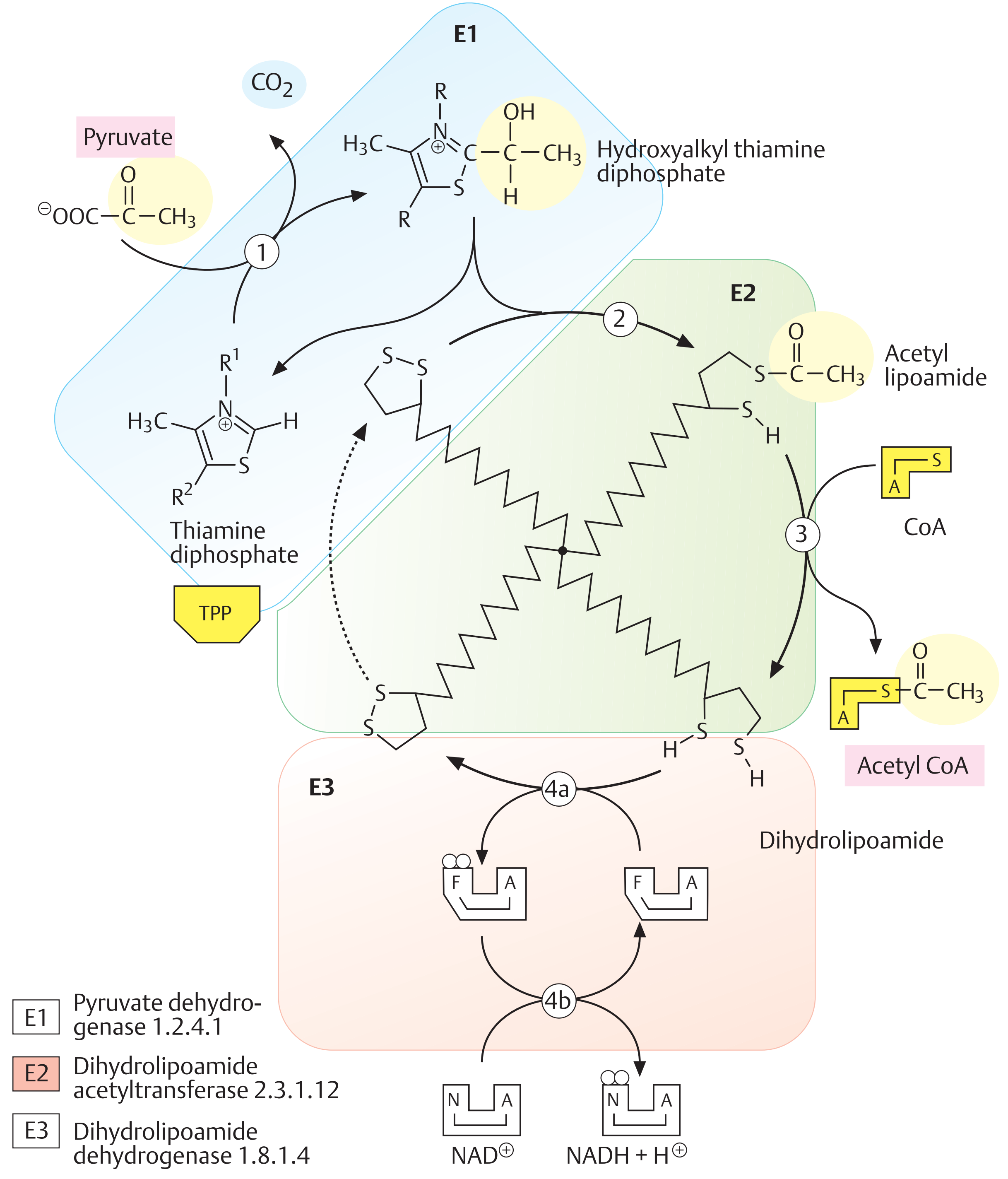
E1
CO2
Pyruvate
$^{\ominus}OOC-C(=O)-CH_3$
1
Hydroxyalkyl thiamine diphosphate
E2
2
Acetyl lipoamide
Thiamine diphosphate
TPP
3
CoA
Acetyl CoA
E3
4a
Dihydrolipoamide
4b
NAD⊕
NADH + H⊕
E1 Pyruvate dehydrogenase 1.2.4.1
E2 Dihydrolipoamide acetyltransferase 2.3.1.12
E3 Dihydrolipoamide dehydrogenase 1.8.1.4

B. PDH complex of *Escherichia coli*

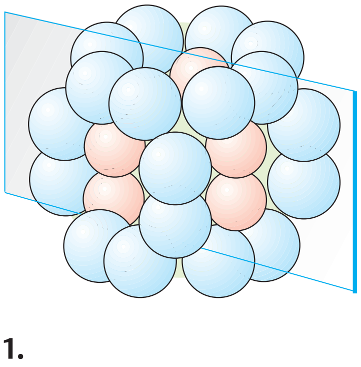
1.

E1	12 · 2 = 24
E2	8 · 3 = 24
E3	6 · 2 =12
E1+ E2 + E3 = 60	

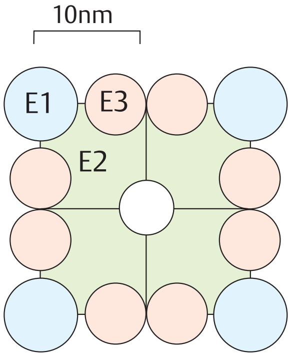
10nm
E1
E3
E2
2.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Tricarboxylic acid cycle: reactions

The **tricarboxylic acid cycle** (TCA cycle, also known as the citric acid cycle or Krebs cycle) is a cyclic metabolic pathway in the mitochondrial matrix (see p. 210). In eight steps, it oxidizes acetyl residues (CH_3-CO-) to carbon dioxide (CO_2). The reducing equivalents obtained in this process are transferred to NAD^+ or ubiquinone, and from there to the respiratory chain (see p. 140). Additional metabolic functions of the cycle are discussed on p. 138.

A. Tricarboxylic acid cycle ◑

The acetyl-CoA that supplies the cycle with acetyl residues is mainly derived from *β-oxidation* of fatty acids (see p. 164) and from the *pyruvate dehydrogenase reaction.* Both of these processes take place in the mitochondrial matrix.

[1] In the first step of the cycle, *citrate synthase* catalyzes the transfer of an acetyl residue from **acetyl CoA** to a carrier molecule, oxaloacetic acid. The product of this reaction, **tricarboxylic acid,** gives the cycle its name.

[2] In the next step, tricarboxylic acid undergoes isomerization to yield **isocitrate.** In the process, only the hydroxyl group is shifted within the molecule. The corresponding enzyme is called *aconitate hydratase* ("aconitase"), because unsaturated *aconitate* arises as an enzyme-bound intermediate during the reaction (not shown; see p. 8). Due to the properties of aconitase, the isomerization is absolutely *stereospecific*. Although citrate is not chiral, isocitrate has two chiral centers, so that it could potentially appear in *four* isomeric forms. However, in the tricarboxylic acid cycle, only one of these stereoisomers, (*2R,3S*)-isocitrate, is produced.

[3] The first oxidative step now follows. *Isocitrate dehydrogenase* oxidizes the hydroxyl group of isocitrate into an oxo group. At the same time, a carboxyl group is released as CO_2, and **2-oxoglutarate** (also known as α-ketoglutarate) and NADH+H^+ are formed.

[4] The next step, the formation of **succinyl CoA,** also involves one oxidation and one decarboxylation. It is catalyzed by *2-oxoglutarate dehydrogenase,* a multienzyme complex closely resembling the PDH complex (see p. 134). NADH+H^+ is once again formed in this reaction.

[5] The subsequent cleavage of the thioester succinylCoA into **succinate** and coenzyme A by *succinic acid-CoA ligase* (succinyl CoA synthetase, succinic thiokinase) is strongly *exergonic* and is used to synthesize a phosphoric acid anhydride bond ("*substrate level phosphorylation* ", see p. 124). However, it is not ATP that is produced here as is otherwise usually the case, but instead **guanosine triphosphate (GTP)**. However, GTP can be converted into ATP by a *nucleoside diphosphate kinase* (not shown).

[6] Via the reactions described so far, the acetyl residue has been completely oxidized to CO_2. At the same time, however, the carrier molecule oxaloacetate has been reduced to succinate. Three further reactions in the cycle now regenerate oxaloacetate from succinate. Initially, *succinate dehydrogenase* oxidizes succinate to **fumarate**. In contrast to the other enzymes in the cycle, succinate dehydrogenase is an integral protein of the inner mitochondrial membrane. It is therefore also assigned to the respiratory chain as complex II. Although succinate dehydrogenase contains FAD as a prosthetic group, **ubiquinone** is the real electron acceptor of the reaction.

[7] Water is now added to the double bond of fumarate by *fumarate hydratase* (*"fumarase"*), and chiral *(2S)*-**malate** is produced.

[8] In the last step of the cycle, malate is again oxidized by *malate dehydrogenase* into **oxaloacetate**, with NADH+H^+ again being produced. With this reaction, the cycle is complete and can start again from the beginning. As the equilibrium of the reaction lies well on the side of malate, the formation of oxaloacetic acid by reaction [8] depends on the strongly exergonic reaction [1], which immediately removes it from the equilibrium.

The **net outcome** is that each rotation of the tricarboxylic acid cycle converts one acetyl residue and two molecules of H_2O into two molecules of CO_2. At the same time, one GTP, three NADH+H^+ and one reduced ubiquinone (QH_2) are produced. By oxidative phosphorylation (see p. 122), the cell obtains around nine molecules of ATP from these reduced coenzymes (see p. 146). Together with the directly formed GTP, this yields a total of 10 ATP per acetyl group.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Tricarboxylic acid cycle

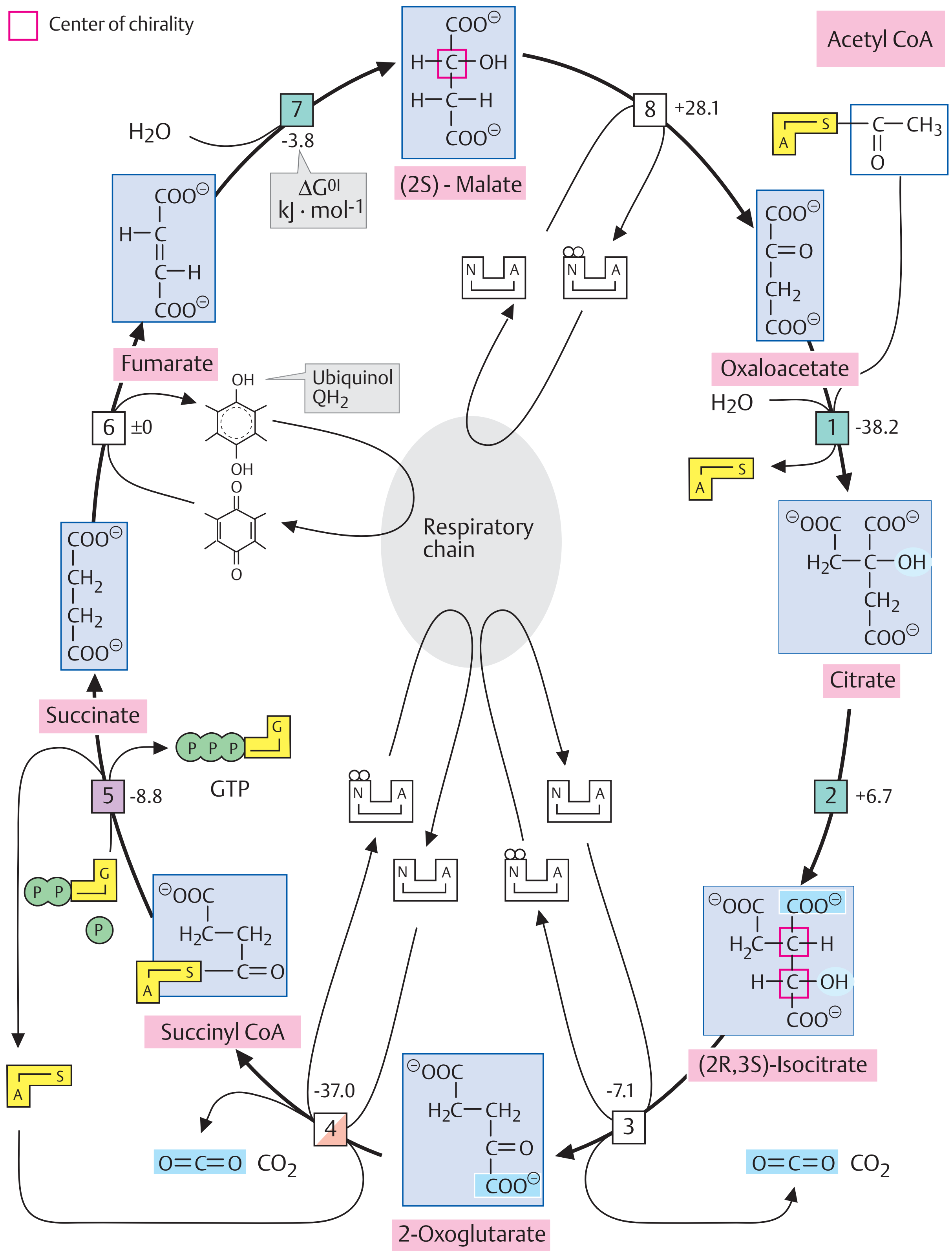
Center of chirality
Acetyl CoA
A
S
C—CH3
O
COO⊖
H—C—OH
H—C—H
COO⊖
(2S) - Malate
8
+28.1
7
-3.8
ΔG0' kJ · mol-1
H2O
Oxaloacetate
COO⊖
C=O
CH2
COO⊖
H2O
1
-38.2
Citrate
⊖OOC
COO⊖
H2C—C—OH
CH2
COO⊖
2
+6.7
(2R,3S)-Isocitrate
⊖OOC
COO⊖
H2C—C—H
H—C—OH
COO⊖
3
-7.1
O=C=O
CO2
2-Oxoglutarate
⊖OOC
H2C—CH2
C=O
COO⊖
4
-37.0
O=C=O
CO2
Succinyl CoA
⊖OOC
H2C—CH2
C=O
5
-8.8
GTP
P
P
G
Succinate
COO⊖
CH2
CH2
COO⊖
6
±0
Fumarate
COO⊖
H—C
C—H
COO⊖
OH
OH
O
O
Ubiquinol QH2
Respiratory chain
N
A

1	Citrate synthase 4.1.3.7	4	2-Oxoglutarate DH complex 1.2.4.2, 1.8.1.4, 2.3.1.61	6	Succinate DH 1.3.5.1 [FAD, Fe_2S_2, Fe_4S_4]
2	Aconitase 4.2.1.3 [Fe_4S_4]	5	Succinate-CoA ligase 6.2.1.4	7	Fumarate hydratase 4.2.1.2
3	Isocitrate DH 1.1.1.41		DH = dehydrogenase	8	Malate DH 1.1.1.37

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Tricarboxylic acid cycle: functions

A. Tricarboxylic acid cycle: functions ◑

The tricarboxylic acid cycle (see p. 136) is often described as the "hub of intermediary metabolism." It has both catabolic and anabolic functions—it is **amphibolic.**

As a **catabolic pathway**, it initiates the **"terminal oxidation"** of energy substrates. Many catabolic pathways lead to intermediates of the tricarboxylic acid cycle, or supply metabolites such as pyruvate and acetyl-CoA that can enter the cycle, where their C atoms are oxidized to CO_2. The reducing equivalents (see p. 14) obtained in this way are then used for *oxidative phosphorylation*—i. e., to aerobically synthesize ATP (see p. 122).

The tricarboxylic acid cycle also supplies important **precursors for anabolic pathways.** Intermediates in the cycle are converted into:

- Glucose (gluconeogenesis; precursors: oxaloacetate and malate—see p. 154)
- Porphyrins (precursor: succinyl-CoA—see p. 192)
- Amino acids (precursors: 2-oxoglutarate, oxaloacetate—see p. 184)
- Fatty acids and isoprenoids (precursor: citrate—see below)

The intermediates of the tricarboxylic acid cycle are present in the mitochondria only in very small quantities. After the oxidation of acetyl-CoA to CO_2, they are constantly regenerated, and their concentrations therefore remain constant, averaged over time. Anabolic pathways, which remove intermediates of the cycle (e. g., gluconeogenesis) would quickly use up the small quantities present in the mitochondria if metabolites did not reenter the cycle at other sites to replace the compounds consumed. Processes that replenish the cycle in this way are called **anaplerotic reactions**.

The degradation of most amino acids is anaplerotic, because it produces either intermediates of the cycle or pyruvate (*glucogenic amino acids;* see p. 180). Gluconeogenesis is in fact largely sustained by the degradation of amino acids. A particularly important anaplerotic step in animal metabolism leads from pyruvate to oxaloacetic acid. This ATP-dependent reaction is catalyzed by *pyruvate carboxylase* [1]. It allows pyruvate yielding amino acids and lactate to be used for gluconeogenesis.

By contrast, *acetyl CoA does not have anaplerotic effects* in animal metabolism. Its carbon skeleton is completely oxidized to CO_2 and is therefore no longer available for biosynthesis. Since fatty acid degradation only supplies acetyl CoA, animals are unable to convert fatty acids into glucose. During periods of hunger, it is therefore not the fat reserves that are initially drawn on, but proteins. In contrast to fatty acids, the amino acids released are able to maintain the blood glucose level (see p. 308).

The tricarboxylic acid cycle not only takes up acetyl CoA from fatty acid degradation, but also supplies the material for the *biosynthesis of fatty acids* and isoprenoids. Acetyl CoA, which is formed in the matrix space of mitochondria by pyruvate dehydrogenase (see p. 134), is not capable of passing through the inner mitochondrial membrane. The acetyl residue is therefore condensed with oxaloacetate by mitochondrial *citrate synthase* to form citrate. This then leaves the mitochondria by antiport with malate (right; see p. 212). In the cytoplasm, it is cleaved again by ATP-dependent *citrate lyase* [4] into acetyl-CoA and oxaloacetate. The oxaloacetate formed is reduced by a cytoplasmic *malate dehydrogenase* to malate [2], which then returns to the mitochondrion via the antiport already mentioned. Alternatively, the malate can be oxidized by *"malic enzyme"* [5], with decarboxylation, to pyruvate. The $NADPH+H^+$ formed in this process is also used for fatty acid biosynthesis.

Additional information

Using the so-called **glyoxylic acid cycle**, plants and bacteria are able to convert acetyl-CoA into succinate, which then enters the tricarboxylic acid cycle. For these organisms, fat degradation therefore functions as an anaplerotic process. In plants, this pathway is located in special organelles, the *glyoxysomes.*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Tricarboxylic acid cycle: functions

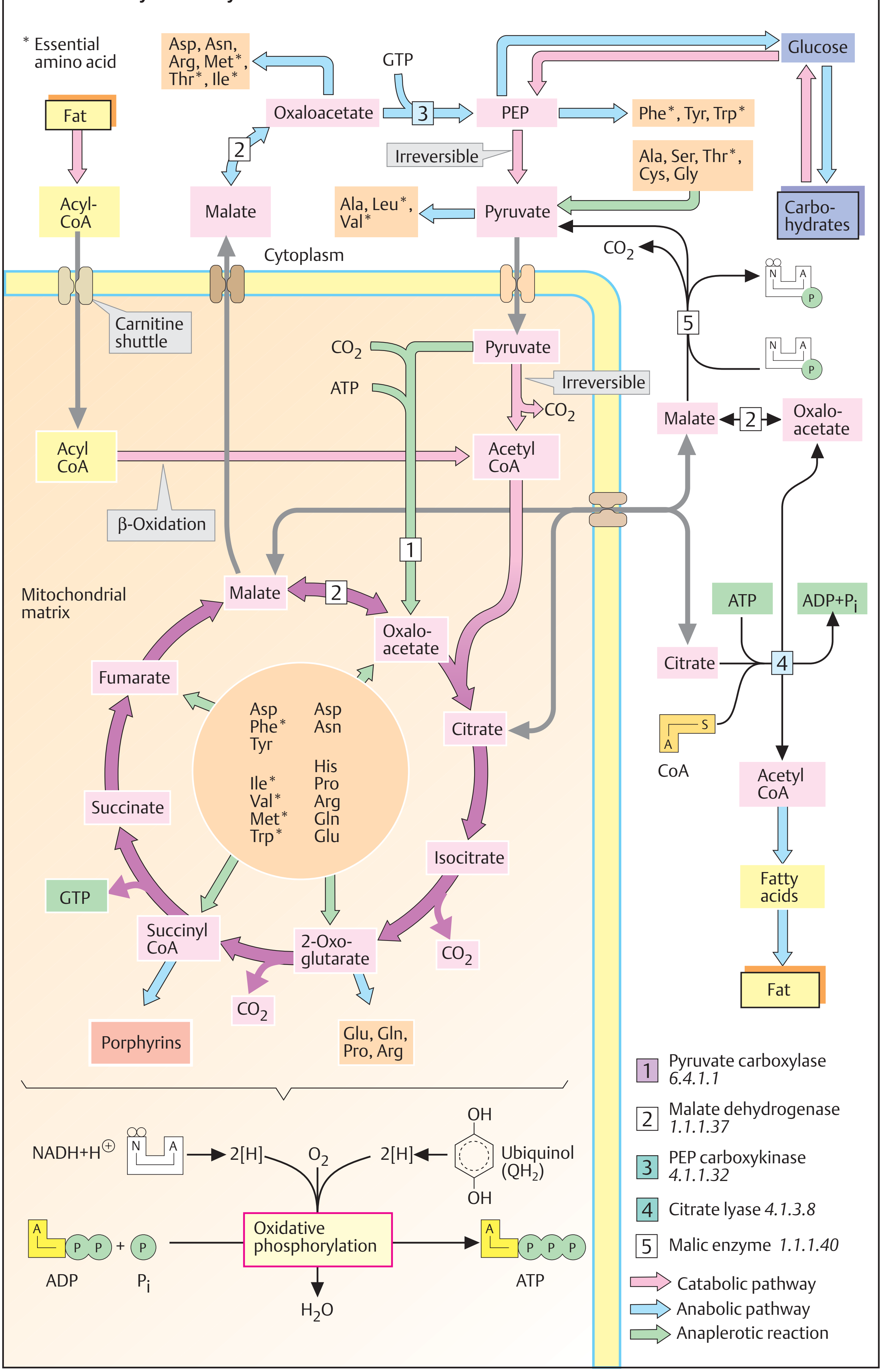
* Essential amino acid
Asp, Asn, Arg, Met*, Thr*, Ile*
GTP
Glucose
Fat
Oxaloacetate
3
PEP
Phe*, Tyr, Trp*
2
Irreversible
Ala, Ser, Thr*, Cys, Gly
Acyl-CoA
Malate
Ala, Leu*, Val*
Pyruvate
Carbo-hydrates
Cytoplasm
CO_2
N A P
5
Carnitine shuttle
CO_2
Pyruvate
ATP
Irreversible
CO_2
Malate
2
Oxalo-acetate
Acyl CoA
Acetyl CoA
β-Oxidation
1
Mitochondrial matrix
Malate
2
Oxalo-acetate
ATP
ADP+P_i
Fumarate
Citrate
4
Citrate
Asp Phe* Tyr
Asp Asn
Ile* Val* Met* Trp*
His Pro Arg Gln Glu
S A
CoA
Acetyl CoA
Succinate
Isocitrate
Fatty acids
GTP
Succinyl CoA
2-Oxo-glutarate
CO_2
CO_2
Fat
Porphyrins
Glu, Gln, Pro, Arg
NADH+H$^{\oplus}$
2[H]
O_2
2[H]
OH
Ubiquinol (QH_2)
OH
Oxidative phosphorylation
A P P + P
ADP
P_i
A P P P
ATP
H_2O
1 Pyruvate carboxylase 6.4.1.1
2 Malate dehydrogenase 1.1.1.37
3 PEP carboxykinase 4.1.1.32
4 Citrate lyase 4.1.3.8
5 Malic enzyme 1.1.1.40
Catabolic pathway
Anabolic pathway
Anaplerotic reaction

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Respiratory chain

The **respiratory chain** is one of the pathways involved in *oxidative phosphorylation* (see p. 122). It catalyzes the steps by which electrons are transported from NADH+H^+ or reduced ubiquinone (QH_2) to molecular oxygen. Due to the wide difference between the redox potentials of the donor (NADH+H^+ or QH_2) and the acceptor (O_2), this reaction is strongly exergonic (see p. 18). Most of the energy released is used to establish a proton gradient across the inner mitochondrial membrane (see p. 126), which is then ultimately used to synthesize ATP with the help of *ATP synthase.*

A. Components of the respiratory chain ◑

The **electron transport chain** consists of three protein complexes (**complexes I, III, and IV**), which are integrated into the inner mitochondrial membrane, and two mobile carrier molecules—**ubiquinone** (coenzyme Q) and **cytochrome c**. *Succinate dehydrogenase,* which actually belongs to the tricarboxylic acid cycle, is also assigned to the respiratory chain as **complex II**. *ATP synthase* (see p. 142) is sometimes referred to as **complex V**, although it is not involved in electron transport. With the exception of complex I, detailed structural information is now available for every complex of the respiratory chain.

All of the complexes in the respiratory chain are made up of numerous polypeptides and contain a series of different protein bound **redox coenzymes** (see pp. 104, 106). These include *flavins* (FMN or FAD in complexes I and II), *iron–sulfur clusters* (in I, II, and III), and *heme groups* (in II, III, and IV). Of the more than 80 polypeptides in the respiratory chain, only 13 are coded by the mitochondrial genome (see p. 210). The remainder are encoded by nuclear genes, and have to be imported into the mitochondria after being synthesized in the cytoplasm (see p. 228).

Electrons enter the respiratory chain in various different ways. In the oxidation of NADH+H^+ by *complex I,* electrons pass via FMN and Fe/S clusters to ubiquinone (Q). Electrons arising during the oxidation of succinate, acyl CoA, and other substrates are passed to ubiquinone by *succinate dehydrogenase* or other *mitochondrial dehydrogenases* via enzyme-bound $FADH_2$ and the electron-transporting flavoprotein (ETF; see p. 164). Ubiquinol passes electrons on to *complex III,* which transfers them via two b-type heme groups, one Fe/S cluster, and heme c_1 to the small heme protein *cytochrome c.* Cytochrome *c* then transports the electrons to complex IV—*cytochrome c oxidase.* Cytochrome *c* oxidase contains redox-active components in the form of two copper centers (Cu_A and Cu_B) and hemes a and a_3, through which the electrons finally reach *oxygen* (see p. 132). As the result of the two-electron reduction of O_2, the strongly basic O^{2-} anion is produced (at least formally), and this is converted into water by binding of two protons. The electron transfer is coupled to the **formation of a proton gradient** by complexes I, III, and IV (see p. 126).

B. Organization ○

Proton transport via complexes I, III, and IV takes place *vectorially* from the matrix into the intermembrane space. When electrons are being transported through the respiratory chain, the H^+ concentration in this space increases—i. e., the pH value there is reduced by about one pH unit. For each H_2O molecule formed, around 10 H^+ ions are pumped into the intermembrane space. If the inner membrane is intact, then generally only *ATP synthase* (see p. 142) can allow protons to flow back into the matrix. This is the basis for the coupling of electron transport to ATP synthesis, which is important for regulation purposes (see p. 144).

As mentioned, although complexes I through V are all integrated into the inner membrane of the mitochondrion, they are not usually in contact with one another, since the electrons are transferred by ubiquinone and cytochrome *c*. With its long apolar side chain, ubiquinone is freely mobile within the membrane. Cytochrome *c* is water-soluble and is located on the *outside* of the inner membrane.

NADH oxidation via complex I takes place on the *inside* of the membrane—i. e., in the matrix space, where the tricarboxylic acid cycle and β-oxidation (the most important sources of NADH) are also located. O_2 reduction and ATP formation also take place in the matrix.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Components of the respiratory chain

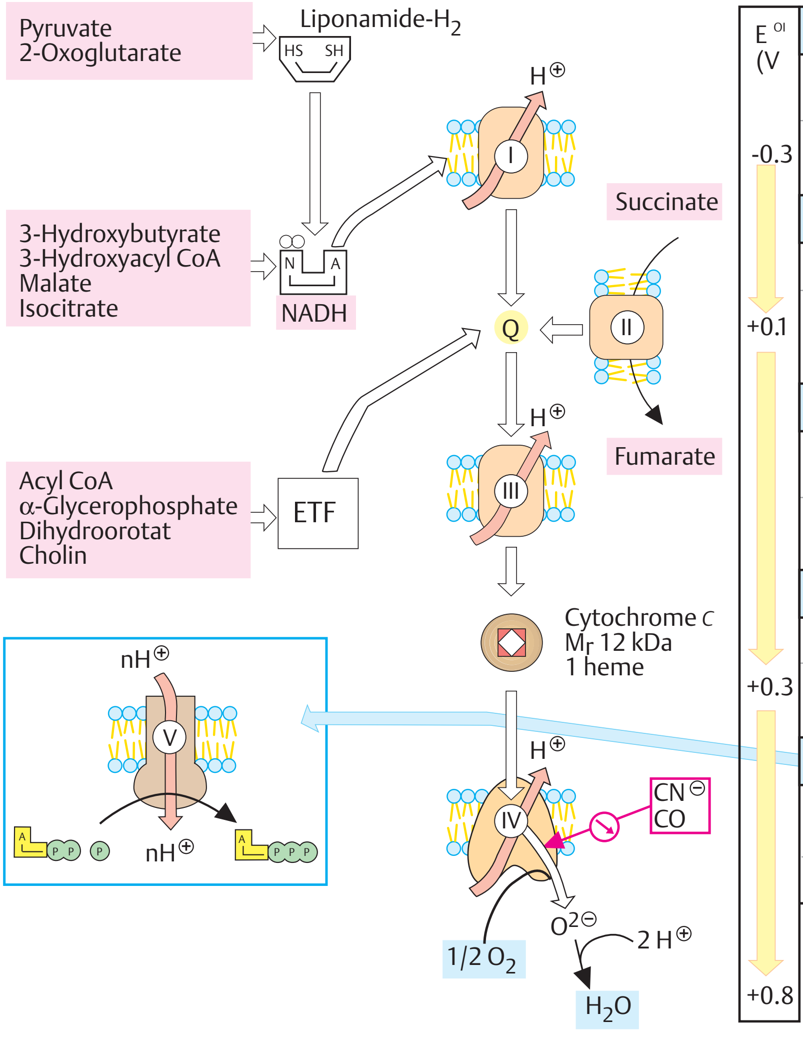
Pyruvate
2-Oxoglutarate
Liponamide-H$_2$
HS SH
3-Hydroxybutyrate
3-Hydroxyacyl CoA
Malate
Isocitrate
N A
NADH
Acyl CoA
α-Glycerophosphate
Dihydroorotat
Cholin
ETF
H⊕
I
Q
II
Succinate
Fumarate
H⊕
III
Cytochrome c
M$_r$ 12 kDa
1 heme
H⊕
IV
CN⊖
CO
O$^{2⊖}$
2 H⊕
1/2 O$_2$
H$_2$O
nH⊕
V
A P P
P
nH⊕
A P P P
E°' (V)
-0.3
+0.1
+0.3
+0.8

Complex I	NADH dehydrogenase (ubiquinone) 1.6.5.3	M_r 700 - 800 kDa, 25 - 30 subunits 1 FMN, 2 Fe_2S_2, 4 - 5 Fe_4S_4
Complex II	Succinate dehydrogenase 1.3.5.1	M_r 125 kDa, 4 - 6 subunits 1 FAD, 1 Fe_2S_2, 1 Fe_4S_4, 1Fe_3S_4 2 ubiquinone, 1 heme b
Complex III	Ubiquinol-cytochrome *c* reductase 1.10.2.2	$M_r \approx$ 400 kDa, 11 subunits 2 Fe_2S_2, 2 heme b, 1 heme c_1
Complex IV	Cytochrome *c* oxidase 1.9.3.1	$M_r \approx$ 200 kDa, 8 - 13 subunits 2 Cu, 1 Zn, 1 heme a, 1 heme a_3
Complex V	H⊕-transporting ATP synthase 3.6.1.34	M_r >400 kDa, >20 subunits

⇨ Electron flow
➡ Proton flow

B. Organization

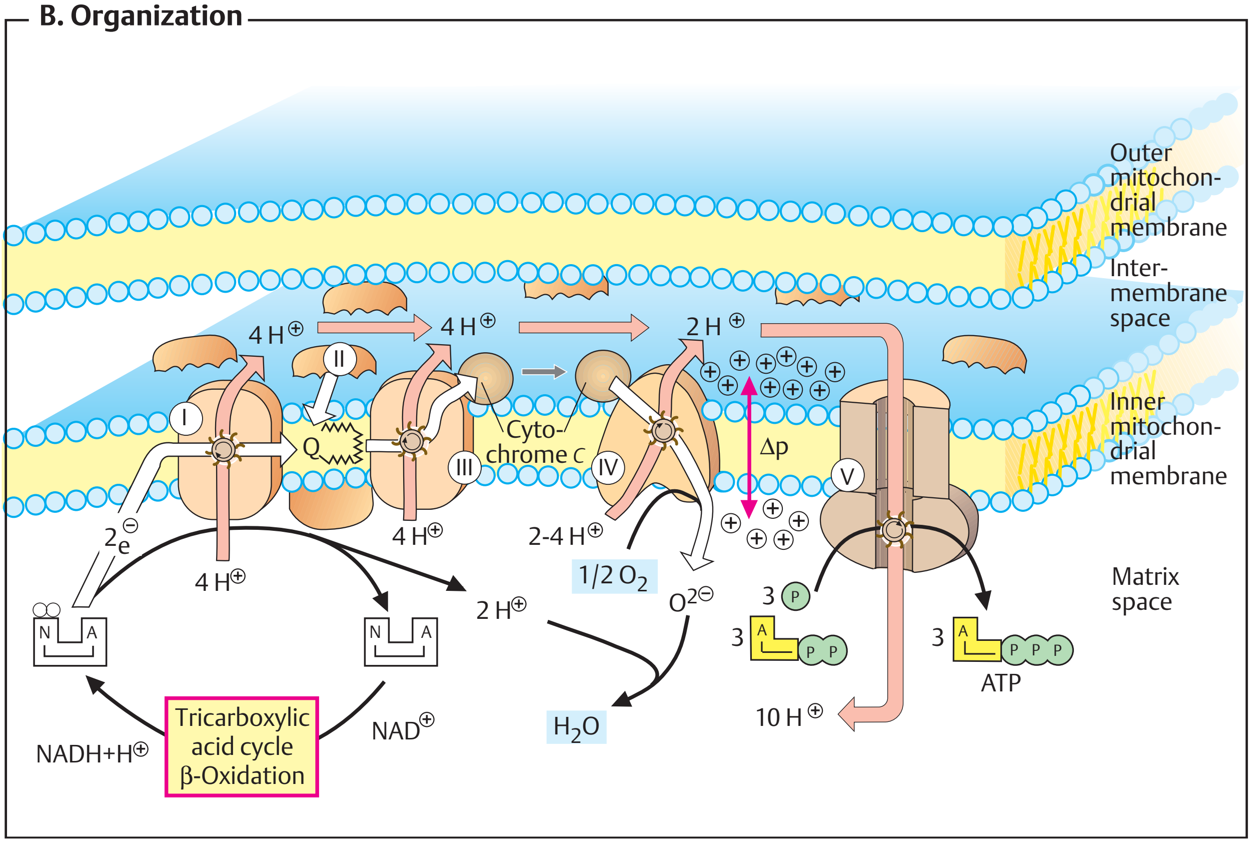
Outer mitochondrial membrane
Intermembrane space
Inner mitochondrial membrane
Matrix space
4 H⊕
4 H⊕
2 H⊕
II
I
Q
III
Cytochrome c
IV
Δp
V
2e⊖
4 H⊕
4 H⊕
2-4 H⊕
1/2 O$_2$
2 H⊕
O$^{2⊖}$
3 P
3 A P P
3 A P P P
ATP
N A
N A
10 H⊕
NADH+H⊕
Tricarboxylic acid cycle β-Oxidation
NAD⊕
H$_2$O

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

ATP synthesis

In the **respiratory chain** (see p. 140), electrons are transferred from NADH or ubiquinol (QH_2) to O_2. The energy obtained in this process is used to establish a proton gradient across the inner mitochondrial membrane. ATP synthesis is ultimately coupled to the return of protons from the intermembrane space into the matrix.

A. Redox systems of the respiratory chain ◑

The electrons provided by NADH do not reach oxygen directly, but instead are transferred to it in various steps. They pass through at least 10 intermediate redox systems, most of which are bound as **prosthetic groups** in complexes I, III, and IV. The large number of coenzymes involved in electron transport may initially appear surprising. However, as discussed on p. 18, in redox reactions, the *change in free enthalpy* ΔG—i. e., the chemical work that is done—depends only on the difference in redox potentials ΔE between the donor and the acceptor. Introducing additional redox systems does not alter the reaction's overall energy yield. In the case of the respiratory chain, the difference between the normal potential of the donor ($NAD^+/NADH+H^+$, $E^{0'} = -0.32$ V) and that of the acceptor (O_2/H_2O, $E^{0'} = +0.82$ V) corresponds to an energy difference $\Delta G^{0'}$ of more than 200 kJ mol^{-1}. This large amount is divided into smaller, more manageable "packages," the size of which is determined by the difference in redox potentials between the respective *intermediates*. It is assumed that this division is responsible for the astonishingly high energy yield (about 60%) achieved by the respiratory chain.

The illustration shows the important redox systems involved in mitochondrial electron transport and their approximate redox potentials. These potentials determine the path followed by the electrons, as the members of a **redox series** have to be arranged in order of increasing redox potential if transport is to occur spontaneously (see p. 32).

In complex 1, the electrons are passed from *NADH+H⁺* first to *FMN* (see p. 104) and then on to several *iron–sulfur (Fe/S) clusters*. These redox systems are only stable in the interior of proteins. Depending on the type, Fe/S clusters may contain two to six iron ions, which form complexes with inorganic sulfide and the SH groups of cysteine residues (see p. 286). *Ubiquinone* (coenzyme Q; see p. 104) is a mobile carrier that takes up electrons from complexes I and II and from reduced ETF and passes them on to complex III. *Heme groups* are also involved in electron transport in a variety of ways. Type b hemes correspond to that found in hemoglobin (see p. 280). Heme c in cytochrome *c* is covalently bound to the protein, while the tetrapyrrole ring of heme a is isoprenylated and carries a formyl group. In complex IV, a *copper ion* (Cu_B) and heme a_3 react directly with oxygen.

B. ATP synthase ◑

The ATP synthase (*EC 3.6.1.34,* complex V) that transports H^+ is a complex molecular machine. The enzyme consists of two parts—a *proton channel* (F_o, for "oligomycin-sensitive") that is integrated into the membrane; and a *catalytic unit* (F_1) that protrudes into the matrix. The F_o part consists of 12 membrane-spanning c-peptides and one a-subunit. The "head" of the F_1 part is composed of three α and three β subunits, between which there are three active centers. The "stem" between F_o and F_1 consists of one γ and one ε subunit. Two more polypeptides, b and δ, form a kind of "stator," fixing the α and β subunits relative to the F_o part.

The catalytic cycle can be divided into three phases, through each of which the three active sites pass in sequence. First, ADP and P_i are bound (**1**), then the anhydride bond forms (**2**), and finally the product is released (**3**). Each time protons pass through the F_o channel protein into the matrix, all three active sites change from their current state to the next. It has been shown that the energy for proton transport is initially converted into a rotation of the γ subunit, which in turn cyclically alters the conformation of the α and β subunits, which are stationary relative to the F_o part, and thereby drives ATP synthesis.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Redox systems of the respiratory chain

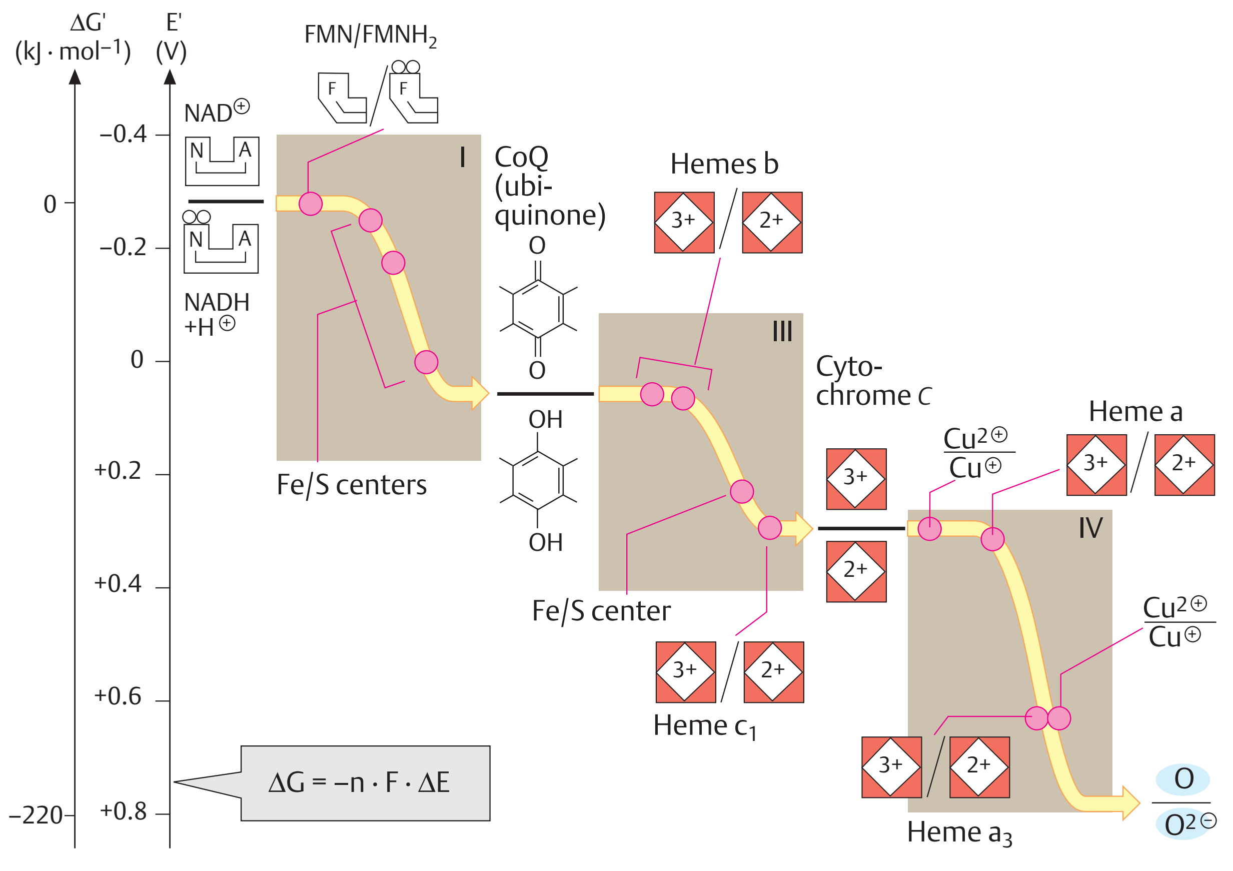
ΔG' (kJ · mol⁻¹)
E' (V)
0
−220
−0.4
−0.2
0
+0.2
+0.4
+0.6
+0.8
NAD⊕
N A
NADH +H⊕
FMN/FMNH₂
F
I
Fe/S centers
CoQ (ubi-quinone)
O
OH
Hemes b
3+
2+
III
Fe/S center
Heme c₁
Cyto-chrome c
Cu²⊕ / Cu⊕
Heme a
IV
Heme a₃
O / O²⊖
ΔG = −n · F · ΔE

B. ATP synthase

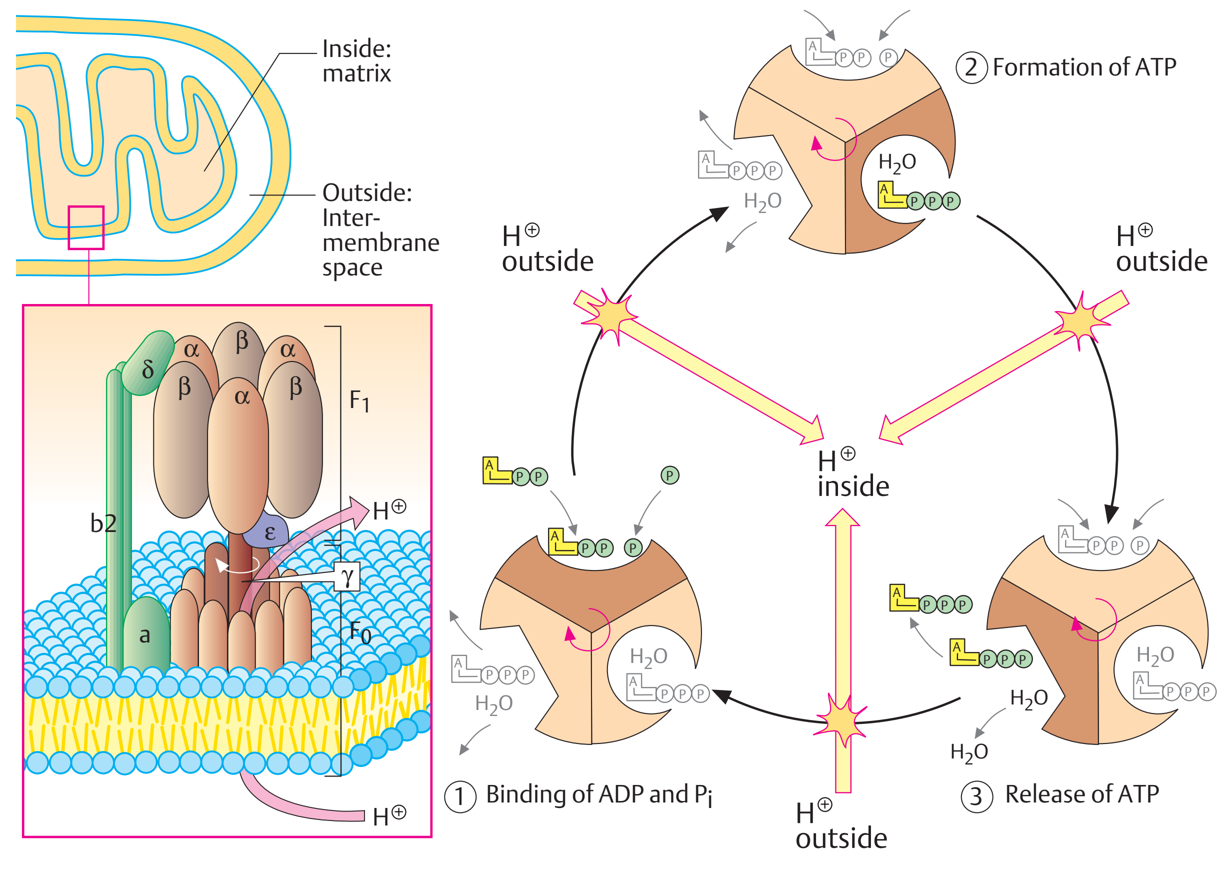
Inside: matrix
Outside: Inter-membrane space
δ
α
β
ε
γ
b2
a
F₁
F₀
H⊕
H⊕ outside
H⊕ inside
A P P P
H₂O
① Binding of ADP and Pᵢ
② Formation of ATP
③ Release of ATP

1. Structure and location

2. Catalytic cycle

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Regulation

The amount of nutrient degradation and ATP synthesis have to be continually adjusted to the body's changing energy requirements. The need to coordinate the production and consumption of ATP is already evident from the fact that the *total amounts* of coenzymes in the organism are low. The human body forms about 65 kg ATP per day, but only contains 3–4 g of adenine nucleotides (AMP, ADP, and ATP). Each ADP molecule therefore has to be phosphorylated to ATP and dephosphorylated again many thousand times a day.

A. Respiratory control ◑

The simple regulatory mechanism which ensures that ATP synthesis is "automatically" coordinated with ATP consumption is known as **respiratory control**. It is based on the fact that the different parts of the oxidative phosphorylation process are *coupled* via shared coenzymes and other factors (left).

If a cell is not using any ATP, hardly any ADP will be available in the mitochondria. Without ADP, *ATP synthase* (**3**) is unable to break down the proton gradient across the inner mitochondrial membrane. This in turn inhibits electron transport in the respiratory chain (**2**), which means that NADH+H^+ can no longer be reoxidized to NAD^+. Finally, the resulting high NADH/NAD^+ ratio inhibits the tricarboxylic acid cycle (**C**), and thus slows down the degradation of the substrate SH_2 (**1**). Conversely, high rates of ATP utilization stimulate nutrient degradation and the respiratory chain via the same mechanism.

If the formation of a proton gradient is prevented (right), substrate oxidation (**1**) and electron transport (**2**) proceed much more rapidly. However, instead of ATP, only heat is produced.

B. Uncouplers ◑

Substances that functionally separate oxidation and phosphorylation from one another are referred to as uncouplers. They break down the proton gradient by allowing H^+ ions to pass from the intermembrane space back into the mitochondrial matrix without the involvement of ATP synthase. Uncoupling effects are produced by **mechanical damage** to the inner membrane (**1**) or by lipid-soluble substances that can transport protons through the membrane, such as **2,4-dinitrophenol** (DNP, **2**). **Thermogenin** (uncoupling protein-1, UCP-1, **3**)—an ion channel (see p. 222) in mitochondria of *brown fat* tissue—is a naturally occurring uncoupler. Brown fat is found, for example, in newborns and in hibernating animals, and serves exclusively to generate heat. In cold periods, norepinephrine activates the *hormone-sensitive lipase* (see p. 162). Increased lipolysis leads to the production of large quantities of free fatty acids. Like DNP, these bind H^+ ions in the intermembrane space, pass the UCP in this form, and then release the protons in the matrix again. This makes fatty acid degradation independent of ADP availability—i. e., it takes place at maximum velocity and only produces heat (**A**). It is becoming increasingly clear that there are also UCPs in other cells, which are controlled by hormones such as thyroxine (see p. 374). This regulates the ATP yield and what is known as the basal metabolic rate.

C. Regulation of the tricarboxylic acid cycle ◑

The most important factor in the regulation of the cycle is the **NADH/NAD^+ ratio**. In addition to *pyruvate dehydrogenase* (PDH) and *oxoglutarate dehydrogenase* (ODH; see p. 134), *citrate synthase* and *isocitrate dehydrogenase* are also inhibited by NAD^+ deficiency or an excess of NADH+H^+. With the exception of isocitrate dehydrogenase, these enzymes are also subject to **product inhibition** by acetyl-CoA, succinyl-CoA, or citrate.

Interconversion processes (see p. 120) also play an important role. They are shown here in detail using the example of the PDH complex (see p. 134). The *inactivating protein kinase* [1a] is inhibited by the substrate pyruvate and is activated by the products acetyl-CoA and NADH+H^+. The p*rotein phosphatase* [1b]—like *isocitrate dehydrogenase* [3] and the *ODH complex* [4]—is activated by Ca^{2+}. This is particularly important during muscle contraction, when large amounts of ATP are needed. *Insulin* also activates the PDH complex (through inhibition of phosphorylation) and thereby promotes the breakdown of glucose and its conversion into fatty acids.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Respiratory control

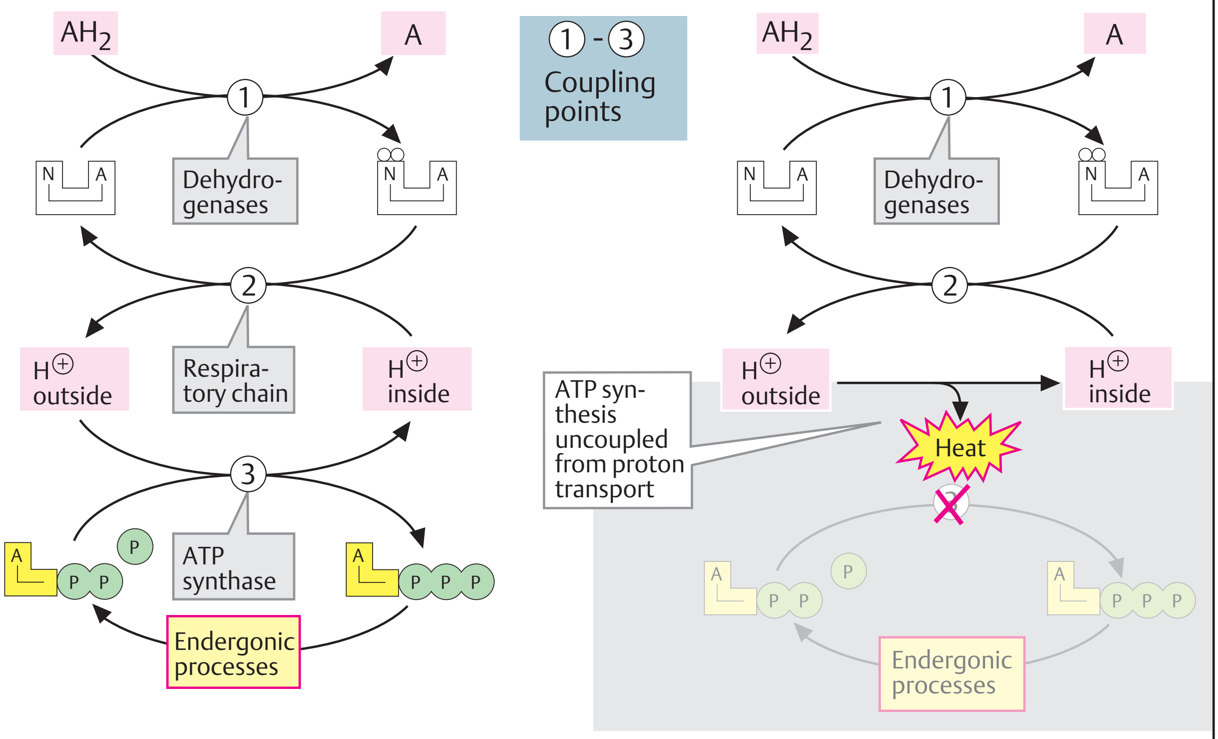
AH$_2$
A
①
Dehydro-genases
N A
②
Respira-tory chain
H^+ outside
H^+ inside
③
ATP synthase
A P P
P
A P P P
Endergonic processes
① - ③ Coupling points
AH$_2$
A
①
Dehydro-genases
②
H^+ outside
H^+ inside
ATP syn-thesis uncoupled from proton transport
Heat
③
P
Endergonic processes

1. Coupled

2. Uncoupled

B. Uncouplers

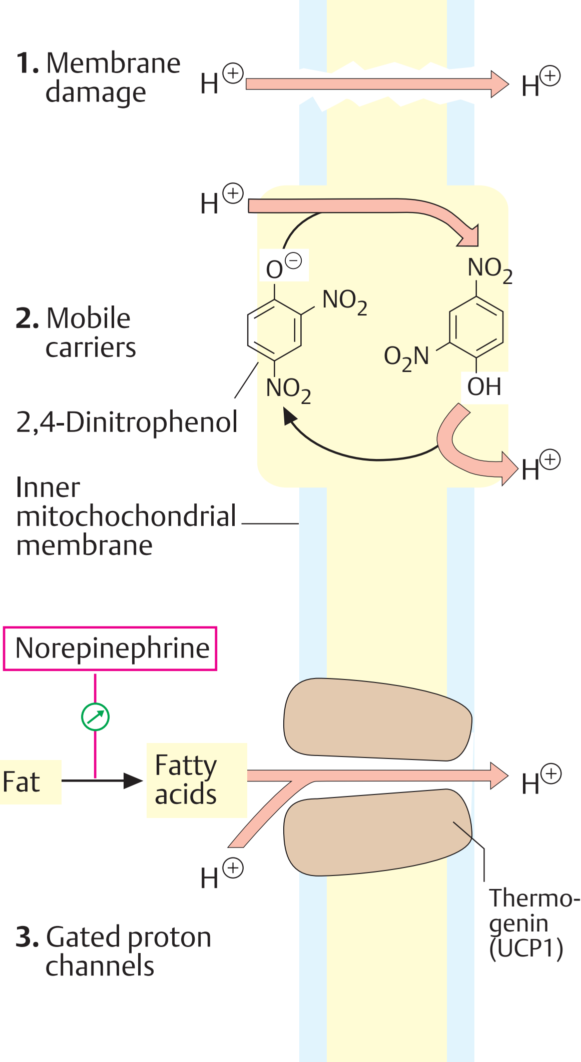
1. Membrane damage
H^+
H^+
H^+
O^-
NO_2
NO_2
NO_2
O_2N
OH
H^+
2. Mobile carriers
2,4-Dinitrophenol
Inner mitochondrial membrane
Norepinephrine
Fat
Fatty acids
H^+
H^+
Thermo-genin (UCP1)
3. Gated proton channels

C. Regulation of the tricarboxylic acid cycle

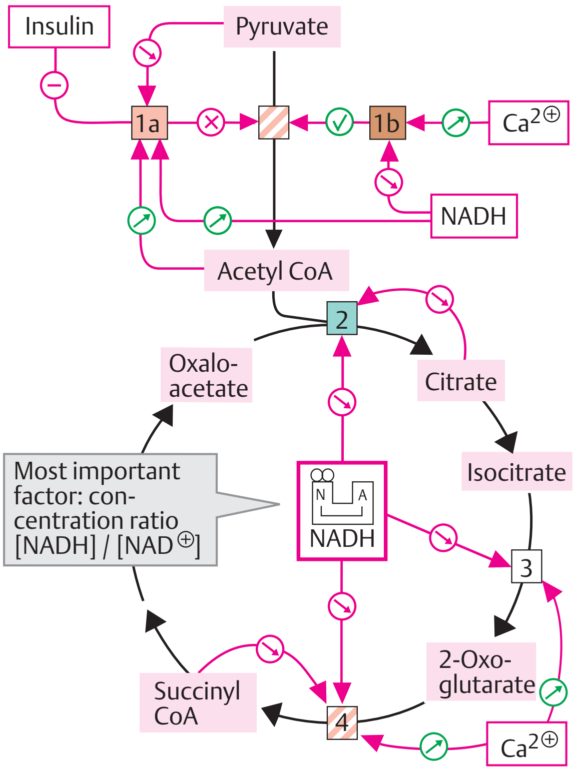
Insulin
Pyruvate
1a
1b
Ca^{2+}
NADH
Acetyl CoA
2
Oxalo-acetate
Citrate
Isocitrate
Most important factor: con-centration ratio [NADH] / [NAD$^+$]
N A
NADH
3
2-Oxo-glutarate
Succinyl CoA
4
Ca^{2+}

1a	PDH kinase *2.7.1.99*	3	Isocitrate dehydrogenase *1.1.1.42*
1b	PDH phosphatase *3.1.3.43*	4	2-Oxoglutarate dehydrogenase *1.2.4.2, 1.8.1.4, 2.3.1.61*
2	Citrate synthase *4.1.3.7*		

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Respiration and fermentation

A. Aerobic and anaerobic oxidation of glucose ◑

In the presence of oxygen (i. e., in *aerobic* conditions), most animal cells are capable of "respiring" various types of nutrient (lipids, amino acids, and carbohydrates)—i. e., using oxidative processes to break them down completely. If oxygen is lacking (i. e., in *anaerobic* conditions), only glucose can be used for ATP synthesis. Although in these conditions glucose breakdown in animals already ends in lactate and only produces small quantities of ATP, it is decisively important for the survival of cells at times of oxygen deficiency.

In **aerobic conditions** (left), ATP is derived almost exclusively from oxidative phosphorylation (see p. 140). **Fatty acids** enter the mitochondria with the help of carnitine (see p. 164), and are broken down there into CoA-bound acetyl residues. **Glucose** is converted into pyruvate by glycolysis (see p. 150) in the cytoplasm. Pyruvate is then also transported into the mitochondrial matrix, where it is oxidatively decarboxylated by the pyruvate dehydrogenase complex (see p. 134) to yield acetyl-CoA. The reducing equivalents (2 $NADH+H^+$ per glucose) that arise in glycolysis enter the mitochondrial matrix via the malate shuttle (see p. 212). The acetyl residues that are formed are oxidized to CO_2 in the tricarboxylic acid cycle (see p. 136). Breakdown of **amino acids** also produces acetyl residues or products that can directly enter the tricarboxylic acid cycle (see p. 180). The reducing equivalents that are obtained are transferred to oxygen via the respiratory chain as required. In the process, chemical energy is released, which is used (via a proton gradient) to synthesize ATP (see p. 140).

In the absence of oxygen—i. e., in **anaerobic conditions**—the picture changes completely. Since O_2 is missing as the electron acceptor for the respiratory chain, $NADH+H^+$ and QH_2 can no longer be reoxidized. Consequently, not only is mitochondrial ATP synthesis halted, but also almost the whole metabolism in the mitochondrial matrix. The main reason for this is the high $NADH+H^+$ concentration and lack of NAD^+, which inhibit the tricarboxylic acid cycle and the pyruvate dehydrogenase reaction (see p. 144). β-Oxidation and the malate shuttle, which are dependent on free NAD^+, also come to a standstill. Since amino acid degradation is also no longer able to contribute to energy production, the cell becomes totally dependent on ATP synthesized via the degradation of glucose by **glycolysis.** For this process to proceed continuously, the $NADH+H^+$ formed in the cytoplasm has to be constantly reoxidized. Since this can no longer occur in the mitochondria, in anaerobic conditions animal cells reduce pyruvate to lactate and pass it into the blood. This type of process is called **fermentation** (see p. 148). The ATP yield is low, with only two ATPs per glucose arising during lactate synthesis.

To estimate the number of ATP molecules formed in an aerobic state, it is necessary to know the **P/O quotient**—i. e., the molar ratio between synthesized ATP ("P") and the water formed ("O"). During transport of two electrons from $NADH+H^+$ to oxygen, about 10 protons are transported into the intermembrane space, while from ubiquinol (QH_2), the number is only six. ATP synthase (see p. 142) probably requires three H^+ to synthesize one ATP, so that maximum P/O quotients of around **3 or 2** are possible. This implies a yield of up to 38 ATP per mol of glucose. However, the actual value is much lower. It needs to be taken into account that the transport of specific metabolites into the mitochondrial matrix and the exchange of ATP^{4-} for ADP^{3-} are also driven by the proton gradient (see p. 212). The P/O quotients for the oxidation of $NADH+H^+$ and QH_2 are therefore more in the range of **2.5 and 1.5**. If the energy balance of aerobic glycolysis is calculated on this basis, the result is a yield of around **32 ATP per glucose**. However, this value is also not constant, and can be adjusted as required by the cell's own uncouplers (UCPs; see p. 144) and other mechanisms.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Aerobic and anaerobic oxidation of glucose

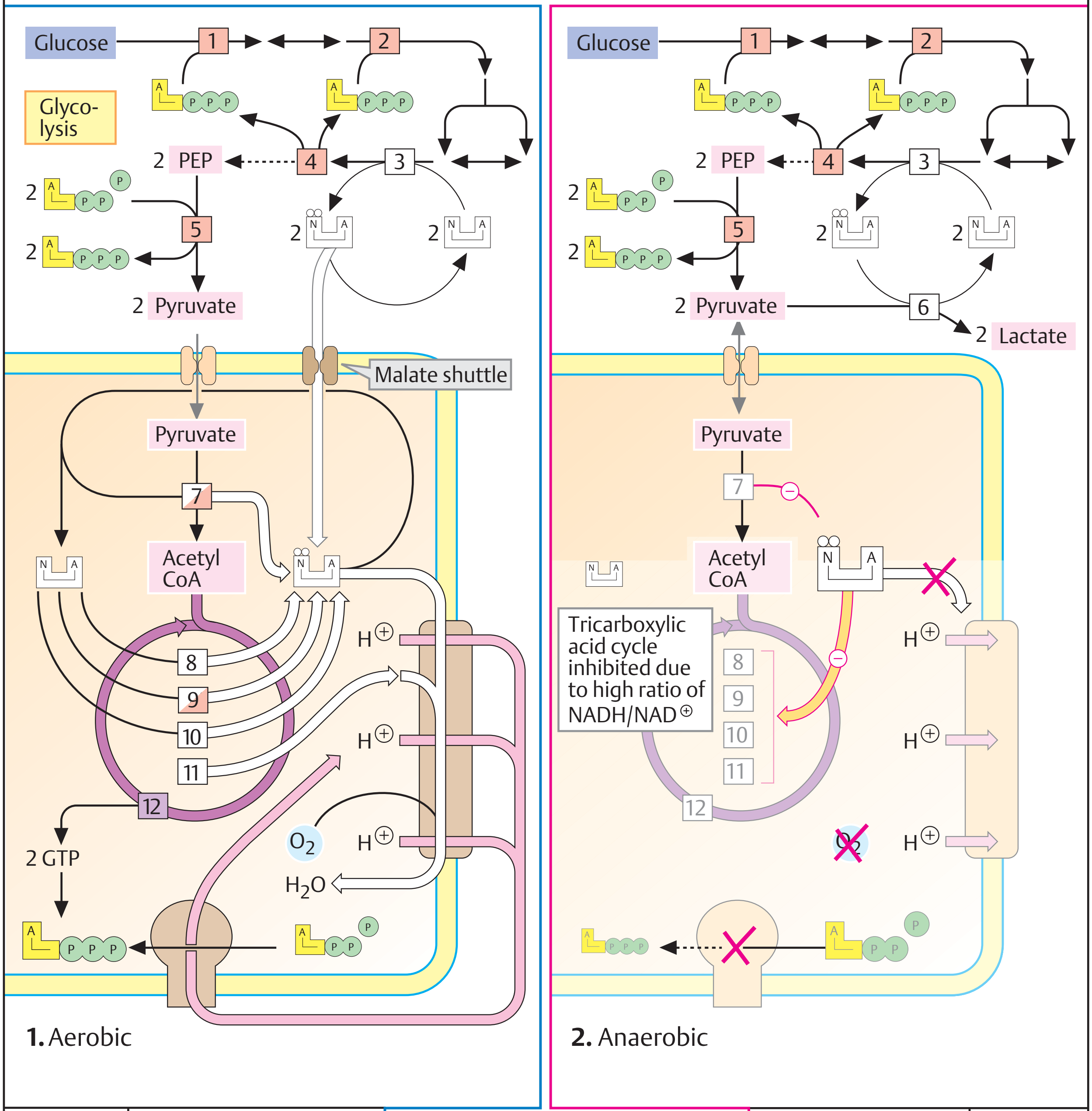
Glucose
1
2
Glyco-lysis
2 PEP
4
3
2
5
2 Pyruvate
Malate shuttle
Pyruvate
7
Acetyl CoA
8
9
10
11
12
2 GTP
$H^{\oplus}$
O_2
H_2O
1. Aerobic
Glucose
1
2
2 PEP
4
3
5
2 Pyruvate
6
2 Lactate
Pyruvate
7
Acetyl CoA
Tricarboxylic acid cycle inhibited due to high ratio of NADH/NAD$^{\oplus}$
8
9
10
11
12
O_2
$H^{\oplus}$
2. Anaerobic

ATP	Coenzymes	Enzymes	Coenzymes	ATP
–1	–1 ATP	1 Hexokinase	–1 ATP	–1
–2	–1 ATP	2 6-Phosphofructokinase	–1 ATP	–2
+3	+5 ATP ← +2 NADH	3 Glyceraldehyde-3Ⓟ DH	+2 NADH ←	–2
+5	+2 ATP	4 Phosphoglycerate kinase	+2 ATP NAD$^{\oplus}$	0
+7	+2 ATP	5 Pyruvate kinase	+2 ATP recycled	+2
		6 Lactate dehydrogenase	–2 NADH ←	
+12	+5 ATP ← +2 NADH	7 Pyruvate dehydrogenase		
+17	+5 ATP ← +2 NADH	8 Isocitrate dehydrogenase		
+22	+5 ATP ← +2 NADH	9 Oxoglutarate dehydrogenase		
+27	+5 ATP ← +2 NADH	10 Malate dehydrogenase		
+30	+3 ATP ← +2 QH_2	11 Succinate dehydrogenase		
+32	+2 ATP ← +2 GTP	12 Succinate-CoA ligase		

Sum: 32 ATP/glucose

DH = dehydrogenase

Sum: 2 ATP/glucose

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Fermentations

As discussed on p. 146, degradation of glucose to pyruvate is the only way for most organisms to synthesize ATP in the *absence of oxygen*. The NADH+H^+ that is also formed in this process has to be constantly reoxidized to NAD^+ in order to maintain glycolysis and thus ATP synthesis. In the animal organism, this is achieved by the reduction of pyruvate to lactate. In microorganisms, there are many other forms of NAD^+ regeneration. Processes of this type are referred to as **fermentations.** Microbial fermentation processes are often used to produce foodstuffs and alcoholic beverages, or to preserve food. Features common to all fermentation processes are that they start with pyruvate and only occur under *anaerobic conditions.*

A. Lactic acid and propionic acid fermentation ○

Many milk products, such as sour milk, yogurt, and cheese are made by *bacterial lactic acid fermentation* (**1**). The reaction is the same as in animals. Pyruvate, which is mainly derived from degradation of the disaccharide *lactose* (see p. 38), is reduced to lactate by *lactate dehydrogenase* [1]. Lactic acid fermentation also plays an important role in the production of sauerkraut and silage. These products usually keep for a long time, because the *pH reduction* that occurs during fermentation inhibits the growth of putrefying bacteria.

Bacteria from the genera *Lactobacillus* and *Streptococcus* are involved in the first steps of dairy production (**3**). The raw materials produced by their effects usually only acquire their final properties after additional fermentation processes. For example, the characteristic taste of Swiss cheese develops during a subsequent propionic acid fermentation. In this process, bacteria from the genus *Propionibacterium* convert pyruvate to propionate in a complex series of reactions (**2**).

B. Alcoholic fermentation ○

Alcoholic beverages are produced by the fermentation of plant products that have a high carbohydrate content. Pyruvate, which is formed from glucose, is initially decarboxylated by *pyruvate decarboxylase* [2], which does not occur in animal metabolism, to produce acetaldehyde (ethanal). When this is reduced by alcohol dehydrogenase [3], with NADH being consumed, *ethanol* [3] is formed.

Yeasts, unicellular fungi that belong to the eukaryotes (**3**), rather than bacteria, are responsible for this type of fermentation. Yeasts are also often used in baking. They produce CO_2 and ethanol, which raise the dough. Brewers' and bakers' yeasts *(Saccharomyces cerevisiae)* are usually haploid and reproduce asexually by budding (**3**). They can live both aerobically and anaerobically. Wine is produced by other types of yeast, some of which already live on the grapes. To promote the formation of ethanol, efforts are made to generally exclude oxygen during alcoholic fermentation—for example, by covering dough with a cloth when it is rising and by fermenting liquids in barrels that exclude air.

C. Beer brewing ○

Barley is the traditional starting material for the brewing of beer. Although cereal grains contain starch, they hardly have any *free* sugars. The barley grains are therefore first allowed to germinate so that starch-cleaving *amylases* are formed. Careful warming of the sprouting grain produces **malt.** This is then ground, soaked in water, and kept warm for a certain time. In the process, a substantial proportion of the starch is broken down into the disaccharide *maltose* (see p. 38). The product (the wort) is then boiled, **yeast** and **hops** are added, and the mixture is allowed to ferment for several days. The addition of hops makes the beer less perishable and gives it its slightly bitter taste. Other substances contained in hops act as sedatives and diuretics.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

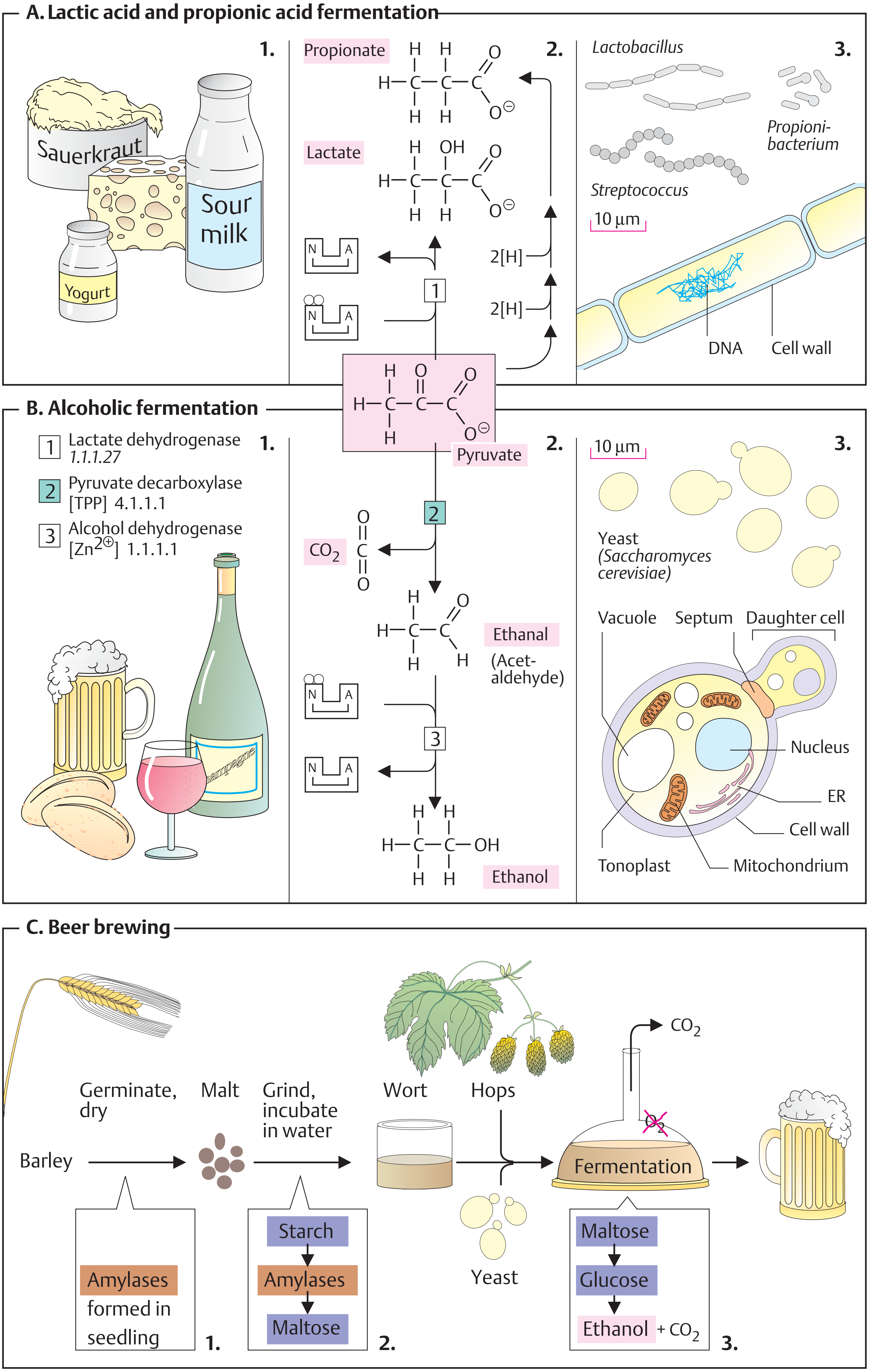

A. Lactic acid and propionic acid fermentation

1.
Sauerkraut
Sour milk
Yogurt
2.
Propionate
Lactate
2[H]
2[H]
1
Pyruvate
3.
Lactobacillus
Propioni-bacterium
Streptococcus
10 µm
DNA
Cell wall

B. Alcoholic fermentation

1.
1 Lactate dehydrogenase 1.1.1.27
2 Pyruvate decarboxylase [TPP] 4.1.1.1
3 Alcohol dehydrogenase [Zn2⊕] 1.1.1.1
Champagne
2.
2
CO2
Ethanal (Acetaldehyde)
3
Ethanol
3.
10 µm
Yeast (Saccharomyces cerevisiae)
Vacuole
Septum
Daughter cell
Nucleus
ER
Cell wall
Tonoplast
Mitochondrium

C. Beer brewing

Barley
Germinate, dry
Malt
Grind, incubate in water
Wort
Hops
Yeast
CO2
O2
Fermentation
Amylases formed in seedling
1.
Starch
Amylases
Maltose
2.
Maltose
Glucose
Ethanol + CO2
3.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Glycolysis

A. Balance ●

Glycolysis is a catabolic pathway in the cytoplasm that is found in almost all organisms—irrespective of whether they live aerobically or anaerobically. The balance of glycolysis is simple: glucose is broken down into two molecules of pyruvate, and in addition two molecules of ATP and two of NADH+H$^+$ are formed.

In the presence of oxygen, pyruvate and NADH+H$^+$ reach the mitochondria, where they undergo further transformation (**aerobic glycolysis**; see p. 146). In anaerobic conditions, *fermentation products* such as lactate or ethanol have to be formed in the cytoplasm from pyruvate and NADH+H$^+$, in order to regenerate NAD$^+$ so that glycolysis can continue (**anaerobic glycolysis**; see p. 146). In the anaerobic state, glycolysis is the only means of obtaining ATP that animal cells have.

B. Reactions ◑

Glycolysis involves ten individual steps, including three isomerizations and four phosphate transfers. The only redox reaction takes place in step [6].

[1] Glucose, which is taken up by animal cells from the blood and other sources, is first phosphorylated to **glucose 6-phosphate**, with ATP being consumed. The glucose 6-phosphate is not capable of leaving the cell.

[2] In the next step, glucose 6-phosphate is isomerized into **fructose 6-phosphate**.

[3] Using ATP again, another phosphorylation takes place, giving rise to **fructose 1,6-bisphosphate.** *Phosphofructokinase* is the most important key enzyme in glycolysis (see p. 144).

[4] Fructose 1,6-bisphosphate is broken down by *aldolase* into the C_3 compounds **glyceraldehyde 3-phosphate** (also known as glyceral 3-phosphate) and **glycerone 3-phosphate** (dihydroxyacetone 3-phosphate).

[5] The latter two products are placed in fast equilibrium by *triosephosphate isomerase.*

[6] Glyceraldehyde 3-phosphate is now oxidized by *glyceraldehyde-3-phosphate dehydrogenase,* with NADH+H$^+$ being formed. In this reaction, *inorganic phosphate* is taken up into the molecule (*substrate-level phosphorylation*; see p. 124), and **1,3-bisphosphoglycerate** is produced. This intermediate contains a *mixed acid–anhydride bond,* the phosphate part of which is at a high chemical potential.

[7] Catalyzed by *phosphoglycerate kinase,* this phosphate residue is transferred to ADP, producing **3-phosphoglycerate** and ATP. The ATP balance is thus once again in equilibrium.

[8] As a result of shifting of the remaining phosphate residue within the molecule, the isomer **2-phosphoglycerate** is formed.

[9] Elimination of water from 2-phosphoglycerate produces the phosphate ester of the *enol form* of pyruvate—**phospho*enol*pyruvate** (PEP). This reaction also raises the second phosphate residue to a high potential.

[10] In the last step, *pyruvate kinase* transfers this residue to ADP. The remaining enol pyruvate is immediately rearranged into **pyruvate**, which is much more stable. Along with step [7] and the thiokinase reaction in the tricarboxylic acid cycle (see p. 136), the pyruvate kinase reaction is one of the three reactions in animal metabolism that are able to produce ATP independently of the respiratory chain.

In glycolysis, two molecules of ATP are initially used for activation ([1], [3]). Later, two ATPs are formed *per C_3 fragment.* Overall, therefore, there is a small net gain of 2 mol ATP per mol of glucose.

C. Energy profile ○

The energy balance of metabolic pathways depends not only on the standard changes in enthalpy ΔG^0, but also on the concentrations of the metabolites (see p. 18). Fig. **C** shows the *actual* enthalpy changes ΔG for the individual steps of glycolysis in erythrocytes.

As can be seen, only three reactions ([1], [3], and [10]), are associated with large changes in free enthalpy. In these cases, the equilibrium lies well on the side of the products (see p. 18). All of the other steps are freely reversible. The same steps are also followed—in the reverse direction—in gluconeogenesis (see p. 154), with the same enzymes being activated as in glucose degradation. The non-reversible steps [1], [3], and [10] are bypassed in glucose biosynthesis (see p. 154).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Glycolysis: balance

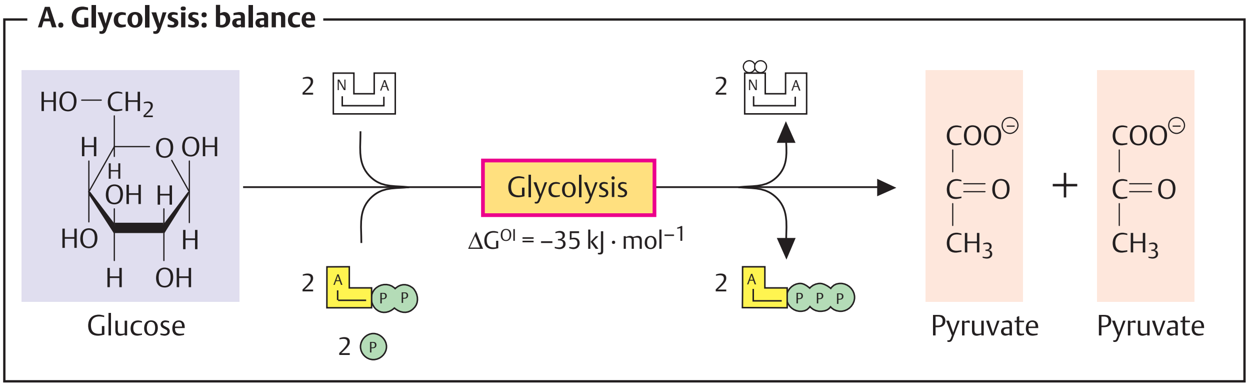
HO–CH2
H
O OH
H
OH H
HO
H
H OH
Glucose
2 NA
2 AP P
2 P
Glycolysis
ΔG°' = –35 kJ · mol⁻¹
2 NA
2 AP P P
COO⊖
C=O
CH3
Pyruvate
+
COO⊖
C=O
CH3
Pyruvate

B. Reactions

C. Energy profile

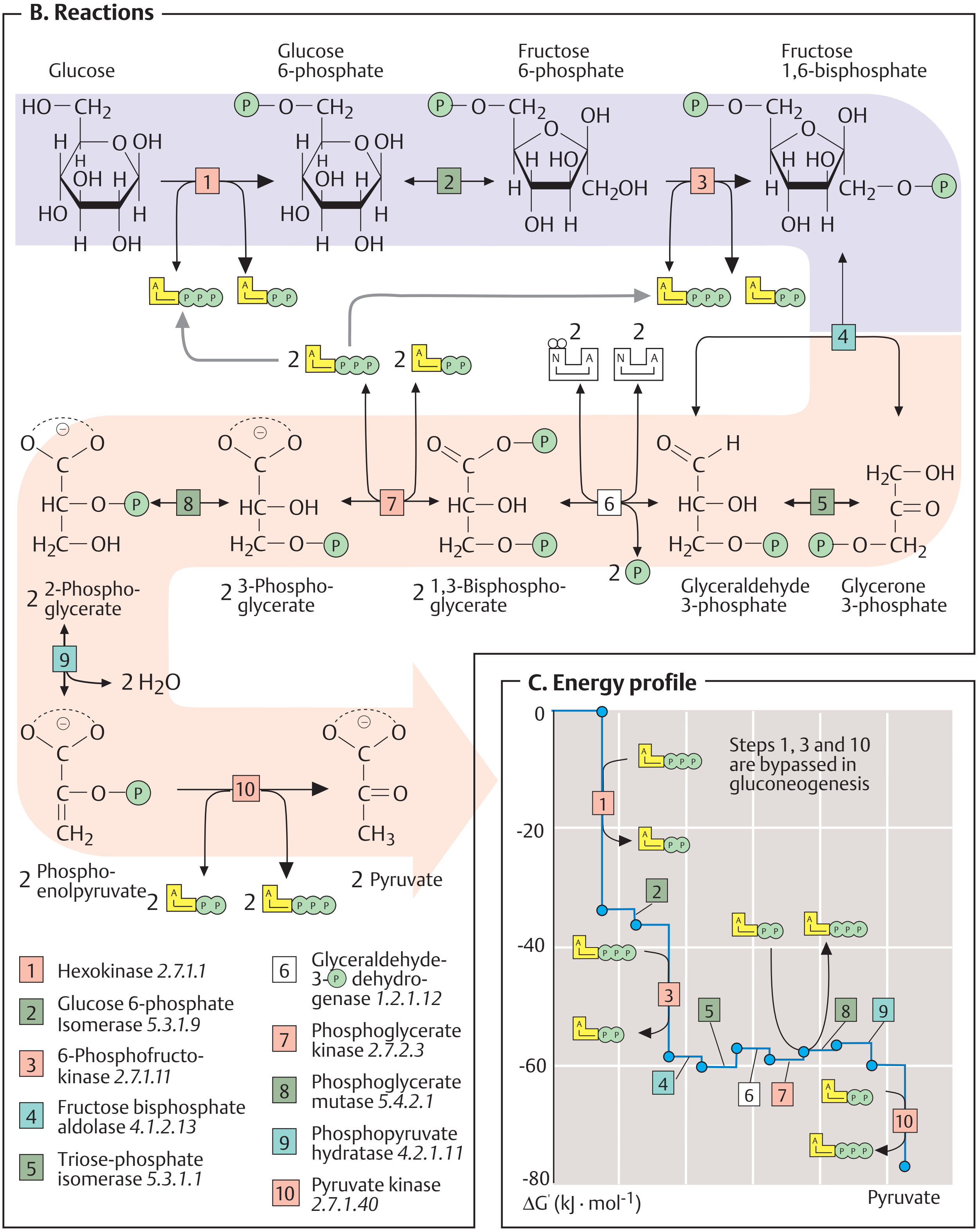
Glucose
Glucose 6-phosphate
Fructose 6-phosphate
Fructose 1,6-bisphosphate
1
2
3
4
5
6
7
8
9
10
2 H2O
2 P
2-Phosphoglycerate
3-Phosphoglycerate
1,3-Bisphosphoglycerate
Glyceraldehyde 3-phosphate
Glycerone 3-phosphate
2 Phosphoenolpyruvate
2 Pyruvate
0
-20
-40
-60
-80
ΔG' (kJ · mol⁻¹)
Pyruvate
Steps 1, 3 and 10 are bypassed in gluconeogenesis

1	Hexokinase *2.7.1.1*
2	Glucose 6-phosphate Isomerase *5.3.1.9*
3	6-Phosphofructokinase *2.7.1.11*
4	Fructose bisphosphate aldolase *4.1.2.13*
5	Triose-phosphate isomerase *5.3.1.1*
6	Glyceraldehyde-3-Ⓟ dehydrogenase *1.2.1.12*
7	Phosphoglycerate kinase *2.7.2.3*
8	Phosphoglycerate mutase *5.4.2.1*
9	Phosphopyruvate hydratase *4.2.1.11*
10	Pyruvate kinase *2.7.1.40*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Pentose phosphate pathway

The pentose phosphate pathway (PPP, also known as the *hexose monophosphate pathway)* is an oxidative metabolic pathway located in the cytoplasm, which, like glycolysis, starts from glucose 6-phosphate. It supplies two important precursors for anabolic pathways: **NADPH+H+,** which is required for the biosynthesis of fatty acids and isoprenoids, for example (see p. 168), and **ribose 5-phosphate,** a precursor in nucleotide biosynthesis (see p. 188).

A. Pentose phosphate pathway: oxidative part ◑

The **oxidative segment** of the PPP converts glucose 6-phosphate to ribulose 5-phosphate. One CO_2 and two $NADPH+H^+$ are formed in the process. Depending on the metabolic state, the much more complex **regenerative part** of the pathway (see **B**) can convert some of the pentose phosphates back to hexose phosphates, or it can pass them on to glycolysis for breakdown. In most cells, less than 10% of glucose 6-phosphate is degraded via the pentose phosphate pathway.

B. Reactions ◑

[1] The **oxidative part** starts with the oxidation of **glucose 6-phosphate** by *glucose-6-phosphate dehydrogenase.* This forms $NADPH+H^+$ for the first time. The second product, **6-phosphogluconolactone**, is an intramolecular ester (*lactone*) of 6-phosphogluconate.

[2] A specific hydrolase then cleaves the lactone, exposing the carboxyl group of **6-phosphogluconate**.

[3] The last enzyme in the oxidative part is *phosphogluconate dehydrogenase* [3], which releases the carboxylate group of 6-phosphogluconate as CO_2 and at the same time oxidizes the hydroxyl group at C_3 to an oxo group. In addition to a second $NADPH+H^+$, this also produces the ketopentose **ribulose 5-phosphate.** This is converted by an isomerase to ribose 5-phosphate, the initial compound for nucleotide synthesis (top).

The **regenerative part** of the PPP is only shown here schematically. A complete reaction scheme is given on p. 408. The function of the regenerative branch is to adjust the *net* production of $NADPH+H^+$ and pentose phosphates to the cell's current requirements. Normally, the demand for $NADPH+H^+$ is much higher than that for pentose phosphates. In these conditions, the reaction steps shown first convert six ribulose 5-phosphates to five molecules of fructose 6-phosphate and then, by isomerization, regenerate five glucose 6-phosphates. These can once again supply $NADPH+H^+$ to the oxidative part of the PPP. Repeating these reactions finally results in the oxidation of one glucose 6-phosphate into six CO_2. Twelve $NADPH+H^+$ arise in the same process. In sum, no pentose phosphates are produced via this pathway.

In the recombination of sugar phosphates in the regenerative part of the PPP, there are two enzymes that are particularly important:

[5] *Transaldolase* transfers C_3 units from sedoheptulose 7-phosphate, a ketose with seven C atoms, to the aldehyde group of glyceraldehyde 3-phosphate.

[4] *Transketolase,* which contains thiamine diphosphate, transfers C_2 fragments from one sugar phosphate to another.

The reactions in the regenerative segment of the PPP are freely reversible. It is therefore easily possible to use the regenerative part of the pathway to convert hexose phosphates into pentose phosphates. This can occur when there is a high demand for pentose phosphates—e. g., during DNA replication in the S phase of the cell cycle (see p. 394).

Additional information

When energy in the form of ATP is required in addition to $NADPH+H^+$, the cell is able to channel the products of the regenerative part of the PPP (fructose 6-phosphate and glyceraldehyde 3-phosphate) into glycolysis. Further degradation is carried out via the tricarboxylic acid cycle and the respiratory chain to CO_2 and water. Overall, the cell in this way obtains 12 mol $NADPH+H^+$ and around 150 mol ATP from 6 mol glucose 6-phosphate. PPP activity is stimulated by *insulin* (see p. 388). This not only increases the rate of glucose degradation, but also produces additional $NADPH+H^+$ for fatty acid synthesis (see p. 168).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Pentose phosphate pathway: oxidative part

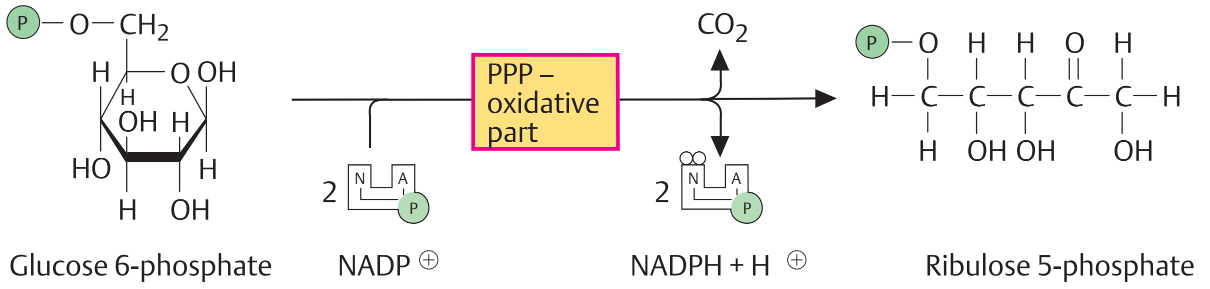
PPP – oxidative part
CO2
2 NADP⊕
2 NADPH + H⊕
Glucose 6-phosphate
Ribulose 5-phosphate

B. Reactions

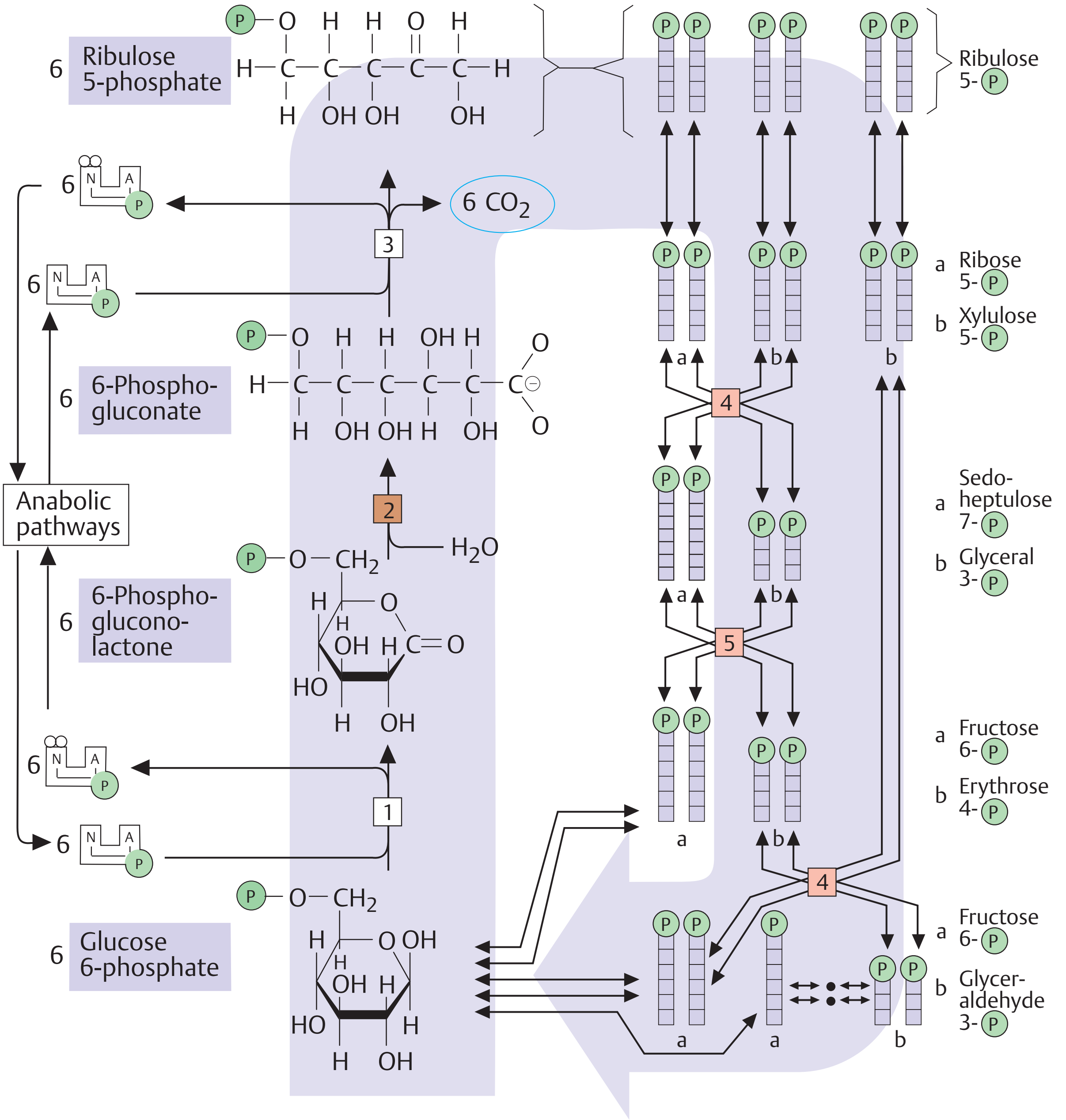
6 Ribulose 5-phosphate
6 CO2
3
6 6-Phospho-gluconate
2
H2O
Anabolic pathways
6 6-Phospho-glucono-lactone
1
6 Glucose 6-phosphate
4
5
a
b
Ribulose 5-P
a Ribose 5-P
b Xylulose 5-P
a Sedo-heptulose 7-P
b Glyceral 3-P
a Fructose 6-P
b Erythrose 4-P
a Fructose 6-P
b Glycer-aldehyde 3-P

1 Glucose 6-phosphate dehydrogenase *1.1.1.49*

2 Gluconolactonase *3.1.1.17*

3 Phosphogluconate dehydrogenase (decarboxylating) *1.1.1.44*

4 Transketolase *2.2.1.1*

5 Transaldolase *2.2.1.2*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Gluconeogenesis

Some tissues, such as *brain* and *erythrocytes*, depend on a constant supply of glucose. If the amount of carbohydrate taken up in food is not sufficient, the blood sugar level can be maintained for a limited time by *degradation of hepatic glycogen* (see p. 156). If these reserves are also exhausted, de-novo synthesis of glucose (**gluconeogenesis**) begins. The **liver** is also mainly responsible for this (see p. 310), but the tubular cells of the **kidney** also show a high level of gluconeogenetic activity (see p. 328). The main precursors for gluconeogenesis are **amino acids** derived from muscle proteins. Another important precursor is **lactate,** which is formed in erythrocytes and muscle proteins when there is oxygen deficiency. **Glycerol** produced from the degradation of fats can also be used for gluconeogenesis. However, the conversion of fatty acids into glucose is *not* possible in animal metabolism (see p. 138). The human organism can synthesize several hundred grams of glucose per day by gluconeogenesis.

A. Gluconeogenesis ◑

Many of the reaction steps involved in gluconeogenesis are catalyzed by the same enzymes that are used in glycolysis (see p. 150). Other enzymes are specific to gluconeogenesis and are only synthesized, under the influence of *cortisol* and *glucagon* when needed (see p. 158). Glycolysis takes place exclusively when needed in the cytoplasm, but gluconeogenesis also involves the *mitochondria* and the *endoplasmic reticulum* (ER). Gluconeogenesis consumes 4 ATP (3 ATP + 1 GTP) per glucose—i. e., twice as many as glycolysis produces.

[1] **Lactate** as a precursor for gluconeogenesis is mainly derived from muscle (see Cori cycle, p. 338) and erythrocytes. LDH (see p. 98) oxidizes lactate to pyruvate, with $NADH+H^+$ formation.

[2] The first steps of actual gluconeogenesis take place in the *mitochondria.* The reason for this “detour” is the equilibrium state of the pyruvate kinase reaction (see p. 150). Even coupling to ATP hydrolysis would not be sufficient to convert pyruvate *directly* into phospho*enol* pyruvate (PEP). **Pyruvate** derived from lactate or amino acids is therefore initially transported into the mitochondrial matrix, and—in a biotin-dependent reaction catalyzed by *pyruvate carboxylase*—is carboxylated there to **oxaloacetate**. Oxaloacetate is also an intermediate in the tricarboxylic acid cycle. *Amino acids* with breakdown products that enter the cycle or supply pyruvate can therefore be converted into glucose (see p. 180).

[3] The oxaloacetate formed in the mitochondrial matrix is initially reduced to **malate**, which can leave the mitochondria via inner membrane transport systems (see p. 212).

[4] In the cytoplasm, oxaloacetate is reformed and then converted into **phosphoenol pyruvate** by a GTP-dependent *PEP carboxykinase.* The subsequent steps up to fructose 1,6-bisphosphate represent the reverse of the corresponding reactions involved in glycolysis. One additional ATP per C_3 fragment is used for the synthesis of 1,3-bisphosphoglycerate.

Two gluconeogenesis-specific phosphatases then successively cleave off the phosphate residues from **fructose 1,6-bisphosphate**. In between these reactions lies the isomerization of fructose 6-phosphate to **glucose 6-phosphate**—another glycolytic reaction.

[5] The reaction catalyzed by *fructose 1,6-bisphosphatase* is an important regulation point in gluconeogenesis (see p. 158).

[6] The last enzyme in the pathway, *glucose 6-phosphatase,* occurs in the liver, but not in muscle. It is located in the interior of the smooth endoplasmic reticulum. Specific transporters allow glucose 6-phosphate to enter the ER and allow the **glucose** formed there to return to the cytoplasm. From there, it is ultimately released into the blood.

Glycerol initially undergoes phosphorylation at C-3 [7]. The **glycerol 3-phosphate** formed is then oxidized by an NAD^+-dependent dehydrogenase to form **glycerone 3-phosphate** [8] and thereby channeled into gluconeogenesis. An FAD-dependent mitochondrial enzyme is also able to catalyze this reaction (known as the “glycerophosphate shuttle”; see p. 212).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Gluconeogenesis

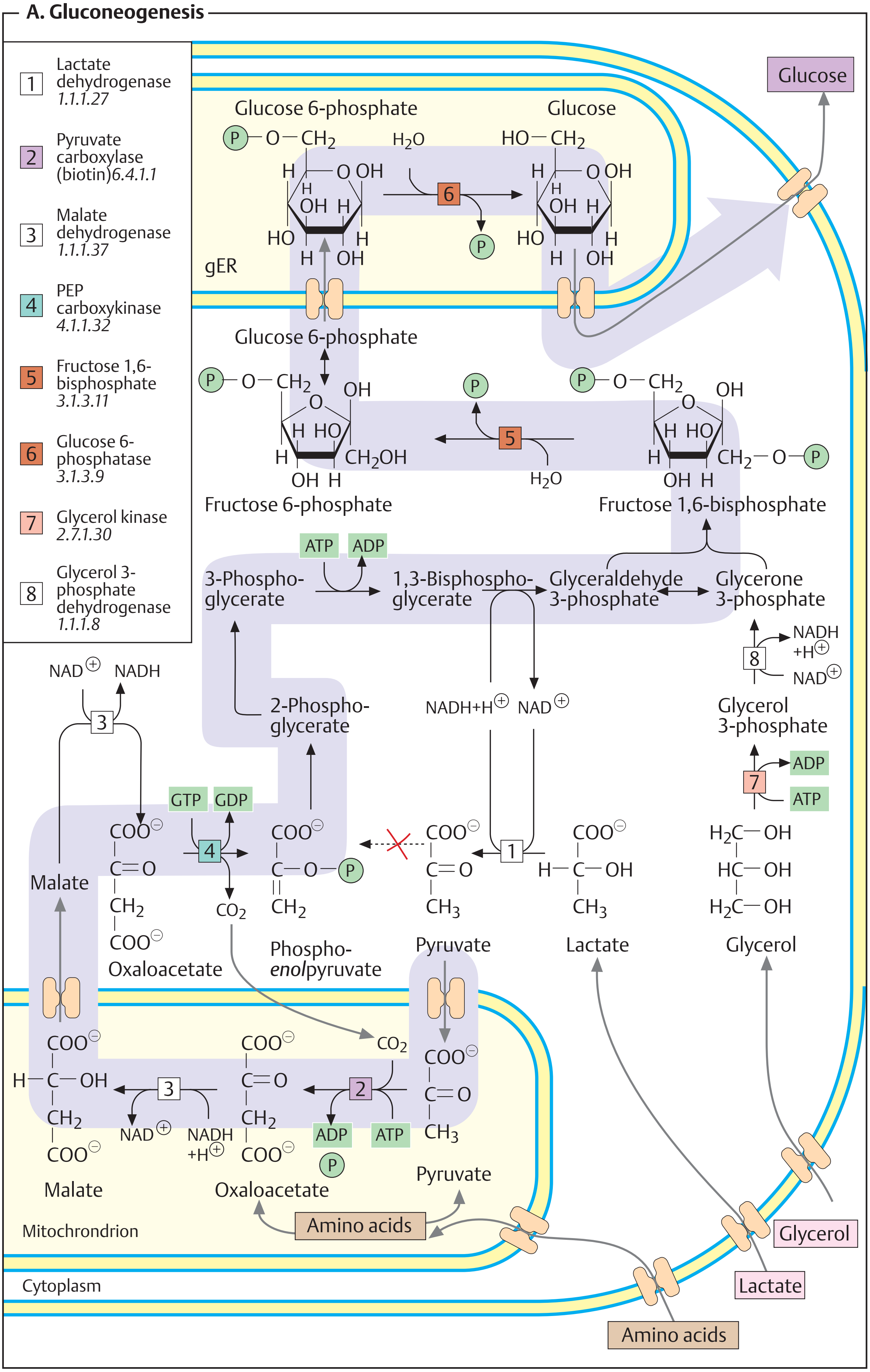
1 Lactate dehydrogenase 1.1.1.27
2 Pyruvate carboxylase (biotin) 6.4.1.1
3 Malate dehydrogenase 1.1.1.37
4 PEP carboxykinase 4.1.1.32
5 Fructose 1,6-bisphosphate 3.1.3.11
6 Glucose 6-phosphatase 3.1.3.9
7 Glycerol kinase 2.7.1.30
8 Glycerol 3-phosphate dehydrogenase 1.1.1.8
Glucose
Glucose 6-phosphate
Glucose
H_2O
P
gER
Glucose 6-phosphate
Fructose 6-phosphate
Fructose 1,6-bisphosphate
H_2O
ATP
ADP
3-Phosphoglycerate
1,3-Bisphosphoglycerate
Glyceraldehyde 3-phosphate
Glycerone 3-phosphate
NADH +H⊕
NAD⊕
Glycerol 3-phosphate
ADP
ATP
2-Phosphoglycerate
NADH+H⊕
NAD⊕
NAD⊕
NADH
GTP
GDP
CO_2
Malate
Oxaloacetate
Phosphoenolpyruvate
Pyruvate
Lactate
Glycerol
CO_2
Malate
NAD⊕
NADH +H⊕
Oxaloacetate
ADP
ATP
P
Pyruvate
Amino acids
Mitochondrion
Cytoplasm
Glycerol
Lactate
Amino acids

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Glycogen metabolism

Glycogen (see p. 40) is used in animals as a **carbohydrate reserve,** from which glucose phosphates and glucose can be released when needed. Glucose storage itself would not be useful, as high concentrations within cells would make them strongly hypertonic and would therefore cause an influx of water. By contrast, insoluble glycogen has only low osmotic activity.

A. Glycogen balance ◑

Animal glycogen, like amylopectin in plants, is a *branched homopolymer of glucose.* The glucose residues are linked by an $\alpha 1 \rightarrow 4$-glycosidic bond. Every tenth or so glucose residue has an additional $\alpha 1 \rightarrow 6$ bond to another glucose. These branches are extended by additional $\alpha 1 \rightarrow 4$-linked glucose residues. This structure produces tree-shaped molecules consisting of up to 50 000 residues (M > 1 10^7 Da).

Hepatic glycogen is never completely degraded. In general, only the nonreducing ends of the "tree" are shortened, or—when glucose is abundant—elongated. The reducing end of the tree is linked to a special protein, **glycogenin**. Glycogenin carries out autocatalytic covalent bonding of the first glucose at one of its tyrosine residues and elongation of this by up to seven additional glucose residues. It is only at this point that *glycogen synthase* becomes active to supply further elongation.

[1] The formation of glycosidic bonds between sugars is *endergonic.* Initially, therefore, the activated form—**UDP-glucose**—is synthesized by reaction of glucose 1-phosphate with UTP (see p. 110).

[2] *Glycogen synthase* now transfers glucose residues one by one from UDP-glucose to the non-reducing ends of the available "branches."

[3] Once the growing chain has reached a specific length (> 11 residues), the *branching enzyme* cleaves an oligosaccharide consisting of 6–7 residues from the end of it, and adds this into the interior of the same chain or a neighboring one with $\alpha 1 \rightarrow 6$ linkage. These **branches** are then further extended by glycogen synthase.

[4] The branched structure of glycogen allows rapid release of sugar residues. The most important degradative enzyme, *glycogen phosphorylase,* cleaves residues from a non-reducing end one after another as **glucose 1-phosphate**. The larger the number of these ends, the more phosphorylase molecules can attack simultaneously. The formation of glucose 1-phosphate instead of glucose has the advantage that no ATP is needed to channel the released residues into glycolysis or the PPP.

[5] [6] Due to the structure of glycogen phosphorylase, degradation comes to a halt four residues away from each branching point. Two more enzymes overcome this blockage. First, a *glucanotransferase* moves a trisaccharide from the side chain to the end of the main chain [5]. A *1,6-glucosidase* [6] then cleaves the single remaining residue as a free glucose and leaves behind an unbranched chain that is once again accessible to phosphorylase.

The **regulation of glycogen metabolism** by interconversion, and the role of hormones in these processes, are discussed on p. 120.

B. Glycogen balance ◑

The human organism can store up to 450 g of glycogen—one-third in the **liver** and almost all of the remainder in **muscle.** The glycogen content of the other organs is low.

Hepatic glycogen is mainly used to maintain the *blood glucose level* in the postresorptive phase (see p. 308). The glycogen content of the liver therefore varies widely, and can decline to almost zero in periods of extended hunger. After this, gluconeogenesis (see p. 154) takes over the glucose supply for the organism. *Muscle glycogen* serves as an *energy reserve* and is not involved in blood glucose regulation. Muscle does not contain any glucose 6-phosphatase and is therefore unable to release glucose into the blood. The glycogen content of muscle therefore does not fluctuate as widely as that of the liver.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Glycogen metabolism

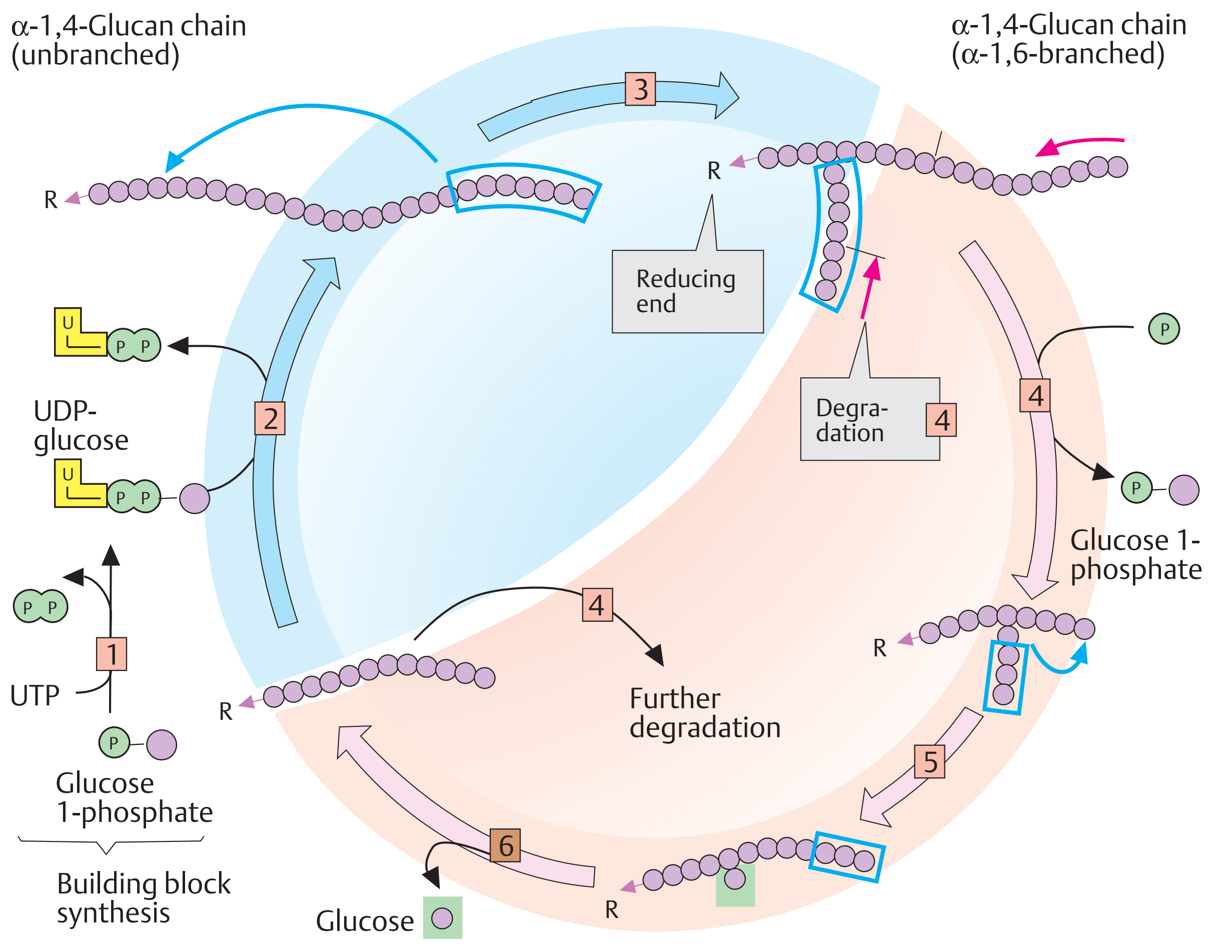
α-1,4-Glucan chain
(unbranched)
α-1,4-Glucan chain
(α-1,6-branched)
R
3
R
Reducing end
2
UDP-glucose
U
P P
Degra-dation 4
4
P
P
Glucose 1-phosphate
1
UTP
P P
P
Glucose 1-phosphate
Building block synthesis
4
Further degradation
R
R
5
6
Glucose
R

- 1 UTP-glucose 1-phosphate uridyltransferase *2.7.7.9*
- 2 Glycogen synthase *2.4.1.11*
- 3 Glucan branching enzyme *2.4.1.18*
- 4 Phosphorylase *2.4.1.1*
- 5 4-α-Glucanotransferase *2.4.1.25*
- 6 Amylo-1,6-glucosidase *3.2.1.33*

B. Glycogen balance

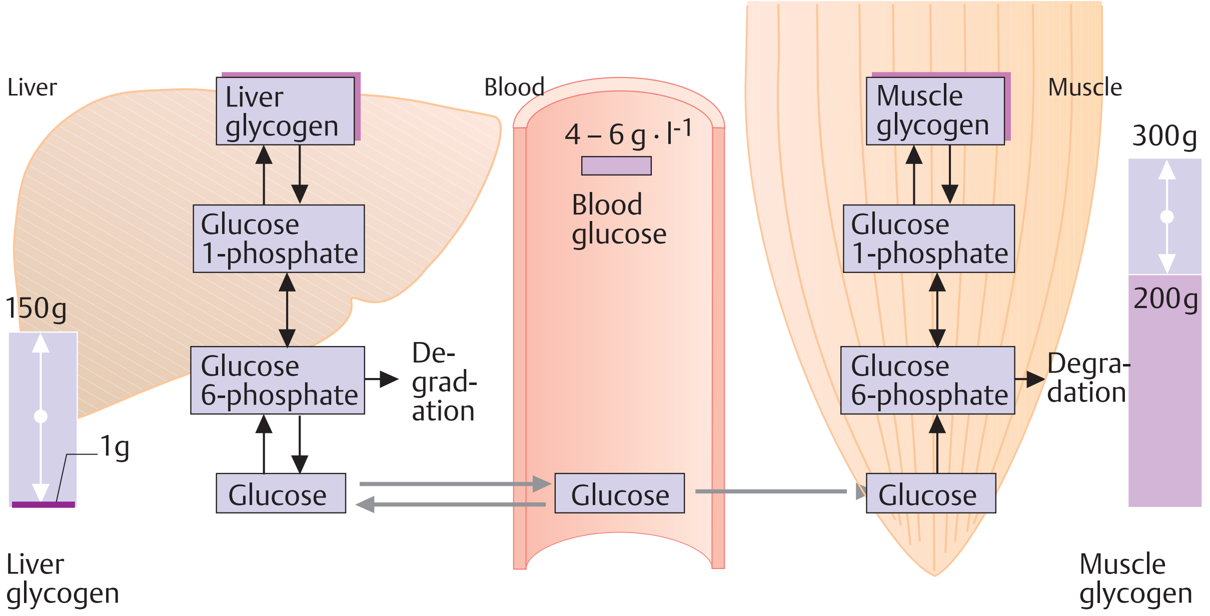
Liver
Liver glycogen
Glucose 1-phosphate
150g
Glucose 6-phosphate
De-grad-ation
1g
Glucose
Liver glycogen
Blood
4 – 6 g · l⁻¹
Blood glucose
Glucose
Muscle glycogen
Muscle
300g
Glucose 1-phosphate
200g
Glucose 6-phosphate
Degra-dation
Glucose
Muscle glycogen

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Regulation

A. Regulation of carbohydrate metabolism ◑

In all organisms, carbohydrate metabolism is subject to complex regulatory mechanisms involving *hormones, metabolites*, and *coenzymes.* The scheme shown here (still a simplified one) applies to the liver, which has central functions in carbohydrate metabolism (see p. 306). Some of the control mechanisms shown here are not effective in other tissues.

One of the liver's most important tasks is to store excess glucose in the form of glycogen and to release glucose from glycogen when required *(buffer function).* When the glycogen reserves are exhausted, the liver can provide glucose by de novo synthesis (*gluconeogenesis*; see p. 154). In addition, like all tissues, the liver breaks glucose down via glycolysis. These functions have to be coordinated with each other. For example, there is no point in glycolysis and gluconeogenesis taking place *simultaneously*, and glycogen synthesis and glycogen degradation should not occur simultaneously either. This is ensured by the fact that two *different* enzymes exist for important steps in both pathways, each of which catalyzes only the anabolic or the catabolic reaction. The enzymes are also regulated differently. Only these key enzymes are shown here.

Hormones. The hormones that influence carbohydrate metabolism include the peptides insulin and glucagon; a glucocorticoid, cortisol; and a catecholamine, epinephrine (see p. 380). **Insulin** activates *glycogen synthase* ([1]; see p. 388), and induces several enzymes involved in glycolysis [3, 5, 7]. At the same time, insulin inhibits the synthesis of enzymes involved in gluconeogenesis (*repression;* [4, 6, 8, 9]). **Glucagon,** the antagonist of insulin, has the opposite effect. It induces gluconeogenesis enzymes [4, 6, 8, 9] and represses *pyruvate kinase* [7], a key enzyme of glycolysis. Additional effects of glucagon are based on the *interconversion* of enzymes and are mediated by the second messenger cAMP. This inhibits glycogen synthesis [1] and activates glycogenolysis [2]. Epinephrine acts in a similar fashion. The inhibition of *pyruvate kinase* [7] by glucagon is also due to interconversion.

Glucocorticoids—mainly **cortisol** (see p. 374)—induce all of the key enzymes involved in gluconeogenesis [4, 6, 8, 9]. At the same time, they also induce enzymes involved in amino acid degradation and thereby provide precursors for gluconeogenesis. Regulation of the expression of *PEP carboxykinase,* a key enzyme in gluconeogenesis, is discussed in detail on p. 244.

Metabolites. High concentrations of **ATP** and **citrate** inhibit glycolysis by allosteric regulation of *phosphofructokinase.* ATP also inhibits *pyruvate kinase.* **Acetyl-CoA,** an inhibitor of *pyruvate kinase,* has a similar effect. All of these metabolites arise from glucose degradation *(feedback inhibition).* **AMP** and **ADP**, signals for ATP deficiency, activate glycogen degradation and inhibit gluconeogenesis.

B. Fructose 2,6-bisphosphate ◑

Fructose 2,6-bisphosphate (Fru-2,6-bP) plays an important part in carbohydrate metabolism. This metabolite is formed in small quantities from fructose 6-phosphate and has purely *regulatory functions.* It stimulates glycolysis by allosteric activation of *phosphofructokinase* and inhibits gluconeogenesis by inhibition of *fructose 1,6-bisphosphatase.*

The synthesis and degradation of Fru-2,6-bP are catalyzed by one and the same protein [10a, 10b]. If the enzyme is present in an unphosphorylated form [10a], it acts as a kinase and leads to the formation of Fru-2,6-bP. After phosphorylation by cAMP-dependent protein kinase A (PK-A), it acts as a phosphatase [10b] and now catalyzes the degradation of Fru-2,6-bP to fructose 6-phosphate. The equilibrium between [10a] and [10b] is regulated by hormones. Epinephrine and glucagon increase the cAMP level (see p. 120). As a result of increased PK-A activity, this reduces the Fru-2,6-bP concentration and inhibits glycolysis, while at the same time activating gluconeogenesis. Conversely, via [10a], insulin activates the synthesis of Fru-2,6-bP and thus glycolysis. In addition, insulin also inhibits the action of glucagon by reducing the cAMP level (see p. 120).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Regulation of carbohydrate metabolism

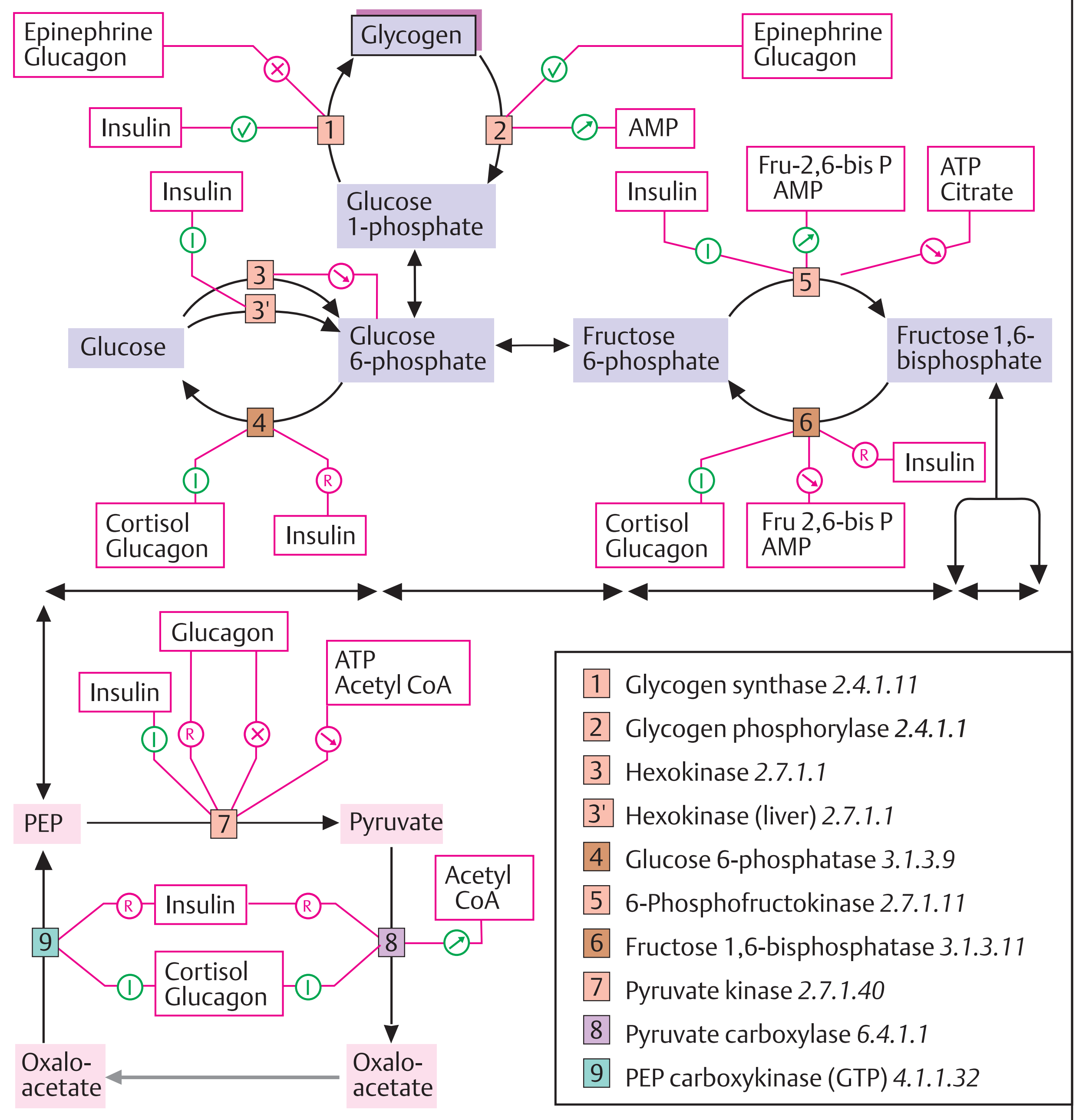
Epinephrine
Glucagon
Glycogen
Epinephrine
Glucagon
Insulin
1
2
AMP
Insulin
Glucose 1-phosphate
Fru-2,6-bis P
AMP
ATP
Citrate
Insulin
3
3'
5
Glucose
Glucose 6-phosphate
Fructose 6-phosphate
Fructose 1,6-bisphosphate
4
6
Insulin
Cortisol
Glucagon
Insulin
Cortisol
Glucagon
Fru 2,6-bis P
AMP
Glucagon
ATP
Acetyl CoA
Insulin
PEP
7
Pyruvate
Acetyl CoA
Insulin
9
8
Cortisol
Glucagon
Oxalo-acetate
Oxalo-acetate

1	Glycogen synthase *2.4.1.11*
2	Glycogen phosphorylase *2.4.1.1*
3	Hexokinase *2.7.1.1*
3'	Hexokinase (liver) *2.7.1.1*
4	Glucose 6-phosphatase *3.1.3.9*
5	6-Phosphofructokinase *2.7.1.11*
6	Fructose 1,6-bisphosphatase *3.1.3.11*
7	Pyruvate kinase *2.7.1.40*
8	Pyruvate carboxylase *6.4.1.1*
9	PEP carboxykinase (GTP) *4.1.1.32*

B. Fructose 2,6-bisphosphate

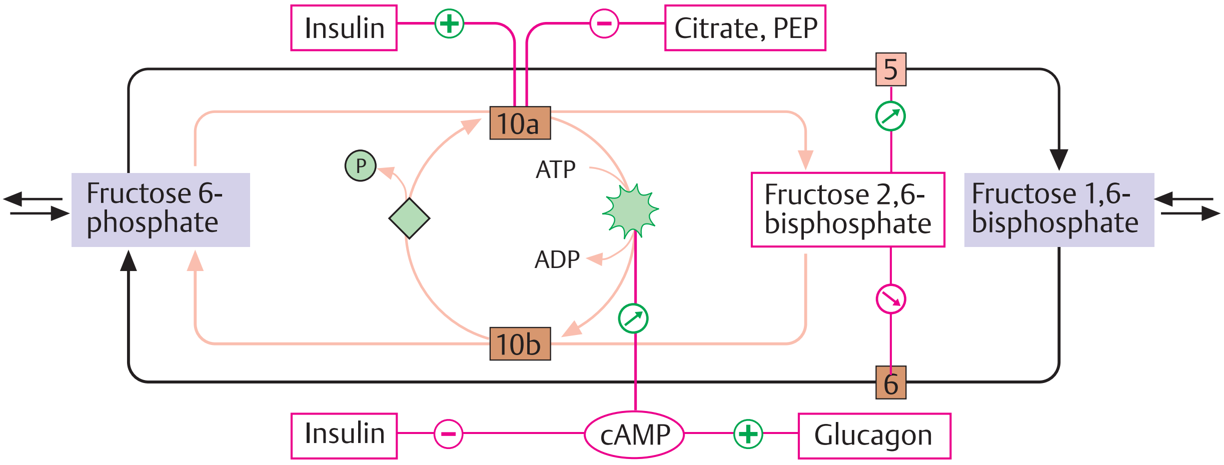
Insulin
Citrate, PEP
5
10a
Fructose 6-phosphate
P
ATP
Fructose 2,6-bisphosphate
Fructose 1,6-bisphosphate
ADP
10b
6
Insulin
cAMP
Glucagon

10a	6-Phosphofructo-2-kinase *2.7.1.105*
10b	Fructose 2,6-bisphosphatase *3.1.3.46*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Diabetes mellitus

Diabetes mellitus is a very common metabolic disease that is caused by absolute or relative insulin deficiency. The lack of this peptide hormone (see p. 76) mainly affects carbohydrate and lipid metabolism. Diabetes mellitus occurs in two forms. In **type 1** diabetes (insulin-dependent diabetes mellitus, IDDM), the insulin-forming cells are destroyed in young individuals by an autoimmune reaction. The less severe **type 2** diabetes (non-insulin-dependent diabetes mellitus, NIDDM) usually has its first onset in elderly individuals. The causes have not yet been explained in detail in this type.

A. Insulin biosynthesis ○

Insulin is produced by the B cells of the *islets of Langerhans* in the pancreas. As is usual with secretory proteins, the hormone's precursor (*preproinsulin*) carries a signal peptide that directs the peptide chain to the interior of the endoplasmic reticulum (see p. 210). *Proinsulin* is produced in the ER by cleavage of the signal peptide and formation of disulfide bonds. Proinsulin passes to the Golgi apparatus, where it is packed into vesicles—the β-*granules.* After cleavage of the *C peptide*, *mature insulin* is formed in the β-granules and is stored in the form of zinc-containing hexamers until secretion.

B. Effects of insulin deficiency ◑

The effects of insulin on **carbohydrate metabolism** are discussed on p. 158. In simplified terms, they can be described as *stimulation of glucose utilization* and *inhibition of gluconeogenesis*. In addition, the transport of glucose from the blood into most tissues is also insulin-dependent (exceptions to this include the liver, CNS, and erythrocytes).

The **lipid metabolism** of adipose tissue is also influenced by the hormone. In these cells, insulin stimulates the reorganization of glucose into fatty acids. This is mainly based on activation of *acetyl CoA carboxylase* (see p. 162) and increased availability of NADPH+H^+ due to increased PPP activity (see p. 152). On the other hand, insulin also inhibits the degradation of fat by hormone-sensitive lipases (see p. 162) and prevents the breakdown of muscle protein.

The effects of insulin *deficiency* on metabolism are shown by arrows in the illustration. Particularly noticeable is the increase in the glucose concentration in the blood, from 5 mM to 9 mM (90 mg dL^{-1}) or more (**hyperglycemia,** elevated blood glucose level). In *muscle* and *adipose tissue* – the two most important glucose consumers—glucose uptake and glucose utilization are impaired by insulin deficiency. Glucose utilization in the *liver* is also reduced. At the same time, gluconeogenesis is stimulated, partly due to increased proteolysis in the muscles. This increases the blood sugar level still further. When the capacity of the *kidneys* to resorb glucose is exceeded (at plasma concentrations of 9 mM or more), glucose is excreted in the urine (**glucosuria**).

The increased degradation of fat that occurs in insulin deficiency also has serious effects. Some of the fatty acids that accumulate in large quantities are taken up by the liver and used for lipoprotein synthesis **(hyperlipidemia)**, and the rest are broken down into acetyl CoA. As the tricarboxylic acid cycle is not capable of taking up such large quantities of acetyl CoA, the excess is used to form **ketone bodies** (*acetoacetate* and *β-hydroxybutyrate* see p. 312). As H^+ ions are released in this process, diabetics not receiving adequate treatment can suffer severe **metabolic acidosis** (diabetic coma). The *acetone* that is also formed gives these patients' breath a characteristic odor. In addition, large amounts of ketone body anions appear in the urine (**ketonuria**).

Diabetes mellitus can have serious secondary effects. A constantly raised blood sugar level can lead in the long term to changes in the blood vessels (diabetic angiopathy), kidney damage (nephropathy) and damage to the nervous system (neuropathy), as well as to cataracts in the eyes.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

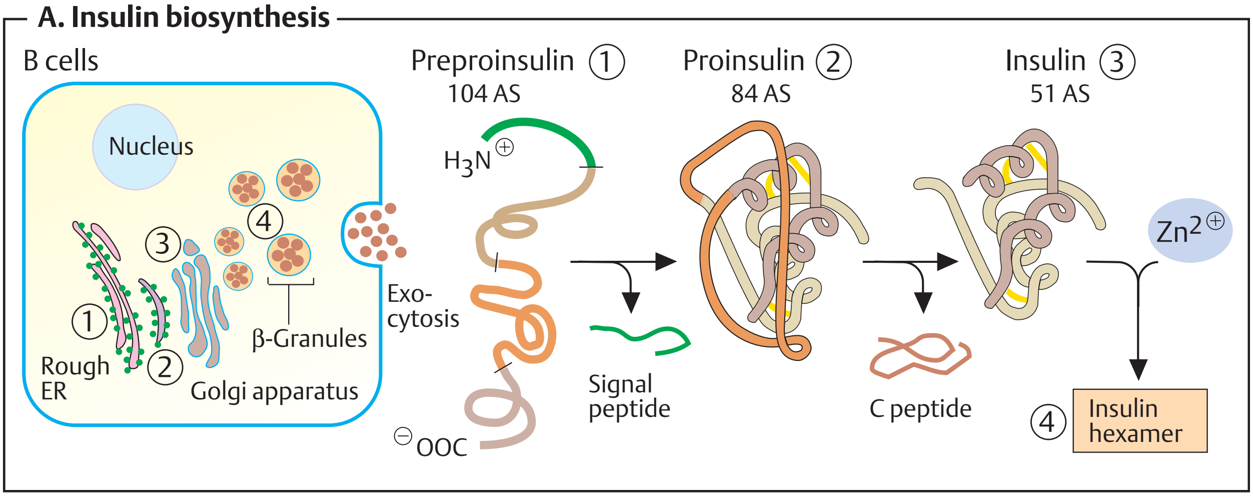

A. Insulin biosynthesis

B cells
Nucleus
① Rough ER
② Golgi apparatus
③
④ β-Granules
Exocytosis
Preproinsulin ① 104 AS
H_3N^+
^-OOC
Signal peptide
Proinsulin ② 84 AS
C peptide
Insulin ③ 51 AS
Zn^{2+}
④ Insulin hexamer

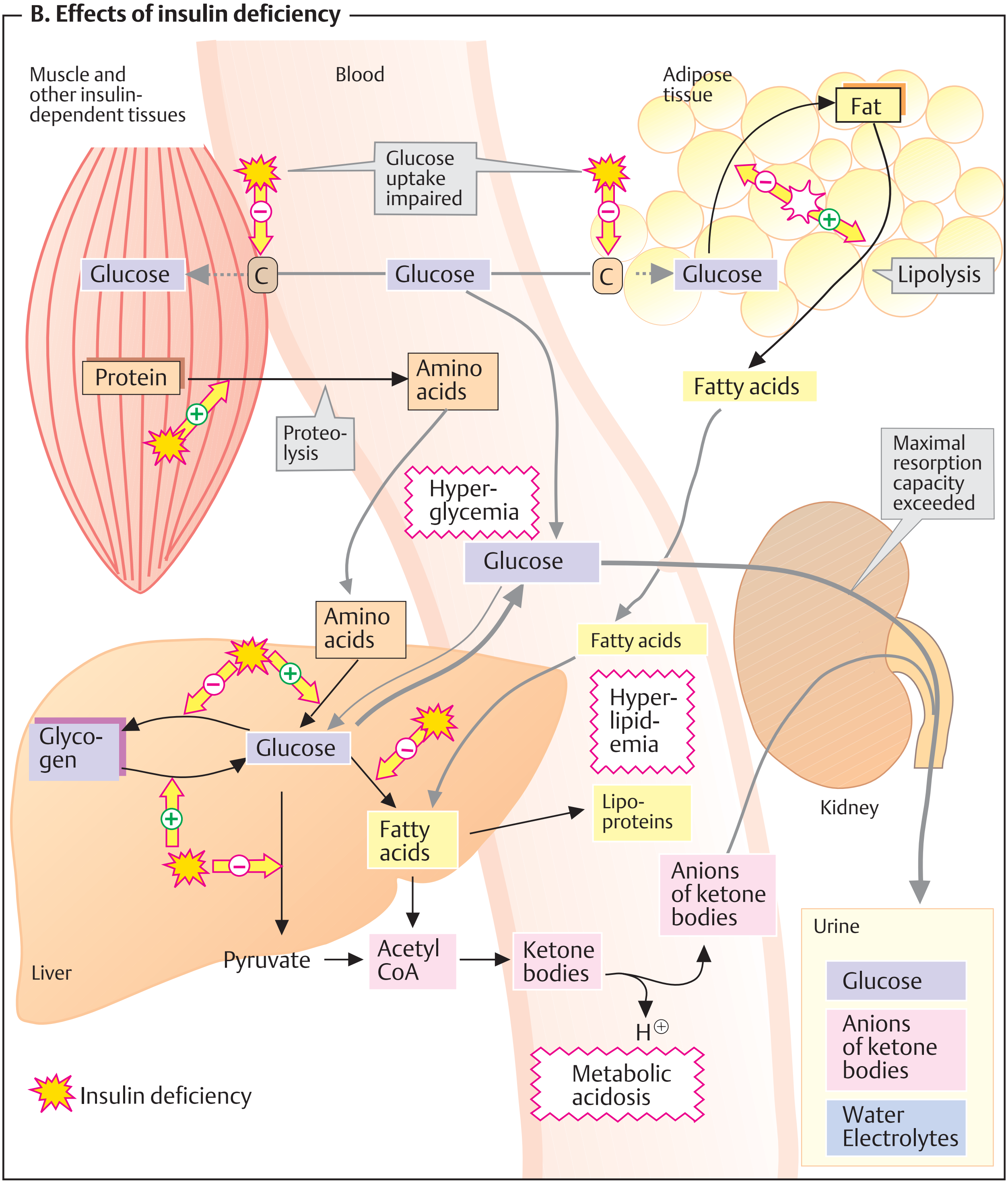

B. Effects of insulin deficiency

Muscle and other insulin-dependent tissues
Blood
Adipose tissue
Fat
Glucose uptake impaired
Glucose
C
Glucose
C
Glucose
Lipolysis
Protein
Amino acids
Proteolysis
Fatty acids
Hyperglycemia
Glucose
Maximal resorption capacity exceeded
Amino acids
Fatty acids
Hyperlipidemia
Glycogen
Glucose
Fatty acids
Lipoproteins
Kidney
Anions of ketone bodies
Pyruvate
Acetyl CoA
Ketone bodies
H^+
Metabolic acidosis
Liver
Urine
Glucose
Anions of ketone bodies
Water Electrolytes
Insulin deficiency

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Overview

A. Fat metabolism ◑

Fat metabolism in adipose tissue (top). Fats (triacylglycerols) are the most important energy reserve in the animal organism. They are mostly stored in insoluble form in the cells of adipose tissue—the *adipocytes*—where they are constantly being synthesized and broken down again.

As precursors for the biosynthesis of fats (**lipogenesis**), the adipocytes use triacylglycerols from lipoproteins (VLDLs and chylomicrons; see p. 278), which are formed in the liver and intestines and delivered by the blood. *Lipoprotein lipase* [1], which is located on the inner surface of the blood capillaries, cleaves these triacylglycerols into glycerol and fatty acids, which are taken up by the adipocytes and converted back into fats.

The degradation of fats (**lipolysis**) is catalyzed in adipocytes by *hormone-sensitive lipase* [2]—an enzyme that is regulated by various hormones by cAMP-dependent *interconversion* (see p. 120). The amount of fatty acids released depends on the activity of this lipase; in this way, the enzyme regulates the plasma levels of fatty acids.

In the blood plasma, fatty acids are transported in free form—i. e., non-esterified. Only short-chain fatty acids are soluble in the blood; longer, less water-soluble fatty acids are transported bound to albumin.

Degradation of fatty acids in the liver (left). Many tissues take up fatty acids from the blood plasma in order to synthesize fats or to obtain energy by oxidizing them. The metabolism of fatty acids is particularly intensive in the hepatocytes in the liver.

The most important process in the degradation of fatty acids is **β-oxidation**—a metabolic pathway in the mitochondrial matrix (see p. 164). Initially, the fatty acids in the cytoplasm are activated by binding to coenzyme A into **acyl CoA** [3]. Then, with the help of a transport system (the carnitine shuttle [4]; see p. 164), the activated fatty acids enter the mitochondrial matrix, where they are broken down into **acetyl CoA**. The resulting acetyl residues can be oxidized to CO_2 in the tricarboxylic acid cycle, producing reduced coenzyme and ATP derived from it by oxidative phosphorylation. If acetyl CoA production exceeds the energy requirements of the hepatocytes—as is the case when there is a high level of fatty acids in the blood plasma (typical in hunger and diabetes mellitus)—then the excess is converted into **ketone bodies** (see p. 312). These serve exclusively to supply other tissues with energy.

Fat synthesis in the liver (right). Fatty acids and fats are mainly synthesized in the liver and in adipose tissue, as well as in the kidneys, lungs, and mammary glands. Fatty acid biosynthesis occurs in the cytoplasm—in contrast to fatty acid degradation. The most important precursor is **glucose**, but certain amino acids can also be used.

The first step is carboxylation of **acetyl CoA** to **malonyl CoA**. This reaction is catalyzed by *acetyl-CoA carboxylase* [5], which is the *key enzyme in fatty acid biosynthesis.* Synthesis into fatty acids is carried out by *fatty acid synthase* [6]. This multifunctional enzyme (see p. 168) starts with one molecule of acetyl-CoA and elongates it by adding malonyl groups in seven reaction cycles until palmitate is reached. One CO_2 molecule is released in each reaction cycle. The fatty acid therefore grows by two carbon units each time. $NADPH+H^+$ is used as the reducing agent and is derived either from the *pentose phosphate pathway* (see p. 152) or from *isocitrate dehydrogenase* and *malic enzyme* reactions.

The elongation of the fatty acid by *fatty acid synthase* concludes at C_{16}, and the product, **palmitate** (16:0), is released. Unsaturated fatty acids and long-chain fatty acids can arise from palmitate in subsequent reactions. Fats are finally synthesized from activated fatty acids (acyl CoA) and glycerol 3-phosphate (see p. 170). To supply peripheral tissues, fats are packed by the hepatocytes into lipoprotein complexes of the VLDL type and released into the blood in this form (see p. 278).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Fat metabolism

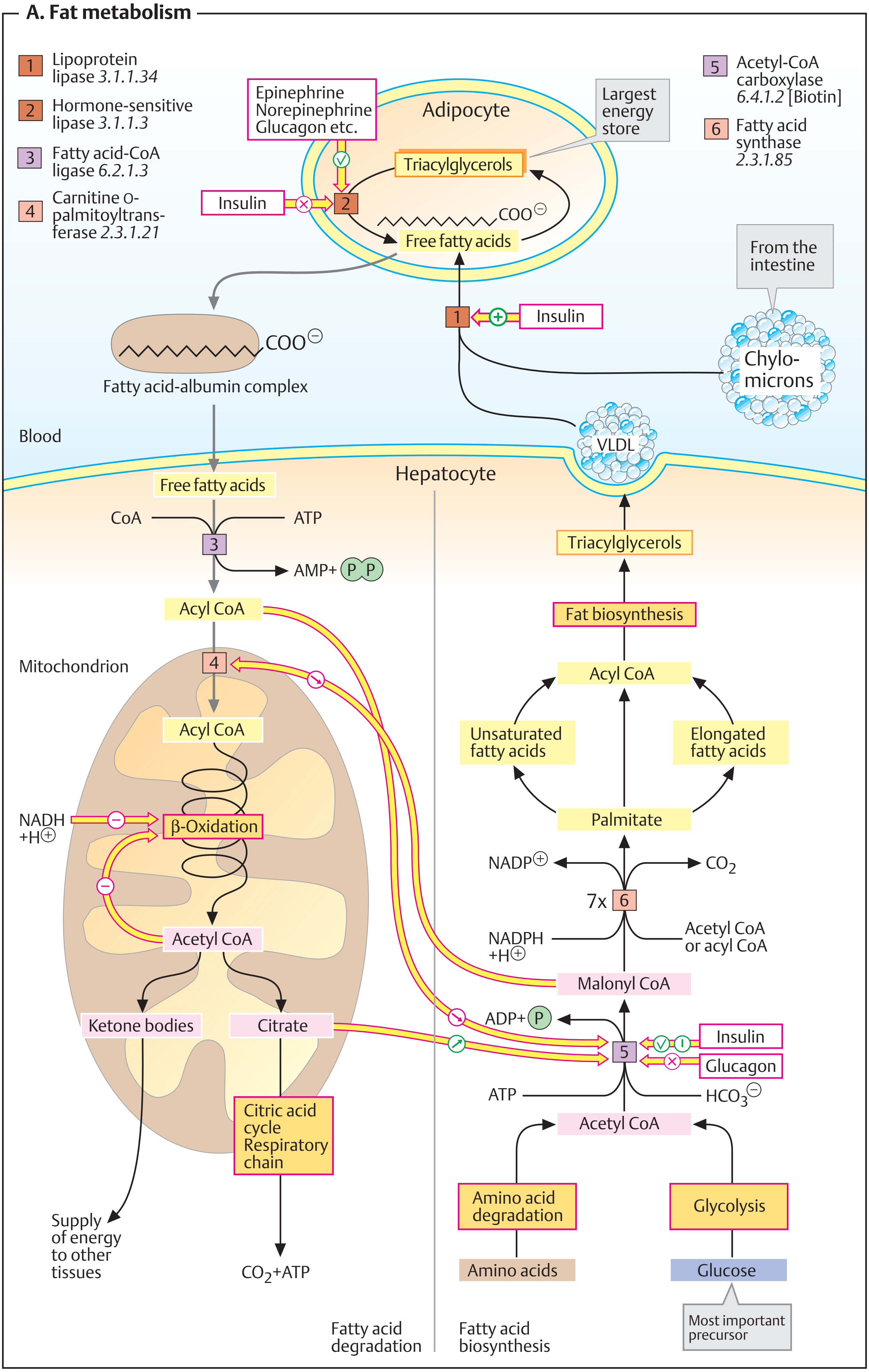
1 Lipoprotein lipase 3.1.1.34
2 Hormone-sensitive lipase 3.1.1.3
3 Fatty acid-CoA ligase 6.2.1.3
4 Carnitine O-palmitoyltransferase 2.3.1.21
5 Acetyl-CoA carboxylase 6.4.1.2 [Biotin]
6 Fatty acid synthase 2.3.1.85
Epinephrine
Norepinephrine
Glucagon etc.
Adipocyte
Largest energy store
Triacylglycerols
Insulin
2
COO⊖
Free fatty acids
From the intestine
1
Insulin
Chylomicrons
COO⊖
Fatty acid-albumin complex
Blood
VLDL
Hepatocyte
Free fatty acids
CoA
ATP
3
AMP+ P P
Triacylglycerols
Acyl CoA
Fat biosynthesis
Mitochondrion
4
Acyl CoA
Acyl CoA
Unsaturated fatty acids
Elongated fatty acids
NADH +H⊕
β-Oxidation
Palmitate
NADP⊕
CO₂
7x 6
Acetyl CoA
NADPH +H⊕
Acetyl CoA or acyl CoA
Malonyl CoA
ADP+ P
Ketone bodies
Citrate
5
Insulin
Glucagon
ATP
HCO₃⊖
Acetyl CoA
Citric acid cycle
Respiratory chain
Amino acid degradation
Glycolysis
Supply of energy to other tissues
CO₂+ATP
Amino acids
Glucose
Most important precursor
Fatty acid degradation
Fatty acid biosynthesis

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Fatty acid degradation

A. Fatty acid degradation: β-oxidation ◑

After uptake by the cell, fatty acids are activated by conversion into their CoA derivatives—**acyl CoA** is formed. This uses up two energy-rich anhydride bonds of ATP per fatty acid (see p. 162). For channeling into the mitochondria, the acyl residues are first transferred to *carnitine* and then transported across the inner membrane as **acyl carnitine** (see **B**).

The degradation of the fatty acids occurs in the mitochondrial matrix through an oxidative cycle in which C_2 units are successively cleaved off as **acetyl CoA** (*activated acetic acid*). Before the release of the acetyl groups, each CH_2 group at C-3 of the acyl residue (the β-C atom) is oxidized to the keto group—hence the term **β-oxidation** for this metabolic pathway. Both spatially and functionally, it is closely linked to the tricarboxylic acid cycle (see p. 136) and to the respiratory chain (see p. 140).

[1] The first step is dehydrogenation of **acyl CoA** at C-2 and C-3. This yields an unsaturated Δ^2-enoyl-CoA derivative with a *trans*-configured double bond. The two hydrogen atoms are initially transferred from FAD-containing *acyl CoA dehydrogenase* to the **electron-transferring flavoprotein (ETF)**. *ETF dehydrogenase* [5] passes them on from ETF to ubiquinone (coenzyme Q), a component of the *respiratory chain* (see p. 140). Other FAD-containing mitochondrial dehydrogenases are also able to supply the respiratory chain with electrons in this fashion.

There are three *isoenzymes* (see p. 98) of *acyl CoA dehydrogenase* that are specialized for long-chain fatty acids (12–18 C atoms), medium-chain fatty acids (4–14), and short-chain fatty acids (4–8).

[2] The next step in fatty acid degradation is the addition of a water molecule to the double bond of the enoyl CoA (*hydration*), with formation of **β-hydroxyacyl CoA**.

[3] In the next reaction, the OH group at C-3 is oxidized to a carbonyl group (*dehydrogenation*). This gives rise to **β-ketoacyl CoA**, and the reduction equivalents are transferred to NAD^+, which also passes them on to the *respiratory chain.*

[4] β-Ketoacyl-CoA is now broken down by an *acyl transferase* into **acetyl CoA** and an **acyl CoA shortened by 2 C atoms** (*"thioclastic cleavage"*).

Several cycles are required for complete degradation of long-chain fatty acids—eight cycles in the case of stearyl-CoA (C18:0), for example. The acetyl CoA formed can then undergo further metabolism in the *tricarboxylic acid cycle* (see p. 136), or can be used for biosynthesis. When there is an excess of acetyl CoA, the liver can also form ketone bodies (see p. 312).

When oxidative degradation is complete, one molecule of palmitic acid supplies around 106 molecules of ATP, corresponding to an energy of 3300 $kJ\ mol^{-1}$. This high energy yield makes fats an ideal form of storage for metabolic energy. Hibernating animals such as polar bears can meet their own energy requirements for up to 6 months solely by fat degradation, while at the same time producing the vital water they need via the respiratory chain ("respiratory water").

B. Fatty acid transport ◑

The inner mitochondrial membrane has a group-specific transport system for fatty acids. In the cytoplasm, the acyl groups of activated fatty acids are transferred to **carnitine** by *carnitine acyltransferase* [1]. They are then channeled into the matrix by an acylcarnitine/carnitine antiport as **acyl carnitine**, in exchange for free carnitine. In the matrix, the mitochondrial enzyme *carnitine acyltransferase* catalyzes the return transfer of the acyl residue to CoA.

The carnitine shuttle is the rate-determining step in mitochondrial fatty acid degradation. Malonyl CoA, a precursor of fatty acid biosynthesis, inhibits *carnitine acyltransferase* (see p. 162), and therefore also inhibits uptake of fatty acids into the mitochondrial matrix.

The most important regulator of β-oxidation is the $NAD^+/NADH+H^+$ ratio. If the respiratory chain is not using any $NADH+H^+$, then not only the tricarboxylic acid cycle (see p. 136) but also β-oxidation come to a standstill due to the lack of NAD^+.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Fatty acid degradation: β-oxidation

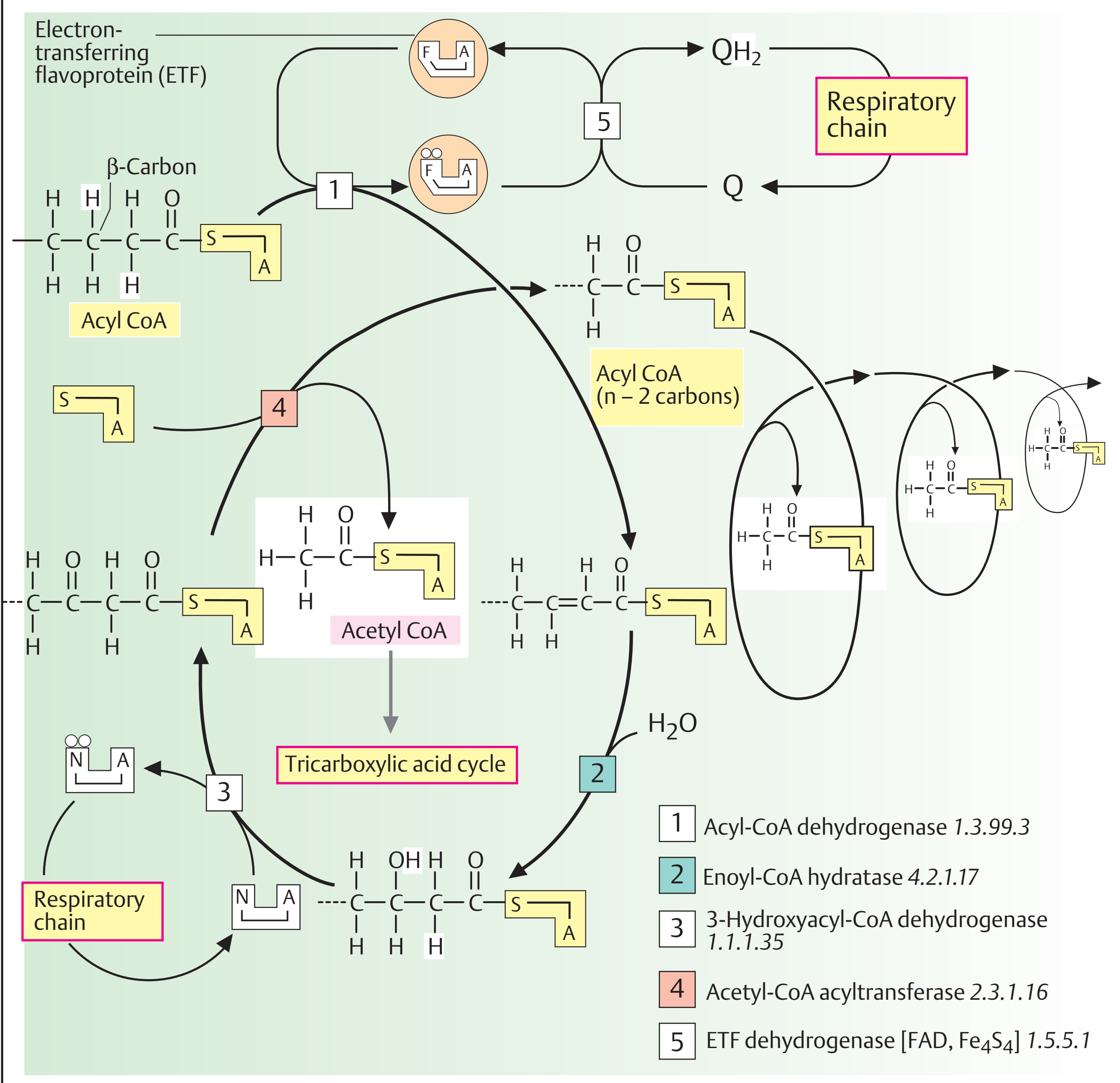
Electron-transferring flavoprotein (ETF)
FAD
QH2
Respiratory chain
Q
β-Carbon
Acyl CoA
Acyl CoA (n – 2 carbons)
Acetyl CoA
Tricarboxylic acid cycle
H2O
NAD
Respiratory chain
1
2
3
4
5

1 Acyl-CoA dehydrogenase *1.3.99.3*

2 Enoyl-CoA hydratase *4.2.1.17*

3 3-Hydroxyacyl-CoA dehydrogenase *1.1.1.35*

4 Acetyl-CoA acyltransferase *2.3.1.16*

5 ETF dehydrogenase [FAD, Fe_4S_4] *1.5.5.1*

B. Fatty acid transport

1 Carnitine *O*-palmitoyltransferase *2.3.1.21*

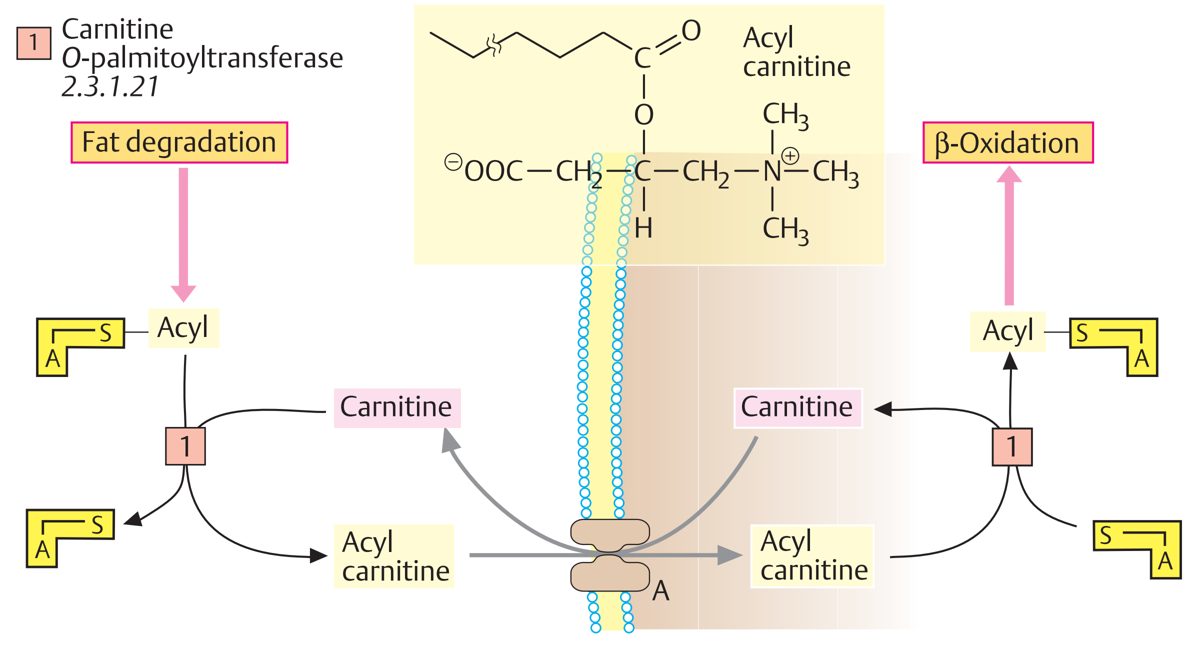
Fat degradation
β-Oxidation
Acyl carnitine
$^{\ominus}OOC-CH_2-C-CH_2-N^{\oplus}-CH_3$
CH_3
CH_3
O
H
Acyl
Carnitine
Acyl carnitine
1
A
S

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Minor pathways of fatty acid degradation

Most fatty acids are saturated and even-numbered. They are broken down via **β-oxidation** (see p.164). In addition, there are special pathways involving degradation of unsaturated fatty acids (**A**), degradation of fatty acids with an odd number of C atoms (**B**), α and ω oxidation of fatty acids, and degradation in peroxisomes.

A. Degradation of unsaturated fatty acids ○

Unsaturated fatty acids usually contain a *cis* double bond at position 9 or 12—e. g., linoleic acid (18:2; 9,12). As with saturated fatty acids, degradation in this case occurs via β-oxidation until the C-9-*cis* double bond is reached. Since *enoyl-CoA hydratase* only accepts substrates with *trans* double bonds, the corresponding enoyl-CoA is converted by an isomerase from the *cis*-Δ^3, *cis*- Δ^6 isomer into the *trans*-Δ^3,*cis*-Δ^6 isomer [1]. Degradation by β-oxidation can now continue until a shortened *trans*-Δ^2, *cis*-Δ^4 derivative occurs in the next cycle. This cannot be isomerized in the same way as before, and instead is reduced in an NADPH-dependent way to the *trans*-Δ^3 compound [2]. After rearrangement by *enoyl-CoA isomerase* [1], degradation can finally be completed via normal β-oxidation.

B. Degradation of oddnumbered fatty acids ○

Fatty acids with an odd number of C atoms are treated in the same way as "normal" fatty acids—i. e., they are taken up by the cell with ATP-dependent activation to acyl CoA and are transported into the mitochondria with the help of the carnitine shuttle and broken down there by β–oxidation (see p. 164). In the last step, **propionyl CoA** arises instead of acetyl CoA. This is first carboxylated by *propionyl CoA carboxylase* into ***(S)*-methylmalonyl CoA** [3], which—after isomerization into the (*R*) enantiomer (not shown; see p. 411)—is isomerized into **succinyl CoA** [4].

Various coenzymes are involved in these reactions. The carboxylase [3] requires *biotin*, and the mutase [4] is dependent on *coenzyme* B_{12} (5′-deoxyadenosyl cobalamin; see p. 108). Succinyl-CoA is an intermediate in the tricarboxylic acid cycle and is available for *gluconeogenesis* through conversion into oxaloacetate. Odd-numbered fatty acids from propionyl-CoA can therefore be used to synthesize glucose.

This pathway is also important for ruminant animals, which are dependent on symbiotic microorganisms to break down their food. The microorganisms produce large amounts of propionic acid as a degradation product, which the host can channel into the metabolism in the way described.

Further information ○

In addition to the degradation pathways described above, there are also additional special pathways for particular fatty acids found in food.

~ **Oxidation** is used to break down methyl-branched fatty acids. It takes place through step-by-step removal of C_1 residues, begins with a hydroxylation, does not require coenzyme A, and does not produce any ATP.

∂ **Oxidation**—i. e., oxidation starting at the end of the fatty acid—also starts with a hydroxylation catalyzed by a *monooxygenase* (see p. 316), and leads via subsequent oxidation to fatty acids with two carboxyl groups, which can undergo β-oxidation from both ends until C_8 or C_6 dicarboxylic acids are reached, which can be excreted in the urine in this form.

Degradation of unusually long fatty acids. An alternative form of β-oxidation takes place in *hepatic peroxisomes*, which are specialized for the degradation of particularly long fatty acids (n > 20). The degradation products are acetyl-CoA and hydrogen peroxide (H_2O_2), which is detoxified by the *catalase* (see p. 32) common in peroxisomes.

Enzyme defects are also known to exist in the minor pathways of fatty acid degradation. In **Refsum disease**, the methyl-branched phytanic acid (obtained from vegetable foods) cannot be degraded by α-oxidation. In **Zellweger syndrome**, a peroxisomal defect means that long-chain fatty acids cannot be degraded.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Degradation of unsaturated fatty acids

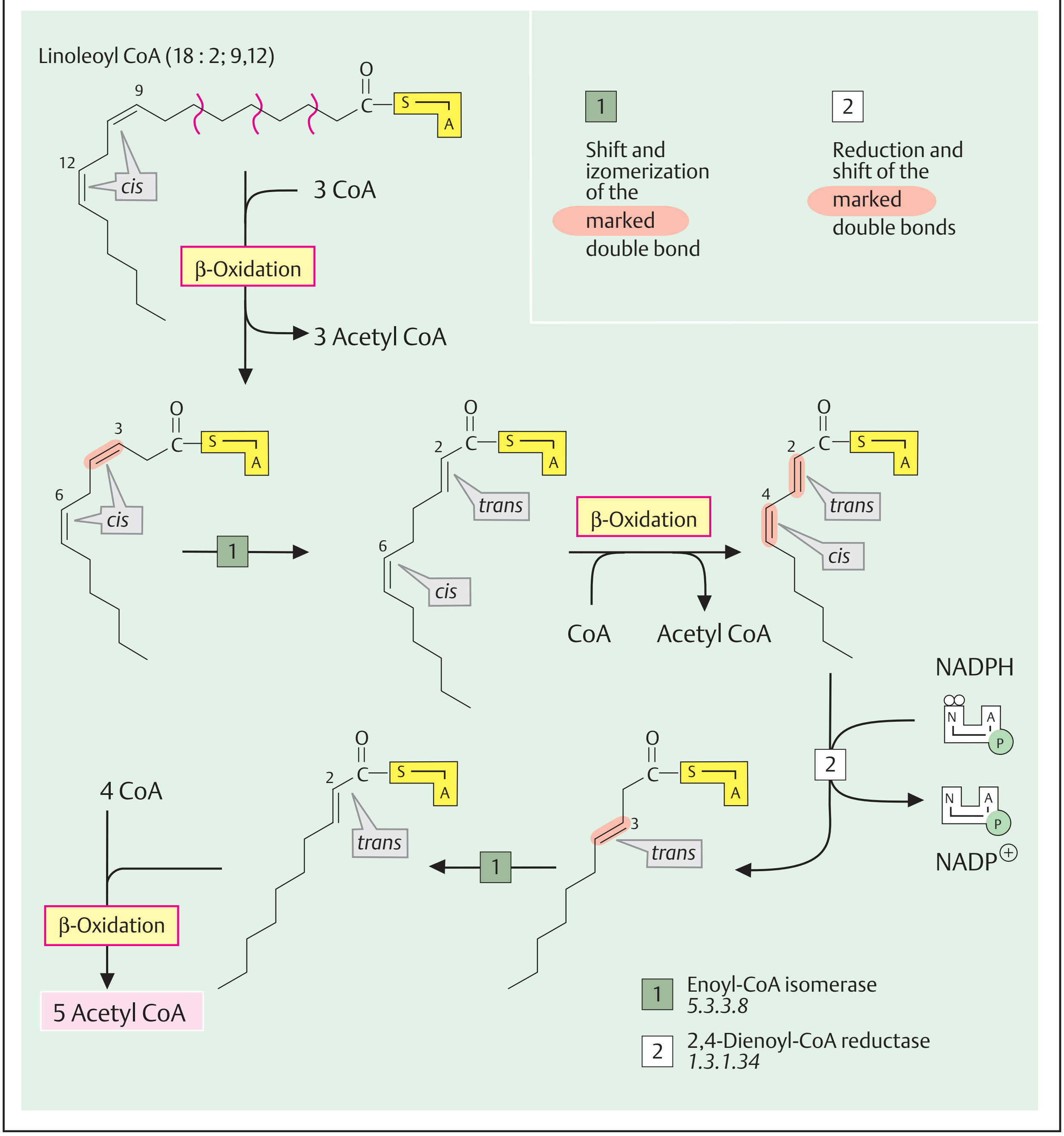
Linoleoyl CoA (18 : 2; 9,12)
9
12
cis
O
C
S
A
3 CoA
β-Oxidation
3 Acetyl CoA
1
Shift and izomerization of the marked double bond
2
Reduction and shift of the marked double bonds
3
6
cis
1
2
trans
6
cis
β-Oxidation
CoA
Acetyl CoA
2
4
trans
cis
NADPH
N
A
P
2
NADP⊕
3
trans
2
trans
1
4 CoA
β-Oxidation
5 Acetyl CoA
1 Enoyl-CoA isomerase
5.3.3.8
2 2,4-Dienoyl-CoA reductase
1.3.1.34

B. Degradation of odd-numbered fatty acids

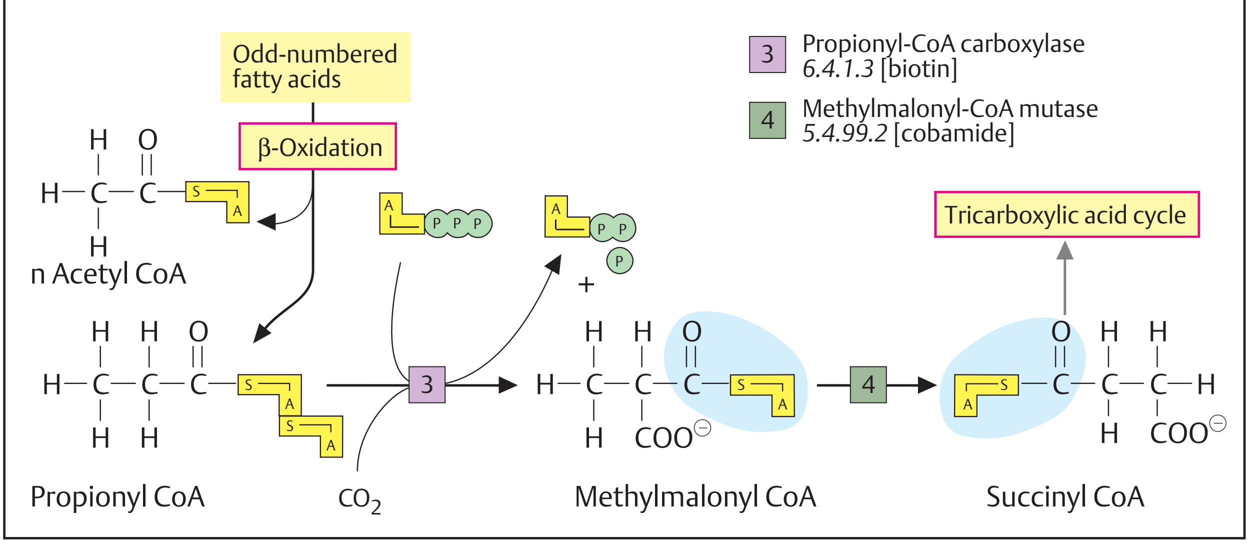
Odd-numbered fatty acids
β-Oxidation
H
O
H—C—C—
H
S
A
n Acetyl CoA
3 Propionyl-CoA carboxylase
6.4.1.3 [biotin]
4 Methylmalonyl-CoA mutase
5.4.99.2 [cobamide]
A
P P P
A
P P
P
+
Tricarboxylic acid cycle
H H O
H—C—C—C—
H H
Propionyl CoA
3
CO2
H H O
H—C—C—C—
H COO⊖
Methylmalonyl CoA
4
O H H
—C—C—C—H
H COO⊖
Succinyl CoA

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Fatty acid synthesis

In the vertebrates, biosynthesis of fatty acids is catalyzed by *fatty acid synthase,* a multifunctional enzyme. Located in the cytoplasm, the enzyme requires acetyl CoA as a starter molecule. In a cyclic reaction, the acetyl residue is elongated by one C_2 unit at a time for seven cycles. NADPH+H^+ is used as a reducing agent in the process. The end product of the reaction is the saturated C_{16} acid, *palmitic acid.*

A. Fatty acid synthase ◑

Fatty acid synthase in vertebrates consists of two identical peptide chains—i. e., it is a homodimer. Each of the two peptide chains, which are shown here as hemispheres, catalyzes all seven of the partial reactions required to synthesize palmitate. The spatial compression of several successive reactions into a single multifunctional enzyme has advantages in comparison with separate enzymes. Competing reactions are prevented, the individual reactions proceed in a coordinated way as if on a production line, and due to low diffusion losses they are particularly efficient.

Each subunit of the enzyme binds acetyl residues as thioesters at two different SH groups: at one peripheral *cysteine residue* (CysSH) and one central *4′-phosphopantetheine group* (Pan-SH). Pan-SH, which is very similar to coenzyme A (see p. 12), is covalently bound to a protein segment of the synthase known as the *acyl-carrier protein* (ACP). This part functions like a long arm that passes the substrate from one reaction center to the next. The two subunits of fatty acid synthase cooperate in this process; the enzyme is therefore only capable of functioning as a dimer.

Spatially, the enzyme activities are arranged into three different domains. **Domain 1** catalyzes the entry of the substrates acetyl CoA and malonyl CoA by *[ACP]-S-acetyltransferase* [1] and *[ACP]-Smalonyl transferase* [2] and subsequent condensation of the two partners by *3-oxoacyl-[ACP]-synthase* [3]. **Domain 2** catalyzes the conversion of the 3-oxo group to a CH_2 group by *3-oxoacyl-[ACP]-reductase* [4], *3-hydroxyacyl-[ACP]-dehydratase* [5], and *enoyl-[ACP]-reductase* [6]. Finally, **domain 3** serves to release the finished product by *acyl-[ACP]-hydrolase* [7] after seven steps of chain elongation.

B. Reactions of fatty acid synthase ◑

The key enzyme in fatty acid synthesis is **acetyl CoA carboxylase** (see p. 162), which precedes the synthase and supplies the malonyl-CoA required for elongation. Like all carboxylases, the enzyme contains covalently bound *biotin* as a prosthetic group and is hormone-dependently *inactivated* by phosphorylation or *activated* by dephosphorylation (see p. 120). The precursor *citrate* (see p. 138) is an allosteric activator, while *palmitoyl-CoA* inhibits the end product of the synthesis pathway.

[1] The first cycle (n = 1) starts with the transfer of an acetyl residue from acetyl CoA to the peripheral cysteine residue (Cys-SH). At the same time,

[2] a malonyl residue is transferred from malonyl CoA to 4-phosphopantetheine (Pan-SH).

[3] By condensation of the acetyl residue—or (in later cycles) the acyl residue—with the malonyl group, with simultaneous decarboxylation, the chain is elongated.

[4]–[6] The following three reactions (reduction of the 3-oxo group, dehydrogenation of the 3-hydroxyl derivative, and renewed reduction of it) correspond in principle to a reversal of β-oxidation, but they are catalyzed by other enzymes and use NADPH+H^+ instead of NADH+H^+ for reduction. They lead to an acyl residue bound at Pan-SH with 2n + 2 C atoms (n = the number of the cycle). Finally, depending on the length of the product,

[1′] The acyl residue is transferred back to the peripheral cysteine, so that the next cycle can begin again with renewed loading of the ACP with a malonyl residue, or:

[7] After seven cycles, the completed **palmitic acid** is hydrolytically released.

In all, one acetyl-CoA and seven malonyl-CoA are converted with the help of 14 NADPH+H^+ into one palmitic acid, 7 CO_2, 6 H_2O, 8 CoA and 14 NADP$^+$. Acetyl CoA carboxylase also uses up seven ATP.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

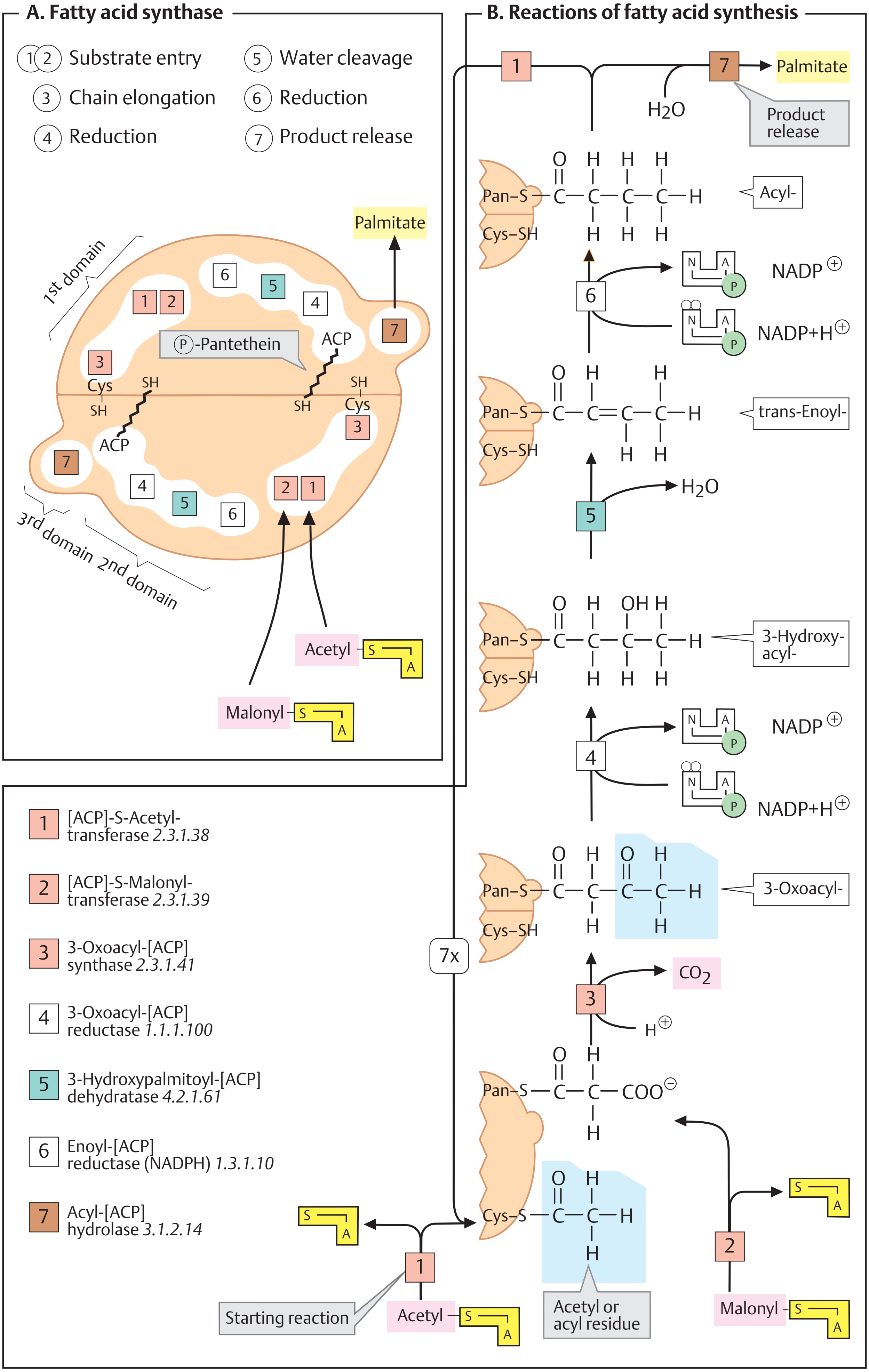

A. Fatty acid synthase

①② Substrate entry
③ Chain elongation
④ Reduction
⑤ Water cleavage
⑥ Reduction
⑦ Product release

Palmitate
1st domain
2nd domain
3rd domain
1
2
3
4
5
6
7
Ⓟ-Pantethein
ACP
Cys
SH
Acetyl
Malonyl
S
A

B. Reactions of fatty acid synthesis

1
7
Palmitate
H2O
Product release
Pan–S
Cys–SH
Acyl-
6
NADP⊕
NADP+H⊕
N
A
P
trans-Enoyl-
5
H2O
3-Hydroxy-acyl-
4
3-Oxoacyl-
7x
3
CO2
H⊕
COO⊖
Cys–S
2
Acetyl or acyl residue
Starting reaction
Acetyl
Malonyl
S
A

1	[ACP]-S-Acetyltransferase *2.3.1.38*
2	[ACP]-S-Malonyltransferase *2.3.1.39*
3	3-Oxoacyl-[ACP] synthase *2.3.1.41*
4	3-Oxoacyl-[ACP] reductase *1.1.1.100*
5	3-Hydroxypalmitoyl-[ACP] dehydratase *4.2.1.61*
6	Enoyl-[ACP] reductase (NADPH) *1.3.1.10*
7	Acyl-[ACP] hydrolase *3.1.2.14*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Biosynthesis of complex lipids

A. Biosynthesis of fats and phospholipids ◑

Complex lipids, such as neutral fats (triacylglycerols), phospholipids, and glycolipids, are synthesized via common reaction pathways. Most of the enzymes involved are associated with the membranes of the smooth endoplasmic reticulum.

The synthesis of fats and phospholipids starts with **glycerol 3-phosphate**. This compound can arise via two pathways:

[1] By reduction from the glycolytic intermediate **glycerone 3-phosphate** (dihydroxyacetone 3-phosphate; enzyme: *glycerol-3-phosphate dehydrogenase (NAD+) 1.1.1.8*), or:

[2] By phosphorylation of **glycerol** deriving from fat degradation (enzyme: *glycerol kinase 2.7.1.30*).

[3] Esterification of glycerol 3-phosphate with a long-chain fatty acid produces a strongly amphipathic **lysophosphatidate** (enzyme: *glycerol-3-phosphate acyltransferase 2.3.1.15*). In this reaction, an acyl residue is transferred from the activated precursor **acyl-CoA** to the hydroxy group at C-1.

[4] A second esterification of this type leads to a **phosphatidate** (enzyme: *1-acylglycerol-3-phosphate acyltransferase 2.3.1.51*). Unsaturated acyl residues, particularly oleic acid, are usually incorporated at C-2 of the glycerol. Phosphatidates (anions of phosphatidic acids) are the key molecules in the biosynthesis of fats, phospholipids, and glycolipids.

[5] To biosynthesize fats (triacylglycerols), the phosphate residue is again removed by hydrolysis (enzyme: *phosphatidate phosphatase 3.1.3.4*). This produces **diacylglycerols (DAG)**.

[6] Transfer of an additional acyl residue to DAG forms **triacylglycerols** (enzyme: *diacylglycerol acyltransferase 2.3.1.20*). This completes the biosynthesis of neutral fats. They are packaged into VLDLs by the liver and released into the blood. Finally, they are stored by adipocytes in the form of insoluble fat droplets.

The biosynthesis of most phospholipids also starts from DAG.

[7] Transfer of a phosphocholine residue to the free OH group gives rise to **phosphatidylcholine** (lecithin; enzyme: *1-alkyl-2-acetylglycerolcholine phosphotransferase 2.7.8.16*). The phosphocholine residue is derived from the precursor CDP-choline (see p. 110). **Phosphatidylethanolamine** is similarly formed from CDP-ethanolamine and DAG. By contrast, **phosphatidylserine** is derived from phosphatidylethanolamine by an exchange of the amino alcohol. Further reactions serve to interconvert the phospholipids—e. g., phosphatidylserine can be converted into phosphatidylethanolamine by decarboxylation, and the latter can then be converted into phosphatidylcholine by methylation with S-adenosyl methionine (not shown; see also p. 409). The biosynthesis of **phosphatidylinositol** starts from phosphatidate rather than DAG.

[8] In the lumen of the intestine, fats from food are mainly broken down into **monoacylglycerols** (see p. 270). The cells of the intestinal mucosa re-synthesize these into neutral fats. This pathway also passes via **DAG** (enzyme: *acylglycerolpalmitoyl transferase 2.3.1.22*).

[9] Transfer of a CMP residue gives rise first to **CDP-diacylglycerol** (enzyme: *phosphatidatecytidyl transferase 2.3.1.22*).

[10] Substitution of the CMP residue by inositol then provides **phosphatidylinositol** (**PtdIns**; enzyme: *CDPdiacylglycerolinositol-3-phosphatidyl transferase 2.7.8.11*).

[12] An additional phosphorylation (enzyme: *phosphatidylinositol-4-phosphate kinase 2.7.1.68*) finally provides **phosphaditylinositol-4,5-bisphosphate** (PIP_2, $PtdIns(4,5)P_2$). PIP_2 is the precursor for the second messengers *2,3-diacylglycerol* (DAG) and *inositol-1,4,5-trisphosphate* ($InsP_3$, IP_3; see p. 367).

The biosynthesis of the **sphingolipids** is shown in schematic form on p. 409.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biosynthesis of fats and phospholipids

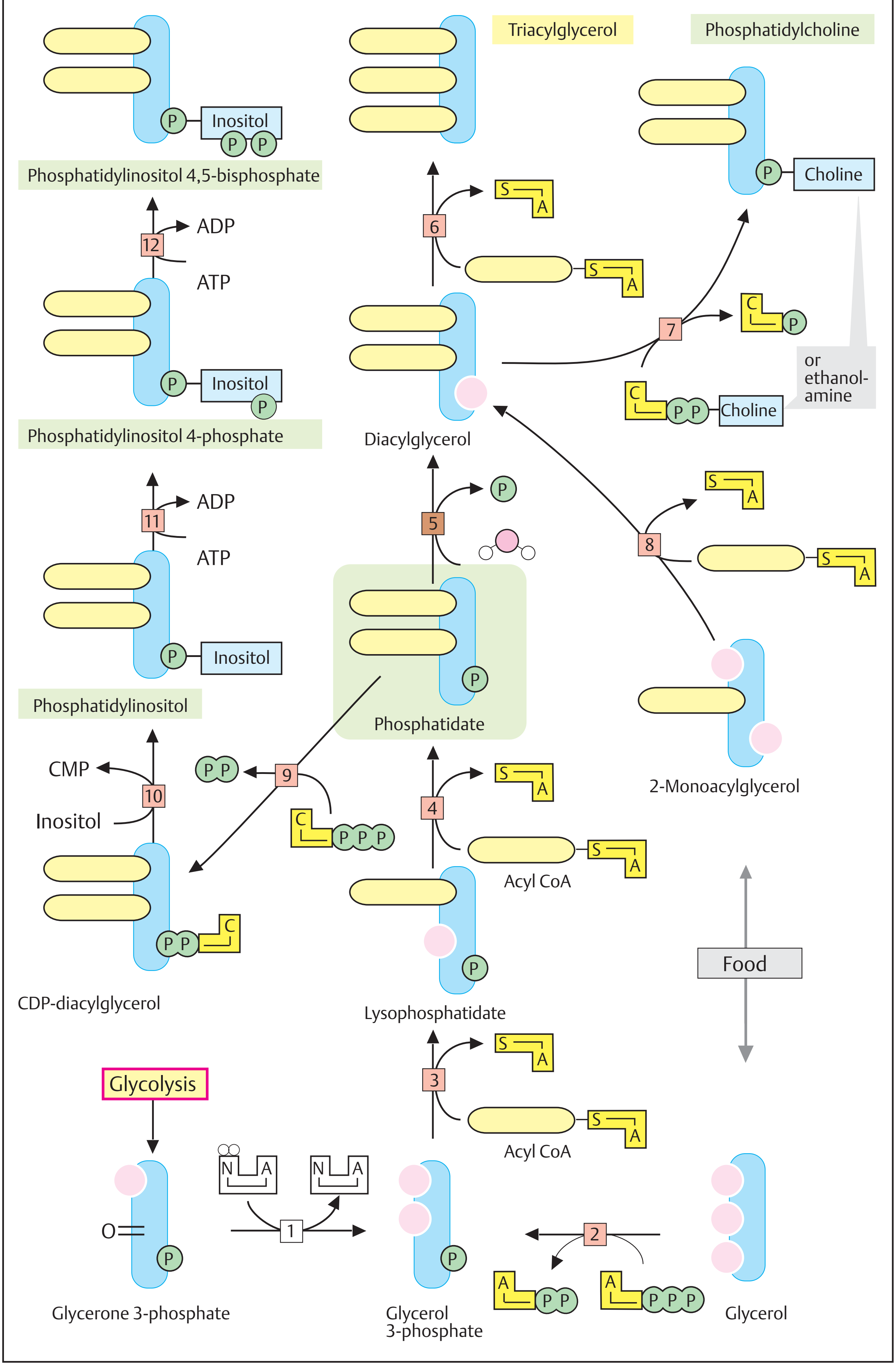
Triacylglycerol
Phosphatidylcholine
Inositol
P
P P
Phosphatidylinositol 4,5-bisphosphate
12
ADP
ATP
6
S A
Choline
Inositol
Phosphatidylinositol 4-phosphate
7
C P
or ethanol-amine
C P P Choline
Diacylglycerol
11
ADP
ATP
5
P
8
Inositol
Phosphatidylinositol
Phosphatidate
2-Monoacylglycerol
CMP
10
Inositol
P P
9
C P P P
4
Acyl CoA
CDP-diacylglycerol
Lysophosphatidate
Food
3
Glycolysis
Acyl CoA
N A
1
O
Glycerone 3-phosphate
Glycerol 3-phosphate
2
A P P
A P P P
Glycerol

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Biosynthesis of cholesterol

Cholesterol is a major constituent of the *cell membranes* of animal cells (see p. 216). It would be possible for the body to provide its full daily cholesterol requirement (ca. 1 g) by synthesizing it itself. However, with a mixed diet, only about half of the cholesterol is derived from *endogenous biosynthesis*, which takes place in the intestine and skin, and mainly in the liver (about 50%). The rest is taken up from *food*. Most of the cholesterol is incorporated into the lipid layer of plasma membranes, or converted into **bile acids** (see p. 314). A very small amount of cholesterol is used for biosynthesis of the **steroid hormones** (see p. 376). In addition, up to 1 g cholesterol per day is released into the *bile* and thus excreted.

A. Cholesterol biosynthesis ○

Cholesterol is one of the isoprenoids, synthesis of which starts from **acetyl CoA** (see p. 52). In a long and complex reaction chain, the C_{27} sterol is built up from C_2 components. The biosynthesis of cholesterol can be divided into four sections. In the first (**1**), **mevalonate,** a C_6 compound, arises from three molecules of **acetyl CoA**. In the second part (**2**), mevalonate is converted into **isopentenyl diphosphate,** the “active isoprene.” In the third part (**3**), six of these C_5 molecules are linked to produce **squalene,** a C_{30} compound. Finally, squalene undergoes cyclization, with three C atoms being removed, to yield cholesterol (**4**). The illustration only shows the most important intermediates in biosynthesis.

(1) Formation of mevalonate. The conversion of acetyl CoA to acetoacetyl CoA and then to *3-hydroxy-3-methylglutaryl CoA* (3-HMG CoA) corresponds to the biosynthetic pathway for *ketone bodies* (details on p. 312). In this case, however, the synthesis occurs not in the mitochondria as in ketone body synthesis, but in the smooth endoplasmic reticulum. In the next step, the 3-HMG group is cleaved from the CoA and at the same time reduced to mevalonate with the help of $NADPH+H^+$. *3-HMG CoA reductase* is the *key enzyme* in cholesterol biosynthesis. It is regulated by *repression* of transcription (effectors: oxysterols such as cholesterol) and by *interconversion* (effectors: hormones). Insulin and thyroxine stimulate the enzyme and glucagon inhibits it by cAMP-dependent phosphorylation. A large supply of cholesterol from food also inhibits 3-HMG-CoA reductase.

(2) Formation of isopentenyl diphosphate. After phosphorylation, mevalonate is decarboxylated to *isopentenyl diphosphate,* with consumption of ATP. This is the component from which all of the isoprenoids are built (see p. 53).

(3) Formation of squalene. Isopentenyl diphosphate undergoes isomerization to form dimethylallyl diphosphate. The two C_5 molecules condense to yield geranyl diphosphate, and the addition of another isopentenyl diphosphate produces farnesyl diphosphate. This can then undergo dimerization, in a *head-to-head reaction,* to yield squalene. Farnesyl diphosphate is also the starting-point for other polyisoprenoids, such as dolichol (see p. 230) and ubiquinone (see p. 52).

(4) Formation of cholesterol. Squalene, a linear isoprenoid, is cyclized, with O_2 being consumed, to form lanosterol, a C_{30} sterol. Three methyl groups are cleaved from this in the subsequent reaction steps, to yield the end product cholesterol. Some of these reactions are catalyzed by *cytochrome P450 systems* (see p. 318).

The endergonic biosynthetic pathway described above is located entirely in the *smooth endoplasmic reticulum.* The energy needed comes from the CoA derivatives used and from ATP. The reducing agent in the formation of mevalonate and squalene, as well as in the final steps of cholesterol biosynthesis, is $NADPH+H^+$.

The division of the intermediates of the reaction pathway into three groups is characteristic: CoA compounds, diphosphates, and highly lipophilic, poorly soluble compounds (squalene to cholesterol), which are bound to *sterol carriers* in the cell.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Cholesterol biosynthesis

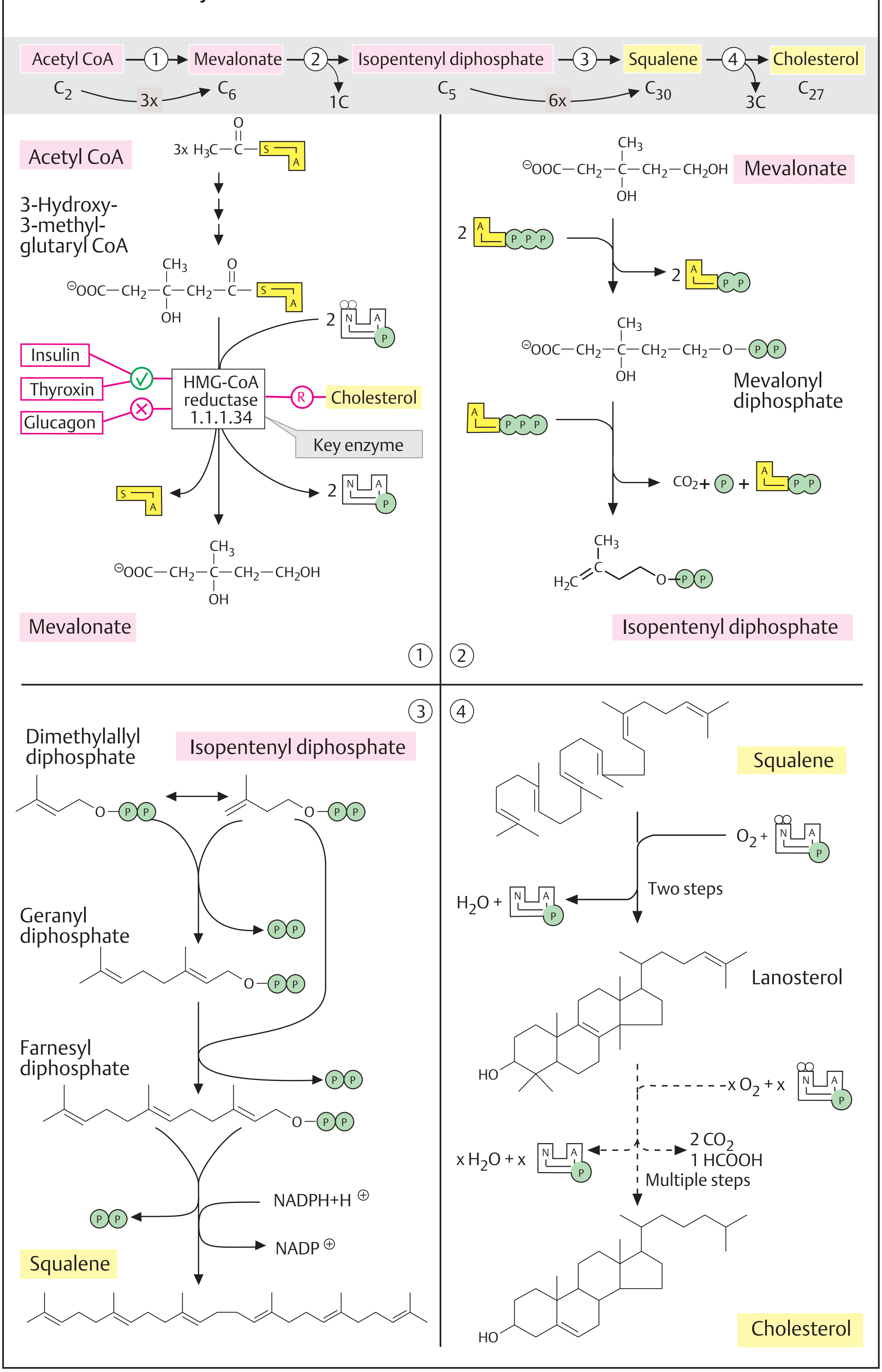
Acetyl CoA
1
Mevalonate
2
Isopentenyl diphosphate
3
Squalene
4
Cholesterol
C2
3x
C6
1C
C5
6x
C30
3C
C27
Acetyl CoA
3x H3C–C(=O)–S-A
3-Hydroxy-3-methyl-glutaryl CoA
⊖OOC–CH2–C(CH3)(OH)–CH2–C(=O)–S-A
2 NADPH
Insulin
Thyroxin
Glucagon
HMG-CoA reductase 1.1.1.34
R
Cholesterol
Key enzyme
S-A
2 NADP
⊖OOC–CH2–C(CH3)(OH)–CH2–CH2OH
Mevalonate
1
2
⊖OOC–CH2–C(CH3)(OH)–CH2–CH2OH
Mevalonate
2 ATP
2 ADP
⊖OOC–CH2–C(CH3)(OH)–CH2–CH2–O–P–P
Mevalonyl diphosphate
ATP
CO2 + P + ADP
H2C=C(CH3)–CH2–CH2–O–P–P
Isopentenyl diphosphate
3
4
Dimethylallyl diphosphate
Isopentenyl diphosphate
O–P–P
O–P–P
P–P
Geranyl diphosphate
O–P–P
P–P
Farnesyl diphosphate
O–P–P
P–P
NADPH+H⊕
NADP⊕
Squalene
Squalene
O2 + NADPH
H2O + NADP
Two steps
Lanosterol
HO
- - - - x O2 + x NADPH
x H2O + x NADP
2 CO2
1 HCOOH
Multiple steps
HO
Cholesterol

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Protein metabolism: overview

Quantitatively, proteins are the most important group of endogenous macromolecules. A person weighing 70 kg contains about 10 kg protein, with most of it located in muscle. By comparison, the proportion made up by other nitrogencontaining compounds is minor. The organism's nitrogen balance is therefore primarily determined by protein metabolism. Several hormones—mainly *testosterone* and *cortisol*—regulate the nitrogen balance (see p. 374).

A. Protein metabolism: overview ●

In adults, the **nitrogen balance** is generally in *equilibrium*—i. e., the quantities of protein nitrogen taken in and excreted per day are approximately equal. If only some of the nitrogen taken in is excreted again, then the balance is *positive.* This is the case during growth, for example. *Negative* balances are rare and usually occur due to disease.

Proteins taken up in food are initially broken down in the gastrointestinal tract into amino acids, which are resorbed and distributed in the organism via the blood (see p. 266). The human body is not capable of synthesizing 8–10 of the 20 proteinogenic amino acids it requires (see p. 60). These amino acids are **essential**, and have to be supplied from food (see p. 184).

Proteins are constantly being lost via the intestine and, to a lesser extent, via the kidneys. To balance these inevitable losses, at least 30 g of protein have to be taken up with food every day. Although this minimum value is barely reached in some countries, in the industrial nations the protein content of food is usually much higher than necessary. As it is not possible to store amino acids, up to 100 g of excess amino acids per day are used for biosynthesis or degraded in the liver in this situation. The nitrogen from this excess is converted into urea (see p. 182) and excreted in the urine in this form. The carbon skeletons are used to synthesize carbohydrates or lipids (see p. 180), or are used to form ATP.

It is thought that adults break down 300–400 g of protein per day into amino acids (**proteolysis**). On the other hand, approximately the same amount of amino acids is reincorporated into proteins (**protein biosynthesis**). The body's high level of protein turnover is due to the fact that many proteins are relatively *short-lived.* On average, their half-lives amount to 2–8 days. The *key enzymes* of the intermediary metabolism have even shorter half-lives. They are sometimes broken down only a few hours after being synthesized, and are replaced by new molecules. This constant process of synthesis and degradation makes it possible for the cells to quickly adjust the quantities, and therefore the activity, of important enzymes in order to meet current requirements. By contrast, structural proteins such as the histones, hemoglobin, and the components of the cytoskeleton are particularly long-lived.

Almost all cells are capable of carrying out **biosynthesis** of proteins (top left). The formation of peptide chains by **translation** at the ribosome is described in greater detail on pp. 250–253. However, the functional forms of most proteins arise only after a series of additional steps. To begin with, supported by auxiliary proteins, the biologically active conformation of the peptide chain has to be formed (**folding**; see pp. 74, 232). During subsequent "post-translational" **maturation**, many proteins remove part of the peptide chain again and attach additional groups—e. g., oligosaccharides or lipids. These processes take place in the endoplasmic reticulum and in the Golgi apparatus (see p. 232). Finally, the proteins have to be transported to their site of action (**sorting**; see p. 228).

Some *intracellular* protein degradation (**proteolysis**) takes place in the lysosomes (see p. 234). In addition, there are protein complexes in the cytoplasm, known as *proteasomes,* in which incorrectly folded or old proteins are degraded. These molecules are recognized by a special **marking** (see p. 176). The proteasome also plays an important part in the presentation of antigens by immune cells (see p. 296).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Protein metabolism: overview

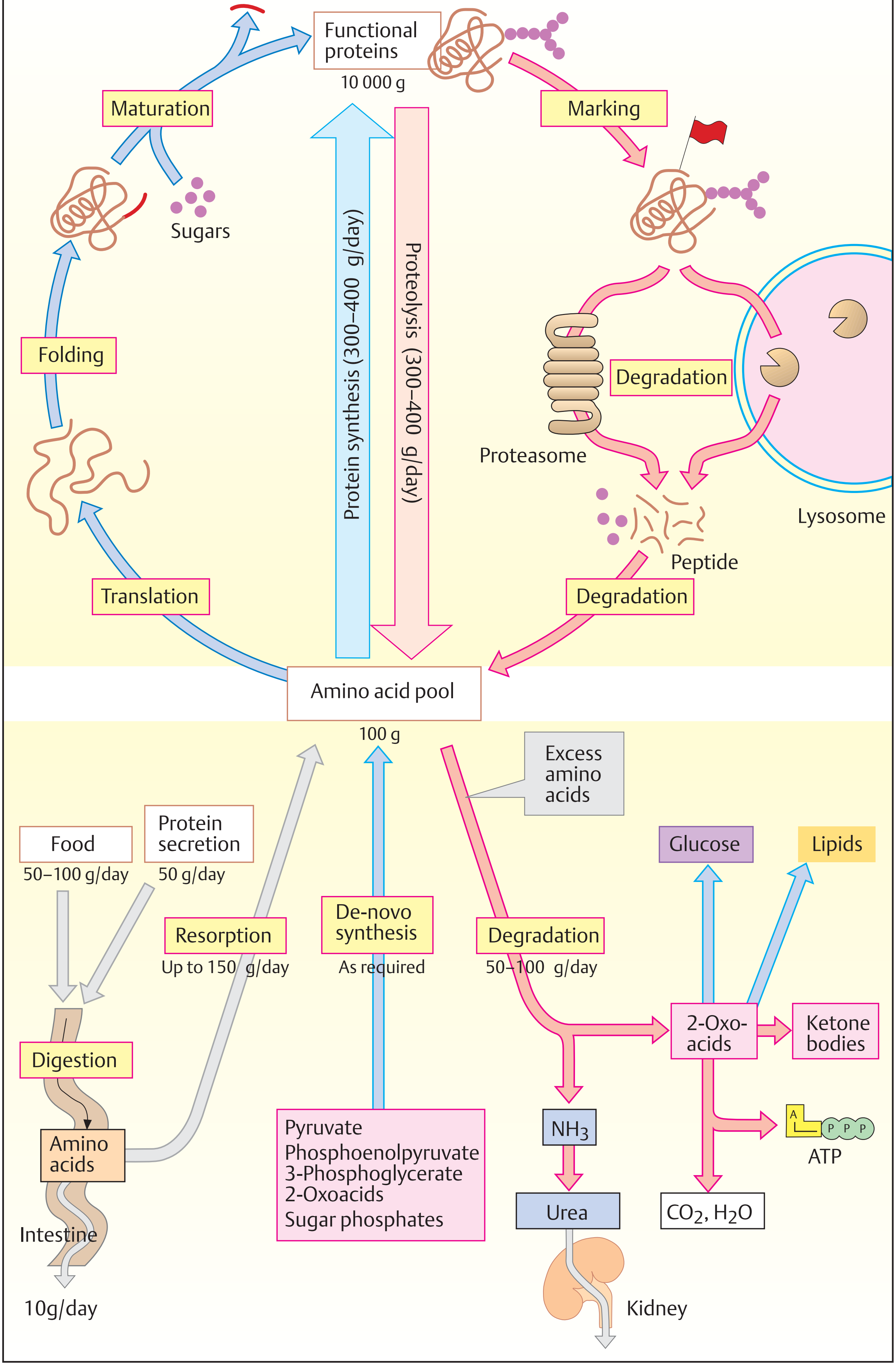
Functional proteins
10 000 g
Maturation
Marking
Sugars
Folding
Protein synthesis (300–400 g/day)
Proteolysis (300–400 g/day)
Degradation
Proteasome
Lysosome
Peptide
Translation
Degradation
Amino acid pool
100 g
Excess amino acids
Food
50–100 g/day
Protein secretion
50 g/day
Resorption
Up to 150 g/day
De-novo synthesis
As required
Degradation
50–100 g/day
Glucose
Lipids
Digestion
2-Oxo-acids
Ketone bodies
Amino acids
NH3
A
P P P
ATP
Pyruvate
Phosphoenolpyruvate
3-Phosphoglycerate
2-Oxoacids
Sugar phosphates
Urea
CO2, H2O
Intestine
10g/day
Kidney

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Proteolysis

A. Proteolytic enzymes ◑

Combinations of several enzymes with different specificities are required for complete degradation of proteins into free amino acids. **Proteinases** and **peptidases** are found not only in the gastrointestinal tract (see p. 268), but also inside the cell (see below).

The proteolytic enzymes are classified into **endopeptidases** and **exopeptidases**, according to their site of attack in the substrate molecule. The *endopeptidases* or *proteinases* cleave peptide bonds *inside* peptide chains. They "recognize" and bind to short sections of the substrate's sequence, and then hydrolyze bonds between particular amino acid residues in a relatively specific way (see p. 94). The **proteinases** are classified according to their reaction mechanism. In *serine proteinases,* for example (see **C**), a serine residue in the enzyme is important for catalysis, while in *cysteine proteinases,* it is a cysteine residue, and so on.

The exopeptidases attack peptides from their termini. Peptidases that act at the N terminus are known as **aminopeptidases**, while those that recognize the C terminus are called **carboxypeptidases**. The **dipeptidases** only hydrolyze dipeptides.

B. Proteasome ○

The functional proteins in the cell have to be protected in order to prevent premature degradation. Some of the intracellularly active proteolytic enzymes are therefore enclosed in lysosomes (see p. 234). The proteinases that act there are also known as **cathepsins**. Another carefully regulated system for protein degradation is located in the cytoplasm. This consists of large protein complexes (mass 2 10^6 Da), the **proteasomes**. Proteasomes contain a barrel-shaped core consisting of 28 subunits that has a sedimentation coefficient (see p. 200) of 20 S. Proteolytic activity (shown here by the scissors) is localized in the interior of the 20-S core and is therefore protected. The openings in the barrel are sealed by 19-S particles with a complex structure that control access to the core.

Proteins destined for degradation in the proteasome (e. g., incorrectly folded or old molecules) are marked by covalent linkage with chains of the small protein **ubiquitin**. The ubiquitin is previously activated by the introduction of reactive thioester groups. Molecules marked with ubiquitin ("ubiquitinated") are recognized by the 19S particle, unfolded using ATP, and then shifted into the interior of the nucleus, where degradation takes place. Ubiquitin is not degraded, but is reused after renewed activation.

C. Serine proteases ○

A large group of proteinases contain serine in their active center. The serine proteases include, for example, the digestive enzymes *trypsin, chymotrypsin,* and *elastase* (see pp. 94 and 268), many *coagulation factors* (see p. 290), and the fibrinolytic enzyme *plasmin* and its *activators* (see p. 292).

As described on p. 270, pancreatic proteinases are secreted as **proenzymes** (zymogens). Activation of these is also based on proteolytic cleavages. This is illustrated here in detail using the example of **trypsinogen**, the precursor of trypsin (**1**). Activation of trypsinogen starts with cleavage of an N-terminal hexapeptide by *enteropeptidase* (enterokinase), a specific serine proteinase that is located in the membrane of the intestinal epithelium. The cleavage product (β-trypsin) is already catalytically active, and it cleaves additional trypsinogen molecules at the sites marked in red in the illustration (autocatalytic cleavage). The precursors of chymotrypsin, elastase, and carboxypeptidase A, among others, are also activated by trypsin.

The active center of trypsin is shown in Fig. **2**. A serine residue in the enzyme (Ser-195), supported by a histidine residue and an aspartate residue (His-57, Asp-102), nucleophilically attacks the bond that is to be cleaved (red arrow). The cleavage site in the substrate peptide is located on the C-terminal side of a lysine residue, the side chain of which is fixed in a special "binding pocket" of the enzyme (left) during catalysis (see p. 94).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Proteolytic enzymes

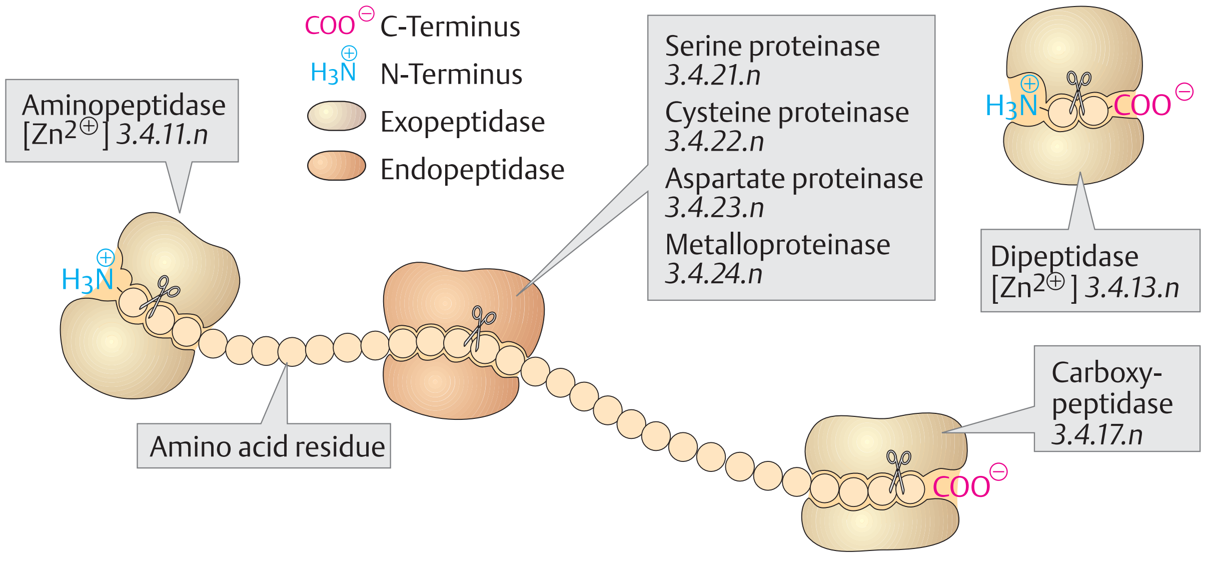
$COO^{\ominus}$ C-Terminus
$H_3\overset{\oplus}{N}$ N-Terminus
Exopeptidase
Endopeptidase
Aminopeptidase [$Zn^{2\oplus}$] 3.4.11.n
Serine proteinase 3.4.21.n
Cysteine proteinase 3.4.22.n
Aspartate proteinase 3.4.23.n
Metalloproteinase 3.4.24.n
Dipeptidase [$Zn^{2\oplus}$] 3.4.13.n
Carboxy-peptidase 3.4.17.n
Amino acid residue
$H_3\overset{\oplus}{N}$
$COO^{\ominus}$

B. Proteasome

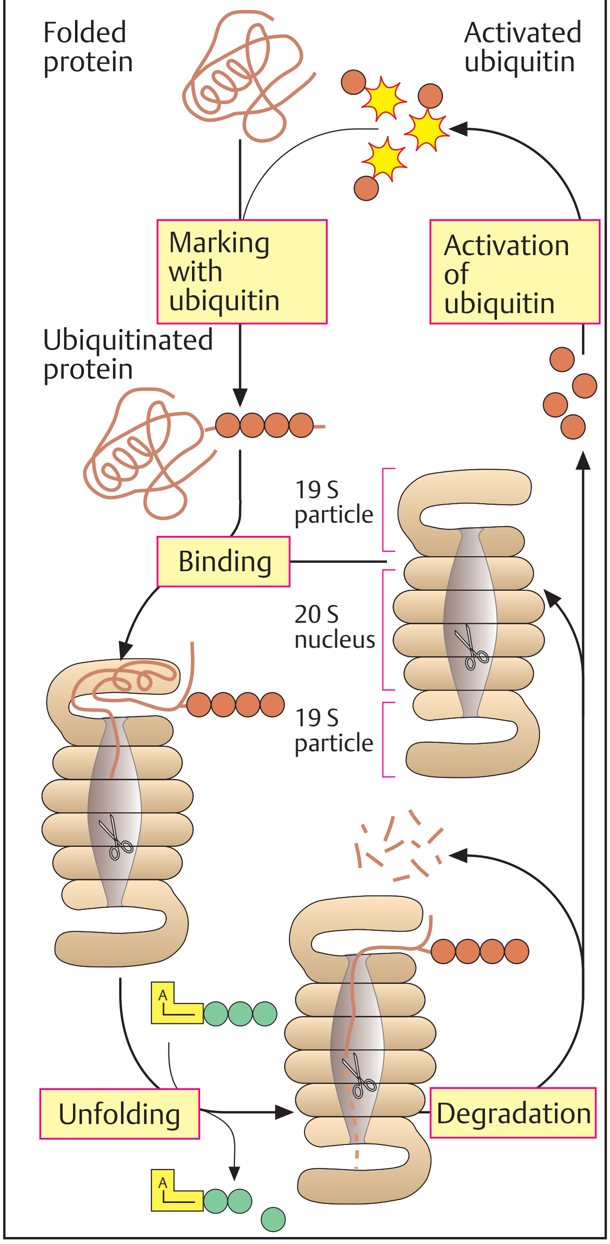
Folded protein
Activated ubiquitin
Marking with ubiquitin
Activation of ubiquitin
Ubiquitinated protein
19 S particle
20 S nucleus
19 S particle
Binding
Unfolding
Degradation
A

C. Serine proteases

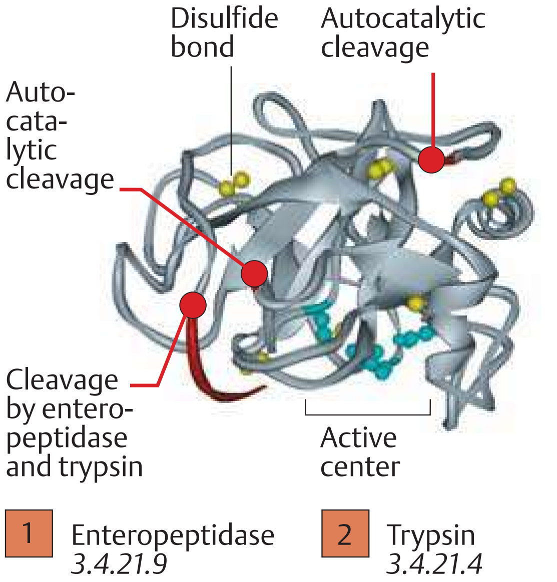
Disulfide bond
Autocatalytic cleavage
Auto-catalytic cleavage
Cleavage by entero-peptidase and trypsin
Active center
1 Enteropeptidase 3.4.21.9
2 Trypsin 3.4.21.4

1. Trypsinogen activation

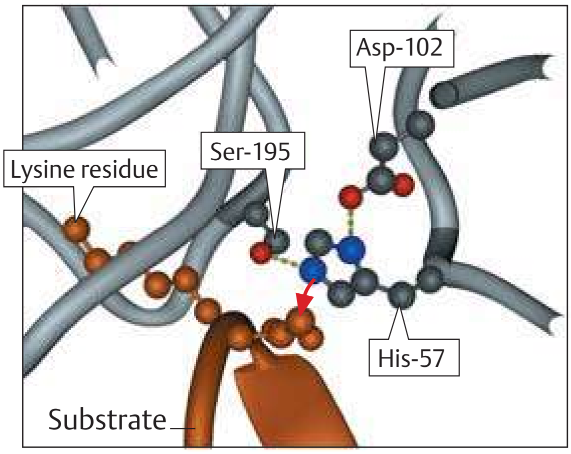
Asp-102
Lysine residue
Ser-195
His-57
Substrate

2. Trypsin: active center

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Transamination and deamination

Amino nitrogen accumulates during protein degradation. In contrast to carbon, amino nitrogen is not suitable for oxidative energy production. If they are not being reused for biosynthesis, the amino groups of amino acids are therefore incorporated into urea (see p. 182) and excreted in this form.

A. Transamination and deamination ◑

Among the NH_2 transfer reactions, **transaminations** (**1**) are particularly important. They are catalyzed by *transaminases*, and occur in both catabolic and anabolic amino acid metabolism. During transamination, the amino group of an amino acid (amino acid 1) is transferred to a 2-oxoacid (oxoacid 2). From the amino acid, this produces a 2-oxoacid (a), while from the original oxoacid, an amino acid is formed (b). The NH_2 group is temporarily taken over by enzyme-bound **pyridoxal phosphate** (PLP; see p. 106), which thus becomes pyridoxamine phosphate.

If the NH_2 is released as ammonia, the process is referred to as **deamination**. There are different mechanisms for this (see p. 180). A particularly important one is **oxidative deamination** (**2**). In this reaction, the α-amino group is initially *oxidized* into an imino group (2a), and the reducing equivalents are transferred to NAD^+ or $NADP^+$. In the second step, the imino group is then cleaved by *hydrolysis*. As in transamination, this produces a 2-oxoacid (**C**). Oxidative deamination mainly takes place in the liver, where glutamate is broken down in this way into 2-oxoglutarate and ammonia, catalyzed by *glutamate dehydrogenase.* The reverse reaction initiates biosynthesis of the amino acids in the glutamate family (see p. 184).

B. Mechanism of transamination ○

In the absence of substrates, the aldehyde group of pyridoxal phosphate is covalently bound to a lysine residue of the transaminase (**1**). This type of compound is known as an **aldimine** or "Schiff's base." During the reaction, amino acid 1 (**A**, **1a**) displaces the lysine residue, and a new aldimine is formed (**2**). The double bond is then shifted by isomerization. The ketimine (**3**) is hydrolyzed to yield the 2-oxoacid and **pyridoxamine phosphate** (**4**).

In the second part of the reaction (see **A**, 1b), these steps take place *in the opposite direction:* pyridoxamine phosphate and the second 2-oxoacid form a ketimine, which is isomerized into aldimine. Finally, the second amino acid is cleaved and the coenzyme is regenerated.

C. NH_3 metabolism in the liver ◑

In addition to urea synthesis itself (see p. 182), the precursors NH_3 and aspartate are also mainly formed in the liver. Amino nitrogen arising in tissue is transported to the liver by the blood, mainly in the form of **glutamine** (Gln) and **alanine** (Ala; see p. 338). In the liver, Gln is hydrolytically deaminated by *glutaminase* [3] into **glutamate** (Glu) and NH_3. The amino group of the alanine is transferred by *alanine transaminase* [1] to **2-oxoglutarate** (2-OG; formerly known as α-ketoglutarate). This transamination (**A**) produces another glutamate. NH_3 is finally released from glutamate by oxidative deamination (**A**). This reaction is catalyzed by *glutamate dehydrogenase* [4], a typical liver enzyme. **Aspartate** (Asp), the second amino group donor in the urea cycle, also arises from glutamate. The *aspartate transaminase* [2] responsible for this reaction is found with a high level of activity in the liver, as is *alanine transaminase* [1].

Transaminases are also found in other tissues, from which they leak from the cells into the blood when injury occurs. Measurement of serum enzyme activity (**serum enzyme diagnosis**; see also p. 98) is an important method of recognizing and monitoring the course of such injuries. Transaminase activity in the blood is for instance important for diagnosing liver disease (e.g., hepatitis) and myocardial disease (cardiac infarction).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Transamination and deamination

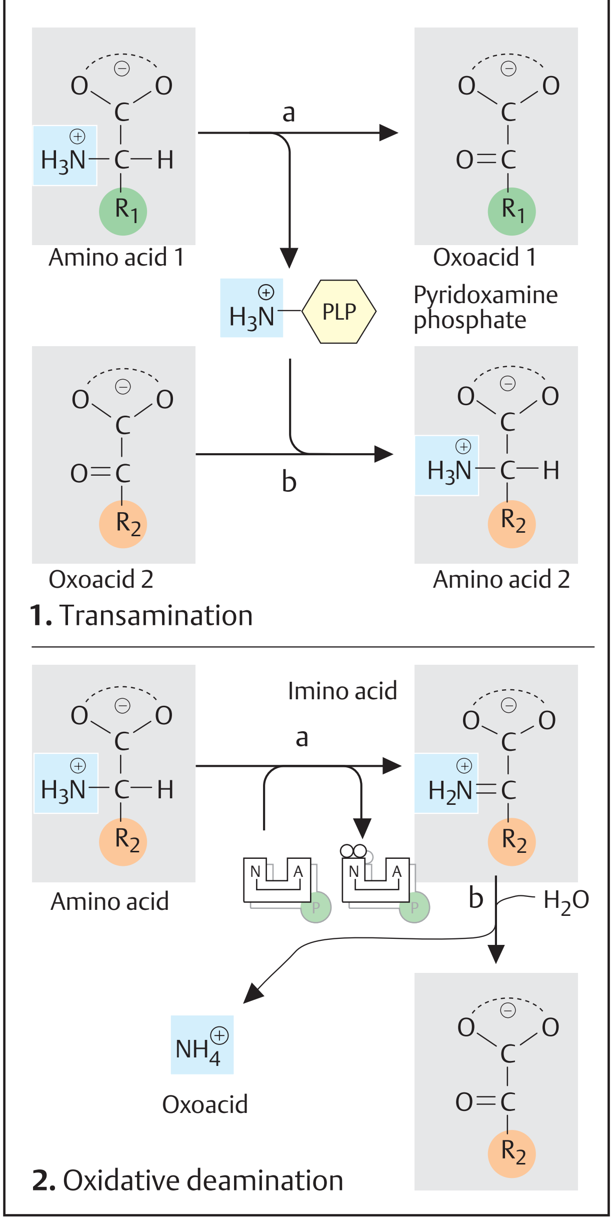
Amino acid 1
a
Oxoacid 1
$H_3N^{\oplus}$
PLP
Pyridoxamine phosphate
Oxoacid 2
b
Amino acid 2
1. Transamination
Imino acid
a
Amino acid
b
H_2O
$NH_4^{\oplus}$
Oxoacid
2. Oxidative deamination

B. Mechanism of transamination

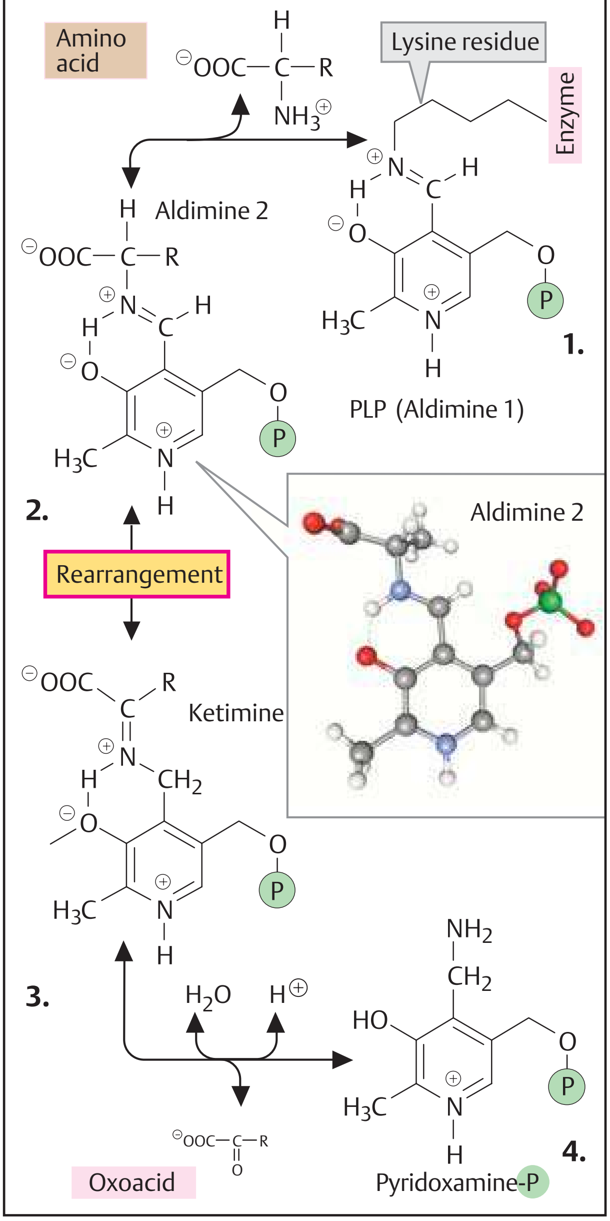
Amino acid
Lysine residue
Enzyme
Aldimine 2
1.
PLP (Aldimine 1)
2.
Rearrangement
Aldimine 2
Ketimine
3.
H_2O
$H^{\oplus}$
4.
Oxoacid
Pyridoxamine-P

C. NH_3 metabolism in the liver

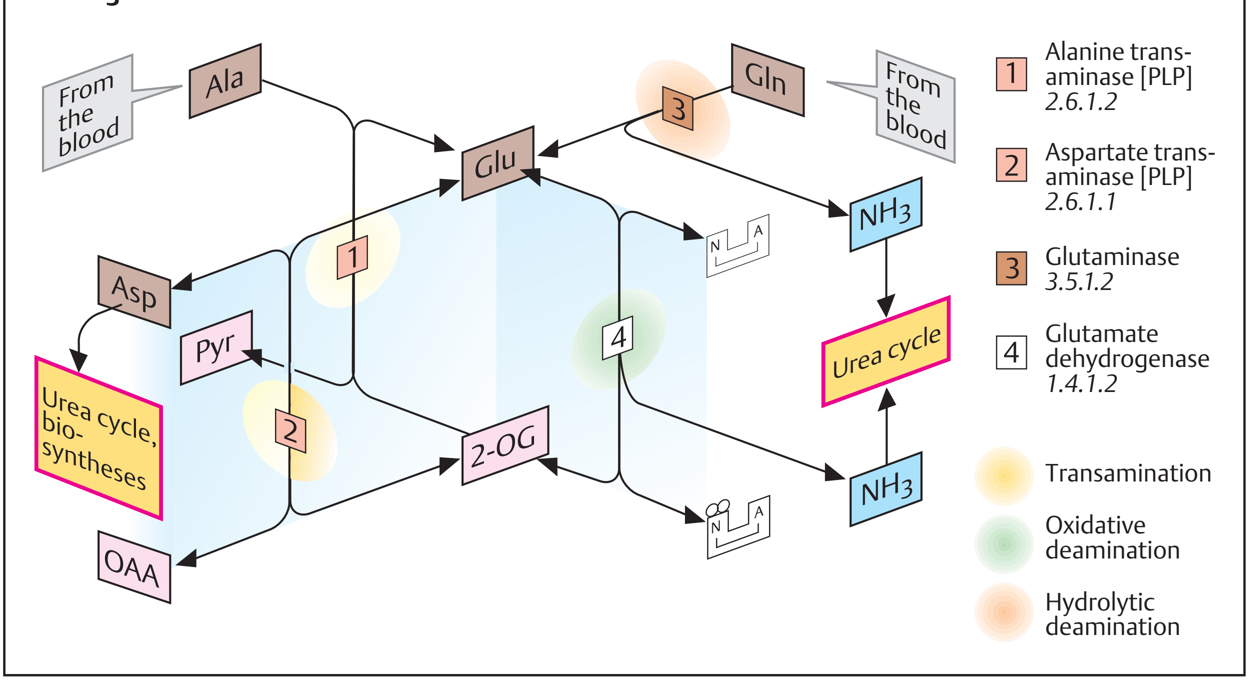
From the blood
Ala
Glu
Gln
From the blood
3
1
NH_3
Asp
Pyr
4
Urea cycle
Urea cycle, bio-syntheses
2
2-OG
NH_3
OAA
1 Alanine transaminase [PLP] 2.6.1.2
2 Aspartate transaminase [PLP] 2.6.1.1
3 Glutaminase 3.5.1.2
4 Glutamate dehydrogenase 1.4.1.2
Transamination
Oxidative deamination
Hydrolytic deamination

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Amino acid degradation

A large number of metabolic pathways are available for amino acid degradation, and an overview of these is presented here. Further details are given on pp. 414 and 415.

A. Amino acid degradation : overview ◑

During the degradation of most amino acids, the α-amino group is initially removed by **transamination** or **deamination**. Various mechanisms are available for this, and these are discussed in greater detail in **B**. The carbon skeletons that are left over after deamination undergo further degradation in various ways.

During degradation, the 20 proteinogenic amino acids produce only seven different **degradation products** (highlighted in pink and violet). Five of these metabolites (2-oxoglutarate, succinyl CoA, fumarate, oxaloacetate, and pyruvate) are precursors for gluconeogenesis and can therefore be converted into glucose by the liver and kidneys (see p. 154). Amino acids whose degradation supplies one of these five metabolites are therefore referred to as **glucogenic amino acids**. The first four degradation products listed are already intermediates in the tricarboxylic acid cycle, while pyruvate can be converted into oxaloacetate by *pyruvate carboxylase* and thus made available for gluconeogenesis (green arrow).

With two exceptions (lysine and leucine; see below), all of the proteinogenic amino acids are also glucogenic. Quantitatively, they represent the most important precursors for gluconeogenesis. At the same time, they also have an **anaplerotic** effect—i. e., they replenish the tricarboxylic acid cycle in order to feed the anabolic reactions that originate in it (see p. 138).

Two additional degradation products (acetoacetate and acetyl CoA) cannot be channeled into gluconeogenesis in animal metabolism, as there is no means of converting them into precursors of gluconeogenesis. However, they can be used to synthesize ketone bodies, fatty acids, and isoprenoids. Amino acids that supply acetyl CoA or acetoacetate are therefore known as **ketogenic amino acids.** Only leucine and lysine are *purely* ketogenic. Several amino acids yield degradation products that are both *glucogenic and ketogenic.* This group includes phenylalanine, tyrosine, tryptophan, and isoleucine.

Degradation of acetoacetate to acetyl CoA takes place in two steps (not shown). First, acetoacetate and succinyl CoA are converted into acetoacetyl CoA and succinate (enzyme: *3-oxoacid-CoA transferase 2.8.3.5*). Acetoacetyl CoA is then broken down by β-oxidation into two molecules of acetyl CoA (see p. 164), while succinate can be further metabolized via the tricarboxylic acid cycle.

B. Deamination ○

There are various ways of releasing ammonia (NH_3) from amino acids, and these are illustrated here using the example of the amino acids glutamine, glutamate, alanine, and serine.

[1] In the branched-chain amino acids (Val, Leu, Ile) and also tyrosine and ornithine, degradation starts with a **transamination.** For alanine and aspartate, this is actually the only degradation step. The mechanism of transamination is discussed in detail on p. 178.

[2] **Oxidative deamination**, with the formation of $NADH+H^+$, only applies to glutamate in animal metabolism. The reaction mainly takes place in the liver and releases NH_3 for urea formation (see p. 178).

[3] Two amino acids—asparagine and glutamine—contain acid–amide groups in the side chains, from which NH_3 can be released by hydrolysis (**hydrolytic deamination**). In the blood, glutamine is the most important transport molecule for amino nitrogen. Hydrolytic deamination of glutamine in the liver also supplies the urea cycle with NH_3.

[4] **Eliminating deamination** takes place in the degradation of histidine and serine. H_2O is first eliminated here, yielding an unsaturated intermediate. In the case of serine, this intermediate is first rearranged into an imine (not shown), which is hydrolyzed in the second step into NH_3 and pyruvate, with H_2O being taken up. H_2O does not therefore appear in the reaction equation.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Amino acid degradation: overview

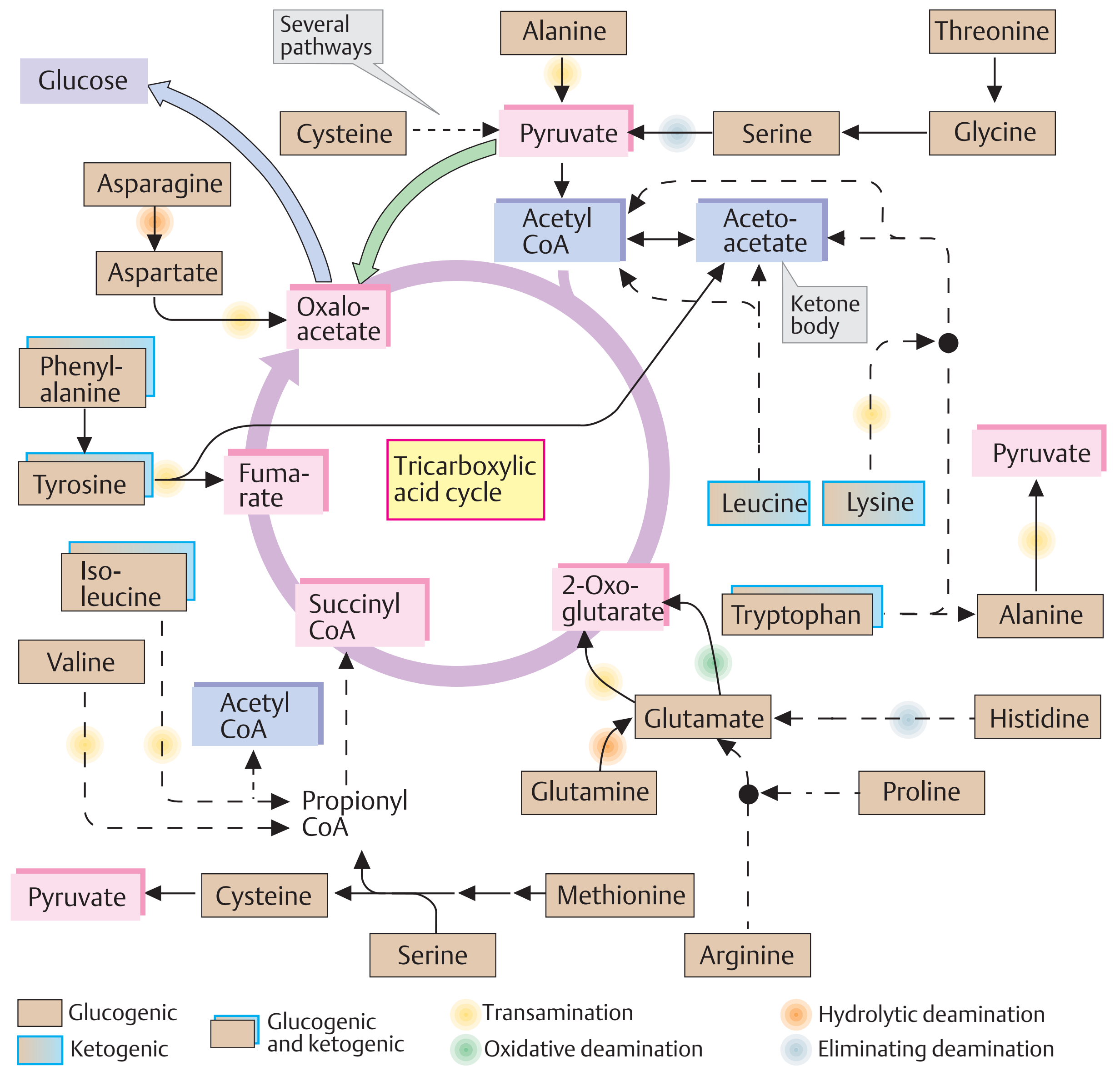
Several pathways
Alanine
Threonine
Glucose
Cysteine
Pyruvate
Serine
Glycine
Asparagine
Acetyl CoA
Aceto-acetate
Aspartate
Oxalo-acetate
Ketone body
Phenyl-alanine
Tyrosine
Fuma-rate
Tricarboxylic acid cycle
Leucine
Lysine
Pyruvate
Iso-leucine
Succinyl CoA
2-Oxo-glutarate
Tryptophan
Alanine
Valine
Acetyl CoA
Glutamate
Histidine
Glutamine
Proline
Propionyl CoA
Pyruvate
Cysteine
Methionine
Serine
Arginine
Glucogenic
Ketogenic
Glucogenic and ketogenic
Transamination
Oxidative deamination
Hydrolytic deamination
Eliminating deamination

B. Deamination

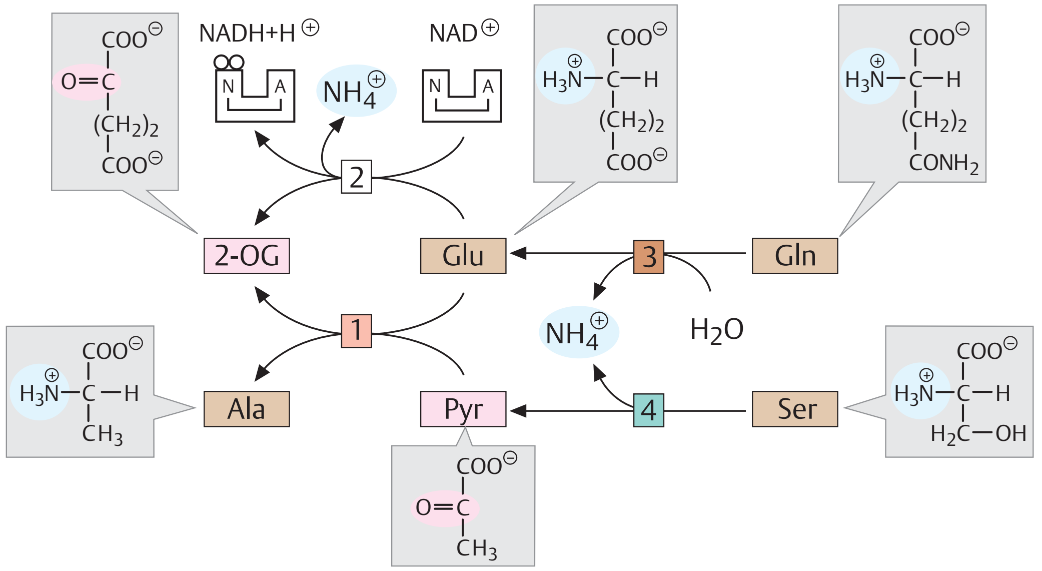
COO⊖
O=C
(CH2)2
COO⊖
NADH+H⊕
N A
NH4⊕
NAD⊕
N A
COO⊖
⊕ H3N–C–H
(CH2)2
COO⊖
COO⊖
⊕ H3N–C–H
(CH2)2
CONH2
2
2-OG
Glu
3
Gln
NH4⊕
H2O
1
COO⊖
⊕ H3N–C–H
CH3
Ala
Pyr
4
Ser
COO⊖
⊕ H3N–C–H
H2C–OH
COO⊖
O=C
CH3

1 Alanine transaminase [PLP] *2.6.1.2*

2 Glutamate dehydrogenase *1.4.1.2*

3 Glutaminase *3.5.1.2*

4 Serine dehydratase [PLP] *4.2.1.13*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Urea cycle

Amino acids are mainly broken down in the liver. Ammonia is released either directly or indirectly in the process (see p. 178). The degradation of nucleobases also provides significant amounts of ammonia (see p. 186).

Ammonia (NH_3) is a relatively strong **base**, and at physiological pH values it is mainly present in the form of the **ammonium ion** $\mathbf{NH_4^+}$ (see p. 30). NH_3 and NH_4^+ are toxic, and at higher concentrations cause brain damage in particular. Ammonia therefore has to be effectively inactivated and excreted. This can be carried out in various ways. Aquatic animals can excrete NH_4^+ directly. For example, fish excrete NH_4^+ via the gills (*ammonotelic animals*). Terrestrial vertebrates, including humans, hardly excrete any NH_3, and instead, most ammonia is converted into urea before excretion (*ureotelic animals*). Birds and reptiles, by contrast, form *uric acid,* which is mainly excreted as a solid in order to save water (*uricotelic animals*).

The reasons for the neurotoxic effects of ammonia have not yet been explained. It may disturb the metabolism of glutamate and its precursor glutamine in the brain (see p. 356).

A. Urea cycle ◑

Urea ($H_2N–CO–NH_2$) is the diamide of carbonic acid. In contrast to ammonia, it is **neutral** and therefore relatively **non-toxic**. The reason for the lack of basicity is the molecule's mesomeric characteristics. The free electron pairs of the two nitrogen atoms are *delocalized* over the whole structure, and are therefore no longer able to bind protons. As a small, uncharged molecule, urea is able to cross biological membranes easily. In addition, it is easily transported in the blood and excreted in the urine.

Urea is produced **only in the liver**, in a cyclic sequence of reactions (the **urea cycle**) that starts in the mitochondria and continues in the cytoplasm. The two nitrogen atoms are derived from $\mathbf{NH_4^+}$ (the second has previously been incorporated into aspartate; see below). The keto group comes from **hydrogen carbonate** (HCO_3^-), or CO_2 that is in equilibrium with HCO_3^-.

[1] In the first step, **carbamoyl phosphate** is formed in the mitochondria from hydrogen carbonate (HCO_3^-) and NH_4^+, with two ATP molecules being consumed. In this compound, the carbamoyl residue ($–O–CO–NH_2$) is at a high chemical potential. In hepatic mitochondria, enzyme [1] makes up about 20% of the matrix proteins.

[2] In the next step, the carbamoyl residue is transferred to the non-proteinogenic amino acid **ornithine**, converting it into **citrulline**, which is also non-proteinogenic. This is passed into the cytoplasm via a transporter.

[3] The second NH_2 group of the later urea molecule is provided by **aspartate**, which condenses with citrulline into **argininosuccinate**. ATP is cleaved into AMP and diphosphate (PP_i) for this endergonic reaction. To shift the equilibrium of the reaction to the side of the product, diphosphate is removed from the equilibrium by hydrolysis.

[4] Cleavage of fumarate from argininosuccinate leads to the proteinogenic amino acid **arginine**, which is synthesized in this way in animal metabolism.

[5] In the final step, isourea is released from the guanidinium group of the arginine by hydrolysis (not shown), and is immediately rearranged into **urea**. In addition, ornithine is regenerated and returns via the ornithine transporter into the mitochondria, where it becomes available for the cycle once again.

The **fumarate** produced in step [4] is converted via malate to oxaloacetate [6, 7], from which **aspartate** is formed again by transamination [9]. The glutamate required for reaction [9] is derived from the glutamate dehydrogenase reaction [8], which fixes the second NH_4^+ in an organic bond. Reactions [6] and [7] also occur in the tricarboxylic acid cycle. However, in urea formation they take place in the cytoplasm, where the appropriate isoenzymes are available.

The rate of urea formation is mainly controlled by reaction [1]. *N*-**acetyl glutamate**, as an allosteric effector, activates *carbamoyl-phosphate synthase.* In turn, the concentration of acetyl glutamate depends on arginine and ATP levels, as well as other factors.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Urea cycle

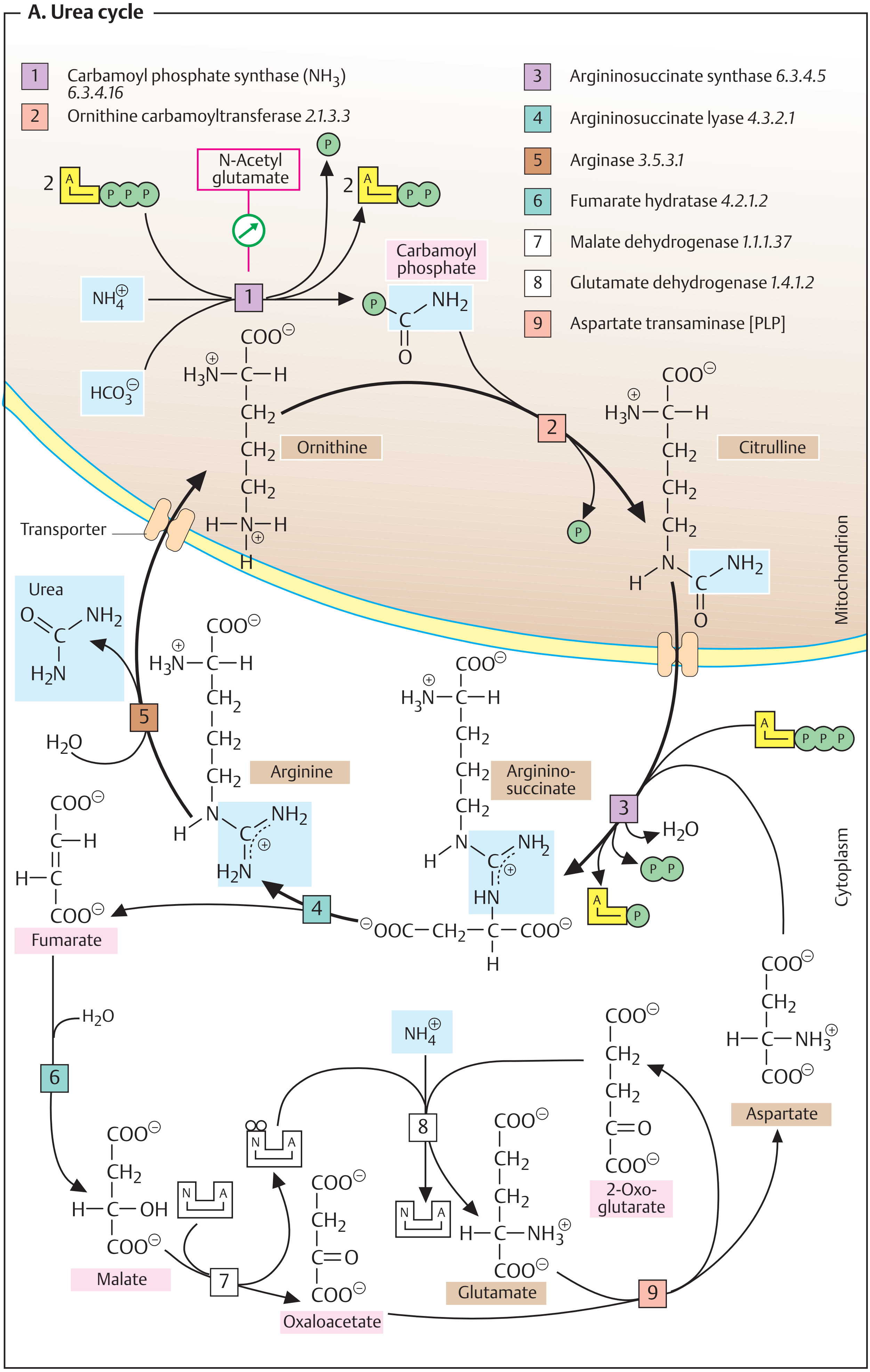
1 Carbamoyl phosphate synthase (NH3) 6.3.4.16
2 Ornithine carbamoyltransferase 2.1.3.3
3 Argininosuccinate synthase 6.3.4.5
4 Argininosuccinate lyase 4.3.2.1
5 Arginase 3.5.3.1
6 Fumarate hydratase 4.2.1.2
7 Malate dehydrogenase 1.1.1.37
8 Glutamate dehydrogenase 1.4.1.2
9 Aspartate transaminase [PLP]
N-Acetyl glutamate
2 A P P P
2 A P P
P
NH4+
HCO3−
Carbamoyl phosphate
P C NH2 O
Ornithine
Citrulline
Transporter
Mitochondrion
Cytoplasm
Urea
O C NH2 H2N
H2O
Arginine
Argininosuccinate
Fumarate
A P P P
H2O
P P
A P
6
H2O
Malate
7
N A
Oxaloacetate
8
NH4+
Glutamate
2-Oxo-glutarate
Aspartate
9

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Amino acid biosynthesis

A. Symbiotic nitrogen fixation ○

Practically unlimited quantities of elementary nitrogen (N_2) are present in the atmosphere. However, before it can enter the natural nitrogen cycle, it has to be reduced to NH_3 and incorporated into amino acids ("fixed"). Only a few species of bacteria and bluegreen algae are capable of fixing atmospheric nitrogen. These exist freely in the soil, or in **symbiosis** with plants. The symbiosis between bacteria of the genus *Rhizobium* and legumes (*Fabales*)—such as clover, beans, and peas—is of particular economic importance. These plants are high in protein and are therefore nutritionally valuable.

In symbiosis with *Fabales*, bacteria live as *bacteroids* in **root nodules** inside the plant cells. The plant supplies the bacteroids with nutrients, but it also benefits from the fixed nitrogen that the symbionts make available.

The N_2-fixing enzyme used by the bacteria is *nitrogenase*. It consists of two components: an *Fe protein* that contains an $[Fe_4S_4]$ cluster as a redox system (see p. 106), accepts electrons from *ferredoxin*, and donates them to the second component, the *Fe–Mo protein*. This molybdenum-containing protein transfers the electrons to N_2 and thus, via various intermediate steps, produces ammonia (NH_3). Some of the reducing equivalents are transferred in a side-reaction to H^+. In addition to NH_3, hydrogen is therefore always produced as well.

B. Amino acid biosynthesis: overview ◑

The proteinogenic amino acids (see p. 60) can be divided into **five families** in relation to their biosynthesis. The members of each family are derived from common precursors, which are all produced in the tricarboxylic acid cycle or in catabolic carbohydrate metabolism. An overview of the biosynthetic pathways is shown here; further details are given on pp. 412 and 413.

Plants and microorganisms are able to synthesize all of the amino acids from scratch, but during the course of evolution, mammals have lost the ability to synthesize approximately half of the 20 proteinogenic amino acids. These **essential amino acids** therefore have to be supplied in food. For example, animal metabolism is no longer capable of carrying out de-novo synthesis of the **aromatic amino acids** (tyrosine is only non-essential because it can be formed from phenylalanine when there is an adequate supply available). The **branched-chain amino acids** (valine, leucine, isoleucine, and threonine) as well as **methionine** and **lysine**, also belong to the essential amino acids. Histidine and arginine are essential in rats; whether the same applies in humans is still a matter of debate. A supply of these amino acids in food appears to be essential at least during growth.

The nutritional value of proteins (see p. 360) is decisively dependent on their essential amino acid content. Vegetable proteins—e. g., those from cereals—are low in lysine and methionine, while animal proteins contain all the amino acids in balanced proportions. As mentioned earlier, however, there are also plants that provide high-value protein. These include the soy bean, one of the plants that is supplied with NH_3 by symbiotic N_2 fixers (**A**).

Non-essential amino acids are those that arise by transamination from 2-oxoacids in the intermediary metabolism. These belong to the **glutamate family** (Glu, Gln, Pro, Arg, derived from 2-oxoglutarate), the **aspartate family** (only Asp and Asn in this group, derived from oxaloacetate), and **alanine**, which can be formed by transamination from pyruvate. The amino acids in the **serine family** (Ser, Gly, Cys) and **histidine**, which arise from intermediates of glycolysis, can also be synthesized by the human body.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Symbiotic nitrogen fixation

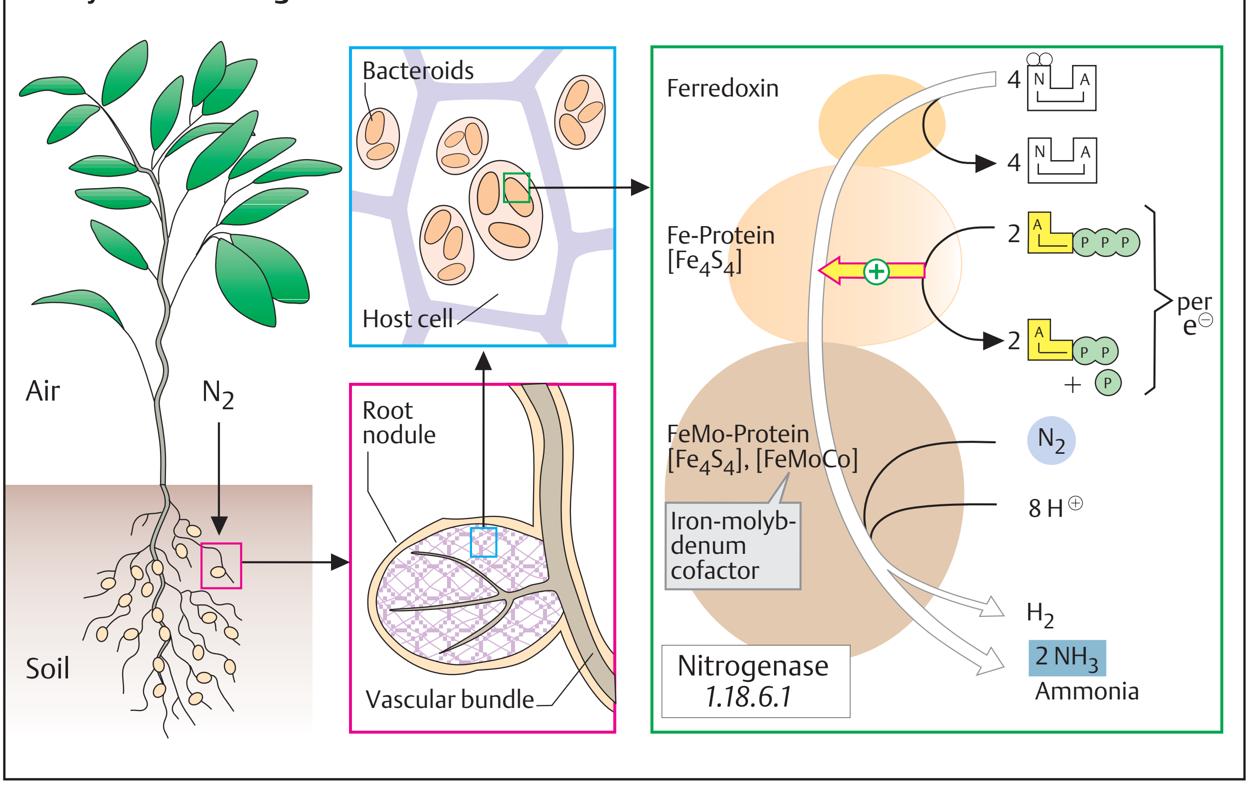
Air
N_2
Soil
Bacteroids
Host cell
Root nodule
Vascular bundle
Ferredoxin
4
4
Fe-Protein $[Fe_4S_4]$
2
2
+
per $e^⊖$
FeMo-Protein $[Fe_4S_4]$, [FeMoCo]
Iron-molybdenum cofactor
N_2
8 $H^⊕$
H_2
2 NH_3
Ammonia
Nitrogenase
1.18.6.1

B. Amino acid biosynthesis: overview

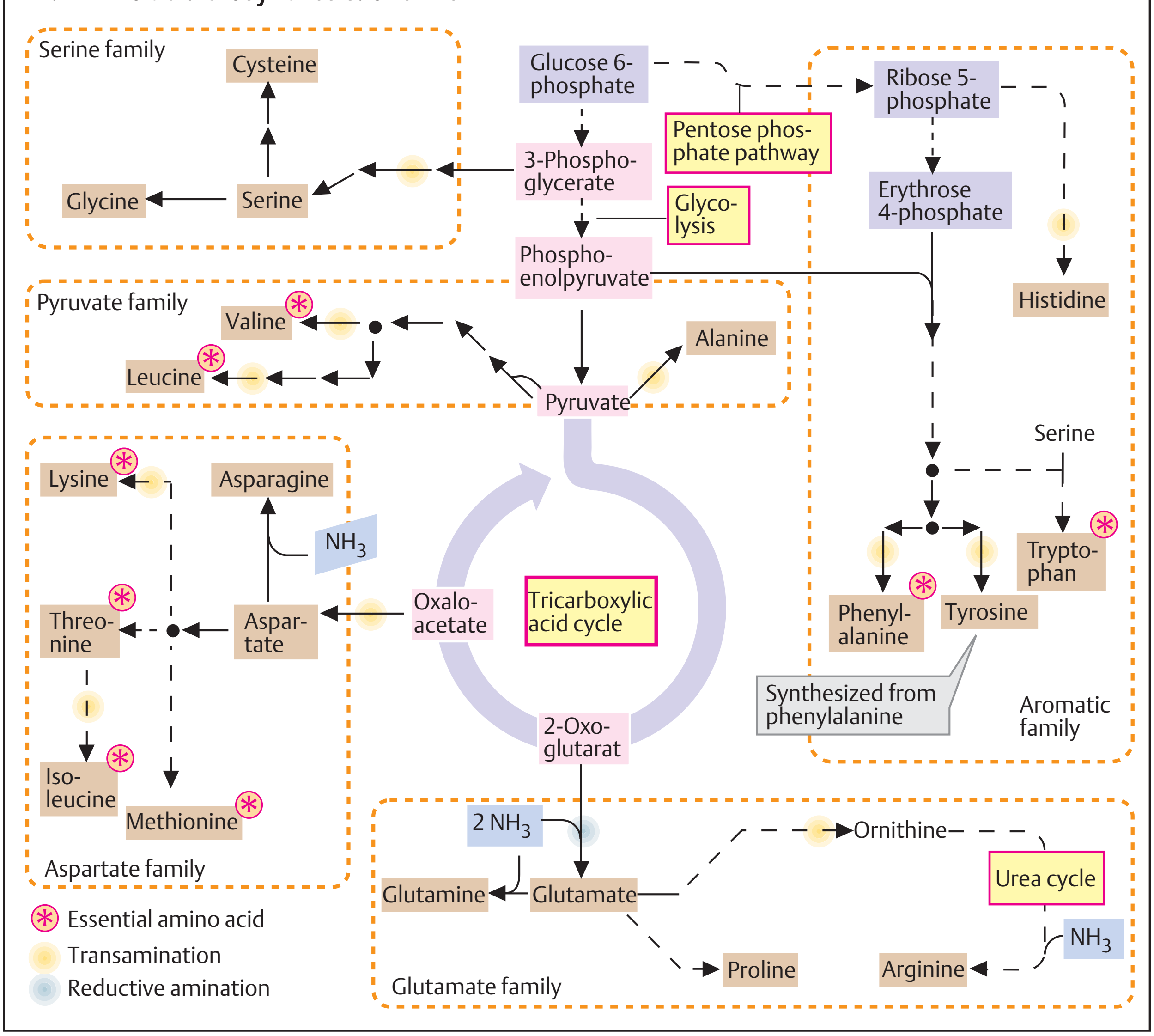
Serine family
Cysteine
Glycine
Serine
Glucose 6-phosphate
3-Phosphoglycerate
Pentose phosphate pathway
Glycolysis
Phosphoenolpyruvate
Ribose 5-phosphate
Erythrose 4-phosphate
Histidine
Pyruvate family
Valine
Leucine
Alanine
Pyruvate
Serine
Lysine
Asparagine
NH_3
Threonine
Aspartate
Oxaloacetate
Tricarboxylic acid cycle
Isoleucine
Methionine
Aspartate family
2-Oxoglutarat
Phenylalanine
Tyrosine
Tryptophan
Synthesized from phenylalanine
Aromatic family
2 NH_3
Glutamine
Glutamate
Ornithine
Urea cycle
NH_3
Proline
Arginine
Glutamate family
Essential amino acid
Transamination
Reductive amination

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Nucleotide degradation

The nucleotides are among the most complex metabolites. Nucleotide biosynthesis is elaborate and requires a high energy input (see p. 188). Understandably, therefore, bases and nucleotides are not completely degraded, but instead mostly recycled. This is particularly true of the purine bases adenine and guanine. In the animal organism, some 90% of these bases are converted back into nucleoside monophosphates by linkage with phosphoribosyl diphosphate (PRPP) (enzymes [1] and [2]). The proportion of pyrimidine bases that are recycled is much smaller.

A. Degradation of nucleotides ◑

The principles underlying the degradation of purines (**1**) and pyrimidines (**2**) differ. In the human organism, purines are degraded into uric acid and excreted in this form. The purine ring remains intact in this process. In contrast, the ring of the pyrimidine bases (uracil, thymine, and cytosine) is broken down into small fragments, which can be returned to the metabolism or excreted (for further details, see p. 419).

Purine (left). The purine nucleotide **guanosine monophosphate** (**GMP**, **1**) is degraded in two steps—first to the *guanosine* and then to *guanine* (Gua). Guanine is converted by deamination into another purine base, *xanthine*.

In the most important degradative pathway for **adenosine monophosphate** (**AMP**), it is the nucleotide that deaminated, and *inosine monophosphate* (IMP) arises. In the same way as in GMP, the purine base *hypoxanthine* is released from IMP. A single enzyme, *xanthine oxidase* [3], then both converts hypoxanthine into xanthine and xanthine into **uric acid**. An oxo group is introduced into the substrate in each of these reaction steps. The oxo group is derived from *molecular oxygen;* another reaction product is *hydrogen peroxide* (H_2O_2), which is toxic and has to be removed by peroxidases.

Almost all mammals carry out further degradation of uric acid with the help of *uricase*, with further opening of the ring to **allantoin**, which is then excreted. However, the primates, including humans, are not capable of synthesizing allantoin. *Uric acid* is therefore the form of the purines excreted in these species. The same applies to birds and many reptiles. Most other animals continue purine degradation to reach allantoic acid or urea and glyoxylate.

Pyrimidine (right). In the degradation of pyrimidine nucleotides (**2**), the free bases *uracil* (Ura) and *thymine* (Thy) are initially released as important intermediates. Both are further metabolized in similar ways. The pyrimidine ring is first reduced and then hydrolytically cleaved. In the next step, *β-alanine* arises by cleavage of CO_2 and NH_3 as the degradation product of uracil. When there is further degradation, *β-alanine* is broken down to yield acetate, CO_2, and NH_3. Propionate, CO_2, and NH_3 arise in a similar way from *γ-aminoisobutyrate,* the degradation product of thymine (see p. 419).

B. Hyperuricemia ○

The fact that purine degradation in humans already stops at the uric acid stage can lead to problems, since—in contrast to allantoin—uric acid is *poorly soluble in water.* When large amounts of uric acid are formed or uric acid processing is disturbed, excessive concentrations of uric acid can develop in the blood (*hyperuricemia*). This can result in the accumulation of uric acid crystals in the body. Deposition of these crystals in the joints can cause very painful attacks of **gout.**

Most cases of hyperuricemia are due to disturbed uric acid excretion via the kidneys (1). A high-purine diet (e. g., meat) may also have unfavorable effects (2). A rare hereditary disease, *Lesch–Nyhan syndrome,* results from a defect in *hypoxanthine phosphoribosyltransferase* (A, enzyme [1]). The impaired recycling of the purine bases caused by this leads to hyperuricemia and severe neurological disorders.

Hyperuricemia can be treated with *allopurinol,* a competitive inhibitor of xanthine oxidase. This substrate analogue differs from the substrate hypoxanthine only in the arrangement of the atoms in the 5-ring.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Degradation of nucleotides

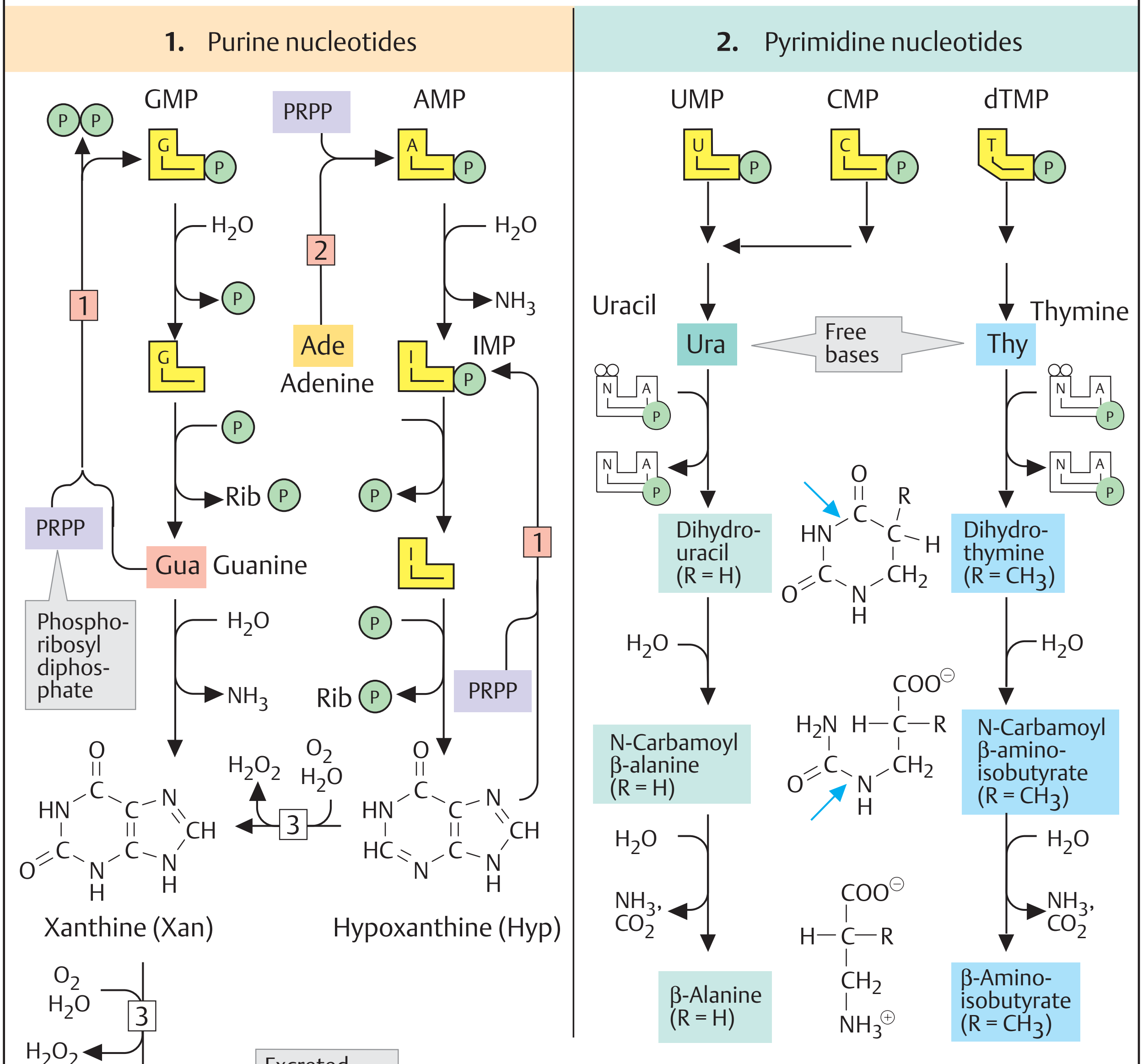
1. Purine nucleotides
2. Pyrimidine nucleotides
GMP
PRPP
AMP
P
G
A
H_2O
NH_3
2
1
Ade
Adenine
IMP
I
Rib
Gua
Guanine
PRPP
Phosphoribosyl diphosphate
H_2O_2
O_2
3
Xanthine (Xan)
Hypoxanthine (Hyp)
UMP
CMP
dTMP
U
C
T
Uracil
Ura
Free bases
Thy
Thymine
N
Dihydrouracil (R = H)
Dihydrothymine (R = CH_3)
N-Carbamoyl β-alanine (R = H)
N-Carbamoyl β-amino-isobutyrate (R = CH_3)
NH_3, CO_2
β-Alanine (R = H)
β-Amino-isobutyrate (R = CH_3)

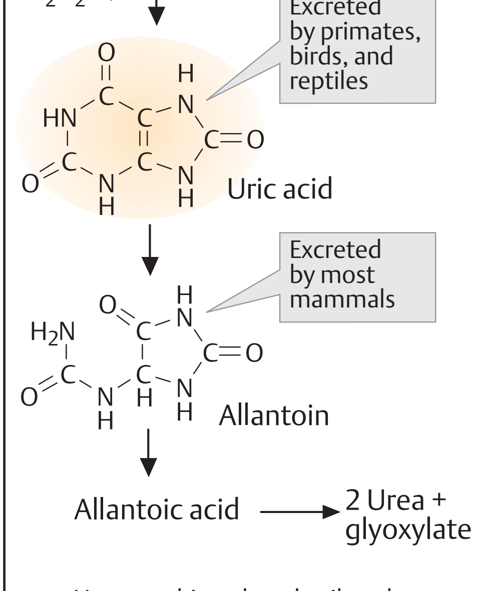
O_2
H_2O
3
H_2O_2
Uric acid
Excreted by primates, birds, and reptiles
Excreted by most mammals
Allantoin
Allantoic acid
2 Urea + glyoxylate

1 Hypoxanthine phosphoribosyltransferase *2.4.2.8*

2 Adenine phosphoribosyltransferase *2.4.2.7*

3 Xanthine oxidase [Fe, Mo, FAD] *1.1.3.22*

B. Hyperuricemia (gout)

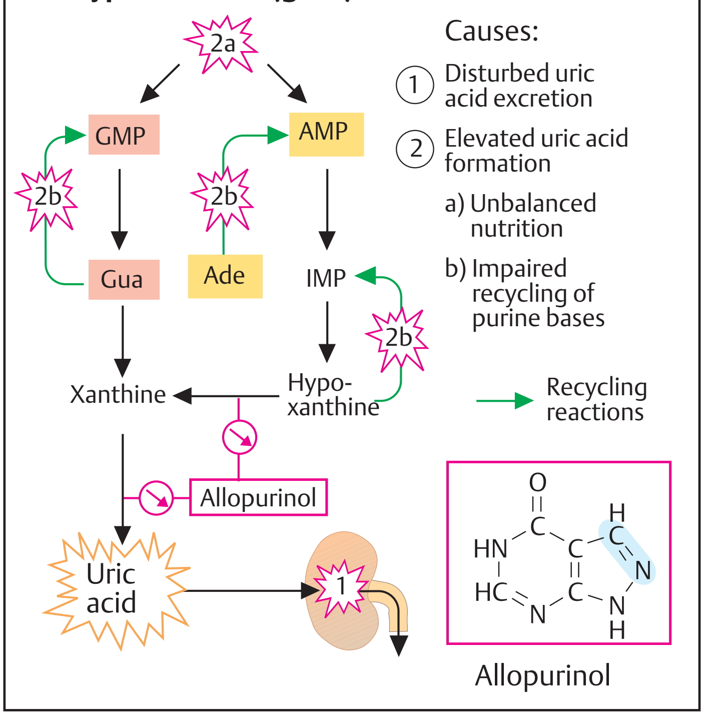
2a
GMP
AMP
2b
Gua
Ade
IMP
Xanthine
Hypoxanthine
Allopurinol
Uric acid
1
Recycling reactions
Allopurinol

Causes:

① Disturbed uric acid excretion

② Elevated uric acid formation

a) Unbalanced nutrition

b) Impaired recycling of purine bases

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Purine and pyrimidine biosynthesis

The bases occurring in nucleic acids are derivatives of the aromatic heterocyclic compounds *purine* and *pyrimidine* (see p. 80). The biosynthesis of these molecules is complex, but is vital for almost all cells. The synthesis of the nucleobases is illustrated here schematically. Complete reaction schemes are given on pp. 417 and 418.

A. Components of nucleobases ○

The **pyrimidine ring** is made up of three components: the nitrogen atom N-1 and carbons C-4 to C-6 are derived from *aspartate,* carbon C-2 comes from HCO_3^-, and the second nitrogen (N-3) is taken from the amide group of *glutamine*.

The synthesis of the **purine ring** is more complex. The only major component is *glycine*, which donates C-4 and C-5, as well as N-7. All of the other atoms in the ring are incorporated individually. C-6 comes from HCO_3^-. Amide groups from *glutamine* provide the atoms N-3 and N-9. The amino group donor for the inclusion of N-1 is *aspartate,* which is converted into fumarate in the process, in the same way as in the urea cycle (see p. 182). Finally, the carbon atoms C-2 and C-8 are derived from formyl groups in N^{10}-formyl-tetrahydrofolate (see p. 108).

B. Pyrimidine and purine synthesis ○

The major intermediates in the biosynthesis of nucleic acid components are the mononucleotides *uridine monophosphate* (UMP) in the pyrimidine series and *inosine monophosphate* (IMP, base: hypoxanthine) in the purines. The synthetic pathways for pyrimidines and purines are fundamentally different. For the pyrimidines, the pyrimidine ring is first constructed and then linked to ribose 5′-phosphate to form a nucleotide. By contrast, synthesis of the purines starts directly from ribose 5′-phosphate. The ring is then built up step by step on this carrier molecule.

The precursors for the synthesis of the pyrimidine ring are **carbamoyl phosphate**, which arises from glutamate and HCO_3^- (**1a**) and the amino acid **aspartate.** These two components are linked to *N*-**carbamoyl aspartate** (**1b**) and then converted into **dihydroorotate** by closure of the ring (**1c**). In mammals, steps 1a to 1c take place in the cytoplasm, and are catalyzed by a single multifunctional enzyme. In the next step (**1d**), dihydroorotate is oxidized to **orotate** by an FMN-dependent dehydrogenase. Orotate is then linked with **phosphoribosyl diphosphate** (PRPP) to form the nucleotide **orotidine 5′-monophosphate (OMP)**. Finally, decarboxylation yields **uridine 5′-monophosphate (UMP)**.

Purine biosynthesis starts with PRPP (the names of the individual intermediates are given on p. 417). Formation of the ring starts with transfer of an amino group, from which the later N-9 is derived (**2a**). Glycine and a formyl group from N^{10}-formyl-THF then supply the remaining atoms of the five-membered ring (**2b**, **2c**). Before the five-membered ring is closed (in step **2f**), atoms N-3 and C-6 of the later six-membered ring are attached (**2d**, **2e**). Synthesis of the ring then continues with N-1 and C-2 (**2g**, **2i**). In the final step (**2j**), the six-membered ring is closed, and **inosine 5′-monophosphate** arises. However, the IMP formed does not accumulate, but is rapidly converted into AMP and GMP. These reactions and the synthesis of the other nucleotides are discussed on p. 190.

Further information

The regulation of bacterial *aspartate carbamoyltransferase* by ATP and CTP has been particularly well studied, and is discussed on p. 116. In animals, in contrast to prokaryotes, it is not ACTase but *carbamoyl-phosphate synthase* that is the key enzyme in pyrimidine synthesis. It is activated by ATP and PRPP and inhibited by UTP.

The biosynthesis of the purines is also regulated by *feedback inhibition.* ADP and GDP inhibit the formation of PRRPP from ribose-5′-phosphate. Similarly, step **2a** is inhibited by AMP and GMP.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Components of nucleobases

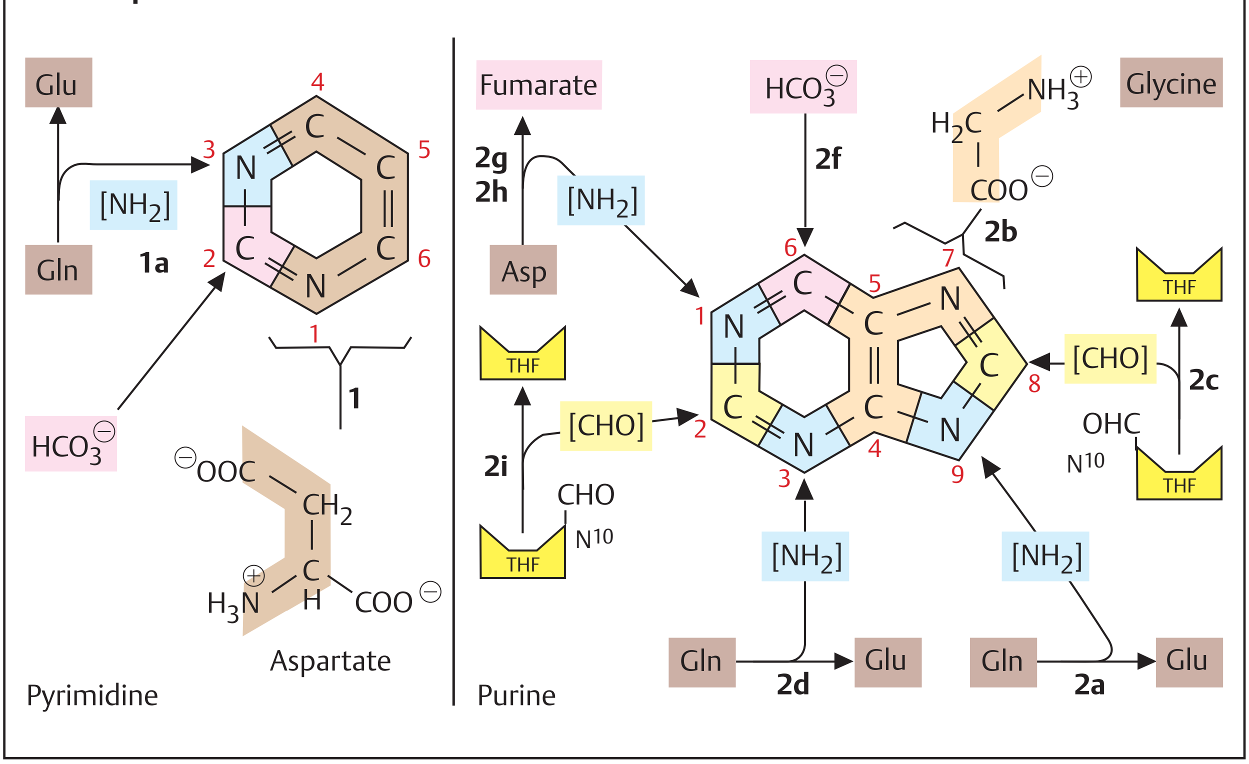
Glu
Gln
[NH$_2$]
1a
HCO$_3^⊖$
4
3
5
2
6
1
N
C
1
⊖OOC
CH$_2$
C
H$_3$N⊕
H
COO⊖
Aspartate
Pyrimidine
Fumarate
2g
2h
[NH$_2$]
Asp
THF
[CHO]
2i
CHO
N10
HCO$_3^⊖$
2f
H$_2$C
NH$_3^⊕$
COO⊖
2b
Glycine
7
6
5
1
8
2
3
4
9
THF
[CHO]
2c
OHC
N10
THF
[NH$_2$]
[NH$_2$]
Gln
Glu
2d
Gln
Glu
2a
Purine

B. Pyrimidine and purine synthesis

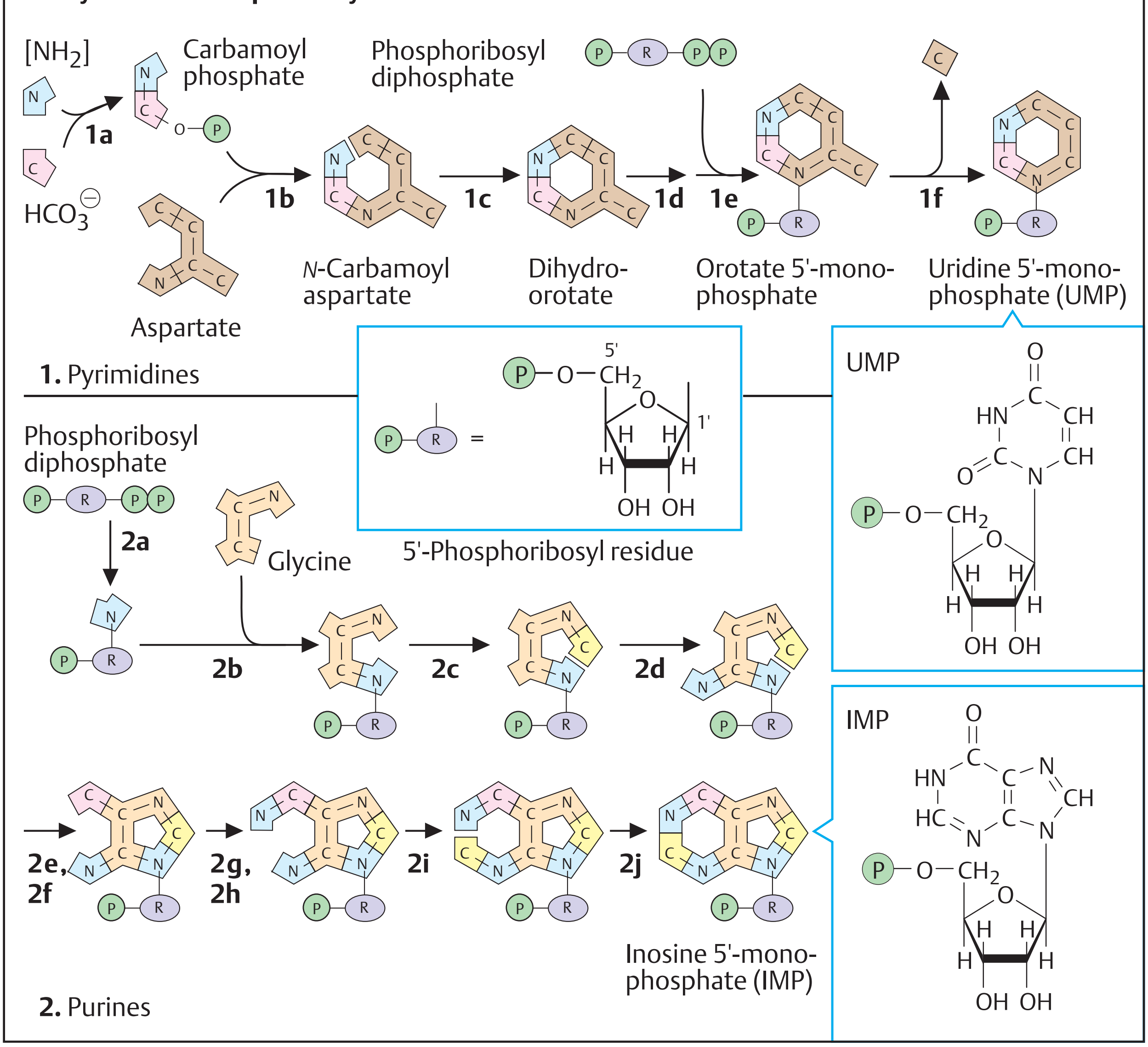
[NH$_2$]
HCO$_3^⊖$
1a
Carbamoyl phosphate
Aspartate
Phosphoribosyl diphosphate
1b
N-Carbamoyl aspartate
1c
Dihydro-orotate
1d
1e
Orotate 5'-mono-phosphate
1f
Uridine 5'-mono-phosphate (UMP)
1. Pyrimidines
5'
1'
OH OH
5'-Phosphoribosyl residue
UMP
HN
CH
O
N
O–CH$_2$
H
OH OH
Phosphoribosyl diphosphate
2a
Glycine
2b
2c
2d
2e, 2f
2g, 2h
2i
2j
Inosine 5'-mono-phosphate (IMP)
IMP
HN
HC
N
CH
O–CH$_2$
OH OH
2. Purines

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Nucleotide biosynthesis

De novo synthesis of purines and pyrimidines yields the monophosphates IMP and UMP, respectively (see p. 188). All other nucleotides and deoxynucleotides are synthesized from these two precursors. An overview of the pathways involved is presented here; further details are given on p. 417. Nucleotide synthesis by recycling of bases (the salvage pathway) is discussed on p. 186.

A. Nucleotide synthesis: overview ◑

The synthesis of **purine nucleotides** (**1**) starts from **IMP**. The base it contains, *hypoxanthine*, is converted in two steps each into adenine or guanine. The nucleoside monophosphates **AMP** and **GMP** that are formed are then phosphorylated by *nucleoside phosphate kinases* to yield the diphosphates **ADP** and **GDP**, and these are finally phosphorylated into the triphosphates **ATP and GTP**. The nucleoside triphosphates serve as components for RNA, or function as coenzymes (see p. 106). Conversion of the ribonucleotides into deoxyribonucleotides occurs at the level of the *diphosphates* and is catalyzed by *nucleoside diphosphate reductase* (**B**).

The biosynthetic pathways for the **pyrimidine nucleotides** (**2**) are more complicated. The first product, **UMP**, is phosphorylated first to the diphosphate and then to the triphosphate, **UTP**. *CTP synthase* then converts UTP into **CTP**. Since pyrimidine nucleotides are also reduced to deoxyribonucleotides at the diphosphate level, CTP first has to be hydrolyzed by a *phosphatase* to yield **CDP** before **dCDP and dCTP** can be produced.

The DNA component deoxythymidine triphosphate (**dTTP**) is synthesized from UDP in several steps. The base thymine, which only occurs in DNA (see p. 80), is formed by methylation of **dUMP** at the nucleoside monophosphate level. *Thymidylate synthase* and its helper enzyme *dihydrofolate reductase* are important target enzymes for cytostatic drugs (see p. 402).

B. Ribonucleotide reduction ○

2′-Deoxyribose, a component of DNA, is not synthesized as a free sugar, but arises at the diphosphate level by reduction of ribonucleoside diphosphates. This reduction is a complex process in which several proteins are involved. The reducing equivalents needed come from **NADPH+H$^+$**. However, they are not transferred directly from the coenzyme to the substrate, but first pass through a *redox series* that has several steps (**1**).

In the first step, *thioredoxin reductase* reduces a small redox protein, **thioredoxin**, via enzyme-bound FAD. This involves cleavage of a disulfide bond in thioredoxin. The resulting SH groups in turn reduce a catalytically active disulfide bond in *nucleoside diphosphate reductase* ("ribonucleotide reductase"). The free SH groups formed in this way are the actual electron donors for the reduction of ribonucleotide diphosphates.

In eukaryotes, ribonucleotide reductase is a tetramer consisting of two R1 and two R2 subunits. In addition to the **disulfide bond** mentioned, a **tyrosine radical** in the enzyme also participates in the reaction (**2**). It initially produces a substrate radical (**3**). This cleaves a water molecule and thereby becomes radical cation. Finally, the deoxyribose residue is produced by reduction, and the tyrosine radical is regenerated.

The regulation of ribonucleotide reductase is complex. The substrate-specificity and activity of the enzyme are controlled by two allosteric binding sites (a and b) in the R1 subunits. ATP and dATP increase or reduce the activity of the reductase by binding at site a. Other nucleotides interact with site b, and thereby alter the enzyme's specificity.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Nucleotide synthesis: overview

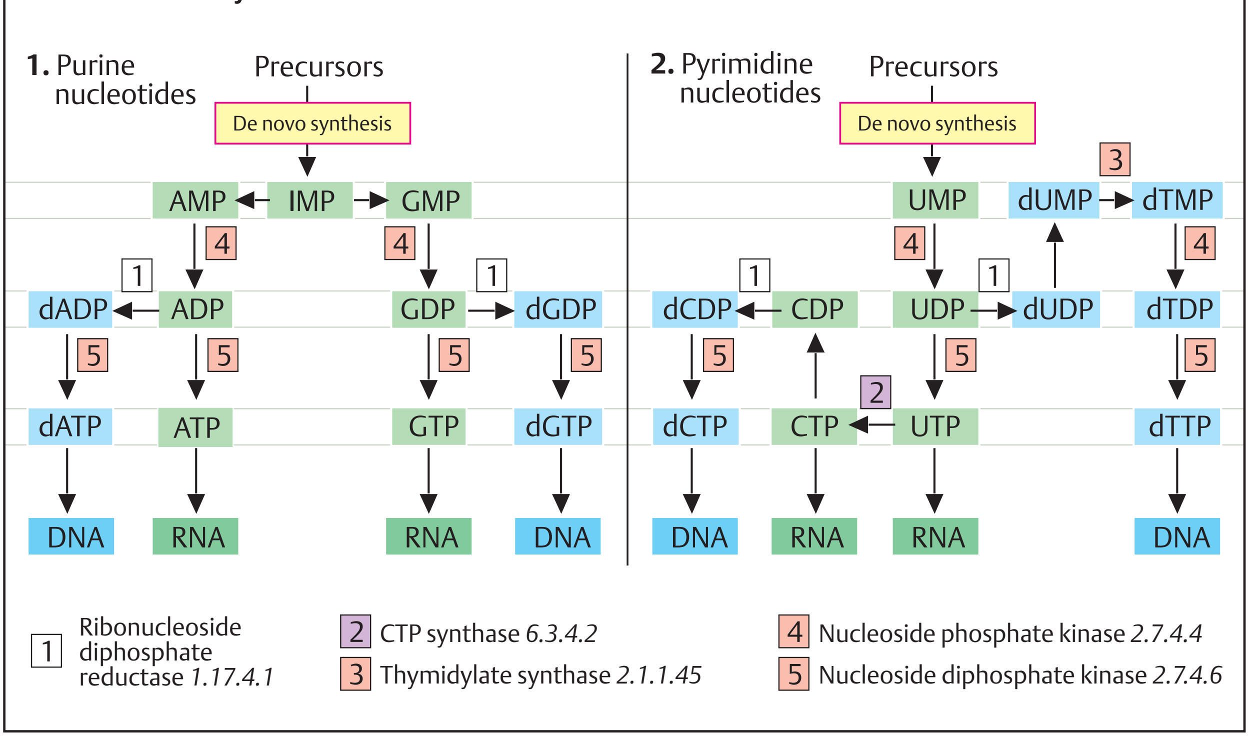
1. Purine nucleotides
Precursors
De novo synthesis
AMP
IMP
GMP
ADP
dADP
GDP
dGDP
ATP
dATP
GTP
dGTP
DNA
RNA
RNA
DNA
2. Pyrimidine nucleotides
Precursors
De novo synthesis
UMP
dUMP
dTMP
CDP
dCDP
UDP
dUDP
dTDP
CTP
dCTP
UTP
dTTP
DNA
RNA
RNA
DNA
1 Ribonucleoside diphosphate reductase 1.17.4.1
2 CTP synthase 6.3.4.2
3 Thymidylate synthase 2.1.1.45
4 Nucleoside phosphate kinase 2.7.4.4
5 Nucleoside diphosphate kinase 2.7.4.6

B. Ribonucleotide reduction

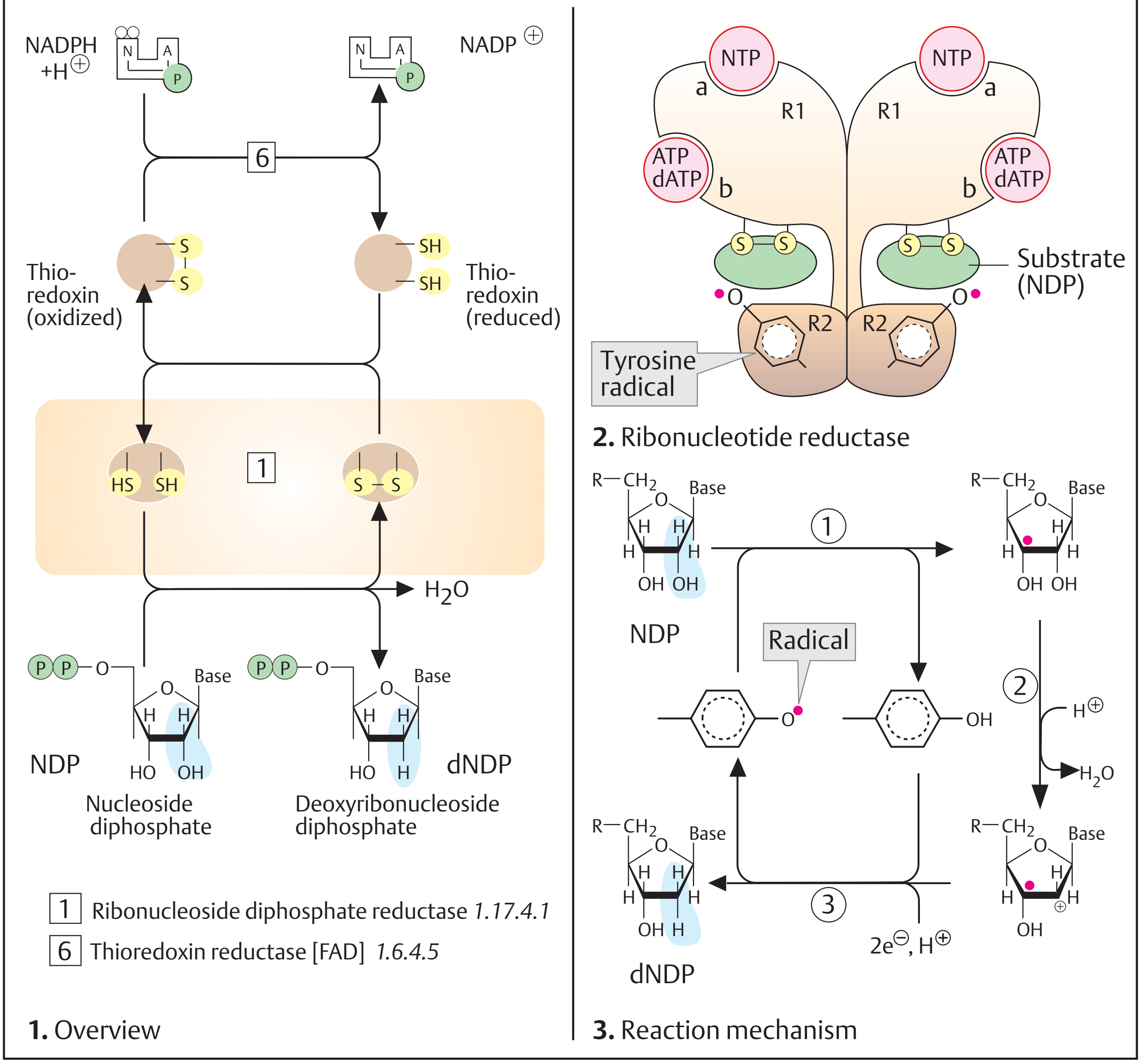
NADPH +H⊕
NADP⊕
N A P
6
Thio-redoxin (oxidized)
Thio-redoxin (reduced)
S S
SH SH
1
HS SH
S – S
H₂O
NDP
dNDP
Base
HO OH
HO H
Nucleoside diphosphate
Deoxyribonucleoside diphosphate
1 Ribonucleoside diphosphate reductase 1.17.4.1
6 Thioredoxin reductase [FAD] 1.6.4.5
1. Overview
NTP
a
R1
ATP dATP
b
S S
Substrate (NDP)
O
R2
Tyrosine radical
2. Ribonucleotide reductase
R–CH₂
Base
H OH OH
NDP
①
Radical
O
OH
②
H⊕
H₂O
③
2e⊖, H⊕
dNDP
OH H
3. Reaction mechanism

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Heme biosynthesis

Heme, an iron-containing tetrapyrrole pigment, is a component of O_2-binding proteins (see p. 106) and a coenzyme of various oxidoreductases (see p. 32). Around 85% of heme biosynthesis occurs in the bone marrow, and a much smaller percentage is formed in the liver. Both mitochondria and cytoplasm are involved in heme synthesis.

A. Biosynthesis of heme ○

Synthesis of the tetrapyrrole ring starts in the mitochondria.

[1] **Succinyl CoA** (upper left), an intermediate in the tricarboxylic acid cycle, undergoes condensation with **glycine** and subsequent decarboxylation to yield **5-aminolevulinate** (ALA). The *ALA synthase* responsible for this step is the key enzyme of the whole pathway. Synthesis of ALA synthase is *repressed* and existing enzyme is inhibited by heme, the end product of the pathway. This is a typical example of end-product or *feedback inhibition.*

[2] 5-Aminolevulinate now leaves the mitochondria. In the cytoplasm, two molecules condense to form **porphobilinogen,** a compound that already contains the pyrrole ring. *Porphobilinogen synthase* is inhibited by lead ions. This is why acute lead poisoning is associated with increased concentrations of ALA in the blood and urine.

[3] The tetrapyrrole structure characteristic of the porphyrins is produced in the next steps of the synthetic pathway. *Hydroxymethylbilane synthase* catalyzes the linkage of four porphobilinogen molecules and cleavage of an NH_2 group to yield **uroporphyrinogen III**.

[4] Formation of this intermediate step requires a second enzyme, *uroporphyrinogen III synthase*. If this enzyme is lacking, the "wrong" isomer, uroporphyrinogen I, is formed.

The tetrapyrrole structure of uroporphyrinogen III is still very different from that of heme. For example, the central iron atom is missing, and the ring contains only eight of the 11 double bonds. In addition, the ring system only carries charged R side chains (four acetate and four propionate residues). As heme groups have to act in the apolar interior of proteins, most of the polar side chains have to be converted into less polar groups.

[5] Initially, the four acetate residues (R_1) are decarboxylated into methyl groups. The resulting **coproporphyrinogen III** returns to the mitochondria again. The subsequent steps are catalyzed by enzymes located either on or inside the *inner mitochondrial membrane.*

[6] An *oxidase* first converts two of the propionate groups (R_2) into vinyl residues. The formation of **protoporphyrinogen IX** completes the modification of the side chains.

[7] In the next step, another oxidation produces the conjugated π-electron system of **protoporphyrin IX**.

[8] Finally, a divalent iron is incorporated into the ring. This step also requires a specific enzyme, *ferrochelatase.* The **heme b** or **Fe-protoporphyrin IX** formed in this way is found in hemoglobin and myoglobin, for example (see p. 280), where it is noncovalently bound, and also in various oxidoreductases (see p. 106).

Further information

There are a large number of hereditary or acquired disturbances of porphyrin synthesis, known as **porphyrias**, some of which can cause severe clinical pictures. Several of these diseases lead to the excretion of heme precursors in feces or urine, giving them a dark red color. Accumulation of porphyrins in the skin can also occur, and exposure to light then causes disfiguring, poorly healing blisters. Neurological disturbances are also common in the porphyrias.

It is possible that the medieval legends about human vampires ("Dracula") originated in the behavior of porphyria sufferers (avoidance of light, behavioral disturbances, and drinking of blood in order to obtain heme—which markedly improves some forms of porphyria).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Heme biosynthesis

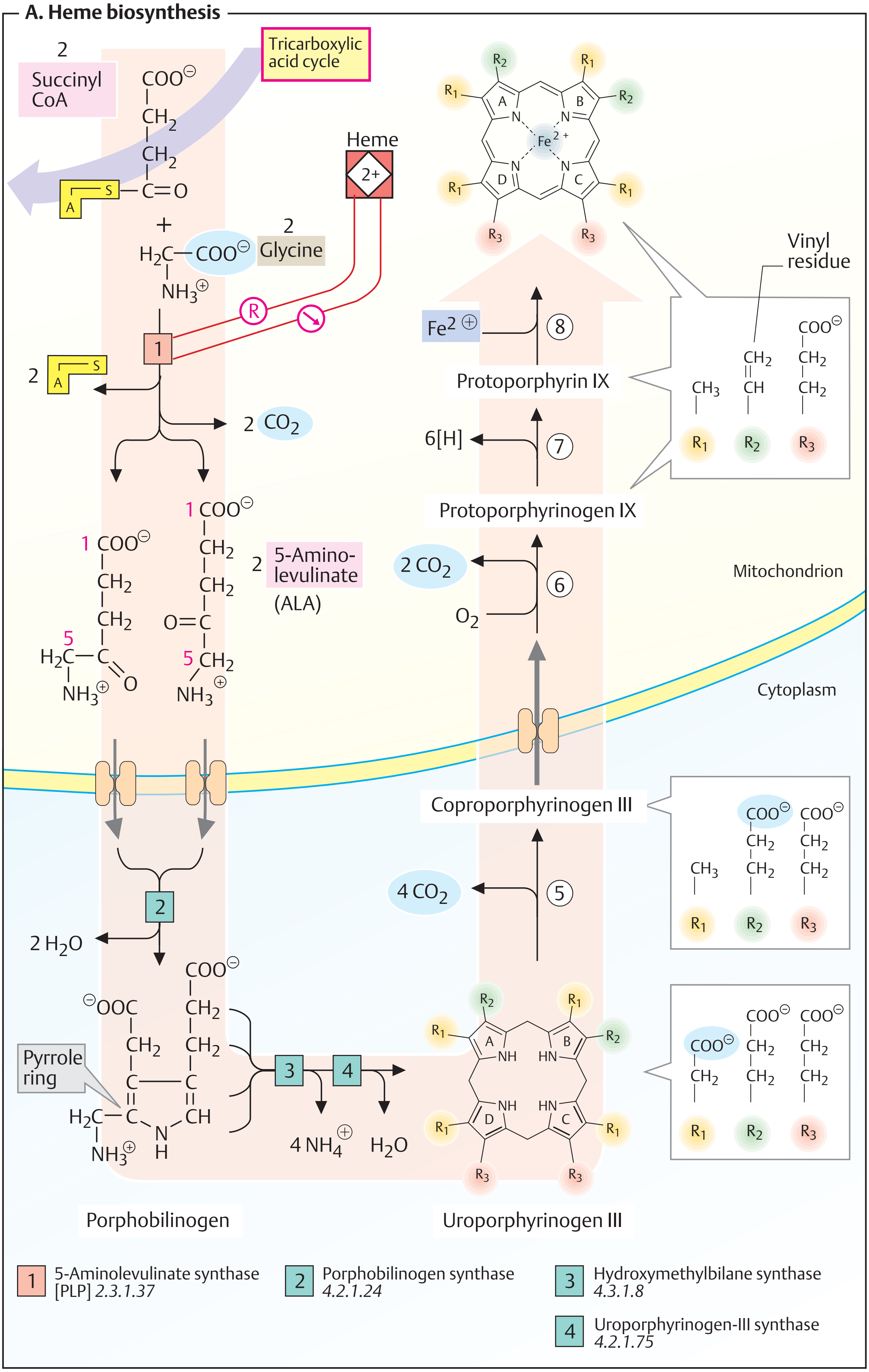
2
Succinyl CoA
Tricarboxylic acid cycle
Heme
2+
+
2
Glycine
R
1
2
2 CO2
1
5
2
5-Aminolevulinate (ALA)
2
2 H2O
Pyrrole ring
Porphobilinogen
3
4
4 NH4+
H2O
Uroporphyrinogen III
5
4 CO2
Coproporphyrinogen III
Cytoplasm
Mitochondrion
6
2 CO2
O2
Protoporphyrinogen IX
7
6[H]
Protoporphyrin IX
8
Fe2+
Vinyl residue
R1
R2
R3
A
B
C
D
Fe2+
1 5-Aminolevulinate synthase [PLP] 2.3.1.37
2 Porphobilinogen synthase 4.2.1.24
3 Hydroxymethylbilane synthase 4.3.1.8
4 Uroporphyrinogen-III synthase 4.2.1.75

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Heme degradation

A. Degradation of heme groups ○

Heme is mainly found in the human organism as a prosthetic group in erythrocyte hemoglobin. Around 100–200 million aged erythrocytes per hour are broken down in the human organism. The degradation process starts in reticuloendothelial cells in the spleen, liver, and bone marrow.

[1] After the protein part (globin) has been removed, the tetrapyrrole ring of heme is oxidatively cleaved between rings A and B by *heme oxygenase.* This reaction requires molecular oxygen and NADPH+H^+, and produces green **biliverdin**, as well as CO (carbon monoxide) and Fe^{2+}, which remains available for further use (see p. 286).

[2] In another redox reaction, biliverdin is reduced by *biliverdin reductase* to the orange-colored **bilirubin.** The color change from purple to green to yellow can be easily observed in vivo in a bruise or hematoma.

The color of heme and the other *porphyrin systems* (see p. 106) results from their numerous conjugated double bonds. Heme contains a cyclic conjugation (highlighted in pink) that is removed by reaction [1]. Reaction [2] breaks the π system down into two smaller separate systems (highlighted in yellow).

For further degradation, bilirubin is transported to the liver via the blood. As bilirubin is poorly soluble, it is bound to **albumin** for transport**.** Some drugs that also bind to albumin can lead to an increase in free bilirubin.

[3] The hepatocytes take up bilirubin from the blood and conjugate it in the endoplasmic reticulum with the help of **UDP-glucuronic acid** into the more easily soluble **bilirubin monoglucuronides** and **diglucuronides**. To do this, *UDP-glucuronosyltransferase* forms ester-type bonds between the OH group at C-1 of glucuronic acid and the carboxyl groups in bilirubin (see p. 316). The glucuronides are then excreted by active transport into the **bile**, where they form what are known as the **bile pigments**.

Glucuronide synthesis is the rate-determining step in hepatic bilirubin metabolism. Drugs such as *phenobarbital,* for example, can induce both conjugate formation and the transport process.

Some of the bilirubin conjugates are broken down further in the intestine by bacterial *γ-glucuronidases.* The bilirubin released is then reduced further via intermediate steps into colorless **stercobilinogen**, some of which is oxidized again into orange to yellow-colored stercobilin. The end products of bile pigment metabolism in the intestine are mostly excreted in feces, but a small proportion is resorbed *(enterohepatic circulation;* see p. 314). When high levels of heme degradation are taking place, stercobilinogen appears as **urobilinogen** in the urine, where oxidative processes darken it to form **urobilin**.

In addition to hemoglobin, other *heme proteins* (myoglobin, cytochromes, catalases, and peroxidases; see p. 32) also supply heme groups that are degraded via the same pathway. However, these contribute only about 10–15% to a total of ca. 250 mg of bile pigment formed per day.

Further information

Hyperbilirubinemias. An elevated bilirubin level (> 10 mg L^{-1}) is known as *hyperbilirubinemia.* When this is present, bilirubin diffuses from the blood into peripheral tissue and gives it a yellow color (jaundice). The easiest way of observing this is in the white conjunctiva of the eyes.

Jaundice can have various causes. If increased erythrocyte degradation (hemolysis) produces more bilirubin, it causes *hemolytic jaundice.* If bilirubin conjugation in the liver is impaired—e. g., due to hepatitis or liver cirrhosis—it leads to *hepatocellular jaundice,* which is associated with an increase in unconjugated (*"indirect"*) bilirubin in the blood. By contrast, if there is a disturbance of bile drainage (*obstructive jaundice,* due to gallstones or pancreatic tumors), then conjugated (*"direct"*) bilirubin in the blood increases. *Neonatal jaundice* (physiologic jaundice) usually resolves after a few days by itself. In severe cases, however, unconjugated bilirubin can cross the blood–brain barrier and lead to brain damage (*kernicterus*).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

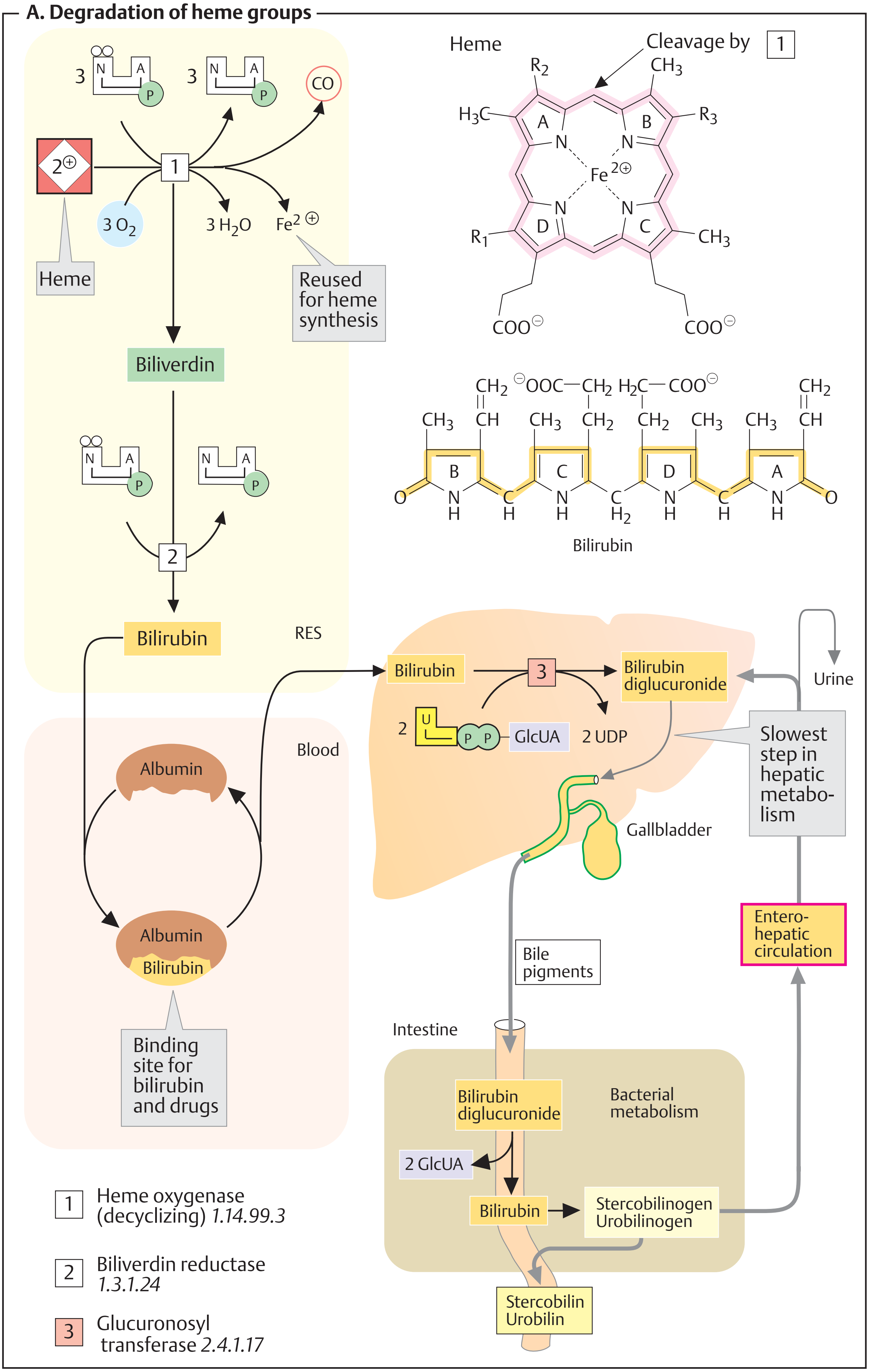

A. Degradation of heme groups

3
N A P
3
N A P
CO
2⊕
1
Heme
3 O_2
3 H_2O
$Fe^{2⊕}$
Reused for heme synthesis
Biliverdin
N A P
N A P
2
Bilirubin
RES
Heme
Cleavage by 1
R_2
CH_3
H_3C
R_3
A
B
N
N
$Fe^{2⊕}$
N
N
D
C
R_1
CH_3
$COO^⊖$
$COO^⊖$
CH_2
$^⊖OOC-CH_2$
$H_2C-COO^⊖$
CH_2
CH_3
CH
CH_3
CH_2
CH_2
CH_3
CH_3
CH
B C D A
O N H C H N H C H_2 N H C H N H O
Bilirubin
Blood
Albumin
Albumin
Bilirubin
Binding site for bilirubin and drugs
Bilirubin
3
Bilirubin diglucuronide
2 U P P GlcUA
2 UDP
Urine
Slowest step in hepatic metabolism
Gallbladder
Entero-hepatic circulation
Bile pigments
Intestine
Bilirubin diglucuronide
Bacterial metabolism
2 GlcUA
Bilirubin
Stercobilinogen
Urobilinogen
Stercobilin
Urobilin
1 Heme oxygenase (decyclizing) 1.14.99.3
2 Biliverdin reductase 1.3.1.24
3 Glucuronosyl transferase 2.4.1.17

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Structure of cells

A. Comparison of prokaryotes and eukaryotes ●

Present-day living organisms can be divided into two large groups—the prokaryotes and eukaryotes. The **prokaryotes** are represented by bacteria *(eubacteria* and *archaebacteria).* These are almost all small unicellular organisms only a few microns (10^{-6} m) in size. The **eukaryotes** include *fungi*, *plants*, and *animals* and comprise both unicellular and multicellular organisms. Multicellular eukaryotes are made up of a wide variety of cell types that are specialized for different tasks. Eukaryotic cells are much larger than prokaryotic ones (volume ratio approximately 2000 : 1). The most important distinguishing feature of these cells in comparison with the prokaryotes is the fact that they have a **nucleus** (*karyon* in Greek—hence the term).

In comparison with the prokaryotes, eukaryotic cells have greater specialization and complexity in their structure and functioning. Eukaryotic cells are structured into *compartments* (see below). The metabolism and synthesis of macromolecules are distributed through these reaction spaces and are separately regulated. In prokaryotes, these functions are organized in a simpler fashion and are spatially closely related.

Although the storage and transfer of genetic information function according to the same principle in the prokaryotes and eukaryotes, there are also differences. Eukaryotic DNA consists of very long, linear molecules with a total of 10^7 to more than 10^{10} base pairs (bp), only a small fraction of which are used for genetic information. In eukaryotes, the genes (20000–50000 per genome) are usually interrupted by non-coding regions (*introns*). Eukaryotic DNA is located in the nucleus, where together with histones and other proteins it forms the chromatin (see p. 238).

In prokaryotes, by contrast, DNA is ring-shaped, much shorter (up to 5 10^6 bp), and located in the cytoplasm. Almost all of it is used for information storage, and it does not contain any introns.

B. Structure of an animal cell ●

In the human body alone, there are at least 200 different cell types. The illustration outlines the basic structures of an animal cell in an extremely simplified way. The details given regarding the proportion of the compartments relative to cell volume (highlighted in yellow) and their numbers per cell frequency (blue) refer to mammalian hepatocytes (liver cells). The figures can vary widely from cell type to cell type.

The eukaryotic cell is subdivided by membranes. On the outside, it is enclosed by a **plasma membrane.** Inside the cell, there is a large space containing numerous components in solution—the **cytoplasm**. Additional membranes divide the internal space into *compartments* (confined reaction spaces). Welldefined compartments of this type are known as **organelles**.

The largest organelle is the **nucleus** (see p. 208). It is easily recognized using the light microscope. The **endoplasmic reticulum** (ER), a closed network of shallow sacs and tubules (see pp. 226ff.), is linked with the outer membrane of the nucleus. Another membrane-bound organelle is the **Golgi apparatus** (see p. 228), which resembles a bundle of layered slices. The **endosomes** and **exosomes** are bubble-shaped compartments *(* vesicles*)* that are involved in the exchange of substances between the cell and its surroundings. Probably the most important organelles in the cell's metabolism are the **mitochondria,** which are around the same size as bacteria (see pp. 210ff.). The **lysosomes** and **peroxisomes** are small, globular organelles that carry out specific tasks. The whole cell is traversed by a framework of proteins known as the **cytoskeleton** (see pp. 204ff.).

In addition to these organelles, plant cells (see p. 43) also have plastids—eg., **chloroplasts**, in which photosynthesis takes place (see p. 128). In their interior, there is a large, fluid-filled **vacuole**. Like bacteria and fungi, plant cells have a rigid cell wall consisting of polysaccharides and proteins.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Comparison of prokaryotes and eukaryotes

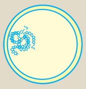

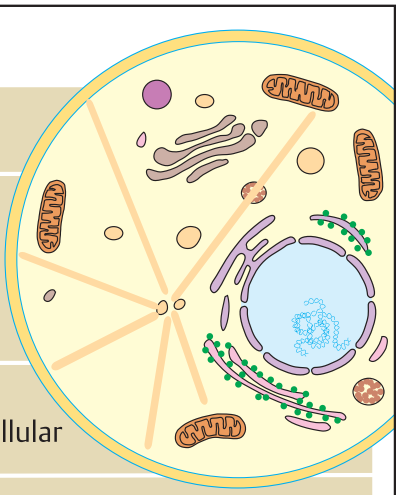

	Prokaryotes	Eukaryotes
Organisms	Eubacteria Archaebacteria	Fungi Plants Animals
Form	Single-celled	Single or multi-cellular
Organelles, cytoskeleton, cell division apparatus	Missing	Present, complicated, specialized
DNA	Small, circular, no introns, plasmids	Large, in nucleus, many introns
RNA: Synthesis and maturation	Simple, in cytoplasm	Complicated, in nucleus
Protein: Synthesis and maturation	Simple, coupled with RNA synthesis	Complicated, in the cytoplasm and the rough endoplasmic reticulum
Metabolism	Anaerobic or aerobic very flexible	Mostly aerobic, compartmented
Endocytosis and Exocytosis	no	yes

B. Structure of an animal cell

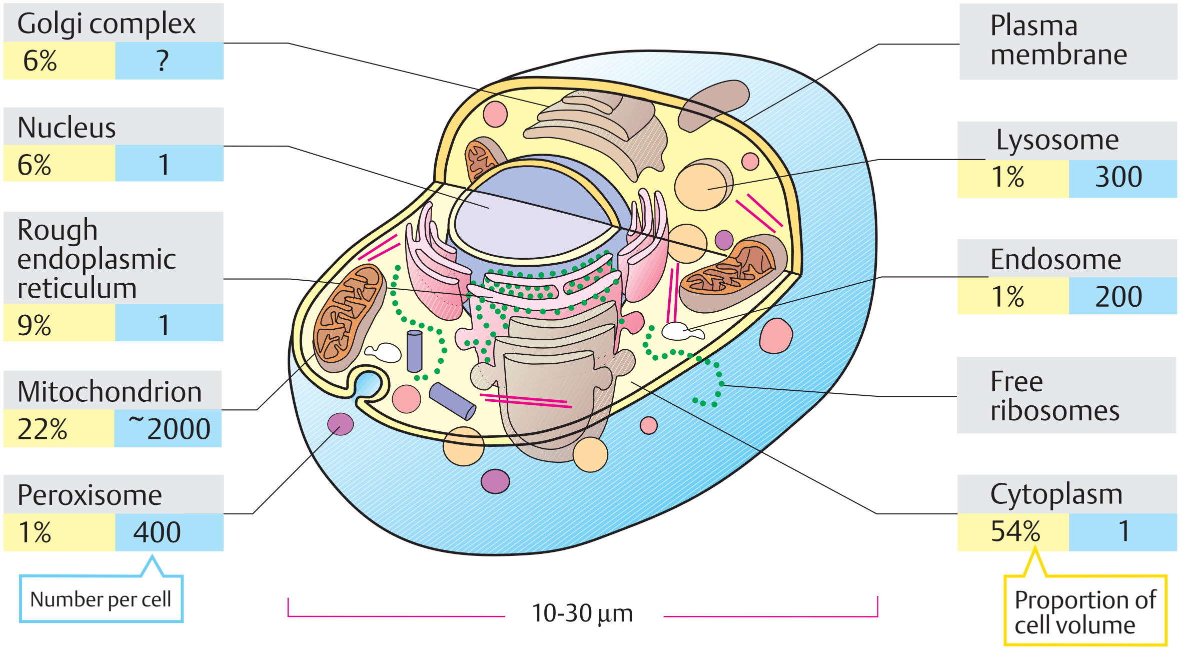
Golgi complex
6%
?
Nucleus
6%
1
Rough endoplasmic reticulum
9%
1
Mitochondrion
22%
~2000
Peroxisome
1%
400
Number per cell
10-30 µm
Plasma membrane
Lysosome
1%
300
Endosome
1%
200
Free ribosomes
Cytoplasm
54%
1
Proportion of cell volume

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Cell fractionation

A. Isolation of cell organelles ○

To investigate the individual compartments of the cell (see p. 196), various procedures have been developed to enrich and isolate cell organelles. These are mainly based on the size and density of the various organelles.

The isolation of cell components starts with **disruption** of the tissue being examined and subsequent **homogenization** of it (breaking down the cells) in a suitable buffer (see below). Homogenization using the "Potter" (the Potter–Elvehjem homogenizer, a rotating Teflon pestle in a glass cylinder) is particularly suitable for animal tissue. This method is very gentle and is therefore used to isolate fragile structures and molecules. Other cell disruption procedures include **enzymatic lysis** with the help of enzymes that break down the cell wall, **mechanical disruption** by grinding frozen tissue, cutting or smashing with rotating knives, large pressure changes, osmotic shock, and repeated freezing and thawing.

To isolate intact organelles, it is important for the homogenization solution to be *isotonic*—i. e., the osmotic value of the buffer has to be the same as that of the interior of the cell. If hypotonic solutions were used, the organelles would take up water and burst, while in hypertonic solutions they would shrink.

Homogenization is followed by coarse **filtration** through gauze to remove intact cells and connective-tissue fragments. The actual fractionation of cellular components is then carried out by **centrifugation steps**, in which the gravitational force (given as multiples of the earth's gravity, $g = 9.81\ \mathrm{m\ s^{-2}}$) is gradually increased (*differential centrifugation*; see p. 200). Due to the different shapes and densities of the organelles, this leads to successive sedimentation of each type out of the suspension.

Nuclei already sediment at low accelerations that can be achieved with bench-top centrifuges. Decanting the residue (the "supernatant") and carefully suspending the sediment (or "pellet") in an isotonic medium yields a fraction that is enriched with nuclei. However, this fraction may still contain other cellular components as contaminants—e. g., fragments of the cytoskeleton.

Particles that are smaller and less dense than the nuclei can be obtained by step-by-step acceleration of the gravity on the supernatant left over from the first centrifugation. However, this requires very powerful centrifuges (high-speed centrifuges and ultracentrifuges). The sequence in which the fractions are obtained is: **mitochondria, membrane vesicles**, and **ribosomes**. Finally, the supernatant from the last centrifugation contains the **cytosol** with the cell's soluble components, in addition to the buffer.

The isolation steps are carried out at low temperatures on principle (usually 0–5 °C), to slow down degradation reactions—e. g., due to released enzymes and other influencing factors. The addition of thiols and chelating agents protects functional SH groups from oxidation. Isolated cell organelles quickly lose their biological activity despite these precautions. Nevertheless, it is possible by working carefully to isolate mitochondria that will still take up substrates for a few hours in the test tube and produce ATP via oxidative phosphorylation.

B. Marker molecules ○

During cell fractionation, it is very important to analyze the purity of the fractions obtained. Whether or not the intended organelle is present in a particular fraction, and whether or not the fraction contains other components, can be determined by analyzing characteristic **marker molecules.** These are molecules that occur exclusively or predominantly in one type of organelle. For example, the activity of organelle-specific enzymes **(marker enzymes)** is often assessed. The distribution of marker enzymes in the cell reflects the compartmentation of the processes they catalyze. These reactions are discussed in greater detail here under the specific organelles.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Isolation of cell organelles

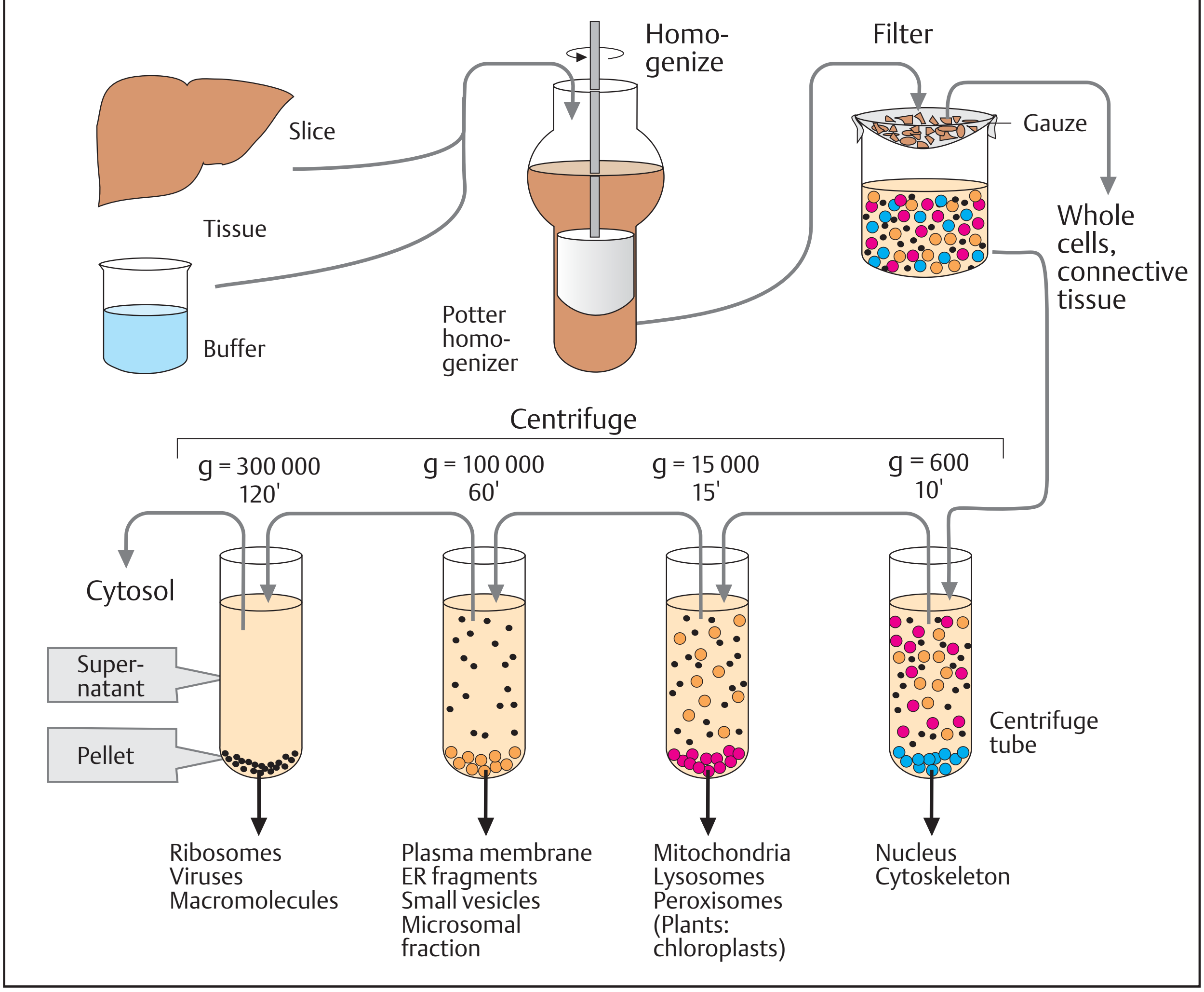

Homo-genize
Filter
Slice
Gauze
Tissue
Whole cells, connective tissue
Buffer
Potter homo-genizer
Centrifuge
g = 300 000
120'
g = 100 000
60'
g = 15 000
15'
g = 600
10'
Cytosol
Super-natant
Pellet
Centrifuge tube
Ribosomes
Viruses
Macromolecules
Plasma membrane
ER fragments
Small vesicles
Microsomal fraction
Mitochondria
Lysosomes
Peroxisomes
(Plants: chloroplasts)
Nucleus
Cytoskeleton

B. Marker molecules

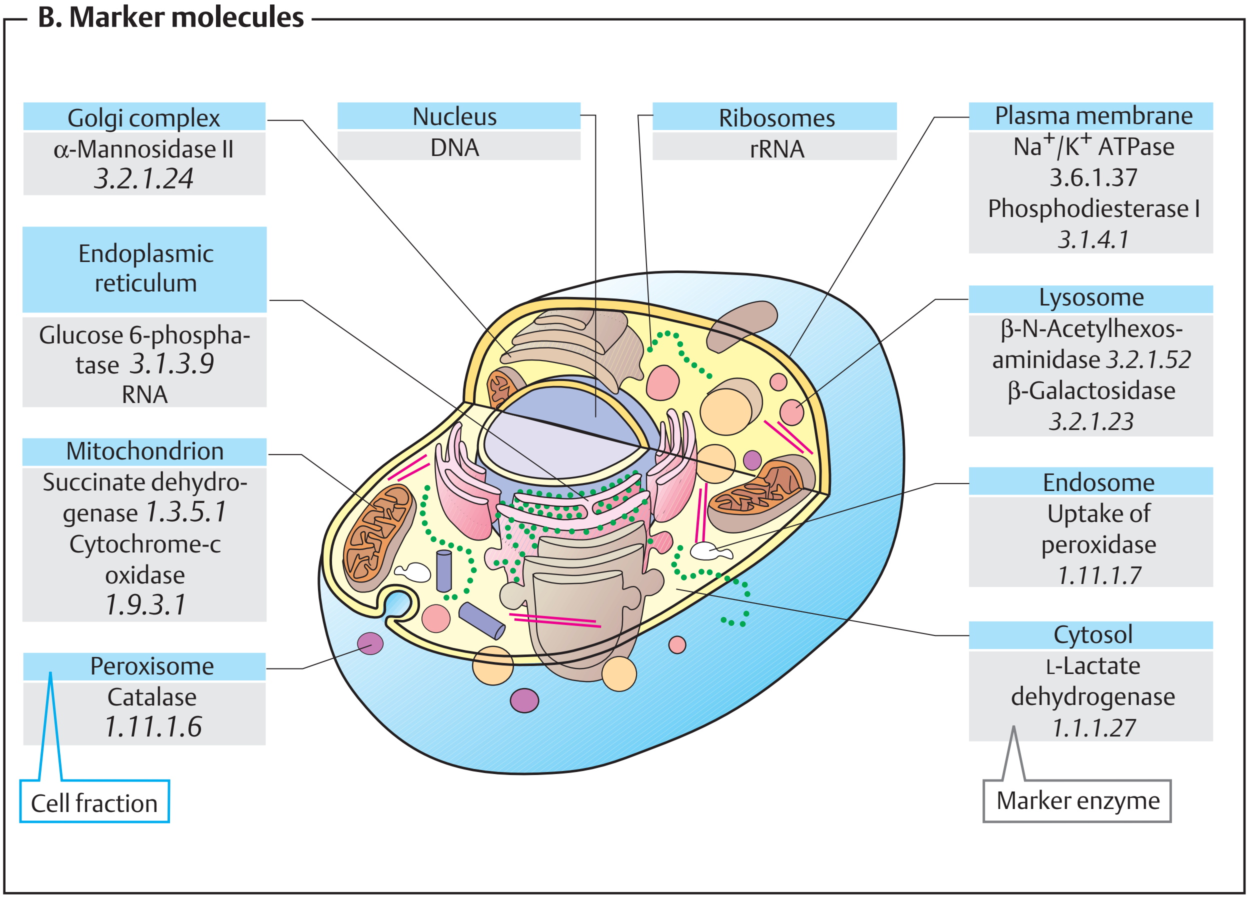

Golgi complex
α-Mannosidase II
3.2.1.24
Nucleus
DNA
Ribosomes
rRNA
Plasma membrane
Na$^+$/K$^+$ ATPase
3.6.1.37
Phosphodiesterase I
3.1.4.1
Endoplasmic reticulum
Glucose 6-phosphatase 3.1.3.9
RNA
Lysosome
β-N-Acetylhexosaminidase 3.2.1.52
β-Galactosidase
3.2.1.23
Mitochondrion
Succinate dehydrogenase 1.3.5.1
Cytochrome-c oxidase
1.9.3.1
Endosome
Uptake of peroxidase
1.11.1.7
Peroxisome
Catalase
1.11.1.6
Cytosol
L-Lactate dehydrogenase
1.1.1.27
Cell fraction
Marker enzyme

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Centrifugation

A. Principles of centrifugation ○

In a solution, particles whose density is higher than that of the solvent sink (*sediment*), and particles that are lighter than it *float* to the top. The greater the difference in density, the faster they move. If there is no difference in density (isopyknic conditions), the particles *hover*. To take advantage of even tiny differences in density to separate various particles in a solution, gravity can be replaced with the much more powerful "centrifugal force" provided by a centrifuge.

Equipment. The acceleration achieved by centrifugation is expressed as a multiple of the earth's gravitational force (g = 9.81 m s^{-2}). Bench-top centrifuges can reach acceleration values of up to 15 000 g, while highspeed refrigerated centrifuges can reach 50 000 g and ultracentrifuges, which operate with refrigeration and in a vacuum, can reach 500 000 g. Two types of rotor are available in high-powered centrifuges: *fixed angle rotors* and *swingout rotors* that have movable bucket containers. The tubes or buckets used for centrifugation are made of plastic and have to be very precisely adjusted to avoid any imbalances that could lead to accidents.

Theory. The velocity (v) of particle sedimentation during centrifugation depends on the angular velocity ω of the rotor, its effective radius (r_{eff}, the distance from the axis of rotation), and the particle's sedimentation properties. These properties are expressed as the **sedimentation coefficient** S (1 Svedberg, = 10^{-13} s). The sedimentation coefficient depends on the mass M of the particle, its shape (expressed as the coefficient of friction, f), and its density (expressed as the reciprocal density $\bar{v}$, "partial specific volume").

At the top right, the diagram shows the densities and sedimentation coefficients for biomolecules, cell organelles, and viruses. Proteins and protein-rich structures have densities of around 1.3 g cm^{-3}, while nucleic acids show densities of up to 2 g cm^{-3}. Equilibrium sedimentation of nucleic acids therefore requires high-density media—e. g., concentrated solutions of cesium chloride (CsCl). To allow comparison of S values measured in different media, they are usually corrected to values for water at 20 °C ("S_{20W}").

B. Density gradient centrifugation ○

Density gradient centrifugation is used to separate macromolecules that differ only slightly in size or density. Two techniques are commonly used.

In **zonal centrifugation,** the sample being separated (e. g., a cell extract or cells) is placed on top of the centrifugation solution as a thin layer. During centrifugation, the particles move through the solution due to their greater density. The rate of movement basically depends on their molecular mass (see **A**, formulae). Centrifugation stops before the particles reach the bottom of the tube. Drilling a hole into the centrifugation tube and allowing the contents to drip out makes it possible to collect the different particles in separate fractions. During centrifugation, the solution tube is stabilized in the tube by a **density gradient**. This consists of solutions of carbohydrates or colloidal silica gel, the concentration of which increases from the surface of the tube to the bottom. Density gradients prevent the formation of convection currents, which would impair the separation of the particles.

Isopyknic centrifugation, which takes much longer, starts with a CsCl solution in which the sample material (e. g., DNA, RNA, or viruses) is homogeneously distributed. A density gradient only forms *during* centrifugation, as a result of sedimentation and diffusion processes. Each particle moves to the region corresponding to its own *buoyant density*. Centrifugation stops once equilibrium has been reached. The samples are obtained by fractionation, and their concentration is measured using the appropriate methods.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Principles of centrifugation

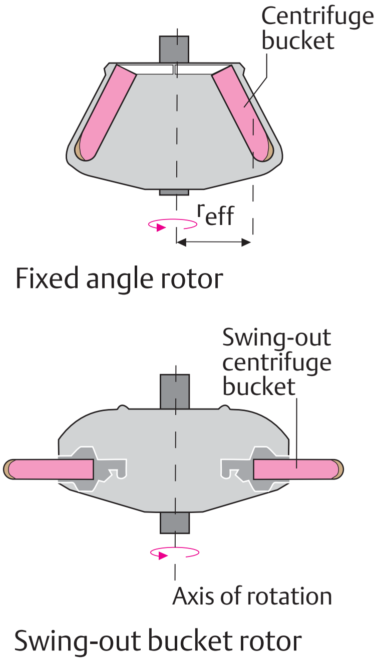
Centrifuge bucket
r_{eff}
Fixed angle rotor
Swing-out centrifuge bucket
Axis of rotation
Swing-out bucket rotor

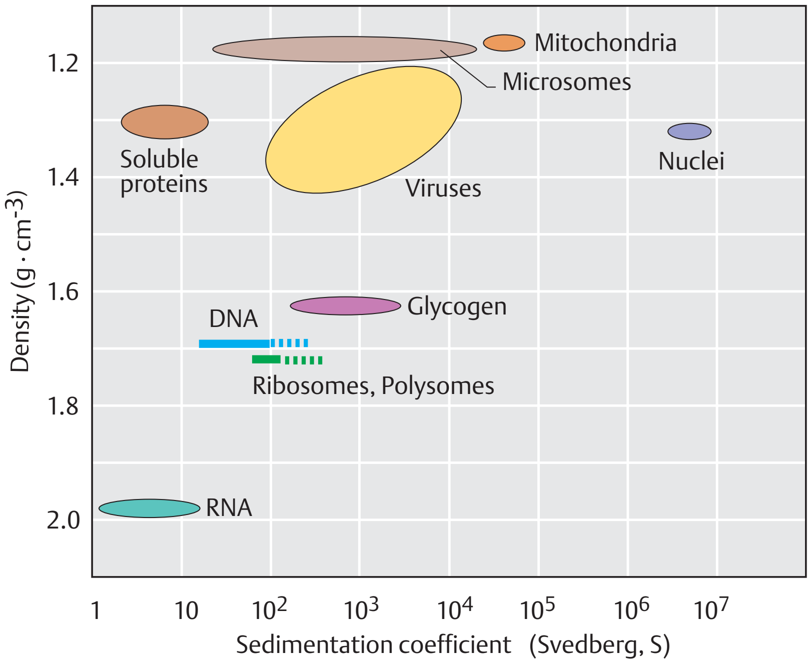

g: Gravitational acceleration

v: Sedimentation velocity ($cm \cdot s^{-1}$)

ω: Angular velocity ($rad \cdot s^{-1}$)

r_{eff}: Effective radius (cm)

$$g = \omega^2 \cdot r_{eff}$$

$$v = \omega^2 \cdot r_{eff} \cdot s$$

$$s = \frac{M \cdot (1 - \bar{v} \cdot r)}{f}$$

s: Sedimentation coefficient ($S = 10^{-13}$ s)

M: Molecular mass

$\bar{v}$: Partial specific particle volume ($cm^3 \cdot g^{-1}$)

r: Density of the solution ($g \cdot cm^3$)

f: Coefficient of friction

B. Density gradient centrifugation

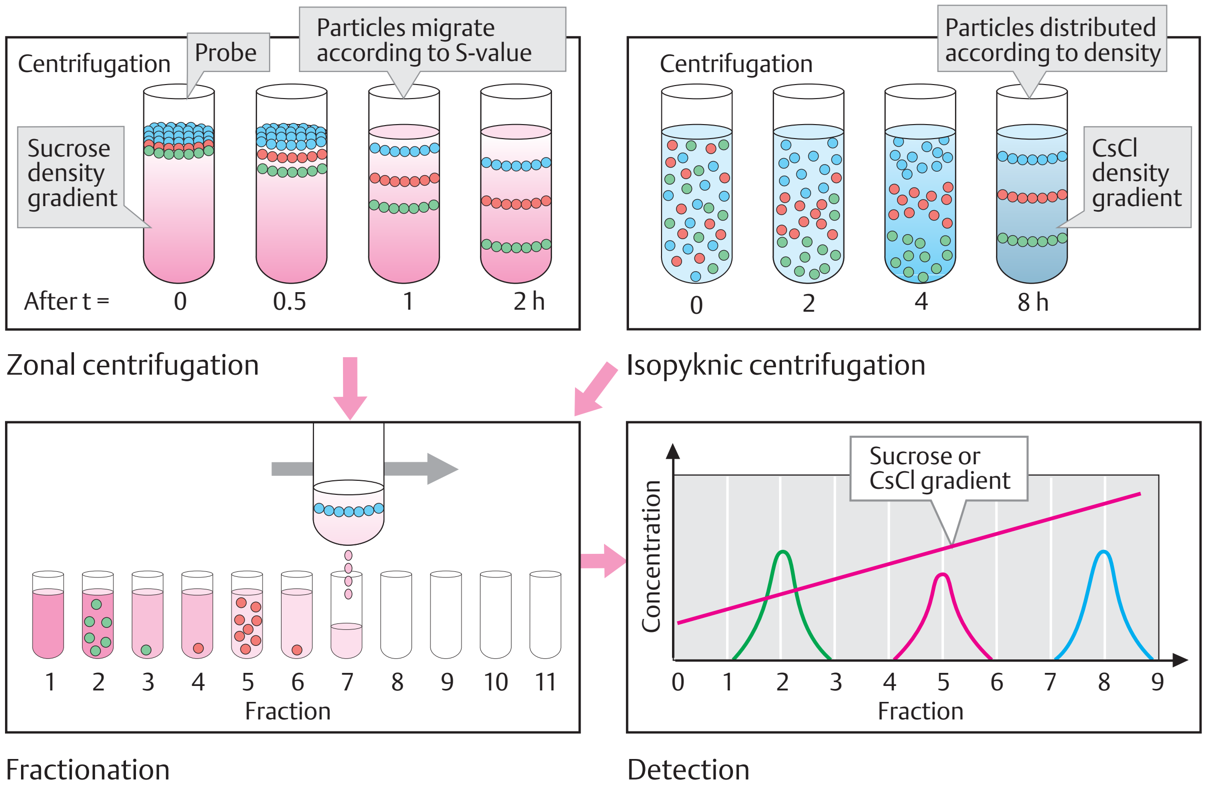
Centrifugation
Probe
Particles migrate according to S-value
Sucrose density gradient
After t =
0
0.5
1
2 h
Zonal centrifugation
Centrifugation
Particles distributed according to density
CsCl density gradient
0
2
4
8 h
Isopyknic centrifugation
1
2
3
4
5
6
7
8
9
10
11
Fraction
Fractionation
Sucrose or CsCl gradient
Concentration
0
1
2
3
4
5
6
7
8
9
Fraction
Detection

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Cell components and cytoplasm

The Gram-negative bacterium *Escherichia coli (E. coli)* is a usually harmless symbiont in the intestine of mammals. The structure and characteristics of this organism have been particularly well characterized. *E. coli* is also frequently used in genetic engineering (see p. 258).

A. Components of a bacterial cell ◑

A single *E. coli* cell has a **volume** of about 0.88 μm^3. One-sixth of this consists of membranes and one-sixth is DNA (known as the "nucleoid"). The rest of the internal space of the cell is known as **cytoplasm** (not "cytosol"; see p. 198).

The main component of *E. coli*—as in all cells—is **water** (70%). The other components are **macromolecules** (proteins, nucleic acids, polysaccharides), **small organic molecules**, and **inorganic ions**. The majority of the macromolecules are proteins, which represent ca. 55% of the dry mass of the cell. When a number of assumptions are made about the distribution and size (average mass 40 kDa) of proteins, it can be estimated that there are approximately 250 000 protein molecules in the cytoplasm of an *E. coli* cell. In eukaryotic cells, which are about a thousand times larger, it is estimated that the number of protein molecules is in the order of several billion.

B. Looking inside a bacterial cell ○

The illustration shows a schematic view inside the **cytoplasm** of *E. coli*, magnified approximately one million times. At this magnification, a single carbon atom would be the size of a grain of salt, and an ATP molecule would be as large as a grain of rice. The detail shown is 100 nm long, corresponding to about 1/600th of the volume of a cell in *E. coli*. To make the macromolecules clearer, small molecules such as water, cofactors, and metabolites have all been omitted from the illustration. The section of the cytoplasm shown contains:

- Several hundred **macromolecules**, which are needed for protein biosynthesis—i. e., 30 ribosomes, more than 100 protein factors, 30 aminoacyl–tRNA synthases, 340 tRNA molecules, 2–3 mRNAs (each of which is 10 times the length of the section shown), and six molecules of RNA polymerase.
- About 330 other enzyme molecules, including 130 glycolytic enzymes and 100 enzymes from the tricarboxylic acid cycle.
- 30 000 **small organic molecules** with masses of 100–1000 Da—e. g., metabolites of the intermediary metabolism and coenzymes. These are shown at a magnification 10 times higher in the bottom right corner.
- And finally, 50 000 **inorganic ions.** The rest consists of water.

The illustration shows that the cytoplasm of cells is a compartment densely packed with macromolecules and smaller organic molecules. The distances between organic molecules are small. They are only separated by a few water molecules.

All of the molecules are in motion. Due to constant collisions, however, they do not advance in a straight path but move in zigzags. Due to their large mass, proteins are particularly slow. However, they do cover an average of 5 nm in 1 ms—a distance approximately equal to their own length. Statistically, a protein is capable of reaching any point in a bacterial cell in less than a second.

C. Biochemical functions of the cytoplasm ◑

In eukaryotes, the cytoplasm, representing slightly more than 50% of the cell volume, is the most important cellular compartment. It is the *central reaction space of the cell.* This is where many important pathways of the intermediary metabolism take place—e. g., glycolysis, the pentose phosphate pathway, the majority of gluconeogenesis, and fatty acid synthesis. Protein biosynthesis (translation; see p. 250) also takes place in the cytoplasm. By contrast, fatty acid degradation, the tricarboxylic acid cycle, and oxidative phosphorylation are located in the mitochondria (see p. 210).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Components of a bacterial cell

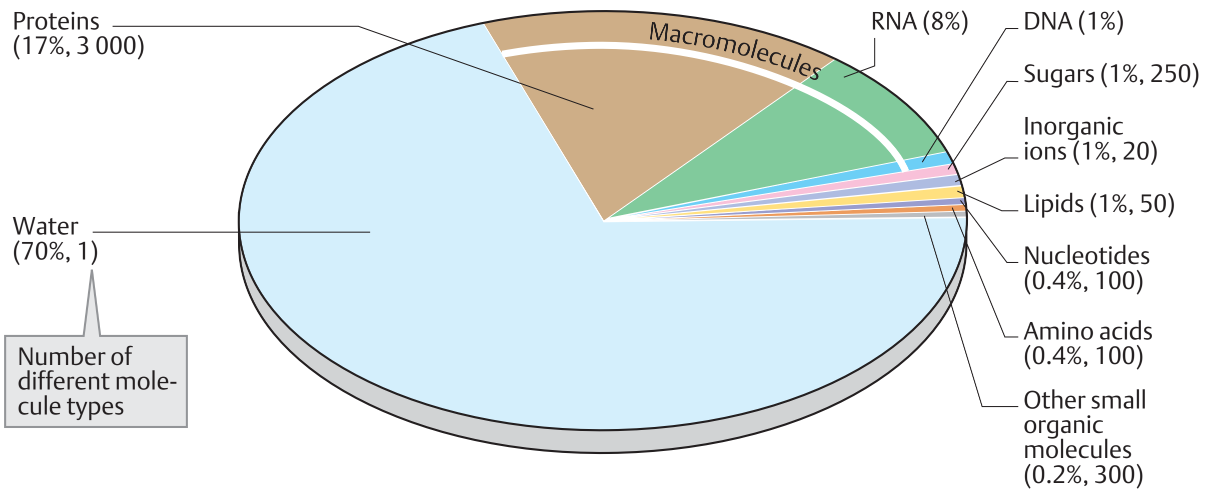

B. View into a bacterial cell

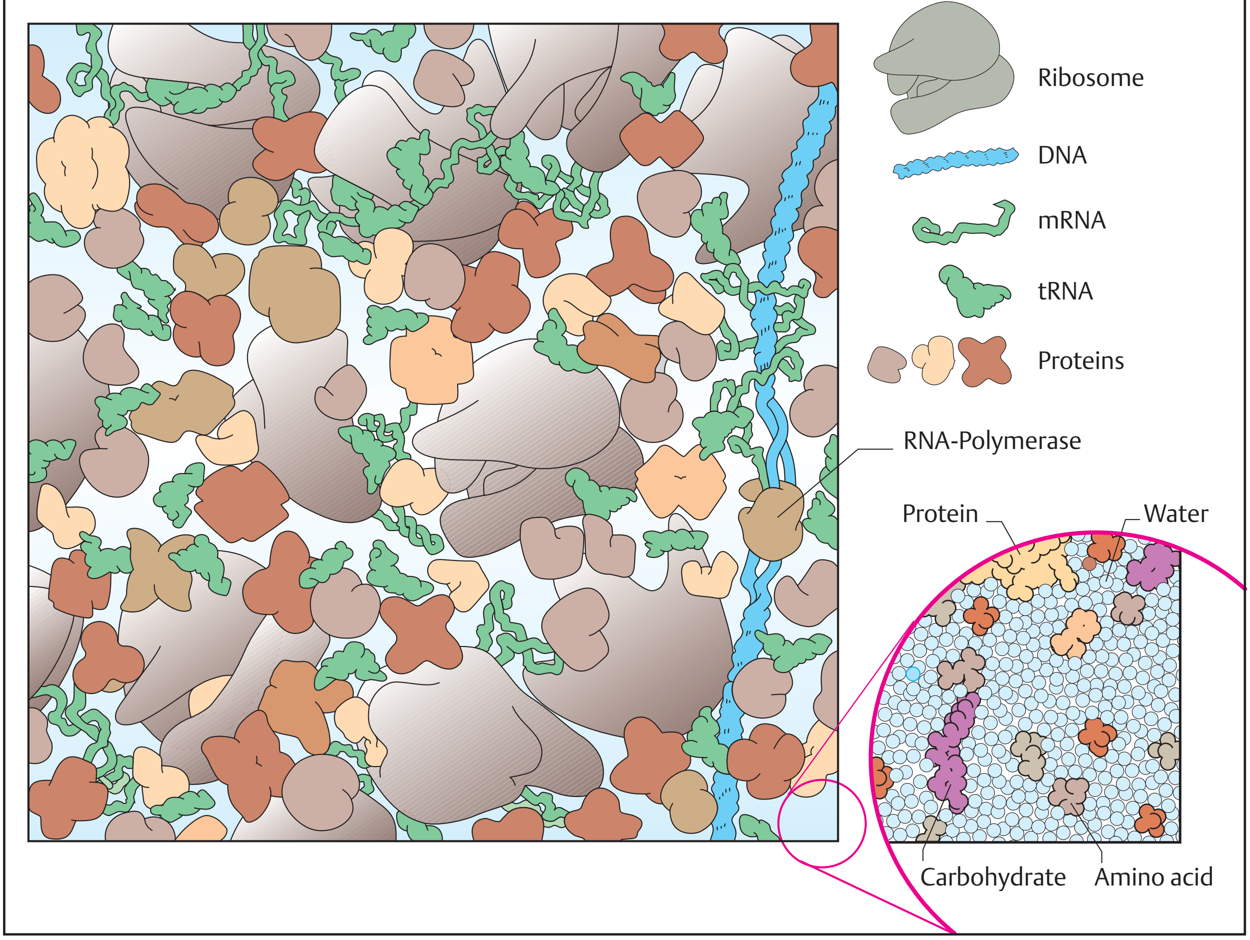
Ribosome
DNA
mRNA
tRNA
Proteins
RNA-Polymerase
Protein
Water
Carbohydrate
Amino acid

C. Biochemical functions of the cytoplasm

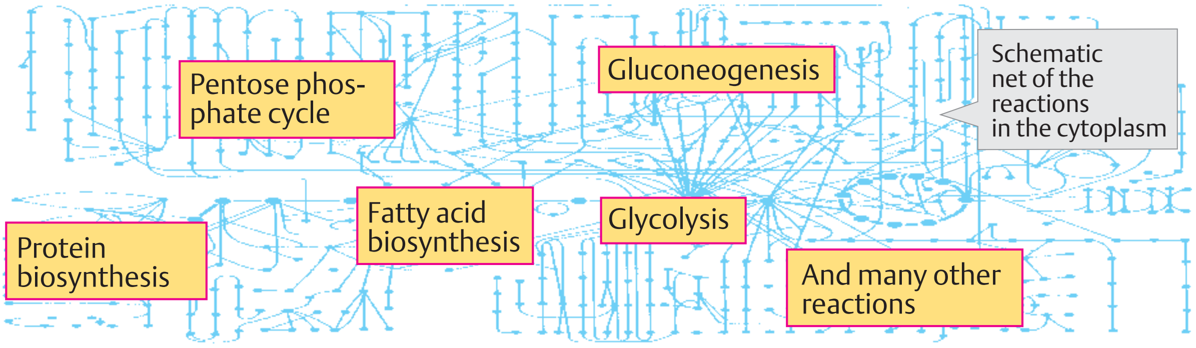
Pentose phos-
phate cycle
Gluconeogenesis
Schematic
net of the
reactions
in the cytoplasm
Protein
biosynthesis
Fatty acid
biosynthesis
Glycolysis
And many other
reactions

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Cytoskeleton: components

The cytoplasm of eukaryotic cells is traversed by three–dimensional scaffolding structures consisting of filaments (long protein fibers), which together form the **cytoskeleton**. These filaments are divided into three groups, based on their *diameters*: **microfilaments** (6–8 nm), **intermediate filaments** (ca. 10 nm), and **microtubules** (ca. 25 nm). All of these filaments are polymers assembled from protein components.

A. Actin ◑

Actin, the most abundant protein in eukaryotic cells, is the protein component of the **microfilaments** (actin filaments). Actin occurs in two forms—a monomolecular form (**G actin**, globular actin) and a polymer (**F actin**, filamentous actin). G actin is an asymmetrical molecule with a mass of 42 kDa, consisting of two domains. As the ionic strength increases, G actin aggregates reversibly to form F actin, a helical homopolymer. G actin carries a firmly bound ATP molecule that is slowly hydrolyzed in F actin to form ADP. Actin therefore also has enzyme properties (*ATPase* activity).

As individual G actin molecules are always oriented in the same direction relative to one another, F actin consequently has *polarity*. It has two different ends, at which polymerization takes place at different rates. If the ends are not stabilized by special proteins (as in muscle cells), then at a critical concentration of G actin the (+) end of F actin will constantly grow, while the (–) end simultaneously decays. These partial processes can be blocked by fungal toxins experimentally. **Phalloidin**, a toxin contained in the *Amanita phalloides* mushroom, inhibits decay by binding to the (–) end. By contrast, **cytochalasins**, mold toxins with cytostatic effects, block polymerization by binding to the (+) end.

Actin–associated proteins. The cytoplasm contains more than 50 different proteins that bind specifically to G actin and F actin. Their actin uptake has various different functions. This type of bonding can serve to regulate the G actin pool (example: *profilin*), influence the polymerization rate of G actin (*villin*), stabilize the chain ends of F actin (*fragin, ϒ–actinin*), attach filaments to one another or to other cell components (*villin, α–actinin, spectrin),* or disrupt the helical structure of F actin (*gelsolin*). The activity of these proteins is regulated by protein kinases via Ca^{2+} and other second messengers (see p. 386).

B. Intermediate filaments ◑

The components of the intermediate filaments belong to five related protein families. They are specific for particular cell types. Typical representatives include the *cytokeratins, desmin, vimentin, glial fibrillary acidic protein* (GFAP), and *neurofilament*. These proteins all have a rod–shaped basic structure in the center, which is known as a *superhelix* ("coiled coil"; see keratin, p. 70). The dimers are arranged in an antiparallel fashion to form tetramers. A staggered head-to–head arrangement produces **protofilaments**. Eight protofilaments ultimately form an intermediary filament.

Free protein monomers of intermediate filaments rarely occur in the cytoplasm, in contrast to microfilaments and microtubules. Their polymerization leads to stable polymers that have no polarity.

C. Tubulins ◑

The basic components of the tube-shaped **microtubules** are α– and β–**tubulin** (53 and 55 kDa). These form α,β-heterodimers, which in turn polymerize to form linear protofilaments. Thirteen protofilaments form a ring-shaped complex, which then grows into a long tube as a result of further polymerization.

Like microfilaments, microtubules are dynamic structures with (+) and (–) ends. The (–) end is usually stabilized by bonding to the centrosome. The (+) end shows *dynamic instability.* It can either grow slowly or shorten rapidly. GTP, which is bound by the microtubules and gradually hydrolyzed into GDP, plays a role in this. Various proteins can also be associated with microtubules.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Actin

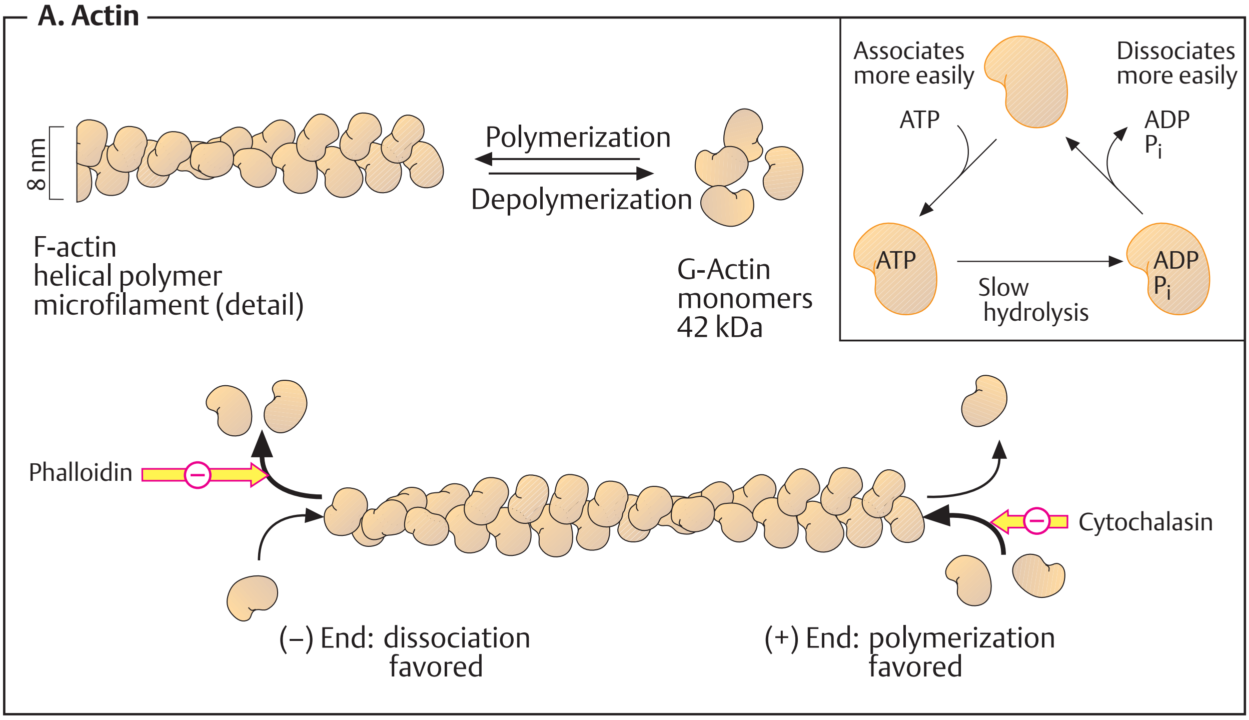
8 nm
F-actin
helical polymer
microfilament (detail)
Polymerization
Depolymerization
G-Actin
monomers
42 kDa
Associates
more easily
Dissociates
more easily
ATP
ADP
P_i
ATP
Slow
hydrolysis
ADP
P_i
Phalloidin
Cytochalasin
(–) End: dissociation
favored
(+) End: polymerization
favored

B. Intermediate filaments

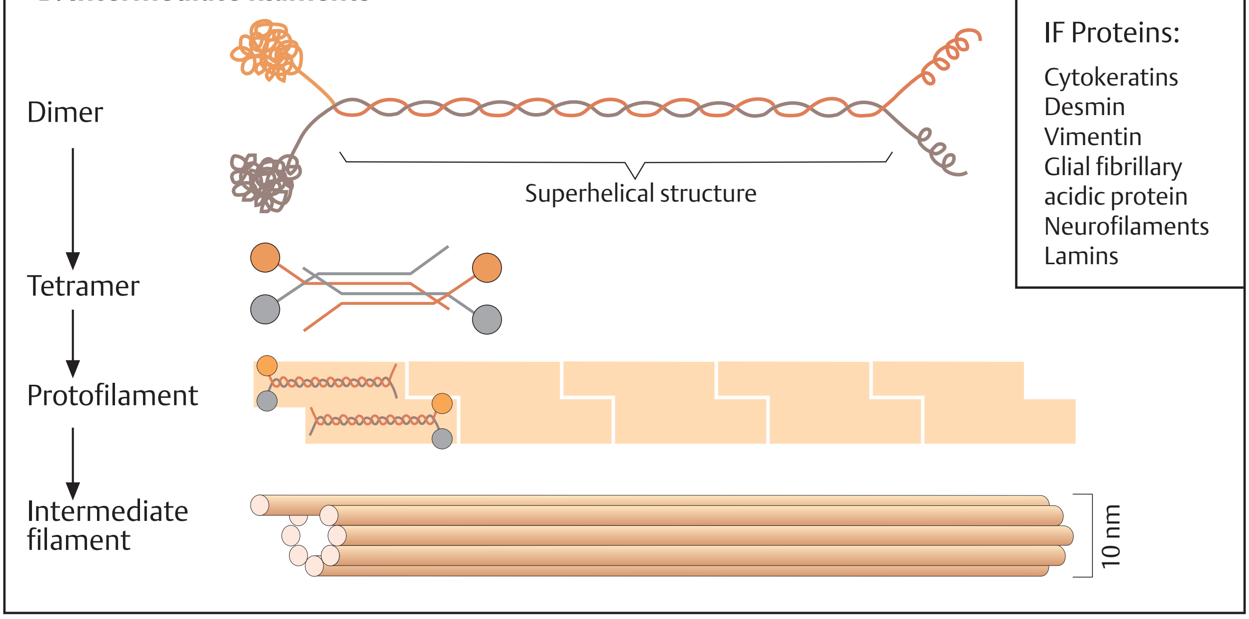
Dimer
Superhelical structure
Tetramer
Protofilament
Intermediate
filament
10 nm
IF Proteins:
Cytokeratins
Desmin
Vimentin
Glial fibrillary
acidic protein
Neurofilaments
Lamins

C. Tubulins

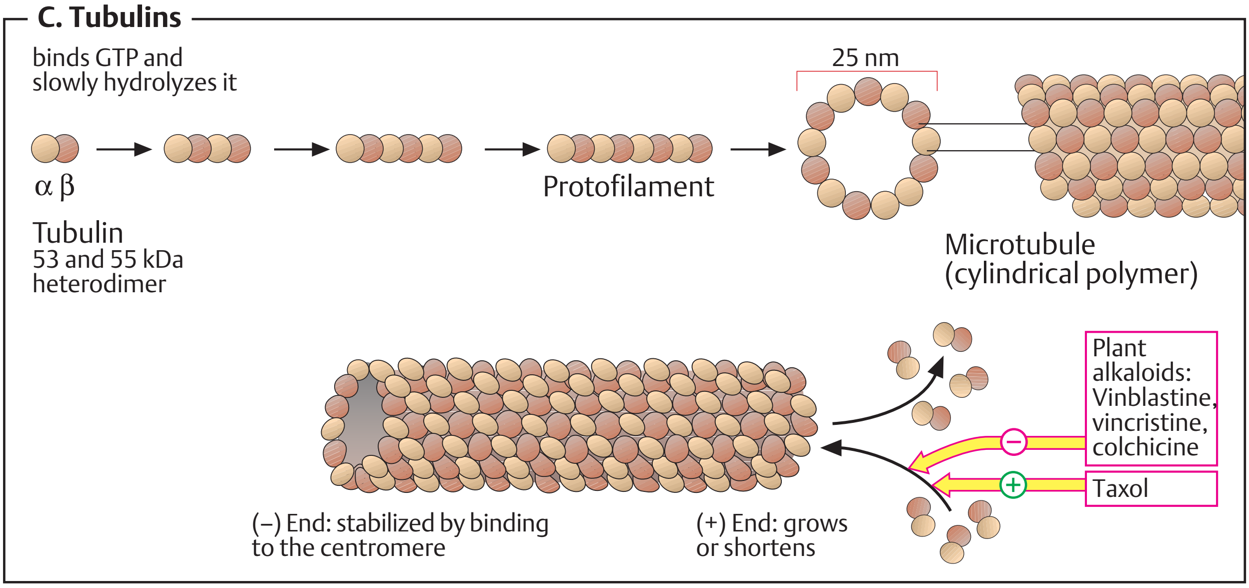
binds GTP and
slowly hydrolyzes it
α β
Tubulin
53 and 55 kDa
heterodimer
Protofilament
25 nm
Microtubule
(cylindrical polymer)
Plant
alkaloids:
Vinblastine,
vincristine,
colchicine
Taxol
(–) End: stabilized by binding
to the centromere
(+) End: grows
or shortens

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Structure and functions

The cytoskeleton carries out three major tasks:

- It represents the cell's **mechanical scaffolding**, which gives it its typical shape and connects membranes and organelles to each other. This scaffolding has dynamic properties; it is constantly being synthesized and broken down to meet the cell's requirements and changing conditions.
- It acts as the **motor for movement** of animal cells. Not only muscle cells (see p. 332), but also cells of noncontractile tissues contain many different *motor proteins,* which they use to achieve coordinated and directed movement. Cell movement, shape changes during growth, cytoplasmic streaming, and cell division are all made possible by components of the cytoskeleton.
- It serves as a **transport track** within the cell. Organelles and other large protein complexes can move along the filaments with the help of the motor proteins.

A. Microfilaments and intermediate filaments ○

The illustration schematically shows a detail of the **microvilli** of an intestinal epithelial cell as an example of the structure and function of the components of the cytoskeleton (see also **C1**).

Microfilaments of *F actin* traverse the microvilli in ordered bundles. The microfilaments are attached to each other by actin–associated proteins, particularly *fimbrin* and *villin. Calmodulin* and a myosin–like *ATPase* connect the microfilaments laterally to the plasma membrane. *Fodrin,* another microfilament–associated protein, anchors the actin fibers to each other at the base, as well as attaching them to the cytoplasmic membrane and to a network of **intermediate filaments.** In this example, the microfilaments have a mainly static function. In other cases, actin is also involved in dynamic processes. These include muscle contraction (see p. 332), cell movement, phagocytosis by immune cells, the formation of microspikes and lamellipodia (cellular extensions), and the acrosomal process during the fusion of sperm with the egg cell.

B. Microtubules ○

Only the cell's **microtubules** are shown here. They radiate out in all directions from a center near the nucleus, the **centrosome**. The tube-shaped microtubules are constantly being synthesized and broken down at their (+) ends. In the centriole, the (–) end is blocked by associated proteins (see p. 204). The (+) end can also be stabilized by associated proteins—e. g., when the microtubules have reached the cytoplasmic membrane.

The microtubules are involved in defining the shape of the cell and also serve as guiding tracks for the transport of organelles. Together with associated proteins (*dynein, kinesin*), microtubules are able to carry out mechanical work—e. g., during the transport of mitochondria, the movement of cilia (hair-like cell protrusions in the lungs, intestinal epithelium, and oviduct) and the beating of the flagella of sperm. Microtubules also play a special role in the mitotic period of cell division (see p. 394).

C. Architecture ○

The complex structure and net-like density of the cytoskeleton is illustrated here using three examples in which the cytoskeletal components are visualized with the help of antibodies.

1. The border of an *intestinal epithelial cell* is seen here (see also **B**). There are **microfilaments** (**a**) passing from the interior of the cell out into the microvilli. The filaments are firmly held together by **spectrin** (**b**), an associated protein, and they are anchored to **intermediate filaments** (**c**).

2. Only **microtubules** are seen in this *fibroblast cell.* They originate from the microtubule organizing center (centrosome) and radiate out as far as the plasma membrane.

3. **Keratin filaments** are visible here in an *epithelial cell.* Keratin fibers belong to the group of **intermediate filaments** (see pp. 70, 204; **d** = nucleus).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Microfilaments and intermediate filaments

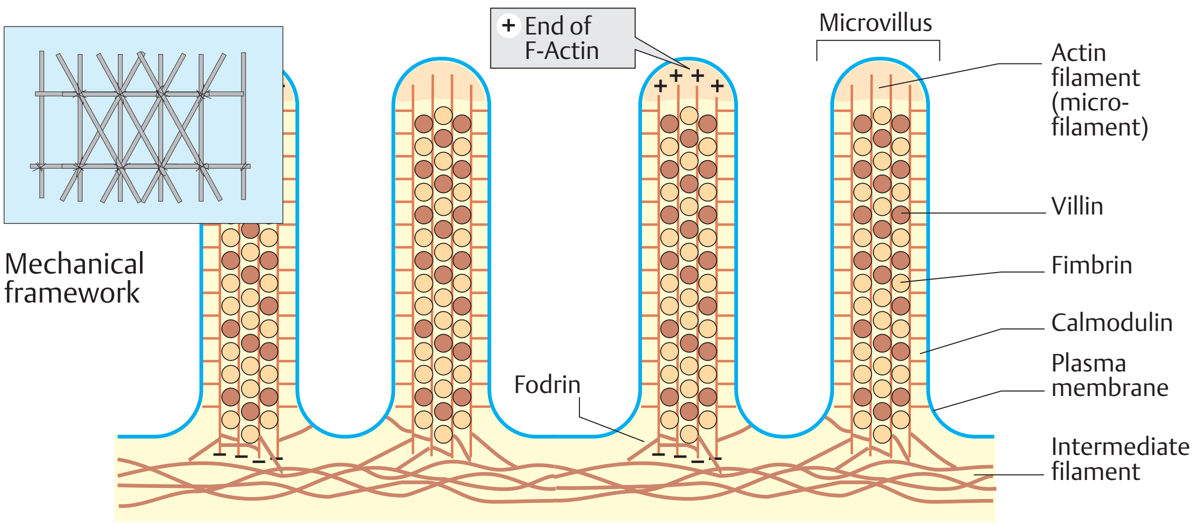
Mechanical framework
+ End of F-Actin
Microvillus
Actin filament (microfilament)
Villin
Fimbrin
Calmodulin
Plasma membrane
Fodrin
Intermediate filament
+ + + +
– – – –

B. Microtubules

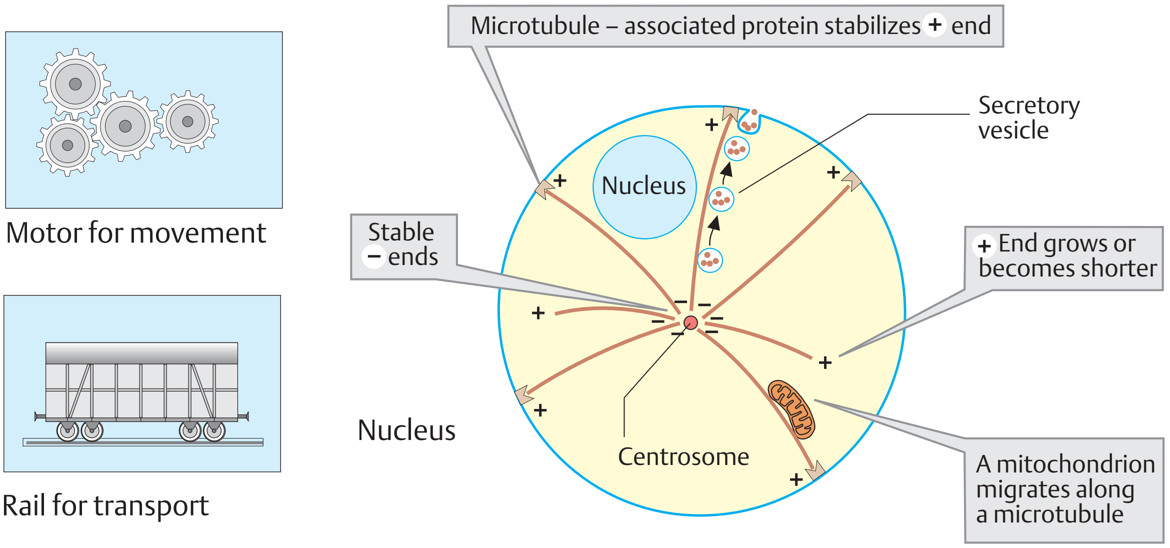
Motor for movement
Rail for transport
Microtubule – associated protein stabilizes + end
Secretory vesicle
Nucleus
Stable – ends
+ End grows or becomes shorter
Nucleus
Centrosome
A mitochondrion migrates along a microtubule
+
–

C. Architecture

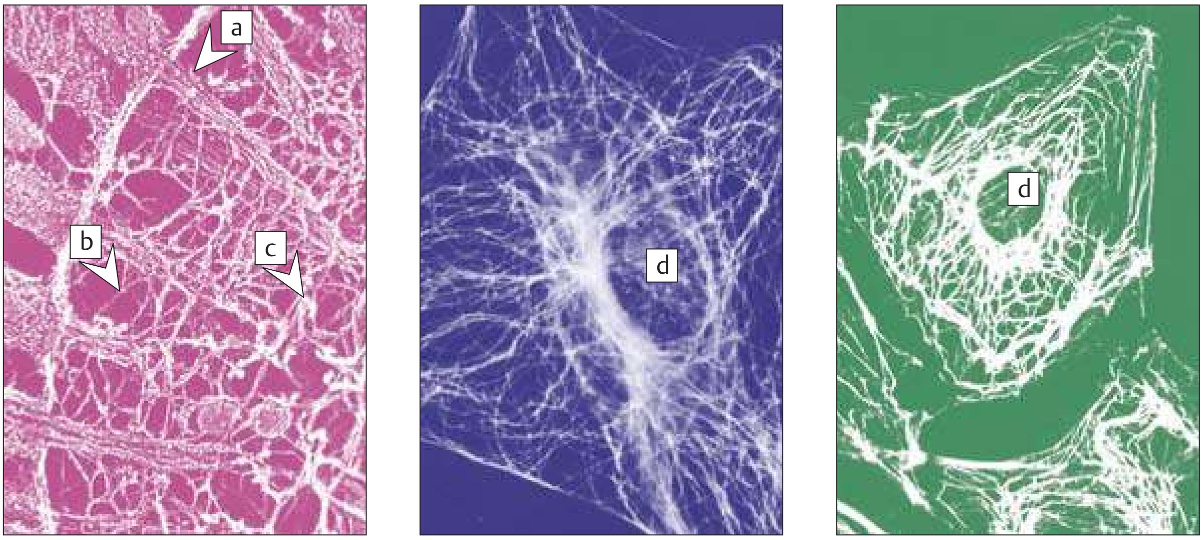

1. Microfilaments

2. Microtubules

3. Intermediate filaments

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Nucleus

A. Nucleus ●

The nucleus is the largest organelle in the eukaryotic cell. With a diameter of about 10 µm, it is easily recognizable with the light microscope. This is the location for *storage, replication,* and *expression* of genetic information.

The nucleus is separated from the cytoplasm by the **nuclear envelope**, which consists of the **outer** and **inner nuclear membranes**. Each of the two nuclear membranes has two layers, and the membranes are separated from each other by the **perinuclear space**. The outer nuclear membrane is continuous with the rough endoplasmic reticulum and is covered with ribosomes. The inner side of the membrane is covered with a protein layer (the nuclear lamina), in which the nuclear structures are anchored.

The nucleus contains almost all of the cell's **DNA** (around 1% of which is mitochondrial DNA). Together with histones and structural proteins, the nuclear DNA forms the **chromatin** (see p. 238). It is only during cell division that chromatin condenses into *chromosomes,* which are also visible with the light microscope. During this phase, the nuclear membrane temporarily disintegrates.

During the phase between cell divisions, the *interphase,* it is possible to distinguish between the more densely packed **heterochromatin** and loose **euchromatin** using an electron microscope. Active *transcription* of DNA into mRNA takes place in the region of the euchromatin. A particularly electron-dense region is noticeable in many nuclei—the **nucleolus** (several nucleoli are sometimes present). The DNA in the nucleolus contains numerous copies of the genes for rRNAs (see p. 242). They are constantly undergoing transcription, leading to a high local concentration of RNA.

B. Nuclear pores ◑

The exchange of substances between the nucleus and the cytoplasm is mediated by **pore complexes** with complicated structures, which traverse the nuclear membrane. The nuclear pores consist of numerous proteins that form several connected rings of varying diameter. Low-molecular structures and small proteins can enter the nucleus without difficulty. By contrast, larger proteins (over 40 kDa) can only pass through the nuclear pores if they carry a **nuclear localization sequence** consisting of four successive basic amino acids inside their peptide chains (see p. 228). mRNAs and rRNAs formed in the nucleus cross the pores into the cytoplasm as complexes with proteins (see below).

C. Relationships between the nucleus and cytoplasm ◑

Almost all of the RNA in the cell is synthesized in the nucleus. In this process, known as **transcription**, the information stored in DNA is transcribed into RNA (see p. 242). As mentioned above, *ribosomal RNA* (rRNA) is mainly produced in the nucleolus, while *messenger* and *transfer RNA* (mRNA and tRNA) are formed in the region of the euchromatin. Enzymatic duplication of DNA—**replication**—also only takes place in the nucleus (see p. 240).

The nucleotide components required for transcription and replication have to be imported into the nucleus from the cytoplasm. Incorporation of these components into RNA leads to primary products, which are then altered by cleavage, excision of introns, and the addition of extra nucleotides (**RNA maturation**; see p. 242). It is only once these process have been completed that the RNA molecules formed in the nucleus can be exported into the cytoplasm for protein synthesis (**translation**; see p. 250).

The nucleus is not capable of synthesizing proteins. All of the nuclear proteins therefore have to be imported—the *histones* with which DNA is associated in chromatin, and also the so-called *non-histone proteins* (DNA polymerases and RNA polymerases, auxiliary and structural proteins, transcription factors, and ribosomal proteins). Ribosomal RNA (rRNA) already associates with proteins in the nucleolus to form ribosome precursors.

A special metabolic task carried out by the nucleus is **biosynthesis of NAD^+**. The immediate precursor of this coenzyme, *nicotinamide mononucleotide* (NMN^+), arises in the cytoplasm and is then transported into the nucleolus, where it is enzymatically converted into the dinucleotide NAD^+. Finally, NAD^+ then returns to the cytoplasm.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Nucleus

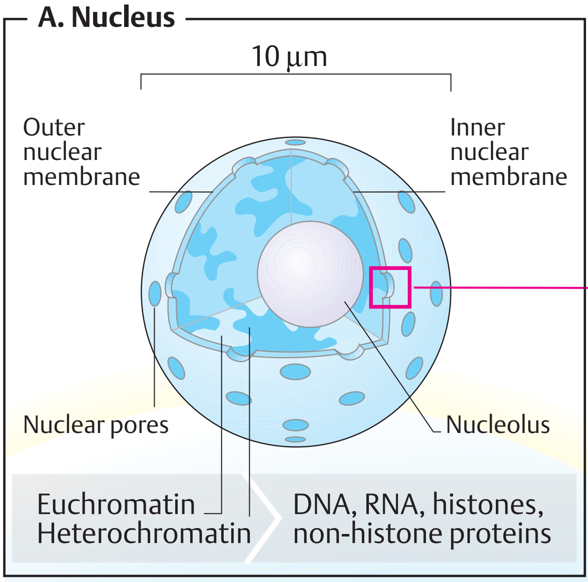
10 µm
Outer nuclear membrane
Inner nuclear membrane
Nuclear pores
Nucleolus
Euchromatin
Heterochromatin
DNA, RNA, histones, non-histone proteins

B. Nuclear pores

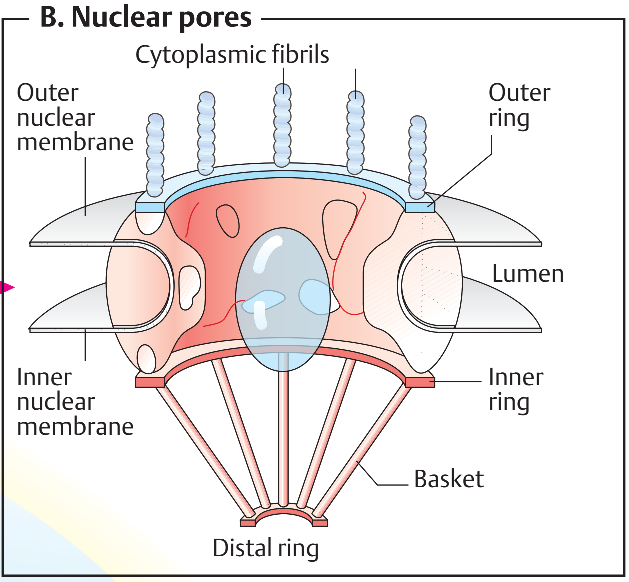
Cytoplasmic fibrils
Outer nuclear membrane
Outer ring
Lumen
Inner nuclear membrane
Inner ring
Basket
Distal ring

C. Interactions between nucleus and cytoplasm

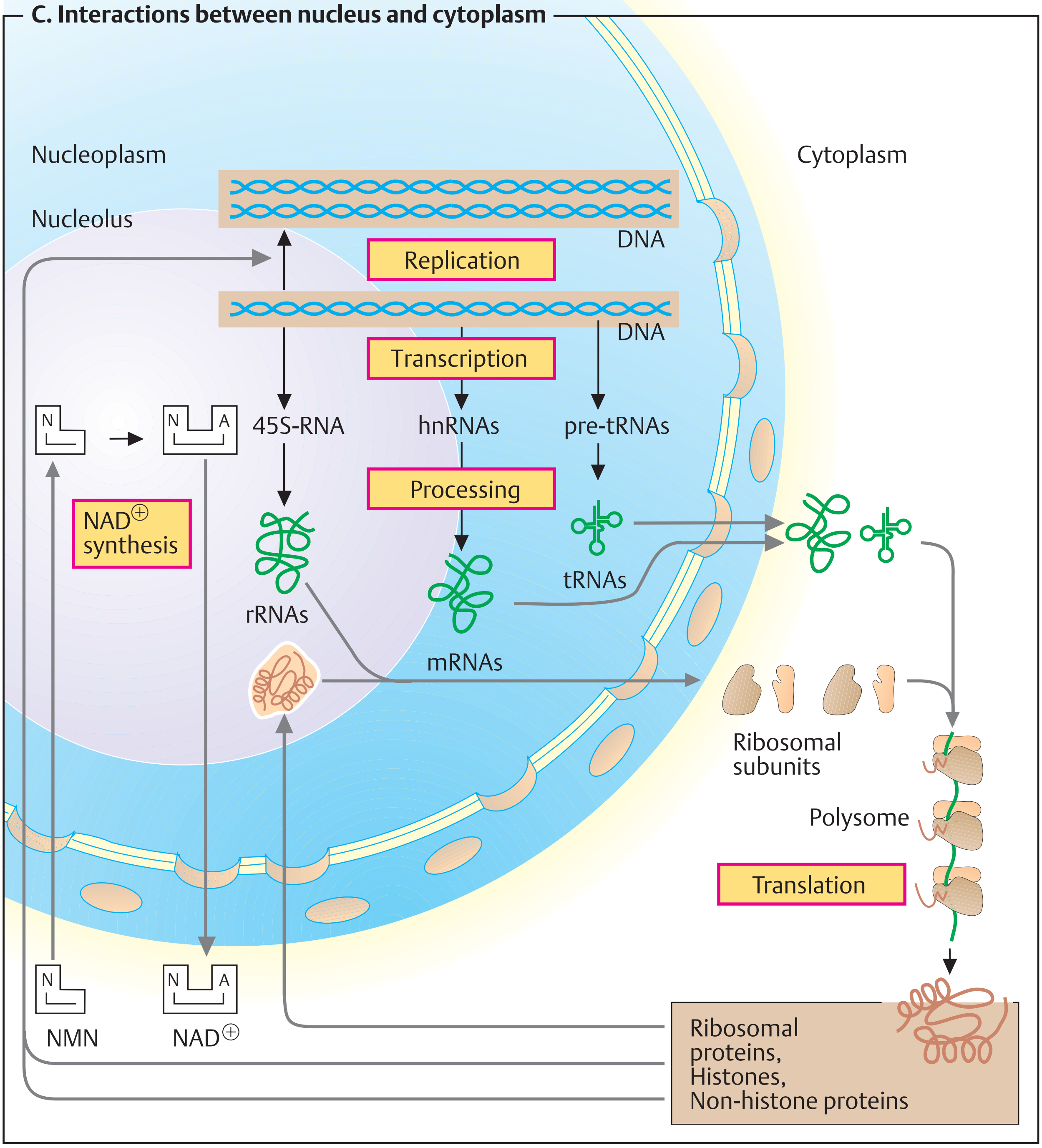
Nucleoplasm
Nucleolus
Cytoplasm
DNA
Replication
DNA
Transcription
N
N A
45S-RNA
hnRNAs
pre-tRNAs
NAD⊕ synthesis
Processing
rRNAs
mRNAs
tRNAs
Ribosomal subunits
Polysome
Translation
N
N A
NMN
NAD⊕
Ribosomal proteins, Histones, Non-histone proteins

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Structure and functions

A. Mitochondrial structure ◑

Mitochondria are bacteria-sized organelles (about 1 × 2 µm in size), which are found in large numbers in almost all eukaryotic cells. Typically, there are about 2000 mitochondria per cell, representing around 25% of the cell volume. Mitochondria are enclosed by two membranes—a smooth **outer** membrane and a markedly folded or tubular **inner mitochondrial membrane**, which has a large surface and encloses the **matrix space**. The folds of the inner membrane are known as **cristae**, and tube-like protrusions are called **tubules**. The **intermembrane space** is located between the inner and the outer membranes.

The number and shape of the mitochondria, as well as the numbers of cristae they have, can differ widely from cell type to cell type. Tissues with intensive oxidative metabolism—e. g., heart muscle—have mitochondria with particularly large numbers of cristae. Even within one type of tissue, the shape of the mitochondria can vary depending on their functional status. Mitochondria are mobile, plastic organelles.

Mitochondria probably developed during an early phase of evolution from aerobic bacteria that entered into symbiosis with primeval anaerobic eukaryotes. This **endosymbiont theory** is supported by many findings. For example, mitochondria have a ring-shaped DNA (four molecules per mitochondrion) and have their own ribosomes. The mitochondrial genome became smaller and smaller during the course of evolution. In humans, it still contains 16 569 base pairs, which code for two rRNAs, 22 tRNAs, and 13 proteins. Only these 13 proteins (mostly subunits of respiratory chain complexes) are produced in the mitochondrion. All of the other mitochondrial proteins are coded by the nuclear genome and have to be imported into the mitochondria after translation in the cytoplasm (see p. 228). The mitochondrial envelope consisting of two membranes also supports the endosymbiont theory. The inner membrane, derived from the former symbiont, has a structure reminiscent of prokaryotes. It contains the unusual lipid cardiolipin (see p. 50), but hardly any cholesterol (see p. 216).

Both mitochondrial membranes are very rich in proteins. **Porins** (see p. 214) in the outer membrane allow small molecules (< 10 kDa) to be exchanged between the cytoplasm and the intermembrane space. By contrast, the inner mitochondrial membrane is completely impermeable even to small molecules (with the exception of O_2, CO_2, and H_2O). Numerous **transporters** in the inner membrane ensure the import and export of important metabolites (see p. 212). The inner membrane also transports **respiratory chain** complexes, **ATP synthase**, and other enzymes. The matrix is also rich in enzymes (see **B**).

B. Metabolic functions ◑

Mitochondria are also described as being the cell's *biochemical powerhouse*, since—through **oxidative phosphorylation** (see p. 112)—they produce the majority of cellular ATP. **Pyruvate dehydrogenase** (PDH), the **tricarboxylic acid cycle**, **β-oxidation** of fatty acids, and parts of the **urea cycle** are located in the matrix. The **respiratory chain**, **ATP synthesis**, and enzymes involved in **heme biosynthesis** (see p. 192) are associated with the inner membrane.

The inner membrane itself plays an important part in oxidative phosphorylation. As it is impermeable to protons, the respiratory chain—which pumps protons from the matrix into the intermembrane space via complexes I, III, and IV—establishes a **proton gradient** across the inner membrane, in which the chemical energy released during NADH oxidation is conserved (see p. 126). ATP synthase then uses the energy stored in the gradient to form ATP from ADP and inorganic phosphate. Several of the transport systems are also dependent on the H^+ gradient.

In addition to the endoplasmic reticulum, the mitochondria also function as an intracellular **calcium reservoir**. The mitochondria also play an important role in "programmed cell death"—**apoptosis** (see p. 396).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Mitochondrial structure

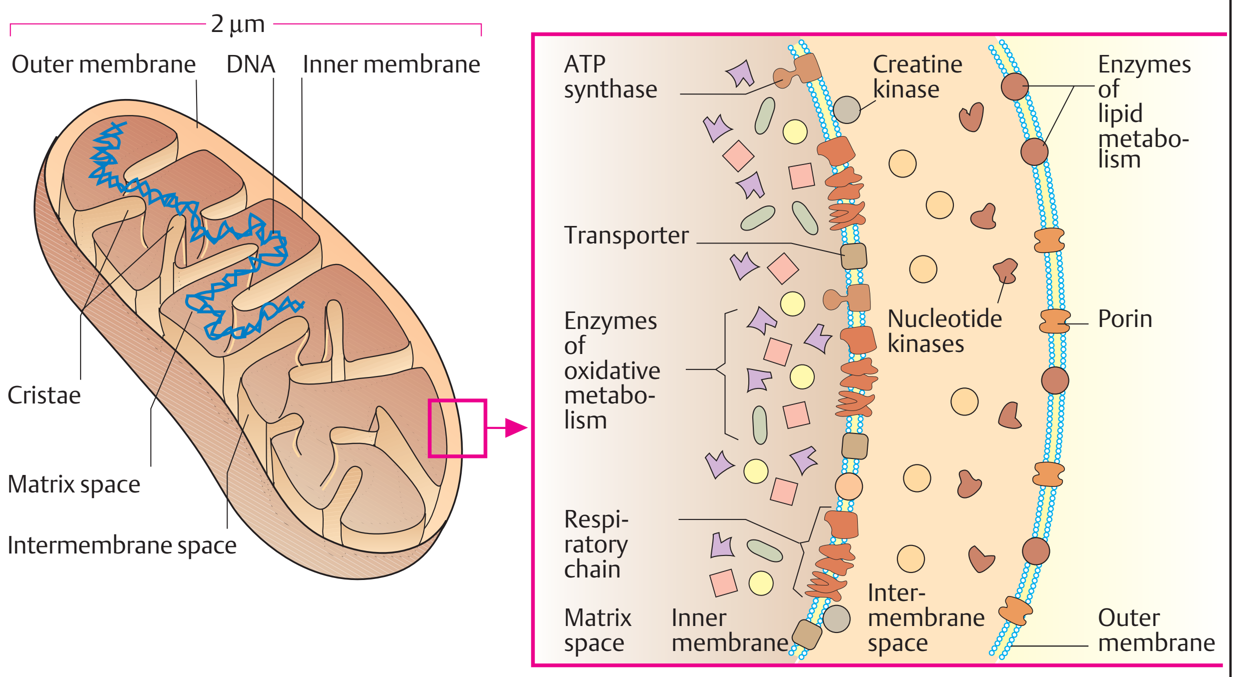
2 µm
Outer membrane
DNA
Inner membrane
Cristae
Matrix space
Intermembrane space
ATP synthase
Creatine kinase
Enzymes of lipid metabolism
Transporter
Enzymes of oxidative metabolism
Nucleotide kinases
Porin
Respiratory chain
Matrix space
Inner membrane
Intermembrane space
Outer membrane

B. Metabolic functions

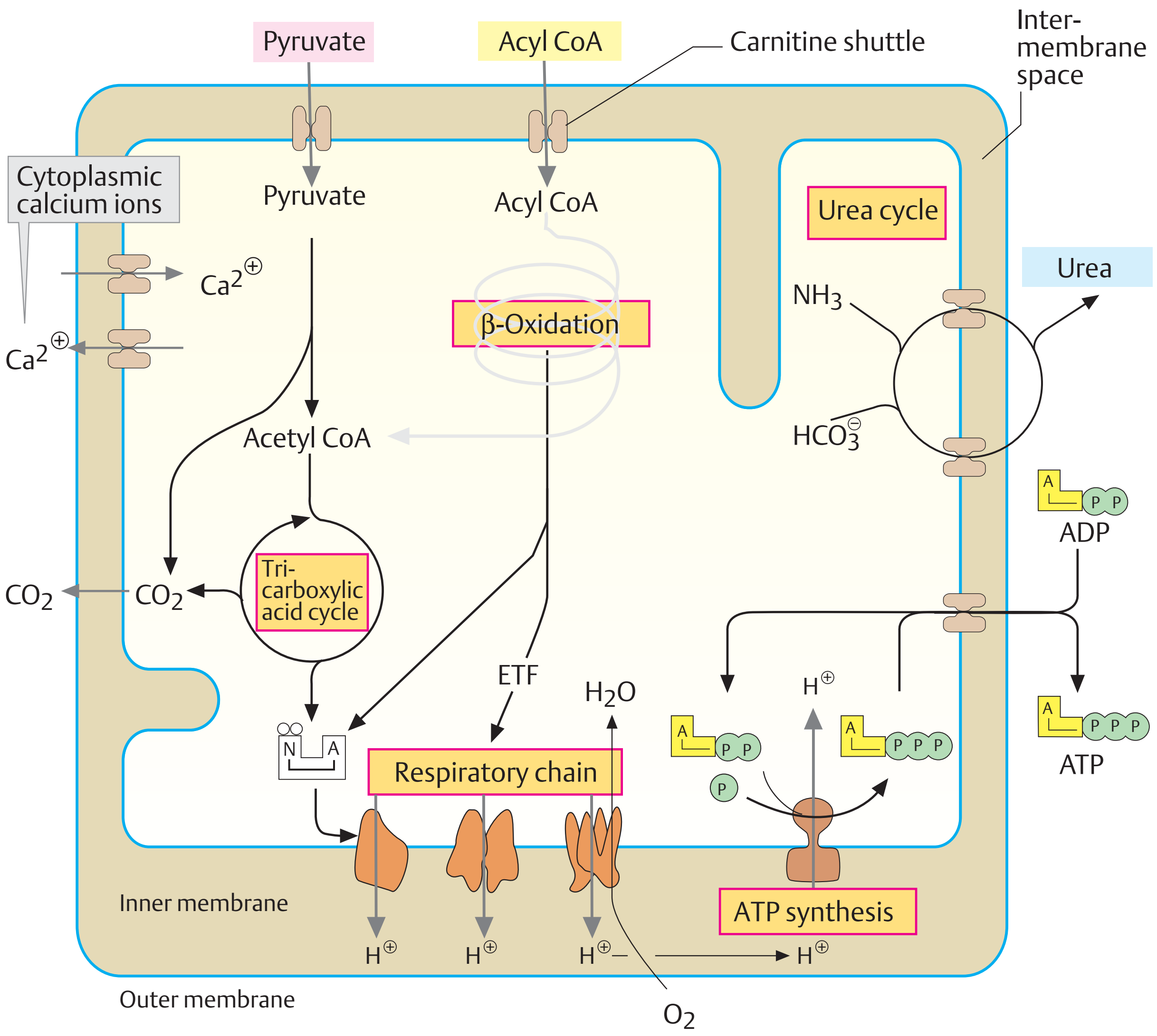
Pyruvate
Acyl CoA
Carnitine shuttle
Intermembrane space
Cytoplasmic calcium ions
Pyruvate
Acyl CoA
Urea cycle
Urea
$Ca^{2\oplus}$
$Ca^{2\oplus}$
β-Oxidation
NH_3
$HCO_3^{\ominus}$
Acetyl CoA
ADP
Tricarboxylic acid cycle
CO_2
CO_2
ETF
H_2O
$H^{\oplus}$
ATP
NAD
Respiratory chain
Inner membrane
ATP synthesis
$H^{\oplus}$
$H^{\oplus}$
$H^{\oplus}$
$H^{\oplus}$
Outer membrane
O_2

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Transport systems

Mitochondria are surrounded by an inner and an outer membrane (see p. 210). The **outer membrane** contains porins, which allow smaller molecules up to 10 kDa in size to pass. By contrast, the **inner membrane** is also impermeable to small molecules (with the exception of water and the gases O_2, CO_2, and NH_3). All of the other substrates of mitochondrial metabolism, as well as its products, therefore have to be moved through the inner membrane with the help of special **transporters**.

A. Transport systems ◑

The transport systems of the inner mitochondrial membrane use various mechanisms. Metabolites or ions can be transported alone (uniport, **U**), together with a second substance (symport, **S**), or in exchange for another molecule (antiport, **A**). Active transport—i. e., transport coupled to ATP hydrolysis—does not play an important role in mitochondria. The driving force is usually the **proton gradient** across the inner membrane (blue star) or the general membrane potential (red star; see p. 126).

The **pyruvate** (left) formed by glycolysis in the cytoplasm is imported into the matrix in antiport with OH^-. The OH^- ions react in the intermembrane space with the H^+ ions abundantly present there to form H_2O. This maintains a concentration gradient of OH^-. The import of **phosphate** ($H_2PO_4^-$) is driven in a similar way. The exchange of the **ATP** formed in the mitochondrion for **ADP** via an *adenine nucleotide translocase* (center) is also dependent on the H^+ gradient. ATP with a quadruple negative charge is exchanged for ADP with a triple negative charge, so that overall one negative charge is transported into the H^+-rich intermembrane space. The import of malate by the *tricarboxylate transporter*, which is important for the malate shuttle (see **B**) is coupled to the export of citrate, with a net export of one negative charge to the exterior again. In the opposite direction, malate can leave the matrix in antiport for phosphate. When Ca^{2+} is imported (right), the metal cation follows the membrane potential. An electroneutral antiport for two H^+ or two Na^+ serves for Ca^{2+} export.

B. Malate and glycerophosphate shuttles ◑

Two systems known as "shuttles" are available to allow the **import of reducing equivalents** that arise from glycolysis in the cytoplasm in the form of NADH+H^+. There is no transporter in the inner membrane for NADH+H^+ itself.

In the **malate shuttle** (left)—which operates in the heart, liver, and kidneys, for example—oxaloacetic acid is reduced to malate by malate dehydrogenase (MDH, [2a]) with the help of NADH+H^+. In antiport for 2-oxoglutarate, malate is transferred to the matrix, where the mitochondrial isoenzyme for MDH [2b] regenerates oxaloacetic acid and NADH+H^+. The latter is reoxidized by complex I of the respiratory chain, while oxaloacetic acid, for which a transporter is not available in the inner membrane, is first transaminated to aspartate by aspartate aminotransferase (AST, [3a]). Aspartate leaves the matrix again, and in the cytoplasm once again supplies oxaloacetate for step [2a] and glutamate for return transport into the matrix [3b]. On balance, only NADH+H^+ is moved from the cytoplasm into the matrix; ATP is not needed for this.

The **glycerophosphate shuttle** (right) was discovered in insect muscle, but is also active in the skeletal musculature and brain in higher animals. In this shuttle, NADH+H^+ formed in the cytoplasm is used to reduce glycerone 3-phosphate, an intermediate of glycolysis (see p. 150) to glycerol 3-phosphate. Via porins, this enters the intermembrane space and is oxidized again there on the exterior side of the inner membrane by the flavin enzyme *glycerol 3-phosphate dehydrogenase* back into glycerone 3-phosphate. The reducing equivalents are passed on to the respiratory chain via **ubiquinone (coenzyme Q)**.

The **carnitine shuttle** for transporting acyl residues into the mitochondrial matrix is discussed on p. 164.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Transport systems

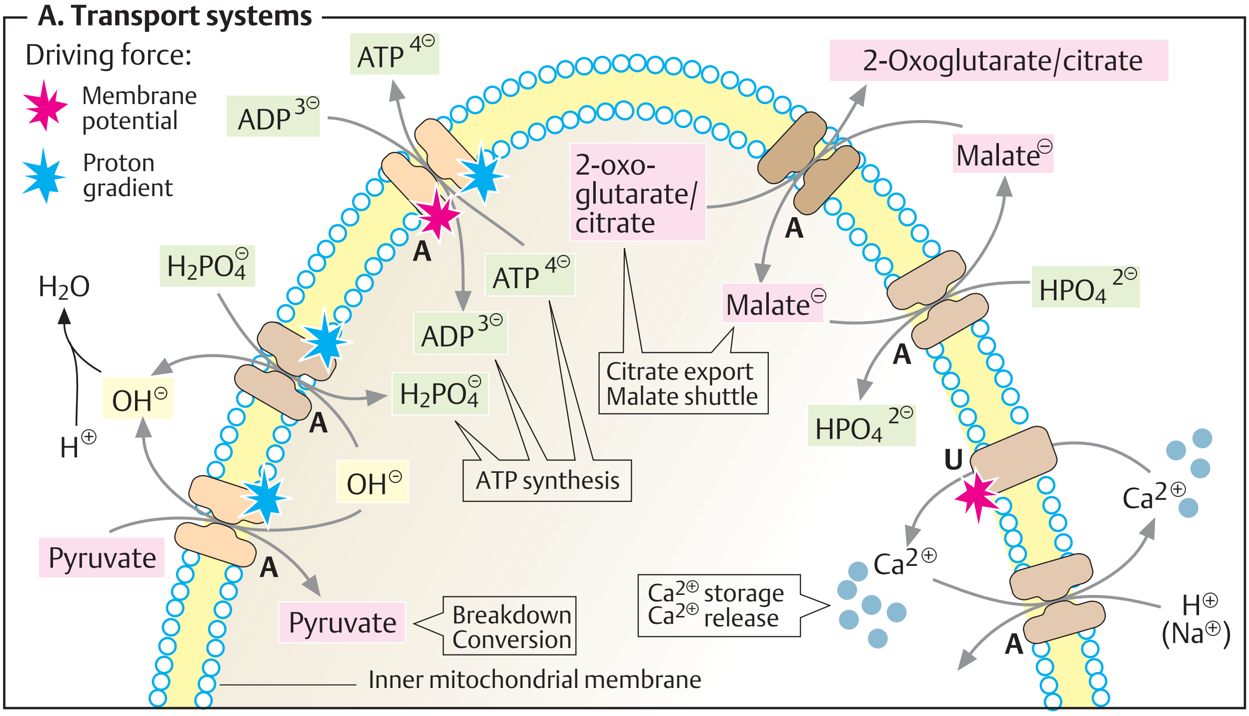
Driving force:
Membrane potential
Proton gradient
$ATP^{4\ominus}$
$ADP^{3\ominus}$
2-Oxoglutarate/citrate
2-oxo-glutarate/citrate
Malate⊖
A
H_2O
$H_2PO_4^{\ominus}$
$ATP^{4\ominus}$
$ADP^{3\ominus}$
Malate⊖
$HPO_4^{2\ominus}$
Citrate export
Malate shuttle
$OH^{\ominus}$
$H^{\oplus}$
$H_2PO_4^{\ominus}$
$HPO_4^{2\ominus}$
U
$OH^{\ominus}$
ATP synthesis
$Ca^{2\oplus}$
Pyruvate
$Ca^{2\oplus}$
Pyruvate
Breakdown
Conversion
$Ca^{2\oplus}$ storage
$Ca^{2\oplus}$ release
$H^{\oplus}$
$(Na^{\oplus})$
Inner mitochondrial membrane

B. Malate and glycerophosphate shuttle

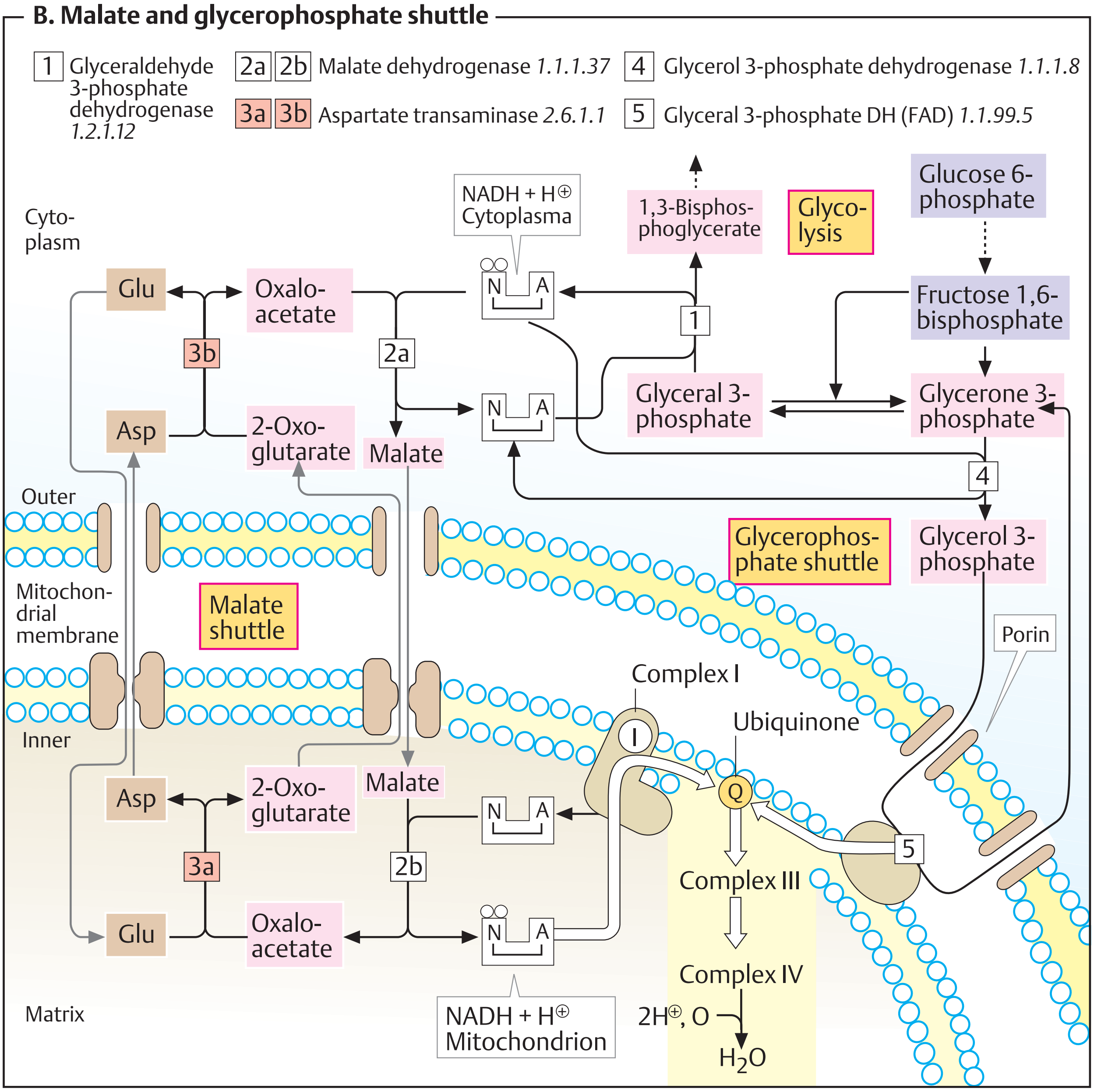
1 Glyceraldehyde 3-phosphate dehydrogenase *1.2.1.12*
2a 2b Malate dehydrogenase *1.1.1.37*
3a 3b Aspartate transaminase *2.6.1.1*
4 Glycerol 3-phosphate dehydrogenase *1.1.1.8*
5 Glyceral 3-phosphate DH (FAD) *1.1.99.5*
Cyto-plasm
NADH + H⊕ Cytoplasma
1,3-Bisphos-phoglycerate
Glyco-lysis
Glucose 6-phosphate
Glu
Oxalo-acetate
N A
1
Fructose 1,6-bisphosphate
3b
2a
Asp
2-Oxo-glutarate
Malate
Glyceral 3-phosphate
Glycerone 3-phosphate
4
Outer
Glycerophos-phate shuttle
Glycerol 3-phosphate
Mitochon-drial membrane
Malate shuttle
Porin
Complex I
Ubiquinone
Inner
I
Q
Asp
2-Oxo-glutarate
Malate
5
3a
2b
Complex III
Glu
Oxalo-acetate
Complex IV
Matrix
NADH + H⊕ Mitochondrion
$2H^{\oplus}$, O
H_2O

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Structure and components

A. Structure of the plasma membrane ◑

All biological membranes are constructed according to a standard pattern. They consist of a continuous **bilayer of amphipathic lipids** approximately 5 nm thick, into which **proteins** are embedded. In addition, some membranes also carry **carbohydrates** (mono- and oligosaccharides) on their exterior, which are bound to lipids and proteins. The proportions of lipids, proteins, and carbohydrates differ markedly depending on the type of cell and membrane (see p. 216).

Membrane lipids are strongly *amphipathic molecules* with a polar hydrophilic "head group" and an apolar hydrophobic "tail." In membranes, they are primarily held together by the hydrophobic effect (see p. 28) and weak Van der Waals forces, and are therefore mobile relative to each other. This gives membranes a more or less fluid quality.

The **fluidity** of membranes primarily depends on their lipid composition and on temperature. At a specific **transition temperature**, membranes pass from a semicrystalline state to a more fluid state. The double bonds in the alkyl chains of unsaturated acyl residues in the membrane lipids disturb the semicrystalline state. The *higher* the proportion of unsaturated lipids present, therefore, the *lower* the transition temperature. The cholesterol content also influences membrane fluidity. While cholesterol increases the fluidity of semicrystalline, closely-packed membranes, it *stabilizes* fluid membranes that contain a high proportion of unsaturated lipids.

Like lipids, proteins are also mobile within the membrane. If they are not fixed in place by special mechanisms, they float within the lipid layer as if in a two-dimensional liquid; biological membranes are therefore also described as being a "fluid mosaic."

Lipids and proteins can shift easily *within* one layer of a membrane, but switching between the two layers (*"flip/flop"*) is not possible for proteins and is only possible with difficulty for lipids (with the exception of cholesterol). To move to the other side, phospholipids require special auxiliary proteins (translocators, "flipases").

B. Membrane lipids ◑

The illustration shows a model of a small section of a membrane. The **phospholipids** are the most important group of membrane lipids. They include *phosphatidylcholine* (lecithin), *phosphatidylethanolamine*, *phosphatidylserine, phosphatidylinositol,* and *sphingomyelin* (for their structures, see p. 50). In addition, membranes in animal cells also contain **cholesterol** (with the exception of inner mitochondrial membranes). **Glycolipids** (a *ganglioside* is shown here) are mainly found on the outside of the plasma membrane. Together with the *glycoproteins,* they form the exterior coating of the cell (the *glycocalyx*).

C. Membrane proteins ◑

Proteins can be anchored in or on membranes in various ways. **Integral membrane proteins** cross right through the lipid bilayer. The sections of the peptide chains that lie within the bilayer usually consist of 20 to 25 mainly hydrophobic amino acid residues that form a right-handed α-helix.

Type I and II membrane proteins only contain *one* **transmembrane helix** of this type, while type III proteins contain several. Rarely, type I and II polypeptides can aggregate to form a type IV transmembrane protein. Several groups of integral membrane proteins—e. g., the porins (see p. 212)—penetrate the membrane with antiparallel β-sheet structures. Due to its shape, this tertiary structure is known as a "β-barrel."

Type V and VI proteins carry **lipid anchors.** These are fatty acids (palmitic acid, myristic acid), isoprenoids (e. g., farnesol), or glycolipids such as glycosyl phosphatidylinositol (GPI) that are covalently bound to the peptide chain.

Peripheral membrane proteins are associated with the head groups of phospholipids or with another integral membrane protein (not shown).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Structure of the plasma membrane

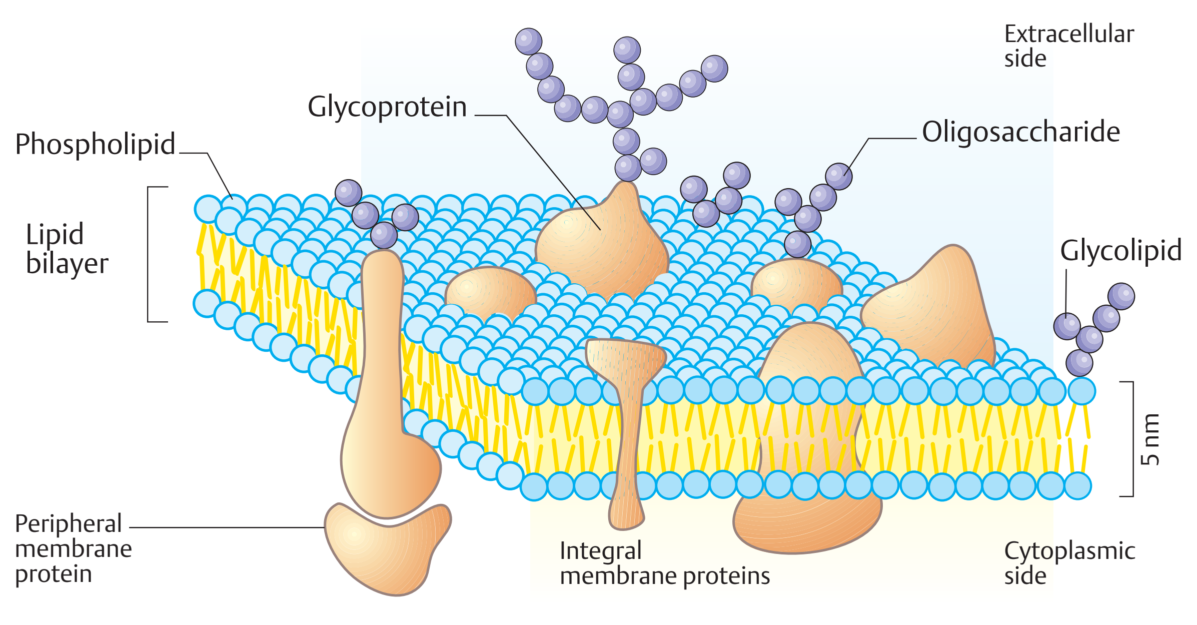
Extracellular side
Glycoprotein
Oligosaccharide
Phospholipid
Lipid bilayer
Glycolipid
5 nm
Peripheral membrane protein
Integral membrane proteins
Cytoplasmic side

B. Membrane lipids

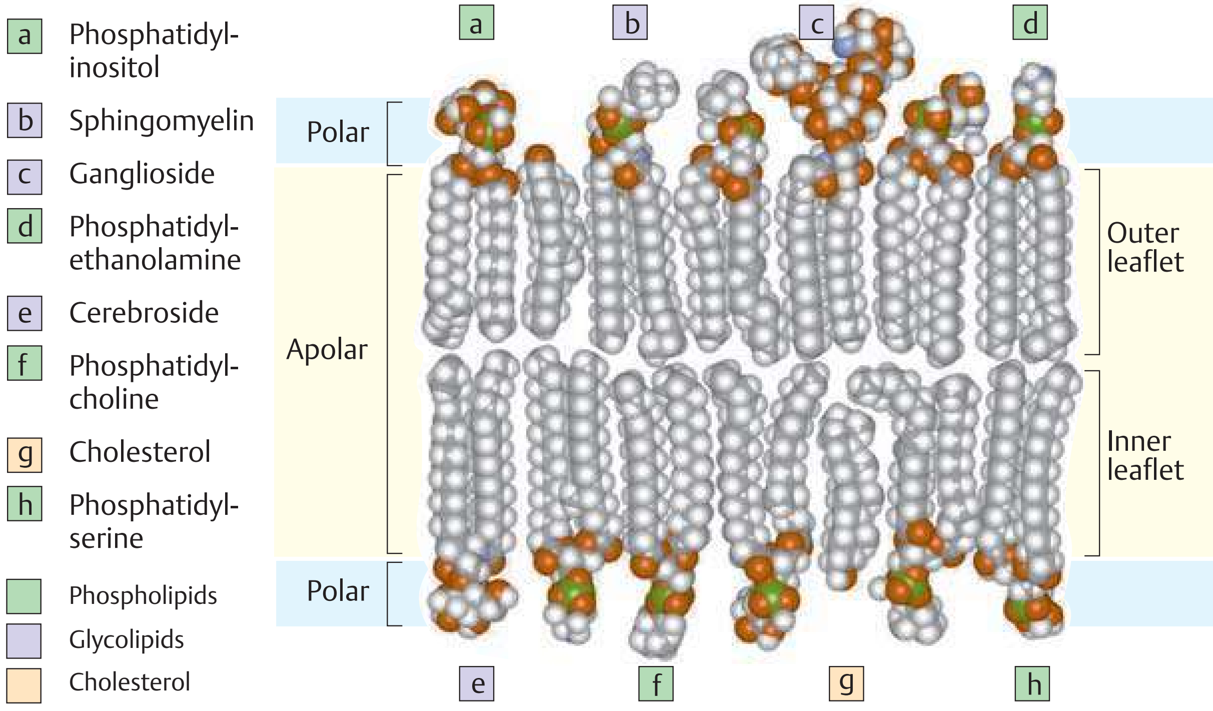
a Phosphatidyl-inositol
b Sphingomyelin
c Ganglioside
d Phosphatidyl-ethanolamine
e Cerebroside
f Phosphatidyl-choline
g Cholesterol
h Phosphatidyl-serine
🟩 Phospholipids
🟪 Glycolipids
🟧 Cholesterol
a
b
c
d
Polar
Apolar
Polar
Outer leaflet
Inner leaflet
e
f
g
h

C. Membrane proteins

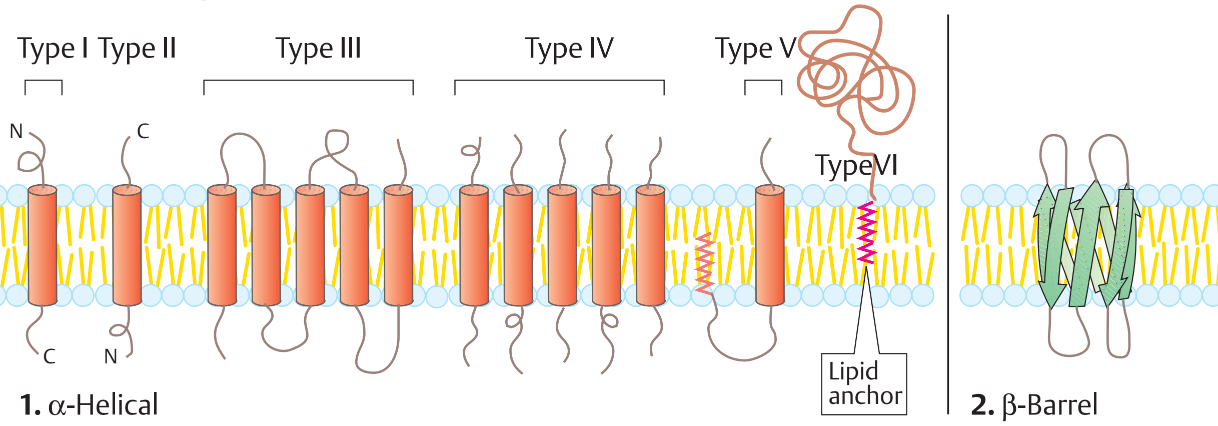
Type I
Type II
Type III
Type IV
Type V
Type VI
N
C
C
N
Lipid anchor
1. α-Helical
2. β-Barrel

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Functions and composition

The most important membranes in animal cells are the *plasma membrane,* the inner and outer *nuclear membranes,* the membranes of the *endoplasmic reticulum* (ER) and the *Golgi apparatus,* and the inner and outer *mitochondrial membranes. Lysosomes, peroxisomes,* and various *vesicles* are also separated from the cytoplasm by membranes. In plants, additional membranes are seen in the plastids and vacuoles. All membranes show *polarity*—i. e., there is a difference in the composition of the inner layer (facing toward the cytoplasm) and the outer layer (facing away from it).

A. Functions of membranes ●

Membranes and their components have the following functions:

1. **Enclosure and insulation** of cells and organelles. The *enclosure* provided by the plasma membrane protects cells from their environment both mechanically and chemically. The plasma membrane is essential for maintaining differences in the concentration of many substances between the intracellular and extracellular compartments.

2. **Regulated transport of substances**, which determines the *internal milieu* and is a precondition for *homeostasis*—i. e., the maintenance of constant concentrations of substances and physiological parameters. Regulated and selective transport of substances through pores, channels, and transporters (see p. 218) is necessary because the cells and organelles are enclosed by membrane systems.

3. **Reception of extracellular signals** and transfer of these signals to the inside of the cell (see pp. 384ff.), as well as the production of signals.

4. **Enzymatic catalysis** of reactions. Important enzymes are located in membranes at the interface between the lipid and aqueous phases. This is where reactions with apolar substrates occur. Examples include *lipid biosynthesis* (see p. 170) and the *metabolism of apolar xenobiotics* (see p. 316). The most important reactions in energy conversion—i. e., *oxidative phosphorylation* (see p. 140) and *photosynthesis* (see p. 128)—also occur in membranes.

5. **Interactions with other cells** for the purposes of cell fusion and tissue formation, as well as communication with the extracellular matrix.

6. **Anchoring of the cytoskeleton** (see p. 204) to maintain the shape of cells and organelles and to provide the basis for movement processes.

B. Composition of membranes ◑

Biological membranes consist of **lipids**, **proteins**, and **carbohydrates** (see p. 214). These components occur in varying proportions (left). Proteins usually account for the largest proportion, at around half. By contrast, carbohydrates, which are only found on the side facing away from the cytoplasm, make up only a few percent. An extreme composition is seen in *myelin,* the insulating material in nerve cells, three-quarters of which consists of lipids. By contrast, the inner *mitochondrial membrane* is characterized by a very low proportion of lipids and a particularly high proportion of proteins.

When the individual proportions of **lipids** in membranes are examined more closely (right part of the illustration), typical patterns for particular cells and tissues are also found. The illustration shows the diversity of the membrane lipids and their approximate quantitative composition. *Phospholipids* are predominant in membrane lipids in comparison with *glycolipids* and *cholesterol.* Triacylglycerols (neutral fats) are not found in membranes.

Cholesterol is found almost exclusively in eukaryotic cells. Animal membranes contain substantially more cholesterol than plant membranes, in which cholesterol is usually replaced by other sterols. There is no cholesterol at all in prokaryotes (with a few exceptions). The inner mitochondrial membrane of eukaryotes is also low in cholesterol, while it is the only membrane that contains large amounts of cardiolipin. These facts both support the endosymbiotic theory of the development of mitochondria (see p. 210).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Functions of membranes

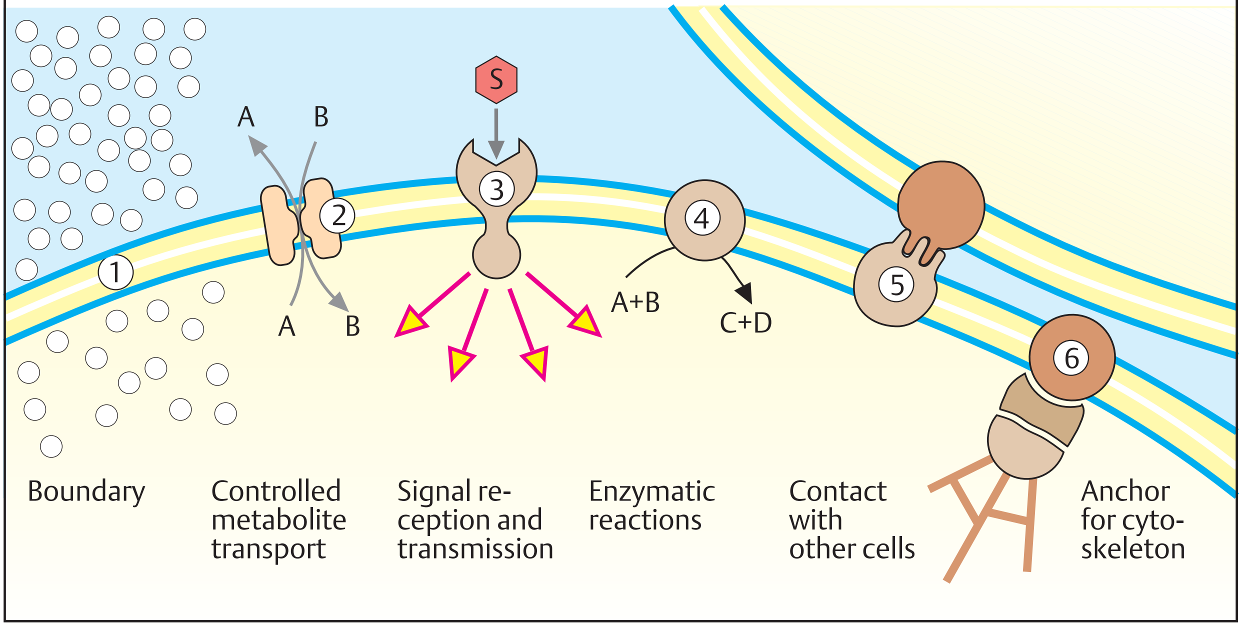
S
A
B
A
B
1
2
3
4
5
6
A+B
C+D
Boundary
Controlled metabolite transport
Signal reception and transmission
Enzymatic reactions
Contact with other cells
Anchor for cytoskeleton

B. Composition of membranes

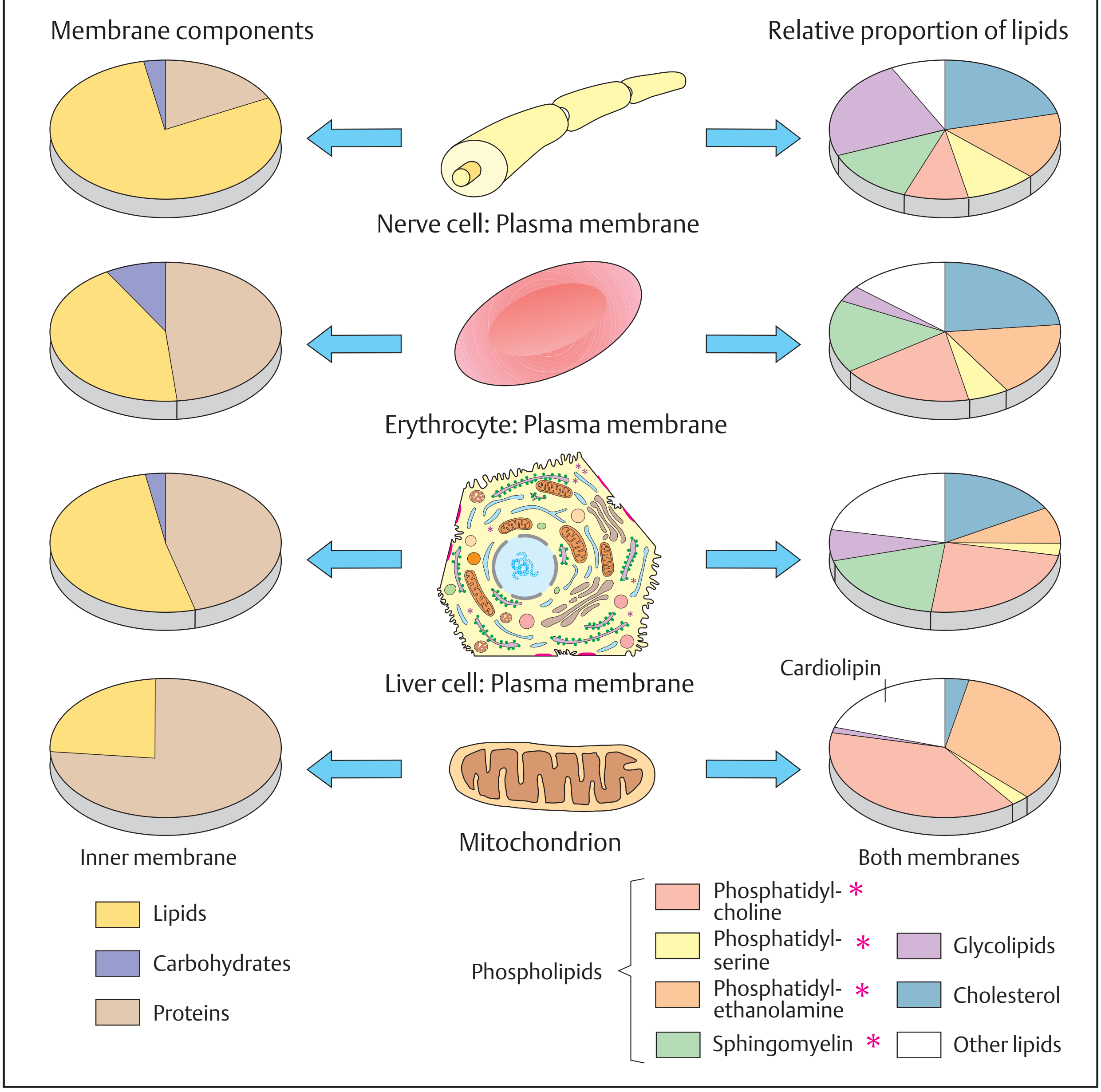

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Transport processes

A. Permeability ●

Only small, uncharged molecules such as gases, water, ammonia, glycerol, or urea are able to pass through biological membranes by *free diffusion*. With increasing **size**, even compounds of this type are no longer able to pass through. Membranes are impermeable to glucose and other sugars, for example.

The **polarity** of a molecule is also important. Apolar substances, such as benzene, ethanol, diethyl ether, and many narcotic agents are able to enter biological membranes easily. By contrast, membranes are impermeable to strongly polar compounds, particularly those that are electrically charged. To be able to take up or release molecules of this type, cells have specialized *channels* and *transporters* in their membranes (see below).

B. Passive and active transport ●

Free diffusion is the simplest form of membrane transport. When it is supported by integral membrane proteins, it is known as **facilitated diffusion** (or facilitated transport).

1. **Channel proteins** have a polar pore through which ions and other hydrophilic compounds can pass. For example, there are channels that allow selected ions to pass (**ion channels**; see p. 222) and **porins** that allow molecules below a specific size to pass in a more or less nonspecific fashion (see p. 212).
2. **Transporters** recognize and bind the molecules to be transported and help them to pass through the membrane as a result of a conformational change. These proteins (permeases) are thus comparable with enzymes—although with the difference that they "catalyze" vectorial transport rather than an enzymatic reaction. Like enzymes, they show a certain *affinity* for each molecule transported (expressed as the dissociation constant, K_d in mol L^{-1}) and a *maximum transport capacity* (V).

Free diffusion and transport processes facilitated by ion channels and transport proteins always follow a *concentration gradient*—i. e., the direction of transport is from the site of higher concentration to the site of lower concentration. In ions, the *membrane potential* also plays a role; the processes are summed up by the term *"electrochemical gradient"* (see p. 126). These processes therefore involve **passive transport**, which runs "downhill" on the slope of a gradient.

By contrast, **active transport** can also run "uphill"—i. e., against a concentration or charge gradient. It therefore requires an input of *energy*, which is usually supplied by the hydrolysis of ATP (see p. 124). The transporter first binds its "cargo" on one side of the membrane. ATP-dependent phosphorylation then causes a conformation change that releases the cargo on the other side of the membrane (see p. 220). A non-spontaneous transport process can also take place through coupling to another active transport process (known as *secondary active transport;* see p. 220).

Using the transport systems in the membranes, cells regulate their volume, internal pH value, and ionic environment. They concentrate metabolites that are important for energy metabolism and biosynthesis, and exclude toxic substances. Transport systems also serve to establish *ion gradients*, which are required for oxidative phosphorylation and stimulation of muscle and nerve cells, for example (see p. 350).

C. Transport processes ◑

Another classification of transport processes is based on the number of particles transported and the direction in which they move. When a *single* molecule or ion passes through the membrane with the help of a channel or transporter, the process is described as a **uniport** (example: the transport of glucose into liver cells). Simultaneous transport of *two different* particles can take place either as a **symport** (example: the transport of amino acids or glucose together with Na^+ ions into intestinal epithelial cells) or as an **antiport**. Ions are often transported in an antiport in exchange for another similarly charged ion. This process is **electroneutral** and therefore more energetically favorable (example: the exchange of HCO_3^- for Cl^- at the erythrocyte membrane).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

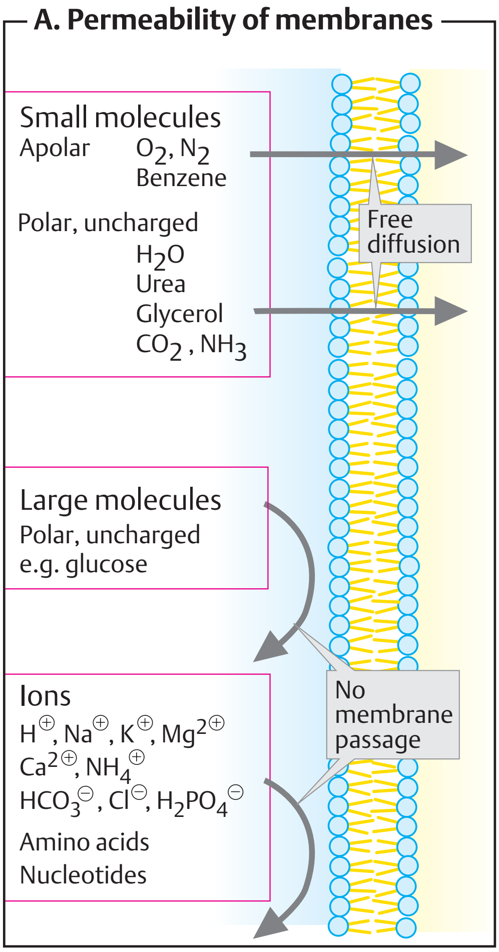
A. Permeability of membranes
Small molecules
Apolar
O_2, N_2
Benzene
Polar, uncharged
H_2O
Urea
Glycerol
CO_2, NH_3
Free diffusion
Large molecules
Polar, uncharged
e.g. glucose
Ions
$H^{\oplus}$, $Na^{\oplus}$, $K^{\oplus}$, $Mg^{2\oplus}$
$Ca^{2\oplus}$, $NH_4^{\oplus}$
$HCO_3^{\ominus}$, $Cl^{\ominus}$, $H_2PO_4^{\ominus}$
Amino acids
Nucleotides
No membrane passage

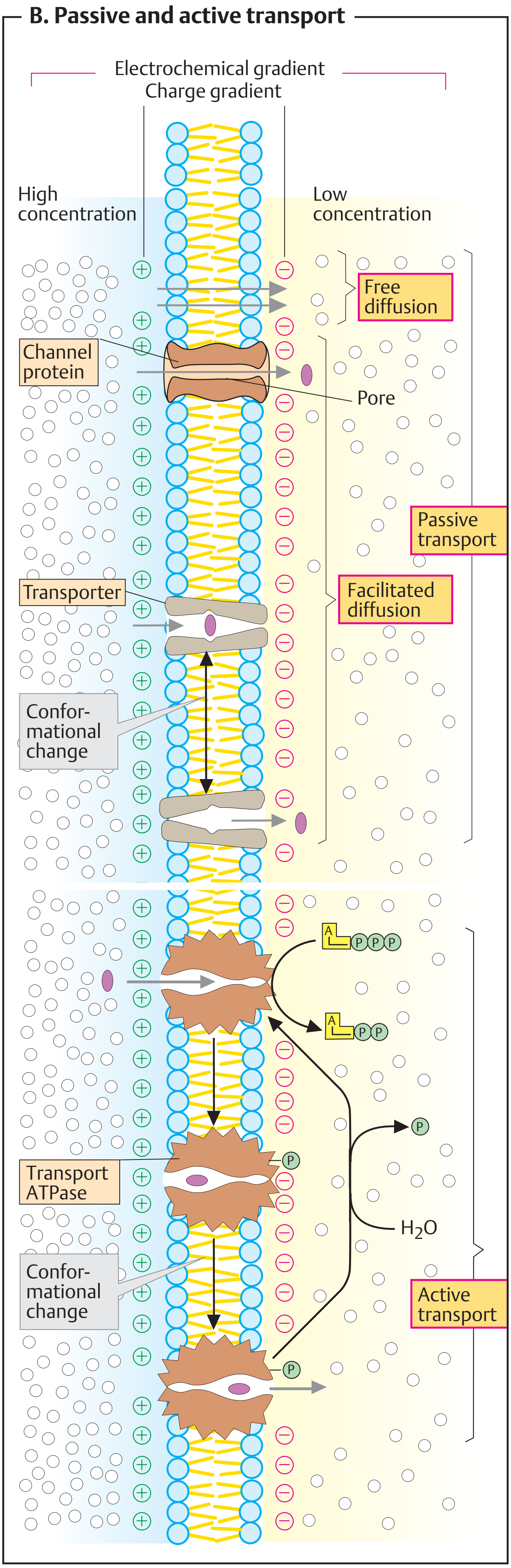
B. Passive and active transport
Electrochemical gradient
Charge gradient
High concentration
Low concentration
Free diffusion
Channel protein
Pore
Passive transport
Transporter
Facilitated diffusion
Confor-mational change
A P P P
A P P
P
H_2O
Transport ATPase
P
Confor-mational change
P
Active transport

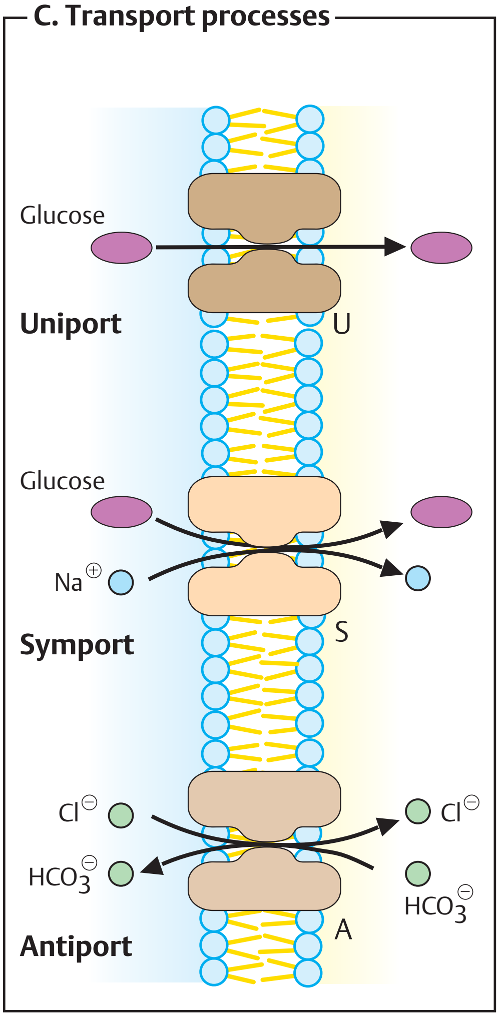
C. Transport processes
Glucose
Uniport
U
Glucose
$Na^{\oplus}$
Symport
S
$Cl^{\ominus}$
$HCO_3^{\ominus}$
$Cl^{\ominus}$
$HCO_3^{\ominus}$
Antiport
A

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Transport proteins

Illustrations **B–D** show transporters whose structure has been determined experimentally or established on analogy with other known structures. They all belong to group III of the α-helical transmembrane proteins (see p. 214).

A. Transport mechanisms ◑

Some cells couple the "pure" transport forms discussed on p. 218—i. e., passive transport (**1**) and active transport (**2**)—and use this mechanism to take up metabolites. In **secondary active transport** (**3**), which is used for example by epithelial cells in the small intestine and kidney to take up glucose and amino acids, there is a **symport** (S) located on the luminal side of the membrane, which takes up the metabolite M together with an Na^+ ion. An ATP-dependent Na^+ transporter (Na^+/K^+ ATPase; see p. 350) on the other side keeps the intracellular Na+ concentration low and thus indirectly drives the uptake of M. Finally, a uniport (U) releases M into the blood.

B. Glucose transporter Glut-1 ○

The glucose transporters (Glut) are a family of related membrane proteins with varying distribution in the organs. Glut-1 and Glut-3 have a relatively high affinity for glucose (K_d = 1 mM). They occur in nearly all cells, and ensure continuous glucose uptake. Glut-2 is found in the liver and pancreas. This form has a lower affinity (K_d = 15–20 mM). The rate of glucose uptake by Glut-2 is therefore strongly dependent on the blood glucose level (normally 4–8 mM). Transport by Glut-4 (K_d = 5 mM), which is mainly expressed in muscle and fat cells, is controlled by insulin, which increases the number of transporters on the cell surface (see p. 388). Glut-5 mediates secondary active resorption of glucose in the intestines and kidney (see **A**).

Glut-1 consists of a single peptide chain that spans the membrane with 12 α-helices of different lengths. The glucose is bound by the peptide loops that project on each side of the membrane.

C. Aquaporin-1 ○

Aquaporins help water to pass through biological membranes. They form hydrophilic pores that allow H_2O molecules, but not hydrated ions or larger molecules, to pass through. Aquaporins are particularly important in the kidney, where they promote the reuptake of water (see p. 328). Aquaporin-2 in the renal collecting ducts is regulated by **antidiuretic hormone** (ADH, vasopressin), which via cAMP leads to shifting of the channels from the ER into the plasma membrane.

Aquaporin-1, shown here, occurs in the proximal tubule and in Henle's loop. It contains eight transmembrane helices with different lengths and orientations. The yellow-colored residues form a narrowing that only H_2O molecules can overcome.

D. Sarcoplasmic Ca^{2+} pump ○

Transport ATPases transport cations—they are "ion pumps." ATPases of the **F type**—e. g., mitochondrial ATP synthase (see p. 142)—use H^+ transport for *ATP synthesis*. Enzymes of the **V type**, using up ATP, "pump" protons into lysosomes and other acidic cell compartments (see p. 234). **P type** transport ATPases are particularly numerous. These are ATP-driven cation transporters that undergo covalent phosphorylation during the transport cycle.

The **Ca^{2+} ATPase** shown also belongs to the P type. In muscle, its task is to pump the Ca^{2+} released into the cytoplasm to trigger muscle contraction back into the sarcoplasmic reticulum (SR; see p. 334). The molecule (**1**) consists of a single peptide chain that is folded into various domains. In the transmembrane part, which is formed by numerous α-helices, there are binding sites for two Ca^{2+} ions (blue) ATP is bound to the cytoplasmic N domain (green).

Four different stages can be distinguished in the enzyme's catalytic cycle (**2**). First, binding of ATP to the N domain leads to the uptake of two Ca^{2+} into the transmembrane part (**a**). Phosphorylation of an aspartate residue in the P domain (**b**) and dissociation of ADP then causes a conformation change that releases the Ca^{2+} ions into the SR (**c**). Finally, dephosphorylation of the aspartate residue restores the initial conditions (**d**).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Transport mechanisms

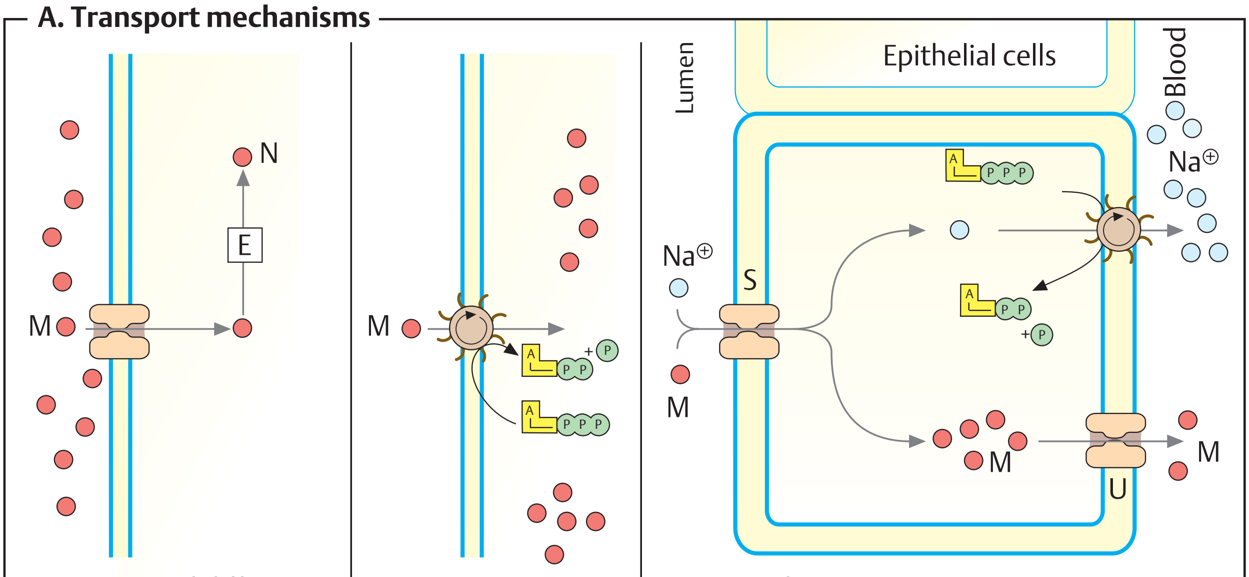
N
E
M
M
Lumen
Epithelial cells
Blood
Na⊕
Na⊕
S
M
M
M
U
A
P P P
P P
P

1. Facilitated diffusion **2.** Active transport **3.** Secondary active transport

B. Glucose transporter Glut-1

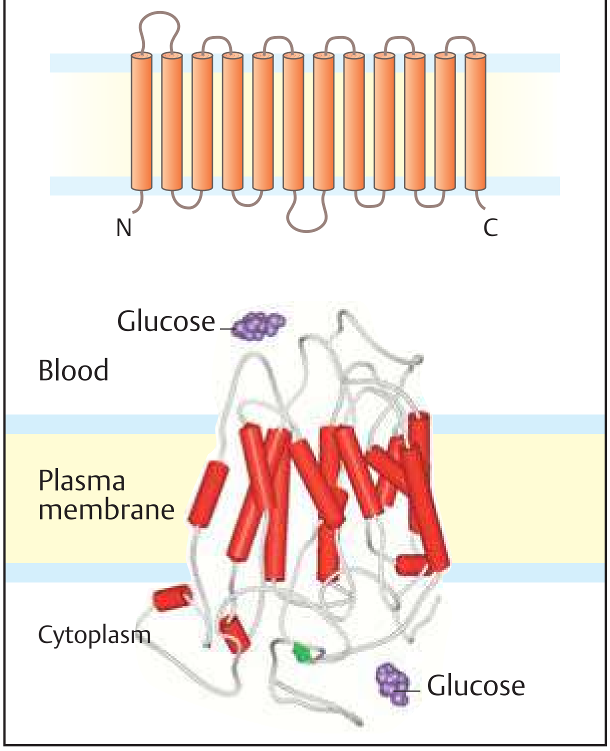
N
C
Glucose
Blood
Plasma membrane
Cytoplasm
Glucose

C. Aquaporin-1

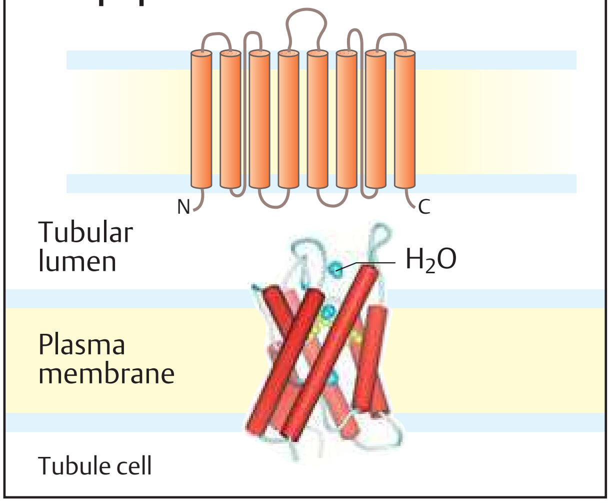
N
C
Tubular lumen
H₂O
Plasma membrane
Tubule cell

D. Sarcoplasmic Ca$^{2\oplus}$ pump

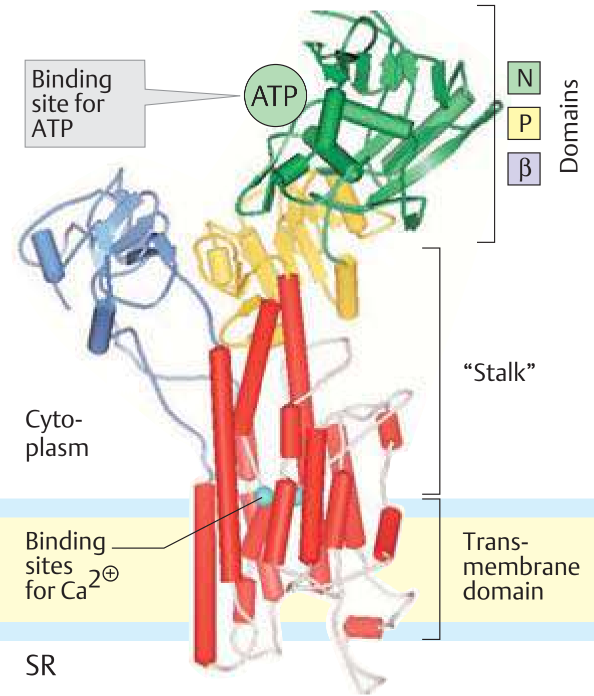
Binding site for ATP
ATP
N
P
β
Domains
"Stalk"
Cyto-plasm
Binding sites for Ca2⊕
Trans-membrane domain
SR

1. Structure

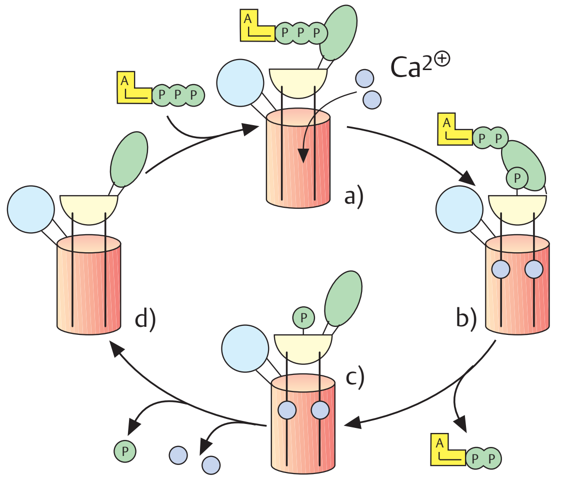
A
P P P
Ca2⊕
a)
b)
c)
d)
P

2. Catalytic cycle

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Ion channels

Ion channels facilitate the diffusion of ions through biological membranes. Some ion channels open and close depending on the membrane potential (**voltage-gated channels, A**) or in response to specific ligands (**ligand-gated channels, B**). Other channels operate passively. In these cases, transport depends only on the concentration gradient (**C**).

A. Voltage-gated Na^+ channel ○

Voltage-gated Na^+ channels play a decisive part in the conduction of electrical impulses in the nervous system (see p. 350). These channels open when the membrane potential in their environment reverses. Due to the high equilibrium potential for Na^+ (see p. 126), an inflow of Na^+ ions takes place, resulting in local **depolarization** of the membrane, which propagates by activation of neighboring voltage-dependent Na^+ channels. A spreading depolarization wave of this type is known as an **action potential** (see p. 350). Externally directed K^+ channels are involved in the repolarization of the membrane. In their functioning, these resemble the much more simply structured K^+ channels shown in **C**. The Ca^{2+} channels that trigger exocytosis of vesicles (see p. 228) are also controlled by the action potential.

The voltage-gated Na^+ channels in higher animals are large complexes made up of several subunits. The α-subunit shown here mediates Na^+ transport. It consists of a very long peptide chain (around 2000 amino acid residues), which is folded into four domains, each with six transmembrane helices (left). The S6 helices of all the domains (blue) together form a centrally located hydrophilic pore which can be made narrow or wide depending on the channel's functional status. The six S4 helices (green) function as voltage sensors.

The current conception of the way in which the opening and closing mechanism functions is shown in a highly simplified form on the right. For the sake of clarity, only one of the four domains (domain IV) is shown. The S4 helices contain several positively charged residues. When the membrane is polarized (**a**), the surplus negative charges on the inner side keep the helix in the membrane. If this attraction is removed as a result of local depolarization, the S4 helices are thought to snap upwards like springs and thus open the central pore (**b**).

B. Nicotinic acetylcholine receptor ○

Many receptors for neurotransmitters function as ligand-gated channels for Na^+, K^+ and/or Ca^{2+} ions (see p. 354). The ones that have been studied in the greatest detail are the nicotinic receptors for **acetylcholine** (see p. 352). These consist of five separate but structurally closely related subunits. Each forms four transmembrane helices, the second of which is involved in the central pore in each case. The type of monomer and its arrangement in the complex is not identical in all receptors of this type. In the neuromuscular junction (see p. 334), the arrangement αβγαδ is found (**1**).

In the interior of the structure, acetylcholine binds to the two α-subunits and thus opens the pore for a short time (1–2 ms). Negatively charged residues are arranged in three groups in a ring shape inside it. They are responsible for the receptor's ion specificity. It is thought that binding of the neurotransmitter changes the position of the subunits in such a way that the pore expands (**3**). The bound acetylcholine dissociates again and is hydrolytically inactivated (see p. 356). The receptor is thus able to function again.

C. K^+ channel in *Streptomyces lividans* ○

The only detailed structures of ion channels established so far are those of potassium channels like that of an outwardly directed K^+ channel in the bacterium *Streptomyces lividans*. It consists of four identical subunits (blue, yellow, green, and red), each of which contains two long α-helices and one shorter one. In the interior of the cell (bottom), the K^+ ions (violet) enter the structure's central channel. Before they are released to the outside, they have to pass through what is known as a "selectivity filter." In this part of the channel, several C=O groups in the peptide chain form a precisely defined opening that is only permeable to non-hydrated K^+ ions.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Voltage-gated Na⊕ channel

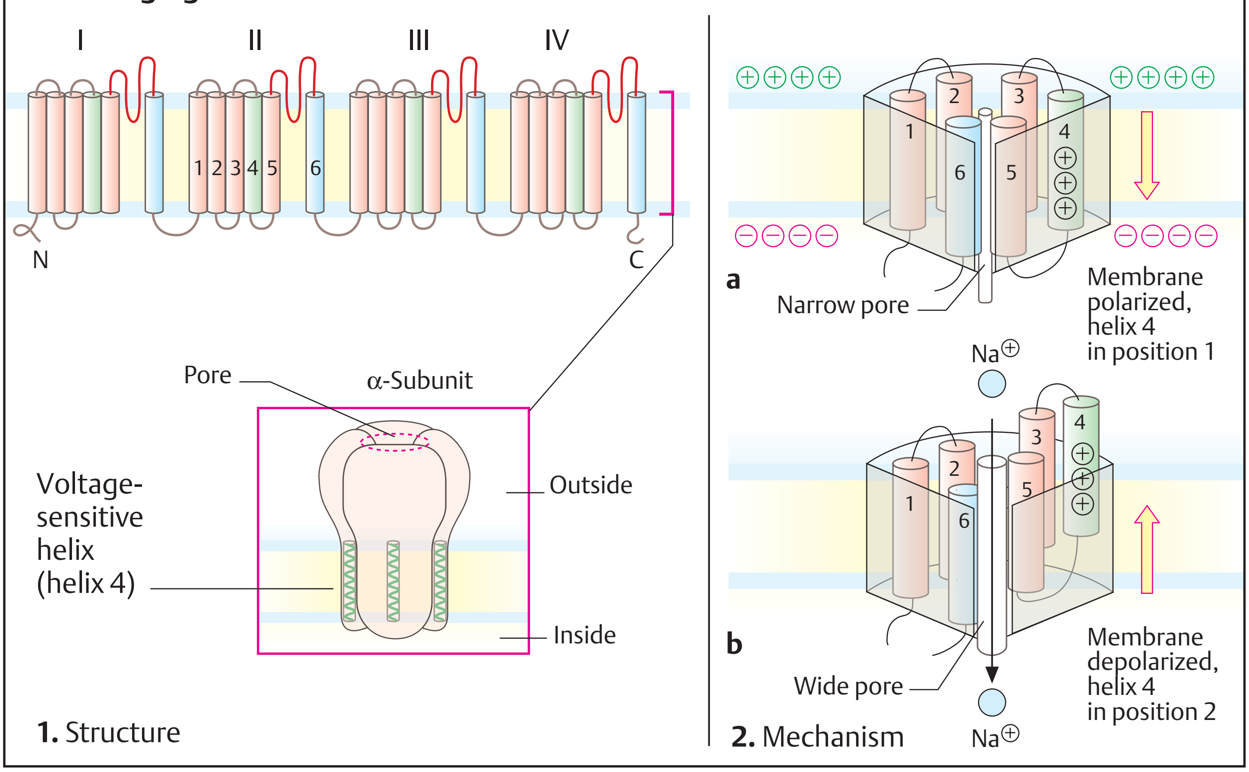
I
II
III
IV
1 2 3 4 5
6
N
C
Pore
α-Subunit
Voltage-sensitive helix (helix 4)
Outside
Inside
1. Structure
⊕⊕⊕⊕
⊕⊕⊕⊕
⊖⊖⊖⊖
⊖⊖⊖⊖
1
2
3
4
5
6
a
Narrow pore
Membrane polarized, helix 4 in position 1
Na⊕
b
Wide pore
Membrane depolarized, helix 4 in position 2
Na⊕
2. Mechanism

B. Nicotinic acetylcholine receptor

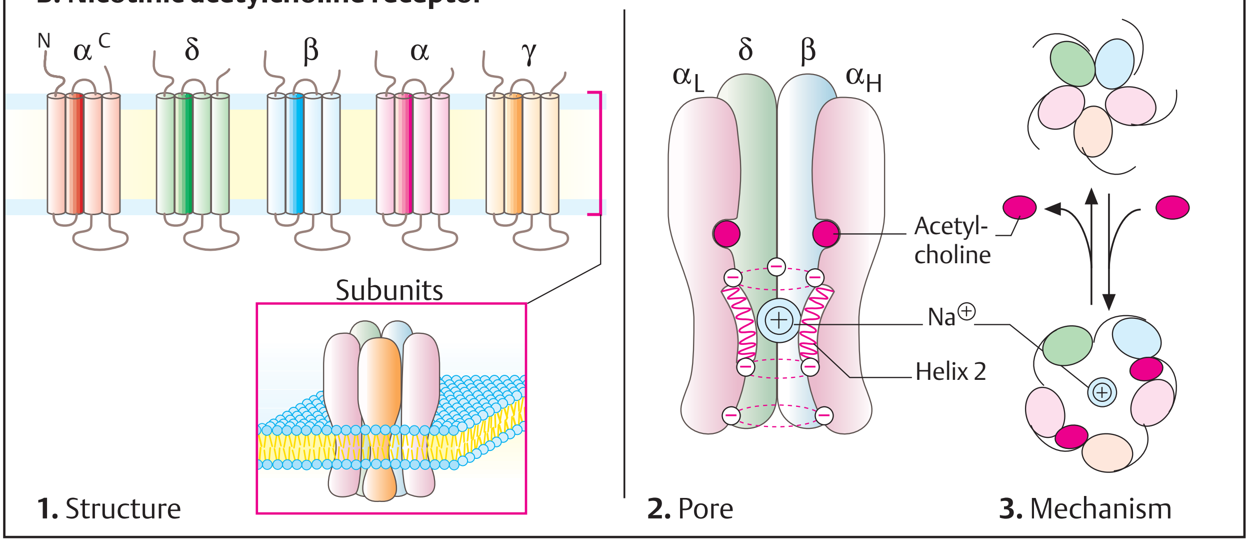
N
α
C
δ
β
α
γ
Subunits
1. Structure
α_L
δ
β
α_H
Acetyl-choline
Na⊕
Helix 2
2. Pore
3. Mechanism

C. K⊕ channel in *Streptomyces lividans*

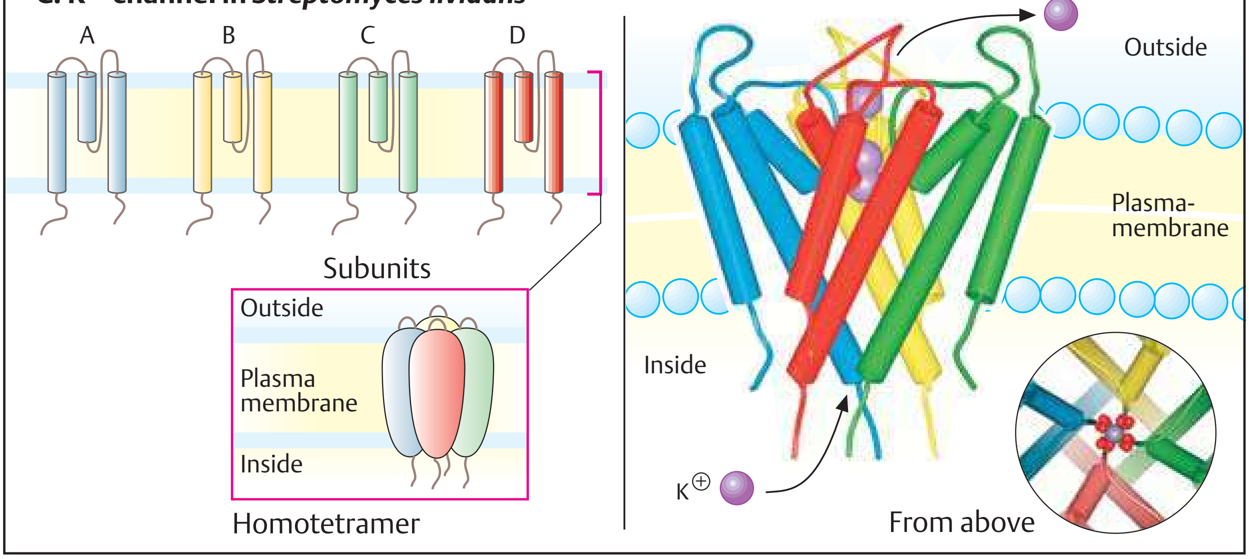
A
B
C
D
Subunits
Outside
Plasma membrane
Inside
Homotetramer
Outside
Plasma-membrane
Inside
K⊕
From above

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Membrane receptors

To receive and pass on chemical or physical signals, cells are equipped with **receptor proteins**. Many of these are integral membrane proteins in the plasma membrane, where they receive signals from their surroundings. Other receptor proteins are located in intercellular membranes. The receptors for lipophilic hormones are among the few that function in a soluble form. They regulate gene transcription in the nucleus (see p. 378).

A. Principle of receptor action ●

Membrane-located receptors can be divided into three parts, which have different tasks. The **receptor domain** reacts specifically to a given signal. Signals of this type can be of a purely physical nature. For example, many organisms react to light. In this way, plants adapt growth and photosynthesis to light conditions, while animals need light receptors for visual processing (**C**; see p. 358). Mechanoreceptors are involved in hearing and in pressure regulation, among other things. Channels that react to action potentials (see p. 350) can be regarded as receptors for electrical impulses.

However, most receptors do not react to physical stimuli, but rather to signal molecules. Receptors for these chemical signals contain binding sites in the receptor domain that are complementary to each ligand. In this respect, they resemble enzymes (see p. 94). As the **effector domain** of the receptor is usually separated by a membrane, a mechanism for **signal transfer** between the domains is needed. Little is yet known regarding this. It is thought that conformation changes in the receptor protein play a decisive part. Some receptors dimerize after binding of the ligand, thereby bringing the effector domains of two molecules into contact (see p. 392).

The way in which the effector works differs from case to case. By binding or interconversion, many receptors activate special **mediator proteins**, which then trigger a signal cascade (signal transduction; see p. 384). Other receptors function as **ion channels**. This is particularly widespread in receptors for neurotransmitters (see p. 354).

B. Insulin receptor ◑

The receptor for the hormone insulin (see p. 76) belongs to the family of **1-helix receptors**.

These molecules span the membrane with only one α-helix. The subunits of the dimeric receptor (red and blue) each consist of two polypeptides (α and β) bound by disulfide bonds. The α-chains together bind the insulin, while the β-chains contain the transmembrane helix and, at the C-terminus, domains with **tyrosine kinase** activity. In the activated state, the kinase domains phosphorylate themselves and also mediator proteins (receptor substrates) that set in motion cascades of further phosphorylations (see pp. 120 and 388).

C. 7-helix receptors ◑

A large group of receptors span the membrane with α-helices seven times. These are known as **7-helix receptors**. Via their effector domains, they bind and activate trimeric proteins, which in turn bind and hydrolyze GTP and are therefore called G proteins. Most G proteins, in turn, activate or inhibit enzymes that create secondary signaling molecules (**second messengers**; see p. 386). Other G proteins regulate ion channels. The illustration shows the complex of the light receptor **rhodopsin**, with the associated G protein **transducin** (see p. 358). The GTP-binding α-subunit (green) and the γ-subunit (violet) of transducin are anchored in the membrane via lipids (see p. 214). The β-subunit is shown in detail on p. 72.

D. T-cell receptor ◑

The cells of the immune system communicate with each other particularly intensively. The **T-cell receptor** plays a central role in the activation of T lymphocytes (see p. 296). The cell at the top has been infected with a virus, and it indicates this by presenting a viral peptide (violet) with the help of a class I **MHC protein** (yellow and green). The combination of the two molecules is recognized by the dimeric T-cell receptor (blue) and converted into a signal that activates the T cell (bottom) and thereby enhances the immune response to the virus.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Principle of receptor action

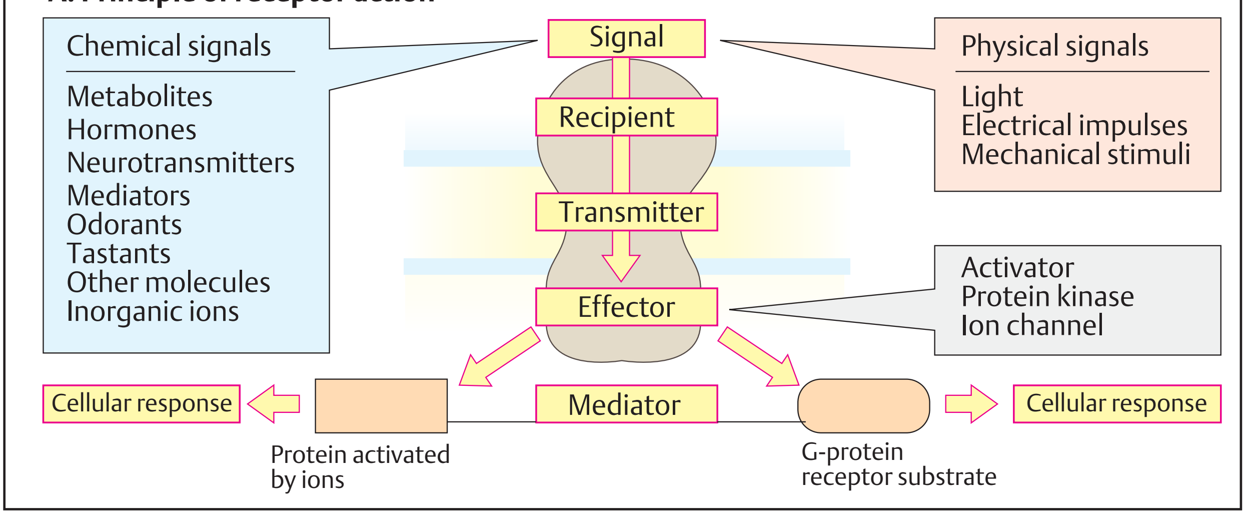
Chemical signals
Metabolites
Hormones
Neurotransmitters
Mediators
Odorants
Tastants
Other molecules
Inorganic ions
Signal
Recipient
Transmitter
Effector
Physical signals
Light
Electrical impulses
Mechanical stimuli
Activator
Protein kinase
Ion channel
Cellular response
Mediator
Cellular response
Protein activated by ions
G-protein receptor substrate

B. Insulin receptor

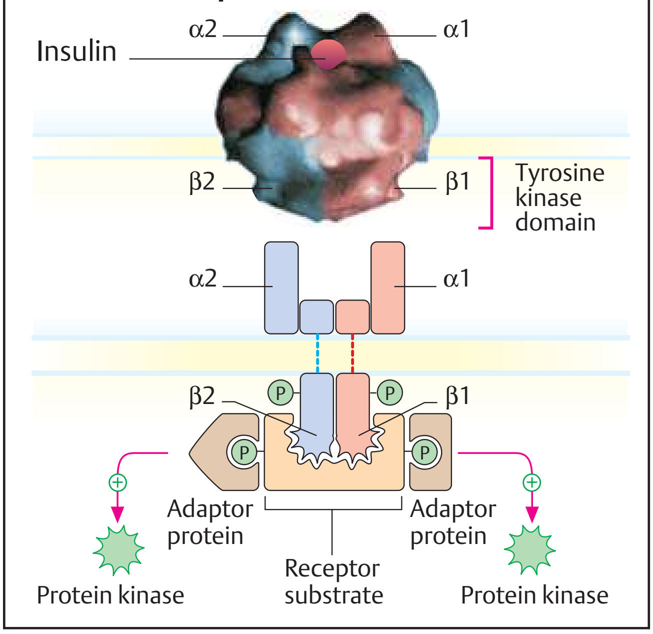
Insulin
α2
α1
β2
β1
Tyrosine kinase domain
α2
α1
β2
β1
P
P
P
P
Adaptor protein
Adaptor protein
Receptor substrate
Protein kinase
Protein kinase

C. 7-Helix receptors

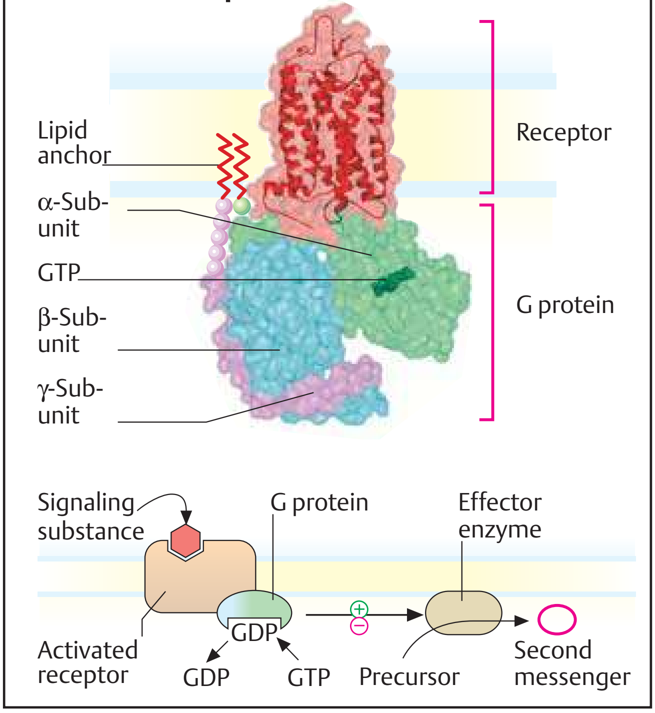
Lipid anchor
α-Sub-unit
GTP
β-Sub-unit
γ-Sub-unit
Receptor
G protein
Signaling substance
G protein
Effector enzyme
Activated receptor
GDP
GDP
GTP
Precursor
Second messenger

D. T-cell receptor

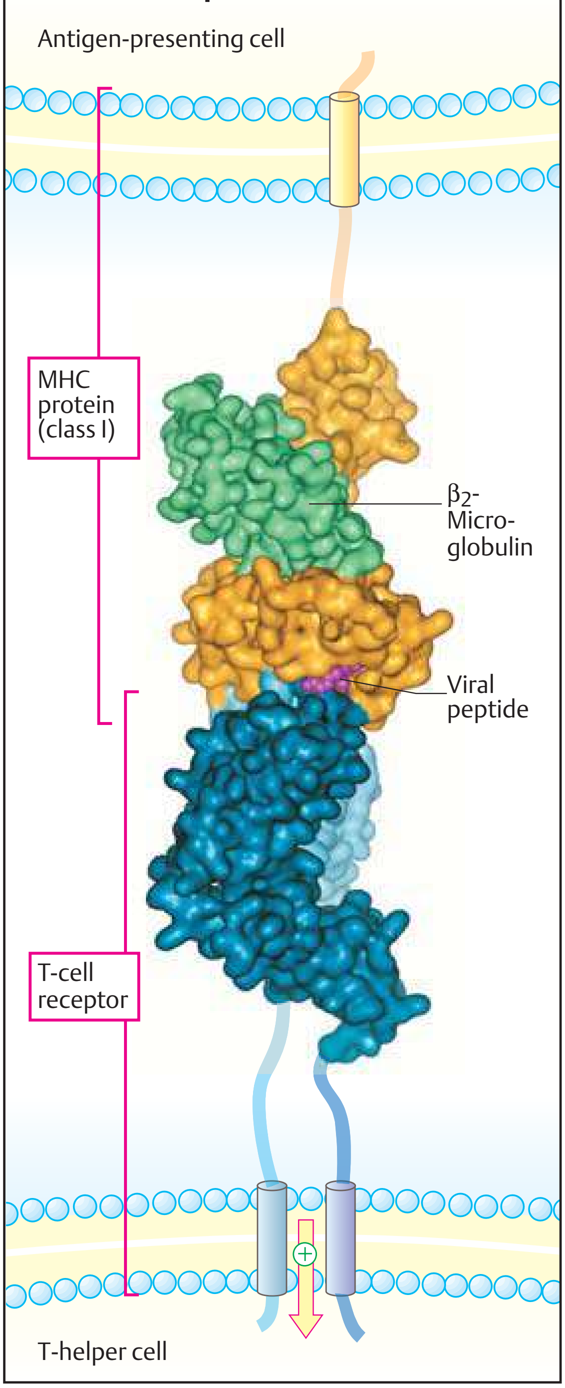
Antigen-presenting cell
MHC protein (class I)
β2-Micro-globulin
Viral peptide
T-cell receptor
T-helper cell

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

ER: structure and function

The endoplasmic reticulum (ER) is an extensive closed membrane system consisting of tubular and saccular structures. In the area of the nucleus, the ER turns into the external nuclear membrane. Morphologically, a distinction is made between the **rough ER** (rER) and the **smooth ER** (sER). Large numbers of ribosomes are found on the membranes of the rER, which are lacking on the sER. On the other hand, the sER is rich in membrane-bound **enzymes**, which catalyze partial reactions in the lipid metabolism as well as biotransformations.

A. Rough endoplasmic reticulum and the Golgi apparatus ◑

The **rER** (**1**) is a site of active *protein biosynthesis.* This is where proteins destined for membranes, lysosomes, and export from the cell are synthesized. The remaining proteins are produced in the cytoplasm on ribosomes that are not bound to membranes.

Proteins synthesized at the rER (**1**) are folded and modified after translation (protein maturation; see p. 230). They remain either in the rER as membrane proteins, or pass with the help of transport vesicles (**2**) to the Golgi apparatus (**3**). Transport vesicles are formed by budding from existing membranes, and they disappear again by fusing with them (see p. 228).

The **Golgi apparatus** (**3**) is a complex network, also enclosed, consisting of flattened membrane saccules ("cisterns"), which are stacked on top of each other in layers. Proteins mature here and are sorted and packed. A distinction is made between the *cis, medial,* and *trans* Golgi regions, as well as a *trans* Golgi network (tGN). The *post–translational modification of proteins,* which starts in the ER, continues in these sections.

From the Golgi apparatus, the proteins are transported by vesicles to various targets in the cells—e. g., to lysosomes (**4**), the plasma membrane (**6**), and secretory vesicles (**5**) that release their contents into the extracellular space by fusion with the plasma membrane (**exocytosis**; see p. 228). Protein transport can either proceed continuously (*constitutive*), or it can be *regulated* by chemical signals. The decision regarding which pathway a protein will take and whether its transport will be constitutive or regulated depends on the signal sequences or signal structures that proteins carry with them like address labels (see p. 228). In addition to proteins, the Golgi apparatus also transports membrane lipids to their targets.

B. Smooth endoplasmic reticulum ◑

Regions of the ER that have no bound ribosomes are known as the **smooth endoplasmic reticulum** (sER). In most cells, the proportion represented by the sER is small. A marked sER is seen in cells that have an active lipid metabolism, such as hepatocytes and Leydig cells. The sER is usually made up of branching, closed tubules.

Membrane-located enzymes in the sER catalyze **lipid synthesis**. Phospholipid synthesis (see p. 170) is located in the sER, for example, and several steps in cholesterol biosynthesis (see p. 172) also take place there. In endocrine cells that form *steroid hormones,* a large proportion of the reaction steps involved also take place in the sER (see p. 376).

In the liver's hepatocytes, the proportion represented by the sER is particularly high. It contains enzymes that catalyze so-called **biotransformations**. These are reactions in which apolar foreign substances, as well as endogenous substances—e. g., steroid hormones—are chemically altered in order to inactivate them and/or prepare them for conjugation with polar substances (phase I reactions; see p. 316). Numerous *cytochrome P450 enzymes* are involved in these conversions (see p. 318) and can therefore be regarded as the major molecules of the sER.

The sER also functions as an intracellular **calcium store**, which normally keeps the Ca^{2+} level in the cytoplasm low. This function is particularly marked in the *sarcoplasmic reticulum*, a specialized form of the sER in muscle cells (see p. 334). For release and uptake of Ca^{2+}, the membranes of the sER contain signal-controlled Ca^{2+} channels and energy-dependent Ca^{2+} ATPases (see p. 220). In the lumen of the sER, the high Ca^{2+} concentration is buffered by Ca^{2+}-binding proteins.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Rough endoplasmic reticulum and Golgi apparatus

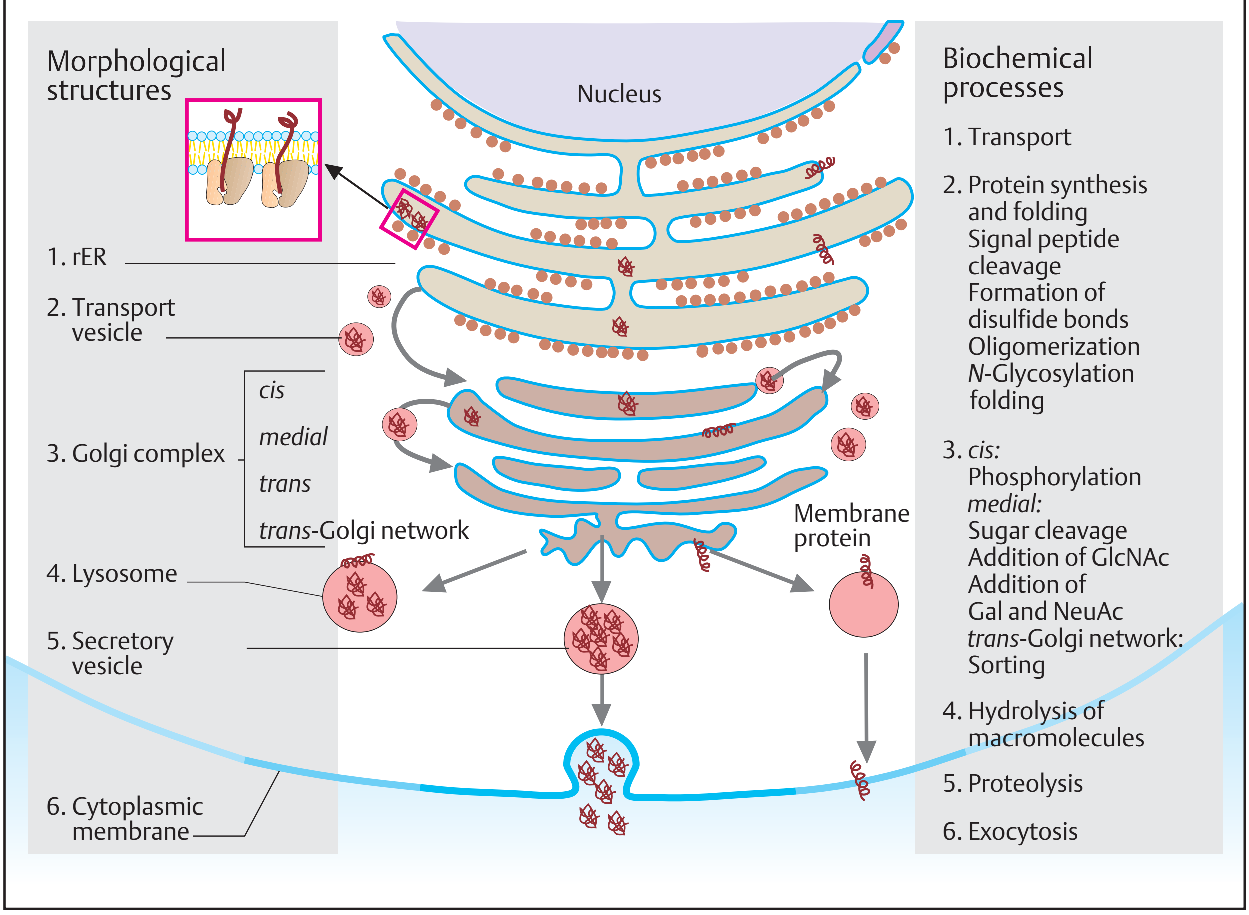
Morphological structures
1. rER
2. Transport vesicle
3. Golgi complex
cis
medial
trans
trans-Golgi network
4. Lysosome
5. Secretory vesicle
6. Cytoplasmic membrane
Nucleus
Membrane protein
Biochemical processes
1. Transport
2. Protein synthesis and folding
Signal peptide cleavage
Formation of disulfide bonds
Oligomerization
N-Glycosylation
folding
3. *cis:*
Phosphorylation
medial:
Sugar cleavage
Addition of GlcNAc
Addition of Gal and NeuAc
trans-Golgi network:
Sorting
4. Hydrolysis of macromolecules
5. Proteolysis
6. Exocytosis

B. Smooth endoplasmic reticulum

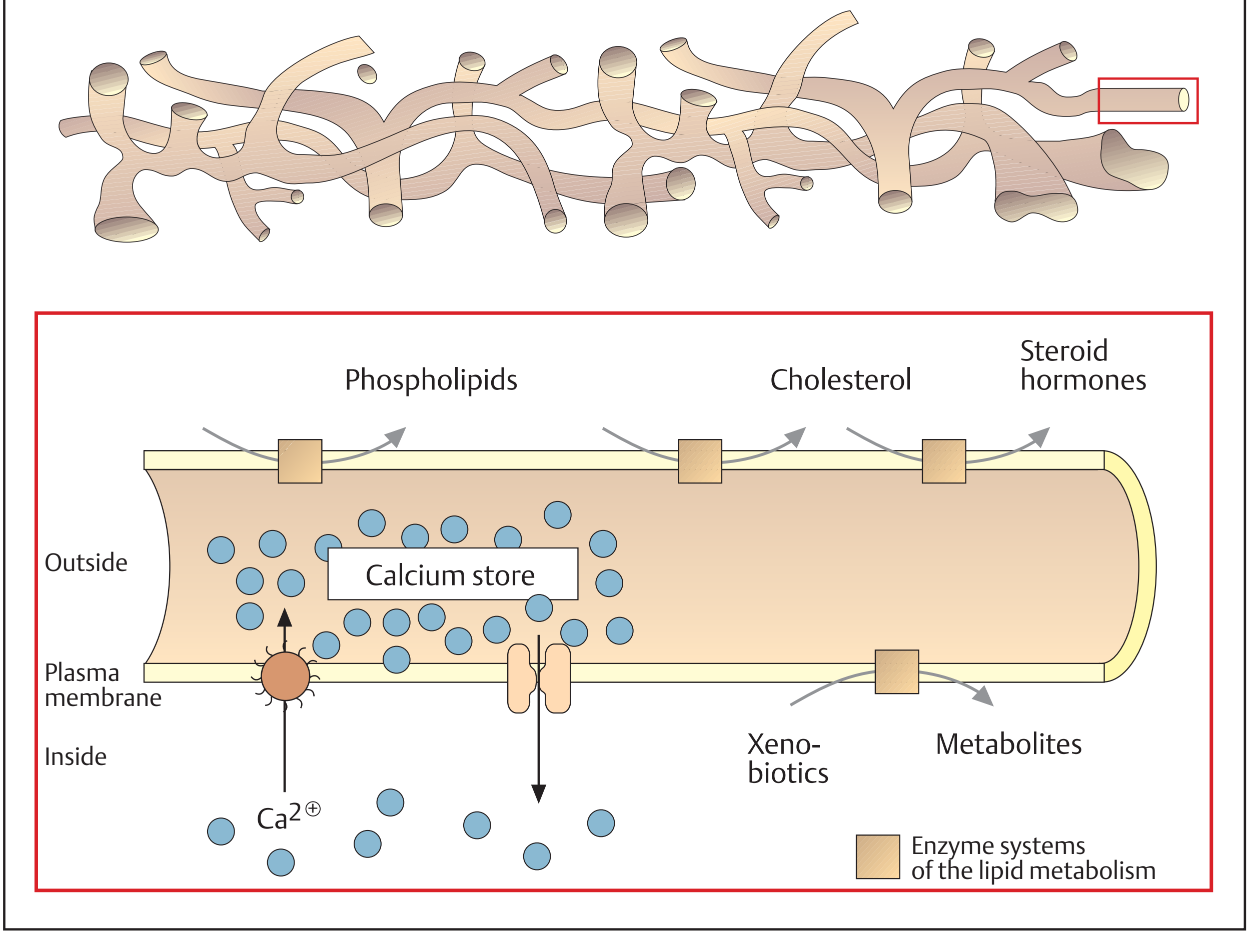
Phospholipids
Cholesterol
Steroid hormones
Outside
Calcium store
Plasma membrane
Inside
Ca2⊕
Xeno-biotics
Metabolites
Enzyme systems of the lipid metabolism

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Protein sorting

A. Protein sorting ◑

The biosynthesis of all proteins starts on free ribosomes (top). However, the paths that the proteins follow soon diverge, depending on which target they are destined for. Proteins that carry a *signal peptide for the ER* (**1**) follow the *secretory pathway* (right). Proteins that do not have this signal follow the *cytoplasmic pathway* (left).

Secretory pathway. Ribosomes that synthesize a protein with a signal peptide for the ER settle on the ER (see p. 228). The peptide chain is transferred into the lumen of the rER. The presence or absence of other signal sequences and signal regions determines the subsequent transport pathway.

Proteins that have *stop-transfer sequences* (**4**) remain as integral membrane proteins in the ER membrane. They then pass into other membranes via vesicular transport (see p. 226). From the rER, their pathway then leads to the Golgi apparatus and then on to the plasma membrane. Proteins destined to remain in the rER—e. g., enzymes—find their way back from the Golgi apparatus to the rER with the help of a *retention signal* (**2**). Other proteins move from the Golgi apparatus to the lysosomes (**3**; see p. 234), to the cell membrane (integral membrane proteins or constitutive exocytosis), or are transported out of the cell (**9;** signal-regulated exocytosis) by secretory vesicles (**8**).

Cytoplasmic pathway. Proteins that do not have a signal peptide for the ER are synthesized in the cytoplasm on free ribosomes, and remain in that compartment. Special signals mediate further transport into the mitochondria (**5**; see p. 232), the nucleus (**6**; see p. 208) or peroxisomes (**7**).

B. Translocation signals ○

Signal peptides are short sections at the N or C terminus, or within the peptide chain. Areas on the protein surface that are formed by various sections of the chain or by various chains are known as **signal regions**. Signal peptides and signal regions are *structural signals* that are usually recognized by *receptors* on organelles (see **A**). They move the proteins, with the help of additional proteins, into the organelles (*selective protein transfer*). Structural signals can also activate *enzymes* that modify the proteins and thereby determine their subsequent fate. Examples include lysosomal proteins (see p. 234) and membrane proteins with lipid anchors (see p. 214).

After they have been used, signal peptides at the N terminus are cleaved off by specific hydrolases (symbol: scissors). In proteins that contain several successive signal sequences, this process can expose the subsequent signals. By contrast, signal peptides that have to be read several times are not cleaved.

C. Exocytosis ○

Exocytosis is a term referring to processes that allow cells to expel substances (e. g., hormones or neurotransmitters) quickly and in large quantities. Using a complex protein machinery, secretory vesicles fuse completely or partially with the plasma membrane and release their contents. Exocytosis is usually *regulated* by chemical or electrical signals. As an example, the mechanism by which neurotransmitters are released from synapses (see p. 348) is shown here, although only the most important proteins are indicated.

The decisive element in exocytosis is the interaction between proteins known as SNAREs that are located on the vesicular membrane (v-SNAREs) and on the plasma membrane (t-SNAREs). In the resting state (**1**), the v-SNARE *synaptobrevin* is blocked by the vesicular protein *synaptotagmin.* When an action potential reaches the presynaptic membrane, voltage-gated Ca^{2+} channels open (see p. 348). Ca^{2+} flows in and triggers the machinery by conformational changes in proteins. Contact takes place between synaptobrevin and the t-SNARE *synaptotaxin* (**2**). Additional proteins known as SNAPs bind to the SNARE complex and allow fusion between the vesicle and the plasma membrane (**3**). The process is supported by the hydrolysis of GTP by the auxiliary protein *Rab.*

The toxin of the bacterium *Clostridium botulinum,* one of the most poisonous substances there is, destroys components of the exocytosis machinery in synapses through enzymatic hydrolysis, and in this way blocks neurotransmission.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Protein sorting

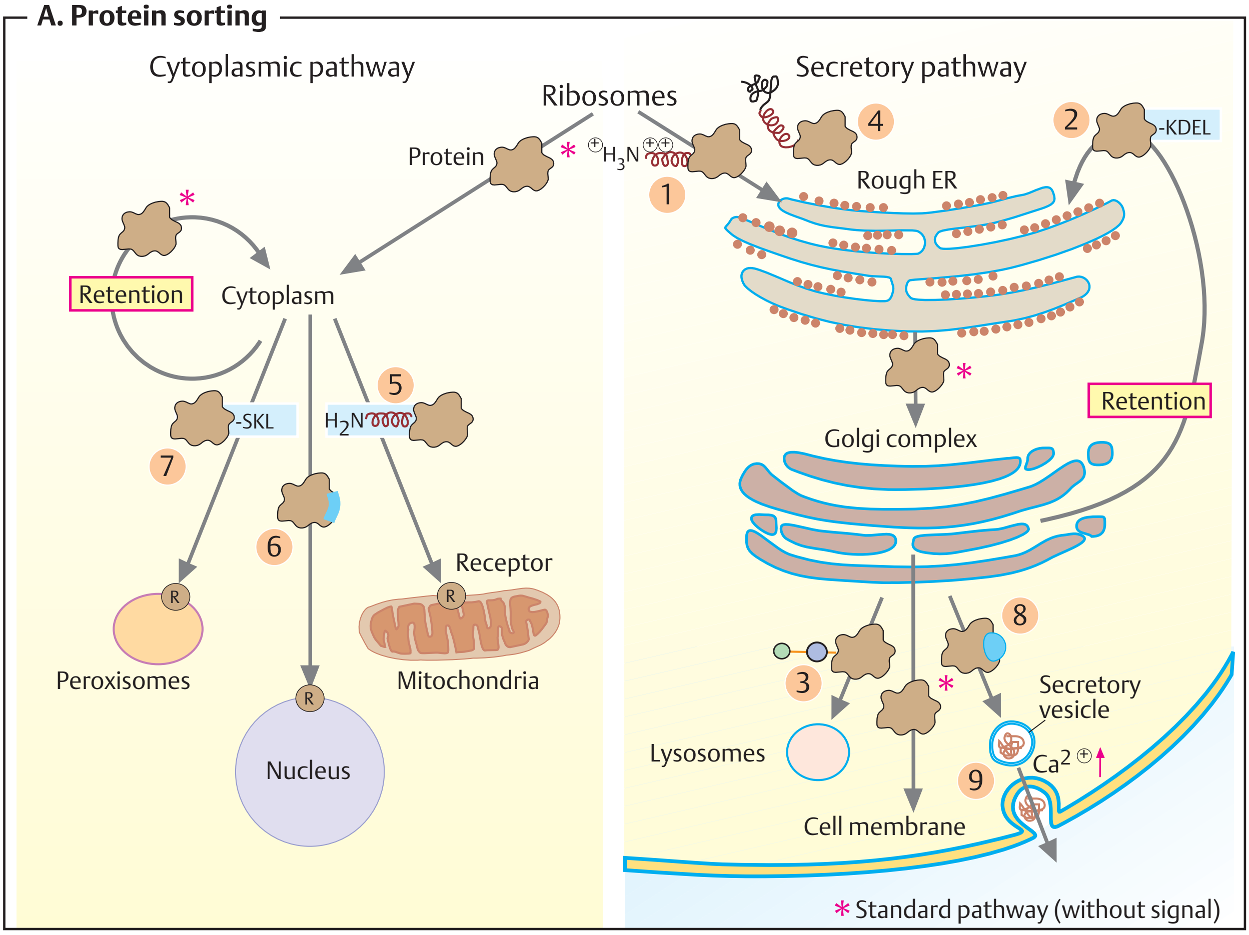
Cytoplasmic pathway
Ribosomes
Secretory pathway
Protein
*
$^{\oplus}H_3N^{\oplus\oplus}$
1
4
2
-KDEL
Rough ER
*
Retention
Cytoplasm
Retention
*
Golgi complex
-SKL
7
5
H_2N
6
Receptor
R
Peroxisomes
R
Mitochondria
R
Nucleus
8
3
*
Secretory vesicle
Lysosomes
9
$Ca^{2\oplus}$
Cell membrane
* Standard pathway (without signal)

B. Translocation signals

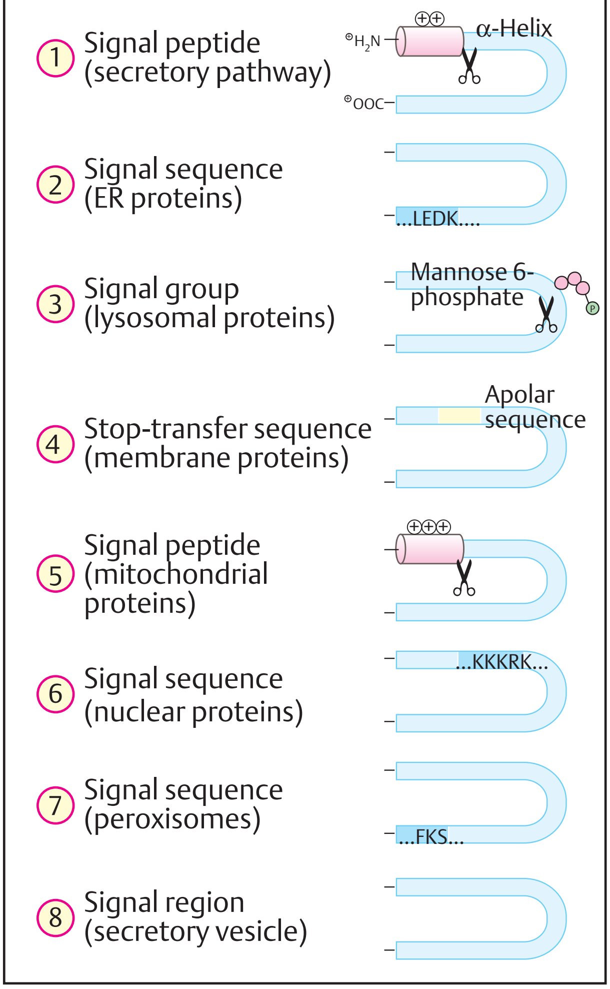
① Signal peptide (secretory pathway)
$^{\oplus}H_2N$
⊕⊕
α-Helix
$^{\ominus}OOC$
② Signal sequence (ER proteins)
...LEDK....
③ Signal group (lysosomal proteins)
Mannose 6-phosphate
P
④ Stop-transfer sequence (membrane proteins)
Apolar sequence
⑤ Signal peptide (mitochondrial proteins)
⊕⊕⊕
⑥ Signal sequence (nuclear proteins)
...KKKRK...
⑦ Signal sequence (peroxisomes)
...FKS...
⑧ Signal region (secretory vesicle)

C. Exocytosis

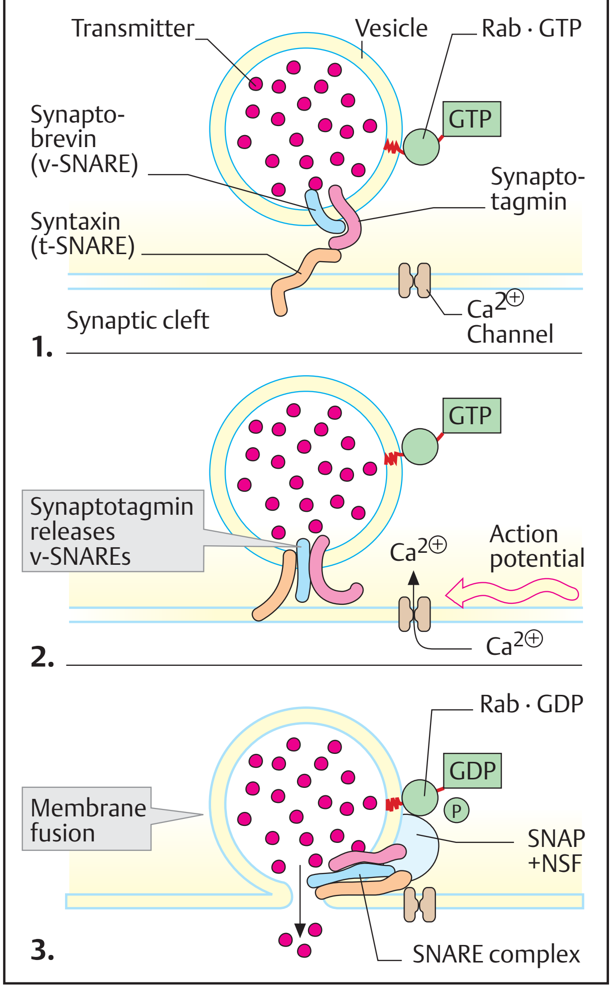
Transmitter
Vesicle
Rab · GTP
GTP
Synapto-brevin (v-SNARE)
Synapto-tagmin
Syntaxin (t-SNARE)
$Ca^{2\oplus}$ Channel
Synaptic cleft
1.
GTP
Synaptotagmin releases v-SNAREs
$Ca^{2\oplus}$
Action potential
$Ca^{2\oplus}$
2.
Rab · GDP
GDP
P
Membrane fusion
SNAP +NSF
SNARE complex
3.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Protein synthesis and maturation

A. Protein synthesis in the rER ◑

With all proteins, protein biosynthesis (**Translation**; for details, see p. 250) starts on free ribosomes in the cytoplasm (**1**). Proteins that are exported out of the cell or into lysosomes, and membrane proteins of the ER and the plasma membrane, carry a **signal peptide for the ER** at their N-terminus. This is a section of 15–60 amino acids in which one or two strongly basic residues (Lys, Arg) near the N-terminus are followed by a strongly hydrophobic sequence of 10–15 residues (see p. 228).

As soon as the signal peptide (red) appears on the surface of the ribosome (**2**), an RNA-containing *signal recognition particle* (**SRP**, green) binds to the sequence and initially interrupts translation (**3**). The SRP then binds to an **SRP receptor** in the rER membrane, and in this way attaches the ribosome to the ER (**4**). After this, the SRP dissociates from the signal peptide and from the SRP receptor and is available again for step 3. This endergonic process is driven by GTP hydrolysis (**5**). Translation now resumes. The remainder of the protein, still unfolded, is gradually introduced into a channel (the **translocon**) in the lumen of the rER (**6**), where a *signal peptidase* located in the inner ER membrane cleaves the signal peptide while translation is still taking place (**7**). This converts the **preprotein** into a **proprotein**, from which the mature protein finally arises after additional post-translational modifications (**8**) in the ER and in the Golgi apparatus.

If the growing polypeptide contains a **stop-transfer signal** (see p. 228), then this hydrophobic section of the chain remains stuck in the membrane outside the translocon, and an *integral membrane protein* arises. In the course of translation, an additional signal sequence can re-start the transfer of the chain through the translocon. Several repetitions of this process produce integral membrane proteins with several transmembrane helices (see p. 214).

B. Protein glycosylation ○

Most extracellular proteins contain covalently bound oligosaccharide residues. For example, all plasma proteins with the exception of albumin are glycosylated. Together with glycolipids, numerous **glycoproteins** on the cell surface form the **glycocalyx**. Inside the ER, the carbohydrate parts of the glycoproteins are cotranslationally transferred to the growing chain, and are then converted into their final form while passing through the ER and Golgi apparatus.

N-bound oligosaccharides (see p. 44) are always bound to the acid-amide group of asparagine residues. If a **glycosylation sequence** (–Asn–X–Ser(Thr)–, where X can be any amino acid) appears in the growing peptide chain, then a *transglycosylase* in the ER membrane [1] transfers a previously produced **core oligosaccharide** consisting of 14 hexose residues *en-bloc* from the carrier molecule **dolichol diphosphate** to the peptide.

Dolichol is a long-chain isoprenoid (see p. 52) consisting of 10–20 isoprene units, which is embedded in the ER membrane. A hydroxyl group at the end of the molecule is bound to diphosphate, on which the nuclear oligosaccharide is built up in an extended reaction sequence (not shown here in detail). The core structure consists of two residues of N-acetylglucosamine (GlcNAc), a branched group of nine mannose residues (Man) and three terminal glucose resides (Glc).

As the proprotein passes through the ER, *glycosidases* [2] remove the glucose residues completely and the mannoses partially ("trimming"), thereby producing the **mannose-rich type** of oligosaccharide residues. Subsequently, various *glycosyltransferases* [3] transfer additional monosaccharides (e.g., GlcNAc, galactose, fucose, and N-acetylneuraminic acid; see p. 38) to the mannose-rich intermediate and thereby produce the **complex type** of oligosaccharide. The structure of the final oligosaccharide depends on the type and activity of the glycosyltransferases present in the ER of the cell concerned, and is therefore genetically determined (although indirectly).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Protein synthesis in the rough endoplasmic reticulum

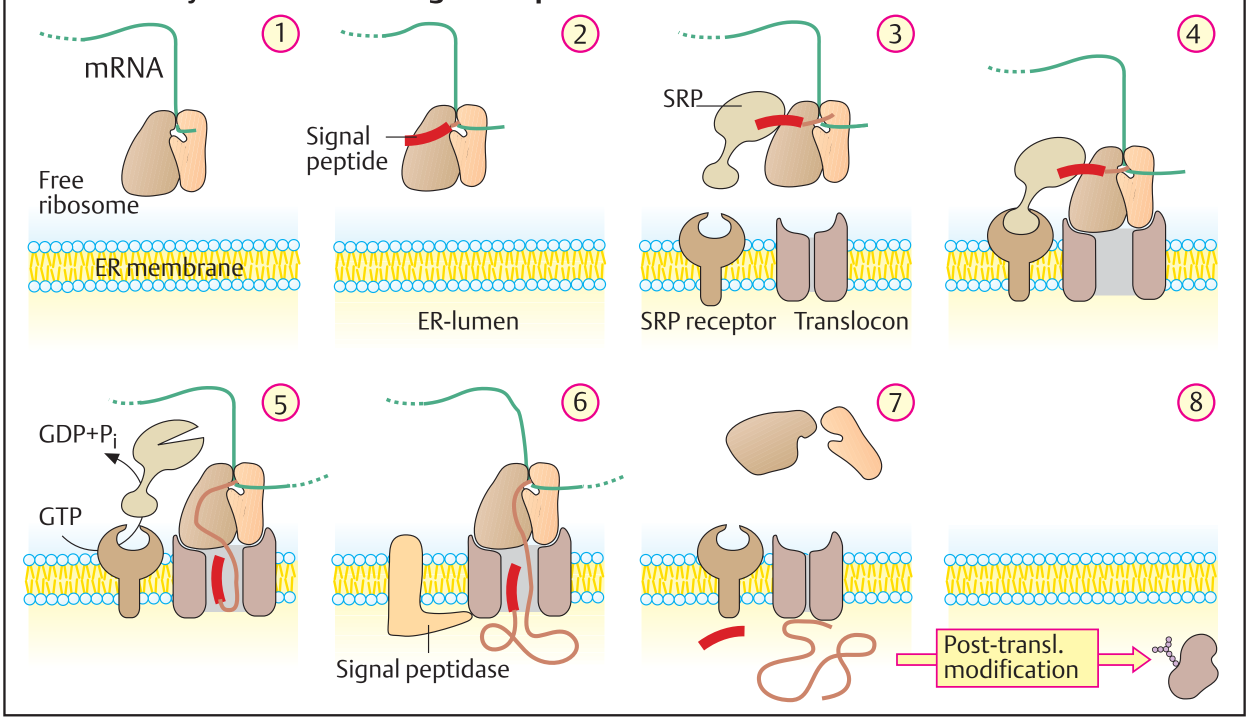
1
2
3
4
mRNA
Free ribosome
ER membrane
Signal peptide
ER-lumen
SRP
SRP receptor
Translocon
5
6
7
8
GDP+P_i
GTP
Signal peptidase
Post-transl. modification

B. Protein glycosylation

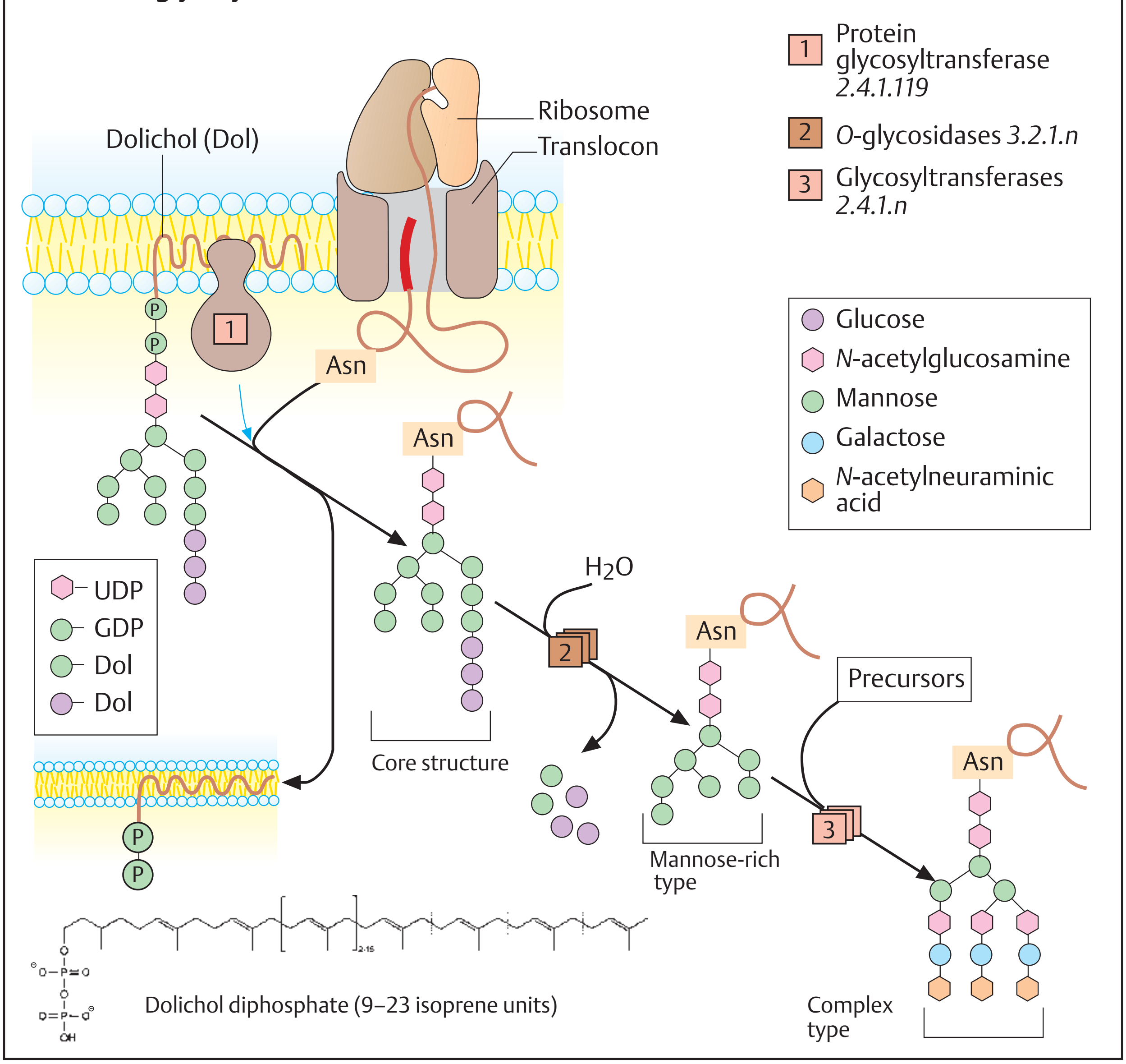
1 Protein glycosyltransferase *2.4.1.119*
2 *O*-glycosidases *3.2.1.n*
3 Glycosyltransferases *2.4.1.n*
Ribosome
Translocon
Dolichol (Dol)
P
P
1
Asn
Glucose
N-acetylglucosamine
Mannose
Galactose
N-acetylneuraminic acid
UDP
GDP
Dol
Dol
Asn
Core structure
H_2O
2
Asn
Precursors
Mannose-rich type
3
Asn
P
P
Complex type
Dolichol diphosphate (9–23 isoprene units)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Protein maturation

After translation, proteins destined for the secretory pathway (see p. 228) first have to fold into their native conformation within the rER (see p. 230). During this process they are supported by various *auxiliary proteins.*

A. Protein folding in the rER ●

To prevent incorrect folding of the growing protein during protein biosynthesis, **chaperones** (see **B**) in the lumen of the rER bind to the peptide chain and stabilize it until translation has been completed. Binding protein (**BiP**) is an important chaperone in the ER.

Many secretory proteins—e. g., pancreatic ribonuclease (RNAse; see p. 74)—contain several disulfide bonds that are only formed oxidatively from SH groups after translation. The eight cysteine residues of the RNAse can in principle form 105 different pairings, but only the combination of the four disulfide bonds shown on p. 75 provides active enzyme. Incorrect pairings can block further folding or lead to unstable or insoluble conformations. The enzyme *protein disulfide isomerase* [1] accelerates the equilibration between paired and unpaired cysteine residues, so that incorrect pairs can be quickly split before the protein finds its final conformation.

Most peptide bonds in proteins take on the *trans* conformation (see p. 66). Only bonds with proline residues (–X–Pro–) can be present in both *cis* and *trans* forms.

In the protein's native conformation, every X–Pro bond has to have the correct conformation (*cis* or *trans*). As the uncatalyzed transition between the two forms is very slow, there is a *proline* cis–trans *isomerase* [2] in the ER that accelerates the conversion.

B. Chaperones and chaperonins ●

Most proteins fold spontaneously into their native conformation, even in the test tube. In the cell, where there are very high concentrations of proteins (around 350 g L^{-1}), this is more difficult. In the unfolded state, the apolar regions of the peptide chain (yellow) tend to aggregate—due to the hydrophobic effect (see p. 28)—with other proteins or with each other to form insoluble products (2). In addition, unfolded proteins are susceptible to proteinases. To protect partly folded proteins, there are auxiliary proteins called **chaperones** because they guard immature proteins against damaging contacts. Chaperones are formed increasingly during temperature stress and are therefore also known as **heat-shock proteins** (hsp). Several classes of hsp are distinguished. Chaperones of the hsp70 type (Dna K in bacteria) are common, as are type hsp60 **chaperonins** (GroEL/ES in bacteria). Class hsp90 chaperones have special tasks (see p. 378).

While small proteins can often reach their native conformation without any help (1), larger molecules require hsp70 proteins for protection against aggregation which bind as monomers and can dissociate again, dependent on ATP (3). By contrast, type hsp60 chaperonins form large, barrel-shaped complexes with 14 subunits in which proteins can fold independently while shielded from their environment (4). The function of hsp60 has been investigated in detail in the bacterial chaperonin **GroEL** (right). The barrel has two chambers, which are closed with a lid (**GroES**) during folding of the guest protein. Driven by ATP hydrolysis, the chambers open and close alternately—i. e., the release of the fully folded protein from one chamber is coupled to the uptake of an unfolded peptide in the second chamber.

C. Protein import in mitochondria ●

Class hsp70 chaperones are also needed for translocation of nuclear-coded proteins from the cytoplasm into the mitochondria (see p. 228). As two membranes have to be crossed to reach the matrix, there are two translocator complexes: **TOM (**"transport outer membrane") and **TIM** ("transport inner membrane"). For transport, proteins are unfolded in the cytoplasm and protected by hsp70. TOM recognizes the positively charged signal sequence at the protein's N terminus (see p. 228) and with the help of the membrane potential threads the chains through the central pores of the two complexes. Inside TIM, further hsp70 molecules bind and pull the chain completely into the matrix. As with import into the ER, the signal peptide is proteolytically removed by a signal peptidase during translocation.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Protein folding in the rER

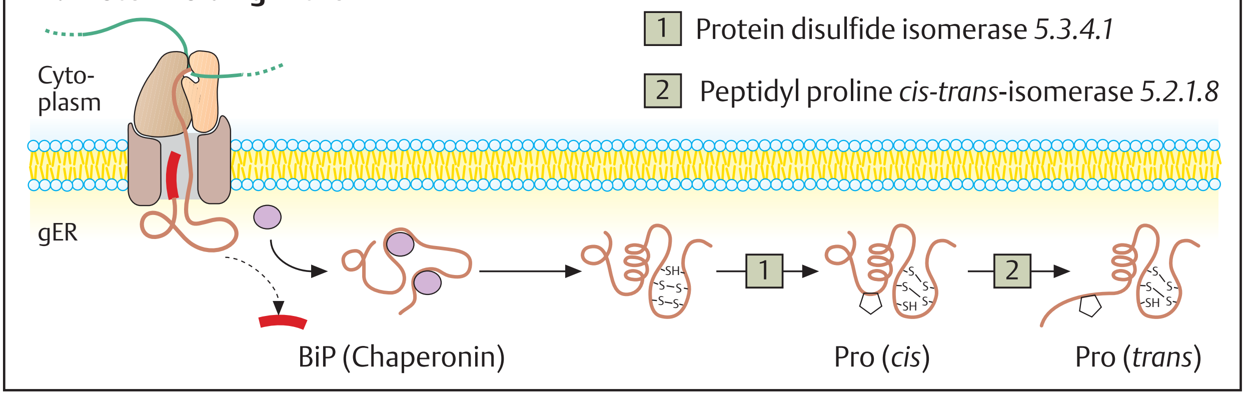
1 Protein disulfide isomerase 5.3.4.1
2 Peptidyl proline *cis-trans*-isomerase 5.2.1.8
Cyto-
plasm
gER
-SH
-S-S-
-S-S-
1
2
BiP (Chaperonin)
Pro (*cis*)
Pro (*trans*)

B. Chaperones and chaperonins

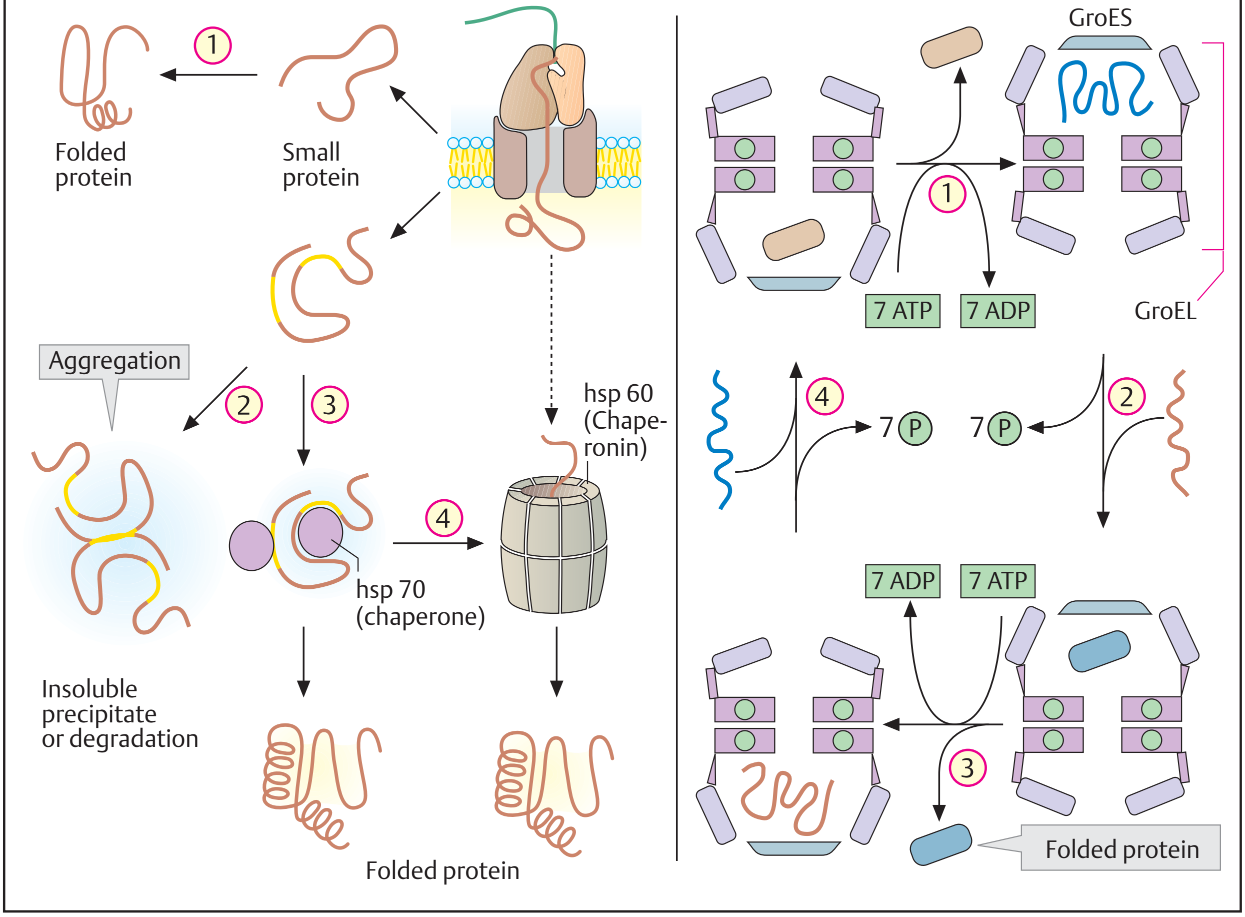
Folded
protein
1
Small
protein
Aggregation
2
3
Insoluble
precipitate
or degradation
4
hsp 70
(chaperone)
hsp 60
(Chape-
ronin)
Folded protein
GroES
GroEL
1
7 ATP
7 ADP
4
7 P
7 P
2
7 ADP
7 ATP
3
Folded protein

C. Protein import in mitochondria

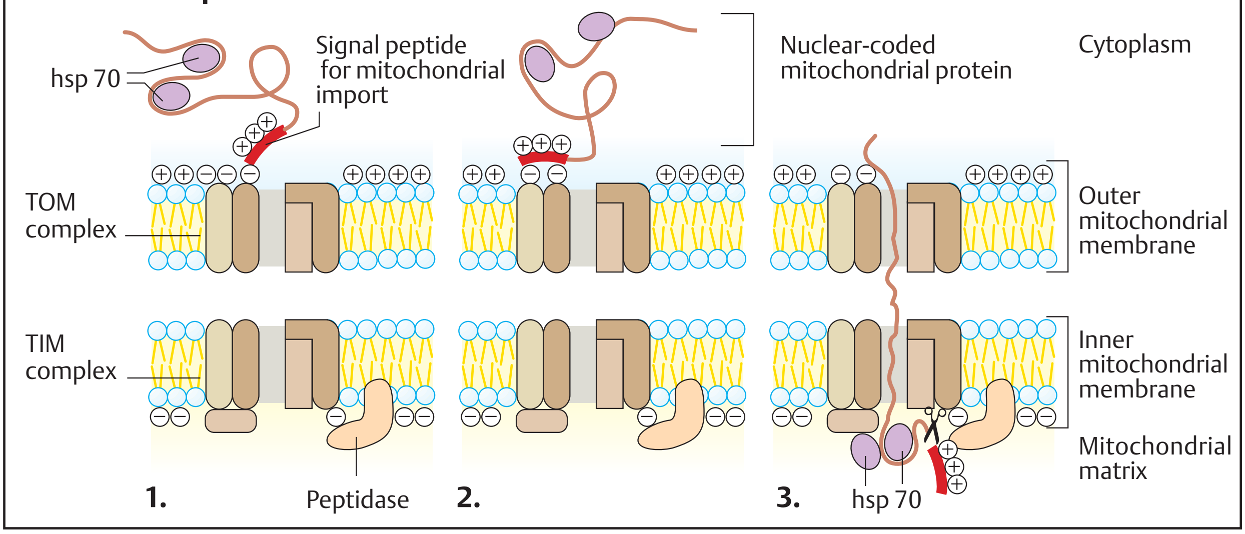
hsp 70
Signal peptide
for mitochondrial
import
Nuclear-coded
mitochondrial protein
Cytoplasm
TOM
complex
Outer
mitochondrial
membrane
TIM
complex
Inner
mitochondrial
membrane
Mitochondrial
matrix
1.
Peptidase
2.
3.
hsp 70

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Lysosomes

A. Structure and contents ◑

Animal lysosomes are organelles with a diameter of 0.2–2.0 µm with various shapes that are surrounded by a single membrane. There are usually several hundred lysosomes per cell. ATP-driven V-type proton pumps are active in their membranes (see p. 220). As these accumulate H^+ in the lysosomes, the content of lysosomes with pH values of 4.5–5 is much more acidic than the cytoplasm (pH 7–7.3).

The lysosomes are the cell's "stomach," serving to break down various cell components. For this purpose, they contain some 40 different types of **hydrolases**, which are capable of breaking down every type of macromolecule. The marker enzyme of lysosomes is *acid phosphatase.* The pH optimum of lysosomal enzymes is adjusted to the acid pH value and is also in the range of pH 5. At neutral pH, as in the cytoplasm, lysosomal enzymes only have low levels of activity. This appears to be a mechanism for protecting the cells from digesting themselves in case lysosomal enzymes enter the cytoplasm at any time. In plants and fungi, the **cell vacuoles** (see p. 43) have the function of lysosomes.

B. Functions ◑

Lysosomes serve for enzymatic degradation of macromolecules and cell organelles, which are supplied in various ways. The example shows the degradation of an overaged mitochondrion by *autophagy.* To accomplish this, the lysosome encloses the organelle (**1**). During this process, the **primary lysosome** converts into a **secondary lysosome**, in which the hydrolytic degradation takes place (**2**). Finally, **residual bodies** contain the indigestible residues of the lysosomal degradation process. Lysosomes are also responsible for the degradation of macromolecules and particles taken up by cells via *endocytosis* and *phagocytosis*—e. g., lipoproteins, proteohormones, and bacteria (*heterophagy*). In the process, lysosomes fuse with the **endosomes** (**3**) in which the endocytosed substances are supplied.

C. Synthesis and transport of lysosomal proteins ◑

Primary lysosomes arise in the region of the Golgi apparatus. **Lysosomal proteins** are synthesized in the rER and are glycosylated there as usual (**1**; see p. 228). The next steps are specific for lysosomal proteins (right part of the illustration). In a two-step reaction, terminal mannose residues (Man) are phosphorylated at the C–6 position of the mannose. First, *N*-acetylglucosamine 1-phosphate is transferred to the OH group at C-6 in a terminal mannose residue, and *N*-acetylglucosamine is then cleaved again. Lysosomal proteins therefore carry a terminal **mannose 6–phosphate** (Man6–P; **2**).

The membranes of the Golgi apparatus contain receptor molecules that bind Man 6–P. They recognize lysosomal proproteins by this residue and bind them (**3**). With the help of *clathrin,* the receptors are concentrated locally. This allows the appropriate membrane sections to be pinched off and transported to the endolysosomes with the help of transport vesicles (**4**), from which primary lysosomes arise through maturation (**5**). Finally, the phosphate groups are removed from Man 6–P (**6**).

The *Man 6–P receptors* are reused. The fall in the pH value in the endolysosomes releases the receptors from the bound proteins (**7**) which are then transported back to the Golgi apparatus with the help of transport vesicles.

Further information

Many hereditary diseases are due to genetic defects in lysosomal enzymes. The metabolism of glycogen (→ *glycogenoses*), lipids (→ *lipidoses*), and proteoglycans (→ *mucopolysaccharidoses*) is particularly affected. As the lysosomal enzymes are indispensable for the intracellular breakdown of macromolecules, unmetabolized macromolecules or degradation products accumulate in the lysosomes in these diseases and lead to irreversible cell damage over time. In the longer term, enlargement takes place, and in severe cases there may be failure of the organ affected—e. g., the liver.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

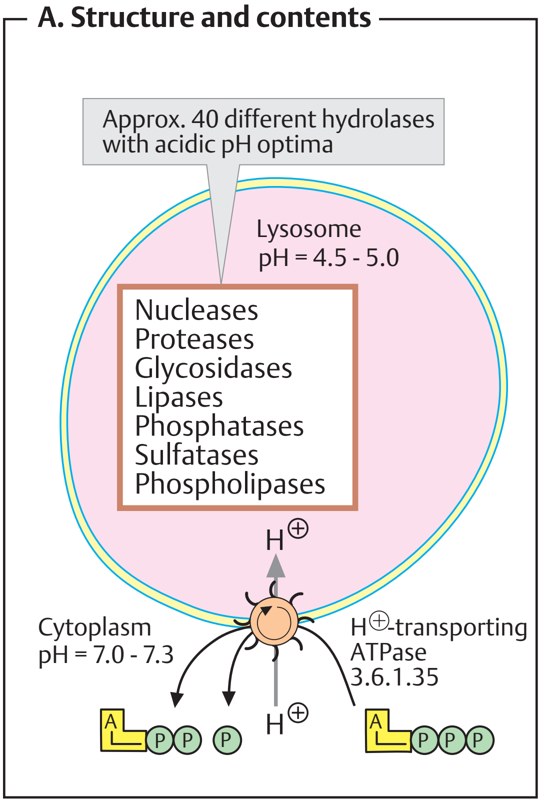
A. Structure and contents
Approx. 40 different hydrolases with acidic pH optima
Lysosome
pH = 4.5 - 5.0
Nucleases
Proteases
Glycosidases
Lipases
Phosphatases
Sulfatases
Phospholipases
H⊕
Cytoplasm
pH = 7.0 - 7.3
H⊕-transporting ATPase 3.6.1.35
H⊕
A P P P
A P P P

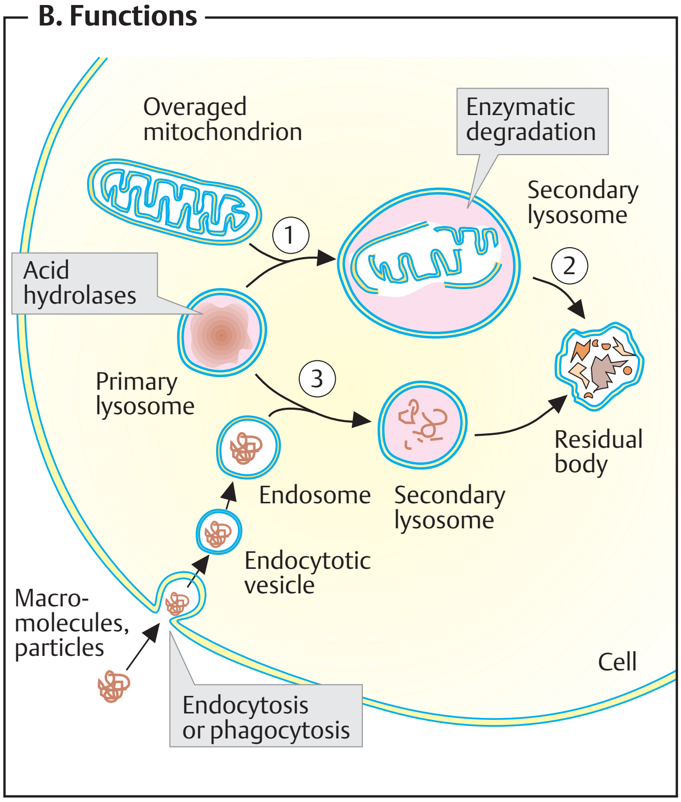
B. Functions
Overaged mitochondrion
Enzymatic degradation
Secondary lysosome
1
2
Acid hydrolases
Primary lysosome
3
Residual body
Endosome
Secondary lysosome
Endocytotic vesicle
Macro-molecules, particles
Endocytosis or phagocytosis
Cell

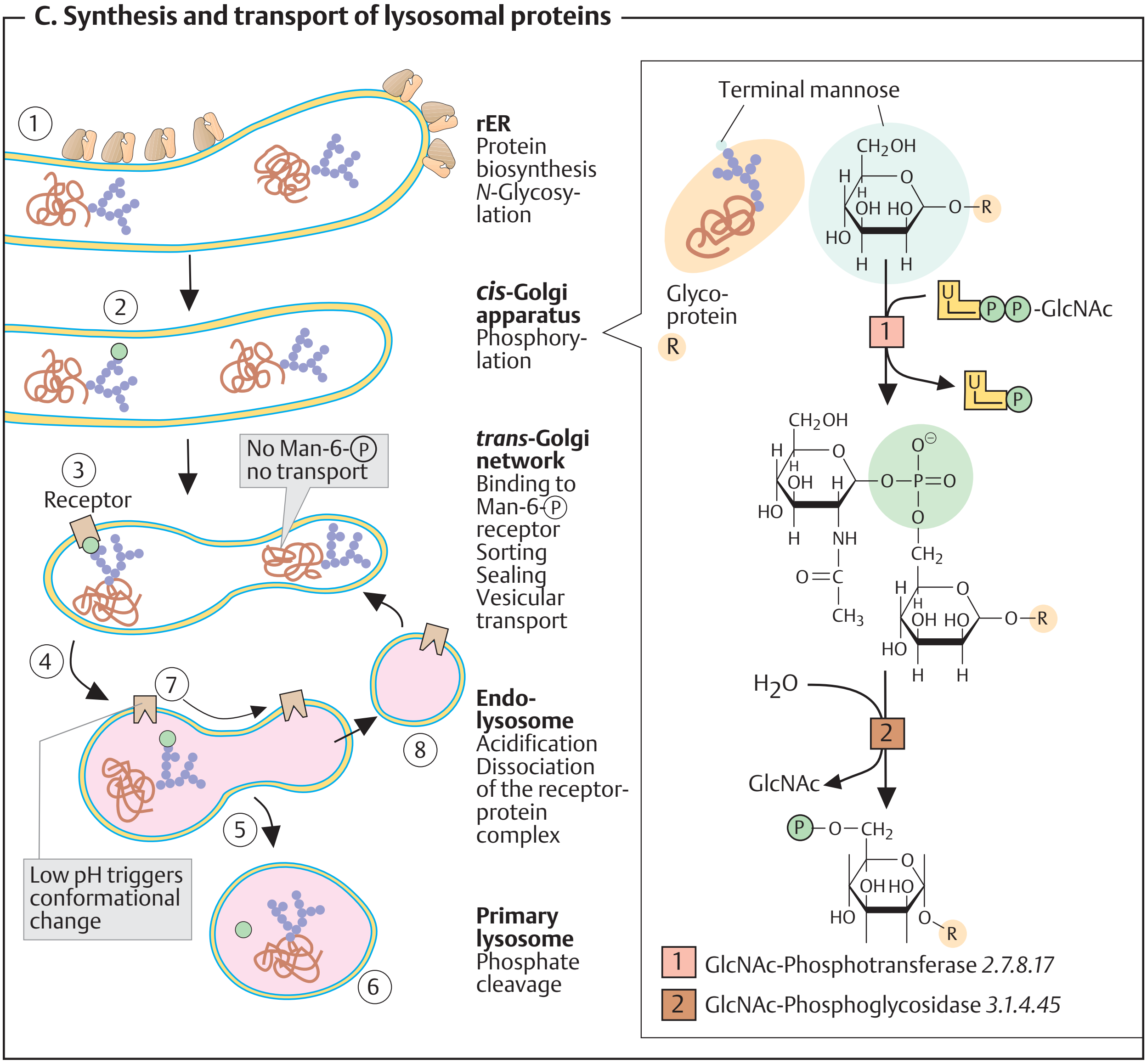
C. Synthesis and transport of lysosomal proteins
1
rER
Protein biosynthesis
N-Glycosylation
2
cis-Golgi apparatus
Phosphorylation
3
Receptor
No Man-6-P no transport
trans-Golgi network
Binding to Man-6-P receptor
Sorting
Sealing
Vesicular transport
4
7
8
Endo-lysosome
Acidification
Dissociation of the receptor-protein complex
5
Low pH triggers conformational change
6
Primary lysosome
Phosphate cleavage
Terminal mannose
Glyco-protein
R
CH2OH
H
O
OH HO
O–R
HO
H H
1
U P P -GlcNAc
U P
CH2OH
O⊖
O–P=O
O
NH
O=C
CH3
CH2
H2O
2
GlcNAc
P–O–CH2
OH HO
O–R
1 GlcNAc-Phosphotransferase 2.7.8.17
2 GlcNAc-Phosphoglycosidase 3.1.4.45

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Molecular genetics: overview

Nucleic acids (**DNA** and various **RNAs**) are of central importance in the storage, transmission, and expression of genetic information. The decisive factor involved is their ability to enter into specific base pairings with each other (see p. 84). The individual processes involved, which are summed up in an overview here, are discussed in more detail on the following pages.

A. Expression and transmission of genetic information ●

Storage. The genetic information of all cells is stored in the base sequence of their DNA (RNA only occurs as a genetic material in viruses; see p. 404). Functional sections of DNA that code for inheritable structures or functions are referred to as **genes**. The 30 000–40 000 human genes represent only a few percent of the **genome**, which consists of approximately 5 10^9 base pairs (bp). Most genes code for proteins—i. e., they contain the information for the sequence of amino acid residues of a protein (its sequence). Every amino acid residue is represented in DNA by a code word (a **codon**) consisting of a sequence of three base pairs (a triplet). At the level of DNA, codons are defined as sequences of the sense strand read in the 5′→3′ direction (see p. 84). A DNA codon for the amino acid phenylalanine, for example, is thus *TTC* (**2**).

Replication. During cell division, all of the genetic information has to be passed on to the daughter cells. To achieve this, the whole of the DNA is copied during the S phase of the cell cycle (see p. 394). In this process, each strand serves as a matrix for the synthesis of a complementary second strand (**1**; see p. 240).

Transcription. For *expression* of a gene—i. e., synthesis of the coded protein—the DNA sequence information has to be converted into a protein sequence. As DNA itself is not involved in protein synthesis, the information is transferred from the nucleus to the site of synthesis in the cytoplasm. To achieve this, the *template strand* in the relevant part of the gene is transcribed into an **RNA** (hnRNA). The sequence of this RNA is thus complementary to that of the *template strand* (**3**), but—with the exception of the exchange of thymine for uracil—it is identical to that of the sense strand. In this way, the DNA triplet *TTC* gives rise in hnRNA to the RNA codon *UUC*.

RNA maturation. In eukaryotes, the hnRNA initially formed is modified several times before it can leave the nucleus as **messenger RNA** (mRNA, **4**). During RNA maturation, superfluous (“intervening”) sequences (introns) are removed from the molecule, and both ends of the transcript are protected by the addition of further nucleotides (see p. 246).

Translation. Mature mRNA enters the cytoplasm, where it binds to **ribosomes**, which convert the RNA information into a peptide sequence. The ribosomes (see p. 250) consist of more than 100 proteins and several RNA molecules (rRNA; see p. 82). rRNA plays a role as a ribosomal structural element and is also involved in the binding of mRNA to the ribosome and the formation of the peptide bond.

The actual information transfer is based on the interaction between the mRNA codons and another type of RNA, **transfer RNA** (tRNA; see p. 82). tRNAs, of which there are numerous types, always provide the correct amino acid to the ribosome according to the sequence information in the mRNA. tRNAs are loaded with an amino acid residue at the 3′ end. Approximately in the middle, they present the triplet that is complementary to each mRNA codon, known as the *anticodon* (*GAA* in the example shown). If the codon *UUC* appears on the mRNA, the anticodon binds a molecule of Phe-t-RNA^{Phe} to the mRNA (**5**) and thus brings the phenylalanine residue at the other end of the molecule into a position in which it can take over the growing polypeptide chain from the neighboring tRNA (**6**).

Amino acid activation. Before binding to the ribosomes, tRNAs are loaded with the correct amino acids by specific ligases (**7**; see p. 248). It is the amino acid tRNA ligases that carry out the transfer (translation) of the genetic information from the nucleic acid level to the protein level.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Expression and transmission of genetic information

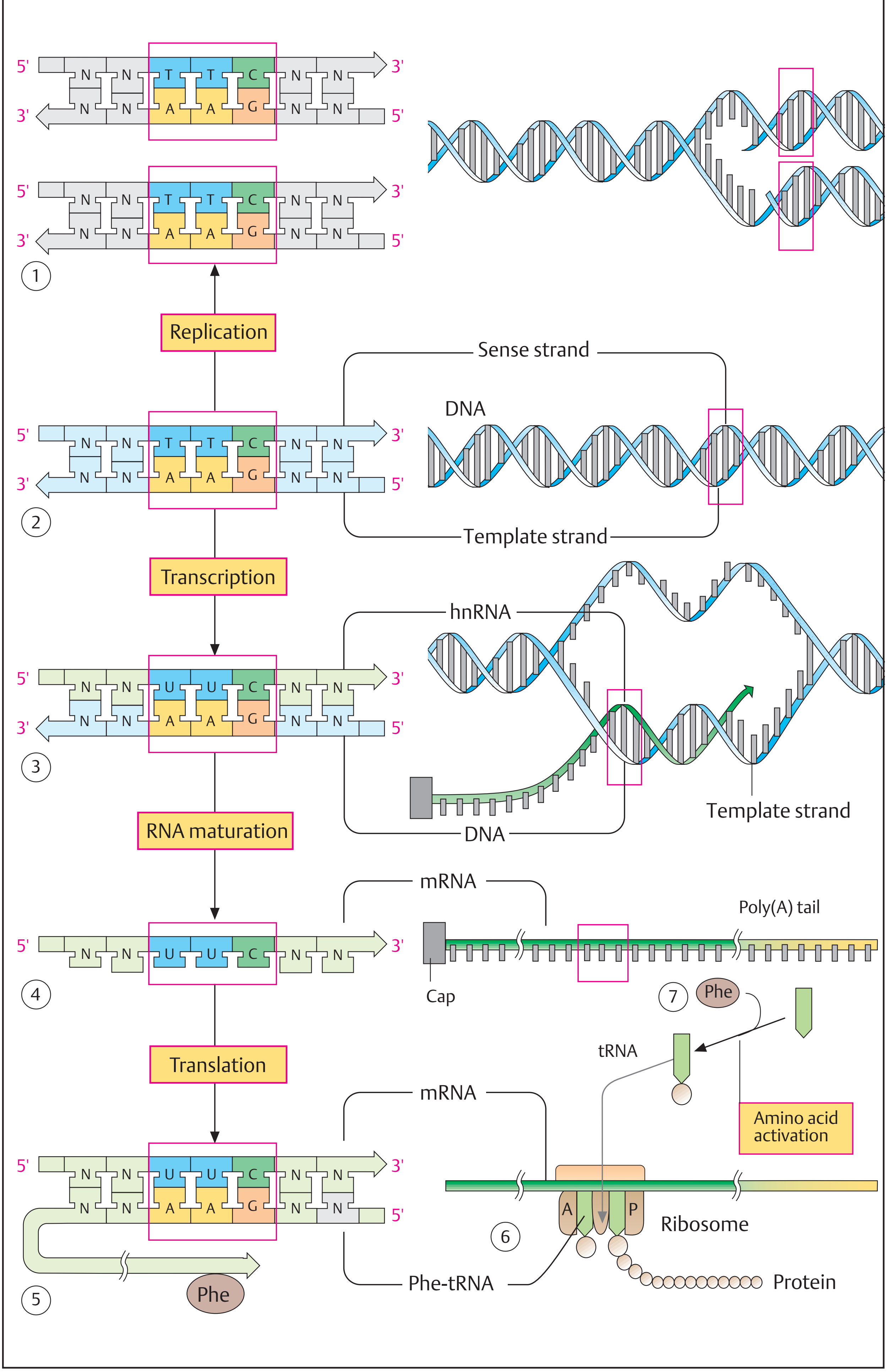
5' N N T T C N N 3'
3' N N A A G N N 5'
5' N N T T C N N 3'
3' N N A A G N N 5'
1
Replication
Sense strand
DNA
5' N N T T C N N 3'
3' N N A A G N N 5'
2
Template strand
Transcription
hnRNA
5' N N U U C N N 3'
3' N N A A G N N 5'
3
Template strand
DNA
RNA maturation
mRNA
Poly(A) tail
5' N N U U C N N 3'
4
Cap
7
Phe
tRNA
Amino acid activation
Translation
mRNA
5' N N U U C N N 3'
3' N N A A G N N 5'
A
P
Ribosome
6
5
Phe
Phe-tRNA
Protein

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Genome

A. Chromatin ◑

In the nuclei of eukaryotes (see p. 196), DNA is closely associated with proteins and RNA. These nucleoprotein complexes, with a DNA proportion of approximately one-third, are known as **chromatin**. It is only during cell division (see p. 394) that chromatin condenses into **chromosomes** that are visible under light microscopy. During interphase, most of the chromatin is loose, and in these conditions a morphological distinction can be made between tightly packed **heterochromatin** and the less dense **euchromatin**. Euchromatin is the site of active transcription.

The proteins contained in chromatin are classified as either histone or non-histone proteins. **Histones** (**B**) are small, strongly basic proteins that are directly associated with DNA. They contribute to the structural organization of chromatin, and their basic amino acids also neutralize the negatively charged phosphate groups, allowing the dense packing of DNA in the nucleus. This makes it possible for the 46 DNA molecules of the diploid human genome, with their 5×10^9 base pairs (bp) and a total length of about 2 m, to be accommodated in a nucleus with a diameter of only 10 μm. Histones also play a central role in regulating transcription (see p. 244).

Two histone molecules each of types **H2A** (blue), **H2B** (green), **H3** (yellow), and **H4** (red) form an octameric complex, around which 146 bp of DNA are wound in 1.8 turns. These particles, with a diameter of 7 nm, are referred to as **nucleosomes**. Another histone (**H1**) binds to DNA segments that are not directly in contact with the histone octamers ("linker" DNA). It covers about 20 bp and supports the formation of spirally wound superstructures with diameters of 30 nm, known as **solenoids**. When chromatin condenses into chromosomes, the solenoids form **loops** about 200 nm long, which already contain about 80 000 bp. The loops are bound to a protein framework (the **nuclear scaffolding**), which in turn organizes some 20 loops to form **minibands**. A large number of stacked minibands finally produces a chromosome. In the chromosome, the DNA is so densely packed that the smallest human chromosome already contains more than 50 million bp.

The **non-histone proteins** are very heterogeneous. This group includes *structural proteins* of the nucleus, as well as many *enzymes* and *transcription factors* (see p. 118), which selectively bind to specific segments of DNA and regulate gene expression and other processes.

B. Histones ○

The histones are remarkable in several ways. With their high proportions of lysine and arginine (blue shading), they are strongly basic, as mentioned above. In addition, their amino acid sequence has hardly changed at all in the course of evolution. This becomes clear when one compares the histone sequences in mammals, plants, and fungi (yeasts are single-celled fungi; see p. 148). For example, the H4 histones in humans and wheat differ only in a single amino acid residue, and there are only a few changes between humans and yeast. In addition, all of these changes are "conservative"—i. e., the size and polarity barely differ. It can be concluded from this that the histones were already "optimized" when the last common predecessor of animals, plants, and fungi was alive on Earth (more than 700 million years ago). Although countless mutations in histone genes have taken place since, almost all of these evidently led to the extinction of the organisms concerned.

The histones in the octamer carry N-terminal mobile "tails" consisting of some 20 amino acid residues that project out of the nucleosomes and are important in the regulation of chromatin structure and in controlling gene expression (see **A2**; only two of the eight tails are shown in full length). For example, the condensation of chromatin into chromosomes is associated with *phosphorylation* (P) of the histones, while the transcription of genes is initiated by *acetylation* (A) of lysine residues in the N-terminal region (see p. 244).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Chromatin

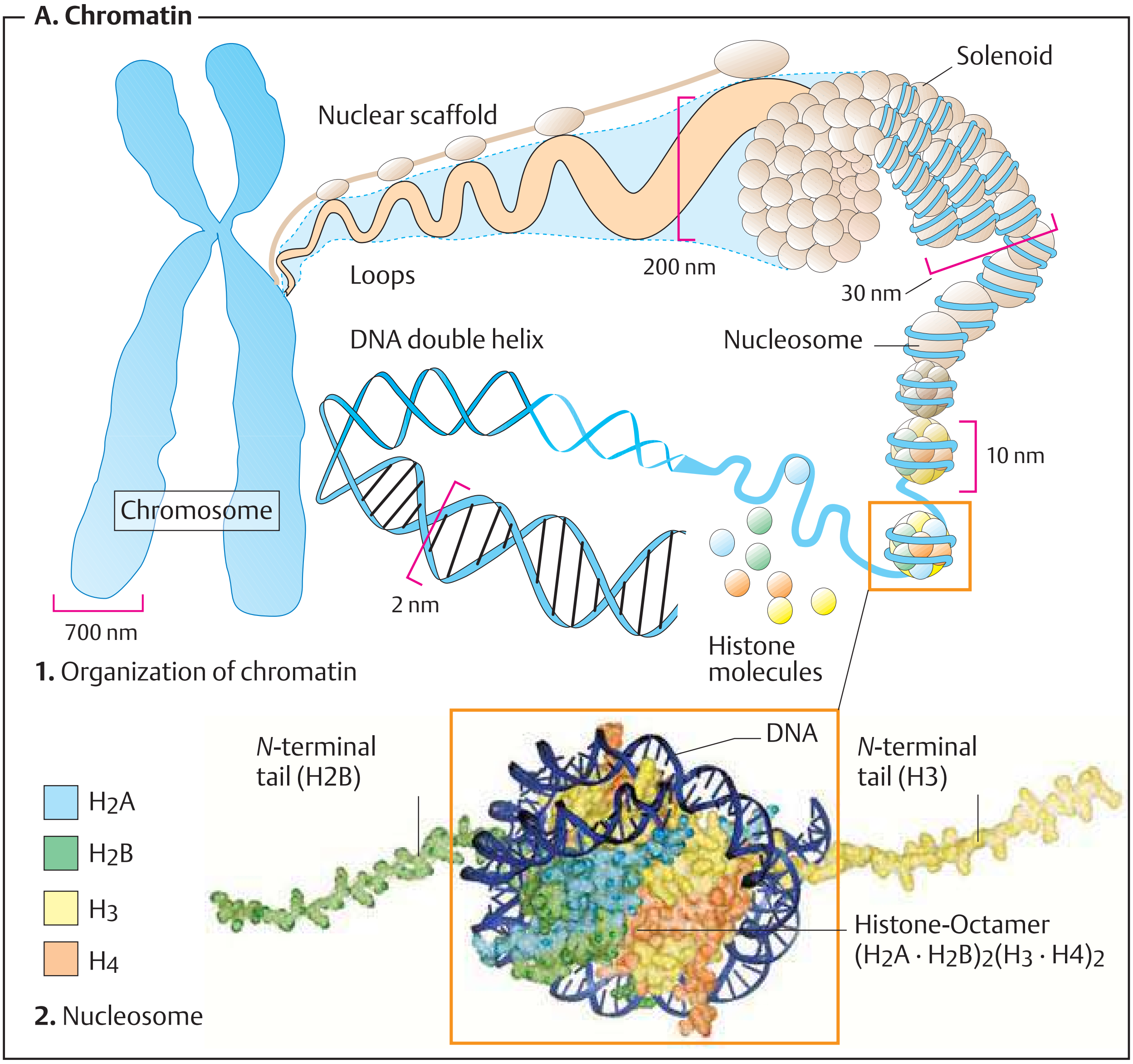
Chromosome
700 nm
Nuclear scaffold
Loops
200 nm
Solenoid
30 nm
Nucleosome
10 nm
DNA double helix
2 nm
Histone molecules
1. Organization of chromatin
H2A
H2B
H3
H4
N-terminal tail (H2B)
DNA
N-terminal tail (H3)
Histone-Octamer (H2A · H2B)2(H3 · H4)2
2. Nucleosome

B. Histone

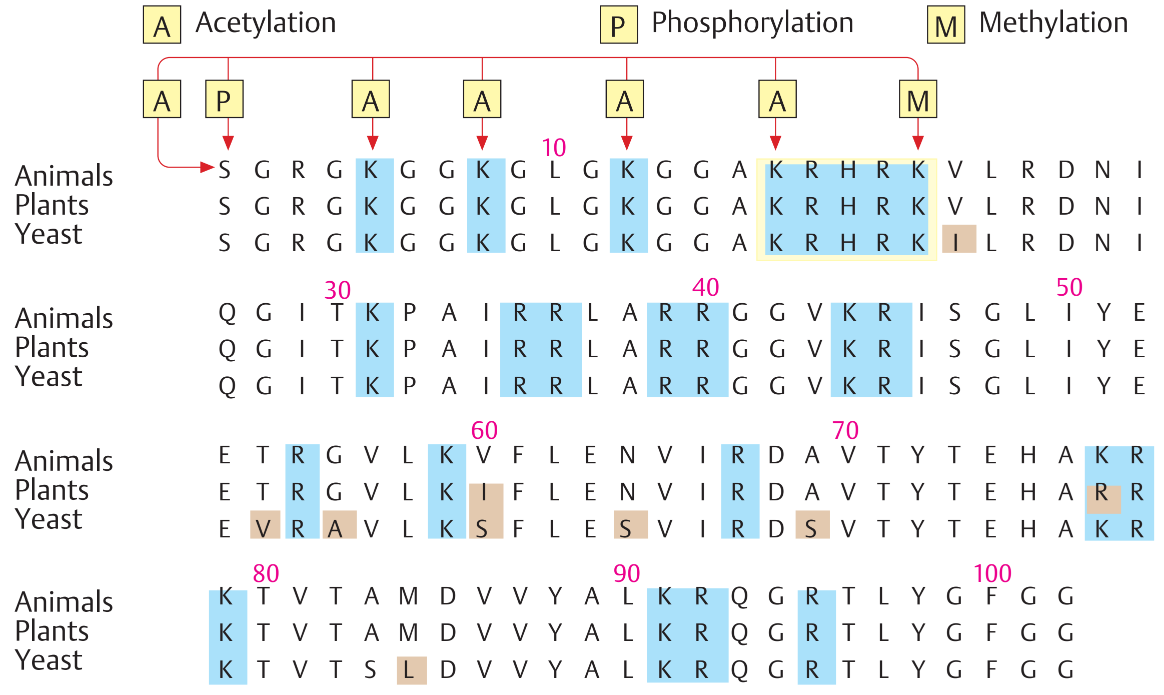
A Acetylation
P Phosphorylation
M Methylation
A
P
A
A
A
A
M
10
30
40
50
60
70
80
90
100

Animals S G R G K G G K G L G K G G A K R H R K V L R D N I
Plants S G R G K G G K G L G K G G A K R H R K V L R D N I
Yeast S G R G K G G K G L G K G G A K R H R K I L R D N I

Animals Q G I T K P A I R R L A R R G G V K R I S G L I Y E
Plants Q G I T K P A I R R L A R R G G V K R I S G L I Y E
Yeast Q G I T K P A I R R L A R R G G V K R I S G L I Y E

Animals E T R G V L K V F L E N V I R D A V T Y T E H A K R
Plants E T R G V L K I F L E N V I R D A V T Y T E H A R R
Yeast E V R A V L K S F L E S V I R D S V T Y T E H A K R

Animals K T V T A M D V V Y A L K R Q G R T L Y G F G G
Plants K T V T A M D V V Y A L K R Q G R T L Y G F G G
Yeast K T V T S L D V V Y A L K R Q G R T L Y G F G G

Amino acid sequence of histone H4

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Replication

For genetic information to be passed on during cell division, a complete copy of the genome has to be produced before each mitosis. This process is known as DNA **replication**.

A. Mechanism of DNA polymerases ◑

Replication is catalyzed by *DNA-dependent DNA polymerases.* These enzymes require a single strand of DNA, known as the **template**. Beginning at a short starting sequence of RNA (the **primer**), they synthesize a second complementary strand on the basis of this template, and thus create a complete DNA double helix again. The substrates of the DNA polymerases are the four deoxynucleoside triphosphates **dATP**, **dGTP**, **dCTP**, and **dTTP**. In each step, base pairing first binds the nucleotide that is complementary to the current base in the template strand. The α–phosphate residue of the newly bound nucleoside triphosphate is then subjected to nucleophilic attack by the 3′–OH group of the nucleotide incorporated immediately previously. This is followed by the elimination of diphosphate and the formation of a new phosphoric acid diester bond. These steps are repeated again for each nucleotide. The mechanism described means that the matrix can only be read in the 3′→5′ direction. In other words, the newly synthesized strand always grows in the **5′→3′ direction**. The same mechanism is also used in transcription by *DNA-dependent RNA polymerases* (see p. 242). Most DNA and RNA polymerases consist of more than 10 subunits, the role of which is still unclear to some extent.

B. Replication in *E. coli* ◑

Although replication in prokaryotes is now well understood, many details in eukaryotes are still unclear. However, it is certain that the process is in principle similar. A simplified scheme of replication in the bacterium *Escherichia coli* is shown here.

In bacteria, replication starts at a specific point in the circular DNA—the **origin of replication**—and proceeds in both directions. This results in two diverging **replication forks**, in which the two strands are replicated simultaneously. Numerous proteins are involved in the processes taking place in this type of fork, and only the most important are shown here. The two strands of the initial DNA (**1**) are shown in blue and violet, while the newly formed strands are pink and orange.

Each fork (**2**) contains two molecules of **DNA polymerase III** and a number of helper proteins. The latter include **DNA topoisomerases** and **single–strand-binding proteins**. Topoisomerases are enzymes that unwind the superhelical DNA double strand (*gyrase, topoisomerase II*) and then separate it into the two individual strands (*helicase, topoisomerase I*). Since the template strand is always read from 3′ to 5′ (see above), only one of the strands (known as the **leading** strand; violet/pink) can undergo *continuous* replication. For the **lagging strand** (light blue), the reading direction is the *opposite* of the direction of movement of the fork. In this matrix, the new strand is first synthesized in individual pieces, which are known as **Okazaki fragments** after their discoverer (green/orange).

Each fragment starts with a short RNA primer (green), which is necessary for the functioning of the DNA polymerase and is synthesized by a special *RNA polymerase* (*"primase,"* not shown). The primer is then extended by *DNA polymerase III* (orange). After 1000–2000 nucleotides have been included, synthesis of the fragment is interrupted and a new one is begun, starting with another RNA primer that has been synthesized in the interim. The individual Okazaki fragments are initially not bound to one another and still have RNA at the 5′ end (**3**). At some distance from the fork, *DNA polymerase I* therefore starts to remove the RNA primer and replace it with DNA components. Finally, the gaps still remaining are closed by a *DNA ligase.* In DNA double helices formed in this way, only *one* of the strands has been newly synthesized—i. e., replication is *semiconservative.*

In bacteria, some 1000 nucleotides are replicated per second. In eukaryotes, replication takes place more slowly (about 50 nucleotides s^{-1}) and the genome is larger. Thousands of replication forks are therefore active simultaneously in eukaryotes.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Mechanism of DNA polymerases

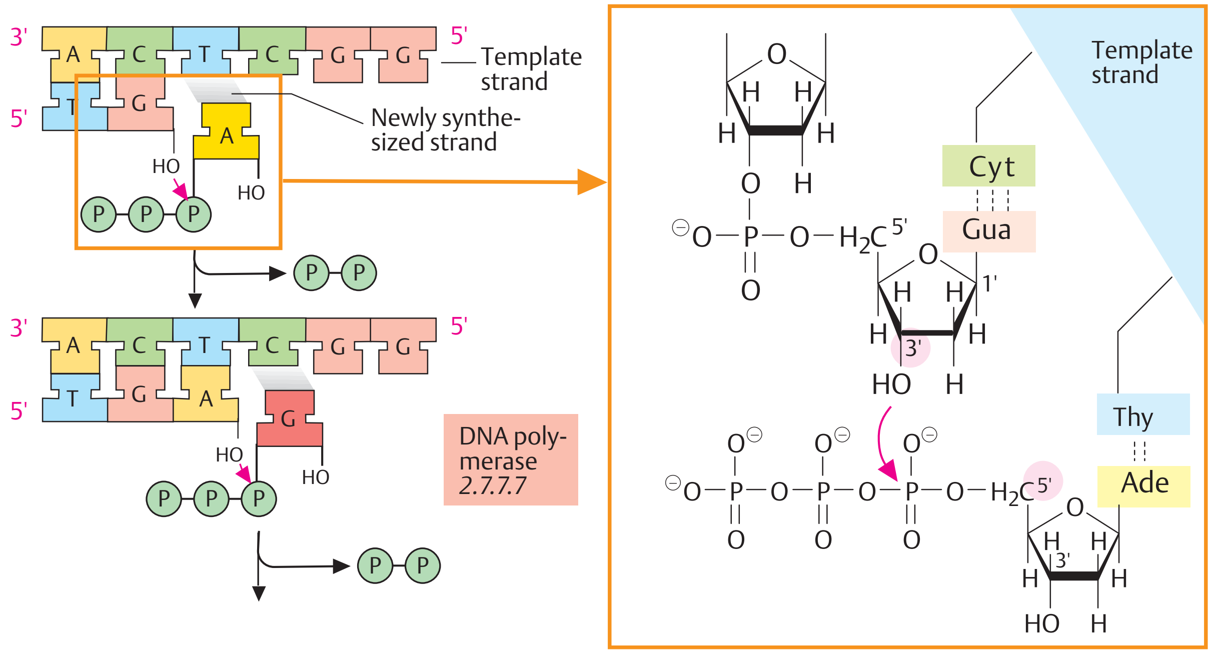
3'
A C T C G G
5'
Template strand
5'
T G
A
Newly synthesized strand
HO
HO
P P P
P P
3'
A C T C G G
5'
5'
T G A
G
HO
HO
P P P
P P
DNA polymerase 2.7.7.7
Template strand
O
H H
H H
O H
⊖O–P–O–H₂C 5'
O
Cyt
Gua
O
H H 1'
H 3' H
HO H
Thy
Ade
O⊖ O⊖ O⊖
⊖O–P–O–P–O–P–O–H₂C 5'
O O O
O
H H
3'
H H
HO H

B. Replication in *E. coli*

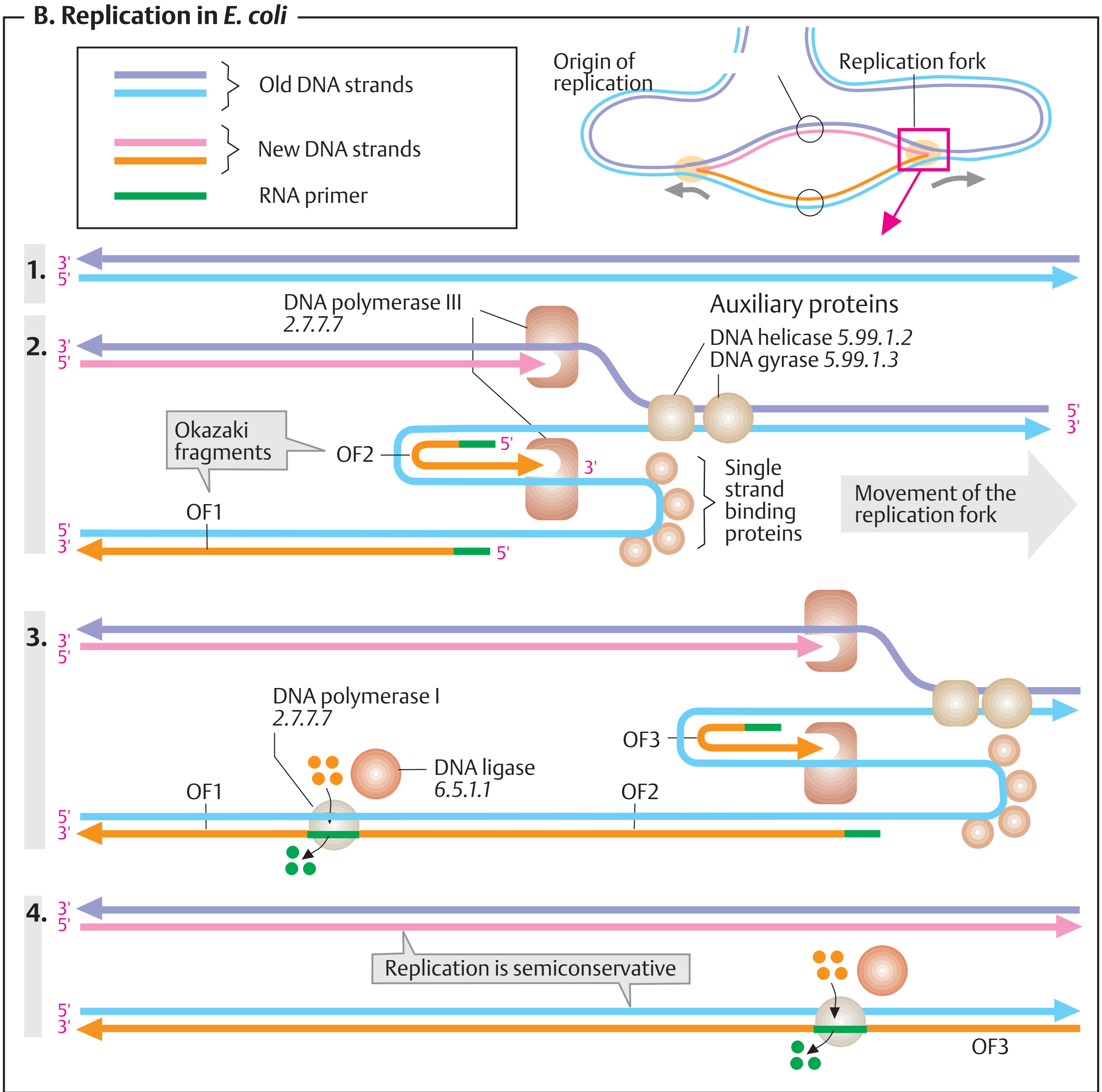
Old DNA strands
New DNA strands
RNA primer
Origin of replication
Replication fork
1.
3'
5'
2.
DNA polymerase III 2.7.7.7
Auxiliary proteins
DNA helicase 5.99.1.2
DNA gyrase 5.99.1.3
5'
3'
Okazaki fragments
OF2
5'
3'
OF1
5'
3'
5'
Single strand binding proteins
Movement of the replication fork
3.
3'
5'
DNA polymerase I 2.7.7.7
DNA ligase 6.5.1.1
OF3
OF1
OF2
5'
3'
4.
3'
5'
Replication is semiconservative
5'
3'
OF3

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Transcription

For the genetic information stored in DNA to become effective, it has to be rewritten (transcribed) into RNA. DNA only serves as a template and is not altered in any way by the **transcription** process. Transcribable segments of DNA that code for a defined product are called **genes.** It is estimated that the mammalian genome contains 30 000–40 000 genes, which together account for less than 5 % of the DNA.

A. Transcription and maturation of RNA: overview ◑

Transcription is catalyzed by *DNA–dependent RNA polymerases.* These act in a similar way to DNA polymerases (see p. 240), except that they incorporate *ribonucleotides* instead of deoxyribonucleotides into the newly synthesized strand; also, they do not require a primer. Eukaryotic cells contain at least three different types of RNA polymerase. *RNA polymerase I* synthesizes an RNA with a sedimentation coefficient (see p. 200) of 45 S, which serves as precursor for three ribosomal RNAs. The products of *RNA polymerase II* are hnRNAs, from which mRNAs later develop, as well as precursors for snRNAs. Finally, *RNA polymerase III* transcribes genes that code for tRNAs, 5S rRNA, and certain snRNAs. These precursors give rise to functional RNA molecules by a process called **RNA maturation** (see p. 246). Polymerases II and III are inhibited by *α–amanitin,* a toxin in the *Amanita phalloides* mushroom.

B. Organization of the PEP-CK gene ○

The way in which a typical eukaryotic gene is organized is illustrated here using a gene that codes for a key enzyme in gluconeogenesis (see p. 154)—the *phosphoenolpyruvate carboxykinase* (PEP-CK).

In the rat, the PEP-CK gene is nearly 7 kbp (kilobase pairs) long. Only 1863 bp, distributed over 10 coding segments (**exons**, dark blue) carry the information for the protein's 621 amino acids. The remainder is allotted to the promoter (pink) and intervening sequences (**introns**, light blue). The gene's promoter region (approximately 1 kbp) serves for regulation (see p. 188). Transcription starts at the 3′ end of the promoter ("transcription start") and continues until the polyadenylation sequence (see below) is reached. The primary transcript (**hnRNA**) still has a length of about 6.2 kbp. During RNA maturation, the noncoding sequences corresponding to the introns are removed, and the two ends of the hnRNA are modified. The translatable mRNA still has half the length of the hnRNA and is modified at both ends (see p. 246).

In many eukaryotic genes, the proportion of introns is even higher. For example, the gene for *dihydrofolate reductase* (see p. 402) is over 30 kbp long. The information is distributed over six exons, which together have a length of only about 6 kbp.

C. Transcription process ◑

As mentioned above, *RNA polymerase II* (green) binds to the 3′ end of the promoter region. A sequence that is important for this binding is known as the **TATA box**—a short A– and T–rich sequence that varies slightly from gene to gene. A typical base sequence *("consensus sequence")* is ...TATAAA... Numerous proteins known as *basal transcription factors* are necessary for the interaction of the polymerase with this region. Additional factors can promote or inhibit the process (transcriptional control; see p. 244). Together with the polymerase, they form the **basal transcription complex**.

At the end of **initiation** (**2**), the polymerase is repeatedly phosphorylated, frees itself from the basal complex, and starts moving along the DNA in the 3′ direction. The enzyme separates a short stretch of the DNA double helix into two single strands. The complementary nucleoside triphosphates are bound by base pairing in the *template strand* and are linked step by step to the hnRNA as it grows in the 5′→3′ direction (**3**). Shortly after the beginning of **elongation**, the 5′ end of the transcript is protected by a "cap" (see p. 246). Once the polyadenylation sequence has been reached (typical sequence: ...AATAA...), the transcript is released (**4**). Shortly after this, the RNA polymerase stops transcribing and dissociates from the DNA.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Transcription and maturation of RNA: overview

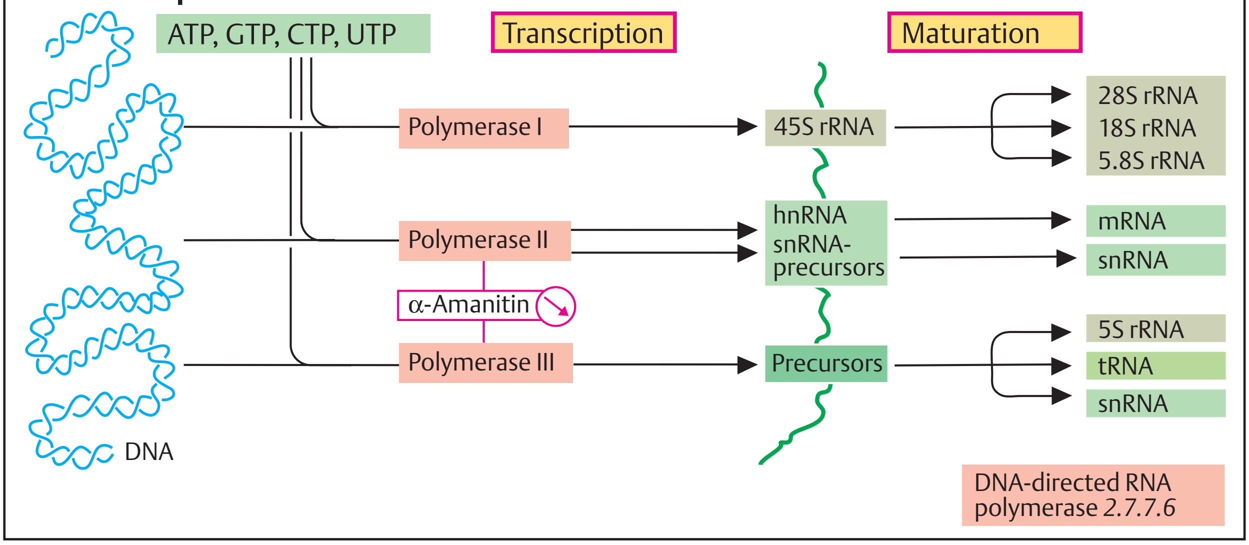
ATP, GTP, CTP, UTP
Transcription
Maturation
Polymerase I
45S rRNA
28S rRNA
18S rRNA
5.8S rRNA
Polymerase II
hnRNA
snRNA-
precursors
mRNA
snRNA
α-Amanitin
Polymerase III
Precursors
5S rRNA
tRNA
snRNA
DNA
DNA-directed RNA polymerase 2.7.7.6

B. Organization of the PEP-CK gene

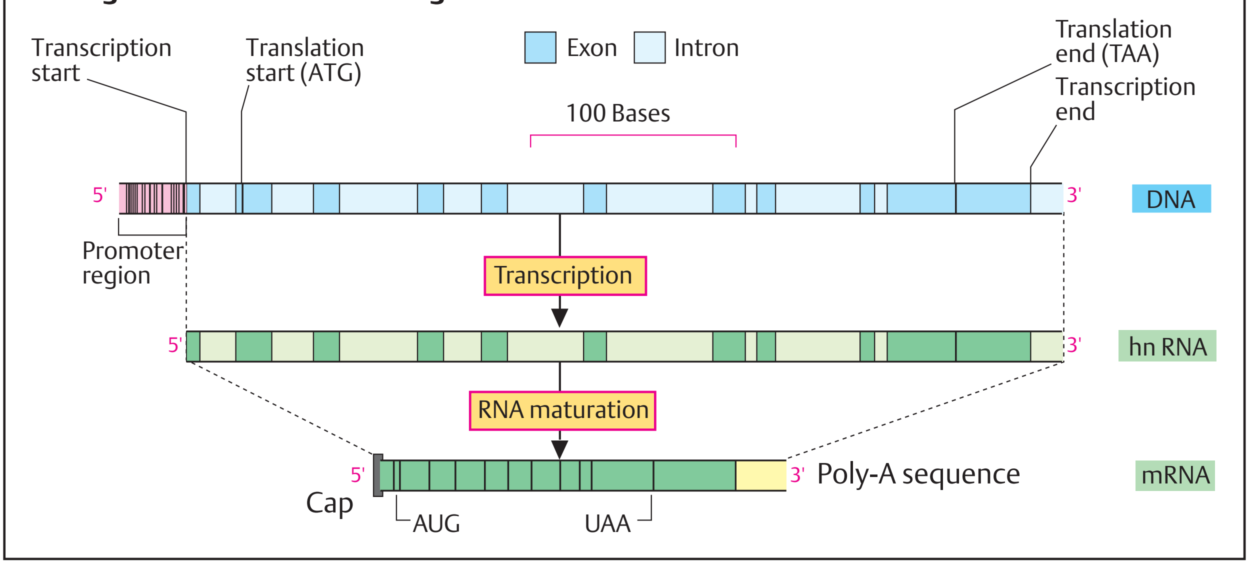
Transcription start
Translation start (ATG)
Exon
Intron
Translation end (TAA)
Transcription end
100 Bases
5'
3'
DNA
Promoter region
Transcription
5'
3'
hn RNA
RNA maturation
5'
Cap
3'
Poly-A sequence
AUG
UAA
mRNA

C. Process of transcription

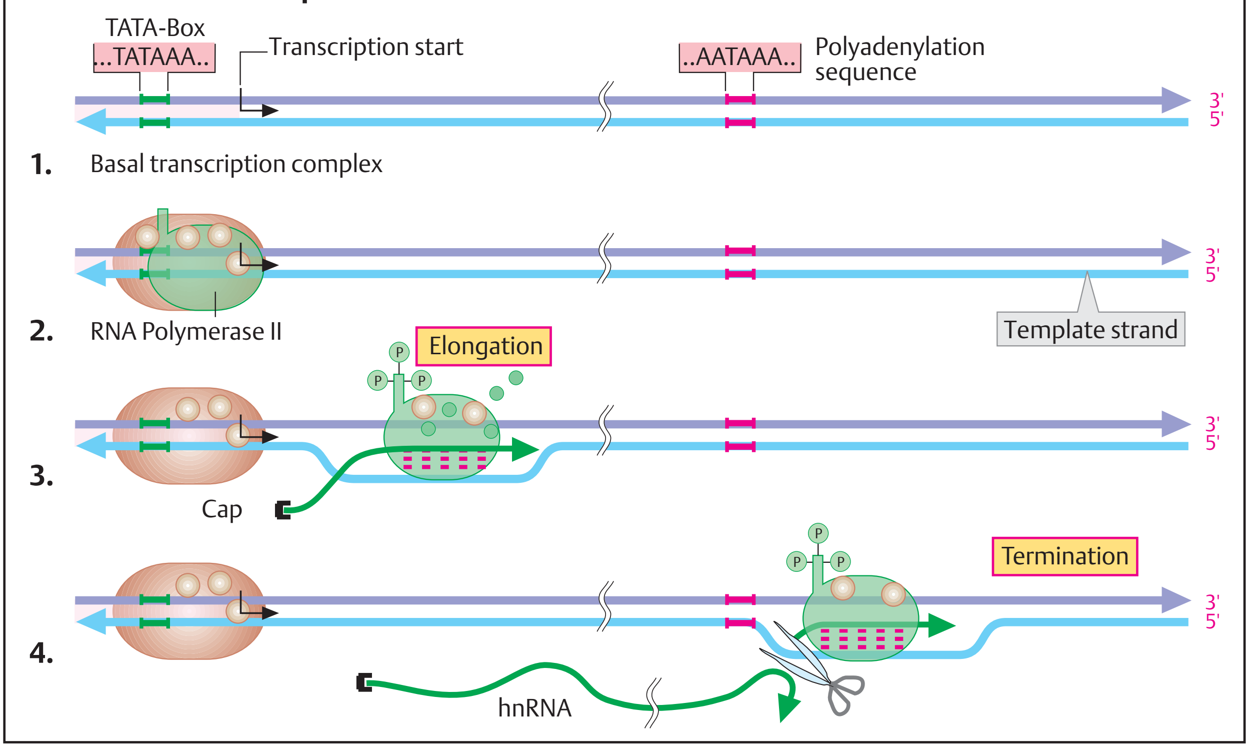
TATA-Box
...TATAAA..
Transcription start
..AATAAA..
Polyadenylation sequence
3'
5'
1.
Basal transcription complex
3'
5'
2.
RNA Polymerase II
Template strand
Elongation
P
P
P
3'
5'
3.
Cap
Termination
P
P
P
3'
5'
4.
hnRNA

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Transcriptional control

Although all cells contain the complete genome, they only use a fraction of the information in it. The genes known as “housekeeping genes,” which code for structural molecules and enzymes of intermediate metabolism, are the only ones that undergo constant transcription. The majority of genes are only active in certain cell types, in specific metabolic conditions, or during differentiation. Which genes are transcribed and which are not is regulated by **transcriptional control** (see also p. 118). This involves *control elements* (*cis*-active elements) in the gene’s promoter region and gene-specific regulatory proteins (transcription factors, *trans* -active factors), which bind to the control elements and thereby activate or inhibit transcription.

A. Initiation of transcription ○

In the higher organisms, DNA is blocked by histones (see p. 238) and is therefore not capable of being transcribed without special positive regulation. In eukaryotes, it is therefore histones that play the role of *repressors* (see p. 118). For transcription to be set in motion at all, the chromatin first has to be restructured.

In the resting state, the lysine residues in the N-terminal “tail” of the histones (see p. 238) are not acetylated. In this state, which can be produced by *histone deacetylases* [1], the nucleosomes are stable. It is only the interaction of activator and regulator proteins with their control elements that allows the binding of coactivator complexes that have *histone acetylase* activity [2]. They acetylate the histone tails and thereby loosen the nucleosome structure sufficiently for the basal transcription complex to form.

This consists of *DNA-dependent RNA polymerase II* and basal transcription factors (TFIIX, X = A – H). First, the basal factor TFIID binds to the promoter. TFIID, a large complex of numerous proteins, contains *TATA box-binding protein* (TBP) and so-called TAFs (TBP-associated factors). The polymerase is attached to this core with the help of TFIIB. Before transcription starts, additional TFs have to bind, including TFIIH, which has *helicase* activity and separates the two strands of DNA during elongation. In all, some 35 different proteins are involved in the basal complex. This alone, however, is still not sufficient for transcription to start. In addition, positive signals have to be emitted by more distant *trans*-active factors, integrated by the coactivator/mediator complex, and passed on to the basal complex (see **B**).

The actual signal for starting elongation consists of the multiple phosphorylation of a domain in the C-terminal region of the polymerase. In phosphorylated form, it releases itself from the basal complex along with a few TFs and starts to synthesize hnRNA.

B. Regulation of PEP-CK transcription ○

Phosphoenolpyruvate carboxykinase (PEP-CK), a key enzyme in gluconeogenesis, is regulated by several hormones, all of which affect the transcription of the PEP-CK gene. Cortisol, glucagon, and thyroxin induce PEP-CK, while insulin inhibits its induction (see p. 158).

More than ten **control elements** (dark red), distributed over approximately 1 kbp, have so far been identified in the promoter of the PEP-CK gene (top). These include response elements for the glucocorticoid receptor (GRE; see p. 378), for the thyroxin receptor (TRE), and for the steroid-like retinoic acid (AF-1). Additional control elements (CRE, cAMP-responsive element) bind the transcription factor C/EBP, which is activated by cAMP-dependent protein kinase A through phosphorylation. This is the way in which glucagon, which raises the cAMP level (see p. 158), works. Control element P1 binds the hormone-independent factor NF-1 (nuclear factor-1). All proteins that bind to the control elements mentioned above are in contact with a **coactivator/mediator complex** (CBP/p300), which integrates their input like a computer and transmits the result in the form of stronger or weaker signals to the basal transcription complex. Inhibition of PEP-CK transcription by insulin is mediated by an insulin-responsive element (IRE) in the vicinity of the GRE. Binding of an as yet unknown factor takes place here, inhibiting the binding of the glucocorticoid receptor to the GRE.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Initiation of transcription

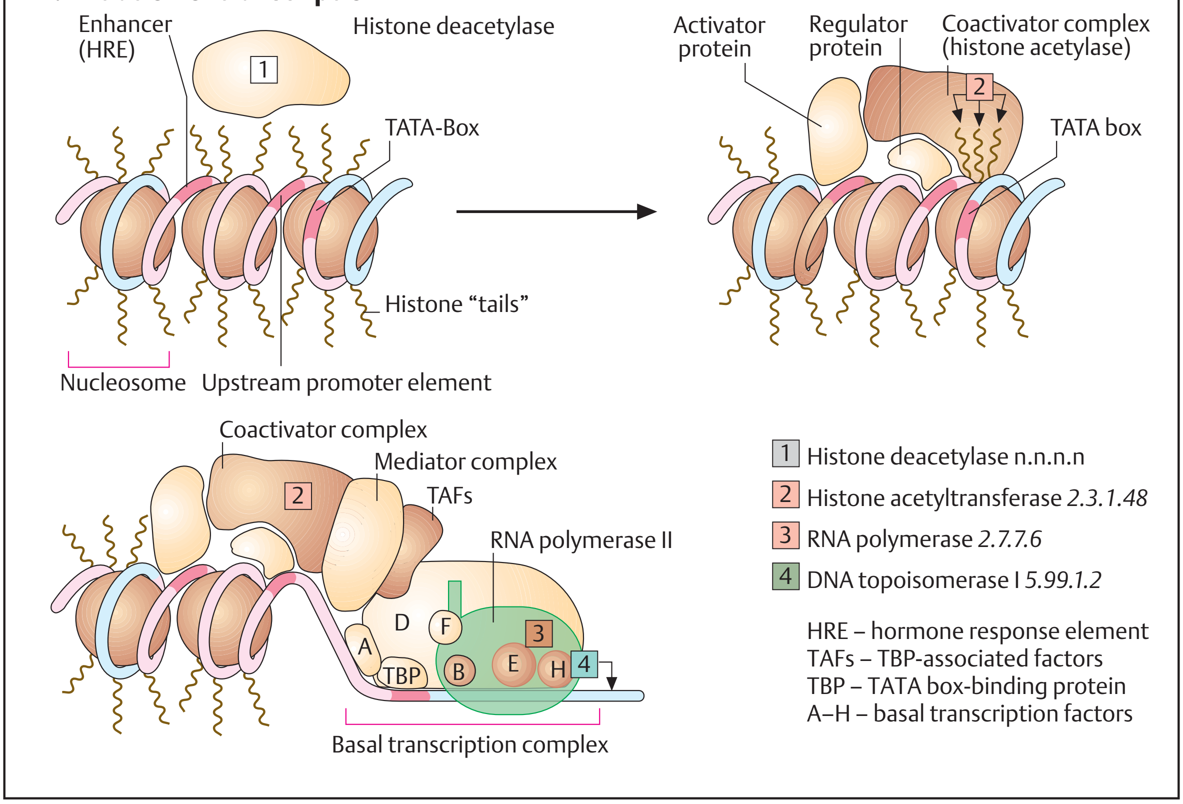
Enhancer (HRE)
Histone deacetylase
1
TATA-Box
Histone "tails"
Nucleosome
Upstream promoter element
Activator protein
Regulator protein
Coactivator complex (histone acetylase)
2
TATA box
Coactivator complex
Mediator complex
2
TAFs
RNA polymerase II
D
F
A
3
TBP
B
E
H
4
Basal transcription complex
1 Histone deacetylase n.n.n.n
2 Histone acetyltransferase 2.3.1.48
3 RNA polymerase 2.7.7.6
4 DNA topoisomerase I 5.99.1.2
HRE – hormone response element
TAFs – TBP-associated factors
TBP – TATA box-binding protein
A–H – basal transcription factors

B. Regulation of PEP-CK transcription

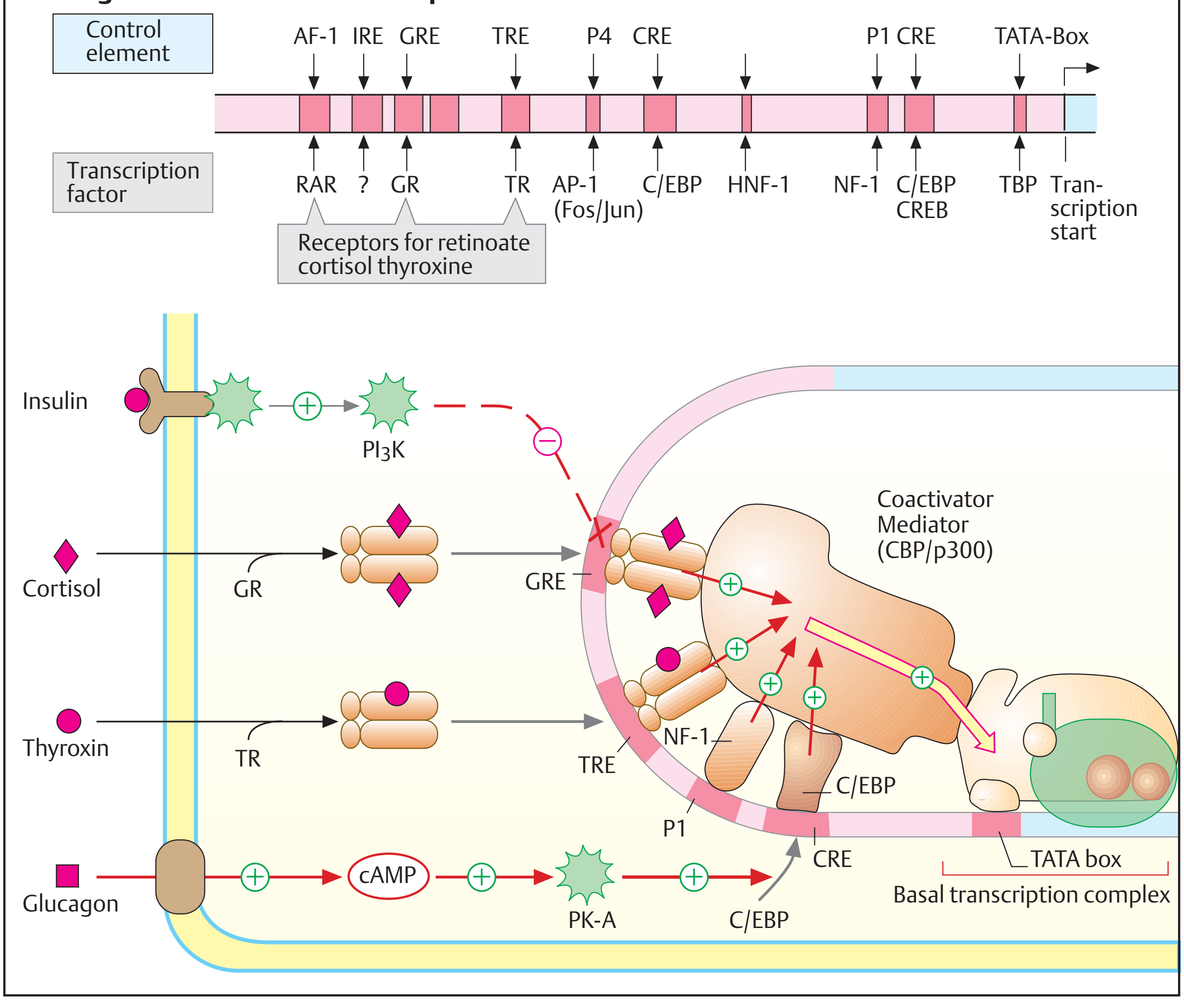
Control element
AF-1
IRE
GRE
TRE
P4
CRE
P1
CRE
TATA-Box
Transcription factor
RAR
?
GR
TR
AP-1 (Fos/Jun)
C/EBP
HNF-1
NF-1
C/EBP CREB
TBP
Tran-scription start
Receptors for retinoate cortisol thyroxine
Insulin
PI3K
Cortisol
GR
GRE
Thyroxin
TR
TRE
Coactivator Mediator (CBP/p300)
NF-1
C/EBP
P1
CRE
TATA box
Basal transcription complex
Glucagon
cAMP
PK-A
C/EBP

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

RNA maturation

Before the hnRNA produced by RNA polymerase II (see p. 242) can leave the nucleus in order to serve as a template for protein synthesis in the cytoplasm, it has to undergo several modifications first. Even during transcription, the two ends of the transcript have additional nucleotides added (**A**). The sections that correspond to the intervening gene sequences in the DNA (introns) are then cut out (splicing; see **B**). Other transcripts—e. g., the 45 S precursor of rRNA formed by polymerase I (see p. 242)—are broken down into smaller fragments by nucleases before export into the cytoplasm.

A. 5′ and 3′ modification of mRNA ◑

Shortly after transcription begins in eukaryotes, the end of the growing RNA is blocked in several reaction steps by a structure known as a “**cap**.” In hnRNAs, this consists of a GTP residue that is methylated at N-7 of the guanine ring. The β-phosphate residue of the cap is linked to the free 5′-OH group of the terminal ribose via an ester bond. After the “polyadenylation signal” has been reached (typical sequence: ...AAUAAA...; see p. 242), a **polyadenylate “tail”** consisting of up to 200 AMP nucleotides is also added at the free 3′ end of the transcript. This reaction is catalyzed by a special *polyadenylate polymerase*. It is only at this point that the mRNA leaves the nucleus as a complex with RNA-binding proteins.

Both the cap and the poly-A tail play a vital part in initiating eukaryotic translation (see p. 250). They help position the ribosome correctly on the mRNA near to the starting codon. The protection which the additional nucleotides provide against premature enzymatic degradation appears to be of lesser importance.

B. Splicing of hnRNA ◑

Immediately after transcription, the hnRNA introns are removed and the exons are linked to form a continuous coding sequence. This process, known as **splicing**, is supported by complicated RNA–protein complexes in the nucleus, the so-called **spliceosomes**. The components of these macromolecular machines are called **snRNPs** (*small nuclear ribonucleoprotein particles,* pronounced “snurps”). SnRNPs occur in five different forms (U1, U2, U4, U5, and U6). They consist of numerous *proteins* and one molecule of *snRNA* each (see p. 82).

To ensure that the RNA message is not destroyed, splicing has to take place in a very precise fashion. The start and end of the hnRNA introns are recognized by a characteristic sequence (...AGGT... at the 5′ end or ...[C,U]AGG... at the 3′ end). Another important structure is the so-called *branching point* inside the intron. Its sequence is less conserved than the terminal splicing sites, but it always contains one adenosine residue (A). During splicing, the 2′-OH group of this residue—supported by the spliceosome (see **C**)—attacks the phosphoric acid diester bond at the 5′ end of the intron and cleaves it (**b**). Simultaneously, an unusual 2′→5′ bond is formed inside the intron, which thereby takes on a *lasso shape* (**c**; see formula). In the second step of the splicing process, the free 3′-OH group at the end of the 5′ terminal exon attacks the A–G bond at the 3′ end of the intron. As a result, the two exons are linked and the intron is released, still in a lasso shape (d).

C. Spliceosome ○

As described above, it is residues of the hnRNA that carry out bond cleavage and bond formation during the splicing process. It is therefore not a protein enzyme that acts as a catalyst here, but rather an RNA. Catalytic RNAs of this type are called *ribozymes* (see also p. 88). The task of the spliceosomes is to fix and orientate the reacting groups by establishing base pairings between snRNAs and segments of the hnRNA. The probable situation before the adenosine attack at the branching point on the 5′ splicing site (see **B**, Fig. **b**) is shown schematically on the right side of the illustration. In this phase, the U1 snRNA fixes the 5′ splicing site, U2 fixes the branching site, and U5 fixes the ends of the two exons.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. 5' and 3' modification of mRNA

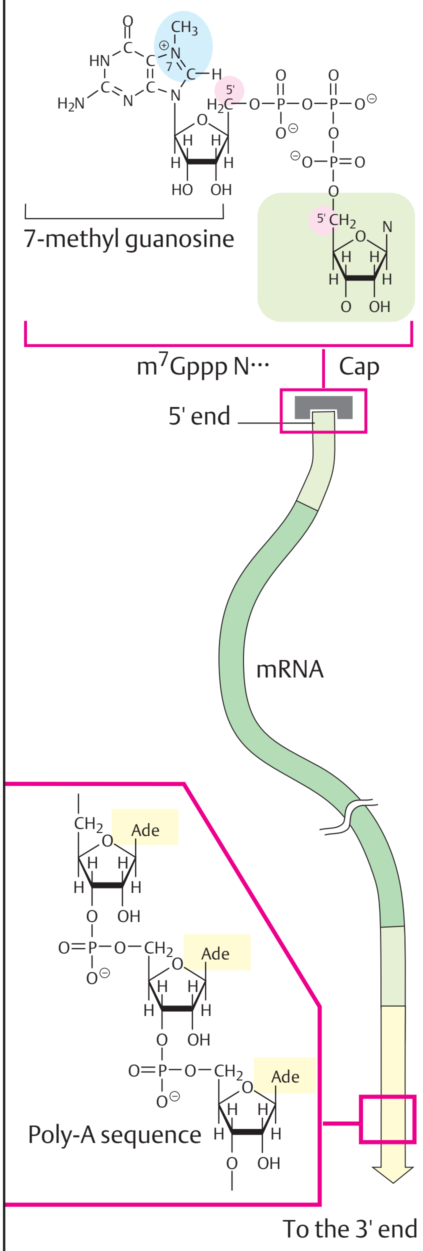
7-methyl guanosine
m⁷Gppp N···
Cap
5' end
mRNA
Poly-A sequence
Ade
To the 3' end

B. Splicing of hnRNA: mechanism

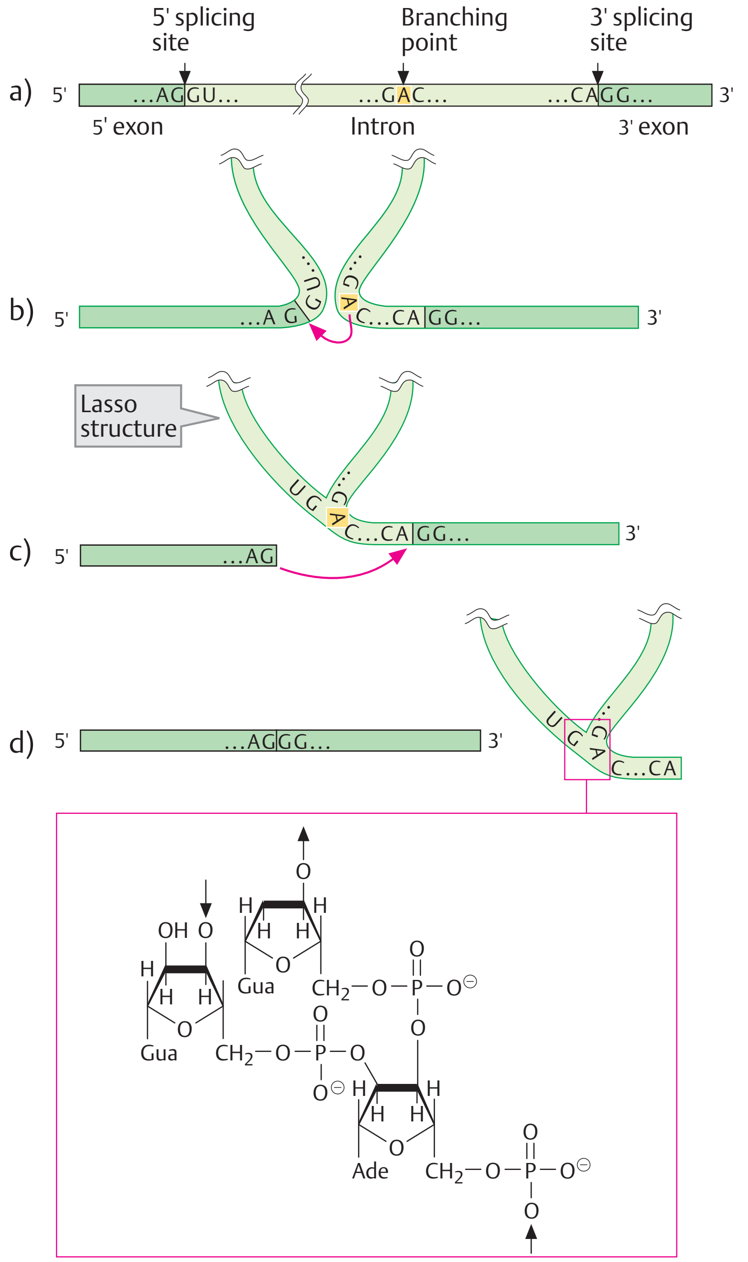
5' splicing site
Branching point
3' splicing site
a)
5'
...AG|GU...
...GAC...
...CA|GG...
3'
5' exon
Intron
3' exon
b)
...A G
GU...
...GA
C...CA|GG...
Lasso structure
c)
...AG
UG
C...CA|GG...
d)
...AG|GG...
UG
GA
C...CA
OH O
Gua
Ade

C. Spliceosome

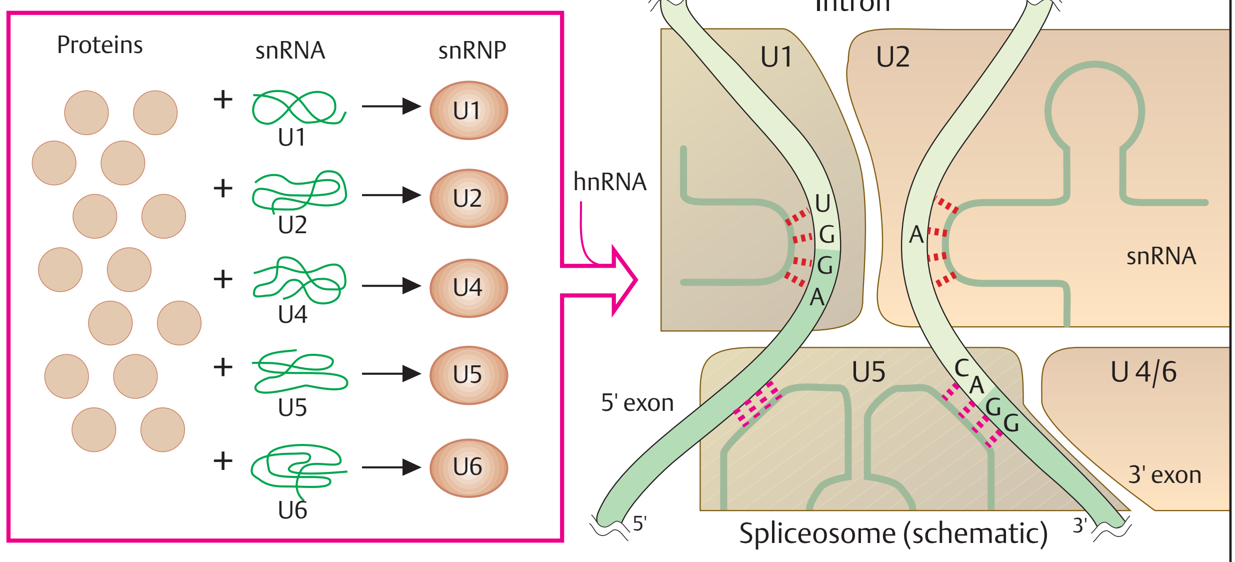
Proteins
snRNA
snRNP
+
U1
U2
U4
U5
U6
hnRNA
Intron
U1
U2
U5
U 4/6
snRNA
UGGA
A
CAGG
5' exon
3' exon
5'
3'
Spliceosome (schematic)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Amino acid activation

A. The genetic code ◑

Most of the genetic information stored in the genome codes for the amino acid sequences of proteins. For these proteins to be expressed, a text in “nucleic acid language” therefore has to be translated into “protein language.” This is the origin of the use of the term **translation** to describe protein biosynthesis. The dictionary used for the translation is the **genetic code.**

As there are 20 proteinogenic amino acids (see p. 60), the nucleic acid language has to contain at least as many words **(codons)**. However, there are only four letters in the nucleic acid alphabet (A, G, C, and U or T). To obtain 20 different words from these, each word has to be at least three letters long (with two letters, there would only be $4^2 = 16$ possibilities). And in fact the codons do consist of three sequential bases **(triplets)**.

Figure 1 shows the standard code in “DNA language” (i. e., as a sequence of triplets in the *sense strand* of DNA, read in the 5′→3′ direction; see p. 84), represented as a circular diagram. The scheme is read from the inside to the outside. For example, the triplet CAT codes for the amino acid histidine. With the exception of the exchange of U for T, the DNA codons are identical to those of mRNA.

As the genetic code provides $4^3 = 64$ codons for the 20 amino acids, there are several synonymous codons for most amino acids—the code is **degenerate**. Three triplets do not code for amino acids, but instead signal the end of translation **(stop codons)**. Another special codon, the **start codon**, marks the start of translation. The code shown here is almost universally applicable; only the mitochondria (see p. 210) and a few microorganisms deviate from it slightly.

As an example of the way in which the code is read, Fig. 2 shows small sections from the normal and a mutated form of the β-globin gene (see p. 280), as well as the corresponding mRNA and protein sequences. The **point mutation** shown, which is relatively frequent, leads to replacement of a glutamate residue in position 6 of the β-chain by valine (GAG → GTG). As a consequence, the mutated hemoglobin tends to aggregate in the deoxygenated form. This leads to sickle-shaped distortions of the erythrocytes and disturbances of O_2 transport (sickle-cell anemia).

B. Amino acid activation ◑

Some 20 different *amino acid tRNA ligases* in the cytoplasm each bind one type of tRNA (see p. 82) with the corresponding amino acid. This reaction, known as **amino acid activation**, is endergonic and is therefore coupled to ATP cleavage in two steps.

First, the amino acid is bound by the enzyme and reacts there with ATP to form diphosphate and an “energy–rich” mixed acid anhydride (**aminoacyl adenylate**). In the second step, the 3′-OH group (in other ligases it is the 2′-OH group) of the terminal ribose residue of the tRNA takes over the amino acid residue from the aminoacyl adenylate. In **aminoacyl tRNAs**, the carboxyl group of the amino acid is therefore esterified with the ribose residue of the terminal adenosine of the sequence ...CCA-3′.

The accuracy of translation primarily depends on the specificity of the amino acid tRNA ligases, as incorrectly incorporated amino acid residues are not recognized by the ribosome later. A *“proofreading mechanism”* in the active center of the ligase therefore ensures that incorrectly incorporated amino acid residues are immediately removed again. On average, an error only occurs once every 1300 amino acid residues. This is a surprisingly low rate considering how similar some amino acids are—e. g., leucine and isoleucine.

C. Asp–tRNA ligase (dimer) ○

The illustration shows the ligase responsible for the activation of aspartate. Each subunit of the dimeric enzyme (protein parts shown in orange) binds one molecule of tRNA^{Asp} (blue). The active centers can be located by the bound ATP (green). They are associated with the 3′ end of the tRNA. Another domain in the protein (upper left) is responsible for “recognition” of the tRNA anticodon.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. The genetic code

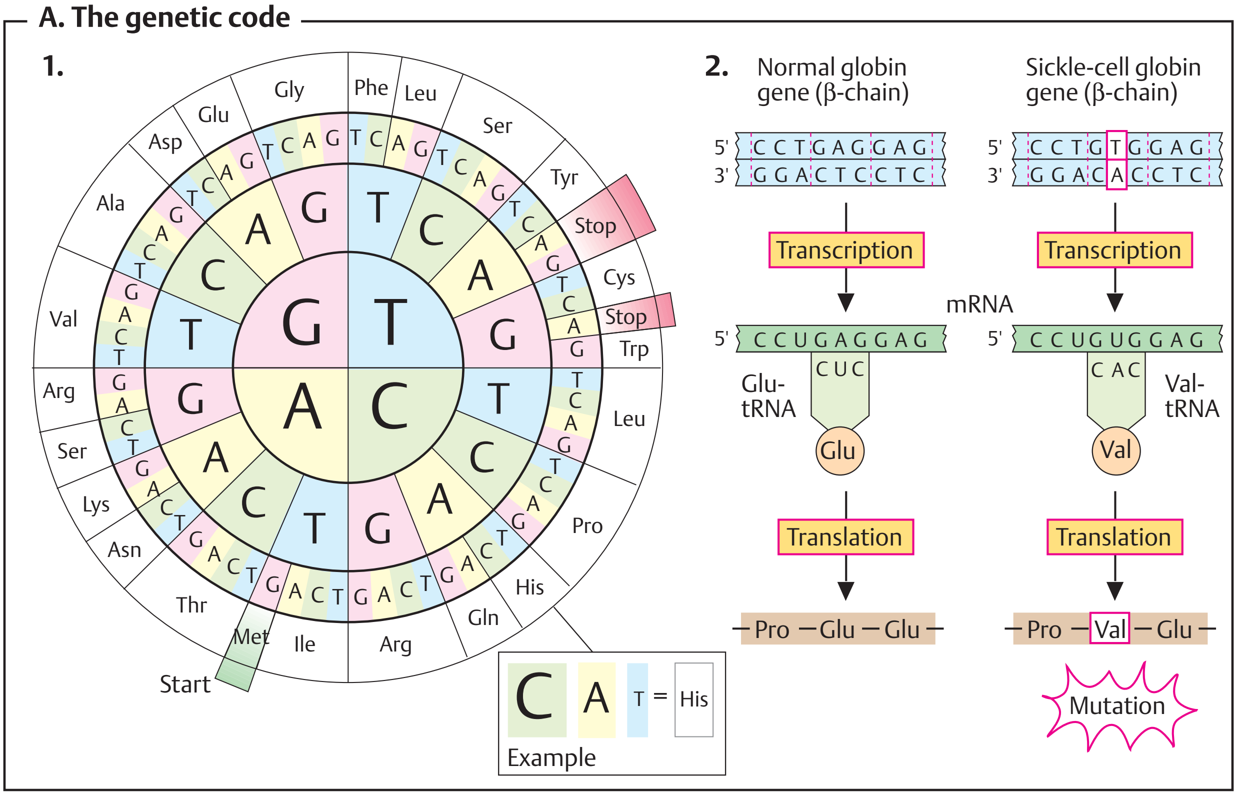
1.
G
T
A
C
Gly
Phe
Leu
Ser
Tyr
Stop
Cys
Stop
Trp
Leu
Pro
His
Gln
Arg
Ile
Met
Start
Thr
Asn
Lys
Ser
Arg
Val
Ala
Asp
Glu
C
A
T = His
Example
2.
Normal globin gene (β-chain)
Sickle-cell globin gene (β-chain)
5' C C T G A G G A G
3' G G A C T C C T C
5' C C T G T G G A G
3' G G A C A C C T C
Transcription
Transcription
mRNA
5' C C U G A G G A G
5' C C U G U G G A G
C U C
Glu-tRNA
Glu
C A C
Val-tRNA
Val
Translation
Translation
Pro – Glu – Glu
Pro – Val – Glu
Mutation

B. Amino acid activation

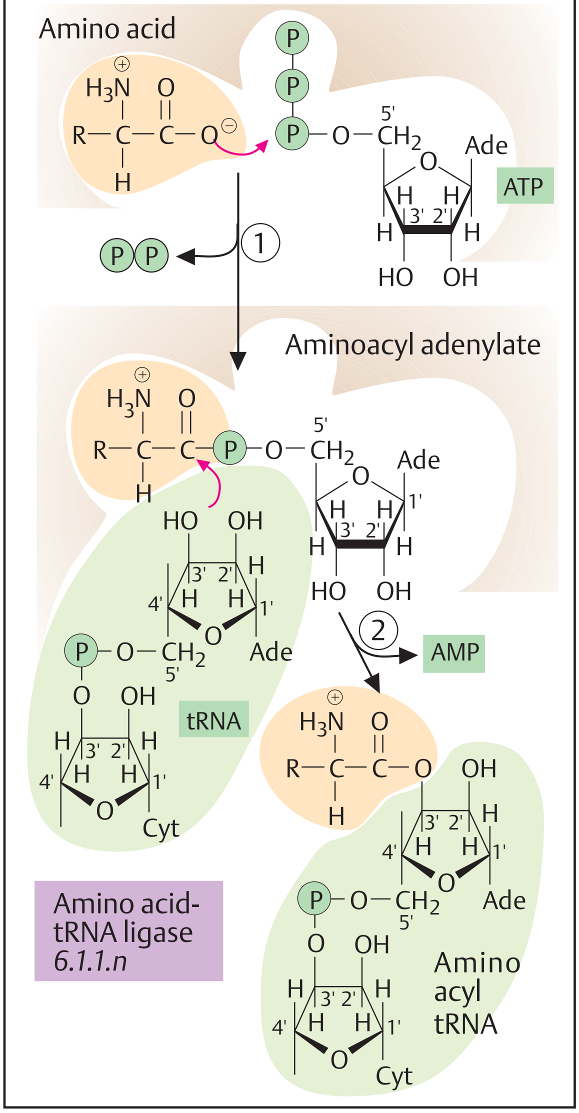
Amino acid
$H_3N^{\oplus}$
R–C–C–O$^{\ominus}$
P
P
P
O–CH_2
5'
Ade
ATP
HO OH
①
P P
Aminoacyl adenylate
HO OH
tRNA
Cyt
②
AMP
Amino acyl tRNA
Amino acid-tRNA ligase 6.1.1.n

C. Asp-tRNA-Ligase (Dimer)

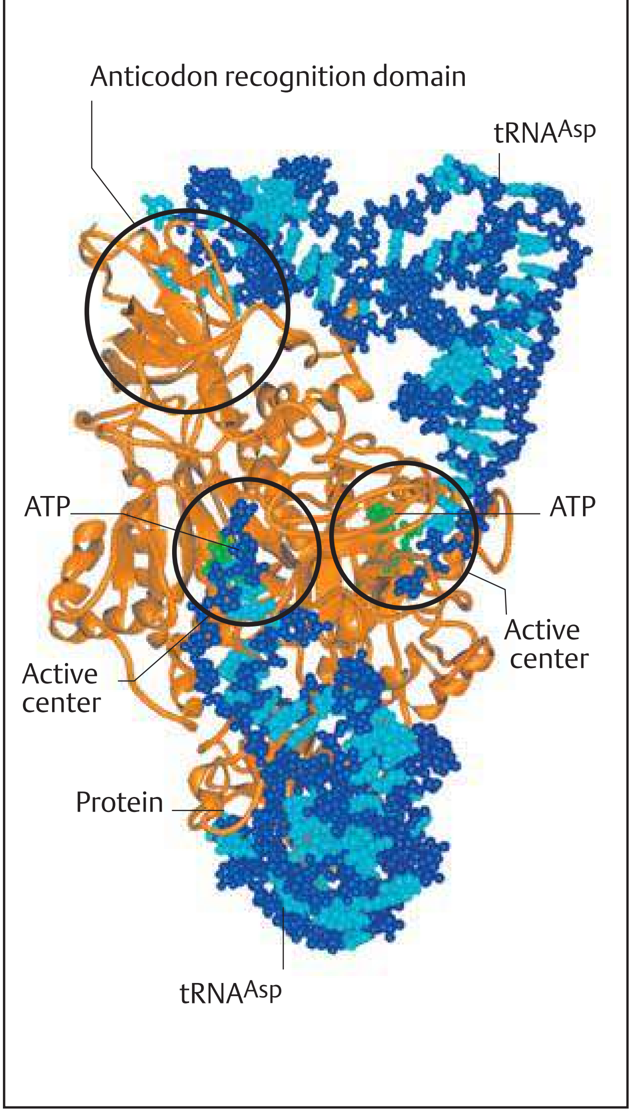
Anticodon recognition domain
tRNA^{Asp}
ATP
ATP
Active center
Active center
Protein
tRNA^{Asp}

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Translation I: initiation

Like amino acid activation (see p. 248), protein biosynthesis (**translation**) takes place in the cytoplasm. It is catalyzed by complex nucleoprotein particles, the **ribosomes**, and mainly requires GTP to cover its energy requirements.

A. Structure of the eukaryotic ribosome ◑

Ribosomes consist of two subunits of different size, made up of **ribosomal RNA (rRNA)** and nearly 80 **proteins** (the number of proteins applies to rat liver ribosomes). It is customary to give the sedimentation coefficients (see p. 200) of ribosomes and their components instead of their masses. For example, the eukaryotic ribosome has a sedimentation coefficient of 80 Svedberg units (80 S), while the sedimentation coefficients of its subunits are 40 S and 60 S (S values are not additive).

The smaller 40 S subunit consists of one molecule of 18 S rRNA and 33 protein molecules. The larger 60 S subunit contains three types of rRNA with sedimentation coefficients of 5 S, 5.8 S, and 28 S and 47 proteins. In the presence of mRNA, the subunits assemble to form the complete ribosome, with a mass about 650 times larger than that of a hemoglobin molecule.

The arrangement of the individual components of a ribosome has now been determined for prokaryotic ribosomes. It is known that filamentous mRNA passes through a cleft between the two subunits near the characteristic “horn” on the small subunit. tRNAs also bind near this site. The illustration shows the size of a tRNA molecule for comparison.

Prokaryotic ribosomes have a similar structure, but are somewhat smaller than those of eukaryotes (sedimentation coefficient 70 S for the complete ribosome, 30 S and 50 S for the subunits). Mitochondrial and chloroplast ribosomes are comparable to prokaryotic ones.

B. Polysomes ◑

In cells that are carrying out intensive protein synthesis, ribosomes are often found in a linear arrangement like a string of pearls; these are known as **polysomes**. This arrangement arises because several ribosomes are translating a single mRNA molecule simultaneously. The ribosome first binds near the **start codon** (AUG; see p. 248) at the 5′ end of the mRNA (top). During translation, the ribosome moves in the direction of the 3′ end until it reaches a **stop codon** (UAA, UAG, or UGA). At this point, the newly synthesized chain is released, and the ribosome dissociates again into its two subunits.

C. Initiation of translation in E. coli ◑

Protein synthesis in prokaryotes is in principle the same as in eukaryotes. However, as the process is simpler and has been better studied in prokaryotes, the details involved in translation are discussed here and on p. 252 using the example of the bacterium *Escherichia coli.*

The first phase of translation, **initiation**, involves several steps. First, two proteins, **initiation factors** IF–1 and IF–3, bind to the 30 S subunit (**1**). Another factor, IF–2, binds as a complex with GTP (**2**). This allows the subunit to associate with the mRNA and makes it possible for a special tRNA to bind to the start codon (**3**). In prokaryotes, this starter tRNA carries the substituted amino acid *N-formylmethionine* (fMet). In eukaryotes, it carries an unsubstituted *methionine*. Finally, the 50 S subunit binds to the above complex (**4**). During steps 3 and 4, the initiation factors are released again, and the GTP bound to IF–2 is hydrolyzed to GDP and P_i.

In the **70 S initiation complex**, formylmethionine tRNA is initially located at a binding site known as the **peptidyl site (P)**. A second binding site, the **acceptor site (A)**, is not yet occupied during this phase of translation. Sometimes, a third tRNA binding site is defined as an *exit site (E),* from which uncharged tRNAs leave the ribosome again (see p. 252; not shown).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Structure of eukaryotic ribosomes

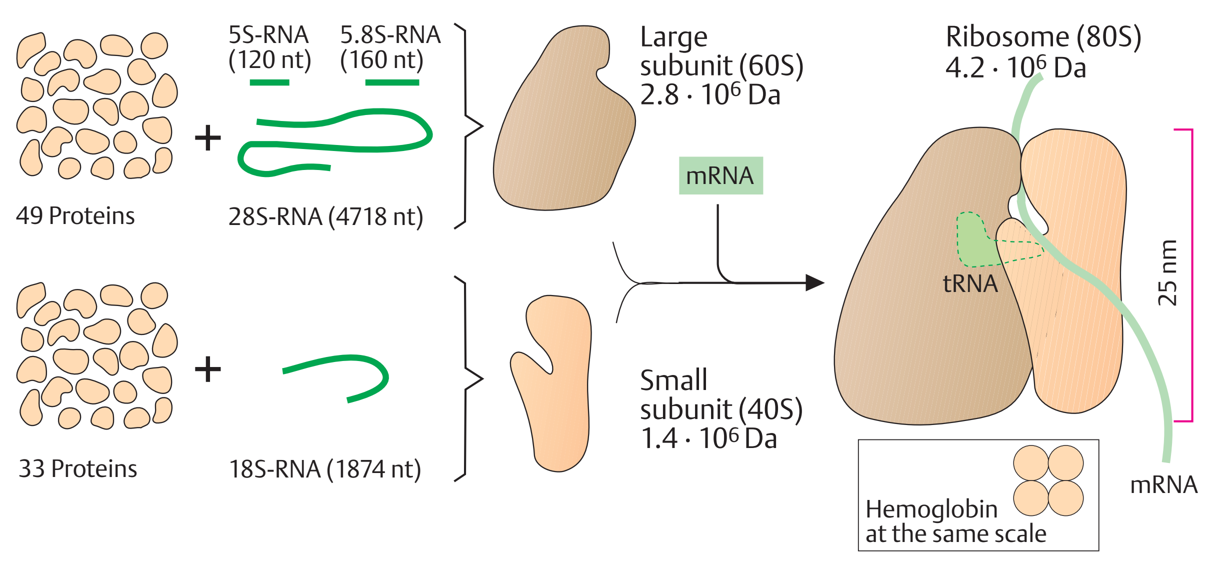
49 Proteins
5S-RNA (120 nt)
5.8S-RNA (160 nt)
28S-RNA (4718 nt)
+
Large subunit (60S) 2.8 · 10⁶ Da
33 Proteins
+
18S-RNA (1874 nt)
Small subunit (40S) 1.4 · 10⁶ Da
mRNA
Ribosome (80S) 4.2 · 10⁶ Da
tRNA
25 nm
mRNA
Hemoglobin at the same scale

B. Polysome

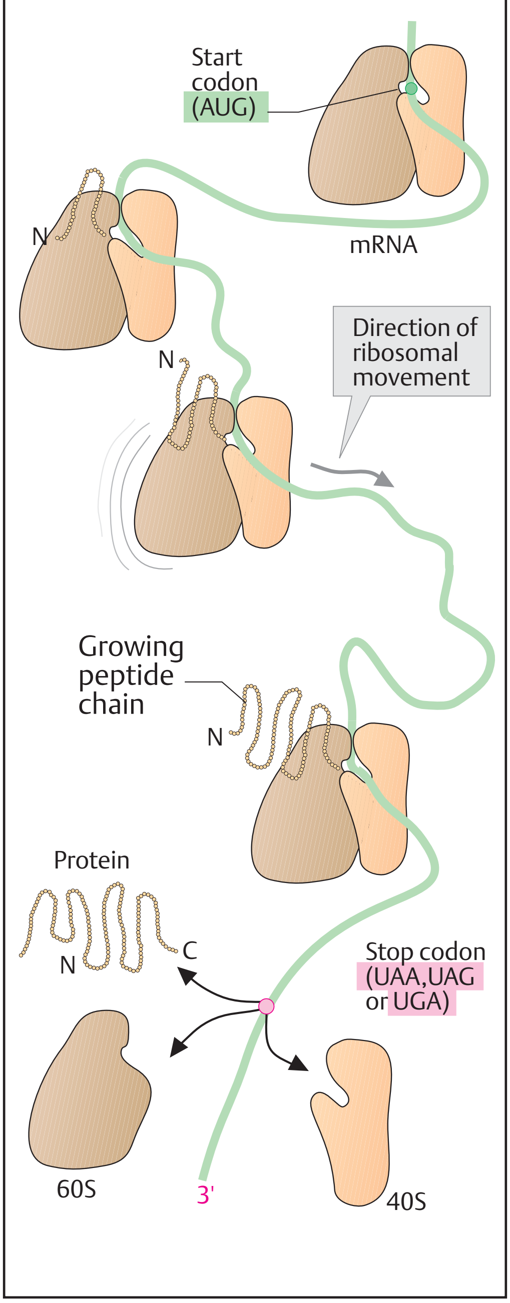
Start codon (AUG)
N
mRNA
N
Direction of ribosomal movement
Growing peptide chain
N
Protein
N
C
Stop codon (UAA,UAG or UGA)
60S
3'
40S

C. Initiation of translation in *E. coli*

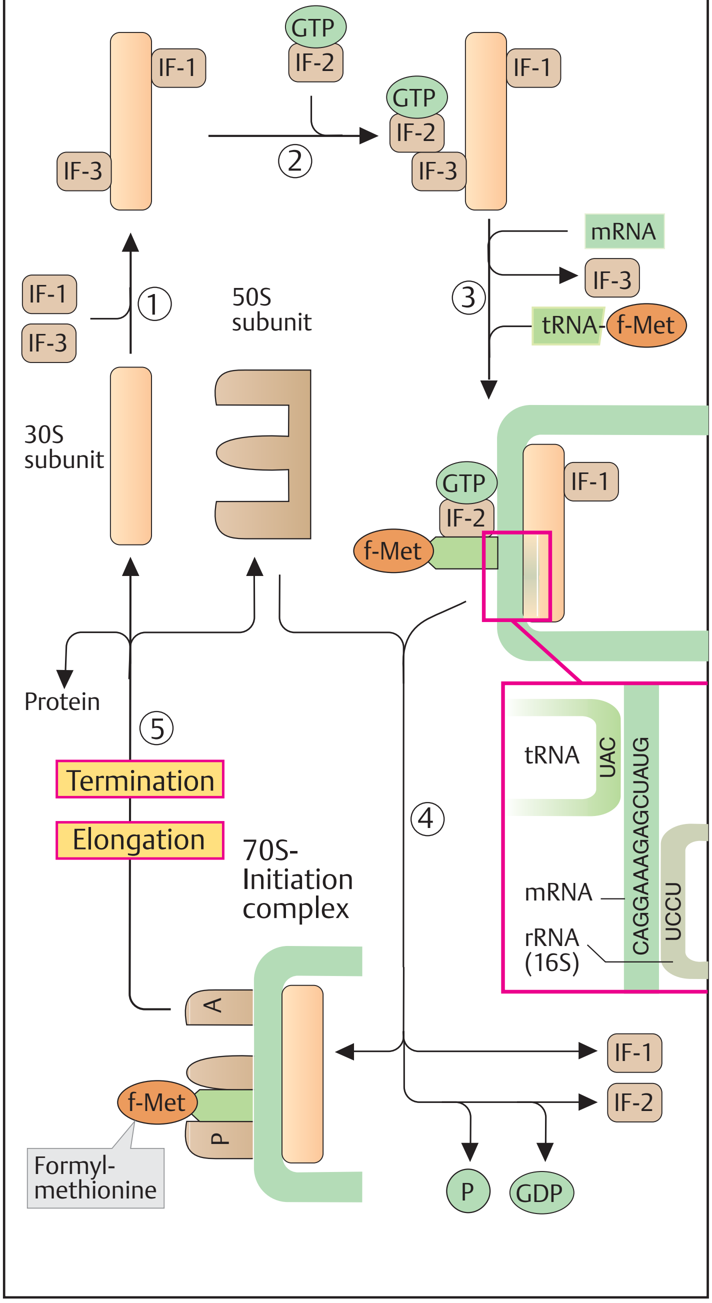
IF-1
IF-3
GTP
IF-2
②
GTP
IF-2
IF-1
IF-3
mRNA
IF-3
③
tRNA-f-Met
IF-1
IF-3
①
50S subunit
30S subunit
GTP
IF-2
IF-1
f-Met
Protein
⑤
Termination
Elongation
70S-Initiation complex
④
tRNA
UAC
mRNA
CAGGAAAGAGCUAUG
rRNA (16S)
UCCU
A
f-Met
P
Formyl-methionine
IF-1
IF-2
P
GDP

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Translation II: elongation and termination

After translation has been initiated (see p. 250), the peptide chain is extended by the addition of further amino acid residues (**elongation**) until the ribosome reaches a stop codon on the mRNA and the process is interrupted (**termination**).

A. Elongation and termination of protein biosynthesis in E. coli ◑

Elongation can be divided into four phases:

[1] **Binding of aminoacyl tRNA.** First, the peptidyl site (P) of the ribosome is occupied by a tRNA that carries at its 3′ end the complete peptide chain formed up to this point (top left). A second tRNA, loaded with the next amino acid (Val–tRNA^{Val} in the example shown), then binds via its complementary anticodon (see p. 82) to the mRNA codon exposed at the acceptor site (in this case GUG). The tRNA binds as a complex with a GTP-containing protein, the *elongation factor Tu* (EF–Tu) (**1a**). It is only after the bound GTP has been hydrolyzed to GDP and phosphate that EF–Tu dissociates again (**1b**). As the binding of the tRNA to the mRNA is still loose before this, GTP hydrolysis acts as a delaying factor, making it possible to check whether the correct tRNA has been bound. A further protein, the *elongation factor Ts* (EF-Ts), later catalyzes the exchange of GDP for GTP and in this way regenerates the EF–Tu GTP complex. EF-Tu is related to the G proteins involved in signal transduction (see p. 384).

[2] **Synthesis of the peptide bond** takes place in the next step. Ribosomal *peptidyltransferase* catalyzes (without consumption of ATP or GTP) the transfer of the peptide chain from the tRNA at the P site to the NH$_2$ group of the amino acid residue of the tRNA at the A site. The ribosome's peptidyltransferase activity is not located in one of the ribosomal proteins, but in the 28 S rRNA. Catalytically active RNAs of this type are known as *ribozymes* (cf. p. 246). It is thought that the few surviving ribozymes are remnants of the *"RNA world"*—an early phase of evolution in which proteins were not as important as they are today.

[3] After the transfer of the growing peptide to the A site, the free tRNA at the P site dissociates and another GTP–containing elongation factor (EF-G GTP) binds to the ribosome. Hydrolysis of the GTP in this factor provides the energy for **translocation** of the ribosome. During this process, the ribosome moves three bases along the mRNA in the direction of the 3′ end. The tRNA carrying the peptide chain is stationary relative to the mRNA and reaches the ribosome's P site during translocation, while the next mRNA codon (in this case GUG) appears at the A site.

[4] The uncharged Val-tRNA then dissociates from the E site. The ribosome is now ready for the next elongation cycle.

When one of the three stop codons (UAA, UAG, or UGA) appears at the A site, **termination** starts.

[5] There are no complementary tRNAs for the stop codons. Instead, two *releasing factors* bind to the ribosome. One of these factors (RF–1) catalyzes hydrolytic cleavage of the ester bond between the tRNA and the C–terminus of the peptide chain, thereby releasing the protein.

[6] Hydrolysis of GTP by factor RF–3 supplies the energy for the dissociation of the whole complex into its component parts.

Energy requirements in protein synthesis are high. Four energy-rich phosphoric acid anhydride bonds are hydrolyzed for each amino acid residue. Amino acid activation uses up two energy-rich bonds per amino acid (ATP → AMP + PP; see p. 248), and two GTPs are consumed per elongation cycle. In addition, initiation and termination each require one GTP per chain.

Further information

In eukaryotic cells, the number of initiation factors is larger and initiation is therefore more complex than in prokaryotes. The cap at the 5′ end of mRNA and the polyA tail (see p. 246) play important parts in initiation. However, the elongation and termination processes are similar in all organisms. The individual steps of bacterial translation can be inhibited by antibiotics (see p. 254).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Elongation and termination of protein biosynthesis in *E. coli*

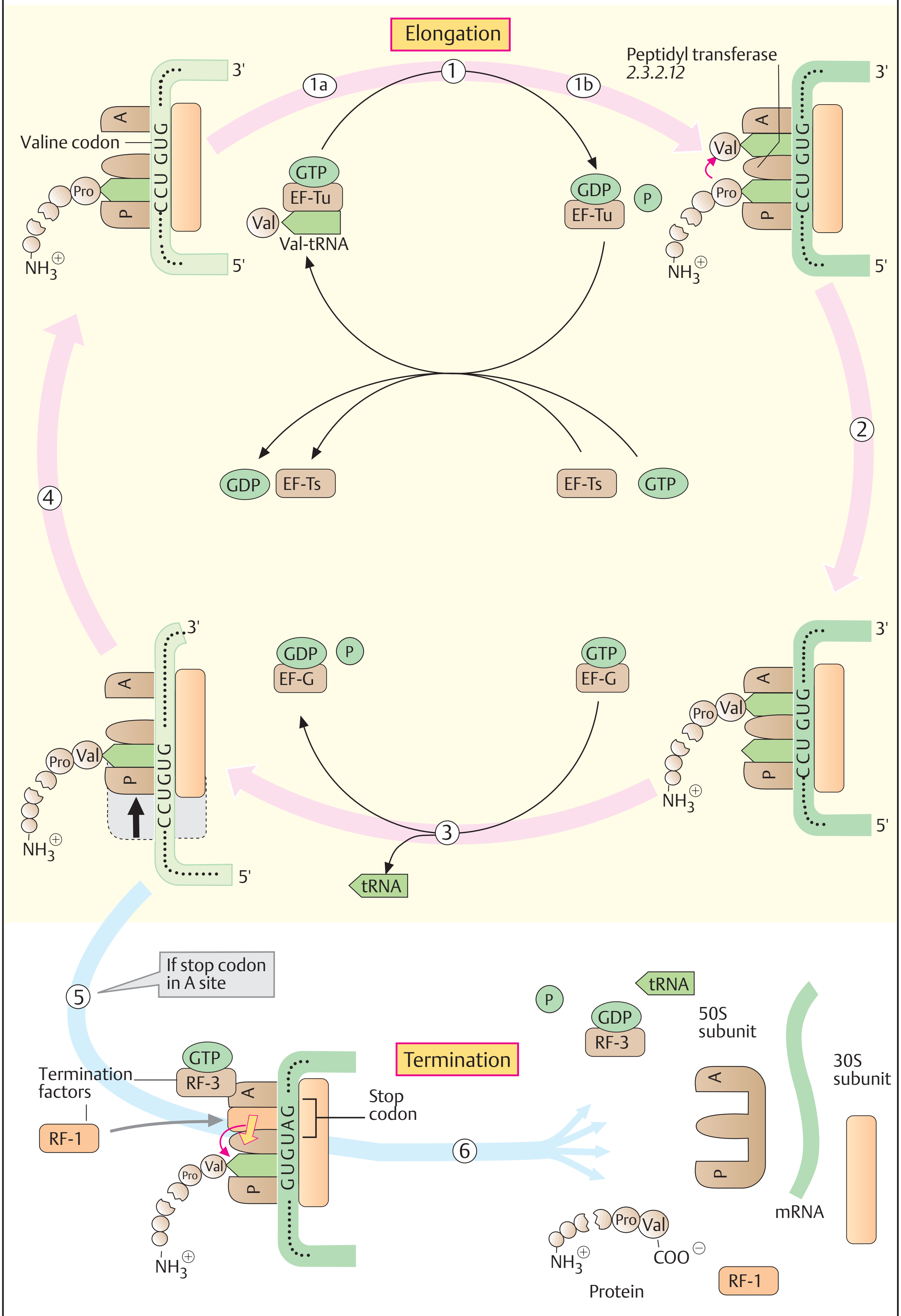
Elongation
1
1a
1b
Peptidyl transferase
2.3.2.12
3'
5'
A
P
Valine codon
CCU GUG
Pro
Val
NH3⊕
GTP
EF-Tu
Val-tRNA
GDP
EF-Tu
P
2
GDP
EF-Ts
EF-Ts
GTP
4
3
GDP
P
EF-G
GTP
EF-G
tRNA
CCUGUG
5
If stop codon in A site
Termination factors
GTP
RF-3
RF-1
GUGUAG
Stop codon
6
Termination
P
tRNA
GDP
RF-3
50S subunit
30S subunit
mRNA
Protein
COO⊖
RF-1

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Antibiotics

A. Antibiotics: overview ◑

Antibiotics are substances which, even at low concentrations, inhibit the growth and reproduction of bacteria and fungi. The treatment of infectious diseases would be inconceivable today without antibiotics. Substances that only restrict the reproduction of bacteria are described as having *bacteriostatic* effects (or *fungistatic* for fungi). If the target cells are killed, then the term *bactericidal* (or *fungicidal*) is used. Almost all antibiotics are produced by microorganisms—mainly bacteria of the genus *Streptomyces* and certain fungi. However, there are also synthetic antibacterial substances, such as sulfonamides and gyrase inhibitors.

A constantly increasing problem in antibiotic treatment is the development of resistant pathogens that no longer respond to the drugs available. The illustration shows a few of the therapeutically important antibiotics and their sites of action in the bacterial metabolism.

Substances known as **intercalators**, such as *rifamycin* and *actinomycin D* (bottom) are deposited in the DNA double helix and thereby interfere with replication and transcription (**B**). As DNA is the same in all cells, intercalating antibiotics are also toxic for eukaryotes, however. They are therefore only used as cytostatic agents (see p. 402). Synthetic inhibitors of DNA topoisomerase II (see p. 240), known as **gyrase inhibitors** (center), restrict replication and thus bacterial reproduction.

A large group of antibiotics attack bacterial ribosomes. These **inhibitors of translation** (left) include the *tetracyclines*—broad-spectrum antibiotics that are effective against a large number of different pathogens. The *aminoglycosides*, of which *streptomycin* is the best-known, affect all phases of translation. *Erythromycin* impairs the normal functioning of the large ribosomal subunit. *Chloramphenicol,* one of the few natural nitro compounds, inhibits ribosomal peptidyltransferase. Finally, *puromycin* mimics an aminoacyl tRNA and therefore leads to premature interruption of elongation.

The **β–lactam antibiotics** (bottom right) are also frequently used. The members of this group, the *penicillins* and *cephalosporins,* are synthesized by fungi and have a reactive β-lactam ring. They are mainly used against Gram-positive pathogens, in which they inhibit cell wall synthesis (**C**).

The first synthetic antibiotics were the **sulfonamides** (right). As analogues of p–aminobenzoic acid, these affect the synthesis of *folic acid,* an essential precursor of the coenzyme THF (see p. 108). **Transport antibiotics** (top center) have the properties of ion channels (see p. 222). When they are deposited in the plasma membrane, it leads to a loss of ions that damages the bacterial cells.

B. Intercalators ○

The effects of intercalators (see also p. 262) are illustrated here using the example of the **daunomycin–DNA complex**, in which two daunomycin molecules (red) are inserted in the double helix (blue). The antibiotic's ring system inserts itself between G/C base pairs (bottom), while the sugar moiety occupies the minor groove in the DNA (above). This leads to a localized change of the DNA conformation that prevents replication and transcription.

C. Penicillin as a "suicide substrate" ○

The site of action in the β–lactam antibiotics is *muramoylpentapeptide carboxypeptidase,* an enzyme that is essential for cross–linking of bacterial cell walls. The antibiotic resembles the substrate of this enzyme (a peptide with the C-terminal sequence D-Ala–D–Ala) and is therefore reversibly bound in the active center. This brings the β–lactam ring into proximity with an essential serine residue of the enzyme. Nucleophilic substitution then results in the formation of a stable covalent bond between the enzyme and the inhibitor, blocking the active center (see p. 96). In dividing bacteria, the loss of activity of the enzyme leads to the formation of unstable cell walls and eventually death.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Antibiotics: overview

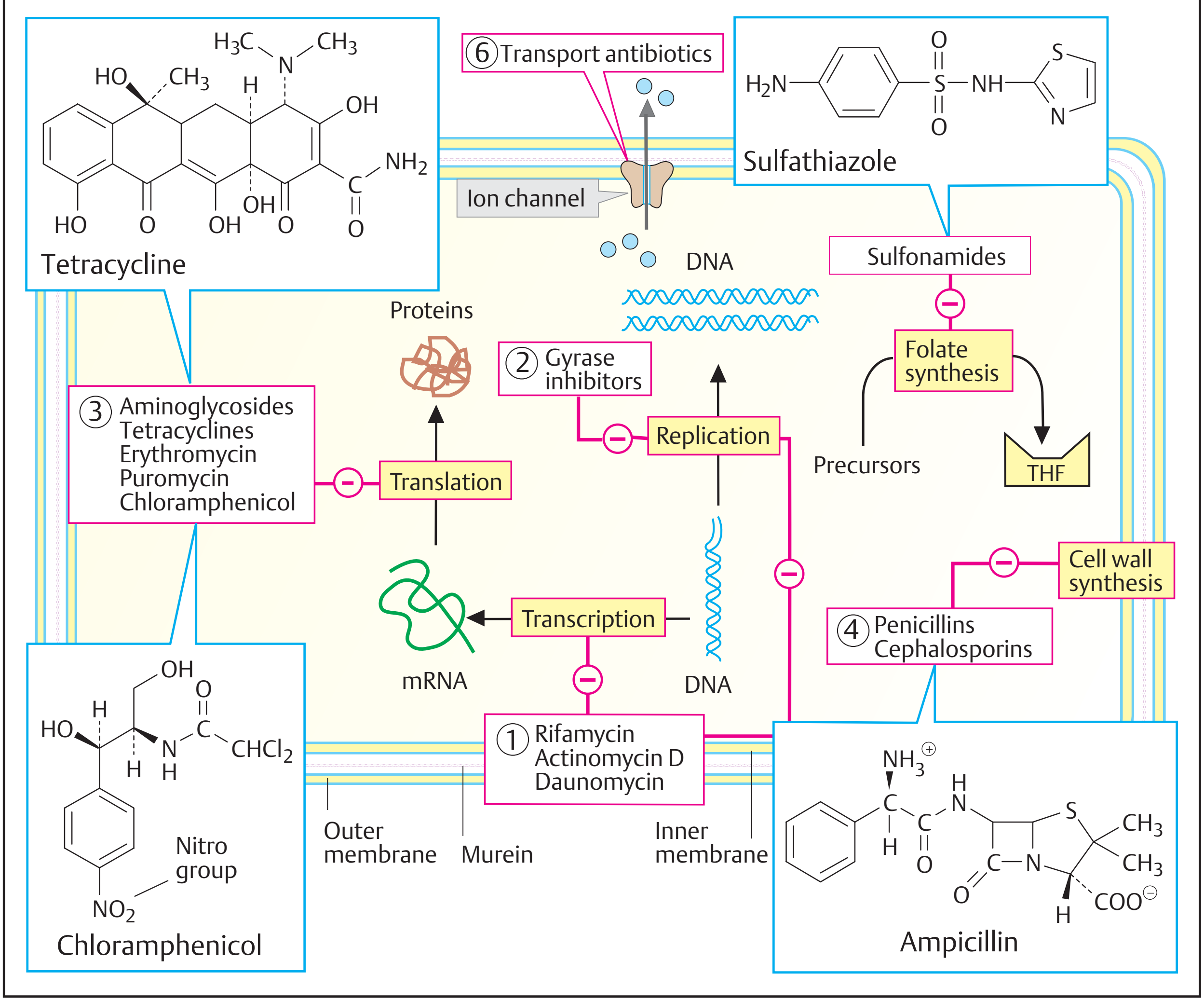
Tetracycline
⑥ Transport antibiotics
Ion channel
Sulfathiazole
Sulfonamides
DNA
Proteins
② Gyrase inhibitors
Folate synthesis
Precursors
THF
Replication
③ Aminoglycosides
Tetracyclines
Erythromycin
Puromycin
Chloramphenicol
Translation
Cell wall synthesis
④ Penicillins
Cephalosporins
Transcription
mRNA
DNA
① Rifamycin
Actinomycin D
Daunomycin
Outer membrane
Murein
Inner membrane
Nitro group
Chloramphenicol
Ampicillin

B. Intercalators

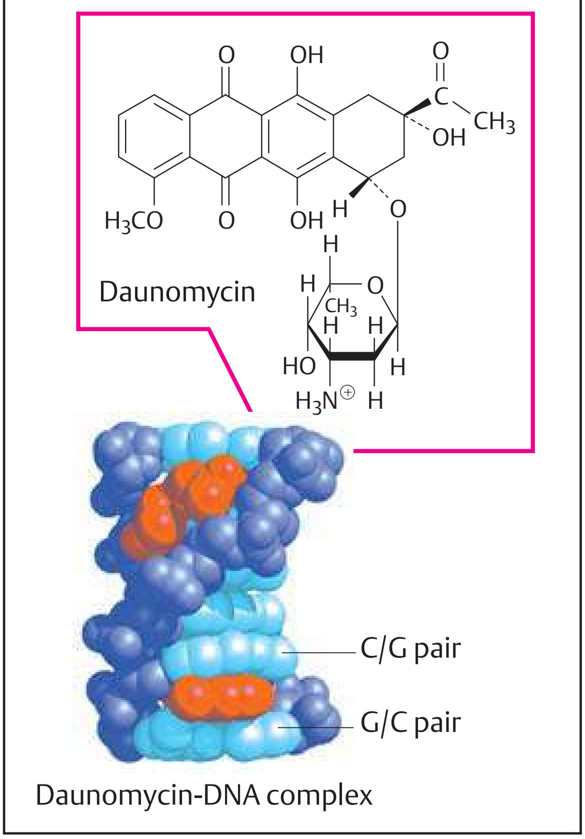
Daunomycin
C/G pair
G/C pair
Daunomycin-DNA complex

C. Penicillin as “suicide substrate”

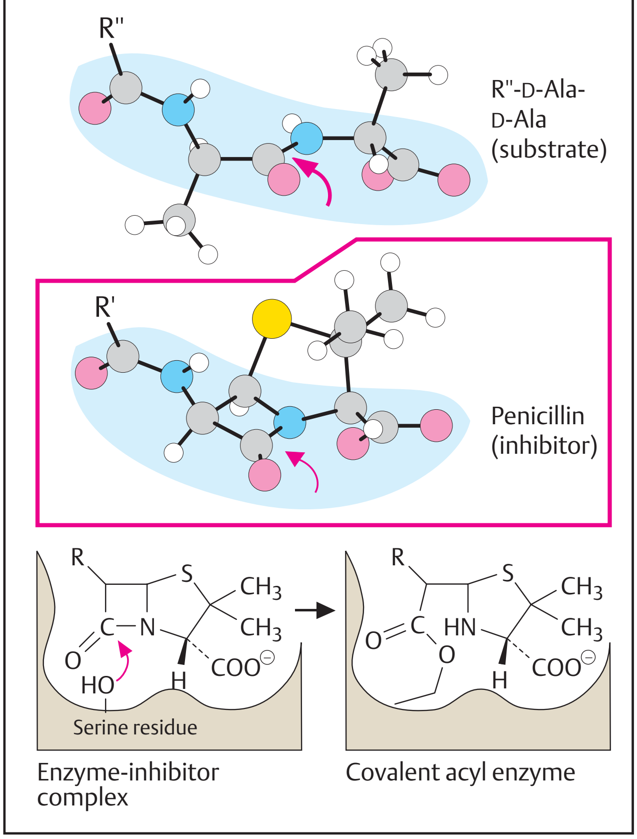
R''
R''-D-Ala-D-Ala (substrate)
R'
Penicillin (inhibitor)
R
Serine residue
Enzyme-inhibitor complex
Covalent acyl enzyme

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Mutation and repair

Genetic information is set down in the base sequence of DNA. Changes in the DNA bases or their sequence therefore have *mutagenic* effects. Mutagens often also damage growth regulation in cells, and they are then also *carcinogenic* (see p. 400). Gene alterations (**mutations**) are one of the decisive positive factors in biological evolution. On the other hand, an excessive mutation frequency would threaten the survival of individual organisms or entire species. For this reason, every cell has **repair mechanisms** that eliminate most of the DNA changes arising from mutations (**C**).

A. Mutagenic agents ◑

Mutations can arise as a result of physical or chemical effects, or they can be due to accidental errors in DNA replication and recombination.

The principal physical mutagen is **ionizing radiation** (α, β, and γ radiation, X-rays). In cells, it produces **free radicals** (molecules with unpaired electrons), which are extremely reactive and can damage DNA. Short-wavelength **ultraviolet light** (UV light) also has mutagenic effects, mainly in skin cells (sunburn). The most common chemical change due to UV exposure is the formation of **thymine dimers**, in which two neighboring thymine bases become covalently linked to one another (**2**). This results in errors when the DNA is read during replication and transcription.

Only a few examples of the group of **chemical mutagens** are shown here. *Nitrous acid* (HNO_2; salt: nitrite) and *hydroxylamine* (NH_2OH) both deaminate bases; they convert cytosine to uracil and adenine to inosine.

Alkylating compounds carry reactive groups that can form covalent bonds with DNA bases (see also p. 402). *Methylnitrosamines* (**3**) release the reactive methyl cation (CH_3^+), which methylates OH and NH_2 groups in DNA. The dangerous carcinogen *benzo* [a]*pyrene* is an aromatic hydrocarbon that is only converted into the active form in the organism (**4**; see p. 316). Multiple hydroxylation of one of the rings produces a reactive epoxide that can react with NH_2 groups in guanine residues, for example. Free radicals of benzo[*a*]pyrene also contribute to its toxicity.

B. Effects ◑

Nitrous acid causes **point mutations** (**1**). For example, C is converted to U, which in the next replication pairs with A instead of G. The alteration thus becomes permanent. Mutations in which a number of nucleotides not divisible by three are inserted or removed lead to reading errors in whole segments of DNA, as they move the reading frame (**frameshift mutations**). This is shown in Fig. **2** using a simple example. From the inserted C onwards, the resulting mRNA is interpreted differently during translation, producing a completely new protein sequence.

C. Repair mechanisms ○

An important mechanism for the removal of DNA damage is **excision repair** (**1**). In this process, a specific *excision endonuclease* removes a complete segment of DNA on both sides of the error site. Using the sequence of the opposite strand, the missing segment is then replaced by a *DNA polymerase*. Finally, a *DNA ligase* closes the gaps again.

Thymine dimers can be removed by **photoreactivation** (**2**). A specific *photolyase* binds at the defect and, when illuminated, cleaves the dimer to yield two single bases again.

A third mechanism is **recombination repair** (**3**, shown in simplified form). In this process, the defect is omitted during replication. The gap is closed by shifting the corresponding sequence from the correctly replicated second strand. The new gap that results is then filled by polymerases and ligases. Finally, the original defect is corrected by excision repair as in Fig. **1** (not shown).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Mutagenic agents

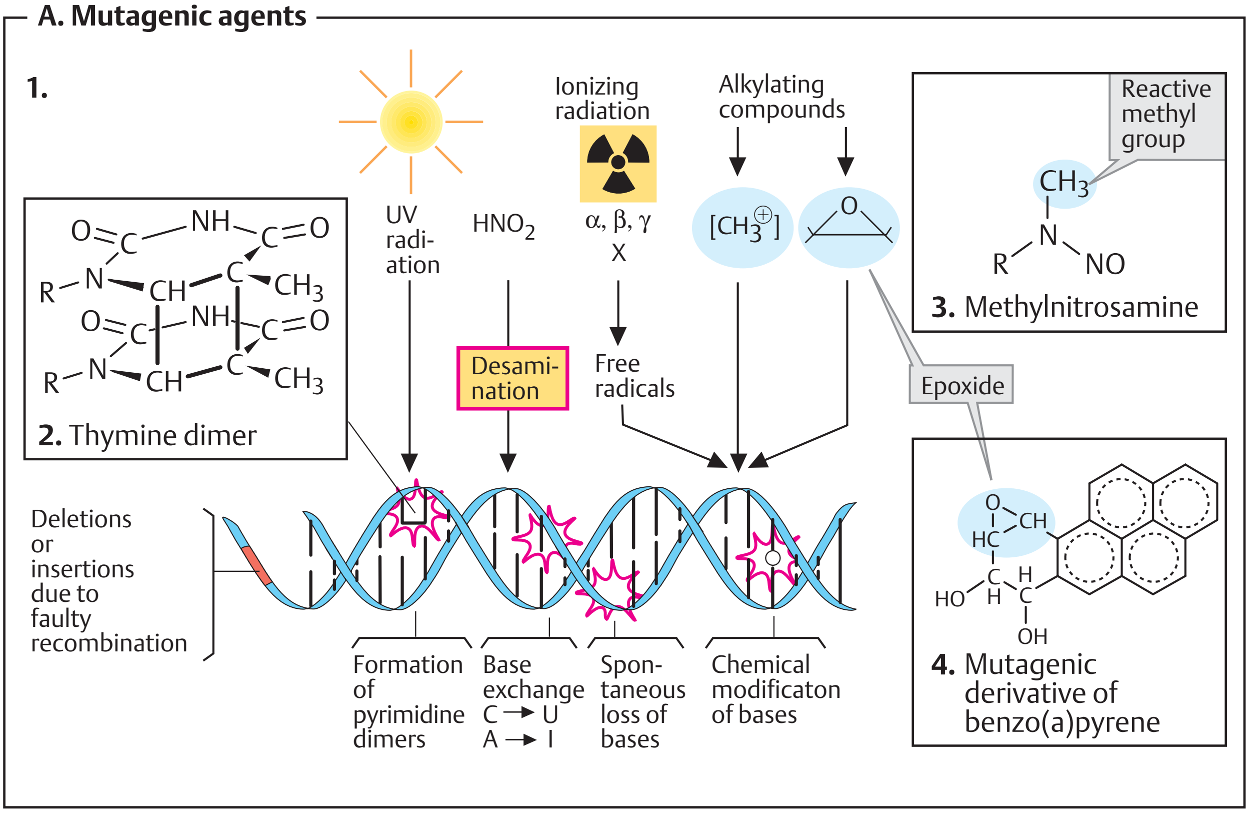
1.
UV radi-ation
HNO$_2$
Ionizing radiation
α, β, γ
X
Alkylating compounds
[CH$_3^{\oplus}$]
O
Desami-nation
Free radicals
Reactive methyl group
CH$_3$
N
R
NO
3. Methylnitrosamine
Epoxide
O=C—NH—C=O
R—N—CH—C—CH$_3$
O=C—NH—C=O
R—N—CH—C—CH$_3$
2. Thymine dimer
Deletions or insertions due to faulty recombination
Formation of pyrimidine dimers
Base exchange
C → U
A → I
Spon-taneous loss of bases
Chemical modificaton of bases
O
CH
HC
C
H
HO
H
C
OH
4. Mutagenic derivative of benzo(a)pyrene

B. Effects

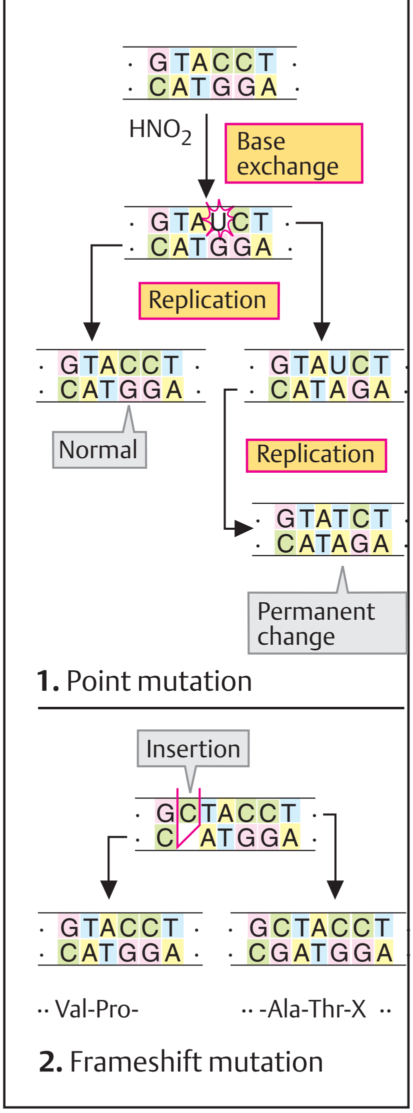
GTACCT
CATGGA
HNO$_2$
Base exchange
GTAUCT
CATGGA
Replication
GTACCT
CATGGA
Normal
GTAUCT
CATAGA
Replication
GTATCT
CATAGA
Permanent change
1. Point mutation
Insertion
GCTACCT
C ATGGA
GTACCT
CATGGA
GCTACCT
CGATGGA
·· Val-Pro-
·· -Ala-Thr-X ··
2. Frameshift mutation

C. Repair mechanisms

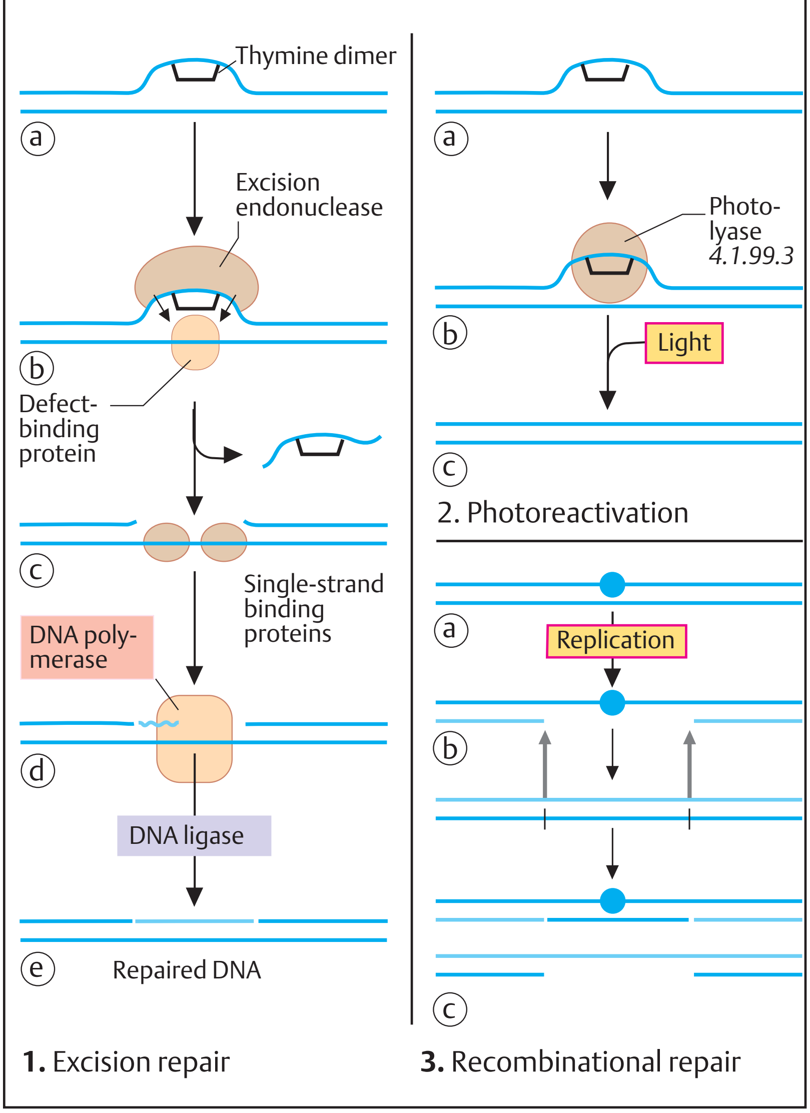
Thymine dimer
a
Excision endonuclease
b
Defect-binding protein
c
Single-strand binding proteins
DNA poly-merase
d
DNA ligase
e
Repaired DNA
1. Excision repair
a
Photo-lyase
4.1.99.3
b
Light
c
2. Photoreactivation
a
Replication
b
c
3. Recombinational repair

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

DNA cloning

The growth of molecular genetics since 1970 has mainly been based on the development and refinement of methods of analyzing and manipulating DNA. **Genetic engineering** has practical applications in many fields. For example, it has provided new methods of diagnosing and treating diseases, and it is now also possible to create targeted changes in specific characteristics of organisms. Since biological risks cannot be completely ruled out with these procedures, it is particularly important to act responsibly when dealing with genetic engineering. A short overview of important methods involved in genetic engineering is provided here and on the following pages.

A. Restriction endonucleases ◑

In many genetic engineering procedures, defined DNA fragments have to be isolated and then newly combined with other DNA segments. For this purpose, enzymes are used that can cut DNA and join it together again inside the cell. Of particular importance are *restriction endonucleases*—a group of bacterial enzymes that cleave the DNA double strand in a sequence–specific way. The numerous restriction enzymes known are named using abbreviations based on the organism from which they originate. The example used here is *Eco*RI, a nuclease isolated from the bacterium *Escherichia coli.*

Like many other restriction endonucleases, EcoRI cleaves DNA at the site of a *palindrome*—i. e., a short segment of DNA in which both the strand and counter-strand have the same sequence (each read in the 5′→3′ direction). In this case, the sequence is 5′-GAATTC-3′. *Eco*RI, a homodimer, cleaves the phosphoric acid diester bonds in both strands between G and A. This results in the formation of complementary *overhanging* or *"sticky" ends* (AATT), which are held together by base pairing. However, they are easily separated—e. g. by heating. When the fragments are cooled, the overhanging ends hybridize again in the correct arrangement. The cleavage sites can then be sealed again by a *DNA ligase*.

B. DNA cloning ○

Most DNA segments—e. g., genes—occur in very small quantities in the cell. To be able to work with them experimentally, a large number of identical copies (**"clones"**) first have to be produced. The classic procedure for cloning DNA takes advantage of the ability of bacteria to take up and replicate short, circular DNA fragments known as **plasmids**.

The segment to be cloned is first cut out of the original DNA using restriction endonucleases (see above; for the sake of simplicity, cleavage using *Eco*RI alone is shown here, but in practice two different enzymes are usually used). As a vehicle (*"vector"*), a plasmid is needed that has only *one Eco*RI cleavage site. The plasmid rings are first opened by cleavage with *Eco*RI and then mixed with the isolated DNA fragments. Since the fragment and the vector have the same overhanging ends, some of the molecules will hybridize in such a way that the fragment is incorporated into the vector DNA. When the cleavage sites are now closed again using *DNA ligase,* a newly combined *("recombinant") plasmid* arises.

By pretreating a large number of host cells, one can cause some of them to take up the plasmid (a process known as **transformation**) and replicate it along with their own genome when reproducing. To ensure that only host bacteria that contain the plasmid replicate, plasmids are used that give the host *resistance* to a particular antibiotic. When the bacteria are incubated in the presence of this antibiotic, only the cells containing the plasmid will replicate. The plasmid is then isolated from these cells, cleaved with *Eco*RI again, and the fragments are separated using agarose gel electrophoresis (see p. 262). The desired fragment can be identified using its size and then extracted from the gel and used for further experiments.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Restriction endonucleases

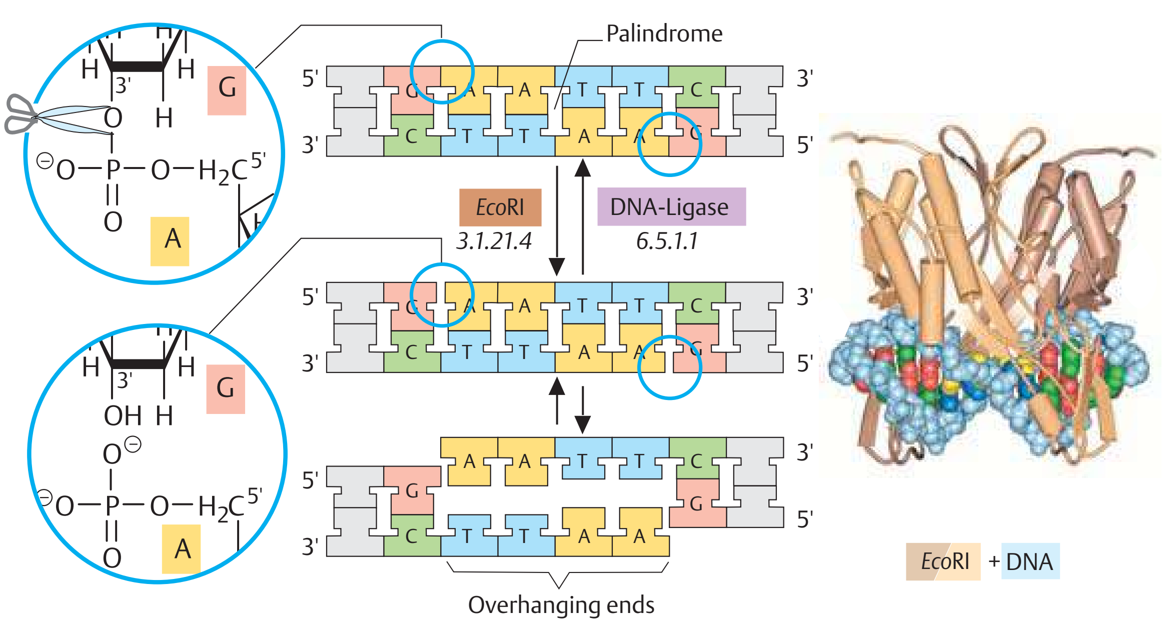
H
3'
H
G
O
H
O
P
O
$H_2C^{5'}$
O
A
H
3'
H
G
OH H
O
O
P
O
$H_2C^{5'}$
O
A
Palindrome
5'
3'
3'
5'
G
A
A
T
T
C
C
T
T
A
A
G
EcoRI
3.1.21.4
DNA-Ligase
6.5.1.1
Overhanging ends
EcoRI + DNA

B. DNA cloning

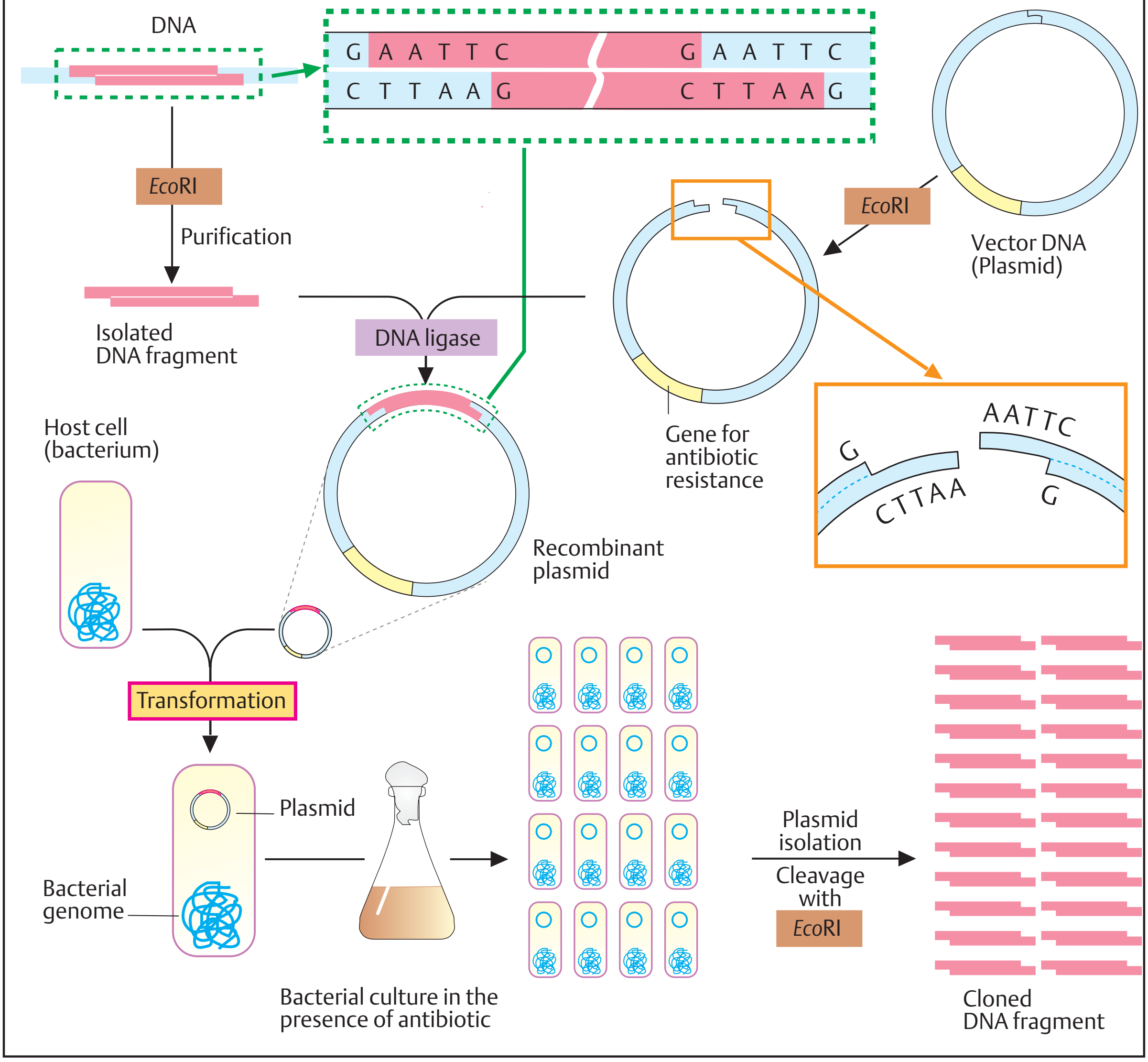
DNA
G A A T T C
C T T A A G
G A A T T C
C T T A A G
EcoRI
Purification
Isolated
DNA fragment
DNA ligase
EcoRI
Vector DNA
(Plasmid)
Gene for
antibiotic
resistance
AATTC
G
CTTAA
G
Recombinant
plasmid
Host cell
(bacterium)
Transformation
Plasmid
Bacterial
genome
Bacterial culture in the
presence of antibiotic
Plasmid
isolation
Cleavage
with
EcoRI
Cloned
DNA fragment

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

DNA sequencing

A. Gene libraries ○

It is often necessary in genetic engineering to isolate a DNA segment when its details are not fully known—e. g., in order to determine its nucleotide sequence. In this case, one can use what are known as **DNA libraries**. A DNA library consists of a large number of *vector DNA molecules* containing different fragments of *foreign DNA*. For example, it is possible to take all of the mRNA molecules present in a cell and transcribe them into DNA. These DNA fragments (known as copy DNA or **cDNA**) are then randomly introduced into vector molecules.

A library of genomic DNA can be established by cleaving the total DNA from a cell into small fragments using restriction endonucleases (see p. 258), and then incorporating these into vector DNA. Suitable vectors for gene libraries include **bacteriophages**, for example ("phages" for short). Phages are viruses that only infect bacteria and are replicated by them (see p. 404). Gene libraries have the advantage that they can be searched for specific DNA segments, using hybridization with oligonucleotides.

The first step is to strongly dilute a small part of the library (10^5–10^6 phages in a small volume), mix it with host bacteria, and plate out the mixture onto nutrient medium. The bacteria grow and form a continuous cloudy layer of cells. Bacteria infected by phages grow more slowly. In their surroundings, the bacterial "lawn" is less dense, and a clearer circular zone known as a **plaque** forms. The bacteria in this type of plaque exclusively contain the offspring of a single phage from the library.

The next step is to make an impression of the plate on a plastic foil, which is then heated. This causes the phage DNA to adhere to the foil. When the foil is incubated with a DNA fragment that hybridizes to the DNA segment of interest (a **gene probe**), the probe binds to the sites on the imprint at which the desired DNA is attached. Binding of the gene probe can be detected by prior radioactive or other labeling of the probe. Phages from the positive plaques in the original plate are then isolated and replicated. Restriction cleavage finally provides large amounts of the desired DNA.

B. Sequencing of DNA ○

The nucleotide sequence of DNA is nowadays usually determined using the so-called **chain termination method**. In single-strand sequencing, the DNA fragment (**a**) is cloned into the DNA of **phage M13** (see p. 404), from which the coded single strand can be easily isolated. This is *hybridized* with a **primer**—a short, synthetically produced DNA fragment that binds to 3′ end of the introduced DNA segment (**b**).

Based on this hybrid, the missing second strand can now be generated in the test tube by adding the four *deoxyribonucleoside triphosphates* (*dNTP*) and a suitable *DNA polymerase* (**c**). The trick lies in also adding small amounts of *dideoxynucleoside triphosphates (ddNTP)*. Incorporating a ddNTP leads to the *termination of second-strand synthesis*. This can occur whenever the corresponding dNTP ought to be incorporated. The illustration shows this in detail using the example of ddGTP. In this case, fragments are obtained that each include the primer plus three, six, eight, 13, or 14 additional nucleotides. *Four separate reactions*, each with a different ddNTP, are carried out (**c**), and the products are placed side by side on a supporting material. The fragments are then separated by *gel electrophoresis* (see p. 76), in which they move in relation to their length.

Following *visualization* (**d**), the sequence of the fragments in the individual lanes is simply read from bottom to top (**e**) to directly obtain the nucleotide sequence. A detail from such a sequencing gel and the corresponding protein sequence are shown in Fig. **2**.

In a more modern procedure, the four ddNTPs are covalently marked with fluorescent dyes, which produce a different color for each ddNTP on laser illumination. This allows the sequence in which the individual fragments appear at the lower end of the gel to be continuously recorded and directly stored in digital form.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Gene libraries

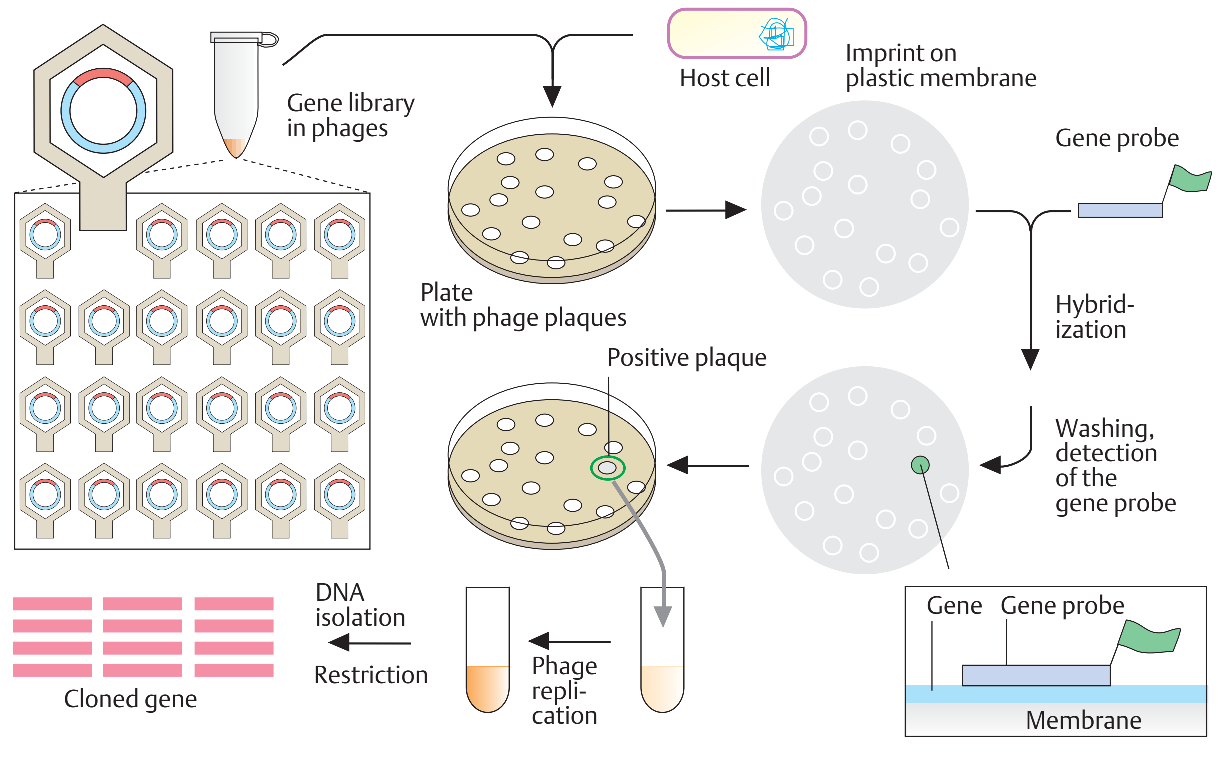
Gene library
in phages
Host cell
Imprint on
plastic membrane
Gene probe
Plate
with phage plaques
Hybrid-
ization
Positive plaque
Washing,
detection
of the
gene probe
DNA
isolation
Restriction
Phage
repli-
cation
Cloned gene
Gene
Gene probe
Membrane

B. Sequencing of DNA

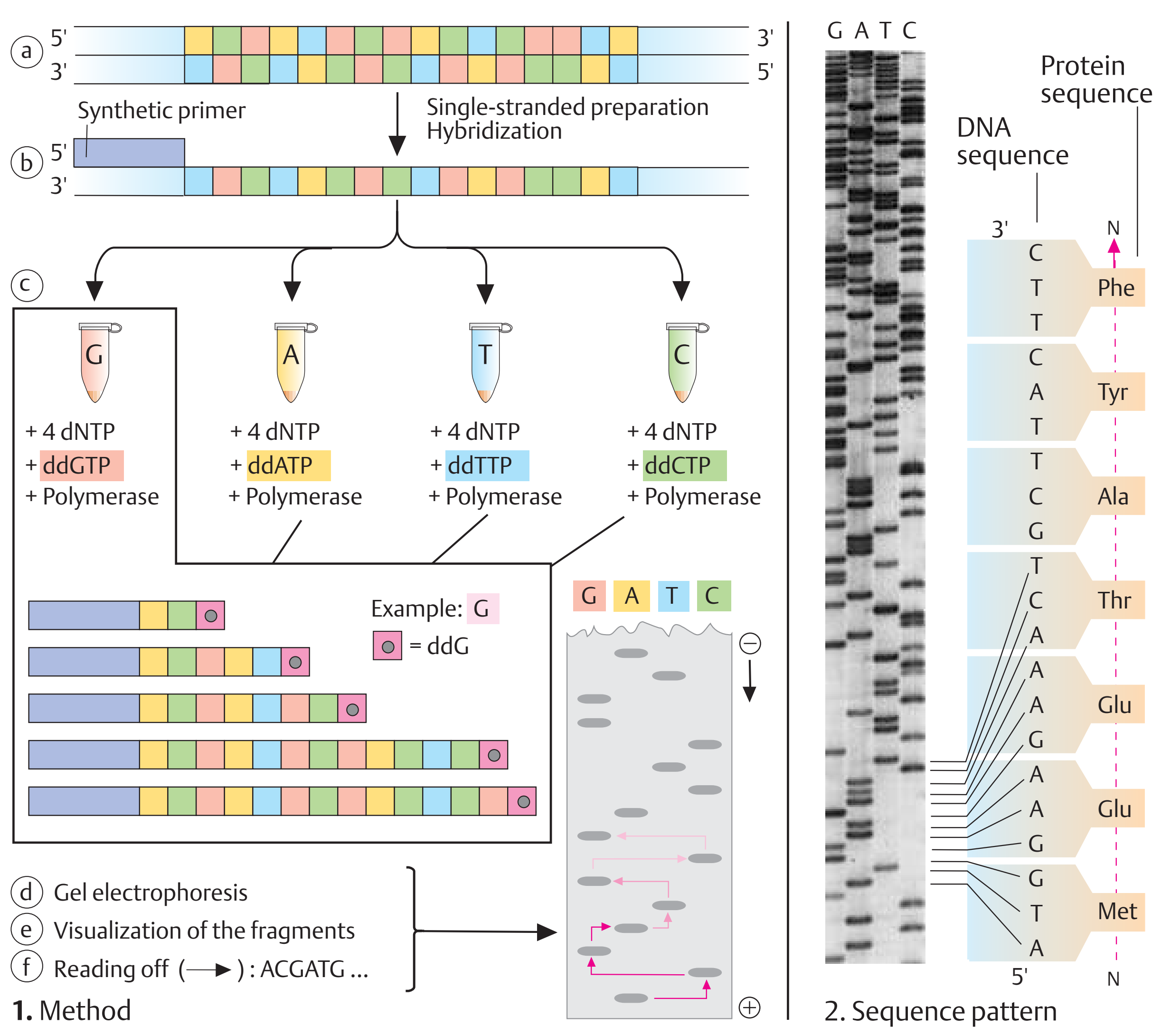
a
5'
3'
3'
5'
Synthetic primer
Single-stranded preparation
Hybridization
b
5'
3'
c
G
A
T
C
+ 4 dNTP
+ ddGTP
+ Polymerase
+ 4 dNTP
+ ddATP
+ Polymerase
+ 4 dNTP
+ ddTTP
+ Polymerase
+ 4 dNTP
+ ddCTP
+ Polymerase
Example: G
= ddG
G A T C
⊖
⊕
d Gel electrophoresis
e Visualization of the fragments
f Reading off (⟶) : ACGATG ...
1. Method
G A T C
Protein
sequence
DNA
sequence
3'
N
C
T
T
Phe
C
A
T
Tyr
T
C
G
Ala
T
C
A
Thr
A
A
G
Glu
A
A
G
Glu
G
T
A
Met
5'
N
2. Sequence pattern

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

PCR and protein expression

A. Polymerase chain reaction (PCR) ◑

The polymerase chain reaction (PCR) is an important procedure in genetic engineering that allows any DNA segment to be replicated (**amplified**) without the need for restriction enzymes, vectors, or host cells (see p. 258). However, the nucleotide sequence of the segment has to be known. Two oligonucleotides (**primers**) are needed, which each hybridize with one of the strands at each end of the DNA segment to be amplified; also needed are sufficient quantities of the four **deoxyribonucleoside triphosphates** and a special heat-tolerant DNA polymerase. The primers are produced by chemical synthesis, and the polymerase is obtained from thermostable bacteria.

First, the starter is heated to around 90 °C to separate the DNA double helix into single strands (**a**; cf. p. 84). The mixture is then cooled to allow hybridization of the primers (**b**). Starting from the primers, complementary DNA strands are now synthesized in both directions by the polymerase (**c**). This *cycle* (cycle 1) is *repeated 20–30 times* with the same reaction mixture (cycle 2 and subsequent cycles). The cyclic heating and cooling are carried out by *computer-controlled thermostats.*

After only the third cycle, double strands start to form with a length equal to the distance between the two primers. The proportion of these approximately doubles during each cycle, until almost all of the newly synthesized segments have the correct length.

B. DNA electrophoresis ○

The separation of DNA fragments by electrophoresis is technically simpler than protein electrophoresis (see p. 78). The mobility of molecules in an electrical field of a given strength depends on the size and shape of the molecules, as well as their charge. In contrast to proteins, in which all three factors vary, the ratio of mass to charge in nucleic acids is constant, as all of the nucleotide components have similar masses and carry one negative charge. When electrophoresis is carried out in a wide-meshed support material that does not separate according to size and shape, the mobility of the molecules depends on their mass alone. The supporting material generally used in genetic engineering is a gel of the polysaccharide **agarose** (see p. 40). Agarose gels are not very stable and are therefore poured horizontally into a plastic chamber in which they are used for separation (top).

To make the separated fragments visible, after running the procedure the gels are placed in solutions of **ethidium bromide**. This is an intercalator (see p. 254) that shows strong fluorescence in UV light after binding to DNA, although it barely fluoresces in an aqueous solution. The result of separating two PCR amplificates (lanes 1 and 2) is shown in the lower part of the illustration. Comparing their distances with those of polynucleotides of known lengths (lane 3; bp = base pairs) yields lengths of approximately 800 bp for fragment 1 and 1800 bp for fragment 2. After staining, the bands can be cut out of the gel and the DNA can be extracted from them and used for further experiments.

C. Overexpression of proteins ◑

To treat some diseases, proteins are needed that occur in such small quantities in the organism that isolating them on a large scale would not be economically feasible. Proteins of this type can be obtained by *overexpression* in bacteria or eukaryotic cells. To do this, the corresponding *gene* is isolated from human DNA and cloned into an **expression plasmid** as described on p. 258. In addition to the gene itself, the plasmid also has to contain DNA segments that allow replication by the host cell and transcription of the gene. After *transformation* and *replication* of suitable host cells, **induction** is used in a targeted fashion to trigger efficient *transcription* of the gene. *Translation* of the mRNA formed in the host cell then gives rise to large amounts of the desired protein. Human insulin (see p. 76), plasminogen activators for dissolving blood clots (see p. 292), and the growth hormone somatotropin are among the proteins produced in this way.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Polymerase chain reaction (PCR)

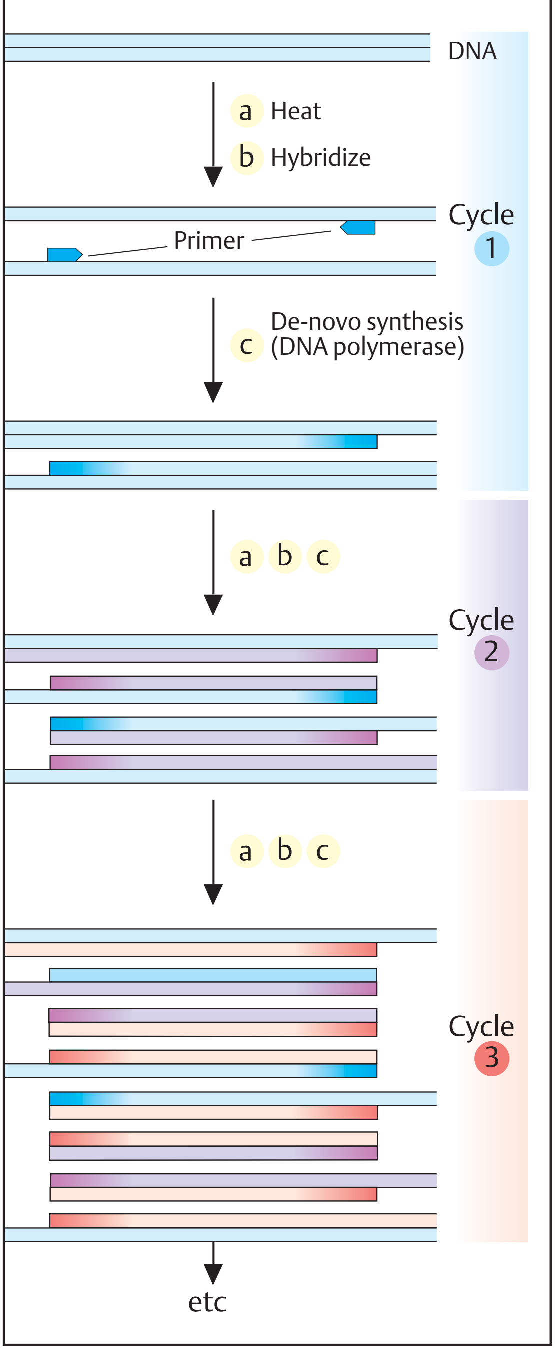
DNA
a Heat
b Hybridize
Cycle 1
Primer
c De-novo synthesis (DNA polymerase)
a b c
Cycle 2
a b c
Cycle 3
etc

B. DNA electrophoresis

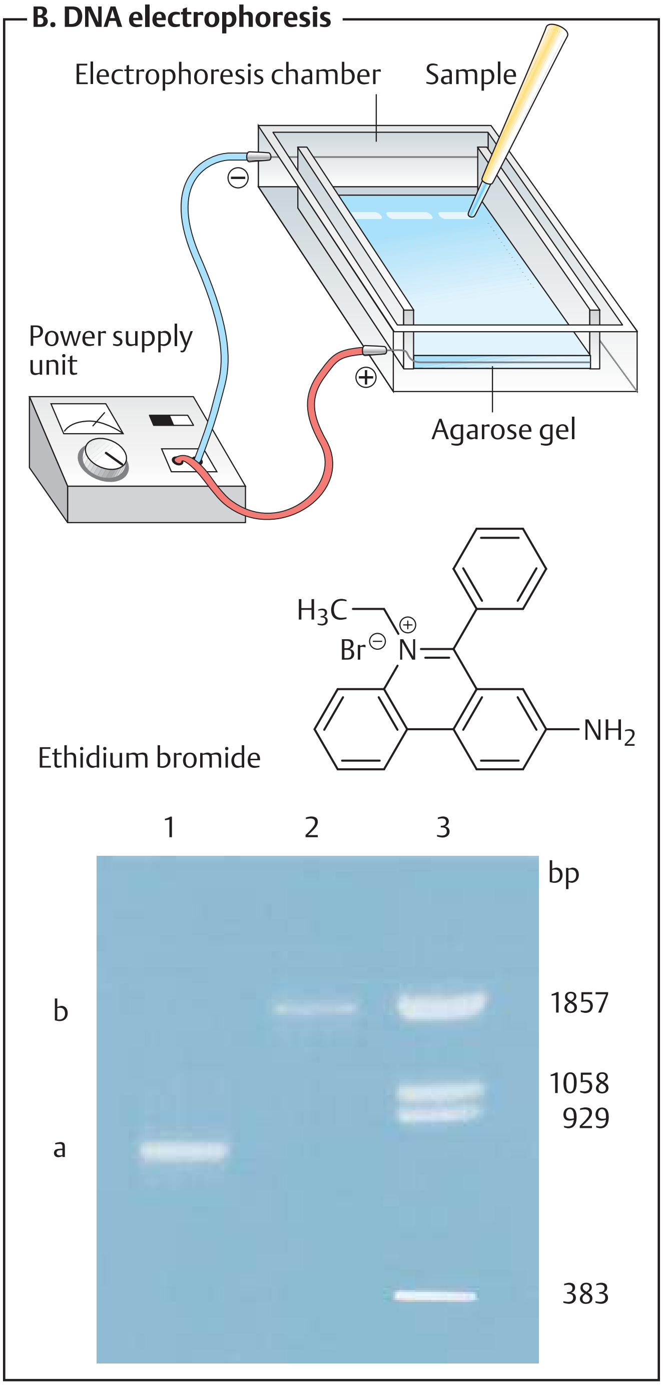
Electrophoresis chamber
Sample
⊖
Power supply unit
⊕
Agarose gel
H_3C
$Br^{\ominus}$
$N^{\oplus}$
NH_2
Ethidium bromide
1
2
3
bp
b
a
1857
1058
929
383

C. Overexpression of proteins

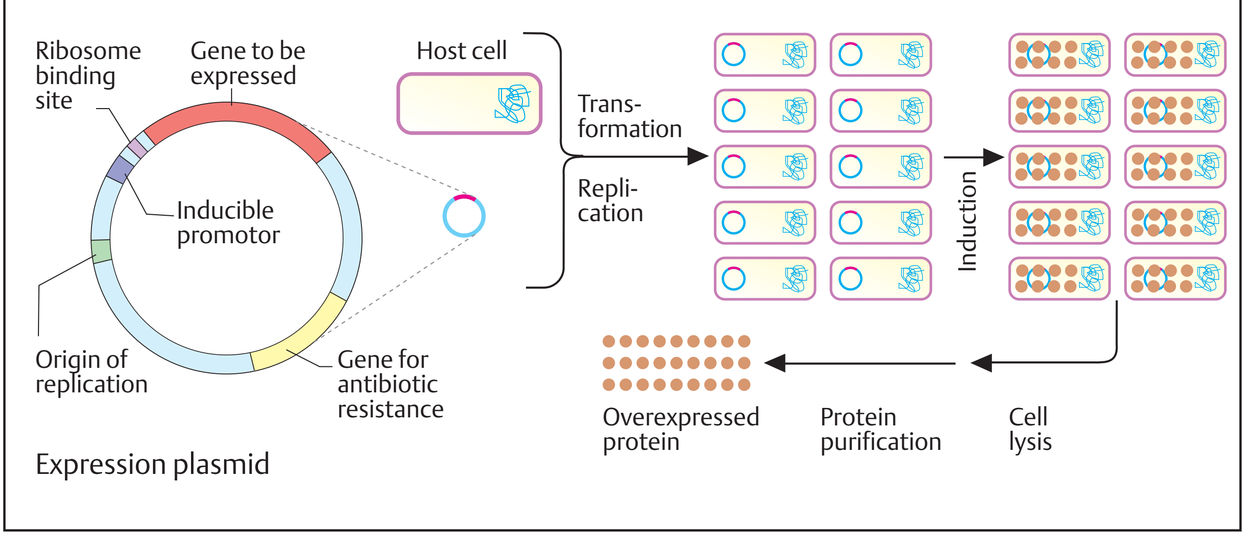
Ribosome binding site
Gene to be expressed
Host cell
Inducible promotor
Origin of replication
Gene for antibiotic resistance
Expression plasmid
Trans-formation
Repli-cation
Induction
Overexpressed protein
Protein purification
Cell lysis

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Genetic engineering in medicine

Genetic engineering procedures are becoming more and more important in medicine for diagnostic purposes (**A**–**C**). New genetic approaches to the treatment of severe diseases are still in the developmental stage (“gene therapy,” **D**).

A. DNA fingerprinting ○

DNA fingerprinting is used to link small amounts of biological material—e. g., traces from the site of a crime—to a specific person. The procedure now used is based on the fact that the human genome contains non-coding repetitive DNA sequences, the length of which varies from individual to individual. **Short tandem repeats** (**STRs**) thus exist in which dinucleotides (e. g., -T-X-) are frequently repeated. Each STR can occur in five to 15 different lengths (alleles), of which one individual possesses only one or two. When the various allele combinations for several STRs are determined after PCR amplification of the DNA being investigated, a “genetic fingerprint” of the individual from whom the DNA originates is obtained. Using comparative material—e. g., saliva samples—definite identification is then possible.

B. Diagnosis of sickle-cell anemia using RFLP ○

This example illustrates a procedure for diagnosing a point mutation in the β-globin gene that leads to sickle-cell anemia (see p. 248). The mutation in the first exon of the gene destroys a cleavage site for the restriction endonuclease *Mst*II (see p. 258). When the DNA of healthy and diseased individuals is cleaved with *Mst*II, different fragments are produced in the region of the β-globin gene, which can be separated by electrophoresis and then demonstrated using specific probes (see p. 260). In addition, heterozygotic carriers of the sickle-cell gene can be distinguished from homozygotic ones.

C. Identification of viral DNA using RT-PCR ○

In viral infections, it is often difficult to determine the species of the pathogen precisely. **RT-PCR** can be used to identify RNA viruses. In this procedure, reverse transcriptase (see p. 404) is used to transcribe the viral RNA into dsDNA, and then PCR is employed to amplify a segment of this DNA with virus-specific primers. In this way, an amplificate with a characteristic length can be obtained for each pathogen and identified using gel electrophoresis as described above.

D. Gene therapy ◑

Many diseases, such as hereditary metabolic defects and tumors, can still not be adequately treated. About 10 years ago, projects were therefore initiated that aimed to treat diseases of this type by transferring genes into the affected cells (gene therapy). The illustration combines conceivable and already implemented approaches to gene therapy for metabolic defects (left) and tumors (right). None of these procedures has yet become established in clinical practice.

If a mutation leads to failure of an enzyme E1 (left), its substrate B will no longer be converted into C and will accumulate. This can lead to cell damage by B itself or by a toxic product formed from it. Treatment with intact E1 is not possible, as the proteins are not capable of passing through the cell membrane. By contrast, it is in principle possible to introduce **foreign genes** into the cell using viruses as vectors (adenoviruses or retroviruses are mainly used). Their gene products could replace the defective E1 or convert B into a harmless product. Another approach uses the so-called **antisense DNA** (bottom right). This consists of polynucleotides that hybridize with the mRNA for specific cellular proteins and thereby prevent their translation. In the case shown, the synthesis of E2 could be blocked, for example.

The main problem in chemotherapy for **tumors** is the lack of tumor-specificity in the highly toxic cytostatic agents used (see p. 402). Attempts are therefore being made to introduce into tumor cells genes with products that are only released from a precursor to form active cytostatics once they have reached their target (left). Other gene products are meant to force the cells into apoptosis (see p. 396) or make them more susceptible to attack by the immune system. To steer the viral vectors to the tumor (targeting), attempts are being made to express proteins on the virus surface that are bound by tumor-specific receptors. Fusion with a tumor-specific promoter could also help limit the effect of the foreign gene to the tumor cells.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. DNA fingerprinting

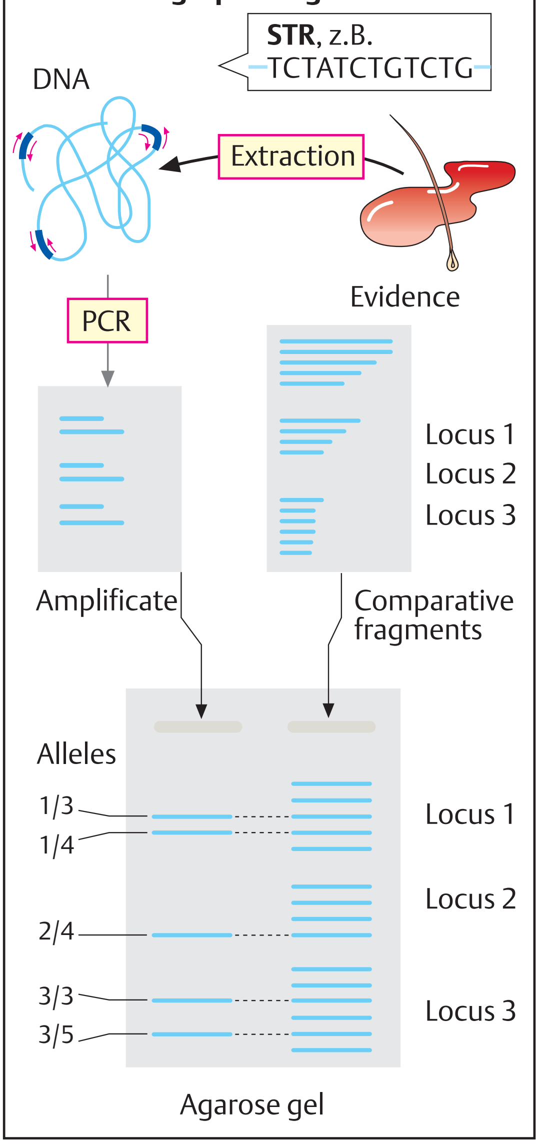
STR, z.B.
TCTATCTGTCTG
DNA
Extraction
Evidence
PCR
Amplificate
Locus 1
Locus 2
Locus 3
Comparative fragments
Alleles
1/3
1/4
2/4
3/3
3/5
Locus 1
Locus 2
Locus 3
Agarose gel

B. Diagnosis of sickle-cell anemia using RFLP

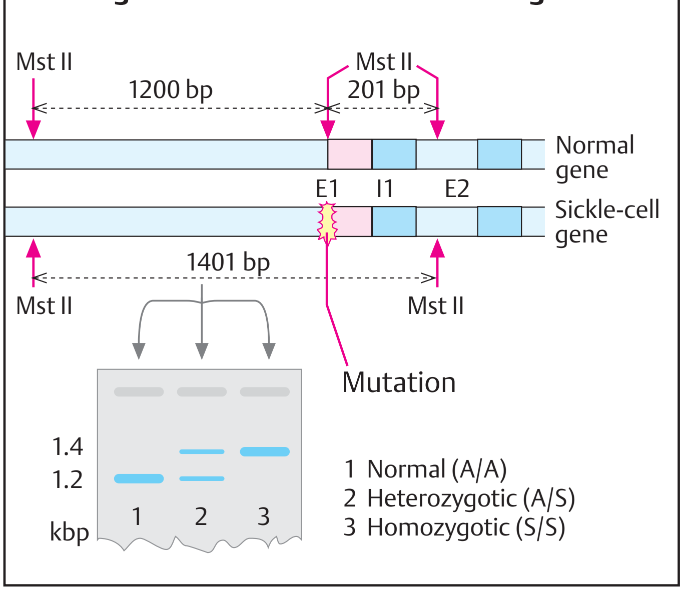
Mst II
1200 bp
Mst II
201 bp
Normal gene
E1
I1
E2
Sickle-cell gene
1401 bp
Mst II
Mst II
Mutation
1.4
1.2
kbp
1
2
3
1 Normal (A/A)
2 Heterozygotic (A/S)
3 Homozygotic (S/S)

C. Evidence of viral DNA using RT-PCR

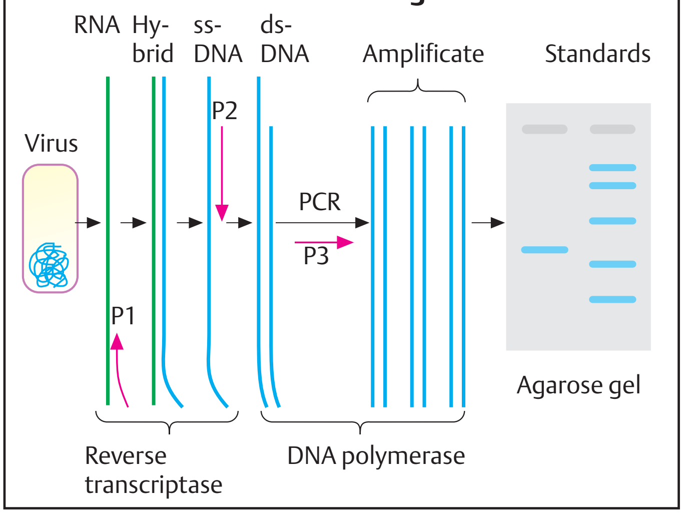
RNA
Hy-brid
ss-DNA
ds-DNA
Amplificate
Standards
Virus
P2
PCR
P3
P1
Agarose gel
Reverse transcriptase
DNA polymerase

D. Gene therapy

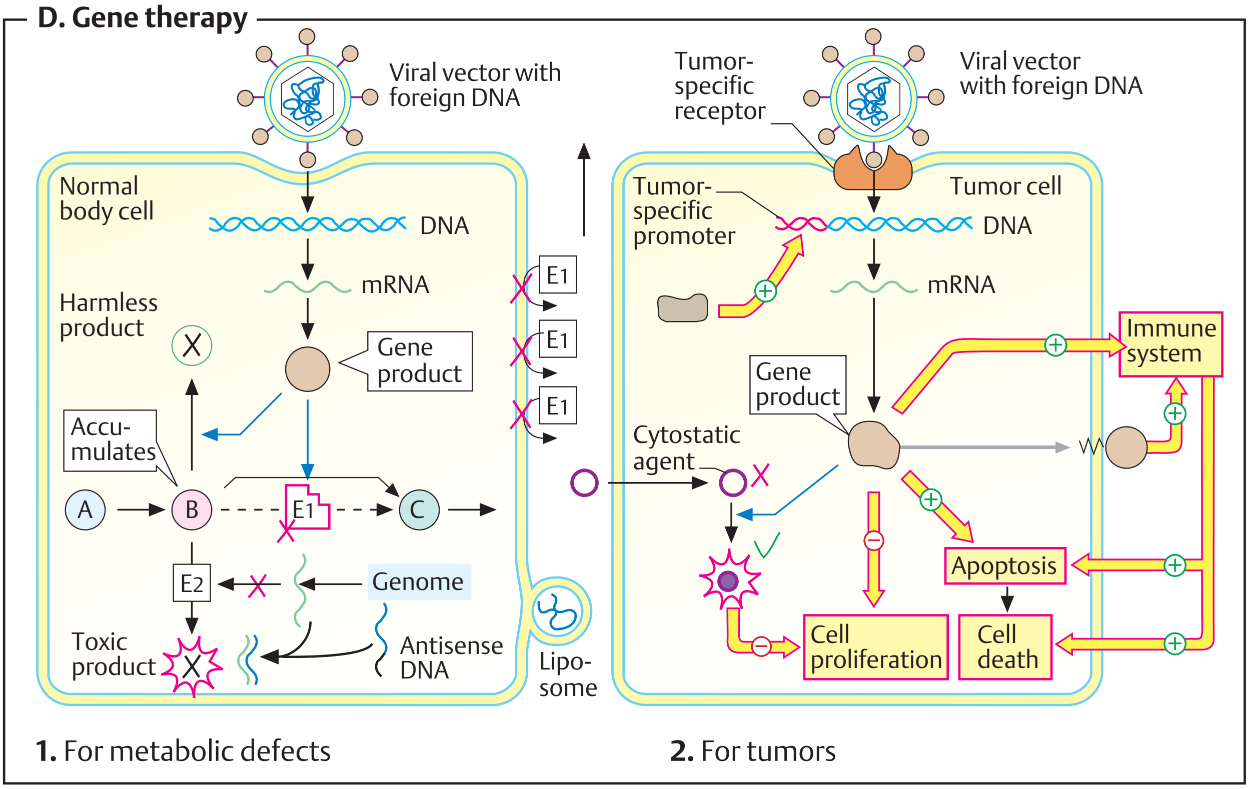
Viral vector with foreign DNA
Normal body cell
DNA
mRNA
Harmless product
Gene product
Accu-mulates
A
B
E1
C
E2
Genome
Toxic product
Antisense DNA
E1
E1
E1
Lipo-some
Tumor-specific receptor
Viral vector with foreign DNA
Tumor-specific promoter
Tumor cell
DNA
mRNA
Gene product
Immune system
Cytostatic agent
Apoptosis
Cell proliferation
Cell death
1. For metabolic defects
2. For tumors

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Digestion: overview

Most components of food (see p. 360) cannot be resorbed directly by the organism. It is only after they have been broken down into smaller molecules that the organism can take up the essential nutrients. *Digestion* refers to the mechanical and enzymatic breakdown of food and the resorption of the resulting products.

A. Hydrolysis and resorption of food components ●

Following mechanical fragmentation of food during chewing in the mouth, the process of enzymatic degradation starts in the stomach. For this purpose, the chyme is mixed with *digestive enzymes* that occur in the various digestive secretions or in membrane-bound form on the surface of the intestinal epithelium (see p. 268). Almost all digestive enzymes are *hydrolases* (class 3 enzymes; see p. 88); they catalyze the cleavage of composite bonds with the uptake of water.

Proteins are first denatured by the stomach's *hydrochloric acid* (see p. 270), making them more susceptible to attack by the *endopeptidases* (proteinases) present in gastric and pancreatic juice. The peptides released by endopeptidases are further degraded into amino acids by *exopeptidases*. Finally, the amino acids are resorbed by the intestinal mucosa in cotransport with Na^+ ions (see p. 220). There are separate transport systems for each of the various groups of amino acids.

Carbohydrates mainly occur in food in the form of polymers (starches and glycogen). They are cleaved by *pancreatic amylase* into oligosaccharides and are then hydrolyzed by *glycosidases,* which are located on the surface of the intestinal epithelium, to yield monosaccharides. Glucose and galactose are taken up into the enterocytes by secondary active cotransport with Na^+ ions (see p. 220). In addition, monosaccharides also have passive transport systems in the intestine.

Nucleic acids are broken down into their components by *nucleases* from the pancreas and small intestine (ribonucleases and deoxyribonucleases). Further breakdown yields the nucleobases (purine and pyrimidine derivatives), pentoses (ribose and deoxyribose), phosphate, and nucleosides (nucleobase pentose). These cleavage products are resorbed by the intestinal wall in the region of the jejunum.

Lipids are a special problem for digestion, as they are not soluble in water. Before enzymatic breakdown, they have to be emulsified by *bile salts* and *phospholipids* in the bile (see p. 314). At the water–lipid interface, *pancreatic lipase* then attacks triacylglycerols with the help of *colipase* (see p. 270). The cleavage products include *fatty acids, 2–monoacylglycerols, glycerol,* and *phosphate* from phospholipid breakdown. After resorption into the epithelial cells, fats are resynthesized from fatty acids, glycerol and 2–monoacylglycerols and passed into the lymphatic system (see p. 272). The lipids in milk are more easily digested, as they are already present in emulsion; on cleavage, they mostly provide short-chain fatty acids.

Inorganic components such as water, electrolytes, and *vitamins* are directly absorbed by the intestine.

High-molecular-weight indigestible components, such as the fibrous components of plant cell walls, which mainly consist of cellulose and lignin, pass through the bowel unchanged and form the main component of feces, in addition to cells shed from the intestinal mucosa. Dietary fiber makes a positive contribution to digestion as a *ballast material* by binding water and promoting intestinal peristalsis.

The food components resorbed by the epithelial cells of the intestinal wall in the region of the jejunum and ileum are transported directly to the liver via the *portal vein.* Fats, cholesterol, and lipid–soluble vitamins are exceptions. These are first released by the enterocytes in the form of *chylomicrons* (see p. 278) into the *lymph system,* and only reach the blood via the thoracic duct.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Hydrolysis and resorption of dietary constituents

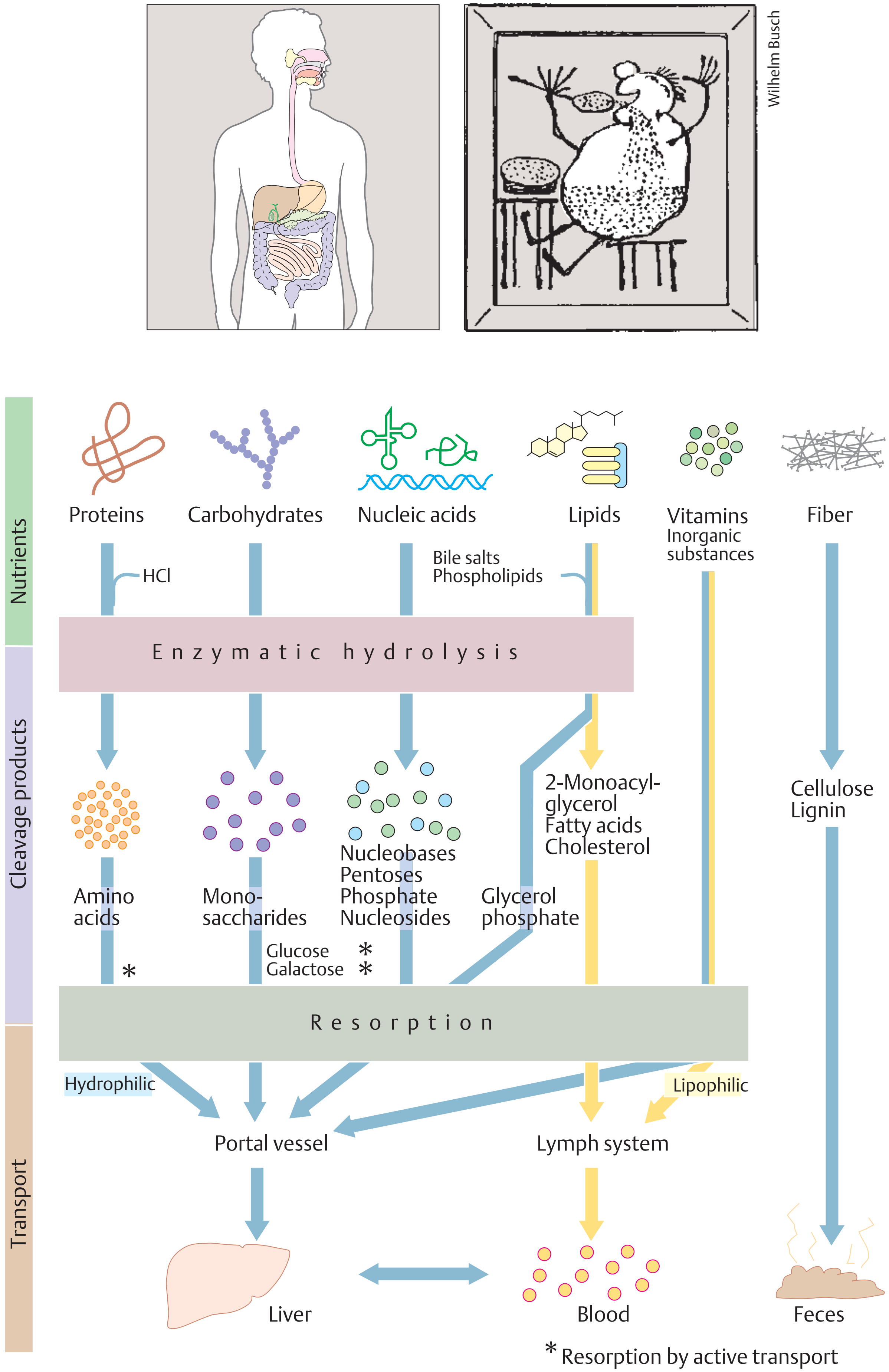
Wilhelm Busch
Nutrients
Cleavage products
Transport
Proteins
Carbohydrates
Nucleic acids
Lipids
Vitamins
Inorganic substances
Fiber
HCl
Bile salts
Phospholipids
E n z y m a t i c h y d r o l y s i s
Amino acids
Mono-saccharides
Nucleobases
Pentoses
Phosphate
Nucleosides
Glycerol phosphate
2-Monoacyl-glycerol
Fatty acids
Cholesterol
Cellulose
Lignin
*
Glucose *
Galactose *
R e s o r p t i o n
Hydrophilic
Lipophilic
Portal vessel
Lymph system
Liver
Blood
Feces
* Resorption by active transport

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Digestive secretions

A. Digestive juices ◑

Saliva. The salivary glands produce a slightly alkaline secretion which—in addition to water and salts—contains *glycoproteins* (mucins) as lubricants, *antibodies,* and *enzymes. α–Amylase* attacks polysaccharides, and a lipase hydrolyzes a small proportion of the neutral fats. α–Amylase and *lysozyme,* a murein-cleaving enzyme (see p. 40), probably serve to regulate the oral bacterial flora rather than for digestion (see p. 340).

Gastric juice. In the stomach, the chyme is mixed with gastric juice. Due to its hydrochloric acid content, this secretion of the gastric mucosa is strongly acidic (pH 1–3; see p. 270). It also contains *mucus* (mainly glycoproteins known as mucins), which protects the mucosa from the hydrochloric acid, *salts,* and *pepsinogen*—the proenzyme ("zymogen") of the aspartate proteinase *pepsin* (see pp. 176, 270). In addition, the gastric mucosa secretes what is known as *"intrinsic factor"*—a glycoprotein needed for resorption of vitamin B_{12} ("extrinsic factor") in the bowel.

In the stomach, pepsin and related enzymes initiate the enzymatic digestion of proteins, which takes 1–3 hours. The acidic gastric contents are then released into the duodenum in batches, where they are neutralized by alkaline pancreatic secretions and mixed with cystic bile.

Pancreatic secretions. In the acinar cells, the pancreas forms a secretion that is alkaline due to its HCO_3^- content, the buffer capacity of which is sufficient to neutralize the stomach's hydrochloric acid. The pancreatic secretion also contains many *enzymes* that catalyze the hydrolysis of high–molecular-weight food components. All of these enzymes are hydrolases with pH optimums in the neutral or weakly alkaline range. Many of them are formed and secreted as proenzymes and are only activated in the bowel lumen (see p. 270).

Trypsin, chymotrypsin, and *elastase* are endopeptidases that belong to the group of serine proteinases (see p. 176). Trypsin hydrolyzes specific peptide bonds on the C side of the basic amino acids Arg and Lys, while chymotrypsin prefers peptide bonds of the apolar amino acids Tyr, Trp, Phe, and Leu (see p. 94). Elastase mainly cleaves on the C side of the aliphatic amino acids Gly, Ala, Val, and Ile. Smaller peptides are attacked by *carboxypeptidases,* which as exopeptidases cleave individual amino acids from the C–terminal end of the peptides (see p. 176).

∼–Amylase, the most important endoglycosidase in the pancreas, catalyzes the hydrolysis of α1→4 bonds in the polymeric carbohydrates starch and glycogen. This releases maltose, maltotriose, and a mixture of other oligosaccharides.

Various pancreatic enzymes hydrolyze lipids, including *lipase* with its auxiliary protein *colipase* (see p. 270), *phospholipase* A_2, and *sterol esterase.* Bile salts activate the lipid-cleaving enzymes through micelle formation (see below).

Several hydrolases—particularly *ribonuclease* (RNAse) and *deoxyribonuclease* (DNAse)—break down the nucleic acids contained in food.

Bile. The liver forms a thin secretion (bile) that is stored in the gallbladder after water and salts have been extracted from it. From the gallbladder, it is released into the duodenum. The most important constituents of bile are *water* and inorganic *salts, bile acids* and *bile salts* (see p. 314), *phospholipids, bile pigments,* and *cholesterol.* Bile salts, together with phospholipids, emulsify insoluble food lipids and activate the lipases. Without bile, fats would be inadequately cleaved, if at all, resulting in "fatty stool" (steatorrhea). Resorption of fat-soluble vitamins would also be affected.

Small-intestinal secretions. The glands of the small intestine (the Lieberkühn and Brunner glands) secrete additional digestive enzymes into the bowel. Together with enzymes on the microvilli of the intestinal epithelium (peptidases, glycosidases, etc.), these enzymes ensure almost complete hydrolysis of the food components previously broken down by the endoenzymes.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Digestive juices

↓ Endoenzyme
← Exoenzyme

Compo-nents | Function or substrate

Saliva

Daily secretion 1.0–1.5 l
pH 7

Water	Moistens food	
Salts		
Mucus	Lubricant	
Antibodies	Bind to bacteria	
α-Amylases (3.2.1.1)	Cleave starch	↓
Lysozyme (3.2.1.17)	Attacks bacterial cell walls	↓

Gastric juice

Daily secretion 2–3 l
pH 1

Water		
Salts		
HCl	Denatures proteins, kills bacteria	
Mucus	Protects stomach lining	
Pepsins (3.4.23.1-3)	Cleave proteins	↓
Chymosin (3.4.23.4)	Precipitates casein	↓
Triacylglycerol lipase (3.1.1.3)	Cleaves fats	
Intrinsic factor	Protects vitamin B_{12}	

Bile

Daily secretion 0.6 l
pH 6.9–7.7

Water	
$HCO_3^{\ominus}$	Neutralizes gastric juice
Bile salts	Facilitate lipid digestion
Phospholipids	Facilitate lipid digestion
Bile pigments	Waste products
Cholesterol	Waste product

Pancreatic secretions

Daily secretion 0.7–2.5 l
pH 7.7 (7.5–8.8)

Water		
$HCO_3^{\ominus}$	Neutralizes gastric juice	
Trypsin (3.4.21.4)	Proteins	↓
Chymotrypsin (3.4.21.1)	Proteins	↓
Elastase (3.4.21.36)	Proteins	↓
Carboxypeptidases (3.4.n.n)	Peptides	←
α-Amylase (3.2.1.1)	Starch and glycogen	↓
Triacylglycerol lipase (3.1.1.3)	Fats	
Co-Lipase	Cofactor for lipase	
Phospholipase A_2 (3.1.1.4)	Phospholipids	
Sterol esterase (3.1.1.13)	Cholesterol esters	
Ribonuclease (3.1.27.5)	RNA	
Deoxyribonuclease I (3.1.21.1)	DNA	

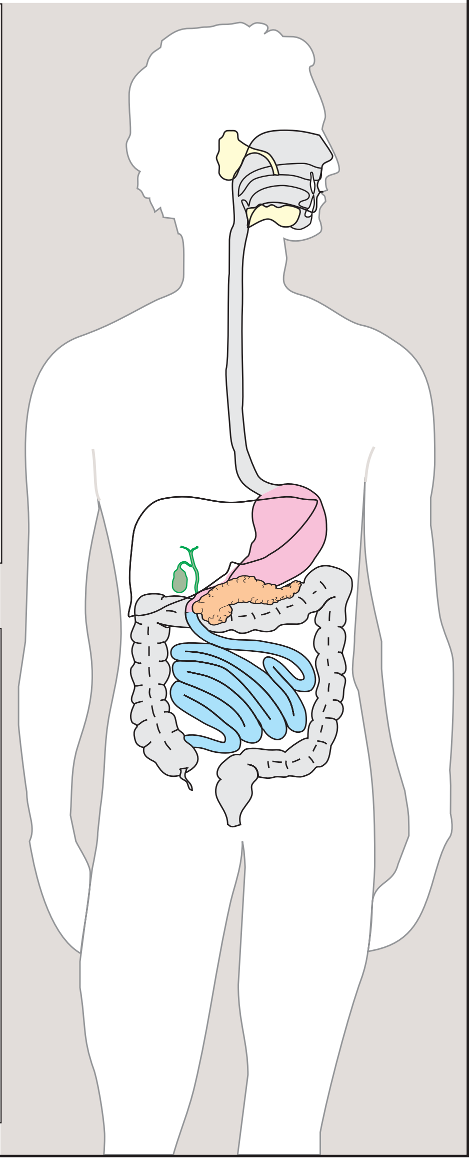

Secretions of the small intestine

Daily secretion unknown
pH 6.5–7.8

Aminopeptidases (3.4.11.n)	Peptides	←
Dipeptidases (3.4.13.n)	Dipeptides	
α-Glucosidase (3.2.1.20)	Oligosaccharides	↓
Oligo-1,6-glucosidase (3.2.1.10)	Oligosaccharides	↓
β-Galactosidase (3.2.1.23)	Lactose	
Sucrose α-glucosidase (3.2.1.48)	Sucrose	
α, α-Trehalase (3.2.1.28)	Trehalose	
Alkaline phosphatase (3.1.3.1)	Phosphoric acid esters	
Polynucleotidases (3.1.3.n)	Nucleic acids, nucleotides	↓
Nucleosidases (3.2.2.n)	Nucleosides	←
Phospholipases (3.1.n.n)	Phospholipids	

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Digestive processes

Gastric juice is the product of several cell types. The *parietal cells* produce hydrochloric acid, *chief cells* release pepsinogen, and *accessory cells* form a mucin-containing mucus.

A. Formation of hydrochloric acid ◑

The secretion of **hydrochloric acid** (H^+ and Cl^-) by the parietal cells is an active process that uses up ATP and takes place against a concentration gradient (in the gastric lumen, with a pH of 1, the H^+ concentration is some 10^6 times higher than in the parietal cells, which have a pH of 7).

The precursors of the exported H^+ ions are carbon dioxide (CO_2) and water (H_2O). CO_2 diffuses from the blood into the parietal cells, and in a reaction catalyzed by *carbonate dehydratase* (carbonic anhydrase [2]), it reacts with H_2O to form H^+ and hydrogen carbonate (HCO_3^-). The H^+ ions are transported into the gastric lumen in exchange for K^+ by a membrane-bound *H^+/K^+-exchanging ATPase* [1] (a transport ATPase of the P type; see p. 220). The remaining hydrogen carbonate is released into the interstitium in electroneutral antiport in exchange for chloride ions (Cl^-), and from there into the blood. The Cl^- ions follow the secreted protons through a channel into the gastric lumen.

The hydrochloric acid in gastric juice is important for digestion. It activates pepsinogen to form pepsin (see below) and creates an optimal pH level for it to take effect. It also denatures food proteins so that they are more easily attacked by proteinases, and it kills micro-organisms.

Regulation. HCl secretion is stimulated by the peptide hormone *gastrin,* the mediator *histamine* (see p. 380), and—via the neurotransmitter *acetylcholine*—by the autonomous nervous system. The peptide *somatostatin* and certain *prostaglandins* (see p. 390) have inhibitory effects. Together with cholecystokinin, secretin, and other peptides, gastrin belongs to the group of **gastrointestinal hormones** (see p. 370). All of these are formed in the gastrointestinal tract and mainly act in the vicinity of the site where they are formed—i. e., they are paracrine hormones (see p. 372). While gastrin primarily enhances HCl secretion, *cholecystokinin* and *secretin* mainly stimulate pancreatic secretion and bile release.

B. Zymogen activation ◑

To prevent self–digestion, the pancreas releases most proteolytic enzymes into the duodenum in an inactive form as *proenzymes* (zymogens). Additional protection from the effects of premature activation of pancreatic proteinases is provided by *proteinase inhibitors* in the pancreatic tissue, which inactivate active enzymes by complex formation (right).

Trypsinogen plays a key role among the proenzymes released by the pancreas. In the bowel, it is proteolytically converted into active trypsin (see p. 176) by **enteropeptidase**, a membrane enzyme on the surface of the enterocytes. Trypsin then autocatalytically activates additional trypsinogen molecules and the other proenzymes (left).

C. Fat digestion ◑

Due to the "hydrophobic effect" (see p. 28), water-insoluble neutral fats in the aqueous environment of the bowel lumen would aggregate into drops of fat in which most of the molecules would not be accessible to pancreatic lipase. The amphipathic substances in bile (bile acids, bile salts, phospholipids) create an **emulsion** in which they occupy the surface of the droplets and thereby prevent them from coalescing into large drops. In addition, the bile salts, together with the auxiliary protein colipase, mediate binding of *triacylglycerol lipase* [1] to the emulsified fat droplets. Activation of the lipase is triggered by a conformation change in the C-terminal domain of the enzyme, which uncovers the active center.

During passage through the intestines, the active lipase breaks down the triacylglycerols in the interior of the droplets into free fatty acids and amphipathic monoacylglycerols. Over time, smaller **micelles** develop (see p. 28), in the envelope of which monoacylglycerols are present in addition to bile salts and phospholipids. Finally, the components of the micelles are resorbed by the enterocytes in ways that have not yet been explained.

Monoacylglycerols and fatty acids are reassembled into fats again (see p. 272), while the bile acids return to the liver (enterohepatic circulation; see p. 314).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Formation of hydrochloric acid

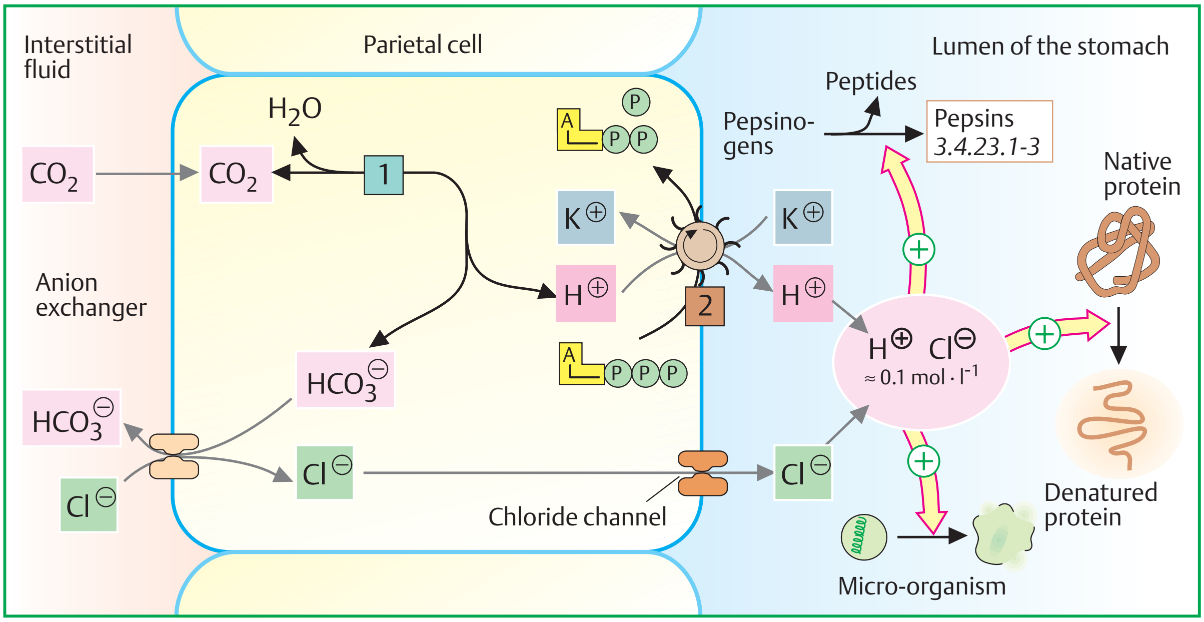
Interstitial fluid
Parietal cell
Lumen of the stomach
CO_2
CO_2
H_2O
1
Anion exchanger
HCO_3^-
HCO_3^-
Cl^-
Cl^-
Cl^-
Chloride channel
A P P
A P P P
P
K^+
K^+
H^+
H^+
2
Pepsinogens
Peptides
Pepsins 3.4.23.1-3
Native protein
H^+ Cl^-
≈ 0.1 mol · l^{-1}
Denatured protein
Micro-organism

1 Carbonate dehydratase *4.2.1.1* [Zn^{2+}]   2 H^+/K^+-exchanging ATPase *3.6.1.36*

B. Zymogen activation

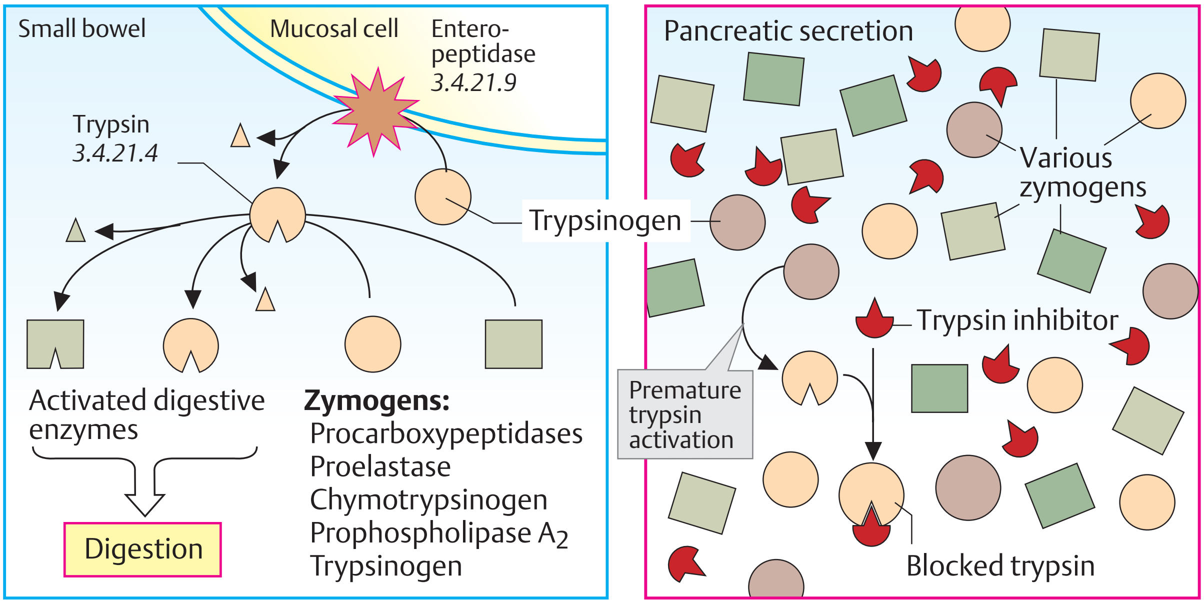
Small bowel
Mucosal cell
Enteropeptidase 3.4.21.9
Trypsin 3.4.21.4
Trypsinogen
Activated digestive enzymes
Digestion
Zymogens:
Procarboxypeptidases
Proelastase
Chymotrypsinogen
Prophospholipase A_2
Trypsinogen
Pancreatic secretion
Various zymogens
Trypsin inhibitor
Premature trypsin activation
Blocked trypsin

C. Fat digestion

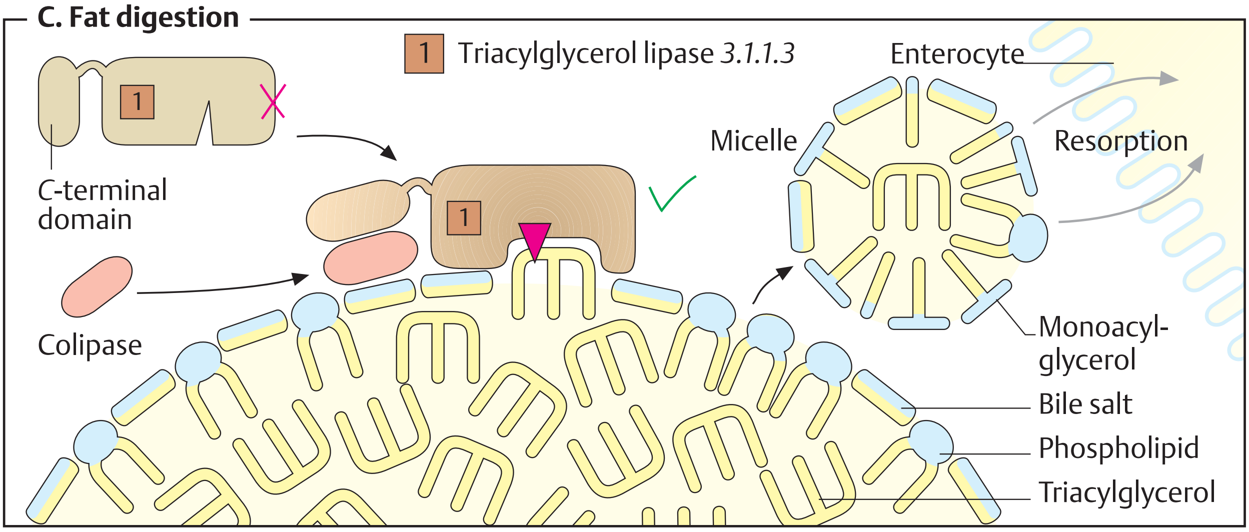
1 Triacylglycerol lipase 3.1.1.3
1
C-terminal domain
Colipase
1
Micelle
Enterocyte
Resorption
Monoacylglycerol
Bile salt
Phospholipid
Triacylglycerol

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Resorption

Enzymatic hydrolysis in the digestive tract breaks down foodstuffs into their resorbable components. *Resorption* of the cleavage products takes place primarily in the small intestine. Only ethanol and short–chain fatty acids are already resorbed to some extent in the stomach.

The resorption process is facilitated by the large inner surface of the intestine, with its brush–border cells. Lipophilic molecules penetrate the plasma membrane of the mucosal cells by simple diffusion, whereas polar molecules require *transporters* (facilitated diffusion; see p. 218). In many cases, carrier-mediated cotransport with Na$^+$ ions can be observed. In this case, the difference in the concentration of the sodium ions (high in the intestinal lumen and low in the mucosal cells) drives the import of nutrients against a concentration gradient (secondary active transport; see p. 220). Failure of carrier systems in the gastrointestinal tract can result in diseases.

A. Monosaccharides ◑

The cleavage of polymeric carbohydrates by *∼–amylase* [1] leads to **oligosaccharides**, which are broken down further by *exoglycosidases* (oligosaccharidases and disaccharidases [2]) on the membrane surface of the brush border. The monosaccharides released in this way then pass with the help of various *sugar–specific transporters* into the cells of intestinal epithelium. *Secondary active transport* serves for the uptake of **glucose** and **galactose**, which are transported against a concentration gradient in cotransport with Na$^+$. The Na$^+$ gradient is maintained on the basal side of the cells by *Na$^+$/K$^+$-ATPase* [3]. Another passive transporter then releases glucose and galactose into the blood. **Fructose** is taken up by a special type of transporter using facilitated diffusion.

Amino acids (not illustrated)

Protein degradation is initiated by *proteinases*—by pepsins in the stomach and by trypsin, chymotrypsin, and elastase in the small intestine. The resulting peptides are then further hydrolyzed by various *peptidases* into amino acids. Individual amino acid groups have *group–specific amino acid transporters*, some of which transport the amino acids into the enterocytes in cotransport with Na$^+$ ions (secondary active transport), while others transport them in an Na$^+$–independent manner through facilitated diffusion. Small peptides can also be taken up.

B. Lipids ◑

Fats and other lipids are poorly soluble in water. The larger the accessible surface is—i. e., the better the fat is emulsified—the easier it is for enzymes to hydrolyze it (see p. 270). Due to the special properties of milk, milk fats already reach the gastrointestinal tract in emulsified form. Digestion of them therefore already starts in the oral cavity and stomach, where lipases in the saliva and gastric juice are available. Lipids that are less accessible—e. g., from roast pork—are emulsified in the small intestine by *bile salts* and *bile phospholipids*. Only then are they capable of being attacked by *pancreatic lipase* [4] (see p. 270).

Fats (triacylglycerols) are mainly attacked by pancreatic lipase at positions 1 and 3 of the glycerol moiety. Cleavage of two fatty acid residues gives rise to **fatty acids** and **2-monoacylglycerols**, which are quantitatively the most important products. However, a certain amount of **glycerol** is also formed by complete hydrolysis. These cleavage products are resorbed by a non-ATP-dependent process that has not yet been explained in detail.

In the mucosal cells, *long-chain fatty acids* are resynthesized by an ATP-dependent ligase [5] to form acyl-CoA and then triacylglycerols (fats; see p. 170). The fats are released into the lymph in the form of **chylomicrons** (see p. 278) and, bypassing the liver, are deposited in the thoracic duct—i. e., the blood system. *Cholesterol* also follows this route.

By contrast, *short-chain fatty acids* (with chain lengths of less than 12 C atoms) pass directly into the blood and reach the liver via the portal vein. Resorbed glycerol can also take this path.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

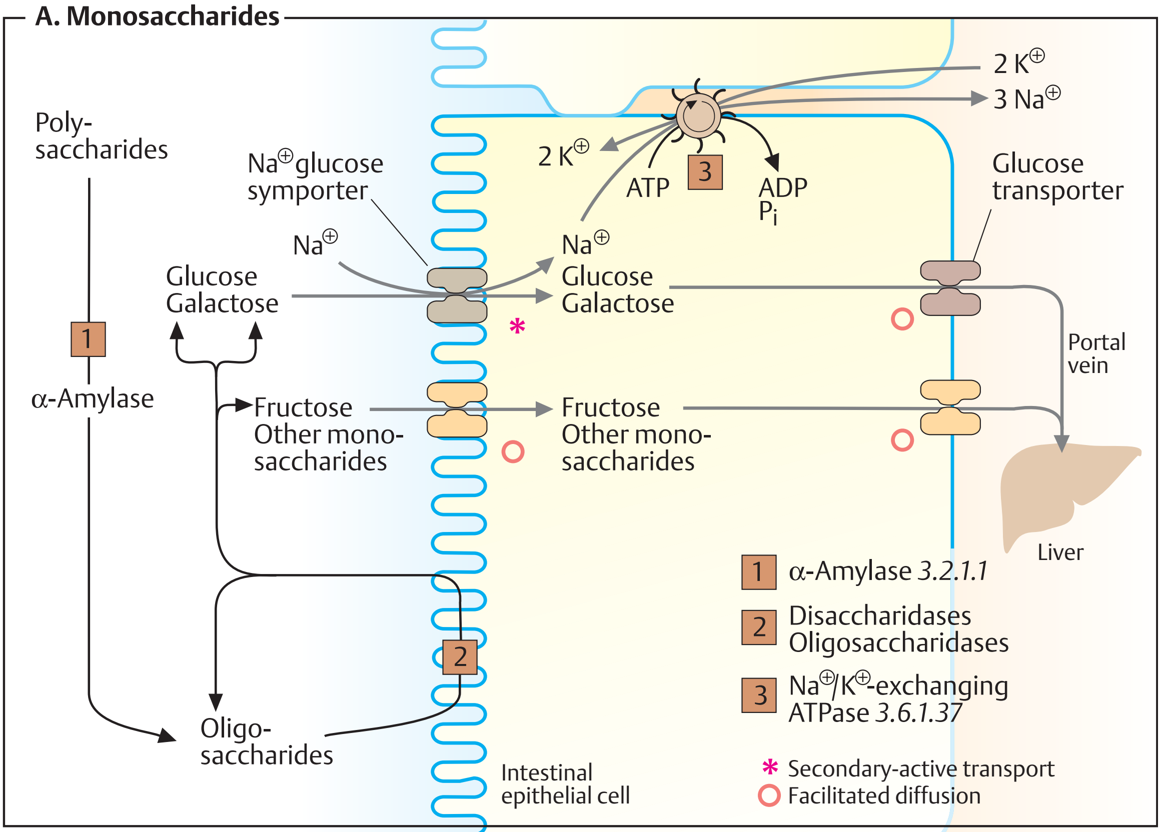
A. Monosaccharides
Poly-saccharides
1
α-Amylase
Oligo-saccharides
Glucose Galactose
Fructose Other mono-saccharides
Na⊕glucose symporter
Na⊕
2
Na⊕ Glucose Galactose
*
Fructose Other mono-saccharides
2 K⊕
ATP
3
ADP Pi
2 K⊕
3 Na⊕
Glucose transporter
Portal vein
Liver
Intestinal epithelial cell
1 α-Amylase 3.2.1.1
2 Disaccharidases Oligosaccharidases
3 Na⊕/K⊕-exchanging ATPase 3.6.1.37
* Secondary-active transport
○ Facilitated diffusion

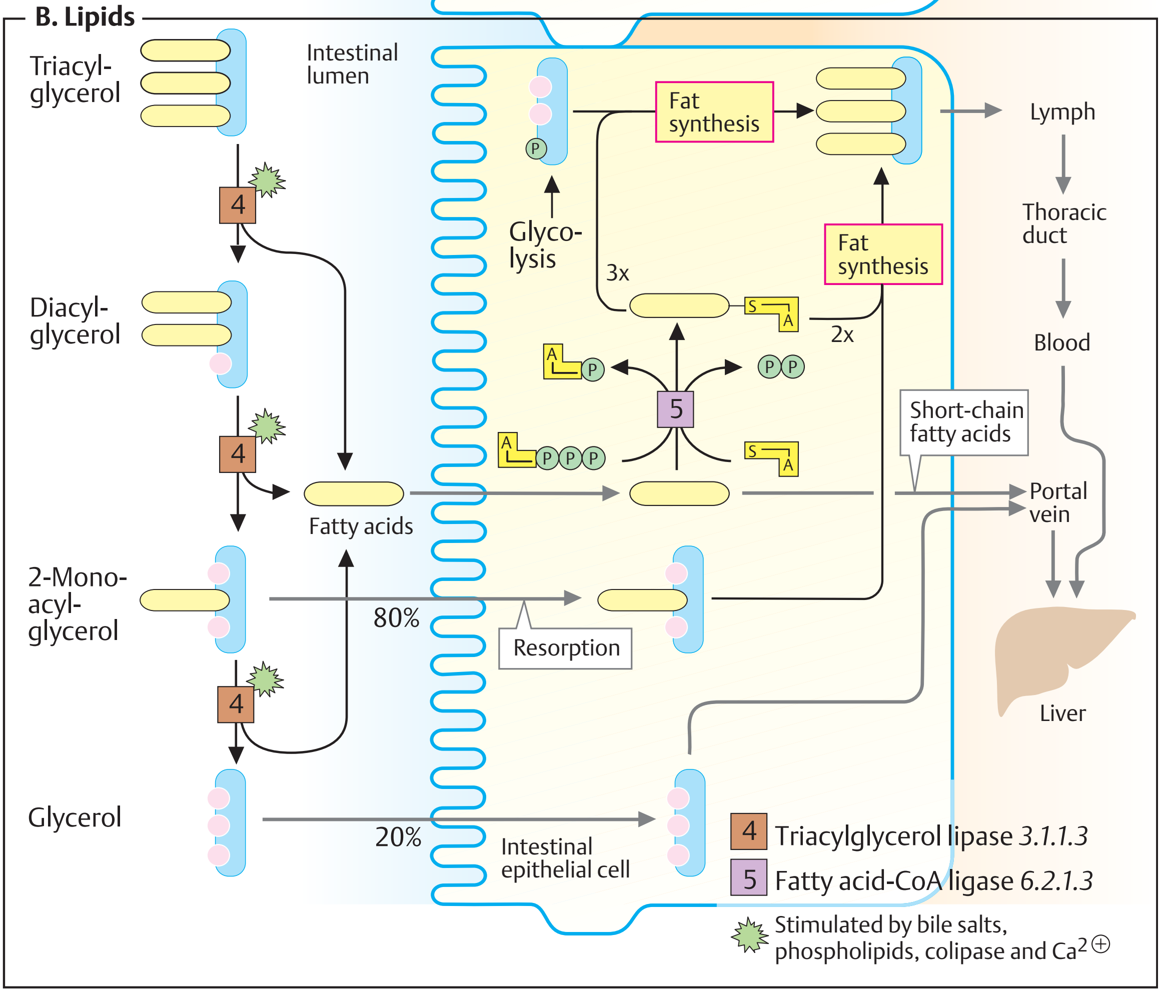
B. Lipids
Triacyl-glycerol
4
Diacyl-glycerol
4
Fatty acids
2-Mono-acyl-glycerol
80%
4
Glycerol
20%
Intestinal lumen
P
Glyco-lysis
Fat synthesis
3x
S A
2x
Fat synthesis
A P
P P
5
A P P P
S A
Resorption
Lymph
Thoracic duct
Blood
Short-chain fatty acids
Portal vein
Liver
Intestinal epithelial cell
4 Triacylglycerol lipase 3.1.1.3
5 Fatty acid-CoA ligase 6.2.1.3
Stimulated by bile salts, phospholipids, colipase and Ca2⊕

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Blood: composition and functions

Human blood constitutes about 8% of the body's weight. It consists of **cells** and cell fragments in an aqueous medium, the **blood plasma.** The proportion of cellular elements, known as *hematocrit,* in the total volume is approximately 45%.

A. Functions of the blood ●

The blood is the most important transport medium in the body. It serves to keep the "internal milieu" constant (homeostasis) and it plays a decisive role in defending the body against pathogens.

Transport. The *gases* oxygen and carbon dioxide are transported in the blood. The blood mediates *the exchange of substances between organs* and takes up *metabolic end products* from tissues in order to transport them to the lungs, liver, and kidney for excretion. The blood also distributes *hormones* throughout the organism (see p. 370).

Homeostasis. The blood ensures that a balanced distribution of water is maintained between the vascular system, the cells (intracellular space), and the extracellular space. The *acid–base balance* is regulated by the blood in combination with the lungs, liver, and kidneys (see p. 288). The regulation of *body temperature* also depends on the controlled transport of heat by the blood.

Defense. The body uses both non-specific and specific mechanisms to defend itself against pathogens. The defense system includes the *cells of the immune system* and certain *plasma proteins* (see p. 294).

Self-protection. To prevent blood loss when a vessel is injured, the blood has systems for stanching blood flow and coagulating the blood (hemostasis; see p. 290). The dissolution of blood clots (fibrinolysis) is also managed by the blood itself (see p. 292).

B. Cellular elements ◑

The solid elements in the blood are the erythrocytes (red blood cells), leukocytes (white blood cells), and thrombocytes (platelets).

The **erythrocytes** provide for gas transport in the blood. They are discussed in greater detail on pp. 280–285.

The **leukocytes** include various types of granulocyte, monocyte, and lymphocyte. All of these have immune defense functions (see p. 294). The *neutrophil granulocytes, monocytes,* and the *macrophages* derived from monocytes are phagocytes. They can ingest and degrade invading pathogens. The *lymphocytes* are divided into two groups, B lymphocytes and T lymphocytes. B lymphocytes produce *antibodies,* while T lymphocytes regulate the immune response and destroy virus-infected cells and tumor cells. *Eosinophilic* and *basophilic granulocytes* have special tasks for defense against animal parasites.

Thrombocytes are cell fragments that arise in the bone marrow from large precursor cells, the megakaryocytes. Their task is to promote hemostasis (see p. 290).

C. Blood plasma: composition ◑

The **blood plasma** is an aqueous solution of electrolytes, nutrients, metabolites, proteins, vitamins, trace elements, and signaling substances. The fluid phase of coagulated blood is known as **blood serum**. It differs from the plasma in that it lacks fibrin and other coagulation proteins (see p. 290).

Laboratory assessment of the composition of the blood plasma is often carried out in clinical chemistry. Among the electrolytes, there is a relatively high concentration of Na^+, Ca^{2+}, and Cl^- ions in the blood in comparison with the cytoplasm. By contrast, the concentrations of K^+, Mg^{2+}, and phosphate ions are higher in the cells. Proteins also have a higher intracellular concentration. The electrolyte composition of blood plasma is similar to that of seawater, due to the evolution of early forms of life in the sea. The solution known as *"physiological saline"* (NaCl at a concentration of 0.15 mol L^{-1}) is almost isotonic with blood plasma.

A list of particularly important **metabolites** in the blood plasma is given on the right.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Functions of the blood

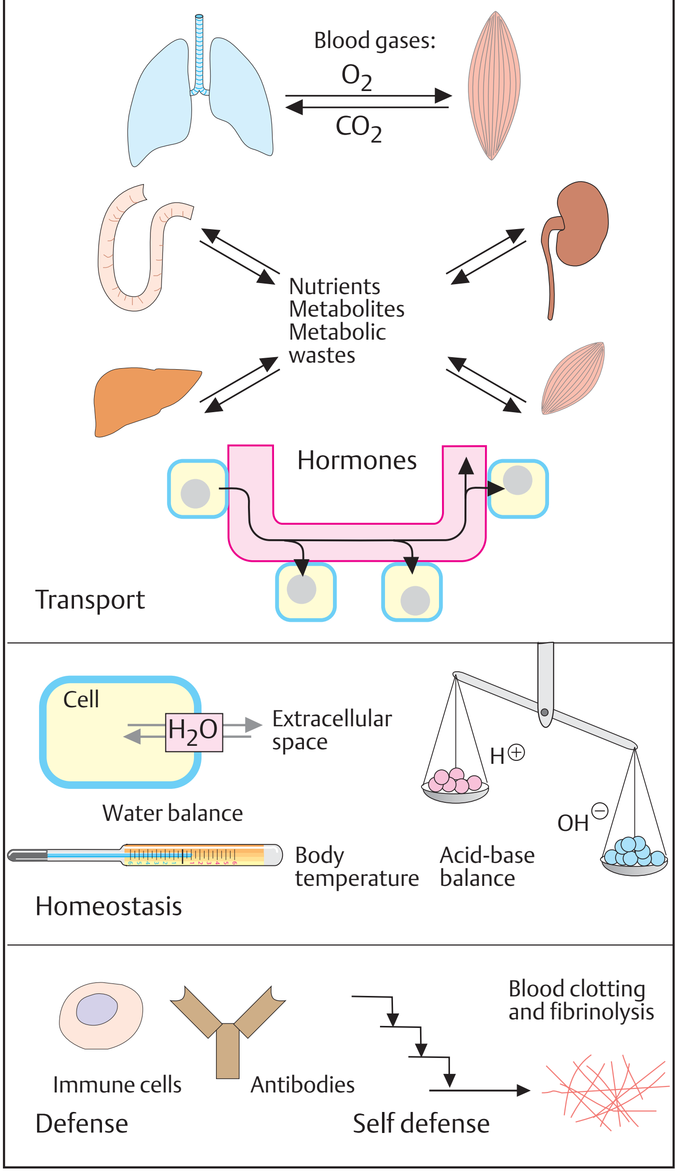
Blood gases:
O_2
CO_2
Nutrients
Metabolites
Metabolic
wastes
Hormones
Transport
Cell
H_2O
Extracellular space
Water balance
$H^\oplus$
$OH^\ominus$
Body temperature
Acid-base balance
Homeostasis
Immune cells
Antibodies
Defense
Blood clotting and fibrinolysis
Self defense

B. Cellular elements

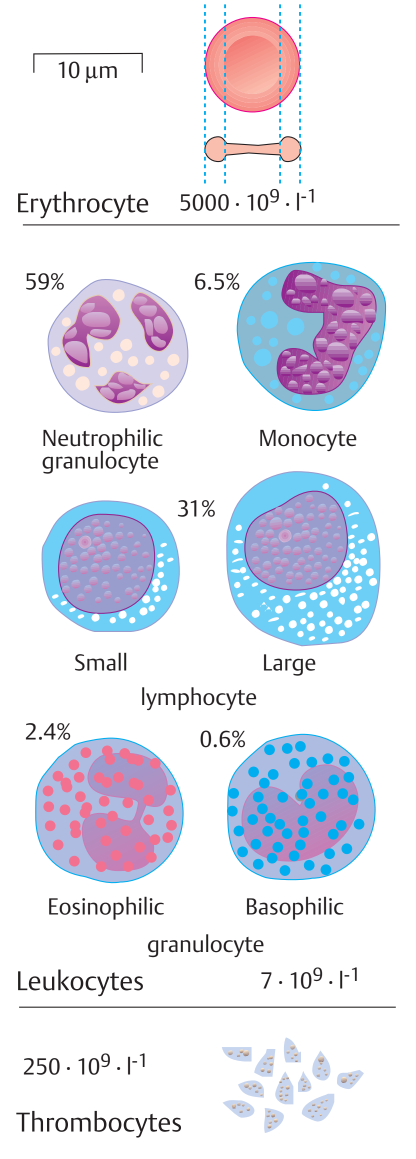
10 µm
Erythrocyte
$5000 \cdot 10^9 \cdot l^{-1}$
59%
6.5%
Neutrophilic granulocyte
Monocyte
31%
Small
Large
lymphocyte
2.4%
0.6%
Eosinophilic
Basophilic
granulocyte
Leukocytes
$7 \cdot 10^9 \cdot l^{-1}$
$250 \cdot 10^9 \cdot l^{-1}$
Thrombocytes

C. Blood plasma: composition

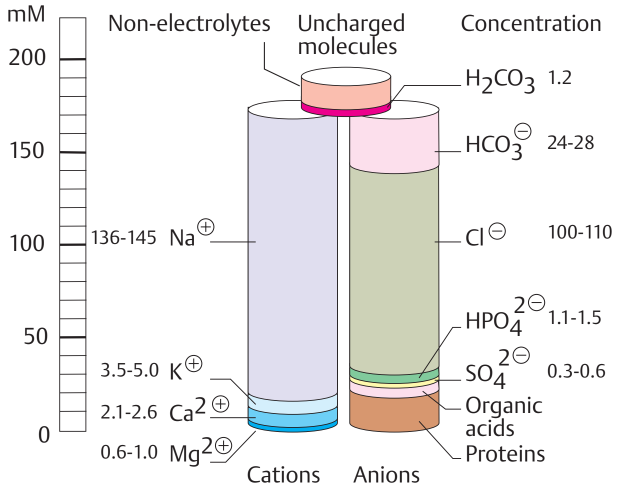

Metabolite	Concentration (mM)
Glucose	3.6 – 6.1
Lactate	0.4 – 1.8
Pyruvate	0.07 – 0.11
Urea	3.5 – 9.0
Uric acid	0.18 – 0.54
Creatinin	0.06 – 0.13
Amino acids	2.3 – 4.0
Ammonia	0.02 – 0.06
Lipids (total)	5.5 – 6.0 $g \cdot l^{-1}$
Triacylglycerols	1.0 – 1.3 $g \cdot l^{-1}$
Cholesterol	1.7 – 2.1 $g \cdot l^{-1}$

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Plasma proteins

Quantitatively, proteins are the most important part of the soluble components of the blood plasma. With concentrations of between 60 and 80 g L^{-1}, they constitute approximately 4% of the body's total protein. Their tasks include transport, regulation of the water balance, hemostasis, and defense against pathogens.

A. Plasma proteins ◑

Some 100 different proteins occur in human blood plasma. Based on their behavior during electrophoresis (see below), they are broadly divided into *five fractions:* **albumins** and α_1–, α_2–, β– and γ-**globulins**. Historically, the distinction between the albumins and globulins was based on differences in the proteins' solubility –albumins are soluble in pure water, whereas globulins only dissolve in the presence of salts.

The most frequent protein in the plasma, at around 45 g L^{-1}, is **albumin**. Due to its high concentration, it plays a crucial role in maintaining the blood's colloid osmotic pressure and represents an important amino acid reserve for the body. Albumin has binding sites for apolar substances and therefore functions as a transport protein for long-chain fatty acids, bilirubin, drugs, and some steroid hormones and vitamins. In addition, serum albumin binds Ca^{2+} and Mg^{2+} ions. It is the only important plasma protein that is not glycosylated.

The albumin fraction also includes *transthyretin* (prealbumin), which together with other proteins transports the hormone thyroxine and its metabolites.

The table also lists important **globulins** in blood plasma, with their mass and function. The α- and β-globulins are involved in the transport of lipids (lipoproteins; see p. 278), hormones, vitamins, and metal ions. In addition, they provide coagulation factors, protease inhibitors, and the proteins of the complement system (see p. 298). Soluble antibodies (immunoglobulins; see p. 300) make up the γ-globulin fraction.

Synthesis and degradation. Most plasma proteins are synthesized by the liver. Exceptions to this include the immunoglobulins, which are secreted by B lymphocytes known as plasma cells (see p. 302) and peptide hormones, which derive from endocrine gland cells.

With the exception of albumin, almost all plasma proteins are *glycoproteins.* They carry oligosaccharides in *N*-and *O*-glycosidic bonds (see p. 44). *N*-acetylneuraminic acid (sialic acid; see p. 38) often occurs as a terminal carbohydrate among sugar residues. *Neuraminidases* (sialidases) on the surface of the vascular endothelia gradually cleave the sialic acid residues and thereby release galactose units on the surfaces of the proteins. These *asialoglycoproteins* ("asialo-" = without sialic acid) are recognized and bound by galactose receptors on hepatocytes. In this way, the liver takes up aged plasma proteins by endocytosis and breaks them down. The oligosaccharides on the protein surfaces thus determine the half-life of plasma proteins, which is a period of days to weeks.

In healthy individuals, the concentration of plasma proteins is constant. Diseases in organs that are involved in protein synthesis and breakdown can shift the protein pattern. For example, via cytokines (see p. 392), severe injuries trigger increased synthesis of *acute-phase proteins,* which include C-reactive protein, haptoglobin, fibrinogen, complement factor C-3, and others. The concentrations of individual proteins are altered in some diseases (known as *dysproteinemias*).

B. Carrier electrophoresis ◑

Proteins and other electrically charged macromolecules can be separated using electrophoresis (see also pp. 78, 262). Among the various procedures used, **carrier electrophoresis** on cellulose acetate foil (CAF) is particularly simple. Using this method, serum proteins—which at slightly alkaline pH values all move towards the anode, due to their excess of negative charges—can be separated into the five fractions mentioned. After the proteins have been stained with dyes, the resulting bands can be quantitatively assessed using densitometry.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Plasma proteins

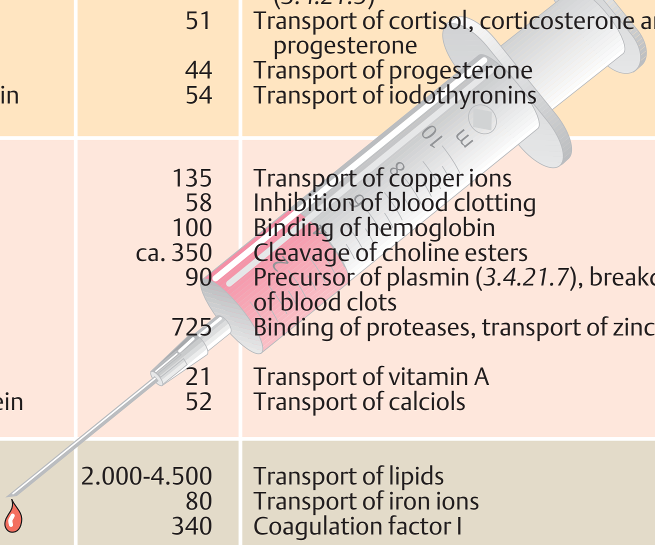

Group	Protein	M_r in kDa	Function
Albumins:	Transthyretin	50-66	Transport of thyroxin and triiodothyronin
	Albumin: 45 g · l^{-1}	67	Maintenance of osmotic pressure; transport of fatty acids, bilirubin, bile acids, steroid hormones, pharmaceuticals and inorganic ions.
α_1-Globulins:	Antitrypsin	51	Inhibition of trypsin and other proteases
	Antichymotrypsin	58-68	Inhibition of chymotrypsin
	Lipoprotein (HDL)	200-400	Transport of lipids
	Prothrombin	72	Coagulation factor II, thrombin precursor (*3.4.21.5*)
	Transcortin	51	Transport of cortisol, corticosterone and progesterone
	Acid glycoprotein	44	Transport of progesterone
	Thyroxin-binding globulin	54	Transport of iodothyronins
α_2-Globulins:	Ceruloplasmin	135	Transport of copper ions
	Antithrombin III	58	Inhibition of blood clotting
	Haptoglobin	100	Binding of hemoglobin
	Cholinesterase (3.1.1.8)	ca. 350	Cleavage of choline esters
	Plasminogen	90	Precursor of plasmin (*3.4.21.7*), breakdown of blood clots
	Macroglobulin	725	Binding of proteases, transport of zinc ions
	Retinol-binding protein	21	Transport of vitamin A
	Vitamin D-binding protein	52	Transport of calciols
β-Globulins:	Lipoprotein (LDL)	2.000-4.500	Transport of lipids
	Transferrin	80	Transport of iron ions
	Fibrinogen	340	Coagulation factor I
	Sex hormone-binding globulin	65	Transport of testosterone and estradiol
	Transcobalamin	38	Transport of vitamin B_{12}
	C-reactive protein	110	Complement activation
γ-Globulins:	IgG	150	Late antibodies
	IgA	162	Mucosa-protecting antibodies
	IgM	900	Early antibodies
	IgD	172	B-lymphocyte receptors
	IgE	196	Reagins

B. Electrophoresis

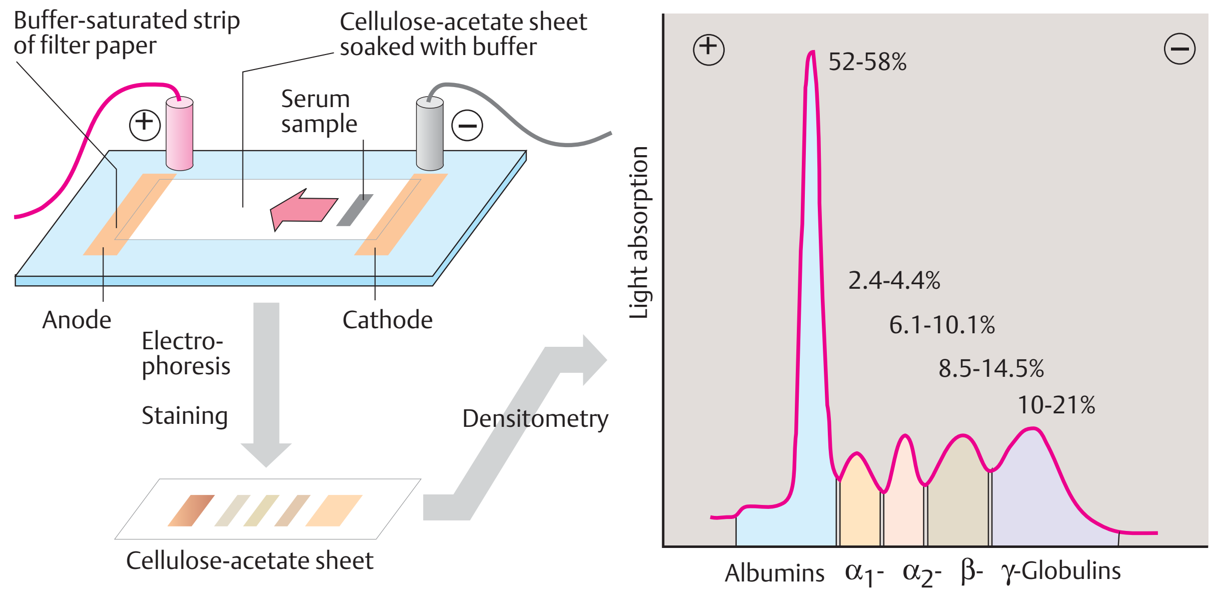
Buffer-saturated strip of filter paper
Cellulose-acetate sheet soaked with buffer
Serum sample
+
−
Anode
Cathode
Electro-phoresis
Staining
Densitometry
Cellulose-acetate sheet
Light absorption
+
−
52-58%
2.4-4.4%
6.1-10.1%
8.5-14.5%
10-21%
Albumins
α1-
α2-
β-
γ-Globulins

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Lipoproteins

Most lipids are barely soluble in water, and many have amphipathic properties. In the blood, free triacylglycerols would coalesce into drops that could cause fat embolisms. By contrast, amphipathic lipids would be deposited in the blood cells' membranes and would dissolve them. Special precautions are therefore needed for lipid transport in the blood. While long-chain fatty acids are bound to albumin and short-chain ones are dissolved in the plasma (see p. 276), other lipids are transported in **lipoprotein complexes**, of which there several types in the blood plasma, with different sizes and composition.

A. Composition of lipoprotein complexes ◑

Lipoproteins are spherical or discoid aggregates of **lipids** and **apoproteins**. They consist of a nucleus of apolar lipids (triacylglycerols and cholesterol esters) surrounded by a single-layered shell approximately 2 nm thick of amphipathic lipids (phospholipids and cholesterol; the example shown here is LDL). The shell, in which the apoproteins are also deposited, gives the surfaces of the particles polar properties and thereby prevents them from aggregating into large particles. The larger the lipid nucleus of a lipoprotein is—i. e., the larger the number of apolar lipids it contains—the lower its density is.

Lipoproteins are classified into five groups. In order of decreasing size and increasing density, these are: *chylomicrons, VLDLs* (very-low-density lipoproteins), *IDLs* (intermediate-density lipoproteins), *LDLs* (low-density lipoproteins), and *HDLs* (high-density lipoproteins). The proportions of apoproteins range from 1% in chylomicrons to over 50% in HDLs. These proteins serve less for solubility purposes, but rather function as recognition molecules for the membrane receptors and enzymes that are involved in lipid exchange.

B. Transport functions

The classes of lipoproteins differ not only in their composition, but also in the ways in which they originate and function.

The **chylomicrons** take care of the transport of triacylglycerols from the intestine to the tissues. They are formed in the intestinal mucosa and reach the blood via the lymphatic system (see p. 266). In the peripheral vessels—particularly in muscle and adipose tissue—*lipoprotein lipase* [1] on the surface of the vascular endothelia hydrolyzes most of the triacylglycerols. Chylomicron breakdown is activated by the transfer of apoproteins E and C from HDL. While the fatty acids released and the glycerol are taken up by the cells, the chylomicrons gradually become converted into **chylomicron remnants**, which are ultimately removed from the blood by the liver.

VLDLs, **IDLs**, and **LDLs** are closely related to one another. VLDLs formed in the liver (see p. 312) transport triacylglycerols, cholesterol, and phospholipids to other tissues. Like chylomicrons, they are gradually converted into IDL and LDL under the influence of *lipoprotein lipase* [1]. This process is also stimulated by HDL. Cells that have a demand for cholesterol bind LDL through an interaction between their **LDL receptor** and ApoB-100, and then take up the complete particle through **receptor-mediated endocytosis**. This type of transport is mediated by depressions in the membrane (*"coated pits"*), the interior of which is lined with the protein *clathrin.* After LDL binding, clathrin promotes invagination of the pits and pinching off of vesicles (*"coated vesicles"*). The clathrin then dissociates off and is reused. After fusion of the vesicle with lysosomes, the LDL particles are broken down (see p. 234), and cholesterol and other lipids are used by the cells.

The **HDLs** also originate in the liver. They return the excess cholesterol formed in the tissues to the liver. While it is being transported, cholesterol is acylated by *lecithin cholesterol acyltransferase* (LCAT). The cholesterol esters formed are no longer amphipathic and can be transported in the core of the lipoproteins. In addition, HDLs promote chylomicron and VLDL turnover by exchanging lipids and apoproteins with them (see above).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Composition of lipoprotein complexes

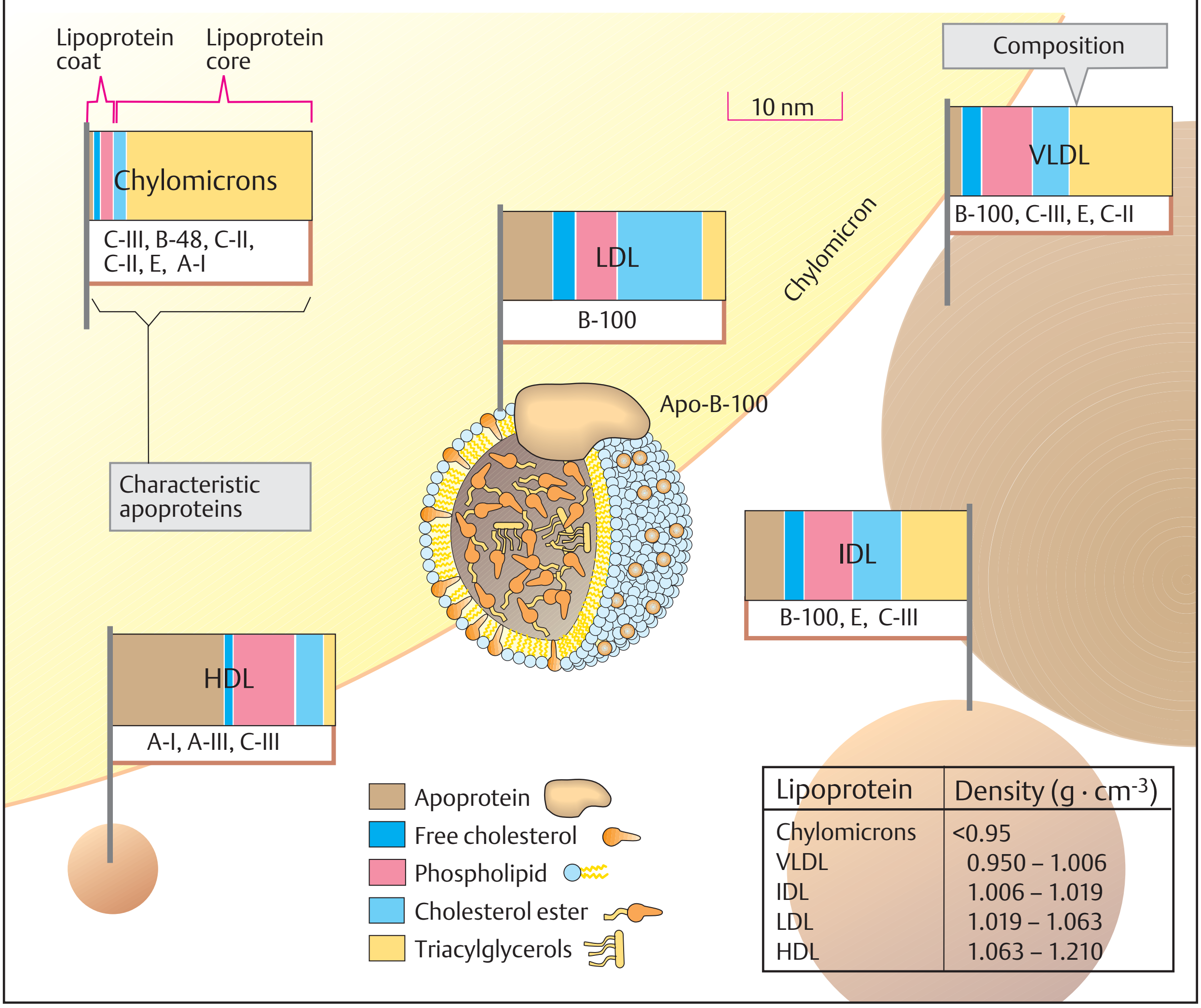
Lipoprotein coat
Lipoprotein core
Composition
10 nm
Chylomicrons
C-III, B-48, C-II, C-II, E, A-I
VLDL
B-100, C-III, E, C-II
LDL
B-100
Chylomicron
Characteristic apoproteins
Apo-B-100
IDL
B-100, E, C-III
HDL
A-I, A-III, C-III
Apoprotein
Free cholesterol
Phospholipid
Cholesterol ester
Triacylglycerols

Lipoprotein	Density (g · cm^{-3})
Chylomicrons	<0.95
VLDL	0.950 – 1.006
IDL	1.006 – 1.019
LDL	1.019 – 1.063
HDL	1.063 – 1.210

B. Transport functions

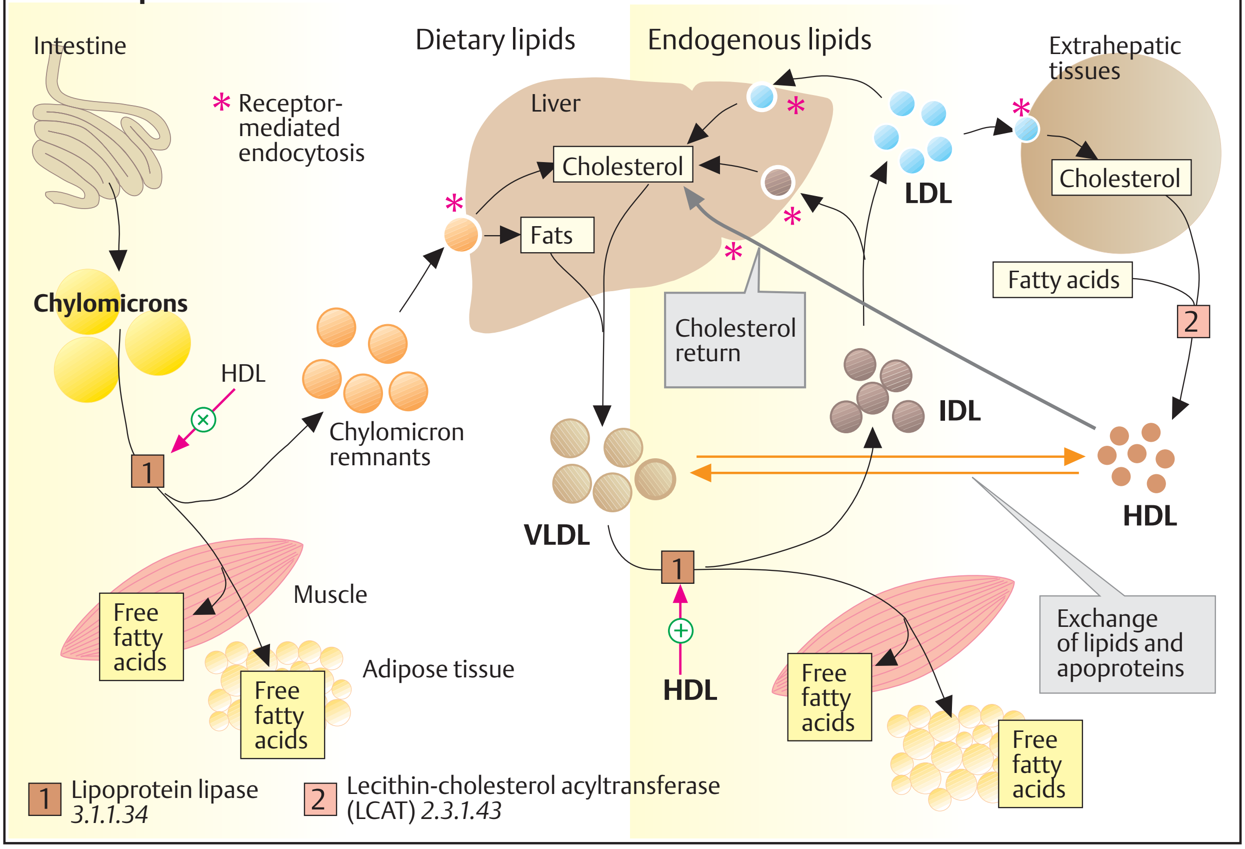
Intestine
Dietary lipids
Endogenous lipids
Extrahepatic tissues
* Receptor-mediated endocytosis
Liver
Cholesterol
Fats
LDL
Cholesterol
Chylomicrons
HDL
Cholesterol return
Fatty acids
2
Chylomicron remnants
IDL
1
VLDL
HDL
Muscle
Free fatty acids
Adipose tissue
Free fatty acids
1
HDL
Free fatty acids
Free fatty acids
Exchange of lipids and apoproteins

1 Lipoprotein lipase *3.1.1.34*

2 Lecithin-cholesterol acyltransferase (LCAT) *2.3.1.43*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Hemoglobin

The most important task of the red blood cells (erythrocytes) is to **transport** molecular oxygen (**O_2**) from the lungs into the tissues, and carbon dioxide (**CO_2**) from the tissues back into the lungs. To achieve this, the higher organisms require a special transport system, since O_2 is *poorly soluble in water.* For example, only around 3.2 mL O_2 is soluble in 1 L blood plasma. By contrast, the protein **hemoglobin** (Hb), contained in the erythrocytes, can bind a maximum of 220 mL O_2 per liter—70 times the physically soluble amount.

The Hb content of blood, at 140–180 g L^{-1} in men and 120–160 g L^{-1} in women, is twice as high as that of the plasma proteins (50–80 g L^{-1}). Hb is therefore also responsible for the majority of the blood proteins' pH buffer capacity (see p. 288).

A. Hemoglobin: structure ◑

In adults, hemoglobin (**HbA**; see below) is a *heterotetramer* consisting of two α-chains and two β-chains, each with masses of 16 kDa. The α- and β-chains have different sequences, but are similarly folded. Some 80% of the amino acid residues form *α-helices,* which are identified using the letters A–H.

Each subunit carries a **heme group** (formula on p. 106), with a central bivalent **iron ion**. When O_2 binds to the heme iron (**Oxygenation** of Hb) and when O_2 is released (**Deoxygenation**), the oxidation stage of the iron does *not* change. Oxidation of Fe^{2+} to Fe^{3+} only occurs occasionally. The oxidized form, *methemoglobin,* is then no longer able to bind O_2. The proportion of Met-Hb is kept low by reduction (see p. 284) and usually amounts to only 1–2%.

Four of the six coordination sites of the iron in hemoglobin are occupied by the nitrogen atoms of the pyrrol rings, and another is occupied by a histidine residue of the globin (the *proximal histidine*). The iron's sixth site is coordinated with oxygen in oxyhemoglobin and with H_2O in deoxyhemoglobin.

B. Hemoglobin: allosteric effects ◑

Like aspartate carbamoyltransferase (see p. 116), Hb can exist in two different states (*conformations*), known as the T form and the R form. The **T form** (for *tense;* left) and has a much *lower O_2* affinity than the R form (for *relaxed;* right).

Binding of O_2 to one of the subunits of the T form leads to a local conformational change that weakens the association between the subunits. Increasing O_2 partial pressure thus means that more and more molecules convert to the higher–affinity R form. This **cooperative interaction** between the subunits increases the O_2 affinity of Hb with increasing O_2 concentrations—i. e., the O_2 saturation curve is **sigmoidal** (see p. 282).

Various **allosteric effectors** influence the equilibrium between the T and R forms and thereby regulate the O_2 binding behavior of hemoglobin (yellow arrows). The most important effectors are *CO_2, H^+*, and *2,3-bisphosphoglycerate* (see p. 282).

Further information

As mentioned above, hemoglobin in adults consists of two α- and two β-chains. In addition to this main form (**HbA_1**, $\alpha_2\beta_2$), adult blood also contains small amounts of a second form with a higher O_2 affinity in which the β-chains are replaced by δ-chains (**HbA_2**, $\alpha_2\delta_2$). Two other forms occur during embryonic and fetal development. In the first three months, **embryonic hemoglobins** are formed, with the structure $\zeta_2\varepsilon_2$ and $\alpha_2\varepsilon_2$. Up to the time of birth, **fetal hemoglobin** then predominates (HbF, $\alpha_2\gamma_2$), and it is gradually replaced by HbA during the first few months of life. Embryonic and fetal hemoglobins have higher O_2 affinities than HbA, as they have to take up oxygen from the maternal circulation.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Hemoglobin: structure

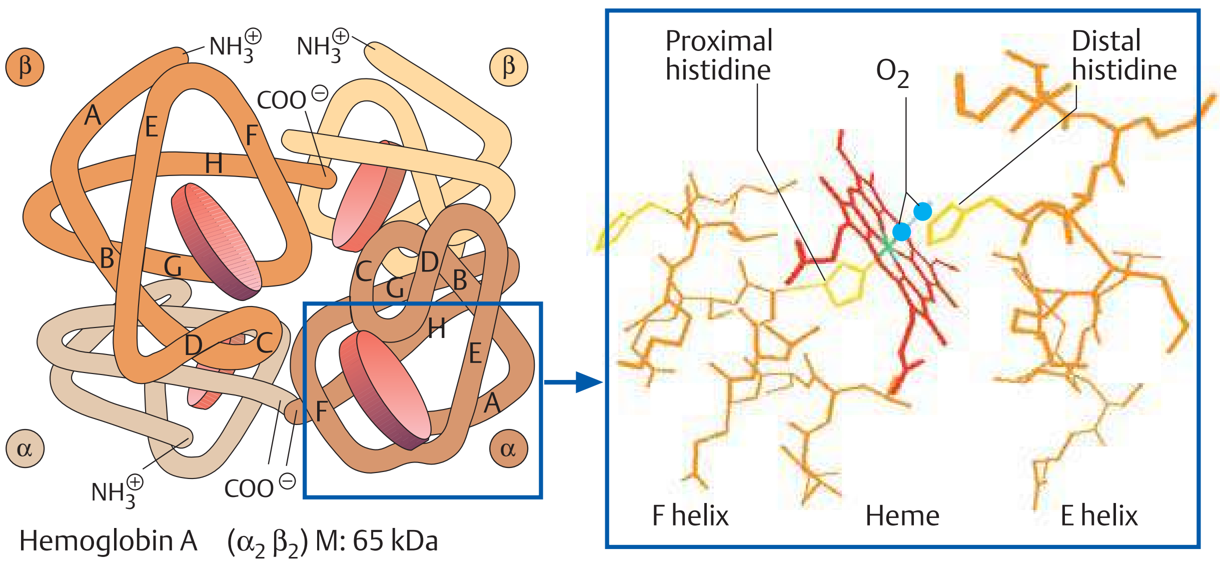
β
NH$_3^{\oplus}$
NH$_3^{\oplus}$
β
COO$^{\ominus}$
A
E
F
H
B
G
C
G
D
B
H
E
D
C
F
A
α
α
NH$_3^{\oplus}$
COO$^{\ominus}$
Hemoglobin A (α$_2$ β$_2$) M: 65 kDa
Proximal histidine
O$_2$
Distal histidine
F helix
Heme
E helix

B. Hemoglobin: allosteric effects

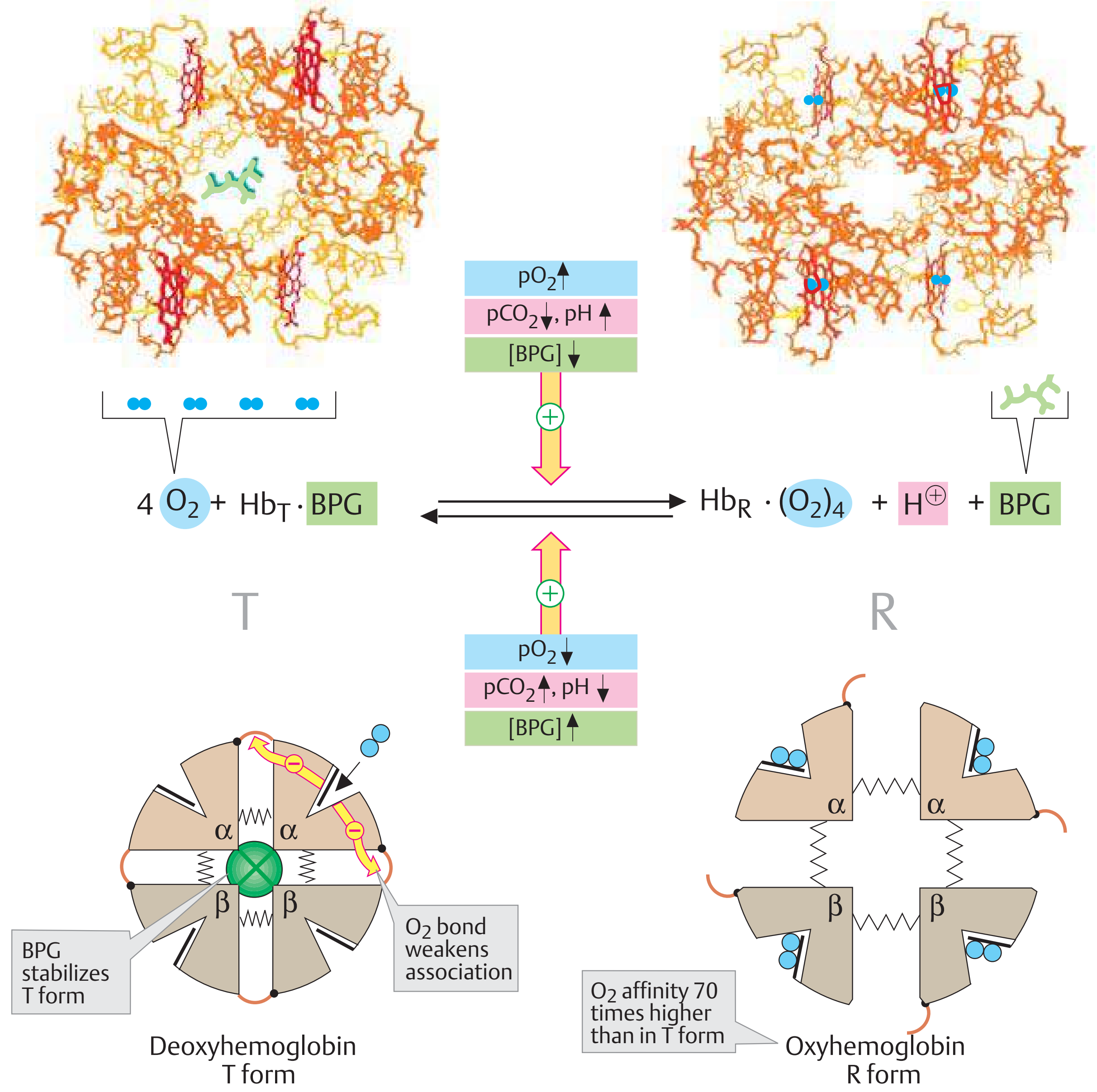
pO$_2$↑
pCO$_2$↓, pH↑
[BPG]↓
4 O$_2$ + Hb$_T$ · BPG ⇄ Hb$_R$ · (O$_2$)$_4$ + H$^{\oplus}$ + BPG
T
R
pO$_2$↓
pCO$_2$↑, pH↓
[BPG]↑
α
α
β
β
BPG stabilizes T form
O$_2$ bond weakens association
Deoxyhemoglobin T form
α
α
β
β
O$_2$ affinity 70 times higher than in T form
Oxyhemoglobin R form

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Gas transport

Most tissues are constantly dependent on a supply of molecular oxygen (O_2) to maintain their oxidative metabolism. Due to its poor solubility, O_2 is bound to hemoglobin for transport in the blood (see p. 280). This not only increases the oxygen transport capacity, but also allows regulation of O_2 uptake in the lungs and O_2 release into tissues.

A. Regulation of O_2 transport ◑

When an enzyme reacts to effectors (substrates, activators, or inhibitors) with conformational changes that increase or reduce its activity, it is said to show *allosteric behavior* (see p. 116). Allosteric enzymes are usually oligomers with several subunits that mutually influence each other.

Although **hemoglobin** is not an enzyme (it releases the bound oxygen without changing it), it has all the characteristics of an allosteric protein. Its effectors include oxygen, which as a *positive homotropic effector* promotes its own binding. The **O_2 saturation curve** of hemoglobin is therefore markedly sigmoidal in shape (**2**, curve 2). The non-sigmoidal saturation curve of the muscular protein **myoglobin** is shown for comparison (curve 1). The structure of myoglobin (see p. 336) is similar to that of a subunit of hemoglobin, but as a monomer it does not exhibit any allosteric behavior.

CO_2, H^+, and a special metabolite of erythrocytes—**2,3-bisphosphoglycerate (BPG)**—act as *heterotropic effectors* of hemoglobin. BPG is synthesized from *1,3-bisphosphoglycerate,* an intermediate of glycolysis (see p. 150), and it can be returned to glycolysis again by breakdown into 2-phosphoglycerate (**1**), with loss of an ATP.

BPG binds selectively to *deoxy-Hb,* thereby increasing its amount of equilibrium. The result is *increased O_2 release* at constant pO_2. In the diagram, this corresponds to a **right shift** of the saturation curve (**2**, curve 3). CO_2 and H^+ act in the same direction as BPG. Their influence on the position of the curve has long been known as the **Bohr effect**.

The effects of CO_2 and BPG are *additive.* In the presence of both effectors, the saturation curve of isolated Hb is similar to that of whole blood (curve 4).

B. Hemoglobin and CO_2 transport ◑

Hemoglobin is also decisively involved in the transport of carbon dioxide (CO_2) from the tissues to the lungs.

Some 5% of the CO_2 arising in the tissues is covalently bound to the N terminus of hemoglobin and transported as *carbaminohemoglobin* (not shown). About 90% of the CO_2 is first converted in the periphery into *hydrogen carbonate* (HCO_3^-), which is more soluble (bottom). In the lungs (top), CO_2 is regenerated again from HCO_3^- and can then be exhaled.

These two processes are coupled to the oxygenation and deoxygenation of Hb. *Deoxy-Hb* is a stronger base than oxy-Hb. It therefore binds additional protons (about 0.7 H^+ per tetramer), which promotes the formation of HCO_3^- from CO_2 in the peripheral tissues. The resulting HCO_3^- is released into the plasma via an antiporter in the erythrocyte membrane in exchange for Cl^-, and passes from the plasma to the lungs. In the lungs, the reactions described above then proceed in reverse order: deoxy-Hb is oxygenated and releases protons. The protons shift the HCO_3/CO_2 equilibrium to the left and thereby promote CO_2 release.

O_2 binding to Hb is regulated by H^+ ions (i.e., by the pH value) via the same mechanism. High concentrations of CO_2 such as those in tissues with intensive metabolism locally increase the H^+ concentration and thereby reduce hemoglobin's O_2 affinity (Bohr effect; see above). This leads to increased O_2 release and thus to an improved oxygen supply.

The adjustment of the equilibrium between CO_2 and HCO_3^- is relatively slow in the uncatalyzed state. It is therefore accelerated in the erythrocytes by *carbonate dehydratase* (carbonic anhydrase) [1])—an enzyme that occurs in high concentrations in the erythrocytes.

All rights reserved. Usage subject to terms and conditions of license.

A. Regulation of O_2 transport

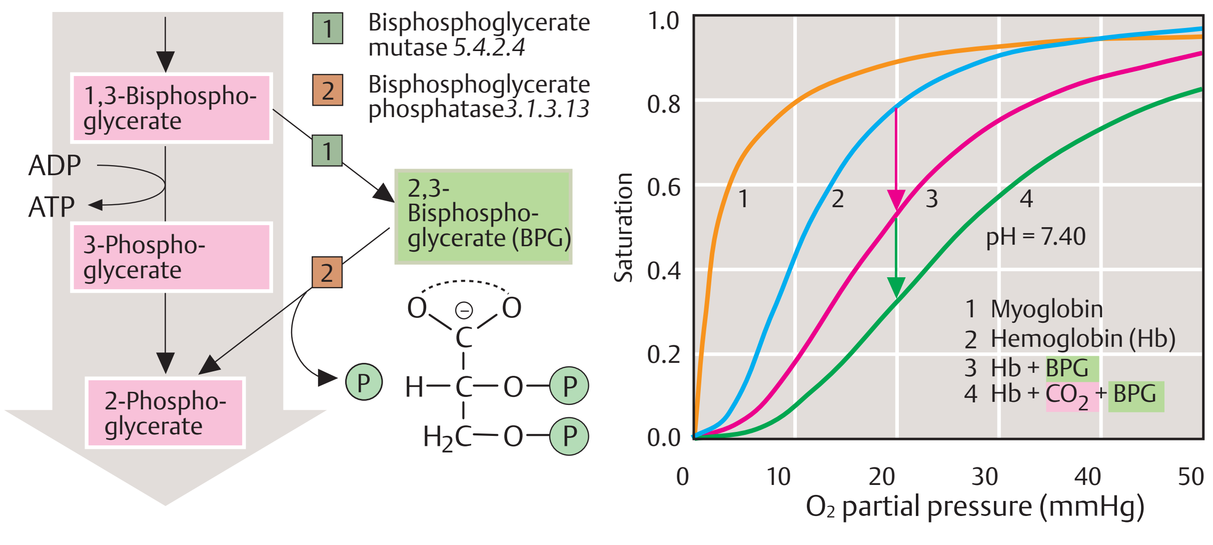
1 Bisphosphoglycerate mutase 5.4.2.4
2 Bisphosphoglycerate phosphatase 3.1.3.13
1,3-Bisphosphoglycerate
ADP
ATP
3-Phosphoglycerate
2-Phosphoglycerate
1
2
2,3-Bisphosphoglycerate (BPG)
P
O ⊖ O
C
H—C—O—P
H2C—O—P
1.0
0.8
0.6
0.4
0.2
0.0
Saturation
0
10
20
30
40
50
O2 partial pressure (mmHg)
1
2
3
4
pH = 7.40
1 Myoglobin
2 Hemoglobin (Hb)
3 Hb + BPG
4 Hb + CO2 + BPG

1. BPG metabolism

2. Saturation curves

D. Hemoglobin and CO_2 transport

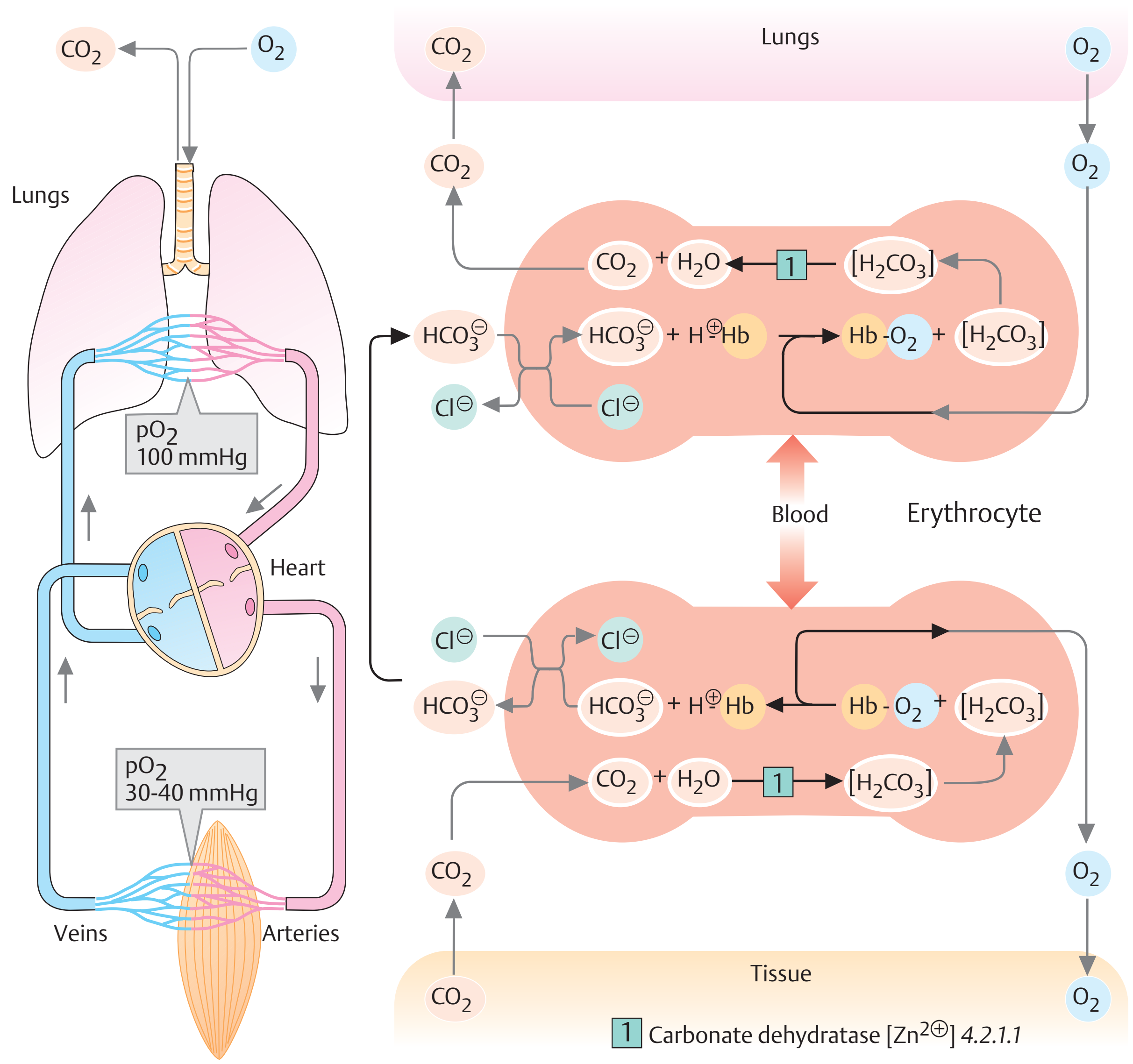
CO2
O2
Lungs
pO2 100 mmHg
Heart
pO2 30-40 mmHg
Veins
Arteries
Lungs
CO2
O2
CO2 + H2O
1
[H2CO3]
HCO3⊖
HCO3⊖ + H⊕Hb
Hb-O2 + [H2CO3]
Cl⊖
Cl⊖
Blood
Erythrocyte
Cl⊖
Cl⊖
HCO3⊖
HCO3⊖ + H⊕Hb
Hb-O2 + [H2CO3]
CO2 + H2O
1
[H2CO3]
CO2
O2
Tissue
1 Carbonate dehydratase [Zn2⊕] 4.2.1.1

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Erythrocyte metabolism

Cells living in aerobic conditions are dependent on molecular oxygen for energy production. On the other hand, O_2 constantly gives rise to small quantities of toxic substances known as **reactive oxygen species** (ROS). These substances are powerful *oxidation agents* or extremely reactive *free radicals* (see p. 32), which damage cellular structures and functional molecules. Due to their role in O_2 transport, the erythrocytes are constantly exposed to high concentrations of O_2 and are therefore particularly at risk from ROS.

A. Reactive oxygen species ◑

The dioxygen molecule (O_2) contains two unpaired electrons—i. e., it is a *diradical.* Despite this, O_2 is relatively stable due to its special electron arrangement. However, if the molecule takes up an extra electron (**a**), the highly reactive **superoxide radical** (O_2^-) arises. Another reduction step (**b**) leads to the **peroxide anion** (O_2^{2-}), which easily binds protons and thus becomes **hydrogen peroxide** (H_2O_2). Inclusion of a third electron (**c**) leads to cleavage of the molecule into the ions O^{2-} and O^-. While O^{2-} can form **water** by taking up two protons, protonation of O^- provides the extremely dangerous **hydroxy radical** (OH). A fourth electron transfer and subsequent protonation also convert O^- into water.

The synthesis of ROS can be catalyzed by iron ions, for example. Reaction of O_2 with FMN or FAD (see p. 32) also constantly produces ROS. By contrast, reduction of O_2 by *cytochrome c-oxidase* (see p. 140) is "clean," as the enzyme does not release the intermediates. In addition to antioxidants (**B**), **enzymes** also provide protection against ROS: *superoxide dismutase* [1] breaks down ("disproportionates") two superoxide molecules into O_2 and the less damaging H_2O_2. The latter is in turn disproportionated into O_2 and H_2O by heme-containing *catalase* [2].

B. Biological antioxidants ◑

To protect them against ROS and other radicals, all cells contain **antioxidants**. These are *reducing agents* that react easily with oxidative substances and thus protect more important molecules from oxidation. Biological antioxidants include vitamins C and E (see pp. 364, 368), coenzyme Q (see p. 104), and several carotenoids (see pp. 132, 364). Bilirubin, which is formed during heme degradation (see p. 194), also serves for protection against oxidation.

Glutathione, a tripeptide that occurs in high concentrations in almost all cells, is particularly important. Glutathione (sequence: Glu–Cys–Gly) contains an atypical γ-peptide bond between Glu and Cys. The thiol group of the cysteine residue is redox-active. Two molecules of the reduced form (GSH, top) are bound to the disulfide (GSSG, bottom) during oxidation.

C. Erythrocyte metabolism ◑

Erythrocytes also have systems that can inactivate ROS (superoxide dismutase, catalase, GSH). They are also able to repair damage caused by ROS. This requires products that are supplied by the erythrocytes' **maintenance metabolism**, which basically only involves anaerobic glycolysis (see p. 150) and the pentose phosphate pathway (PPP; see p. 152).

The **ATP** formed during glycolysis serves mainly to supply *Na^+/K^+-ATPase,* which maintains the erythrocytes' membrane potential. The allosteric effector **2,3-BPG** (see p. 282) is also derived from glycolysis. The PPP supplies **NADPH+H$^+$**, which is needed to regenerate **glutathione** (GSH) from GSSG with the help of *glutathione reductase* [3]. GSH, the most important antioxidant in the erythrocytes, serves as a coenzyme for *glutathione peroxidase* [5]. This selenium-containing enzyme detoxifies H_2O_2 and hydroperoxides, which arise during the reaction of ROS with unsaturated fatty acids in the erythrocyte membrane. The reduction of methemoglobin (Hb Fe^{3+}) to Hb (Hb Fe^{2+}, [4]) is carried out by GSH or ascorbate by a non-enzymatic pathway; however, there are also NAD(P)H-dependent Met-Hb reductases.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Reactive oxygen species

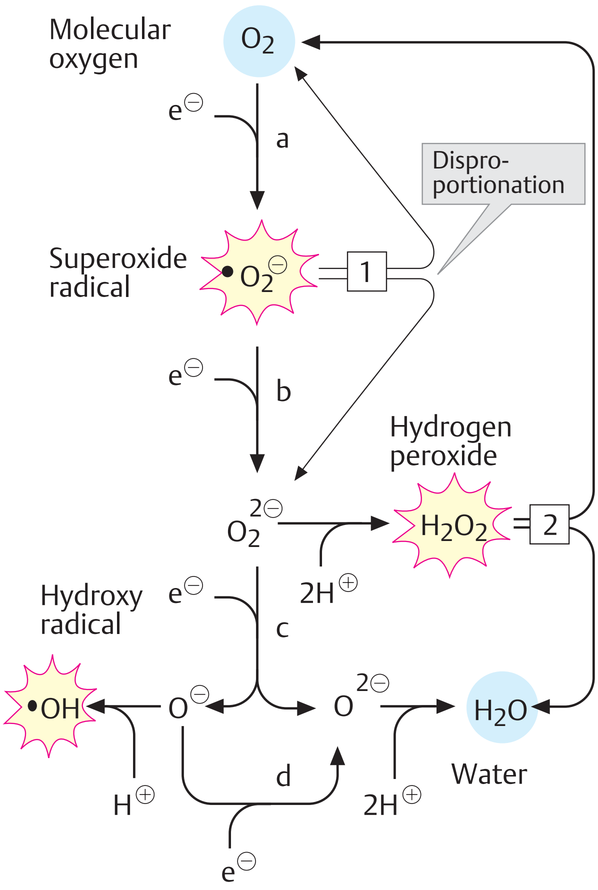
Molecular oxygen
O_2
$e^{\ominus}$
a
Superoxide radical
$\bullet O_2^{\ominus}$
1
Dispro-portionation
$e^{\ominus}$
b
Hydrogen peroxide
$O_2^{2\ominus}$
H_2O_2
2
$2H^{\oplus}$
Hydroxy radical
$e^{\ominus}$
c
$\bullet OH$
$O^{\ominus}$
$O^{2\ominus}$
H_2O
Water
d
$H^{\oplus}$
$2H^{\oplus}$
$e^{\ominus}$

1 Superoxide dismutase *1.15.1.1*

2 Catalase *1.11.1.6*

B. Biological antioxidants

Quinols and enols	α-Tocopherol (vitamin E) Ubiquinol (coenzyme Q) Ascorbic acid (vitamin C)
Carotenoids	β-Carotin Lycopin
Others	Glutathione Bilirubin

1. Examples

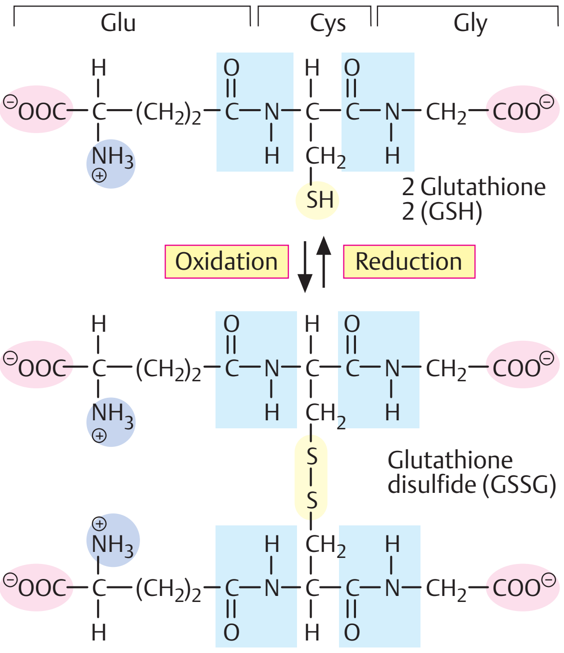
Glu
Cys
Gly
$^{\ominus}OOC-C-(CH_2)_2-C-N-C-C-N-CH_2-COO^{\ominus}$
$NH_3^{\oplus}$
SH
2 Glutathione
2 (GSH)
Oxidation
Reduction
S
S
Glutathione disulfide (GSSG)

2. Glutathione

C. Erythrocyte metabolism

3 Glutathione reductase [FAD] *1.6.4.2*

4 Methemoglobin reductase

5 Glutathione peroxidase [Se] *1.11.1.9/12*

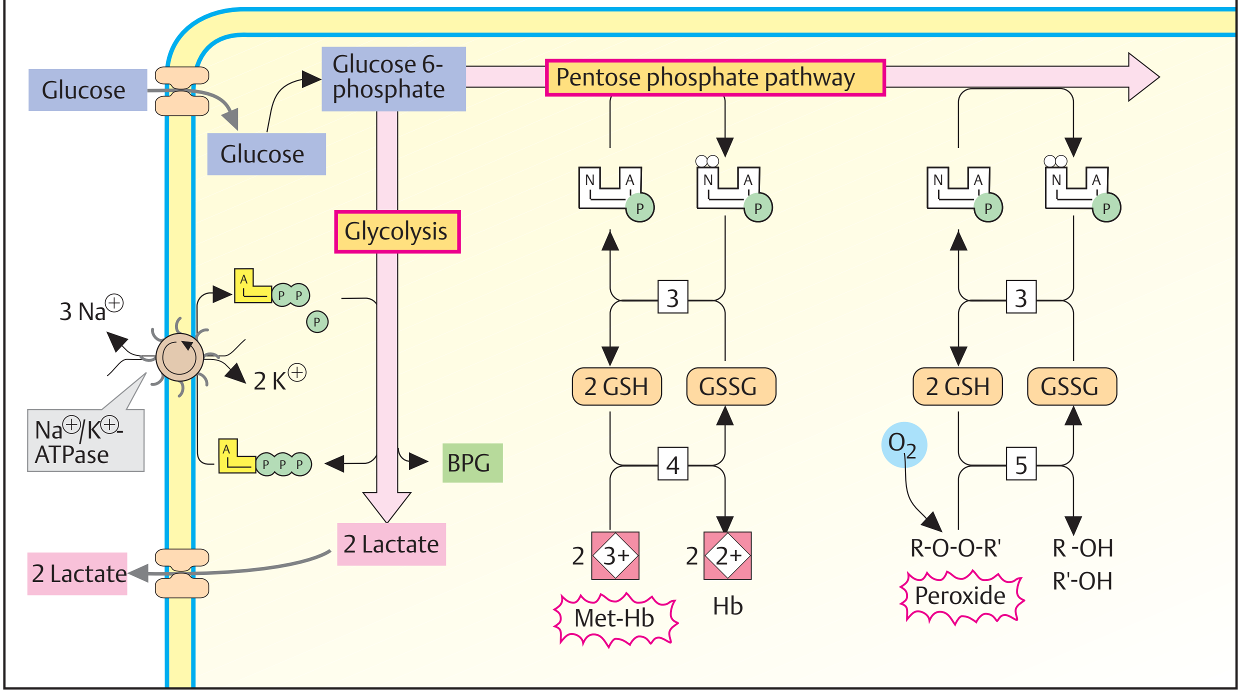
Glucose
Glucose 6-phosphate
Glucose
Pentose phosphate pathway
Glycolysis
N A P
3
3 $Na^{\oplus}$
2 $K^{\oplus}$
$Na^{\oplus}/K^{\oplus}$-ATPase
BPG
2 GSH
GSSG
4
5
O_2
2 3+
2 2+
Met-Hb
Hb
R-O-O-R'
Peroxide
R-OH
R'-OH
2 Lactate
2 Lactate

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Iron metabolism

A. Distribution of iron ◑

Iron (Fe) is quantitatively the most important trace element (see p. 362). The human body contains 4–5 g iron, which is almost exclusively present in protein-bound form. Approximately three-quarters of the total amount is found in **heme proteins** (see pp. 106, 192), mainly hemoglobin and myoglobin. About 1% of the iron is bound in **iron–sulfur clusters** (see p. 106), which function as cofactors in the respiratory chain, in photosynthesis, and in other redox chains. The remainder consists of iron in transport and storage proteins (transferrin, ferritin; see **B**).

B. Iron metabolism ◑

Iron can only be resorbed by the bowel in bivalent form (i. e., as Fe^{2+}). For this reason, reducing agents in food such as ascorbate (vitamin C; see p. 368) promote **iron uptake**. Via transporters on the luminal and basal side of the enterocytes, Fe^{2+} enters the blood, where it is bound by *transferrin*. Part of the iron that is taken up is stored in the bowel in the form of *ferritin* (see below). Heme groups can also be resorbed by the small intestine.

Most of the resorbed iron serves for the formation of red blood cells in the bone marrow (**erythropoiesis**, top). As discussed on p. 192, it is only in the final step of hem biosynthesis that Fe^{2+} is incorporated by *ferrochelatase* into the previously prepared tetrapyrrol framework.

In the blood, 2.5–3.0 g of hemoglobin iron circulates as a component of the erythrocytes (top right). Over the course of several months, the flexibility of the red blood cells constantly declines due to damage to the membrane and cytoskeleton. Old erythrocytes of this type are taken up by macrophages in the spleen and other organs and broken down. The organic part of the heme is oxidized into bilirubin (see p. 194), while the iron returns to the plasma pool. The quantity of heme iron recycled per day is much larger than the amount resorbed by the intestines.

Transferrin, a β-globulin with a mass of 80 kDa, serves to transport iron in the blood. This monomeric protein consists of two similar domains, each of which binds an Fe^{2+} ion very tightly. Similar iron transport proteins are found in secretions such as saliva, tears, and milk; these are known as lactoferrins (bottom right). Transferrin and the lactoferrins maintain the concentration of *free* iron in body fluids at values below 10^{-10} mol L^{-1}. This low level prevents bacteria that require free iron as an essential growth factor from proliferating in the body. Like LDLs (see p. 278), transferrin and the lactoferrins are taken up into cells by *receptor-mediated endocytosis.*

Excess iron is incorporated into **ferritin** and stored in this form in the liver and other organs. The ferritin molecule consists of 24 subunits and has the shape of a hollow sphere (bottom left). It takes up Fe^{2+} ions, which in the process are oxidized to Fe^{3+} and then deposited in the interior of the sphere as *ferrihydrate.* Each ferritin molecule is capable of storing several thousand iron ions in this way. In addition to ferritin, there is another storage form, **hemosiderin**, the function of which is not yet clear.

Further information

Disturbances of the iron metabolism are frequent and can lead to severe disease pictures.

Iron deficiency is usually due to blood loss, or more rarely to inadequate iron uptake. During pregnancy, increased demand can also cause iron deficiency states. In severe cases, reduced hemoglobin synthesis can lead to **anemia** (“iron-deficiency anemia”). In these patients, the erythrocytes are smaller and have less hemoglobin. As their membrane is also altered, they are prematurely eliminated in the spleen.

Disturbances resulting from raised iron concentrations are less frequent. Known as **hemochromatoses**, these conditions can have genetic causes, or may be due to repeated administration of blood transfusions. As the body has practically no means of excreting iron, more and more stored iron is deposited in the organs over time in patients with untreated hemochromatosis, ultimately leading to severe disturbances of organ function.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Distribution of iron

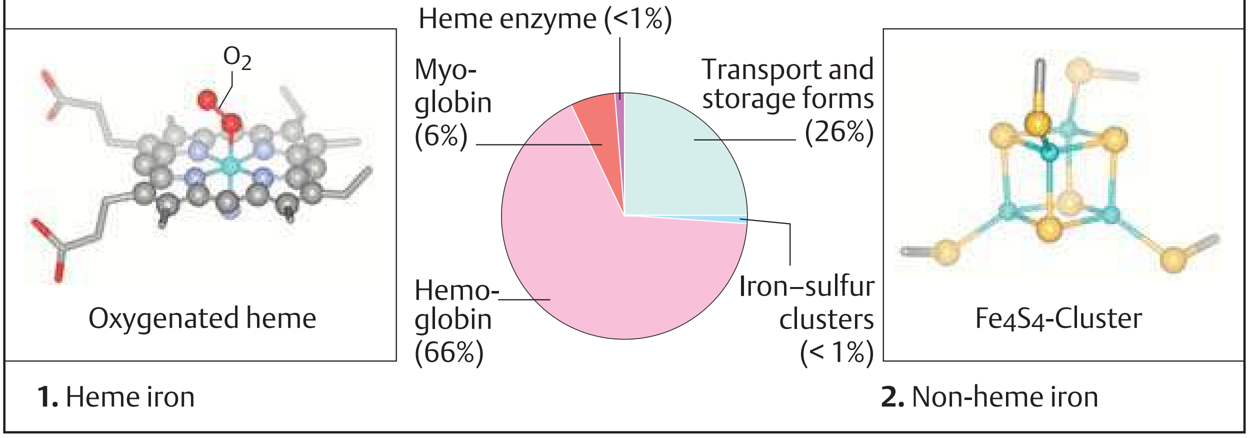
Heme enzyme (<1%)
O_2
Oxygenated heme
Myo-globin (6%)
Transport and storage forms (26%)
Hemo-globin (66%)
Iron–sulfur clusters (< 1%)
Fe_4S_4-Cluster
1. Heme iron
2. Non-heme iron

B. Iron metabolism

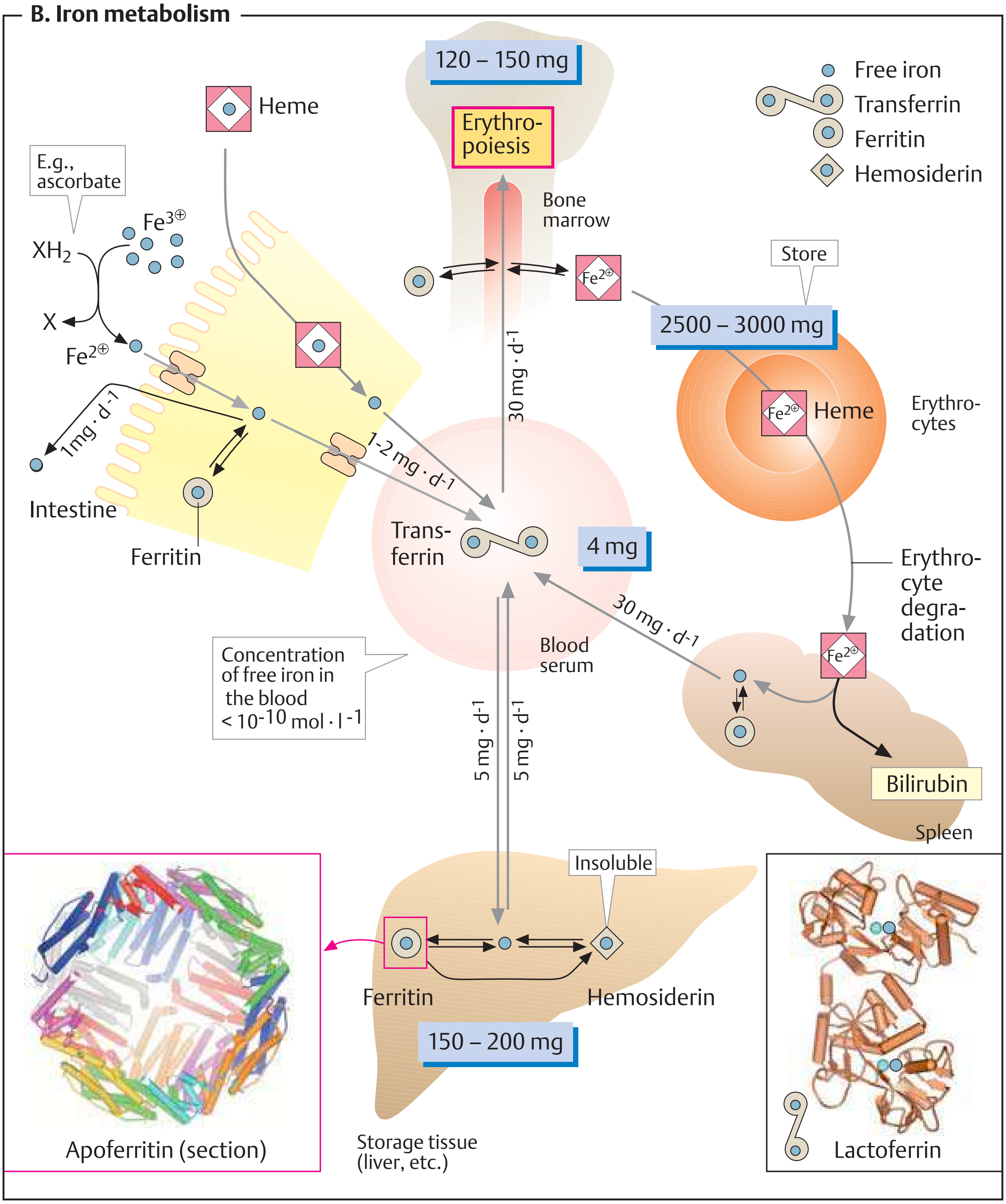
120 – 150 mg
Erythro-poiesis
Bone marrow
Heme
Free iron
Transferrin
Ferritin
Hemosiderin
E.g., ascorbate
XH_2
X
$Fe^{3⊕}$
$Fe^{2⊕}$
$1 mg \cdot d^{-1}$
Intestine
Ferritin
$1-2 mg \cdot d^{-1}$
$30 mg \cdot d^{-1}$
$Fe^{2⊕}$
Store
2500 – 3000 mg
$Fe^{2⊕}$
Heme
Erythro-cytes
Trans-ferrin
4 mg
Erythro-cyte degra-dation
$30 mg \cdot d^{-1}$
Blood serum
Concentration of free iron in the blood $< 10^{-10} mol \cdot l^{-1}$
$Fe^{2⊕}$
$5 mg \cdot d^{-1}$
$5 mg \cdot d^{-1}$
Bilirubin
Spleen
Insoluble
Ferritin
Hemosiderin
150 – 200 mg
Storage tissue (liver, etc.)
Apoferritin (section)
Lactoferrin

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Acid–base balance

A. Hydrogen ion concentration in the blood plasma ◑

The H^+ concentration in the blood and extracellular space is approximately 40 nM (4 10^{-8} mol L^{-1}). This corresponds to a pH of 7.40. The body tries to keep this value constant, as large shifts in pH are incompatible with life.

The pH value is kept constant by **buffer systems** that cushion minor disturbances in the *acid–base balance* (**C**). In the longer term, the decisive aspect is maintaining a balanced equilibrium between H^+ production and uptake and H^+ release. If the blood's buffering capacity is not sufficient, or if the acid–base balance is not in equilibrium—e. g., in kidney disease or during *hypoventilation* or *hyperventilation*—shifts in the plasma pH value can occur. A reduction by more than 0.03 units is known as **acidosis,** and an increase is called **alkalosis.**

B. Acid–base balance ◑

Protons are mainly derived from two sources—free acids in the diet and sulfur–containing amino acids. *Acids* taken up with food—e. g., citric acid, ascorbic acid, and phosphoric acid—already release protons in the alkaline pH of the intestinal tract. More important for proton balance, however, are the amino acids **methionine** and **cysteine**, which arise from protein degradation in the cells. Their S atoms are oxidized in the liver to form sulfuric acid, which supplies protons by dissociation into sulfate.

During anaerobic glycolysis in the muscles and erythrocytes, glucose is converted into **lactate**, releasing protons in the process (see p. 338). The synthesis of the **ketone bodies** acetoacetic acid and 3–hydroxybutyric acid in the liver (see p. 312) also releases protons. Normally, the amounts formed are small and of little influence on the proton balance. If acids are formed in large amounts, however (e. g., during *starvation* or in *diabetes mellitus;* see p. 160), they strain the buffer systems and can lead to a reduction in pH (**metabolic acidoses**; *lactacidosis* or *ketoacidosis*).

Only the **kidney** is capable of excreting protons in exchange for Na^+ ions (see p. 326). In the urine, the H^+ ions are buffered by NH_3 and phosphate.

C. Buffer systems in the plasma ◑

The **buffering capacity** of a buffer system depends on its concentration and its pK_a value. The strongest effect is achieved if the pH value corresponds to the buffer system's pK_a value (see p. 30). For this reason, weak acids with pK_a values of around 7 are best suited for buffering purposes in the blood.

The most important buffer in the blood is the **CO_2/bicarbonate buffer**. This consists of water, carbon dioxide (CO_2, the anhydride of carbonic acid H_2CO_3), and hydrogen carbonate (HCO_3^-, bicarbonate). The adjustment of the balance between CO_2 and HCO_3^- is accelerated by the zinc-containing enzyme *carbonate dehydratase* (carbonic anhydrase [1]; see also p. 282). At the pH value of the plasma, HCO_3^- and CO_2 are present in a ratio of about 20 : 1. However, the CO_2 in solution in the blood is in equilibrium with the gaseous CO_2 in the pulmonary alveoli. The CO_2/HCO_3^- system is therefore a powerful *open buffer system,* despite having a not entirely optimal pK_a value of 6.1. Faster or slower respiration increases or reduces CO_2 release in the lungs. This shifts the CO_2/HCO_3^- ratio and thus the plasma pH value (respiratory acidosis or alkalosis). In this way, respiration can compensate to a certain extent for changes in plasma pH values. However, it does *not* lead to the excretion of protons.

Due to their high concentration, **plasma proteins**—and **hemoglobin** in the erythrocytes in particular—provide about one-quarter of the blood's buffering capacity. The buffering effect of proteins involves contributions from all of the ionizable side chains. At the pH value of blood, the acidic amino acids (Asp, Glu) and histidine are particularly effective.

The second dissociation step in **phosphate** (H_2PO_4/HPO_4^{2-}) also contributes to the buffering capacity of the blood plasma. Although the pK_a value of this system is nearly optimal, its contribution remains small due to the low total concentration of phosphate in the blood (around 1 mM).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Hydrogen ion concentration in the blood plasma

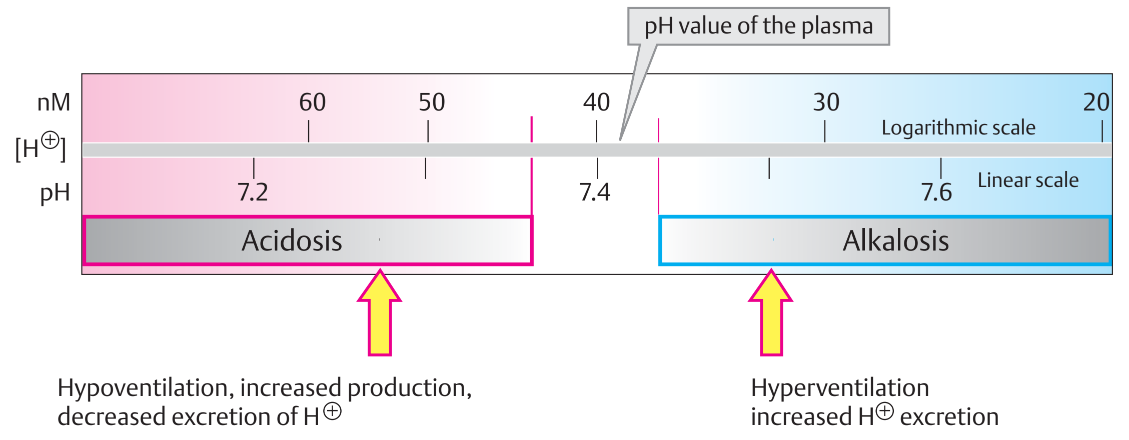
pH value of the plasma
nM
[H⊕]
pH
60
50
40
30
20
Logarithmic scale
7.2
7.4
7.6
Linear scale
Acidosis
Alkalosis
Hypoventilation, increased production,
decreased excretion of H⊕
Hyperventilation
increased H⊕ excretion

B. Acid–base balance

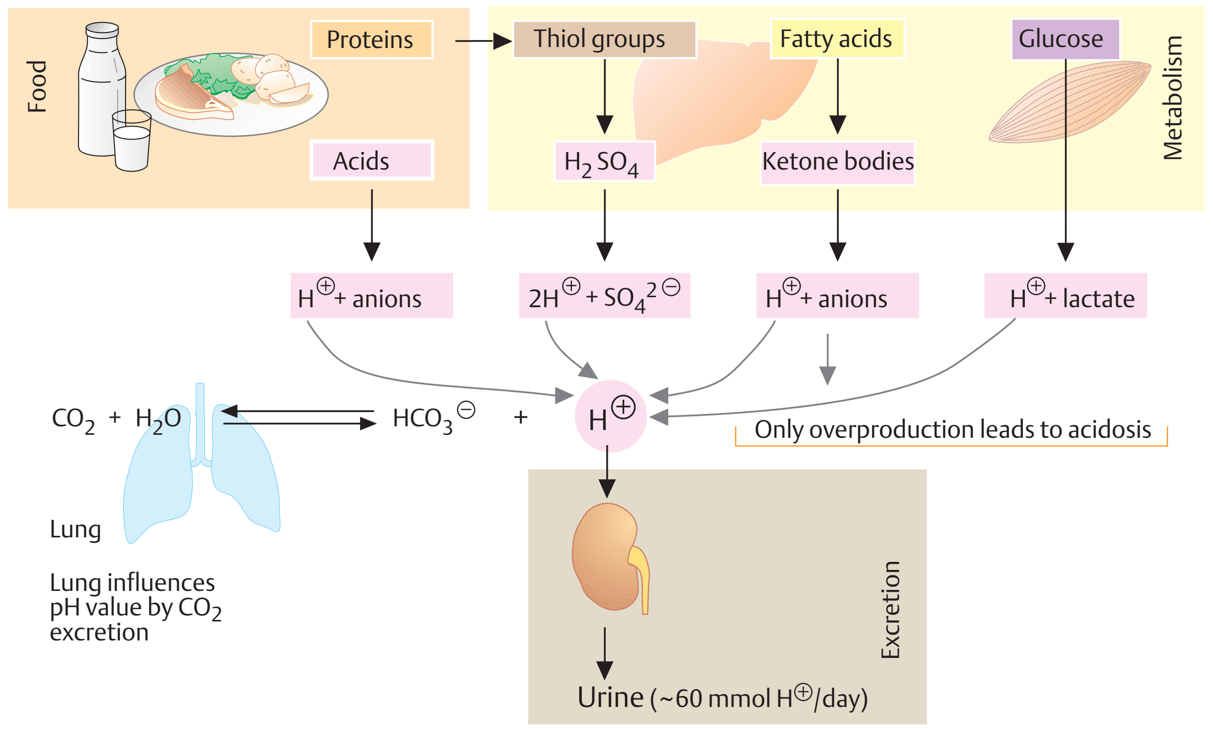
Food
Proteins
Acids
Thiol groups
H_2SO_4
Fatty acids
Ketone bodies
Glucose
Metabolism
H⊕ + anions
2H⊕ + SO_4^{2-}
H⊕ + anions
H⊕ + lactate
CO_2 + H_2O
HCO_3^-
+
H⊕
Only overproduction leads to acidosis
Lung
Lung influences
pH value by CO_2
excretion
Excretion
Urine (~60 mmol H⊕/day)

C. Buffer systems in the plasma

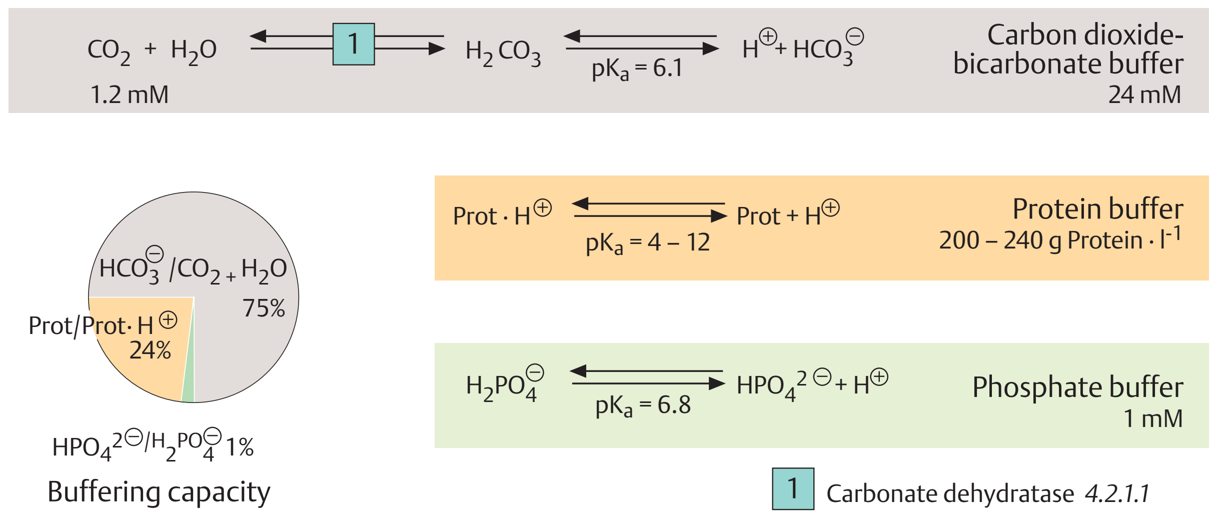
$CO_2 + H_2O$
1.2 mM
1
H_2CO_3
$pK_a = 6.1$
$H^+ + HCO_3^-$
Carbon dioxide-bicarbonate buffer
24 mM
$Prot \cdot H^+$
$Prot + H^+$
$pK_a = 4 - 12$
Protein buffer
200 – 240 g Protein · l^{-1}
$H_2PO_4^-$
$HPO_4^{2-} + H^+$
$pK_a = 6.8$
Phosphate buffer
1 mM
$HCO_3^-/CO_2 + H_2O$
75%
$Prot/Prot \cdot H^+$
24%
$HPO_4^{2-}/H_2PO_4^-$ 1%
Buffering capacity
1 Carbonate dehydratase 4.2.1.1

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Blood clotting

Following injury to blood vessels, **hemostasis** ensures that blood loss is minimized. Initially, thrombocyte activation leads to contraction of the injured vessel and the formation of a loose clot consisting of thrombocytes (**hemostasis**). Slightly later, the action of the enzyme *thrombin* leads to the formation and deposition in the thrombus of polymeric fibrin (**coagulation, blood clotting**). The coagulation process is discussed here in detail.

A. Blood clotting ◑

The most important reaction in blood clotting is the conversion, catalyzed by *thrombin,* of the soluble plasma protein *fibrinogen* (factor I) into polymeric *fibrin,* which is deposited as a fibrous network in the primary thrombus. Thrombin (factor IIa) is a serine proteinase (see p. 176) that cleaves small peptides from fibrinogen. This exposes binding sites that spontaneously allow the fibrin molecules to aggregate into polymers. Subsequent covalent cross-linking of fibrin by a *transglutaminase* (factor XIII) further stabilizes the thrombus.

Normally, thrombin is present in the blood as an inactive proenzyme (see p. 270). Prothrombin is activated in two different ways, both of which represent cascades of enzymatic reactions in which inactive proenzymes (zymogens, symbol: circle) are proteolytically converted into active **proteinases** (symbol: sector of a circle). The proteinases activate the next proenzyme in turn, and so on. Several steps in the cascade require additional **protein factors** (factors III, Va and VIIIa) as well as anionic **phospholipids** (PL; see below) and **Ca^{2+}** ions. Both pathways are activated by injuries to the vessel wall.

In the **extravascular pathway** (right), *tissue thromboplastin* (factor III), a membrane protein in the deeper layers of the vascular wall, activates coagulation factor VII. The activated form of this (VIIa) autocatalytically promotes its own synthesis and also generates the active factors IXa and Xa from their precursors. With the aid of factor VIIIa, PL, and Ca^{2+}, factor IXa produces additional Xa, which finally—with the support of Va, PL, and Ca^{2+}—releases active thrombin.

The **intravascular pathway** (left) is probably also triggered by vascular injuries. It leads in five steps via factors XIIa, XIa, IXa, and Xa to the activation of prothrombin. The significance of this pathway in vivo has been controversial since it was found that a genetic deficiency in factor XII does not lead to coagulation disturbances.

Both pathways depend on the presence of activated **thrombocytes**, on the surface of which several reactions take place. For example, the *prothrombinase complex* (left) forms when factors Xa and II, with the help of Va, bind via Ca^{2+} ions to anionic phospholipids in the thrombocyte membrane. For this to happen, factors II and X have to contain the non-proteinogenic amino acid **γ-carboxyglutamate** (Gla; see p. 62), which is formed in the liver by post-translational carboxylation of the factors. The Gla residues are found in groups in special domains that create contacts to the Ca^{2+} ions. Factors VII and IX are also linked to membrane phospholipids via Gla residues.

Substances that bind Ca^{2+} ions (e. g., *citrate*) prevent Gla-containing factors from attaching to the membrane and therefore inhibit coagulation. Antagonists of *vitamin K,* which is needed for synthesis of the Gla residues (see p. 364) also have anticoagulatory effects. These include dicumarol, for example.

Active **thrombin** not only converts fibrinogen into fibrin, but also indirectly promotes its own synthesis by catalyzing the activation of factors V and VIII. In addition, it catalyzes the activation of factor XIII and thereby triggers the cross-linking of the fibrin.

Regulation of blood clotting (not shown). To prevent the coagulation reaction from becoming excessive, the blood contains a number of anticoagulant substances, including highly effective proteinase inhibitors. For example, *antithrombin III* binds to various serine proteinases in the cascade and thereby inactivates them. *Heparin,* an anticoagulant glycosaminoglycan (see p. 346), potentiates the effect of antithrombin III. *Thrombomodulin,* which is located on the vascular endothelia, also inactivates thrombin. A glycoprotein known as *Protein C* ensures proteolytic degradation of factors V and VIII. As it is activated by thrombin, coagulation is shut down in this way.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Blood clotting

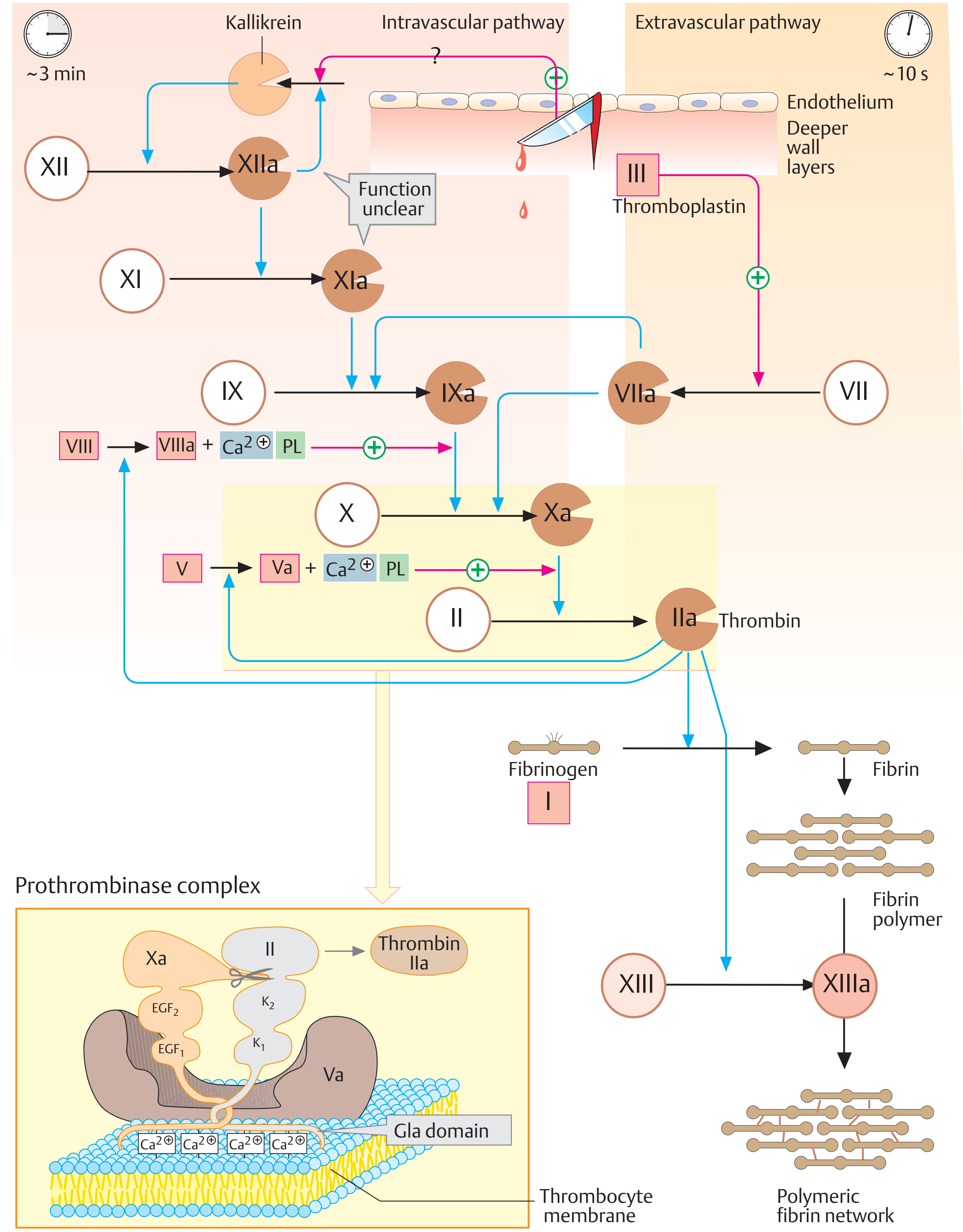
~3 min
Kallikrein
Intravascular pathway
?
Extravascular pathway
~10 s
Endothelium
Deeper wall layers
XII
XIIa
Function unclear
III
Thromboplastin
XI
XIa
IX
IXa
VIIa
VII
VIII
VIIIa + Ca2⊕ PL
X
Xa
V
Va + Ca2⊕ PL
II
IIa
Thrombin
Fibrinogen
I
Fibrin
Fibrin polymer
Prothrombinase complex
Xa
II
Thrombin IIa
EGF2
K2
EGF1
K1
Va
Gla domain
Ca2⊕
Ca2⊕
Ca2⊕
Ca2⊕
Thrombocyte membrane
XIII
XIIIa
Polymeric fibrin network

Coagulation factors

I Fibrinogen
◆ II Prothrombin* *3.4.21.5*
III Tissue factor/thromboplastin
IV $Ca^{2\oplus}$
V Proaccelerin
VI Synonym for Va
◆ VII Proconvertin* *3.4.21.21*
VIII Antihemophilic factor A

◆ IX Christmas factor* *3.4.21.22*
◆ X Stuart–Prower factor* *3.4.21.6*
XI Plasma thromboplastin antecedent* (PTA) *3.4.21.27*
XII Hageman factor* *3.4.21.38*
XIII Fibrin-stabilizing factor* *2.3.2.13*

* Proenzyme
◆ Contains γ-carboxyglutamate

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Fibrinolysis, blood groups

A. Fibrinolysis ◑

The fibrin thrombus resulting from blood clotting (see p. 290) is dissolved again by *plasmin,* a serine proteinase found in the blood plasma. For this purpose, the precursor *plasminogen* first has to be proteolytically activated by enzymes from various tissues. This group includes the *plasminogen activator* from the kidney (*urokinase*) and *tissue plasminogen activator* (t-PA) from vascular endothelia. By contrast, the plasma protein α_2-antiplasmin, which binds to active plasmin and thereby inactivates it, inhibits fibrinolysis.

Urokinase, t-PA, and streptokinase, a bacterial proteinase with similar activity, are used clinically to dissolve thrombi following *heart attacks.* All of these proteins are expressed recombinantly in bacteria (see p. 262).

B. Blood groups: the ABO system ◑

During blood transfusions, immune reactions can occur that destroy the erythrocytes transfused from the donor. These reactions result from the formation of antibodies (see p. 300) directed to certain surface structures on the erythrocytes. Known as **blood group antigens**, these are *proteins* or *oligosaccharides* that can differ from individual to individual. More than 20 different blood group systems are now known. The ABO system and the Rh system are of particular clinical importance.

In the **ABO system**, the carbohydrate parts of glycoproteins or glycolipids act as antigens. In this relatively simple system, there are four *blood groups* (A, B, AB, and 0). In individuals with blood groups A and B, the antigens consist of tetrasaccharides that only differ in their terminal sugar (galactose or *N*-acetylgalactosamine). Carriers of the AB blood group have both antigens (A *and* B). Blood group 0 arises from an oligosaccharide (the H antigen) that lacks the terminal residue of antigens A and B. The molecular causes for the differences between blood groups are mutations in the *glycosyl transferases* that transfer the terminal sugar to the core oligosaccharide.

Antibodies are only formed against antigens that the individual concerned does *not* possess. For example, carriers of blood group A form antibodies against antigen B (“anti-B”), while carriers of group B form antibodies against antigen A (“anti-A”). Individuals with blood group 0 form both types, and those with blood group AB do not form any of these antibodies.

If blood from blood group A is transfused into the circulation of an individual with blood group B, for example, then the anti-A present there binds to the A antigens. The donor erythrocytes marked in this way are recognized and destroyed by the complement system (see p. 298). In the test tube, *agglutination* of the erythrocytes can be observed when donor and recipient blood are incompatible.

The recipient’s serum should not contain any antibodies against the donor erythrocytes, and the donor serum should not contain any antibodies against the recipient’s erythrocytes. Donor blood from blood group 0 is unproblematic, as its erythrocytes do not possess any antibodies and therefore do not react with anti-A or anti-B in the recipient’s blood. Conversely, blood from the AB group can only be administered to recipients with the AB group, as these are the only ones without antibodies.

In the **Rh system** (not shown), proteins on the surface of the erythrocytes act as antigens. These are known as “rhesus factors,” as the system was first discovered in rhesus monkeys.

The rhesus D antigen occurs in 84% of all white individuals, who are therefore *“Rh-positive.”* If an Rh-positive child is born to an Rh-negative mother, fetal erythrocytes can enter the mother’s circulation during birth and lead to the formation of antibodies (IgG) against the D antigen. This initially has no acute effects on the mother or child. Complications only arise when there is a second pregnancy with an Rh-positive child, as maternal anti-D antibodies cross the placenta to the fetus even before birth and can trigger destruction of the child’s Rh-positive erythrocytes (*fetal erythroblastosis*).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Fibrinolysis

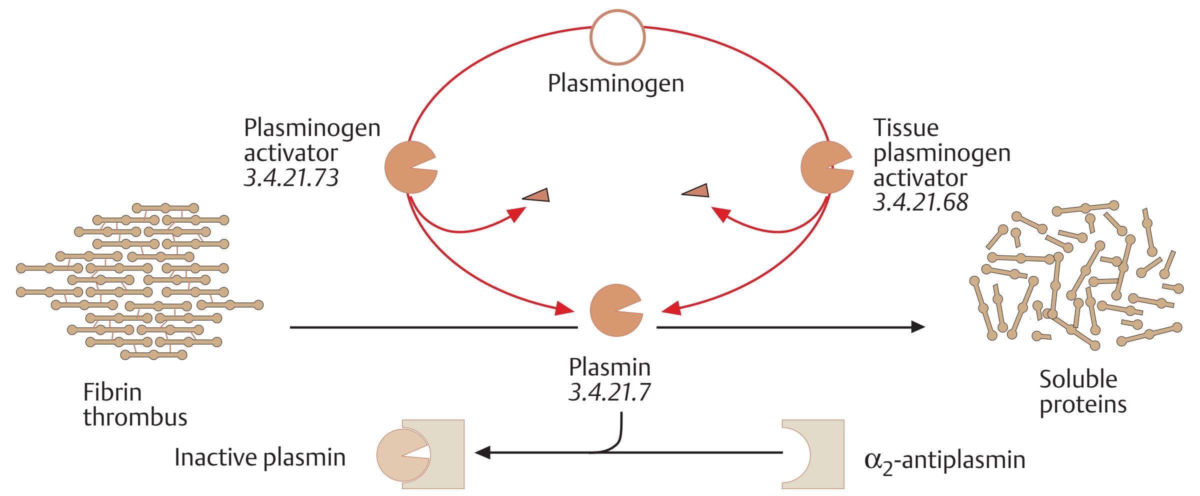
Plasminogen
Plasminogen activator 3.4.21.73
Tissue plasminogen activator 3.4.21.68
Fibrin thrombus
Plasmin 3.4.21.7
Soluble proteins
Inactive plasmin
α2-antiplasmin

B. Blood groups: the AB0 system

Blood group	A	B	AB	0
Genotypes	AA and A0	BB and B0	AB	00
Antigens	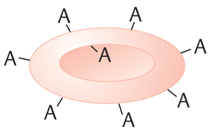	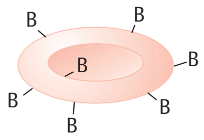	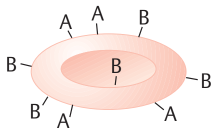	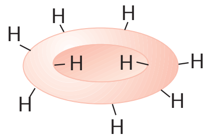
Antibodies in blood	anti-B	anti-A	—	anti-A anti-B
Frequency in central Europe	40%	16%	4%	40%

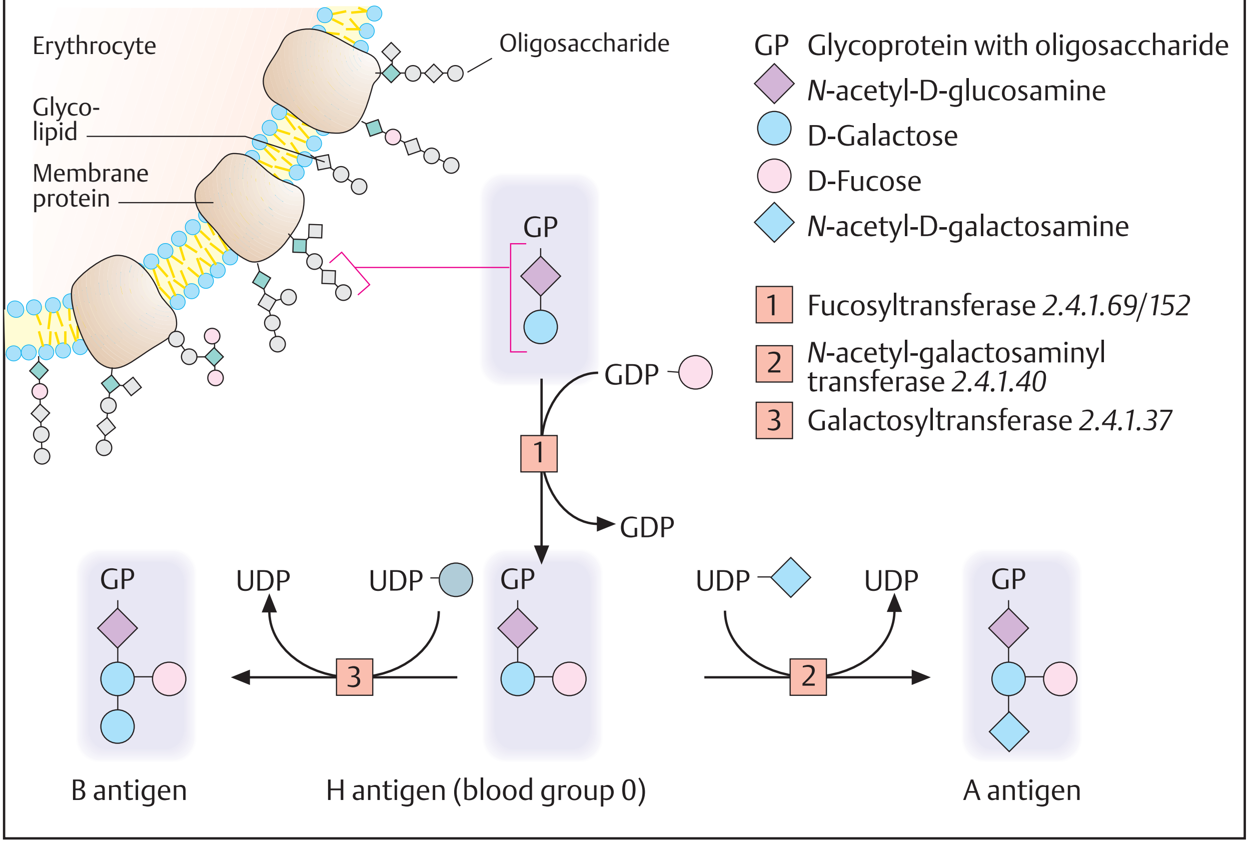
Erythrocyte
Glyco-lipid
Membrane protein
Oligosaccharide
GP
GP Glycoprotein with oligosaccharide
N-acetyl-D-glucosamine
D-Galactose
D-Fucose
N-acetyl-D-galactosamine
1 Fucosyltransferase 2.4.1.69/152
2 N-acetyl-galactosaminyl transferase 2.4.1.40
3 Galactosyltransferase 2.4.1.37
GDP
1
GDP
UDP
3
UDP
UDP
2
UDP
B antigen
H antigen (blood group 0)
A antigen

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Immune response

Viruses, bacteria, fungi, and parasites that enter the body of vertebrates of are recognized and attacked by the **immune system**. Endogenous cells that have undergone alterations—e. g., tumor cells—are also usually recognized as foreign and destroyed. The immune system is supported by physiological changes in infected tissue, known as **inflammation**. This reaction makes it easier for the immune cells to reach the site of infection.

Two different systems are involved in the immune response. The **innate immune system** is based on receptors that can distinguish between bacterial and viral surface structures or foreign proteins (known as *antigens*) and those that are endogenous. With the help of these receptors, *phagocytes* bind to the pathogens, absorb them by endocytosis, and break them down. The complement system (see p. 298) is also part of the innate system.

The **acquired** (adaptive) **immune system** is based on the ability of the *lymphocytes* to form highly specific antigen receptors "on suspicion," without ever having met the corresponding antigen. In humans, there are several billion different lymphocytes, each of which carries a different antigen receptor. If this type of receptor recognizes "its" cognate antigen, the lymphocyte carrying it is activated and then plays its special role in the immune response.

In addition, a distinction is made between cellular and humoral immune responses. The *T lymphocytes* (T cells) are responsible for **cellular immunity**. They are named after the thymus, in which the decisive steps in their differentiation take place. Depending on their function, another distinction is made between *cytotoxic T cells* (green) and *helper T cells* (blue). **Humoral immunity** is based on the activity of the *B lymphocytes* (B cells, light brown), which mature in the bone marrow. After activation by T cells, B cells are able to release soluble forms of their specific antigen receptors, known as *antibodies* (see p. 300), into the blood plasma. The immune system's "memory" is represented by memory cells. These are particularly long–lived cells that can arise from any of the lymphocyte types described.

A. Simplified diagram of the immune response ◑

Pathogens that have entered the body—e. g., viruses (top)—are taken up by **antigen-presenting cells** (APCs) and proteolytically degraded (**1**). The viral fragments produced in this way are then presented on the surfaces of these cells with the help of special membrane proteins (MHC proteins; see p. 296) (**2**). The APCs include B lymphocytes, macrophages, and dendritic cells such as the skin's Langerhans cells.

The complexes of MHC proteins and viral fragments displayed on the APCs are recognized by T cells that carry a receptor that matches the antigen ("T-cell receptors"; see p. 296) (**3**). Binding leads to activation of the T cell concerned and selective replication of it (**4**, *"clonal selection"*). The proliferation of immune cells is stimulated by *interleukins* (IL). These are a group of more than 20 signaling substances belonging to the cytokine family (see p. 392), with the help of which immune cells communicate with each other. For example, activated macrophages release IL-1 (**5**), while T cells stimulate their own replication and that of other immune cells by releasing IL-2 (**6**).

Depending on their type, activated T cells have different functions. **Cytotoxic T cells** (green) are able to recognize and bind virus-infected body cells or tumor cells (**7**). They then drive the infected cells into apoptosis (see p. 396) or kill them with *perforin*, a protein that perforates the target cell's plasma membrane (**8**).

B lymphocytes, which as APCs present viral fragments on their surfaces, are recognized by **helper T cells** (blue) or their T cell receptors (**9**). Stimulated by interleukins, selective clonal replication then takes place of B cells that carry antigen receptors matching those of the pathogen (**10**). These mature into **plasma cells** (**11**) and finally secrete large amounts of soluble **antibodies** (**12**).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Simplified scheme of the immune response

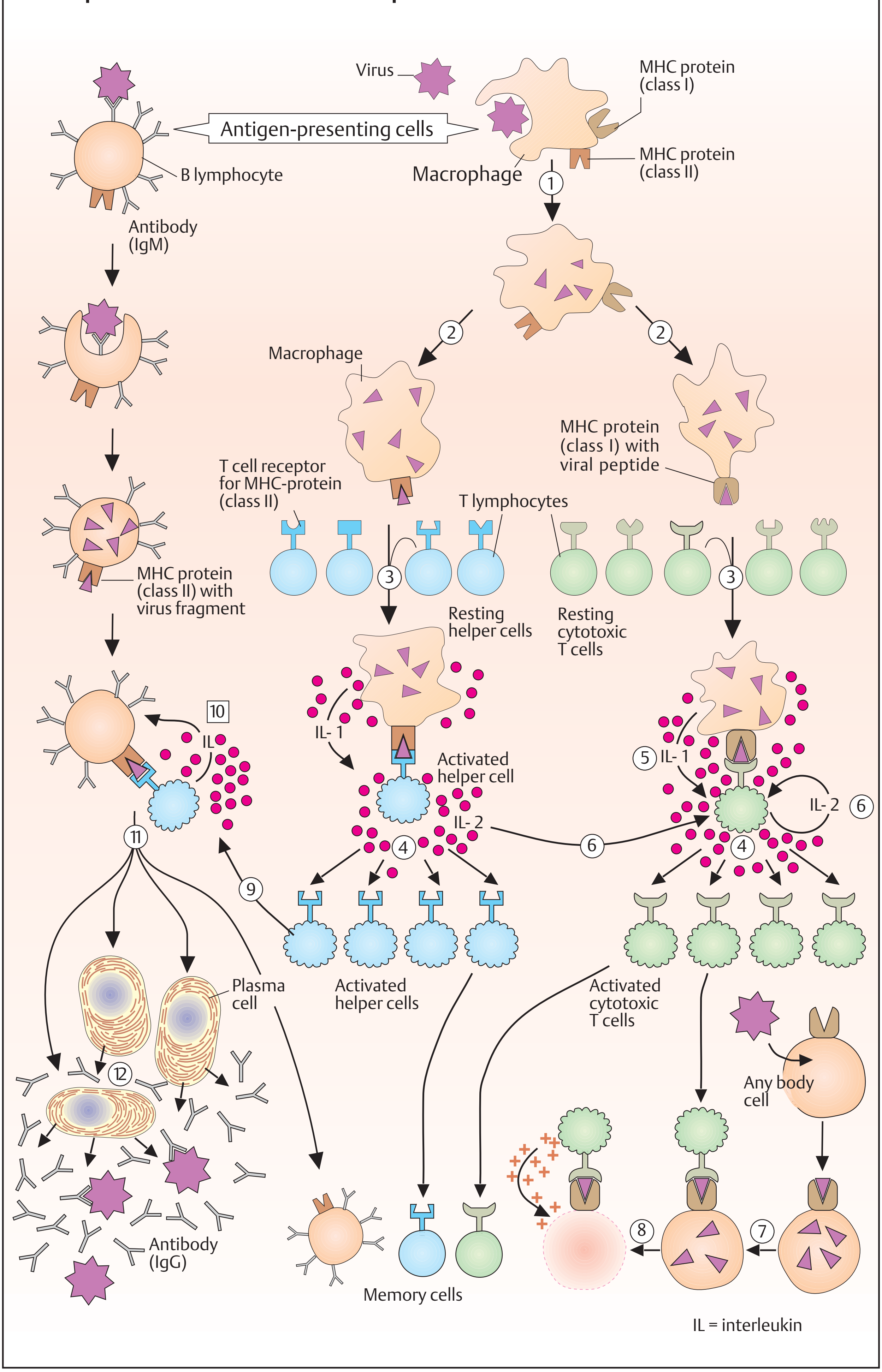
Virus
MHC protein
(class I)
Antigen-presenting cells
B lymphocyte
Macrophage
MHC protein
(class II)
①
Antibody
(IgM)
②
②
Macrophage
MHC protein
(class I) with
viral peptide
T cell receptor
for MHC-protein
(class II)
T lymphocytes
MHC protein
(class II) with
virus fragment
③
③
Resting
helper cells
Resting
cytotoxic
T cells
10
IL
IL-1
Activated
helper cell
⑤ IL-1
IL-2
IL-2 ⑥
④
⑥
④
⑪
⑨
Plasma
cell
Activated
helper cells
Activated
cytotoxic
T cells
Any body
cell
⑫
Antibody
(IgG)
Memory cells
⑧
⑦
IL = interleukin

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

T-cell activation

For the selectivity of the immune response (see p. 294), the cells involved must be able to recognize foreign antigens and proteins on other immune cells safely and reliably. To do this, they have antigen receptors on their cell surfaces and co-receptors that support recognition.

A. Antigen receptors ○

Many antigen receptors belong to the **immunoglobulin superfamily**. The common characteristic of these proteins is that they are made up from "immunoglobulin domains." These are characteristically folded substructures consisting of 70–110 amino acids, which are also found in soluble immunoglobulins (Ig; see p. 300). The illustration shows schematically a few of the important proteins in the Ig superfamily. They consist of constant regions (brown or green) and variable regions (orange). Homologous domains are shown in the same colors in each case. All of the receptors have transmembrane helices at the C terminus, which anchor them to the membranes. Intramolecular and intermolecular disulfide bonds are also usually found in proteins belonging to the Ig family.

Immunoglobulin M (IgM), a membrane protein on the surface of B lymphocytes, serves to bind free antigens to the B cells. By contrast, **T cell receptors** only bind antigens when they are presented by another cell as a complex with an MHC protein (see below). Interaction between MHC-bound antigens and T cell receptors is supported by **co-receptors**. This group includes **CD8**, a membrane protein that is typical in cytotoxic T cells. T helper cells use **CD4** as a co-receptor instead (not shown). The abbreviation "CD" stands for "cluster of differentiation." It is the term for a large group of proteins that are all located on the cell surface and can therefore be identified by antibodies. In addition to CD4 and CD8, there are many other co-receptors on immune cells (not shown).

The **MHC proteins** are named after the *"major histocompatibility complex"*—the DNA segment that codes for them. Human MHC proteins are also known as HLA antigens ("human leukocyte-associated" antigens). Their polymorphism is so large that it is unlikely that any two individuals carry the same set of MHC proteins—except for monozygotic twins.

Class I MHC proteins occur in almost all nucleated cells. They mainly interact with cytotoxic T cells and are the reason for the rejection of transplanted organs. Class I MHC proteins are heterodimers ($\alpha\beta$). The β subunit is also known as β_2-microglobulin.

Class II MHC proteins also consist of two peptide chains, which are related to each other. MHC II molecules are found on all antigen-presenting cells in the immune system. They serve for interaction between these cells and CD4-carrying T helper cells.

B. T-cell activation ◑

The illustration shows an interaction between a virus-infected body cell (bottom) and a CD8-carrying cytotoxic T lymphocyte (top). The infected cell breaks down viral proteins in its cytoplasm (**1**) and transports the peptide fragments into the endoplasmic reticulum with the help of a special transporter (*TAP*) (**2**). Newly synthesized class I MHC proteins on the endoplasmic reticulum are loaded with one of the peptides (**3**) and then transferred to the cell surface by vesicular transport (**4**). The viral peptides are bound on the surface of the α_2 domain of the MHC protein in a depression formed by an insertion as a "floor" and two helices as "walls" (see smaller illustration).

Supported by CD8 and other co-receptors, a T cell with a matching T cell receptor binds to the MHC peptide complex (**5**; cf. p. 224). This binding activates protein kinases in the interior of the T cell, which trigger a chain of additional reactions (*signal transduction;* see p. 388). Finally, destruction of the virus-infected cell by the cytotoxic T lymphocytes takes place.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Antigen receptors

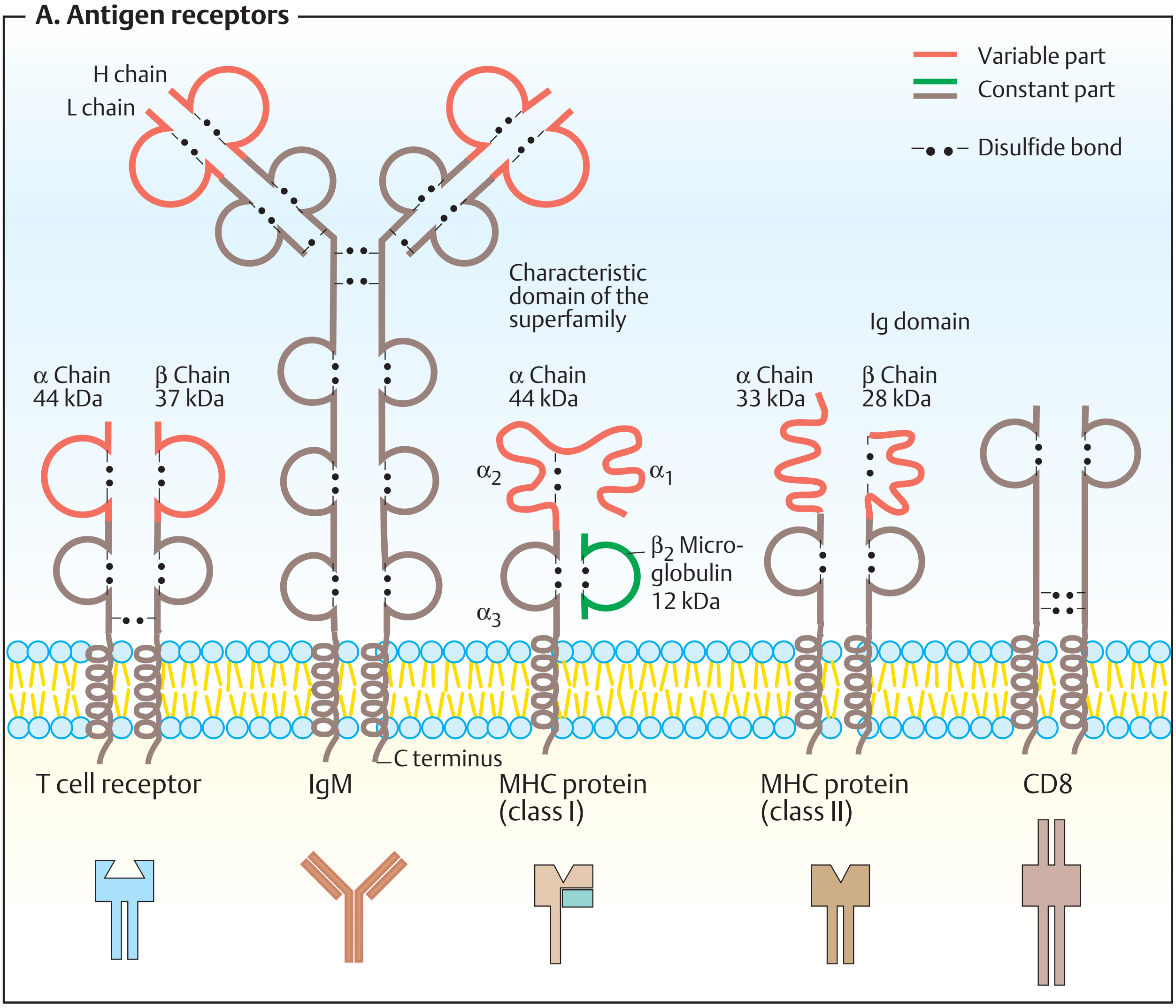
H chain
L chain
Variable part
Constant part
Disulfide bond
Characteristic domain of the superfamily
Ig domain
α Chain 44 kDa
β Chain 37 kDa
α Chain 44 kDa
α2
α1
β2 Micro-globulin 12 kDa
α3
α Chain 33 kDa
β Chain 28 kDa
C terminus
T cell receptor
IgM
MHC protein (class I)
MHC protein (class II)
CD8

B. T cell activation

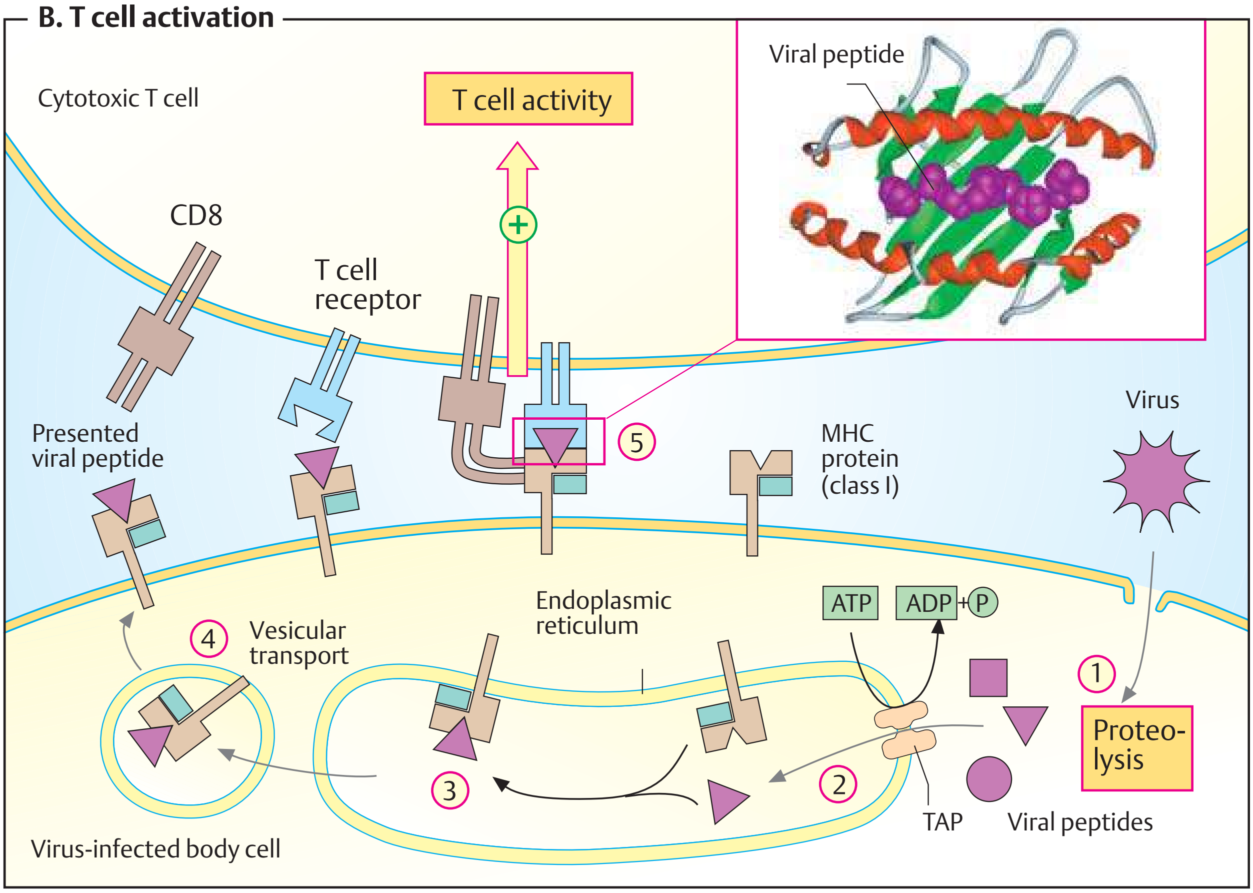
Cytotoxic T cell
T cell activity
Viral peptide
CD8
T cell receptor
+
Presented viral peptide
5
MHC protein (class I)
Virus
Endoplasmic reticulum
ATP
ADP
P
4
Vesicular transport
1
Proteo-lysis
3
2
TAP
Viral peptides
Virus-infected body cell

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Complement system

The complement system is part of the innate immune system (see p. 294). It supports *non-specific defense* against microorganisms. The system consists of some 30 different proteins, the *"complement factors,"* which are found in the blood and represent about 4% of all plasma proteins there. When inflammatory reactions occur, the complement factors enter the infected tissue and take effect there.

The complement system works in three different ways:

Chemotaxis. Various complement factors attract immune cells that can attack and phagocytose pathogens.

Opsonization. Certain complement factors ("opsonins") bind to the pathogens and thereby mark them as targets for phagocytosing cells (e. g., macrophages).

Membrane attack. Other complement factors are deposited in the bacterial membrane, where they create pores that lyse the pathogen (see below).

A. Complement activation ◑

The reactions that take place in the complement system can be initiated in several ways. During the early phase of infection, lipopolysaccharides and other structures on the surface of the pathogens trigger the *alternative pathway* (right). If antibodies against the pathogens become available later, the antigen–antibody complexes formed activate the *classic pathway* (left). Acute-phase proteins (see p. 276) are also able to start the complement cascade (*lectin pathway,* not shown).

Factors C1 to **C4** (for "complement") belong to the classic pathway, while **factors B** and **D** form the reactive components of the alternative pathway. Factors **C5** to **C9** are responsible for membrane attack. Other components not shown here regulate the system.

As in blood coagulation (see p. 290), the *early components* in the complement system are *serine proteinases,* which mutually activate each other through limited proteolysis. They create a self-reinforcing **enzyme cascade**. Factor **C3**, the products of which are involved in several functions, is central to the complement system.

The **classic pathway** is triggered by the formation of factor C1 at IgG or IgM on the surface of microorganisms (left). C1 is an 18-part molecular complex with three different components (C1q, C1r, and C1s). C1q is shaped like a bunch of tulips, the "flowers" of which bind to the F_c region of antibodies (left). This activates C1r, a *serine proteinase* that initiates the cascade of the classic pathway. First, C4 is proteolytically activated into C4b, which in turn cleaves C2 into C2a and C2b. C4B and C2a together form *C3 convertase* [1], which finally catalyzes the cleavage of C3 into C3a and C3b. Small amounts of C3b also arise from non-enzymatic hydrolysis of C3.

The **alternative pathway** starts with the binding of factors C3b and B to bacterial lipopolysaccharides (endotoxins). The formation of this complex allows cleavage of B by factor D, giving rise to a second form of *C3 convertase (C3bBb).*

Proteolytic cleavage of factor **C3** provides two components with different effects. The reaction exposes a highly *reactive thioester group* in C3b, which reacts with hydroxyl or amino groups. This allows C3b to bind covalently to molecules on the bacterial surface (*opsonization,* right). In addition, C3b initiates a chain of reactions leading to the formation of the *membrane attack complex* (see below). Together with C4a and C5a (see below), the smaller product C3a promotes the inflammatory reaction and has chemotactic effects.

The "late" factors C5 to C9 are responsible for the development of the **membrane attack complex** (bottom). They create an ion-permeable pore in the bacterial membrane, which leads to lysis of the pathogen. This reaction is triggered by *C5 convertase* [2]. Depending on the type of complement activation, this enzyme has the structure *C4b2a3b* or *C3bBb3b,* and it cleaves C5 into C5a and C5b. The complex of C5b and C6 allows deposition of C7 in the bacterial membrane. C8 and numerous C9 molecules—which form the actual pore—then bind to this core.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Complement activation

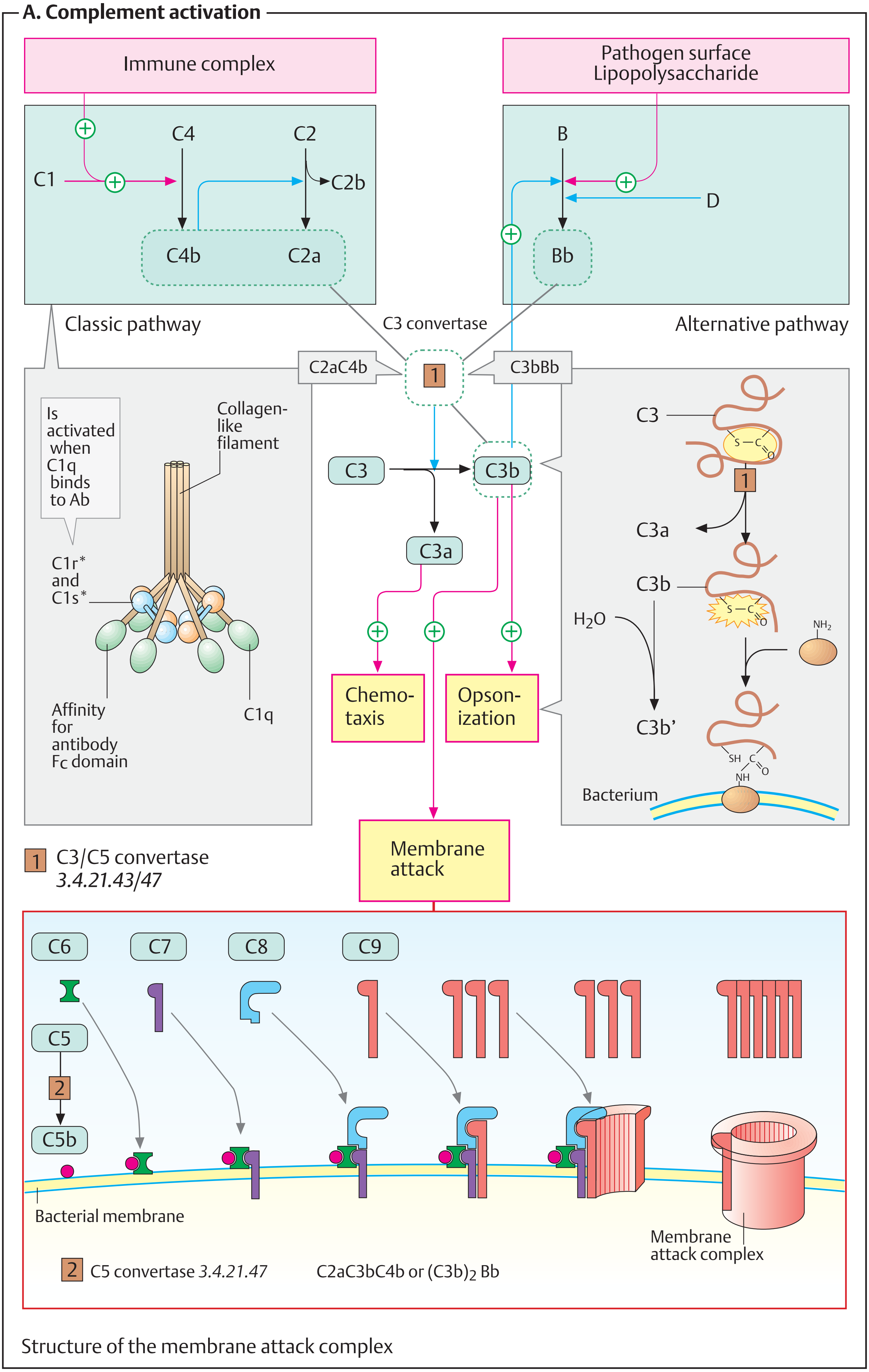
Immune complex
Pathogen surface
Lipopolysaccharide
C1
C4
C2
C2b
C4b
C2a
B
D
Bb
Classic pathway
C3 convertase
Alternative pathway
C2aC4b
1
C3bBb
Is activated when C1q binds to Ab
Collagen-like filament
C1r* and C1s*
Affinity for antibody Fc domain
C1q
C3
C3b
C3a
Chemo-taxis
Opson-ization
Membrane attack
C3
S—C=O
1
C3a
C3b
S—C=O
H_2O
NH_2
C3b'
SH
C=O
NH
Bacterium
1 C3/C5 convertase 3.4.21.43/47
C6
C7
C8
C9
C5
2
C5b
Bacterial membrane
Membrane attack complex
2 C5 convertase 3.4.21.47
C2aC3bC4b or $(C3b)_2$ Bb

Structure of the membrane attack complex

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Antibodies

Soluble antigen receptors, which are formed by activated B cells (plasma cells; see p. 294) and released into the blood, are known as **antibodies**. They are also members of the immunoglobulin family (Ig; see p. 296). Antibodies are an important part of the humoral immune defense system. They have no antimicrobial properties themselves, but support the cellular immune system in various ways:

1. They bind to antigens on the surface of pathogens and thereby prevent them from interacting with body cells (*neutralization;* see p. 404, for example).

2. They link single-celled pathogens into aggregates (immune complexes), which are more easily taken up by phagocytes (*agglutination*).

3. They activate the complement system (see p. 298) and thereby promote the innate immune defense system (*opsonization*).

In addition, antibodies have become indispensable aids in medical and biological diagnosis (see p. 304).

A. Domain structure of immunoglobulin G ◑

Type G immunoglobulins (**IgG**) are quantitatively the most important antibodies in the blood, where they form the fraction of γ-globulins (see p. 276). IgGs (mass 150 kDa) are tetramers with two **heavy chains** (H chains; red or orange) and two **light chains** (L chains; yellow). Both H chains are glycosylated (violet; see also p. 43).

The proteinase *papain* cleaves IgG into two F_{ab} fragments and one F_c fragment. The $\mathbf{F_{ab}}$ ("antigen-binding") fragments, which each consist of one L chain and the N-terminal part of an H chain, are able to bind antigens. The $\mathbf{F_c}$ ("crystallizable") fragment is made up of the C-terminal halves of the two H chains. This segment serves to bind IgG to cell surfaces, for interaction with the complement system and antibody transport.

Immunoglobulins are constructed in a modular fashion from several **immunoglobulin domains** (shown in the diagram on the right in Ω form). The H chains of IgG contain four of these domains ($\mathbf{V}_H$, $\mathbf{C}_H1$, $\mathbf{C}_H2$, and $\mathbf{C}_H3$) and the L chains contain two ($\mathbf{C}_L$ and $\mathbf{V}_L$). The letters C and V designate constant or variable regions.

Disulfide bonds link the two heavy chains to each other and also link the heavy chains to the light chains. Inside the domains, there are also disulfide bonds that stabilize the tertiary structure. The domains are approximately 110 amino acids (AA) long and are homologous with each other. The antibody structure evidently developed as a result of gene duplication. In its central region, known as the "hinge" region, the antibodies are highly mobile.

B. Classes of immunoglobulins ◑

Human immunoglobulins are divided into five classes. **IgA** (with two subgroups), **IgD, IgE, IgG** (with four subgroups), and **IgM** are defined by their H chains, which are designated by the Greek letters α, δ, ε, γ, and μ. By contrast, there are only two types of **L chain** (κ and λ). IgD and IgE (like IgG) are tetramers with the structure H_2L_2. By contrast, soluble IgA and IgM are multimers that are held together by disulfide bonds and additional **J peptides** (joining peptides).

The antibodies have different tasks. **IgMs** are the first immunoglobulins formed after contact with a foreign antigen. Their early forms are located on the surface of B cells (see p. 296), while the later forms are secreted from plasma cells as pentamers. Their action targets microorganisms in particular. Quantitatively, **IgGs** are the most important immunoglobulins (see the table showing serum concentrations). They occur in the blood and interstitial fluid. As they can pass the placenta with the help of receptors, they can be transferred from mother to fetus. **IgAs** mainly occur in the intestinal tract and in body secretions. **IgEs** are found in low concentrations in the blood. As they can trigger degranulation of mast cells (see p. 380), they play an important role in allergic reactions. The function of **IgDs** is still unexplained. Their plasma concentration is also very low.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Domain structure of immunoglobulin G

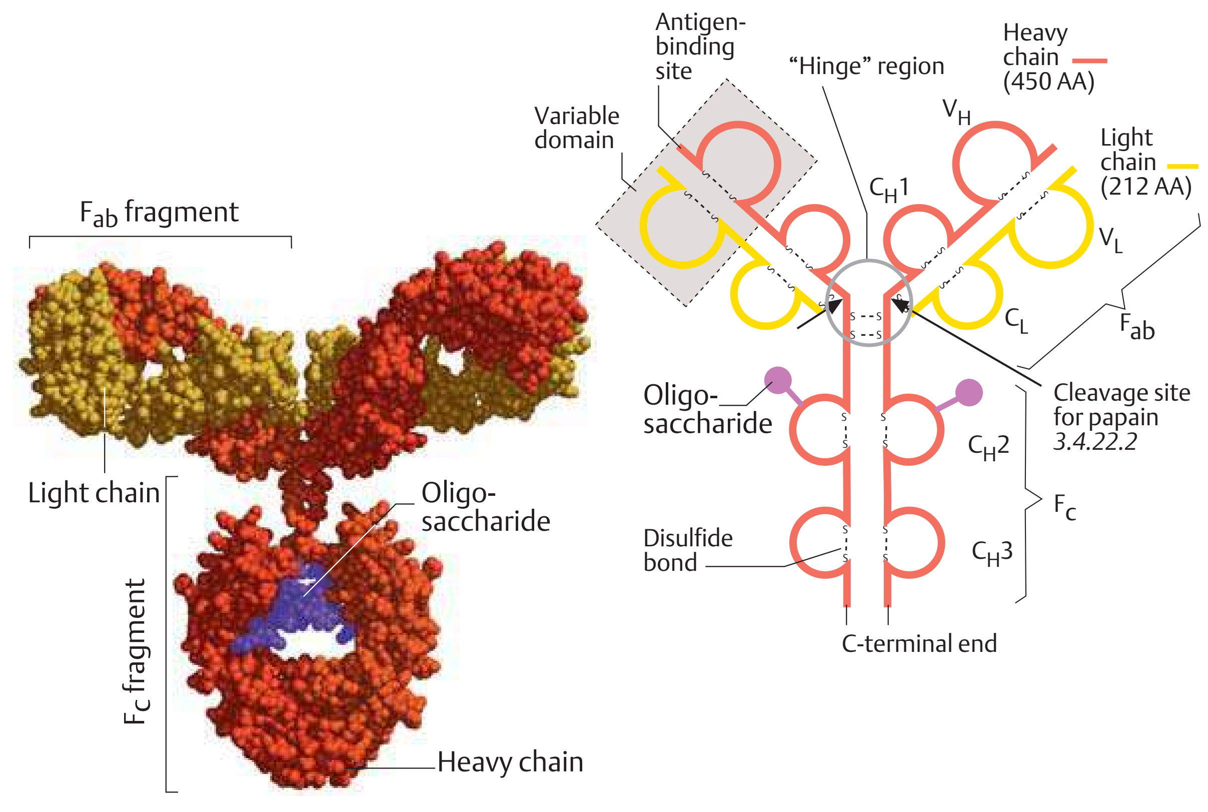
Fab fragment
Light chain
Fc fragment
Oligo-saccharide
Heavy chain
Antigen-binding site
Variable domain
"Hinge" region
Heavy chain (450 AA)
Light chain (212 AA)
VH
CH1
VL
CL
Fab
Cleavage site for papain 3.4.22.2
Oligo-saccharide
CH2
Fc
Disulfide bond
CH3
C-terminal end

B. Classes of immunoglobulins

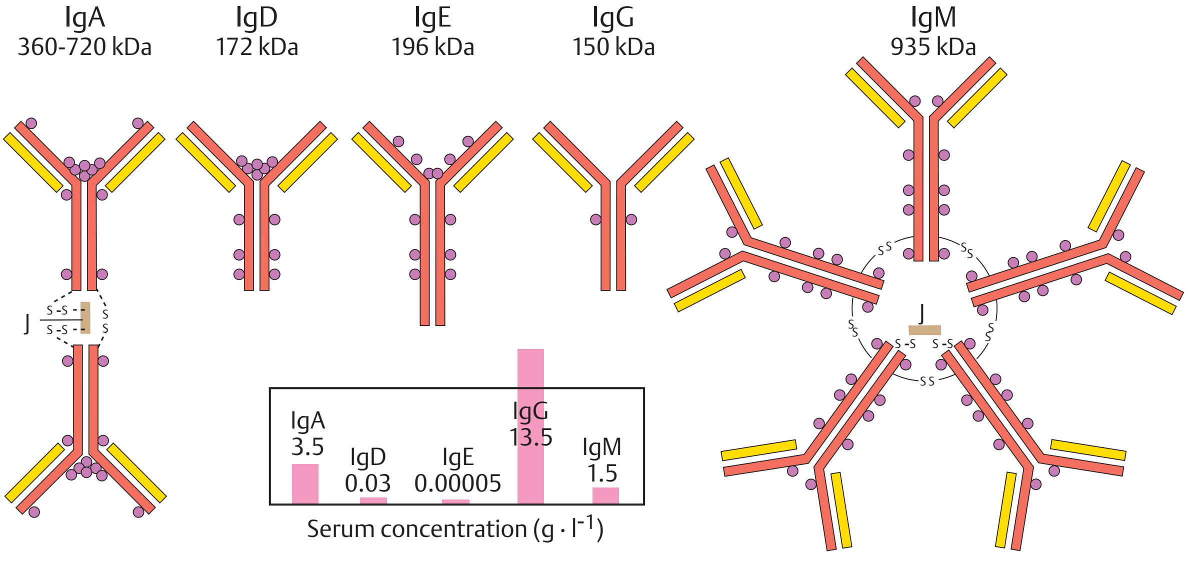
IgA
360-720 kDa
IgD
172 kDa
IgE
196 kDa
IgG
150 kDa
IgM
935 kDa
J
J
IgA 3.5
IgD 0.03
IgE 0.00005
IgG 13.5
IgM 1.5
Serum concentration (g · l-1)

Chain:

H	α	δ	ε	γ	μ
L	κ or λ	κ or λ	κ or λ	κ or λ	κ or λ

Structure:

$(\alpha_2\,\kappa_2)_n J$	$\delta_2\,\kappa_2$	$\varepsilon_2\,\kappa_2$	$\gamma_2\,\kappa_2$	$(\mu_2\,\kappa_2)_5 J$
$(\alpha_2\,\lambda_2)_n J$	$\delta_2\,\lambda_2$	$\varepsilon_2\,\lambda_2$	$\gamma_2\,\lambda_2$	$(\mu_2\,\lambda_2)_5 J$

n = 1, 2 or 3

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Antibody biosynthesis

The acquired (adaptive) immune system (see p. 294) is based on the ability of the lymphocytes to keep an extremely large repertoire of antigen receptors and soluble antibodies ready for use, so that even infections involving new types of pathogen can be combated. The wide range of immunoglobulins (Ig) are produced by genetic recombination and additional mutations during the development and maturation of the individual lymphocytes.

A. Variability of immunoglobulins ◑

It is estimated that more than 10^8 different antibody variants occur in every human being. This **variability** affects both the heavy and the light chains of immunoglobulins.

There are five different types of heavy (H) chain, according to which the antibody classes are defined (α, δ, ε, γ, μ), and two types of light (L) chain (κ and λ; see p. 300). The various Ig types that arise from combinations of these chains are known as **isotypes**. During immunoglobulin biosynthesis, plasma cells can switch from one isotype to another ("gene switch"). **Allotypic variation** is based on the existence of various alleles of the same gene—i. e., genetic differences between individuals. The term **idiotypic variation** refers to the fact that the antigen binding sites in the F_{ab} fragments can be highly variable. Idiotypic variation affects the *variable domains* (shown here in pink) of the light and heavy chains. At certain sites—known as the *hypervariable regions* (shown here in red)—variation is particularly wide; these sequences are directly involved in the binding of the antigen.

B. Causes of antibody variety ◑

There are three reasons for the extremely wide variability of antibodies:

1. **Multiple genes.** Various genes are available to code for the variable protein domains. Only one gene from among these is selected and expressed.

2. **Somatic recombination.** The genes are divided into several segments, of which there are various versions. Various ("untidy") combinations of the segments during lymphocyte maturation give rise to randomly combined new genes ("mosaic genes").

3. **Somatic mutation.** During differentiation of B cells into plasma cells, the coding genes mutate. In this way, the "primordial" *germ-line genes* can become different *somatic genes* in the individual B cell clones.

C. Biosynthesis of a light chain ○

We can look at the basic features of the genetic organization and synthesis of immunoglobulins using the biosynthesis of a mouse κ chain as an example. The gene segments for this light chain are designated L, V, J, and C. They are located on chromosome 6 in the **germ-line DNA** (on chromosome 2 in humans) and are separated from one another by introns (see p. 242) of different lengths.

Some 150 identical **L segments** code for the signal peptide ("leader sequence," 17–20 amino acids) for secretion of the product (see p. 230). The **V segments**, of which there are 150 different variants, code for most of the variable domains (95 of the 108 amino acids). L and V segments always occur in pairs—in tandem, so to speak. By contrast, there are only five variants of the **J segments** (joining segments) at most. These code for a peptide with 13 amino acids that links the variable part of the κ chains to the constant part. A single **C segment** codes for the constant part of the light chain (84 amino acids).

During the differentiation of B lymphocytes, individual *V/J combinations* arise in each B cell. One of the 150 L/V tandem segments is selected and linked to one of the five J segments. This gives rise to a *somatic gene* that is much smaller than the germline gene. Transcription of this gene leads to the formation of the **hnRNA** for the κ chain, from which introns and surplus J segments are removed by splicing (see p. 246). Finally, the completed **mRNA** still contains one each of the L–V–J–C segments and after being transported into the cytoplasm is available for translation. The subsequent steps in Ig biosynthesis follow the rules for the synthesis of membrane-bound or secretory proteins (see p. 230).

All rights reserved. Usage subject to terms and conditions of license.

A. Variability of immunoglobulins

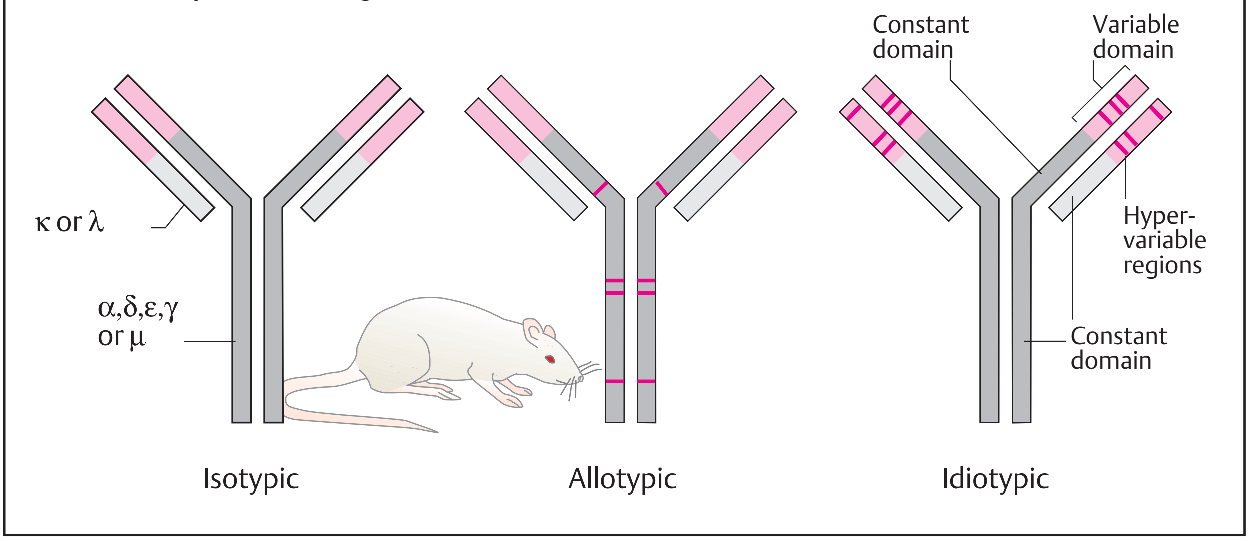
κ or λ
α,δ,ε,γ or μ
Constant domain
Variable domain
Hyper-variable regions
Constant domain
Isotypic
Allotypic
Idiotypic

B. Origins of antibody variety

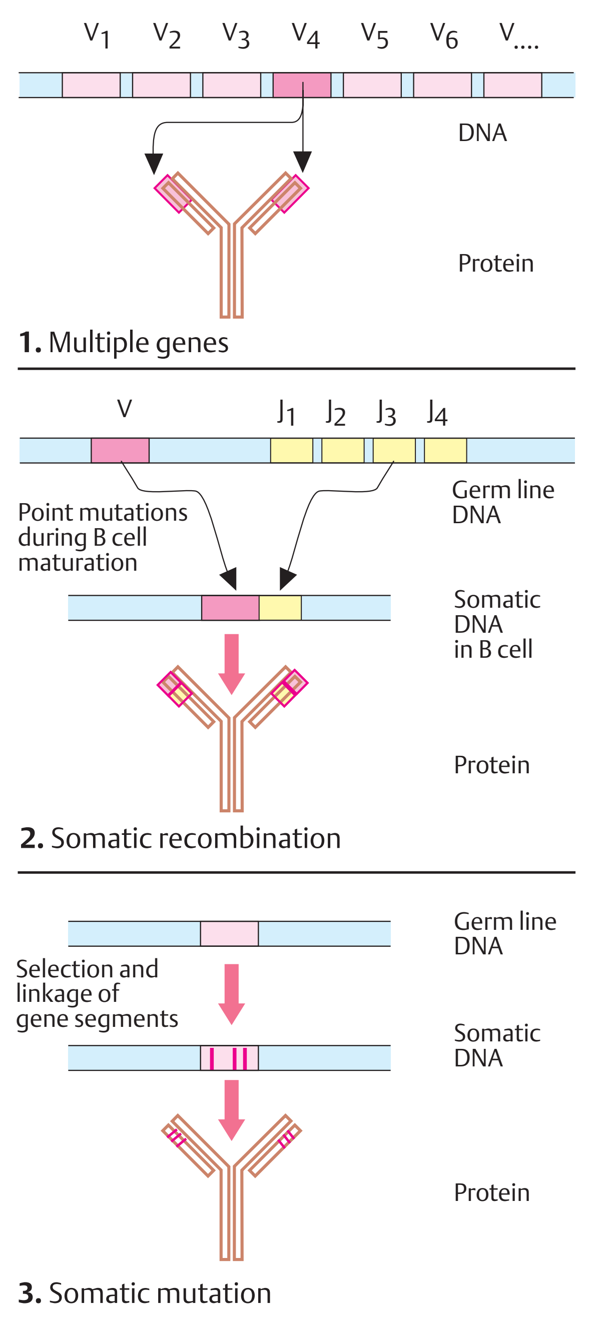
V_1 V_2 V_3 V_4 V_5 V_6 V....
DNA
Protein
1. Multiple genes
V
J_1 J_2 J_3 J_4
Germ line DNA
Point mutations during B cell maturation
Somatic DNA in B cell
Protein
2. Somatic recombination
Germ line DNA
Selection and linkage of gene segments
Somatic DNA
Protein
3. Somatic mutation

C. Biosynthesis of a light chain

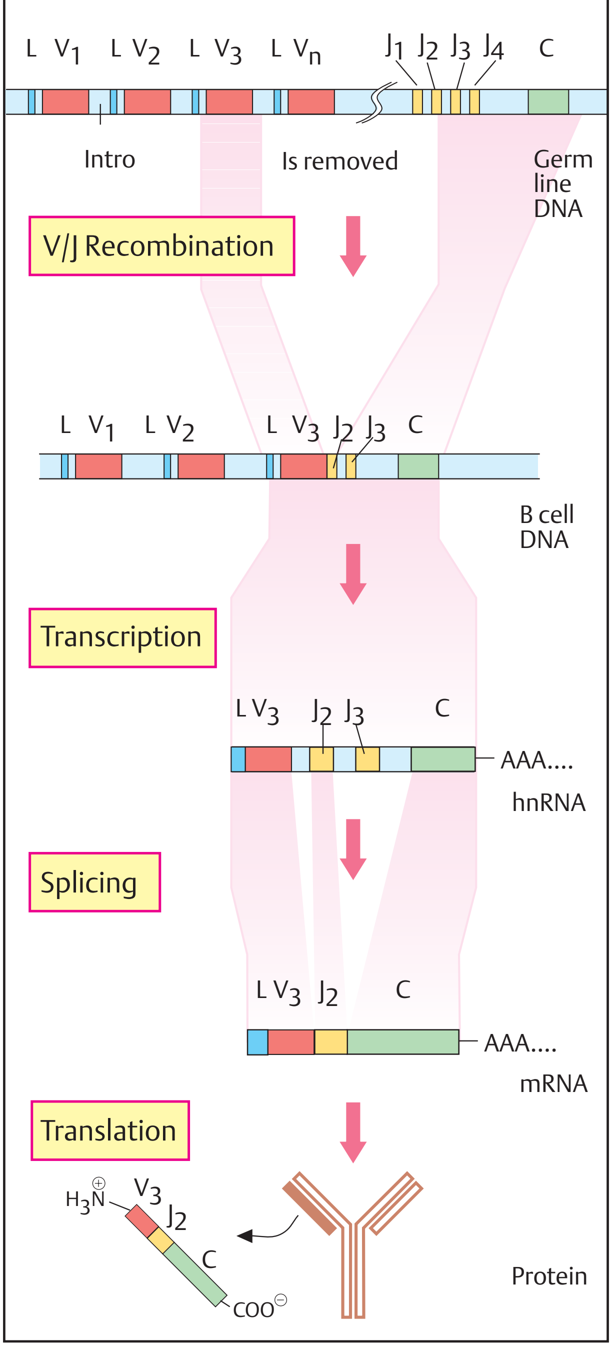
L V_1 L V_2 L V_3 L V_n
J_1 J_2 J_3 J_4 C
Intro
Is removed
Germ line DNA
V/J Recombination
L V_1 L V_2 L V_3 J_2 J_3 C
B cell DNA
Transcription
L V_3 J_2 J_3 C
AAA....
hnRNA
Splicing
L V_3 J_2 C
AAA....
mRNA
Translation
$H_3\overset{\oplus}{N}$
V_3
J_2
C
$COO^{\ominus}$
Protein

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Monoclonal antibodies, immunoassay

A. Monoclonal antibodies ○

Monoclonal antibodies (MABs) are secreted by immune cells that derive from a single antibody-forming cell (from a single cell *clone*). This is why each MAB is directed against only one specific *epitope* of an immunogenic substance, known as an *"antigenic determinant."* Large molecules contain several epitopes, against which various antibodies are formed by various B cells. An antiserum containing a mixture of all of these antibodies is described as being *polyclonal.*

To obtain MABs, **lymphocytes** isolated from the spleen of immunized mice (**1**) are fused with mouse tumor cells (**myeloma cells**, **2**). This is necessary because antibody-secreting lymphocytes in culture have a lifespan of only a few weeks. Fusion of lymphocytes with tumor cells gives rise to cell hybrids, known as **hybridomas,** which are potentially immortal.

Successful *fusion* (**2**) is a rare event, but the frequency can be improved by adding polyethylene glycol (PEG). To obtain only successfully fused cells, incubation is required for an extended period in a **primary culture** with *HAT medium* (**3**), which contains hypoxanthine, aminopterin, and thymidine. *Aminopterin*, an analogue of dihydrofolic acid, competitively inhibits *dihydrofolate reductase* and thus inhibits the synthesis of dTMP (see p. 402). As dTMP is essential for DNA synthesis, myeloma cells cannot survive in the presence of aminopterin. Although spleen cells are able to circumvent the inhibitory effect of aminopterin by using *hypoxanthine* and *thymidine,* they have a limited lifespan and die. Only hybridomas survive culture in HAT medium, because they possess both the immortality of the myeloma cells and the spleen cells' metabolic side pathway.

Only a few fused cells actually produce antibodies. To identify these cells, the hybridomas have to be isolated and replicated by **cloning** (**4**). After the clones have been tested for antibody formation, positive cultures are picked out and selected by further cloning (**5**). This results in hybridomas that synthesize *monoclonal antibodies.* Finally, MAB production is carried out in vitro using a bioreactor, or in vivo by producing ascites fluid in mice (**6**).

B. Immunoassay ○

Immunoassays are semiquantitative procedures for assessing substances with low concentrations. In principle, immunoassays can be used to assess any compound against which antibodies are formed.

The basis for this procedure is the *antigen–antibody "reaction"*—i. e., specific binding of an antibody to the molecule being assayed. Among the many different immunoassay techniques that have been developed—e. g., *radioimmunoassay* (RIA), and *chemoluminescence immunoassay* (CIA)—a version of the **enzyme-linked immunoassay** (EIA) is shown here.

The substance to be assayed—e. g., the hormone thyroxine in a serum sample—is pipetted into a microtiter plate (**1**), the walls of which are coated with *antibodies* that specifically bind the hormone. At the same time, a small amount of thyroxine is added to the incubation to which an enzyme known as the *"tracer"* (**1**) has been chemically coupled. The tracer and the hormone being assayed compete for the small number of antibody binding sites available. After binding has taken place (**2**), all of the unbound molecules are rinsed out. The addition of a substrate solution for the enzyme (a *chromogenic solution*) then triggers an indicator reaction (**3**), the products of which can be assessed using photometry (**4**).

The larger the amount of enzyme that can bind to the antibodies on the container's walls, the larger the amount of dye that is produced. Conversely, the larger the amount of the substance being assayed that is present in the sample, the *smaller* the amount of tracer that can be bound by the antibodies. Quantitative analysis can be carried out through parallel measurement using standards with a known concentration.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Monoclonal antibodies

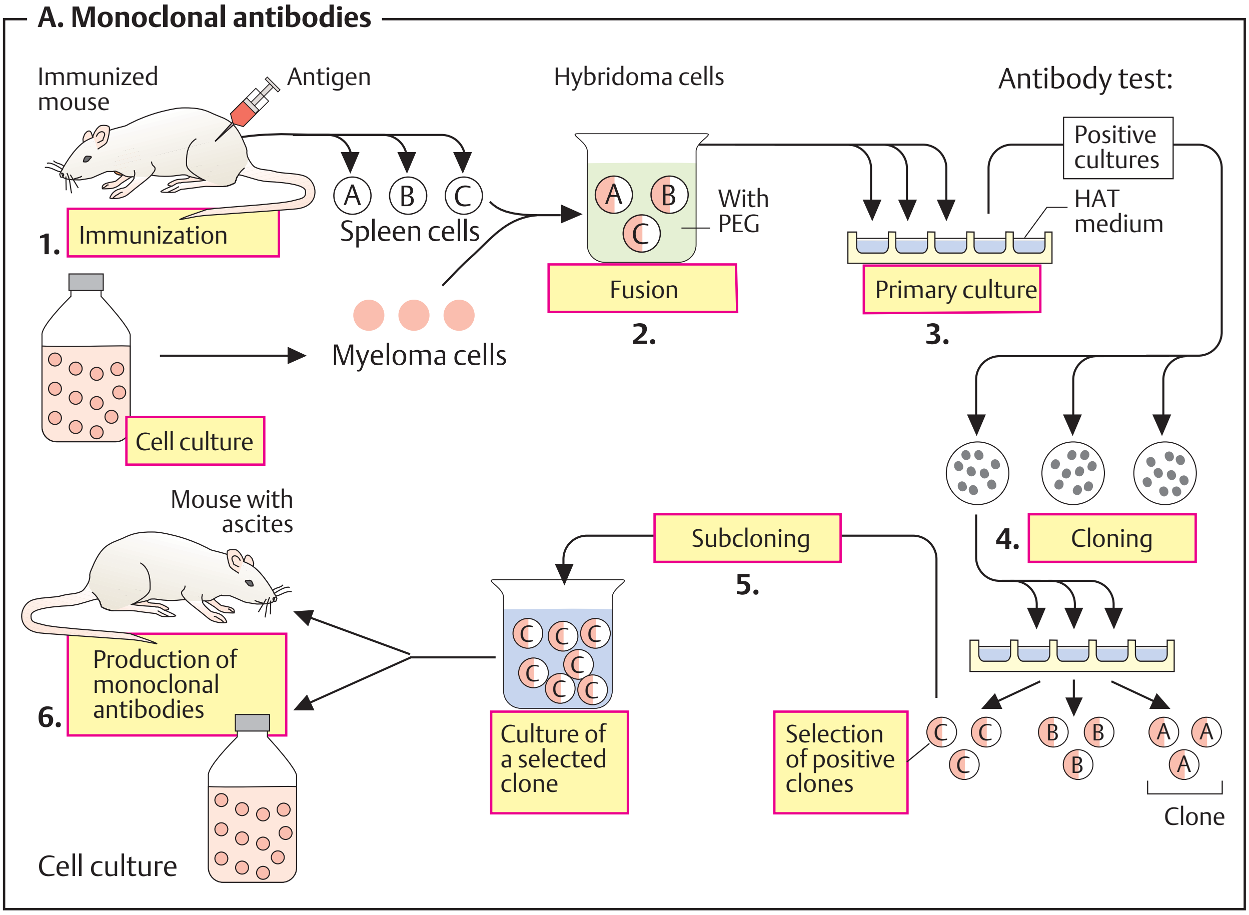
Immunized mouse
Antigen
Hybridoma cells
Antibody test:
Positive cultures
HAT medium
A B C
Spleen cells
A B C
With PEG
1. Immunization
Fusion
2.
Primary culture
3.
Myeloma cells
Cell culture
Mouse with ascites
Subcloning
5.
4. Cloning
6. Production of monoclonal antibodies
Culture of a selected clone
Selection of positive clones
C C C
B B B
A A A
Clone
Cell culture

B. Immunoassay

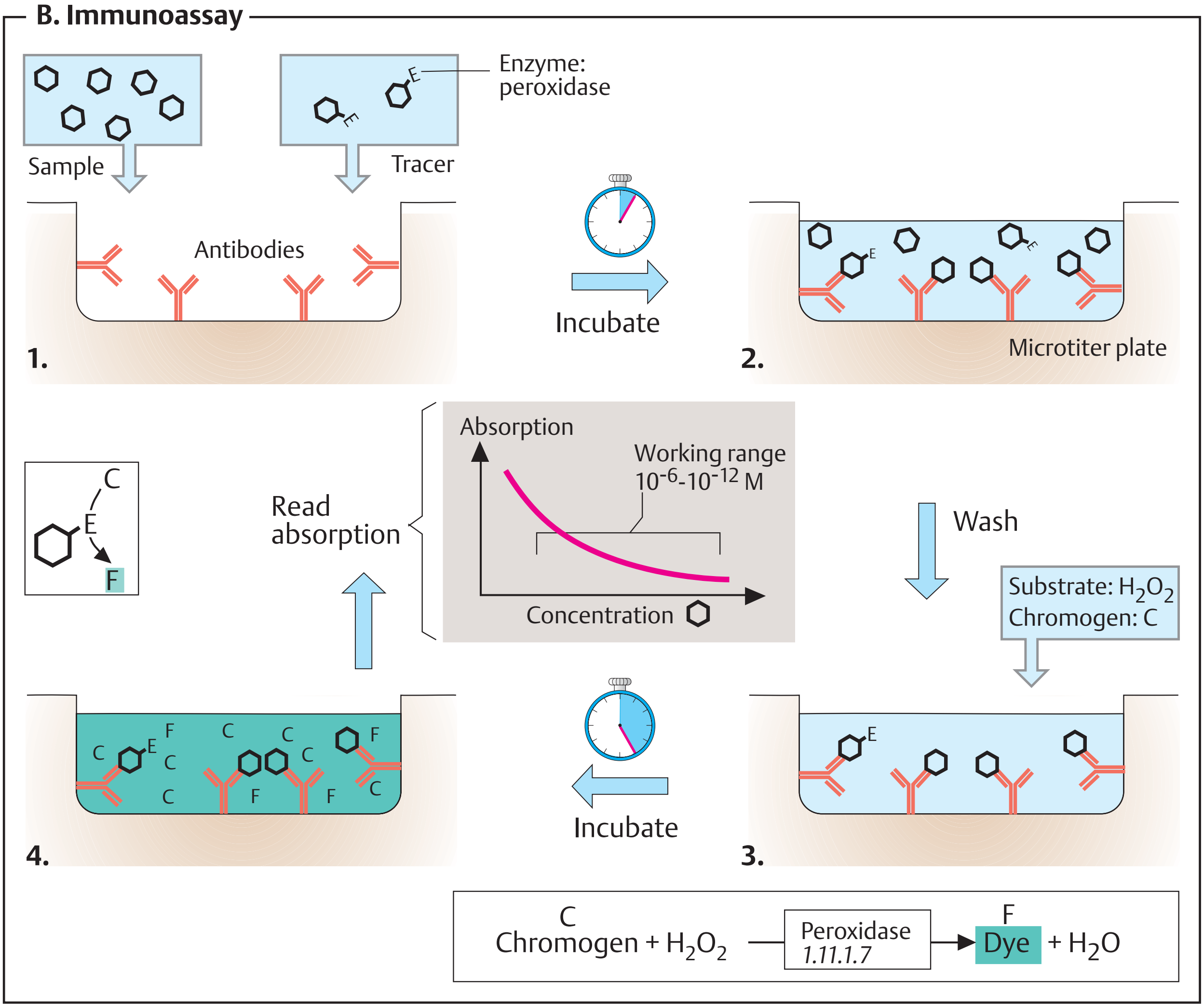
Sample
Tracer
E
Enzyme: peroxidase
Antibodies
1.
Incubate
2.
Microtiter plate
C
E
F
Read absorption
Absorption
Working range 10^{-6}-10^{-12} M
Concentration
Wash
Substrate: H_2O_2
Chromogen: C
4.
Incubate
3.
C
Chromogen + H_2O_2
Peroxidase 1.11.1.7
F
Dye
+ H_2O

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Liver: functions

Weighing 1.5 kg, the liver is one of the largest organs in the human body. Although it only represents 2–3% of the body's mass, it accounts for 25–30% of oxygen consumption.

A. Diagram of a hepatocyte ◑

The 3 10^{11} cells in the liver—particularly the **hepatocytes**, which make up 90% of the cell mass—are the central location for the body's *intermediary metabolism.* They are in close contact with the blood, which enters the liver from the portal vein and the hepatic arteries, flows through capillary vessels known as sinusoids, and is collected again in the central veins of the hepatic lobes. Hepatocytes are particularly rich in endoplasmic reticulum, as they carry out intensive protein and lipid synthesis. The cytoplasm contains granules of insoluble glycogen. Between the hepatocytes, there are bile capillaries through which bile components are excreted.

B. Functions of the liver ●

The most important functions of the liver are:

1. **Uptake** of nutrients supplied by the intestines via the portal vein.
2. Biosynthesis of endogenous compounds and storage, conversion, and degradation of them into excretable molecules (**metabolism**). In particular, the liver is responsible for the biosynthesis and degradation of almost all plasma proteins.
3. **Supply** of the body with metabolites and nutrients.
4. **Detoxification** of toxic compounds by biotransformation.
5. **Excretion** of substances with the bile.

C. Hepatic metabolism ●

The liver is involved in the metabolism of practically all groups of metabolites. Its functions primarily serve to cushion fluctuations in the concentration of these substances in the blood, in order to ensure a constant supply to the peripheral tissues (*homeostasis*).

Carbohydrate metabolism. The liver takes up glucose and other monosaccharides from the plasma. Glucose is then either stored in the form of the polysaccharide *glycogen* or converted into fatty acids. When there is a drop in the blood glucose level, the liver releases glucose again by breaking down glycogen. If the glycogen store is exhausted, glucose can also be synthesized by *gluconeogenesis* from lactate, glycerol, or the carbon skeleton of amino acids (see p. 310).

Lipid metabolism. The liver synthesizes fatty acids from acetate units. The fatty acids formed are then used to synthesize fats and phospholipids, which are released into the blood in the form of *lipoproteins.* The liver's special ability to convert fatty acids into *ketone bodies* and to release these again is also important (see p. 312).

Like other organs, the liver also synthesizes cholesterol, which is transported to other tissues as a component of lipoproteins. Excess cholesterol is converted into bile acids in the liver or directly excreted with the bile (see p. 314).

Amino acid and protein metabolism. The liver controls the plasma levels of the amino acids. Excess amino acids are broken down. With the help of the urea cycle (see p. 182), the nitrogen from the amino acids is converted into urea and excreted via the kidneys. The carbon skeleton of the amino acids enters the intermediary metabolism and serves for glucose synthesis or energy production. In addition, most of the plasma proteins are synthesized or broken down in the liver (see p. 276).

Biotransformation. Steroid hormones and bilirubin, as well as drugs, ethanol, and other xenobiotics are taken up by the liver and inactivated and converted into highly polar metabolites by conversion reactions (see p. 316).

Storage. The liver not only stores energy reserves and nutrients for the body, but also certain mineral substances, trace elements, and vitamins, including iron, retinol, and vitamins A, D, K, folic acid, and B_{12}.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Diagram of a hepatocyte

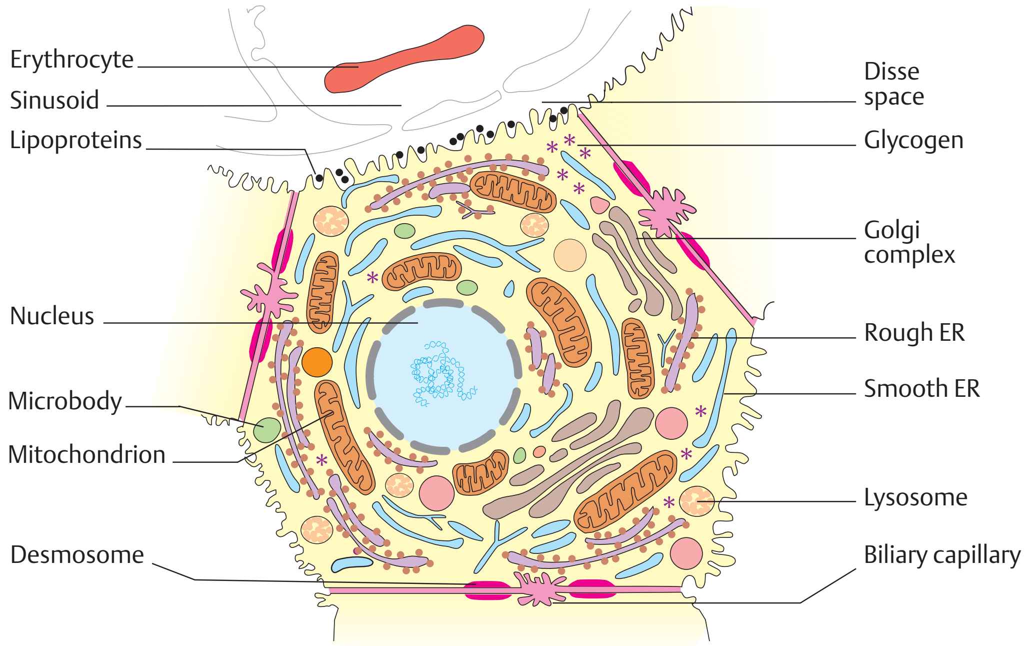
Erythrocyte
Sinusoid
Lipoproteins
Nucleus
Microbody
Mitochondrion
Desmosome
Disse space
Glycogen
Golgi complex
Rough ER
Smooth ER
Lysosome
Biliary capillary

B. Functions of the liver

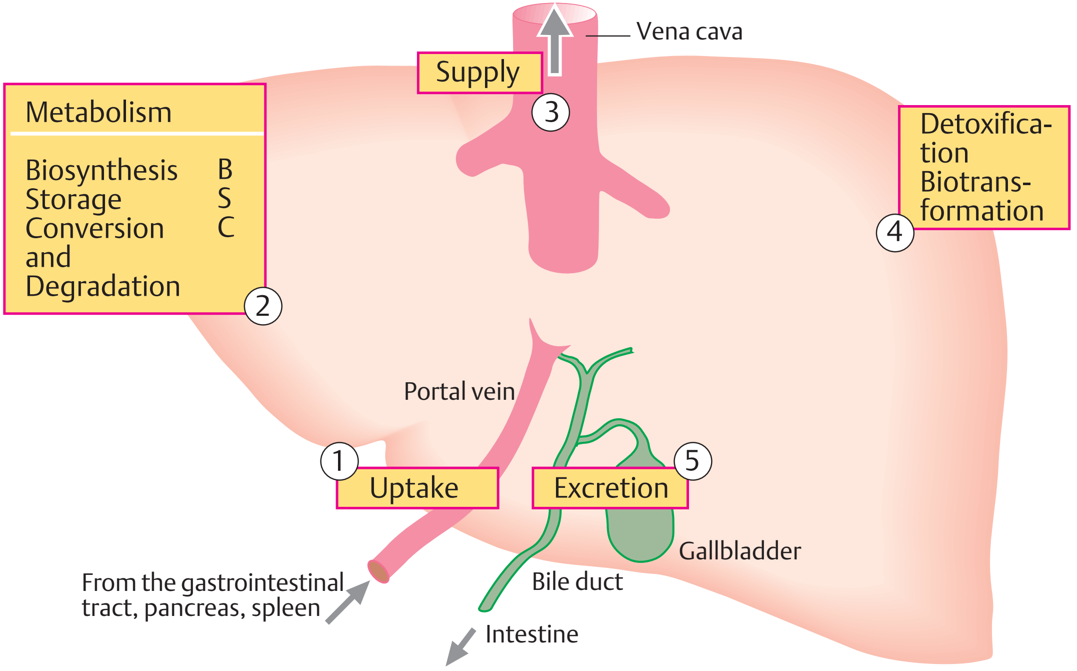
Vena cava
Supply
③
Metabolism
Biosynthesis B
Storage S
Conversion and Degradation C
②
Detoxification
Biotransformation
④
Portal vein
①
Uptake
Excretion
⑤
Gallbladder
Bile duct
From the gastrointestinal tract, pancreas, spleen
Intestine

C. Liver metabolism

Carbohydrate metabolism	
Glucose	BSC
Galactose	C
Fructose	C
Mannose	C
Pentoses	BC
Lactate	C
Glycerol	BC
Glycogen	BSC

Lipid metabolism	
Fatty acids	BC
Fats	BC
Ketone bodies	B
Cholesterol	BEC
Bile acids	BE
Vitamins	SC

Amino acid metabolism	
Amino acids	C
Urea	B

Plasma proteins	
Lipoproteins	BC
Albumin	BC
Coagulation factors	BC
Hormones	C
Enzymes	BC

Biotransformation	
Steroid hormones	EC
Bile pigments	EC
Ethanol	C
Drugs	EC

B	Biosynthesis
C	Conversion and degradation
E	Excretion
S	Storage

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Buffer function in organ metabolism

All of the body's tissues have a constant requirement for energy substrates and nutrients. The body receives these metabolites with food, but the supply is irregular and in varying amounts. The liver acts here along with other organs, particularly adipose tissue, as a balancing *buffer* and *storage organ.*

In the metabolism, a distinction is made between the *absorptive state (well-fed state)* immediately after a meal and the *postabsorbtive state (state of starvation)*, which starts later and can merge into hunger. The switching of the organ metabolism between the two phases depends on the concentration of energy-bearing metabolites in the blood (plasma level). This is regulated jointly by hormones and by the autonomic nervous system.

A. Absorptive state ◑

The absorptive state continues for 2–4 hours after food intake. As a result of food digestion, the plasma levels of glucose, amino acids, and fats (triacylglycerols) temporarily increase.

The endocrine **pancreas** responds to this by altering its hormone release—there is an increase in *insulin* secretion and a reduction in *glucagon* secretion. The increase in the insulin/glucagon quotient and the availability of substrates trigger an *anabolic phase* in the tissues—particularly liver, muscle, and adipose tissues.

The **liver** forms increased amounts of glycogen and fats from the substrates supplied. Glycogen is stored, and the fat is released into the blood in very low density lipoproteins (VLDLs).

Muscle also refills its glycogen store and synthesizes proteins from the amino acids supplied.

Adipose tissue removes free fatty acids from the lipoproteins, synthesizes triacylglycerols from them again, and stores these in the form of insoluble droplets.

During the absorptive state, the **heart** and **neural tissue** mainly use glucose as an energy source, but they are unable to establish any substantial energy stores. Heart muscle cells are in a sense "omnivorous," as they can also use other substances to produce energy (fatty acids, ketone bodies). By contrast, the central nervous system (CNS) is dependent on glucose. It is only able to utilize ketone bodies after a prolonged phase of hunger (**B**).

B. Postabsorptive state ◑

When the food supply is interrupted, the postabsorbtive state quickly sets in. The pancreatic A cells now release increased amounts of *glucagon,* while the B cells reduce the amount of *insulin* they secrete. The reduced insulin/glucagon quotient leads to switching of the intermediary metabolism. The body now falls back on its energy reserves. To do this, it breaks down *storage substances* (glycogen, fats, and proteins) and shifts energy-supplying metabolites between the organs.

The **liver** first empties its glycogen store (*glycogenolysis;* see p. 156). It does not use the released glucose itself, however, but supplies the other tissues with it. In particular, the brain, adrenal gland medulla, and erythrocytes depend on a constant supply of glucose, as they have no substantial glucose reserves themselves. When the liver's glycogen reserves are exhausted after 12–24 hours, *gluconeogenesis* begins (see p. 154). The precursors for this are derived from the musculature (amino acids) and adipose tissue (glycerol from fat degradation). From the fatty acids that are released (see below), the liver starts to form ketone bodies (*ketogenesis;* see p. 312). These are released into the blood and serve as important energy suppliers during the hunger phase. After 1–2 weeks, the CNS also starts to use ketone bodies to supply part of its energy requirements, in order to save glucose.

In **muscle**, the extensive glycogen reserves are exclusively used for the muscles' own requirements (see p. 320). The slowly initiated protein breakdown in muscle supplies amino acids for gluconeogenesis in the liver.

In **adipose tissue**, glucagon triggers *lipolysis,* releasing fatty acids and glycerol. The fatty acids are used as energy suppliers by many types of tissue (with the exception of brain and erythrocytes). An important recipient of the fatty acids is the liver, which uses them for ketogenesis.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Absorptive state

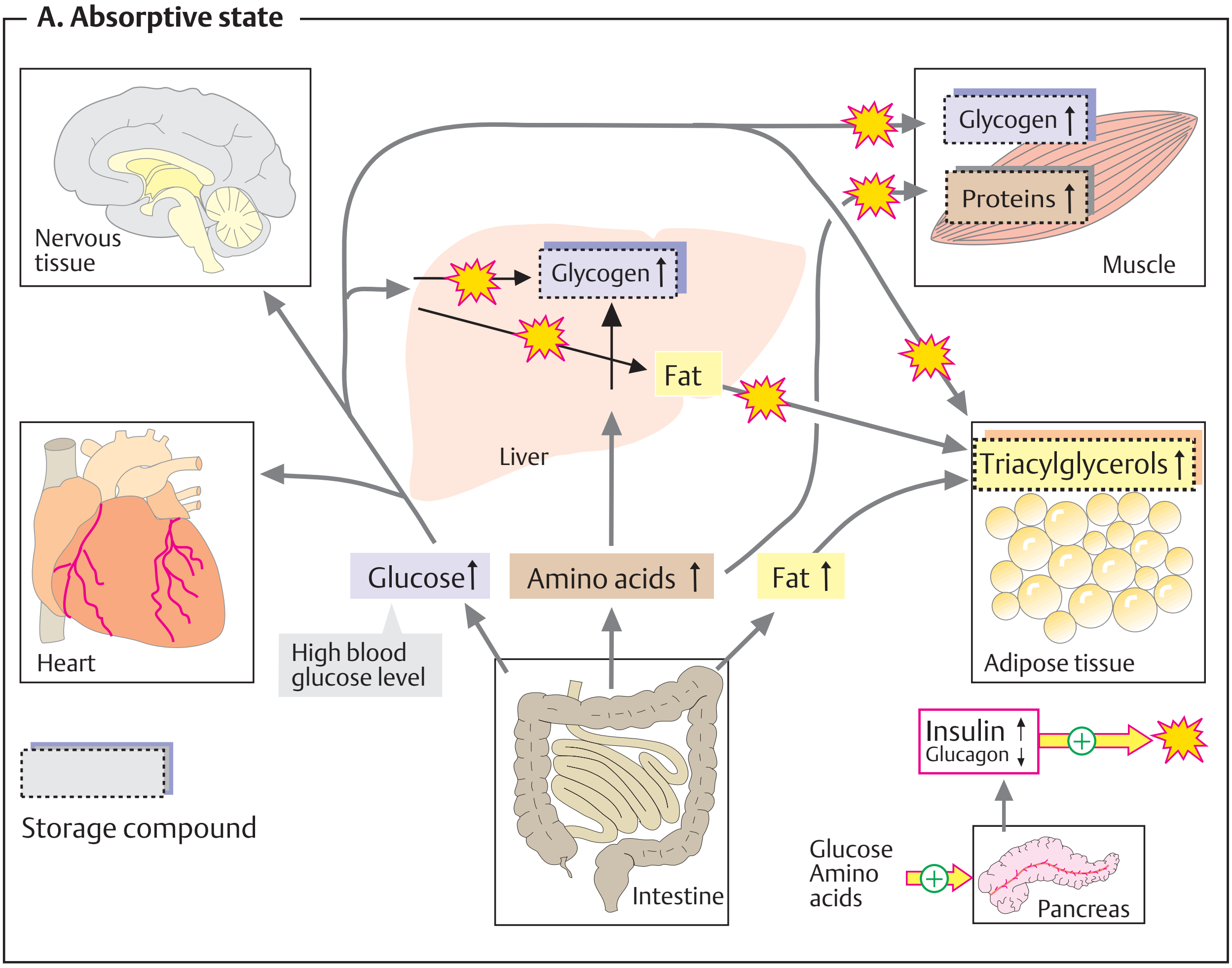
Nervous tissue
Glycogen ↑
Proteins ↑
Muscle
Glycogen ↑
Fat
Liver
Heart
Triacylglycerols ↑
Glucose ↑
Amino acids ↑
Fat ↑
High blood glucose level
Adipose tissue
Storage compound
Insulin ↑
Glucagon ↓
Glucose
Amino acids
Intestine
Pancreas

B. Postabsorptive state

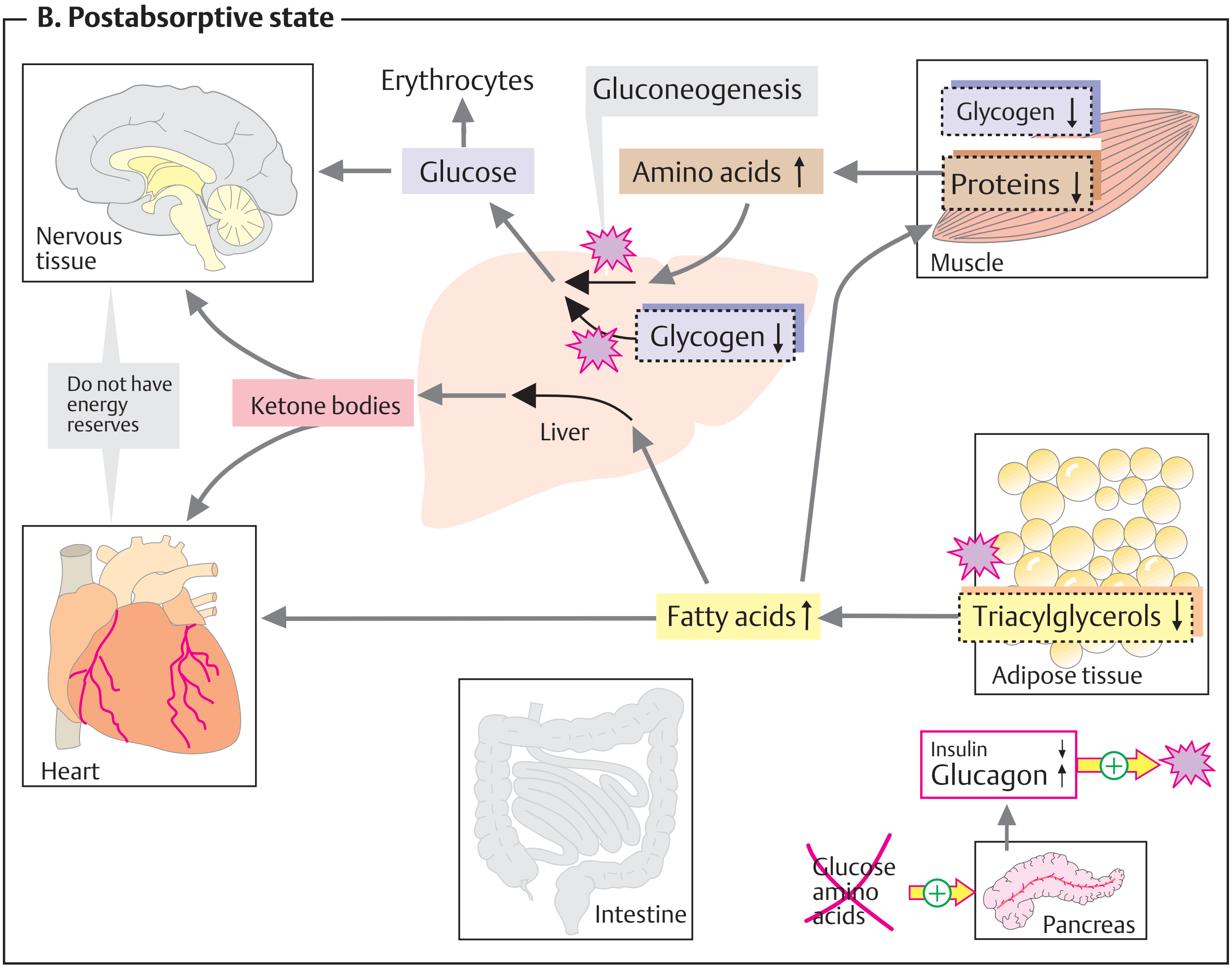
Erythrocytes
Gluconeogenesis
Nervous tissue
Glucose
Amino acids ↑
Glycogen ↓
Proteins ↓
Muscle
Glycogen ↓
Do not have energy reserves
Ketone bodies
Liver
Heart
Fatty acids ↑
Triacylglycerols ↓
Adipose tissue
Intestine
Insulin ↓
Glucagon ↑
~~Glucose amino acids~~
Pancreas

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Carbohydrate metabolism

Besides fatty acids and ketone bodies, glucose is the body's most important energy supplier. The concentration of glucose in the blood (the *"blood glucose level"*) is maintained at 4–6 mM (0.8–1.0 g L^{-1}) by precise regulation of glucosesupplying and glucose-utilizing processes. Glucose suppliers include the intestines (glucose from food), liver, and kidneys. The liver plays the role of a "glucostat" (see p. 308).

The liver is also capable of forming glucose by converting other sugars—e. g., *fructose* and *galactose*—or by synthesizing from other metabolites. The conversion of lactate to glucose in the *Cori cycle* (see p. 338) and the conversion of alanine to glucose with the help of the *alanine cycle* (see p. 338) are particularly important for the supply of erythrocytes and muscle cells.

Transporters in the plasma membrane of hepatocytes allow insulin-independent transport of glucose and other sugars in both directions. In contrast to muscle, the liver possesses the enzyme *glucose-6-phosphatase,* which can release glucose from glucose-6-phosphate.

A. Gluconeogenesis: overview ◑

Regeneration of glucose (up to 250 g per day) mainly takes place in the liver. The tubule cells of the kidney are also capable of carrying out gluconeogenesis, but due to their much smaller mass, their contribution only represents around 10% of total glucose formation. Gluconeogenesis is regulated by hormones. *Cortisol, glucagon,* and *epinephrine* promote gluconeogenesis, while *insulin* inhibits it (see pp. 158, 244).

The main precursors of gluconeogenesis in the liver are *lactate* from anaerobically working muscle cells and from erythrocytes, glucogenic *amino acids* from the digestive tract and muscles (mainly alanine), and *glycerol* from adipose tissue. The kidney mainly uses amino acids for gluconeogenesis (Glu, Gln; see p. 328).

In mammals, fatty acids and other suppliers of acetyl CoA are not capable of being used for gluconeogenesis, as the acetyl residues formed during β-oxidation in the tricarboxylic acid cycle (see p. 132) are oxidized to CO_2 and therefore cannot be converted into oxaloacetic acid, the precursor for gluconeogenesis.

B. Fructose and galactose metabolism ◑

Fructose is mainly metabolized by the liver, which channels it into glycolysis (left half of the illustration).

A special *ketohexokinase* [1] initially phosphorylates fructose into **fructose 1-phosphate**. This is then cleaved by an *aldolase* [2], which is also fructose-specific, to yield **glycerone 3-phosphate** (dihydroxyacetone phosphate) and **glyceraldehyde**. Glycerone 3-phosphate is already an intermediate of glycolysis (center), while glyceraldehyde can be phosphorylated into glyceraldehyde 3-phosphate by *triokinase* [3].

To a smaller extent, glyceraldehyde is also reduced to glycerol [4] or oxidized to glycerate, which can be channeled into glycolysis following phosphorylation (not shown). The reduction of glyceraldehyde [4] uses up NADH. As the rate of degradation of alcohol in the hepatocytes is limited by the supply of NAD^+, fructose degradation accelerates alcohol degradation (see p. 320).

Outside of the liver, fructose is channeled into the sugar metabolism by reduction at C-2 to yield sorbitol and subsequent dehydration at C-1 to yield glucose (the *polyol pathway;* not shown).

Galactose is also broken down in the liver (right side of the illustration). As is usual with sugars, the metabolism of galactose starts with a phosphorylation to yield **galactose 1-phosphate** [5]. The connection to the glucose metabolism is established by C-4 epimerization to form **glucose 1-phosphate**. However, this does not take place directly. Instead, a *transferase* [6] transfers a uridine 5′-monophosphate (UMP) residue from uridine diphosphoglucose (UDPglucose) to galactose 1-phosphate. This releases glucose 1-phosphate, while galactose 1-phosphate is converted into uridine diphosphogalactose (UDP-galactose). This then is isomerized into UDP-glucose. The *biosynthesis* of galactose also follows this reaction pathway, which is freely reversible up to reaction [5]. Genetic defects of enzymes [5] or [6] can lead to the clinical picture of *galactosemia.*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Gluconeogenesis: overview

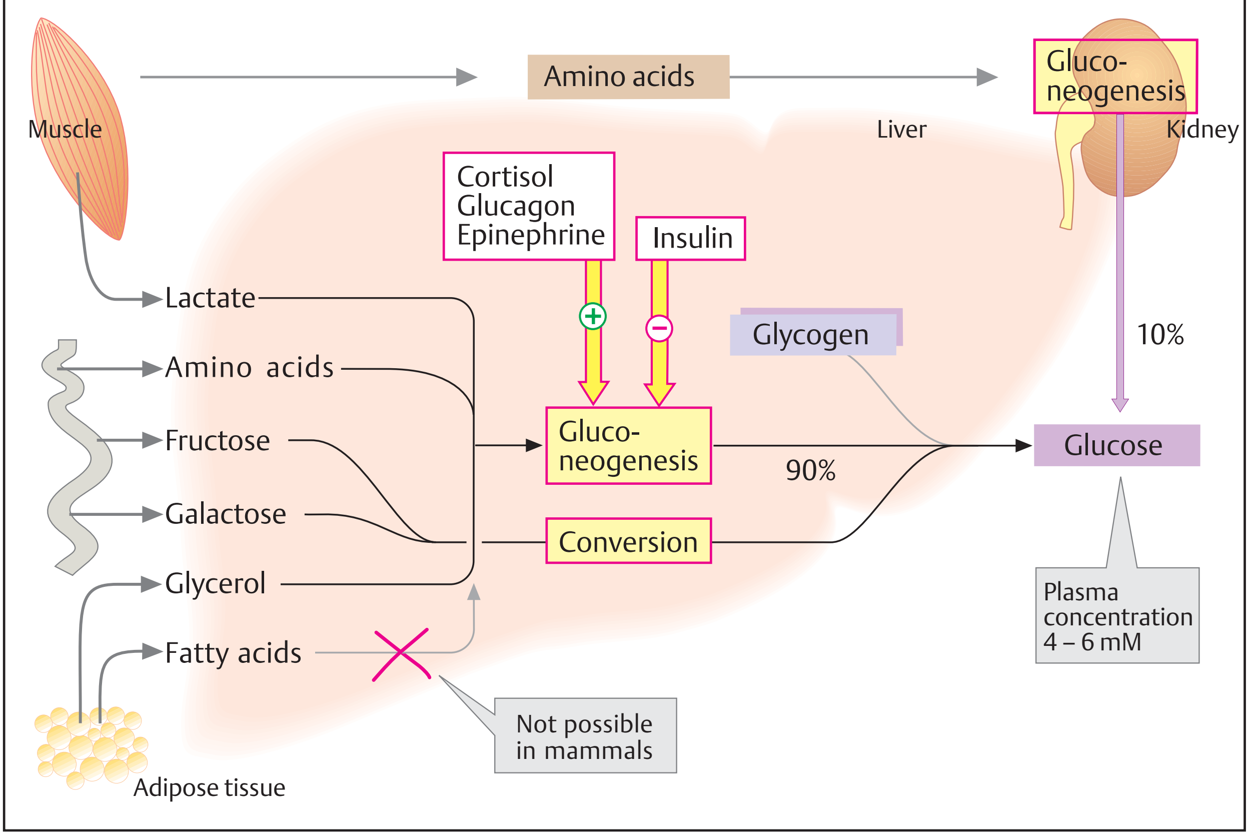
Amino acids
Gluco-neogenesis
Muscle
Liver
Kidney
Cortisol
Glucagon
Epinephrine
Insulin
+
−
Lactate
Amino acids
Fructose
Galactose
Glycerol
Fatty acids
Glycogen
10%
Gluco-neogenesis
90%
Glucose
Conversion
Plasma concentration 4 – 6 mM
Not possible in mammals
Adipose tissue

B. Fructose and galactose metabolism

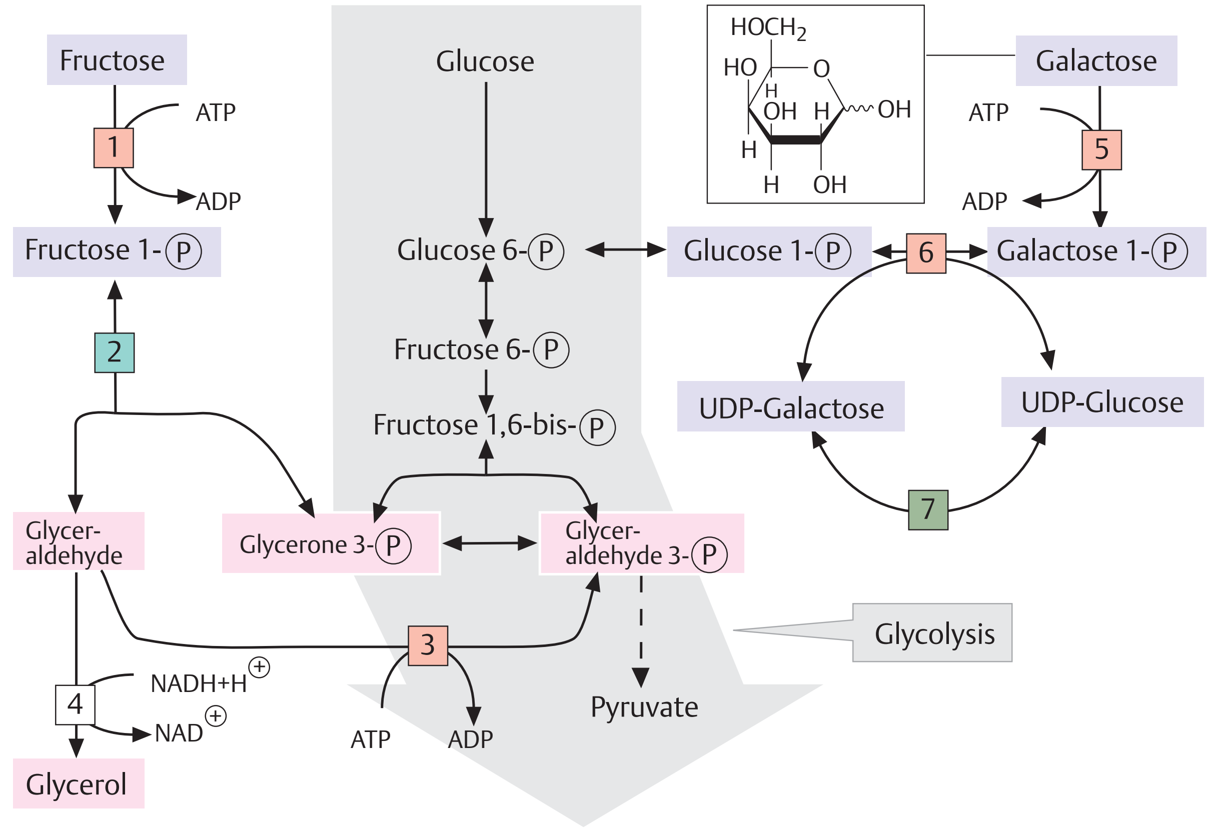
Fructose
ATP
1
ADP
Fructose 1-Ⓟ
2
Glucose
Glucose 6-Ⓟ
Fructose 6-Ⓟ
Fructose 1,6-bis-Ⓟ
HOCH2
HO
O
H
OH
H
OH
H
H
OH
Galactose
ATP
5
ADP
Glucose 1-Ⓟ
6
Galactose 1-Ⓟ
UDP-Galactose
UDP-Glucose
7
Glycer-aldehyde
Glycerone 3-Ⓟ
Glycer-aldehyde 3-Ⓟ
3
4
NADH+H+
NAD+
ATP
ADP
Pyruvate
Glycolysis
Glycerol

1 Ketohexokinase *2.7.1.3*

2 Fructose-bisphosphate aldolase *4.1.2.13*

3 Triokinase *2.7.1.28*

4 Aldehyde reductase *1.1.1.21*

5 Galactokinase *2.7.1.6*

6 Hexose-1-phosphate uridyltransferase *2.7.7.12*

7 UDPglucose 4-epimerase *5.1.3.2*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Lipid metabolism

The liver is the most important site for the formation of *fatty acids, fats (triacylglycerols), ketone bodies,* and *cholesterol.* Most of these products are released into the blood. In contrast, the triacylglycerols synthesized in adipose tissue are also stored there.

A. Lipid metabolism ◑

Lipid metabolism in the liver is closely linked to the carbohydrate and amino acid metabolism. When there is a good supply of nutrients in the *resorptive (wellfed) state* (see p. 308), the liver converts glucose via acetyl CoA into **fatty acids**. The liver can also take up fatty acids from *chylomicrons,* which are supplied by the intestine, or from fatty acid–albumin complexes (see p. 162). Fatty acids from both sources are converted into **fats** and **phospholipids**. Together with apoproteins, they are packed into very-low-density lipoproteins (**VLDLs**; see p. 278) and then released into the blood by exocytosis. The VLDLs supply extrahepatic tissue, particularly adipose tissue and muscle.

In the *postresorptive state* (see p. 292)—particularly during fasting and starvation—the lipid metabolism is readjusted and the organism falls back on its own reserves. In these conditions, adipose tissue releases fatty acids. They are taken up by the liver and are mainly converted into **ketone bodies** (**B**).

Cholesterol can be derived from two sources—food or endogenous synthesis from acetyl-CoA. A substantial percentage of endogenous cholesterol synthesis takes place in the liver. Some cholesterol is required for the synthesis of **bile acids** (see p. 314). In addition, it serves as a building block for cell membranes (see p. 216), or can be esterified with fatty acids and stored in lipid droplets. The rest is released together into the blood in the form of lipoprotein complexes (VLDLs) and supplies other tissues. The liver also contributes to the cholesterol metabolism by taking up from the blood and breaking down lipoproteins that contain cholesterol and cholesterol esters (HDLs, IDLs, LDLs; see p. 278).

B. Biosynthesis of ketone bodies ◑

At high concentrations of acetyl-CoA in the liver mitochondria, two molecules condense to form **acetoacetyl CoA** [1]. The transfer of another acetyl group [2] gives rise to 3-hydroxy-3-methylglutaryl-CoA (**HMG CoA**), which after release of acetyl CoA [3] yields free **acetoacetate** (*Lynen cycle*). Acetoacetate can be converted to **3-hydroxybutyrate** by reduction [4], or can pass into **acetone** by nonenzymatic decarboxylation [5]. These three compounds are together referred to as *"ketone bodies,"* although in fact 3-hydroxybutyrate is not actually a ketone. As reaction [3] releases an H^+ ion, metabolic acidosis can occur as a result of increased ketone body synthesis (see p. 288).

The ketone bodies are released by the liver into the blood, in which they are easily soluble. Blood levels of ketone bodies therefore rise during periods of hunger. Together with free fatty acids, 3-hydroxybutyrate and acetoacetate are then the most important energy suppliers in many tissues (including heart muscle). Acetone cannot be metabolized and is exhaled via the lungs or excreted with urine.

To channel ketone bodies into the energy metabolism, acetoacetate is converted with the help of succinyl CoA into succinic acid and acetoacetyl CoA, which is broken down by β-oxidation into acetyl CoA (not shown; see p. 180).

If the production of ketone bodies exceeds the demand for them outside the liver, there is an increase in the concentration of ketone bodies in the plasma (*ketonemia*) and they are also eventually excreted in the urine (*ketonuria*). Both phenomena are observed after prolonged starvation and in inadequately treated *diabetes mellitus.* Severe ketonuria with ketoacidosis can cause electrolyte shifts and loss of consciousness, and is therefore life-threatening (*ketoacidotic coma*).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Lipid metabolism

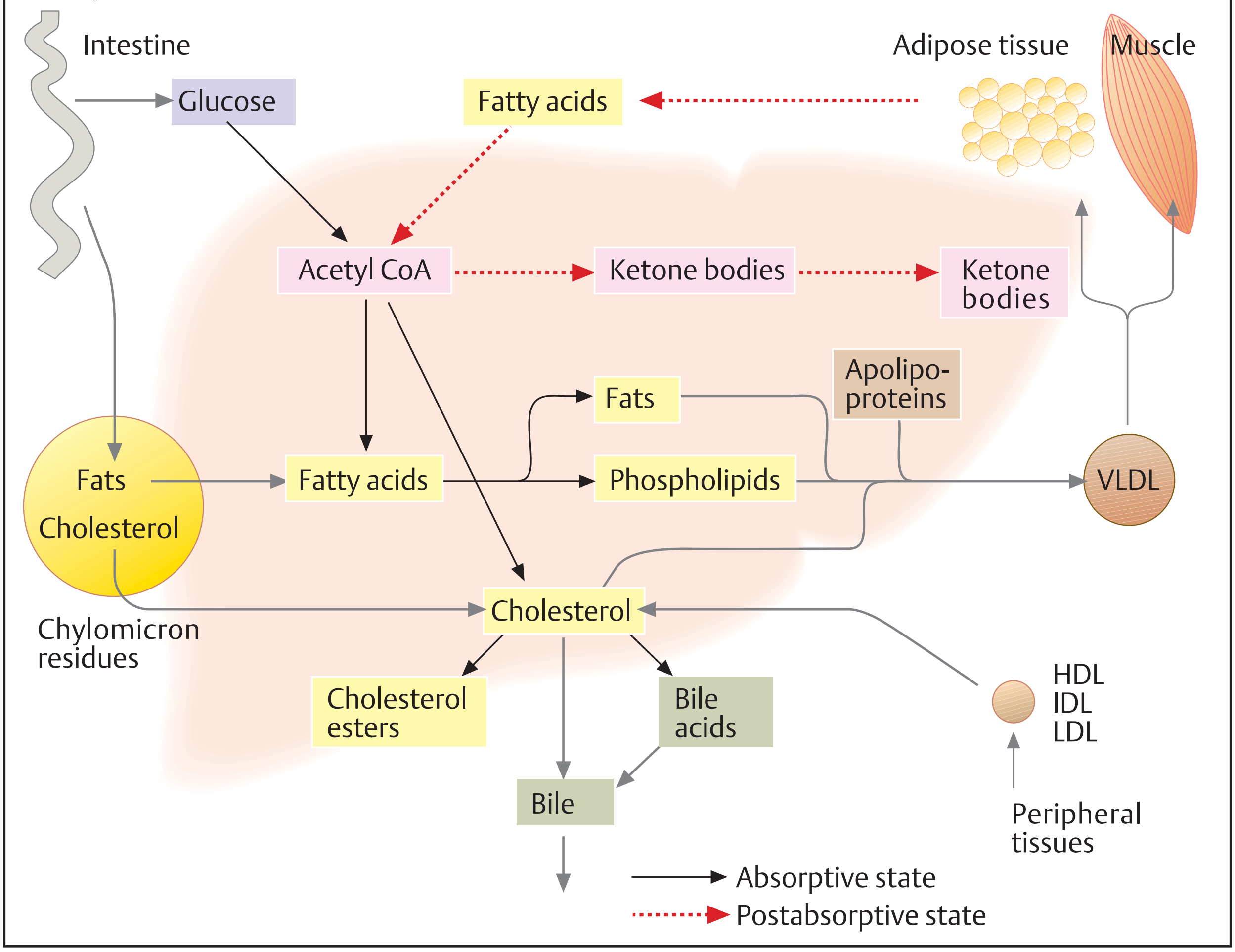
Intestine
Glucose
Fatty acids
Adipose tissue
Muscle
Acetyl CoA
Ketone bodies
Ketone bodies
Apolipo-proteins
Fats
Fats
Cholesterol
Fatty acids
Phospholipids
VLDL
Chylomicron residues
Cholesterol
Cholesterol esters
Bile acids
HDL
IDL
LDL
Bile
Peripheral tissues
Absorptive state
Postabsorptive state

A. Biosynthesis of ketone bodies

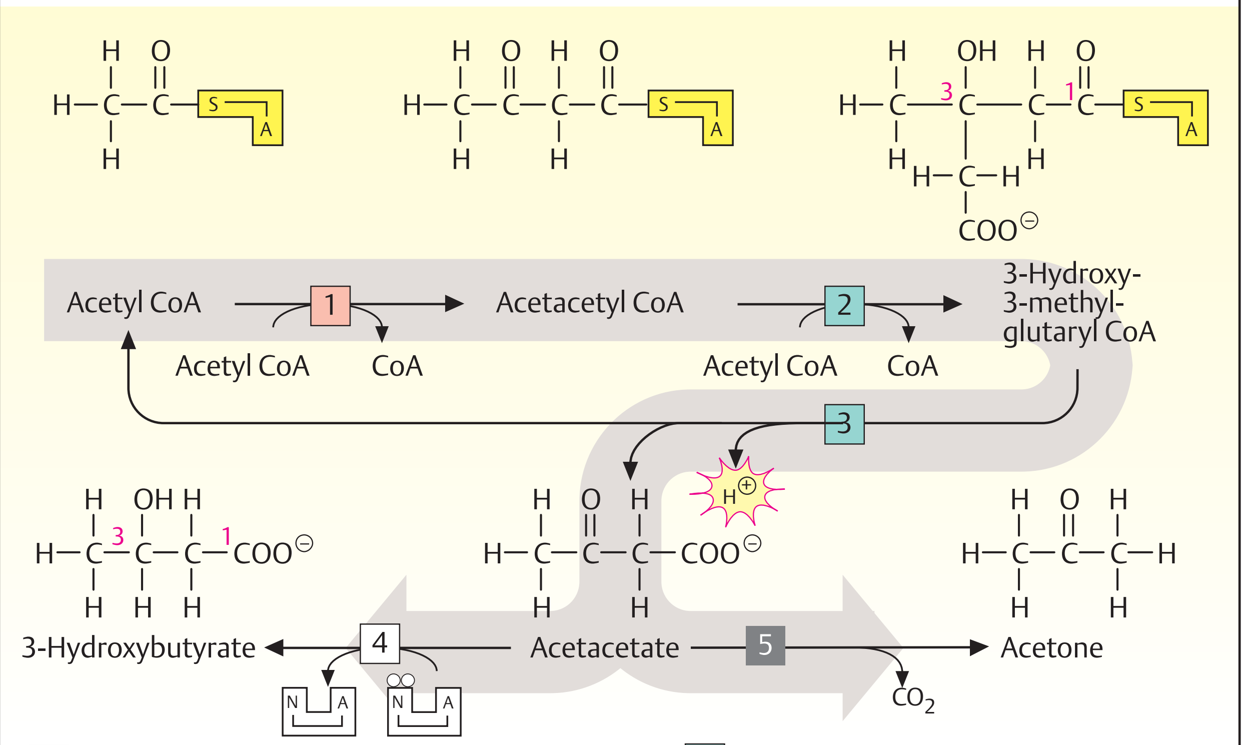
H O
H–C–C–S A
H
H O H O
H–C–C–C–C–S A
H H
H OH H O
H–C–3–C–C–1–C–S A
H H–C–H H
COO⊖
Acetyl CoA
1
Acetacetyl CoA
2
3-Hydroxy-3-methyl-glutaryl CoA
Acetyl CoA
CoA
Acetyl CoA
CoA
3
H⊕
H OH H
H–C–3–C–C–1–COO⊖
H H H
H O H
H–C–C–C–COO⊖
H H
H O H
H–C–C–C–H
H H
3-Hydroxybutyrate
4
Acetacetate
5
Acetone
CO2
N A
N A

- 1 Acetyl-CoA-C-acyltransferase *2.3.1.16*
- 2 Hydroxymethylglutaryl-CoA synthase *4.1.3.5*
- 3 Hydroxymethylglutaryl-CoA lyase *4.1.3.4*
- 4 3-Hydroxybutyrate dehydrogenase *1.1.1.30*
- 5 Nonenzymatic reaction

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Bile acids

Bile is an important product released by the hepatocytes. It promotes the digestion of fats from food by *emulsifying* them in the small intestine (see p. 2770). The emulsifying components of bile, apart from phospholipids, mainly consist of **bile acids** and **bile salts** (see below). The bile also contains free cholesterol, which is excreted in this way (see p. 312).

A. Bile acids and bile salts ◑

Bile acids are steroids consisting of 24 C atoms carrying one carboxylate group and several hydroxyl groups. They are formed from cholesterol in the liver via an extensive reaction pathway (top). Cytochrome P450 enzymes in the sER of hepatocytes are involved in many of the steps (seep. 318). Initially, the cholesterol double bond is removed. Monooxygenases then introduce one or two additional OH groups into the sterane framework. Finally, the side chain is shortened by three C atoms, and the terminal C atom is oxidized to a carboxylate group.

It is important that the arrangement of the A and B rings is altered from *trans* to *cis* during bile acid synthesis (see p. 54). The result of this is that all of the hydrophilic groups in the bile acids lie on one side of the molecule. Cholesterol, which is weakly amphipathic (top), has a small polar "head" and an extended apolar "tail." By contrast, the much more strongly amphipathic bile acid molecules (bottom) resemble disks with polar top sides and apolar bottom sides. At physiological pH values, the carboxyl groups are almost completely dissociated and therefore negatively charged.

Cholic acid and *chenodeoxycholic acid,* known as the **primary bile acids**, are quantitatively the most important metabolites of cholesterol. After being biosynthesized, they are mostly activated with coenzyme A and then conjugated with *glycine* or the non-proteinogenic amino acid *taurine* (see p. 62). The acid amides formed in this way are known as **conjugated bile acids** or **bile salts**. They are even more amphipathic than the primary products.

Deoxycholic acid and *lithocholic acid* are only formed in the intestine by enzymatic cleavage of the OH group at C-7 (see **B**). They are therefore referred to as **secondary bile acids**.

B. Metabolism of bile salts ◑

Bile salts are exclusively synthesized in the liver (see **A**). The slowest step in their biosynthesis is hydroxylation at position 7 by a *7-α-hydroxylase.* Cholic acid and other bile acids inhibit this reaction (*end-product inhibition*). In this way, the bile acids present in the liver regulate the rate of cholesterol utilization.

Before leaving the liver, a large proportion of the bile acids are activated with CoA and then conjugated with the amino acids *glycine* or *taurine* (**2**; cf. **A**). In this way, cholic acid gives rise to *glycocholic acid* and *taurocholic acid.* The *liver bile* secreted by the liver becomes denser in the gallbladder as a result of the removal of water (*bladder bile*; **3**).

Intestinal bacteria produce enzymes that can chemically alter the bile salts (**4**). The acid amide bond in the bile salts is cleaved, and dehydroxylation at C-7 yields the corresponding secondary bile acids from the primary bile acids (**5**). Most of the intestinal bile acids are resorbed again in the ileum (**6**) and returned to the liver via the portal vein (*enterohepatic circulation*). In the liver, the secondary bile acids give rise to primary bile acids again, from which bile salts are again produced. Of the 15–30g bile salts that are released with the bile per day, only around 0.5g therefore appears in the feces. This approximately corresponds to the amount of daily de novo synthesis of cholesterol.

Further information

The cholesterol excreted with the bile is poorly water-soluble. Together with phospholipids and bile acids, it forms micelles (see p. 270), which keep it in solution. If the proportions of phospholipids, bile acids and cholesterol shift, *gallstones* can arise. These mainly consist of precipitated cholesterol (cholesterol stones), but can also contain Ca^{2+} salts of bile acids and bile pigments (pigment stones).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Bile acids and bile salts

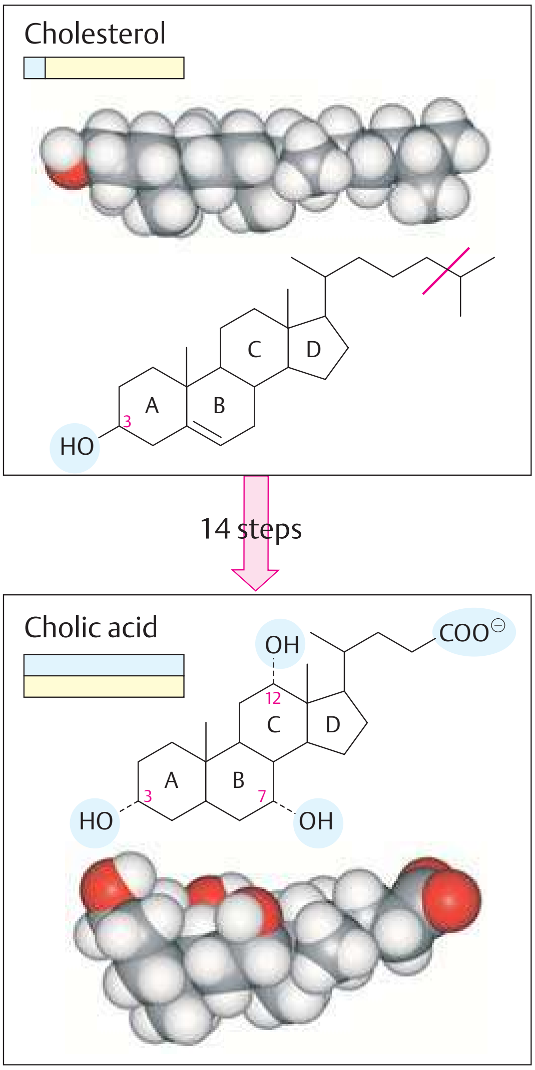
Cholesterol
HO
3
A
B
C
D
14 steps
Cholic acid
OH
12
7
COO⊖

Bile acids	HO– in position		
Primary bile acids			
— Cholic acid	C-3	C-7	C-12
— Chenodeoxycholic acid	C-3	C-7	-
Secondary bile acids			
— Deoxycholic acid	C-3	-	C-12
— Lithocholic acid	C-3	-	-

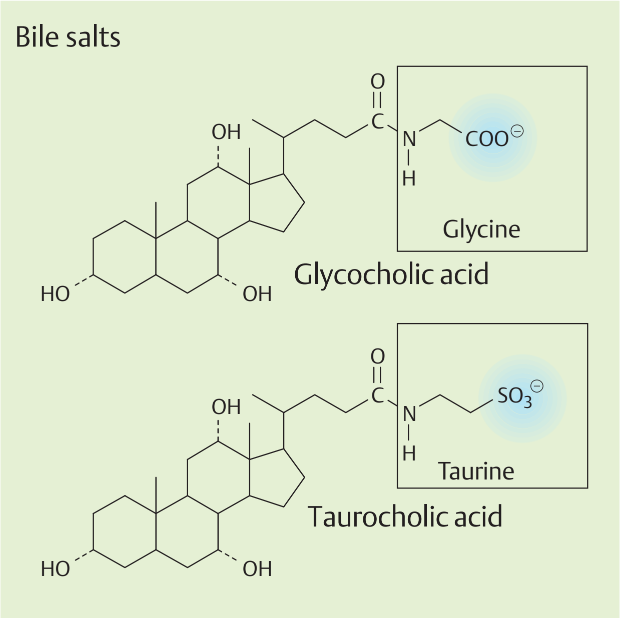
Bile salts
OH
O
C
N
H
COO⊖
Glycine
HO
OH
Glycocholic acid
OH
O
C
N
H
SO$_3$⊖
Taurine
Taurocholic acid
HO
OH

Bile salts = conjugated bile acids

B. Metabolism of bile salts

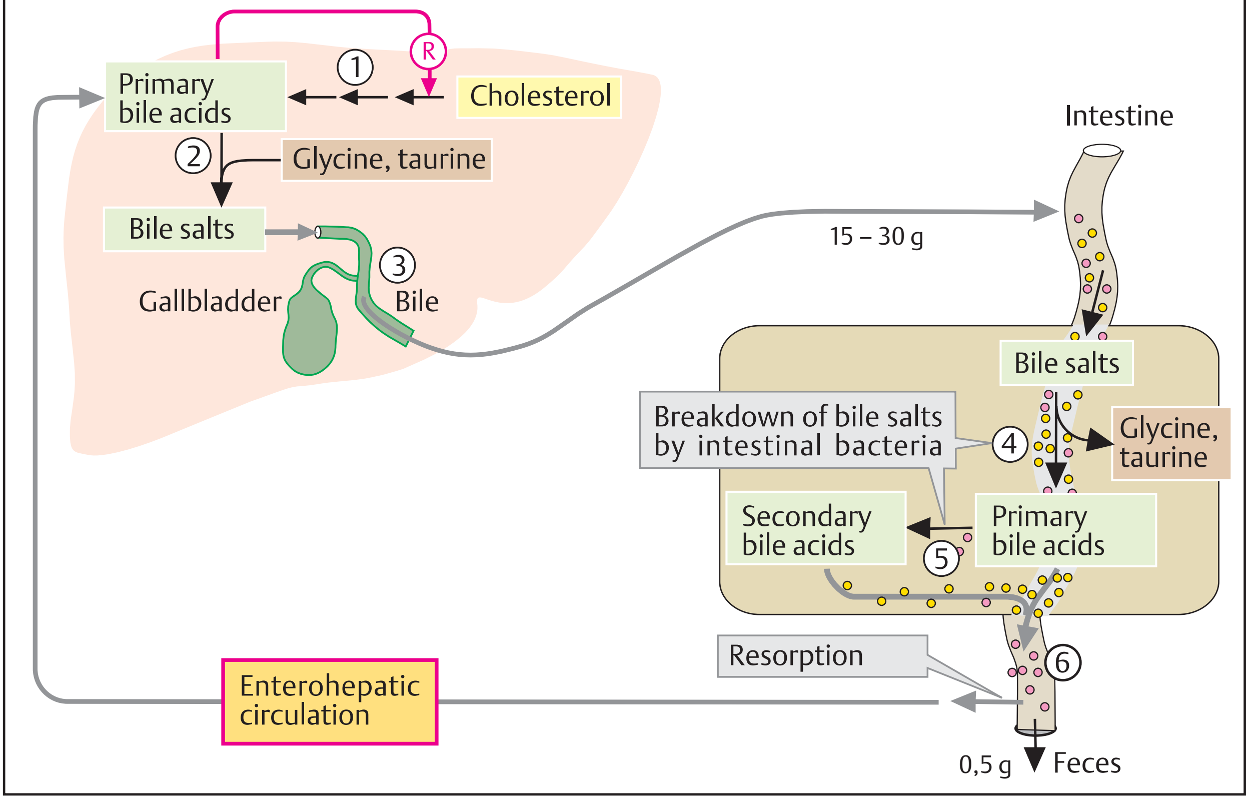
R
①
Primary bile acids
Cholesterol
②
Glycine, taurine
Bile salts
③
Gallbladder
Bile
15 – 30 g
Intestine
Bile salts
Breakdown of bile salts by intestinal bacteria
④
Glycine, taurine
Secondary bile acids
⑤
Primary bile acids
Resorption
⑥
Enterohepatic circulation
0,5 g
Feces

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Biotransformations

The body is constantly taking up **foreign substances** (= *xenobiotics*) from food or through contact with the environment, via the skin and lungs. These substances can be natural in origin, or may have been synthetically produced by humans. Many of these substances are toxic, particularly at high concentrations. However, the body has effective mechanisms for inactivating and then excreting foreign substances through *biotransformations.* The mechanisms of biotransformation are similar to those with which **endogenous substances** such as bile pigments and steroid hormones are enzymatically converted. Biotransformations mainly take place in the *liver.*

A. Biotransformations ◑

Phase I reactions (interconversion reactions). Type I reactions introduce functional groups into inert, apolar molecules or alter functional groups that are already present. In many cases, this is what first makes it possible for foreign substances to conjugate with polar molecules via phase II reactions (see below). Phase I reactions usually reduce the *biological activity* or *toxicity* of a substance ("detoxification"). However, some substances only become biologically active as a result of the interconversion reaction (see, for example, benzo[*a*]pyrene, p. 256) or become more toxic after interconversion than the initial substance ("toxification").

Important phase I biotransformation reactions include:

- **Hydrolytic cleavages** of ether, ester, and peptide bonds. Example (**1**) shows hydrolysis of the painkiller *acetylsalicylic acid.*
- **Oxidations.** Hydroxylations, epoxide formation, sulfoxide formation, dealkylation, deamination. For example, benzene is oxidized into phenol, and toluene (methylbenzene) is oxidized into benzoic acid.
- **Reductions.** Reduction of carbonyl, azo-, or nitro- compounds, dehalogenation.
- **Methylations.** Example (**2**) illustrates the inactivation of the catecholamine *norepinephrine* by methylation of a phenolic OH group (see p. 334).
- **Desulfurations.** The reactions take place in the hepatocytes on the smooth endoplasmic reticulum.

Most oxidation reactions are catalyzed by **cytochrome P450 systems** (see p. 318). These monooxygenases are induced by their substrates and show wide specificity. The substrate-specific enzymes of the steroid metabolism (see p. 376) are exceptions to this.

Phase II reactions (conjugate formation). Type II reactions couple their substrates (bilirubin, steroid hormones, drugs, and products of phase I reactions) via ester or amide bonds to highly polar negatively charged molecules. The enzymes involved are transferases, and their products are known as **conjugates.**

The most common type of conjugate formation is coupling with *glucuronate* (GlcUA) as an *O*-or *N*-glucuronide. The coenzyme for the reaction is uridine diphosphate glucuronate, the *"active glucuronate"* (see p. 110). Coupling with the polar glucuronate makes an apolar (hydrophobic) molecule more strongly polar, and it becomes sufficiently water-soluble and capable of being excreted. Example (**3**) shows the glucuronidation of *tetrahydrocortisol,* a metabolite of the glucocorticoid cortisol (see p. 374).

The biosynthesis of sulfate esters with the help of *phosphoadenosine phosphosulfate* (PAPS), the *"active sulfate",* (see p. 110) and amide formation with *glycine* and *glutamine* also play a role in conjugation. For example, benzoic acid is conjugated with glycine to form the more soluble and less toxic *hippuric acid* (*N*-benzoylglycine; see p. 324).

In contrast with unconjugated compounds, the conjugates are much more water-soluble and capable of being excreted. The conjugates are eliminated from the liver either by the *biliary* route—i. e., by receptor-mediated excretion into the bile—or by the *renal* route, via the blood and kidneys by filtration.

Further information

To detoxify **heavy metals**, the liver contains *metallothioneins,* a group of cysteine-rich proteins with a high affinity for divalent metal ions such as Cd^{2+}, Cu^{2+}, Hg^{2+}, and Zn^{2+}. These metal ions also induce the formation of metallothioneins via a special metal-regulating element (MRE) in the gene's promoter (see p. 244).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biotransformations

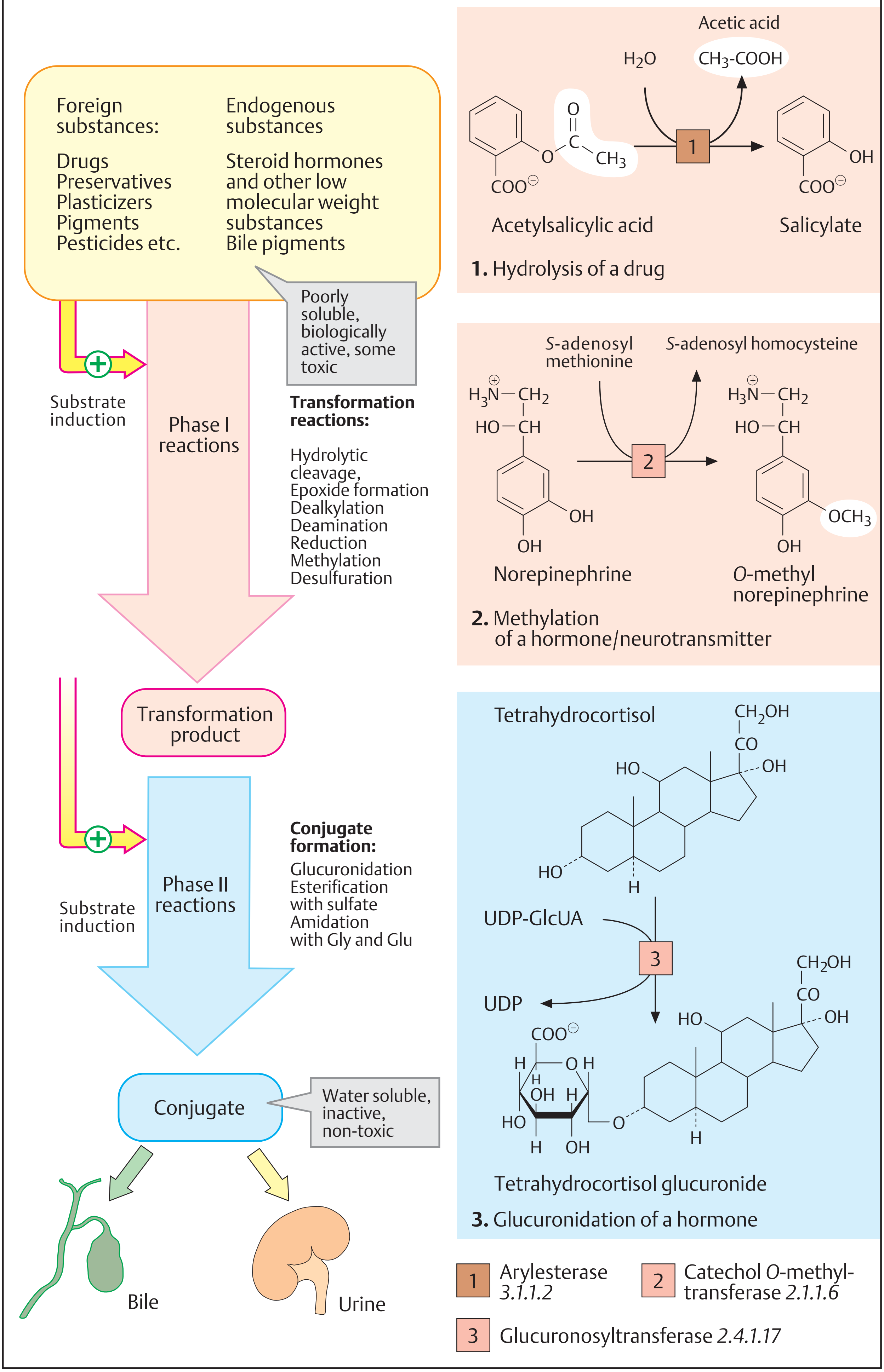
Foreign substances:
Drugs
Preservatives
Plasticizers
Pigments
Pesticides etc.
Endogenous substances
Steroid hormones and other low molecular weight substances
Bile pigments
Poorly soluble, biologically active, some toxic
Substrate induction
Phase I reactions
Transformation reactions:
Hydrolytic cleavage,
Epoxide formation
Dealkylation
Deamination
Reduction
Methylation
Desulfuration
Transformation product
Substrate induction
Phase II reactions
Conjugate formation:
Glucuronidation
Esterification with sulfate
Amidation with Gly and Glu
Conjugate
Water soluble, inactive, non-toxic
Bile
Urine
Acetic acid
H2O
CH3-COOH
O
C
O
CH3
COO⊖
OH
1
Acetylsalicylic acid
Salicylate
1. Hydrolysis of a drug
S-adenosyl methionine
S-adenosyl homocysteine
H3N⊕—CH2
HO—CH
OH
OH
2
OCH3
Norepinephrine
O-methyl norepinephrine
2. Methylation of a hormone/neurotransmitter
Tetrahydrocortisol
CH2OH
CO
OH
HO
HO
H
UDP-GlcUA
3
UDP
COO⊖
H
O H
H
OH
HO
H
OH
O
Tetrahydrocortisol glucuronide
3. Glucuronidation of a hormone
1 Arylesterase 3.1.1.2
2 Catechol O-methyltransferase 2.1.1.6
3 Glucuronosyltransferase 2.4.1.17

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Cytochrome P450 systems

During the first phase of biotransformation in the liver, compounds that are weakly chemically reactive are enzymatically hydroxylated (see p. 316). This makes it possible for them to be conjugated with polar substances. The hydroxylating enzymes are generally *monooxygenases* that contain a **heme** as the redox-active coenzyme (see p. 106). In the reduced form, the heme can bind carbon monoxide (CO), and it then shows characteristic light absorption at 450 nm. This was what led to this enzyme group being termed **cytochrome P450** (Cyt P450).

Cyt P450 systems are also involved in many other metabolic processes—e. g., the biosynthesis of steroid hormones (see p. 172), bile acids (see p. 314), and eicosanoids (see p. 390), as well as the formation of unsaturated fatty acids (see p. 409). The liver's reddish-brown color is mainly due to the large amounts of P450 enzymes it contains.

A. Cytochrome P450-dependent mono oxygenases: reactions ◑

Cyt P450-dependent monooxygenases catalyze reductive cleavage of molecular oxygen (O_2). One of the two oxygen atoms is transferred to the substrate, while the other is released as a water molecule. The necessary reducing equivalents are transferred to the actual monooxygenase by an FAD-containing auxiliary enzyme from the coenzyme NADPH+H^+.

Cyt P450 enzymes occur in numerous forms in the liver, steroid-producing glands, and other organs. The substrate specificity of liver enzymes is low. Apolar compounds containing aliphatic or aromatic rings are particularly easily converted. These include endogenous substances such as steroid hormones, as well as medical drugs, which are inactivated by phase I reactions. This is why Cyt P450 enzymes are of particular interest in pharmacology. The degradation of ethanol in the liver is also partly catalyzed by Cyt P450 enzymes (the "microsomal ethanol-oxidizing system"; see p. 304). As alcohol and drugs are broken down by the same enzyme system, the effects of alcoholic drinks and medical drugs can sometimes be mutually enhancing—even sometimes to the extent of becoming life-threatening.

Only a few examples of the numerous Cyt P450-dependent reactions are shown here. *Hydroxylation* of aromatic rings (**a**) plays a central part in the metabolism of medicines and steroids. Aliphatic methyl groups can also be oxidized to hydroxyl groups (**b**). *Epoxidation* of aromatics (**c**) by Cyt P450 yields products that are highly reactive and often toxic. For example, the mutagenic effect of benzo[*a*]pyrene (see p. 244) is based on this type of interconversion in the liver. In Cyt P450 dependent *dealkylations* (**d**), alkyl substituents of *O, N,* or *S* atoms are released as aldehydes.

B. Reaction mechanism ○

The course of Cyt P450 catalysis is in principle well understood. The most important function of the *heme group* consists of converting molecular oxygen into an especially reactive atomic form, which is responsible for all of the reactions described above.

[1] In the resting state, the heme iron is trivalent. Initially, the substrate binds near the heme group.

[2] Transfer of an electron from $FADH_2$ reduces the iron to the divalent form that is able to bind an O_2 molecule (2).

[3] Transfer of a second electron and a change in the valence of the iron reduce the bound O_2 to the peroxide.

[4] A hydroxyl ion is now cleaved from this intermediate. Uptake of a proton gives rise to H_2O and the reactive form of oxygen mentioned above. In this ferryl radical, the iron is formally tetravalent.

[5] The activated oxygen atom inserts itself into a C–H bond in the substrate, thereby forming an OH group.

[6] Dissociation of the product returns the enzyme to its initial state.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Cytochrome P450-dependent monooxygenases: reactions

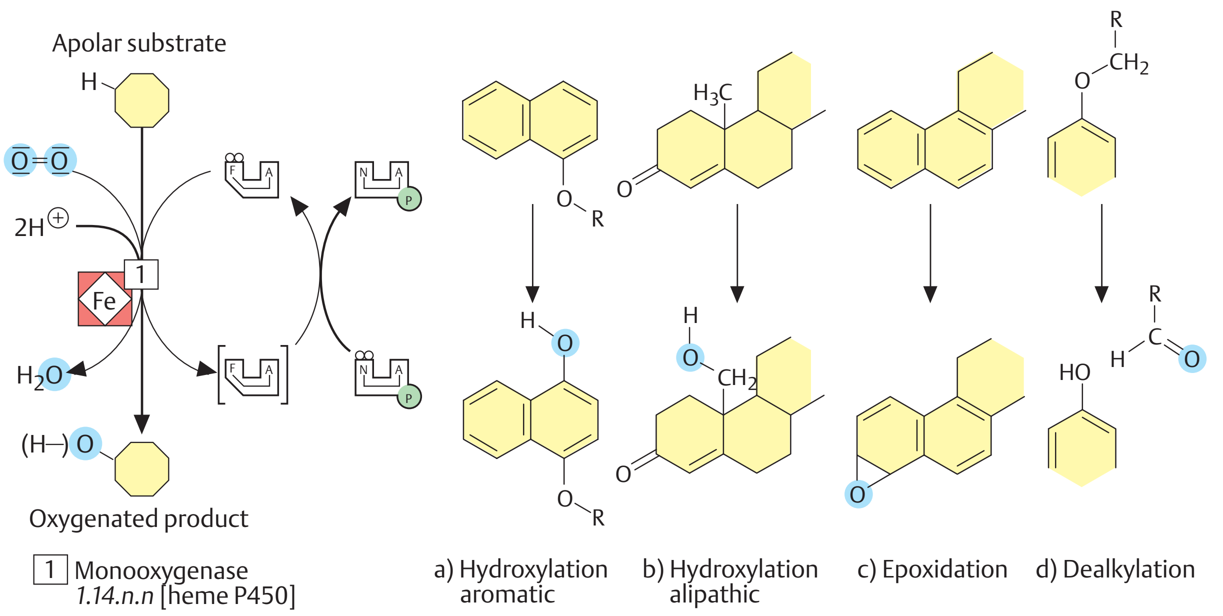
Apolar substrate
H
O=O
2H⊕
1
Fe
H2O
(H–) O
Oxygenated product
1 Monooxygenase
1.14.n.n [heme P450]
F A
N A
P
O R
H O
O R
H3C
O
H
O
CH2
O
O
R
O CH2
R
H C O
HO
a) Hydroxylation aromatic
b) Hydroxylation alipathic
c) Epoxidation
d) Dealkylation

B. Reaction mechanism

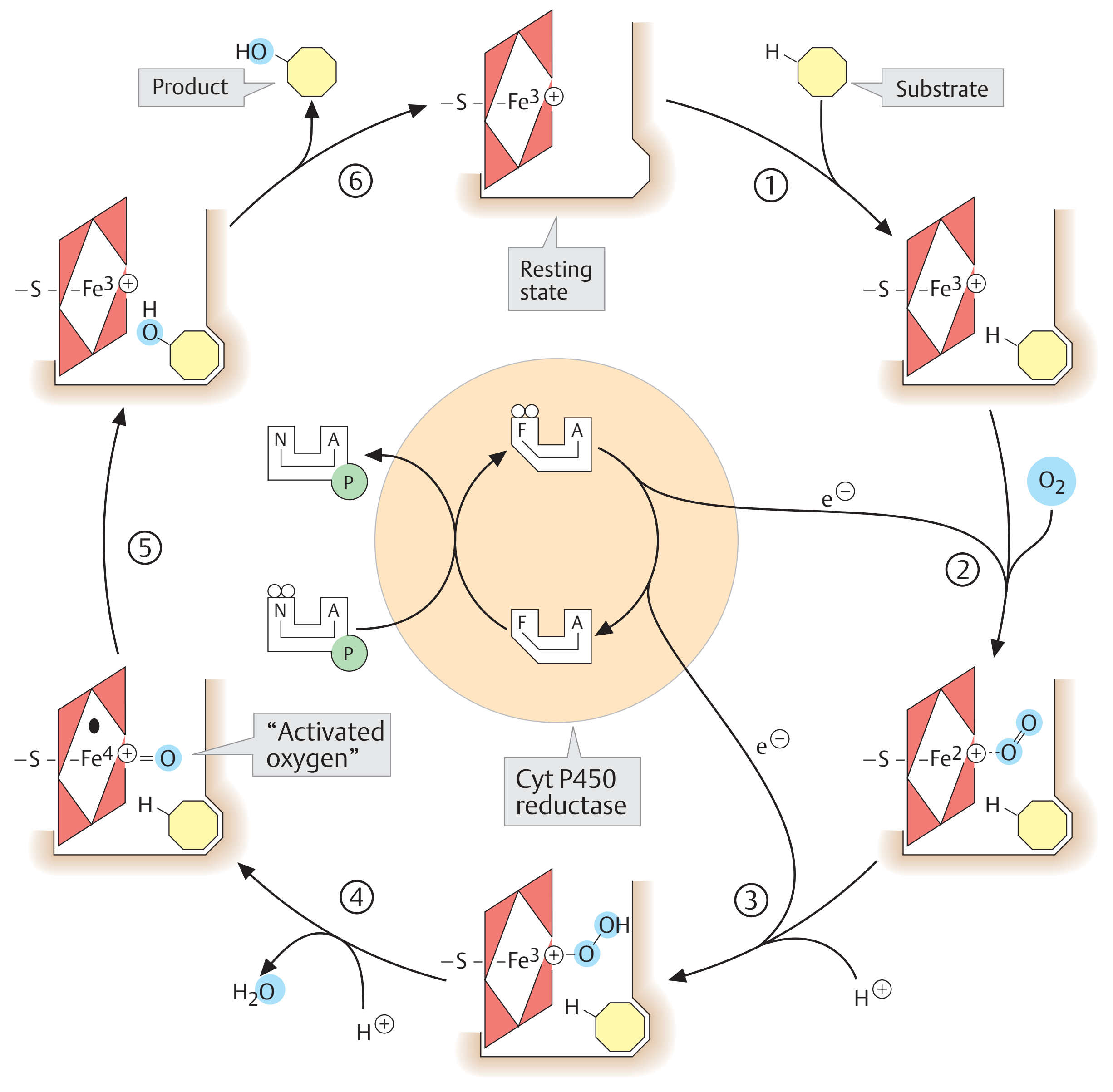
HO
Product
–S – –Fe3⊕
H
Substrate
①
②
③
④
⑤
⑥
Resting state
–S – –Fe3⊕
H
O
O2
e⊖
F A
N A
P
Cyt P450 reductase
–S – –Fe4⊕=O
"Activated oxygen"
–S – –Fe2⊕··O O
–S – –Fe3⊕–O OH
H2O
H⊕
e⊖
H⊕

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Ethanol metabolism

A. Blood ethanol level ◑

Ethanol (EtOH, "alcohol") naturally occurs in fruit in small quantities. Alcoholic drinks contain much higher concentrations. Their alcohol content is usually given as percent by volume. To estimate alcohol uptake and the blood alcohol level, it is useful to convert the amount to grams of ethanol (density 0.79 kg L^{-1}). For example, a bottle of beer (0.5 L at 4% v/v alcohol) contains 20 mL = 16 g of ethanol, while a bottle of wine (0.7 L at 12% v/v alcohol) contains 84 mL = 66 g ethanol.

Ethanol is membrane-permeable and is quickly resorbed. The maximum blood level is already reached within 60–90 min after drinking. The *resorption rate* depends on various conditions, however. An empty stomach, a warm drink (e. g., mulled wine), and the presence of sugar and carbonic acid (e. g., in champagne) promote ethanol resorption, whereas a heavy meal reduces it. Ethanol is rapidly distributed throughout the body. A large amount is taken up by the muscles and brain, but comparatively little by adipose tissue and bones. Roughly 70% of the body is accessible to alcohol. Complete resorption of the ethanol contained in one bottle of beer (16 g) by a person weighing 70 kg (distribution in 70 kg 70/100 = 49 kg) leads to a blood alcohol level of 0.33 per thousand (7.2 mM). The lethal concentration of alcohol is approximately 3.5 per thousand (76 mM).

B. Ethanol metabolism ◑

The major site of ethanol degradation is the liver, although the stomach is also able to metabolize ethanol. Most **ethanol** is initially oxidized by *alcohol dehydrogenase* to form **ethanal** (acetaldehyde). A further oxidization, catalyzed by *aldehyde dehydrogenase,* leads to acetate. Acetate is then converted with the help of *acetate-CoA ligase* to form **acetyl CoA**, using ATP and providing a link to the intermediary metabolism. In addition to cytoplasmic alcohol dehydrogenase, *catalase* and inducible *microsomal alcohol oxidase* ("MEOS"; see p. 318) also contribute to a lesser extent to ethanol degradation. Many of the enzymes mentioned above are induced by ethanol.

The rate of ethanol degradation in the liver is limited by alcohol dehydrogenase activity. The amount of NAD$^+$ available is the limiting factor. As the maximum degradation rate is already reached at low concentrations of ethanol, the ethanol level therefore declines at a constant rate (zero-order kinetics). The *calorific value* of ethanol is 29.4 kJ g^{-1}. Alcoholic drinks—particularly in alcoholics—can therefore represent a substantial proportion of dietary energy intake.

C. Liver damage due to alcohol ◑

Alcohol is a socially accepted drug of abuse in Western countries. Due to the high potential for addiction to develop, however, it is actually a "hard" drug and has a much larger number of victims than the opiate drugs, for example. In the brain, ethanol is deposited in membranes due to its amphipathic properties, and it influences receptors for neurotransmitters (see p. 352). The effect of GABA is enhanced, while that of glutamate declines.

High ethanol consumption over many years leads to liver damage. For a healthy man, the limit is about 60 g per day, and for a woman about 50 g. However, these values are strongly dependent on body weight, health status, and other factors.

Ethanol-related high levels of NADH+H$^+$ and acetyl-CoA in the liver lead to increased synthesis of neutral fats and cholesterol. However, since the export of these in the form of VLDLs (see p. 278) is reduced due to alcohol, storage of lipids occurs (**fatty liver**). This increase in the fat content of the liver (from less than 5% to more than 50% of the dry weight) is initially reversible. However, in chronic alcoholism the hepatocytes are increasingly replaced by connective tissue. When **liver cirrhosis** occurs, the damage to the liver finally reaches an irreversible stage, characterized by progressive loss of liver functions.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Blood ethanol level

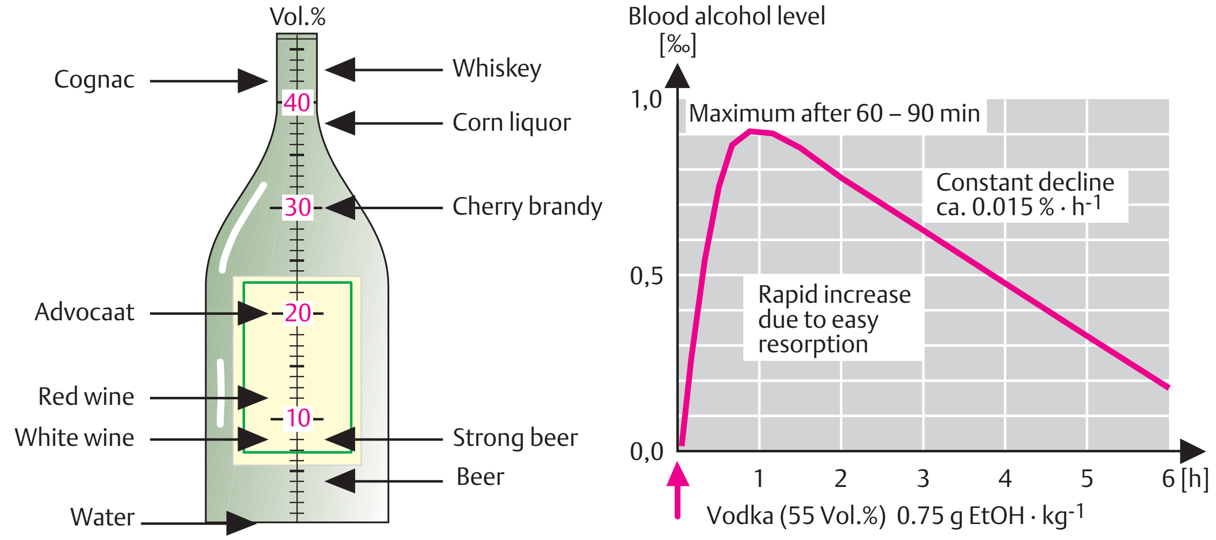

B. Ethanol metabolism

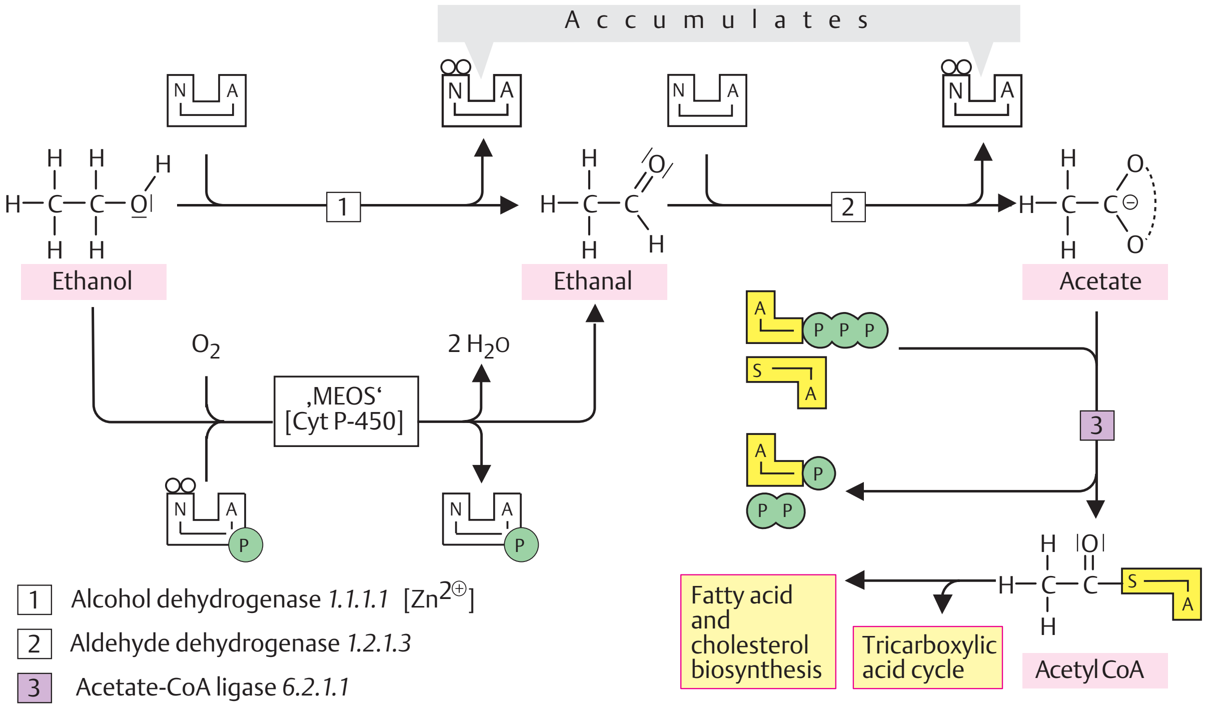
Accumulates
N A
H–C–C–O–H
Ethanol
1
Ethanal
2
Acetate
O2
2 H2O
‚MEOS'
[Cyt P-450]
P
A PPP
S A
3
A P
P P
Acetyl CoA
Fatty acid and cholesterol biosynthesis
Tricarboxylic acid cycle
1 Alcohol dehydrogenase 1.1.1.1 [Zn2⊕]
2 Aldehyde dehydrogenase 1.2.1.3
3 Acetate-CoA ligase 6.2.1.1

C. Liver damage due to alcohol

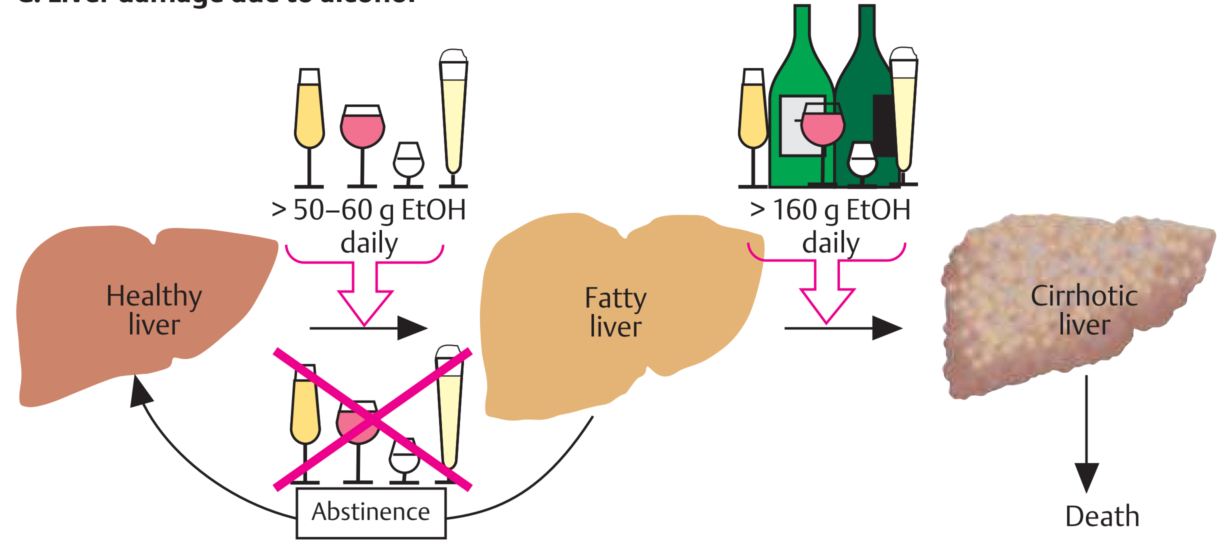
> 50–60 g EtOH daily
> 160 g EtOH daily
Healthy liver
Fatty liver
Cirrhotic liver
Abstinence
Death

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Kidney: functions

A. Functions of the kidneys ●

The kidneys' main function is **excretion** of water and water-soluble substances (**1**). This is closely associated with their role in regulating the body's electrolyte and acid–base balance (**homeostasis, 2**; see pp. 326 and 328). Both excretion and homeostasis are subject to hormonal control. The kidneys are also involved in synthesizing several **hormones** (**3**; see p. 315). Finally, the kidneys also play a role in the **intermediary metabolism** (**4**), particularly in amino acid degradation and gluconeogenesis (see p. 154).

The kidneys are extremely well-perfused organs, with about 1500 L of blood flowing through them every day. Approximately 180 L of primary urine is filtered out of this. Removal of water leads to extreme concentration of the primary urine (to approximately one-hundredth of the initial volume). As a result, only a volume of 0.5–2.0 L of **final urine** is excreted per day.

B. Urine formation ◑

The functional unit of the kidney is the **nephron**. It is made up of the Malpighian bodies or renal corpuscles (consisting of Bowman's capsules and the glomerulus), the proximal tubule, Henle's loop, and the distal tubule, which passes into a collecting duct. The human kidney contains around one million nephrons. The nephrons form urine in the following three phases.

Ultrafiltration. Ultrafiltration of the blood plasma in the glomerulus gives rise to *primary urine*, which is *isotonic* with plasma. The pores in the glomerular basal membrane, which are made up of type IV collagen (see p. 344), have an effective mean diameter of 2.9 nm. This allows all plasma components with a molecular mass of up to about 15 kDa to pass through unhindered. At increasing masses, molecules are progressively held back; at masses greater than 65 kDa, they are completely unable to enter the primary urine. This applies to almost all plasma proteins—which in addition, being anions, are repelled by the negative charge in the basal membrane.

Resorption. All low-molecular weight plasma components enter the primary urine via glomerular filtration. Most of these are transported back into the blood by resorption, to prevent losses of valuable metabolites and electrolytes. In the proximal tubule, organic metabolites (e. g., glucose and other sugars, amino acids, lactate, and ketone bodies) are recovered by secondary active transport (see p. 220). There are several group-specific transport systems for resorbing amino acids, with which hereditary diseases can be associated (e. g., *cystinuria, glycinuria,* and *Hartnup's disease*). HCO_3^-, Na^+, phophate, and sulfate are also resorbed by ATP-dependent (active) mechanisms in the proximal tubule. The later sections of the nephron mainly serve for additional water recovery and regulated resorption of Na^+ and Cl^- (see pp. 326, 328). These processes are controlled by hormones (aldosterone, vasopressin).

Secretion. Some excretable substances are released into the urine by *active transport* in the renal tubules. These substances include H^+ and K^+ ions, urea, and creatinine, as well as drugs such as penicillin.

Clearance. Renal clearance is used as a quantitative measure of renal function. It is defined as the plasma volume cleared of a given substance per unit of time. *Inulin,* a fructose polysaccharide with a mass of ca. 6 kDa (see p. 40) that is neither actively excreted nor resorbed but is freely filtered, has a clearance of 120 mL min^{-1} in healthy individuals.

Further information

Concentrating urine and transporting it through membranes are processes that require large amounts of energy. The kidneys therefore have very high energy demands. In the proximal tubule, the ATP needed is obtained from oxidative metabolism of *fatty acids, ketone bodies,* and several *amino acids.* To a lesser extent, lactate, glycerol, and citric acid are also used. In the distal tubule and Henle's loop, *glucose* is the main substrate for the energy metabolism. The endothelial cells in the proximal tubule are also capable of *gluconeogenesis.* The substrates for this are mainly the carbohydrate skeletons of amino acids. Their amino groups are used as ammonia for buffering urine (see p. 311). Enzymes for peptide degradation and the amino acid metabolism occur in the kidneys at high levels of activity (e. g., amino acid oxidases, amine oxidases, glutaminase).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Functions of the kidneys

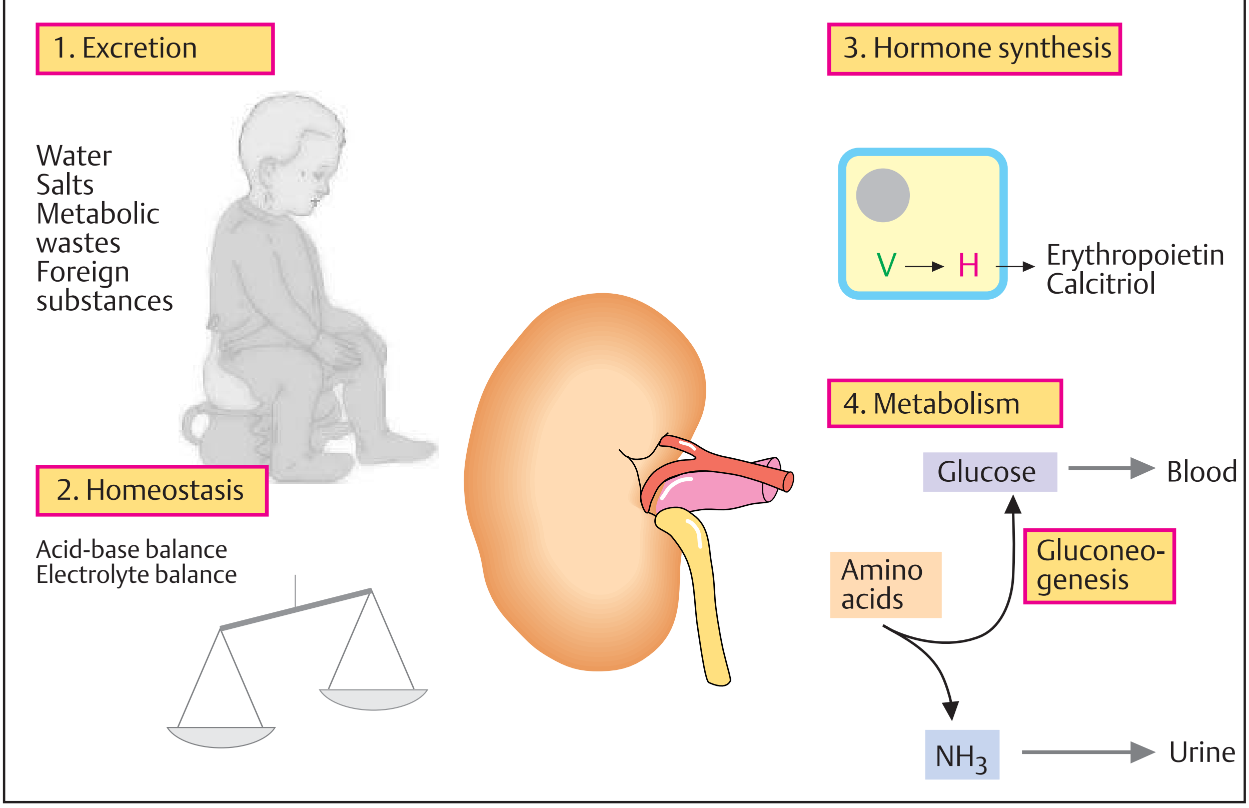
1. Excretion
Water
Salts
Metabolic wastes
Foreign substances
2. Homeostasis
Acid-base balance
Electrolyte balance
3. Hormone synthesis
V → H
Erythropoietin
Calcitriol
4. Metabolism
Glucose
Blood
Amino acids
Gluconeo-genesis
NH_3
Urine

B. Urine formation

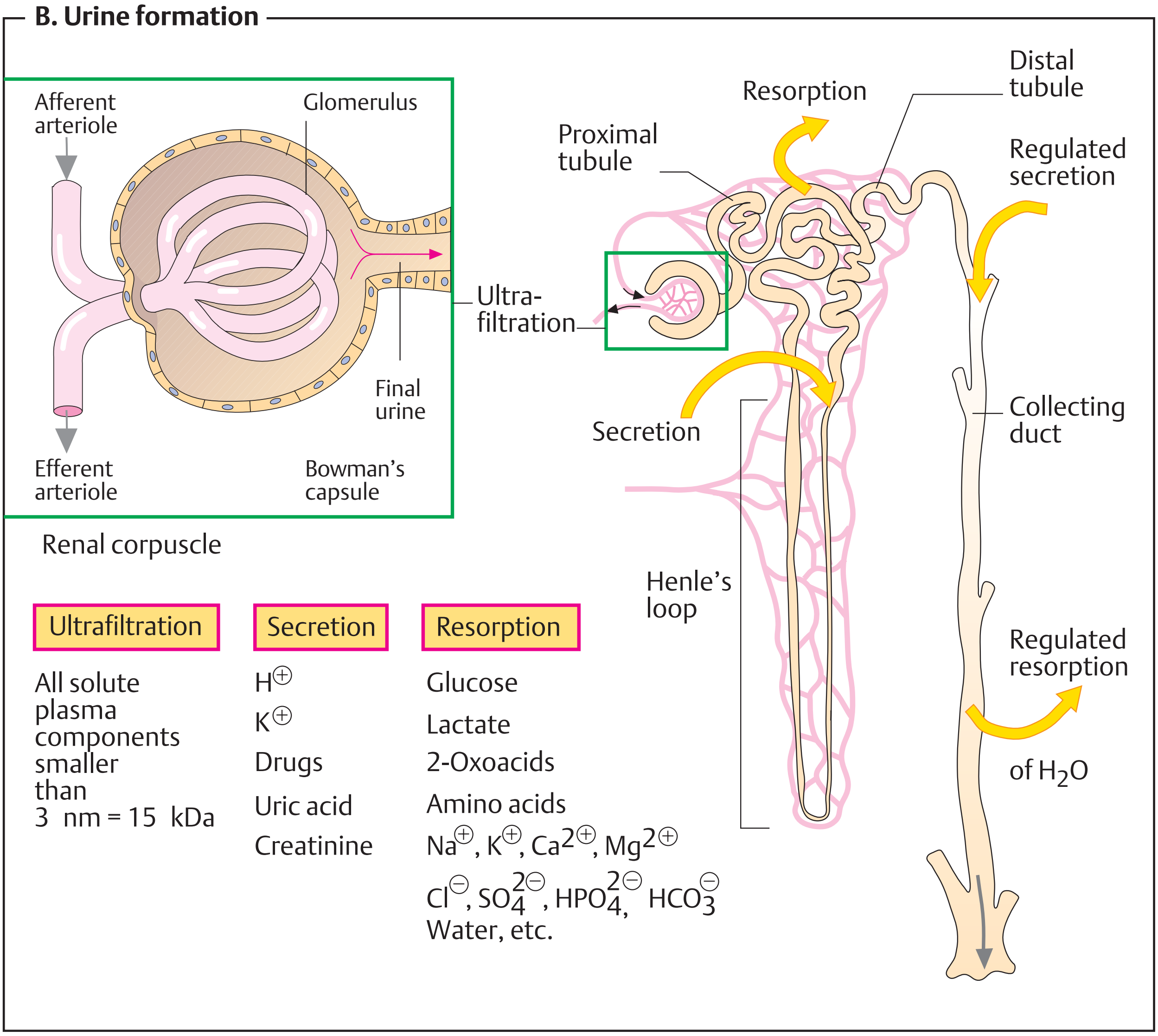
Afferent arteriole
Glomerulus
Efferent arteriole
Final urine
Bowman's capsule
Renal corpuscle
Proximal tubule
Resorption
Distal tubule
Regulated secretion
Ultra-filtration
Secretion
Collecting duct
Henle's loop
Regulated resorption
of H_2O
Ultrafiltration
All solute plasma components smaller than 3 nm = 15 kDa
Secretion
$H^{\oplus}$
$K^{\oplus}$
Drugs
Uric acid
Creatinine
Resorption
Glucose
Lactate
2-Oxoacids
Amino acids
$Na^{\oplus}$, $K^{\oplus}$, $Ca^{2\oplus}$, $Mg^{2\oplus}$
$Cl^{\ominus}$, $SO_4^{2\ominus}$, $HPO_4^{2\ominus}$, $HCO_3^{\ominus}$
Water, etc.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Urine

A. Urine ◑

Water and water-soluble compounds are excreted with the urine. The volume and composition of urine are subject to wide variation and depend on food intake, body weight, age, sex, and living conditions such as temperature, humidity, physical activity, and health status. As there is a marked circadian rhythm in urine excretion, the amount of urine and its composition are usually given relative to a 24-hour period.

A human adult produces 0.5–2.0 L urine per day, around 95% of which consists of water. The urine usually has a slightly acidic pH value (around 5.8). However, the pH value of urine is strongly affected by metabolic status. After ingestion of large amounts of plant food, it can increase to over 7.

B. Organic components ◑

Nitrogen-containing compounds are among the most important organic components of urine. **Urea,** which is mainly synthesized in the liver (urea cycle; see p. 182), is the form in which nitrogen atoms from amino acids are excreted. Breakdown of pyrimidine bases also produces a certain amount of urea (see p. 190). When the nitrogen balance is constant, as much nitrogen is excreted as is taken up (see p. 174), and the amount of urea in the urine therefore reflects protein degradation: 70 g protein in food yields approximately 30 g urea in the urine.

Uric acid is the end product of the purine metabolism. When uric acid excretion via the kidneys is disturbed, gout can develop (see p. 190). **Creatinine** is derived from the muscle metabolism, where it arises spontaneously and irreversibly by cyclization of creatine and creatine phosphate (see p. 336). Since the amount of creatinine an individual excretes per day is constant (it is directly proportional to muscle mass), creatinine as an endogenous substance can be used to measure the glomerular filtration rate. The amount of **amino acids** excreted in free form is strongly dependent on the diet and on the efficiency of liver function. Amino acid derivatives are also found in the urine (e. g., **hippurate**, a detoxification product of benzoic acid). Modified amino acids, which occur in special proteins such as *hydroxyproline* in collagen and *3-methylhistidine* in actin and myosin, can be used as indicators of the degradation of these proteins.

Other components of the urine are conjugates with sulfuric acid, glucuronic acid, glycine, and other polar compounds that are synthesized in the liver by biotransformation (see p. 316). In addition, metabolites of many hormones (catecholamines, steroids, serotonin) also appear in the urine and can provide information about hormone production. The proteohormone *chorionic gonadotropin* (hCG, mass ca. 36 kDa), which is formed at the onset of pregnancy, appears in the urine due to its relatively small size. Evidence of hCG in the urine provides the basis for an immunological *pregnancy test.*

The yellow color of urine is due to *urochromes,* which are related to the bile pigments produced by hemoglobin degradation (see p. 194). If urine is left to stand long enough, oxidation of the urochromes may lead to a darkening in color.

C. Inorganic components ◑

The main inorganic components of the urine are the *cations* Na^+, K^+, Ca^{2+}, Mg^{2+}, and NH_4^+ and the *anions* Cl^-, SO_4^{2-}, and HPO_4^{2-}, as well as traces of other ions. In total, Na^+ and Cl^- represent about two-thirds of all the electrolytes in the final urine. Calcium and magnesium occur in the feces in even larger quantities. The amounts of the various inorganic components of the urine also depend on the composition of the diet. For example, in acidosis there can be a marked increase in the excretion of ammonia (see p. 326). Excretion of Na^+, K^+, Ca^{2+}, and phosphate via the kidneys is subject to hormonal regulation (see p. 330).

Further information

Shifts in the concentrations of the *physiological components* of the urine and the appearance of *pathological urine components* can be used to diagnose diseases. Important examples are glucose and ketone bodies, which are excreted to a greater extent in *diabetes mellitus* (see p. 160).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Urine

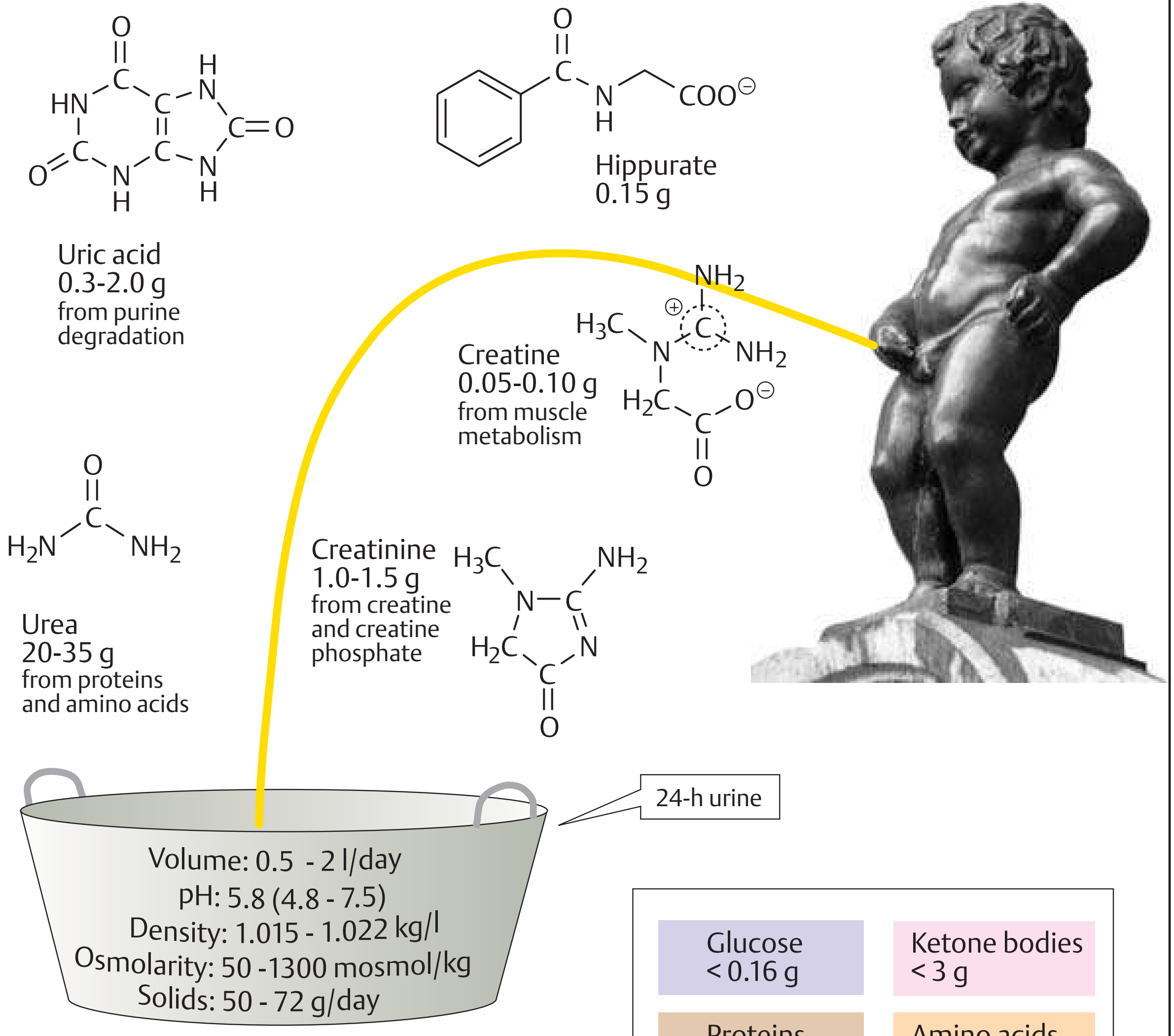
Uric acid
0.3-2.0 g
from purine degradation
Hippurate
0.15 g
Creatine
0.05-0.10 g
from muscle metabolism
Urea
20-35 g
from proteins and amino acids
Creatinine
1.0-1.5 g
from creatine and creatine phosphate
24-h urine
Volume: 0.5 - 2 l/day
pH: 5.8 (4.8 - 7.5)
Density: 1.015 - 1.022 kg/l
Osmolarity: 50 -1300 mosmol/kg
Solids: 50 - 72 g/day

Glucose < 0.16 g	Ketone bodies < 3 g
Proteins < 0.15 g	Amino acids 1-3 g

B. Organic constituents

C. Inorganic constituents

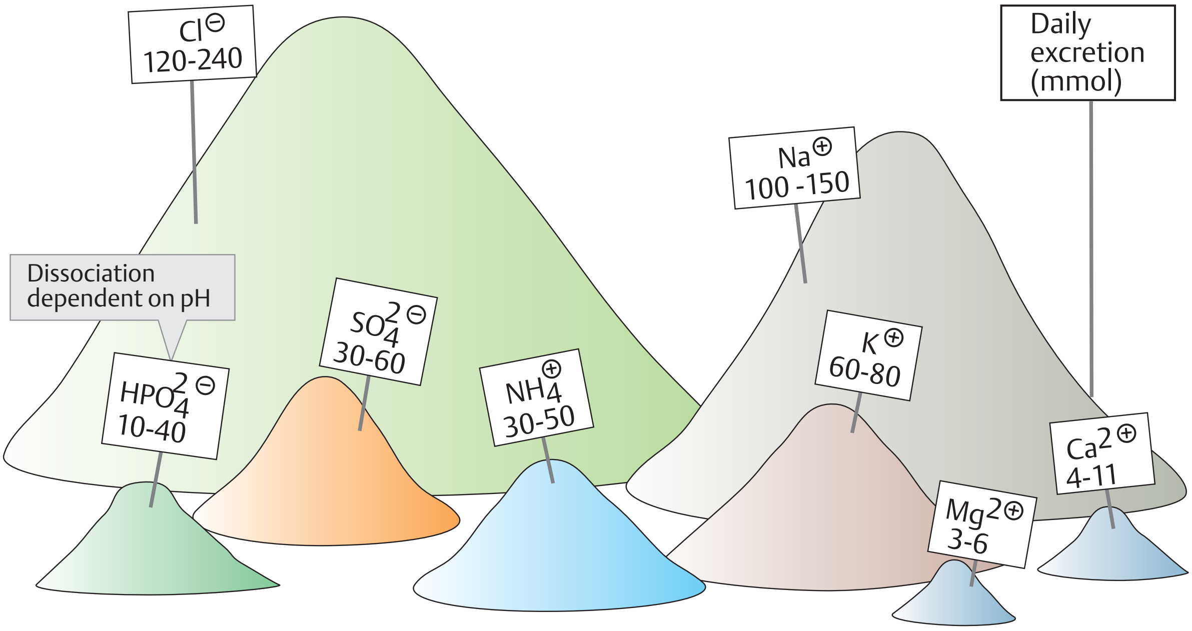
Cl⊖
120-240
Daily excretion (mmol)
Na⊕
100 -150
Dissociation dependent on pH
HPO$_4^{2\ominus}$
10-40
SO$_4^{2\ominus}$
30-60
NH$_4^{\oplus}$
30-50
K⊕
60-80
Ca$^{2\oplus}$
4-11
Mg$^{2\oplus}$
3-6

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Functions in the acid–base balance

Along with the lungs, the kidneys are particularly involved in keeping the pH value of the extracellular fluid constant (see p. 288). The contribution made by the kidneys particularly involves resorbing HCO_3^- and actively excreting protons.

A. Proton excretion ◑

The renal tubule cells are capable of secreting protons (H^+) from the blood into the urine against a concentration gradient, despite the fact that the H^+ concentration in the urine is up to a thousand times higher than in the blood. To achieve this, carbon dioxide (CO_2) is taken up from the blood and—together with water (H_2O) and with the help of *carbonate dehydratase* (carbonic anhydrase, [1])—converted into hydrogen carbonate ("bicarbonate," HCO_3^-) and one H^+. Formally, this yields carbonic acid H_2CO_3 as an intermediate, but it is not released during the reaction.

The hydrogen carbonate formed in carbonic anhydrase returns to the plasma, where it contributes to the blood's base reserve. The proton is exported into the urine by *secondary active transport* in antiport for Na^+ (bottom right). The driving force for **proton excretion**, as in other secondary active processes, is the Na^+ gradient established by the ATPase involved in the *Na^+/K^+* exchange ("Na^+/K^+ ATPase", see p. 220). This integral membrane protein on the basal side (towards the blood) of tubule cells keeps the Na^+ concentration in the tubule cell low, thereby maintaining Na^+ inflow. In addition to this secondary active H^+ transport mechanism, there is a V-type H^+-transporting ATPase in the distal tubule and collecting duct (see p. 220).

An important function of the secreted H^+ ions is to promote **HCO_3^- resorption** (top right). Hydrogen carbonate, the most important buffering base in the blood, passes into the primary urine quantitatively, like all ions. In the primary urine, HCO_3^- reacts with H^+ ions to form water and CO_2, which returns by free diffusion to the tubule cells and from there into the blood. In this way, the kidneys also influence the CO_2/HCO_3^- buffering balance in the plasma.

B. Ammonia excretion ◑

Approximately 60 mmol of protons are excreted with the urine every day. Buffering systems in the urine catch a large proportion of the H^+ ions, so that the urine only becomes weakly acidic (down to about pH 4.8).

An important buffer in the urine is the *hydrogen phosphate/dihydrogen phosphate system* ($HPO_4^{2-}/H_2PO_4^-$). In addition, *ammonia* also makes a vital contribution to buffering the secreted protons.

Since plasma concentrations of *free* ammonia are low, the kidneys release NH_3 from **glutamine** and other amino acids. At 0.5–0.7 mM, glutamine is the most important amino acid in the plasma and is the preferred form for ammonia transport in the blood. The kidneys take up glutamine, and with the help of *glutaminase* [4], initially release NH_3 from the amide bond hydrolytically. From the **glutamate** formed, a second molecule of NH_3 can be obtained by oxidative deamination with the help of *glutamate dehydrogenase* [5] (see p. 178). The resulting **2-oxoglutarate** is further metabolized in the tricarboxylic acid cycle. Several other amino acids—*alanine* in particular, as well as *serine, glycine,* and *aspartate*—can also serve as suppliers of ammonia.

Ammonia can diffuse freely into the urine through the tubule membrane, while the **ammonium ions** that are formed in the urine are charged and can no longer return to the cell. Acidic urine therefore promotes ammonia excretion, which is normally 30–50 mmol per day. In metabolic **acidosis** (e. g., during fasting or in *diabetes mellitus*), after a certain time increased *induction of glutaminase* occurs in the kidneys, resulting in increased NH_3 excretion. This in turn promotes H^+ release and thus counteracts the acidosis. By contrast, when the plasma pH value shifts towards alkaline values (*alkalosis*), renal excretion of ammonia is reduced.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Proton secretion

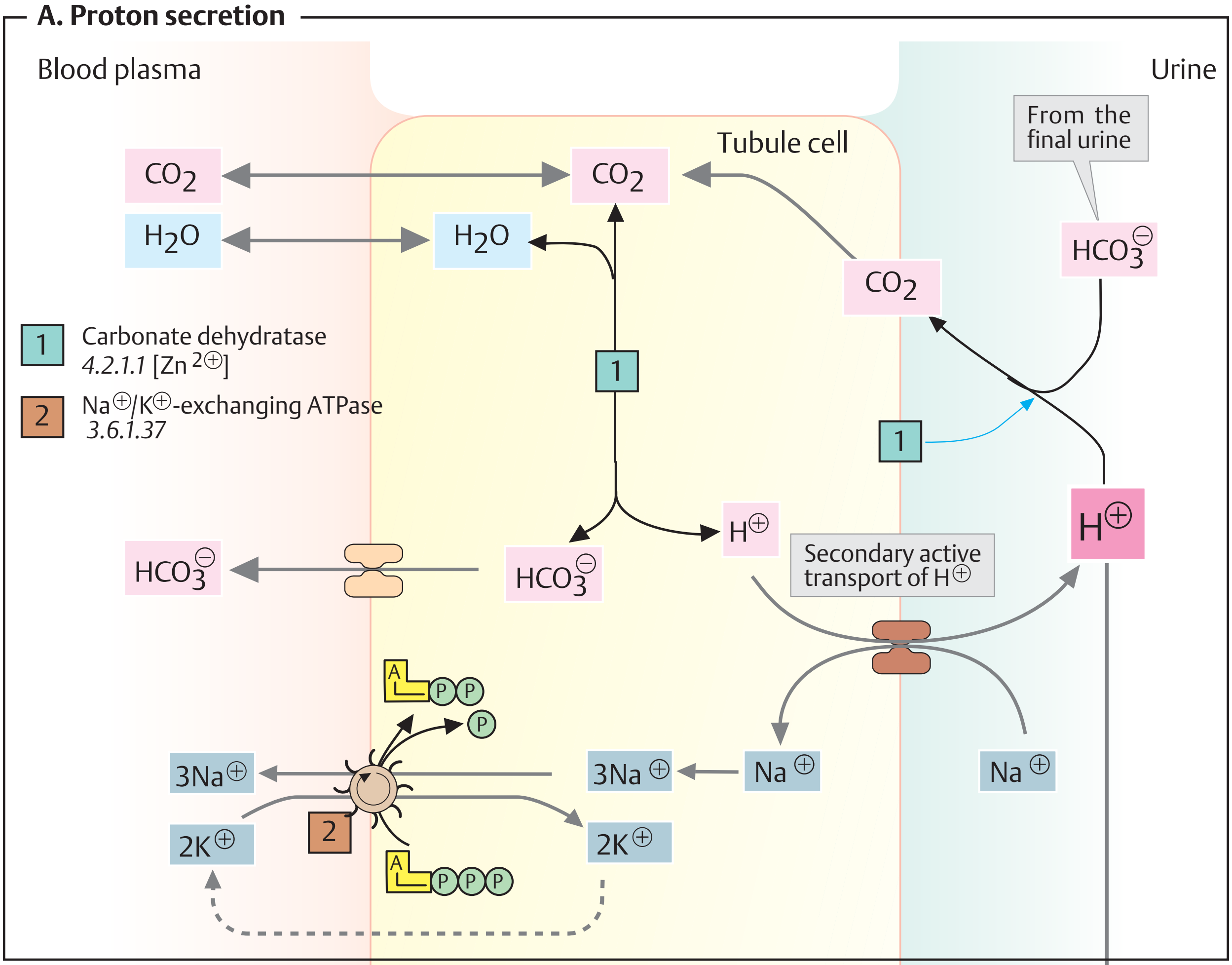
Blood plasma
Tubule cell
Urine
From the final urine
CO_2
CO_2
CO_2
H_2O
H_2O
HCO_3^-
HCO_3^-
HCO_3^-
1
1
H^+
H^+
Secondary active transport of H^+
1 Carbonate dehydratase 4.2.1.1 [Zn^{2+}]
2 Na^+/K^+-exchanging ATPase 3.6.1.37
A P P
P
A P P P
$3Na^+$
$3Na^+$
Na^+
Na^+
$2K^+$
$2K^+$
2

B. Ammonia excretion

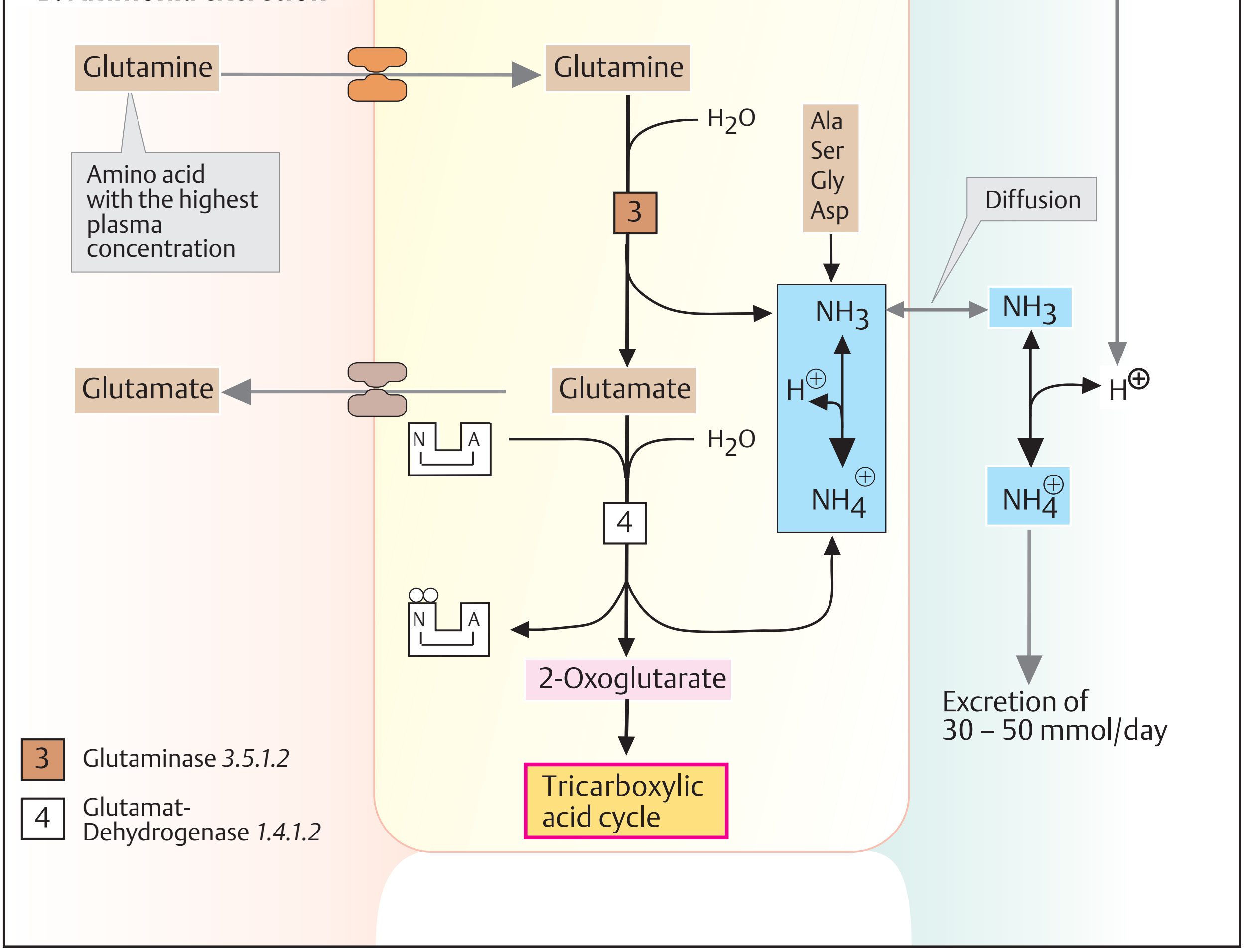
Glutamine
Glutamine
Amino acid with the highest plasma concentration
H_2O
3
Ala Ser Gly Asp
Diffusion
NH_3
NH_3
H^+
H^+
NH_4^+
NH_4^+
Glutamate
Glutamate
N A
H_2O
4
N A
2-Oxoglutarate
Tricarboxylic acid cycle
Excretion of 30 – 50 mmol/day
3 Glutaminase 3.5.1.2
4 Glutamat-Dehydrogenase 1.4.1.2

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Electrolyte and water recycling

A. Electrolyte and water recycling ◑

Electrolytes and other plasma components with low molecular weights enter the *primary urine* by ultrafiltration (right). Most of these substances are recovered by energy-dependent resorption (see p. 322). The extent of the resorption determines the amount that ultimately reaches the final urine and is excreted. The illustration does not take into account the *zoning* of transport processes in the kidney (physiology textbooks may be referred to for further details).

Calcium and phosphate ions. Calcium (Ca^{2+}) and phosphate ions are almost completely resorbed from the primary urine by active transport (i. e., in an ATP-dependent fashion). The proportion of Ca^{2+} resorbed is over 99%, while for phosphate the figure is 80–90%. The extent to which these two electrolytes are resorbed is regulated by the three hormones parathyrin, calcitonin, and calcitriol.

The peptide hormone *parathyrin* (PTH), which is produced by the parathyroid gland, stimulates Ca^{2+} resorption in the kidneys and at the same time inhibits the resorption of phosphate. In conjunction with the effects of this hormone in the bones and intestines (see p. 344), this leads to an *increase in the plasma level of* Ca^{2+} and a reduction in the level of phosphate ions.

Calcitonin, a peptide produced in the C cells of the thyroid gland, inhibits the resorption of both calcium and phosphate ions. The result is an overall *reduction in the plasma level* of both ions. Calcitonin is thus a parathyrin antagonist relative to Ca^{2+}.

The steroid hormone *calcitriol,* which is formed in the kidneys (see p. 304), stimulates the resorption of both calcium and phosphate ions and thus increases the plasma level of both ions.

Sodium ions. Controlled resorption of Na^+ from the primary urine is one of the most important functions of the kidney. Na^+ resorption is highly effective, with more than 97% being resorbed. Several mechanisms are involved: some of the Na^+ is taken up passively in the proximal tubule through the junctions between the cells (paracellularly). In addition, there is secondary active transport together with glucose and amino acids (see p. 322). These two pathways are responsible for 60–70% of total Na^+ resorption. In the ascending part of Henle's loop, there is another transporter (shown at the bottom right), which functions electroneutrally and takes up one Na^+ ion and one K^+ ion together with two Cl^- ions. This symport is also dependent on the activity of Na^+/K^+ ATPase [2], which pumps the Na^+ resorbed from the primary urine back into the plasma in exchange for K^+.

The steroid hormone *aldosterone* (see p. 55) increases Na^+ reuptake, particularly in the distal tubule, while *atrial natriuretic peptide* (ANP) originating from the cardiac atrium reduces it. Among other effects, aldosterone induces Na^+/K^+ ATPase and various Na^+ transporters on the luminal side of the cells.

Water. Water resorption in the proximal tubule is a *passive process* in which water follows the osmotically active particles, particularly the Na^+ ions. Fine regulation of water excretion (**diuresis**) takes place in the collecting ducts, where the peptide hormone *vasopressin* (antidiuretic hormone, ADH) operates. This promotes recovery of water by stimulating the transfer of aquaporins (see p. 220) into the plasma membrane of the tubule cells via V_2 receptors. A lack of ADH leads to the disease picture of *diabetes insipidus,* in which up to 30 L of final urine is produced per day.

B. Gluconeogenesis ◑

Apart from the liver, the kidneys are the only organs capable of producing glucose by neosynthesis *(gluconeogenesis;* see p.154). The main substrate for gluconeogenesis in the cells of the proximal tubule is **glutamine**. In addition, other amino acids and also **lactate**, **glycerol**, and **fructose** can be used as precursors. As in the liver, the key enzymes for gluconeogenesis are induced by *cortisol* (see p. 374). Since the kidneys also have a high level of glucose consumption, they only release very little glucose into the blood.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Electrolyte and water recycling

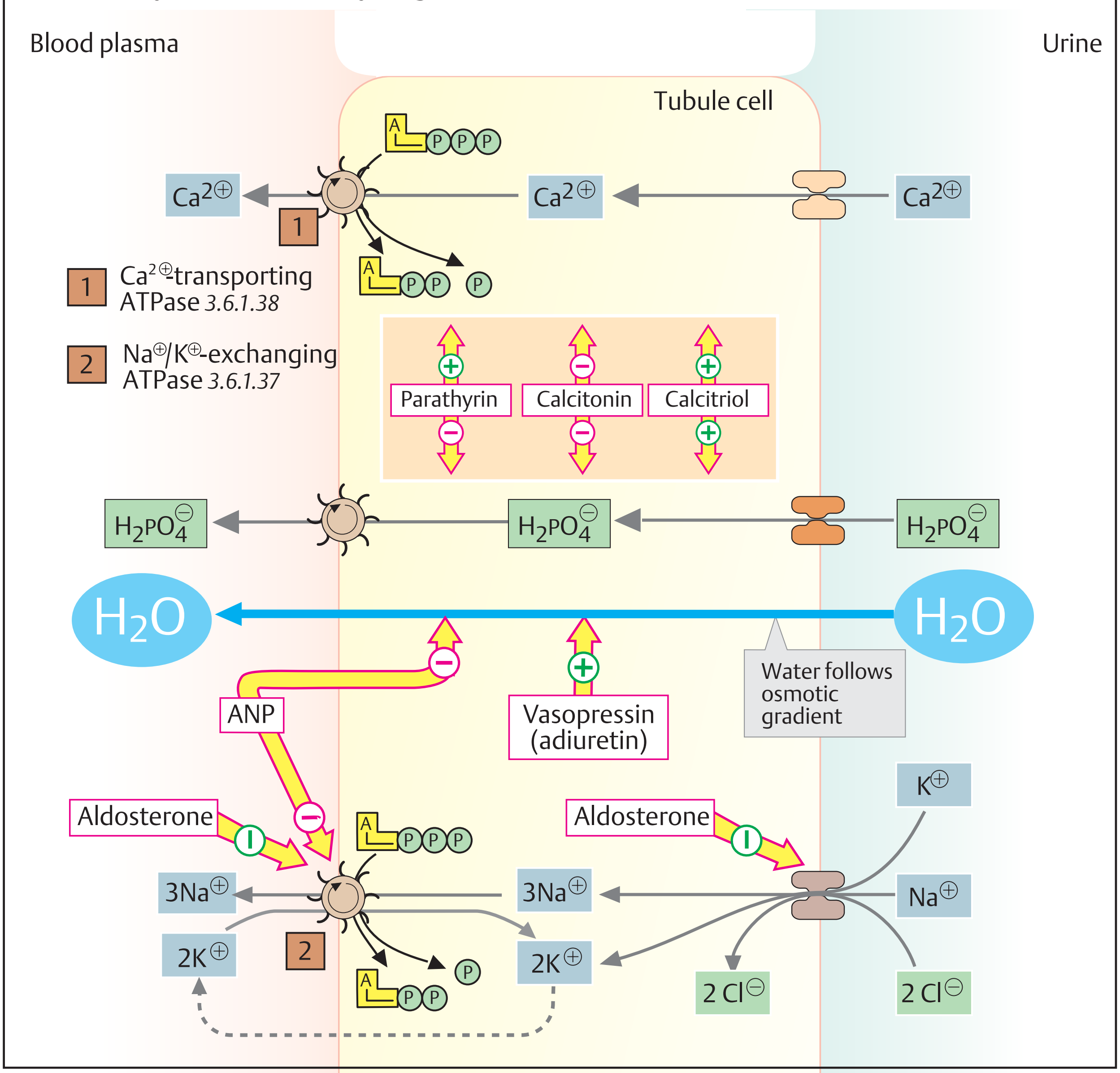
Blood plasma
Urine
Tubule cell
$Ca^{2\oplus}$
$Ca^{2\oplus}$
$Ca^{2\oplus}$
1
1 $Ca^{2\oplus}$-transporting ATPase 3.6.1.38
2 $Na^{\oplus}/K^{\oplus}$-exchanging ATPase 3.6.1.37
A P P P
A P P P
Parathyrin
Calcitonin
Calcitriol
$H_2PO_4^{\ominus}$
$H_2PO_4^{\ominus}$
$H_2PO_4^{\ominus}$
H_2O
H_2O
Water follows osmotic gradient
ANP
Vasopressin (adiuretin)
Aldosterone
Aldosterone
A P P P
$K^{\oplus}$
$3Na^{\oplus}$
$3Na^{\oplus}$
$Na^{\oplus}$
$2K^{\oplus}$
$2K^{\oplus}$
2
A P P P
2 $Cl^{\ominus}$
2 $Cl^{\ominus}$

B. Gluconeogenesis

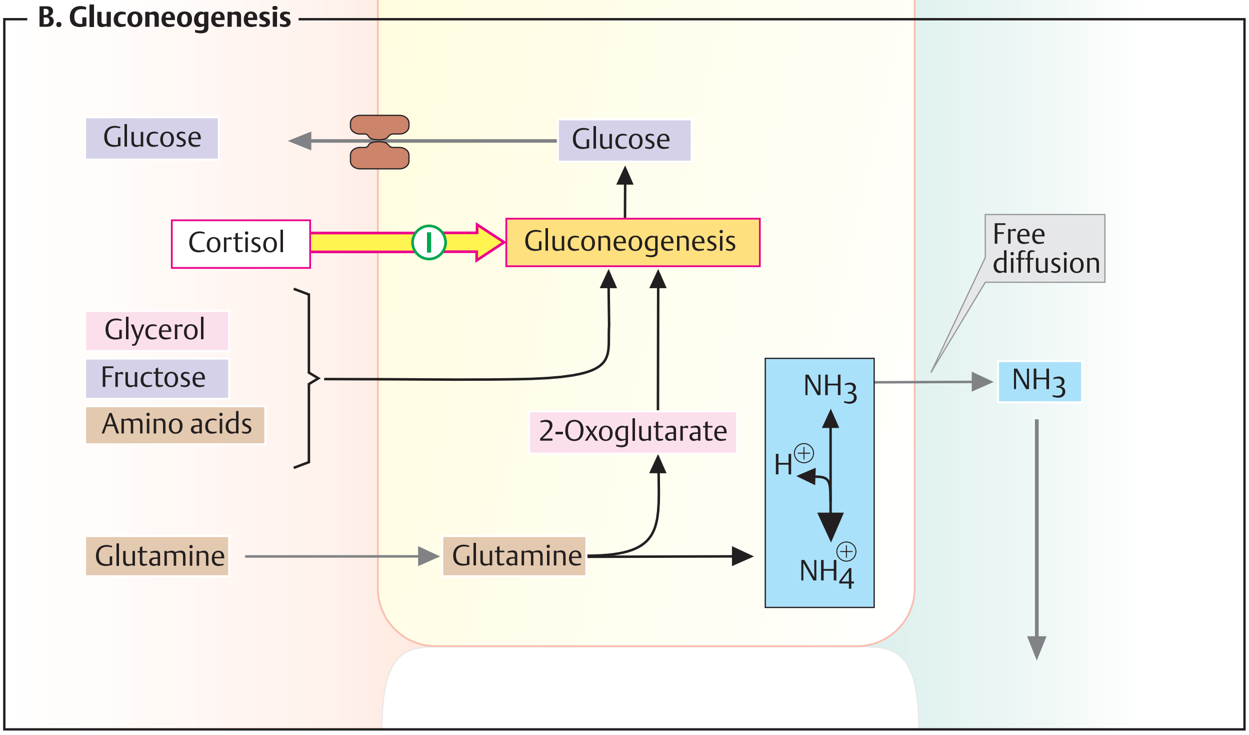
Glucose
Glucose
Cortisol
Gluconeogenesis
Free diffusion
Glycerol
Fructose
Amino acids
NH_3
NH_3
$H^{\oplus}$
2-Oxoglutarate
$NH_4^{\oplus}$
Glutamine
Glutamine

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Renal hormones

A. Renal hormones ◑

In addition to their involvement in excretion and metabolism, the kidneys also have endocrine functions. They produce the hormones **erythropoietin** and **calcitriol** and play a decisive part in producing the hormone **angiotensin II** by releasing the enzyme *renin.* Renal prostaglandins (see p. 390) have a local effect on Na^+ resorption.

Calcitriol (vitamin D hormone, 1α,25-dihydroxycholecalciferol) is a hormone closely related to the steroids that is involved in Ca^{2+} homeostasis (see p. 342). In the kidney, it is formed from calcidiol by hydroxylation at C-1. The activity of *calcidiol-1-monooxygenase* [1] is enhanced by the hormone *parathyrin* (PTH).

Erythropoietin is a peptide hormone that is formed predominantly by the kidneys, but also by the liver. Together with other factors known as *"colony-stimulating factors"* (CSF; see p. 392), it regulates the differentiation of stem cells in the bone marrow.

Erythropoietin release is stimulated by hypoxia (low pO_2). Within hours, the hormone ensures that erythrocyte precursor cells in the bone marrow are converted to erythrocytes, so that their numbers in the blood increase. Renal damage leads to reduced erythropoietin release, which in turn results in *anemia.* Forms of anemia with renal causes can now be successfully treated using erythropoietin produced by genetic engineering techniques. The hormone is also administered to dialysis patients. Among athletes and sports professionals, there have been repeated cases of erythropoietin being misused for doping purposes.

B. Renin–angiotensin system ◑

The peptide hormone angiotensin II is not synthesized in a hormonal gland, but in the blood. The kidneys take part in this process by releasing the enzyme renin.

Renin [2] is an aspartate proteinase (see p. 176). It is formed by the kidneys as a precursor (prorenin), which is proteolytically activated into renin and released into the blood. In the blood plasma, renin acts on **angiotensinogen,** a plasma glycoprotein in the α_2-globulin group (see p. 276), which like almost all plasma proteins is synthesized in the liver. The decapeptide cleaved off by renin is called **angiotensin I.** Further cleavage by *peptidyl dipeptidase A* (*angiotensin-converting enzyme,* ACE), a membrane enzyme located on the vascular endothelium in the lungs and other tissues, gives rise to the octapeptide **angiotensin II** [3], which acts as a hormone and neurotransmitter. The lifespan of angiotensin II in the plasma is only a few minutes, as it is rapidly broken down by other peptidases (angiotensinases [4]), which occur in many different tissues.

The plasma level of angiotensin II is mainly determined by the rate at which renin is released by the kidneys. Renin is synthesized by juxtaglomerular cells, which release it when sodium levels decline or there is a fall in blood pressure.

Effects of angiotensin II. Angiotensin II has effects on the kidneys, brain stem, pituitary gland, adrenal cortex, blood vessel walls, and heart via membrane-located receptors. It increases blood pressure by triggering *vasoconstriction* (narrowing of the blood vessels). In the kidneys, it promotes the *retention of* Na^+ and water and reduces potassium secretion. In the brain stem and at nerve endings in the sympathetic nervous system, the effects of angiotensin II lead to increased tonicity (neurotransmitter effect). In addition, it triggers the *sensation of thirst.* In the pituitary gland, angiotensin II stimulates *vasopressin release* (antidiuretic hormone) and corticotropin (ACTH) release. In the adrenal cortex, it increases the *biosynthesis and release of aldosterone,* which promotes sodium and water retention in the kidneys. All of the effects of angiotensin II lead directly or indirectly to *increased blood pressure,* as well as increased *sodium and water retention.* This important hormonal system for blood pressure regulation can be pharmacologically influenced by *inhibitors* at various points:

- Using angiotensinogen analogs that inhibit renin.
- Using angiotensin I analogs that competitively inhibit the enzyme ACE [3].
- Using hormone antagonists that block the binding of angiotensin II to its receptors.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Renal hormones

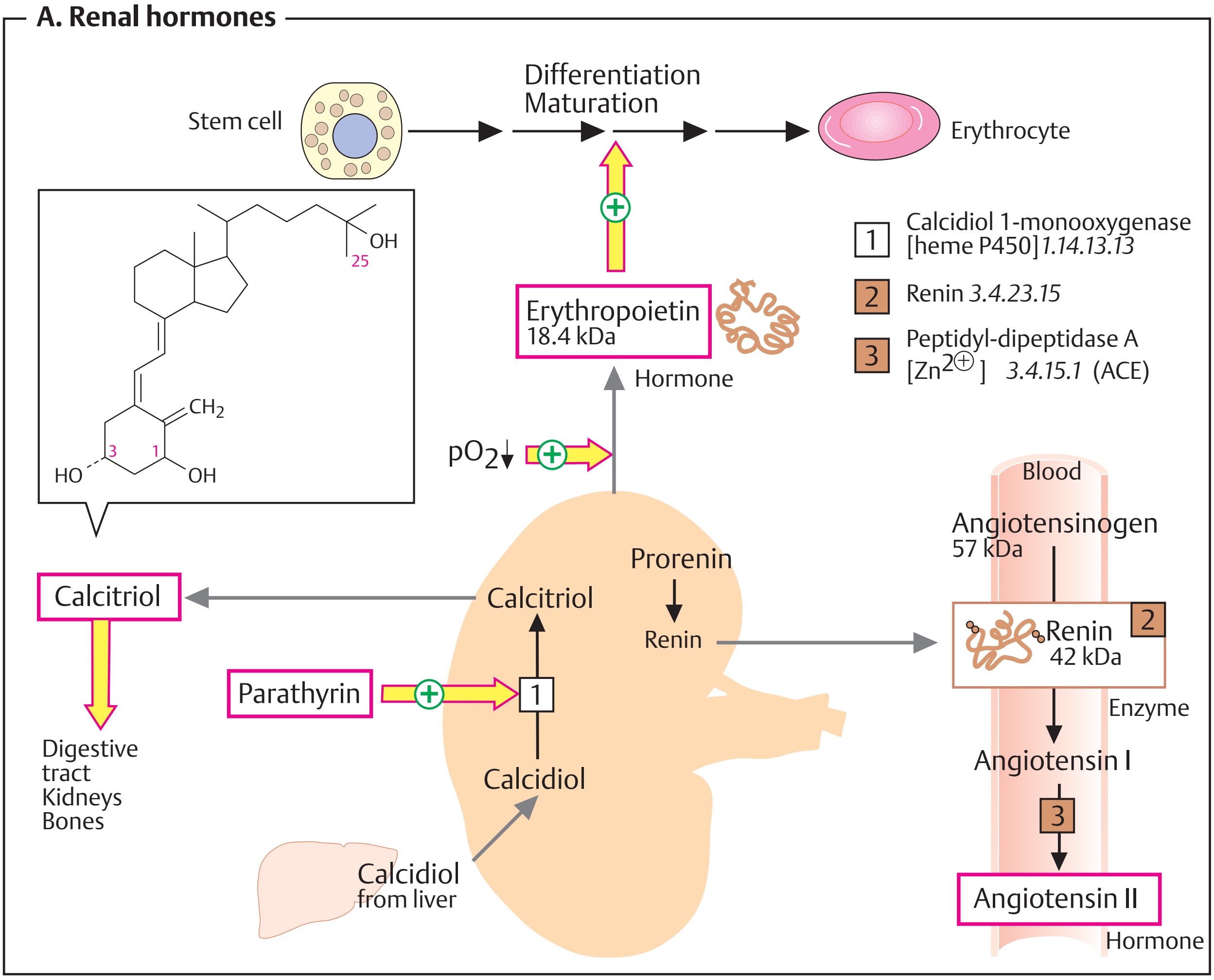
Stem cell
Differentiation
Maturation
Erythrocyte
OH
25
CH_2
HO
3
1
OH
Erythropoietin
18.4 kDa
Hormone
1 Calcidiol 1-monooxygenase [heme P450] *1.14.13.13*
2 Renin *3.4.23.15*
3 Peptidyl-dipeptidase A [Zn^{2+}] *3.4.15.1* (ACE)
pO_2↓
Blood
Angiotensinogen
57 kDa
Prorenin
Renin
Renin
42 kDa
2
Enzyme
Calcitriol
Calcitriol
Parathyrin
1
Digestive
tract
Kidneys
Bones
Calcidiol
Angiotensin I
3
Calcidiol
from liver
Angiotensin II
Hormone

B. Renin angiotensin system

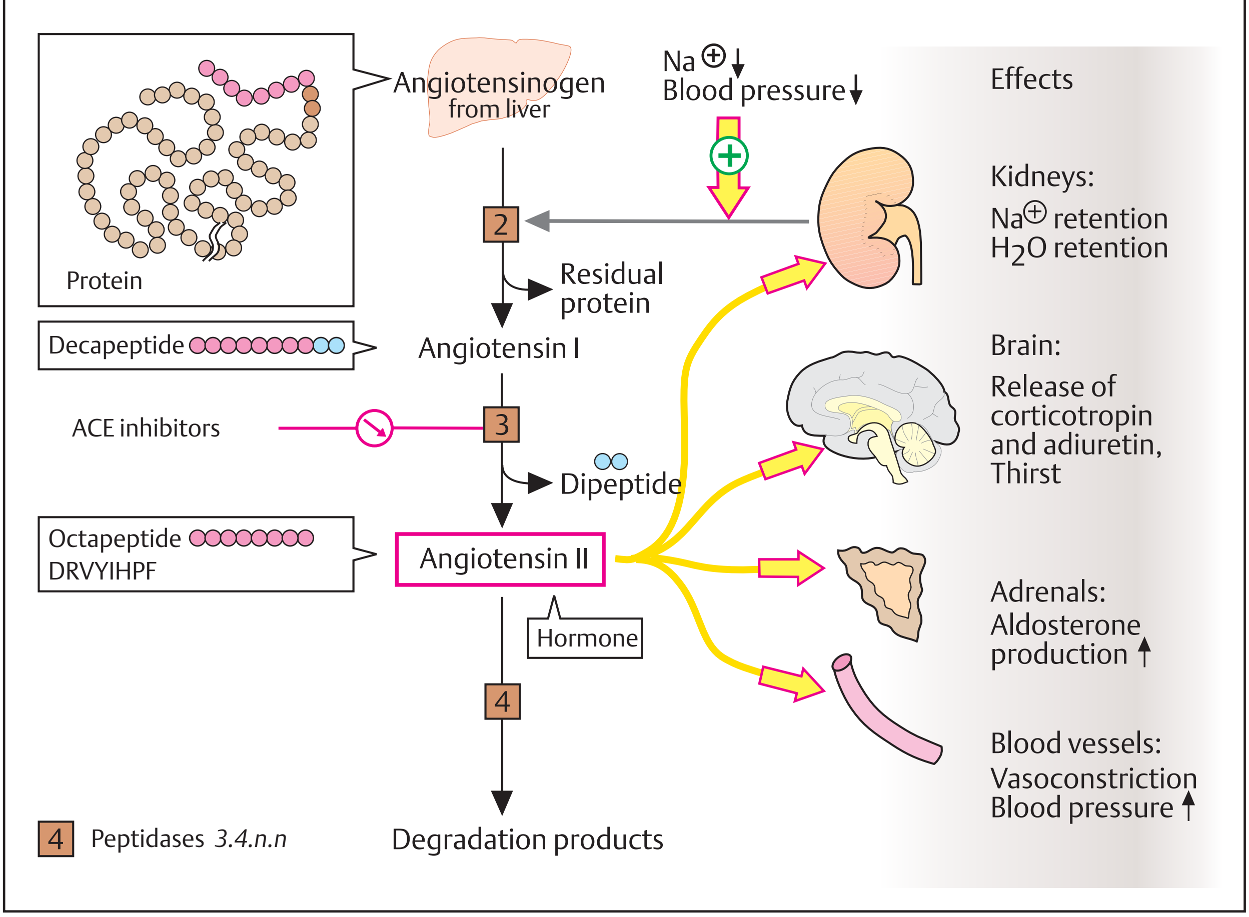
Protein
Angiotensinogen
from liver
Na^+↓
Blood pressure↓
Effects
Kidneys:
Na^+ retention
H_2O retention
2
Residual
protein
Decapeptide
Angiotensin I
Brain:
Release of
corticotropin
and adiuretin,
Thirst
ACE inhibitors
3
Dipeptide
Octapeptide
DRVYIHPF
Angiotensin II
Adrenals:
Aldosterone
production ↑
Hormone
4
Blood vessels:
Vasoconstriction
Blood pressure ↑
4 Peptidases *3.4.n.n*
Degradation products

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Muscle contraction

The musculature is what makes movements possible. In addition to the **skeletal muscles**, which can be contracted voluntarily, there are also the autonomically activated **heart muscle** and **smooth muscle**, which is also involuntary. In all types of muscle, contraction is based on an interplay between the proteins actin and myosin.

A. Organization of skeletal muscle ◑

Striated muscle consists of parallel bundles of **muscle fibers.** Each fiber is a single large multinucleate cell. The cytoplasm in these cells contains **myofibrils** 2–3 μm thick that can extend over the full length of the muscle fiber.

The *striation* of the muscle fibers is characteristic of skeletal muscle. It results from the regular arrangement of molecules of differing density. The repeating contractile units, the **sarcomeres**, are bounded by Z lines from which thin filaments of **F-actin** (see p. 204) extend on each side. In the A bands, there are also thick parallel filaments of **myosin**. The H bands in the middle of the A bands only contain myosin, while only actin is found on each size of the Z lines.

Myosin is quantitatively the most important protein in the myofibrils, representing 65% of the total. It is shaped like a golf club (bottom right). The molecule is a hexamer consisting of two identical *heavy chains* (2 × 223 kDa) and four *light chains* (each about 20 kDa). Each of the two heavy chains has a globular "head" at its amino end, which extends into a "tail" about 150 nm long in which the two chains are intertwined to form a superhelix. The small subunits are attached in the head area. Myosin is present as a bundle of several hundred stacked molecules in the form of a *"thick myosin filament."* The head portion of the molecule acts as an *ATPase,* the activity of which is modulated by the small subunits.

Actin (42 kDa) is the most important component of the *"thin filaments."* It represents ca. 20–25% of the muscle proteins. **F-actin** is also an important component of the cytoskeleton (see p. 204). This filamentous polymer is held in equilibrium with its monomer, **G-actin.** The other protein components of muscle include tropomyosin and troponin. **Tropomyosin** (64 kDa) attaches to F-actin as a rod-like dimer and connects approximately seven actin units with each other. The heterotrimer **troponin** (78 kDa) is bound to one end of tropomyosin.

In addition to the above proteins, a number of other proteins are also typical of muscle—including *titin* (the largest known protein), *α-* and *β-actinin, desmin,* and *vimentin.*

B. Mechanism of muscle contraction ◑

The *sliding filament model* describes the mechanism involved in muscle contraction. In this model, sarcomeres become shorter when the thin and thick filaments slide alongside each other and telescope together, with ATP being consumed. During contraction, the following reaction cycle is repeated several times:

[**1**] In the initial state, the myosin heads are attached to actin. When ATP is bound, the heads detach themselves from the actin (the "plasticizing" effect of ATP).

[**2**] The myosin head hydrolyzes the bound ATP to ADP and P_i, but initially withholds the two reaction products. ATP cleavage leads to allosteric tension in the myosin head.

[**3**] The myosin head now forms a new bond with a neighboring actin molecule.

[**4**] The actin causes the release of the P_i, and shortly afterwards release of the ADP as well. This converts the allosteric tension in the myosin head into a conformational change that acts like a rowing stroke.

The cycle can be repeated for as long as ATP is available, so that the thick filaments are constantly moving along the thin filaments in the direction of the Z disk. Each rowing stroke of the 500 or so myosin heads in a thick filament produces a contraction of about 10 nm. During strong contraction, the process is repeated about five times per second. This leads to the whole complex of thin filaments moving together; the H band becomes shorter and the Z lines slide closer together.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Organization of striated muscle

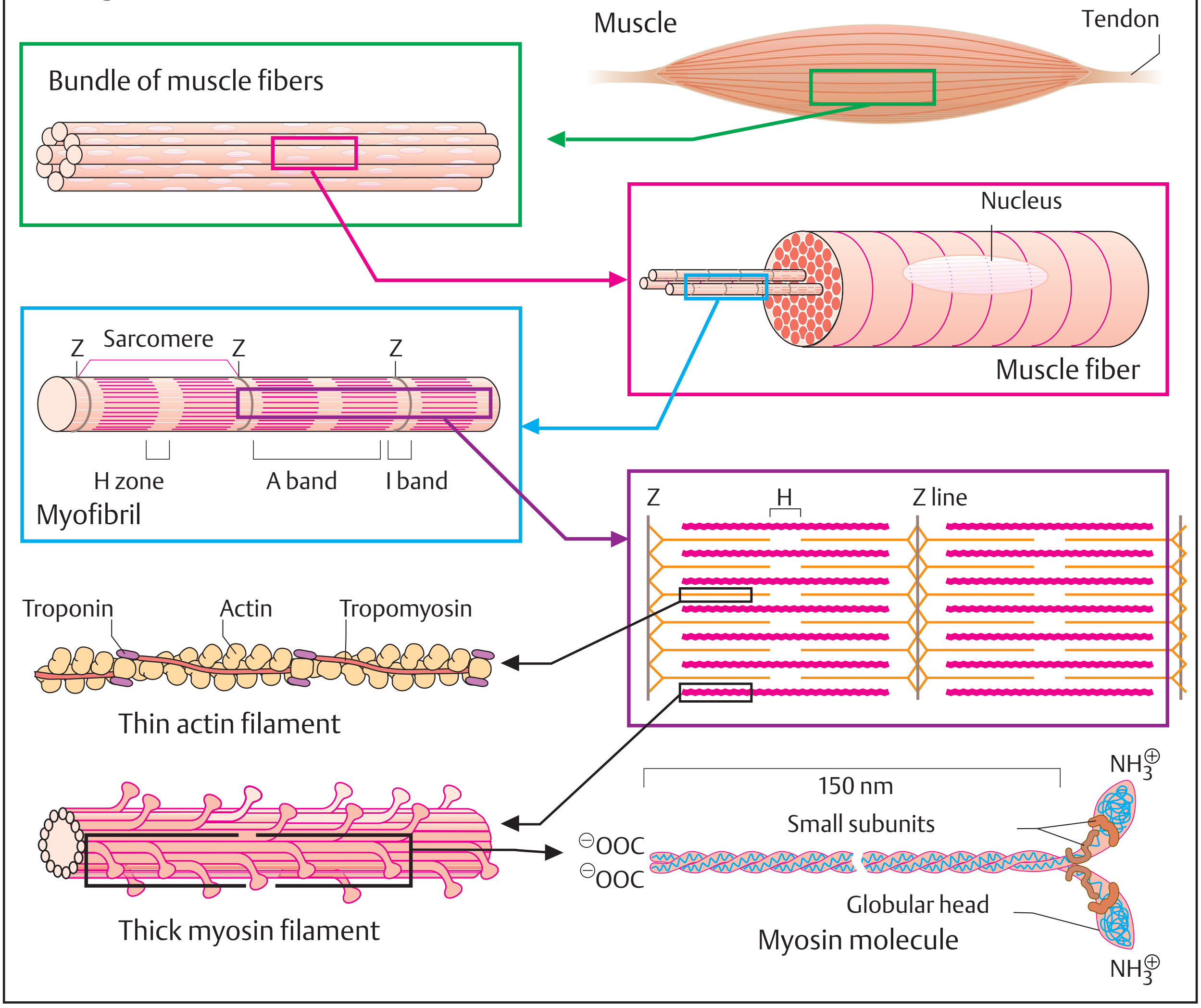
Muscle
Tendon
Bundle of muscle fibers
Nucleus
Muscle fiber
Z
Sarcomere
Z
Z
H zone
A band
I band
Myofibril
Z
H
Z line
Troponin
Actin
Tropomyosin
Thin actin filament
150 nm
Small subunits
$NH_3^{\oplus}$
$^{\ominus}OOC$
$^{\ominus}OOC$
Globular head
Myosin molecule
$NH_3^{\oplus}$
Thick myosin filament

B. Mechanism of muscle contraction

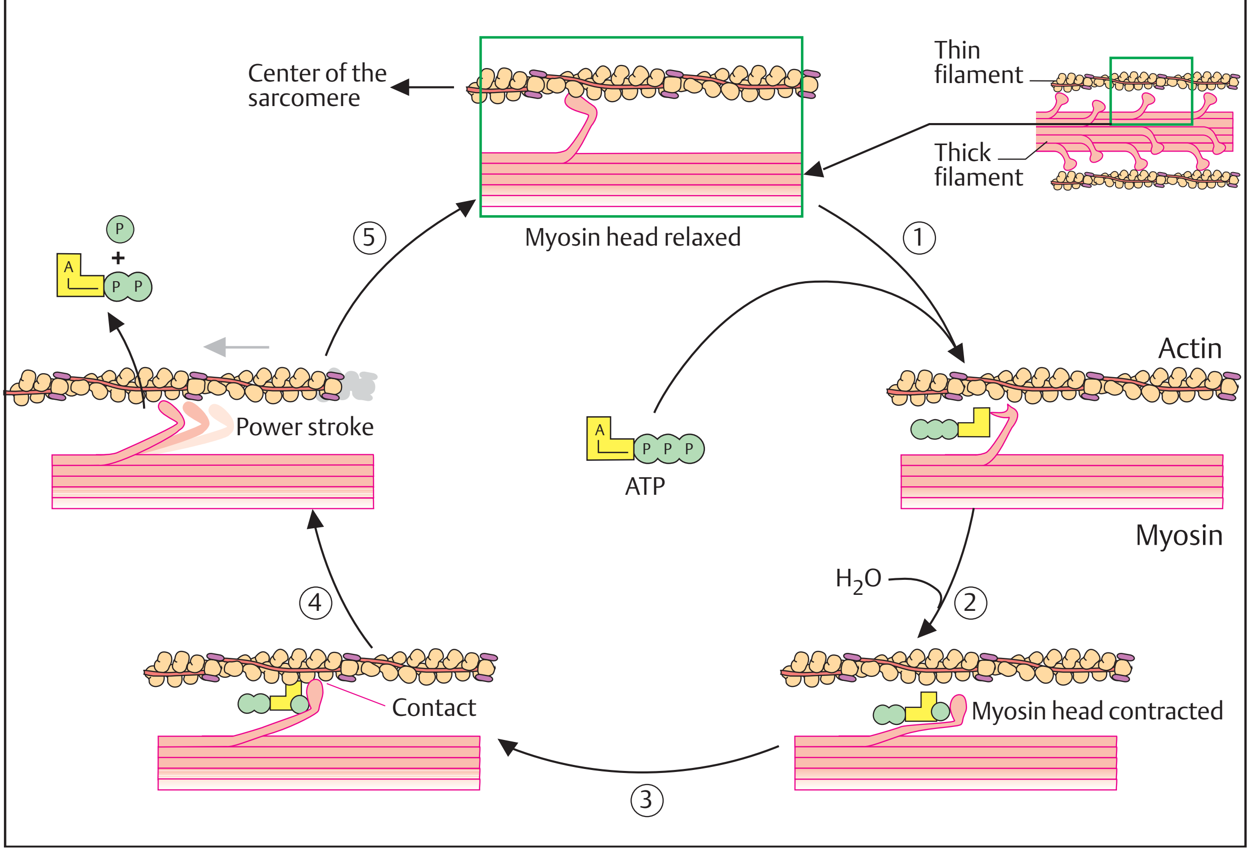
Center of the sarcomere
Thin filament
Thick filament
Myosin head relaxed
⑤
①
P
+
A
P P
Power stroke
A
P P P
ATP
Actin
Myosin
H_2O
②
④
Contact
Myosin head contracted
③

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Control of muscle contraction

A. Neuromuscular junction ◑

Muscle contraction is triggered by *motor neurons* that release the neurotransmitter *acetylcholine* (see p. 352). The transmitter diffuses through the narrow synaptic cleft and binds to nicotinic *acetylcholine receptors* on the plasma membrane of the muscle cell (the *sarcolemma*), thereby opening the ion channels integrated into the receptors (see p. 222). This leads to an inflow of Na^+, which triggers an *action potential* (see p. 350) in the sarcolemma. The action potential propagates from the end plate in all directions and constantly stimulates the muscle fiber. With a delay of a few milliseconds, the contractile mechanism responds to this by contracting the muscle fiber.

B. Sarcoplasmic reticulum (SR) ◑

The action potential (**A**) produced at the neuromuscular junction is transferred in the muscle cell into a transient increase in the Ca^{2+} concentration in the cytoplasm of the muscle fiber (the *sarcoplasm*).

In the resting state, the Ca^{2+} level in the sarcoplasm is very low (less than 10^{-7} M). By contrast, the **sarcoplasmic reticulum** (SR), which corresponds to the ER, contains Ca^{2+} ions at a concentration of about 10^{-3} M. The SR is a branched organelle that surrounds the myofibrils like a net stocking inside the muscle fibers (illustrated at the top using the example of a heart muscle cell). The high Ca^{2+} level in the SR is maintained by Ca^{2+}-transporting ATPases (see p. 220). In addition, the SR also contains *calsequestrin,* a protein (55 kDa) that is able to bind numerous Ca^{2+} -ions via acidic amino acid residues.

The transfer of the action potential to the SR is made possible by **transverse tubules** (T tubules), which are open to the extracellular space and establish a close connection with the SR. There is a structure involved in the contact between the T tubule and the SR that was formerly known as the "SR foot" (it involves parts of the *ryanodine receptor;* see p. 386).

At the point of contact with the SR, the action potential triggers the opening of the *Ca^{2+} channels* on the surface of the sarcolemma. Calcium ions then leave the SR and enter the sarcoplasm, where they lead to a rapid increase in Ca^{2+} concentrations. This in turn causes the myofibrils to contract (**C**).

C. Regulation by calcium ions ◑

In relaxed skeletal muscle, the complex consisting of **troponin** and **tropomyosin** blocks the access of the myosin heads to actin (see p. 332). Troponin consists of three different subunits (**T**, **C**, and **I**). The rapid increase in cytoplasmic Ca^{2+} concentrations caused by opening of the calcium channels in the SR leads to binding of Ca^{2+} to the C subunit of troponin, which closely resembles calmodulin (see p. 386). This produces a conformational change in troponin that causes the whole troponin–tropomyosin complex to slip slightly and expose a binding site for myosin (red). This initiates the contraction cycle. After contraction, the sarcoplasmic Ca^{2+} concentration is quickly reduced again by active transport back into the SR. This results in troponin losing the bound Ca^{2+} ions and returning to the initial state, in which the binding site for myosin on actin is blocked. It is not yet clear whether the mechanism described above is the only one that triggers binding of myosin to actin.

When **triggering of contraction** in striated muscle occurs, the following sequence of processes thus takes place:

1. The sarcolemma is depolarized.
2. The action potential is signaled to Ca^{2+} channels in the SR.
3. The Ca^{2+} channels open and the Ca^{2+} level in the sarcoplasm increases.
4. Ca^{2+} binds to troponin C and triggers a conformational change.
5. Troponin causes tropomyosin to slip, and the myosin heads bind to actin.
6. The actin–myosin cycle takes place and the muscle fibers contract.

Conversely, at the **end of contraction**, the following processes take place:

1. The Ca^{2+} level in the sarcoplasm declines due to transport of Ca^{2+} back into the SR.
2. Troponin C loses Ca^{2+} and tropomyosin returns to its original position on the actin molecule.
3. The actin–myosin cycle stops and the muscle relaxes.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Neuromuscular junction

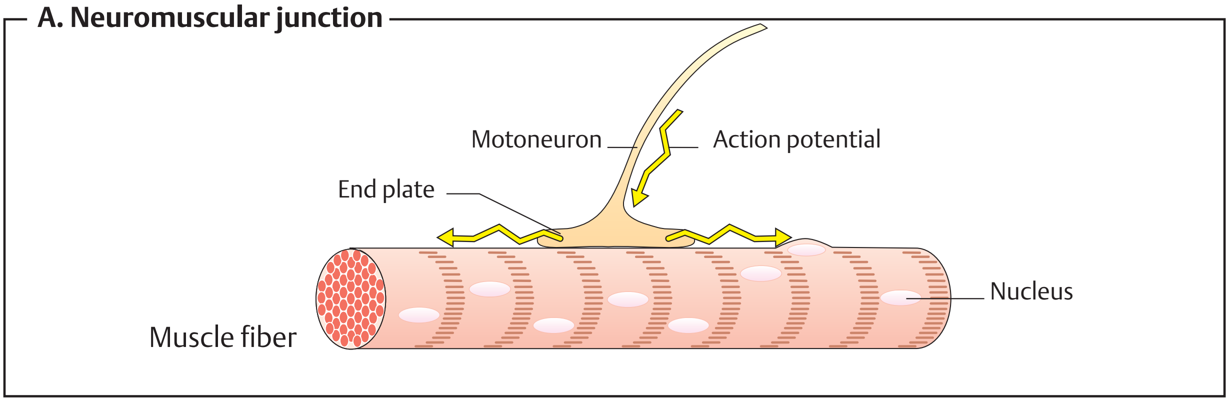
Motoneuron
Action potential
End plate
Nucleus
Muscle fiber

B. Sarcoplasmatic reticulum (SR)

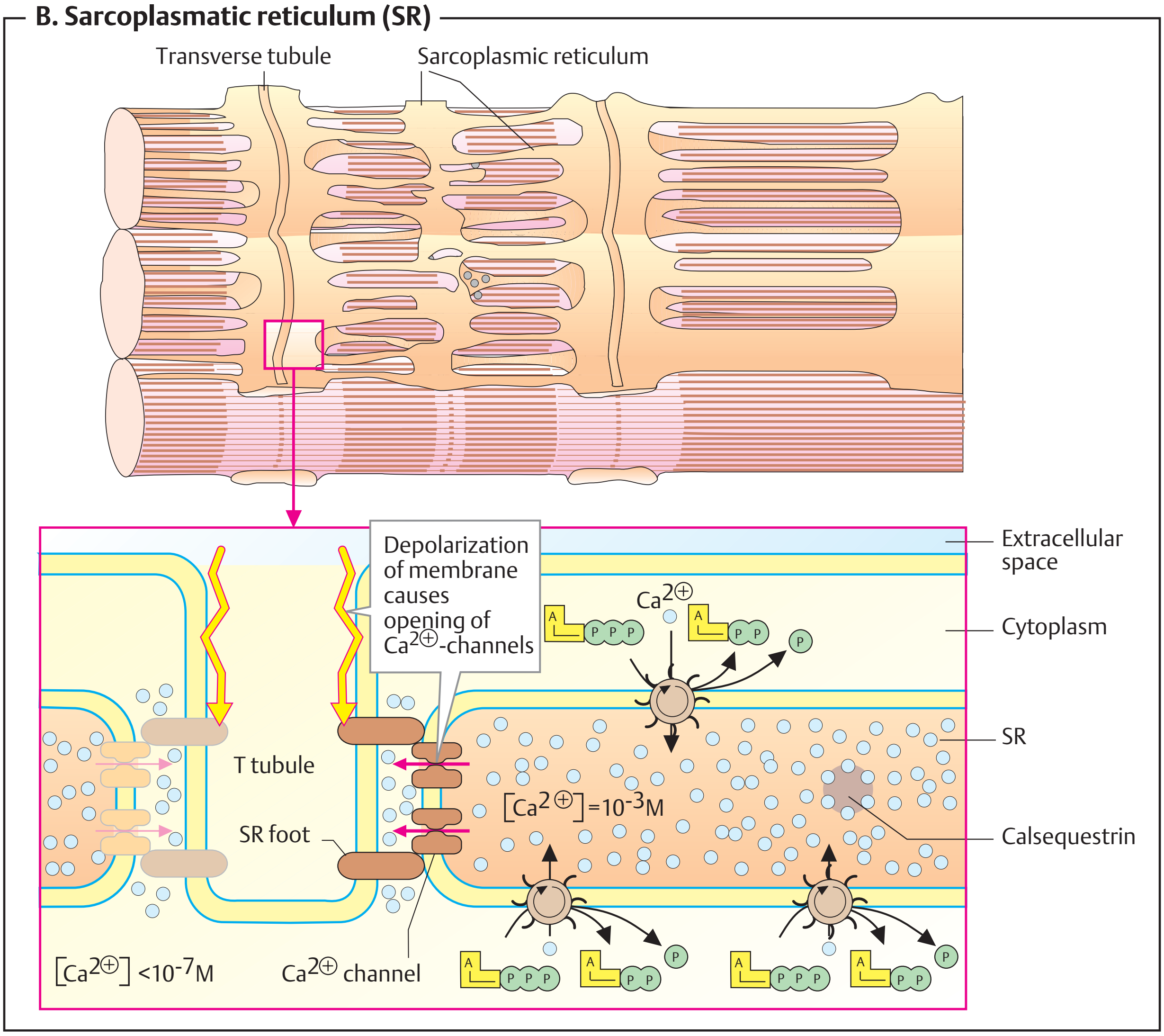
Transverse tubule
Sarcoplasmic reticulum
Depolarization of membrane causes opening of $Ca^{2\oplus}$-channels
$Ca^{2\oplus}$
Extracellular space
Cytoplasm
SR
Calsequestrin
T tubule
SR foot
$[Ca^{2\oplus}] = 10^{-3}M$
$[Ca^{2\oplus}] < 10^{-7}M$
$Ca^{2\oplus}$ channel
A
P
P
P

C. Regulation by calcium ions

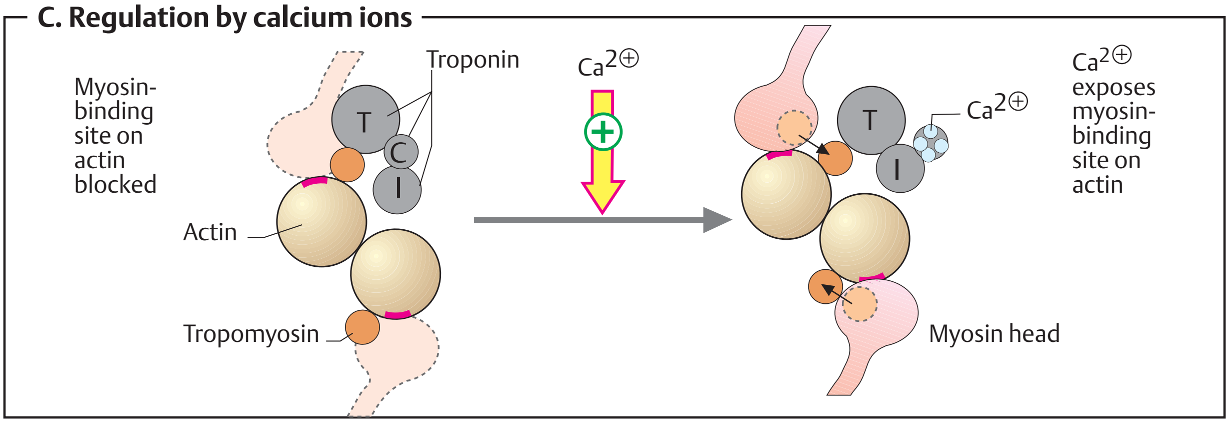
Myosin-binding site on actin blocked
Troponin
T
C
I
$Ca^{2\oplus}$
Actin
Tropomyosin
$Ca^{2\oplus}$
$Ca^{2\oplus}$ exposes myosin-binding site on actin
Myosin head

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Muscle metabolism I

Muscle contraction is associated with a high level of ATP consumption (see p. 332). Without constant resynthesis, the amount of ATP available in the resting state would be used up in less than 1 s of contraction.

A. Energy metabolism in the white and red muscle fibers ◑

Muscles contain two types of fibers, the proportions of which vary from one type of muscle to another. **Red fibers** (type I fibers) are suitable for prolonged effort. Their metabolism is mainly aerobic and therefore depends on an adequate supply of O_2. **White fibers** (type II fibers) are better suited for fast, strong contractions. These fibers are able to form sufficient ATP even when there is little O_2 available. With appropriate training, athletes and sports participants are able to change the proportions of the two fiber types in the musculature and thereby prepare themselves for the physiological demands of their disciplines in a targeted fashion. The expression of functional muscle proteins can also change during the course of training.

Red fibers provide for their ATP requirements mainly (but not exclusively) from **fatty acids**, which are broken down via β-oxidation, the tricarboxylic acid cycle, and the respiratory chain (right part of the illustration). The red color in these fibers is due to the monomeric heme protein **myoglobin**, which they use as an O_2 reserve. Myoglobin has a much higher affinity for O_2 than hemoglobin and therefore only releases its O_2 when there is a severe drop in O_2 partial pressure (cf. p. 282).

At a high level of muscular effort—e. g., during weightlifting or in very fast contractions such as those carried out by the eye muscles—the O_2 supply from the blood quickly becomes inadequate to maintain the aerobic metabolism. White fibers (left part of the illustration) therefore mainly obtain ATP from **anaerobic glycolysis**. They have supplies of **glycogen** from which they can quickly release glucose-1-phosphate when needed (see p. 156). By isomerization, this gives rise to glucose-6-phosphate, the substrate for glycolysis. The $NADH+H^+$ formed during glycolysis has to be reoxidized into NAD^+ in order to maintain glucose degradation and thus ATP formation. If there is a lack of O_2, this is achieved by the formation of **lactate**, which is released into the blood and is resynthesized into glucose in the liver (Cori cycle; see p. 338).

Muscle-specific auxiliary reactions for ATP synthesis exist in order to provide additional ATP in case of emergency. **Creatine phosphate** (see **B**) acts as a buffer for the ATP level. Another ATP-supplying reaction is catalyzed by *adenylate kinase* [1] (see also p. 72). This disproportionates two molecules of ADP into ATP and AMP. The AMP is deaminated into IMP in a subsequent reaction [2] in order to shift the balance of the reversible reaction [1] in the direction of ATP formation.

B. Creatine metabolism ◑

Creatine (*N*-methylguanidoacetic acid) and its phosphorylated form **creatine phosphate** (a guanidophosphate) serve as an ATP buffer in muscle metabolism. In creatine phosphate, the phosphate residue is at a similarly high chemical potential as in ATP and is therefore easily transferred to ADP. Conversely, when there is an excess of ATP, creatine phosphate can arise from ATP and creatine. Both processes are catalyzed by *creatine kinase* [5].

In resting muscle, creatine phosphate forms due to the high level of ATP. If there is a risk of a severe drop in the ATP level during contraction, the level can be maintained for a short time by synthesis of ATP from creatine phosphate and ADP. In a nonenzymatic reaction [6], small amounts of creatine and creatine phosphate cyclize constantly to form **creatinine**, which can no longer be phosphorylated and is therefore excreted with the urine (see p. 324).

Creatine does not derive from the muscles themselves, but is synthesized in two steps in the kidneys and liver (left part of the illustration). Initially, the guanidino group of arginine is transferred to glycine in the kidneys, yielding **guanidino acetate** [3]. In the liver, *N*-methylation of guanidino acetate leads to the formation of creatine from this [4]. The coenzyme in this reaction is *S-adenosyl methionine* (SAM; see p. 110).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Energy metabolism in the white and red muscle fibers

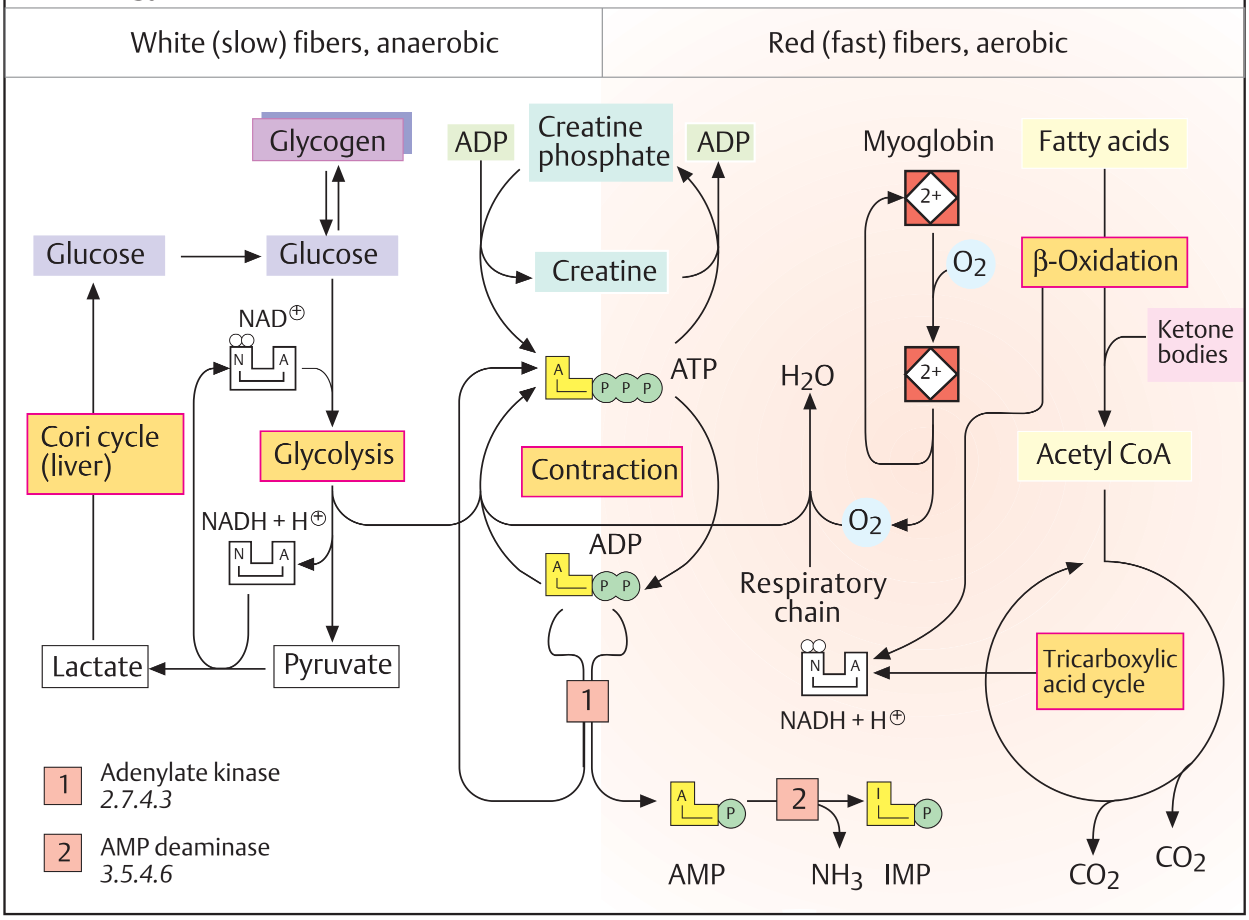
White (slow) fibers, anaerobic
Red (fast) fibers, aerobic
Glycogen
ADP
Creatine phosphate
ADP
Myoglobin
Fatty acids
Glucose
Glucose
Creatine
2+
O_2
β-Oxidation
$NAD^{\oplus}$
N A
ATP
H_2O
2+
Ketone bodies
Cori cycle (liver)
Glycolysis
Contraction
Acetyl CoA
$NADH + H^{\oplus}$
O_2
ADP
Respiratory chain
Lactate
Pyruvate
1
Tricarboxylic acid cycle
$NADH + H^{\oplus}$
1 Adenylate kinase 2.7.4.3
2 AMP deaminase 3.5.4.6
2
AMP
NH_3
IMP
CO_2
CO_2

B. Creatine metabolism

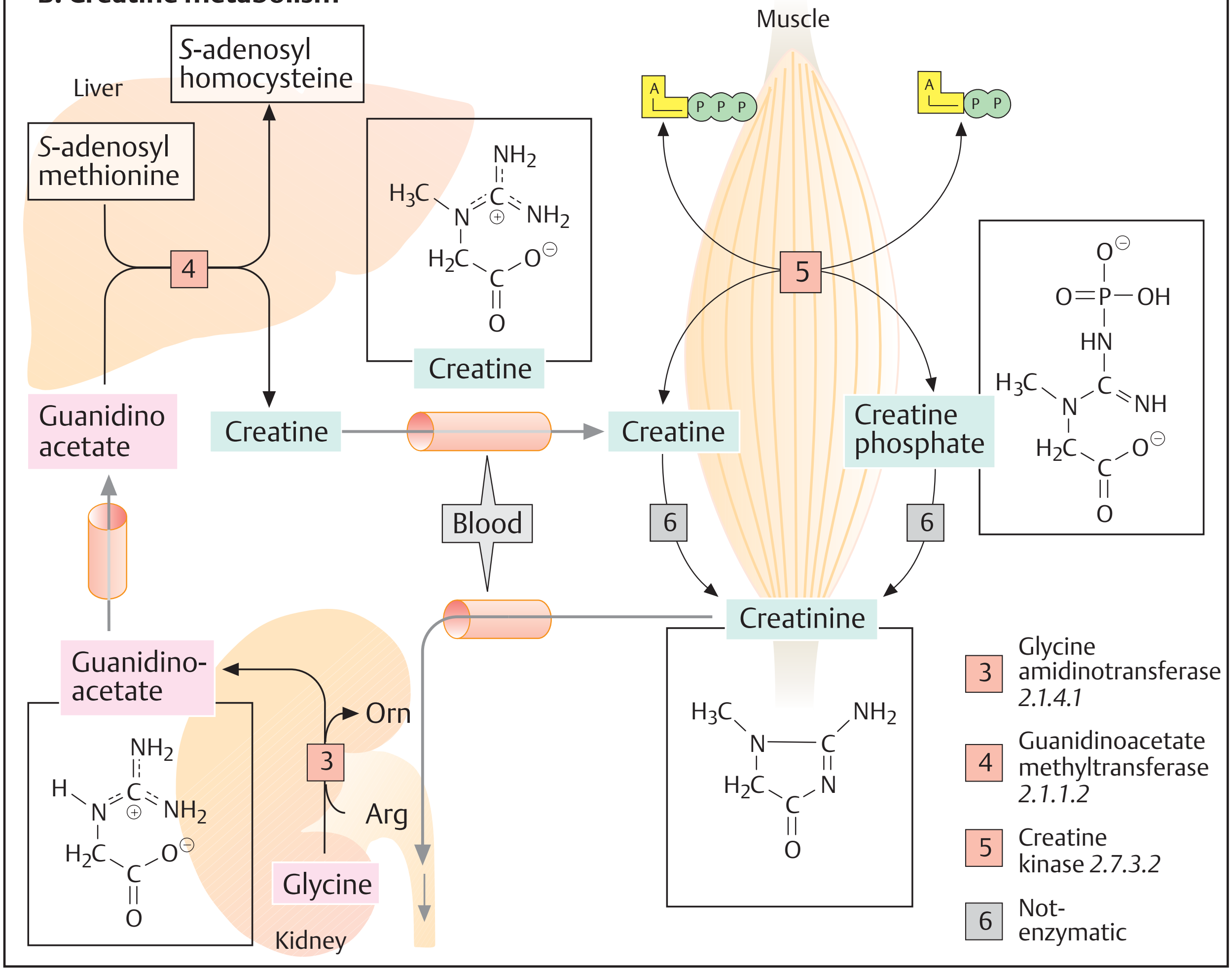
Muscle
Liver
S-adenosyl homocysteine
S-adenosyl methionine
4
NH_2
H_3C
N
C
NH_2
H_2C
C
$O^{\ominus}$
O
Creatine
5
Guanidino acetate
Creatine
Creatine
Creatine phosphate
$O^{\ominus}$
O=P—OH
HN
H_3C
N
C
NH
H_2C
C
$O^{\ominus}$
O
Blood
6
6
Creatinine
Guanidino-acetate
Orn
3
Arg
NH_2
H
N
C
NH_2
H_2C
C
$O^{\ominus}$
O
Glycine
Kidney
H_3C
NH_2
N
C
N
H_2C
C
O
3 Glycine amidinotransferase 2.1.4.1
4 Guanidinoacetate methyltransferase 2.1.1.2
5 Creatine kinase 2.7.3.2
6 Not-enzymatic

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Muscle metabolism II

A. Cori and alanine cycle ◑

White muscle fibers (see p. 336) mainly obtain ATP from **anaerobic glycolysis**—i. e., they convert glucose into lactate. The **lactate** arising in muscle and, in smaller quantities, its precursor **pyruvate** are released into the blood and transported to the liver, where lactate and pyruvate are resynthesized into glucose again via *gluconeogenesis,* with ATP being consumed in the process (see p. 154). The glucose newly formed by the liver returns via the blood to the muscles, where it can be used as an energy source again. This circulation system is called the **Cori cycle**, after the researchers who first discovered it. There is also a very similar cycle for erythrocytes, which do not have mitochondria and therefore produce ATP by anaerobic glycolysis (see p. 284).

The muscles themselves are not capable of gluconeogenesis. Nor would this be useful, as gluconeogenesis requires much more ATP than is supplied by glycolysis. As O_2 deficiencies do not arise in the liver even during intensive muscle work, there is always sufficient energy there available for gluconeogenesis.

There is also a corresponding circulation system for the amino acid alanine. The **alanine cycle** in the liver not only provides alanine as a precursor for gluconeogenesis, but also transports to the liver the amino nitrogen arising in muscles during protein degradation. In the liver, it is incorporated into urea for excretion.

Most of the amino acids that arise in muscle during proteolysis are converted into glutamate and 2-oxo acids by *transamination* (not shown; cf. p. 180). Again by transamination, glutamate and pyruvate give rise to alanine, which after glutamine is the second important form of transport for amino nitrogen in the blood. In the liver, alanine and 2-oxoglutarate are resynthesized into pyruvate and glutamate (see p. 178). Glutamate supplies the urea cycle (see p. 182), while pyruvate is available for gluconeogenesis.

B. Protein and amino acid metabolism ◑

The skeletal muscle is the most important site for degradation of the *branched-chain amino acids* (Val, Leu, Ile; see p. 414), but other amino acids are also broken down in the muscles. **Alanine** and **glutamine** are resynthesized from the components and released into the blood. They transport the nitrogen that arises during amino acid breakdown to the liver (*alanine cycle;* see above) and to the kidneys (see p. 328).

During periods of hunger, **muscle proteins** serve as an energy reserve for the body. They are broken down into amino acids, which are transported to the liver. In the liver, the carbon skeletons of the amino acids are converted into intermediates in the tricarboxylic acid cycle or into acetoacetyl-CoA (see p. 175). These amphibolic metabolites are then available to the energy metabolism and for gluconeogenesis. After prolonged starvation, the brain switches to using ketone bodies in order to save muscle protein (see p. 356).

The synthesis and degradation of muscle proteins are regulated by hormones. **Cortisol** leads to muscle degradation, while **testosterone** stimulates protein formation. Synthetic **anabolics** with a testosterone-like effect have repeatedly been used for doping purposes or for intensive muscle-building.

Further information

Smooth muscle differs from skeletal muscle in various ways. Smooth muscles—which are found, for example, in blood vessel walls and in the walls of the intestines—do not contain any muscle fibers. In smooth-muscle cells, which are usually spindle-shaped, the contractile proteins are arranged in a less regular pattern than in striated muscle. Contraction in this type of muscle is usually not stimulated by nerve impulses, but occurs in a largely spontaneous way. Ca^{2+} (in the form of Ca^{2+}-calmodulin; see p. 386) also activates contraction in smooth muscle; in this case, however, it does not affect troponin, but activates a protein kinase that phosphorylates the light chains in myosin and thereby increases myosin's ATPase activity. Hormones such as epinephrine and angiotensin II (see p. 330) are able to influence vascular tonicity in this way, for example.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Cori and alanine cycle

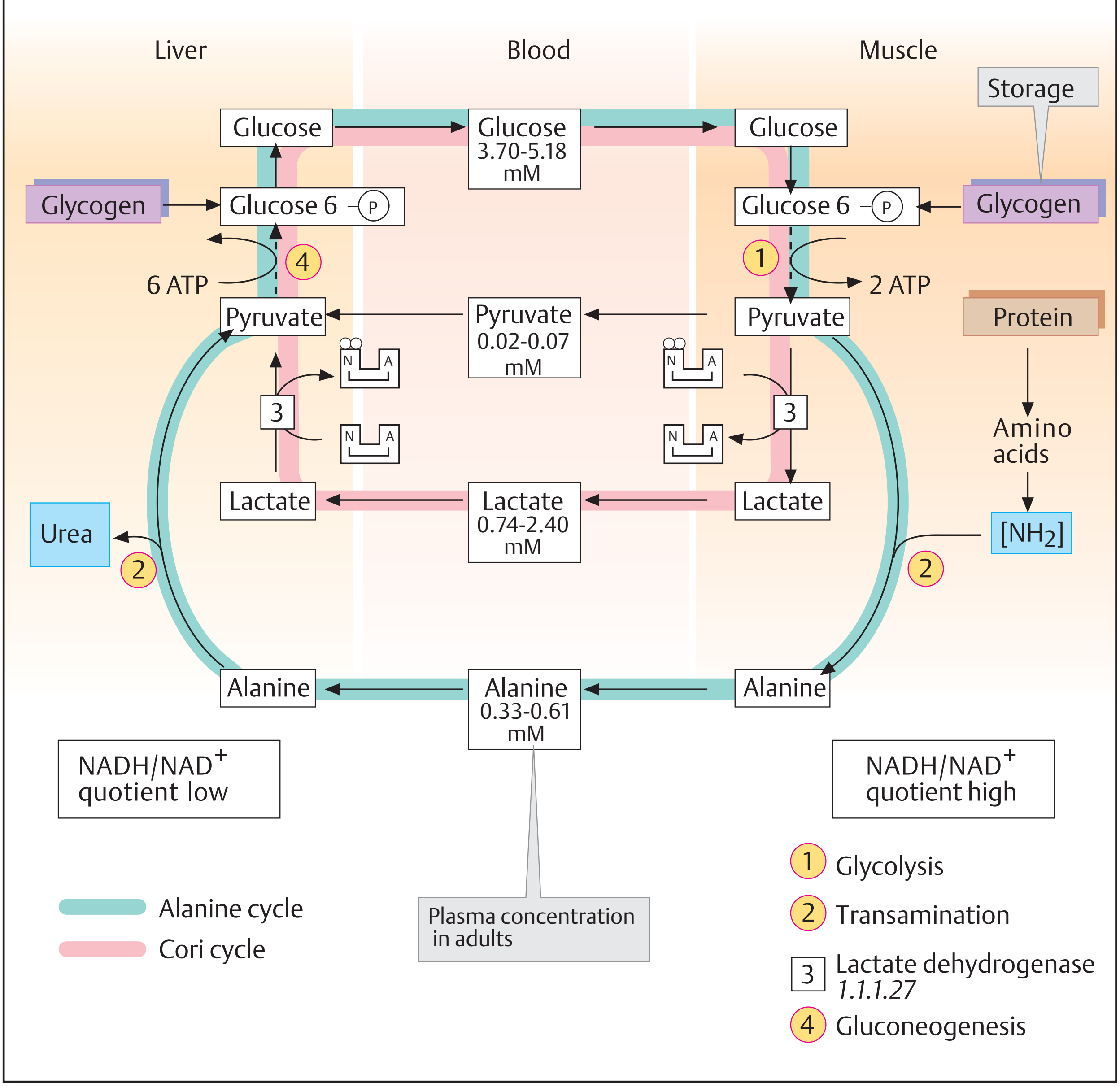
Liver
Blood
Muscle
Storage
Glucose
Glucose 3.70-5.18 mM
Glucose
Glycogen
Glucose 6 -Ⓟ
Glucose 6 -Ⓟ
Glycogen
6 ATP
④
①
2 ATP
Pyruvate
Pyruvate 0.02-0.07 mM
Pyruvate
Protein
N A
N A
3
3
N A
N A
Amino acids
Lactate
Lactate 0.74-2.40 mM
Lactate
Urea
②
②
[NH2]
Alanine
Alanine 0.33-0.61 mM
Alanine
NADH/NAD+ quotient low
NADH/NAD+ quotient high
Alanine cycle
Cori cycle
Plasma concentration in adults
① Glycolysis
② Transamination
3 Lactate dehydrogenase 1.1.1.27
④ Gluconeogenesis

B. Protein and amino acid metabolism

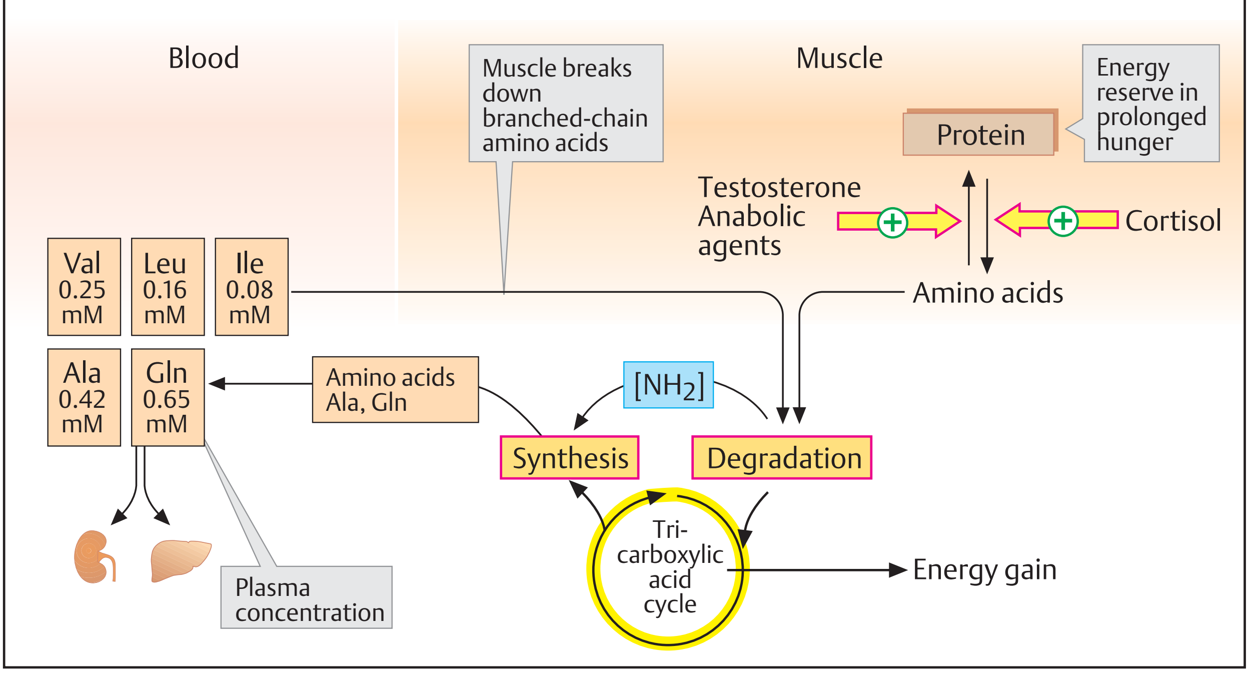
Blood
Muscle breaks down branched-chain amino acids
Muscle
Energy reserve in prolonged hunger
Protein
Testosterone Anabolic agents
Cortisol
Val 0.25 mM
Leu 0.16 mM
Ile 0.08 mM
Amino acids
Ala 0.42 mM
Gln 0.65 mM
Amino acids Ala, Gln
[NH2]
Synthesis
Degradation
Tri-carboxylic acid cycle
Energy gain
Plasma concentration

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Bone and teeth

The family of connective-tissue cells includes *fibroblasts, chondrocytes* (cartilage cells), and *osteoblasts* (bone-forming cells). They are specialized to secrete extracellular proteins, particularly collagens, and mineral substances, which they use to build up the *extracellular matrix* (see p. 346). By contrast, *osteoclasts* dissolve bone matter again by secreting H^+ and collagenases (see p. 342).

A. Bone ◑

Bone is an extremely dense, specialized form of connective tissue. In addition to its supportive function, it serves to store calcium and phosphate ions. In addition, blood cells are formed in the bone marrow. The most important mineral component of bone is **apatite**, a form of crystalline *calcium phosphate.*

Apatites are complexes of cationic Ca^{2+} matched by HPO_4^{2-}, CO_3^{2-}, OH^-, or F^- as anions. Depending on the counter-ion, apatite can occur in the forms *carbonate apatite* $Ca_{10}(PO_4)_6CO_3$, as *hydroxyapatite* $Ca_{10}(PO_4)_6(OH)_2$, or *fluoroapatite* $Ca_{10}(PO_4)_6F_2$. In addition, alkaline earth carbonates also occur in bone. In adults, more than 1 kg calcium is stored in bone.

Osteoblast and *osteoclast* activity is constantly incorporating Ca^{2+} into bone and removing it again. There are various hormones that regulate these processes: *calcitonin* increases deposition of Ca^{2+} in the bone matrix, while *parathyroid hormone* (PTH) promotes the mobilization of Ca^{2+}, and *calcitriol* improves mineralization (for details, see p. 342).

The most important *organic components* of bone are **collagens** (mainly type I; see p. 344) and **proteoglycans** (see p. 346). These form the extracellular matrix into which the apatite crystals are deposited (*biomineralization*). Various proteins are involved in this not yet fully understood process of bone formation, including collagens and phosphatases. *Alkaline phosphatase* is found in osteoblasts and *acid phosphatase* in osteoclasts. Both of these enzymes serve as *marker enzymes* for bone cells.

B. Teeth ◑

The illustration shows a longitudinal section through an incisor, one of the 32 permanent teeth in humans. The majority of the tooth consists of **dentine**. The crown of the tooth extends beyond the gums, and it is covered in **enamel**. By contrast, the root of the tooth is coated in dental **cement**.

Cement, dentin, and enamel are bone-like substances. The high proportion of inorganic matter they contain (about 97% in the dental enamel) gives them their characteristic hardness. The organic components of cement, dentin, and enamel mainly consist of *collagens* and *proteoglycans;* their most important mineral component is *apatite,* as in bone (see above).

A widespread form of dental disease, **caries**, is caused by acids that dissolve the mineral part of the teeth by neutralizing the negatively charged counter-ions in apatite (see **A**). Acids occur in food, or are produced by microorganisms that live on the surfaces of the teeth (e. g., *Streptococcus mutans*).

The main product of anaerobic degradation of sugars by these organisms is lactic acid. Other products of bacterial carbohydrate metabolism include extracellular dextrans (see p. 40)—insoluble polymers of glucose that help bacteria to protect themselves from their environment. Bacteria and dextrans are components of *dental plaque,* which forms on inadequately cleaned teeth. When Ca^{2+} salts and other minerals are deposited in plaque as well, *tartar* is formed.

The most important form of protection against caries involves avoiding sweet substances (foods containing saccharose, glucose, and fructose). Small children in particular should not have very sweet drinks freely available to them. Regular removal of plaque by cleaning the teeth and hardening of the dental enamel by fluoridization are also important. Fluoride has a protective effect because fluoroapatite (see **A**) is particularly resistant to acids.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Bone

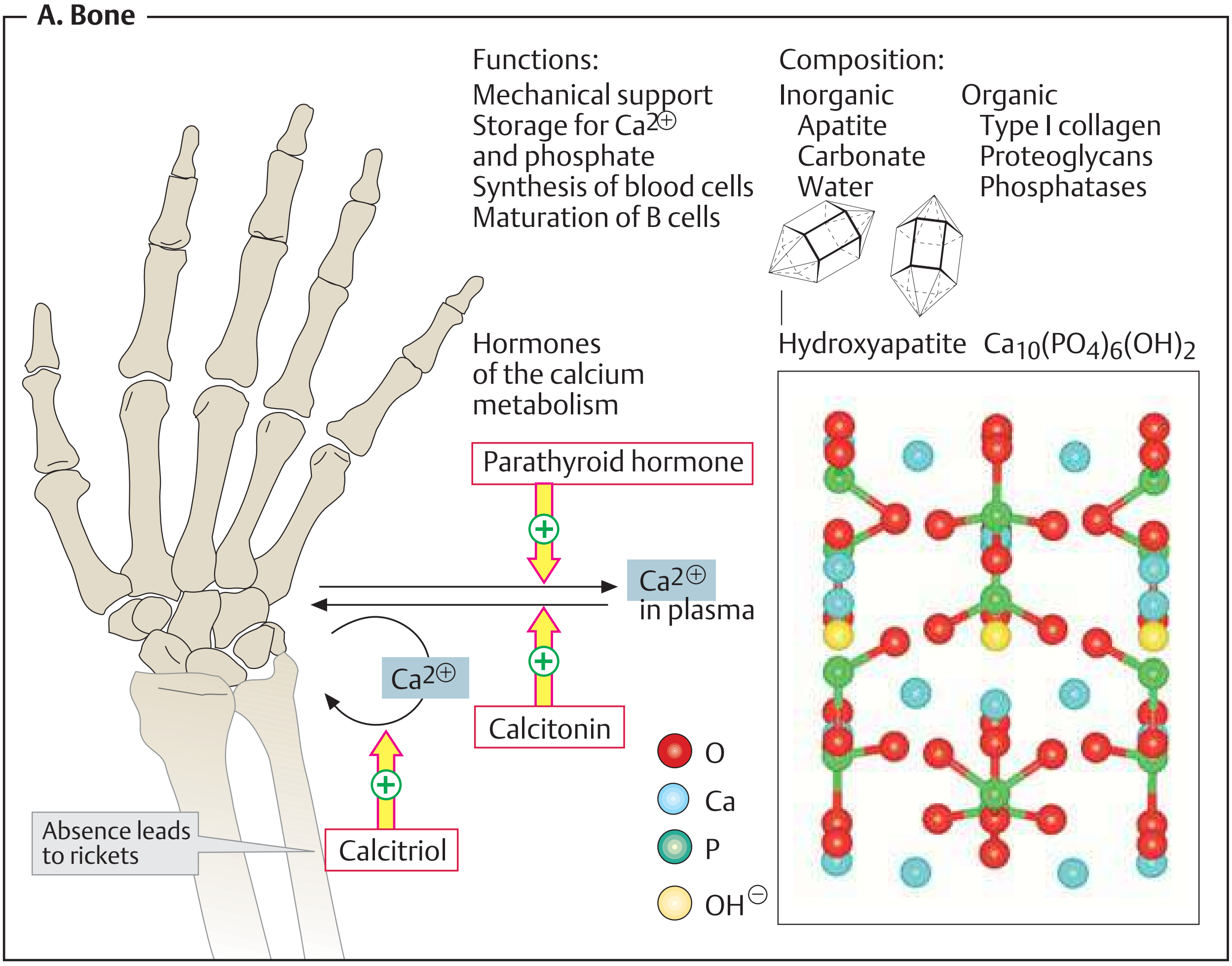
Functions:
Mechanical support
Storage for $Ca^{2\oplus}$
and phosphate
Synthesis of blood cells
Maturation of B cells
Composition:
Inorganic
Apatite
Carbonate
Water
Organic
Type I collagen
Proteoglycans
Phosphatases
Hydroxyapatite $Ca_{10}(PO_4)_6(OH)_2$
Hormones
of the calcium
metabolism
Parathyroid hormone
$Ca^{2\oplus}$
in plasma
$Ca^{2\oplus}$
Calcitonin
Calcitriol
Absence leads
to rickets
O
Ca
P
$OH^{\ominus}$

B. Teeth

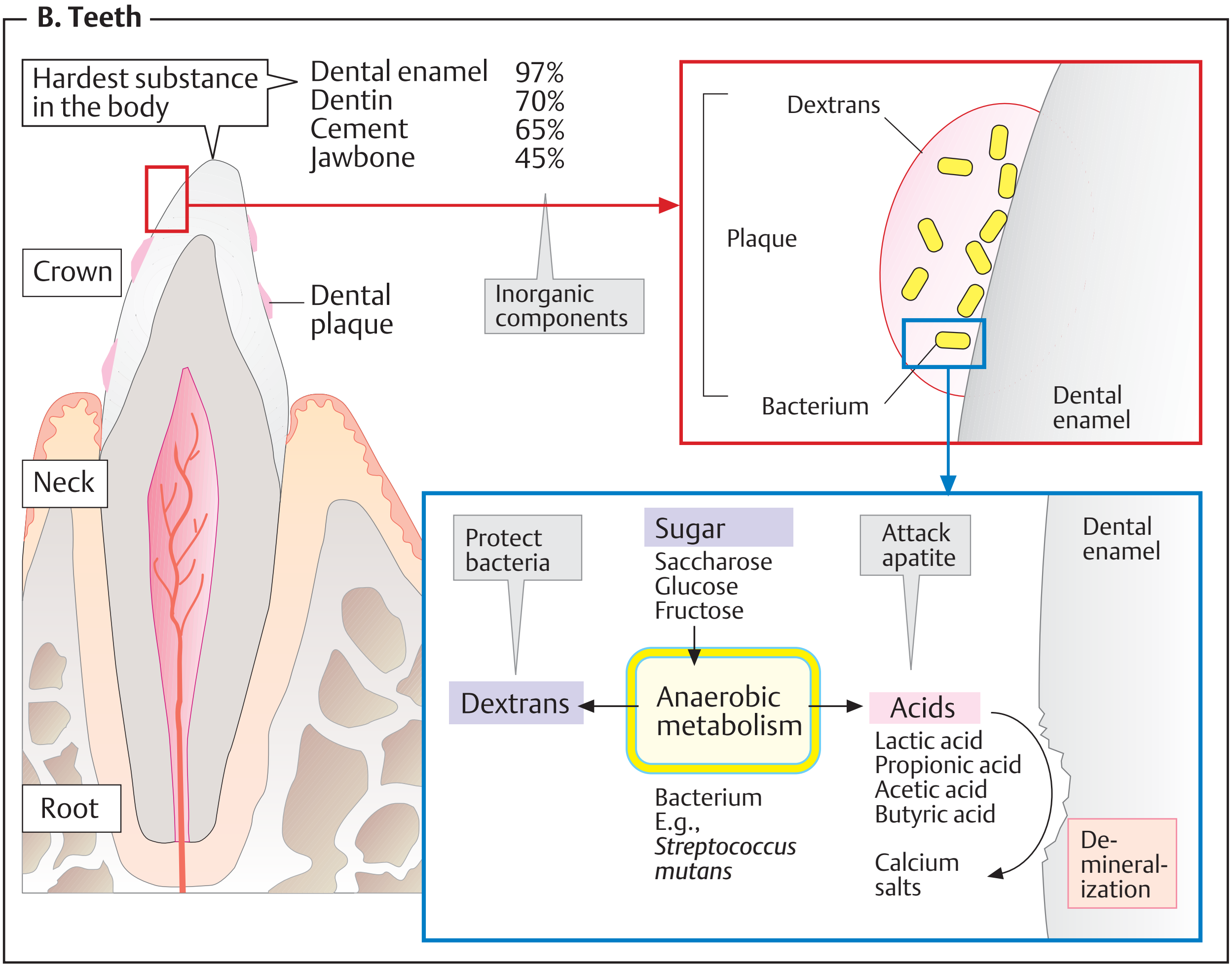
Hardest substance
in the body
Dental enamel 97%
Dentin 70%
Cement 65%
Jawbone 45%
Crown
Dental
plaque
Inorganic
components
Neck
Root
Dextrans
Plaque
Bacterium
Dental
enamel
Protect
bacteria
Sugar
Saccharose
Glucose
Fructose
Attack
apatite
Dental
enamel
Dextrans
Anaerobic
metabolism
Acids
Lactic acid
Propionic acid
Acetic acid
Butyric acid
Bacterium
E.g.,
Streptococcus
mutans
Calcium
salts
De-
mineral-
ization

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Calcium metabolism

A. Functions of calcium ●

The human body contains 1–1.5 kg Ca^{2+}, most of which (about 98%) is located in the mineral substance of bone (see p. 362).

In addition to its role as a **bone component**, calcium functions as a **signaling substance**. Ca^{2+} ions act as *second messengers* in signal transduction pathways (see p. 386), they trigger *exocytosis* (see p. 228) and *muscle contraction* (see p. 334), and they are indispensable as cofactors in *blood coagulation* (see p. 290). Many *enzymes* also require Ca^{2+} for their activity. The intracellular and extracellular concentrations of Ca^{2+} are strictly regulated in order to make these functions possible (see **B**, **C**, and p. 388).

Proteins bind Ca^{2+} via oxygen ligands, particularly carboxylate groups and carbonyl groups of peptide bonds. This also applies to the structure illustrated here, in which a Ca^{2+} ion is coordinated by the oxygen atoms of carboxylate and acid amide groups.

B. Bone remodeling ◑

Deposition of Ca^{2+} in bone (*mineralization*) and Ca^{2+} mobilization from bone are regulated by at least 15 hormones and hormone-like signaling substances. These mainly influence the maturation and activity of bone cells.

Osteoblasts (top) deposit collagen, as well as Ca^{2+} and phosphate, and thereby create new bone matter, while **osteoclasts** (bottom) secrete H^+ ions and collagenases that locally dissolve bone (*bone remodeling*). Osteoblasts and osteoclasts mutually activate each other by releasing **cytokines** (see p. 392) and **growth factors**. This helps keep bone formation and bone breakdown in balance.

The Ca^{2+}-selective hormones calcitriol, parathyroid hormone, and calcitonin influence this interaction in the bone cells. **Parathyroid hormone** promotes Ca^{2+} release by promoting the release of cytokines by osteoblasts. In turn, the cytokines stimulate the development of mature osteoclasts from precursor cells (bottom). **Calcitonin** inhibits this process. At the same time, it promotes the development of osteoblasts (top). *Osteoporosis,* which mainly occurs in women following the menopause, is based (at least in part) on a reduction in **estrogen** levels. Estrogens normally inhibit the stimulation of osteoclast differentiation by osteoblasts. If the effects of estrogen decline, the osteoclasts predominate and excess bone removal occurs.

The effects of the steroid hormone **calcitriol** (see p. 330) in bone are complex. On the one hand, it promotes bone formation by stimulating osteoblast differentiation (top). This is particularly important in small children, in whom calcitriol deficiency can lead to mineralization disturbances (*rickets;* see p. 364). On the other hand, calcitriol increases blood Ca^{2+} levels through increased Ca^{2+} mobilization from bone. An overdose of vitamin D (cholecalciferol), the precursor of calcitriol, can therefore have unfavorable effects on the skeleton similar to those of vitamin deficiency (*hypervitaminosis;* see p. 364).

C. Calcium homeostasis ◑

Ca^{2+} metabolism is balanced in healthy adults. Approximately 1 g Ca^{2+} is taken up per day, about 300 mg of which is resorbed. The same amount is also excreted again. The amounts of Ca^{2+} released from bone and deposited in it per day are much smaller. Milk and milk products, especially cheese, are particularly rich in calcium.

Calcitriol and parathyroid hormone, on the one hand, and calcitonin on the other, ensure a more or less constant level of Ca^{2+} in the blood plasma and in the extracellular space (80–110 mg 2.0–2.6 mM). The peptide **parathyroid hormone** (PTH; 84 AA) and the steroid **calcitriol** (see p. 374) promote direct or indirect processes that raise the Ca^{2+} level in blood. Calcitriol increases Ca^{2+} resorption in the intestines and kidneys by inducing transporters. Parathyroid hormone supports these processes by stimulating calcitriol biosynthesis in the kidneys (see p. 330). In addition, it directly promotes resorption of Ca^{2+} in the kidneys (see p. 328) and Ca^{2+} release from bone (see **B**). The PTH antagonist **calcitonin** (32 AA) counteracts these processes.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

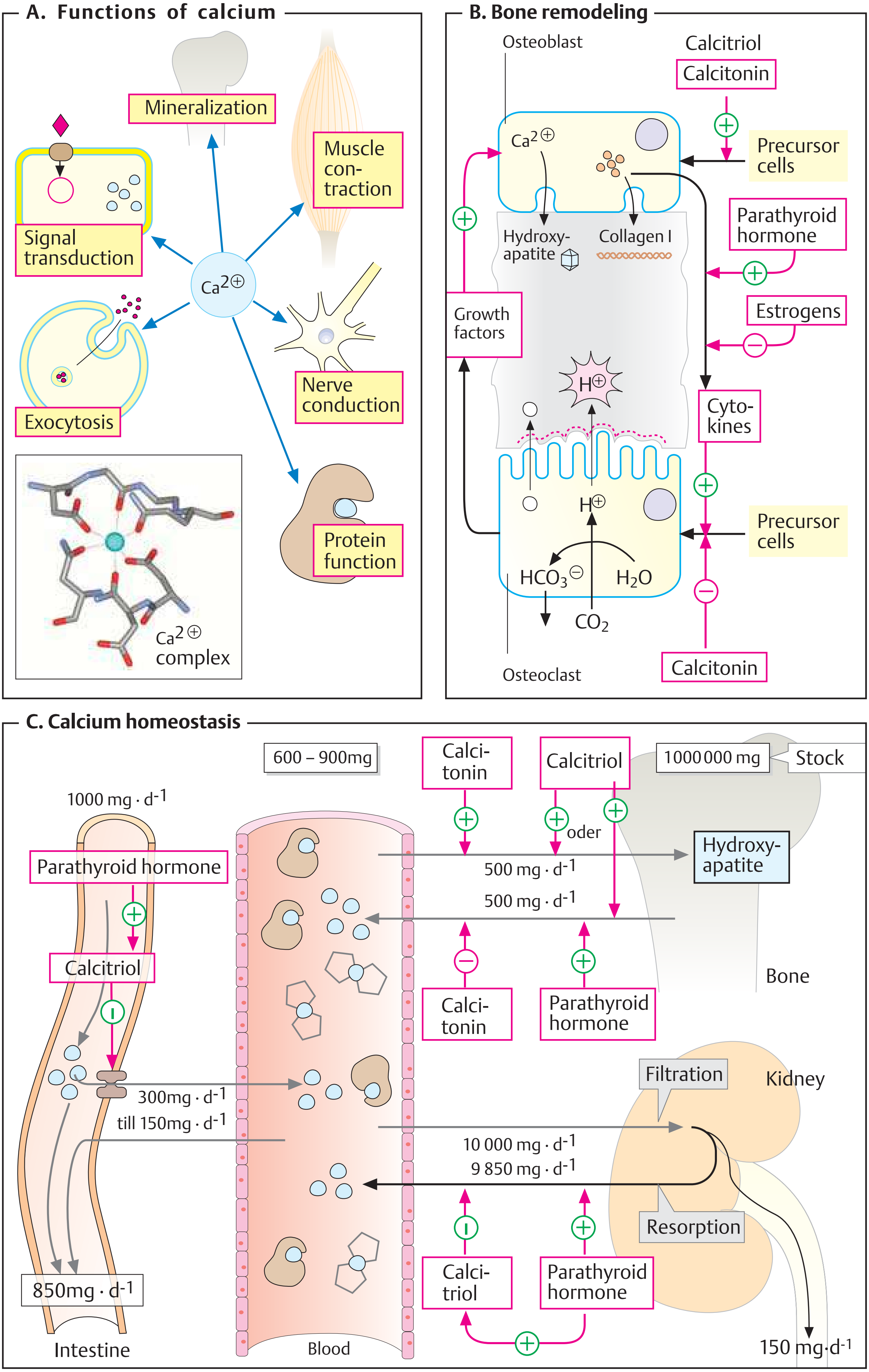
A. Functions of calcium
Mineralization
Muscle con-traction
Signal transduction
$Ca^{2\oplus}$
Exocytosis
Nerve conduction
Protein function
$Ca^{2\oplus}$ complex
B. Bone remodeling
Osteoblast
Calcitriol
Calcitonin
$Ca^{2\oplus}$
Precursor cells
Hydroxy-apatite
Collagen I
Parathyroid hormone
Estrogens
Growth factors
$H^{\oplus}$
Cyto-kines
Precursor cells
$H^{\oplus}$
$HCO_3^{\ominus}$
H_2O
CO_2
Osteoclast
Calcitonin
C. Calcium homeostasis
600 – 900mg
Calci-tonin
Calcitriol
1000000 mg
Stock
1000 mg · d-1
oder
Parathyroid hormone
Hydroxy-apatite
500 mg · d-1
500 mg · d-1
Calcitriol
Calci-tonin
Parathyroid hormone
Bone
Filtration
Kidney
300mg · d-1
till 150mg · d-1
10 000 mg · d-1
9 850 mg · d-1
Resorption
850mg · d-1
Calci-triol
Parathyroid hormone
Intestine
Blood
150 mg·d-1

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Collagens

Collagens are quantitatively the most abundant of animal proteins, representing 25% of the total. They form insoluble tensile fibers that occur as structural elements of the extracellular matrix and connective tissue throughout the body. Their name (which literally means "glue-producers") is derived from the gelatins that appear as a decomposition product when collagen is boiled.

A. Structure of collagens ◑

Nineteen different collagens are now known, and they are distinguished using roman numerals. They mostly consist of a dextrorotatory **triple helix** made up of three polypeptides (α-chains) (see p. 70).

The triplet **Gly-X-Y** is constantly repeated in the sequence of the triple-helical regions—i. e., every third amino acid in such sequences is a *glycine. Proline* (Pro) is frequently found in positions X or Y; the Y position is often occupied by *4-hydroxyproline* (4Hyp), although *3-hydroxyproline* (3Hyp) and *5-hydroxylysine* (5Hyl) also occur. These hydroxylated amino acids are characteristic components of collagen. They are only produced after protein biosynthesis by **hydroxylation** of the amino acids in the peptide chain (see p. 62).

The formation of Hyp and Hyl residues in procollagen is catalyzed by iron-containing *oxygenases* ("proline and lysine hydroxylase," *EC 1.14.11.1/2*). *Ascorbate* is required to maintain their function. Most of the symptoms of the vitamin C deficiency disease *scurvy* (see p. 368) are explained by disturbed collagen biosynthesis.

The hydroxyproline residues stabilize the triple helix by forming hydrogen bonds between the α-chains, while the hydroxyl groups of hydroxylysine are partly **glycosylated** with a disaccharide (–Glc–Gal).

The various types of collagen consist of different combinations of α-chains (α1 to α3 and other subtypes). Types I, II, and III represent 90% of collagens. The **type I** collagen shown here has the structure $[\alpha 1(I)]_2\alpha 2(1)$.

Numerous **tropocollagen** molecules (mass 285 kDa, length 400 nm) aggregate extracellularly into a defined arrangement, forming cylindrical **fibrils** (20–500 nm in diameter). Under the electron microscope, these fibrils are seen to have a characteristic banding pattern of elements that are repeated every 64–67 nm.

Tropocollagen molecules are firmly linked together, particularly at their ends, by covalent networks of altered lysine side chains. The number of these links increases with age. **Type IV** collagens form networks with a defined mesh size. The size-selective filtering effect of the basal membranes in the renal glomeruli is based on this type of structure (see p. 322).

B. Biosynthesis ◑

The precursor molecule of collagen (*preprocollagen*), formed in the rER, is subject to extensive **post-translational modifications** (see p. 232) in the ER and Golgi apparatus.

Cleavage of the signal peptide gives rise to **procollagen**, which still carries large *propeptides* at each end [1]. During this phase, most proline residues and some lysine residues of procollagen are hydroxylated [2]. The procollagen is then glycosylated at hydroxylysine residues [3]. Intramolecular and intermolecular disulfide bonds form in the propeptides [4], allowing correct positioning of the peptide strands to form a triple helix [5]. It is only after these steps have been completed that procollagen is secreted into the extracellular space by exocytosis. This is where the *N*- and *C*-terminal propeptides are removed proteolytically [6], allowing the staggered aggregation of the **tropocollagen** molecules to form fibrils [7]. Finally, several ε-amino groups in lysine residues are oxidatively converted into aldehyde groups [8]. Covalent links between the molecules then form as a result of condensation [9]. In this way, the fibrils reach their final structure, which is characterized by its high *tensile strength* and *proteinase resistance.*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Structure of collagens

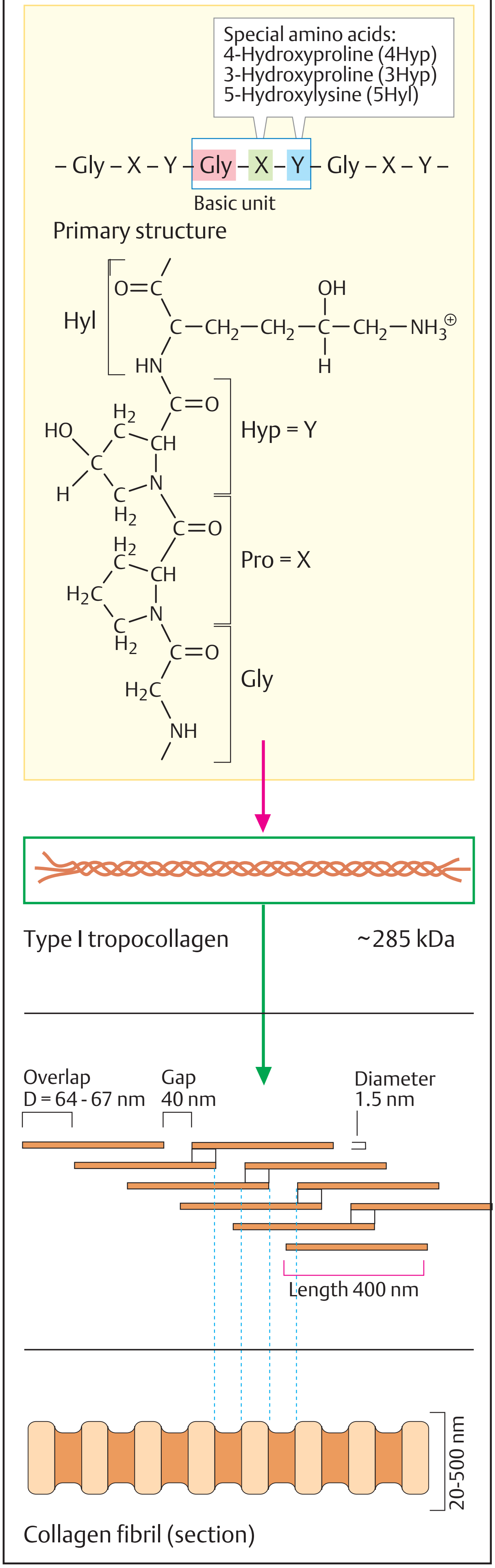
Special amino acids:
4-Hydroxyproline (4Hyp)
3-Hydroxyproline (3Hyp)
5-Hydroxylysine (5Hyl)
– Gly – X – Y – Gly – X – Y – Gly – X – Y –
Basic unit
Primary structure
Hyl
O=C
OH
C–CH2–CH2–C–CH2–NH3⊕
HN
H
C=O
HO
H2
C
CH
C
N
H
C
H2
Hyp = Y
C=O
H2
C
CH
H2C
N
C
H2
Pro = X
C=O
H2C
NH
Gly
Type I tropocollagen
~285 kDa
Overlap
D = 64 - 67 nm
Gap
40 nm
Diameter
1.5 nm
Length 400 nm
20-500 nm
Collagen fibril (section)

B. Biosynthesis

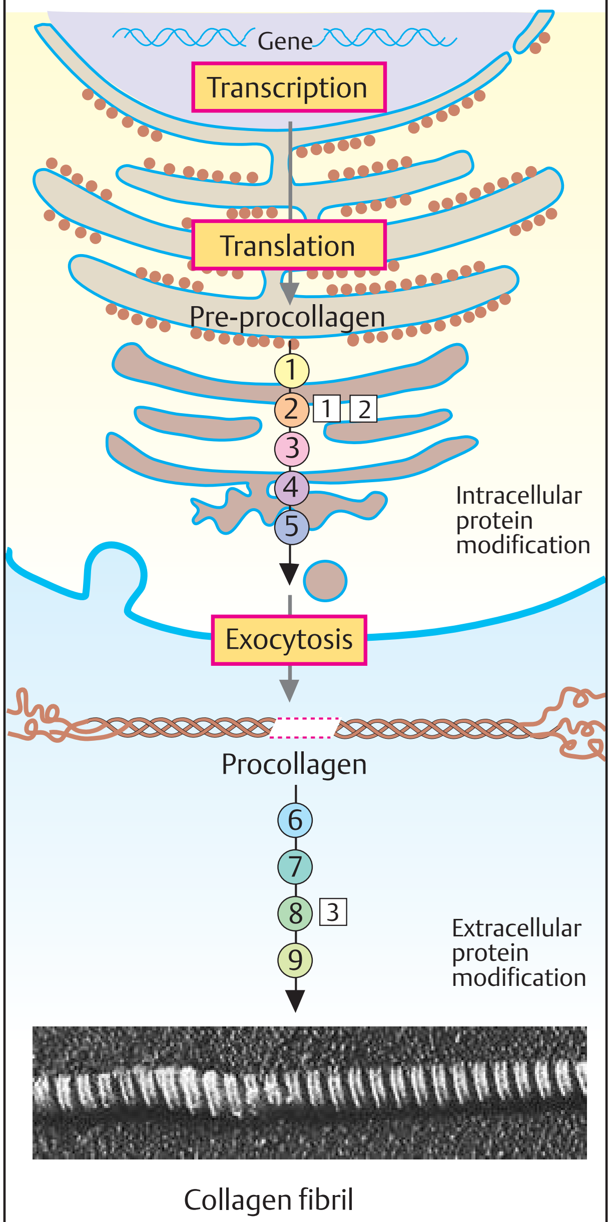
Gene
Transcription
Translation
Pre-procollagen
1
2
1
2
3
4
5
Intracellular protein modification
Exocytosis
Procollagen
6
7
8
3
9
Extracellular protein modification
Collagen fibril

1. Removal of the prepeptide
2. Hydroxylation of Pro and Lys residues
3. Glycosylation of 5Hyl and Asn
4. Oxidation of Cys in propeptides
5. Assemblage to form triple helix
6. Removal of the propeptide
7. Staggered deposition to form fibrils
8. Oxidation of Lys and 5Hyl to aldehydes
9. Cross-linking to form supramolecules

[1] Procollagen-proline 4-dioxygenase *1.14.11.2* [ascorbate, Fe]

[2] Procollagen-lysine 5-dioxygenase *1.14.11.4* [ascorbate, Fe]

[3] Protein-lysine 6-oxidase *1.4.3.13* [Cu]

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Extracellular matrix

A. Extracellular matrix ◑

The space between the cells (the interstitium) is occupied by a substance with a complex composition known as the **extracellular matrix** (ECM). In many types of tissue—e. g., muscle and liver—the ECM is only a narrow border between the cells, while in others it forms a larger space. In *connective tissue, cartilage,* and *bone,* the ECM is particularly strongly marked and is actually the functional part of the tissue (see p. 340). The illustration shows the three main constituents of the extracellular matrix in a highly schematic way: collagen fibers, network-forming adhesive proteins, and space-filling proteoglycans.

The ECM has a very wide variety of functions: it establishes mechanical connections between cells; it creates structures with special mechanical properties (as in bone, cartilage, tendons, and joints); it creates filters (e. g., in the basal membrane in the renal corpuscles; see p. 322); it separates cells and tissues from each other (e. g., to allow the joints to move freely); and it provides pathways to guide migratory cells (important for embryonic development). The chemical composition of the ECM is just as diverse as its functions.

Collagens (see p. 344), of which there are at least 19 different varieties, form fibers, fibrils, networks, and ligaments. Their characteristic properties are *tensile strength* and *flexibility.* **Elastin** is a fiber protein with a high degree of elasticity.

Adhesive proteins provide the connections between the various components of the extracellular matrix. Important representatives include **laminin** and **fibronectin** (see **B**). These multifunctional proteins simultaneously bind to several other types of matrix component. Cells attach to the cell surface receptors in the ECM with the help of the adhesive proteins.

Due to their polarity and negative charge, **proteoglycans** (see **C**) bind water molecules and cations. As a homogeneous “cement,” they fill the gaps between the ECM fibers.

B. Fibronectins ○

Fibronectins are typical representatives of adhesive proteins. They are filamentous dimers consisting of two related peptide chains (each with a mass of 250 kDa) linked to each other by disulfide bonds. The fibronectin molecules are divided into different *domains,* which bind to cell-surface receptors, collagens, fibrin, and various proteoglycans. This is what gives fibronectins their *“molecular glue”* characteristics.

The domain structure in fibronectins is made up of a few types of *peptide module* that are repeated numerous times. Each of the more than 50 modules is coded for by one exon in the fibronectin gene. *Alternative splicing* (see p. 246) of the hnRNA transcript of the fibronectin gene leads to fibronectins with different compositions. The module that causes adhesion to cells contains the characteristic amino acid sequence –Arg–Gly–Asp–Ser–. It is these residues that enable fibronectin to bind to cell-surface receptors, known as **integrins**.

C. Proteoglycans ◑

Proteoglycans are giant molecule complexes consisting of carbohydrates (95%) and proteins (5%), with masses of up to 2 10^6 Da. Their bottlebrush-shaped structure is produced by an axis consisting of **hyaluronate**. This thread-like polysaccharide (see p. 44) has **proteins** attached to it, from which in turn long polysaccharide chains emerge. Like the central hyaluronate, these terminal polysaccharides belong to the glycosaminoglycan group (see p. 44).

The **glycosaminoglycans** are made up of repeating disaccharide units, each of which consists of one *uronic acid* (glucuronic acid or iduronic acid) and one *amino sugar* (*N*-acetylglucosamine or *N*-acetylgalactosamine) (see p. 38). Many of the amino sugars are also esterified with sulfuric acid (sulfated), further increasing their polarity. The proteoglycans bind large amounts of water and fill the gaps between the fibrillar components of the ECM in the form of a hydrated gel. This inhibits the spread of pathogens in the ECM, for example.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

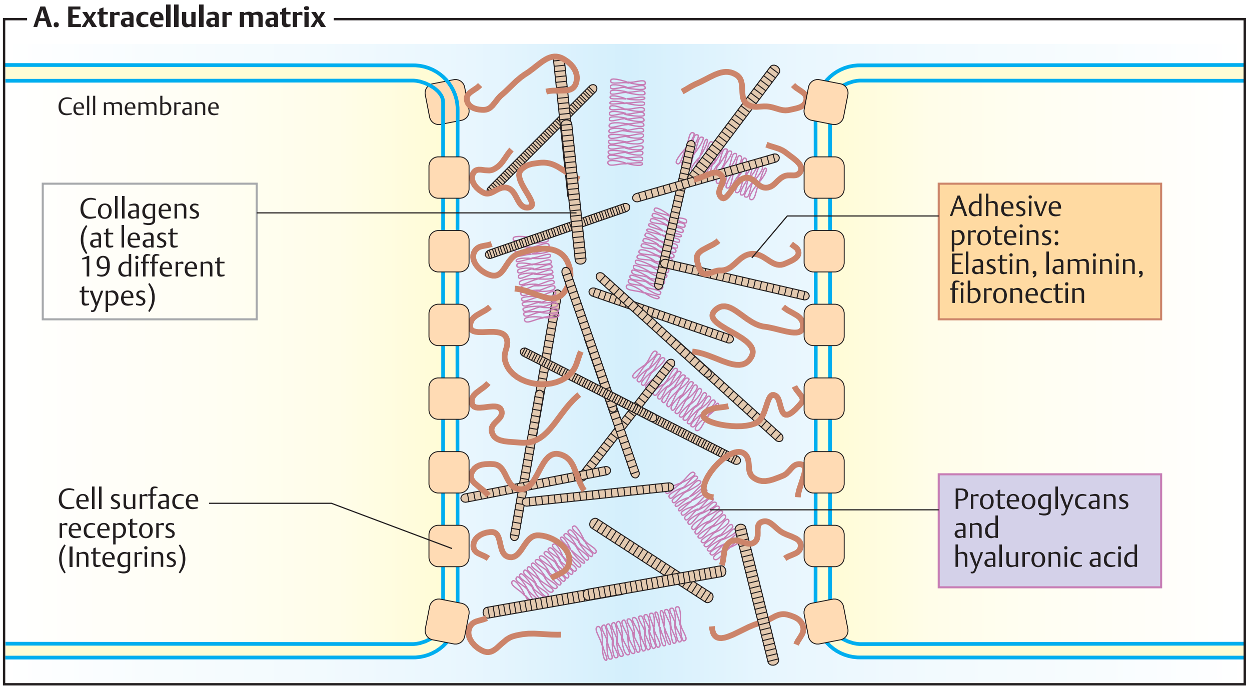

A. Extracellular matrix

Cell membrane
Collagens (at least 19 different types)
Adhesive proteins: Elastin, laminin, fibronectin
Cell surface receptors (Integrins)
Proteoglycans and hyaluronic acid

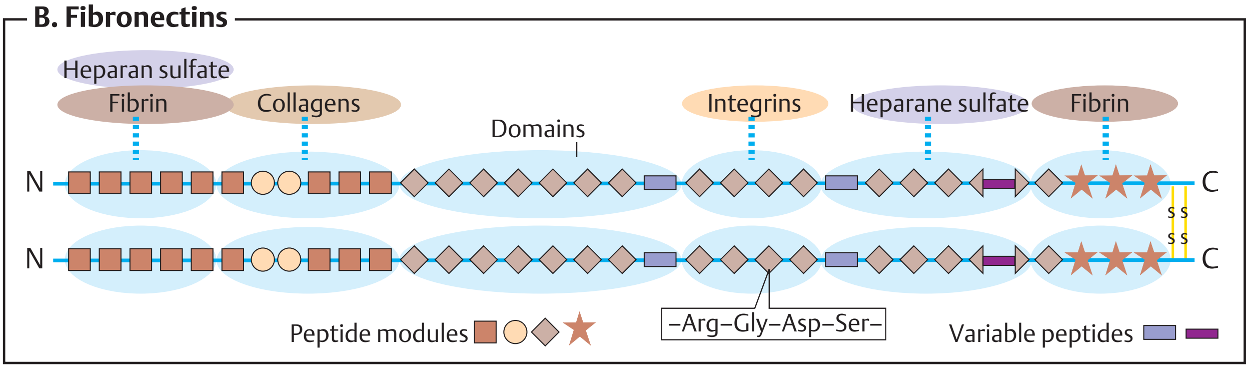

B. Fibronectins

Heparan sulfate
Fibrin
Collagens
Domains
Integrins
Heparane sulfate
Fibrin
N
C
N
C
S S
S S
Peptide modules
–Arg–Gly–Asp–Ser–
Variable peptides

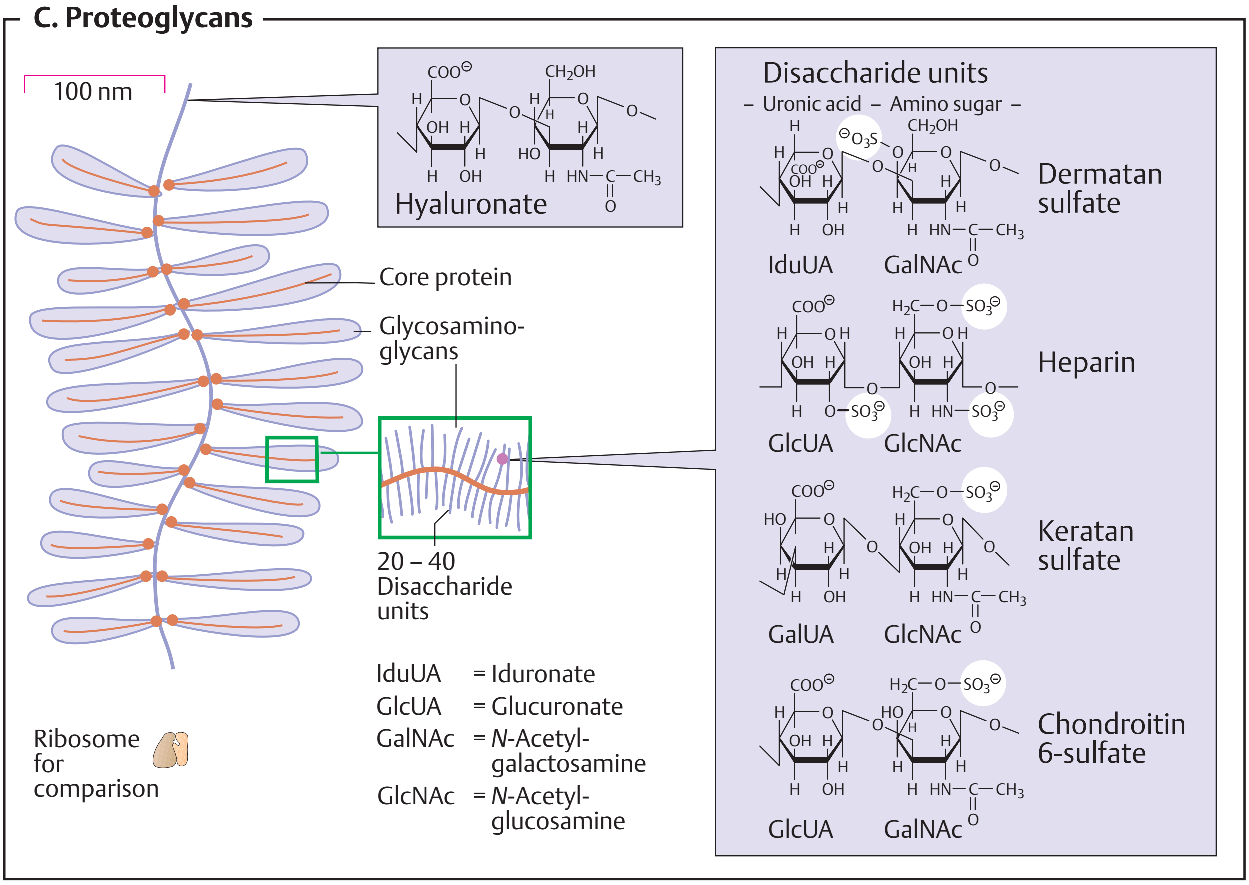

C. Proteoglycans

100 nm
Hyaluronate
Core protein
Glycosamino-glycans
20 – 40 Disaccharide units
Ribosome for comparison
Disaccharide units
– Uronic acid – Amino sugar –
Dermatan sulfate
IduUA
GalNAc
Heparin
GlcUA
GlcNAc
Keratan sulfate
GalUA
GlcNAc
Chondroitin 6-sulfate
GlcUA
GalNAc
IduUA = Iduronate
GlcUA = Glucuronate
GalNAc = N-Acetyl-galactosamine
GlcNAc = N-Acetyl-glucosamine

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Signal transmission in the CNS

A. Structure of nerve cells ●

Nerve cells (neurons) are easily excitable cells that produce electrical signals and can react to such signals as well. Their structure is markedly different from that of other types of cell. Numerous branching processes project from their **cell body** (soma). Neurons are able to receive signals via **dendrites** and to pass them on via **axons**. The axons, which can be up to 1 m long, are usually surrounded by *Schwann cells,* which cover them with a lipid-rich myelin sheath to improve their electrical insulation.

The transfer of stimuli occurs at the **synapses,** which link the individual neurons to each other as well as linking neurons functionally to muscle fibers. *Neurotransmitters* (see p. 352) are stored in the axonal nerve endings. These signaling substances are released in response to electrical signals in order to excite neighboring neurons (or muscle cells). It is estimated that each neuron in the brain is in contact via synapses with approximately 10 000 other neurons.

There is a noticeably high proportion of lipids in the composition of nerve cells, representing about 50% of their dry weight. In particular, there is a very wide variety of phospholipids, glycolipids, and sphingolipids (see p. 216).

B. Neurotransmitters and neurohormones ◑

Neurosecretions are classed into two groups: **neurotransmitters** are released into the *synaptic cleft* in order to influence neighboring cells (**C**). They have a short range and a short lifespan. By contrast, **neurohormones** are released into the blood, allowing them to cover larger distances. However, the distinction between the two groups is a fluid one; some neurotransmitters simultaneously function as neurohormones.

C. Synaptic signal transmission ◑

All chemical synapses function according to a similar principle. In the area of the synapse, the surface of the signaling cell (*presynaptic membrane*) is separated from the surface of the receiving cell (*postsynaptic membrane*) only by a narrow *synaptic cleft.* When an **action potential** (see p. 350) reaches the presynaptic membrane, *voltage-gated* Ca^{2+} *channels* integrated into the membrane open and trigger **exocytosis** of the neurotransmitter stored in the presynaptic cell (for details, see p. 228).

Each neuron usually releases *only one type* of **neurotransmitter**. Neurons that release dopamine are referred to as "dopaminergic," for example, while those that release acetylcholine are "cholinergic," etc. The transmitters that are released diffuse through the synaptic cleft and bind on the other side to **receptors** on the postsynaptic membrane. These receptors are integral membrane proteins that have binding sites for neurotransmitters on their exterior (see p. 224).

The receptors for neurotransmitters are divided into two large groups according to the effect produced by binding of the transmitter (for details, see p. 354).

Ionotropic receptors (bottom left) are *ligand-gated ion channels.* When they open as a result of the transmitter's influence, ions flow in due to the membrane potential (see p. 126). If the inflowing ions are cations (Na^+, K^+, Ca^{2+}), **depolarization** of the membrane occurs and an action potential is triggered on the surface of the postsynaptic cell. This is the way in which stimulatory transmitters work (e. g., acetylcholine and glutamate). By contrast, if anions flow in (mainly Cl^-), the result is **hyperpolarization** of the postsynaptic membrane, which makes the production of a postsynaptic action potential more difficult. The action of inhibitory transmitters such as glycine and GABA is based on this effect.

A completely different type of effect is observed in **metabotropic receptors** (bottom right). After binding of the transmitter, these interact on the inside of the postsynaptic membrane with *G proteins* (see p. 384), which in turn activate or inhibit the synthesis of **second messengers**. Finally, second messengers activate or inhibit **protein kinases**, which phosphorylate cellular proteins and thereby alter the behavior of the postsynaptic cells (*signal transduction;* see p. 386).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Structure of nerve cells

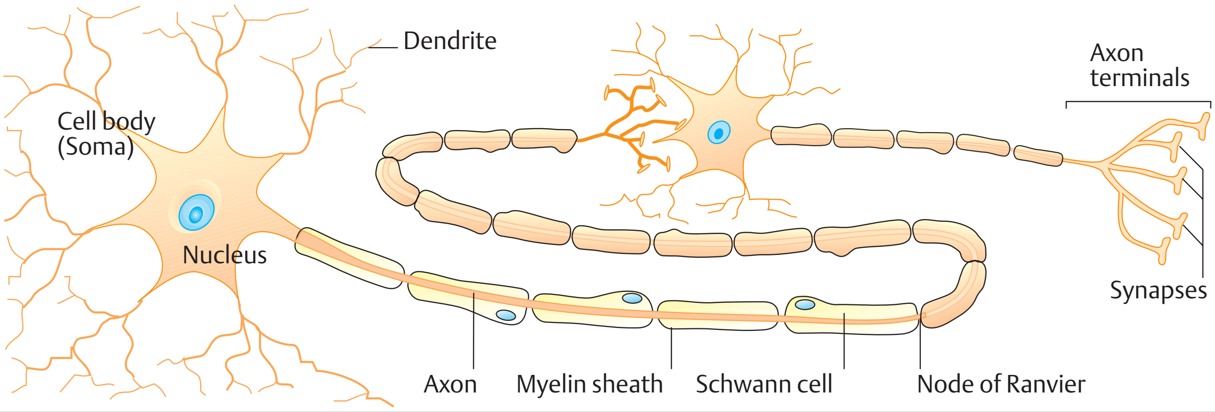
Dendrite
Axon terminals
Cell body (Soma)
Nucleus
Synapses
Axon
Myelin sheath
Schwann cell
Node of Ranvier

A. Neurotransmitters and neurohormones

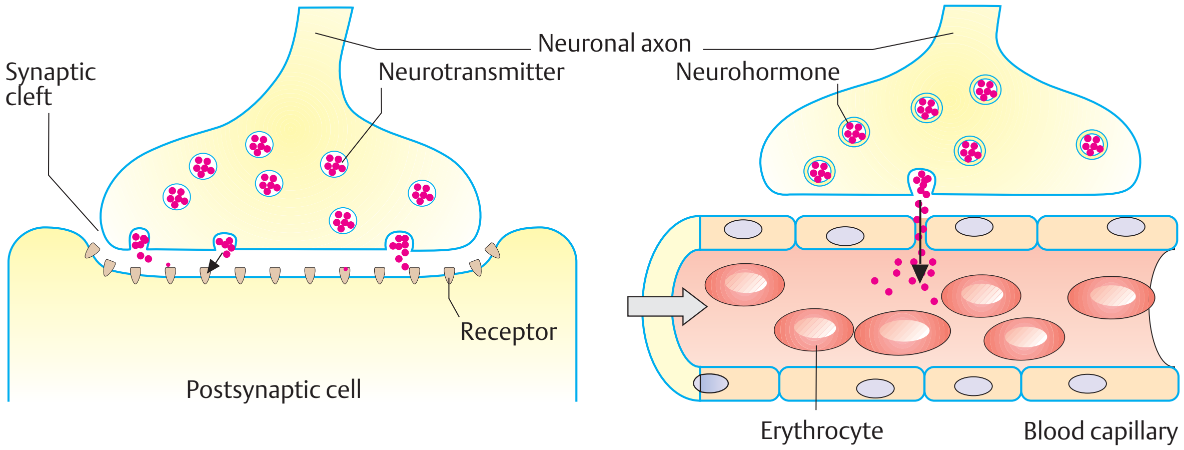
Neuronal axon
Synaptic cleft
Neurotransmitter
Neurohormone
Receptor
Postsynaptic cell
Erythrocyte
Blood capillary

C. Synaptic signal transmission

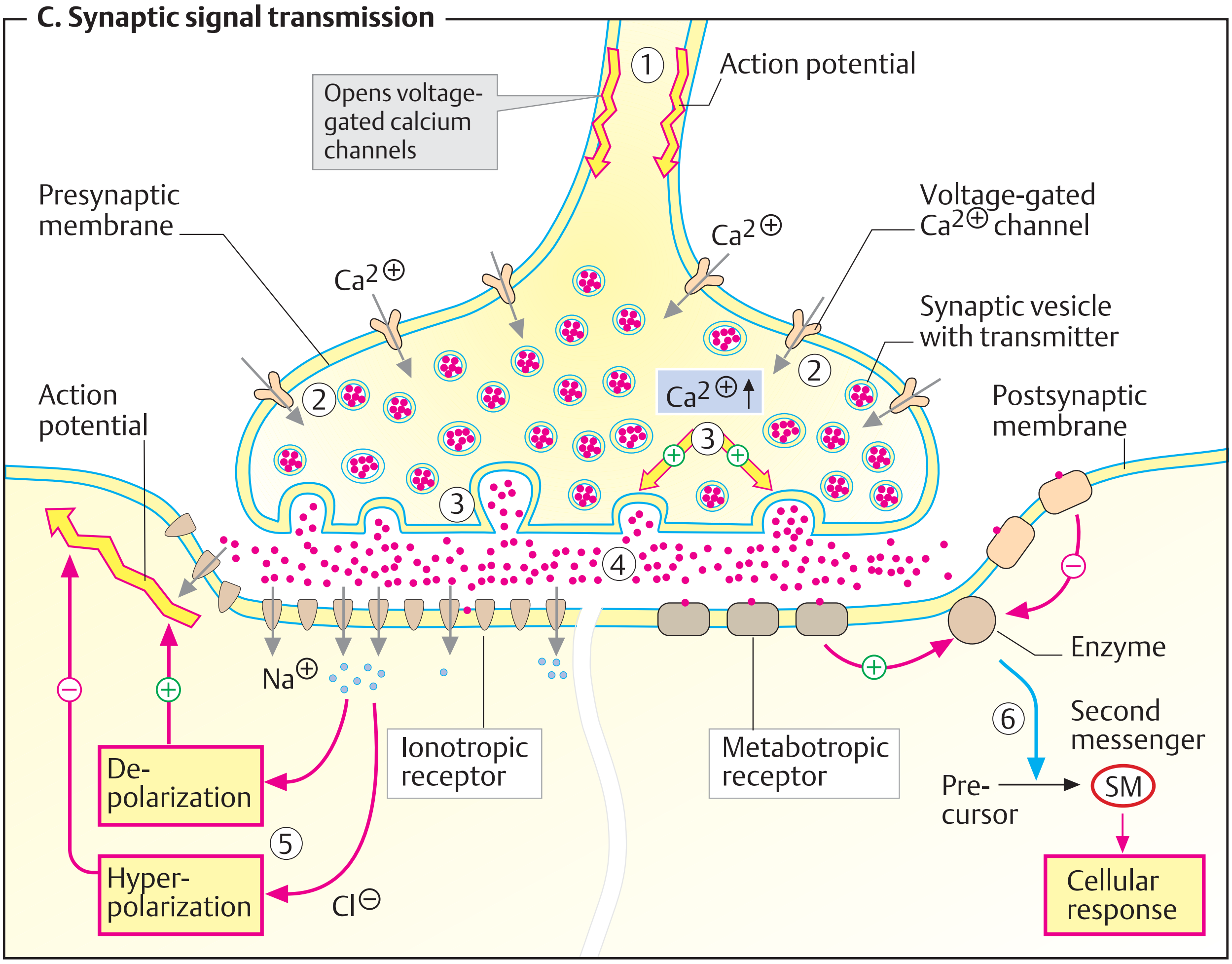
①
Action potential
Opens voltage-gated calcium channels
Presynaptic membrane
Voltage-gated Ca2⊕ channel
Ca2⊕
Ca2⊕
Synaptic vesicle with transmitter
2
2
Ca2⊕↑
3
3
Postsynaptic membrane
Action potential
4
Na⊕
Enzyme
Ionotropic receptor
Metabotropic receptor
6
Second messenger
Pre-cursor
SM
De-polarization
5
Hyper-polarization
Cl⊖
Cellular response

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Resting potential and action potential

A. Resting potential ◑

A characteristic property of living cells is the uneven distribution of positively and negatively charged ions on the inside and outside of the plasma membrane. This gives rise to a **membrane potential** (see p. 126)—i. e., there is electrical voltage between the two sides of the membrane, which can only balance out when *ion channels* allow the unevenly distributed ions to move.

At rest, the membrane potential in most cells is –60 to –90 mV. It mainly arises from the activity of *Na^+/K^+ transporting ATPase* ("Na^+/K^+ ATPase"), which occurs on practically all animal cells. Using up ATP, this P-type enzyme (see p. 220) "pumps" three Na^+ ions out of the cell in exchange for two K^+ ions. Some of the K^+ ions, following the concentration gradient, leave the cell again through *potassium channels*. As the protein anions that predominate inside the cell cannot follow them, and inflow of Cl^- ions from the outside is not possible, the result is an excess of positive charges outside the cell, while anions predominate inside it.

An **equilibrium potential** exists for each of the ions involved. This is the value of the membrane potential at which there is no net inflow or outflow of the ions concerned. For K^+ ions, the resting potential lies in the range of the membrane potential, while for Na^+ ions it is much higher at +70 mV. At the first opportunity, Na^+ ions will therefore spontaneously flow into the cell. The occurrence of action potentials is based on this (see **B**).

Nerve cell membranes contain **ion channels** for Na^+, K^+, Cl^-, and Ca^{2+}. These channels are usually closed and open only briefly to let ions pass through. They can be divided into channels that are regulated by membrane potentials ("*voltage-gated*"—e. g., fast Na^+ channels; see p. 222) and those regulated by ligands *("ligand-gated"*—e. g., nicotinic acetylcholine receptors; see p. 222).

B. Action potential ◑

Action potentials are special signals that are used to transmit information in the nervous system. They are triggered by chemical stimuli (or more rarely electrical stimuli). Binding of a neurotransmitter to an ionotropic receptor results in a brief local increase in the membrane potential from –60 mV to about +30 mV. Although the membrane potential quickly returns to the initial value within a few milliseconds (ms) at its site of origin, the depolarization is propagated because neighboring membrane areas are activated during this time period.

[1] The process starts with the opening of voltage-gated Na^+ channels (see p. 222). Due to their high equilibrium potential (see **A**), Na^+ ions flow into the cell and reverse the local membrane potential (**depolarization**).

[2] The Na^+ channels immediately close again, so that the inflow of positive charges is only very brief.

[3] Due to the increase in the membrane potential, voltage-dependent K^+ channels open and K^+ ions flow out. In addition, Na^+/K^+ ATPase (see **A**) pumps the Na+ ions that have entered back out again. This leads to **repolarization** of the membrane.

[4] The two processes briefly lead to the charge even falling below the resting potential (**hyperpolarization**). The K^+ channels also close after a few milliseconds. The nerve cell is then ready for re-stimulation.

Generally, it is always only a very small part of the membrane that is depolarized during an action potential. The process can therefore be repeated again after a short refractory period, when the nerve cell is stimulated again. Conduction of the action potential on the surface of the nerve cell is based on the fact that the local increase in the membrane potential causes neighboring voltage-gated ion channels to open, so that the membrane stimulation spreads over the whole cell in the form of a *depolarization wave*.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Resting potential

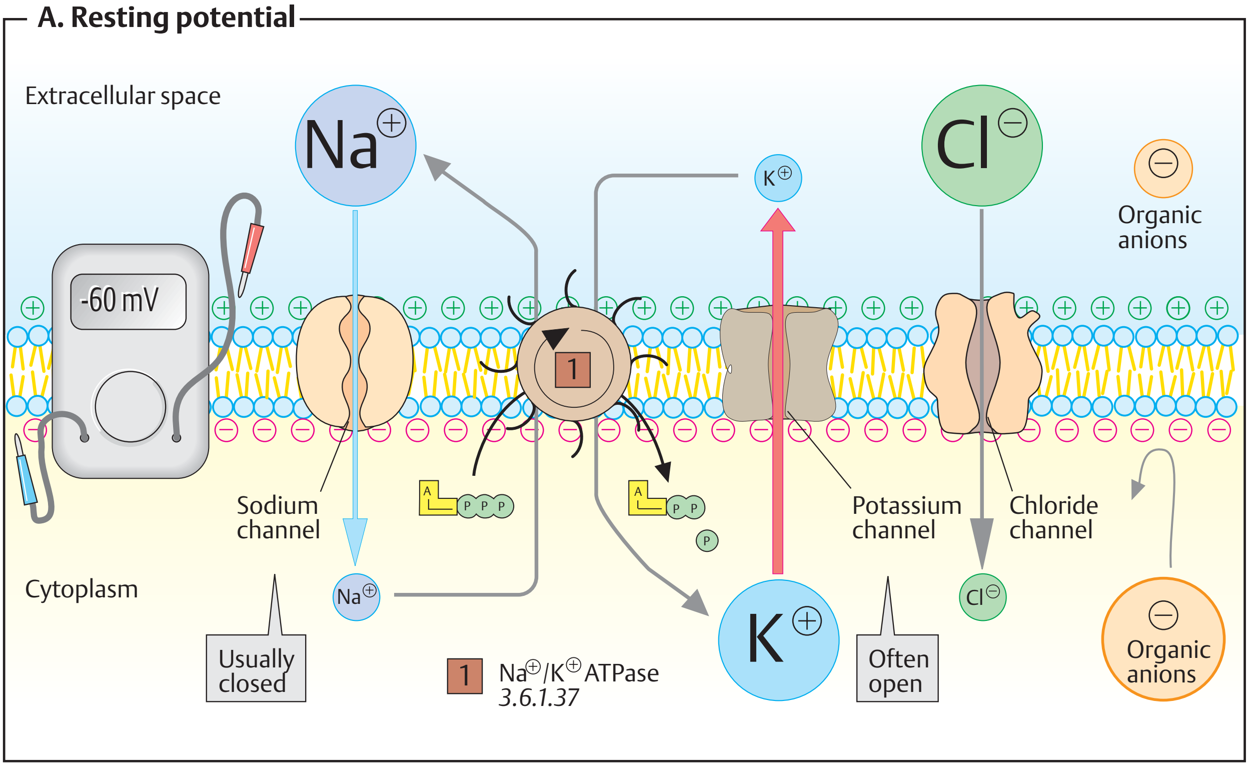
Extracellular space
Na⊕
K⊕
Cl⊖
Organic anions
-60 mV
Sodium channel
Potassium channel
Chloride channel
Cytoplasm
Usually closed
Often open
Na⊕
K⊕
Cl⊖
Organic anions
1
Na⊕/K⊕ATPase 3.6.1.37

B. Action potential

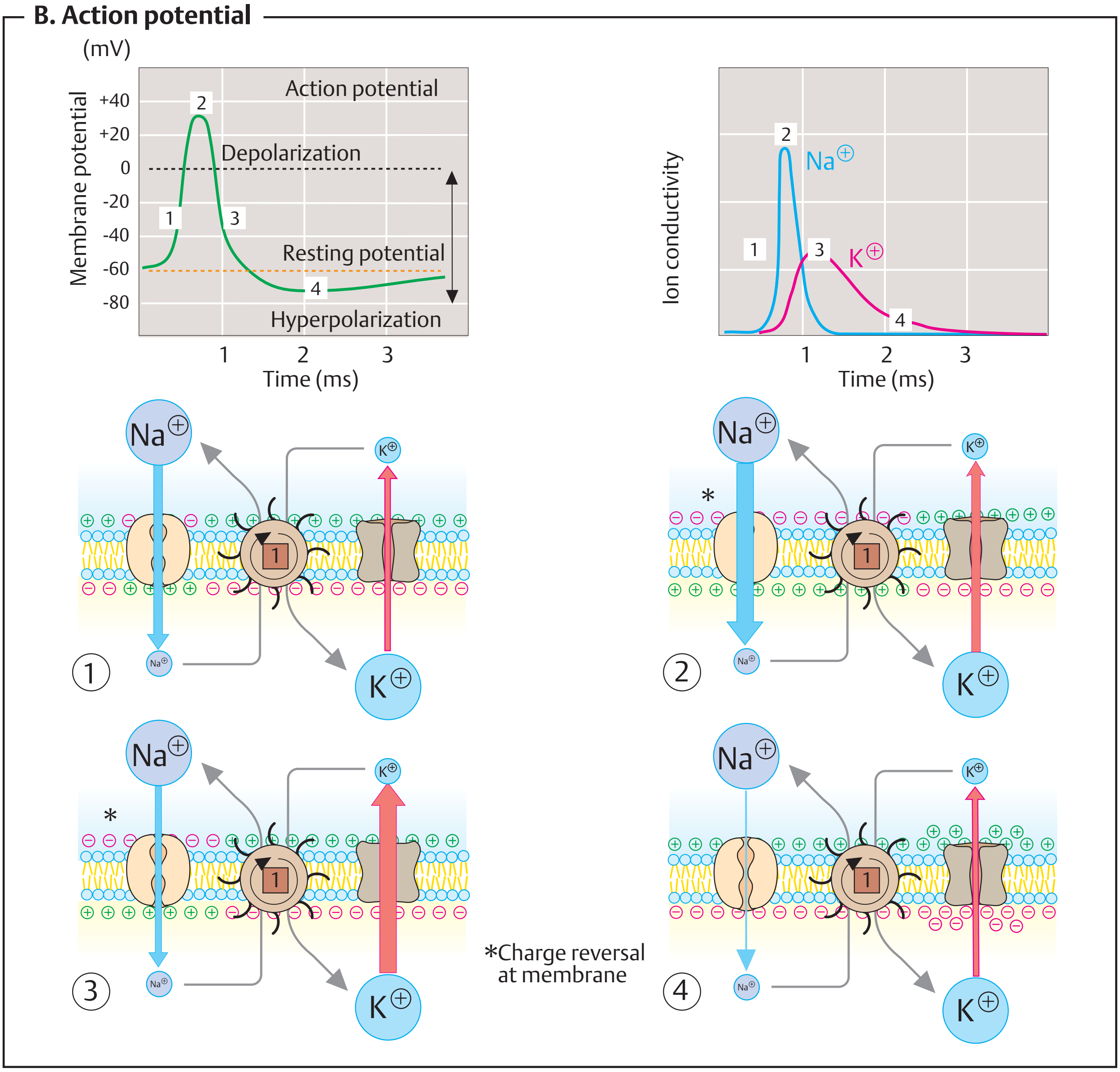
(mV)
Membrane potential
+40
+20
0
-20
-40
-60
-80
Action potential
Depolarization
Resting potential
Hyperpolarization
1
2
3
4
Time (ms)
Ion conductivity
Na⊕
K⊕
1
2
3
4
Time (ms)
①
②
③
④
Na⊕
K⊕
*
*Charge reversal at membrane

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Neurotransmitters

Neurotransmitters in the strict sense are substances that are produced by neurons, stored in the synapses, and released into the synaptic cleft in response to a stimulus. At the postsynaptic membrane, they bind to special receptors and affect their activity.

A. Important neurotransmitters ◑

Neurotransmitters can be classified into several groups according to their chemical structure. The table lists the most important representatives of this family, which has more than 100 members.

Acetylcholine, the acetic acid ester of the cationic alcohol choline (see p. 50) acts at neuromuscular junctions, where it triggers muscle contraction (see p. 334), and in certain parts of the brain and in the autonomous nervous system.

Several proteinogenic **amino acids** (see p. 60) have neurotransmitter effects. A particularly important one is *glutamate,* which acts as a stimulatory transmitter in the CNS. More than half of the synapses in the brain are glutaminergic. The metabolism of glutamate and that of the amine GABA synthesized from it (see below) are discussed in more detail on p. 356. *Glycine* is an inhibitory neurotransmitter with effects in the spinal cord and in parts of the brain.

Biogenic amines arise from amino acids by decarboxylation (see p. 62). This group includes *4-aminobutyrate* (γ-aminobutyric acid, GABA), which is formed from glutamate and is the most important inhibitory transmitter in the CNS. The *catecholamines* norepinephrine and epinephrine (see **B**), *serotonin,* which is derived from tryptophan, and *histamine* also belong to the biogenic amine group. All of them additionally act as hormones or mediators (see p. 380).

Peptides make up the largest group among the neurosecretions. Many peptide hormones—e. g., thyroliberin (TRH) and angiotensin II—simultaneously act as transmitters. Most neuropeptides are small (3–15 AA). At their N-terminus, many of them have a glutamate residue that has been cyclized to form *pyroglutamate* (5-oxoproline, <G), while the C-terminus is often an acid amide ($-NH_2$). This provides better protection against breakdown by peptidases.

Endorphins, dynorphins, and *enkephalins* are a particularly interesting group of neuropeptides. They act as "endogenous opiates" by producing analgetic, sedative, and euphoriant effects in extreme situations. Drugs such as morphine and heroin activate the receptors for these peptides (see p. 354).

Purine derivatives with neurotransmitter function are all derived from adenine-containing nucleotides or nucleosides. ATP is released along with acetylcholine and other transmitters, and among other functions it regulates the emission of transmitters from its synapse of origin. The stimulatory effect of *caffeine* is mainly based on the fact that it binds to adenosine receptors.

B. Biosynthesis of catecholamines ◑

The catecholamines are *biogenic amines* that have a catechol group. Their biosynthesis in the adrenal cortex and CNS starts from **tyrosine**.

[1] Hydroxylation of the aromatic ring initially produces **dopa** (3,4-dihydroxyphenylalanine). This reaction uses the unusual coenzyme *tetrahydrobiopterin* (THB). Dopa (cf. p. 6) is also used in the treatment of Parkinson's disease.

[2] Decarboxylation of dopa yields **dopamine**, an important transmitter in the CNS. In dopaminergic neurons, catecholamine synthesis stops at this point.

[3] The adrenal gland and adrenergic neurons continue the synthesis by hydroxylating dopamine into **norepinephrine** (noradrenaline). *Ascorbic acid* (vitamin C; see p. 368) acts as a hydrogen-transferring coenzyme here.

[4] Finally, *N*-methylation of norepinephrine yields **epinephrine** (adrenaline). The coenzyme for this reaction is *S*-adenosylmethionine (SAM; see p. 110).

The physiological effects of the catecholamines are mediated by a large number of different receptors that are of particular interest in pharmacology. Norepinephrine acts in the autonomic nervous system and certain areas of the brain. Epinephrine is also used as a transmitter by some neurons.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Important neurotransmitters

Acetylcholine

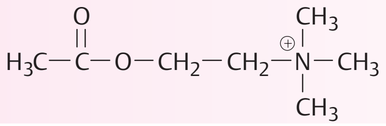

$$H_3C-\overset{O}{\overset{\|}{C}}-O-CH_2-CH_2-\overset{CH_3}{\overset{|}{\underset{\underset{CH_3}{|}}{\overset{\oplus}{N}}}}-CH_3$$

Amino acids:

Glutamate
Glycine
Dopa

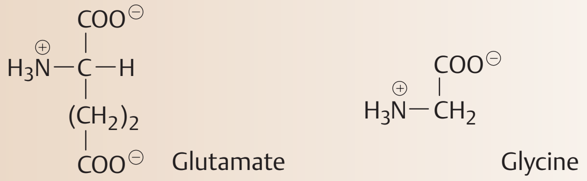
Glutamate
Glycine

Biogenic amines:

γ-Aminobutyrate (GABA)
Dopa
Dopamine
Norepinephrine
Epinephrine
Serotonine
Histamine

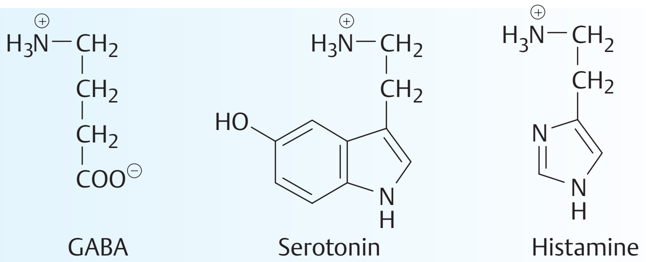
GABA
Serotonin
Histamine

Peptides:

β-Endorphin	YGGFMTSEKSQTPLVTLFKNAITKNAYKKGE
Met- and Leu-enkephalin	YGGFM und YGGFL
Thyroliberin (TRH)	<GHP-NH_2
Gonadoliberin (GnRH)	<GHWSYGLRPG-NH_2
Substance P	RPKPQQFFGLM
Somatostatin	AGCKNFFWKTFTSC
Angiotensin II	DRVYIHPF
Cholecystokinin (CCK-4)	WMDF-NH_2
and many others	

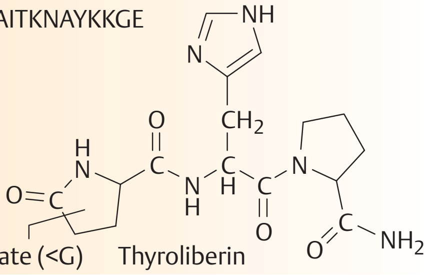
Pyroglutamate (<G)
Thyroliberin

Purine derivatives:

ATP
ADP
AMP
Adenosine

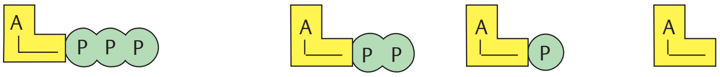
A
P P P
A
P P
A
P
A

B. Biosynthesis of the catecholamines

1 Tyrosine 3-monooxygenase [$Fe^{2\oplus}$,THB] *1.14.16.2*

2 Aromatic-L-amino-acid decarboxylase (Dopa decarboxylase) [PLP] *4.1.1.28*

3 Dopamine β-monooxygenase[Cu] *1.14.17.1*

4 Phenylethanolamine *N*-methyltransferase *2.1.1.28*

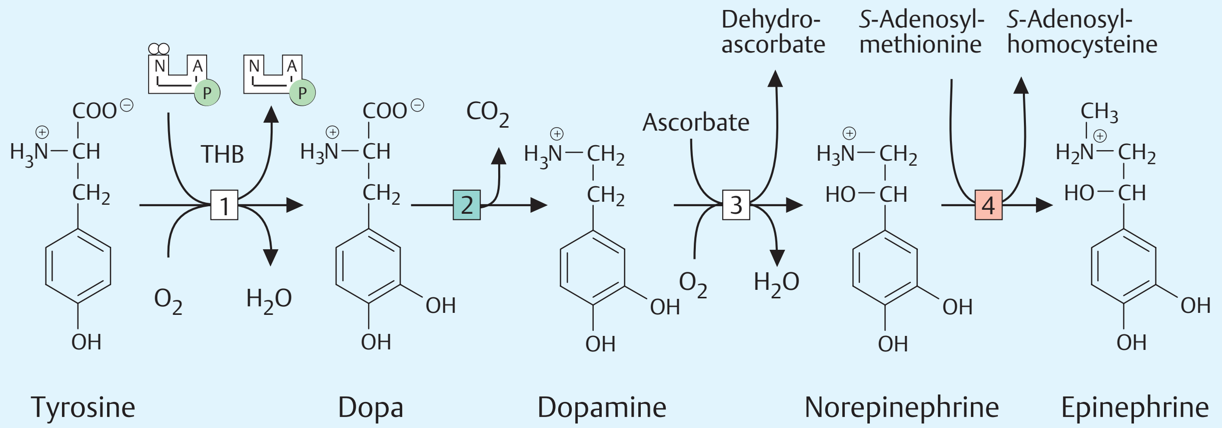
THB
1
O_2
H_2O
N
A
P
2
CO_2
Ascorbate
3
Dehydro-ascorbate
O_2
H_2O
S-Adenosyl-methionine
4
S-Adenosyl-homocysteine
Tyrosine
Dopa
Dopamine
Norepinephrine
Epinephrine

Receptors for neurotransmitters

Like all signaling substances, neurotransmitters (see p. 352) act via receptor proteins. The receptors for neurotransmitters are integrated into the membrane of the postsynaptic cell, where they trigger ion inflow or signal transduction processes (see p. 348).

A. Receptors for neurotransmitters ◑

A considerable number of receptors for neurotransmitters are already known and new ones are continuing to be discovered. The table only lists the most important examples. They are classified into two large groups according to their mode of action.

Ionotropic receptors are *ligand-gated ion channels* (left half of the table). The receptors for stimulatory transmitters (indicated in the table by a ⊕) mediate the inflow of cations (mainly Na^+). When these open after binding of the transmitter, local *depolarization* of the postsynaptic membrane occurs. By contrast, inhibitory neurotransmitters (GABA and glycine) allow Cl^- to flow in. This increases the membrane's negative resting potential and hinders the action of stimulatory transmitters (*hyperpolarization,* ⊖).

Metabotropic receptors (right half of the table) are coupled to G proteins (see p. 386), through which they influence the *synthesis of second messengers.* Receptors that work with type G_s proteins (see p. 386) increase the cAMP level in the postsynaptic cell ([cAMP] ↑), while those that activate G_i proteins reduce it ([cAMP] ↓). Via type G_q proteins, other receptors increase the intracellular Ca^{2+} concentration ($[Ca^{2+}]$ ↑).

There are several **receptor subtypes** for most neurotransmitters. These are distinguished numerically (e. g., D_1 to D_5) or are named after their agonists—i. e., after molecules experimentally found to activate the receptor. For example, one specific subtype of glutamate receptors reacts to NMDA (*N*-methyl-D-aspartate), while another subtype reacts to the compound AMPA, etc.

B. Acetylcholine receptors ◑

Acetylcholine (ACh) was the neurotransmitter first discovered, at the beginning of the last century. It binds to two types of receptor.

The **nicotinic ACh receptor** responds to the alkaloid *nicotine* contained in tobacco (many of the physiological effects of nicotine are based on this). The nicotinic receptor is ionotropic. Its properties are discussed in greater detail on p. 222.

The **muscarinic ACh receptors** (of which there are at least five subtypes) are metabotropic. Their name is derived from the alkaloid *muscarine,* which is found in the fly agaric mushroom (*Amanita muscaria*), for example. Like ACh, muscarine is bound at the receptor, but in contrast to ACh (see **C**), it is not broken down and therefore causes permanent stimulation of muscle.

The muscarinic ACh receptors influence the cAMP level in the postsynaptic cells (M_1, M_3 and M_5 increase it, while subtypes M_2 and M_4 reduce it).

C. Metabolism of acetylcholine ◑

Acetylcholine is synthesized from acetyl-CoA and choline in the cytoplasm of the presynaptic axon [1] and is stored in **synaptic vesicles**, each of which contains around 1000–10 000 ACh molecules. After it is released by exocytosis (see p. 228), the transmitter travels by diffusion to the receptors on the postsynaptic membrane. Catalyzed by *acetylcholinesterase,* hydrolysis of ACh to acetate and choline immediately starts in the synaptic cleft [2], and within a few milliseconds, the ACh released has been eliminated again. The cleavage products **choline** and **acetate** are taken up again by the presynaptic neuron and reused for acetylcholine synthesis [3].

Substances that block the serine residue in the active center of acetylcholinesterase [2]—e. g., the neurotoxin E605 and other *organophosphates*—prevent ACh degradation and thus cause prolonged stimulation of the postsynaptic cell. This impairs nerve conduction and muscle contraction. *Curare,* a paralyzing arrow-poison used by South American Indians, competitively inhibits binding of ACh to its receptor.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Receptors for neurotransmitters

Ionotropic				Metabotropic		
Receptor	**Transmitter**	**Ion(s)**	**Effect**	**Receptor**	**Transmitter**	**Effect**
Acetylcholine (nicotinic)	Acetylcholine	$Na^{\oplus}$	$\oplus$	Acetylcholine (muscarinic) M1, M3, M5, M2, M4	Acetylcholine	$[Ca^{2\oplus}]\uparrow$ $[cAMP]\downarrow$
5HT3	Serotonin	$Na^{\oplus}$	$\oplus$	$5HT_1$ $5HT_2$ $5HT_4$	Serotonin „ „	$[Ca^{2\oplus}]\uparrow$ $[cAMP]\uparrow$ $[cAMP]\downarrow$
GABAa	GABA	$Cl^{\ominus}$	$\ominus$	α_1 α_2 $\beta_1, \beta_2, \beta_3$	Norepinephrine „ „	$[Ca^{2\oplus}]\uparrow$ $[cAMP]\uparrow$ $[cAMP]\downarrow$
Glycine	Glycine	$Cl^{\ominus}$	$\ominus$			
AMPA NMDA Kainate	Glutamate Glutamate Glutamate	$Na^{\oplus}$ $K^{\oplus}$ $Na^{\oplus}$ $K^{\oplus}$ $Ca^{2\oplus}$ $Na^{\oplus}$ $K^{\oplus}$	$\oplus$ $\oplus$ $\oplus$	D_1, D_5 D_2, D_3, D_4	Dopamine „	$[cAMP]\uparrow$ $[cAMP]\downarrow$
				δ, κ, μ	Opioids	$[cAMP]\downarrow$

B. Acetylcholine receptors

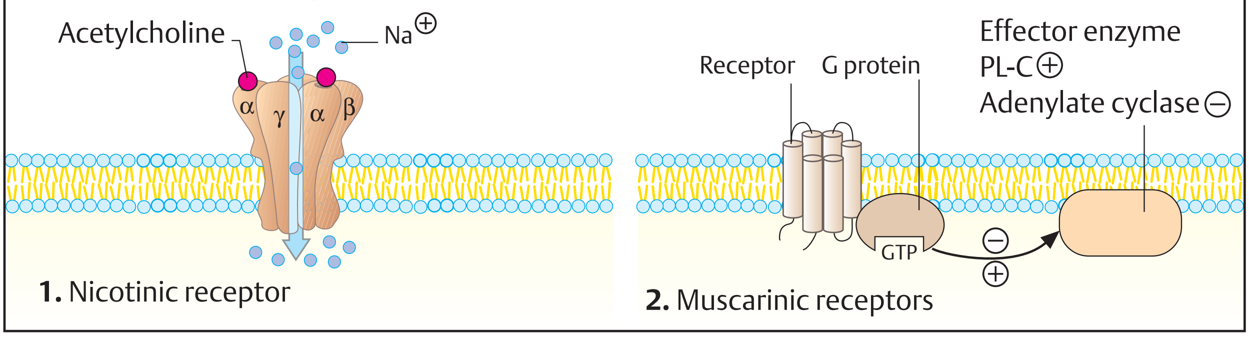
Acetylcholine
$Na^{\oplus}$
α
γ
α
β
Receptor
G protein
Effector enzyme
PL-C ⊕
Adenylate cyclase ⊖
GTP
⊖
⊕

1. Nicotinic receptor

2. Muscarinic receptors

C. Metabolism of acetylcholine

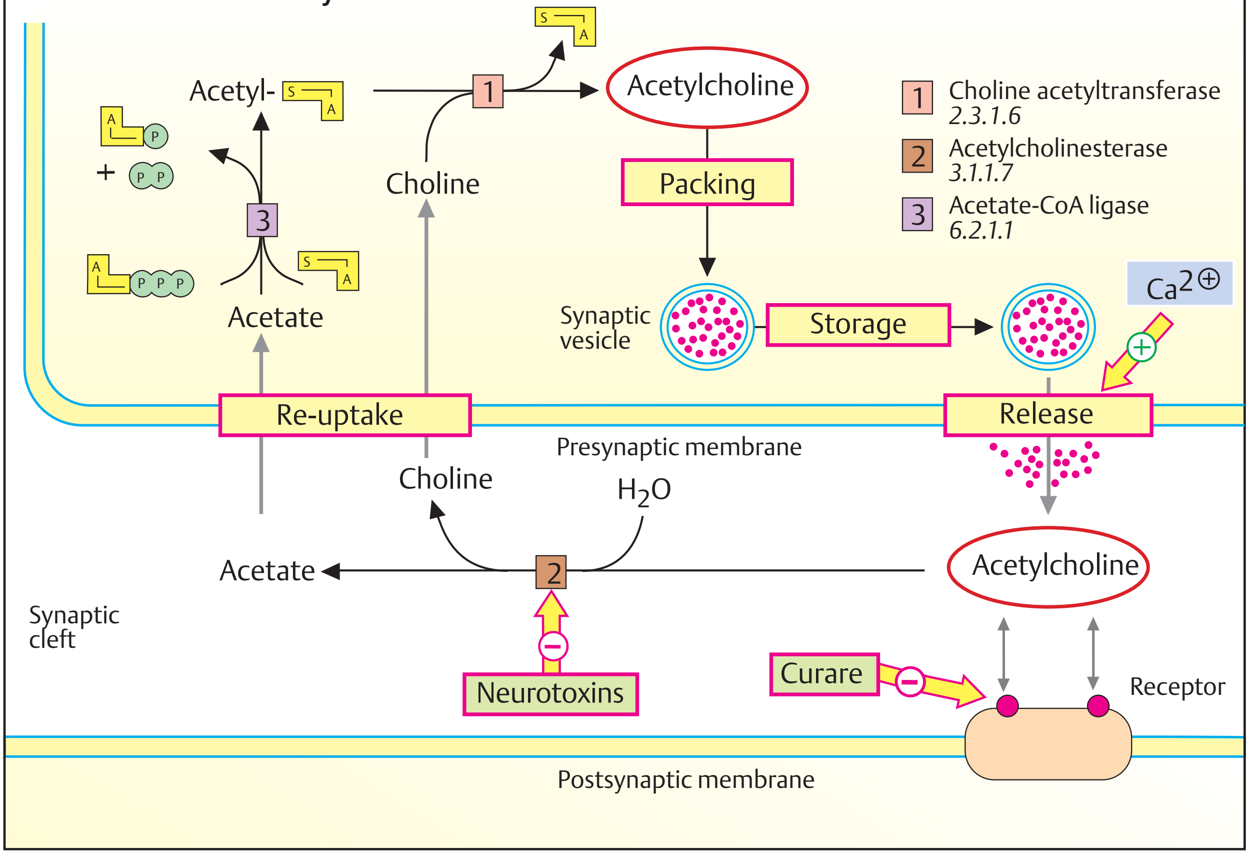
Acetyl-
S A
A P
+ P P
A P P P
3
Acetate
Choline
1
Acetylcholine
Packing
Synaptic vesicle
Storage
1 Choline acetyltransferase 2.3.1.6
2 Acetylcholinesterase 3.1.1.7
3 Acetate-CoA ligase 6.2.1.1
$Ca^{2\oplus}$
+
Re-uptake
Release
Presynaptic membrane
Choline
H_2O
Acetate
2
Acetylcholine
Synaptic cleft
–
Neurotoxins
Curare
–
Receptor
Postsynaptic membrane

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Metabolism

The brain and other areas of the central nervous system (CNS) have high ATP requirements. Although the brain only represents about 2% of the body's mass, it consumes around 20% of the metabolized oxygen and ca. 60% of the glucose. The neurons' high energy requirements are mainly due to ATP-dependent ion pumps (particularly Na^+/K^+ ATPase) and other active transport processes that are needed for nerve conduction (see p. 350).

A. Energy metabolism of the brain ●

Glucose is normally the only metabolite from which the brain is able to obtain adequate amounts of ATP through aerobic glycolysis and subsequent terminal oxidation to CO_2 and H_2O. Lipids are unable to pass the **blood–brain barrier**, and amino acids are also only available in the brain in limited quantities (see **B**). As neurons only have minor glycogen reserves, they are dependent on a constant supply of glucose from the blood. A severe drop in the blood glucose level—as can occur after insulin overdosage in diabetics, for example—rapidly leads to a drop in the ATP level in the brain. This results in loss of consciousness and neurological deficits that can lead to death. Oxygen deficiency (hypoxia) also fint affects the brain. The effects of a brief period of hypoxia are still reversible, but as time progresses irreversible damage increasingly occurs and finally complete loss of function ("brain death").

During periods of starvation, the brain after a certain time acquires the ability to use **ketone bodies** (see p. 312) in addition to glucose to form ATP. In the first weeks of a starvation period, there is a strong increase in the activities of the enzymes required for this in the brain. The degradation of ketone bodies in the CNS saves glucose and thereby reduces the breakdown of muscle protein that maintains gluconeogenesis in the liver during starvation. After a few weeks, the extent of muscle breakdown therefore declines to one-third of the initial value.

B. Glutamate, glutamine, and GABA ◑

The proteinogenic amino acid **glutamate** (Glu) and the biogenic amine **4-aminobutyrate** derived from it are among the most important neurotransmitters in the brain (see p. 352). They are both synthesized in the brain itself. In addition to the neurons, which use Glu or GABA as transmitters, *neuroglia* are also involved in the metabolism of these substances.

Since glutamate and GABA as transmitters must not appear in the extracellular space in an unregulated way, the cells of the neuroglia (center) supply "glutaminergic" and "GABAergic" neurons with the precursor **glutamine** (Gln), which they produce from glutamate with the help of *glutamine synthetase* [1].

GABA neurons (left) and glutamate neurons (right) initially hydrolyze glutamine with the help of *glutaminase* [1] to form glutamate again. The glutamate neurons store this in vesicles and release it when stimulated. The GABA neurons continue the degradation process by using *glutamate decarboxylase* [3] to convert glutamate into the transmitter GABA.

Both types of neuron take up their transmitter again. Some of it also returns to the neuroglia, where glutamate is amidated back into glutamine.

Glutamate can also be produced again from GABA. The reaction sequence needed for this, known as the **GABA shunt**, is characteristic of the CNS. A *transaminase* [4] first converts GABA and 2-oxoglutarate into glutamate and succinate semialdehyde ($^-OOC–CH_2–CH_2–CHO$). In an NAD^+-dependent reaction, the aldehyde is oxidized to succinic acid [5], from which 2-oxoglutarate can be regenerated again via tricarboxylic acid cycle reactions.

The function of glutamate as a stimulatory transmitter in the brain is the cause of what is known as the "Chinese restaurant syndrome." In sensitive individuals, the monosodium glutamate used as a flavor enhancer in Chinese cooking can raise the glutamate level in the brain to such an extent that transient mild neurological disturbances can occur (dizziness, etc.).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Energy metabolism of the brain

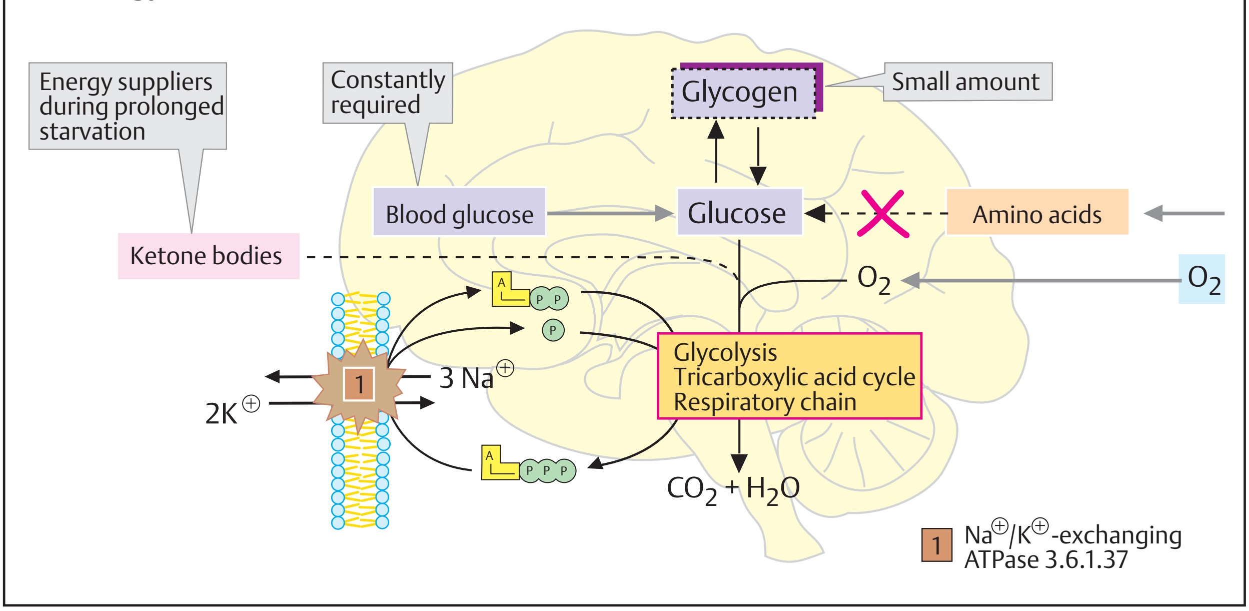
Energy suppliers during prolonged starvation
Constantly required
Glycogen
Small amount
Ketone bodies
Blood glucose
Glucose
Amino acids
O_2
O_2
$2K^+$
$3\ Na^+$
1
Glycolysis
Tricarboxylic acid cycle
Respiratory chain
$CO_2 + H_2O$
A
P P
P
P P P
1 Na^+/K^+-exchanging ATPase 3.6.1.37

B. Glutamate, glutamine, and GABA

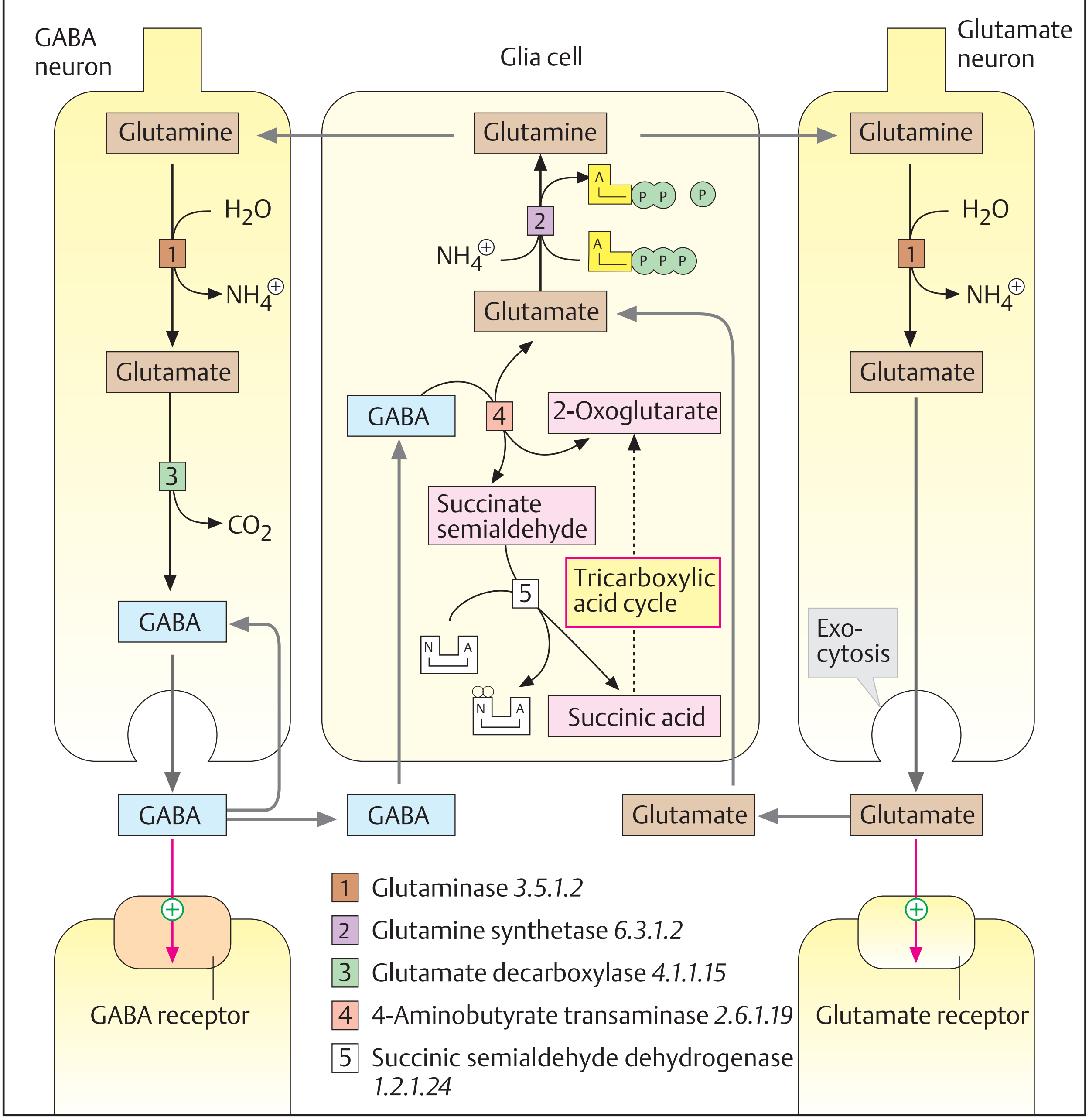
GABA neuron
Glia cell
Glutamate neuron
Glutamine
Glutamine
Glutamine
H_2O
H_2O
1
1
NH_4^+
NH_4^+
NH_4^+
2
A
P P
P
P P P
Glutamate
Glutamate
Glutamate
3
CO_2
GABA
4
2-Oxoglutarate
Succinate semialdehyde
5
N A
Tricarboxylic acid cycle
Succinic acid
Exo-cytosis
GABA
GABA
Glutamate
Glutamate
GABA receptor
Glutamate receptor
1 Glutaminase 3.5.1.2
2 Glutamine synthetase 6.3.1.2
3 Glutamate decarboxylase 4.1.1.15
4 4-Aminobutyrate transaminase 2.6.1.19
5 Succinic semialdehyde dehydrogenase 1.2.1.24

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Sight

Two types of photoreceptor cell are found in the human retina—*rods* and *cones.* Rods are sensitive to low levels of light, while the cones are responsible for color vision at higher light intensities.

Signaling substances and many proteins are involved in visual processes. Initially, a **light-induced** *cis–trans* **isomerization** of the pigment retinal triggers a conformational change in the membrane protein *rhodopsin.* Via the G protein *transducin,* which is associated with rhodopsin, an enzyme is activated that breaks down the second messenger *cGMP.* Finally, the cGMP deficiency leads to *hyperpolarization* of the light-sensitive cell, which is registered by subsequent neurons as *reduced neurotransmitter release.*

A. Photoreceptor ◑

The cell illustrated opposite, a **rod**, has a structure divided by membrane discs into which the 7-helix receptor **rhodopsin** is integrated (see p. 224). In contrast to other receptors in the 7-helix class (see p. 384), rhodopsin is a light-sensitive *chromoprotein.* Its protein part, **opsin**, contains the aldehyde **retinal** (see p. 364)—an isoprenoid which is bound to the ε-amino group of a lysine residue as an *aldimine.*

The light absorption of rhodopsin is in the visible range, with a maximum at about 500 nm. The absorption properties of the visual pigment are thus optimally adjusted to the spectral distribution of sunlight.

Absorption of a photon triggers isomerization from the 11-*cis* form of retinal to all-*trans*-retinal (top right). Within milliseconds, this *photochemical process* leads to an allosteric conformational change in rhodopsin. The active conformation (**rhodopsin***) binds and activates the G protein **transducin**. The *signal cascade* (**B**) that now follows causes the rod cells to release less neurotransmitter (glutamate) at their synapses. The adjoining bipolar neurons register this change and transmit it to the brain as a signal for light.

There are several different rhodopsins in the **cones**. All of them contain retinal molecules as light-sensitive components, the absorption properties of which are modulated by the different proportions of opsin they contain in such a way that colors can also be perceived.

B. Signal cascade ◑

Dark (bottom left). Rod cells that are not exposed to light contain relatively high concentrations (70 μM) of the cyclic nucleotide **cGMP** (3′,5′-cycloGMP; cf. cAMP, p. 386), which is synthesized by a *guanylate cyclase* ([2], see p. 388). The cGMP binds to an ion channel in the rod membrane (bottom left) and thus keeps it open. The inflow of cations (Na^+, Ca^{2+}) depolarizes the membrane and leads to release of the neurotransmitter glutamate at the synapse (see p. 356).

Light (bottom right). When the G protein transducin binds to light-activated rhodopsin* (see **A**, on the structure of the complex; see p. 224), it leads to the GDP that is bound to the transducin being exchanged for GTP. In transducin* that has been activated in this way, the GTP-containing α-subunit breaks off from the rest of the molecule and in turn activates a membrane *cGMP phosphodiesterase* [1]. This hydrolyzes cGMP to GMP and thus reduces the level of free cGMP within milliseconds. As a consequence, the cGMP bound at the ion channel dissociates off and the channel closes. As cations are constantly being pumped out of the cell, the membrane potential falls and **hyperpolarization** of the cell occurs, which interrupts glutamate release.

Regeneration. After exposure to light, several processes restore the initial conditions:

1. The α-subunit of transducin* inactivates itself by GTP hydrolysis and thus terminates the activation of cGMP esterase.
2. The reduced Ca^{2+} concentration causes activation of guanylate cyclase, which increases the cGMP level until the cation channels reopen.
3. An isomerase [3] transfers all-*trans* -retinal to the 11-*cis* -form, in which it is available for the next cycle. A dehydrogenase [4] can also allow retinal to be supplied from vitamin A (retinol).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

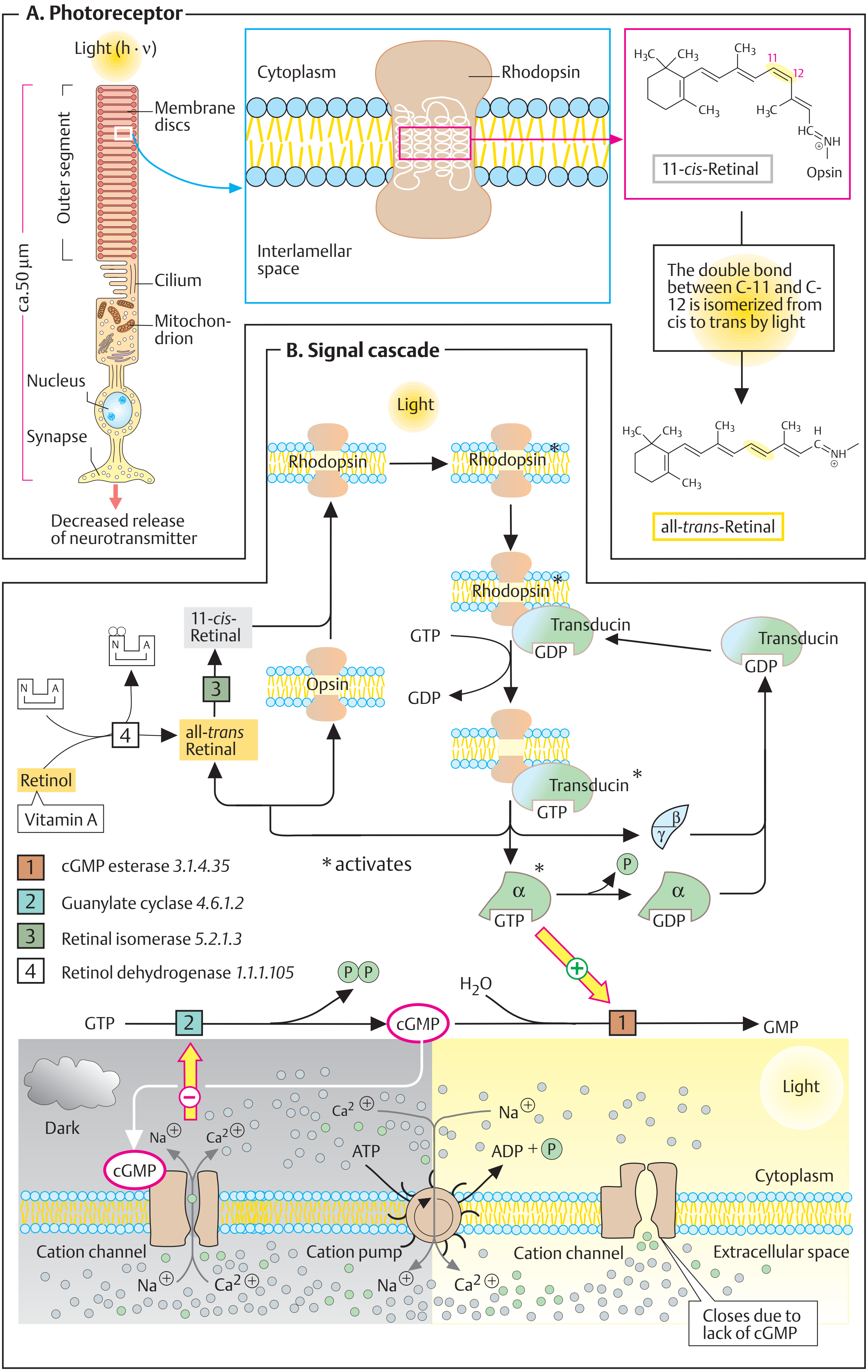
A. Photoreceptor
Light (h · ν)
ca. 50 µm
Outer segment
Membrane discs
Cilium
Mitochondrion
Nucleus
Synapse
Decreased release of neurotransmitter
Cytoplasm
Rhodopsin
Interlamellar space
H_3C
CH_3
11
12
HC
NH
Opsin
11-*cis*-Retinal
The double bond between C-11 and C-12 is isomerized from cis to trans by light
all-*trans*-Retinal
B. Signal cascade
Light
Rhodopsin
Rhodopsin*
Transducin
GDP
GTP
Transducin*
Opsin
11-*cis*-Retinal
all-*trans* Retinal
Retinol
Vitamin A
N
A
1
2
3
4
β
γ
α
P
* activates
1 cGMP esterase *3.1.4.35*
2 Guanylate cyclase *4.6.1.2*
3 Retinal isomerase *5.2.1.3*
4 Retinol dehydrogenase *1.1.1.105*
P P
H_2O
cGMP
GMP
Dark
Light
Na^+
Ca^{2+}
ATP
ADP + P
Cation channel
Cation pump
Cytoplasm
Extracellular space
Closes due to lack of cGMP

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Organic substances

A balanced human diet needs to contain a large number of different components. These include *proteins, carbohydrates, fats, minerals* (including water), and *vitamins.* These substances can occur in widely varying amounts and proportions, depending on the type of diet. As several components of the diet are *essential for life,* they have to be regularly ingested with food. Recommended daily minimums for nutrients have been published by the World Health Organization (WHO) and a number of national expert committees.

A. Energy requirement ●

The amount of energy required by a human is expressed in kJ d^{-1} (kilojoule per day). An older unit is the kilocalorie (kcal; 1 kcal = 4.187 kJ). The figures given are recommended values for adults with a normal body weight. However, actual requirements are based on age, sex, body weight, and in particular on physical activity. In those involved in competitive sports, for example, requirements can increase from 12 000 to 17 000 kJ d^{-1}.

It is recommended that about half of the energy intake should be in the form of carbohydrates, a third at most in the form of fat, and the rest as protein. The fact that *alcoholic beverages* can make a major contribution to daily energy intake is often overlooked. Ethanol has a caloric value of about 30 kJ g^{-1} (see p. 320).

B. Nutrients ●

Proteins provide the body with amino acids, which are used for endogenous protein biosynthesis. Excess amino acids are broken down to provide energy (see p. 174). Most amino acids are *glucogenic*—i. e., they can be converted into glucose (see p. 180).

Proteins are essential components of the diet, as they provide **essential amino acids** that the human body is not capable of producing on its own (see the table). Some amino acids, including cysteine and histidine, are not absolutely essential, but promote growth in children. Some amino acids are able to substitute for each other in the diet. For example, humans can form tyrosine, which is actually essential, by hydroxylation from phenylalanine, and cysteine from methionine.

The *minimum daily requirement* of protein is 37 g for men and 29 g for women, but the recommended amounts are about twice these values. Requirements in pregnant and breastfeeding women are even higher. Not only the quantity, but also the quality of protein is important. Proteins that lack several essential amino acids or only contain small quantities of them are considered to be of low value, and larger quantities of them are therefore needed. For example, pulses only contain small amounts of methionine, while wheat and corn proteins are poor in lysine. In contrast to vegetable proteins, most animal proteins are high-value (with exceptions such as collagen and gelatin).

Carbohydrates serve as a general and easily available energy source. In the diet, they are present as *monosaccharides* in honey and fruit, or as *disaccharides* in milk and in all foods sweetened with sugar (sucrose). Metabolically usable *polysaccharides* are found in vegetable products (starch) and animal products (glycogen). Carbohydrates represent a substantial proportion of the body's energy supply, but they are not essential.

Fats are primarily important energy suppliers in the diet. Per gram, they provide more than twice as much energy as proteins and carbohydrates. Fats are essential as suppliers of *fat-soluble vitamins* (see p. 364) and as sources of *polyunsaturated fatty acids,* which are needed to biosynthesize eicosanoids (see pp. 48, 390).

Mineral substances and trace elements, a very heterogeneous group of essential nutrients, are discussed in more detail on p. 362. They are usually divided into macrominerals and microminerals.

Vitamins are also indispensable components of the diet. The animal body requires them in very small quantities in order to synthesize coenzymes and signaling substances (see pp. 364–369).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Energy requirement

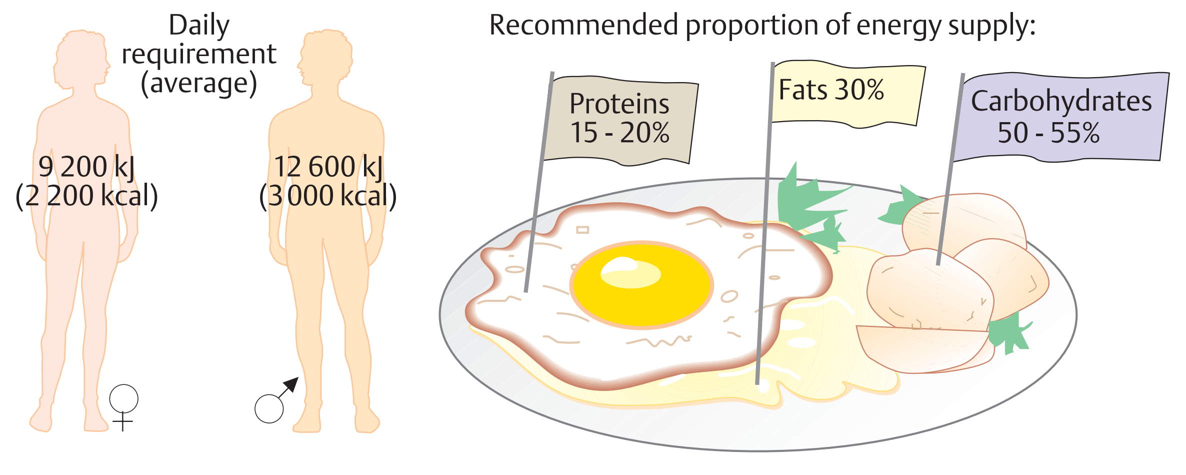

B. Nutrients

	Quantity in body (kg)	Energy content kJ · g^{-1} (kcal · g^{-1})	Daily requirement (g)			General function in metabolism	Essential constituents
			a	b	c		
Proteins	10	17 (4.1)	♂ 37 ♀ 29	55 45	92 75	Supplier of amino acids Energy source Daily requirement in mg per kg body weight	Essential amino acids:: Val (14) Leu (16) Ile (12) Lys (12) Phe (16) Trp (3) Met (10) Thr (8) Cys and His stimulate growth
Carbohydrates	1	17 (4.1)	0	390	240-310	General source of energy (glucose) Energy reserve (glycogen) Roughage (cellulose) Supporting substances (bones, cartilage, mucus)	Non-essential nutritional constituent
Fats	10-15	39 (9.3)	10	80	130	General energy source Most important energy reserve Solvent for vitamins Supplier of essential fatty acids	Polyunsaturated fatty acids: Linoleic acid Linolenic acid Arachidonic acid (together 10 g/day)
Water	35-40	0	2 400	–	–	Solvent Cellular building block Dielectric Reaction partner Temperature regulator	
Minerals	3	0				Building blocks Electrolytes Cofactors of enzymes	Macrominerals Microminerals (trace elements)
Vitamins	–	–				Often precursors of coenzymes	Lipid-soluble vitamins Water-soluble vitamins

a: Minimum daily requirement

b: Recommended daily intake

c: Actual daily intake in industrialized nations

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Minerals and trace elements

A. Minerals ◑

Water is the most important essential inorganic nutrient in the diet. In adults, the body has a daily requirement of 2–3 L of water, which is supplied from drinks, water contained in solid foods, and from the *oxidation water* produced in the respiratory chain (see p. 140). The special role of water for living processes is discussed in more detail elsewhere (see p. 26).

The elements essential for life can be divided into **macroelements** (daily requirement > 100 mg) and **microelements** (daily requirement < 100 mg). The macroelements include the **electrolytes** sodium (Na), potassium (K), calcium (Ca), and magnesium (Mg), and the nonmetals chlorine (Cl), phosphorus (P), sulfur (S), and iodine (I).

The essential microelements are only required in trace amounts (see also p. 2). This group includes iron (Fe), zinc (Zn), manganese (Mn), copper (Cu), cobalt (Co), chromium (Cr), selenium (Se), and molybdenum (Mo). Fluorine (F) is not essential for life, but does promote healthy bones and teeth. It is still a matter of controversy whether vanadium, nickel, tin, boron, and silicon also belong to the essential trace elements.

The second column in the table lists the average **amounts** of mineral substances in the body of an adult weighing 65 kg. The **daily requirements** listed in the fourth column also apply to an adult, and are *average values.* Children, pregnant and breast-feeding women, and those who are ill generally have higher mineral requirements relative to body weight than men.

As the human body is able to store many minerals, deviations from the daily ration are balanced out over a given period of time. Minerals stored in the body include water, which is distributed throughout the whole body; calcium, stored in the form of apatite in the bones (see p. 340); iodine, stored as thyroglobulin in the thyroid; and iron, stored in the form of ferritin and hemosiderin in the bone marrow, spleen, and liver (see p. 286). The storage site for many trace elements is the liver. In many cases, the metabolism of minerals is regulated by *hormones*—for example, the uptake and excretion of H_2O, Na^+, Ca^{2+}, and phosphate (see p. 328), and storage of Fe^{2+} and I^-.

Resorption of the required mineral substances from food usually depends on the body's requirements, and in several cases also on the composition of the diet. One example of dietary influence is calcium (see p. 342). Its resorption as Ca^{2+} is promoted by lactate and citrate, but phosphate, oxalic acid, and phytol inhibit calcium uptake from food due to complex formation and the production of insoluble salts.

Mineral deficiencies are not uncommon and can have quite a variety of causes—e. g., an unbalanced diet, resorption disturbances, and diseases. *Calcium deficiency* can lead to rickets, osteoporosis, and other disturbances. *Chloride deficiency* is observed as a result of severe Cl^- losses due to vomiting. Due to the low content of iodine in food in many regions of central Europe, *iodine deficiency* is widespread there and can lead to goiter. *Magnesium deficiency* can be caused by digestive disorders or an unbalanced diet—e. g., in alcoholism. Trace element deficiencies often result in a disturbed blood picture—i. e., forms of anemia.

The last column in the table lists some of the functions of minerals. It should be noted that almost all of the **macroelements** in the body function either as *nutrients* or *electrolytes.* Iodine (as a result of its incorporation into iodothyronines) and calcium act as *signaling substances.* Most **trace elements** are *cofactors for proteins,* especially for enzymes. Particularly important in quantitative terms are the *iron proteins* hemoglobin, myoglobin, and the cytochromes (see p. 286), as well as more than 300 different *zinc proteins.*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Minerals

Mineral	Content * (g)	Major source	Daily requirement (g)	Functions/Occurrence
Water	35 000- 40 000	Drinks Water in solid foods From metabolism 300g	1200 900	Solvent, cellular building block, dielectric, coolant, medium for transport, reaction partner
Macroelements (daily requirement >100 mg)				
Na	100	Table salt	1.1-3.3	Osmoregulation, membrane potential, mineral metabolism
K	150	Vegetables, fruit, cereals	1.9-5.6	Membrane potential, mineral metabolism
Ca	1 300	Milk, milk products	0.8	Bone formation, blood clotting, signal molecule
Mg	20	Green vegetables	0.35	Bone formation, cofactor for enzymes
Cl	100	Table salt	1.7-5.1	Mineral metabolism
P	650	Meat, milk, cereals, vegetables	0.8	Bone formation, energy metabolism, nucleic acid metabolism
S	200	*S*-containing amino acids (Cys and Met)	0.2	Lipid and carbohydrate metabolism, conjugate formation
Microelements (trace elements)			(mg	
Fe	4-5	Meat, liver, eggs, vegetables, potatoes, cereals	10	Hemoglobin, myoglobin, cytochromes, Fe/S clusters
Zn	2-3	Meat, liver, cereals	15	Zinc enzymes
Mn	0.02	Found in many foodstuffs	2-5	Enzymes
Cu	0.1-0.2	Meat, vegetables, fruit, fish	2-3	Oxidases
Co	<0.01	Meat	Traces	Vitamin B_{12}
Cr	<0.01		0.05-0.2	Not clear
Mo	0.02	Cereals, nuts, legumes	0.15-0.5	Redox enzymes
Se		Vegetables, meat	0.05-0.2	Selenium enzymes
I	0.03	Seafood, iodized salt, drinking water	0.15	Thyroxin
Requirement not known				🟦 Metals 🟨 Non-metals
F		Drinking water (fluoridated), tea, milk	0.0015-0.004	Bones, dental enamel

* Content in the body of a 65 kg adult

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Lipid-soluble vitamins

Vitamins are essential organic compounds that the animal organism is not capable of forming itself, although it requires them in small amounts for metabolism. Most vitamins are **precursors of coenzymes**; in some cases, they are also precursors of **hormones** or act as **antioxidants.** Vitamin requirements vary from species to species and are influenced by age, sex, and physiological conditions such as pregnancy, breast-feeding, physical exercise, and nutrition.

A. Vitamin supply ●

A healthy diet usually covers average daily vitamin requirements. By contrast, malnutrition, malnourishment (e. g., an unbalanced diet in older people, malnourishment in alcoholics, ready meals), or resorption disturbances lead to an inadequate supply of vitamins from which **hypovitaminosis,** or in extreme cases avitaminosis, can result. Medical treatments that kill the intestinal flora—e. g., antibiotics—can also lead to vitamin deficiencies (K, B_{12}, H) due to the absence of bacterial vitamin synthesis.

Since only a few vitamins can be stored (A, D, E, B_{12}), a lack of vitamins quickly leads to **deficiency diseases.** These often affect the skin, blood cells, and nervous system. The causes of vitamin deficiencies can be treated by improving nutrition and by administering vitamins in tablet form. An overdose of vitamins only leads to **hypervitaminoses**, with toxic symptoms, in the case of vitamins A and D. Normally, excess vitamins are rapidly excreted with the urine.

B. Lipid-soluble vitamins ◑

Vitamins are classified as either lipid-soluble or water-soluble. The lipid-soluble vitamins include vitamins A, D, E, and K, all of which belong to the isoprenoids (see p. 52).

Vitamin A (retinol) is the parent substance of the *retinoids,* which include *retinal* and *retinoic acid.* The retinoids also can be synthesized by cleavage from the provitamin β-carotene. Retinoids are found in meat-containing diets, whereas β-carotene occurs in fruits and vegetables (particularly carrots). **Retinal** is involved in visual processes as the pigment of the chromoprotein *rhodopsin* (see p. 358). **Retinoic acid**, like the steroid hormones, influences the transcription of genes in the cell nucleus. It acts as a differentiation factor in growth and development processes. Vitamin A deficiency can result in *night blindness, visual impairment,* and *growth disturbances.*

Vitamin D (calciol, cholecalciferol) is the precursor of the hormone *calcitriol* (1α,25-dihydroxycholecalciferol; see p. 320). Together with two other hormones (parathyrin and calcitonin), calcitriol regulates the calcium metabolism (see p. 342). Calciol can be synthesized in the skin from 7-dehydrocholesterol, an endogenous steroid, by a photochemical reaction. Vitamin D deficiencies only occur when the skin receives insufficient exposure to ultraviolet light and vitamin D is lacking in the diet. Deficiency is observed in the form of *rickets* in children and *osteomalacia* in adults. In both cases, bone mineralization is disturbed.

Vitamin E (tocopherol) and related compounds only occur in plants (e. g., wheat germ). They contain what is known as a *chroman ring.* In the lipid phase, vitamin E is mainly located in biological membranes, where as an *antioxidant* it protects unsaturated lipids against ROS (see p. 284) and other radicals.

Vitamin K (phylloquinone) and similar substances with modified side chains are involved in carboxylating glutamate residues of coagulation factors in the liver (see p. 290). The form that acts as a cofactor for carboxylase is derived from the vitamin by enzymatic reduction. Vitamin K antagonists (e. g., coumarin derivatives) inhibit this reduction and consequently carboxylation as well. This fact is used to inhibit blood coagulation in *prophylactic treatment against thrombosis.* Vitamin K deficiency occurs only rarely, as the vitamin is formed by bacteria of the intestinal flora.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

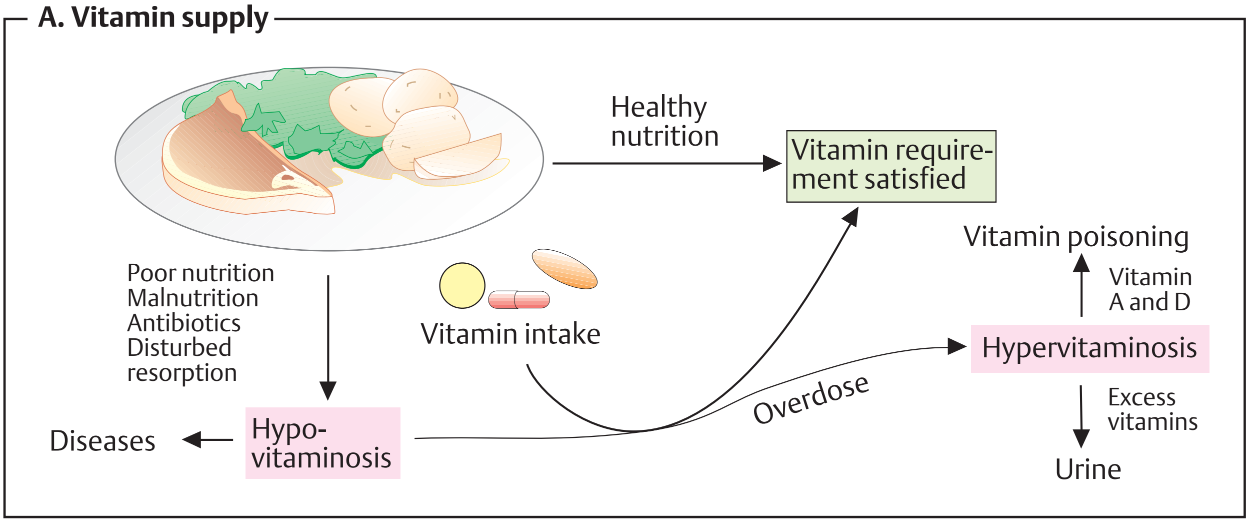

A. Vitamin supply

Healthy nutrition
Vitamin requirement satisfied
Vitamin poisoning
Vitamin A and D
Poor nutrition
Malnutrition
Antibiotics
Disturbed resorption
Vitamin intake
Hypervitaminosis
Overdose
Excess vitamins
Diseases
Hypo-vitaminosis
Urine

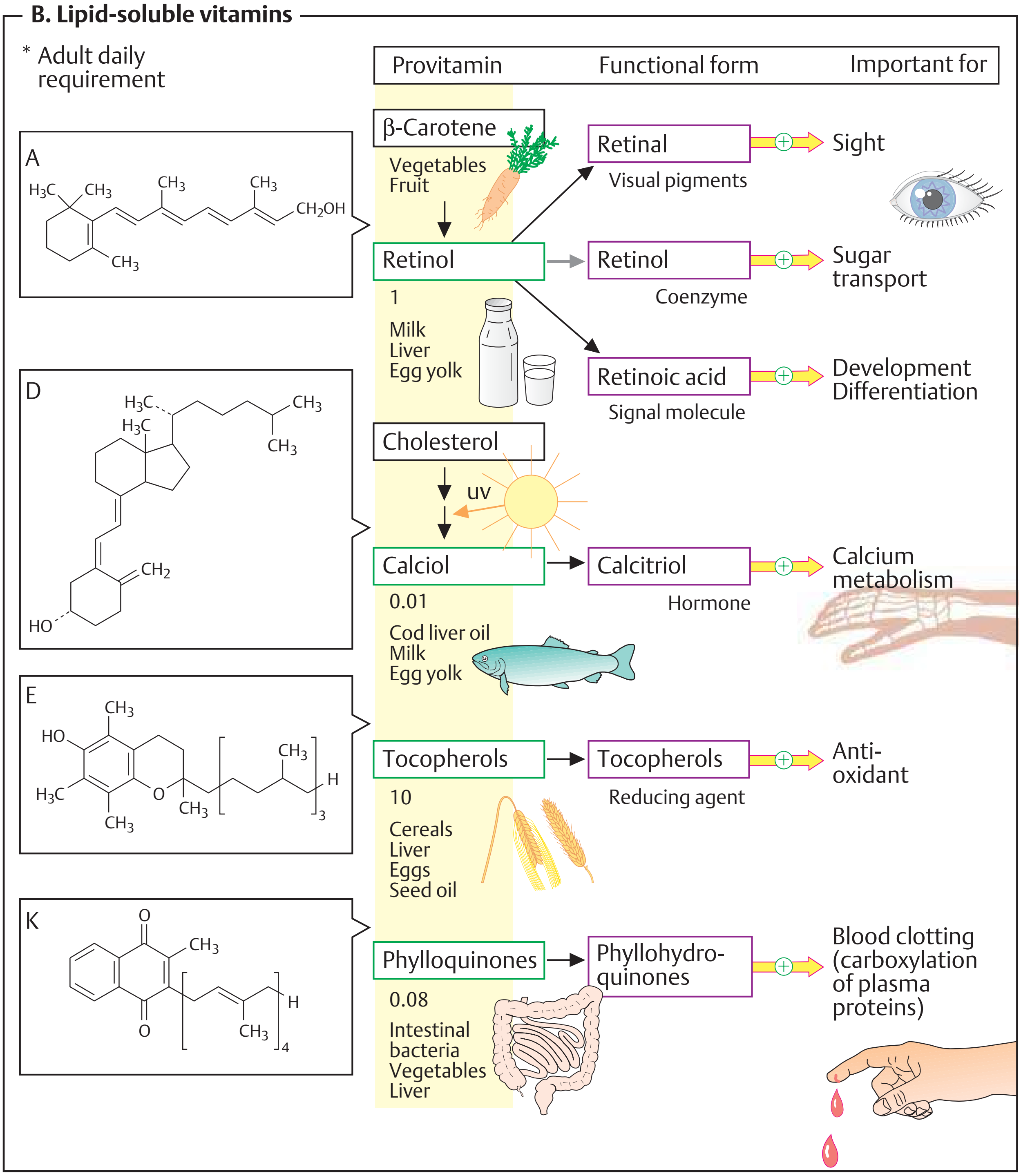

B. Lipid-soluble vitamins

* Adult daily requirement
Provitamin
Functional form
Important for
A
H3C
CH3
CH3
CH3
CH2OH
CH3
β-Carotene
Vegetables
Fruit
Retinal
Visual pigments
Sight
Retinol
Retinol
Coenzyme
Sugar transport
1
Milk
Liver
Egg yolk
Retinoic acid
Signal molecule
Development
Differentiation
D
H3C
H3C
CH3
CH3
CH2
HO
Cholesterol
uv
Calciol
Calcitriol
Hormone
Calcium metabolism
0.01
Cod liver oil
Milk
Egg yolk
E
CH3
HO
CH3
H3C
O
CH3
H
CH3
3
Tocopherols
Tocopherols
Reducing agent
Anti-oxidant
10
Cereals
Liver
Eggs
Seed oil
K
O
CH3
H
O
CH3
4
Phylloquinones
Phyllohydro-quinones
Blood clotting (carboxylation of plasma proteins)
0.08
Intestinal bacteria
Vegetables
Liver

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Water-soluble vitamins I

The B group of vitamins covers water-soluble vitamins, all of which serve as precursors for coenzymes. Their numbering sequence is not continuous, as many substances that were originally regarded as vitamins were not later confirmed as having vitamin characteristics.

A. Water-soluble vitamins I ◑

Vitamin B_1 (thiamine) contains two heterocyclic rings—a *pyrimidine ring* (a six-membered aromatic ring with two Ns) and a *thiazole ring* (a five-membered aromatic ring with N and S), which are joined by a methylene group. The active form of vitamin B_1 is **thiamine diphosphate** (TPP), which contributes as a coenzyme to the transfer of hydroxyalkyl residues (active aldehyde groups). The most important reactions of this type are oxidative decarboxylation of 2-oxoacids (see p. 134) and the transketolase reaction in the pentose phosphate pathway (see p. 152). Thiamine was the first vitamin to be discovered, around 100 years ago. Vitamin B_1 deficiency leads to *beriberi,* a disease with symptoms that include neurological disturbances, cardiac insufficiency, and muscular atrophy.

Vitamin B_2 is a complex of several vitamins: riboflavin, folate, nicotinate, and pantothenic acid.

Riboflavin (from the Latin *flavus,* yellow) serves in the metabolism as a component of the redox coenzymes flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD; see p. 104). As prosthetic groups, **FMN** and **FAD** are cofactors for various oxidoreductases (see p. 32). No specific disease due to a deficiency of this vitamin is known.

Folate, the anion of folic acid, is made up of three different components—a *pteridine derivative, 4-aminobenzoate,* and one or more *glutamate* residues. After reduction to tetrahydrofolate (**THF**), folate serves as a coenzyme in the C_1 metabolism (see p. 418). Folate deficiency is relatively common, and leads to disturbances in nucleotide biosynthesis and thus cell proliferation. As the precursors for blood cells divide particularly rapidly, disturbances of the blood picture can occur, with increased amounts of abnormal precursors for megalocytes (*megaloblastic anemia*). Later, general damage ensues as phospholipid synthesis and the amino acid metabolism are affected.

In contrast to animals, microorganisms are able to synthesize folate from their own components. The growth of microorganisms can therefore be inhibited by *sulfonamides,* which competitively inhibit the incorporation of 4-aminobenzoate into folate (see p. 254). Since folate is not synthesized in the animal organism, sulfonamides have no effect on animal metabolism.

Nicotinate and **nicotinamide,** together referred to as "niacin," are required for biosynthesis of the coenzymes nicotinamide adenine dinucleotide (**NAD^+**) and nicotinamide adenine dinucleotide phosphate (**$NADP^+$**). These both serve in energy and nutrient metabolism as carriers of *hydride ions* (see pp. 32, 104). The animal organism is able to convert *tryptophan* into nicotinate, but only with a poor yield. Vitamin deficiency therefore only occurs when nicotinate, nicotinamide, and tryptophan are all simultaneously are lacking in the diet. It manifests in the form of skin damage (*pellagra*), digestive disturbances, and depression.

Pantothenic acid is an acid amide consisting of β-alanine and 2,4-dihydroxy-3,3′-dimethylbutyrate (pantoic acid). It is a precursor of *coenzyme A,* which is required for activation of acyl residues in the lipid metabolism (see pp. 12, 106). *Acyl carrier protein* (ACP; see p. 168) also contains pantothenic acid as part of its prosthetic group. Due to the widespread availability of pantothenic acid in food (Greek *pantothen* = "from everywhere"), deficiency diseases are rare.

Further information

The requirement for vitamins in humans and other animals is the result of mutations in the enzymes involved in biosynthetic coenzymes. As intermediates of coenzyme biosynthesis are available in sufficient amounts in the diet of heterotrophic animals (see p. 112), the lack of endogenous synthesis did not have unfavorable effects for them. Microorganisms and plants whose nutrition is mainly autotrophic have to produce all of these compounds themselves in order to survive.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Water-soluble vitamins I

* Adult daily requirement

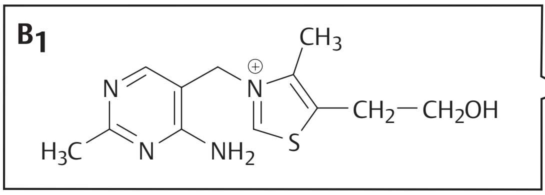
B_1
CH_3
N
$N^{\oplus}$
CH_2—CH_2OH
H_3C
N
NH_2
S

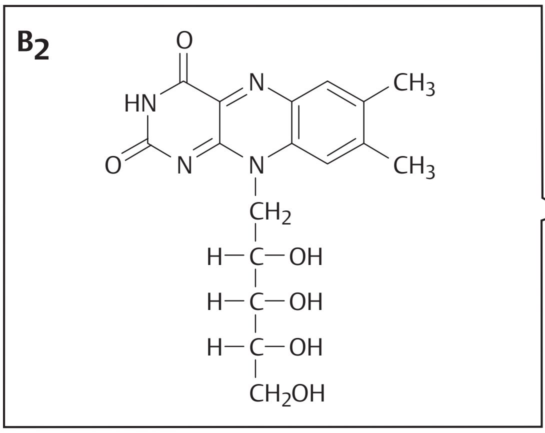
B_2
O
HN
N
CH_3
CH_3
O
N
N
CH_2
H—C—OH
H—C—OH
H—C—OH
CH_2OH

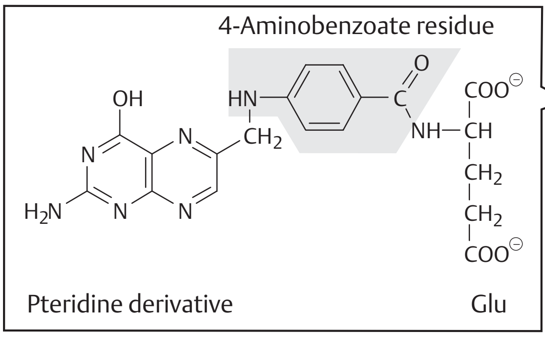
4-Aminobenzoate residue
OH
HN
O
C
$COO^{\ominus}$
N
N
CH_2
NH—CH
CH_2
CH_2
$COO^{\ominus}$
H_2N
N
N
Pteridine derivative
Glu

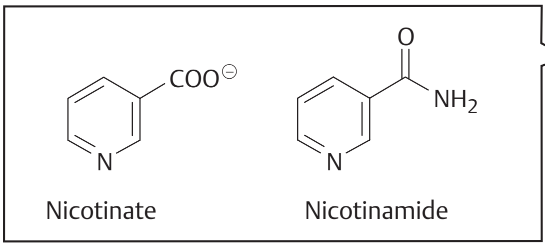
$COO^{\ominus}$
N
O
NH_2
N

Nicotinate Nicotinamide

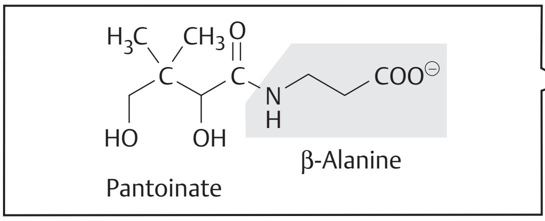
H_3C
CH_3
O
C
C
N
H
$COO^{\ominus}$
HO
OH
β-Alanine

Pantoinate

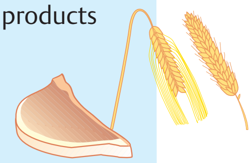

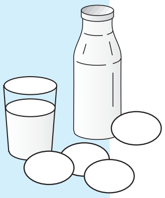

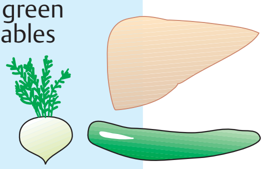

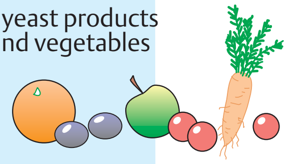

Vitamin	Active form: coenzyme	Function in metabolism
Thiamine 1.5 mg * Grain Yeast products Pork	TPP Thiamine diphosphate	Transfer of hydroxyalkyl residues
Riboflavin 1.8 mg* Milk Eggs	FMN FAD	Hydrogen transfer
Folate 0.2 mg* Fresh green vegetables Liver	THF Tetrahydrofolate	C_1-metabolism
Nicotinate Nicotinamide 20 mg* (or 1.2 g tryptophan) Meat, yeast products Fruit and vegetables	NADP⊕ NAD⊕	Hydride transfer
Pantothenate 7 mg* Widely distributed	CoA	Activation of carboxylic acids

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Water-soluble vitamins II

A. Water-soluble vitamins II ◑

Vitamin B_6 consists of three substituted pyridines—**pyridoxal**, **pyridoxol**, and **pyridoxamine.** The illustration shows the structure of pyridoxal, which carries an aldehyde group (–CHO) at C-4. Pyridoxol is the corresponding alcohol ($–CH_2OH$), and pyridoxamine the amine ($–CH_2NH_2$).

The active form of vitamin B_6, **pyridoxal phosphate,** is the most important coenzyme in the amino acid metabolism (see p. 106). Almost all conversion reactions involving amino acids require pyridoxal phosphate, including transaminations, decarboxylations, dehydrogenations, etc. *Glycogen phosphorylase,* the enzyme for glycogen degradation, also contains pyridoxal phosphate as a cofactor. Vitamin B_6 deficiency is rare.

Vitamin B_{12} (cobalamine) is one of the most complex low-molecular-weight substances occurring in nature. The core of the molecule consists of a tetrapyrrol system (*corrin*), with cobalt as the central atom (see p. 108). The vitamin is exclusively synthesized by microorganisms. It is abundant in liver, meat, eggs, and milk, but not in plant products. As the intestinal flora synthesize vitamin $B_{12,}$ strict vegetarians usually also have an adequate supply of the vitamin.

Cobalamine can only be resorbed in the small intestine when the gastric mucosa secretes what is known as *intrinsic factor*—a glycoprotein that binds cobalamine (the *extrinsic factor*) and thereby protects it from degradation. In the blood, the vitamin is bound to a special protein known as *transcobalamin.* The liver is able to store vitamin B_{12} in amounts sufficient to last for several months. Vitamin B_{12} deficiency is usually due to an absence of intrinsic factor and the resulting resorption disturbance. This leads to a disturbance in blood formation known as *pernicious anemia.*

In animal metabolism, derivatives of cobalamine are mainly involved in rearrangement reactions. For example, they act as coenzymes in the conversion of methylmalonyl-CoA to succinyl-CoA (see p. 166), and in the formation of methionine from homocysteine (see p. 418). In prokaryotes, cobalamine derivatives also play a part in the reduction of ribonucleotides.

Vitamin C is L-**ascorbic acid** (chemically: 2-oxogulonolactone). The two hydroxyl groups have acidic properties. By releasing a proton, ascorbic acid therefore turns into its anion, **ascorbate**. Humans, apes, and guinea pigs require vitamin C because they lack the enzyme L-*gulonolactone oxidase (1.1.3.8),* which catalyzes the final step in the conversion of glucose into ascorbate.

Vitamin C is particularly abundant in fresh fruit and vegetables. Many soft drinks and foodstuffs also have synthetic ascorbic acid added to them as an antioxidant and flavor enhancer. Boiling slowly destroys vitamin C. In the body, ascorbic acid serves as a reducing agent in variations reactions (usually hydroxylations). Among the processes involved are *collagen synthesis, tyrosine degradation, catecholamine synthesis,* and *bile acid biosynthesis.* The daily requirement for ascorbic acid is about 60 mg, a comparatively large amount for a vitamin. Even higher doses of the vitamin have a protective effect against infections. However, the biochemical basis for this effect has not yet been explained. Vitamin C deficiency only occurs rarely nowadays; it becomes evident after a few months in the form of *scurvy,* with connective-tissue damage, bleeding, and tooth loss.

Vitamin H (**biotin)** is present in liver, egg yolk, and other foods; it is also synthesized by the intestinal flora. In the body, biotin is covalently attached via a lysine side chain to enzymes that catalyze carboxylation reactions. Biotin-dependent carboxylases include *pyruvate carboxylase* (see p. 154) and *acetyl-CoA carboxylase* (see p. 162). CO_2 binds, using up ATP, to one of the two N atoms of biotin, from which it is transferred to the acceptor (see p. 108).

Biotin binds with high affinity ($K_d = 10^{-15}$ M) and specificity to *avidin*, a protein found in egg white. Since boiling denatures avidin, biotin deficiency only occurs when egg whites are eaten raw.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Water-soluble vitamins II

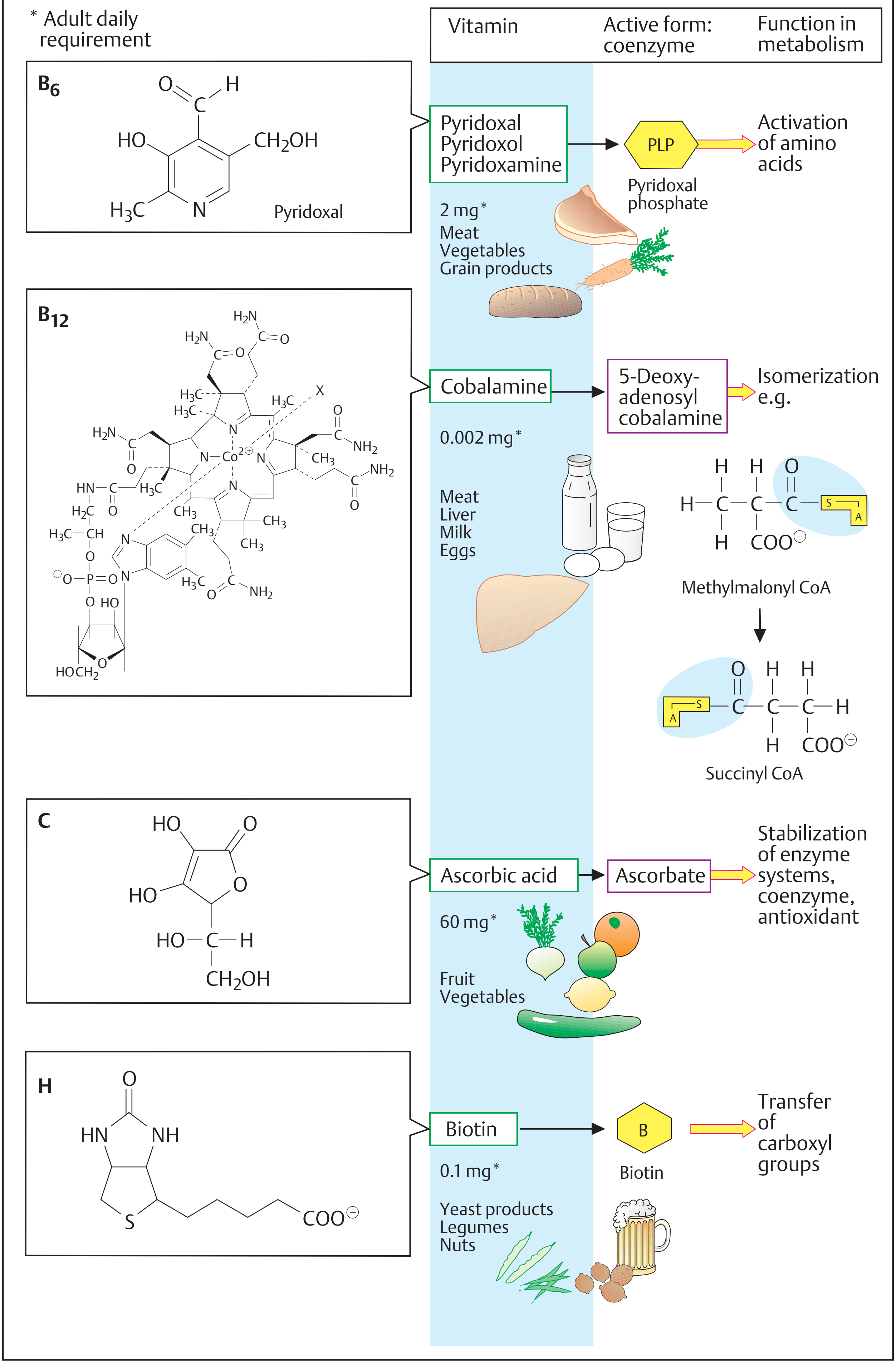
* Adult daily requirement
Vitamin
Active form: coenzyme
Function in metabolism
B6
O
C
H
HO
CH2OH
H3C
N
Pyridoxal
Pyridoxal
Pyridoxol
Pyridoxamine
PLP
Pyridoxal phosphate
Activation of amino acids
2 mg*
Meat
Vegetables
Grain products
B12
H2N
C=O
H3C
X
N
N–Co2⊕
CH3
NH2
HN–C
H2C
H3C–CH
O
⊖O–P=O
O HO
HOCH2
Cobalamine
5-Deoxy-adenosyl cobalamine
Isomerization e.g.
0.002 mg*
Meat
Liver
Milk
Eggs
H H O
H–C–C–C–
S
A
H COO⊖
Methylmalonyl CoA
O H H
–C–C–C–H
H COO⊖
Succinyl CoA
C
HO
O
HO
O
HO–C–H
CH2OH
Ascorbic acid
Ascorbate
Stabilization of enzyme systems, coenzyme, antioxidant
60 mg*
Fruit
Vegetables
H
O
HN
NH
S
COO⊖
Biotin
B
Biotin
Transfer of carboxyl groups
0.1 mg*
Yeast products
Legumes
Nuts

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Basics

Hormones are *chemical signaling substances.* They are synthesized in specialized cells that are often associated to form *endocrine glands.* Hormones are released into the *blood* and transported with the blood to their *effector organs.* In the organs, the hormones carry out physiological and biochemical regulatory functions. In contrast to endocrine hormones, **tissue hormones** are only active in the immediate vicinity of the cells that secrete them.

The distinctions between hormones and other signaling substances (mediators, neurotransmitters, and growth factors) are fluid. **Mediators** is the term used for signaling substances that do not derive from special hormone-forming cells, but are form by many cell types. They have hormone-like effects in their immediate surroundings. *Histamine* (see p. 352) and *prostaglandins* (see p. 390) are important examples of these substances. **Neurohormones** and **neurotransmitters** are signaling substances that are produced and released by nerve cells (see p. 348). **Growth factors** and **cytokines** mainly promote cell proliferation and cell differentiation (see p. 392).

A. Hormones: overview ●

The animal organism contains more than 100 hormones and hormone-like substances, which can be classified either according to their structure or according to their function. In chemical terms, most hormones are *amino acid derivatives, peptides* or *proteins,* or *steroids.* Hormones regulate the following processes:

- **Growth and differentiation of cells, tissues, and organs**
  These processes include cell proliferation, embryonic development, and sexual differentiation—i. e., processes that require a prolonged time period and involve proteins de novo synthesis. For this reason, mainly steroid hormones which function via transcription regulation are active in this field (see p. 244).
- **Metabolic pathways**
  Metabolic regulation requires rapidly acting mechanisms. Many of the hormones involved therefore regulate *interconversion* of enzymes (see p. 120). The main processes subject to hormonal regulation are the uptake and degradation of storage substances (glycogen, fat), metabolic pathways for biosynthesis and degradation of central metabolites (glucose, fatty acids, etc.), and the supply of metabolic energy.
- **Digestive processes**
  Digestive processes are usually regulated by locally acting peptides (paracrine; see p. 372), but mediators, biogenic amines, and neuropeptides are also involved (see p. 270).
- **Maintenance of ion concentrations (homeostasis)**
  Concentrations of Na^+, K^+, and Cl^- in body fluids, and the physiological variables dependent on these (e. g. blood pressure), are subject to strict regulation. The principal site of action of the hormones involved is the kidneys, where hormones increase or reduce the resorption of ions and recovery of water (see pp. 326–331). The concentrations of Ca^{2+} and phosphate, which form the mineral substance of bone and teeth, are also precisely regulated.

Many hormones influence the above processes only indirectly by regulating the synthesis and release of other hormones (*hormonal hierarchy;* see p. 372).

B. Hormonal regulation system ●

Each hormone is the center of a hormonal regulation system. Specialized glandular cells synthesize the hormone from precursors, store it in many cases, and release it into the bloodstream when needed (**biosynthesis**). For **transport**, the poorly water-soluble lipophilic hormones are bound to plasma proteins known as hormone carriers. To stop the effects of the hormone again, it is inactivated by enzymatic reactions, most of which take place in the liver (**metabolism**). Finally, the hormone and its metabolites are expelled via the excretory system, usually in the kidney (**excretion**). All of these processes affect the concentration of the hormone and thus contribute to regulation of the hormonal signal.

In the effector organs, target cells receive the hormone's message. These cells have hormone receptors for the purpose, which bind the hormone. Binding of a hormone passes information to the cell and triggers a response (**effect**).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Hormones: overview

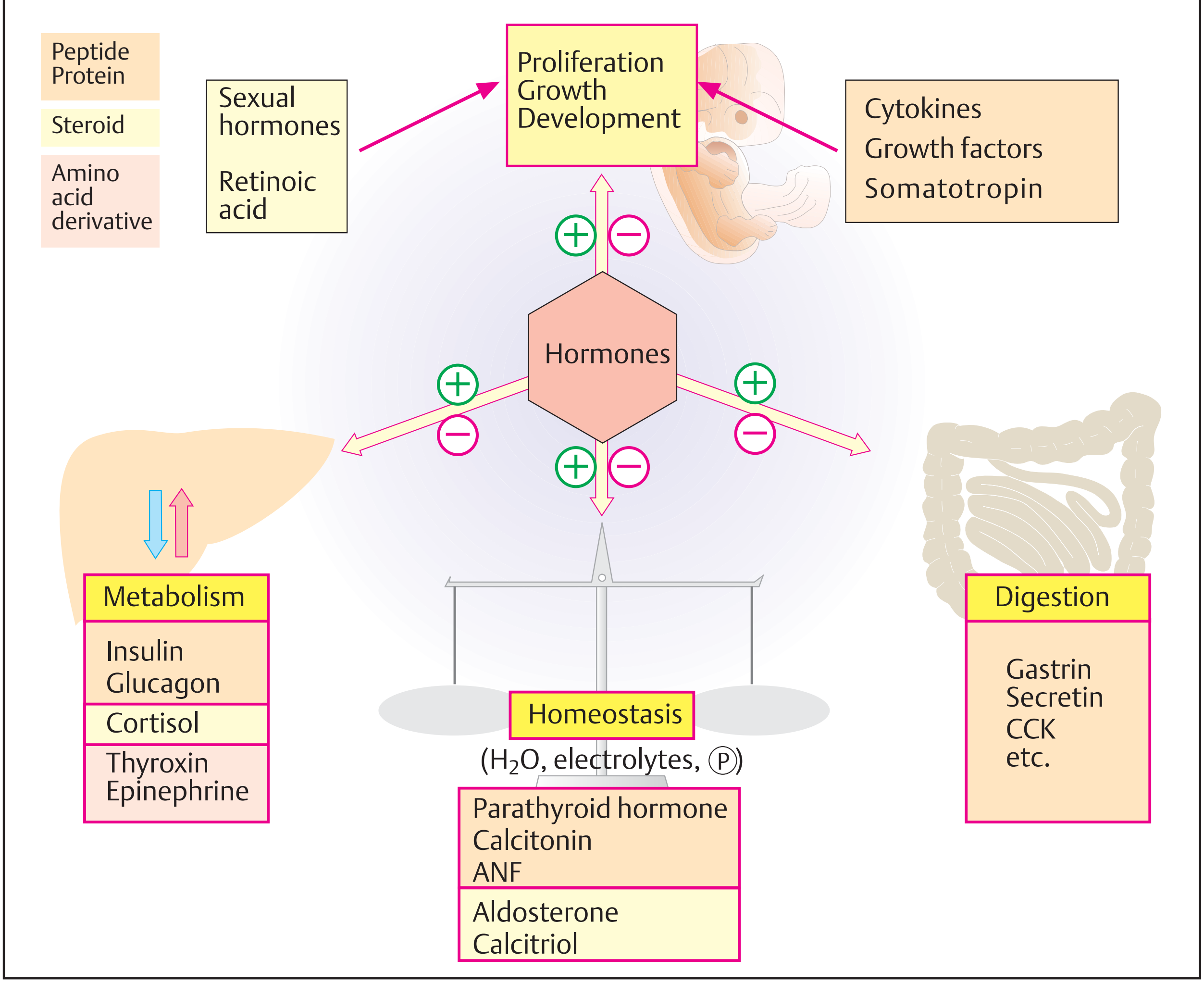
Peptide
Protein
Steroid
Amino acid derivative
Sexual hormones
Retinoic acid
Proliferation
Growth
Development
Cytokines
Growth factors
Somatotropin
Hormones
+
−
Metabolism
Insulin
Glucagon
Cortisol
Thyroxin
Epinephrine
Homeostasis
(H_2O, electrolytes, Ⓟ)
Parathyroid hormone
Calcitonin
ANF
Aldosterone
Calcitriol
Digestion
Gastrin
Secretin
CCK
etc.

A. Hormonal regulation system

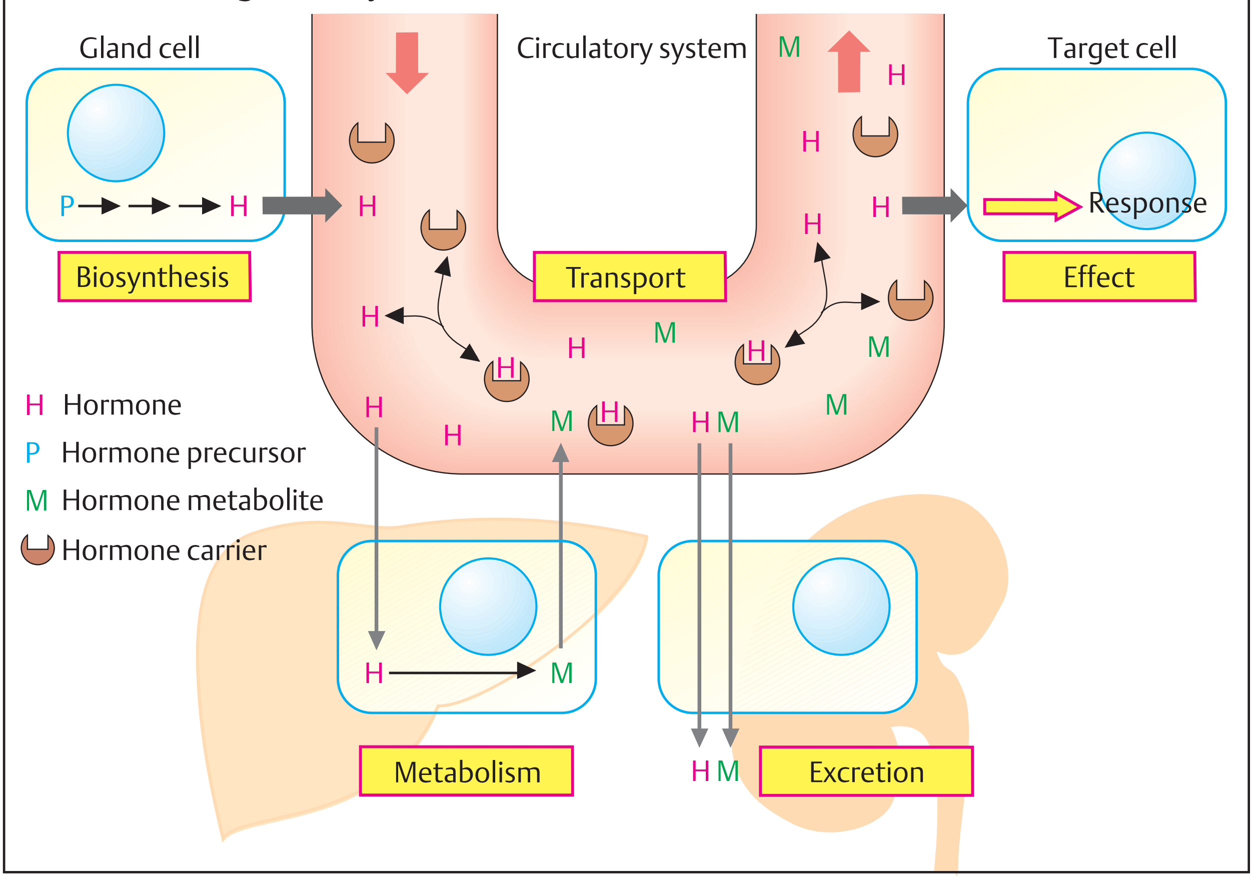
Gland cell
P
H
Biosynthesis
Circulatory system
Transport
H
M
Target cell
Response
Effect
H Hormone
P Hormone precursor
M Hormone metabolite
Hormone carrier
Metabolism
Excretion
H M

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Plasma levels and hormone hierarchy

A. Endocrine, paracrine, and autocrine hormone effects ◑

Hormones transfer signals by migrating from their site of synthesis to their site of action. They are usually transported in the blood. In this case, they are said to have an **endocrine effect** (**1**; example: *insulin*). By contrast, *tissue hormones,* the target cells for which are in the immediate vicinity of the glandular cells that produce them, are said to have a **paracrine effect** (**2**; example: *gastrointestinal tract hormones*). When signal substances also pass effects back to the cells that synthesize them, they are said to have an **autocrine effect** (**3**; example: *prostaglandins*). Autocrine effects are often found in tumor cells (see p. 400), which stimulate their own proliferation in this way.

Insulin, which is formed in the B cells of the pancreas, has both endocrine and paracrine effects. As a hormone with endocrine effects, it regulates glucose and fat metabolism. Via a paracrine mechanism, it inhibits the synthesis and release of *glucagon* from the neighboring A cells.

B. Dynamics of the plasma level ◑

Hormones circulate as signaling substances in the blood at very low concentrations (10^{-12} to between 10^{-7} mol L^{-1}). These values change periodically in rhythms that depend on the time of day, month, or year, or on physiological cycles.

The first example shows the **circadian** rhythm of the cortisol level. As an activator of gluconeogenesis (see p. 158), cortisol is mainly released in the early morning, when the liver's glycogen stores are declining. During the day, the plasma cortisol level declines.

Many hormones are released into the blood in a spasmodic and irregular manner. In this case, their concentrations change in an **episodic** or **pulsatile** fashion. This applies, for instance, to luteinizing hormone (LH, lutropin).

Concentrations of other hormones are **event-regulated**. For example, the body responds to increased blood sugar levels after meals by releasing *insulin.* Regulation of hormone synthesis, release, and degradation allows the blood concentrations of hormones to be precisely adjusted. This is based either on simple feedback control or on hierarchically structured regulatory systems.

C. Closed-loop feedback control ◑

The biosynthesis and release of *insulin* by the pancreatic B cells (see p. 160) is stimulated by high blood glucose levels (> 5 mM). The insulin released then stimulates increased uptake and utilization of glucose by the cells of the muscle and adipose tissues. As a result, the blood glucose level falls back to its normal value, and further release of insulin stops.

D. Hormone hierarchy ◑

Hormone systems are often linked to each another, giving rise in some cases to a hierarchy of higher-order and lower-order hormones. A particularly important example is the *pituitary–hypothalamic axis,* which is controlled by the central nervous system (CNS).

Nerve cells in the hypothalamus react to stimulatory or inhibitory signals from the CNS by releasing activating or inhibiting factors, which are known as **liberins** (*"releasing hormones"*) and **statins** (*"inhibiting hormones"*). These neurohormones reach the adenohypophysis by short routes through the bloodstream. In the adenohypophysis, they stimulate (liberins) or inhibit (statins) the biosynthesis and release of tropines. **Tropines** (*glandotropic hormones*) in turn stimulate peripheral glands to synthesize glandular hormones. Finally, the **glandular hormone** acts on its target cells in the organism. In addition, it passes effects back to the higher-order hormone systems. This (usually negative) feedback influences the concentrations of the higher-order hormones, creating a feedback loop.

Many steroid hormones are regulated by this type of axis—e. g., thyroxin, cortisol, estradiol, progesterone, and testosterone. In the case of the glucocorticoids, the hypothalamus releases corticotropin-releasing hormone (CRH or corticoliberin, a peptide consisting of 41 amino acids), which in turn releases corticotropin (ACTH, 39 AAs) in the pituitary gland. Corticotropin stimulates synthesis and release of the glandular steroid hormone cortisol in the adrenal cortex.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Endocrine, paracrine and autocrine hormone effects

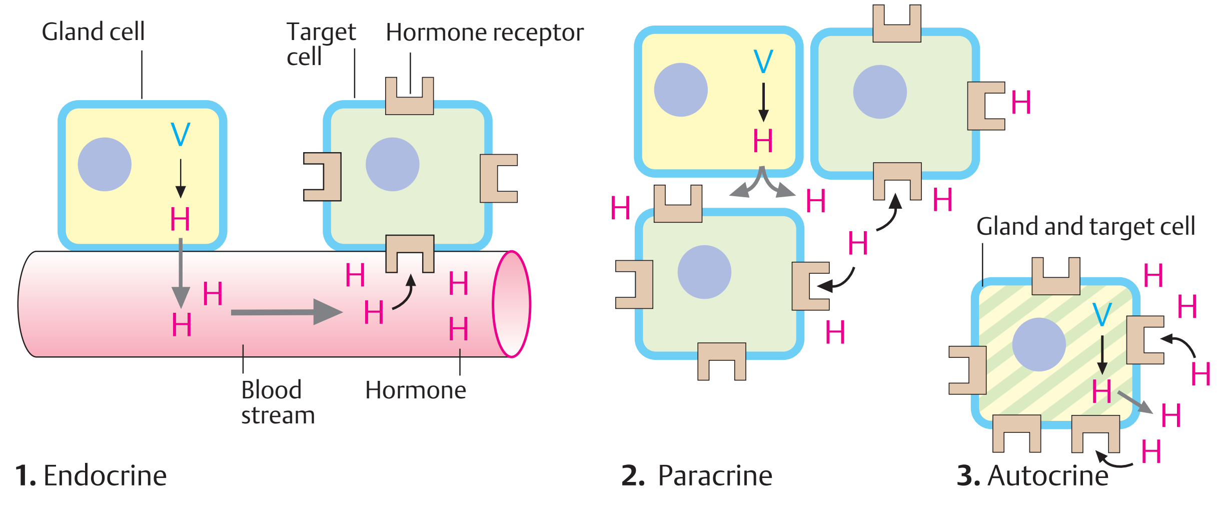
Gland cell
Target cell
Hormone receptor
V
H
Blood stream
Hormone
1. Endocrine
2. Paracrine
Gland and target cell
3. Autocrine

B. Plasma level dynamics

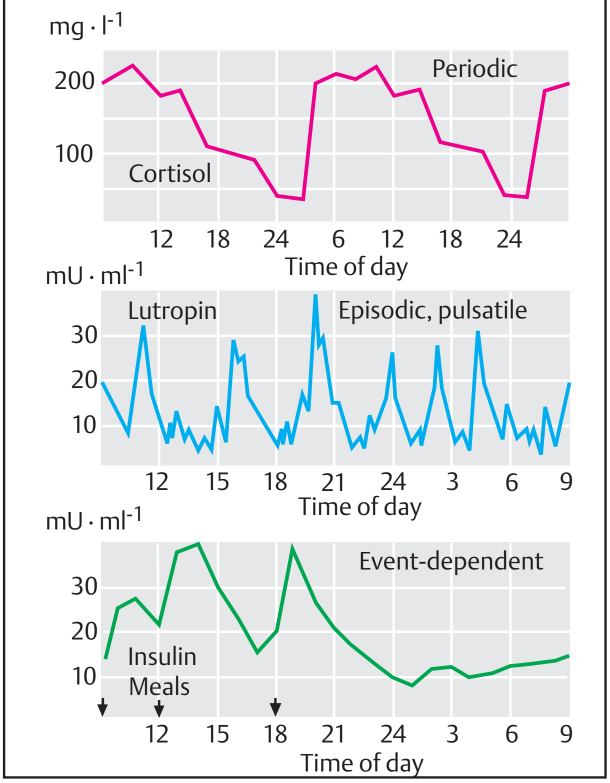

C. Regulatory circuit

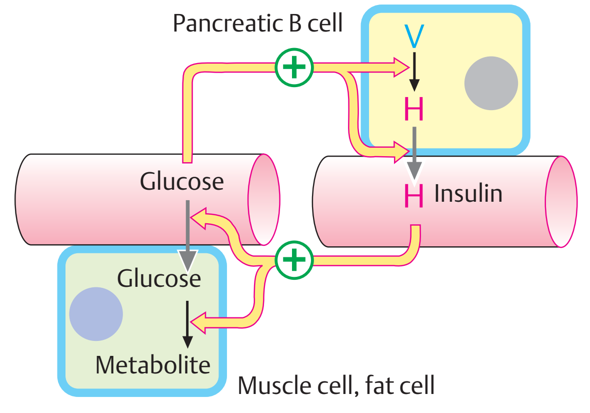
Pancreatic B cell
V
H
Glucose
Insulin
Glucose
Metabolite
Muscle cell, fat cell

D. Hormone hierarchy

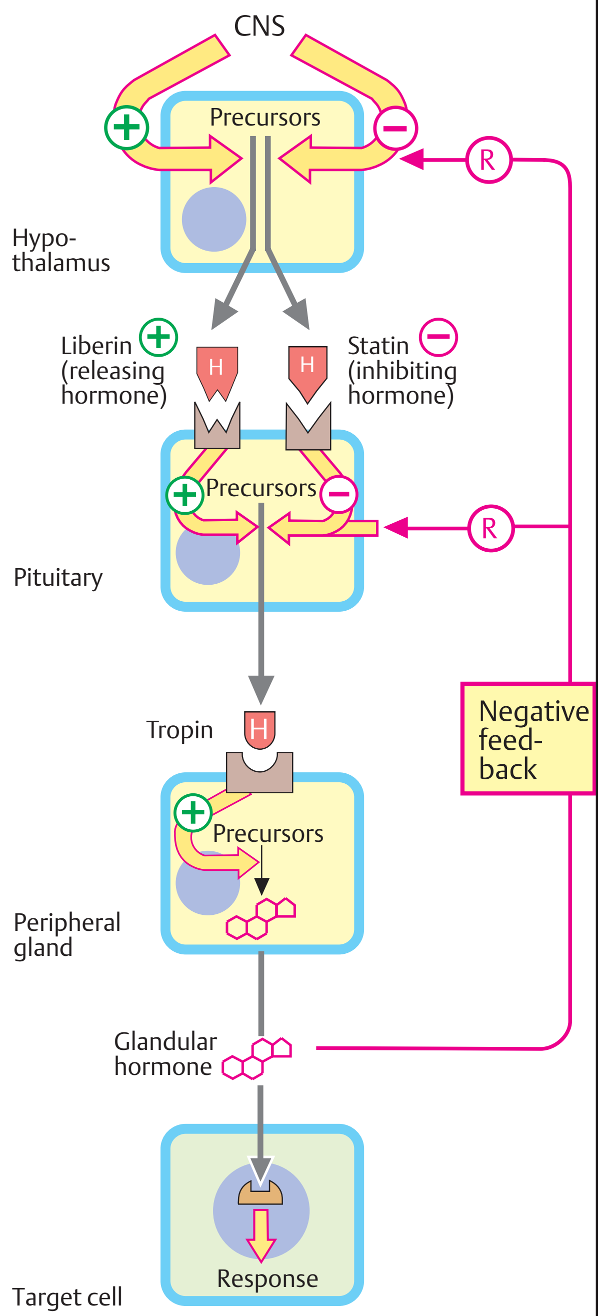
CNS
Precursors
Hypo-
thalamus
Liberin (releasing hormone)
Statin (inhibiting hormone)
H
R
Precursors
Pituitary
Tropin
Negative feed-back
Precursors
Peripheral gland
Glandular hormone
Response
Target cell

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Lipophilic hormones

Classifying hormones into hydrophilic and lipophilic molecules indicates the chemical properties of the two groups of hormones and also reflects differences in their mode of action (see p. 120).

A. Lipophilic hormones ◑

Lipophilic hormones, which include *steroid hormones, iodothyronines,* and *retinoic acid,* are relatively small molecules (300–800 Da) that are poorly soluble in aqueous media. With the exception of the iodothyronines, they are not stored by hormone-forming cells, but are released immediately after being synthesized. During transport in the blood, they are bound to specific carriers. Via intracellular receptors, they mainly act on transcription (see p. 358). Other effects of steroid hormones—e. g., on the immune system—are not based on transcriptional control. Their details have not yet been explained.

Steroid hormones

The most important steroid hormones in vertebrates are listed on p. 57. *Calcitriol* (vitamin D hormone) is also included in this group, although it has a modified steroid structure. The most important steroid hormone in invertebrates is *ecdysone.*

Progesterone is a female sexual steroid belonging to the progestin *(gestagen)* family. It is synthesized in the corpus luteum of the ovaries. The blood level of progesterone varies with the menstrual cycle. The hormone prepares the uterus for a possible pregnancy. Following fertilization, the placenta also starts to synthesize progesterone in order to maintain the pregnant state. The development of the mammary glands is also stimulated by progesterone.

Estradiol is the most important of the *estrogens.* Like progesterone, it is synthesized by the ovaries and, during pregnancy, by the placenta as well. Estradiol controls the menstrual cycle. It promotes proliferation of the uterine mucosa, and is also responsible for the development of the female secondary sexual characteristics (breast, fat distribution, etc.).

Testosterone is the most important of the male sexual steroids (*androgens*). It is synthesized in the Leydig intersitial cells of the testes, and controls the development and functioning of the male gonads. It also determines secondary sexual characteristics in men (muscles, hair, etc.).

Cortisol, the most important *glucocorticoid,* is synthesized by the adrenal cortex. It is involved in regulating protein and carbohydrate metabolism by promoting protein degradation and the conversion of amino acids into glucose. As a result, the blood glucose level rises (see p. 152). Synthetic glucocorticoids (e. g., dexamethasone) are used in drugs due to their anti-inflammatory and immunosuppressant effects.

Aldosterone, a *mineralocorticoid,* is also synthesized in the adrenal gland. In the kidneys, it promotes Na^+ resorption by inducing Na^+/K^+ ATPase and Na^+ channels (see p. 328). At the same time, it leads to increased K^+ excretion. In this way, aldosterone indirectly increases blood pressure.

Calcitriol is a derivative of vitamin D (see p. 364). On exposure to ultraviolet light, a precursor of the hormone can also arise in the skin. Calcitriol itself is synthesized in the kidneys (see p. 330). Calcitriol promotes the resorption of calcium in the intestine and increases the Ca^{2+} level in the blood.

Iodothyronines

The thyroid hormone **thyroxine** (tetraiodothyronine, T_4) and its active form **triiodothyronine** (T_3) are derived from the amino acid *tyrosine.* The iodine atoms at positions 3 and 5 of the two phenol rings are characteristic of them. Post-translational synthesis of thyroxine takes place in the thyroid gland from tyrosine residues of the protein *thyroglobulin,* from which it is proteolytically cleaved before being released. Iodothyronines are the only organic molecules in the animal organism that contain iodine. They increase the basal metabolic rate, partly by regulating mitochondrial ATP synthesis. In addition, they promote embryonic development.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Lipophilic hormones

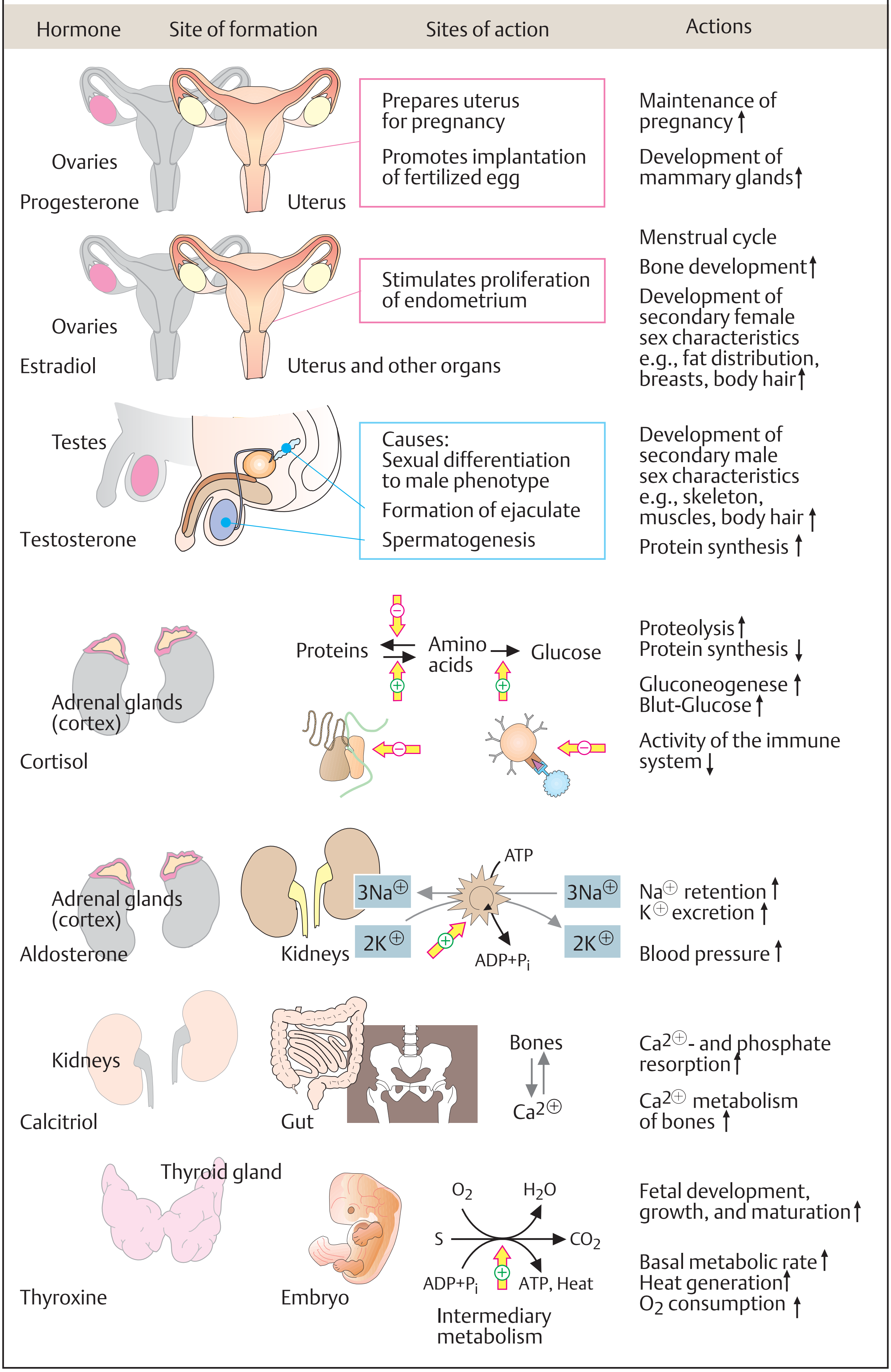
Hormone
Site of formation
Sites of action
Actions
Ovaries
Progesterone
Uterus
Prepares uterus for pregnancy
Promotes implantation of fertilized egg
Maintenance of pregnancy ↑
Development of mammary glands ↑
Ovaries
Estradiol
Uterus and other organs
Stimulates proliferation of endometrium
Menstrual cycle
Bone development ↑
Development of secondary female sex characteristics e.g., fat distribution, breasts, body hair ↑
Testes
Testosterone
Causes: Sexual differentiation to male phenotype
Formation of ejaculate
Spermatogenesis
Development of secondary male sex characteristics e.g., skeleton, muscles, body hair ↑
Protein synthesis ↑
Adrenal glands (cortex)
Cortisol
Proteins
Amino acids
Glucose
Proteolysis ↑
Protein synthesis ↓
Gluconeogenese ↑
Blut-Glucose ↑
Activity of the immune system ↓
Adrenal glands (cortex)
Aldosterone
Kidneys
ATP
$3Na^{\oplus}$
$2K^{\oplus}$
$ADP+P_i$
$3Na^{\oplus}$
$2K^{\oplus}$
$Na^{\oplus}$ retention ↑
$K^{\oplus}$ excretion ↑
Blood pressure ↑
Kidneys
Calcitriol
Gut
Bones
$Ca^{2\oplus}$
$Ca^{2\oplus}$- and phosphate resorption ↑
$Ca^{2\oplus}$ metabolism of bones ↑
Thyroid gland
Thyroxine
Embryo
O_2
H_2O
S
CO_2
$ADP+P_i$
ATP, Heat
Intermediary metabolism
Fetal development, growth, and maturation ↑
Basal metabolic rate ↑
Heat generation ↑
O_2 consumption ↑

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Metabolism of steroid hormones

A. Biosynthesis of steroid hormones ○

All steroid hormones are synthesized from **cholesterol.** The gonane core of cholesterol consists of 19 carbon atoms in four rings (A–D). The D ring carries a side chain of eight C atoms (see p. 54).

The cholesterol required for biosynthesis of the steroid hormones is obtained from various sources. It is either taken up as a constituent of LDL lipoproteins (see p. 278) into the hormone-synthesizing glandular cells, or synthesized by glandular cells themselves from acetyl-CoA (see p. 172). Excess cholesterol is stored in the form of fatty acid esters in lipid droplets. Hydrolysis allows rapid mobilization of the cholesterol from this reserve again.

Biosynthetic pathways. Only an overview of the synthesis pathways that lead to the individual hormones is shown here. Further details are given on p. 410.

Among the reactions involved, hydroxylations (H) are particularly numerous. These are catalyzed by specific *monooxygenases* ("hydroxylases") of the cytochrome P450 family (see p. 318). In addition, there are NADPH-dependent and $NADP^+$-dependent *hydrogenations* (Y) and *dehydrogenations* (D), as well as *cleavage* and *isomerization reactions* (S, I). The estrogens have a special place among the steroid hormones, as they are the only ones that contain an aromatic A ring. When this is formed, catalyzed by *aromatase,* the angular methyl group (C-19) is lost.

Pregnenolone is an important intermediate in the biosynthesis of most steroid hormones. It is identical to cholesterol with the exception of a shortened and oxidized side chain. Pregnenolone is produced in three steps by hydroxylation and cleavage in the side chain. Subsequent dehydrogenation of the hydroxyl group at C-3 (b) and shifting of the double bond from C-5 to C-4 results in the gestagen **progesterone**.

With the exception of calcitriol, all steroid hormones are derived from progesterone. Hydroxylations of progesterone at C atoms 17, 21, and 11 lead to the glucocorticoid **cortisol**. Hydroxylation at C-17 is omitted during synthesis of the mineralocorticoid **aldosterone.** Instead, the angular methyl group (C-18) is oxidized to the aldehyde group. During synthesis of the androgen **testosterone** from progesterone, the side chain is completely removed. Aromatization of the A ring, as mentioned above, finally leads to **estradiol**.

On the way to **calcitriol** (vitamin D hormone; see p. 342), another double bond in the B ring of cholesterol is first introduced. Under the influence of UV light on the skin, the B ring is then photochemically cleaved, and the secosteroid *cholecalciferol* arises (vitamin D_3; see p. 364). Two Cyt P450-dependent hydroxylations in the liver and kidneys produce the active vitamin D hormone (see p. 330).

B. Inactivation of steroid hormones ○

The steroid hormones are mainly inactivated in the liver, where they are either reduced or further hydroxylated and then conjugated with *glucuronic acid* or *sulfate* for excretion (see p. 316). The *reduction reactions* attack oxo groups and the double bond in ring A. A combination of several inactivation reactions gives rise to many different steroid metabolites that have lost most of their hormonal activity. Finally, they are excreted with the *urine* and also partly via the *bile.* Evidence of steroids and steroid metabolites in the urine is used to investigate the hormone metabolism.

Further information

Congenital defects in the biosynthesis of steroid hormones can lead to severe developmental disturbances. In the *adrenogenital syndrome* (AGS), which is relatively common, there is usually a defect in *21-hydroxylase,* which is needed for synthesis of cortisol and aldosterone from progesterone. Reduced synthesis of this hormone leads to increased formation of testosterone, resulting in masculinization of female fetuses. With early diagnosis, this condition can be avoided by providing the mother with hormone treatment before birth.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biosynthesis of steroid hormones

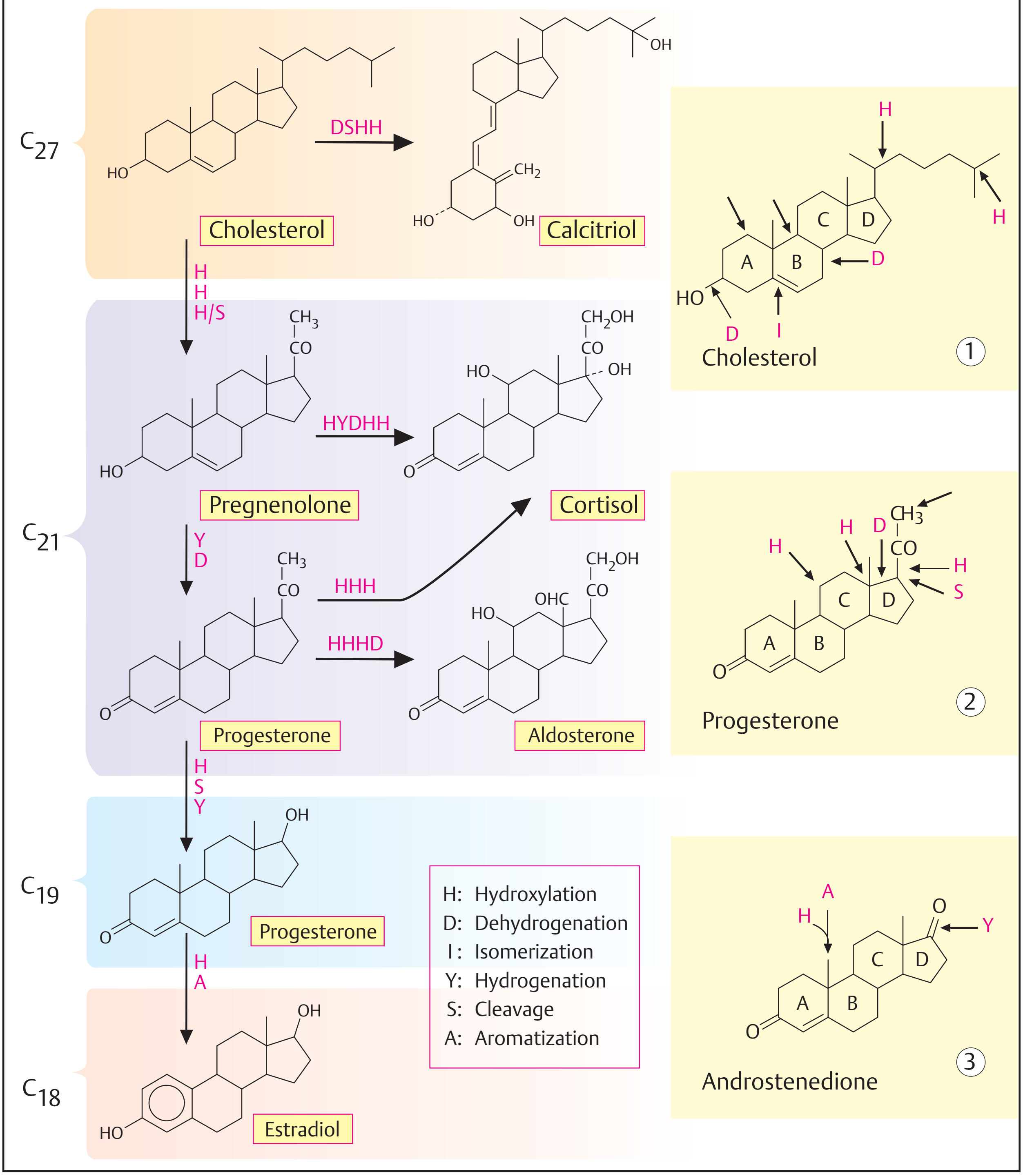
C27
Cholesterol
DSHH
Calcitriol
H
H
H/S
C21
Pregnenolone
HYDHH
Cortisol
Y
D
HHH
HHHD
Progesterone
Aldosterone
H
S
Y
C19
Progesterone
H
A
C18
Estradiol
Cholesterol
①
Progesterone
②
Androstenedione
③

H: Hydroxylation
D: Dehydrogenation
I: Isomerization
Y: Hydrogenation
S: Cleavage
A: Aromatization

B. Inactivation of steroid hormones

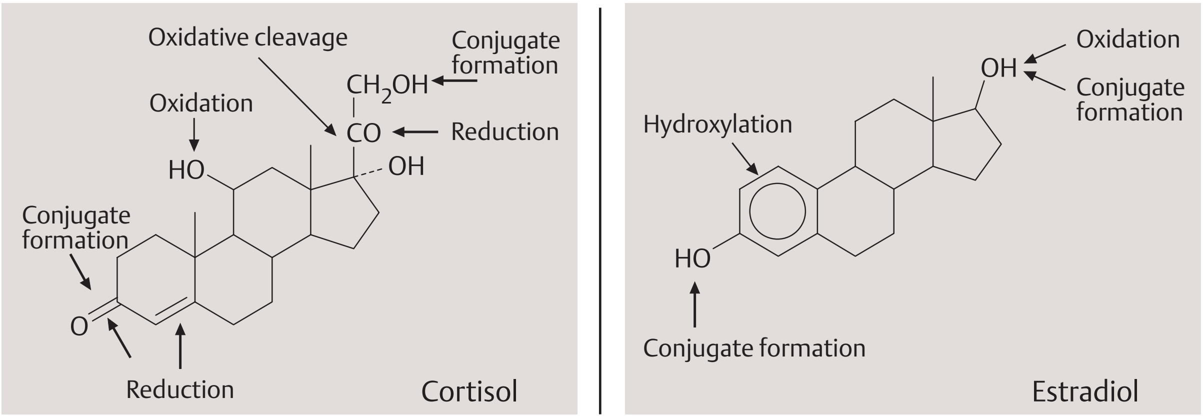
Oxidative cleavage
Conjugate formation
Oxidation
Reduction
Conjugate formation
Reduction
Cortisol
Hydroxylation
Oxidation
Conjugate formation
Conjugate formation
Estradiol

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Mechanism of action

A. Mechanism of action of lipophilic hormones ◑

Lipophilic signaling substances include the *steroid hormones, calcitriol,* the *iodothyronines* (T_3 and T_4), and *retinoic acid.* These hormones mainly act in the *nucleus* of the target cells, where they regulate gene transcription in collaboration with their receptors and with the support of additional proteins (known as coactivators and mediators; see p. 244). There are several effects of steroid hormones that are not mediated by transcription control. These alternative pathways for steroid effects have not yet been fully explained.

In the blood, there are a number of transport proteins for lipophilic hormones (see p. 276). Only the free hormone is able to penetrate the membrane and enter the cell. The hormone encounters its receptor in the nucleus (and sometimes also in the cytoplasm).

The **receptors** for lipophilic hormones are rare proteins. They occur in small numbers (10^3–10^4 molecules per cell) and show marked *specificity* and high *affinity* for the hormone ($K_d = 10^{-8}$–10^{-10} M). After binding to the hormone, the steroid receptors are able to bind as homodimers or heterodimers to *control elements* in the promoters of specific genes, from where they can influence the transcription of the affected genes—i. e., they act as *transcription factors.*

The illustration shows the particularly well-investigated mechanism of action for **cortisol**, which is unusual to the extent that the hormone–receptor complex already arises in the cytoplasm. The free receptor is present in the cytoplasm as a monomer in complex with the chaperone **hsp90** (see p. 232). Binding of cortisol to the complex leads to an *allosteric conformational change* in the receptor, which is then released from the hsp90 and becomes capable of DNA binding as a result of *dimerization.*

In the nucleus, the hormone–receptor complex binds to nucleotide sequences known as **hormone response elements** (HREs). These are short palindromic DNA segments that usually promote transcription as enhancer elements (see p. 244). The illustration shows the HRE for glucocorticoids (GRE; "n" stands for any nucleotide). Each hormone receptor only recognizes its "own" HRE and therefore only influences the transcription of genes containing that HRE. Recognition between the receptor and HRE is based on interaction between the amino acid residues in the DNA-binding domain (**B**) and the relevant bases in the HRE (emphasized in color in the structure illustrated).

As discussed on p. 244, the hormone receptor does not interact directly with the RNA polymerase, but rather—along with other transcription factors—with a coactivator/mediator complex that processes all of the signals and passes them on to the polymerase. In this way, hormonal effects lead within a period of minutes to hours to altered levels of mRNAs for key proteins in cellular processes ("cellular response").

B. Steroid receptors ○

The receptors for lipophilic signaling substances all belong to one *protein superfamily.* They are constructed in a modular fashion from **domains** with various lengths and functions. Starting from the N terminal, these are: the *regulatory domain,* the *DNA-binding domain,* a *nuclear localization sequence* (see p. 228), and the *hormone-binding domain* (see p. 73D).

The homology among receptors is particularly great in the area of the DNA-binding domain. The proteins have cysteine-rich sequences here that coordinatively bind zinc ions (**A**, Cys shown in yellow, Zn^{2+} in light blue). These centers, known as "zinc fingers" or "**zinc clusters**," stabilize the domains and support their dimerization, but do not take part in DNA binding directly. As in other transcription factors (see p. 118), "recognition helices" are responsible for that.

In addition to the receptors mentioned in **A**, the family of steroid receptors also includes the product of the oncogene *erb*-A (see p. 398), the receptor for the environmental toxin *dioxin,* and other proteins for which a distinct hormone ligand has not been identified (known as "orphan receptors"). Several steroid receptors—e. g., the retinoic acid receptor—form functional heterodimers with orphan receptors.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Mechanism of action of lipophilic hormones

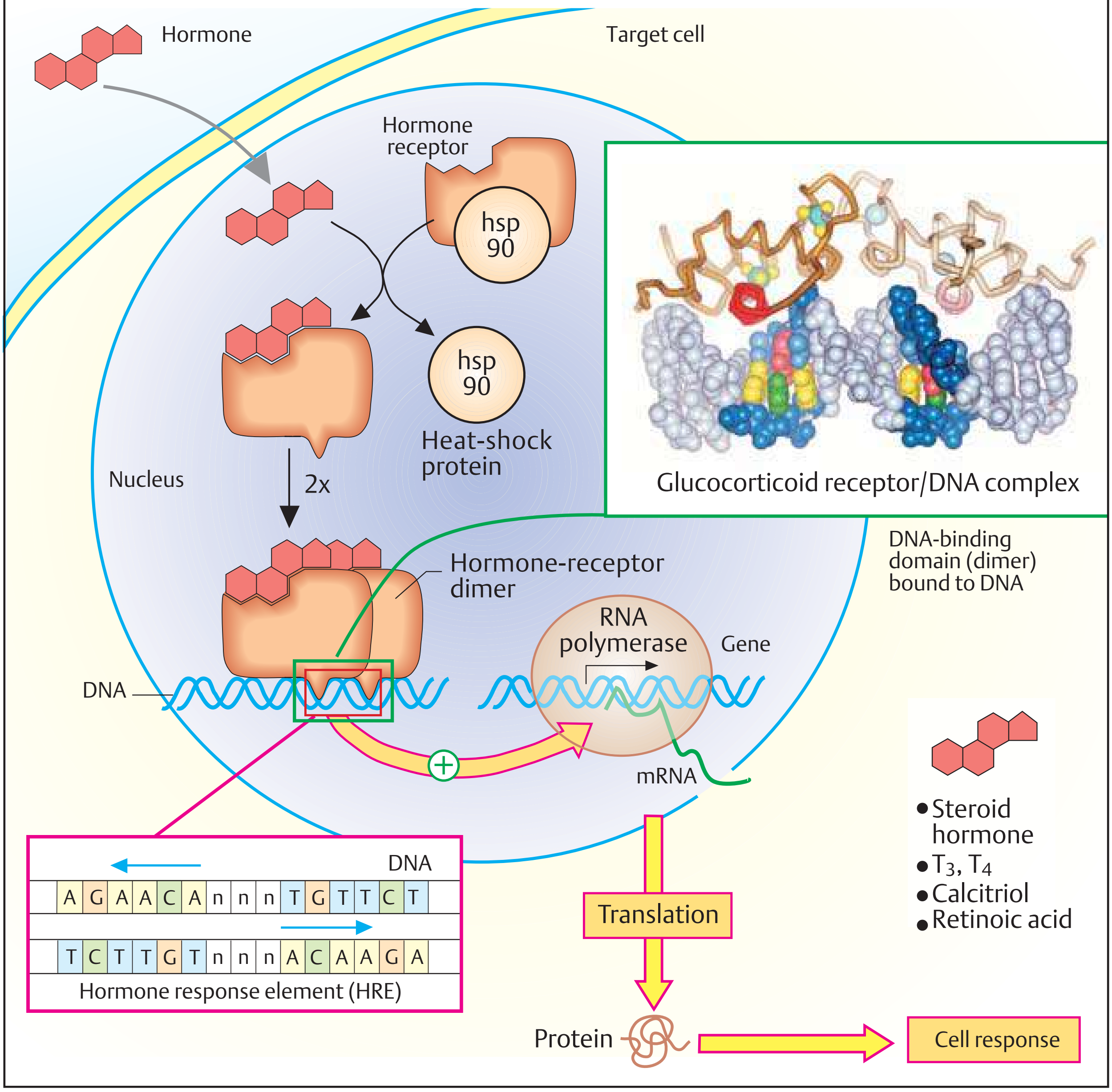
Hormone
Target cell
Hormone receptor
hsp 90
hsp 90
Heat-shock protein
Nucleus
2x
Glucocorticoid receptor/DNA complex
DNA-binding domain (dimer) bound to DNA
Hormone-receptor dimer
RNA polymerase
Gene
DNA
mRNA
DNA
A G A A C A n n n T G T T C T
T C T T G T n n n A C A A G A
Hormone response element (HRE)
Translation
Protein
Cell response
Steroid hormone
T_3, T_4
Calcitriol
Retinoic acid

B. Receptors of lipophilic hormones

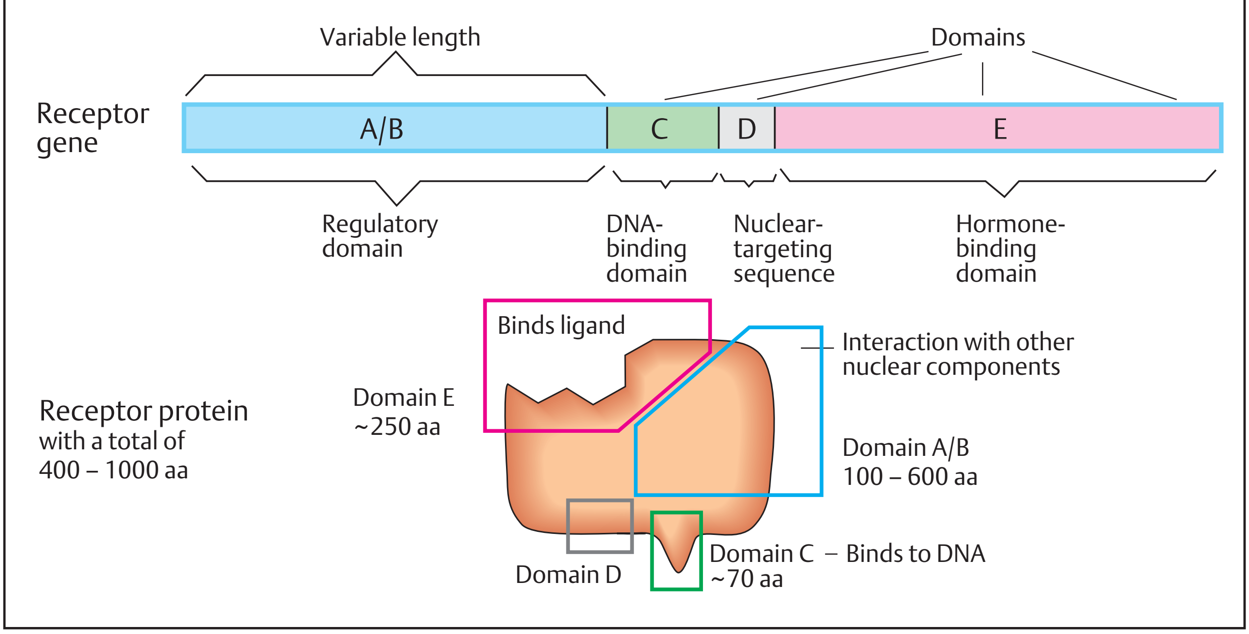
Variable length
Domains
Receptor gene
A/B
C
D
E
Regulatory domain
DNA-binding domain
Nuclear-targeting sequence
Hormone-binding domain
Binds ligand
Interaction with other nuclear components
Receptor protein with a total of 400 – 1000 aa
Domain E ~250 aa
Domain A/B 100 – 600 aa
Domain D
Domain C – Binds to DNA ~70 aa

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Hydrophilic hormones

The hydrophilic hormones are derived from amino acids, or are peptides and proteins composed of amino acids. Hormones with endocrine effects are synthesized in glandular cells and stored there in vesicles until they are released. As they are easily soluble, they do not need carrier proteins for transport in the blood. They bind on the plasma membrane of the target cells to receptors that pass the hormonal signal on (signal transduction; see p. 384). Several hormones in this group have paracrine effects—i. e., they only act in the immediate vicinity of their site of synthesis (see p. 372).

A. Signaling substances derived from amino acids ◑

Histamine, serotonin, melatonin, and the catecholamines dopa, dopamine, norepinephrine, and epinephrine are known as *"biogenic amines."* They are produced from amino acids by decarboxylation and usually act not only as hormones, but also as neurotransmitters.

Histamine, an important *mediator* (local signaling substance) and *neurotransmitter,* is mainly stored in tissue mast cells and basophilic granulocytes in the blood. It is involved in inflammatory and allergic reactions. "Histamine liberators" such as tissue hormones, type E immunoglobulins (see p. 300), and drugs can release it. Histamine acts via various types of receptor. Binding to H_1 receptors promotes contraction of smooth muscle in the bronchia, and dilates the capillary vessels and increases their permeability. Via H_2 receptors, histamine slows down the heart rate and promotes the formation of HCl in the gastric mucosa. In the brain, histamine acts as a neurotransmitter.

Epinephrine is a hormone synthesized in the adrenal glands from tyrosine (see p. 352). Its release is subject to neuronal control. This "emergency hormone" mainly acts on the blood vessels, heart, and metabolism. It constricts the blood vessels and thereby increases blood pressure (via α_1 and α_2 receptors); it increases cardiac function (via β_2 receptors); it promotes the degradation of glycogen into glucose in the liver and muscles (via β_2 receptors); and it dilates the bronchia (also via β_2 receptors).

B. Examples of peptide hormones and proteohormones ◑

Numerically the largest group of signaling substances, these arise by protein biosynthesis (see p. 382). The smallest peptide hormone, thyroliberin (362 Da), is a tripeptide. Proteohormones can reach masses of more than 20 kDa—e. g., thyrotropin (28 kDa). Similarities in the primary structures of many peptide hormones and proteohormones show that they are related to one another. They probably arose from common predecessors in the course of evolution.

Thyroliberin (thyrotropin-releasing hormone, TRH) is one of the neurohormones of the hypothalamus (see p. 330). It stimulates pituitary gland cells to secrete thyrotropin (TSH). TRH consists of three amino acids, which are modified in characteristic ways (see p. 353).

Thyrotropin (thyroid-stimulating hormone, TSH) and the related hormones *lutropin* (luteinizing hormone, LH) and *follitropin* (follicle-stimulating hormone, FSH) originate in the adenohypophysis. They are all dimeric glycoproteins with masses of around 28 kDa. Thyrotropin stimulates the synthesis and secretion of thyroxin by the thyroid gland.

Insulin (for the structure, see p. 70) is produced and released by the B cells of the pancreas and is released when the glucose level rises. Insulin reduces the blood sugar level by promoting processes that consume glucose—e. g., glycolysis, glycogen synthesis, and conversion of glucose into fatty acids. By contrast, it inhibits gluconeogenesis and glycogen degradation. The transmission of the insulin signal in the target cells is discussed in greater detail on p. 388.

Glucagon, a peptide of 29 amino acids, is a product of the A cells of the pancreas. It is the antagonist of insulin and, like insulin, mainly influences carbohydrate and lipid metabolism. Its effects are each opposite to those of insulin. Glucagon mainly acts via the second messenger cAMP (see p. 384).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Signaling substances derived from amino acids

Hormone	Sites of formation	Sites of action	Actions
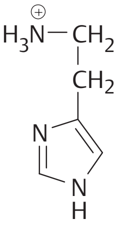 Histamine	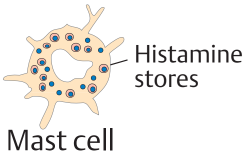 Mast cell 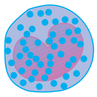 Basophilic granulocyte	Lungs 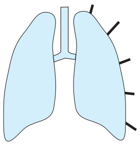 Stomach 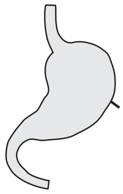	Width of bronchi ↓ Capillaries: width ↑ permeability ↑ Gastric acid secretion by parietal cells ↑
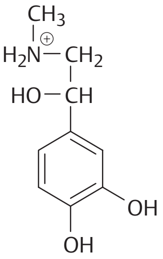 Epinephrine	Adrenal glands (medulla) 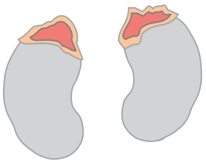	Heart 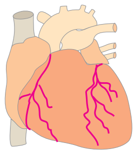 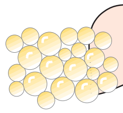 Adipose tissue 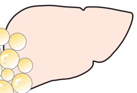 Liver Muscle	Cardiac output ↑ Width of blood vessels ↓ Blood pressure ↑ Metabolism: Glycogenolysis ↑ Blood glucose ↑ Lipolysis ↑

B. Examples of peptide hormones and proteohormones

Thyroliberin (TRH) 3 AA	Hypothalamus 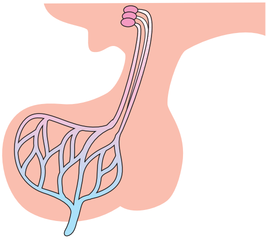	Pituitary 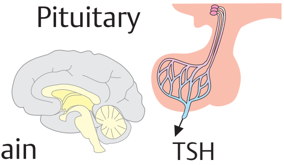 Brain	Thyrotropin secretion ↑ Neurotransmitter action
Thyrotropin (TSH) α chain 92 AA β chain 112 AA	Adeno-hypophysis	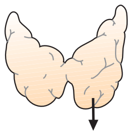 Thyroid gland Thyroxine	Synthesis and secretion of thyroxine ↑
Insulin A chain 21 AA B chain 30 AA	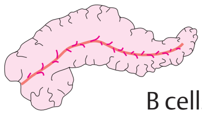 B cells Pancreas	Glucose → 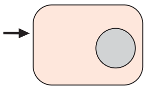 Glycogen ↑↓ Glucose Proteins ↑↓ Amino acids Fats ↑↓ Fatty acids	Glucose uptake by cells ↑ Blood glucose ↓ Storage compounds: formation ↑ degradation ↓
Glucagon 29 AA	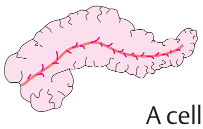 A cells Pancreas	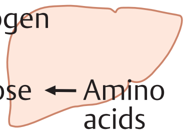 Glycogen ↓ Glucose ← Amino acids Fats ↓ Fatty acids ↓ Ketone bodies	Glycogenolysis ↑ Gluconeogenesis ↑ Blood glucose ↑ Ketone body formation ↑

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Metabolism of peptide hormones

Hydrophilic hormones and other water-soluble signaling substances have a variety of biosynthetic pathways. Amino acid derivatives arise in special *metabolic pathways* (see p. 352) or through *post-translational modification* (see p. 374). Proteohormones, like all proteins, result from *translation* in the ribosome (see p. 250). Small peptide hormones and neuropeptides, most of which only consist of 3–30 amino acids, are released from precursor proteins by *proteolytic degradation*.

A. Biosynthesis ○

The illustration shows the synthesis and processing of the precursor protein **proopiomelanocortin** (POMC) as an example of the biosynthesis of small peptides with signaling functions. POMC arises in cells of the adenohypophysis, and after processing in the rER and Golgi apparatus, it supplies the opiate-like peptides *met-enkephalin* and *β-endorphin* (implying "opio-"; see p. 352), three *melanocyte-stimulating hormones* (α-, β- and γ-MSH, implying "melano-"), and the glandotropic hormone *corticotropin* (ACTH, implying "-cortin"). Additional products of POMC degradation include two *lipotropins* with catabolic effects in the adipose tissue (β- and γ-LPH).

Some of the peptides mentioned are overlapping in the POMC sequence. For example, additional cleavage of ACTH gives rise to α-MSH and corticotropin-like intermediary peptide (CLIP). Proteolytic degradation of β-LPH provides γ-LPH and β-endorphin. The latter can be further broken down to yield met-enkephalin, while γ-LPH can still give rise to β-MSH (not shown). Due to the numerous derivative products with biological activity that it has, POMC is also known as a **polyprotein**. Which end product is formed and in what amounts depends on the activity of the proteinases in the ER that catalyze the individual cleavages.

The principles underlying protein synthesis and protein maturation (see pp. 230–233) can be summed up once again using the example of POMC:

[1] As a result of *transcription* of the POMC gene and *maturation of the hnRNA*, a mature **mRNA** consisting of some 1100 nucleotides arises, which is modified at both ends (see p. 246). This mRNA codes for prepro-POMC—i. e., a POMC protein that still has a signal peptide for the ER at the N terminus (see p. 230).

[2] **Prepro-POMC** arises through *translation* in the rough endoplasmic reticulum (rER). The growing peptide chain is introduced into the ER with the help of a signal peptide.

[3] Cleavage of the signal peptide and other modifications in the ER (formation of disulfide bonds, glycosylation, phosphorylation) give rise to the mature prohormone ("**pro-POMC**").

[4] The **neuropeptides** and **hormones** mentioned are now formed by *limited proteolysis* and stored in vesicles. Release from these vesicles takes place by exocytosis when needed.

The biosynthesis of peptide hormones and proteohormones, as well as their secretion, is controlled by higher-order regulatory systems (see p. 372). Calcium ions are among the substances involved in this regulation as *second messengers;* an increase in calcium ions stimulates synthesis and secretion.

B. Degradation and inactivation ◑

Degradation of peptide hormones often starts in the blood plasma or on the vascular walls; it is particularly intensive in the kidneys.

Several peptides that contain disulfide bonds (e. g., insulin) can be inactivated by reductive *cleavage of the disulfide bonds* (**1**). Peptides and proteins are also cleaved by *peptidases,* starting from one end of the peptide by *exopeptidases* (**2**), or in the middle of it by proteinases (endopeptidases, **3**). *Proteolysis* gives rise to a variety of hormone fragments, several of which are still biologically active. Some peptide hormones and proteohormones are removed from the blood by binding to their receptors with subsequent *endocytosis* of the hormone–receptor complex (**4**). They are then broken down in the lysosomes. All of the degradation reactions lead to amino acids, which become available to the metabolism again.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

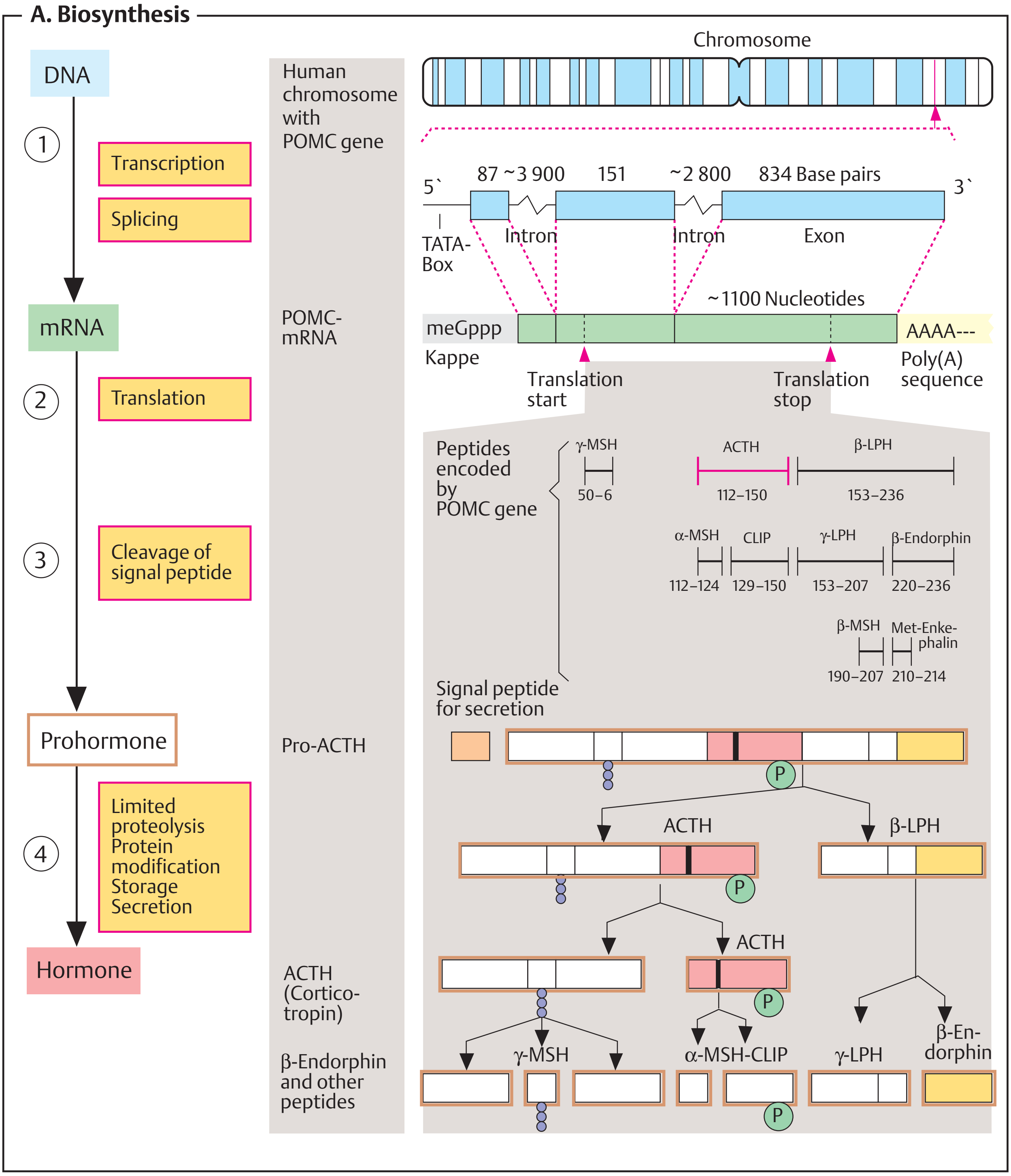

A. Biosynthesis

DNA
①
Transcription
Splicing
mRNA
②
Translation
③
Cleavage of signal peptide
Prohormone
④
Limited proteolysis
Protein modification
Storage
Secretion
Hormone
Human chromosome with POMC gene
Chromosome
5`
87 ~3 900
151
~2 800
834 Base pairs
3`
TATA-Box
Intron
Intron
Exon
POMC-mRNA
~1100 Nucleotides
meGppp
Kappe
AAAA---
Poly(A) sequence
Translation start
Translation stop
Peptides encoded by POMC gene
γ-MSH
50–6
ACTH
112–150
β-LPH
153–236
α-MSH
112–124
CLIP
129–150
γ-LPH
153–207
β-Endorphin
220–236
β-MSH
190–207
Met-Enkephalin
210–214
Signal peptide for secretion
Pro-ACTH
P
ACTH
β-LPH
P
ACTH
P
ACTH (Corticotropin)
γ-MSH
α-MSH-CLIP
γ-LPH
β-Endorphin
β-Endorphin and other peptides
P

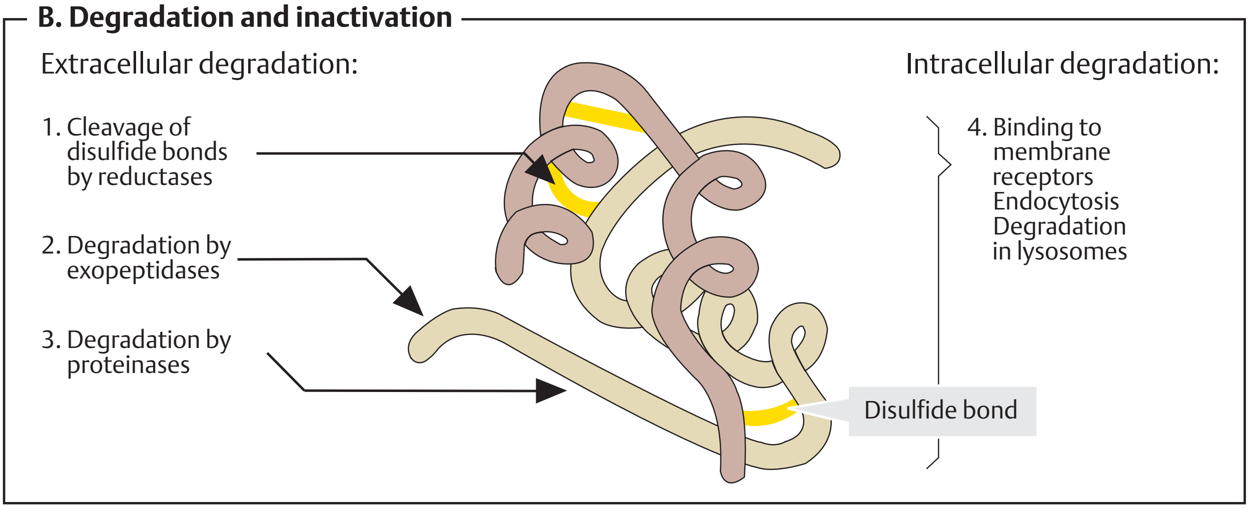

B. Degradation and inactivation

Extracellular degradation:
1. Cleavage of disulfide bonds by reductases
2. Degradation by exopeptidases
3. Degradation by proteinases
Intracellular degradation:
4. Binding to membrane receptors
Endocytosis
Degradation in lysosomes
Disulfide bond

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Mechanisms of action

The messages transmitted by hydrophilic signaling substances (see p. 380) are sent to the interior of the cell by *membrane receptors.* These bind the hormone on the outside of the cell and trigger a new second signal on the inside by altering their conformation. In the interior of the cell, this secondary signal influences the activity of enzymes or ion channels. Via further steps, *switching of the metabolism,* changes in the *cytoskeleton,* and activation or inhibition of *transcription factors* can occur *("signal transduction")* can occur.

A. Mechanisms of action ◑

Receptors are classified into three different types according to their structure (see also p. 224):

1. 1-Helix receptors (left) are proteins that span the membrane with only one α-helix. On their inner (cytoplasmic) side, they have domains with *allosterically activatable enzyme activity.* In most cases, these are *tyrosine kinases.*

Insulins (see p. 388), *growth factors,* and *cytokines* (see p. 392), for example, act via 1-helix receptors. Binding of the signaling substance leads to activation of internal kinase activity (in some cases, dimerization of the receptor is needed for this). The activated kinase phosphorylates itself using ATP (*autophosphorylation*), and also phosphorylates tyrosine residues of other proteins (known as *receptor substrates*). Adaptor proteins that recognize the phosphotyrosine residues bind to the phosphorylated proteins (see pp. 388, 392). They pass the signal on to other protein kinases.

2. Ion channels (center). These receptors contain *ligand-gated ion channels.* Binding of the signaling substance opens the channels for ions such as Na^+, K^+, Ca^{2+}, and Cl^-. This mechanism is mainly used by *neurotransmitters* such as *acetylcholine* (nicotinic receptor; see p. 224) and *GABA* (A receptor; see p. 354).

3. 7-Helix receptors (serpentine receptors, right) represent a large group of membrane proteins that transfer the hormone or transmitter signal, with the help of *G proteins* (see below), to *effector proteins* that alter the concentrations of *ions* and *second messengers* (see **B**).

B. Signal transduction by G proteins ◑

G proteins transfer signals from 7-helix receptors to effector proteins (see above). G protein are *heterotrimers* consisting of three different types of subunit (α, β, and γ; see p. 224). The α-subunit can bind GDP or GTP (hence the name "G protein") and has *GTPase activity.* Receptor-coupled G proteins are related to other GTP-binding proteins such as *Ras* (see pp. 388, 398) and *EF-Tu* (see p. 252).

G proteins are divided into several types, depending on their effects. *Stimulatory G proteins* (**G_s**) are widespread. They activate adenylate cyclases (see below) or influence ion channels. *Inhibitory G proteins* (**G_i**) *inhibit* adenylate cyclase. G proteins in the *G_q family* activate another effector enzyme—phospholipase c (see p. 386).

Binding of the signaling substance to a 7-helix receptor alters the receptor conformation in such a way that the corresponding G protein can attach on the inside of the cell. This causes the α-subunit of the G protein to exchange bound GDP for GTP (**1**). The G protein then separates from the receptor and dissociates into an α-subunit and a βγ-unit. Both of these components bind to other membrane proteins and alter their activity; *ion channels* are opened or closed, and *enzymes* are activated or inactivated.

In the case of the β_2-catecholamine receptor (illustrated here), the α-subunit of the G_s protein, by binding to adenylate cyclase, leads to the synthesis of the **second messenger** cAMP. cAMP activates protein kinase A, which in turn activates or inhibits other proteins (**2**; see p. 120).

The βγ-unit of the G protein stimulates a kinase (βARK, not shown), which phosphorylates the receptor. This reduces its affinity for the hormone and leads to binding of the blocking protein *arrestin.* The internal GTPase activity of the α-subunit hydrolyzes the bound GTP to GDP within a period of seconds to minutes, and thereby terminates the action of the G protein on the adenylate cyclase (**3**).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Mechanisms of action

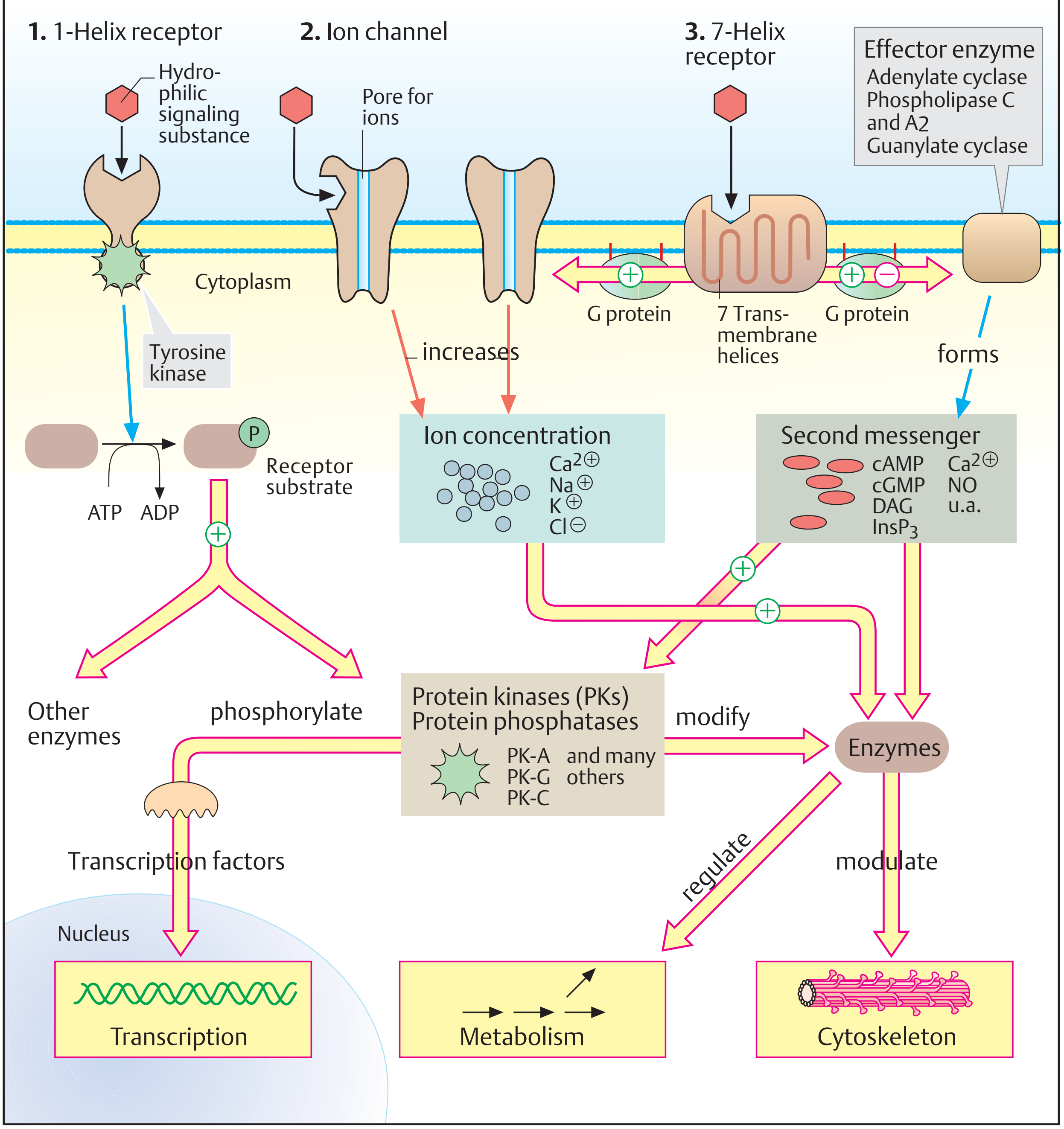
1. 1-Helix receptor
Hydro-philic signaling substance
Cytoplasm
Tyrosine kinase
ATP
ADP
P
Receptor substrate
2. Ion channel
Pore for ions
increases
Ion concentration
Ca^{2+}
Na^{+}
K^{+}
Cl^{-}
3. 7-Helix receptor
Effector enzyme
Adenylate cyclase
Phospholipase C and A2
Guanylate cyclase
G protein
7 Trans-membrane helices
G protein
forms
Second messenger
cAMP
cGMP
DAG
$InsP_3$
Ca^{2+}
NO
u.a.
Other enzymes
phosphorylate
Protein kinases (PKs)
Protein phosphatases
PK-A
PK-G
PK-C
and many others
modify
Enzymes
Transcription factors
Nucleus
regulate
modulate
Transcription
Metabolism
Cytoskeleton

B. Signal transduction by G proteins

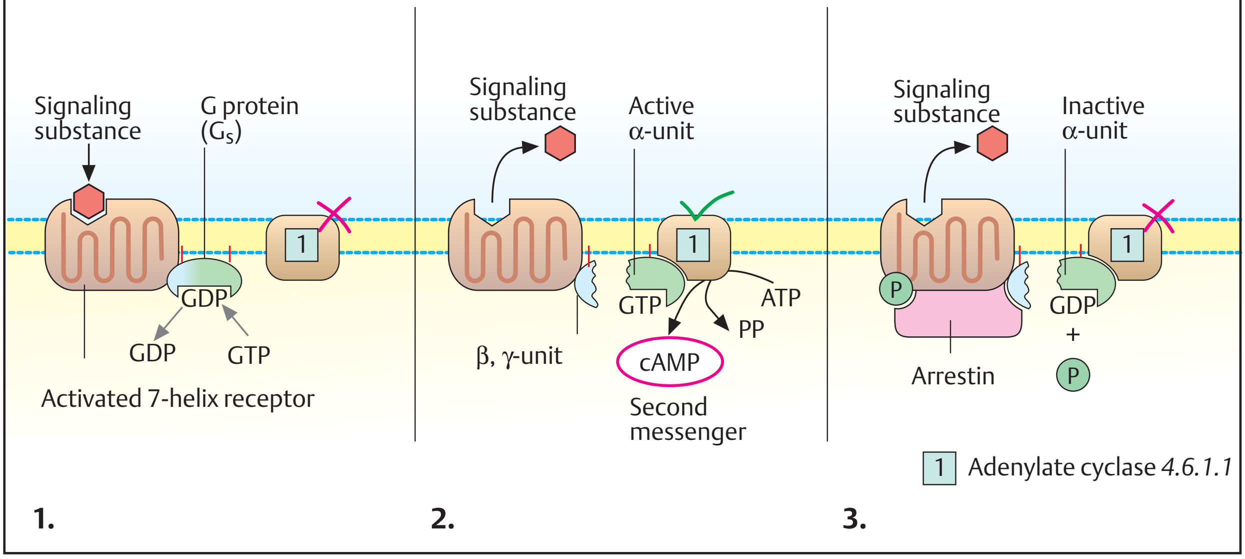
Signaling substance
G protein (G_s)
1
GDP
GDP
GTP
Activated 7-helix receptor
Signaling substance
Active α-unit
1
GTP
ATP
PP
cAMP
β, γ-unit
Second messenger
Signaling substance
Inactive α-unit
1
P
Arrestin
GDP
+
P
1 Adenylate cyclase 4.6.1.1
1.
2.
3.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Second messengers

Second messengers are *intracellular* chemical signals, the concentration of which is regulated by hormones, neurotransmitters, and other extracellular signals (see p. 384). They arise from easily available substrates and only have a short half-life. The most important second messengers are cAMP, cGMP, Ca^{2+}, inositol triphosphate (InsP$_3$), diacylglycerol (DAG), and nitrogen monoxide (NO).

A. Cyclic AMP ◑

Metabolism. The nucleotide **cAMP** (adenosine 3′,5′-cyclic monophosphate) is synthesized by membrane-bound *adenylate cyclases* [1] on the inside of the plasma membrane. The adenylate cyclases are a family of enzymes that cyclize ATP to cAMP by cleaving diphosphate (PP$_i$). The degradation of cAMP to AMP is catalyzed by *phosphodiesterases* [2], which are inhibited by *methylxanthines* such as caffeine, for example. By contrast, *insulin* activates the esterase and thereby reduces the cAMP level (see p. 388).

Adenylate cyclase activity is regulated by *G proteins* (G$_s$ and G$_i$), which in turn are controlled by extracellular signals via *7-helix receptors* (see p. 384). Ca^{2+}-calmodulin (see below) also activates specific adenylate cyclases.

Action. cAMP is an allosteric effector of *protein kinase A* (PK-A, [3]). In the inactive state, PK-A is a heterotetramer (C$_2$R$_2$), the catalytic subunits of which (C) are blocked by regulatory units (R; autoinhibition). When cAMP binds to the regulatory units, the C units separate from the R units and become enzymatically active. Active PK-A phosphorylates serine and threonine residues of more than 100 different proteins, enzymes, and transcription factors. In addition to cAMP, **cGMP** also acts as a second messenger. It is involved in sight (see p. 358) and in the signal transduction of NO (see p. 388).

B. Inositol 1,4,5-trisphosphate and diacylglycerol ◑

Type G$_q$ G proteins activate *phospholipase C* [4]. This enzyme creates two second messengers from the double-phosphorylated membrane phospholipid *phosphatidylinositol bisphosphate* (PInsP$_2$), i. e., inositol 1,4,5-trisphosphate (**InsP3**), which is soluble, and diacylglycerol (**DAG**). InsP$_3$ migrates to the endoplasmic reticulum (ER), where it opens Ca^{2+} channels that allow Ca^{2+} to flow into the cytoplasm (see **C**). By contrast, DAG, which is lipophilic, remains in the membrane, where it activates type C *protein kinases*, which phosphorylate proteins in the presence of Ca^{2+} ions and thereby pass the signal on.

C. Calcium ions ◑

Calcium level. Ca^{2+} (see p. 342) is a signaling substance. The concentration of Ca^{2+} ions in the cytoplasm is normally very low (10–100 nM), as it is kept down by ATP-driven Ca^{2+} pumps and Na^{+}/Ca^{2+} exchangers. In addition, many proteins in the cytoplasm and organelles bind calcium and thus act as Ca^{2+} buffers.

Specific signals (e. g., an action potential or second messenger such as InsP$_3$ or cAMP) can trigger a sudden increase in the cytoplasmic Ca^{2+} level to 500–1000 nM by opening Ca^{2+} channels in the plasma membrane or in the membranes of the *endoplasmic* or *sarcoplasmic reticulum. Ryanodine,* a plant substance, acts in this way on a specific channel in the ER. In the cytoplasm, the Ca^{2+} level always only rises very briefly (Ca^{2+} "spikes"), as prolonged high concentrations in the cytoplasm have cytotoxic effects.

Calcium effects. The biochemical effects of Ca^{2+} in the cytoplasm are mediated by special Ca^{2+}-binding proteins (*"calcium sensors"*). These include the *annexins, calmodulin,* and *troponin C* in muscle (see p. 334). **Calmodulin** is a relatively small protein (17 kDa) that occurs in all animal cells. Binding of four Ca^{2+} ions (light blue) converts it into a *regulatory element.* Via a dramatic conformational change (cf. 2a and 2b), Ca^{2+}-calmodulin enters into interaction with other proteins and modulates their properties. Using this mechanism, Ca^{2+} ions regulate the activity of enzymes, ion pumps, and components of the cytoskeleton.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Cyclic AMP

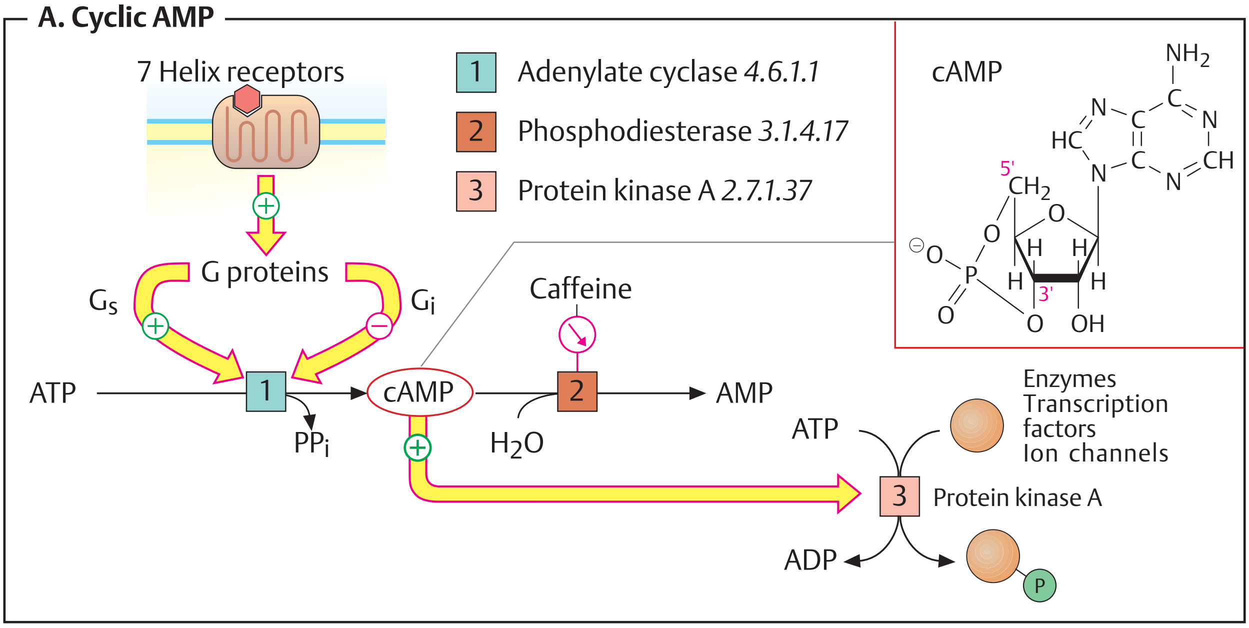
7 Helix receptors
1 Adenylate cyclase 4.6.1.1
2 Phosphodiesterase 3.1.4.17
3 Protein kinase A 2.7.1.37
cAMP
NH_2
N
C
HC
CH
5'
CH_2
O
P
H
3'
OH
G proteins
G_s
G_i
Caffeine
ATP
1
cAMP
2
AMP
PP_i
H_2O
Enzymes
Transcription factors
Ion channels
3
Protein kinase A
ADP
P

B. Inositol 1,4,5-trisphosphate and diacylglycerol

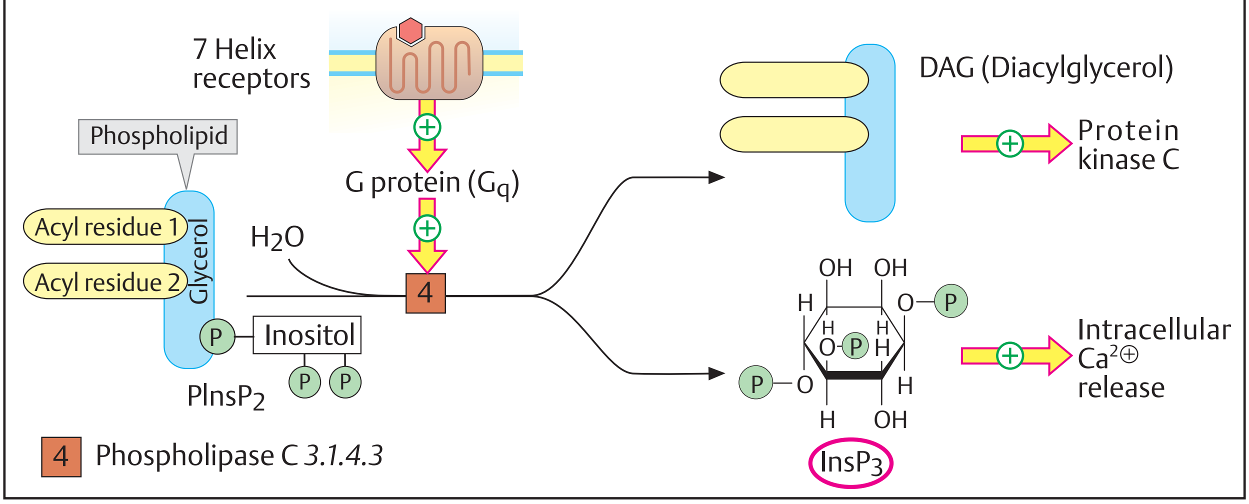
7 Helix receptors
Phospholipid
G protein (G_q)
Acyl residue 1
Acyl residue 2
Glycerol
H_2O
4
P
Inositol
$PInsP_2$
DAG (Diacylglycerol)
Protein kinase C
OH
H
O
P
Intracellular $Ca^{2⊕}$ release
$InsP_3$
4 Phospholipase C 3.1.4.3

C. Calcium ions

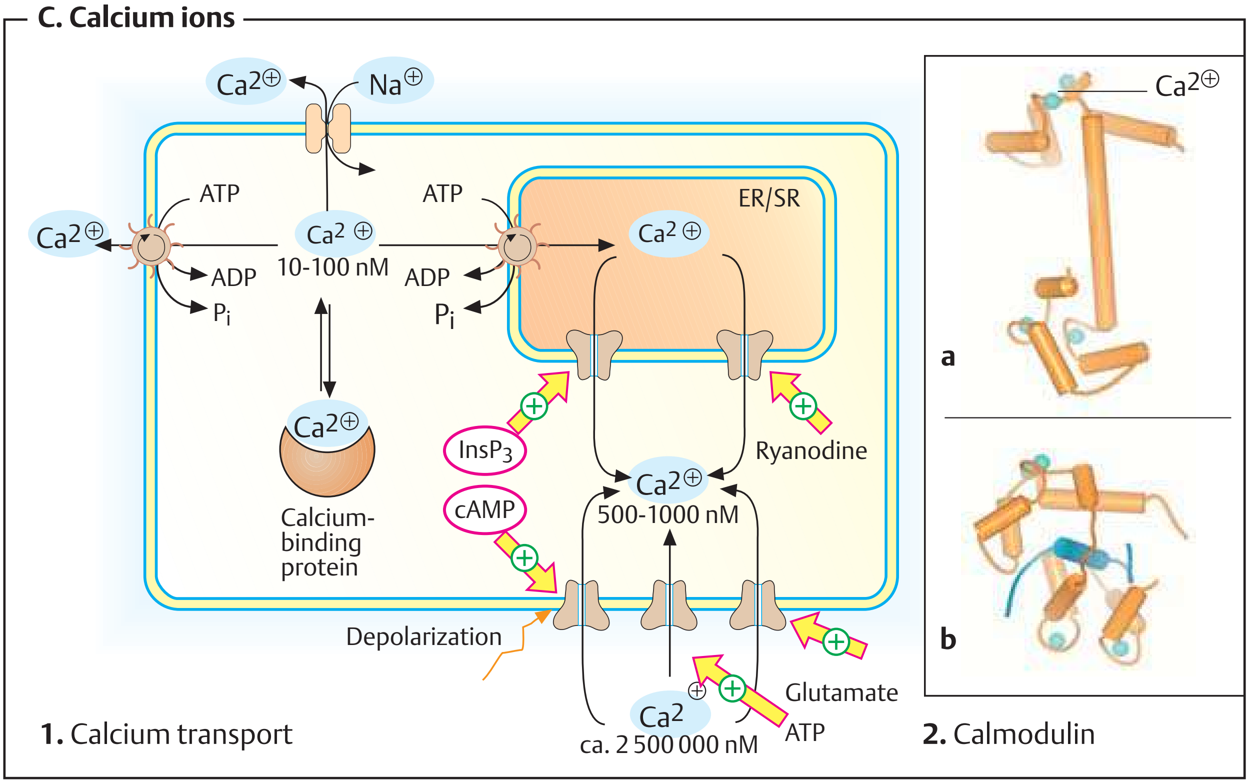
$Ca^{2⊕}$
$Na^⊕$
ATP
ADP
P_i
10-100 nM
ER/SR
Calcium-binding protein
$InsP_3$
cAMP
Ryanodine
500-1000 nM
Depolarization
Glutamate
ca. 2 500 000 nM
1. Calcium transport
a
b
2. Calmodulin

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Signal cascades

The signal transduction pathways that mediate the effects of the metabolic hormone **insulin** are of particular medical interest (see **A**). The mediator **nitrogen monoxide** (NO) is also clinically important, as it regulates vascular caliber and thus the body's perfusion with blood (see **B**).

A. Insulin: signal transduction ○

The diverse effects of insulin (see p. 160) are mediated by protein kinases that mutually activate each other in the form of enzyme cascades. At the end of this chain there are kinases that influence gene transcription in the nucleus by phosphorylating target proteins, or promote the uptake of glucose and its conversion into glycogen. The signal transduction pathways involved have not yet been fully explained. They are presented here in a simplified form.

The **insulin receptor** (top) is a dimer with subunits that have activatable tyrosine kinase domains in the interior of the cell (see p. 224). Binding of the hormone increases the tyrosine kinase activity of the receptor, which then phosphorylates itself and other proteins (**receptor substrates**) at various tyrosine residues. **Adaptor proteins**, which conduct the signal further, bind to the phosphotyrosine residues.

The effects of insulin on transcription are shown on the left of the illustration. Adaptor proteins **Grb-2** and **SOS** ("son of sevenless") bind to the phosphorylated **IRS** (insulin-receptor substrate) and activate the G protein **Ras** (named after its gene, the oncogene *ras*; see p. 398). Ras activates the protein kinase **Raf** (another oncogene product). Raf sets in motion a phosphorylation cascade that leads via the kinases **MEK** and **ERK** (also known as MAPK, "mitogen-activated protein kinase") to the phosphorylation of transcription factors in the nucleus.

Some of the effects of insulin on the **carbohydrate metabolism** (right part of the illustration) are possible without protein synthesis. In addition to Grb-2, another dimeric adaptor protein can also bind to phosphorylated IRS. This adaptor protein thereby acquires *phosphatidylinositol-3-kinase* activity (**PI_3K**) and, in the membrane, phosphorylates phospholipids from the phosphatidylinositol group (see p. 50) at position 3. Protein kinase **PDK-1** binds to these reaction products, becoming activated itself and in turn activating protein kinase B (**PK-B**).

This has several effects. In a manner not yet fully understood, PK-B leads to the fusion with the plasma membrane of vesicles that contain the glucose transporter Glut-4. This results in inclusion of Glut-4 in the membrane and thus to increased glucose uptake into the muscles and adipose tissue (see p. 160). In addition, PK-B inhibits glycogen synthase kinase 3 (**GSK-3**) by phosphorylation. As GSK-3 in turn inhibits glycogen synthase by phosphorylation (see p. 120), its inhibition by PK-B leads to *increased* glycogen synthesis. Protein phosphatase-1 (**PP-1**) converts glycogen synthase into its active form by dephosphorylation (see p. 120). PP-1 is also activated by insulin.

B. Nitrogen monoxide (NO) as a mediator ○

Nitrogen monoxide (NO) is a short-lived radical that functions as a locally acting mediator (see p. 370).

In a complex reaction, NO arises from arginine in the endothelial cells of the blood vessels [1]. The trigger for this is Ca^{2+}-calmodulin (see p. 386), which forms when there is an increase in the cytoplasmic Ca^{2+} level.

NO diffuses from the endothelium into the underlying vascular muscle cells, where it leads, as a result of activation of *guanylate cyclase* [2], to the formation of the second messenger **cGMP** (see pp. 358, 384). Finally, by activating a special protein kinase (**PK-G**), cGMP triggers relaxation of the smooth muscle and thus dilation of the vessels. The effects of atrionatriuretic peptide (**ANP**; see p. 328) in reducing blood pressure are also mediated by cGMP-induced vasodilation. In this case, cGMP is formed by the guanylate cyclase activity of the ANP receptor.

Further information

The drug *nitroglycerin* (glyceryl trinitrate), which is used in the treatment of *angina pectoris*, releases NO in the bloodstream and thereby leads to better perfusion of cardiac muscle.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Insulin: signal transduction

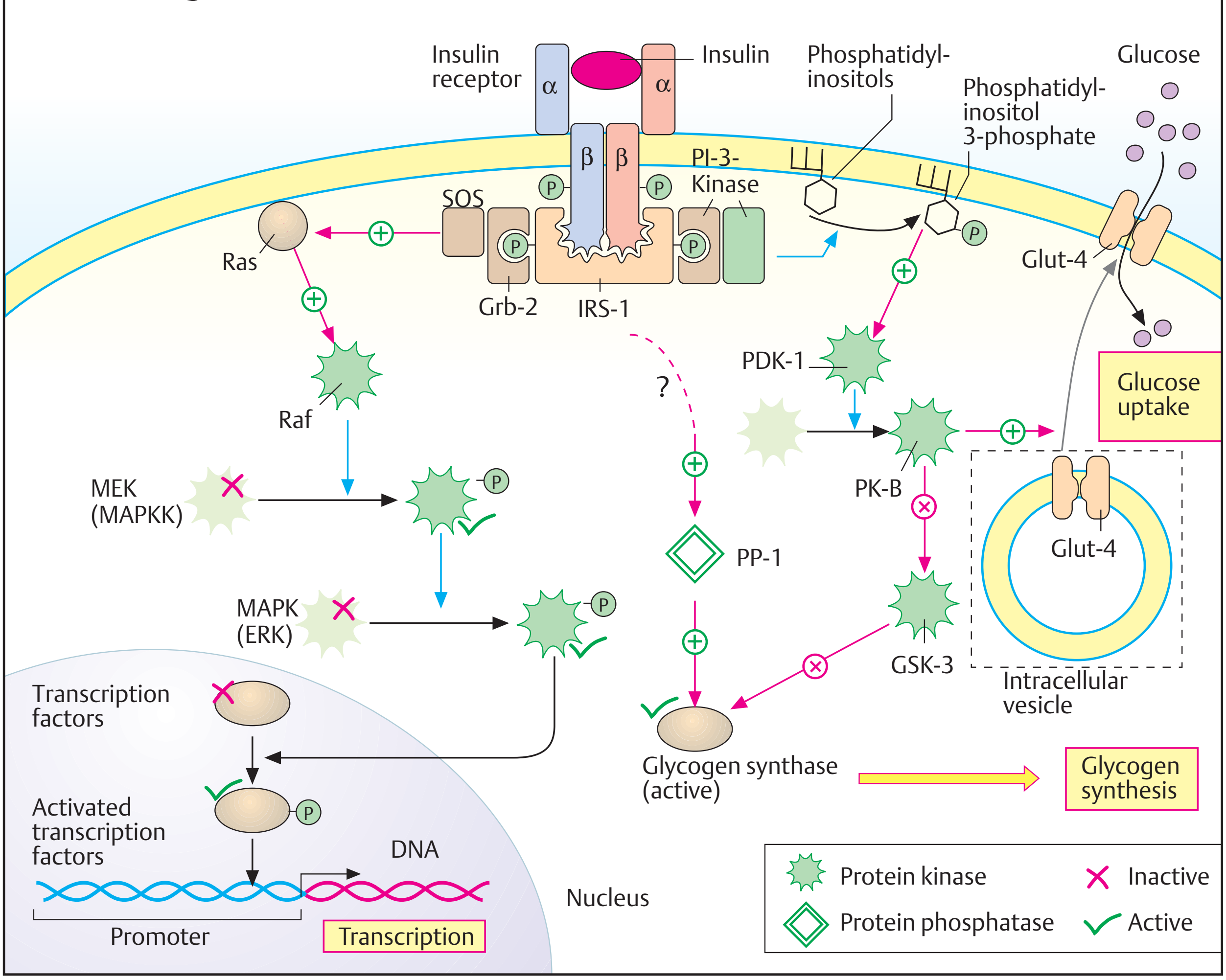
Insulin receptor
α
α
Insulin
β
β
P
P
Phosphatidyl-inositols
Glucose
Phosphatidyl-inositol 3-phosphate
PI-3-Kinase
SOS
P
P
P
Ras
Grb-2
IRS-1
Glut-4
PDK-1
Glucose uptake
?
Raf
PK-B
MEK (MAPKK)
P
PP-1
Glut-4
MAPK (ERK)
P
GSK-3
Intracellular vesicle
Transcription factors
Glycogen synthase (active)
Glycogen synthesis
Activated transcription factors
P
DNA
Nucleus
Promoter
Transcription
Protein kinase
Inactive
Protein phosphatase
Active

B. Nitrogen monoxide (NO) as a mediator

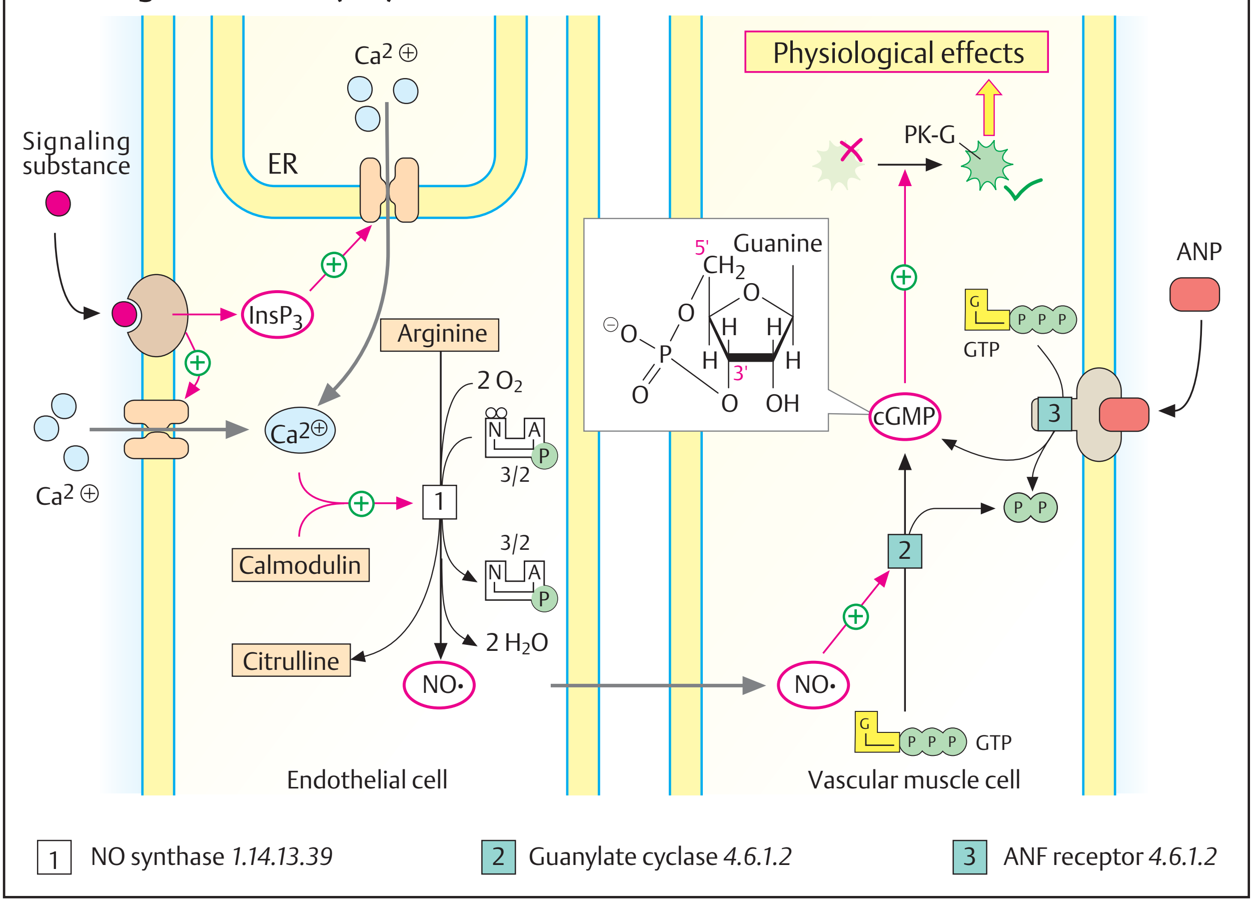
Ca2⊕
Physiological effects
Signaling substance
ER
PK-G
ANP
5'
Guanine
CH2
O
O
⊖O
P
H
H
H
H
3'
O
O
OH
InsP3
Arginine
G
P P P
GTP
2 O2
Ca2⊕
N A
P
3/2
cGMP
3
Ca2⊕
1
Calmodulin
3/2
N A
P
P P
2
Citrulline
2 H2O
NO·
NO·
G
P P P
GTP
Endothelial cell
Vascular muscle cell
1 NO synthase 1.14.13.39
2 Guanylate cyclase 4.6.1.2
3 ANF receptor 4.6.1.2

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Eicosanoids

The eicosanoids are a group of signaling substances that arise from the C-20 fatty acid *arachidonic acid* and therefore usually contain 20 C atoms (Greek *eicosa* = 20). As mediators, they influence a large number of physiological processes (see below). Eicosanoid metabolism is therefore an important drug target. As short-lived substances, eicosanoids only act in the vicinity of their site of synthesis (paracrine effect; see p. 372).

A. Eicosanoids ○

Biosynthesis. Almost all of the body's cells form eicosanoids. Membrane phospholipids that contain the polyunsaturated fatty acid **arachidonic acid** (20:4; see p. 48) provide the starting material.

Initially, *phospholipase* A_2 [1] releases the arachidonate moiety from these phospholipids. The activity of phospholipase A_2 is strictly regulated. It is activated by hormones and other signals via *G proteins*. The arachidonate released is a signaling substance itself. However, its metabolites are even more important.

Two different pathways lead from arachidonate to **prostaglandins**, **prostacyclins**, and **thromboxanes**, on the one hand, or **leukotrienes** on the other. The key enzyme for the first pathway is *prostaglandin synthase* [2]. Using up O_2, it catalyzes in a two-step reaction the cyclization of arachidonate to prostaglandin H_2, the parent substance for the prostaglandins, prostacyclins, and thromboxanes. Acetylsalicylic acid (aspirin) irreversibly acetylates a serine residue near the active center of prostaglandin synthase, so that access for substrates is blocked (see below).

As a result of the action of *lipoxygenases* [3], hydroxyfatty acids and hydroperoxyfatty acids are formed from arachidonate, from which elimination of water and various conversion reactions give rise to the leukotrienes. The formulae only show one representative from each of the various groups of eicosanoids.

Effects. Eicosanoids act via membrane receptors in the immediate vicinity of their site of synthesis, both on the synthesizing cell itself (*autocrine* action) and on neighboring cells (*paracrine* action). Many of their effects are mediated by the second messengers cAMP and cGMP.

The eicosanoids have a very wide range of physiological effects. As they can stimulate or inhibit smooth-muscle contraction, depending on the substance concerned, they affect blood pressure, respiration, and intestinal and uterine activity, among other properties. In the stomach, prostaglandins inhibit HCl secretion via G_i proteins (see p. 270). At the same time, they promote mucus secretion, which protects the gastric mucosa against the acid. In addition, prostaglandins are involved in bone metabolism and in the activity of the sympathetic nervous system. In the immune system, prostaglandins are important in the inflammatory reaction. Among other things, they attract leukocytes to the site of infection. Eicosanoids are also decisively involved in the development of pain and fever. The thromboxanes promote thrombocyte aggregation and other processes involved in hemostasis (see p. 290).

Metabolism. Eicosanoids are inactivated within a period of seconds to minutes. This takes place by enzymatic reduction of double bonds and dehydrogenation of hydroxyl groups. As a result of this rapid degradation, their range is very limited.

Further information

Acetylsalicylic acid and related non-steroidal anti-inflammatory drugs (NSAIDs) selectively inhibit the *cyclooxygenase activity* of prostaglandin synthase [2] and consequently the synthesis of most eicosanoids. This explains their analgesic, antipyretic, and antirheumatic effects. Frequent side effects of NSAIDs also result from inhibition of eicosanoid synthesis. For example, they impair hemostasis because the synthesis of thromboxanes by thrombocytes is inhibited. In the stomach, NSAIDs increase HCl secretion and at the same time inhibit the formation of protective mucus. Long-term NSAID use can therefore damage the gastric mucosa.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Eicosanoids

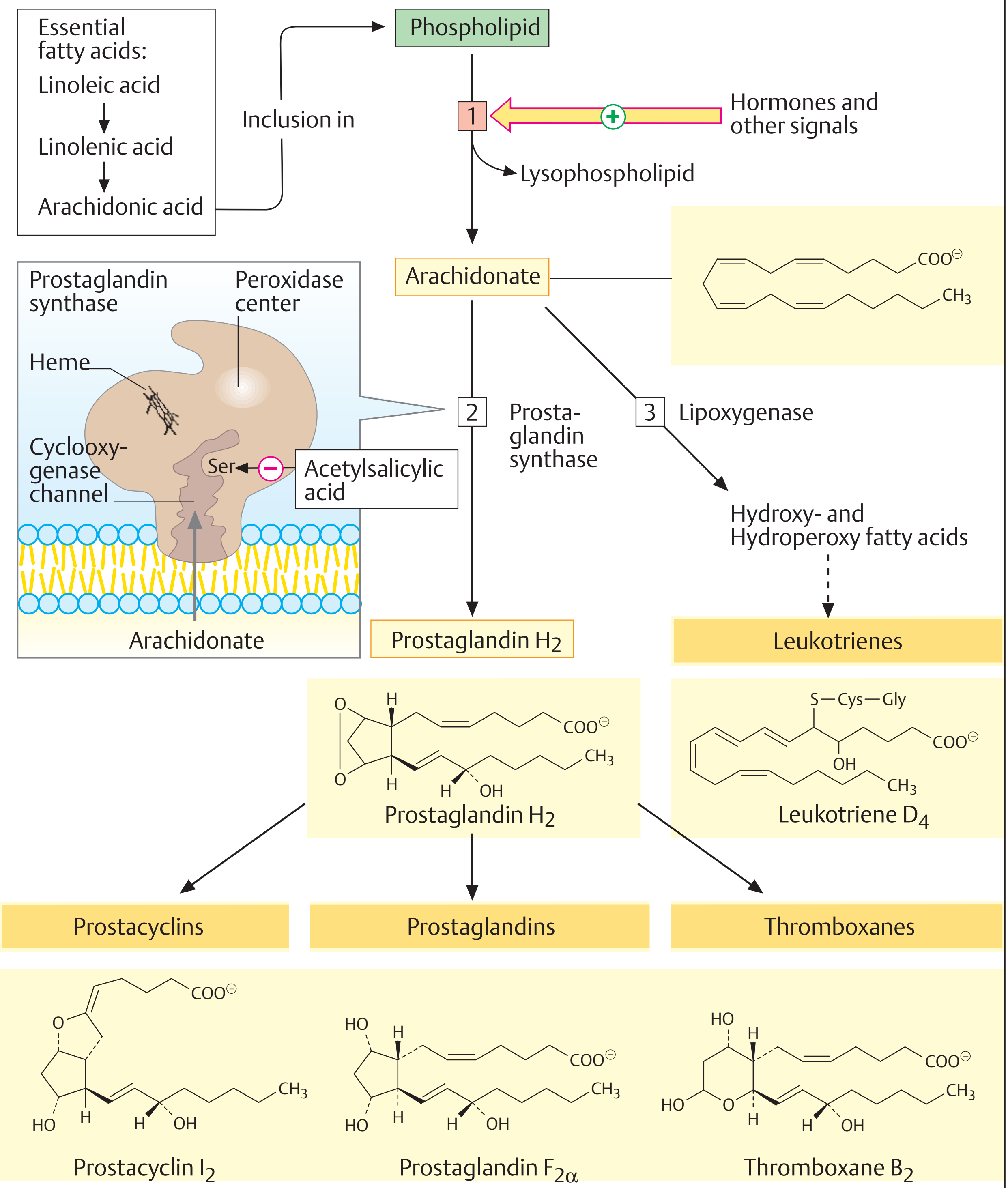
Essential fatty acids:
Linoleic acid
Linolenic acid
Arachidonic acid
Inclusion in
Phospholipid
1
Hormones and other signals
Lysophospholipid
Arachidonate
COO⊖
CH3
Prostaglandin synthase
Peroxidase center
Heme
Cyclooxygenase channel
Ser
Acetylsalicylic acid
Arachidonate
2
Prostaglandin synthase
3
Lipoxygenase
Hydroxy- and Hydroperoxy fatty acids
Prostaglandin H2
Leukotrienes
O
H
O
H
H
OH
COO⊖
CH3
Prostaglandin H2
S—Cys—Gly
COO⊖
OH
CH3
Leukotriene D4
Prostacyclins
Prostaglandins
Thromboxanes
COO⊖
O
CH3
HO
H
H
OH
Prostacyclin I2
HO
H
COO⊖
CH3
HO
H
H
OH
Prostaglandin F2α
HO
H
COO⊖
CH3
HO
O
H
H
OH
Thromboxane B2

Effects: Stimulation of

Contraction of smooth muscle
Biosynthesis of steroid hormones
Gastric juice secretion

Hormone-controlled lipases
Thrombocyte aggregation
Pain production
Inflammatory response

1 Phospholipase A_2 *3.1.1.4*

2 Prostaglandin H-synthase [heme] (dioxygenase + peroxidase) *1.14.99.1*

3 Arachidonate lipoxygenases *1.13.11. n*

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Cytokines

A. Cytokines ◑

Cytokines are hormone-like *peptides* and *proteins* with signaling functions, which are synthesized and released by cells of the immune system and other cell types. Their numerous biological functions operate in three areas: they regulate the *development and homeostasis of the immune system;* they control the *hematopoietic system;* and they are involved in *non-specific defense,* influencing inflammatory processes, blood coagulation, and blood pressure. In general, cytokines regulate the growth, differentiation, and survival of cells. They are also involved in regulating apoptosis (see p. 396).

There is an extremely large number of cytokines; only the most important representatives are listed opposite. The cytokines include *interleukins* (IL), *lymphokines, monokines, chemokines, interferons* (IFN), and *colony-stimulating factors* (CSF). Via interleukins, immune cells stimulate the proliferation and activity of other immune cells (see p. 294). Interferons are used medically in the treatment of viral infections and other diseases.

Although cytokines rarely show structural homologies with each other, their effects are often very similar. The cytokines differ from *hormones* (see p. 370) only in certain respects: they are released by many different cells, rather than being secreted by defined glands, and they regulate a wider variety of target cells than the hormones.

B. Signal transduction in the cytokines ○

As peptides or proteins, the cytokines are **hydrophilic signaling substances** that act by binding to *receptors* on the cell surface (see p. 380). Binding of a cytokine to its receptor (**1**) leads via several intermediate steps (**2** –**5**) to the activation of transcription of specific genes (**6**).

In contrast to the receptors for insulin and growth factors (see p. 388), the **cytokine receptors** (with a few exceptions) have *no* tyrosine kinase activity. After binding of cytokine (**1**), they associate with one another to form homodimers, join together with other signal transduction proteins (STPs) to form dimers, or promote dimerization of other STPs (2). Class I cytokine receptors interact with three different STPs (gp130, β_c, and γ_c). The STPs themselves do not bind cytokines, but conduct the signal to tyrosine kinases (**3**). The fact that different cytokines can activate the same STP via their receptors explains the overlapping biological activity of some cytokines.

As an example of the signal transduction pathway in cytokines, the illustration shows the way in which the **IL-6 receptor**, after binding its ligand **IL-6** (**1**), induces the dimerization of the STP **gp130** (**2**). The dimeric gp130 binds cytoplasmic *tyrosine kinases* from the Jak family ("**Janus kinases**," with two kinase centers) and activates them (**3**). The Janus kinases phosphorylate cytokine receptors, STPs, and various cytoplasmic proteins that conduct the signal further. In addition, they phosphorylate transcription factors known as **STATs** ("**s**ignal **t**ransducers and **a**ctivators of **t**ranscription"). STATs are among the proteins that have an *SH2 domain* and are able to bind phosphotyrosine residues (see p. 388). They therefore bind to cytokine receptors that have been phosphorylated by Janus kinases. When STATs are then also phosphorylated themselves (**4**), they are converted into their active form and become dimers (**5**). After transfer to the nucleus, they bind—along with auxiliary proteins as transcription factors—to the promoters of inducible genes and in this way regulate their transcription (**6**).

The activity of the cytokine receptors is terminated by *protein phosphatases,* which hydrolytically cleave the phosphotyrosine residues. Several cytokine receptors are able to lose their ligand-binding extracellular domain by proteolysis (not shown). The extracellular domain then appears in the blood, where it competes for cytokines. This reduces the effective cytokine concentration.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Cytokines

IL-1	Interleukin 1	G-CSF	Granulocyte colony-stimulating factor
IL-2	Interleukin 2	GM-CSF	Granulocyte/macrophage colony-stimulating factor
IL-3	Interleukin 3		
IL-4	Interleukin 4	MIF	Macrophage migration inhibitory factor
IL-5	Interleukin 5	M-CSF	Monocyte colony-stimulating factor
IL-6	Interleukin 6	TNFα	Tumor necrosis factor- α
IFN-α	Interferon α	TNFβ	Tumor necrosis factor- β
IFN-β	Interferon β		
IFN-γ	Interferon γ	and others	

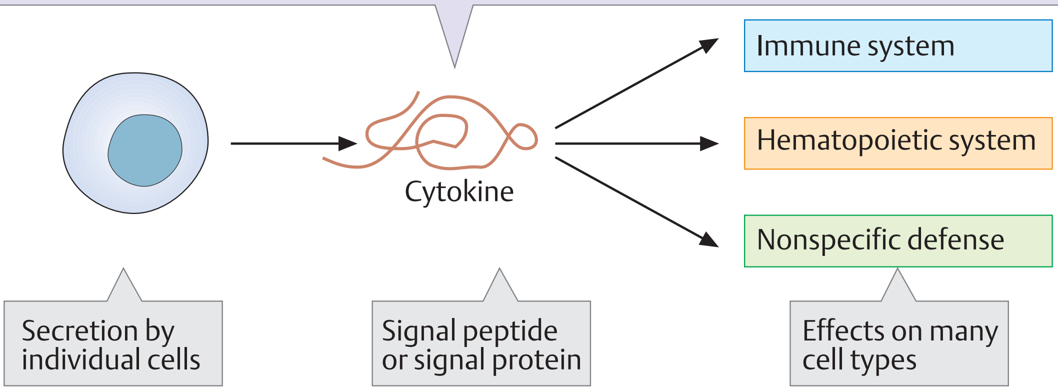
Cytokine
Immune system
Hematopoietic system
Nonspecific defense
Secretion by individual cells
Signal peptide or signal protein
Effects on many cell types

B. Signal transduction in the cytokines

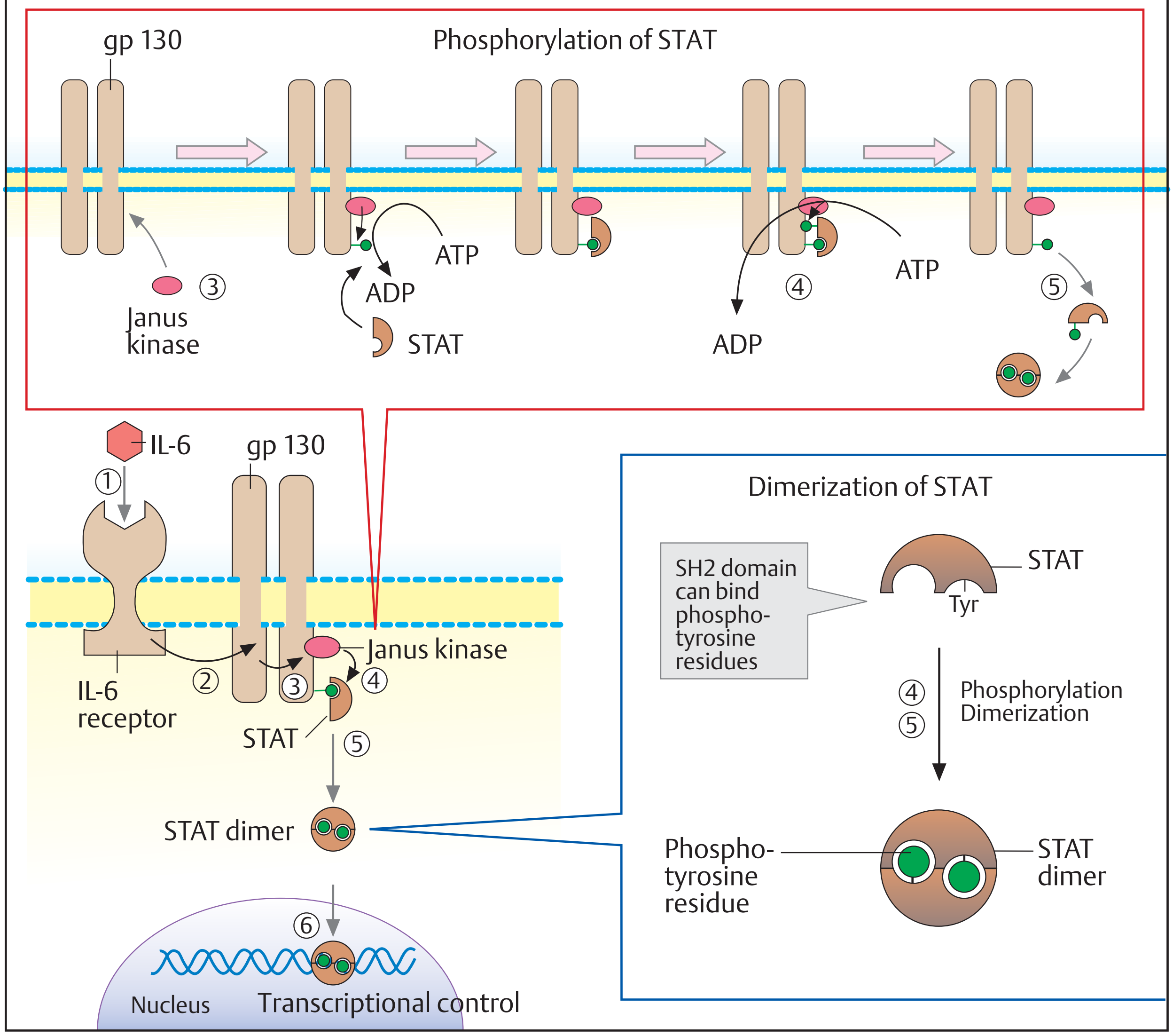
gp 130
Phosphorylation of STAT
③
Janus kinase
ATP
ADP
STAT
④
ATP
ADP
⑤
IL-6
①
gp 130
IL-6 receptor
②
③
Janus kinase
④
STAT
⑤
STAT dimer
⑥
Nucleus
Transcriptional control
Dimerization of STAT
SH2 domain can bind phospho-tyrosine residues
STAT
Tyr
④
⑤
Phosphorylation
Dimerization
Phospho-tyrosine residue
STAT dimer

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Cell cycle

A. Cell cycle ◑

Proliferating cells undergo a cycle of division (the cell cycle), which lasts approximately 24 hours in mammalian cells in cell culture. The cycle is divided into four different phases (G_1, S, G_2, and M—in that sequence).

Fully differentiated animal cells only divide rarely. These cells are in the so-called **G_0 phase**, in which they can remain permanently. Some G_0 cells return to the **G_1 phase** again under the influence of mitogenic signals (growth factors, cytokines, tumor viruses, etc.), and after crossing a *control point* (G_1 to S), enter a new cycle. DNA is replicated (see p. 240) during the **S phase**, and new chromatin is formed. Particularly remarkable in morphological terms is the actual mitosis (**M phase**), in which the chromosomes separate and two daughter cells are formed. The M and S phases are separated by two segments known as the **G_1** and **G_2** phases (the G stands for "gap"). In the G_1 phase, the duration of which can vary, the cell grows by de novo synthesis of cell components. Together, the G_1, G_0, S, and G_2 phases are referred to as the **interphase**, which alternates in the cell cycle with the short M phase.

B. Control of the cell cycle ○

The progression of the cell cycle is regulated by **interconversion processes**. In each phase, special *Ser/Thr-specific protein kinases* are formed, which are known as cyclin-dependent kinases (**CDKs**). This term is used because they have to bind an activator protein (**cyclin**) in order to become active. At each control point in the cycle, specific CDKs associate with equally phase-specific cyclins. If there are no problems (e.g., DNA damage), the CDK–cyclin complex is activated by phosphorylation and/or dephosphorylation. The activated complex in turn phosphorylates **transcription factors**, which finally lead to the formation of the proteins that are required in the cell cycle phase concerned (enzymes, cytoskeleton components, other CDKs, and cyclins). The activity of the CDK–cyclin complex is then terminated again by proteolytic **cyclin degradation**.

The above outline of cell cycle progression can be examined here in more detail using the G_2–M transition as an example.

Entry of animal cells into mitosis is based on the "mitosis-promoting factor" (**MPF**). MPF consists of **CDK1** (cdc2) and **cyclin B**. The intracellular concentration of cyclin B increases constantly until mitosis starts, and then declines again rapidly (top left). MPF is initially inactive, because CDK1 is phosphorylated and cyclin B is dephosphorylated (top center). The M phase is triggered when a *protein phosphatase* [1] dephosphorylates the CDK while cyclin B is phosphorylated by a *kinase* [2]. In its active form, MPF phosphorylates various proteins that have functions in mitosis—e.g., *histone H1* (see p. 238), components of the cytoskeleton such as the *laminins* in the nuclear membrane, *transcription factors, mitotic spindle* proteins, and various *enzymes.*

When mitosis has been completed, cyclin B is marked with **ubiquitin** and broken down proteolytically by *proteasomes* (see p. 176). Protein phosphatases then regain control and dephosphorylate the proteins involved in mitosis. This returns the cell to the interphase.

Further information

The G_1–S transition (not shown) is particularly important for initiating the cell cycle. It is triggered by the CDK4–cyclin D complex, which by phosphorylating the protein **pRb** releases the transcription factor E2F previously bound to pRb. This activates the transcription of genes needed for DNA replication.

If the DNA is damaged by mutagens or ionizing radiation, the protein **p53** initially delays entry into the S phase. If the DNA repair system (see p. 256) does not succeed in removing the DNA damage, p53 forces the cell into apoptosis (see p. 396). The genes coding for pRb and p53 belong to the **tumor-suppressor genes** (see p. 398). In many tumors (see p. 400), these genes are in fact damaged by mutation.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Cell cycle

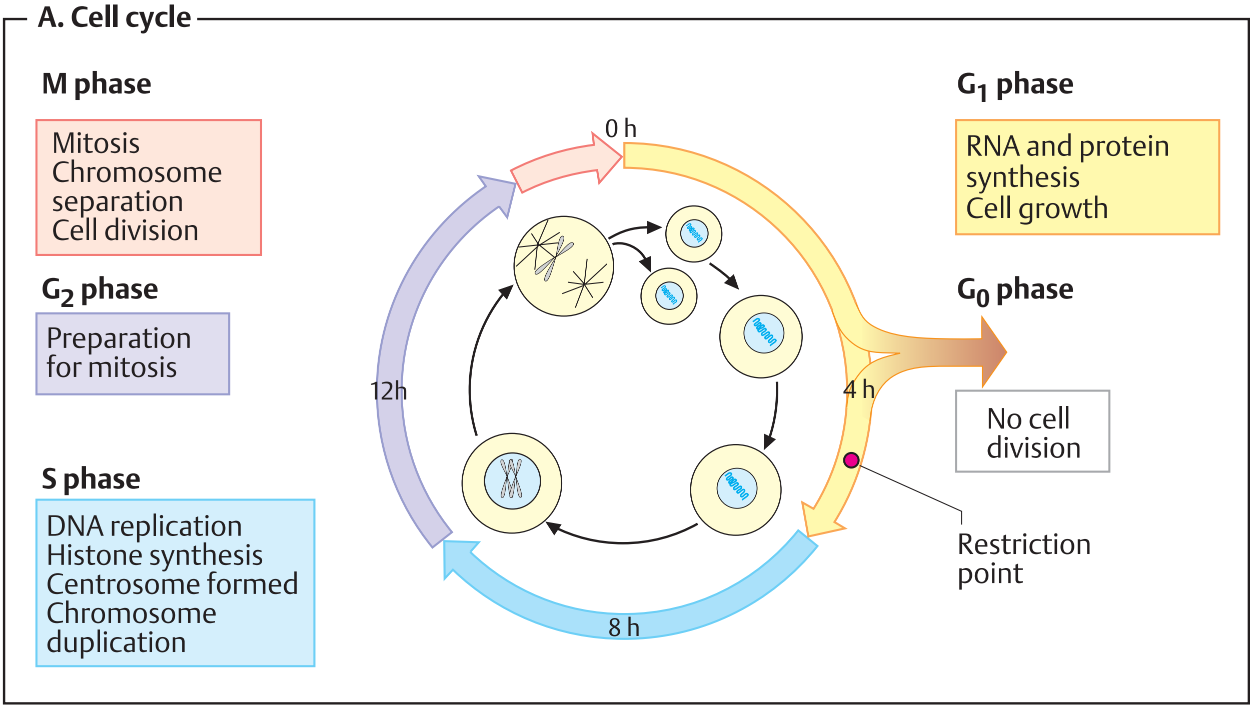
M phase
Mitosis
Chromosome separation
Cell division
G$_2$ phase
Preparation for mitosis
S phase
DNA replication
Histone synthesis
Centrosome formed
Chromosome duplication
0 h
4 h
8 h
12h
G$_1$ phase
RNA and protein synthesis
Cell growth
G$_0$ phase
No cell division
Restriction point

B. Control of the cell cycle

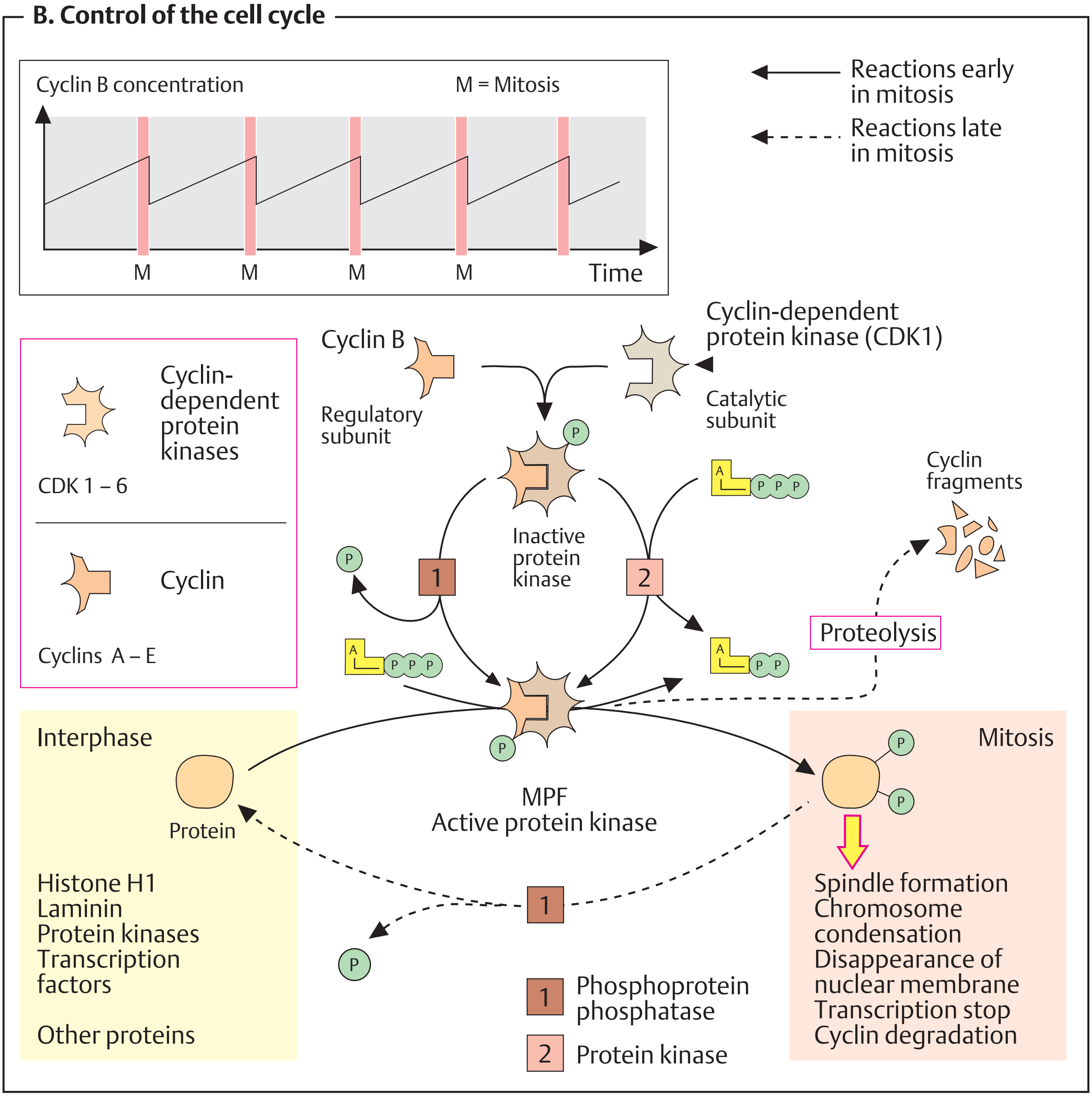
Cyclin B concentration
M = Mitosis
M
M
M
M
Time
Reactions early in mitosis
Reactions late in mitosis
Cyclin-dependent protein kinases
CDK 1 – 6
Cyclin
Cyclins A – E
Cyclin B
Regulatory subunit
Cyclin-dependent protein kinase (CDK1)
Catalytic subunit
P
Inactive protein kinase
A
P P P
P P
1
2
Cyclin fragments
Proteolysis
Interphase
Protein
Histone H1
Laminin
Protein kinases
Transcription factors
Other proteins
MPF
Active protein kinase
Mitosis
Spindle formation
Chromosome condensation
Disappearance of nuclear membrane
Transcription stop
Cyclin degradation
1 Phosphoprotein phosphatase
2 Protein kinase

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Apoptosis

A. Cell proliferation and apoptosis ◑

The number of cells in any tissue is mainly regulated by two processes—cell **proliferation** and *physiological cell death*, **apoptosis**. Both of these processes are regulated by stimulatory and inhibitory factors that act in solute form (growth factors and cytokines) or are presented in bound form on the surface of neighboring cells (see below).

Apoptosis is genetically programmed cell death, which leads to "tidy" breakdown and disposal of cells. Morphologically, apoptosis is characterized by changes in the cell membrane (with the formation of small blebs known as "apoptotic bodies"), shrinking of the nucleus, chromatin condensation, and fragmentation of DNA. *Macrophages* and other phagocytic cells recognize apoptotic cells and remove them by phagocytosis without inflammatory phenomena developing.

Cell necrosis (not shown) should be distinguished from apoptosis. In cell necrosis, cell death is usually due to physical or chemical damage. Necrosis leads to swelling and bursting of the damaged cells and often triggers an inflammatory response.

The growth of tissue (or, more precisely, the number of cells) is actually regulated by apoptosis. In addition, apoptosis allows the elimination of unwanted or superfluous cells—e. g., during embryonic development or in the immune system. The contraction of the uterus after birth is also based on apoptosis. Diseased cells are also eliminated by apoptosis—e. g., tumor cells, virus-infected cells, and cells with irreparably damaged DNA. An everyday example of this is the peeling of the skin after sunburn.

B. Regulation of apoptosis ○

Apoptosis can be triggered by a number of different signals that use various transmission pathways. Other signaling pathways prevent apoptosis.

At the center of the apoptotic process lies a group of specialized *cysteine-containing aspartate proteinases* (see p. 176), known as **caspases**. These mutually activate one another, creating an *enzyme cascade* resembling the cascade involved in blood coagulation (see p. 290). Other enzymes in this group, known as **effector caspases**, cleave cell components after being activated—e. g., laminin in the nuclear membrane and snRP proteins (see p. 246)—or activate special DNases which then fragment the nuclear DNA.

An important trigger for apoptosis is known as the **Fas system**. This is used by cytotoxic T cells, for example, which eliminate infected cells in this way (top left). Most of the body's cells have *Fas receptors* (CD 95) on their plasma membrane. If a T cell is activated by contact with an MHC presenting a viral peptide (see p. 296), binding of its *Fas ligands* occurs on the target cell's Fas receptors. Via the mediator protein FADD ("Fas-associated death domain"), this activates *caspase-8* inside the cell, setting in motion the apoptotic process.

Another trigger is provided by **tumor necrosis factor-α** (TNF-α), which acts via a similar protein (TRADD) and supports the endogenous defense system against tumors by inducing apoptosis.

Caspase-8 activates the effector caspases either directly, or indirectly by promoting the **cytochrome c** (see p. 140) from mitochondria. Once in the cytoplasm, cytochrome c binds to and activates the protein Apaf-1 (not shown) and thus triggers the caspase cascade. Apoptotic signals can also come from the cell nucleus. If irreparable DNA damage is present, the **p53 protein** (see p. 394)—the product of a *tumor suppressor gene*—promotes apoptosis and thus helps eliminate the defective cell.

There are also inhibitory factors that oppose the signals that activate apoptosis. These include **bcl-2** and related proteins. The genomes of several viruses include genes for this type of protein. The genes are expressed by the host cell and (to the benefit of the virus) prevent the host cell from being prematurely eliminated by apoptosis.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Cell proliferation and apoptosis

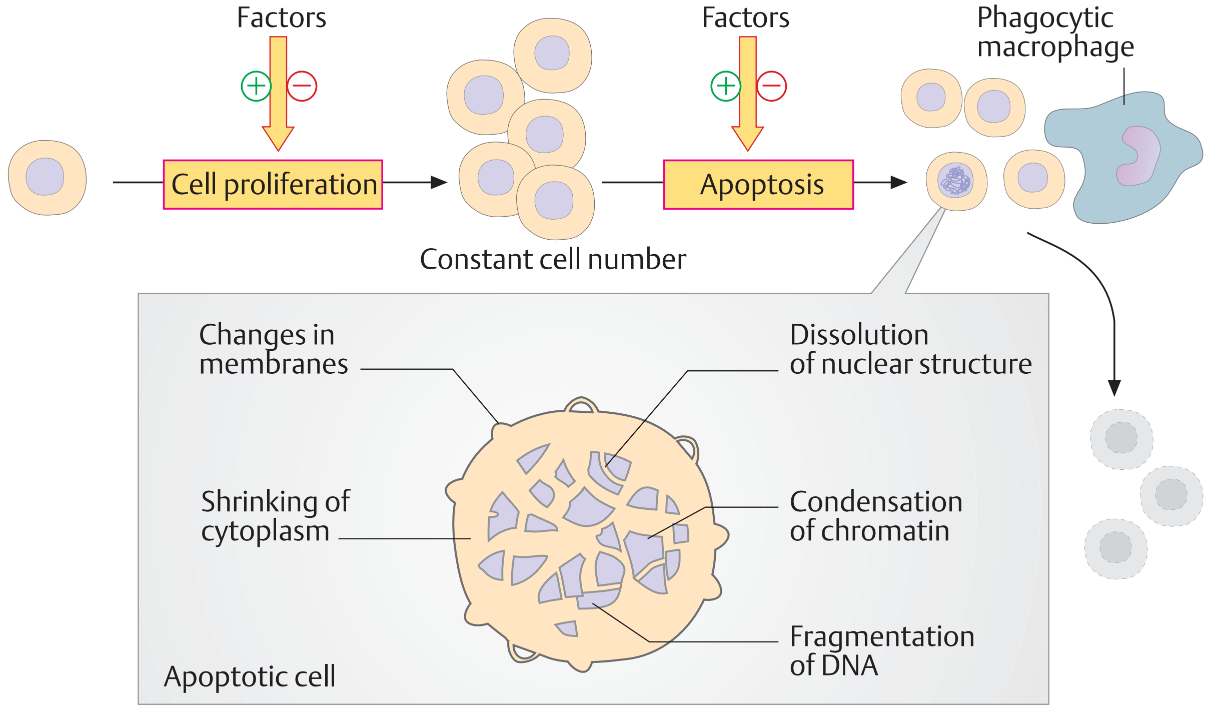
Factors
Cell proliferation
Constant cell number
Factors
Apoptosis
Phagocytic macrophage
Changes in membranes
Dissolution of nuclear structure
Shrinking of cytoplasm
Condensation of chromatin
Fragmentation of DNA
Apoptotic cell

B. Regulation of apoptosis

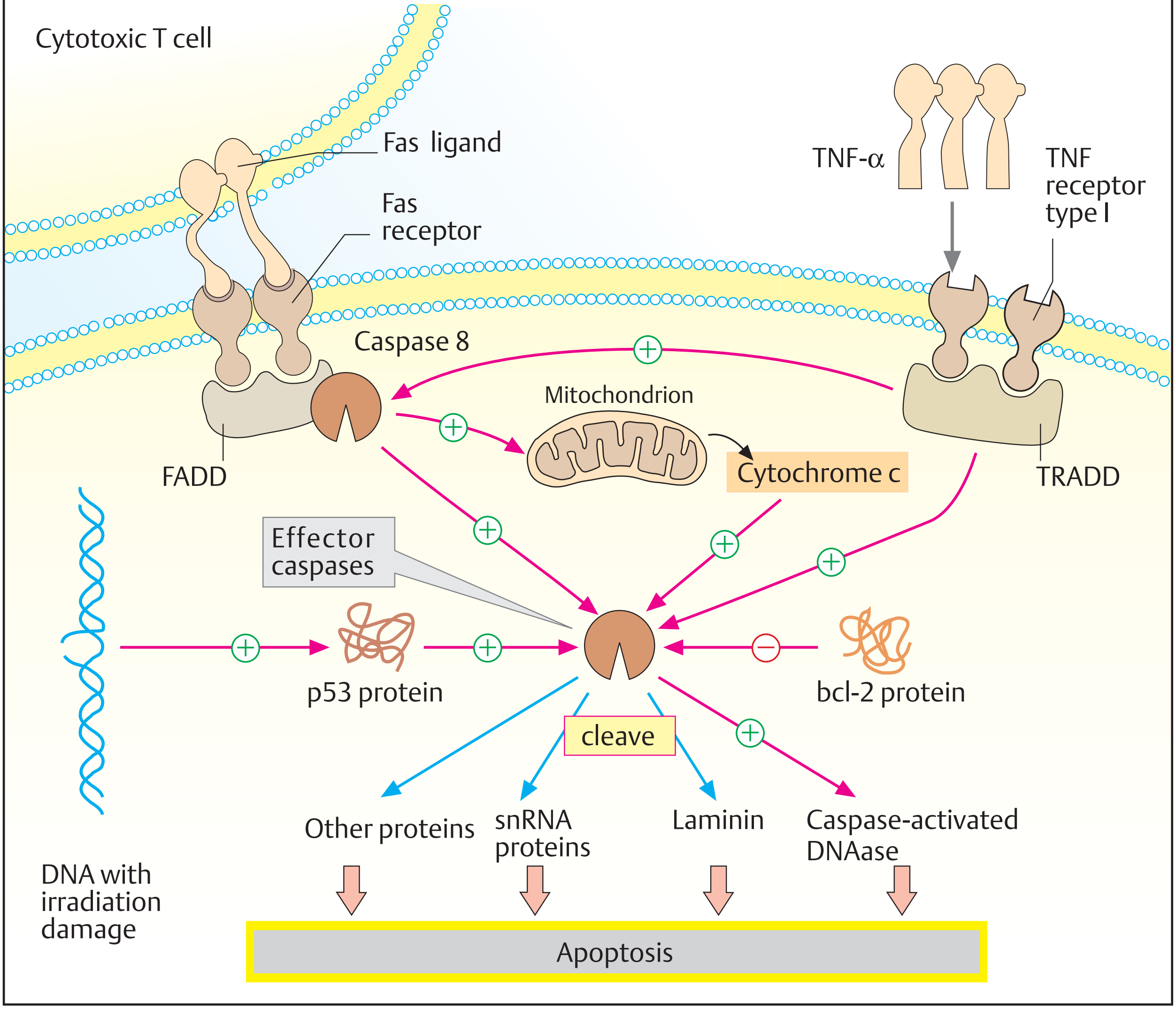
Cytotoxic T cell
Fas ligand
Fas receptor
TNF-α
TNF receptor type I
Caspase 8
Mitochondrion
Cytochrome c
FADD
TRADD
Effector caspases
p53 protein
bcl-2 protein
cleave
Other proteins
snRNA proteins
Laminin
Caspase-activated DNAase
DNA with irradiation damage
Apoptosis

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Oncogenes

Oncogenes are cellular genes that can trigger uncontrolled cell proliferation if their sequence is altered or their expression is incorrectly regulated. They were first discovered as *viral (v-) oncogenes* in retroviruses that cause tumors (tumor viruses). Viruses of this type (see p. 404) sometimes incorporate genes from the host cell into their own genome. If these genes are reincorporated into the host DNA again during later infection, tumors can then be caused in rare cases. Although virus-related tumors are rare, research into them has made a decisive contribution to our understanding of oncogenes and their functioning.

A. Proto-oncogenes: biological role ◑

The cellular form of oncogenes (known as c-oncogenes or **proto-oncogenes**) code for proteins involved in controlling growth and differentiation processes. They only become oncogenes if their sequence has been altered by *mutations* (see p. 256), *deletions,* and other processes, or when excessive amounts of the gene products have been produced as a result of *overexpression.*

Overexpression can occur when **amplification** leads to numerous functional copies of the respective gene, or when the gene falls under the influence of a highly active promoter (see p. 244). If the control of oncogene expression by **tumor suppressor genes** (see p. 394) is also disturbed, **transformation** and unregulated proliferation of the cells can occur. A single activated oncogene does not usually lead to a loss of growth control. It only occurs when over the course of time mutations and regulation defects accumulate in one and the same cell. If the immune system does not succeed in eliminating the transformed cell, it can over the course of months or years grow into a macroscopically visible **tumor**.

B. Oncogene products: biochemical functions ◑

A feature common to all oncogenes is the fact that they code for proteins involved in *signal transduction processes.* The genes are designated using three-letter abbreviations that usually indicate the origin of the viral gene and are printed in italics (e. g., *myc* for myelocytomatosis, a viral disease in birds). Oncogene products can be classified into the following groups according to their functions.

1. **Ligands** such as *growth factors* and *cytokines,* which promote cell proliferation.
2. **Membrane receptors** of the 1-helix type with tyrosine kinase activity, which can bind growth factors and hormones (see p. 394).
3. **GTP-binding proteins.** This group includes the G proteins in the strict sense and related proteins such as Ras (see p. 388), the product of the oncogene c-*ras.*
4. **Receptors for lipophilic hormones** mediate the effects of steroid hormones and related signaling substances. They regulate the transcription of specific genes (see p. 378). The products of several oncogenes (e. g., *erbA*) belong to this superfamily of *ligand-controlled transcription factors.*
5. **Nuclear tumor suppressors** inhibit return to the cell cycle in fully differentiated cells. The genes that code for these proteins are referred to as *anti-oncogenes* due to this function. On the role of p53 and pRb, see p. 394.
6. **DNA-binding proteins.** A whole series of oncogenes code for *transcription factors.* Particularly important for cell proliferation are *myc,* as well as *fos* and *jun.* The protein products of the latter two genes form the transcription factor AP-1 as a heterodimer (see p. 244).
7. **Protein kinases** play a central role in intracellular signal transduction. By phosphorylating proteins, they bring about alterations in biological activity that can only be reversed again by the effects of *protein phosphatases.* The interplay between protein phosphorylation by protein kinases and dephosphorylation by protein phosphatases (*interconversion*) serves to regulate the cell cycle (see p. 394) and other important processes. The protein kinase Raf is also involved in the signal transduction of insulin (see p. 388).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

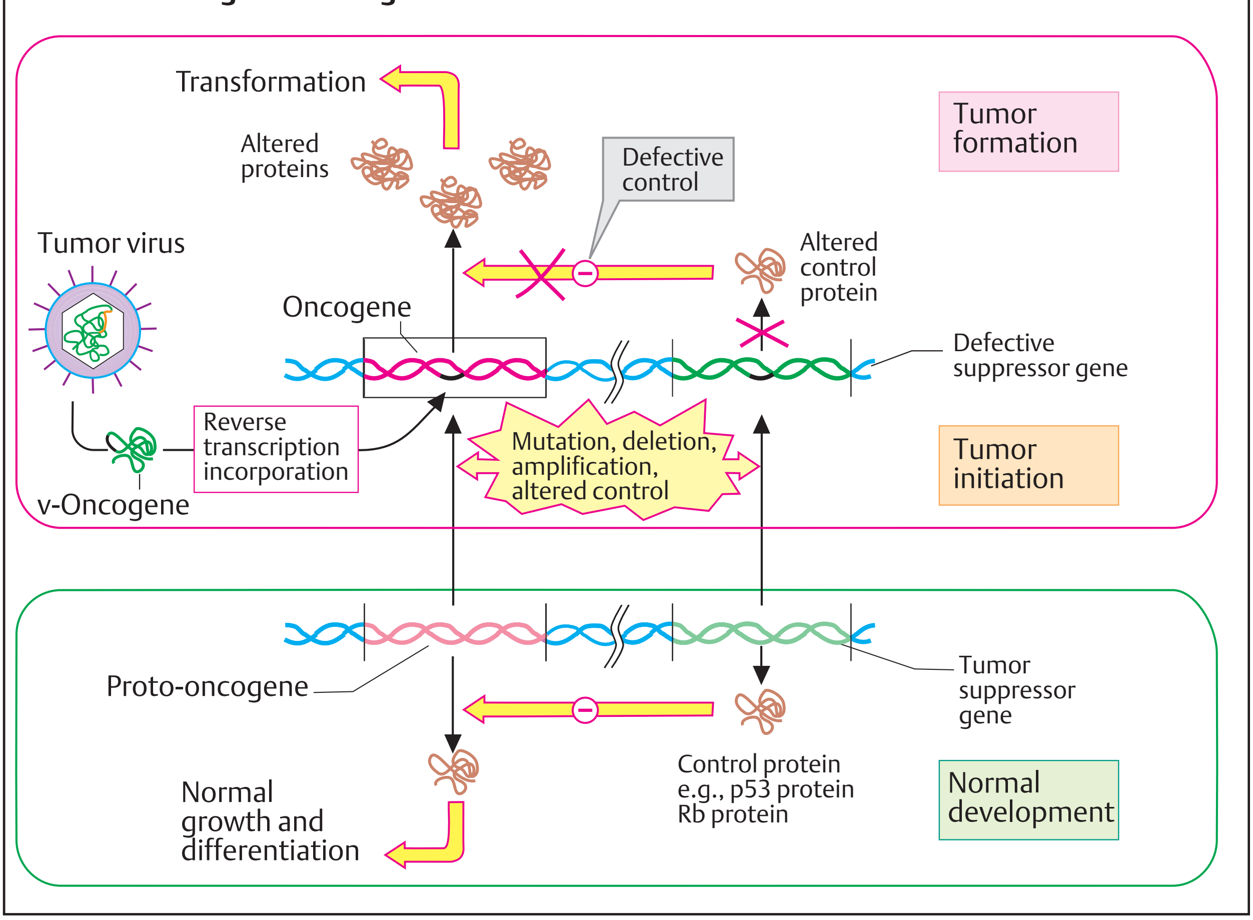

A. Proto-oncogenes: biological role

Transformation
Altered proteins
Defective control
Tumor formation
Tumor virus
Oncogene
Altered control protein
Defective suppressor gene
Reverse transcription incorporation
Mutation, deletion, amplification, altered control
Tumor initiation
v-Oncogene
Proto-oncogene
Tumor suppressor gene
Normal growth and differentiation
Control protein e.g., p53 protein Rb protein
Normal development

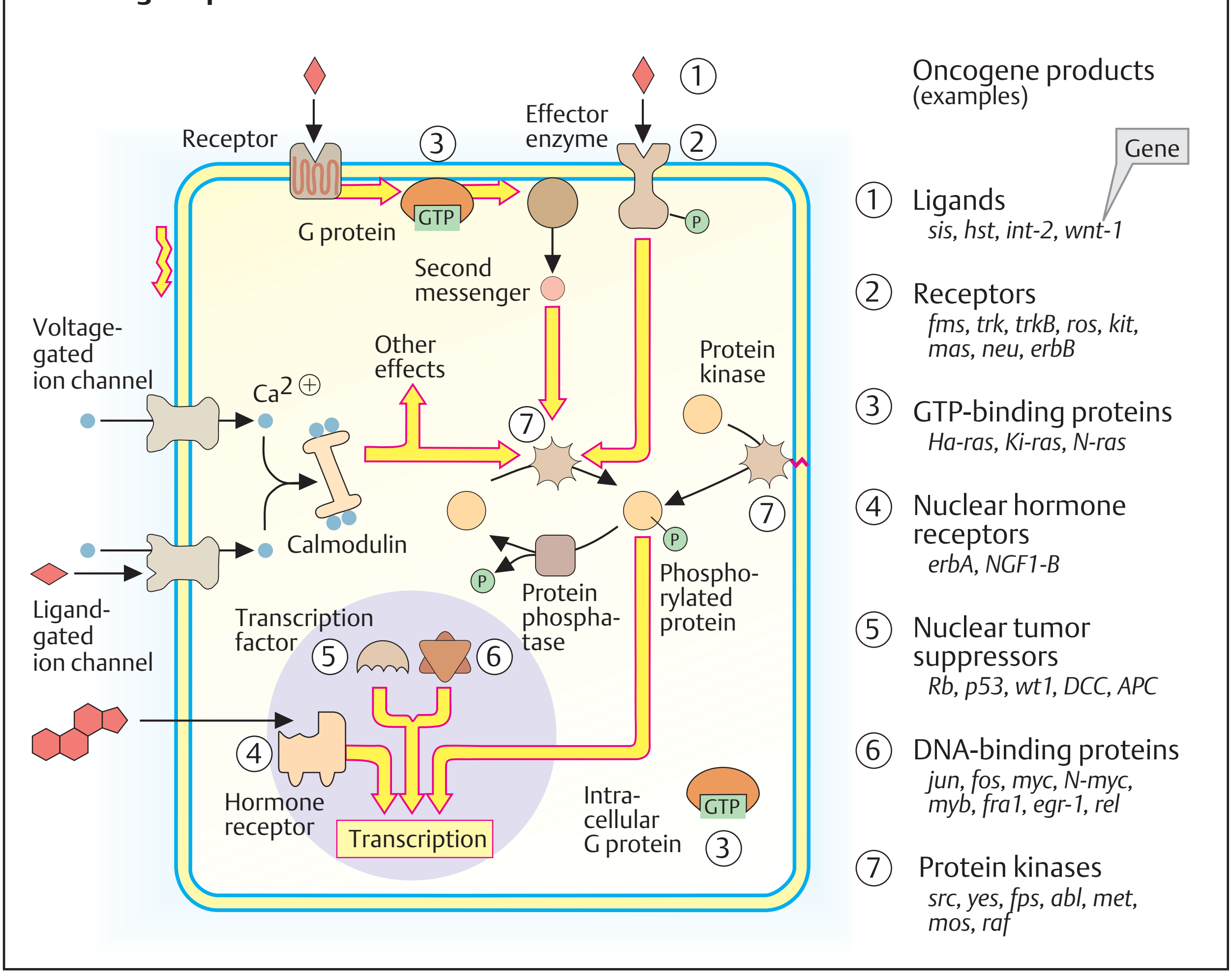

B. Oncogene products: biochemical functions

Receptor
③
Effector enzyme
①
②
G protein
GTP
P
Second messenger
Voltage-gated ion channel
Ca2⊕
Other effects
Protein kinase
⑦
Calmodulin
⑦
P
Ligand-gated ion channel
P
Protein phosphatase
Phosphorylated protein
Transcription factor
⑤
⑥
④
Hormone receptor
Transcription
Intracellular G protein
GTP
③
Oncogene products (examples)
Gene
① Ligands
sis, hst, int-2, wnt-1
② Receptors
fms, trk, trkB, ros, kit, mas, neu, erbB
③ GTP-binding proteins
Ha-ras, Ki-ras, N-ras
④ Nuclear hormone receptors
erbA, NGF1-B
⑤ Nuclear tumor suppressors
Rb, p53, wt1, DCC, APC
⑥ DNA-binding proteins
jun, fos, myc, N-myc, myb, fra1, egr-1, rel
⑦ Protein kinases
src, yes, fps, abl, met, mos, raf

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Tumors

A. Division behavior of cells ◑

The body's cells are normally subject to strict "social" control. They only divide until they come into contact with neighboring cells; cell division then ceases due to *contact inhibition.* Exceptions to this rule include embryonic cells, cells of the intestinal epithelium (where the cells are constantly being replaced), cells in the bone marrow (where formation of blood cells takes place), and **tumor cells.** *Uncontrolled cell proliferation* is an important indicator of the presence of a tumor. While normal cells in cell culture only divide 20–60 times, tumor cells are potentially immortal and are not subject to contact inhibition.

In medicine, a distinction is made between *benign* and *malignant tumors.* Benign tumors consist of slowly growing, largely differentiated cells. By contrast, malignant tumors show rapid, invasive growth and tend to form *metastases* (dissemination of daughter lesions). The approximately 100 different types of tumor that exist are responsible for more than 20% of deaths in Europe and North America.

B. Transformation ◑

The transition of a normal cell into a tumor cell is referred to as **transformation.**

Normal cells have all the characteristics of fully differentiated cells specialized for a particular function. Their division is inhibited and they are usually in the G_0 phase of the cell cycle (see p. 394). Their external shape is variable and is determined by a strongly structured cytoskeleton.

In contrast, *tumor cells* divide without inhibition and are often de-differentiated—i. e., they have acquired some of the properties of embryonic cells. The surface of these cells is altered, and this is particularly evident in a disturbance of contact inhibition by neighboring cells. The cytoskeleton of tumor cells is also restructured and often reduced, giving them a rounded shape. The nuclei of tumor cells can be atypical in terms of shape, number, and size.

Tumor markers are clinically important for detecting certain tumors. These are proteins that are formed with increasing frequency by tumor cells (group 1) or are induced by them in other cells (group 2). Group 1 tumor markers include *tumor-associated antigens,* secreted hormones, and enzymes. The table lists a few examples.

The transition from a normal to a transformed state is a process involving several steps.

1. Tumor initiation. Almost every tumor begins with damage to the DNA of an individual cell. The genetic defect is almost always caused by environmental factors. These can include tumor-inducing chemicals (*carcinogens*—e. g., components of tar from tobacco), physical processes (e. g., UV light, X-ray radiation; see p. 256), or in rare cases tumor viruses (see p. 398). Most of the approximately 10^{14} cells in the human body probably suffer this type of DNA damage during the average lifespan, but it is usually repaired again (see p. 256). It is mainly defects in **proto-oncogenes** (see p. 398) that are relevant to tumor initiation; these are the decisive cause of *transformation*. Loss of an **anti-oncogene** (a tumor-suppressor gene) can also contribute to tumor initiation.

2. Tumor promotion is preferential proliferation of a cell damaged by transformation. It is a very slow process that can take many years. Certain substances are able to strongly accelerate it—e. g., *phorbol esters.* These occur in plants (e. g., *Euphorbia* species) and act as activators of protein kinase C (see p. 386).

3. Tumor progression finally leads to a macroscopically visible tumor as a result of growth. When solid tumors of this type exceed a certain size, they form their own vascular network that supplies them with blood (angiogenesis). *Collagenases* (matrix metalloproteinases, MMPs) play a special role in the metastatic process, by loosening surrounding connective tissue and thereby allowing tumor cells to disseminate and enter the bloodstream. New approaches to combating tumors have been aimed at influencing tumor angiogenesis and metastatic processes.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

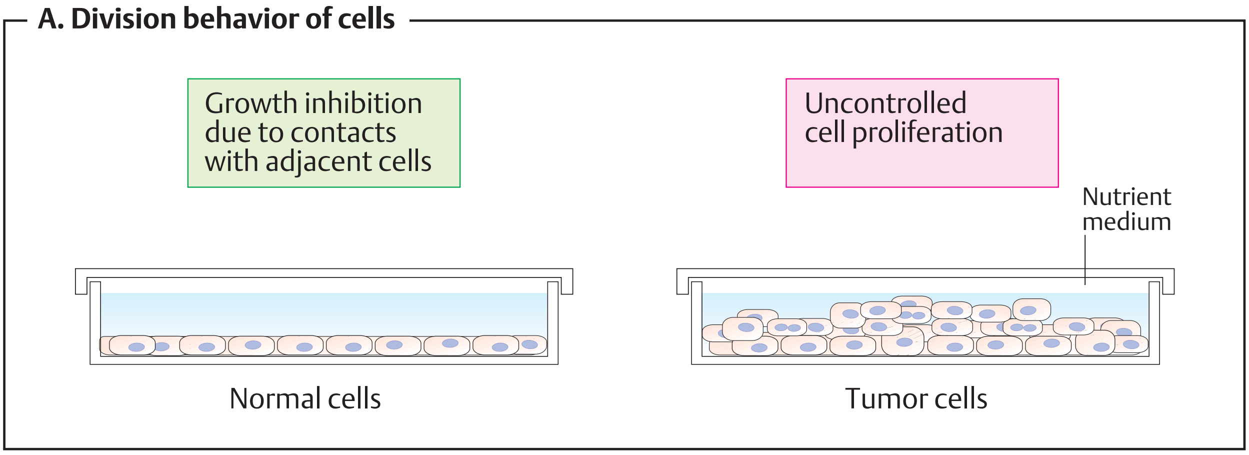
A. Division behavior of cells
Growth inhibition due to contacts with adjacent cells
Uncontrolled cell proliferation
Nutrient medium
Normal cells
Tumor cells

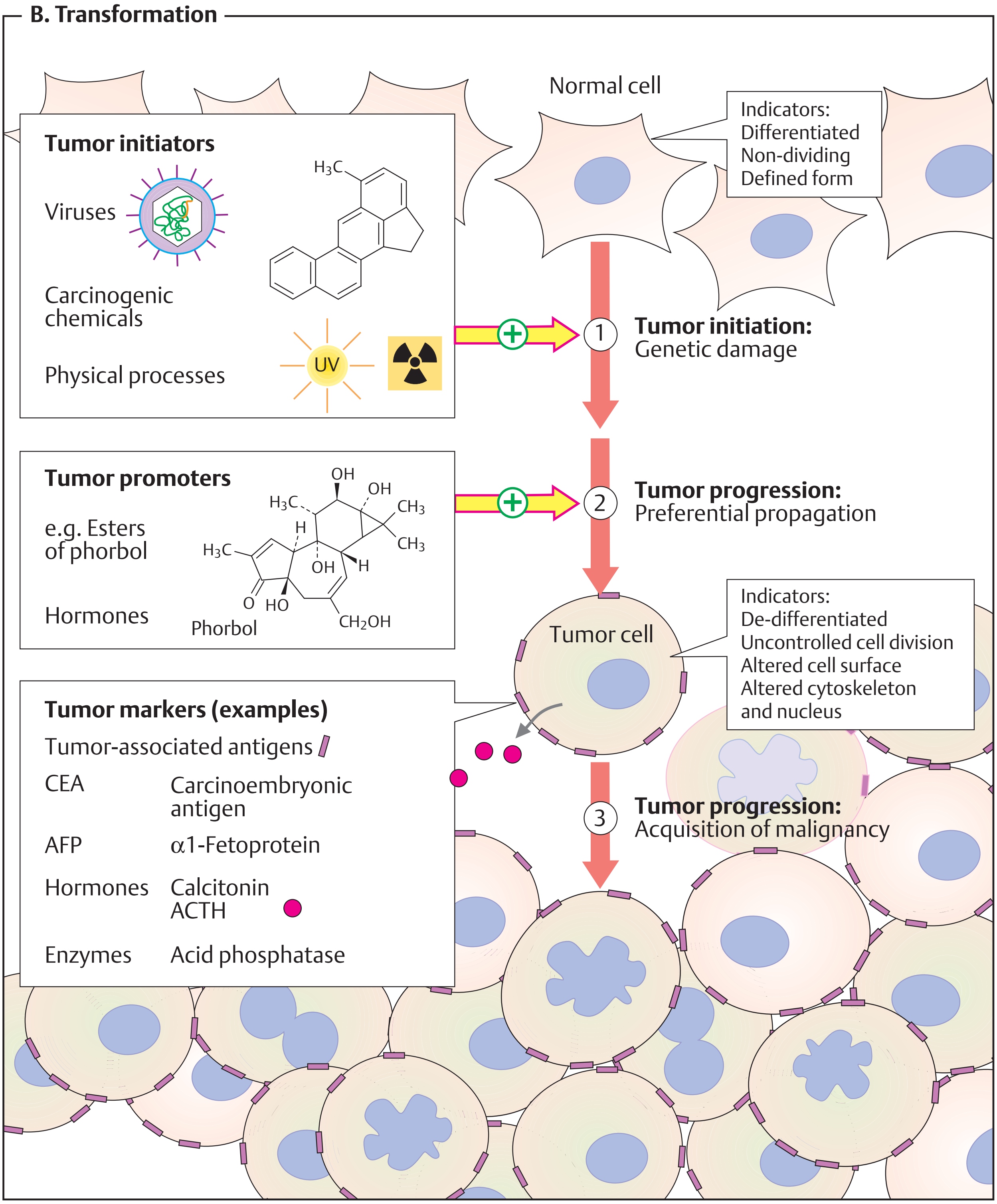
B. Transformation
Normal cell
Indicators:
Differentiated
Non-dividing
Defined form
Tumor initiators
Viruses
Carcinogenic chemicals
H_3C
Physical processes
UV
1
Tumor initiation:
Genetic damage
Tumor promoters
e.g. Esters of phorbol
Hormones
OH
OH
H_3C
CH_3
H
CH_3
H_3C
OH
H
O
HO
CH_2OH
Phorbol
2
Tumor progression:
Preferential propagation
Tumor cell
Indicators:
De-differentiated
Uncontrolled cell division
Altered cell surface
Altered cytoskeleton and nucleus
Tumor markers (examples)
Tumor-associated antigens
CEA Carcinoembryonic antigen
AFP α1-Fetoprotein
Hormones Calcitonin ACTH
Enzymes Acid phosphatase
3
Tumor progression:
Acquisition of malignancy

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Cytostatic drugs

Tumors (see p. 400) arise from degenerated (transformed) cells that grow in an uncontrolled way as a result of genetic defects. Most transformed cells are recognized by the immune system and eliminated (see p. 294). If endogenous defense is not sufficiently effective, rapid tumor growth can occur. Attempts are then made to inhibit growth by physical or chemical treatment.

A frequently used procedure is targeted irradiation with γ-rays, which block cell reproduction due their mutagenic effect (see p. 256). Another approach is to inhibit cell growth by chemotherapy. The growth-inhibiting substances used are known as **cytostatic drugs**. Unfortunately, neither radiotherapy nor chemotherapy act selectively—i. e., they damage normal cells as well, and are therefore often associated with severe side effects.

Most cytostatic agents directly or indirectly inhibit DNA replication in the S phase of the cell cycle (see p. 394). The first group (**A**) lead to chemical changes in cellular DNA that impede transcription and replication. A second group of cytostatic agents (**B**) inhibit the synthesis of DNA precursors.

A. Alkylating agents, anthracyclines ○

Alkylating agents are compounds capable of reacting covalently with DNA bases. If a compound of this type contains *two* reactive groups, intramolecular or intermolecular *crosslinking* of the DNA double helix and "bending" of the double strand occurs. Examples of this type shown here are **cyclophosphamide** and the inorganic complex **cisplatin**. Anthracyclines such as **doxorubicin** (adriamycin) insert themselves non-covalently between the bases and thus lead to local alterations in the DNA structure (see p. 254 B).

B. Antimetabolites ○

Antimetabolites are enzyme inhibitors (see p. 96) that selectively block metabolic pathways. The majority of clinically important cytostatic drugs act on *nucleotide biosynthesis*. Many of these are modified nucleobases or nucleotides that *competitively* inhibit their target enzymes (see p. 96). Many are also incorporated into the DNA, thereby preventing replication.

The cytostatic drugs administered (indicated by a syringe in the illustration) are often not active themselves but are only converted into the actual active agent in the metabolism. This also applies to the adenine analogue **6-mercaptopurine**, which is initially converted to the mononucleotide tIMP (thioinosine monophosphate). Via several intermediate steps, tIMP gives rise to tdGTP, which is incorporated into the DNA and leads to crosslinks and other anomalies in it. The second effective metabolite of 6-mercaptopurine is *S*-methylated tIMP, an inhibitor of *amidophosphoribosyl transferase* (see p. 188).

Hydroxyurea selectively inhibits *ribonucleotide reductase* (see p. 190). As a radical scavenger, it removes the tyrosine radicals that are indispensable for the functioning of the reductase.

Two other important cytostatic agents target the synthesis of DNA-typical thymine, which takes place at the level of the deoxymononucleotide (see p. 190). The deoxymononucleotide formed by **5-fluorouracil** or the corresponding nucleoside inhibits *thymidylate synthase.* This inhibition is based on the fact that the fluorine atom in the pyrimidine ring cannot be substituted by a methyl group. In addition, the fluorine analogue is also incorporated into the DNA.

Dihydrofolate reductase acts as an auxiliary enzyme for thymidylate synthase. It is involved in the regeneration of the coenzyme N^5,N^{10}-methylene-THF, initially reducing DHF to THF with NADPH as the reductant (see p. 418). The folic acid analogue **methotrexate**, a frequently used cytostatic agent, is an extremely effective competitive inhibitor of dihydrofolate reductase. It leads to the depletion of N^5,N^{10}-methylene-THF in the cells and thus to cessation of DNA synthesis.

Further information

To reduce the side effects of cytostatic agents, new approaches are currently being developed on the basis of **gene therapy** (see p. 264). Attempts are being made, for example, to administer drugs in the form of precursors (known as prodrugs), which only become active in the tumor itself ("tumor targeting").

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Alkylating agents, anthracyclines

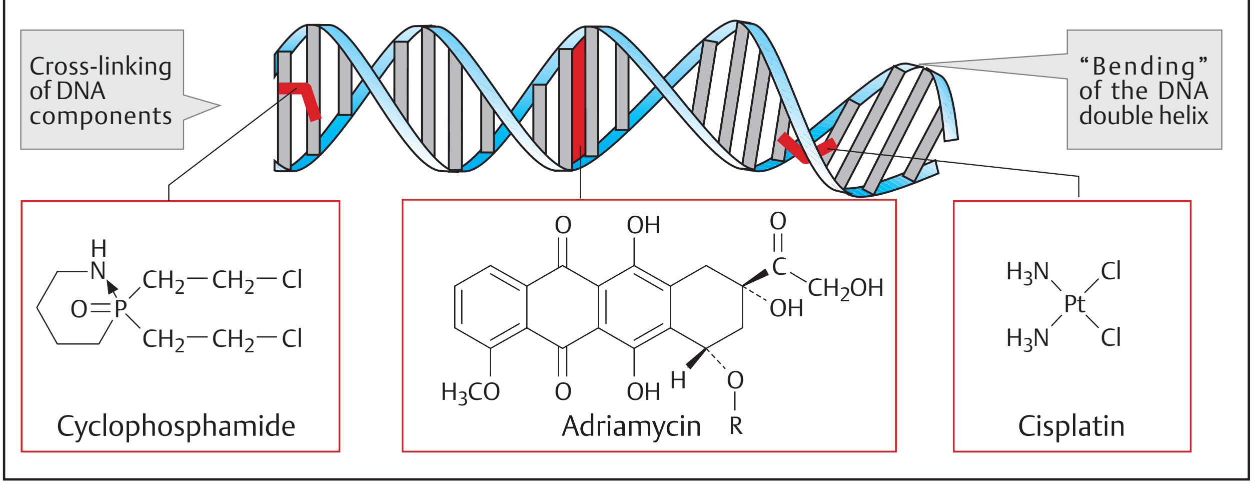
Cross-linking of DNA components
"Bending" of the DNA double helix
Cyclophosphamide
Adriamycin
Cisplatin

B. Antimetabolites

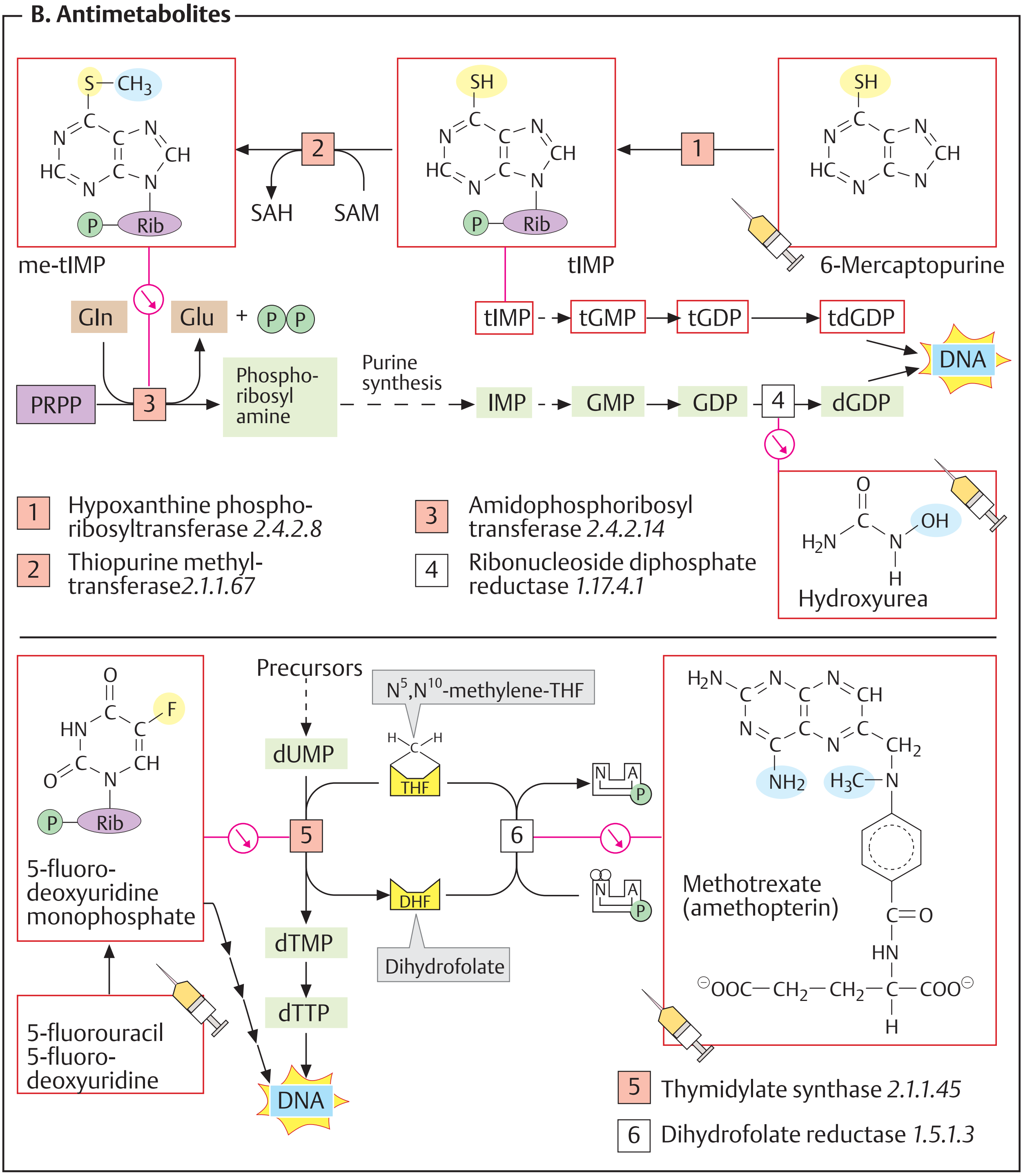
me-tIMP
tIMP
6-Mercaptopurine
SAH
SAM
Gln
Glu
PRPP
Phospho-ribosyl amine
Purine synthesis
tIMP
tGMP
tGDP
tdGDP
IMP
GMP
GDP
dGDP
DNA
Hydroxyurea
1 Hypoxanthine phosphoribosyltransferase 2.4.2.8
2 Thiopurine methyltransferase2.1.1.67
3 Amidophosphoribosyl transferase 2.4.2.14
4 Ribonucleoside diphosphate reductase 1.17.4.1
Precursors
N5,N10-methylene-THF
dUMP
THF
DHF
Dihydrofolate
dTMP
dTTP
DNA
5-fluoro-deoxyuridine monophosphate
5-fluorouracil 5-fluoro-deoxyuridine
Methotrexate (amethopterin)
5 Thymidylate synthase 2.1.1.45
6 Dihydrofolate reductase 1.5.1.3

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Viruses

Viruses are *parasitic nucleoprotein complexes.* They often consist of only a single nucleic acid molecule (DNA or RNA, never both) and a protein coat. Viruses have *no metabolism of their own,* and can therefore only replicate themselves with the help of host cells. They are therefore not regarded as independent organisms. Viruses that damage the host cell when they replicate are *pathogens.* Diseases caused by viruses include AIDS, rabies, poliomyelitis, measles, German measles, smallpox, influenza, and the common cold.

A. Viruses: examples ◑

Only a few examples from the large number of known viruses are illustrated here. They are all shown on the same scale.

Viruses that only replicate in bacteria are known as **bacteriophages** (or "phages" for short). An example of a phage with a simple structure is **M13.** It consists of a single-stranded DNA molecule (ssDNA) of about 7000 bp with a coat made up of 2700 helically arranged protein subunits. The coat of a virus is referred to as a *capsid,* and the complete structure as a *nucleocapsid.* In genetic engineering, M13 is important as a *vector* for foreign DNA (see p. 258).

The phage **T4** (bottom left), one of the largest viruses known, has a much more complex structure with around 170 000 base pairs (bp) of double-stranded DNA (dsDNA) contained within its "head."

The **tobacco mosaic virus** (center right), a plant pathogen, has a structure similar to that of M13, but contains ssRNA instead of DNA. The **poliovirus**, which causes poliomyelitis, is also an RNA virus. In the **influenza virus**, the pathogen that causes viral flu, the nucleocapsid is additionally surrounded by a *coat* derived from the plasma membrane of the host cell (**C**). The coat carries viral proteins that are involved in the infection process.

B. Capsid of the rhinovirus ○

Rhinoviruses cause the common cold. In these viruses, the capsid is shaped like an *icosahedron*—i. e., an object made up of 20 equilateral triangles. Its surface is formed from three different proteins, which associate with one another to form pentamers and hexamers. In all, 60 protein molecules are involved in the structure of the capsid.

C. Life cycle of HIV ◑

The *human immunodeficiency virus (HIV)* causes the immunodeficiency disease known as **AIDS** (*acquired immune deficiency syndrome*). The structure of this virus is similar to that of the influenza virus (**A**).

The HIV genome consists of two molecules of ssRNA (each 9.2 kb). It is enclosed by a double-layered capsid and a protein-containing coating membrane. HIV mainly infects T helper cells (see p. 294) and can thereby lead to failure of the immune system in the longer term.

During infection (**1**), the virus's coating membrane fuses with the target cell's plasma membrane, and the core of the nucleocapsid enters the cytoplasm (**2**). In the cytoplasm, the viral RNA is initially transcribed into an RNA/DNA hybrid (**3**) and then into dsDNA (**4**). Both of these reactions are catalyzed by *reverse transcriptase,* an enzyme deriving from the virus. The dsDNA formed is integrated into the host cell genome (**5**), where it can remain in an inactive state for a long time.

When viral replication occurs, the DNA segment corresponding to the viral genome is first transcribed by host cell enzymes (**6**). This gives rise not only to viral ssRNA, but also to transcription of mRNAs for precursors of the viral proteins (**7**). These precursors are integrated into the plasma membrane (**8**, **9**) before undergoing proteolytic modification (**10**). The cycle is completed by the release of new virus particles (**11**).

The group of RNA viruses to which HIV belongs are called **retroviruses**, because DNA is produced from RNA in their replication cycle—the reverse of the usual direction of transcription (DNA → RNA).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Viruses: examples

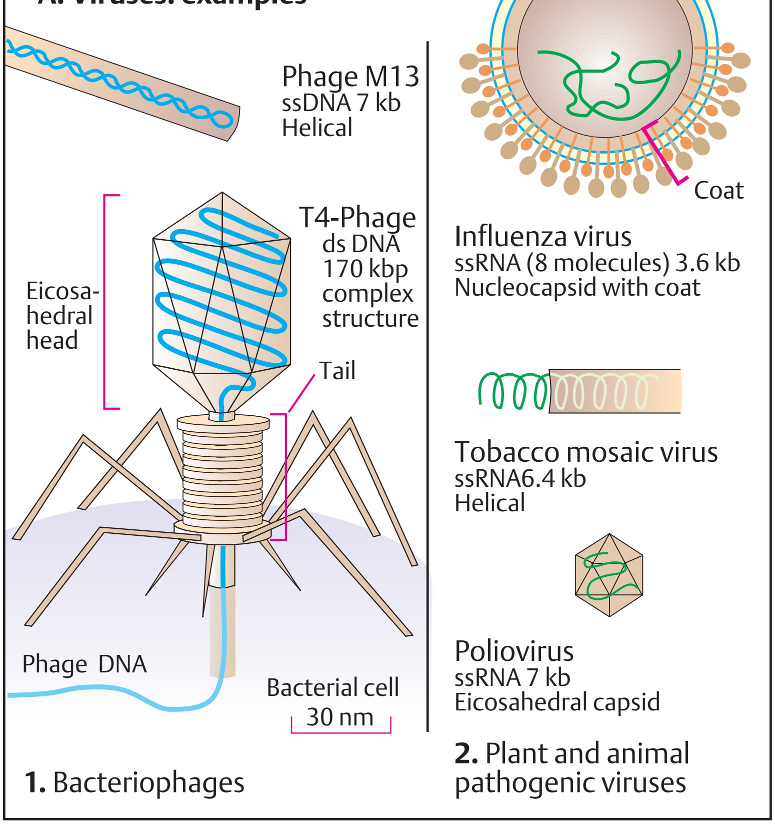
Phage M13
ssDNA 7 kb
Helical
T4-Phage
ds DNA
170 kbp
complex
structure
Eicosa-
hedral
head
Tail
Phage DNA
Bacterial cell
30 nm
1. Bacteriophages
Coat
Influenza virus
ssRNA (8 molecules) 3.6 kb
Nucleocapsid with coat
Tobacco mosaic virus
ssRNA6.4 kb
Helical
Poliovirus
ssRNA 7 kb
Eicosahedral capsid
2. Plant and animal
pathogenic viruses

B. Capsid of the rhinovirus

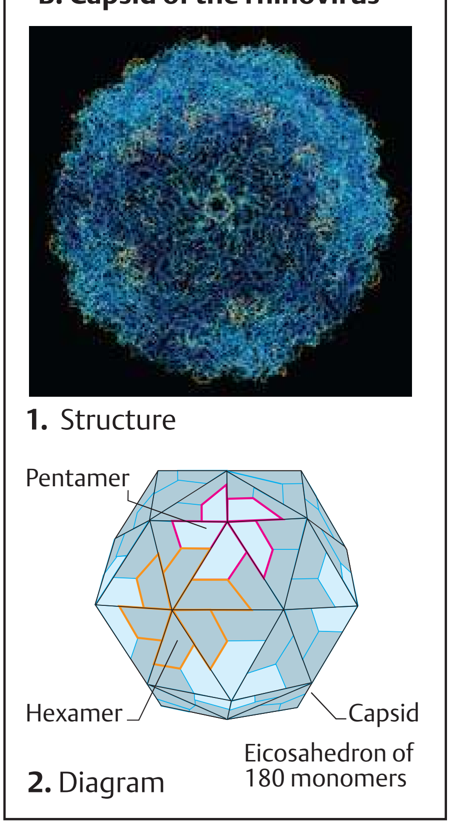
1. Structure
Pentamer
Hexamer
Capsid
Eicosahedron of
180 monomers
2. Diagram

C. Life cycle of the human immunodeficiency virus (HIV)

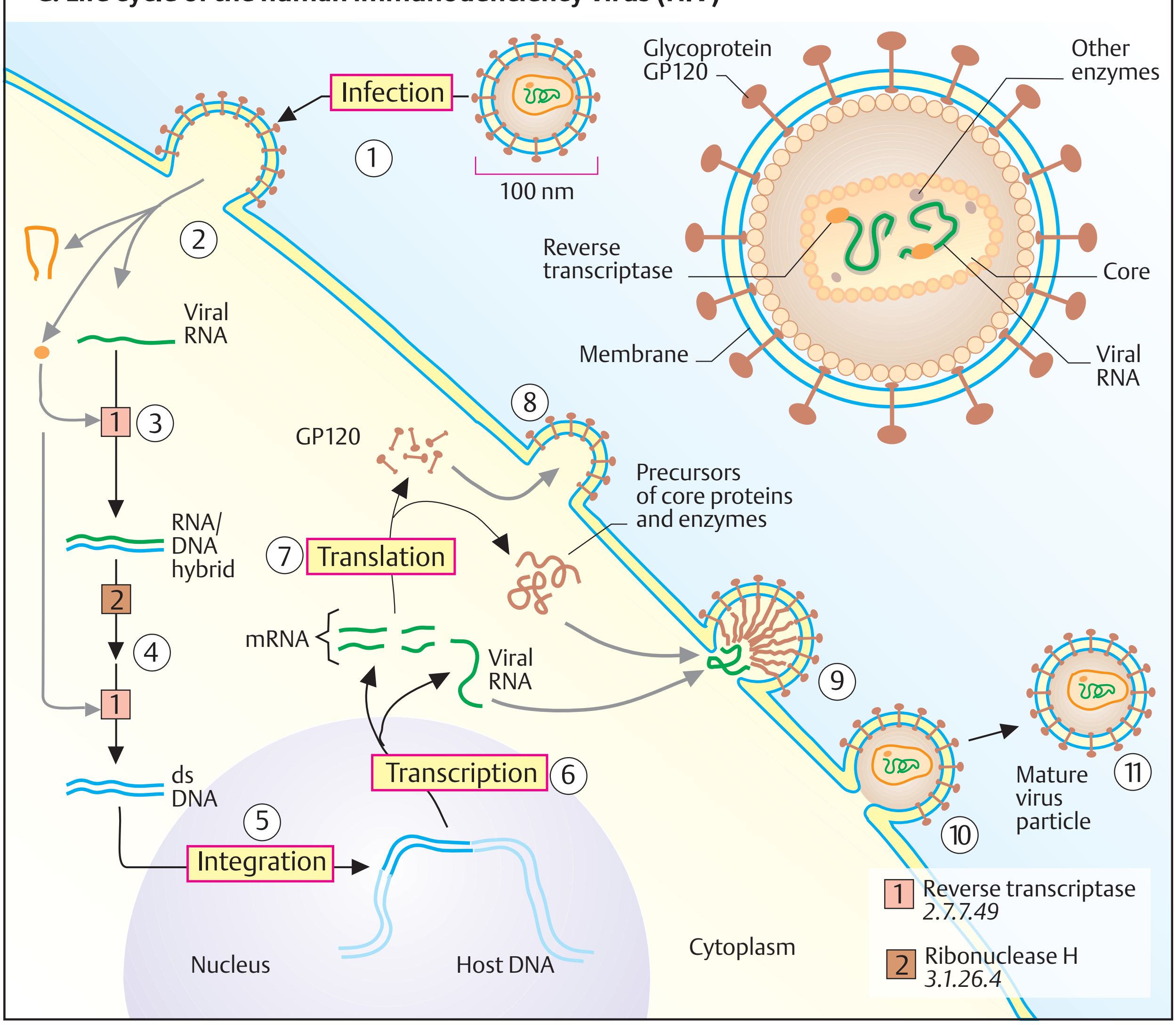
Infection
①
100 nm
Glycoprotein
GP120
Other
enzymes
Reverse
transcriptase
Core
Membrane
Viral
RNA
②
Viral
RNA
1 ③
RNA/
DNA
hybrid
2
④
1
ds
DNA
⑤
Integration
Nucleus
Host DNA
Transcription
⑥
mRNA
Viral
RNA
⑦
Translation
GP120
⑧
Precursors
of core proteins
and enzymes
⑨
⑩
Mature
virus
particle
⑪
Cytoplasm
1 Reverse transcriptase
2.7.7.49
2 Ribonuclease H
3.1.26.4

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Metabolic charts

Explanations

The following 13 plates (pp. 407–419) provide a concise schematic overview of the most important metabolic pathways. Explanatory text is deliberately omitted from them.

These "charts":
- Contain details of metabolic pathways that are only shown in outline in the main text for reasons of space. This applies in particular to the synthesis and degradation of the amino acids and nucleotides, and for some aspects of carbohydrate and lipid metabolism.
- Offer a quick overview of a specific pathway, the metabolites that arise in it, and the enzymes involved.
- Can be used for reference purposes and for revising material previously learned.

The most important **intermediates** are shown with numbers in the charts. The corresponding compounds can be identified using the table on the same page.

In addition, at each step the four-figure EC number (see p. 88) for the **enzyme** responsible for a reaction is given in *italics*. The enzyme name and its systematic classification in the system used by the *Enzyme Catalogue* are available in the following **annotated enzyme list** (pp. 420–430), in which all of the enzymes mentioned in this book are listed according to their EC number. The book's index is helpful when looking for a specific enzyme in the text.

In reactions that involve **coenzymes**, the coenzyme names are also given (sometimes in simplified form). Particularly important starting, intermediate, or end products are given with the full name, or as formulae.

Example

On p. 407, the initial step of the dark reactions in plant photosynthesis (in the Calvin cycle) is shown at the top left.

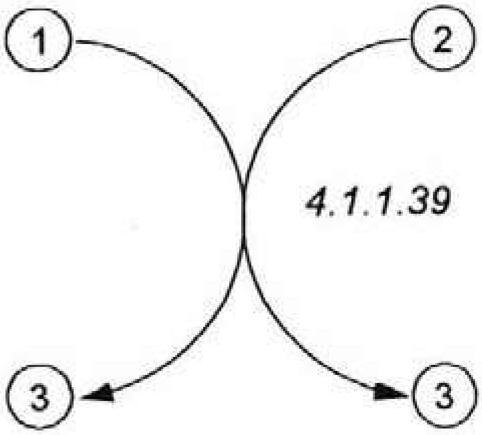

In this reaction, one molecule of ribulose-1,5-bisphosphate (metabolite 1) and one molecule of CO_2 (metabolite 2) give rise to two molecules of 3-phosphoglycerate (metabolite 3). The enzyme responsible has the EC number *4.1.1.39.* The annotated enzyme list shows that this refers to *ribulose bisphosphate carboxylase* ("rubisco" for short). Rubisco belongs to enzyme class 4 (the lyases) and, within that group, to subclass 4.1 (the carboxy-lyases). It contains copper as a cofactor ([Cu]).

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Calvin cycle (plant chloroplasts)

$$6\,CO_2 + 18\,ATP + 12\,NADPH + 12\,H^{\oplus} \longrightarrow Hexose + 18\,ADP + 18\,P + 12\,NADP^{\oplus}$$

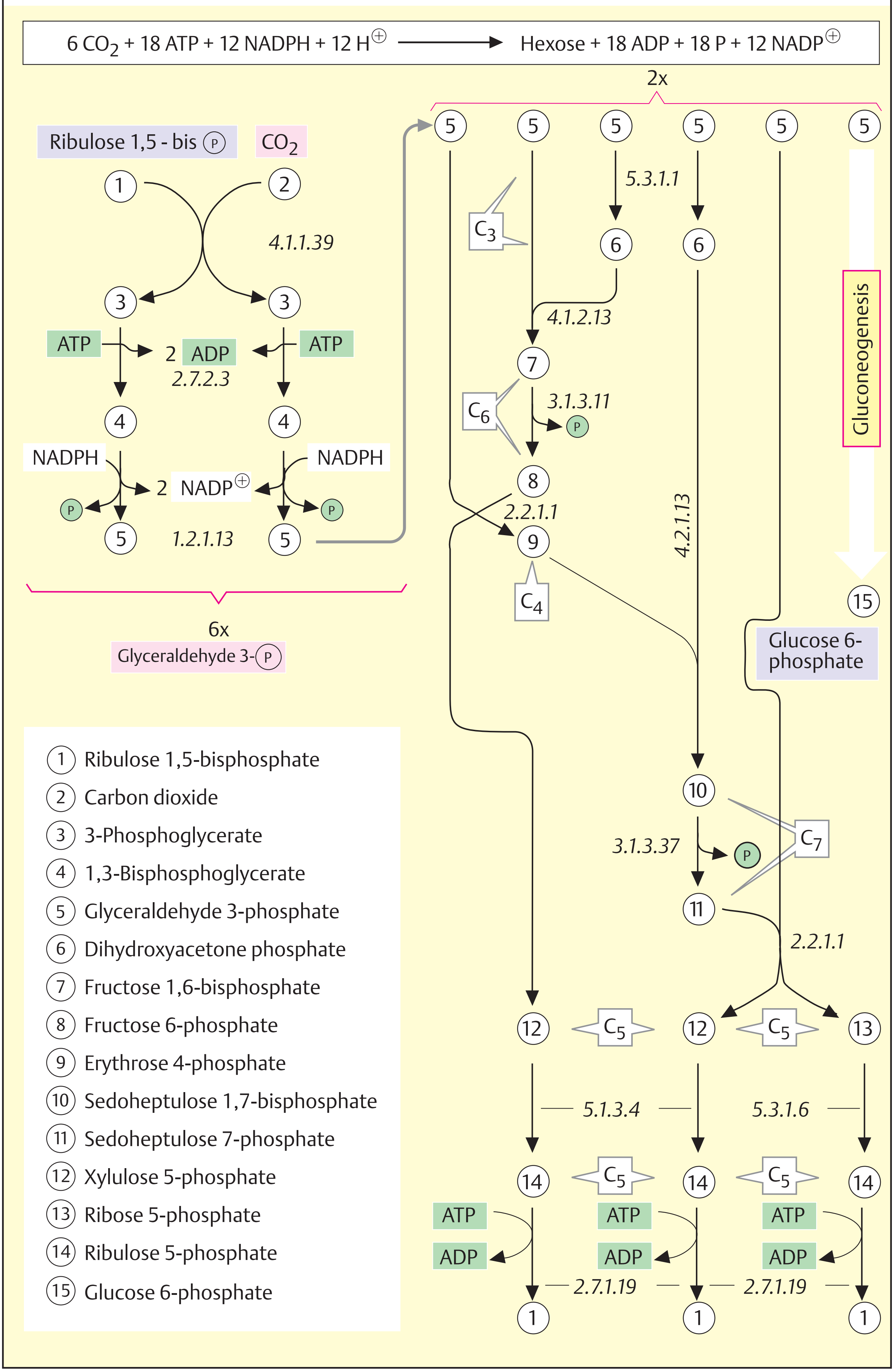
Ribulose 1,5 - bis Ⓟ
CO₂
4.1.1.39
ATP
2 ADP
2.7.2.3
NADPH
2 NADP⊕
1.2.1.13
6x
Glyceraldehyde 3-Ⓟ
2x
5.3.1.1
C₃
4.1.2.13
C₆
3.1.3.11
2.2.1.1
C₄
4.2.1.13
Gluconeogenesis
Glucose 6-phosphate
3.1.3.37
C₇
2.2.1.1
C₅
5.1.3.4
5.3.1.6
2.7.1.19

① Ribulose 1,5-bisphosphate
② Carbon dioxide
③ 3-Phosphoglycerate
④ 1,3-Bisphosphoglycerate
⑤ Glyceraldehyde 3-phosphate
⑥ Dihydroxyacetone phosphate
⑦ Fructose 1,6-bisphosphate
⑧ Fructose 6-phosphate
⑨ Erythrose 4-phosphate
⑩ Sedoheptulose 1,7-bisphosphate
⑪ Sedoheptulose 7-phosphate
⑫ Xylulose 5-phosphate
⑬ Ribose 5-phosphate
⑭ Ribulose 5-phosphate
⑮ Glucose 6-phosphate

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Carbohydrate metabolism

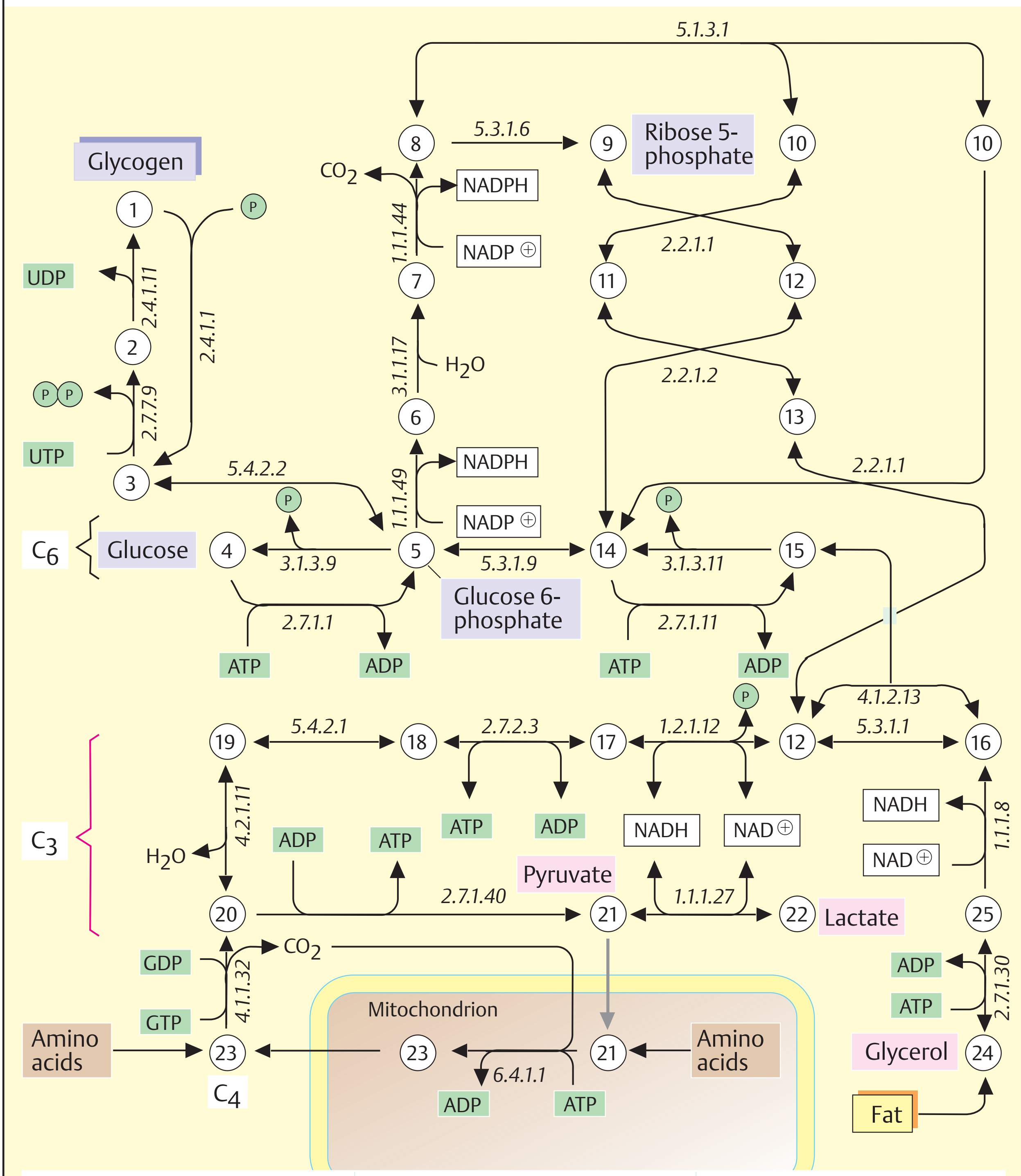
5.1.3.1
5.3.1.6
Ribose 5-phosphate
Glycogen
CO2
NADPH
NADP⊕
1.1.1.44
2.2.1.1
UDP
2.4.1.11
2.4.1.1
H2O
3.1.1.17
2.2.1.2
P P
2.7.7.9
UTP
5.4.2.2
1.1.1.49
NADPH
NADP⊕
2.2.1.1
C6
Glucose
3.1.3.9
5.3.1.9
3.1.3.11
Glucose 6-phosphate
2.7.1.1
2.7.1.11
ATP
ADP
ATP
ADP
4.1.2.13
5.4.2.1
2.7.2.3
1.2.1.12
5.3.1.1
ATP
ADP
NADH
NAD⊕
NADH
NAD⊕
1.1.1.8
C3
4.2.1.11
H2O
ADP
ATP
Pyruvate
2.7.1.40
1.1.1.27
Lactate
GDP
CO2
GTP
4.1.1.32
ADP
ATP
2.7.1.30
Mitochondrion
Amino acids
Amino acids
6.4.1.1
Glycerol
C4
ADP
ATP
Fat

①	Glycogen	⑨	Ribose 5-phosphate	⑰	1,3-Bisphosphoglycerate
②	UDP-Glucose	⑩	Xylulose 5-phosphate	⑱	3-Phosphoglycerate
③	Glucose 1-phosphate	⑪	Sedoheptulose 7-phosphate	⑲	2-Phosphoglycerate
④	Glucose	⑫	Glyceraldehyde 3-phosphate	⑳	Phosphoenolpyruvate
⑤	Glucose 6-phosphate	⑬	Erythrose 4-phosphate	㉑	Pyruvate ㉒ Lactate
⑥	Gluconolactone 6-Ⓟ	⑭	Fructose 6-phosphate	㉓	Oxaloacetate
⑦	Gluconate 6-phosphate	⑮	Fructose 1,6-bisphosphate	㉔	Glycerol
⑧	Ribulose 5-phosphate	⑯	Glycerone-3-phosphate	㉕	Glycerol 3-phosphate

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biosynthesis of fats and membrane lipids

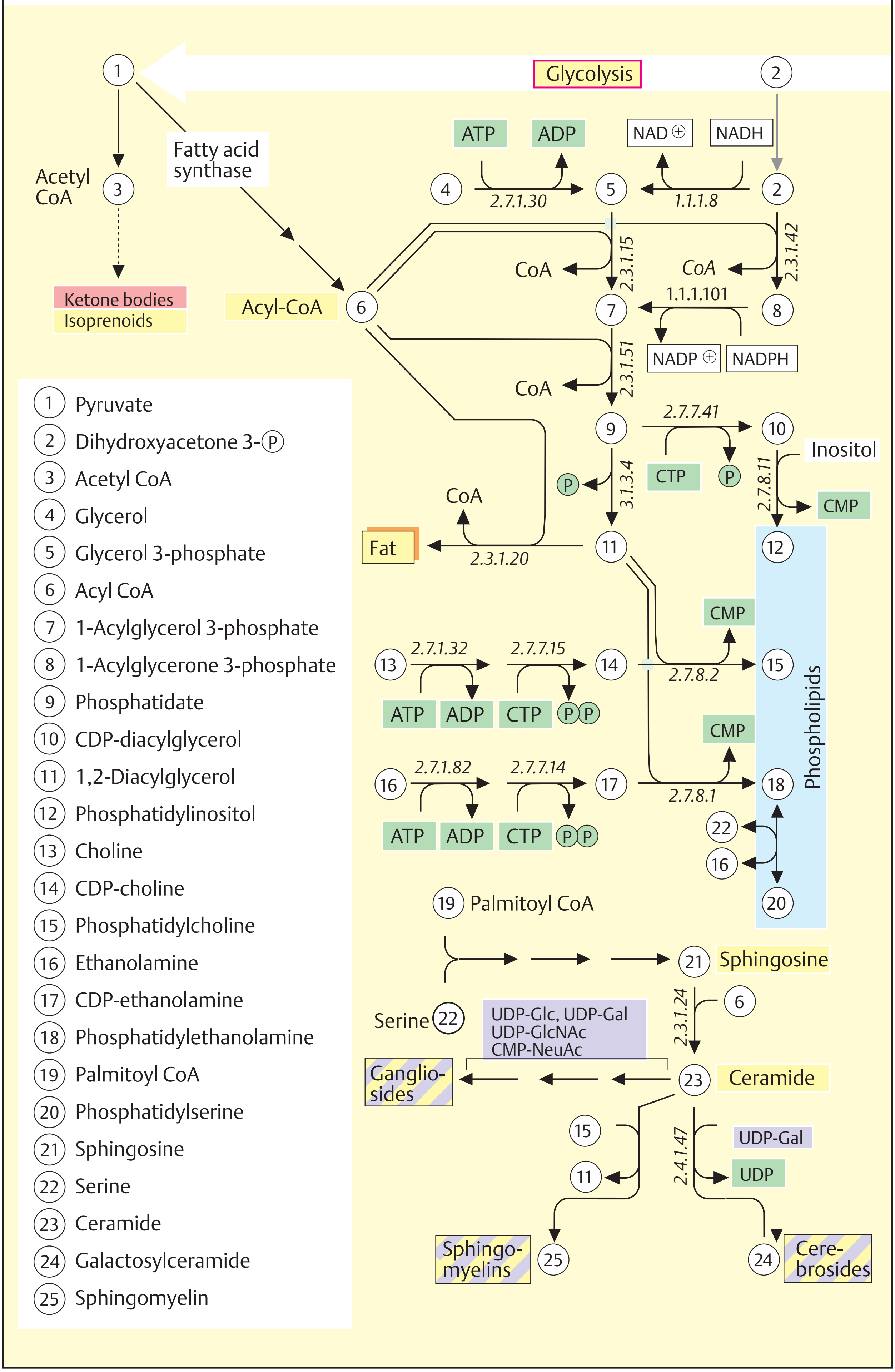
1
Glycolysis
2
ATP
ADP
NAD ⊕
NADH
Fatty acid synthase
Acetyl CoA
3
4
2.7.1.30
5
1.1.1.8
2
Ketone bodies
Isoprenoids
Acyl-CoA
6
CoA
2.3.1.15
CoA
2.3.1.42
7
1.1.1.101
8
NADP ⊕
NADPH
CoA
2.3.1.51
9
2.7.7.41
10
Inositol
Ⓟ
3.1.3.4
CTP
Ⓟ
2.7.8.11
CMP
CoA
Fat
2.3.1.20
11
12
CMP
13
2.7.1.32
2.7.7.15
14
2.7.8.2
15
ATP
ADP
CTP
ⓅⓅ
CMP
Phospholipids
16
2.7.1.82
2.7.7.14
17
2.7.8.1
18
ATP
ADP
CTP
ⓅⓅ
22
16
20
19 Palmitoyl CoA
21 Sphingosine
6
2.3.1.24
Serine 22
UDP-Glc, UDP-Gal
UDP-GlcNAc
CMP-NeuAc
Ganglio-sides
23 Ceramide
15
11
2.4.1.47
UDP-Gal
UDP
Sphingo-myelins
25
24
Cere-brosides

- ① Pyruvate
- ② Dihydroxyacetone 3-Ⓟ
- ③ Acetyl CoA
- ④ Glycerol
- ⑤ Glycerol 3-phosphate
- ⑥ Acyl CoA
- ⑦ 1-Acylglycerol 3-phosphate
- ⑧ 1-Acylglycerone 3-phosphate
- ⑨ Phosphatidate
- ⑩ CDP-diacylglycerol
- ⑪ 1,2-Diacylglycerol
- ⑫ Phosphatidylinositol
- ⑬ Choline
- ⑭ CDP-choline
- ⑮ Phosphatidylcholine
- ⑯ Ethanolamine
- ⑰ CDP-ethanolamine
- ⑱ Phosphatidylethanolamine
- ⑲ Palmitoyl CoA
- ⑳ Phosphatidylserine
- ㉑ Sphingosine
- ㉒ Serine
- ㉓ Ceramide
- ㉔ Galactosylceramide
- ㉕ Sphingomyelin

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Synthesis of ketone bodies and steroids

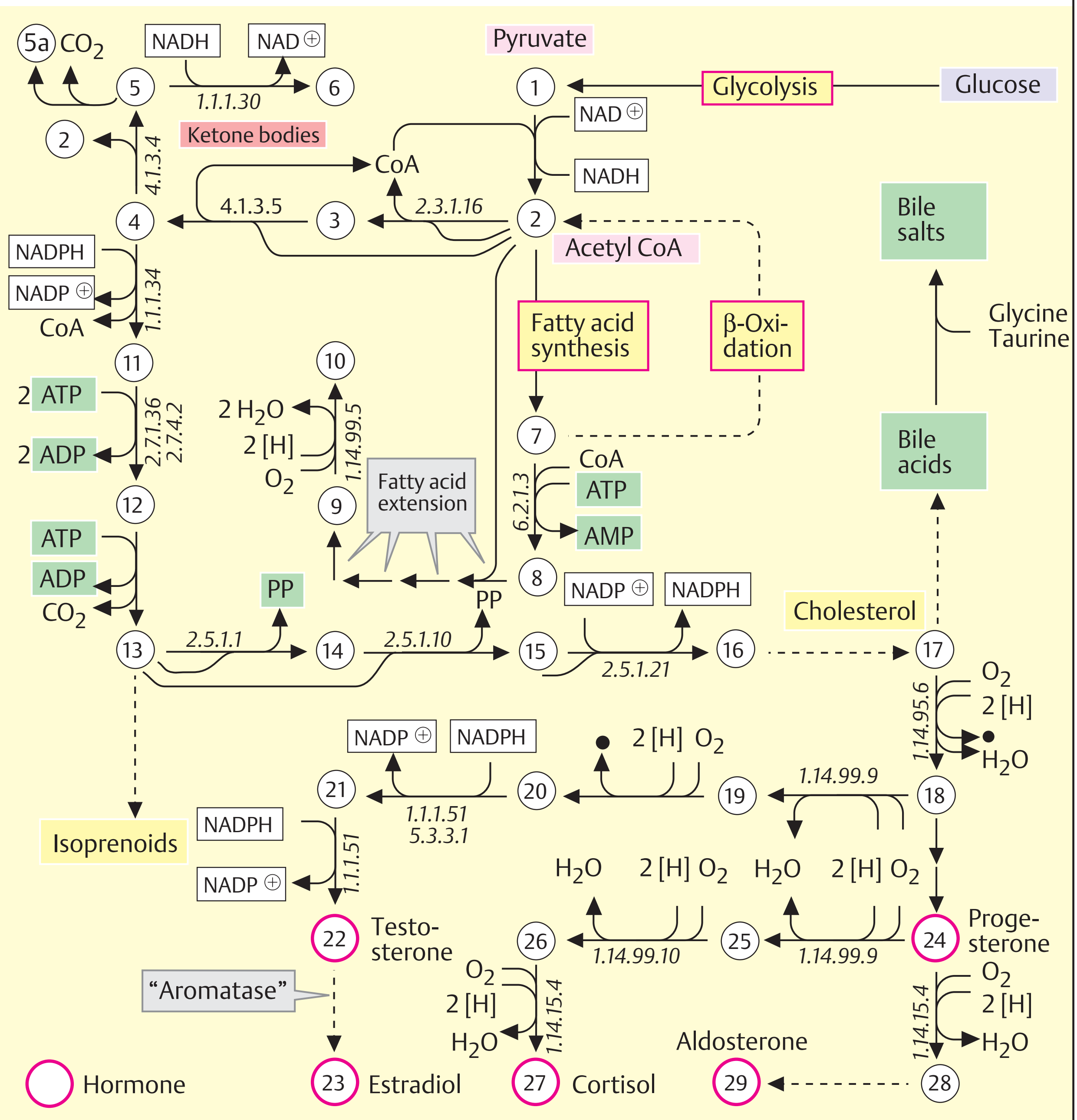
5a
CO2
NADH
NAD⊕
Pyruvate
5
1.1.1.30
6
1
Glycolysis
Glucose
2
Ketone bodies
4.1.3.4
NAD⊕
CoA
NADH
4
4.1.3.5
3
2.3.1.16
2
Acetyl CoA
Bile salts
NADPH
NADP⊕
CoA
1.1.1.34
Fatty acid synthesis
β-Oxidation
Glycine
Taurine
11
10
2 ATP
2 H2O
2.7.1.36
2.7.4.2
2 [H]
1.14.99.5
2 ADP
O2
7
Bile acids
Fatty acid extension
9
CoA
ATP
6.2.1.3
12
AMP
ATP
ADP
PP
8
CO2
PP
NADP⊕
NADPH
Cholesterol
13
2.5.1.1
14
2.5.1.10
15
16
17
2.5.1.21
O2
1.14.95.6
2 [H]
NADP⊕
NADPH
2 [H]
O2
H2O
1.14.99.9
21
20
19
18
1.1.1.51
5.3.3.1
Isoprenoids
NADPH
1.1.1.51
H2O
2 [H] O2
H2O
2 [H] O2
NADP⊕
22
Testosterone
26
25
24
Progesterone
1.14.99.10
1.14.99.9
O2
O2
"Aromatase"
2 [H]
1.14.15.4
1.14.15.4
2 [H]
H2O
Aldosterone
H2O
Hormone
23
Estradiol
27
Cortisol
29
28

①	Pyruvate	⑩	Oleyl CoA	⑳	Dehydroepiandrosterone
②	Acetyl CoA	⑪	Mevalonate	㉑	Androstene-3,17-dione
③	Acetoacetyl CoA	⑫	Mevalonate 5-diphosphate	㉒	Testosterone
④	3-Hydroxy-3-methylglutaryl CoA	⑬	Isopentenyl diphosphate	㉓	Estradiol
⑤	Acetoacetate	⑭	Geranyl diphosphate	㉔	Progesterone
(5a)	Acetone	⑮	Farnesyl diphosphate	㉕	17-OH-Progesterone
⑥	3-Hydroxybutyrate	⑯	Squalene	㉖	11-Deoxycortisol
⑦	Palmitate	⑰	Cholesterol	㉗	Cortisol
⑧	Palmitoyl CoA	⑱	Pregnenolone	㉘	11-Deoxycorticosterone
⑨	Stearoyl CoA	⑲	17-OH-Pregnenolone	㉙	Aldosterone

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Degradation of fats and phospholipids

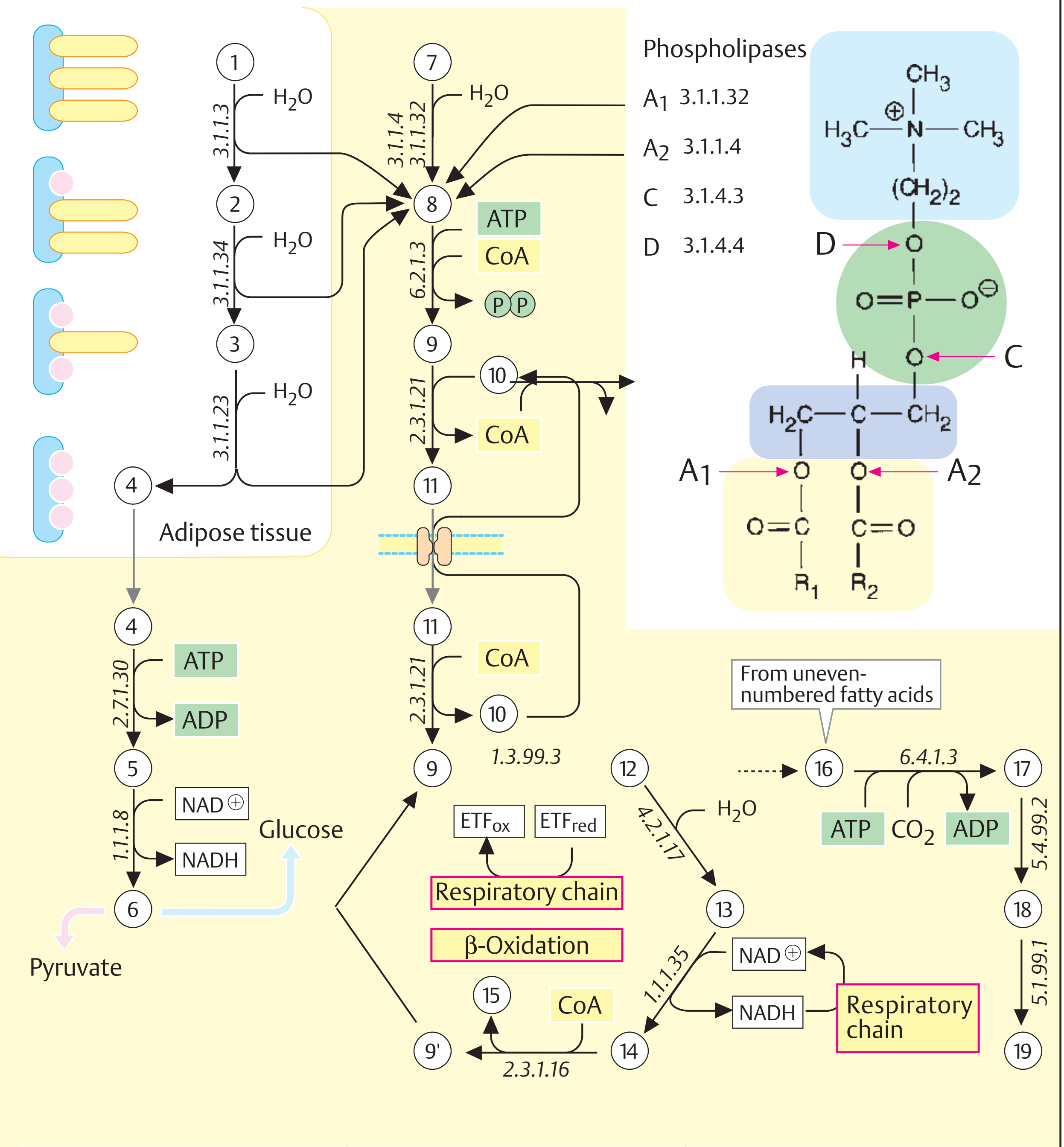
①
3.1.1.3
H_2O
②
3.1.1.34
H_2O
③
3.1.1.23
H_2O
④
Adipose tissue
⑦
3.1.1.4
3.1.1.32
H_2O
⑧
6.2.1.3
ATP
CoA
P P
⑨
2.3.1.21
⑩
CoA
⑪
⑪
2.3.1.21
CoA
⑩
⑨
Phospholipases
A1 3.1.1.32
A2 3.1.1.4
C 3.1.4.3
D 3.1.4.4
CH_3
H_3C
N
CH_3
$(CH_2)_2$
D
O
O=P—O⊖
O
C
H
H_2C—C—CH_2
A1
O
O
A2
O=C
C=O
R_1
R_2
④
2.7.1.30
ATP
ADP
⑤
1.1.1.8
NAD⊕
NADH
⑥
Glucose
Pyruvate
1.3.99.3
ETFox
ETFred
Respiratory chain
β-Oxidation
⑫
4.2.1.17
H_2O
⑬
1.1.1.35
NAD⊕
NADH
Respiratory chain
⑭
2.3.1.16
CoA
⑮
⑨'
From uneven-numbered fatty acids
⑯
6.4.1.3
ATP
CO_2
ADP
⑰
5.4.99.2
⑱
5.1.99.1
⑲

①	Triacylglycerol
②	Diacylglycerol
③	Monoacylglycerol
④	Glycerol
⑤	Glycerol 3-phosphate
⑥	Glycerone 3-phosphate
⑦	Phospholipid
⑧	Free fatty acid
⑨	Acyl CoA
(9')	Shortened acyl CoA
⑩	Carnitine
⑪	Acyl carnitine
⑫	2,3-Dehydroacyl CoA
⑬	3-Hydroxyacyl CoA
⑭	3-Oxoacyl CoA
⑮	Acetyl CoA
⑯	Propionyl CoA
⑰	(S)-Methylmalonyl CoA
⑱	(R)-Methylmalonyl CoA
⑲	Succinyl CoA

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biosynthesis of the essential amino acids

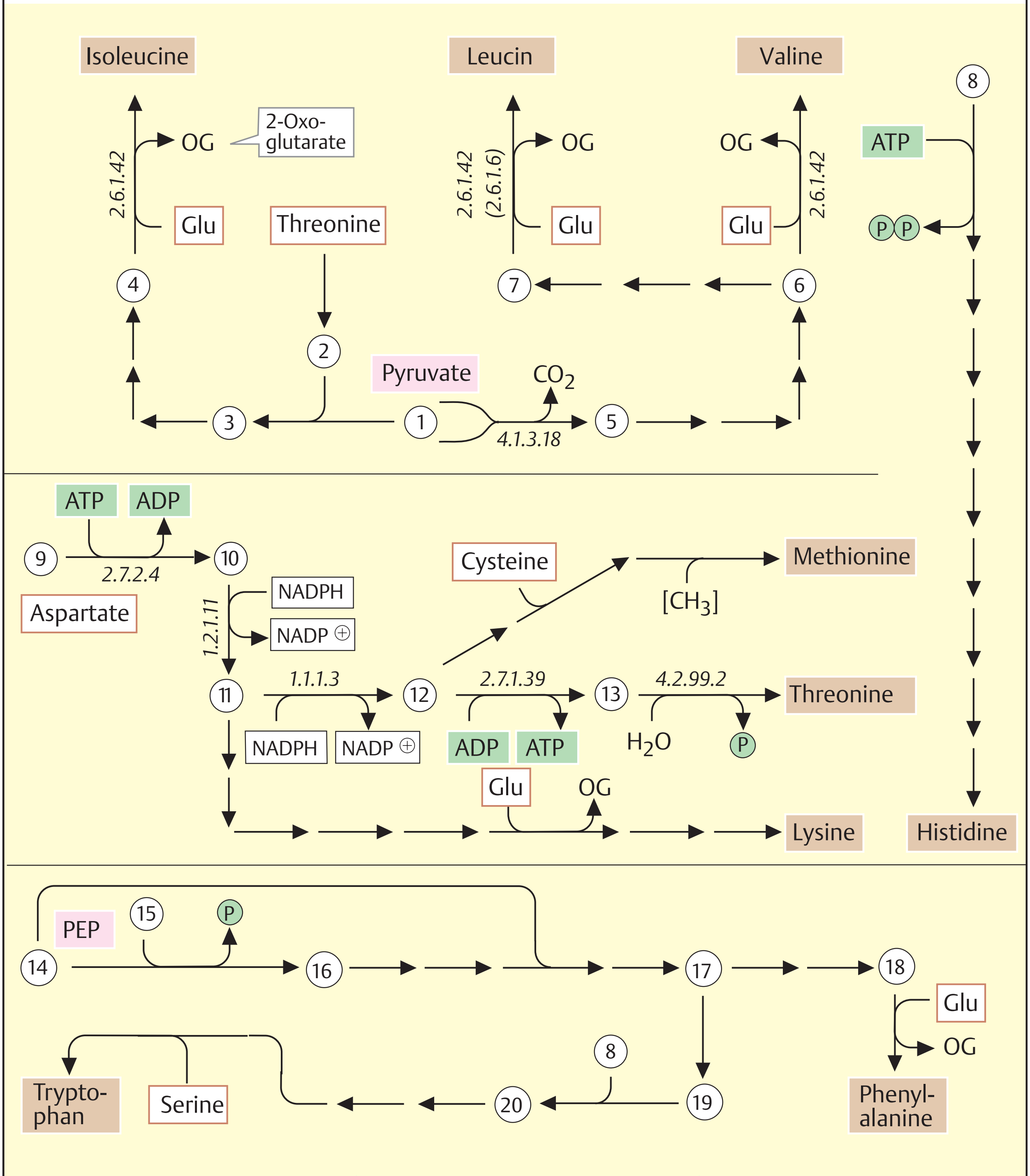
Isoleucine
Leucin
Valine
2-Oxo-glutarate
OG
Glu
Threonine
2.6.1.42
(2.6.1.6)
ATP
Ⓟ Ⓟ
⑧
④
②
⑦
⑥
Pyruvate
CO_2
③
①
⑤
4.1.3.18
ATP
ADP
⑨
⑩
2.7.2.4
Aspartate
Cysteine
Methionine
NADPH
NADP ⊕
1.2.1.11
$[CH_3]$
⑪
1.1.1.3
⑫
2.7.1.39
⑬
4.2.99.2
Threonine
NADPH
NADP ⊕
ADP
ATP
H_2O
Ⓟ
Glu
OG
Lysine
Histidine
PEP
⑮
Ⓟ
⑭
⑯
⑰
⑱
Glu
OG
⑧
Trypto-phan
Serine
⑳
⑲
Phenyl-alanine

① Pyruvate
② 2-Oxobutyrate
③ 2-Aceto-2-hydroxybutyrate
④ 2-Oxo-4-methylvalerate
⑤ 2-Acetolactate
⑥ 2-Oxoisovalerate
⑦ 2-Oxoisocaproate
⑧ Phosphoribosyl diphosphate
⑨ Aspartate
⑩ Aspartyl 4-phosphate
⑪ Aspartate 4-semialdehyde
⑫ Homoserine
⑬ Phosphohomoserine
⑭ Phosphoenolpyruvate
⑮ Erythrose 4-phosphate
⑯ 2-Oxo-3-deoxy-arabinoheptulosonate 7-phosphate
⑰ Chorismate
⑱ Phenylpyruvate
⑲ Anthranilate
⑳ *N*-(Phosphoribosyl)-anthranilate

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biosynthesis of the non-essential amino acids

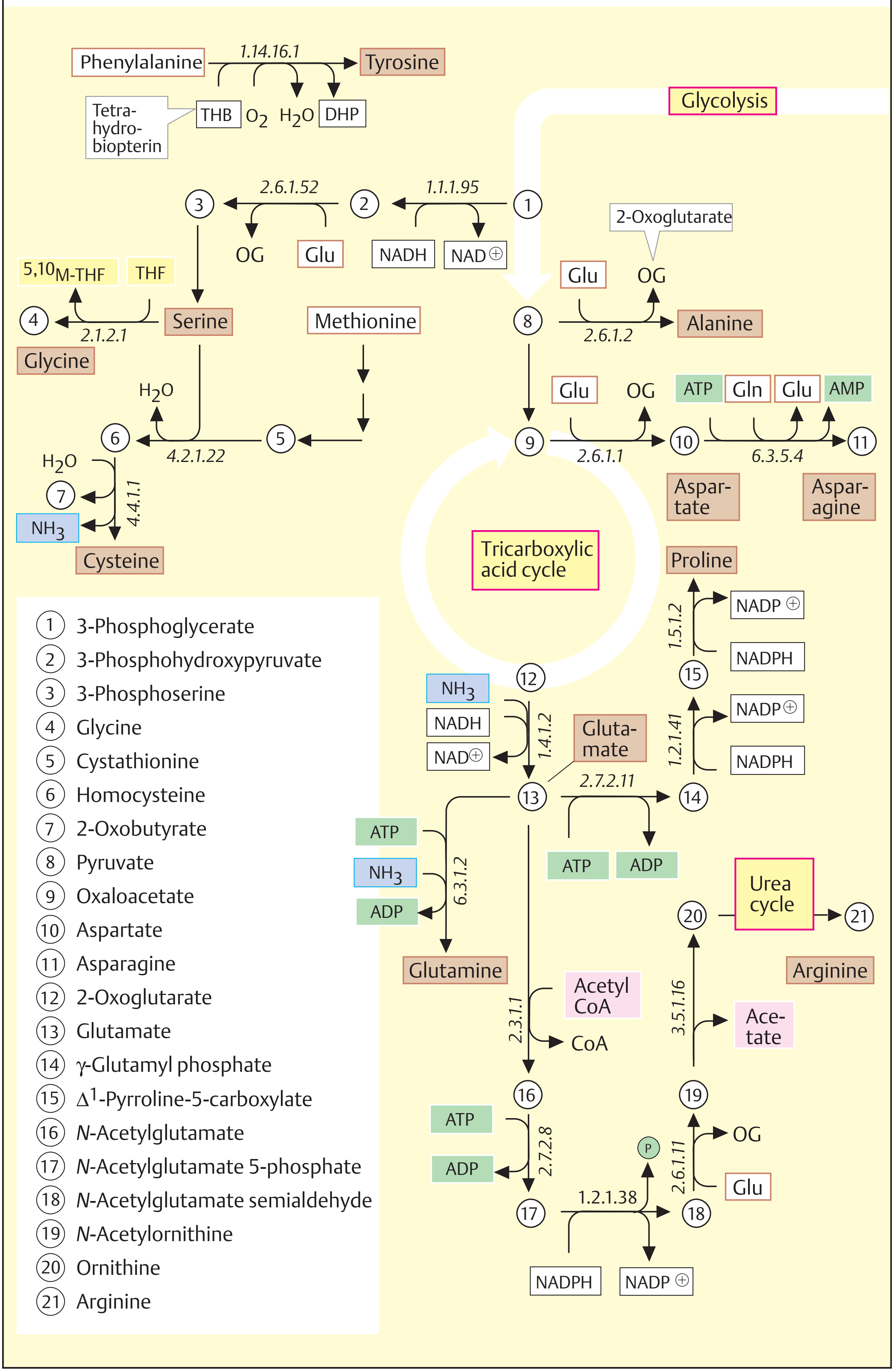
Phenylalanine
1.14.16.1
Tyrosine
Tetra-hydro-biopterin
THB
O_2
H_2O
DHP
Glycolysis
③
2.6.1.52
②
1.1.1.95
①
OG
Glu
NADH
NAD⊕
2-Oxoglutarate
Glu
OG
5,10M-THF
THF
④
2.1.2.1
Serine
Methionine
⑧
2.6.1.2
Alanine
Glycine
H_2O
⑥
4.2.1.22
⑤
Glu
OG
⑨
2.6.1.1
⑩
ATP
Gln
Glu
AMP
6.3.5.4
⑪
H_2O
⑦
NH_3
4.4.1.1
Cysteine
Aspar-tate
Aspar-agine
Tricarboxylic acid cycle
Proline
1.5.1.2
NADP ⊕
NADPH
⑮
⑫
NH_3
NADH
NAD⊕
1.4.1.2
Gluta-mate
NADP⊕
NADPH
1.2.1.41
⑬
2.7.2.11
⑭
ATP
NH_3
ADP
6.3.1.2
ATP
ADP
Glutamine
Urea cycle
⑳
㉑
Arginine
Acetyl CoA
2.3.1.1
CoA
3.5.1.16
Ace-tate
⑯
⑲
ATP
ADP
2.7.2.8
P
OG
2.6.1.11
Glu
⑰
1.2.1.38
⑱
NADPH
NADP ⊕

① 3-Phosphoglycerate
② 3-Phosphohydroxypyruvate
③ 3-Phosphoserine
④ Glycine
⑤ Cystathionine
⑥ Homocysteine
⑦ 2-Oxobutyrate
⑧ Pyruvate
⑨ Oxaloacetate
⑩ Aspartate
⑪ Asparagine
⑫ 2-Oxoglutarate
⑬ Glutamate
⑭ γ-Glutamyl phosphate
⑮ Δ^1-Pyrroline-5-carboxylate
⑯ *N*-Acetylglutamate
⑰ *N*-Acetylglutamate 5-phosphate
⑱ *N*-Acetylglutamate semialdehyde
⑲ *N*-Acetylornithine
⑳ Ornithine
㉑ Arginine

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Amino acid degradation I

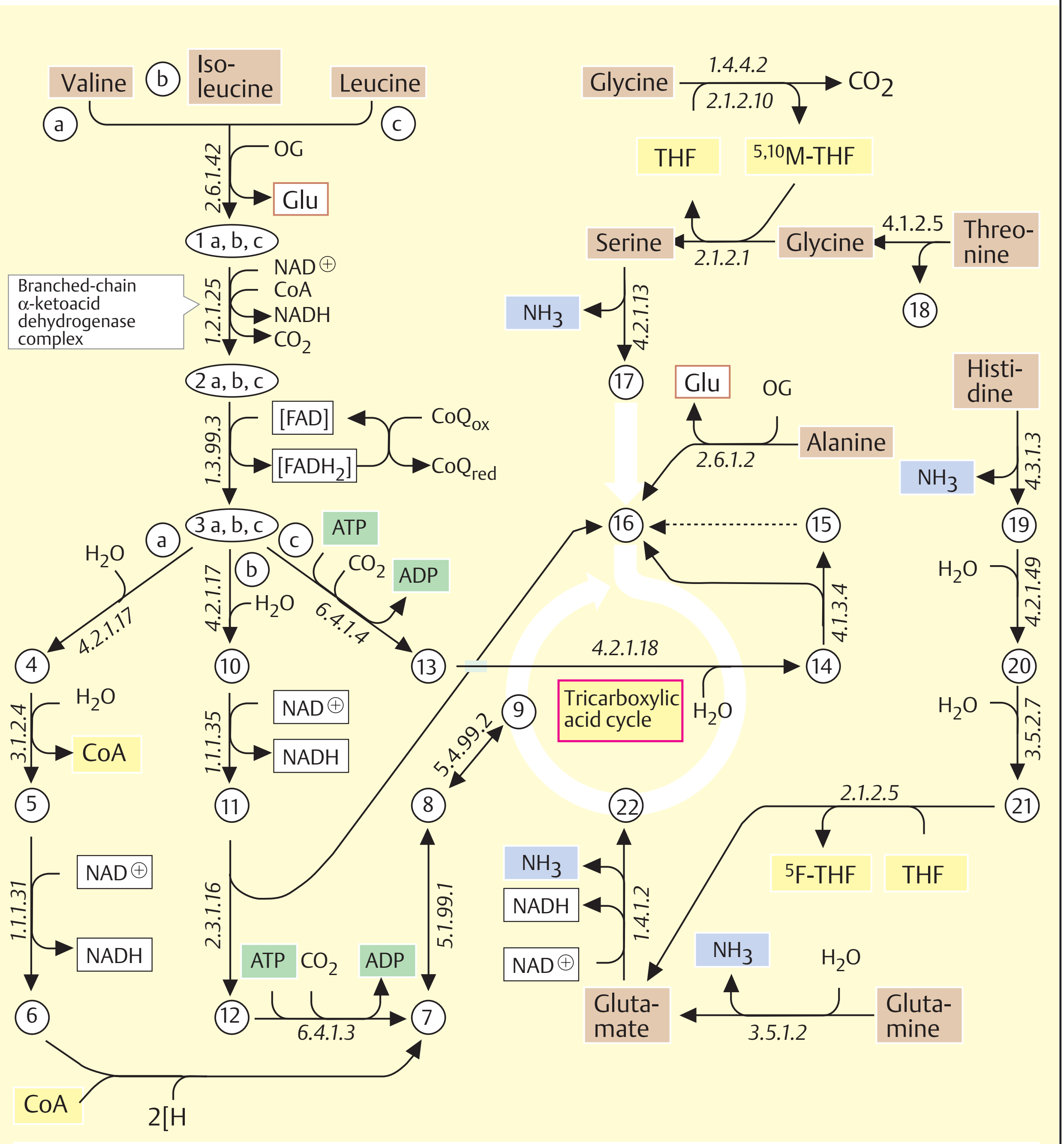
Valine
b
Iso-leucine
Leucine
a
c
2.6.1.42
OG
Glu
1 a, b, c
Branched-chain α-ketoacid dehydrogenase complex
1.2.1.25
NAD⊕
CoA
NADH
CO_2
2 a, b, c
1.3.99.3
[FAD]
[$FADH_2$]
CoQ_{ox}
CoQ_{red}
3 a, b, c
a
b
c
ATP
CO_2
ADP
H_2O
4.2.1.17
4.2.1.17
H_2O
6.4.1.4
4
10
13
3.1.2.4
H_2O
CoA
1.1.1.35
NAD⊕
NADH
5.4.99.2
9
8
5
11
1.1.1.31
NAD⊕
NADH
2.3.1.16
ATP
CO_2
ADP
5.1.99.1
6
12
6.4.1.3
7
CoA
2[H
Glycine
1.4.4.2
2.1.2.10
CO_2
THF
5,10M-THF
Serine
2.1.2.1
Glycine
4.1.2.5
Threo-nine
18
NH_3
4.2.1.13
17
Glu
OG
Alanine
2.6.1.2
Histi-dine
NH_3
4.3.1.3
16
15
19
4.1.3.4
H_2O
4.2.1.49
4.2.1.18
14
20
Tricarboxylic acid cycle
H_2O
H_2O
3.5.2.7
22
2.1.2.5
21
NH_3
NADH
NAD⊕
1.4.1.2
5F-THF
THF
NH_3
H_2O
Gluta-mate
3.5.1.2
Gluta-mine

1a	2-Oxoisovalerate
1b	2-Oxo-3-methylvalerate
1c	2-Oxoisocapronate
2a	Isobutyryl CoA
2b	2-Methylbutyryl CoA
2c	Isovaleryl CoA
3a	Methylacrylyl CoA
3b	Tiglyl CoA
3c	3-Methylcrotonyl CoA
4	3-Hydroxyisobutyryl CoA
5	3-Hydroxyisobutyrate
6	Methylmalonyl-semialdehyde
7	(S)-Methylmalonyl CoA
8	(R)-Methylmalonyl CoA
9	Succinyl CoA
10	2-Methyl-3-hydroxybutyryl CoA
11	2-Methylacetoacetyl CoA
12	Propionyl CoA
13	3-Methylglutaconyl CoA
14	3-Hydroxy-3-methylglutaryl CoA
15	Acetoacetate
16	Acetyl CoA
17	Pyruvate
18	Acetaldehyde
19	Urocanate
20	Imidazolone-5-propionate
21	*N*-Formimino-glutamate
22	2-Oxoglutarate

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Amino acid degradation II

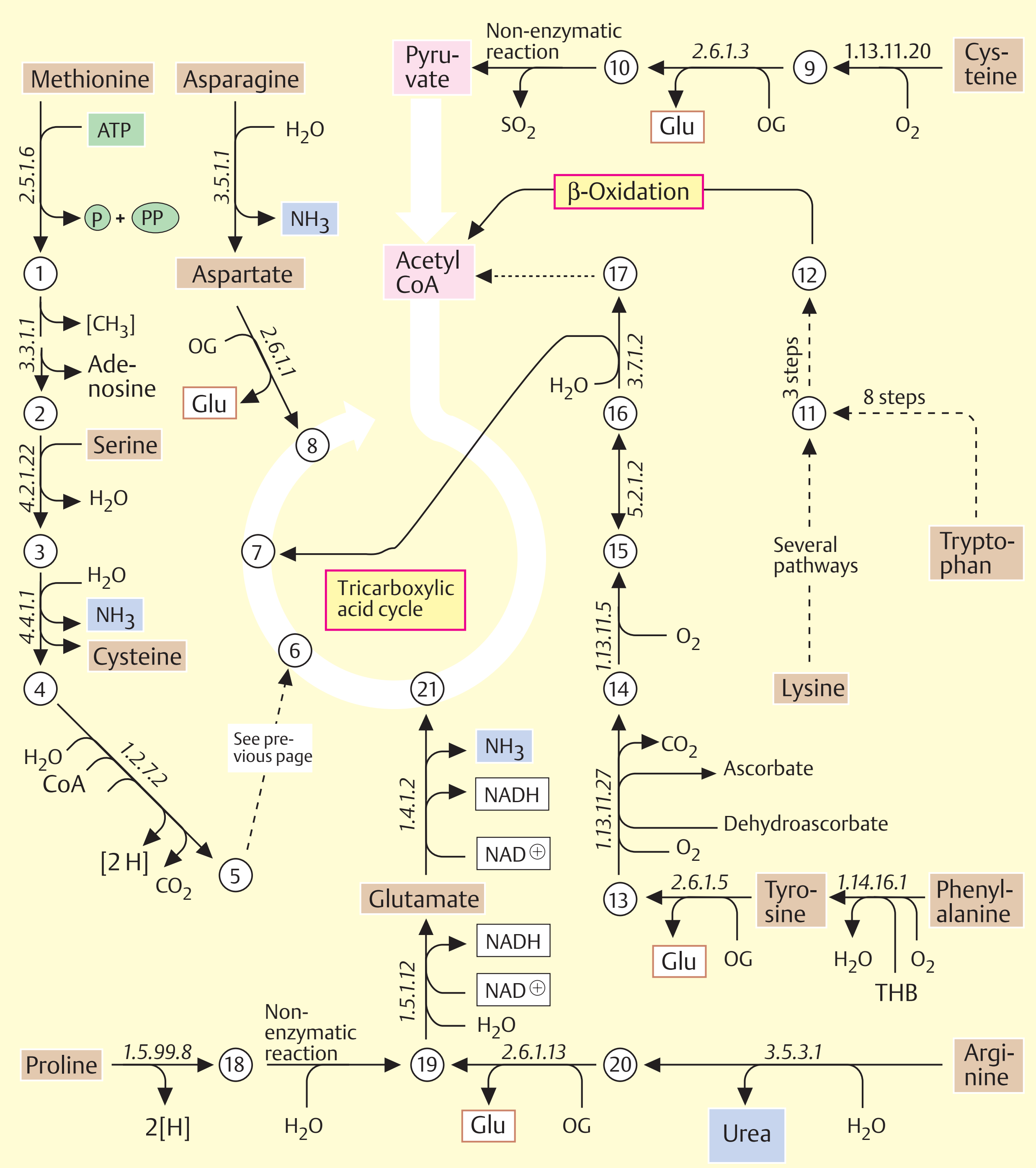
Methionine
Asparagine
Pyruvate
Non-enzymatic reaction
2.6.1.3
1.13.11.20
Cysteine
SO_2
Glu
OG
O_2
ATP
2.5.1.6
P + PP
H_2O
3.5.1.1
NH_3
β-Oxidation
Aspartate
Acetyl CoA
3.3.1.1
$[CH_3]$
Adenosine
OG
2.6.1.1
Glu
3.7.1.2
H_2O
3 steps
8 steps
Serine
4.2.1.22
H_2O
5.2.1.2
Several pathways
Tryptophan
H_2O
NH_3
4.4.1.1
Cysteine
Tricarboxylic acid cycle
1.13.11.5
O_2
Lysine
H_2O
CoA
1.2.7.2
See previous page
[2 H]
CO_2
1.4.1.2
NADH
$NAD^{\oplus}$
1.13.11.27
CO_2
Ascorbate
Dehydroascorbate
O_2
Glutamate
2.6.1.5
Tyrosine
1.14.16.1
Phenylalanine
Glu
OG
H_2O
O_2
THB
1.5.1.12
NADH
$NAD^{\oplus}$
H_2O
Proline
1.5.99.8
Non-enzymatic reaction
2.6.1.13
3.5.3.1
Arginine
2[H]
H_2O
Glu
OG
Urea
H_2O
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

① *S*-Adenosylmethionine
② Homocysteine
③ Cystathionine
④ 2-Oxobutyrate
⑤ Propionyl CoA
⑥ Succinyl CoA
⑦ Fumarate
⑧ Oxaloacetate
⑨ Cysteine sulfinate
⑩ 3-Sulfinylpyruvate
⑪ 2-Oxoadipate
⑫ Crotonyl CoA
⑬ 4-Hydroxyphenylpyruvate
⑭ Homogentisate
⑮ 4-Maleylacetoacetate
⑯ Fumarylacetoacetate
⑰ Acetoacetate
⑱ Glutamate 4-semialdehyde
⑲ Δ^1-Pyrroline-5-carboxylate
⑳ Ornithine
㉑ 2-Oxoglutarate

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Ammonia metabolism

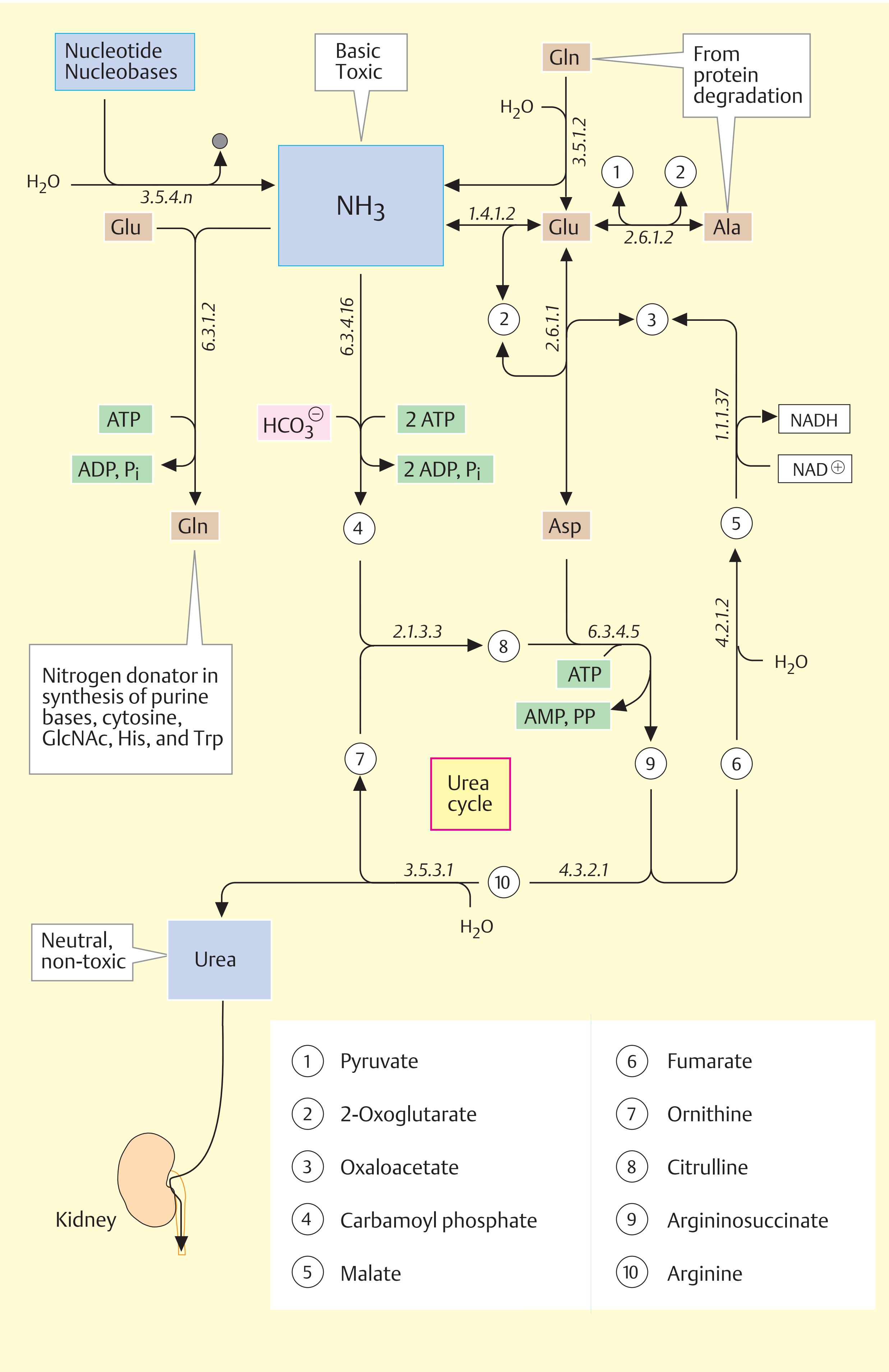
Nucleotide
Nucleobases
Basic
Toxic
Gln
From protein degradation
H_2O
3.5.1.2
H_2O
3.5.4.n
NH_3
1
2
1.4.1.2
Glu
2.6.1.2
Ala
Glu
6.3.1.2
6.3.4.16
2
2.6.1.1
3
1.1.1.37
ATP
ADP, P_i
HCO_3^-
2 ATP
2 ADP, P_i
NADH
NAD^+
Gln
4
Asp
5
4.2.1.2
H_2O
2.1.3.3
8
6.3.4.5
ATP
AMP, PP
Nitrogen donator in synthesis of purine bases, cytosine, GlcNAc, His, and Trp
7
Urea cycle
9
6
3.5.3.1
10
4.3.2.1
H_2O
Neutral, non-toxic
Urea
Kidney

1	Pyruvate	6	Fumarate
2	2-Oxoglutarate	7	Ornithine
3	Oxaloacetate	8	Citrulline
4	Carbamoyl phosphate	9	Argininosuccinate
5	Malate	10	Arginine

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biosynthesis of purine nucleotides

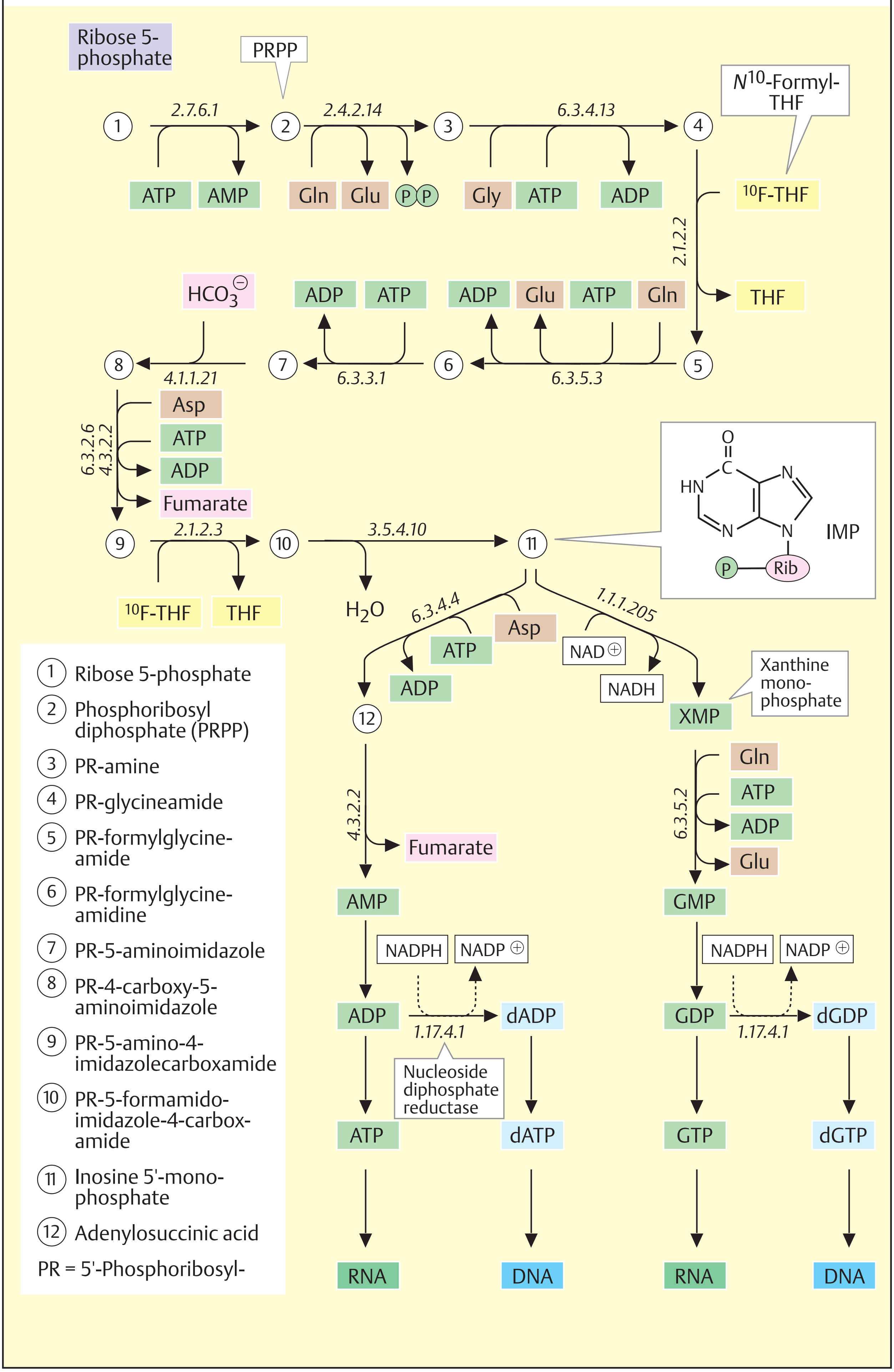
Ribose 5-phosphate
PRPP
N10-Formyl-THF
① 2.7.6.1 ② 2.4.2.14 ③ 6.3.4.13 ④
ATP
AMP
Gln
Glu
P P
Gly
ATP
ADP
10F-THF
2.1.2.2
THF
HCO3⊖
ADP
ATP
ADP
Glu
ATP
Gln
⑧ 4.1.1.21 ⑦ 6.3.3.1 ⑥ 6.3.5.3 ⑤
6.3.2.6
4.3.2.2
Asp
ATP
ADP
Fumarate
⑨ 2.1.2.3 ⑩ 3.5.4.10 ⑪
10F-THF
THF
H2O
O
C
HN
N
N
N
P
Rib
IMP
6.3.4.4
Asp
ATP
ADP
1.1.1.205
NAD⊕
NADH
⑫
XMP
Xanthine mono-phosphate
4.3.2.2
Fumarate
6.3.5.2
Gln
ATP
ADP
Glu
AMP
GMP
NADPH
NADP⊕
NADPH
NADP⊕
ADP
dADP
1.17.4.1
GDP
dGDP
1.17.4.1
Nucleoside diphosphate reductase
ATP
dATP
GTP
dGTP
RNA
DNA
RNA
DNA

① Ribose 5-phosphate
② Phosphoribosyl diphosphate (PRPP)
③ PR-amine
④ PR-glycineamide
⑤ PR-formylglycine-amide
⑥ PR-formylglycine-amidine
⑦ PR-5-aminoimidazole
⑧ PR-4-carboxy-5-aminoimidazole
⑨ PR-5-amino-4-imidazolecarboxamide
⑩ PR-5-formamido-imidazole-4-carbox-amide
⑪ Inosine 5'-mono-phosphate
⑫ Adenylosuccinic acid

PR = 5'-Phosphoribosyl-

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Biosynthesis of the pyrimidine nucleotides and C_1 metabolism

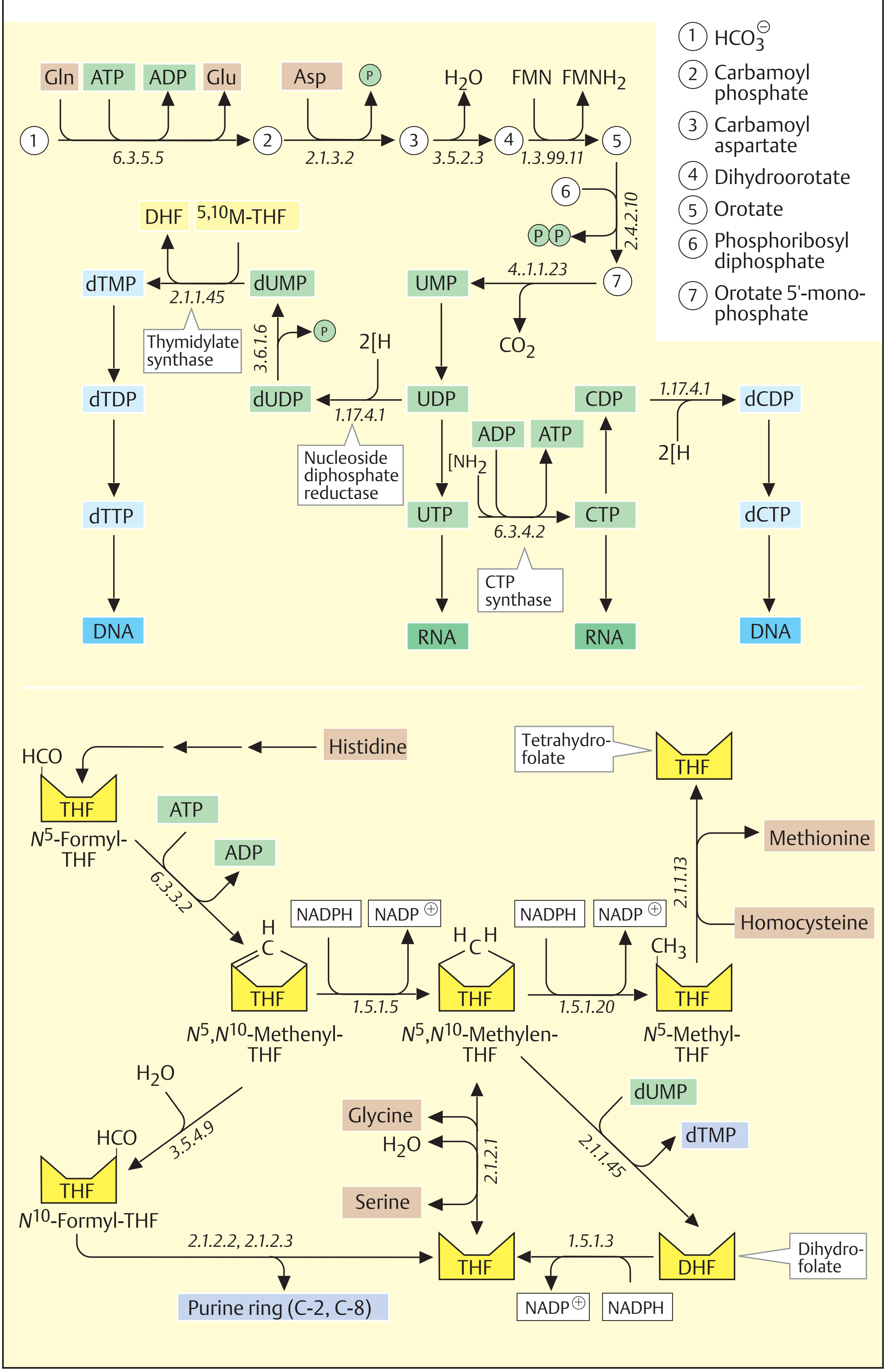
① HCO_3^-
② Carbamoyl phosphate
③ Carbamoyl aspartate
④ Dihydroorotate
⑤ Orotate
⑥ Phosphoribosyl diphosphate
⑦ Orotate 5'-monophosphate
Gln
ATP
ADP
Glu
6.3.5.5
Asp
P
2.1.3.2
H_2O
3.5.2.3
FMN
$FMNH_2$
1.3.99.11
2.4.2.10
P P
4..1.1.23
CO_2
DHF
5,10M-THF
dTMP
2.1.1.45
dUMP
UMP
Thymidylate synthase
3.6.1.6
P
2[H
dTDP
dUDP
1.17.4.1
UDP
CDP
1.17.4.1
dCDP
Nucleoside diphosphate reductase
ADP
ATP
[NH2
2[H
dTTP
UTP
6.3.4.2
CTP
dCTP
CTP synthase
DNA
RNA
RNA
DNA
HCO
Histidine
Tetrahydrofolate
THF
THF
N^5-Formyl-THF
ATP
ADP
6.3.3.2
Methionine
2.1.1.13
NADPH
NADP⊕
NADPH
NADP⊕
Homocysteine
H
C
THF
1.5.1.5
H H
C
THF
CH_3
1.5.1.20
THF
N^5,N^{10}-Methenyl-THF
N^5,N^{10}-Methylen-THF
N^5-Methyl-THF
H_2O
dUMP
3.5.4.9
Glycine
dTMP
HCO
H_2O
2.1.2.1
2.1.1.45
THF
Serine
N^{10}-Formyl-THF
2.1.2.2, 2.1.2.3
1.5.1.3
THF
DHF
Dihydrofolate
Purine ring (C-2, C-8)
NADP⊕
NADPH

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

A. Nucleotide degradation

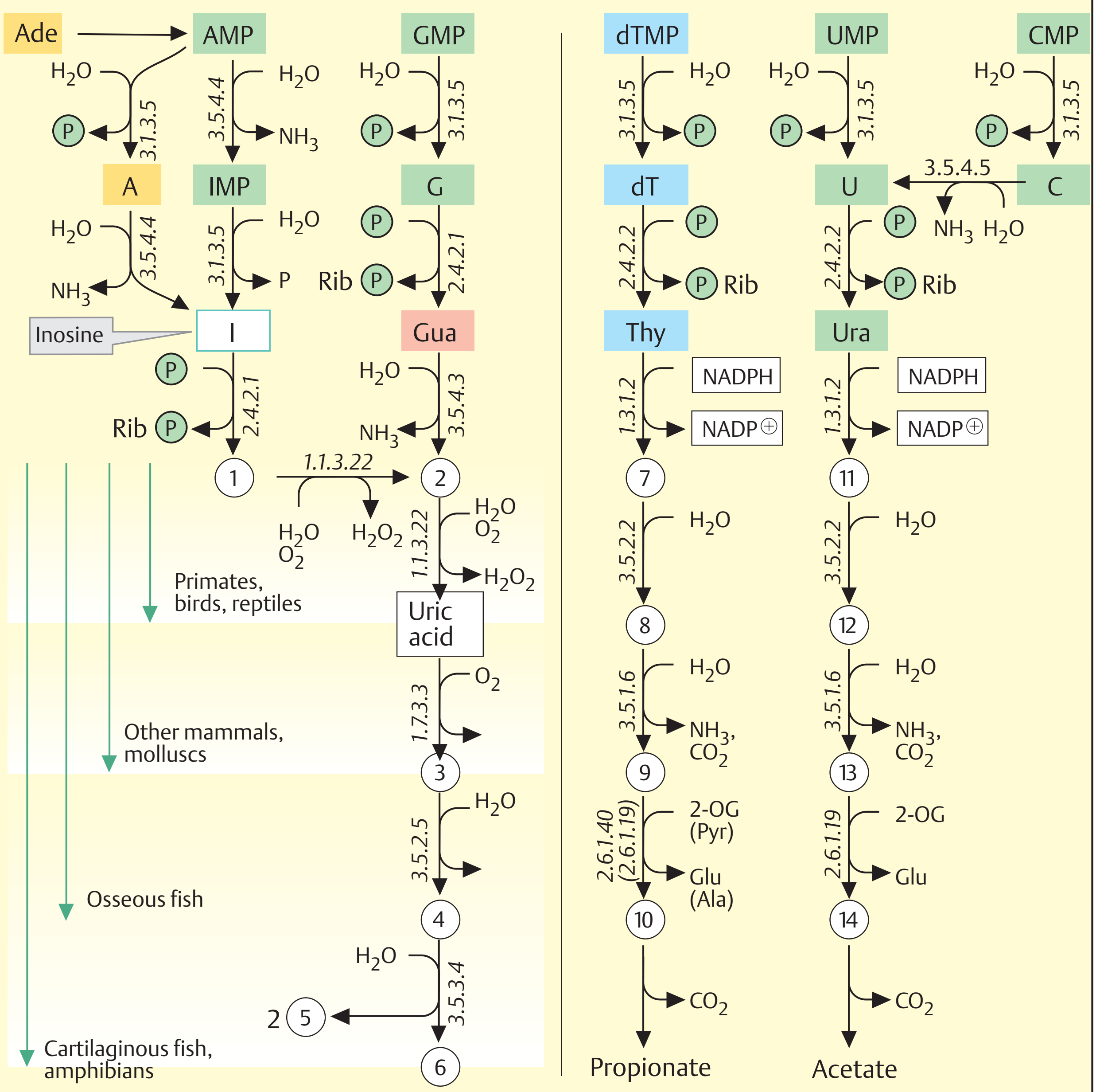
Ade
AMP
GMP
dTMP
UMP
CMP
H_2O
Ⓟ
3.1.3.5
3.5.4.4
NH_3
A
IMP
G
dT
U
C
3.5.4.5
NH_3 H_2O
P
Rib Ⓟ
2.4.2.1
2.4.2.2
Ⓟ Rib
Inosine
I
Gua
Thy
Ura
3.5.4.3
1.3.1.2
NADPH
NADP⊕
① ② ⑦ ⑪
1.1.3.22
H_2O O_2
H_2O_2
Primates, birds, reptiles
Uric acid
O_2
1.7.3.3
Other mammals, molluscs
③
3.5.2.2
⑧ ⑫
3.5.1.6
NH_3, CO_2
⑨ ⑬
2.6.1.40 (2.6.1.19)
2.6.1.19
2-OG (Pyr)
Glu (Ala)
2-OG
Glu
3.5.2.5
④
Osseous fish
⑩ ⑭
CO_2
3.5.3.4
2 ⑤
⑥
Cartilaginous fish, amphibians
Propionate
Acetate

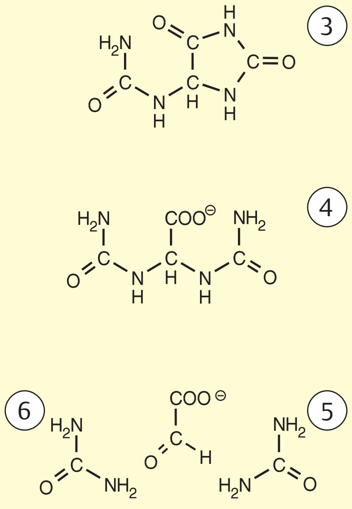
③
④
⑤
⑥
H_2N
NH_2
$COO^⊖$
C=O
N H

①	Hypoxanthine (Hyp)
②	Xanthine (Xan)
③	Allantoin
④	Allantoic acid
⑤	Glyoxylate
⑥	Urea
⑦	Dihydrothymine
⑧	Carbamoyl-β-amino-isobutyrate
⑨	β-Aminoisobutyrate
⑩	2-Methyl-3-oxopropionate
⑪	Dihydrouracil
⑫	Carbamoyl-β-alanine
⑬	β-alanine
⑭	3-oxopropionate

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Annotated enzyme list

Only the enzymes mentioned in this atlas are listed here, from among the more than 2000 enzymes known. The enzyme names are based on the IUBMB's official *Enzyme nomenclature 1992*. The additions shown in round brackets belong to the enzyme name, while prosthetic groups and other cofactors are enclosed in square brackets. Common names of enzyme groups are given in italics, and trivial names are shown in quotation marks.

Class 1: Oxidoreductases (catalyze reduction-oxidation reactions)

Subclass 1.n: What is the electron *donor*?
Sub-subclass 1.n.n: What is the electron *acceptor*?

1.1 A –CH–OH group is the donor

1.1.1	**NAD(P)$^+$ is the acceptor** (*dehydrogenases, reductases*)
1.1.1.1	Alcohol dehydrogenase [Zn^{2+}]
1.1.1.3	Homoserine dehydrogenase
1.1.1.8	Glycerol 3-phosphate dehydrogenase (NAD$^+$)
1.1.1.21	Aldehyde reductase
1.1.1.27	Lactate dehydrogenase
1.1.1.30	3-Hydroxybutyrate dehydrogenase
1.1.1.31	3-Hydroxyisobutyrate dehydrogenase
1.1.1.34	Hydroxymethylglutaryl-CoA reductase (NADPH)
1.1.1.35	3-Hydroxyacyl-CoA dehydrogenase
1.1.1.37	Malate dehydrogenase
1.1.1.40	Malate dehydrogenase (oxaloacetate-decarboxylating, NADP$^+$)—"malic enzyme"
1.1.1.41	Isocitrate dehydrogenase (NAD$^+$)
1.1.1.42	Isocitrate dehydrogenase (NADP$^+$)
1.1.1.44	Phosphogluconate dehydrogenase (decarboxylating)
1.1.1.49	Glucose 6-phosphate 1-dehydrogenase
1.1.1.51	3(or 17)β-Hydroxysteroid dehydrogenase
1.1.1.95	Phosphoglycerate dehydrogenase
1.1.1.100	3-Oxoacyl-[ACP] reductase
1.1.1.101	Acylglycerone phosphate reductase
1.1.1.105	Retinol dehydrogenase
1.1.1.145	3β-Hydroxy-Δ^5-steroid dehydrogenase
1.1.1.205	IMP dehydrogenase

1.1.3	**Molecular oxygen is the acceptor** (*oxidases*)
1.1.3.4	Glucose oxidase [FAD]
1.1.3.8	L-Gulonolactone oxidase
1.1.3.22	Xanthine oxidase [Fe, Mo, FAD]
1.1.99.5	Glycerol-3-phosphate dehydrogenase (FAD)

1.2 An aldehyde or keto group is the donor

1.2.1	**NAD(P)$^+$ is the acceptor** *(dehydrogenases)*
1.2.1.3	Aldehyde dehydrogenase (NAD$^+$)
1.2.1.11	Aspartate semialdehyde dehydrogenase
1.2.1.12	Glyceraldehyde 3-phosphate dehydrogenase
1.2.1.13	Glyceraldehyde 3-phosphate dehydrogenase (NADP$^+$) (phosphorylating)
1.2.1.24	Succinate semialdehyde dehydrogenase
1.2.1.25	2-Oxoisovalerate dehydrogenase (acylating)
1.2.1.38	*N*-Acetyl-γ-glutamylphosphate reductase
1.2.1.41	Glutamylphosphate reductase

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

1.2.4	**A disulfide is the acceptor**
1.2.4.1	Pyruvate dehydrogenase (lipoamide) [TPP]
1.2.4.2	Oxoglutarate dehydrogenase (lipoamide) [TPP]
1.2.7	**An Fe/S protein is the acceptor**
1.2.7.2	2-Oxobutyrate synthase
1.3	**A –CH–CH– group is the donor**
1.3.1.10	Enoyl-[ACP] reductase (NADPH)
1.3.1.24	Biliverdin reductase
1.3.1.34	2,4-Dienoyl-CoA reductase
1.3.5.1	Succinate dehydrogenase (ubiquinone) [FAD, Fe_2S_2, Fe_4S_4], “complex II”
1.3.99.3	Acyl-CoA dehydrogenase [FAD]
1.3.99.11	Dihydroorotate dehydrogenase [FMN]
1.4	**A –CH–NH2 group is the donor**
1.4.1.2	Glutamate dehydrogenase
1.4.3.4	Amine oxidase [FAD], “monoamine oxidase (MAO)”
1.4.3.13	Protein lysine 6-oxidase [Cu]
1.4.4.2	Glycine dehydrogenase (decarboxylating) [PLP]
1.5	**A –CH–NH group is the donor**
1.5.1.2	Pyrroline-5-carboxylate reductase
1.5.1.3	Dihydrofolate reductase
1.5.1.5	Methylenetetrahydrofolate dehydrogenase ($NADP^+$)
1.5.1.12	1-Pyrroline-5-carboxylate dehydrogenase
1.5.1.20	Methylenetetrahydrofolate reductase (NADPH) [FAD]
1.5.5.1	Electron-transferring flavoprotein (ETF) dehydrogenase [Fe_4S_4]
1.5.99.8	Proline dehydrogenase [FAD]
1.6	**NAD(P)H is the donor**
16.4.2	Glutathione reductase (NADPH) [FAD]
1.6.4.5	Thioredoxin reductase (NADPH) [FAD]
1.6.5.3	NADH dehydrogenase (ubiquinone) [FAD, Fe_2S_2, Fe_4S_4]—“complex I”
1.8	**A sulfur group is the donor**
1.8.1.4	Dihydrolipoamide dehydrogenase [FAD]
1.9	**A heme group is the donor**
1.9.3.1	Cytochrome *c* oxidase [heme, Cu, Zn] – “cytochrome oxidase,” “complex IV”
1.10	**A diphenol is the donor**
1.10.2.2	Ubiquinol cytochrome *c* reductase [heme, Fe_2S_2]—“complex III”
1.11	**A peroxide is the acceptor** (*peroxidases*)
1.11.1.6	Catalase [heme]
1.11.1.7	Peroxidase [heme]
1.11.1.9	Glutathione peroxidase [Se]
1.11.1.12	Lipid hydroperoxide glutathione peroxidase [Se]
1.13	**Molecular oxygen is incorporated into the electron donor** (*oxygenases*)
1.13.11	**One donor, both O atoms are incorporated** (*dioxygenases*)
1.13.11.5	Homogentisate 1,2-dioxygenase [Fe]
1.13.11.20	Cysteine dioxygenase [Fe]
1.13.11.27	4-Hydroxyphenylpyruvate dioxygenase [ascorbate]
1.13.11.n	Arachidonate lipoxygenases

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

1.14	**Two donors, one O atom is incorporated into both** (*monooxygenases, hydroxylases*)
1.14.11.2	Procollagen proline 4-dioxygenase [Fe, ascorbate]—“proline hydroxylase”
1.14.11.4	Procollagen lysine 5-dioxygenase [Fe, ascorbate]—“lysine hydroxylase”
1.14.13.13	Calcidiol 1-monooxygenase [heme]
1.14.15.4	Steroid 11β-monooxygenase [heme]
1.14.15.6	Cholesterol monooxygenase (side-chain-cleaving) [heme]
1.14.16.1	Phenylalanine 4-monooxygenase [Fe, tetrahydrobiopterin]
1.14.16.2	Tyrosine 3-monooxygenase [Fe, tetrahydrobiopterin]
1.14.17.1	Dopamine β-monooxygenase [Cu]
1.14.99.1	Prostaglandin H-synthase [heme]
1.14.99.3	Heme oxygenase (decyclizing) [heme]
1.14.99.5	Stearoyl-CoA desaturase [heme]
1.14.99.9	Steroid 17α-monooxygenase [heme]
1.14.99.10	Steroid 21-monooxygenase [heme]
1.15	**A superoxide radical is the acceptor**
1.15.1.1	Superoxide dismutase
1.17	**A $-CH_2$ group is the donor**
1.17.4.1	Ribonucleoside diphosphate reductase [Fe]—“ribonucleotide reductase”
1.18	**Reduced ferredoxin is the donor**
1.18.1.2	Ferredoxin-$NADP^+$ reductase [FAD]
1.18.6.1	Nitrogenase [Fe, Mo, Fe_4S_4]

Class 2: Transferases (catalyze the transfer of groups from one molecule to another)

Subclass 2.n: *Which group* is transferred?

2.1	**A C_1 group is transferred**
2.1.1	**A methyl group**
2.1.1.2	Guanidinoacetate *N*-methyltransferase
2.1.1.6	Catechol *O*-methyltransferase
2.1.1.13	5-Methyltetrahydrofolate-homocysteine *S*-methyltransferase
2.1.1.28	Phenylethanolamine *N*-methyltransferase
2.1.1.45	Thymidylate synthase
2.1.1.67	Thiopurine methyltransferase
2.1.2	**A formyl group**
2.1.2.1	Glycine hydroxymethyltransferase [PLP]
2.1.2.2	Phosphoribosylglycinamide formyltransferase
2.1.2.3	Phosphoribosylaminoimidazolecarboxamide formyltransferase
2.1.2.5	Glutamate formiminotransferase [PLP]
2.1.2.10	Aminomethyltransferase
2.1.3	**A carbamoyl group**
2.1.3.2	Aspartate carbamoyltransferase [Zn^{2+}]
2.1.3.3	Ornithine carbamoyltransferase
2.1.4	**An amidino group**
2.1.4.1	Glycine amidinotransferase
2.2	**An aldehyde or ketone residue is transferred**
2.2.1.1	Transketolase [TPP]
2.2.1.2	Transaldolase

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

2.3 An acyl group is transferred

2.3.1 With acyl-CoA as donor

2.3.1.1	Amino acid *N*-acetyltransferase
2.3.1.6	Choline *O*-acetyltransferase
2.3.1.12	Dihydrolipoamide acetyltransferase [lipoamide]
2.3.1.15	Glycerol 3-phosphate *O*-acyltransferase
2.3.1.16	Acetyl-CoA acyltransferase
2.3.1.20	Diacylglycerol *O*-acyltransferase
2.3.1.21	Carnitine *O*-palmitoyltransferase
2.3.1.22	Acylglycerol *O*-palmitoyltransferase
2.3.1.24	Sphingosine *N*-acyltransferase
2.3.1.37	5-Aminolevulinate synthase [PLP]
2.3.1.38	[ACP] *S*-acetyltransferase
2.3.1.39	[ACP] *S*-malonyltransferase
2.3.1.41	3-Oxoacyl-[ACP] synthase
2.3.1.42	Glycerone phosphate *O*-acyltransferase
2.3.1.43	Phosphatidylcholine-sterol acyltransferase—“lecithin-cholesterol acyltransferase (LCAT)”
2.3.1.51	Acylgylcerol-3-phosphate *O*-acyltransferase
2.3.1.61	Dihydrolipoamide succinyltransferase
2.3.1.85	Fatty-acid synthase

2.3.2 An aminoacyl group is transferred

2.3.2.2	γ-glutamyltransferase
2.3.2.12	Peptidyltransferase (*a ribozyme*)
2.3.2.13	Protein-glutamine γ-glutamyltransferase [Ca]—“fibrin-stabilizing factor”

2.4 A glycosyl group is transferred

2.4.1 A hexose residue

2.4.1.1	Phosphorylase [PLP]—“glycogen (starch) phosphorylase”
2.4.1.11	Glycogen (starch) synthase
2.4.1.17	Glucuronosyltransferase
2.4.1.18	1,4-α-Glucan branching enzyme
2.4.1.25	4-α-Glucanotransferase
2.4.1.47	*N*-Acylsphingosine galactosyltransferase
2.4.1.119	Protein glycotransferase

2.4.2 A pentose residue

2.4.2.7	Adenine phosphoribosyltransferase
2.4.2.8	Hypoxanthine phosphoribosyltransferase
2.4.2.10	Orotate phosphoribosyltransferase
2.4.2.14	Amidophosphoribosyl transferase

2.5 An alkyl or aryl group is transferred

2.5.1.1	Dimethylallyltransferase
2.5.1.6	Methionine adenosyltransferase
2.5.1.10	Geranyltransferase
2.5.1.21	Farnesyl diphosphate farnesyltransferase

2.6 A nitrogen-containing group is transferred

2.6.1 An amino group (*transaminases*)

2.6.1.1	Aspartate transaminase [PLP]—“GOT”
2.6.1.2	Alanine transaminase [PLP]—“GPT”

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

2.6.1.3	Cysteine transaminase [PLP]
2.6.1.5	Tyrosine transaminase [PLP]
2.6.1.6	Leucine transaminase (PLP]
2.6.1.11	Acetylornithine transaminase [PLP]
2.6.1.13	Ornithine transaminase [PLP]
2.6.1.19	4-Aminobutyrate transaminase [PLP]
2.6.1.42	Branched-chain amino acid transaminase [PLP]
2.6.1.52	Phosphoserine transaminase [PLP]
2.7	**A phosphorus-containing group is transferred** (*kinases*)
2.7.1	**With –CH–OH as acceptor**
2.7.1.1	Hexokinase
2.7.1.3	Ketohexokinase
2.7.1.6	Galactokinase
2.7.1.11	6-Phosphofructokinase
2.7.1.19	Phosphoribulokinase
2.7.1.28	Triokinase (triosekinase)
2.7.1.30	Glycerol kinase
2.7.1.32	Choline kinase
2.7.1.36	Mevalonate kinase
2.7.1.37	Protein kinase
2.7.1.38	Phosphorylase kinase
2.7.1.39	Homoserine kinase
2.7.1.40	Pyruvate kinase
2.7.1.67	1-Phosphatidylinositol-4-kinase
2.7.1.68	1-Phosphatidylinositol 4-phosphate kinase
2.7.1.82	Ethanolamine kinase
2.7.1.99	[Pyruvate dehydrogenase] kinase
2.7.1.105	6-Phosphofructo-2-kinase
2.7.1.112	Protein tyrosine kinase
2.7.2	**With –CO–OH as acceptor**
2.7.2.3	Phosphoglycerate kinase
2.7.2.4	Aspartate kinase
2.7.2.8	Acetylglutamate kinase
2.7.2.11	Glutamate 5-kinase
2.7.3	**With a nitrogen-containing group as acceptor**
2.7.3.2	Creatine kinase
2.7.4	**With a phosphate group as acceptor**
2.7.4.2	Phosphomevalonate kinase
2.7.4.3	Adenylate kinase
2.7.4.4	Nucleoside phosphate kinase
2.7.4.6	Nucleoside diphosphate kinase
2.7.6	**A diphosphate residue is transferred**
2.7.6.1	Ribose phosphate pyrophosphokinase
2.7.7	**A nucleotide is transferred**
2.7.7.6	DNA-directed RNA polymerase—“RNA polymerase”
2.7.7.7	DNA-directed DNA polymerase—“DNA polymerase”
2.7.7.9	UTP-glucose-l-phosphate uridyltransferase
2.7.7.12	Hexose-1-phosphate uridyltransferase
2.7.7.14	Ethanolamine phosphate cytidyltransferase

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

2.7.7.15	Choline phosphate cytidyltransferase
2.7.7.41	Phosphatidate cytidyltransferase
2.7.7.49	RNA-directed DNA polymerase—“reverse transcriptase”
2.7.8	**Another substituted phosphate is transferred**
2.7.8.1	Ethanolaminephosphotransferase
2.7.8.2	Diacylglycerol cholinephosphotransferase
2.7.8.11	CDPdiacylglycerol-inositol 3-phosphatidyltransferase
2.7.8.16	1-Alkyl-2-acetylglycerol cholinephosphotransferase
2.7.8.17	*N*-Acetylglucosaminephosphotransferase

Class 3: Hydrolases (catalyze bond cleavage by hydrolysis)

Subclass 3.n: *What kind of bond* is hydrolyzed?

3.1	**An ester bond is hydrolyzed** (*esterases*)
3.1.1	**In carboxylic acid esters**
3.1.1.2	Arylesterase
3.1.1.3	Triacylglycerol lipase
3.1.1.4	Phospholipase A_2
3.1.1.7	Acetylcholinesterase
3.1.1.13	Cholesterol esterase
3.1.1.17	Gluconolactonase
3.1.1.32	Phospholipase A_1
3.1.1.34	Lipoprotein lipase, diacylglycerol lipase
3.1.2	**In thioesters** 3.1.2.4
	3-Hydroxyisobutyryl-CoA hydrolase
3.1.2.14	Acyl-[ACP] hydrolase
3.1.3	**In phosphoric acid monoesters** (*phosphatases*)
3.1.3.1	Alkaline phosphatase [Zn^{2+}]
3.1.3.2	Acid phosphatase
3.1.3.4	Phosphatidate phosphatase
3.1.3.9	Glucose 6-phosphatase
3.1.3.11	Fructose bisphosphatase
3.1.3.13	Bisphosphoglycerate phosphatase
3.1.3.16	Phosphoprotein phosphatase
3.1.3.37	Sedoheptulose bisphosphatase
3.1.3.43	[Pyruvate dehydrogenase] phosphatase
3.1.3.46	Fructose-2,6-bisphosphate 2-phosphatase
3.1.3.n	Polynucleotidases
3.1.4	**In phosphoric acid diesters** (*phosphodiesterases*)
3.1.4.1	Phosphodiesterase
3.1.4.3	Phospholipase C
3.1.4.4	Phospholipase D
3.1.4.17	3′,5′-cNMP phosphodiesterase
3.1.4.35	3′,5′-cGMP phosphodiesterase
3.1.4.45	*N*-Acetylglucosaminyl phosphodiesterase
3.1.21	**In DNA**
3.1.21.1	Deoxyribonuclease I
3.1.21.4	Site-specific deoxyribonuclease (type II)—“restriction endonuclease”

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

3.10.26–7	**In RNA**
3.1.26.4	Ribonuclease H
3.1.27.5	Pancreatic ribonuclease
3.2	**A glycosidic bond is hydrolyzed** (*glycosidases*)
3.2.1	**In *O*-glycosides**
3.2.1.1	α-Amylase
3.2.1.10	Oligo-1,6-glucosidase
3.2.1.17	Lysozyme
3.2.1.18	Neuraminidase
3.2.1.20	α-Glucosidase
3.2.1.23	β-Galactosidase
3.2.1.24	α -Mannosidase
3.2.1.26	β-Fructofuranosidase—“saccharase,” “invertase”
3.2.1.28	α,α-Trehalase
3.2.1.33	Amylo-1,6-glucosidase
3.2.1.48	Sucrose α-glucosidase
3.2.1.52	β-*N*-Acetylhexosaminidase
3.2.2.n	Nucleosidases
3.3	**An ether bond is hydrolyzed**
3.3.1.1	Adenosylhomocysteinase
3.4	**A peptide bond is hydrolyzed** (*peptidases*)
3.4.11	**Aminopeptidases** (*N*-terminal exopeptidases)
3.4.11.n	Various aminopeptidases [Zn^{2+}]
3.4.13	**Dipeptidases** (act on dipeptides only)
3.4.13.n	Various dipeptidases [Zn^{2+}]
3.4.15	**Peptidyl dipeptidases** (*C*-terminal exopeptidases, releasing dipeptides)
3.4.15.1	Peptidyl-dipeptidase A [Zn^{2+}]—“angiotensin-converting enzyme (ACE)”
3.4.17	**Carboxypeptidases** (*C*-terminal exopeptidases)
3.4.17.1	Carboxypeptidase A [Zn^{2+}]
3.4.17.2	Carboxypeptidase B [Zn^{2+}]
3.4.17.8	Muramoylpentapeptide carboxypeptidase
3.4.21	**Serine proteinases** (endopeptidases)
3.4.21.1	Chymotrypsin
3.4.21.4	Trypsin
3.4.21.5	Thrombin
3.4.21.6	Coagulation factor Xa—“Stuart–Prower factor”
3.4.21.7	Plasmin
3.4.21.9	Enteropeptidase—“enterokinase”
3.4.21.21	Coagulation factor VIIa—“proconvertin”
3.4.21.22	Coagulation factor IXa—“Christmas factor”
3.4.21.27	Coagulation factor XIa—“plasma thromboplastin antecedent”
3.4.21.34	Plasma kallikrein
3.4.21.35	Tissue kallikrein
3.4.21.36	Elastase
3.4.21.38	Coagulation factor XIIa—“Hageman factor”
3.4.21.43	C3/C5 convertase (complement—classical pathway)
3.4.21.47	C3/C5 convertase (complement—alternative pathway)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

3.4.21.68	Plasminogen activator (tissue)—“tissue plasminogen activator (t-PA)”
3.4.21.73	Plasminogen activator (urine)—“urokinase”

3.4.22	**Cysteine proteinases** (*endopeptidases*)
3.4.22.2	Papain

3.4.23	**Aspartate proteinases** (*endopeptidases*)
3.4.23.1	Pepsin A
3.4.23.2	Pepsin B
3.4.23.3	Gastricsin (pepsin C)
3.4.23.4	Chymosin
3.4.23.15	Renin

3.4.24	**Metalloproteinases** (*endopeptidases*)
3.4.24.7	Collagenase

3.4.99	**Other peptidases**
3.4.99.36	Signal peptidase

3.5	**Another amide bond is hydrolyzed** (*amidases*)
3.5.1.1	Asparaginase
3.5.1.2	Glutaminase
3.5.1.16	Acetylornithine deacetylase [Zn^{2+}]
3.5.2.3	Dihydroorotase
3.5.2.7	Imidazolonepropionase
3.5.3.1	Arginase
3.5.4.6	AMP deaminase
3.5.4.9	Methylenetetrahydrofolate cyclohydrolase
3.5.4.10	IMP cyclohydrolase

3.6	**An anhydride bond is hydrolyzed**
3.6.1.6	Nucleoside diphosphatase
3.6.1.32	Myosin ATPase
3.6.1.34	H$^+$-transporting ATP synthase—“ATP synthase,” “complex V”
3.6.1.35	H$^+$-transporting ATPase
3.6.1.36	H$^+$/K$^+$-exchanging ATPase
3.6.1.37	Na$^+$/K$^+$-exchanging ATPase—“Na$^+$/K$^+$-ATPase”
3.6.1.38	Ca^{2+}-transporting ATPase

3.7	**A C–C bond is hydrolyzed**
3.7.1.2	Fumarylacetoacetase

Class 4: Lyases (cleave or form bonds without oxidative or hydrolytic steps)

Subclass 4.n: *What kind of bond* is formed or cleaved?

4.1 **A C–C bond is formed or cleaved**

4.1.1	**Carboxy-lyases** (*carboxylases, decarboxylases*)
4.1.1.1	Pyruvate decarboxylase [TPP]
4.1.1.15	Glutamate decarboxylase [PLP]
4.1.1.21	Phosphoribosylaminoimidazole carboxylase
4.1.1.23	Orotidine-5′-phosphate decarboxylase
4.1.1.28	Aromatic L-amino acid decarboxylase [PLP]
4.1.1.32	Phosphoenolpyruvate carboxykinase (GTP)
4.1.1.39	Ribulose bisphosphate carboxylase [Cu]—“rubisco”

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

4.1.2	**Acting on aldehydes or ketones**
4.1.2.5	Threonine aldolase [PLP]
4.1.2.13	Fructose bisphosphate aldolase—“aldolase”
4.1.3.4	Hydroxymethylglutaryl-CoA lyase
4.1.3.5	Hydroxymethylglutaryl-CoA synthase
4.1.3.7	Citrate synthase
4.1.3.8	ATP-citrate lyase
4.1.3.18	Acetolactate synthase [TPP, flavin]
4.1.99	**Other C–C lyases**
4.1.99.3	Deoxyribodipyrimidine photolyase [FAD]—“photolyase”
4.2	**A C–O bond is formed or cleaved**
4.2.1	**Hydrolyases** (*hydratases, dehydratases*)
4.2.1.1	Carbonate dehydratase [Zn^{2+}]—“carbonic anhydrase”
4.2.1.2	Fumarate hydratase—“fumarase”
4.2.1.3	Aconitate hydratase [Fe_4S_4]—“aconitase”
4.2.1.11	Phosphopyruvate hydratase—“enolase”
4.2.1.13	Serine dehydratase
4.2.1.17	Enoyl-CoA hydratase
4.2.1.18	Methylglutaconyl-CoA hydratase
4.2.1.22	Cystathionine β-synthase [PLP]
4.2.1.24	Porphobilinogen synthase
4.2.1.49	Urocanate hydratase
4.2.1.61	3-Hydroxypalmitoyl-[ACP] dehydratase
4.2.1.75	Uroporphyrinogen III synthase
4.2.99	**Other C–O lyases**
4.2.99.2	Threonine synthase [PLP]
4.3	**A C–N bond is formed or cleaved**
4.3.1	**Ammonia lyases** 4.3.1.3
	Histidine ammonia lyase
4.3.1.8	Hydroxymethylbilane synthase
4.3.2	**Amidine lyases** 4.3.2.1
	Argininosuccinate lyase
4.3.2.2	Adenylosuccinate lyase
4.4	**A C–S bond is formed or cleaved**
4.4.1.1	Cystathionine γ-lyase [PLP]
4.6	**A P–O bond is formed or cleaved**
4.6.1.1	Adenylate cyclase
4.6.1.2	Guanylate cyclase

Class 5: Isomerases (catalyze changes within one molecule)

Subclass 5.n: *What kind of isomerization* is taking place?

5.1	**A racemization or epimerization** (*epimerases*)
5.1.3.1	Ribulose phosphate 3-epimerase
5.1.3.2	UDPglucose 4-epimerase
5.1.3.4	L-Ribulose phosphate 4-epimerase
5.1.99.1	Methylmalonyl-CoA epimerase

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

5.2	**A *cis–trans* isomerization**
5.2.1.2	Maleylacetoacetate isomerase
5.2.1.3	Retinal isomerase
5.2.1.8	Peptidyl proline *cis–trans*-isomerase

5.3	**An intramolecular electron transfer**
5.3.1.1	Triose phosphate isomerase
5.3.1.6	Ribose 5-phosphate isomerase
5.3.1.9	Glucose 6-phosphate isomerase
5.3.3.1	Steroid Δ-isomerase
5.3.3.8	Enoyl-CoA isomerase
5.3.4.1	Protein disulfide isomerase

5.4	**An intramolecular group transfer** (*mutases*)
5.4.2.1	Phosphoglycerate mutase
5.4.2.2	Phosphoglucomutase
5.4.2.4	Bisphosphoglycerate mutase
5.4.99.2	Methylmalonyl-CoA mutase [cobamide]

5.99	**Another kind of isomerization**
5.99.1.2	DNA topoisomerase (type I)—“DNA helicase”
5.99.1.3	DNA topoisomerase (ATP-hydrolyzing, type II)—“DNA gyrase”

Class 6: Ligases (join two molecules with hydrolysis of an “energy-rich” bond)

Subclass 6.n: *What kind of bond* is formed?

6.1	**A C–O bond is formed**
6.1.1.n	(Amino acid)-tRNA ligases (*aminoacyl-tRNA synthetases*)

6.2	**A C–S bond is formed**
6.2.1.1	Acetate-CoA ligase
6.2.1.3	Long-chain fatty-acid-CoA ligase
6.2.1.4	Succinate-CoA ligase (GDP-forming)—“thiokinase”

6.3	**A C–N bond is formed**
6.3.1.2	Glutamate-NH_3 ligase—“glutamine synthetase”
6.3.2.6	Phosphoribosylaminoimidazolesuccinocarboxamide synthase *(sorry!)*
6.3.3.1	Phosphoribosylformylglycinamidine cycloligase
6.3.3.2	5-Formyltetrahydrofolate cycloligase
6.3.4.2	CTP synthase
6.3.4.4	Adenylosuccinate synthase
6.3.4.5	Argininosuccinate synthase
6.3.4.13	Phosphoribosylamine glycine ligase
6.3.4.16	Carbamoylphosphate synthase (NH_3)
6.3.5.2	GMP synthase (glutamine-hydrolyzing)
6.3.5.3	Phosphoribosylformylglycinamidine synthase
6.3.5.4	Asparagine synthase (glutamine-hydrolyzing)
6.3.5.5	Carbamoylphosphate synthase (glutamine-hydrolyzing)

6.4	**A C–C bond is formed**
6.4.1.1	Pyruvate carboxylase [biotin]
6.4.1.2	Acetyl-CoA carboxylase [biotin]
6.4.1.3	Propionyl-CoA carboxylase [biotin]
6.4.1.4	Methylcrotonyl-CoA carboxylase [biotin]

6.5	**A P–O bond is formed**
6.5.1.1	DNA ligase (ATP)

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Abbreviations

Abbreviations for amino acids, p. 60
For bases and nucleosides, p. 80
For monosaccharides, p. 38

AA	Amino acid
ACE	Angiotensin-converting enzyme (peptidyl-dipeptidase A)
ACP	Acyl carrier protein
ACTH	Adrenocorticotropic hormone (corticotropin)
ADH	Antidiuretic hormone (adiuretin, vasopressin)
ADP	Adenosine 5′-diphosphate
AIDS	Acquired immunodeficiency syndrome
ALA	5-Aminolevulinic acid
AMP	Adenosine 5′-monophosphate
ANF	Atrial natriuretic factor
ANP	Atrial natriuretic peptide (= ANF)
ATP	Adenosine 5′-triphosphate
AVP	Arginine vasopressin
b	Base
bp	Base pair
BPG	2,3-Bisphosphoglycerate
cAMP	3′,5′-Cyclic AMP
CAP	Catabolite activator protein
CDK	Cyclin-dependent protein kinase (in cell cycle)
cDNA	Complementary DNA
CDP	Cytidine 5′-diphosphate
cGMP	3′,5′-Cyclic GMP
CIA	Chemoluminescence immunoassay
CMP	Cytidine 5′-monophosphate
CoA	Coenzyme A
CoQ	Coenzyme Q (ubiquinone)
CSF	colony-stimulating factor
CTP	Cytidine 5′-triphosphate
d	Deoxy-
Da	Dalton (atomic mass unit)
DAG	Diacylglycerol
dd	Dideoxy-
DH	Dehydrogenase
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
EA	Ethanolamine
EIA	Enzyme-linked immunoassay
ER	Endoplasmic reticulum
FAD	Flavin adenine dinucleotide
Fd	Ferredoxin
FFA	Free fatty acid
fMet	*N*-formylmethionine
FMN	Flavin mononucleotide
Fp	Flavoprotein (containing FMN or FAD)
GABA	γ-Aminobutyric acid
GDP	Guanosine 5′-diphosphate
Glut	Glucose transporter
GMP	Guanosine 5′-monophosphate
GSH	Reduced glutathione
GSSG	Oxidized glutathione
GTP	Guanosine 5′-triphosphate
h	hour
HAT medium	Medium containing hypoxanthine, aminopterin, and thymidine
Hb	Hemoglobin
HDL	High-density lipoprotein
HIV	Human immunodeficiency virus
HLA	Human leukocyte-associated antigen
HMG-CoA	3-Hydroxy-3-methylglutaryl-CoA
HMP	Hexose monophosphate pathway
hnRNA	Heterogeneous nuclear ribonucleic acid
HPLC	High-performance liquid chromatography
hsp	Heat-shock protein
IDL	Intermediate-density lipoprotein
IF	Intermediary filament
IFN	Interferon
Ig	Immunoglobulin
IL	Interleukin
$InsP_3$	Inositol 1,4,5-trisphosphate
IPTG	Isopropylthiogalactoside
IRS	Insulin-receptor substrate
kDa	Kilodalton (10^3 atomic mass units)
K_m	Michaelis constant

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

LDH	Lactate dehydrogenase
LDL	Low-density lipoprotein
M	Molarity (mol L^{-1})
Mab	Monoclonal antibody
MAP kinase	Mitogen-activated protein kinase
MHC	Major histocompatability complex
MPF	Maturation-promoting factor
mRNA	Messenger ribonucleic acid
N	Nucleotide with any base
NAD^+	Oxidized nicotinamide adenine dinucleotide
NADH	Reduced nicotinamide adenine dinucleotide
$NADP^+$	Oxidized nicotinamide adenine dinucleotide phosphate
NADPH	Reduced nicotinamide adenine dinucleotide phosphate
NeuAc	*N*-acetylneuraminic acid
nm	Nanometer (10^{-9} m)
ODH	2-Oxoglutarate dehydrogenase
PAGE	Polyacrylamide gel electrophoresis
Pan	Pantetheine
PAPS	Phosphoadenosine phosphosulfate
PCR	Polymerase chain reaction
PDH	Pyruvate dehydrogenase
PEG	Polyethylene glycol
PEP	Phosphoenolpyruvate
pH	pH value
P_i	Inorganic phosphate
PK	Protein kinase
PLP	Pyridoxal phosphate
PP	Protein phosphatase
PPP	Pentose phosphate pathway
PQ	Plastoquinone
PRPP	5-Phosphoribosyl 1-diphosphate
PS	Photosystem
PTH	Parathyroid hormone
Q	Oxidized coenzyme Q (ubiquinone)
QH_2	Reduced coenzyme Q (ubiquinol)
R	Gas constant
rER	Rough endoplasmic reticulum
RES	Reticuloendothelial system
RFLP	Restriction fragment length polymorphism
RIA	Radioimmunoassay
RNA	Ribonucleic acid
ROS	Reactive oxygen species
RP	Reversed phase (of silica gel)
rRNA	Ribosomal ribonucleic acid
S	Svedberg (unit of sedimentation coefficient)
SAH	*S*-adenosyl L-homocysteine
SAM	*S*-adenosyl L-methionine
SDS	Sodium dodecylsulfate
sER	Smooth endoplasmic reticulum
sn	Stereospecific numbering
snRNA	Small nuclear ribonucleic acid
SR	Sarcoplasmic reticulum
ssDNA	Single-stranded DNA
TBG	Thyroxine-binding globulin
THB	Tetrahydrobiopterin
THF	Tetrahydrofolate
TLC	Thin-layer chromatography
TPP	Thiamine diphosphate
TRH	Thyrotropin-releasing hormone (thyroliberin)
Tris	Tris(hydroxymethyl)aminomethane
tRNA	Transfer ribonucleic acid
TSH	Thyroid-stimulating hormone (thyrotropin)
UDP	Uridine 5′-diphosphate
UMP	Uridine 5′-monophosphate
UTP	Uridine 5′-triphosphate
UV	Ultraviolet radiation
V_{max}, V	Maximal velocity (of an enzyme)
VLDL	Very-low-density lipoprotein

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

Quantities and units

1. SI base units

Quantity	SI unit	Symbol	Remarks
Length	Meter	m	1 yard (yd) = 0.9144 m 1 inch (in) = 0.0254 m 1 Å = 10^{-10} m = 0.1 nm
Mass	Kilogram	kg	1 pound (lb) = 0.4536 kg
Time	Second	s	
Current strength	Ampere	A	
Temperature	Kelvin	K	°C (degree Celsius) = K – 273.2 Fahrenheit: °C = 5/9 (°F – 32)
Light	Candela	Cd	
Amount of substance	Mol	mol	

2 Derived units

Quantity	Unit	Symbol	Derivation	Remarks
Frequency	Hertz	Hz	s^{-1}	
Volume	Liter	L	$10^{-3}\ \ m^3$	1 U.S. gallon (gal) = 3.785 L
Force	Newton	N	$kg\ \ m\ \ s^{-2}$	
Pressure	Pascal	Pa	$N\ \ m^{-2}$	1 bar = 10^5 Pa 1 mmHg = 133.3 Pa
Energy, work, heat	Joule	J	N m	1 calorie (cal) = 4.1868 J
Power	Watt	W	$J\ \ s^{-1}$	
Electrical charge	Coulomb	C	A s	
Voltage	Volt	V	$W\ \ A^{-1}$	
Concentration	Molarity	M	$mol\ \ L^{-1}$	
Molecular mass	Dalton	Da	$1.6605\ \ 10^{-24}$ g	
Molar mass	–	–	g	
Molecular weight	–	M_r	–	Nondimensional
Reaction rate		v	$mol\ \ s^{-1}$	
Catalytic activity	Katal	kat	$mol\ \ s^{-1}$	1 unit (U) = $1.67\ \ 10^{-8}$ kat
Specific activity	–	–	kat (kg enzyme)$^{-1}$	Usually: U (mg enzyme)$^{-1}$
Sedimentation coefficient	Svedberg	S	10^{-13} s	
Radioactivity	Becquerel	Bq	Decays s^{-1}	1 curie (Ci) = $3.7\ \ 10^{10}$ Bq

3 Multiples and fractions

Factor	Prefix	Symbol	Example
10^9	Giga	G	GHz = 10^9 hertz
10^6	Mega	M	MPa = 10^6 pascal
10^3	Kilo	k	kJ = 10^3 joule
10^{-3}	Milli	m	mM = 10–3 mol L^{-1}
10^{-6}	Micro	µ	µV = 10^{-6} volt
10^{-9}	Nano	n	nkat = 10^{-9} katal
10^{-12}	Pico	p	pm = 10^{-12} meter

4 Important constants

General gas constant, R	R = 8.314 J mol^{-1} K^{-1}
Loschmidt (Avogadro) number, N (number of particles per mol)	N = $6.0225\ \ 10^{23}$
Faraday constant F	F = 96480 C mol^{-1}

Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thieme
All rights reserved. Usage subject to terms and conditions of license.

LANGE

MICHAEL J. PARMELY

USMLE ROAD MAP

IMMUNOLOGY

- HIGH-YIELD FACTS
- ILLUSTRATIONS
- CLINICAL PROBLEMS
- CLINICAL CORRELATIONS

The Best Route to Step 1 Success

CONTENTS

Using the Road Map Series for Successful Review vii

Acknowledgments viii

MECHANISMS AND CONSEQUENCES OF IMMUNE RECOGNITION

1. Innate Immunity 1

2. Adaptive Immunity 14

3. Antigens and Antibodies 28

4. Immunoglobulin Gene Expression 40

5. Antigen Recognition by Antibody 52

6. T Cell Recognition of and Response to Antigen 64

7. Major Histocompatibility Complex 77

DEVELOPMENT OF IMMUNE EFFECTOR MECHANISMS

8. Complement 91

9. B Cell Differentiation and Function 104

10. T Cell Differentiation and Function 118

11. Regulation of Immune Responses 130

12. Cytokines 137

IMMUNITY IN HEALTH AND DISEASE

13. Immune Tissue Injury 149

14. Protective Immunity and Vaccines 164

15. Immune Deficiency States 175

16. Autotolerance and Autoimmunity 192

17. Transplantation 204

Appendices 217

Index 219

CHAPTER 1

INNATE IMMUNITY

I. Immunity is distinguished by the following features.

A. The immune system, which protects the body against microbial invaders and environmental agents, takes two forms.

1. **Innate immunity** is available at birth and protects the newborn from pathogenic microbes.
2. **Adaptive or acquired immunity** arises in the host as a consequence of exposure to a microbe or foreign substance.

B. The life-style of the microbe determines the nature of the protective immune response.

1. **Extracellular microbes** can be neutralized by antibodies and other soluble immune mediators.
2. Elimination of **intracellular pathogens** requires their recognition by immune cells that can destroy pathogen-infected host cells.

C. Both forms of immunity require a specific recognition of the pathogen or environmental agent and an ability to distinguish it from "self."

D. Innate immunity is a phylogenetically ancient defense mechanism designed for rapidly recognizing, lysing, or phagocytozing pathogenic microbes and signaling their presence to the host.

1. The innate immune system recognizes **microbial patterns** that are widely distributed across genera, rather than the discrete antigenic determinants that characterize a particular species of microbe (Chapter 2).
2. Innate immunity does not require prior exposure to the offending agent and is not altered by a previous encounter with it.
3. Innate immunity is expressed within minutes to hours, representing the first response of the host to microbial pathogens.

E. The **effector mechanisms** used by the innate immune system to eliminate foreign invaders (eg, phagocytosis) are often the same as those used for immune elimination during an adaptive immune response (Chapter 2).

F. Many of the responses we consider to be part of the innate immune system also play a central role in **inflammatory responses** to tissue injury (Table 1–1).

II. First lines of defense limit microbial survival.

A. **Physical and chemical barriers** provide some of the first lines of innate defense by preventing microbial attachment, entry, or local tissue survival in a nonspecific manner.

Table I–1. Components shared by the innate immune and inflammatory systems.

Component	Innate Immune Effects	Inflammatory Effects
Phagocytic leukocytes	Intracellular killing of microbes	Elimination of damaged host cells
Complement system	Chemoattraction of leukocytes Lysis, opsonophagocytosis and clearance of microbes	Chemoattraction of leukocytes Increased vascular permeability
Fibrinolysis system	Complement activation Leukocyte chemotaxis	Increased vascular permeability Leukocyte chemotaxis
Vascular endothelium	Delivery of immune mediators to sites of infection	Delivery of inflammatory mediators to sites of damaged tissues
Cytokines	Danger signaling Phagocyte activation Respiratory burst Fever	Leukocyte adhesion, chemotaxis, and uptake of cellular debris Phagocyte activation Fever Tissue healing
Neutrophil granules	Antimicrobial cationic peptides	Extracellular matrix degradation

1. The **epithelium** of the skin and mucous membranes provides a physical barrier.
2. The **mucocilliary movement** of the lung epithelium and the **peristalsis** of the gastrointestinal tract move microbes and other foreign agents across mucosal surfaces and out of the body.
3. The **low pH** and high fatty acid content of the skin inhibit microbial growth.
4. The low pH of the stomach damages essential structures of microbes and limits their survival.
5. **Mucins** associated with mucosal epithelia prevent microbial penetration and bind soluble immune factors (eg, antibody molecules).
6. A variety of **iron-binding proteins** (eg, lactoferrin) compete with microbes for extracellular iron.
   a. Lactoferrin competes for iron in the extracellular space.
   b. The ***Nramp1*** gene product enables host cells to acquire the Fe^{2+} ions necessary to generate reactive oxygen species.
   c. ***Nramp2*** aids cells in depleting Fe^{2+} from the phagosome, thus inhibiting microbial survival.

B. The **normal flora** found at epithelial surfaces provides a **biological barrier** to pathogenic microbes that attempt to survive at that site.
1. Normal microbial flora competes with pathogens for nutrients and environmental niches, especially at external body surfaces, such as the skin, intestines, and lungs.
2. Normal flora can induce innate immune responses in the epithelium that limit the survival of pathogenic microorganisms.

III. Pathogens that breach the primary barriers initiate an innate immune response.

A. **Pathogen-associated molecular patterns** (PAMP) are recognized by innate immune cells and soluble mediators.

1. PAMP are often highly charged surface structures or unique spatial arrangements of chemical groups (eg, sugar moieties) that are not seen on host tissues.
2. PAMP are phylogenetically conserved structures that are essential for the survival of microorganisms.
3. Host cell receptors capable of recognizing PAMP are encoded within the germline and are phylogenetically conserved.
    - a. Relatively few host cell surface receptors are required to recognize a wide range of pathogens.
    - b. The **Toll-like receptor** (TLR) family is an important example of phylogenetically conserved PAMP-specific host molecules (Table 1–2).
4. A number of soluble host proteins also recognize PAMP.
    - a. **Mannose-binding protein** (MBP) (also called **mannose-binding lectin**) binds to mannose residues of a particular spacing that is seen on microbial, but not mammalian, cells.
        - *(1)* MBP serves as an opsonin promoting phagocytosis.
        - *(2)* MBP promotes lysis and phagocytosis of microbes by activating complement (Chapter 8).
    - b. **Lysozyme** degrades the peptidoglycan layer of bacterial cell walls.

Table 1–2. The Toll-like receptor (TLR) family.

TLR	Microbial Ligands
TLR1	Bacterial lipopeptides
TLR2	Bacterial peptidoglycan, lipoteichoic acid, lipoarabinomannan, glycolipids, porins
TLR3	Viral double-stranded RNA
TLR4	Bacterial lipopolysaccharide, viral proteins
TLR5	Bacterial flagellin
TLR6	Bacterial lipopeptides; fungal cell wall
TLR7	Viral single-stranded RNA
TLR8	Viral single-stranded RNA
TLR9	Bacterial CpG-containing DNA

B. The recognition of PAMP activates leukocyte functions.

- **1.** Phagocytic **leukocytes** (blood neutrophils and tissue macrophages) can recognize microbes directly through their **mannose receptors, scavenger receptors, Toll-like receptors**, or **chemotactic receptors**.
  - **a.** The recognition of microbial **chemotactic factors** directs leukocytes to the site of infection.
    - *(1)* Chemotactic factors can be either microbial or host in origin (Table 1–3).
    - *(2)* Chemotactic factors are recognized by seven transmembrane G protein-coupled **receptors**.
  - **b.** **Opsonic receptors** on leukocytes recognize host components that have bound to the surface of microbes.
  - **c.** Attachment of a microbe to the surface of a phagocyte is followed by its **uptake** by membrane invagination (Figure 1–1).
    - *(1)* The microbe is ingested into a **phagosome**.
    - *(2)* The **phagosome** fuses with an organelle called the **lysosome** to form a **phagolysosome**.
  - **d.** Intracellular killing of the microbe occurs within the phagolysosome.
    - *(1)* Lysosomal **hydrolytic enzymes** (acidic proteases, lipases, and nucleases) degrade microbial structures.
    - *(2)* Leukocyte cytoplasmic granules containing cationic antimicrobial peptides (**defensins** and **cathelicidins**) fuse with the phagolysosome.
      - *(a)* These peptides act as disinfectants by disrupting the membrane functions of microorganisms.
      - *(b)* Defensins recognize the highly charged phospholipids on the outer membranes of microbes.
      - *(c)* Antimicrobial peptides of very similar structure have been found both in the vernix caseosa covering the skin of newborn humans and the skin secretions of frogs.

Table 1–3. Chemotactic factors that attract innate immune cells.

Cell Type	Chemotactic Factors
Neutrophil	Bacterial lipoteichoic acid Bacterial formyl-methionyl peptides Complement peptide C5a Fibrinogen-derived peptides Leukotriene B_4 Mast cell-derived chemotactic peptide NCF-A Cytokines: interleukin-8
Macrophage	Cytokines: transforming growth factor-β, monocyte chemotactic protein-1
Lymphocyte	Cytokines: macrophage inflammatory protein-1

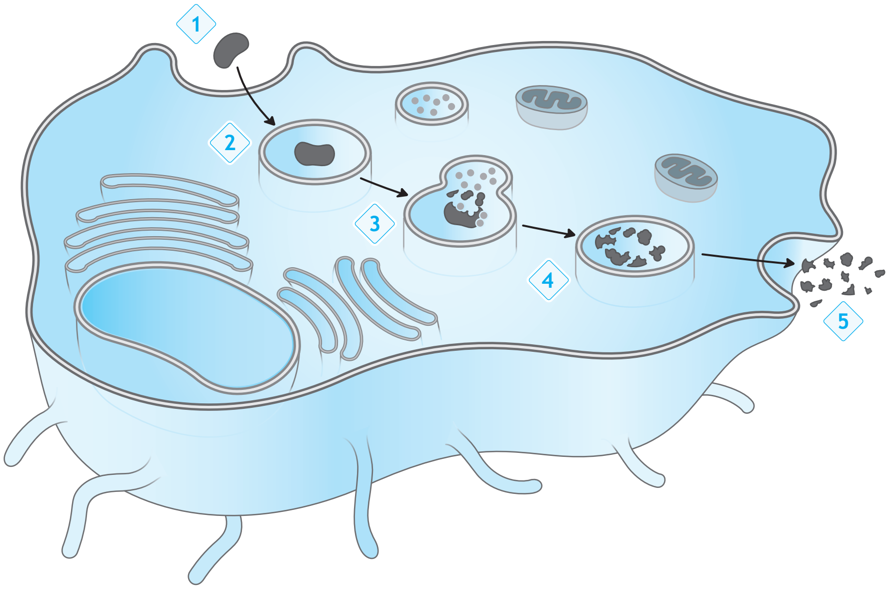

Figure 1–1. Opsonophagocytosis and intracellular killing of a pathogen by a phagocytic cell. 1, Attachment; 2, ingestion (phagosome); 3, phagolysosome; 4, killing, digestion; 5, release.

(3) In the presence of adequate oxygen, microbe recognition at the phagocytic cell surface can initiate a **respiratory burst**, the one electron reduction of molecular oxygen (Figure 1–2).

- *(a)* **Reactive oxygen intermediates** (oxidants and radicals) produced during this process irreversibly damage essential microbial structures.
- *(b)* The reaction begins with the **respiratory burst oxidase**, a multicomponent membrane-associated enzyme.
- *(c)* This oxidase catalyzes the reduction of oxygen (O_2) to the radical **superoxide** ($O_2^{\bullet}$).
- *(d)* The dismutation of superoxide to form **hydrogen peroxide** (H_2O_2) is catalyzed by the enzyme **superoxide dismutase** (SOD).
- *(e)* In the presence of a halide (eg, chloride ion), neutrophil-specific **myeloperoxidase** catalyzes the production of hypohalite (eg, **hypochlorite** or bleach) and organic chloramines.
- *(f)* In the presence of ferric ion, the highly reactive **hydroxyl radical** ($OH^{\bullet}$) is formed from superoxide and hydrogen peroxide.

CHRONIC GRANULOMATOUS DISEASE (CGD) IS A MUTATION OF THE RESPIRATORY BURST OXIDASE

CLINICAL CORRELATION

- *Mutations in the subunits of the **respiratory burst oxidase** (also called NADPH oxidase) can lead to a decreased production of the superoxide radical by phagocytes.*

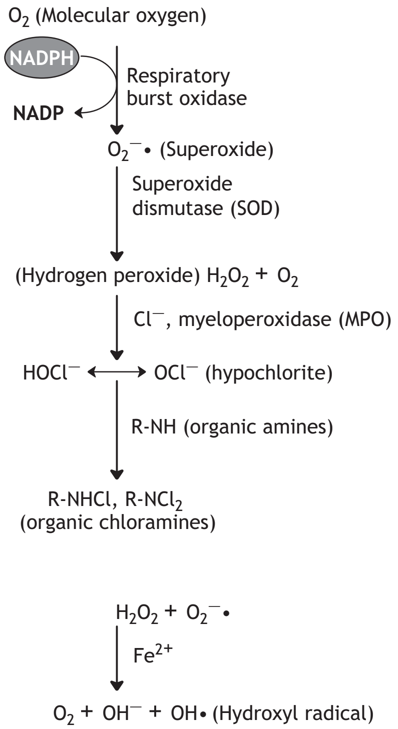
O_2 (Molecular oxygen)
NADPH
NADP
Respiratory burst oxidase
$O_2^{-}\bullet$ (Superoxide)
Superoxide dismutase (SOD)
(Hydrogen peroxide) $H_2O_2 + O_2$
Cl^-, myeloperoxidase (MPO)
$HOCl^-$
OCl^- (hypochlorite)
R-NH (organic amines)
R-NHCl, R-NCl_2 (organic chloramines)
$H_2O_2 + O_2^{-}\bullet$
Fe^{2+}
$O_2 + OH^- + OH\bullet$ (Hydroxyl radical)

Figure I–2. The respiratory burst and reactive oxygen intermediates.

- *Leukocytes of CGD patients fail to produce many of the oxidants that mediate killing of microorganisms within the phagolysosome.*
- *CGD patients are at risk for acquiring* ***opportunistic infections*** *with microbes that would otherwise show low virulence in normal individuals.*
- *Because the phagocytosis of microbes is normal in these patients, some pathogens that are not killed replicate within the phagolysosome.*
- *The host attempts to wall off leukocytes containing viable microbes by forming a structure called a* ***granuloma*** *in the lungs and liver.*

(4) Oxygen-independent intracellular killing is essential when tissue oxygen is limited, as in deep tissue abscesses.

(5) Some phagocytic cells (eg, tissue macrophages) produce the radical **nitric oxide (NO$^\bullet$)**, which can damage microbial structures.

  (a) NO$^\bullet$ is formed from L-arginine and oxygen through a reaction catalyzed by **nitric oxide synthase (NOS)**:

$$\text{L-arginine-NH}_2 + \text{NADPH} + O_2 \rightarrow \text{NO}^\bullet + \text{L-citruline} + \text{NADP}$$

  (b) In macrophages and hepatocytes, the **inducible form of NOS (iNOS)** catalyzes high level, sustained production of NO$^\bullet$ that functions as an antimicrobial agent.

  (c) Only a few microbes (eg, *Mycobacteria* and *Listeria* species) are highly susceptible to NO$^\bullet$.

        (d) When $NO^{\bullet}$ and $O_2^{\bullet}$ combine, they form **peroxynitrite** ($ONOO^-$), an especially potent oxidant.

2. **Epithelial cells also produce the defensins.**
    - a. Defensins limit microbial survival at the mucosal surface of the lung, intestine, and genitourinary tract.
    - b. Defensins are chemotactic for dendritic cells, monocytes, and T lymphocytes that mediate mucosal defense.
3. **Intraepithelial T lymphocytes** are found in the skin, lung, and small intestine.
    - a. These cells bear germline gene-encoded antigen receptors (Chapter 6) that recognize conserved microbial glycolipids.
    - b. Intraepithelial T cells mediate host protection by the secretion of cytokines that can activate phagocytic cells.
4. **Natural killer (NK) cells** recognize host cells that are infected with intracellular pathogens, such as viruses.
    - a. NK cells bear two types of receptors, one for activating the cell and another for inhibiting its activation.
        - *(1)* NK cell **activating receptors** are specific for host and microbial ligands.
        - *(2)* NK cell **inhibitory receptors** are specific for **major histocompatibility complex (MHC)** molecules that are widely distributed on host tissues (Chapter 7).
        - *(3)* When the inhibitory receptor binds host MHC molecules, activation of the NK cell is blocked.
        - *(4)* When the expression of MHC molecules is decreased on host tissues, NK cells become activated through their activating receptors.
        - *(5)* The expression of MHC is often decreased on virus-infected cells.
    - b. Upon activation, NK cells can eliminate microbial pathogens by secreting cytokines, which activate macrophages.
    - c. NK cells can also lyse infected host cells.
    - d. NK cells also synthesize **interferons** (Chapter 12) that block the replication of viruses within infected cells.
5. **Natural killer T (NKT) cells** bear many of the surface receptors present on NK cells as well as an unconventional form of the **T cell antigen receptor** (Chapter 6).
    - a. Most NKT cells are specific for microbial glycolipids.
    - b. NKT cells can produce cytokines capable of activating macrophages.
    - c. NKT cells can express cytotoxic activity, although the role of this function in host defense is still unclear.

C. The recognition of microbial pathogens signals “danger” to the host.

1. **TLRs** (Table 1–2) initiate danger signaling when they bind microbial PAMP.
    - a. Intracellular signal transduction initiated by TLR leads to the activation of transcription factors.
    - b. For example, **TLR4** mediates the recognition of bacterial **lipopolysaccharides (LPS)**, which are common components of the outer membrane of gram-negative bacteria (Figure 1–3)
    - c. TLR4 signaling results in the activation of the **nuclear factor-κB (NFκB)** and **AP-1** transcription factors.
    - d. Among the genes regulated by NFκB and AP-1 are those encoding proinflammatory cytokines and their receptors, cell adhesion molecules, immunoglobulins, and antigen receptors.

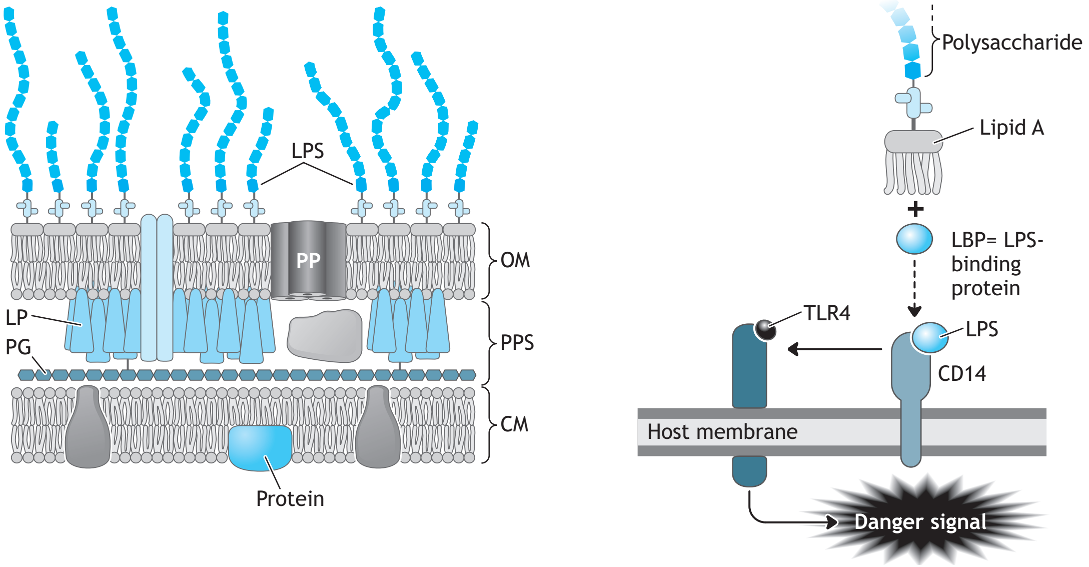

Figure 1–3. Cellular responses to bacterial lipopolysaccharide (LPS) are mediated by toll-like receptor 4 (TLR4). CM, cytoplasmic membrane; LP, lipoprotein; LPS, lipopolysaccharide; OM, outer membrane; PG, peptidoglycan; PP, porin protein; PPS, periplasmic space.

EXCESSIVE DANGER SIGNALING AND SEPSIS

CLINICAL CORRELATION

- ***Sepsis*** *is a systemic host response to disseminated infection characterized by fever, tachycardia, tachypnea, hemodynamic dysfunction, coagulopathy, and multiorgan damage.*
- *These processes result from microvascular changes, diminished tissue perfusion, and inadequate tissue oxygenation.*
- *Sepsis represents excessive danger signaling on the part of the host; soluble and cellular mediators of innate immunity are produced in excess.*
- *The cytokines interleukin (IL)-1, interferon (IFN)-γ, and tumor necrosis factor (TNF)-α are important early mediators of sepsis.*
- *Clinical trials using reagents (eg, antibodies) designed to neutralize any one of these mediators have been disappointing, probably owing to mediator redundancy.*

2. Cytokine genes are induced by danger signaling and are essential for appropriate innate immune responses to infection.
   a. Cytokines are peptide hormone-like mediators of immunity and inflammation (Chapter 12).
   b. Cytokines are produced by a variety of immune cells and induce gene expression, cell growth, and differentiation.
   c. Cytokines act through specific **cytokine receptors**, many of which activate gene transcription.
   d. Among the important effects of cytokines are fever, hematopoiesis, chemotaxis, increased cell adhesion, changes in blood vessel function, antibody production, and apoptosis (Table 1–4).

Table 1–4. Cytokines that act as danger signals.[a]

Cytokine	Functions Related to Danger Signaling
TNF-α	Fever, leukocyte adhesion to endothelium, acute phase protein synthesis, respiratory burst, cachexia, cardiac suppression, disseminated intravascular coagulation and shock
IL-1, IL-6	Fever, leukocyte adhesion to endothelium, acute phase protein synthesis, B lymphocyte coactivation
Chemokines	Lymphocyte and leukocyte migration to sites of infection
IL-4	Lymphocyte coactivation and antibody production
IL-12	Lymphocyte coactivation and cell-mediated immunity
IFN-α, IFN-β	Antiviral state, coactivation of macrophages and NK cells, increased MHC expression
IFN-γ	Coactivation of macrophages, increased MHC expression

[a] TNF, tumor necrosis factor; IL, interleukin; IFN, interferon; NK, natural killer; MHC, major histocompatibility complex.

e. The **interferons (IFN)** are a family of cytokines first noted for their antiviral activity.
(1) IFN-α and IFN-β block virus replication within cells.
(2) IFN-γ is a potent activator of macrophages for the killing of intracellular bacteria and fungi.

DEFECTIVE IFN-γ RECEPTOR FUNCTION LEADS TO OPPORTUNISTIC INFECTIONS

CLINICAL CORRELATION

- *The killing of intracellular microbial pathogens by macrophages requires that the cells be activated by microbial or host signals, including cytokines.*
- *IFN-γ is a potent macrophage activating cytokine that acts on cells through its receptor.*
- *Point mutations in the human IFN-γ receptor 1 gene impair signaling and macrophage activation for the killing of intracellular pathogens.*
- *Life-threatening infections with Mycobacterium and Salmonella species are common and can become widely disseminated throughout the body.*
- *Because the receptors for IFN-α and IFN-β are distinct from those that bind IFN-γ, the affected individuals do not suffer from increased viral infections.*

3. PAMP can activate the **complement** system of serum proteins.
a. Several complement components can recognize highly charged microbial structures, such as bacterial LPS and surface mannose residues.

b. Peptides produced during complement activation mediate host defense and inflammatory functions, such as the chemotaxis of neutrophils, opsonization, and the lysis of microbial membranes.

c. The activation of mast cell degranulation by complement peptides, called **anaphylatoxins**, leads to the release of an additional wave of inflammatory mediators that are stored in mast cell cytoplasmic granules (Chapter 13).

4. The synthesis of **acute phase proteins** is a response to danger signaling.

a. Many acute phase proteins are produced in the liver in response to the cytokines IL-1, IL-6, and TNF-α.

b. **C-reactive protein (CRP)** binds to bacterial surface phospholipids, activates complement, and serves as an opsonin.

c. Increased **fibrinogen** in plasma increases the **erythrocyte sedimentation rate (ESR)**, a clinical laboratory test indicative of acute inflammation.

5. The **coagulation** and **fibrinolysis** systems are activated during acute infections and inflammation.

a. Coagulation serves to localize infection by retaining microbes within a fibrin clot.

b. Peptides derived from fibrinogen during fibrinolysis are chemotactic for neutrophils.

c. **Plasmin** generated during fibrinolysis can activate the complement system.

DYSREGULATION OF THE COMPLEMENT SYSTEM RESULTS IN ACUTE INFLAMMATION

CLINICAL CORRELATION

- *Unabated activation of the complement system is potentially harmful to the host due to the production of inflammatory mediators.*
- *An important regulator of the classic pathway of complement activation is the protease inhibitor **C1 inhibitor (C1 Inh)**.*
- *Patients with **hereditary angioedema (HAE)** have significantly decreased levels of plasma C1 Inh.*
- *Episodic activation of complement in HAE patients results in the production of complement peptides that increase vascular permeability.*
- *The resulting subcutaneous and submucosal edema can lead to airway obstruction, asphyxiation, and severe abdominal pain.*

IV. Danger signals can promote the activation of antigen-specific T and B lymphocytes of the adaptive immune system.

A. The nature of danger signaling depends on the type of microbe.

1. Intracellular pathogens often induce innate signals (eg, IL-12) that promote the development of cellular immunity.

2. Extracellular pathogens often favor induction of antibody responses to microbial antigens (Chapter 2).

B. Danger signals activate T and B lymphocytes through their cell surface **coreceptors**.

1. The complement peptide C3d generated during an innate immune response is the ligand for the **B cell coreceptor CR2**.

2. C3d costimulates B cells that have bound antigen through their antigen receptors (Chapter 8).

C. Cytokines produced by innate immune cells are important regulators of lymphocyte activation during adaptive immune responses.
  1. IFN-α and IFN-β enhance T lymphocyte responses to microbial antigens by controlling the expression of MHC molecules (Chapter 7).
  2. IL-4 and IL-5 promote the production of certain classes of antibodies by B lymphocytes.
  3. IL-12 promotes differentiation of T lymphocytes.

D. Adjuvants are substances that promote adaptive immune responses.
  1. Most adjuvants act by inducing danger signaling.
  2. Adjuvants can increase the expression of lymphocyte coreceptors.
  3. Adjuvants can induce the expression of ligands for lymphocyte coreceptors.
  4. Adjuvants can induce cytokine production or increased cytokine receptor expression.

CLINICAL PROBLEMS

Ms. Jones is a retired secretary who has been admitted to the hospital for treatment of an apparent urinary tract infection. She is administered a third-generation cephalosporin antibiotic at approximately 1:00 PM, at which time she has a fever of 101°F, blood pressure of 110/60, and a pulse of 115. The patient tolerates the antibiotic well during the first hour, but when the nurse returns to her room at 3:00 PM, Ms. Jones' vital signs have deteriorated. Her blood pressure has decreased to 80/50, her pulse is now 128, and she no longer responds when called by name. Her physician concludes that Ms. Jones is septic.

1. Which of the following treatments should be administered immediately?

A. Increase the dose of antibiotic to control the infection.

B. Administer a vasodilator, such as verapamil.

C. Discontinue the antibiotic and administer intravenous fluids.

D. Administer TNF-α to control the infection.

E. Administer complement components to control the systemic inflammatory response.

Johnny is a 1-month-old healthy child who has not, as yet, received any childhood immunizations. He presents with his first episode of otitis media (middle ear infection) that is successfully treated with a 3-week course of antibiotics.

2. Which one of the following immune components contributed the most to his clearing the infectious agent during the first few days of his infection?

A. Antigen receptors on his B lymphocytes

B. Toll-like receptors on his neutrophils

C. Cytokines that promoted antibody formation

D. T cell responses to bacterial antigens

E. Memory B cells

Recently a patient was identified who had a defect in IL-1 receptor-associated kinase (IRAK)-dependent cellular signaling associated with her TLR4 receptor.

3. Which one of the following groups of pathogens would be expected to cause recurrent infections in this individual?

A. Retroviruses, such as HIV-1

B. Fungi that cause vaginal yeast infections

C. Gram-negative bacteria

D. Gastrointestinal viruses

E. Insect-borne parasites

Anaerobic bacteria are often cultured from infected deep tissue abscesses.

4. If you were a neutrophil recruited to an anaerobic site to kill such a bacterium, which of the following substances would you most likely use?

A. IL-12

B. Nitric oxide

C. Interferon-α

D. Respiratory burst oxidase

E. Cathelicidin

You are part of a research team that is attempting to design a better vaccine for the prevention of tuberculosis, which is caused by the intracellular bacterial pathogen *Mycobacterium tuberculosis*. One of your colleagues suggests that you include an adjuvant in the vaccine formulation.

5. Based on your knowledge of protective immunity to this pathogen, which one of the following would be a reasonable choice of an adjuvant component?

A. A cytokine that promotes an IFN-γ response to mycobacterial antigens

B. The complement peptide C3d, which will ensure adequate antibody production.

C. Interleukin-10

D. Bacterial lipopolysaccharide

E. Lactoferrin

ANSWERS

1. The answer is C. Sepsis is a systemic inflammatory response to infection that results from exposure of the host to diverse microbial components expressing PAMP. With antibiotic treatment, large quantities of bacteria die and release these proinflammatory components, including bacterial LPS. The most important initial step is to discontinue antibiotic treatment until the septic episode has passed. Fluids are given to correct hypotension, and severe cases may need to be treated with pressors (eg, dopamine) to maintain blood pressure. The patient's symptoms (hypotension, tachycardia, and hypoxia) are indicative of extreme vasodilation, the loss of fluid to the extravascular tissues, and inadequate tissue oxygenation. TNF-α is thought to be a major central mediator of systemic septic shock. The activation of complement would be expected to aggravate the systemic inflammatory response by further inducing vascular changes, hypotension, and hypoxia.

2. The correct answer is B. In a child of this age who has not previously been exposed to this bacterial pathogen or immunized against its antigens host defense is primarily mediated by the innate immune system. Neutrophils play a central role in clearing bacteria and recognize molecular patterns on these pathogens via their TLR. By contrast, T and B lymphocytes mediate adaptive immunity (eg, antibody formation), which requires several days to develop in an immunologically naive individual.

3. The correct answer is C. TLR4 is the signaling receptor for bacterial LPS, a component of the outer membrane of gram-negative bacteria. Patients with impaired TLR4 signaling are at risk for recurrent, life-threatening infections with gram-negative bacteria. TLR4 is not known to mediate protective responses to viruses, fungi, or parasites.

4. The correct answer is E. In the absence of molecular oxygen, neither reactive oxygen species (eg, superoxide) nor NO can be produced in sufficient quantities to kill bacteria. Under these conditions, the neutrophil must rely on oxygen-independent killing mechanisms, such as the action of its antimicrobial granule peptides.

5. The correct answer is A. Any cytokine that would promote the development of antigen-specific, IFN-γ-producing lymphocytes would probably have a favorable effect. Patients who cannot produce IFN-γ are at risk for developing mycobacterial infections. Interleukin-12 is a good example of an IFN-γ-inducing cytokine. Because this pathogen resides within tissue macrophages in chronically infected individuals, macrophage activation for intracellular killing is an essential protective response to infection.

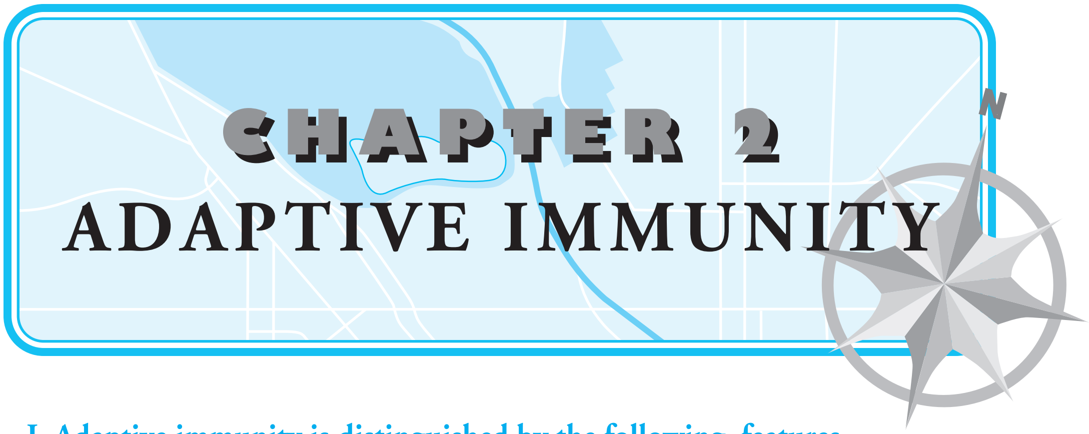

CHAPTER 2

ADAPTIVE IMMUNITY

I. Adaptive immunity is distinguished by the following features.

A. Unlike innate immunity, **adaptive immunity** is an *acquired* response to antigen that is initiated by the recognition of discrete *antigenic determinants* on foreign invaders (Table 2–1).

B. The host is changed by its exposure to antigen; the individual becomes **"immunized"** against a particular antigen.

  1. The **primary response** to an antigen takes several days and requires antigen recognition, the activation and proliferation of T and B lymphocytes, and the differentiation of these cells into populations of **effector lymphocytes**.

  2. Sets of antigen-specific **memory T and B cells** are also generated that mediate **secondary responses** to the antigen at a later time.

  3. The long-term maintenance of memory and the return of lymphocytes to a nonactivated state are carefully controlled.

C. T and B lymphocytes are the primary mediators of adaptive immunity and recognize antigenic determinants by their cell surface **antigen receptors**.

  1. Receptors with specificity for **autoantigens** are expressed during the development of the adaptive immune system.

    a. Most lymphocytes with autoreactive receptors are deleted.

    b. Some autoreactive lymphocytes survive, but their activation is carefully controlled in the periphery.

  2. Whereas B cell receptors predominantly recognize soluble native antigens, T cell receptors recognize foreign antigens only on the surfaces of other host cells.

D. Many of the effector cells and molecules that mediate antigen clearance in adaptive immunity are the same as those that mediate protective innate immune responses.

II. Primary and secondary adaptive immune responses differ.

A. Evidence of a **primary immune response** to an antigen appears only after an initial **lag phase** (Figure 2–1).

B. Antibody produced following **active immunization** is **specific** for the immunizing antigen.

C. The host has enormous **diversity** in its capacity to respond to different antigens.

  1. Estimates of the number of different antigen receptors potentially expressed by B or T cells range from 10^8 to 10^9.

Table 2–1. Comparison of the properties of innate andadaptive immunity.

Property	Innate Immunity	Adaptive Immunity
Time	Immediate	Delayed
Recognition	Conserved, widely distributed microbial components	Discrete, diverse antigenic determinants Antigen presentation
Cells	Many cells: phagocytes, some lymphocytes, epithelial cells	Lymphocytes
Response	Uptake and clearance, danger signaling	Clearance, lysis, memory

2. A diverse **immune repertoire** exists at birth in human beings and undergoes further changes based on the immunological experiences of the individual.

D. Immunity mediated by lymphocytes or the antibodies they produce can be transferred from an immune host to a naive recipient.

1. The transfer of antibodies is called **passive immunization**.
2. The transfer of immune cells is called **adaptive immunization**.

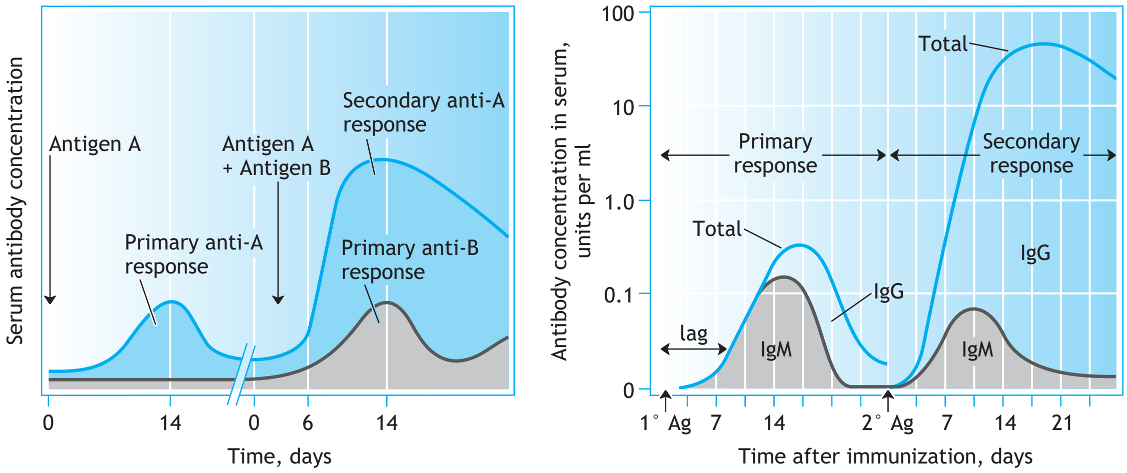

Figure 2–1. Time course of a typical primary and secondary antibody response. Ig, immunoglobulin.

NEWBORNS ACQUIRE MATERNAL IMMUNITY BY PASSIVE IMMUNIZATION

- *Newborns are passively immunized when their mothers transfer protective antibodies to them either across the placenta or through the colostrum and milk.*
- *The relative importance of these two routes in various mammalian species is determined by the structure of their placentas.*
- *Human beings have two cell layers separating fetal and maternal blood and actively transport maternal antibodies across the placenta.*
- *The class of antibody that is transported is immunoglobulin G (IgG) (Chapter 3) and it provides systemic antibody protection to the newborn.*
- *Colostral antibodies in humans are predominantly immunoglobulin A (IgA) and they protect the newborn intestinal tract from infectious pathogens.*
- *By contrast, the fetal calf receives no immunoglobulin from its mother during gestation, and is highly dependent upon suckling colostrum containing IgG antibodies during its first few days of life.*

E. **Secondary responses** to an antigen in immune animals differ from primary adaptive immune responses (Figure 2–1).
  1. The lag period is shorter for the secondary response.
  2. The overall amount of antibody produced is greater in the secondary response.
  3. The class of antibody differs in the two responses.
    a. Primary antibody responses are mostly **immunoglobulin M (IgM)**.
    b. Secondary antibody responses consist primarily of non-IgM classes, especially **IgG**.
    c. The change that occurs in the class of antibody produced is called **isotype switching** and results in new functions being associated with the same antibody specificity (Chapter 3).
    d. The longer duration of the secondary response reflects the large number of memory B cells that are activated and the longer half-life of IgG in circulation compared to IgM.

III. Cells of the adaptive immune system are found in discrete lymphoid tissues and organs.

A. **Primary lymphoid organs** are organs in which the antigen-independent development of lymphocytes occurs.
  1. The **bone marrow** is a major site of hematopoiesis and lymphopoiesis in both young and adult animals.
  2. Under the influence of the bone marrow stromal cells and growth factors, the various blood lineages develop.

HEMATOPOIETIC GROWTH FACTORS CAN CORRECT IMMUNE DEFICIENCIES

- ***Cyclic neutropenia*** *is a 3-week oscillating deficiency in the production of blood neutrophils that can leave patients at risk for infections (Table 2–2).*
- *The inherited form of the disease results from point mutations in the neutrophil elastase gene, suggesting that this protease participates in myelopoiesis.*
- *Treatment with* ***colony-stimulating factor for granulocytes (CSF-G)*** *replenishes neutrophil numbers in the blood during the neutropenic phase.*
- *By contrast,* ***severe congenital neutropenia****, which is due to a mutation in the gene for the receptor for CSF-G, is not responsive to CSF-G therapy.*

Table 2–2. Congenital leukocyte and lymphocyte deficiencies affecting immunity.

Cell Type	Subsets	Normal Cell Numbers (10^3 per μL) in the Blood of Adults	Congenital Deficiencies Affecting the Number of Cells in the Periphery
Leukocytes	Total	4–11	
	Neutrophils	2–7	Neonatal and cyclic neutropenia
	Monocytes	0.2–1.2	
Lymphocytes	Total	1.0–4.8	Severe combined immune deficiency
	B cells	0.1–0.7	X-linked agammaglobulinemia
	T cells	0.5–3.6	DiGeorge syndrome
	CD4$^+$ T cells	0.2–1.8	MHC class II deficiency
	CD8$^+$ T cells	0.1–0.9	MHC class I deficiency

- *Colony-stimulating factors for other hematopoietic progenitor cells, including erythropoietin and interleukin-3, have become standard treatments for many selective hematopoietic deficiencies.*

3. Lymphocytes are also derived from a common self-renewing **hematopoietic stem cell** that gives rise to all blood cell lineages.
   a. In the appropriate inductive microenvironment, the pluripotent stem cell differentiates into a **lymphoid progenitor cell (LPC)**.
   b. The LPC can become a **progenitor T (pro-T) cell** or a **progenitor B (pro-B) cell**.
   c. In human beings, B lymphopoiesis occurs primarily in the bone marrow (Chapter 9), but T lymphocyte development moves to the thymus at the pro-T cell stage (Chapter 10).
   d. An important process that accompanies T and B cell differentiation in the bone marrow and thymus is the expression of surface antigen receptors.

DIGEORGE SYNDROME PATIENTS LACK A THYMUS

CLINICAL CORRELATION

- *DiGeorge syndrome is a congenital condition arising from defective embryogenesis of the third and fourth pharyngeal pouches.*
- *Patients with DiGeorge syndrome show abnormalities in the structure of their major blood vessels, heart, and parathyroids and evidence thymic hypoplasia.*
- *Immunological abnormalities include severe lymphopenia (Table 2–2) at birth and early onset infections by opportunistic viruses and fungi.*
- *Antibody levels are normal at birth due to the transplacental passage of antibodies from the mother in utero.*
- *The immune deficiencies of these children can be cured by thymic transplantation from a suitable donor.*

B. **Secondary lymphoid organs** are sites at which the host mounts adaptive immune responses to foreign invaders.

1. Secondary lymphoid organs and tissues include the spleen, lymph nodes, Peyer's patches, and widely distributed lymphoid follicles.
2. The **lymph node** is an encapsulated organ that receives antigens from subcutaneous and submucosal tissues via its **afferent lymphatics** (Figure 2–2).
   - a. B cells are concentrated in discrete **primary and secondary follicles** within the cortex of lymph nodes, where they undergo antigen-driven differentiation.
   - b. Memory B cells develop within the **germinal centers** of the cortex.
   - c. T cells are located primarily in the **diffuse cortex** (or **paracortex**), where they associate with dendritic cells.
   - d. Cells enter the lymph nodes in large numbers by crossing the high endothelial layer of **postcapillary venules** located in the diffuse cortex.
   - e. Terminally differentiated B cells, called **plasma cells**, are found in the **medullary cords**, where they produce large amounts of antibody during their limited life-span.
   - f. Secreted antibodies exit the lymph nodes via the **efferent lymphatics** and eventually enter the blood stream.
3. The **spleen** receives antigens through the blood circulation and contains areas functionally equivalent to those of the lymph nodes (Table 2–3).
4. Other important peripheral lymphoid tissues include the **Peyer's patches** and submucosa of the small intestine, which are sites in which mucosal antibody responses are induced (Chapter 9).

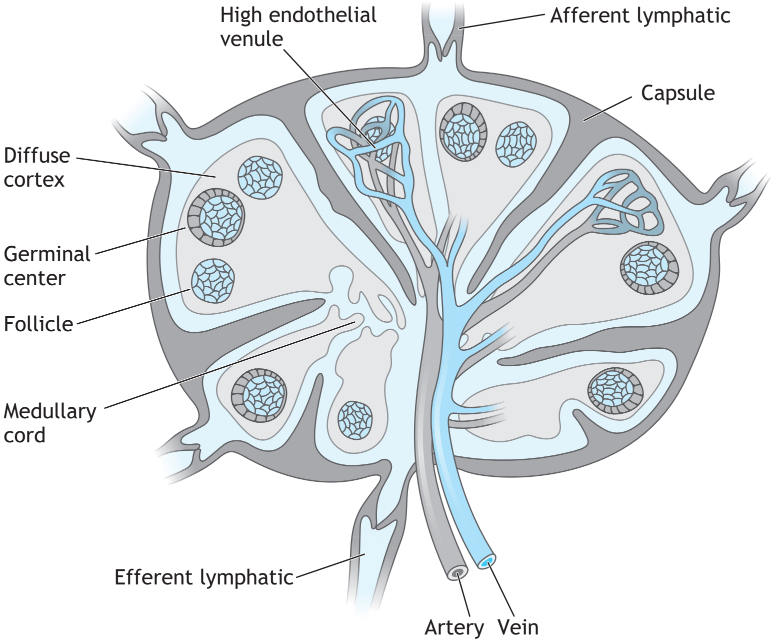

Figure 2–2. Schematic structure of a lymph node.

Table 2–3. Comparison of analogous structures in the lymph nodes and spleen.

Lymph Node	Spleen	Function
Follicles	Follicles	B cell activation and differentiation
Diffuse cortex	Periarteriolar lymphoid sheath	T cell activation and differentiation
Medullary cords	Red pulp cords	Antibody production by plasma cells
High endothelium of postcapillary venules	Marginal sinuses	Site of entry of lymphocytes from the recirculating lymphocyte pool
Afferent lymphatics	Marginal sinuses	Site of entry of antigens
Efferent lymphatics	Marginal sinuses	Point of exit of effector lymphocytes to join the recirculating pool

C. A **recirculating lymphocyte pool** of long-lived small lymphocytes continually travels between the various lymphoid tissues by a route that includes the blood and lymph (Figure 2–3).

1. Recirculating lymphocytes enter the lymph nodes from the blood by crossing the **high endothelial venules** in the diffuse cortex.
2. The cells exit the lymph nodes via the **efferent lymphatics** and migrate via the **common thoracic duct** to the blood stream.
3. Lymphocytes can exit the blood circulation by crossing the endothelium at many locations.

IV. The **clonal selection theory of adaptive immunity** proposes that an antigen selects and activates the appropriate clone of lymphocytes from a preformed diverse pool.

A. The theory predicts the following:

1. Each lymphocyte is precommitted to a particular antigen prior to encountering that antigen (Figure 2–4).
2. Lymphocytes recognize their antigens with cell surface antigen receptors.
3. The receptor on a given lymphocyte is specific for only one antigen.
4. Antigen binding to the receptor induces the expansion of a **clone** of cells, all with identical receptor specificity.
5. The antigen receptor on a given lymphocyte is uniform and identical to the antibody molecules secreted by the cell.

B. The theory has proven correct for both B and T lymphocytes with the following exceptions:

1. The B cell antigen receptor is actually a modified form of an antibody molecule (Chapter 4).
2. The T cell antigen receptor is not an antibody molecule (Chapter 6).
3. T cells do not secrete large amounts of their receptor molecules.

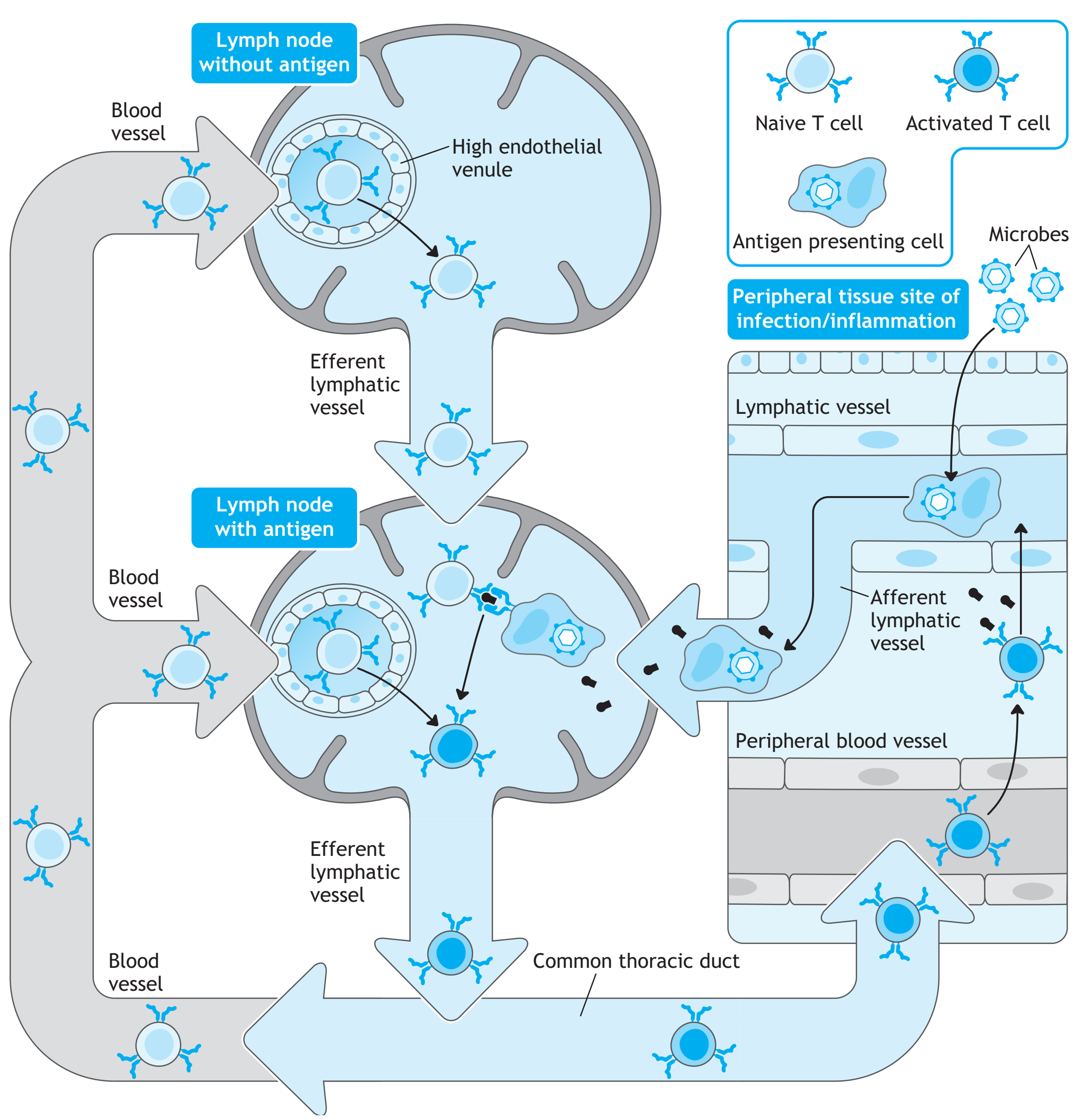
Lymph node without antigen
Blood vessel
High endothelial venule
Efferent lymphatic vessel
Lymph node with antigen
Blood vessel
Efferent lymphatic vessel
Blood vessel
Common thoracic duct
Naive T cell
Activated T cell
Antigen presenting cell
Microbes
Peripheral tissue site of infection/inflammation
Lymphatic vessel
Afferent lymphatic vessel
Peripheral blood vessel

Figure 2–3. Route of recirculation of lymphocytes from blood to lymph.

C. The production of a pool of **memory lymphocytes** ensures that a higher concentration of specific antibody is produced in a more rapid fashion during a secondary response.

D. Because lymphocytes with receptors specific for autoantigens (**"forbidden clones"**) are mostly deleted or inactivated during their differentiation, **autoimmunity** is rare.

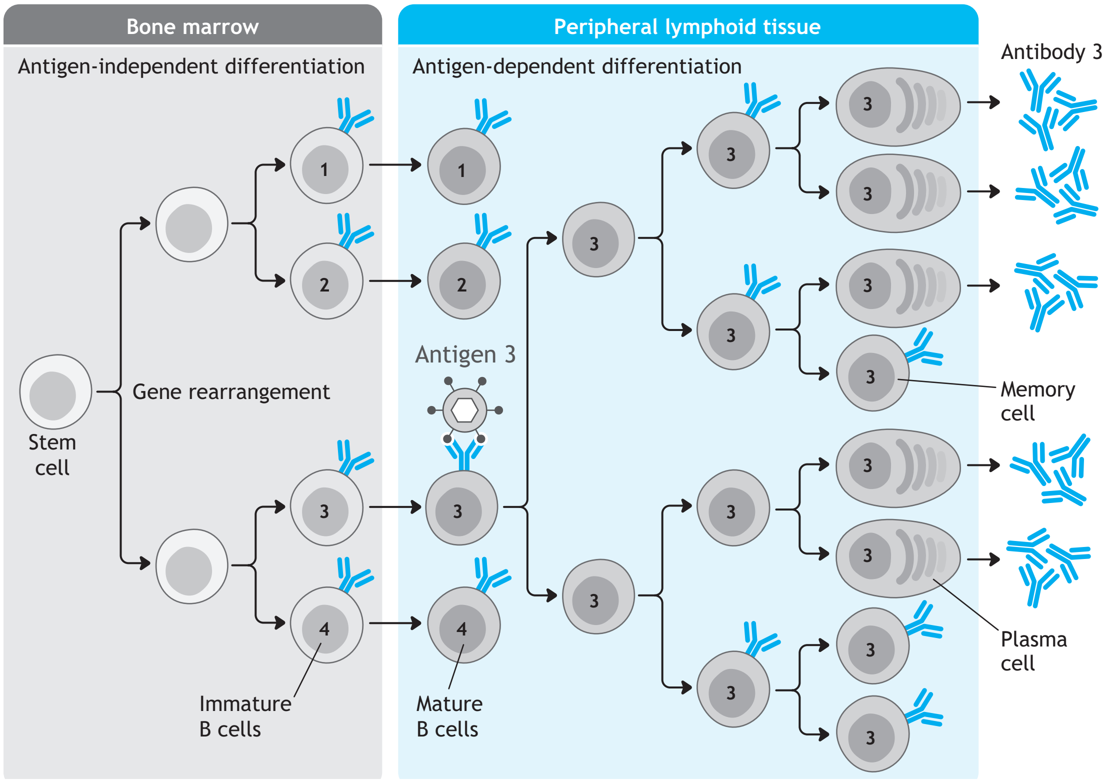
Bone marrow
Antigen-independent differentiation
Peripheral lymphoid tissue
Antigen-dependent differentiation
Antibody 3
Stem cell
Gene rearrangement
Antigen 3
Immature B cells
Mature B cells
Memory cell
Plasma cell
1
2
3
4

Figure 2–4. Clonal selection theory of adaptive immunity.

PERNICIOUS ANEMIA

- ***Pernicious anemia (PA)*** *is an organ-specific autoimmune disease characterized by a decreased absorption of dietary vitamin* B_{12}*.*
- *Vitamin* B_{12} *is normally absorbed in the ileum as a complex with* ***intrinsic factor****, a protein synthesized by gastric parietal cells.*
- *Failure to absorb vitamin* B_{12} *in PA results from an autoimmune attack on gastric parietal cells and the clearance of intrinsic factor by autoantibodies.*
- *The resulting vitamin* B_{12} *deficiency results in impaired erythropoiesis (megaloblastic anemia).*

CLINICAL CORRELATION

V. Lymphocytes express antigen receptors.

A. The **antigen receptor complexes** of T and B lymphocytes are similar in structure and function (Figure 2–5).

1. The B cell receptor (BCR) for antigen consists of a membrane form of antibody and the accessory peptides **Ig**α and **Ig**β.

   a. The BCR recognizes discrete antigenic determinants (epitopes) on soluble antigens.

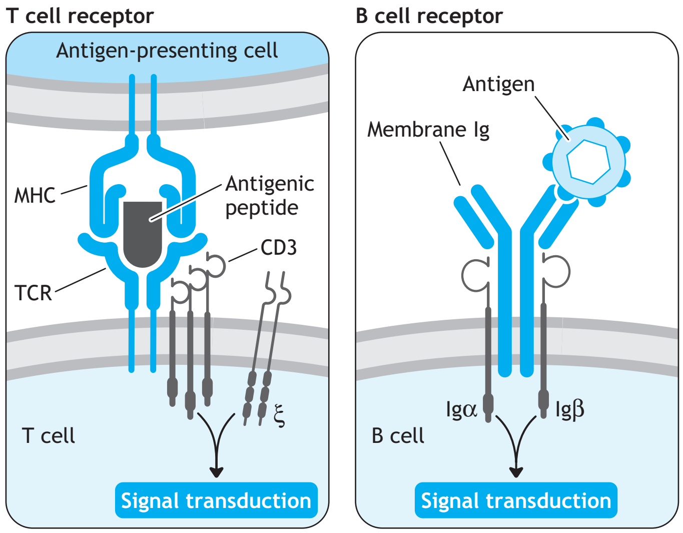
T cell receptor
Antigen-presenting cell
MHC
Antigenic peptide
CD3
TCR
ξ
T cell
Signal transduction
B cell receptor
Antigen
Membrane Ig
Igα
Igβ
B cell
Signal transduction

Figure 2–5. Antigen receptors on B and T lymphocytes. MHC, major histocompatibility complex; TCR, T cell receptor; Ig, immunoglobulin.

B. Igα and Igβ mediate antigen-induced transmembrane signaling in B cells.

1. The **T cell receptor (TCR)** complex consists of antibody-like peptides and signaling peptides.
   - **a.** Unlike the BCR, the TCR can recognize only foreign antigens that are displayed on the surfaces of other cells.
     - *(1)* T cells recognize surface-bound small peptides that are derived by proteolysis from native protein antigens.
     - *(2)* The TCR recognizes determinants of the foreign peptide plus host cell surface molecules called **major histocompatibility complex (MHC) molecules**.
     - *(3)* The cells on which these processed foreign peptides are displayed are collectively referred to as **antigen-presenting cells (APC)**.
   - **b.** The MHC regulates antigen recognition by T cells.
     - *(1)* Only peptides that can bind to the host's MHC molecules are recognized by T cells.
     - *(2)* Two classes of MHC molecules control T cell antigen recognition in this fashion (Figure 2–6).
       - *(a)* **MHC class I** is expressed on nearly all nucleated cells.
       - *(b)* **MHC class II** is expressed on dendritic cells, macrophages, and B cells.
     - *(3)* MHC restriction ensures that T cells will be activated by antigen only in proximity to other host cells.
       - *(a)* **T helper (Th) cells** recognize antigen presented by MHC class II-expressing **dendritic cells (DC)**, B cells, and macrophages.

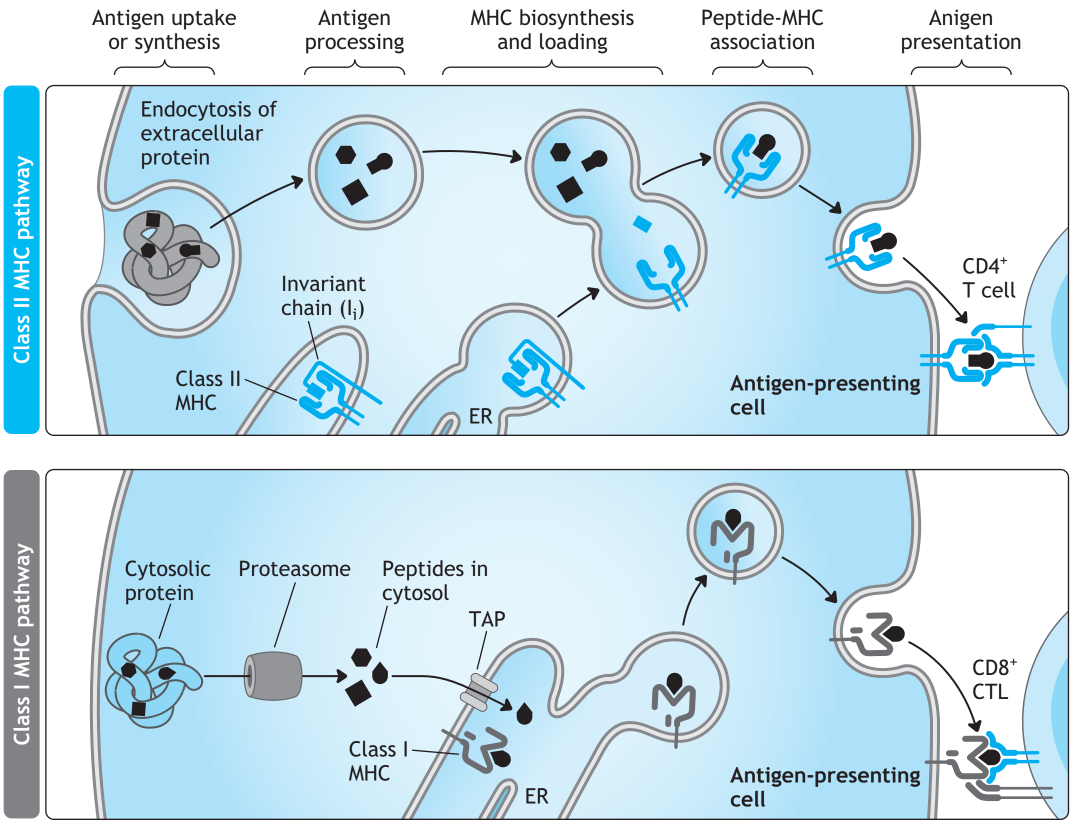
Antigen uptake or synthesis
Antigen processing
MHC biosynthesis and loading
Peptide-MHC association
Anigen presentation
Class II MHC pathway
Endocytosis of extracellular protein
Invariant chain (I$_i$)
Class II MHC
ER
CD4$^+$ T cell
Antigen-presenting cell
Class I MHC pathway
Cytosolic protein
Proteasome
Peptides in cytosol
TAP
Class I MHC
ER
CD8$^+$ CTL
Antigen-presenting cell

Figure 2–6. Schematic of antigen presentation. MHC, major histocompatibility complex; ER, endoplasmic reticulum; CTL, cytotoxic T lymphocyte; TAP, transporter of antigenic peptide.

   (b) **Cytotoxic T (Tc) cells** recognize antigen-expressing (eg, viral antigens), MHC class I-expressing host cells.

c. The signaling peptides of the TCR are collectively referred to as **CD3**.

d. The TCR complex is first expressed during T cell development in the thymus (Chapter 10).

T CELL DEFICIENCIES DUE TO ABNORMAL CD3 EXPRESSION

CLINICAL CORRELATION

- *Mutations in two different CD3 peptides (CD3γ and CD3ε) have been described that decrease TCR expression.*
- *These patients show few peripheral T cells, susceptibility to viral and fungal infections, and autoimmunity.*
- *In addition to treatment for infections, these patients are candidates for hematopoietic stem cell transplantation, which can correct the defects.*
- *This and other congenital immune deficiencies are described in detail at the Online Mendelian Inheritance in Man web site maintained by the National Library of Medicine (www.ncbi.nlm.nih.gov/omim).*

C. **Antigen presentation** requires the processing and MHC-dependent display of antigenic determinants (Figure 2–6).

  1. Any nucleated cell can potentially present peptide antigens through MHC class I.
    a. T cells that recognize peptides + MHC class I bear a distinguishing cell surface coreceptor called **CD8**.
    b. Most CD8$^+$ T cells are cytotoxic.
  2. DC, B cells, monocytes, and macrophages express MHC class II and can present antigenic peptides via class II molecules.
    a. T cells that recognize peptides + MHC class II bear a distinguishing cell surface coreceptor called **CD4**.
    b. Most CD4$^+$ T cells secrete cytokines that regulate the activation of other immune cells.

VI. Lymphocytes bear a number of additional cell surface molecules that control their activation, migration, or effector functions.

Table 2–4. Properties of CD markers.[a]

CD Category	Examples	Expression	Major Functions
Antigen presentation	CD1	Dendritic cells	Presents glycolipid antigens
Adhesion molecules	CD18	Leukocytes	Adhesion to endothelium
Coreceptors	CD4	T cells	Coactivates with TCR-CD3
Cytokine receptors	CD25	B cells	Binding of IL-2
Ig-binding receptors	CD32	Macrophages	Binding IgG–antigen complexes
Signal transduction	CD3	T cells	Mediates signaling of TCR
Homing receptors	CD62L	B and T cells	Homing to lymph nodes
Death-inducing receptors	CD120	Many cells	TNF-α-induced apoptosis
Enzymes	CD45	T cells	Phosphatase; regulates TCR signaling
Complement receptors	CD55	Many cells	Regulates complement activation
Blood group markers	CD240D	Erythrocytes	Major Rh antigen

[a] For a more complete list of CD markers, see Appendix I. TCR, T cell receptor; IL-2, interleukin 2; IgG, immunoglobulin G; TNF, tumor necrosis factor.

A. **Coreceptors** on lymphocytes promote signaling through the TCR and BCR.
   1. Coreceptors lower the threshold for lymphocyte activation through their antigen receptors.
   2. Coreceptors can act by modulating intracellular signaling pathways or increasing the expression of other receptors.

B. T and B cells express a range of **cytokine receptors** that provides additional signals for cell activation.

C. Receptors that bind **IgG molecules** present in antigen–antibody complexes (**Fc receptors**) typically inhibit B cell activation.

D. **Cell adhesion molecules** mediate lymphocyte migration between and within tissues and increase the binding between lymphocyte subsets.

E. One method for cataloging cell surface molecules is the **cluster of differentiation (CD)** scheme that assigns a CD number to each unique cell surface molecule (Table 2–4, Appendix I, and www.hlda8.org/).

PHENOTYPING OF LYMPHOCYTE SUBSETS

CLINICAL CORRELATION

- *Human lymphocyte subsets are routinely enumerated in the clinical laboratory by a technique known as* ***flow cytometry*** *(Chapter 5).*
- *Monoclonal antibodies to the* ***CD markers*** *of interest (Table 2–4) are labeled with a fluorochrome and used to stain cells.*
- *The flow cytometer detects labeled antibody binding to the cell and thereby enumerates surface CD molecules on blood cells or cells prepared from solid tissues (eg, tumors).*
- *Abnormal lymphocyte numbers are associated with congenital or acquired immune deficiencies, infections, and neoplastic conditions.*
- *Leukemias and lymphomas can be typed and staged by defining the CD markers they express.*

CLINICAL PROBLEMS

A 2-month-old male child presents with thrush (a yeast infection in the oral cavity), diarrhea, and failure to thrive. His complete blood count reveals a severe lymphopenia. Flow cytometry demonstrates a very low number of $CD3^+$ lymphocytes in his blood, but normal numbers of membrane IgM^+ lymphocytes when compared to age-matched controls.

1. Which one of the following represents the most likely underlying disease in this child?

   A. X-linked agammaglobulinemia (XLA)
   B. DiGeorge syndrome
   C. Neonatal neutropenia
   D. Myeloperoxidase deficiency
   E. Aplastic anemia

A patient has a history of recurrent pneumonias that reappear within a week following completion of antibiotic therapy. She is found to have a deficiency in the expression of CD18.

2. Which of the following two clinical abnormalities would you most expect to find in such a patient? More than one answer may be correct.

    A. Lymphopenia
    B. Leukocytosis
    C. Recurrent viral infections
    D. Recurrent bacterial infections
    E. Abnormal BCR cell signaling
    F. Agammaglobulinemia
    G. Reduced T lymphocyte receptor expression

3. Which of the following is a “pattern recognition receptor”?

    A. BCR
    B. Interleukin-1 receptor
    C. Fc receptor
    D. CD4
    E. Mannose receptor

A patient presents with a lymphocytosis and enlarged lymph nodes, and a biopsy of the bone marrow is performed. Ninety percent of the patient’s bone marrow cells stain with a fluorescent antibody specific for CD3.

4. Which of the following is a reasonable differential diagnosis?

    A. AIDS
    B. DiGeorge syndrome
    C. T cell leukemia
    D. Cytomegalovirus infection
    E. This is a normal laboratory finding.

Patients with deficiencies in antibody production can often present with the same types of infections as are seen in patients with phagocytic cell deficiencies.

5. Which of the following statements **best explains** this observation?

    A. Autoantibodies can remove phagocytic cells from the blood circulation.
    B. Plasma cells are the direct progenitors of certain phagocytic cells.
    C. Macrophages can differentiate into antibody-producing plasma cells.
    D. Antibodies are important opsonins that promote microbe recognition by phagocytes.
    E. Antibodies are essential for the continued production of phagocytic cells by the bone marrow.

ANSWERS

1. The correct answer is B. The finding of low lymphocyte counts in the blood (lymphopenia) affecting only T cells suggests a deficiency in cell-mediated immunity secondary to decreased T cell numbers, such as in thymic dysplasia (DiGeorge syndrome). This would be expected to lead to opportunistic infections by intracellular pathogens, such as the yeast *Candida albicans*.
2. The correct answers are B and D. The CD18 gene encodes for an adhesion molecule, and patients with decreased CD18 expression show poor leukocyte adhesion to the vascular endothelium. The increase in leukocyte production seen during infections results in an increased number of leukocytes in the blood (leukocytosis). The resulting decreased leukocyte migration to infection sites impairs clearance of extracellular bacterial pathogens.
3. The correct answer is E. The mannose receptor recognizes the spatial arrangement of mannose residues that is found only on microbial surfaces.
4. The correct answer is C. Finding large numbers of $CD3^+$ cells in the bone marrow is most likely indicative of neoplastic T cells (leukemia), because CD3 is normally expressed only after the pro-T cell migrates to the thymus.
5. The answer is D. Antibodies of certain classes can bind to microbial surface antigens and mark them for uptake by phagocytic cells. This is because phagocytes bear opsonic receptors that bind antibody-coated particles and signal an increased rate of uptake. Thus, patients who lack certain antibodies will show diminished phagocytic cell function due to impaired opsonophagocytosis.

CHAPTER 3

ANTIGENS AND ANTIBODIES

I. Antigens

- **A.** An **antigen** is a substance that can elicit an adaptive immune response.
- **B.** Most natural and medically important antigens are macromolecules or substances that can bind covalently to them.
- **C.** Naturally occurring antigens include proteins, polysaccharides, lipids, and nucleic acids.
- **D.** The region of an antigen that is recognized by the immune system is called an **antigenic determinant** or **epitope**.
- **E.** Antigens generally have two properties.
  - **1. Immunogenicity** is the capacity to induce an immune response.
    - **a.** Immunogenicity is determined by the molecular mass, molecular complexity (number of potential determinants), and conformation of an antigen.
    - **b.** A high degree of phylogenetic disparity between an antigen and the host generally promotes immunogenicity.
  - **2. Antigenicity** is the ability to bind specifically to antibody molecules or antigen receptors on lymphocytes.
  - **3.** A **hapten** is a molecule that is antigenic, but not immunogenic.
    - **a.** A hapten generally has a molecular mass of less than 10,000 Da.
    - **b.** A hapten can become immunogenic if it is covalently attached (conjugated) to a **carrier molecule**.
    - **c.** Some haptens can cause **allergic dermatitis** if they bind to proteins in the skin.
    - **d.** Conjugating microbial epitopes (haptens) to carrier proteins is an effective approach to producing highly immunogenic vaccines.

PENICILLIN IS A HAPTEN THAT CAN INDUCE IMMUNE-MEDIATED TISSUE DAMAGE

CLINICAL CORRELATION

- *Penicillin G is a small drug (MW 356) that can bind to a variety of host proteins, including those on the surface of human erythrocytes.*
- *Antipenicillin antibodies can be produced that cause **autoimmune hemolytic anemia**.*
- *The **Coombs test** is used to determine if an anemia has an immune basis by determining whether immunoglobulin G (IgG) antibodies are present on the patient's erythrocytes.*
- *The treatment of Coombs-positive anemias includes discontinuing the drug (hapten) and transfusing normal ABO-matched erythrocytes.*

F. Antigenic determinants that bind to antibodies and B cell antigen receptors are often different from those that bind to T cell antigen receptors (Table 3–1).
  1. Determinants recognized by antibodies or B cells are typically found on the exposed **hydrophilic** surface of an antigen molecule.
  2. Antigens that bind to the T cell antigen receptor typically do not possess their native conformation.
    a. Most antigens that stimulate T cells are proteins.
    b. For a protein to be recognized by a T cell it must be degraded into peptides and presented by an antigen-presenting cell (APC).
    c. T cell determinants are not limited to surface exposed regions of proteins and can be derived from internal hydrophobic protein domains.

II. Antibodies

A. All antibodies are **immunoglobulins (Igs)**.

B. Igs are a homologous family of proteins with considerable structural heterogeneity.

C. Most Igs migrate in the "gamma" region upon **serum protein electrophoresis** and are therefore called **γ-globulins** (Figure 3–1).

INTRAVENOUS IMMUNOGLOBULIN (IVIG)

CLINICAL CORRELATION

- *Immunoglobulins (γ-globulins) are routinely given to patients for the prevention or treatment of various diseases.*
- *IVIGs are often used to treat congenital antibody deficiencies and must be given repeatedly.*
- *Because IVIG is greater than 95% IgG, its primary immune benefit in immunodeficient patients is to provide protection against extracellular microbial pathogens or their toxic products.*
- *Some infectious diseases in normal individuals, such as rabies and hepatitis, can be effectively prevented with γ-globulins prepared from specific immune donors that are administered shortly after a suspected infection.*

D. The structure of antibodies was first determined by studying Igs purified from the sera of patients with **multiple myeloma**.

Table 3–1. Comparison of the antigenic determinants recognized by B and T lymphocytes.

Property	B cells	T cells
Conformation dependent	+	–
Hydrophillic/hydrophobic	Hydrophillic	Hydrophillic and hydrophobic
Location of determinants	Surface	Surface and internal
MHC dependent	–	+
Examples	Native proteins and carbohydrates	Peptides

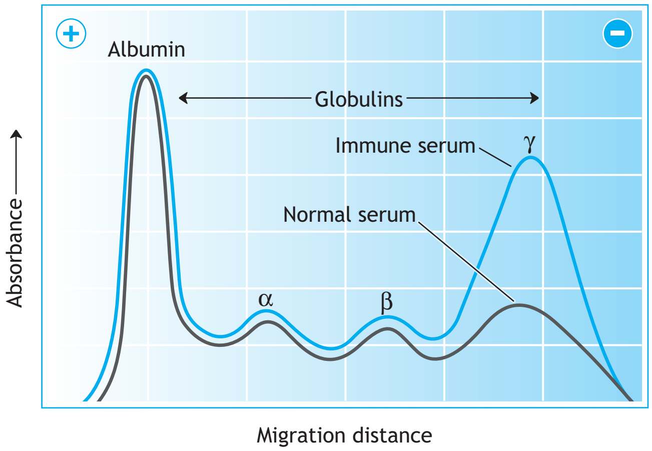

Figure 3–1. Serum protein electrophoresis.

1. Multiple myeloma is a malignancy of a clone of Ig-producing plasma cells.
2. Myeloma Igs (also called **M proteins**) are monoclonal and structurally homogeneous.
3. All M proteins belong to one of five structurally distinct Ig **classes** or **isotypes**.
4. The most frequently encountered isotype among myeloma proteins is IgG.
   - **a.** IgG molecules have the molecular formula H_2L_2 and show **bilateral symmetry** (Figure 3–2).
     - *(1)* Each IgG molecule contains two identical **heavy (H) chains**.
     - *(2)* The H chains of IgG molecules are designated γ chains and have a molecular mass of approximately 50,000 Da.
     - *(3)* H chains are linked to one another by *inter*-chain disulfide bonds.
     - *(4)* Each IgG molecule contains two identical **light (L) chains**, each with a molecular mass of approximately 25,000 Da.
     - *(5)* L chains are covalently linked to H chains by *inter*-chain disulfide bonds.
   - **b.** *Intra*-chain disulfide bonds in both H and L chains maintain protein folds that define **Ig domains** of approximately 110 amino acids in length (Figure 3–2).
   - **c.** There are two such domains in L chains and four in γ H chains.
5. There are a total of five **classes or isotypes** of human Ig, designated IgG, IgM, IgA, IgE, and IgD (Table 3–2).
   - **a.** Each Ig has the core molecular H_2L_2 structure and contains H chains unique to that isotype.
   - **b.** IgG, IgM, IgA, IgE, and IgD have γ, μ, α, ε, and δ, H chains, respectively.
     - *(1)* γ, α, and δ chains have molecular masses of 50,000–60,000 Da.
     - *(2)* μ and ε chains, which have an additional Ig domain, have molecular masses of 65,000–70,000 Da (Table 3–2).

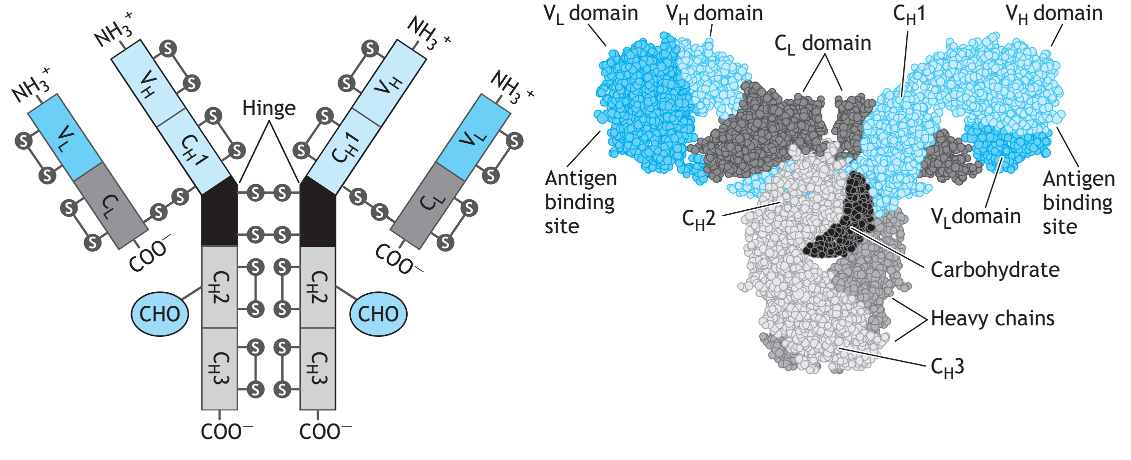

Figure 3–2. Structure of an immunoglobulin (Ig) G molecule.

  c. The IgG class is further subdivided into the IgG_1, IgG_2, IgG_3, and IgG_4 subclasses.
  d. The IgA class is further subdivided into the IgA_1 and IgA_2 subclasses.

6. All Igs can occur in two forms, secreted and membrane bound.
7. There are two human L chain **types**, termed κ and λ.
   a. In any one immunoglobulin molecule, both L chains are either κ or λ.
   b. The overall κ:λ ratio for the Igs found in human serum is approximately 60:40.
8. **Oligosaccharides** are covalently attached to the carboxyl-terminal half of H chains, especially within μ and ε chains.

MULTIPLE MYELOMA

CLINICAL CORRELATION

- *Patients with multiple myeloma can have extremely high levels of Igs in their plasma that are often first detected by serum protein electrophoresis (Figure 3–1).*
- *Because their cancers are monoclonal and have a homogeneous charge, the abnormal M protein appears on electrophoresis as a "spike."*
- *Each M protein has a single H chain class and a single L chain type.*
- *Many patients with myeloma also have high concentrations of Ig L chains in their urine, which are produced in excess by their myeloma cells.*
- *In a given patient, these **Bence–Jones proteins** are monoclonal, dimeric, and either κ or λ, but not both.*
- *Finding a significant departure from the normal 60:40 κ:λ ratio in serum or urinary L chains is diagnostic of a **monoclonal gammopathy**, such as myeloma.*

E. IgM and IgA molecules can exist as multimers (Figure 3–3).
   1. Serum IgM is a pentamer of the core H_2L_2 structure that is joined together by intersubunit disulfide bonds and stabilized by a 15-kDa **joiner (J) chain**.

Table 3–2. Properties of human immunoglobulin isotypes.

Property	IgG	IgM	IgA	IgE	IgD
Subclasses	IgG$_{1-4}$	None	IgA$_{1-2}$	None	None
Heavy chain	γ	μ	α	ε	δ
C_H domains	3	4	3	4	3
L chain	κ or λ	κ or λ	κ or λ	κ or λ	κ or λ
Molecular mass[a] (secreted form)	150,000	900,000	150,000 570,000[b]	190,000	170,000
Serum concentration (mg/dL)	1350	150	300	0.05	3
Serum half-life (days)	23	5	6	2	3
Activates complement (classical pathway)	+	++	–	–	–
Placental transport	+	–	–	–	–
Mucosal epithelial transport	–	+	++	–	–
Mast cell degranulation	–	–	–	+	–
Binding to phagocyte Fc receptors	+	–[c]	–	–	–
Binding to Fc receptors on natural killer cells	+	–	–	–	–
Antigen receptor of naive B cells	–	+	–	–	+

[a] Refers to the predominant forms in serum and external secretions, not the form found on B cell surfaces.

[b] Secretory IgA found in external secretions is a dimer with an associated J chain and secretory component.

[c] Recent evidence suggests Fc$_{\mu}$ receptors on neutrophils may mediate host defense against spirochetes.

2. Serum IgA is primarily a monomer, but secretory IgA that is found in external secretions (eg, saliva and colostrum) is dimeric and contains a J chain polypeptide.
    a. Dimeric IgA is synthesized in the intestinal submucosa and efficiently translocated across the mucosal epithelium.

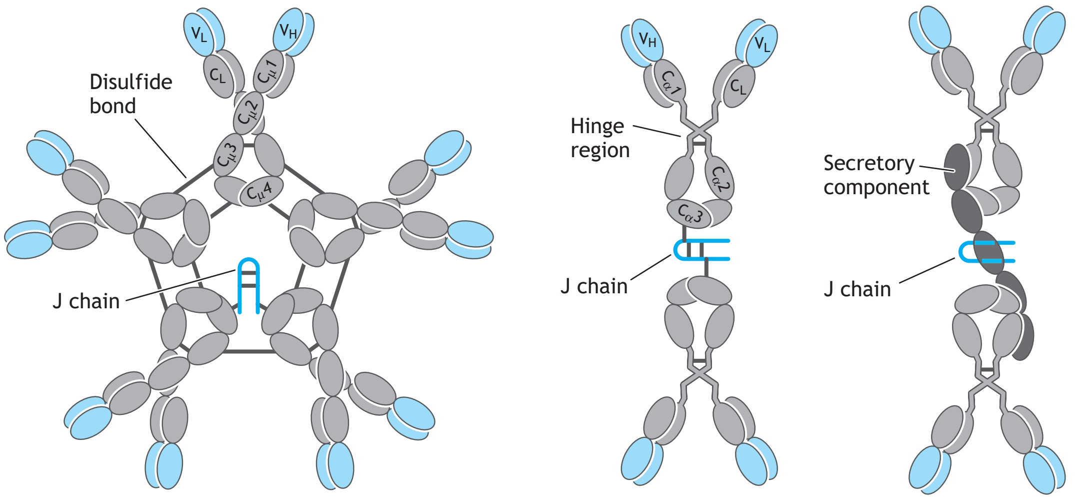

Pentameric IgM Dimeric IgA Secretory IgA

Figure 3–3. Polymeric immunoglobulins (Ig).

- **b.** During the process of translocation, epithelial cells add a 70-kDa peptide, called a **secretory component (SC)**, to the IgA molecule (Figure 3–3).
    - ***(1)*** SC protects secretory IgA from proteolytic digestion.
    - ***(2)*** SC promotes the binding of secretory IgA to the mucous layer of the epithelium.

F. Amino acid sequencing of myeloma H and L chains shows that they have constant (C) and variable (V) regions.

- **1.** The amino-terminal half of an L chain is designated the **variable light (V_L) domain.**
    - **a.** Amino acid sequence diversity in V_L domains is greatest within the three **hypervariable regions**, HV1, HV2, and HV3.
    - **b.** HV regions are also called **complementarity determining regions (CDR)**, because they comprise the walls of the antigen-binding cleft.
- **2.** The carboxyl-terminal half of an L chain is relatively constant and is designated as the **constant light (C_L)** domain (C_κ or C_λ).
- **3.** As with V_L domains, the amino-terminal **V_H** domains of H chains are highly variable when several H chains are compared.
    - **a.** V_H domains contain three hypervariable regions.
    - **b.** V_H domains that have different amino acid sequences generally have different specificities.
- **4.** The antigen-binding activity of an Ig depends on both its V_H and V_L domains.
- **5.** The three carboxyl-terminal **C_H** domains of IgG H chains are designated **C_H**1, **C_H2**, and **C_H**3.

G. Structure and function in Igs are related.

- **1.** The distribution of Ig function within Ig domains was first revealed by studying proteolytic fragments of the molecules.

2. Enzymatic digestion of rabbit IgG antibodies with the protease **papain** yields three fragments.
    - **a.** Two are identical **antigen-binding fragments (Fab)** consisting of the V_H + C_H1 domains.
    - **b.** Each Fab can bind one molecule of antigen.
    - **c.** One **Fc fragment** is also generated and consists of the carboxyl-terminal half of the two disulfide-linked H chains (ie, the C_H2 + C_H3 domains in IgG).
    - **d.** The Fc region mediates the opsonic activity of IgG antibodies by binding them to **Fc receptors** on phagocytic cells.
3. The **effector functions** of antibodies depend on the structures of their C_H domains.
    - **a.** Many functions of antibodies are not expressed until they have bound their antigens, which induces conformational changes in the C_H domains.
    - **b.** Transplacental transport, transepithelial transport, and mast cell binding of Igs do not require prior binding of the Ig to its antigen.
    - **c.** A **hinge region** in many C_H domains is responsible for the flexibility in antibody conformation necessary for cooperative antigen binding.
4. The principal functions of IgG antibodies are to serve as opsonins, activate the complement system, provide fetal protection following transplacental transport, and inhibit B cell activation (Table 3–2).
5. The principal functions of IgM antibodies are to serve as antigen receptors on naive B cells, activate the complement system, clear circulating antigen, and provide host defense at mucosal surfaces.
6. The principal function of IgA antibodies is to provide antibody defense at mucosal surfaces.
7. The principal function of IgE antibodies is to mediate allergic reactions by promoting mast cell degranulation.
8. The principal function of IgD antibodies is to serve as an antigen receptor on naive B cells.

ISOTYPE SWITCHING AND CONJUGATE VACCINES

CLINICAL CORRELATION

- *Antibody production undergoes **isotype switching** late in the primary adaptive immune response (Chapter 2).*
- *This associates new antibody functions with the same antibody specificity, a principle illustrated by the Haemophilus influenzae type b (Hib) conjugate vaccine.*
- *Hib is one of the leading causes of bacterial nasopharyngeal infections in children and can progress to pneumonia and life-treatening meningitis.*
- *Immune elimination of H influenzae requires antibodies to the polysaccharide capsule of this organism.*
- *H influenzae capsular polysaccharide vaccines have not proven effective at eliciting opsonic IgG antibodies in children.*
- *By contrast, the conjugate vaccine consisting of Hib capsular polysaccharides covalently linked to a foreign carrier protein (eg, the tetanus toxoid vaccine protein) is quite effective at inducing protective IgG antibodies in children.*
- *A description of the Hib conjugate vaccine and other recommended pediatric vaccines can be found at www.cdc.gov/nip/recs/child-schedule.pdf.*

H. **Natural antibodies** appear early in life.

1. Most natural antibodies belong to the **IgM class** of Igs.

2. Many natural antibodies are produced in response to carbohydrate antigens present on microbial flora.
3. Some natural antibodies also cross-react with carbohydrate structures on host cells.
    - a. The **isohemagglutinins**, natural antibodies to the **ABO blood group antigens**, can mediate the destruction of transfused mismatched blood.
    - b. Isohemagglutinins recognize the terminal sugar groups of the ABO polysaccharide antigens expressed by erythrocytes and endothelial cells (Figure 3–4).
        - *(1)* Three blood group antigens are defined by this sytem: A, B, and H.
        - *(2)* Because the ABO genes are expressed in a codominant fashion, the potential ABO types are A, B, AB, and O.
        - *(3)* Individuals produce isohemagglutinins specific for the ABO antigens that they lack (Table 3–3).
        - *(4)* With regard to blood transfusions, the **universal donor** is type O, because this individual lacks antigens that would be recognized by ABO-incompatible recipients.
        - *(5)* The **universal recipient** is type AB, because this individual lacks antibodies that would destroy ABO-incompatible donor erythrocytes.
    - c. Natural antibodies can also mediate the rejection of **xenotransplants**, organs exchanged between species (Chapter 17).

III. Additional Ig heterogeneity

A. Slight sequence variations can exist between immunoglobulin molecules of the same isotype (eg, IgG_1) from different individuals.

1. These inherited genetic markers reside within the C_H or C_L domains of human Igs and are called **allotypes**.
2. Antiallotype antibodies are sometimes produced following pregnancy, blood transfusions, or γ-globulin therapy.

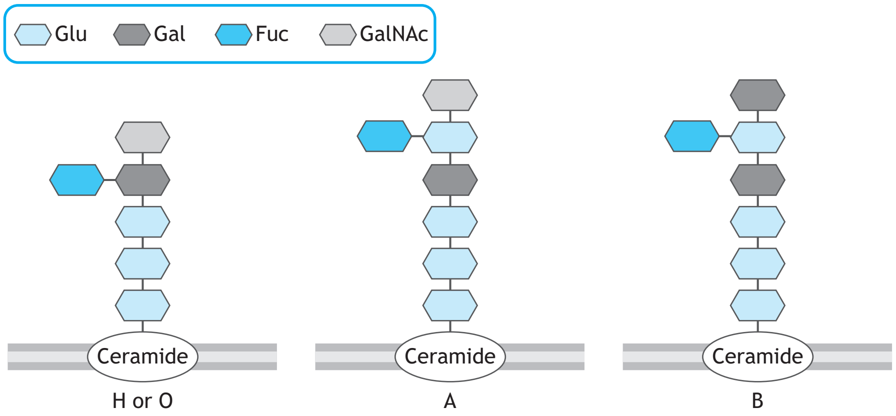

Figure 3–4. Structures of the ABO blood group antigens.

Table 3–3. Antigens and antibodies of the ABO blood group system.

Blood Group Type	Genotypes	Antigens	Isohemagglutinins
A	AA, AO	A	Anti-B
B	BB, BO	B	Anti-A
AB	AB	A and B	None
O	OO	H	Anti-A and anti-B

B. Igs of the same isotype can also be distinguished from one another by their V regions.
1. An antibody of a given specificity has unique $V_L + V_H$ domains, the combination of which specifies its **idiotype**.
2. An antibody molecule and the B cell receptor of the clone of B cells that produced it have the same idiotype (Chapter 2).
3. Antiidiotype antibodies can block the binding of antigen to its antibody.

IV. Monoclonal Igs

A. Monoclonal antibodies are only rarely produced in response to active immunization with natural antigens.

B. In abnormal conditions called **monoclonal gammopathies** a single clone of B cells or plasma cells produces a monoclonal Ig.
1. Malignant monoclonal gammopathies include **multiple myeloma, non-Hodgkin's lymphoma**, and **Waldenström's macroglobulinemia**.
2. Benign monoclonal gammopathies include **light chain disease** and **monoclonal gammopathy of undetermined significance (MGUS)**, a condition that is common in the elderly.
3. **Bence–Jones proteins** are monoclonal Ig L chains that are found in the urine.
   a. The exclusion properties of the renal glomeruli permit passage of L chains, but not H chains.
   b. These Bence–Jones proteins can accumulate in renal tubular epithelial cells, where they can cause necrosis.

ABNORMAL IG SYNTHESIS AND CRYOPATHIES

CLINICAL CORRELATION

- *The excess production of Ig L chains can result in a number of systemic effects, including **cryoglobulinemia**, the microprecipitation of circulating proteins in the cold.*
- *Complications of cryoglobulinemia involving Ig proteins include cyanosis of the digits, hemorrhage, thrombosis, and gangrene.*
- *In B cell malignancies, cryoglobulins are monoclonal and can include intact M components or free L chains.*
- *In addition to treatment of the primary disease (eg, myeloma), plasma exchange and **cryophoresis** (reinfusion of autologous plasma following removal of the cryoglobulins) have met with some success.*

C. **Hybridoma antibody technology** allows production of a monoclonal antibody with a desired specificity.
   1. Hybridomas are produced by fusing a nonsecreting myeloma cell with B cells from an animal that has been immunized with an antigen of interest.
   2. Monoclonal antibody-producing hybridomas with the desired specificity are then identified, selected, and expanded in cell culture.
   3. Hybridoma antibodies show less cross-reactivity than conventional antisera and are preferred for diagnostic tests (Chapter 5).

NOVEL ANTIBODY-MEDIATED THERAPIES

CLINICAL CORRELATION

- ***Immunotoxins*** *are conjugates composed of lethal plant or microbial toxins covalently linked to antibodies that target the conjugate to specific cells or tissues (eg, leukemia cells).*
- ***Bispecific antibodies*** *are genetically engineered to contain Fab regions specific for two different antigens (eg, CD3 on T cells and a tumor target antigen).*
- ***Humanized mouse hybridoma antibodies*** *are genetically engineered to contain mouse Ig hypervariable regions embedded in a human IgG antibody framework.*
- *This minimizes the likelihood that the host will mount an immune response to the foreign mouse hybridoma antibody sequences.*
- ***Fusion proteins*** *produced by covalently linking cytokine receptor domains to Ig molecules have extended half-lives in circulation and have proven effective for the treatment of cytokine-mediated inflammatory diseases.*

CLINICAL PROBLEMS

You are considering treatment options for a 1-year-old child with a congenital immune deficiency. He has very low serum IgG levels and recurrent bacterial infections, and he lacks B cells in his peripheral lymph nodes. You decide to treat with IVIGs, which you realize may be required throughout his life.

1. How often will you have to administer IVIG to maintain a consistent level of protective immunity in this child?

   A. Twice per week
   B. Once per week
   C. Once every 3 weeks
   D. Once every 6 months
   E. Yearly

A child is immunized by the intramuscular route at 2, 4, and 6 months of age with tetanus toxoid, a protein antigen. One week following each immunization her serum is collected and analyzed for antitoxoid antibodies.

2. Which of the following properties would characterize the serum collected after the third immunization compared to that collected after the first immunization?

A. Increased specificity for tetanus toxoid

B. Increased ability of the serum to promote the uptake of the toxoid by phagocytic cells

C. Increased complement activation by antigen–antibody complexes formed with the serum

D. Increased reactivity with autoantigens of the host

E. Decreased binding of the toxoid

An 80-year-old man presents with slight pain on urination, and his physician orders a test for protein in his urine. An elevated urinary protein level prompts the physician to request a urine protein electrophoresis, which shows a protein spike.

3. Which of the following additional findings would be most consistent with a diagnosis of a benign monoclonal gammopathy?

A. A high concentration of serum IgG and elevated β and γ region proteins on protein electrophoresis

B. Elevated concentrations of κ and λ L chains in his serum

C. A high κ:λ ratio (100:1) among his urinary L chains

D. An elevated erythrocyte sedimentation rate

E. Recurrent bacterial infections

Johnny is a 3-month-old infant in need of a blood transfusion. The blood bank has determined that Johnny's parents are both blood group type AB.

4. Which of the following findings would indicate that Johnny has an ABO type different from that of his parents?

A. His serum reacts with type O erythrocytes.

B. His cells react with anti-A.

C. His cells react with both anti-A and anti-B.

D. His serum reacts with his father's cells.

E. His cells react with his mother's serum.

A researcher has produced an idiotype-specific antibody by immunizing a rabbit with an M protein purified from the serum of a myeloma patient. This antibody does not react with any other M protein tested.

5. With which of the following structures would this antibody react?

A. Fab fragment

B. J chain

C. C_H1 domain

D. Fc fragment

E. Fc receptor binding site

ANSWERS

1. The correct answer is C. IVIG is a γ-globulin preparation derived from pooled immune donor plasma and contains primarily IgG, which is what this patient lacks. The half-life of human IgG in the circulation is 23 days (approximately 3 weeks).
2. The correct answer is B. One of the principal reasons for repeated immunizations is to induce Ig class switching (eg, a change from predominantly IgM to predominantly IgG production). This has the effect of introducing new functions associated with the additional isotypes. For example, IgG (but not IgM) is a potent opsonin due to the presence of Fc receptors on phagocytic cells that recognize γ heavy chains.
3. The correct answer is C. To diagnose a monoclonal gammopathy the clinical laboratory must demonstrate a predominance of one L chain (κ or λ) over the other, which would alter the normal 60:40 ratio. Monoclonal gammopathies appear on electrophoresis as a "spike," whereas polyclonal gammopathies appear as an elevated amount of protein across a wide electrophoretic range (ie, the entire γ region). Monoclonal gammopathy of undetermined significance is a common finding in patients in this particular age group.
4. The correct answer is D. Johnny could theoretically be either A, B, or AB. If his serum reacts with his father's cells, he must have either anti-A or anti-B in his serum. Either finding would indicate he has a blood type that is different from his parents, because AB individuals have no antibodies to ABO antigens.
5. The correct answer is A. If this antibody is indeed specific for only this patient's M protein, it must be recognizing the V region of the Ig. The only structure listed that contains V region domains is the Fab fragment.

CHAPTER 4

IMMUNOGLOBULIN GENE EXPRESSION

I. A fundamental problem posed by the unusual structure of immunoglobulin (Ig) molecules is solved by a novel genetic mechanism.

A. Igs are unusual in that millions of different V region sequences exist among the pool of Ig polypeptide chains.

B. These diverse V regions are found within polypeptides that show limited diversity in their C region sequences.

C. The germ line genome of a species is not large enough to encode all of the potential Ig V region sequences.

D. To form the pool of Ig V region sequences a limited number of inherited **Ig gene segments** undergoes genetic recombination and random reassortment.

1. The genetic capacity to generate the diverse family of Ig proteins exists within the germ line of every cell.
2. Rearrangements of Ig DNA by somatic recombination occurs only in B cells and is a random pretranscriptional process.
3. Diversity is generated prior to exposure to foreign antigen.
4. Each member of the species has a similar **potential repertoire** of Igs.

II. Ig genes are initially organized in a germ line configuration.

A. There are three **gene families** (also referred to as **loci**) that encode Ig polypeptides, the κ, λ, and H chain gene families.

1. There are only 10–100 **gene segments** at each locus.
2. Both the κ and λ chain loci contain **constant (C), variable (V),** and **joiner (J) gene segments** (Figure 4–1).
3. The H chain locus has a similar arrangement with a $5'$ cluster of V_H gene segments, several J_H segments, a group of **diversity (D_H)** segments (not seen in the L chain loci), and nine C_H segments.
   - **a.** The C_H gene segments correspond to the nine classes and subclasses of human H chains.
   - **b.** Each of the C_H segments consists of a cluster of three or four exons encoding the different C_H domains of secreted Ig H chains (Figure 3–2).
   - **c.** One or two short exons follow these C_H exon clusters that code the transmembrane and cytoplasmic regions of membrane H chains.
4. Latent promoters are found upstream of each V_κ, V_λ, and V_H segment.

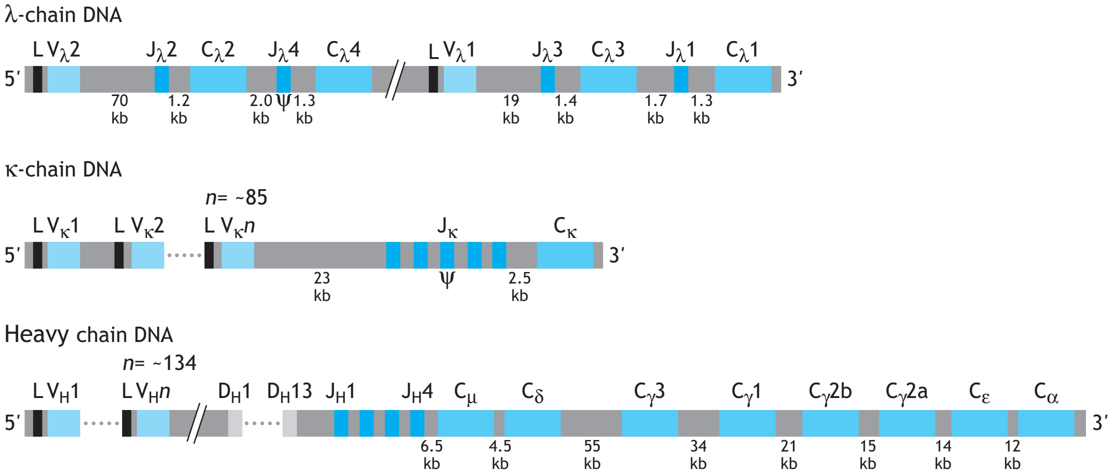
λ-chain DNA
L Vλ2
Jλ2
Cλ2
Jλ4
Cλ4
L Vλ1
Jλ3
Cλ3
Jλ1
Cλ1
5′
3′
70 kb
1.2 kb
2.0 kb
ψ
1.3 kb
19 kb
1.4 kb
1.7 kb
1.3 kb
κ-chain DNA
n= ~85
L Vκ1
L Vκ2
L Vκn
Jκ
Cκ
5′
3′
23 kb
ψ
2.5 kb
Heavy chain DNA
n= ~134
L VH1
L VHn
DH1
DH13
JH1
JH4
Cμ
Cδ
Cγ3
Cγ1
Cγ2b
Cγ2a
Cε
Cα
5′
3′
6.5 kb
4.5 kb
55 kb
34 kb
21 kb
15 kb
14 kb
12 kb

Figure 4–1. Germ line configuration of the three Ig gene families.

Ig LOCUS DELETIONS

CLINICAL CORRELATION

- *Deletion mutations affecting large regions of the human Ig loci are known to exist.*
- *Among a consanguineous population of Tunisians, several individuals with homozygous deletions in the H chain locus have been identified.*
- *One haplotype has a multigene deletion of the entire C_γ1-to-C_ε segment (Figure 4–1).*
- *An affected individual showed normal numbers of B cells in his peripheral lymphoid tissues, but his serum contained no detectable IgG_1, IgG_2, IgG_4, or IgA_1.*
- *In some instances of Ig locus deletions one Ig class can apparently compensate for the absence of another, thus preventing any increased susceptibility to infection.*

III. The recombination of Ig gene segments precedes the expression of a complete Ig gene.

A. Somatic recombination rearranges the order of the various segments within the L chain loci to create a functional Ig gene.

B. Recombination generates diversity within the V regions of L chain polypeptides.
  1. For example, recombination at the κ locus brings one V_κ segment, one J_κ segment, and one C_κ segment into proximity with one another (Figure 4–2).
  2. Recombination involves the cleavage of double-stranded DNA and its religation.
  3. A **V(D)J recombinase** catalyzes cleavage and is expressed only in lymphocytes.
    a. V(D)J recombinase consists of two components encoded by **recombination activating gene 1 (*Rag1*)** and **recombination activating gene 2 (*Rag2*)**.
    b. Mutations within *Rag1* or *Rag2* prevent the rearrangement of the Ig and T cell receptor (TCR) loci.

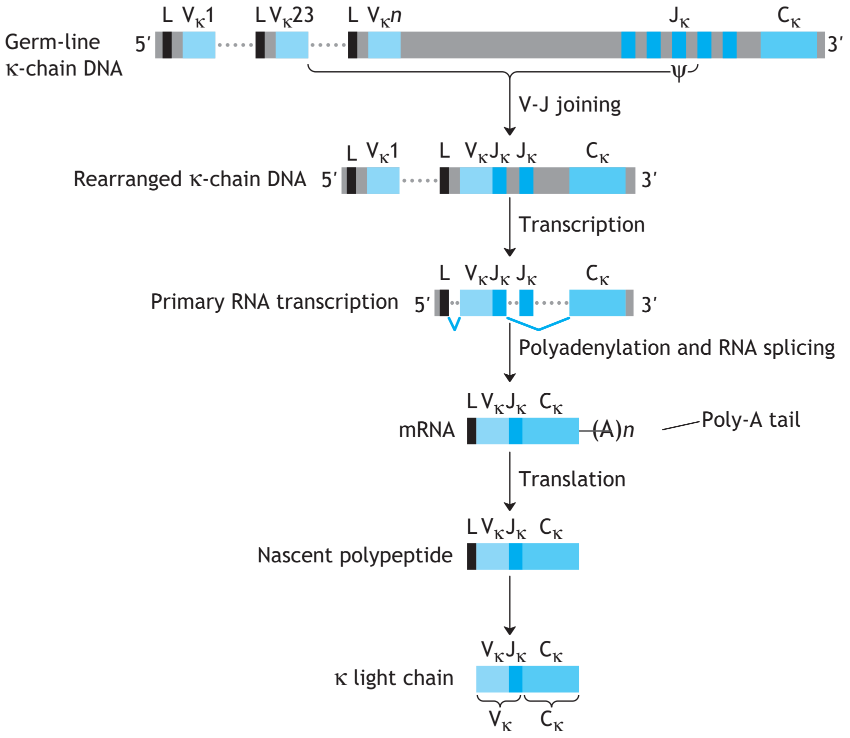
Germ-line
κ-chain DNA
5′
L $V_\kappa 1$
L $V_\kappa 23$
L $V_\kappa n$
J_κ
C_κ
3′
ψ
V-J joining
Rearranged κ-chain DNA
5′
L $V_\kappa 1$
L $V_\kappa J_\kappa$ J_κ
C_κ
3′
Transcription
Primary RNA transcription
5′
L $V_\kappa J_\kappa$ J_κ
C_κ
3′
Polyadenylation and RNA splicing
mRNA
L $V_\kappa J_\kappa$ C_κ
(A)*n*
Poly-A tail
Translation
Nascent polypeptide
L $V_\kappa J_\kappa$ C_κ
κ light chain
$V_\kappa J_\kappa$ C_κ
V_κ
C_κ

Figure 4–2. Somatic recombination and rearrangement of the κ locus.

RAG DEFICIENCY

CLINICAL CORRELATION

- *Mutations in the Rag1 or Rag2 genes dramatically alter lymphocyte differentiation and cause* ***severe combined immune deficiency (SCID)****.*
- *SCID is characterized by* $T^-B^-NK^+$ *lymphopenia, because neither immature B cells nor thymocytes survive differentiation if they cannot successfully express their antigen receptor genes.*
- *SCID is a pediatric emergency due to the rapid onset of fatal opportunistic infections by intracellular pathogens (primarily viral and fungal) early in postnatal life.*
- *Hematopoietic stem cell transplantation in the early neonatal period is a highly successful treatment for SCID.*

4. Sequences within the noncoding introns, called **recombination signal sequences**, determine which gene segments can be juxtaposed by recombination.
5. Additional enzymes catalyze the repair of these double-stranded DNA breaks.
6. The joining of a V segment to a J segment (**V–J joining**) results in the looping out and loss of intervening V and J segments.
7. Any J segments not joined to the V segment as well as intervening introns are excised at the RNA processing stage.
8. The same processes occur at the λ locus.

C. Recombination at the κ and λ loci occurs primarily in the bone marrow independent of any exposure to foreign antigen.

D. Once rearranged, the κ locus can be transcribed into a primary RNA transcript (Figure 4–2).

1. Rearrangement permits the expression of the κ locus by positioning promoters and enhancers close to one another and the transcriptional start site.
2. Transcription ends at a stop codon that follows the rearranged C_κ segment.
3. Intervening RNA and nonjoined J_κ segments are removed by processing the primary RNA transcript.
4. A messenger RNA (mRNA) is produced by splicing and is then polyadenylated and translated into a κ chain polypeptide.

E. The same process permits the expression of a rearranged λ locus.

CHROMOSOMAL TRANSLOCATIONS TO IG LOCI

CLINICAL CORRELATION

- *Many malignant lymphoid cells, including Burkitt s lymphoma and multiple myeloma, show chromosomal translocations that involve the Ig loci.*
- *The high level of genetic recombination that occurs at Ig loci during lymphocyte development makes them particularly susceptible to these genetic abnormalities.*
- *Particularly frequent are translocations that include oncogenes, such as c-myc.*
- *Translocating genes for growth factors, various growth factor receptors, antiapoptosis genes (eg, bcl-2), and transcription factors to the transcriptionally active Ig loci can promote uncontrolled cell proliferation.*

F. The process of gene rearrangement at the H chain locus is slightly different (Figure 4–3).

1. The first rearrangement involves the joining of a D segment with a J segment (**D–J joining**).
2. Following D–J joining, the V region is joined to the D–J segment by a second recombination.
3. Intervening D, J, and V regions are deleted during these recombinations.
4. H locus rearrangement occurs only in pre-B cells (Chapter 9).
5. Transcription yields a primary RNA transcript consisting of joined $V_H D_H J_H C_\mu C_\delta$ segments.
    - **a.** Transcription is initiated by promoters upstream of the joined V_H segment.
    - **b.** The first downstream stop codon for transcription is found downstream of the C_δ segment.
        - ***(1)*** There is no stop codon following the C_μ segment.
        - ***(2)*** Primary transcripts from the rearranged H locus have the capacity to encode complete μ and δ polypeptide chains.
        - ***(3)*** The same $V_H D_H J_H$-encoding region is shared by the μ and δ mRNA (see below).
6. **A primary RNA transcript is synthesized, processed, polyadenylated, and translated into an Ig H chain.**
7. Assembly of H and L polypeptide chains occurs in the rough endoplasmic reticulum.
8. Glycosylation and packaging for secretion occur within the Golgi.

G. The random process of recombination that occurs within each B cell adds to the diversity of the entire B cell pool.

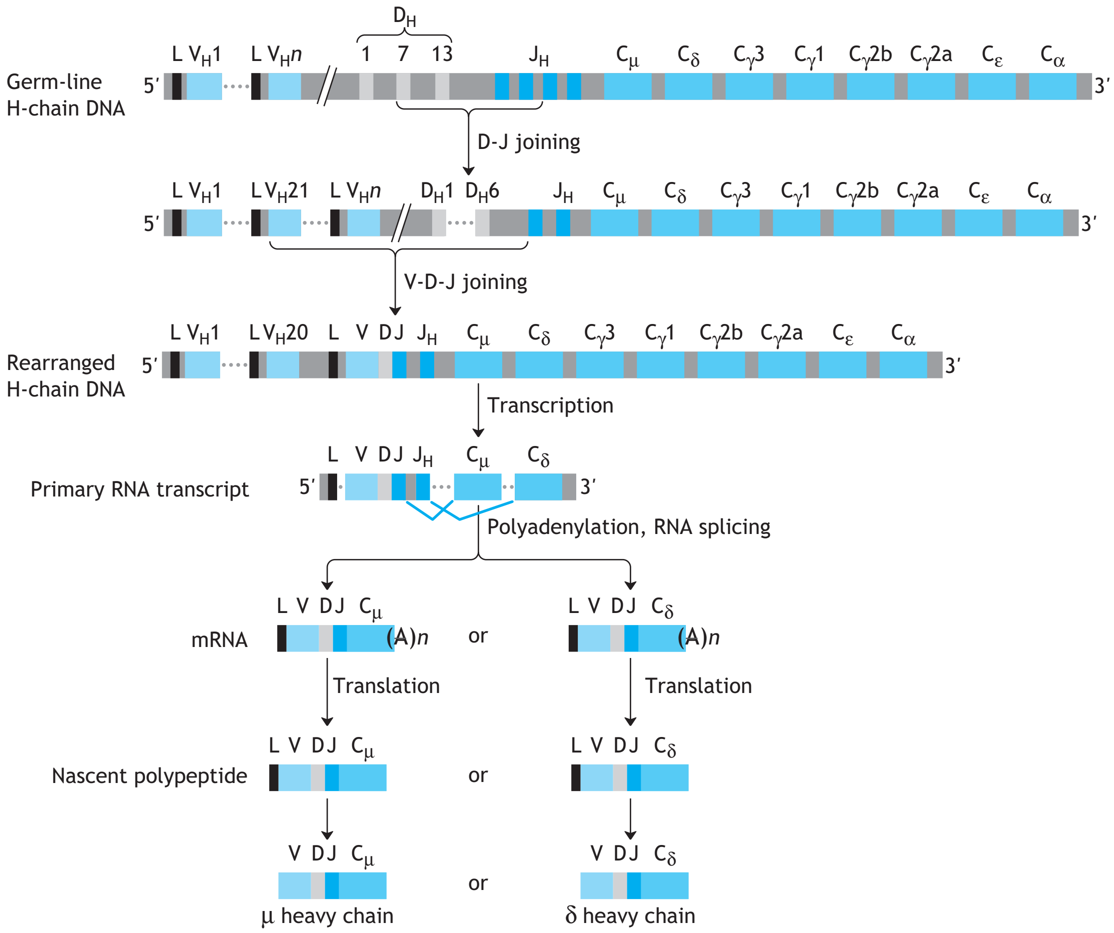
D_H
L V_H1
L V_Hn
1
7
13
J_H
C_μ
C_δ
$C_\gamma 3$
$C_\gamma 1$
$C_\gamma 2b$
$C_\gamma 2a$
C_ε
C_α
Germ-line H-chain DNA
5′
3′
D-J joining
L V_H1
L V_H21
L V_Hn
D_H1
D_H6
J_H
C_μ
C_δ
$C_\gamma 3$
$C_\gamma 1$
$C_\gamma 2b$
$C_\gamma 2a$
C_ε
C_α
5′
3′
V-D-J joining
L V_H1
L V_H20
L V DJ J_H
C_μ
C_δ
$C_\gamma 3$
$C_\gamma 1$
$C_\gamma 2b$
$C_\gamma 2a$
C_ε
C_α
Rearranged H-chain DNA
5′
3′
Transcription
L V DJ J_H
C_μ
C_δ
Primary RNA transcript
5′
3′
Polyadenylation, RNA splicing
L V DJ C_μ
mRNA
(A)n
or
L V DJ C_δ
(A)n
Translation
Translation
L V DJ C_μ
Nascent polypeptide
or
L V DJ C_δ
V DJ C_μ
or
V DJ C_δ
μ heavy chain
δ heavy chain

Figure 4–3. Ig H locus rearrangements.

H. A relatively small pool consisting of several hundred V, D, and J segments recombines to generate a very large potential pool (> 10^{10}) of V region-encoding genes (Table 4–1).

I. **Allelic exclusion** ensures that a single B cell expresses no more than one BCR.

1. Diploid species have an opportunity to express both alleles at a given locus.
2. Successful rearrangements at one allele signal inhibition of rearrangements at the homologous allele.
    a. If rearrangement at the first H-chain locus is nonproductive, rearrangement is attempted at the second H locus.
    b. If the first allele rearranges successfully, rearrangement at the second allele is inhibited.
    c. If both alleles fail to productively rearrange, the cell may undergo apoptosis.

Table 4–1. Various mechanisms contribute to the diversity in Ig molecules.

Mechanism	κ	λ	**H**
A pool of V, D, and J segments	40 V segments 5 J segments	30 V segments 4 J segments	50 V segments 27 D segments 6 J segments
Combinations of VDJ joining	$40 \times 5 = 200$	$30 \times 4 = 120$	$51 \times 27 \times 6 = 8262$
Imprecise joining	Significant amplification	Significant amplification	Significant amplification
Nucleotide addition or deletion	Significant amplification	Significant amplification	Significant amplification
Combinations of H + L polypeptides[a]	(200	+ 120)	$\times 8262 = 2.64 \times 10^6$
BCR editing[b]	+	+	+
Somatic hypermutation[b]	+	+	+

[a] This estimate of the potential Ig repertoire does not include the amplification derived from imprecise joining and nucleotide additions or deletions.
[b] The effects of these processes are difficult to estimate, can depend upon exposure to antigen, and affect both L and H chain gene expression (Chapter 9). BCR, B cell receptor.

- **d.** Successful rearrangement of either κ allele inhibits rearrangements at both of the λ loci.
- **e.** Failing to rearrange both κ alleles signals rearrangement at the λ locus.

IV. Antibody diversity has a genetic basis.

- **A.** The **immune repertoire** of an individual is the sum of his or her Ig and TCR V regions.
- **B.** Several mechanisms contribute to the generation of the **potential Ig repertoire** (Table 4–1).
  - **1.** Hundreds of V, D, and J segments are present in the germ line that can combine to code for Ig V regions.
  - **2.** **Combinatorial association** of these segments in a random fashion adds diversity.
  - **3.** **Imprecise joining (joining flexibility)** can generate several sequence variations spanning the recombined region.
  - **4.** **Nucleotide addition or deletion** can occur at joining sites during recombination.
    - **a.** Nucleotides are added by the enzyme **terminal deoxynucleotidyltransferase (TdT)** to generate **junctional diversity**.

        b. Because the HV3 region of Ig polypeptides is encoded by the sites of V(D)J joining, junctional diversity is greatest within this hypervariable region (Figure 4–3).

    5. **Combinatorial pairing** of H and L polypeptide chains further expands the potential repertoire of Ig molecules.

    6. **Somatic hypermutation** of V region sequences and **BCR editing** (Chapter 9) add to the diversity of mature B cell specificities.

C. The **utilized repertoire** of an individual can be different from the potential repertoire of the species.

1. Many recombinations are nonproductive and signal either cell death or a second recombination event.
2. Some V regions show specificity for self antigens resulting in the deletion of the clone expressing those antigen receptors.
3. Some H and L polypeptide chains do not pair to form functional antigen receptors.
4. At any one time the host may not possess sufficient numbers of B cells to accommodate expression of all potential BCR of the species.
5. Any expansion in a clone of effector or memory B cells can reduce the size of the available repertoire by limiting the space for newly emerging naive B cells with novel specificities.
6. The generation of memory B cells can increase the repertoire by signaling somatic hypermutation of V regions (Chapter 9).

V. A single B cell can express BCRs of different Ig classes.

A. A mature naive B cell in the periphery expresses both μ- and δ-containing membrane BCRs.

1. The primary RNA transcribed from a rearranged H locus gene contains exons for both μ and δ constant regions (Figure 4–4).
    - **a.** Differential processing of the RNA produces mRNAs that code for either μ or δ H chains.
    - **b.** The VDJ sequences of the two mRNA species are the same.
2. These two BCRs differ from one another only in their C_H domains.
    - **a.** The V regions and L chains of the two BCRs are identical.
    - **b.** The idiotypes and specificities of the two BCRs are identical.
3. At different stages in the differentiation of B cells (Chapter 9), RNA processing and translation determine whether membrane IgM, membrane IgD, or both are expressed by the cell.

BICLONAL MYELOMA

- *Very rarely patients with multiple myeloma will synthesize two different M components in their plasma.*
- *These monoclonal Igs belong to different Ig classes (eg, IgM and IgG) but typically have identical V region sequences.*
- *Antiidiotype antibodies specific for the combining site of one M protein react with the combining site of the second M protein.*
- *The disease process of myeloma is thought to begin at the pre-B or immature B cell stage, at which time V(D)J sequences are being established by recombination.*
- *Commitment to the H chain class occurs later, during which these rare multiple isotypic neoplasias arise.*

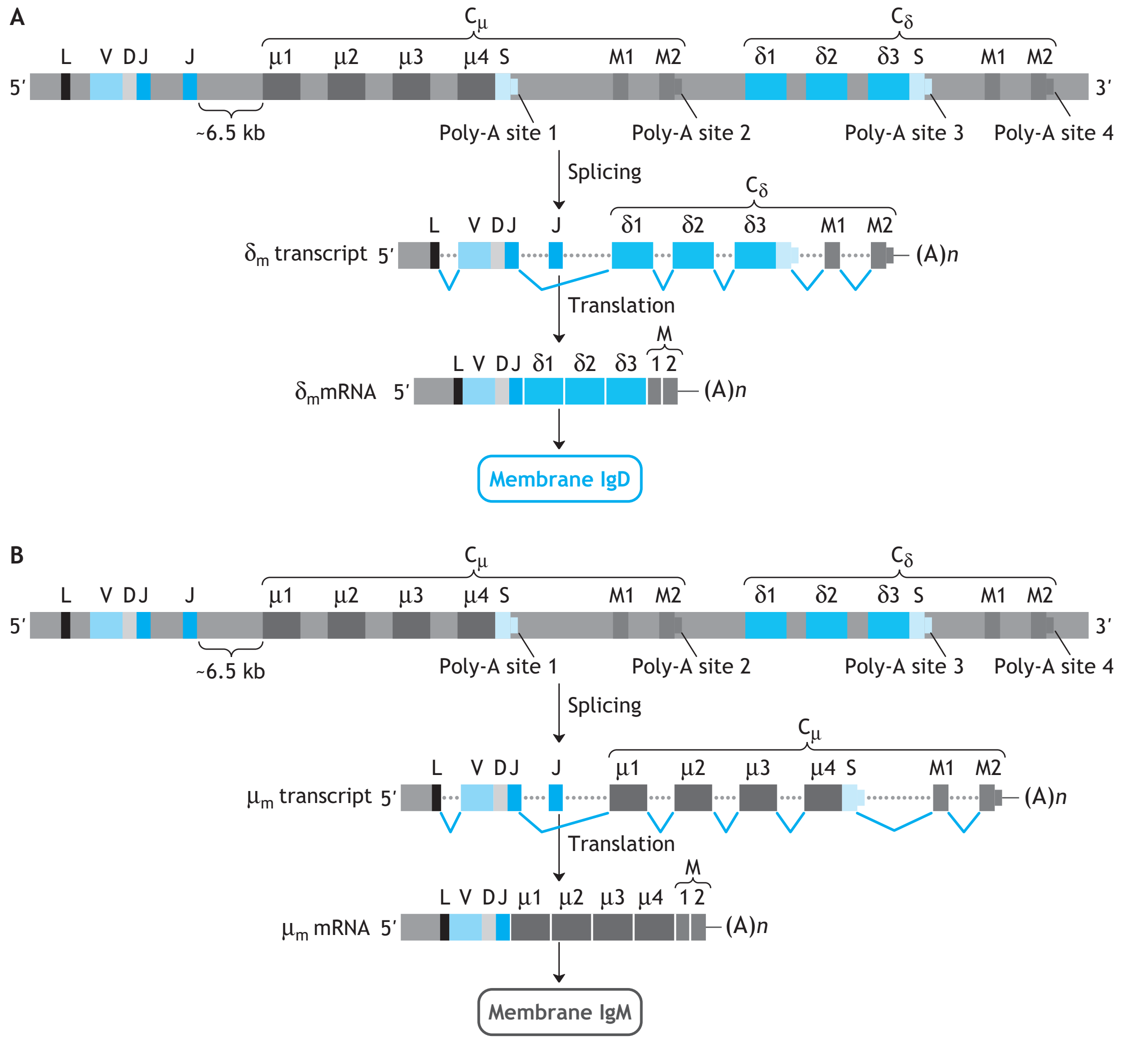
A
Cμ
L V DJ J μ1 μ2 μ3 μ4 S M1 M2
Cδ
δ1 δ2 δ3 S M1 M2
5′
3′
~6.5 kb
Poly-A site 1
Poly-A site 2
Poly-A site 3
Poly-A site 4
Splicing
δm transcript 5′
L V DJ J δ1 δ2 δ3 M1 M2
(A)n
Translation
δmmRNA 5′
L V DJ δ1 δ2 δ3 M 1 2
(A)n
Membrane IgD
B
Cμ
L V DJ J μ1 μ2 μ3 μ4 S M1 M2
Cδ
δ1 δ2 δ3 S M1 M2
5′
3′
~6.5 kb
Poly-A site 1
Poly-A site 2
Poly-A site 3
Poly-A site 4
Splicing
μm transcript 5′
L V DJ J μ1 μ2 μ3 μ4 S M1 M2
(A)n
Translation
μm mRNA 5′
L V DJ μ1 μ2 μ3 μ4 M 1 2
(A)n
Membrane IgM

Figure 4–4. B cell receptors (BCRs) containing μ and/or δ heavy chains.

B. Normal B cells maintain their specificity as they undergo Ig class switching following antigen stimulation.

1. **Class** (or **isotype) switching** refers to a change from the expression of IgM and IgD to the expression of one of the other isotypes (ie, IgG).
   a. Class switching involves a change in the C_H region of the heavy polypeptide chain of the Ig molecule.
   b. During class switching, the L chain remains the same.
   c. The V_H and V_L regions of the Ig, which determine specificity, remain the same during class switching.

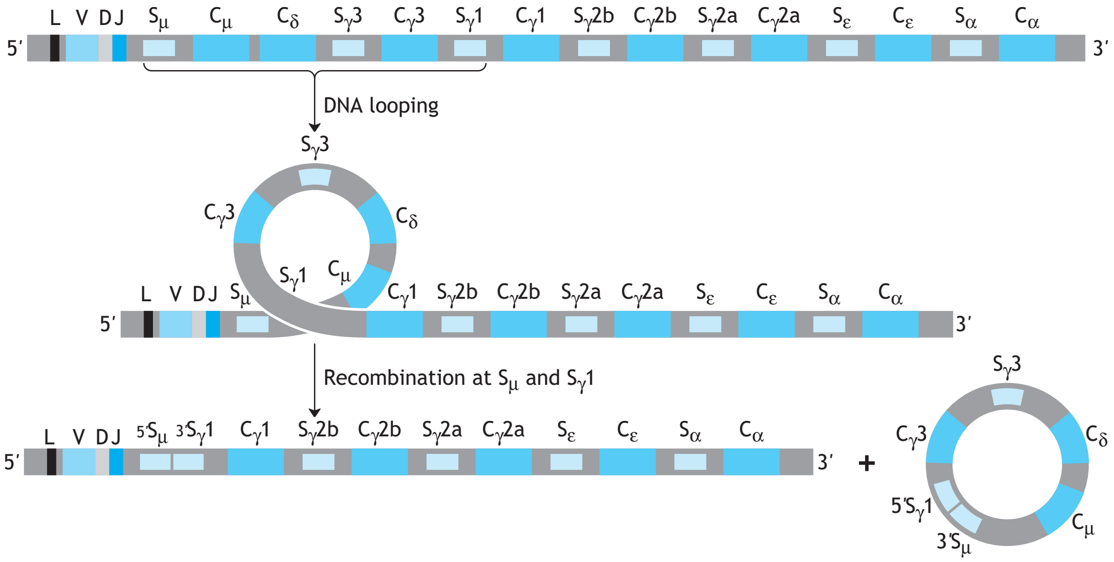
L V DJ S_μ C_μ C_δ $S_\gamma3$ $C_\gamma3$ $S_\gamma1$ $C_\gamma1$ $S_\gamma2b$ $C_\gamma2b$ $S_\gamma2a$ $C_\gamma2a$ S_ε C_ε S_α C_α
5′
3′
DNA looping
$S_\gamma3$
$C_\gamma3$
C_δ
$S_\gamma1$
C_μ
L V DJ S_μ
$C_\gamma1$ $S_\gamma2b$ $C_\gamma2b$ $S_\gamma2a$ $C_\gamma2a$ S_ε C_ε S_α C_α
5′
3′
Recombination at S_μ and $S_\gamma1$
L V DJ $5'S_\mu$ $3'S_\gamma1$ $C_\gamma1$ $S_\gamma2b$ $C_\gamma2b$ $S_\gamma2a$ $C_\gamma2a$ S_ε C_ε S_α C_α
5′
3′
+
$S_\gamma3$
$C_\gamma3$
C_δ
$5'S_\gamma1$
$3'S_\mu$
C_μ

Figure 4–5. Genetic basis for Ig class switching.

2. At the DNA level, Ig class switching involves a change in the order of the C_H segments (Figure 4–5).
   a. DNA between the rearranged VDJ segment and the new C_H segment is looped out by recombination.
   b. Class switching is irreversible, because the intervening DNA is excised and discarded.
   c. Isotype switching always involves a new *downstream* (3′) C_H gene segment.
   d. A switch from IgM and IgD to IgG expression, for example, can be followed by a second switch to another isotype (eg, IgE).

C. Membrane Ig H chains differ from the secreted form of the same polypeptide only in the presence of membrane-spanning domains at their carboxy-termini (Figure 4–4).
   1. A single primary RNA transcript is differentially processed to yield two different mRNAs that encode for either membrane or secreted heavy chains.
   2. All of the Ig isotypes have membrane and secreted forms.
   3. The membrane antigen receptor (BCR) and secreted antibody molecules of a clone of B cells have the same idiotype and specificity.

VI. Defining the genetic basis for Ig diversity and expression explains much of the unique behavior of B cells (Table 4–2).

A. B cells express antigen receptors that are membrane-bound forms of secreted antibody molecules.

B. Allelic exclusion ensures that the BCR expressed by a B cell is monospecific.

Table 4–2. Molecular correlates of the clonal selection theory.

Prediction	Molecular Explanation
Lymphocytes are precommitted in their antigen specificity.	Gene rearrangement occurs in each lymphocyte during its differentiation in the bone marrow and thymus.
A single cell expresses receptors with a uniform antigen specificity.	Rearrangements that lead to BCR[a] or TCR[a] expression are followed by the inhibition of further Ig or TCR locus rearrangements (allelic exclusion).
All antigen receptors on a cell are uniform and identical.	Each antigen receptor variant on a cell uses the same rearranged V region segments.
Daughter cells retain the antigen specificity of the parental cell.	Once expressed, V region genes are relatively stable in lymphocytes.
The antigen receptor of a cell has a specificity identical to the antibody the cell produces.	The same V regions are used to encode for membrane and secreted Ig molecules.

[a] BCR, B cell receptor; TCR, T cell receptor.

C. The antigenic specificity of a B cell is maintained throughout its life span despite clonal expansion, memory cell production, antibody secretion, and isotype switching.

D. T cells utilize many of the same mechanisms to express an even more diverse set of antigen receptors.

CLINICAL PROBLEMS

Mary is a 3-year-old patient who presents with a fever, labored breathing, and shortness of breath. Her history includes recurrent bacterial infections (sinusitis and otitis media) since 1 year of age. She has been prescribed oral antibiotics as often as six times per year. After each course of treatment, the infections subside, but a recurrence of symptoms often follows within several weeks. Laboratory tests indicate a mild neutropenia, but no other hematological abnormalities. Her serum IgG and IgA levels are far below those of age-matched controls, whereas her serum IgM levels are significantly elevated. When immunized with tetanus toxoid, a vaccine she has received before, she makes no detectable IgG antibody response.

1. Which of the following is the most likely cellular or molecular defect in this patient?

A. Failure of her B cells to undergo Ig class switching
B. A block in all B cell differentiation
C. Rag1 deficiency
D. Toll-like receptor deficiency
E. DiGeorge syndrome

A 2-year-old patient presents with a history of infections that suggests an immune defect, and the clinical laboratory reports serum IgG levels one-third that of age-matched controls. The child is ABO blood type A.

2. Which of the following findings would be more consistent with a diagnosis of H locus deletion affecting only C_γ segments than a complete absence of B cells?

A. Decreased IgM levels in her serum
B. The absence of follicles in her splenic white pulp
C. The absence of Ig in her saliva
D. The presence of B cells in the follicles of her lymph nodes
E. The lack of isohemagglutinins in her serum

You have isolated a population of plasma cells from a patient who was recently immunized with a protein vaccine. Her cells secrete IgG in vitro, but lack any membrane IgG.

3. Which of the following accounts for this finding?

A. DNA rearrangement following antigen stimulation deletes BCR-encoding gene segments.
B. The Ig H chain locus does not code for a membrane form of γ chains.
C. Differential RNA processing favors the synthesis of a secreted form of γ heavy chains.
D. Plasma cells that secrete IgG express IgA on their surface.
E. Only surface L chains are expressed by B cells.

You have a patient with an $\mathrm{IgG}_1\kappa$ myeloma. Analysis of DNA from the myeloma cells indicates that the H and κ alleles that he inherited from his father (paternal) have been rearranged.

4. Which of the following loci could still be in the germ line configuration in these cells?

A. Maternal κ allele
B. Maternal λ allele
C. Paternal λ allele
D. Maternal H allele
E. All of the above loci could be in the germ line configuration.

You have a patient with multiple myeloma and examine a preparation of cellular mRNA from the myeloma cells by Northern blotting. The probe you use detects a $V_HD_HJ_H$ sequence that is unique to this particular myeloma cell. The results of this experiment indicate that the myeloma cells have four mRNA species of different sizes, each containing the target sequence.

5. Which one of the following lists of Ig polypeptides is probably encoded by these four mRNAs?
   - A. μ, δ, γ, and α H chains
   - B. κ, λ, μ, and δ chains
   - C. Membrane μ, membrane δ, secreted μ, and secreted δ H chains
   - D. Secreted μ, δ, κ, and λ chains
   - E. Secreted μ, δ, α, and ε H chains

ANSWERS

1. The correct answer is A. Mary produces IgM, which rules out an absence of B cells or a Rag deficiency. The latter would cause a complete absence of both T and B cells, which would have resulted in the onset of viral and fungal infections shortly after birth. Mary's history of infections, serum IgG and IgA levels, and diminished antibody response suggest that she lacks the ability to class switch from IgM production to the more "mature" isotypes (IgG and IgA).
2. The correct answer is D. With γ chain gene deletions, mature B cells are still produced and can be found in the periphery. They simply fail to synthesize IgG antibodies after antigen stimulation. With a complete absence of B cells, follicles in the lymph nodes and spleen fail to develop.
3. The correct answer is C. Whereas mature B lymphocytes express membrane BCR molecules, plasma cells do not normally express membrane Ig. This decision in the B cell lineage to express membrane Ig, secreted Ig, or both is made at the level of RNA processing.
4. The correct answer is E. The H locus rearranges first, followed by either the κ or λ locus. If successful rearrangements had first been undertaken at the paternal H and κ alleles, then the maternal H and κ alleles would remain in the germ line arrangement. Because the κ locus was successfully rearranged in this cell, both λ alleles could have been suppressed and could have remained in the germ line configuration.
5. The correct answer is C. Because the probe detects only mRNAs with VDJ sequences of H chains, light chain mRNA should not be detected. Messenger RNA for μ and δ chains could be expressed by one myeloma cell by differential splicing, and these mRNA would differ only in their C_H regions. μ and δ H chains have different molecular masses, and their membrane forms, which are expressed by a portion of the myeloma cells, are longer than their secreted forms.

CHAPTER 5

ANTIGEN RECOGNITION BY ANTIBODY

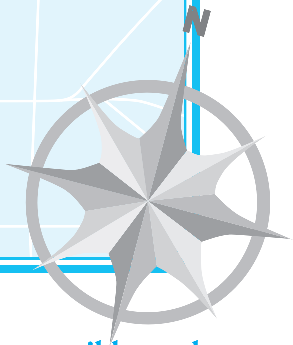

I. Interactions between antibodies and antigens are specific, reversible, and reach a state of equilibrium defined by the following equation:

$$\mathrm{Ab} + \mathrm{Ag} \leftrightarrows \mathrm{AbAg}$$

A. The molecular forces that contribute to the strength of antibody–antigen (Ab–Ag) binding include ionic bonding, hydrophobic interactions, hydrogen bonding, and van der Waals interactions.
  1. **Affinity** is a measure of the strength of binding between an antibody and a monovalent antigen.
    a. Antibody affinities are reflected in **dissociation constants** (K_d) and typically range from 10^{-6} to 10^{-11} M.
    b. The average affinity of an antiserum generally increases with repeated immunizations, a phenomenon referred to as **affinity maturation** (Chapter 9).
    c. As a general rule, immunoglobulin (Ig) G antibodies have higher affinities than IgM antibodies.
  2. **Antibody avidity** refers to the aggregate strength of Ab–Ag binding when multiple antibody-combining sites of an antibody molecule interact with more than one determinant of a multivalent antigen.
    a. High avidity reactions are common with natural antigens that contain multiple copies of the same antigenic determinant (eg, polysaccharides).
    b. Because of their pentameric structure, IgM antibodies generally have a higher avidity than IgG antibodies.

B. The **specificity** of Ab–Ag reactions is similar to the specificity that enzymes show for their substrates.
  1. Antibodies can distinguish between sterioisomeric forms of chemical groups.
  2. **Cross-reactivity** is the tendency of an antibody to react with more than one antigen.
    a. Cross-reactive antigens (eg, bovine and equine serum albumin) share the same or similar antigenic determinants.
    b. In general, **polyclonal antisera** are more cross-reactive than **monoclonal antibodies**.
    c. B cell and T cell receptors (BCR and TCR) can also show cross-reactivity (Chapter 6).

CROSS-REACTIVITY AND CROSS-PROTECTION

CLINICAL CORRELATION

- *Immunity induced by vaccination, infection, or environmental exposure to antigens can lead to either beneficial or deleterious immunological cross-reactions (Table 5–1).*
- *Edward Jenner developed an effective means of eliciting cross-protective immunity to the variola (smallpox) virus by immunizing individuals with vaccinia virus, which causes cowpox.*
- *The nonpathogenic mycobacterium bacille Calmette-Guérin (BCG) induces in vaccinees cross-protective immunity against the causative agent of tuberculosis, Mycobacterium tuberculosis.*
- *Allergic responses to cross-reactive determinants on birch pollen can lead to food allergies to apples, plums, and hazelnuts.*

C. When antibodies bind to their respective antigens, neither is covalently modified, although both can undergo important conformational changes.
  1. Some cell surface **Fcγ receptors** bind IgG Fc regions only after they have bound antigen.
    a. These Fcγ receptors are generally of low affinity (K_d = 10^{-6}–10^{-7} M) and mediate opsonization by phagocytic cells.
    b. High affinity Fcγ and Fcε receptors (K_d = 10^{-8}–10^{-10} M) can bind monomeric IgG and IgE, respectively, in the absence of antigen.
    c. High-affinity Fc receptors mediate placental transfer of monomeric IgG.
  2. The activation of **complement** by IgG and IgM antibodies requires antigen binding that reveals cryptic complement-binding sites within the C_H2 domains (Chapter 8).
  3. **Membrane Igs** that serve as antigen receptors on B cells undergo conformational changes when their antigens are bound.
    a. This permits protein–protein interactions between the C-terminal domains of membrane H chains and two other members of the receptor complex, Igα and Igβ.
    b. BCR–Igα–Igβ interactions initiate transmembrane signaling in B cells (Chapter 9).

Table 5–1. Clinically important immunological crossreactions.

Antibody Specificity	Cross-reactive Antigens	Effect
Microbial carbohydrates	ABO erythrocyte antigens	Mediate transfusion reactions
Steptococcal M protein	Cardiac antigens	Rheumatic fever
Beef heart cardiolipin	Human cardiolipin	False positive tests for syphilis
House dust mite allergens	Shrimp	Food allergies
Timothy grass allergens	June grass allergens	Seasonal allergies
Methyldopa	Rh antigens	Autoimmune hemolytic anemia

D. A complex formed between an antibody and its antigen is often referred to as an **immune complex**.

1. Both circulating and tissue-bound immune complexes have proinflammatory effects (Chapter 13).
2. Circulating soluble immune complexes typically form in antigen excess and deposit in the skin, blood vessel walls, joint spaces, and basement membranes of renal glomeruli.
3. A number of important laboratory tests can detect immune complexes in body fluids or tissues and aid in diagnosing **immune complex diseases** (Chapters 13 and 16).

II. The measurement of Ab–Ag reactions is the basis for many clinical diagnostic tests in immunology and microbiology (Table 5–2).

A. Serology refers to techniques that measure the reactivity of serum antibodies with their antigens.

Table 5–2. Selected applications of immunoassays in the clinical laboratory.[a]

Measurement	Assay	Specific Examples
Ig concentration	Nephelometry	Serum IgG, IgM, IgA, and IgE
Monoclonal gammopathy	Immunoelectrophoresis Immunofixation	M component
Complement		
Individual components	ELISA	C1, C2, C3, C4
Total hemolytic activity	Complement fixation test	CH$_{50}$
Antibodies to infectious agents		
HIV	ELISA, Western blot	Anti-gp24, gp41, gp120, gp160
Streptococcus	Agglutination	Anticapsular polysaccharide
Treponema pallidum	Agglutination	VDRL test for syphilis
Autoimmunity		
Anti-Ig antibody	Passive hemagglutination	Rheumatoid factor
Antinuclear antibodies	Immunofluorescence	ANA, anti-dsDNA
Cell surface markers		
Normal lymphocytes	Flow cytometry	CD4, CD8, CD19
Tumor cells	Flow cytometry	CD10 on leukemia cells

[a] Ig, immunoglobulin; ELISA, enzyme-linked immunosorbent assay; HIV, human immunodeficiency virus; VDRL, Venereal Disease Research Laboratory; ANA, antinuclear antibody; dsDNA, double-stranded DNA.

B. The relative concentration of an antibody is often expressed as a **titer**, which is the reciprocal of the greatest dilution of the antiserum that shows a reaction with its antigen.

C. Particulate antigens agglutinate when combined with their antibodies.
  1. Hemagglutination occurs when antibodies to erythrocyte surface antigens cause aggregation of cells bearing the respective antigen.
    a. In **blood typing**, agglutinating antibodies of known specificities (eg, anti-RhD) are used to determine the antigenic phenotypes of prospective donors and recipients.
    b. In the **major cross-match**, cells from the donor are mixed with serum from the recipient as a means of detecting preformed recipient antidonor antibodies.
  2. The **Coombs test** (Chapter 16), a specialized hemagglutination test, detects antibodies on erythrocytes or platelets.
    a. In the Coombs test, the patient's cells are mixed with the **"Coombs reagent,"** which is an antihuman IgG.
    b. IgG antibody-coated erythrocytes or platelets agglutinate in the presence of the Coombs reagent.
    c. Agglutination can indicate the presence of antibodies against haptens (eg, drugs) bound to the cell surface or antibodies specific for endogenous erythrocyte antigens (eg, anti-RhD).

RH DISEASE OF THE NEWBORN

CLINICAL CORRELATION

- *Rh incompatibility between Rh-negative mothers and their Rh-positive children can lead to the production of antibodies by the mother that cause hemolytic anemia in the fetus.*
- *First pregnancies are generally not affected, because the initial maternal IgM antibodies do not cross the placenta.*
- *Subsequent pregnancies are at risk due to isotype switching to IgG in the mother.*
- *A definitive diagnosis of Rh disease in an affected newborn can be made, in part, with the Coombs test, which detects IgG on the surface of the neonate's erythrocytes.*

  3. The species or strain of a bacterium can be determined by serological agglutination tests.
    a. Serological typing provides a more rapid identification of a bacterial species or group (eg, group A *Streptococcus* or *Salmonella* serogroups) than does culturing the specimen.
    b. Agglutination assays can also reveal serotypes associated with specific diseases (eg, *Escherichia coli* serotype O157:H7 and hemorrhagic colitis).

D. Soluble antigens present at high concentrations form visible precipitates when combined with their antibodies.
  1. Immunoprecipitation techniques are relatively insensitive and require microgram per milliliter to milligram per milliliter concentrations of antigen.
  2. Large insoluble Ab–Ag precipitates can be detected by light scatter using an instrument known as a **nephelometer**.
  3. The amount of immune precipitate that forms is a function of the Ab:Ag ratio (Figure 5–1).
    a. At an optimum ratio, called **equivalence**, the maximum amount of precipitate is observed and neither free Ab nor free Ag is found in solution.

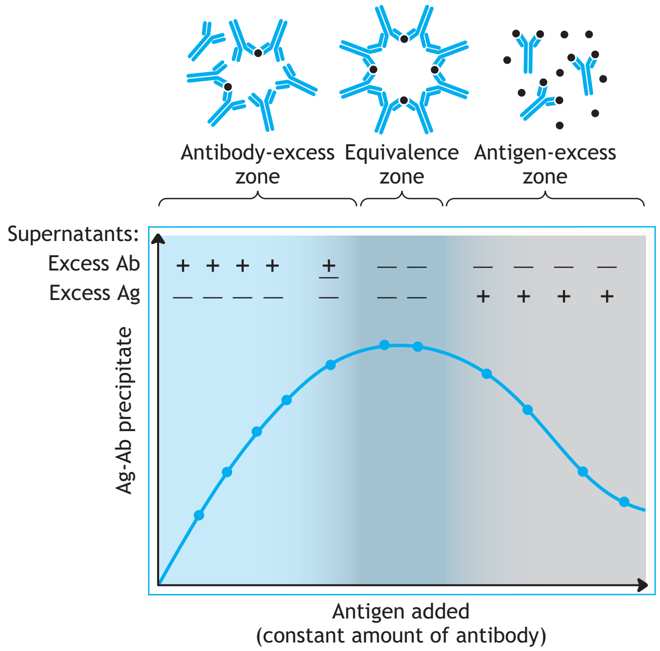

Figure 5–1. The precipitin curve. Ag, antigen; Ab, antibody.

- **b.** In the range of **Ag excess**, small soluble Ab_1Ag_2 complexes are formed.
- **c.** Large complexes are typically cleared by phagocytic cells, whereas smaller soluble complexes can cause **immune complex disease**.

4. Precipitated Ab–Ag complexes can also be visualized in agar gels.

- **a.** In **agar gel immunodiffusion**, solutions of Ab and Ag diffuse toward one another resulting in the formation of a visible precipitate at the diffusion interface (Figure 5–2).
  - ***(1)*** A pattern of **identity** indicates that two antigens are identical with respect to the determinant(s) recognized by the antiserum.
  - ***(2)*** A pattern of **partial identity** occurs when two antigens share some, but not all, of their antigenic determinants.
  - ***(3)*** A pattern of **nonidentity** forms when two antigens have unique determinants not found in the other.
- **b.** **Immunofixation**, which combines the technique of electrophoresis with immunoprecipitation, is used to analyze antigens in complex biological samples, such as serum (Figure 5–3).
  - ***(1)*** Test samples are first electrophoresed in an agar gel.
  - ***(2)*** The electrophoresed proteins are then detected with specific antibodies that precipitate the antigen.
  - ***(3)*** These insoluble immune precipitates are stained with a protein stain.

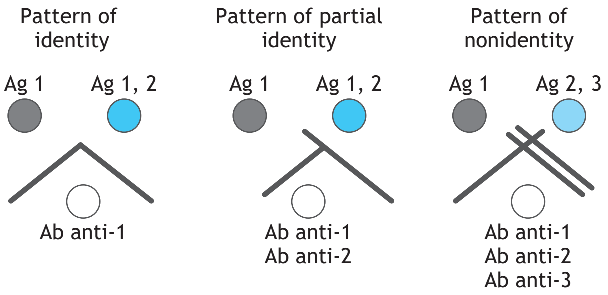
Pattern of identity
Ag 1
Ag 1, 2
Ab anti-1
Pattern of partial identity
Ag 1
Ag 1, 2
Ab anti-1
Ab anti-2
Pattern of nonidentity
Ag 1
Ag 2, 3
Ab anti-1
Ab anti-2
Ab anti-3

Figure 5–2. Agar gel immunodiffusion. Ag, antigen; Ab, antibody.

POLYCLONAL AND MONOCLONAL GAMMOPATHIES

CLINICAL CORRELATION

- ***Hypergammaglobulinemia*** *can occur following repeated immunization, infection, or abnormal B cell differentiation, including neoplasia.*
- *A* ***polyclonal gammopathy*** *generally signals infection, whereas a* ***monoclonal gammopathy*** *indicates a benign or malignant condition affecting a single clone of B cells.*
- *Benign* ***monoclonal gammopathy of undetermined significance (MGUS)*** *is common in the elderly, occurring at a frequency of approximately 3% in persons over 70 years of age.*
- *Because many monoclonal conditions are malignant, it is important to determine whether a hypergammaglobulinemia is monoclonal.*
- *Immunofixation is used to ascertain whether the elevated γ-globulins have a single H chain class and L chain type, which means they are likely monoclonal.*

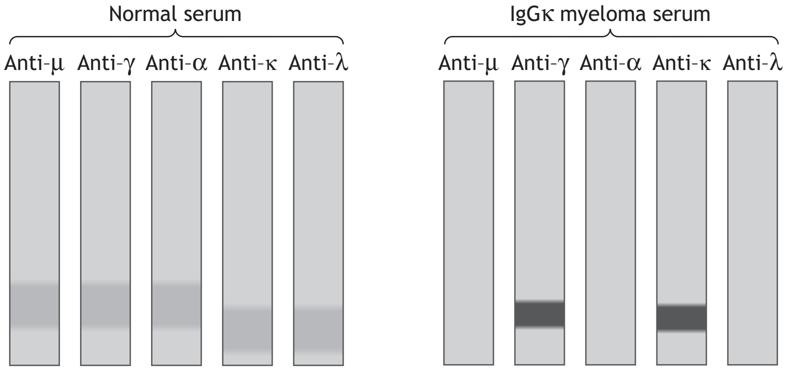
Normal serum
Anti-μ
Anti-γ
Anti-α
Anti-κ
Anti-λ
IgGκ myeloma serum
Anti-μ
Anti-γ
Anti-α
Anti-κ
Anti-λ

Figure 5–3. Detection of a monoclonal gammopathy by immunofixation.

E. Sensitive immunoassays are required to detect antigens that exist at low concentrations (picogram per milliliter to nanogram per milliliter) in clinical samples.

- **1. Radioimmunoassays** use radiolabeled antigens or antibodies.
    - **a.** The competitive radioimmunoassay measures the ability of a test sample (eg, patient serum) to compete with a radiolabeled antigen for binding to an immobilized antibody.
    - **b.** Immunoassays of this type require that dilutions of purified antigen be included as standards.
    - **c.** The **radioallergosorbant test** is a radioimmunoassay that measures the concentration of serum IgE antibody to a known allergen (Chapter 13).
- **2. Enzyme-linked immunosorbent assays (EIA or ELISA)** use enzyme-conjugated antibodies (E-Ab) to detect antigens.
    - **a.** The E-Ab–Ag complexes are detected by their ability to convert a chromogenic substrate into a colored product.
    - **b.** Like radioimmunoassay (RIA), the ELISA can be used to measure antigen or antibody concentrations and shows sensitivity for antigens in solution in the picogram per milliliter range.
    - **c.** Many **home pregnancy tests** use ELISA technology to detect human chorionic gonadotropin in the urine.
- **3. Immunoblotting (Western blotting)** uses either radiolabeled or enzyme-conjugated antibodies to detect antigens within a complex sample, such as serum or a tissue extract.
    - **a.** The sample is first electrophoresed in a polyacrylamide gel under conditions that separate antigens according to their molecular weights.
    - **b.** Separated proteins are then transferred from the gel to a membrane.
    - **c.** The membrane is treated with an E-Ab.
    - **d.** E-Ab–Ag complexes are detected by incubating the membrane with a chromogenic substrate, which forms an insoluble colored precipitate on the complex.
    - **e.** Western blots are often used to confirm initial screening tests, such as an ELISA.

F. Immunofluorescence assays use fluorescenated antibodies to detect tissue- or cell-bound antigens.

- **1.** Because of their specificity, the most popular antibodies for this purpose are monoclonal antibodies.
- **2.** The availability of multiple fluorochromes (dyes) enables polychromatic analysis of more than one antigen at a time.
- **3.** Specialized equipment (eg, a fluorescence or confocal microscope or flow cytometer) is required to detect the fluorescence emitted from the stained sample.
- **4. Tissue immunofluorescence** is routinely used to enumerate certain markers in tissue sections (eg, tumors).
- **5. Flow cytometry** is a specialized application of immunofluorescence in which individual cells in suspension are analyzed for the expression of a given marker(s)(Figure 5–4).
    - **a.** The cell suspension (eg, blood) is first treated with the fluorescenated antibody to label the cell subset of interest.
    - **b.** The labeled cell suspension is then sprayed across the path of a laser to excite the fluorochrome.

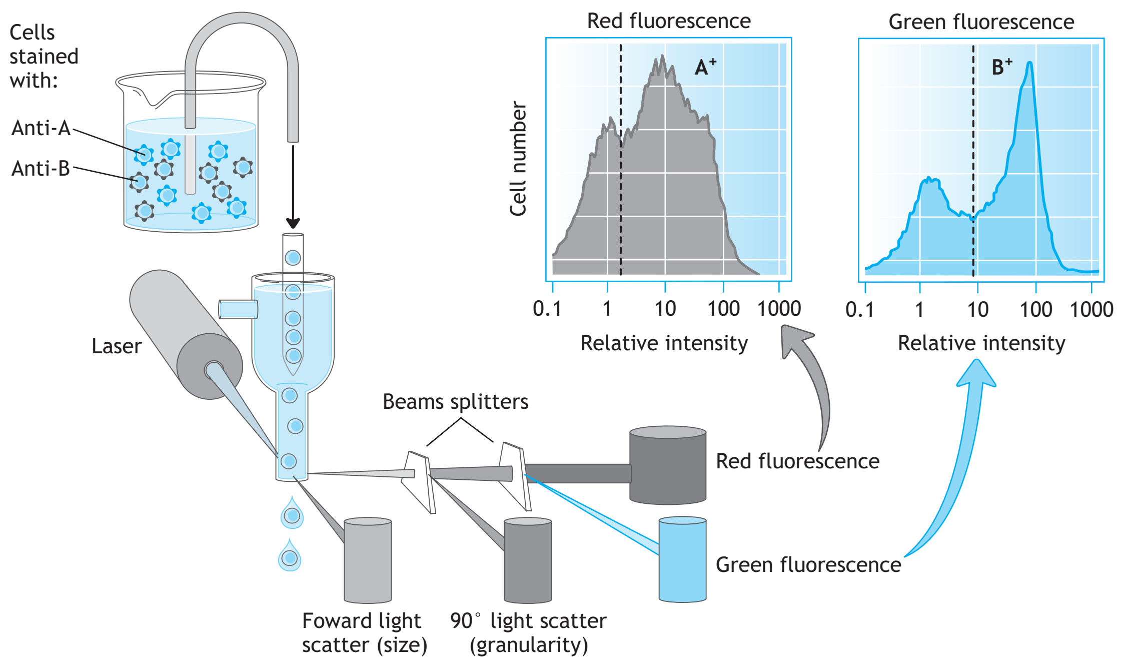
Cells
stained
with:
Anti-A
Anti-B
Laser
Red fluorescence
Green fluorescence
Cell number
A+
B+
0.1 1 10 100 1000
Relative intensity
0.1 1 10 100 1000
Relative intensity
Beams splitters
Red fluorescence
Green fluorescence
Foward light
scatter (size)
90° light scatter
(granularity)

Figure 5–4. Schematic of a flow cytometer.

c. Individual cells are scored for their intensity of fluorescence emission.
d. Flow cytometry is the preferred technique for enumerating CD markers on human cells (Appendix I).
e. **Polychromatic flow cytometry** simultaneously detects multiple markers on a single cell (Figure 5–5).
f. Individual cells expressing these markers can be physically purified by a modification of flow cytometry called **fluorescence-activated cell sorting (FACS)**.

PHENOTYPING LYMPHOMAS AND LEUKEMIAS BY FLOW CYTOMETRY

CLINICAL CORRELATION

- *An important application of flow cytometry is the phenotyping of hematological malignancies using monoclonal antibodies to various CD markers.*
- *This approach can establish the number of different cell populations present in a sample, their lineage, and their stage of differentiation.*
- *For example, a typical B cell chronic lymphocytic leukemia (B cell-CLL) expresses membrane Ig, CD5, CD19, CD20, and CD22.*
- *By comparison, a plasmacytoma typically does not express membrane Ig, CD19, CD20, or CD22, but contains intracytoplasmic Ig.*

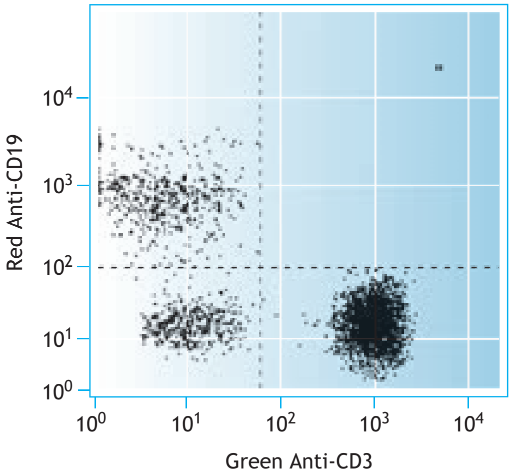

Figure 5–5. Multicolor flow cytometry analysis of human blood lymphocytes.

CLINICAL PROBLEMS

A 1-year-old child with a history of chronic sinusitis, otitis media, and pneumonia is suspected of having a primary immune deficiency. He has benefited from antibiotics and intravenous immunoglobulins and shows a very low concentration of serum IgG and no IgM. Analysis of the patient's lymph node cells by flow cytometry shows normal numbers of $CD3^+$ lymphocytes, but an absence of cells expressing membrane IgM.

1. Which of the following diagnoses is most consistent with the data presented for this patient?

   A. Severe combined immune deficiency (SCID)
   B. X-linked agammaglobulinemia (XLA)
   C. Selective IgG deficiency
   D. Light chain disease
   E. B cell chronic lymphocytic leukemia

You decide to determine whether the patient in question 1 can mount a secondary antibody response to tetanus toxoid. Ten days after immunizing with the toxoid, you measure the titer of serum antibodies to the foreign protein.

2. Which of the following laboratory tests would be most appropriate for measuring this response?

   A. Flow cytometry
   B. ELISA
   C. Immunofluorescence

D. Immunoblotting (Western blotting)

E. Immunodiffusion in gels

You have a patient with recurrent infections and decide to examine a biopsy of her lymph node for potential structural abnormalities. Immunohistological staining reveals very few cells in the diffuse cortex that stain with a peroxidase-conjugated antibody to CD3 followed by a chromogenic peroxidase substrate. Unstained cells are present in the superficial cortex of the lymph node.

3. Which of the following diseases is consistent with this finding?

A. Toll-like receptor deficiency

B. Severe combined immunodeficiency (SCID)

C. Chronic granulomatous disease

D. DiGeorge syndrome

E. Multiple myeloma

A 32-year-old patient presents with lymphopenia, but no other hematological abnormalities. He has mucocutaneous candidiasis, a yeast infection of the mucous membranes, and is seronegative for HIV-1 by ELISA.

4. Which of the following flow cytometry findings with blood lymphocytes [peripheral blood mononuclear cells (PMBCs)] would nonetheless still support a diagnosis of HIV infection in this patient?

A

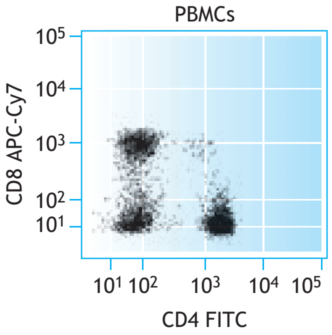

B

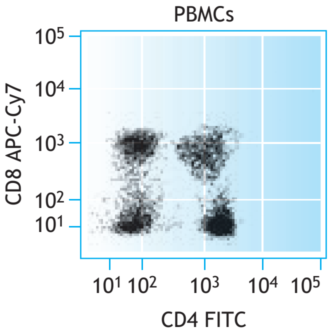

C

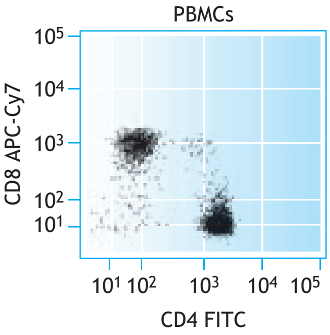

D

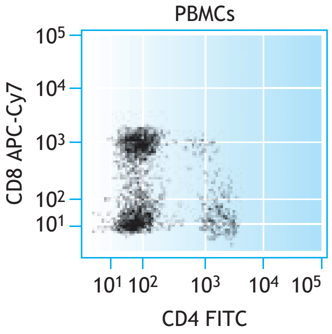

E

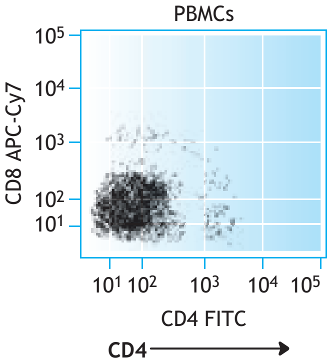

A. Panel A
B. Panel B
C. Panel C
D. Panel D
E. Panel E

A patient with multiple myeloma shows proteinuria, and you wish to characterize the urinary protein. Using immunodiffusion, you react a rabbit antiserum specific for human IgG with the patient's serum and her urine.

5. Which of the patterns shown below would suggest that the urinary protein is an Ig L chain?

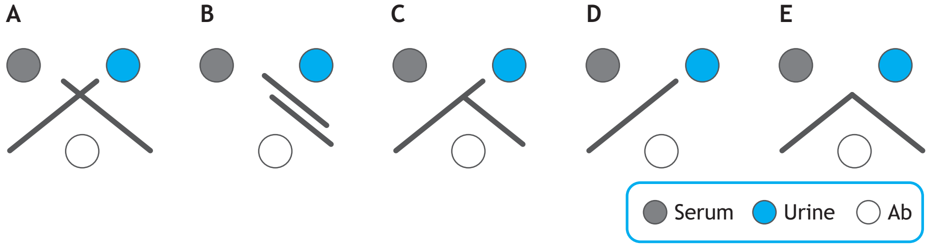

A. Pattern A
B. Pattern B
C. Pattern C
D. Pattern D
E. Pattern E

ANSWERS

1. The correct answer is B. The history and flow cytometry data are consistent with an antibody deficiency that results from the absence of peripheral B cells. Although SCID would also show aberrant B cell differentiation, T cells would be absent in this disease as well.

2. The correct answer is B. Among the choices, only ELISA is a quantitative assay. Immunoblotting is not quantitative, and immunodiffusion requires high concentrations of antibody and antigen to visualize immune precipitates.

3. The correct answer is D. DiGeorge syndrome (thymic dysplasia) results in a selective deficiency of T cells and an acellular appearance in T cell-dependent areas of the lymph nodes.

4. The correct answer is D. There is a selective depletion of CD4$^+$ T cells in patients with HIV infections. For a comparison with the normal pattern, see panel A. The pattern in panel B is not typical for blood because it contains double-positive cells, which are normally seen only in the thymus.

5. The correct answer is C. The excess urinary proteins in myeloma are generally Ig L chains (Bence–Jones proteins), which would show a cross-reaction with the patient's serum monoclonal Ig (M component). Because the serum Ig would contain antigenic determinants on their H chains not found on the urinary L chains, a pattern of partial identity would appear.

CHAPTER 6

T CELL RECOGNITION OF AND RESPONSE TO ANTIGEN

I. The T cell receptor (TCR) for antigen is a complex consisting of an antigen-specific heterodimer (TCR) and four additional peptides of the CD3 family (Figure 2–5).

A. TCR structure is similar to that of the B cell receptor (BCR) (Table 6–1).

1. The α, β, γ, and δ polypeptides are transmembrane proteins with short cytoplasmic domains.
2. TCR peptides have amino acid sequence homology with Ig peptides and belong to the **immunoglobulin gene superfamily**.
3. Each TCR peptide has a V domain and a C domain (eg, $V_\alpha C_\alpha$ in α chains), which are stabilized by intrachain disulfide bonds.
4. The two V domains (eg, $V_\alpha + V_\beta$) form a monovalent ligand-binding site much like an Ig Fab fragment.
5. The membrane-distal face of the TCRαβ and TCRγδ receptors that contact their ligands has a flat conformation, rather than the cleft seen in immunoglobulins (Igs) (Figure 6–1).
6. Of peripheral T cells 95% bear TCRs composed of αβ heterodimers; the remaining T cells bear TCRs composed of γδ heterodimers.
7. **CD3 γ, δ, ε, and ζ** polypeptide chains are transmembrane signaling components of the receptor complex.
8. The TCR is never expressed without CD3 peptides; the complete CD3 complex is never expressed on the cell surface without a TCR heterodimer.

CD3 DEFICIENCY

CLINICAL CORRELATION

- *Rare immune deficiencies affecting the expression of the TCR result from natural mutations in the CD3 γ and ε chains.*
- *These patients present with severe lymphopenia at birth due to an inability of their thymocytes to signal through an early form of the TCR called the* ***pre-TCR****, which requires CD3 for its expression.*
- *Without a functional pre-TCR, immature thymocytes do not develop properly and cellular immunity is impaired (Chapter 10).*
- *Opportunistic viral and fungal infections occur early in life and require aggressive treatment.*

B. The expression of TCR genes requires their rearrangement by recombination during cell differentiation in the thymus (Figure 6–2).

Table 6–1. Distinguishing features of the BCR and TCR.[a]

Feature	BCR	TCR
Antigen-specific component	Membrane Ig	αβ or γδ heterodimer
Receptor valence	Bivalent	Monovalent
Signaling peptides	Igα, Igβ	CD3 γ, δ, ε, ζ
Ligand	Native epitopes of proteins, carbohydrates, lipids, nucleic acids	Processed peptides plus MHC or glycolipids plus CD1
Isotype switching	Yes	No

[a] BCR, B cell receptor; TCR, T cell receptor; Ig, immunoglobulin; MHC, major histocompatibility complex.

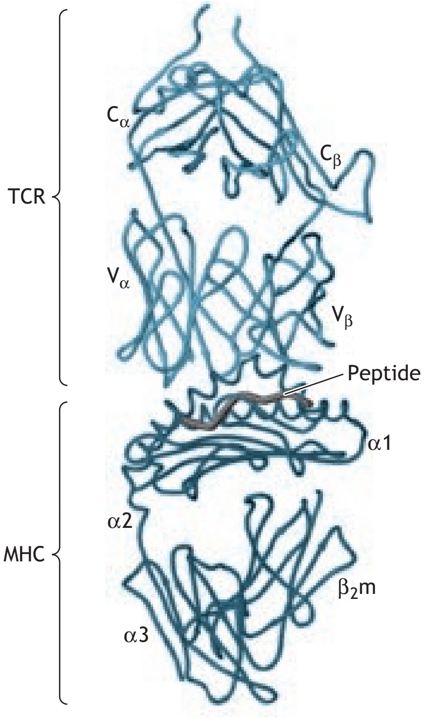

Figure 6–1. Model of a typical T cell receptor (TCR) showing the structure of the ligand-binding surface. MHC, major histocompatibility complex. (Adapted, with permission, from Garcia KC: An αβT cell receptor structure at 2.5 A and its Orientation in the TCR-MHC complex. Science 1996;274:209.)

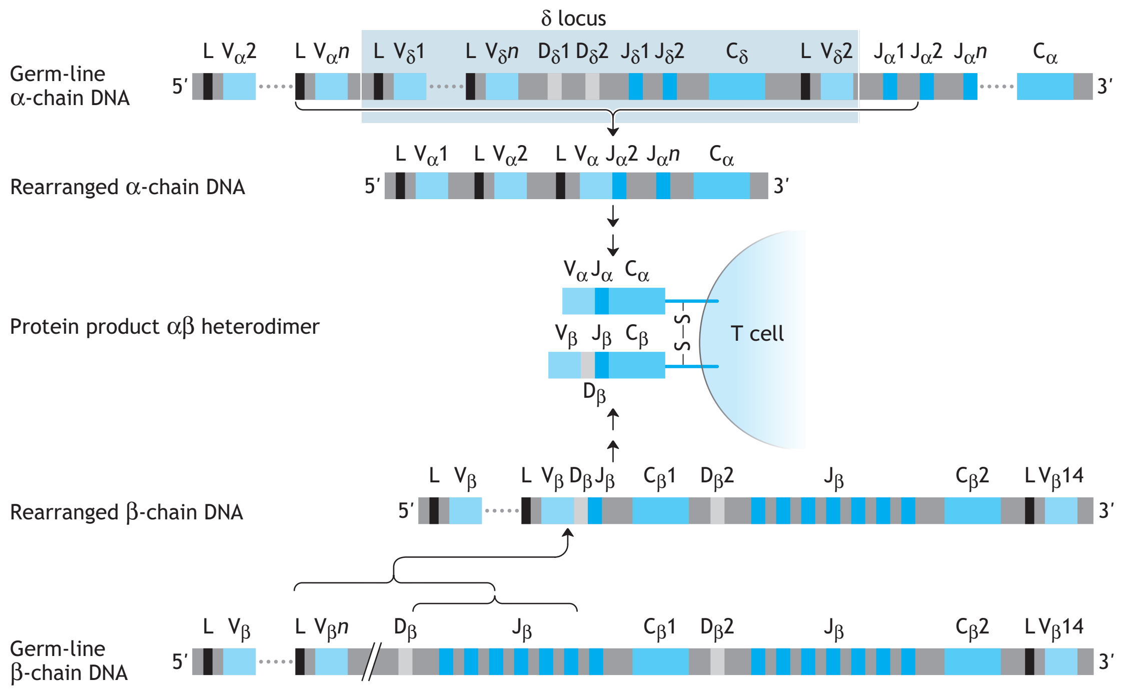
δ locus
Germ-line α-chain DNA
5′
L Vα2
L Vαn
L Vδ1
L Vδn
Dδ1 Dδ2
Jδ1 Jδ2
Cδ
L Vδ2
Jα1 Jα2 Jαn
Cα
3′
Rearranged α-chain DNA
5′
L Vα1
L Vα2
L Vα Jα2 Jαn
Cα
3′
Protein product αβ heterodimer
Vα Jα Cα
Vβ Jβ Cβ
Dβ
S–S
T cell
Rearranged β-chain DNA
5′
L Vβ
L Vβ Dβ Jβ
Cβ1
Dβ2
Jβ
Cβ2
L Vβ14
3′
Germ-line β-chain DNA
5′
L Vβ
L Vβn
Dβ
Jβ
Cβ1
Dβ2
Jβ
Cβ2
L Vβ14
3′

Figure 6–2. Germ line and rearranged α and β T cell receptor (TCR) gene loci.

1. The TCR genes are found at three genetic loci, α/δ, β, and γ.
2. Variable (V), joining (J), and constant (C) gene segments are found at each locus (Table 6–2).
3. Diversity (D) segments are present within the β and δ loci.

C. The diversity of the TCR repertoire arises from genetic mechanisms that are similar to those used to create a diverse set of Ig molecules.

Table 6–2. TCR locus gene segments.[a]

Segment	TCRα	TCRβ	TCRγ	TCRδ
V	50	57	14	3
D	0	2	0	3
J	70	13	5	3
C	1	2	2	1

[a] TCR, T cell receptor.

1. There are several hundred V, D, and J segments that recombine randomly to create a pool of rearranged TCR V genes.
2. Imprecise joining and nucleotide addition add diversity.
3. Combinatorial association between the α and β or the γ and δ chains further expands the repertoire.
4. TCR editing occurs infrequently, and there is no somatic hypermutation of TCR V region sequences as there is for BCR V region genes.

TCR REARRANGEMENTS IN CANCER

CLINICAL CORRELATION

- *The molecular diagnosis of human T cell leukemias and lymphomas is greatly aided by showing that they are monoclonal and express a TCR with a uniform V_α or V_β region.*
- *This can be accomplished by molecular techniques, such as the polymerase chain reaction (PCR), that demonstrate the proliferating cells have rearranged a uniform TCR V_α or V_β segment.*
- *This type of molecular approach aids in early diagnosis and monitoring the status of the neoplastic clone during therapy.*

D. The natural ligand for the TCR is a foreign peptide plus a host MHC molecule.
  1. The TCR recognizes peptides derived from foreign protein antigens that are bound to a major histocompatibility complex (MHC) molecule and presented by an antigen-presenting cell (APC).
  2. The complementarity determining regions (CDRs) of the TCR make contact with amino acid residues of both the peptide and the MHC molecule (Figure 6–1).

E. **Superantigens** produced by microbes can bind the TCR and the MHC outside the peptide-binding site.
  1. *Staphylococcus* and *Streptococcus* species produce exotoxins that act as superantigens.
  2. Microbial superantigens cross-link TCR V_β regions to MHC class II molecules.
  3. Binding of superantigens to the TCR induces polyclonal activation of T cells.

TOXIC SHOCK SYNDROME

CLINICAL CORRELATION

- ***Toxic shock syndrome (TSS)** is a systemic disease associated with vaginal or surgical wound infections with the gram-positive bacterium Staphylococcus aureus.*
- *S aureus produces a superantigen, called **toxic shock syndrome toxin (TSST)**, which activates $CD4^+$ T cells to produce massive quantities of inflammatory cytokines.*
- *The systemic nature of the inflammatory response manifests as fever, coagulopathies, extreme hypotension, skin rash and exfoliation, and diarrhea.*

F. **Natural killer T (NKT) cells** express an unusual form of the TCR.
  1. **Invariant NKT cells (iNKT cells)** express a uniform α chain composed of a single V_α region called $V_\alpha 24$.
  2. The β chain of the iNKT cell TCR is created by recombination using only a subset of the V_β gene segments.
  3. The known ligands of the iNKT cell TCR are glycolipids that are presented by CD1 molecules, not MHC molecules (Chapter 7).
  4. iNKT cells produce cytokines and become cytotoxic when activated through their TCR.

II. Optimum T cell activation by antigen requires TCR stimulation and signals derived from the APC.

- **A.** Compared to antibodies, TCRs bind their ligands with relatively low affinity (K_d 10^4–10^5 M).
- **B.** The binding between a T cell and its APC is strengthened by the T cell **adhesion molecules** LFA-1 and CD2 (Table 6–3).
- **C.** **Coreceptors** on T cells augment the intracellular signal transduction pathways that are initiated by the TCR–CD3 complex.
  - **1.** Many coreceptors promote TCR signaling by phosphorylating **immunotyrosine activation motifs (ITAMs)** of CD3.
  - **2.** T cells that express TCR specific for peptide-MHC *class II* complexes use **CD4** as a signaling coreceptor.
  - **3.** T cells that express TCR specific for peptide-MHC *class I* complexes use **CD8** as a coreceptor.
  - **4.** **CD28** provides important coactivation signaling for T cells.
  - **5.** **CD154** is a ligand for the CD40 coreceptor found on B cells.
- **D.** **Cytokines** provide a second source of costimulatory signals (Chapter 12).

HYPER IGM SYNDROME

CLINICAL CORRELATION

- *Children with X-linked **hyper IgM syndrome** show elevated levels of serum IgM, but extremely low concentrations of IgG, IgA, and IgE.*
- *Their activated T cells lack **CD154**, which results in impaired Th cell function and B cell signaling.*
- *Impaired Ig class switching in B cells results in elevated IgM antibody levels.*
- *With the failure to produce opsonizing IgG antibodies, infections with pyogenic (pus-forming) bacteria are common in these patients.*

Table 6–3. Adhesion molecules and coreceptors on T cells.[a]

T Cell Coreceptor	Ligand on APC	Cells That Express the Receptor Ligand
LFA-1	ICAM-1	Endothelial cells, many others
CD2	CD58	Many cells
CD4	MHC class II	Dendritic cells, B cells, macrophages
CD8	MHC class I	Nearly all nucleated cells
CD28	B7	Dendritic cells, activated B cells, macrophages
CD154	CD40	Dendritic cells, B cells, macrophages, others

[a] APC, antigen-presenting cell; LFA-1, lymphocyte function-associated antigen 1; ICAM, intercellular adhesion molecule; MHC, major histocompatibility complex.

III. Signaling through the TCR leads to gene transcription.

A. The binding of the TCR to it ligand recruits the CD3 peptides to the αβ chains via charge interactions in their transmembrane domains.

B. **ITAMs** in the cytoplasmic tails of CD3 peptides are sites for the assembly of signaling intermediates, including the kinase **ζ-associated protein-70 kDa (ZAP-70)** (Figure 6–3).

1. ZAP-70 becomes phosphorylated and, in turn, phosphorylates **phospholipase Cγ (PLCγ)**.
    - a. PLCγ hydrolyzes the membrane phospholipid phosphatidylinositol 4,5-biphosphate (PIP$_2$) producing **diacylglyceride (DAG)** and **inositol 1,4,5-triphosphate (IP$_3$)**.
        - *(1)* DAG activates **protein kinase C (PKC)**, which activates the **nuclear factor-κB (NFκB)** transcription factor pathway (Figure 6–4).
        - *(2)* IP$_3$ mobilizes intracellular Ca^{2+}, which activates **calmodulin**.
        - *(3)* Ca^{2+}/calmodulin-dependent enzymes, including the phosphatase **calcineurin**, are activated.
        - *(4)* Calcineurin mediates *de*phosphorylation of the latent transcription factor **nuclear factor of activated T cells (NFAT)**.
2. ZAP-70 also phosphorylates adapter proteins that lead to the activation of the G protein-mediated Ras/Rac pathways.
    - a. **Mitogen-activated protein (MAP) kinases** are then activated.

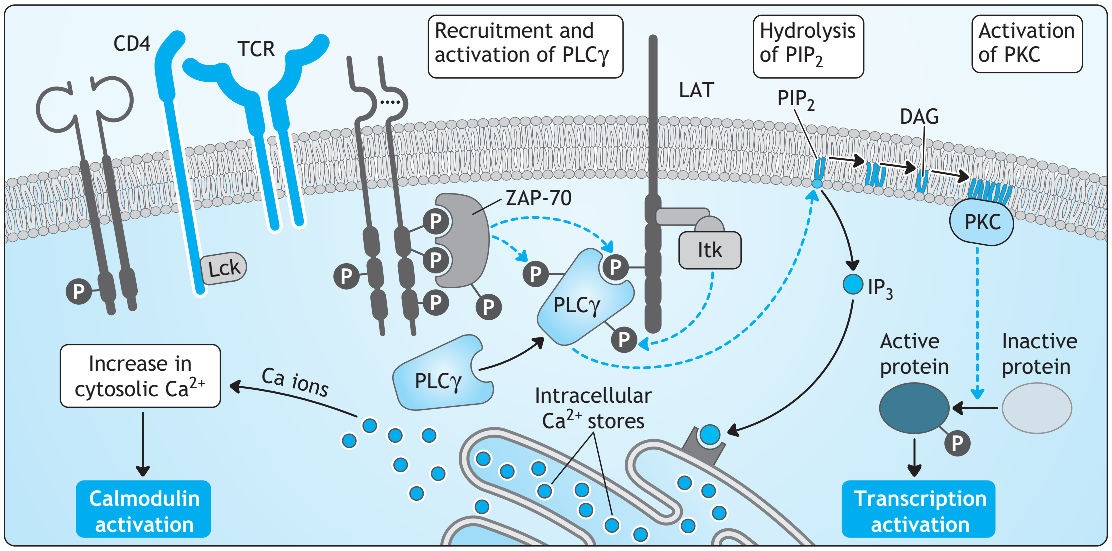
CD4
TCR
Recruitment and activation of PLCγ
LAT
Hydrolysis of PIP$_2$
Activation of PKC
PIP$_2$
DAG
ZAP-70
P
Lck
Itk
PLCγ
PKC
IP$_3$
Increase in cytosolic Ca^{2+}
Ca ions
PLCγ
Intracellular Ca^{2+} stores
Active protein
Inactive protein
Calmodulin activation
Transcription activation

Figure 6–3. Intracellular signaling events accompanying T cell activation. PLC, phospholipase C; PIP$_2$, phosphatidylinositol 4,5-biphosphate; PKC, protein kinase C; ZAP-70, ζ-associated protein-70 kDa; LAT, linker for activation of T cells; DAG, diacylglyceride; IP$_3$, inositol 1,4,5-triphosphate.

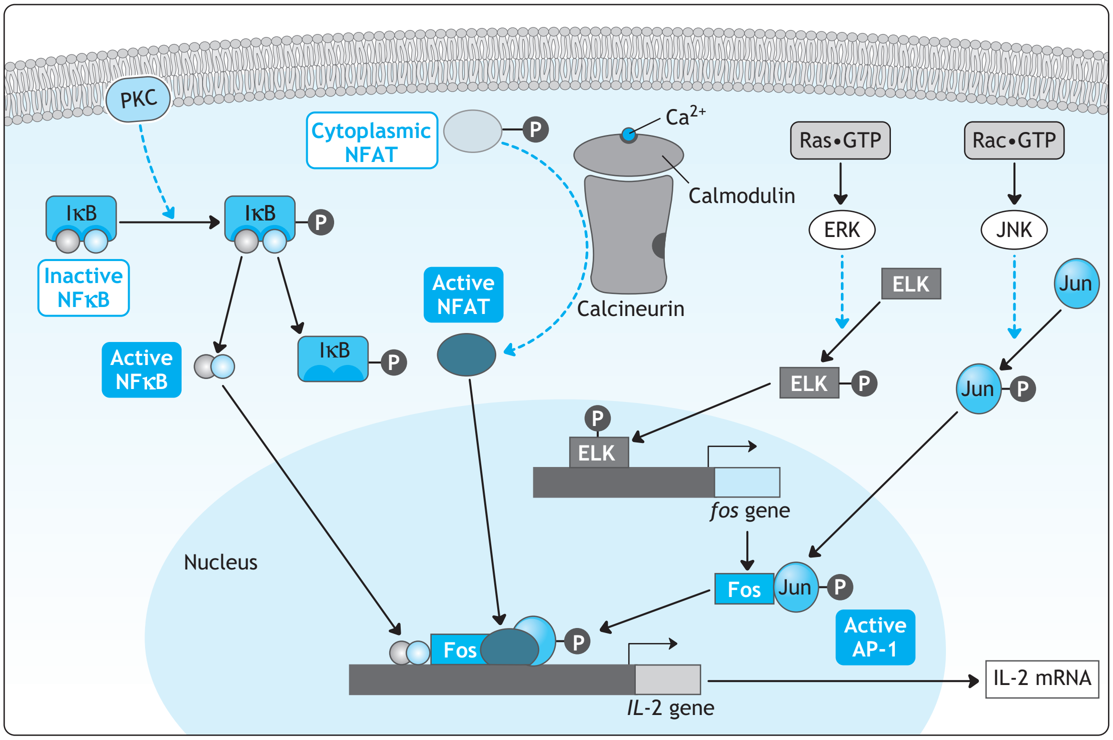
PKC
Cytoplasmic NFAT
P
Ca2+
Calmodulin
Ras•GTP
Rac•GTP
IκB
IκB
P
Inactive NFκB
ERK
JNK
Active NFAT
Calcineurin
ELK
Jun
Active NFκB
IκB
P
ELK
P
Jun
P
P
ELK
fos gene
Nucleus
Fos
Jun
P
Fos
P
Active AP-1
IL-2 mRNA
IL-2 gene

Figure 6–4. Transcription factors that mediate T cell signaling. NF, nuclear factor; NFAT, nuclear factor of activated T cells; GTP, guanosine triphosphate; ERK, extracellular signal-regulated kinase; JNK, Jun N-terminal kinase; IL, interleukin.

b. The MAP kinases phosphorylate and activate the **AP-1** family of transcription factors (Figure 6–4).

C. The transcription factors NFκB, NFAT, and AP-1 act coordinately to initiate the transcription of important T cell genes.

1. NFAT regulates the transcription of the interleukin (IL)-2, IL-4, interferon (IFN)-γ, and tumor necrosis factor (TNF)-α genes.
2. NFκB transcription factors induce the expression of the IL-2, IFN-γ, and TNF-α genes.
3. AP-1 activates transcription of the IL-2 and many other cytokine genes.

CYCLOSPORINE AND TRANSPLANTATION

CLINICAL CORRELATION

- *Cyclosporine and the related drug FK-506 (Tracolimus) have proven effective in managing allograft rejection episodes that are mediated by activated T cells.*
- *Both drugs inhibit the cellular phosphatase calcineurin and thereby diminish NFAT activation and cytokine gene expression.*
- *The use of these powerful immunosuppressive drugs has allowed clinical transplantation to be performed successfully even when some genetic disparity exists between donor and host (Chapter 17).*

D. Coreceptors augment signaling initiated by the TCR (Figure 6–3).

1. CD4 and CD8 binding to MHC molecules recruits the kinase **Lck** to the TCR complex.
2. Lck then phosphorylates the CD3ζ ITAM residues, which promotes ZAP-70 recruitment to the complex.
3. CD28 coreceptor binding to its ligand, B7 on dendritic cells, recruits PLCγ to the cell membrane.
4. TCR ligation in the absence of coreceptor signaling can lead to **anergy** (unresponsiveness) in T cells.
    - **a.** Clonal anergy is important in maintaining unresponsiveness to certain self antigens (Chapter 16).
    - **b.** Conversely, the loss of clonal anergy among T cells that bear receptors for autoantigens can result in autoimmunity.

IV. The functions of antigen-activated T cells emerge after several cycles of cell division and differentiation.

A. Most activated $CD4^+$ T cells express helper activity by secreting cytokines that act on other immune system cells.

1. The $CD4^+$ T helper population can be further subdivided into the Th1 and Th2 cell subsets based upon the cytokines they secrete (Table 6–4 and Figure 6–5).

Table 6–4. Th1 and Th2 $CD4^+$ T cell subsets.[a]

Cytokine	Th1 Cells	Th2 Cells	Principal Functions
IL-2	++	–	Lymphocyte proliferation
IFN-γ	++	–	Macrophage activation Inhibition of IL-4 effects
LT (TNF-β)	++	–	Cytotoxicity
GM-CSF	++	+	Monocyte and dendritic cell differentiation
IL-4	–	++	B cell activation; Ig class switching; mast cell growth; inhibition of Th1 pathway
IL-5	–	++	B cell activation; Ig class switching; eosinophil growth and differentiation
IL-10	–	++	Inhibition of macrophage and lymphocyte activation
IL-13	–	++	Induction of mucus secretion; inhibition of IFN-γ action

[a] IL, interleukin; IFN, interferon; LT, lymphotoxin; TNF, tumor necrosis factor; GM-CSF, granulocyte macrophage colony-stimulating factor; Ig, immunoglobulin.

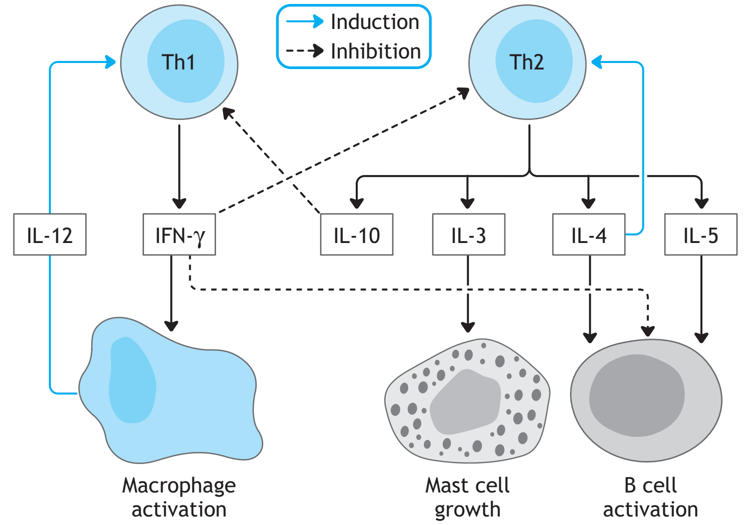

Figure 6–5. Cross-regulation of the Th1 and Th2 cell subsets. IL, interleukin.

   a. The **Th1 cell** subset produces cytokines primarily associated with **cell-mediated immunity** (ie, cytotoxic T cell growth and macrophage activation).
   b. The **Th2 cell** subset produces cytokines primarily associated with **humoral immunity** (ie, B cell activation, proliferation, differentiation, and Ig class switching).
2. Some of the cytokines produced by the two Th cell subsets promote the development of their respective subset.
   a. IL-4 is an important inducer of the Th2 response.
   b. IFN-γ favors the development of a Th1 response.
3. Each subset produces at least one cytokine that inhibits the functions of the other subset pathway (Figure 6–5).
   a. IL-10 and IL-4 produced by Th2 cells inhibit IL-12 production by APCs.
   b. IL-4 antagonizes the macrophage-activating effects of IFN-γ.
   c. IFN-γ produced by Th1 cells inhibits the proliferation of the Th2 cell subset.
   d. IFN-γ antagonizes the B cell-activating effects of IL-4.
4. Dedicated transcription factors mediate the genetic control of Th1/Th2 development.
   a. **T-bet** is activated by IFN-γ and promotes Th1 cell development.
   b. **GATA-3** favors the development of a Th2 cell response.

B. Several populations of T cells express cytotoxic activity against foreign target cells (Table 6–5).
   1. Conventional CD8$^+$ **cytotoxic T lymphocytes (CTL)** kill targets expressing foreign peptides and MHC class I molecules.
   2. Some CD4$^+$ T cells are cytotoxic for targets expressing foreign peptides and MHC class II molecules.

Table 6–5. Cytotoxic lymphocytes.[a]

Property	CTL	NKT Cell	NK Cell
Receptors	TCRαβ; TCRγδ	Invariant TCRαβ	a
Receptor ligands	Peptides + MHC class I	Lipids and glycolipids + CD1	a
Surface markers	CD8 > CD4	CD8, CD4, and NK cell markers	NK cell markers (eg, CD161)
Functions	Cytotoxicity and cytokine production	Cytotoxicity and cytokine production	Cytotoxicity and cytokine production
Pathogens recognized	Viruses and intracellular bacteria	Unknown	Viruses

[a] For a description of the diverse NK cell receptors and their ligands, see Chapter 10. CTL, cytotoxic T cells; NKT, natural killer T; NK, natural killer; TCR, T cell receptor.

3. iNKT cells are cytotoxic against cells expressing glycolipid antigens presented by CD1 molecules.
4. All of these cytotoxic T cells require direct contact with their target cells and induce death by apoptosis (Figure 6–6).
    - a. CTL reorient their killing components toward the cellular pole in contact with their target cell.
    - b. Membrane-bound cytosolic **granules** of the cytotoxic cell then fuse with the cell membrane.
        - *(1)* **Perforin** is a granule protein that polymerizes in the target cell membrane and facilitates the delivery of granzymes.
        - *(2)* **Granzymes** are serine proteases that cleave caspases and initiate apoptosis in the target cell.
    - c. **Fas ligand**, which is expressed on activated T cells and NKT cells, can induce apoptosis.
        - *(1)* Fas ligand binds to a widely distributed receptor called **Fas**.
        - *(2)* Fas contains death domains that recruit and activate the cytosolic caspases within the target cell.
    - d. Cytotoxic cells secrete TNF-α and **lymphotoxin (LT)**, which bind to **TNF receptors**.
        - *(1)* TNF receptors contain death domains that recruit adaptor proteins and activate caspases.
        - *(2)* TNF receptors can also activate transcription factors that regulate gene expression (Chapter 12).
5. After delivering a lethal hit, the cytotoxic cell dissociates from its target cell and repeats the process with another target cell.

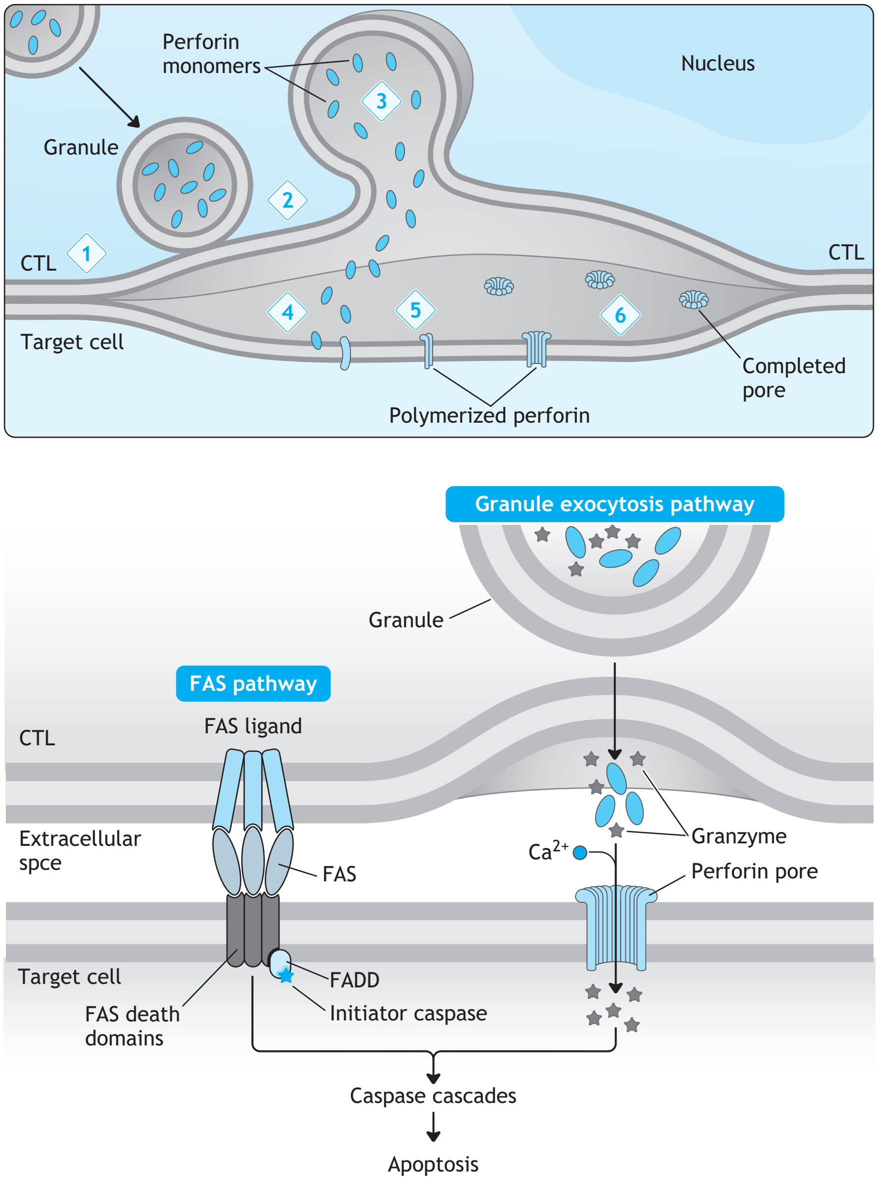
Perforin monomers
Nucleus
Granule
3
2
CTL
1
CTL
Target cell
4
5
6
Completed pore
Polymerized perforin
Granule exocytosis pathway
Granule
FAS pathway
FAS ligand
CTL
Extracellular spce
FAS
Ca^{2+}
Granzyme
Perforin pore
Target cell
FADD
Initiator caspase
FAS death domains
Caspase cascades
Apoptosis

Figure 6–6. Conjugate formation and cell-mediated lysis by a cytotoxic T cell (CTL). (1) CTL-target cell contact, (2) granule fusion, (3) degranulation, (4) perforin insertion into target cell membrane, (5) polymerization of perforin, (6) pore formation

CLINICAL PROBLEMS

Infection of cells with certain herpes viruses leads to a decrease in the expression of class I MHC molecules by the infected cells.

1. For which of the following cytotoxic cells would this *decrease* its ability to recognize the virus-infected cell?

A. $CD4^+$ T cell

B. $CD8^+$ T cell

C. NK cell

D. NKT cell

E. NK and NKT cell

A patient in renal failure received a kidney transplant from an unrelated donor. The patient was treated immediately with cyclosporine. Three weeks posttransplant, he developed an infection with cytomegalovirus that was attributed to his immunosuppressive therapy.

2. A decrease in which of the following responses would best explain this type of opportunistic infection?

A. Transcriptional activation in B cells

B. Ca^{2+} mobilization in macrophages

C. Cytokine gene expression in T cells

D. Neutrophil chemotaxis

E. Signaling through the BCR

Ronald is a 2-year-old boy who presents with a 1-year history of recurrent sinus infections caused by gram-positive bacteria, including Group A streptococci. He has not had infections with viruses or fungi. He was successfully treated with antibiotics in the past, but his infections often returned after treatment ended. Laboratory tests indicate an elevated titer of anti-A and anti-B isohemagglutinins, but an IgG level that is less than 5% of the normal level for his age.

3. Which of the following mechanisms or diseases would be most consistent with this clinical picture?

A. Bruton's X-linked agammaglobulinemia

B. A deficiency of ZAP-70

C. DiGeorge syndrome

D. A lack of CD154 expression

E. HIV-1 infection

Mary is an 11-year-old child with asthma. Allergy skin testing has revealed allergic reactions to 10 different household and outdoor allergens.

4. If you were to design an experimental therapy directed at reducing cytokine production in this disease, which of the following cytokines would be the best target of this therapy?
   A. IL-1
   B. IL-2
   C. IL-4
   D. IL-8
   E. IL-12

A patient with recurrent viral infections has normal numbers of blood lymphocytes, but the cells do not express granzymes when activated in vitro.

5. A mutation in the gene encoding which of the following signaling intermediates would produce this phenotype?
   A. Lck
   B. Fas
   C. MyD88
   D. ZAP-70
   E. Btk

ANSWERS

1. The correct answer is B. Among the choices, only CD8$^+$ T cells require MHC class I molecules for their recognition of viral peptide antigens. The expression of this protein can be inhibited by viruses that interfere with the synthesis of β_2-microglobulin, which is shared by MHC class I and CD1 molecules (Chapter 7).

2. The correct answer is C. Cyclosporine inhibits the phosphatase calcineurin, which is important in activating the transcription factor NFAT. NFAT regulates a number of cytokine genes in T cells, including those that coinduce virus-specific CD8$^+$ cytotoxic T cells to differentiate (eg, IL-2).

3. The correct answer is D. Elevated IgM antibody together with very low or absent IgG is an indication that the patient's B cells do not switch Ig class efficiently. This is typical of hyper IgM syndrome type 1, the congenital absence of CD154. The patient does not lack B cells, but fails to receive the proper signal from Th cells.

4. The correct answer is C. IL-4 plays a central role in the development of allergy by promoting Ig class switching to IgE and inducing more Th2 helper cells that produce IL-4. Mast cells are the source of inflammatory mediators in asthma, and their growth is also promoted by IL-4 (Chapter 13).

5. The correct answer is D. ZAP-70 deficiency causes a decreased signaling in both CD4$^+$ and CD8$^+$ T cells. Granzyme expression is a marker for activated cytotoxic lymphocytes.

CHAPTER 7

MAJOR HISTOCOMPATIBILITY COMPLEX

I. The principal function of the major histocompatibility complex (MHC) of genes is to regulate antigen recognition by T cells.

- **A.** In the thymus, MHC molecules determine the content of the T cell receptor (TCR) repertoire.
  - **1.** Clones of thymocytes expressing TCRs that recognize foreign peptide antigens presented by the host's own MHC molecules are retained (Chapter 10).
  - **2.** Clones of thymocytes with TCRs that recognize peptide antigens presented by nonhost MHC molecules are deleted.
  - **3.** Mature thymocytes leave the thymus bearing TCRs capable of recognizing foreign peptides + self MHC in the periphery.
- **B.** In the postthymic environment, the MHC controls antigen recognition by mature T cells.
  - **1.** The TCR recognizes antigenic peptides in the periphery only when they are complexed with a self MHC molecule.
    - **a.** Antigen recognition by T cells is "MHC restricted."
    - **b.** The TCR has a dual specificity for "altered self" (ie, self MHC + a foreign peptide).
  - **2.** MHC genes are **immune response genes** that determine the magnitude of T cell activation by foreign protein antigens.
- **C.** MHC genes control the activation of natural killer (NK) cells by two mechanisms (Chapter 10).
  - **1.** MHC gene products present peptides to NK cell **activating receptors**.
  - **2.** MHC glycoproteins are also recognized by **inhibitory receptors** that suppress NK cell activation.
- **D.** The MHC serves as the most important genetic barrier to organ transplantation between two unrelated individuals of a species.
  - **1.** In organ transplantation, the MHC molecules on a donor's tissues can be viewed by the recipient's T cells as foreign.
    - **a.** In this context MHC molecules are **alloantigens**, antigens expressed by some, but not all, members of a species.
    - **b.** MHC alloantigens stimulate vigorous T cell responses leading to graft rejection (Chapter 17).

6193ch07.qxd_mg 2/6/06 12:55 PM Page 78

II. MHC genes encode two classes of polypeptides that control antigen recognition by T cells.

A. **MHC class I** gene products present foreign peptides to $CD8^+$ T cells.
  1. Antigen recognition by $CD8^+$ T cells is restricted by MHC class I molecules.
  2. Only peptides bound to MHC class I molecules activate $CD8^+$ T cells.
  3. The three MHC class I genetic loci in human beings are called *human leukocyte antigen-A (HLA-A), HLA-B,* and *HLA-C.*
  4. Each HLA class I gene encodes an ~45-kDa α polypeptide chain.
    a. Class I α chains are always expressed on the cell surface noncovalently associated with a 12-kDa invariant chain called **β_2**-microglobulin (Figure 7–1).
    b. Class I α chains are integral membrane proteins with three external domains.
      (1) Amino acids within the membrane-distal α1 and α2 domains are highly variable.

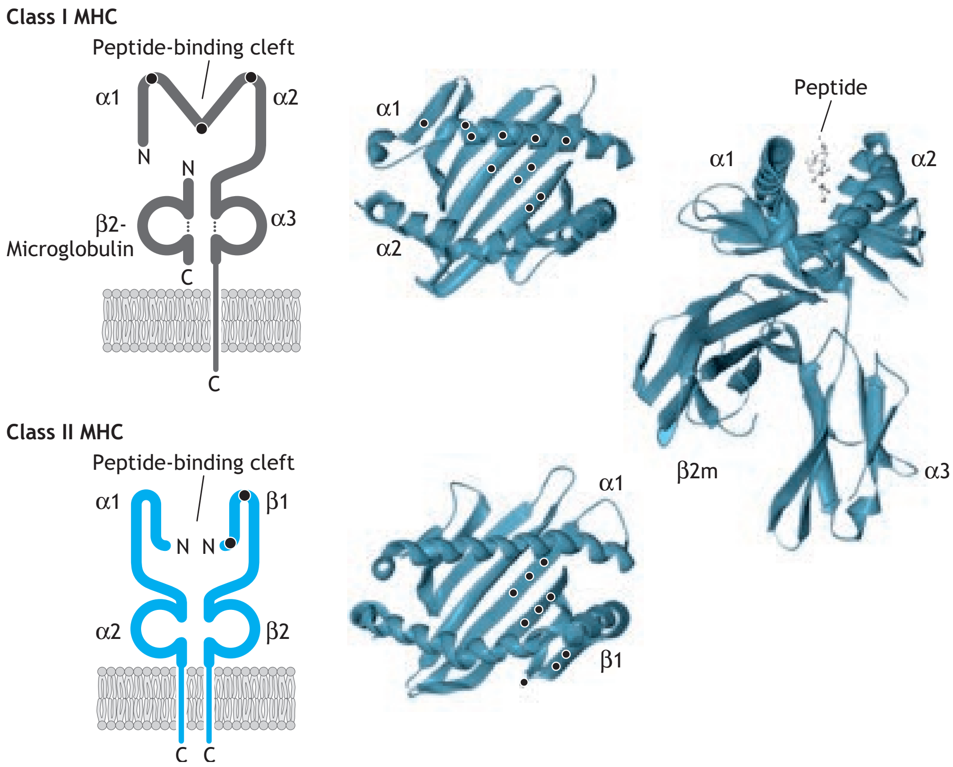

Class I MHC

Class II MHC

Figure 7–1. Structure of major histocompatibility complex (MHC) molecules. β2m, β_2-microglobulin. (Top: Adapted, with permission, from Bjorkman PJ: Structure of the human class I histocompatibility antigen, HLA-A2. Nature 1987;329:506.) (Bottom: Adapted, with permission, from Brown JH: Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 1993;364:33.)

    - *(2)* The α1 and α2 domains determine the peptide-binding specificity and alloantigenicity of class I.
    - *(3)* The membrane proximal α3 domain and β_2-microglobulin stabilize the overall structure.
  - **5.** MHC class I molecules possess a single **peptide-binding groove** within the α1 and α2 domains (Figure 7–1).
    - **a.** Contact residues within the groove determine specificity.
    - **b.** The typical peptide that binds to a class I molecule is 9–11 amino acids in length.
    - **c.** MHC class I-binding peptides contain terminal hydrophobic residues that anchor them into the groove.
    - **d.** Each MHC class I molecule shows **"promiscuous binding"** in that it can bind several different peptides.
- **B.** The principal function of **MHC class II molecules** in the periphery is to present peptide antigens to CD4$^+$ T cells.
  - **1.** Antigen recognition by CD4$^+$ T cells is MHC class II restricted.
  - **2.** Only foreign peptides bound to MHC class II molecules can activate CD4$^+$ T cells.
  - **3.** The MHC class II genes in human beings are found in three genetic regions called ***HLA-DP, HLA-DQ,*** and ***HLA-DR***.
  - **4.** Each class II genetic region contains one or more α and β genes.
    - **a.** The class II α and β loci encode for 33- and 28-kDa polypeptide chains, respectively.
    - **b.** The class II molecules form αβ heterodimeric cell surface glycoproteins (Figure 7–1).
    - **c.** Each chain has a membrane-distal domain (ie, α1 or β1) that is highly variable.
    - **d.** The α1 and β1 domains of class II molecules contain the peptide-binding groove.
    - **e.** These regions of sequence variability also prescribe the alloantigenic nature of class II molecules.
  - **5.** The peptide-binding groove of a typical class II molecule is similar in structure to that of a class I molecule.
    - **a.** Contact residues in the binding groove determine binding specificity.
    - **b.** MHC class II molecules typically bind 12–15 amino acid peptides that can extend beyond the limits of the groove.
    - **c.** Like class I, class II molecules show binding promiscuity.
- **C.** Class I and class II MHC genes belong to the Ig supergene family.
- **D.** MHC class III genes do not control T cell activation directly.
  - **1.** Class III genes include those for the complement proteins C2, C4, and Factor B (Chapter 8).
  - **2.** The gene for tumor necrosis factor (TNF)-α is also found in the class III MHC family.

ALTERED MHC-BINDING PEPTIDE LIGANDS

CLINICAL CORRELATION

- *The nature of a T cell response to a peptide antigen can be altered by changing only a few of the contact residues of the peptide that interact with the TCR.*

- *These **altered peptide ligands** include "superagonistic" peptides that can enhance T cell activation, which would be useful in designing antitumor immunotherapy.*
- *Similarly, altered peptide ligands can theoretically be designed that would inhibit unwanted T cell activation, such as that seen in autoimmune diseases.*

III. MHC molecules show extensive functional diversity due to polygeny and genetic polymorphism.

A. **Polygeny** refers to the existence of multiple genetic loci that have similar functions.
  1. Three MHC class I gene loci, *HLA-A, HLA-B,* and *HLA-C,* control antigen presentation to $CD8^+$ T cells.
  2. Many MHC class II genetic loci control antigen presentation to $CD4^+$ T cells (eg, $DP\alpha1$, $DP\alpha2$, $DP\beta1$, and $DP\beta2$).
  3. The existence of multiple MHC gene products increases the number of foreign peptides that can be presented to T cells.

B. **Polymorphism** within the MHC refers to the existence of multiple **alleles** of a gene within a population (Table 7–1).
  1. Each class I allele encodes an α chain.
  2. Each class II allele encodes either an α or a β chain.
  3. A diploid individual can express at most two allelic products at each MHC locus (ie, in the heterozygote).
  4. The β_2-microglobulin gene is *not* polymorphic.
  5. Polymorphism has a greater effect on HLA diversity within the species than within an individual (Table 7–2).
    a. Thus, in the human population there is the capacity to present an enormous number of antigenic peptides.
    b. Each individual of that population has a much narrower potential repertoire of MHC-encoded molecules.

C. HLA diversity is sufficient for recognizing most protein antigens.
  1. Each allelic product (eg, HLA-A1) can bind multiple peptides.
  2. Each cell expresses multiple allelic forms (Figure 7–2).

Table 7–1. Examples of polymorphism at the HLA loci.

MHC Class	HLA Locus	Number of Alleles
I	*HLA-A*	250
I	*HLA-B*	490
I	*HLA-C*	119
II	*HLA-DPA1*	20
II	*HLA-DPB1*	99
II	*HLA-DQA1*	22
II	*HLA-DQB1*	53
II	*HLA-DRB1*	315
II	*HLA-DRB3*	38

Table 7–2. The basis for diversity among immunoglobulins and MHC molecules.

Feature	Immunoglobulin Genes	MHC Genes
Repertoire of an individual	$\sim 10^4$ genes; $\sim 10^9$ Igs	6 class I alleles 6 class II alleles
Repertoire of the species	$\sim 10^4$ genes; $\sim 10^9$ Igs	> 400 class I alleles > 500 class II alleles
Diversity mechanisms	Large pools of germ-line V, D, J segments Combinatorial joining of V, D, J segments Imprecision in joining sites Nucleotide additions at joining sites Somatic hypermutation Combinatorial association of H and L polypeptides	Heterozygosity Codominant expression Polygeny Polymorphism

3. Mounting a successful response to a limited number of antigenic epitope–MHC complexes is apparently adequate to protect the host.

IV. MHC genes are inherited as a closely linked group of loci and are codominantly expressed (Figure 7–3).

A. HLA class I loci are clustered at the 3′ end of the complex on chromosome 6.

1. Each class I locus encodes an α chain, which is paired with β_2-microglobulin prior to surface expression (Figure 7–4).

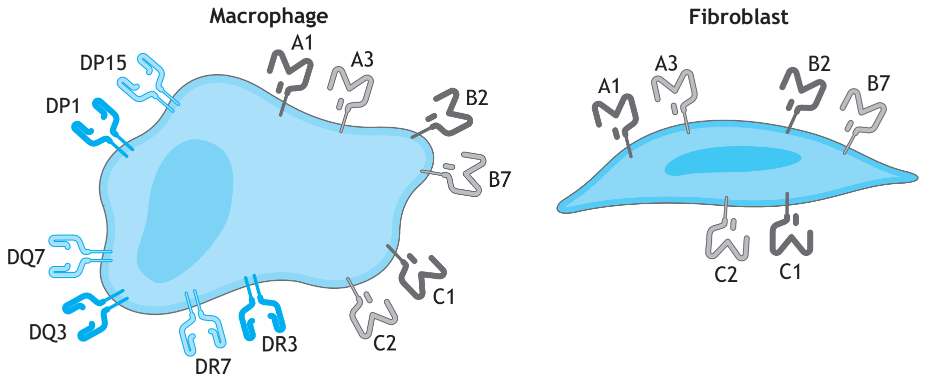

Figure 7–2. Patterns of expression of *HLA* class I and class II genes.

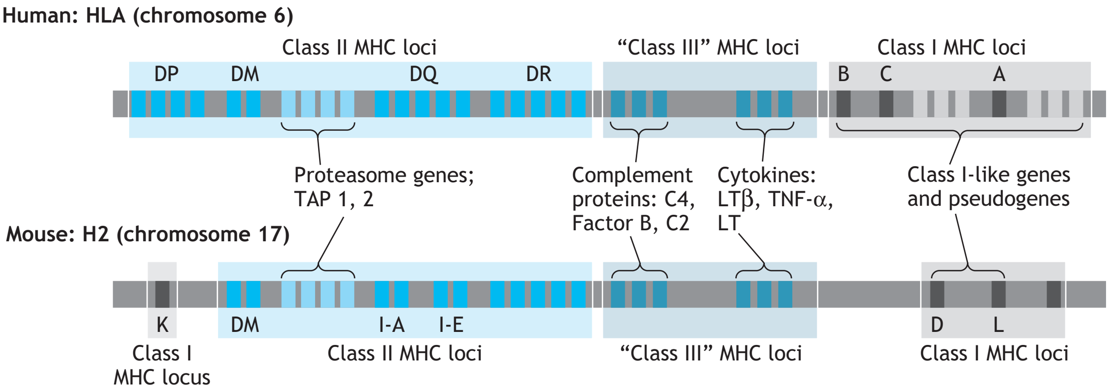
Human: HLA (chromosome 6)
Class II MHC loci
DP
DM
DQ
DR
"Class III" MHC loci
Class I MHC loci
B
C
A
Proteasome genes; TAP 1, 2
Complement proteins: C4, Factor B, C2
Cytokines: LTβ, TNF-α, LT
Class I-like genes and pseudogenes
Mouse: H2 (chromosome 17)
K
Class I MHC locus
DM
I-A
I-E
Class II MHC loci
"Class III" MHC loci
D
L
Class I MHC loci

Figure 7–3. Genetic map of the *HLA* complex. MHC, major histocompatibility complex; TAP, transporter in antigen processing; LT, lymphotoxin; TNF, tumor necrosis factor.

B. The class II loci are clustered at the 5′ end of the complex and code for either an α or a β chain.
   1. Different α and β gene products from a given region (eg, DP) can pair with one another to form heterodimeric products.
   2. Additional loci between the DP and DQ regions encode proteins that function in antigen presentation (see below).

C. The closely linked genes of the MHC are inherited in a Mendelian fashion.
   1. The constellation of linked MHC genes on a given chromosome is called a **haplotype** (Table 7–3).
   2. Haplotype inheritance ensures that certain predictions can be made about HLA relationships within families (Figure 7–5).
      a. Assuming no genetic recombination has occurred, parents are always **haploidentical** with their children.
      b. Each child has a 75% chance of haploidentity with a sibling.
      c. Each child has a 25% chance of complete **HLA identity** with a sibling.

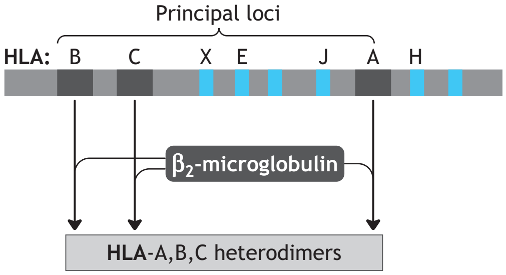
Principal loci
HLA:
B
C
X
E
J
A
H
β_2-microglobulin
HLA-A,B,C heterodimers

Figure 7–4. HLA class I heterodimers.

Table 7–3. Nomenclature of the HLA genetic system.

Term	Definition	Example
Region	A region within the HLA complex containing related genes	HLA-DR
Locus	An open reading frame that encodes for an HLA peptide	HLA-DRB1 (encodes a β chain of HLA-DR)
Allele	One of the alternative sequences at a locus	HLA-DRB1*0101
Antigen	A cell surface heterodimeric protein	HLA-DR1
Haplotype	The constellation of linked genes across the entire HLA complex on one chromosome	HLA-A1, B2, C5, DP2, DQ1, DR4

D. Individuals inherit parental MHC genes in their germ line DNA.
  1. The specificity of an individual's T cells is an inherited trait.
  2. The MHC loci do *not* rearrange prior to expression (Table 7–2).

E. MHC loci are expressed in a **codominant** fashion (Figure 7–2).
  1. All six allelic products of the HLA class I loci are expressed in the heterozygote.
    a. Each MHC class I heterodimer is expressed on all nucleated cells, except cells of the trophoblast.
    b. Any nucleated cell that expresses MHC class I molecules can present peptide antigens to CD8$^+$ T cells.
    c. Each MHC class II heterodimer is expressed on all dendritic cells, B cells, and macrophages.
    d. Any such cell can present peptide antigens to CD4$^+$ T cells.

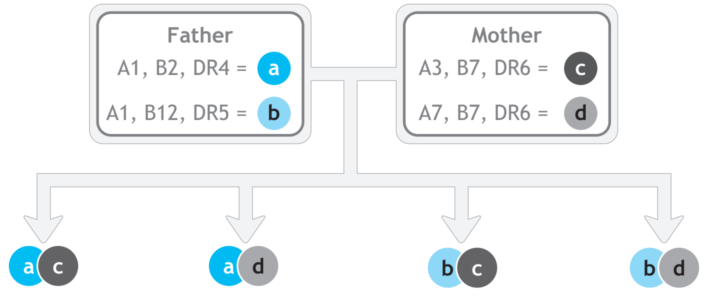

Figure 7–5. Inheritance of *HLA* genes within a family.

F. The expression of MHC class II genes is coordinately controlled by tissue-specific transcription factors, including the **class II transactivator (CIITA)**.

1. Constitutive expression of MHC class II is seen in B cells, monocytes, macrophages, and dendritic cells (“professional antigen-presenting cells”).
2. Several other cell types (eg, epithelial cells) can be induced to express class II molecules by inflammatory signals.

COORDINATED EXPRESSION OF MHC CLASS II

CLINICAL CORRELATION

- *An autosomal recessive condition expressed in infants called* ***bare lymphocyte syndrome (BLS) type 2*** *is characterized by lymphopenia, diarrhea, and a variety of opportunistic infections.*
- *These patients have significantly decreased numbers of blood* $CD4^+$ *T cells, and show decreased cellularity in their peripheral lymphoid tissues.*
- *HLA-DP, DQ, and DR molecules are essentially absent from their tissues.*
- *In many BLS type II patients, the expression of CIITA is defective.*

V. Two distinct pathways of antigen presentation ensure that foreign peptide antigens are presented by the proper MHC molecule to the appropriate T cell subset (Table 7–4 and Figure 7–6).

A. Intracellular microbes (eg, viruses and intracellular bacteria) express protein antigens that are presented on class I molecules to $CD8^+$ T cells.

1. Antigen presentation involves the degradation of these foreign proteins following their synthesis within the **cytosol**.
2. The resulting peptides are directed to newly synthesized class I molecules in the endoplasmic reticulum.

Table 7–4. Pathways of antigen processing and presentation.[a]

Pathway	Source of Antigen	Essential Cellular Components	Typical Antigens Processed
Class I	Endogenous	MHC class I α chains β_2-Microglobulin Ubiquitin Proteasome TAP-1,2 Chaperones (calnexin, tapasin, calreticulin)	Viral capsule proteins and nucleoproteins Proteins derived from intracellular bacteria
Class II	Exogenous	MHC class II α and β chains Ii chain CLIP HLA-DM	Proteins derived from extracellular pathogens Traditional protein vaccines

[a] TAP, transporter associated with antigen processing; CLIP, class II-associated invariant chain peptide.

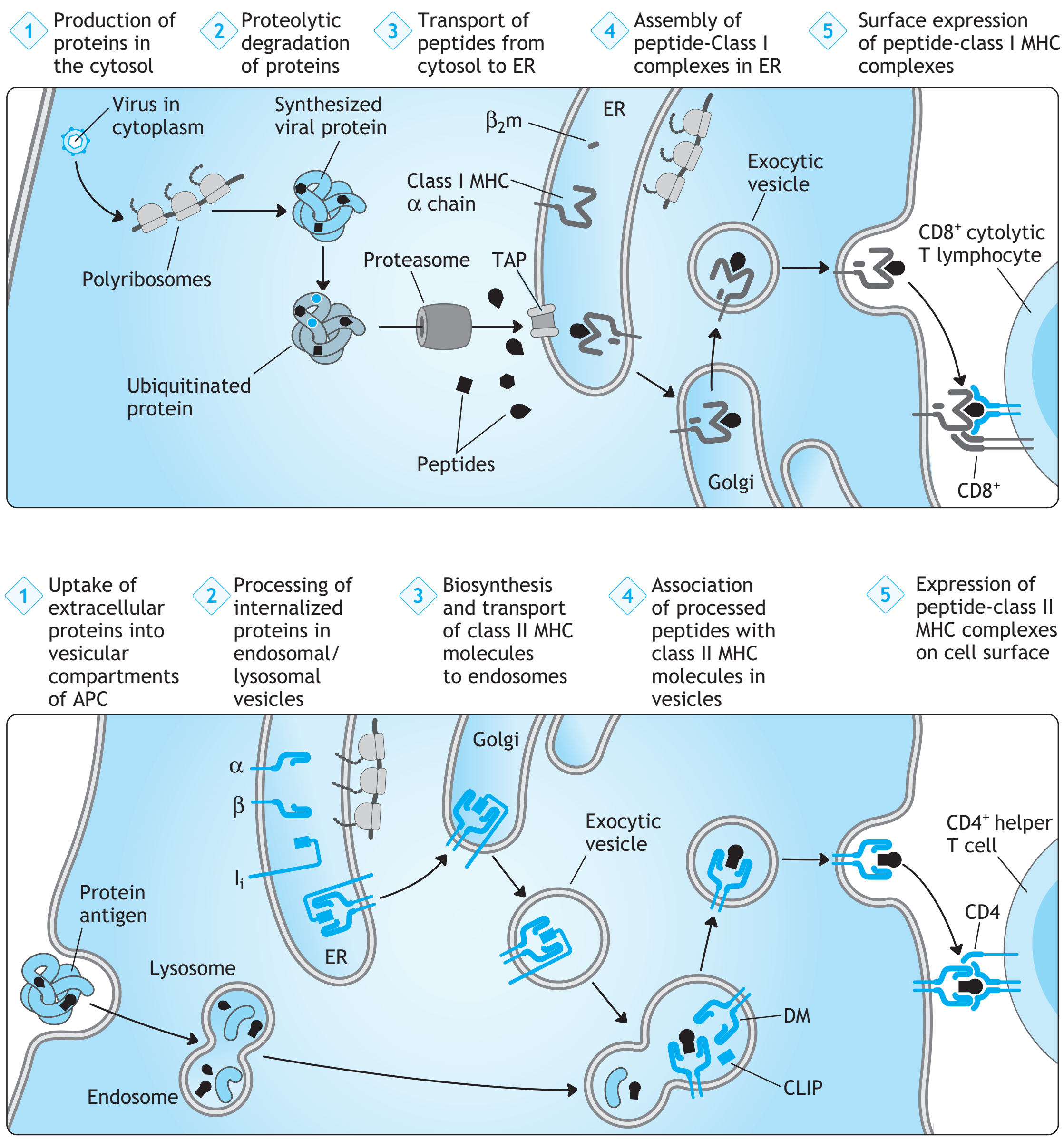
1 Production of proteins in the cytosol
2 Proteolytic degradation of proteins
3 Transport of peptides from cytosol to ER
4 Assembly of peptide-Class I complexes in ER
5 Surface expression of peptide-class I MHC complexes
Virus in cytoplasm
Synthesized viral protein
β_2m
ER
Class I MHC α chain
Exocytic vesicle
CD8$^+$ cytolytic T lymphocyte
Polyribosomes
Proteasome
TAP
Ubiquitinated protein
Peptides
Golgi
CD8$^+$
1 Uptake of extracellular proteins into vesicular compartments of APC
2 Processing of internalized proteins in endosomal/ lysosomal vesicles
3 Biosynthesis and transport of class II MHC molecules to endosomes
4 Association of processed peptides with class II MHC molecules in vesicles
5 Expression of peptide-class II MHC complexes on cell surface
α
β
Golgi
I_i
Exocytic vesicle
CD4$^+$ helper T cell
Protein antigen
ER
CD4
Lysosome
DM
Endosome
CLIP

Figure 7–6. Antigen processing by the class I and class II pathways. ER, endoplasmic reticulum; TAP, transporter associated with antigen processing; MHC, major histocompatibility complex; CLIP, class II-associated invariant chain peptide.

   3. These **endogenous protein antigens** are expressed on MHC class I and recognized by CD8$^+$ T cells.

B. **Extracellular microbes** activate CD4$^+$ T cells by expressing antigenic peptides that are presented on MHC class II molecules.
   1. The uptake of these exogenous antigens can be accomplished by **phagocytosis, endocytosis,** or **receptor-mediated endocytosis**.
   2. Antigen processing and presentation require an intracellular pathway that degrades proteins within an **endosome**.
   3. The resulting peptides are directed to an **exocytic vesicle** containing newly synthesized class II molecules.
   4. These exogenous protein epitopes are then presented by the **class II molecules** to CD4$^+$ T cells.

C. Components of the **class I (endogenous or cytosolic) antigen presentation pathway** are dedicated to proteolysis, peptide sizing, peptide transport, and protein targeting.
   1. The foreign protein antigen is first covalently attached to the small peptide **ubiquitin** via the ε-amino group of lysine residues of the antigen molecule (Figure 7–6).
   2. Ubiquitination targets the protein for degradation by the multicatalytic **proteasome**.
      a. The proteasome subunits LMP-2 and LMP-7, in part, determine substrate cleavage specificity.
      b. LMP-2 and LMP-7 are encoded by genes in the MHC (Figure 7–3).
      c. Peptides that leave the cylindrical proteasome vary in size and charge.
   3. Peptides that will fit the binding groove of MHC class I molecules are transported from the cytosol to the endoplasmic reticulum.
      a. The **transporter associated with antigen processing (TAP)** complex mediates peptide transport across the ER membrane.
      b. TAP selects peptides of the appropriate length (9–11 amino acids) with hydrophobic ends.
      c. The genes encoding the TAP complex, ***TAP-1*** and ***TAP-2***, reside in the MHC (Figure 7–3).
   4. Newly synthesized MHC class I α chains are present in the ER membrane associated with β_2-microglobulin and the **chaperone proteins calnexin, tapasin,** and **calreticulum**.
      a. Chaperones direct the trafficking of MHC class I molecules to the TAP complex.
      b. Dissociation of the chaperones enables the loading of peptides into the MHC class I molecules.
   5. The MHC class I–peptide complex is then directed to the membrane for expression.
      a. The absence of β_2-microglobulin precludes surface expression of MHC class I molecules.
      b. The absence of a peptide in the binding groove of MHC class I prevents its surface expression.
      c. Pathogenic intracellular microbes often block the class I pathway by inhibiting β_2-microglobulin synthesis or TAP function.

6. In the absence of *foreign* antigenic peptides, MHC class I molecules are always loaded with *self* peptides.

BARE LYMPHOCYTE SYNDROME TYPE 1

CLINICAL CORRELATION

- *Mutations in either the TAP1 or TAP2 gene prevent the transport of cytosolic peptides into the ER and result in incomplete peptide loading.*
- *Patients with these mutations do not express HLA-A, B, and C molecules on their nucleated cells, because class I molecules cannot be expressed unless they are loaded with peptide.*
- *These patients also fail to develop CD8$^+$ T cells due to a defect in positive selection in the thymus (Chapter 10).*
- *The resulting immunodeficiency is not as severe as severe combined immunodeficiency (SCID), but can lead to opportunistic viral infections.*
- *Patients with BLS type 1 generally show normal numbers of circulating lymphocytes due to the normal development of the CD4$^+$ T cell subset.*

D. Peptides presented by MHC class II molecules are generally derived from extracellular proteins that are taken up by antigen-presenting cells.
   1. Following uptake into an endocytic or phagocytic vesicle, the organelle becomes increasingly acidic with time.
   2. Hydrolytic enzymes of the endolysosomes and phagolysosomes mediate proteolysis of the antigen.
   3. These vesicles then fuse with exocytic vesicles containing newly synthesized MHC class II molecules.
      a. Newly synthesized class II molecules are associated with a chaperone membrane protein called the **invariant chain (Ii)** .
      b. Ii ensures proper folding and protects the binding groove of the class II heterodimer.
   4. When the vesicles fuse, the Ii chain is degraded into a small peptide, called a **class II-associated invariant chain peptide (CLIP)**.
   5. **HLA-DM** catalyzes the removal of CLIP and the binding of appropriately sized antigenic peptides.
   6. The loaded MHC class II molecule is then directed to the cell surface.

VI. There are notable associations between specific HLA phenotypes and the susceptibility for certain diseases (Table 7–5).

A. The relative risk of **hereditary hemochromatosis** is higher in individuals with the HLA-A3 + B14 haplotype than in the general population.
   1. Persons with this haplotype have an elevated incidence of mutations in a nonclassical MHC class I molecule called **HLA-H** (also known as HFE).
   2. HLA-H appears to regulate the function of the transferrin receptor, thereby affecting the absorption of dietary iron in the stomach and small intestine.
   3. Thus, certain mutations in HLA-H result in a chronic systemic exposure to excess iron resulting in hereditary hemochomatosis.

B. HLA associations are particularly common among autoimmune diseases.
   1. **HLA-B27** expression is associated with an increased relative risk of **spondylarthropathies**.

Table 7–5. Disease associations with HLA phenotypes.

Disease	HLA Antigen or Haplotype	Relative Risk
Hereditary hemochromatosis	HLA-A3 + HLA-B14	90
Ankylosing spondylitis	HLA-B27	90
Reiter's syndrome	HLA-B27	37
Gluten enteropathy	HLA-DR3	12
Rheumatoid arthritis	HLA-DR4	14

a. Spondylarthropathies are characterized by inflammation of the joints and characteristic extraarticular tissues.
b. Individuals expressing HLA-B27 have an increased relative risk of developing the autoimmune conditions ankylosing spondylitis and Reiter's syndrome (Table 7–5).
c. **Relative risk** is the prevalence of the disease in those carrying the HLA marker compared to the prevalence of the disease among the general population.
d. A number of microbial antigens cross-react with HLA-B27–peptide complexes, suggesting an infectious etiology for these autoimmune diseases.

CLINICAL PROBLEMS

The HLA genetic system was used in a civil case to resolve a paternity suit. The case involved a woman who sued an acquaintance for child support for her newborn twins. The court ordered tissue typing (serological identification of the HLA markers; Chapter 17) of the mother, the two children, and the defendant. Partial results of the typing are shown below.

Individual	HLA-A Markers	HLA-B Markers
Mother	1,2	8,12
Son	1,3	8,7
Daughter	2,5	12,9
Defendant	3,3	7,28

1. Assuming there has been no recombination within the HLA complex, which haplotype did the son inherit from his mother?

A. HLA-A1 + HLA-A2
B. HLA-A1 + HLA-B7
C. HLA-A1 + HLA-B8
D. HLA-A3 + HLA-B7
E. HLA-B8 + HLA-B7

2. Assuming no recombination within the HLA complex, which haplotype did the daughter inherit from her mother?
A. HLA-2 + HLA-B12
B. HLA-2 + HLA-A5
C. HLA-B12 + HLA-B9
D. HLA-A2 + HLA-B9
E. HLA-A5 + HLA-B12

3. Which of the following is a paternal haplotype that could have been inherited by one of the children from the defendant?
A. HLA-A3 + HLA-B7
B. HLA-A1 + HLA-B8
C. HLA-A2 + HLA-B9
D. HLA-A5 + HLA-B9
E. HLA-A1 + HLA-B7

4. Based on these data alone, which of the following could be true?
A. The son cannot be related to the defendant.
B. The daughter cannot be related to the defendant.
C. Neither child is related to the defendant.
D. Both children could be related to the defendant.
E. The two children have different fathers.

Below is a pedigree showing the HLA phenotypes of the Smith family. This is the second marriage for both the father and the mother.

5. Which of the five children shown here is most likely the child of a previous marriage involving the mother?

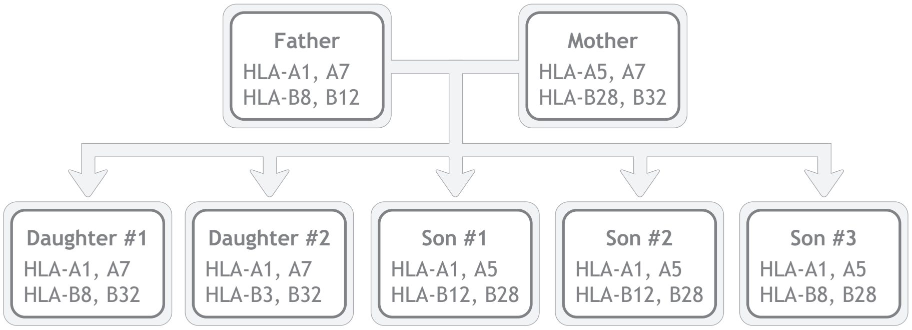
Father
HLA-A1, A7
HLA-B8, B12
Mother
HLA-A5, A7
HLA-B28, B32
Daughter #1
HLA-A1, A7
HLA-B8, B32
Daughter #2
HLA-A1, A7
HLA-B3, B32
Son #1
HLA-A1, A5
HLA-B12, B28
Son #2
HLA-A1, A5
HLA-B12, B28
Son #3
HLA-A1, A5
HLA-B8, B28

James is a 13-year-old child in need of a kidney transplant. He is the third of 14 children all with the same parents.

6. What are the chances that one of his siblings will be fully HLA compatible with James?

A. There is almost no chance this will occur.

B. 1 chance in 13

C. 1 chance in 4

D. 1 chance in 2

E. All of his siblings would be HLA compatible.

ANSWERS

1. The correct answer is C. A haplotype is a group of linked genes (eg, HLA-A1 + HLA-B8) on one chromosome. In this case, the maternal haplotype can be derived by noting that these are the only HLA antigens shared between the mother and the son.

2. The correct answer is A. As with the son, these are the only HLA antigens shared with the mother. Thus, the son and daughter define the two maternal haplotypes, HLA-A1 + B8 and HLA-A2 + B12.

3. The correct answer is A. If A1, B8, A2, and B12 are of maternal origin, then the remaining HLA antigens were derived from the father. What is surprising in this case is that only the HLA-A3 + HLA-B7 haplotype is represented in the defendant's type.

4. The correct answers are B and E. Based on HLA typing it is safe to conclude that the daughter is *not* related to the defendant. Because these children are fraternal twins, it can only be concluded that the two children had different fathers. This is a case of apparent superfecundity, which occurs when two ova are produced during one ovulation cycle and become fertilized by sperm from two different fathers.

5. The correct answer is daughter #2. Analysis of the pedigree indicates that the haplotypes of the father are A1 + B8 and A7 + B12. The mother's haplotypes are A7 + B32 and A5 + B28. Daughter #2 bears the maternal haplotype A7 + B32, but the paternal haplotype does not correspond to those of the father. Therefore, the child must be from a previous marriage.

6. The correct answer is C. Based on Medelian inheritance in a codominant system, there is a 25% chance of full haplotype identity between siblings, assuming no recombinations or mutations.

CHAPTER 8

COMPLEMENT

I. Complement is a primitive immune effector and amplification system consisting of more than two dozen serum proteins.

- A. Complement components are designated by a **"C"** followed by a number (eg, C1) or the word **"Factor"** followed by an upper case letter (eg, Factor B).
- B. The **biological functions** of complement include cell membrane lysis, opsonization, activation of inflammatory responses, clearance of immune complexes, and B cell activation.
  - **1.** Many of the biological activities are mediated by cell surface **complement receptors**.
  - **2.** The complement system can be activated during either innate or adaptive immune responses.
- C. Complement system proteins become enzymatically or biologically active when they are cleaved, form multimers, or bind to other biomolecules.

II. The complement system can be activated by three different pathways (Table 8–1).

- A. In each pathway, complement is activated by the recognition of molecular patterns, on either antibody molecules or microbial components.
- B. All three activation pathways share a set of **core elements** that includes C3, C5, and the so-called **terminal components** C6–C9 (Figure 8–1).
  - **1.** Initiation of each of the pathways requires recruitment or assembly of a serine protease.
  - **2.** The critical second step is the cleavage of C3 by **C3 convertase** to form the peptide fragments C3a + C3b.
  - **3.** Then C5 is cleaved by a **C5 convertase**.
  - **4.** **Amplification** at each step results from the cleavage of multiple substrate molecules by these two convertases.
- C. The three pathways can be distinguished by their processes of initiation, their unique components, their unique functions, and the manner by which they are regulated (Table 8–1).
- D. The **classical pathway of complement activation** is initiated by antigen–antibody complexes.
  - **1.** Only immunoglobulin (Ig) M and IgG antibodies participate in classical pathway activation.

Table 8–1. Three pathways of complement activation.[a]

Property	Classical Pathway	Lectin Pathway	Alternative Pathway
Initiating factors	IgG- or IgM-containing immune complexes	Microbial mannose residues	Highly charged surfaces
Unique components	C1, C2, C4	MBL, MASP	Factor B, Factor D, Factor P, Factor H
C3 convertase	C4b2a	C4b2a	C3bBb
C5 convertase	C4b2a3b	C4b2a3b	C3bBb3b
Unique functions	Adaptive immunity	Innate immunity	Innate immunity; positive feedback amplification; C3 tickover function
Unique regulators	C1 Inh		Factor H

[a] Ig, immunoglobulin; MBL, mannose-binding lectin; MASP, MBL-associated serine protease; C1 Inh, C1 inhibitor.

2. Activation begins with the binding of the **C1q** peptide to the C_H domains of μ and γ heavy chains (Figure 8–2).
   - **a.** C1q is a hexamer that cross-links two adjacent binding sites located on the C_H domains of μ or γ Ig heavy chains.
   - **b.** IgM antibodies are more efficient than IgG antibodies in binding C1q, because the pentameric nature of IgM ensures that two C_H domains will be in close proximity to one another.

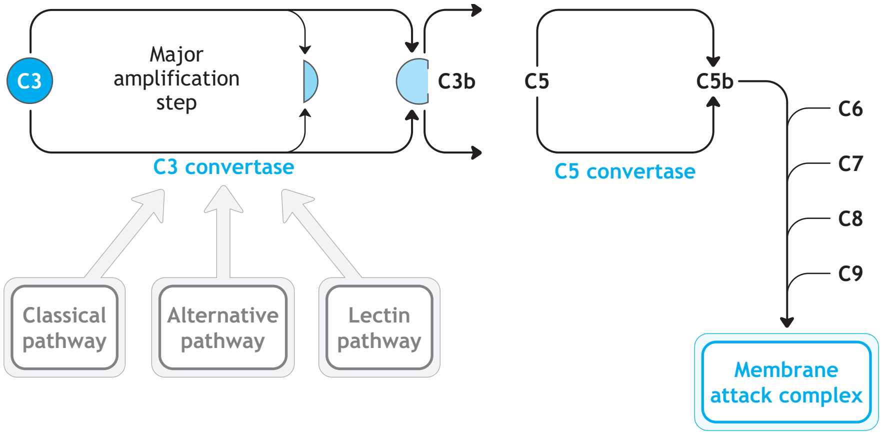
C3
Major amplification step
C3b
C3 convertase
C5
C5b
C5 convertase
C6
C7
C8
C9
Classical pathway
Alternative pathway
Lectin pathway
Membrane attack complex

Figure 8–1. The core elements shared by the three pathways of complement activation.

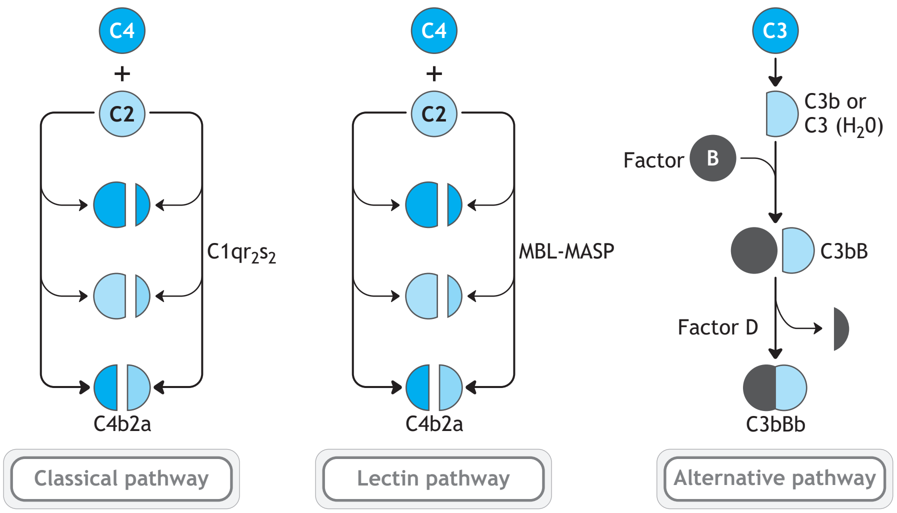

Figure 8–2. The initiation steps in the three pathways of complement activation. MBL–MASP, mannose-binding lectin–MBL-associated serine protease.

3. Once bound, C1q recruits two copies each of **C1r** and **C1s** to form an active enzyme.
4. The **C1qr$_2$s$_2$** complex can cleave C4 to yield peptides C4a and C4b.[1]
5. C1qr$_2$s$_2$ then cleaves C2 to yield the C2a + C2b peptides.
   a. C2a binds to C4b to produce C4b2a, the **classical C3 convertase**.
   b. This enzyme complex can cleave C3 to yield C3a + C3b.
6. Multiple copies of peptides are produced at each cleavage step.

HEREDITARY ANGIOEDEMA

CLINICAL CORRELATION

- ***Hereditary angioedema*** *is a autosomal dominant trait characterized by acute, nonpainful, nonpruritic, and nonerythematous swelling of the skin and mucous membranes.*
- *Edema of the bowel wall can be painful, and laryngeal swelling is potentially life-threatening due to asphyxiation.*
- *The genetic defect is a decrease or absence of the complement regulatory protein,* ***C1 Inhibitor (C1 Inh)****, that regulates C1qr$_2$s$_2$, Hageman factor, kallikrein, and plasmin.*
- *Unabated complement activation through the classical pathway appears to contribute to the acute angioedema, although the precise biochemical mediator has not been defined.*

E. The **lectin pathway of complement activation** shares properties and components with the classical pathway (Figure 8–2).

---

[1] A common convention in the complement field is the use of an overbar to designate enzymatically active complement components or complexes (eg, C1$\overline{qr_2s_2}$ or C$\overline{4b2a}$). For simplicity, this convention has been omitted here.

1. Activation is initiated by the binding of **mannose-binding lectin (MBL)** to microbial carbohydrate residues.
   a. MBL is another name for **mannose-binding protein** (Chapter 1).
   b. MBL recruits and binds **MBL-associated serine protease (MASP)**, an enzyme with substrate specificity similar to that of $C1qr_2s_2$.
2. MASP cleaves C4 and C2 to form the **lectin pathway C3 convertase**, C4b2a.

F. The **alternative pathway of complement activation** is initiated by at least three different mechanisms (Figure 8–2).
1. Highly charged microbial surface patterns, including polysaccharides, can bind C3.
2. In fluid phase, C3 can spontaneously hydrolyze with water to form an intermediate designated $C3(H_2O)$.
3. The C3b peptide, derived from the classical or lectin pathways, can activate the alternative pathway directly.
4. Bound C3, $C3(H_2O)$, or C3b then binds **Factor B**.
5. Bound B is then cleaved by the constitutively active enzyme **Factor D** to produce Ba + Bb.
6. C3Bb, $C3(H_2O)Bb$, and C3bBb are the **alternative pathway C3 convertases**.
   a. Whereas $C3(H_2O)Bb$ is unstable, the other two complexes are relatively stable.
   b. C3bBb can be further stabilized by binding **Factor P**.
7. The ability of the alternative pathway C3 convertases to generate more C3b constitutes a **feedback amplification loop** (Figure 8–3).

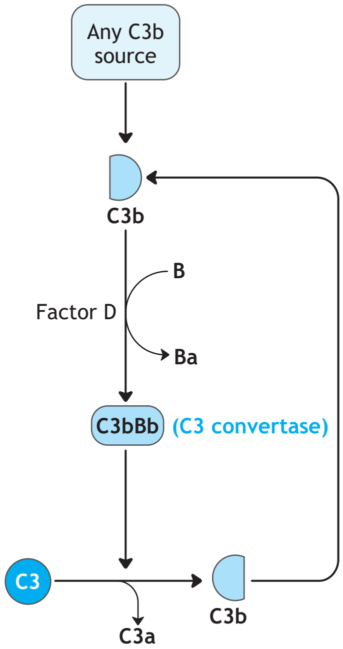

Figure 8–3. The positive feedback amplification loop of the alternative pathway.

8. Another unique feature of the alternative pathway is called **C3 tickover**.
  a. C3 constantly interacts with H_2O in fluid phase to form C3(H_2O), an initiator of alternative pathway activation.
  b. However, the C3(H_2O) intermediate is very unstable.

G. The peptides C3b and C4b can establish stable sites for further complement activation by covalently binding to appropriate surfaces.
  1. C3b and C4b contain internal **thioester bonds** that can be rearranged to form ester or amide linkages with biomolecules.
  2. These sites also provide covalently attached ligands for complement receptors.

III. The terminal steps in complement activation are shared by the three pathways.

A. C3 convertases are converted to **C5 convertases** by the binding of C3b (Table 8–1).

B. The C5 convertases cleave C5 generating C5a + C5b.
  1. **C5a** is a biologically active peptide that mediates chemotaxis and mast cell activation (see below).
  2. **C5b** is the first component of the **membrane attack complex (MAC)** (Figure 8–4).
    a. The assembly of the MAC does not require proteolysis.
    b. The assembled C5b678 complex can intercalate into the lipid membranes of both microbial and host cells.
    c. The C5b678 complex recruits multiple copies of C9 to form a stable membrane pore.
    d. Membrane disruption and rapid cell lysis result.

IV. The activation of complement is controlled at several levels (Table 8–2).

A. Some regulatory elements **inhibit** the formation of or **dissociate** active enzyme complexes.
  1. Many such inhibitors (eg, C1 Inh and **C4b-BP**) block early pathway initiation.
  2. Other inhibitors prevent the inadvertent lysis of host cells.
  3. Components, such as **HRF, CD59**, or **decay accelerating factor (DAF)**, are species specific and block complement activation on host, but not microbial, membranes.

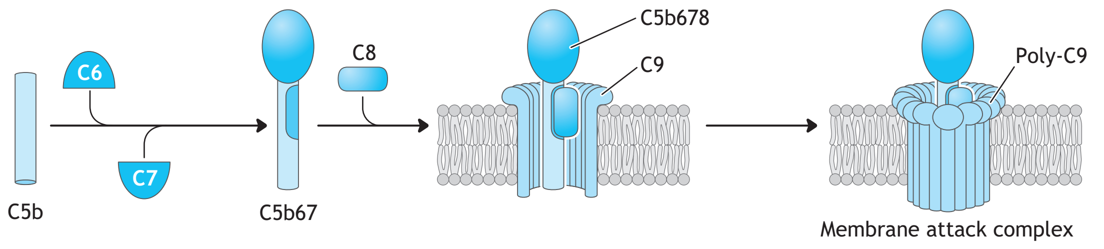

Figure 8–4. Terminal steps in complement activation.

Table 8–2. Regulation of the complement system.

Regulator	Pathway Affected	Mechanism of Action
C1 inhibitor (C1 Inh)	Classical	A serine protease inhibitor that dissociates C1r$_2$s$_2$ from C1q
C4b-binding protein (C4bBP)	Classical	Prevents formation of the classical and lectin pathway C3 convertase by blocking binding of C4b to C2a
CR1	All	Prevents formation of C3 convertases by blocking the binding of C3b to Factor B or the binding of C4b to C2a
Factor P	Alternative	Stabilizes C3bBb
Factor H	Alternative	Blocks formation of C3bBb by binding to C3b
Decay accelerating factor (DAF)	All	Accelerates the decay of the C3 convertases C4b2a and C3bBb
Factor I	All	Cleaves C3b and C4b to produce iC3b and iC4b, respectively
Anaphylatoxin inhibitor (AI)	All	Cleaves and inactivates the anaphylatoxins C3a, C4a, and C5a
S protein	All	Prevents membrane insertion of the C5b67 complex
Hemologous restriction factor (HRF)	All	Prevents binding of poly-C9 to the C5b678 complex

B. Several regulatory factors **inactivate** biologically active complement peptides by proteolysis.

1. **Anaphylatoxin inhibitor (AI)** cleaves the terminal arginine residues of peptides C3a, C4a, and C5a, rendering them inactive.
2. **Factor I** cleaves C3b and C4b, destroying their ability to form C3 convertases.

C. Some factors (eg, S protein) prevent membrane attack by **blocking the assembly** of a mature MAC.

CLINICAL CORRELATION

TRANSGENIC PIGS, MEMBRANE ATTACK, AND XENOTRANSPLANTATION

- *The worldwide shortage of organs suitable for clinical transplantation has stimulated research on the transplantation of organs between species (**xenotransplantation**).*

- *The early rejection of xenografts, including pig grafts in human beings, appears to be mediated by natural human antibodies and complement.*
- *The membranes of pig cells appear to be particularly susceptible to membrane attack, because their complement regulatory proteins do not inhibit human complement components (eg, C3 convertases and MAC).*
- *A novel approach to this dilemma has been to create transgenic pigs expressing human complement regulatory proteins (eg, DAF).*
- *If successful, these animals could be developed as a source of organs for clinical transplantation.*

V. The biological activities of the complement system depend on the action of its peptides and assembled complexes (Table 8–3).

A. Membrane damage is caused by the assembly of the **MAC** in a lipid bilayer.
  1. This is an important mechanism of host cell lysis in transfusion reactions, transplantation rejection, and autoimmunity.
  2. The MAC can lyse the exposed outer membranes of Gram-negative bacteria.
  3. Membrane damage caused by MAC is similar to that caused by perforin produced by cytotoxic T cells (Chapter 6).
    a. Poly-C9 and poly-perforin produce membrane pores.
    b. Perforin and C9 show considerable amino acid sequence homology.

FACTOR I DEFICIENCY

CLINICAL CORRELATION

- *Because Factor I destroys C3 convertases by cleaving C3b and C4b, it regulates complement activation at a key amplification step.*
- *The importance of **Factor I** in this regard is illustrated by individuals with hereditary Factor I deficiency.*

Table 8–3. Summary of important complement components and complexes.[a]

	Classical Pathway	Lectin Pathway	Alternative Pathway
Enzymatic components	$C1qr_2s_2$ C4b2a C4b2a3b	MBL-MASP C4b2a C4b2a3b	Factor D C3bBb C3bBb3b
Biologically active peptides or complexes	C3b C4b C3d C3a C4a C5a $C5b\text{-}9_n$	C3b C4b C3d C3a C4a C5a $C5b\text{-}9_n$	C3b C4b C3d C3a C4a C5a $C5b\text{-}9_n$
Important regulatory components	C1 Inh C4b-BP Factor I	C4b-BP Factor I	Factor P Factor I Factor H

[a] MBL-MASP, mannose-binding lectin-MBL-associated serine proteases; C1 Inh, C1 inhibitor; C4b-BP, C4b binding protein.

- *These patients suffer from recurrent pyogenic infections, including meningococcal meningitis.*
- *The treatment of these patients with recombinant Factor I provides some relief during acute infections, but its long-term clinical effectiveness is unknown.*

B. Many complement peptides exert their biological effects through cell surface **complement receptors (CR)** (Table 8–4).

1. **Complement receptor 1 (CR1)** mediates the clearance of circulating immune complexes that bear C3b or C4b.
   - **a.** Erythrocytes bear the highest densities of CR1 and mediate the greatest share of immune complex clearance.
   - **b.** The complexes are delivered to the liver, where they are taken up by Kupffer cells.
2. **CR2** mediates B cell coactivation when C3d-coated antigens cross-link CR2 and BCR (Figure 8–5).
3. The **opsonic** activity of C3b, iC3b, C4b, and iC4b is due to the binding of particles coated with these peptides to **CR1, CR3,** and **CR4**.
4. The **C3a/4a receptor (CR3a/4aR)** mediates mast cell degranulation (Chapter 13).
5. **C5aR** is a G protein-coupled seven transmembrane chemotactic receptor that mediates the chemoattraction of neutrophils.

C. Microbes have developed elaborate mechanisms of immune evasion by targeting the complement system and diminishing its biological activities.

1. Herpes simplex virus produces a glycoprotein (C) that promotes C3 decay.

Table 8–4. Complement receptors.

Complement Receptor	Ligands	Cell Expressing the Receptor	Principal Biological Activities
CR1	C3b, iC3b C4b, iC4b	Erythrocytes, phagocytic cells	Clearance of immune complexes
CR2	C3d	B cells	B cell coactivation
CR3	C3b, iC3b C4b, iC4b	Phagocytic cells	Immune complex binding, opsonophagocytosis, cell adhesion
CR4	C3b, iC3b C4b, iC4b	Phagocytic cells	Immune complex binding, opsonophagocytosis, cell adhesion
C3a/C4a R	C3a, C4a	Mast cells	Degranulation leading to inflammatory mediator release
C5aR	C5a	Neutrophils Mast cells	Chemotaxis Degranulation and mediator release

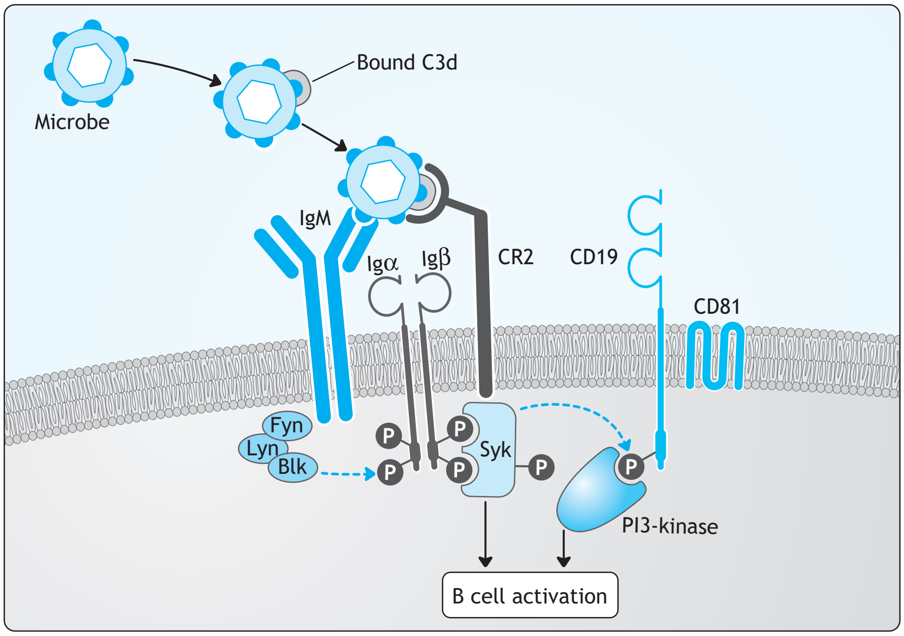

Figure 8–5. Signaling through the B cell coreceptor by binding C3d. Ig, immunoglobulin; PI3-kinase, phosphatidylinositol-3-kinase.

2. The long, branching polysaccharide chains of bacterial surface lipopolysaccharides are thought to direct complement activation away from the outer lipid membrane (Figure 1–3).
3. *Pseudomonas aeruginosa* produces a protease that cleaves C3b.

VI. Complement plays a central role in the pathogenesis of human disease.

A. Serum complement levels can be markedly decreased in **immune complex diseases**.
   1. The circulating levels of C1, C2, C3, and C4 decline when immune complexes containing IgM or IgG antibodies activate the classical pathway.
      a. Total complement activity is measured in **complement hemolytic 50% (CH_{50})** units.
      b. Standard CH_{50} assays measure the activity of the classical pathway.
   2. Regularly measuring total serum complement can be an effective means of monitoring the therapy of immune complex diseases.

B. C3b often deposits within tissues and induces inflammation (Chapter 13).

C. **Primary deficiencies of complement** components cause inflammatory and infectious diseases.
   1. Inherited deficiencies (generally autosomal recessive) of most of the individual complement components have been described.

2. Deficiencies of the early classical and alternative pathways (eg, C1, C2, C4, and Factor B) are most often associated with an increased risk of bacterial infections.
3. The absence of C9 has only a minimum effect on infection rates.
4. Patients with congenital deficiencies of C5, C6, C7, or C8 have an increased risk of disseminated *Neisseria* infections (eg, meningitis).
5. These observations would suggest that complement-dependent lysis of bacteria is *less* important to the host than is its role in the chemotaxis and opsonophagocytosis by neutrophils.

D. **Paroxysmal nocturnal hemoglobinuria (PNH)** is a condition characterized by episodes of intravascular hemolysis and hemoglobinuria.

1. Patients' erythrocytes are particularly sensitive to complement-mediated lysis.
2. PNH patients lack the GPI-linked membrane proteins DAF and CD59.
    - **a.** Membrane C3 convertases are not well controlled, resulting in the deposition of C3b (Table 8–2).
    - **b.** Assembly of the MAC is also not regulated at the cell membrane.

E. **Leukocyte adhesion defect type 1 (LAD-1)** is a deficiency in the expression of complement receptors CR3 and CR4.

1. LAD-1 involves a mutation in the β_2 integrin CD18 (Table 8–5).
2. **CD18** normally forms heterodimers with **CD11a, CD11b,** or **CD11c**.
    - **a.** The CD11a–CD18 dimer is the adhesion molecule **LFA-1**, which facilitates neutrophil binding to endothelial cells and leukocyte emigration from the blood.
    - **b.** The CD11b–CD18 and CD11c–CD18 dimers are **CR3** and **CR4**, respectively, which are adhesion and opsonic receptors on phagocytic cells.
3. LAD-1 patients suffer from recurrent pyogenic infections.

COMPLEMENT DEFICIENCIES AND AUTOIMMUNITY

- *Patients with deficiencies of the early classical pathway components (eg, C2 and C4) have an increased incidence of autoimmune disorders, especially lupus-like nephritis.*
- *This may reflect their inability to clear immune complexes that contain C3b generated by the classical pathway.*
- *Conversely, patients with immune complex-mediated autoimmunity often show depressed levels of C1, C2, C3, and C4 during acute attacks.*
- *C1qr$_2$s$_2$, C4b, C2a, and C3b are removed from the circulation when immune complexes are cleared by CR1-expressing erythrocytes.*

Table 8–5. Leukocyte adhesion defect type 1.[a]

CD18 Heterodimeric Partner	Receptor Formed	Receptor Ligands
CD11a	LFA-1	ICAM-1
CD11b	CR3	iC3b, iC4b
CD11c	CR4	iC3b, iC4b

[a] LFA-1, lymphocyte function-associated antigen-1; ICAM-1, intercellular adhesion molecule 1.

CLINICAL PROBLEMS

A patient presents with recurrent joint pain and renal insufficiency, which is associated with immune complex deposition in these organs. It is determined that the patient does not clear IgG-containing immune complexes from the circulation at a normal rate.

1. A defect in which of the following complement receptors would most contribute to this latter finding?
   A. CR1
   B. CR2
   C. CR3
   D. CR4
   E. CR5

Mary is a 2-year-old child with a history of recurrent upper respiratory tract infections. She has normal serum Ig levels for her age, and her complete blood count (CBC) and tests for neutrophil functions are normal. The total complement activity of her serum is normal in terms of supporting the lysis of antibody-coated erythrocytes. However, her serum complement is not activated by bacterial lipopolysaccharide (LPS).

2. Which of the following primary deficiencies of the complement system would be most consistent with these findings?
   A. C2 deficiency
   B. Factor D deficiency
   C. MBL deficiency
   D. C1 Inh deficiency
   E. DAF deficiency

A 4-year-old child is referred to you with a diagnosis of leukocyte adhesion defect. She has a history of recurrent bacterial infections, including pneumonias and skin abscesses. Radiology demonstrates granuloma-like lesions in her lung and liver. However, the data in the table below suggest this condition has not been properly diagnosed.

3. Which element of the table is **inconsistent** with a diagnosis of LAD?

CD Marker	Patient (%)	Normal Range (%)
CD2	76	75–93
CD3	75	60–85
CD4	52	30–65
CD8	26	20–52
CD11a	65	62–85
CD18	73	70–86
CD19	12	7–17
CD45	52	44–62

A. CD2
B. CD3
C. CD19
D. CD11a
E. CD45

A pediatric patient presents with recurrent staphylococcal infections (pneumonias, sinusitis, and otitis media). Because many of her immune parameters are normal, you hypothesize that she has a defect in her complement system.

4. For which of the following complement components or receptors would a deficiency produce this pattern of recurrent infections?
A. C2
B. C6
C. C8
D. CR1
E. CD59

A patient with the immune complex autoimmune disease systemic lupus erythematosus (SLE) is experiencing significant symptoms, including a worsening of her skin rash and evidence of impaired renal function.

5. If you were to test her serum complement levels during this acute phase of disease, which of the following would be most consistent with a worsening clinical course?
A. A decline in C9 levels
B. A decline in C3 and C4 levels
C. An increase in Factor P levels
D. An increase in C1 Inh levels
E. A decline in C7 levels

ANSWERS

1. The correct answer is A. Complement receptor CR1 is responsible for clearing immune complexes from the circulation by binding to C3b or iC3b. While CR3 and CR3 show the same ligand specificity, the substantially larger number of CR1 receptors expressed by numerous erythrocytes in the blood make this receptor the primary vehicle of immune complex clearance.

2. The correct answer is B. Of the choices, only Factor D is unique to the alternative pathway of complement, which is activated by bacterial LPS. A deficiency of Factor D prevents cleavage of Factor B, which is essential for producing the alternative pathway C3 convertase.

3. The correct answer is D. This patient appears to express normal levels of all of the CD markers shown here. This would *not* be expected of a patient with LAD type 1. Such patients lack CD18, which impairs the expression of CD11a, CD11b, and CD11c.

4. The correct answer is A. Of the components listed, only deficiencies in C2, C6, and C8 result in an increased susceptibility to bacterial infections. With C6 and C8, the pathogens are members of the *Neisseria* genus and certain other gram-negative bacteria. Defects in early classical components, such as C2, result in recurrent pyogenic infections by gram-positive bacteria, such as staphylococci.

5. The correct answer is B. Lupus is an autoimmune disease in which complement is activated by IgG autoantibodies via the classical pathway. This can lead to an acute depletion in the early classical components, such as C1, C2, C3, and C4. The concentrations of the terminal components are affected much less.

CHAPTER 9

B CELL DIFFERENTIATION AND FUNCTION

I. The development of a diverse, self-tolerant population of antigen-specific B cells is central to creating an adaptive immune system.

A. The initial events of **lymphopoiesis** occur in the fetal liver and bone marrow and do not require exposure to foreign antigens.

1. **Hematopoietic stem cells (HSC)**, which express **CD34**, are **pluripotent** and can become any of the blood cell lineages (eg, erythroid, lymphoid, myeloid).
2. Lineage commitment is determined by the **hematopoietic inductive microenvironment**, which includes stromal cell ligands and growth factors (Chapter 12).
3. **Interleukin-7 (IL-7)** is an important growth factor for B and T lymphocyte development.
4. Lymphocyte lineage commitment is indicated by the formation of the **lymphoid progenitor cell.**
5. Commitment to the B lymphocyte lineage occurs at the **progenitor B (pro-B) cell** stage (Table 9–1).
    - a. Most pro-B cells express **CD19** as do all subsequent cells of the B cell lineage.
    - b. **Recombination activating genes 1 and 2 (Rag1, Rag2)** and **terminal deoxynucleotidyltransferase (TdT)** are expressed in preparation for immunoglobulin (Ig) locus rearrangements.
    - c. The Ig H chain locus is rearranged by VDJ recombination, but H chain polypeptides are not expressed.
6. The pro-B cell differentiates into a **precursor-B (pre-B) cell**, which expresses a pre-B cell receptor (BCR).
    - a. The **pre-BCR** consists of a membrane μ chain, an invariant **surrogate L chain,** and Igα and Igβ polypeptides.
    - b. Signaling through the pre-BCR induces allelic exclusion at the Ig H locus and rearrangement of the κ locus.
7. The pre-B cell differentiates into an **immature B cell** expressing an authentic BCR.
    - a. The BCR complex consists of membrane μ chains, κ or λ light chains, and the Igα and Igβ polypeptides.
    - b. Immature B cells exit the bone marrow and complete their differentiation into mature B cells in the periphery.
    - c. Most immature B cells die in the periphery unless they encounter antigen.

Table 9–1. Stages of antigen-independent B cell differentiation.[a]

	HSC	Pro-B	Pre-B	Immature B	Mature B
Igα, Igβ	–	+	+	+	+
Rag1, 2	–	+	+	+	–
TdT	–	+	+	–	–
Surrogate L chain	–	+	+	–	–
Membrane μ chain	–	–	+	+	+
Membrane δ chain				–	+
Membrane κ or λ chain	–	–	–	+	+
Btk kinase	–	+	+	+	+
Positive selection	-	–	+	+	–
Negative selection	–	–	–	+	–

[a] HSC, hematopoietic stem cells; Ig, immunoglobulin; Rag1, Rag2, recombination activating genes 1 and 2; Tdt, terminal deoxynucleotidyltransferase; Btk, Bruton's tyrosine kinase.

8. **Mature B cells** express two forms of the BCR, membrane IgM and membrane IgD (Chapter 4).
   a. **B-1 B cells** are among the first peripheral B cells and are found predominantly in the peritoneal and pleural cavities.
   b. B-1 B cells produce natural antibodies thought to be induced by microbial flora.
   c. B-1 B cells have a limited BCR repertoire.
      (1) Most B-1 B cells produce low-affinity IgM antibodies specific for **polysaccharide antigens**.
      (2) The BCRs of B-1 B cells contain relatively conserved V regions.

B. **Selection** occurs at several differentiation checkpoints and determines the peripheral B cell repertoire.
   1. **Positive selection** rescues bone marrow B cells from apoptosis ("death by neglect") (Table 9–2).
      a. Positive selection occurs at the pre-B and immature B cell stages.
      b. Selection requires signaling through the pre-BCR or BCR.
      c. The pre-BCR signals differentiation to the immature B cell stage and the BCR promotes differentiation into mature B cells.
      d. Positive selection through the BCR probably involves low affinity binding to self-ligands.

Table 9–2. Positive and negative selection of developing B cells.[a]

	Positive Selection	Negative Selection
Stage of B cell development	Pre-B cell Immature B cell	Immature B cell
Receptors	Pre-BCR BCR	BCR
Ligands	Unknown	Self antigens
Events triggered by selection	Allelic exclusion of H locus Proliferation of pre-B cells κ locus rearrangement	Apoptosis BCR editing Secondary κ locus rearrangement
Outcome	Advancement to immature B cell stage Monoclonal expression of μ chain BCR expression	Elimination of autoreactive B cell clones Replacement of autoreactive BCRs

[a] BCR, B cell receptor.

X-LINKED AGAMMAGLOBULINEMIA

- ***X-linked agammaglobulinemia (XLA)*** *is a congenital immune deficiency characterized by the lack of peripheral B cells.*
- ***Agammaglobulinemia*** *becomes apparent only after maternal IgG disappears during the first year of life.*
- *XLA patients have a mutation in the gene coding for* ***Bruton's tyrosine kinase (Btk)****.*
- *Btk is first required for signaling by the pre-BCR at the pre-B cell stage and mediates positive selection.*
- *A similar phenotype exists in patients with H locus deletions that prevent functional μ chain synthesis and pre-BCR expression.*

2. **Negative selection** mediates removal of autoreactive B cell clones.
   - a. Negative selection establishes **self-tolerance**.
   - b. Negative selection occurs at the immature B cell stage and is mediated by high-affinity BCR binding of self antigens.
   - c. Negative selection can signal either apoptosis or anergy within B cells.
   - d. Negative selection can stimulate **BCR editing**.
     - *(1)* The cell undergoes a second light chain locus rearrangement.

(2) The second rearrangement replaces the L chain of the self-reactive BCR.
(3) The B cell undergoes another round of selection based on its new BCR.

II. In the periphery, antigen induces further differentiation of mature B cells into antibody-producing plasma cells and memory cells.

- **A.** Antigen enters the spleen, lymph nodes, and submucosal lymphoid follicles via the blood, the lymph, or by transport across the mucosal epithelium, respectively.
- **B.** A single antigen-reactive B cell can give rise to thousands of daughter cells through 10–-12 cell doublings.
- **C.** Activated B cells differentiate into sessile **plasma cells** that live for only a few days.
    - **1.** Differential RNA processing ensures that plasma cells synthesize secreted, rather than membrane, Igs (Chapter 4).
    - **2.** The progeny of a single B cell can synthesize up to 10^{12} antibody molecules.
- **D.** Antigen-specific, long-lived **memory B cells** also arise during B cell clonal expansion in lymph node **germinal centers.**
    - **1.** **Follicular dendritic cells** promote memory B cell development by retaining antigens over long periods of time.
    - **2.** Germinal center development is T cell dependent.
    - **3.** Memory B cells undergo Ig isotype switching.
    - **4.** Memory B cells mediate **secondary responses** characterized by the following:
        - **a.** A requirement for less antigen to induce the response.
        - **b.** A shorter lag period before antibody is detected.
        - **c.** Higher average affinity of the antibodies produced.
        - **d.** The presence of additional Ig isotypes.
- **E.** **Affinity maturation** accompanies memory B cell development.
    - **1.** Affinity maturation is an increase in the average affinity of an antibody response over time.
    - **2.** Affinity maturation is T cell dependent.
    - **3.** Affinity maturation requires **somatic hypermutation** of Ig V region genes.
        - **a.** Proliferating B cell clones bear mutations in the complementarity determining regions of their BCRs.
        - **b.** The average affinities of the antibodies these cells produce increase by 10- to 100-fold.
        - **c.** Cells that express mutated, high-affinity BCRs are **positively selected** by antigen for additional cycles of proliferation.

III. Antigen-induced activation of B cells is mediated through the BCR, coreceptors, and cytokine receptors (Figure 9–1).

- **A.** The BCRs on mature naive B cells are membrane IgM and IgD.
- **B.** Memory B cells utilize membrane IgG, IgA, or IgE as their BCRs.
- **C.** Each of these BCRs signals through Igα and Igβ and intermediates that are similar to those used by the T cell receptor (TCR) (Chapter 6) (Table 9–3).
- **D.** Signaling is initiated by clustering of the BCR complex.
    - **1.** Polyvalent antigens with repeating identical determinants can activate B cells without coreceptor signals.
    - **2.** Most native protein antigens contain univalent epitopes that do not mediate BCR cross-linking.

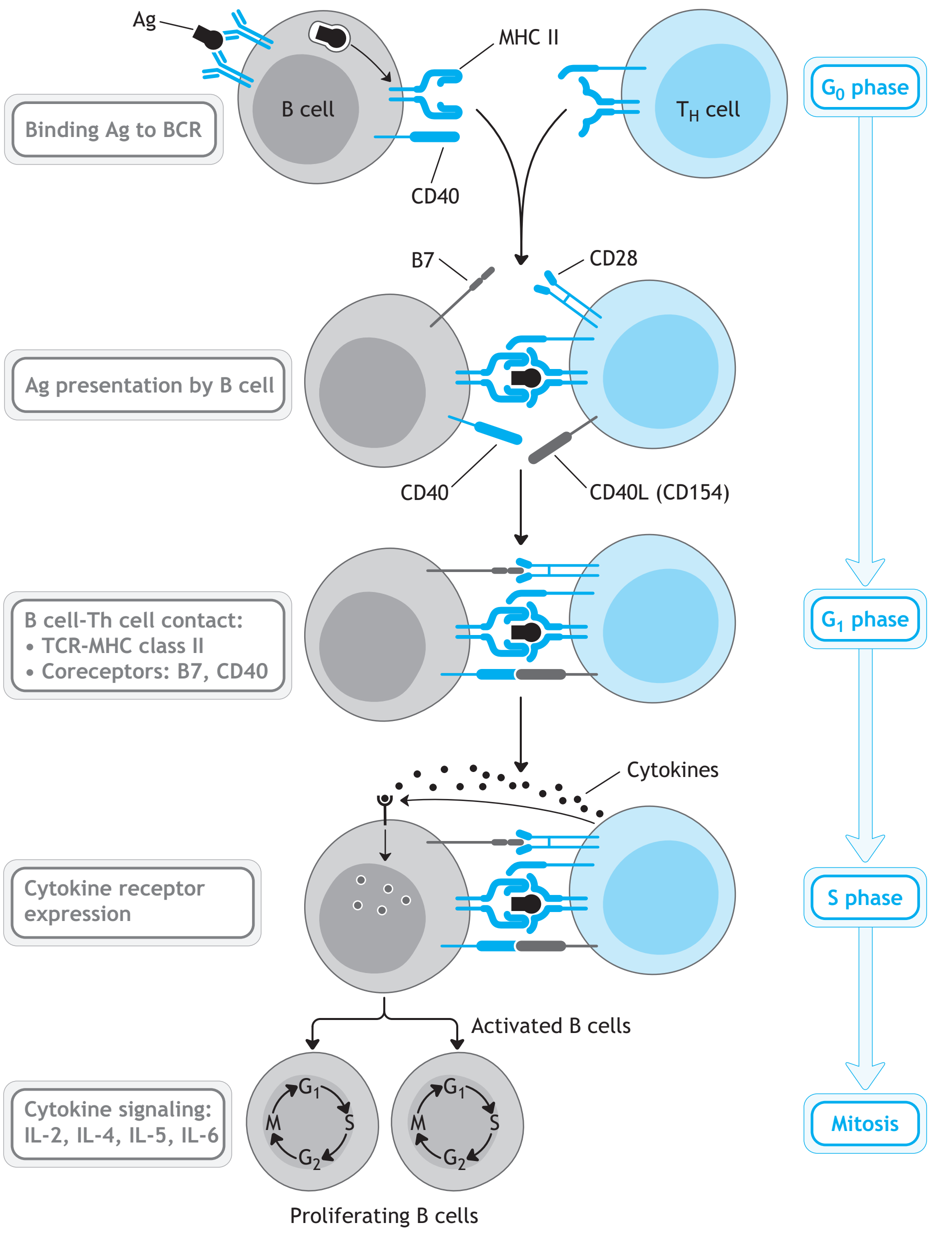
Ag
MHC II
B cell
T_H cell
Binding Ag to BCR
CD40
G_0 phase
B7
CD28
Ag presentation by B cell
CD40
CD40L (CD154)
B cell-Th cell contact:
• TCR-MHC class II
• Coreceptors: B7, CD40
G_1 phase
Cytokines
Cytokine receptor expression
S phase
Activated B cells
Cytokine signaling: IL-2, IL-4, IL-5, IL-6
G_1
M
S
G_2
G_1
M
S
G_2
Mitosis
Proliferating B cells

Figure 9–1. Cooperative signaling for B cell activation by antigen. Ag, antigen; BCR, B cell receptor; TCR, T cell receptor; MHC, major histocompatibility complex; IL, interleukin.

Table 9–3. Analogous components of the TCR and BCR signaling pathways.[a]

Signaling Event or Intermediate	TCR Associated	BCR Associated	Effects
Clustering of receptors in the membrane	TCRαβ or TCRγδ with CD3 peptides	BCR with Igα and Igβ	Concentrate subsequent signaling mediators
Phosphorylation of ITAMs	Phosphorylation of CD3 ITAMs by Lck	Phosphorylation of Igα and Igβ ITAMs by Src kinases	Binding sites for downstream adaptor proteins and kinases
Effector kinase recruitment and activation	ZAP-70	Syk	Phosphorylates downstream adapters and kinases
Phospholipase activation	PLCγ	PLCγ	PIP$_2$ hydrolysis
IP$_3$ and DAG	Ca^{2+} mobilization PKC activation	Ca^{2+} mobilization PKC activation	Calcineurin activation
Calcineurin activation	NFAT dephosphorylation	NFAT dephosphorylation	Transcriptional activation through NFAT
PKC activation of IκB kinase	IκB phosphorylation	IκB phosphorylation	Transcriptional activation through NFκB
Rac, Ras activation of MAP kinases	Fos/Jun phosphorylation	Fos/Jun phosphorylation	Transcriptional activation through NFκB

[a] TCR, T cell receptor; BCR, B cell receptor; Ig, immunoglobulin; ITAM, immunotyrosine activation motif; ZAP-70, ζ-associated protein-70 kDa; PLC, phospholipase C; PIP$_2$, phosphatidylinositol 4,5-biphosphate; IP$_3$, inositol triphosphate; DAG, diacylglycerol; PKC, protein kinase C; MAP, mitogen-activated protein.

3. Igα and Igβ transmit BCR signals across the cell membrane.
    a. The cytoplasmic domains of Igα and Igβ contain **immunotyrosine activation motifs (ITAM)**(Figure 9–2).
    b. **Src family kinases** phosphorylate ITAM tyrosine residues.
4. The tyrosine kinase **Syk** is recruited to the phosphorylated ITAMs, becomes phosphorylated, and phosphorylates downstream adapter proteins (eg, BLNK) and latent kinases.
    a. Btk kinase is recruited to BLNK and activates **phospholipase Cγ (PLCγ)**.
        (1) PLCγ hydrolyzes the membrane phospholipid phosphatidylinositol 4,5-biphosphate (PIP$_2$) to form inositol triphosphate (IP$_3$) and diacylglycerol (DAG).
        (2) IP$_3$ mobilizes intracellular Ca^{2+} and activates **calcineurin**.
        (3) Calcineurin activates the transcription factor **NFAT** by dephosphorylation.

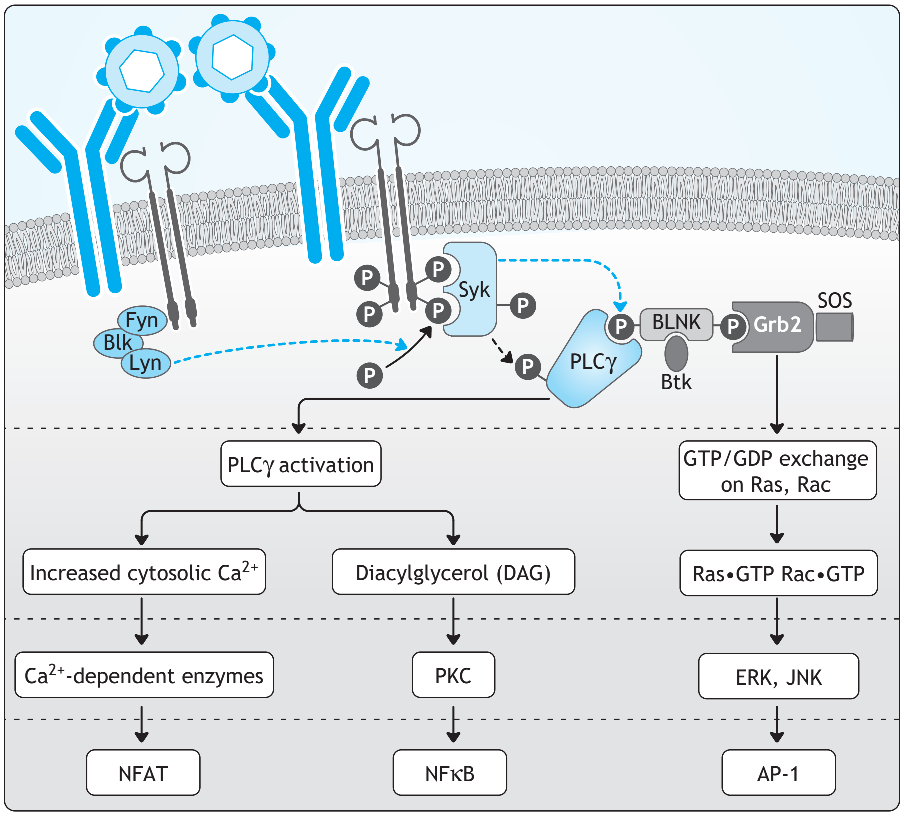
Fyn
Blk
Lyn
P
Syk
P
BLNK
P
Grb2
SOS
Btk
PLCγ
PLCγ activation
GTP/GDP exchange on Ras, Rac
Increased cytosolic Ca2+
Diacylglycerol (DAG)
Ras•GTP Rac•GTP
Ca2+-dependent enzymes
PKC
ERK, JNK
NFAT
NFκB
AP-1

Figure 9–2. B cell receptor (BCR) signaling. PLC, phospholipase C; GTP, guanosine triphosphate; GDP, guanosine diphosphate; PKC, protein kinase C.

(4) **Protein kinase C** is activated by DAG, which indirectly induces the degradation of **IκB**, the inhibitor of **NFκB**.

(5) The transcription factor NFκB is activated.

b. The guanosine triphosphate/guanosine diphosphate (GTP/GDP) exchange proteins **Rac** and **Ras** are activated.

(1) Rac and Ras activate mitogen-activated protein (MAP) family kinases.

(2) The MAP kinases activate the **AP-1** family of transcription factors (eg, Fos and Jun) by phosphorylation.

5. NFAT, NFκB, and AP-1 translocate to the nucleus and initiate gene transcription by binding to their respective enhancers.

6. The transcription of Ig, coreceptor, and cytokine receptor genes is initiated or increased.

E. Coreceptors enhance signals delivered through the BCR (Table 9–4).

1. Contact with T helper cells is required for coreceptor signaling.

Table 9–4. Important receptors and coreceptors on B cells.[a]

Receptor or Coreceptor	Ligand	Biological Response
Pre-BCR	Unknown	Positive selection of pre-B cells
BCR	Self antigens	Positive and negative selection of immature B cells
	Foreign antigens	Activation of peripheral mature B cells
CR2	C3d	Coactivation of mature B cells Enhancement of BCR signaling
CD40	CD154 (CD40L)	Coactivation of mature B cells Enhancement of BCR signaling
IL-2R	IL-2	Growth of activated B cells
IL-4R	IL-4	Switching to IgE
IL-5R	IL-5	Switching to IgA
IL-6R	IL-6	Differentiation of cycling B cells Increased Ig synthesis
CCR7	CCL7	Chemotaxis of germinal center B cells

[a] BCR, B cell receptor; IL, interleukin; Ig, immunoglobulin.

   a. The chemokine **CCL7** mediates chemoattraction of B cells to the outer edge of germinal centers where they bind Th cells.
   b. Contact-dependent signaling is promoted by **major histocompatibility complex (MHC) class II,** on B cells, which is bound by the TCR.
   c. T cell contact-dependent signaling promotes B cell **proliferation**, increases **MHC class II** expression, induces **coreceptor**, **cytokine receptor**, and **chemokine receptor** expression, promotes **affinity maturation**, and induces **Ig class switching**.

2. **CD154** (CD40 ligand or CD40L) on CD4$^+$ Th cells coactivates B cells by binding to CD40.
   a. **CD40** signals B cells to switch Ig class by H chain gene locus rearrangement.
   b. Mutations in the CD154 gene can block T cell-induced Ig class switching in B cells **(hyper-IgM syndrome type 1)**.
3. Complement activation during innate and adaptive immune responses can generate the **B cell coreceptor** ligand **C3d** (Figure 8–5).
   a. C3d can covalently bind to antigens.
   b. C3d–antigen complexes cross-link the BCR with **CR2**, which is composed of the CD19 and CD21 peptides.

  - **c.** Src kinases associated with CR2 promote BCR signaling through phosphorylation.

SWITCH RECOMBINASE DEFICIENCY

- *Ig class switching involves a DNA recombination event mediated by the recombinase* ***activation-induced cytidine deaminase (AID)****.*
- *Mutations in the AID gene have been described and result in impaired antigen-induced isotype switching, somatic hypermutation in B cells, and affinity maturation of the antibody response.*
- *The resulting phenotype, designated* ***hyper-IgM syndrome type 2****, resembles the X-linked CD154 deficiency known as* ***hyper-IgM syndrome type 1****.*

F. Cytokines produced by T helper cells promote B cell activation.
  1. Interleukin (IL)-2, IL-4, and IL-5 promote B cell proliferation.
  2. IL-6 enhances the differentiation of activated B cells into antibody-producing plasma cells.
  3. IL-2, IL-4, and IL-6 promote antibody synthesis by activated B cells and plasma cells.
  4. Several cytokines promote Ig class switching.
    a. **Switch cytokines** promote specific **switch recombinase** interactions with specific switch sites within the Ig H gene locus (Chapter 4).
    b. IL-4 and IL-13 promote switching to IgE.
    c. Interferon (IFN)-γ promotes switching to IgG_1 and IgG_3.
    d. IL-5 and transforming growth factor-β (TGF-β) induce switching to IgA.

ANTICYTOKINE THERAPIES FOR CONTROLLING B CELL ACTIVATION

- *Monoclonal antibodies capable of neutralizing cytokines or blocking their receptors have potential for the treatment of allergic or neoplastic diseases involving B cells.*
- *For example, atopic allergies could theoretically be treated by blocking B cell switching to IgE synthesis with anti-IL-4 or anti-IL-13.*
- *Another application undergoing clinical trials is the use of anti-IL-6 or anti-IL-6 receptor antibodies to inhibit the growth and Ig production of myeloma cells.*

IV. Foreign polysaccharides, glycolipids, and nucleic acids induce antibody production without the need for T cell help.

A. These **T-independent (TI) antigens** contain repeating epitopes that cross-link multiple BCRs on a single B cell.

B. Some TI antigens (eg, bacterial LPS) also coactivate B cells through Toll-like receptors (Chapter 1).

C. Some TI antigens [eg, lipopolysaccharide (LPS)] can activate complement and coactivate B cells through CR2.

D. Because TI antigens are not presented by antigen-presenting cells (APCs), they do not activate $CD4^+$ Th cells.

E. The responses to TI antigens differ from responses to foreign proteins.
  1. There is little **Ig class switching** in TI antibody responses; IgM and IgG_2 antibodies predominate.
  2. A limited repertoire of antibody-mediated **effector functions** results.

3. **Memory B cell** populations are not formed and **affinity maturation** through somatic hypermutation does not occur.
4. **High-titered antibody** responses are not seen on secondary challenge.

SELECTIVE IgG$_2$ DEFICIENCY

CLINICAL CORRELATION

- *IgG$_2$ is a common subclass of antibody produced in response to T-independent antigens in humans.*
- *Antibody responses to protein antigens are predominantly of the IgG$_1$ subclass.*
- *Children with selective IgG$_2$ subclass deficiency have difficulty clearing bacteria that express polysaccharide capsules (eg, Streptococcus pneumoniae and Haemophilus influenzae).*
- *The increased survival of encapsulated bacteria in IgG$_2$-deficient patients suggests that other (sub)classes of Igs do not provide sufficient host defense against these organisms.*

V. Antibody responses at mucosal surfaces are mediated by a specialized set of B cells that synthesizes IgA antibodies.

A. Most of the IgA in the body is synthesized in the small intestine.

B. IgA-secreting plasma cells are most abundant within the lamina propria of the submucosum and produce 2 g of Ig per day.

C. The secretory form of IgA is the central mediator of mucosal humoral immunity.
   1. **Secretory IgA** is dimeric and contains J chain and **secretory component (SC)**(Chapter 3).
      a. The α, κ, λ, and J chains of dimeric IgA are produced by mucosal B cells.
      b. Secretory component is synthesized by the intestinal **epithelial cell**.
   2. Secretory component mediates transepithelial transport of IgA.
      a. On the basolateral surface of epithelial cells, a precursor of SC called **poly-Ig receptor** is expressed (Figure 9–3).
      b. The poly-Ig receptor binds the polymeric Igs (IgA and IgM).
      c. The loaded receptor is internalized into endosomes, which are translocated to the apical cell surface.

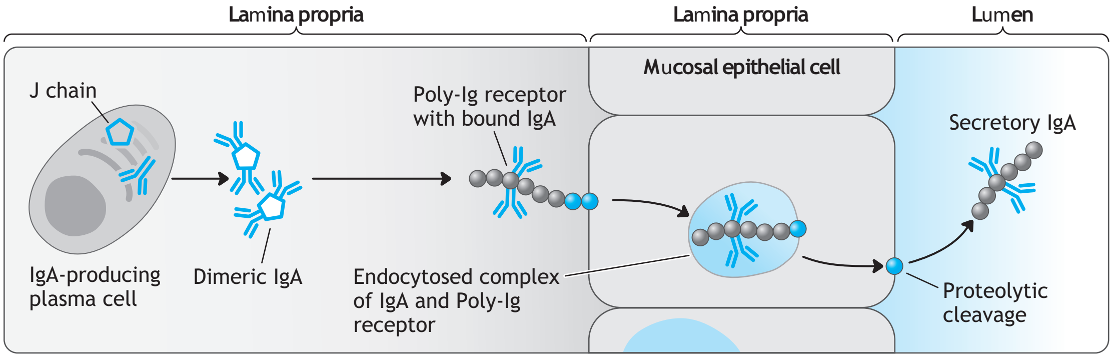
Lamina propria
Lamina propria
Lumen
Mucosal epithelial cell
J chain
Poly-Ig receptor with bound IgA
Secretory IgA
IgA-producing plasma cell
Dimeric IgA
Endocytosed complex of IgA and Poly-Ig receptor
Proteolytic cleavage

Figure 9–3. Transport of dimeric immunoglobulin A (IgA) by poly-Ig receptor.

   d. Proteolytic cleavage of the poly-Ig receptor releases secretory component with its polymeric Ig attached.

D. At the mucosal surface, secretory IgA neutralizes toxins and allergens and prevents microbial entry.

SELECTIVE IGA DEFICIENCY

CLINICAL CORRELATION

- *Selective IgA antibody deficiency is the most common antibody deficiency in humans with frequencies as high as 1:333 reported in some populations.*
- *This deficiency is characterized by recurrent bacterial and viral infections originating at mucosal surfaces (respiratory, gastrointestinal, and genitourinary tract infections).*
- *Increased incidences of food allergies, autoimmunity, and certain types of cancers have also been reported in these patients.*
- *However, half of all persons with this deficiency are asymptomatic.*

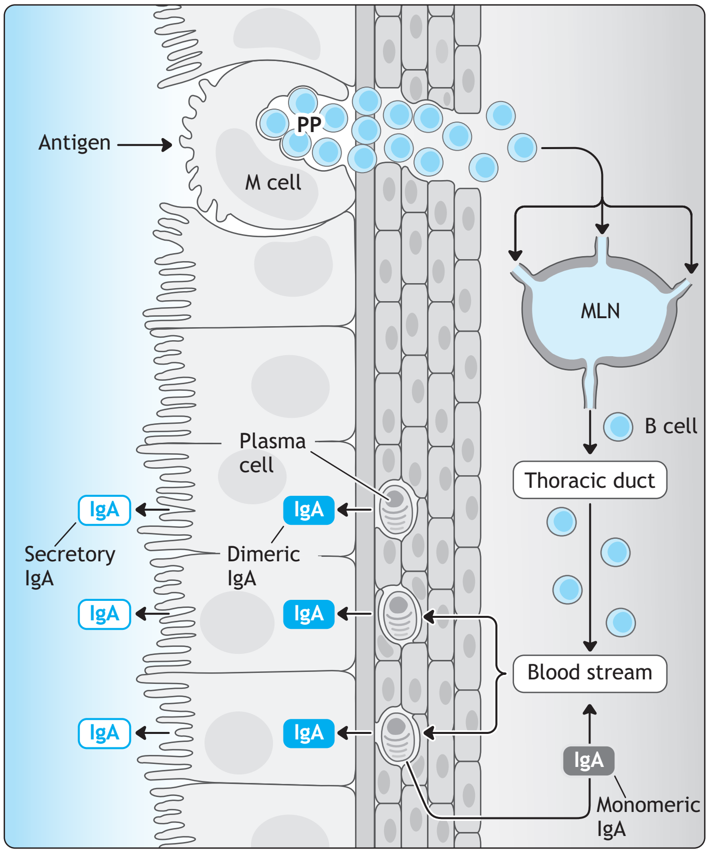
Antigen
PP
M cell
MLN
B cell
Thoracic duct
Plasma cell
IgA
IgA
Secretory IgA
Dimeric IgA
IgA
IgA
Blood stream
IgA
IgA
IgA
Monomeric IgA

Figure 9–4. Induction of a mucosal immunoglobulin A (IgA) antibody response. M, microfold; PP, Peyer's patch; MLN, mesenteric lymph node.

- *The absence of increased rates of infection in the unaffected subset of IgA-deficient individuals probably reflects the translocation of secretory IgM by the poly-Ig receptor.*

E. Mucosal IgA antibody responses are initially induced within submucosal Peyer's patches of the small intestine (Figure 9–4).

1. The epithelium overlying the patch is composed of specialized nonvillous epithelial cells, called **M cells**, which efficiently transport antigens to lymphoid cells within the patch.
2. Peyer's patch Th cells produce IL-5 and TGF-β, two IgA switch cytokines.
3. B cells activated in the Peyer's patches migrate via draining lymph nodes into the lymph and blood.
4. Circulating IgA-expressing B cells then seed the intestine and other **mucosal-associated lymphoid tissues (MALT)**.
5. By this process, the recognition of enteric antigens leads to the production and transport of secretory IgA at multiple mucosal tissue sites.

MUCOSAL AND PARENTERAL VACCINES

CLINICAL CORRELATION

- *In the 1950s two competing vaccines were produced against the polio virus.*
- *The parenteral (Salk) vaccine was injected by the intramuscular route and induced IgG antibody production.*
- *These antibodies prevented disease by neutralizing virus particles in transit from the gastrointestinal tract to the nervous tissues.*
- *The Sabin vaccine was an inactivated virus, and oral immunization resulted in secretory IgA responses in the gut-associated lymphoid tissues.*
- *Secretory IgA antibodies prevented viral attachment and entry into intestinal epithelial cells.*
- *Current recommendations are for children to receive the parenteral vaccine at 2, 4, and 6 months of age (www.cdc.gov/nip/recs/child-schedule.PDF).*

CLINICAL PROBLEMS

A patient is shown to have a mutation in his *Btk* tyrosine kinase gene.

1. Which of the following stages of lymphocyte differentiation is blocked in this patient?
   A. Differentiation of mature B cells into plasma cells
   B. Generation of memory B cells
   C. Differentiation of pre-B cells into immature B cells
   D. Negative selection of mature B cells
   E. Activation of mature B cells

A patient presents with recurrent bacterial infections and is found to have low serum levels of IgA and IgG_1, but elevated concentrations of serum IgM. Flow cytometry of his lymph node cells reveals an absence of CD154.

2. In which of the following areas of the lymph node would this cell surface marker normally be found?
   A. Medullary cords
   B. Diffuse cortex
   C. Germinal centers
   D. High endothelial cells of the venules
   E. Subcapsular sinuses

Lymphocytes from an immunodeficient patient are found to lack mRNA for AID.

3. Which of the following lymphocyte subsets would be expected to be absent in this patient?
   A. Lymph node B cells expressing membrane IgG
   B. CD4^{+} T lymphocytes in the blood
   C. Natural killer (NK) cells in the spleen
   D. All lymphocytes in peripheral lymphoid tissues
   E. CD19^{+} B cells in the bone marrow

You have a patient with recurrent infections, but normal T cell function.

4. Which of the following laboratory tests would help you determine if this immune deficiency is due to an *absence* of B cells?
   A. Radioimmunoassay
   B. Flow cytometry
   C. Enzyme-linked immunosorbent assay (ELISA)
   D. Immunofixation
   E. Coombs test

A lymph node biopsy stained with hematoxylin and eosin shows numerous germinal centers in the superficial cortex.

5. Which of the following best describes the source of this biopsy tissue?
   A. A patient who lacks CD154
   B. A patient who lacks Rag1
   C. A patient who has been immunized repeatedly with a protein vaccine
   D. A patient who has been immunized repeatedly with streptococcal polysaccharides
   E. A newborn

ANSWERS

1. The correct answer is C. Whereas the Btk kinase is used for both BCR and pre-BCR signaling, an absence of Btk is manifest first in pre-B cells. The mutation causes defective positive selection of pre-B cells and apoptotic death. There are no peripheral immature or mature B cells in Btk-deficient individuals, a condition known as XLA.
2. The correct answer is B. CD154 would normally be found on Th cells in the diffuse or deep cortex. Its absence is indicative of hyper-IgM syndrome type 1.
3. The correct answer is A. The switch recombinase AID is normally expressed in B cells that are undergoing Ig isotype switching.
4. The correct answer is B. Only flow cytometry would be able to demonstrate the absence of B cells. The other techniques would not distinguish between an absence of B cells and the inability of those cells to secrete antibody.
5. The correct answer is C. Germinal center development is antigen, Th cell, and CD40 dependent. The antigen must be a T-dependent antigen, such as a protein. Polysaccharide antigens do not induce germinal center formation, and newborns have generally not encountered protein antigens in utero that elicit this type of response.

CHAPTER 10

T CELL DIFFERENTIATION AND FUNCTION

I. The development of T lymphocytes is similar to that of B cells, although unique events occur in the inductive environment of the thymus.

- **A.** Thymic maturation does not require exposure to foreign antigen.
- **B.** Like B cells, T cells are derived from pluripotent **hematopoietic stem cells** in the fetal liver and bone marrow, which differentiate into common **lymphoid progenitor cells** and **progenitor T (pro-T) cells** (Table 10–1).
- **C.** The pro-T cell begins its migration from the bone marrow to the thymus during the first trimester of human pregnancy.
- **D.** In the thymus the cells are called **thymocytes** and undergo considerable cell division and apoptotic death.
    - **1.** Differentiation is induced by **stromal cells** and **growth factors**.
        - **a.** Epithelial cells, macrophages, and dendritic cells provide both contact-dependent and secreted signals.
        - **b.** **Interleukin (IL)-7** is a key growth-promoting factor for thymocytes.
        - **c.** Major histocompatibility complex (MHC) molecules promote thymocyte growth and apoptosis.
    - **2.** Thymocytes are subjected to both positive and negative selection based on the specificity of their T cell receptors (TCRs).
- **E.** The production of mature T cells occurs throughout life, but the thymus becomes less important as it atrophies with age.

THE BUBBLE BOY

CLINICAL CORRELATION

- *In the 1976 movie "Boy in a Bubble," a child with a severe immunodeficiency avoided life-threatening infections by living in a pathogen-free plastic tent.*
- *A number of molecular defects can cause this condition, known as **severe combined immune deficiency**.*
- *These include mutations in the genes for the Rag recombinases, adenosine deaminase, and the γ chain of the IL-2 receptor (Chapter 16).*
- *The microbial pathogens that establish early opportunistic infections in these children include intracellular fungi, viruses, and bacteria.*
- *While B cell differentiation can also be directly or indirectly affected (depending on the mutation), maternal IgG provides immune protection early in life.*
- *Hematopoietic stem cell transplantation provides the only practical cure.*

Table 10–1. Stages of T cell differentiation.[a]

	HSC	**Pro-T**	**Pre-T**	**Double Positive**	**Single Positive**	**Naive Mature T Cell**
Rag	–	+	–	+	–	–
TdT	–	+	–	–	–	–
TCR DNA	Germline	Germline	Rearranged β chain	Rearranged β and α chain	Rearranged β and α chain	Rearranged β and α chain
TCR	None	None	Pre-TCR	TCR	TCR	TCR
CD3	–	–	+	+	+	+
Surface markers	–	–	–	CD4$^+$8$^+$	CD4$^+$ or CD8$^+$	CD4$^+$ or CD8$^+$
Anatomic site	Bone marrow	Thymus	Thymus	Thymus	Thymus	Periphery
Response to TCR ligation	None	None	Positive selection	Positive and negative selection	Negative selection	Activation

[a] Tdt, terminal deoxynucleotidyltransferase.

II. Thymocyte differentiation is accompanied by the generation of the TCR repertoire and the establishment of functional cell subsets.

A. Four important events in the development of the T cell lineage occur.
  1. TCRαβ and TCRγδ genes are rearranged by recombination (Chapter 6) and are expressed.
  2. The MHC restriction specificity of the T cell lineage is established.
  3. Thymocytes with autoreactive TCRs are eliminated.
  4. Two major functional T cell subsets (CD4$^+$ and CD8$^+$) are generated.

B. Clonal proliferation, apoptotic death, and differential gene expression accompany thymocyte differentiation.
  1. **Pro-T** cells migrate to the superficial cortex and become part of the pool of **double-negative (CD4$^-$CD8$^-$)** thymocytes (Table 10–1 and Table 10–2).
     - **a.** Pro-T cells begin to express recombination activating genes 1 and 2 (Rag1, Rag2) and terminal deoxynucleotidyltransferase (TdT) in preparation for TCRβ locus rearrangements.
     - **b.** In most pro-T cells, the β locus undergoes D–J rearrangement.
     - **c.** Approximately 5% of the descendants of these cells will express γδ TCRs and the remainder will become TCRαβ$^+$.
  2. Pro-T cells then differentiate into **pre-T cells**, in which V–DJ$_\beta$ joining occurs and a β chain polypeptide is expressed.

Table 10–2. The major subsets of the TCR$\alpha\beta$ lineage of thymocytes.[a]

Thymocyte Subset	Surface Phenotype	Major Events or Properties
Double negative		
Pro-T cell	CD4$^-$CD8$^-$TCR$^-$	TdT$^+$Rag1,2$^+$; D-J$_\beta$ rearrangement
Pre-T cell	CD4$^-$CD8$^-$TCR$^-$CD3$^+$pre-TCR$^+$	Rag1,2$^+$; V-DJ$_\beta$ rearrangement
Double positive	CD4$^+$CD8$^+$CD3$^+$TCR$\alpha\beta^+$	V-J$_\alpha$ rearrangement Positive and negative selection
Single positive		
CD4$^+$	CD4$^+$CD8$^-$CD3$^+$TCR$\alpha\beta^+$	Negative selection
CD8$^+$	CD4$^-$CD8$^+$CD3$^+$TCR$\alpha\beta^+$	Negative selection

[a] TCR, T cell receptor.

  - **a.** A **pre-TCR** is expressed that contains a β chain plus a **pre-Tα chain** and CD3 peptides.
  - **b.** Activation of pre-T cells through the pre-TCR stimulates cell proliferation, β locus allelic exclusion, α locus rearrangement, and the expression of CD4 and CD8.

3. **Double positive** thymocytes express TCR$\alpha\beta$ with CD3 peptides.
   - **a.** Rag1,2 are expressed for a second time to facilitate rearrangement of the α locus.
   - **b.** TCR$\alpha\beta$ binding to MHC molecules signals further maturation.
     - *(1)* Rag1,2 genes become silent.
     - *(2)* Positive and negative selection occurs as the cells transition to a single positive (CD4$^+$8$^-$ or CD4$^-$8$^+$) state.
4. **Single positive** cells emerge from the thymus and undergo further maturation in the peripheral lymphoid tissues.
5. The maturation of thymocytes can be monitored by polychromatic flow cytometry (Figure 10–1).
   - **a.** Using antibodies to CD4 and CD8, each with a different fluorochrome, four cell subsets can be identified.
     - *(1)* Double negative (CD4$^-$CD8$^-$) cells are the pro-T and pre-T cells.
     - *(2)* Double positive (CD4$^+$CD8$^+$) cells express TCR$\alpha\beta$ or TCR$\gamma\delta$ and are positively and negatively selected based on their TCR specificities.
     - *(3)* Single positive (CD4$^+$CD8$^-$ or CD4$^-$CD8$^+$) cells are fairly mature and ready to exit the thymus.
   - **b.** A similar analysis of peripheral T cells would identify only the two single-positive subsets.

CLINICAL CORRELATION

SIGNIFICANCE OF ORAL THRUSH IN INFANTS

- *A cardinal clinical sign of defective cellular immunity in a neonate is the appearance of mucocutaneous candidiasis or* ***thrush****.*

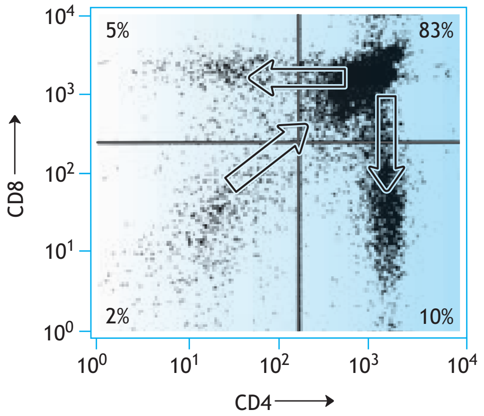

Figure 10–1. Identification of the major subsets of thymocytes by flow cytometry. (Courtesy of Thomas Yankee, University of Kansas Medical Center)

- *Oral thrush is uncommon in immunocompetent individuals.*
- *This painful condition is caused by a superficial infection by the opportunistic yeast* Candida albicans, *which is normal flora in the oral cavity.*
- *The most frequent immune deficiency causing neonatal thrush is AIDS.*
- *However, any cellular immune deficiency, including severe combined immunodeficiency (SCID), selective T cell deficiencies (eg, DiGeorge syndrome, CD3 mutations), and MHC deficiencies, can increase the risk of thrush.*

III. Host MHC molecules determine the specificities of the TCRs that survive thymic selection.

A. The **potential repertoire** of TCRαβ and TCRγδ receptors is determined by the quasirandom recombination events within the DNA that codes for TCR V regions (Chapter 6).

B. The **utilized repertoire** of TCRs results from the selection of receptors with the ability to recognize foreign peptides presented by the MHC of the host.

C. The affinity with which thymocyte TCRs bind MHC molecules in the thymus determines whether positive or negative selection will occur.
   1. In the absence of binding, thymocytes undergo death by neglect.
   2. Low-affinity TCR–MHC interactions trigger positive selection.
   3. High-affinity TCR binding to MHC signals negative selection.

D. Positive selection promotes the development of thymocytes with TCRs specific for foreign peptides plus the host's own MHC molecules.
   1. Cells bearing TCRs specific for non-self MHC (ie, that of another individual) die by neglect.
   2. **Positive selection** occurs at the double-positive thymocyte stage and induces differentiation to the single-positive phenotype.
      a. Double-positive thymocytes with TCRs that bind to self MHC *class I* molecules are induced to become CD4$^-$CD8$^+$.

b. Double positive thymocytes with TCRs that bind to MHC *class II* molecules become CD4$^+$CD8$^-$ single positive.
c. Clones that express TCRs specific for nonhost MHC class I or II are not positively selected and die by default.
d. The majority of developing thymocytes are *not* positively selected and undergo apoptosis.

MHC DEFICIENCIES PREVENT POSITIVE SELECTION

CLINICAL CORRELATION

- *Two important groups of immune deficiency diseases result from inefficient positive selection of thymocytes.*
- ***Bare lymphocyte syndrome (BLS)*** *type 1 and type 2 (Chapter 7) are congenital deficiencies in which MHC class I and class II molecules, respectively, are not expressed.*
- *The absence of MHC molecules precludes positive selection of double positive thymocytes and the corresponding T cell subset fails to develop.*
- *For example, BLS type 2 patients are lymphopenic, because CD4$^+$ T cells, which normally constitute the majority of T cells in the blood, do not develop.*
- *The differential diagnosis of BLS type 2 requires exclusion of HIV-1 infection, which is a much more common cause of decreased CD4:CD8 ratios (Chapter 15).*

E. **Negative selection** eliminates potentially harmful thymocyte clones that bear high affinity receptors for self peptides plus self MHC molecules.
  1. Negative selection establishes **central tolerance** to self antigens (Chapter 16).
  2. Negative selection occurs at the double positive-to-single positive transition.
  3. Both MHC class I and class II can signal negative selection.
  4. Coreceptor signaling through CD4 and CD8 promotes negative selection.
  5. The death signal during negative selection is delivered through **Fas–Fas ligand** interactions between thymocytes and stromal cells of the thymus.

F. Four potential outcomes result from the binding of peptide–MHC complexes by TCRs.
  1. Thymocyte clones are induced to proliferate and differentiate (positive selection).
  2. Thymocyte clones undergo apoptosis (negative selection).
  3. Peripheral T cells proliferate and differentiate in response to foreign antigens.
  4. Peripheral T cells can become anergic (**peripheral tolerance**) if coreceptor signaling is lacking (Chapter 16).

DEFECTIVE THYMIC APOPTOSIS

CLINICAL CORRELATION

- *Apoptotic death regulates thymocyte selection and T cell activation in the periphery.*
- ***Autoimmune lymphoproliferative syndrome (ALPS)*** *(also known as Canale–Smith syndrome) is a defect in the apoptosis of activated T cells.*
- *In most patients a mutation exists in the gene coding for* ***Fas*** *(CD95).*
- *Patients present with greatly enlarged lymph nodes, splenamegaly, and autoimmunity (eg, Coombs-positive hemolytic anemias).*
- *Affected individuals also show lymphocytosis and an elevated number of double negative ($TCR\alpha\beta^+CD4^-CD8^-$) T cells in the blood.*
- *Similar phenotypes occur with patients bearing mutations in Fas ligand or certain caspase genes.*

IV. Cell surface adhesion molecules, coreceptors, and cytokine receptors regulate T cell activation in the periphery.

A. **Cell adhesion molecules** promote T cell homing to specific tissues by binding to ligands on vascular endothelial cells and matrix proteins.
  1. T cells express **integrins**, such as **lymphocyte function-associated antigen-1 (LFA-1)**, that mediate lymphocyte homing to sites of infection and inflammation.
  2. Specialized integrins direct T cells to mucosal lymphoid tissues.
  3. Integrins increase the avidity of binding between T cells and their antigen-presenting cells.
  4. Integrins promote binding between cytotoxic T cells and their target cells.

B. **Coreceptors** are induced and recruited to the **immunological synapse** that forms between a T cell and its antigen-presenting cell.
  1. The cell surface expression of coreceptors and their ligands is often regulated to avoid the unwanted activation of resting T cells.
  2. Receptor clustering brings coreceptors in proximity to TCRs.
  3. Receptor clustering promotes synergy between intracellular signaling pathways.
  4. **CD4** and **CD8** are coreceptors that bind to nonpolymorphic residues of MHC class II and class I molecules, respectively.
    a. The T cell-specific kinase **Lck** is associated with CD4 and CD8 in the membrane.
    b. Lck phosphorylates immunotyrosine activation motif (ITAM) residues on CD3 peptides.
    c. Thymocytes lacking Lck are not positively selected at the double-positive stage.
  5. **CD28** and **cytotoxic T lymphocyte-associated protein 4 (CTLA-4)** (CD152) on T cells modify TCR signaling.
    a. The **costimulatory ligands** for these coreceptors are members of the B7 family and are expressed on antigen-presenting cells.
    b. Ligation of CD28 on naive T cells by B7 induces phosphatidylinositol-3-kinase activation.
    c. Stimulation of T cells through CD28 activates NFκB.
    d. CTLA-4 is expressed on activated T cells and mediates negative signaling when B7 is bound.
  6. **CD154** is a coreceptor for $CD4^+$ Th cells.

C. **Cytokine receptors** augment or inhibit T cell responses to antigen.
  1. The expression of cytokine receptors is often induced by antigen and coreceptor stimulation.
  2. **IL-2** is a growth factor for activated $CD4^+$ and $CD8^+$ T cells.
  3. **IL-4** promotes the differentiation of the Th2 subset from naive, antigen-stimulated $CD4^+$ T cells.
  4. **Interferon (IFN)-γ** promotes the differentiation of the Th1 subset of T cells.
  5. **IL-10** and **transforming growth factor (TGF-β)** inhibit the activation of Th1 cells.
  6. **IFN-γ** inhibits the activation of Th2 cells.

D. Different types of costimulatory and cytokine signals are derived from different types of antigen-presenting cells (Table 10–3).

Table 10–3. Differences between APCs related to T cell activation.[a]

Property	Dendritic Cell	B Cell	Macrophage
Antigen uptake	Active endocytosis	BCR-mediated endocytosis	Endocytosis and phagocytosis
MHC class II expression	Constitutive	Constitutive	Inducible
Costimulatory ligand expression	Constitutive B7	Inducible B7	Inducible B7
T cells activated	Naive, effector and memory	Effector and memory	Effector and memory

[a] APC, antigen-presenting cells; BCR, B cell receptor.

1. **Dendritic cells** are particularly effective at activating naive $CD4^+$ T cells, because they express B7 and MHC class II constitutively.
2. **B cells** have the advantage of being able to capture limited quantities of antigen through their BCR, which aids in activating memory T cells with high-affinity TCRs.
3. **Macrophages** can phagocytize particulate antigens and present their epitopes to T cells.

V. The activation of mature T cells by antigen in the periphery leads to clonal expansion and differentiation into effector and memory cell subsets.

A. The frequency of T cells specific for a given antigen changes during an immune response.
   1. Clonal frequencies among resting naive T cells is approximately 10^{-6}.
   2. The frequency of antigen-specific T cells increases to 10^{-2} at the peak of the expansion phase.
   3. Memory cells for a given antigen exist at a frequency of 10^{-4}.

B. $CD4^+$ helper T cells (Th cells) can be subdivided into two subsets (Figure 6–5).
   1. **Th1 cells** mediate cellular immunity to intracellular microbial pathogens by the secretion of cytokines [eg, IL-2, IFN-γ, and tumor necrosis factor (TNF)-α] that promote T cell growth and activate macrophages and neutrophils.
   2. **Th2 cells** promote humoral immunity to extracellular microbial pathogens by producing cytokines (IL-4, IL-5, and IL-13) that activate B cells.

C. **Memory T cells** divide at a low rate and recirculate for decades.
   1. The chemokine receptor CCR7 and certain adhesion molecules facilitate memory T cell migration into lymph nodes.
   2. The antigen-presenting cell (APC) signals required to activate a memory T cell with antigen are different than those required by naive T cells (Table 10–2).
   3. Memory T cells become a greater proportion of the T cell pool with age.

D. **Regulatory T (Treg) cells** inhibit immune responses (Chapter 11).
  1. Treg cells are most often CD4$^+$ and express **CD25**, the α chain of the high-affinity IL-2 receptor.
  2. Treg cells produce the inhibitory cytokines **IL-10** and **TGF-β**.
  3. Treg cells contribute to the maintenance of self tolerance (Chapter 16).

E. Several subsets of T cells and related lymphocytes are **cytotoxic** (Table 10–4).
  1. **Cytotoxic T lymphocytes (CTL)** are induced by antigen, terminally differentiated and short lived.
    - a. Most CTL are CD8$^+$ and MHC class I restricted.
    - b. Cytotoxic CD4$^+$ cells exist, but are MHC class II restricted.
    - c. Both TCR$\alpha\beta$ and TCR$\gamma\delta$ T cells can be cytotoxic.
    - d. Cytotoxicity is mediated by the granzyme–perforin pathway, the Fas–Fas ligand pathway, and the TNF receptor pathway (Chapter 6).
  2. The **$\gamma\delta$ T cell** subset shows specificity for pathogens associated with epithelial surfaces.
    - a. $\gamma\delta$ T cells are **intraepithelial lymphocytes** that accumulate in the skin, small intestine, lung, and genitourinary tract.
    - b. Most TCR$\gamma\delta^+$ cells do not express CD4 or CD8.
    - c. TCR$\gamma\delta$ is specific for nonpeptides, including glycolipids that represent pathogen-associated molecular patterns.
    - d. TCR$\gamma\delta$ is not MHC restricted.

Table 10–4. A comparison of conventional T cells, $\gamma\delta$ T cells, NKT cells, and NK cells.[a]

Property	Conventional T Cell	$\gamma\delta$ T Cell	NKT Cell	NK Cell
Tissue location	Spleen, lymph nodes, lymphoid follicles	Intraepithelial	Thymus, liver, spleen	Liver > spleen
Receptor type	TCR$\alpha\beta$ + CD3	TCR$\gamma\delta$ + CD3	Invariant TCR$\alpha\beta$ + CD3	Activating FcγR NKG2D Inhibitory KIR Lectins
Receptor specificity	Peptide + MHC class I or class II	Microbial glycolipids	Microbial glycolipids Stress-induced host ligands There are two receptor types and two receptor specificities: activating and inhibitory	Activating IgG antibody Stress-induced host ligands Inhibitory HLA-A,B,C

[a] NKT, natural killer T; TCR, T cell receptor; MHC, major histocompatibility complex; KIR, killer inhibitory receptor; IgG, immunoglobulin G.

e. Signaling through **TCRγδ** is similar to that of TCRαβ.
f. γδ T cells can be cytotoxic or cytokine producing.
3. Natural killer T (NKT) cells express both TCRs and NK cell markers.
a. The invariant TCRαβ of NKT cells is restricted by CD1 and specific for microbial glycolipids.
b. NKT cells respond rapidly to antigen and produce IL-4 and IFN-γ.

F. Although not thymus derived, **NK cells** share a number of properties with T cells and NKT cells.
1. NK cells recognize host cells infected with intracellular microbial pathogens using unique receptors.
a. NK cell **inhibitory receptors** recognize MHC class I molecules and deliver signals that are dominant over NK cell-**activating receptor** signals.
b. NK cell-**activating receptors** recognize host cell ligands that are present on infected or stressed cells.
c. When host cells fail to express MHC class I (as during virus infections), the inhibitory signal is lost and the NK cell becomes activated.
d. Activating receptors signal through their ITAMs and inhibitory receptors utilize **immunotyrosine *inhibitory*** motifs (ITIMs) to transmit intracellular signals.
e. The functions of NK cells are promoted by cytokines.
2. NK cells mediate innate immunity by producing cytokines (eg, IFN-γ) and killing infected target cells without the need for clonal expansion.
3. Important coactivating signals for NK cells include IL-12 and IFN-α/β.

CHEDIAK–HIGASHI SYNDROME

CLINICAL CORRELATION

- *Chediak–Higashi Syndrome (CHS) is a rare autosomal recessive condition characterized by recurrent infections and poor NK cell activity.*
- *The genetic defect resides in the gene for **lysosomal trafficking regulator (LYST)** and causes a fusion of cytoplasmic granules in NK cells, neutrophils, monocytes, and other granule-containing cells.*
- *CHS NK cells bind normally to their target cells, but killing is absent.*
- *The immune deficiencies of CHS patients are probably more related to defective neutrophil function than impaired NK cell activity.*

CLINICAL PROBLEMS

You have a patient with enlarged lymph nodes, splenamegaly, and anemia. During an immunological work-up you find that the patient has a slight lymphocytosis. You perform multicolor flow cytometry after staining the patient's blood lymphocytes with antibodies to CD3, CD4, and CD8. Shown below are $CD3^+$ blood lymphocytes stained for CD4 (x-axis) and CD8 (y-axis) expression.

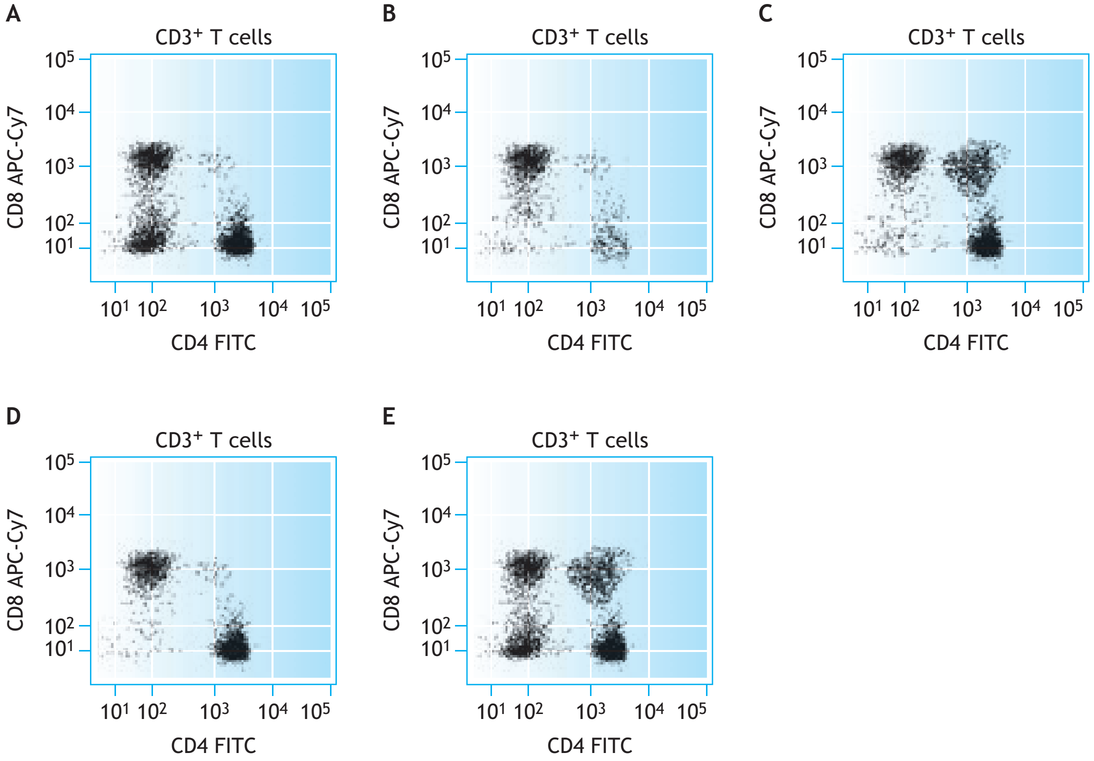

1. Which of the patterns shown above would suggest an apoptosis defect is responsible for this patient's disease?
   - A. A
   - B. B
   - C. C
   - D. D
   - E. E

A patient with BLS type 1 presents with a herpes virus infection. His blood lymphocytes are found to kill autologous virus-infected target cells rapidly in vitro.

2. Which of the following cell types is probably mediating this killing?
   - A. CD8$^+$ T cells
   - B. γδ T cells
   - C. Macrophages
   - D. NK cells
   - E. B cells

You are a member of a cardiac transplantation team in a medical school and meet with second-year medical students to explain the surgical procedure and posttransplantation therapy. One of the drugs you review is cyclosporine.

3. Which of the following would be a reasonable summary of its immunosuppressive effect?
   - A. The drug kills dividing T and B cells in metaphase.
   - B. The drug inhibits antigen processing by dendritic cells.
   - C. The drug inhibits IL-2 gene transcription in T cells.
   - D. The drug blocks NK cell recognition of foreign HLA antigens.
   - E. The drug inhibits HLA class I expression by cardiac muscle cells.

Mutations in either Rag1 or Rag2 cause severe immunodeficiencies in human beings.

4. Which of the following cell types would show normal numbers in a patient with a Rag1 deficiency?
   - A. CD3$^+$ cells with a CD4 coreceptor
   - B. CD19$^+$ cells
   - C. Single positive thymocytes
   - D. Lymphocytes with a coreceptor specific for C3d
   - E. Cells with an inhibitory receptor specific for MHC class I

Johnny is an 8-month-old child with recurrent viral and fungal infections. His blood lymphocytes can bind IL-2, and his macrophages can bind IFN-γ in vitro. It has been determined that his parents are both heterozygous for a mutation in the ZAP-70 gene known to disrupt the activity of this kinase.

5. Assuming Johnny has inherited this mutation from both parents, which of his cells should be *normal*?
   - A. Monocytes and macrophages
   - B. TCRαβ T lymphocytes
   - C. TCRγδ T lymphocytes
   - D. NK cells
   - E. NKT cells

ANSWERS

1. The correct answer is A. This pattern is abnormal in the sense that the blood contains large numbers of double negative T lymphocytes (CD3$^+$CD4$^-$CD8$^-$). Normally, double negative CD3$^+$ cells are found only in the thymus (panel E). This finding suggests altered thymic selection and peripheral homeostasis of T cell subsets secondary to defective apoptosis induced by the Fas–Fas ligand system (ie, autoimmune lymphoproliferative syndrome). The normal pattern for blood CD3$^+$ cells is shown in panel D.

2. The correct answer is D. NK cells recognize virus-infected and neoplastic cells that lack MHC class I, which is characteristic of cells from patients with BLS type 1.

3. The correct answer is C. Cyclosporine inhibits the Ca^{2+}/calmodulin-dependent phosphatase calcineurin, which leads to decreased TCR-initiated dephosphorylation of the latent transcription factor NFAT. With diminished NFAT activation, IL-2 and IL-4 gene transcription is decreased, which has a general immunosuppressive effect on activated T cells.

4. The correct answer is E. $T^{-}B^{-}NK^{+}$ SCID is caused by Rag recombinase deficiency. The patients have normal numbers of NK cells, which bear inhibitory receptors for MHC class I molecules.

5. The correct answer is A. All of the lymphocyte subsets listed use ZAP-70 as a signaling intermediate for receptor-induced cell activation.

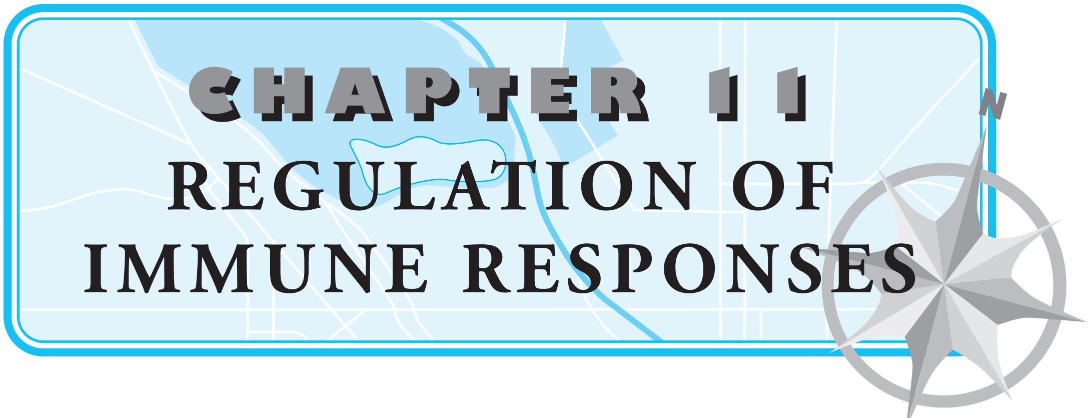

CHAPTER 11

REGULATION OF IMMUNE RESPONSES

I. The immune response is regulated at the level of **antigen recognition** by mechanisms designed to distinguish self from nonself.

- A. The innate immune system recognizes **pathogen-associated molecular patterns** that are not seen on host tissues.
- B. The adaptive immune system recognizes nonself **epitopes** using T and B cell receptor repertoires that have been selected for their specificity during lymphocyte differentiation.
- C. Both systems are imperfect, and immune responses to self components can arise even in healthy individuals (Chapter 16).
  1. Toll-like receptors recognize self components derived from damaged tissues (eg, heat shock proteins).
  2. Antibody production against self antigens occurs frequently without any adverse effect.

II. Antigen-presenting cells (APCs) control the induction of adaptive immune responses.

- A. The constellation of major histocompatibility complex (MHC) molecules expressed by an APC determines the peptides it can present.
- B. **Coreceptor signals** derived from APCs influence the type of immune response induced.
  1. Whereas mature dendritic cells (DC) activate $CD4^+$ Th cells, immature DC activate $CD4^+$ Treg cells.
  2. CD28 binding of B7 molecules induces Th cell activation, whereas cytotoxic T lymphocyte-associated protein 4 (CTLA-4) binding of B7 molecules inhibits Th cell activation.
- C. **Cytokines** influence APC functions and lymphocyte activation.
  1. Interferon (IFN)-γ and tumor necrosis factor (TNF)-α increase MHC class II expression and APC function.
  2. Macrophage- and DC-derived interleukin (IL)-12 promotes Th1 cell activation.
  3. IL-4 produced by T cells and mast cells promotes Th2 cell differentiation.

III. When an immune response does occur, its duration is limited, and the immune system returns to a **basal homeostatic state.**

- A. As antigen is cleared from the body, lymphocyte activation subsides.

B. Antigen-activated lymphocytes undergo **activation-induced cell death**.
  1. Activated T lymphocytes express both **Fas** and **Fas ligand**, which signals apoptotic death.
    a. The cytoplamic domains of Fas contain death domains that bind adapter proteins, recruit and activate caspases, and trigger apoptosis.
    b. Fas deficiency in humans causes **autoimmune lymphoproliferative syndrome (ALPS)** (Chapter 10), the failure of activated T cells to undergo apoptosis.
  2. Oxidant production by activated cells initiates cell death through the endogenous apoptosis pathway.

C. Effector lymphocytes, neutrophils, and macrophages are relatively **short lived**.

D. Memory B and T cells show only **low levels of proliferation**.

E. **Antibodies** regulate immune responses to their antigens.
  1. Maternal antibodies that cross the placenta can interfere with the immunization of the newborn in the first few months of life.
  2. Complement-activating antibodies enhance B cell activation through the **CR2 coreceptor** (Chapter 9).
  3. Immune complexes containing immunoglobulin G (IgG) can also inhibit B cell activation (Figure 11–1).
    a. IgG antibody–antigen complexes can simultaneously bind to the **Fcγ receptor IIB (FcγRIIB)** and BCR.
    b. BCR cross-linking to FcγRIIB blocks B cell receptor (BCR) signaling.
      (1) The cytoplasmic tail of FcγRIIB contains **immunotyrosine inhibitory motifs (ITIMs)**, which recruit **Src homology 2 domain-containing inositol polyphosphate 5-phosphatase (SHIP)**.
      (2) SHIP disrupts phospholipase Cγ (PLCγ) and Bruton's tyrosine kinase (Btk) activation.

RHOGAM THERAPY

CLINICAL CORRELATION

- ***Hemolytic disease of the newborn (erythroblastosis fetalis)*** *is caused when an Rh-negative mother produces IgG antibodies to the Rh antigens of her fetus.*
- *Fetal anemia and jaundice occur when these antibodies cross the placenta.*
- *To prevent this condition, Rh-negative mothers are given* ***Rhogam*** *(Rh-specific γ-globulin) during and immediately after pregnancy.*
- *Rhogam prevents immunization of the mother by clearing fetal erythrocytes from her blood circulation.*
- *Rhogam also cross-links the BCR and FcγRIIB on maternal B cells and inhibits maternal B cell activation.*

  4. IgG immune complex binding to macrophage **FcγRI** (CD64) also induces the release of **IL-10, transforming growth factor (TGF)-β**, and **prostaglandin E_2 (PGE_2),** all of which inhibit lymphocyte activation.
  5. Antibodies produced against the idiotypic determinants of a BCR can inhibit B cell activation.
    a. Autoantibodies against idiotypes are produced regularly during adaptive immune responses.
    b. Antiidiotype antibodies can block B cell activation by cross-linking the BCR to FcγRIIB.

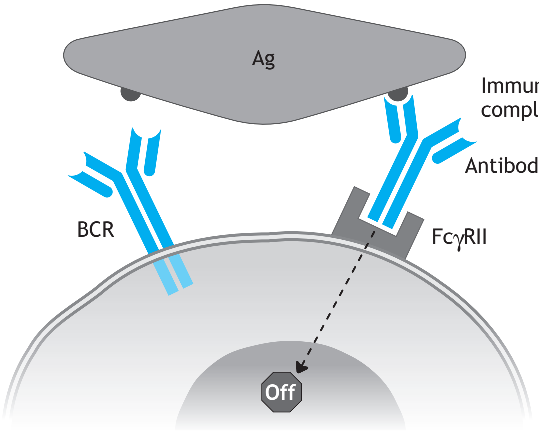

Figure 11–1. Mechanism by which immune complexes inhibit B cell activation. BCR, B cell receptor.

6. When high circulating antibody concentrations are achieved, the rate of Ig catabolism also increases.

INTRAVENOUS IMMUNOGLOBULIN (IVIG) THERAPY IN MULTIPLE MYELOMA

CLINICAL CORRELATION

- *Multiple myeloma is characterized by pancytopenia due to the depression of hematopoiesis by the malignant cells within the bone marrow.*
- *This causes a functional **hypogammaglobulinemia** (ie, decreased production of useful antibodies) and increases the risk of infections.*
- *Prophylactic treatment with IVIG has not always proven useful for correcting this defect.*
- *The extremely high circulating levels of the M component cause an elevated clearance rate for all Ig, which limits the lifespan of exogenous IVIG.*

IV. CD4⁺CD25⁺ regulatory T cells (Treg cells) mediate immune homeostasis.

A. Treg cells suppress immune responses by several mechanisms.
  1. They consume IL-2 and deprive other lymphocytes of its growth-promoting activity.
  2. Treg cells inhibit activated T cells by cell–cell contact and the secretion of the inhibitory cytokines **IL-10** and **TGF-β**.
  3. Treg cells are induced late in immune responses and act by suppressing T cell effector functions.

B. Treg cells prevent immune-mediated diseases.
  1. Treg cells limit host tissue damage that might otherwise occur during exuberant immune responses to microbes.
  2. Treg cells assist in the maintenance of self tolerance and prevent autoimmunity.

THE IMMUNOLOGICAL ENIGMA OF PREGNANCY

- *In human pregnancies, the fetus is generally considered a* ***semiallograft*** *(ie, half of its tissue antigens are genetically foreign to the mother) (Chapter 17).*
- *It is not clear why such a foreign tissue graft is not rejected by the mother, but several mechanisms appear to contribute to fetal survival.*
- *Trophoblastic cells that form the fetal interface with the maternal circulation are devoid of most HLA antigens and certain costimulatory molecules, such as B7.*
- *Inhibitory factors, including IL-10, B7 isoforms, death receptor ligands, and PGE_2, are expressed at high levels at sites of implantation.*
- *Some fetal cells do cross the placenta and are tolerated by the mother's immune system for years after pregnancy, suggesting a central immune anergy.*

V. Polymorphisms exist in genes that regulate immune functions.

- **A.** Dozens of genes that control immune cell differentiation and activation have been described, some of which will be discussed in Chapter 16.
  - **1.** The majority of heritable defects in adaptive immunity have been genetically mapped, and the gene products have been identified.
  - **2.** Much less is known about human genetic polymorphisms or mutations linked to defects in innate immunity.
- **B.** Functionally significant mutations and polymorphisms in genes affecting complement components, components of antigen presentation pathways, cytokines, cytokine receptors, and Ig switch recombinases have been described.

VI. Dysregulation of the immune system can cause immunological diseases.

- **A.** **Autoimmunity** occurs when self-tolerance is lost (Chapter 16).
  - **1.** T cells specific for autoantigens can be activated when coreceptor ligands (eg, B7) are abnormally expressed.
  - **2.** B cells that produce high affinity autoantibodies can be activated by costimulatory signals derived from infection.
  - **3.** Defects in apoptosis (eg, Fas mutations) can result in a failure to delete autoreactive T cells.
- **B.** Adaptive immune responses to microbial antigens can **cross-react** with self tissue antigens and cause immune-mediated tissue damage.

RHEUMATIC FEVER

- *Rheumatic fever in school age children is a complication secondary to pharyngitis caused by the bacterium* Streptococcus pyogenes.
- *Damage to the heart and joints appears 2–4 weeks after the pharyngitis.*
- *Antibodies produced against the streptococcal M protein* ***cross-react*** *with cardiac autoantigens.*
- *Autoantibodies to myosin, keratin, and laminin play a key role in cardiac valve damage.*

- **C.** **Atopic allergies** result when excessive immune responses are made to relatively innocuous environmental antigens (Chapter 13).
  - **1.** Immune dysregulation causes **Th2 polarization.**
  - **2.** IgE antibody production favors acute inflammatory responses.
- **D.** **Immune responses to microbes** can exceed what is necessary to clear the pathogen.

1. **Inflammatory bowel disease** results from immune responses to microbial flora that initiate chronic inflammation.
2. **Infectious mononucleosis** is caused by an excessive $CD8^+$ T cell response to Epstein–Barr virus antigens that damages host tissues.

SEPSIS REVISITED

CLINICAL CORRELATION

- ***Sepsis** is a systemic inflammatory response to infection and one of the leading causes of death among hospitalized patients.*
- *The disease also illustrates what happens when immune mechanisms designed to protect the host are excessively activated (Figure 11–2).*
- *A complex pathogenesis results from the activation of multiple host defense systems by pathogenic and opportunistic microbes.*
- *Bacterial components, including lipopolysaccharide (LPS), can activate the complement, coagulation, and fibrinolysis systems.*
- *Macrophage activation for the production of proinflammatory cytokines (eg, TNF-α) and arachidonic acid metabolites results from Toll-like receptor stimulation.*
- *Endothelial and smooth muscle cells are activated, resulting in coagulopathies, severe hypotension, impaired organ perfusion, and hypoxic tissue damage.*

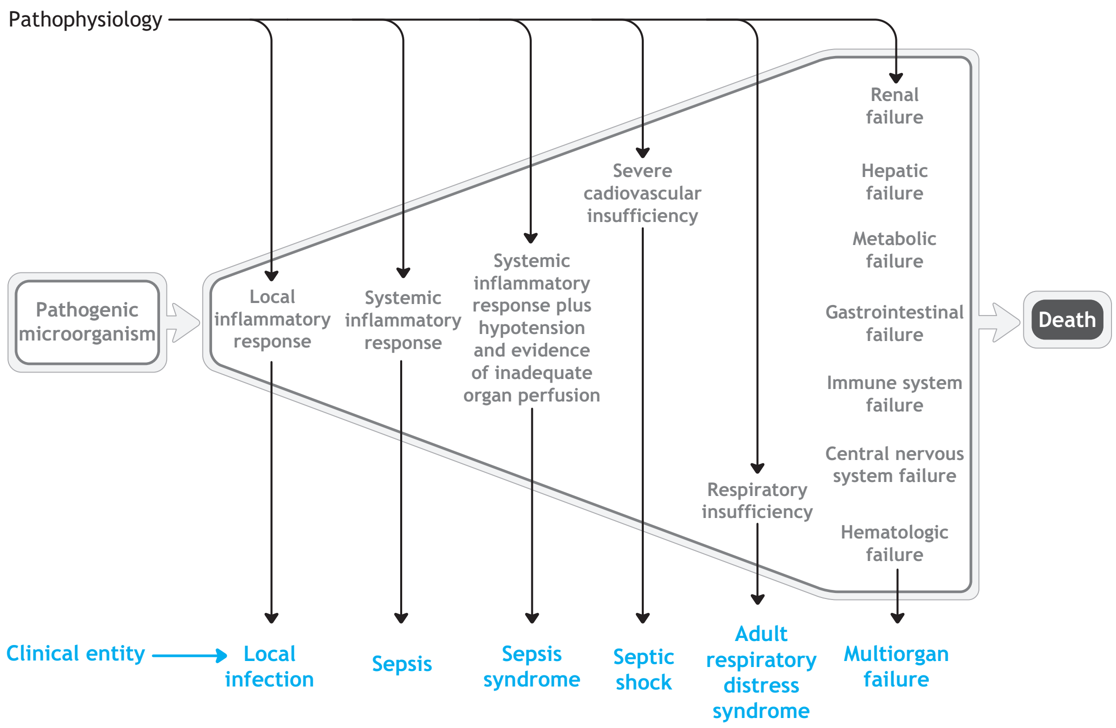
Pathophysiology
Pathogenic microorganism
Local inflammatory response
Systemic inflammatory response
Systemic inflammatory response plus hypotension and evidence of inadequate organ perfusion
Severe cardiovascular insufficiency
Respiratory insufficiency
Renal failure
Hepatic failure
Metabolic failure
Gastrointestinal failure
Immune system failure
Central nervous system failure
Hematologic failure
Death
Clinical entity
Local infection
Sepsis
Sepsis syndrome
Septic shock
Adult respiratory distress syndrome
Multiorgan failure

Figure 11–2. Pathogenesis and clinical definitions associated with the sepsis process.

CLINICAL PROBLEMS

Clinical vaccine trials using a nine-amino acid peptide immunogen from an HIV-1 glycoprotein have been performed. The vaccine is effective at inducing high-affinity IgG virus-neutralizing antibodies in only a subset of healthy adult vaccinees.

1. Which of the following immune properties probably characterizes the responsive individuals?
   - A. They share HLA class II genes.
   - B. They lack Treg cells.
   - C. They are high producers of IL-10.
   - D. They lack IL-2 receptors.
   - E. They are all over 70 years of age.

A patient with lupus presents with impaired renal function, and a biopsy shows IgG and C3b deposits in the renal glomerulus. The patient's serum contains antibodies to her DNA and erythrocytes. Her total IgG is significantly elevated (1900 mg/dL).

2. Which of the following best explains the hypergammaglobulinemia seen in this patient?
   - A. Reduced clearance of IgG by the damaged kidney
   - B. Destruction of Treg cells by anti-DNA antibodies
   - C. Polyclonal B cell activation by numerous self antigens
   - D. An opportunistic virus infection
   - E. Depletion of serum complement

Mary is a 2-year-old child with an unremarkable medical history, but presents today with enlarged cervical lymph nodes. Her complete blood count shows lymphocytosis (10,500/μL) and her serum contains elevated concentrations of IgG and IgM. Flow cytometry indicates that her $CD3^+$ lymphocytes are 25% $CD4^+$ and 20% $CD8^+$. She has no history of recurrent infections and no evidence of a current infection or cancer.

3. Which of the following is the most likely basis for Mary's immune abnormalities?
   - A. HIV-1 infection
   - B. IgA deficiency
   - C. Thymic hypoplasia (DiGeorge syndrome)
   - D. An apoptosis defect (Fas deficiency)
   - E. A monoclonal gammopathy

A 2-month-old healthy child is immunized with a new bacterial protein vaccine as part of a clinical trial. His preimmunization titer of IgG antibody to the vaccine protein is 1:16, and the titer is unchanged 2 weeks postimmunization.

4. Which of the following is a likely explanation for this observation?

A. The child has had an infection with the bacterium during gestation.
B. Maternal antibody to the bacterial protein blocked active immunization of the child.
C. The child has severe combined immune deficiency.
D. The child has a Th cell defect that prevents Ig class switching.
E. Children cannot make antibodies at this age.

ANSWERS

1. The correct answer is A. The high responding vacinees probably share HLA antigens that can present the peptide vaccine. Because this is an exogenous antigen, class II molecules would be involved. Presumably, the HLA class II molecules of the nonresponders failed to bind and present the peptide.

2. The correct answer is C. Elevated IgG levels in this patient are probably due to polyclonal B cell activation by autoantigens. Most patients of this type produce autoantibodies to a wide range of self antigens, suggesting that they have defects in general immune regulation mechanisms.

3. The correct answer is D. Mary has an unusual pattern of T cell markers. Only 45% of her T cells express either CD4 or CD8; the remaining T cells are apparently double negative. This is a classic sign of ALPS (Chapter 10). The condition arises from a defect in apoptosis among T cells resulting from a Fas mutation. Her elevated immunoglobulins probably represent autoantibodies against a range of self antigens, because self-reactive T cells are not controlled by Fas-mediated apoptosis.

4. The correct answer is B. The IgG antibodies present prior to immunization were probably maternally derived, because children of this age do not synthesize much IgG. The fact that the titer did not increase following immunization seems to indicate that the maternal antibodies interfered with B cell activation. This would occur if the vaccine and preformed maternal antibodies formed immunosuppressive immune complexes and cross-linked the BCR and FcγRIIB.

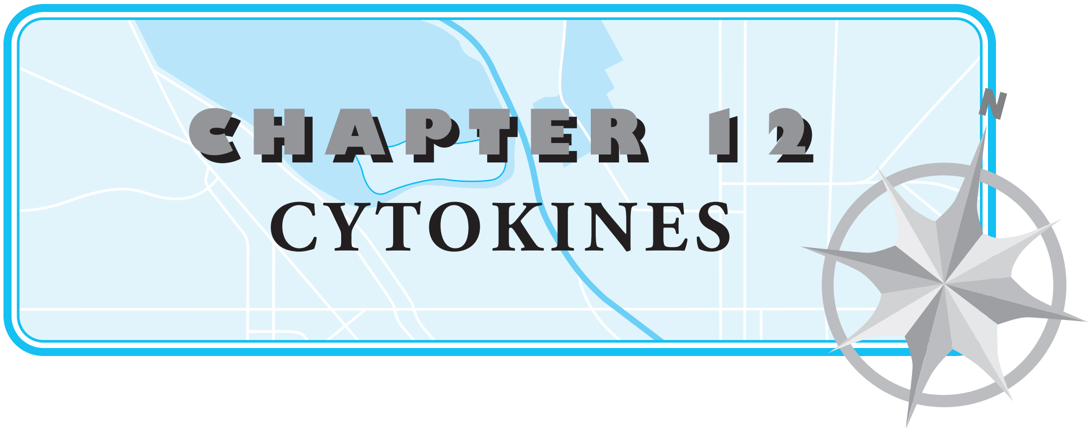

CHAPTER 12

CYTOKINES

I. Cytokines mediate cell–cell communication in the immune system.

- **A.** Cytokines are 8- to 25-kDa polypeptides that include **interleukins (IL), chemokines, growth factors (GF), interferons (IFN)**, and **colony-stimulating factors (CSF)** (Appendix II).
- **B.** Most cytokines are designed for rapid, short-term effects.
  - **1.** They are often secreted immediately after their synthesis.
  - **2.** Their short half-lives in circulation reflect the action of inactivators and binding proteins.
- **C.** Cytokines typically act over short distances.
  - **1.** Many cytokines (eg, IL-2) have **autocrine** effects; they act on the cells that produce them.
  - **2.** Other cytokines have localized **paracrine** effects; one cell produces a cytokine that acts on a nearby cell.
  - **3.** Cell–cell contact promotes the action of cytokines.
  - **4.** A few cytokines show **endocrine** effects; they act at a distance after secretion into the blood stream.
    - **a.** Tumor necrosis factor (TNF)-α produced by tissue macrophages during bacterial infections can have systemic inflammatory effects.
    - **b.** IL-1β produced during inflammation or infections causes fever by its endocrine effects on the hypothalamus.
- **D.** Cytokines activate cells through specific high-affinity receptors.
  - **1.** Dissociation constants for cytokine receptors are in the picomolar range.
  - **2.** Receptors for cytokines that mediate acute inflammatory or innate immune responses are typically constitutively expressed.
  - **3.** Receptors for lymphocyte-activating cytokines are often induced.
  - **4.** The same receptor for a given cytokine may have significantly different effects when expressed on different cell types.
- **E.** The most common cellular response to a cytokine is the rapid initiation of gene transcription.
- **F.** Cytokines have **redundant** biological effects (Table 12–1).
  - **1.** This ensures that a deficiency of one cytokine does not eliminate an essential biological function.
  - **2.** Cytokine redundancy frustrates anticytokine therapies.
- **G.** Most cytokines are **pleiotropic**; they exhibit several functions.

Table 12–1. Redundancies between TNF-α and IL-1β.[a]

Response	TNF-α	IL-1β
Fever induction	+	+
Acute phase protein induction through IL-6	+	+
Induction of vascular adhesion molecule expression	+	+
Induction of IL-8	+	+
Activation of NFκB	+	+
Joint inflammation in rheumatoid arthritis	+	+
Systemic inflammatory response syndrome (SIRS)	+	+
Cachexia	+	–
Induction of apoptosis	+	–

[a] TNF, tumor necrosis factor; IL, interleukin.

H. When combined, cytokines can have **synergistic** effects.
  1. One cytokine may induce the expression of receptors for the second cytokine.
  2. Distinct cytokine receptor signaling pathways can converge at the same cellular response (eg, NFκB activation).

I. Cytokines can **amplify** a biological response by inducing additional cytokine synthesis (Figure 12–1).

J. Certain cytokines **antagonize** the effects of other cytokines.
  1. IFN-γ inhibits the effects of IL-4 on B cells.
  2. IL-10 and transforming growth factor (TGF)-β antagonize the effects of IFN-γ on macrophages.

K. There are dedicated **inhibitors** of cytokines.
  1. **IL-1 receptor antagonist (IL-1ra)** is produced by the same cells that secrete IL-1 and binds to the same receptor.
  2. The amino-terminus of TGF-β maintains the secreted cytokine in a latent form until it is removed by proteolysis.

TUBERCULOSIS IN RHEUMATOID ARTHRITIS

CLINICAL CORRELATION

- ***Rheumatoid arthritis (RA)*** *is an autoimmune disease characterized by inflammation of the joints and a progressive loss of joint function.*
- *New therapies for RA include the use of anticytokine reagents to block the action of TNF-α or IL-1β.*
- *For example, Etanercept is a recombinant protein consisting of a portion of a TNF receptor (p75) fused to the Fc portion of human immunoglobulin (Ig) G to prolong its circulating half-life.*

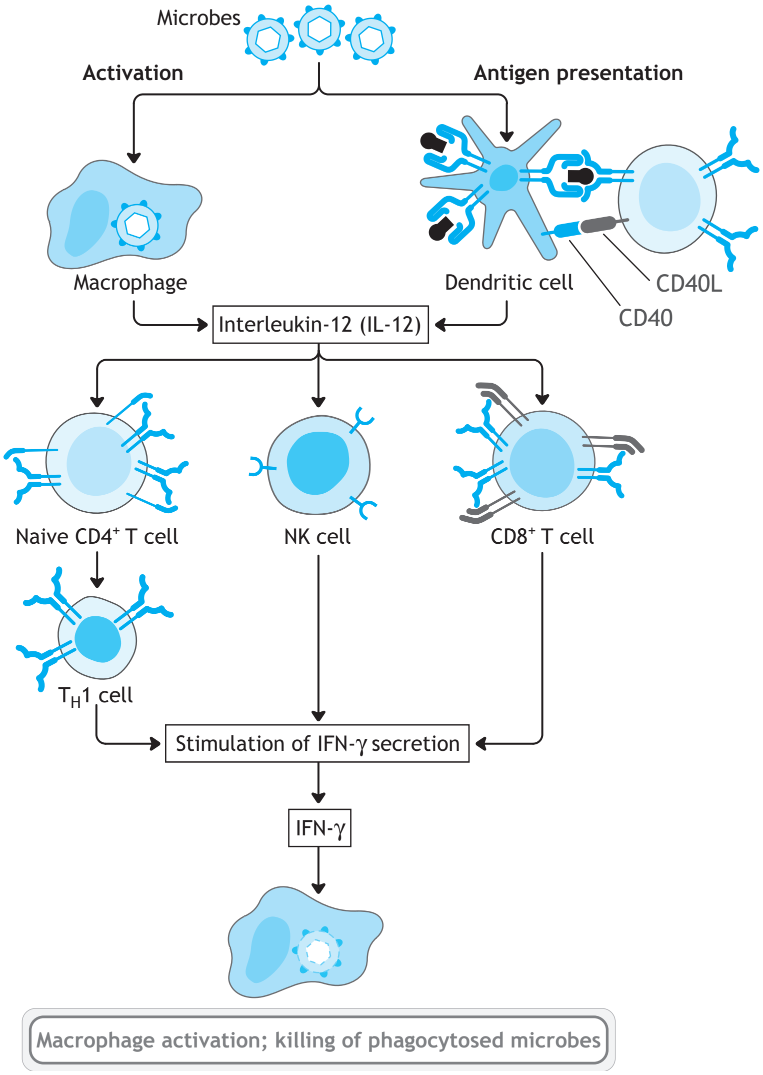
Microbes
Activation
Antigen presentation
Macrophage
Dendritic cell
CD40L
CD40
Interleukin-12 (IL-12)
Naive CD4⁺ T cell
NK cell
CD8⁺ T cell
T_H1 cell
Stimulation of IFN-γ secretion
IFN-γ
Macrophage activation; killing of phagocytosed microbes

Figure 12–1. A cytokine cascade. NK, natural killer; IFN, interferon.

- *Combining an anti-TNF-α and an anti-IL-1β drug in these patients increases the risk of pulmonary tuberculosis.*
- *This finding emphasizes the importance of using anticytokines under conditions that inhibit inflammation without impairing innate host defenses.*

II. Macrophages, natural killer (NK) cells, and NKT cells are important sources of **proinflammatory cytokines** that also mediate **innate immunity** (Table 12–2).

Table 12–2. Cytokines that mediate innate immunity and acute inflammation.[a]

Biological Function	Important Cytokines	Specific Examples
Leukocyte adhesion	TNF-α	Induces ICAM-1 expression on endothelium
Chemotaxis	TGF-β IL-8	Recruits blood monocytes Recruits blood neutrophils
Induction of other inflammatory mediators	IFN-γ	Promotes macrophage activation for IL-1 and TNF-α production
Fever induction	IL-1β	Activates cells of the hypothalamus
Acute phase protein synthesis	IL-6, IL-1, TNF-α	Activates hepatocytes and Kupffer cells to produce fibrinogen and complement components
Oxidant production	TNF-α IFN-γ	Activates neutrophils for superoxide production Activates macrophages for NO production
Inhibition of immunity	IL-10, TGF-β	Inhibits TNF-α synthesis by IFN-γ-activated macrophages
Apoptosis	TNF-α LT, IL-2	Apoptosis of tumor cells Activation induced cell death in T cells

[a] TNF, tumor necrosis factor; ICAM, intercellular adhesion molecule; TGF, transforming growth factor; IL, interleukin; IFN, interferon; LT, leukotriene.

A. These cells bear receptors that recognize both conserved microbial structures and endogenous ligands derived from damaged host cells.
  1. Macrophages express Toll-like receptors (TLRs) (Table 1–2).
  2. NK cells and NKT cells recognize exogenous and endogenous ligands from infected, stressed, or injured host cells (Table 10–4).

B. The cytokines produced by these cells signal **danger** to phagocytes and endothelial cells.
  1. TNF-α has the following functions.
    a. It stimulates the expression of **adhesion molecules** and **integrins** on endothelial cells and leukocytes, respectively.
    b. It primes **neutrophils** and **macrophages** for respiratory bursts and NO production.
    c. It induces the synthesis of **acute phase proteins**.
    d. It induces **apoptosis** through its death domain-containing receptor.

2. **IL-1β** has the following functions.
   - **a.** It is redundant with TNF-α (Table 12–1).
   - **b.** It synergizes the effects of TNF-α.
   - **c.** It induces **fever** by stimulating the hypothalmus.
3. **IL-12** has the following functions.
   - **a.** It coactivates NK cells and T cells.
   - **b.** It coinduces **IFN-γ production** and cytotoxicity by NK cells and NKT cells.
   - **c.** It induces the differentiation of $CD4^+$ Th1 cells.
4. **IFN-γ** (type II interferon) has the following functions.
   - **a.** It **activates macrophages**.
   - **b.** It induces the production of other **proinflammatory cytokines**, including TNF-α, IL-1β, IL-6, and the chemokine CCL4.
   - **c.** It promotes **major histocompatibility complex (MHC) class II** gene expression.
   - **d.** It promotes the differentiation of $CD4^+$ Th1 cells.
5. **Type I interferons (IFN-α/β)** have the following functions.
   - **a.** They **block virus replication** by autocrine signaling.
   - **b.** They increase **MHC class I expression**.
   - **c.** At low levels they **coactivate macrophages**.
   - **d.** At high concentrations they inhibit macrophage and lymphocyte activation.
6. **Chemokines** have the following characteristics.
   - **a.** Chemokines are chemotactic cytokines that are produced rapidly in response to the danger signals associated with infection and inflammation.
   - **b.** They **attract cells** to sites of infection or inflammation.

C. The proinflammatory effects of these cytokines are balanced by **antiinflammatory cytokines (eg, IL-10), cytokine-binding proteins**, and **receptor antagonists** (see above).

D. The amount of a cytokine that is secreted determines its overall effects.

1. In **low to moderate levels** of production, proinflammatory cytokines initiate important innate mechanisms of host defense.
2. When produced at **high concentrations**, proinflammatory cytokines cause the **systemic inflammatory response syndrome (SIRS)**.
   - **a. Sepsis** is a form of SIRS initiated by infection (Chapter 11).
   - **b.** SIRS can be induced by any stimulus that activates large numbers of inflammatory cells and mediators.

SYSTEMIC INFLAMMATORY RESPONSE SYNDROME (SIRS)

CLINICAL CORRELATION

- *SIRS is defined as a constellation of clinical signs and symptoms, including **hyperthermia** (> 38°C) or **hypothermia** (< 36°C), **tachycardia** (> 90 bpm), **tachypnea** (> 20 breaths/min), and/or **altered white blood cell (WBC) counts** (> 12,000/µL or < 4,000/µL).*
- ***Septicemia** (microbes in the blood stream) and microbial **toxemia** (toxins in the blood stream) are among the leading causes, in which case the syndrome is called **sepsis** or **septic shock**.*
- *Regardless of the initiating event, proinflammatory cytokines play a central role.*
- *Multiorgan dysfunction, including acute respiratory distress syndrome, is a common sequela.*
- *Mortality rates when multiorgan failure is present approach 90%.*

E. Cytokines of innate immunity are secreted in a temporal sequence.

- **1.** TNF-α is produced early and promotes leukocyte localization at sites of infection.
- **2.** Chemokines are produced early as a means of recruiting leukocytes to affected tissue sites.
- **3.** Cytokines that activate phagocytes (eg, IFN-γ) are produced somewhat later.

III. Cytokines of adaptive immunity provide coactivating signals for lymphocytes and regulate antigen-presenting cells.

A. Cytokines that induce **humoral immunity** are concerned with B cell recruitment, activation, growth, and differentiation.

- **1. IL-2** from CD4$^+$ Th1 cells induces B cell proliferation and increases Ig and J chain synthesis.
- **2. IL-4** from **CD4$^+$ Th2 cells** promotes **Ig class switching** to IgE, induces the differentiation of naive CD4$^+$ T cells into **Th2 cells**, and **antagonizes** IFN-γ-driven Ig class switching to IgG$_3$.
- **3. IL-5** and **TGF-β** promote the synthesis of IgA by B cells.
- **4.** The chemokine **CCL7** directs antigen-stimulated B cells to T-dependent areas of the lymph nodes.
- **5.** The coactivating effects of IL-4 on B cells are balanced by the inhibitory effects of IFN-γ.

B. Cytokines that promote **cellular immunity** induce the activation, growth, and/or differentiation of T cells, NK cells, NKT cells, and macrophages.

- **1. IL-2** has the following characteristics.
  - **a.** It is an autocrine **growth factor** for many CD4$^+$ and CD8$^+$ T cells.
  - **b.** It enhances the cytotoxic activity of NK cells and conventional CD8$^+$ T cells.
  - **c.** It costimulates T cells for the production of IL-4, IL-5, and IFN-γ.
  - **d.** It promotes the development of **Treg cells**.
  - **e.** It induces autocrine **activation-induced cell death** in T cells (Chapter 11).
- **2. IFN-γ** has the following characteristics.
  - **a.** It is primarily produced by CD4$^+$ Th1 cells and CD8$^+$ T cells during adaptive immune responses.
  - **b.** It activates macrophage **intracellular killing** of microbes.
  - **c.** It increases both class I and class II **antigen presentation** pathways by increasing MHC, transporter associated with antigen processing (TAP) protein, the proteasome subunits, and HLA-DM expression (Chapter 7).
  - **d.** It promotes Th1 polarization by naive CD4$^+$ T cells.
  - **e.** It inhibits Th2-associated humoral immunity.

DEFICIENCY IN IFN-γ SIGNALING

CLINICAL CORRELATION

- *Genetic deficiencies have been described for either chain of the heterodimeric IFN-γ receptor.*
- *Patients with these defects are susceptible to infections by viruses and* Mycobacterium *and* Salmonella *species, two intracellular bacterial pathogens.*
- *The same phenotype is associated with defects in the expression of either IL-12 or the IL-12 receptor, illustrating the importance of IL-12 as an inducer of IFN-γ.*

- **3. Lymphotoxin (LT** or **TNF-β)** has the following characteristics.
  - **a.** It is produced by activated T cells.

   b. It has many of the same proinflammatory and apoptosis-inducing effects as TNF-α.

4. **TGF-β** is chemotactic for blood monocytes.

IV. Cytokines, referred to as **colony-stimulating factors (CSFs)** or **poietins,** coordinate the diverse process of **hematopoiesis** (Figure 12–2).

A. **Stem cell factor (SCF)** and **IL-3** have broad growth-promoting effects on multiple blood cell lineages.

B. **Lymphopoiesis** is regulated by **IL-7** (T and B cells), **IL-2** (T cells), and **IL-15** (NK cells).

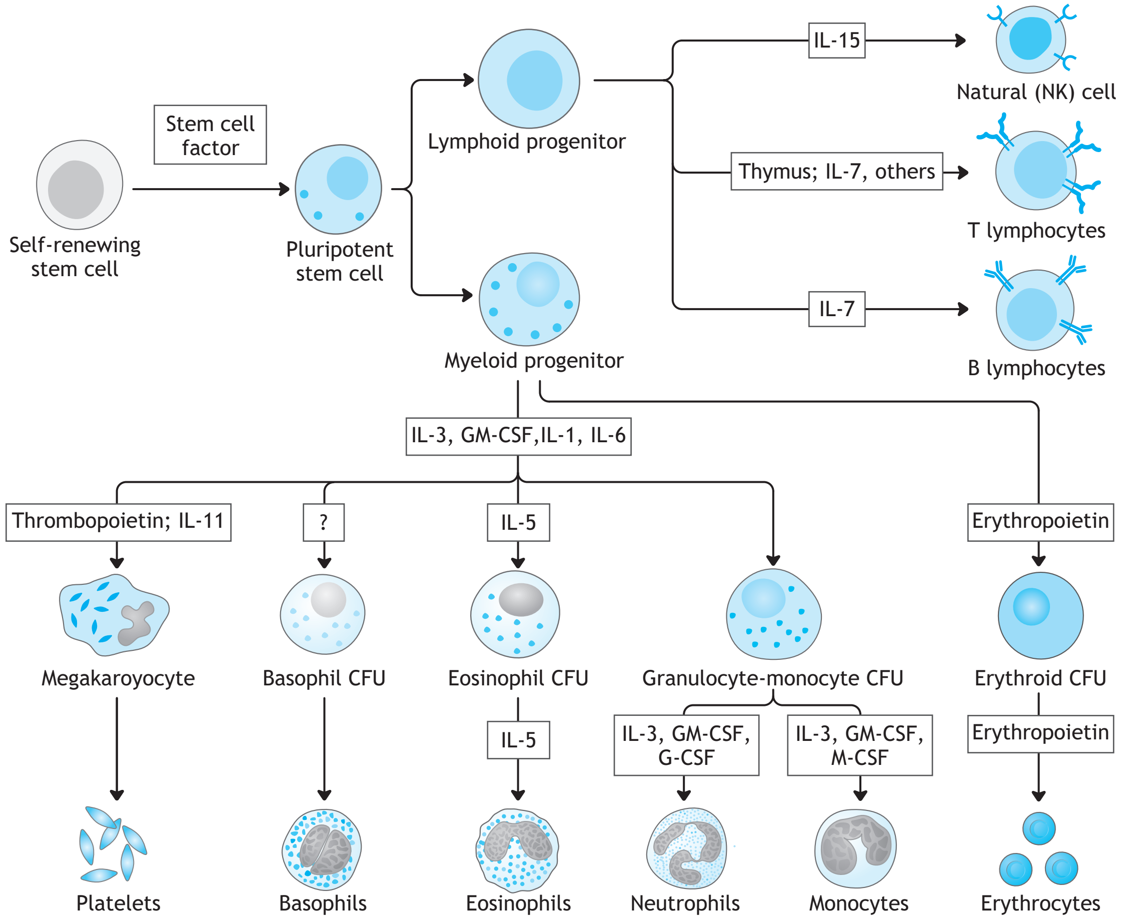
Stem cell factor
Self-renewing stem cell
Pluripotent stem cell
Lymphoid progenitor
IL-15
Natural (NK) cell
Thymus; IL-7, others
T lymphocytes
IL-7
B lymphocytes
Myeloid progenitor
IL-3, GM-CSF,IL-1, IL-6
Thrombopoietin; IL-11
?
IL-5
Erythropoietin
Megakaroyocyte
Basophil CFU
Eosinophil CFU
Granulocyte-monocyte CFU
Erythroid CFU
IL-5
IL-3, GM-CSF, G-CSF
IL-3, GM-CSF, M-CSF
Erythropoietin
Platelets
Basophils
Eosinophils
Neutrophils
Monocytes
Erythrocytes

Figure 12–2. Cytokines that regulate hematopoiesis. IL, interleukin; GM-CSF, granulocyte/macrophage colony-stimulating factor; CFU, colony-forming unit.

C. Myelopoiesis is controlled by **erythropoietin (EPO)**(erythrocytes), **thrombopoietin** (platelets), **IL-5** (eosinophils), **granulocyte/macrophage (GM)-CSF** and **G-CSF** (neutrophils), and GM-CSF and **M-CSF** (monocytes).

D. Recombinant CSFs are now used routinely to correct certain defects in hematopoiesis.

ANEMIA AND ERYTHROPOIETIN

CLINICAL CORRELATION

- ***Anemia*** *is a reduction in red cell mass and has a number of causes, including depressed erythropoiesis.*
- *Chronic diseases, such as renal failure, malignancy, autoimmune diseases, and certain infections, are often accompanied by decreased EPO production.*
- *For the treatment of anemia associated with end-stage renal disease, EPO has replaced blood transfusions as a therapy.*
- *Repeated transfusions of whole blood to prospective renal transplant recipients often immunizes them against the HLA antigens of the blood donors.*
- *This can result in antibody-mediated hyperacute rejection of a subsequent transplant (Chapter 17).*

V. Chemokines, a diverse family of chemotactic cytokines, regulate normal cell traffic, tissue architecture, and inflammatory cell recruitment.

A. Chemokines establish a chemical concentration gradient that is sensed by the cell.

B. They induce the expression of leukocyte integrins that mediate leukocyte binding to and migration between vascular endothelial cells.

C. Chemokines induce cytoskeletal changes that mediate cell migration.

D. The two largest chemokine families are designated **CCL#** or **CXCL#** (distinguished by the arrangement of their cysteine residues)(Table 12–3).

1. CCL chemokines attract monocytes, lymphocytes, and eosinophils.
2. Most CXCL chemokines attract neutrophils, although some act on lymphocytes.

Table 12–3. Examples of CC and CXC chemokines and their receptors.[a]

Chemokine (Ligand)	Also Known as	Receptors	Cells Attracted
CCL4	MIP-1β	CCR5	T cells, NK cells, DC, and monocytes
CCL11	Eotaxin	CCR3	Eosinophils, basophils, Th2 cells
CXCL2	ENA-78	CXCR2	Neutrophils
CXCL8	IL-8	CXCR1, CXCR2	Neutrophils
CXCL10	IP-10	CXCR3	T cells

[a] MIP, macrophage inflammatory protein; NK, natural killer; DC, dendritic cells; ENA-78, epithelial neutrophil-activating protein 78; IL, interleukin; IP-10, interferon-γ-inducible protein 10.

E. Chemokine production is induced by a variety of immune and inflammatory stimuli, including TNF-α, IL-1β, IFN-γ, and IL-4.

F. Cellular responses to chemokines are mediated by seven-transmembrane, G protein-coupled cell surface receptors designated CCR# or CXCR#.

G. Some chemokine receptors serve as coreceptors for virus entry.
  1. HIV-1 uses both CCR5 and CXCR4 as coreceptors with CD4.
  2. The tissue distribution of these chemokine receptors specifies the tissue tropism of a particular HIV-1 clone.

VI. Cytokines activate cells by binding to high-affinity receptors (Table 12–4)

A. Cytokine receptors induce gene expression by activating latent cytosolic transcription factors.

B. The **type I (hematopoietic) receptors** activate the **Janus kinase (Jak)** signaling pathway.
  1. Jak is a family of receptor-associated tyrosine kinases that is activated upon receptor clustering.
  2. Jak phosphorylation of cytokine receptors recruits **signal transducers and activators of transcription (STAT)** peptides.
  3. Jak phosphorylates STAT peptides and the peptides dimerize.
  4. Dimerization of STAT results in nuclear localization, DNA binding, and transcriptional activation.

Table 12–4. Cytokine receptor families.[a]

Receptor Family	Ligands	Signaling Mechanism
Type I (hematopoietic)	IL-2, 3, 4, 5, 6, 7, 12, 13, 15, G-CSF, GM-CSF	Jak activation of latent, preformed STAT transcription factors
Type II (interferon-like)	IFN-α/β, IFN-γ, IL-10	Jak activation of latent, preformed STAT transcription factors
TNF	TNF, LT, CD40L, FasL	Binding of adapter proteins and activation of NFκB and AP-1 Binding of adapter proteins and activation of caspases
Ig superfamily	IL-1, M-CSF, SCF	Binding of adapter proteins and activation of latent NFκB
Seven-transmembrane G protein coupled	Chemokines	GTP exchange Activation of PLC and PI3K

[a] IL, interleukin; Jak, Janus kinase; STAT, signal transducers and activators of transcription; GM-CSF, granulocyte/macrophage colony-stimulating factor; IFN, interferon; TNF, tumor necrosis factor; LT, leukotriene; SCF, stem cell factor; PLC, phospholipase C; PI3K, phosphatidylinositol-3′-kinase.

5. The receptors for **IL-2, 3, 4, 6, 7, 12, and 13, G-CSF, and GM-CSF** utilize the Jak/STAT pathway for signaling.

C. **Type II interferon-like receptors** for IFN-α/β, IFN-γ, and IL-10 also signal through Jak/STAT.

D. The **TNF receptor** family utilizes adapter proteins to activate diverse signaling pathways.

1. When adapter proteins with **death domains** are recruited, caspase activation signals apoptosis.
2. When adapters without death domains are recruited, **kinase activation** results in the activation of NFκB.

E. The **Ig superfamily** of receptors, such as the IL-1 receptor, activates NFκB using specialized adapter proteins (eg, MyD88) (Figure 1–3).

F. G protein-coupled **chemokine receptors** signal through phospholipase C (PLC), phosphatidylinositol-3′-kinase (PI3K), and guanosine triphosphate (GTP) exchange proteins.

G. Different forms of the IL-2 receptor show different affinities for IL-2.

1. Resting T cells, B cells, and NK cells express the β and γc chains of IL-2R, which shows a moderate affinity for IL-2 (10^{-9} M).
2. Activated T cells express an additional α chain in their IL-2R, which increases receptor affinity to 10^{-11} M.

COMMON γ CHAIN (γC) DEFICIENCY

CLINICAL CORRELATION

- *Whereas cytokine deficiencies are extremely rare, an X-linked recessive cytokine receptor defect is the most common cause of* ***severe combined immune deficiency****.*
- *Mutations in the common γ chain (γc) impair the development of T cells and NK cells, because the receptors for IL-2, IL-4, IL-7, and IL-15 share the γc subunit.*
- *The resulting $T^-B^+NK^-$ phenotype places these infants at risk for infections with intracellular pathogens.*

CLINICAL PROBLEMS

A 5-month-old male patient presents with enlarged lymph nodes, and a biopsy shows acid-fast bacteria, a feature of infections with mycobacteria. The patient shows a normal complete blood count (CBC) and differential, normal T and B cell subsets by flow cytometry, and normal serum Ig levels when referenced to age-matched controls.

1. For which of the following cytokines would a mutation in its receptor explain this clinical picture?

   A. IL-1
   B. IL-2R
   C. TNF-α
   D. IFN-γ
   E. IL-10

2. Which of the following additional cytokine deficiencies would produce a similar phenotype?
   A. IL-12 deficiency
   B. IL-7 deficiency
   C. IL-15 deficiency
   D. SCF deficiency
   E. IL-4 deficiency

A patient who received a cardiac transplant from an unrelated donor is given the drug cyclosporine to control rejection of his allograft.

3. Which of the following receptors signals through an intermediate that is sensitive to the drug cyclosporine?
   A. IL-1 receptor
   B. IL-2 receptor
   C. T cell receptor (TCR)
   D. TLR4
   E. CXCR4

The blood lymphocytes of a 6-month-old child with recurrent, complicated viral and fungal infections do not respond normally to IL-2 and IL-4 in vitro. The patient currently has thrush and a cytomegalovirus (CMV) infection. The patient has a single gene mutation that explains all of these findings.

4. Which of the following immune functions might you expect to be **normal** in this child?
   A. Neutrophil ingestion of bacteria
   B. NK cell killing of virus-infected cells
   C. T cell activation by antigen + antigen-presenting cells (APCs)
   D. Ig class switching by B cells cultured with the patient's T cells
   E. Killing of CMV-infected cells by $CD8^+$ cells

A menstruating teenager presents to the emergency room in shock and is diagnosed as suffering from toxic shock syndrome secondary to a vaginal infection with the bacterium *Staphylococcus aureus*.

5. Which of the following cytokines is both key to host defense against this organism and a central mediator of toxic shock pathogenesis?
   A. IL-10
   B. TNF-α
   C. IL-3
   D. TGF-β
   E. IL-5

ANSWERS

1. The correct answer is D. Mutations in the IFN-γR result in susceptibility to intracellular pathogens (bacterial and fungal) early in life. This probably results from the failure to activate infected macrophages, a cell type commonly infected by these pathogens.
2. The correct answer is A. Because of its role as a coinducer of IFN-γ, IL-12 is also essential for the induction of cellular immunity to intracellular bacterial pathogens, such as the mycobacteria. Defects in the IL-12R produce a similar clinical picture.
3. The correct answer is C. None of the cytokine receptors signals through a cyclosporine-sensitive pathway, although the production of IL-2 and IL-4 by activated T cells is inhibited by the drug. Cyclosporine inhibits calcineurin and the activation of nuclear factor of activated T cells (NFAT), which induces transcription of genes in T and B cells.
4. The correct answer is A. The only known gene defect that could interfere with responses to both IL-2 and IL-4 is a mutation in the γc chain of the IL-2R and IL-4R. This defect also blocks IL-15R signaling, which would result in an absence of NK cells. The overall phenotype in the periphery would be $T^-B^+NK^-$.
5. The correct answer is B. TNF-α is a key mediator of both toxic (Chapter 6) and septic shock (Chapters 1 and 12). Toxic shock is initiated when superantigens produced by certain microbes, including the gram-positive *Staphylococcus*, activate Th cells and macrophages by cross-linking TCRs to MHC class II molecules. Massive cytokine production results when both cell types are activated.

6193ch13.qxd_mg 2/6/06 01:11 PM Page 149

CHAPTER 13

IMMUNE TISSUE INJURY

I. Robust immune responses, even to relatively nonthreatening antigens, can cause host tissue damage.

A. An **allergy** (or **hypersensitivity**) is an adaptive immune response to an environmental antigen that damages tissues and causes organ dysfunction.
  1. Allergy is mediated by specific antibodies and/or effector T cells.
  2. The signs and symptoms of allergy include watery eyes, runny nose, sneezing, coughing, wheezing, and pruritic (itchy) hives.
  3. An **allergen** is an antigen that can produce these signs and symptoms (Table 13–1).
  4. Typical allergens include pollens, animal dander, insect venoms, and foods.

B. **Atopic allergies** are familial, rapid, and strong reactions mediated by immunoglobulin (Ig) E.
  1. **Atopy** is a heritable predisposition to produce IgE antibodies and other allergic mediators in response to environmental antigens.
  2. Atopic conditions are more common in developed countries with high levels of air pollutants and closed living spaces.
  3. Common manifestations of atopic allergies are **allergic rhinitis, allergic asthma, atopic dermatitis,** and **allergic anaphylaxis**.

C. **Anaphylaxis** is a systemic form of allergy involving the cardiovascular, respiratory, cutaneous, and gastrointestinal systems.
  1. **Anaphylactic shock** includes hypotension, tissue hypoxia, and the loss of consciousness.
  2. **Anaphylactoid shock** is similar in its presentation, but is not mediated by IgE.
    a. IgG antibodies mediate systemic anaphylactoid reactions.
    b. Complement-mediated mast cell activation produces anaphylaxis-like symptoms.

D. Allergies can develop to inhaled, ingested, or injected antigens or antigens encountered by contact with the skin (Table 13–1).
  1. Tree and grass pollens induce seasonal **rhinitis** and **conjunctivitis**.
  2. Peanuts, shellfish, eggs, and dairy products often cause **food allergies**.
  3. Insect and snake venoms and injected drugs (eg, penicillin) can cause **allergic anaphylaxis**.
  4. Plant resins (eg, poison ivy pentadecacatechol), industrial chemicals, and metals can cause **allergic contact dermatitis**.

Table 13–1. Common allergies and allergens.

Clinical Manifestations	Typical Sources of Allergens
Atopic allergic rhinitis and allergic asthma	Tree and grass pollens, mold spores, mites in house dust, pet dander
Allergic gastroenteritis	Food products (milk, eggs, shellfish)
Occupational allergies	Pharmaceuticals, organic and inorganic compounds, enzymes, microbial agents, dyes, plant products, shellfish, animal products
Urticaria and anaphylaxis	Foods (shellfish, nuts, seeds), drugs (polypeptide hormones, enzymes, haptenic drugs, opiates), insect venoms
Contact dermatitis	Nickel and other metals, formaldehyde, epoxy resins, neomycin cream, benzocaine, plant resins
Drug allergies	β-Lactam antibiotics, NSAIDs[a]

[a] NSAIDs, nonsteroidal antiinflammatory drugs.

E. Gell and Coombs first described the four major mechanisms of immune tissue injury that underlie human hypersensitivities (Table 13–2).

  1. The mechanisms differ in terms of the nature of the antigen, the immune components involved in their pathogenesis, and the nature and onset of clinical symptoms.

  2. Antigenic specificity is mediated by antibodies and/or T cells.

  3. Tissue damage results from the activation of complement, mast cells, neutrophils, macrophages, or lymphocytes.

  4. Immune tissue injury mediated by the **innate immune system** is not included in the Gell and Coombs classification scheme.

F. Immune tissue injury mediated by adaptive immunity requires two phases.

  1. **Sensitization** occurs on first exposure to the allergen (Figure 13–1).

    a. Antibody production and/or the proliferation of T cells occur during the sensitization phase.

    b. Antibody can either bind to cellular Fc receptors or remain in the circulation, depending on Ig class.

    c. Tissue damage is not generally seen during the sensitization phase.

  2. During the **elicitation phase**, antibodies or immune T cells recognize the allergen and trigger effector mechanisms.

    a. For example, large numbers of memory $CD4^+$ Th1 cells are activated to produce cytokines.

    b. Allergen–antibody complexes activate complement to produce inflammatory peptides.

Table 13–2. The Gell and Coombs classification of immune tissue injury mediated by adaptive immunity.[a]

Type	Time from Exposure to Response	Form of the Antigen	Antibodies	Cellular Immune Mediators	Soluble Immune Mediators	Clinical Examples
I	Minutes	Protein or hapten-protein	IgE	Mast cells, basophils, eosinophils	Histamine, proteases, eicosinoids, cytokines, chemokines	Atopic allergies Allergic asthma
II	Hours to days	Cell or tissue bound	IgG or IgM	Neutrophils	C3b, C3a, C4a, C5a	Transfusion reactions
III	Hours to days	Soluble	IgG or IgM	Neutrophils	C3b, C3a, C4a, C5a	Lupus nephritis
IV	Days	MHC-bound peptide	None	APCs, T cells, NK cells, macrophages	IL-2, IFN-γ, chemokines	Allergic contact dermatitis

[a] Ig, immunoglobulin; MHC, major histocompatibility complex; APC, antigen-presenting cells; NK, natural killer; IL, interleukin; IFN, interferon.

c. Mast cells with IgE antibodies bound to their surface Fc receptors degranulate when allergen is encountered.

G. Innate immunity also causes localized tissue damage, but does not require a prior exposure to the foreign agent (ie, a sensitization phase).
  1. The tissue damage seen in **septic arthritis** is caused by the activation of neutrophils recruited to an infected joint.
  2. Innate immune tissue injury can be elicited by relatively nontoxic microbial products, such as bacterial DNA or cell wall lipopolysaccharide (LPS).
    a. **Sepsis** is an acute systemic inflammatory response to infection mediated by **danger signals** and the excessive production of inflammatory mediators.
    b. Responses to components of nonpathogenic microbial flora can induce organ-specific inflammatory tissue damage.
      (1) **Crohn's disease** is thought to be initiated by innate immune responses to intestinal microbial flora.
      (2) **Ulcerative colitis** appears to involve a similar imbalance of the mucosal immune system.

CROHN'S DISEASE

CLINICAL CORRELATION

- *Crohn's disease (CD) is a chronic transmural inflammation of the large and small bowel that begins with the loss of epithelial barrier function in the intestine.*
- *Some CD patients have mutations in the Nod2 antimicrobial protein, which diminishes their ability to clear bacteria that breach the epithelium.*

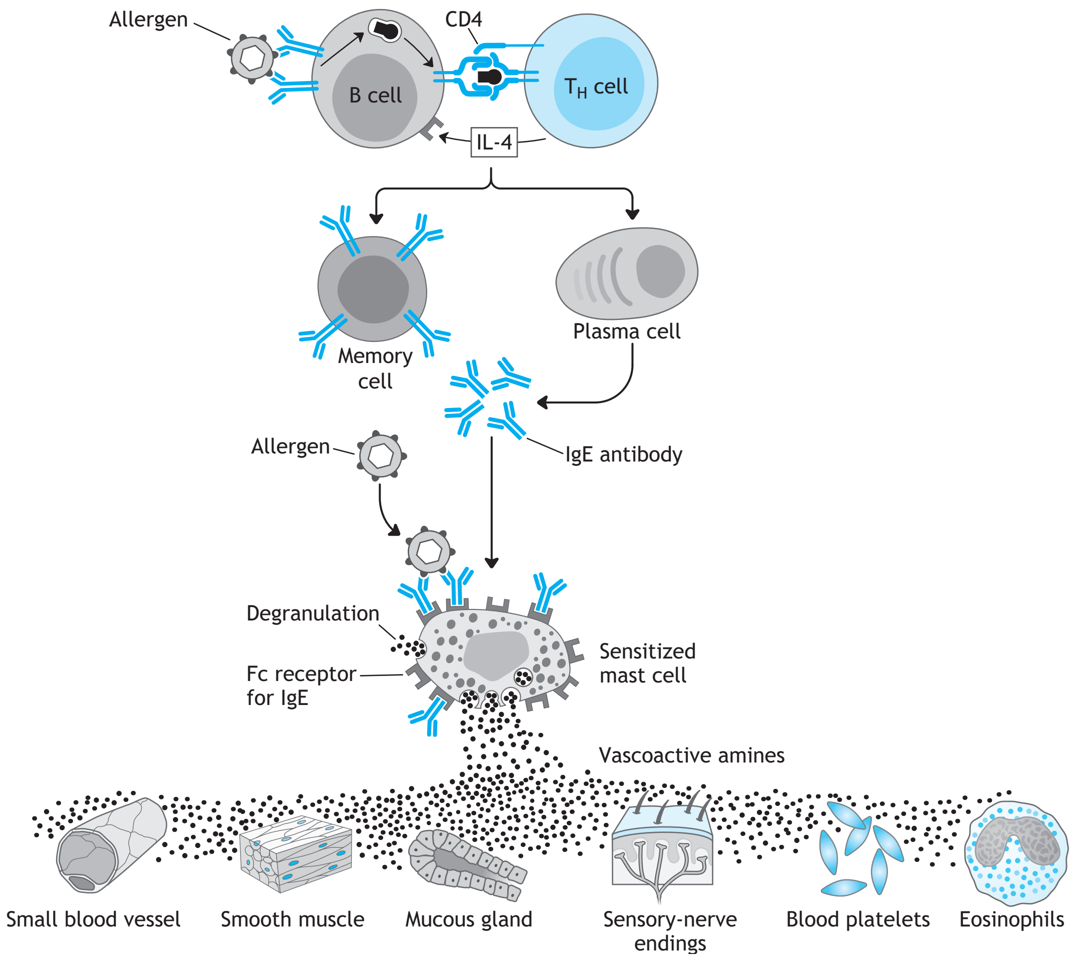
Allergen
CD4
B cell
T$_H$ cell
IL-4
Memory cell
Plasma cell
Allergen
IgE antibody
Degranulation
Fc receptor for IgE
Sensitized mast cell
Vascoactive amines
Small blood vessel
Smooth muscle
Mucous gland
Sensory-nerve endings
Blood platelets
Eosinophils

Figure 13–1. The sensitization and elicitation phase of immediate type immune tissue injury. IL-4, interleukin 4; IgE, immunoglobulin E.

- *Aberrant innate immune responses develop in the gastrointestinal tract to the intestinal flora.*
- *Activated macrophages and Th1 cells produce excessive interleukin (IL)-12, interferon (IFN)-γ, and tumor necrosis factor (TNF)-α.*
- *Sustained macrophage and neutrophil activation contributes to chronic inflammation.*
- *Blocking the action of TNF-α with monoclonal antibodies to the cytokine has proven effective in the treatment of CD.*

II. Type I (immediate-type) immune tissue injury (Table 13–2) is a rapid response to environmental antigens mediated by IgE antibodies and vasoactive inflammatory mediators released by mast cells.

A. The antigens that elicit these responses are **proteins** or **haptens** (eg, drugs) that can bind to proteins.
   1. The allergens are presented by major histocompatibility complex (MHC) class II molecules and activate **CD4$^+$** Th2 cells.
   2. Th2 cells produce IL-4 and IL-13, which induce class switching to IgE.

B. The propensity to make high-titer IgE antibodies to environmental allergens is, in part, genetically determined.
   1. Multiple genes control susceptibility to allergy and asthma.
   2. Among the traits that appear to be genetically controlled are total IgE levels, Th2 polarization, and mast cell and eosinophil growth.

C. Allergen-specific IgE antibodies bind to **Fcε receptor I (FcεRI)** on tissue **mast cells**.
   1. FcεRI is a very high-affinity receptor (K_d = 10 pM).
   2. The activation of mast cells through FcεRI requires receptor cross-linking by multideterminant antigens.
   3. Because the antibodies displayed on mast cells are specific for many different epitopes, polyvalent protein antigens can cross-link surface FcεRI molecules.

D. The tissue distribution and density of mast cells as well as the surface density of FcεRI determine the distribution of tissue injury.
   1. The route of immunization (eg, respiratory versus gastrointestinal) determines the local distribution of antibody specificities on tissue mast cells.
   2. Mucosal and connective tissue mast cells produce distinct patterns of inflammatory mediators.

HYPER-IgE SYNDROME

CLINICAL CORRELATION

- ***Hyper-IgE syndrome*** *or* ***Job's syndrome*** *is an immune deficiency characterized by recurrent and severe staphylococcal skin abscesses and pneumonia.*
- *Th1:Th2 imbalance results in a decreased production of IFN-γ, a cytokine that normally inhibits Th2 responses, such as switching to IgE production.*
- *These patients can also show pruritic dermatitis, but increased allergy is not common, despite exceedingly high serum IgE levels (> 2000 IU/mL).*

E. Signaling through FcεRI results in three mechanisms of inflammatory mediator release (Figure 13–2).
   1. **Preformed mediators** stored in cytoplasmic granules are released when mast cells **degranulate**.
      a. Allergen binding triggers FcεRI clustering.
      b. **Phospholipase Cγ (PLCγ)** is recruited to the immunotyrosine activation motifs (ITAMs) of the FcεRI and is activated.
      c. **Protein kinase C (PKC)** is then activated by diacylglycerol and phosphorylates myosin light chains.
      d. Cytoskeleton changes are induced that promote granule exocytosis and the release of granule constituents.
         (1) The preformed vasoactive amine **histamine** (Table 13–3) is released.
         (2) **Heparin**, the proteases **chymase** and **tryptase**, and **chemotactic peptides** that recruit neutrophils and eosinophils are also released from granules.
      e. Degranulation can be inhibited by maintaining elevated levels of **intracellular cyclic AMP (cAMP)**.

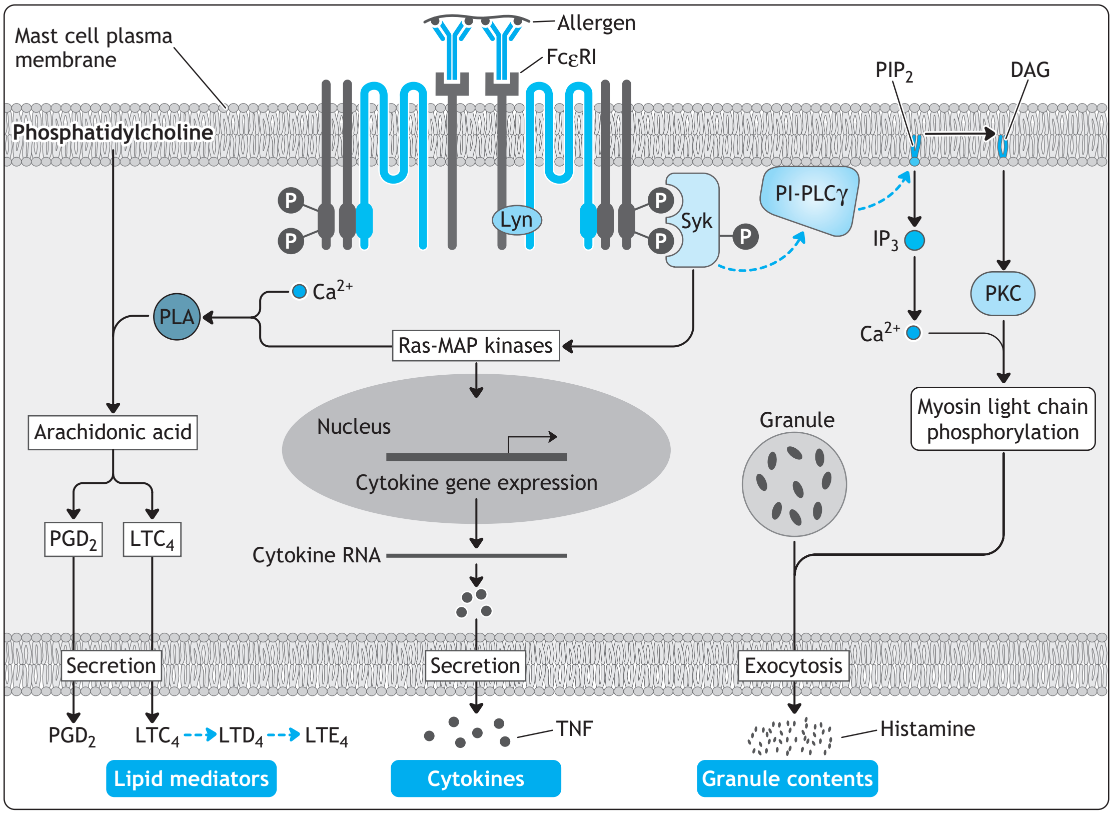
Mast cell plasma membrane
Allergen
FcεRI
PIP$_2$
DAG
Phosphatidylcholine
P
P
Lyn
P
P
Syk
P
PI-PLCγ
IP$_3$
Ca^{2+}
PLA
Ras-MAP kinases
PKC
Ca^{2+}
Myosin light chain phosphorylation
Arachidonic acid
Nucleus
Cytokine gene expression
Granule
PGD$_2$
LTC$_4$
Cytokine RNA
Secretion
Secretion
Exocytosis
PGD$_2$
LTC$_4$ → LTD$_4$ → LTE$_4$
TNF
Histamine
Lipid mediators
Cytokines
Granule contents

Figure 13–2. Fcε receptor I (FcεRI) signaling in mast cells. PIP$_2$, phosphatidylinositol 4,5-biphosphate; DAG, diacylglyceride; PLCγ, phospholipase Cγ; IP$_3$, inositol 1,4,5-triphosphate; PLA$_2$, phospholipase A$_2$; MAP, mitogen-activated protein; PKC, protein kinase C; PGD$_2$, prostaglandin D$_2$; LT, leukotriene; TNF, tumor necrosis factor.

2. FcεRI cross-linking also activates **phospholipase A$_2$** (PLA$_2$).
   a. PLA$_2$ cleaves phosphatidylcholine to form **arachidonic acid**.
   b. Arachidonic acid is metabolized by the **cyclooxygenase** and **lipoxygenase** pathways to form **prostaglandins (PGs)** and **leukotrienes (LTs)**, respectively.

3. The de novo expression of **cytokine genes** is induced by the activation of transcription factors.
   a. **Nuclear factor of activated T cells (NFAT), nuclear factor-κB (NFκB),** and **AP-1** are activated in mast cells.
   b. The genes for **IL-3, IL-4, IL-5, IL-6, IL-13,** and **TNF-α** are transcribed.

Table 13–3. Inflammatory mediators derived from activated mast cells.[a]

Preformed or Newly Synthesized	Mediators	Proinflammatory Effects	Mast Cell Signaling Intermediates
Preformed mediators	Histamine Heparin Tryptase Chymase E-CFA N-CFA	Itchy sensation, vasodilation, increased vascular permeability, bronchoconstriction, mucus secretion, leukocyte chemotaxis, gut peristalsis	PLCγ, Ca^{2+} mobilization, PKC activation, myosin phosphorylation, granule exocytosis
Newly synthesized mediators	PAF, prostaglandins (PGD_2), leukotrienes (LTC_4) IL-3, IL-4, IL-5, IL-6, TNF-α, eotaxin	Vasodilation, increased vascular permeability, bronchoconstriction, eosinophil chemotaxis, mucus secretion, platelet aggregation, ICAM-1 expression, mast cell growth, Th2 cell differentiation	PLA_2 activation, arachidonic acid, COX-2 NFκB, NFAT, and AP-1 activation

[a] E-CFA, eosinophil chemotactic factor of anaphylaxis; N-CFA, neutrophil chemotactic factor of anaphylaxis; PLCγ, phospholipase Cγ; PKC, protein kinase C; PAF, platelet-activating factor; IL, interleukin; TNF, tumor necrosis factor; ICAM, intercellular adhesion molecule; PLA_2, phospholipase A_2; COX, cyclooxygenase; NFκB, nuclear factor κB; NFAT, nuclear factor of activated T cells.

IgE-INDEPENDENT MAST CELL ACTIVATION

CLINICAL CORRELATION

- *Several agents can activate mast cells without binding as allergens to cell surface IgE.*
- *Certain **drugs** (morphine, codeine, calcium ionophores, vancomycin, radiocontrast agents) can activate mast cells directly and induce anaphylaxis.*
- *The **anaphylatoxins** C3a, C4a, and C5a trigger mast cell degranulation.*
- *The **neuropeptides** substance P and somatostatin mediate cholinergic triggering of asthma by activating mast cells.*
- *Lectins, such as those found in some foods, can bind carbohydrate residues on IgE and induce adverse reactions to foods.*

F. The signs and symptoms of allergy reflect the action of mast cell mediators on target tissues (Table 13–3).
  1. Most of the inflammatory effects of **histamine** in allergy are mediated through binding of histamine to H_1 **histamine receptors**.
    a. H_1 receptors are G protein coupled and activate adenylate cyclase.
    b. Adenylate cyclase activation leads to a *transient* increase in intracellular cAMP concentrations.
  2. The effects of histamine depend on the cell type affected.

   - **a.** Histamine relaxes smooth muscle cells in peripheral blood vessels causing **vasodilation**.
   - **b.** Histamine induces **bronchospasms** in the lungs and **peristalsis** in the intestines by stimulating smooth muscle contraction.
   - **c.** Histamine causes contraction of vascular endothelial cells resulting in **increased vascular permeability** and edema.

3. **Tryptase** and **chymase** are serine proteases stored in granules.
   - **a.** The circulating levels of tryptase are a clinical measure of recent mast cell degranulation.
   - **b.** Tryptase induces **mucus secretion** in the lungs.
   - **c.** Tryptase can activate certain matrix metalloproteases.
   - **d.** Tryptase cleaves plasminogen to form plasmin, which cleaves C3.

4. Mast cell granules contain two chemotactic peptides.
   - **a.** **Eosinophil chemotactic factor of anaphylaxis (E-CFA)** attracts eosinophils to the lung in asthma and to the intestine in certain parasitic infections.
   - **b.** **Neutrophil chemotactic factor of anaphylaxis (N-CFA)** recruits neutrophils that contribute to tissue damage.

5. Proteoglycans, including **heparin** and **chondroitin sulfate**, are granule constituents to which other mast cell mediators are bound.

SYSTEMIC ALLERGIC ANAPHYLAXIS

CLINICAL CORRELATION

- *The allergens that typically induce systemic anaphylaxis include drugs, insect venoms, and certain foods.*
- *Key target organs include the cardiovascular, pulmonary, gastrointestinal, and cutaneous systems.*
- *Typical signs and symptoms of anaphylatic shock include generalized pruritis, apprehension, hypotension, abdominal cramping, tachycardia, tachypnea, and hives.*
- *Timely injection of epinephrine is life saving.*
- *Drugs that inhibit the late phase of anaphylaxis are essential to avoid a secondary shock response several hours later.*

6. Three groups of **lipid mediators** contribute to allergic responses.
   - **a.** **Platelet activating factor (PAF)** is derived from the PLA$_2$-mediated hydrolysis of membrane phospholipids.
      - ***(1)*** PAF is produced by mast cells, basophils, and endothelial cells.
      - ***(2)*** PAF induces **vasodilation, bronchoconstriction,** and increased **vascular permeability**.
   - **b.** **PGs** are derived from the metabolism of arachidonic acid via the **cyclooxygenase (COX) pathway**.
      - ***(1)*** **COX-2** is the principal mast cell cyclooxygenase.
      - ***(2)*** **PGD$_2$** is the principal prostaglandin produced.
      - ***(3)*** PGD$_2$ causes **vasodilation** and **bronchoconstriction**.
   - **c.** **LTs** are derived from arachidonic acid by the **lipoxygenase pathway**.
      - ***(1)*** **LTC$_4$**, **LTD$_4$**, and **LTE$_4$** are the principal leukotriene mediators of allergy.
      - ***(2)*** LTC$_4$ is produced by mucosal mast cells, but not connective tissue mast cells (eg, in the skin).
      - ***(3)*** LTC$_4$ is a major cause of **bronchoconstriction**.

7. **Cytokines** are expressed by mast cells following transcription.
    a. **IL-3** and **IL-4** promote the growth of mast cells.
    b. **IL-4** and **IL-13** induce B cell class switching to IgE.
    c. **IL-5** is a chemotactic factor for eosinophils.
    d. **IL-4** promotes eosinophil growth.
    e. **TNF-α** facilitates leukocyte chemotaxis by inducing vascular intercellular adhesion molecule (ICAM)-1 and leukocyte integrin expression.
    f. Cytokines mediate many of the **late-phase** allergic responses.

ALLERGIC ASTHMA

CLINICAL CORRELATION

- *The allergic form of asthma is caused by IgE-mediated immunity to inhaled allergens.*
- *Th2 polarization probably begins early in life.*
- *Asthmatic "attacks" can be triggered by either intrinsic signals (eg, cold outside air or exercise) or extrinsic signals (ie, allergens).*
- *Chronic mast cell activation leads to persistent mediator release, bronchoconstriction (chest tightness), mucus secretion, and eosinophil recruitment and activation.*
- *Eosinophils damage the mucosal epithelium, cause perivascular edema, and initiate lung tissue remodeling.*
- *Mucus plugging of the bronchi results in the hyperinflation of the lungs, an inability to expire air, and expiratory wheezing.*
- *Early intervention is designed to diagnose and treat the underlying allergic conditions.*
- *As the disease progresses, more aggressive treatment to control chronic inflammation (eg, corticosteroids) must be considered.*

G. The proper diagnosis of atopic allergies requires a careful patient history and allergy skin testing.
  1. The diagnostic gold standard for allergy is the **skin test**.
      a. Allergen injected into the skin induces histamine release, localized edema (**wheal**), and a zone of erythema (**flare**) within 20 minutes.
      b. Negative and positive controls (eg, saline and histamine, respectively) are essential to validate the test.
  2. The **radioallergosorbant test (RAST)** is useful for measuring the circulating levels of IgE antibodies to known allergens (Chapter 5).

HAY FEVER

CLINICAL CORRELATION

- ***Allergic rhinitis** or **hay fever** can be a perennial or seasonal condition.*
- *For example, in the Midwest, ragweed pollenosis shows a well-delineated season with symptoms appearing first in mid-August and lasting through the first frost.*
- *Rhinitis presents as sneezing, nasal congestion, and watery, itchy eyes.*
- *Typical allergens include pollens, fungal spores, animal dander, and house mites.*
- *The accurate identification of the allergen requires a careful patient history combined with skin testing using standardized allergens.*

H. Prophylaxis and treatment of allergic conditions take many forms.
  1. The first approach to prophylaxis is to avoid the allergen.
  2. **Immunotherapy** is an effort to divert the immune system away from IgE-mediated triggering of mast cells.

   - **a.** The patient receives repeated intradermal injections of the allergen over several months.
   - **b.** Some patients produce IgG and IgA antibodies that compete with allergenic IgE.
   - **c.** The IgE antibody levels in some patients decrease by unknown mechanisms.
   - **d.** A Th2 to Th1 shift in immunity to the allergen may occur.

3. Antibodies to the Fc portion of IgE (omalizumab) are beneficial for treating severe allergic asthma (Table 13–4).
   - **a.** The antibodies decrease circulating levels of IgE.
   - **b.** They decrease the density of FcεRI on tissue mast cells.

4. **Desensitization** is a therapy designed to establish a temporary block of allergen-induced mast cell mediator release.
   - **a.** Desensitization is most often used when no alternative therapy exists for a patient with a drug allergy.
   - **b.** The patient is given the allergenic drug in incrementally increasing doses beginning with a subclinical dose.
   - **c.** Desensitization probably results from the gradual, controlled release of mast cell mediators.
   - **d.** The desensitized state persists only while the drug is given.

Table 13–4. Pharmacological treatment of allergy.[a]

Group	Examples	Mechanisms of Action	Clinical Uses
Anti-IgE	Omalizumab	Decreasing circulating IgE	Asthma
Antihistamines	Diphenhydramine, fexofenadine	H_1 receptor antagonists	Allergic rhinitis, urticaria, anaphylaxis
β_2-Agonists	Epinephrine, albuterol	β_2-Adrenergic agonist	Anaphylaxis, asthma, urticaria
NSAIDs	Indomethacin	Blocks PG synthesis	Urticaria
Corticosteroids	Beclomethasone, methylprednisolone	Blocks PLA_2 and cytokine expression	Allergic rhinitis, asthma, anaphylaxis
Mast cell stabilizers	Cromolyn sodium	Prevents granule fusion	Allergic rhinitis Asthma
Methylxanthines	Theophylline	Inhibits phosphodiesterase	Asthma
Leukotriene receptor antagonists	Zafirlukast	Blocks LT action	Asthma

[a] IgE, immunoglobulin E; NSAIDs, nonsteroidal antiinflammatory drugs; PG, prostaglandin; PLA_2, phospholipase A_2; LT, leukotriene.

5. Inhibiting mast cell degranulation is effective against allergy.
   - a. **Cromolyn sodium** inhibits Ca^{2+} influx, an important step in FcεRI signaling.
   - b. **Theophylline** maintains high intracellular cAMP levels in mast cells.
   - c. **Nonsteroidal antiinflammatory drugs (NSAIDs)** (eg, indomethasone) inhibit arachidonic acid metabolism.
   - d. **Hydrocortisone** acts at multiple steps in mast cells.
     - *(1)* It inhibits histamine synthesis from histidine.
     - *(2)* It inhibits the activation of PLA_2.
     - *(3)* The drug blocks NFκB activation by increasing the expression of its inhibitor IκB.

FOOD ALLERGIES

CLINICAL CORRELATION

- *Food intolerance can be IgE dependent or IgE independent.*
- *Most adults show the IgE-dependent form, whereas significant numbers of infants or children have food intolerance that is not IgE mediated.*
- *Common allergens include proteins of peanuts, soybeans, shellfish, milk, and eggs.*
- *Oropharyngeal symptoms (pruritis and urticaria in the mouth) appear first followed by nausea, cramping, vomiting, flatulence, and diarrhea.*
- *Urticarial rashes and even anaphylaxis can result from allergic reactions to foods.*
- *IgA deficiency increases the childhood risk of developing IgE-mediated food allergies.*

6. Some drugs block the effects of allergic mediators.
   - a. The β-adrenergic receptor agonist **epinephrine** blocks histamine action.
   - b. Epinephrine is life saving in acute allergic anaphylaxis.
   - c. The β_2-receptor agonist **albuterol** is an effective bronchodilator used for the treatment of allergic asthma.
   - d. Antihistamines include the H_1 receptor antagonists (eg, **diphenhydramine** and **chlorpheniramine**).
   - e. Leukotriene receptor antagonists are used to treat asthma.
   - f. **Hydrocortisone** and other corticosteroids also block inflammatory gene expression in target cells.

III. Type II (cytotoxic) immune tissue injury results from the recognition of cell- or tissue-bound antigens by IgG and IgM antibodies (Table 13–2).

A. Tissue damage generally requires hours or days to become evident.

B. **IgM** and **IgG antibodies** activate **complement** through the classical pathway.
   1. Tissue damage can result from the formation of the **membrane attack complex** C5b-9.
   2. C5a is **chemotactic** for neutrophils.
   3. C3a, C4a, and C5a are **anaphylatoxins** that directly activate mast cells.
   4. Phagocytic cells can cause tissue damage when they attach to IgG- or iC3b-coated tissues and cells.

C. Anemia is a common effect of cytotoxic immune tissue injury directed toward erythrocyte antigens.
   1. **Transfusion reactions** result when mismatched blood is transfused to a patient with a preformed antibody.

2. **Rh disease** of the newborn is due to maternal IgG antibodies that cross the placenta and bind to fetal Rh antigens.
3. A number of **autoimmune diseases** (eg, Coombs-positive hemolytic anemias and thrombocytopenias) are mediated by host antibodies reacting with autoantigens (Chapter 16).

D. Damage to solid tissues and extracellular matrix is also mediated by cytotoxic immune tissue injury.

1. **Hyperacute graft rejection** is mediated by anti-HLA antibodies.
2. Autoantibodies in **Goodpasture syndrome** react with type IV collagen of the glomerular basement membrane.

E. The **diagnosis** of type II immune tissue injury is often based on detecting an abnormal distribution of IgG.

1. The Coombs test detects Ig on erythrocytes.
2. Immunofluorescence can detect IgG deposition within tissues.

IV. Type III (immune complex) tissue injury results when soluble antigen–antibody complexes form in the blood and deposit in tissues.

A. Immune complex deposition in the kidneys causes **nephritis**.

B. Immune complexes in the joints result in **arthralgia** and **arthritis**.

C. Immune complexes in the skin cause a range of lesions, including **rashes**.

D. Immune complex deposition in blood vessel walls causes **vasculitis, purpura,** and **edema**.

E. IgG-containing immune complexes activate complement in tissues.

1. C5a attracts neutrophils to sites of immune complex deposition.
2. iC3b serves as an adhesion ligand promoting neutrophil binding.
3. The membrane attack complex damages cells.
4. Immune complex formation often causes a transient decrease in the serum levels C1, C2, C3, and C4.

F. **Diagnosis** involves the detection of antibodies in the serum or affected tissues.

SERUM SICKNESS

CLINICAL CORRELATION

- ***Serum sickness*** *was first described in the preantibiotic era as a reaction to horse antitoxins given repeatedly for the treatment of toxin-mediated infectious diseases.*
- *The condition can arise after antibody production to either protein [eg, intravenous immunoglobulins (IVIG), foreign monoclonal antibodies, vaccines] or nonprotein (eg, penicillin or sulfonamides) antigens.*
- *Serum sickness typically becomes evident 1–3 days after exposure to the antigen.*
- *Signs and symptoms include fever, rash, hypotension, anaphylactoid purpura, lymphadenopathy, myalgia, arthralgia, and nephritis.*
- *The treatment of* ***anaphylactoid syndrome*** *is the same as that for anaphylaxis, ie, discontinuance of the antigen and administration of epinephrine, antihistamines, and corticosteroids.*

V. Type IV (delayed-type) immune tissue injury is a manifestation of cell-mediated immunity to haptens or protein antigens.

A. The immunologically specific component of **delayed-type hypersensitivity (DTH)** is the Th1 cell, rather than antibody.

1. Allergens must be proteins or chemically reactive groups capable of binding covalently to proteins.
2. Important cytokines include **IL-1, TNF-α, IL-2, IL-12, IFN-γ**, and the **CCL** group of chemokines (Chapter 12).

B. Host tissue damage is mediated by cytotoxic T lymphocytes, natural killer (NK) cells, macrophages, and a range of soluble inflammatory mediators.
   1. Macrophages and Th1 cells produce the proinflammatory cytokines IL-1, IL-2, IFN-γ, and TNF-α.
   2. Macrophages produce eicosinoids (LTs and PGs), hydrolytic enzymes, and reactive oxygen species.

C. The important clinical manifestations of type IV hypersensitivity show delayed onset (2–3 days) and include **contact dermatitis**.
   1. Plant resins (eg, poison ivy pentadecacatechol), metals (eg, nickel) and chemicals (eg, cosmetics) cause a variety of rashes.
   2. Industrial exposure to certain drugs, including antibiotics, can cause type IV hypersensitivities.

D. Delayed-type hypersensitivity (DTH) skin lesions are initially characterized by their **indurated** (firm) nature, rather than the edematous (fluid-filled) lesions of atopic allergy.

E. The diagnosis of allergic contact dermatitis is aided by the **patch test**.
   1. Solutions of test allergens are placed on absorbent skin patches.
   2. The extent of induration and erythema beneath each patch is determined 48–72 hours later.

CLINICAL PROBLEMS

A 44-year-old school teacher who paints houses in the summer presents in early June with erythematous, indurated, and blistering lesions limited to his hands and forearms. The lesions typically appear each summer within several days of beginning to use painting products, including solvents and epoxy resins. He indicates that he does not wear protective gloves when working with these chemicals. The lesions disappear several weeks after he returns to his teaching position in the fall.

1. Which of the following represents a likely step in the pathogenesis of this disease?
   A. Activation of C1
   B. Chemotaxis of neutrophils
   C. Induction of IgE antibodies
   D. Antigen presentation by the MHC class II pathway
   E. Immune complex deposition

2. Which of the following test results would best fit the clinical findings in this case?
   A. A positive skin test after 20 minutes with a dilute solution of epoxy resin
   B. A positive skin patch test with a dilute solution of epoxy resin

C. Decreased serum levels of C1, C2, C3, and C4

D. A decreased CD4:CD8 ratio

E. Elevated serum tryptase

Jim received penicillin for pneumonia several years ago and presented last week with what appeared to be an allergic reaction to the drug when it was taken for a sinus infection. The next day after beginning oral penicillin, Jim's eyelids became puffy and his abdomen was covered with itchy, nearly confluent hives. After the administration of diphenhydramine and hydrocortisone his condition improved. Laboratory tests performed on his acute serum showed an elevated level of tryptase and decreased C1 and C3. A skin test with penicillin was negative.

3. Which of the elements of this case suggests this reaction was caused by circulating IgG antibody–drug complexes, rather than IgE antibody to penicillin?

A. His plasma tryptase level was elevated.

B. His symptoms were corrected by administering antihistamines.

C. His C1 and C3 levels were decreased.

D. His skin lesions were pruritic (itchy).

E. He showed evidence of altered vascular permeability.

Jesse is a 10-year-old child who complains of chest tightness and wheezing after gym class and upon exposure to cold outside air. He experiences sneezing, nasal itching, and nasal congestion indoors and following contact with his friend's dog. His allergist has determined by skin testing that he is allergic to oak pollen, Bermuda grass, house dust, and dog dander. His pulmonary function tests indicate a forced expiratory volume in 1 second (FEV_1) < 60% of the expected value for his age and a peak flow rate of 180 L/min (normal > 350 L/min).

4. Which of the following drugs would be most beneficial in alleviating Jesse's pulmonary symptoms?

A. Albuterol

B. Anti-TNF-α monoclonal antibody

C. IVIG

D. Cyclosporin

E. Recombinant IL-4

Anna is a 7-year-old child who presents with itchy, erythematous, indurated lesions on her forearms and legs 3 days after a camping trip with her classmates. Her mother treats Anna with antihistamines, which reduces the itching sensation, but does not correct the skin lesions.

5. What additional therapy would probably benefit this child?

A. Albuterol by inhaler

B. Topical corticosteroids

C. Epinephrine

D. Anti-IgE
E. Cromolyn sodium

ANSWERS

1. The correct answer is D. The patient's history suggests this is an occupational exposure to a contact allergen in solvents. The indurated nature of the lesions suggests mononuclear cell infiltration into the skin, and the time necessary for their development (several days) is consistent with allergic contact dermatitis.
2. The correct answer is B. The standard diagnostic test for allergic contact dermatitis is the patch test. Dilute solutions of well-standardized test allergens are available that can be placed on the skin. It is essential that the test samples be dilute enough to not cause primary irritant reactions on the skin.
3. The correct answer is C. This clinical picture would normally be attributed to a type I hypersensitivity reaction mediated by IgE antibodies. However, two aspects of his reaction suggest this is not the pathogenic mechanism. First, he did not test positive to the allergen in an immediate-type skin test. Second, his complement levels were decreased, suggesting the activation and depletion of complement by circulating antigen–antibody complexes. It should also be noted that his symptoms did not appear immediately.
4. The correct answer is A. Albuterol is a β_2-receptor agonist and is used as a bronchodilator in asthma. It acts by antagonizing histamine-induced signaling in the lung. Jesse's pulmonary dysfunction results from his lungs being hyperinflated. In addition, his breathing is impaired by bronchoconstriction and mucus plugging.
5. The correct answer is B. The time to onset of symptoms and the indurated nature of the lesions suggest this is a case of poison ivy rash. Corticosteroids are administered topically to inhibit macrophage and lymphocyte activation in the skin, which are two features of delayed-type contact dermatitis.

CHAPTER 14

PROTECTIVE IMMUNITY AND VACCINES

I. Only a subset of the immune responses we make to microbes is protective.

A. **Neutralizing antibodies** are very effective at controlling infectious agents and limiting the effects of their toxins.
   1. Antibodies can prevent viral entry into and uncoating within host cells.
   2. Neutralizing antibodies against microbial exotoxins (eg, tetanus toxin) block toxin binding to host cells.
   3. **Seroconversion** refers to the appearance of antibodies to a microbial pathogen in a patient.
      a. Seroconversion can aid in the diagnosis of a specific infectious disease.
      b. Alternatively, seropositivity may simply indicate a prior encounter with the organism (or a cross-reacting species), rather than an ongoing or recent infection.

B. The **clearance** of extracellular microbes is primarily mediated by opsonophagocytosis.
   1. Opsonins include antibodies, complement peptides, and soluble pattern recognition molecules (Chapter 1).
   2. Opsonic receptors also initiate microbial killing mechanisms.

C. **Killing** of microbes can occur either outside or inside immune cells.
   1. Extracellular pathogens that are small enough to be phagocytized (eg, bacteria) are killed within phagocytic cells.
   2. Large extracelluar pathogens (eg, filamentous fungi) are killed when phagocytes adhere to the surface of the microbe.

D. The **resolution** of an infection requires the complete eradication of the infectious agent.
   1. Patients with underlying cellular immune defects often do not resolve infections well.
   2. Infections often reappear in these patients after antimicrobial therapy is discontinued.
   3. Unresolved infections can also persist in normal individuals if the microbe (eg, *Mycobacterium tuberculosis*) establishes a **latent infection**.

E. Innate and adaptive immunity provide distinct forms of protection.
   1. **Innate immunity** to microbes is essential early in an infection and lowers the early microbial burden in the host.

2. Innate immunity is more important in immunologically naive hosts (eg, neonates) who lack **adaptive immunity** to the pathogen.
3. Adaptive immunity arises late in primary infections and often mediates resolution of infections.
4. Adaptive immunity can mediate lifelong specific protection against a particular pathogen.
5. Adaptive immunity is the objective of our current **vaccines**.

CLINICAL CORRELATION

WORLDWIDE INFECTION RATES

- *Each year infectious diseases account for one-third of all deaths worldwide.*
- *Infectious diseases that are prevalent in developing countries are often not common in the United States.*
- *For example, tens of millions of individuals worldwide have schistosomiasis, leishmaniasis, and tuberculosis.*
- *Malaria, tuberculosis, and AIDS each causes several million deaths each year.*
- *Because of the complexity of these organisms and the relatively low level of research on these pathogens, effective vaccines have been slow to emerge.*

II. The specific protective immune response to a microbial pathogen depends on the life-style of the microorganism (Tables 14–1 and 14–2).

A. Protection against **extracellular pathogens** (eg, most bacteria) requires killing of the microbe within an intracellular organelle by phagocytic cells.

B. Large **extracellular pathogens** (eg, filamentous fungi) are killed when phagocytes direct these antimicrobial responses to the extracellular environment.

C. Protective responses to **obligate intracellular microbes** limit microbial replication or kill the infected cell as a means of destroying the microbe.

D. While in the extracellular phase of their life cycle, **obligate and facultative intracellular microbes** can be neutralized and cleared from the host by antibodies and complement.

III. At least eight distinct immune mechanisms are known to provide protection against microbial pathogens.

A. **Membrane lysis** is most effective against microbes with outer membranes (eg, gram-negative bacteria) or lipid envelopes (eg, enveloped viruses).

1. Among the gram-negative bacteria, *Neisseria* species are among the most susceptible to complement-mediated membrane attack.
2. **Antimicrobial peptides** appear to act by damaging the outer membranes of bacteria.

B. **Opsonophagocytosis** is primarily directed at extracellular bacteria, fungi, and parasites.

1. Most opsonic receptors signal phagocytes for increased antimicrobial activity.
2. **Pyogenic bacteria** are particularly susceptible to opsonophagocytic clearance and killing.

C. **Microbial killing** by neutrophils is essential for elimination of extracellular microbes.

Table 14–1. Protective immune responses to infectious agents.[a]

Protective Mechanisms	**Immune Components**	**Groups of Pathogens**	**Specific Pathogens**
Membrane attack	Defensins, cathelicidins Complement	Gram-negative bacteria	*Pseudomonas aeruginosa* *Neisseria* species
Opsonophagocytosis	Neutrophils, macrophages Opsonic receptors (CR3, FcR) and ligands (iC3b, IgG)	Extracellular pyogenic bacteria and yeasts	*Staphylococcus, Streptococcus, Pseudomonas* species
Intracellular killing	Phagolysosomes; reactive oxygen	Extracellular bacteria and yeasts	Many catalase-negative bacteria and fungi
Toxin neutralization	IgM, IgG, IgA antibodies Immune complex clearance	Exotoxin-producing microbes	*Clostridium tetani, Staphylococcus aureus, Escherichia coli*
Virus neutralization	IgM, IgG, IgA antibodies Phagocytic cells, mucus	Respiratory and gastrointestinal viruses	Rhinoviruses, adenoviruses, rotaviruses
Extracellular killing	Neutrophils, macrophages	Filamentous fungi, parasites	*Histoplasma capsulatum, Candida albicans*
Macrophage-mediated intracellular killing	Th1 cells, NK cells, IFN-γ, nitric oxide	Obligate or facultative intracellular pathogens	*Mycobacterium tuberculosis, Listeria monocytogenes, Candida albicans, Histoplasma capsulatum*
Cytotoxic T cell killing	CD8$^+$ T cells, MHC class I, CD4$^+$ Th cells, and IL-2	Viruses and many intracellular bacteria and fungi	Herpes simplex virus, *Listeria monocytogenes*
Eosinophil-mediated cytotoxicity	Eosinophils, IgE antibodies, FcεR	Intestinal parasites (helminths)	*Taenia saginata*
Inhibition of intracellular replication	Type I interferons	Most viruses	Herpesviruses
NK cell-mediated cytotoxicity	NK cells, MHC class I	Many viruses	Herpesviruses

[a] Ig, immunoglobulin; NK, natural killer; IFN, interferon; MHC, major histocompatibility complex.

Table 14–2. Important protective immune components organized by microbe class.[a]

Immune Component	Extracellular Bacteria	Intracellular Bacteria	Viruses	Fungi	Parasites
Antibody	++	Minimal	++	–	++
Complement	++	Minimal	+ (enveloped)	–	–
Neutrophils	++	– or +	–	+	– (eosinophils)
Th2 cells	++	–	++	–	–
Th1 cells	–	++	++	++	++
Macrophages	+	++	+	++	++
Cytotoxic T cells	–	+	++	–	–
NK cells	++ (cytokines)	+ (cytokines)	+ (lysis)	Unknown	Unknown

[a] NK, natural killer.

1. **Reactive oxygen species** (eg, hydrogen peroxide) mediate killing of many aerobic microbes.
2. A more limited range of microbes (eg, *Lysteria, Mycobacterium* species) is killed by **reactive nitrogen species** (eg, nitric oxide).
3. For the killing of anaerobic microbes, oxidants are less important than antimicrobial peptides and lysosomal hydrolases.

D. Intracellular microbes (eg, *Mycobacterium, Legionella, Listeria, Salmonella,* and *Francisella* species) that replicate within mononuclear phagocytes are particularly susceptible to **Th1-type cellular immunity**.
   1. Macrophage activation by Th1 cytokines is a key step.
   2. A combination of reactive oxygen and nitrogen metabolites and oxygen-independent mediators mediates killing.

E. **Toxin neutralization** is important in decreasing the severity of infectious diseases mediated by endo- and exotoxins.

F. **Virus neutralization** is particularly important in blocking virus attachment to mucosal surfaces and in clearing viruses from the blood stream.
   1. Secretory immunoglobulin (Ig) A antibodies protect against inhaled and ingested viruses.
   2. IgG antibodies reduce viremia and block virus spread.
   3. Antibodies are particularly important for viruses (eg, rabies, polio) that require hematogenous dissemination.

G. **Cytotoxic T cells (CTL)** provide protection against intracellular pathogens whose antigens are presented by major histocompatibility complex (MHC) class I molecules.
  1. Most virus-infected cells are susceptible to CTL attack.
  2. Microbes that block the class I pathway of antigen presentation (eg, herpesviruses) can evade CTL-mediated clearance.
  3. Intracellular bacteria that escape the phagosome (eg, *Listeria, Francisella*) are also subject to detection by CTL.

H. **Natural killer (NK) cells** also protect against pathogen-infected host cells.
  1. NK cells can recognize and kill infected host cells in which the microbe has inhibited **MHC class I** expression.
  2. NK cells produce **type I interferons** (IFN-α, β) early in infection.
    a. IFN-α, β establishes an antiviral state within infected cells.
    b. IFN-α, β induces RNase L and protein kinase R, which degrade viral mRNA and block translation, respectively.

I. **Eosinophils** use **IgE antibodies** to attack extracellular parasites (eg, intestinal worms) by antibody-dependent cellular cytotoxicity (ADCC).

III. The nature of opportunistic infections in immunocompromised patients can reveal important protective immune responses (Table 14–3).

A. Infections with intracellular pathogens are common in patients lacking **Th1 cell, CTL, NK cell,** or **macrophage** functions.

B. Microbes that replicate within the cytosol (eg, viruses) are particularly problematic for patients with **CTL** or **MHC class I** defects.

Table 14–3. Common infections in the immunocompromised patient.[a]

	T Cells	B Cells	Phagocytic Cells	Complement
Bacteria	*Mycobacterium, Listeria*	*Stahpylococcus, Streptococcus, Haemophilus*	*Staphylococcus, Streptococcus, Haemophilus*	*Staphylococcus, Streptococcus, Neisseria*
Fungi	*Candida, Histoplasma, Pneumocystis*		*Aspergillus, Candida*	
Viruses	CMV, HSV, VZV, EBV, JC	Enteroviruses, influenza, RSV		
Principal mechanisms	Macrophage, NK cell, and CTL activation	Virus neutralization, opsonophagocytosis	Phagocytosis, intracellular killing	Membrane attack, chemotaxis and opsonophagocytosis

[a] CMV, cytomegalovirus; HSV, herpes simplex virus; VZV, varicella-zoster virus; EBV, Epstein–Barr virus; RSV, respiratory syncytial virus; NK, natural killer; CTL, cytotoxic T cell.

C. Opportunistic infections with extracellular bacteria (eg, *Staphylococcus aureus*) and some cytolytic viruses (eg, influenza viruses) are common in **antibody-deficiency** states.

D. Extracellular bacteria and fungi commonly infect patients with **neutrophil** defects, including deficiencies in oxidant production or granule expression.

E. Patients with **complement** deficiencies cannot lyse gram-negative bacteria and enveloped viruses or clear extracellular microbes.

IV. Microbial immune evasion mechanisms have evolved to circumvent the most important protective responses by the host.

A. **Antigenic variation** allows bacteria, viruses, and parasites to avoid immune recognition.

1. Point mutations in antigenic proteins **(antigenic drift)** occur in influenza and human immunodeficiency virus (HIV)-1 by error-prone nucleic acid replication.
2. Other organisms, including *Neisseria gonorrhea* and *Trypanosoma cruzii*, scramble the genes that encode their major surface antigens.
3. More substantial changes in antigens **(antigenic shift)** occur by genetic recombination between related viruses.

B. Adopting an **intracellular life cycle** is a microbial response to immune elimination mediated by antibody and complement.

1. The host response to this microbial strategy has been to evolve CTL and NK cells capable of recognizing a microbe-infected cell.
2. Antigen presentation can be viewed as a host adaptation to intracellular microbial parasitism.

C. **Disguising microbial antigens** prevents immune recognition and attack.

1. Schistosomes coat themselves with host MHC molecules.
2. *Plasmodium* species (etiological agents of malaria) infect erythrocytes that do not express MHC molecules.
3. Many bacteria (eg, *Streptococcus pneumoniae* and *Haemophilus influenzae*) express thick polysaccharide capsules that disguise their more antigenic cell wall components.
4. The long polysaccharide chains of bacterial lipopolysaccharide (LPS) molecules direct complement activation away from the bacterial outer membrane.

ANTIGENIC SHIFT IN THE INFLUENZA A VIRUS

CLINICAL CORRELATION

- *The influenza A virus is responsible for periodic worldwide pandemics.*
- *Protective immunity to this virus depends on the production of neutralizing antibodies to two surface glycoproteins, neuraminidase (N) and hemagglutinin (H).*
- ***Antigenic drift** is the gradual change in the N and H serotypes by mutation.*
- *Regular monitoring of the N and H gene sequences determines the serotype of the predominant influenza virus each flu season, and an appropriate vaccine is produced.*
- *Occasionally, a major **antigenic shift** occurs due to reassortment of genomes between different influenza virus strains, including those of animals.*
- *The 1957 "Asian flu" pandemic resulted from the sudden emergence of new H and N serotypes by genetic reassortment between antigenically distinct influenza viruses.*

D. **Suppression of the immune response** is another common approach to evasion.

1. HIV-1 infects and depletes $CD4^+$ T cells.
2. Herpesviruses, HIV-1, and adenoviruses inhibit MHC expression or transporter associated with antigen processing (TAP) function and thereby depress CTL activation.
3. Many viruses produce proteins that mimic inhibitory cytokines or bind to activating cytokines (eg, Epstein–Barr virus and poxviruses, respectively).
4. Protein A produced by staphylococci binds IgG Fc regions.
5. *Streptococcus pyogenes* produces a C5a peptidase.
6. *N gonorrhea* produces a protease that degrades secretory IgA.
7. *Legionella* and *Listeria* species avoid intracellular killing by blocking phagosome–lysosome fusion or escaping from the phagosome, respectively.

V. Effective vaccines augment protective host responses to infection.

A. Vaccines do not typically prevent infections, although they reduce the severity of infectious diseases.

B. Most existing vaccines are **preemptive**; they must be given prior to infection (Table 14–4).

Table 14–4. Recommended immunizations for children.[a]

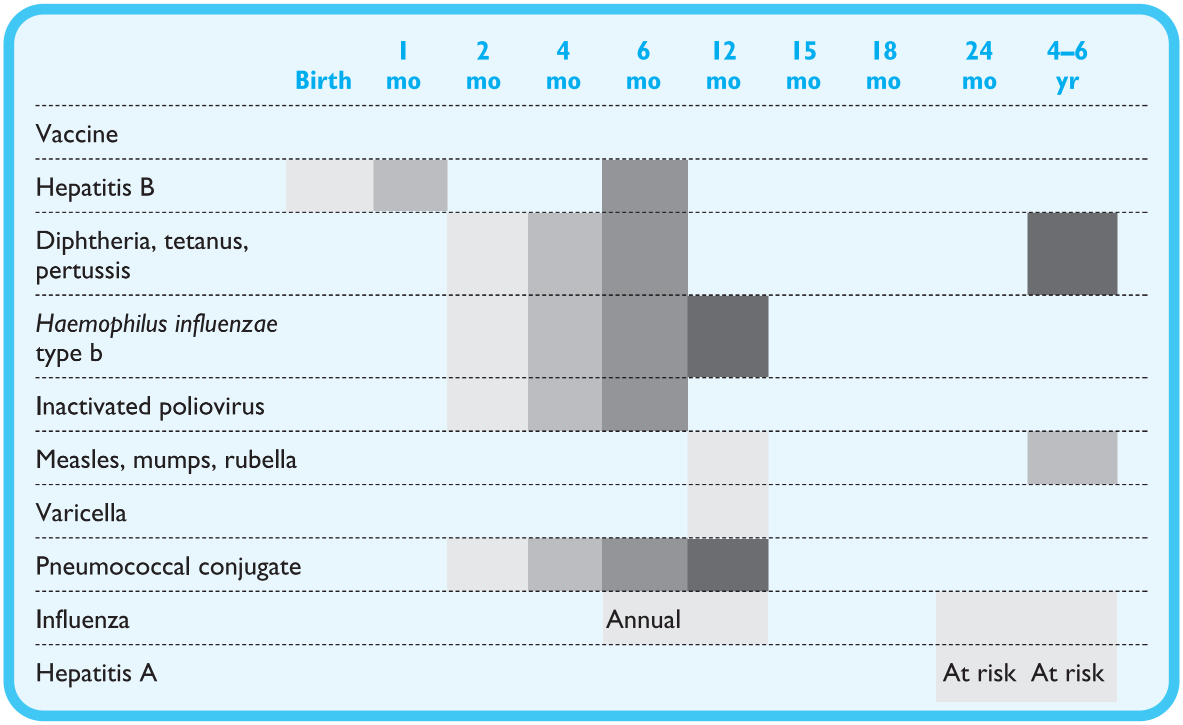

Vaccine	Birth	1 mo	2 mo	4 mo	6 mo	12 mo	15 mo	18 mo	24 mo	4–6 yr
Hepatitis B	░	▒			▓					
Diphtheria, tetanus, pertussis			░	▒	▓					█
Haemophilus influenzae type b			░	▒	▓	█				
Inactivated poliovirus			░	▒	▓					
Measles, mumps, rubella						░				▒
Varicella						░				
Pneumococcal conjugate			░	▒	▓	█				
Influenza					░ Annual	░			░	░
Hepatitis A									░ At risk	░ At risk

[a] Adapted from www.cdc.gov/nip/recs/child-schedule.htm#printable. The different shades denote the different number of immunizations (eg, children should receive three immunizations with the hepatitis B vaccine).

   1. Only a few vaccines (eg, rabies, hepatitis) are **therapeutic vaccines** that can be used to treat an infected patient.
   2. Rabies is a slowly progressing infection that does not outpace the immune response.

C. All existing vaccines are directed at **epitopes**, not microbial patterns, and induce adaptive immunity.

D. We have very few vaccines against complex microbes (eg, parasites).

E. Most of our best vaccines induce neutralizing antibodies.

F. Routine **pediatric vaccines** (Table 14–4) are designed to prevent life-threatening infectious diseases of childhood.
   1. These diseases typically outpace the ability of the neonate to mount an adequate protective immune response.
   2. Immunization is also valuable when the dose of a microbial toxin (eg, tetanus toxin) sufficient to cause disease is far less than the immunizing dose.
   3. Some vaccines (eg, *H influenzae* type b conjugate vaccine) are designed to elicit opsonic IgG antibodies that would not normally be made during an infection in children.

G. Practical issues dictate the design and use of vaccines.
   1. Live attenuated vaccines can cause disease in an immunodeficient individual and should be avoided.
   2. Repeated immunizations are required to attain high-titered, long-lasting protective immunity or to compensate for antigenic shift.
   3. Multiple immunizations are necessary to induce Ig isotype switching (eg, IgG antibody formation).

COMMON COLDS IN CHILDREN

CLINICAL CORRELATION

- *Children normally contract five or six colds (upper respiratory tract infections) each year.*
- *This frequency is increased by exposure to day care centers and cigarette smoke, allergies and asthma, cystic fibrosis, congenital immune defects, and ciliary dyskinesia.*
- *Several hundred different viruses can cause the common cold, including the rhinoviruses, influenza viruses, adenoviruses, and respiratory syncytial virus.*
- *The development of a vaccine for this disease has been hampered by the wide variety of antigenically distinct agents responsible for these infections.*

   4. The design of a vaccine should theoretically reflect the principal protective response desired in the host (Table 14–5).
      a. Immunization against intracellular pathogens requires a live vaccine that duplicates the intracellular infection.
      b. **Live attenuated vaccines** are useful when inducing a CTL response.
      c. **Polyvalent vaccines** (eg, *Neisseria* species) are useful when multiple serotypes of a pathogen cause disease.
      d. **Subunit vaccines** (eg, hepatitis B) are effective when inducing immunity to surface antigens or toxins.
      e. **Conjugate vaccines** (eg, *H influenzae* type b) promote the production of IgG antibodies to T-independent epitopes.
      f. **Mucosal immunization** (eg, influenza nasal vaccine) can protect against respiratory or gastrointestinal pathogens.

Table 14–5. Types of vaccines.

Type of Vaccine	Examples
Live attenuated	Oral polio, varicella; mumps, measles, rubella, bacillus Calmette–Gúerin (BCG)
Killed or inactivated	Inactivated polio
Subunit	Diphtheria toxoid, tetanus toxoid
Recombinant subunit	Hepatitis B
Conjugate	*Haemophilus influenzae* type b, *Streptococcus pneumoniae*
Polyvalent	*S pneumoniae*

NOVEL VACCINE STRATEGIES

CLINICAL CORRELATION

- *Molecular biological techniques provide the means for isolating genes from microbial pathogens and expressing them in relatively safe viral vectors.*
- *This permits the construction of **polyvalent vaccines** carrying genes for a large number of microbial antigens.*
- ***Adjuvant** activity can be included by also expressing genes for human cytokines (eg, switch cytokines or Th1 cell-inducing cytokines) and costimulatory molecules.*
- ***DNA vaccines** are another approach that could be used to induce T cell-mediated immunity by directing microbial protein expression associated with host MHC.*

CLINICAL PROBLEMS

A 15-month-old child is seen by his pediatrician for a cough, congestion, and fever of 39°C. He has a rapid pulse, appears ill, and is unresponsive. His white blood cell count is elevated. His mother seems unsure of his immunization history, but recalls only two visits to the county health clinic for immunizations. The pediatrician prescribes the antibiotic ampicillin, and the child's condition improves substantially over the next few days.

1. A routine pediatric vaccine for which of the following pathogens might have prevented this condition?

A. Rabies

B. Poliovirus

C. *Legionella pneumophila*

D. *Haemophilus influenzae* type b

E. Hepatitis B

Harry is aware that he has a congenital deficiency of the C6 complement peptide.

2. For which of the following infectious agents is Harry particularly susceptible?
   - A. *Streptococcus pneumoniae*
   - B. *Neisseria meningitidis*
   - C. *Clostridium tetani*
   - D. *Mycobacterium tuberculosis*
   - E. *Clostridium difficile*

Donald has a defect in his IFN-γ receptor resulting from an inherited mutation in one of the receptor-encoding genes.

3. Which of the following pathogens would most likely cause recurrent infections in this child?
   - A. *Pseudomonas aeruginosa*
   - B. *Streptococcus pyogenes*
   - C. *Staphylococcus aureus*
   - D. *Legionella pneumophila*
   - E. *Neisseria meningitidis*

A 1-day-old child born of an uncomplicated full-term pregnancy presents with jaundice. The mother is Rh^+, and the newborn is not anemic. Suspecting a congenital virus infection, viral serology on a sample of the child's serum is ordered. High-titered IgG antibodies to cytomegalovirus are found.

4. What can be concluded from these clinical and laboratory findings?
   - A. The patient has a congenital cytomegalovirus (CMV) infection.
   - B. The infant suffers from Rh disease of the newborn.
   - C. The child should be immunized immediately to limit CMV infection.
   - D. The infant should be given broad-spectrum antibiotics.
   - E. The positive CMV serology reflects the immune status of the mother.

Milind is a 56-year-old immigrant from India who is brought to the emergency room with a cough that produces blood-containing sputum. He has a positive skin test with tuberculin purified protein derivative (PPD), and his chest x-ray shows lesions typical of mycobacterial pneumonia. He indicates that he was immunized with bacillus Calmette–Gúerin (BCG) as a child.

5. What is the nature of this vaccine?
   - A. Subunit vaccine
   - B. Conjugate vaccine
   - C. Live attenuated vaccine
   - D. Recombinant vaccine
   - E. Polyvalent vaccine

ANSWERS

1. The correct answer is D. *H influenzae* is a gram-negative bacterium that can cause mild nasopharyngeal infections. In young children, these conditions can progress to meningitis, and vaccination has proven effective at reducing morbidity and mortality. Multiple immunizations in children are recommended by 15 months of age (Table 14–4).
2. The correct answer is B. *Neisseria* species are among the most susceptible to complement-mediated lysis by the C5b-9 membrane attack complex due to the structure of their outer membranes. The other organisms listed do not have classic outer membranes.
3. The correct answer is D. Only *Legionella pneumophila* is an intracellular pathogen. Defective IFN-γ signaling would impair Th1 cellular immunity, especially the activation of macrophages. This bacterium establishes an intracellular infection within resting macrophages and is killed when macrophages become activated.
4. The correct answer is E. High-titered antibodies to CMV in a newborn's serum most likely reflect IgG that has crossed the placenta from the mother. Vaccinating the child at this point would not limit the infection, and there currently is no CMV vaccine.
5. The correct answer is C. The BCG strain is a live bovine mycobacterium that induces cross-protective immunity to other *Mycobacterium* species in humans. Its effectiveness derives from its ability to establish life-long Th1-type cellular immunity against mycobacterial antigens, but protection is variable among vaccines.

CHAPTER 15

IMMUNE DEFICIENCY STATES

I. Primary immune deficiency diseases in human beings are generally rare, inherited conditions affecting a defined immune component (Figure 15–1 and Table 15–1).

- **A.** The term **congenital** is also used to describe these diseases, because they are inherited and become apparent early in life.
- **B.** **Recurrent infections** are the most common clinical presentation.
  - **1.** The nature of the infections in an immunocompromised patient can provide important clues for establishing a diagnosis.
    - **a.** These infections are often with **opportunistic pathogens**, which do not pose problems for healthy individuals.
      - ***(1)*** Opportunistic agents generally show low virulence.
      - ***(2)*** Many opportunistic microbes are normal flora.
    - **b.** The infections may resist aggressive antimicrobial therapy or may not **resolve** (Chapter 14) after the completion of the therapy.
    - **c.** The infection may **progress more rapidly** than it would in an immunocompetent individual of the same age.
    - **d.** The course of the infections may be unusual in **severity** or **organ involvement**.
    - **e.** The infections may occur much **earlier in life** (eg, during the first week) than would be seen in a normal child.
    - **f.** The infections may show **delayed onset** in infancy (eg, after 6 months of life), reflecting the normal decline of maternal antibody acquired during pregnancy.
    - **g.** Similar infections may have occurred **within a family**, especially among male infants.
    - **h.** The infections may follow shortly **after immunization** with a live vaccine.
  - **2.** In adults, recurrent infections more often indicate an **acquired** or **secondary immune deficiency**.
    - **a.** A history of **high risk behavior for human immunodeficiency virus (HIV)-1 infection** (intravenous drug abuse, unprotected sex with an infected partner) supports a diagnosis of an acquired immune deficiency.
    - **b.** **Chemotherapy** for cancer or **occupational exposure** to ionizing radiation increases the risk of immunodeficiency.
    - **c.** Patients who are receiving **immunosuppressive therapy** for autoimmune disease or following transplantation are at risk for recurrent infections.

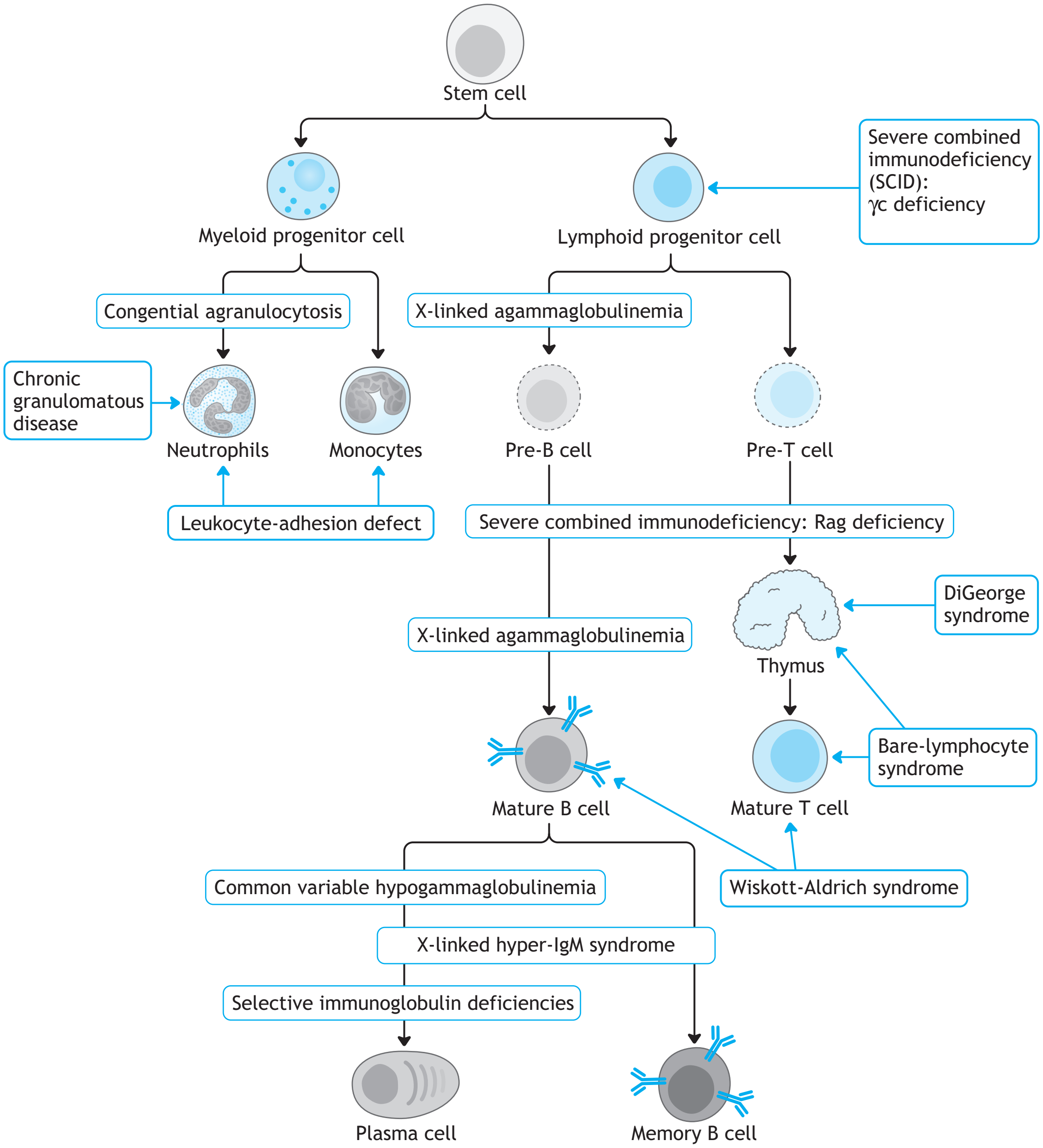
Stem cell
Myeloid progenitor cell
Lymphoid progenitor cell
Severe combined immunodeficiency (SCID): γc deficiency
Congential agranulocytosis
X-linked agammaglobulinemia
Chronic granulomatous disease
Neutrophils
Monocytes
Pre-B cell
Pre-T cell
Leukocyte-adhesion defect
Severe combined immunodeficiency: Rag deficiency
DiGeorge syndrome
Thymus
X-linked agammaglobulinemia
Bare-lymphocyte syndrome
Mature B cell
Mature T cell
Common variable hypogammaglobulinemia
Wiskott-Aldrich syndrome
X-linked hyper-IgM syndrome
Selective immunoglobulin deficiencies
Plasma cell
Memory B cell

Figure 15–1. Primary immune deficiencies.

Table 15–1. Examples of selective primary T and B cell deficiencies.[a]

Category of Deficiency	Examples	Molecular or Cellular Defect	Clinical Presentation	Therapies
Antibody or humoral immunity	XID	Btk kinase mutation	Infections with extracellular bacteria commencing at 6–9 months of age	Antibiotics, IVIG
	Selective IgA deficiency	Varied	Sinopulmonary infections and diarrhea; increased incidence of atopic allergies	Antibiotics
	Hyper-IgM syndrome	CD154 or AID mutations	Elevated IgM; little or no IgG or IgA; absence of Ig class switching; bacterial infections	Antibiotics, IVIG
T cell or cellular immunity	DiGeorge syndrome	Thymic epithelium	Parathyroid deficiency; heart defects; early onset of fungal and viral infections; lymphopenia	Antimicrobial agents; thymic transplant
	BLS-II	Transcription of HLA class II genes	Absence of HLA-D expression; opportunistic viral and fungal infections; $CD4^+$ T cell lymphopenia	Antimicrobial agents; stem cell transplant
	Deficiency of TCR signaling	ZAP-70 mutation	Recurrent viral and fungal infections; depressed cellular immunity	Antimicrobial agents; stem cell transplant

[a] XID, X-linked immunodeficiency; Btk, Bruton's tyrosine kinase; Ig, immunoglobulin; IVIG, intravenous immunoglobulin; AID, activation-induced cytidine deaminase; BLS-II, bare lymphocyte syndrome type II; TCR, T cell receptor; ZAP-70, ζ-associated protein-70.

3. Conversely, the successful resolution of an infection aids in establishing a differential diagnosis of immune deficiency diseases.
  - **a.** Children who resolve a viral infection (eg, chickenpox) in a normal manner are not likely to be deficient in T cells.
  - **b.** Patients who do not have a history of recurrent bacterial infections probably have normal neutrophil function.

4. The nature of the infectious agent can suggest the type of selective immune deficiency.
   - a. Patients with primary antibody, phagocytic cell, or complement deficiencies present with recurrent infections involving **extracellular bacteria** and certain **fungi**.
   - b. Patients with primary T cell or combined [T cell plus antibody and/or natural killer (NK) cell] immune deficiencies present with recurrent infections involving **viruses, intracellular bacteria, fungi,** and certain **parasites**.

C. A differential diagnosis is essential for determining appropriate therapy.
   1. Antibody, phagocytic cell, and complement deficiencies present with similar types of infections, but require distinct therapies.
   2. Similarly, deficiencies in recombination activating gene 1 (Rag1), transporter associated with antigen processing (TAP)-1, and thymus development present with similar opportunistic virus infections.

II. Antibody deficiencies are the most common immune deficiency state in infants and children, accounting for nearly half of all conditions (Table 15–2).

A. The prevalence of these diseases ranges from a few reported cases to 1 case per 350 individuals.

B. Antibody deficiencies range from generalized **agammaglobulinemia** to selective immunoglobulin (Ig) (sub)class deficiencies (eg, IgG_2 deficiency).

C. Some patients with normal serum levels of Igs fail to make antibodies to important microbial antigens.

D. The most frequent antibody deficiencies represent a failure to synthesize or secrete Ig, rather than an absence of B cells.
   1. **Selective IgA deficiency** is the most common primary immune deficiency with a frequency of 1 in 350.
      - a. In some patients, IgA is synthesized but not secreted.

Table 15–2. Distinguishing features of primary antibody deficiencies.[a]

Disease	Distinguishing Features
Selective IgA deficiency	Most common primary immune deficiency; many affected individuals are asymptomatic; IVIG contraindicated
Selective IgG_2 deficiency	Impaired antibody responses to bacterial polysaccharide antigens; recurrent infections with encapsulated bacteria
Transient hypogammaglobulinemia of infancy	Delayed onset of IgM and IgG production; IVIG contraindicated; recurrent sinopulmonary infections
XLA	Absence of peripheral $CD19^+$ B cells; delayed onset to 6 months of age; X-linked; pyogenic bacterial infections

[a] Ig, immunoglobulin; IVIG, intravenous immunoglobulin; XLA, X-linked agammaglobulinemia.

    - **b.** Nearly half of IgA-deficient patients are asymptomatic due to the translocation of IgM across the mucosal epithelium.
  - **2. Selective IgG_2 deficiency** is a common form of IgG deficiency.
    - **a.** These patients do not respond to polysaccharide vaccines.
    - **b.** Infections with encapsulated bacteria are common.
  - **3. Transient hypogammaglobulinemia of infancy** is a condition that can lead to infections toward the end of the first year of life.
    - **a.** These infants have normal numbers of mature B cells.
    - **b.** They show delayed IgG secretion that becomes apparent as maternal IgG levels subside (ie, at 6–8 months of age).
    - **c.** Otitis media, pneumonia, and diarrhea are common.
- **E.** B cells are absent in only a few antibody deficiency diseases.
  - **1. X-linked agammaglobulinemia (XLA)** is a mutation in the gene for Bruton's tyrosine kinase (**Btk**).
    - **a.** B cell development is blocked at the pre-B cell stage.
    - **b.** XLA patients have a peripheral phenotype of $T^+B^-NK^+$.
    - **c.** Normal levels of serum IgG are present at birth and wane over time.
    - **d.** XLA patients suffer from recurrent infections with pyogenic (pus-forming) extracellular bacteria.
  - **2.** A similar phenotype is seen in patients with **Ig H chain locus deletions** that prevent the expression of a functional μ chain.
  - **3. Igα deficiency** blocks B cell development at the pre-B stage.
  - **4. κ-chain deficiency** blocks the differentiation of κ-bearing B cells.
- **F.** Most antibody deficiencies can be diagnosed using simple laboratory tests.
  - **1.** Except for certain rare conditions (eg, XLA), the measurement of B cell numbers is not essential to establishing a diagnosis.
  - **2.** Ig concentrations can be measured by nephelometry (Chapter 5).
  - **3.** Isohemagglutinin titers reflect the ability to produce antibody.
  - **4.** Responses to the tetanus toxoid and pneumococcal polysaccharide vaccines provide information on memory and naive B cell function.
  - **5.** A differential diagnosis requires the elimination of phagocytic cell and complement deficiencies.
- **G.** The treatment of B cell deficiencies addresses both infections and the underlying hypogammaglobulinemia.
  - **1. Antibiotics** are given to treat bacterial infections.
  - **2. Intravenous immunoglobulins (IVIG)** restore antibody levels.
    - **a.** Because the half-life of IgG is approximately 3 weeks, IVIG must be given frequently.
    - **b.** IVIG is prepared from the pooled sera of immune donors.
    - **c.** IVIG therapy is not recommended for children with transient hypogammaglobulinemia of infancy or patients with selective IgA deficiency.
  - **3.** Because IVIG is relatively safe and effective, stem cell transplantation is rarely performed to correct B cell deficiencies.

IVIG AND SELECTIVE IGA DEFICIENCY

CLINICAL CORRELATION

- *Selective IgA deficiency is an important exception to the general use of IVIG to treat antibody deficiencies.*
- *The use of IVIG in these patients carries a risk of anaphylactic or anaphylactoid shock.*

- *This is thought to result from the tendency of these patients to produce IgE and IgG antibodies to IgA that is present in the IVIG preparations.*
- *Such antibodies mediate type I and type III hypersensitivity responses to IVIG.*

III. Patients with primary T cell ($T^-B^+NK^+$) and combined ($T^-B^-NK^+$, $T^-B^+NK^-$, or $T^-B^-NK^-$) immune deficiencies acquire infections in their first few weeks of life.

A. These infections are **pediatric emergencies**.
  1. The infectious agents include **viruses, fungi,** and other **intracellular pathogens**.
  2. Many of the organisms that infect T cell-deficient patients (eg, *Cryptosporidium, Pneumocystis, Leishmania* species) show low virulence and only rarely cause disease in the general population.
  3. The infections in T cell-deficient patients are often disseminated and rapidly progressing.

GRAFT-VERSUS-HOST DISEASE (GVHD) IN PREGNANCY

CLINICAL CORRELATION

- ***GVHD** is typically seen only in stem cell transplantation (Chapter 17).*
- *T cells of the donor recognize and attack the recipient's tissues.*
- *Because pregnancy provides an opportunity for maternal T cells to recognize fetal HLA antigens inherited from the father, GVHD can also occur in this setting.*
- *Because the fetal immune system is sufficiently mature to destroy maternal T cells that cross the placenta, the disease is rarely seen in normal pregnancies.*
- *However, when fetuses have congenital T cell deficiencies, GVHD can occur.*
- *GVHD also occurs when fresh whole blood is transfused to a T cell-deficient child.*

B. Selective T cell deficiencies are due to blocks in thymocyte differentiation and/or receptor-initiated cell signaling (Table 15–3).
  1. **DiGeorge syndrome** is a rare thymic hypoplasia due to dysmorphogenesis of the third and fourth pharyngeal pouches.
    a. Embryonic abnormalities are also seen in the heart, major vessels, parathyroid, and facial structures.
    b. Antibody deficiency results from the absence of Th cells.
    c. T cell lymphopenia and opportunisitic viral and fungal (eg, *Pneumocystis*) infections result.
  2. Defects in **T cell receptor (TCR)-initiated signaling** (Chapter 6) cause deficiencies in cell-mediated immunity.
    a. Loss-of-function mutations exist in **CD3ζ, Lck,** and **ζ-associated protein-70 (ZAP-70)**.
    b. Defective positive selection in the thymus causes peripheral lymphopenia with a **$T^-B^+NK^+$** phenotype.
  3. **Bare lymphocyte syndrome type I (BLS-I)** is a selective deficiency of $CD8^+$ T cells.
    a. This condition is caused by mutations in the **TAP** proteins.
    b. **Major histocompatibility complex (MHC) class I** expression and positive selection in the thymus are impaired.
    c. The peripheral lymphocyte phenotype is **$CD4^+CD8^-B^+NK^+$.**

Table 15–3. Distinguishing features of primary T cell deficiencies.[a]

Disease	Distinguishing Features
DiGeorge syndrome	Lymphopenia with hypoparathyroidism, cardiac and facial defects; corrected by thymic transplant
ZAP-70 or CD3ζ deficiency	Absence of both $CD4^+$ and $CD8^+$ T cell function with normal numbers of B cells and NK cells
BLS type I	Absence of HLA class I expression and decreased numbers of $CD8^+$ T cells
BLS type II	Absence of HLA class II expression and decreased numbers of $CD4^+$ T cells
Th1 deficiency	Decreased IL-12 and IFN-γ signaling; mycobacterial pneumonia
Hyper-IgM syndrome type I	Absence of CD154; elevated IgM antibody responses; decreased IgG and IgA
AID deficiency	Elevated IgM antibody responses; normal expression of CD154 and CD40

[a] ZAP-70, ζ-associated protein-70; NK, natural killer; BLS, bare lymphocyte syndrome; IL, interleukin; IFN, interferon; Ig, immunoglobulin; AID, activation-induced cytidine deaminase.

4. **Bare lymphocyte syndrome type II (BLS-II)** results in a $\mathbf{CD4^-CD8^+B^+NK^+}$ phenotype.
   a. These patients have mutations in transcription factors, such as **CIITA**, that coordinately regulate MHC class II gene expression.
   b. Positive selection of $CD4^+$ thymocytes is impaired.
   c. The activation of peripheral $CD4^+$ T cells is blocked.
5. A family of related **Th 1 cell deficiencies** is caused by deficient cytokine signaling.
   a. **Interleukin (IL)-12** and **IL-12 receptor** defects block Th1 cell differentiation and activation.
   b. **Interferon (IFN)-γ receptor deficiency** blocks macrophage activation.
   c. Recurrent infections with intracellular bacteria (eg, *Mycobacterium* and *Salmonella* species) result.
6. The absence of the coreceptor CD40 or its ligand, CD154, causes **hyper-IgM syndrome type 1**.
   a. This leads to a deficiency of the Th2 signals necessary for Ig class switching in B cells (Chapter 6).
   b. Patients have elevated levels of IgM antibodies, but no IgG, IgA, or IgE.
7. **Hyper-IgM syndrome type 2** results from a mutation in the gene for **activation-induced cytidine deaminase (AID)** (Chapter 9).

   a. T helper cell signaling for Ig class switching is absent.
   b. Somatic hypermutation and affinity maturation in B cells are also impaired.

C. **Severe combined immune deficiency (SCID)** is caused by a number of different genetic mutations (Table 15–4).
   1. The X-linked recessive form of SCID is most prevalent.
      a. It is caused by an absence of the **common γ chain (γc)** of the receptors for IL-2, IL-4, IL-7, IL-9, and IL-15.
      b. Abnormal development of T cells and NK cells results in a **$T^-B^+NK^-$** phenotype.
   2. The same phenotype (**$T^-B^+NK^-$**) is seen in deficiencies of **Janus kinase 3 (Jak3)**, which mediates signaling from these receptors.
   3. Mutations in the gene coding for the enzyme **adenosine deaminase (ADA)** cause SCID.
      a. ADA catalyzes the breakdown of intracellular adenosine and deoxyadenosine.
      b. Excess adenosine inhibits methyltransferase reactions in T cells, B cells, and NK cells resulting in a **$T^-B^-NK^-$** phenotype.
   4. Mutations in the **Rag1** and **recombination activating gene 2 (Rag2)** genes block T and B cell development by preventing TCR and B cell receptor (BCR) gene rearrangements.
      a. Lymphocyte differentiation is blocked at the pre-T and pre-B cell stages.
      b. The lymphocyte phenotype in this condition is **$T^-B^-NK^+$.**
   5. **Wiscott–Aldrich syndrome (WAS)** is a condition characterized by thrombocytopenia and immune deficiency.
   6. **Perforin and granzyme deficiencies** manifest as decreased T cell and NK cell cytotoxic activity.

Table 15–4. Distinguishing features of primary SCID.[a]

Gene Mutation	Overall Cellular Phenotype	Distinguishing Features
γc	$T^-B^+NK^-$	Most frequent cause; impaired signaling through multiple cytokine receptors; normal B cell numbers
Jak3	$T^-B^+NK^-$	Rare, but otherwise indistinguishable from γc deficiency
ADA	$T^-B^-NK^-$	Decline in the numbers of all lymphocyte subsets over time
Rag1, Rag2	$T^-B^-NK^+$	Absence of peripheral T and B cells from birth; normal NK cells

[a] Jak3, Janus kinase 3; ADA, adenosine deaminase; Rag1, Rag2, recombination activating genes 1 and 2; NK, natural killer.

D. The **diagnosis** of T cell and combined immune deficiencies depends on recognizing the typical infections that characterize these conditions and measuring lymphocyte numbers and functions.

1. Infections begin in the first few months of life.
2. Lymphopenia is common.
3. Measuring anti-HIV antibodies or viral load is essential when a T cell deficiency is suspected.
4. Measuring CD3, CD4, CD8, CD19, and CD154 expression by flow cytometry can help detect lymphocyte subset deficiencies.
5. T cell proliferation in response to polyclonal activators (the plant mitogen phytohemagglutinin) can be measured in vitro.
6. Delayed-type hypersensitivity (DTH) skin tests (eg, intradermal or patch testing) are used to evaluate $CD4^+$ Th1 cell responses to common microbes.

E. **Therapy** includes antiviral and antifungal agents and hematopoietic stem cell replacement.

1. **Thymus transplantation** is effective in treating DiGeorge syndrome.
2. **Stem cell transplantation** (Chapter 17) is effective for SCID and selective T cell deficiency patients.
3. **Gene therapy** is being tested for some of these immune deficiencies (eg, ADA deficiency).
4. Other therapies are **contraindicated**.
    - a. T cell-deficient patients should not be immunized with **live viral vaccines**.
    - b. All blood products (whole blood, platelets, and plasma) should be irradiated prior to transfusion.

GENE THERAPY FOR IMMUNE DEFICIENCIES

CLINICAL CORRELATION

- *The identification of specific genetic mutations that cause immunodeficiency diseases provides the opportunity to correct these conditions by gene therapy.*
- *The most promising approach appears to be "repairing" the defective gene within the patient's own hematopoietic stem cells by retroviral transfer followed by reinfusion.*
- *ADA deficiency was the first human disease treated by gene therapy in the 1990s.*
- *Other immune deficiencies that have been considered include γc deficiency, leukocyte adhesion defect (LAD; CD18 deficiency), and chronic granulomatous disease (respiratory burst deficiency).*

IV. Primary phagocytic cell deficiencies result from decreased cell number, abnormal cell adhesion and migration, defective intracellular granules, or impaired intracellular killing mechanisms (Table 15–5).

A. **Cyclic neutropenia** is caused by a mutation in the **neutrophil elastase** gene.

1. Neutropenia (< 200 cells/μL) occurs periodically.
2. Treatment with recombinant **colony-stimulating factor-granulocyte (CSF-G)** has proven effective.

B. **LAD** is characterized by poor leukocyte adhesion to the vascular endothelium, delayed wound healing, and defective complement-dependent opsonophagocytosis.

1. **LAD type 1** is a mutation in the gene for **CD18**, a β2 integrin family member (Chapter 8).
    - a. Patients show chronic leukocytosis involving neutrophils and fail to phagocytize iC3b-coated particles.

Table 15–5. Distinguishing features of phagocytic cell deficiencies.[a]

Disease	Distinguishing Features
Cyclic neutropenia	Oscillating neutropenia; responsiveness to CSF-G
LAD type 1	Chronic leukocytosis; absence of CD18-containing receptors; limited neutrophil extravasation and pus formation; impaired iC3b-mediated phagocytosis
LAD type 2	Leukocytosis; infections; normal CD18
LAD type 3	Decreased selectin expression
LAD type 4	Impaired CD18 initiated signaling
Chediak–Higashi syndrome	Formation of giant fused granules in neutrophils and NK cells
CGD	Absence of superoxide production; infections with catalase-positive microbes; granulomas
Job's syndrome	Decreased IFN-γ synthesis; IgE levels, skin abscesses without erythema or tenderness
Th1 pathway defects	Impaired macrophage activation

[a] CSF-G, colony-stimulating factor-granulocyte; IFN, interferon; LAD, leukocyte adhesion defect; NK, natural killer; CGD, chronic granulomatous disease.

   b. Diagnosis is aided by demonstrating the absence of CD11 or CD18 on blood leukocytes.
2. **LAD type 2** is a defect in fucose metabolism that impairs selectin function.
3. **LAD type 3** is a defect in selectin expression.
4. **LAD type 4** is caused by a mutation in the rac signaling molecule activated by CD18.

C. **Chediak–Higashi syndrome (CHS)** results from the excessive fusion of granules in a number of granulated cells, including leukocytes.
   1. The defect is caused by a mutation in a regulator of lysosomal trafficking.
   2. CHS cells show poor chemotaxis and impaired NK cell function.

D. **Neutrophil-specific granule deficiencies**, including the absence of α and β defensins, result in impaired intracellular killing of microbes.

IMMUNE DEFICIENCY IN JOB'S SYNDROME

CLINICAL CORRELATION

- ***Job's syndrome*** *(hyper-IgE syndrome; Chapter 13) is characterized by highly elevated levels of serum IgE, eczema, and sinopulmonary and skin infections.*
- *Given the elevated IgE levels, atopic dermatitis is a common misdiagnosis.*
- *Infections can cause deep skin and lung abscesses.*

- *An unusual characteristic of the skin lesions is the absence of erythema, tenderness, and warmth, similar to the appearance of the skin boils of the biblical character Job.*
- *A marked decrease of IFN-γ production by these patients may explain both the elevated IgE levels and the impaired phagocyte responses to infection.*

E. **Chronic granulomatous disease (CGD)** is a heterogeneous condition caused by a defective respiratory burst oxidase.
  1. Mutations in any of four subunit genes have been reported.
  2. Diagnosis is based on the nitroblue tetrazolium (NBT) dye reduction test for oxidant production.
  3. CGD is characterized by bacterial and fungal infections, especially *Staphylococcus* and *Aspergillus* species.
  4. **Granulomas** form in the lungs, liver, and spleen when microbes that cannot be killed are "walled off."
  5. Prophylactic antibiotics, recombinant IFN-γ, and granulocyte transfusions are beneficial.
  6. Stem cell transplantation is potentially curative.

F. **Myeloperoxidase (MPO) deficiency** causes a block in the production of hypohalous acids.

G. Aggressive **treatment** with antibiotics is necessary to control infections in phagocytic cell deficiencies.
  1. Some conditions benefit from prophylactic antimicrobial agents.
  2. Recombinant cytokines are of value in treating cyclic neutropenia and CGD.
  3. Leukocyte transfusions during acute infections are beneficial.
  4. Stem cell transplantation is potentially curative in CGD, Chediak–Higashi syndrome, and LAD.

ANTIBIOTICS AND PHAGOCYTE DEFICIENCIES

- *Neutrophil deficiencies tend to manifest as gingivitis and periodontal diseases, cutaneous infections and abscesses, pneumonias, and/or granulomas caused by multiple infectious agents.*
- *While IFN-γ deficiency affects macrophage function and results in infections primarily with intracellular bacterial pathogens, diverse types of pathogens cause infections in Chediak–Higashi syndrome or cyclic neutropenia.*
- *Prophylactic antimicrobial therapy is less effective when infections are caused by diverse microbial species.*

V. Inherited deficiencies in the **complement system** cause both infectious and noninfectious diseases (Table 15–6).

A. Rare deficiencies in all of the individual components have been described.
  1. A deficiency of **mannose-binding protein (MBP)** is the most common single component defect.
  2. **C2 deficiency** is the most common deficiency affecting the classical pathway.

B. Deficiencies in the classical, lectin, or alternative pathways can increase the susceptibility to bacterial and some viral infections.
  1. Early pathway deficiencies result in recurrent pyogenic infections.
  2. Defects in the terminal components C5–C8 lead to disseminated *Neisseria* infections.

Table 15–6. Primary deficiencies in the complement system.

Type of Deficiency	Components Involved	Clinical Characteristics
Classical pathway	C1q,r,s, C4, C2	Lupus nephritis Bacterial, viral, and fungal infections
Alternative pathway	P, B, D	*Neisseria* infections Pyogenic infections
Lectin pathway	MBL	Diverse infections
Common components	C3	Pyogenic infections
Terminal components	C5–C8	Disseminated *Neisseria* infections
Inhibitors	C1 Inh, H, I, C4b-BP	Angioedema, nephritis, infections
Cell surface inhibitors	DAF, CD59	Paroxysmal nocturnal hemoglobinuria
Receptors	CR3, CR4	Leukocyte adhesion, pyogenic infections, delayed wound healing

C. Defects in the expression of the **complement regulatory components** cause both infectious and noninfectious diseases.

1. **Factor I deficiency** increases the risk of pyogenic infections, including meningococcal meningitis.
   a. Unrestrained complement activation leads to C3 consumption (Chapter 8).
   b. Exogenous Factor I provides relief in acute disease.
2. **Factor H deficiency** also results in increased complement activation, C3 depletion, and an increased risk of infection.
3. **C1 inhibitor (C1 Inh) deficiency (hereditary angioedema)** and **C4b-BP deficiency** are characterized by episodic cutaneous and submucosal edema.
   a. Both conditions result in unabated activation of the classical complement pathway.
   b. A C2-derived peptide is thought to cause increased vascular permeability and localized edema in hereditary angioedema.
   c. C1 Inh also regulates the production of bradykinin, which can have a similar effect on endothelial permeability.
4. **Paroxysmal nocturnal hemoglobinuria (PNH)** is due to a failure of cells to express glycosylphosphatidylinositol (GPI)-linked membrane proteins, including CD59 and CD55 (decay accelerating factor).
   a. CD59 blocks insertion of C8 and C9 into the membrane.
   b. CD55 inhibits C3 and C5 convertases on the cell surface.
   c. Erythrocytes of PNH patients are highly susceptible to complement-mediated lysis.

D. Complement deficiencies and **autoimmunity** are often associated with one another.
  1. Patients with defects in the early classical pathway components often develop a lupus-like nephritis.
  2. Circulating immune complexes in autoimmunity deplete the early classical components.
  3. Autoantibodies to complement components (eg, anti-C1q) are sometimes produced by patients with systemic lupus erythematosus (SLE) (Chapter 16).
  4. Patients with SLE, Sjögren's syndrome, and autoimmune hemolytic anemias often show decreased expression of **complement receptor 1 (CR1)** on their erythrocytes.

E. The treatment of complement deficiencies focuses primarily on controlling infections and correcting acute symptoms.
  1. Aggressive antibiotic therapy is the first objective for infections.
  2. Plasma infusion or injection of individual components (eg, recombinant C1 Inh) has some clinical value for acute disease.
  3. The various treatments for autoimmunity associated with complement deficiencies are discussed in Chapter 17.

VI. Inherited innate immune defects can cause inflammatory or infectious diseases.

A. An increased susceptibility to infections results from mutations in **interleukin-1 receptor-associated kinase (IRAK)** of the Toll-like receptors.

B. Mutations in **mannose-binding lectin** increase the risk of infection, especially in the presence of other chronic diseases (eg, cystic fibrosis).

C. Deficiencies in the antimicrobial **Nod proteins** (Chapter 1) are thought to contribute to the induction of inflammatory bowel disease.

VII. Acquired or secondary immune deficiency states typically show adult onset.

A. Acquired immune deficiencies can also result from **infections**.
  1. **Acquired immune deficiency syndrome (AIDS)** is the most prevalent immune deficiency in the United States.
     a. CD4$^+$ T cells are depleted by the lymphotrophic virus **human immunodeficiency virus-1 (HIV-1)**.
     b. Th1 cell functions are particularly affected.
     c. Opportunistic infections with intracellular pathogens (Table 15–7) are the primary causes of morbidity and mortality.
  2. Transient neutropenia in children is seen in some viral (eg, measles) and bacterial (eg, tuberculosis) infections.
  3. A number of microbial pathogens (eg, herpesviruses) have developed immune evasion mechanisms that cause selective immunosuppression of the host (Chapter 14).

B. Immune deficiency can arise secondary to **therapy**.
  1. **Splenectomy** impairs the ability to produce antibodies.
     a. Splenectomized patients are at increased risk for sepsis.
     b. Splenectomy decreases IgG antibody responses to polysaccharide vaccines.
  2. **Immunosuppressive drugs** increase the risk of infections.

Table 15–7. Opportunistic pathogens causing infections in AIDS.

Bacteria	Fungi	Viruses	Parasites
Mycobacteria	*Pneumocystis*	Herpes simplex	*Toxoplasma*
Salmonella	*Cryptococcus*	Cytomegalovirus	*Cryptosporidium*
	Candida	Varicella-Zoster	*Leishmania*

    a. Many **chemotherapeutic agents** used to treat malignancies are immunosuppressive.
    b. Drugs that block allograft rejection (Chapter 17) also inhibit immune responses to microbial pathogens.
    c. **Corticosteroids** increase the risk of infection when used at high doses or over long periods of time.
    d. **Disease-modifying antirheumatic drugs (DMARDS)** used for treating chronic autoimmune diseases carry a similar risk of increasing infection rates.
    e. The use of certain **anticytokine reagents** can inhibit protective immunity to specific pathogens [eg, antitumor necrosis factor (TNF)-α antibodies and tuberculosis].

C. **Cancer** is a common cause of secondary immune deficiencies.
  1. Malignancies that grow within the bone suppress hematopoiesis, including lymphopoiesis and granulopoiesis.
  2. Lymphoid malignancies (eg, lymphomas) inhibit normal lymphocyte functions in the peripheral lymphoid tissues.
  3. Cancer cells often produce excessive quantities of cytokines and growth factors that cause immune imbalance.

D. **Protein-losing enteropathies, extensive burns**, and **nephrotic syndrome** result in protein loss, hypogammaglobulinemia, and hypocomplementemia.

E. **Common variable immune deficiency** is a heterogeneous syndrome that shows early or adult onset.
  1. B cell numbers are normal, but plasma cells are infrequent and antibody production is impaired.
  2. T cell numbers and subsets are also typically normal, but cellular immunity can be decreased.
  3. Bacterial agents constitute the greatest threats of infection, which resemble those seen in XLA.

CLINICAL PROBLEMS

A 4-year-old child is referred with the diagnosis of type 1 LAD. She has a history of recurrent bacterial infections, including pneumonias and skin abscesses. Radiography demonstrates granuloma-like lesions in her lungs and liver. However, the data in the following table suggest this condition has not been properly diagnosed.

1. Which of the elements in the table is **inconsistent** with a diagnosis of type 1 LAD?

CD Marker	Patient (%)	Normal Range (%)
CD2	76	75–92
CD3	75	60–85
CD4	52	30–65
CD8	26	20–52
CD11a	65	65–85
CD18	73	70–86
CD19	12	7–17
CD45	52	44–62

A. The CD2 value
B. The CD19 value
C. The CD11a value
D. The CD45 value
E. The CD3 value

A 3-month-old child presents with a history of persistent diarrhea and oral thrush beginning at age 3 weeks and has developed signs and symptoms consistent with *Pneumocystis carinii* pneumonia. Both he and his parents are HIV negative based on serological and polymerase chain reaction (PCR) testing. Complete blood counts demonstrate profound lymphopenia.

2. Which of the following blood cell types should show normal numbers if this condition was caused by a Rag deficiency?

A. $CD3^+$ cells with a CD4 coreceptor
B. $CD19^+$ cells
C. Single-positive thymocytes
D. Memory B cells
E. NK cells

An 8-month-old male child has an apparent antibody deficiency. The diagnosis has been narrowed to two possibilities.

3. Measuring which of the following immune parameters would distinguish hyper-IgM syndrome (type 1) from X-linked agammaglobulinemia?

A. IgG levels in the serum
B. $CD19^+$ cells in the blood
C. IL-2 production by the patient's lymphocytes in vitro

D. NK cell-mediated cytotoxicity in vitro
E. IgA levels in the saliva

A pediatric patient has a history of recurrent infections, and it is hypothesized that she has a defect in the neutrophil-mediated killing of phagocytized bacteria.

4. Which of the following laboratory assays would most directly test this hypothesis?

A. Complement fixation test
B. Quantitative test for serum Igs
C. NBT dye reduction test
D. Flow cytometry for CD18 expression
E. Complete blood count and differential

An adolescent with a history of childhood upper respiratory tract infections and diarrhea continues to have recurring sinusitis, which is attributed to *Pseudomonas aeruginosa*, a gram-negative bacterium. Laboratory tests over the past 5 years have repeatedly indicated that he has normal T cell numbers (based on CD3, CD4, and CD8), normal serum complement levels and activity, and normal serum IgM and IgG levels for his age. Isohemagglutinin titers are normal. Neutrophil numbers and function (chemotaxis, opsonophagocytosis, and respiratory burst) are all normal. His IgG antibody response to tetanus toxoid challenge is also within the normal range. He shows 4+ reactions in delayed hypersensitivity skin tests against several common fungal antigens.

5. What therapy is appropriate for this patient?

A. Antibiotics only
B. Antibiotics and IVIG
C. Antibiotics, IVIG, and stem cell transplantation
D. Stem cell transplantation
E. Antibiotics and IFN-γ

Johnny is an 8-month-old child with recurrent fungal and viral infections. His blood lymphocyte numbers are normal and they bind IL-2 in vitro. It has been determined that his parents are both heterozygous for a mutation in their ZAP-70 genes that can explain Johnny's disease.

6. In addition to antimicrobial agents to treat his infections, what therapy should be considered for Johnny?

A. Thymus transplant
B. IVIG and all of the pediatric vaccines
C. Recombinant IL-2
D. Stem cell transplantation
E. Besides antimicrobial agents, there is no effective therapy for this patient.

ANSWERS

1. The correct answer is C. The diagnosis of LAD type 1 requires demonstrating that the patient lacks CD18 or any of the α chains that pair with the CD18 β chain (ie, CD11a, CD11b, or CD11c). This patient has normal levels of CD11a, which means she does not have LAD type 1.
2. The correct answer is E. Rag-deficient SCID patients show a $T^-B^-NK^+$ overall lymphocyte phenotype.
3. The correct answer is B. In hyper-IgM syndrome, normal numbers of B cells are present in the periphery. However, they do not receive the normal signaling from Th cells due to a defect in the coreceptor ligand CD154. In XLA the absence of the Btk kinase blocks B cell differentiation at the pre-B cell stage, and the patients have no peripheral CD19$^+$ B cells.
4. The correct answer is C. The NBT dye reduction test measures cellular production of oxidants, which are important in the oxygen-dependent killing of microbes by neutrophils (Chapter 1). The other tests could confirm phagocyte dysfunction as well, but would not directly relate to intracellular killing function.
5. The correct answer is A. This patient lacks any clear evidence of an immune deficiency that would explain his recurrent bacterial infections. There is no rationale for stem cell transplantation, and there is no antibody defect that would suggest giving IVIG. Only conventional antibiotic therapy is indicated for the treatment of the infection.
6. The correct answer is D. ZAP-70 deficiency affects TCR-initiated signaling in his T cells. Replacing the defective T cells with normal T cell precursors would theoretically be of value here.

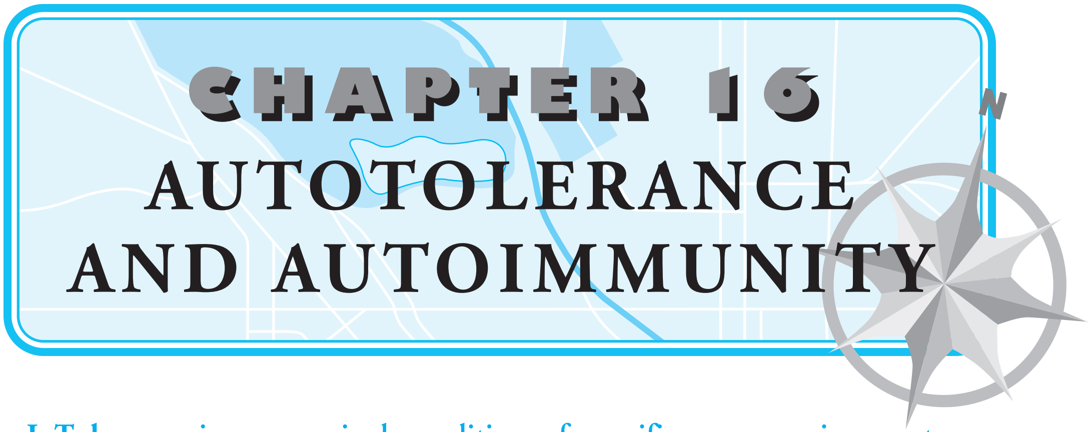

CHAPTER 16

AUTOTOLERANCE AND AUTOIMMUNITY

I. Tolerance is an acquired condition of specific unresponsiveness to an antigen.

A. **Autotolerance** (unresponsiveness to self antigens) is established within B and T lymphocytes as they encounter self antigens during development.
  1. Self tolerance is antigen specific and can be induced either in a primary lymphoid organ or in the periphery.
  2. Autotolerance continues to be established throughout life, even after the thymus atrophies.

B. Three mechanisms contribute to the induction of tolerance.
  1. **Clonal deletion** results in the apoptosis of specific B and T cell clones.
    a. Clonal deletion generally occurs among immature B and T cells through a process called **negative selection** (Chapters 9 and 10).
    b. Deletion requires antigen recognition by immature lymphocytes through their B cell receptor (BCR) or T cell receptor (TCR).
  2. **Clonal anergy** is a form of unresponsiveness in lymphocytes caused by antigen recognition in the absence of essential stimulatory coreceptor signals.
    a. Antigen presentation by antigen-presenting cells (APCs) that lack the B7 ligand for the CD28 coreceptor induces **T cell anergy**.
    b. Antigen presentation by cells that engage the inhibitory cytotoxic T lymphocyte-associated protein 4 (CTLA-4) coreceptor can also induce T cell anergy.
  3. Tolerance, including autotolerance, can also be mediated by active suppressor cells in the periphery.
    a. **Regulatory T cells** (Treg cells) mediate unresponsiveness to specific antigens.
    b. Treg cells are induced in both the thymus and the periphery by autoantigens.
    c. Suppression is generally mediated by the production of inhibitory cytokines, such as interleukin (IL)-10 and transforming growth factor (TGF)-β.

C. Autotolerance is established at two levels.
  1. **Central tolerance** to self antigens is established during the differentiation of B and T cells in the bone marrow and thymus, respectively.
    a. Central tolerance results in the deletion of autoreactive clones (negative selection).

    b. Self-specific B cells can be rescued from clonal deletion by undergoing secondary immunoglobulin (Ig) locus rearrangements (BCR editing.)
  2. **Peripheral tolerance** involves all three mechanisms of tolerance induction—deletion, anergy, and suppression.
    a. The induction of autotolerance in the periphery continues throughout life.
    b. Because quiescent autoreactive B and T cell clones reside in normal individuals, autoimmunity is common.

TRANSPLANTATION TOLERANCE

CLINICAL CORRELATION

- *It is not uncommon for transplant recipients to become tolerant to their foreign organ grafts over time, despite significant genetic disparities with the organ donor.*
- *Pretransplantation lymphoablation appears to favor the induction of tolerance, as does the establishment of donor–recipient chimerism when donor bone marrow cells are provided.*
- *The continual exposure of the recipient to the HLA antigens of the donor may also induce active suppressor cells, including Treg cells.*
- *Blocking CD28/B7 and CD40/CD154 costimulatory signaling may be another mechanism for inducing specific unresponsiveness to allografts.*
- *The ultimate goal of inducing transplantation tolerance is to reduce the dependence on powerful immunosuppressive therapies in the posttransplant period.*

II. Autoimmunity results from the loss of self tolerance.

A. The **molecular mimicry theory of autoimmunity** proposes that autoimmunity occurs when microbial antigens cross-react with self antigens.
  1. Antibodies to the M protein of *Streptococcus pyogenes* that are induced during infection can react with cardiac autoantigens of the sarcolemma and heart valves and cause **rheumatic fever**.
  2. Antibodies to *Treponema pallidum* induced during syphilis can cross-react with human fibronectin and collagen.
  3. Antibodies to the insulin receptor are induced during papilloma virus infections.

B. The **costimulatory theory of autoimmunity** proposes that silent autospecific lymphocytes are activated when tissues inappropriately express stimulatory coreceptor ligands.
  1. Major histocompatibility complex (MHC) class II molecules are expressed by pancreatic β cells in type 1 diabetes mellitus.
  2. Microglial cells of patients with multiple sclerosis express B7.

C. Autoimmunity is induced when cryptic or **sequestered autoantigens** that are normally not seen by the immune system are released to the periphery.
  1. Autoimmune uveitis is thought to develop when ocular antigens released from a damaged eye induce immune damage to the contralateral eye.
  2. Autoantigens of spermatogonia can be recognized as nonself if they are later released from the testes by vasectomy.

D. **Modified self antigens** can induce autoimmune responses.
  1. Autoimmune hemolytic anemias result when drugs bind to erythrocyte surface proteins.
  2. Autoantibodies to the Fc domains of IgG molecules (**rheumatoid factors**) are often produced in **rheumatoid arthritis**.

VASECTOMY AND AUTOIMMUNE ORCHITIS

- *Vasectomy is the most common means by which men elect permanent sterilization for the purpose of contraception.*
- *Vasectomy does not halt spermatogenesis, and sperm can become accessible to the immune system following the procedure.*
- *Sperm-agglutinating antibodies develop in over 50% of vasectomized males.*
- *Epididymitis and orchitis mediated by autoantibodies occur at a rate of 5–6%.*

III. Autoimmunity often exists without overt autoimmune disease.

A. The elderly commonly produce low-affinity autoantibodies that do not cause disease.

B. **Witebsky's postulates** establish a set of criteria for designating a disease as autoimmune in nature.

   1. Autoantibodies or autoreactive T cells must be demonstrated in the patient.
   2. The autoantigen(s) must be identified.
   3. Immunizing a laboratory animal with the autoantigen must induce a comparable autoimmune response and autoimmune disease.
   4. The disease manifestations must be transferable to a naive recipient with autoantibody or autospecific T cells.

C. Several human diseases appear to satisfy Witebsky's postulates.

   1. Myelin-specific T cells of the type found in **multiple sclerosis** can transfer disease to laboratory rats and mice.
   2. The lesions found in **myasthenia gravis** can be duplicated in animals by the transfer of patient serum.
   3. Harrington's experiments with **idiopathic thrombocytopenia purpura** clearly identify this condition as an autoimmune disease.

WILLIAM HARRINGTON AND IDIOPATHIC THROMBOCYTOPENIA PURPURA

- *In 1951, a young hematologist named William Harrington was caring for a group of patients with bleeding disorders and low platelet counts.*
- *Dr. Harrington proposed that their thrombocytopenia was immune based.*
- *After injecting himself with the plasma of one of his patients, he developed severe, but transient thrombocytopenia.*
- *Because his megakaryocytes were spared destruction, he eventually recovered.*
- *This experiment firmly established that autoantibodies can cause human disease.*

D. The diagnosis of autoimmune diseases is aided by laboratory tests.

   1. Some tests detect **tissue-bound autoantibodies**.
      a. Antibodies to glomerular basement membrane antigens in **Goodpasture's disease** show a characteristic linear distribution in the kidney by immunofluorescence.
      b. By contrast, the "lumpy-bumpy" appearance of IgG in the glomerulus is typical of immune complex autoimmune diseases (eg, **systemic lupus erythematosus or SLE**).
   2. Many laboratory tests are designed to detect **circulating autoantibodies**.
      a. **Antinuclear antibodies (ANA)** are detected by immunofluorescence.

    - *(1)* The autoantigens include DNA, histones, and nucleolar proteins.
    - *(2)* These antibodies are seen in SLE, Sjögren's syndrome, and scleroderma.
  - **b.** **Antineutrophil cytoplasmic antigen (ANCA)** is characteristic of certain forms of autoimmune nephritis and vasculitis (eg, Wegener's granulomatosis).
  - **c.** **Rheumatoid factor (RF)** is common to rheumatoid arthritis and a number of other autoimmune diseases.
    - *(1)* RF is an autoantibody to the Fc region of IgG.
    - *(2)* RF is mostly IgM anti-IgG, but IgG and IgA autoantibodies have also been described.
  - **d.** The **Coombs test** detects IgG antibodies on erythrocytes and platelets.
    - *(1)* These antibodies can be specific for endogenous or exogenous (ie, haptenic) cell surface antigens.
    - *(2)* "Coombs-positive" is synonymous with immune-based when used to describe an anemia.
- **3.** Tests that measure the serum concentrations of **complement** can be used to follow the status of autoimmune diseases.
  - **a.** Complement activity is often depressed during acute exacerbations of immune complex autoimmune diseases.
  - **b.** Measuring complement levels can be useful for monitoring the success of therapy.

IV. The spectrum of **autoimmune diseases** ranges from those that are organ specific to those that are essentially systemic in nature.

- **A.** In **organ-specific autoimmune diseases** symptoms reflect the distribution of the target autoantigen(s).
  - **1.** Hashimoto's thyroiditis, pernicious anemia, and Addison's disease are examples of organ-specific autoimmune diseases.
  - **2.** The organ-specific disease manifestations of these conditions (eg, thyroid dysfunction) are important clues to their diagnosis.
- **B.** **Systemic autoimmune diseases** (ie, the **rheumatological disorders**) typically involve the skin, kidneys, muscles, joints, and blood vessels.
  - **1.** Tissue damage is widely distributed.
  - **2.** Symptoms can reflect the deposition of soluble immune complexes, which is an important step in disease pathogenesis.

AUTOIMMUNE POLYENDOCRINE SYNDROME

CLINICAL CORRELATION

- *Autoimmune polyendocrine syndrome (APS) type 1 is a rare condition caused by a mutation in the gene for the transcription factor autoimmune regulator (AIRE).*
- *Patients present with chronic mucocutaneous candidiasis, hypoparathyroidism, and/or adrenal insufficiency.*
- *AIRE is expressed in thymic epithelial cells and may control induction of self tolerance.*
- *Twenty percent of affected individuals develop type 1 diabetes mellitus.*

- **C.** Tissue damage and symptoms in autoimmune diseases can change with time, because autoantibody responses become increasingly polyspecific.
  - **1.** The initial autoantibody is specific for a dominant self epitope.
  - **2.** Autoimmunity broadens with time by **epitope spreading** (Figure 16–1).

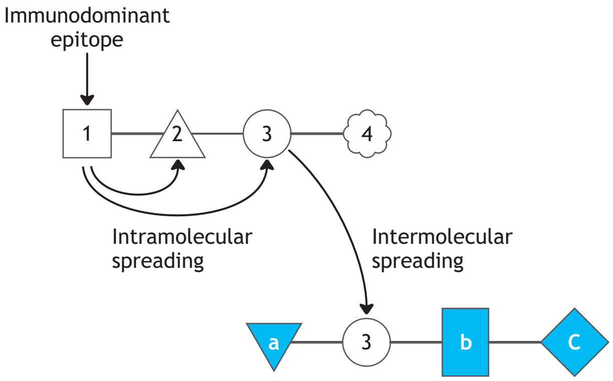

Figure 16–1. Epitope spreading in autoimmunity.

- a. B cells specific for the initial epitope elicit new responses by presenting additional epitopes on the same antigen (**intramolecular spreading**).
- b. B cells specific for the epitopes of the first autoantigen can cross-react with a second autoantigen (**intermolecular spreading**).

V. One of four pathogenic mechanisms underlies the development of most autoimmune diseases (Table 16–1).

A. **Antireceptor autoantibodies** can cause autoimmune disease.
- 1. **Myasthenia gravis** is a neurological condition presenting as muscle weakness and fatigue.
  - a. Manifestations include diplopia, dysphonia, and difficulty swallowing, breathing, and walking.
  - b. Antagonistic antibodies to the **acetylcholine receptor** block neurotransmission at motor end plates.
- 2. **Grave's disease** is mediated by agonistic antibodies to the thyroid-stimulating hormone (TSH) receptor.
  - a. Anti-TSH receptor antibodies are diagnostic.
  - b. Hyperthyroidism and thyroid hyperplasia are common.
  - c. Females are affected seven times more often than males.
  - d. Many patients require thyroid ablation by radiotherapy or surgery.

B. **Cytotoxic immune tissue injury** is the basis for many forms of organ-specific autoimmune disease.
- 1. **Coombs-positive hemolytic anemias** and thrombocytopenias result from the immune lysis and/or clearance of blood cells.
  - a. Complement plays a central role in cell destruction.
  - b. "**Cold reactive antibodies**" are IgM and show high avidity.
    - *(1)* They are more commonly found after certain infections (eg, mycoplasma).
    - *(2)* They typically induce intravascular hemolysis.
  - c. "**Warm reactive antibodies**" are IgG and have a lower avidity.

Table 16–1. Major pathogenic mechanisms underlying autoimmune diseases.

Mechanism of Disease	Examples	Initial Target Autoantigen	Distinguishing Clinical Features
Antireceptor autoantibody	Myasthenia gravis	Acetylcholine receptor	Muscle fatiguability
	Grave's disease	Thyroid-stimulating hormone receptor	Hyperthyroidism; goiter
	Pernicious anemia	Intrinsic factor	Vitamin B_{12} deficiency
	Addison's disease	Adrenal cell antigens	Adrenal insufficiency
Cytotoxic immune injury	Coombs-positive hemolytic anemia	Endogenous erythrocyte antigens or haptens	Intravascular hemolysis, clearance, and anemia
	Crohn's disease	Unknown	Transmural colitis and ileitis, ulceration
	Goodpasture's disease	Basement membrane collagen	Nephritis, pneumonitis
	Pemphigus vulgaris	Desmosome antigens in the skin	Blistering skin lesions, immunoglobulin G staining on epidermal cells
	Rheumatic fever	Cardiac muscle antigens	Heart valve damage
	Sjögren's syndrome	Nucleoproteins Ro and La	Dry eyes and mouth
	Thrombocytopenia purpura	Platelet integrins	Hemorrhagic purpura
Immune complex tissue injury	Drug-induced serum sickness	Drug hapten	Vasculitis, hemorrhage, arthralgia, nephritis
	Rheumatoid arthritis	Unknown	Synovitis, arthritis, cartilage and bone loss
	Systemic lupus erythematosus	Nucleic acids, nucleoproteins	Malar skin rash, arthralgia, anemia, arthritis, nephritis
Cellular immune tissue injury	Diabetes mellitus (type 1)	β cell antigens	Hyperglycemia; pancreatic β cell loss
	Hashimoto's thyroiditis	Thyroglobulin; thyroid peroxidase	Hypothyroidism
	Multiple sclerosis	Myelin proteins	Progressing sensory deficits, fatigue, weakness

   (1) These antibodies cause Fc receptor-mediated clearance of erythrocytes.
   (2) They are often induced by drugs that bind to cell surface proteins.

2. The **direct Coombs test** detects IgG antibodies on a patient's erythrocytes.
3. The **indirect Coombs test** detects circulating IgG autoantibodies that can react with endogenous erythrocyte antigens (eg, Rh).

4. **Drug-induced autoimmune hemolytic anemias** are caused by several mechanisms.
    - **a.** Penicillin induces **direct Coombs-positive** anemias.
        - *(1)* The drug binds directly to the erythrocyte surface and induces an antidrug antibody.
        - *(2)* These conditions improve when the drug is discontinued.
    - **b.** Methyldopa induces anemias that are **direct and indirect Coombs positive**.
        - *(1)* This drug induces an antidrug antibody that cross-reacts with an Rh antigen.
        - *(2)* Treatment may require immunosuppression and/or plasmapheresis to remove the autoantibodies.
    - **c.** Other drugs can induce autoantibodies that form immune complexes with the drug.
        - *(1)* The immune complexes can bind to the erythrocyte surface through CR1 (Chapter 8).
        - *(2)* Such conditions are direct and indirect Coombs positive.
        - *(3)* Treatment may require immunosuppression and/or plasmapheresis to remove the immune complexes.
5. **Goodpasture's disease** is mediated by IgG antibodies to basement membrane antigens, which together with neutrophils cause nephritis and pneumonitis.
    - **a.** Autoantibodies in this disease are directed at epitopes of type IV collagen that are found in basement membranes.
    - **b.** Treatment involves antiinflammatory and immunosuppressive drugs and plasmapheresis.
    - **c.** End-stage renal disease may require kidney transplantation.

INFECTIONS AND AUTOIMMUNITY

CLINICAL CORRELATION

- *An increased incidence of infection is not always seen in autoimmune diseases.*
- *However, when autoimmunity adversely affects the integrity of an epithelial barrier (eg, Crohn's disease or Sjögren's syndrome) opportunistic infections can occur.*
- *Complement consumption in immune complex autoimmune diseases can also impair host responses to infectious agents.*
- *Immunosuppressive therapies for autoimmunity increase the risks of infection.*
- *Paradoxically, with a decrease in the overall infection rate in developed countries, an increase in the incidence of allergy and autoimmune disease has occurred.*
- *This may reflect the failure of a microbe-poor environment to induce adequate inhibitory immune responses (eg, Treg cells) that prevent allergy and autoimmunity.*

C. Immune complex-mediated diseases share certain clinical features.

1. **Drug-induced serum sickness** results when antibodies to drugs (eg, penicillin) form soluble immune complexes with their antigens.
    - **a.** These complexes deposit in the skin, blood vessel walls, joints, and kidneys, activate complement, and recruit inflammatory cells.
    - **b.** Vasculitis, hemorrhage, arthralgia, skin rashes, and nephritis are among the principal clinical findings.
2. **SLE** is one of the most common autoimmune disease in women 20–40 years of age, especially African-Americans (Table 16–2).

Table 16–2. Epidemiology of autoimmune diseases.

Disease	HLA Associations	Female to Male Ratios	Approximate Incidence (per 100,000)
Myasthenia gravis	A1, B8, DR4		3
Grave's disease		7	100–300
Addison's disease	DR3, DR4		4
Sjögren's syndrome	B8, DRw52, DR3	9	<600
Crohn's disease	DR1 with DQw5		200
Pemphigus	DR4, DRw6		3
SLE[a]	DR2, DR3	5	50
Diabetes mellitus (type 1)	DR4 or DR1 with DQ1	<1	250
Rheumatoid arthritis	DR4, DR1	3	1000–3000
Hashimoto's thyroiditis	DR3, DR4, DR5	10	500

[a] SLE, systemic lupus erythematosus.

  a. The condition is characterized by fever, arthralgia, nonerosive arthritis, myalgia, photosensitivity, skin rashes, anemia, and glomerulonephritis.

  b. The autoantibodies produced in patients with SLE become increasingly polyspecific with epitope spreading.

    (1) Antibodies to **double-stranded DNA** are diagnostic.

    (2) Additional autoantibodies are produced against nuclear antigens (histones, ribonuclear proteins) and erythrocyte surface antigens.

  c. Immunofluorescence staining for immune complexes or C3b in the kidney shows a "lumpy-bumpy" pattern.

  d. Therapy for this condition is directed at reducing pain, inflammation, and autoantibody production (Table 16–3).

3. **Rheumatoid arthritis** is the most common autoimmune disease in the United States today.

  a. Patients usually first present with morning stiffness and joint pain.

  b. Synovitis results from the infiltration of lymphocytes and macrophages and synovial growth in the form of a **pannus**.

  c. Th1 cells infiltrate the joint and contribute to macrophage and osteoclast activation.

Table 16–3. Therapies used to treat autoimmune diseases.[a]

Treatment	Intent	Diseases
Plasmapheresis	Reduce circulating autoantibodies and immune complexes	Grave's disease, hemolytic anemias, SLE
NSAIDs	Reduce pain and inflammation	Rheumatic diseases in general
Corticosteroids	Antiinflammatory effects	RA, SLE
Cytotoxic drugs (azathioprine)	Inhibit lymphocyte activation and proliferation	SLE
Anticytokine reagents	Block proinflammatory cytokine effects	Crohn's disease (TNF-α), RA, (IL-1β or TNF-α)
Corrective surgery	Removal or replacement of damaged tissues	Crohn's disease, Hashimoto's thyroiditis, RA
Transfusion and transplantation	Organ or tissue replacement	Hemolytic anemias, type I diabetes mellitus

[a] SLE, systemic lupus erythematosus; NSAIDs, nonsteroidal antiinflammatory drugs; RA, rheumatoid arthritis; TNF, tumor necrosis factor; IL, interleukin.

- **d.** Macrophages activate matrix metalloproteases and osteoclasts, which leads to cartilage and bone erosion.
- **e.** Immune complexes, including rheumatoid factor, are present in the joint spaces.
- **f.** Therapy includes immunosuppression, synovectomy, and joint replacement (Table 16–3).

D. Cell-mediated autoimmune diseases do not require autoantibodies.

  1. Multiple sclerosis is a neurodegenerative condition resulting from central nervous system (CNS) demyelination by activated macrophages.
    - **a.** Inflammation is initiated by $CD4^+$ T cells specific for myelin autoantigens (myelin basic protein, myelin oligodendrocyte glycoprotein).
    - **b.** Th1 cytokines recruit and activate T cells and macrophages.
    - **c.** MHC class II and the B7 coreceptor ligand are induced on astrocytes and microglia.

  2. Type 1 (immune-mediated) diabetes mellitus involves the destruction of pancreatic islet β cells and hyperglycemia.
    - **a.** Target autoantigens include the insulin receptor, β cell granule proteins, and insulin.
    - **b.** Increased HLA-DR expression on β cells is common.

   c. CD4$^+$ Th1 cells, CD8$^+$ T cells, and macrophages infiltrate the islets and induce β cell apoptosis.
   d. Insulin replacement by injection, pancreas transplantation, or islet cell transplantation is the preferred therapy.
3. **Sjögren's syndrome** is a disease involving T cell-mediated damage to lacrimal and salivary glands that leads to dry eyes and dry mouth (sicca symptoms).

CLINICAL PROBLEMS

A patient recently diagnosed with SLE is experiencing significant symptoms of 3 weeks duration that include edema and skin rash. Her serum blood urea nitrogen (BUN) and creatinine levels are increased, and a biopsy of her kidney shows IgG and C3b at the glomerular basement membrane. Her serum complement levels are monitored over time.

1. Which of the following findings would be most consistent with a worsening clinical course?
   A. A decline in C9 levels
   B. A decline in C3 and C4 levels
   C. An increase in Factor P levels
   D. An increase in C1 inhibitor (C1 Inh) levels
   E. A decline in C7 levels
2. Of the following laboratory test findings, which would have been most useful in making the initial diagnosis of this disease?
   A. A positive ANA
   B. A positive cross-match
   C. Bence–Jones urinary proteins
   D. A positive mixed lymphocyte reaction
   E. Antibodies to myelin basic protein

Helen's physician has prescribed methyldopa for the treatment of her hypertension. She visits the outpatient medicine clinic several months later complaining of fatigue and has a noticeable pale color in the palms of her hands. Her hemoglobin is 9.5 g/dL (normal range = 13–15 g/dL).

3. What laboratory test should be ordered to determine if her anemia is due to an antibody?
   A. HLA typing
   B. Antithyroglobulin
   C. Coombs test
   D. ANA
   E. Major cross-match

Bill is a 30-year-old man who presents with lower right quadrant abdominal pain, a 10-pound weight loss over the past month, fever, and nonbloody diarrhea. Barium enema contrast imaging indicates mucosal swelling of the wall of the ileum and ascending colon. A biopsy of the colon shows the presence of transmural granulomatous inflammation. Laboratory tests indicate mild anemia, leukocytosis, and decreased total serum protein and albumin, which suggest anorexia.

4. Which of the following therapies would be expected to be most beneficial in treating this condition?

A. Plasmapheresis
B. Anti-tumor necrosis factor (TNF)-α
C. Recombinant C1 Inh
D. Appendectomy
E. Antifungal agents

A 72-year-old man complains of bilateral swelling and pain in the joints of his hands and feet. Several of his proximal phalangeal joints are noticeably deformed. A radiograph of the right hand shows a narrowing of the joint spaces in three digits.

5. Which of the following additional findings would support a diagnosis of rheumatoid arthritis?

A. A decreased erythrocyte sedimentation rate
B. Serum antibodies to double-stranded DNA
C. Antibodies to IgG in the serum
D. A positive response to antibiotics
E. Autoantibody to the TSH receptor

ANSWERS

1. The correct answer is B. Acute episodes in lupus are accompanied by complement depletion by activation of the classical pathway. This leads to a decline in the serum levels of C1, C2, C3, and C4.

2. The correct answer is A. Over 95% of lupus patients have antinuclear antibodies. The other choices do not fit a diagnostic approach to autoimmunity, except choice E. Antibodies to myelin are typical of multiple sclerosis, but not SLE.

3. The correct answer is C. The most appropriate test for establishing an immune basis for an anemia is the Coombs test, which detects IgG antibodies on erythrocytes. This patient's fatigue is most likely related to her low erythrocyte mass.

4. The correct answer is B. This case fits the classic signs and symptoms of Crohn's disease and the imaging and laboratory data support this diagnosis. Chronic inflammation is maintained by the activation of a Th1 type of cytokine production, including IL-12, interferon (IFN)-γ, and TNF-α. Anti-TNF-α has proven effective in treating Crohn's disease in many patients.

5. The correct answer is C. Although not diagnostic of this condition, the presence of rheumatoid factor certainly distinguishes rheumatoid arthritis from several other forms of arthritis. Lupus patients can develop arthritis, but the condition of not erosive in nature. *Increased* sedimentation rates are common in acute disease.

CHAPTER 17
TRANSPLANTATION

I. The principles of transplantation predict the survival of transplanted tissues.

- **A.** Tissue or organs transplanted within an individual are accepted.
  - **1.** Such grafts are called **autografts**.
  - **2.** Examples of autografts include skin transplants to treat local burns and cosmetic hair follicle transplantation to treat hair loss.
- **B.** Grafts exchanged between genetically identical individuals are accepted.
  - **1.** These grafts are called **isografts**.
  - **2.** The donor and recipient are **isogeneic** or **syngeneic** to one another.
- **C.** Grafts between genetically dissimilar individuals are rejected.
  - **1.** The donor and recipient are **allogeneic** to one another.
  - **2.** The grafts are referred to as **allografts**.
- **D.** Grafts exchanged between species are normally rejected.
  - **1.** The donor and recipient are **xenogeneic**.
  - **2.** The grafts are called **xenografts**.
- **E.** Foreign **histocompatibility antigens** initiate graft rejection.
  - **1. Major histocompatibility complex (MHC) alloantigens** (ie, human HLA) are the strongest stimuli for inducing allograft rejection.
  - **2. Minor histocompatibility antigens** and **tissue-specific antigens** can mediate graft rejection.
    - **a.** Vascular endothelial cells express ABO antigens that can stimulate hyperacute rejection.
    - **b.** Endothelium-specific antigens can trigger graft rejection.
  - **3. MHC alloantigens** are recognized directly by large numbers of cross-reacting CD4$^+$ and CD8$^+$ T cells (Figure 17–1).
    - **a.** Most T cell receptors (TCRs) cross-react with allogeneic forms of MHC.
    - **b.** Up to 15% of all peripheral T cells can be activated by a single MHC alloantigen.
  - **4.** Multiple MHC incompatibilities lead to more vigorous graft rejection than does a disparity at a single locus (Figure 17–2).
  - **5.** Disparities at MHC class I + MHC class II loci induce greater responses than does either a class I or a class II difference alone.
  - **6.** Prior immunization can increase the pace of allograft rejection.

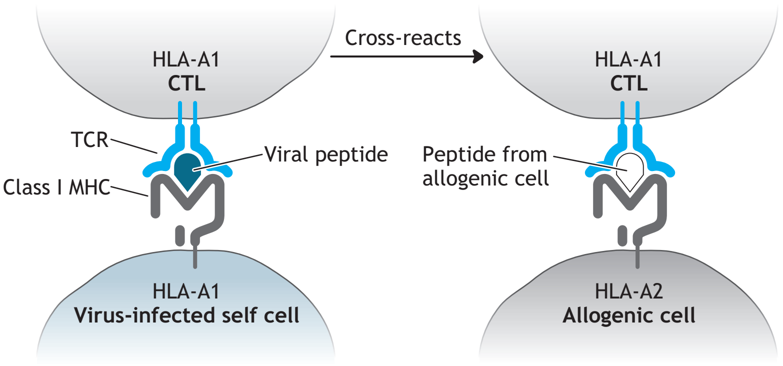
Cross-reacts
HLA-A1
CTL
HLA-A1
CTL
TCR
Viral peptide
Class I MHC
Peptide from allogenic cell
HLA-A1
Virus-infected self cell
HLA-A2
Allogenic cell

Figure 17–1. Recognition of an allogeneic major histocompatibility complex (MHC) molecule by T cells. CTL, cytotoxic T lymphocyte; TCR, T cell receptor.

a. **"First set" rejection** occurs within days because the number of T cells activated by a foreign graft is large.

b. Repeated exposure to the same MHC alloantigens (eg, through blood transfusions) can cause even more rapid rejection (**"second set" rejection**).

F. The fate of allografts varies somewhat depending on the nature of the tissue or organ transplanted (Table 17–1).

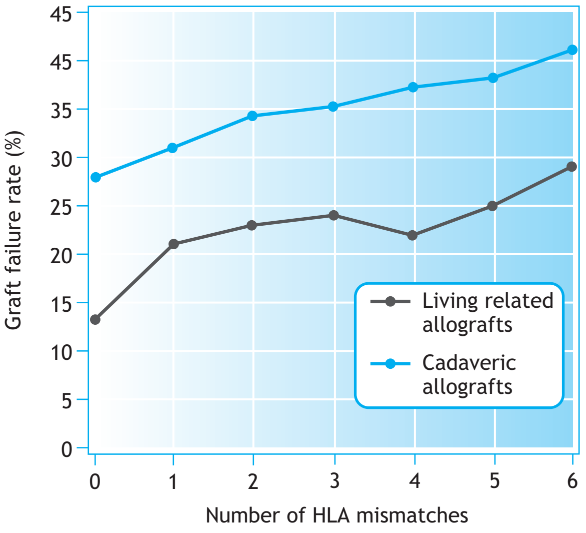

Figure 17–2. Influence of HLA on 5-year allograft failure rates.

Table 17–1. Special considerations related to particular forms of transplantation.

Organ or Tissue Transplanted	Considerations
Kidney	95% overall success rate after 1 year when living related kidneys are used; diabetes is the most frequent cause of end-stage renal failure
Heart	Reserved for imminently life-threatening cardiac disease; often combined with lung transplantation; relaxed criteria for donor selection due to organ shortage and effectiveness of cyclosporin
Lung	Low success rate compared to other solid organs
Blood	Transfusions are typically tolerated despite HLA disparity; preformed anti-HLA antibodies can cause transfusion reactions
Liver	Less requirement for HLA typing; evidence that tolerance to allogeneic transplants develops with time; contraindicated in certain malignancies, viral hepatitis, and human immunodeficiency virus (HIV) infection
Pancreas	Prevention of secondary effects of diabetes; whole organ, segmental, or isolated islet cell transplantation possible
Hematopoietic stem cells	Concern for graft-versus-host disease and infections; graft-versus-tumor effect beneficial
Cornea	Benefits from the privileged status of the anterior chamber

CLINICAL CORRELATION

CORNEAL TRANSPLANTS

- *Approximately 40,000 corneal transplants are performed in the United States each year.*
- *The donors are not routinely matched with recipients for HLA, and recipients do not receive systemic immunosuppression following transplantation.*
- *The immunologically privileged status of the anterior chamber of the eye results in a low level of graft rejection (~10%).*
- *Immune privilege is maintained by a lack of lymphatic drainage and the expression of* ***Fas ligand*** *(Chapter 6) on corneal endothelial and epithelial cells.*
- *Invading activated T cells expressing Fas are killed by binding to ocular Fas ligand.*
- *When they occur, rejection episodes can be treated with topical corticosteroids.*

II. Allograft rejection involves a diverse set of immune mediators (Figure 17–3).

A. **CD4$^+$** Th1 cells activate various effector cells by secreting cytokines.

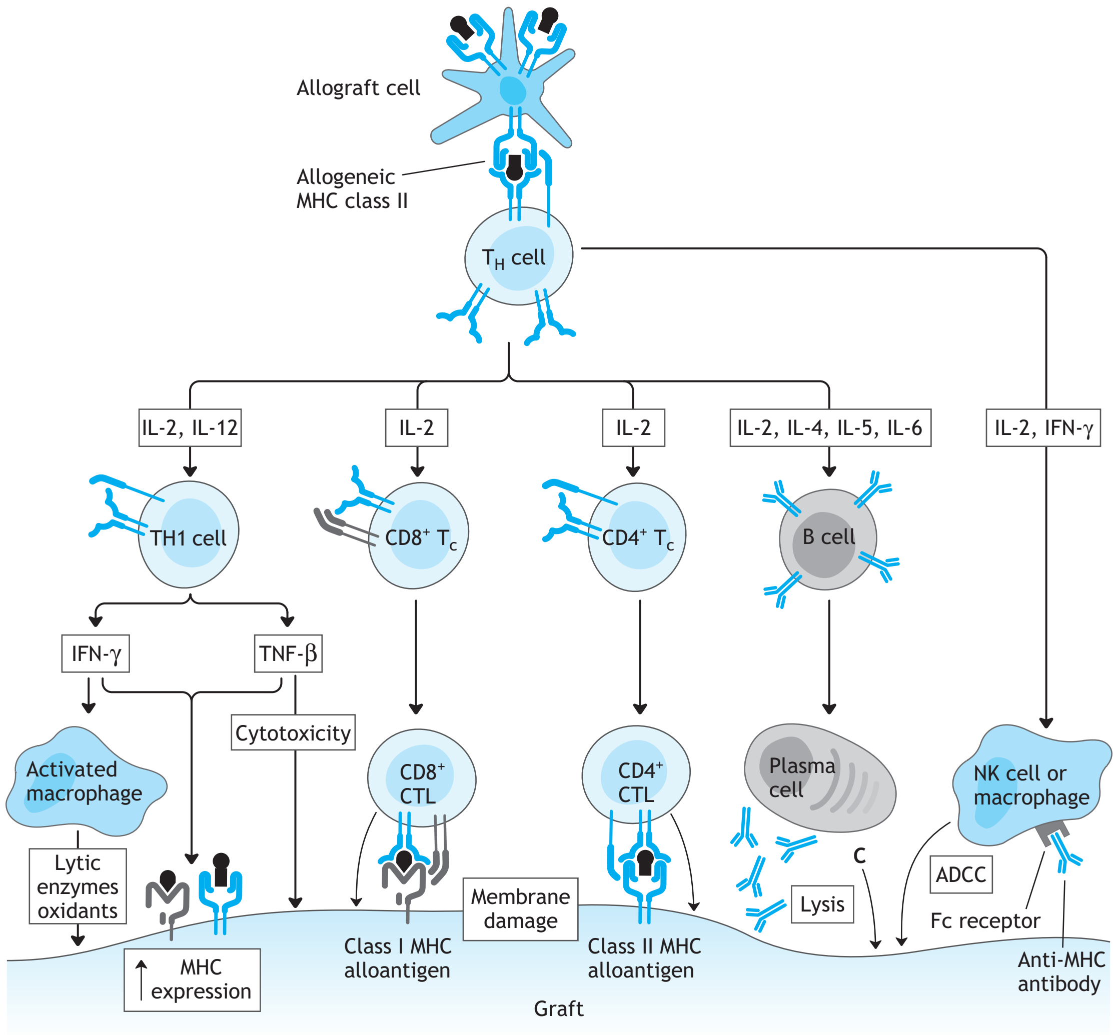
Allograft cell
Allogeneic MHC class II
T_H cell
IL-2, IL-12
IL-2
IL-2
IL-2, IL-4, IL-5, IL-6
IL-2, IFN-γ
TH1 cell
CD8$^+$ T_C
CD4$^+$ T_C
B cell
IFN-γ
TNF-β
Cytotoxicity
Activated macrophage
Lytic enzymes oxidants
MHC expression
CD8$^+$ CTL
CD4$^+$ CTL
Plasma cell
NK cell or macrophage
Membrane damage
Class I MHC alloantigen
Class II MHC alloantigen
Lysis
C
ADCC
Fc receptor
Anti-MHC antibody
Graft

Figure 17–3. Mediators of allograft rejection. MHC, major histocompatibility complex; T_H, T helper; IL, interleukin; T_C, cytotoxic T cell; IFN, interferon; TNF, tumor necrosis factor; NK, natural killer; ADCC, antibody-dependent cell-mediated cytotoxicity.

1. Th1 cells recognize **allogeneic HLA-DP, DQ, and DR** molecules.
2. The Th1 cell subset produces interleukin (IL)-2, interferon (IFN)-γ, and tumor necrosis factor (TNF)-α, which activate cytotoxic T cells, natural killer (NK) cells, and macrophages.

B. **CD4$^+$ Th2 cells** are activated by allogeneic MHC class II molecules.
   1. The Th2 cytokines IL-4 and IL-5 induce growth, class switching, and antibody production by B cells.

    2. **Antibodies** specific for allogeneic donor MHC molecules can mediate **hyperacute rejection** (see below).
    3. Immunoglobulin (Ig) G alloantibodies can promote **antibody-dependent cellular cytotoxicity (ADCC)** by NK cells and macrophages.

- C. **Cytokines** that mediate allograft rejection are also produced by NK cells (IFN-γ), CD8$^+$ T cells (IFN-γ), and macrophages (TNF-α).
- D. **CD8$^+$ cytotoxic T cells** recognize allogeneic MHC class I molecules.
    1. Killing is mediated by the perforin–granzyme, Fas–Fas ligand, and TNF receptor pathways (Chapter 10).
    2. Rejection is due to apoptotic cell death in the transplanted tissues.
- E. **Macrophages** kill donor cells with oxidants, hydrolytic enzymes, and TNF-α.
- F. **NK cells** kill by recognizing the absence of self MHC class I molecules.

III. Three mechanisms of allograft rejection exist.

- A. **Hyperacute allograft rejection** occurs within minutes to hours and is mediated by preformed antibodies in the recipient.
    1. The principal target antigens include allogeneic HLA and ABO, both of which are expressed on the donor's vascular endothelial cells.
    2. Anti-HLA antibodies are produced in response to pregnancy, blood transfusions, or prior transplants.
    3. These IgM and IgG antibodies activate complement through the classical pathway.
        - a. C5a attracts neutrophils to the allograft.
        - b. iC3b promotes neutrophil attachment and activation.
        - c. C3a, C4a, and C5a activate mast cells, which increases the inflammatory response within the graft.
    4. Activated neutrophils, monocytes, and NK cells damage the endothelium.
    5. Activation and damage to the vascular endothelium cause important hemodynamic changes in the allograft.
        - a. Tissue factor is expressed by endothelial cells and initiates thrombosis.
        - b. Endothelial cell damage promotes platelet activation, thrombosis, infarction, edema, and hemorrhage.
- B. **Acute allograft rejection** occurs over weeks to months and is mediated by CD4$^+$ Th1 cells, CD8$^+$ cytotoxic T cells, NK cells, and macrophages that infiltrate the graft.
- C. **Chronic allograft rejection** occurs over months to years.
    1. **Fibrosis, infarction, and ischemia** are common features.
    2. Chronic rejection is difficult to control by immunosuppression.

IV. Graft rejection can be predicted or prevented by matching the donor and recipient histocompatibility antigens (Table 17–2).

- A. **Tissue typing** determines the tissue antigens expressed by prospective donors and recipients and any preformed immunity to those antigens.
    1. **ABO matching** is essential for preventing endothelial damage by isohemagglutinins.
    2. Preformed antibodies to ABO are detected with the **major cross-match** (incubating donor cells in recipient serum).

Table 17–2. Prospective HLA typing for the transplantation of solid organ allografts.

	Mediator of Rejection		
	CD4$^+$ Th Cells	CD8$^+$ Cytotoxic T Cells	Antibodies
HLA antigens recognized	HLA-DP, DQ, DR	HLA-A, B, C	ABO; HLA-A, B, C, DP, DQ, DR
Important typing procedures	Type and match HLA-DP, DQ, DR Perform mixed lymphocyte reaction (MLR)	Type and match HLA-A, B, C	Major cross-match to detect anti-ABO and anti-HLA

TRANSPLANTATION ACROSS ABO INCOMPATIBILITIES

CLINICAL CORRELATION

- *Conventional wisdom advises against transplanting organs when an ABO incompatibility exists between the donor and recipient.*
- *This restriction has limited the number of cardiac transplants, especially those involving infants with congenital heart defects.*
- *Recently, it was discovered that ABO incompatibility need not prevent the transplantation of HLA-matched hearts into very young children.*
- *The isohemagglutinins that mediate hyperacute rejection develop slowly over time in the neonate.*
- *The antigenic stimulus for inducing isohemagglutinins appears to be the cross-reacting polysaccharides of microbial flora.*
- *The delayed appearance of these natural antibodies reflects a gradual acquisition of commensal organisms during the initial months of postnatal life.*

3. **HLA typing** is performed by molecular techniques [eg, polymerase chain reaction (PCR)].
   a. The best donor–recipient pair is selected based on similarities between their HLA genotypes.
   b. HLA typing and matching can minimize acute rejection.
   c. Extensive matching of HLA types is not as important in liver transplants as in kidney transplants (Table 17–1).
   d. The shortage of heart allografts and the effectiveness of cyclosporine in cardiac transplantation obviate the need for extensive HLA typing.
   e. HLA typing is important for identifying haplotype segregation within a family (Figure 7–7).
   f. Transplantation using organs from deceased donors may not afford an opportunity for detailed tissue typing.
   g. High-resolution typing of HLA-D region antigens is important in stem cell transplantation to avoid graft-versus-host disease (GVHD).
4. Preformed anti-HLA antibodies are detected with a **major cross-match**.

- **a.** Such antibodies can be formed against either class I or class II alloantigens.
- **b.** The major cross-match can prevent hyperacute rejection.

5. The **mixed lymphocyte reaction (MLR)** detects T cell-mediated reactivity between donor and recipient (Figure 17–4).

- **a.** The assay measures the recognition of allogeneic MHC by coculturing the lymphocytes from two individuals.
- **b.** The MLR also predicts the potential for **GVHD** following stem cell transplantation.

TRANSPLANTATION WAITING LISTS

CLINICAL CORRELATION

- *Recent national data (www.ustransplant.org) indicate that over 80,000 patients in the United States are waiting for organ transplants, a number that has doubled in less than 10 years.*
- *Nearly 60,000 patients await kidney transplantation, while only 15,000 renal transplants are performed each year.*
- *Two-thirds of prospective kidney transplant recipients spend over a year on the waiting list.*
- *The median time until transplantation is over 3 years.*

V. Xenotransplantation has the potential to address the shortage of available donor organs and tissues.

A. The species providing a xenotransplant to a human patient is determined by organ physiology as well as social and ethical concerns.

1. Kidneys from minipigs are suitable for transplantation into humans, based on anatomical and physiological considerations.
2. Pigs are also suitable for genetic manipulation (see the Clinical Correlation on transgenic pigs in Chapter 8).

B. Xenotransplantation carries the risk of transmitting **zoonotic infections**.

1. Retroviruses similar to human immunodeficiency virus type 1 (HIV-1) are a particular concern.

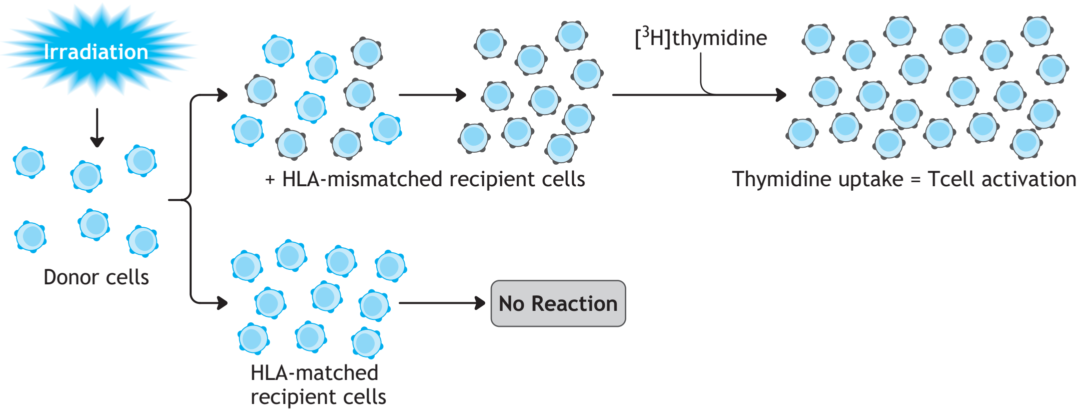
Irradiation
Donor cells
+ HLA-mismatched recipient cells
[^{3}H]thymidine
Thymidine uptake = Tcell activation
No Reaction
HLA-matched recipient cells

Figure 17–4. Mixed lymphocyte reaction. *Blood mononuclear cells.

  2. Immunosuppressed transplant recipients are at increased risk for such infections.

C. The primary immunological barrier to xenotransplantation appears to be **hyperacute rejection**.
  1. Natural IgM antibodies against widely distributed ABO-like carbohydrate xenoantigens are present in nearly all species.
  2. These antibodies activate complement and initiate neutrophil-mediated damage to vascular endothelial cells.
  3. Because the membrane complement regulatory proteins [eg, decay-accelerating factor (DAF)] of the pig do not inhibit the human C3 convertases, uncontrolled complement-mediated damage also occurs.

D. T cells that recognize the MHC molecules of the xenogeneic species are also activated and mediate strong rejection responses.

VI. Hematopoietic stem cell (HSC) transplantation provides a therapy for hematopoietic disorders, immune deficiencies, and cancer (Table 17–3).

A. HSCs bear the **CD34** marker and can be purified from bone marrow cells.
  1. HSCs can also be purified from peripheral blood after mobilization in the donor by injection of **colony-stimulating factor-granulocyte (CSF-G)**.
  2. For the treatment of malignant conditions, autologous HSCs can be collected during remission and transplanted at a later time.

B. Several potential outcomes of stem cell transplantation must be balanced.
  1. **Rejection** of an HSC transplant can be minimized by tissue typing.
  2. **GVHD** represents a significant risk in HSC transplantation beginning at the first month posttransplantation.
    a. Acute GVHD presents as a sunburn-like rash, hepatosplenomegaly, jaundice, elevated billirubin, and diarrhea.
    b. GVHD occurs when the recipient expresses HLA antigens not present on the donor.

Table 17–3. Diseases treated by hematopoietic stem cell transplantation.

Defects in Hematopoiesis	Immune Deficiencies	Malignancies
Sickle cell disease	Severe combined immune deficiency	Acute myeloid leukemia
Thalassemia	Wiscott–Aldrich syndrome	Acute lymphoblastic leukemia
Aplastic anemia	Leukocyte adhesion defect	Chronic myelogenous leukemia
Pure red cell aplasia	Chediak–Higashi syndrome Hyper-IgM syndrome	Multiple myeloma

   c. GVHD requires donor T cells capable of responding to allogeneic HLA of the recipient.
   d. GVHD is most evident when the recipient is immunosuppressed or genetically incapable of rejecting donor T cells.
   e. GVHD can be anticipated by tissue typing in the direction of donor anti-host.

3. Despite the negative effects of GVHD, donor T cells can also have a beneficial **graft-versus-tumor effect** in malignant diseases.
4. The use of preconditioning regimens (eg, chemotherapy prior to transplantation) and powerful immunosuppressive drugs increases the risks of **opportunistic infections**.
   a. Depending on the degree of myeloablation, early infections can be severe and can be caused by viruses, fungi, and bacteria.
   b. Common pathogens include cytomegalovirus, herpes simplex virus, *Candida albicans,* and *Aspergillus* species.
   c. Neutrophil functions are generally restored first.
   d. Prophylactic therapies include antimicrobial agents, laminar flow rooms, intravenous immunoglobulin (IVIG), and pediatric vaccines after lymphocyte reconstitution.

VII. Posttransplantation immunosuppressive therapy primarily targets lymphocyte activation, growth, and differentiation and inhibits inflammation (Table 17–4).

Table 17–4. Immunosuppressive drugs used in transplantation.[a]

Agents	Targets	Clinical Use
Irradiation	Many cells	Conditioning for stem cell transplantation
Methotrexate, azathioprine	Many cells	Organ transplantation, autoimmunity
Mycofenolate mofetil	Lymphocytes	Organ transplantation
Cyclosporine A (CsA), tacrolimus (FK506), sirolimus (rapamycin)	Lymphocytes	Organ transplantation
Corticosteroids	Many cells	Inflammatory diseases, organ transplantation
OKT3	T cells	Organ transplantation
ALG, ATG, daclizumab	Lymphocytes	Organ transplantation

[a] ALG, antilymphocytic globulin; ATG, antithymocyte globulin.

A. **Cyclosporine** and related drugs block calcineurin-dependent activation of nuclear factor of activated T cells (NFAT) in T cells and inhibit cytokine (eg, IL-2) production (Chapter 6).

B. Azathioprine and methotrexate are cytotoxic to lymphocytes.
   1. **Azathioprine** inhibits DNA synthesis.
   2. **Methotrexate** is a folate antagonist.
   3. The drugs are also used to treat autoimmune diseases (Chapter 16).

C. Antibodies against lymphocytes or their receptors block signaling and cause lymphocyte death.
   1. **OKT3** is an antibody to CD3 used in pancreas transplantation.
   2. **Antilymphocyte globulin** and **antithymocyte globulin** deplete T cells.
   3. **Daclizumab** is an antibody against the IL-2 receptor (CD25), which inhibits lymphocyte proliferation.

D. **Corticosteroids** inhibit inflammation (eg, neutrophil and macrophage activation) associated with graft destruction.

E. **Mycophenolate mofetil** inhibits purine metabolism and blocks lymphocyte proliferation.

F. All of these drugs are essentially nonspecific in the sense that they inhibit protective immune responses while suppressing graft rejection.

CLINICAL PROBLEMS

Mary is a 52-year-old woman with end-stage renal disease secondary to diabetes. She received an allogeneic renal transplant 2 months ago and has progressed to a maintenance dose of immunosuppressive therapy. However, within the last week she has been experiencing a progressive decline in renal function [decreased urine output, elevated blood urea nitrogen (BUN) and creatine] and currently has a temperature of 39°C. Her kidney is now tender, painful, and swollen. A renal biopsy shows a dense interstitial mononuclear infiltration.

1. Which of the following is most likely responsible for the change in her clinical course over the past week?

   A. An opportunistic viral infection
   B. Renal toxicity due to her immunosuppressive drugs
   C. Graft-versus-host disease
   D. Acute allograft rejection
   E. Diabetic nephropathy

Two months after having received an allogeneic stem cell transplant for the treatment of acute lymphoblastic leukemia, a patient presented with diarrhea and a blistering, erythematous appearance to the skin of his anterior neck and back. Serum chemistry revealed elevated liver enzymes and billirubin, and the liver and spleen appeared enlarged on physical examination.

2. Which of the following diseases or pathogenic mechanisms is most likely causing this clinical picture?
   A. Rejection of the transplant
   B. Autoantibody production against the patient's lymphocytes
   C. Deposition of immune complexes in the skin and liver
   D. A reaction to his immunosuppressive drugs
   E. Graft-versus-host disease

As a treatment for end-stage renal disease, xenotransplantation involves unique clinical challenges not seen in allotransplantation.

3. Which of the following mechanisms is more common in the rejection of renal *xeno*grafts than renal allografts?
   A. Production of IgE antibodies
   B. Activation of HLA class I-restricted $CD4^+$ T cells
   C. Binding of anti-HLA antibodies to donor tissues
   D. Activation of complement by natural antibodies
   E. Attraction of eosinophils into the graft

The HLA phenotypes of a family, including a patient who requires a kidney transplant, are listed below.

Family Member	HLA-A Alleles	HLA-B Alleles	HLA-DR Alleles
Mother	1,7	8,12	4,5
Father	5,7	28,32	4,6
Daughter #1	1,7	8,32	5,6
Daughter #2	1,7	8,28	4,4
Son #1	1,5	12,28	4,5
Son #2	1,7	12,32	5,6
Patient	**1,5**	**8,28**	**4,4**

4. All other considerations being equal, which of the family members is best suited as the transplant donor?
   A. Daughter #1
   B. Daughter #2
   C. Mother
   D. Son #1
   E. Son #2

5. Which of the donors above would be most appropriate to donate *stem cells* to the patient if that were the indicated therapy?

A. Daughter #1
B. Daughter #2
C. Mother
D. Son #1
E. Son #2

ANSWERS

1. The correct answer is D. This is a typical description of acute renal allograft rejection. The histopathology confirms rejection, but does not fit a virus infection or diabetic nephropathy (hyaline thickening of the glomerular capillaries).

2. The correct answer is E. GVHD is common in allogeneic stem cell transplantation and involves the liver, spleen, gastrointestinal tract, and skin.

3. The correct answer is D. Although allografts can also undergo hyperacute rejection initiated by isohemagglutinins, this rarely occurs due to the now routine use of the major cross-match. Hyperacute rejection is currently the major initial barrier to successful xenografting.

4. The correct answer is B. Daughter #2 has only a single disparity at one MHC class I locus. Particularly important in selecting a donor–recipient pair is avoiding disparities at both class I and class II loci.

5. The correct answer is B. The most important consideration in selecting a stem cell donor is avoiding GVHD, which is primarily induced by donor cells that recognize HLA-D region alloantigens of the recipient. The second consideration in selecting a donor for stem cell transplantation is the survival of the graft, which is dictated by the same rules that govern solid organ transplantation.

Appendix I. CD markers and their functions.[a]

CD Marker	Principal Functions
CD1	Presentation of glycolipids to NKT cells
CD3	Signaling chains of the TCR
CD4	Coreceptor for MHC class II-restricted T cells
CD8	Coreceptor for MHC class I-restricted T cells
CD11	α chain of the β_2 integrin family
CD14	Nonsignaling component of the LPS receptor
CD18	β chain of the β_2 integrin family
CD19	Signal transducer on B cells
CD20	Chain of the CR2 coreceptor
CD21	Chain of the CR2 coreceptor
CD25	α chain of the IL-2 receptor
CD28	Coreceptor on T cells that binds B7
CD34	Marker on hematopoietic stem cells
CD40	Coreceptor on B cells
CD45	Tyrosine phosphatase of T and B cells
CD55	Membrane decay accelerating factor
CD59	Membrane regulator of MAC assembly
CD80	B7-1 coreceptor on APCs
CD154	Ligand on T cells for CD40

[a] NKT, natural killer T; TCR, T cell receptor; MHC, major histocompatibility complex; LPS, lipopolysaccharide; IL, interleukin; MAC, membrane attack complex; APC, antigen-presenting cell.

Appendix II. Cytokines.[a]

Cytokine	Principal Functions
IL-1 α, β	Similar to TNF-α; fever, leukocyte adhesion, lymphocyte costimulation
IL-2	T, B, and NK cell proliferation
IL-3	Proliferation of progenitor cells; growth factor for mast cells
IL-4	Differentiation of Th2 cells and B cells; Ig class switching to IgE.
IL-5	Differentiation of B cells and eosinophils; Ig class switching to IgA
IL-6	Hematopoiesis, acute phase response
IL-7	Progenitor B and T cell growth
IL-8	Chemotaxis of neutrophils
IL-10	Inhibits the Th1 pathway
IL-12	Coactivator of the Th1 pathway; induces IFN-γ production
IL-13	Ig class switching to IgE
IL-15	NK cell growth
IL-18	Coactivator for IFN-γ production
TGF-β	Antiinflammatory; promotes wound healing; chemotaxis; Ig class switching to IgA
TNF-α	Proinflammatory; neutrophil and endothelial cell activation; proapoptotic
IFN-α,β	Inhibits viral replication; primes macrophages
IFN-γ	Macrophage activation; inhibition of Th2 pathway
CSF-M	Growth and differentiation of monocyte/macrophage lineage
CSF-G	Growth and differentiation of granulocytes
CSF-GM	Growth and differentiation of myeloid lineage cells

[a] IL, interleukin; TNF, tumor necrosis factor; NK, natural killer; Ig, immunoglobulin; IFN, interferon; TGF, transforming growth factor; CSF, colony-stimulating factor; M, macrophage; G, granulocyte. See also Table 12-3 for a list of important chemokines.

Abiogenesis

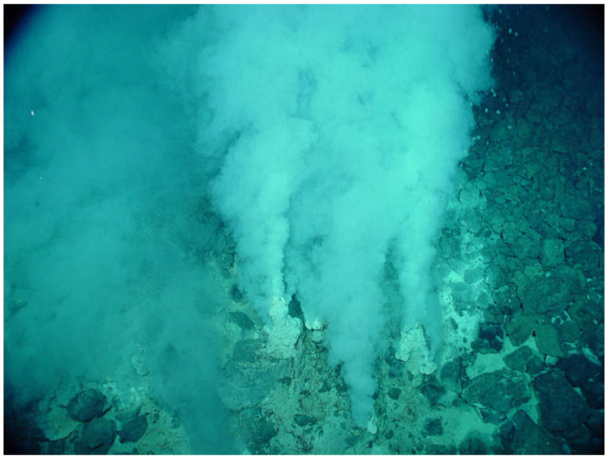

The earliest known life-forms on Earth are putative fossilized microorganisms, found in hydrothermal vent precipitates, that may have lived as early as 4.28 billion years ago, relatively soon after the oceans formed 4.41 billion years ago, and not long after the formation of the Earth 4.54 billion years ago.[1][2]

Abiogenesis, or informally the **origin of life**,[3][4][5][note 1] is the natural process by which life arises from non-living matter, such as simple organic compounds.[6][4][7][8] While the details of this process are still unknown, the prevailing scientific hypothesis is that the transition from non-living to living entities was not a single event, but a gradual process of increasing complexity that involved molecular self-replication, self-assembly, autocatalysis, and the emergence of cell membranes.[9][10][11] Although the occurrence of abiogenesis is uncontroversial among scientists, there is no single, generally accepted model for the origin of life, and this article presents several principles and hypotheses for *how* abiogenesis could have occurred.

Researchers study abiogenesis through a combination of molecular biology, paleontology, astrobiology, oceanography, biophysics, geochemistry and biochemistry, and aim to determine how pre-life chemical reactions gave rise to life.[12] The study of abiogenesis can be geophysical, chemical, or biological,[13] with more recent approaches attempting a synthesis of all three,[14] as life arose under conditions that are strikingly different from those on Earth today. Life functions through the specialized chemistry of carbon and water and builds largely upon four key families of chemicals: lipids (fatty cell walls), carbohydrates (sugars, cellulose), amino acids (protein metabolism), and nucleic acids (self-replicating DNA and RNA). Any successful theory of abiogenesis must explain the origins and interactions of these classes of molecules.[15] Many approaches to abiogenesis investigate how self-replicating molecules, or their components, came into existence. Researchers generally think that current life on Earth descends from an RNA world,[16] although RNA-based life may not have been the first life to have existed.[17][18]

The classic 1952 Miller–Urey experiment and similar research demonstrated that most amino acids, the chemical constituents of the proteins used in all living organisms, can be synthesized from inorganic compounds under conditions intended to replicate those of the early Earth. Scientists have proposed various external sources of energy that may have triggered these reactions, including lightning and radiation. Other approaches ("metabolism-first" hypotheses) focus on understanding how catalysis in chemical systems on the early Earth might have provided the precursor molecules necessary for self-replication.[19] Complex organic molecules occur in the Solar System and in interstellar space, and these molecules may have provided starting material for the development of life on Earth.[20][21][22][23]

The biochemistry of life may have begun shortly after the Big Bang, 13.8 billion years ago, during a habitable epoch when the age of the universe was only 10 to 17 million years.[24][25] The panspermia hypothesis suggests that microscopic life was distributed to the early Earth by space dust,[26] meteoroids,[27] asteroids and other small Solar System bodies and that life may exist throughout the universe.[28] The panspermia hypothesis proposes that life originated outside the Earth, but does not definitively explain its origin.

Nonetheless, Earth remains the only place in the universe known to harbour life,[29][30] and fossil evidence from the Earth informs most studies of abiogenesis. The age of the Earth is about 4.54 billion years;[31][32][33] the earliest undisputed evidence of life on Earth dates from at least 3.5 billion years ago,[34][35][36] and possibly as early as the Eoarchean Era (between 3.6 and 4.0 billion years ago), after geological crust started to solidify following the molten Hadean Eon. In May 2017 scientists found possible evidence of early life on land in 3.48-billion-year-old geyserite and other related mineral deposits (often found around hot springs and geysers) uncovered in the Pilbara Craton of Western Australia.[37][38][39][40] However, a number of discoveries suggest that life may have appeared on Earth even earlier. As of 2017, microfossils, or fossilised microorganisms, within hydrothermal-vent precipitates dated from 3.77 to 4.28 billion years old found in Quebec, Canadian rocks may harbor the oldest record of life on Earth, suggesting life started soon after ocean formation 4.4 billion years ago.[1][2][41][42][43] According to biologist Stephen Blair Hedges, "If life arose relatively quickly on Earth ... then it could be common in the universe."[44][45][46]

Contents

Early geophysical conditions on Earth
- Earliest biological evidence for life

Conceptual history
- Spontaneous generation
- Etymology
- Louis Pasteur and Charles Darwin
- "Primordial soup" hypothesis
- Proteinoid microspheres

Current models

Chemical origin of organic molecules
- Chemical synthesis
- Autocatalysis
- Homochirality

Self-enclosement, reproduction, duplication and the RNA world
- Protocells
- RNA world
  - Viral origins
- RNA synthesis and replication
- Pre-RNA world

Origin of biological metabolism
- Iron–sulfur world
- Zn-world hypothesis
- Deep sea vent hypothesis
- Thermosynthesis

Other models
- Clay hypothesis
- Gold's "deep-hot biosphere" model
- Panspermia
- Extraterrestrial organic molecules
- Lipid world
- Polyphosphates
- PAH world hypothesis
- Radioactive beach hypothesis

Thermodynamic dissipation
Multiple genesis
Fluctuating hydrothermal pools on volcanic islands or proto-continents
Information theory
See also
Notes
References
Bibliography
External links

Early geophysical conditions on Earth

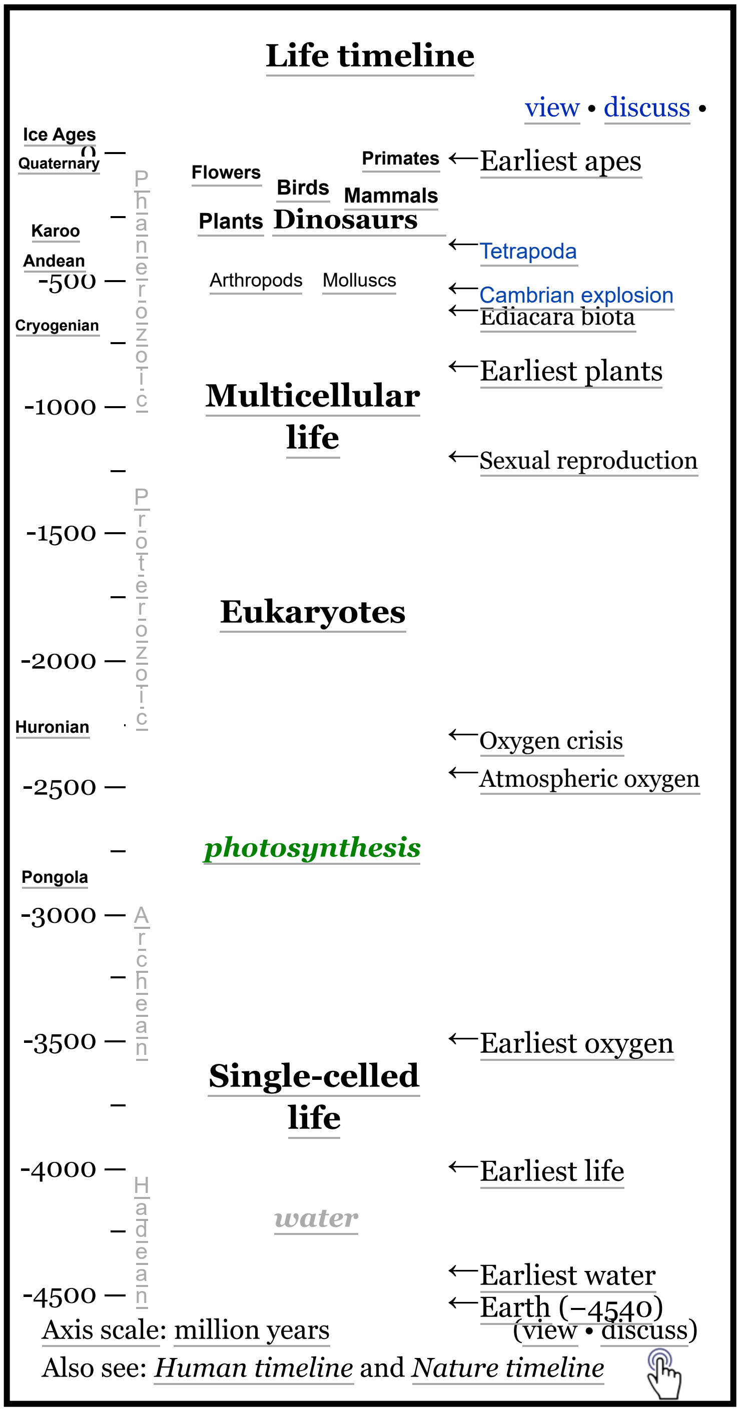

The Hadean Earth is thought to have had a secondary atmosphere, formed through degassing of the rocks that accumulated from planetesimal impactors. At first, it was thought that the Earth's atmosphere consisted of hydrogen compounds—methane, ammonia and water vapour—and that life began under such reducing conditions, which are conducive to the formation of organic molecules. According to later models, suggested by study of ancient minerals, the atmosphere in the late Hadean period consisted largely of water vapour, nitrogen and carbon dioxide, with smaller amounts of carbon monoxide, hydrogen, and sulfur compounds.[47] During its formation, the Earth lost a significant part of its initial mass, with a nucleus of the heavier rocky elements of the protoplanetary disk remaining.[48] As a consequence, Earth lacked the gravity to hold any molecular hydrogen in its atmosphere, and rapidly lost it during the Hadean period, along with the bulk of the original inert gases. The solution of carbon dioxide in water is thought to have made the seas slightly acidic, giving it a pH of about 5.5. The atmosphere at the time has been characterized as a "gigantic, productive outdoor chemical laboratory."[49] It may have been similar to the mixture of gases released today by volcanoes, which still support some abiotic chemistry.[49]

Oceans may have appeared first in the Hadean Eon, as soon as two hundred million years (200 Ma) after the Earth was formed, in a hot 100 °C (212 °F) reducing environment, and the pH of about 5.8 rose rapidly towards neutral.[50] This has been supported by the dating of 4.404 Ga-old zircon crystals from metamorphosed quartzite of Mount Narryer in the

Western Australia Jack Hills of the Pilbara, which are evidence that oceans and continental crust existed within 150 Ma of Earth's formation.[51] Despite the likely increased volcanism and existence of many smaller tectonic "platelets," it has been suggested that between 4.4 and 4.3 Ga (billion year), the Earth was a water world, with little if any continental crust, an extremely turbulent atmosphere and a hydrosphere subject to intense ultraviolet (UV) light, from a T Tauri stage Sun, cosmic radiation and continued bolide impacts.[52]

The Hadean environment would have been highly hazardous to modern life. Frequent collisions with large objects, up to 500 kilometres (310 mi) in diameter, would have been sufficient to sterilize the planet and vaporize the ocean within a few months of impact, with hot steam mixed with rock vapour becoming high altitude clouds that would completely cover the planet. After a few months, the height of these clouds would have begun to decrease but the cloud base would still have been elevated for about the next thousand years. After that, it would have begun to rain at low altitude. For another two thousand years, rains would slowly have drawn down the height of the clouds, returning the oceans to their original depth only 3,000 years after the impact event.[53]

Earliest biological evidence for life

The most commonly accepted location of the root of the tree of life is between a monophyletic domain Bacteria and a clade formed by Archaea and Eukaryota of what is referred to as the "traditional tree of life" based on several molecular studies starting with C. Woese.[54] A very small minority of studies have concluded differently, namely that the root is in the Domain Bacteria, either in the phylum Firmicutes[55] or that the phylum Chloroflexi is basal to a clade with Archaea+Eukaryotes and the rest of Bacteria as proposed by Thomas Cavalier-Smith.[56] More recently Peter Ward has established an alternative view which is rooted in abiotic RNA synthesis which becomes enclosed within a capsule and then creates RNA ribozyme replicates. It is proposed that this then bifurcates between Dominion Ribosa (RNA life), and after the loss of ribozymes RNA viruses as Domain Viorea, and Dominion Terroa, which after creating a large cell within a lipid wall, creating DNA the 20 based amino acids and the triplet code, is established as the last universal common ancestor or LUCA, of earlier phylogenic trees.[57]

Precambrian stromatolites in the Siyeh Formation, Glacier National Park. In 2002, a paper in the scientific journal *Nature* suggested that these 3.5 Ga (billion years) old geological formations contain fossilized cyanobacteria microbes. This suggests they are evidence of one of the earliest life forms on Earth.

The earliest life on Earth existed more than 3.5 billion years ago,[34][35][36] during the Eoarchean Era when sufficient crust had solidified following the molten Hadean Eon. The earliest physical evidence so far found consists of microfossils in the Nuvvuagittuq Greenstone Belt of Northern Quebec, in "banded iron formation" rocks at least 3.77 billion and possibly 4.28 billion years old.[1][58] This finding suggested that there was almost instant development of life after oceans were formed. The structure of the microbes was noted to be similar to bacteria found near hydrothermal vents in the modern era, and provided support for the hypothesis that abiogenesis began near hydrothermal vents.[42][1]

Also noteworthy is biogenic graphite in 3.7 billion-year-old metasedimentary rocks from southwestern Greenland[59] and microbial mat fossils found in 3.48 billion-year-old sandstone from Western Australia.[60][61] Evidence of early life in rocks from Akilia Island, near the Isua supracrustal belt in southwestern Greenland, dating to 3.7 billion years ago have shown biogenic carbon isotopes.[62][63] In other parts of the Isua supracrustal belt, graphite inclusions trapped within garnet crystals are connected to the other elements of life: oxygen, nitrogen, and possibly phosphorus in the form of phosphate, providing further evidence for life 3.7 billion years ago.[64] At Strelley Pool, in the Pilbara region of Western Australia, compelling evidence of early life was found in pyrite-bearing sandstone in a fossilized beach, that showed rounded tubular cells that oxidized sulfur by photosynthesis in the absence of oxygen.[65][66][67] Further research on zircons from Western Australia in 2015 suggested that life likely existed on Earth at least 4.1 billion years ago.[44][68][69]

Traditionally it was thought that during the period between 4.28[1][2] and 3.8 Ga, changes in the orbits of the giant planets may have caused a heavy bombardment by asteroids and comets[70] that pockmarked the Moon and the other inner planets (Mercury, Mars, and presumably Earth and Venus). This would likely have repeatedly sterilized the planet, had life appeared before that time.[49] Geologically, the Hadean Earth would have been far more active than at any other time in its history. Studies of meteorites suggests that radioactive isotopes such as aluminium-26 with a half-life of 7.17×10^5 (717 thousand) years, and potassium-40 with a half-life of 1.250×10^9 (1.25 billion) years, isotopes mainly produced in supernovae, were much more common.[71] Internal heating as a result of gravitational sorting between the core and the mantle would have caused a great deal of mantle convection, with the probable result of many more smaller and more active tectonic plates than now exist.

The time periods between such devastating environmental events give time windows for the possible origin of life in the early environments. If the deep marine hydrothermal setting was the site for the origin of life, then abiogenesis could have happened as early as 4.0 to 4.2 Ga. If the site was at the surface of the Earth, abiogenesis could only have occurred between 3.7 and 4.0 Ga.[72]

In 2016, a set of 355 genes likely present in the Last Universal Common Ancestor (LUCA) of all organisms living on Earth was identified.[73] A total of 6.1 million prokaryotic protein coding genes from various phylogenic trees were sequenced, identifying 355 protein clusters from amongst 286,514 protein clusters that were probably common to LUCA. The results "depict LUCA as anaerobic, CO_2-fixing, H_2-dependent with a Wood–Ljungdahl pathway, N_2-fixing and thermophilic. LUCA's biochemistry was replete with FeS clusters and radical reaction mechanisms. Its cofactors reveal dependence upon transition metals, flavins, S-adenosyl methionine, coenzyme A, ferredoxin, molybdopterin, corrins and selenium. Its genetic code required nucleoside modifications and S-adenosylmethionine-dependent methylations." The results depict methanogenic clostridia as a basal clade in the 355 phylogenies examined, and suggest that LUCA inhabited an anaerobic hydrothermal vent setting in a geochemically active environment rich in H_2, CO_2 and iron.[74] M.D. Brazier has shown that early micro-fossils came from a hot world of gases such as methane, ammonia, carbon dioxide and hydrogen sulphide, which are toxic to much current life.[75] Another analysis of the conventional threefold tree of life shows thermophilic and hyperthermophilic bacteria and archaea are closest to the root, suggesting that life may have evolved in a hot environment.[76]

Conceptual history

Spontaneous generation

Belief in spontaneous generation of certain forms of life from non-living matter goes back to Aristotle and ancient Greek philosophy and continued to have support in Western scholarship until the 19th century.[77] This belief was paired with a belief in heterogenesis, i.e., that one form of life derived from a different form (e.g., bees from flowers).[78] Classical notions of spontaneous generation held that certain complex, living organisms are generated by decaying organic substances. According to Aristotle, it was a readily observable truth that aphids arise from the dew that falls on plants, flies from putrid matter, mice from dirty hay, crocodiles from rotting logs at the bottom of bodies of water, and so on.[79] In the 17th century, people began to question such assumptions. In 1646, Sir Thomas Browne published his *Pseudodoxia Epidemica* (subtitled *Enquiries into Very many Received Tenets, and commonly Presumed Truths*), which was an attack on false beliefs and "vulgar errors." His contemporary, Alexander Ross, erroneously refuted him, stating: "To question this [spontaneous generation], is to question Reason, Sense, and Experience: If he doubts of this, let him go to *Ægypt*, and there he will finde the fields swarming with mice begot of the mud of *Nylus*, to the great calamity of the Inhabitants."[80][81]

In 1665, Robert Hooke published the first drawings of a microorganism. Hooke was followed in 1676 by Antonie van Leeuwenhoek, who drew and described microorganisms that are now thought to have been protozoa and bacteria.[82] Many felt the existence of microorganisms was evidence in support of spontaneous generation, since microorganisms seemed too simplistic for sexual reproduction, and asexual reproduction through cell division had not yet been observed. Van Leeuwenhoek took issue with the ideas common at the time that fleas and lice could spontaneously result from putrefaction, and that frogs could likewise arise from slime. Using a broad range of experiments ranging from sealed and open meat incubation and the close study of insect reproduction he became, by the 1680s, convinced that spontaneous generation was incorrect.[83]

The first experimental evidence against spontaneous generation came in 1668 when Francesco Redi showed that no maggots appeared in meat when flies were prevented from laying eggs. It was gradually shown that, at least in the case of all the higher and readily visible organisms, the previous sentiment regarding spontaneous generation was false. The alternative seemed to be biogenesis: that every living thing came from a pre-existing living thing (*omne vivum ex ovo*, Latin for "every living thing from an egg").

In 1768, Lazzaro Spallanzani demonstrated that microbes were present in the air, and could be killed by boiling. In 1861, Louis Pasteur performed a series of experiments that demonstrated that organisms such as bacteria and fungi do not spontaneously appear in sterile, nutrient-rich media, but could only appear by invasion from without.

The belief that self-ordering by spontaneous generation was impossible begged for an alternative. By the middle of the 19th century, the theory of biogenesis had accumulated so much evidential support, due to the work of Pasteur and others, that the alternative theory of spontaneous generation had been effectively disproven. John Desmond Bernal, a pioneer in X-ray crystallography, suggested that earlier theories such as spontaneous generation were based upon an explanation that life was continuously created as a result of chance events.[84]

Etymology

The term *biogenesis* is usually credited to either Henry Charlton Bastian or to Thomas Henry Huxley.[85] Bastian used the term around 1869 in an unpublished exchange with John Tyndall to mean "life-origination or commencement". In 1870, Huxley, as new president of the British Association for the Advancement of Science, delivered an address entitled *Biogenesis and Abiogenesis*.[86] In it he introduced the term *biogenesis* (with an opposite meaning to Bastian's) as well as *abiogenesis*:

> And thus the hypothesis that living matter always arises by the agency of pre-existing living matter, took definite shape; and had, henceforward, a right to be considered and a claim to be refuted, in each particular case, before the production of living matter in any other way could be admitted by careful reasoners. It will be necessary for me to refer to this hypothesis so frequently, that, to save circumlocution, I shall call it the hypothesis of *Biogenesis*; and I shall term the contrary doctrine–that living matter may be produced by not living matter–the hypothesis of *Abiogenesis*.[86]

Subsequently, in the preface to Bastian's 1871 book, *The Modes of Origin of Lowest Organisms*,[87] Bastian referred to the possible confusion with Huxley's usage and explicitly renounced his own meaning:

> A word of explanation seems necessary with regard to the introduction of the new term *Archebiosis*. I had originally, in unpublished writings, adopted the word *Biogenesis* to express

> the same meaning—viz., life-origination or commencement. But in the mean time the word *Biogenesis* has been made use of, quite independently, by a distinguished biologist [Huxley], who wished to make it bear a totally different meaning. He also introduced the word *Abiogenesis*. I have been informed, however, on the best authority, that neither of these words can—with any regard to the language from which they are derived—be supposed to bear the meanings which have of late been publicly assigned to them. Wishing to avoid all needless confusion, I therefore renounced the use of the word *Biogenesis*, and being, for the reason just given, unable to adopt the other term, I was compelled to introduce a new word, in order to designate the process by which living matter is supposed to come into being, independently of pre-existing living matter.[88]

Louis Pasteur and Charles Darwin

Louis Pasteur remarked, about a finding of his in 1864 which he considered definitive, "Never will the doctrine of spontaneous generation recover from the mortal blow struck by this simple experiment."[89][90] One alternative was that life's origins on Earth had come from somewhere else in the universe. Periodically resurrected (see Panspermia, above) Bernal said that this approach "is equivalent in the last resort to asserting the operation of metaphysical, spiritual entities... it turns on the argument of creation by design by a creator or demiurge."[91] Such a theory, Bernal said, was unscientific. A theory popular around the same time was that life was the result of an inner "life force", which in the late 19th century was championed by Henri Bergson.

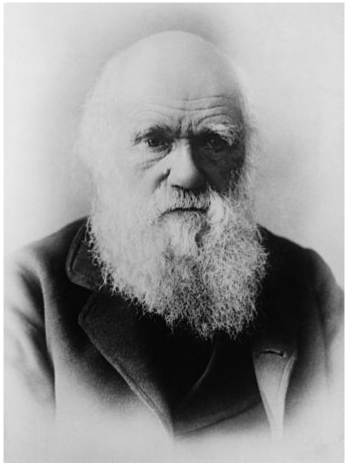
Charles Darwin in 1879

The idea of evolution by natural selection proposed by Charles Darwin put an end to these metaphysical theologies. In a letter to Joseph Dalton Hooker on 1 February 1871,[92] Darwin discussed the suggestion that the original spark of life may have begun in a "warm little pond, with all sorts of ammonia and phosphoric salts, light, heat, electricity, &c., present, that a proteine compound was chemically formed ready to undergo still more complex changes." He went on to explain that "at the present day such matter would be instantly devoured or absorbed, which would not have been the case before living creatures were formed." He had written to Hooker in 1863 stating that, "It is mere rubbish, thinking at present of the origin of life; one might as well think of the origin of matter." In *On the Origin of Species*, he had referred to life having been "created", by which he "really meant 'appeared' by some wholly unknown process", but had soon regretted using the Old Testament term "creation".[93]

"Primordial soup" hypothesis

No new notable research or hypothesis on the subject appeared until 1924, when Alexander Oparin reasoned that atmospheric oxygen prevents the synthesis of certain organic compounds that are necessary building blocks for the evolution of life. In his book *The Origin of Life*,[94][95] Oparin proposed that the "spontaneous generation of life" that had been attacked by Louis Pasteur did in fact occur once, but was now impossible because the conditions found on the early Earth had changed, and preexisting organisms would immediately consume any spontaneously generated organism. Oparin argued that a "primeval soup" of organic molecules could be created in an oxygenless atmosphere through the action of sunlight. These would combine in ever more complex ways until they formed coacervate droplets. These droplets would "grow" by fusion with other droplets, and "reproduce" through fission into daughter droplets, and so have a primitive metabolism in which factors that promote "cell integrity" survive, and those that do not become extinct. Many modern theories of the origin of life still take Oparin's ideas as a starting point.

Robert Shapiro has summarized the "primordial soup" theory of Oparin and J. B. S. Haldane in its "mature form" as follows:[96]

1. The early Earth had a chemically reducing atmosphere.
2. This atmosphere, exposed to energy in various forms, produced simple organic compounds ("monomers").
3. These compounds accumulated in a "soup" that may have concentrated at various locations (shorelines, oceanic vents etc.).
4. By further transformation, more complex organic polymers – and ultimately life – developed in the soup.

About this time, Haldane suggested that the Earth's prebiotic oceans (quite different from their modern counterparts) would have formed a "hot dilute soup" in which organic compounds could have formed. Bernal called this idea *biopoiesis* or *biopoesis*, the process of living matter evolving from self-replicating but non-living molecules,[84][97] and proposed that biopoiesis passes through a number of intermediate stages.

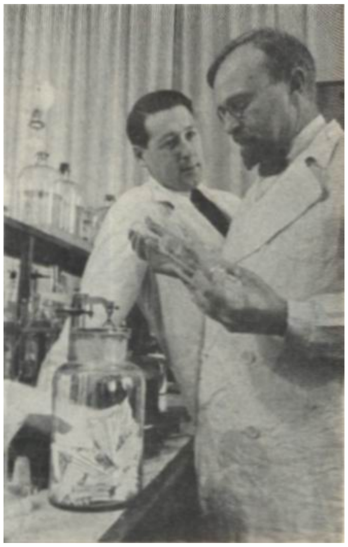

Alexander Oparin (right) in his laboratory, 1938

One of the most important pieces of experimental support for the "soup" theory came in 1952. Stanley L. Miller and Harold C. Urey performed an experiment that demonstrated how organic molecules could have spontaneously formed from inorganic precursors under conditions like those posited by the Oparin-Haldane hypothesis. The now-famous Miller–Urey experiment used a highly reducing mixture of gases – methane, ammonia, and hydrogen, as well as water vapour – to form simple organic monomers such as amino acids.[98] The mixture of gases was cycled through an apparatus that delivered electrical sparks to the mixture. After one week, it was found that about 10% to 15% of the carbon in the system was then in the form of a racemic mixture of organic compounds, including amino acids, which are the building blocks of proteins. This provided direct experimental support for the second point of the "soup" theory, and it is around the remaining two points of the theory that much of the debate now centres.

Bernal showed that based upon this and subsequent work there is no difficulty in principle in forming most of the molecules we recognize as the necessary molecules for life from their inorganic precursors. The underlying hypothesis held by Oparin, Haldane, Bernal, Miller and Urey, for instance, was that multiple conditions on the primeval Earth favoured chemical reactions that synthesized the same set of complex organic compounds from such simple precursors. A 2011 reanalysis of the saved vials containing the original extracts that resulted from the Miller and Urey experiments, using current and more advanced analytical equipment and technology, has uncovered more biochemicals than originally discovered in the 1950s. One of the more important findings was 23 amino acids, far more than the five originally found.[99] However, Bernal said that "it is not enough to explain the formation of such molecules, what is necessary, is a physical-chemical explanation of the origins of these molecules that suggests the presence of suitable sources and sinks for free energy."[100]

More recent studies, in October 2017, support the notion that life may have begun right after the Earth was formed as RNA molecules emerging from "warm little ponds".[46]

Proteinoid microspheres

In trying to uncover the intermediate stages of abiogenesis mentioned by Bernal, Sidney W. Fox in the 1950s and 1960s studied the spontaneous formation of peptide structures (small chains of amino acids) under conditions that might plausibly have existed early in Earth's history. In one of his experiments, he allowed amino acids to dry out as if puddled in a warm, dry spot in prebiotic conditions. He found that, as they dried, the amino acids formed long, often cross-linked, thread-like, submicroscopic polypeptide molecules now named "proteinoid microspheres".[101]

In another experiment to set suitable conditions for life to form, Fox collected volcanic material from a cinder cone in Hawaii. He discovered that the temperature was over 100 °C (212 °F) just 4 inches (100 mm) beneath the surface of the cinder cone, and suggested that this might have been the environment in which life was created—molecules could have formed and then been washed through the loose volcanic ash into the sea. He placed lumps of lava over amino acids derived from methane, ammonia and water, sterilized all materials, and baked the lava over the amino acids for a few hours in a glass oven. A brown, sticky substance formed over the surface, and when the lava was drenched in sterilized water, a thick, brown liquid leached out. The amino acids had combined to form proteinoids, and the proteinoids had combined to form small globules that Fox called "microspheres". His proteinoids were not cells, although they formed clumps and chains reminiscent of cyanobacteria, but they contained no functional nucleic acids or any encoded information. Based upon such experiments, Colin S. Pittendrigh stated in December 1967 that "laboratories will be creating a living cell within ten years," a remark that reflected the typical contemporary naivety about the complexity of cell structures.[102]

Current models

There is no single, generally accepted model for the origin of life. Scientists have proposed several plausible hypotheses, which share some common elements. While differing in the details, these hypotheses are based on the framework laid out by Alexander Oparin (in 1924) and by J. B. S. Haldane (in 1925), who postulated the molecular or chemical evolution theory of life.[103] According to them, the first molecules constituting the earliest cells "were synthesized under natural conditions by a slow process of molecular evolution, and these molecules then organized into the first molecular system with properties with biological order".[103] Oparin and Haldane suggested that the atmosphere of the early Earth may have been chemically reducing in nature, composed primarily of methane (CH_4), ammonia (NH_3), water (H_2O), hydrogen sulfide (H_2S), carbon dioxide (CO_2) or carbon monoxide (CO), and phosphate (PO_4^{3-}), with molecular oxygen (O_2) and ozone (O_3) either rare or absent. According to later models, the atmosphere in the late Hadean period consisted largely of nitrogen (N_2) and carbon dioxide, with smaller amounts of carbon monoxide, hydrogen (H_2), and sulfur compounds;[104] while it did lack molecular oxygen and ozone,[105] it was not as chemically reducing as Oparin and Haldane supposed. In the atmosphere proposed by Oparin and Haldane, electrical activity can produce certain small molecules (monomers) of life, such as amino acids. The Miller–Urey experiment reported in 1953 demonstrated this.

Bernal coined the term *biopoiesis* in 1949 to refer to the origin of life.[106] In 1967, he suggested that it occurred in three "stages":

1. the origin of biological monomers
2. the origin of biological polymers
3. the evolution from molecules to cells

Bernal suggested that evolution commenced between stages 1 and 2. Bernal regarded the third stage – discovering methods by which biological reactions were incorporated behind a cell's boundary – as the most difficult. Modern work on the way that cell membranes self-assemble, and the work on micropores in various substrates may be a halfway house towards the development of independent free-living cells.[107][108][109]

The chemical processes that took place on the early Earth are called *chemical evolution*. Since the end of the nineteenth century, 'evolutive abiogenesis' means increasing complexity and evolution of matter from inert to living state.[110] Both Manfred Eigen and Sol Spiegelman demonstrated that evolution, including replication, variation, and natural selection, can occur in populations of molecules as well as in organisms.[49] Spiegelman took advantage of natural selection to synthesize the Spiegelman Monster, which had a genome with just 218 nucleotide bases, having deconstructively evolved from a 4500-base bacterial RNA. Eigen built on Spiegelman's work and produced a similar system further degraded to just 48 or 54 nucleotides – the minimum required for the binding of the replication enzyme.[111]

Following on from chemical evolution came the initiation of biological evolution, which led to the first cells.[49] No one has yet synthesized a "protocell" using simple components with the necessary properties of life (the so-called "bottom-up-approach"). Without such a proof-of-principle, explanations have tended to focus on chemosynthesis.[112] However, some researchers work in this field, notably Steen Rasmussen and Jack W. Szostak. Others have argued that a "top-down approach" is more feasible. One such approach, successfully attempted by Craig Venter and others at J. Craig Venter Institute, involves engineering existing prokaryotic cells with progressively fewer genes, attempting to discern at which point the most minimal requirements for life are reached.[113][114][115]

The NASA strategy on abiogenesis states that it is necessary to identify interactions, intermediary structures and functions, energy sources, and environmental factors that contributed to the diversity, selection, and replication of evolvable macromolecular systems.[116] Emphasis must continue to map the chemical landscape of potential primordial informational polymers. The advent of polymers that could replicate, store genetic information, and exhibit properties subject to selection likely was a critical step in the emergence of prebiotic chemical evolution.[116]

In October 2018, researchers at McMaster University announced the development of a new technology, called a *Planet Simulator*, to help study the origin of life on planet Earth and beyond.[117][118][119][120] It consists of a sophisticated climate chamber to study how the building blocks of life were assembled and how these prebiotic molecules transitioned into self-replicating RNA molecules.[117]

Chemical origin of organic molecules

The elements, except for hydrogen and helium, ultimately derive from stellar nucleosynthesis. On 12 October 2016, astronomers reported that the very basic chemical ingredients of life — the carbon-hydrogen molecule (CH, or methylidyne radical), the carbon-hydrogen positive ion (CH+) and the carbon ion (C+) — are largely the result of ultraviolet light from stars, rather than other forms of radiation from supernovae and young stars, as thought earlier.[121] Complex molecules, including organic molecules, form naturally both in space and on planets.[20] There are two possible sources of organic molecules on the early Earth:

1. Terrestrial origins – organic molecule synthesis driven by impact shocks or by other energy sources (such as UV light, redox coupling, or electrical discharges; e.g., Miller's experiments)
2. Extraterrestrial origins – formation of organic molecules in interstellar dust clouds, which rain down on planets.[122][123] (See pseudo-panspermia)

Based on recent computer model studies, the complex organic molecules necessary for life may have formed in the protoplanetary disk of dust grains surrounding the Sun before the formation of the Earth.[124][125] According to the computer studies, this same process may also occur around other stars that acquire planets. (Also see Extraterrestrial organic molecules).

Estimates of the production of organics from these sources suggest that the Late Heavy Bombardment before 3.5 Ga within the early atmosphere made available quantities of organics comparable to those produced by terrestrial sources.[126][127]

It has been estimated that the Late Heavy Bombardment may also have effectively sterilized the Earth's surface to a depth of tens of metres. If life evolved deeper than this, it would have also been shielded from the early high levels of ultraviolet radiation from the T Tauri stage of the Sun's evolution. Simulations of geothermically heated oceanic crust yield far more organics than those found in the Miller-Urey experiments (see below). In the deep hydrothermal vents, Everett Shock has found "there is an enormous thermodynamic drive to form organic compounds, as seawater and hydrothermal fluids, which are far from equilibrium, mix and move towards a more stable state."[128] Shock has found that the available energy is maximized at around 100 – 150 degrees Celsius, precisely the temperatures at which the hyperthermophilic bacteria and thermoacidophilic archaea have been found, at the base of the phylogenetic tree of life closest to the Last Universal Common Ancestor (LUCA).[129]

The accumulation and concentration of organic molecules on a planetary surface is also considered an essential early step for the origin of life.[116] Identifying and understanding the mechanisms that led to the production of prebiotic molecules in various environments is critical for establishing the inventory of ingredients from which life originated on Earth, assuming that the abiotic production of molecules ultimately influenced the selection of molecules from which life emerged.[116]

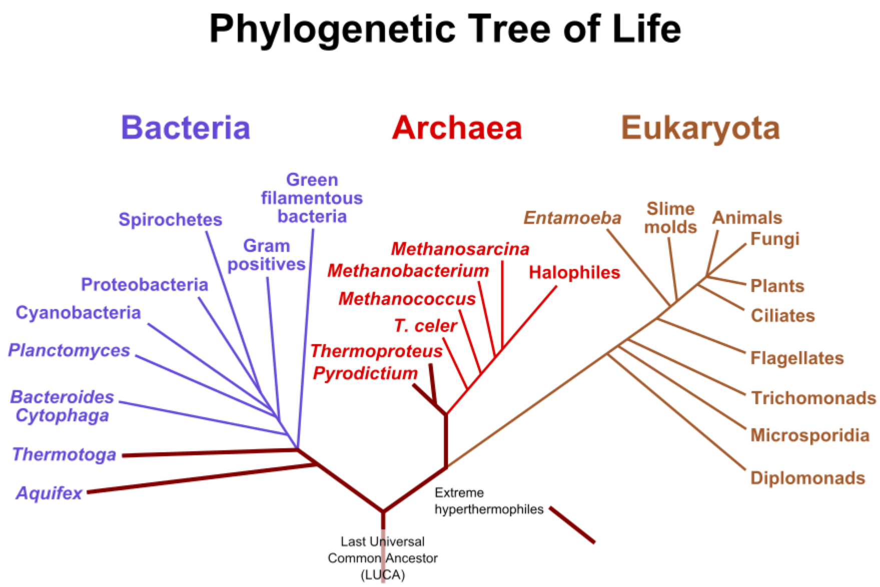
Phylogenetic Tree of Life
Bacteria
Archaea
Eukaryota
Green filamentous bacteria
Spirochetes
Gram positives
Proteobacteria
Cyanobacteria
Planctomyces
Bacteroides Cytophaga
Thermotoga
Aquifex
Methanosarcina
Methanobacterium
Halophiles
Methanococcus
T. celer
Thermoproteus
Pyrodictium
Entamoeba
Slime molds
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia
Diplomonads
Extreme hyperthermophiles
Last Universal Common Ancestor (LUCA)

A cladogram demonstrating extreme hyperthermophiles at the base of the phylogenetic tree of life.

Chemical synthesis

While features of self-organization and self-replication are often considered the hallmark of living systems, there are many instances of abiotic molecules exhibiting such characteristics under proper conditions. Stan Palasek suggested based on a theoretical model that self-assembly of ribonucleic acid (RNA) molecules can occur spontaneously due to physical factors in hydrothermal vents.[130] Virus self-assembly within host cells has implications for the study of the origin of life,[131] as it lends further credence to the hypothesis that life could have started as self-assembling organic molecules.[132][133]

Multiple sources of energy were available for chemical reactions on the early Earth. For example, heat (such as from geothermal processes) is a standard energy source for chemistry. Other examples include sunlight and electrical discharges (lightning), among others.[49] Computer simulations also suggest that cavitation in primordial water reservoirs such as breaking sea waves, streams and oceans can potentially lead to the synthesis of biogenic compounds.[134] Unfavourable reactions can also be driven by highly favourable ones, as in the case of iron-sulfur chemistry. For example, this was probably important for carbon fixation (the conversion of carbon from its inorganic form to an organic one).[note 2] Carbon fixation via iron-sulfur chemistry is highly favourable, and occurs at neutral pH and 100 °C (212 °F). Iron-sulfur surfaces, which are abundant near hydrothermal vents, are also capable of producing small amounts of amino acids and other biological metabolites.[49]

As early as the 1860s, experiments have demonstrated that biologically relevant molecules can be produced from interaction of simple carbon sources with abundant inorganic catalysts. In particular, experiments by Butlerov (the formose reaction) showed that tetroses, pentoses, and hexoses are produced when formaldehyde is heated under basic conditions with divalent metal ions like calcium. The reaction was scrutinized and subsequently proposed to be autocatalytic by Breslow in 1959. Similar experiments (see below) demonstrate that nucleobases like guanine and adenine could be synthesized from simple carbon and nitrogen sources like hydrogen cyanide and ammonia.

Formamide produces all four ribonucleotides and other biological molecules when warmed in the presence of various terrestrial minerals. Formamide is ubiquitous in the Universe, produced by the reaction of water and hydrogen cyanide (HCN). It has several advantages as a biotic precursor, including the ability to easily become concentrated through the evaporation of water.[135][136] Although HCN is poisonous, it only affects aerobic organisms (eukaryotes and aerobic bacteria), which did not yet exist. It can play roles in other chemical processes as well, such as the synthesis of the amino acid glycine.[49]

In 1961, it was shown that the nucleic acid purine base adenine can be formed by heating aqueous ammonium cyanide solutions.[137] Other pathways for synthesizing bases from inorganic materials were also reported.[138] Leslie E. Orgel and colleagues have shown that freezing temperatures are advantageous for the synthesis of purines, due to the concentrating effect for key precursors such as hydrogen cyanide.[139] Research by Stanley L. Miller and colleagues suggested that while adenine and guanine require freezing conditions for synthesis, cytosine and uracil may require boiling temperatures.[140] Research by the Miller group notes the formation of seven different amino acids and 11 types of nucleobases in ice when ammonia and cyanide were left in a freezer from 1972 to 1997.[141][142] Other work demonstrated the formation of s-triazines (alternative nucleobases), pyrimidines (including cytosine and uracil), and adenine from urea solutions subjected to freeze-thaw cycles under a reductive atmosphere (with spark discharges as an energy source).[143] The explanation given for the unusual speed of these reactions at such a low temperature is eutectic freezing. As an ice crystal forms, it stays pure: only molecules of water join the growing crystal, while impurities like salt or cyanide are excluded. These impurities become crowded in microscopic pockets of liquid within the ice, and this crowding causes the molecules to collide more often. Mechanistic exploration using quantum chemical methods provide a more detailed understanding of some of the chemical processes involved in chemical evolution, and a partial answer to the fundamental question of molecular biogenesis.[144]

At the time of the Miller–Urey experiment, scientific consensus was that the early Earth had a reducing atmosphere with compounds relatively rich in hydrogen and poor in oxygen (e.g., CH_4 and NH_3 as opposed to CO_2 and nitrogen dioxide (NO_2)). However, current scientific consensus describes the primitive atmosphere as either weakly reducing or neutral[145][146] (see also Oxygen Catastrophe). Such an atmosphere would diminish both the amount and variety of amino acids that could be produced, although studies that include iron and carbonate minerals (thought present in early oceans) in the experimental conditions have again produced a diverse array of amino acids.[145] Other scientific research has focused on two other potential reducing environments: outer space and deep-sea thermal vents.[147][148][149]

The spontaneous formation of complex polymers from abiotically generated monomers under the conditions posited by the "soup" theory is not at all a straightforward process. Besides the necessary basic organic monomers, compounds that would have prohibited the formation of polymers were also formed in high concentration during the Miller–Urey and Joan Oró experiments.[150] The Miller–Urey experiment, for example, produces many substances that would react with the amino acids or terminate their coupling into peptide chains.[151]

A research project completed in March 2015 by John D. Sutherland and others found that a network of reactions beginning with hydrogen cyanide and hydrogen sulfide, in streams of water irradiated by UV light, could produce the chemical components of proteins and lipids, as well as those of RNA,[152][153] while not producing a wide range of other compounds.[154] The researchers used the term "cyanosulfidic" to describe this network of reactions.[153]

Autocatalysis

Autocatalysts are substances that catalyze the production of themselves and therefore are "molecular replicators." The simplest self-replicating chemical systems are autocatalytic, and typically contain three components: a product molecule and two precursor molecules. The product molecule joins together the precursor molecules, which in turn produce more product molecules from more precursor molecules. The product molecule catalyzes the reaction by providing a complementary template that binds to the precursors, thus bringing them together. Such systems have been demonstrated both in biological macromolecules and in small organic molecules.[155][156] Systems that do not proceed by template mechanisms, such as the self-reproduction of micelles and vesicles, have also been observed.[156]

It has been proposed that life initially arose as autocatalytic chemical networks.[157] British ethologist Richard Dawkins wrote about autocatalysis as a potential explanation for the origin of life in his 2004 book *The Ancestor's Tale*.[158] In his book, Dawkins cites experiments performed by Julius Rebek Jr. and his colleagues in which they combined amino adenosine and pentafluorophenyl esters with the autocatalyst amino adenosine triacid ester (AATE). One product was a variant of AATE, which catalyzed the synthesis of themselves. This experiment demonstrated the possibility that autocatalysts could exhibit competition within a population of entities with heredity, which could be interpreted as a rudimentary form of natural selection.[159][160]

In the early 1970s, Manfred Eigen and Peter Schuster examined the transient stages between the molecular chaos and a self-replicating hypercycle in a prebiotic soup.[161] In a hypercycle, the information storing system (possibly RNA) produces an enzyme, which catalyzes the formation of another information system, in sequence until the product of the last aids in the formation of the first information system. Mathematically treated, hypercycles could create quasispecies, which through natural selection entered into a form of Darwinian evolution. A boost to hypercycle theory was the discovery of ribozymes capable of catalyzing their own chemical reactions. The hypercycle theory requires the existence of complex biochemicals, such as nucleotides, which do not form under the conditions proposed by the Miller–Urey experiment.

Geoffrey W. Hoffmann has shown that an early error-prone translation machinery can be stable against an error catastrophe of the type that had been envisaged as problematical for the origin of life, and was known as "Orgel's paradox".[162][163][164]

Hoffmann has furthermore argued that a complex nucleation event as the origin of life involving both polypeptides and nucleic acid is compatible with the time and space available in the primitive oceans of Earth[165] Hoffmann suggests that volcanic ash may provide the many random shapes needed in the postulated complex nucleation event. This aspect of the theory can be tested experimentally.

Homochirality

Homochirality refers to a geometric uniformity of some materials composed of chiral units. Chiral refers to nonsuperimposable 3D forms that are mirror images of one another, as are left and right hands. Living organisms use molecules that have the same chirality ("handedness"): with almost no exceptions,[166] amino acids are left-handed while nucleotides and sugars are right-handed. Chiral molecules can be synthesized, but in the absence of a chiral source or a chiral catalyst, they are formed in a 50/50 mixture of both enantiomers (called a racemic mixture). Known mechanisms for the production of non-racemic mixtures from racemic starting materials include: asymmetric physical laws, such as the electroweak interaction; asymmetric environments, such as those caused by circularly polarized light, quartz crystals, or the Earth's rotation, statistical fluctuations during racemic synthesis,[167] and spontaneous symmetry breaking.[168][169][170]

Once established, chirality would be selected for.[171] A small bias (enantiomeric excess) in the population can be amplified into a large one by asymmetric autocatalysis, such as in the Soai reaction.[172] In asymmetric autocatalysis, the catalyst is a chiral molecule, which means that a chiral molecule is catalyzing its own production. An initial enantiomeric excess, such as can be produced by polarized light, then allows the more abundant enantiomer to outcompete the other.[173]

Clark has suggested that homochirality may have started in outer space, as the studies of the amino acids on the Murchison meteorite showed that L-alanine is more than twice as frequent as its D form, and L-glutamic acid was more than three times prevalent than its D counterpart. Various chiral crystal surfaces can also act as sites for possible concentration and assembly of chiral monomer units into macromolecules.[174][175] Compounds found on meteorites suggest that the chirality of life derives from abiogenic synthesis, since amino acids from meteorites show a left-handed bias, whereas sugars show a predominantly right-handed bias, the same as found in living organisms.[176]

Self-enclosement, reproduction, duplication and the RNA world

Protocells

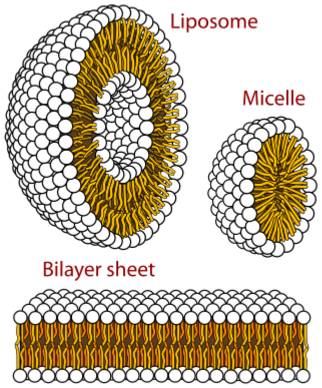

The three main structures phospholipids form spontaneously in solution: the liposome (a closed bilayer), the micelle and the bilayer.

A protocell is a self-organized, self-ordered, spherical collection of lipids proposed as a stepping-stone to the origin of life.[177] A central question in evolution is how simple protocells first arose and differed in reproductive contribution to the following generation driving the evolution of life. Although a functional protocell has not yet been achieved in a laboratory setting, there are scientists who think the goal is well within reach.[178][179][180]

Self-assembled vesicles are essential components of primitive cells.[177] The second law of thermodynamics requires that the universe move in a direction in which entropy increases, yet life is distinguished by its great degree of organization. Therefore, a boundary is needed to separate life processes from non-living matter.[181] Researchers Irene A. Chen and Jack W. Szostak amongst others, suggest that simple physicochemical properties of elementary protocells can give rise to essential cellular behaviours, including primitive forms of differential reproduction competition and energy storage. Such cooperative interactions between the membrane and its encapsulated contents could greatly simplify the transition from simple replicating molecules to true cells.[179] Furthermore, competition for membrane molecules would favour stabilized membranes, suggesting a selective advantage for the evolution of cross-linked fatty acids and even the phospholipids of today.[179] Such micro-encapsulation would allow for metabolism within the membrane, the exchange of small molecules but the prevention of passage of large substances across it.[182] The main advantages of encapsulation include the increased solubility of the contained cargo within the capsule and the storage of energy in the form of a electrochemical gradient.

A 2012 study led by Armen Y. Mulkidjanian of Germany's University of Osnabrück, suggests that inland pools of condensed and cooled geothermal vapour have the ideal characteristics for the origin of life.[183] Scientists confirmed in 2002 that by adding a montmorillonite clay to a solution of fatty acid micelles (lipid spheres), the clay sped up the rate of vesicles formation 100-fold.[180]

Another protocell model is the Jeewanu. First synthesized in 1963 from simple minerals and basic organics while exposed to sunlight, it is still reported to have some metabolic capabilities, the presence of semipermeable membrane, amino acids, phospholipids, carbohydrates and RNA-like molecules.[184][185] However, the nature and properties of the Jeewanu remains to be clarified.

Electrostatic interactions induced by short, positively charged, hydrophobic peptides containing 7 amino acids in length or fewer, can attach RNA to a vesicle membrane, the basic cell membrane.[186]

RNA world

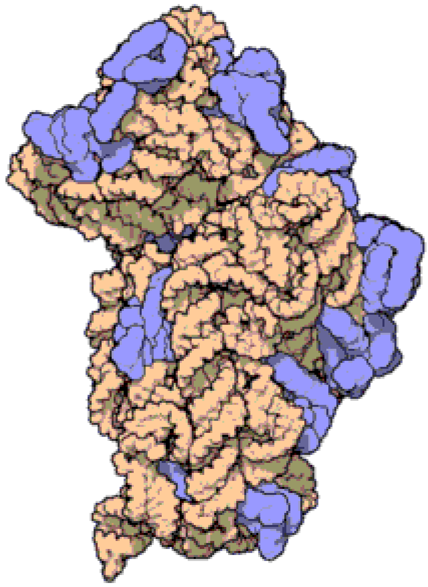
Molecular structure of the ribosome 30S subunit from *Thermus thermophilus*.[187] Proteins are shown in blue and the single RNA chain in orange.

The RNA world hypothesis describes an early Earth with self-replicating and catalytic RNA but no DNA or proteins.[188] It is widely accepted that current life on Earth descends from an RNA world,[16][189] although RNA-based life may not have been the first life to exist.[17][18] This conclusion is drawn from many independent lines of evidence, such as the observations that RNA is central to the translation process and that small RNAs can catalyze all of the chemical groups and information transfers required for life.[18][190] The structure of the ribosome has been called the "smoking gun," as it showed that the ribosome is a ribozyme, with a central core of RNA and no amino acid side chains within 18 angstroms of the active site where peptide bond formation is catalyzed.[17] The concept of the RNA world was first proposed in 1962 by Alexander Rich,[191] and the term was coined by Walter Gilbert in 1986.[18][192]

Possible precursors for the evolution of protein synthesis include a mechanism to synthesize short peptide cofactors or form a mechanism for the duplication of RNA. It is likely that the ancestral ribosome was composed entirely of RNA, although some roles have since been taken over by proteins. Major remaining questions on this topic include identifying the selective force for the evolution of the ribosome and determining how the genetic code arose.[193]

Eugene Koonin said, "Despite considerable experimental and theoretical effort, no compelling scenarios currently exist for the origin of replication and translation, the key processes that together comprise the core of biological systems and the apparent pre-requisite of biological evolution. The RNA World concept might offer the best chance for the resolution of this conundrum but so far cannot adequately account for the emergence of an efficient RNA replicase or the translation system. The MWO ["many worlds in one"] version of the cosmological model of eternal inflation could suggest a way out of this conundrum because, in an infinite multiverse with a finite number of distinct macroscopic histories (each repeated an infinite number of times), emergence of even highly complex systems by chance is not just possible but inevitable."[194]

Viral origins

Recent evidence for a "virus first" hypothesis, which may support theories of the RNA world, has been suggested.[195] [196] One of the difficulties for the study of the origins of viruses is their high rate of mutation; this is particularly the case in RNA retroviruses like HIV.[197] A 2015 study compared protein fold structures across different branches of the tree of life, where researchers can reconstruct the evolutionary histories of the folds and of the organisms whose genomes code for those folds. They argue that protein folds are better markers of ancient events as their three-dimensional structures can be maintained even as the sequences that code for those begin to change.[195] Thus, the viral protein repertoire retain traces of ancient evolutionary history that can be recovered using advanced bioinformatics approaches. Those researchers think that "the prolonged pressure of genome and particle size reduction eventually reduced virocells into modern viruses (identified by the complete loss of cellular makeup), meanwhile other coexisting cellular lineages diversified into modern cells.[198] The data suggest that viruses originated from ancient cells that co-existed with the ancestors of modern cells.

These ancient cells likely contained segmented RNA genomes.[195][199] Although the virus-first hypothesis is highly controversial today, some astrobiologists have suggested looking for viruses on other celestial bodies such as Mars if they do emerge before cells.[200]

RNA synthesis and replication

A number of hypotheses of formation of RNA have been put forward. As of 1994, there were difficulties in the explanation of the abiotic synthesis of the nucleotides cytosine and uracil.[201] Subsequent research has shown possible routes of synthesis; for example, formamide produces all four ribonucleotides and other biological molecules when warmed in the presence of various terrestrial minerals.[135][136] Early cell membranes could have formed spontaneously from proteinoids, which are protein-like molecules produced when amino acid solutions are heated while in the correct concentration of aqueous solution. These are seen to form micro-spheres which are observed to behave similarly to membrane-enclosed compartments. Other possible means of producing more complicated organic molecules include chemical reactions that take place on clay substrates or on the surface of the mineral pyrite.

Factors supporting an important role for RNA in early life include its ability to act both to store information and to catalyze chemical reactions (as a ribozyme); its many important roles as an intermediate in the expression of and maintenance of the genetic information (in the form of DNA) in modern organisms; and the ease of chemical synthesis of at least the components of the RNA molecule under the conditions that approximated the early Earth. Relatively short RNA molecules have been synthesized, capable of replication.[202] Such replicase RNA, which functions as both code and catalyst provides its own template upon which copying can occur. Jack W. Szostak has shown that certain catalytic RNAs can join smaller RNA sequences together, creating the potential for self-replication. If these conditions were present, Darwinian natural selection would favour the proliferation of such autocatalytic sets, to which further functionalities could be added.[203] Such autocatalytic systems of RNA capable of self-sustained replication have been identified.[204] The RNA replication systems, which include two ribozymes that catalyze each other's synthesis, showed a doubling time of the product of about one hour, and were subject to natural selection under the conditions that existed in the experiment.[205] In evolutionary competition experiments, this led to the emergence of new systems which replicated more efficiently.[17] This was the first demonstration of evolutionary adaptation occurring in a molecular genetic system.[205]

Depending on the definition, life started when RNA chains began to self-replicate, initiating the three mechanisms of Darwinian selection: heritability, variation of type, and differential reproductive output. The fitness of an RNA replicator (its per capita rate of increase) would likely be a function of its intrinsic adaptive capacities, determined by its nucleotide sequence, and the availability of resources.[206][207] The three primary adaptive capacities may have been: (1) replication with moderate fidelity, giving rise to both heritability while allowing variation of type, (2) resistance to decay, and (3) acquisition of process resources.[206][207] These capacities would have functioned by means of the folded configurations of the RNA replicators resulting from their nucleotide sequences.

Carl Zimmer has speculated that the chemical conditions, including the presence of boron, molybdenum and oxygen needed for the initial production of RNA, may have been better on early Mars than on early Earth.[208][209][210] If so, life-suitable molecules originating on Mars may have later migrated to Earth via meteor ejections.

Pre-RNA world

It is possible that a different type of nucleic acid, such as PNA, TNA or GNA, was the first to emerge as a self-reproducing molecule, only later replaced by RNA.[211][212] Larralde *et al.*, say that "the generally accepted prebiotic synthesis of ribose, the formose reaction, yields numerous sugars without any selectivity."[213] and they conclude that their "results suggest that the backbone of the first genetic material could not have contained ribose or other sugars because of their instability." The ester linkage of ribose and phosphoric acid in RNA is known to be prone to hydrolysis.[214]

Pyrimidine ribonucleosides and their respective nucleotides have been prebiotically synthesized by a sequence of reactions which by-pass the free sugars, and are assembled in a stepwise fashion by using nitrogenous or oxygenous chemistries. Sutherland has demonstrated high yielding routes to cytidine and uridine ribonucleotides built from small 2 and 3 carbon fragments such as glycolaldehyde, glyceraldehyde or glyceraldehyde-3-phosphate, cyanamide and cyanoacetylene. One of the steps in this sequence allows the isolation of enantiopure ribose aminooxazoline if the enantiomeric excess of glyceraldehyde is 60% or greater.[215] This can be viewed as a prebiotic purification step, where the said compound spontaneously crystallized out from a mixture of the other pentose aminooxazolines. Ribose aminooxazoline can then react with cyanoacetylene in a mild and highly efficient manner to give the alpha cytidine ribonucleotide. Photoanomerization with UV light allows for inversion about the 1' anomeric centre to give the correct beta stereochemistry.[216] In 2009 they showed that the same simple building blocks allow access, via phosphate controlled nucleobase elaboration, to 2',3'-cyclic pyrimidine nucleotides directly, which are known to be able to polymerize into RNA. This paper also highlights the possibility for the photo-sanitization of the pyrimidine-2',3'-cyclic phosphates.[217]

Origin of biological metabolism

Metabolism-like reactions could have occurred naturally in early oceans, before the first organisms evolved.[19][218] Metabolism may predate the origin of life, which may have evolved from the chemical conditions in the earliest oceans. Reconstructions in laboratories show that some of these reactions can produce RNA, and some others resemble two essential reaction cascades of metabolism: glycolysis and the pentose phosphate pathway, that provide essential precursors for nucleic acids, amino acids and lipids.[218] A study at the University of Düsseldorf created phylogenic trees based upon 6 million genes from bacteria and archaea, and identified 355 protein families that were probably present in the LUCA. They were based upon an anaerobic metabolism fixing carbon dioxide and nitrogen. It suggests that the LUCA evolved in an environment rich in hydrogen, carbon dioxide and iron.[219] Following are some observed discoveries and related hypotheses.

Iron–sulfur world

In the 1980s, Günter Wächtershäuser, encouraged and supported by Karl R. Popper,[220][221][222] postulated his iron–sulfur world, a theory of the evolution of pre-biotic chemical pathways as the starting point in the evolution of life. It systematically traces today's biochemistry to primordial reactions which provide alternative pathways to the synthesis of organic building blocks from simple gaseous compounds.

In contrast to the classical Miller experiments, which depend on external sources of energy (simulated lightning, ultraviolet irradiation), "Wächtershäuser systems" come with a built-in source of energy: sulfides of iron (iron pyrite) and other minerals. The energy released from redox reactions of these metal sulfides is available for the synthesis of organic molecules, and such systems may have evolved into autocatalytic sets constituting self-replicating, metabolically active entities predating the life forms known today.[19][218] Experiments with such sulfides in an aqueous environment at 100 °C produced a relatively small yield of dipeptides (0.4% to 12.4%) and a smaller yield of tripeptides (0.10%) although under the same conditions, dipeptides were quickly broken down.[223]

Several models reject the self-replication of a "naked-gene", postulating instead the emergence of a primitive metabolism providing a safe environment for the later emergence of RNA replication. The centrality of the Krebs cycle (citric acid cycle) to energy production in aerobic organisms, and in drawing in carbon dioxide and hydrogen ions in biosynthesis of complex organic chemicals, suggests that it was one of the first parts of the metabolism to evolve.[224] Concordantly, geochemist Michael Russell has proposed that "the purpose of life is to hydrogenate carbon dioxide" (as part of a "metabolism-first," rather than a "genetics-first," scenario).[225][226] Physicist Jeremy England of MIT has proposed that life was inevitable from general thermodynamic considerations: "... when a group of atoms is driven by an external source of energy (like the sun or chemical fuel) and surrounded by a heat bath (like the ocean or atmosphere), it will often gradually restructure itself in order to dissipate increasingly more energy. This could mean that under certain conditions, matter inexorably acquires the key physical attribute associated with life."[227][228]

One of the earliest incarnations of this idea was put forward in 1924 with Oparin's notion of primitive self-replicating vesicles which predated the discovery of the structure of DNA. Variants in the 1980s and 1990s include Wächtershäuser's iron–sulfur world theory and models introduced by Christian de Duve based on the chemistry of thioesters. More abstract and theoretical arguments for the plausibility of the emergence of metabolism without the presence of genes include a mathematical model introduced by Freeman Dyson in the early 1980s and Stuart Kauffman's notion of collectively autocatalytic sets, discussed later that decade.

Orgel summarized his analysis by stating, "There is at present no reason to expect that multistep cycles such as the reductive citric acid cycle will self-organize on the surface of FeS/FeS_2 or some other mineral."[229] It is possible that another type of metabolic pathway was used at the beginning of life. For example, instead of the reductive citric acid cycle, the "open" acetyl-CoA pathway (another one of the five recognized ways of carbon dioxide fixation in nature today) would be compatible with the idea of self-organization on a metal sulfide surface. The key enzyme of this pathway, carbon monoxide dehydrogenase/acetyl-CoA synthase, harbours mixed nickel-iron-sulfur clusters in its reaction centres and catalyzes the formation of acetyl-CoA (similar to acetyl-thiol) in a single step. There are increasing concerns, however, that prebiotic thiolated and thioester compounds are thermodynamically and kinetically unfavourable to accumulate in presumed prebiotic conditions (i.e. hydrothermal vents).[230] It has also been proposed that cysteine and homocysteine may have reacted with nitriles resulting from the Stecker reaction, readily forming catalytic thiol-reach poplypeptides.[231]

Zn-world hypothesis

The Zn-world (zinc world) theory of Armen Y. Mulkidjanian[232] is an extension of Wächtershäuser's pyrite hypothesis. Wächtershäuser based his theory of the initial chemical processes leading to informational molecules (RNA, peptides) on a regular mesh of electric charges at the surface of pyrite that may have facilitated the primeval polymerization by attracting reactants and arranging them appropriately relative to each other.[233] The Zn-world theory specifies and differentiates further.[232][234] Hydrothermal fluids rich in H_2S interacting with cold primordial ocean (or Darwin's "warm little pond") water leads to the precipitation of metal sulfide particles. Oceanic vent systems and other hydrothermal systems have a zonal structure reflected in ancient volcanogenic massive sulfide deposits (VMS) of hydrothermal origin. They reach many kilometres in diameter and date back to the Archean Eon. Most abundant are pyrite (FeS_2), chalcopyrite ($CuFeS_2$), and sphalerite (ZnS), with additions of galena (PbS) and alabandite (MnS). ZnS and MnS have a unique ability to store radiation energy, e.g. from UV light. During the relevant time window of the origins of replicating molecules, the primordial atmospheric pressure was high enough (>100 bar, about 100 atmospheres) to precipitate near the Earth's surface, and UV irradiation was 10 to 100 times more intense than now; hence the unique photosynthetic properties mediated by ZnS provided just the right energy conditions to energize the synthesis of informational and metabolic molecules and the selection of photostable nucleobases.

The Zn-world theory has been further filled out with experimental and theoretical evidence for the ionic constitution of the interior of the first proto-cells before archaea, bacteria and proto-eukaryotes evolved. Archibald Macallum noted the resemblance of body fluids such as blood and lymph to seawater;[235] however, the inorganic composition of all cells differ from that of modern seawater, which led Mulkidjanian and colleagues to reconstruct the "hatcheries" of the first cells combining geochemical analysis with phylogenomic scrutiny of the inorganic ion requirements of universal components of modern cells. The authors conclude that ubiquitous, and by inference primordial, proteins and functional systems show affinity to and functional requirement for K^+, Zn^{2+}, Mn^{2+}, and phosphate. Geochemical reconstruction shows that the ionic composition conducive to the origin of cells could not have existed in what we today call marine settings but is compatible with emissions of vapour-dominated zones of what we today call inland geothermal systems. Under the oxygen depleted, CO_2-dominated primordial atmosphere, the chemistry of water condensates and exhalations near geothermal fields would resemble the internal milieu of modern cells. Therefore, the precellular stages of evolution may have taken place in shallow "Darwin ponds" lined with porous silicate minerals mixed with metal sulfides and enriched in K^+, Zn^{2+}, and phosphorus compounds.[236][237]

Deep sea vent hypothesis

The deep sea vent, or alkaline hydrothermal vent, theory posits that life may have begun at submarine hydrothermal vents,[238][239] William Martin and Michael Russell have suggested "that life evolved in structured iron monosulphide precipitates in a seepage site hydrothermal mound at a redox, pH, and temperature gradient between sulphide-rich hydrothermal fluid and iron(II)-containing waters of the Hadean ocean floor. The naturally arising, three-dimensional compartmentation observed within fossilized seepage-site metal sulphide precipitates indicates that these inorganic compartments were the precursors of cell walls and membranes found in free-living prokaryotes. The known capability of FeS and NiS to catalyze the synthesis of the acetyl-methylsulphide from carbon monoxide and methylsulphide, constituents of hydrothermal fluid, indicates that pre-biotic syntheses occurred at the inner surfaces of these metal-sulphide-walled compartments,..."[240] These form where hydrogen-rich fluids emerge from below the sea floor, as a result of serpentinization of ultra-mafic olivine with seawater and a pH interface with carbon dioxide-rich ocean water. The vents form a sustained chemical energy source derived from redox reactions, in which electron donors (molecular hydrogen) react with electron acceptors (carbon dioxide); see Iron–sulfur world theory. These are highly exothermic reactions.[238][note 3]

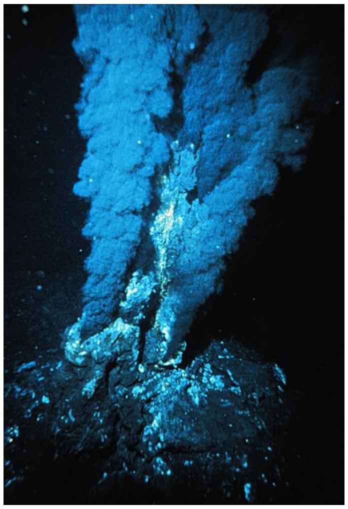

Deep-sea hydrothermal vent or black smoker

Michael Russell demonstrated that alkaline vents created an abiogenic proton motive force (PMF) chemiosmotic gradient,[240] in which conditions are ideal for an abiogenic hatchery for life. Their microscopic compartments "provide a natural means of concentrating organic molecules," composed of iron-sulfur minerals such as mackinawite, endowed these mineral cells with the catalytic properties envisaged by Wächtershäuser.[224] This movement of ions across the membrane depends on a combination of two factors:

1. Diffusion force caused by concentration gradient—all particles including ions tend to diffuse from higher concentration to lower.
2. Electrostatic force caused by electrical potential gradient—cations like protons H^+ tend to diffuse down the electrical potential, anions in the opposite direction.

These two gradients taken together can be expressed as an electrochemical gradient, providing energy for abiogenic synthesis. The proton motive force can be described as the measure of the potential energy stored as a combination of proton and voltage gradients across a membrane (differences in proton concentration and electrical potential).

Jack W. Szostak suggested that geothermal activity provides greater opportunities for the origination of life in open lakes where there is a buildup of minerals. In 2010, based on spectral analysis of sea and hot mineral water, Ignat Ignatov and Oleg Mosin demonstrated that life may have predominantly originated in hot mineral water. The hot mineral water that contains bicarbonate and calcium ions has the most optimal range.[241] This case is similar to the origin of life in hydrothermal vents, but with bicarbonate and calcium ions in hot water. This water has a pH of 9–11 and is possible to have the reactions in seawater. According to Melvin Calvin, certain reactions of condensation-dehydration of amino acids and nucleotides in individual blocks of peptides and nucleic acids can take place in the primary hydrosphere with pH 9-11 at a later evolutionary stage.[242] Some of these compounds like hydrocyanic acid (HCN) have been proven in the experiments of Miller. This is the environment in which the stromatolites have been created. David Ward of Montana State University described the formation of stromatolites in hot mineral water at the Yellowstone National Park. Stromatolites survive in hot mineral water and in proximity to areas with volcanic activity.[243] Processes have evolved in the sea near geysers of hot mineral water. In 2011, Tadashi Sugawara from the University of Tokyo created a protocell in hot water.[244]

Experimental research and computer modelling suggest that the surfaces of mineral particles inside hydrothermal vents have catalytic properties similar to those of enzymes and are able to create simple organic molecules, such as methanol (CH_3OH) and formic, acetic and pyruvic acid out of the dissolved CO_2 in the water.[245][246]

The research reported above by William F. Martin in July 2016 supports the thesis that life arose at hydrothermal vents,[247][248] that spontaneous chemistry in the Earth’s crust driven by rock–water interactions at disequilibrium thermodynamically underpinned life’s origin[249][250] and that the founding lineages of the archaea and bacteria were H2-dependent autotrophs that used CO2 as their terminal acceptor in energy metabolism.[251] Martin suggests, based upon this evidence that LUCA "may have depended heavily on the geothermal energy of the vent to survive".[252]

Thermosynthesis

Today's bioenergetic process of fermentation is carried out by either the aforementioned citric acid cycle or the Acetyl-CoA pathway, both of which have been connected to the primordial Iron–sulfur world. In a different approach, the thermosynthesis hypothesis considers the bioenergetic process of chemiosmosis, which plays an essential role in cellular respiration and photosynthesis, more basal than fermentation: the ATP synthase enzyme, which sustains chemiosmosis, is proposed as the currently extant enzyme most closely related to the first metabolic process.[253][254]

First, life needed an energy source to bring about the condensation reaction that yielded the peptide bonds of proteins and the phosphodiester bonds of RNA. In a generalization and thermal variation of the binding change mechanism of today's ATP synthase, the "first protein" would have bound substrates (peptides, phosphate, nucleosides, RNA 'monomers') and condensed them to a reaction product that remained bound until after a temperature change it was released by thermal unfolding.

The energy source under the thermosynthesis hypothesis was thermal cycling, the result of suspension of protocells in a convection current, as is plausible in a volcanic hot spring; the convection accounts for the self-organization and dissipative structure required in any origin of life model. The still ubiquitous role of thermal cycling in germination and cell division is considered a relic of primordial thermosynthesis.

By phosphorylating cell membrane lipids, this "first protein" gave a selective advantage to the lipid protocell that contained the protein. This protein also synthesized a library of many proteins, of which only a minute fraction had thermosynthesis capabilities. As proposed by Dyson,[13] it propagated functionally: it made daughters with similar capabilities, but it did not copy itself. Functioning daughters consisted of different amino acid sequences.

Whereas the Iron–sulfur world identifies a circular pathway as the most simple, the thermosynthesis hypothesis does not even invoke a pathway: ATP synthase's binding change mechanism resembles a physical adsorption process that yields free energy,[255] rather than a regular enzyme's mechanism, which decreases the free energy. It has been claimed that the emergence of cyclic systems of protein catalysts is implausible.[256]

Other models

Clay hypothesis

Montmorillonite, an abundant clay, is a catalyst for the polymerization of RNA and for the formation of membranes from lipids.[257] A model for the origin of life using clay was forwarded by Alexander Graham Cairns-Smith in 1985 and explored as a plausible mechanism by several scientists.[258] The clay hypothesis postulates that complex organic molecules arose gradually on pre-existing, non-organic replication surfaces of silicate crystals in solution.

At the Rensselaer Polytechnic Institute, James P. Ferris' studies have also confirmed that clay minerals of montmorillonite catalyze the formation of RNA in aqueous solution, by joining nucleotides to form longer chains.[259]

In 2007, Bart Kahr from the University of Washington and colleagues reported their experiments that tested the idea that crystals can act as a source of transferable information, using crystals of potassium hydrogen phthalate. "Mother" crystals with imperfections were cleaved and used as seeds to grow "daughter" crystals from solution. They then examined the distribution of imperfections in the new crystals and found that the imperfections in the mother crystals were reproduced in the daughters, but the daughter crystals also had many additional imperfections. For gene-like behaviour to be observed, the quantity of inheritance of these imperfections should have exceeded that of the mutations in the successive generations, but it did not. Thus Kahr concluded that the crystals "were not faithful enough to store and transfer information from one generation to the next."[260]

Nature timeline

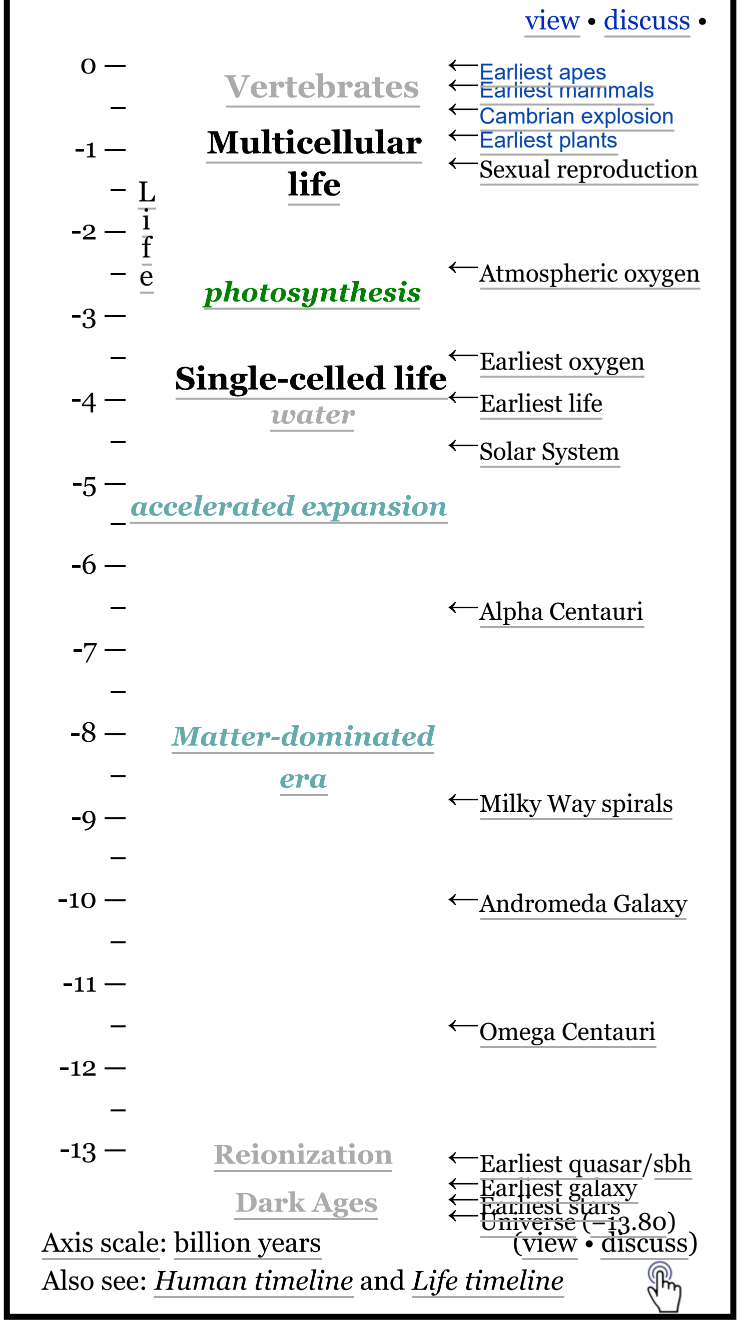
view • discuss •
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
Life
Vertebrates
Multicellular life
photosynthesis
Single-celled life
water
accelerated expansion
Matter-dominated era
Reionization
Dark Ages
←Earliest apes
←Earliest mammals
←Cambrian explosion
←Earliest plants
←Sexual reproduction
←Atmospheric oxygen
←Earliest oxygen
←Earliest life
←Solar System
←Alpha Centauri
←Milky Way spirals
←Andromeda Galaxy
←Omega Centauri
←Earliest quasar/sbh
←Earliest galaxy
←Earliest stars
←Universe (−13.80)
(view • discuss)
Axis scale: billion years
Also see: Human timeline and Life timeline

Gold's "deep-hot biosphere" model

In the 1970s, Thomas Gold proposed the theory that life first developed not on the surface of the Earth, but several kilometres below the surface. It is claimed that discovery of microbial life below the surface of another body in our Solar System would lend significant credence to this theory. Thomas Gold also asserted that a trickle of food from a deep, unreachable, source is needed for survival because life arising in a puddle of organic material is likely to consume all of its food and become extinct. Gold's theory is that the flow of such food is due to out-gassing of primordial methane from the Earth's mantle; more conventional explanations of the food supply of deep microbes (away from sedimentary carbon compounds) is that the organisms subsist on hydrogen released by an interaction between water and (reduced) iron compounds in rocks.

Panspermia

Panspermia is the hypothesis that life exists throughout the universe, distributed by meteoroids, asteroids, comets,[261] planetoids,[262] and, also, by spacecraft in the form of unintended contamination by microorganisms.[263][264]

The panspermia hypothesis does not attempt to explain how life first originated, but merely shifts it to another planet or a comet. The advantage of an extraterrestrial origin of primitive life is that life is not required to have formed on each planet it occurs on, but rather in a single location, and then spread about the galaxy to other star systems via cometary and/or meteorite impact.[265] Evidence to support the hypothesis is scant, but it finds support in studies of Martian meteorites found in Antarctica and in studies of extremophile microbes' survival in outer space tests.[266][267][268][269] (See also: List of microorganisms tested in outer space.)

Extraterrestrial organic molecules

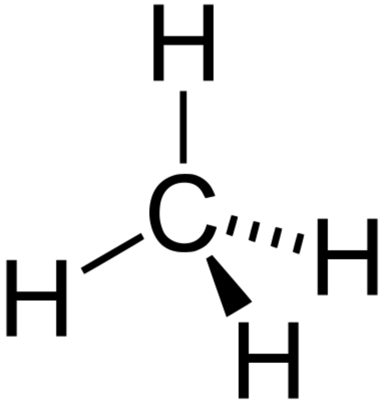

Methane is one of the simplest organic compounds

An organic compound is any member of a large class of gaseous, liquid, or solid chemicals whose molecules contain carbon. Carbon is the fourth most abundant element in the Universe by mass after hydrogen, helium, and oxygen.[270] Carbon is abundant in the Sun, stars, comets, and in the atmospheres of most planets.[271] Organic compounds are relatively common in space, formed by "factories of complex molecular synthesis" which occur in molecular clouds and circumstellar envelopes, and chemically evolve after reactions are initiated mostly by ionizing radiation.[20][272][273][274] Based on computer model studies, the complex organic molecules necessary for life may have formed on dust grains in the protoplanetary disk surrounding the Sun before the formation of the Earth.[124] According to the computer studies, this same process may also occur around other stars that acquire planets.[124]

Observations suggest that the majority of organic compounds introduced on Earth by interstellar dust particles are considered principal agents in the formation of complex molecules, thanks to their peculiar surface-catalytic activities.[275][276] Studies reported in 2008, based on $^{12}C/^{13}C$ isotopic ratios of organic compounds found in the Murchison meteorite, suggested that the RNA component uracil and related molecules, including xanthine, were formed extraterrestrially.[277][278] On 8 August 2011, a report based on NASA studies of meteorites found on Earth was published suggesting DNA components (adenine, guanine and related organic molecules) were made in outer space.[275][279][280] Scientists also found that the cosmic dust permeating the universe contains complex organics ("amorphous organic solids with a mixed aromatic–aliphatic structure") that could be created naturally, and rapidly, by stars.[281][282][283] Sun Kwok of The University of Hong Kong suggested that these compounds may have been related to the development of life on Earth said that "If this is the case, life on Earth may have had an easier time getting started as these organics can serve as basic ingredients for life."[281]

Glycolaldehyde, the first example of an interstellar sugar molecule, was detected in the star-forming region near the centre of our galaxy. It was discovered in 2000 by Jes Jørgensen and Jan M. Hollis.[284] In 2012, Jørgensen's team reported the detection of glycolaldehyde in a distant star system. The molecule was found around the protostellar binary IRAS 16293-2422 400 light years from Earth.[285][286][287] Glycolaldehyde is needed to form RNA, which is similar in function to DNA. These findings suggest that complex organic molecules may form in stellar systems prior to the formation of planets, eventually arriving on young planets early in their formation.[288] Because sugars are associated with both metabolism and the genetic code, two of the most basic aspects of life, it is thought the discovery of extraterrestrial sugar increases the likelihood that life may exist elsewhere in our galaxy.[284]

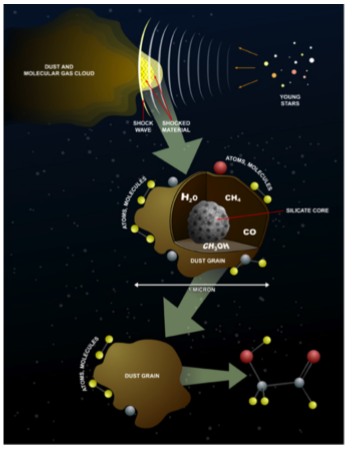

Formation of glycolaldehyde in stardust

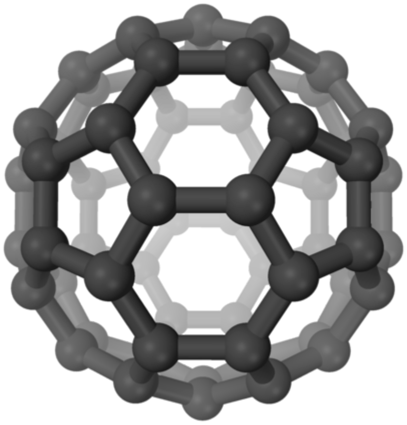

The C_{60} buckyball is a complex molecule that has been detected in nebulae.

NASA announced in 2009 that scientists had identified another fundamental chemical building block of life in a comet for the first time, glycine, an amino acid, which was detected in material ejected from comet Wild 2 in 2004 and grabbed by NASA's *Stardust* probe. Glycine has been detected in meteorites before. Carl Pilcher, who leads the NASA Astrobiology Institute commented that "The discovery of glycine in a comet supports the idea that the fundamental building blocks of life are prevalent in space, and strengthens the argument that life in the universe may be common rather than rare."[289] Comets are encrusted with outer layers of dark material, thought to be a tar-like substance composed of complex organic material formed from simple carbon compounds after reactions initiated mostly by ionizing radiation. It is possible that a rain of material from comets could have brought significant quantities of such complex organic molecules to Earth.[290][291][292] Amino acids which were formed extraterrestrially may also have arrived on Earth via comets.[49] It is estimated that during the Late Heavy Bombardment, meteorites may have delivered up to five million tons of organic prebiotic elements to Earth per year.[49]

Polycyclic aromatic hydrocarbons (PAH) are the most common and abundant of the known polyatomic molecules in the observable universe, and are considered a likely constituent of the primordial sea.[293][294][295] In 2010, PAHs, along with fullerenes (or "buckyballs"), have been detected in nebulae.[296][297] In March 2015, NASA scientists reported that, for the first time, complex DNA and RNA organic compounds of life, including uracil, cytosine and thymine, have been formed in the laboratory under outer space conditions, using starting chemicals, such as pyrimidine, found in meteorites. Pyrimidine, like PAHs, the most carbon-rich chemical found in the Universe, may have been formed in red giant stars or in interstellar dust and gas clouds.[298] A group of Czech scientists reported that all four RNA-bases may be synthesized from formamide in the course of high-energy density events like extraterrestrial impacts.[299]

Lipid world

The lipid world theory postulates that the first self-replicating object was lipid-like.[300][301] It is known that phospholipids form lipid bilayers in water while under agitation—the same structure as in cell membranes. These molecules were not present on early Earth, but other amphiphilic long-chain molecules also form membranes. Furthermore, these bodies may expand (by insertion of additional lipids), and under excessive expansion may undergo spontaneous splitting which preserves the same size and composition of lipids in the two progenies. The main idea in this theory is that the molecular composition of the lipid bodies is the preliminary way for information storage, and evolution led to the appearance of polymer entities such as RNA or DNA that may store information favourably. Studies on vesicles from potentially prebiotic amphiphiles have so far been limited to systems containing one or two types of amphiphiles. This in contrast to the output of simulated prebiotic chemical reactions, which typically produce very heterogeneous mixtures of compounds.[177] Within the hypothesis of a lipid bilayer membrane composed of a mixture of various distinct amphiphilic compounds there is the opportunity of a huge number of theoretically possible combinations in the arrangements of these amphiphiles in the membrane. Among all these potential combinations, a specific local arrangement of the membrane would have favoured the constitution of a hypercycle,[302][303] actually a positive feedback composed of two mutual catalysts represented by a membrane site and a specific compound trapped in the vesicle. Such site/compound pairs are transmissible to the daughter vesicles leading to the emergence of distinct lineages of vesicles which would have allowed Darwinian natural selection.[304]

Polyphosphates

A problem in most scenarios of abiogenesis is that the thermodynamic equilibrium of amino acid versus peptides is in the direction of separate amino acids. What has been missing is some force that drives polymerization. The resolution of this problem may well be in the properties of polyphosphates.[305][306] Polyphosphates are formed by polymerization of ordinary monophosphate ions PO_4^{-3}. Several mechanisms of organic molecule synthesis have been investigated. Polyphosphates cause polymerization of amino acids into peptides. They are also logical precursors in the synthesis of such key biochemical compounds as adenosine triphosphate (ATP). A key issue seems to be that calcium reacts with soluble phosphate to form insoluble calcium phosphate (apatite), so some plausible mechanism must be found to keep calcium ions from causing precipitation of phosphate. There has been much work on this topic over the years, but an interesting new idea is that meteorites may have introduced reactive phosphorus species on the early Earth.[307]

PAH world hypothesis

Polycyclic aromatic hydrocarbons (PAH) are known to be abundant in the universe,[293][294][295] including in the interstellar medium, in comets, and in meteorites, and are some of the most complex molecules so far found in space.[271]

The Cat's Paw Nebula lies inside the Milky Way Galaxy and is located in the constellation Scorpius.
Green areas show regions where radiation from hot stars collided with large molecules and small dust grains called "polycyclic aromatic hydrocarbons" (PAHs), causing them to fluoresce. (Spitzer space telescope, 2018)

Other sources of complex molecules have been postulated, including extraterrestrial stellar or interstellar origin. For example, from spectral analyses, organic molecules are known to be present in comets and meteorites. In 2004, a team detected traces of PAHs in a nebula.[308] In 2010, another team also detected PAHs, along with fullerenes, in nebulae.[296] The use of PAHs has also been proposed as a precursor to the RNA world in the PAH world hypothesis. The Spitzer Space Telescope has detected a star, HH 46-IR, which is forming by a process similar to that by which the Sun formed. In the disk of material surrounding the star, there is a very large range of molecules, including cyanide compounds, hydrocarbons, and carbon monoxide. In September 2012, NASA scientists reported that PAHs, subjected to interstellar medium conditions, are transformed, through hydrogenation, oxygenation and hydroxylation, to more complex organics—"a step along the path toward amino acids and nucleotides, the raw materials of proteins and DNA, respectively."[309][310] Further, as a result of these transformations, the PAHs lose their spectroscopic signature which could be one of the reasons "for the lack of PAH detection in interstellar ice grains, particularly the outer regions of cold, dense clouds or the upper molecular layers of protoplanetary disks."[309][310]

NASA maintains a database for tracking PAHs in the universe.[271][311] More than 20% of the carbon in the universe may be associated with PAHs,[271][271] possible starting materials for the formation of life. PAHs seem to have been formed shortly after the Big Bang, are widespread throughout the universe,[293][294][295] and are associated with new stars and exoplanets.[271]

Radioactive beach hypothesis

Zachary Adam claims that tidal processes that occurred during a time when the Moon was much closer may have concentrated grains of uranium and other radioactive elements at the high-water mark on primordial beaches, where they may have been responsible for generating life's building blocks.[312] According to computer models,[313] a deposit of such radioactive materials could show the same self-sustaining nuclear reaction as that found in the Oklo uranium ore seam in Gabon. Such radioactive beach sand might have provided sufficient energy to generate organic molecules, such as amino acids and sugars from acetonitrile in water. Radioactive monazite material also has released soluble phosphate into the regions between sand-grains, making it biologically "accessible." Thus amino acids, sugars, and soluble phosphates might have been produced simultaneously, according to Adam. Radioactive actinides, left behind in some concentration by the reaction, might have formed part of organometallic complexes. These complexes could have been important early catalysts to living processes.

John Parnell has suggested that such a process could provide part of the "crucible of life" in the early stages of any early wet rocky planet, so long as the planet is large enough to have generated a system of plate tectonics which brings radioactive minerals to the surface. As the early Earth is thought to have had many smaller plates, it might have provided a suitable environment for such processes.[314]

Thermodynamic dissipation

The 19th-century Austrian physicist Ludwig Boltzmann first recognized that the struggle for existence of living organisms was neither over raw material nor energy, but instead had to do with entropy production derived from the conversion of the solar spectrum into heat by these systems.[315] Boltzmann thus realized that living systems, like all irreversible processes, were dependent on the dissipation of a generalized chemical potential for their existence. In his book "What is Life", the 20th-century Austrian physicist Erwin Schrödinger[316] emphasized the importance of Boltzmann's deep insight into the irreversible thermodynamic nature of living systems, suggesting that this was the physics and chemistry behind the origin and evolution of life. However, irreversible processes, and much less living systems, could not be conveniently analyzed under this perspective until Lars Onsager,[317] and later Ilya Prigogine,[318] developed an elegant mathematical formalism for treating the "self-organization" of material under a generalized chemical potential. This formalism became known as Classical Irreversible Thermodynamics and Prigogine was awarded the Nobel Prize in Chemistry in 1977 "for his contributions to non-equilibrium thermodynamics, particularly the theory of dissipative structures". The analysis of Prigogine showed that if a system were left to evolve under an imposed external potential, material could spontaneously organize (lower its entropy) forming what he called "dissipative structures" which would increase the dissipation of the externally imposed potential (augment the global entropy production). Non-equilibrium thermodynamics has since been successfully applied to the analysis of living systems, from the biochemical production of ATP [319] to optimizing bacterial metabolic pathways [320] to complete ecosystems.[321][322][323]

In his "Thermodynamic Dissipation Theory of the Origin and Evolution of Life",[324][325][326][327] Karo Michaelian has taken the insight of Boltzmann and the work of Prigogine to its ultimate consequences regarding the origin of life. This theory postulates that the hallmark of the origin and evolution of life is the microscopic dissipative structuring of organic pigments and their proliferation over the entire Earth surface.[327] Present day life augments the entropy production of Earth in its solar environment by dissipating ultraviolet and visible photons into heat through organic pigments in water. This heat then catalyzes a host of secondary dissipative processes such as the water cycle, ocean and wind currents, hurricanes, etc.[325][328] Michaelian argues that if the thermodynamic function of life today is to produce entropy through photon dissipation in organic pigments, then this probably was its function at its very beginnings. It turns out that both RNA and DNA when in water solution are very strong absorbers and extremely rapid dissipaters of ultraviolet light within the 230–290 nm wavelength (UV-C) region, which is a part of the Sun's spectrum that could have penetrated the prebiotic atmosphere.[329] In fact, not only RNA and DNA, but many fundamental molecules of life (those common to all three domains of life) are also pigments that absorb in the UV-C, and many of these also have a chemical affinity to RNA and DNA.[330][331] Nucleic acids may thus have acted as acceptor molecules to the UV-C photon excited antenna pigment donor molecules by providing an ultrafast channel for dissipation. Michaelian has shown using the formalism of non-linear irreversible thermodynamics that there would have existed during the Archean a thermodynamic imperative to the abiogenic UV-C photochemical synthesis and proliferation of these pigments over the entire Earth surface if they acted as catalysts to augment the dissipation of the solar photons.[332] By the end of the Archean, with life-induced ozone dissipating UV-C light in the Earth's upper atmosphere, it would have become ever more improbable for a completely new life to emerge that didn't rely on the complex metabolic pathways already existing since now the free energy in the photons arriving at Earth's surface would have been insufficient for direct breaking and remaking of covalent bonds. It has been suggested, however, that such changes in the surface flux of ultraviolet radiation due to geophysical events affecting the atmosphere could have been what promoted the development of complexity in life based on existing metabolic pathways, for example during the Cambrian explosion [333]

Many salient characteristics of the fundamental molecules of life (those found in all three domains) all point directly to the involvement of UV-C light in the dissipative structuring of incipient life.[326] Some of the most difficult problems concerning the origin of life, such as enzyme-less replication of RNA and DNA, homochirality of the fundamental molecules, and the origin of information encoding in RNA and DNA, also find an explanation within the same dissipative thermodynamic framework by considering the probable existence of a relation between primordial replication and UV-C photon dissipation. Michaelian suggests that it is erroneous to expect to describe the emergence, proliferation, or even evolution, of life without overwhelming reference to entropy production through the dissipation of a generalized chemical potential, in particular, the prevailing solar photon flux.

Multiple genesis

Different forms of life with variable origin processes may have appeared quasi-simultaneously in the early history of Earth.[334] The other forms may be extinct (having left distinctive fossils through their different biochemistry—e.g., hypothetical types of biochemistry). It has been proposed that:

> The first organisms were self-replicating iron-rich clays which fixed carbon dioxide into oxalic and other dicarboxylic acids. This system of replicating clays and their metabolic phenotype then evolved into the sulfide rich region of the hotspring acquiring the ability to fix nitrogen. Finally phosphate was incorporated into the evolving system which allowed the synthesis of nucleotides and phospholipids. If biosynthesis recapitulates biopoiesis, then the synthesis of amino acids preceded the synthesis of the purine and pyrimidine bases. Furthermore the polymerization of the amino acid thioesters into polypeptides preceded the directed polymerization of amino acid esters by polynucleotides.[335]

Fluctuating hydrothermal pools on volcanic islands or proto-continents

Armid Mulkidjanian and co-authors think that the marine environments did not provide the ionic balance and composition universally found in cells, as well as of ions required by essential proteins and ribozymes found in virtually all living organisms, especially with respect to K^+/Na^+ ratio, Mn^{2+}, Zn^{2+} and phosphate concentrations. The only known environments that mimic the needed conditions on Earth are found in terrestrial hydrothermal pools fed by steam vents.[238] Additionally, mineral deposits in these environments under an anoxic atmosphere would have suitable pH (as opposed to current pools in an oxygenated atmosphere), contain precipitates of sulfide minerals that block harmful UV radiation, have wetting/drying cycles that concentrate substrate solutions to concentrations amenable to spontaneous formation of polymers of nucleic acids, polyesters[336] and depsipeptides,[337] both by chemical reactions in the hydrothermal environment, as well as by exposure to UV light during transport from vents to adjacent pools. Their hypothesized pre-biotic environments are similar to the deep-oceanic vent environments most commonly hypothesized, but add additional components that help explain peculiarities found in reconstructions of the Last Universal Common Ancestor (LUCA) of all living organisms.[338]

Bruce Damer and David Deamer have come to the conclusion that cell membranes cannot be formed in salty seawater, and must therefore have originated in freshwater. Before the continents formed, the only dry land on Earth would be volcanic islands, where rainwater would form ponds where lipids could form the first stages towards cell membranes. These predecessors of true cells are assumed to have behaved more like a superorganism rather than individual structures, where the porous membranes would house molecules which would leak out and enter other protocells. Only when true cells had evolved would they gradually adapt to saltier environments and enter the ocean.[339]

Colín-García *et al.* (2016) discuss the advantages and disadvantages of hydrothermal vents as primitive environments.[238] They mention the exergonic reactions in such systems could have been a source of free energy that promoted chemical reactions, additional to their high mineralogical diversity which implies the induction of important chemical gradients, thus favoring the interaction between electron donors and acceptors. Colín-García *et al.* (2016) also summarize a set of experiments proposed to test the role of hydrothermal vents in prebiotic synthesis.[238]

Information theory

A theory that speaks to the origin of life on Earth and other rocky planets posits life as an information system in which information content grows because of selection. Life must start with minimum possible information, or minimum possible departure from thermodynamic equilibrium, and it requires thermodynamically free energy accessible by means of its information content. The most benign circumstances, minimum entropy variations with abundant free energy, suggest the pore space in the first few kilometres of the surface. Free energy is derived from the condensed products of the chemical reactions taking place in the cooling nebula.[340]

See also

- Anthropic principle
- Artificial cell
- Artificial life
- Astrochemistry
- Biological immortality
- Carbon Mineral Challenge
- Common descent
- Entropy and life
- Formamide-based prebiotic chemistry
- GADV protein world
- Mediocrity principle
- Mycoplasma laboratorium
- Nexus for Exoplanet System Science
- Noogenesis
- Planetary habitability
- Rare Earth hypothesis
- Shadow biosphere
- Tholin

Notes

1. Also occasionally called biopoiesis.
2. The reactions are:

$$FeS + H_2S \rightarrow FeS_2 + 2H^+ + 2e^-$$
$$FeS + H_2S + CO_2 \rightarrow FeS_2 + HCOOH$$

3. The reactions are:
Reaction 1: Fayalite + water → magnetite + aqueous silica + hydrogen

OPINION

A new perspective on radiation resistance based on *Deinococcus radiodurans*

Michael J. Daly

Abstract | In classical models of radiation toxicity, DNA is the molecule that is most affected by ionizing radiation (IR). However, recent data show that the amount of protein damage caused during irradiation of bacteria is better related to survival than to DNA damage. In this Opinion article, a new model is presented in which proteins are the most important target in the hierarchy of macromolecules affected by IR. A first line of defence against IR in extremely radiation-resistant bacteria might be the accumulation of manganese complexes, which can prevent the production of iron-dependent reactive oxygen species. This would allow an irradiated cell to protect sufficient enzymatic activity needed to repair DNA and survive.

One early goal of radiobiology was to explain why most organisms are so sensitive to X-rays and γ-rays. A whole-body exposure of just 10 Gy (gray; absorbed radiation dose) is lethal to most vertebrate animals, including humans[1], and most bacteria[1] cannot survive 200 Gy. Invertebrate animals are more resistant if the adult form has no somatic cell division (for example, fruit flies, which are largely post-mitotic)[2], and can survive 500 Gy. The freshwater invertebrate animal *Philodina roseola*[3], the water bear *Milnesium tardigradum*[4] and the roundworm *Caenorhabditis elegans*[5] can tolerate 3,000–5000 Gy, but are rendered sterile. As a haploid, the basidiomycete fungus *Ustilago maydis* carries a single set of chromosomes per cell in the G2 phase (after its chromosomes have been duplicated but before cell division) and can withstand ~3,600 Gy[6]. The archaeal species *Pyrococcus furiosus*[7] and *Halobacterium* sp. NRC-1 (REF. 8) can resist ~3,000 and ~5,000 Gy, respectively, whereas some bacteria from the *Deinococcus–Thermus* group can survive doses of more than 12,000 Gy[9]. Perhaps if we knew why some cells are so resistant to radiation, we could find ways to protect people from atomic radiation.

In the 1950s, DNA, lipids and proteins were proposed to be the main targets of ionizing radiation (IR) in cells. By the mid 1960s, 'death by DNA damage' became the predominant paradigm of radiation toxicity[10]. In the decades that followed, radiation biologists attempted to explain the extreme IR resistance of radioresistant organisms that can withstand extraordinary genome fragmentation and DNA base damage. This damage is caused by reactive oxygen species (ROS), the chemical agents that are principally responsible for cellular radiation damage[11] (BOX 1).

Because the fate of an irradiated cell ultimately depends on whether its genome is preserved, most studies have focused on how DNA that is damaged during irradiation is repaired[11–14]. DNA double-strand breaks (DSBs) are presumed to be the most lethal damage, although they are significantly less frequent than single-strand breaks (SSBs) and DNA base damages. IR damage that affects only one strand (for example, SSBs or damage to DNA bases) can be repaired using the undamaged complementary strand as a template, but DSBs provide little guidance for mending without mutagenesis, as neither of the two strands is intact. Non-mutagenic repair of DSBs is dependent on the presence of at least two sets of chromosomes per cell[12–15], although genome multiplicity by itself is insufficient for radioresistance. For example, *Escherichia coli* contains 4–8 haploid genomes per cell during exponential growth but cannot survive 200 Gy, which causes approximately 4 DSBs per haploid genome[14]. Because the linear density of genomic DSBs inflicted per Gy per Mbp (0.004–0.01) is similar for diverse organisms[3,9,12,16–18], during irradiation, cells with small genomes suffer fewer DSBs than cells with large genomes. For example, an acute dose of 1,000 Gy that inflicts ~30 DSBs in a 3 Mbp chromosome would be expected to inflict ~150 DSBs in a 15 Mbp chromosome. Therefore, cell survival would be predicted to decline as genome size increases, but significantly improve as ploidy increases to more than two. A similar result emerged for radiation-induced DNA base damage: the differences in base lesion yields for resistant and sensitive bacteria exposed to ultraviolet C (UVC) radiation (265 nm) or γ-rays were not nearly sufficient to account for the differences in bacterial survival[19,20].

UVC causes substantial direct damage to both DNA and proteins *in vivo*[19]. For proteins, UVC disrupts certain types of disulphide bonds and causes protein aggregation[21]. Antioxidants can increase the survival of cells exposed to UVC[21]. Thus, although cellular macromolecules are clearly not protected from direct damage caused by either UVC[19] or IR[9,16], any antioxidants in the cells would also be expected to help avert ROS-mediated toxicity that is secondary to the radiation itself; for example, caused by damaged proteins involved in energy metabolism[22]. Irradiated *Deinococcus radiodurans* is predicted to rid itself of damaged proteins using an expanded family of subtilisin-like proteases[23] that are expected to be highly expressed.

Radiation-resistant organisms suffer from similar levels of genomic damage following irradiation as sensitive organisms, but they can survive the formation of hundreds of IR-induced DSBs per genome[3,9,12,14,16–18]. The first reports of homologous recombination in an extremely IR-resistant organism

© 2009 Macmillan Publishers Limited. All rights reserved

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **MAR 2009**	2. REPORT TYPE	3. DATES COVERED **00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE **A new perspective on radiation resistance based on Deinococcus radiodurans**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Uniformed Services University of the Health Sciences (USUHS),Department of Pathology,4301 Jones Bridge Road,Bethesda,MD,20814**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In classical models of radiation toxicity, DNA is the molecule that is most affected by ionizing radiation (IR). However, recent data show that the amount of protein damage caused during irradiation of bacteria is better related to survival than to DNA damage. In this Opinion article, a new model is presented in which proteins are the most important target in the hierarchy of macromolecules affected by IR. A first line of defence against IR in extremely radiation-resistant bacteria might be the accumulation of manganese complexes, which can prevent the production of iron-dependent reactive oxygen species. This would allow an irradiated cell to protect sufficient enzymatic activity needed to repair DNA and survive.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Box 1 | Primary reactive oxygen species generated and their cellular targets

H_2O is the most abundant chemical found in living cells. The primary source of reactive oxygen species (ROS) generated in irradiated cells is the radiolysis of H_2O, which generates $HO^{\bullet}$, H^+ and e^-_{aq} (REFS 11,69) (see the figure, Equation 1). $HO^{\bullet}$ radicals react with each other to generate H_2O_2 (REFS 11,69) (see the figure, Equation 2), which can diffuse throughout the cell[11,34]. Dissolved O_2, either derived from the atmosphere or generated endogenously, reacts with e^-_{aq} to form $O_2^{\bullet-}$ (REFS 11,34,69) (see the figure, Equation 3), which can reform H_2O_2 in the presence of H^+ (REF. 69) (see the figure, Equations 4 and 5). Whereas H_2O_2 and $O_2^{\bullet-}$ are relatively inert, $HO^{\bullet}$ are extremely oxidizing, indiscriminately reacting with organic molecules and oxidizing DNA, RNA, lipids and proteins[11,34]. However, the short lifetime of the $HO^{\bullet}$ precludes damage to molecules beyond a few Angstroms from where $HO^{\bullet}$ is formed. For DNA, as the dose of ionizing radiation (IR) increases, the linear density of base damages and single strand breaks (SSBs) increases on both strands, which gives rise to double-strand breaks (DSBs). In addition to the widespread 'indirect damage' caused in cells by IR-induced ROS, IR also can inflict 'direct damage' when macromolecules absorb X-ray and γ-ray photons[11]. A dose of IR typically causes 40 times more SSBs than DSBs[9,11]. A secondary source of $HO^{\bullet}$ in cells during irradiation is the Fenton reaction, which is one of the most powerful oxidizing reactions known and involves the catalytic decomposition of H_2O_2 by ferrous ions ($H_2O_2 + Fe^{2+} \rightarrow Fe^{3+} + OH^- + HO^{\bullet}$)[61] (FIG. 3); the analogous reaction with Mn(II) does not occur[51]. In contrast to $HO^{\bullet}$, the reactivity of $O_2^{\bullet-}$ is high only for selected targets: small inorganic prosthetic Fe–S groups of some proteins[57,61]. The damaging potential of $O_2^{\bullet-}$ is greater than that of H_2O_2 for Fe–S groups because $O_2^{\bullet-}$ is negatively charged.

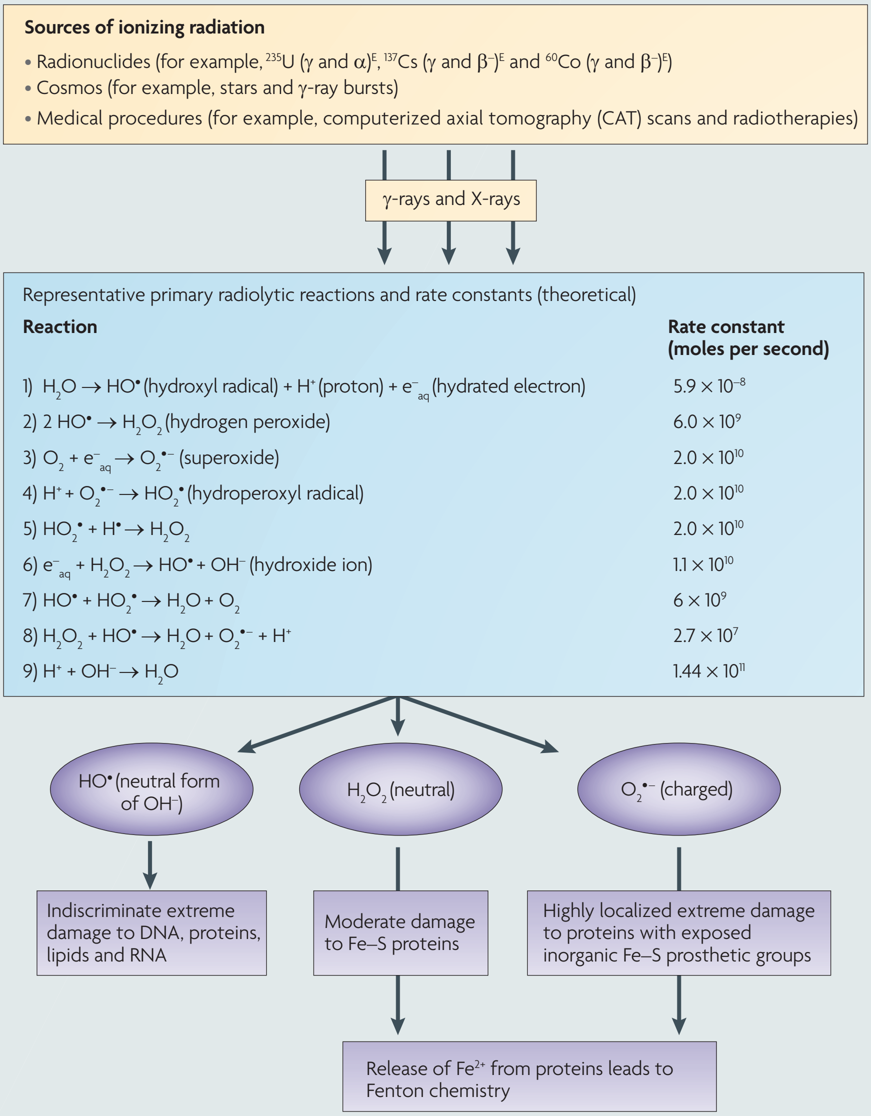

Sources of ionizing radiation

- Radionuclides (for example, ^{235}U (γ and α)^{E}, ^{137}Cs (γ and β^-)^{E} and ^{60}Co (γ and β^-)^{E})
- Cosmos (for example, stars and γ-ray bursts)
- Medical procedures (for example, computerized axial tomography (CAT) scans and radiotherapies)

Representative primary radiolytic reactions and rate constants (theoretical)

Reaction	Rate constant (moles per second)
1) $H_2O \rightarrow HO^{\bullet}$ (hydroxyl radical) + H^+ (proton) + e^-_{aq} (hydrated electron)	5.9×10^{-8}
2) $2\ HO^{\bullet} \rightarrow H_2O_2$ (hydrogen peroxide)	6.0×10^{9}
3) $O_2 + e^-_{aq} \rightarrow O_2^{\bullet-}$ (superoxide)	2.0×10^{10}
4) $H^+ + O_2^{\bullet-} \rightarrow HO_2^{\bullet}$ (hydroperoxyl radical)	2.0×10^{10}
5) $HO_2^{\bullet} + H^{\bullet} \rightarrow H_2O_2$	2.0×10^{10}
6) $e^-_{aq} + H_2O_2 \rightarrow HO^{\bullet} + OH^-$ (hydroxide ion)	1.1×10^{10}
7) $HO^{\bullet} + HO_2^{\bullet} \rightarrow H_2O + O_2$	6×10^{9}
8) $H_2O_2 + HO^{\bullet} \rightarrow H_2O + O_2^{\bullet-} + H^+$	2.7×10^{7}
9) $H^+ + OH^- \rightarrow H_2O$	1.44×10^{11}

HO• (neutral form of OH⁻)
H₂O₂ (neutral)
O₂•⁻ (charged)
Indiscriminate extreme damage to DNA, proteins, lipids and RNA
Moderate damage to Fe–S proteins
Highly localized extreme damage to proteins with exposed inorganic Fe–S prosthetic groups
Release of Fe²⁺ from proteins leads to Fenton chemistry

were for *U. maydis*, which carries two sets of chromosomes per cell and can survive ~6,000 Gy as a diploid[6,13]. The only other extremely IR-resistant organism in which the physical products of homologous recombination during genome reassembly have been mapped following irradiation is the bacterium *D. radiodurans* (BOX 2), which as a polyploid (4–10 haploid genome copies per cell) can survive ~17,000 Gy[12,15,24] (FIG. 1a). *U. maydis* and *D. radiodurans* rely on efficient systems for genome reassembly and require the induction of DNA-repair genes[13,25]. Homologous recombination processes in these organisms depend on proteins that actively promote pairing of DNA, which results in efficient exchange of genetic information between homologous DNA regions[12–15]. Over the past two decades, several hypotheses to explain the efficiency of homologous recombination in *D. radiodurans* have been proposed, building on the idea that *D. radiodurans* maintains a novel chromosome alignment that keeps homologous regions together and facilitates access of broken ends to homologous templates[14,26,27]. However, a series of molecular studies on irradiated *D. radiodurans* cells showed that high levels of recombination between homologous DSB fragments originated from widely separated genomic locations[12,15,24,26], and cryoelectron microscopy of vitreous sections of *D. radiodurans* showed that DNA fragments in the cells were mobile[28]. Thus, IR-induced DSB fragments in *D. radiodurans* are not immobilized and the structural form of its nucleoids does not seem to have a key role in radioresistance[22]. Indeed, sensitivity to DSB damage can increase when the mobility of broken DNA ends is restricted[29].

Early on, it became evident that the enzymes which mediate DNA repair in *D. radiodurans* were probably not unique[30,31]. For example, a highly IR-sensitive mutant of *D. radiodurans* that contains a mutated DNA polymerase I gene (*polA*) was fully restored by expression of the corresponding gene from the IR-sensitive *E. coli*[30].

The nature of extreme IR resistance in *Deinococcus* spp. thus remains unclear and is made even more curious by their impressive ability to resist desiccation[14,16,32] and UV radiation[30,31]. Hypotheses that explain why repair proteins (either native or introduced by genetic engineering from IR-sensitive bacteria) in *D. radiodurans* cells function better after irradiation than repair proteins in other organisms generally fall into two categories. First, a subset of uncharacterized genes encode proteins that somehow facilitate the homologous recombination systems

© 2009 Macmillan Publishers Limited. All rights reserved

of *D. radiodurans*[14,23,33], or second, the accumulation of near-millimolar concentrations of Mn^{2+} in cells somehow prevents protein oxidation during irradiation, with the result that sufficient repair enzymes survive radiation damage and allow subsequent DNA repair[16,22,26,34].

This Opinion article considers how, after exposure to huge doses of γ-rays, or to months of desiccation and exposure to UV sunlight in a desert, some bacteria retain sufficient protein activity to reconstitute their DNA. I propose that the accumulation of Mn^{2+} complexes in resistant organisms decreases the cellular concentration of ROS formed during irradiation. This in turn prevents IR-induced protein oxidation, which leaves sufficient recombination activity in place to repair the DSBs in the genome and enable survival.

The *Deinococcus* paradox

D. radiodurans was the first extremely IR-resistant organism to be sequenced[35]. Surprisingly, DNA-repair proteins in *D. radiodurans* seemed to be unremarkable, as the *D. radiodurans* genome encoded approximately the same number and types of DNA-repair proteins as IR-sensitive bacteria[23,26]. Similarly, the sequencing of genomes of the IR-resistant organisms *Bacillus pumilus*[36], *U. maydis*[6], *P. roseola*[3], *P. furiosus*[37] and *Halobacterium* sp. NRC-1 (REF. 38) did not provide clear insights into IR resistance mechanisms. Although the global transcriptional response to IR in *D. radiodurans* involved the upregulation of many novel genes, hundreds of characterized genes that were unrelated to DNA repair were also induced[25,39]. At least 12 novel *D. radiodurans* genes with discernible functional relevance to the preservation of genome integrity were knocked out, and the resulting mutants were characterized for IR resistance. Remarkably, there was no drastic change in the level of IR resistance of most of the mutants, indicating that few of the putative resistance genes, at least individually, make a substantial contribution to the recovery of irradiated *D. radiodurans*[26]. For example, DdrA is a protein that binds to and protects the 3′ ends of DNA from nuclease degradation *in vitro*, but 50% of *D. radiodurans ddrA*$^{-}$ mutant cells survived 10,000 Gy in an undefined rich medium[40]. Furthermore, a whole-genome analysis of gene gain and gene loss between *Deinococcus geothermalis* (BOX 2) and *D. radiodurans*, which are equally resistant to radiation[26], failed to produce a coherent picture of the systems that were thought to be important for survival. Instead, the number of novel genes that were thought to be implicated in recovery from IR in these *Deinococcus* species was substantially reduced[26].

Challenging radiation dogma

Because individual proteins are typically present at much higher levels than their corresponding genes, IR damage to one protein is not a lethal event, unlike an unrepaired DSB. As a dose of IR that causes indiscriminate cellular DNA damage should damage a similar portion of proteins in the same cell[11], the tacit assumption has been that IR-induced cell death is mainly mediated by DNA damage. However, recent investigations contradict this radiation dogma[34]. Whereas a specific dose of IR causes similar numbers of DSBs in sensitive and resistant bacteria[9,16,17], sensitive bacteria are more susceptible to oxidative protein damage than resistant bacteria[34] (FIG. 2). Based on this finding, I propose that the ability of IR-resistant organisms to minimize protein damage from IR allows them to retain sufficient DNA-repair functions to restore the chromosomal damage. I further argue that extremely IR-sensitive bacteria, such as *Shewanella oneidensis*, which are killed by IR doses (40 Gy) that cause less than 1 DSB per genome, do not survive mainly owing to protein damage[22,34,41]. Only for the few organisms that are extremely resistant to IR owing to their ability to protect proteins

Box 2 | Where *Deinococcus* bacteria roam

Representatives of the extremely radiation-resistant family Deinococcaceae can typically survive acute exposures to ionizing radiation (≥12,000 Gy (gray; absorbed radiation dose)), ultraviolet (≥1,000 J per m^2), and desiccation (years)[14,23,26,32,70], and can grow under harsh conditions of chronic irradiation (50 Gy per hour)[16]. Only approximately 30 distinct species have been described, despite their apparent ancient derivation[71]. The first member to be isolated was *Deinococcus radiodurans* (see the figure), originally named *Micrococcus radiodurans*[72]. This bacterium belongs to the *Deinococcus–Thermus* group, which is putatively related to cyanobacteria and is deeply branched in bacterial phylogenetic trees[23]. To date, the natural distribution of the deinococci has still not been explored systematically. Members have been isolated worldwide but have diverse and patchy distributions[23].

A wide range of natural and man-made environments on the Earth are characterized by their physical extremes of temperature, pressure and radiation. These harsh environments are different from the preferred environments of life typically encountered by humans, but can be colonized by *Deinococcus*. Some species live in highly radioactive soils at nuclear waste sites[73] and alpine environments[74], some have settled on sandstone, marble and ice in Antarctica[75], and others are ubiquitous microbial inhabitants of deserts[32]. High temperatures and pressures also do not seem to be an obstacle to their survival. *Deinococcus geothermalis* was originally isolated from a hot spring in Italy[76], and *D. geothermalis* DNA has been extracted from deep-ocean subsurface environments (68–118 meters below the sea floor)[77].

The astonishing survival of *Deinococcus* bacteria following irradiation has given rise to some whimsical descriptions of their derivation, including that they had an extraterrestrial origin[78]. Whole-genome comparisons of *D. geothermalis* with *D. radiodurans* and two *Thermus* species, however, have firmly established that the evolutionary steps which led to the *Deinococcus–Thermus* group occurred in their terrestrial ancestors[26]. *D. geothermalis* requires little more than some carbohydrates and oxygen to grow[79]. By contrast, *D. radiodurans* is metabolically less proficient. *D. radiodurans* requires a rich source of amino acids and vitamin B3 analogues in addition to sugars and oxygen for growth[79], representing biosynthetic deficiencies predicted by its genome sequence[23,35]. It is remarkable, therefore, that *D. radiodurans* has been isolated from nutrient-poor radioactive environmental waste sites and from deserts[32,73]. Indeed, numerous other *Deinococcus* species isolated from an elephant, a llama, a duck and a fish have similar growth requirements[23]. The picture that has emerged for the life cycle of most *Deinococcus* species is one that comprises a cell-replication phase that requires nutrient-rich conditions, such as in the gut of an animal[23], followed by release, drying and dispersal. Dried deinococci can endure for years, and if blown by winds through the atmosphere would be expected to survive and land worldwide — some becoming encased in ice, some becoming entombed in dried desert soils and some even ending up in the stomachs of humans[80].

© 2009 Macmillan Publishers Limited. All rights reserved

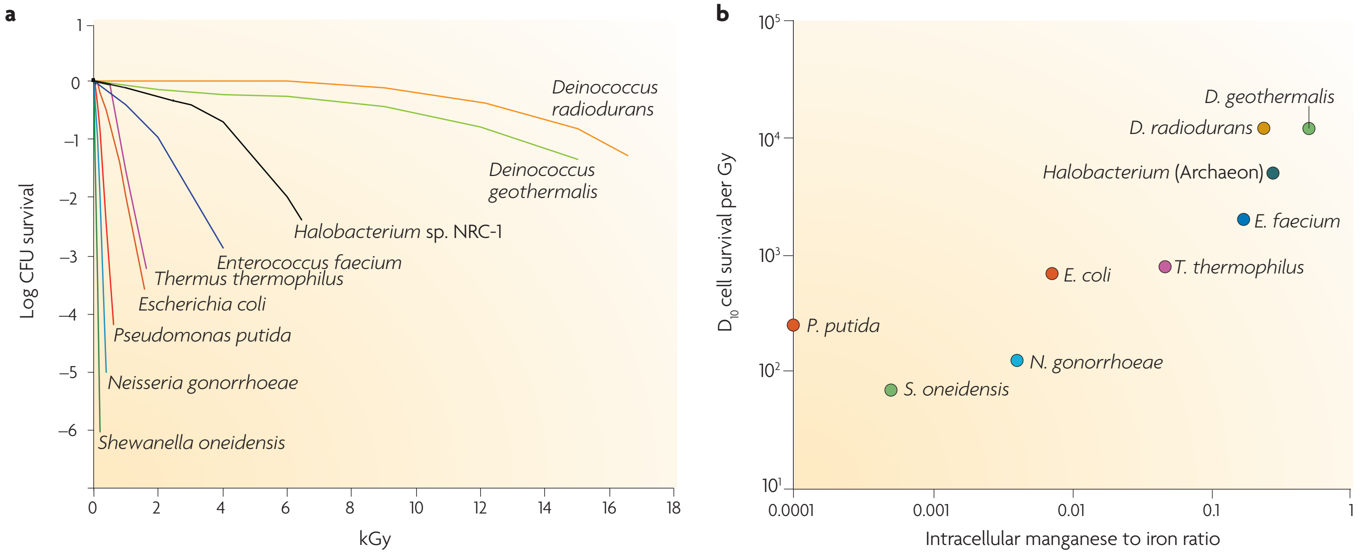

Figure 1 | **Relationship between survival following exposure to ionizing radiation and intracellular manganese and iron contents. a** | Ionizing radiation (IR) survival curves for whole-genome sequenced strains that encode a similar repertoire of DNA-repair proteins[8,16,38]. Standard growth, irradiation (^{60}Co at 0 °C) and recovery conditions were used[16]. **b** | Intracellular manganese to iron concentration ratios. For cells cultured under standard growth conditions, total manganese and iron contents were determined by inductively coupled plasma mass spectrometry[16]. D_{10} represents the IR dose that reduces the number of viable cells by 90%. Other manganese to iron ratios and D_{10} survival values reported recently for novel IR-resistant desert bacteria[32] include *Deinococcus radiodurans* 7b-1 (D_{10} = 13 kGy; manganese:iron = 0.14), *Deinococcus* sp. 1A1 (D_{10} = 17 kGy; manganese:iron = 0.15), *Deinococcus* sp. 5A5 (D_{10} = 15 kGy; manganese:iron = 0.38), *Deinococcus* sp. 1A6 (D_{10} = 7 kGy; manganese:iron = 0.15), *Deinococcus* sp. 3B1 (D_{10} = 10.5 kGy; manganese:iron = 0.12), alphaproteobacterium 4A4 (D_{10} = 1.5 kGy; manganese:iron = 0.07) and alphaproteobacterium 4A6 (D_{10} = 1.5 kGy; manganese:iron = 0.15). We did not investigate the distribution of manganese or iron in the desert strains, nor did we evaluate the extent of cell grouping before irradiation, which can significantly increase D_{10} survival values based on colony forming unit (CFU) assays[16,22]. Part **b** is modified, with permission, from REF. 81 © (2006) Elsevier Science.

from the indirect effects of IR (BOX 1) is survival ultimately determined by the level of DNA damage[12]. In the early 1970s, Holliday's work on IR-induced homologous recombination strongly supported the notion that the most radiosensitive targets in extremely IR-resistant *U. maydis* cells were single genes[13], an inference that has been applied by others to IR toxicity in general.

Resistance to DNA-damaging agents

In addition to IR and UV, *D. radiodurans* is profoundly resistant to the lethal and mutagenic effects of many redox-active (ROS-mediated) xenobiotics, such as mitomycin C[42], an antibiotic produced by *Streptomyces* that is known for its ability to cross-link DNA. Whereas *E. coli* is readily mutated by mitomycin C, *D. radiodurans* cells that survive mitomycin C treatment show approximately the same low level of mutagenesis that occurs during one normal round of replication[31,42]. Thus, *E. coli* has an error-prone DNA-repair pathway but *D. radiodurans* apparently does not[31]. Under our model, I attribute error-prone DNA repair to ROS-damaged enzymes, which passively promote mutations by repair malfunction[34]. By contrast, *D. radiodurans* and *E. coli* are equally sensitive to the mutagenic effects of *N*-methyl-*N*′-nitro-*N*-nitrosoguanidine, a direct mutagen[31,42] that was used to construct most *D. radiodurans* mutants prior to 1990 (REF. 31). Because DNA-repair-defective *D. radiodurans* mutants (such as *recA* and *polA* mutants) are approximately as sensitive to IR, UV and indirect (ROS-mediated) mutagens as repair-defective *E. coli*[9,30,31,43], accurate repair of DNA in *D. radiodurans* might require little more than protection against protein oxidation and genome multiplicity. I attribute protein protection in *D. radiodurans* to the presence of high levels of manganese complexes[16,34]. In radiation-resistant fungi and cyanobacteria, it is possible that resistance is conferred by different sorts of antioxidants, perhaps based on melanin[44] and trehalose[45], respectively.

Manganese within resistant bacteria

As no clear explanation has been inferred from the genome sequences of IR-resistant organisms for their survival capabilities, there have been no reliable physiological predictors of the ability of a cell to tolerate IR. In 1976, Bruce and colleagues reported that *D. radiodurans* accumulates substantially more manganese than the IR-sensitive bacterium *Micrococcus luteus*[46]. Furthermore, the IR-resistant bacterium *Lactobacillus plantarum*[16,47], which lacks the enzyme superoxide dismutase, and *Synechocystis* sp. PCC 68034 (REF. 48) accumulated exceptionally high levels of manganese[49,50]. The growth of IR-resistant bacteria was reported to be relatively independent of iron compared with IR-sensitive bacteria[16]. This set of reports led my research team to examine the relationship between manganese and iron accumulation and bacterial radioresistance[16]. FIGURE 1b shows the correlation between IR resistance and the ratio of intracellular manganese to iron concentrations for eight bacteria and one archaeon. For example, *D. radiodurans* (manganese:iron = 0.24) accumulates 157 times more manganese and 3.3 times less iron than the IR-sensitive *S. oneidensis* (manganese:iron = 0.0005)[16]. Bacteria with high manganese to iron ratios are extremely resistant to IR-induced protein oxidation, whereas bacteria with low manganese to iron ratios are hypersensitive to protein oxidation[16,32,34] (FIG. 2). Measurement of the accumulation of protein carbonyl groups[34] revealed that Mn^{2+} accumulation in bacteria specifically prevented protein oxidation, but did not affect DSB levels[16]. This led to the hypothesis that manganese-accumulating bacteria have an enhanced capacity to

© 2009 Macmillan Publishers Limited. All rights reserved

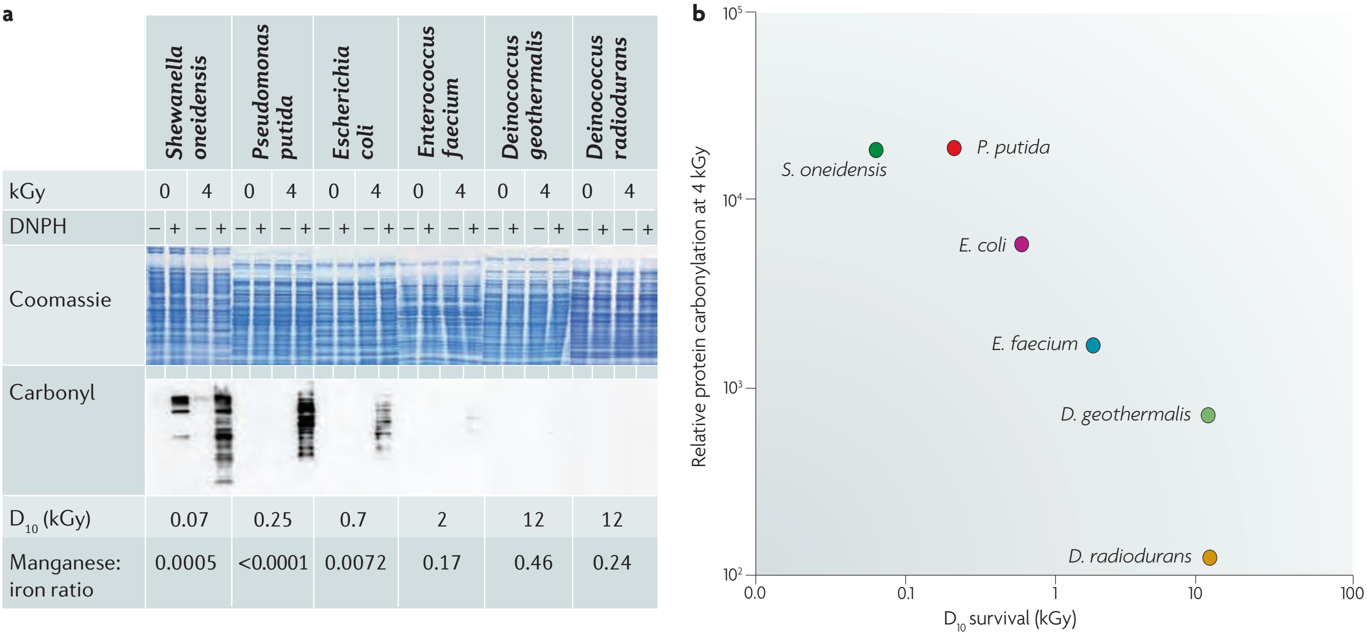

Figure 2 | ***In vivo* ionizing radiation-induced oxidative protein damage.** **a** | Western blot immunoassay of protein-bound carbonyl groups in cell extracts prepared from bacteria irradiated to 4 kGy (^{60}Co at 0 °C; 20 µg of a protein sample (soluble fraction) was loaded per lane). Coomassie represents a coomassie-stained polyacrylamide denaturing gel, whereas carbonyl represents the corresponding western blot, which reveals the presence (black) or absence of protein oxidation (no signal). Carbonyl groups (aldehydes and ketones) are widely used as markers of irreversible protein damage, which cannot be repaired and represents a small fraction of the total oxidative protein damage[34]. **b** | Relationship between bacterial survival (D_{10}; the ionizing radiation dose that reduces the number of viable cells by 90%) and relative protein carbonylation following exposure of the strains to 4 kGy. Carbonylation in relative units was quantified: the intensity profile of a particular lane (part **a**; carbonyl) was generated from previously published digitized membrane images[34]. DNPH, 2,4-dinitrophenylhydrazine. Part **a** is modified from REF. 34.

prevent the formation of iron-dependent ROS (through the Fenton reaction)[16,22,34] (FIG. 3). As the ability of manganese to protect the cell becomes depleted, DNA-repair systems and other essential cellular enzymes are expected to become increasingly subjected to damage[22,34]. This model explains the ability of orthologues from radiosensitive bacteria to complement *D. radiodurans* DNA-repair mutants, in which manganese complexes in *D. radiodurans* shield proteins irrespective of their origin[34]. It also explains the long 'shoulders' of IR-response curves for extremely resistant organisms (FIG. 1a), which are ultimately overwhelmed and killed owing to the hundreds of recalcitrant DSBs per cell[3,12] rather than protein damage[34].

Biological evidence

Direct evidence for a biological role of manganese accumulation in IR resistance comes from studies of *D. radiodurans*, which actively transports manganese into the cell[16]. Mn^{2+} sensing occurs widely in bacteria and influences both Mn^{2+} homeostasis and genes involved in the oxidative stress response[51]. *D. radiodurans* possesses two of the three types of known Mn^{2+} transporters: one from the Nramp (natural resistance-associated macrophage) family and one from the ATP-dependent ABC-type transporter family[16]. The third type of manganese transporter, the unique P-type ATPase with a high specificity for Mn^{2+}, has been detected in *L. plantarum*, but has not been found in *D. radiodurans*. Manganese transport in *D. radiodurans* is predicted to be regulated by a transcriptional regulator of the MntR–DtxR (manganese transport regulator–diphtheria toxin repressor) family by a metal-binding motif that is more closely related to the Mn^{2+}-binding configuration than the Fe^{2+} form[16,51]. Repression of manganese transport in *D. radiodurans* is predicted to be controlled by TroR[16].

When *D. radiodurans* was grown in conditions that limit manganese accumulation, the manganese to iron ratio of the cells decreased from 0.24 to 0.04, and the cells became highly sensitive to IR and highly susceptible to IR-induced protein oxidation, but there were no other obvious effects on other traits[16,34]. Furthermore, when *D. radiodurans* cells with normal manganese levels were irradiated under conditions that were known to perturb manganese-dependent ROS scavenging *in vitro*, such as high pH, the cells lost their ability to prevent IR-induced protein oxidation and became sensitive to IR[34]. Increasing evidence indicates that accumulated manganese ions can act as chemical antioxidant protectants by decreasing the levels of ROS to prevent oxidative stress. This mode of ROS control is not specific for radiation-resistant organisms, but is applicable to diverse settings. Addition of manganese can rescue yeast and bacterial mutants that are deficient in ROS-scavenging enzymes[51–53], and manganese supplementation can restore the lifespan of short-lived *C. elegans* mutants[54]. Mitochondria accumulate high Mn^{2+} concentrations[55], and the level of carbonylated proteins, which is recognized as a marker of oxidative stress, is reduced in animal cells[56] treated with Mn^{2+}.

Manganese and iron in IR resistance

The effects of radiation in cells are mediated primarily through ROS. Dissolved oxygen, either derived from the atmosphere, generated endogenously from IR-induced H_2O_2 by redox-cycling of iron, manganese and/or other biologically relevant transition metals, or produced enzymatically (for example, by the intracellular decomposition of H_2O_2 by catalase), reacts rapidly with electrons released during water radiolysis[11] to form superoxide ($O_2^{\bullet-}$) (BOX 1; FIG. 3). Because $O_2^{\bullet-}$ does not easily cross membranes, it can

© 2009 Macmillan Publishers Limited. All rights reserved

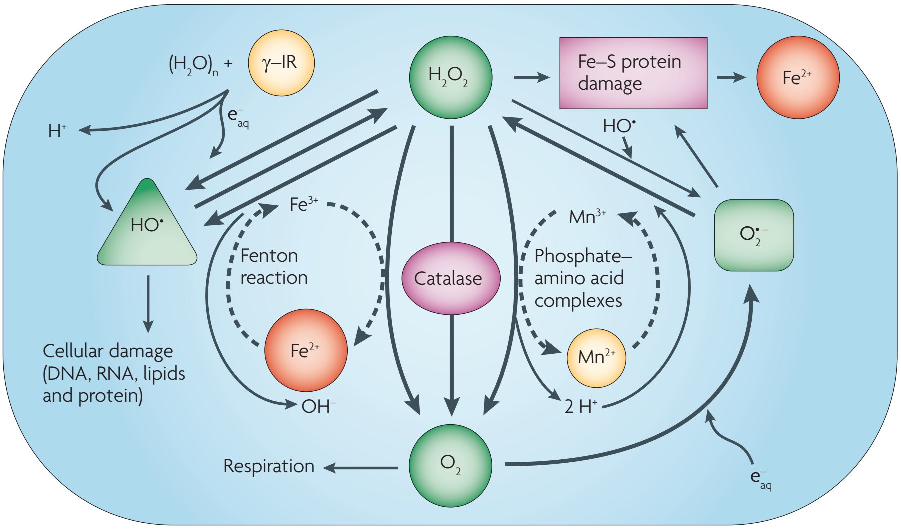

Figure 3 | **Model of ionizing radiation-driven manganese and iron redox cycling.** Under ionizing radiation (IR), Fe^{2+} and Fe^{3+} redox cycling is predicted to generate $HO^{\bullet}$ and $O_2^{\bullet-}$, whereas Mn^{2+} and Mn^{3+} redox cycling is predicted to favour $O_2^{\bullet-}$ scavenging without $HO^{\bullet}$ production. Catalase is a common enzyme that catalyses the decomposition of H_2O_2 to O_2 and H_2O. Thus, under anaerobic or aerobic irradiation conditions, in a cascade of events, $O_2^{\bullet-}$ would damage and inactivate enzymes with exposed Fe–S clusters[57,61], thereby triggering the release of Fe^{2+}, and 'free' Fe^{2+} would promote IR-driven Fe^{2+} and Fe^{3+} redox cycling. Fe–S clusters participate in a range of biochemical processes, including electron transfer, substrate binding and activation, redox catalysis, DNA replication and repair, regulation of gene expression, and tRNA modification. Thus, Mn^{2+} and Mn^{3+} redox cycling is predicted to prevent the proliferation of iron-dependent reactive oxygen species and protect diverse cellular functions[22,34]. Radiolysis of water exposed to ionizing radiation[11,69] (BOX 1), $H_2O + \gamma\text{-IR} \rightarrow HO^{\bullet} + H^+ + e^-_{aq}$; primary radiolytic reaction yielding H_2O_2 (REF. 11), $2\ HO^{\bullet} \rightarrow H_2O_2$; IR-induced superoxide[11], $O_2 + e^-_{aq} \rightarrow O_2^{\bullet-}$; Fenton reaction[61], $H_2O_2 + Fe^{2+} \rightarrow HO^{\bullet} + OH^- + Fe^{3+}$; Haber–Weiss reaction[11], $Fe^{3+} + HO_2^{\bullet} \rightarrow Fe^{2+} + O_2 + H^+$ and $2\ Fe^{3+} + H_2O_2 \rightarrow 2\ Fe^{2+} + O_2 + 2\ H^+$; manganese oxidation[34,49,51,58], $Mn^{2+} + O_2^{\bullet-} + 2\ H^+ \rightarrow H_2O_2 + Mn^{3+}$; manganese reduction[34,49,51], $2\ Mn^{3+} + H_2O_2 \rightarrow 2\ Mn^{2+} + O_2 + 2\ H^+$. Figure modified from REF. 34.

transiently build up in cells under aerobic or anaerobic irradiation conditions[34]. Although $O_2^{\bullet-}$ does not react with DNA or most proteins, it can damage and inactivate enzymes with exposed 2Fe–2S or 4Fe–4S clusters, which contain Fe^{2+} and can therefore trigger the release of ferrous ions (Fe^{2+})[57]. In turn, when Fe^{2+} ions are made accessible to H_2O_2 they can catalyse Fenton reactions[57] (FIG. 3). Thus, $O_2^{\bullet-}$ is not readily consumed by other cellular targets and the damaging potential of $O_2^{\bullet-}$ lingers for Fe–S clusters. I have proposed that Mn^{2+} ions that accumulate in resistant bacteria form complexes that shield Fe–S cluster-containing proteins from $O_2^{\bullet-}$-related ROS generated during irradiation (FIG. 3). By preventing the release of Fe^{2+} from Fe–S-cluster containing proteins, the global collateral damaging effects of Fenton chemistry during irradiation would be minimized, allowing enzyme systems involved in recovery to survive and function efficiently[22,34].

Because Mn^{2+} has a higher reduction potential than Fe^{2+} (REF. 51), Mn^{2+} is less likely than Fe^{2+} to donate an electron to reduce another molecule. Fe^{2+} thus catalyses the Fenton reaction, whereas the analogous reaction between Mn^{2+} and H_2O_2 does not occur[51]. All of the proposed benefits of manganese accumulation compared with iron accumulation in irradiated cells are related to the limitation of Fenton chemistry (FIG. 3). Like iron, manganese can cycle between the divalent and trivalent oxidation states[34,51]. In contrast to Fe^{2+} and Fe^{3+} redox cycling, Mn^{2+} and Mn^{3+} redox cycling during IR is predicted to favour $O_2^{\bullet-}$ scavenging without hydroxyl radical ($HO^{\bullet}$) production[34] (FIG. 3). Because the intrinsic reduction potentials of Mn^{2+} and Fe^{2+} are similar to those of many common biological compounds, each metal can be recruited, by carefully choosing the liganding environment, to perform biologically useful redox catalysis[51]. For example, as Mn^{2+} increases to millimolar concentrations, complex formation with ligands such as phosphate promotes catalytic $O_2^{\bullet-}$ scavenging during IR[34,58]. Indeed, Mn^{2+} efficiently scavenges for ROS at near-millimolar concentrations *in vitro*. Fridovich and Archibald demonstrated that Mn^{2+} scavenges $O_2^{\bullet-}$ in phosphate buffer[49], and Stadtman and colleagues detected Mn^{2+}-dependent disproportionation of H_2O_2 in bicarbonate buffer that contained amino acids[59], which endowed the mixtures with peroxidatic and catalatic activities[51]. Because Mn^{2+} is innocuous under conditions (notably aerobic environments) where Fe^{2+} tends to promote ROS, cells can tolerate high cytoplasmic concentrations of Mn^{2+} with virtually no negative redox consequences[51], unlike other biologically relevant redox-active metals. The IR-resistant bacteria *L. plantarum*, *Synechocystis* sp. PCC 6803 and *D. radiodurans* have intracellular Mn^{2+} concentrations that range from 1 to 30 mM[16,50,51,60]. Electronic paramagnetic resonance spectroscopy[16,46] and X-ray-absorption near-edge structure spectroscopy[34] revealed that manganese exists predominantly as divalent ions in *D. radiodurans*. An ample supply of Mn^{2+} would also ensure that the demands for this cation by manganese-dependent ROS-scavenging enzymes are met[61], and in cells in which iron is limited or sequestered, accumulated Mn^{2+} ions might functionally replace Fe^{2+} as mononuclear cofactors in enzymes, thereby protecting the active sites[61]. Thus, the accumulation of intracellular Mn^{2+} and its cofactors (for example, phosphate) might constitute chemical antioxidant protectants.

Location, location, location

The manganese-based protection hypothesis (FIG. 3) requires a global distribution of this metal within cells. X-ray fluorescence microspectroscopy showed that manganese is dispersed throughout *D. radiodurans* cells, but surprisingly, iron, as well as copper, which also catalyses the conversion of H_2O_2 to $HO^{\bullet}$ *in vitro*, is sequestered between dividing cells (FIG. 4). By contrast, in *S. oneidensis* cells iron is dispersed throughout the cell[62]. Because of the sequestration of iron, the manganese to iron ratio in the cytosol of *D. radiodurans* is probably even higher than 0.24.

Banding patterns for oxidized proteins in carbonyl assays for sensitive bacteria exposed to IR[34] (FIG. 2a) revealed that some proteins are highly oxidized, whereas other proteins in the same cells are undamaged. By contrast, all proteins in *D. radiodurans* cells are resistant to IR-induced oxidation. However, proteins purified from *D. radiodurans* are not resistant to oxidation when irradiated *in vitro*[34]. As H_2O_2 generated during irradiation diffuses throughout a cell, any Fe^{2+} encountered (bound or unbound) would produce a torrent of ROS that would react with their immediate neighbours until either the H_2O_2 was exhausted or the iron was removed[57,61]. Owing to the short lifetime of $HO^{\bullet}$ (BOX 1), $HO^{\bullet}$ cannot damage

© 2009 Macmillan Publishers Limited. All rights reserved

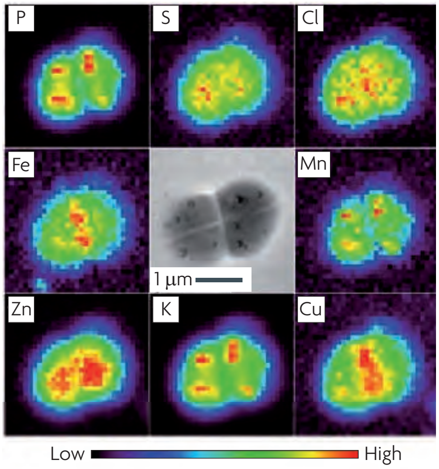

Figure 4 | **X-ray fluorescence maps of the qualitative distribution and concentration gradients of various metals and non-metals in *Deinococcus radiodurans*.** The tetracoccus was harvested from a mid-logarithmic culture in undefined rich medium, imaged and quantified as described previously[26,34]. The central panel is a transmission electron micrograph (TEM) of the tetracoccus after X-ray fluorescence (XRF) imaging. The panels that surround the TEM are the element distribution images of indicated metals and non-metals, which qualitatively depict high and low X-ray fluorescent intensities as indicated by the colour key (bottom): red represents the highest element concentration and black represents the lowest element concentration. A mathematical model of the original morphology of the cells was constructed in approximate likeness to the tetracoccus to determine the distribution of the elements[34]. XRF analysis measurements were made using the hard X-ray microprobe beamline 2ID-D at the Advanced Photon Source, Argonne National Laboratory, Chicago, USA, under previously described conditions[34,62]. Figure modified from REF. 34.

molecules beyond a few Angstroms from the ferrous ion where they are generated. One predicted repercussion is that proteins with predominantly negatively charged surfaces would become the focal points of Fenton chemistry during irradiation because they attract Fe^{2+}. This indicates that iron-dependent IR damage in sensitive cells is highly localized at the molecular level. By contrast, $HO^\bullet$ generated directly by the radiolysis of water would cause global, indiscriminate cellular damage (FIG. 3). The proposed perspective on IR toxicity has converged with a theory of eukaryotic ageing that postulates a causal relationship between oxidative protein damage and the decline of physiological functions and metabolic processes: carbonylation of specific proteins increases almost exponentially as animals age[63]. This model could also help us to understand the toxicity of redox-active heavy metals (for example, Cr^{6+} and Hg^{2+}) and hypochlorous acid (bleach), both of which would also be expected to knock down the activity of proteins they encounter in a selective manner[64].

This leaves the question of why chromosomal and plasmid DNA in extremely resistant bacteria are not similarly protected but instead are damaged with a similar dose dependence as DNA in sensitive bacteria[9]. Only 3% of a typical bacterial nucleoid by mass is DNA; the multitude of small proteins in the nucleoid that are involved in chromosomal condensation provide substantial shielding from $HO^\bullet$ generated by radiolysis. In the presence of cations (for example, Ca^{2+} and K^+), the genome is compacted by 1,000-fold by reducing the electrostatic repulsion between the negative charges of DNA[65,66]. When packaged in IR-resistant or IR-sensitive bacteria, DNA is more than 20 times as resistant to IR damage as when the same DNA is purified and irradiated in water[9]. Thus, DNA that is highly condensed in different organisms might be similarly susceptible to IR-induced DNA damage because of a combination of excluded water volume effects and the abundance of proteins[65,66]. In this context, DNA would remain prone to the direct damaging effects of IR, but might be out of reach of most of the indirect damage caused by $HO^\bullet$ generated by radiolysis (BOX 1). Provided that any iron atoms associated with DNA-binding proteins were tightly coordinated, and any Fe^{2+} or Fe^{3+} released within nucleoids was quickly removed and caged inside proteins, such as DPS (DNA protection during starvation)[67], DNA in bacterial nucleoids would be generally less prone to Fenton chemistry than regions of a cell where proteins with exposed Fe–S groups predominate.

Summary

Based on the radiochemistry of Mn^{2+} *in vivo* and *in vitro*[34,58], I propose that a combination of Mn^{2+} with ligands, such as phosphate[49,58], and perhaps other small molecules, spontaneously form intracellular complexes that act as potent catalytic scavengers[34,49,51,58,59] of $O_2^{\bullet-}$ and related ROS (FIG. 3). Mn^{2+} complexes are therefore expected to provide immediate cytosolic protection from IR-induced ROS. By shielding proteins with exposed Fe–S groups in particular, Mn^{2+} complexes would prevent the proliferation of iron-dependent ROS during irradiation in a way that is not highly dependent on the induction of enzymes, stage of growth or temperature over a range at which cells are metabolically active. It must be emphasized, however, that the fate of an irradiated cell in this 'death by protein damage' model rests not only on the level of protein oxidation, but also on the number of genome copies per cell and genome size. Cells that are proficient at homologous recombination but lack systems that can rejoin random DSB ends[15] (for example, non-homologous end joining) would not benefit from high levels of protein protection during irradiation if their genome copy number was less than two, as one DSB would be lethal. For polyploid cells, however, the presence of chemical antioxidant protectants of DNA-repair proteins would be expected to substantially increase the number of DSBs an organism could survive following irradiation.

A recent report shows that the dose-response relationship for desiccation killing in bacteria isolated from desert environments parallels the levels of protein damage and manganese to iron ratios[32]. Therefore, desiccation resistance also seems to depend on the abundance of cellular manganese complexes that protect proteins. The relative contribution to survival of accumulated Mn^{2+} ions compared with ROS-scavenging processes that have been classically attributed to enzymes has been explored for IR. Based on experimental analyses in *D. radiodurans*, Mn^{2+} accumulation trumps enzymatic ROS defence systems by far[16]; the genes that encode the constitutively and highly expressed ROS-scavenging enzymes *sodA* and *katA* can be disrupted in *D. radiodurans* with little effect on IR resistance, but the Mn^{2+} transporter gene (*nramp*) is essential[26]. The possibility that proteins with exposed Fe–S clusters are the most vulnerable macromolecules in irradiated cells (FIG. 3) warrants careful further investigation because of the practical implications. If the predictive power of IR-induced protein damage on bacterial survival extends to human cells exposed to low doses of IR, various analytical techniques could be applied to biodosimetry, which could be used to identify and quantify proteins that are covalently modified by ROS.

The ability of a complex eukaryotic genome to undergo extensive repair and rebuilding from the most extreme genomic insults is underscored by a recent report that diploid yeast cells can survive 250 IR-induced DSBs per cell[18]. This has led to the idea that higher eukaryotes could be made more resistant to IR. In 1968, Serianni and Bruce reported radioprotection of

© 2009 Macmillan Publishers Limited. All rights reserved

E. coli (strain *B/r*) by a radioresistant factor extracted from *D. radiodurans*, but the nature of the protective agents and their cellular targets were not identified[68]. Based on insights presented in this Opinion article, delivery of *D. radiodurans* Mn^{2+} complexes into cells would be expected to protect proteins from oxidative damage and thereby increase survival. Several prominent questions now remain: can the putative antioxidant Mn^{2+} complexes of *D. radiodurans* be purified and identified and can they be used to prevent radiation injury and other forms of oxidative stress in human cells?

Michael J. Daly is at the Uniformed Services University of the Health Sciences, Bethesda, Maryland 20814, USA.
e-mail: mdaly@usuhs.edu or mdaly@usuhs.mil

doi:10.1038/nrmicro2073
Published online 27 January 2009

1. Thornley, M. J. Radiation resistance among bacteria. *J. Appl. Bacteriol.* **26**, 334–345 (1963).
2. Parashar, V., Frankel, S., Lurie, A. G. & Rogina, B. The effects of age on radiation resistance and oxidative stress in adult *Drosophila melanogaster*. *Radiat. Res.* **169**, 707–711 (2008).
3. Gladyshev, E. & Meselson, M. Extreme resistance of bdelloid rotifers to ionizing radiation. *Proc. Natl Acad. Sci. USA* **105**, 5139–5144 (2008).
4. Horikawa, D. D. *et al.* Radiation tolerance in the tardigrade *Milnesium tardigradum*. *Int. J. Radiat. Biol.* **82**, 843–848 (2006).
5. Johnson, T. E. & Hartman, P. S. Radiation effects on life span in *Caenorhabditis elegans*. *J. Gerontol.* **43**, B137–B141 (1988).
6. Holloman, W. K., Schirawski, J. & Holliday, R. Towards understanding the extreme radiation resistance of *Ustilago maydis*. *Trends Microbiol.* **15**, 525–529 (2007).
7. DiRuggiero, J., Santangelo, N., Nackerdien, Z., Ravel, J. & Robb, F. T. Repair of extensive ionizing-radiation DNA damage at 95 °C in the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.* **179**, 4643–4645 (1997).
8. Kottemann, M., Kish, A., Iloanusi, C., Bjork, S. & DiRuggiero, J. Physiological responses of the halophilic archaeon *Halobacterium* sp. strain NRC1 to desiccation and gamma irradiation. *Extremophiles* **9**, 219–227 (2005).
9. Daly, M. J., Ouyang, L., Fuchs, P. & Minton, K. W. *In vivo* damage and *recA*-dependent repair of plasmid and chromosomal DNA in the radiation-resistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **176**, 3508–3517 (1994).
10. Hutchinson, F. The molecular basis for radiation effects on cells. *Cancer Res.* **26**, 2045–2052 (1966).
11. von Sonntag, C. *The Chemical Basis of Radiation Biology* (Taylor & Francis, London, 1987).
12. Daly, M. J. & Minton, K. W. Interchromosomal recombination in the extremely radioresistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **177**, 5495–5505 (1995).
13. Holliday, R. Early studies on recombination and DNA repair in *Ustilago maydis*. *DNA Repair* **3**, 671–682 (2004).
14. Cox, M. M. & Battista, J. R. *Deinococcus radiodurans* — the consummate survivor. *Nature Rev. Microbiol.* **3**, 882–892 (2005).
15. Daly, M. J. & Minton, K. W. An alternative pathway of recombination of chromosomal fragments precedes *recA*-dependent recombination in the radioresistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **178**, 4461–4471 (1996).
16. Daly, M. J. *et al.* Accumulation of Mn(II) in *Deinococcus radiodurans* facilitates gamma-radiation resistance. *Science* **306**, 1025–1028 (2004).
17. Gerard, E., Jolivet, E., Prieur, D. & Forterre, P. DNA protection mechanisms are not involved in the radioresistance of the hyperthermophilic archaea *Pyrococcus abyssi* and *P. furiosus*. *Mol. Genet. Genomics* **266**, 72–78 (2001).
18. Argueso, J. L. *et al.* Double-strand breaks associated with repetitive DNA can reshape the genome. *Proc. Natl Acad. Sci. USA* **105**, 11845–11850 (2008).
19. Setlow, D. M. & Duggan, D. E. The resistance of *Micrococcus radiodurans* to ultraviolet radiation: ultraviolet-induced lesions in the cell's DNA. *Biochim. Biophys. Acta* **87**, 664–668 (1964).
20. Kish, A. *et al.* Salt shield: intracellular salts provide cellular protection against ionizing radiation in the halophilic archaeon, *Halobacterium salinarum* NRC-1. *Environ. Microbiol.* 5 Jan 2009 (doi:10.1111/j.1462-2920.2008.01828.x).
21. Chan, H. L. *et al.* Proteomic analysis of UVC irradiation-induced damage of plasma proteins: serum amyloid P component as a major target of photolysis. *FEBS Lett.* **580**, 3229–3236 (2006).
22. Ghosal, D. *et al.* How radiation kills cells: survival of *Deinococcus radiodurans* and *Shewanella oneidensis* under oxidative stress. *FEMS Microbiol. Rev.* **29**, 361–375 (2005).
23. Makarova, K. S. *et al.* Genome of the extremely radiation resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol. Mol. Biol. Rev.* **65**, 44–79 (2001).
24. Daly, M. J. & Minton, K. W. Recombination between a resident plasmid and the chromosome following irradiation of the radioresistant bacterium *Deinococcus radiodurans*. *Gene* **187**, 225–229 (1997).
25. Liu, Y. *et al.* Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. *Proc. Natl Acad. Sci. USA* **100**, 4191–4196 (2003).
26. Makarova, K. S. *et al. Deinococcus geothermalis*: the pool of extreme radiation resistance genes shrinks. *PLoS ONE* **9**, e955 (2007).
27. Daly, M. J. & Minton, K. W. Resistance to radiation. *Science* **270**, 1318 (1995).
28. Eltsov, M. & Dubochet, J. Fine structure of the *Deinococcus radiodurans* nucleoid revealed by cryoelectron microscopy of vitreous sections. *J. Bacteriol.* **187**, 8047–8054 (2005).
29. Dimitrova, N., Chen, Y. C., Spector, D. L. & de Lange, T. 53BP1 promotes non-homologous end joining of telomeres by increasing chromatin mobility. *Nature* **456**, 524–528 (2008).
30. Gutman, P. D., Fuchs, P. & Minton, K. W. Restoration of the DNA damage resistance of *Deinococcus radiodurans* DNA polymerase mutants by *Escherichia coli* DNA polymerase I and Klenow fragment. *Mutat. Res.* **314**, 87–97 (1994).
31. Minton, K. W. DNA repair in the extremely radioresistant bacterium *Deinococcus radiodurans*. *Mol. Microbiol.* **13**, 9–15 (1994).
32. Fredrickson, J. K. *et al.* Protein oxidation: key to bacterial desiccation resistance? *ISME J.* **2**, 393–403 (2008).
33. Zahradka, K. *et al.* Reassembly of shattered chromosomes in *Deinococcus radiodurans*. *Nature* **443**, 569–573 (2006).
34. Daly, M. J. *et al.* Protein oxidation implicated as the primary determinant of bacterial radioresistance. *PLoS Biol.* **5**, e92 (2007).
35. White, O. *et al.* Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**, 1571–1577 (1999).
36. Gioia, J. *et al.* Paradoxical DNA repair and peroxide resistance gene conservation in *Bacillus pumilus* SAFR-032. *PLoS ONE* **2**, e928 (2007).
37. Cohen, G. N. *et al.* An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. *Mol. Microbiol.* **47**, 1495–1512 (2003).
38. Ng, W. V. *et al.* Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl Acad. Sci. USA* **97**, 12176–12181 (2000).
39. Koonin, E. V. Chance and necessity in cellular response to challenge. *Mol. Syst. Biol.* **3**, 107 (2007).
40. Harris, D. R., Ngo, K. V. & Cox, M. M. The stable, functional core of DdrA from *Deinococcus radiodurans* R1 does not restore radioresistance *in vivo*. *J. Bacteriol.* **190**, 6475–6482 (2008).
41. Qiu, X. *et al.* Transcriptome analysis applied to survival of *Shewanella oneidensis* MR-1 exposed to ionizing radiation. *J. Bacteriol.* **188**, 1199–1204 (2006).
42. Sweet, D. M. & Moseley, B. E. The resistance of *Micrococcus radiodurans* to killing and mutation by agents which damage DNA. *Mutat. Res.* **34**, 175–186 (1976).
43. Sommers, C. H. & Rajkowski, K. T. Inactivation of *Escherichia coli* JM109, DH5α, and 0157:H7 suspended in Butterfield's phosphate buffer by gamma radiation. *J. Food Sci.* **73**, M87–M90 (2008).
44. Dadachova, E. & Casadevall, A. Ionizing radiation: how fungi cope, adapt, and exploit with the help of melanin. *Curr. Opin. Microbiol.* **11**, 525–531 (2008).
45. Shirkey, B. *et al.* Genomic DNA of *Nostoc commune* (Cyanobacteria) becomes covalently modified during long-term (decades) desiccation but is protected from oxidative damage and degradation. *Nucleic Acids Res.* **31**, 2995–3005 (2003).
46. Leibowitz, P. J., Schwartzberg, L. S. & Bruce, A. K. The *in vivo* association of manganese with the chromosome of *Micrococcus radiodurans*. *Photochem. Photobiol.* **23**, 45–50 (1976).
47. Hastings, J. W., Holzapfel, W. H. & Niemand, J. G. Radiation resistance of lactobacilli isolated from radurized meat relative to growth and environment. *Appl. Environ. Microbiol.* **52**, 898–901 (1986).
48. Domain, F., Houot, L., Chauvat, F. & Cassier-Chauvat, C. Function and regulation of the cyanobacterial genes *lexA*, *recA* and *ruvB*: LexA is critical to the survival of cells facing inorganic carbon starvation. *Mol. Microbiol.* **53**, 65–80 (2004).
49. Archibald, F. S. & Fridovich, I. The scavenging of superoxide radical by manganous complexes: *in vitro*. *Arch. Biochem. Biophys.* **214**, 452–463 (1982).
50. Keren, N., Kidd, M. J., Penner-Hahn, J. E. & Pakrasi, H. B. A light-dependent mechanism for massive accumulation of manganese in the photosynthetic bacterium *Synechocystis* sp. PCC 6803. *Biochemistry* **41**, 15085–15092 (2002).
51. Kehres, D. G. & Maguire, M. E. Emerging themes in manganese transport, biochemistry and pathogenesis in bacteria. *FEMS Microbiol. Rev.* **27**, 263–290 (2003).
52. Sanchez, R. J. *et al.* Exogenous manganous ion at millimolar levels rescues all known dioxygen-sensitive phenotypes of yeast lacking CuZnSOD. *J. Biol. Inorg. Chem.* **10**, 913–923 (2005).
53. Jakubovics, N. S. & Jenkinson, H. F. Out of the iron age: new insights into the critical role of manganese homeostasis in bacteria. *Microbiology* **147**, 1709–1718 (2001).
54. Lin, Y. T. *et al.* Manganous ion supplementation accelerates wild type development, enhances stress resistance, and rescues the life span of a short-lived *Caenorhabditis elegans* mutant. *Free Radic. Biol. Med.* **40**, 1185–1193 (2006).
55. Puskin, J. S. & Gunter, T. E. Ion and pH gradients across the transport membrane of mitochondria following Mn(II) uptake in the presence of acetate. *Biochem. Biophys. Res. Commun.* **51**, 797–803 (1973).
56. Oweson, C., Sköld, H., Pinsino, A., Matranga, V. & Hernroth, B. Manganese effects on haematopoietic cells and circulating coelomocytes of *Asterias rubens* (Linnaeus). *Aquat. Toxicol.* **89**, 75–81 (2008).
57. Imlay, J. A. Iron–sulphur clusters and the problem with oxygen. *Mol. Microbiol.* **59**, 1073–1082 (2006).
58. Barnese, K., Gralla, E. B., Cabelli, D. E. & Valentine. J. S. Manganous phosphate acts as a superoxide dismutase. *J. Am. Chem. Soc.* **130**, 4604–4606 (2008).
59. Berlett, B. S., Chock, P. B., Yim, M. B. & Stadtman, E. R. Manganese(II) catalyzes the bicarbonate-dependent oxidation of amino acids by hydrogen peroxide and the amino acid-facilitated dismutation of hydrogen peroxide. *Proc. Natl Acad. Sci. USA* **87**, 389–393 (1990).
60. Archibald, F. S. & Fridovich, I. Manganese, superoxide dismutase, and oxygen tolerance in some lactic acid bacteria. *J. Bacteriol.* **146**, 928–936 (1981).
61. Imlay, J. A. Cellular defenses against superoxide and hydrogen peroxide. *Annu. Rev. Biochem.* **77**, 755–776 (2008).
62. Marshall, M. J. *et al.* c-Type cytochrome-dependent formation of U(IV) nanoparticles by *Shewanella oneidensis*. *PLoS Biol.* **4**, e268 (2006).
63. Levine R. L. & Stadtman E. R. Oxidative modification of proteins during aging. *Exp. Gerontol.* **36**, 1495–1502 (2001).
64. Winter, J., Ilbert, M., Graf, P. C., Ozcelik, D. & Jakob, U. Bleach activates a redox-regulated chaperone by oxidative protein unfolding. *Cell* **135**, 691–701 (2008).
65. Reyes-Lamothe, R., Wang, W. & Sherratt, D. *Escherichia coli* and its chromosome. *Trends Microbiol.* **16**, 238–245 (2008).
66. Maeshima, K. & Eltsov, M. Packaging the genome: the structure of mitotic chromosomes. *J. Biochem.* **143**, 145–153 (2008).

© 2009 Macmillan Publishers Limited. All rights reserved

67. Cuypers, M. G., Mitchell, E. P., Romão, C. V. & McSweeney, S. M. The crystal structure of the Dps2 from *Deinococcus radiodurans* reveals an unusual pore profile with a non-specific metal binding site. *J. Mol. Biol*. **371**, 787–799 (2007).
68. Serianni, R. W. & Bruce, A. K. Role of sulphur in radioprotective extracts of *Micrococcus radiodurans*. *Nature* **218**, 485–487 (1968).
69. Lin, L.-H. *et al.* The yield and isotopic composition of radiolytic H_2, a potential energy source for the deep subsurface biosphere. *Geochim. Cosmochim. Acta* **69**, 893–903 (2005).
70. Rainey, F. A. *et al.* Extensive diversity of ionizing-radiation-resistant bacteria recovered from Sonoran desert soil and description of nine new species of the genus *Deinococcus* obtained from a single soil sample. *Appl. Environ. Microbiol*. **71**, 5225–5235 (2005).
71. Blasius, M., Sommer, S. & Hubscher, U. *Deinococcus radiodurans*: what belongs to the survival kit? *Crit. Rev. Biochem. Mol. Biol*. **43**, 221–238 (2008).
72. Anderson, A., Nordan, H., Cain, R., Parrish, G. & Duggan, D. Studies on a radioresistant micrococcus. I. Isolation, morphology, cultural characteristics, and resistance to gamma radiation. *Food Technol*. **10**, 575–578 (1956).
73. Fredrickson, J. K. *et al.* Geomicrobiology of high level nuclear waste contaminated vadose sediments at the Hanford Site, Washington State. *Appl. Environ. Microbiol*. **70**, 4230–4241 (2004).
74. Callegan, R. P. *et al.* Description of four novel psychrophilic, ionizing radiation-sensitive *Deinococcus* species from alpine environments. *Int. J. Syst. Evol. Microbiol*. **58**, 1252–1258 (2008).
75. Hirsch, P. *et al. Deinococcus frigens* sp. nov., *Deinococcus saxicola* sp. nov., and *Deinococcus marmoris* sp. nov., low temperature and draught-tolerating, UV-resistant bacteria from continental Antarctica. *Syst. Appl. Microbiol*. **27**, 636–645 (2004).
76. Ferreira, A. C. *et al. Deinococcus geothermalis* sp. nov. and *Deinococcus murrayi* sp. nov., two extremely radiation resistant and slightly thermophilic species from hot springs. *Int. J. Syst. Bacteriol*. **47**, 939–947 (1997).
77. Kimura, H., Asada, R., Masta, A. & Naganuma, T. Distribution of microorganisms in the subsurface of the manus basin hydrothermal vent field in Papua New Guinea. *Appl. Environ. Microbiol*. **69**, 644–648 (2003).
78. Pavlov, A. K., Kalinin, V. L., Konstantinov, A. N., Shelegedin, V. N. & Pavlov, A. A. Was earth ever infected by martian biota? Clues from radioresistant bacteria. *Astrobiology* **6**, 911–918 (2006).
79. Brim, H., Venkateswaran, A., Kostandarithes, H. M., Fredrickson, J. K. & Daly, M. J. Genetic development of *D. geothermalis* for bioremediation of high temperature radioactive waste environments. *Appl. Environ. Microbiol*. **69**, 4575–4582 (2003).
80. Bik, E. M. *et al.* Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl Acad. Sci. USA* **103**, 732–737 (2006).
81. Daly, M. J. Modulating radiation resistance: insights based on defenses against reactive oxygen species in the radioresistant bacterium *Deinococcus radiodurans*. *Clin. Lab. Med*. **26**, 491–504 (2006).

Acknowledgements
Work on protein oxidation in the Daly laboratory is supported by grant FA9550-07-1-0218 to M.J.D. from the Air Force Office of Scientific Research. The author thanks E.K. Gaidamakova and V.Y. Matrosova for assistance with the figures.

DATABASES
Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Bacillus pumilus | Caenorhabditis elegans | Deinococcus geothermalis | Deinococcus radiodurans | Escherichia coli | Halobacterium sp. NRC-1 | Lactobacillus plantarum | Micrococcus luteus | Pyrococcus furiosus | Shewanella oneidensis | Synechocystis sp. PCC 68034 | Ustilago maydis

FURTHER INFORMATION
Michael J. Daly's homepage: http://www.usuhs.mil/pat/deinococcus/index_20.htm

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

OPINION

Antimicrobial peptides: linking partition, activity and high membrane-bound concentrations

Manuel N. Melo, Rafael Ferre and Miguel A. R. B. Castanho

Abstract | An increasing amount of information on the action of antimicrobial peptides (AMPs) at the molecular level has not yet been translated into a comprehensive understanding of effects in bacteria. Although some biophysical attributes of AMPs have been correlated with macroscopic features, the physiological relevance of other properties has not yet been addressed. Pertinent and surprising conclusions have therefore been left unstated. Strong membrane-binding and micromolar therapeutic concentrations of AMPs indicate that membrane-bound concentrations may be reached that are higher than intuitively expected, triggering disruptive effects on bacteria.

Antimicrobial peptides (AMPs) represent a wide range of short, cationic, gene-encoded peptide antibiotics that can be found in virtually every organism[1]. Different AMPs display different properties, and many peptides in this class are being intensively researched not only as antibiotics, but also as antivirals[2,3], templates for cell-penetrating peptides[4], immunomodulators[5] and antitumoural drugs[6].

Despite sharing a few common features (for example, cationicity, amphipathicity and short size), AMP sequences vary greatly, and at least four structural groups (α-helical, β-sheet, extended and looped) have been proposed to accommodate the diversity of the observed AMP conformations[7,8]. Likewise, several modes of action as antibiotics have been proposed, and there is debate about whether the primary target of many of these peptides is the cell membrane or whether the primary target is cytoplasmic invasion and disruption of core metabolic functions[9].

Several bilayer interaction and disruption models have been proposed for those AMPs that depend on membrane interference for their antimicrobial activity[10–12] (FIG. 1). However, it is now becoming obvious that such models might be too rigid to account fully for the many interactions that these small molecules can establish in a complex environment, such as the cell membrane. The limitations of the previously proposed models have been exposed in molecular dynamics simulations of AMPs interacting with phospholipid bilayers. Observations from these studies included multiple coexistent structures (frequently unrelated to clean α-helices or β-sheets), nonspecific peptide–peptide interactions and membrane perturbation dictated by stochastic events[13–15] (FIG. 1). The advantage of this indefinite behaviour is that bacteria seem to find it hard to circumvent AMP action, which is certainly a reason behind the multistep mutations usually required for resistance to evolve[16].

Independently of the chosen membrane perturbation model, an implicit concentration threshold is always required for disruption (FIG. 1). This is supported by several observations, in model systems, of phenomena related to such threshold crossings[13,17–19]. Nevertheless, AMP concentrations that are close to full membrane coverage, which are often considered unphysiological conditions, are a frequent requisite for such observations (TABLE 1).

In this Opinion, we survey reports of threshold events of AMPs and propose a correlation between these events and properties such as bactericidal concentration and membrane binding. We use this relationship to support our view, which reconciles existing interaction models with

> Independently of the chosen membrane perturbation model, an implicit concentration threshold is always required for disruption.

© 2009 Macmillan Publishers Limited. All rights reserved

JOURNAL OF BACTERIOLOGY, Dec. 2005, p. 8047–8054 Vol. 187, No. 23
0021-9193/05/$08.00+0 doi:10.1128/JB.187.23.8047–8054.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Fine Structure of the *Deinococcus radiodurans* Nucleoid Revealed by Cryoelectron Microscopy of Vitreous Sections

Mikhail Eltsov* and Jacques Dubochet

Laboratoire d'Analyse Ultrastructurale, Bâtiment de Biologie, Université de Lausanne, CH-1015 Lausanne, Switzerland

Received 7 July 2005/Accepted 26 August 2005

Transmission electron microscopy revealed that the nucleoid of the extremely radioresistant bacteria *Deinococcus radiodurans* may adopt an unusual ring shape. This led to the hypothesis that the tight toroidal package of the *D. radiodurans* genome might contribute to radioresistance by preventing diffusion of ends of double-stranded DNA breaks. The molecular arrangement of DNA in the nucleoid, which must be determined to test this hypothesis, is not discernible by conventional methods of electron microscopy. We have applied cryoelectron microscopy of vitreous sections and found that the DNA arrangement in *D. radiodurans* differs from toroidal spooling. Diffuse coralline nucleoids of exponentially growing *D. radiodurans* do not reveal any particular molecular order. Electron-dense granules are generally observed in the centers of nucleoids. In stationary-phase cells, the nucleoid segregates from cytoplasm and DNA filaments show locally parallel arrangements, with increasing aspects of cholesteric liquid crystalline phase upon prolonged starvation. The relevance of the observed nucleoid organization to the radiation resistance of *D. radiodurans* is discussed.

Deinococcus radiodurans is a gram-positive, nonsporulating bacterium which usually grows in tetrad form. This organism is interesting because of its extreme resistance to DNA damage induced by ionizing radiation (4). It exhibits detectable survival after irradiation of 15,000 Gy (4) and grows continuously at 60 Gy/h (25). The unusual radiation resistance of *D. radiodurans* results from its ability to repair the genome, containing more than a hundred double-stranded DNA breaks, without mutations and loss of genome integrity (4, 32). The double-stranded DNA break repair in *D. radiodurans* is *recA*-dependent (10), but an explanation for its remarkable efficiency is yet to be found.

Initial analysis of *D. radiodurans* by transmission electron microscopy was performed in the 1970s (39, 40). Analysis focused mainly on the structure of the cell wall, in which a periodic S layer was described and characterized in detail (5, 38, 40). Methods available at that time did not reveal an unusual structure which could contribute to the radioresistance of *D. radiodurans*.

Interest in the organization of *D. radiodurans* has reawakened with recent advances in transmission electron microscopy techniques. Levin-Zaidman et al. (29). applied rapid freezing and freeze-substitution for their study of *D. radiodurans* nucleoids. They found that its nucleoid adopts an unusual ring shape. It is known that in vitro DNA tends to form toroidal aggregates in the presence of condensing agents (21). In these toroids, DNA molecules are condensed in a closely packed hexagonal arrangement (20). Levin-Zaidman at al. assumed that the ring-shaped nucleoid of *D. radiodurans* has basically the same structure. The hypothesis has been made that the toroidal shape facilitates repair of the fragmented genome because its densely packed structure prevents double-stranded DNA break ends from diffusing (13, 29).

The molecular arrangement of the DNA in the *D. radiodurans* nucleoid must be determined in order to test this hypothesis. Conventional methods of electron microscopy do not reveal the arrangement of DNA in bacteria, even when rapid freezing and freeze-substitution study are used because the bacterial chromatin aggregates during dehydration in the organic solvents, which are necessary in such preparations (23). Obtaining the fine structure of bacterial chromatin requires that fully hydrated specimens be observed.

Cryoelectron microscopy of vitreous sections (CEMOVIS) enables transmission electron microscopy observation of fully hydrated biological material. This method relies on vitrification of biological samples by rapid cooling, similar to the technique used in freeze-substitution. Then, instead of replacing the solidified water with an organic solvent and embedding the specimen in resin, native, unstained vitreous samples are cut at −140°C into thin sections which are observed in a cryoelectron microscope at even lower temperatures (2). CEMOVIS has been successfully applied to different biological samples, including bacteria. A number of new structural details not preserved by other methods of electron microscopy have been visualized in the cell envelope and extracellular matrixes of bacteria (3, 30, 31).

Single isolated DNA molecules can be visualized in a thin layer of vitreous water (12), but tracing DNA inside the crowded environment of a living cell remains a challenging task even if the structure is perfectly preserved by vitrification. Only when the DNA filaments are ordered and favorably oriented it is possible to determine their arrangement. CEMOVIS resolved the local DNA package in human and horse spermatozoa (36). It therefore seems probable that, if there is an ordered DNA package in *D. radiodurans*, it should be revealed in vitreous sections.

In this study we used CEMOVIS to examine the structure

---

* Corresponding author. Mailing address: Laboratoire d'Analyse Ultrastructurale, Bâtiment de Biologie, Université de Lausanne, CH-1015 Lausanne, Switzerland. Phone: 41 21 6924289. Fax: 41 21 6924285. E-mail: Mikhail.Eltsov@unil.ch.

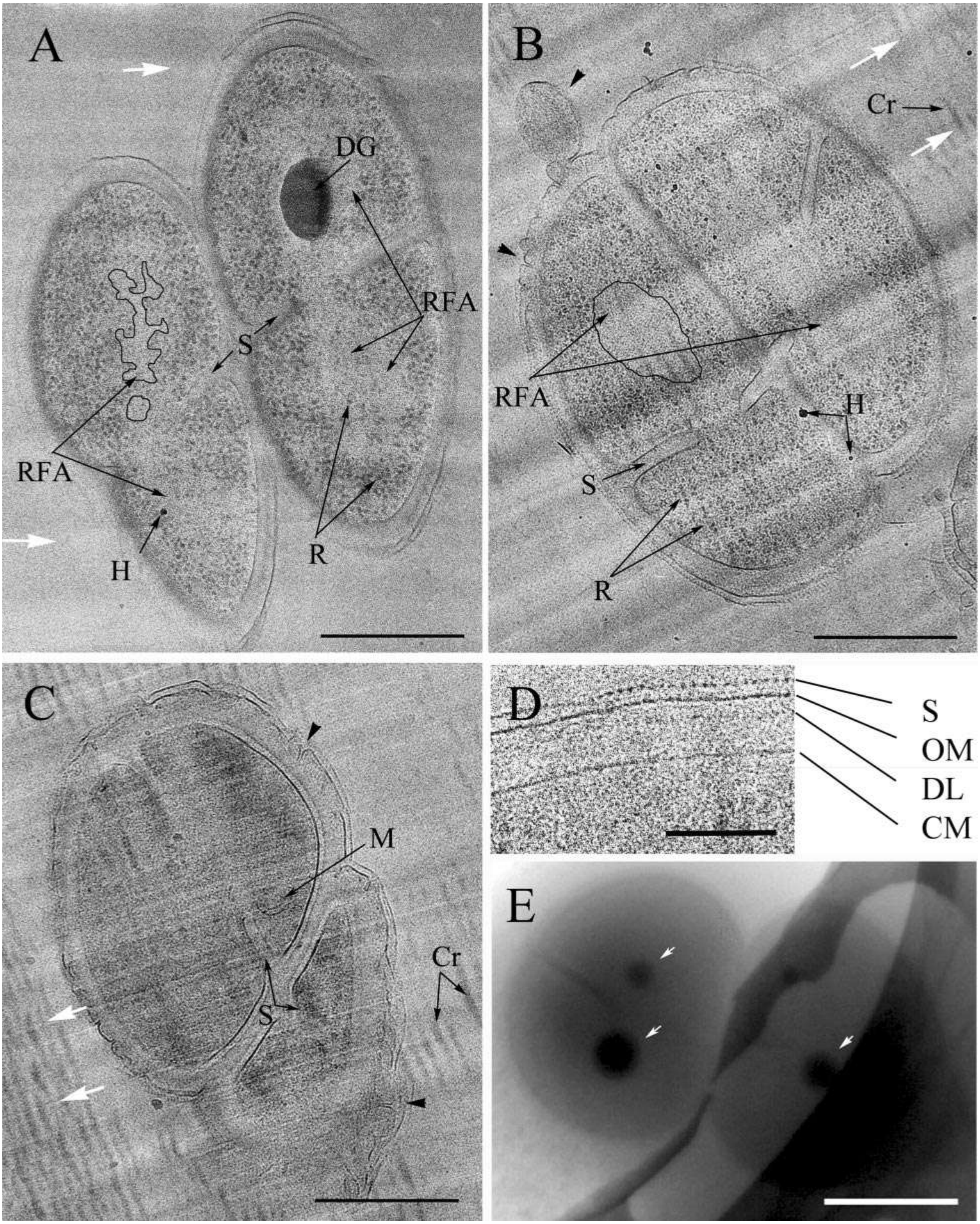

FIG. 1. (A, B, C) Vitreous sections of typical *D*. radiodurans tetrads from exponentially growing (A), stationary-phase (B), and long-stationary-phase (C) cultures were imaged with high defocus (5 to 10 μm) to obtain strong contrast favorable for general morphology mapping. Knife marks (white arrows) and crevasses (Cr) are cutting artifacts. H, surface contamination with hexagonal ice. The cytoplasm of some adjacent cells in the tetrad is interconnected through incomplete septa (S). A large electron-dense granule (DG) is visible inside the exponentially growing cell. The small granules are ribosomes (R). In the central part of the exponentially growing and stationary-phase cells, ribosome-free areas (RFA) are seen and are outlined in one cell of the tetrad. Note the dispersed coralline shape of the ribosome-free areas in exponentially growing cells (A) and the compact roundish shape in stationary-phase cells (B). Some membranous structures (M) are the only distinguishable structures in the highly dense

of nucleoids in exponentially growing and stationary-phase *D. radiodurans*. We found that nucleoids of exponentially growing bacteria have diffuse coralline shapes and do not show a visible molecular order. Electron-dense granules are generally observed in the center of nucleoids. In stationary-phase cells, the nucleoid segregates from the cytoplasm and DNA filaments show locally parallel arrangements, with increasing aspects of cholesteric liquid crystalline phase upon prolonged starvation. The possible relevance of the observed nucleoid organization to the radiation resistance of *D. radiodurans* and to the hypothesis of Levin-Zaidman et al. will be discussed.

MATERIALS AND METHODS

D. radiodurans strain SARK was grown in TGY broth (0.8% Bacto tryptone, 0.1% glucose, 0.4% yeast extract) at 30°C with vigorous shaking. Exponentially growing bacteria were collected after 10 to 12 h of growth at an optical density at 600 nm of 0.4 to 0.6. Stationary-phase and long-stationary-phase bacteria were collected after 4 days and 12 days of incubation, respectively. Bacteria were harvested by centrifugation with a Sorvall RS28S centrifuge for 5 min at 3,000 rpm. A soft pellet of bacteria was suspended with the same volume of 30% dextran (42 kDa, Sigma-Aldrich, St. Louis, MO), mixed by pipetting, and high-pressure frozen within 5 min in a Leica EMPACT (Leica, Vienna, Austria) apparatus. Thin frozen hydrated sections were obtained with a diamond cryoknife (Diatome, Biel, Switzerland) in a Leica FCS Ultracut S cryomicrotome with a nominal cutting feed of 50 nm at −140°C as described previously (3). For plunge freezing, 5 μl of bacterial suspension was placed on a holey carbon grid and frozen by plunging into liquid ethane, as described (1).

Grids containing thin frozen-hydrated sections or vitrified thin layers were mounted in a Gatan 626 cryospecimen holder (Gatan, Warrendale, PA) and observed below −180°C in Philips CM12 and CM100 cryoelectron microscopes (FEI, Eindhoven, The Netherlands) operating at an acceleration voltage of 80 kV. Electron micrographs were recorded on Kodak SO-136 electron image films or a 1K Multiscan charge-coupled device camera (Gatan, Warrendale, PA) or 2K TemCam charge-coupled device camera (TVIPS GmbH, Gauting, Germany). Negatives were digitalized using an Imacon Flextight Precision III scanner (Imacon, Redmond, WA). The contrast of micrographs was adjusted with Adobe Photoshop. No other image correction was performed. Fourier transforms of images were made with the Image J program (National Institutes of Health [http://rsb.info.nih.gov/ij/]).

RESULTS

General morphology of *D. radiodurans*. The cell morphology of *D. radiodurans* was checked immediately before vitrification by phase-contrast light microscopy. In all cultures *D. radiodurans* cells were observed mainly in tetrads, but diplococci were also seen occasionally.

D. radiodurans was well visible in vitreous sections. Figure 1 shows typical tetrads from exponentially growing (A), stationary-phase (B), and long-stationary-phase (C) cultures. The cytoplasm of the adjacent cells in the tetrad was either separated or interconnected through incomplete septa. Cutting artifacts such as compression, knife marks (white arrows), and crevasses are seen in the sections. Compression appears as a shortening of cell dimensions along the direction of cutting, resulting in the elliptical shape of the bacteria. The cutting direction is revealed by series of knife marks (Fig. 1A, B, and C, white arrows), which are irregularities in the thickness of sections originating from unevenness of the cutting edge of the knife. Crevasses are fissures perpendicular to the cutting direction. These cutting-induced artifacts are still frequent in vitreous sections (3) and must be considered during the interpretation of images.

A cell envelope is clearly visible around *D. radiodurans* cells. In regions of favorable orientation three sharp, high-contrast layers and one diffuse dense layer are visible in the cell envelope (Fig. 1A, B, and C). Such organization of the cell envelope is observed in all types of cultures. Figure 1D shows a magnified fragment of the cell envelope. The innermost sharp layer is the cytoplasmic membrane, while the intermediate one is the outer membrane. The outermost layer is the S-layer in which a periodic organization is seen on cross-sections and a hexagonal order on tangential sections (not shown). The diffuse dense layer is seen in the periplasmic space. It is most pronounced in long-stationary-phase cells. The outer membrane and the S-layer surround the entire tetrad and do not enter the septa. Irregularities in the structure of the outer part of the cell envelope, appearing as bubbles and wrinkles of the outer membrane and S-layer, are often observed in vitreous sections (Fig. 1C, black arrowheads).

Internal organization of cells differs in exponentially growing, stationary-phase, and long-stationary-phase cells. Multiple small granules and one or several large electron-dense granules are visible inside exponentially growing cells (Fig. 1A). The small granules are identified as ribosomes from their characteristic size and appearance. Ribosomes are distributed unevenly within the cell volume. The highest concentration of ribosomes is seen at the cell periphery. In the central part of the cell, individual ribosomes or small groups of ribosomes are spaced by ribosome-free areas. In a majority of exponentially growing cells these ribosome-free areas have a dispersed coralline shape (Fig. 1A, outlined).

Electron-dense granules are located in the central part of exponentially growing cells. Electron-dense granules have an elliptical shape in vitreous sections (Fig. 1A, dg). The large diameter of electron-dense granules rarely exceeds 400 nm. The small diameter of the ellipse is oriented along the cutting direction. This suggests that electron-dense granules are spherical bodies deformed into an elliptical shape by cutting-induced compression. In order to check whether electron-dense granules are an artifact of incubation with the cryoprotector, we performed plunge-freezing of whole-mount bacteria directly in culture medium. The thickness of the bacteria frozen in vitreous thin layers is too large for high-resolution imaging with a normal-voltage microscope, but due to their high electron density, electron-dense granules are nevertheless well visible (Fig. 1E, white arrows). This proves that electron-dense granules are a natural component of exponentially growing cells.

content of long-stationary-phase cells (C) in these imaging conditions. Irregularities in the structure of the outer part of the cell envelope (black arrowheads) are often observed in vitreous sections. (D) Magnified fragment of the cell envelope. The cytoplasmic membrane (CM), the diffuse dense layer (DL), the outer membrane (OM), and the outermost periodic S-layer (S) of the cell wall are distinguishable in vitreous sections. (E) Whole-mount exponentially growing tetrad plunge-frozen in culture medium. Note electron-dense granules (white arrows). Scale bars: A, B, and C, 500 nm; D, 100 nm; E, 1 μm.

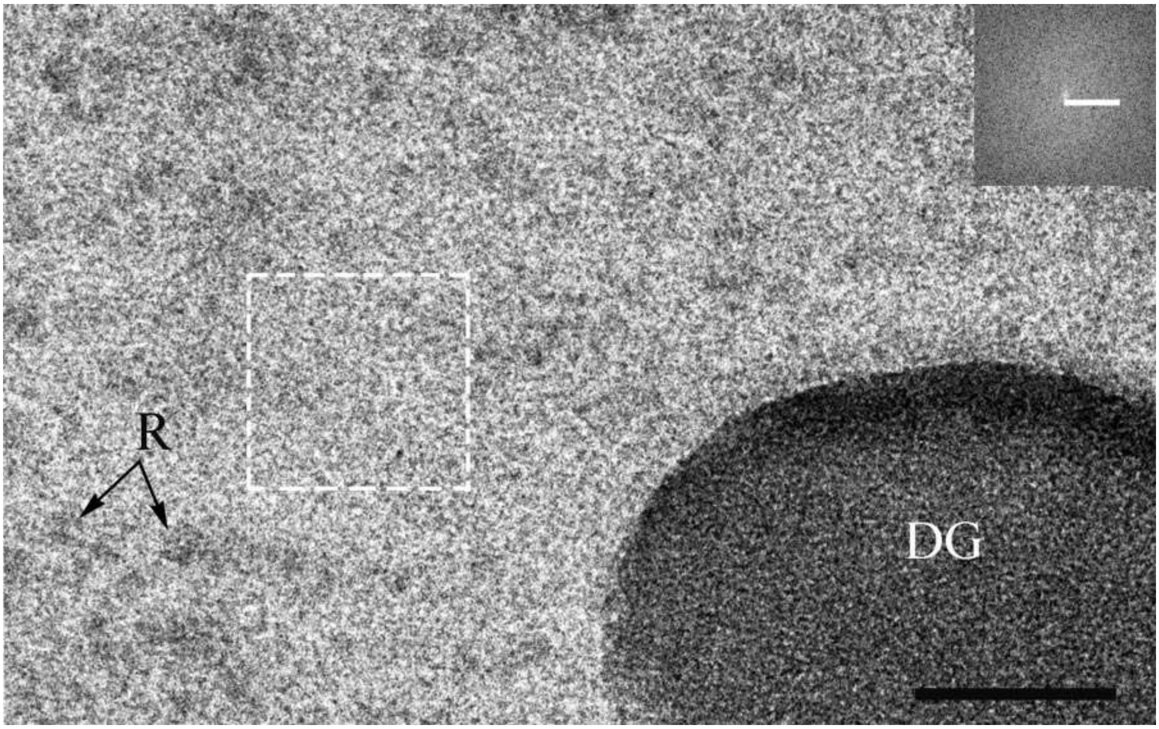

FIG. 2. High-resolution view of the central part of an exponentially growing cell (defocus, 1.8 μm). Electron-dense granule (DG), ribosomes (R), and other granules are visible. The space between the granules is disordered. Inset: Fourier transform of the outlined region. Scale bars: image, 100 nm; Fourier transform, (5 nm)$^{-1}$.

Vitreous biological material is prone to beam damage. Irradiation with the electron beam changes the structure of the biological material. The ultimate state of damage is apparent as bubbling. We found that electron-dense granules begin to bubble simultaneously or soon after cell membranes, whereas ribosome-free areas are the most bubbling-resistant regions of cells.

In typical stationary-phase cells (Fig. 1B) electron-dense granules are not observed and ribosomes are excluded from the central part of the cells, forming large unified ribosome-free areas. These ribosome-free areas have a roundish shape and homogenous contents. Ribosome-free areas in adjacent stationary-phase cells can be interconnected (Fig. 1B).

On a few occasions, in both exponentially growing and stationary-phase cells, we found ring-shaped ribosome-free areas with an island of ribosome-rich cytoplasm in the center.

The content of a typical long-stationary-phase cell (Fig. 1C) has a high electron density, which makes it difficult to identify the internal components. Ribosomes and electron-dense granules are not distinguishable in long-stationary-phase cells. Occasionally some membranous structures are seen inside long-stationary-phase cells.

Interestingly, the bacterial populations in stationary-phase and long-stationary-phase cultures are not completely uniform. About 1% of the cells in a stationary-phase culture and few cells in a long-stationary-phase culture present aspects typical of exponentially growing cells; diffuse ribosome-free areas and electron-dense granules.

Fine structure of the nucleoid seen on vitreous sections. It is known from studies of freeze-substituted bacteria stained with osmium amine (19) that the bacterial nucleoid is located in ribosome-free areas. Hence, the ribosome-free regions visible in vitreous sections of exponentially growing and stationary-phase cells of *D. radiodurans* can be identified with the nucleoid.

Contrary to what happens in stained embedded specimens, the average density of the various regions of *D. radiodurans* is about the same everywhere, except in the dense granules. Imaging conditions favoring phase contrast over large areas must be used in order to obtain a significant contrast between the nucleoid and the rest of the bacteria.

Such conditions, implying high defocus (5 to 10 μm), have been used for recording Fig. 1 A to C. They are, however, not adequate for high-resolution observations. In order to reveal the molecular arrangement of the DNA in the nucleoid, a defocus compatible with the expected distances of few nanometers has been used (<3 μm).

Exponentially growing cells. Figure 2 shows a high-resolution view of the central region of an exponentially growing cell. Electron-dense granules are recognized because of the high contrast. As well as electron-dense granules, ribosomes and other granules, which can be transcription or replication complexes, are visible. The space between granules is disordered, and Fourier transform of the image does not reveal any predominant distance (Fig. 2, inset). Together these results suggest a dispersed, actively transcribed nucleoid without ordered compaction.

Stationary-phase cells. At high resolution, compact, roughly circular ribosome-free areas of stationary-phase cells do not appear to be homogenous any more, and specific textures become apparent in them (Fig. 3A). The "dotted" pattern (Fig. 3B) consists of highly contrasting dots and lines which are clustered and, in several cases, form regularly spaced arrays (Fig. 3B, underlined with white). The "stripy" pattern is formed from longer and partially parallel lines (Fig. 3C) with lower contrast than the dots in the dotted pattern. Figures 3D and 3E show an intermediate case between dotted and stripy patterns which simultaneously contains dots and short and longer lines.

The high contrast of the dots suggests an accumulation of density along most of the section with a thickness averaging 70 nm. The diameter of the dots is in the range of 2 to 2.5 nm. Taken together, these aspects are characteristic of a DNA fragment seen along the viewing direction (6, 27). The dotted pattern can only be seen as a bundle of locally parallel DNA filaments whose direction is perpendicular to the section plan. The stripy patterns confirm the local parallel orientation of the DNA. They appear when the bundle is tilted with respect to the viewing direction. The Fourier transform reveals no reinforced dimensions in the ribosome-rich area (cytoplasm, Fig. 3A, left inset), whereas distances corresponding to 4 to 6 nm are reinforced in the ribosome-free regions. This characteristic order is best seen in the dotted regions (maximum at ca. 4.8 nm; Fig. 3A, right inset) and in the stripy areas.

The regions with local order do not entirely occupy ribosome-free areas. The clusters of dots and lines are spaced by regions without definable structure (Fig. 3B, C, and D, asterisks). The aspects of the local order are less pronounced in some nucleoids (Fig. 3E).

Long-stationary-phase cells. The crowded interior of long-stationary-phase cells does not allow us to distinguish the ribosome-containing area from the ribosome-free area at low magnification. Nevertheless, the specific organization of the central region of the cell revealed at higher resolution is similar to the dotted-stripy pattern observed in stationary phase (Fig. 4). The lines are frequently organized in arches characteristic of a cholesteric arrangement (8). This suggests that the DNA in the nucleoids of long-stationary-phase cells has a higher degree of order than in stationary-phase cells. The Fourier transform shows that the average interfilament distance corresponds to 4 nm (Fig. 4, inset), which is slightly shorter than in stationary-phase cells.

DISCUSSION

CEMOVIS revealed aspects of *D. radiodurans* structure which are not resolved by conventional electron microscopy embedding and sectioning techniques. For example, the periodic structure and hexagonal order of the S-layer of the cell envelope are directly seen in vitreous sections, whereas they were previously observed only on biochemically isolated S-layers or on freeze-etched cells (5, 37, 38).

Another characteristic aspect of native *D. radiodurans* structure is the dense spherical granules (electron-dense granules). Recent structural studies did not report electron-dense granules (9, 29), although early freeze-etching studies described spherical structures reminiscent of electron-dense granules in their size and location (37). Thornley et al. also observed dense granules in resin sections on a few occasions, but their content was often lost during cutting (39). It seems, therefore, that electron-dense granules do exist in *D. radiodurans* but they are optimally preserved only in frozen-hydrated material.

Electron-dense granules are located in the central part of exponentially growing bacteria in the nucleoid region. This addresses a question about the possible chromatin nature of electron-dense granule. However, their high sensitivity to beam-induced bubbling suggests that they do not have a high nucleic acid concentration, since aromatic-rich materials are radiation resistant (11). Due to structural similarity, Thornley et al. (39) associated electron-dense granules with polyphosphate granules observed in other species of bacteria. Phosphate, in combination with proteins, would account for the high density and radiation sensitivity of the granules. The precise composition of electron-dense granules, however, remains to be identified.

CEMOVIS successfully revealed aspects of the molecular arrangement of DNA within *D. radiodurans* nucleoids. The diffuse coralline nucleoids of exponentially growing cells do not show any particular order. We attribute this to the high transcriptional and replication activity required for active growth. A local order first appears in stationary-phase cells in the form of bundles of locally parallel DNA filaments with an average interfilament distance of 4.8 nm. Upon prolonged starvation the aspect of cholesteric liquid crystalline order is observed and the average interfilament distance shortens to 4.0 nm. The gradual increase in the local order together with reduction of the interfilament distance suggests a liquid crystalline organization of the *D. radiodurans* nucleoid.

It is known that liquid crystalline phases of DNA spontaneously assemble in vitro with increasing DNA concentration (26). In bacteria, the increased DNA concentration leading to liquid crystallization may result from accumulation of high-copy plasmids (34), but this is not the case for stationary-phase cells of *D. radiodurans*, in which DNA content per cell is lower than in exponentially growing cells (18). Since at stationary phase the nucleoid segregates to a compact round domain excluding ribosomes, the crowding of DNA can be originated by its redistribution into the confined part of the cell volume. This effect can be related to the fact that cholesteric liquid crystalline nucleoids were found in starving *Escherichia coli* lacking Dps, an abundant starvation-induced unspecific DNA binding protein (16). In contrast, wild-type *E. coli* and strains overexpressing Dps show nucleoid compaction by formation of DNA-Dps cocrystals with a specific structural appearance different from that observed in *D. radiodurans* (15, 16). It can therefore be concluded that the segregation and compaction of the *D. radiodurans* nucleoid can be driven by a decrease of protein-DNA binding in the stationary phase. In addition, an unusually high concentration of Mn^{2+} ions found in *D. radiodurans* (9, 28) can facilitate liquid crystalline compaction by compensating for repulsive forces between DNA molecules.

A minority of cells in stationary-phase cultures of *D. radiodurans* have the typical morphology of exponentially growing cells. We attribute this morphological polymorphism to the presence of mutants gaining a growth advantage during the extended stationary phase (14).

Nucleoid shapes that can be interpreted as rings have been found in only a few cases. This observation argues against the results of other researchers showing an abundance of ring-like nucleoids, which were considered DNA toroids (29). We sus-

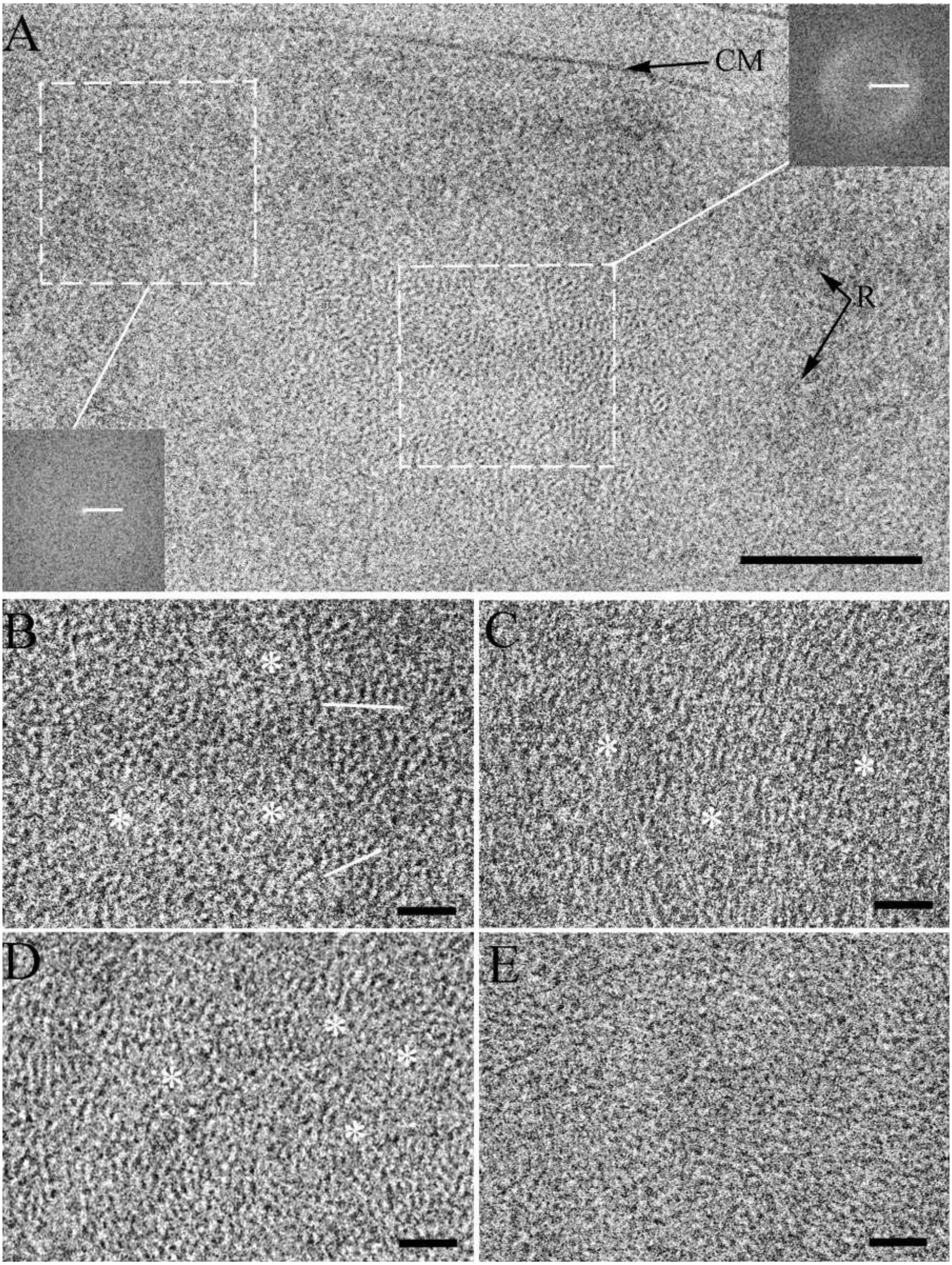

FIG. 3. Fine structure of the nucleoid in stationary-phase cells. A, High-resolution view of stationary-phase cell (defocus, 2 μm). At this defocus the contrast of ribosomes (R) is faded, but fine structural details are visible in the nucleoid. Insets show Fourier transforms of the corresponding marked regions of the cell. CM, cytoplasmic membrane. B, C, D, and E, Different structural textures of the nucleoid. B, Dotted pattern. Regularly spaced arrays of dots are underlined in white. C, Stripy pattern. D and E, Intermediate case between dotted and stripy patterns which simultaneously contains dots and short and longer lines. The clusters of dots and lines are spaced by the regions without definable structure (asterisks). The aspects of the local order are less pronounced in some nucleoids (E). Scale bars: A, 100 nm; B, C, D, and E, 20 nm; Fourier transforms, $(5\ \mathrm{nm})^{-1}$.

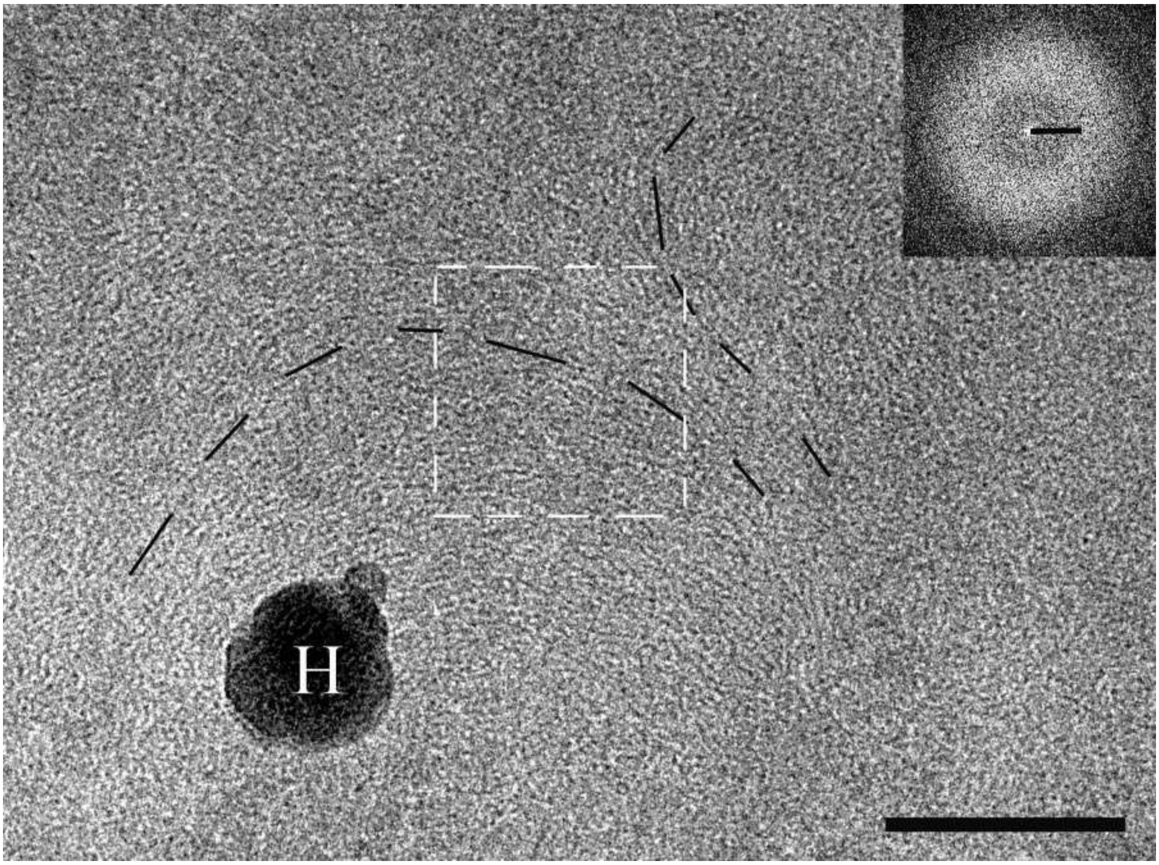

FIG. 4. Nucleoid of long-stationary-phase cells shows the aspect of cholesteric arrangement: arches (dashed lines) formed by dots and lines (defocus, 1.6 μm). H, surface contamination with hexagonal ice. Inset: Fourier transform of the marked region. Scale bars: image, 100 nm; Fourier transform, $(5\ \text{nm})^{-1}$.

pect that conventional studies have misinterpreted the dispersed nucleoids as DNA toroids when the central electron-dense granule is lost during sample preparation and the resolution is insufficient to define the molecular arrangement. Our high-resolution analysis reveals that the DNA arrangement in nucleoids of *D. radiodurans* differs from toroidal DNA spooling both in actively growing and in stationary-phase cultures.

Diffuse coralline nucleoids without a specific molecular order observed in exponentially growing *D. radiodurans* are similar to those of nonradioresistant bacteria (7, 19, 30). Nevertheless, exponentially growing cells of *D. radiodurans* are known to tolerate 5,000 Gy (33). Stationary-phase cultures have a radioresistance approximately three times higher, but this increase already occurs at the beginning of the stationary-phase (24) and is therefore independent of the ordered genome compaction which appears with the aging of the culture. This suggests that the arrangement of the nucleoid does not play a key role in the radioresistance of *D. radiodurans*. Even if it does, the mechanism must differ from the one proposed by Levin-Zaidman et al. (29), in which the dense toroidal package prevents separation of double-stranded DNA break ends, thus favoring efficient repair.

We have found that the dense toroidal package probably does not exist in *D. radiodurans*. Furthermore, we observed that the ordered condensation of DNA, leading to cholesteric organization, always remains changing and dynamic. It might well be that the diffusibility of DNA fragments is reduced in liquid crystals but that is the nature of this type of order; they still remain mobile. This is confirmed by experiments in λ phage DNA cyclization in the presence of polyamines, which show that DNA ends remains mobile in most condensed liquid crystalline phases (22). The hypothesis that the liquid crystalline order of DNA is not directly related to radioresistance is also supported by the fact that dinoflagellata, whose genome is normally in the form of cholesteric liquid crystal (35), are not unusually radioresistant.

The fact that the segregation of the nucleoid from the ribosome-rich cytoplasm is already complete at the early phase of DNA ordering suggests that nucleoid separation is the basic structural change accompanying the transition to the stationary phase. Therefore it could be relevant to the increase in radioresistance at the beginning of stationary phase. This idea is supported by fact that the nucleoids of radiation-sensitive bacteria remain coralline and dispersed at stationary phase (17, 41), whereas the nucleoids of radioresistant bacterial species are more localized (41). We speculate that the segregation of nucleoids reduces the damage caused by free radicals generated in the cytoplasm by radiation (17).

The molecular arrangement of DNA revealed in the nucleoid of *D. radiodurans* cannot directly serve as a structural support for DNA repair. We believe that the unusual efficiency of the latter in *D. radiodurans* is more likely to have a physiological than a structural basis. Nevertheless, nucleoid segregation at the stationary phase can be protective and requires further study.

ACKNOWLEDGMENTS

This work was supported by the 3D-EM Network of Excellence within Research Framework Programme 6 of the European Commission.

We are grateful to H. Engelhardt for supplying bacteria and to A. Minsky for stimulating discussions.

REFERENCES

1. **Adrian, M., J. Dubochet, J. Lepault, and A. W. McDowall.** 1984. Cryo-electron microscopy of viruses. Nature **308:**32–36.
2. **Al-Amoudi, A., J. J. Chang, A. Leforestier, A. McDowall, L. M. Salamin, L. P. Norlen, K. Richter, N. S. Blanc, D. Studer, and J. Dubochet.** 2004. Cryo-electron microscopy of vitreous sections. EMBO J. **23:**3583–3588.
3. **Al-Amoudi, A., L. P. Norlen, and J. Dubochet.** 2004. Cryo-electron microscopy of vitreous sections of native biological cells and tissues. J. Struct. Biol. **148:**131–135.
4. **Battista, J. R.** 1997. Against all odds: the survival strategies of *Deinococcus radiodurans*. Annu. Rev. Microbiol. **51:**203–224.
5. **Baumeister, W., M. Barth, R. Hegerl, R. Guckenberger, M. Hahn, and W. O. Saxton.** 1986. Three-dimensional structure of the regular surface layer (HPI layer) of *Deinococcus radiodurans*. J. Mol. Biol. **187:**241–250.
6. **Bednar, J., P. Furrer, A. Stasiak, J. Dubochet, E. H. Egelman, and A. D. Bates.** 1994. The twist, writhe and overall shape of supercoiled DNA change during counterion-induced transition from a loosely to a tightly interwound superhelix. Possible implications for DNA structure in vivo. J. Mol. Biol. **235:**825–847.
7. **Bohrmann, B., W. Villiger, R. Johansen, and E. Kellenberger.** 1991. Coralline shape of the bacterial nucleoid after cryofixation. J. Bacteriol. **173:**3149–3158.
8. **Bouligand, Y.** 1978. Liquid crystalline order in biological materials, p 261–297. *In* A. Blumstein (ed.), Liquid crystalline order in polymers. Academic Press, New York, N.Y.
9. **Daly, M. J., E. K. Gaidamakova, V. Y. Matrosova, A. Vasilenko, M. Zhai, A. Venkateswaran, M. Hess, M. V. Omelchenko, H. M. Kostandarithes, K. S. Makarova, L. P. Wackett, J. K. Fredrickson, and D. Ghosal.** 2004. Accumulation of Mn(II) in *Deinococcus radiodurans* facilitates gamma-radiation resistance. Science **306:**1025–1028.
10. **Daly, M. J., and K. W. Minton.** 1997. Recombination between a resident plasmid and the chromosome following irradiation of the radioresistant bacterium *Deinococcus radiodurans*. Gene **187:**225–229.
11. **Dubochet, J.** 1975. Carbon loss during irradiation of T4 bacteriophages and *E. coli* bacteria in electron microscopes. J. Ultrastruct. Res. **52:**276–288.
12. **Dubochet, J., J. Bednar, P. Furrer, A. Z. Stasiak, A. Stasiak, and A. A. Bolshoy.** 1994. Determination of the DNA helical repeat by cryo-electron microscopy. Nat. Struct. Biol. **1:**361–363.
13. **Englander, J., E. Klein, V. Brumfeld, A. K. Sharma, A. J. Doherty, and A. Minsky.** 2004. DNA toroids: framework for DNA repair in *Deinococcus radiodurans* and in germinating bacterial spores. J. Bacteriol. **186:**5973–5977.
14. **Finkel, S. E., and R. Kolter.** 1999. Evolution of microbial diversity during prolonged starvation. Proc. Natl. Acad. Sci. USA **96:**4023–4027.
15. **Frenkiel-Krispin, D., I. Ben-Avraham, J. Englander, E. Shimoni, S. G. Wolf, and A. Minsky.** 2004. Nucleoid restructuring in stationary-state bacteria. Mol. Microbiol. **51:**395–405.
16. **Frenkiel-Krispin, D., S. Levin-Zaidman, E. Shimoni, S. G. Wolf, E. J. Wachtel, T. Arad, S. E. Finkel, R. Kolter, and A. Minsky.** 2001. Regulated phase transitions of bacterial chromatin: a non-enzymatic pathway for generic DNA protection. EMBO J. **20:**1184–1191.
17. **Ghosal, D., M. V. Omelchenko, E. K. Gaidamakova, V. Y. Matrosova, A. Vasilenko, A. Venkateswaran, M. Zhai, H. M. Kostandarithes, H. Brim, and K. S. Makarova.** 2005. How radiation kills cells: Survival of *Deinococcus radiodurans* and *Shewanella oneidensis* under oxidative stress. FEMS Microbiol. Rev. **29:**361–375.
18. **Hansen, M. T.** 1978. Multiplicity of genome equivalents in the radiation-resistant bacterium *Micrococcus radiodurans*. J. Bacteriol. **134:**71–75.
19. **Hobot, J. A., W. Villiger, J. Escaig, M. Maeder, A. Ryter, and E. Kellenberger.** 1985. Shape and fine structure of nucleoids observed on sections of ultrarapidly frozen and cryosubstituted bacteria. J. Bacteriol. **162:**960–971.
20. **Hud, N. V., and K. H. Downing.** 2001. Cryoelectron microscopy of lambda phage DNA condensates in vitreous ice: The fine structure of DNA toroids. Proc. Natl. Acad. Sci. USA **98:**14925–14930.
21. **Hud, N. V., and I. D. Vilfan.** 2005. Toroidal DNA condensates: unraveling the fine structure and the role of nucleation in determining Size. Annu. Rev. Biophys. Biomol. Struct. **34:**295–318.
22. **Jary, D., and J. L. Sikorav.** 1999. Cyclization of globular DNA. Implications for DNA-DNA interactions in vivo. Biochemistry **38:**3223–3227.
23. **Kellenberger, E., and B. Arnold-Schulz-Gahmen.** 1992. Chromatins of low-protein content: special features of their compaction and condensation. FEMS Microbiol. Lett. **79:**361–370.
24. **Keller, L. C., and R. B. Maxcy.** 1984. Effect of physiological age on radiation resistance of some bacteria that are highly radiation resistant. Appl. Environ. Microbiol. **47:**915–918.
25. **Lange, C. C., L. P. Wackett, K. W. Minton, and M. J. Daly.** 1998. Engineering a recombinant *Deinococcus radiodurans* for organopollutant degradation in radioactive mixed waste environments. Nat. Biotechnol. **16:**929–933.
26. **Leforestier, A., and F. Livolant.** 1993. Supramolecular ordering of DNA in the cholesteric liquid crystalline phase: an ultrastructural study. Biophys. J. **65:**56–72.
27. **Leforestier, A., H. U. Nissen, and J. Dubochet.** 1995. DNA-DNA interaction in thin layer analysed by cryo-electron microscopy. C. R. Acad. Sci. Ser. III Sci. Vie **318:**1015–1020.
28. **Leibowitz, P. J., L. S. Schwartzberg, and A. K. Bruce.** 1976. The in vivo association of manganese with the chromosome of *Micrococcus radiodurans*. Photochem. Photobiol. **23:**45–50.
29. **Levin-Zaidman, S., J. Englander, E. Shimoni, A. K. Sharma, K. W. Minton, and A. Minsky.** 2003. Ringlike structure of the *Deinococcus radiodurans* genome: a key to radioresistance? Science **299:**254–256.
30. **Matias, V. R., A. Al-Amoudi, J. Dubochet, and T. J. Beveridge.** 2003. Cryo-transmission electron microscopy of frozen-hydrated sections of *Escherichia coli* and *Pseudomonas aeruginosa*. J. Bacteriol. **185:**6112–6118.
31. **Matias, V. R., and T. J. Beveridge.** 2005. Cryo-electron microscopy reveals native polymeric cell wall structure in *Bacillus subtilis* 168 and the existence of a periplasmic space. Mol. Microbiol. **56:**240–251.
32. **Minton, K. W., and M. J. Daly.** 1995. A model for repair of radiation-induced DNA double-strand breaks in the extreme radiophile *Deinococcus radiodurans*. Bioessays **17:**457–464.
33. **Moseley, B. E., and A. Mattingly.** 1971. Repair of irradiation transforming deoxyribonucleic acid in wild type and a radiation-sensitive mutant of *Micrococcus radiodurans*. J. Bacteriol. **105:**976–983.
34. **Reich, Z., E. J. Wachtel, and A. Minsky.** 1994. Liquid-crystalline mesophases of plasmid DNA in bacteria. Science **264:**1460–1463.
35. **Rill, R. L., F. Livolant, H. C. Aldrich, and M. W. Davidson.** 1989. Electron microscopy of liquid crystalline DNA: direct evidence for cholesteric-like organization of DNA in dinoflagellate chromosomes. Chromosoma **98:**280–286.
36. **Sartori Blanc, N., A. Senn, A. Leforestier, F. Livolant, and J. Dubochet.** 2001. DNA in human and stallion spermatozoa forms local hexagonal packing with twist and many defects. J. Struct. Biol. **134:**76–81.
37. **Sleytr, U. B., M. Kocur, A. M. Glauert, and M. J. Thornley.** 1973. A study by freeze-etching of the fine structure of *Micrococcus radiodurans*. Arch. Microbiol. **94:**77–87.
38. **Sleytr, U. B., M. T. Silva, M. Kocur, and N. F. Lewis.** 1976. The fine structure of *Micrococcus radiophilus* and *Micrococcus radioproteolyticus*. Arch. Microbiol. **107:**313–320.
39. **Thornley, M. J., R. W. Horne, and A. M. Glauert.** 1965. The fine structure of *Micrococcus radiodurans*. Arch. Mikrobiol. **51:**267–289.
40. **Work, E., and H. Griffiths.** 1968. Morphology and chemistry of cell walls of *Micrococcus radiodurans*. J. Bacteriol. **95:**641–657.
41. **Zimmerman, J. M., and J. R. Battista.** 2005. A ring-like nucleoid is not necessary for radioresistance in the *Deinococcaceae*. BMC Microbiol. **5:**17.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/7506200

Deinococcus radiodurans – The consummate survivor

Article *in* Nature Reviews Microbiology · December 2005

DOI: 10.1038/nrmicro1264 · Source: PubMed

CITATIONS
408

READS
1,544

2 authors:

Michael M Cox
University of Wisconsin–Madison
284 PUBLICATIONS **16,416** CITATIONS
SEE PROFILE

John R Battista
Louisiana State University
70 PUBLICATIONS **4,833** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project
Pedagogy View project

Project
Radiation resistant E. coli View project

All content following this page was uploaded by Michael M Cox on 14 August 2015.

The user has requested enhancement of the downloaded file.

DEINOCOCCUS RADIODURANS — THE CONSUMMATE SURVIVOR

Michael M. Cox and John R. Battista*[‡]

Abstract | Relatively little is known about the biochemical basis of the capacity of *Deinococcus radiodurans* to endure the genetic insult that results from exposure to ionizing radiation and can include hundreds of DNA double-strand breaks. However, recent reports indicate that this species compensates for extensive DNA damage through adaptations that allow cells to avoid the potentially detrimental effects of DNA strand breaks. It seems that *D. radiodurans* uses mechanisms that limit DNA degradation and that restrict the diffusion of DNA fragments that are produced following irradiation, to preserve genetic integrity. These mechanisms also increase the efficiency of the DNA-repair proteins.

IONIZING RADIATION
Any electromagnetic or particulate radiation powerful enough to strip electrons from atoms to produce ions.

In 1956, Anderson *et al.*[1] isolated a novel vegetative bacterium from canned ground meat that had been γ-irradiated at 4,000 Gray (Gy), a dose that is approximately 250 times higher than that typically used to kill *Escherichia coli*. The authors named this species *Micrococcus radiodurans* because of its superficial morphological similarity to members of the genus *Micrococcus*. However, research on *M. radiodurans* over the next 30 years resulted in reclassification of this species and its closest relatives into a distinct phylum within the domain Bacteria[2–7]. The genus name — *Deinococcus* — was based on the Greek adjective 'deinos', which means strange or unusual[8]; an apt description for an organism with an ability to survive genetic damage that sets it apart from much of the life on Earth.

Treatment of *Deinococcus radiodurans* with high levels of IONIZING RADIATION can produce hundreds of genomic double-strand breaks (TABLE 1), but the genome is reassembled accurately before initiation of the next cycle of cell division. The extraordinary capacity of *Deinococcus* spp. to reconstitute their genomes has inspired a small but growing community of researchers who are interested in the relevant mechanisms that are used to achieve this. In recent years, investigation of the biology and biochemistry of *Deinococcus* spp. has accelerated, catalysed by the availability of genome information and the development of genetic tools. Evidence has accumulated for both passive and enzymatic contributions to genome restitution (FIG. 1), which provides a framework for our discussion in this review.

Ionizing-radiation resistance

Of all the phenotypes associated with prokaryotes, resistance to ionizing radiation is one of the most difficult to rationalize in terms of natural selection. As there are no naturally occurring environments known that result in exposures exceeding 400 mGy per year[9], it is unlikely that species evolved mechanisms to protect themselves against the effects of high-dose ionizing radiation *per se*. Instead, it seems that the damage introduced by γ-irradiation shares features with the damage that results from other stresses to which bacteria have adapted. For example, desiccation introduces many DNA double-strand breaks into the genomes of *D. radiodurans*[10] and members of the cyanobacterial genus *Chroococcidiopsis*[11]. Both organisms are tolerant to desiccation and are resistant to the potentially lethal effects of ionizing radiation, which might indicate that the radioresistance of these species is a fortuitous consequence of their ability to tolerate desiccation-induced strand breaks.

TABLE 2 lists members of six bacterial phyla that are resistant to ionizing radiation. The Deinococcaceae are the best-known family on this list, and have been

**Department of Biochemistry, University of Wisconsin–Madison, Wisconsin 53706-1544, USA.*
[‡]Department of Biological Sciences, Louisiana State University and A&M College, Baton Rouge, Louisiana 70803, USA.
Correspondence to J.R.B.
e-mail: jbattis@lsu.edu
doi:10.1038/nrmicro1264

© 2005 Nature Publishing Group

Table 1 | **Double-strand breaks formed in *E. coli* and *D. radiodurans***

Species*	Genomes per cell	DNA double-strand breaks‡ at D_{37}	Average distance between lesions
Escherichia coli K12	4–5	8–9	530,000 bp
Deinococcus radiodurans R1	8–10	275	10,000 bp

*In each study, actively growing aerobic cultures were irradiated and the number of double-strand breaks determined using neutral sucrose-gradient centrifugation. ‡For *D. radiodurans* this is lesions per unit length of chromosome I (2.64 Mb). Chromosome I is used because it is the largest target within the cell. Calculations are based on a measured rate of 1 DNA double-strand break per 10 Gray per 5×10^9 Da of double-stranded DNA. D_{37}, the dose at which 37% of the cells survive.

studied for five decades. Little has been learnt about the other bacterial species listed, other than their capacity to tolerate ionizing radiation. Draft genome sequences of *Rubrobacter xylanophilus* and *Kineococcus radiotolerans* have been released in the past year, but detailed analyses of these sequences have not yet been reported.

There is no clear pattern of evolution among ionizing-radiation-resistant species (FIG. 2). The scattered appearance of ionizing-radiation resistance among distinct prokaryotic lineages indicates two possibilities. First, radioresistance could be a vestige of DNA-repair mechanisms that were present in ancestral species and have been retained in those organisms that continue to require this phenotype. This explanation assumes that the ancestor's ability to cope with DNA damage has been lost by most descendents, and predicts that the molecular mechanisms of radioresistance should be similar among ionizing-radiation-resistant species. Second, given the infrequent occurrence of ionizing-radiation resistance, it is possible that this phenotype has arisen in unrelated species through horizontal gene transfer, or possibly convergent evolution. Much as birds and bats evolved wings despite distinct evolutionary origins, it is possible that these diverse prokaryotic cells adapted differently to a similar selective pressure and that there might be multiple mechanisms of radioresistance. This latter possibility seems more probable, as fewer independent evolutionary events (the loss of resistance from each species being considered a separate event) would be required to generate the handful of known radioresistant species.

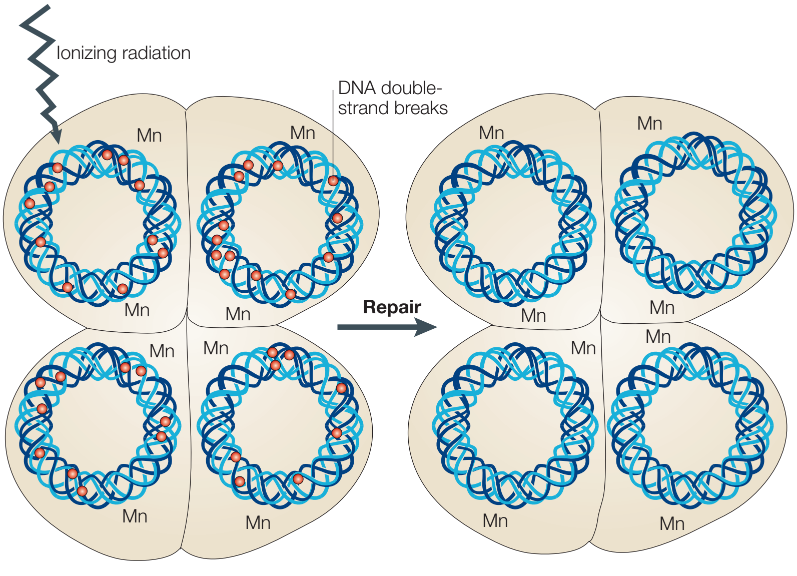

Figure 1 | **Potential contributions to the recovery from radiation damage in *Deinococcus radiodurans*.** The schematic depicts a *D. radiodurans* tetracoccus. The nucleoid in each compartment is highly condensed and maintains its overall architecture after irradiation. High levels of Mn(II) might contribute to the recovery from DNA damage. A wide range of enzymes probably also contribute to genome reconstitution.

Radiation resistance is not restricted to the domain Bacteria. Several HYPERTHERMOPHILIC archaea (members of the Euryarchaeota and Crenarchaeota) show extreme ionizing-radiation resistance[12–15], in some cases (for example, *Thermococcus gammatolerans*) comparable to that of *Deinococcus radiodurans*[14].

The Deinococcaceae

The Deinococcaceae family comprises 11 validly described species — *D. radiodurans*, *Deinococcus proteolyticus*, *Deinococcus radiopugnans*, *Deinococcus grandis*, *Deinococcus radiophilus*, *Deinococcus geothermalis*, *Deinococcus murrayi*, *Deinococcus indicus*, *Deinococcus frigens*, *Deinococcus saxicola* and *Deinococcus marmoris* — all grouped in a single genus, *Deinococcus*[16–18]. Deinococci do not form spores and are non-motile. Most species grow best in rich media at temperatures between 30 and 37°C, with a doubling time between 1.5 and 3 hours. However, *D. geothermalis* and *D. murrayi* are true thermophiles, with optimal growth temperatures of 45–55°C. With the exception of *D. grandis*, which is rod-shaped, all *Deinococcus* species are spherical cells that exist singly or as pairs and tetrads in liquid culture.

To date, *D. radiodurans* has received more attention than the other deinococci. The genome of *D. radiodurans* strain R1 (ATCC BAA-816) has been sequenced[19,20] and can be accessed at the TIGR Comprehensive Microbial Resource database (see Online links box). The *D. radiodurans* chromosome is 3.28 Mb, with a GC content of 66.6%. There are nine types of short nucleotide repeats, ranging in size from 60 to 215 bp, found at 295 sites that are randomly scattered in the genome. The genome is segmented and consists of a 2.64-Mb chromosome (chromosome I), a 0.41-Mb chromosome (chromosome II), a 0.18-Mb megaplasmid and a 0.045-Mb plasmid[21]. *D. radiodurans* has between 4 and 10 genome copies per cell, depending on the stage of the bacterial growth phase[22,23].

The γ-irradiation survival curves of actively growing cultures of *D. radiodurans* R1 have a shoulder of resistance up to 5,000 Gy[24], and until this dose is achieved there is no measurable loss of viability in the irradiated culture. Under these conditions, the D_{37} dose (BOX 1) for *D. radiodurans* R1 is approximately 6,500 Gy. Assuming that there are eight genome copies per cell in these cultures, a 5,000-Gy dose will introduce approximately 1,600 double-strand breaks per cell. As this dose is sublethal, it can be inferred that potentially catastrophic deletions and genome rearrangements occur at low frequencies. Although there is no formal proof, it seems that the process of double-strand-break repair in *D. radiodurans* is error-free.

© 2005 Nature Publishing Group

Table 2 | **Species of ionizing-radiation-resistant bacteria**

Species	Representative D_{10} value*	Phylum	Refs
Methylobacterium radiotolerans	1,000 Gray	α-Proteobacteria	80,81
Kocuria rosea	2,000 Gray	Actinobacteria	4,8
Acinetobacter radioresistens	2,000 Gray	γ-Proteobacteria	82
Kineococcus radiotolerans	2,000 Gray	Actinobacteria	83
Hymenobacter actinosclerus	3,500 Gray	Flexibacter–Cytophaga–Bacteroides	84
Chroococcidiopsis spp.	4,000 Gray	Cyanobacteria	11
Rubrobacter xylanophilus	5,500 Gray	Actinobacteria	85
Deinococcus radiodurans R1	10,000 Gray	Deinococcus–Thermus	86
Rubrobacter radiotolerans	11,000 Gray	Actinobacteria	85

*The D_{10} defines the dose that is needed to eradicate 90% of the irradiated population. The D_{10} values for a given species vary substantially depending on growth conditions. The values provided are for actively growing oxygenated cultures.

TABLE 1 compares the number of double-strand DNA breaks that occur in the large chromosome of *D. radiodurans* R1 (REF. 25) and the *E. coli* K12 (REF. 26) genome at the D_{37} radiation dose. The difference is striking: compared with *E. coli* K12, *D. radiodurans* tolerates approximately 30-fold more DNA double-strand breaks before succumbing to the damage. Clearly, *D. radiodurans* R1 cells have mechanisms to avoid the lethal effects of double-strand breaks that are absent in *E. coli*.

Whatever the mechanisms that *D. radiodurans* uses, they do not prevent DNA damage. Recent evidence from two groups of researchers has shown that measurable double-strand breaks form at the same rate in *E. coli* and *D. radiodurans* if cultures are irradiated under identical conditions[27,28], which makes the existence of a mechanism that blocks the formation of strand breaks in the *D. radiodurans* genome unlikely. In contrast to most other cells, *D. radiodurans* has the ability to efficiently and accurately repair that damage.

The observation that *D. radiodurans* cannot passively prevent DNA damage does not rule out the possibility that other passive mechanisms contribute to the capacity of the cell to tolerate DNA damage. Experimental evidence indicates that the recovery of *D. radiodurans* from substantial DNA damage relies on both passive features of deinococcal physiology and a robust complement of repair enzymes. There are several mechanisms that have the potential to contribute to ionizing-radiation resistance. For example, there is a clear requirement for RecA-dependent homologous recombination — this process has been examined and reviewed extensively[29–36].

In this review, we do not attempt to attribute the contribution to recovery from DNA damage that is made by each different mechanism, and the reader should bear in mind that new recovery mechanisms might still await discovery.

Passive contributions to radioresistance

Genome copy number. Cells with increased numbers of genome copies have enhanced resistance to ionizing radiation[26,37]. The extra genetic material is thought to protect the cell in two ways. First, when multiple genomes are present, there are additional copies of crucial loci that improve the chance of the cell surviving irradiation. The probability of inactivating a specific gene in an organism is given by $P=1-(1-1/M)^L$ (where M is the total number of genes present, and L is the number of inactivating lesions per genome). For a specific dose of radiation, the probability of inactivating all the copies of a specific gene is equivalent to P^N, where N is the number of gene copies present (assuming that

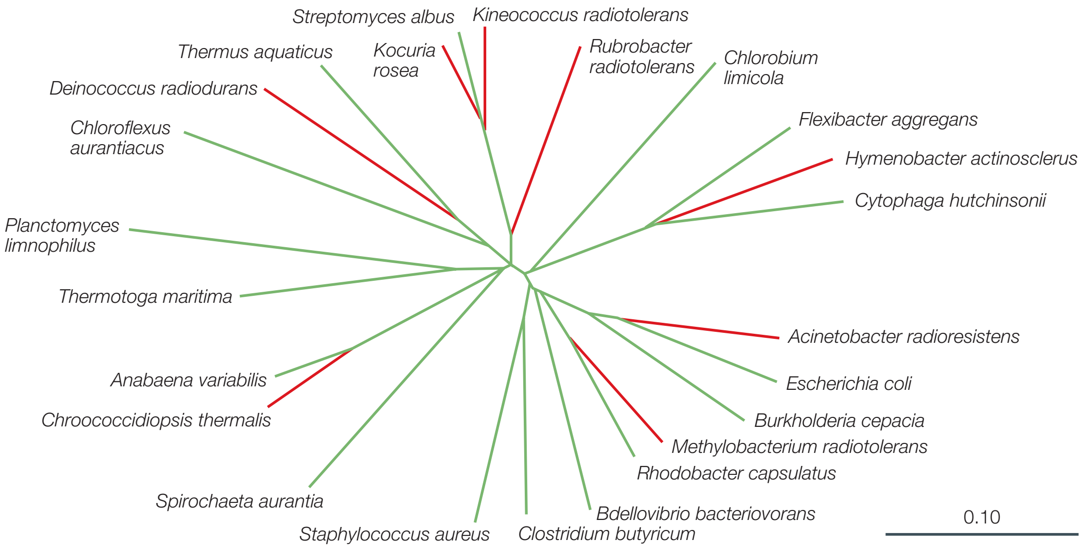
Streptomyces albus
Kineococcus radiotolerans
Thermus aquaticus
Kocuria rosea
Rubrobacter radiotolerans
Chlorobium limicola
Deinococcus radiodurans
Flexibacter aggregans
Chloroflexus aurantiacus
Hymenobacter actinosclerus
Cytophaga hutchinsonii
Planctomyces limnophilus
Thermotoga maritima
Acinetobacter radioresistens
Anabaena variabilis
Escherichia coli
Chroococcidiopsis thermalis
Burkholderia cepacia
Methylobacterium radiotolerans
Rhodobacter capsulatus
Spirochaeta aurantia
Bdellovibrio bacteriovorans
Staphylococcus aureus
Clostridium butyricum
0.10

Figure 2 | **A 16S-rRNA-gene-sequence-based phylogeny of the main lineages of the domain Bacteria.** The branches in red are those in which ionizing-radiation-resistant taxa have been described. The scale bar represents 10 inferred nucleotide substitutions per 100 nucleotides.

HYPERTHERMOPHILIC
Organisms that have an optimal growth temperature above 80°C.

© 2005 Nature Publishing Group

Box 1 | High-energy radiation and its effect on DNA

Ionizing radiation is radiation with sufficient energy to ionize molecules. There are two types of ionizing radiation, both produced by the decay of radioactive elements: electromagnetic (X- and γ-radiation that form part of the electromagnetic spectrum that includes visible light and radio waves) and particulate (α- and β-particles). Different types of ionizing radiation deposit energy in matter at different rates. α- and β-particles produce ionization by collisions, depositing their energy within a short range after entering matter. γ-rays are photons that generate ions by several types of energy-absorption events (most commonly by the Compton effect, an increase in the wavelength of electromagnetic radiation when it collides with electrons in matter, see the figure) and can penetrate deeply into a cell or tissue. Ion production is accompanied by the release of energetic electrons (see the figure), and multiple ions and electrons can be generated in one event. The figure shows the tracks of three different types of ionizing radiation. Small dots indicate energy deposition events. The inset depicts the ejection of an electron from an atom to generate an ion, mediated by an encounter with a γ-ray photon. The γ-ray transfers part of its energy to a valance electron, which is thereby ejected from the nucleus to create an ion. The scattered γ-ray can undergo additional Compton effects within the matter.

Ions react with other molecules to produce free radicals on a nanosecond timescale. In water, the products include highly reactive hydroxyl radicals. Typically, each energy deposition event can generate 2–5 radical pairs within a radius of a few nm.

The energy deposited (the dose absorbed by the cell, tissue or other matter) is measured in rads or in Gray (SI measurement; 100 rad = 1 Gray). The radioresistance of microorganisms is compared by measurement of the D_{37} dose (the dose at which 37% of the cells survive). At the D_{37} dose, each cell on average has experienced a lethal event (those that survive are balanced by other cells with two or more lethal events).

The ions and free radicals produced as radiation passes through matter react rapidly and modify molecules. Of all the effects, genome damage probably has the greatest impact on cell viability. The reasons for this are threefold. First, genomic DNA occupies the largest fraction of the cell volume, and will therefore be 'hit' most often. Second, the genome is present in lower copy numbers than other molecules — there is little redundancy. Third, and most importantly, the genome regulates all cellular functions, so loss of any portion is catastrophic for a single-celled organism.

Ionizing radiation generates multiple types of DNA damage. As much of the damage results from the action of hydroxyl radicals, the spectrum of damage is similar to that produced by oxidative damage associated with endogenous aerobic metabolism. In DNA, the nucleobase is most often modified, and many dozens of different structural modifications to the bases have been characterized. However, 10–20% of the time the sugar-phosphate moiety is affected, which can lead to a single-strand break. The 5′ and 3′ ends at these breaks are usually phosphorylated, indicating that one or more nucleotides have been excised at the break site. The 3′ ends sometimes include a glycolate moiety derived from fragmentation of the deoxyribose.

Most of the lesions are accurately repaired by robust DNA-repair systems present in all cells, but double-strand breaks are the most difficult to repair, and therefore the most lethal. They can arise when single-strand breaks occur by chance in close proximity on opposite strands, when a cluster of hydroxyl radicals introduce strand breaks in both strands at one location, or when the cell attempts to enzymatically excise damaged bases present in close proximity on both strands. Double-strand breaks can result in significant loss of genetic information and, if not repaired, will prevent replication of the prokaryotic genome. How an organism deals with double-strand breaks lies at the heart of that organism's capacity to survive radiation exposure.

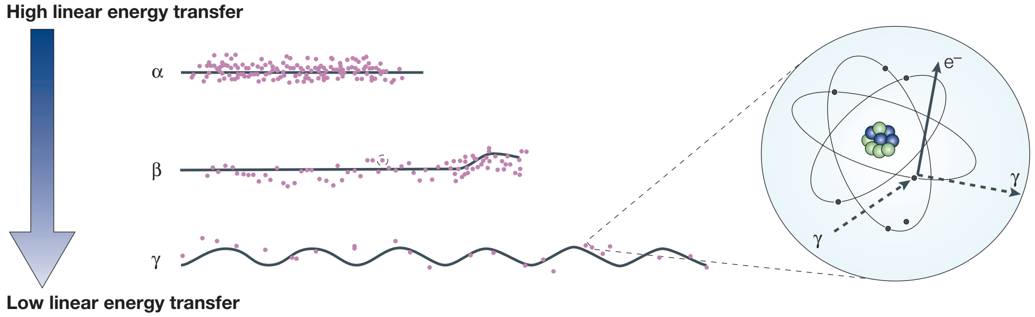

there is one copy of each gene on each genome copy, N is equal to the number of genome copies). If 100 inactivating lesions are randomly introduced into a single copy of the *D. radiodurans* genome (2,897 genes), the probability of inactivating any specific gene on that genome copy is 3.4%. The probability of inactivating all the copies of the same gene is reduced to 0.12% when two genomes are present, and to 0.004% when there are three genomes.

Redundant genetic information also functions as a reserve that can be used to repair DNA segments that are damaged (or degraded) beyond repair. In *E. coli* K12, repair of DNA double-strand breaks only occurs during the exponential growth phase in rich media, when cells contain multiple chromosomes[26], and diploid and tetraploid yeast are more resistant to ionizing radiation than isogenic haploid strains[37]. It has been hypothesized that genetic redundancy contributes

© 2005 Nature Publishing Group

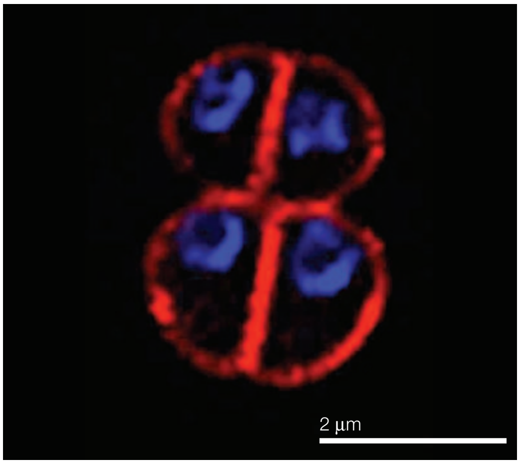

Figure 3 | **A tetrad of *Deinococcus radiodurans*.** An epifluorescence image of stationary-phase *D. radiodurans* R1. DNA is stained with DAPI and appears blue in the figure; the membrane is stained with the dye FM4-64 and appears red.

to the radioresistance of *Deinococcus* spp., although this idea has not been tested. The number of genome copies in *D. radiodurans* has never been reported to be less than four, and cells have the same survival rate whether they contain four or ten genome copies[23]. The contribution of genome redundancy to radioresistance in *D. radiodurans* (relative to other mechanisms) is unclear. Many bacteria contain more than one genome copy, particularly during exponential growth. For example, *E. coli* contains between five and 18 genome equivalents during exponential-phase growth[38], and *Azotobacter vinelandii* can accumulate more than 100 genome copies[39]. Neither species is radioresistant, so genetic redundancy alone does not account for radioresistance. Nevertheless, repair of DNA double-strand breaks by recombination requires the presence of more than one genome[26,37], so it seems probable that genome redundancy is necessary for genome repair.

Nucleoid organization. The nucleoids of stationary-phase *D. radiodurans* cells (FIG. 3) are arranged as a tightly structured ring[40] that remains unaltered by high-dose irradiation[41]. Minsky and colleagues[40] have suggested that this structure passively contributes to *D. radiodurans* radioresistance by preventing the fragments that are formed by double-strand breaks from diffusing apart during repair, which maintains the linear continuity of the genome even when it is fragmented. Although intellectually appealing, this hypothesis is controversial and has been criticized because the prevalence of ring-like nucleoids under different growth conditions does not always correlate with ionizing-radiation resistance[27,41]. A recent examination of the nucleoids of members of the radioresistant genera *Deinococcus* and *Rubrobacter* revealed a high degree of genome condensation — regardless of nucleoid shape — relative to the more radiosensitive species *E. coli* and *Thermus aquaticus*[41], which could indicate that species with a condensed genome might be better protected from ionizing radiation.

Manganese content. Daly and colleagues examined the effects of the high concentrations of Mn(II) that can accumulate in *D. radiodurans*[27] on the capacity of these cells to survive irradiation. When *D. radiodurans* cultures are starved of Mn(II), their resistance to ionizing radiation decreases[27]. The numbers of DNA double-strand breaks formed are the same for a specific dose of ionizing radiation both in the presence or absence of Mn(II), so Mn(II) does not prevent DNA damage[27]. Instead, cellular damage that results from exposure to high radiation doses is better tolerated if Mn(II) is present.

Most DNA damage that occurs after exposure to ionizing radiation results from the generation of reactive oxygen species (ROS) and the hydrolytic cleavage of water. Intracellular Mn(II) can be protective by scavenging ROS. For example, *Lactobacillus plantarum* lacks the protective enzyme superoxide dismutase, and instead substitutes with intracellular Mn(II) concentrations of 20–25 mM[42,43]. As the levels of double-strand breaks in *D. radiodurans* seem to be unaffected by Mn(II), any scavenging of ROS must protect macromolecules other than DNA. Daly *et al.*[27,44] have proposed that Mn(II) accumulation prevents superoxide and related ROS that are produced during irradiation from damaging proteins. If this is the case, bacteria that do not accumulate sufficient Mn(II) might succumb to ionizing-radiation-induced protein damage before DNA is significantly damaged[44].

Alternatively, the increased Mn(II) concentration could contribute to the condensation of the *D. radiodurans* genome[40,45]. DNA can be condensed *in vitro* by adding multivalent cations to an aqueous solution of DNA — the cations neutralize the repulsion of phosphate groups in the DNA backbone[46]. In this way, the proposals of Daly[27,44] and Minsky[40,45] could be related and have similar positive consequences in the context of genome reconstitution.

Regulation of cellular responses to extensive radiation damage. When *D. radiodurans* is exposed to ionizing radiation, a well characterized sequence of physiological events take place, including rapid cessation of DNA replication. At sublethal doses, the duration of the replication delay always exceeds the time that is required to repair the DNA damage that caused inhibition of replication[47–49]. The capacity to inhibit DNA replication is not unlike the DNA-damage checkpoints of eukaryotes — mechanisms that sense DNA damage and initiate a delay in the cell cycle until the damage is repaired[50]. However, the existence of a DNA-damage checkpoint operating in *D. radiodurans* has not been formally established.

Potential enzymatic contributions to repair

Ultimately, DNA strand breaks must be enzymatically repaired, and *Deinococcus* spp. can use novel repair processes. In 1996, Daly and Minton provided evidence for a genome-repair process that involved a temporal progression through at least two distinct stages[31]. Substantial chromosome repair occurs during the first

© 2005 Nature Publishing Group

Table 3 | **Selected *Deinococcus radiodurans* proteins with putative functions in DNA repair**

Protein	Region of homology	% identity	% similarity	Comments
Ligase (DnlJ)	11–669/671 *Escherichia coli*, 29–681/708 *D. radiodurans*	42	57	-
PolA	5–924/928 *E. coli*, 43–952/965 *D. radiodurans*	35	49	-
PriA	202–729/731 *E. coli*, 404–922/924 *D. radiodurans*	26	42	-
RecA	4–324/353 *E. coli*, 16–336/362 *D. radiodurans*	57	72	-
RecD	30–598/609 *E. coli*, 218–705/716 *D. radiodurans*	27	40	N-terminal extension in *D. radiodurans*
RecF	3–330/358 *E. coli*, 6–326/360 *D. radiodurans*	28	43	-
RecG	6–693/694 *E. coli*, 107–777/785 *D. radiodurans*	39	53	N-terminal extension in *D. radiodurans*
RecJ	68–570/579 *E. coli*, 3–461/685 *D. radiodurans*	34	51	C-terminal extension in *D. radiodurans*
RecN	2–553/553 *E. coli*, 34–564/564 *D. radiodurans*	31	49	-
RecO	7-157/242 *E. coli*, 10-159/224 *D. radiodurans*	*18*	*34*	Low homology; required PSI-BLAST
RecQ	9–600/608 *E. coli*, 8–605/824 *D. radiodurans*; 557–605/608 *E. coli*, 680–728/824 *D. radiodurans;* 549–606/608 *E. coli*, 768–825/824 *D. radiodurans*	46, 36, 33	63, 64, 59	HRDC domain repeated three times in *D. radiodurans*
RecR	1–199/202 *E. coli*, 1–196/220 *D. radiodurans*	42	55	C-terminal extension in *D. radiodurans*
RuvA	1–199/203 *E. coli*, 1–197/201 *D. radiodurans*	33	49	-
RuvB	13–332/337 *E. coli*, 2–321/333 *D. radiodurans*	56	75	A second orthologue with weaker similarity is present
RuvC	4–168/174 *E. coli*, 3–166/179 *D. radiodurans*	33	51	-
SbcC	27–1032/1049 *E. coli*, 22–896/909 *D. radiodurans*	21	35	-
SbcD	1–293/400 *E. coli*, 24–319/417 *D. radiodurans*	28	46	C-terminal changes

For each protein, the region of homology is indicated as residue numbers, followed by the total number of residues in the protein. The gene for the single-stranded-DNA-binding protein (SSB) is described in the text. No homology found for DnaC, DnaT, PriB, PriC, RecB, RecC, RecE, RecT or SbcB.

1.5 hours after *D. radiodurans* is treated with a high dose of ionizing radiation, through RecA-independent repair processes. Almost one-third of the double-strand breaks were repaired in this phase. RecA-dependent recombinational DNA repair becomes important several hours after irradiation, and predominates in the later stages of genome reconstitution[31]. Proteins that are known to be potentially important in genome-repair processes are listed in TABLE 3.

A search for novel genes that are induced in response to ionizing radiation and desiccation, using genome-based microarrays, provided new evidence for both RecA-independent and RecA-dependent pathways of double-strand-break repair[51]. In exponentially growing cells, 72 genes are induced three-fold or more after γ-irradiation. Seventy-three loci were induced during recovery from extended desiccation, and 33 of these genes were also induced following irradiation. The five genes most highly induced in response to both stresses were identical and encode proteins of unknown function. Inactivation of these loci — *ddrA*, *ddrB*, *ddrC*, *ddrD* and *pprA* — produces phenotypes that are relevant for genome repair. Genetic analyses defined three EPISTASIS GROUPS that affect ionizing-radiation resistance, and established that two of the loci (*ddrA* and *ddrB*) contribute to radioresistance through different RecA-independent processes. The *pprA* and *recA* loci form a third epistasis group, indicating that the *pprA* gene product interacts with RecA. Identification of these novel loci indicates that there are new mechanisms with an important role in genome repair of *D. radiodurans* post-irradiation.

EPISTASIS GROUP
This occurs when two or more genes control a phenotype. The combined effect of mutations in these genes on a phenotype deviates from the sum of their individual effects.

© 2005 Nature Publishing Group

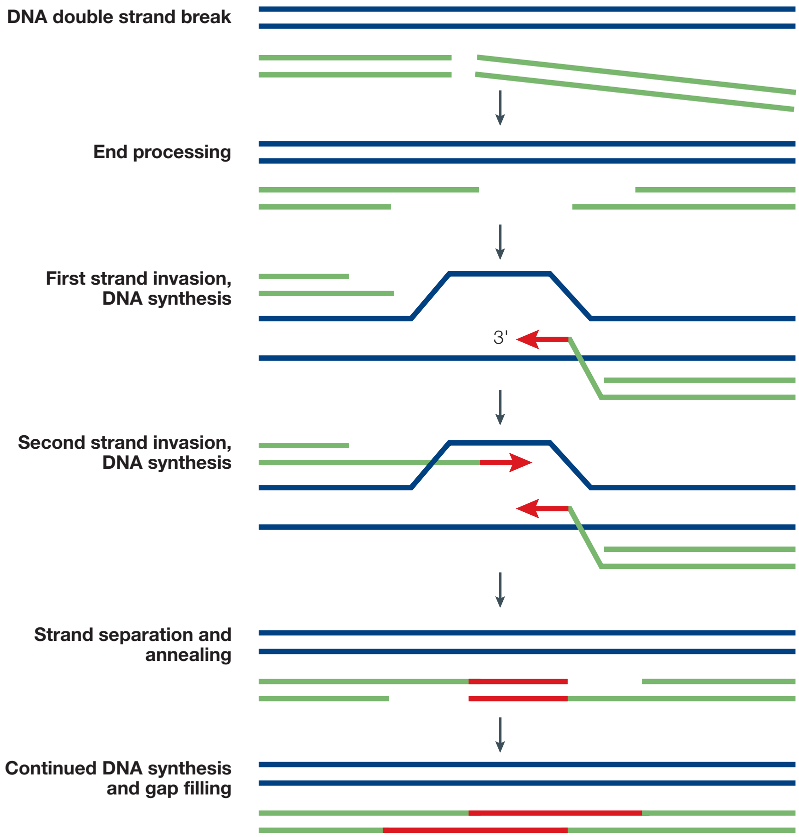
DNA double strand break
End processing
First strand invasion, DNA synthesis
3'
Second strand invasion, DNA synthesis
Strand separation and annealing
Continued DNA synthesis and gap filling

Figure 4 | **Synthesis-dependent strand annealing.** A mechanism of error-free double-strand-break repair that is initiated by creating 3′ overhangs from the ends of the broken DNA duplex (green in the figure). One of these 3′ ends invades a homologous region on an undamaged sister duplex (blue in the figure), priming DNA synthesis and creating a D-loop that acts as a template or DNA synthesis primed by the other 3′ end. If displaced, the newly synthesized DNA can anneal, closing the double-strand break. Newly synthesized DNA is coloured red.

DNA end protection. The *ddrA* locus was the first of the five new genes to be characterized[52]. Deletion of *ddrA* function results in a modest increase in radiation sensitivity in cells grown in rich media. However, when cells are irradiated and then starved, deletion of *ddrA* results in a 100-fold loss of viability over 5 days compared with wild-type cells. The loss in viability of the *ddrA* mutant is accompanied by a dramatic decrease in genomic DNA content by nucleolytic degradation. The DdrA protein binds to the 3′ ends of single-stranded DNA *in vitro* and protects them from nucleolytic degradation[52].

DdrA seems to function as a DNA-end-protection system. As double-strand breaks occur, DdrA (and perhaps other proteins) binds to the exposed DNA ends and prevents nuclease digestion of the chromosomal DNA. This strategy is particularly useful in the genome repair that occurs after desiccation. DNA repair uses a lot of metabolic energy, but cells recovering after desiccation in an environment that lacks nutrients would not have the opportunity to repair DNA damage. As DNA strand breaks occur, nuclease action could degrade genomic DNA. By protecting the broken DNA ends, cells could preserve genomic DNA until conditions become suitable for cell growth and DNA repair.

Bioinformatics initially failed to identify any DdrA homologues, but further bioinformatics investigation revealed that DdrA was distantly related to the eukaryotic Rad52 protein and to the prokaryotic Redβ, Erf and RecT proteins[52,53], so additional functions for DdrA have not been ruled out.

RecA-independent double-strand-break repair. Both NON-HOMOLOGOUS END JOINING (NHEJ) and single-strand annealing (SSA) pathways have been hypothesized to function in *D. radiodurans*. Minsky and colleagues[40,45] suggested that NHEJ would be a useful process for the repair of double-strand breaks in the context of a condensed chromosome, in which ends might not be free to diffuse away from each other. An NHEJ system has been identified in *Bacillus subtilis*[54] and is probably present in other bacteria. Two other laboratories also recently suggested that NHEJ occurs in *D. radiodurans*[55,56]. PprA and PolX are two proteins with predicted activities that are consistent with the existence of NHEJ[55,56]. However, classical NHEJ systems are generally error-prone[57] and seem unsuited to the accurate genome repair that is observed in *Deinococcus* spp. Patterns of recombination between plasmids and the re-circularization of integrated plasmids in irradiated *Deinococcus* cells are not consistent with NHEJ[31].

Plasmid repair and re-circularization of genome-integrated plasmids during the RecA-independent phase of DNA double-strand-break repair in *D. radiodurans* is dependent on homology, indicating that SSA might have a role[31]. Recent research has suggested a compelling model for genome restitution in this species, in which the related but more efficient process called synthesis-dependent single-strand annealing (SDSA) plays a necessary part (M. Radman, personal communication). It provides evidence that *D. radiodurans* R1 uses SDSA as a first step in genome re-assembly. During SDSA, the 3′ end of a strand derived from a DNA double-strand break invades the homologous region of a sister duplex (FIG. 4). The invading 3′ end is used to prime DNA synthesis, unwinding the sister duplex and enlarging the D-loop. The displaced strand in the undamaged complex anneals to the remaining free 3′ end created by the double-strand break. As each 3′ end primes complementary DNA synthesis, the resulting newly synthesized strands can anneal, sealing the breach in the damaged duplex in an error-free manner. This work shows that all RecA-independent genome assembly requires extensive *polA*-dependent DNA synthesis, and that the pattern of nucleotide incorporation, as measured by density labelling post-irradiation, indicates a distributive mode of replication that is consistent with SDSA. The amount of newly synthesized DNA recorded during this phase of post-irradiation recovery indicates that the tails are longer than expected based on precedent in other species that exhibit SDSA. The authors suggest that the long tails assure precise annealing, ensuring an error-free recovery.

NON-HOMOLOGOUS END JOINING
One of several pathways that can be used to repair chromosomal double-strand DNA breaks. The process is non-homologous because adjacent broken strands are fused by direct end-to-end contact without regard to sequence homology. Therefore, non-homologous end joining is error-prone because it results in joining of the breaks without a template.

© 2005 Nature Publishing Group

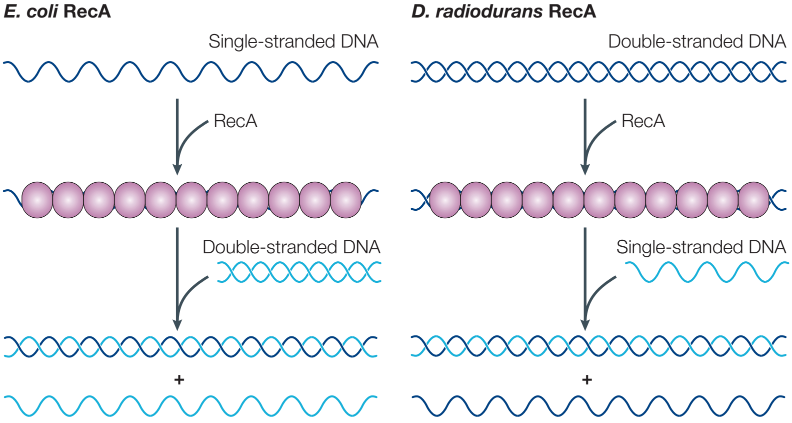

Figure 5 | **Inverse DNA-strand exchange promoted by the *Deinococcus radiodurans* RecA protein.** Instead of forming a filament on single-stranded DNA and then reacting with a homologous duplex, strand exchange is initiated by a filament formed on duplex DNA.

At least one additional RecA-independent pathway mediated by the DdrB protein is present in *D. radiodurans*[51]. Although DdrB has not yet been characterized *in vitro*, deletion of *ddrA* and *ddrB* produces a mutant that is significantly more sensitive to ionizing radiation than either *ddrA* or *ddrB* mutants, indicating that DdrA and DdrB have complementary activities.

Recombinational DNA repair. Any ends that lack overlapping sequence or other means to guide accurate repair require recombinational repair using redundant genome information. Studies of the roles of the classical proteins involved in bacterial recombinational DNA repair in *D. radiodurans* have been initiated (TABLE 3), including studies of recombinase A (RecA), single-strand-binding protein (SSB), recombinase D (RecD), DNA polymerase I, recombinase R (RecR) and recombinase O (RecO).

The *D. radiodurans* RecA protein (361 amino acids, M_r 38,013) is 57% identical (72% similar) to the *E. coli* RecA protein (352 amino acids, M_r 37,842). *In vitro*, the protein promotes all of the key recombinogenic activities of RecA-class recombinases. It forms filaments on DNA, hydrolyses ATP and dATP in a DNA-dependent fashion and promotes DNA-strand exchange[58]. However, the *D. radiodurans* RecA protein has one distinct function. The DNA strand-exchange reactions of the *E. coli* RecA protein, and all other homologues examined to date, are ordered so that the single-stranded DNA is generally bound first, before the double-stranded DNA is bound. This order of DNA binding makes sense from a biological standpoint, as the RecA protein must be targeted to single-strand gaps at stalled replication forks and other damaged DNA sites. By contrast, the *D. radiodurans* RecA protein promotes an obligate inverse DNA-strand-exchange reaction[59], binding the duplex DNA first and the homologous single-stranded DNA substrate second (FIG. 5). It is probable that this reaction pathway reflects the function of *D. radiodurans* RecA in double-strand-break repair, although its significance is not yet clear.

Based on the properties of an E142K mutant of *D. radiodurans* RecA (also known as RecA424), Satoh *et al.*[60] suggested that the effect of *D. radiodurans* RecA on double-strand-break repair primarily reflected a regulatory rather than a recombination function. The mutant protein does not complement the null phenotype of a *recA* knockout in *E. coli*, but does have residual regulatory functions associated with RecA. Strains harbouring the E142K mutation seem to retain resistance to γ-irradiation. However, the E142K mutation does retain significant DNA-strand-exchange activity in some assays, and more work is needed to define the function of RecA in *D. radiodurans*. Assuming that *D. radiodurans* must locate and splice together overlapping segments of its chromosomes to reconstruct a functional genome, a DNA-pairing activity such as that provided by RecA would be at the centre of such a process.

The *ssb* gene in *D. radiodurans* was originally annotated as a tripartite gene, with two frameshifts required to form a functional open reading frame (ORF) for translation[20,61–63]. However, this locus seems to be specific to the *D. radiodurans* R1 strain that was sequenced (ATCC BAA-816), as it differs substantially from that of the *D. radiodurans* R1 type strain (ATCC 13939). In the type strain, the *ssb* locus is a single continuous ORF that encodes the largest bacterial SSB polypeptide identified to date. It has two oligonucleotide/oligosaccharide-binding (OB) folds rather than the one present in most bacterial SSB proteins, and it functions as a dimer rather than a tetramer[64]. The gene is highly homologous to *ssb* genes found in *Thermus* species, to which *D. radiodurans* is closely related. The *D. radiodurans* SSB protein is efficient at stimulating the DNA-strand exchange that is promoted by RecA proteins from both *E. coli* and *D. radiodurans*, being more active than the *E. coli* SSB in both cases. The atomic structure of *D. radiodurans* SSB has been solved[65]. This protein could be important in chromosome repair. The concentration of the *D. radiodurans* SSB is an order of magnitude higher than the normal *in vivo* levels of the SSB protein in *E. coli*.

Less is known about other classical repair proteins in *D. radiodurans*. In the absence of a *recB* or *recC* gene, the *recD* gene product has been characterized. The *D. radiodurans* RecD protein has a 5′ to 3′ helicase activity, in common with the RecD subunit of the *E. coli* RecBCD enzyme[66]. It can unwind short (20 bp) duplexes if a 5′ single-strand tail is adjacent. The atomic structure of the *D. radiodurans* RecO protein has been solved[67], but nothing is known about its function. Similarly, the structure of the *Deinococcus* RecR protein has been determined[68].

Narumi and colleagues have shown that purified PprA protein can bind to duplex DNA with strand breaks *in vitro*, protect the strands from nuclease digestion and facilitate the ligation of duplex DNA fragments[56]. They speculate that PprA is part of a deinococcal NHEJ system, although this role seems inconsistent with the absence of mutations observed after irradiation in this species. As with most *D. radiodurans* proteins, PprA characterization is at an early stage.

© 2005 Nature Publishing Group

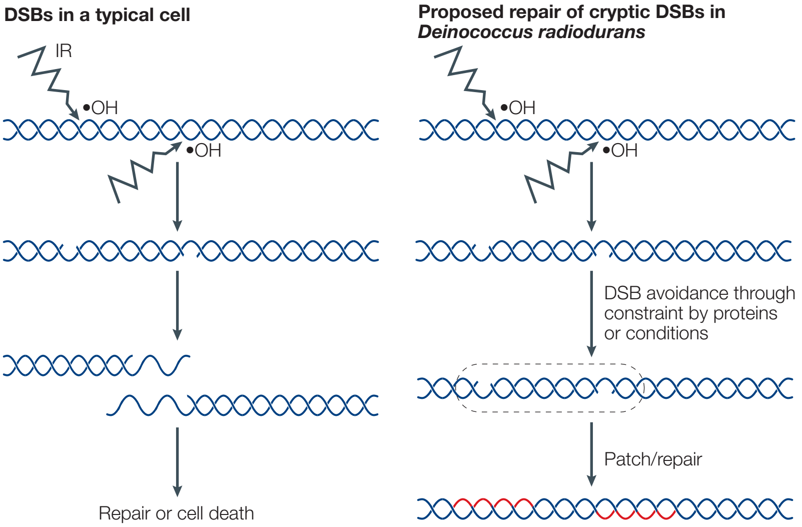
DSBs in a typical cell
IR
•OH
•OH
Repair or cell death
Proposed repair of cryptic DSBs in *Deinococcus radiodurans*
•OH
•OH
DSB avoidance through constraint by proteins or conditions
Patch/repair

Figure 6 | **A proposed mechanism that might contribute to the tolerance of radiation damage in *Deinococcus radiodurans*.** Potential double-strand breaks (DSBs) are constrained, held together by proteins and/or their local environment, so that many breaks scored as DSBs in *in vitro* assays are cryptic in the cell. IR, ionizing radiation; •OH, hydroxyl radical.

A new proposal for radiation tolerance

To the list above, we add one more mechanism, which is a new proposal (FIG. 6). It is particularly difficult to explain the accurate repair of the hundreds of scattered repeat elements in the *D. radiodurans* genome[20,63,69] after the introduction of hundreds of double-strand breaks. How does *D. radiodurans* avoid the potentially disruptive recombination events that could occur between these repeated elements to maintain the continuity of its genome? We speculate that *D. radiodurans* uses a strategy that potentially combines elements of many of the systems described in this review to avoid this problem — including the condensed nucleoid structure first highlighted by Minsky[40,45] and the high concentrations of Mn(II) ions discussed by Daly[27,44] — thereby improving the cell's capacity to tolerate ionizing radiation.

Many DNA double-strand breaks arise from the juxtaposition of two single-strand breaks that form as a function of the distance in base pairs between these breaks (BOX 1). In general, the further apart the single-strand breaks, the less likely it is that the DNA ends will separate to form a double-strand break. We suggest that a significant contribution to the observed tolerance of ionizing radiation in *D. radiodurans* could arise if many of the measured double-strand breaks that occur are cryptic *in vivo*. In other words, some fraction of the measured double-strand breaks could actually be held together so that the separation of the DNA ends never really occurs in the cell. This proposal assumes that the organism has a mechanism to stabilize opposed breaks, constraining the intervening base pairs so that actual separation of the two DNA ends does not occur. This idea is distinct from proposals[40,45] in which the ends separate and are repaired by NHEJ. The cryptic breaks we propose would be scored as double-strand breaks during the processing and analysis of genomic DNA because the conditions necessary to stabilize the paired single-strand breaks are lost when the cell is disrupted. However, in *situ* the DNA strands are not separated. Unlike classical NHEJ, repair of these cryptic breaks would be effectively templated, as genomic continuity would never be lost.

This proposal envisions a specialized system operating in the condensed chromosome that prevents the formation of double-strand breaks at those sites where some base pairing is present. Opposed single-strand breaks would be repaired using a set of enzymes that could deal with the close proximity of the breaks in opposing strands without affecting this base pairing. By this mechanism, the linear continuity of the genome sequence, including regions rich in short nucleotide repeats, would be preserved at many potential sites of double-strand breaks in a manner that is error-free. Experimentally, this process would be indistinguishable from a non-mutagenic type of NHEJ. If such a system exists, the repair would be accurate and RecA-independent, passively reducing the cell's dependence on recombinational DNA repair and the accompanying homology search at any break site that is stabilized in this manner.

Examples from the literature indicate that the cell might use two alternative, but not mutually exclusive, mechanisms to stabilize base pairing between opposed single-strand breaks. First, *D. radiodurans* might encode proteins that hold the DNA together. Second, the intracellular ionic composition could be sufficient to physically limit dissociation of DNA base pairs. If proteins are responsible, we assume that they will be functionally analogous to the structural-maintenance-of-chromosomes (SMC) proteins that are present in many eukaryotic and prokaryotic species[70–72]. In eukaryotes, these proteins are referred to as cohesins and condensins, and the significance of their role in genome stabilization and DNA repair is becoming apparent[73]. The stability of annealed complementary DNA is dependent on the ionic strength of the medium in which the DNA is dissolved, and increased intracellular Mn(II) concentrations[27] might help to hold DNA that contains several single-strand breaks together. Freifelder and Trumbo[74] have shown that high-ionic-strength media stabilize opposed breaks separated by as little as two base pairs

A system for the repair of cryptic double-strand breaks might exist in many organisms. Conceptually, the damage that is caused by high-dose ionizing radiation is similar to the damage that would occur if a type II restriction enzyme was expressed *in vivo*. A large number of opposed single-strand breaks would be formed and would produce double-strand breaks unless stabilized. In yeast, the prolonged expression of *Eco*RI produces thousands of breaks but results in a surprisingly modest (2–3 fold) loss of viability[75–77]. Similarly, prolonged artificial expression of the yeast mating-type-specific HO endonuclease results in only a 35% loss of viability[78] in strains that have four HO-sensitive chromosomal loci. Survival

© 2005 Nature Publishing Group

in these strains does not depend on homologous-recombination systems. Lethal fragmentation of chromosomes only occurs in cells that have defective NHEJ or checkpoint-control systems. Even *E. coli* cells are surprisingly resistant to *Eco*RI-mediated chromosomal cleavage as long as the appropriate DNA ligases are not inactivated[79]. Once the cryptic breaks are repaired, breaks where the ends have separated could be repaired by additional processes such as SSA and recombinational DNA repair.

Conclusion

The extraordinary phenotypes of *D. radiodurans* have encouraged a host of rather fanciful descriptions of the origin of this organism. Stories of the arrival of this species on earth on a comet, or arising through mutations owing to mankind's attempts to harness nuclear power, are readily available on the internet (See Online links box). These ideas are likely to flourish until this remarkable organism is better understood. *D. radiodurans* has a readily documented evolutionary origin within the domain Bacteria. It is clearly related to the rest of life on this planet, carrying out all of the fundamental processes that have been characterized in prokaryotes.

Our interpretation of the literature published to date on the Deinococcaceae suggests no extraordinary survival strategies, or at least no single 'magic bullet' that provides a complete explanation for the phenomenon. Instead, enhanced radioresistance seems to be the consequence of an evolutionary process that has coordinated various passive and active mechanisms, enabling survival from stresses such as desiccation. The almost ubiquitous classical DNA-repair pathways, perhaps with some specialized properties, seem to be augmented by novel features of deinococcal nucleoid structure and metabolism. It is currently difficult to predict which mechanism(s) will be most important in radioresistance, or even whether all of the contributing mechanisms have been discovered. Each of the many enhancements could individually have a modest effect, but collectively they mediate radioresistance. These changes need not be unique to *D. radiodurans*, but might be present in other organisms.

Therefore, the capacity of the deinococci to deal with stresses that cause massive genetic damage might not be as strange or as unusual as it once seemed. The combination of repair strategies is sophisticated and effective, but the individual strategies delineated so far are based on mechanisms present in many other organisms.

1. Anderson, A. W., Nordon, H. C., Cain, R. F., Parrish, G. & Duggan, D. Studies on a radio-resistant micrococcus. I. Isolation, morphology, cultural characteristics, and resistance to γ radiation. *Food Technol.* **10**, 575–578 (1956).
2. Hensel, R., Demharter, W., Kandler, O., Kroppenstedt, R. M. & Stackebrandt, E. Chemotaxonomic and molecular-genetic studies of the genus *Thermus*: evidence for a phylogenetic relationship of *Thermus aquaticus* and *Thermus ruber* to the genus *Deinococcus*. *Int. J. Syst. Bacteriol.* **36**, 444–453 (1986).
3. Murray, R. G. E. in *Bergey's Manual of Systematic Bacteriology* (eds Sneath, P. H. A., Mair, N. S., Sharpe, M. E. & Holt, J. G.) 1035–1043 (Williams & Wilkins, Baltimore, 1986).
4. Rainey, F. A., Nobre, M. F., Schumann, P., Stackebrandt, E. & da Costa, M. S. Phylogenetic diversity of the deinococci as determined by 16S ribosomal DNA sequence comparison. *Int. J. Syst. Bacteriol.* **47**, 510–514 (1997).
5. Weisburg, W. G., Giovannoni, S. J. & Woese, C. R. The Deinococcus–Thermus phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst. Appl. Microbiol.* **11**, 128–134 (1989).
6. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
7. Woese, C. R., Stackebrandt, E., Macke, T. J. & Fox, G. E. A phylogenetic definition of the major eubacterial taxa. *Syst. Appl. Microbiol.* **61**, 143–151 (1985).
8. Brooks, B. W. & Murray, R. G. E. Nomenclature for "*Micrococcus radiodurans*" and other radiation-resistant cocci: *Deinococcaceae* fam. nov. and *Deinococcus* gen. nov., including five species. *Int. J. Syst. Bacteriol.* **31**, 353–360 (1981).
9. United Nations Scientific Committee on the Effects of Atomic Radiation. Effects of atomic radiation. Report A/RES/37/87. (New York, 1982).
10. Mattimore, V. & Battista, J. R. Radioresistance of *Deinococcus radiodurans*: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. *J. Bacteriol.* **178**, 633–637 (1996).
11. Billi, D., Friedmann, E. I., Hofer, K. G., Caiola, M. G. & Ocampo-Friedmann, R. Ionizing-radiation resistance in the desiccation-tolerant cyanobacterium *Chroococcidiopsis*. *Appl. Environ. Microbiol.* **66**, 1489–1492 (2000).
12. DiRuggiero, J., Santangelo, N., Nackerdien, Z., Ravel, J. & Robb, F. T. Repair of extensive ionizing-radiation DNA damage at 95 degrees C in the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.* **179**, 4643–4645 (1997).
13. Jolivet, E., Corre, E., L'Haridon, S., Forterre, P. & Prieur, D. *Thermococcus marinus* sp. nov. and *Thermococcus radiotolerans* sp. nov., two hyperthermophilic archaea from deep-sea hydrothermal vents that resist ionizing radiation. *Extremophiles* **8**, 219–227 (2004).
14. Jolivet, E., L'Haridon, S., Corre, E., Forterre, P. & Prieur, D. *Thermococcus gammatolerans* sp. nov., a hyperthermophilic archaeon from a deep-sea hydrothermal vent that resists ionizing radiation. *Int. J. Syst. Evol. Microbiol.* **53**, 847–851 (2003).
15. Jolivet, E. *et al.* Physiological responses of the hyperthermophilic archaeon "*Pyrococcus abyssi*" to DNA damage caused by ionizing radiation. *J. Bacteriol.* **185**, 3958–3961 (2003).
16. Battista, J. R. & Rainey, F. A. in *Bergey's Manual of Systematic Bacteriology* (eds Boone, D. R. & Castenholz, R. W.) 395–414 (Springer, New York, 2001).
17. Hirsch, P. *et al. Deinococcus frigens* sp. nov., *Deinococcus saxicola* sp. nov., and *Deinococcus marmoris* sp. nov., low temperature and draught-tolerating, UV-resistant bacteria from continental Antarctica. *Syst. Appl. Microbiol.* **27**, 636–645 (2004).
18. Suresh, K., Reddy, G. S., Sengupta, S. & Shivaji, S. *Deinococcus indicus* sp. nov., an arsenic-resistant bacterium from an aquifer in West Bengal, India. *Int. J. Syst. Evol. Microbiol.* **54**, 457–461 (2004).
19. Corrections and Clarifications, *Science* **303**, 766 (2004).
20. White, O. *et al.* Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**, 1571–1517 (1999).
The description of the *D. radiodurans* R1 genome sequence; it has served as a reference for almost all research on the species since 1999.
21. Lin, J. *et al.* Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* **285**, 1558–1562 (1999).
22. Hansen, M. T. Multiplicity of genome equivalents in the radiation-resistant bacterium *Micrococcus radiodurans*. *J. Bacteriol.* **134**, 71–75 (1978).
23. Harsojo, Kitayama, S. & Matsuyama, A. Genome multiplicity and radiation resistance in *Micrococcus radiodurans*. *J. Biochem. (Tokyo)* **90**, 877–880 (1981).
24. Moseley, B. E. & Mattingly, A. Repair of irradiation transforming deoxyribonucleic acid in wild type and a radiation-sensitive mutant of *Micrococcus radiodurans*. *J. Bacteriol.* **105**, 976–983 (1971).
25. Krasin, F. & Hutchinson, F. Repair of DNA double-strand breaks in *Escherichia coli*, which requires *recA* function and the presence of a duplicate genome. *J. Mol. Biol.* **116**, 81–98 (1977).
26. Burrell, A. D., Feldschreiber, P. & Dean, C. J. DNA-membrane association and the repair of double breaks in X-irradiated *Micrococcus radiodurans*. *Biochim. Biophys. Acta* **247**, 38–53 (1971).
27. Daly, M. J. *et al.* Accumulation of Mn(II) in *Deinococcus radiodurans* facilitates γ-radiation resistance. *Science* **306**, 1025–1028 (2004).
The authors show that Mn(II) is necessary for the ionizing-radiation resistance of *D. radiodurans* and provide evidence that elevated intracellular Mn might also protect other species.
28. Gerard, E., Jolivet, E., Prieur, D. & Forterre, P. DNA protection mechanisms are not involved in the radioresistance of the hyperthermophilic archaea *Pyrococcus abyssi* and *P. furiosus*. *Mol. Genet. Genomics* **266**, 72–78 (2001).
29. Daly, M. J., Ling, O. & Minton, K. W. Interplasmidic recombination following irradiation of the radioresistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **176**, 7506–7515 (1994).
30. Daly, M. J. & Minton, K. W. Interchromosomal recombination in the extremely radioresistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **177**, 5495–5505 (1995).
31. Daly, M. J. & Minton, K. W. An alternative pathway of recombination of chromosomal fragments precedes recA-dependent recombination in the radioresistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **178**, 4461–4471 (1996).
32. Daly, M. J. & Minton, K. W. Recombination between a resident plasmid and the chromosome following irradiation of the radioresistant bacterium *Deinococcus radiodurans*. *Gene* **187**, 225–259 (1997).
33. Daly, M. J., Ouyang, L., Fuchs, P. & Minton, K. W. *In vivo* damage and recA-dependent repair of plasmid and chromosomal DNA in the radiation-resistant bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **176**, 3508–3517 (1994).
34. Minton, K. W. DNA repair in the extremely radioresistant bacterium *Deinococcus radiodurans*. *Mol. Microbiol.* **13**, 9–15 (1994).
35. Minton, K. W. Repair of ionizing-radiation damage in the radiation resistant bacterium *Deinococcus radiodurans*. *Mutat. Res.* **363**, 1–7 (1996).

© 2005 Nature Publishing Group

36. Minton, K. W. & Daly, M. J. A model for repair of radiation-induced DNA double-strand breaks in the extreme radiophile *Deinococcus radiodurans*. *Bioessays* **17**, 457–464 (1995).
37. Mortimer, R. K. Radiobiological and genetic studies on a polyploid series (haploid to hexaploid) of *Saccharomyces cerevisiae*. *Radiat. Res.* **66**, 158–169 (1958).
38. Akerlund, T., Nordstrom, K. & Bernander, R. Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of *Escherichia coli*. *J. Bacteriol.* **177**, 6791–6797 (1995).
39. Maldonado, R., Jimenez, J. & Casadesus, J. Changes of ploidy during the *Azotobacter vinelandii* growth cycle. *J. Bacteriol.* **176**, 3911–3919 (1994).
40. Levin-Zaidman, S. *et al.* Ringlike structure of the *Deinococcus radiodurans* genome: a key to radioresistance? *Science* **299**, 254–256 (2003).
The first formal discussion of the unusual nucleoid architecture associated with this species. Some of the conclusions have been controversial, stimulating several subsequent studies.
41. Zimmerman, J. M. & Battista, J. R. A ring-like nucleoid is not necessary for radioresistance in the *Deinococcaceae*. *BMC Microbiol.* **5**, 17 (2005).
42. Archibald, F. S. & Fridovich, I. Manganese, superoxide dismutase, and oxygen tolerance in some lactic acid bacteria. *J. Bacteriol.* **146**, 928–936 (1981).
43. Archibald, F. S. & Fridovich, I. Manganese and defenses against oxygen toxicity in *Lactobacillus plantarum*. *J. Bacteriol.* **145**, 442–451 (1981).
44. Ghosal, D. *et al.* How radiation kills cells: survival of *Deinococcus radiodurans* and *Shewanella oneidensis* under oxidative stress. *FEMS Microbiol. Rev.* **29**, 361–375 (2005).
45. Englander, J. *et al.* DNA toroids: framework for DNA repair in *Deinococcus radiodurans* and in germinating bacterial spores. *J. Bacteriol.* **186**, 5973–5977 (2004).
46. Wilson, R. W. & Bloomfield, V. A. Counterion-induced condesation of deoxyribonucleic acid. a light-scattering study. *Biochemistry* **18**, 2192–2196 (1979).
47. Dean, C. J., Feldschreiber, P. & Lett, J. T. Repair of X-ray damage to the deoxyribonucleic acid in *Micrococcus radiodurans*. *Nature* **209**, 49–52 (1966).
48. Lett, J. T., Feldschreiber, P., Little, J. G., Steele, K. & Dean, C. J. The repair of X-ray damage to the deoxyribonucleic acid in *Micrococcus radiodurans*: a study of the excision process. *Proc. R. Soc. Lond. B Biol. Sci.* **167**, 184–201 (1967).
49. Moseley, B. E. & Copland, H. J. R. Involvement of a recombination repair function in disciplined cell division of *Micrococcus radiodurans*. *J. Gen. Microbiol.* **86**, 343–357 (1975).
50. Kastan, M. B. & Bartek, J. Cell-cycle checkpoints and cancer. *Nature* **432**, 316–323 (2004).
51. Tanaka, M. *et al. Deinococcus radiodurans*' transcriptional response to ionizing radiation and desiccation reveals novel proteins that contribute to extreme radioresistance. *Genetics* **168**, 21–33 (2004).
This manuscript identifies five 'hypothetical proteins' that contribute significantly to *D. radiodurans* radioresistance in RecA-dependent and RecA-independent processes, indicating that this species uses novel mechanisms to facilitate DNA repair.
52. Harris, D. R. *et al.* Preserving genome integrity: the DdrA protein of *Deinococcus radiodurans* R1. *PLoS Biol.* **2**, 1629–1639 (2004).
The DdrA protein protects the 3′ ends of DNA fragments *in vitro* and *in vivo*. This manuscript provides evidence that mechanisms that limit DNA degradation post-irradiation contribute to radioresistance in this species.
53. Iyer, L. M., Koonin, E. V. & Aravind, L. Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. *BMC Genomics* **3**, 8 (2002).
54. Weller, G. R. *et al.* Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* **297**, 1686–1689 (2002).
55. Lecointe, F., Shevelev, I. V., Bailone, A., Sommer, S. & Hubscher, U. Involvement of an X family DNA polymerase in double-stranded break repair in the radioresistant organism *Deinococcus radiodurans*. *Mol. Microbiol.* **53**, 1721–1730 (2004).
56. Narumi, I. *et al.* PprA: a novel protein from *Deinococcus radiodurans* that stimulates DNA ligation. *Mol. Microbiol.* **54**, 278–285 (2004).
PprA is a protein of unknown function that is necessary for radioresistance in *D. radiodurans*. The protein will bind to the ends of double-stranded DNA and stimulate DNA ligation. The authors suggest that PprA helps mediate NHEJ in this species.
57. Haber, J. E. Partners and pathways repairing a double-strand break. *Trends Genet.* **16**, 259–264 (2000).
58. Kim, J. I. *et al.* RecA protein from the extremely radioresistant bacterium *Deinococcus radiodurans*: expression, purification, and characterization. *J. Bacteriol.* **184**, 1649–1660 (2002).
59. Kim, J. I. & Cox, M. M. The RecA proteins of *Deinococcus radiodurans* and *Escherichia coli* promote DNA strand exchange via inverse pathways. *Proc. Natl Acad. Sci. USA* **99**, 7917–7921 (2002).
60. Satoh, K. *et al.* Characterization of RecA424 and RecA670 proteins from *Deinococcus radiodurans*. *J. Biochem. (Tokyo)* **131**, 121–129 (2002).
61. Lipton, M. S. *et al.* Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl Acad. Sci. USA* **99**, 11049–11054 (2002).
62. Liu, Y. *et al.* Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. *Proc. Natl Acad. Sci. USA* **100**, 4191–4196 (2003).
63. Makarova, K. S. *et al.* Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol. Mol. Biol. Rev.* **65**, 44–79 (2001).
64. Eggington, J. M., Haruta, N., Wood, E. A. & Cox, M. M. The single-stranded DNA-binding protein of *Deinococcus radiodurans*. *BMC Microbiol.* **4**, 2 (2004).
65. Bernstein, D. A. *et al.* Crystal structure of the *Deinococcus radiodurans* single-stranded DNA-binding protein suggests a mechanism for coping with DNA damage. *Proc. Natl Acad. Sci. USA* **101**, 8575–8580 (2004).
66. Wang, J. & Julin, D. A. DNA helicase activity of the RecD protein from *Deinococcus radiodurans*. *J. Biol. Chem.* **279**, 52024–52032 (2004).
67. Makharashvili, N., Koroleva, O., Bera, S., Grandgenett, D. P. & Korolev, S. A novel structure of DNA repair protein RecO from *Deinococcus radiodurans*. *Structure (Camb.)* **12**, 1881–1889 (2004).
68. Leiros, I., Timmins, J., Hall, D. R. & McSweeney, S. Crystal structure and DNA-binding analysis of RecO from *Deinococcus radiodurans*. *EMBO J.* **24**, 906–918 (2005).
69. Makarova, K. S., Wolf, Y. I., White, O., Minton, K. & Daly, M. J. Short repeats and IS elements in the extremely radiation-resistant bacterium *Deinococcus radiodurans* and comparison to other bacterial species. *Res. Microbiol.* **150**, 711–724 (1999).
70. Cobbe, N. & Heck, M. M. Review: SMCs in the world of chromosome biology — from prokaryotes to higher eukaryotes. *J. Struct. Biol.* **129**, 123–143 (2000).
71. Jessberger, R. SMC proteins at the crossroads of diverse chromosomal processes. *IUBMB Life* **55**, 643–652 (2003).
72. Volkov, A., Mascarenhas, J., Andrei-Selmer, C., Ulrich, H. D. & Graumann, P. L. A prokaryotic condensin/cohesin-like complex can actively compact chromosomes from a single position on the nucleoid and binds to DNA as a ring-like structure. *Mol. Cell. Biol.* **23**, 5638–5650 (2003).
73. Lisby, M. & Rothstein, R. DNA repair: keeping it together. *Curr. Biol.* **14**, R994–R996 (2004).
74. Freifelder, D. & Trumbo, B. Matching of single-strand breaks to form double-strand breaks in DNA. *Biopolymers* **7**, 681–693 (1969).
75. Lewis, L. K., Kirchner, J. M. & Resnick, M. A. Requirement for end-joining and checkpoint functions, but not RAD52-mediated recombination, after *Eco*RI endonuclease cleavage of *Saccharomyces cerevisiae* DNA. *Mol. Cell. Biol.* **18**, 1891–1902 (1998).
76. Lewis, L. K. & Resnick, M. A. Tying up loose ends: nonhomologous end-joining in *Saccharomyces cerevisiae*. *Mutat. Res.* **451**, 71–89 (2000).
77. Lewis, L. K., Westmoreland, J. W. & Resnick, M. A. Repair of endonuclease-induced double-strand breaks in *Saccharomyces cerevisiae*: essential role for genes associated with nonhomologous end-joining. *Genetics* **152**, 1513–1529 (1999).
78. Haber, J. E. & Leung, W. Y. Lack of chromosome territoriality in yeast: promiscuous rejoining of broken chromosome ends. *Proc. Natl Acad. Sci. USA* **93**, 13949–13954 (1996).
79. Heitman, J., Zinder, N. D. & Model, P. Repair of the *Escherichia coli* chromosome after *in vivo* scission by the *Eco*RI endonuclease. *Proc. Natl Acad. Sci. USA* **86**, 2281–2285 (1989).
80. Green, P. N. & Bousfield, I. J. Emendation of *Methylobacterium* (Patt, Cole, and Hanson 1976); *Methylobacterium rhodinum* (Heumann 1962) comb. nov. corrig. ; *Methylobacterium radiotolerans* (Ito and Iizuka 1971) comb. nov. corrig.; and *Methylobacterium mesophilicum* (Austin and Goodfellow 1979) comb. nov. *Int. J. Syst. Bacteriol.* **33**, 875–877 (1983).
81. Ito, H. & Iizuka, H. Taxonomic studies on a radio-resistant *Pseudomonas*. XII. Studies on the microorganisms of cereal grain. *Agric. Biol. Chem.* **35**, 1566–1571 (1971).
82. Nishimura, Y., Uchida, K., Tanaka, K., Ino, T. & Ito, H. Radiation sensitivities of *Acinetobacter* strains isolated from clinical sources. *J. Basic Microbiol.* **34**, 357–360 (1994).
83. Phillips, R. W. *et al. Kineococcus radiotolerans* sp. nov., a radiation-resistant, Gram-positive bacterium. *Int. J. Syst. Evol. Microbiol.* **52**, 933–938 (2002).
84. Collins, M. D., Hutson, R. A., Grant, I. R. & Patterson, M. F. Phylogenetic characterization of a novel radiation-resistant bacterium from irradiated pork: description of *Hymenobacter actinosclerus* sp. nov. *Int. J. Syst. Evol. Microbiol.* **50**, 731–734 (2000).
85. Ferreira, A. C. *et al.* Characterization and radiation resistance of new isolates of *Rubrobacter radiotolerans* and *Rubrobacter xylanophilus*. *Extremophiles* **3**, 235–238 (1999).
86. Ito, H., Watanabe, H., Takeshia, M. & Iizuka, H. Isolation and identification of radiation-resistant cocci belonging to the genus *Deinococcus* from sewage sludges and animal feeds. *Agric. Biol. Chem.* **47**, 1239–1247 (1983).

Acknowledgements

Work on genome repair in *Deinococcus radiodurans* in the Battista and Cox laboratories is supported a grant from the National Institutes of Health. The authors thank F.A. Rainey for producing the phylogenetic tree in Figure 2. We also thank J. Haber for helpful discussions.

Competing interests statement

The authors declare no competing financial interests

Online links

DATABASES

The following terms in this article are linked online to:

Entrez: http://www.ncbi.nlm.nih.gov/Entrez
Azotobacter vinelandii | *Bacillus subtilis* | *Deinococcus geothermalis* | *Deinococcus radiodurans strain* R1 | *E. coli* K12 | *Kineococcus radiotolerans* | *Lactobacillus plantarum* | *Rubrobacter xylanophilus*

SwissProt: http://www.expasy.ch
DNA polymerase I | *D. radiodurans* RecA | *D. radiodurans* SSB | *E. coli* RecA | *E. coli* SSB | PprA | RecR

FURTHER INFORMATION

John Battista's laboratory: http://www.biology.lsu.edu/faculty_listings/fac_pages/jbattista.html
Michael Cox's laboratory: http://www.biochem.wisc.edu/cox/
Speculation on the origins of *D. radiodurans*: http://www.newscientist.com/article.ns?id=dn2844
TIGR Comprehensive Microbial Resource database: http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
Access to this interactive links box is free online.

© 2005 Nature Publishing Group

View publication stats

PRL 109, 160401 (2012)

Selected for a Viewpoint in Physics

PHYSICAL REVIEW LETTERS

week ending
19 OCTOBER 2012

Quantum Time Crystals

Frank Wilczek

Center for Theoretical Physics Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(Received 29 March 2012; published 15 October 2012)

Some subtleties and apparent difficulties associated with the notion of spontaneous breaking of time-translation symmetry in quantum mechanics are identified and resolved. A model exhibiting that phenomenon is displayed. The possibility and significance of breaking of imaginary time-translation symmetry is discussed.

DOI: 10.1103/PhysRevLett.109.160401 PACS numbers: 11.30.−j, 03.75.Lm, 05.45.Xt

Symmetry and its spontaneous breaking is a central theme in modern physics. Perhaps no symmetry is more fundamental than time-translation symmetry, since time-translation symmetry underlies both the reproducibility of experience and, within the standard dynamical frameworks, the conservation of energy. So it is natural to consider the question, whether time-translation symmetry might be spontaneously broken in a closed quantum-mechanical system. That is the question we will consider, and answer affirmatively, here.

Here we are considering the possibility of time crystals, analogous to ordinary crystals in space. They represent spontaneous emergence of a clock within a time-invariant dynamical system. Classical time crystals are considered in a companion Letter [1]; here the primary emphasis is on quantum theory.

Several considerations might seem to make the possibility of quantum time crystals implausible. The Heisenberg equation of motion for an operator with no intrinsic time dependence reads

$$\langle\Psi|\dot{\mathcal{O}}|\Psi\rangle = i\langle\Psi|[H,\mathcal{O}]|\Psi\rangle \underset{\Psi=\Psi_E}{\longrightarrow} 0, \tag{1}$$

where the last step applies to any eigenstate Ψ_E of H. This seems to preclude the possibility of an order parameter that could indicate the spontaneous breaking of infinitesimal time-translation symmetry. Also, the very concept of "ground state" implies the state of lowest energy, but in any state of definite energy (it seems) the Hamiltonian must act trivially. Finally, a system with spontaneous breaking of time-translation symmetry in its ground state must have some sort of motion in its ground state, and is therefore perilously close to fitting the definition of a perpetual motion machine.

Ring particle model.—And yet there is a familiar physical phenomenon that almost does the job. A superconductor, in the right circumstances, can support a stable current-carrying ground state. Specifically, this occurs if we have a superconducting ring threaded by a flux that is a fraction of the flux quantum. If the current is constant then nothing changes in time, so time-translation symmetry is not broken, but clearly there is a sense in which something is moving.

We can display the essence of this situation in a simple model, that displays its formal structure clearly. Consider a particle with charge q and unit mass, confined to a ring of unit radius that is threaded by flux $2\pi\alpha/q$. The Lagrangian, canonical (angular) momentum, and Hamiltonian for this system are, respectively,

$$L = \tfrac{1}{2}\dot{\phi}^2 + \alpha\dot{\phi}, \qquad \pi_\phi = \dot{\phi} + \alpha, \qquad H = \tfrac{1}{2}(\pi_\phi - \alpha)^2. \tag{2}$$

π_ϕ, through its role as generator of (angular) translations, and in view of the Heisenberg commutation relations, is realized as $-i\frac{\partial}{\partial\phi}$. Its eigenvalues are integers l, associated with the states $|l\rangle = e^{il\phi}$. For these states we have

$$\langle l|\dot{\phi}|l\rangle = l - \alpha, \qquad \langle l|H|l\rangle = \tfrac{1}{2}(l-\alpha)^2. \tag{3}$$

The lowest energy state will occur for the integer l_0 that makes $l - \alpha$ smallest. If α is not an integer, we will have

$$\langle l_0|\dot{\phi}|l_0\rangle = l_0 - \alpha \neq 0. \tag{4}$$

The case when α is half an odd integer requires special consideration. In that case we will have two distinct states $|\alpha \pm \frac{1}{2}\rangle$ with the minimum energy. We can clarify the meaning of that degeneracy by combining two simple observations. First, that the combined operation G_k of multiplying wave functions by $e^{ik\phi}$ and changing $\alpha \to \alpha + k$, for integer k, in the Lagrangian leaves the dynamics invariant. Indeed, if we interpret α in L as embodying a constant gauge potential, G_k is a topologically nontrivial gauge transformation on the ring, corresponding to the multiply valued gauge function $A \to A + \nabla\Lambda$, $\Lambda = k\phi/q$. Note that the total flux is *not* invariant under this topologically nontrivial gauge transformation, which cannot be extended smoothly off the ring, so L is modified. Second, that the operation of time-reversal T, implemented by complex conjugation of wave functions, takes $|l\rangle \to |-l\rangle$ and leaves the dynamics invariant if simultaneously $\alpha \to -\alpha$. Putting these observations together, we see that the combined operation

© 2012 American Physical Society

$$\tilde{T} = G_{2\alpha}T \tag{5}$$

leaves the Lagrangian invariant; it is a symmetry of the dynamics and maps $|l\rangle \rightarrow |2\alpha - l\rangle$. $\tilde{T}$ interchanges $|l \pm \alpha\rangle \rightarrow |l \mp \alpha\rangle$. Thus the degeneracy between those states is a consequence of a modified time-reversal symmetry. We can choose combinations $|\alpha + \frac{1}{2}\rangle \pm |\alpha - \frac{1}{2}\rangle$ that simultaneously diagonalize H and $\tilde{T}$; for these combinations the expectation value of $\dot{\phi}$ vanishes.

Returning to the generic case: For α that are not half-integral time-reversal symmetry is not merely modified, but simply broken, and there is no degeneracy. How do we reconcile $\langle l_0|\dot{\phi}|l_0\rangle \neq 0$ with Eq. (1)? The point is that $\dot{\phi}$, despite appearances, is neither the time-derivative of a legitimate operator nor the commutator of the Hamiltonian with one, since ϕ, acting on wave functions in Hilbert space, is multivalued. By way of contrast, operators corresponding to single-valued functions of ϕ, spanned by trigonometric functions $\mathcal{O}_k = e^{ik\phi}$, do satisfy Eq. (1) for the eigenstates $|\Psi\rangle = |l\rangle$.

Wave functions of the quantized ring particle model correspond to the (classical) wave functions that appear in the Landau-Ginzburg theory of superconductivity. Those wave functions, in turn, heuristically describe the wave function for macroscopic occupation of the single-particle quantum state appropriate to a Cooper pair, regarded as a particle. Under this correspondence, the nonvanishing expectation value of $\dot{\phi}$ for the ground state of the ring particle subject to fractional flux maps onto the persistent current in a superconducting ring.

Symmetry breaking and observability.—The choice of a ground state that violates time-translation symmetry τ must be based on some criterion other than energy minimization. But what might seem to be a special difficulty with breaking τ, because of its connection to the Hamiltonian, actually arises in only a slightly different form for all cases of spontaneous symmetry breaking. Consider, for example, the breaking of number (or dually, phase) symmetry. We characterize such breaking through a complex order parameter, Φ, that acquires a nonzero expectation value, which we can take to be real:

$$\langle 0|\Phi|0\rangle = v \neq 0. \tag{6}$$

We also have states $|\sigma\rangle$ related to $|0\rangle$ by the symmetry operation. These are all energetically degenerate and mutually orthogonal in the appropriate "infinite volume" limit (see immediately below), and satisfy

$$\langle \sigma|\Phi|\sigma\rangle = ve^{i\sigma}. \tag{7}$$

The superposition

$$|\Omega\rangle \propto \int_0^{2\pi} d\sigma|\sigma\rangle \tag{8}$$

is energetically degenerate with all the $|\sigma\rangle$, and it is symmetric, with

$$\langle \Omega|\Phi|\Omega\rangle = 0. \tag{9}$$

[Normalization of $|\Omega\rangle$ depends on how the limit is taken. If we arrange $\langle \sigma|\sigma'\rangle \rightarrow \delta(\sigma - \sigma')$, then the proportionality constant is $(2\pi)^{-1}$.]

Why then do we prefer one of the states $|\sigma\rangle$ as a description of the physical situation? The reason is closely related to the emergent orthogonality of the different $|\sigma\rangle$ states, as we now recall. We envisage that our system extends over a large number N of identical subsystems having correlated values of the long-range order parameter σ, but otherwise essentially uncorrelated. Then we can express the total wave function in the form

$$\Psi_\sigma(x_1, \ldots, x_N) \approx \prod_{j=1}^{N} \psi_\sigma(x_j). \tag{10}$$

For different values σ, σ' we have therefore

$$\langle \Psi_{\sigma'}|\Psi_\sigma\rangle \approx \prod_{j=1}^{N} \langle \psi_{\sigma'}(x_j)|\psi_\sigma(x_j)\rangle = (f_{\sigma'\sigma})^N \rightarrow 0 \tag{11}$$

for $\sigma' \neq \sigma$ and large N, since $|f_{\sigma'\sigma}| < 1$. Similarly, for any finite set of local observables (that is, observables whose arguments include only upon a finite subset of the x_j), we have

$$\langle \Psi_{\sigma'}|\mathcal{O}_1(x_a)\mathcal{O}_2(x_b)\ldots|\Psi_\sigma\rangle \propto (f_{\sigma'\sigma})^{N-\text{finite}} \rightarrow 0 \tag{12}$$

for $\sigma' \neq \sigma$. Since the off-diagonal matrix elements vanish, any world of local observations (including "observations" by the environment) can be described using a single $|\sigma\rangle$ state. Averaging over them, to produce $|\Omega\rangle$, is a purely formal operation. Measurement of a nonsinglet observable will project onto a $|\sigma\rangle$ state.

This analysis, which elaborates [2], brings out several relevant points. The physical criterion that identifies useful "ground states" is not simply energy, but also robust observability—that is, relevance to the description of observations in a world of mutually communicating observers. Mathematically, that requirement is reflected in the orthogonality of the Hilbert spaces built upon $|\sigma\rangle$ states by the action of physical observables. The large N limit is crucial for spontaneous symmetry breaking. It is only in that infinite degree of freedom, or (as it is usually called) infinite volume, limit, that the $|\sigma\rangle$ states and their Fourier transforms $|j\rangle \propto \int d\sigma e^{ij\sigma}|\sigma\rangle$, with definite charge j, become degenerate, and the former are preferred. Important for present purposes: The preceding discussion applies, with only symbolic changes, when we consider possible breaking of time-translation τ in place of phase symmetry.

Soliton model.—After these preparations, it is not difficult to construct an appropriate model. We consider a large number of ring-particles with an attractive interaction. Heuristically, we can expect that they will want to form a lump and, in view of Eq. (4), that they will want to move. So we can expect that the physical ground state features a moving lump, which manifestly breaks τ.

To make contact with the argument of the previous section, we need an appropriate notion of locality. We assume that the particles have an additional integer label, besides the common angle ϕ, and that the local physical observables are of finite range in that additional label. One can imagine an array of separate rings, displaced along an axis, so that the coordinates of particle j are $(\phi, x = ja)$. Note that with this interpretation, the basic interaction is infinitely long ranged, and would have to be specially engineered. I will revisit this issue below, after describing the construction.

An appropriate Hamiltonian is

$$\begin{aligned} H &= \sum_{j=1}^{N} \frac{1}{2}(\pi_j - \alpha)^2 - \frac{\lambda}{N-1} \sum_{j \neq k, 1}^{N} \delta(\phi_j - \phi_k) \\ &\equiv \sum_{j=1}^{N} \frac{1}{2}(\pi_j - \alpha)^2 + V(\phi_1, \ldots, \phi_N), \end{aligned} \tag{13}$$

with the understanding that H acts on periodic functions, so the δ interaction is well defined. (Here the discrete index appears as a subscript.)

We work in the mean field approximation, taking a product ansatz

$$\Psi(\phi_1, \ldots, \phi_n) = \prod_{j=1}^{N} \psi(\phi_j), \tag{14}$$

and solving an approximate one-body equation for ψ. To get such an equation, we define an effective potential

$$\begin{aligned} V_{\text{eff.}}(\phi_1, \ldots, \phi_N) &= \sum_{j=1}^{N}, \\ W(\phi_j)W(\phi_j) &= \int \prod_{k \neq j} d\phi_k \psi^*(\phi_k) V \psi(\phi_k), \end{aligned} \tag{15}$$

so that

$$\langle \Psi | V_{\text{eff.}} | \Psi \rangle = \langle \Psi | V | \Psi \rangle. \tag{16}$$

Then the effective Schrödinger equation for Ψ,

$$i \frac{\partial \Psi}{\partial t} = \left(\sum_{j=1}^{N} \frac{1}{2}(\pi_j - \alpha)^2 + V_{\text{eff.}} \right) \Psi, \tag{17}$$

reduces to the one-body nonlinear Schrödinger equation

$$i \frac{\partial \psi}{\partial t} = \frac{1}{2}(\pi_\phi - \alpha)^2 \psi - \lambda |\psi|^2 \psi \tag{18}$$

for ψ.

Consider first the case $\alpha = 0$. Eq. (18) can be solved for a stationary state in terms of the Jacobi dn elliptic function, with

$$\begin{aligned} \psi(\phi, t) = e^{-i\mathcal{E}t} \psi_0(\phi + \beta), \qquad \psi_0(\phi) = r\,\text{dn}(r\sqrt{\lambda}\phi, k^2), \\ \mathcal{E} = -r^2 \lambda \left(1 - \frac{k^2}{2} \right), \end{aligned} \tag{19}$$

with β a disposable parameter. To fix the parameters k, r we must impose 2π periodicity in ϕ and normalize ψ_0. Those conditions become

$$E(k^2) = \frac{\sqrt{\lambda}}{2r}, \qquad K(k^2) = \pi r \sqrt{\lambda} \tag{20}$$

in terms of the complete elliptic integrals $E(k^2)$, $K(k^2)$. We can solve $E(k^2)K(k^2) = \frac{\pi\lambda}{2}$ for k^2, given λ. The minimum value of the left-hand side occurs at $k = 0$ and corresponds to $\lambda = \frac{\pi}{2}$. Here $\text{dn}(u, 0)$ reduces to a constant, and $\mathcal{E} = -1/4$. As λ increases beyond that value k rapidly approaches 1, as does $E(k^2)$. $\text{dn}(u, k^2) \rightarrow \text{sech} u$ and $\mathcal{E} \rightarrow -\lambda^2/8$ in that limit. Of course the constant solution with $\mathcal{E} = -\lambda/2\pi$ exists for any value of λ, but when λ exceeds the critical value the inhomogeneous solution is more favorable energetically. These results have simple qualitative interpretations. The hyperbolic secant is the famous soliton of the nonlinear Schrödinger equation on a line. If that soliton is not too big it can be deformed, without prohibitive energy cost, to fit on a unit circle. The parameter β reflects spontaneous breaking of (ordinary) translation symmetry. Here that breaking is occurring through a kind of phase separation.

Our Hamiltonian is closely related, formally, to the Lieb-Liniger model [3], but because we consider ultraweak ($\sim 1/N$) attraction instead of repulsion, the ground state physics is very different. In general low-dimensional models of spontaneous symmetry breaking are subject to derangement by fluctuations [4]. Since our extremely inhomogeneous approximate ground state does not support low-energy, long-wavelength modes (apart from overall translation, but note that the mass of the lump is growing with N), it has no serious infrared sensitivity. It would be interesting to the model with attractive couplings more deeply, and at finite coupling. In any case, it is not difficult to realize the same ideas in higher-dimensional models, such as the Wigner crystal briefly mentioned below (and now analyzed in depth as a proposed experiment [5]). In finite systems the correlation time will be finite, of course, but in interesting cases it becomes very long. Its growth with system size might, by analogy with more familiar cases [6], be algebraic rather than exponential for some low-dimensional systems.

Now since nonzero α can be interpreted as magnetic flux through the ring, we might anticipate, from Faraday's law, that as we turn it on, starting from $\alpha = 0$, our lump of charge will feel a simple torque. (Note that since Faraday's law is a formal consequence of the mathematics of gauge potentials, its use does not require additional hypotheses.) We can also apply "gauge transformations", as in the discussion around Eq. (5). These observations are reflected mathematically in the following construction: For any l, we solve

$$i \frac{\partial \psi_l}{\partial t} = \frac{1}{2}(-i\partial_\phi - \alpha)^2 \psi_l - \lambda |\psi_l|^2 \psi_l, \tag{21}$$

with

$$\psi_l(\phi, t) = e^{-il\phi}\tilde{\psi}(\phi + (l+\alpha)t, t) i\frac{\partial\tilde{\psi}}{\partial t}$$
$$= \frac{1}{2}(-i\partial_\phi)^2\tilde{\psi} - \lambda|\tilde{\psi}|^2\tilde{\psi} + \frac{(l+\alpha)^2}{2}\tilde{\psi}. \quad (22)$$

As in the noninteracting ring particle model, the lowest energy is obtained by minimizing $l_0 + \alpha$, for integral l_0. This will supply appropriate $\tilde{\psi}$. If α is not an integer $\psi_{l_0}(\phi, t)$ will be a moving lump, and time-translation symmetry will have been spontaneously broken. If α is half an odd integer, then its $\tilde{T}$ symmetry is spontaneously broken too.

This example exhibits several characteristic features of natural τ breaking [1]. The lump moves along a constant energy trajectory. The parameter β, which parameterizes an orbit of (ordinary) translation symmetry, changes at a constant rate; both τ and translation symmetry are broken, but a combination remains intact.

Now let us return to address the conceptual issues alluded to earlier, around locality. Our model Hamiltonian was nonlocal, but we required observables to be local. That schizophrenic distinction can be appropriate, since the Hamiltonian might be—and, for our rather artificial dynamics, would have to be—carefully engineered, as opposed to being constructed from easily implemented, natural observables and interactions, which are local. However it is not unlikely that the assumption of all-to-all coupling, adopted for mathematical convenience, could be relaxed, in particular, by locating the rings at the nodes of a multidimensional lattice and limiting the couplings to a finite range.

Were we literally considering charged particles confined to a common ring, and treating the electromagnetic field dynamically, our moving lump of charge would radiate. The electromagnetic field provides modes that couple to all the particles, and in effect provide observers who manifestly violate the framework of Eq. (12). That permits, and enforces, relaxation to a $|k\rangle$ state. Simple variations can ameliorate this issue, e.g., use of multipoles in place of single charges, embedding the system in a cavity, or simply arranging that the motion is slow. A more radical variation, that also addresses the unrealistic assumption of attraction among the charges, while still obtaining spatial nonuniformity, would be to consider charged particles on a ring that form—through repulsion—a Wigner lattice.

Imaginary-time crystals.—In the standard treatment of finite temperature quantum systems using path integral techniques, one considers configurations whose arguments involve imaginary values of the time, and imposes imaginary-time periodicity in the inverse temperature $\beta = 1/T$. In this setup the whole action is converted, in effect, into a potential energy: time derivatives map onto gradients in imaginary time, which is treated on the same footing as the spatial variables.

At the level of the action, there is symmetry under translations in imaginary time (iTime). But since iTime appears, in this formulation, on the same footing as the spatial variables, it is natural to consider the possibility that for appropriate systems the dominant configurations in the path integral are iTime crystals. Let the iTime crystal have preferred period λ. When β is an integer multiple of λ the crystal will fit without distortion, but otherwise it must be squeezed or stretched, or incorporate defects. Periodic behavior of thermodynamics quantities in $1/T$, with period λ, arise, and provide an experimental diagnostic. Integration over the collective coordinate for the broken symmetry contributes to the entropy, even at zero temperature. Inspired by the spatial crystal—iTime crystal analogy, one might also consider the possibility of iTime glasses (iGlasses), which would likewise have residual entropy, but no simple order, or iQuasicrystals.

Comments.—(i). It is interesting to speculate that a (considerably) more elaborate quantum-mechanical system, whose states could be interpreted as collections of qubits, might be engineered to traverse, in its ground configuration, a programmed landscape of structured states in Hilbert space over time.

(ii). Fields or particles in the presence of a time crystal background will be subject to energy-changing processes, analogous to crystalline Umklapp processes. In either case the apparent nonconservation is in reality a transfer to the background. (In our earlier model, $O(1/N)$ corrections to the background motion arise.)

(iii). Many questions that arise in connection with any spontaneous ordering, including the nature of transitions into or out of the order at finite temperature, critical dimensionality, defects and solitons, and low-energy phenomenology, likewise pose themselves for time crystallization. There are also interesting issues around the classification of space-time periodic orderings (roughly speaking, four dimensional crystals [7]).

(iv). The ac Josephson effect is a semimacroscopic oscillatory phenomenon related in spirit to time crystallization. It requires, however, a voltage difference that must be sustained externally, so it is not a ground state effect.

(v). Quantum time crystals based on the classical time crystals of [1], which use singular Hamiltonians, can be constructed by combining the ideas of this Letter with those of [8,9]. The appearance of swallowtail band structures in [10], and emergence of complicated frequency dependence in modeling finite response times [1], as in [11], suggest possible areas of application.

I thank B. Halperin, Hong Liu, J. Maldacena, and especially Al Shapere for helpful comments. This work is supported in part by DOE grant DE-FG02-05ER41360.

[1] A. Shapere and F. Wilczek, following Letter, Phys. Rev. Lett., **109**, 160402 (2012).

[2] Compare F. Strocchi, *Symmetry Breaking* (Springer, New York, 2008), 2nd ed.; P. Anderson, *Basic Notions of*

Condensed Matter Physics (Addison-Wesley, Reading, MA, 1997).
[3] E. Lieb and W. Liniger, Phys. Rev. **130**, 1605 (1963).
[4] H. Mermin and H. Wagner, Phys. Rev. Lett. **17**, 1133 (1966).
[5] L. Zhao, P. Yu, and W. Xu, arXiv:1206.2983.
[6] P. Chaikin and T. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, England, 1995).
[7] H. Brown, R. Bülow, J. Neubüser, H. Wondratschek, and H. Zassenhaus, *Crystallographic Groups of Four-Dimensional Space* (Wiley, New York, 1978).
[8] M. Henneaux, C. Teitelboim, and J. Zanelli, Phys. Rev. A **36**, 4417 (1987).
[9] A. Shapere and F. Wilczek, arXiv:1207.2677.
[10] B. T. Seaman, L. D. Carr, and M. J. Holland, Phys. Rev. A **72**, 033602 (2005).
[11] G. Georges, G. Kotliar, W. Krauth, and M. Rozenberg, Rev. Mod. Phys. **68**, 13 (1996).

Helium-3, ^{3}He.

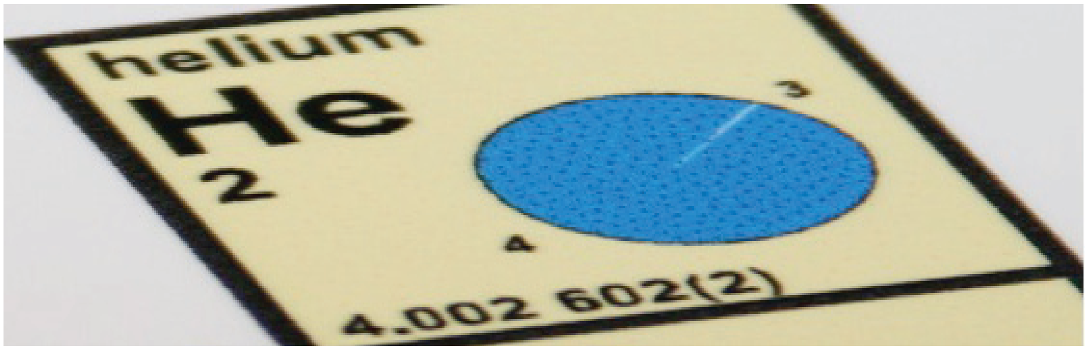

Chemical Purity ≥99.9995%
Isotopic Enrichment ≥99.9 atom%

^{3}He is the very rare stable isotope of ordinary helium. All ^{3}He produced today comes from the slow degradation of the radioisotope tritium, ^{3}He is principally shipped and used in gaseous form in neutron detectors, spectrometers, medical imaging, ring lasers, dilution refrigerators, and ultra low temperature physics research. Helium-3 is an odorless, colorless, and tasteless high pressure gas.

Physical and chemical properties

Chemical purity	≥99.9995% Total impurities as specified below ≤5ppm
Isotopic enrichment ^{3}He atomic fraction (%)	≥99.9 Atom%
Molecular weight	3.016g/mol
Gas density @ 21.1°C (70°F) ("NTP"):	0.0078 Lb/ft$_3$ (0.125 Kg/m3) (0.125 G/ltr)
@ 0°C ("STP"):	0.0084 Lb/ft$_3$ (0.135 Kg/m3) (0.135 G/ltr)
Boiling point @ 1 atm	-269.96°C
Freezing/Melting point @ 1 atm:	None
Solubility in water vol/vol at 20°c (68°F) and 1 atm:	8.61 M$_3$/1 kg water
Gas Density	0.125g/L
Critical pressure:	-2.7 Psia (0.186 Bar)
Coefficient water/Oil distribution:	Not applicable
Odor threshold:	Helium-3 is odorless
Appearance, odor and state:	Colorless, odorless gas

Chemical Purity

COMPONENT	SPECIFICATIONS (MOLE/MOLE)
Oxygen (ppm)	≤ 1
Nitrogen (ppm)	≤ 1
Carbon Dioxide (ppm)	≤0.5
Carbon Monoxide (ppm)	≤1
Argon (ppm)	≤0.5
Water (ppm)	≤0.1
Activity (uCi/cc)	≤1.3 x 10-7

Cylinder Specifications

Size	Water Volume	Maximum Gas Contained
Lecture Bottle (Linde 2)	0.42L	50L
6 (Linde 6)	1.45L	200L
5 (Linde 10)	3L	400L
4 (Linde 30)	8L	1100L

U.S. Shipping information

U.S. Dot proper shipping name	Helium, compressed
Hazard class number and description:	2.2 (Non-flammable gas)
UN identification number:	UN 1046
U.S. DOT shipping label(s) required:	Non-flammable gas
CAS number	7440-59-7

Linde Electronics and Specialty Gases
One Greenwich Street, Suite 100, Stewartsville, NJ 08886
Phone +1.908.329-9700 or +1.800.932-0624, Fax +1.908.329-9740, www.lindeus.com

Linde North America Inc. is a member of The Linde Group. Linde is a trading name used by companies within the Linde Group. The Linde logo is a trademark of The Linde Group. © The Linde Group 2012.

WIKIPEDIA

Helium-3

Helium-3 (^{3}He, see also helion) is a light, non-radioactive isotope of helium with two protons and one neutron (common helium having two protons and two neutrons). Its hypothetical existence was first proposed in 1934 by the Australian nuclear physicist Mark Oliphant while he was working at the University of Cambridge Cavendish Laboratory. Oliphant had performed experiments in which fast deuterons collided with deuteron targets (incidentally, the first demonstration of nuclear fusion).[1] Helium-3 was thought to be a radioactive isotope until it was also found in samples of natural helium, which is mostly helium-4, taken both from the terrestrial atmosphere and from natural gas wells.[2] Other than protium (ordinary hydrogen), helium-3 is the only stable isotope of any element with more protons than neutrons.

Helium-3 occurs as a primordial nuclide, escaping from the Earth's crust into the atmosphere and into outer space over millions of years. Helium-3 is also thought to be a natural nucleogenic and cosmogenic nuclide, one produced when lithium is bombarded by natural neutrons, which can be released by spontaneous fission and by nuclear reactions with cosmic rays. Some of the helium-3 found in the terrestrial atmosphere is also a relic of atmospheric and underwater nuclear weapons testing.

Much speculation has been made over the possibility of helium-3 as a future energy source. Unlike most other nuclear fusion reactions, the fusion of helium-3 atoms releases large amounts of energy without causing the surrounding material to become radioactive. However, the temperatures required to achieve helium-3 fusion reactions are much higher than in traditional fusion reactions,[3] and the process may unavoidably create other reactions that themselves would cause the surrounding material to become radioactive.[4]

The abundance of helium-3 is thought to be greater on the Moon than on Earth, having been embedded in the upper layer of regolith by the solar wind over billions of years,[5] though still lower in abundance than in the solar system's gas giants.[6][7]

Helium-3, ^{3}He

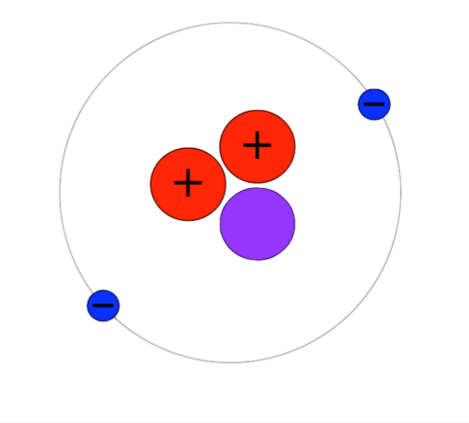

General	
Name, symbol	Helium-3, He-3, ^{3}He
Neutrons	1
Protons	2
Nuclide data	
Natural abundance	0.000137% (% He on Earth)
Half-life	stable
Parent isotopes	^{3}H (beta decay of tritium)
Isotope mass	3.0160293 u
Spin	½
Complete table of nuclides	

Contents

Physical properties
Fusion reactions
Neutron detection
Cryogenics
Medical lung imaging

Industrial production
Natural abundance
  Solar nebula (primordial) abundance
  Terrestrial abundance
Extraction from extraterrestrial sources
Power generation
References in science fiction
See also
Notes and references
  Bibliography
External links

Physical properties

Because of its low atomic mass of 3.02 atomic mass units, helium-3 has some physical properties different from those of helium-4, with a mass of 4.00 atomic mass units. Because of the weak, induced dipole–dipole interaction between the helium atoms, their microscopic physical properties are mainly determined by their zero-point energy. Also, the microscopic properties of helium-3 cause it to have a higher zero-point energy than helium-4. This implies that helium-3 can overcome dipole–dipole interactions with less thermal energy than helium-4 can.

The quantum mechanical effects on helium-3 and helium-4 are significantly different because with two protons, two neutrons, and two electrons, helium-4 has an overall spin of zero, making it a boson, but with one fewer neutron, helium-3 has an overall spin of one half, making it a fermion.

Helium-3 boils at 3.19 K compared with helium-4 at 4.23 K, and its critical point is also lower at 3.35 K, compared with helium-4 at 5.2 K. Helium-3 has less than half the density of Helium-4 when it is at its boiling point: 59 gram per liter compared to the 125 gram per liter of helium-4—at a pressure of one atmosphere. Its latent heat of vaporization is also considerably lower at 0.026 kilojoules per mole compared with the 0.0829 kilojoules per mole of helium-4.[8][9]

Fusion reactions

${}^3\text{He}$ can be produced by the low temperature fusion of ${}^2\text{H} + {}^1\text{p} \xrightarrow{(\text{D-p})} {}^3\text{He} + \gamma + 4.98$ MeV. If the fusion temperature is below that for the helium nuclei to fuse, the reaction produces a high energy alpha particle which quickly acquires an electron producing a stable light helium ion which can be utilized directly as a source of electricity without producing dangerous neutrons.

${}^3\text{He}$ can be used in fusion reactions by either of the reactions ${}^2\text{H} + {}^3\text{He} \rightarrow {}^4\text{He} + {}^1\text{p} + 18.3$ MeV, or ${}^3\text{He} + {}^3\text{He} \rightarrow {}^4\text{He} + 2\ {}^1\text{p} + 12.86$ MeV.

The conventional deuterium + tritium ("D-T") fusion process produces energetic neutrons which render reactor components radioactive with activation products. The appeal of helium-3 fusion stems from the aneutronic nature of its reaction products. Helium-3 itself is non-radioactive. The lone high-energy by-product, the proton, can be contained using electric and magnetic fields. The momentum energy of

Comparison of neutronicity of reactions[10][11][12][13][14]

Reactants		Products	*Q*	n/MeV
First-generation fusion fuels				
${}^2\text{D} + {}^2\text{D}$	→	${}^3\text{He} + {}^1_0\text{n}$	3.268 MeV	0.306
${}^2\text{D} + {}^2\text{D}$	→	${}^3\text{T} + {}^1_1\text{p}$	4.032 MeV	0
${}^2\text{D} + {}^3\text{T}$	→	${}^4\text{He} + {}^1_0\text{n}$	17.571 MeV	0.057
Second-generation fusion fuel				
${}^2\text{D} + {}^3\text{He}$	→	${}^4\text{He} + {}^1_1\text{p}$	18.354 MeV	0

this proton (created in the fusion process) will interact with the containing electromagnetic field, resulting in direct net electricity generation.[15]

Because of the higher Coulomb barrier, the temperatures required for ^{2}H + ^{3}He fusion are much higher than those of conventional D-T fusion. Moreover, since both reactants need to be mixed together to fuse, reactions between nuclei of the same reactant will occur, and the D-D reaction (^{2}H + ^{2}H) does produce a neutron. Reaction rates vary with temperature, but the D-^{3}He reaction rate is never greater than 3.56 times the D-D reaction rate (see graph). Therefore, fusion using D-^{3}He fuel at the right temperature and a D-lean fuel mixture, can produce a much lower neutron flux than D-T fusion, but is not clean, negating some of its main attraction.

Third-generation fusion fuels				
$^{3}He + ^{3}He$	→	$^{4}He + 2^{1}_{1}p$	12.86 MeV	0
$^{11}B + ^{1}_{1}p$	→	$3\ ^{4}He$	8.68 MeV	0
Net result of D burning (sum of first 4 rows)				
6 D	→	2(^{4}He + n + p)	43.225 MeV	0.046
Current nuclear fuel				
^{235}U + n	→	2 FP+ 2.5n	~200 MeV	0.001

The second possibility, fusing ^{3}He with itself (^{3}He + ^{3}He), requires even higher temperatures (since now both reactants have a +2 charge), and thus is even more difficult than the D-^{3}He reaction. However, it does offer a possible reaction that produces no neutrons; the charged protons produced can be contained using electric and magnetic fields, which in turn results in direct electricity generation. ^{3}He + ^{3}He fusion is feasible as demonstrated in the laboratory and has immense advantages, but commercial viability is many years in the future.[16]

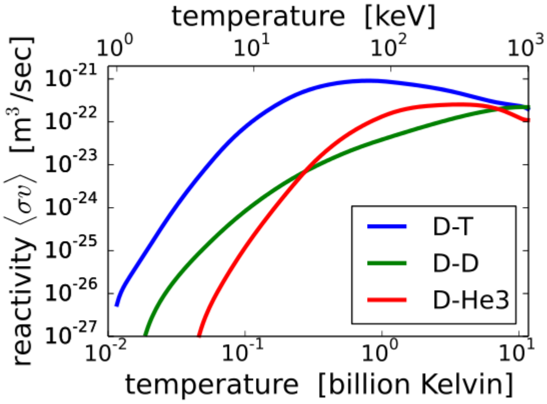

The fusion reaction rate increases rapidly with temperature until it maximizes and then gradually drops off. The DT rate peaks at a lower temperature (about 70 keV, or 800 million kelvins) and at a higher value than other reactions commonly considered for fusion energy.

The amounts of helium-3 needed as a replacement for conventional fuels are substantial by comparison to amounts currently available. The total amount of energy produced in the ^{2}D + ^{3}He reaction is 18.4 MeV, which corresponds to some 493 megawatt-hours (4.93×10^{8} W·h) per three grams (one mole) of ^{3}He If the total amount of energy could be converted to electrical power with 100% efficiency (a physical impossibility), it would correspond to about 30 minutes of output of a gigawatt electrical plant per mole of ^{3}He. Thus, a year's production (at 6 grams for each operation hour) would require 52.5 kilograms of helium-3. The amount of fuel needed for large-scale applications can also be put in terms of total consumption: electricity consumption by 107 million U.S. households in 2001[17] totaled 1,140 billion kW·h (1.14×10^{15} W·h). Again assuming 100% conversion efficiency, 6.7 tonnes per year of helium-3 would be required for that segment of the energy demand of the United States, 15 to 20 tonnes per year given a more realistic end-to-end conversion efficiency.

Neutron detection

Helium-3 is a most important isotope in instrumentation for neutron detection. It has a high absorption cross section for thermal neutron beams and is used as a converter gas in neutron detectors. The neutron is converted through the nuclear reaction

$$n + {}^{3}He \rightarrow {}^{3}H + {}^{1}H + 0.764\ \mathrm{MeV}$$

into charged particles tritium ions (T, ^3H) and Hydrogen ions, or protons (p, ^1H) which then are detected by creating a charge cloud in the stopping gas of a proportional counter or a Geiger-Müller tube.[18]

Furthermore, the absorption process is strongly spin-dependent, which allows a spin-polarized helium-3 volume to transmit neutrons with one spin component while absorbing the other. This effect is employed in neutron polarization analysis, a technique which probes for magnetic properties of matter.[19][20][21][22]

The United States Department of Homeland Security had hoped to deploy detectors to spot smuggled plutonium in shipping containers by their neutron emissions, but the worldwide shortage of helium-3 following the drawdown in nuclear weapons production since the Cold War has to some extent prevented this.[23] As of 2012, DHS determined the commercial supply of boron-10 would support converting its neutron detection infrastructure to that technology.[24]

Cryogenics

A helium-3 refrigerator uses helium-3 to achieve temperatures of 0.2 to 0.3 kelvin. A dilution refrigerator uses a mixture of helium-3 and helium-4 to reach cryogenic temperatures as low as a few thousandths of a kelvin.[25]

An important property of helium-3, which distinguishes it from the more common helium-4, is that its nucleus is a fermion since it contains an odd number of spin $1/2$ particles. Helium-4 nuclei are bosons, containing an even number of spin $1/2$ particles. This is a direct result of the addition rules for quantized angular momentum. At low temperatures (about 2.17 K), helium-4 undergoes a phase transition: A fraction of it enters a superfluid phase that can be roughly understood as a type of Bose–Einstein condensate. Such a mechanism is not available for helium-3 atoms, which are fermions. However, it was widely speculated that helium-3 could also become a superfluid at much lower temperatures, if the atoms formed into *pairs* analogous to Cooper pairs in the BCS theory of superconductivity. Each Cooper pair, having integer spin, can be thought of as a boson. During the 1970s, David Lee, Douglas Osheroff and Robert Coleman Richardson discovered two phase transitions along the melting curve, which were soon realized to be the two superfluid phases of helium-3.[26][27] The transition to a superfluid occurs at 2.491 millikelvins on the melting curve. They were awarded the 1996 Nobel Prize in Physics for their discovery. Tony Leggett won the 2003 Nobel Prize in Physics for his work on refining understanding of the superfluid phase of helium-3.[28]

In a zero magnetic field, there are two distinct superfluid phases of ^3He, the A-phase and the B-phase. The B-phase is the low-temperature, low-pressure phase which has an isotropic energy gap. The A-phase is the higher temperature, higher pressure phase that is further stabilized by a magnetic field and has two point nodes in its gap. The presence of two phases is a clear indication that ^3He is an unconventional superfluid (superconductor), since the presence of two phases requires an additional symmetry, other than gauge symmetry, to be broken. In fact, it is a *p*-wave superfluid, with spin one, $\mathbf{S}=1$, and angular momentum one, $\mathbf{L}=1$. The ground state corresponds to total angular momentum zero, $\mathbf{J}=\mathbf{S}+\mathbf{L}=0$ (vector addition). Excited states are possible with non-zero total angular momentum, $\mathbf{J}>0$, which are excited pair collective modes. Because of the extreme purity of superfluid ^3He (since all materials except ^4He have solidified and sunk to the bottom of the liquid ^3He and any ^4He has phase separated entirely, this is the most pure condensed matter state), these collective modes have been studied with much greater precision than in any other unconventional pairing system.

Medical lung imaging

Helium-3 nuclei have an intrinsic nuclear spin of $1/2$, and a relatively high magnetogyric ratio. Helium-3 can be hyperpolarized using non-equilibrium means such as spin-exchange optical pumping.[29] During this process, circularly polarized infrared laser light, tuned to the appropriate wavelength, is used to excite electrons in an alkali metal, such as caesium or rubidium inside a sealed glass vessel. The angular momentum is transferred from the alkali metal electrons to the noble gas nuclei through collisions. In essence, this process effectively aligns the nuclear spins with the magnetic field in order to enhance the NMR signal. The hyperpolarized gas may then be stored at pressures of 10 atm, for up to 100 hours. Following inhalation, gas mixtures containing the hyperpolarized helium-3 gas can be imaged with an MRI scanner to produce anatomical and functional images of lung ventilation. This technique is also able to produce images of the airway tree, locate unventilated defects, measure the alveolar oxygen partial pressure, and measure the ventilation/perfusion ratio. This technique may be critical for the diagnosis and treatment management of chronic respiratory diseases such as chronic obstructive pulmonary disease (COPD), emphysema, cystic fibrosis, and asthma.[30]

Industrial production

Production, sales and distribution of helium-3 in the United States are managed by the US Department of Energy (DOE) Isotope Program.[31] Virtually all helium-3 used in industry today is produced from the radioactive decay of tritium. Tritium is a radioactive isotope of hydrogen and is typically produced by bombarding lithium-6 with neutrons in a nuclear reactor. The lithium nucleus absorbs a neutron and splits into helium-4 and tritium. Tritium decays into helium-3 with a half-life of 12.3 years, so helium-3 can be produced by simply storing the tritium until it undergoes radioactive decay.

Tritium is a critical component of nuclear weapons and historically it was produced and stockpiled primarily for this application. The decay of tritium into helium-3 reduces the explosive power of the fusion warhead, so periodically the accumulated helium-3 must be removed from warhead reservoirs and tritium in storage. Helium-3 removed during this process is marketed for other applications.

For decades this has been, and remains, the principal source of the world's helium-3.[32] However, since the signing of the START I Treaty in 1991 the number of nuclear warheads that are kept ready for use has decreased[33][34] This has reduced the quantity of helium-3 available from this source. Helium-3 stockpiles have been further diminished by increased demand,[35] primarily for use in neutron radiation detectors and medical diagnostic procedures. US industrial demand for helium-3 reached a peak of 70,000 liters (approximately 8 kg) per year in 2008. Price at auction, historically about $100/liter, reached as high as $2000/liter[36] Since then, demand for helium-3 has declined to about 6000 liters per year due to the high cost and efforts by the DOE to recycle it and find substitutes.

The DOE recognized the developing shortage of both tritium and helium-3. and began producing tritium by lithium irradiation at the Tennessee Valley Authority's Watts Bar Nuclear Generating Station in 2010.[35] In this process tritium-producing burnable absorber rods (TPBARs) containing lithium in a ceramic form are inserted into the reactor in place of the normal boron control rods[37] Periodically the TPBARs are replaced and the tritium extracted.

Currently only one reactor is used for tritium production but the process could, if necessary, be vastly scaled up to meet any conceivable demand simply by utilizing more of the nation's power reactors. Substantial quantities of tritium and helium-3 could also be extracted from the heavy water moderator in CANDU nuclear reactors.[35][38]

Natural abundance

Solar nebula (primordial) abundance

One early estimate of the primordial ratio of ^{3}He to ^{4}He in the solar nebula has been the measurement of their ratio in the atmosphere of Jupiter, measured by the mass spectrometer of the Galileo atmospheric entry probe. This ratio is about 1:10,000,[39] or 100 parts of ^{3}He per million parts of ^{4}He. This is roughly the same ratio of the isotopes as in lunar regolith, which contains 28 ppm helium-4 and 2.8 ppb helium-3 (which is at the lower end of actual sample measurements, which vary from about 1.4 to 15 ppb). However, terrestrial ratios of the isotopes are lower by a factor of 100, mainly due to enrichment of helium-4 stocks in the mantle by billions of years of alpha decay from uranium and thorium.

Terrestrial abundance

^{3}He is a primordial substance in the Earth's mantle, considered to have become entrapped within the Earth during planetary formation. The ratio of ^{3}He to ^{4}He within the Earth's crust and mantle is less than that for assumptions of solar disk composition as obtained from meteorite and lunar samples, with terrestrial materials generally containing lower ^{3}He/^{4}He ratios due to ingrowth of ^{4}He from radioactive decay.

^{3}He has a cosmological ratio of 300 atoms per million atoms of ^{4}He (at. ppm),[40] leading to the assumption that the original ratio of these primordial gases in the mantle was around 200-300 ppm when Earth was formed. A lot of ^{4}He was generated by alpha-particle decay of uranium and thorium, and now the mantle has only around 7% primordial helium,[40] lowering the total ^{3}He/^{4}He ratio to around 20 ppm. Ratios of ^{3}He/^{4}He in excess of atmospheric are indicative of a contribution of ^{3}He from the mantle. Crustal sources are dominated by the ^{4}He which is produced by the decay of radioactive elements in the crust and mantle.

The ratio of helium-3 to helium-4 in natural Earth-bound sources varies greatly.[41][42] Samples of the lithium ore spodumene from Edison Mine, South Dakota were found to contain 12 parts of helium-3 to a million parts of helium-4. Samples from other mines showed 2 parts per million.[41]

Helium is also present as up to 7% of some natural gas sources,[43] and large sources have over 0.5% (above 0.2% makes it viable to extract).[44] The fraction of ^{3}He in helium separated from natural gas in the U.S. was found to range from 70 to 242 parts per billion.[35][45] Hence the US 2002 stockpile of 1 billion normal m^3[44] would have contained about 12 to 43 kilograms of helium-3. According to one expert, about 26 m^3 or almost 5 kg of ^{3}He is available annually for separation from the US natural gas stream. If the process of separating out the ^{3}He could employ as feedstock the liquefied helium typically used to transport and store bulk quantities, estimates for the incremental energy cost range from US$34 to $300 per liter NTP, excluding the cost of infrastructure and equipment.[35] Algeria's annual gas production is assumed to contain 100 million normal cubic metres[44] and this would contain between 7 and 24 m^3 of helium-3 (about 1 to 4 kilograms) assuming a similar ^{3}He fraction.

^{3}He is also present in the Earth's atmosphere. The natural abundance of ^{3}He in naturally occurring helium gas is 1.38×10^{-6} (1.38 parts per million). The partial pressure of helium in the Earth's atmosphere is about 0.52 Pa, and thus helium accounts for 5.2 parts per million of the total pressure (101325 Pa) in the Earth's atmosphere, and ^{3}He thus accounts for 7.2 parts per trillion of the atmosphere. Since the atmosphere of the Earth has a mass of about 5.14×10^{15} tonnes,[46] the mass of ^{3}He in the Earth's atmosphere is the product of these numbers, or about 37,000 tonnes of ^{3}He.

^{3}He is produced on Earth from three sources: lithium spallation, cosmic rays, and beta decay of tritium (^{3}H). The contribution from cosmic rays is negligible within all except the oldest regolith materials, and lithium spallation reactions are a lesser contributor than the production of ^{4}He by alpha particle emissions.

The total amount of helium-3 in the mantle may be in the range of 0.1–1 million tonnes. However, most of the mantle is not directly accessible. Some helium-3 leaks up through deep-sourced hotspot volcanoes such as those of the Hawaiian Islands, but only 300 grams per year is emitted to the atmosphere. Mid-ocean ridges emit another 3 kilogram per year. Around subduction zones, various sources produce helium-3 in natural gas deposits which possibly contain a thousand tonnes of helium-3 (although there may be 25 thousand tonnes if all ancient subduction zones have such deposits). Wittenberg estimated that United States crustal natural gas sources may have only half a tonne total.[47] Wittenberg cited Anderson's estimate of another 1200 metric tonnes in interplanetary dust particles on the ocean floors.[48] In the 1994 study, extracting helium-3 from these sources consumes more energy than fusion would release.[49]

Extraction from extraterrestrial sources

Materials on the Moon's surface contain helium-3 at concentrations between 1.4 and 15 ppb in sunlit areas,[50][51] and may contain concentrations as much as 50 ppb in permanently shadowed regions.[7] A number of people, starting with Gerald Kulcinski in 1986,[52] have proposed to explore the moon, mine lunar regolith and use the helium-3 for fusion. Because of the low concentrations of helium-3, any mining equipment would need to process extremely large amounts of regolith (over 150 million tonnes of regolith to obtain one gram of helium 3),[53] and some proposals have suggested that helium-3 extraction be piggybacked onto a larger mining and development operation.

The primary objective of Indian Space Research Organisation's first lunar probe called Chandrayaan-I, launched on October 22, 2008, was reported in some sources to be mapping the Moon's surface for helium-3-containing minerals.[54] However, no such objective is mentioned in the project's official list of goals, although many of its scientific payloads have noted helium-3-related applications.[55][56]

Cosmochemist and geochemist Ouyang Ziyuan from the Chinese Academy of Sciences who is now in charge of the Chinese Lunar Exploration Program has already stated on many occasions that one of the main goals of the program would be the mining of helium-3, from which operation "each year, three space shuttle missions could bring enough fuel for all human beings across the world."[57] In January 2006, the Russian space company RKK Energiya announced that it considers lunar helium-3 a potential economic resource to be mined by 2020,[58] if funding can be found.[59][60]

Mining gas giants for helium-3 has also been proposed.[61] The British Interplanetary Society's hypothetical Project Daedalus interstellar probe design was fueled by helium-3 mines in the atmosphere of Jupiter, for example. Jupiter's high gravity makes this a less energetically favorable operation than extracting helium-3 from the other gas giants of the solar system, however.

Not all authors feel the extraterrestrial extraction of helium-3 is feasible. Dwayne Day, writing in The Space Review, identifies some major obstacles to helium-3 extraction from extraterrestrial sources for use in fusion, and questions the feasibility of extraterrestrial extraction when compared to production on Earth.[62]

Power generation

A second-generation approach to controlled fusion power involves combining helium-3 ($^{3}_{2}\mathrm{He}$) and deuterium ($^{2}_{1}\mathrm{H}$). This reaction produces a helium-4 ion ($^{4}_{2}\mathrm{He}$) (like an alpha particle, but of different origin) and a high-energy proton (positively charged hydrogen ion) ($^{1}_{1}\mathrm{p}$). The most important potential advantage of this fusion reaction for power production as well as other applications lies in its compatibility with the use of electrostatic fields to control fuel ions and the fusion protons. High speed protons, as positively charged particles, can have their kinetic energy converted directly into electricity, through use of solid-state conversion materials as well as other techniques. Potential conversion efficiencies of 70% may be possible, as there is no need to convert proton energy to heat in order to drive a turbine-powered electrical generator.

There have been many claims about the capabilities of helium-3 power plants. According to proponents, fusion power plants operating on deuterium and helium-3 would offer lower capital and operating costs than their competitors due to less technical complexity, higher conversion efficiency, smaller size, the absence of radioactive fuel, no air or water pollution, and only low-level radioactive waste disposal requirements. Recent estimates suggest that about $6 billion in investment capital will be required to develop and construct the first helium-3 fusion power plant. Financial breakeven at today's wholesale electricity prices (5 US cents per kilowatt-hour) would occur after five 1-gigawatt plants were on line, replacing old conventional plants or meeting new demand.[63]

The reality is not so clear-cut. The most advanced fusion programs in the world are inertial confinement fusion (such as National Ignition Facility) and magnetic confinement fusion (such as ITER and Wendelstein 7-X). In the case of the former, there is no solid roadmap to power generation. In the case of the latter, commercial power generation is not expected until around 2050.[64] In both cases, the type of fusion discussed is the simplest: D-T fusion. The reason for this is the very low Coulomb barrier for this reaction; for D+^{3}He, the barrier is much higher, and it is even higher for ^{3}He–^{3}He. The immense cost of reactors like ITER and National Ignition Facility are largely due to their immense size, yet to scale up to higher plasma temperatures would require reactors far larger still. The 14.7 MeV proton and 3.6 MeV alpha particle from D–^{3}He fusion, plus the higher conversion efficiency, means that more electricity is obtained per kilogram than with D-T fusion (17.6 MeV), but not that much more. As a further downside, the rates of reaction for helium-3 fusion reactions are not particularly high, requiring a reactor that is larger still or more reactors to produce the same amount of electricity.

To attempt to work around this problem of massively large power plants that may not even be economical with D-T fusion, let alone the far more challenging D–^{3}He fusion, a number of other reactors have been proposed – the Fusor, Polywell, Focus fusion, and many more, though many of these concepts have fundamental problems with achieving a net energy gain, and generally attempt to achieve fusion in thermal disequilibrium, something that could potentially prove impossible,[65] and consequently, these long-shot programs tend to have trouble garnering funding despite their low budgets. Unlike the "big", "hot" fusion systems, however, if such systems were to work, they could scale to the higher barrier "aneutronic" fuels, and therefore their proponents tend to promote p-B fusion, which requires no exotic fuels like helium-3.

References in science fiction

Several science fiction works have featured helium-3 extraction on the moon, including the films *Moon* (2009) and *Iron Sky* (2012), the manga and corresponding anime Planetes, the video game *Anno 2205* (2015) and the novel *Luna: New Moon* (2015). The novel *Morning Star* (Pierce Brown, 2016) features helium-3 mining on Phobos (a moon of Mars), while his novel *Red Rising* (2014) features helium-3 extraction from Mars itself. The titular mech of Godzilla vs. Mechagodzilla II is powered by heavy hydrogen and helium-3 in pellet form.

See also

- Colonization of the Moon

Notes and references

1. Oliphant, M. L. E.; Harteck, P.; Rutherford, E. (1934). "Transmutation Effects Observed with Heavy Hydrogen". *Proceedings of the Royal Society A*. **144** (853): 692–703. Bibcode:1934RSPSA.144..692O (http://adsabs.harvard.edu/abs/1934RSPSA.144..692O). doi:10.1098/rspa.1934.0077 (https://doi.org/10.1098%2Frspa.1934.0077). JSTOR 2935553 (https://www.jstor.org/stable/2935553).
2. "Lawrence and His Laboratory: Episode: A Productive Error" (http://www.lbl.gov/Science-Articles/Research-Review/Magazine/1981/81fepi1.html). Newsmagazine Publication. 1981. Retrieved 2009-09-01.
3. Matson, John (12 Jun 2009). "Is MOON's Sci-Fi Vision of Lunar Helium 3 Mining Based in Reality?" (https://blogs.scientificamerican.com/news-blog/is-moons-sci-fi-vision-of-lunar-hel-2009-06-12/#). *Scientific American – News Blog*. Retrieved 29 Aug 2017.
4. Close, Frank (August 2007). "Fears Over Factoids" (http://cds.cern.ch/record/1055767/files/CM-PRS00002036.pdf) (PDF). *CERN Document Server*. Physicsworld.com. Retrieved 8 July 2018.
5. Fa WenZhe & Jin YaQiu (December 2010). "Global inventory of Helium-3 in lunar regoliths estimated by a multi-channel microwave radiometer on the Chang-E 1 lunar satellite" (http://lunarnetworks.blogspot.in/2010/12/change-1-maps-moons-helium-3-inventory.html).

RESEARCH, SCIENCE & ENVIRONMENT, TECHNOLOGY & ENGINEERING

Scientists unveil new form of matter: time crystals

By **Robert Sanders**, Media relations | JANUARY 26, 2017

121

To most people, crystals mean diamond bling, semiprecious gems or perhaps the jagged amethyst or quartz crystals beloved by collectors.

To Norman Yao, these inert crystals are the tip of the iceberg.

If crystals have an atomic structure that repeats in space, like the carbon lattice of a diamond, why can't crystals also have a structure that repeats in time? That is, a time crystal?

In a paper published online last week in the journal *Physical Review Letters*, the UC Berkeley assistant professor of physics describes exactly how to make and measure the properties of such a crystal, and even predicts what the various phases surrounding the time crystal should be — akin to the liquid and gas phases of ice.

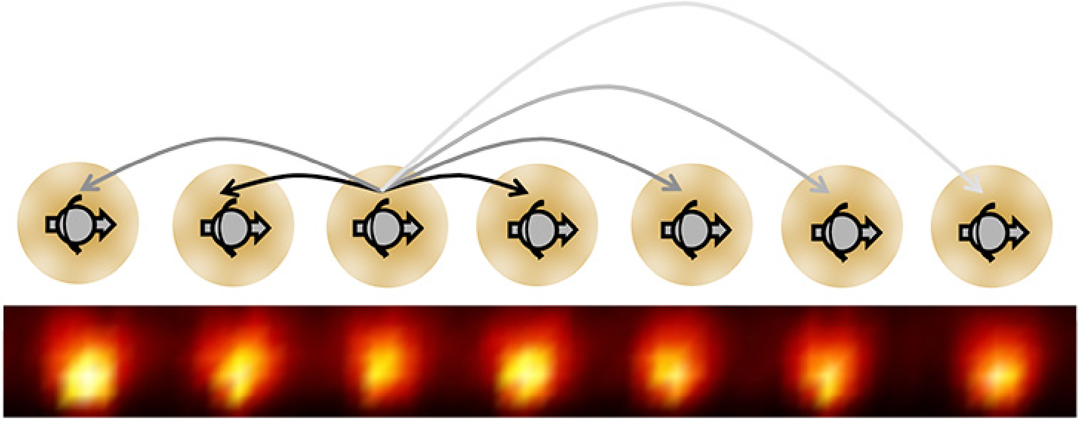

A one-dimensional chain of ytterbium ions was turned into a time crystal by physicists at the University of Maryland, based on a blueprint provided by UC Berkeley's Norman Yao. Each ion behaves like an electron spin and exhibits long-range interactions indicated as arrows. (Image courtesy of Chris Monroe)

This is not mere speculation. Two groups followed Yao's blueprint and have already created the first-ever time crystals. The groups at the University of Maryland and Harvard University reported their successes, using two totally different setups, in papers posted online last year, and have submitted the results for publication. Yao is a co-author on both papers.

Time crystals repeat in time because they are kicked periodically, sort of like tapping Jell-O repeatedly to get it to jiggle, Yao said. The big breakthrough, he argues, is less that these particular crystals repeat in time than that they are the first of a large class of new materials that are intrinsically out of equilibrium, unable to settle down to the motionless equilibrium of, for example, a diamond or ruby.

"This is a new phase of matter, period, but it is also really cool because it is one of the first examples of non-equilibrium matter," Yao said. "For the last half-century, we have been exploring equilibrium matter, like metals and insulators. We are just now starting to explore a whole new landscape of non-equilibrium matter."

While Yao is hard put to imagine a use for a time crystal, other proposed phases of non-equilibrium matter theoretically hold promise as nearly perfect memories and may be useful in quantum computers.

An ytterbium chain

The time crystal created by Chris Monroe and his colleagues at the University of Maryland employs a conga line of 10 ytterbium ions whose electron spins interact, similar to the qubit systems being tested as quantum computers. To keep the ions out of equilibrium, the researchers alternately hit them with one laser to create an effective magnetic field and a second laser to partially flip the spins of the atoms, repeating the sequence many times. Because the spins interacted, the atoms settled into a stable, repetitive pattern of spin flipping that defines a crystal.

Time crystals were first proposed in 2012 by Nobel laureate Frank Wilczek, and last year theoretical physicists at Princeton University and UC Santa Barbara's **Station Q** independently proved that such a crystal could be made. According to Yao, the UC Berkeley group was "the bridge between the theoretical idea and the experimental implementation."

From the perspective of quantum mechanics, electrons can form crystals that do not match the underlying spatial translation symmetry of the orderly, three-dimensional array of atoms, Yao said. This breaks the symmetry of the material and leads to unique and stable properties we define as a crystal.

A time crystal breaks time symmetry. In this particular case, the magnetic field and laser periodically driving the ytterbium atoms produce a repetition in the system at twice the period of the drivers, something that would not occur in a normal system.

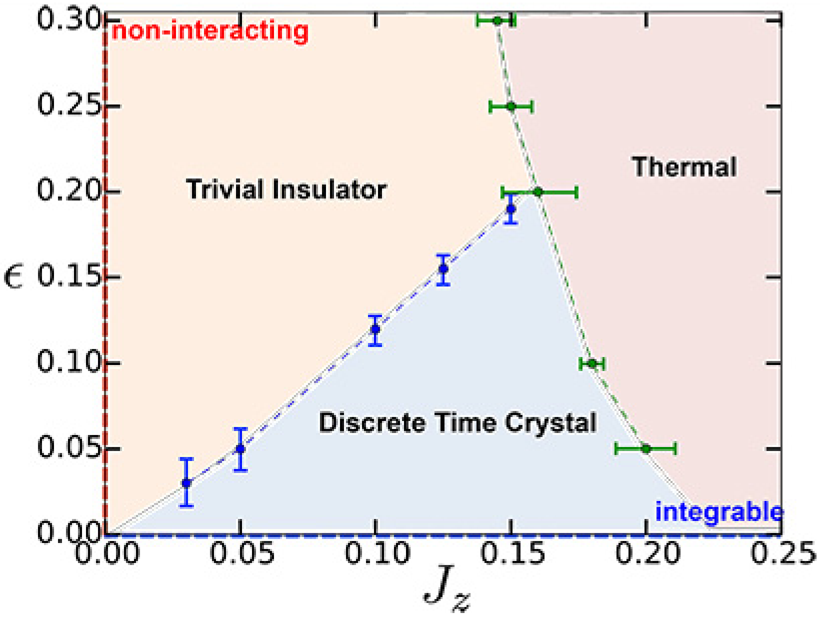

This phase diagram shows how changing the experimental parameters can 'melt' a time crystal into a normal insulator or heat up a time crystal to a high temperature thermal state. Norman Yao graphic.

"Wouldn't it be super weird if you jiggled the Jell-O and found that somehow it responded at a different period?" Yao said. "But that is the essence of the time crystal. You have some periodic driver that has a period 'T', but the system somehow synchronizes so that you observe the system oscillating with a period that is larger than 'T'."

Yao worked closely with Monroe as his Maryland team made the new material, helping them focus on the important properties to measure to confirm that the material was in fact a stable or rigid time crystal. Yao also described how the time crystal would change phase, like an ice cube melting, under different magnetic fields and laser pulsing.

The Harvard team, led by Mikhail Lukin, set up its time crystal using densely packed nitrogen vacancy centers in diamonds.

FORENSIC CHEMISTRY HANDBOOK

P
S
32.066
Ge
33
As
74.9216
34
Se
78.96
51
Sb
121.75
52
Te
127.60
53
I
84
Br
79.90

Edited by

LAWRENCE F. KOBILINSKY

WILEY

CONTENTS

Preface **xv**

Contributors **xxi**

1. Forensic Environmental Chemistry **1**
Anthony Carpi and Andrew J. Schweighardt

1.1 Introduction 2
1.2 Chemical Fingerprinting 4
  1.2.1 Hydrocarbon Mixtures 4
  1.2.2 Polycyclic Aromatic Hydrocarbons 6
  1.2.3 Biomarkers 11
  1.2.4 Additives 11
  1.2.5 Isotopes 12
  1.2.6 Tracers 13
  1.2.7 Methods of Detection 16
  1.2.8 Weathering 18
1.3 Spatial Association of Environmental Incidents 18
References 20

2. Principles and Issues in Forensic Analysis of Explosives **23**
Jimmie C. Oxley, Maurice Marshall, and Sarah L. Lancaster

2.1 Introduction 24
2.2 Sample Collection 25
2.3 Packaging 29
2.4 Sorting 30
2.5 Documentation 31
2.6 Environmental Control and Monitoring 31
2.7 Storage 33
2.8 Analysis 33
2.9 Records 36
2.10 Quality Assurance 36
2.11 Safety and Other Issues 37
Conclusion 37
References 38

3. Analysis of Fire Debris **41**
John J. Lentini

3.1 Introduction 42
3.2 Evolution of Separation Techniques 43
3.3 Evolution of Analytical Techniques 47
3.4 Evolution of Standard Methods 49
3.5 Isolating the Residue 51
  3.5.1 Initial Sample Evaluation 51
  3.5.2 ILR Isolation Method Selection 51
  3.5.3 Solvent Selection 54
  3.5.4 Internal Standards 54
  3.5.5 Advantages and Disadvantages of Isolation Methods 56
3.6 Analyzing the Isolated ILR 56
  3.6.1 Criteria for Identification 63
  3.6.2 Improving Sensitivity 90
  3.6.3 Estimating the Degree of Evaporation 95
  3.6.4 Identity of Source 98
3.7 Reporting Procedures 101
3.8 Record Keeping 102
3.9 Quality Assurance 105
Conclusion 105
References 106

4. Forensic Examination of Soils **109**
Raymond C. Murray

4.1 Introduction 110
4.2 Murder and the Pond 111
4.3 Oil Slicks and Sands 113
4.4 Medical Link 114
4.5 Examination Methods 114
  4.5.1 Color 115
  4.5.2 Particle-Size Distribution 117
  4.5.3 Stereo Binocular Microscope 120
  4.5.4 Petrographic Microscope 122
  4.5.5 Refractive Index 124
  4.5.6 Cathodoluminescence 124
  4.5.7 Scanning Electron Microscope 125
  4.5.8 X-Ray Diffraction 126
4.6 Chemical Methods 127
  4.6.1 FTIR and Raman Spectroscopy 128
4.7 Looking Ahead 129
References 130

5. Analysis of Paint Evidence **131**
Scott G. Ryland and Edward M. Suzuki
5.1 Introduction 132
5.2 Paint Chemistry and Color Science 134
    5.2.1 Binders 134
    5.2.2 Pigments 136
5.3 Types of Paint 139
    5.3.1 Automotive Finish Systems 139
    5.3.2 Architectural Coatings (Structural Paints or House Paints) 140
    5.3.3 Other Coatings 141
5.4 Paint Evidence Interpretation Considerations 141
5.5 Analytical Methods 142
    5.5.1 Microscopic Examinations 143
    5.5.2 Physical Nature of the Transfer 147
    5.5.3 Microscopy 149
    5.5.4 Microspectrophotometry 152
    5.5.5 Infrared Spectroscopy 158
    5.5.6 Raman Spectroscopy 175
    5.5.7 Pyrolysis Gas Chromatography and Pyrolysis Gas Chromatography–Mass Spectrometry 178
    5.5.8 Elemental Analysis Methods 188
    5.5.9 Other Methods 205
5.6 Examples 208
    5.6.1 Example 1 208
    5.6.2 Example 2 210
    5.6.3 Example 3 213
References 217

6. Analysis Techniques Used for the Forensic Examination of Writing and Printing Inks **225**
Gerald M. LaPorte and Joseph C. Stephens
6.1 Introduction 226
6.2 Ink 226
    6.2.1 Ink Composition 227
6.3 Ink Analysis 230
    6.3.1 Physical Examinations 233
    6.3.2 Optical Examinations 236
    6.3.3 Chemical Examinations 238
    6.3.4 Ink Dating 240
6.4 Office Machine Systems 242
    6.4.1 Inkjet Ink 242

6.4.2 Inkjet Ink Analysis 243
6.4.3 Toner Printing 245
6.4.4 Toner Analysis 246
Conclusion 247
References 248

7. The Role of Vibrational Spectroscopy in Forensic Chemistry 251
Ali Koçak

7.1 Introduction to Vibrational Spectroscopy 252
7.2 Infrared Spectroscopy 253
7.3 Infrared Sampling Techniques 255
7.3.1 Transmission Spectroscopy 255
7.3.2 External Reflection Spectroscopy 255
7.3.3 Attenuated Total Reflectance 256
7.3.4 Diffuse Reflectance Spectroscopy 258
7.3.5 Infrared Microspectroscopy 259
7.4 Raman Spectroscopy 260
7.5 Raman Spectroscopic Techniques 262
7.5.1 Surface-Enhanced Raman Spectroscopy 262
7.5.2 Resonance Raman Scattering 263
7.5.3 Coherent anti-Stokes Raman Spectroscopy 263
7.5.4 Confocal Raman Spectroscopy 263
7.6 Applications of Vibrational Spectroscopy in Forensic Analysis 264
References 265

8. Forensic Serology 269
Richard Li

8.1 Introduction 270
8.2 Identification of Blood 271
8.2.1 Oxidation–Reduction Reactions 272
8.2.2 Microcrystal Assays 275
8.2.3 Other Assays for Blood Identification 275
8.3 Species Identification 278
8.3.1 Immunochromatographic Assays 278
8.3.2 Ouchterlony Assay 280
8.3.3 Crossed-Over Immunoelectrophoresis 281
8.4 Identification of Semen 282
8.4.1 Visual Examination 282
8.4.2 Acid Phosphatase Assays 283
8.4.3 Microscopic Examination of Spermatozoa 284
8.4.4 Immunochromatographic Assays 285

8.4.5 RNA-Based Assays 286
8.5 Identification of Saliva 286
8.5.1 Visual and Microscopic Examination 287
8.5.2 Identification of Amylase 287
8.5.3 RNA-Based Assays 289
References 289

9. Forensic DNA Analysis 291
Henrietta Margolis Nunno

9.1 Introduction 292
9.1.1 Background on DNA Typing 292
9.1.2 DNA Structure 294
9.1.3 Nuclear and Mitochondrial DNA Organization 295
9.2 Methodology 296
9.2.1 Sample Collection and DNA Extraction 296
9.2.2 DNA Quantification 297
9.2.3 Polymerase Chain Reaction 298
9.2.4 Short Tandem Repeats 298
9.2.5 PCR of STRs 300
9.2.6 Separation and Sizing of STR Alleles 301
9.2.7 Combined DNA Index System (CODIS) Database 305
9.2.8 Frequency and Probability 306
9.3 Problems Encountered in STR Analysis 307
9.3.1 Low-Copy-Number DNA 307
9.3.2 Degraded DNA and Reduced-Size (mini) STR Primer Sets 308
9.3.3 PCR Inhibition 310
9.3.4 Interpretation of Mixtures of DNA 310
9.3.5 Null Alleles and Allele Dropout 311
9.3.6 Factors Causing Extra Peaks in Results Observed 312
9.3.7 Stutter Product Peaks 312
9.3.8 Nontemplate Addition (Incomplete Adenylation) 313
9.3.9 Technological Artifacts 313
9.3.10 Single-Nucleotide Polymorphism Analysis of Autosomal DNA SNPs 313
9.3.11 Methods Used for SNP Analysis 314
9.3.12 Mitochondrial DNA Analysis 315
9.4 Methodology for mtDNA Analysis 316
9.4.1 Preparation of Samples 316

9.4.2 MtDNA Sequencing Methods 316
9.4.3 Reference Sequences 317
9.4.4 Screening Assays for mtDNA 318
9.4.5 Interpretation of mtDNA Sequencing Results 319
9.4.6 Statistics: The Meaning of a Match for mtDNA 320
9.4.7 Heteroplasmy 320
9.4.8 The Future of DNA Analysis 321
References 322

10. Current and Future Uses of DNA Microarrays in Forensic Science 327
Nathan H. Lents

10.1 Introduction 328
10.2 What is a DNA Microarray? 328
10.2.1 cDNA Microarray 329
10.2.2 Other Types of DNA Arrays 330
10.2.3 The Birth of "-omics" 331
10.3 DNA Microarrays in Toxicogenomics 332
10.3.1 Sharing Information 333
10.3.2 Forensic Application 333
10.4 Detection of Microorganisms Using Microarrays 334
10.4.1 Historical Perspective 334
10.4.2 DNA Fingerprinting 335
10.4.3 DNA Fingerprinting by Microarrays 336
10.4.4 DNA Sequence-Based Detection 337
10.4.5 Where DNA Microarrays Come In 337
10.4.6 Looking Forward: Genetic Virulence Signatures 338
10.5 Probing Human Genomes by DNA Microarrays 340
10.5.1 STR Analysis 340
10.5.2 SNP Analysis 343
10.5.3 Exploring an Unknown Genome? 344
Conclusion 345
References 345

11. Date-Rape Drugs with Emphasis on GHB 355
Stanley M. Parsons

11.1 Introduction 357
11.2 Molecular Mechanisms of Action 357
11.2.1 Receptors and Transporters 357

- 11.2.2 Real GHB Receptors 359
11.3 Societal Context of Date-Rape Agents 361
- 11.3.1 Acute Effects of Date-Rape Agents on Cognition and Behavior 361
- 11.3.2 Medicinal Uses of Date-Rape Drugs 361
- 11.3.3 Self-Abuse 362
- 11.3.4 Date Rape, Death, and Regulation 363
11.4 Metabolism Fundamentals 363
- 11.4.1 Complexity in Unraveling Metabolism of GHB-Related Compounds 363
- 11.4.2 Isozymes in GHB-Related Metabolism 364
- 11.4.3 Subcellular Compartmentalization of Enzymes, Transporters, and Substrates 364
- 11.4.4 Dynamics and Equilibria for Enzymes and Transporters 365
- 11.4.5 Thermodynamics-Based Analysis of Metabolic Flux 366
- 11.4.6 Metabolism of Endogenous GHB Versus Ingested GHB and Prodrugs 367
- 11.4.7 Directionality of in Vivo and in Vitro Enzymatic Activity 367
- 11.4.8 Transporters and Enzymes Mediating GHB-Related Metabolism 367
11.5 Biosynthesis of Endogenous GHB 368
- 11.5.1 First Step for GHB Biosynthesis in the Known Pathway 368
- 11.5.2 Second Step for GHB Biosynthesis in the known Pathway 368
- 11.5.3 Third Step for GHB Biosynthesis in the known Pathway 371
- 11.5.4 Which Step in GHB Biosynthesis is Rate Limiting? 373
- 11.5.5 Are There Other Biosynthetic Pathways to Endogenous GHB? 374
11.6 Absorption and Distribution of Ingested GHB 376
- 11.6.1 Gastrointestinal Tract 376
- 11.6.2 Blood 377
11.7 Initial Catabolism of GHB 377
- 11.7.1 Transport into Mitochondria 377
- 11.7.2 Iron-Dependent Alcohol Dehydrogenase ADHFe1 377
- 11.7.3 Poorly Characterized Catabolism of GHB 379

11.8 Chemistry of GHB and Related Metabolites not Requiring Enzymes 380
11.9 Experimental Equilibrium Constants for Redox Reactions of GHB 380
11.10 Estimated Equilibrium Constants for Redox Reactions of GHB in Vivo 381
11.11 Different Perspectives on Turnover of Endogenous GHB are Consistent 384
11.12 Disposition of Succinic Semialdehyde 385
11.13 Conversion of Prodrugs to GHB and Related Metabolites 386
  11.13.1 γ-Butyrolactone 386
  11.13.2 1,4-Butanediol 387
11.14 Subcellular Compartmentalization of GHB-Related Compounds 388
11.15 Comparative Catabolism of Ethanol, 1,4-Butanediol, Fatty Acids, and GHB 389
11.16 Catabolism of MDMA, Flunitrazepam, and Ketamine 390
11.17 Detection of Date-Rape Drugs 390
  11.17.1 Compounds Diagnostic for Dosing by Synthetic Date-Rape Drugs 390
  11.17.2 Compounds Diagnostic for Dosing by GHB 390
  11.17.3 Gold-Standard Testing 391
  11.17.4 Many Applications for Reliable Field Tests 392
  11.17.5 Hospital Emergency Department Example 392
  11.17.6 Preparation of a Sample for Delayed Analysis 393
  11.17.7 Time Window Available to Detect Dosing 393
  11.17.8 Extending the Time Window 394
11.18 Special Circumstances of GHB 395
  11.18.1 Industrial Connection 395
  11.18.2 Enzymes Acting on GHB in Bacteria, Yeast, and Plants 395
  11.18.3 Possible Accidental Intoxication by GHB in the Future 395
11.19 Considerations During Development of Field Tests 396
  11.19.1 Shortcomings of Antibody-Based Screens for Simple Analytes 396
  11.19.2 Advantages of Enzyme-Based Screens for Simple Natural Analytes 397
11.20 Development of an Enzymatic Test for GHB 399
  11.20.1 Sensitivity Required for the Hospital Emergency Department 399

11.20.2 Choice of Enzyme 399
11.20.3 Reliable Field Test for GHB 400
Conclusion 402
Notes 404
References 406

12. Forensic and Clinical Issues in Alcohol Analysis 435
Richard Stripp

12.1 Introduction 436
12.2 Blood Alcohol Concentration 437
12.3 Alcohol Impairment and Driving Skills 441
12.4 Field Sobriety Tests 443
12.5 Blood Alcohol Measurements 444
12.5.1 Enzymatic Methods 444
12.5.2 Headspace Gas Chromatography 445
12.5.3 Breath Alcohol Testing 446
12.5.4 Breath Alcohol Instrumentation 447
12.5.5 Extrapolation from BrAC to BAC 449
12.5.6 Urine and Saliva 450
12.5.7 Ethyl Glucuronide 450
12.5.8 Postmortem Determination of Alcohol 451
12.5.9 Quality Assurance of Alcohol Testing 452
References 453

13. Fundamental Issues of Postmortem Toxicology 457
Donald B. Hoffman, Beth E. Zedeck, and Morris S. Zedeck

13.1 Introduction 458
13.2 Tissue and Fluid Specimens 460
13.2.1 Blood 460
13.2.2 Urine 461
13.2.3 Vitreous Humor and Cerebrospinal Fluid 461
13.2.4 Gastric Contents 462
13.2.5 Meconium 463
13.2.6 Brain 464
13.2.7 Liver and Bile 464
13.2.8 Lung, Spleen, Kidney, and Skin 465
13.2.9 Muscle 465
13.2.10 Bone, Teeth, Nails, and Hair 465
13.2.11 Other Materials for Analysis 466
13.3 Specimen Collection and Storage 466
13.4 Extraction Procedures 467
13.5 Analytical Techniques 467
13.6 Interpretation 470

13.6.1 Postmortem Redistribution 470
13.6.2 Pharmacogenomics 471
13.6.3 Drug Interactions 472
13.6.4 Drug Stability and Decomposed Tissue 473
13.6.5 Effects of Embalming Fluid 474
Conclusion 475
References 476

14. Entomotoxicology: Drugs, Toxins, and Insects 483
Jason H. Byrd and Michelle R. Peace

14.1 Introduction 484
14.2 The Fly and Forensic Science 484
14.2.1 History of Forensic Entomology, Toxicology, and the Rise of Entomotoxicology 485
14.2.2 Drugs and the Fly Life Cycle 488
14.2.3 Why Use Insects as a Toxicological Specimen? 490
14.2.4 Drug Extraction Methods 492
14.2.5 Qualitative Versus Quantitative 493
14.2.6 Changes in Insect Development: Toxins and Drugs 494
14.2.7 The Future of Entomotoxicology 494
References 495

Index 501

PREFACE

In February 2009 a report entitled *Strengthening Forensic Science in the United States: A Path Forward* was issued by the National Research Council (NRC) of the National Academy of Sciences. The committee members who wrote the report included scientists, judges, lawyers, statisticians, and forensic scientists. The authors of the report recognized that there is an ongoing need to assure that evidence analysis is held to the highest standards and that what is reported in writing and in testimony must be reliable and credible. Forensic science is a large umbrella science consisting of many subdisciplines, including serology, forensic DNA analysis, toxicology, document examination, hair and fiber analysis, arson investigation, firearms and toolmarks, explosives analysis, blood spatter pattern analysis, digital evidence, impression evidence, forensic pathology, forensic anthropology, forensic odontology, and others. Crime scene personnel are trained to identify and collect biological and physical evidence for subsequent laboratory analysis. This evidence can shed light on events leading up to and during the commission of a crime. Often, evidence collected at a crime scene can be associated with either a victim or a suspect. Criminals can often be linked to the crime scene and/or to the victim. At the same time, those falsely accused can be excluded or exonerated based on reliable analysis of physical evidence. Following scientific analysis, the criminalist writes a report and will testify in a court of law. The testimony of an expert witness must be unbiased, accurate, and based on a sound scientific foundation. The admissibility of novel scientific evidence must be determined prior to expert testimony. The use of a novel scientific method to analyze evidence and the subsequent testimony that describes the results of such testing can be challenged for a variety of reasons.

The *Strengthening Forensic Science* (SFS) report recognized that not all forensic disciplines were at the same level with respect to standards and reliability. This problem is based in part on insufficient funding for forensic research and lack of oversight by any national organization. Although ASCLD-LAB accredits forensic laboratories, American Board of Criminalistics (ABC) certifies forensic analysts, and the National Institute of Justice provides funding for some forensic research projects, techniques and principles used in many forensic disciplines are not at the same level of reliability as that achieved by DNA scientists. The science of using DNA for human identification dates back to the work of Alec Jeffries, who applied restriction fragment length polymorphism analysis for this purpose in the mid-1980s. This major breakthrough was complemented by a second breakthrough technique, polymerase chain reaction, developed by Kary Mullis. He used this in vitro method to amplify a relatively small number of template nucleic acid molecules (derived from biological evidence) into billions of copies for subsequent identification of the DNA donor. These procedures are considered highly reliable and are admissible in all courts in the United States.

The SFS report spells out 13 recommendations that if accomplished would certainly expand funded research and establish the reliability of non-DNA forensic evidence analysis.

- Congress should establish a new National Institute of Forensic Science (NIFS), which would involve itself with research and education, the forensic science disciplines, physical and life sciences, measurements and standards, and testing and evaluation.
- NIFS should establish standard terminology and standards to be used in reporting and testifying about the results of forensic testing.
- NIFS should competitively fund research in forensic science to address the accuracy, reliability, and validity of the subdisciplines.
- NIFS, through congressional funding, should provide funding to remove public forensic labs from the administrative control of law enforcement agencies or prosecutors' offices.
- NIFS should encourage research on observer bias and sources of human error in forensic examinations.
- Congress should fund (through NIFS) the development of tools to advance measurement, validation, reliability, information sharing, and proficiency testing in forensic science and to establish protocols for forensic examinations, methods, and practices.
- Laboratory accreditation and analyst certification should be mandatory.
- Forensic laboratories should establish routine quality control and quality assurance programs to ensure the accuracy of forensic analyses.
- NIFS should establish a national code of ethics for all forensic disciplines.
- Congress should provide funding to NIFS to work with appropriate organizations and educational institutions to improve and develop graduate education programs and to provide scholarships and fellowships with emphasis on developing research methods and methodologies applicable to forensic science practice.
- NIFS should support replacing coroner systems with medical examiner systems to improve medicolegal death investigation.
- Congress should provide funds for NIFS to launch a new effort to achieve nationwide fingerprint data interoperability.
- Congress should provide funding for NIFS to help forensic scientists manage and analyze evidence from events that affect homeland security.

The full SFS report, which describes all of the above, is available through the National Academies Press, 500 Fifth Street N.W., Washington, DC 20001.

Some of the issues faced by forensic experts who write official reports and present courtroom testimony about their analysis are (1) whether or not the techniques of the discipline are founded on a sound scientific foundation, so that the experimental findings are accurate and reliable; (2) whether or not there is any significant possibility of human error or analyst bias that could potentially taint the results; and (3) whether or not rigorous standards have been established for interpreting the results.

The authors of the SFS report believe that with the exception of DNA, no subdiscipline of forensic science is sufficiently reliable that a unique identification can be made of physical evidence: for example, "This bullet was fired from this gun and from no other gun in the world"; or "This fingerprint was left by this person and by no other person"; or "This tire track impression was made by this tire and by no other tire". What is the statistical basis for such a statement? What is the measurement error rate? What is the human error rate? How should conclusions be expressed in a report or in testimony?

With all of this in mind, I invite you to read through the various chapters in this book and keep in mind that the subject matter is about science and technology and not about art. Keep in mind how the expert comes to a conclusion and how that conclusion is reported. What is the statistical basis for the expert's report and testimony? Can the expert testify to a high degree of scientific certainty that the questioned and known specimens have a common origin? I leave that to the reader to decide.

This handbook begins with a review of forensic environmental chemistry which involves the use of trace chemical techniques for investigating environmental spills in an effort to determine if there is any civil or criminal liability. The field can be broken down into two broad areas based on the techniques used to determine liability: chemical fingerprinting and spatial association. In chemical fingerprinting, complex mixtures of chemicals or chemical isotopes are used to associate a spill or environmental release with a specific source. In spatial association, a geographical information system and geochemical techniques are used to attribute the location of a contaminant with a possible source in physical space.

Chapter 2 addresses the principles and issues that exist in the forensic analysis of explosives. It lays out the foundation for proper handling of evidence, which is critical to identifying and convicting the criminal. Evidence at the scene of an explosion, especially a large explosion, offers some unique challenges. Basic principles of evidence collection, handling, storing, and identifying various explosives are discussed herein.

Chapter 3 is a review of arson and fire debris analysis. The isolation and identification of ignitable liquid residues (ILRs) from fire debris is a critically important aspect of arson investigation. This chapter covers common techniques for the isolation and identification of ILRs. Analytical procedures have become more sensitive, and results of testing play a very important role during litigation in a criminal or civil court. Quality control is an important component in fire debris analysis. Reports of findings should be written in a scientific manner describing the fire under investigation, evidence handling, a description of the evidence and where it was collected, the isolation procedure and what testing was done and with what kind of equipment, observations made, and conclusions, with a discussion of the meaning of the results.

Chapter 4 reviews the forensic examination of soils. Soils and sediment are excellent sources of trace evidence in both criminal and civil cases because there are an almost unlimited number of identifiable soil types based on the content of rocks, minerals, glasses, and human-made particles and chemicals. Forensic examination commonly identifies the original geographic location of soils associated with a crime, thus assisting an investigation. Studies of soil and related material samples associated with a suspect and crime scene can produce evidence that the samples had or did not have a common source, thus indicating whether or not a suspect was ever at a particular location. Gathering intelligence for criminal and civil investigations, as well as gem and art fraud studies, often use the methods of forensic geology

Chapter 5 deals with the analysis of paint evidence. Paint and coatings often appear in criminal, civil, and art-authenticity investigations. This chapter reviews the current methodologies and approaches used by forensic paint examiners to analyze this type of physical evidence as well as the problems that they may encounter. Fragments of multilayered in-service paint are one of the most complex types of materials encountered in the forensic science laboratory. They consist of both organic and inorganic components heterogeneously distributed in very small samples, often on the order of only 1 square millimeter. These characteristics dictate the requirements of the analytical chemistry approaches to be used, and they can present a formidable challenge to the forensic analyst responsible for classification of the materials and an evaluation of their evidential significance. Several case examples are presented to illustrate these concepts.

Chapter 6 describes analytical techniques used for the forensic examination of writing and printing inks. The analysis and identification of writing and printing inks and toners are generally very important in document examination, especially when used in conjunction with a reference library. Inks can be differentiated based on the chemistry of colorants, solvents, resins, and additives. Instrumental analysis, including GC–MS, HPLC, and FT–IR and Raman spectroscopy, can often be used following visual examination, microscopic observation, and thin-layer chromatography. Analysis of toners can be performed with XRF, SEM–EDS, or pyrolysis GC. Although chemical analysis of materials used to create documents can provide vast amounts of relevant information and strongly support associations between questioned and known materials, in nearly all cases, the data obtained will not support a conclusion that identifies a particular writing instrument of printing device.

Chapter 7 describes the role of vibrational spectroscopy in forensic chemistry. Spectroscopy is the study of the interaction of electromagnetic radiation with matter to determine the molecular structure of a solid sample or one dissolved in a specific solvent. This interaction depends on the intrinsic properties of the sample material and can be classified by the energy of the probing electromagnetic radiation. Energy can be in the form of ultraviolet, visible, or infrared light, as well as other forms of energy. Infrared spectroscopy is a good technique to use to identify such fibers as acrylics, nylons, or polyesters or paints or alkyds, acrylics, or nitrocellulose. The size of the sample may require the use of microscopic infrared spectroscopy, and the nature of the sample may indicate the use of external reflection spectroscopy or attenuated total reflectance spectroscopy. These techniques are reviewed as well as related methods of sample identification.

Chapter 8 discusses the important science of forensic serology, an important area of modern forensic science. The primary activity of the forensic serologist is the identification of bodily fluids, as these fluid stains are commonly associated with violent criminal cases. Proving the presence of blood, semen, saliva, and so on, can often confirm alleged violent acts.

Chapter 9 reviews the field of forensic DNA analysis. It describes how DNA became a valuable forensic tool in identifying the source of physical evidence left at a crime scene. The use of restriction fragment length polymorphism analysis in the mid-1980s was replaced by the use of the polymerase chain reaction (PCR) method, which is more sensitive, requiring far less high-molecular-weight DNA, uses less hazardous materials, and is faster and more economical. PCR-STR-based genetic profile typing methods have improved in sensitivity over the past 20 years and have become a basic tool in the crime lab. Where nuclear DNA is insufficient to generate a full genetic profile, mitochondrial DNA can sometimes be used to provide identifying information. Also described are low-copy-number procedures and the typing of single-nucleotide polymorphisms within the human genome.

Chapter 10 reviews current and future uses of DNA microarrays in forensic science. DNA microarrays have revolutionized basic research in molecular and cellular biology, biochemistry, and genetics. Through hybridization of labeled probes, this high-throughput technology allows the screening of tens or even hundreds of thousands of data points in a single run. The technology is most advanced with nucleic acids, but protein and antibody microarrays are coming of age as well. Because of the unique ability to screen for large numbers of molecules, such as DNA sequences, simultaneously, the potential utility to forensic investigations is tremendous. Indeed, progress has been made demonstrating that microarrays are powerful tools for use in the forensic laboratory. As the technology matures and associated costs come down, the day that microarray analysis becomes a routine part of the forensic toolkit draws nearer.

Chapter 11 reviews the problem of date-rape drugs such as MDMA, flunitrazepam, and ketamine, with an emphasis on GHB. Recreational, predatory, and lethal doses, metabolism, and diagnostic metabolites are described. Similarities to and differences from the effects and metabolism of ethanol are also discussed. The advantages of field tests to detect date-rape drugs, and limitations of antibodies and advantages of enzymes for field testing, are discussed. The development of a rapid enzymatic test for the detection of GHB is described.

Chapter 12 covers forensic and clinical issues in alcohol analysis. Ethanol, a clear volatile liquid that is soluble in water and has a characteristic taste and odor, is a central nervous system (CNS) depressant and causes most of its effects on the body by depressing brain function. CNS depression is correlated directly with the concentration of alcohol in the blood (BAC). The Estimation of a person's blood alcohol concentration is based on important parameters such as body weight, ethanol concentration of the beverage consumed and number of beverages consumed, and length of time and pattern of the drinking. Because men and women have different body water amounts (men average 68% and women 55%), there are differences between the ethanol concentration achieved in men and women of similar weight for the same amount of alcohol. Various methods are described that can help to determine BAC in the field as well as in the laboratory.

Chapter 13 discusses fundamental issues of postmortem toxicology. The basic principles of forensic postmortem toxicology are presented. This chapter covers the acquisition and usefulness of different specimens, current analytical techniques, and the interpretation of findings. Special problems associated with the interpretation of drug levels include the conditions of the specimens and the effects of postmortem redistribution, postmortem drug changes, pharmacogenomics, drug interactions, and embalming fluid.

Chapter 14 reviews a field of growing importance in forensic science, entomotoxicology: drugs, toxins, and insects. Forensic entomology is gaining widespread acceptance within the forensic sciences as one method of estimating a portion of the postmortem interval by utilizing the time of insect colonization of a body, also known as the period of insect activity. Additionally, insect evidence can be utilized as alternative toxicology samples in cases where no other viable specimens exist. This subfield, known as entomotoxicology, can provide useful qualitative information to investigators as to the presence of drugs in the tissues at the time of larval feeding. The presence of drugs can alter the developmental period of the insects and should always be taken into consideration by the forensic entomologist. The relationship between toxicology and forensic entomology is also examined.

LAWRENCE KOBILINSKY

CONTRIBUTORS

Jason H. Byrd, University of Florida, College of Medicine, Gainesville, Florida

Anthony Carpi, Environmental Toxicology, John Jay College of Criminal Justice, The City University of New York, New York

Donald B. Hoffman, Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Lawrence Kobilinsky, Chairman, Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Ali Koçak, Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Sarah L. Lancaster, Forensic Explosives Laboratory, Defence Science and Technology Laboratory, Fort Halstead, Sevenoaks, Kent, UK

Gerald M. LaPorte, National Institute of Justice, Office of Investigative and Forensic Sciences, Washington, DC

John J. Lentini, Scientific Fire Analysis, LLC, Big Pine Key, Florida

Nathan H. Lents, Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Richard Li, Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Henrietta Margolis Nunno, Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Maurice Marshall, (Formerly) Forensic Explosives Laboratory, Defence Science and Technology Laboratory, Fort Halstead, Sevenoaks, Kent, UK

Raymond C. Murray, University of Montana, Missoula, Montana

Jimmie C. Oxley, University of Rhode Island, Chemistry Department, DHS Center of Excellence for Explosives Detection, Mitigation, and Response, Kingston, Rhode Island

Stanley M. Parsons, Department of Chemistry and Biochemistry, Program in Biomolecular Science and Engineering, Neuroscience Research Institute, University of California, Santa Barbara, California

Michelle R. Peace, Department of Forensic Science. Virginia Commonwealth University, Richmond, Virginia

Scott G. Ryland, Florida Department of Law Enforcement, Orlando, Florida

Andrew J. Schweighardt, John Jay College of Criminal Justice, The City University of New York, New York

Joseph C. Stephens, United States Secret Service, Forensic Services Division, Questioned Document Branch, Instrumental Analysis Section, Washington, DC

Richard Stripp, Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Edward M. Suzuki, Washington State Crime Laboratory, Washington State Patrol, Seattle, Washington

Beth E. Zedeck, Pediatric Nurse Practitioner, New York

Morris S. Zedeck, (Retired) Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

CHAPTER 1

Forensic Environmental Chemistry

ANTHONY CARPI and ANDREW J. SCHWEIGHARDT
John Jay College of Criminal Justice, The City University of New York, New York

Summary Forensic environmental chemistry involves the use of trace chemical techniques for investigating environmental spills in an effort to determine civil or criminal liability. The field can be broken down into two broad areas based on the techniques used to determine liability: chemical fingerprinting and spatial association. In chemical fingerprinting, complex mixtures of chemicals or chemical isotopes are used to associate a spill or environmental release with a source. In spatial association, geographical information systems and geochemical techniques are used to attribute the location of a contaminant with a possible source in physical space.

1.1 Introduction 2
1.2 Chemical fingerprinting 4
  1.2.1 Hydrocarbon mixtures 4
  1.2.2 Polycyclic aromatic hydrocarbons 6
  1.2.3 Biomarkers 11
  1.2.4 Additives 11
  1.2.5 Isotopes 12
  1.2.6 Tracers 13
  1.2.7 Methods of detection 16
  1.2.8 Weathering 18
1.3 Spatial association of environmental incidents 18
References 20

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

1.1 INTRODUCTION

As technology for trace chemical analysis has expanded in recent decades, so has its application to criminal and civil casework. This has transformed traditional forensic investigations and has expanded their applicability to less traditional areas, such as those involving environmental crimes. Prior to 1950, environmental law in the United States was based on tort and property law and was applied to a very limited number of incidents. Driven by growing environmental awareness in the 1950s and 1960s, the U.S. Congress passed the first Clean Air Act in 1963. This was followed by a slow but steady string of further developments, including the founding of the Environmental Protection Agency (EPA) in 1970 and the passage of the Clean Water Act in 1972, the Endangered Species Act in 1973, and the Comprehensive Environmental Response, Compensation, and Liability Act (Superfund) in 1980. International law began to address environmental issues with the signing of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) in 1975 and other international treaties. These early milestones have been bolstered by recent amendments, new agencies, and renewed funding, all of which make up a series of laws and regulations that define criminal practices and govern civil liability cases involving the environment. Increased legislation and improved enforcement have led to a significant decrease in easily identifiable environmental disasters, such as when the Cuyahoga River in Cleveland, Ohio burst into flames in 1969 as a result of industrial discharge. As these visible issues have diminished, environmental scientists have found themselves faced with questions that are more difficult to identify and are more intractable in nature. This has led, in turn, to advances in the investigative techniques used to investigate environmental crimes.

It is impossible to pinpoint the exact birth date of forensic environmental science. However, one source attributes the origin of the term *environmental forensics* to the scientific contractor Battelle in the late 1990s (Haddad, 2004). One of the company's specialties is forensic environmental chemistry, and the company provides services in hydrocarbon fingerprinting, contamination identification, and product identification. Regardless of when the field was named, most sources would agree that the field began gathering momentum about 30 years ago. Since that time, various subdivisions have emerged. Some of these divisions have their roots in diverse areas such as geology, toxicology, biology, physics, and chemistry. As such, the term *environmental forensics* might be considered a misnomer for two reasons. The first is the tendency of the word *forensics* to be semantically confusing, because it has no real meaning when used in this context. The second is the loss of the word *science*, for this serves as a necessary reminder of the field's vast and diverse capabilities, spanning across not just one but many sciences.

The term *environmental forensics* is often misapplied to what should rightfully be called *forensic environmental chemistry*. For example, environmental forensics has been defined as "the systematic investigation of a contaminated site or an event that has impacted the environment," a definition that is clearly biased toward the chemistry perspective (Stout et al., 1998). The broad capabilities of the field are unnecessarily simplified to the question: Who caused the contamination, and when did it occur? (Ram et al., 1999). Surely this is not the only question that environmental forensics is capable of answering. Nevertheless, this mindset has persisted because it is acknowledged and reaffirmed repeatedly. Many of the shortfalls of the earlier definitions of environmental forensics have been identified and amended in subsequent definitions. Many of these revisions offer a more generic, all-inclusive definition. One source defines forensic environmental science simply as "litigation science" (Murphy, 2000); another as "environmental 'detective work' . . . operating at the interface junction points of several main sciences including chemistry and biochemistry, biology, geology and hydrogeology, physics, statistics, and modeling" (Petrisor, 2005). Vives-Rego (2004) defines it not just as the environmental application of chemistry, biology, and geology, but as "science and the art of deduction." Finally, Carpi and Mital (2000) define it as "the scientific investigation of a criminal or civil offense against the environment." These updated definitions more accurately reflect the capabilities of forensic environmental science beyond the chemistry realm. In particular, the definition provided by Carpi and Mital (2000) specifically includes the use of DNA to solve crimes perpetrated against wildlife and plant life. In this chapter we focus on the specific subarea of forensic environmental chemistry and leave to another source the broader description of the methods and techniques that apply to environmental forensics.

However one chooses to define this growing field, one thing is certain: Forensic environmental science is filling the significant niche left void by forensic science and environmental science. Due in large part to its close association with the core sciences, forensic environmental science has experienced significant growth since its inception, especially in recent years. Aside from technological achievements in the past 30 years, several important advances have helped propel forensic environmental science from a burgeoning offshoot of forensic science to a scientific discipline in its own right. One such advancement was the founding of the journal *Environmental Forensics* in 2000 (Taylor & Francis, London). Although research pertaining to forensic environmental science occurred before the journal existed, the journal can be credited with offering a place for environmental research that falls under the forensic science umbrella. Thus, *Environmental Forensics* provides a forum to facilitate the exchange of information, ideas, and investigations unique to forensic environmental science (Wenning and Simmons, 2000).

Forensic environmental science has become such a diverse field that it is difficult to find a single work that adequately covers all its subdisciplines. The literature on the subject that enjoys the most success does so because it focuses on a specific area of forensic environmental science. As such, in this chapter we focus on forensic environmental chemistry. Our aim is to elaborate on several key areas of forensic environmental chemistry, perhaps where other resources have been unable to or have failed to do so. In particular, we focus on chemical fingerprinting and its subsidiaries, such as hydrocarbon fingerprinting, isotope fingerprinting, and complex mixture fingerprinting. Chemical fingerprinting attempts to individualize a chemical and trace it back to its origin. This technique has become increasingly important not only to identify that a chemical spill has indeed occurred, but also to identify the party responsible. We also focus on spatial analysis for the purpose of source attribution. Several cases are discussed that are illustrative of the capabilities of spatial analysis and chemical fingerprinting as they pertain to forensic environmental chemistry.

1.2 CHEMICAL FINGERPRINTING

Chemical fingerprinting is a subsidiary of forensic environmental chemistry that examines the constituents of a mixture for the purpose of creating a unique chemical signature that can be used to attribute the chemicals to their source. At one time it was sufficient to arrive at a generic classification and quantitation of the chemical mixture so that appropriate remediation measures could be designed and implemented. However, modern analytical techniques that are focused on individualizing and associating a mixture with a source have become increasingly popular, both for liability reasons and because of the recognition and attempt to apportion liability when multiple and/or temporally distant parties may be responsible for chemical contamination. The main objectives of chemical fingerprinting are to characterize, quantitate, and individualize a chemical mixture (Alimi et al., 2003). In this section we provide the reader with a review of some of the constituents of a mixture that are useful for assembling a chemical fingerprint as well as the techniques used to screen for these constituents. The efficacy of these analytes and of detection techniques are evaluated by illustrating their application in several cases.

1.2.1 Hydrocarbon Mixtures

The majority of chemical spills involve hydrocarbon mixtures; as a result, many techniques are tailored for these mixtures (Sauer and Uhler, 1994). Early techniques were used simply to quantify the total petroleum hydrocarbon concentration, but modern techniques must be capable of quantification as well as identification and individualization (Zemo et al., 1995). The latter two are especially important for litigation purposes. However, identification and individualization may also provide for the design of a more effective remediation plan that accounts for dispersal, weathering, and degradation of the chemical mixture (Zemo et al., 1995).

Petroleum hydrocarbon mixtures may be broadly classified into three general groups. *Petrogenic hydrocarbons* are present in crude oil or its refined products. *Pyrogenic hydrocarbons* are the combusted remnants of petrogenic hydrocarbons and other by-products. *Biogenic hydrocarbons* are those that arise from more recent natural processes: for example, swamp gas or the volatile hydrocarbon mixtures released by decaying plant or animal tissue exposed to anaerobic conditions. Within each of these three broad groups, hydrocarbons are generally separated into three types: saturated aliphatics (alkanes), unsaturated aliphatics (alkenes, etc.), and aromatic hydrocarbons. Aromatic hydrocarbons include both light petroleum products [e.g., benzene, toluene, ethylbenzene, and xylenes (BTEX)] and heavier products such as polycyclic aromatic hydrocarbons.

Analytical techniques such as gas chromatography are usually adequate for differentiating among petrogenic, pyrogenic, and biogenic hydrocarbon mixtures because of the unique ratios of alkanes, alkenes, and aromatic structures that can be expected in these mixtures. Furthermore, gas chromatography can be used to differentiate different grades of petrogenic hydrocarbons because crude mixtures have a variety of hydrocarbon components (i.e., unresolved complex mixtures), which often present themselves as a "hump" on a chromatogram, whereas more refined mixtures have less variety in their components. Retention time for various petrogenic compounds is affected by the structure of these compounds: for example, gasoline elutes first (C_4 to C_{12}), along with Stoddard solvents (C_7 to C_{12}), which are followed by middle distillate fuels (C_{10} to C_{24}), and crude mixtures (up to C_{40}) (Zemo et al., 1995). Crude oil mixtures contain a diverse array of hydrocarbons and, on average, are comprised of 15 to 60% paraffins and isoparaffins, 30 to 60% naphthenes, and 3 to 30% aromatics, with the remainder of the mixture being composed of asphaltenes and various trace compounds (Bruce and Schmidt, 1994). Pyrogenic hydrocarbon mixtures can be recognized on the chromatogram because large molecules undergo combustion first, leaving behind a disproportion of smaller molecules. However, pyrogenic compounds are more difficult to attribute to a source because the chemical signature is further removed from the original petrogenic source (Bruce and Schmidt, 1994). Steranes and hopanes are often used as target analytes when the focus of a study is biogenic hydrocarbons, because these analytes are more resistant to many more forms of weathering than are other biogenic components (Alimi et al., 2003).

Although it is worthwhile to classify a mixture as petrogenic, pyrogenic, or biogenic in origin, this is commonly not enough. To arrive at a unique chemical signature, the analysis must extend beyond identifying the class characteristics of a mixture. Modern methods often involve the examination of ancillary components of a mixture, such as dyes, additives, stable isotopes, radioactive isotopes, biomarkers, polycyclic aromatic hydrocarbons (PAHs). PAH homologs, and metabolized PAHs. It is customary to screen for many of these analytes with the intent of providing the most comprehensive chemical signature possible. Before selecting a suite of analytes, it is wise first to consider if these analytes may already have been present at a location (due to a prior contamination or natural processes) and if these analytes are highly susceptible to degradation. Indeed, the characteristics of a good chemical marker are that it is resistant to degradation and that it can uniquely identify the hydrocarbons released from other sources (Sauer and Uhler, 1994).

The first step in confirming hydrocarbon contamination is accomplished by screening for saturated hydrocarbon molecules such as pristane and phytane. These are isoparaffins that are resistant to degradation and are highly indicative of hydrocarbon contamination (Sauer and Uhler, 1994). Pristane and phytane usually represent themselves to the right of C_{17} and C_{18} peaks on a chromatogram. Fresh hydrocarbon mixtures have prominent C_{17} and C_{18} peaks in relation to pristane and phytane peaks, whereas the converse is true for degraded mixtures (Bruce and Schmidt, 1994; Morrison, 2000b). Due to the proportionality between the ratios of these compounds and the extent of degradation, the ratios of pristane and phytane to the C_{17} and C_{18} peaks are often used to estimate the degree of weathering.

1.2.2 Polycyclic Aromatic Hydrocarbons

Polycyclic aromatic hydrocarbons (PAHs) are hydrocarbon compounds with two to six rings. Homologs of the PAH compounds may be similar to the parent compound except that they are substituted for by one or more alkyl groups. The ratio of two PAHs to two other PAHs is sometimes expressed in double-ratio plots, in which certain regions of the plot are diagnostic for one source or another. It is also becoming increasingly common to screen for metabolized PAHs (as well as BTEX compounds), whose structures differ predictably from the original PAH. PAHs are often very useful in studies involving weathered mixtures, because the complex structure of PAHs makes them more resistant to degradation. The rate of degradation is proportional to the complexity of the ring structure, with the compounds having the fewest number of rings degrading first (Alimi et al., 2003). Some target parent PAH compounds and their alkyl homologs are shown in Table 1.1.

One of the most prominent applications of PAH analysis has been the study of the *Exxon Valdez* oil spill. The spill occurred when the tanker hull was punctured as it ran aground on March 24, 1989, releasing some 10.8 million gallons of oil into Prince William Sound, Alaska. The oil released was dispersed by water currents and a windstorm that followed the spill a few days later, and concern

TABLE 1.1 Target Parent PAH Compounds and Their Alkyl Homologs

Naphthalenes	Chrysenes
C_0-naphthalene (N)	C_0-chrysene (C)
C_1-naphthalenes (N1)	C_1-chrysenes (C1)
C_2-naphthalenes (N2)	C_2-chrysenes (C2)
C_3-naphthalenes (N3)	C_3-chrysenes (C3)
C_4-naphthalenes (N4)	C_4-chrysenes (C4)
Phenanthrenes	EPA priority pollutant
C_0-phenanthrene (P)	Biphenyl (Bph)
C_1-phenanthrenes (P1)	Acenaphthylene (Acl)
C_2-phenanthrenes (P2)	Acenaphthiene (Ace)
C_3-phenanthrenes (P3)	Anthracene (An)
C_4-phenanthrenes (P4)	Fluoranthene (Fl)
Dibenzothiophenes	Pyrene (Py)
C_0-dibenzothiophene (D)	Benzo[*a*]anthracene (BaA)
C_1-dibenzothiophenes (D1)	Benzo[*b*]fluoranthene (BbF)
C_2-dibenzothiophenes (D2)	Benzo[*k*]fluoranthene (BkF)
C_3-dibenzothiophenes (D3)	Benzo[*e*]pyrene (BeP)
Fluorenes	Benzo[*a*]pyrene (BaP)
C_0-fluorene (F)	Perylene (Pe)
C_1-fluorenes (F1)	Indeno[1,2,3-*cd*]pyrene (IP)
C_2-fluorenes (F2)	Dibenz[*a*,*h*]anthracene (DA)
C_3-fluorenes (F3)	Benzo[*ghi*]perylene (BP)

Source: Alimi et al. (2003).

was raised over the dispersal of the oil into adjacent bodies of water (Galt et al., 1991). Many of these concerns were seemingly corroborated by the detection of oil in neighboring bays. However, it was speculated that some of the oil detected in these neighboring waters may have been from biogenic sources, petrogenic sources from previous spills, or pyrogenic sources from hydrocarbons that had previously undergone combustion. A massive effort was mounted to identify the extent to which Exxon was responsible for oil detected in these adjacent waters.

The study immediately focused on components of oil that were the most resistant to degradation, such as PAHs and biomarkers. A substantial part of the investigation focused on evaluating the effects of weathering on the *Exxon Valdez* cargo if it was to be accurately differentiated from other sources (Figure 1.1). As expected, lighter components of the oil matrix were preferentially lost to weathering. With the effects of weathering understood, the investigation then turned to PAH analysis. Two PAHs that were focused on for distinguishing different crude mixtures were phenanthrenes and dibenzothiophenes; chrysenes were used to differentiate crude from refined mixtures because chrysenes are removed during the refining process (Boehm et al., 1997). The ratios of the PAH compounds to one another were particularly useful

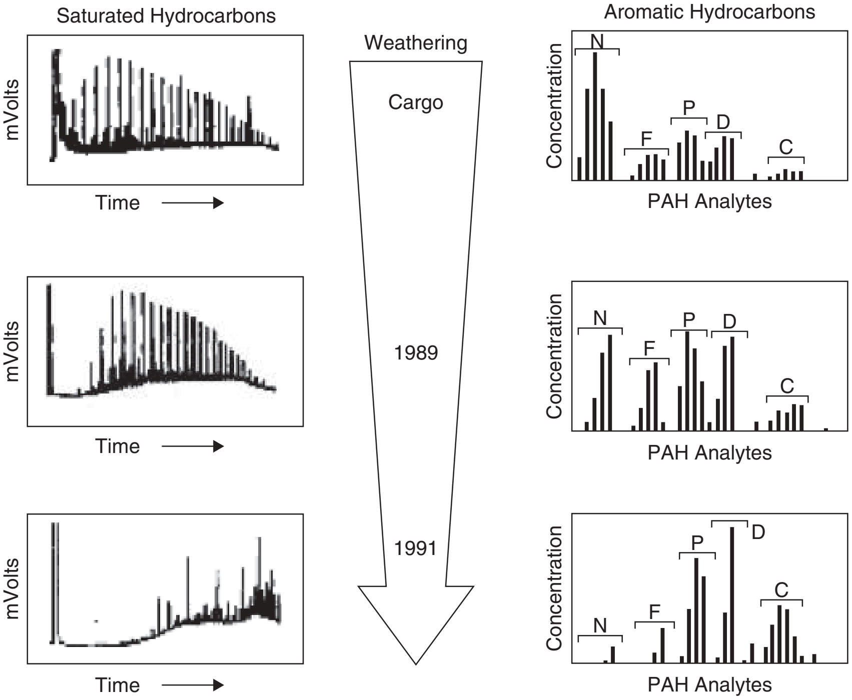

Figure 1.1 Effects of weathering on (a) saturated hydrocarbons and (b) aromatic hydrocarbons from the *Exxon Valdez* spill. N, naphthalenes; F, fluorenes; P, phenanthrenes; D, dibenzothiophenes; C, chrysenes (Boehm et al., 1997).

because the concentrations of the PAHs will change with weathering; however, the ratio of one PAH to another generally remains constant (Boehm et al., 1997). In this case, researchers created a double-ratio plot comparing dibenzothiophenes to phenanthrenes in order to distinguish PAHs of the *Exxon Valdez* spill from PAHs of other sources (Figure 1.2). As seen in the figure, the double-ratio plot showed distinct clustering of oil samples from different sources, allowing a differentiation to be made. When the PAHs in neighboring bays were analyzed, some were attributed to the *Exxon Valdez* spill, but many were found to have originated from other sources, both natural and anthropogenic (Boehm et al., 1998).

PAHs have also been used to study contamination at former manufactured gas plant (MGP) facilities. Prior to the use of natural gas, MGPs made coal gas to use as fuel. Former MGP sites are evaluated for contamination by screening for PAHs that would have been introduced to the environment as coal tar, which is a by-product of the coal gas manufacturing process. This can sometimes be a difficult task because the sites often contain PAHs that may be unrelated to the MGP, having been introduced via other natural and anthropogenic avenues. The investigations are further complicated because similar PAH signatures are obtained for MGP coal tar residues and background residues. Although the composition of the PAHs contained in MGPs and background sources may be similar, the PAH ratios and patterns (i.e., petrogenic or pyrogenic) can be used to differentiate PAHs from different sources. One study examined the ratios and patterns of PAHs for the purpose of distinguishing MGP PAHs from background PAHs in soil samples collected in and around a stream near an MGP (Costa et al., 2004).

PAHs may be present either as unsubstituted parent compounds or as a substituted alkyl homolog (see Table 1.1 for examples). Petrogenic patterns of PAHs are recognized because they contain a bell-shaped distribution of the parent PAH and its homologs where concentration of the single- or double-substituted homologs

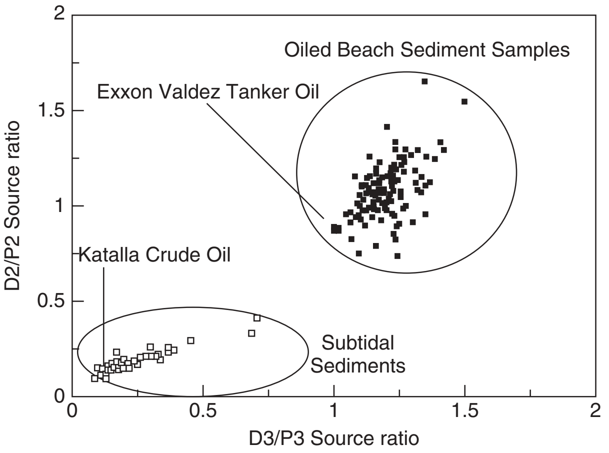

Figure 1.2 Double-ratio plot showing how the ratio of PAHs (dibenzothiophenes to phenanthrenes) can be diagnostic for one source or another (Boehm et al., 1997).

are highest, and concentrations decrease as one moves in either direction toward the unsubstituted parent or toward the complex, multisubstituted homolog (see Figure 1.4 for an example). Pyrogenic patterns of PAHs are recognized because they contain a distribution in which the parent PAH is more abundant, due to preferential combustion of the substituted homologs. Researchers observed a pyrogenic pattern in PAH residues derived from the MGP site in question (Figure 1.3), but samples collected from an adjacent stream indicated a mix of petrogenic and pyrogenic PAHs (Figure 1.4).

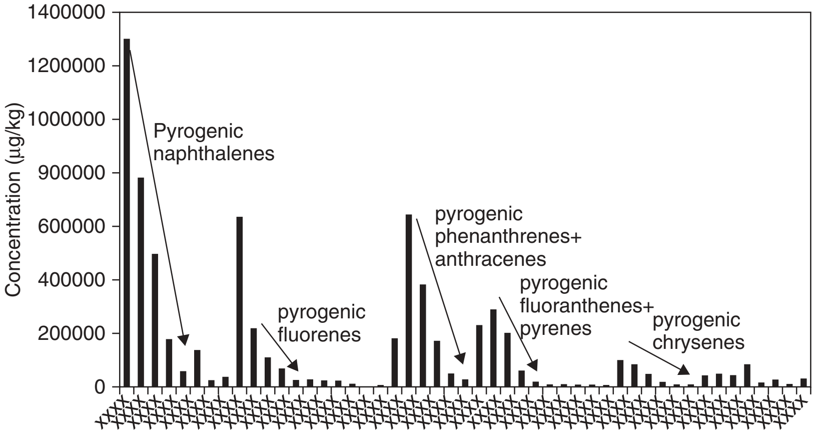

Figure 1.3 PAH composition of residues derived from an MGP site. Decreasing concentrations of substituted homologs of the parent compounds indicate the pyrogenic origin of the sample (Costa et al., 2004).

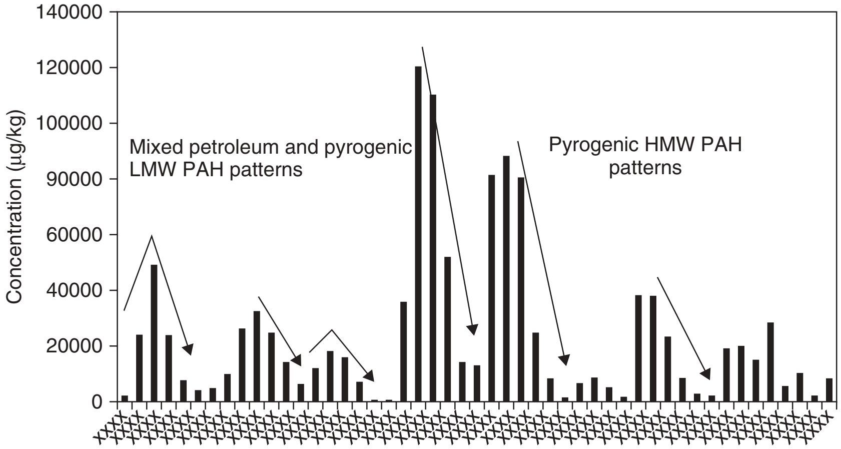

Figure 1.4 PAH composition of residues derived from a streambed. Mixed patterns indicate the presence of a mixture of pyrogenic and petrogenic hydrocarbons (Costa et al., 2004).

The researchers then turned to high-molecular-weight PAH ratios to determine if the pyrogenic pattern observed in the stream was from weathered MGP residues or from recent background contamination. Several PAHs were chosen to create double-ratio plots in which certain sections of the plot were diagnostic for either the MGP, background sources, or a mix of the two. A comparison of samples from the streambed surface (Figure 1.5) and samples from the streambed subsurface (Figure 1.6) indicated that most of the surface (i.e., newer) PAHs were derived

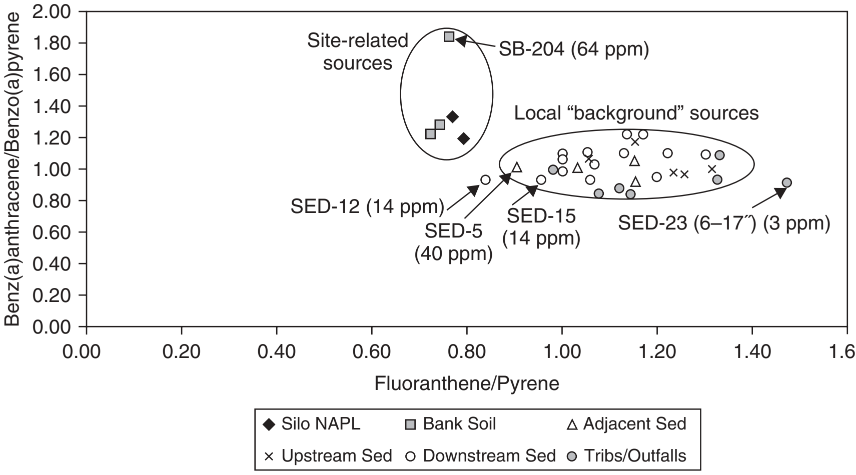

Figure 1.5 Double-ratio plot used to distinguish site-related, background, and mixed PAH signatures in streambed surface samples (Costa et al., 2004).

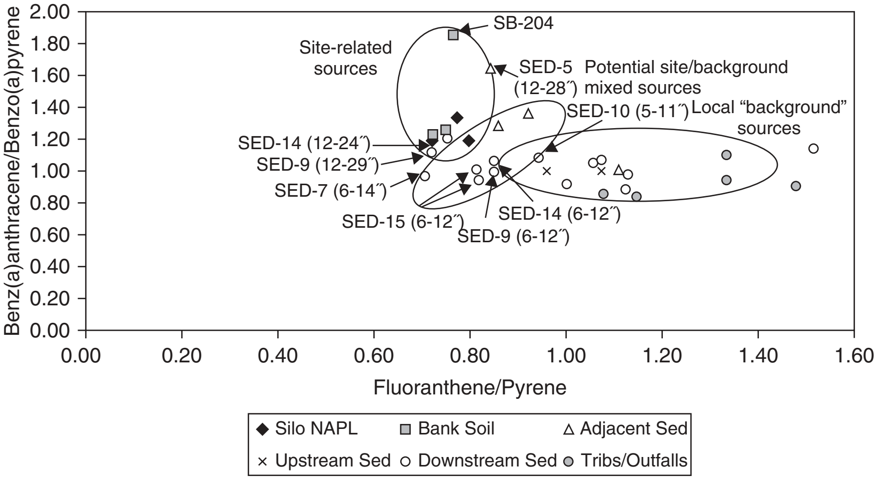

Figure 1.6 Double-ratio plot used to distinguish site-related, background, and mixed PAH signatures in streambed subsurface samples (Costa et al., 2004).

from background sources, whereas most of the subsurface (i.e., older) PAHs were derived from the MGP site. Results of studies such as this can help to draw attention to other potential sources of contamination in order achieve the most efficacious remediation effort.

1.2.3 Biomarkers

Biomarkers such as steranes and hopanes are hydrocarbon remnants of deceased organisms that are useful in chemical fingerprinting because they are extremely resistant to weathering (Alimi et al., 2003). Thus, biomarkers can often be useful to individualize a hydrocarbon mixture when saturated hydrocarbons and PAHs have already been degraded (Sauer and Uhler, 1994). A comprehensive list of biomarkers that are useful in hydrocarbon mixture studies is provided by Alimi et al. (2003).

1.2.4 Additives

Inorganic compounds are often added to hydrocarbon mixtures to serve as anti-knock agents, octane boosters, corrosion inhibitors, and anti-icers (Kaplan, 2003). Additives are not present in crude mixtures, of course, so their presence is indicative of a refined mixture (Bruce and Schmidt, 1994). Because refining practices and additives change over time, and since these changes have been well documented, the presence of particular additives in a hydrocarbon mixture can be highly indicative of a certain time frame during which a sample was produced. For example, lead was first added to gasoline in 1923, and its concentration in gasoline decreased steadily until it was phased out in U.S. automobile fuels in 1995 (Kaplan, 2003). Other gasoline additives that have predictably appeared and disappeared throughout history are methylcyclopentadienyl manganese tricarbonyl (MMT) and methyl *tert*-butyl ether (MTBE). The chronology of some popular additives has been thoroughly documented in several sources (Morrison, 2000a,b; Kaplan, 2003). Some additives, such as lead, have been used over large time frames, but the concentration of lead in gasoline has varied predictably over the years. Although this can be used to arrive at a reasonable estimate of time of manufacture of the hydrocarbon mixture in question, it is not an infallible method because additive concentrations are often reported based on a pooled standard, which ignores batch-to-batch variation (Morrison, 2000a).

The use of additives for dating a release can be complicated by the fact that additives may be discontinued in certain countries or for certain applications, but may still be used in others. Additives that are supposedly absent in a mixture may also be present in very dilute amounts. The utility of additives in dating manufacture or release dates is greater than their capacity to individualize a mixture. This is because many companies often purchase additives from the same manufacturer. These additives are then added unaltered into various hydrocarbon mixtures, so many different mixtures may have the same additives present (Morrison, 2000a). Further complications when screening for old additives may be encountered because these compounds often contain oxygen, which contributes to their rapid weathering over time (Morrison, 2000b).

1.2.5 Isotopes

When complex hydrocarbon mixtures cannot be identified by analysis of stable components, the mixtures can be identified by analysis of stable isotopes within the mixture. Stable isotopes are often analyzed with respect to one another. In other words, the ratio of one stable isotope to another stable isotope within the same mixture can often be unique, thereby allowing for the creation of an isotope signature. In contrast to stable isotopes, unstable isotopes decay predictably such that the degree of decay can be correlated with the age of the mixture. Analysis of unstable isotope decay is often referred to as a *long-term method* because it is capable of estimating release dates thousands of years prior. Unstable isotopes are also useful because their decay is independent of environmental factors such as weathering (Kaplan, 2003).

Isotopes can be useful in chemical fingerprinting in two ways. The ratio of two isotopes can be compared as a means of individualization because no two mixtures will have exactly the same ratio of two isotopes. Carbon and lead isotope ratios are commonly used for source identification. Radioactive isotopes are also useful for dating a release because these isotopes have known rates of decay that are independent of environmental conditions.

Carbon isotopes were used in one study to determine the origin of soil gas methane near the site of a prior gasoline spill (Lundegard et al., 2000). The investigation was triggered by the detection of high methane levels near a service station where approximately 80,000 gallons of gasoline had been spilled 20 years earlier. Initially, it was speculated that the methane was due to the bacterial degradation of the gasoline, but the investigators were considering other possibilities. Suspicion was raised because high levels of methane were detected outside the original gasoline plume, and in some cases the levels detected outside the plume were higher than those within the plume (Figure 1.7).

The initial hypothesis of methane generation by bacterial degradation of the gasoline was also challenged because this is not a common degradation pathway. For gasoline to be fermented to methane, it would first have to be converted to the necessary precursor compounds for methanogenesis by fermentation (Lundegard et al., 2000). Although the generation of methane via this pathway is possible, the investigators were considering more plausible origins of the methane that, coincidentally, were unrelated to the gasoline spill. One of the potential origins considered was the biodegradation of organic matter.

The methanogenesis of petrogenic compounds can be distinguished from that of organic compounds from biogenic origins through the use of ^{13}C, which is a stable carbon isotope. Differentiating the methanogenesis of petrogenic and organic compounds is accomplished based on the idea that older, petrogenic compounds have lower quantities of ^{13}C isotopes than does newer organic matter. The process of methanogenesis significantly reduces the amount of ^{13}C present in the original

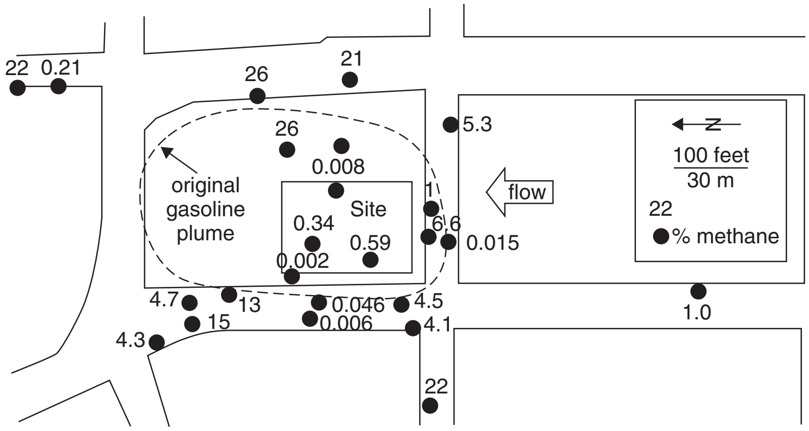

Figure 1.7 Service station map showing methane concentrations within and surrounding the original plume (Lundegard et al., 2000).

organic matter, but the ^{13}C in the nascent methane remains stable regardless of environmental conditions (Lundegard et al., 2000). The study indicated that wood fill from beneath the service station site and gasoline from within the original plume had indistinguishable ^{13}C quantities.

Another way of differentiating methane from petrogenic and biogenic sources is through the use of ^{14}C, which is a naturally occurring radioactive isotope of carbon taken up by all living organisms. The age of the source from which the methane was formed can be predicted because ^{14}C has a half-life of about 5700 years, and therefore it will still be detectable in methane formed from biogenic organic matter less than 50,000 years old. The hypothesis that methane originated from the degradation of biogenic organic matter was corroborated by the ^{14}C analysis, which indicated that the highest ^{14}C levels were detected outside the original plume (Figure 1.8). The level of ^{14}C in petrogenic hydrocarbons is zero, so the researchers concluded that the methanogenesis must be of biogenic origins. This hypothesis was further supported because a review of the site history indicated that the area consisted of organic fill, including wood and sawdust.

1.2.6 Tracers

When none of the analytes discussed previously are amenable to the case at hand, techniques that rely on tracers can sometimes be used for forensic tracking of environmental chemicals. A tracer can be any molecule that is diagnostic of one source but not others. Sometimes, multiple tracers are used to augment the significance of the results. One study used organic tracers to determine the origin of gas- and particle-phase air pollutants in two California cities (Schauer and Cass, 2000). The objective of the study was to determine the primary source(s) of air pollutants in Fresno and Bakersfield, California. The results and chemical composition of samples collected at the two locations were compared to those collected at a remote

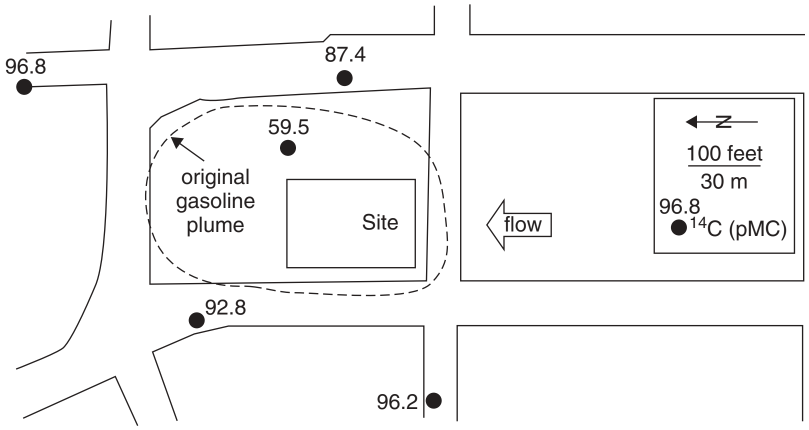

Figure 1.8 ^{14}C levels detected within and surrounding the original plume of a gasoline spill exhibiting high methane concentrations. High ^{14}C concentrations indicated methanogenesis of biogenic hydrocarbons and countraindicated gasoline as the source of origin (Lundegard et al., 2000).

site located at the Kern Wildlife Refuge and distant from anthropogenic sources of air pollutants. Previous tests for air pollutants have used generic compounds to draw connections between air pollutants and their sources (Harley et al., 1992). However, some of the analytes used in these other studies are not exclusive to a particular source. The researchers in the California study aimed to develop a more accurate method for tracing the origin of the air pollutants. Atmospheric samples in the two cities were collected, as well as single-source control samples that consisted of combustion emissions from gasoline-powered motor vehicles, diesel engines, hardwood combustion, softwood combustion, and meat-cooking operations. Tracers that were unique to sources and those that were common between multiple sources were chosen both to fingerprint and then apportion emissions in particulate samples with mixed origins. Further criteria used to choose tracers were (1) that they were not selectively removed from the environment, and (2) that they were not formed by atmospheric reactions to any significant extent. The researchers used direct measurements of these tracer compounds to draw conclusions about the source of particulate pollutants in the areas indicated.

Specific tracers were used to apportion the results obtained from specific sources. For example, the compound levoglucosan was found to be specific to wood combustion, so the concentration of levoglucosan in proportion to other constituents in a mixed sample could be used to apportion the contribution of wood combustion to particulate loading in an area (see Figure 1.9). Based on the concentrations of the other tracers in the samples, the relative contributions of each source (e.g., automobiles, wood combustion) were apportioned (Figure 1.10). Low levels of pollutants derived from anthropogenic sources at the remote site were noteworthy. The researchers concluded that local anthropogenic emissions (particularly automobile exhaust) were responsible for the majority of air pollutants in the two urban

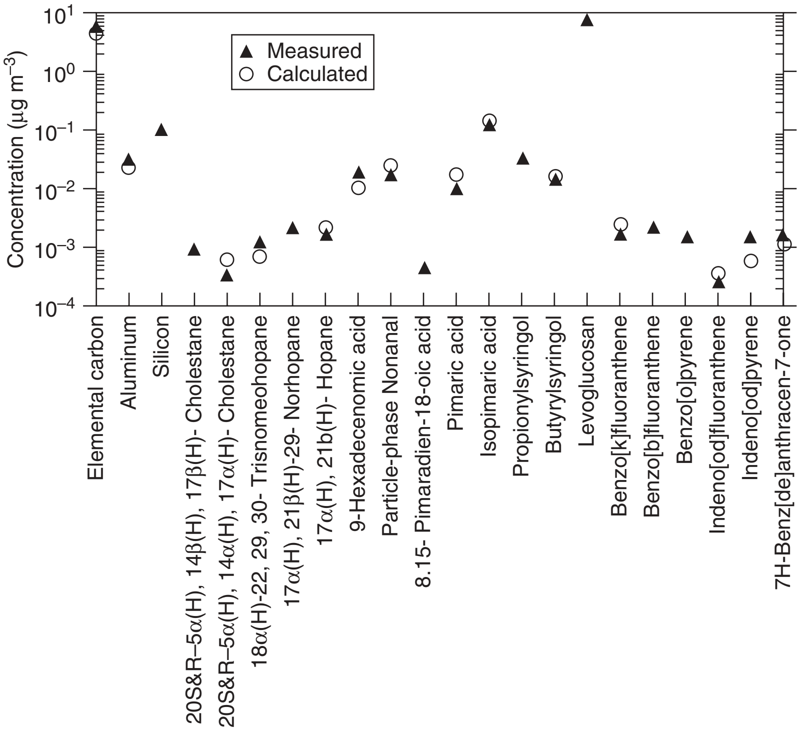

Figure 1.9 Ambient concentrations of various air pollution tracer compounds in samples (Schauer and Cass, 2000).

environments, whereas naturally occurring dusts were primary contributors at the remote site.

Inorganic compounds can also be used for forensic chemistry purposes. One study used metal tracers to identify dust that resulted from the collapse of the World Trade Center (WTC) buildings (Scott et al., 2007). The analysis of particulate matter arising from this catastrophic event has been an area of great interest because there are significant health implications associated with inhalation of the dust. Substantial amounts of the dust were transferred from the collapse to nearby buildings, so the objective of the research was to develop a method based on metal tracer detection to determine which buildings were most affected, and for those that were severely affected, to determine if appropriate remediation efforts had been undertaken.

Techniques employed included screening for human-made vitreous fibers, as well as trace metals, including As, Cd, Cr, Cu, Pb, Mn, Ni, V, and Zn. Trace metal detection was found to be more applicable to the identification of WTC dust because the atmosphere and buildings around the WTC were probed routinely for these metals after the collapse (Scott et al., 2007). Although these metals can originate from other sources as well, the researchers expected trace metals to be detected in quantities and ratios that were unique to WTC dust.

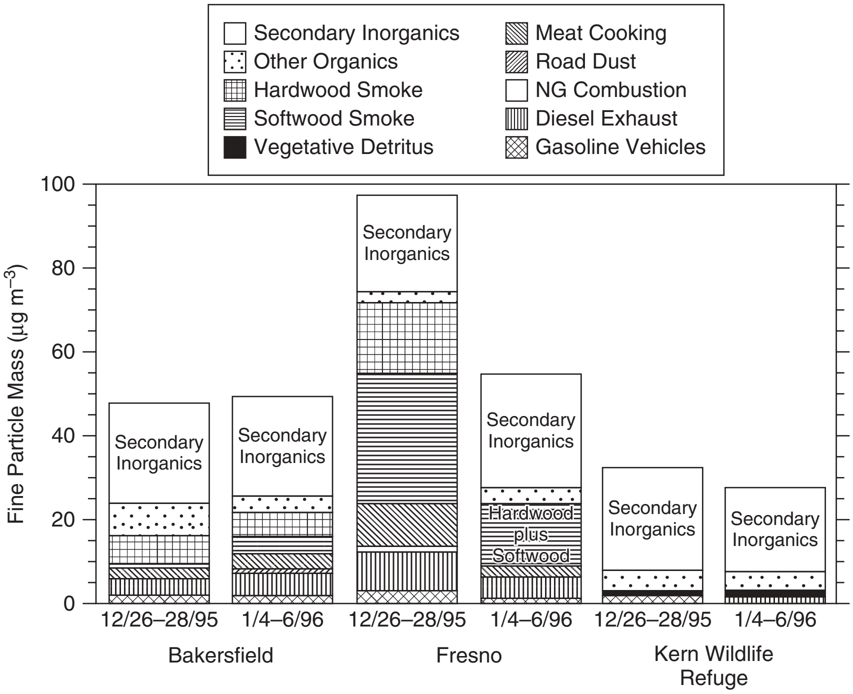

Figure 1.10 Source apportionment to particulate air pollution in two urban areas and one remote site in California (Schauer and Cass, 2000).

Concentrations of the nine metals in WTC dust as reported by four studies were compared to concentrations in background dust collected from Arizona. A discriminant analysis model was used to classify each sample as having originated from WTC or background dust based on the relative concentrations of the nine metals (Figure 1.11). The analysis indicated that WTC dust had elevated levels of Cr and Mn and low levels of As, Cd, and Cu compared to background dust. The researchers were able to demonstrate that trace metals could be used to distinguish pure WTC dust from background dust with 94% accuracy; however, mixed dust samples had lower levels of accuracy (Scott et al., 2007).

1.2.7 Methods of Detection

One of the most widely used techniques in chemical fingerprinting for hydrocarbons is gas chromatography. This is based on the concept that each compound has a unique structure and will therefore be retained differentially in the gas chromatograph before being eluted. As long as other parameters (e.g., temperature, column length, column packing) are held constant, any differences in retention time can be attributed to the structure of the compound (Bruce and Schmidt, 1994). Mixtures contain many different compounds, so a gas chromatogram represents a chemical fingerprint of all the chemical constituents in a mixture. Gas chromatography

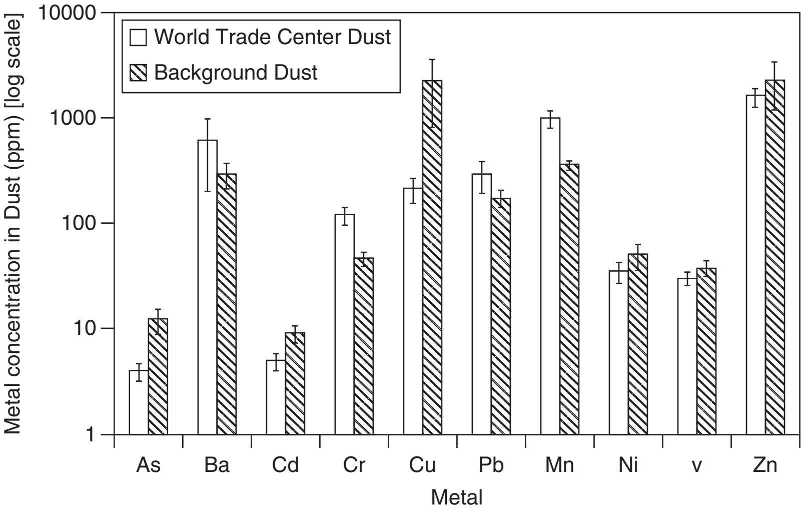

Figure 1.11 Comparison of trace metal fingerprints in dust from the WTC (white bars) and from background (striped) (Scott et al., 2007).

is often combined with other techniques to achieve a more detailed analysis. For example, a gas chromatograph is commonly used as a preliminary separation technique that is followed by detection using mass spectroscopy. Some researchers even use two-dimensional gas chromatography (GC × GC) to achieve superior resolution (Gaines et al., 1999). A good review of the literature focusing on these techniques is provided by Suggs et al. (2002). The potential weaknesses and vulnerabilities of these techniques are discussed by Morrison (2000b).

Although many of the aforementioned techniques are highly effective, they often have a deleterious impact on the sample. That is, substantial portions of the sample are often destroyed in the course of the analysis. Sample destruction may not be a major concern in other disciplines, but evidence is sometimes limited in forensic investigations, and what little sample may be available often attains the status of a precious and rare commodity. Another example of when the destruction of a sample is avoided is when the sample itself is, quite literally, a rare commodity, such as an archaeological treasure or artifact. For samples of limited quantity or prized value, less invasive methods of analysis are often sought.

One study that warranted the use of a minimally invasive technique involved the analysis of ancient tools made of obsidian (Tykot, 2002). The purpose of the analysis was to evaluate the Mediterranean sources and trade routes of obsidian tools without damaging them. To bolster the results and to compensate for the potential weaknesses of certain techniques, this study relied on a series of methods, including scanning electron microscopy (SEM), x-ray fluorescence (XRF), neutron activation analysis (NAA), and inductively coupled plasma mass spectroscopy (ICP-MS). The elemental compositions indicated by the four techniques were used to construct possible sources and distributions of obsidian. The results helped lend credence to the theory of a vast distribution network for the tools rather than the lone source theory that was once promulgated.

1.2.8 Weathering

When screening for various analytes, one variable that must be kept in mind is weathering. This is the process by which the chemical signature of a mixture is altered due to evaporation, dispersal, biodegradation, or oxidation of certain components of the mixture. Short-chain hydrocarbons are most vulnerable to weathering mainly because their simple structure makes them susceptible to degradation, particularly to biodegradation (Alimi et al., 2003). Compounds with more complex structures are generally more resistant to weathering. Weathering generally occurs at predictable rates such that the age of the mixture can be estimated accurately based on the relative amount of weathering of short-chain hydrocarbons to larger, more resistant molecules. However, weathering can vary because of site-specific differences in environmental conditions (Morrison, 2000b). The compounds that exhibit the greatest longevity are generally the most useful for estimating the degree of weathering and therefore the age of a particular compound. Isotopes, BTEX compounds, PAHs, and biomarkers have all been used with varying degrees of success for determining the extent of weathering of hydrocarbon samples.

Certain studies have focused on families of constituents of oil that are resistant to degradation, specifically paraffins, isoparaffins, aromatics, naphthenes, and olefins, all of which generally range from three to 13 carbons (Kaplan et al., 1997). These compounds, commonly referred to as PIANO compounds, are useful because their ratios vary among different hydrocarbon mixtures. For example, fuels of different grades and octane levels are composed of unique ratios of PIANO compounds. PIANO analysis is particularly useful to spatial analysis because the concentration of the various PIANO constituents and additives has changed over time, due to evolving oil refining practices, varying octane levels, and increasingly stringent legal regulations (Davis et al., 2005).

1.3 SPATIAL ASSOCIATION OF ENVIRONMENTAL INCIDENTS

Spatial analysis is used to associate a pollutant release or plume with a source by tracing back its geographic point of origin to a particular place in space or time. This approach is often thought of as one of the more esoteric areas of forensic environmental chemistry; however, this reputation is, for the most part, undeserved because spatial analysis is inherently simple and straightforward. Spatial analysis applies all of the usual tools of forensic environmental chemistry to a spatial problem. For example, spatial analysis relies on many of the analyses with which we are familiar, such as those that screen for various hydrocarbons, fuel additives, isotopes, and biomarkers. However, spatial analysis attempts to determine more than the source and content of a particular contamination. Spatial analysis goes further by elucidating not only the “what” and “who” of a chemical contamination incident, but also the “where” and “for how long.”

Spatial analysis of chemical transport involves integrating the results of chemical analyses with spatial and sometimes historical information about a site. For example, Carpi et al. (1994) examined the spatial distribution of airborne mercury pollution around a municipal solid-waste incinerator. The study used transplanted and prepared samples of sphagnum moss as biological monitors of air pollutants. Clean moss samples were distributed to 16 sites within a 5-km (3.1-mi) radius of the waste incinerator, plus one remote site about 20 km (12.4 mi) away. Samples of the moss at each of these stations were collected in duplicate every 2 weeks for about three months. The samples were analyzed by cold vapor atomic absorption spectroscopy for mercury contamination in two ways: Each sample was split and half was analyzed as received and the other half was first oven-dried at 105°C for 24 h before analysis.

Higher concentrations of mercury were correlated with sites closest to the incinerator, which then led the researchers to use meteorological data from the nearest weather service station [38 km (23.6 mi) away] to determine if mercury near the incinerator could indeed be traced back to the incinerator. It was determined that plants from sites with the highest levels of mercury were downwind of the incinerator. A locally weighted spatial statistics technique called *kriging* was used to develop regression surfaces for the pollutants over the area and these regression surfaces showed that proximity to the incinerator accounted for a high degree of the variability in mercury concentrations with location (Opsomer et al., 1995). The study benefited by incorporating topographical data to support the conclusions.

Interestingly, the mercury concentration in moss from the remote site was approximately equivalent to the mercury detected in some of the samples within the 5-km radius of the incinerator. However, comparison of the undried to dried moss samples demonstrated that the volatility of mercury at the remote site, and thus chemical species of mercury accumulating in samples at the remote site, was significantly different than it was near the incinerator. Samples collected near the incinerator demonstrated relatively low volatility, and indeed incinerators are known to emit high levels of $HgCl_2$, which has a low volatility (Carpi, 1997). In contrast, mercury at the remote site demonstrated high volatility, indicating that the form of mercury in the samples was primarily volatile elemental mercury. The authors conducted a site history at the remote site that revealed that whereas the site was distant from any anthropogenic source of mercury, it was close to a recently flooded reservoir system, and flooding of land is known to release naturally occurring elemental mercury from soil.

Spatial analysis can sometimes be complicated if samples have been collected sporadically or randomly, or if the data are otherwise incomplete. Such was the case in a study that investigated unusually high radiocesium (^{137}Cs) levels in a river basin near the Chernobyl nuclear power plant (Burrough et al., 1999). An explosion occurred at the power plant in April 1986 and released radioactive materials to large areas surrounding the plant. After the explosion, ^{137}Cs levels in the contaminated areas generally decreased due to radioactive decay and various environmental factors that resulted in dispersal and dilution. However, some locations near a river basin exhibited high or increasing levels of ^{137}Cs after the explosion.

The research attempted to find a correlation between hydrological events (e.g., flooding) and spatial and temporal variations in ^{137}Cs contamination by using a method of statistical analysis involving *geographical information systems* (GISs) (Burrough, 2001).

GIS was used to construct various maps showing soil types, land cover, and proximity of the flooded areas to main rivers. Using the maps to inspect the spatial distribution of ^{137}Cs over the contaminated area was complicated because samples were collected only in 1988, 1993, and 1994. The sparsely collected samples necessitated data interpolation, in which new data points were created to augment existing data points. However, interpolation of GIS data can sometimes be unrealistic unless the propagation of errors is understood through the use of geostatistics. Ultimately, the use of a GIS and geostatistics in this study helped to establish that there was a relationship between flood events in the river basin and high concentrations of ^{137}Cs.

GISs have been used for many years, but it has only recently been suggested that a GIS be combined with geostatistics, for the precise reasons illustrated in the case above. The GIS technique was developed to automate the mapmaking process by aiding in storage, retrieval, analysis, and display of spatial data (Burrough, 2001). The flaw in the GIS method is that it analyzes the attributes of an object or surface but does not consider spatial variation. Geostatistics has been proposed as the perfect complement to GISs because geostatistics is more realistic, in that it considers chance, uncertainty, and incompleteness in a data set (Burrough, 2001). Unfortunately, the fundamental differences between the two fields, combined with the fact that GISs were not designed with geostatistics in mind, have often caused some recalcitrance when the combination of GISs and geostatistics is suggested (Wise et al., 2001). The benefits of using both methods have only begun to be appreciated, mostly because the advantages can no longer be ignored. Geostatistics stands to profit from the union primarily because a visual interpretation of the data is made available. GISs will benefit because they often rely on data sets that are incomplete and uncertain, and geostatistics offers a method for interpolating such data and understanding the error (Burrough, 2001).

Spatial analysis is a blend of many techniques that are not necessarily exclusive to spatial analysis. At times, its boundaries may even seem amorphous. Spatial analysis may be regarded as superior to other disciplines that share these techniques because only spatial analysis attempts to discern both the spatial and temporal extent of a contamination. Other disciplines merely apply these techniques to determine the identity of the contaminant. Spatial analysis is likely to be relied on increasingly in the future as a supplement to more conventional techniques when there is a need to identify the source and dispersal pattern of widespread contamination.

REFERENCES

Alimi, H., Ertel, T., et al. (2003). Fingerprinting of hydrocarbon fuel contaminants: literature review. *Environ. Forensics*, **4**(1):25–38.

Boehm, P. D., Douglas, G. S., et al. (1997). Application of petroleum hydrocarbon chemical fingerprinting and allocation techniques after the *Exxon Valdez* oil spill. *Mar. Pollut. Bull*., **34**(8):599–613.

Boehm, P. D., Page, D. S., et al. (1998). Study of the fates and effects of the *Exxon Valdez* oil spill on benthic sediments in two bays in Prince William Sound, Alaska: 1. Study design, chemistry, and source fingerprinting. *Environ. Sci. Technol*., **32**(5):567–576.

Bruce, L. G., and Schmidt, G. W. (1994). Hydrocarbon fingerprinting for application in forensic geology; review with case studies. *AAPG Bull*., **78**(11):1692–1710.

Burrough, P. A. (2001). GIS and geostatistics: essential partners for spatial analysis. *Environ. Ecol. Stat*., **8**(4):361–377.

Burrough, P. A., van der Perk, M., et al. (1999). Environmental mobility of radiocaesium in the Pripyat Catchment, Ukraine/Belarus. *Water Air Soil Pollut*., **110**(1):35–55.

Carpi, A. (1997). Mercury from combustion sources: a review of the chemical species emitted and their transport in the atmosphere. *Water Air Soil Pollut*., **98**(3–4): 241–254.

Carpi, A., and Mital, J. (2000). Expanding use of forensics in environmental science. *Environ. Sci. Technol*., **34**(11):A262–A266.

Carpi, A., Weinstein, L. H., et al. (1994). Bioaccumulation of mercury by sphagnum moss near a municipal solid waste incinerator. *J. Air Waste Manag. Assoc*., **44**(5):669–672.

Costa, H. J., White, K. A., et al. (2004). Distinguishing PAH background and MGP residues in sediments of a freshwater creek. *Environ. Forensics*, **5**(3):171–182.

Davis, A., Howe, B., et al. (2005). Use of geochemical forensics to determine release eras of petrochemicals to groundwater, Whitehorse, Yukon. *Environ. Forensics*, **6**(3): 253–271.

Gaines, R. B., Frysinger, G. S., et al. (1999). Oil spill source identification by comprehensive two-dimensional gas chromatography. *Environ. Sci. Technol*., **33**(12):2106–2112.

Galt, J. A., Lehr, W. J., et al. (1991). Fate and transport of the *Exxon Valdez* oil spill: 4. *Environ. Sci. Technol*. **25**(2):202–209.

Haddad, R. I. (2004). Invited editorial: What is environmental forensics? *Environ. Forensics*, **5**(1):3.

Harley, R. A., Hannigan, M. P., et al. (1992). Respeciation of organic gas emissions and the detection of excess unburned gasoline in the atmosphere. *Environ. Sci. Technol*., **26**(12):2395–2408.

Kaplan, I. R. (2003). Age dating of environmental organic residues. *Environ. Forensics*, **4**(2):95–141.

Kaplan, I. R., Galperin, Y., et al. (1997). Forensic environmental geochemistry: differentiation of fuel-types, their sources and release time. *Org. Geochem*., **27**(5–6):289–317.

Lundegard, P. D., Sweeney, R. E., et al. (2000). Soil gas methane at petroleum contaminated sites: forensic determination of origin and source. *Environ. Forensics*, **1**(1):3–10.

Morrison, R. D. (2000a). Application of forensic techniques for age dating and source identification in environmental litigation. *Environ. Forensics*, **1**(3):131–153.

Morrison, R. D. (2000b). Critical review of environmental forensic techniques: II. *Environ. Forensics*, **1**(4):175–195.

Murphy, B. L. (2000). Editorial. *Environ. Forensics*, **1**(4):155.

Opsomer, J. D., Agras, J., et al. (1995). An application of locally weighted regression to airborne mercury deposition around an incinerator site. *Environmetrics*, **6**(2):205–219.

Petrisor, I. (2005). Sampling and analyses: key steps of a forensics investigation. *Environ. Forensics*, **6**(1):1.

Ram, N. M., Leahy, M., et al. (1999). Environmental sleuth at work. *Environ. Sci. Technol*., **33**(21):464–469.

Sauer, T. C., and Uhler, A. D. (1994). Pollutant source identification and allocation: advances in hydrocarbon fingerprinting. *Remediation*, **5**(1):25–46.

Schauer, J. J., and Cass, G. R. (2000). Source apportionment of wintertime gas-phase and particle-phase air pollutants using organic compounds as tracers. *Environ. Sci. Technol*., **34**(9):1821–1832.

Scott, P. K., Unice, K. M., et al. (2007). Statistical evaluation of metal concentrations as a method for identifying World Trade Center dust in buildings. *Environ. Forensics*, **8**(4):301–311.

Stout, S. A., Uhler, A. D., et al. (1998). Environmental forensics unraveling site liability: an interdisciplinary analytical approach can unravel environmental liability at contaminated sites. *Environ. Sci. Technol*., **32**(11):260A–264A.

Suggs, J. A., Beam, E. W., et al. (2002). Guidelines and resources for conducting an environmental crime investigation in the United States. *Environ. Forensics*, **3**(2):91–113.

Tykot, R. H. (2002). Chemical fingerprinting and source tracing of obsidian: the central Mediterranean trade in black gold. *Acc. Chem. Res*., **35**(8):618–627.

Vives-Rego, J. (2004). Environmental forensics: a scientific service at the service of justice and society. *Environ. Forensics*, **5**(3):123–124.

Wenning, R. J., and Simmons, K. (2000). Editorial. *Environ. Forensics*, **1**(1):1.

Wise, S., Haining, R., et al. (2001). Providing spatial statistical data analysis functionality for the GIS user: the SAGE project. *Int. J. Geogr. Inf. Sci*., **15**:239–254.

Zemo, D. A., Bruya, J. E., et al. (1995). The application of petroleum hydrocarbon fingerprint characterization in site investigation and remediation. *Ground Water Monit. Remediation*, **15**(2):147–156.

CHAPTER 2

Principles and Issues in Forensic Analysis of Explosives

JIMMIE C. OXLEY

University of Rhode Island, Chemistry Department, DHS Center of Excellence for Explosives Detection, Mitigation, and Response, Kingston, Rhode Island

MAURICE MARSHALL

(Formerly) Forensic Explosives Laboratory, Defence Science and Technology Laboratory, Fort Halstead, Sevenoaks, Kent, UK

SARAH L. LANCASTER

Forensic Explosives Laboratory, Defence Science and Technology Laboratory, Fort Halstead, Sevenoaks, Kent, UK

Summary Proper handling of evidence is critical to identifying and convicting a criminal. Evidence at the scene of an explosion, especially a large explosion, offers some unique challenges. Basic principles of evidence collection, handling, storing, and identifying are discussed herein.

2.1 Introduction 24
2.2 Sample collection 25
2.3 Packaging 29
2.4 Sorting 30
2.5 Documentation 31
2.6 Environmental control and monitoring 31
2.7 Storage 33
2.8 Analysis 33
2.9 Records 36
2.10 Quality assurance 36
2.11 Safety and other issues 37
Conclusion 37
References 38

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

2.1 INTRODUCTION

Many types of laboratories engage in chemical analysis. In forensic science laboratories, a wide variety of chemical, biological, and physical analyses are undertaken. The principal difference between forensic analysis and more general analysis is the degree of certainty required of the results. The technical issues depend on the nature of the sample: in particular, on whether bulk samples or invisible traces are being sought. Bulk samples are considered to be anything that is visible to the naked eye and can range from micrograms to several kilograms of material.

The most obvious difference between analyses of bulk versus trace samples is the relationship between the sample and the environment. Sometimes the desired analyte may be in the environment, and sometimes a species in the environment (e.g., water, oxygen, iron particles) may degrade the sample or affect the results. The analyte in the environment is not generally an issue in bulk samples. In trace samples, the amount of contamination may be large enough to distort results. Thus, if the sample contains a tiny amount of the analyte sought and the environment contains a large concentration of that species, extreme precautions will need to be taken to protect the sample and exclude contact with the environment. Conversely, if the sample contains a large concentration of the analyte and the environment a tiny amount, the issue is trivial. An understanding of the composition of the background environment is therefore highly desirable, but not always possible. This needs to be considered when reporting results.

A robust and well-designed trace analysis protocol is likely to involve (1) physical separation between the analyst and the sample; (2) the use of disposable items for handling, packaging, and containment; (3) appropriate blank and control samples; and (4) environmental monitoring. The precise detail of the measures will depend on the environmental challenge to the integrity of the analysis, and it is often possible to strengthen one protective technique to counter the weakness or absence of another.

Although explosives are, of course, simply chemicals or chemical mixtures, in some respects their analysis is easier because many explosives (e.g., the organic explosives) are rarely found in the general public environment. In response to the Provisional Irish Republican Army (PIRA) bombing campaign on the mainland (1970s to 1990s), the United Kingdom (UK) led the way in protocols pertaining to explosive evidence. Over the course of a decade four studies were produced documenting explosives in the environment: two on background levels of military explosives and two on levels of inorganic ions (Crowson et al., 1996; Walker et al., 2001; Cullum et al., 2004; Sykes and Salt, 2004). Of 670 samples collected on the British mainland, only eight showed traces of organic explosives. A recent repeat of this study in the United States showed that only three out of 333 samples had traces of high explosives (Laboda et al., 2008). Both the UK and U.S. studies showed nitrates at the microgram level in 20 to 30% of the samples.

Much attention has quite correctly been paid to issues of cross-contamination in the analysis of explosive traces because of the generally very serious nature of the criminal offences involved. However, in reality, all forensic trace analyses need to be protected against the risk of compromise by ill-founded suspicions of all types of cross-contamination. There should, of course, be no such suspicion that is well founded! Trace analysis procedures need to be designed, tested, and validated to ensure that positive evidence is produced showing the integrity of the results. This applies whether the sample is a few nanograms of explosive or a few nanograms of DNA.

This is a field where contamination of evidence can easily occur due to the wide range of vapor pressures exhibited by explosive formulations. Did the Madrid bombing use Goma 2 ECO or Titadyn? The answer is critical because it would point to one or another terrorist group. Dinitrotoluene (DNT) and nitroglycerin were found as part of the evidence. Was it in the explosives used to make the terrorist devices, or was it a result of cross-contamination during storage since DNT is highly volatile?

2.2 SAMPLE COLLECTION

Unfortunately, forensic chemists do not always have control over the vital aspect of collection and packaging of the materials they must examine. There is a world of difference between the effort and preplanning required for dealing with a large bombing attack and that required for the investigation of a bombing of a mailbox or single residence. Most forensic scientists will only deal with the latter. Nonetheless, preplanning will be worthwhile. Clean containers and packaging materials should be procured and stockpiled ready for use. Examples of such items are disposable scoops, scrapers, dustpans, and brushes, as well as metal cans and nylon bags of various sizes. Similarly, collection devices such as brushes, scoops, scrapers, and vacuum pumps and filters should be obtained. Minivacuums can be constructed from disposable plastic tubing, syringe filters, and plastic syringes.

Preferably all items used for collection should be subjected to quality assurance tests before use. The easiest way to ensure the cleanliness of tools used for collection of trace explosive evidence is to use disposable items from a known supplier which have just come from the box. If possible, a statistical sample of each item should be prescreened before operational employment. However, if a prescreen is not possible, a more rigorous regime of analysis of blank and control samples can be substituted. It is to be understood that this may entail the risk of loss of evidence if a control is analyzed as being positive. It should also be noted that suitable control samples should still be obtained, even with the use of prescreened materials.

Swabbing to collect trace explosive evidence is a common practice. Swabs may be pre-prepared using solvent-washed cotton balls that are either dry or have been wetted with a solvent (Jenkins and Yallop, 1970). Although ideally the swabs should be premade and preanalyzed, necessity may drive the investigator to use improvised swabs (e.g., alcohol-wetted hand-wipes, facial tissues, paper towels). Although swabbing is a superficially simple technique, in fact a plethora of interacting variables and issues need to be considered. A key issue is to identify the type of explosive being sought: for example, inorganic or organic? Another issue is practicality and generality. Although it is arguably possible to design swabbing protocols that are optimized for the collection of particular explosives, our experience is that it is better to design for the widest possible application. Bomb scenes are generally places of chaos where decision making is handicapped by lack of information, and it is operationally much better to avoid the need to make choices by providing sample collection kits of general application.

In the UK in the early 1970s, dry swabs, water-wetted swabs, and solvent-wetted swabs were all used for collecting different types of residue. Subsequently, it was realized that the choice of dry or water- or solvent-wetted swabs was usually little better than guesswork prior to laboratory analysis. Furthermore, the choice of solvent or water was less significant than it first appeared. Although the solubility of diverse inorganic and organic explosive species varies dramatically in water and organic solvents, the small amounts present in trace samples means that the actual concentration that has to be dissolved in the swabbing solvent is rather low. Moreover, too strong a solvent will not necessarily recover more explosive residue; rather, it will merely pick up more unwanted background material, thereby complicating the subsequent preanalytical cleanup and concentration in the laboratory. Another very important consideration in the choice of a swabbing solvent is toxicity. These various issues led the UK to adopt a 50 : 50 ethanol–water mixture in their swab kits. This was found to provide recovery of a broad spectrum of both inorganic and organic explosive traces and to be compatible with subsequent laboratory protocols (Douse, 1985; Warren et al., 1999).

Very large samples do not lend themselves to solvent washing. Large containers for solvent extraction may be available, but plastics may allow interferents to elute. Generally, borosilicate glass rather than plastic containers are preferred, but extraction of ions from soda glass can also be a problem. Since large containers are generally not available, large samples should be swabbed. The swabs found in premise kits could be used. (Premise kits are useful if prepared ahead of time. They contain clean, validated swabs; disposable gloves; tweezers; and a solvent such as ethanol–water 50 : 50. A control in a premise kit would be a solvent-wetted but unused swab. Like hand kits, premise kits are heat sealed under positive pressure.)

Driven by the pressures of dealing with the long-term PIRA campaign, the UK has pioneered the use of premade, screened kits for both hand testing and premise screening. In addition, in a pilot program hair kits have been prepared specifically for explosives, and some laboratories include materials for hair sampling in kits used for the recovery of gunshot residue.

Every effort should be made to ensure that the investigator does not contaminate the scene. This includes using fresh, disposable tweezers to handle and manipulate swabbing materials and donning disposable outerwear before entering the scene. Ideally, different investigators should obtain samples from the crime scene and a suspect's premise to eliminate the possibility of cross contamination (of explosives traces and other types of evidence, such as DNA or fiber). If this is not possible and the same investigators are screening the scene and the suspect's premises, they need to take steps to ensure that (1) they do not inadvertently transfer contamination between scenes, and (2) they provide objective documented evidence to prove the efficacy of their measures to prevent cross-contamination. Such measures might include, for example, use of disposable overalls, gloves, hats, hoods or hairnets, and bootees or overshoes; handwashing, hairwashing, or whole-body showers, and possibly the taking of personal control swabs or personal checking with airport-style explosive screening instruments. In addition, attention needs to be paid to the cleaning or overwrapping of personal items such as spectacles, hair bands, wristwatches, and jewelry. These are excellent potential sources of trace evidence from suspects and present a risk of cross contamination for the scene examiner and the laboratory scientist. The simplest approach is, of course, the best; do not wear such items to a scene.

Searching a scene for bulk and trace evidence involves somewhat different approaches. The investigator generally is called upon by law enforcement to identify bulk explosives based on visual identification. This is likely to come about because police have entered a suspect's premises and found something that seems suspicious, have intercepted an attempted bombing and disrupted the device, or have discovered a device that fails to function or functions inefficiently, such as a poorly constructed pipe bomb. Trace explosive evidence is not evident; that is, it is invisible to the investigator or the collector of evidence. Both the collector and the investigator must rely on good training, sense, and good luck. It is vital that collection material and people be shown to be free of contamination prior to entering a scene.

At the scene of an explosion, the investigator will start by attempting to identify the seat of the explosion. Generally, this will be indicated by the presence of a crater. A crater is not always formed, such as in the case of a low level of explosives; if a disperse-phase (gas-phase) explosion has occurred; or if a high-explosive condensed-phase explosion has occurred either in midair or on a frangible surface that has been destroyed in the process. The location of the explosion seat may also be identified by tracking back from the patterns of physical damage at the scene, such as the marks made by objects that have been moved, or from patterns of fragment penetration in nearby items.

Once the explosion seat has been located, the investigator can begin to look for likely places to find explosive residue. The best candidates will be substrates that were close to the explosion when it happened. The obvious first step is to take samples from the crater left by the explosion. However, most of the residue is expected to be found “near,” not “in,” the crater where it has been thrown, spalled from the exploding charge, or where it has condensed out of the fireball onto relatively colder surfaces. Nearby metallic items such as door frames and furniture, window frames, and light fittings are such surfaces and should be examined. In general, if a surface shows evidence of explosive damage or disruption [i.e., microcratering, pitting, or sooting (either visible to the naked eye or through a low-power lens)], it may be a good source of explosive residue. In the search for explosives at a bomb scene, impermeable surfaces offer the best chance of finding explosive residue for the majority of explosives. However, for very volatile explosives, such as triacetone triperoxide (TATP), the issue is the rapid loss of the explosive from surfaces due to evaporation or sublimation. In these circumstances, an absorbent, porous substrate (e.g., carpet or wood) is likely to retain the volatile explosive much longer than is a hard surface such as metal. In any case, sealed packaging is critical for retaining this evidence. As discussed below, for most explosives evidence nylon bags are preferable. Because one does not know a priori which type of explosive has been used, the on-scene investigator must take samples that cover either possibility.

In the case of a car bombing, such as a booby trap attack intended to kill the occupants, the preferred option would be to package the vehicle and take it to the forensic laboratory. Whether the vehicle is examined at the scene or in the lab, the first step is to identify the seat of the explosion and then take samples from both the nonporous surfaces, such as plastic and metal, and the porous surfaces, such as carpet and upholstery. If a vehicle, a car or truck, was used to contain the bomb, explosive residue will probably be found on those scattered vehicle parts that were in close proximity or contact with the explosive before detonation; thus, depending on the initial location of the bomb, residues are likely on parts from the trunk or passenger compartment, whereas they are generally not to be found on the engine block.

When searching for evidence that explosives may have been stored or hidden in a location or a suspect's premises, the investigator will look for explosive traces in likely storage places. Apart from cupboards and shelving, other locations, such as hidden spaces behind removable paneling in bathrooms, should also be sampled. In some instances, dusty surfaces may reveal the outline of containers that have been removed. If it is suspected that a premises has been used for explosives manufacture or processing, it will also be worthwhile to remove the pipes and waste traps from sinks and washbasins to see if any explosive residue has collected there.

If manufacture or chemical synthesis is suspected, clearly any obvious items of laboratory equipment will be seized for examination. However, amateur experimenters and makers of improvised explosives are also quite likely to press ordinary domestic equipment into service: for example, kitchen mixing bowls, cooking spatulas, measuring cylinders or jugs, food processors and stirrers, and scales or balances. All of these should be examined for explosive residue. In the absence of any obvious laboratory work area at a suspect's premises, it is often worthwhile to swab the kitchen work surfaces for traces of explosives residue, paying particular attention to cracks or crevices where residues might have been missed by any cleaning efforts. Discovery of clandestine laboratories presents a number of difficult issues for an investigator. First and foremost is the likelihood that unknown hazardous materials are present, often in unlabeled containers. Second is the very real risk of booby traps and antihandling devices. The investigator will wish at an early stage to assess whether a clandestine laboratory is likely to be drug or explosives related. The presence of technical literature dealing with one or the other subject can help. In addition, the chemical synthesis of most illicit drugs is chemically more complex than that of the common organic explosives. Hence, as an initial guide, an elaborate laboratory with complex equipment is more likely to be drug than explosive related. It should be noted that there have been examples of drug laboratories being booby trapped with explosive devices precisely to guard against the arrival of law enforcement. For example, packets wrapped to appear as illegal drugs were in reality Armstrong's mixture, an extremely sensitive formulation used in toy caps and party poppers. Moreover, the first encounter with TATP in a criminal context in the United States was in booby traps in a drug laboratory (Christian, 2005).

Frequently, the investigator needs to collect evidence from a suspect's person. Typically, the options are hands, hair, clothing, jewelry, and glasses. The law enforcement investigator and forensic examiner will be familiar with the use of pre-prepared kits for examining hands for gunshot residue. The principle can be extended to explosive residues, but the design details must be modified. Hair is washed much less frequently than hands; furthermore, it has been shown that explosive handlers can be allowed to wash their hair and in some instances still exhibit detectible levels of contamination. Examination of hair for trace explosives can be performed simply by combing the hair with a comb on which clean gauze has been threaded. After multiple passes through hair, the gauze can be disengaged from the comb and extracted with solvent for chemical analysis (Sanders et al., 2002; Oxley et al., 2005,2007a,b). Chemical digestion of hair is performed when looking for evidence of drug use, but such procedures have not yet been investigated for explosives.

Clothing evidence has been pivotal in a number of high-profile cases. The Oklahoma City bomber, Timothy McVeigh, had traces of pentaerythritol tetranitrate (PETN) on his clothing; and Ressam Amad, the would-be millennium bomber, had acid burns and other traces of explosives on his trousers. In high-profile cases it is usual not only to confiscate and examine the clothing worn when the suspect was apprehended, but also the entire contents of his wardrobe as well as bags, backpacks, and holdalls, which may have been used to transport devices. It is commonplace when a person is handling items for him to wipe his hands on the seat of his trousers. Other common hand movements that potentially transfer residue from hands to clothing are adjusting the knees of one's trousers when sitting down or putting hands in one's pockets. All these spots are good places to look for explosives traces.

As mentioned previously, jewelry, rings, bracelets, watch straps, hair bands, and glasses are likely to be worn for long periods of time with no rigorous attempts at cleaning. These readily become contaminated with any explosive handled by a suspect and are likely to retain traces in crevices over long periods of time. This is a positive feature in terms of finding evidence of malfeasance and a negative feature in preventing the investigator from contaminating the evidence. Generally, these items are seized and solvent washed, preferably with the aid of an ultrasonic bath, although items such as leather watch straps can be wiped with a solvent-wetted swab rather than being solvent washed.

2.3 PACKAGING

Without the proper packaging, all the effort made in the collection of evidence is for naught. Selection of container material should be done well in advance of need and should consider the size of the object to be collected and the degree of isolation required between the sample and the environment. For example, does the sample need to be protected against sorption of external vapors? Or is it a potential source of vapors that might be lost as evidence or contaminate other samples? Some samples, of course, should not be sealed in nonpermeable bags; these included dried or wet blood, which may be subject to attack by bioorganisms, as well as items likely to undergo chemical decomposition and generate gaseous decomposition products.

Small samples are usually sealed in nylon bags with the neck twisted and taped, doubled over twice, and taped again. Of the plastics commonly available, nylon is the most impervious to vapors; commercial evidence bags are usually polyethylene. Large samples, such as cars, light fixtures, or window and door frames, which may have trace residue and might show evidence of explosive cratering, should be sealed in polymer sheeting (tarpaulins) and taped together with adhesive tape. Medium-sized samples that will not be analyzed for trace explosives can be packed in plastic tubs or crates with lids, evidence seals, and labels.

It is likely that the collection of evidence at the scene will result in contamination of the outside of the collection container. For this reason, after the initial bagging at the scene, a second bagging should occur well away from the scene at an established workstation. Some laboratories actually perform a third bagging upon receipt at reception to contain any external contamination picked up in transit (or at the scene) and to prevent any contamination from the nontrace areas of the laboratory itself. An on-site workstation outside the target premises can be produced simply by laying out clean disposable paper at a convenient spot protected from the elements [e.g., the back of a police van (assuming it is known and can be shown not to have contained explosives or firearms)]. If it is known that evidence of trace explosives is to be sought and operational constraints allow time for preplanning, it may be worthwhile considering hiring clean vehicles and prescreening them to ensure the absence of explosive contamination before the operation.

Opening the evidence at the forensic laboratory requires as much thought as does packaging it. If the sample has been triple-wrapped, its wrappings need to be removed sequentially as it moves to progressively cleaner areas. The multiple layers of packaging are pealed off sequentially with gloved hands in a manner that ensures no transfer of contamination from the outside world into the trace area. This means that fresh gloves are used at each step. The innermost layer of packaging is not opened until the item is in a location guaranteed to be free from contamination.

2.4 SORTING

The Forensic Explosive Laboratory in the UK, which has had a great deal of experience processing debris from large bombing events, has found it useful to use large, lidded plastic bins for large pieces of evidence. At the height of the PIRA bombing campaign, stashes of these bins were maintained at various locations in England. With huge volumes of debris it was found much easier to locate the anomalies if the debris were sorted into bits of similar size. Therefore, in the field or directly on returning to the analytical facility, they presorted what was considered the most promising debris. The goal of such sorting was to find pieces derived from the bomb or its container.

2.5 DOCUMENTATION

A great deal of material is required to fully document a crime scene: for example, photographs and drawings as well as notes and logbooks. Every item of evidence that is collected needs to be uniquely labeled and controlled. Each person involved in evidence collection and labeling must have a notebook to keep track of the exhibit references he or she has generated. For forensic purposes the collection and packaging procedure needs to ensure evidential continuity and integrity, with a properly documented chain of custody. This is usually accomplished by attaching a signature seal and receipt that is countersigned as the item passes from hand to hand. As one might expect, modern electronic solutions have been tested in some jurisdictions with the aim of supplanting the traditional paper-and-ink systems. Although barcode labels and electronic scanners are widely used in retailing, this technique is more easily implemented in the laboratory than in the field.

Adequate documentation is essential to the satisfactory progress of a case. Conversely, it is all too easy to generate superfluous paperwork, with accompanying nugatory effort. In the late 1980s it was normal in the UK to divide a bomb scene into numerous zones, often on a grid pattern, and expend much effort recording the fact that a particular item was found in a specific zone. Eventually, the pressure of work generated by a massive terrorist bombing campaign forced an evaluation of such practices. It was realized that it was rarely, if ever, possible to draw any useful evidential conclusion from such detailed information about the location of bomb debris. The key piece of information for a court was that a specific item had been found at the bomb scene. Consequently, the practice of zoning at bomb scenes was curtailed dramatically. Typically, a scene might use the area around the crater and other natural boundaries for zoning, or if the scene was a building, perhaps divided according to rooms and whether the evidence was recovered from inside or outside the building.

2.6 ENVIRONMENTAL CONTROL AND MONITORING

As stated in Section 2.1, the precautions necessary to protect sample integrity depend on the challenges presented by the environment. For example, does the environment contain significant amounts of the target analyte so that contamination of the sample becomes a real rather than a theoretical risk? This issue is particularly relevant when handling trace rather than bulk evidence. To protect the integrity of the sample, a rigorous protocol is needed to ensure laboratory cleanliness, complemented by a measuring regime to demonstrate that the requisite standards of cleanliness have actually been achieved and maintained. The process begins with the initial design of the laboratory. The laboratory layout should impose barriers between the unknown outside environment and the controlled trace environment. Physical features include airlocks, filtered air with positive pressure, and a separate antechamber through which personnel must enter or exit, in the process changing into clean, disposable overgarments, removing watches, jewelry, and eyeglasses.

The antechamber. In the antechamber, personnel will wash hands and other exposed skin and don hair covers, disposable overalls, shoe covers, and gloves. They should also ensure cleanliness by a quick screen using explosive detection instrumentation [e.g. ion mobility spectrometry (IMS)]. Pens, papers, watches or other jewelry, and eyeglasses should be left here. Duplicates should be waiting on the other side of the airlock in the trace examination room. Administrative controls include access control to ensure that only appropriately trained or supervised personal can enter the trace laboratory. In fact, since every entry and egress represents a theoretical contamination risk, the number of people entering the analysis area should be minimized. A log of visits should be maintained. It is from this antechamber that the forensic chemist and the sample, still wrapped in its innermost layer of packaging, enter the trace laboratory.

The trace laboratory. A regular cleaning schedule is necessary. The primary defense is a filtered air system, whose performance is monitored both by differential pressure gauges and atmospheric particulate monitors. Floors should be washed at least once a week with strong alkaline detergent. One approach to ensuring cleanliness is to predefine a number of squares (e.g., 50 cm × 50 cm) on the floor which are regularly swabbed for contamination (e.g., weekly). The tops and insides of cupboards and drawers and air intakes need a less frequent scheduled cleaning and screening. The goal is to see that every surface is cleaned and tested for contamination at a known interval. It is important that the forensic chemist hypothesizes where contamination might accumulate. Analytical instruments may be elevated on small blocks to facilitate cleaning under each instrument. It is useful if instruments and materials can be dedicated to a particular stage in the analysis. As far as possible, disposable tools should be used. For example, the trace laboratory should have its own paper and pens, which are used only once. Wrenches, screwdrivers, and beakers should be cleaned by heating in an oven ($\sim 250^{\circ}$C) and either storing the materials in the oven or removing then and wrapping them individually in nylon bags until use.

At the beginning of the work cycle, the analytical bench should be cleaned: for example, with alkaline detergent solution and ethanol, and wiped down with paper towels. A fresh disposable benchtop surface, such as butcher paper, aluminum foil, or Benchkote, should be used for each sample. The forensic chemist should change gloves frequently, at least between samples. To provide evidence that no contamination is present, the forensic chemist should take control swabs of the work surface, the front of her suit and gloved hands, and the disposable benchtop surface. These are considered control samples; they are necessary to show that contamination has not occurred (Hiley, 1998; Crowson et al., 2001; Beardah et al., 2007).

2.7 STORAGE

Like most evidence, aging can result in loss or contamination of explosives evidence. A rapid turnaround time is important. In the UK during the PIRA terrorist campaign, the Forensic Explosives Laboratory, responsible for analysis of trace and other evidence, pledged a 6 hour turnaround time for a single hand-test kit and two weeks for a preliminary report, regardless of the type of case.

Storage conditions are also critical; for example, the first bombings using TATP were in Israel. Even in cases where intelligence pointed to the use of TATP, forensic examiners found no residue. Subsequent studies showed that TATP is so highly volatile that evidence should be thoroughly sealed and, ideally, examined immediately, to maximize the chances of recovery and detection. Sealing and storing at subambient temperatures are not always necessary but can hardly hurt explosive evidence. (*Note:* Samples with moist blood or other biological media present may be subject to bacterial growth if sealed without the additional precaution of subambient temperature. However, consideration should then be given to the effect of thawing on the other materials.)

2.8 ANALYSIS

Bulk analysis may be of visible particles of an explosive and/or its combustion products, but it may be of bomb components, such as timers, fuses, wiring, or tape. The analysis is usually performed to identify the material and to make comparisons, for example, between a material recovered from a crime scene and one recovered from a suspect’s premises. At best, it is only possible to say that two materials are different or to list those aspects in which they are similar. However, to make such a statement, it is necessary to have a well-founded understanding of the likely variations in a material and the practical significance of such variations. This highlights the value of reference collections and surveys of items and materials likely to be encountered in case work (e.g., carpets, car paints, accelerants, and explosives). Comprehensive databases or reference collections are useful for all sorts of bomb-making gear (e.g., pipes, timers, tapes, and batteries).

In some cases the explosive material used in the bomb can be seen. Inefficient devices such as those using smokeless powder or black powder often leave visible residue. High explosives usually leave residue, too, but since it is not visible to the human eye, it may not make the multiple down-selects to end up in the final analytical protocol. In efficient bombs, the residual explosive has probably been thrown out from the surface of the bomb—spalled off the surface (Kelleher, 2002). Thus, looking just outside the crater may be more productive than looking at evidence inside the crater.

Visible explosive particles should first be examined intact (i.e., with nondestructive techniques, as clues are available at this point). Although prilled ammonium nitrate is made at several hundred locations around the world, there are wide variations in prill shape (to be determined visually) and prill coating (to be determined analytically). With the proper reference collection, such variations may make it possible to trace the material back to its original manufacturer.

For bomb scene materials with no visible residue, prescreening for trace explosives may be possible using an explosive detector. In practice, this approach is less labor-saving than it first appears and, more important, runs the risk of inadvertent contamination of valuable trace evidence or the inadvertent exclusion of evidence with traces of material not detected by the specific detector. Otherwise, the more accurate but laborious approach is a multistep extraction of the debris to separate organic from polar species. One approach is to extract the debris with the aid of an ultrasonic bath using a solvent for both polar and nonpolar species. If one uses a strong solvent, such as acetone or acetonitrile, all of the explosive residue will dissolve, but so will many unwanted species, which serve to obscure the analysis and clog the system. The Forensic Explosives Laboratory in the UK has settled on a technique using a relatively weak solvent system.

The debris is extracted with ethanol–water (50 : 50) and run through a glass pipette containing a few grams of a solid adsorbent such as Chromosorb 104. The packing retains the common organic explosives (e.g., TNT, PETN) and allows the polar species to elute. This polar fraction is analyzed by ion chromatography looking for species such as nitrates (ammonium or uronium) or chlorates. The Chromosorb column is then eluted with ethyl acetate to extract the organic species. These are analyzed by gas chromatograph (GC) or liquid chromatography (LC), depending on the anticipated volatility of organic explosives suspected (see Figure 2.1). For example, the thermal stability of hexamethylene triperoxide diamine (HMTD) is so poor that LC is the preferred technique; otherwise, samples break down in the inlet port of the GC. Obviously, the GC or LC can be attached to a variety of detectors. For GC, an electron capture detector (ECD) is the most sensitive detector, and a thermal energy analyzer (TEA), which analyzes for nitro groups, is the most specific. If LC is used, a mass selective detector (MSD) is usually employed. This has the added feature of providing orthogonal confirmation: not only retention time but also fragmentation patterns. While the field of mass spectroscopy (MS) is developing rapidly, for the common benchtop GC–electron ionization–MS, most explosives rarely exhibit a parent peak (P). Identification is based on fragmentation patterns, where P–OH is often observed. Yinon has discussed this extensively (Yinon and Zitrin, 1981, 1993; Yinon, 2007). An MS database of explosives is available online at http://expdb.chm.uri.edu, but it is essential that forensic scientists also create their own libraries using their own experimental conditions, as mass spectral fragmentation patterns are highly dependent on the exact conditions under which they are obtained.

Orthogonal techniques are defined as two techniques for which sample identification is based on completely different principles (e.g., infrared spectroscopy and mass spectroscopy). In practice, finding two absolutely orthogonal techniques is rarely achievable; rather, in practice, significantly different techniques are used (e.g., using two different GC columns which yield significantly different retention times). The more orthogonal the techniques, the higher the level of confidence in the assignment of identity.

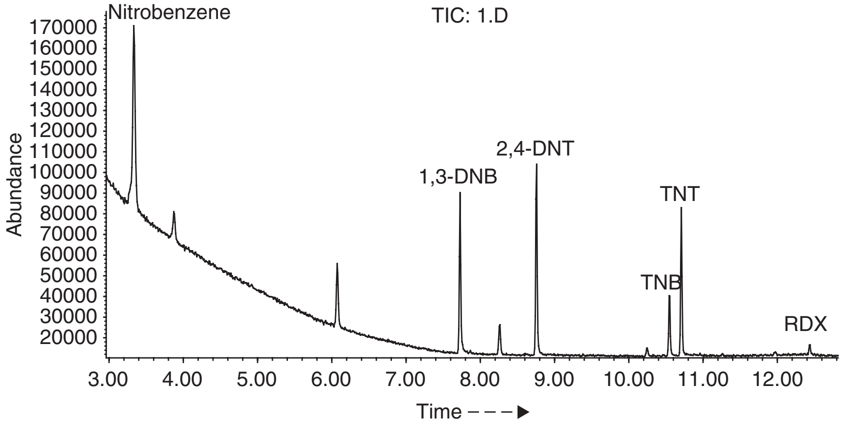

Restek 8330 Calibration Mix #1

Shipped as 1000 ppm in ACN of:

HMX	RDX	1,3-DNB	TNB
NB	TNT	2,4-DNT	

Diluted to 20 ppm in ACN

GC/MS Conditions
HP-5MS: 0.25mm x 30m x 0.25μm
Inlet: 20.94 psi, 225°C, Constant Flow
Splitless 15 sec., purge flow 20 mL/min,
Flow Rate: 2 mL/min Average velocity: 53 cm/sec
Oven Program: Initial 100°C x 2 min.,
Ramp 10°C/min. to 250°C, Hold 250°C 8 min.
Transfer Line: 270°C

Figure 2.1 Chromatogram of explosive calibration standard.

In developing the analytical protocols discussed above, it is important to use standard solutions with known concentrations of explosives. Mixed standards are useful. Not only do they demonstrate the system suitability and performance across the entire range of explosive species being sought, but the mixed standard can never be confused as the real result. For high precision in retention results, a retention time marker [e.g., musk tibetine or fluoronitrotoluene (for GC)] should be added to both sample and standard solutions.

Although forensic samples are rarely quantified, if accurate quantification is desired, explosive recoveries should be determined. A good approach is to use the method of standard additions. This involves dividing the sample into two parts, spiking one part of the sample with a reasonable amount of explosive, typically between one-third and two-thirds of the amount estimated to be present in the sample, then performing the standard workup on the original and spiked samples and determining what percentage of the spike was recovered.

Isotope ratio mass spectroscopy (IRMs) applied to forensic evidence is a vibrant research area at the time of this writing. Much work is needed to establish reproducibility of results and degree of variability among "identical" samples. In principle it is easy to say that if two samples differ in some significant respect, they are in fact different. However, the converse is not true. One can never state with absolute forensic certainty that two items are identical. All one can do is to list all those features in which the two samples match. In practice there comes a point at which a reasonable person will conclude that a match has been established beyond a reasonable doubt. Thus, it may be possible, in a few years, that with the aid of foreseeable improvements in isotope ratio mass spectrometry, one could testify that a sample of explosive recovered from a pipe bomb at a scene matches a sample of explosive recovered from a mixing bowl at a suspect's premises and that it is highly probable that the samples came from the same original batch of explosive. This issue of the definition of a match is a more general one for forensic science, and would benefit from wider consideration, and preferably agreement within the judicial system. The obvious parallel is the long-established UK criterion that for two fingerprints to match, at least 16 points of agreement must be identified. Other jurisdictions accept a lower figure; for example, in the United States, 12 points are considered sufficient. For this to become possible, large reference collections and databases will become essential, and possibly the very expensive IRMS instrument will become so widely purchased that the price will become affordable to every forensic lab (much as has happened with GC–MS). Rapid improvements in electronic instrumentation suggest that the limitation in precision currently encountered in this technique may be attributed to the chemical methods used in the sample preparation.

2.9 RECORDS

While keeping good records of work is always useful, it is absolutely essential in a forensic laboratory. The written record becomes a legal document. Whether the employee leaves the laboratory or is not called to testify for five years, every step of the procedure and every aspect of the results, including negative results, are extremely important to document. Given the fugacity of electronic records, it is wise to archive all materials in the form of printed paper. Such paper records should, of course, be signed and dated by the scientist producing them. Records should include the printouts from various instruments: chromatograms with standards, mass spectral fragmentation patterns, infrared spectra.

ISO 9001 requires that all quality-related records be retained for a minimum of seven years. Although this would apply to forensic evidence, some jurisdictions require much longer document retention (e.g., 20 years in the UK for explosive offenses).

2.10 QUALITY ASSURANCE

For forensic work only techniques and methods of proven validity should be used. These methods should be documented in the form of standard methods or standard operating procedures (SOPs) to which the laboratory staff can refer readily. Similarly, only persons of proven competence should be allowed to report forensic results, and these people should be subjected to periodic proficiency tests. Requirements for personnel training and method validation are found, for example, in ASCLD (American Society of Crime Laboratory Directors) protocols and ISO 17025 (for laboratory work) and ISO 17020 (for scenes).

To create a new laboratory protocol, the developer should undertake any necessary research and precision and validation studies, producing both a draft method and a report documenting his or her findings. Once the method is accepted, the originator would be expected to become the first trainer in this protocol; it would be sensible to bear the latter requirement in mind when selecting a scientist to undertake the method research and development. Elements that should be part of the new method are a clear statement of precision based on a statistical study of replicate analyses; procedures for calibration, preparation of standards, traceability and uncertainty of measurements; and guidance as to the interpretation of results for forensic purposes.

2.11 SAFETY AND OTHER ISSUES

For a newly set up explosive laboratory, a number of issues will arise. These include acquisition of pure explosive standards, explosive storage, licensing and regulatory compliance, and disposal of hazardous materials.

Generally, licensing and approved storage are required before any explosive can be acquired or stored. Once storage is available, it will be necessary to find an entity that can provide pure explosive standards. Most commercial explosives have been formulated with plasticizers, stabilizers, phlegmatizers, and so on. In consequence, commercial explosives will contain significantly less than 100% of the desired explosive. It may be necessary to synthesize the desired explosive or recrystallize a commercial explosive. The physical properties of this explosive must be characterized and recorded. Differential scanning calorimetry (DSC) allows the determination of material purity, provided that the materials have observable melting points (see Figure 2.2). It should be noted that the purified explosive, freed from the additives present in commercial products, may be significantly more hazardous.

Waste from the explosive laboratory will represent various types of hazards: explosive; flammable; toxic; biological; sharps. Proper arrangements need to be made for disposal for each of these waste streams. In general this will require the services of specialist contractors. Explosive laboratories attached to explosive firing ranges may be able to send their explosive waste for disposal. In large facilities, this is typically achieved by incineration. However, for small amounts of explosive and in most laboratories, the best option is chemical destruction. Suitable methods may be found in the literature.

One point that has yet to be mentioned is the fact that many explosives are toxic. This should be obvious from a historic point of view since nitrate esters are used to treat angina, and cyclonite (RDX) was patented for use as a rat poison.

CONCLUSION

Forensic analysis differs from routine chemical analysis in the critical nature of the results and the fact that the results and documentation of it may come under severe

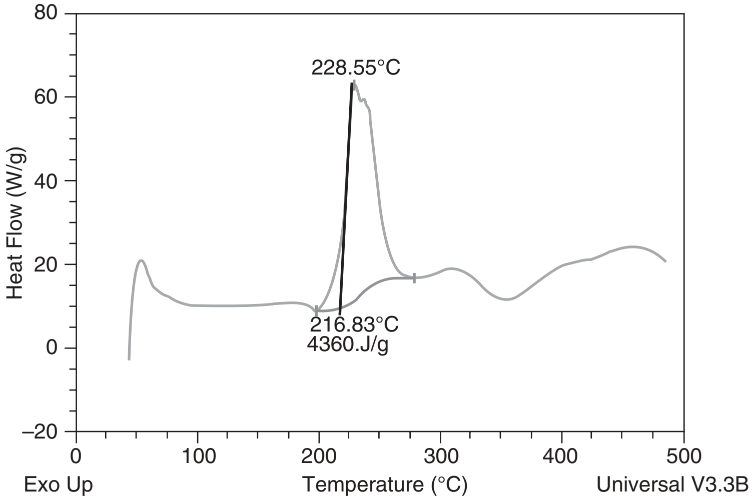

Figure 2.2 Differential scanning calorimetric scan of unknown black solid (suspected of being an explosive). *(See insert for color representation.)*

scrutiny. For trace analysis, every pathway for contamination must be anticipated and mitigated. The differences in analyte concentration between sample and environment must be characterized and proper barriers put in place. There are only two standards in forensic analysis: perfect and useless.

REFERENCES

Beardah, M. S., Doyle, S. P., and Hendey, C. E. (2007). Effectiveness of contamination prevention procedures in a trace explosives laboratory. *Sci. Justice*, **47**:120–124.

Christian, D. R. (2005). The multifaceted demands and dangers you face when investigating clandestine laboratories. *Evidence Technol. Mag*., May–June, pp. 18–21.

Crowson, C. A., Cullum, H. E., Hiley, R. W., and Lowe A. M. (1996). A survey of high explosives traces in public places. *J. Forensic Sci*., **41(6)**:980–989.

Crowson, A., Hiley, R. W., and Todd, C. C. (2001). Quality assurance testing of an explosive trace analysis laboratory. *J. Forensic Sci*., **46(1)**:53–56.

Cullum, H. E., McGavigan, C., Uttley, C. Z., Stroud, M. A. M., and Warren, D. C. (2004). A second survey of high explosives traces in public places. *J. Forensic Sci*., **49(4)**:684–690.

Douse, J. M. F. (1985). Trace analysis of explosives at the low nanogram level in handswab extracts using columns of Amberlite XAD-7 porous polymer beads and silica capillary column gas chromatorgraphy with thermal energy analysis and electron capture detection. *J. Chromatogr*., **328**:155–165.

Hiley, R. W. (1998). Quality control in the detection and identification of traces of organic high explosives. In: Beveridge, A. (Ed.), *Forensic Investigation of Explosions*. London: Taylor & Francis, pp. 315–342.

Jenkins, R., and Yallop, H. J. (1970). The identification of explosives in trace quantities on objects near an explosion. *Explosivstoffe*, **6**:139–141.

Kelleher, J. D. (2002). Explosives residue: origin and distribution. *Forensic Sci. Commun*., 4.

Laboda, K. G., Collin, O. L., Mathis, J. A., LeClair, H. E., Wise, S. H., and McCord, B. R. (2008). A survey of background levels of explosives and related compounds in the environment. *J. Forensic Sci.*, **53(4)**:1556–4029.

Oxley, J. C., Smith, J. L., Kirschenbaum, L., Shinde, K. P., and Marimganti, S. (2005). Accumulation of explosives in hair. *J. Forensic Sci*., **50(4)**:826–831.

Oxley, J. C., Smith, J. L., Kirschenbaum, L., Shinde, K. P., and Marimganti, S. (2007a). Accumulation of explosives in hair: II. Factors affecting sorption. *J. Forensic Sci*., **52(6)**:1291–1296.

Oxley, J. C., Smith, J. L., Bernier, E., Moran, J. S., and Luongo, J. (2007b). Hair as forensic evidence of explosives handling. *Propellants Explosives Pyrotech*., Nov.

Sanders, K. P., Marshall, M., Oxley, J. C., Smith, J. L., and Egee, L. (2002). Preliminary investigation into the recovery of explosives from hair. *Sci. Justice*, **42(3)**:137–142.

Sykes, A.-M., and Salt, I. (2004). Survey of inorganic traces in the environment. *Proceedings of the 8th International Symposium on Analysis and Detection of Explosives*, Ottawa, Ontario, Canada, June 6–10.

Walker, C. J., Cullum, H. E., and Hiley, R. W. (2001). An environmental survey relating to improvised and emulsion/gel explosives. *J. Forensic Sci*., **46(2)**:254–267.

Warren, D., Hiley, R. W., Phillips, S. A., and Ritchie, K. (1999). Novel technique for the combined recovery, extraction and clean-up of forensic organic and inorganic trace explosives samples. *Sci. Justice*, **39**:11–18.

Yinon, J. (Ed.) (2007). *Counterterrorist Detection Techniques of Explosives*. Amsterdam: Elsevier.

Yinon, J., and Zitrin, S. (1981). *The Analysis of Explosives*. Oxford, UK: Pergamon Press.

Yinon, J., and Zitrin, S. (1993). *Modern Methods and Applications in Analysis of Explosives*. New York: Wiley.

CHAPTER 3

Analysis of Fire Debris†

JOHN J. LENTINI

Scientific Fire Analysis, LLC, Big Pine Key, Florida

Summary The isolation and identification of ignitable liquid residues (ILRs) from fire debris are critically important aspects of arson investigation. In this chapter we describe common techniques for the isolation and identification of ILRs. Analytical procedures have become more sensitive, and results of testing play a very important role during litigation in a criminal or civil court. Quality control is an important component in fire debris analysis. Reports of findings should be written in a scientific manner, describing the fire under investigation, evidence handling, a description of the evidence and where it was collected, the isolation procedure and what testing was done with what kind of equipment, observations made, and conclusions, with a discussion of the meaning of the results.

3.1 Introduction 42
3.2 Evolution of separation techniques 43
3.3 Evolution of analytical techniques 47
3.4 Evolution of standard methods 49
3.5 Isolating the residue 51
  3.5.1 Initial sample evaluation 51
  3.5.2 ILR isolation method selection 51
  3.5.3 Solvent selection 54
  3.5.4 Internal standards 54
  3.5.5 Advantages and disadvantages of isolation methods 56
3.6 Analyzing the isolated ILR 56
  3.6.1 Criteria for identification 63
  3.6.2 Improving sensitivity 90

---

†This chapter is essentially an updated version of the chapter entitled "Analysis of Ignitable Liquid Residues" that appears in the author's textbook, *Scientific Protocols for Fire Investigation* (CRC Press, 2006), and is used here with permission.

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

3.6.3 Estimating the degree of evaporation 95
3.6.4 Identity of source 98
3.7 Reporting procedures 101
3.8 Record keeping 102
3.9 Quality assurance 105
Conclusion 105
References 106

3.1 INTRODUCTION

The laboratory analysis of fire debris is one of the most important hypothesis tests that can be performed in an investigation, especially when the investigator forms a hypothesis that the fire was set using ignitable liquids. It has been widely acknowledged that making a determination that ignitable liquids were used in a fully involved compartment is not valid based on visual observation alone (NFPA 921, 2011). A laboratory analysis is necessary. Even when the compartment is not fully involved, and there is an exceptional burn pattern similar to the one shown in Figure 3.1, the investigator still needs to determine the identity of the ignitable liquid.

In the past, the laboratory analysis was referred to as “the icing on the cake,” because by the time samples were collected, the fire investigator had already decided what caused the fire, and the stated purpose of the laboratory analysis was merely to help determine the identity of the flammable or combustible liquid used to start the fire. Investigators were accustomed to receiving negative reports from their laboratory, even when they “knew” that a fire had been intentionally set with ignitable liquids.

In those days, the term *ignitable liquid* had not yet been coined. *Flammable or combustible liquids* were generally referred to as *accelerants*, even by people in the laboratory who had no idea how such liquids may have been used. Many of the findings were, in fact, false negatives, because laboratory methods were not sensitive enough to detect ignitable liquid residues at low levels. The sensitivity has improved dramatically since this author began distilling fire debris in 1974, so much so that it is now possible to detect the petroleum products that are a natural part of the background. Many fire investigators, however, still distrust negative reports from laboratories, based on this earlier experience.

The development of analytical procedures has paralleled the development of standards for fire debris analysis. The first vague outline of a standard was not published until 1982 (AA Notes, 1982), and prior to that, analysts would report that a sample contained “an oily liquid” that exhibited “sufficient similarities” to a known sample of gasoline or kerosene or diesel fuel. What was sufficient, however, was not clearly defined until there were standards.

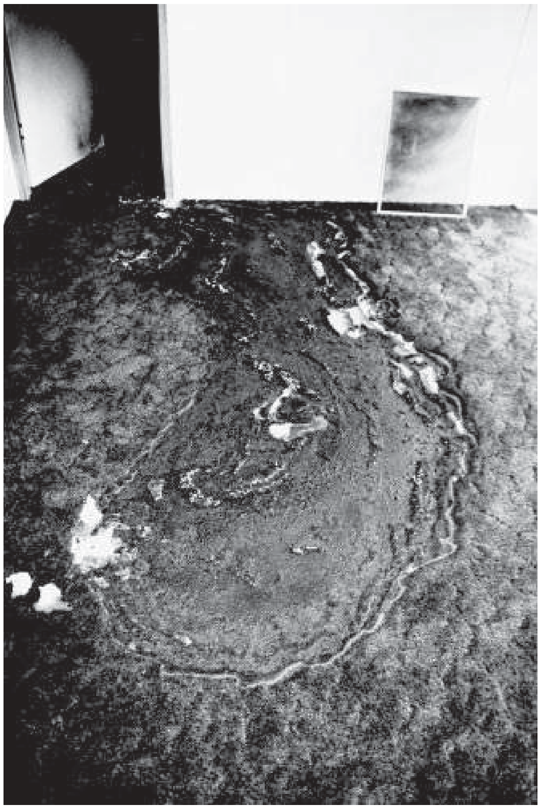

Figure 3.1 "Obvious pour pattern." The only surface that was burned in this mobile home was the floor. There were no furnishings in the house. The carpet was tested and found to be positive for the presence of a medium petroleum distillate such as mineral spirits or charcoal lighter fluid. This is a unique case where visual observation alone can lead to valid conclusions about what caused the pattern. *(See insert for color representation.)*

3.2 EVOLUTION OF SEPARATION TECHNIQUES

Paul Kirk reported in 1969 that the normal manner of isolating a liquid accelerant from other materials is "to distill the liquid from a solid residue in a current of steam." Kirk reported that the distillate would then be subjected to fractional distillation, flash point, refractive index, or density determination, but that a better procedure was to employ a gas–liquid chromatograph. Kirk stated, "From all of these laboratory procedures, the most important single piece of information that is made available is that a foreign flammable liquid was present at the fire scene. This alone is strong evidence for arson, at least after the possibility of accidental placing of the liquid is eliminated" (Kirk, 1969). The operative word here is *foreign*.

Figure 3.2 shows a steam distillation apparatus for the isolation of ignitable liquid residue from fire debris. In this classical separation technique, the debris is

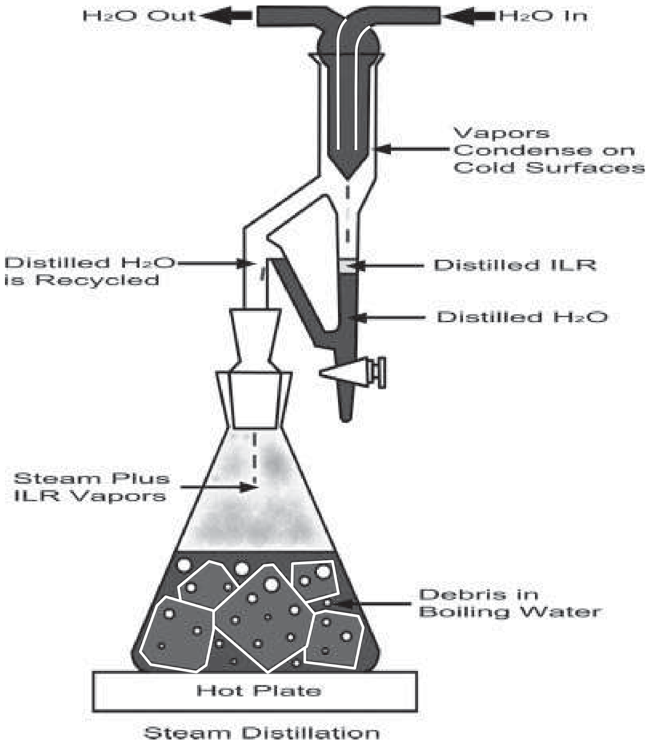
H₂O Out
H₂O In
Vapors Condense on Cold Surfaces
Distilled H₂O is Recycled
Distilled ILR
Distilled H₂O
Steam Plus ILR Vapors
Debris in Boiling Water
Hot Plate
Steam Distillation

Figure 3.2 Steam distillation apparatus. The sample is boiled. Vapors condense on a "cold finger" and fall into the trap, which allows the water to recycle while the immiscible oil layer builds up on top of the water column.

covered with water and boiled, and the steam and other vapors are condensed in a trap that recycles the water and allows any nonmiscible oily liquids to float on top. There is actually a visible layer of liquid isolated from the sample. More often than not, this layer consisted of a drop or two, or simply a rainbow film on top of the column of water. This could be extracted with a solvent and analyzed, but even so, steam distillation was not a very sensitive technique. If the sample did not exhibit a detectable petroleum odor, steam distillation was almost always ineffective at isolating any ignitable liquid. Despite its lack of sensitivity, steam distillation did have the advantage, given a sufficiently concentrated sample, to produce a visible layer of liquid that could be shown to a jury.

In 1969, Kirk suggested that the debris could be heated in a closed container and the internal gaseous phase could be sampled and analyzed by gas chromatography, but indicated that he was unaware of its use in routine analyses (Kirk, 1969). By the mid-1970s, this technique, known as *heated headspace* (Figure 3.3), was used routinely but had sensitivity limitations similar to those of steam distillation and, because of the low volatility of higher-molecular-weight compounds, was ineffective at isolating the heavier components of common combustible liquids

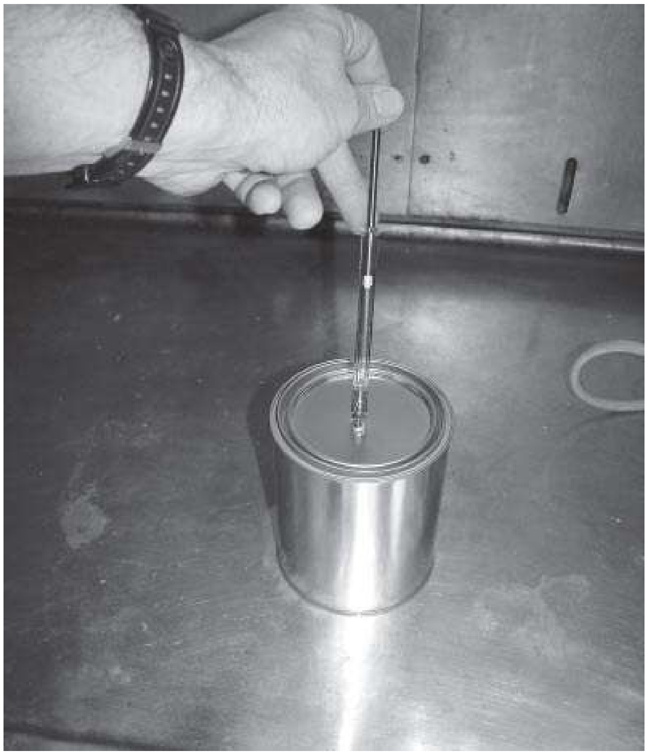

Figure 3.3 Using a gastight syringe to withdraw a sample of headspace from a fire debris sample. The headspace sample, about 500 μL, is injected directly into the GC–MS injection port.

such as diesel fuel. Heated headspace (Kirk called it a "shortcut") is still used as a screening tool in some laboratories today.

The first quantum jump in sensitivity took place in 1979, when two chemists from the Bureau of Alcohol and Firearms Philadelphia laboratory reported in the *Arson Analysis Newsletter* on the separation process known as *dynamic headspace concentration*. They used a dry nitrogen purge and a vacuum pump to draw ignitable liquid vapors from a heated sample through a Pasteur pipette filled with activated coconut charcoal (Chrostowski and Holmes, 1979). The vapors were rinsed off the charcoal using carbon disulfide and analyzed by gas chromatography. Over the next decade, the apparatus for conducting this type of analysis resulted in the publication of many articles describing newer and more wonderful apparatuses. Dynamic headspace concentration is still a recognized analytical technique, but because it is both destructive and complicated, it is not used by many laboratories.

In 1982, Juhala made the case for using *passive headspace concentration* on fire debris, wherein an adsorbent package is placed in the sample container and heated up. He used charcoal-coated copper wires and Plexiglas beads, and reported an increase in sensitivity of two orders of magnitude over distillation and heated headspace analysis (Juhala, 1982). Many laboratories had just completed setting up their dynamic systems; however, adoption of passive headspace concentration took some time, but gradually its advantages made it the dominant method of separation. Dietz reported on an improved package for the adsorbent called C-bags (Dietz, 1993), but these quickly gave way to activated carbon strips (ACSs), which required much less preparation. In 1993, Waters and Palmer reported on the essentially nondestructive nature of ACS analysis, performing up to five consecutive analyses on the same sample with little discernible change and no change in the ultimate classification of the residue (Waters and Palmer, 1991). This separation technique is the method of choice in most laboratories today. Figure 3.4 is a conceptual drawing of the procedure for passive headspace concentration, and Figure 3.5 is a photograph of a typical adsorption device, which consists of a 10 mm × 10 mm square of finely divided activated charcoal impregnated on a polytetrafluoroethylene (PTFE) strip. The technique was actually adapted from the industrial hygiene industry. Charcoal disk badges are worn by employees to determine their exposure to hazardous chemicals. The charcoal adsorbs a wide variety of organic compounds.

Solid-phase microextraction (SPME) represents yet another kind of passive headspace concentration technique. The SPME fiber is a more active adsorber of most ignitable liquid residues than is an activated carbon strip. Exposing an ACS to the headspace of a sample at elevated temperatures for 16 h allows for the isolation of less than 0.1 μL of ignitable liquid residue if there is no competition from

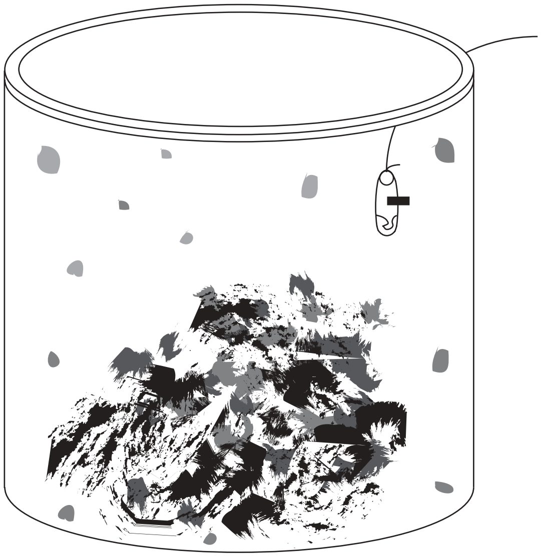

Figure 3.4 Schematic drawing of passive headspace concentration using an activated carbon strip. Vapors are produced by heating the container with debris to 80°C. The ACS adsorbs the vapors for 16 h, then is rinsed with diethyl ether spiked with 100 ppm, and the resulting solution is analyzed by GC–MS.

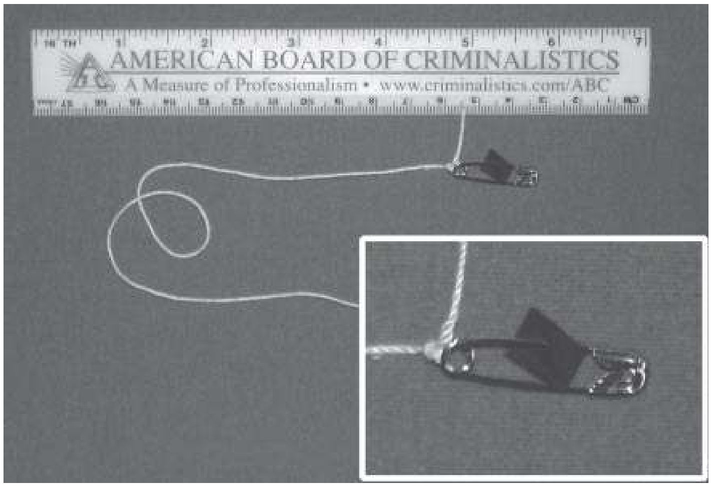
AMERICAN BOARD OF CRIMINALISTICS
A Measure of Professionalism • www.criminalistics.com/ABC

Figure 3.5 Close-up view of an activated carbon strip.

the substrate. A SPME fiber can accomplish the same task in 20 min or less. The advantages and disadvantages of the various separation techniques are discussed later in the chapter.

3.3 EVOLUTION OF ANALYTICAL TECHNIQUES

As the ability of the separation step to isolate smaller quantities of ignitable liquid residue improved, the sensitivity of the analytical instruments improved as well. In the 1950s and 1960s, extracts were analyzed by infrared (IR) or ultraviolet (UV) spectroscopy (Midkiff, 1982), but because most extracts were mixtures, these techniques were neither very sensitive nor very specific. The IR spectrum of gasoline looks very much like the IR spectrum of kerosene. Gas chromatography (GC) using pattern recognition techniques became the analytical method of choice beginning in the late 1960s. GC is actually a separation technique rather than an identification technique, but unlike separating the ignitable liquid residue from the sample matrix, GC works by separating similar compounds in an extract from each other.

In the 1970s, GC columns, the engine that makes the technique work, were glass or metal tubes, $\frac{1}{4}$ in. in diameter by 6 to 10 ft long. Chemists typically purchased empty columns and packed the columns themselves, using a coated powdery substance (stationary phase). It was known that $\frac{1}{8}$-in. columns provided better resolution than $\frac{1}{4}$-in. columns, but these had to be made from metal, and the chemist could not see inside the tube to check for gaps. These columns were usually purchased already packed. As the column manufacturers experimented with narrower and narrower columns, they went back to drawn-glass tubes coated on the inside with the oily stationary phase. There were problems with these early capillary columns, not the least of which was the forensic science community's resistance to change. As the bugs were worked out, capillary columns became the standard choice, but as late as 1990, packed columns were contemplated in ASTM E1387, *Standard Test Method for Flammable or Combustible Liquid Residues in Extracts from Fire Debris Samples by Gas Chromatography*. In the 1995 edition the standard "recommended," but did not require, capillary columns. The 2001 edition requires "A capillary, bonded phase, methylsilicone or phenylmethylsilicone column or equivalent. Any column length or temperature program conditions may be used provided that each component of the test mixture is adequately separated" (ASTM E1387-01, 2001). Today, very few laboratories use packed columns.

Gas chromatography detectors originally measured the change in thermal conductivity (TC) of the effluent from the column. Flame ionization detection (FID) improved the sensitivity by a couple of orders of magnitude over TC detectors. In a flame ionization detector, there is a hydrogen flame burning between two charged plates. As indicated by the flow of current between the two plates, the electrical conductivity changes when a hydrocarbon compound comes through the hydrogen flame and is burned.

Some laboratories were using gas chromatography–mass spectrometry as early as 1976 (GC–MS) (Stone, 1976). At that time, mass spectrometers were expensive, not terribly reliable, and required a computer (this was before the days of the personal computer, when computers took up half of the room), and GC–MS was the exception rather than the rule. With the earlier instruments, the operator watched a stripchart recorder and pushed a button to collect a mass spectrum when a peak indicated that a compound was coming off. This was a very labor-intensive process. There were people who argued that chemists had an obligation to use the best technology available, and the advantages of GC–MS over GC–FID required that MS be used. One of the leading proponents of GC–MS was Jack Nowicki, who also noted that GC–MS would make the previous accelerant classification system obsolete (Nowicki, 1990). He was eventually proved correct. Most laboratories stayed with the FID methods because of the difficulties with implementing GC–MS and because they felt comfortable with their ability to read patterns using FID alone. By the early 1980s, mass spectrometry was still expensive, but its use had become more widespread in fire debris analysis, particularly in the better-funded laboratories. Public laboratories acquired GC–MS instruments for use in drug identification, and this was another reason that they became available for fire debris analysis. The instrumentation of the 1980s was more automated and could collect a mass spectrum several times per second, even if no peak was eluting. This resulted in a much more efficient process, but the data files were very large. Today's GC–MS, collecting data every tenth of a second, uses sophisticated software to keep the file size to around a megabyte, an amazing feat considering that 18,000 spectra may be collected during a single run.

In 1982, Martin Smith published an article about a technique he had developed, called *mass chromatography*, which utilized a computer to separate the mass spectral signals according to the functional groups of the compounds that produced them. This technique allowed chemists to view many simple and easy-to-recognize patterns as opposed to looking at one large complicated pattern (Smith, 1982).

Today this powerful analytical tool is known as *extracted ion profiling* or *extracted ion chromatography* and forms the basis of most identifications.

The development of the personal computer made it possible for average laboratories to control a mass spectrometer, and the instrument manufacturers responded to the demand by producing benchtop models with increasing sensitivity and extraordinary robustness. The *quadrupole mass filter*, which is at the heart of the most popular mass selective detectors, has no moving parts. Only those parts of the instrument required to maintain a vacuum are subject to mechanical problems.

3.4 EVOLUTION OF STANDARD METHODS

As fire debris analysis technology improved, so did the approach of the forensic science community to the problems encountered by analysts. *Arson Analysis Newsletter* continued publication through 1984, and much valuable information was exchanged. Through this informal journal, forensic scientists analyzing fire debris had a means of communicating with each other that was unavailable to scientists in many other disciplines. In 1982, based on work conducted at the Center for Fire Research at the National Bureau of Standards (NBS) (now the NIST) and the ATF National Laboratory, an *accelerant classification system* was published (AA Notes, 1982). Not only was there a description of five classes recognized as "usually identifiable by GC–FID patterning alone when recovered from fire debris," but the authors (who were not named in the publication but are believed to include Philip Wineman and Mary Lou Fultz) also published minimum requirements for class identification. This was the first time that anyone explained what "sufficient similarities" should mean and was a watershed moment in the history of fire debris analysis. Although the original publication of the classification system stated that the final report was not yet available and that the results of the evaluation would be printed in a future issue of the *Arson Analysis Newsletter*, that never happened. The classification system was used informally for the next six years.

The International Association of Arson Investigators (IAAI) has had, almost since its beginning, a standing committee of forensic scientists and engineers that it called upon to advise fire investigators about laboratory analysis issues. In 1987, IAAI President John Primrose approached the Forensic Science Committee and requested that it produce a position paper on what should appear in a laboratory report. It quickly became apparent to members of the committee that in order to prescribe the contents of a report, it would first be necessary to set down an acceptable method of analysis. Four sample preparation techniques (steam distillation, headspace analysis, solvent extraction, and dynamic headspace–purge and trap) were the separation techniques described. Gas chromatography with flame ionization detection, photo ionization, or mass spectral detection was required. Although the publication was called a "guideline," it contained the following sentence: "Unless a petroleum distillate has been identified by the pattern recognition techniques described below, it has not been sufficiently identified" (IAAI Forensic Science Committee, 1988). The guidelines then reproduced the NBS–ATF classification and identification scheme and described how some materials would not fall within the guidelines. Isoparaffinic hydrocarbons were cited as one example of an ignitable liquid not described in the guidelines. At that time, the only place that a fire debris chemist would be likely to encounter isoparaffinic hydrocarbons was in Gulflite charcoal starter fluid. With the introduction of the IAAI guidelines, fire debris chemists became accustomed to the idea that they needed to follow standard methods. In the private sector, laboratories advertised to the membership of the IAAI, their main source of clients, that they followed the guidelines. Laboratories that did not follow the guidelines lost business.

In 1990, ASTM Committee E30 on Forensic Sciences took the IAAI guidelines and formulated them into six different standards for the preparation, cleanup, and analysis of fire debris extracts. ASTM E1387, the *Standard Test Method for Flammable or Combustible Liquid Residues in Extracts from Fire Debris Samples by Gas Chromatography*, was published originally in 1990. In 1995, the cumbersome phrase "flammable or combustible liquid" was changed to "ignitable liquid." In 1993, Committee E30 enlisted the aid of Martin Smith, Jack Nowicki, and several other prominent chemists to draft a *Guide for Fire Debris Analysis by Gas Chromatography–Mass Spectrometry*. The guide was revised in 1997, and in 2001 it was "promoted" to a standard test method.

The ignitable liquid classification scheme from 1982 was updated in an attempt to keep up with the ever-changing output of the petrochemical industry. New products were constantly being brought on line, including "environmentally friendly" alternatives to solvents such as mineral spirits and fuels such as charcoal lighter fluid. When ASTM E1387 was first published, a "Class 0" was added to the original classification scheme to account for the liquids that did not fit into one of the five original classes. Further classification within Class 0 was, however, possible, so Classes 0.1 through 0.5 appeared in ASTM E1387-95. By the time the next revision was due, Class 0.6 had been created for de-aromatized distillates, but the committee realized it was time for a change, as the miscellaneous classes now outnumbered the original classes. In 2000, the system was completely redesigned, resulting in nine differently named but no longer numbered classes, with subdivisions of light (C_4–C_9), medium (C_8–C_{13}), and heavy (C_8–C_{20+}) in eight of those nine classes (ASTM E1618-10, 2010).

In 2009, after much discussion, Committee E30 on Forensic Sciences made the decision that analysis of ILRs by GC alone no longer represented the "best practice." ASTM E1387 was allowed to expire and become a "historical" standard. Most fire debris analysts had long ago made the transition to GS-MS. As of this writing, gas chromatography-mass spectrometry is the only generally accepted method for analyzing ignitable liquid residues.

The last 30 years have seen dramatic improvements in separation technology, in analytical technology, and in the scientific community's approach to fire debris analysis. In 1999, the U.S. Department of Justice Office of Law Enforcement Standards produced a report entitled *Forensic Sciences: Review of Status and Needs*, compiled by more than 40 eminent forensic scientists. In reviewing the state of the art, fire debris analysis was described as "a sub discipline of trace analysis that is in good standing because there is sufficient published work on the analysis and interpretation of the material involved. Standard guides for the examination and interpretation of chemical residues in fire debris have been published through the consensus process of ASTM Committee E30 on Forensic Science. These standardization documents are often quoted in the scientific literature, helping to meet the requirements of the legal community" (U.S. Department of Justice, 1999). Most of the other forensic disciplines discussed in this review were reported as still needing standardization and/or validation of standard methods. Fire debris analysts can point to a history of standardization that existed even before the *Daubert* court made it a necessity.

3.5 ISOLATING THE RESIDUE

3.5.1 Initial Sample Evaluation

Once the receipt of a sample has been documented, and the chain of custody protected (as described in ASTM E1492-05, 2005) the first critical step in the analysis of fire debris is the selection of a separation technique. Choosing an inappropriate technique could result in a false negative, a misidentification, or the destruction of evidence. The first step in this selection process (and the first step in any chemical analysis process) is to *look at the sample*. One purpose of looking at the sample is to ensure that it is what it purports to be. Sample characteristics will determine the most appropriate method for isolating any ignitable liquid residues (ILRs) that may be present. Once the visual examination has taken place, the next step is a "nasal appraisal." Occupational safety experts will no doubt frown on this recommendation, but it can be done carefully. There is no need to put one's nose in the can, even though the analyst can be reasonably certain that the fire investigator who collected the sample has already done exactly that. Unless the sample is a liquid sample for comparison purposes, it can be safely appraised by removing the lid, and waving the hand gently over the top of the sample to see if it exhibits any obvious odors. If there is an odor present, it becomes possible to do a rapid and accurate analysis by removing a small piece of the sample and extracting it with solvent. If the odor is very strong, it is advisable to remove a small piece of the sample and place it into a separate container for analysis.

3.5.2 ILR Isolation Method Selection

Solvent extraction according to ASTM E1386 is an appropriate method for rinsing out empty containers, for extracting small aliquots of samples with a high concentration of ILR, and for isolating residues from very small samples. Not every investigator who comes to the laboratory will have the experience and knowledge to know how to find the best samples, and it is not unusual for an inexperienced investigator to bring in a sample in a film canister.

The vast majority of samples, however, are likely to be samples of burned building materials, floor coverings, and furnishings that do not exhibit a strong odor and are best analyzed by passive headspace concentration, as described in ASTM E1412 (ASTM E1412-07, 2007). This technique is essentially nondestructive. If the analyst decides to use another technique later, running passive headspace concentration will not interfere with that.

Other methods of isolation have been studied thoroughly over the last 30 years, and although they have some utility, none match the advantages of passive headspace concentration using an activated carbon strip. Headspace sampling, described in ASTM E1388, which involves warming the container and sampling the vapors in the headspace directly and then injecting those vapors into the gas chromatograph, is a good screening technique, but it does not result in the production of an archiveable extract, nor does it detect compounds much heavier than C_{15} (ASTM E1388-05, 2005). Dynamic headspace concentration, described in ASTM E1413, was useful in demonstrating the effectiveness of adsorption or elution as a valid approach to ILR isolation, but it is destructive, requires far more attention than passive headspace concentration, and the apparatus can be cumbersome and finicky. It is no more sensitive than passive headspace concentration, and it is possible for *breakthrough*, loss of analyte out the effluent end of the tube, to occur (ASTM E1413-07, 2007). The only advantage that it offers is speed. If the results can be reported in 24 h, there is no advantage at all. Most laboratories (unfortunately) take several days or weeks to report on a fire debris analysis, due to sample backlogs. Solid-phase microextraction is another alternative, but like dynamic headspace concentration, it is very labor intensive and, like headspace sampling, does not have the potential to produce an archiveable sample that can be analyzed again (ASTM E2154-01, 2008) Using passive headspace concentration results in a solution that can be injected many times, and when the carbon strip is left in the solution, it will gradually readsorb ignitable liquid residues as the eluting solvent evaporates. (Juhala reported the readsorption by small portions of the activated charcoal that fell off his Plexiglas beads in 1982.) The solution can be reconstituted years later if a second look at the sample is desired. Because of the transient nature of many fire debris sample containers, the archived activated carbon strip is often the best evidence after a few years have passed.

The only equipment required for passive headspace concentration is a convection oven, vials, ACS strips, and a solvent dispenser. Caseload will determine the required oven size. In our laboratory, the convection oven can hold up to ten 1-gallon cans and more than twenty 1-quart cans.

Every laboratory should optimize the parameters in its ACS procedure to make sure that they are getting the best results possible. “Good” results from ACS are those where the chromatogram of the concentrated headspace vapors of a standard closely matches the chromatogram of that same standard in the eluting solvent.

A 10 mm × 10 mm carbon strip is the minimum size recommended. This 100-mm^2 strip can easily accommodate the headspace vapors from 10 μL of any ignitable liquid placed on a Kimwipe in a 1-quart can. It is possible to overload carbon strips, and this results in the preferential adsorption of heavier hydrocarbons over light hydrocarbons and of aromatics over aliphatics, but this effect is generally not large enough to affect the identification. Usually, samples that are capable of overloading the carbon strip will exhibit a strong odor, and the analyst can take an aliquot of the sample or reduce the analysis time.

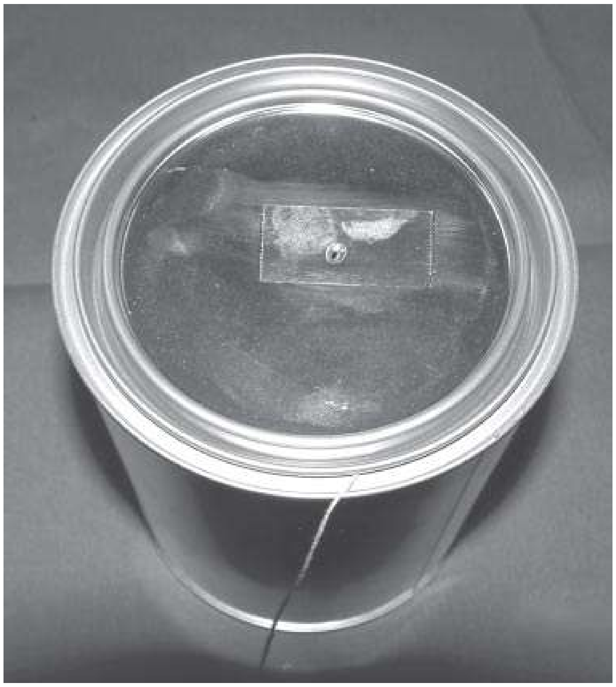

Figure 3.6 One-quart can equipped with a pressure relief device, a short strip of cellophane tape over a small hole pierced in the lid.

For samples with a high water content, there exists a danger that the vapor pressure in the sample container will cause the lid to pop off. This has the potential to contaminate the sample oven. A pressure relief device is easy to construct for such samples. Puncture a small hole in the lid, and cover with cellophane tape. Figure 3.6 shows a can so equipped.

The adsorption time of a typical ACS procedure is 16 h. The analyst is encouraged to experiment with different adsorption times and temperatures, with the goal of finding a balance between the maximum recovery and the minimum time necessary. Sixteen hours is convenient because the samples can be put in the oven at 4:00 P.M. and taken out at 8:00 A.M. the next day. One of the major advantages of ACS is that it requires very little attention from the analyst. Once the strip is in the can and the can is in the oven, nothing is going to happen until it is time to take the strip out of the can, put it in the vial, and add the eluting solvent. At this point in the analysis, however, the analyst must be extremely attentive to the procedure. Once the carbon strips are taken out of the sample container, they have an identical appearance. It is at the point of placing the strips into their vials that an unrecoverable error can occur—a strip can be misidentified. The analyst should not allow himself or herself to be distracted by phone calls or other people in the lab. The operative word here is *focus*.

Many laboratories use a pre-concentration step, wherein they add approximately 500 μL of solvent to the strip in a vial; then after the strip has had a chance to equilibrate, it is removed and the solvent is evaporated down to 100 μL or so. This will result in a fivefold increase in concentration, but that increase can usually be achieved electronically with very little loss of signal or increase in noise. The first analysis, in this analyst's view, does not require the pre-concentration step. That can be accomplished at a later time, if necessary, but pre-concentration runs the risk of skewing the results if not done very carefully. The only safe way to evaporate the solvent is to blow a stream of dry nitrogen over it. Heating the solvent is a bad idea, because lower-molecular-weight compounds are likely to be lost.

3.5.3 Solvent Selection

The solvent used for the elution is another critical choice, not so much in terms of the quality of results but in terms of the analyst's quality of life. The most popular eluting solvent is carbon disulfide, a highly toxic, carcinogenic, teratogenic, smelly, nasty liquid that will ignite upon exposure to boiling water. It does work very well to elute aromatics and aliphatics approximately equally from ACSs, but so does diethyl ether. Studies indicate that carbon disulfide is superior to diethyl ether, or to pentane, the other solvent recommended by ASTM Committee E30 (Newman and Dolan, 2001), but if standards are prepared using diethyl ether, the slight change in the chromatographic profile does not affect the identification. Armstrong and Lentini (1997) found only marginal differences between diethyl ether and carbon disulfide when applied to carbon strips exposed to 10-μL samples of ignitable liquid residues. Comparing the relative health risks makes diethyl ether an obvious choice for this analysis.

Carbon disulfide was originally selected as a fire debris solvent because of its high desorption efficiency and its relatively quiet signal when passing through a flame ionization detector. When using a mass spectral detector, the advantage of its low signal disappears because the detector is turned off while the solvent is passing through.

Some concerns have been expressed about the capability of diethyl ether to form explosive peroxides, but that will not occur if the ether is kept in a refrigerator and used on a regular basis. Explosions of cans of ether have only been reported when those cans have been allowed to sit for years, unused, in the back of an unrefrigerated stockroom. Carbon disulfide, diethyl ether, and pentane are all highly flammable; however, with respect to fire, carbon disulfide poses the greatest risk, as it has the lowest ignition temperature and the broadest flammable limits (NFPA 325, 1994). Finally, carbon disulfide costs almost 10 times what the other solvents cost. A comparison of the properties of the three solvents recommended by ASTM E1412 is shown in Table 3.1.

3.5.4 Internal Standards

There are two places in the analysis of fire debris where the use of internal standards is appropriate. Addition of an internal standard to the sample itself allows the analyst to develop at least a qualitative feel for the "tenacity" of the sample and for

TABLE 3.1 Elution Solvent Comparison[a]

	Carbon Disulfide	Diethyl Ether	Pentane
Flash point	−30°C (−22°F)	−45°C (−49°F)	−40°C (−40°F)
Lower elution limit (LEL) (vol% in air)	1.0	1.9	1.5
Upper elution limit (UEL) (vol% in air)	50.0	48.0	7.8
Specific gravity	1.3	0.7	0.9
Boiling point	40°C (115°F)	35°C (95°F)	36°C (97°F)
Autoignition temp.	90°C (194°F)	180°C (356°F)	260°C (500°F)
Exposure limit, TWA[b]	4 ppm	400 ppm	600 ppm
Exposure limit, STEL[c]	12 ppm	500 ppm	750 ppm
Carcinogenic	yes	no	no
Teratogenic	yes	no	no
IDLH	500 ppm	19,000 ppm (LEL)	15,000 ppm (LEL)
FID signal	small	large	very large
Cost per liter[d]	$399.7.70	$44.25	$38.10

[a] Refer to your laboratory's hazard communication literature or MSDS for more complete information.
[b] Time-weighted average for an 8-h exposure.
[c] Short-term exposure limit (up to 15 min).
[d] Source: J.T. Baker, online catalogue, Aug. 18, 2007. All prices are for case quantities of Ultra Resi–analyzed grade.

the effectiveness of the isolation procedure. In our laboratory, this is accomplished by the addition of 0.5 μL of 3-phenyltoluene (actually, 20 μL of a 2.5% solution of 3-phenyltoluene in ether). In the eluting solvent, we use a second internal standard consisting of 0.1% (100 ppm) perchloroethylene. If the 3-phenyltoluene does not appear in the chromatogram, this means that we have an exceedingly tenacious sample (some samples come equipped with their own active sites, which can compete very effectively with the ACS for analyte molecules) and suggests that the tenacity of the sample might be the reason that the chromatogram appears so flat. If the perchloroethylene peak does not appear, or is significantly reduced in height, something has gone wrong with the injection.

Comparison of the signal from the sample to the perchloroethylene signal allows for a semiquantitative determination of the amount of ILR present. For 10-μL standards of known ignitable liquids isolated according to ASTM E1412, the two internal standard peaks are roughly on the same order of magnitude as the sample peaks. Essential blanks include the 3-phenyltoluene applied to filter paper and the blank strip eluted with the spiked solvent. Some analysts may perceive the danger of being accused of "contaminating" a sample, but this is easily overcome by having a proper blank in the file. The advantages of using internal standards far outweigh this disadvantage.

3.5.5 Advantages and Disadvantages of Isolation Methods

Two drawbacks have been cited for use of the ACS method, the time required to perform the adsorption, and the relative lack of sensitivity compared to SPME. If laboratories were in the habit of providing same-day service, the first argument might have merit. Turnaround times generally range from two days to two months. In that context, a 16-h versus a 15-min analysis time is meaningless. With respect to sensitivity, the ACS method is capable of routinely detecting 0.1 μL of ignitable liquid from a nontenacious background, which should be low enough. Our goal is to help the fire investigator understand whether a foreign ignitable liquid was present at the fire scene. We now have the ability to detect the solvent in polyurethane finish five years or more after it has been applied to a hardwood floor. We have no need to be more sensitive than that. A technique that is capable of adsorbing significant quantities of ignitable liquid residue in 15 min is also capable of becoming contaminated much more easily than a carbon strip that might take 16 h to come to equilibrium with a dilute sample.

The use of a "screening" technique for fire debris samples is an issue that each laboratory should address. There is usually not a need to screen samples, since an ACS separation is likely to be more sensitive. A sample that tests negative on a quick headspace analysis (per ASTM E1388) needs to be tested further anyway. If a request is made to look for light oxygenates (alcohols, acetone) or to get a ballpark estimate of analyte concentration (other than through a "nasal appraisal"), running a headspace can be useful. Screening techniques also allow for a swift, if less than definitive, result. For routine analyses, however, passive headspace concentration, conducted according to ASTM E1412, should be the norm. Solvent extraction, as described in ASTM E1386, is appropriate for sampling aliquots of very strong samples or for extracting very small samples or empty containers (ASTM E1386-10, 2010).

Steam distillation may be selected in the odd case where it is desired to produce a neat liquid extract of the fire debris. The benefit of this is that a vial of the liquid can be brought into a courtroom, shown to a jury, lit on fire on a Q-tip, and passed around. Because steam distillation is appropriate only on very concentrated samples, however, it is preferable to make sure that the sample is preserved, and the sample itself can be passed around for the jury to smell. For the most part, steam distillation is a technique whose time has come and gone. The last reapproval of an ASTM steam distillation standard took place in 2001, and the Committee on Forensic Sciences in 2006 took the decision to allow the standard to expire. A comparison of the advantages and disadvantages of the various isolation techniques is shown in Table 3.2.

3.6 ANALYZING THE ISOLATED ILR

Despite the improvements in separation and detection technology, the overall approach to identification of ignitable liquid residues is the same as it was in the early 1970s. A chromatogram from the sample is compared with chromatograms

TABLE 3.2 Comparison of ILR Isolation Techniques

Method	Advantages	Disadvantages
E1385, steam distillation	Produces a visible liquid, simple to explain	Labor intensive, destructive, not sensitive, requires expensive glassware
E1386, solvent extraction	Useful for small samples and empty containers, does not cause significant fractionation, useful for distinguishing heavy petroleum distillates from each other	Labor intensive, expensive, co-extracts nonvolatile substances, increased risk of fire, solvent exposure, destructive
E1388, headspace sampling	Rapid, more sensitive to lower alcohols, nondestructive	No archiveable sample, not sensitive to heavier compounds, poor reproducibility
E1412, passive headspace	Requires little analyst attention, sensitive, nondestructive, produces archiveable sample, inexpensive	Requires overnight sampling time
E1413, dynamic headspace using activated charcoal	Rapid, sensitive, produces archiveable sample, inexpensive	Labor intensive, subject to breakthrough, destructive
E1413, dynamic headspace using Tenax	Rapid, sensitive	Labor intensive, requires thermal desorption, no archiveable sample, destructive
E2154, solid-phase microextraction	Rapid, highly sensitive, useful for field sampling with portable GC–MS	Labor intensive, expensive, requires special injection port, reuse of fibers, no archiveable sample

from known standards, and the analyst determines whether there are “sufficient similarities” to make an identification. What has changed is the quantity of information available because of the increased resolution provided by capillary columns and the ability to obtain a mass spectrum up to 10 times per second, as well as reaching a consensus on the meaning of “sufficient.” So, although there is more information to compare, the technique is still one of pattern recognition and pattern matching.

An argument can be made that when one looks at a mass spectrum, one is looking at structural details rather than simply matching patterns, but the patterns still have to match. The same argument has been made about structural elucidation in the use of FTIR for drug identification. Although it would be nice to think that analysts routinely consider molecular structure, the day-to-day operation is one of pattern matching.

There exists a specific skill set that is required to compare chromatographic patterns. This skill set must be learned carefully and used routinely if it is to remain sharp. Pattern recognition is one “scientific” skill that has historically been problematic. Although fire debris analysts use the same equipment as drug analysts, the drug analyst typically looks for a single peak with a particular retention time. The fire debris analyst must compare an entire pattern of peaks produced by a sample to the pattern of peaks produced by a standard. This exercise includes a comparison of inter- and intragroup peak heights and can be quite complex. Although it is true that many drug analysts also conduct ignitable liquid residue analyses, a different set of skills is involved.

Fire debris analysis also involves a set of skills different from those employed by environmental scientists, who are typically trying to quantify the components of an oil spill or contaminants at a Superfund site. Environmental methods typically assume the presence of gasoline or other petroleum products, then look for benzene, toluene, ethylbenzene, and xylene (BTEX) to quantify the amount present. Unless they are trying to identify the source of the spill, environmental analysts are usually not employing the same skill set used by fire debris analysts.

If one thinks about the nature of many petroleum products and the processes that are going on when those products are isolated from debris samples, one can begin to understand why chromatographic patterns look the way they do. Many petroleum products are straight-run distillates, particularly the medium and heavy petroleum distillates. The overall pattern of these products is a Gaussian (bell-shaped) distribution of peaks, dominated by the normal alkanes. The patterns produced by kerosene and diesel fuel have been likened to a stegosaurus, because the chromatograms bear a passing resemblance to the dinosaur’s dorsal fin. (Actually, a spinosaurus’ sail back fin more closely resembles the pattern, but stegosaurus is the term of art.). A medium petroleum distillate produces the same pattern, shrunken and coming off early.

A fractionation process similar to distillation occurs in the isolation of an ignitable liquid residue from a sample. This is caused by the very low vapor pressure of ILR components above C_{18} and by the selective adsorption of the heavier hydrocarbons on complex substrates. Hydrocarbons up to C_{23} can be captured from diesel fuel placed on a noncompeting substrate such as filter paper. That same fuel placed on charred wood, however, may exhibit a pattern that ends at C_{18}. If compounds do not get into the air in the headspace, they will not be adsorbed onto the carbon strip. The tenacity of the sample needs to be considered, particularly when trying to distinguish between kerosene and diesel fuel.

This author learned pattern recognition the old-fashioned way: running dozens, then hundreds of standards, and learning what the patterns looked like. Today’s analysts are fortunate in that the ASTM standards give examples of many of the patterns that an analyst is likely to see in positive samples, and there exists a detailed compilation of literally hundreds of patterns available in a standard text (Newman et al., 1998). Both the ASTM standards and the *GC–MS Guide to Ignitable Liquids* provide sufficient information to allow the analyst to set up an instrument to provide patterns that look very much like the ones in the texts. Although these texts are an important resource, it is imperative that every fire debris analysis laboratory have its own library of ignitable liquid residues. This provides for patterns with exactly the

TABLE 3.3 Ions Used for Extracted Ion Chromatography

Ion	Compounds	Ion	Compounds
57	Alkanes	119	C_4 alkylbenzenes
83	Cycloalkanes	131	Methyl, dimethylindans
91	Toluene, xylenes	142	Methylnaphthalenes
105	C_3 alkylbenzenes	156	Dimethylnaphthalenes
117	Indan, methylindans		

same retention times and mass spectra with exactly the same fragmentation patterns. The in-house library is also a quality assurance tool, which lets the analyst know when there has been some drift in the instrument, when a sensitivity loss has taken place, and when it is time to run a new set of standards. Whatever approach is taken, it will take time to develop the ability to recognize ILR patterns.

The mass spectrometer provides the ability to simplify what can be very complex and confusing patterns. This is a result of the ability of the data analysis software to separate out only those peaks having particular ions present in the pattern. For example, if one wants to look at the alkanes, one only needs to obtain an extracted ion chromatogram for m/z (mass/charge ratio) 57, hereafter referred to as ion 57, and most of the balance of the components will disappear. It is thus possible to break down a total ion chromatogram (TIC) into its component parts—see Table 3.3.

Although there are more patterns to learn, they are simpler patterns and easier to remember. This approach to data analysis, known as *mass chromatography*, was first proposed by Smith in 1982. There are basically two ways to approach mass chromatography: the single-ion approach and the multiple-ion approach. Dolan (2004) has proposed referring to the single-ion chromatograms as *extracted ion chromatograms* (EICs) and multiple-ion chromatograms as *extracted ion profiles* (EIPs).

When working with software that scales the extracted ion chromatogram or profile to the tallest peak, some caution is advised when using multiple ions. To the extent that the second, third, and fourth ions contribute to a pattern, they tend to make it more complicated, thus defeating some of the purpose of extracting the ions in the first place. To the extent that these additional ions do not change the pattern, they may convince the analyst that he or she is seeing more than is actually present. Finally, to the extent that these additional ions are present at substantially lower concentrations, they would be better observed on their own rather than in the profile. Figure 3.7 is a comparison of ion 57 from a kerosene standard versus a profile based on ions 57, 71, 85, and 99. The profile is slightly more complicated, but since ion 57 is the base peak for almost every component in the chromatogram, the ion 57 profile is the tallest.

Another example is shown in Figure 3.8, which presents ions 128, 142, and 156, the naphthalenes, from a gasoline sample. In the top chart, the ions are combined into a profile, which although it gives the analyst an idea about the relative abundance of the three ions, shows very little detail for the dimethylnaphthalenes represented by ion 156. When the extracted ions are presented separately, the analyst still gets the quantitative data by reading the abundance numbers next to the

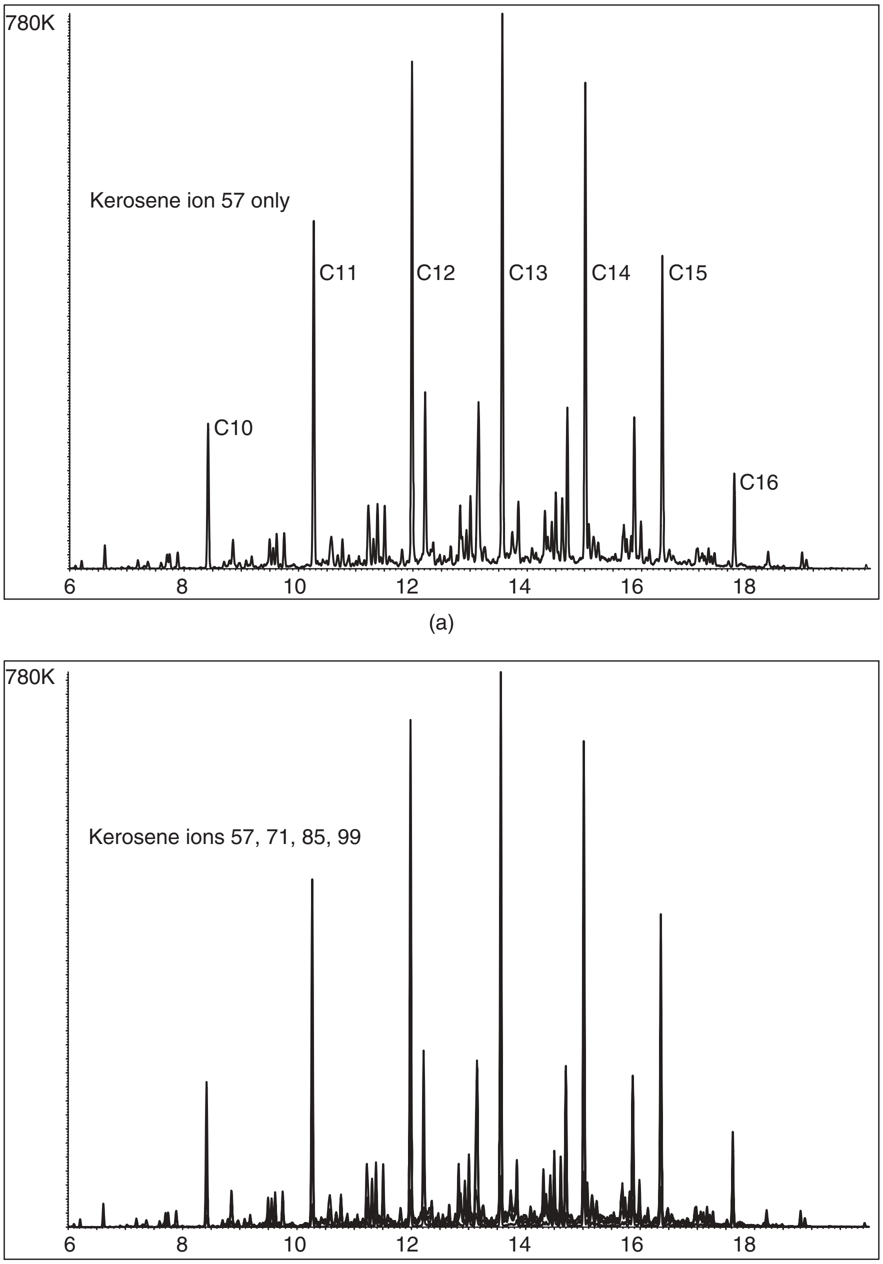

(a)

(b)

Figure 3.7 Ion profiling versus ion chromatography. Comparison of (a) the ion 57 chromatogram with (b) the ion profile combining ions 57, 71, 85, and 99 from a kerosene standard. The four ions in the profile are plotted in the merged format. Plotted individually, they are all very similar, with ion 57 presenting the tallest peaks.

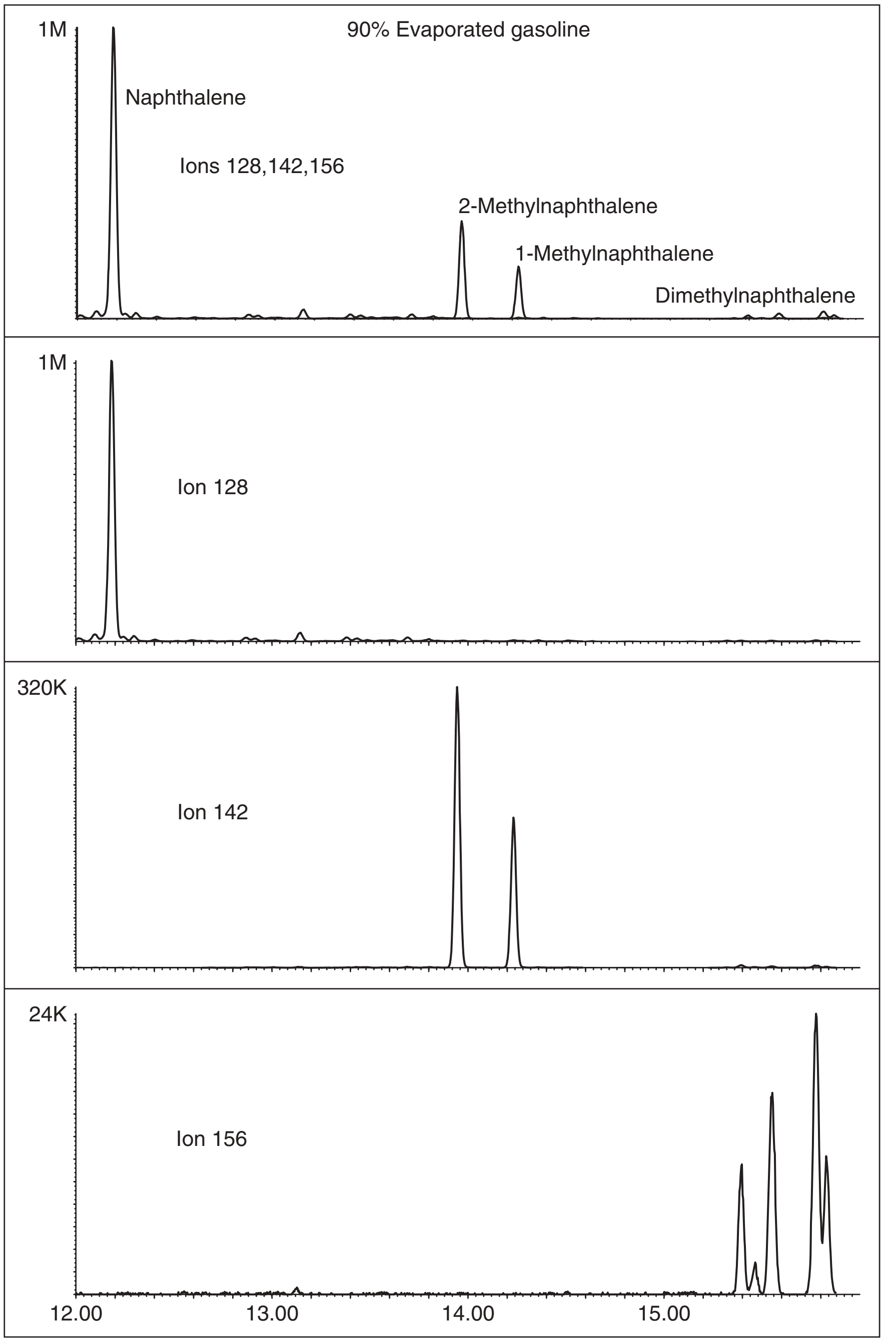

Figure 3.8 Ion chromatography versus ion profiling. Comparison of the ion profile of the naphthalenes from a gasoline standard (top chart). Compare with the detail provided by presenting the three ions (128, 142, and 156) separately.

Y-axis, but also gets to see the fine details in the shape of the peaks at the right side of the chart.

A similar example is shown in Figure 3.9(a). These charts show ions 91, 105, and 119 from a 75% evaporated gasoline standard. The effect is even more pronounced at 90% evaporation, shown in Figure 3.9(b), when the ion 91 peaks are smaller. Whether an analyst chooses extracted ion chromatography or extracted ion profiling is largely a matter of taste. It is easier for this author to use the simpler patterns

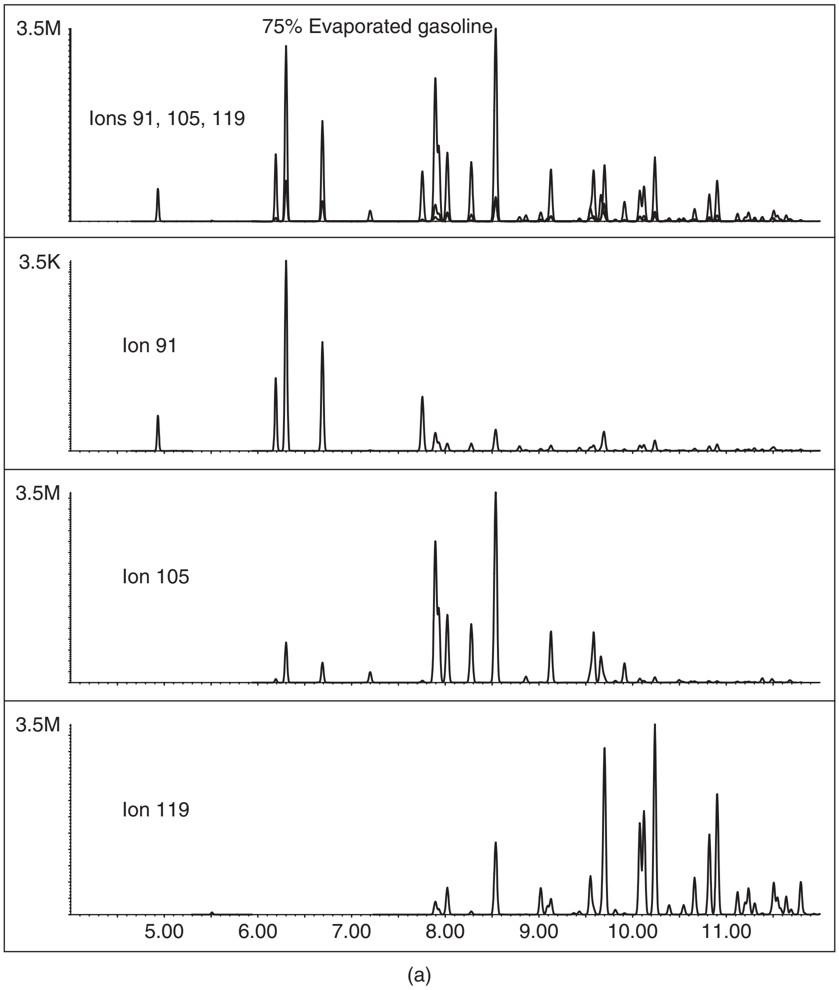

(a)

Figure 3.9 Ion profiling versus ion chromatography. (a) Comparison of ion profiles from 75%-evaporated gasoline. The top chart shows a combination of ions 91, 105, and 119. The next three charts show those ion chromatograms plotted independently. (b) Comparison of ion profiles from 90%-evaporated gasoline. The top chart shows a combination of ions 91, 105, and 119. The next three charts show those ion chromatograms plotted independently. Note the improvement in the level of detail presented for the xylenes.

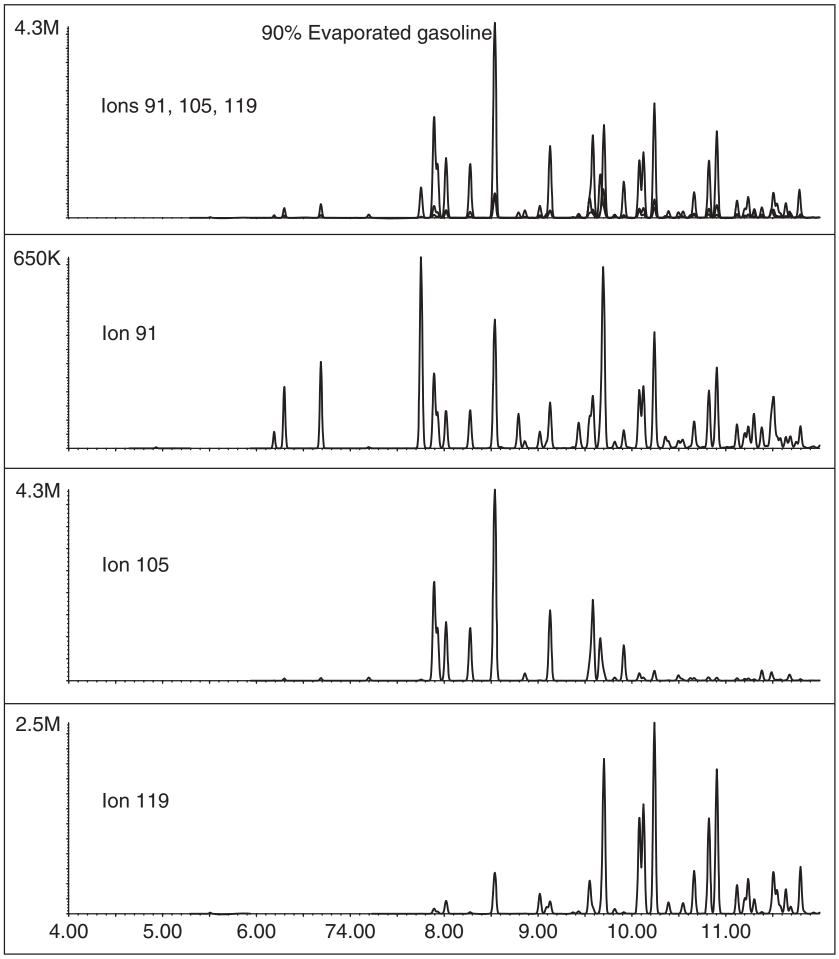

(b)

Figure 3.9 (*Continued*)

and avoid the eyestrain required to see the smaller peaks in the profiles. The typical set of six ion chromatograms used in our laboratory to document the presence of gasoline is shown in Figure 3.10, and the set of three ion chromatograms used for distillates is shown in Figure 3.11. Details of each are discussed below.

3.6.1 Criteria for Identification

Most of the chromatograms that an analyst uses to make a positive identification of an ignitable liquid residue will match (exhibit “sufficient similarities” to the standard) at the level of the TIC. Samples with low concentrations of ILR, high backgrounds, or both, can sometimes yield a positive identification for ILR if the

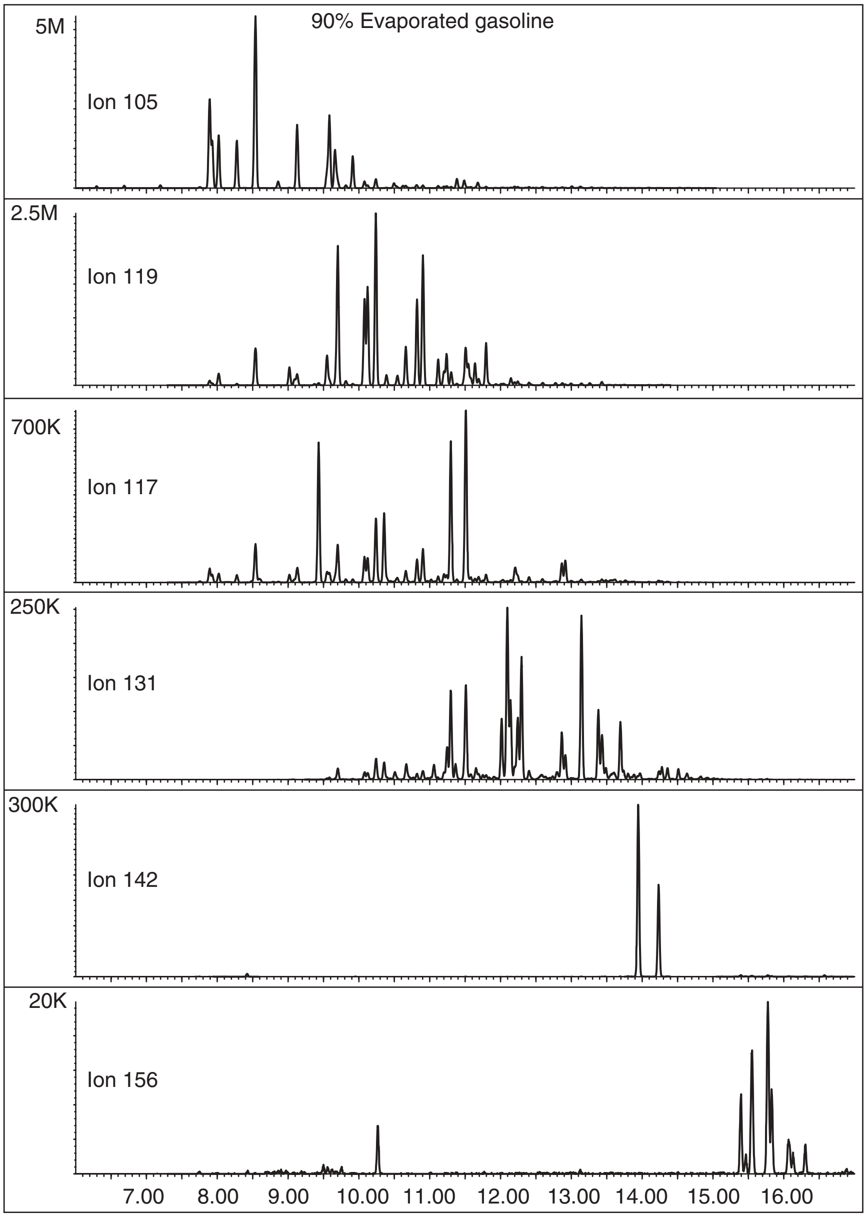

Figure 3.10 Typical set of six ion chromatograms used in our laboratory to document the presence of gasoline. Ions we use are 105 for C_3 alkylbenzenes, 119 for C_4 alkylbenzenes, 117 for indan and methyl indans, 131 for dimethyl indans, 142 for 2- and 1-methylnaphthalene, and 156 for dimethylnaphthalenes.

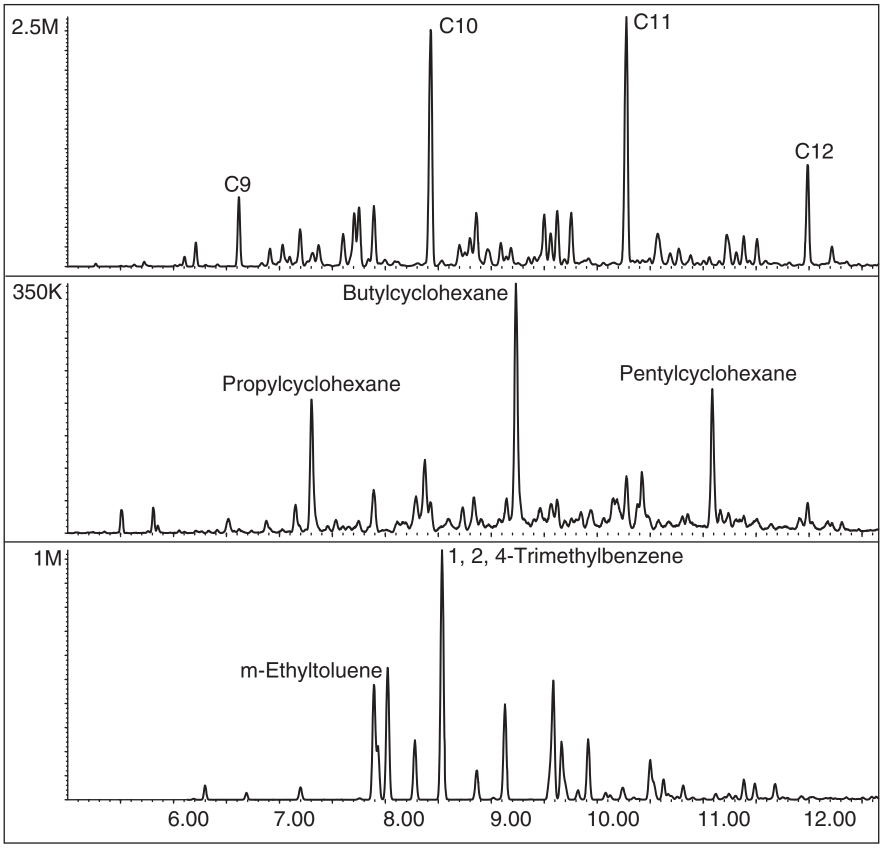

Figure 3.11 Typical set of three ion chromatograms used in our laboratory to document the presence of distillates. These are the ion 57, 83, and 105 chromatograms from Smokey Bear charcoal lighter, a medium petroleum distillate.

analyst is very careful. Usually, but not always, if the TIC from the sample does not match the TIC from the standard, the sample is likely to be negative (determined to contain no detectible ignitable liquid residue). Examples of cases where this is not true are presented following this general discussion of criteria for identification of an ILR in routine cases.

ASTM E1618 identifies eight classes of ignitable liquids identifiable by GC–MS:

1. Gasoline
2. Petroleum distillates
3. Isoparaffinic products
4. Aromatic products, including dearomatized distillates
5. Naphthenic–paraffinic products

6. *N*-Alkane products
7. Oxygenated solvents
8. Miscellaneous

With the exception of gasoline and oxygenated solvents, each of the classes above can be placed into one of three ranges: light (C_4–C_9), medium (C_8–C_{13}), and heavy (C_8–C_{20+}). This standard presents criteria for identifying the various kinds of compounds (alkanes, cycloalkanes, aromatics, and condensed ring aromatics) found in each one of the nine classes. The alkanes include both straight-chain and branched hydrocarbons, and can be extracted using ion 57. The cycloalkanes are mostly substituted cyclohexanes, which can be seen by extracting ion 83. Cycloalkanes also have a strong ion at 55, and as the length of the substituted alkane increases, ion 57 begins to dominate the mass spectra. *Aromatics* means alkyl-substituted benzenes with a single ring. These can be extracted using ions 91, 105, and 119. Ion 91 will show toluene and xylenes, ion 105 will show xylenes and C_3 alkylbenzenes, and ion 119 will show C_4 alkylbenzenes. There will be some overlap in the extracted ion chromatograms, as shown in Figure 3.9. *Condensed ring aromatics* refers to indans and naphthalenes. Indans have a five-membered ring attached to a benzene ring, and may be substituted. The naphthalenes that we usually see are naphthalene itself, 2- and 1-methyl naphthalene (written in that order because that is the order of elution from a nonpolar column), and the dimethyl naphthalenes. Naphthalene presents a single peak when ion 128 is extracted. There are two peaks for the methyl naphthalenes seen when ion 142 is extracted, and there are eight peaks in the dimethyl naphthalene chromatogram seen when ion 156 is extracted. The indans can be visualized by extracting ions 117 and 131.

Identification of Gasoline The composition of petroleum products as found in fire debris is influenced by three factors: crude oil parentage, the effects of petroleum refining processes, and the effects of weathering. Gasoline is the way it is largely because of the second of these influences, petroleum refining. Although all crude stocks contain aromatics, aliphatic hydrocarbons comprise the bulk of most crude oils. Because the aliphatics cause knocking in gasoline engines, the value of the crude stock is enhanced by *cracking*, making small molecules out of larger ones, and by *reformation* through a process of dehydrogenation. Toluene and xylenes are the most abundant compounds produced in these processes. Gasoline does contain numerous light aliphatics (as light as butane) present when gasoline is first pumped, but by the time it has been through a fire, most of the lighter ($<C_7$) aliphatics have evaporated. Consequently, toluene is usually one of the first tall peaks seen in a sample of gasoline, and once the gasoline has weathered to 50% or more, the toluene peak is much shorter than the C_3 alkyl benzene or xylene peaks. Gasoline changes considerably as it evaporates, which is what makes it one of the more difficult classes to identify. When a Gaussian distribution of normal alkanes changes because of evaporation, it is still a Gaussian distribution of normal alkanes, just a heavier one. Most of the samples of gasoline that this author has

seen in samples of fire debris have been more than 75% evaporated. (Estimating the degree of evaporation is discussed later.)

Gasoline is, in this author's experience, the most frequently misidentified ignitable liquid residue. That is because many of the compounds present in gasoline as it comes from the pump are also produced when polymers, such as poly(vinyl chloride) (PVC) and polystyrene, degrade as a result of exposure to heat. The key to avoiding misidentifications is making sure that the ratios between groups of compounds and *within* groups of compounds are consistent with the standard. Toluene is a very common pyrolysis product. Note that the polymer need not be an aromatic one to produce aromatic pyrolysis products. In fact, the pyrolysis products of PVC, because of the way it responds to heat, consist almost entirely of aromatics (Stauffer, 2003). It is an unusual fire debris sample that does not contain toluene at some level. Since it is one of the first gasoline compounds to evaporate, one does not expect to see a tall toluene peak in the absence of equally tall xylene peaks in a sample that is positive for gasoline. Figure 3.12 shows the chromatogram of a 10-μL standard of gasoline adsorbed using an activated charcoal strip and eluted with diethyl ether spiked with 100-ppm perchloroethylene. The toluene and xylene peaks are almost equally tall. If a fire debris sample contains toluene from gasoline, it will be accompanied by xylenes

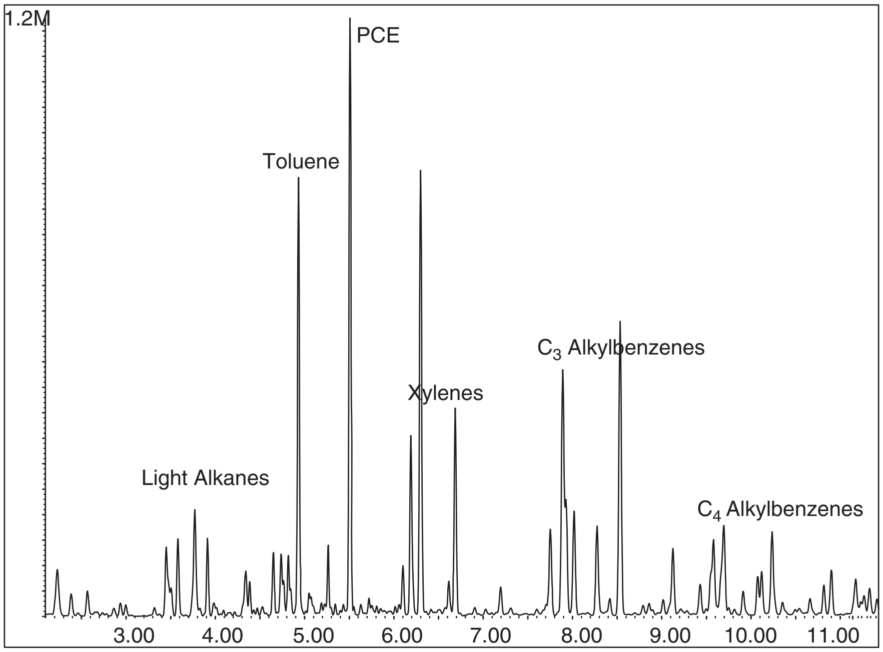

Figure 3.12 Total ion chromatogram of fresh gasoline; 10 μL was spotted on a piece of filter paper and the headspace was concentrated using ASTM E1412. The eluting solvent was diethyl ether spiked with 100 ppm of perchloroethylene, shown here eluting between toluene and the xylenes.

and the higher peak groupings of gasoline. Toluene that is not so accompanied comes from something other than gasoline.

Xylenes are also produced by the decomposition of plastics, but unlike toluene, the xylenes can be examined for correct intergroup ratios. Figure 3.13 shows ion 91, the base ion for xylenes, from gasoline in three different stages of evaporation, kerosene, and a medium petroleum distillate. The relative ratios for the three peaks are almost indistinguishable. Note that the ethylbenzene peak in the 50% evaporated gasoline is slightly lower. This trend continues as the degree of evaporation increases, but generally, one will find that if the xylenes in a sample are from an ignitable liquid, they will exhibit this characteristic ratio. Note that there are three isomers of xylene, *ortho*-, *meta*-, and *para*-, but that the three peaks seen in the chromatograms actually represent four compounds, because *meta*- and *para*-xylene cannot be resolved except in the longest columns (ethyl benzene followed by the three xylenes). If xylenes are found in ratios other than the one shown in Figure 3.13, particularly if ethyl benzene is the tallest peak in the group, one can safely conclude that neither gasoline nor any other petroleum product was the sole source of the xylene. If styrene is present at a higher concentration than the ethylbenzene, one is certainly looking at polymer decomposition products.

The next group to consider is the C_3 alkyl benzenes. This is by far the most important of the patterns in any sample of gasoline. Like the xylenes, it is also found in all petroleum products from which the aromatics have not been removed, and unless evaporation has decreased the concentration of the lighter compounds, the peak ratios will always be the same. This is the group identified in the 1982 guidelines (and in every standard since) as “the *m*-ethyltoluene pseudocumene five-peak group.” This group appears in many samples that contain no gasoline, and it is the failure to exhibit the proper ratio of peaks that is sometimes the analyst’s only clue that he or she is looking at data that can easily be misinterpreted. The group of peaks must be present to identify gasoline, but sometimes it is overemphasized, and sometimes it is “seen” where it really is not present. Figure 3.14 shows one such sample, where there was a misidentification of gasoline. The figure shows the suspect sample at the top, as well as four standards that exhibit the proper ratios of components in the C_3 alkyl benzene group. As with the xylenes, *meta*- and *para*-ethyl toluene cannot be resolved, although the resolution is somewhat better here than in the xylenes—we see a shoulder. In nearly all cases where the C_3 alkyl benzene group is, in fact, present as a component of a petroleum product, the pseudocumene peak will be the tallest. Because most of the residues encountered in gasoline will be highly evaporated, the pseudocumene peak will be the tallest peak on the chart—both the TIC and the EIC. Note that in the questioned sample at the top of Figure 3.14(a), it is the second peak that is the tallest, not the pseudocumene peak. The first peak has the same retention time as *m*- and *p*-ethyltoluene, but the shape is wrong, and the mass spectrum indicates that the peak actually represents benzaldehyde. The mass spectrometer unequivocally identified the second peak in the questioned sample as 1,3,5-trimethylbenzene, but the third peak, which may have represented *o*-ethyltoluene, also contained a significant 118 ion, indicating the

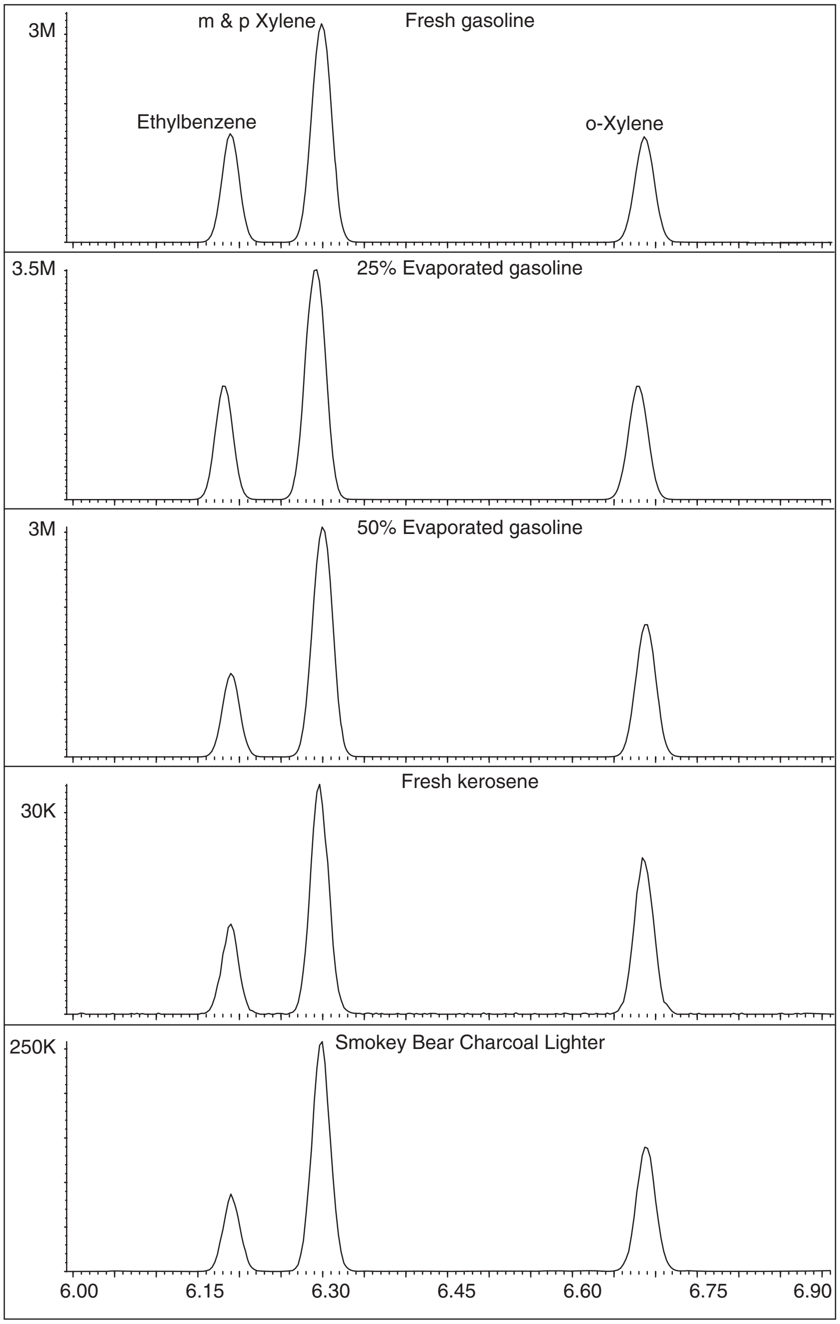

Figure 3.13 Ion chromatogram for ion 91 from gasoline in three different stages of evaporation, kerosene, and a medium petroleum distillate. Note that the relative ratios for the three peaks in the xylene group are almost indistinguishable. Evaporation causes the ethylbenzene peak in the 50%-evaporated gasoline to be slightly shorter.

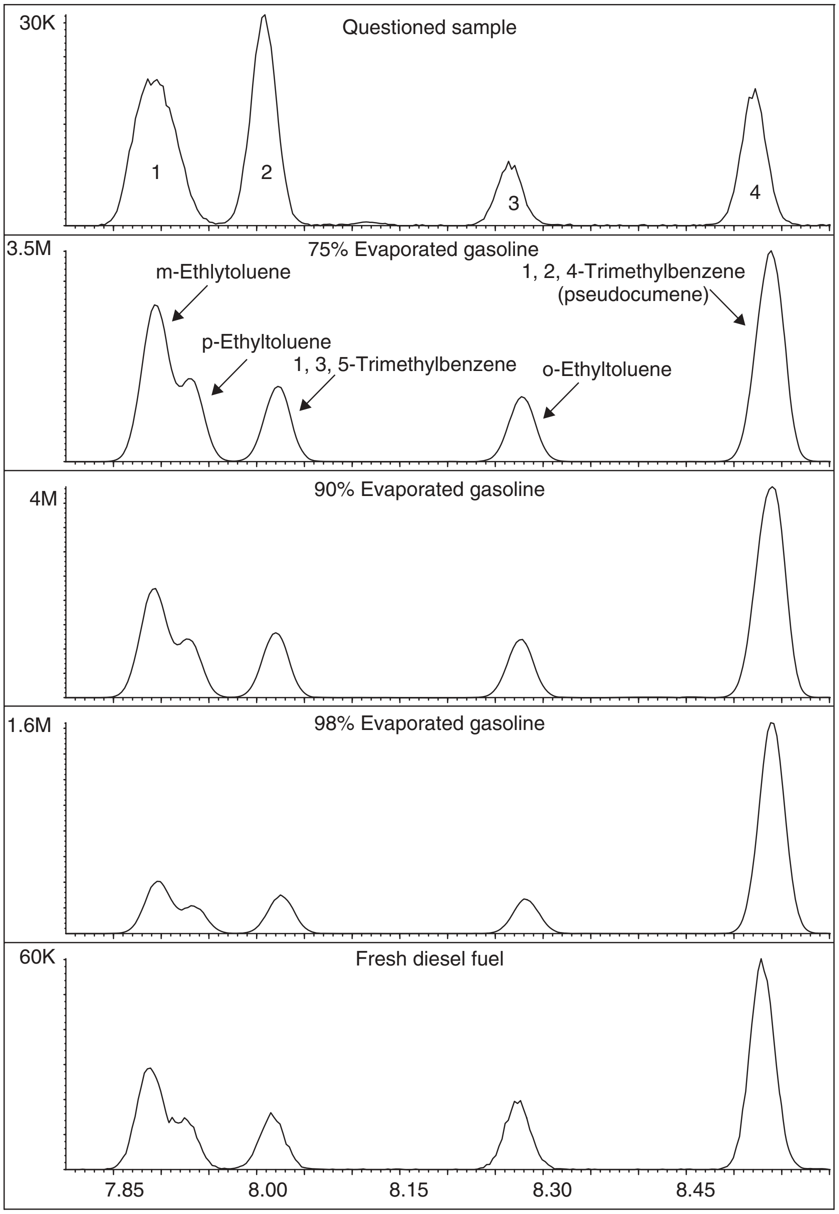

(a)

Figure 3.14 Study of the C_3 alkylbenzene group. (a) The top chromatogram shows an unknown sample. This is followed by gasoline in three stages of evaporation, and fresh diesel fuel. The TIC of the unknown did not look like gasoline, but based on this EIP and a few other ion chromatography comparisons, an analyst called the sample positive for gasoline.

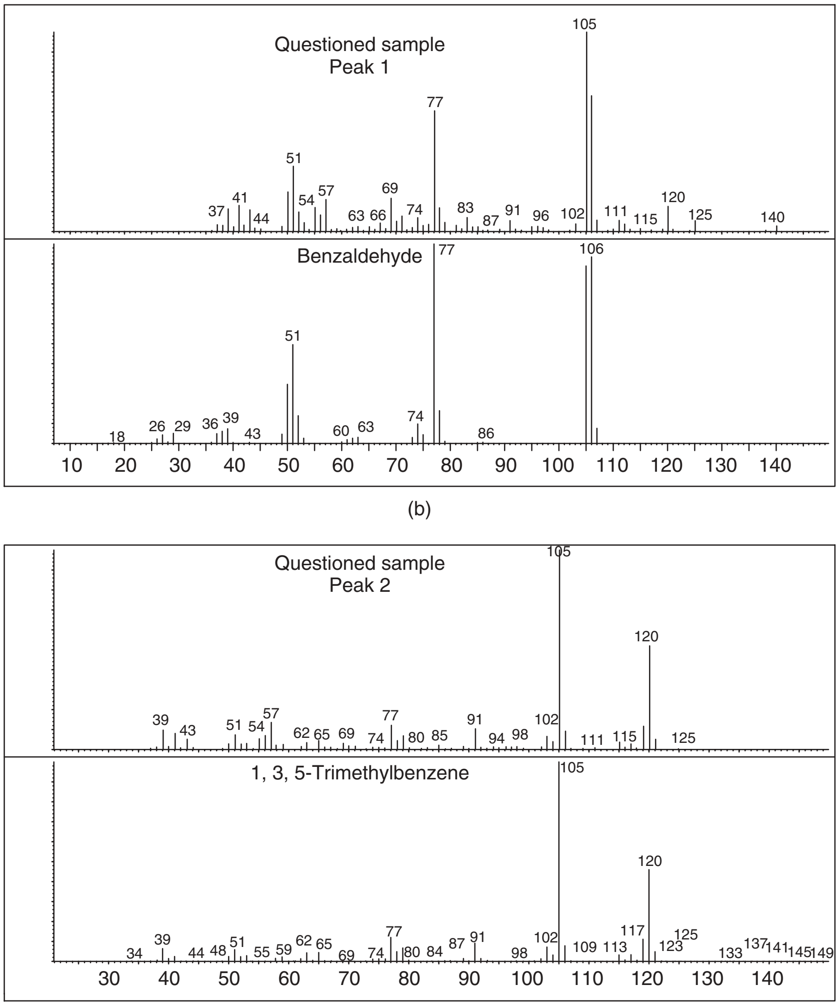

(b)

(c)

Figure 3.14 (*Continued*) The mass spectra shown in parts (b) to (e) show that two of the peaks, 1 and 3, were not gasoline components (b). Mass spectrum and library match of the first peak in the top chart shown in part (a). This peak actually represents benzaldehyde, not *m*-ethyltoluene. (c) Mass spectrum of the second fully resolved peak in the top chart shown in part (a). As with gasoline, this peak is identified as 1,3,5-trimethylbenzene (d). Mass spectrum of the third major peak in the chart shown at the top of part (a). This mass spectrum changed across the peak but clearly contained α-methylstyrene. 1-Decene, a common decomposition product, was also suggested by the spectrum. There may be some *o*-ethyltoluene coeluting as well, but that cannot be demonstrated (e). Comparison of the mass spectrum fourth peak in the chart shown in part (a) with 1,2,4-trimethyl benzene (pseudocumene). This is a reasonably straightforward match.

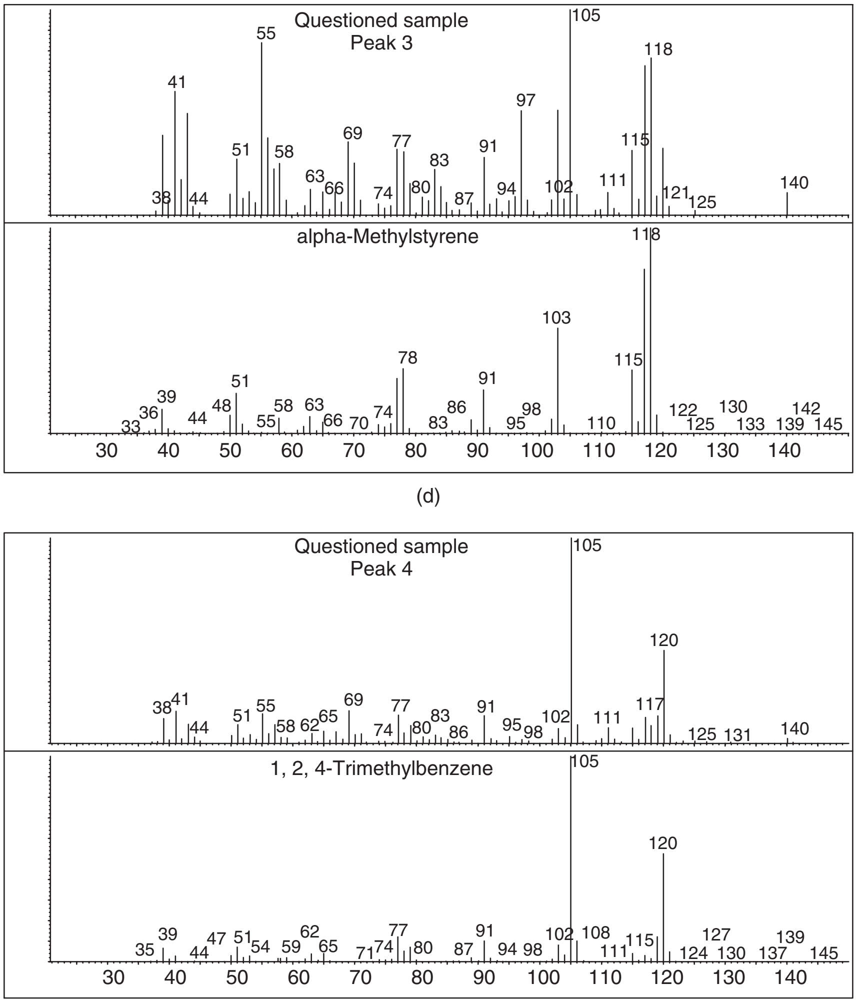

Figure 3.14 (*Continued*)

presence of ∝-methylstyrene, a common pyrolysis product. Phenol, which exhibits a base peak at 94, is also a confounding compound that elutes in the range of the five-peak group. *Simply matching the components when the ratios are not right can lead to misidentifications*. The odds against three background components coeluting in exactly the right concentrations to skew the peaks in this group are pretty high. Three other samples from the same fire scene were similarly misidentified, and exhibited similar peak ratios and mass spectral characteristics. As with almost all misidentifications, there was not a good pattern match with the TIC. Extracted ion chromatography or extracted ion profiling can be very useful; however, the analyst should remember that it is a spectral as well as a chromatographic technique. The individual mass spectra should be examined, especially when the peak ratios are “off.”

A word about the use of the mass spectrometer beyond generating mass chromatograms is in order. Most analysts will use the mass selective detector or mass spectrometer as a tool for generating extracted ion chromatograms and extracted ion profiles. Obviously, if one is looking at very simple mixtures or single components, the mass spectrum is necessary in order to make an identification. A sometimes-overlooked function of the mass spectrometer is the evaluation of extracted ion chromatograms and profiles. As we saw in Figure 3.14(a), an extracted ion chromatogram with the intragroup ratios just slightly “off” has the potential to be misleading. Figure 3.14(b) to (e) show the mass spectra of each of those five peaks, as well as the library’s best match. The first peak, which is coincident with *m*-ethyl toluene, has a large peak at m/z 77, but no such ion is present in the spectrum of *m*-ethyl toluene. There may be some *m*-ethyl toluene hidden under this peak, but there is definitely some benzaldehyde as well.

One way to determine whether there are coeluting compounds under a chromatographic peak is to examine the mass spectra at different points across the peak. If the peak represents a pure compound, the mass spectrum will change little, if at all. In Figure 3.14(c), the second peak in the five-peak group is a nearly-perfect match for 1,3,5-trimethylbenzene, which is the third peak in the five-peak group required to identify gasoline. The third peak, whose mass spectrum is shown in Figure 3.14(d), clearly contains more than one substance, but the strong peak at 118, as well as the retention time, indicates that α-methylstyrene is coeluting. Like the second peak, the fourth peak is a pure compound and a nearly perfect match for pseudocumene. (The mass spectra of all of the aromatics in this group are quite similar. In fact, using a library search usually results in a list of match candidates that includes all four members of the group.) Because the peak ratios are off, and especially because two of the four peaks in the ion 105 chromatogram represent compounds not found in gasoline, it must be concluded that the first identification was in error.

Background subtraction is a tool found in most data analysis software and can frequently resolve questions as to whether a particular compound is really there. A detailed evaluation of the quality of the spectra underlying a mass chromatogram, whether it is a single-ion extraction or a multiple-ion profile, should be carried out periodically just to keep the analyst in practice, but it should be carried out routinely on any sample where either the peak ratios or retention times are “just a little off.” Note in Figure 3.14(a) that when a sample is sufficiently evaporated, the peaks at the left side of the chart begin to diminish. This result is expected. The initial guidelines for identifying gasoline stated that the *m*-ethyl toluene/pseudocumene five-peak group was still present in gasolines that had lost 90% of their fresh weight. In fact, this group does not disappear until the gasoline is more than 98% evaporated. If this group cannot be positively identified, an analyst is on very thin ice indeed when identifying a sample as containing gasoline. In such a case, there must be a peak-for-peak match of all of the higher peak groupings. As with many of the other components of gasoline, this five-peak group is present in almost all petroleum products from which the aromatics have not been removed. The bottom chart in Figure 3.14(a) is from unevaporated diesel fuel.

The C_4 alkyl benzenes are best viewed by extracting ion 119. Because there are more ways to build a C_4 alkyl benzene than a C_3, this is a more complex pattern, but it is present in almost all petroleum products. Figure 3.15 shows the patterns found in highly evaporated gasoline, as well as in kerosene and diesel fuel, which are produced by the presence of the C_4 alkyl benzenes. Note that in kerosene and diesel fuel, this pattern is more complex than in gasoline.

The next group of compounds that must be present in order to make a solid identification of gasoline is the indans. These can be extracted using ions 117 and 131. The doublet at 11.3 and 11.5 min in the ion 117 chart is, in this analyst's experience, always present. It has traditionally not been an absolute requirement of the ASTM standards, but it probably should be. The standard states, "Indan (dihydroindene) and methylindans are usually present." The first tall peak in the 117 chart is indan, appearing at 9.4 min. Figure 3.16 shows the peak groupings characteristic of gasoline for the indans, methylindans, and dimethyl indans. Like the C_3 alkyl benzenes, the indans and alkyl-substituted indans can be found in roughly the same proportions in gasoline that is 98% evaporated. Other petroleum products, particularly the distillates, contain indans and methylindans, but their mass chromatographic patterns are more complex than the ones for gasoline.

The next group of compounds that should be present in a sample identified as containing gasoline is the methyl- and dimethylnaphthalenes. Naphthalene is also present, but naphthalene is so common that its presence in a sample is meaningless. 2-Methylnaphthalene elutes before 1-methylnaphthalene and is almost always more abundant than the 1-methylnaphthalene. In any case where this ratio is reversed, the extract should be considered suspect. The boiling point of the methylnaphthalenes is high enough that the ratio, unlike xylenes or C_3 alkyl benzenes, is unlikely to be affected by evaporation. All petroleum products are likely to contain both the naphthalenes and the dimethylnaphthalenes, and the ratios should be highly comparable to each other.

There are some aromatic products in the marketplace that meet almost all of the ASTM criteria for the identification of gasoline. Such products are used as solvents for stains, insecticides, adhesives, and industrial and commercial products. The way to differentiate these aromatic products from gasoline is to look for the presence of alkanes. All gasolines contain high percentages of alkanes when fresh, but even highly evaporated gasolines will contain some branched alkanes. Straight-chain hydrocarbons are an unusual finding in gasoline, because they are undesirable components that cause knocking when burned in gasoline engines. [Knocking is

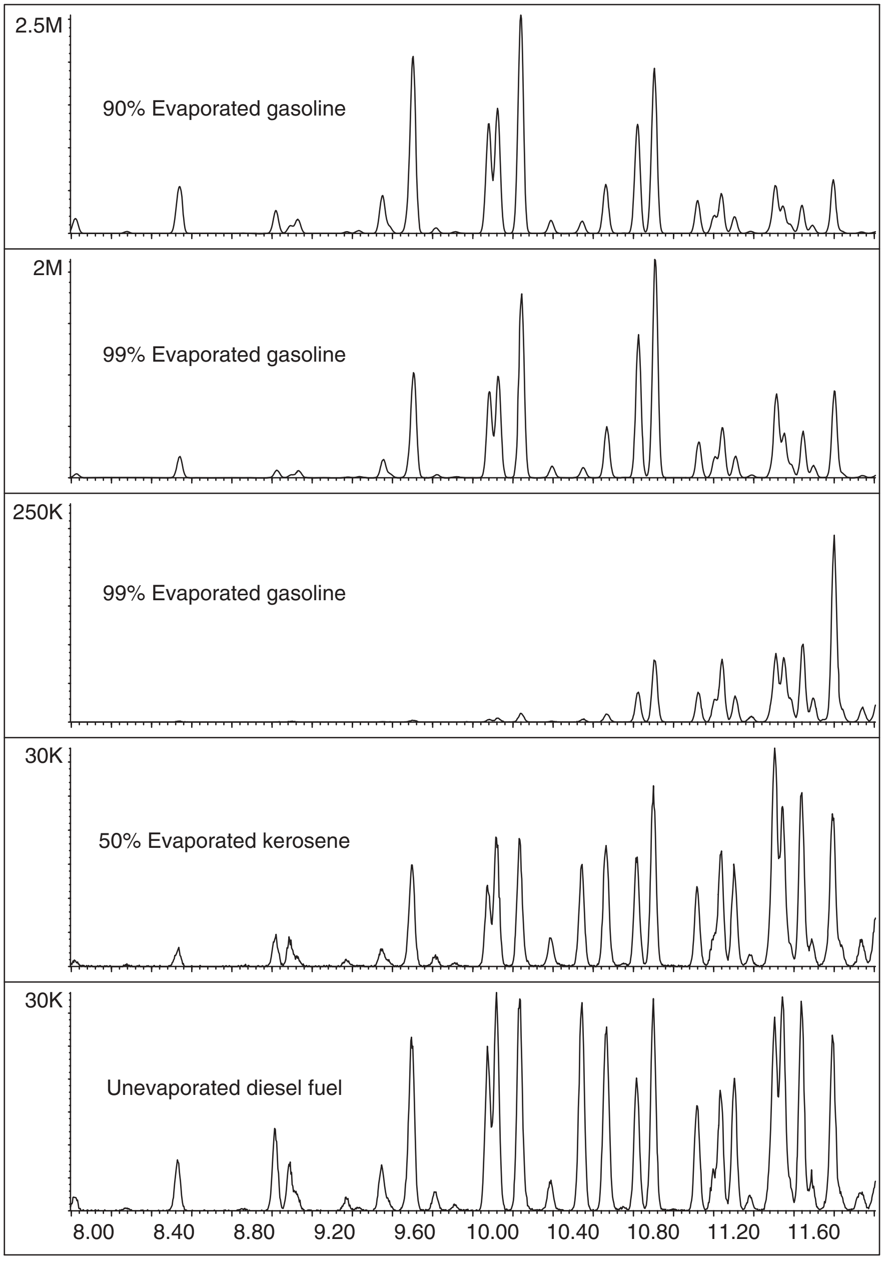

Figure 3.15 C_4 alkylbenzenes represented by ion 119 in gasoline in three different stages of evaporation, 50%-evaporated kerosene, and unevaporated diesel fuel. All samples were prepared by spotting 10 μL on a piece of filter paper, and processed using ACS adsorption–elution.

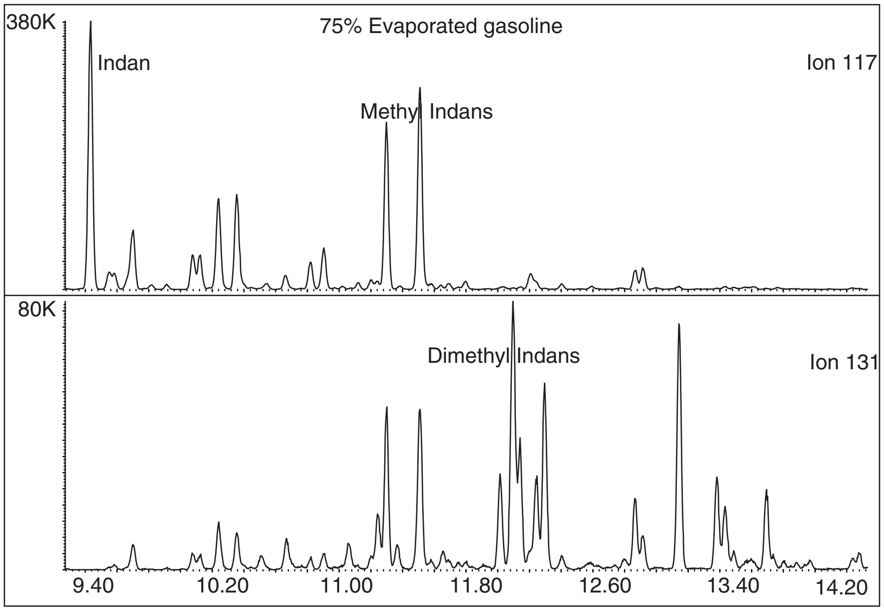

Figure 3.16 Ion chromatograms for indan, the methylindans, and dimethylindans found in 75%-evaporated gasoline.

the premature detonation of a fuel and occurs more readily with straight-chain hydrocarbons than with branched-chain hydrocarbons or aromatics. Resistance to knock is known as the *octane rating*. Octane rating is based on a scale of 0 to 100, where 0 is the resistance to knock of 100% *n*-heptane and 100 is the resistance to knock of 100% iso-octane (2,2,4-trimethylpentane). *N*-Octane actually has an octane rating below zero.]

Identification of Distillates Most distillates likely to be encountered in fire debris are straight-run distillates from crude oil. They have not been subjected to cracking or reforming, so they lack the high aromatic content found in gasoline, but aromatics are still present. The distillates are characterized by an abundance of normal alkanes, along with branched alkanes and cycloalkanes.

Distillates are usually easy to recognize, except for light petroleum distillates, which might be missed unless the analyst checks for them in every sample before calling it negative for ILR. Because of their high volatility, they tend to be present at low concentrations in fire debris samples and are typically found on the left side of the chart, where they may be mistaken for decomposition products. This is a particular hazard when the higher-boiling components are decomposition products. Light petroleum distillates (LPDs) do not generally exhibit the Gaussian distribution seen in medium and heavy petroleum distillates (MPDs and HPDs). When the peaks that elute prior to 8 min are examined, and the analyst sees a mixture of cycloalkanes, branched alkanes, and perhaps some normal alkanes in the C_7

through C_9 range, an LPD should be suspected. The analyst should have in the library as many LPDs as possible, because unlike the heavier distillates, the patterns tend to vary from one to the other,. Figure 3.17(a) shows three different brands of cigarette lighter fluid, each exhibiting a different pattern of peaks. The recommended extracted ion profiles for LPDs are 57, 55, 83, and 91, which can be seen in Figure 3.17(b). This particular LPD is unusual in that it contains xylenes but no toluene is present. Because of their high volatility, LPDs are not persistent in the environment and are not usually expected to be found as background material. There are some cleaning agents and automotive products that contain LPDs, but if the sample was collected from somewhere other than the garage, the workshop, or under the kitchen sink, a finding of LPD generally indicates the presence of a foreign ignitable liquid.

This is not the case with medium or heavy petroleum distillates, which are far more common in our environment (Lentini et al., 2000). A typical medium petroleum distillate, Sparky charcoal lighter fluid, is shown in Figure 3.18(a). The ions necessary to make a valid identification of a medium petroleum distillate are 57 and 83, shown in Figure 3.18(b). The analyst should also check for ion 105 or 91, to be certain that the aromatics have not been removed, in which case the ILR can be identified as a dearomatized distillate. One does not notice much difference between the TIC and the ion 57 EIC, because distillates are dominated by the normal and branched alkanes. The alkylcyclohexanes, which elute about midway between the normal alkanes, present an overall appearance similar to that of the alkane chromatogram. If the normal alkanes are not present, or if they are present

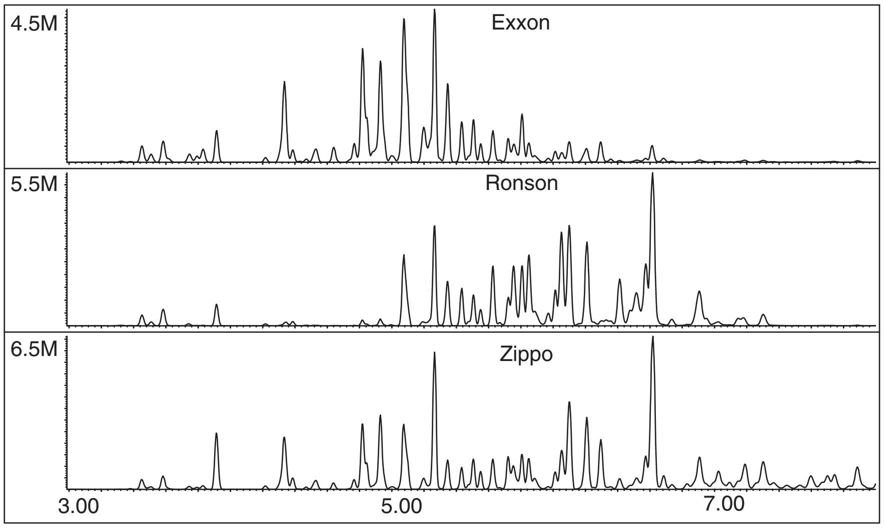

(a)

Figure 3.17 (a) Total ion chromatograms of three different brands of cigarette lighter fluid. (b) Ion chromatograms of ions 57, 55, 83, and 91 from a standard of Ronson lighter fluid.

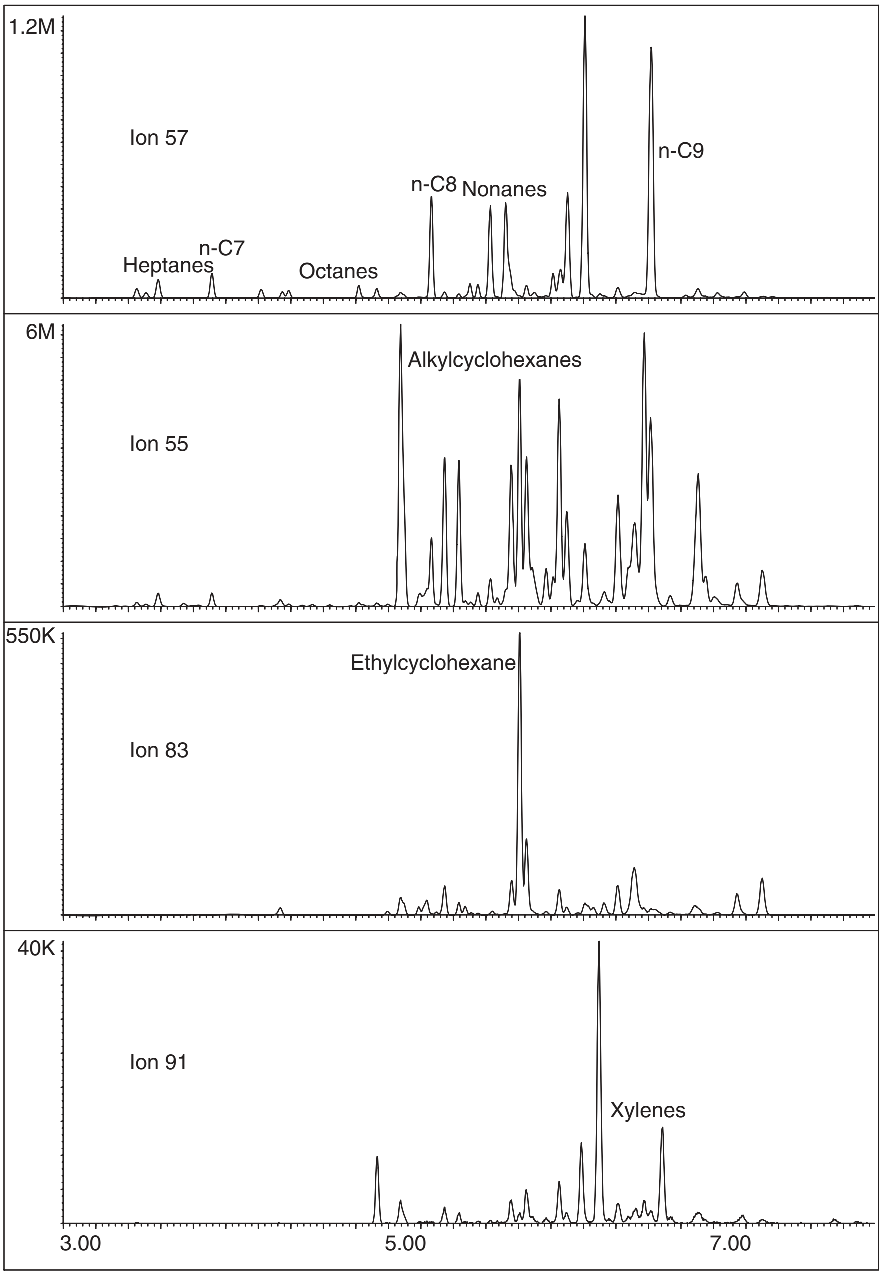

(b)

Figure 3.17 (*Continued*)

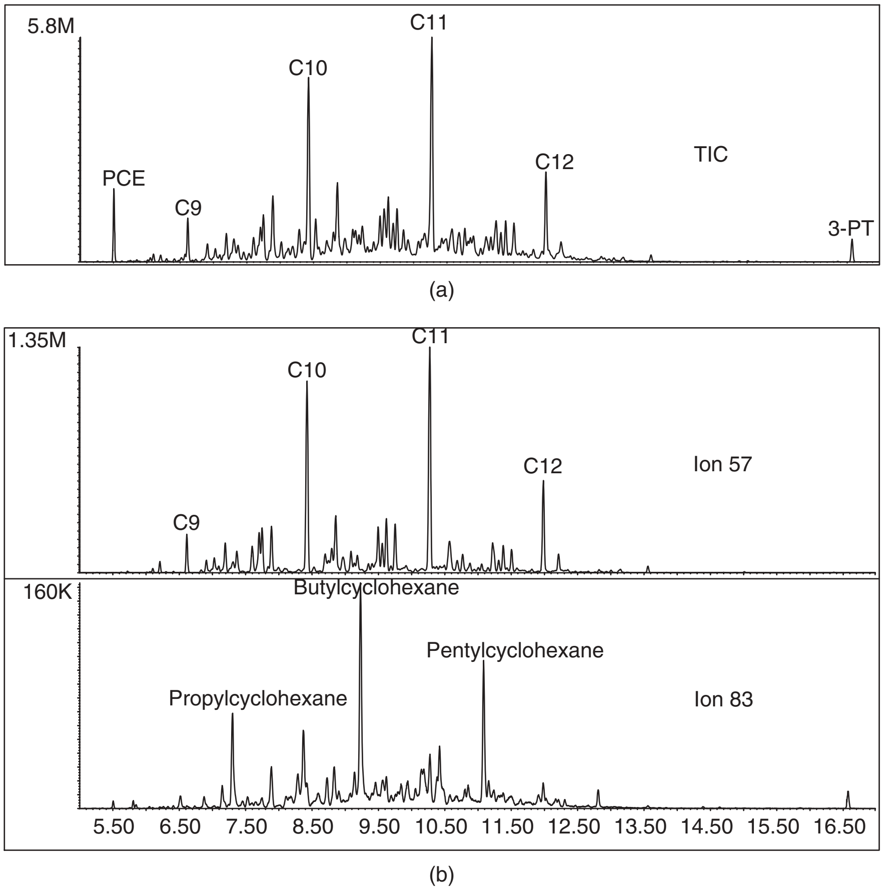

Figure 3.18 Total ion chromatogram of (a) a typical medium petroleum distillate, Sparky charcoal lighter fluid, and (b) extracted ion profiles for ions 57 and 83.

at approximately the same concentration as the branched alkanes, an isoparaffinic product should be suspected. When the cycloalkanes are present at an abundance greater than about 20% of the abundance of the alkanes, a naphthenic–paraffinic source is indicated. In most distillates, the cycloalkanes will be present at about 5 to 10% of the concentration of the normal and branched alkanes. Medium petroleum distillates cover a wide range of products and may be used as fuels, such as lamp oil or charcoal starter, or as solvents, such as mineral spirits or insecticide carriers. This wide range of products and formulations requires that the library of reference materials include numerous MPDs.

Heavy petroleum distillates (HPD) include kerosene and diesel fuel. Except for HPDs that are formulated for specific applications, such as jet fuel, the carbon number range of heavy petroleum distillates can vary. Thus, unless one has a sample of the unevaporated liquid, it is difficult to determine the degree of evaporation

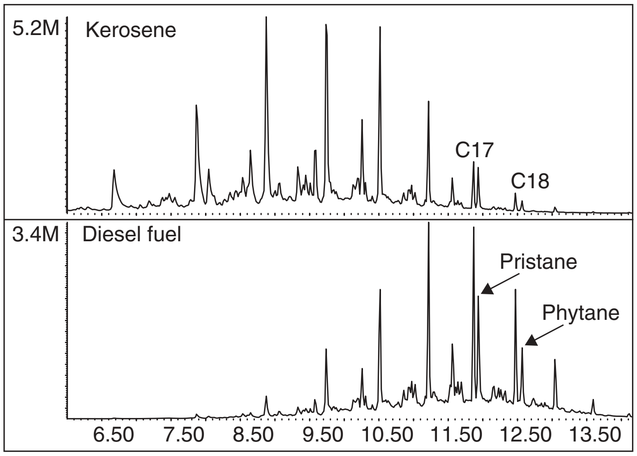

Figure 3.19 Total ion chromatograms showing a comparison of kerosene and diesel fuel.

of the residue isolated from a sample. In northern parts of the country, diesel fuel sold in the winter may actually be kerosene.

Figure 3.19 shows a comparison of kerosene and diesel fuel. These distillates come with internal carbon number markers, in the form of pristane and phytane. Pristane is 2,6,10,14-tetramethyl pentadecane ($C_{19}H_{40}$) and it elutes immediately after normal heptadecane. Phytane is 2,6,10,14-tetramethyl hexadecane ($C_{20}H_{42}$) and it elutes immediately after normal octadecane. (Although pristane and phytane elute immediately after two normal hydrocarbons, they come from different sources. The n-alkanes started out as fatty acids, but pristane and phytane started out as parts of the chlorophyll molecule.) Thus, one can identify the two doublets on the high side of the bell-shaped curve and can count carbons up and down from there. In Figure 3.19 these doublets occur at 12 and 12.5 min. Kerosene, which is a known standard that has been evaporated to 50% of its original volume, covers the range C_{11} to C_{19}, while the diesel fuel range is C_{12} to C_{21}. If the kerosene is evaporated further, it will look more like the diesel fuel. Had the diesel fuel been evaporated less, it would look more like the kerosene.

As with the medium petroleum distillates, interpretation of a finding of heavy petroleum distillates should be approached with caution. There are numerous household products that contain HPDs, including many of the same kinds of products in which MPDs are found. The safer charcoal lighters are made from kerosene rather than mineral spirits. They are safer because of their higher flash point. Figure 3.20 shows a total ion chromatogram of lemon oil furniture polish. This polish has a carbon number range from C_{12} to C_{22}, with C_{17} being the tallest peak on the chart. The limonene peak at the left side of the chart could easily be attributed to a pine substrate. Pinenes and limonene are very common in samples containing structural (coniferous) wood. If found in a fire debris sample, this particular furniture polish could easily be misreported as diesel fuel.

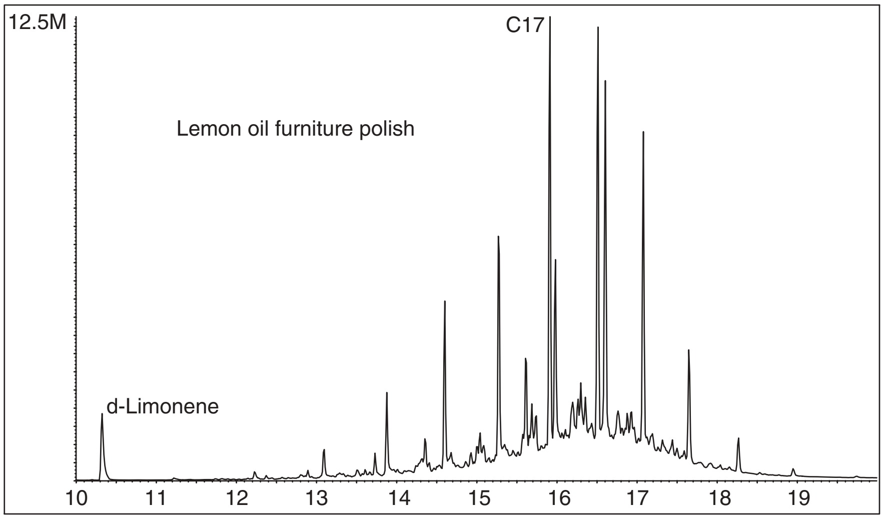

Figure 3.20 Total ion chromatogram of lemon oil furniture polish. This sample could easily be reported out as diesel fuel.

It is not necessary for a liquid to be present in order for a distillate to be detected. Figure 3.21(a) shows the total ion chromatogram of a piece of pine wood that was stained with Minwax finish 10 months before it was subjected to headspace concentration. The naturally occurring terpenes, α- and β-pinene and *d*-limonene, are the dominant peaks on the chart, but the mineral spirits solvent is still clearly visible. When ions 57 and 83 are extracted, the terpenes disappear, to yield the charts shown in Figure 3.21(b). These results show the critical necessity of asking for comparison samples, particularly when samples of flooring are submitted for analysis. The flooring does not need to be recently painted to exhibit the distillate solvent used to apply the floor coating. This author has reported finding distillates 24 months after application (Lentini, 2001) and has detected distillates in samples of finished flooring and furniture up to 10 years old. There is no reason to believe that these solvents do not persist indefinitely, trapped in either the wood matrix or in the polymer coating matrix.

In addition to true distillates, there are some distillate-like residues produced as the result of decomposition of other products. Asphalt is what is left at the bottom of the distillation pot after all of the volatiles have been distilled from crude oil. It contains hydrocarbons ranging from C_{30} to C_{60}. When these long-chain hydrocarbons undergo pyrolysis, they do so in much the same way as that of the long-chain hydrocarbons in polyethylene, via random scission. This results in the production of normal alkanes in the range C_9 to C_{18} (Lentini, 1998). When these pyrolysis products are present in fire debris and subjected to headspace concentration, they produce a chromatogram that can be and has been mistaken for the chromatogram of a heavy petroleum distillate. Such a chromatogram is shown in Figure 3.22(a). In 1982 it was reported that roof shingles could produce “accelerant-like residues,”

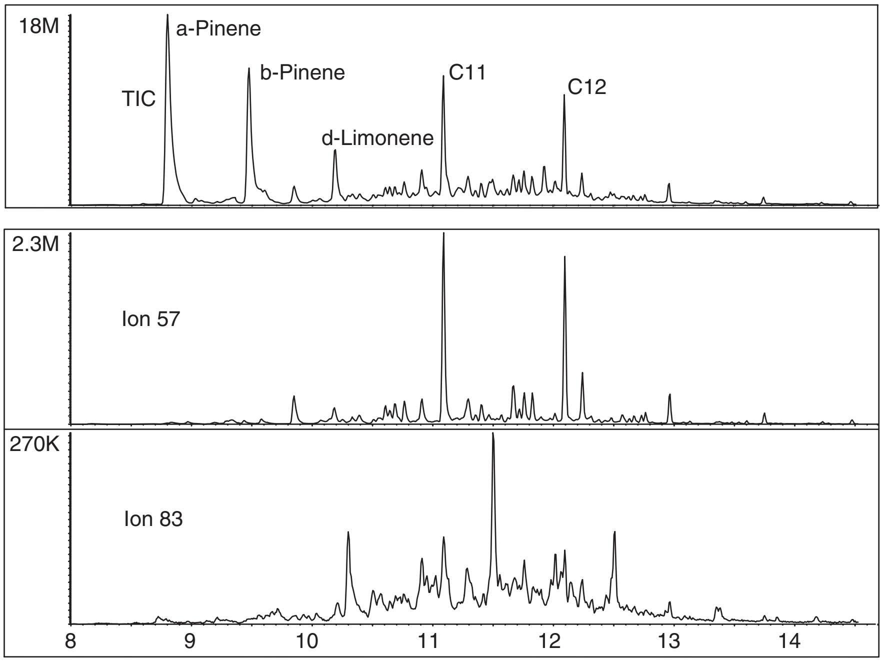

Figure 3.21 Total ion chromatogram of (a) a piece of pine wood stained with Minwax Finish 10 months prior to its analysis, and (b) extracted ion profiles for ions 57 and 83.

and at that time, there existed no reliable methods for distinguishing HPDs from asphalt shingle residues (Lentini and Waters, 1982). The increased use of capillary columns allowed for the occasional visualization of a "double-peak kerosene," sometimes called "pseudokerosene," but even with capillary columns, it was not uncommon for asphalt residue to be misidentified as a liquid petroleum distillate. This author learned how to make the differentiation in 1995 in connection with the investigation of an insurance claim that had been erroneously denied because of a finding of HPDs where none should have been present. Figure 3.22(b) shows how the distinction is made. One compares ion 57 with ion 55. If a second peak appears in front of the n-alkane peak, or if a small peak grows larger, we can conclude (particularly after we collect the mass spectrum) that the second peak is the 1-olefin. A sample such as this must be classified as asphalt smoke condensate or asphalt decomposition residue.

The amount by which the olefin peak "grows" will vary depending on the sample. When one looks at kerosene and diesel fuel, the only difference in appearance between the ion 55 chart and the ion 57 chart is that the abundance of ion 55 is lower. The relative abundances of the individual peaks with respect to each other do not change.

Another way to distinguish asphalt decomposition products from HPDs is to look for the cycloalkanes. They are not present in asphalt decomposition products. A Gaussian pattern will be observed when ion 83 is extracted, but this is due

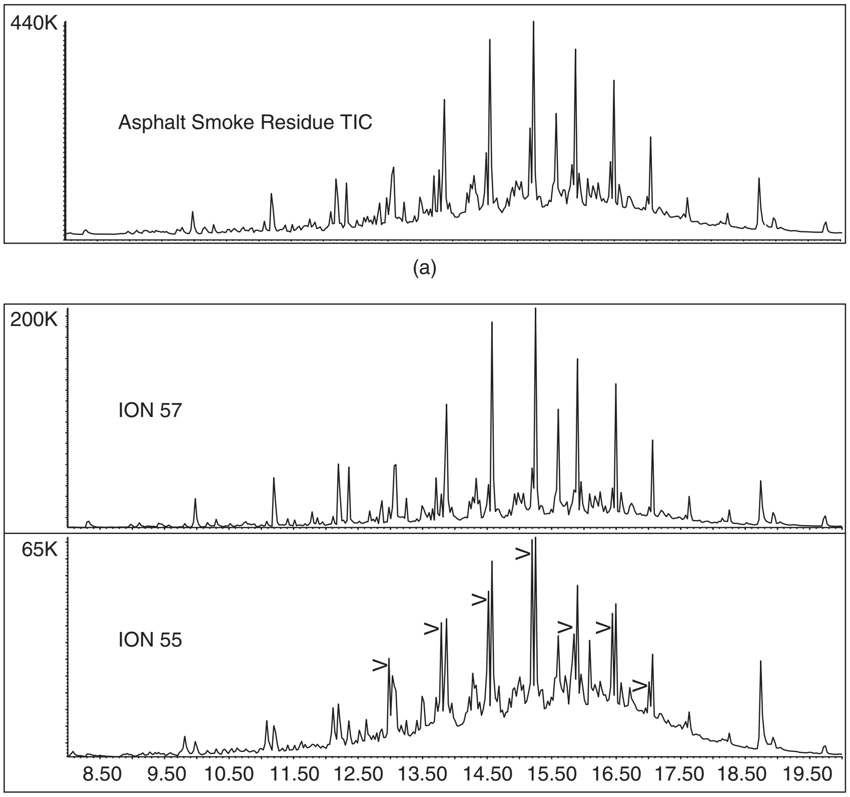

(a)

(b)

Figure 3.22 Total ion chromatogram of (a) asphalt smoke residue of the type that can be mistaken for a petroleum distillate, and (b) extracted ion profiles for ions 57 and 55. The growth of the peak in front of the normal alkane when comparing 55 to 57 demonstrates the presence of alkenes, which allows the identification of the residue as a decomposition product rather than a foreign petroleum distillate.

entirely to the presence of the olefins. The ion 83 and ion 55 ion chromatograms show the same peaks. Asphalt smoke condensates also contain little if any pristane and phytane.

Polyethylene is another substance that produces a distillate-like appearance in the chart, but if the capillary column has any resolution at all, it will be obvious that one is looking at polyethylene residue rather than at a distillate. Figure 3.23(a) shows the total ion chromatogram of polyethylene smoke condensate. All of the peaks are doublets and most are actually triplets, owing to the presence of the 1,$(n - 1)$-diene in the mixture. When one looks at ion 57 and compares it with ion 55, as shown in Figure 3.23(b), the growth in the olefin peak is obvious relative to the alkane peak. The diene peak also grows, because the dienes contain more ion 55 than ion 57. Lubricating oils are subject to the same decomposition processes as asphalt and polyethylene, and the smoke condensates of lubricating oils appear similar.

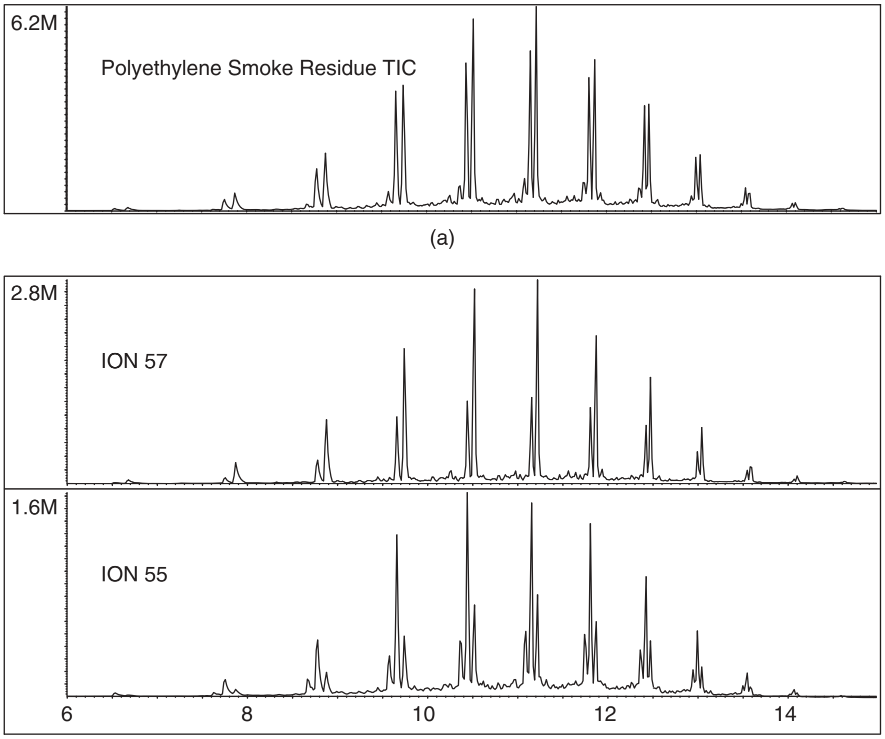

(a)

(b)

Figure 3.23 Total ion chromatogram of (a) polyethylene smoke condensate. The first peak in each doublet is the alkene; the second is the alkane. Pristane, phytane cycloalkanes, and aromatics are all absent. (b) Ion 57 EIC, and (c) ion 55 EIC. In the ion 55 chromatogram, a third peak, representing the n, $(n - 1)$-diene, appears.

A careful examination of the chromatographic and mass spectral data will prevent the analyst from misidentifying decomposition products as distillates, but automation of pattern recognition is coming and presents some dangers if not handled properly. As with all computer “answers,” one should always do a “reality check.” The same pattern recognition software that allows a mass spectrum from an unknown compound to be matched against the spectra of 100,000 compounds can be applied to chromatograms. The chromatogram is converted using Microsoft Excel into a bar graph that has the same general appearance as a mass spectrum. This graph can then be compared against a library of known ignitable liquids that have been transformed similarly. As with the mass spectral libraries, “extra” peaks do not necessarily keep the database from recognizing a “match.” The polyethylene smoke condensate chart shown in Figure 3.23(a), matched up against a database that contains only ignitable liquids, will yield a match of exceptionally high quality for diesel fuel. This is why it is necessary to populate ignitable liquid libraries used for this purpose with known background chromatograms. The chart would make an even better match for polyethylene smoke condensate, but only if it is available in the database.

Identifying Other Classes of Products The remaining classes of ignitable liquid residues can frequently be identified by what is and is not present. A dearomatized distillate, for example, will have a signal from the aromatics that is less than 1% of the signal from the aliphatics. Otherwise, it will look the same with respect to both the alkanes and the cycloalkanes. Because of environmental regulations governing the aromatic content of distillates, there has been, since 1973, an ASTM standard for this determination. ASTM D3257-07 is entitled *Standard Test Methods for Aromatics in Mineral Spirits by Gas Chromatography* (ASTM D3257-07, 2007) and uses a specified test blend for calibration. If the calculation of aromatic content is an issue, a forensic scientist might avoid a *Daubert* challenge by using this established method rather than by devising a new one.

The normal alkane products are very easy to recognize, as long as one makes sure that one is looking at a homologous series of normal alkanes as opposed to a homologous series of aldehydes or some other group of homologues that are pyrolysis or decomposition products. Groups of compounds that differ from each other only in that they have one additional CH_2 group look pretty much the same as a series of normal alkanes, and require some caution. Figure 3.24 shows a series of normal alkane products marketed by Exxon, as well as a sample of Lamplight Farms ultrapure lamp oil, which also consists of normal alkanes. One common source of normal alkanes is carbonless forms (NCR paper). The bottom chromatogram in Figure 3.24 shows the chromatogram of concentrated headspace vapors from a 2-in. × 2-in. square of a carbonless form. These forms contain microspheres filled with normal alkane solvent that, when broken, causes color to develop in the ink. There are only a few microliters per square foot, but this concentration is easily detectable. Figure 3.25 is a scanning electron micrograph, showing the bottom side of a carbonless form. It should be noted that the alkanes found in carbonless forms are usually accompanied by a pair of substituted biphenyls. Normal alkane products are also found in some brands of linoleum floor covering. Any time that a sample is collected from a floor likely to have linoleum in its structure (kitchen, bathroom, laundry room), a finding of normal alkanes is probably not meaningful. A comparison sample is an absolute necessity in such cases.

Isoparaffinic hydrocarbons are made by removing the normal hydrocarbons with a molecular sieve. These liquids are becoming more common as petrochemical manufacturers move to more environmentally friendly and odor-free replacements for the straight-run distillates, such as mineral spirits. Similar to the isoparaffinic hydrocarbons are the naphthenic–paraffinic products, which are characterized by an abundance of cycloalkanes. Whereas one might expect to find cycloalkanes present at less than 5% in the isoparaffinic hydrocarbons, they may be present at up to 30% in naphthenic–paraffinic products. The distinction requires looking at the ion profiles for the cycloalkanes versus the alkanes, as well as looking at the abundance numbers on the left side of the chart. Figure 3.26 shows a comparison

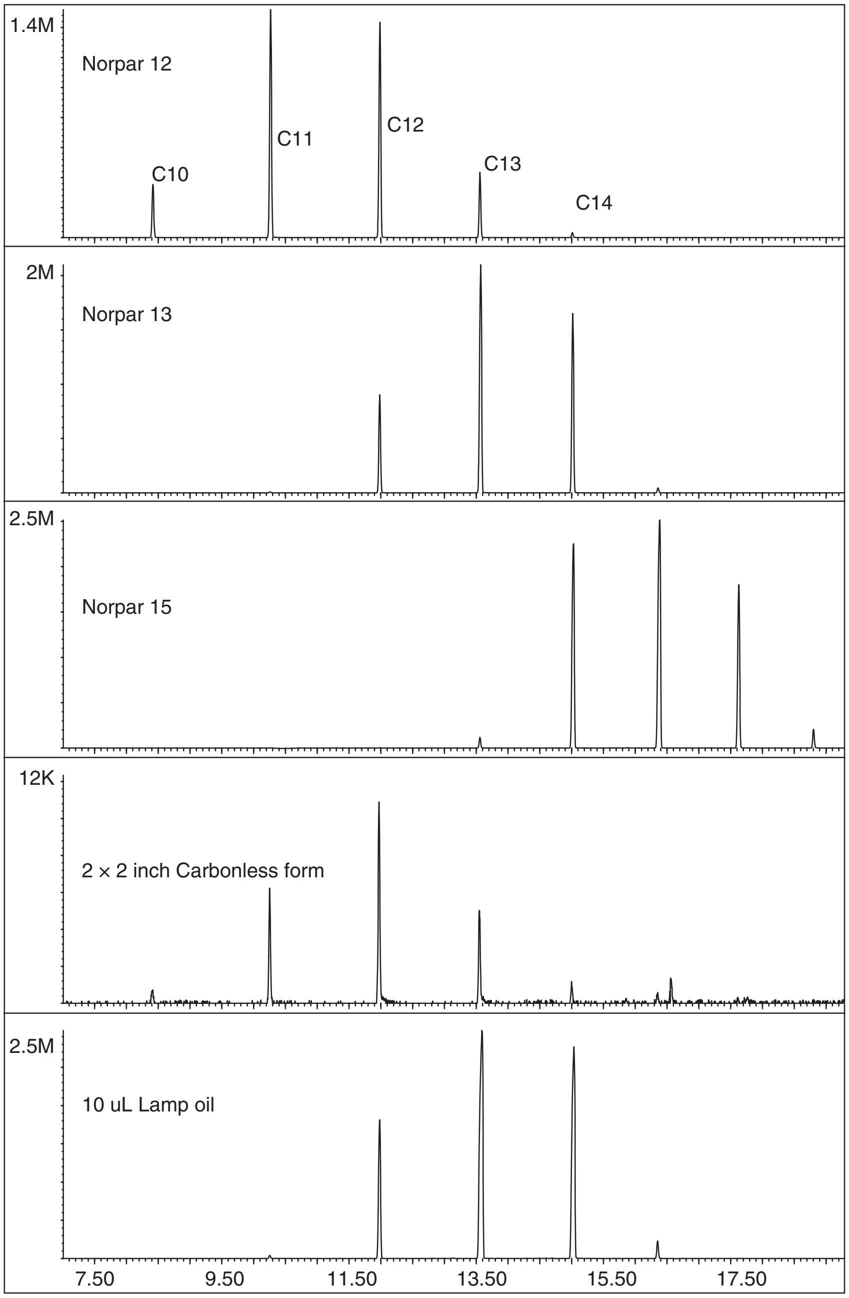

Figure 3.24 Ion 57 chromatogram of three Exxon Isopar products, compared with an ACS extract from a carbonless form, and the ACS extract of a 10-μL sample of Lamplight Farms ultrapure lamp and candle oil.

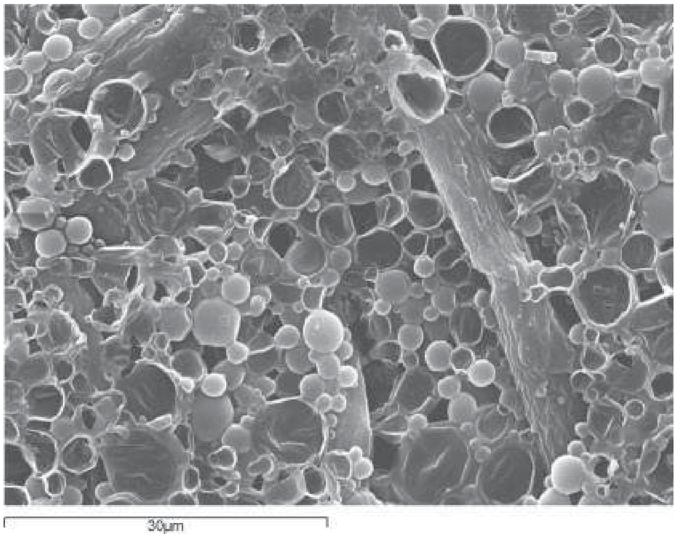

Figure 3.25 Scanning electron micrograph of the underside of a carbonless form. The microspheres range in size from 2 to 20 μm and are filled with normal alkanes.

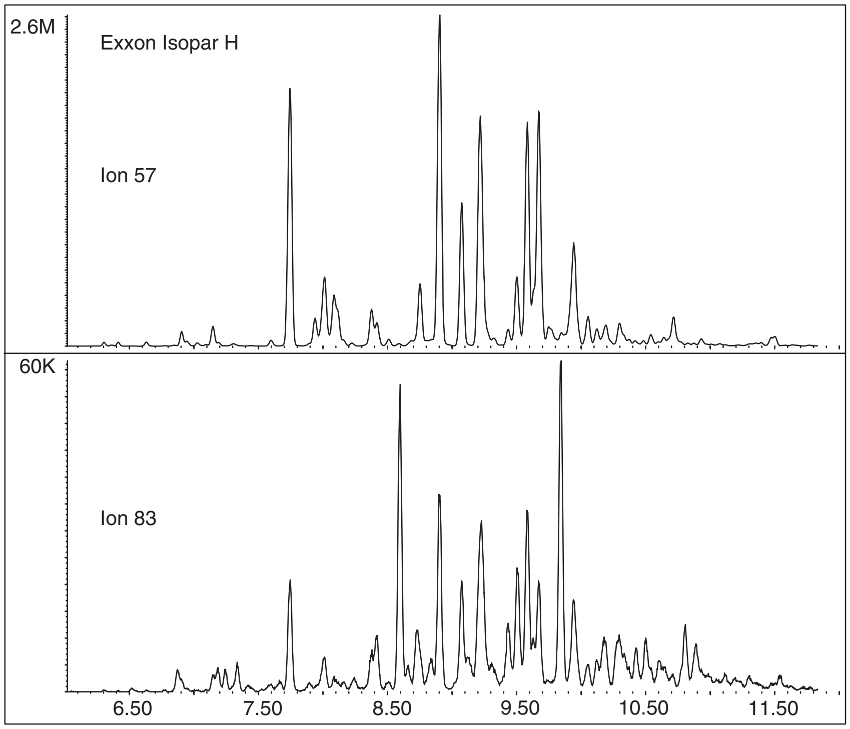

Figure 3.26 Comparison of the alkane (ion 57) and cycloalkane (ion 83) chromatograms from Exxon Isopar H. The cycloalkane peaks are only about 2% of the height of the branched alkane peaks.

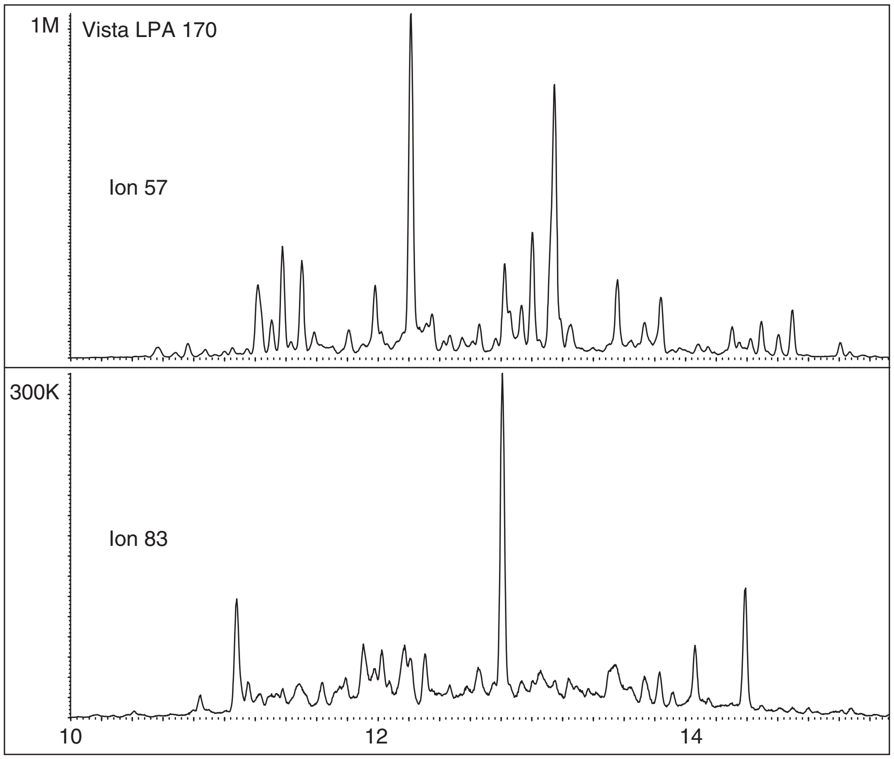

Figure 3.27 Comparison of the alkane (ion 57) and cycloalkane (ion 83) chromatograms from Vista LPA 170, a naphthenic–paraffinic product. The height of the cycloakane peaks is about 30% of the height of the branched alkane peaks.

of the ion 57 and ion 83 chromatograms from Isopar H, a common solvent used in cosmetics and other household products. Compare this to Figure 3.27, which compares those same ion chromatograms from a naphthenic–paraffinic solvent, Vista LPA 170. Instead of having an ion 83 chromatogram that is 2% the height of the ion 57 chromatogram, the ion 83 is more than 30% of the height of the ion 57 chromatogram in the naphthenic–paraffinic solvent. Also, note that the tall peaks in the ion 57 chromatogram from the naphthenic–paraffinic product are not normal alkanes, but are branched alkanes.

Another way to make the distinction between the isoparaffinic products and the naphthenic–paraffinic products is to take an average mass spectrum. Representative isoparaffinic products are shown in Figure 3.28. A average mass spectrum of any of these looks very much like the mass spectrum of a single alkane. Ion 57 is the base peak, and the other fragments are spread out in a Gaussian distribution with a spacing of 14 mass units between them. A typical isoparaffin average mass spectrum, that of Isopar L, taken over a 3-min portion of the chromatogram, is shown in Figure 3.29.

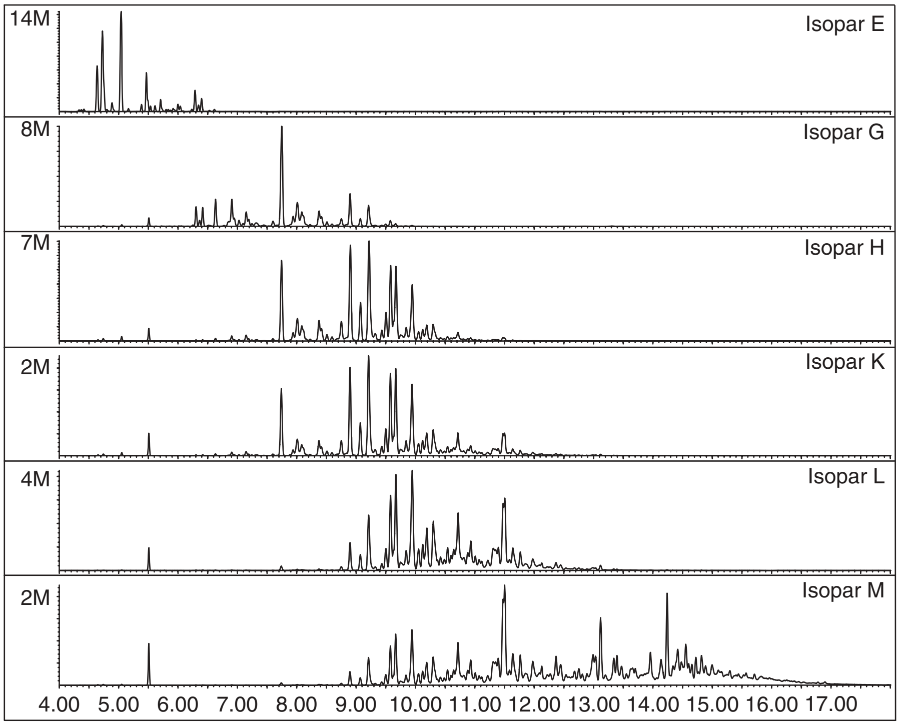

Figure 3.28 Total ion chromatograms of six isoparaffinic hydrocarbon products, Exxon Isopars E, G, H, K, L, and M.

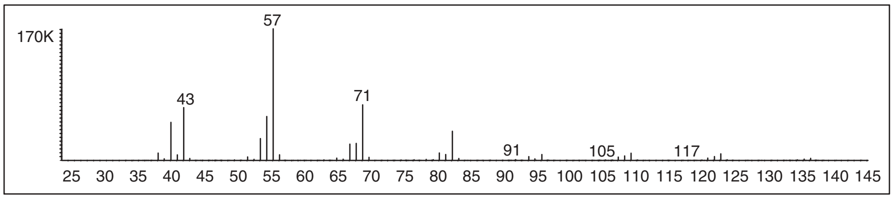

Figure 3.29 Average mass spectrum taken from Isopar L, over the range of 9 to 12 min. This spectrum has an appearance very similar to that of a branched alkane.

The same exercise can be performed on the naphthenic–paraffinic products, several of which are shown in Figure 3.30. The average mass spectrum of a typical naphthenic–paraffinic product, shown in Figure 3.31, looks quite different from that of the isoparaffinic product. While it still exhibits a base peak of 57, there will be a greater abundance of ions 55, 69, and 83 than in the isoparaffinic product.

The identification of other ignitable liquid residues requires an individual examination of the peaks in the chromatogram. It is possible to find just about any single compound in just about any sample. Findings of alcohols, turpentines,

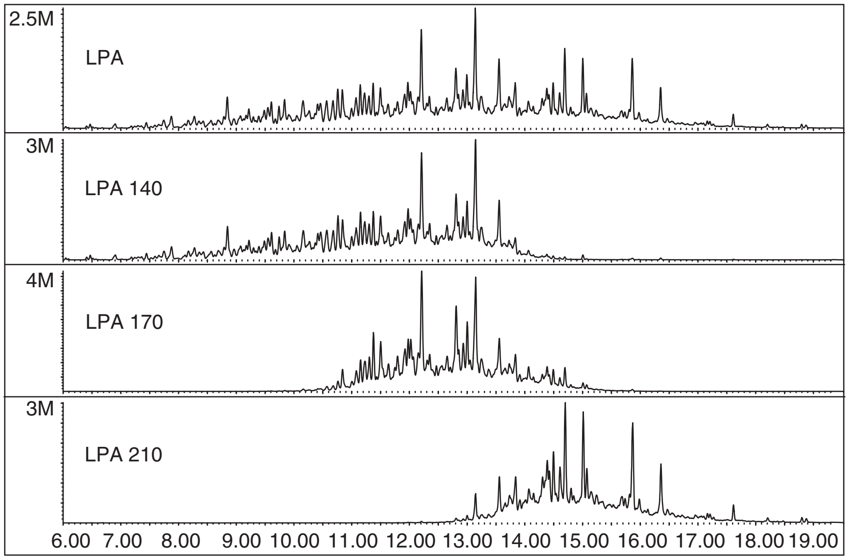

Figure 3.30 Total ion chromatograms of four naphthenic–paraffinic products: Vista LPA, LPA 140, LPA 170, and LPA 210. The numbers in the product names correspond approximately to flash points.

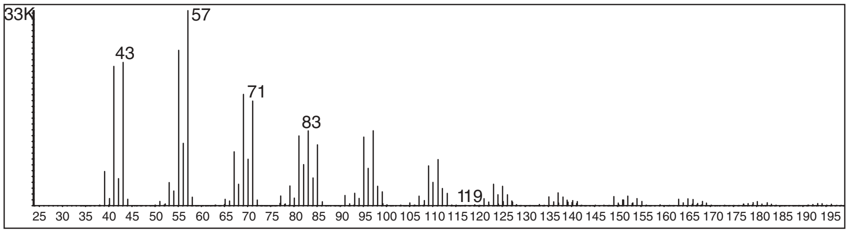

Figure 3.31 Average mass spectrum of Vista LPA taken from 8 to 17 min..

aromatic solvents, and other “flammable” liquids need to be undertaken with great caution. ASTM E1618 recommends not making an identification of these single compounds unless they are present in such concentrations that the signal is *at least two orders of magnitude greater than the background*. Just about every fire debris sample will contain methanol. Just about every fire debris sample will contain toluene. Unless these substances are present at concentrations sufficiently high for the analyst to feel comfortable saying they are not native to the background, they should not be reported.

3.6.2 Improving Sensitivity

The techniques used routinely to isolate and analyze ignitable liquid residues are quite sensitive. In samples with nontenacious substrates, it is possible to detect as little as one-tenth of 1 μL of ignitable liquid residue such as gasoline in 1 kg or more of debris. The task gets a little more difficult when there are active surfaces in the fire debris, such as charcoal, but an activated carbon strip is more tenacious than most debris that one is likely to encounter.

Before attempting to improve the sensitivity of the detection method, it is useful to have some idea about how much sensitivity one already has. This is best accomplished by preparing serial dilutions of an ignitable liquid with an internal standard, and comparing the signals to those generated by extracts from known amounts of ignitable liquid run through the laboratory's adsorption or elution process. In our laboratory we typically use 10-μL standards. The liquid is applied to a piece of filter paper in a 1-quart can, and the filter paper is subjected to ACS headspace concentration, just like any debris sample. When exposed to 10 μL of 75% evaporated gasoline, a 10 mm × 10 mm carbon strip eluted with 500 μL of diethyl ether spiked with 100 ppm of perchloroethylene yields a concentration of approximately 1000 ppm. This is easily detectable. (This translates to a recovery rate of about 5%; 500 μL at 1000 ppm equals 0.5 μL recovered out of 10 μL placed in the sample can.) One microliter in the sample can yields a concentration of approximately 100 ppm in the eluate: again, very easily detectable. At 0.1 μL (10 ppm in the eluate), the signal is noisy and it becomes difficult to extract acceptable ion chromatograms except for ion 105, showing the C_3 alkyl benzenes. To obtain better ion chromatograms, a number of different strategies can be followed. The lower limit of detection is formally defined as that concentration of analyte that will give a signal that is twice the background level; for this analyst, the detection limit is that concentration of analyte that produces a signal large enough to feel comfortable about a determination. One can lower the detection limit either by increasing the signal or by decreasing the background noise. Increasing the signal can be accomplished in a number of ways. One obvious way is to increase the concentration of the analyte. In the case of fire debris eluates from C-strips, this is accomplished by evaporating off some of the solvent. This strategy is not without its costs. It is time consuming and must be done very carefully if one is to avoid losing the analyte along with the solvent. No heat should be used in this process. At some point, the volume of the solution becomes too small to be handled by an autosampler, and this requires manual injection. Although it can be done, easier ways exist to increase the signal. Another obvious strategy is simply to increase the size of the injection. Normally, we inject 1 μL. When we are looking at samples of low concentration, we inject 2 μL.

One way to increase the signal dramatically is to run the gas chromatograph in the *splitless mode*. (Running splitless is one reason that SPME techniques are so sensitive.) Most laboratories use a split ratio of 20 : 1 to 50 : 1. This dilution of the sample improves resolution. Turning off the split will increase the amount of analyte presented to the detector, and the signal will increase by up to a factor of 10. The only cost is likely to be a small decrease in resolution.

Another way to increase the signal from the ions of interest in order to produce better ion chromatograms is to increase the dwell time, the time that the detector spends looking at a particular ion. This is accomplished by using the selected ion monitoring (SIM) mode on a quadrupole instrument. In the full-scan mode, we typically look from 33 to 300 amu, resulting in a dwell time of less than 1 ms per ion. If we ask the instrument to look only at the ions in which we are interested, we can increase the dwell time dramatically. The detector response will increase in proportion to the square root of the dwell time (i.e., if one increases the dwell time by a factor of 4, the signal strength will increase by a factor of 2. Our instrument performs 5.24 scans/s in the full-scan mode, resulting in a dwell time of 0.7 ms. In the SIM mode, there is only about one scan per second, resulting in somewhat rougher peaks, but with a dwell time of 50 ms (an increase of about 70-fold), the signal will increase by about a factor of 8. It is not the increasing signal that is the most attractive feature of SIM, however. It is the reduction in noise. "Changing channels" 267 times per scan (1400 times per second) generates an abundance of noise in the instrument. This is known as *housekeeping noise*. If one looks only at the 23 ions listed in Table 3.4, the noise is reduced by a factor of 60 or more.

Figure 3.32 shows a comparison of the noise generated in the full-scan mode versus the noise generated in the SIM mode. It is easy to see that even with a much smaller signal, the peaks of interest will be plainly visible.

The price to be paid for this increase in sensitivity is a decrease in specificity. (There is no such thing as a free lunch.) The mass spectra produced in the SIM mode do not contain enough ions for a library to match. The analyst can, however, make certain that the ion of interest is the base peak for a particular compound, and make sure that a "qualifier ion" is present. The abbreviated spectrum from a SIM peak can be compared with the SIM spectrum from a known compound, but that is as far as it can go. For this reason, the analyst should apply more rigid criteria when trying to determine whether a SIM pattern exhibits "sufficient similarities" to a standard pattern.

Figure 3.33 shows a comparison of a 10-ppm standard of 75% evaporated gasoline run in the full-scan mode, in the full-scan splitless mode, and in the SIM mode. Looking at just the total ion chromatograms, one might conclude that the SIM mode does not add value compared to the splitless mode, but when one generates the

TABLE 3.4 Ions Used for SIM

Ion	Compounds	Ion	Compounds
31	Methanol	117	Indan, methylindans
43	Alkanes	118	Methylstyrene
45	Ethanol	119	C_4 alkylbenzenes
55	Alkenes, cycloalkanes	120	C_3 alkylbenzenes
57	Alkanes	128	Naphthalene
69	Alkenes	131	Methyl, dimethylindans
71	Alkanes	133	C_5 alkylbenzenes
83	Cycloalkanes, alkenes	134	C_4 alkylbenzenes
85	Alkanes	142	Methylnaphthalenes
91	C_1, C_2 alkylbenzenes	156	Dimethylnaphthalenes
104	Styrene	168	3-phenyltoluene
105	C_3 alkylbenzenes		

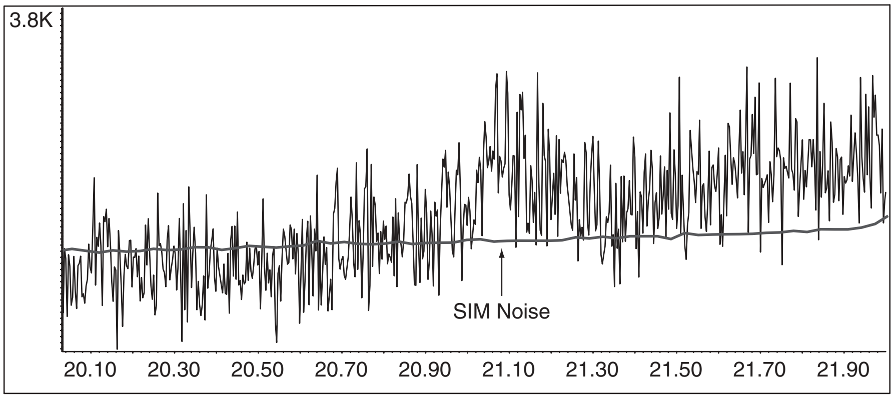

Figure 3.32 Comparison of the baseline noise generated by a full scan with that generated in the SIM mode. *(See insert for color representation.)*

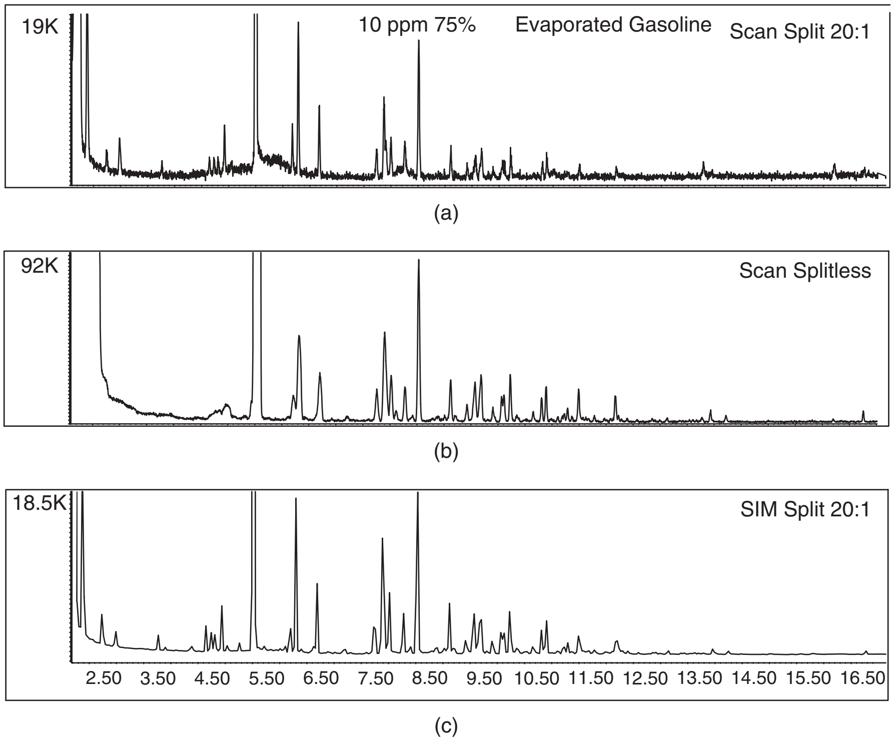

Figure 3.33 10-ppm solution of 75%-evaporated gasoline. (a) Run in the full-scan mode with a 20 : 1 split ratio. In each of these three charts the off-scale peak at 5.5 min is the perchloroethylene internal standard. (b) Run in the full-scan mode with a splitless injection. The abundance is increased by about a factor of 5 over the split injection. (c) Run in the SIM mode with a 20 : 1 split ratio. The abundance is not increased compared to the scan mode, but the noise is reduced significantly.

six extracted ion profiles for gasoline, the SIM chromatograms clearly outshine those produced in the full-scan mode, even with no split. This is demonstrated in Figure 3.34.

This process of increasing sensitivity can be taken even one step further, by using a splitless injection and selected ion monitoring, although in doing so, we are approaching territory where findings of lower and lower amounts may not be meaningful. Figure 3.35 shows the six gasoline ions from a solution of 0.1-ppm standard of 75% evaporated gasoline. This is roughly equivalent to a sample containing 0.001 μL of gasoline (1 nL!). Even at a concentration in the eluate of 0.01 ppm (10 ppb), the baseline does not go completely flat in the SIM splitless mode.

By simply using the tools available, we can see that the sensitivity of the analysis can be increased by four orders of magnitude. It does not seem to this analyst that lowering the detection limit further than this would be worthwhile. Table 3.5 shows

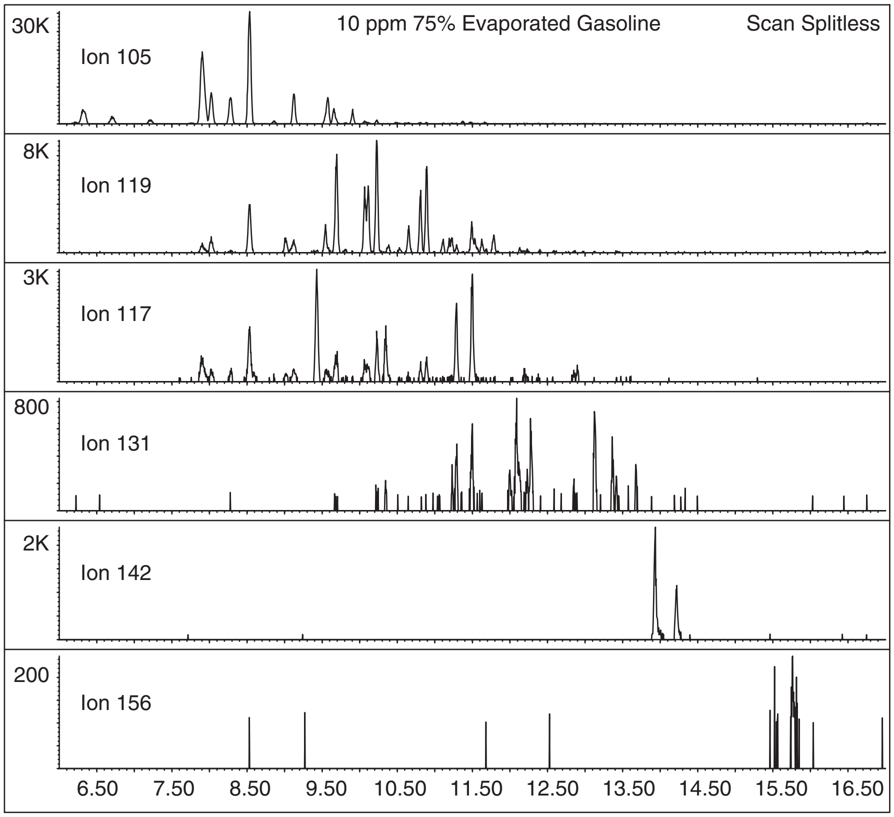

(a)

Figure 3.34 Six ion chromatograms produced by a 10-ppm solution of 75% evaporated gasoline run in the full-scan mode, with splitless injection; (b) run in the SIM mode, with a 20 : 1 split ratio..

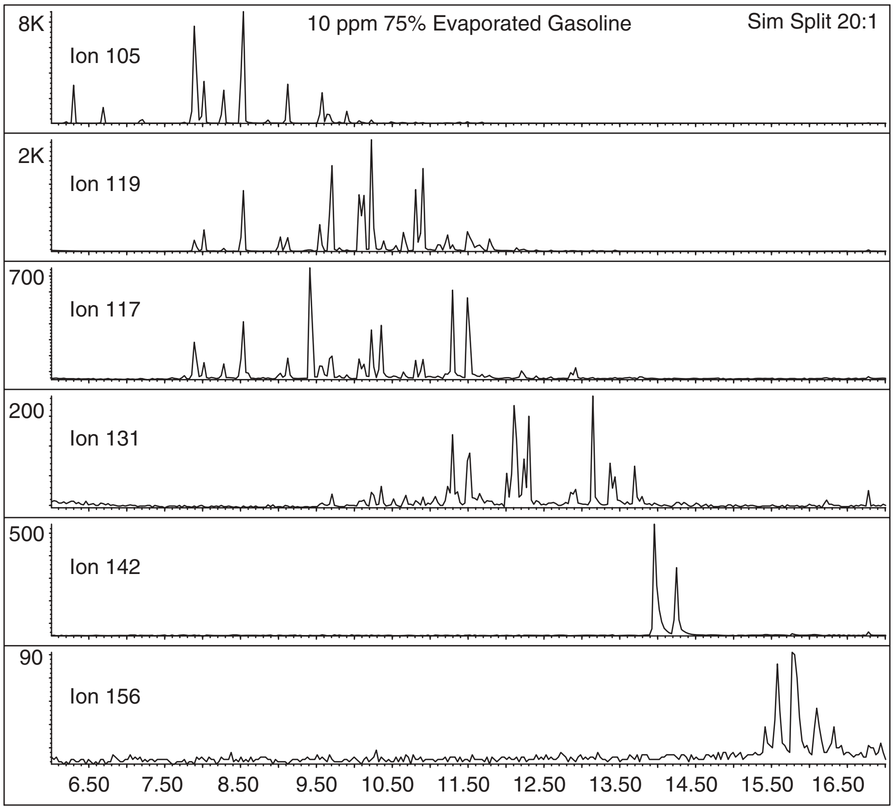

(b)

Figure 3.34 (*Continued*)

a rough order-of-magnitude calculation of the detection limits available using the techniques described above.

3.6.3 Estimating the Degree of Evaporation

Estimating the degree of evaporation is a task that an analyst is occasionally asked to perform. Usually, evaporation is not an issue, and frequently, even if it is an issue, estimating how much of the original volume of a sample may have evaporated is difficult. Comparing degrees of evaporation may be helpful in source or common source determinations, as will be discussed later.

Unless we know what the ignitable liquid was to start with, it is next to impossible to determine how much it has evaporated. This is particularly true for distillates, which may range from C_7 through C_{12} in one batch and C_9 through C_{13} in the next batch of the same product. Similarly, kerosene and diesel fuel formulations change with latitudes and seasons. So, unless there is a container that is suspected of being

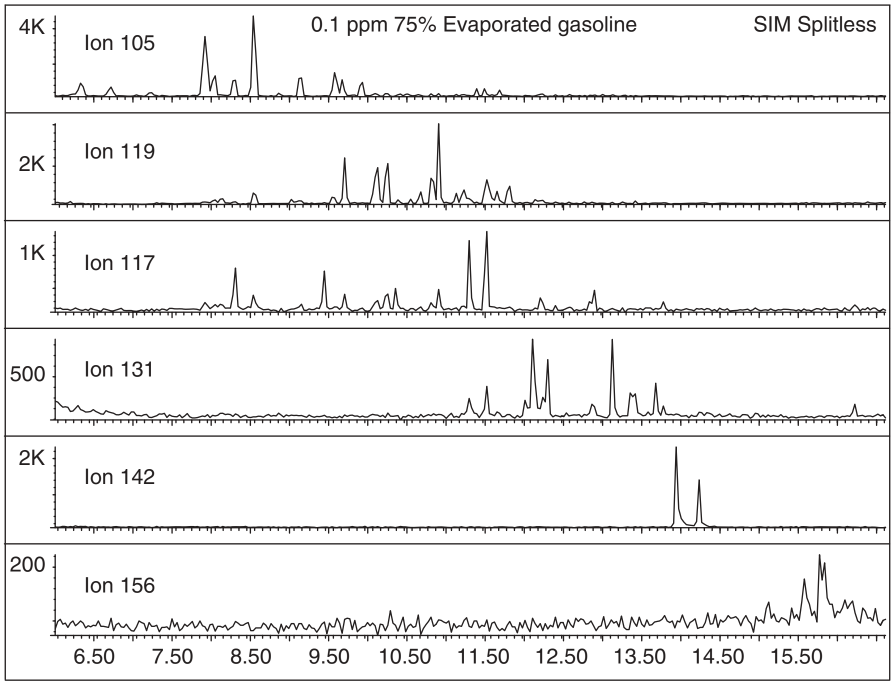

Figure 3.35 Six gasoline ions produced by a 2-μL splitless injection of a solution of 0.1 ppm 75%-evaporated gasoline.

TABLE 3.5 Detection Limits for Four Instrumental Configurations

Method	LLD (Solution)[a]	LLD (Sample)[b]
Full scan, split 20 : 1, 1-μL injection	100 (ppm)	1.0(μL)
Full scan, splitless, 1-μL injection	10	0.1
SIM, split 20 : 1, 2-μL injection	1	0.01
SIM, splitless, 2-μL injection	0.1	0.001

[a] Concentration at which this analyst feels comfortable calling a sample positive.
[b] Assumes little or no competition from the sample, and a 5% recovery rate.

the source of the residue, this is a task that simply cannot be accomplished, with the possible exception of gasoline.

Fresh from the pump, gasolines look pretty much alike, particularly with respect to major components, and they change in a reasonably predictable fashion as a result of exposure to fire. The issue of evaporation frequently arises in the context of trying to decide whether gasoline was present in a sample before the fire or introduced during or after the fire. It sometimes happens that firefighters bring gasoline-powered equipment to fire scenes, and it is sometimes necessary to eliminate such equipment as the source of residue found in a debris sample. A number of different techniques can be brought to bear on this question. The simplest way to make the estimation is to compare the total ion chromatograms. In fresh gasoline, toluene will be the tallest peak, and there will be numerous light (C_5 to C_8) alkane peaks to the left of the toluene peak. By the time that gasoline is 25% evaporated, many of these light alkane peaks have disappeared. Also, in a 25%-evaporated gasoline sample, the tallest peak is the *m*- and *p*-xylene peak.

The analyst's library of gasolines should include gasoline in many stages of evaporation, as well as fresh gasoline. In our laboratory, we keep gasoline standards on file at 0-, 25-, 50-, 75-, 90-, 98- and 99%-evaporated. Performing a rough estimation of the degree of evaporation happens every time we identify gasoline, simply because it is necessary to put a hard copy of at least one standard in the file (ASTM E1618), and it is desirable for the reference chromatogram to match the sample chromatogram as closely as possible.

By the time gasoline is 50% evaporated, the toluene peak is significantly shorter than the *m*- and *p*-xylene peak, and in fact, is shorter than the ethylbenzene peak. The pseudocumene peak, meanwhile, has grown so that it is almost as tall as the *m*- and *p*-xylene peak. At 75% evaporation, the pseudocumene peak is the tallest peak in the chart, and the toluene peak has almost disappeared. At 90% evaporation, the toluene peak is gone and the xylenes are far shorter than the C_3 alkylbenzenes. The C_4 alkylbenzenes, meanwhile, have grown as tall as most of the C_3 alkylbenzenes, but pseudocumene is still the tallest peak in the chart. By the time the sample is 90% evaporated, the C_4 alkylbenzenes are taller than the C_3 alkylbenzenes.

A more quantitative technique for estimating evaporation involves use of the mass spectrometer. Collecting the average mass spectrum from C_6 to C_{13} (about 3 to 13 min in our laboratory) yields an average that changes predictably from one stage of evaporation to the next. In fresh gasoline, ions 91 and 105 are the tallest, followed by 119, 43, 57, and 71. The 43 fragment is produced by the light branched alkanes to the left of toluene. At 25% evaporation, 91 and 105 still dominate the mass spectrum, but the 57 ion is now taller than the 43. At 50% evaporation, ion 91 is no longer equal to 105, and 119 and 120 exceed the height of ion 57.

At 90% evaporation, the 119 ion is almost equal in size to the 105 ion, and ion 134 is beginning to catch up with ion 91. At 98% evaporation, ion 119 is the tallest, followed by 105, 134, and 91. Figure 3.36 shows these six average mass spectra. Keep in mind that the substrate can contribute to the spectra, thus making this approach more difficult in some cases.

Any attempt to estimate evaporation should be done by comparing "apples to apples." Although it is hoped that the analyst has optimized the isolation process so that an eluate from an ACS looks like the liquid in solution that has not gone through the ACS process, standards should be eluates rather than solutions prepared simply by diluting the liquids. Further, samples and standards should have approximately equal concentrations. More concentrated samples tend to appear more evaporated, due to the effects of displacement (Newman et al., 1996).

There are certain substrates that hold certain compounds preferentially, and substrate effects need to be considered before reporting a sample as having a problematic evaporation level. Concrete, for example, tends to retain lighter hydrocarbons, making residues appear less evaporated than the fire scene conditions

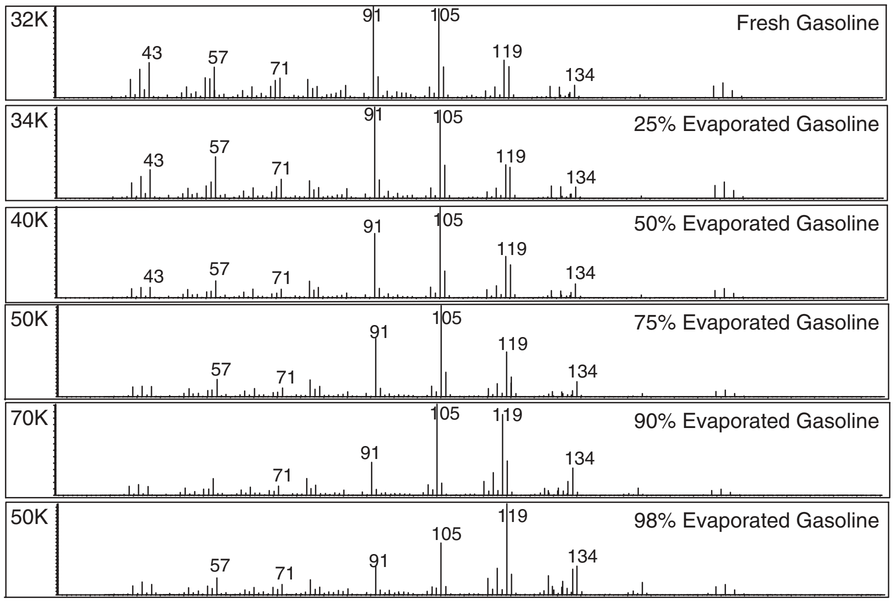

Figure 3.36 Average mass spectra from 3 to 13 min of fresh gasoline, followed by the same spectrum taken for gasoline evaporated 25, 50 75, 90, and 98%.

would indicate. One of the more important reasons for estimating the degree of evaporation is to allow for an “apples to apples” comparison when attempting to match a questioned sample to a known source.

3.6.4 Identity of Source

The ability to match a weathered ignitable liquid residue with a proposed source of unburned ignitable liquid has eluded fire debris analysts until just recently. The environmental forensics community has developed some tools to identify the sources of petroleum spills, but they have the advantage of having huge quantities of material available for characterization. Petroleum hydrocarbons contain trace quantities of biomarkers, substances that have changed little since they were first synthesized inside a living organism. The relative amounts of these biomarkers frequently allows for the identification of the source of a major spill. Polycyclic aliphatic and aromatic hydrocarbons are also useful in “fingerprinting” hydrocarbons. Examination of additives, such as oxygenates, or in the case of old gasolines, the alkyl lead compounds, can also provide clues as to the identity of the source of a spill.

There are three levels of “control” on the composition of petroleum. The primary control is the genesis of the petroleum: its “genetic” features, including its geographic location, the original source of the fuel (terrestrial or marine), and the conditions in the rock where the oil resided for millions of years.

The second level of control affecting the hydrocarbon fingerprint is processes imposed on the hydrocarbon by the refiner. Activities such as distillation, hydrocarbon cracking, isomerization, alkylation, and blending change the composition of the fuel. The tertiary controls are those that occur after the petroleum product leaves the refiner. The most important tertiary control of interest to fire debris analysts is, of course, weathering, or evaporation. Mixing in the service station tank is also critical. Because the tanks are seldom empty, each time a new delivery is made, the fuel service station tank assumes a new temporary identity. For a more extended discussion on chemical fingerprinting in the environmental forensic science arena, see *Introduction to Environmental Forensics* (Murphy and Morrison, 2002).

Some of the early work to identify the source of ignitable liquid residues from fire debris was done by Dale Mann (1987a,b). Using a 60-m column, Mann learned that it was possible to make comparisons on samples of fresh gasoline and to correctly identify the source by examining the relative ratios of the light hydrocarbons, ranging from *n*-pentane to *n*-octane. It is, unfortunately, exactly these hydrocarbons that are lost before a gasoline is even 25% evaporated. Mann's second paper describes the limitations of making comparisons between ignitable liquid residues and fresh samples, due to the contamination of residues with pyrolysis products, changes introduced by the isolation method, and the loss of volatiles resulting from weathering. Comparison of residues with fresh sources was, therefore, not studied frequently after Mann's extensive work. More recently, however, Dolan and Ritacco reported that they had devised a way to measure the relative abundances of 20 peak pairs in gasolines and were able to identify the source of 30 samples that were evaporated to 25% and 50%. This process uses the relative abundances of branched alkanes that occur generally in the center of the gasoline chromatogram. Because the peaks are sequential, eluting only a few seconds apart, the relative ratios are not affected significantly by evaporation. The peaks examined are all minor components. The major components of gasoline are similar enough to each other that they are not useful in discriminating between sources (Dolan and Ritacco, 2002).

Following up on this work, Wintz and Rankin applied principal components analysis only to learn that while the ratios within the pairs show little if any change between unevaporated and 50%-evaporated gasoline, only a few of those pairs were actually useful in distinguishing between one gasoline source and another, and some of those pairs occurred in the light petroleum distillate range, and so are unlikely to be found in gasolines evaporated to more than 50% (Wintz and Rankin, 2004).

Further work by Barnes and Dolan identified six ratios of sequentially eluting minor components in 50% evaporated gasoline and four ratios in 75% evaporated gasoline. Using these ratios, they were able to discriminate successfully among 16 gasoline samples evaporated 50% and 10 gasoline samples evaporated 75%. They used a blind study to confirm the ability of the technique to identify the source of three gasolines correctly matched against a library of 10 candidate sources. The technique uses a 60-m column, and is automated using a target compound program. The program takes the abundance of the base peak from the mass spectrum of each compound of interest and divides the base peak abundance of the later eluting compound by the base peak abundance of the earlier eluting compound. As well as this technique appears to work, the authors state that it is most useful for *eliminating* a source, and caution that the analyst's statement of conclusions reflect that a common origin is indicated but is not a certainty. (Barnes et al., 2004).

A somewhat simpler approach, using an analysis of naphthalene, methyl, and dimethylnaphthalenes has been reported by Sandercock and Du Pasquier in a three-part series published in 2003 and 2004 (Sandercock and Du Pasquier, 2003, 2004a,b). The column used was an everyday 30-m HP-5 column of the type used for routine fire debris analysis. A SIM method was used, and in order to have all of the peaks capable of being integrated at the same approximate level of precision, the signal for naphthalene and the methylnaphthalenes was selectively reduced by monitoring ions 127 and 129 rather than 128 for naphthalene, and 141 and 143 rather than 142 for methylnaphthalene. The base peak of 156 was used for the dimethylnaphthalenes. Sandercock and Du Pasquier started their research by pretreating the gasoline with an alumina column to separate out both polynuclear aromatic hydrocarbons and the polar alkylphenols, but learned that the alkylphenols did not provide much help in discriminating among different samples. The PAH ratios do not change significantly upon evaporation and seem to be a characteristic imparted by the refinery.

One way to compare any type of ILR to a suspected source, or to see if two ILRs within the same class came from the same source is to compare the average molecular weights. When exposed to a fire, ignitable liquid evaporates, and its average molecular weight necessarily increases. No predictable experience will cause the average molecular weight of the fire-exposed liquid to decrease. If the suspected source of an ignitable liquid, which was not exposed to the fire, exhibits a higher average molecular weight than the residue extracted from samples collected at the scene, the suspected source can be conclusively eliminated.

In doing comparisons of gasolines or any ignitable liquid, the best approach takes the proposed source liquid, evaporates some of it so that it matches the residue in terms of carbon number range, and then subjects it to the same separation procedure that was applied to the questioned sample. Once this has been accomplished, a detailed examination of the finer points in the chromatogram can be carried out.

A really good match between chromatograms not only matches the peaks but also matches the valleys. Examining multiple extracted ion chromatograms for both quality and quantity is a necessary step when doing comparisons. This practice has been codified with respect to the analysis of distillates heavier than kerosene in ASTM D5739-06, *Standard Practice for Oil Spill Source Identification by Gas Chromatography and Positive Ion Electron Impact Low Resolution Mass Spectrometry* (ASTM D5739-06, 2006). The approach in this method compares 24 extracted ion chromatograms of polynuclear aromatics and certain biomarkers. Although this method is not applicable to kerosene and lighter mixtures such as gasoline, the general approach seems workable. Given the state of the art, it is probably a good thing that comparisons are not frequently requested.

3.7 REPORTING PROCEDURES

The analyst's report can be one of the most important documents generated during the investigation of a fire. The report should therefore be written carefully enough so that readers are not misled. The objective of fire debris analysis is to determine whether any foreign ignitable liquid residue is present in a sample.

The report from a forensic science laboratory is supposed to be a scientific report. As such, it should include an introduction, a section detailing the test methods and results, and a discussion and conclusion section, if necessary. The laboratory report should also state what was done with the evidence. There is an unfortunate tendency in some agencies to provide checklists and merely state "positive for gasoline" or "negative for accelerants" rather than preparing a real laboratory report. Reading such a report, a reviewer cannot tell whether the analyst used a GC–MS or a Ouija board!

Given that it is now possible to store templates with all possible results in them, the excuse that it is "too time-consuming" to prepare an understandable narrative report is unacceptable. Although laboratory reports are intended to assist investigators, they should also be understandable to people asked to review the work. At a minimum, a report should include the following:

- Identification of the fire in question.
- A description of the how the sample was delivered to the laboratory, when, and by whom.
- A description of the samples, including container size, substrate material, and a reported location from where the sample was collected.
- A description of the isolation procedure used to separate the ILR from the sample substrate.
- A description of the analytical technique applied to the sample extract.
- The results of the analysis of the data.
- A discussion of the meaning of the results if there is any chance of misinterpretation or misuse.

In the discussion section, the analyst can provide examples of potential sources for whatever ignitable liquid residues may have been identified. This is also an appropriate place to put in a disclaimer about the possibility that the ILR may not be foreign to the background.

- A conclusion or bottom line understandable even to an attorney is a helpful thing.
- A sentence stating what has happened to the sample.

If a substance is found that is natural or incidental to the background, it is the analyst's job to say so. ASTM E1618 allows for the inclusion of disclaimers on both positive and negative reports. With respect to negative reports, a disclaimer to the effect that negative results do not preclude the possibility that ignitable liquids were present at the fire scene is suggested as an aid to help readers avoid misunderstanding the report. Similarly, in the case of a positive report, the standard states: "It may be appropriate to add a disclaimer to the effect that the identification of an ignitable liquid residue in a fire scene does not necessarily lead to the conclusion that a fire was incendiary in nature. Further investigation may reveal a legitimate reason for the presence of ignitable liquid residues." Certainly, a finding of gasoline in the living room is noteworthy. It is not so noteworthy to find gasoline in the basement near the chain saw. On occasion, it seems that the fire debris analyst is too eager to "help" with an arson investigation. Reporting a sample as being "positive" when all the analyst has identified is background compounds is neither helpful nor morally defensible. Karen and Paul Stanley of Akron, Ohio, were falsely accused of setting a fire that killed their infant son. The fire debris samples were submitted to a laboratory that reported finding turpentine in a sample consisting of charred Douglas fir. The laboratory analyst had absolutely no business reporting turpentine, or at the very least, should have reported the finding of turpentine with a huge disclaimer stating that the turpentine was indistinguishable from naturally occurring turpentine found in coniferous woods. The case was dismissed after the prosecutor understood the significance (or lack thereof) of the chemist's findings (Trexler, 2002).

A similar case of an analyst being too "helpful" occurred in a homicide case in Georgia, where the laboratory analyst reported finding "toluene, a flammable liquid" on a suspect's "clothing." Actually, the "clothing" included a pair of tennis shoes, and the examination of a new pair of tennis shoes directly from the shoestore revealed the presence of a high concentration of toluene. A closer examination of the first analysis revealed that in addition to toluene, diethylene glycol and butylated hydroxytoluene (BHT) were also present in the suspect's shoes and the exemplar shoes in identical relative proportions. Less than a year earlier, there had been a presentation on the analysis of suspects' shoes in arson cases at an AAFS seminar (Cherry, 1996). Perhaps if the analyst had attended that presentation (or read the proceedings), the toluene would have been reported properly as being native to the sample background, or more appropriately, not reported at all. Typical positive and negative reports from our laboratory are shown as Figure 3.37.

3.8 RECORD KEEPING

Each case file that includes a positive identification of an ignitable liquid residue should contain not only the sample charts but also a standard to which that sample can be compared. This means that both the sample and the standard should be printed with the same time scale, so that the data can easily be reviewed. Even if there is no criminal prosecution, the analyst should be aware that many fire cases involve civil litigation, so the case file should be kept for a reasonable period of time.

Electronic data files should also be protected and stored. There are only a few versions of GC–MS software in widespread use, so it is possible for one analyst

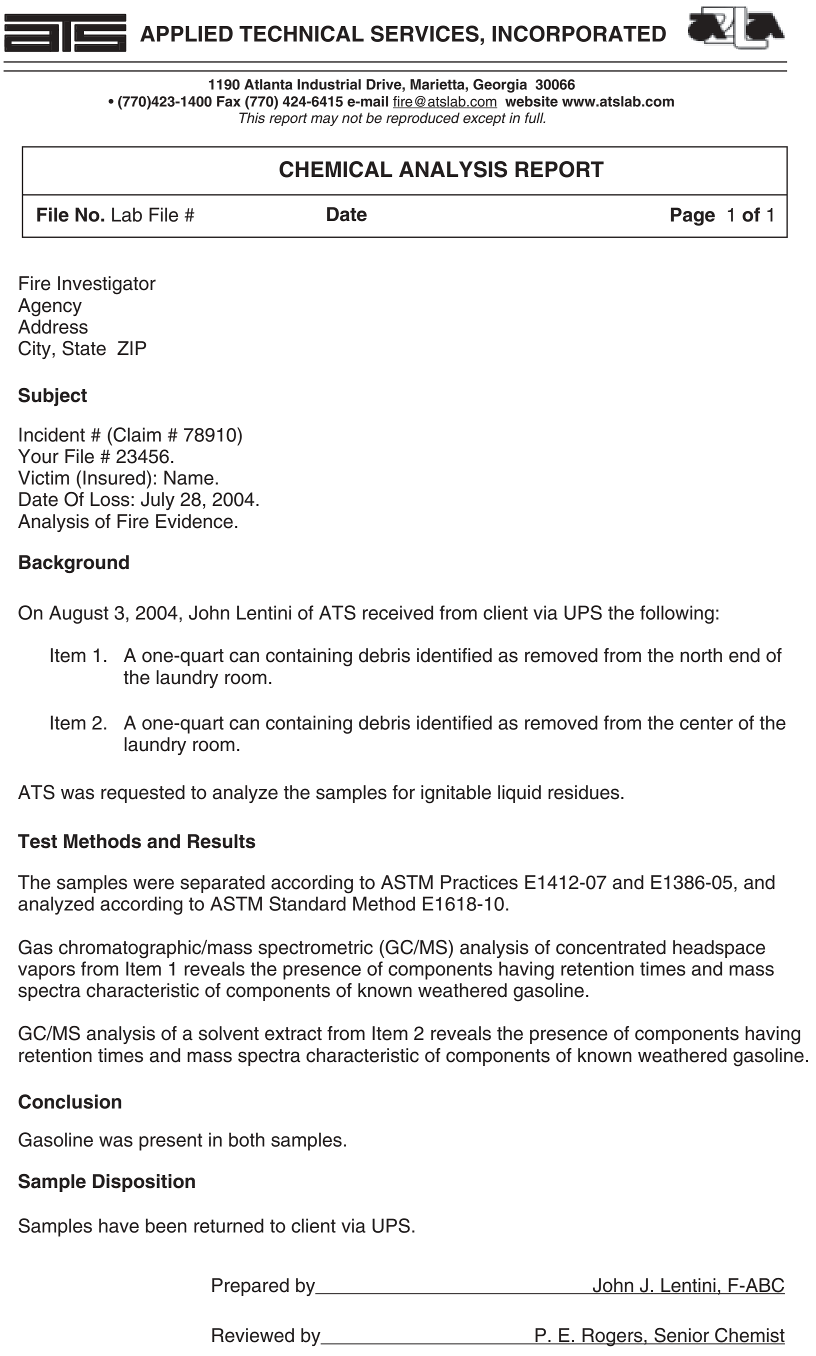

APPLIED TECHNICAL SERVICES, INCORPORATED

1190 Atlanta Industrial Drive, Marietta, Georgia 30066
• (770)423-1400 Fax (770) 424-6415 e-mail fire@atslab.com **website www.atslab.com**
This report may not be reproduced except in full.

CHEMICAL ANALYSIS REPORT		
File No. Lab File #	**Date**	**Page** 1 **of** 1

Fire Investigator
Agency
Address
City, State ZIP

Subject

Incident # (Claim # 78910)
Your File # 23456.
Victim (Insured): Name.
Date Of Loss: July 28, 2004.
Analysis of Fire Evidence.

Background

On August 3, 2004, John Lentini of ATS received from client via UPS the following:

Item 1. A one-quart can containing debris identified as removed from the north end of the laundry room.

Item 2. A one-quart can containing debris identified as removed from the center of the laundry room.

ATS was requested to analyze the samples for ignitable liquid residues.

Test Methods and Results

The samples were separated according to ASTM Practices E1412-07 and E1386-05, and analyzed according to ASTM Standard Method E1618-10.

Gas chromatographic/mass spectrometric (GC/MS) analysis of concentrated headspace vapors from Item 1 reveals the presence of components having retention times and mass spectra characteristic of components of known weathered gasoline.

GC/MS analysis of a solvent extract from Item 2 reveals the presence of components having retention times and mass spectra characteristic of components of known weathered gasoline.

Conclusion

Gasoline was present in both samples.

Sample Disposition

Samples have been returned to client via UPS.

Prepared by John J. Lentini, F-ABC

Reviewed by P. E. Rogers, Senior Chemist

Figure 3.37 (a) Typical positive report from our laboratory; (b) typical negative report from our laboratory.

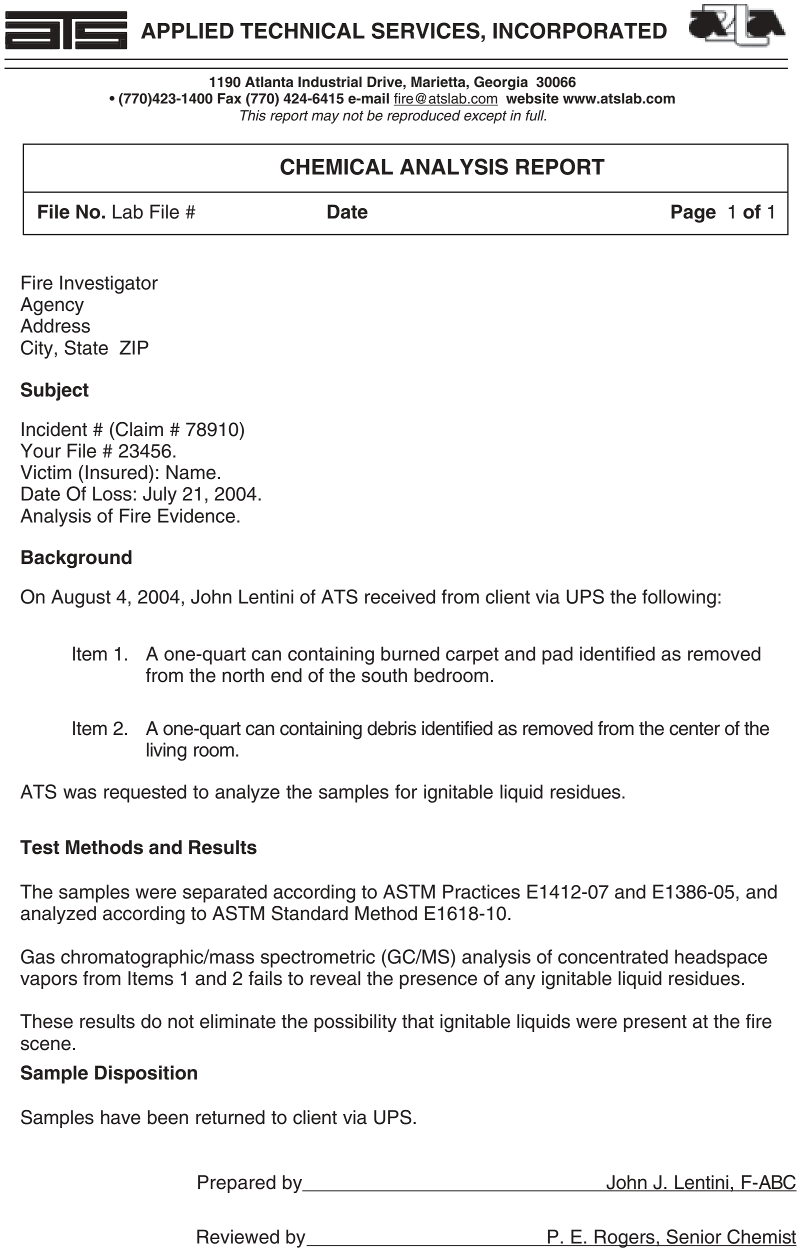

APPLIED TECHNICAL SERVICES, INCORPORATED

1190 Atlanta Industrial Drive, Marietta, Georgia 30066
• (770)423-1400 Fax (770) 424-6415 e-mail fire@atslab.com **website www.atslab.com**
This report may not be reproduced except in full.

CHEMICAL ANALYSIS REPORT		
File No. Lab File #	**Date**	**Page** 1 **of** 1

Fire Investigator
Agency
Address
City, State ZIP

Subject

Incident # (Claim # 78910)
Your File # 23456.
Victim (Insured): Name.
Date Of Loss: July 21, 2004.
Analysis of Fire Evidence.

Background

On August 4, 2004, John Lentini of ATS received from client via UPS the following:

Item 1. A one-quart can containing burned carpet and pad identified as removed from the north end of the south bedroom.

Item 2. A one-quart can containing debris identified as removed from the center of the living room.

ATS was requested to analyze the samples for ignitable liquid residues.

Test Methods and Results

The samples were separated according to ASTM Practices E1412-07 and E1386-05, and analyzed according to ASTM Standard Method E1618-10.

Gas chromatographic/mass spectrometric (GC/MS) analysis of concentrated headspace vapors from Items 1 and 2 fails to reveal the presence of any ignitable liquid residues.

These results do not eliminate the possibility that ignitable liquids were present at the fire scene.

Sample Disposition

Samples have been returned to client via UPS.

Prepared by______________________________John J. Lentini, F-ABC

Reviewed by________________________P. E. Rogers, Senior Chemist

Figure 3.37 (*Continued*)

to review another's raw data. Such reviews should be facilitated by preservation of the data. Instrumental data such as tune reports or spectrum scans should be kept, if only to keep track of the performance of the instrument. Certainly, if a case file has matching standards and samples, there is no need to review the tune report for that particular week. No matter what happened, the instrument tune parameters would not cause a false positive. Blanks, on the other hand, should be stored somewhere, particularly for those cases where the detection limit of the procedure is being reduced.

3.9 QUALITY ASSURANCE

The factors affecting the overall quality of a laboratory's work product are numerous. For this reason, accreditation programs have been set up so that an outside agency can verify that a laboratory is performing appropriately. Each laboratory should have a detailed written procedure for the examination of fire debris. As part of the quality assurance program, fire debris analysts should participate in proficiency testing at least once a year. This can be internal or external, but certainly, external proficiency testing carries more credibility. Proficiency tests are available from a number of commercial suppliers, but it is not necessary to actually purchase proficiency tests. Groups of analysts can form round-robin groups and take turns preparing samples for each other. Because it is not necessary to demonstrate one's ability to place a strip in a can, samples can be prepared by exposing 10 strips to the same sample, then just mailing the strips. By doing regular proficiency tests or round robins, analysts can monitor their own performance, and laboratory directors can be assured that their analysts are performing adequately.

Individual professional development and continuing education is an important component of any quality assurance program. Analysts should take the time to keep up with the literature and should attend professional meetings when possible. Although the science of fire debris analysis is reasonably settled, this is an interesting field with a large cadre of professionals who perform and publish research on a continuing basis. Keeping up with this research is a professional responsibility.

The ultimate sign that an analyst is keeping up with the profession is the decision to become certified. The American Board of Criminalistics offers a certification in fire debris analysis. Certification can be maintained only by continuing education and annual participation in external proficiency testing. Becoming certified is a way for an analyst to demonstrate that he or she cares about professional development. Supporting individual certification is a way that a laboratory director can assure his or her employees that the agency also cares.

CONCLUSION

The isolation and identification of ignitable liquid residues from fire debris samples is an important part of the fire investigation process. The fire debris analyst should be familiar with the common techniques of fire investigation and understand the language used by investigators. The analyst must also understand the stakes involved. If an ignitable liquid residue is identified, a hypothesis that a fire was incendiary may be supported, and a long litigation process may ensue in either civil or criminal court. The laboratory results are often the deciding factor in a prosecutor's decision to indict or an insurance company's decision to resist a claim.

The technology for fire debris analysis has improved dramatically over the last 20 years, to the point where our methods are now as sensitive as they need to be, and possibly more sensitive than they need to be. The analytical procedure requires focus and creativity, but also adherence to the generally accepted criteria for making an identification. The analysis should be conducted and reported in such a way that it is capable of being reviewed by another analyst. There is no reason that a reviewing analyst should reach a different conclusion, and such reviews should not only be expected but be welcomed. Science is based on "multiple witnessing." Fire debris analysts need to be acutely aware of the stakes involved in what they are doing and of the need to communicate effectively about the meaning of their findings.

REFERENCES

AA Notes (1982). *Arson Anal. Newsl*., **6(3)**:57.

Armstrong, A., and Lentini, J. (1997). Comparison of the eluting efficiency of carbon disulfide with diethyl ether: the case for laboratory safety. *J. Forensic Sci.*, **42(2)**:307.

ASTM D3257-07 (2007). *Standard Test Methods for Aromatics in Mineral Spirits by Gas Chromatography*. Annual Book of Standards, Vol. 6.03. West Conshohocken, PA: American Society for Testing and Materials.

ASTM D5739-06 (2006). *Standard Practice for Oil Spill Source Identification by Gas Chromatography and Positive Ion Electron Impact Low Resolution Mass Spectrometry*. Annual Book of Standards, Vol. **11.02**. West Conshohocken, PA: American Society for Testing and Materials.

ASTM E1386-10 (2010). *Standard Practice for Separation and Concentration of Ignitable Liquid Residues from Fire Debris Samples by Solvent Extraction*. Annual Book of Standards, Vol. **14.02**. West Conshohocken, PA: American Society for Testing and Materials.

ASTM E1387-01 (2001). *Historical Standard Test Method for Ignitable Liquid Residues in Extracts from Fire Debris Samples by Gas Chromatography*. Annual Book of Standards, Vol. **14.02**. West Conshohocken, PA: American Society for Testing and Materials.

ASTM E1388-05 (2005). *Standard Practice for Sampling of Headspace Vapors from Fire Debris Samples*. Annual Book of Standards, Vol. **14.02**. West Conshohocken, PA: American Society for Testing and Materials.

ASTM E1412 (2007). *Standard Practice for Separation of Ignitable Liquid Residues from Fire Debris Samples by Passive Headspace Concentration with Activated Charcoal*. Annual Book of Standards, Vol. **14.02**. West Conshohocken, PA: American Society for Testing and Materials.

CHAPTER 4

Forensic Examination of Soils

RAYMOND C. MURRAY
University of Montana, Missoula, Montana

Summary Soils and sediment constitute excellent trace evidence in both criminal and civil cases because there are an almost unlimited number of identifiable soil types based on the content of rocks, minerals, glasses, human-made particles, and chemicals. Forensic examination commonly identifies the original geographic location of soils associated with a crime, thus assisting an investigation. Studies of soil and related material samples associated with a suspect and crime scene can produce evidence that the samples had or did not have a common source, thus indicating whether or not the suspect was ever at a particular place. Gathering intelligence and mine, gem, and art fraud studies often use the methods of forensic geology.

4.1 Introduction 110
4.2 Murder and the pond 111
4.3 Oil slicks and sands 113
4.4 Medical link 114
4.5 Examination methods 114
  4.5.1 Color 115
  4.5.2 Particle-size distribution 117
  4.5.3 Stereo binocular microscope 120
  4.5.4 Petrographic microscope 122
  4.5.5 Refractive index 124
  4.5.6 Cathodoluminescence 124
  4.5.7 Scanning electron microscope 125
  4.5.8 X-Ray diffraction 126
4.6 Chemical methods 127
  4.6.1 FTIR and Raman spectroscopy 128
4.7 Looking ahead 129
References 130

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

4.1 INTRODUCTION

As with so many other types of physical evidence, forensic geology began with the writings of Sir Arthur Conan Doyle, who wrote the Sherlock Holmes series between 1887 and 1927. He was a physician who apparently had two motives: writing salable literature and using his scientific expertise to encourage the use of science as evidence.

In 1893, Hans Gross, an Austrian forensic scientist, wrote a handbook for examining magistrates (Gross, 1893) in which he suggested that “perhaps the dirt on someone’s shoes could tell more about where a person had last been than toilsome inquiries.” It was only a matter of time before these ideas from an author of fiction and a criminalists’ handbook would appear in a courtroom.

In 1908, a German chemist, Georg Popp, undertook the study of soil evidence in the homicide case of Margarethe Filbert. In this case Popp was able to demonstrate that soil samples from two locations associated with the crime had characteristics similar to the sample collected from the suspect’s shoes. In addition, using soil from the suspect’s shoes he was able to show a sequence of events in soil accumulation that was consistent with the theory of the crime, and he found no soil evidence that supported the suspect’s alibi (Thorwald, 1967).

A century later, the use of geologic materials in criminal and civil cases is commonplace (Pye and Croft, 2004). Public and private laboratories around the world have trace evidence examiners qualified to examine soil, glass, and related material (Murray, 2011; Pye, 2007; Ruffell and McKinley, 2008).

Forensic geology studies vary in scope. A common type of investigation involves identifying a material that is key to a case: for example, examining pigments in a painted picture or material in a sculpture when authenticity or value is at issue. Identification is also important in questions of mining, mineral, or gem fraud to determine if the material is what its sellers claim it to be. Identification of fire-resistant safe insulation on a person or individual’s property may provide probable cause for further investigation.

Beyond identification, forensic geologists can also look at the source of particular material. Here the examiner needs a broad knowledge of the geology and the best geologic and soil maps to answer questions. For example, if the soil on a body does not match the location where the body is found, from where was the body moved? Similarly, examiners can compare two samples, one associated with the suspect and the other collected from the crime scene, to see if they had a common source. For example, does the soil on the suspect’s shoe have characteristics similar to those of the soil type collected at the crime scene?

Another new developing area of forensic geology is its use in intelligence work. For example, a person may claim never to have been to a particular location, but is then found with rocks from that spot, thus linking the person to the location. Remember the outcrop you saw behind Osama bin Laden on your television screen after 9/11? What was the location? A geologist who has done field work in the area would be able to locate that outcrop, and that actually happened: A geologist, John Shroder, was able to identify the region where bin Laden had been sighted in

Afghanistan in 2001. As with all class evidence, geologic evidence rarely provides a unique solution for which the geologic mind cannot imagine another possibility (Murray, 2005). But there are some exceptions, as illustrated in the following two cases.

4.2 MURDER AND THE POND

The murder of John Bruce Dodson produced one of the most interesting cases in the entire history of forensic geology. Here, the geologic evidence is unequivocal, in that it tied the suspect directly to the crime and eliminated the suspect's alibi. Most important, the investigator of the crime recognized the potential importance of the geologic evidence and arranged for an examination of that evidence. The testimony of the forensic geologist was critical to the prosecution of the case. The case began on October 15, 1995, when John Dodson was found dead while on a hunting trip with his wife of three months, Janice. The scene was high in the Uncompahgre Mountains of western Colorado.

At first glance it appeared to be a hunting accident. However, the autopsy revealed two bullet wounds to the body and one bullet hole through John's orange vest. The investigation showed that the Dodsons were camped near other hunters, one of whom was a Texas law enforcement officer. He responded to Janice's frantic call that her husband had been shot. She was standing about 200 yards from the camp in a grassy field along a fence line. The officer determined that John was dead and started the process of getting help. Prior to calling for help, Janice had returned to her camp and removed her hunting coveralls, which were covered with mud from the knees down. She later told investigators that she had stepped into a mud bog along the fence near the camp. Investigators found a. 308-caliber shell case approximately 60 yards from the body. In addition, they found a. 308-caliber bullet in the ground on the other side of the fence, which created a direct line from the location of the case to the body to the bullet.

Janice's ex-husband, J. C. Lee, was camped three-fourths of a mile from the Dodsons. Janice knew that the site was his favorite camp location. He naturally came under suspicion. However, Lee was hunting far away from camp with his boss at the time of the shooting. Most important, Lee reported to investigators that while he was out hunting, someone had stolen his. 308 rifle and a box of. 308 cartridges from his tent. Winter comes early at 9000 ft in the Umcompahgre, and little more could be done at the scene. However, investigators Bill Booth, Dave Martinez and Wayne Bryant returned during the summers of 1996, 1997, and 1998 and searched for the rifle and other evidence. They tried to search every place a weapon could have been hidden. They combed the entire area, including ponds, with metal detectors in the hope of finding the rifle; it has never been found. During the final search of the pond near Janice's ex-husband's camp, Al Bieber of NecroSearch International commented that the mud in and around a cattle pond near Lee's camp was bentonite, a clay that someone brought to the pond to stop the water from seeping out of the bottom. That evening, Booth and Martinez were camped near the crime scene. They were discussing the evidence in the

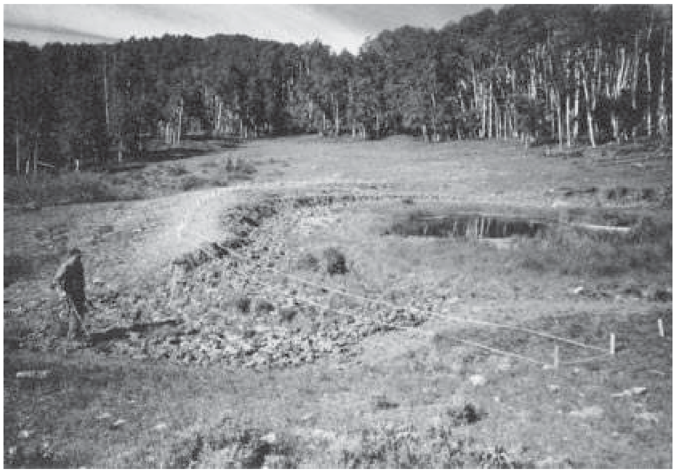

Figure 4.1 Pond lined with bentonite clay where Janice Dodson got clay on her shoes and pants when she went to steal the rifle that she used to kill her husband, John. *(See insert for color representation.)*

case when investigator Booth said, “The mud.” He was referring to the dried mud that was found on Janice Dodson’s clothing. If Janice had obtained the rifle from Lee’s camp, she would probably have stepped or fallen into the bentonite clay that drained across the road from the cattle pond (see Figure 4.1). Remembering Janice’s statement that she was returning to camp on the morning of the crime and stepped into a mud bog near her camp, Booth and Martinez decided they needed to obtain dried mud samples from the bog near the Dodsons’ camp, the area around a pond nearby the camp, and the human-made pond and runoff near Lee’s camp.

Booth and Martinez packaged the dried mud from each location and sent the samples along with the dried mud that had been recovered from Janice’s overalls to the laboratory section of the Colorado Bureau of Investigation in Denver, where it was examined by Jacqueline Battles, a forensic scientist and lab agent. She concluded and later testified to the fact that the dried mud found on Janice Dodson’s clothing was consistent with the dried mud recovered from the pond near Lee’s camp. The dried mud that had been recovered from Janice’s overalls was found not to be consistent with the mud bog or the pond near her camp. This was a breaking point in the case that allowed Booth and Martinez to put Janice Dodson in her ex-husband’s camp around the time that his rifle had been stolen. There are no other bentonite-lined ponds in the area and no bentonite deposits.

Booth and Martinez went to Texas and served an arrest warrant on Janice. She was extradited to Colorado, tried in court, and convicted in the murder of John Bruce Dodson. The jury understood the results that followed Booth’s insightful “the mud” exclamation. Janice is now serving a life sentence without the possibility of parole in Colorado’s state prison for women. An appreciation of the value of soil evidence and the collection of that evidence was critical in this case.

4.3 OIL SLICKS AND SANDS

A case that illustrates many of the issues comparing soil and related material occurred in Canada a few years ago. The body of 8-year-old Gupta Rajesh was found alongside a road outside Scarboro, Ontario. The back of his shirt had a smear of oily material, and the preliminary conclusion was that he was the victim of a hit-and-run accident, with the oily material coming from the undercarriage of a vehicle. But examination of the oily material and the particles suspended in it by forensic geologist William Graves of the Centre of Forensic Sciences in Toronto told a different story.

Investigators had collected samples of oily material on the floor of an indoor concrete parking garage where a suspect, Sarbjit Minhas, parked her Honda automobile. Analysis of the samples showed that the sand and other particles within the oil from the victim's clothes and the parking garage were similar. Analysis of the oil from the victim's shirt and garage floor showed them to be both similar and different from oil collected on the floor of 10 other garages in the area.

Particles in samples from the victim's clothes and the suspect's parking place provided considerable information. The sand from both samples was sieved and subsamples produced of the various size grades for the two samples. When compared after the oil had been removed, the color of each pair of subsamples was identical.

Additionally, the heavy minerals in both samples were similar, and three distinct types of glass were found in the two samples: amber glass, tempered glass, and lightbulb glass. Each of the different glasses was identical in refractive index value (the amount that a ray of light bends when passing through the glass into another medium). Small particles of yellow paint with attached glass beads were found in both samples. This type of paint is often found on center stripes of highways and reflects light.

Graves concluded that there was a high probability that the body of Gupta Rajesh had been in contact with the concrete floor of the garage at the place where the suspect parked her car. Interestingly, the same oil and particles were found in the suspect's Honda. Whether the oil and particles on the victim came from inside the vehicle or the floor of the garage, the presence and distinctiveness of the samples still strongly associated those two areas with the victim. Minhas was tried in the Superior Court of the Province of Ontario in November 1983 and convicted, with help from testimony by Graves.

This case illustrates an important concept in the presentation of soil evidence and perhaps all physical evidence, except DNA. We have become awed and impressed by the high probabilities that result from DNA evidence. Some people expect that other types of evidence should have similar statistical information. But in the Minhas case, we see a conclusion based on at least 10 different materials and observations. Because we do not know the probability of a tempered glass fragment, a particular group of heavy minerals, or sand of the same color being on a particular parking place in a concrete garage in Scarboro, Ontario—and in all likelihood we will never know—a frequency statistic cannot be generated. A useful database of sands, particles, glass, oils, and heavy minerals would be too difficult to generate. Additionally, it may not apply to any one specific case because of the variability of mineral particles—the very distinctiveness that makes geologic materials such good evidence. Thus, we rely on the skilled and honest examiner to reach a conclusion expressed in words rather than in numbers to inform the jury or judge so that they can reach a verdict. In this way the expert is a teacher, instructing the judge, attorneys, and jury in the basic concepts and premises that allow them to do the work they do. The triers of fact must be schooled in the methods of production of the evidence (how light bulb glass is made, for example), the procedures used to analyze it, and what makes the evidence significant. That understanding will lead the courts to an appreciation of unquantifiable evidence and give the jury a basis for weighing its significance (Houck, 1999).

4.4 MEDICAL LINK

A recent case does not fit the pattern of most soil evidence but clearly illustrates the contribution being made by forensic geologists. Washington State Patrol forensic geologist Bill Schneck became involved in the investigation into the serious illness of a small child caused by arsenic poisoning. The suspected person was absolved when an examination of the child's house revealed a number of mineral specimens left in the house and the yard by a former occupant who was a mineral collector. Many of those specimens were arsenopyrite, an iron arsenic sulfide. The child had been eating and chewing on the material. This case is a good reminder that lead is not the only material that can cause health problems in children.

4.5 EXAMINATION METHODS

The value of any type of physical evidence depends in large part on how many different types that can be characterized exist on this Earth and how they are distributed. The value of soil evidence rests on the fact that there are an almost unlimited number of rock types, mineral types, fossils, and artificial rock material, including glass and brick and concrete. In addition, it is not uncommon for soil material to have incorporated human-made particles such as plastics, metals, and chemicals such as fertilizer and pesticides.

Various instruments, methods, and procedures are used to study minerals, rocks, soils, and related materials for forensic purposes. These methods are used to collect data from the various questioned and known samples, which are then used to make a judgment as to comparison or lack of comparison. The evidential value of some of these methods is greater than others. However, they are all standard methods long accepted for use in the identification of material. Because of the extreme diversity of soils and related material, the examiner may decide that in a particular case some other method will provide information that will assist in reaching a conclusion. Because soils often have added particles, such as fibers, hair, and/or paint chips, these can be collected for examination by experts in those forensic science fields (Junger, 1996). One of the reasons that microscopic examination is so important is that it is only way that particles of other forms of evidence can be found and unusual mineral particles discovered. We must remember that examination is focused on determining whether two samples have the same properties and thus have the possibility of having a common source. That determination is greatly strengthened when rare and unusual minerals and particles are found. The experienced forensic geologist can express an opinion based on experience and education (Stam, 2002). The strength of that opinion is often based on knowledge of the rarity of particular minerals, rocks, particles, and related properties.

4.5.1 Color

Color is one of the most important identifying characteristics of minerals and soils (Dudley, 1975; Sugita and Marumo, 1996; Guedes et al., 2009). Minerals form a mosaic of grays, yellows, browns, reds, blacks, and even greens and brilliant purples. Virtually all possible colors of the visible light spectrum are represented. With most geologic materials and soils, the native minerals contribute directly to the soil color. This is particularly true with stream deposits, windblown silts, and other recent formations that have been in place a comparatively short period of time. If sands along a river channel are examined, the color of each sand grain can generally be recognized individually; however, after a deposit has weathered for a long period of time, there is a degree of leaching, accumulation, and/or movement of substances within the soil. Soil particles become stained, coated, and impregnated with mineral and organic substances, giving the soil an appearance different from its original one. The mineral grains, especially the larger ones, are generally coated. In most situations the coatings on the soil particles consist of iron, aluminum, organic matter, clay, and other substances. The coloring of the coatings alone can give some indication as to the history of the sample.

The “redness” of a soil depends not only on the amount of iron present but also on its state of oxidation, with a highly oxidized condition tending to have a more reddish color. The iron on the coatings of the particles probably is in the form of hematite, limonite, goethite, lepidocrocite, and other iron-rich mineral forms. Black mineral colors in the soil are generally related to manganese or various iron and manganese combinations. Green colors are generally due to concentrations of specific minerals rather than of the mineral coatings. For example, some copper minerals, chlorite, and glauconite are usually green. Deep blue to purple coloration in the soil is generally due to the mineral vivianite, an iron phosphate.

Apart from the mineral colors in the soil are those that result from organic matter. The organic litter on the soil surface is generally black. Humus percolates through the mineral horizons, giving various dark colors. In some instances the iron and humic acids combine to form a dark reddish brown to nearly black color. To have some uniformity in descriptions of the color of geologic materials and soils, certain standards have been established. The color standards most frequently used in the United States are those of the Munsell Color Company. The color standards are established on three factors: hue, value, and chroma. *Hue* is the dominant spectral color, *value* is the lightness color, and *chroma* is the relative purity of the spectral color. Soil and rock colors are generally recorded as, for example, 7.5YR5/2 (brown). The 7.5YR refers to the hue, 5 the value, and 2 the chroma. This standardization of colors offers some degree of uniformity, but moisture content will also affect the color of the soil, as will light intensity and wavelength. Soil color is different in natural light than in fluorescent and different still in incandescent light. If a soil is air dry, it may be recorded as yellow, but if moist the recording may be yellowish brown. Moisture added to a dry soil will usually result in a more brilliant appearance. It is therefore important to record not only the color of the soil but also an estimate of the "wetness factor" at the time of the recording.

In the 1970s, investigators at the UK Home Office Forensic Science Service studied the use of color as an examination method and made many contributions to establishing the study of color as an important first step. Soil, being a mixture of materials of various sizes and compositions, contains individual minerals of different colors. If soil is fractionated into various sizes—coarse sand, medium sand, fine sand, silt, and clay—there is a tendency for the finer-sized particles to exhibit more red or reddish-brown colors as opposed to grays and yellows in the coarser fractions. In considering coarser sand particles, the matrix will commonly have a speckled appearance, with the quartz and feldspar particles being gray or yellowish; but the heavy minerals such as ilmenite and magnetite will generally be black. Sand particles from soils of recent origin, such as recent glacial or stream deposits, usually retain their original mineral appearance and we can usually detect a mosaic of colors. But sand fractions from the old landscapes commonly have coatings of clay, and the sand grains may be iron stained, which results in a more uniform color of the entire matrix. Soil grains veneered with organic matter give the particles a dark gray appearance. It is important first to record the color of the untreated soil sample and then to remove the organics so that the true color and appearance of the sand grains can be studied.

In studying soil samples for forensic purposes, the sample is normally dried at approximately 100°C and viewed with natural light, preferably coming from a northerly direction. A north-facing window is a good location for such observations. Such studies should be made on samples that have the same general size distribution of particles. The color of samples prepared from the individual sieved-out particle-size ranges gives important additional data. Two or more samples, collected for study, can be compared directly by the observer (see Figure 4.2). It is then possible to use a color chart with the samples to determine the Munsell color numbers for precise description of the color.

Instrumental determination of color provides more quantitative information and can now be done with a precision which exceeds that the human eye. Pye Associates of Great Britain have found that rapid, reliable, and highly reproducible results are achieved using the Minolta CM-2002 photospectrometer. This instrument covers the visible wavelength range 400 to 700 nm, providing a number of light source types and observer angles, and producing a range of color indices, including Munsell hue,

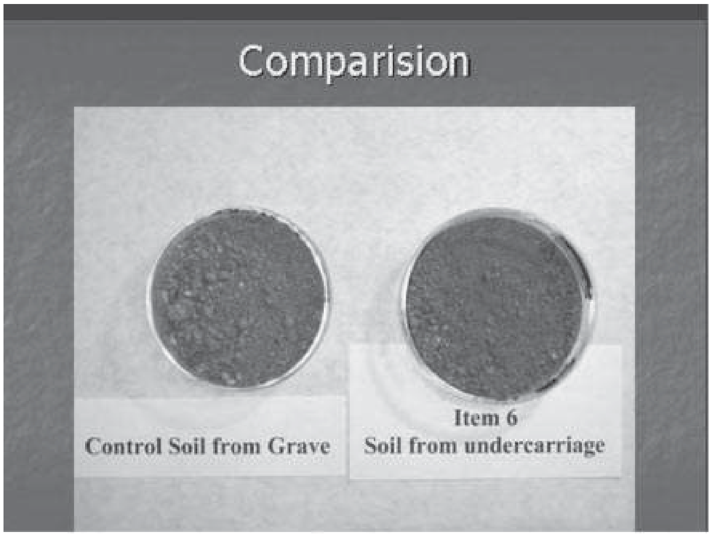

Figure 4.2 Two soil samples available for examination for color comparison. *(See insert for color representation.)*

value, and chroma; L*a*b* indices; and reflectance data and curves. Calibration of the instrument is carried out for every analysis to black, white, and CERAM II color standard materials.

4.5.2 Particle-Size Distribution

The determination of the distribution of particle sizes in a sample can often provide significant evidence. This determination is often produced for a variety of reasons: (1) to produce samples for comparison studies that are similar, in which case the control sample may contain some larger or smaller particles that are not present in the sample being questioned or an associated sample, and they must be removed; (2) the samples may be broken down into subsamples in which all the particles are in the same size range for mineral or color studies; or (3) a determination of the distribution of particle sizes may be produced as a method of comparison. A diagram showing the distribution of grain sizes can be used as a presentation and in many cases may be of evidential value (see Figure 4.3). For example, when abrasive particles have been introduced into machinery for the purpose of sabotage, the size distribution of the particles may be diagnostic of the material, assuming that changes in particle size have not taken place in the machinery.

The basic methods used for the separation of sizes are (1) passing the sample through a nest of wire sieves, with the size of the openings decreasing from top to bottom; (2) determining the rate of settling of the grains in a fluid, which is a measure of the size of the particles; and (3) instruments that measure the size of particles in a microscopic view and record the number of particles of each size. The distribution of particle sizes is then plotted on a diagram.

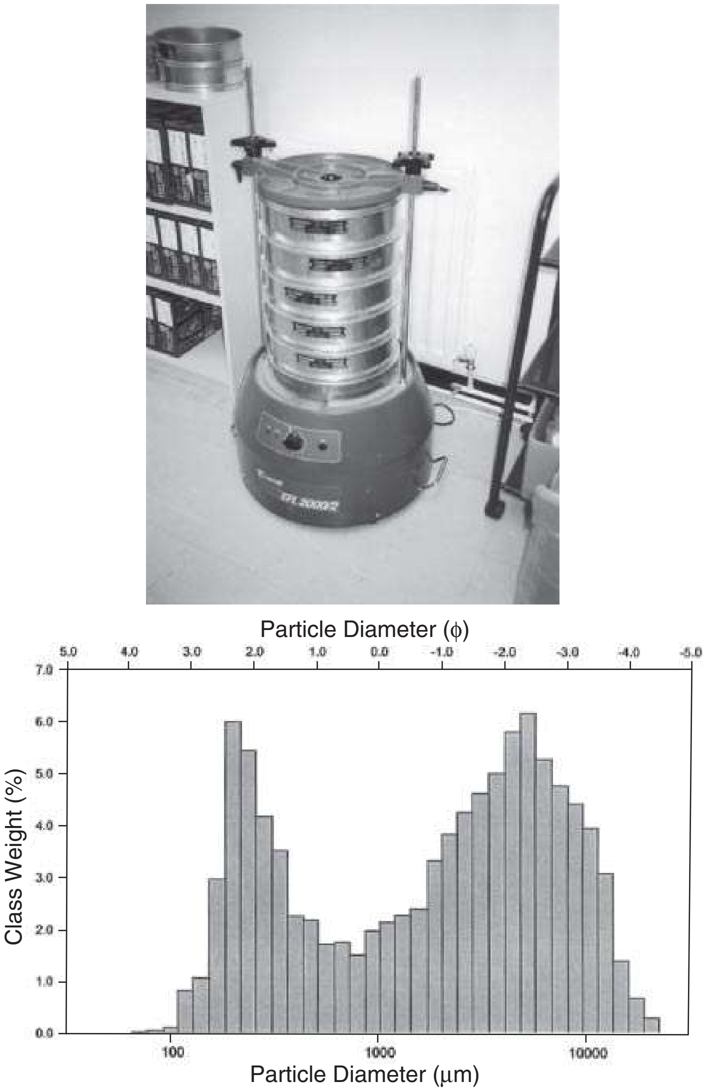

Figure 4.3 Nest of sieves used to separate particles into different size grades and a diagram showing the size distribution of particles. *(See insert for color representation.)*

Before making a mechanical analysis to determine the size distribution of particles, it is necessary to disperse the soil. Individual soil particles tend to stick together in the form of aggregates. Cementing agents of the aggregates must be removed; otherwise, a cluster of silt and clay particles would have the physical dimensions of sand or gravel. Cementing agents consist of organic matter, accumulated carbonates, and iron oxide coatings, and in some situations there is a mutual attraction of particles by physicochemical forces.

If carbonates have cemented the particles together, it is desirable to pretreat the sample with dilute hydrochloric acid to remove the carbonates. The sample is then treated with hydrogen peroxide to remove the organic cementing agents. All samples must be treated in the same way, and it must be determined before treatment that important information will not be lost, such as dissolving carbonate cement from grains that should be treated as single grains. It is almost always desirable to determine the size distribution of soil by sieving in a liquid, usually water. Dry sieving of the entire sample is generally unsatisfactory because the small particles tend to cluster together, and clay tends to adhere to larger particles. Sometimes a dispersing agent is added to the water.

A number of methods can then be used to determine the size distribution of the finer particles in a dispersed suspension. The hydrometer method is a rapid method for determining the percentage of sand, silt, and clay in a sample. It is based on the principle of a decreasing density of the suspension as the solid particles settle out. This method, although rapid and accurate, is unsatisfactory if we want to make a subsequent examination of the various size ranges, because there is actually no physical separation of the various-sized particles.

One of the most accurate and satisfactory procedures for fractionating soil samples is by the pipette method. This consists of pretreating the sample as is done in the hydrometer method, dispersing the soil in water, and calculating the time required for various-sized particles to settle out from the suspension. The principle is based on the fact that the rate of settling depends on the size of the mineral matter, with larger particles settling at a more rapid rate. The procedure is based on Stokes' law: $V = (2/9)gr^2(d - d')/n$, where V is the velocity of fall in centimeters per second g the acceleration due to gravity, r the radius of the particle in centimeters, d the density of the particle in grams per cubic centimeter, d' is the density of the fluid in grams per cubic centimeter, and n the viscosity of the fluid in poise. Although this method is generally considered the most satisfactory in regard to accuracy, it is not infallible. Several assumptions are made: that all particles have the same shape and that all the soil particles have the same density, neither of which is usually the case. Nevertheless, the pipette method is generally considered to be the best available mechanical method.

In making a mineral analysis of a sample, one fact becomes apparent: the sample contains different groups of minerals within various size ranges. Sands are made up of a set of minerals that are usually completely different from those within the clay size range. Therefore, in comparative analysis it is important to make comparisons within the same size ranges. It can be quite deceptive to compare the minerals found in one size range in one sample with a different size range in another sample.

When a transfer is made from a soil it is unlikely that the transferred sample is truly representative of the distribution of sizes that existed at that place. For this reason two samples that are to be compared using grain size distribution may not show exactly the same distribution of grain sizes. The method can be useful in coming to a conclusion of comparison or lack of comparison but is seldom definitive in itself.

Figure 4.4 Forensic scientist using a binocular microscope to identify particles in a soil sample.

4.5.3 Stereo Binocular Microscope

The stereo binocular microscope can be a most useful tool to the forensic geologist (see Figure 4.4). After an analysis of color has been made, the information that can be obtained by examination of soil samples with this instrument makes it the logical first step in the direct examination of particles.

Light microscopes are generally of two types, transmitted light and reflected light. In transmitted light microscopes, the light source is placed beneath the specimen, which must be transparent. Biological microscopes that are used in studying tissue are of this type. Reflected light microscopes have a light source above the object, and the surface features of the particle are viewed. Such a microscope is essentially a stationary, higher-power magnifying glass. Most of these microscopes have two sets of lenses, and thus the object is viewed in three dimensions, that is, in stereo. The magnification of a microscope is determined by multiplying the magnification of the ocular lens, commonly 10×, by the magnification of the objective lens, which differs from microscope to microscope but seldom exceeds 10×. This gives a maximum magnification of approximately 100×. The objectives may be individual lenses of fixed magnification, or in some microscopes a zoom objective is used. Such lenses can change magnification continuously from less than 1× to about 5×. Most viewing with stereo binocular microscopes is done at magnifications between 10× and 40×. Some of these microscopes are seated on a base that contains a second light source so that objects can be viewed in both transmitted and reflected light. When the transmitted light is polarized, the microscope may be used for both stereo reflected light viewing and low-power transmitted polarizing light studies.

Objects as small as approximately 10 μm in diameter can be viewed with a stereo binocular microscope. The upper limit is determined by how large a sample can fit under the instrument so that the surface of pebbles and cobbles can easily be viewed. The sample is normally placed on a tray having a dull black finish for ease of viewing light-colored minerals or a white finish for viewing dark-colored minerals. Various inserts are available for these microscopes that permit measurement of the size of objects or provide grids that aid in counting the various particles. The sample tray may have an etched grid that serves a similar purpose. Trays are available with various gummed surfaces that hold the grains in place for ease of counting.

In examining a soil sample or similar material, the scientist will commonly first examine the entire sample as it is received and observe the types of grains and particles. Recording a general impression of the material is normally done at this time. It is not uncommon to observe in soil samples nonmineral materials such as fibers, metals, paint, glass, and plastics. These objects can have important evidential value. In some situations they can be the most important parts of the sample. Materials such as metals, hair, fibers, paint, and plastic are removed for further examination by specialists. Plant particles can be of great value. The total amount of plant material is in general relatively useless for forensic purposes. However, identification of the individual grasses, seeds, leaves, and the like can be most useful.

Preliminary examination of the entire sample with the binocular microscope is normally very difficult. The mixture of particles of all sizes commonly obscures the grains and makes identification difficult. The presence of organic material contributes to this problem. The sample must be cleaned for the study of minerals and rock. Sieving of the samples removes the larger particles and most of the larger organic fragments. If the sample is washed carefully in water, the lighter organic particles will generally float and can be removed and saved for study. Treatment with hydrogen peroxide will remove the fine organic matter and clean the sample. The use of ultrasonic cleaners is dangerous in many cases. Many times the scientist has taken a sample that contained chips of red shale before ultrasonic cleaning and found that the sample contained thousands of silt-sized quartz grains and clay minerals after cleaning and that the shale chips had been broken up into its components. If the rocks and minerals are of a type that would not be disaggregated by ultrasonic cleaning, the method can be most useful.

With a clean sample, the experienced scientist can identify the rocks and minerals at sight or by using simple tests. In addition, it is possible to observe the texture and coatings on the surface of the grains. Properties of grains such as shape, rounding, weathering, inclusions, color, and polish can be observed and recorded. The counting of different types of grains is especially important. When we record our information in numbers, it is normally more useful than qualitative impressions. However, the samples may be so different on first examination that further work is not useful because a determination of comparison could never be made. The number or percentage of different types of grains is an extremely important tool in determining comparison or lack of comparison. In counting grains of different types, it is important that the sample counted be representative of the whole sample, that the identification be consistent and accurate, and that the same grain not be counted twice because the sample moved. The latter problem can usually be avoided by placing the sample on a gummed surface or by removing the grains as they are counted and placing each grain, as it is removed, in a container or gummed individual tray. Most important is the judgment and caution used by the scientist, whatever method is used.

4.5.4 Petrographic Microscope

A petrographic microscope differs in detail from an ordinary compound microscope (see Figure 4.5). However, its primary function is the same: to produce an enlarged image of an object placed on the stage. The magnification is produced by a combination of two sets of lenses, the objective and the ocular. The function of the objective lens, at the lower end of the microscope tube, is to produce an image that is sharp and clear. The ocular lens merely enlarges this image. For mineralogical work, three objectives—low, medium, and high power—are normally used. The magnification produced by objectives is usually 2× (low), 10× (medium), and 50× (high). Oculars have different magnifications, usually 5×, 7×, 10×, 15×, and 20×. The total magnification of the image is determined by multiplying the magnification of the objective by that of the ocular as follows: 50× times 10× = 500× (see Figure 4.2).

Oculars normally contain a crosshair that is useful for locating grains under high power when changing objectives. A condensing lens system is normally provided under the stage for use with high magnification and for assisting the viewing of the various optical effects produced by minerals. Petrographic microscopes have a rotating stage and a polarizing filter under the stage that transmits light usually

Figure 4.5 Forensic scientist using a petrographic microscope to identify particles.

vibrating in a N-S (front-to-back) direction. Above the stage a second, removable polarizing filter is placed in the tube of the microscope. It transmits light usually in an E-W (left-to-right) direction. When the upper filter is inserted, light is blocked out from passing through the microscope. In this case the filters, which are called *polars*, are said to be crossed. Only when an *anisotropic* material, that is, a material that is not isotropic (meaning that it forms in the isometric crystal system or is amorphous), is placed on the microscope stage under crossed polars can it be seen. The effect of the anisotropic mineral is to rotate the N-S vibrating light from the lower polarizing filter, thus permitting some of it to pass the upper E-W polarizing filter. When the stage is rotated, there will be four positions when the vibration directions in an anisotropic crystal will line up with the N-S and E-W direction. At these positions the crystal is said to be at extinction and will appear black and thus not be seen.

In identifying mineral grains under the petrographic microscope, it is common to use the immersion method. Mineral grains are placed on a microscope slide in a liquid of known refractive index. These liquids are available commercially. The range 1.46 to 1.62, with a difference of 0.02 between adjacent liquids, serves most purposes. When the grain is viewed, a narrow line of light is commonly seen surrounding the grain. If the distance between the objective and the sample is increased slightly, usually by raising the tube of the microscope, the line of light, called a *Becke line*, will move in the direction of the higher refractive index. If the mineral has a higher refractive index, the Becke line will move into the grain. If the liquid of known refractive index is higher, the Becke line will move away from the grain into the liquid.

By trial and error with different liquids, a match is found. At this point the grain will be almost invisible in the liquid. In most cases the refractive index of the grain is found to fall between the refractive indexes of two liquids, and the value can be estimated by an experienced observer.

The petrographic microscope is an important tool in many aspects of forensic work and is the best method for a study of the optical properties of rocks and minerals. A study of individual mineral grains or thin sections of rocks and related material is easily accomplished by anyone trained in the use of the instrument. A thin section is a thin slice of rock mounted on a glass slide. The slice is normally 30 μm in thickness and may be prepared from a solid rock or loose material impregnated with plastic. The rock is cut with a diamond saw and the surface polished. This polished surface is cemented to a glass microscope slide with an adhesive of known refractive index such as epoxy or Canada balsam. A saw cut is then made parallel to the glass, leaving a wafer of rock cemented on the slide. Grinding of the wafer proceeds until it is thinned to approximately 30 μm. A thin class cover is then glued onto the polished rock surface to protect the rock and improve viewing with the microscope. Most rocks are transparent at this thickness and can be viewed in transmitted light (see Figure 4.6). Loose mineral grains of the same general size, also commonly mounted in epoxy or Canada balsam on a microscope slide, are covered with a thin platelet of glass (cover glass) and studied. This is the method used when the heavy minerals, that is, those minerals with high

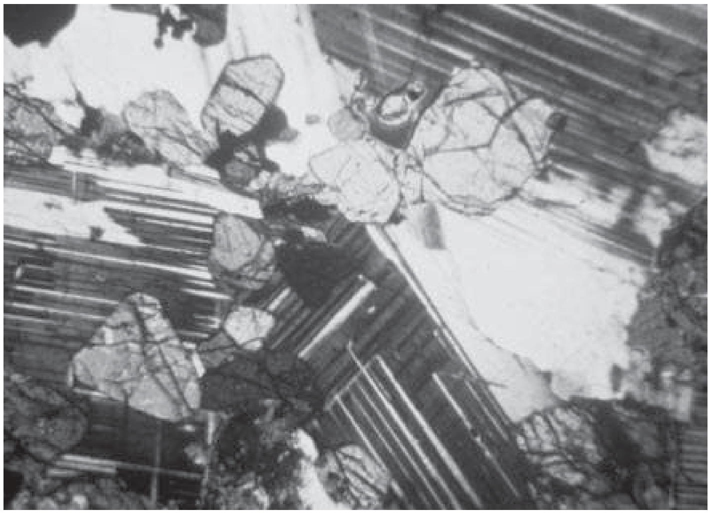

Figure 4.6 Photomicrograph of a rock thin section as viewed through a petrographic microscope, Note the different minerals with different sizes and shapes. *(See insert for color representation.)*

specific gravity (such as rutile, garnet, zircon, and tourmaline), are separated from the more common lighter minerals (such as quartz and feldspar) by settling in a heavy liquid such as sodium polytungstate or one of the other tungsten-based heavy liquids and studied.

4.5.5 Refractive Index

The index of refraction of a transparent material is the ratio of the velocity of light in a vacuum, normally considered to be 1, to the velocity of light in the material being analyzed. Thus, a refractive index of 2.4553 means that light travels 2.4553 times as fast in a vacuum than in the transparent material. The measurement of refractive index, which is one of the most important methods for the comparison of glass, may be made using the Becke line method discussed for minerals. Most forensic laboratories today use the semiautomated refractive index instrument GRIM 11 for glass identification (see Figure 4.7).

4.5.6 Cathodoluminescence

The instrument used for cathodoluminescence is a luminoscope that is attached as a stage on a microscope or a scanning electron microscope. The specimen—for example, mineral grains or a thin section—is bombarded with a beam of electrons generated by the instrument. When the electrons strike the surface of the specimen, an optical luminescence is produced, which is seen as a display of colors. The colors and their intensity depend in large part on very small changes in the concentration of

Figure 4.7 Forensic geologist using GRIM 11 instrumentation to determine the refractive index of a glass.

trace impurities, the minerals present, and where the trace impurities are located in the structure of the minerals. Thus, the method has wide application in determining or observing a variety of differences in mineral grains that otherwise appear similar.

4.5.7 Scanning Electron Microscope

The scanning electron microscope (SEM) has a wide range of magnifications, generally from 25× to over 650,000× and can record something as small as 1.5 nm. Needless to say, most of the work of the forensic geologist falls well within these limits. This instrument became commercially available in the mid-1960s and has been introduced rapidly in forensic work, especially in the study of gunshot residues and various other very small particles. It has the advantage that the surface of a sample may be viewed directly. However, an ultrathin coating of carbon or gold plated on the specimen improves the quality of the picture. The depth of field is very large, and most SEM pictures have an excellent three-dimensional appearance. In using the instrument it is possible to change magnification easily and thus study the appearance of the surface from very low to very high magnification. Differences in very small fossils that were not previously known can now be seen during routine examination. Surface features of individual grains of minerals such as quartz can be seen and shown to have many different types of scratches, pitting, and mineral growth. Some of these features may be useful in telling us the past history of the individual grain. It is not uncommon to observe other minerals, such as clay flakes, filling the scratches and thus adding another characteristic that can be useful in comparing the minerals.

When using these powerful instruments in forensic work, it is well to keep in mind that no two objects are ever exactly the same. No two sand grains are ever exactly alike when studied under the high magnifications of an electron microscope. This is true even if they have been side by side for the past million years. Observations made with these instruments can be very useful for establishing similarity or dissimilarity between samples. However, the very power of the instruments permits the possibility of their abuse in the hands of the unscrupulous. If we were to do a complete chemical analysis of a total person by the most modern methods in the morning and repeat the analysis on the same person in the afternoon, we would find chemical differences. However, this would not demonstrate that we had analyzed two different people. Similarly, the demonstration of small differences in soil does not prove in itself that they do not compare. It is equally true that showing a common similarity among soil samples, such as their both containing quartz, the most common mineral in soil and sediment, is poor evidence on which to base a comparison. The professional judgment of the scientist thus becomes increasingly important when these powerful instruments are used.

Scanning electron microscopes have the ability to determine the elemental composition of the particles being examined. This is possible because x-rays are produced when the electron beam of the microscope strikes a target. The scanning electron microscope can be coupled to an x-ray analyzer. The emitted x-rays are sorted by their energy or wavelength values, which are related to specific elements, and the analyzer produces information that identifies the elements present in the material being viewed. The amount of each element present is determined by the intensity of the emitted x-rays. Thus, the examiner can determine the chemical composition of the individual particle or particles being viewed.

4.5.8 X-Ray Diffraction

X-ray diffraction is one of the most important and reliable methods of identifying the composition of geologic, soil, and other crystalline substances (see Figure 4.8). The method is based on the arrangement of atoms, ions, and molecules within the specimen. The sample is analyzed by passing x-rays through a crystal and measuring the angle of the diffracted x-rays. Each crystalline material has its own distinctive x-ray pattern. The x-ray diffraction pattern of a sample is controlled by the internal structure of the specimen. The diffraction pattern can be collected on film, on an image plate, or by using an electronic detector. The interpretation of x-ray patterns under normal situations is a comparatively simple matter. Two factors are of prime importance in the interpretation: the d-spacing from d in Bragg's law, which is expressed in angstroms (Å; 1×10^{-8} centimeter or 0.0039 millionths of an inch), and the intensity.

There are at least two avenues for interpreting x-ray diffraction data: The first involves measuring the d values and intensities and comparing this information with published lists of data on minerals. The second involves comparing the x-ray pattern directly with the pattern produced by a known mineral. If a comparison between samples is to be made, there are situations in which the x-ray diffractograms themselves may be compared without actual identification of the substance. This is obviously less useful as evidence than actual identification.

Figure 4.8 Forensic geologist using an x-ray diffractometer to identify minerals.

Among the strong points favoring the use of x-ray diffraction patterns is that they record the crystal structure. If two substances, diamond and graphite, were analyzed chemically, they would be identical because both are composed of pure carbon, but x-ray diffractograms of the two minerals would be quite different. Many samples are mixtures of two or more substances. If substances are analyzed chemically, some difficulties may be encountered because the actual chemical form of the substances at times cannot be established. As an example, we can use a mixture of two salts, sodium chloride and potassium nitrate. If the composition of the sample were determined by the usual chemical methods, it would reveal sodium, potassium, chloride, and nitrate; but what were the original compounds? Were they sodium chloride and potassium nitrate, sodium nitrate and potassium chloride, or a mixture of four salts? An x-ray diffractogram of such a salt mixture would tell us specifically the form of the salt.

X-ray diffraction is used as the principal tool in the modern identification of clay minerals. The chemical composition of clays generally tells us very little as to the nature of clay substances, but the possibilities of identification of clays by x-ray diffraction are almost unlimited. Clays, as well as other crystalline substances, can be x-rayed, and if identification is desired, the composition of the sample can be ascertained by measuring the diffraction patterns from a standard reference book or from reference cards.

4.6 CHEMICAL METHODS

There are a large number of instruments and methods that measure the chemical composition of organic and inorganic materials. There are times where these methods can be very useful and provide valuable information. For example, the chemical composition of glass is often a valuable addition to the optical and physical properties. Identification of organic compounds in a soil such a fertilizer adds an entirely new dimension to the list of properties used in determining comparison.

Methods for determining the amounts and types of elements in a sample may rely on the fact that as one of its properties, an element selectively emits or adsorbs light. This is the basis for emission spectroscopy and atomic adsorption spectrophotometry. Neutron activation analyses is a nondestructive method that produces a detection sensitivity of one-billionth of a gram. This method requires a nuclear reactor to bombard the sample with neutrons. The resulting gamma-ray radioactivity is measured to identify the elements present and their amounts. Needless to say, this method is expensive to operate and maintain and leaves the samples radioactive. Organic compounds contain carbon, and their identification requires different methods. There are several techniques for separating out the various compounds in a mixture. Generally, these methods depend on the relative amounts of the gas phase of a compound and the liquid phase under fixed conditions. That amount is a characteristic property of each compound. Because those compounds that have a higher tendency to go to the gas phase will move faster away in a given time, the distance they move in that time identifies and separates the compounds. These methods are generally called chromatography. Identification is also done using spectrophotometers, which measure the light-adsorbing properties of a compound. Mass spectrometers have the ability to uniquely identify compounds if analyzed under proper laboratory procedures. This instrument bombards the sample with high-energy electrons, causing the molecules to lose electrons and become positively charged. This is an unstable state and the molecules immediately break up into fragments. The instrument then passes the fragments through an electric or magnetic field where they are separated according to their masses. This permits the specific identification of the compounds because the distribution of masses is a unique property.

4.6.1 FTIR and Raman Spectroscopy

Fourier transform infrared spectroscocopy (FTIR) and Raman spectroscopy are nondestructive analytical tools that are commonly applied to the identification of minerals and material of organic origin. They are often combined with a microscope, thus facilitating the identification of a very small object or part of a sample. The nondestructive aspect is important because most owners do not appreciate removing part of their art object or gems for analyses. In addition, in criminal cases the sample may be very small and removal of material for analysis might leave little remaining for additional study or verification. In applying these methods, a source generates light across the spectrum of interest. The sample absorbs light according to its chemical properties. A detector collects the radiation that passes through the sample. Computer software analyzes the data collected and the results are compared with known spectra of inorganic and inorganic materials.

4.7 LOOKING AHEAD

The last several years have witnessed a tremendous increase in both the quantity and quality of forensic examination of soil and related material. There are now three comprehensive books on the subject, and the reader is encouraged to use them as source material (Murray, 2011; Pye, 2007; Ruffell and McKinley, 2008). The future will depend on how well we address a series of issues that include the following:

1. New methods are being developed that take advantage of the discriminating power inherent in earth materials. The opportunities appear to lie primarily in quantitative mineral identification. For this reason, the development of quantitative x-ray diffraction techniques that will provide quantitative data on the mineral composition of a sample would seem to be an important direction for research.
2. Considerable effort must be devoted to defining appropriate sampling methods and the training of those who collect samples. Communicating to law enforcement personnel the potential evidential value of soil and related material is particularly important, because unless the evidence is collected there will never be an opportunity for forensic examination.
3. There is a tremendous need for studies that attempt to demonstrate the diversity of soils. Such studies provide the opportunity for establishing worthwhile information on the distribution of the various soil types. The information should be data based and widely distributed.
4. There must be a continuing effort at all levels to improve the qualifications of examiners in forensic geology.
5. A problem has developed that result from the availability of new instruments that provide increasingly detailed measurements or observations, such as the scanning electron microscope. These instruments are capable of discriminating between individual grains. When you discriminate between individual grains, you lose the ability to say that the sample being examined was once part of, and thus had a common source with, another sample. When this is done, the entire concept of comparison is eliminated and the evidential value is lost. The forensic geologist must choose methods that provide the maximum discriminating power between samples without falsely excluding samples that are in fact associated.

The real future of the science lies with education: the education of the investigator and evidence collector that earth materials can make a major contribution to justice if they are both collected and studied properly. The real challenge of yesterday, today, and tomorrow is the production of the best evidence by the most skillful and objective people to serve the cause of justice. Our system of justice is still run by people, people who are human and in many cases people who are trained advocates. Those who the courts honor by allowing them to express an opinion, the expert scientific witnesses, must rise to the highest standard. If they don’t, the advocates may find ways to remove the privilege and return to a legal world populated only by human witnesses reciting their stories from memory.

REFERENCES

Dudley, R. J. (1975). The use of color in the discrimination between soils. *J. Forensic Sci. Soc*., **15**:209–218.

Gross, H. (1893). *Handbuch für Untersuchungsrichter*. Munich, Germany: N.P.

Guedes, A., et al. (2009). Quantative colour analysis of beach and dune sediments for forensic purposes: a Portuguese example. *Forensic Sci. Int*., **190**:42–51.

Houck, M. M. (1999). Statistics and trace evidence: the tyranny of numbers, Oct., Vol. 1, No. 3, http://www.fbi.gov/hq/lab/fsc/current/index.htm.

Junger, E. P. (1996). Assessing the unique characteristics of close-proximity soil samples: Just how useful is soil evidence? *J. Forensic Sci*., **41**:27–34.

Murray, R. C. (2005). Collecting crime evidence from soil. *Geotimes*, Jan., pp. 18–22.

Murray, R. C. (2011). *Evidence from the Earth* 2nd edition, Missoula, MT: Mountain Press.

Pye, K. (2007). *Geological and Soil Evidence*. Boca Raton, FL: CRC Press.

Pye, K., and Croft, D. J. (Eds.) (2004). *Forensic Geoscience*. Special Publication 232. London: Geological Society of London.

Ruffell, A., and McKinley, J. (2008). *Geoforensics*. Hoboken, NJ: Wiley-Blackwell.

Stam, M. (2002). The dirt on you. *Calif. Assoc. Criminalist's Newsl*., 2nd quarter, pp. 8–11.

Sugita, R., and Marumo, Y. (1996). Validity of color examination for forensic soil identification. *Forensic Sci. Int*., **83**:201–210.

Thorwald, J. (1967). *Crime and Science: The New Frontier in Criminology*. New York: Harcourt Brace Jovanovich.

CHAPTER 5

Analysis of Paint Evidence

SCOTT G. RYLAND
Florida Department of Law Enforcement, Orlando, Florida
EDWARD M. SUZUKI
Washington State Crime Laboratory, Washington State Patrol, Seattle, Washington

Summary Paint and coatings often appear in criminal, civil, and art authenticity investigations. In this chapter we review the current methodologies and approaches used by forensic paint examiners to analyze this type of physical evidence as well as the problems that they may encounter. Fragments of multilayered in-service paint are one of the most complex types of materials encountered in the forensic science laboratory. They consist of both organic and inorganic components heterogeneously distributed in very small samples, often on the order of only 1 mm^2 in size. Each layer is an entity onto itself and its autonomy must be maintained throughout the analysis. These characteristics dictate the requirements of the analytical chemistry approaches to be used, and they can present a formidable challenge to the forensic analyst responsible for the classification of the materials and an evaluation of their evidential significance. Several case examples are presented to illustrate these concepts.

5.1 Introduction 132
5.2 Paint chemistry and color science 134
  5.2.1 Binders 134
  5.2.2 Pigments 136
5.3 Types of paint 139
  5.3.1 Automotive finish systems 139
  5.3.2 Architectural coatings (structural paints or house paints) 140
  5.3.3 Other coatings 141
5.4 Paint evidence interpretation considerations 141
5.5 Analytical methods 142
  5.5.1 Microscopic examinations 143
  5.5.2 Physical nature of the transfer 147

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

5.5.3 Microscopy 149
5.5.4 Microspectrophotometry 152
5.5.5 Infrared spectroscopy 158
5.5.6 Raman spectroscopy 175
5.5.7 Pyrolysis gas chromatography and pyrolysis gas chromatography–mass spectrometry 178
5.5.8 Elemental analysis methods 188
5.5.9 Other methods 205
5.6 Examples 208
5.6.1 Example 1 208
5.6.2 Example 2 210
5.6.3 Example 3 213
References 217

5.1 INTRODUCTION

The body of Wendy Coffield, 16, was discovered in the Green River, south of Seattle, on July 15, 1982. In the months to follow, the bodies of several more young women were found in or near the Green River. By the early 1990s, the total number of victims believed to be associated with this unsolved case had reached 60 or more, and investigators were faced with the very real possibility that the most prolific serial killer in U.S. history would never be apprehended. In 2001, based on recently developed DNA technology, semen samples collected from three of the victims were found to be those of Gary L. Ridgway, a paint detailer at a local truck manufacturing plant. Ridgway pleaded not guilty when charged with homicide, claiming that he only had sex with the victims. Although usable DNA evidence was not recovered from any of the other bodies, hundreds of very small colored spherical particles were found on the clothing of four of them as well as on the clothing of two of the victims associated with Ridgway's DNA (Palenik, 2007). Similar particles were found on Ridgway's work clothing and in his home and work environments. These particles were identified as dried droplets produced from the plume of spray paint, and the paint was found to have an acrylic urethane composition. In the early 1980s, this formulation was not used at all on North American original automotive finishes and it would also have been quite unusual for architectural coatings, with the possible exception of certain varnishes, but these would not have been pigmented as were the samples found on the bodies. The paint was identified as Imron, a DuPont high-end specialty product not sold to the general public; more specifically, it was found to be Imron that had been manufactured before 1984. This product had been used at the truck plant where Ridgway was employed, and the paint evidence was the basis for charging him with four additional counts of homicide. In 2003, Gary L. Ridgway pleaded guilty to 48 counts of homicide, prompted by his desire to avoid the death penalty and his fear of a conviction in the face of more physical evidence linking him to the victims. Ridgway's attorney was quoted in the press as stating that the paint evidence was crucial in his client's decision to change his plea.

This example illustrates the value of paint as physical evidence and as an investigative tool, and although the majority of cases involving paint may not be in the national limelight, this important category of trace evidence provides crucial links in criminal investigations and prosecutions on a daily basis. Typically, forensic paint examiners are asked to perform a comparison of two or more samples to determine if they could share a common origin. The samples may result from transfer of paint from one item to another, as may occur when a vehicle strikes another vehicle, a building or road structure (such as a guardrail), or an individual. If the object that is struck also has a painted surface, a double transfer may result, with each item now having paint from the other. Other circumstances may also facilitate a transfer, as when a crowbar is used to pry open a window, safe, or some other item with a painted surface; a bullet or other projectile strikes a painted surface; a person or object comes into contact with wet paint; or when paint particles are transferred from the clothing of one person to another due to contact. The latter two transfer scenarios are believed to have occurred in the Green River cases.

In other instances, an unknown paint chip is recovered from the scene of a hit-and-run accident or from clothing of a victim struck by a vehicle, and the forensic paint examiner is asked to try to identify the type of vehicle involved. On a few occasions, the paint examiner may be asked to determine if a material *is* paint, and if so, the type of the paint and its possible relevance to the circumstances of the investigation.

In the Green River cases, elements of all three types of examinations (comparison, vehicle identification, and determination if paint) were involved. After the unknown colored particles were removed from the victims' clothing, the trace analyst first had to determine what they were. Finding them to be dried droplets of spray paint, the analyst next had to determine their chemical composition and whether the specimens were similar to those associated with the suspect. Based on chemical composition, an original automotive finish—which usually serves as the basis for identifying the type of vehicle involved in a hit-and-run incident—appeared very unlikely. The nature of the paint, dried droplets of spray paint, also suggested that an original automotive finish was probably not involved, as the nearest automotive assembly plant was thousands of miles away. Nonetheless, an identification of the paint was sought to assess the significance of the match that was found between the specimens from the suspect and the victims. In this case, this resulted in the determination not only of the specific brand of paint that was involved, but also of a time frame for when it was manufactured.

One other type of examination that the forensic paint examiner may be asked to perform concerns whether a painting or some other artifact is a forgery. Although authenticity in such cases cannot be established by paint composition alone, in certain instances, forgeries can. If a painting or other artifact that was supposed to have been made on a certain date or during a certain time period contains pigments or other paint components that were first available only after this date or period, the painting or artifact could not have been produced when purported (assuming that no modern restorations were performed).

How the forensic paint examiner goes about accomplishing these varied tasks is described in this chapter. We begin with an overview of paint chemistry and color science, briefly describe the types of coatings encountered, and discuss some issues and concerns specific to paint evidence of which analysts must be cognizant. The various analytical methods that are used to examine paint evidence and the types of information that these methods provide are then described. Since correct interpretation of analytical data is the cornerstone of formulating appropriate conclusions, we discuss analytical methods in some depth. Because data collected for two or more samples are frequently compared, differences that may occur and *their significance* are central issues in forensic science. Aspects of a particular analysis that can significantly affect data—such as sample preparation, instrument parameters used, nature of the samples, and possible alterations of samples caused by the analysis process—are therefore emphasized. We conclude by presenting three more case examples to illustrate the concepts discussed and to demonstrate further the wide variety of analyses that an examination of paint evidence may entail.

5.2 PAINT CHEMISTRY AND COLOR SCIENCE

Paint is a film or coating that serves both decorative and protective functions. In the coatings industry, *paint* usually refers to a pigmented coating, but in this chapter it is synonymous with the term *coating*. Dried paint consists of three components: a binder, pigments, and additives. The *binder* comprises the matrix that holds the paint together, within which are suspended very small particles of pigments and additives. The binder consists of an organic polymer that not only forms a strong film but also adheres to the substrate to which it is applied. A *pigment* is a finely powdered compound or mixture that imparts color, opacity, or other essential properties (such as luster or texture) to the paint. An *additive* is a substance added to a paint (usually in a small quantity) to improve its properties; it includes components such as corrosion inhibitors, catalysts, ultraviolet absorbers, and *plasticizers*. Plasticizers are moderate-weight organic compounds (often esters) or polymers that serve to fill the spaces between adjacent polymer chains, making the product more flexible and less brittle.

5.2.1 Binders

The polymers that are used for paint binders include acrylics, alkyds, polyesters, urethanes, epoxies, vinyls, cellulosics, and silicones (Morgans, 1990; Ryland, 1995; Thornton, 2002). Binder polymers may consist of single repeating units (homopolymers); two or more units, usually in a random order (copolymers); different homopolymer or copolymer units joined together (block copolymers); or homopolymers, copolymers, or block copolymers associated in the same binder matrix but without covalent bonds between them. Depending on the binder, the polymer chains may be linear, branched, cross-linked, or a combination of these, and many of the polymers cited may constitute the main backbone chain, a side chain, or a cross-linking chain (Oil and Colour Chemists' Association, 1983). Binders are typically manufactured by a chemical company, such as Rohm & Haas or Union Carbide, and then sold to a paint manufacturer, such as Sherwin-Williams or Inmont, for incorporation into the paint batch. There are exceptions to this trend with large companies such as DuPont, which has facilities to manufacture the binders as well as divisions devoted solely to paint manufacture.

Several different processes are used to create binders from polymers or monomers. One straightforward method involves a preformed polymer, usually linear, which is dissolved in an appropriate solvent. A film of the paint is formed by simple solvent evaporation and this type of paint is referred to as a *lacquer*. The adjacent polymer chains of the binder are held together by relatively weak forces (van der Waals and polar forces), so such paints can be redissolved in appropriate solvents.

One lacquer that many readers may be familiar with is nitrocellulose, since it is the main binder of most fingernail polishes. Nitrocellulose is a linear polymer that is produced by forming nitric acid esters with roughly two-thirds of the free hydroxyl groups of cotton (cellulose). It is soluble in acetone, ethyl acetate, and other organic solvents but forms a rather brittle film by itself, so a plasticizer is also used in the formulation. An additional polymer, such as tosylamide–formaldehyde, may be included to improve adhesion to the fingernail surface (Drahl, 2008). As a lacquer, fingernail polish is readily removed using acetone or other solvents. A nitrocellulose lacquer was used for the first production-line automobiles, manufactured in the 1920s (Fettis, 1995).

A second method of forming binders also involves a mostly preformed polymer, but in this case, small *micelles* of the polymer are suspended in the paint solvent (similar to fat globules dispersed in milk) rather than forming a true solution. These colloidal systems are known as *emulsion* or *dispersion finishes*, although when water is used as the medium and architectural coatings are involved, they are more commonly referred to as *latex paints*. Upon evaporation of the solvent, the micelles coalesce to form a film and further polymerization occurs. Depending on the specific polymer used, the film formed from this process may be an *enamel* (a finish that is not soluble in most solvents) or a lacquer. Solubility usually reflects polymer structure, with lacquers normally comprised of linear chains whereas enamels exhibit considerable cross-linking. Other methods of forming binder polymers are also used.

One of the most common automotive paint binders is the acrylic melamine enamel, which has served as a staple for North American automobile finish layers during the past 40 years or so (Rodgers et al., 1976a; Ryland, 1995; Suzuki, 1996a; Ryland et al., 2001). The polymer chain of this binder is comprised of a sequence of acrylic copolymers, and methylated, butylated, or isobutylated melamine–formaldehyde is used to cross-link the acrylic backbone (Nylen and

Sunderland, 1965; McBane, 1987; Fettis, 1995). This composition is shown schematically in Figure 5.1; eight different acrylic copolymers are depicted to illustrate the variety that may be used, although in practice, no more than five are normally found in a single formulation (Nylen and Sunderland, 1965). The various copolymers are chosen to impart specific properties, such as hardness, flexibility, durability, adhesion, gloss, and water and chemical resistance. Although not an acrylic, styrene is also frequently incorporated into the backbone. In addition to specific copolymer composition, the properties of this binder are determined by the extent of cross-linking. Original automotive paints are baked to cure the finish, that is, to promote the cross-linking process, and also to increase gloss.

5.2.2 Pigments

Pigments are distinguished from dyes in that they are insoluble in most solvents, and in particular they are insoluble in the solvent used for the paint itself. Dyes, in contrast, are soluble, and they are not used in most paints, as they are less durable than pigments and also because they are usually too transparent; it would be difficult and probably cost prohibitive to try to incorporate enough dye into a paint to make it opaque. Dyes also suffer from one other serious drawback regarding tint control. The optical properties of a color-imparting agent (such as hue and *tinctorial strength*) depend on particle size, particle-size distributions, and crystal structure (Braun, 1993). Many common paint pigments exhibit *polymorphism*; that is, they can occur in more than one crystal structure, and in some cases, two different crystal structures of the same pigment will exhibit quite different hues. Particle size, particle-size distributions, and specific crystal structures cannot be controlled when a dye recrystallizes from solution; hence, a color-imparting agent that is insoluble in the paint must be used.

Even with the use of pigments, paint formulators must be aware of the effects of polymorphism. Copper Phthalocyanine Blue (Figure 5.2) is a very common organic pigment used in automotive and other paints and it can occur in any of four different crystal forms (Lewis, 1995). The polymorph of this pigment that is used in most paints is treated to prevent its slow conversion to another crystal form that has a green rather than a blue hue, as most consumers do not appreciate their blue vehicle turning green with age.

The most common pigment used in paint is titanium dioxide. This compound occurs in three polymorphic forms, rutile, anatase, and brookite, but only the first two are used in paints. Rutile is, by far, the more common of the two, both because of its slightly greater index of refraction and, especially, its greater durability to light exposure and heat. Rutile serves to make a paint opaque, that is, to make it white (or a lighter color—adding a large amount of rutile to a red paint, for example, makes it pink). A binder that is cured to form a dried film is normally transparent, but the presence of numerous small particles of rutile causes light incident on this medium to scatter in all directions. Rutile is effective at scattering light because its index of refraction ($\sim$2.7) is quite high compared to that of paint binders ($\sim$1.5).

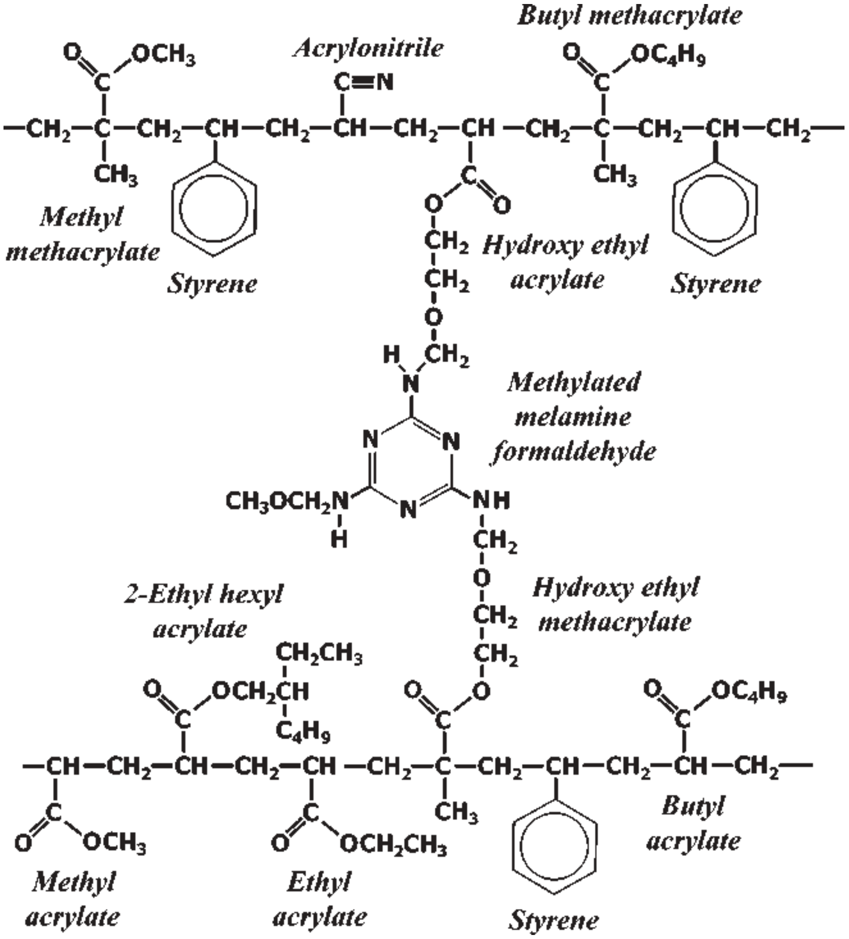
Butyl methacrylate
Acrylonitrile
Methyl methacrylate
Styrene
Hydroxy ethyl acrylate
Styrene
Methylated melamine formaldehyde
2-Ethyl hexyl acrylate
Hydroxy ethyl methacrylate
Methyl acrylate
Ethyl acrylate
Styrene
Butyl acrylate

Figure 5.1 Composition of an acrylic melamine enamel binder commonly used in automotive finish layers.

Particles of a substance that have the same index of refraction as the paint binder will not produce any optical effects, but they do serve to add bulk to a paint. Pigments that have indices of refraction that are close to that of the binder are thus referred to as *extender pigments*. They are sometimes used to modify the optical, physical, or chemical characteristics of paint, such as gloss, surface texture, viscosity, strength, chemical resistance (weatherability), or abrasion resistance. Some of the more common extender pigments used in paints include talc, kaolin (clay), calcite (calcium carbonate), quartz (silicon dioxide), mica, and barytes (barium sulfate), all of which have indices of refraction near 1.6.

The vast majority of pigments used in paint produce color by selective absorption of certain wavelengths of visible light. Most inorganic pigments are comprised of transition metal complexes or salts, and their visible absorption properties arise from electronic transitions of valence electrons in d orbitals (Judd and Wyszecki, 1975). Electronic transitions between d orbitals of isolated atoms are not allowed by *selection rules*, but they may become “weakly allowed” when these orbitals form ligand bonds.

Organic pigments are flat planar systems with extensive conjugation (Figure 5.2) and light absorption involves π-electron transitions. These are allowed transitions, and organic pigments are strong absorbers of visible light. Organic pigments therefore have higher tinctorial strengths than inorganic pigments. In regard to other pigment properties, organic pigments tend to produce brighter and more vivid colors, have smaller particle sizes, are less opaque, are less durable to long-term light exposure, and are more expensive than inorganic pigments (Oil and Colour Chemists’ Association, 1983).

In attempting to mimic the iridescence of pearls, butterfly wings, and some other visually striking objects in nature, chemists have devised several specialized pigments for use in cosmetics, paints, and other products. Most such *effect pigments* produce colors by means of light interference (Droll, 1999). A common example of this is the change in color of an oil slick when viewed from different angles. The oil slick consists of a very thin layer of hydrocarbons on water, and interference between light waves reflected from the top of this film and from the oil–water interface produces different colors, depending on the angle at which the film is viewed, the thickness of the film, and its index of refraction. This results from constructive and destructive interference between the two reflected rays. Pearlescent and certain colored and color-shifting pigments are based on this principle, and they consist of refractive laminated flakes that are oriented mostly horizontally in the paint. Color-shifting pigments based on very small

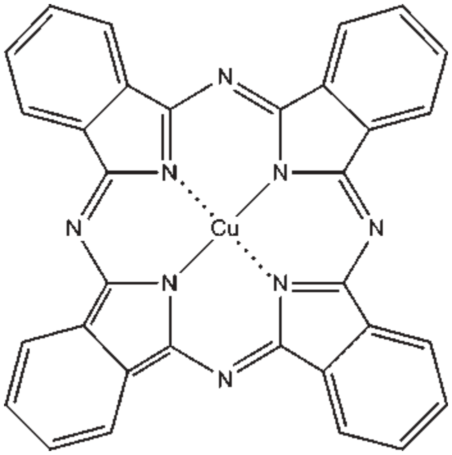

Figure 5.2 Structure of Copper Phthalocyanine Blue (Pigment Blue 15), a very common organic pigment used in automotive and other paints. Note that this structure is only one of several canonical structures and that the four isoindole ring systems are equivalent.

striated plates acting as miniature diffraction gratings have also been introduced recently.

The types of pigments and pigment combinations that are used in a particular paint depend very heavily on the applications of the paint and the visual effects that are desired. All white and light-colored exterior paints normally contain large amounts of rutile, while varnishes, automotive clearcoats, and other transparent finishes contain little or no pigmentation. Metallic paints must have a semitransparent finish so that the metal flakes, which are usually composed of aluminum metal, can be observed. Organic pigments are usually more prevalent in such paints, as they are more transparent than inorganic pigments, owing to their smaller particle sizes and lower indices of refraction. Extender pigments are normally used to control luster in paints, and architectural finishes with low lusters contain large amounts of such pigments. More details regarding paint composition, manufacture, and use are described by Nylen and Sunderland (1965), Oil and Colour Chemists' Association (1983, 1984), Morgans (1990), Lambourne and Strivens (1999), and Bentley (2001).

5.3 TYPES OF PAINT

There is no such thing as a universal finish, as even when discounting consumer tastes regarding color and type of finish, there is no single binder or binder and pigment combination that will meet all of the varied demands of every paint application. Certainly, this is a beneficial situation for the forensic scientist since, as we saw, the composition of a paint can tell us a lot about its uses, and it is this very wide diversity that makes paint so valuable as evidence.

Paint composition is ultimately a compromise between various conflicting performance characteristics, economic issues, and environmental and safety concerns. It is not possible, for example, for a paint to have an extremely hard finish and to be quite flexible at the same time. An epoxy with a bisphenol A component is often used for automotive undercoats and many indoor applications because of its excellent adhesion and other desirable properties, but it is not durable enough with respect to light exposure to be used in automotive finish layers. Acrylics and alkyds have been used extensively for automotive finish layers, providing a hard durable finish on rigid metal and plastic substrates, but they are not used (by themselves) on flexible plastic automobile bumper covers because they are too brittle and would crack readily when subjected to the minor collisions that the bumper itself is designed to withstand.

5.3.1 Automotive Finish Systems

The demands of an automotive finish are among the most stringent of any type produced in the coatings industry. Appearance is paramount for a consumer product, where utility often takes a back seat to status, and visually, automotive finishes are the most diverse in the industry (Panush, 1975; Ehlich, 1988). Such paints span the entire spectrum of colors from black to white (both of which are also popular choices), with virtually everything in between, and more than half of the original finishes contain some sort of “effect” pigment flake (metallic, pearlescent, interference, or diffraction). About the only color that is not used in an original automotive finish is nonmetallic pink (“shocking pink”), but even that hue may be found in some fleet or special-order original finishes. Flat and satin finishes are rare and the overwhelming majority of automotive finishes have high gloss. Durability (color and gloss retention) is a very important consideration, and vehicle paints also have very demanding mechanical requirements, such as chip, corrosion, and chemical resistance. They are carefully applied as highly engineered layer systems, with each layer serving a specific purpose. Binder formulations are highly complex and a broad variety of polymer types are employed.

5.3.2 Architectural Coatings (Structural Paints or House Paints)

Architectural paints are the second most common type of paint encountered in the forensic science laboratory. Unlike automotive paints, they are seldom designed to be applied as layer systems; consequently, layer colors and sequences will vary with the whim of the customer. They are designed either for interior or exterior application, and a good portion of their design is targeted at ease of use. This affects choices as to the binders and pigments to be used, the additives to control viscosity and, consequently, film buildup; and the additives to control stability during storage after use. There are two general chemical classifications for the binders; water-based (latexes) and oil-based. The latexes permit easy cleanup with soap and water, a big selling point for the home user. The oil-based systems require mineral spirits or turpentine for cleanup. Only three major binders are used, although there are many variations within each major class. The latex binders include the poly(vinyl acetate)–acrylic (PVA–acrylic) and acrylic resins, while the oil-based binders are almost solely alkyd resins. There are some poly(vinyl acetate)–polyethylene binders on the latex market, but they command a very small share of the sales (less than 2%), as do some newer latex paints with low concentrations of alkyd resin incorporated. Unlike the case of automotive finishes, appearance other than color is not paramount. This results in a highly competitive market with many more manufacturers than exist for automotive paints. Almost an infinite number of colors are available, but many fall into the off-white or pastel classes, employing very low concentrations of coloring pigments. The paints are pigmented at the retail store and offered in a series of gloss levels, including flat, satin, eggshell, semigloss, and high gloss. The pigment is added to a tint base, which is a 1- or 5-gallon can of white paint of the chosen binder type and gloss type. Even the high-gloss paints do not approach the gloss levels of automotive finish coats. It is extremely unusual for these paints to include effect pigments.

It is difficult to specifically place certain finishes in this general class of coatings; however, some applications deserve mention. Wood coatings, such as colorless varnishes, are of note, with most binders being alkyds or urethanes and some having acrylic copolymers incorporated as well.

5.3.3 Other Coatings

The third most common type of paint encountered in the forensic science laboratory falls into the class of maintenance paints. These finishes are designed primarily for their protective properties and functional characteristics, with color and gloss quality being of secondary importance. They are quite diverse and may be found as coatings on tools, appliances, lawn and garden tools, circuit boxes, safes, ATM devices, valves, bridges, traffic signs, and so on. Considering the applications, it does not take too much imagination to deduce why they are encountered as evidence. The binders and pigments used cover a broad range, with original finishes demanding durability similar to that of automotive coatings. They may be high gloss and may also have effect pigment in them, but rarely interference decorative flake.

Trade-sale spray paints, the type found on the shelves of your local hardware or home improvement store, are yet another general class of end-use product encountered in the forensic science laboratory. Binders typically include alkyds, acrylics, nitrocelluloses, urethanes, and epoxies. Some silicone binders may be found as well. The finishes may have effect pigments incorporated in them; however, application methods are obviously not nearly as controlled as in the automotive industry.

5.4 PAINT EVIDENCE INTERPRETATION CONSIDERATIONS

Although one often thinks of paint as a “wet” medium that is applied and allowed to dry, the forensic paint examiner deals predominately with cured in-service paint films. There may, however, be occasions where a can of paint or spray paint is submitted as evidence and the paint examiner is asked to determine if the contents of the can could have been the source of a paint film. In these instances, the paint examiner will prepare a film of dried paint for comparison.

In a comparison case, two samples are typically received for analysis, referred to as *known* and *questioned* specimens. The former denotes the reference sample, which usually has a known origin (e.g., a vehicle suspected of being involved in a hit-and-run incident), and this is to be compared to sample recovered from a scene (a paint chip from a hit-and-run location) or a suspect, or some other source that may provide a link between events and individuals.

Although it may seem like a straightforward process to compare two paint samples, the forensic paint examiner deals with real-world specimens that are often far from ideal for analysis. Because of the nature of the events that usually produce paint evidence, the questioned specimens are often quite small, such as those recovered in the Ridgway case. The paint may have been in service for many years, subjected to weathering, marring, contamination, repaints, and so on. These processes result in what is decidedly a double-edged sword as far as interpretation of evidence is concerned. They can make obtaining and interpreting data a very demanding task; at the same time, however, some of the unique characteristics that they might confer to a paint can be quite significant if similar features are found on both known and questioned samples.

There are many factors that the forensic paint examiner must consider when interpreting data for paint comparisons. Many of these arise from the inhomogeneities in paint, not only at the microscopic level but at the macroscopic level as well. Many readers have probably seen the effects of weathering on an interior house paint arising from differences in exposure to light: the paint behind a picture on the wall may have a different hue compared to the same paint that has been exposed to sunlight. Bleaching and other effects may be even more pronounced for automotive paints, where, for example, the roof has received considerably more exposure to sunlight and other environmental conditions than has the vertical side of a fender. The original paint systems on many automobiles are likely to be varied to begin with, considering that some have two tone colors, and as noted, paint systems used on metal substrates may differ from those used on plastic bumpers and other plastic parts. Another possible source of inhomogeneity of an original automotive finish commonly arises from factory touch-ups. These occur when a portion of the vehicle is damaged during the assembly process and must be repaired and refinished before leaving the plant. A nonfactory repaint will usually result in even more heterogeneity when only a portion of the vehicle is refinished, as typically occurs. Refinished areas may also differ due to removal of one or more previous layers as a result of sanding. Paint systems therefore cannot automatically be excluded as having a common origin, even if the number of layers does not match. For cases involving paint from a spray can, differences in paint composition can also occur, depending on whether the can has been shaken or not prior to use (Zeichner et al., 1992). Assessing heterogeneity in samples—and its significance—is a major portion of the forensic paint examiner's task.

These considerations emphasize the importance of having appropriate reference samples when comparisons are made, and the various circumstances of which the forensic paint analyst must be cognizant when interpreting paint evidence. Unfortunately, appropriate reference samples are not always available, and the condition of the questioned samples can be an analytical chemist's nightmare. Sample size, sample heterogeneity, layer smearing, or contamination by the substrate onto which the specimen was transferred epitomize some of the difficulties that may confront the paint examiner. Such realities, as well as the lack of comprehensive databases for all types of paints, underscore the reasons why the conclusions that the forensic paint examiner can draw run the gamut from unequivocal to simply unattainable.

5.5 ANALYTICAL METHODS

As discussed, dried paint is a complex matrix comprised of a variety of mostly nonvolatile and inert inorganic and organic ingredients spanning a very wide range of concentrations, and most paints that forensic analysts examine are insoluble enamels. It is thus not surprising that paint examiners typically use a battery of different methods to obtain as much information as practical to characterize, differentiate, identify, and compare such evidence. Methods that provide maximum discrimination between paints are sought and it is desirable that the suite of methods chosen provides complementary information. At the same time, the forensic paint examiner finds it helpful to have some overlap in this information, because by themselves, the methods may provide only limited data for identifying certain components or else provide general class characteristics rather than unequivocal data to identify each component conclusively. Methods that are nondestructive are preferred, both to preserve the evidence and to allow further testing using other techniques. If destructive methods are used, they are normally performed after the nondestructive methods. The choice of methods that are used may also be dictated by the type of paint that is encountered, as well as its size and nature (e.g., the analysis of a paint smear versus the analysis of paint from the entire door of a vehicle).

5.5.1 Microscopic Examinations

As mentioned previously, paint evidence can be found as smears on clothing, smears on objects, minute fragments fused to the surface of an object, or fragments contained in the surface debris recovered from an object. As with any trace evidence, microscopy is necessary in finding and examining the specimen. If you cannot find the material, you obviously cannot analyze it. Furthermore, if you cannot accurately characterize its form and structure, no matter what sophisticated analytical techniques you may have at your disposal, the analysis will ultimately not generate complete and accurate results. The initial instrument of choice for searches and preliminary characterization of recovered specimens is the stereomicroscope. This technique provides a three-dimensional view, with magnifications on the order of 5 to 60 times the object's actual size. It has a broad field of view, permitting relatively large areas to be searched microscopically. Furthermore, it offers long working distances providing ample room to manipulate the specimen for observation and sample preparation for subsequent examination techniques.

Unlike other common types of materials encountered as trace evidence, such as fibers, glass, plastics, adhesives, rubbers, foams, metals, and most composites, paint evidence often has a laminate structure comprised of multiple complex layers. Stereomicroscopy provides the initial view of this layer structure. Layer colors, order, textures, and relative thicknesses are often the initial clue as to the general source of origin of the specimen. Did it originate from a motor vehicle or from an architectural application? Is it more likely a maintenance coating, such as might be found on tools, commercial door frames, or electrical junction boxes? This initial information is important for several reasons. First, it lets the examiner know what to expect from the sample. Thorough background knowledge of all types of coatings is essential. If you are going to perform a forensic comparison of any material, first know your product. In the case of paint, it is important to know how it is manufactured, how it is applied to the substrate, of what is it composed, and what types of common manufacturing irregularities may be encountered. All become important in knowing what to look for and in recognizing what is unusual.

Second, it permits the examiner to deduce from where the sample may have originated. This is commonly an integral question in the investigation, whether it be a hit-and-run investigation or a homicide investigation involving forced entry to a structure. For example, there is usually little benefit in analyzing an architectural paint in detail when it comes from the debris recovered from the surface of a hit-and-run victim. Finally, it permits the examiner to know which pool of background knowledge to access when evaluating the evidential significance of a sample. Just as one does not want to testify that the layer structure of an architectural paint discovered in a homicide investigation is unusual based on their background knowledge of automotive paint layer structures, one equally does not want to attempt to determine the year/make/model vehicle origin of an architectural paint sample discovered in debris recovered from the garments of a hit-and-run victim. All these points may seem unwittingly basic, but striving to accomplish them is anything but.

To appreciate the microscopical methods used in forensic paint examinations, it is probably necessary first to discuss some of the attributes of the various end-use categories described previously (e.g., automotive, architectural, maintenance, and trade-sale spray paints). Due to the specialized nature of the forensic examination of cultural artifacts, this topic is not addressed, although many of the techniques to be discussed are employed in that endeavor as well.

Automotive Paints Automotive paints consist of finish coats applied over primer coats and generally fall into two distinct categories, original finishes and refinishes. Original finishes consist of the multiple paint layers applied at the factory at the time of manufacture. Refinishes commonly refer to either single or multiple layers applied after the vehicle leaves the factory (i.e., aftermarket applications). Refinishes applied at the factory, however, present a third category. Although these systems employ a reapplication of paint layers, the material used commonly has the same composition as that used in the original application. As might be expected, it is valuable to be able to recognize these three categories when assessing the evidential significance of a transfer, and various types of microscopy play an integral part in making those decisions.

Finish coats impart the visual appearance to the paint system and are the uppermost layers of the coating. They are found as either monocoats, clearcoat/basecoats, or tricoats. Monocoats consist of opaque layers designed to have high gloss and color. Both the clearcoat/basecoat and tricoat systems have a transparent layer on top that is typically colorless but may be lightly tinted. This is followed by a single colored and opaque basecoat layer in the former system or a transparent or translucent colored layer followed by the opaque basecoat layer in the latter system. Any one of these finish coat layers other than the clearcoat may have decorative flake dispersed throughout. The decorative flake is commonly either some form of aluminum flake, pearlescent flake, or interference flake. Stoecklein (2002) describes the latter in detail.

Below the finish coats are primers, designed to be an interface between the finish coats and the substrate material. They serve several functions, including adhesion, ease of sanding for leveling out imperfections, and corrosion resistance. There may be anywhere from one to four or five of these layers present; however, the most common layer structures employ only one or two. The primers may usually be distinguished from the finish coats, as they are opaque and include much higher concentrations of extender pigment. This in turn gives them a lower gloss and more particulate appearance. Color-coordinated primers are designed to have a similar color to the basecoat above them and often have slightly lower concentrations of extender pigment present. Figure 5.3 depicts the most common original finish layer systems. The primer-surfacer is a primer that is often present between the base primer and the finish coats. The term *surfacer* is added to the name in order to describe its primary function, a layer that is easily sanded to produce an acceptably smooth surface on which to apply the finish coats. Traditionally, it appeared in a rather limited color palette, consisting of various shades and tints of gray, black, or red-browns. Since the mid-1990s, some vehicle manufacturers began using colors coordinated to the general color of the monocoat or basecoat. Further descriptions of these systems may be found in Fettis (1995), Ryland (1995), Thornton (2002), and Streitberger and Dossel (2008). Anticorrosion treatment of the steel used in automotive panels leaves a characteristic pattern on the bottom of original primers commonly referred to as "orange peel," similar in appearance to that shown in Figure 5.4. Weathering of the surface of the top layer can result in oxidation of the resin and pigment and sometimes cracking of the surface. Improper refinish techniques may result in dirt or foreign material trapped between layers and visible at the interface between the layers. Occasionally, poor painting procedures result in foreign overspray paint droplets deposited on the surface of a topcoat.

Architectural (Structural) Paints Forensic laboratories may also have a substantial caseload, dealing with architectural paints ranging from cases involving

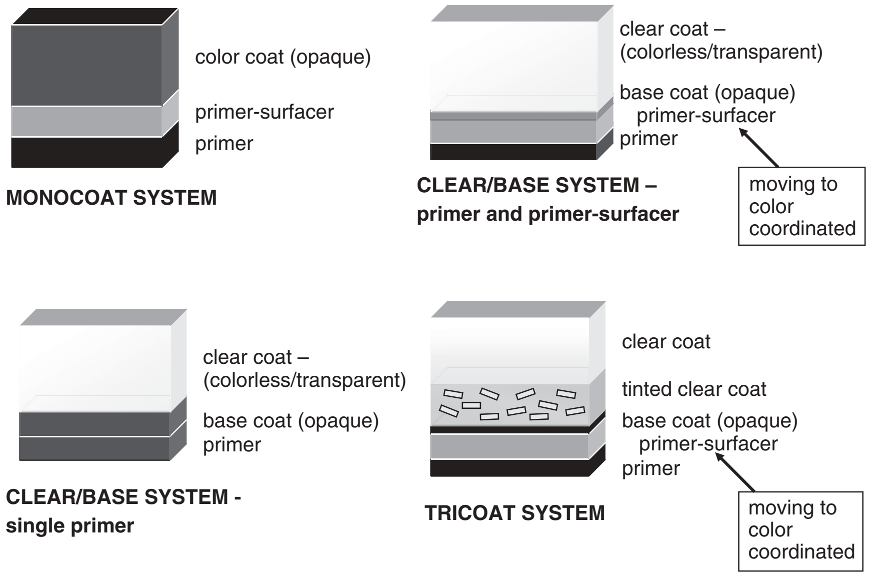
color coat (opaque)
primer-surfacer
primer
MONOCOAT SYSTEM
clear coat –
(colorless/transparent)
base coat (opaque)
primer-surfacer
primer
CLEAR/BASE SYSTEM –
primer and primer-surfacer
moving to
color
coordinated
clear coat –
(colorless/transparent)
base coat (opaque)
primer
CLEAR/BASE SYSTEM -
single primer
clear coat
tinted clear coat
base coat (opaque)
primer-surfacer
primer
TRICOAT SYSTEM
moving to
color
coordinated

Figure 5.3 Diagrammatic representation of common original automotive finish layer structures. *(See insert for color representation.)*

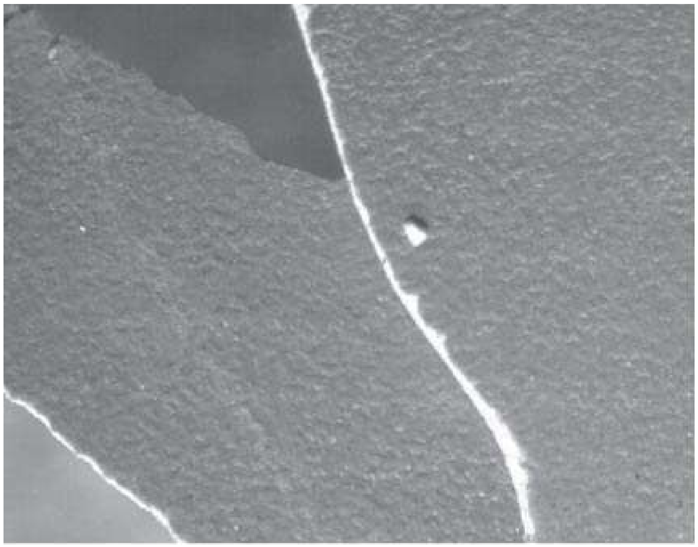

Figure 5.4 Stereomicroscopic view of the underside of the original base primer from an automotive application revealing the characteristic "orange peel" appearance. Magnification of 20×. *(See insert for color representation.)*

forced entry to those dealing with debris recovered from either a suspect's or a victim's garments. There is seldom a predesigned system of layers in these samples—the layer colors, textures, and chemistry changing with the whim of the consumer. The variety of finishes encountered can range from flat wall paints to colorless glossy varnishes on wood. The flat wall paints contain very high concentrations of extender pigments. Even semigloss to high-gloss wall paints are characterized by the presence of some extender pigment, with the grain size being much larger than that encountered in vehicle primers. The surfaces are typically somewhat rough, even though they may have a relatively high gloss, and the pigment dispersion is heterogeneous compared to automotive finishes. One can often see clumps of undispersed coloring pigment throughout the paint layer. Latex paints have a softer texture than automotive paints and when treated with a bit of chloroform form a gummy residue during solvent evaporation, quite unlike automotive paints treated in the same manner. It is unusual to encounter decorative flake, except for the occasional use of retail spray paints for finishing architectural accessories. In those cases, the decorative flake is usually quite dense in distribution compared to that found in most automotive applications. On occasion, a portion of the substrate material will cling to the bottom layer of the paint fragment. Discovery of wood or wallboard obviously indicates that the application was not automotive. Consecutive applications of light pastel or off-white colors that are similar to one another in color and texture are sometimes encountered and present difficulty discerning between layer interfaces. Microscopy is the basic tool for recognizing these characteristics.

Maintenance and Trade-Sale Spray Paints Maintenance and trade-sale spray paints have microscopic characteristics that lie somewhere between architectural finishes and automotive finishes. In the case of tools or commercial metal door frames, they are often coated with both a primer and a finish coat. The finish coats may be high gloss, resembling automotive monocoats. Their surfaces typically will have rough pitted spots over larger areas, but these may not be noticeable in small samples. The chemistry used in the paints is similar to that employed in the automotive industry and hence the textures are quite similar. Often, the layers are quite a bit thinner than automotive paint layers and may be brittle, giving rise to a suspicion that the origin is both nonautomotive and nonarchitectural. In the case of trade-sale spray paints, they are usually encountered as thin layers applied to the surface of existing finish systems and frequently display significant heterogeneity in pigment dispersion. This layer structure suggests nonautomotive applications.

5.5.2 Physical Nature of the Transfer

In addition to the various morphological characteristics attributable to different end-use types of paint, the physical nature of paint transfers encountered in forensic investigations plays a significant role in the choice of examination techniques. In a majority of automotive paint cases, the samples are quite small and multilayered, as depicted in Figures 5.5 and 5.6. Both the microscopical and analytical techniques chosen for comparisons must be compatible with this sample size and morphology, keeping in mind that each layer contains a wealth of morphological and both organic and inorganic chemical information. Some evidence will have numerous layers, suggesting the possibility of encountering another source with a corresponding

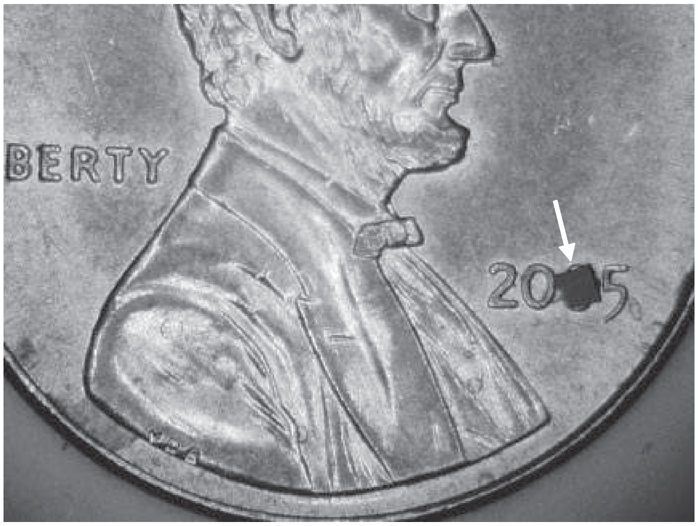

Figure 5.5 Paint fragment placed on the date of a penny for sample size perspective. *(See insert for color representation.)*

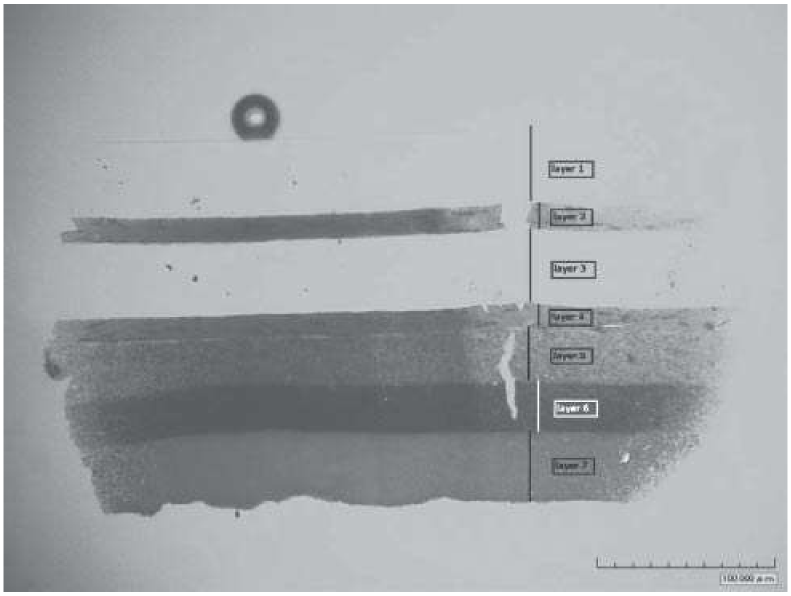

Figure 5.6 Cross section of a seven-layer automotive paint sample mounted on a microscope slide with a 1.56 refractive index mounting media and coverslip and viewed with transmitted brightfield light on a polarized light microscope. Layers 1 and 3 are clearcoats, layers 2 and 4 are basecoats with a decorative flake in them, and layer 5 is a color-coordinated primer followed by two additional primers (layers 6 and 7). *(See insert for color representation.)*

layer structure is extremely remote (Tippett et al., 1968; Gothard, 1976; Ryland and Kopec, 1979; Gothard and Maynard, 1996; Willis et al., 2001; Edmondstone et al., 2004; Wright et al., 2011). In such instances, further chemical analyses may not be deemed necessary (ASTM, 2007). In the case of larger paint fragments, be they automotive or not, the examiner must guard against damaging the edges and surfaces of the evidence since the potential for comparison of the fractured edge contours and the paint chip surface configurations with an appropriate known source hold the possibility of a physical correspondence that is definitive, unlike any compositional comparison. If such a correspondence is found, the analyst will conclude that the paint fragment whose origin is in question originated from and was at one time a part of the paint on the known exhibit.

Paint transfers also occur in the form of smears onto substrates. One such example is a multiple-layer architectural paint smeared onto a tool used to gain entry to premises. Mixing of the smeared layers and contamination of the transfer by the material on the tool's surface become serious analytical issues. Microscopy offers the initial means to evaluate and mitigate the impact of these problems on subsequent analytical techniques. Another such example is an automotive paint smeared onto the clothing of a hit-and-run victim. In such instances, it is often difficult to discern by stereomicroscopic examination alone whether the thin transfer is paint or plastic from a vehicle part. Other microscopic techniques and perhaps analytical techniques are usually necessary to draw a conclusion. Application of those techniques is often thwarted by contamination of the paint smear by the underlying synthetic fabric. Polyester and acrylic fibers have chemical characteristics similar to those of many paint binders, and the abrasion transfer process often generates heat that melts, mixes, and fuses the substrate fiber polymers into the paint transfer.

As described earlier, paint transfers may also occur with uncured paint. If the transferred paint dries following the transfer, microscopy is the only way to ascertain that it was wet when deposited. That little bit of additional information can be extremely valuable in an investigation, as evidenced in the Green River cases. Occasionally, tiny droplets of spray paint will float through the air at an automotive refinish facility and land on the painted surface of a nearby vehicle. This is obviously not good quality control and represents an unusual situation, not to mention that the color and chemistry of the contaminating paint is somewhat random. Recognizing these tiny cured droplets on the surface of both questioned and known paint chips represents a very unusual characteristic that adds immensely to the associative evidence afforded by the corresponding paint fragments alone. The transfer may also occur from a person rubbing against a freshly painted object. A transfer such as this seldom finds its way into the forensic laboratory, but it may. Recognizing this type of transfer is often of value when searching for automotive paint transfers on the clothing of a hit-and-run victim. If a paint smear is found that was transferred in a wet condition, it is probably not from the striking vehicle and routinely turns out to be an architectural paint. As another example, one of the authors (SGR) examined a case where the suspect of a sexual assault rolled over the victim's palette of paints she was using while painting at a secluded area on the beach. The significance of the numerous corresponding colors and chemistries of paints discovered on the suspect's clothing was reinforced further by the fact that the paints were transferred in an uncured condition. All of these transfers demonstrate very smooth surfaces and rounded edges at the perimeters of the deposits upon stereomicroscopic examination. The rounded edges are a result of flow during the curing process and are not present in abraded transfers of cured paint. Even when glossy automotive paint partially melts onto the fabric of a hit-and-run victim's garments as a result of heat generated from friction, the surface of the paint is not smooth nor are the corners at the perimeters rounded from flow.

5.5.3 Microscopy

With an understanding of the sample attributes present in a forensic paint examination, let us now focus in a bit more detail on the types of microscopy and sample preparation available to permit evaluation of those attributes. Stereomicroscopy affords views of the bulk sample on all sides, including the intact bulk layer structure. The small surface area of thin layers, the rough surface of neighboring layers, and the interplay of neighboring layer colors make it difficult to view all layers definitively. Smoothing the surface of the cross-sectional view will help

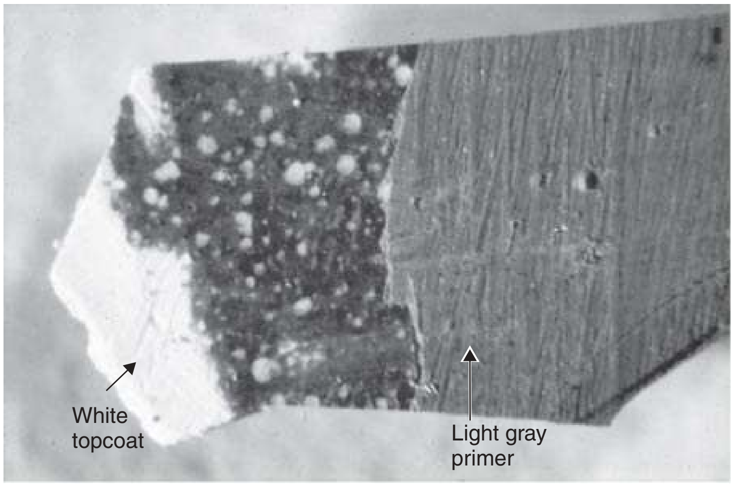

Figure 5.7 Peeling of successive layers to expose larger surface areas permits visualization of unusual characteristics, such as this thin mottled primer–surfacer between the white finish coat and the light gray primer. Stereomicroscopic view at 10× magnification. *(See insert for color representation.)*

somewhat, whether done manually with a sharp blade or polished using abrasives. This does not solve the problem of color rendition of small surface areas or the interplay of neighboring layer colors. Typically, the sample is carved manually from either the top or bottom layer to expose larger areas of each layer, including those on the interior of the sample. Portions of these carvings may be saved for subsequent testing. The carving and subsequent increased layer-area exposure permit stereomicroscopic evaluation of colors, textures, pigment heterogeneity, decorative flake size, shape and color in reflected light, and layer interface materials and their distribution (such as the mottled primer-surfacer between the topcoat and base primer shown in Figure 5.7). However, very thin layers may be missed by this method, especially if they are colorless and transparent. It is also difficult to see the transition from a basecoat to a color-coordinated primer (especially black on black or white on white) or the transition from one layer to another in a multilayered white architectural paint fragment.

To solve this problem, it is advisable to prepare thin cross sections like that in Figure 5.6. This can sometimes be done manually or by embedding a specimen in a plastic medium and then cutting very thin cross sections with a microtome. The sample can be observed on the stereomicroscope, but it is often advisable to view it at greater magnification on a transmitted light compound microscope. This can be accomplished on a simple compound microscope, on a polarized light microscope, or in the case of comparisons on a properly color-balanced compound comparison microscope. The cross sections are placed on a microscope slide and covered with a drop of mounting medium having a refractive index substantially different from the 1.50 to 1.52 of most paint resins. This permits the boundaries of colorless clearcoat layers to be observed distinctly. A coverslip is then placed on top, and the preparation is ready for viewing. It permits critical evaluation of layer structure, layer inclusions, and pigment color and distribution throughout the thickness of the layers. It can often allow classification of an automotive paint layer as a finish coat or a primer, primers having a much more particulate appearance, and is invaluable in discovering tricoat layer structures. In the case of some multiple-layer low-gloss white architectural paints, even this method does not reveal layer interfaces clearly. In that case, it is often of value to observe the cross section using a reflected light fluorescence microscope with a variety of excitation wavelengths, as reported by Allen (1994). This technique may also be of value in distinguishing similarly colored automotive paint primers, as described by Stoecklein (1994).

Aluminum decorative flakes and some interference decorative flakes cannot be distinguished when paints containing them are viewed with reflected light using a stereomicroscope. Transmitted light brightfield microscopy, however, can be used to differentiate the two types by observing thin peels cut from the finish coats that have been placed on a microscope slide with mounting medium and a coverslip. As reported by Stoecklein (2002), many types of interference flakes are based on mica substrates, which are transparent in transmitted light, unlike the opaque aluminum metallic flake (see Figure 5.8). Furthermore, the interference flakes that are opaque typically have sharp-lined geometric shapes, unlike the opaque aluminum flake profiles, which resemble a flake of cornflakes cereal or a sand dollar.

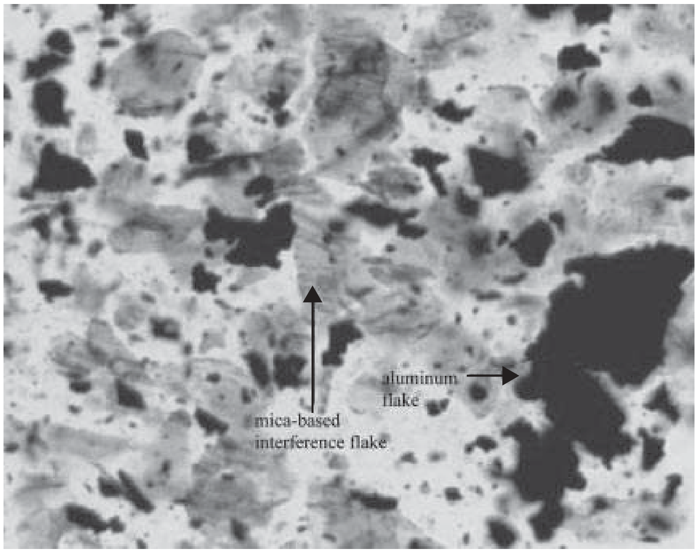

Figure 5.8 Transmitted light microscopic view of a thin peel of an automotive finish coat having two types of decorative flake: mica-based interference flake and aluminum flake. The hand-cut peel is mounted in Norland Type 64 optical adhesive on a microscope slide with coverslip and viewed at 250× magnification on a polarized light microscope with the polars uncrossed. *(See insert for color representation.)*

Observation of thin sections of paint on a polarized light microscope employing crossed polars can also aid in recognizing or even identifying anisotropic extender pigments. If noted in an automotive paint layer, it is a strong indication that the layer is a primer, not a finish coat. If observed in a thin peel of an architectural or spray paint, the grain size may be large enough to permit identifications using basic optical characteristics, as demonstrated by Kilbourn and Marx (1994).

The value of using various microscopic techniques in analyzing paint samples has been demonstrated over the years by a number of authors. McCrone (1979) describes the use of polarized light microscopy in characterizing particulate matter encountered in the forensic science laboratory, while Hammer (1982) describes the use of transmitted light microscopy for automotive paints. Stoecklein (1994) discusses the value of various microscopic techniques for examining automotive paints, and the ASTM *Standard Guide for Forensic Paint Analysis and Comparison* (ASTM, 2007) further details the basic use of microscopy in forensic paint examinations in general.

5.5.4 Microspectrophotometry

Although not all coatings are colored, paint is generally considered to be a pigmented liquid or liquefiable composition which is converted to a solid film after application as a thin layer. Coloring pigments may be either inorganic or organic, and both impart color to the film. Color by its basic definition involves the absorption and reflection of visible light. Beyond observation with the human eye, what more intuitive way is there for the chemist to measure and compare this characteristic than with visible light absorption spectrophotometry?

The comparison of color is one of the first steps taken in a forensic paint comparison. The human eye is a very discriminating sensor, being able to distinguish anywhere from over 100,000 to several million colors (Hunter, 1975; Boynton, 1979); hence, its value as a discriminating tool for paint evidence cannot be overemphasized. However, it is possible to formulate paints that appear the same color to the human eye, yet are comprised of different pigment combinations. Analytical instruments that permit the chemist and industrial technologist to discriminate such colors have been available for six decades. The ultraviolet–visible (UV–Vis) absorption spectrophotometer is known to every undergraduate chemistry and biology major. It is sensitive to the different degrees to which various wavelengths of visible and UV light are absorbed by a material, most typically a solution of the material contained in a UV-transparent cuvette. If two solutions are the same color but contain different chromophores, they will typically produce differing visible absorption spectra. And, of course, the solutions may have no color at all, yet contain organic molecules that have differing aromatic structures and thereby produce differing UV spectra. The chromophores in the organic pigments used in paints are typically rich in aromatic and conjugated ring systems and as such often produce characteristic UV absorption spectra.

UV–Vis spectrophotometry therefore seems to be a natural choice to include in an analytical scheme to compare paint samples. However, the task is not as simple as it may seem. As mentioned previously, the sample is typically dissolved in a solvent to decrease its absorption to a level within the dynamic range of the spectrometer's detector when performing UV–Vis absorption spectrophotometry. Many paint binders, the resins that bind the pigments together and adhere to the substrate, are insoluble in most common organic and aqueous solvents that do not affect the chromophore chemically. In those instances the only way to achieve the aforementioned goal is to make the solid sample very thin, which is often done by slicing or crushing the specimen. However, we also know that it is important to control the pathlength of the samples being compared, because Beer's law states that the absorbance is proportional to the pathlength. This is no easy task when slicing or crushing two samples. Hence, the *Standard Guide for Microspectrophotometry and Color Measurement in Forensic Paint Analysis* (Scientific Working Group for Materials Analysis, 2007) recommends that for detailed comparisons of thin sections of paint the specimens be prepared by cutting the two simultaneously using microtomy. This procedure involves embedding the two samples in a liquid resin, curing the resin, and then mounting the block in a microtome and cutting at controlled thicknesses to produce the same pathlength for both sections. This solves the first problem of dealing with solid specimens; however, most forensic samples present the examiner with a second issue. The samples are quite small and often multilayered. How is one to introduce this sample to a UV–Vis spectrophotometer?

Microspectrophotometers (MSPs) allow for objective measurement of the color of small or microscopic samples and are more precise and quantitative than the more subjective results of visual microscopical color comparisons. As the name suggests, they are UV–Vis spectrophotometers incorporated into microscopes (see Figure 5.9). The design of such instruments is difficult, for typical microscope

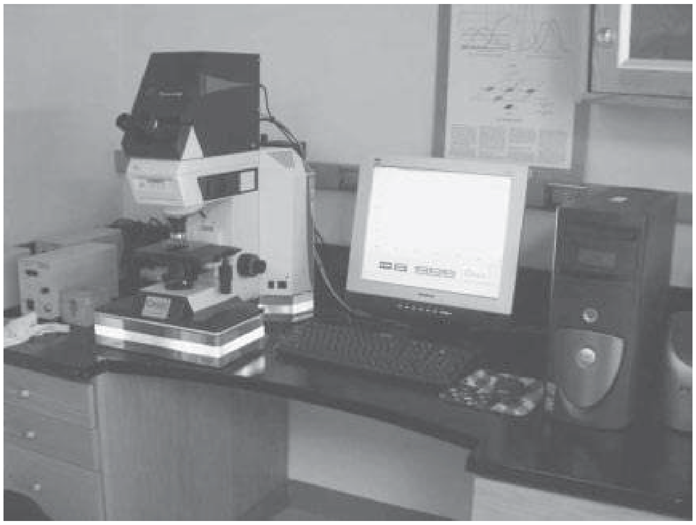

Figure 5.9 CRAIC model QDI 1000 ultraviolet–visible microspectrophotometer with Cassegrainian optics and a semiconductor array detector.

objectives are made of glass, which does not transmit UV light. Additionally, the amount of light transmitted through the small apertures required to define the area of interest in microscopic samples is quite limited and the sensitivity of the photometric detectors must be high while at the same time controlling deterioration of the signal-to-noise ratio. The first problem has been solved either by using non-UV-absorbing glass to make the microscope objectives (which is very expensive) or by using Cassegrainian objectives, which are constructed of mirrors and never pass the beam of light through any glass. The latter suffer from inferior image quality and control of the light path. The second problem has been solved by using very sophisticated focusing optics and highly stabilized photomultiplier tubes or by using fast-response semiconductor array detectors, which permit co-addition and averaging of the signals collected. In either design, the specimen is mounted in a non-UV-light-absorbing mounting medium (typically, glycerin), placed on a quartz microscope slide with a quartz coverslip, and then placed on the MSP's stage. The area of interest is physically defined by adjustable or fixed apertures viewed in the same field of view as the specimen and then background and sample measurements are made. The instrument software ratios the two, and plots of the resulting absorption or percent transmittance spectrum are obtained.

As might be expected, analysis of solid samples is not as straightforward as solution absorption spectrophotometry. There are microscopical concerns, such as light path focusing, sample focusing, and diffraction at specimen edges; spectrometer concerns, such as appropriate calibration standards for wavelength and photometric accuracy; and specimen concerns, such as microheterogeneity and sampling area. Eyring (2002), Stoecklein (2001), and the *Standard Guide for Microspectrophotometry and Color Measurement in Forensic Paint Analysis* (Scientific Working Group for Materials Analysis, 2007) discuss all these concerns in detail. Given the concern for controlling path length and the effect that it has on the interpretation of results, perhaps a brief discussion of sample microtomy might be in order.

Microtomy is a technique employed to prepare thin sections of controlled thickness using a microtome, a mechanical instrument designed especially for this purpose. The discipline finds its primary roots in the histology laboratory; however, its application also finds its way into the field of materials science. For the purpose of paint fragment sectioning, a basic instrument will suffice. Most laboratory or medical supply vendors will carry the array of available equipment, ranging from basic manual models using steel knives to motorized automatic models using cryostatic freezing to attain submicrometer sections. A used rotary microtome with a sharp steel blade will perform adequately, although the use of glass knives will produce sections that do not have striae (from knife blade sharpening) and are much less prone to curling while being cut. The use of glass knives will require a knife maker, special microtomy glass, and a glass knife microtomy stage. A typical rotary microtome with a glass knife can be seen in Figure 5.10 and with a steel knife in Figure 5.11.

The choice of embedding medium is just as important as the choice of microtome. The sample(s) will be placed in this liquid resin and then the resin will be cured to form a rigid block that is placed into the microtome sample jig for

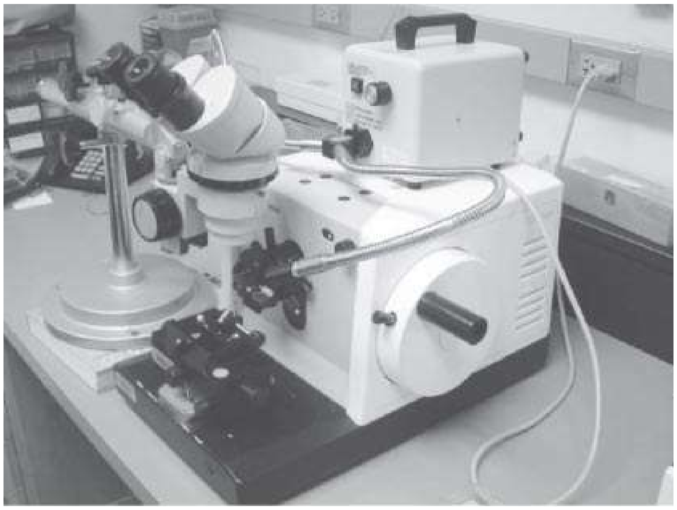

Figure 5.10 Microm manual rotary microtome with glass knife stage and universal sample holder.

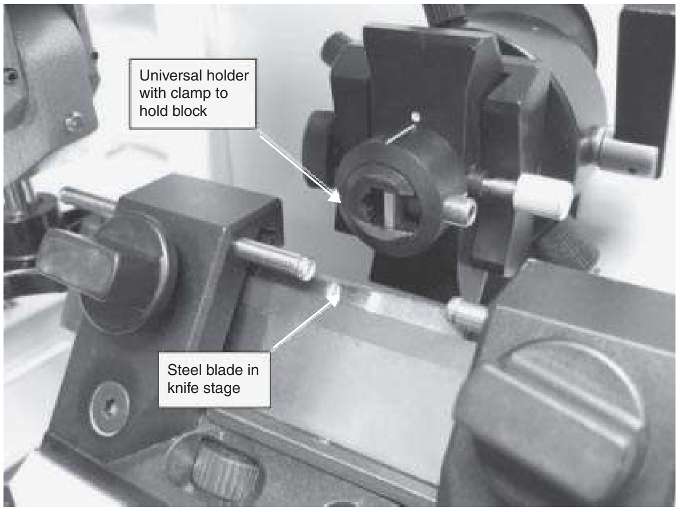

Universal holder with clamp to hold block
Steel blade in knife stage

Figure 5.11 Microm rotary microtome with a steel knife stage and the sample block held in a universal specimen holder.

sectioning. It is recommended that the embedding medium be approximately the same hardness as the specimen to be sectioned for optimum results. If it is too soft, the specimen will be dislodged from the embedding medium when encountering resistance with the blade. If it is too hard the blade will quickly be dulled and produce poor-quality sections. Both Stoecklein (2001) and Derrick (1995) provide guidance on these topics. For those familiar with this technique, one of the authors of this chapter SGR finds that a UV-curing optical adhesive (UV-74 Lens Bond, Summers Optical Co., Fort Washington, PA) works quite nicely, permitting the examiner to place some embedding medium resin in the well of the mold and to cure it for approximately 6 s with a high-intensity UV light to form a thin skin on which to place the first paint fragment. This permits critical positioning of the paint chip so that it will be parallel to the remaining chips placed in the block. After positioning the first chip, some additional embedding resin is added and then flash-cured for about 6 s to form another partially cured surface. The next paint chip is then positioned and more resin added. This process is continued until the block has been completed. After adding the final aliquot of resin, the mold is cured under high-intensity UV light for approximately 6 to 10 min. It is then removed from the mold and all sides are exposed to the UV light to remove the tack from the optical adhesive. The block is then ready to be cut at this time or within a day or two, as depicted in Figure 5.12. Typically, excess resin is trimmed away manually to expose the specimens near the surface. The block is then secured in the microtome jig and trimmed even further using the steel blade until a position very close to the specimens is reached. The microtome knife stage is then switched to the glass knife stage, where the final sections are produced. Three-micrometer sections are cut for brightfield microscopy and UV–Vis microspectrophotometry, while 5-μm sections may be cut for Fourier transfer–infrared (FT–IR) microspectroscopy. Either the remaining block with the exposed paint chip surfaces or a very thick section (20 μm) may be used for subsequent scanning electron microscopy in conjunction with energy-dispersive x-ray spectrometry (SEM–EDS) analyses.

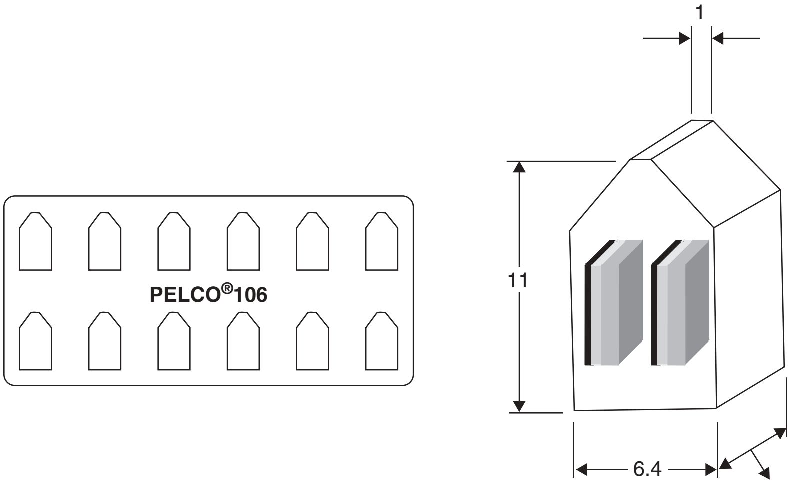

Figure 5.12 Representation of two automotive paint fragments embedded in the same microtome sectioning block (right) and of the top view of the silicone mold used to form the block (left). Dimensions are in mm. *(See insert for color representation.)*

Although this sample preparation approach may seem very appealing given the minimal sample manipulation of the native paint fragment, the production of controlled-thickness paint sections, and the control of surface topography and orientation (critical factors in SEM–EDS analyses), there are drawbacks to the preparation of samples by microtomy alone. First, very small areas of the sample are exposed in cross section. This forces the analyst to deal with microheterogeneity problems, since large analytical areas cannot be used to average heterogeneity. Cured paint is simply not homogeneous on the microscopic scale. Second, difficulties in control of the analytical area may be encountered when using an infrared microscope, due to diffraction produced with the small imaging apertures required, and in SEM–EDS analyses due to excited volumes crossing over into neighboring layers. These concerns are discussed by Ryland (1995), Henson and Jergovich (2001), Bartick (2010), and in the *Standard Guide for Using Scanning Electron Microscopy/X-Ray Spectrometry in Forensic Paint Examination* (Scientific Working Group for Materials Analysis, 2002).

As with most spectrophotometric methods of analysis of solid specimens, visible-light microspectrophotometry can be performed in the reflectance mode as well as in the transmission mode. This approach offers the lure of apparent minimal sample preparation but has notable disadvantages compared to transmission analyses. The signal-to-noise ratio is typically lower in reflectance measurements, and there are sample geometry, spectral distortion, and reproducibility concerns. The sample surface must be sufficiently smooth so as to have ridges and valleys measuring less than the size of the particles contained therein. If not, light is scattered randomly and signal targeting and signal-to-noise quality are lost. To achieve this goal, the sample must be polished on a microscopic level using a series of abrasives typically performed on embedded samples, much as would be encountered in the preparation of samples for microtomy. In addition, it is critical that the specimen(s) surfaces be oriented parallel to the reflectance objective of the microspectrophotometer, as failure to do so results in variation in spectral features. For certain monocoats containing high concentrations of strongly absorbing pigments (such as Copper Phthalocyanine Blue), specular reflection distortion of the measurement can occur (Suzuki, 2010). This process produces *more* reflection for strongly absorbing wavelengths and results in a *bronzing effect* (this phenomenon can also occur for some dark blue inks, which appear to have a metallic bronze sheen when viewed at certain angles). For metallic paints, the incident beam can undergo either direct reflection from a metal flake or diffuse reflectance from the paint matrix, and differences in the ratios of these two processes also affect spectral reproducibility. Hence, it is much more difficult to achieve precision with reflectance techniques than with transmission techniques. Discussion of this shortfall can be found in Eyring (2002), Stoecklein (2001), and in the *Standard Guide for Microspectrophotometry and Color Measurement in Forensic Paint Analysis* (Scientific Working Group for Materials Analysis, 2007).

Despite the intuitive application of microspectrophotometry to forensic paint comparisons, the technique is not widely used. Reports from paint proficiency tests offered by Collaborative Testing Services (2007) over a three-year span indicate only 7 to 15% of the participating laboratories used microspectrophotometry in the examinations. These rates are similar for both architectural and automotive paint samples. The limited application is understandable for either white or pastel flat architectural paints, for they have little color present and are loaded with extender pigments. However, this does not explain the limited use for automotive exemplars. Perhaps this can be traced to the fact that little work has been published to demonstrate the improved discrimination of this method over the classical microscopical and instrumental techniques. This is not surprising given the complexity of automotive paint formulations and layer systems. If different paint manufacturers used different pigments to achieve the same color, it is quite likely that other chemical differences would be present in the formulations as well (Contos and Ryland, 2001). That is not to say that very careful analyses may not have the ability to distinguish between batches, where slight color adjustments have been made to achieve an acceptable final color and appearance, as presented by Stoecklein and Palenik (1998). Similar analyses may even be able to discriminate between the quantity of UV absorbers remaining in the clearcoats of two cars painted with the same paint-layer system but subjected to varying degrees of exposure to the sun (Stoecklein and Fujiwara, 1999). In the case of automotive paints, Kopchick and Bommarito (2006) demonstrated that apparently achromatic finish coats can be discriminated by visible microspectrophotometry, although the study did not evaluate which of those finishes were also distinguishable by other routine analytical examinations. This rationale for automotive finish coats should not be carried over into the realm of other end-use paints, such as household spray paints, where the variation in binders and layer structures is much more limited.

As the number of available organic pigments expands and the diversity in binder formulations appears to be diminishing due to company mergers, international markets, and environmental concerns, it is argued that pigment identification may take on a new role in forensic paint examinations. This may be in the arena of comparisons or even in the task of provenance determinations. Certainly, with some care in sample preparation and access to high-magnification objectives, microspectrophotometry has the ability to focus on individual pigment grains in situ and effect identifications based on collections of known pigment spectra. The future will tell what benefits and acceptance this approach holds.

5.5.5 Infrared Spectroscopy

This technique involves measuring the absorptions of a sample in the infrared region, which consists of electromagnetic radiation with frequencies lower than that detected by the human eye. The frequencies in this region are denoted by the wavenumber, which is the reciprocal of the wavelength given in centimeters (cm^{-1}); the region of primary interest is between 4000 and 200 cm^{-1}.

The infrared absorption process results in an increase in the quantized vibrational energies of the molecules of the sample. Some of the energies are localized in vibrational modes involving specific chemical bonds of the compound, and these absorptions often occur over a narrow and characteristic range in the infrared region. Other vibrational modes may involve the entire molecule or large portions of the molecule, and their absorptions occur over a wide range of frequencies and tend to be characteristic of the molecule as a whole. For organic compounds, the former class of absorptions allows an analyst to determine the presence of functional groups, whereas the latter allows the differentiation—and hence the identification—of even closely related compounds. Organic chemists thus use this technique for both diagnostic and identification purposes.

For inorganic compounds containing predominantly covalent bonds (such as silica and other framework silicates), a somewhat similar situation occurs, although these infrared absorptions tend to occur at lower frequencies than for organic compounds with absorption breadths that are often broader. For salts involving cations or anions with covalent bonds (such as $NH_4{}^+$ or $CO_3{}^{-2}$), characteristic absorptions of these ions also occur, although they tend to be less distinct than those of organic compounds since they are often broader and far fewer in number. The only compounds that do not absorb in the region examined by infrared spectroscopy are homonuclear diatomic molecules (most of which are gases such as N_2 and O_2), and simple inorganic salts that do not contain covalent bonds (such as KBr and CsI, which are used as windows for infrared spectroscopy).

All of the components of paint thus produce infrared absorptions, and infrared spectroscopy is one of the main tools used by both the coatings chemist and the forensic analyst to determine the overall composition of paint. For most paints, the binders that are used can be identified by these means, as well as the pigments that are present in high concentrations. The notable exception to this occurs for some architectural coatings with low lusters, as these typically contain large quantities of inorganic extender pigments, the strong broad absorptions of which typically obscure most binder absorptions.

Paint Analysis Methods The two main sampling accessories that are used with FT–IR spectrometers to obtain paint spectra are the diamond anvil cell (DAC) and the infrared microscope. Less often, attenuated total reflectance (ATR) may also be used. Both high- and low-pressure DACs are used for the analysis of paint evidence, although the latter has become the main device used in the United States, due to its thinner windows, ability to be used on an infrared microscope, and lower cost. Paint samples are pressed between the two 1-mm-thick diamond windows (known as *anvils*) of the DAC, or they may be pressed onto a single window. Although the DAC requires much larger samples than can be analyzed with the infrared microscope, one of the main advantages of this accessory is that when used on an extended-range FT–IR spectrometer, the lower frequency (far-infrared) absorptions of the sample can be observed.

An infrared microscope accessory allows the paint examiner to choose a small rectangular area of the sample to be analyzed by the spectrometer while viewing the paint specimen optically. Areas as small as 20 by 20 μm can be analyzed, and multilayered paint specimens can be cross-sectioned, then each layer analyzed

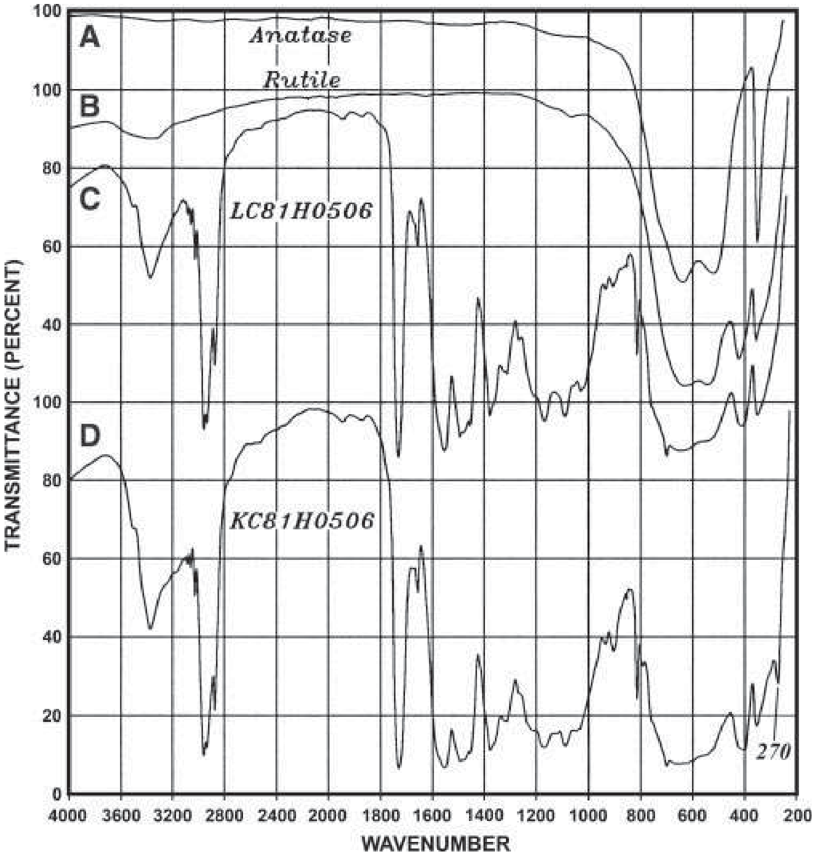

Figure 5.13 Infrared spectra of rutile, anatase, and two yellow nonmetallic automotive monocoats from the *Reference Collection of Automotive Paints*. The monocoats are identified by their *Reference Collection* codes: (A) rutile; (B) anatase; (C) LC81H0506; (D) KC81H0506. Both monocoats have acrylic melamine enamel binders with styrene and contain rutile, Nickel Titanate, and an organic pigment. KC81H0506 also contains hydrous ferric oxide. [From Suzuki (1996b); copyright © ASTM International; reprinted with permission.]

sequentially. The main advantages of the infrared microscope are thus its size and sequential analysis capabilities. Care must be exercised when performing a sequential analysis to set the edge of the aperture as far from the adjacent layer or layers as possible. This minimizes the amount of stray light from an adjacent layer or layers (caused by diffraction) that might contribute to the spectrum of the layer of interest. This is not an easy task in the case of thin layers.

One drawback of the infrared microscope arises from the sensitive detectors that must be used with this accessory, as they have limited spectral ranges. Their low-frequency cutoff points vary between 700 and 450 cm^{-1} and there is an inverse relationship between sensitivity and range; the wider the range, the less sensitive

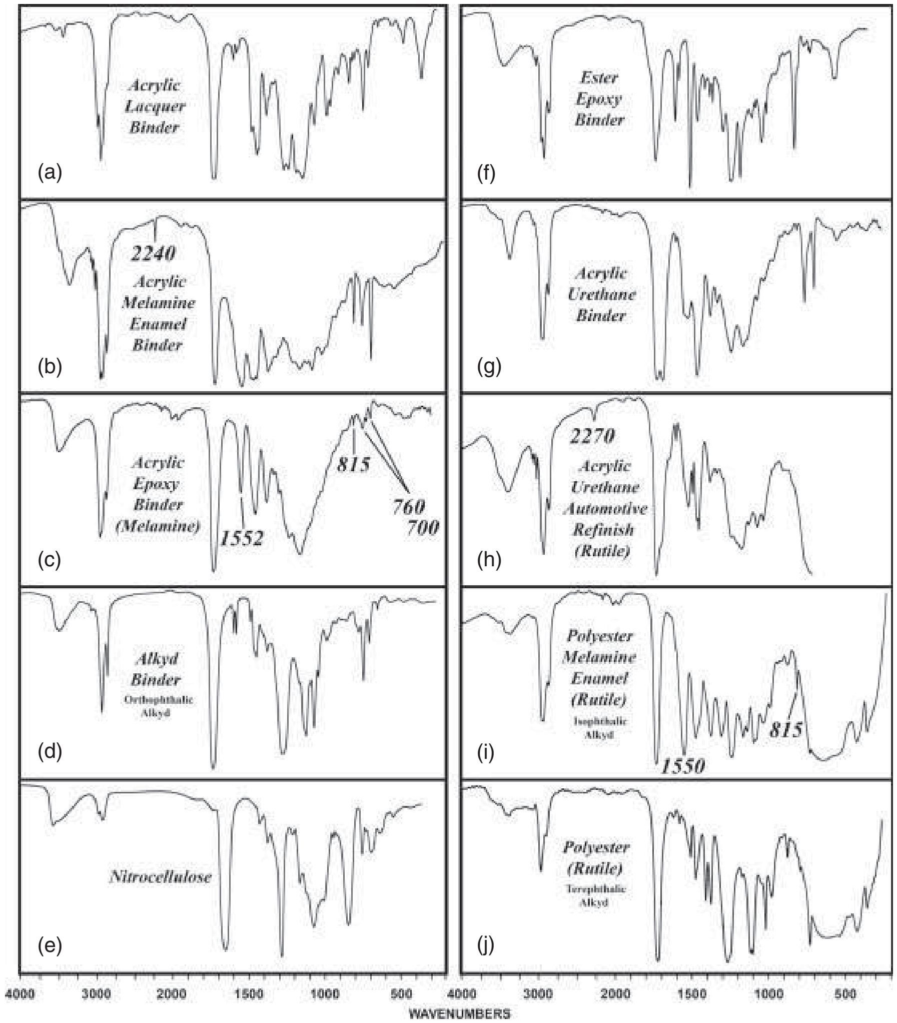

Figure 5.14 Infrared spectra of 10 different paint binders or paints that consist primarily of binders; three of the paints also contain significant amounts of rutile. (A) Acrylic lacquer. (B) Acrylic melamine enamel that contains styrene and an acrylonitrile copolymer. (C) Acrylic epoxy clearcoat (an acrylic melamine enamel with a small amount of styrene; the minor aliphatic epoxy component does not produce a characteristic infrared absorption). (D) Alkyd (orthophthalic alkyd). (E) Nitrocellulose. (F) Epoxy (aromatic) with an ester modifier. (G) Acrylic urethane with a small amount of melamine. (H) An acrylic urethane used in automotive refinishes that contains rutile; the weak absorption at 2270 cm^{-1} is due to unreacted isocyanate starting material. (I) Polyester melamine enamel (isophthalic alkyd) that contains rutile. (J) Polyester (terephthalic alkyd) that contains rutile.

the detector. Many infrared microscopes are thus equipped with the most sensitive detector, even with its narrower range. This does not adversely affect the identification of most paint binders, but it does limit the identification of those inorganic pigments that have most or all of their absorptions below 700 cm^{-1}. Differentiation of the two polymorphs of titanium dioxide used in paints (rutile and anatase), for example, requires the observation of low-frequency absorptions between 500 and 300 cm^{-1} (see Figure 5.13A and 5.13B). The infrared microscope spectrum of a white automotive monocoat that contains rutile is depicted in Figure 5.14H, and only the shoulder of its absorption is observed (the detector used for this analysis has a cutoff point at 680 cm^{-1}). Spectra of two other paints that were collected using a DAC on an extended-range instrument are shown in Figure 5.14I,J, and as can be seen, both contain rutile and not anatase. Another very common inorganic pigment used in paints is ferric oxide (Fe_2O_3, rust-colored). The three absorptions of this pigment occur below 700 cm^{-1} (Figure 5.15A) and they may be observed in the spectrum of a maroon nonmetallic automotive monocoat that contains this pigment (Figure 5.15B). The number of Fe_2O_3 absorptions that are observed using an infrared microscope will depend on the specific range of the detector, but none will be observed using the most sensitive detector.

ATR is a reflectance method that is utilized in either a single-pass accessory or as an objective of an infrared microscope. Paint samples are simply pressed against an ATR element, which is a refractive infrared-transparent material such as zinc selenide or diamond. ATR has a very shallow penetration depth that is proportional to wavelength, so ATR spectra have stronger relative intensities for the lower-frequency absorptions than those of transmission spectra. ATR spectra of paints containing inorganic constituents may also exhibit considerable distortions of the absorptions of these constituents. This occurs because ATR penetration depths also depend on indices of refraction of the sample, and inorganic compounds typically exhibit large changes in their indices of refraction across an absorption band.

Because of the very shallow ATR penetration depths, thin slices of paint do not have to be prepared and paint specimens can be analyzed intact, allowing spectra of the outer layers of a multilayered paint chip to be readily obtained. An ATR objective of an infrared microscope also provides the best means to analyze paint smears that are too thin to be physically removed from the substrate. Such smears can be analyzed in situ, although the spectra obtained by this method should always be compared to those of the substrate since, even with a shallow penetration depth, absorptions of the substrate may be observed in the resulting spectra. If this occurs, it is possible that only the lower-frequency absorptions of the substrate will be observed since these frequencies are more likely to reach the substrate. An ATR spectrum of a thin smear of an automotive clearcoat layer onto a light-colored paint, for example, may well exhibit a “tail” of a rutile absorption (as seen in Figure 5.14h). Although the clearcoat does not contain rutile, the substrate paint does, so the differential nature of the analysis process must always be considered when interpreting such spectra. This point was emphasized in a case involving several smears of automotive paint, which were analyzed using ATR (Giang et al., 2002).

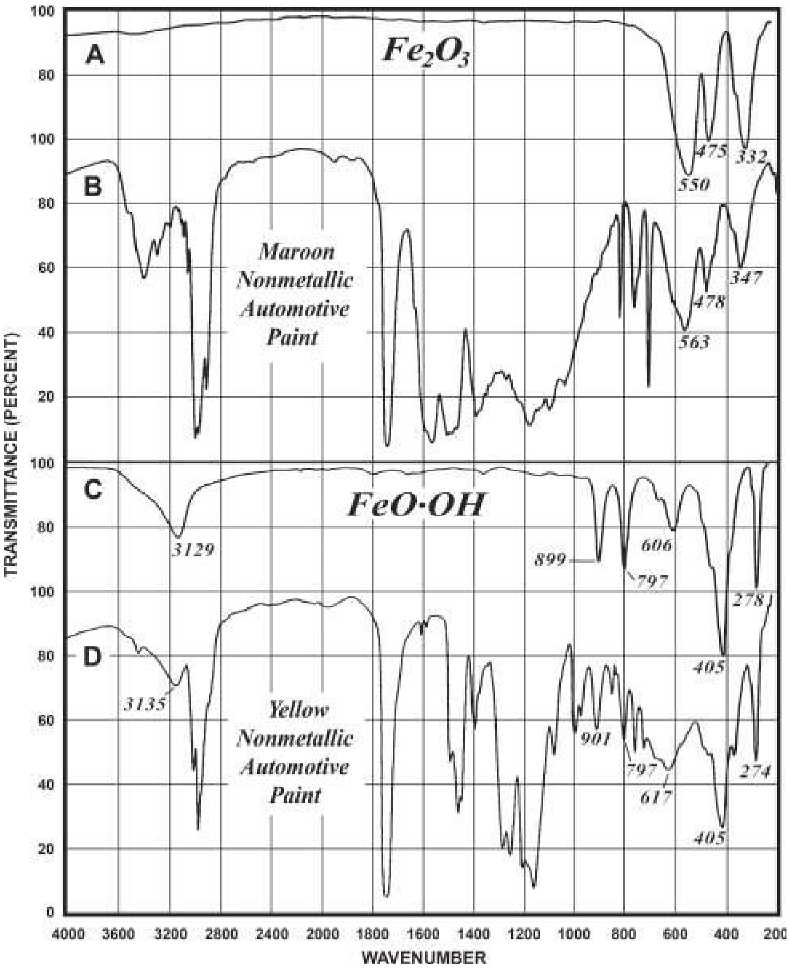

Figure 5.15 Infrared spectra of two iron oxide pigments and two automotive monocoats from the *Reference Collection of Automotive Paints* that contain these pigments. (A) Ferric oxide, Fe_2O_3. (B) Maroon nonmetallic acrylic melamine enamel with styrene that contains ferric oxide. (C) Hydrous ferric oxide, FeO•OH. (D) Yellow nonmetallic acrylic lacquer that contains hydrous ferric oxide. [From Suzuki (1996b); copyright © ASTM International; reprinted with permission.]

More details regarding the ATR analysis process are described by Mirabella (1993), while the differences between ATR and transmission paint data are noted by Ryland et al. (2001) and discussed in general by Suzuki (2010). Bartick et al. (1994) describe some forensic applications using an ATR objective of an infrared microscope. Paint analyses using the DAC (Schiering, 1988; Suzuki, 2010) and the infrared microscope (Allen, 1992; Ryland, 1995; Bartick, 2010) are also described in more detail elsewhere.

Binder Analyses One of the most common binders that the forensic paint examiner is likely to encounter is the acrylic melamine enamel (Figure 5.1) since, as noted, it is often used in automotive finish layers (basecoats and clearcoats). The absorptions of this binder are thus discussed in some detail to illustrate the types of information that forensic paint examiners can deduce from infrared spectra, as well as the limitations of this technique.

The spectrum of a typical acrylic melamine enamel binder is shown in Figure 5.16, and the characteristic absorptions that paint analysts use to identify components of this binder are indicated. The main absorptions of the acrylic component are the ester carbonyl stretching absorption at 1730 cm^{-1} and the series of C—O stretching absorptions between 1300 and 900 cm^{-1}. The latter are somewhat more characteristic of acrylics than the ester carbonyl absorption, since spectra of almost all paint binders contain an ester absorption (produced either by the binder or binder-related components). Melamine produces a broad absorption near 1550 cm^{-1} and a sharper, weaker absorption at 815 cm^{-1} due to vibrations of the triazine ring (Figure 5.1), and these two absorptions are observed in spectra of other binders containing melamine (such as the polyester melamine enamel of Figure 5.14I). Styrene, present in most acrylic melamine enamels, may produce three sets of absorptions (Figure 5.16): (1) weak or very weak ═C—H stretching absorptions above 3000 cm^{-1}; (2) two very weak overtone-combination bands between 2000 and 1800 cm^{-1} which appear as "blips"; and (3) weak to medium ═C—H out-of-plane bending vibrations at 760 and 700 cm^{-1} (Figure 5.16).

Some of the variations that are observed between spectra of acrylic melamine enamels arise from differences in the relative amounts of melamine or styrene present, as well as differences in the acrylic copolymers that are used. One means

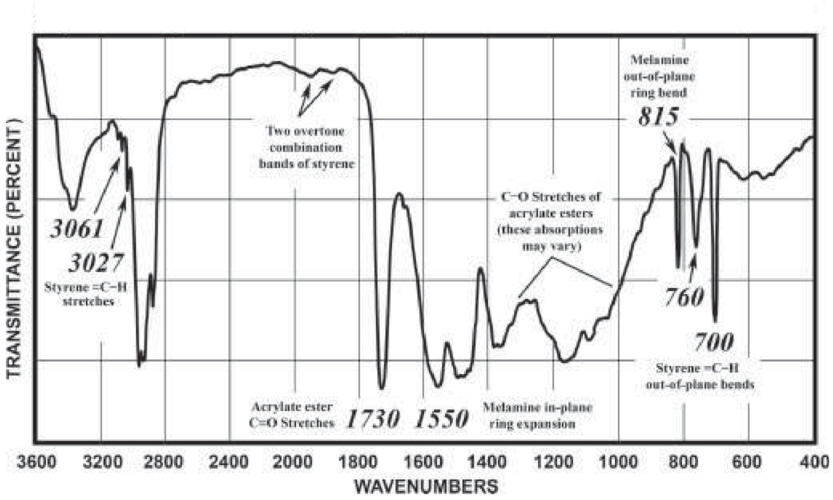

Figure 5.16 Infrared spectrum of a typical acrylic melamine enamel binder with styrene.

to gauge the relative amount of melamine present is to compare the intensity of the carbonyl stretching absorption at 1730 cm^{-1} (which reflects the presence of all of the acrylic copolymers that comprise the backbone) to that of the adjacent melamine 1550 cm^{-1} absorption. Care must be exercised when judging relative infrared intensities, as a nonlinear logarithmic scale is involved in the percent transmittance presentation. Variations may be introduced when working with samples of differing thicknesses or nonpristine samples of limited sizes. Nonetheless, noticeable differences in this ratio may occur, as seen by comparing the spectra of the two automotive monocoats depicted in Figures 5.17A, and 5.17B; a more dramatic difference is seen by comparing Figure 5.14B and 5.14C. Although the binder of Figure 5.14C is referred to as an acrylic epoxy by PPG, the manufacturer of this paint, the aliphatic epoxy component does not produce significant infrared absorptions for characterization, and in terms of its composition, the spectrum is indicative of an acrylic melamine enamel with a relatively small amount of melamine (and a small amount of styrene).

Styrene is found in the vast majority of acrylic melamine enamels used for automotive finishes, and its levels may also vary. The 760/700 cm^{-1} styrene pair may appear to be quite weak when there is a large amount of rutile present, as seen in Figure 5.13C,D. This is a consequence of the logarithmic nature of the percent transmittance scale, which produces a compression of absorption intensities when baselines are low. The other absorptions of styrene ($=$C—H stretches above 3000 cm^{-1} and the two overtone-combination band “blips” between 2000 and 1800 cm^{-1}), however, are easily seen in Figure 5.13C,D. The three sets of styrene absorptions can be seen in spectra of the acrylic melamine enamels depicted in Figures 5.14B, 5.15B, 5.17B and 5.18C. They may not all be observed, however, when small amounts of styrene are present (Figures 5.14C, 5.17A, and 5.18B) or when large amounts of rutile or other pigments are present (Figure 5.19A, 5.19F). Styrene is used in other binders, and this same set of absorptions may be observed in their spectra; styrene $=$C—H stretches and the two “blips” are seen, for example, in the spectrum of the acrylic urethane automotive refinish of Figure 5.14H.

Several different acrylic copolymers are used to create the backbone of an acrylic melamine enamel (Figure 5.1), and in some cases this produces differences in their infrared spectra (Rodgers et al., 1976a). The differences occur mainly in the C—O stretching region between 1300 and 900 cm^{-1} (Figure 5.16), as can be seen by comparing these absorptions in Figure 5.17A and 5.17B and Figure 5.18B and 5.18C. These differences are empirical, that is, they do not permit one to identify which set of copolymers is present. However, acrylonitrile, which contains the —C$\equiv$N group (see Figure 5.1), produces a sharp nitrile stretching peak at 2240 cm^{-1}, which does allow one to identify the presence of this copolymer (Suzuki, 1996a). Although weak, this absorption is readily identified (see Figure 5.14B) because it occurs in a region usually devoid of other significant features.

Another common acrylic binder that was used for many General Motors vehicles from the 1960s to the early 1990s is acrylic lacquer (Ryland, 1995). Although it, too, contains a mixture of acrylic copolymers, the main copolymer is poly(methyl methacrylate) (Plexiglas) and the infrared spectrum of this binder

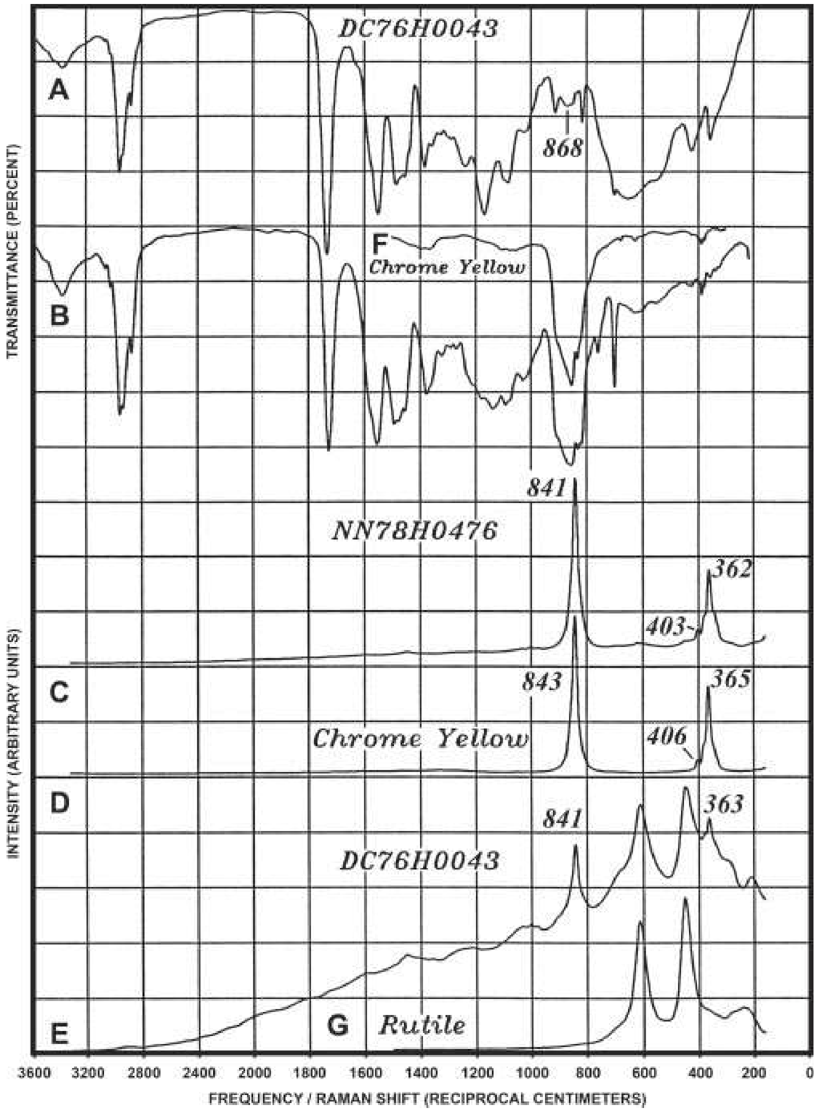

Figure 5.17 (A) Infrared spectrum of a yellow nonmetallic acrylic melamine enamel automotive monocoat with styrene, DC76H0043, which contains a large amount of rutile and a small amount of Chrome Yellow (the Chrome Yellow absorption is marked with its frequency). (B) Infrared spectrum of a yellow nonmetallic acrylic melamine enamel automotive monocoat with styrene, NN78H0476, which contains a large amount of Chrome Yellow. (C) Raman spectrum of NN78H0476. (D) Raman spectrum of Chrome Yellow. (E) Raman spectrum of DC76H0043. (F) Infrared spectrum of Chrome Yellow. G. Raman spectrum of rutile. [From Suzuki and Carrabba (2001); copyright © ASTM International; reprinted with permission.]

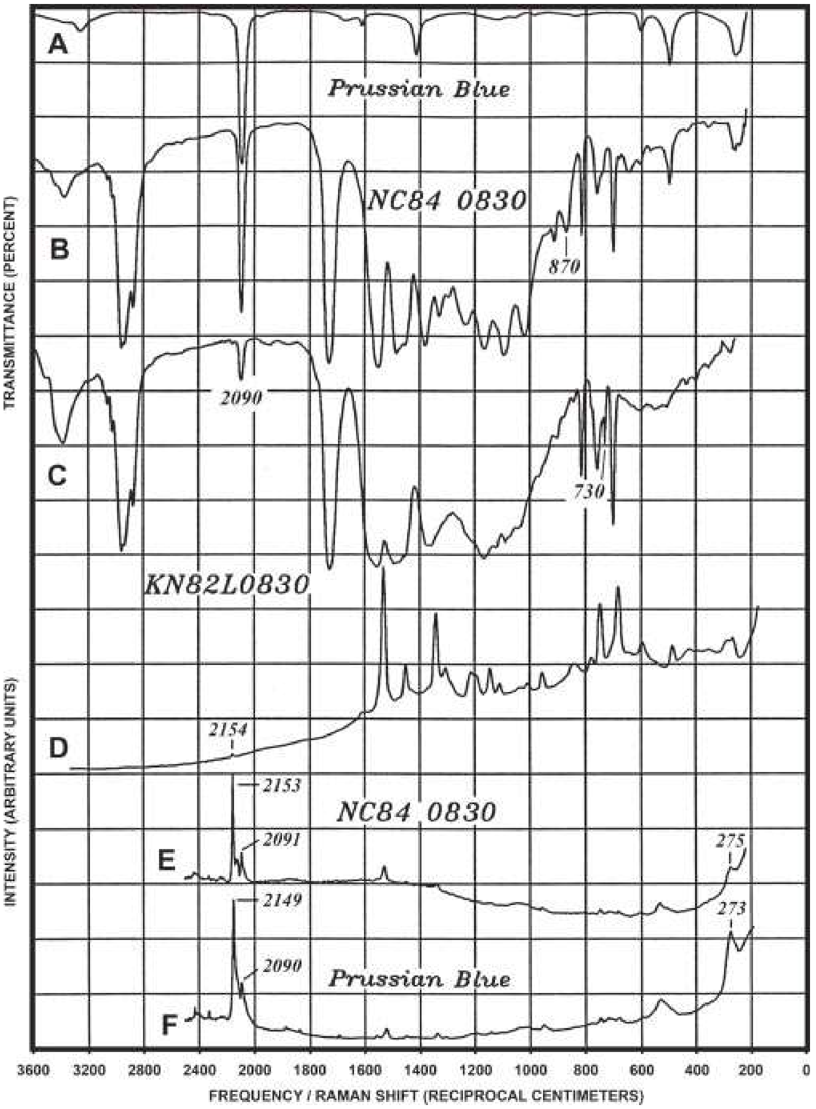

Figure 5.18 (A) Infrared spectrum of Prussian Blue. (B) Infrared spectrum of a dark blue nonmetallic acrylic melamine enamel automotive monocoat with styrene, NC84 0830, which contains a large amount of Prussian Blue. (C) Infrared spectrum of a dark blue nonmetallic acrylic melamine enamel automotive monocoat, KN82L0830, which contains a small amount of Prussian Blue; this paint has the same color as NC84 0830. (D) Raman spectrum of KN82L0830. (E) Raman spectrum of NC84 0830. (F) Raman spectrum of Prussian Blue. [From Suzuki and Carrabba (2001); copyright © ASTM International; reprinted with permission.]

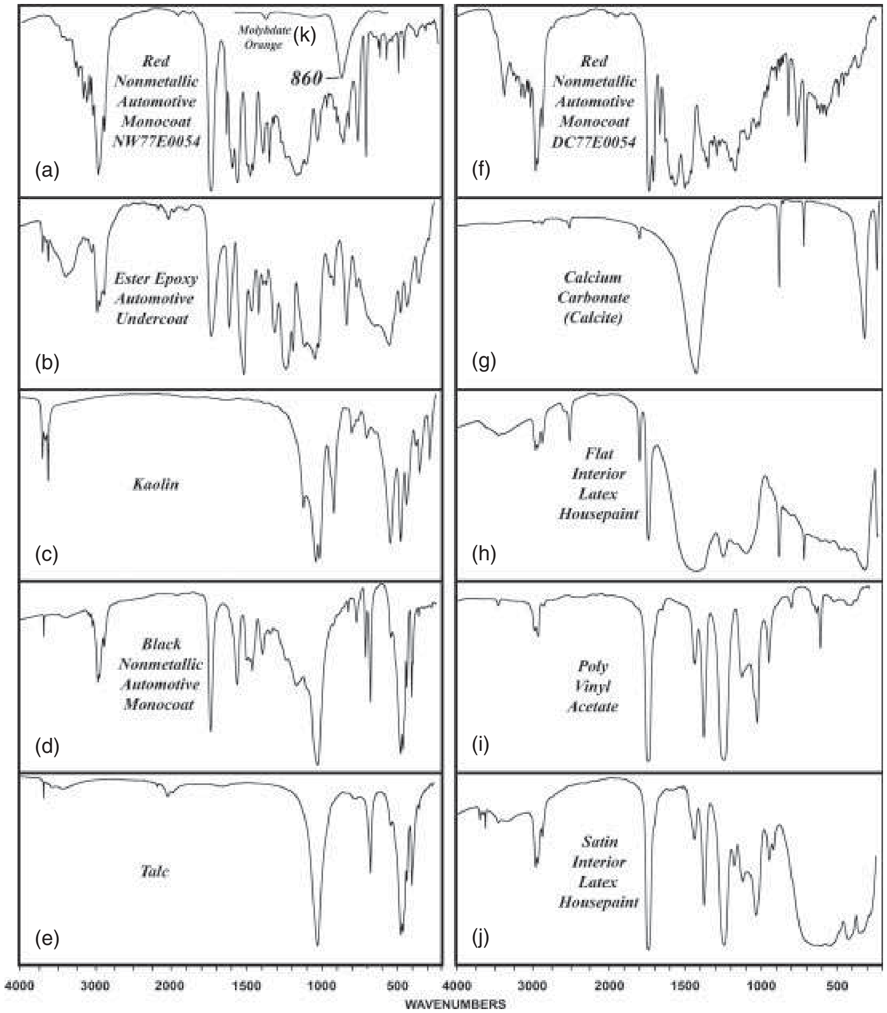

Figure 5.19 Infrared spectra of paints and pigments. (A) Red nonmetallic acrylic melamine enamel monocoat with styrene that contains Molybdate Orange and Quinacridone Red Y. (B) Automotive undercoat with an ester epoxy binder that contains rutile and kaolin. (C) Kaolin. (D) Nonmetallic black automotive monocoat with styrene that contains talc. (E) Talc. (F) Red nonmetallic acrylic melamine enamel monocoat with styrene that contains rutile, ferric oxide, Quinacridone Red Y, and Benzimidazolone Orange. (G) Calcium carbonate (Calcite). (H) Flat interior latex house paint that contains calcite, a silicate, and titanium dioxide or some other oxide. (I) Poly(vinyl acetate). (J) Satin interior latex paint that contains rutile and a small amount of kaolin. (K) Molybdate Orange.

(Figure 5.14A) bears a gross similarity to that of this plastic. Binders referred to simply as “acrylics” are also used in automotive and other paints and have spectra similar to that of Figure 5.14C (the “acrylic epoxy”), except that melamine may not be present, or there may be more styrene (Rodgers et al., 1976a).

The spectra of some other paint binders are depicted in Figure 5.14D to J and Figure 5.19I to illustrate the diversity of their absorption patterns. These are spectra of the binders themselves or of paints that contain little or no observable pigment absorptions (except for the paints of Figure 5.14H to J, which contain large amounts of rutile). Figure 5.14D,I,J are spectra of an alkyd (orthophthalic alkyd) binder, an automotive undercoat with a polyester (isophthalic alkyd) melamine binder and rutile, and a beige paint from a three-ring binder hole punch consisting of a polyester (terephthalic alkyd) binder and rutile. The compositions of the three binders differ primarily in the substitution pattern of two ester groups on an aromatic ring, which are *ortho, meta*, and *para*, respectively, and they produce distinct absorption patterns. All three binders are polyesters, but for historical and chemical reasons, different nomenclatures have been used to describe them, and the term *polyester* has been used for two distinct compositions.

The nominal polymers of three other paint binders—nitrocellulose, epoxy, and urethane—do not by themselves contain an ester carbonyl group, but when used in paints, they almost always have a modifier (copolymer, block copolymer, cross-linking chain, side chain, or plasticizer) that does. The spectrum of nitrocellulose is depicted in Figure 5.14E. As noted, a film of this polymer by itself is quite brittle and it is always used with a block copolymer such as an alkyd or a plasticizer; the absorptions of nitrocellulose, however, are often the strongest in spectra of such binders. The epoxy binders used in paints are aromatic epoxies based on bisphenol A and are actually polyethers; they always contain a modifier such as an acrylic, alkyd, or polyester, so an ester carbonyl absorption is observed in their spectra (Figure 5.14F).

Urethanes are carbamates with the R_1O—(C═O)—NHR_2 functional group, which has a C═O stretching absorption that occurs below 1700 cm^{-1}. This characteristic absorption of urethanes is usually manifested as either a doublet with the ester carbonyl absorption (Figure 5.14G) or as a shoulder to the ester peak (Figure 5.14H), and is accompanied by the urethane C—N stretch absorption at 1520 to 1530 cm^{-1}. The spectra of a few urethanes may also have a weak to medium absorption in the region 2270 to 2260 cm^{-1} (Figure 5.14H) produced by an isocyanate (R—N═C═O) functional group (Suzuki, 1976a). An isocyanate compound is used as the starting material for most urethanes, and its presence in the paint represents unreacted precursor. Like the nitrile group of acrylonitrile (Figure 5.14B), this absorption is usually readily identified because so few compounds absorb in this region. (The very weak triplet of absorptions near 2200 cm^{-1} in Figure 5.14C,G,I, are artifacts; the paints were sandwiched between both DAC anvils, and this may result in what appear to be weak absorptions caused by differences in reflection from the diamond faces.) The acrylic urethane of Figure 5.14G also appears to contain a small amount of melamine, evidenced by the shoulder band near 1550 cm^{-1} and the very weak sharp peak at 815 cm^{-1}. Figure 5.19I is the spectrum of poly(vinyl acetate) (PVA); a PVA–acrylic binder is the most common used in latex house paints, and the spectrum of this binder is primarily that of PVA since the acrylic component is minor and its absorptions generally overlap those of PVA.

The identification of paint binders based on their infrared absorptions is treated in more detail by Ryland (1995), who also presents an automotive paint binder classification flowchart for their identification based on certain key absorptions. Flowcharts are intended primarily for use with paints lacking significant pigment absorptions or for use by analysts who already possess some pattern recognition skills for absorptions of common binders and pigments. Some caution should thus be exercised in their use and it should be remembered that as new products evolve, their spectral characteristics may not be reflected in existing flowcharts. Nonetheless, they may be a useful aid when used with the proper precautions, and an updated chart used for the identification of automotive paint binders is presented in Figure 5.20.

Several new binders were introduced for automotive paints during the period from roughly the mid-1980s to the mid-1990s. A collaborative study using infrared spectroscopy was therefore conducted to determine the discriminating characteristics of this technique for black nonmetallic basecoat/clearcoat finishes used in the 1990s (Ryland et al., 2001). Such finishes lack most of the microscopic discriminating features described previously, and carbon black, which does not produce

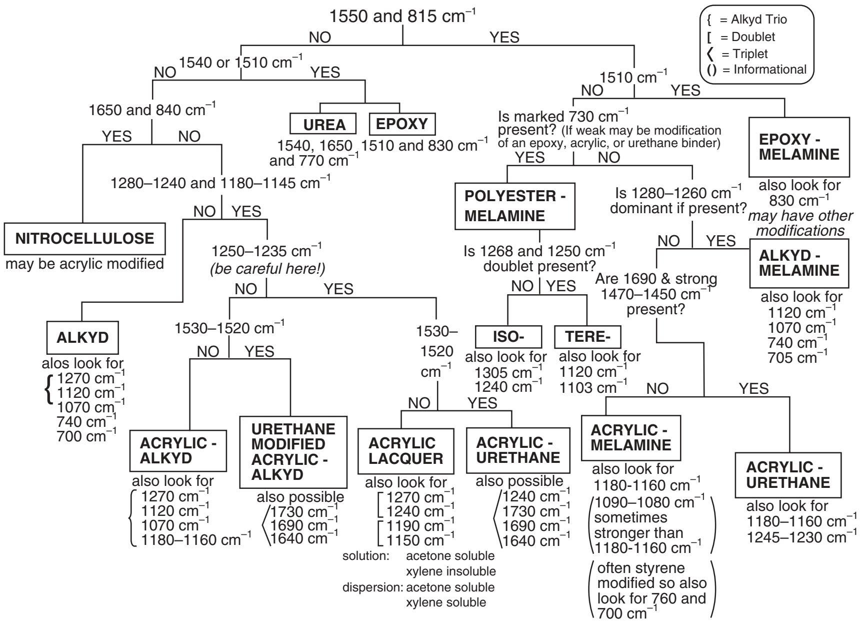

{ = Alkyd Trio
[= Doublet
❬ = Triplet
() = Informational
1550 and 815 cm^{-1}
NO
YES
1540 or 1510 cm^{-1}
NO
YES
1510 cm^{-1}
NO
YES
1650 and 840 cm^{-1}
YES
NO
UREA
1540, 1650 and 770 cm^{-1}
EPOXY
1510 and 830 cm^{-1}
Is marked 730 cm^{-1} present? (If weak may be modification of an epoxy, acrylic, or urethane binder)
YES
NO
EPOXY - MELAMINE
also look for 830 cm^{-1}
may have other modifications
1280–1240 and 1180–1145 cm^{-1}
NO
YES
POLYESTER - MELAMINE
Is 1280–1260 cm^{-1} dominant if present?
NO
YES
NITROCELLULOSE
may be acrylic modified
1250–1235 cm^{-1}
(be careful here!)
NO
YES
Is 1268 and 1250 cm^{-1} doublet present?
NO
YES
ALKYD - MELAMINE
also look for
1120 cm^{-1}
1070 cm^{-1}
740 cm^{-1}
705 cm^{-1}
Are 1690 & strong 1470–1450 cm^{-1} present?
ALKYD
alos look for
1270 cm^{-1}
1120 cm^{-1}
1070 cm^{-1}
740 cm^{-1}
700 cm^{-1}
1530–1520 cm^{-1}
NO
YES
1530–1520 cm^{-1}
ISO-
also look for
1305 cm^{-1}
1240 cm^{-1}
TERE-
also look for
1120 cm^{-1}
1103 cm^{-1}
NO
YES
NO
YES
ACRYLIC - ALKYD
also look for
1270 cm^{-1}
1120 cm^{-1}
1070 cm^{-1}
1180–1160 cm^{-1}
URETHANE MODIFIED ACRYLIC - ALKYD
also possible
1730 cm^{-1}
1690 cm^{-1}
1640 cm^{-1}
ACRYLIC LACQUER
also look for
1270 cm^{-1}
1240 cm^{-1}
1190 cm^{-1}
1150 cm^{-1}
solution: acetone soluble
xylene insoluble
dispersion: acetone soluble
xylene soluble
ACRYLIC - URETHANE
also possible
1240 cm^{-1}
1730 cm^{-1}
1690 cm^{-1}
1640 cm^{-1}
ACRYLIC - MELAMINE
also look for
1180-1160 cm^{-1}
(1090–1080 cm^{-1} sometimes stronger than 1180-1160 cm^{-1})
(often styrene modified so also look for 760 and 700 cm^{-1})
ACRYLIC - URETHANE
also look for
1180–1160 cm^{-1}
1245–1230 cm^{-1}

Figure 5.20 Automotive paint binder classification flowchart based on infrared absorptions. Updated from an earlier version published in Ryland (1995).

discrete infrared absorptions, is typically the only pigment used. Binder composition thus assumes a much greater role as a means to distinguish such finishes, and this collaborative study confirmed the suitability of using infrared spectroscopy for this purpose and demonstrated the reproducibility of infrared spectra acquired by numerous laboratories using different instrumentation.

Inorganic Pigment Analyses Because colored inorganic pigments are generally weak absorbers of visible light and strong absorbers of infrared radiation, high concentrations are used in some paints, and their absorptions may be prominent in the spectra of these paints. In addition to ferric oxide (Figure 5.15A,B), there are two other common iron-containing inorganic pigments, hydrous ferric oxide and Prussian Blue. The spectrum of hydrous ferric oxide ($FeO \cdot OH$), a yellow pigment, is shown in Figure 5.15C. Absorptions of this pigment may be seen in the spectrum (Figure 5.15D) of a yellow nonmetallic automotive paint with an acrylic lacquer binder (compare to Figure 5.14A).

Prussian Blue, which has a dark blue hue (Navy Blue), is comprised of iron and other salts of the $[Fe(CN)_6]^{4-}$ anion; its spectrum is depicted in Figure 5.18A (Suzuki, 1996a). The strongest absorption of this pigment is the cyano ($C{\equiv}N^-$) stretching peak at 2090 cm^{-1}, which like the acrylonitrile (Figure 5.14B) and isocyanate (Figure 5.14H) peaks, occurs in a region where it can readily be observed. It is easily seen in the spectra of two blue nonmetallic automotive monocoats (Figure 5.18B,C) that contain this pigment; both paints have acrylic melamine enamel binders with styrene.

The lead chromates are another important family of colored inorganic pigments, comprised of Chrome Yellow, Molybdate Orange, and silica-encapsulated versions of the two (Suzuki, 1996b). The spectrum of Chrome Yellow, a bright yellow–orange inorganic pigment that consists of lead chromate with a lesser amount of lead sulfate ($PbCrO_4 \cdot x PbSO_4$), is shown in Figure 5.17F. The absorptions of this pigment can be seen in the spectrum (Figure 5.17B) of a bright yellow–orange nonmetallic automotive monocoat, which has an acrylic melamine enamel binder with styrene. Molybdate Orange is comprised of lead chromate, lead molybdate, and lead sulfate ($PbCrO_4 \cdot x PbMoO_4 \cdot y PbSO_4$) and its spectrum consists of one main absorption near 860 cm^{-1} (Figure 5.19K), in the same region as the Chrome Yellow absorption (Figure 5.17F). When present in high concentrations, the minor differences in the absorptions of Chrome Yellow and Molybdate Orange may be observed in some paint spectra, although this is usually not the case when low concentrations are used. Figure 5.19A is the spectrum of a red nonmetallic automotive topcoat with an acrylic melamine enamel binder and styrene that contains Molybdate Orange together with an organic pigment.

Three of the most common inorganic extender pigments used in paints are kaolin (Figure 5.19C), talc (Figure 5.19E), and calcium carbonate (Figure 5.19G). Kaolin and talc are silicates that contain hydroxyl groups involving little or no hydrogen bonding, hence the frequencies of the O—H stretches are relatively high (above 3650 cm^{-1}) and the bands are quite sharp. Although weak, these peaks are thus conspicuous and are often easy to detect in paint spectra. The spectrum of an automotive undercoat with an epoxy binder is depicted in Figure 5.19B, and absorptions of both rutile (Figure 5.13B) and kaolin can be seen. The spectrum of the satin latex paint depicted in Figure 5.19J also contains weak kaolin absorptions, along with strong absorptions of rutile. All of the absorptions of talc (Figure 5.19E) are observed in the spectrum of the black nonmetallic acrylic melamine enamel automotive monocoat with styrene shown in Figure 5.19D. This paint is quite unusual in that it has a flat finish, whereas almost all automotive finishes have a high gloss. The presence of such a large amount of talc is, in fact, the reason for this lack of luster, as this extender pigment is serving as a flatting agent.

Calcium carbonate has two common polymorphic forms, calcite and aragonite, both of which may be used in paints. The spectra of the two can be distinguished (Infrared Spectroscopy Committee of the Chicago Society for Coatings Technology, 1980), and the spectrum of calcite, the more common of the two, is shown in Figure 5.19G. This spectrum lacks the sharp O—H “handles” seen in the spectra of kaolin or talc, but it does have two very weak peaks at 2515 and 1797 cm^{-1}. When calcite is used in high concentrations in paint, the two are often conspicuous in spectra because the former is too low in frequency for a C—H stretch and the latter is quite high for a carbonyl stretching absorption. The two can be seen in the spectrum of a flat interior latex house paint (Figure 5.19H), which contains a very large amount of calcite, a silicate (probably silica), and titanium dioxide or some other oxide. As noted, paints having low lusters typically contain large amounts of inorganic pigments, and their broad and strong absorptions often obscure binder features in paint spectra. Only two of the PVA absorptions below 2000 cm^{-1} (Figure 5.19I) are observed in Figure 5.19H. In contrast, most of the PVA absorptions are observed in the spectrum of an interior latex house paint that has a satin finish (Figure 5.19J). Although characterization of the binder is hampered for such low-luster paints, the prominent absorptions of the extender pigments are an important means to differentiate such paints, since the concentrations and combinations of these pigments vary considerably.

Organic Pigment Analyses Organic pigments can be identified by their absorptions in certain paint infrared spectra, although these features are typically weaker than those of inorganic pigments. Organic pigment peaks are narrower than most binder or inorganic pigment absorptions, and they occur over a wider spectral range than inorganic pigments. They are mostly observed in spectra of some paints that have bright vivid colors (characteristic of many automotive finishes), however, and are rarely seen in spectra of paints with pastel or pale shades and blue hues (Suzuki, 2010).

Figure 5.19A,F show spectra of two bright red automotive monocoats that contain absorptions of both organic and inorganic pigments (Suzuki and Marshall, 1998); both paints have acrylic melamine enamel binders with styrene, and the organic pigment absorptions are the sharp features not observed in the spectrum of this binder (Figure 5.16). Analysts who are not familiar with binder absorptions, however, would probably find it very difficult to distinguish binder features from those of either inorganic pigments or organic pigments. The spectrum of Figure 5.19F, in fact, contains significant absorptions of two inorganic pigments and two organic pigments. Familiarity with binder absorption patterns is thus the most important skill for paint analysts to acquire, followed by familiarity with common inorganic pigment absorptions. Organic pigments, in contrast, are best identified by comparisons to literature reference spectra rather than by trying to memorize their absorptions.

Some of the organic pigments or classes of organic pigments that have been identified in automotive finishes using infrared spectroscopy include benzimidazolones (Suzuki and Marshall, 1997), quinacridones (Suzuki and Marshall, 1998), DPP Red BO and Thioindigo Bordeaux (Suzuki, 1999a), Isoindolinone 3R, Isoindoline Yellow, and Anthrapyrimidine Yellow (Suzuki, 1999b), and perylenes (Massonnet and Stoecklein, 1999a). Other spectra of organic pigments may be found in the various editions of *An Infrared Spectroscopy Atlas for the Coatings Industry* (Infrared Spectroscopy Committee of the Chicago Society for Coatings Technology, 1980; Infrared Spectroscopy Atlas Working Committee, 1991). These volumes are also the most comprehensive compilations of spectra of other paint ingredients, including binders, inorganic pigments, and additives.

Color is the most distinguishing feature of certain types of paints, particularly automotive finishes (Ryland and Kopec, 1979). It is thus desirable that the analytical methods that are used in addition to microscopy and microspectrophotometry provide as much discrimination as possible between paints having very similar colors. For paints of the same color where different binders are used, there will clearly be differences in their infrared spectra. Similar pigment formulations are often used to produce the same color; however, this is not always the case, even when similar binders are used. Different pigments or pigment combinations may be used to produce the same color, as illustrated by the spectra of Figures 5.13C,D,B,C, and Figures 5.19A,F. These are spectra of pairs of original automotive monocoats from two different suppliers used for a specific color on a particular vehicle model. As discussed previously, the monocoats of Figure 5.18B,C both contain Prussian Blue, but with significantly different concentrations. Figure 5.19A, 5.19F show spectra of two automotive monocoats with bright orange–red nonmetallic hues (referred to in the industry as "Fire Engine Red"). The first paint (Figure 5.19A) contains Molybdate Orange and Quinacridone Red Y; the second (Figure 5.19F) contains rutile, ferric oxide, Quinacridone Red Y, and Benzimidazolone Orange. The main difference between the spectra of Figure 5.13C and 5.13D is a peak of hydrous ferric oxide (Figure 5.15C) in the latter spectrum (the other peaks of this pigment are mostly obscured).

Rodgers et al. performed most of the early work on the forensic identification of automotive paint constituents using infrared spectroscopy, and their series of papers serves as an excellent introduction to this subject (Rodgers et al., 1976a–c). A statistical method has been applied to evaluate the significance of minor differences in the infrared spectra of some automotive topcoats to determine their value as evidence (Zieba and Pomianowski, 1981). The infrared spectra of various automotive refinish paints were examined and considerable variations in their compositions were noted, as well as differences in the types of refinish binders compared to those of original finishes (Percy and Audette, 1980). A collaborative study involving the infrared analysis of automotive undercoats has demonstrated that even when two or more pigments are present, it is usually possible to identify the pigments and the major binder components based on their absorptions (Norman et al., 1983). Buzzini et al. (2005) used infrared spectroscopy to analyze paint on 207 crowbars that had been seized as evidence. The purpose of this study was to determine the types and frequencies of such paints. Simulations involving crowbars and painted wood surfaces were also conducted, and it was found that paint transfers, including surface to tool and tool to surface, were common.

The infrared analysis of paint evidence has been described in review treatments by Ryland (1995) and Beveridge et al. (2001). The analysis of paint using FT–IR spectroscopy has also been reviewed from the perspective of the coatings chemist (Hartshorn, 1992). General guidelines for forensic paint analyses are outlined in the *Standard Guide for Using Infrared Spectroscopy in Forensic Paint Examinations* (Scientific Working Group for Materials Analysis, 2009).

Vehicle Identification The identification of an unknown paint recovered from the scene of a hit-and-run incident requires a suitable reference collection or database. Until 1990, paint examiners in the United States relied on the *Reference Collection of Automotive Paints* (Collaborative Testing Services, 1989), a comprehensive collection of original finishes used on American automobiles manufactured between 1974 and 1989. Many of the automotive paints discussed in this chapter are from this collection, and they are referred to by their *Reference Collection of Automotive Paints* identification codes. The codes provide information about the paint manufacturer, binder composition, and the year that formulation was first used. The *Reference Collection of Automotive Paint Technical Data* booklet (Collaborative Testing Services, 1989) provides a list of the vehicle models for which each such paint may have been used.

As an example, the two dark blue nonmetallic monocoats with spectra shown in Figure 5.18B (NC84 0830) and 5.18C (KN82L0830) have the same color, which is indicated by the "0830" numerals. Older paints have a letter color code and the "L" of KN82L0830 signifies that this paint was classified as blue. The "N" of NC84 0830 indicates that the paint was manufactured by BASF Inmont, while the "C" denotes an acrylic enamel binder; this paint was first manufactured in 1984 based on the "84." KN82L0830 was manufactured by Glasurit America, has a nonaqueous dispersion enamel binder, and was first produced in 1982. This paint was used on some 1982 Jeeps and on American Concords, Eagles, and Spirits. NC84 0830 was used on some 1984 Jeep Cherokees and on American Alliances, Eagle 30s, Eagle SX4s, and Encores.

Although a useful aid for identifying domestic vehicles, the *Reference Collection of Automotive Paints* had two major limitations. It did not include the complete finish system, including primers and other undercoats, nor did it include paint from imported vehicles. In addition, after 1989 the only comprehensive original automotive paint finish coat collection in the United States was maintained by the FBI Laboratory. In 1994, a committee of international forensic paint examiners, then known as TWGPAINT (now part of SWGMAT, the Scientific Working Group for Materials Analysis), addressed this issue. TWGPAINT obtained funding to expand an existing database maintained by the Royal Canadian Mounted Police (RCMP) Laboratories, known as the Paint Data Query (PDQ) system.

The goal of the PDQ system is to compile data for the complete original finish systems of all vehicles in North America, domestic and import, with infrared data collected for each individual layer of the finish system. The system includes both text and spectral search capabilities, and users input data about binders and pigments, as well as color information, from an unknown original automotive finish. When all of these factors are included in the search, it has generally been found that the list of possible vehicles is quite short. Often, the list consists only of vehicles manufactured during a short time period at one specific assembly plant. More information regarding the PDQ system is discussed elsewhere (Buckle et al., 1997; Beveridge et al., 2001; Ryland et al., 2006).

5.5.6 Raman Spectroscopy

Like infrared spectroscopy, Raman spectroscopy involves vibrational transitions of the molecules of a sample. Raman spectra are generated by an entirely different mechanism, however, involving *inelastic scattering* of a monochromatic laser source. The laser may have any frequency in the near-ultraviolet, visible, or near-infrared region, and the interaction of this radiation with the molecules of the sample can be viewed as an oscillating electric field driving a dipole (the molecules). This generally results in *elastic scattering* (*Rayleigh scattering*), where the scattered radiation has the same frequency. Inelastic scattering (Raman scattering) occurs when some of the energy is transferred to vibrational modes of the molecules of the sample. As a result, the scattered radiation has less energy, and this difference is equal to the energy of the vibrational transition that occurs. A Raman spectrum is a plot of the intensity of the scattered radiation versus its frequency, with the abscissa depicting the *Raman shift*, the difference in frequency (in wavenumbers) between the laser line and the scattered radiation.

The Raman effect is quite weak, and the vast majority of the scattered light (by a factor or 10^5 or so) occurs as Rayleigh scattering. Florescence is a much more efficient process than Raman scattering, and if it occurs (caused either by the analyte itself or by impurities), it may overwhelm the much weaker Raman peaks. In addition to fluorescence, the other main difficulty that may occur when attempting to obtain a Raman spectrum of paint arises from the absorption of the laser light by the sample. This can cause localized heating and possibly destruction of the sample, so that low laser power levels and defocused beams are normally used.

Dispersive charge-coupled device (CCD) array and Fourier transform (FT) Raman spectrometers are the two main types of Raman instruments commercially available. The former uses gratings to separate different wavelengths of Raman-scattered light, which is then dispersed onto a CCD detector array. Dispersive CCD spectrometers use near-ultraviolet or visible lasers. FT–Raman spectrometers employ an FT–IR interferometer in lieu of a dispersing device, a near-infrared laser, and a separate near-infrared detector. Because Raman scattering is proportional to the fourth power of the frequency of the scattered radiation, Raman spectral peaks are stronger when they occur in the visible region compared to the near-infrared region. CCD detector arrays are also more sensitive than near-infrared detectors, so dispersive CCD instruments are inherently more sensitive devices than are FT–Raman spectrometers. Since fluorescence is the main problem encountered with Raman spectroscopy, however, FT–Raman systems may be more applicable for some samples. Both types of instruments may be fitted with a microscope attachment that allows selection of the area to be examined while viewing the sample optically, analogous to an infrared microscope.

Essentially no sample preparation is required with Raman spectroscopy when examining many paints, other than ensuring that the surface is clean. The specimen is simply positioned at the focal point of the instrument collection optics. Using a Raman microscope, sequential analyses of individual layers of cross-sectioned multilayered paint can be conducted. Unlike infrared microscopy, thin sections do not have to be prepared, and spectral contributions from adjacent layers are much less of a concern. For the most part, the technique is nondestructive, although it is possible to create small craters in the paint when laser power levels are too high. Because clearcoat layers are mostly transparent to the incident laser, Raman spectra of basecoats can often be obtained without removing the clearcoat.

A Raman spectrum is generally quite distinct from an infrared spectrum of the same compound. For a few very symmetric compounds, in fact, the transitions that are observed in a Raman spectrum cannot be observed using infrared spectroscopy, and vice versa. Although this is usually not the case, the two techniques are still very complementary, particularly when they are applied to a complex matrix such as a paint that contains inorganic and organic components spanning a wide range of concentrations. Symmetric vibrational modes may produce no infrared absorptions or only weak ones, whereas they often give strong Raman scattering peaks. As noted for pigments, inorganic compounds may produce broad infrared absorptions, but their Raman scattering peaks are usually narrow. Many inorganic pigments are much stronger Raman scatterers than binders. In certain instances, pigments (usually organic) that can absorb the laser light produce significantly enhanced Raman scattering, known as the *resonance Raman effect* (Johnson and Peticolas, 1976), and even when present in very low concentrations, Raman peaks of such pigments may be observed. Raman spectra of paints are thus typically much simpler than corresponding infrared spectra, as they are usually dominated by pigment features.

Some of the differences that are observed between infrared and Raman spectra of paints can be seen by comparing Figures 5.17A, and 5.17E, 5.17B and 5.17C, 5.18B and 5.18E, and 5.18C and 5.18D. These are infrared and Raman spectra, respectively, of four automotive monocoats, and the Raman spectra were acquired using a dispersive CCD system (Suzuki and Carrabba, 2001). As discussed previously, Figure 5.17B is the infrared spectrum of a yellow acrylic melamine enamel with styrene that contains Chrome Yellow. The Raman spectrum of this paint (Figure 5.17C), however, is predominantly that of Chrome Yellow (Figure 5.17D), and binder peaks are almost totally absent. Figure 5.17A is the infrared spectrum of another yellow acrylic melamine enamel with styrene that contains a large amount of rutile. A small amount of Chrome Yellow is also present, but the weak infrared absorption at 868 cm^{-1} cannot be definitely attributed to this pigment since it may also be due to either Molybdate Orange (Figure 5.19K) or a binder feature (Figure 5.18B). The Raman spectrum of this paint, however, indicates clearly that Chrome Yellow is present along with rutile (Figure 5.17G), as the Raman spectrum of Molybdate Orange has a doublet with equal intensities near 362 cm^{-1} (Suzuki and Carrabba, 2001) rather than a singlet. The sloping baseline of Figure 5.17E is caused by fluorescence.

Copper Phthalocyanine Blue is by far the most common blue pigment used in automotive paint as well as many other paints. Copper Phthalocyanine Blue has a very high tinctorial strength (unlike Prussian Blue), so the concentrations of this pigment used in paint are quite low. Consequently, infrared absorptions of Copper Phthalocyanine Blue are rarely observed in paint spectra. As discussed previously, Figure 5.18B,C show infrared spectra of two dark blue nonmetallic automotive monocoats with the same color, but they have noticeably different intensities for the Prussian Blue cyano absorption. The difference in concentration of Prussian Blue is actually greater than might be inferred from the two spectra, as a logarithmic scale is involved and weak peaks appear stronger than in a linear scale. How can two paints have the same color if there is much more pigment present in one of them? A good guess would be that there is another blue pigment in the paint of Figure 5.18C. In fact, the pigment is Copper Phthalocyanine Blue, but the only manifestation of its presence in the infrared spectrum of Figure 5.18C is a weak sharp absorption at 730 cm^{-1} (sandwiched between the two styrene absorptions at 760 and 700 cm^{-1}). The Raman spectrum of this paint (Figure 5.18D), however, is predominantly that of Copper Phthalocyanine Blue, which produces resonance-enhanced peaks (Palys et al., 1995). Prussian Blue, in contrast, is not a particularly strong Raman scatterer. Its peaks (Figure 5.18F) are seen in the Raman spectrum of the paint with the larger quantity of Prussian Blue (Figure 5.18E), but are barely perceptible in the Raman spectrum with the lesser amount (Figure 5.18D).

Massonnet and Stoecklein (1999b) collected FT–Raman spectra of some yellow, orange, and red organic pigments used in automotive finishes along with spectra of some red paints that contain some of these pigments. De Gelder et al. (2005) also examined automotive paints using FT–Raman spectroscopy and concluded that the Raman peaks of organic pigments in basecoats provide the best discrimination. Raman peaks of rutile, calcite, and barium sulfate in spectra of undercoats were also found useful for characterization. Unlike infrared spectroscopy, however, Raman spectroscopy is not useful for detecting a wide variety of inorganic pigments. In their study of pigments in automotive paint, Suzuki and Carrabba (2001) found that silicates (talc, kaolin, quartz, and others) do not produce significant Raman peaks, even when present in high concentrations.

In three studies, Bell et al. examined the discriminating capabilities of Raman spectroscopy for house paints using a dispersive CCD spectrometer. In the first study (Bell et al., 2005a), the spectra of 51 lilac-colored house paints were collected and compared. The three components producing the strongest Raman peaks of these paints were rutile, Copper Phthalocyanine Blue, and a bluish-violet dioxazine pigment, Pigment Violet 23. In their two other studies, a combination of infrared and Raman spectroscopy was used to study the discrimination between binders used in house paints (Bell et al., 2005b) and between white house paints (Bell et al., 2005c). These studies reiterated the complementary nature of the two techniques.

Kendix et al. (2004) examined some historical (nineteenth century) house paints using a Raman microscope. Some of the inorganic pigments that they identified include calcium carbonate, barium sulfate, lead carbonate, rutile, and anatase. The rutile-to-anatase ratio varied considerably in these older paints, and differences in this ratio were readily detected because the two pigments produce discrete narrow Raman peaks; in contrast, such differences are difficult to observe using infrared spectroscopy—compare Figure 5.13A and 5.13B. Buzzini and Massonnet (2004) analyzed 40 green spray paints using a combination of infrared and dispersive CCD Raman spectroscopy and found that this combination allowed discrimination of most of the 40, although when used alone, infrared spectroscopy provided a higher discrimination power than did Raman spectroscopy. Resonance Raman peaks of two green phthalocyanine pigments, Pigment Green 7 and Pigment Green 36, were identified in some spectra. Buzzini et al. (2006) discussed six case examples involving automotive, household, and spray paints in which the combination of infrared and dispersive CCD Raman spectroscopy also proved useful.

5.5.7 Pyrolysis Gas Chromatography and Pyrolysis Gas Chromatography–Mass Spectrometry

Pyrolysis gas chromatography (PyGC) and pyrolysis gas chromatography in conjunction with mass spectrometry (PyGC–MS) are additional techniques used to examine and compare the organic portion of a cured coating: the binder and additives. In *analytical pyrolysis*, the sample is subjected to a sufficiently high temperature under controlled conditions in an oxygen-free environment so that its constituent molecules break down into smaller fragments. Those molecular fragments (pyrolyzates) are then swept into an analytical instrument, most commonly a gas chromatograph, where they are separated chromatographically and detected by either a flame ionization detector (FID) or a mass selective detector (MSD) upon eluting from the chromatography column. The FID offers a relatively inexpensive, sensitive, and broadly applicable means for the detection of analytes eluting from the column. The data obtained from such a detector can provide a pattern of pyrolyzates that can be used either to identify the class of polymer that is present or to perform a detailed comparison of resin compositions. If a MSD is used in

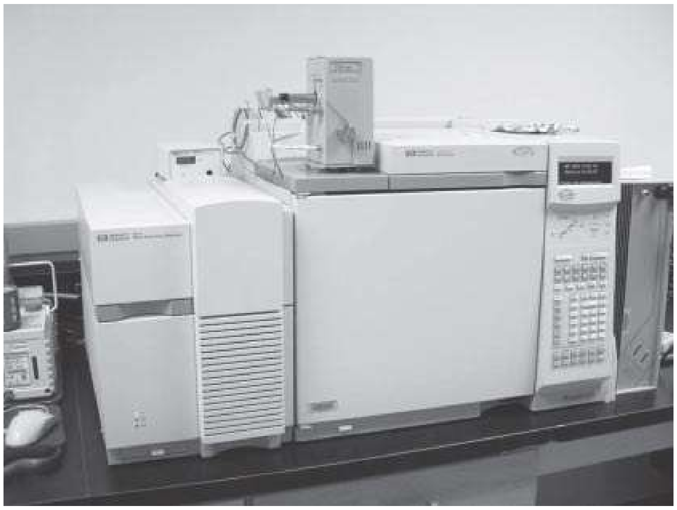

Figure 5.21 Typical pyrolysis gas chromatograph–mass spectrometer (PyGC–MS) equipped with an inductively heated pyrolysis probe, pyrolysis interface, and single capillary column having a mass selective detector (MSD).

place of a FID, much more information can be obtained from the sample. In addition to the pattern information obtained with the FID, the data acquired from a MSD can be used for the potential identification of select compounds of interest. This often aids in identifying the class of polymer that is present. Of course, this additional information comes at a price, as the MSD is considerably more expensive than the FID. A typical PyGC–MS system is shown in Figure 5.21, and an inductively heated pyrolysis probe that inserts into the pyrolysis interface on the gas chromatograph in Figure 5.22.

Why pyrolyze the sample in the first place? Paint binders consist of high-molecular-weight polymers that are simply too large to volatize and introduce into a gas chromatograph. Even if the resin is soluble in a volatile organic solvent such as acetone or chloroform, the polymer chains are still of extremely high molecular weight and cannot be volatized when injected into the standard injection port of a gas chromatograph. By thermally fragmenting the polymers, pyrolyzates are produced that are amenable to volatization and chromatographic separation in a vaporous state. The technique requires little sample (on the order of 10 to 20 μg) and can be performed on most solid polymeric specimens. One may ask why gel permeation chromatography or high-pressure liquid chromatography could not be used, to avoid fragmentation of the polymers. These techniques may be suitable in polymer laboratories but are less so when dealing with real-world forensic samples. First, they require the sample to be in solution, a feat unachievable for a majority of cured, cross-linked paint resins. Second, these techniques typically require a substantial quantity of polymer, which is not available from each layer of most multilayered forensic paint specimens. Finally, these techniques do not

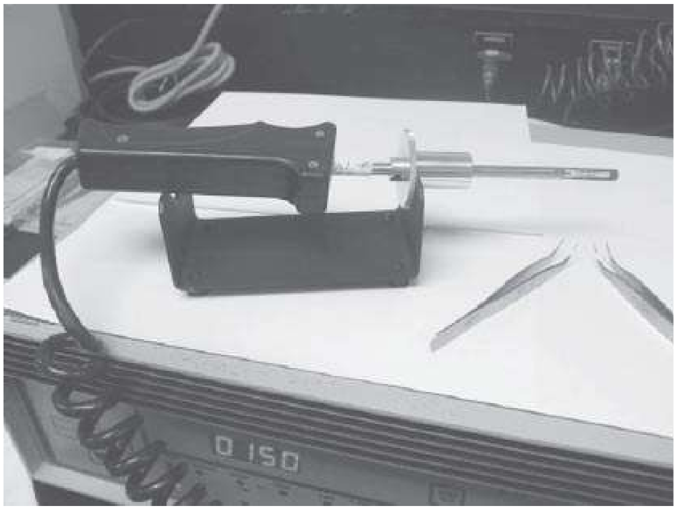

Figure 5.22 Chemical Data Systems inductively heated pyrolysis probe which inserts into the pyrolysis interface mounted on the gas chromatograph injection port.

provide chromatographic characteristics that are specific to the component polymeric species and hence do not maximize discrimination.

Perhaps the more pressing question is why use PyGC or PyGC–MS at all? As we have seen earlier, infrared spectroscopy is a versatile nondestructive technique available to most laboratories which provides spectral data that permit both discrimination and classification of most binders. Furthermore, often it simultaneously provides spectral data characterizing the major inorganic constituents in a paint. Demonstrating its discriminating capabilities, Contos and Ryland (2001) reported 97.5% of the 80 possible pairs of similarly colored *Reference Collection of Automotive Paints* (Collaborative Testing Services, 1989) finish coats having no decorative flake were discriminated by infrared spectroscopy. Edmondstone et al. (2004) reported on a discrimination study of 260 randomly acquired automotive paints, yielding 32,670 possible pairs for comparison. Samples were initially discriminated by microscopical comparison of their finish coat colors, which differentiated all but 28 of the 32,670 possible pairs. Following attenuated total reflectance (ATR) FT–IR microspectroscopic analyses of these 28 indistinguishable pairs' clearcoats, only two pair could not be distinguished. Subsequent transmission FT–IR microspectroscopy of the two pairs' primers served to discriminate another one of the pairs. Obviously, the technique is capable of substantial discrimination.

Infrared spectroscopy does have limitations in the analysis of coating samples. The first is in the discrimination of binders consisting of a mixture of similarly structured polymers. This occurs in the acrylics and alkyds used in automotive paints, in the acrylics used in architectural latex paints, and in the alkyds used in architectural oil-based paints and household spray paints. When performing infrared spectroscopy on a paint resin, the spectrum is comprised of the resulting absorptions of each component in the polymeric mixture, overlaid upon one another. Detection of characteristics arising from minor components in the mixture is difficult to achieve when the spectra of the individual polymers are all similar to one another. PyGC, on the other hand, employs a chromatographic separation of the pyrolyzates generated from the mixture, permitting detection of characteristics produced from even small amounts of organic components. One such example can be seen in Figures 5.23 and 5.24. The transmission infrared spectra of two different acrylic–melamine enamel automotive finish coats are displayed in Figure 5.23.

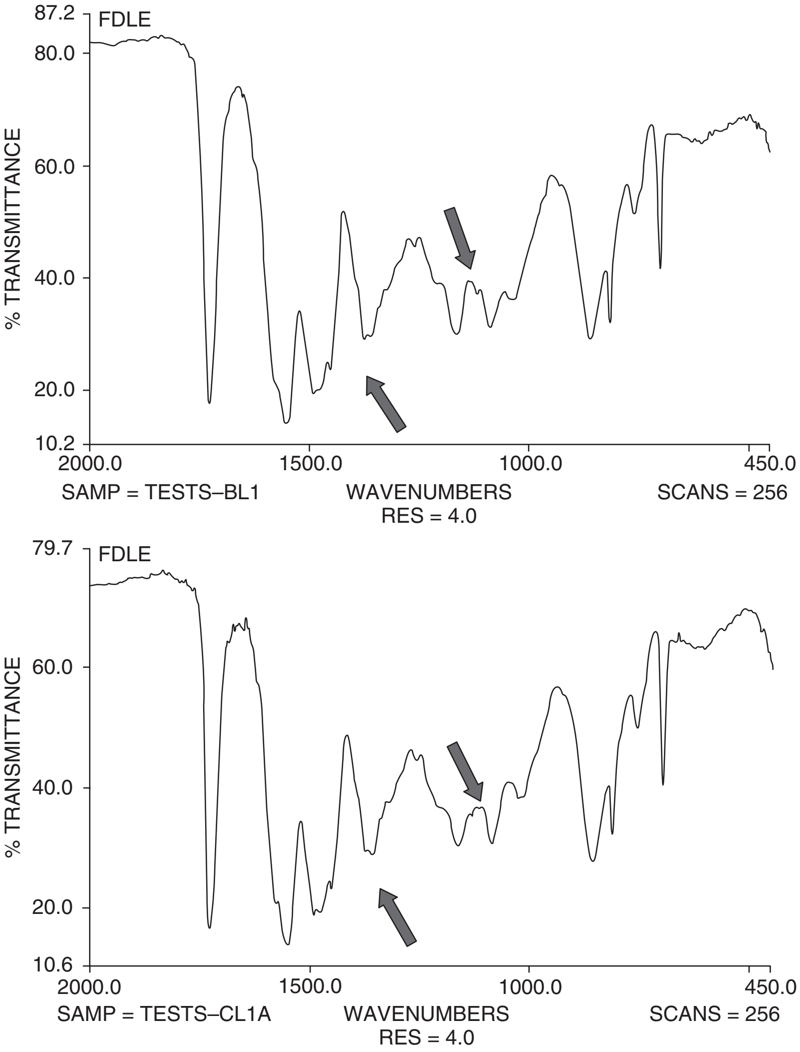

Figure 5.23 Transmission infrared spectra of two different acrylic–melamine enamel automotive finish coats that have the same color. The arrows point out reproducible spectral differences.

They are quite similar, with small differences noted by the arrows. The pyrograms of the same two are displayed in Figure 5.24, presenting no difficulty in differentiating the two. Another example is given in Figures 5.25 and 5.26, these also being two automotive acrylic–melamine enamel finish coats produced by the same paint manufacturer in the same year in the same color. Again, the pyrograms demonstrate little difficulty in differentiating the two. As mentioned previously, the same type of situation can occur in the alkyd enamels, where the same diacid and polyol are used as precursors, but the drying oils are different. Burke et al. (1985), Fukuda (1985), Cassista and Sandercock (1994), and Ryland (1995) all have noted the capacity of PyGC to detect binder variations that go unnoticed, or are at least questionable, when samples are analyzed by infrared spectroscopy alone.

The second major limitation of infrared spectroscopy is a consequence of one of the aforementioned benefits of the technique. In paints that are heavily loaded with extender or coloring pigment, the absorptions resulting from the binder are masked by those that are detected simultaneously from the pigment (Figure 5.19H). Thus, little is present to discriminate between differing binders when performing comparisons. These include some automotive primers, flat architectural paints, flat household spray paints, and some low-gloss maintenance finishes. On the other hand, PyGC in effect performs an automatic extraction prior to analysis. Upon pyrolysis

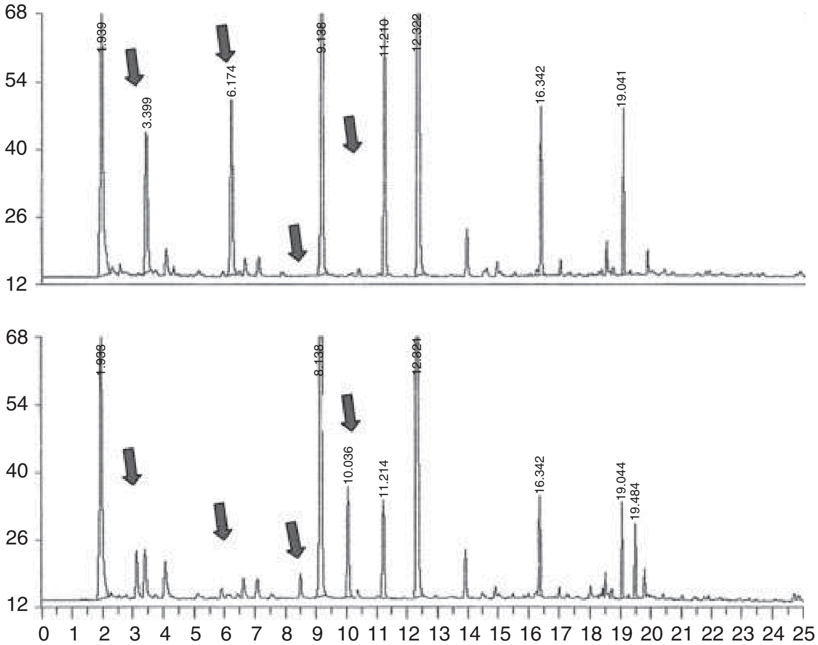

Figure 5.24 Pyrograms of the same two finishes whose infrared spectra are depicted in Figure 5.23. Arrows indicate obvious reproducible differences between the two. The X axis is labeled in minutes and the Y axis is relative intensity.

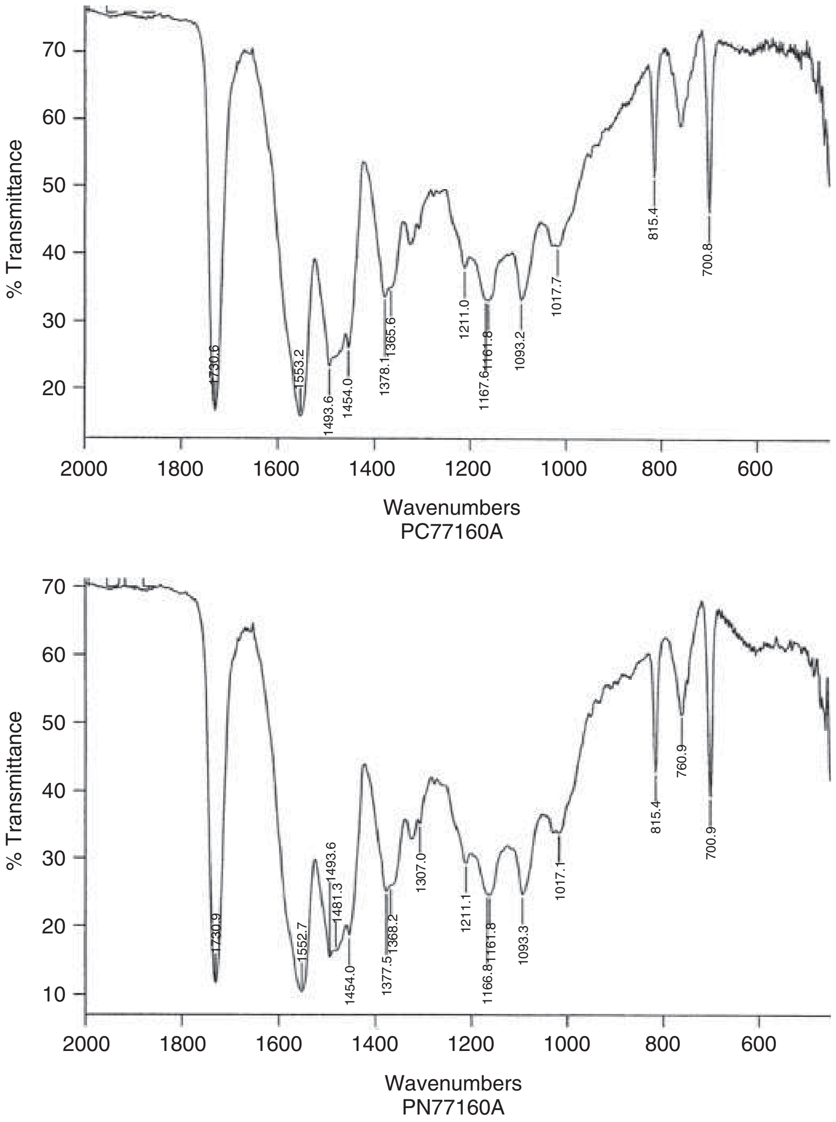

Figure 5.25 Transmission infrared spectra of two different acrylic–melamine enamel automotive finishes having the same color. Both were manufactured by PPG for use on the same year vehicle line; however, one is a low-solvent-emission formulation while the other is not.

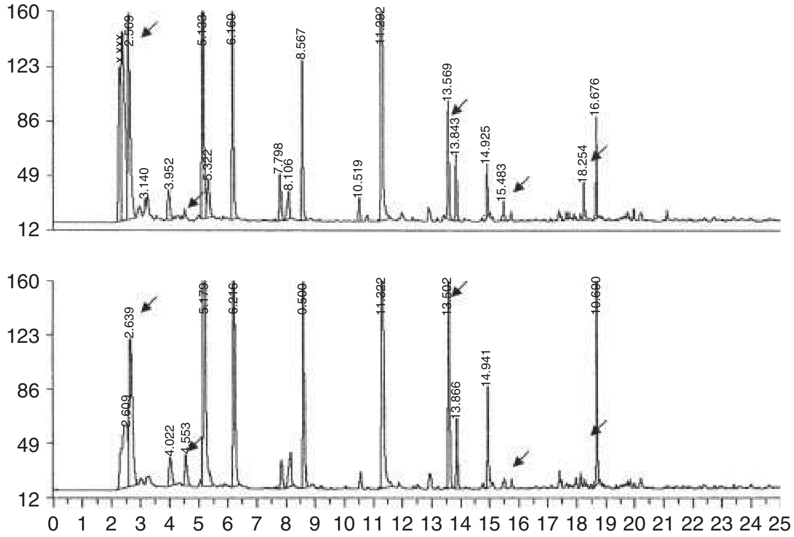

Figure 5.26 Pyrograms of the same two finishes whose infrared spectra are depicted in Figure 5.25. Arrows indicate some of the reproducible differences between the two finishes. The X axis is labeled in minutes and the Y axis is relative intensity.

at an appropriate temperature, the organic portion of the paint is volatilized, fragmented, and passed onto the gas chromatograph while the inorganic portion of the paint remains behind, effectively eliminating any inorganic interference.

It becomes apparent that unlike many other types of trace evidence, paint presents some complex analytical chemistry challenges that demand the use of several analytical techniques to properly characterize its composition. PyGC offers distinct advantages over infrared spectroscopy, but is not without its limitations as well. It requires approximately 10 to 20 μg of sample, on the order of 10 times more than is required for an infrared microscope. In the case of smeared architectural exemplars or very tiny multilayered automotive paint fragments, which are increasingly becoming the norm, PyGC is simply not an option. Furthermore, it is destructive and the organic portion of the sample cannot be reclaimed for further analytical examinations. The inorganic portion of the sample can be reclaimed; however, it has suffered potential deterioration due to exposure to high temperatures. It is much more time consuming than infrared spectroscopy, as the typical run takes approximately 30 to 45 min plus an additional 15 min or more to clear out the column between analyses. It has rather poor long-term reproducibility, making it a poor choice for databasing. However, the discrimination benefits are compelling given certain types of specimens, particularly when used in a comparative mode.

A variety of pyrolysis accessories are available on the market, including inductively heated filament, Curie point, and furnace. Their design, benefits, and shortfalls are discussed by Wampler (2006). In the inductively heated filament type, the sample is placed either on a platinum ribbon probe or in a quartz tube contained within a platinum filament coil probe and inserted into a heated interface attached to the injection port of the gas chromatograph. Heat is supplied by electrical resistance in the platinum element with both the rate of heating and the final temperature being controlled by the electrical power supplied by the control module. Sophisticated electrical circuitry is used to control the rate of heating reproducibly, and the final temperature is calibrated for each pyrolysis probe. Upon completion of analysis, the sample is heated to a very high temperature to clean the platinum ribbon or the quartz tube. An overwhelming majority of instruments used in forensic laboratories in the United States are of the inductively heated filament type, such as that depicted in Figure 5.22. In the Curie point type, the sample is either wedged into a crimped metal alloy wire or placed inside a neatly folded metallic foil pack. The metal carrier along with the sample is then placed into the pyrolysis probe, which houses a radio-frequency (RF) coil. When power is supplied to the RF coil, the metallic carrier quickly heats to its Curie point temperature, that temperature being determined by the metal alloy being used in the carrier. A number of alloys are available, providing an array of fixed final temperatures. Upon completion of analysis, the Curie point wire is either cleaned of any residual inorganic material or the Curie point foil is simple removed and discarded. In the furnace type, the sample is placed in a carrier which drops into an inductively heated furnace held at an adjustable final pyrolysis temperature. Following analysis, the residual sample containing any inorganic residue is removed with the carrier.

No matter what the design, the basic principles of effective, reproducible pyrolysis are the same. The specimen must be small enough to be pyrolyzed completely, yet large enough to provide sufficient pyrolyzates to be detected by the analytical instrument following pyrolysis. Precision will suffer if the samples are not pyrolyzed completely, since differing quantities of pyrolyzates will be produced from run to run. The specimen must be heated to the final pyrolysis temperature both quickly and at a reproducible rate. An overwhelming amount of the fragmentation occurs in the first few milliseconds of the run, and the temperatures experienced by the samples must be reproducible at this point as well as at the final temperature, or the pyrolyzates will differ. The final temperature will determine the amount of energy available for cleaving the bonds and obviously must be reproducible as well. It is typically minimized so that higher-molecular-weight pyrolyzates are produced, being more characteristic of their polymeric source, while being kept high enough to assure complete pyrolysis. The latter can be checked quickly by initiating another run following the sample run and its "bakeout" to see if any additional material is eluted. If pyrolysis was complete, no material should be eluted. Most paint resins are pyrolyzed in the range 650 to 750°C. In addition, small thin specimens are more desirable than large thick specimens. The outer surface of the specimen is exposed to the pyrolysis heat first and this is where the initial pyrolyzates are formed. If the specimen is thick, the inner portion will remain cooler and the newly formed species will have a tendency to condense on these surfaces and form secondary pyrolysis species, which tend to be much less reproducible than the initial pyrolyzate species. It is best that the energy sweep rapidly through sample to prevent the formation of temperature gradients. Finally, the placement of the samples must be reproducible such that they are exposed to the same thermal environment every time. This is quite critical in the inductively heated filament pyrolysis probes.

These concepts should be kept in mind when preparing paint samples for PyGC analysis. Samples should be single-layer peels if possible, not thick, intact layered fragments. Replicate analyses are essential if one is to establish precision and then use that range of variation to judge if one material produces a pyrolysis pattern similar to that of another. Precision will change from one type of resin to another. For example, pyrograms of acrylic resins are much more reproducible than those of alkyds. Samples should also be of approximately the same mass. This can usually be deduced from size as long as the morphology of the specimens is similar.

The data produced are in the form of the familiar chromatogram, with eluted material producing peaks displayed on an x-axis of time. The y-axis is proportional to the quantity of material eluted, although this is reflected more accurately in individual peak areas. The chromatograms are typically quite complex, as can be seen in the Figure 5.27 flame ionization–detected pyrogram of an automotive acrylic–melamine binder. The *thermal fragmentation* of one polymeric species will produce numerous characteristic pyrolyzate fragments, not just one peak. The concept is akin to electron impact mass spectrometry, where a beam of electrons breaks the molecule apart into fragments which are then recorded by the mass spectrometer's analyzer. This fragmentation pattern is quite characteristic of the parent molecule and is often thought of as being a definitive identification of the material. In PyGC the polymer is fragmented by thermal energy and the resulting fragments (*pyrolyzates*) are quite characteristic of the parent polymer. It is not simply gas chromatography with a liquid injection of a solution of intact molecules; it is a fragmentation process that imparts uniqueness to the pattern. This pattern (the pyrogram) can be used for comparison (i.e., to ascertain if one paint is like the other); or it can be used for classification (i.e., to ascertain what kind of material it is). To achieve the latter, the analyst must have pyrograms collected from the full array of materials of interest since the classification is based on pattern similarity to a known material, not primary molecular information. This shortfall of PyGC is often overcome by incorporating a mass spectrometer as the detector. Primary molecular information can then be acquired for the pyrolyzates, which leads to pyrolyzate identification and subsequently, to polymer classification. This is not to say that PyGC–MS is more discriminating than PyGC when comparing two resins, merely that it will produce additional information that often permits classification (identification) of the polymer. Remember, pyrolysis of one polymeric species does not just produce one fragment. It produces many fragments, as demonstrated in the pyrogram of pure polystyrene in Figure 5.28. One would be hard pressed to find another material that produces the same thermal fragmentation pattern.

In using PyGC–MS to classify polymers, it is important to keep in mind that the monomer used to construct the polymer is not necessarily what is detected following pyrolysis. Some copolymers, such as the acrylates, fragment primarily back to their monomeric form. An example can be seen in the Figure 5.29 pyrogram of an acrylic–melamine enamel automotive finish (Figure 5.1). The various acrylates and styrene used to construct the binder are readily apparent. On the other hand,

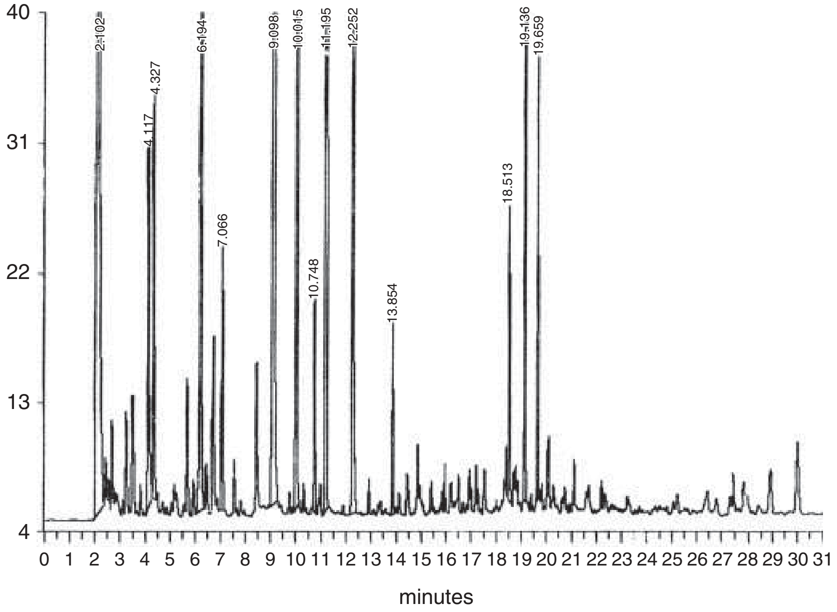

Figure 5.27 Pyrogram of an automotive acrylic–melamine enamel analyzed on a pyrolysis gas chromatograph equipped with an inductively heated pyrolysis probe, a 25-m medium-bore free fatty acid phase (FFAP) fused-silica capillary column and a flame ionization detector.

there is no indication of the presence of melamine other than the presence of *n*-butanol, a cleavage fragment from the butylated melamine cross-linker. Another example can be found in the pyrolysis of poly(vinylacetate). Vinyl acetate is not the major pyrolyzate detected, but instead it is acetic acid. One can deduce why that pyrolysis fragment is produced in the degradation of poly(vinylacetate), but the classification is not straightforward. In other instances, the deduction is not nearly as easily understood. Pyrolysis of an isophthalate-based polyester resin yields benzene, toluene, and benzoic acid as primary pyrolyzates, with no intact indication of the *meta*-substituted aromatic monomer. This point has been demonstrated by numerous authors, such as Challinor (2001, 2006) and Learner (2004).

In an effort to solve this problem to some degree, Challinor (1995) reviewed previous work he published using *tetra*-methylammonium hydroxide (TMAH) in a methanol solution to stabilize pyrolysis frgments as soon as they form (Challinor, 1989, 1991a,b). A small quantity of the solution is placed on the sample prior to introducing it into the pyrolysis chamber, and upon pyrolysis the reagent forms the stabilized methyl esters of the fragile pyrolysis fragments most characteristic of the mother polymer's precursors. For example, this approach permits recognition of the intact methyl esters of the diacids and polyols used to form polyester and alkyd resins as well as the fatty acids comprising the various drying oils used to cross-link alkyd resins. This information is difficult to come by using infrared

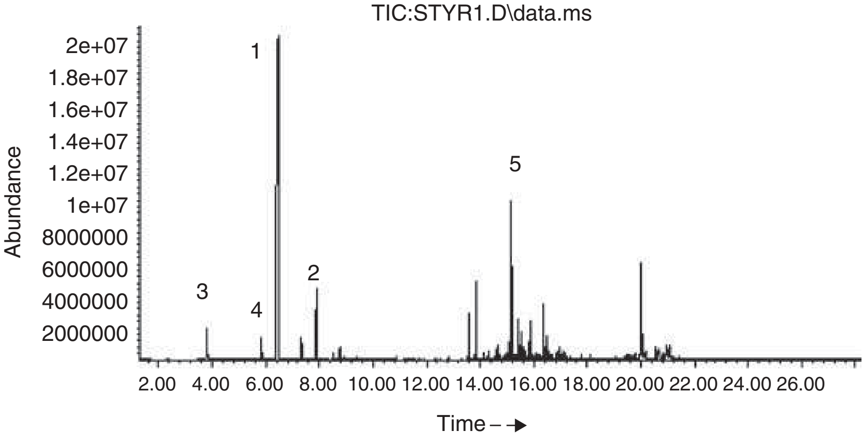

Figure 5.28 Pyrogram of pure polystyrene pyrolyzed at 750°C in an inductively heated pyrolysis probe and chromatographed on a PyGC–MS system using a 30-m narrow-bore 5% phenyl methyl siloxane fused-silica capillary column. The pyrogram consists of many other pyrolyzates in addition to the styrene peak (1), including methyl styrene (2), toluene (3) and ethyl benzene (4). Some of the longer retention time pyrolyzates are dimmers and trimers of these pyrolyzates (5).

spectroscopy alone, especially in the case of polyester resins employing a mixture of ortho-, iso-, and tere- substituted phthalates or alkyd resins formed from similar diacids and polyols but different drying oils.

5.5.8 Elemental Analysis Methods

The two yellow nonmetallic automotive monocoats with spectra depicted in Figure 5.13C and D appear to contain rutile, but both actually contain a mixture of rutile and Nickel Titanate, a yellow inorganic pigment closely related to rutile (Suzuki and McDermot, 2006). Nickel Titanate and other titanate pigments are formed by partial replacement of titanium in the rutile crystal lattice by two other transition metals (Hackman, 1988). For Nickel Titanate (which has the formula $20TiO_2 \cdot Sb_2O_5 \cdot NiO$), antimony and nickel are used. The elements present in some other titanate pigments are apparent from their names: Chromium Antimony Titanate, Nickel Niobium Titanate, Chromium Niobium Titanate, and Manganese Antimony Titanate. The titanates and rutile all have similar infrared spectra, so titanate absorptions in paint spectra can easily be mistaken for those of rutile. Elemental analysis, used in conjunction with infrared spectroscopy, is therefore required to detect and differentiate members of this pigment family.

Based on the infrared data of Figure 5.19H, calcium from the calcium carbonate and silicon from the silicate should be present. If silicates other than silica are present, other silicate elements might be observed, such as aluminum from kaolin, magnesium from talc, or aluminum and potassium from mica. Titanium

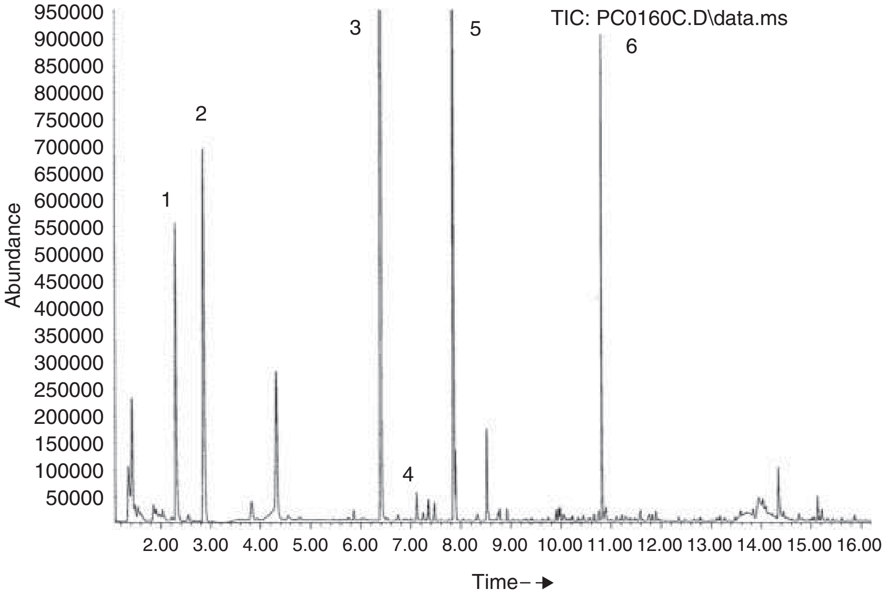

Figure 5.29 PyGC–MS program of an acrylic–melamine enamel automotive finish. The major pyrolyzate peaks are (1) *n*-butanol, (2) methyl methacrylate, (3) styrene, (4) 2-hydroxyethyl acrylate, (5) *n*-butyl methacrylate, and (6) 2-ethylhexyl acrylate.

dioxide (Figure 5.13A,B) appears to be present in this paint, but the additional low-frequency absorptions of calcium carbonate and the silicate render interpretation of this region ambiguous at best. Other oxides, such as zinc oxide and antimony oxide, might be used in house paint, and they also produce broad absorptions in this region.

As noted previously, the detectors used on infrared microscopes have cutoff energies that deny the analyst important information about inorganic pigments (both coloring and extender) in the region below 700 cm^{-1}. Elemental analysis aids in counteracting this limitation by providing comparative data related to the elements comprising these extender pigments. For example, if a paint contains a mixture of kaolin, hydrous silica, and mica, detection of the silica and mica will be difficult, if not impossible, using an infrared microscope. However, elemental analysis will not only reveal differences in the quantities of kaolin and silica present based on the relative intensities of the silicon and aluminum signals, but will also reveal characteristics of the mica based on the potassium signal.

Elemental analyses therefore serve to confirm, augment, and clarify the data obtained by infrared spectroscopy and other methods. Further information about the paint may be provided by the detection of elements not revealed by other methods. When such elements are found, however, the lack of definitive corresponding molecular structural information necessarily limits the degree to which they can be attributed to specific compounds. Since paint data are typically used in a comparative mode, this limitation is not usually a serious issue. At the same time, the presence of some elements can be revealing when trying to determine the type of coating involved, as some elements are found only in certain types. Information about the possible sources of various elements can also aid in interpretation when differences between known and questioned samples occur, as may arise from contaminants or from actual differences. Some knowledge of the possible origins of paint elements is therefore essential, as this can have a bearing on the nature and strengths of the conclusions that can be drawn. In the following discussion, only elements with atomic numbers of 9 (fluorine) or higher are considered since lower atomic number elements are of more limited value for forensic paint examinations.

Detectable elements in paint may arise from binders, although they are much more prevalent in pigments and additives. Except for acrylic silanes and silicones, which contain silicon, other common binders do not contain detectable elements that are an integral component of the polymer structure (although such elements might be present in a binder formulation as a catalyst or other additive). In contrast, virtually all inorganic pigments (both coloring and extender) and many organic pigments contain elements that can be detected. Tabulations of these materials can be found in sources such as Morgans (1990) and Thornton (2002). Some of the elements that are observed for a particular pigment, however, are actually from compounds closely associated with their formulation and not from the pigment itself. Two of the most common paint pigments, rutile and Copper Phthalocyanine Blue, illustrate this point.

The surface of a rutile (TiO_2) particle is capped with hydroxyl groups. In the presence of water and oxygen and with exposure to ultraviolet light, the hydroxyl groups catalyze the production of free radicals, which can lead to degradation of the binder (Morgans, 1990). Individual particles of rutile intended for outdoor use (including all automotive paints) are therefore coated (encapsulated). The most common encapsulating agent is silica (SiO_2), but alumina (Al_2O_3) or zirconia (ZrO_2) may also be used. In addition, rutile is usually treated with alumina to prevent flocculation (the formation of pigment agglomerates or clusters). Therefore, silicon and aluminum will usually be observed along with titanium when analyzing outdoor paints containing rutile. Rutile that is encapsulated with zirconia may also contain tin oxide (Tyler, 2000), and titanium, zirconium, and tin were found in a zirconia-encapsulated rutile product that was analyzed by XRF (Suzuki and McDermot, 2006). Some formulations of lead chromate pigments are also silica-encapsulated to minimize reactions of the chromate anion, an oxidizing agent.

As noted, Copper Phthalocyanine Blue can exist in four different crystal forms, and the most common formulation (known as Pigment Blue 15 : 2) is stabilized to prevent both polymorphic conversion and flocculation. The exact nature of the stabilization process is proprietary, but it usually includes incorporation of chlorinated and sulfonated forms of this pigment (Lewis, 1995). Consequently, in addition to copper, chlorine and sulfur are usually found in paints containing Copper Phthalocyanine Blue. The green phthalocyanine pigments, Pigment Green 7 and Pigment Green 36, have the same structures as Copper Phthalocyanine Blue (Figure 5.2) except that the aromatic rings are highly chlorinated, or highly chlorinated and brominated. These pigments are normally not treated, so paints containing Pigment Green 7 or Pigment Green 36 will contain copper and chlorine, or copper, chlorine, and bromine, respectively.

Encapsulation coatings, antiflocculating agents, and stabilizers are often bound intimately to the pigment particles. Some pigments, however, might also contain small quantities of a diluting agent. These may be added to adjust the tinctorial strength, which is done to ensure that the hue of a particular pigment lot falls within a specified tolerance range. Barium sulfate is commonly used for this, and since it is also a common extender pigment and antisettling additive, its levels in paint can span a very wide concentration range.

Metallic and pearlescent pigments, which are common in automotive finishes, always contain one or more detectable elements. The most common metallic pigment consists of aluminum flakes. Aluminum flakes used in aqueous dispersion basecoats are treated with phosphate surfactants (Bentley, 2001), so phosphorus may be detected in addition to aluminum. Gold-bronze metallic pigments have a composition ranging from pure copper to pure zinc or various alloys (brass) of the two, and these produce various hues (Thornton, 2002). Stainless steel metallic pigments contain chromium, nickel, molybdenum, and manganese, in addition to iron. Zinc metallic pigments may contain traces of iron, lead, and cadmium, and nickel metal pigments may contain traces of iron and sulfur.

Mica-based pearlescent pigments may include laminates of rutile, ferric oxide, chromium oxide, or zirconia, so in addition to silicon and other elements of mica, titanium, iron, zirconium, or a combination may be observed. Some pearlescent pigments are based on bismuth oxychloride, so bismuth and chlorine may be observed in addition to elements of the laminates. The elements of mica (muscovite, the most common form) include potassium, aluminum, and silicon, but micas often include several other minerals, and elements of these might also be present. Chromatic shift interference pigments consist of a core low-refracting dielectric material with a thin coating of a semitransparent highly reflecting layer. The core may transmit light or it may be reflective. Core materials include silicon dioxide, aluminum oxide, magnesium fluoride, and aluminum, while outer coatings include aluminum, chromium, or even molybdenum sulfide, thus expanding the list of detectable elements even further.

Driers are used to catalyze reactions involved in cross-linking between double bonds of a substance, such as unsaturated fatty acids. Elements of driers may thus be found in paints with binders such as alkyds, which use the fatty acids found in drying oils. Driers are probably the source of the greatest variety of elements in paint, and some of the elements that have been used or are currently used include (in order of increasing atomic number) potassium, calcium, vanadium, chromium, manganese, iron, cobalt, zinc, zirconium, cadmium, tin, barium, lanthanum, cerium, and lead. Other transition metals probably have also been used. Unfortunately, their concentrations are quite low, typically 0.1% by weight or less. Some of the elements cited do not function as driers by themselves but do so only when used with others, as discussed by Morgans (1990) and Thornton (2002).

Contaminants that might be introduced during the paint manufacturing process are another possible source of elements. Pigment dispersion techniques rely on various grinding media, which have included steel or porcelain balls, glass or zirconia beads, sand, and other objects (Bentley, 2001). Such media require periodic replacement due to wear, and their fragments are introduced into the paint. Although this might be expected to result in a very minor component, highly variable levels of zirconium were found in a number of automotive paints of various colors using XRF analysis (Suzuki and Marshall, 1998). Various sources of this element were thus considered. Most of the paints with zirconium contained little or no rutile, and tin was not found, so a zirconia-encapsulated rutile product was ruled out. Driers were another possible source of zirconium, but for an automotive paint, they would probably be found in an alkyd; all of the paints involved had either acrylic melamine enamel or acrylic lacquer binders. In addition, if zirconium were used as a drier, it would usually occur with cobalt (Morgans, 1990), but cobalt was not detected. Wear from zirconia dispersion beads thus became the prime suspect, and the paint manufacturers who were contacted concurred that this was the most likely source.

Preservatives, particularly fungicides used in exterior architectural coatings, are another source of elements. Some of the elements that have been used, or are currently used, in fungicides and other preservatives include chlorine, copper, chromium, zinc, iodine, mercury, and other heavy metals, although the recent trend has been to avoid heavy metals due to health concerns.

Silicates (talc, kaolin, quartz, diatomaceous silica, synthetic silica, mica, bentonite, asbestine, wollastonite, montmorillonite, etc.) are very common extender pigments. As noted, extender pigments are frequently used in low-luster finishes and automotive undercoats since in addition to lowering the cost of the paint by adding bulk, they serve as flatting or delustering agents. Synthetic silica and treated montmorillonite are used as thickening agents for paints with low viscosities, such as automotive clearcoats, where they help to keep the wet paint of a thick layer on a vertical surface from dripping; talc is used as an antisettling agent to help keep other pigments suspended in the liquid medium; mica is used in pearlescent paints; silica is used to encapsulate certain pigments; silicates are minor components of some inorganic pigments, such as ochres (which are mainly ferric oxide or hydrous ferric oxide); and so on.

Not surprisingly, silicon is a common element in paint. In some cases, the presence of other elements associated with the source of the silicon might serve to clarify its origin. Magnesium, for example, is present in talc [$Mg_3Si_4O_{10}(OH)_2$], and aluminum in kaolin [$Al_2Si_2O_5(OH)_4$]. For some minerals, including talc, cation substitution is a common occurrence. Magnesium (Mg^{2+}) and iron (Fe^{2+}) can occur interchangeably in these minerals since they have the same charge and the same size. Talc may thus contain iron even though its nominal formula may not indicate this. Since the function of extender pigments is not usually affected adversely by minor impurities, there are probably many other impurities in them as well, although they may be present in levels below the detection limits of the method used for analysis.

Silicon in paint can thus originate from any number of components, including binders (silanes), extender pigments, mica-containing pearlescent pigments, impurities in other pigments, encapsulating agents, products of dispersion bead wear, and additives serving a wide variety of functions. This element can therefore occur in virtually any type of coating, including every layer of an automotive finish. It is also a very common element in soil, so the condition of the sample(s) must be evaluated carefully to decide if it may be present as a result of contamination from material that is *external* to the paint.

Although detectable elements are less common for organic pigments, they do occur. Twenty-nine pigments were examined in four studies of organic pigments used in automotive paint (Suzuki and Marshall, 1997, 1998; Suzuki, 1996a,b). Ten contain chlorine, one chlorine and sulfur, one chlorine and nickel, one sulfur, one nickel, one copper, and one fluorine. Chlorine is therefore relatively common (12 of 29) in organic pigments. This element, which is usually present as a substituent on an aromatic ring, is serving at least two functions (Lewis, 1995). It imparts insolubility to a pigment and is used to modify the color of a base structure by changing its chromophore (as evidenced by the effects of chlorine and bromine substitution on the phthalocyanine ring).

Extensive lists of pigments are presented in compilations by Crown (1968) and Eastaugh et al. (2004), and the detectable elements of pigments can be obtained from their chemical formulas. It should be noted, however, that components that might be associated with some of their formulations are not included, and these may contain additional elements. Nolan and Keeley (1979) also compiled classes of paint elemental constituents based on their frequency encountered in SEM–EDS analysis of random samples.

Scanning Electron Microscopy/Energy-Dispersive X-Ray Spectrometry

SEM–EDS is the main method used in U.S. forensic science laboratories for elemental analysis of paint evidence (Henson and Jergovich, 2001; Ryland et al., 2006). This instrument can be used both as an imaging device with very high magnifications (greater than 100,000× in some cases) and as a means to obtain elemental profiles for selected areas of a sample. Elemental data for individual layers of multilayered paint specimens can be collected sequentially and the high magnifications of the microscope permit the analysis of very small samples. The range of elements that can be detected simultaneously with an EDS analysis is extremely wide and, with an appropriate detector window, all elements of atomic number six (carbon) or higher can be identified. The detector has a broad dynamic range and the technique may be nondestructive to the specimen. It is, however, limited to a minimum detection level of approximately 0.1% by weight for median atomic weight elements.

The terminology *energy-dispersive* x-ray spectrometry is intended to distinguish it from a related technique, wavelength-dispersive x-ray spectrometry (WDS). These two methods employ different mechanisms to measure x-ray energies. Both may be mounted on a scanning electron microscope. In a WDS instrument, x-rays are separated by wavelength using a crystal, which serves as a diffraction grating. Multielemental analysis is much slower than with EDS and higher beam currents are required to produce enough x-rays to be detected; however, sensitivity is increased tenfold and the improved resolution provides a solution to the peak overlap problems experienced in the EDS analysis of some elements. In an EDS system, a solid-state silicon–lithium (SiLi) detector not only counts the number of x-ray photons that arrive at the detector, but also measures their energies. An x-ray photon in the SiLi detector loses its energy in a series of steps, after which a number of electrons are promoted from the valence band to the conduction band. This number is proportional to the original energy of the photon and is measured electronically. The total number of photon counts for a given characteristic energy is proportional to the quantity of the element present in the sample.

For an SEM–EDS analysis, samples that are nonconductive, such as paint, must be coated with a film of a conducting material, usually carbon. This prevents the material from developing a negative charge during analysis, which would cause it to repel further electrons. Most instruments are operated in a high vacuum, as atmospheric gases not only impede the electron beam, but also absorb the lower-energy x-rays. Variable-pressure SEMs are available at a nominal increase in cost and make it unnecessary to coat a specimen. By leaking just a small amount of air or selected gas into the specimen chamber, the charge on the uncoated sample is dissipated. This does, however, result in a slightly diminished signal from the sample and a slight loss in control of the analytical area due to spread of the electron beam diameter.

The operation of an SEM–EDS instrument is not unlike that of a cathode ray tube (CRT) monitor. An electron beam, directed and focused by electric and magnetic fields, traces a raster pattern on the selected rectangular area of interest of the sample. An image is formed by measuring the intensity of electrons ejected by the sample as the beam traverses the selected area. Analogous to an optical image, lighter areas of the image correspond to a higher flux of electrons. Two detectors located at different positions relative to the sample are used to collect *secondary electrons* and *backscattered electrons*. Secondary electrons result from ionization of atoms of the sample near the surface and have relatively low energies. Backscattered electrons, in contrast, have nearly the same energy as that of the incident beam. Backscattering results when an incident electron passes very close to an atomic nucleus, and the strong electric field of the nucleus swings the much lighter electron around, sending it in the opposite direction. Backscattering increases with atomic number, so an SEM image based on backscattered electrons provides a compositional map, with lighter areas representing domains comprised of heavier atoms. Imaging of paint specimens in this mode provides a considerable advantage over other elemental analysis techniques since it provides a rapid visual assessment of heterogeneity (Figure 5.30) and aids in selecting appropriate analytical sizes and locations.

The energy of the electrons striking the sample is set by the analyst, and this is an important parameter for x-ray analysis because it determines the energy range of the x-rays that can be produced. An electron that is accelerated through a voltage

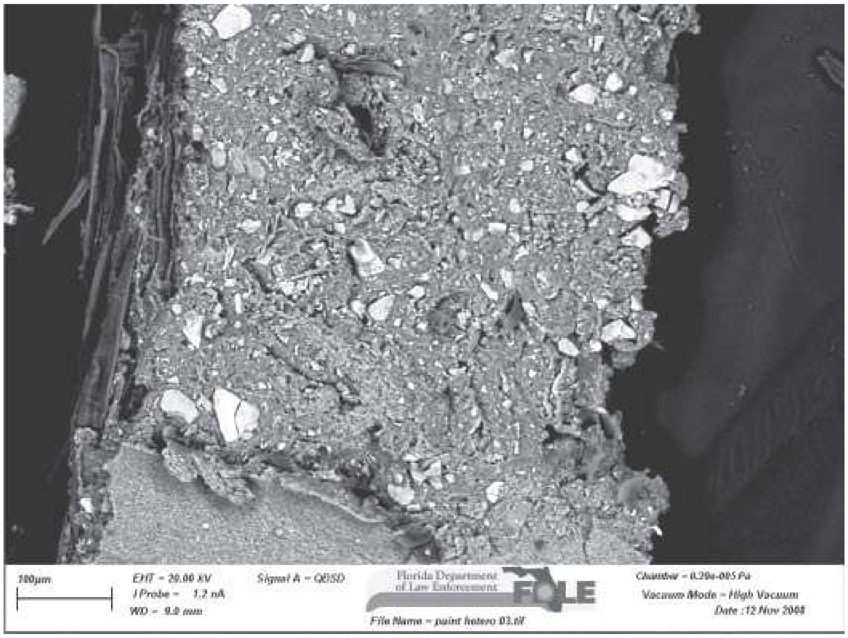

Figure 5.30 SEM backscatter image of an architectural paint layer exposed by peeling back the layer above. Note the visual rendering of the large extender pigment grain heterogeneity.

potential of 20,000 V (20 kV) acquires an energy of 20,000 electron volts (20 keV). The range of possible voltages for an SEM–EDS is typically between 1 and 40 kV, although voltages higher than 25 kV are generally not used with paint because of the excessive heat that may be generated and the increased depth of the beam's penetration. X-ray energies are also measured in electron volts, and the range of an SEM–EDS x-ray detector is typically from 0.1 to 40 keV. The flux of x-rays reaching the detector is controlled by the electron beam current.

When the electron beam strikes the sample, x-rays are generated by two different mechanisms. The first involves a process known as *bremsstrahlung*, a German word meaning "braking radiation." When a charged particle accelerates (meaning that it increases or decreases its velocity or is deflected from a straight path), it can produce radiation. In the case of the electron beam of an SEM–EDS instrument, electrons are decelerated and their paths deflected as they interact with the sample. The bremsstrahlung x-rays that result from this have a continuum of energies, ranging from zero to that of the incident electron, and this is manifested as a broad background (having a roughly skewed semicircular shape) in all EDS spectra. Bremsstrahlung backgrounds can be seen in the EDS spectra of Figures 5.31A,B, 5.32C, and 5.33. In general, a matrix having a low average atomic number composition, such as an automotive finish coat, will have a much higher bremsstrahlung

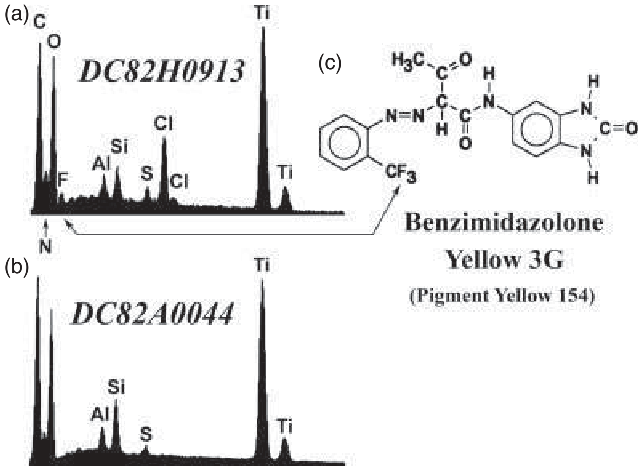

Figure 5.31 (A) SEM–EDS spectrum of a yellow nonmetallic automotive monocoat, DC82H0913, which contains rutile and Benzimidazolone Yellow 3G. (B) SEM–EDS spectrum of a white automotive monocoat, DC82A0044, which contains rutile. An electron beam voltage of 20 kV was used for both SEM–EDS analyses. (C) Structure of Benzimidazolone Yellow 3G. [From Suzuki and Marshall (1997); copyright © ASTM International; reprinted with permission.]

background than that having a relatively higher average atomic number composition, such as an automotive primer or architectural paint. This decreases the elemental minimum detection levels in the former.

The second type of x-rays produced is referred to as *characteristic x-rays*. This process begins with ionization of the atoms of the sample, and the electrons of the incident beam must have sufficient energies to remove electrons from one of the first three shells of an atom. Atomic shells are designated as K, L, or M, corresponding to electrons in orbitals of the first three principal quantum numbers. The K shell is comprised of 1s electrons, the L shell of 2s and 2p electrons, and the M shell of 3s, 3p, and 3d electrons. Following removal of an electron from one of these shells, a "hole" is created in the atom and an electron from a higher-energy shell decays to fill the vacancy and bring the atom to a lower-energy state. In so doing, it emits an x-ray having an energy equal to the energy difference between the two shells. The 2s and 2p electrons have different energies, as do the 3s, 3p, and 3d electrons. Consequently, the L to K shell decays represented by 2s $\rightarrow$ 1s and 2p $\rightarrow$ 1s have different energies, as do the various transitions from the M to the L shell. There is therefore more than one x-ray energy produced when an electron is removed from a particular shell. The x-rays emitted by electrons decaying to fill the K shell are referred to as the K series, with analogous definitions for the L and M series.

X-rays produced by individual transitions of a series may have very similar energies that cannot be resolved by the SiLi detector, In the case of heavier elements, separate peaks occur for the K series and for the L series. The lower-energy peak

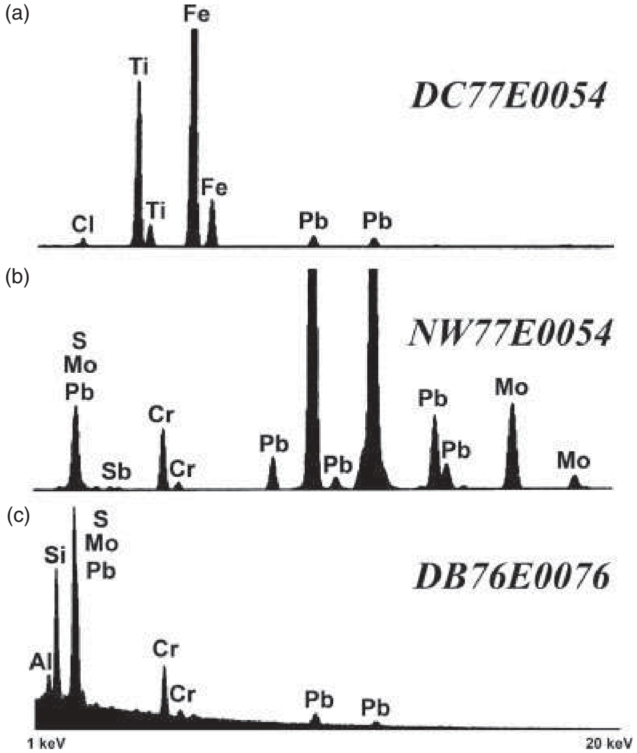

Figure 5.32 XRF and SEM–EDS spectra of three red nonmetallic automotive monocoats depicted between 1 and 20 keV. The two XRF spectra were obtained using a tin secondary target. (A) XRF spectrum of DC77E0054; the infrared spectrum of this paint is shown in Figure 19f. (B) XRF spectrum of NW77E0054; the infrared spectrum of this paint is shown in Figure 19A. (C) SEM–EDS spectrum of DB76E0076, acquired using an electron beam voltage of 22 kV. (From Suzuki and Marshall, 1998; copyright © ASTM International; reprinted with permission.)

of the *K* series (known as K_α) typically has an intensity roughly six times that of the second peak (known as K_β), whereas the two peaks of the *L* series have more nearly equal intensities. All of the peaks observed in Figure 5.31A are from the *K* series, and the two *L*-series peaks of lead can be seen in Figure 5.32C (compare to the same peaks of lead obtained by XRF in Figure 5.32A and 5.32B). *M* series transitions cannot be resolved and only one peak is observed. Experienced paint examiners can usually distinguish between peaks of the three series, but instrument software may also be used to aid in this task, as they include markers that indicate the positions and relative intensities of *K*, *L*, and *M* peaks for each element.

As the atomic number increases, the inner-shell electrons become more tightly bound to the nucleus, lowering their energies, and the energy differences between adjacent shells increase. There is thus a regular increase in the energies of each series as the atomic number increases (*Mosley's law*); the differences in energies between peaks of adjacent elements also increases with atomic number. Consequently, there is no element for which *K*, *L*, and *M* peaks can all be observed in the range 0.1 to 40 keV, although for some elements, two series are observed. As noted, electron energies much above 25 keV may damage the paint specimens or penetrate too deeply into its surface, so only about half of the detectable range is used in any case. Even with this more limited range, detection of a single element present in sufficient concentration is not a problem, and this is also often true when several elements are present. However, due to overlap of *K*, *L*, and *M* peaks in the low-energy region, certain combinations of elements may be difficult to determine.

This is illustrated by the EDS spectrum (Figure 5.32C) of a red nonmetallic automotive monocoat that contains Molybdate Orange ($PbCrO_4 \cdot xPbMoO_4 \cdot yPbSO_4$), for which peaks of lead, chromium, molybdenum, and sulfur are expected. *L*-series peaks of lead and *K*-series peaks of chromium are observed, but the *K*-series peaks of molybdenum are not (compare to Figure 5.32B, which is the XRF spectrum of another paint that contains Molybdate Orange). *L*-series peaks of molybdenum do occur in the range of the instrument, but in this case they are buried in the strongest peak of Figure 5.32C. This peak represents an overlap of the *M* peak of lead, the *L* peaks of molybdenum, and the *K* peaks of sulfur, so the presence of molybdenum and sulfur cannot be definitely established in this case. This confluence of *K*, *L*, and *M* peaks of three elements of a single pigment is *extremely* unusual (if not unique), but it does illustrate the interpretation difficulties that overlapping peaks can present.

Secondary electrons arise from atoms very near the surface of the sample, and backscattered electrons originate from somewhat greater depths. SEM images thus reveal predominately surface characteristics. The electron beam, however, penetrates beyond these relatively shallow depths. As it interacts with the sample, it begins to lose its directionality and individual electron paths become more random. Consequently, the morphology of the volume from which x-rays originate is not a well-defined narrow cylinder, but rather, has a hemispherical or a teardrop shape, depending on beam energy, sample composition, and x-ray energies.

The interactions of the electrons with the sample also affect their energies. An electron beam energy of 22 keV was used to acquire the EDS spectrum of Figure 5.32C, and this energy is sufficient to ionize a *K*-shell electron of molybdenum, but no *K*-series peaks of this element are observed. The sample contains much more lead than molybdenum, and lead requires even less energy to ionize its *L*-shell electrons, yet the lead *L*-series peaks are quite weak. The intensities of the lead *L*-series peaks and the molybdenum *K*-series peaks can be increased slightly by employing a higher beam excitation potential, ideally 2.5 times the minimum required to ionize the electrons in those shells. However, as mentioned earlier, there are drawbacks to doing this.

The preferential excitement of the lower-atomic-number elements of a sample is a characteristic feature of an SEM–EDS analysis. It results from the numerous interactions that the electron beam experiences as it travels through the sample, constantly losing energy. Consequently, most of these electrons can only excite elements of lower atomic numbers. Low-atomic-number elements are thus readily observed in SEM–EDS spectra, as may be seen for the *K*-series peaks of carbon, nitrogen, oxygen, and fluorine of Figure 5.31A (which is the spectrum of a yellow nonmetallic automotive monocoat). Some of the carbon is from the thin coating used to make the sample conductive, while the fluorine is from a yellow organic pigment, Benzimidazolone Yellow 3G (Figure 5.31C).

Figure 5.31B is the SEM–EDS spectrum of a white nonmetallic automotive monocoat that contains a large quantity of rutile, and in addition to the titanium peak of rutile, the silicon and aluminum from the encapsulation coating and the antiflocculating agent, respectively, are also observed. Similar ratios of these three elements are observed for the spectrum of the yellow monocoat (Figure 5.31A), which also contains a large amount of rutile. It is also instructive to note that the aluminum and silicon are not originating from a clay extender pigment that also might be suspected as a source from the SEM–EDS spectrum alone.

Because of the differences in the sample domains that are responsible for image formation and generation of x-rays, SEM–EDS analysts should always remember the caveat, "What you see is not always what you get." It is especially important to remember when analyzing individual layers of an intact multilayered paint.

Four different methods have been used to obtain SEM–EDS data for individual layers of multilayered paint samples, each with its merits, limitations, and degree of skill (and patience) required to obtain optimal results. In the first, each paint layer is excised and analyzed separately. This requires a fair amount of sample preparation, but it is the only method for which analysts can be assured that elements of adjacent layers are not being observed. A second procedure involves what is referred to as the stair-step method since after preparation of a specimen, it has a steplike appearance when viewed from the side. This method also requires a considerable amount of sample preparation, as beginning with what will be the bottom layer, a rectangular portion consisting of all of the overlaying layers is removed, exposing a portion of the bottom layer. An adjacent rectangle is then removed, exposing a portion of the next-to-bottom layer, and so on. The sample is positioned in the SEM–EDS instrument, so that all of the exposed layers are perpendicular to the beam. Analysis of very thin layers may yield x-rays from a underlying layers, and it may help to reduce electron beam energies to determine if this is the case.

The last two methods involve cutting a cross section through the layers and situating this cut surface perpendicular to the SEM–EDS beam. This can be done without any further sample preparation, or the sample can be embedded in a medium that hardens and the surface of the cross section polished. Care must be taken in either case to ensure that the surface is situated horizontally and that the cut was made perpendicular to the layers, so that electrons traveling straight down do not encounter another layer. Embedding and polishing require considerable sample preparation, but this process also yields the most reproducible results regarding relative ratios of EDS x-ray peaks. Embedding, however, makes removal of the sample difficult, if not impossible, should there be a need for further analysis. For either type of cross-sectional analysis, a major concern is the possibility that the EDS analysis volume may extend into an adjacent layer or into two adjacent layers. To minimize this possibility, long narrow rastering areas with edges as far from

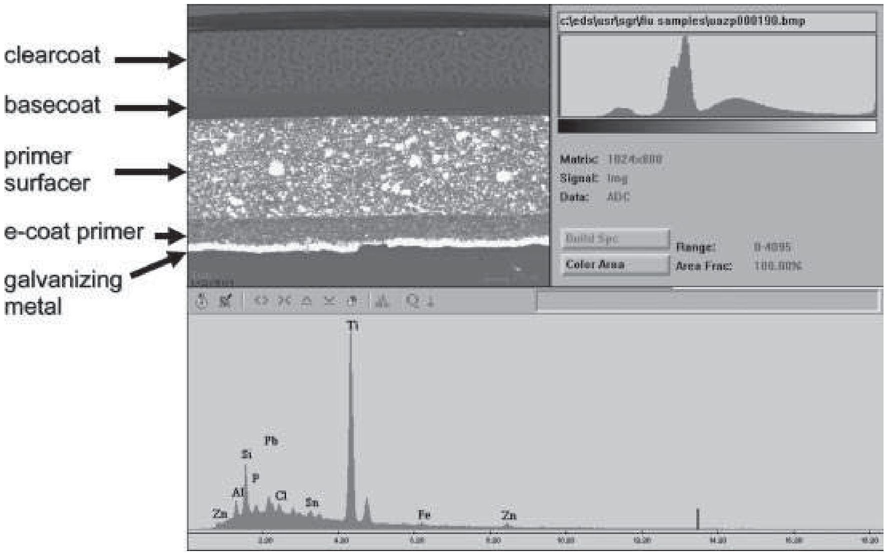

c:\eds\usr\sgr\lu samples\uazp000190.bmp

Matrix: 1024x800
Signal: Img
Data: ADC

Range: 0-4095
Area Frac: 100.00%

Figure 5.33 SEM–EDS analysis of the base primer in an automotive paint cross section block cut by microtomy. The red box superimposed on the e-coat primer defines the visual boundaries of the analytical area. *(See insert for color representation.)*

the adjacent layers as practical are used, as depicted in Figure 5.33. Regardless of which of the four methods is used, it is important that both questioned and known samples be subjected to the same procedures.

As mentioned, the analytical area is typically chosen by scanning the electron beam over a large defined area to average out the effects of heterogeneity, a constant concern in paint samples. The relative success of this endeavor is assessed by comparing spectra collected from one area of a layer to another of the same layer. If the analytical area is large enough to average out the lack of precision introduced by heterogeneity, the spectra will appear like one another. The SEM offers another mode of sample collection called the *spot mode*. This approach is quite valuable in analyzing particulate material in the matrix, such as metallic flake and interference pigment flake (the rich elemental and morphological information contained in these pigments was described previously). The beam is targeted on the particle of interest and is held stationary, as seen in Figure 5.34. The excitation voltage is typically reduced in an attempt to keep the excited volume within the boundaries of the target particle, keeping in mind the necessity to maintain enough energy to effectively excite the elements of interest. In addition to analyzing effect pigment flake in paints, the same method may be used to enhance the inferred identification of other pigments in the paint. If a large area is scanned and barium, sulfur, magnesium, aluminum, and silicon are detected, this merely informs the analyst that these elements are present in the layer in their respective relative ratios. Should spot mode be used in addition to this technique, it can be demonstrated that the barium and sulfur are associated with one particulate population, magnesium and silicon

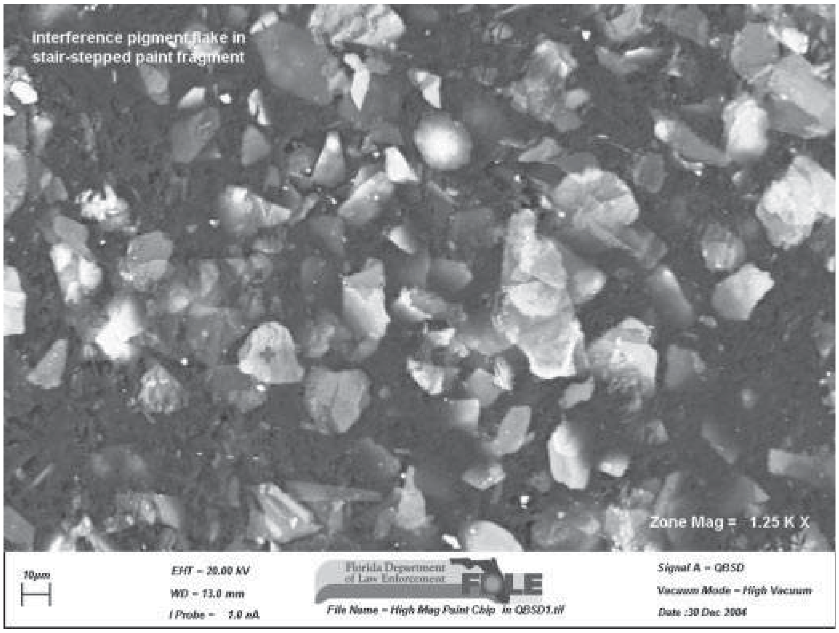
interference pigment flake in
stair-stepped paint fragment
Zone Mag = 1.25 K X
10μm
EHT = 20.00 kV
WD = 13.0 mm
I Probe = 1.0 nA
Florida Department of Law Enforcement
FDLE
File Name = High Mag Paint Chip in QBSD1.tif
Signal A = QBSD
Vacuum Mode = High Vacuum
Date :30 Dec 2004

Figure 5.34 SEM–EDS spot analysis of interference pigment flake in situ. The flake is visualized by viewing the exposed layer in the backscatter mode. The red + indicates the beam target spot. *(See insert for color representation.)*

are associated with another, and aluminum and silicon are associated with a third. This significantly strengthens the deduction that the paint layer contains barium sulfate, talc, and clay extender pigments. Although these are typically recognized in infrared spectra collected down to at least 450 cm^{-1}, it may not be apparent in situations where there is a significant quantity of titanium dioxide present, and one of the extenders is present in a low concentration relative to the others.

More information about SEM–EDS instrumentation is given by Goldstein et al. (1992), while the preparation of paint samples for SEM–EDS analysis is discussed in more detail by Ryland et al. (2006). An introduction to paint analysis using an SEM–EDS analysis is given by Wilson and Judd (1973), and Henson and Jergovich (2001) provide an extensive review of this topic. General guidelines for such analyses are outlined in the *Standard Guide for Using Scanning Electron Microscopy/X-Ray Spectrometry in Forensic Paint Examination* (Scientific Working Group for Materials Analysis, 2002).

X-Ray Fluorescence Spectrometry XRF spectrometry is the second most frequently used method in the United States for the elemental analysis of paint evidence (Henson and Jergovich, 2001; Ryland et al., 2006). *Fluorescence* refers to a process whereby atoms or molecules absorb electromagnetic radiation, then reemit radiation with lower energies. In XRF spectrometry, as opposed to SEM–EDS, x-rays are used to ionize atoms instead of electrons. It should be noted, however, that the emission process is the same for both techniques. XRF spectra might therefore be expected to mirror EDS spectra, but this is generally far from the case. As we saw, the forte of an EDS analysis is the detection of lighter elements. XRF spectrometry, in contrast, is better suited for the identification of heavier elements, so the two are complementary. There is no need to coat samples for an XRF analysis, and if only the heavier elements are sought, sample chambers do not have to be evacuated. XRF instruments with the lowest detection limits (tens of parts per million or lower in some cases) require considerably larger samples than are required for an SEM–EDS analysis, while micro-XRF systems, which have higher detection limits, allow smaller samples. There are concerns regarding the increased beam penetration experienced when using x-rays as opposed to electrons; consequently, the current approach for individual layer analysis favors preparation of individual layer peels. When comparing spectra of layers excised from small multilayered samples, one must also be concerned with the inherent imprecision introduced by critical depth effects experienced in samples of different thicknesses (Howden et al., 1977).

As with an EDS analysis, the energies of the x-rays used for excitation must be sufficient to ionize K, L, or M electrons. With x-rays, however, the most efficient means to do this is with an x-ray energy that is slightly greater than the minimum required (known as an *absorption edge*), and as x-ray energies increase beyond this value, the excitation efficiencies decrease.

X-rays for XRF excitation are generated in an x-ray tube. The tube consists of a metal anode or target, typically rhodium, housed in a vacuum. A voltage between 1 and 50 or 60 kV is used to accelerate a beam of electrons toward the target. Analogous to what occurs in an SEM–EDS instrument, bremsstrahlung and characteristic x-rays of rhodium are produced. Depending on the voltage used, Rhodium L peaks, or K and L peaks, are generated. All XRF instruments may use x-rays from the tube for excitation. Since the distribution of x-ray energies from the tube includes the broad bremsstrahlung continuum, this serves to excite a wide range of elements.

For some instruments, an alternative mode of excitation is possible using *secondary targets*, which consist of several elements spanning a range of atomic numbers. X-rays from the tube are directed toward a selected secondary target, causing it to emit K-series peaks. These x-rays are then used to excite the sample, and since K_α is more intense than K_β, this amounts to using mostly monochromatic radiation for excitation. Since this is the most effective means to excite a particular element or small group of elements with similar atomic numbers, this mode provides the lowest detection limits, particularly for heavier elements. To achieve this for a wide range of elements, however, analyses using several different secondary targets are required. The intensities of x-rays produced by secondary targets are limited, so relatively large samples and long collection times are normally required for paint samples. Secondary targets are not available on micro-XRF spectrometers, which use collimators or focus optics to produce beam diameters between 300 and 10 μm. Micro-XRF instruments with selectable filters that improve the detection limits for selected elemental ranges are now available. Appropriate thin elemental (metal) films are positioned selectively in front of the x-ray tube to reduce the bremsstrahlung continuum and/or characteristic tube x-ray lines striking the sample. This in turn reduces the x-ray scatter background or eliminates the scattered x-ray tube lines emanating from the specimen and still leaves enough higher continuum x-rays to excite the elements of interest. Like secondary target instruments, however, much longer data collection times are required.

The complementary nature of an EDS and XRF analysis can be seen by comparing Figure 5.32B,C, which are spectra of two red automotive monocoats (NW77E0054 and DB76E0076) that contain Molybdate Orange. For the XRF analysis (Figure 5.32B) a tin secondary target ($K_\alpha = 25.2$ keV) was used, while an electron beam energy of 22 keV was used to obtain the EDS spectrum (Figure 5.32C). DB76E0076, which was analyzed by EDS, contains more Molybdate Orange than NW77E0054, judging from relative infrared absorption intensities. Note that even with an XRF analysis (Figure 5.32B), the presence of the sulfur of Molybdate Orange cannot be determined because of the overlap discussed earlier. In these situations additional techniques may be employed, such as wavelength-dispersive x-ray spectrometry (WDS).

The XRF spectrum of a second automotive paint (DC77E0054) that has the same color as NW77E0054 is shown in Figure 5.32A. The infrared spectra of these two paints (Figure 5.19A,F) were discussed previously and it is informative to see how the infrared and elemental analysis data correlate. NW77E0054 (Figures 5.19A and 5.32B) contains Molybdate Orange and Quinacridone Red Y. Quinacridone Red Y does not contain any detectable elements, but peaks of all of the elements of Molybdate Orange, except for sulfur, are observed in the XRF spectrum of NW77E0054. DC77E0054 contains rutile, ferric oxide, Quinacridone Red Y, and Benzimidazolone Orange, and its XRF spectrum (Figure 5.32A) has peaks of titanium (rutile), iron (ferric oxide), and chlorine (Benzimidazolone Orange). A small quantity of lead is also observed, and analyses using other secondary targets and longer collection times detected very small amounts of molybdenum and chromium. The presence of a small quantity of Molybdate Orange is thus indicated, although its absorptions (Figure 5.19K) are not observed in the infrared spectrum of this paint (Figure 5.19F).

The XRF spectra have reiterated the value of elemental analysis when used in conjunction with infrared spectroscopy and other methods, and the additional information about the paint that may be obtained. They were obtained, however, with secondary targets using large pristine samples, and the data were collected overnight. Micro-XRF systems are currently better suited for most case samples, since much smaller specimens can be analyzed with sensitivities intermediate between those of an EDS and a secondary target XRF analysis. These sensitivities are still sufficient to detect low concentrations of tinting pigments or even drier metals having concentrations on the order of 0.005% by weight. Other innovations have continued to allow smaller samples to be subjected to XRF analyses. A micro-XRF attachment for an SEM–EDS instrument, for example, has recently been introduced. It consists of an x-ray tube using poly capillary focus optics to form an excitation beam diameter of 40 μm. Fluorescent x-rays are detected using the SiLi detector of the SEM–EDS system. There are, however, concerns for background scatter levels on samples mounted on carbon substrates and for isolation of layers given the increased penetration of x-rays versus electrons from the SEM.

Fiber optic x-ray guide tubes are also used on a new micro-XRF spectrometer that produces spot sizes as small as 10 μm. Kanngiesser et al. (2005) demonstrated the feasibility of analyzing individual layers of a multilayered paint sample using a micro-XRF instrument with a 10-μm spot size, but this required use of synchrotron radiation. This new system focuses x-rays from a conventional x-ray tube without the need for a synchrotron.

Jenkins (1999) provides more information about XRF spectrometry theory and instrumentation. Howden et al. (1977) examined the discrimination of white, red, and green single-layer household paints using XRF spectrometry. All of the red and green paints could be distinguished by this means and the white paints were divided into nine groups. Haag (1977) used XRF spectrometry to study elemental profiles of automotive paint chips. Zieba-Palus and Borusiewicz (2006) examined several multilayered paint samples using a combination of infrared and Raman microscopy and micro-XRF spectrometry. The discrimination of black (Govaert et al., 2001) and red (Govaert and Bernard, 2004) spray paints was examined using a combination of optical microscopy, infrared spectroscopy, and XRF spectrometry.

Other Elemental Analysis Methods Several other elemental analysis techniques have been used for the analysis of paint. Emission spectrography was one of the first such methods used in forensic science, but it has been supplanted by inductively coupled plasma emission spectroscopy (ICP-ES) and its even more sensitive offspring, inductively coupled plasma mass spectrometry (ICP-MS). ICP methods, however, require samples that are soluble—not an easy task for most coatings, which are enamels that contain pigments designed to be insoluble. Neutron activation analysis (NAA), which is gamma-ray emission spectrometry of nuclei in excited states, is probably the most esoteric of the methods that have been applied to paint. NAA requires a neutron source such as a nuclear reactor, and although not destructive, it produces, at least temporarily, a radioactive sample. Like an SEM, an electron microprobe analysis uses an electron beam to produce characteristic x-rays, but this method has few, if any, advantages over SEM–EDS analysis. A proton beam is used in a related technique known as proton-induced x-ray emission or particle-induced x-ray emission (PIXE). This method has lower detection limits for many elements compared to an SEM–EDS analysis, but like an electron microprobe instrument, it lacks the imaging capabilities of an SEM–EDS system. Thornton (2002) provides a more detailed review of these methods as applied to paint.

Two relatively new techniques show some potential as future tools for the examination of paint evidence, as both allow trace levels of a wide variety of elements to be detected. Major issues regarding reproducibility occur with both techniques, however, and they are currently used primarily for qualitative analyses.

Total-reflection x-ray fluorescence spectrometry (TXRF) is a variation of a conventional XRF analysis where a grazing incident angle ($<0.1^\circ$) of an x-ray beam is used for excitation (Klockenkamper et al., 1992). A standing wave is produced, and this results in significantly enhanced (parts per billion) sensitivities. TXRF has been used for several elemental analyses of artists' paints and other pigmented historical or archeological objects (Klockenkamper et al., 2000).

Laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) is an alternative method for detecting trace elements in materials that cannot easily be subjected to a conventional ICP-MS analysis. A pulse from a laser is used to vaporize a sample and create an aerosol, which is swept into the plasma torch of the ICP-MS instrument. The high temperatures of the plasma ($>6000^\circ$C) produce mostly monoatomic ions, so that the mass spectrometer yields elemental analysis data. Hobbs and Almirall (2003) examined the feasibility of using this method for the analysis of automotive paint. Typically, a suite of suspected elements must be selected prior to analysis. While a number of elements were identified at trace levels, their relative ratios varied considerably between results obtained from successive pulses. More stability was found in primer layers; however, the diameter of the ablation crater had to be increased to approximately 200 μm to improve precision. There is a "charred" discolored zone that extends approximately 300 μm beyond this crater, and precision of analysis within this zone has currently not been assessed. This has a practical impact on sampling, as replicate analyses would require a sample much larger than 1 mm^2. Assuming that the precision concerns can be overcome, there is some potential for batch discrimination of automotive primers resulting from the variation of trace elements contained as contaminants in different lots of the minerals used as extender pigments. The same holds true for architectural paint tint bases, although full-thickness 1-mm^2 samples are seldom encountered in casework.

Deconinck et al. (2006) also examined automotive paint using LA-ICP-MS, although their work focused on demonstrating inaccuracies suffered by quadrapole-based ICP-MS instruments, due to spectral interferences during trace element depth profiling of multilayered paint samples. They concluded that a high-resolution sector field–based mass spectrometer would be required to avoid such problems. Like the earlier study by Hobbs and Almirall, this technique was found useful for qualitative analyses, but it did not resolve issues of reproducibility and quantitative analyses on small samples. Although it was stated that four-layered car paints of the same color produced by different manufacturers could be discriminated from one another using the technique proposed, improved discrimination over current methodology was not documented.

5.5.9 Other Methods

X-Ray Diffraction Like emission spectrography, XRD was one of the first methods used for the analysis of inorganic components in paint evidence. The term *x-ray diffraction* actually refers to two related methods. In one, usually referred to as *x-ray crystallography*, a single crystal is analyzed using an incident monochromatic x-ray beam, and the three-dimensional pattern of spots that results from reflection from various crystal planes is analyzed. The planes are formed by the particular arrangement that individual atoms or ions assume in the crystal, and they are not only peculiar to each compound but also differ for each crystal structure of a polymorphic substance. X-ray crystallography is used predominantly for molecular structural determinations, although it has also been used for analytical chemistry.

The technique that is used for the analysis of paint is *powder XRD*, which is applicable to materials that consist of, or contain, crystalline substances where the orientations of individual crystals of the sample are random. To further ensure that the orientations are completely isotropic, the sample is rotated during analysis. Instead of a pattern of spots, x-rays are reflected from the sample in cones with apices at the sample. The cones are characterized by their angles relative to the direction of the incident beam and their intensities. A graph of an XRF powder diffraction pattern, known as a *diffractogram*, is produced with the abscissa representing the angle of reflection (actually 2θ, or twice this angle), and the ordinate depicting the intensities of the reflected x-rays. In principle, all of the major crystalline components of paint can be identified by powder XRD. Some paint binders may exhibit localized regions of crystallinity, but because they lack large-scale regularities in their structures, binders cannot be identified by these means. Reflections from crystal planes arise from x-ray scattering from individual atoms or ions; such scattering is caused by the electrons of the sample, so the intensity of x-ray scattering increases with atomic number. Organic pigments are therefore weak scatterers of x-rays, so powder XRD analyses are limited primarily to inorganic pigments.

Because of its limited scope and past requirements for relatively large samples and long analysis times, powder XRD has not been widely used for forensic paint examinations. Newer instruments, however, are capable of analyzing paint samples as small as 20 μm in reasonable periods of time with essentially no sample preparation. Unlike an elemental analysis, powder XRD provides molecular structural information, and for some paints, it provides more definitive data than either infrared or Raman spectroscopy. Silicates, for example, cannot be identified by Raman spectroscopy, and when large quantities of extender pigments are present in paint (as typically occurs for low-luster house paints; see Figure 5.19H), their strong broad infrared absorptions usually overlap, hampering identification of individual pigments. Powder XRD diffractogram peaks, in contrast, are quite narrow and because there are no contributions from binders or most organic pigments, relatively simple patterns occur for paint. This is illustrated by Figure 5.35A, which is the diffractogram of a paint that contains significant quantities of four common extender pigments (Snider, 1992). Diffractograms of two of these, kaolin and quartz, are shown in Figure 5.35B and 5.35D, respectively, along with that of hydrous ferric oxide (Figure 5.35C). Different polymorphs of a given compound are readily distinguished by powder XRD, so mixtures of such polymorphs can also be identified easily.

Laser Desorption Mass Spectrometry LDMS is a relatively new technique that, unlike its cousin LA-ICP-MS, provides molecular structural information. Both techniques use a pulse from a laser to volatilize and ionize samples, but in LA-ICP-MS, all molecular species are destroyed in the plasma. In LDMS, much less

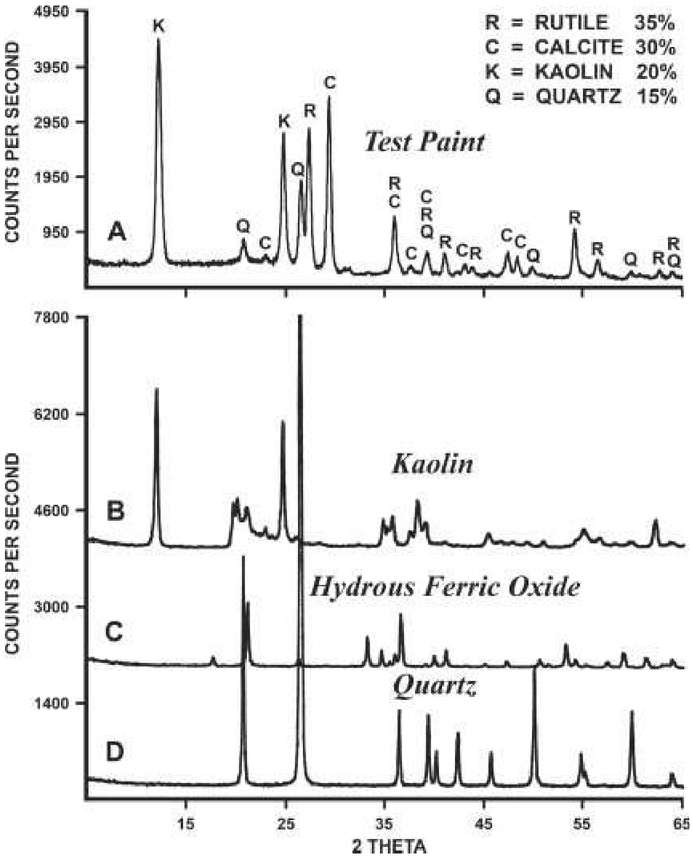

Figure 5.35 Powder XRD diffractograms of a test paint and three common inorganic paint pigments: (A) test paint that contains rutile (35%), calcite (30%), kaolin (20%), and quartz (15%); (B) kaolin; (C) hydrous ferric oxide; (D) quartz. [From Snider (1992); copyright © ASTM International; reprinted with permission.]

vigorous conditions are used to vaporize and ionize paint constituents, which are then introduced directly into a mass spectrometer. No sample preparation is required other than separating individual layers of a multilayered paint if they are to be subjected to analysis.

Stachura et al. (2007) demonstrated the feasibility of using LDMS to identify some organic and inorganic pigments in automotive paints. Positive and negative ions were examined, and both mass spectra were found to be useful since they provided different types of information. Molecular ions of Copper Phthocyanine Blue with expected isotopic ratios were identified in mass spectra obtained for a blue automotive paint. Lesser amounts of molecular ions of the mono- and dichloro-substituted compounds were also observed, from the stabilizers discussed previously. Of all the methods that have been described, this is the only one that has elucidated the nature of these stabilizers.

5.6 EXAMPLES

5.6.1 Example 1

The body of Veronica Hill was found several feet off the side of a little-traveled dirt road in Daytona Beach, Florida. Her clothed body was wrapped in a peach-colored sheet, a floral print white sheet, and an old blanket. The cause of death was determined to be blunt trauma to the head. There was no evidence of sexual assault. Subsequent investigation disclosed that she frequently had sex with men in exchange for money, and several of her recent acquaintances were identified by the clientele in establishments she frequented. Time lines pointed to one primary suspect, David Cartwright. He appeared quite cooperative upon questioning and agreed that the police could search his apartment and garage after obtaining the appropriate paperwork. He acknowledged being with Ms. Hill in the past, but claimed that he was not with her on the night of her death.

Mr. Cartwright's apartment appeared to be in order, with no signs of struggle or evidence of Ms. Hills' suspected presence. Even had such evidence been discovered, it would have been of little value, as Mr. Cartwright already said he had known her and she had previously been to his apartment. In searching the garage, a strong scent of chlorine bleach was noted along with signs of a recent acid wash of the concrete floor. A search for blood resulted in failure. Examination of his 1984 Toyota pickup truck revealed that it was customized both inside and out. It had a relatively new coat of deep purple paint with a distinctive sparkling effect. The interior had the headliner and seats covered with purple imitation fur and there was evidence that the bed area of the truck had at one time also had a similar treatment. The rear of the cab had previously been removed and a truck bed cover was welded to the cab, modifying the pickup truck into somewhat of a sport utility vehicle. The purple faux fur fabric lining was torn from the bed of the truck as well as from the sides of the compartment. The roof's interior still had the fabric present as a custom headliner. Mr. Cartwright explained that he was in the process of refurbishing the interior, and a search failed to reveal the presence of blood stains. Mr. Cartwright shared photos of the interior of the custom vehicle prior to his removal of the appointments and acquaintances related that the work must have been done recently, as the interior was intact just the prior week. He was an avid do-it-yourself customizer as evidenced by the paint and refinish supplies in his garage.

Witness statements and cell phone logs continued to point to Mr. Cartwright and Ms. Hill being together the night of her demise. He held to his claim of not being with her but had no alibi, since he was home alone that night. No murder weapon had been discovered and there was no DNA evidence. Examination of the sheets and blanket in which Ms. Hill's body was wrapped revealed an odd pattern of apparent oversprayed black paint on one of the sheets and the blanket. There were straight lines with no paint on one side and thin overspray on the other. One even had a curve at the top of it, as if something was placed on the fabric and spray-painted. There were two other round spray areas on the blanket, as if a can of spray paint was being tested or the nozzle cleared. These areas were cut out and the adhering surface debris from the remainder of the fabrics was collected and searched stereomicroscopically. This search disclosed the presence of numerous purple synthetic fibers, some appearing to be long and furlike in shape. Comparison of these fibers to various standards taken from the headliner, side compartment liners, and seat covers of Mr. Cartwright's customized truck revealed several corresponding types. There were three different types that were alike generically (acrylic, modacrylic, and polyester) in addition to their microscopical morphology and color. Morphology was compared by polarized light microscopy, fluorescence comparison microscopy, and brightfield comparison microscopy. Color was compared by comparison brightfield microscopy and microspectrophotometry. Of course, one could argue that Ms. Hill had probably been in Mr. Cartwright's truck and could have had some of the fibers on her clothing from previous contact. However, the number of fibers found was striking; and they were found not only on the victim's clothing but on the outer blanket and sheets as well. Also, there was the paint overspray.

Stereomicroscopical examination of the overspray areas on the outer blanket and the floral sheet used to wrap the body revealed tiny droplets of cured paint fused to the fibers comprising the fabrics. The surfaces of the droplets were smooth and tapered down to the shaft of the fibers, indicating that the paint had been applied wet and then cured on the fabric as opposed to being abraded onto the fabric's surface. Examination by polarized light microscopy (PLM) and FT–IR microspectroscopy revealed that the droplets consisted of a black urethane-modified alkyd enamel paint with interference decorative flake present, providing a purple tint. This paint was encased with a clear, colorless coating that was also a urethane-modified alkyd enamel, similar to the underlying black layer in binder chemical characteristics. The binder chemistry and compatible layer system of the two paints suggested a vehicle refinish paint. Refinish paints are particularly valuable evidence, due to the number of manufacturers and the variety of binder types marketed by each manufacturer. Considering the combination of these two factors, there are over 30 differentiable binder formulations for each color offered.

No potential sources of this paint were discovered in Mr. Cartwright's garage; however, there were numerous overspray areas on the concrete walls and cabinetry. One such area consisted of a black urethane-modified alkyd enamel paint containing interference decorative flake, providing a purple tint. It was the same color as the paint on the blanket around the victim's body. Examination of paint samples collected from his 1984 Toyota truck disclosed that the uppermost finish coat consisted of a clear colorless urethane-modified alkyd enamel with either metallic decorative flake or both metallic and pearlescent flake distributed sparsely throughout. This was followed by a black urethane-modified alkyd enamel containing interference decorative flake providing a purple tint. Numerous refinish layers were present below these topcoats, varying in layer structure from one panel of the vehicle to another. Comparison of the top two paints with the paint overspray found on the blanket and the garage wall revealed that they are like one another with respect to their colors, binder characteristics, and pigment characteristics, including the elemental characteristics of the interference decorative flake present in the black basecoat. Stereomicroscopy, comparison brightfield microscopy, comparison polarized light microscopy, infrared spectroscopy, and scanning electron microscopy in conjunction with energy-dispersive x-ray spectroscopy were all used in the comparisons. PyGC was not used, due to the limited quantity of paint present in the overspray transfers.

As with many cases involving trace circumstantial evidence, the two layers of paint overspray discovered on the blanket and sheet wrapped around Ms. Hill's body corresponding to the paint on Mr. Cartwright's vehicle and garage wall was certainly not definitive proof of the association in and of itself. Paint is a mass- and batch-produced material and typically escapes the realm of source individualization. But on the other hand, considering that the paint on the body's wrappings was applied as a wet spray of a compatible two-layer vehicle refinish paint system which coincided with the outermost finish on Mr. Cartwright's vehicle along with the presence of the numerous unusual fibers of varying types found on the body's wrappings that also coincided to the upholstery in Mr. Cartwright's vehicle makes for very compelling evidence of an association.

5.6.2 Example 2

By far the most common type of paint examination encountered in public forensic science laboratories involves vehicle paint transferred in hit-and-run cases. There are two types of investigations. In the first, a victim's clothing and/or possessions are examined for the presence of automotive paint in order to determine the year, make, and model of the striking vehicle. This investigative lead information is provided to authorities to aid in the search for a suspect vehicle. In the second, paint exemplars taken from a suspect vehicle (knowns) are compared to paint transfers found on the victim to determine whether or not the suspect vehicle could have been the source of the paint transferred. Occasionally, the investigations become very challenging, such as in a 2002 "cold case" from a city on the east coast of Florida, where the suspect vehicle of a hit-and-run fatality was a white 1977 GMC van that had been discovered in a salvage yard eight months after the incident occurred.

The suspect had sold the van for parts shortly after the accident, and many accessories were missing (see Figure 5.36). It had been stripped of all interior parts, front-end exterior lamps, lenses, and side rear-view mirrors. All windows had been broken and removed. No fabric marks were found on the vehicle. There was some body damage to the front passenger side of the vehicle, especially the right-front fender and hood; however, other panels on the vehicle were also damaged. The exterior passenger side of the vehicle had been spray painted white by hand, as evidenced by the thin coating where the paint had run down the surface of the door and quarter panel.

The accident scene had pieces of broken amber plastic side marker lens, pieces of broken side-view mirror glass, pieces of broken headlight glass, and a right exterior rear-view mirror assembly strewn about. The victim was injured primarily by blunt trauma to the rear of the head, apparently from being struck by the mirror assembly. Search of the victim's clothing revealed the presence of white paint

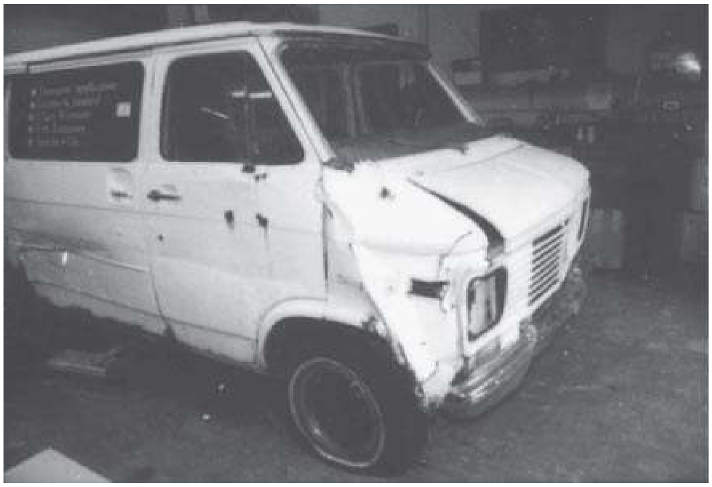

Figure 5.36 The suspect 1977 GMC van recovered from a salvage yard. *(See insert for color representation.)*

deposits fused to the back of the victim's jeans. Numerous borosilicate headlight glass fragments and nontempered window glass fragments were found in the surface debris recovered from the clothing, along with several fragments of multilayered paint fragments having a white topcoat.

Examination of the suspect vehicle revealed rust-encrusted mounting points for an exterior right rear-view mirror, which corresponded in general configuration to the mirror mount found at the scene. However, no detailed characteristics remained at the mounting sites due to the corrosion present. This was associative evidence, but far from definitive. The pieces of amber plastic side maker lens were similar to the red lenses remaining on the rear of the vehicle. Both they and the exterior side-view mirror assembly were consistent with the types used on mid- to late-1970 to early-1980 General Motors (GM) vans and panel trucks. No headlight, mirror, or window glass remained in the vehicle for use as standards to compare to the glass fragments recovered from the victim's garments. Following comparisons by stereomicroscopy, polarized light microscopy, FT–IR microspectroscopy of all layers, PyGC of the finish coats, and SEM–EDS of all layers, it was found that the paint chips fused to the rear of the victim's jeans had two-layer structures corresponding to the finishes on the right front fender and hood of the suspect vehicle. One panel has a three-layer original finish paint system underneath a poorly applied refinish paint (see Figure 5.37), while the other has a two-layer refinish paint system (see Figure 5.38). The two types of corresponding paint, along with the parts found at the scene, provide strong evidence of an association between the victim and the suspect 1977 GMC van.

The forensic examiner's task often does not end with the completion of examinations in the laboratory and authoring a report of the findings. He or she may be called to court to give testimony not only of what was found but also of its

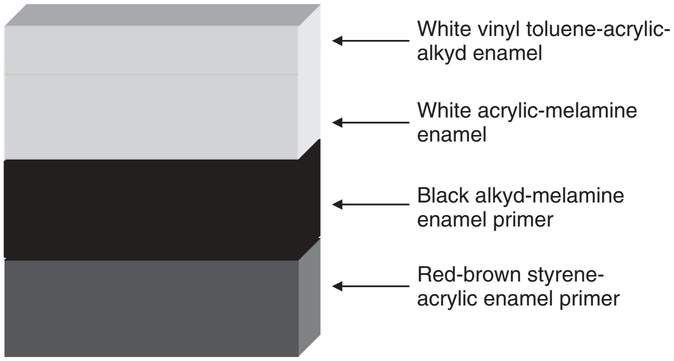
White vinyl toluene-acrylic-alkyd enamel
White acrylic-melamine enamel
Black alkyd-melamine enamel primer
Red-brown styrene-acrylic enamel primer

Figure 5.37 Layer structure of the paint from the passenger side door and quarter panel of the 1977 GMC van. *(See insert for color representation.)*

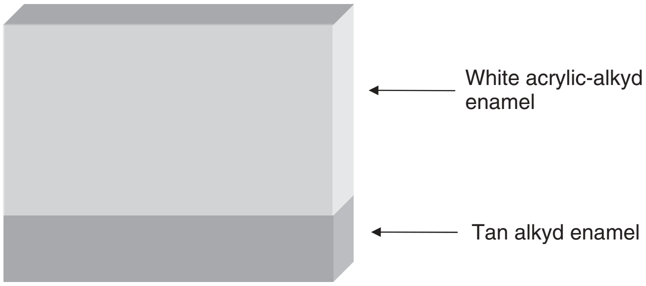
White acrylic-alkyd enamel
Tan alkyd enamel

Figure 5.38 Layer structure of the paint from the right side of the hood of the 1977 GMC van. *(See insert for color representation.)*

significance. Although the work on this case could have stopped at this point, a search of the Paint Data Query (PDQ) International Original Automotive Paint database was undertaken to further evaluate how many other vehicles may have had the same original finish as that on the 1977 GMC van produced in GM's Lordstown, Ohio plant. The original finish paint layer structure consists of a white acrylic–melamine enamel topcoat over a black alkyd–melamine enamel primer-surfacer over a reddish-brown styrene–acrylic enamel primer. There is a thin white vinyl toluene–acrylic–alkyd enamel layer on the surface of the original topcoat. This surface paint binder is typical of a household spray paint. The characteristics of the remaining three original automotive finish layers were entered into the PDQ database and searched. Although there was no record of a finish system exactly like that on the Lordstown, Ohio–manufactured 1977 GMC van in this case, all queries did lead to General Motors products manufactured between 1977 and 1980. A further search of these "hits" removed the possibility of the finish coat being an acrylic lacquer, thus restricting the possibilities to GM trucks and vans, with the exception of several types of cars being manufactured in California. Insistence on the color and chemistry of the base electro-coat primer in the PDQ search, even without the black primer-surfacer, points to 1977 to 1980 GM trucks or vans.

To what end did this additional work bring the examiner? Testimony would surely stress that two types of paint were found on the victim that correspond to the coatings on the front passenger side of the suspect vehicle. One is a refinish and the other is an original finish with a repaint on its surface (determined by their layer structures and chemistry). Furthermore, the binder chemistry of the latter refinish is consistent with a household spray paint. When asked the question, "But how many other vehicles could have an original paint like this common white finish on them?" the examiner can now respond, "In the case of the three-layer original finishes found in the questioned and known samples in this case, a search of an automotive paint database containing over 14,000 original-finish layer systems used over the past 30 years resulted in similar paint layer systems being found only on vehicles produced by General Motors from 1977 to 1980, primarily vans and trucks."

5.6.3 Example 3

Human use of pigments for paintings, decorations, and other ritual purposes has a very long history. The earliest known cave paintings, for example, are well over 30,000 years old. The pigments used for these were predominantly inorganic compounds (especially ferric oxide and hydrous ferric oxide), although a few naturally occurring organic compounds, such as chlorophyll from plants, were also used. The vast majority of synthetic pigments were not available until the early twentieth century, and this is certainly the case for synthetic paint binders. When dealing with art and other painted or pigmented objects, the presence of synthetic pigments and binders is therefore one of the most definitive markers used by art conservation and provenance chemists to indicate either a modern restoration or a forgery. Introduction dates for earlier paint components, however, can also serve to limit the time periods when they could have been used. Although some inorganic pigments have been used for tens of thousands of years, their modern formulations have characteristic signatures that serve to distinguish them from their ground mineral counterparts. Modern pigments, for example, have definite particle size and particle-size distributions, morphologies, specific crystal structures in the case of polymorphic substances, additives, and so on.

The Shroud of Turin There has probably never been an object analyzed by a forensic science laboratory that has generated more controversy than the Shroud of Turin. The shroud is a linen cloth that appears to bear the image of a man. It is believed by some to be the cloth that was placed on Jesus Christ prior to his burial, with the image formed by blood, secretions, or other means. The shroud is kept at the Cathedral of Saint John the Baptist in Turin, Italy.

In 1977, a team of scientists was selected to study the shroud, with this investigation given the designation Shroud of Turin Research Project (STURP). One of the STURP team members was Walter McCrone (1916–2002), a leading microscopist and founder of the McCrone Research Institute and McCrone Associates in Chicago. McCrone's main tool for this investigation was polarized light microscopy, although he confirmed his findings using a variety of other analytical methods. In 1979, based on his analyses, McCrone concluded that there is no blood on the shroud and that the image is produced by a paint consisting of a collagen tempera binder pigmented with red ochre (mainly ferric oxide) and vermilion (mercuric sulfide) (McCrone and Skirius, 1980; McCrone, 1980, 1981). This was a common paint composition during the fourteenth century and McCrone stated that “the Shroud is a beautiful painting by an inspired medieval artist.”

McCrone was the only STURP team member to reach this conclusion, and when other members learned that he disagreed with them, his samples were confiscated and he was removed from the team. When his conclusions became public, he received hate mail and death threats. In 1989, three independent radiocarbon studies of the shroud were conducted at the University of Arizona, Oxford University, and the Swiss Federal Institute of Technology. The data from all three laboratories were consistent and indicated that the shroud originated from the time period 1260 to 1390 (Damon et al., 1989), confirming McCrone’s earlier claim. More details of this case are described in *Judgement Day for the Shroud of Turin* (McCrone, 1999b). In 2000, the American Chemical Society acknowledged McCrone for his contributions to chemical microscopy—specifically citing his studies on the shroud—by honoring him with the American Chemical Society Award in Analytical Chemistry.

The Vinland Map The Vinland map is purported to be a fifteenth-century document that was discovered in 1957. Its authenticity gained some credibility from the fact that at one time it was probably bound together with another ancient document of undisputed authenticity. The map depicts most of the world, including Europe and Asia, Africa, Greenland, and a large island west of Greenland labeled “Vinland” (North America). The map was donated to Yale University, but before accepting it, the university had it examined by two museum curators and a librarian, who pronounced it authentic. Unfortunately, this was done in secret and other specialists were not consulted.

In the scientific community, the Vinland map has probably generated more controversy than the Shroud of Turin, owing to both its content and its chemical composition. The map depicts Greenland as an island with a shape that is quite accurate, considering that in the fifteenth century it was thought to be a peninsula, and there were no recorded passages of ships around the island until the twentieth century. Some of the Latin names cited on the map were not believed to have been used until hundreds of years later.

One of the main factors that has continued to fuel the Vinland map controversy is the radiocarbon dating analysis, which indicates the map parchment to be from the 1423 to 1445 time period, consistent with its purported date of origin. The map, however, has a coating of an unidentified substance from the 1950s. The British Museum examined the map in 1967 and found that it appeared to have been drawn with two superimposed lines consisting of black and yellow inks, and was unlike any that had been examined previously.

In 1972, Yale University sent the map to McCrone Associates for chemical analysis of the ink. Walter McCrone found that the map was produced with a yellow line overlaid with a black line in the middle (McCrone, 1976). McCrone indicated that this was done with considerable skill, but he did find segments where the yellow and black lines were not parallel. The use of two lines was apparently done to simulate the natural weathering of medieval black iron-containing inks, which would develop such an appearance as the ink migrated slowly along the parchment fibers. Using polarized light microscopy, SEM–EDS, transmission electron microscopy (TEM), XRD, and other methods, McCrone identified anatase in the yellow ink, with the anatase having morphologies, particle sizes, and particle-size distributions consistent with those of modern formulations. Anatase was first used as an ink pigment in the 1920s and may have a yellow hue from iron impurities. McCrone thus concluded that the Vinland map could not have been produced when purported.

Proponents of the authenticity of the Vinland map have argued that since anatase is present in trace quantities in some of the components of inks used in the medieval period, McCrone's findings do not support a forgery. However, there have been no investigations into the microscopic characteristics of anatase found in such documents to show that these impurities cannot be distinguished from modern formulations, which was the key to McCrone's findings. In 1991, McCrone performed a second examination of the map ink. Using an infrared microscope, he identified the ink binder as gelatin, probably made from animal skin (McCrone, 1999a). Although this does not rule out a medieval origin, McCrone confirmed that significant quantities of anatase are present. This had become an issue since his first study, as an examination of the ink using PIXE suggested that the titanium levels were too low to be from anatase in the ink (Cahill et al., 1987). However, a 2002 study using a Raman microscope confirmed that a significant amount of anatase is present in the yellow ink, and found that most of the black ink consists of carbon (Brown and Clark, 2002). Such inks would not be expected to produce yellowing with age, which is characteristic of iron-containing inks.

Judging from the titles of two letters to *Analytical Chemistry* and a recent review article—*Evidence That the Vinland Map Is Medieval* (Olin, 2003), *The Vinland Map Ink Is NOT Medieval* (Towe, 2004), and *Analysing the Vinland Map: A Critical Review of a Critical Review* (Towe et al., 2008)—the heated controversy over the authenticity of this document will probably not end any time soon. McCrones's detailed examinations of the anatase particles of the yellow ink, however, provide convincing evidence for a modern pigment rather than a natural contaminant in a medieval pigment.

A Trove of Previously Undiscovered Paintings by Jackson Pollock?

In 2002, Alex Matter reported his discovery of more than 20 previously unknown paintings that he attributed to the American abstract expressionist artist Jackson Pollock (1912–1956). He found them in a storage locker belonging to his deceased father, who had been a friend of Pollock's. Based on the style of the paintings and an inscription on the paper wrapper in which the paintings were said to have been discovered, at least one leading expert on Pollock's work concluded that the collection was genuine. This expert felt that "there are too many things about them that are pure Jackson" (Kennedy, 2005). In preparation for an exhibition that featured the paintings, the owners sent them to the Harvard University Art Museum Laboratory, the Museum of Fine Arts (Boston) Laboratory, and to Orion Analytical LLC in Massachusetts for paint analysis. Scientists at these laboratories also examined previous known paintings by Jackson Pollock as a comparison (Martin, 2008).

The following are some of the paint components that were identified in most of the paintings that were analyzed: (1) an acrylic binder with styrene, (2) Benzimidazolone Yellow 4G (Pigment Yellow 151), (3) Benzimidazolone Orange (Pigment Orange 62), (4) rutile-coated mica pearlescent pigments, (5) iron oxide–coated mica pearlescent pigments, and (6) DPP Red BO (Pigment Red 254). These components were identified by infrared spectroscopy, with additional methods, including microscopy and elemental analysis, used to identify the pearlescent pigments. Benzimidazolone Yellow 4G, Benzimidazolone Orange, and DPP Red BO were first identified in automotive paints. Because they were fairly new, their introduction dates and their first known use for American vehicles were also reported to aid forensic paint examiners interpret data from unknown paint chips recovered from hit-and-run scenes. The art provenance chemists performing the Pollock studies relied on these automotive paint studies, as well as others (Martin, 2008).

None of the six paint components cited was found in previous known works by Pollock. More revealing, however, are the commercial introduction dates for the six: (1) 1961, (2) 1969, (3) 1971, (4) 1978, (5) 1979, and (6) 1983 (Martin, 2008). All six were therefore introduced *after* Pollock's death in 1956, and all three laboratories concluded that the paintings containing these compounds could not have been produced by Pollock.

Acknowledgments

We would like to thank JoAnn Buscaglia (FBI Laboratory Counterterrorism and Forensic Science Research Unit), Vincent Desiderio (New Jersey State Police Forensic Laboratory), Jim Parzych (Washington State Crime Laboratory), and John Reffner (John Jay College of Criminal Justice) for their helpful discussions; and Mr. Desiderio, Helen Griffins (Ventura County Forensic Laboratory), James Martin (Orion Analytical LLC), Erik Neilson (Washington State Crime Laboratory), Skip Palenik (Microtrace), Bill Schneck (Washington State Crime Laboratory), Donna Wilson (Washington State Crime Laboratory), and Daniel Van Wyk (Washington State Crime Laboratory) for taking the time to review this manuscript.

The co-author (EMS) is eternally grateful for the involvement of the principal author, who was extremely reluctant to get involved in yet another writing project. He only consented after considerable arm-twisting and cajoling, as—being a very tactful person—he was concerned about offending those who in the past have sought his considerable knowledge, skills, and writing ability, but whom he was unable to accommodate.

CHAPTER 6

Analysis Techniques Used for the Forensic Examination of Writing and Printing Inks†

GERALD M. LaPORTE

National Institute of Justice, Office of Investigative and Forensic Sciences, Washington, DC

JOSEPH C. STEPHENS

United States Secret Service, Forensic Services Division, Questioned Document Branch,
Instrumental Analysis Section, Washington, DC

Summary The analysis and identification of writing and printing inks and toners is generally very important in document examination, especially when used in conjunction with a reference library. Inks can be differentiated based on the chemistry of colorants, solvents, resins, and additives. Instrumental analysis, including GC–MS, HPLC, and FT–IR and Raman spectroscopy, can often be used following visual examination, microscopic observation, and thin-layer chromatography. Analysis of toners can be performed with XRF, SEM–EDS, or pyrolysis GC. Although chemical analysis of materials used to create documents can provide vast amounts of relevant information and strongly support associations between questioned and known materials, in nearly all cases the data obtained will not support a conclusion that identifies a particular writing instrument or printing device.

6.1 Introduction 226
6.2 Ink 226
    6.2.1 Ink composition 227
6.3 Ink analysis 230
    6.3.1 Physical examinations 233
    6.3.2 Optical examinations 236

---

†All references pertaining to manufacturers and their products do not imply endorsement by the United States government, or the authors.

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

  6.3.3 Chemical examinations 238
  6.3.4 Ink dating 240
6.4 Office machine systems 242
  6.4.1 Inkjet ink 242
  6.4.2 Inkjet ink analysis 243
  6.4.3 Toner printing 245
  6.4.4 Toner analysis 246
Conclusion 247
References 248

6.1 INTRODUCTION

Crimes that involve questioned documents can include matters of national security (National Commission on Terrorist Attacks Upon the United States, 2004), homicides, and various types of document fraud. In all cases it is imperative for investigators and forensic scientists to recognize the realm of examinations that can be conducted on documents and the potential information that may be gleaned from their findings. A questioned document (QD) is defined by Lindblom (2006) as "... any material containing marks, symbols, or signs that convey meaning or message." Documents can present themselves in numerous formats, including letters, envelopes, packages, calendars, diaries, currencies, identification cards, financial documents, contracts, wills, and business records. Forensic document examiners (FDEs) can also evaluate a variety of evidence that does not fall into the traditional category of a document, but requires similar expertise, such as written entries found on the body of a victim, or a threat written or painted on a wall.

General acceptance of handwriting analysis to identify a suspected author dates back several decades. While handwriting evaluations are a valuable QD examination, there is a gamut of nonhandwriting examinations that require chemical analyses. Given the variation of materials encountered on a questioned document and the diversity of analytical techniques available, the purpose of this chapter is to provide an overview of the chemical procedures that may be used for document examination. The authors have chosen to highlight some common procedures that may be employed to gather information from documents submitted for forensic analysis. The appropriate scientific procedure is heavily dependent on the composition of the material being analyzed. Often, documents are created using a combination of inks and paper. Correction fluid, adhesives, stains, and various other materials may be encountered on a document as well.

6.2 INK

Ink is one of the most common materials used in the production of a document. Although there are various types of writing and printing inks, chemical and instrumental techniques remain similar when attempting to characterize their basic components. Typically, the goal of such examinations is to determine the source of the manufacturer, compare questioned and known items, and ascertain when the document was produced and/or decipher alterations, erasures, and obliterations.

6.2.1 Ink Composition

All inks, in their basic form, are composed mainly of a colorant(s) suspended in a vehicle (solvents and resins). There are also other organic and inorganic ingredients that may be present in inks, which can include antioxidants, preservatives, wetting agents, lubricants, and trace elements, but these typically form a small fraction of the overall formulation. Nevertheless, their importance should not be discounted because it is possible that a combination of these components allows otherwise similar inks to be discriminated. Before delving into the various analytical approaches for the examination and characterization of inks, it is necessary to gain an understanding of the major components.

Colorants Colorants are a crucial part of all inks because without them, inks would not be discernible under visible light (approximately 380 to 780 nm). The molecular composition of the colorants will dictate how certain wavelengths of visible light are absorbed and reflected, thereby influencing their color. Depending on the vehicle and its interaction with the colorant, two types of colorants can be used: dyes and/or pigments. Dyes are generally considered to be compounds with highly conjugated resonance structures. Their molecular weights can vary from the low hundreds to the high thousands. Dyes can be classified according to their chemical structure or how they are applied to material. It is beyond our scope in this chapter to adequately review the various types of dyes. The most comprehensive volume regarding dye information is *The Colour Index*, published by the Society of Dyers and Colourists and the American Association of Textile Chemists and Colorists (1971). Dyes may be listed by their International Union of Pure and Applied Chemistry (IUPAC) name, Chemical Abstracts Services (CAS) number, or the Colour Index (C.I.) name. *The Colour Index* divides dyes into a series of large groups: acid dyes, azoic dyes (monoazo-, diazo-, triazo-), basic dyes, developers dyes, direct dyes, disperse dyes, fluorescent brighteners, food dyes, ingrain dyes, leather dyes, mordant dyes, natural dyes, oxidation bases, phthalocyanine dyes, reactive dyes, reducing agents, solvent dyes, sulfur dyes, and vat dyes. These colorants are used in many industries to yield colors on an array of substrates; however, solvent dyes and pigments are the most commonly used colorants in writing instruments. Acid dyes and reactive dyes are also used, but less frequently. The molecular structures of a common phthalocyanine dye and a monoazo reactive dye are shown in Figure 6.1.

The major distinguishing feature between dyes and pigments is that the latter consists of fine particles of insoluble material that are suspended in the vehicle. Pigments are generally considered more stable and long-lasting than dyes because pigments are less prone to photodecomposition (lightfast) and are insoluble in

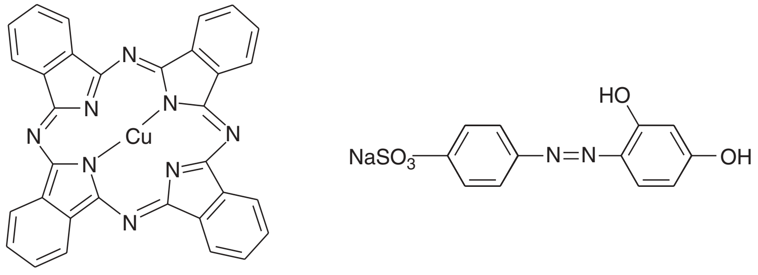

Copper Phthalocyanine

C.I. Acid Orange 6 (mono)-azo dye

Figure 6.1 Molecular structures of copper phthalocyanine and a monoazo reactive dye that can be found in writing and printing inks.

water (waterfast). Their color can be derived from a metal-centered complex and is generally less vibrant than dyestuffs. Additionally, pigments are more opaque than dyes, so the colorant is more efficient at masking any underlying material. There are five major categories of pigments: organic pigments, toners, lakes, extended pigments, and inorganic pigments. Toners, lakes, and extended pigments are all precipitated from an aqueous solution in conjunction with some catalyst(s).

Solvents The fluid portion of ink that suspends and delivers the colorant to the substrate is known as the *vehicle*. Vehicles are necessary to carry the color from the cartridge to the paper. Once on the paper, the solvent undergoes a series of changes over a fixed period of time, causing the colorant to dry onto the paper. Changes that occur to the solvent–ink mixture over time can include polymerization, evaporation, oxidation, or photodecomposition. These modifications to the original chemical composition of the ink have been the focus of methodologies that are employed when attempting to date a document. A more extensive discussion of ink aging will be covered in the ink analysis section of this chapter.

Although much of the specific information on each company's ink formulations is proprietary, there are some standard chemicals that are used as vehicles. Glycols, alcohols, and water are the most commonly found solvents in use for pens today (Brunelle and Crawford, 2003). 2-Phenoxyethanol (2-PE), ethanol, 1-phenoxypropan-2-ol, benzyl alcohol, and many more solvents can also be used as vehicles. The choice of which solvent or solvents to use often relies on properties related to the writing instrument. The writing instrument type (e.g., fountain pen, ballpoint, felt-tip marker), composition of the ink cartridge, region of sale (e.g., dry, humid), and type of colorants and resins are all considerations when attempting to determine the type of solvent(s) to be used. The desired properties for a single formulation are often achieved through a combination of several vehicles. Other considerations, such as the solubility of the colorants in the vehicles, are also critical, since vehicles are the primary means of delivering the colorant to the substrate.

Resins and Additives Resins, which can be natural or synthetic, are incorporated into inks to provide them with a desired viscosity and a means to bond the ink and the substrate as the ink dries. Normally, the resinous material is dissolved into the vehicle to create a solution in which colorants can be added. Phthalates, ketone resins, styrene-type resins, phenol-type resins, rosin-type resins, and poly(vinylpyrrolidones) are just a few plasticizers that find use in ink formulations as binding resins (U.S. patent 7,462,229). Along with the resins, additives such as biocides, surfactants, lubricants, corrosion inhibitors, preservatives, buffers, and diluting agents are included in ink formulations to enhance properties of the ink (Brunelle and Crawford, 2003). Bügler et al. (2005) used thermal desorption and gas chromatography–mass spectrometry (GS–MS) to characterize ballpoint pen inks. Table 6.1 is a summary of the resins and solvents found in a population of 25 inks.

TABLE 6.1 Summary of Resins and Solvents Found in 25 Different Ballpoint Inks Analyzed Using Thermal Desorption and Gas Chromatography–Mass Spectrometry

Manufacturer	Year of Introduction	Color	Resin Class[a]	Solvents[b]
1	1997	Black	AF	PE, PDE, *N*-methylpyrrolidone
2	2000	Blue	AF	PE, PDE
1	2000	Black	AF	PE
3	2000	Blue	AF	PE, PDE
	2000	Black	CF	PE, PDE
4	2000	Black	AF	PE, PDE
5	2001	Blue	CF	PE, dipropylene glycol
	1999	Black	AF	PE, PDE, ethyl diethylene glycol
6	2002	Black	Alkyd	PE
7	2002	Black	Alkyd	PE, PDE, propylene glycol
8	2003	Blue	AF	PE, PDE, diisopropylene glycol
	2003	Black	AF	PE, PDE, ethyl diglycol
9	2003	Blue	CF, Alkyd	PE, PDE, hexylene glycol
	2003	Black	AF	PE, PDE, hexylene glycol
10	2003	Blue	Alkyd	PE, butylene glycol
11	2003	Blue	Alkyd	PE, PDE, benzyl alcohol
12	2003	Blue	AF	PE, PDE, butylene glycol
	2003	Black	AF, Alkyd	PE, PDE, butylene glycol
13	2004	Blue	AF	PE, benzyl alcohol
	2004	Black	Unknown	PE, benzyl alcohol
14	2005	Blue	Alkyd	PE, benzyl alcohol
	2005	Black	Alkyd	PE, benzyl alcohol
15	2005	Blue	CF, Alkyd	PE, PDE
	2005	Black	CF, Alkyd	PE, PDE
	2005	Blue	CF	PE, PDE, trioctylphthalate

Source: Adapted, in part, with permission, from the *Journal of Forensic Sciences*, Vol. 50, No. 5, copyright © ASTM International.

[a] AF, acetophenone–formaldehyde resin; CF, cyclohexanone–formaldehyde resin.

[b] PE, 2-phenoxyethanol; PDE, 2-(2-ethoxy) ethoxybenzene.

TABLE 6.2(a) Ink Formulation Containing Glycol

U.S. patent: 4,077,807	Formula
Ethylene glycol monophenyl ether	20.0
1,3-Butylene glycol	20.0
Fatty acid (oleic acid type)	12.0
Solvent Blue 5 dye (such as Victoria Pure 6 Blue BO base)	5.0
Solvent Violet 8 dye (such as Methyl 4 Violet base)	5.0
Resin, ketone condensation type	35.0
Resin, poly(vinylpyrrolidone) type	3.0
Corrosion inhibitor	1.0
Antioxidant	2.0
Parts by weight	103.0

TABLE 6.2(b) Series of Ink Formulas Containing Benzyl Alcohol, Antioxidants, and Corrosion Inhibitors

	Formula			
U.S. patent: 4,077,807	A	B	C	D
Benzyl alcohol	22.0	22.0	22.0	22.0
Ethylene glycol phenyl ether	22.0	22.0	22.0	22.0
Oleic acid	12.0	12.0	12.0	12.0
1,2-Propylene glycol	4.7	4.7	4.7	4.7
Blau Base KG (Solvent Blue 64, BASF)	15.9	15.9	15.9	15.9
Victoria Blue base F4R (Solvent Blue 2, BASF)	7.5	7.5	7.5	7.5
Hexane triol phthalate resin	15.9	15.9	15.9	15.9
2,2-Methylene bis(4-methyl-6-tertiary-)butylphenol (antioxidant)	0.0	2.0	0.0	2.0
Benzotriazole (corrosion inhibitor)	0.0	0.0	0.1	0.1
Parts by weight	100.0	102.0	100.1	102.1

Additional information on specific ink formulations can often be found in patents. Tables 6.2 and 6.3 depict the chemical profiles of several inks obtained from U.S. patents 4,077,807 and 4,097,290.

6.3 INK ANALYSIS

The information obtained from forensic ink examinations can be a critical component of a case, but the analysis process can be extraordinarily complex. We have chosen to generically describe modern techniques that pertain to more commonly encountered inks such as writing instruments and office machine printers (i.e., photocopiers, laser printers, and inkjet printers). Ink analysis revolves around classifying and identifying the various components of ink: colorants, vehicles, and additives. It is very rare to identify a questioned ink as coming from a specific pen or printer, so in most cases, conclusions are reached about the formulation of the ink. Being a mixture, an analytical profile for the ink formulation can be developed using an assortment of instrumental techniques on the various components.

TABLE 6.3 Ink Formulations for the Gillette Company

U.S. Patent 4,097,290	839-41A Black	814-96A Black	814-92C Red	839-41 Green	839-26C Blue
Natural rubber	22.5%	26%	29%	22.5%	22.5%
Aliphatic diluent No. 6	23.5%	—	—	23.5%	23.5%
Aliphatic solvent 360-66	—	17%			
VM&P naphtha	—	14%	19%		
Terphenyl (HB40)	30%	26%			
DOP	—	—	—	33.5%	27%
Mineral oil	—	—	27%		
Pigment	20%	17%	15%	16.5%	23%
Stearic and lauric acid	4%	—	—	4%	4%
	Carbon black (0.08 µm)	Graphite (2–5 µm)	Organic Red (0.015 µm) suspended in mineral oil (35% pigment dry basis)	Phthalocyanine Green (0.015 µm)	Phthalocyanine Blue (0.015 µm) Victoria Blue 10.5% (0.025 µm)

Numerous formulations of writing inks are available to the general public, thereby greatly decreasing the frequency of occurrence of any particular ink formulation. Personal preferences (e.g., fountain, ballpoint, roller, and gel pens), changes in technological features, cost considerations, and the availability of raw materials are some examples of why so many writing inks exist. Furthermore, most ink companies spend substantial resources on research and development. As a result, many of these formulations are considered proprietary and kept secret. It is also important to recognize the significance when two inks are determined to be chemically indistinguishable or "match" each other. As outlined in Section 9.3 in the ASTM's *Standard Guide for Test Methods for Forensic Writing Ink Comparison* (ASTM E1422-05, 2005):

> When the comparison of two or more ink samples by optical or chemical analyses, or both reveals no significant, reproducible, inexplicable differences and there is significant agreement in all observable aspects of the results, it may be concluded that the ink samples match at the level of analysis and that the results indicate that the ink samples are of the same formula or two similar formulas with the same non-volatile components.

Determining the significance of a match can be further interpolated when the analyst has a reference collection for comparison as outlined in the ASTM's *Standard Guide for Writing Ink Identification* (ASTM E1789-11). The U.S. Secret Service (USSS) and the Internal Revenue Service (IRS) jointly maintain the largest known forensic collection of writing inks in the world. The collection includes more than 10500 samples of ink that date back to the 1920s and has been obtained from various manufacturers throughout the world. Pen and ink manufacturers are contacted on an annual basis and requested to submit any new formulations of inks, along with appropriate information, so that the new standards can be chemically tested and added to the reference collection. In addition, writing pens are obtained on the open market and compared with the library of standards to identify additional inks that may not have been formally submitted by a manufacturer. Maintenance of the library is a formidable task that obviously requires significant resources and is not often practical for most forensic laboratories. Consequently, the USSS generally considers offering forensic assistance (made on a case-by-case request) to law enforcement agencies requesting the analyses of writing inks. The Bavarian State Bureau of Investigation in Munich, Germany also maintains a reference collection of over 6000 writing inks (personal communication, Dec. 4, 2009).

Requests for ink analysis typically comprise three types. The first may involve comparing two or more inks to determine if the formulations match each other. This may help to ascertain if any of the written entries were added or altered. The second is to attempt to identify the writing ink on a questioned document and provide investigative information regarding the possible source of the ink. The third and most challenging request is to establish when the written entries were created to help determine if they are authentic with respect to the purported date of preparation of a document(s). This type of request, usually referred to as *ink dating*, is addressed in further detail in a separate discussion. The first two tasks can be achieved by comparing the questioned ink(s) with an adequate collection of standards, as described previously.

Prior to conducting chemical analysis, it is important to determine how the entries on a document were produced (e.g., writing instrument, inkjet printer, toner-based office machine). Without this knowledge it will be difficult to select an appropriate analytical scheme. Moreover, there are physical and optical characteristics of inks that, when in combination with the chemical properties, allow one to form an analytical profile of the ink. It is important to recognize that it is the combination of physical, optical, and chemical results that will allow one to make more accurate and objective conclusions.

6.3.1 Physical Examinations

Physical examinations include the determination of the color and type of ink. Determining the type of ink (e.g., ballpoint, felt tip, fountain pen) often requires training and experience in the evaluation of morphological characteristics using a stereomicroscope (approximately 5× to 50×). It is important to establish the type of writing instrument used to make a questioned entry, as this dictates the extraction solvent necessary for a chemical ink examination. There are various types of writing instruments and writing inks, but generally, inks are separated into three categories: nonaqueous or glycol-based (ballpoint), fluid (porous, felt or fiber tips, gel, and roller pens), and fountain pens. Definitive microscopic differentiation of these three categories of inks is possible in most cases due to their unique morphological characteristics on a substrate. Although there are a variety of writing instruments that utilize fluid inks (e.g., gel pens, roller pens, felt-tip markers), their physical properties are not always definitively discernible. In addition to class characteristics, individual characteristics may be identified, thereby making the microscopic examination a necessary part of the procedure. As an example, Figure 6.2 illustrates a damaged ballpoint casing and the microscopic scratching that appears on

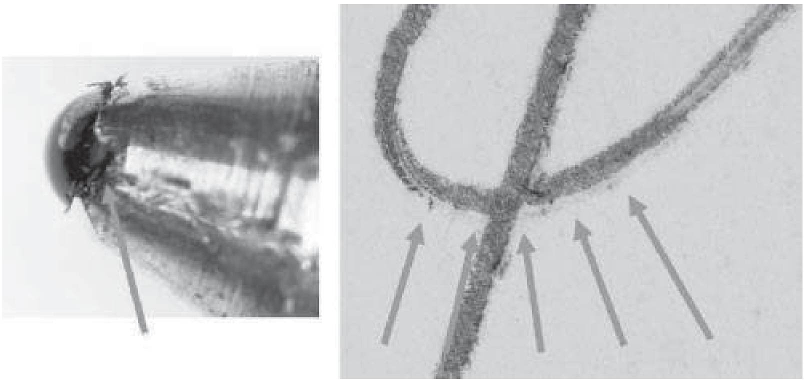
(a) (b)

Figure 6.2 (a) Damaged ballpoint pen casing. (b) Visual defects generated by damaged writing instrument. *(See insert for color representation.)*

the paper. Such damage to the writing tip may be used to help associate a suspect writing instrument with the questioned document.

Morphological Characteristics in the Written Line When examining a written line, the first step is to make a distinction between absorbent and nonabsorbent (or semiabsorbent) ink. Inks that are water based will absorb into the paper fibers, providing nearly uniform coverage of the substrate over the width of the stroke. Semiabsorbent inks such as ballpoint inks generally do not penetrate the fibers as much but merely rest on top of the paper fibers.

Hesitations, or pauses in the flow of writing, are also indicative of a particular type of writing mechanism. While related to vehicle composition, the absorbing inks tend to saturate and bleed into surrounding paper fibers when the pen is left in contact with the paper for a period of time. To a lesser extent, gel pens exhibit some bleeding, while ballpoint pens yield no bleed when left in contact with a piece of paper for an extended period of time. Figure 6.3 shows ballpoint ink, which exhibits no bleed, and a gel ink with minimal bleed. Both the felt-tipped and roller ball inks bleed significantly on paper.

Strokes of each type of writing instrument are shown in Figure 6.4. Both ballpoint ink, and to a lesser extent gel inks, fail to provide complete coverage of the paper fibers because they are semi- or nonabsorbent. Roller ball inks and felt-tipped (porous) inks are absorbent and therefore fully penetrate most porous substrates, creating uniform coverage.

Additional characteristics observed in the written line can further distinguish writing instruments. For example, fountain pens are often noted for their shading and, when the ink is in short supply, for the appearance of two lines. When there is insufficient ink, the two separate nibs will appear on the substrate as two parallel lines. This phenomenon can also be observed if significant pressure is applied to the fountain pen nib.

Ballpoint inks tend to have a sheen and provide incomplete inking across the paper fibers. Gooping, the release of a large volume of ink in a small area, is unique to ballpoint inks. Striations are voids in the inking that are a result of the ball not rolling through the ink supply prior to being rolled over the paper. Striations that

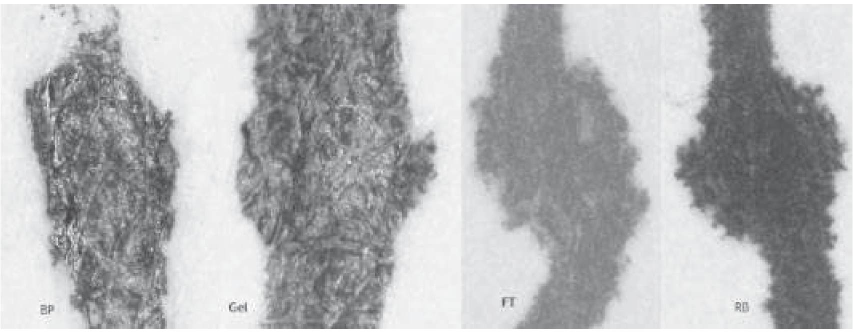

Figure 6.3 Characteristics of the written line from various writing instruments. *(See insert for color representation.)*

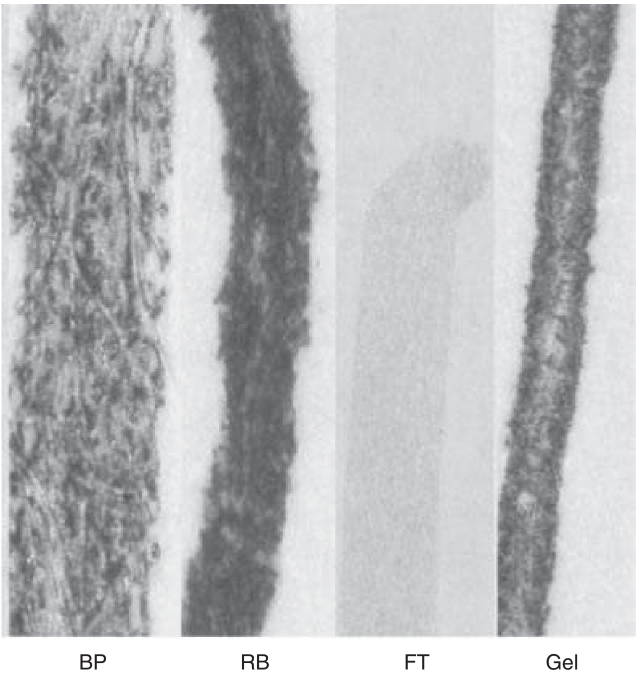

Figure 6.4 Written lines from various writing instruments. *(See insert for color representation.)*

result from the ball are seen only in ballpoint inks because they do not absorb like other ball mechanism inks (gel, roller ball). The ballpoint mechanism typically leaves a writing groove in the substrate.

Non-ballpoint (porous-tipped) inks are noted for their complete coverage across the width of the inked line. No grooves in the substrate are found and the ink tends to bleed if there is hesitation. There are instances where a felt-tipped pen may become degraded and stray fibers on the tip will leave stray markings off the main inked line.

Gel pens, which utilize a ballpoint mechanism, tend to have heavy saturation at the edge of the written line and less saturation in the center. Since little pressure is needed to apply ink to the substrate, the writing groove present is typically not as deep as one would expect to find from a ballpoint pen. Neither gooping nor striations are visualized.

Writing instruments can often leave information that is pertinent to a handwriting examination. Determining the direction of stroke is a valuable forensic examination. The ink saturation on gel pens, porous-tipped pens, steel pens, and fountain pens makes determination of the direction of stroke challenging. Fountain and steel pens may, in some instances, indicate a direction of stroke, as writers tend to apply more pressure on the downstroke (Hilton, 1982). As pressure is applied to the nib-tipped pens, the nibs separate and the one visible line may become two. The striations in

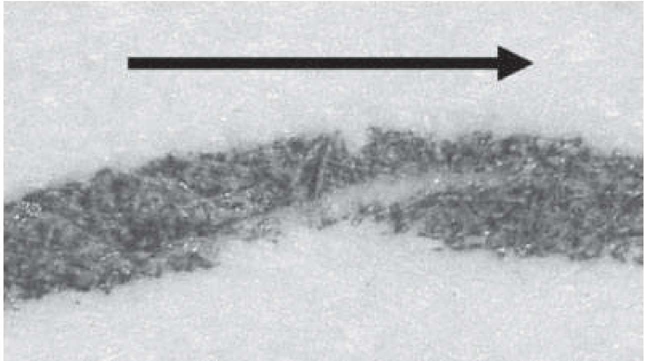

Figure 6.5 Ballpoint pen striations. *(See insert for color representation.)*

ballpoint pens are the best indicator for the direction of stroke. When a ballpoint pen is rotated around the curve of a letter, it often leaves striations on the paper where the inking is incomplete. The striation works its way from the inside of the curve outward. In effect, the thinnest part of the "top" of the curve is always pointing in the direction of the stroke. Figure 6.5 illustrates the direction of stroke of a ballpoint pen. Table 6.4 lists the various physical characteristics that can be used to identify the type of writing instrument.

6.3.2 Optical Examinations

An optical examination, also referred to as a filtered light examination (FLE) or video spectral analysis (VSA), can provide valuable insight regarding the overall composition of an ink. The presence of colorants and other materials will directly affect the manner in which an ink absorbs, reflects, and transmits electromagnetic radiation. Ultraviolet (UV) and infrared (IR) radiation are forms of energy that can be used to examine inks on a document. Near-infrared reflectance (IRR) and infrared luminescence (IRL) properties can help significantly when evaluating properties of an ink. Instruments designed specifically for the purpose of a filtered light examination are best suited for document analysis. Foster & Freeman, Projectina, and other manufacturers provide equipment suited for such analyses.

TABLE 6.4 Summary of the Various Characteristics Typically Found in Writing Inks

	Gooping?	Complete Inking?	Writing Groove?	Ink Bleed?
Fountain pen	—	Some	Yes	Yes
Ballpoint pen	Yes	—	Yes	—
Porous pen	—	Yes	—	Yes
Roller ball pen	—	Yes	Yes	Yes
Gel pen	—	Some	Yes	Some

Drexler and Smith (2002) have described a self-made apparatus that can be manufactured using off-the-shelf equipment, and Richards (1977) provides a basic model for the application of infrared and ultraviolet techniques. Modern commercial equipment typically incorporates the use of a charge-coupled device (CCD), providing users with the opportunity to examine and save images electronically. When illuminating a sample, there are two ways in which the sample can give off energy. The first involves the energy from the illuminant being reflected to the detector. The second way to detect energy from a sample is through luminescence. Luminescence is the absorption of energy at one wavelength and the reradiation of energy at another, typically longer, wavelength. The greatest advantage in examining the luminescence of a sample is the ability to filter out the incident energy on the sample. By filtering out the illuminant, the detector displays only energy that has been reradiated at a longer wavelength. When ink examinations are conducted, both IRL and IRR properties of the ink are recorded. While IRR and IRL properties are often used to determine the regions of interest on a document, they are rarely used to eliminate inks under comparison. Inks that exhibit varying IRR and IRL properties generally contain dyes in the formulation. Typically, dyes do not absorb IR radiation, whereas pigments do. Therefore, pigments will appear black at long wavelengths, whereas dyes often seem to disappear (transmit) in the IR region. When dyes become transparent in the IR region, the underlying substrate often reflects the incident energy, making it seem as though the ink has disappeared. It is important to note that a FLE can be completed before or after chemical analysis is conducted on the ink components. Since ink is a complex combination of chemicals, competing colorants can often reabsorb IRL from other colorants. Some components exhibiting IRL may be visible only after they are separated using chemical analysis, as other components can mask the luminescence. Figure 6.6 demonstrates two inks used for a check alteration with different IRL properties.

An evaluation using optical methods should always be considered, especially when the request is to compare two or more entries, or when evaluating multiple entries on a page. Often, VSA will be used to evaluate if the writing ink on a page is homogeneous. Determining homogeneity may be necessary if further chemical

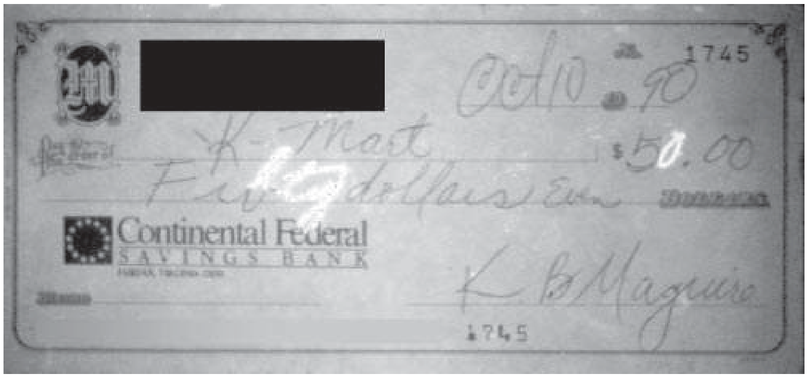

Figure 6.6 Altered check created using two inks with different IRL properties.

testing is warranted. It is not practical to sample and chemically analyze all entries in most cases, so VSA is used as a screening tool. Additionally, it may be possible to detect if there have been any changes to the document or written entries, such as addition of solvents or materials, alterations, insertions, obliterations, and deletions using a FLE. It is critical to emphasize that if two inks cannot be differentiated via their IRR, IRL, and UV properties, this is not conclusive evidence that they are a match as described in the ASTM's *Standard Guide for Test Methods for Forensic Writing Ink Comparison* (ASTM E1422-05, 2005). It is not uncommon that different inks will exhibit similar or even identical optical properties. Furthermore, when performing such examinations, caution must be taken in rendering conclusions if the evaluations are made on different substrates. The substrate can have an overwhelming effect by masking or absorbing the energy that is reradiated from the ink.

6.3.3 Chemical Examinations

Determining the chemical components present in a writing ink is a critical piece for developing an overall profile. The first stage of analysis typically involves extraction of the ink into the appropriate solution. Often, inks with different extractabilities will have different components. Extracting the same inks that have undergone different aging processes, were stored in different environmental conditions, or were found on different substrates may yield different extractabilities. It is important for an analyst to evaluate ink extractability in order to verify that the ink colorants are being extracted appropriately. Occasionally, some non-ballpoint inks may not extract in an aqueous solvent mixture. These are usually pigment-based fluid inks.

Typically, ballpoint inks are extracted with organic solvents such as pyridine, methanol, or acetonitrile. Fluid inks can be extracted with a mixture of ethanol and water (1 : 1), but this is not always the ideal extracting solvent since some fluid inks are non-aqueous-based (e.g., permanent markers). Recently unpublished research suggests that a tetrahydrofuran (THF) and water mixture (4 : 1) may be suitable for both ballpoint and fluid inks. In some instances, inks may have to be extracted in other solvents, depending on the formulation. While it is necessary to maintain a single solvent or solvent system across a reference collection for searching purposes, there are numerous solvents that can be utilized when conducting an intraplate comparison. Figure 6.7 shows the variations in dyes that are extracted based on the assortment of extraction solvents that were used.

Thin-Layer Chromatography Thin-layer chromatography (TLC) is one of the most widely used and generally accepted scientific methodologies employed to compare and help characterize ink formulations. TLC is an effective and efficient method for separating and identifying colorants. For example, two or more questioned inks can be compared to determine if they are the same, or questioned inks can be associated to a known standard. However, it must be emphasized that TLC is only one portion of an analytical scheme, and the "profile" of an ink is achieved only by using the results from a series of physical, optical, and chemical

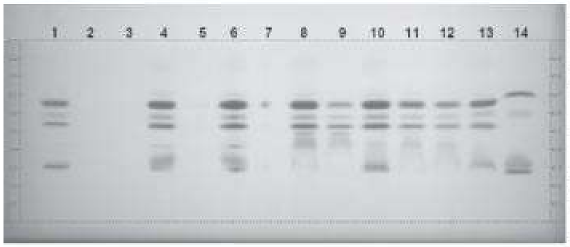

Sample	Extraction solvent
1	Methanol
2	Acetone
3	Tetrahydrofuran
4	Ethanol
5	Ethyl acetate
6	Chloroform
7	tert-Butylmethyl ether
8	Methanol, water (8:2)
9	Methanol, water (1:1)
10	Methanol, acetone (1:1)
11	Ethanol, water (6:4)
12	Ethanol, water (1:1)
13	Ethanol, water (8:2)
14	Dimethylformamide, water (1:1)

Figure 6.7 Different banding patterns based on changes to extraction solvent. *(See insert for color representation.)*

examinations. In 1999, Roux et al. conducted studies on black and blue ballpoint inks using a FLE, TLC, and reflectance visible microspectrophotometry (MSP). The authors concluded that "the power of the individual techniques to discriminate inks between and within brands, models and batches varied, the most informative techniques being TLC > FLE > MSP." They showed that filtered light examinations of blue inks ($n = 49$) and black inks ($n = 42$) resulted in a discriminating power (DP) of 0.83 and 0.96, respectively. After conducting TLC and in conjunction with FLE, the DP was 0.98 for blue and 0.99 for black, where

$$\mathrm{DP} = \frac{\text{number of discriminated pairs}}{\text{number of possible pairs}}$$

TLC for writing inks has been discussed extensively by Brunelle and Crawford (2003), Witte (1963), Brunelle and Pro (1972), Brunelle and Reed (1984), Kelly and Cantu (1975), and Aginsky (1993a). TLC analysis begins by removing an ink sample from a document and subsequently extracting the ink in an appropriate solvent. The extract is then applied to a silica-coated TLC plate (glass or plastic backing) and placed in a solvent-equilibrated glass chamber containing a solvent or mixture of solvents. A mixture of solvents, often referred to as a solvent system, is often employeed to gain timely separation of ink components across the retardation factor range. The sample components then migrate up the plate via capillary action. Typically, the colorants (e.g., dye components) that are present in the ink sample will separate into colored bands or spots.

Other Instrumentation There are numerous research papers regarding the use of instrumental methodologies for the analysis of ink that have been published. The use of gas chromatography–mass spectrometry (GC–MS) and high-performance liquid chromatography (HPLC) for the characterization and/or dating of inks has been discussed by Tebbett et al. (1992), Aginsky (1994, 1996), Andrasko (2001a), Gaudreau and Brazeau (2002), Hofer (2004), LaPorte et al. (2004), Wilson et al. (2004), and Bügler et al. (2005). In addition, spectroscopic techniques such as Fourier transform–infrared (FT–IR) and Raman spectroscopy have been studied for the characterization of inks by Humecki (1985), Merrill and Bartick (1992), and Andermann and Neri (1998). All of these instrumental methods may help further delineate an ink formulation into its individual components, but it is not always necessary. One must take into consideration the environmental changes that can significantly affect the noncolorant components over a period of time immediately following their application to the paper. In cases where a standard reference collection is available, an ink may be determined to match a standard from a database to the exclusion of all others, and further discrimination may not be warranted.

It is critical to understand some fundamental concepts regarding the analysis of inks submitted for forensic examination. First, many inks are composed of volatile and semivolatile components that will evaporate as an ink ages. Therefore, some components that are present in a fresh sample of ink may not be detectable in an aged sample using GC–MS. Although obtaining a supplemental volatile profile may increase the degree of discrimination, limitations may include solvent loss over time or other external factors, such as exposure to high temperatures, light, and/or humidity. Therefore, caution is warranted when making interpretations of qualitative, and especially quantitative, data from instrumental techniques. In addition, there are some commonly utilized dyes that can photodecompose. Triarylmethanes (e.g., methyl violet) are one such group of dyes. They undergo a photochemical reaction that can be “influenced by the concentration of the reactants, the reaction medium, the temperature, the wavelength and intensity of the light” (Weyermann, 2005). Second, when inks are applied to a substrate such as paper, there can be numerous interfering components in the paper that can affect data interpretation. Modern papers can contain optical brighteners, resins, and fillers that can inhibit the analysis of inks. Many inks absorb into the substrate, making FT–IR a difficult methodology, due to the significant contribution of some paper components to IR spectra. Even with spectral subtraction of a paper blank, there may still be an overwhelming effect from some paper ingredients. However, these matrix challenges do not preclude use of the aforementioned technologies. Rather, the implication is that the examiner must consider these variables and the type of request when choosing the appropriate instrumental technique.

In summary, developing a single standardized analytical approach to ink analysis can be challenging, but physical, optical, and TLC examinations should be routine. GC–MS, HPLC, and FT–IR should be considered, but will be dependent on the type of ink, the necessity for further discrimination, and other practical considerations, such as the age of the document.

6.3.4 Ink Dating

Cantu (1995, 1996) has outlined analytical approaches for determining the age of an ink on a questioned document through both static and dynamic methods. The static approach to ink dating generally applies to methods that are based on comparisons with a standard reference collection of inks to determine the first date of production. Numerous ink formulations have been introduced commercially in the past several decades. With the appropriate information and documentation acquired from an ink manufacturer, a database can be used to render more reliable conclusions.

The dynamic approach includes methods that incorporate procedures for the purpose of measuring the physical and/or chemical properties of an ink that change with time. The changes that occur over a given period of time can generally be referred to as *aging characteristics*. Different approaches to measuring the age of an ink once it has been placed on a document have been discussed in the literature by Gaudreau and Brazeau (2002), Hofer (2004), Humecki (1985), Cantu (1987), Brunelle (1992, 1995), Aginsky (1993a, 1998), Andermann and Neri (1998), Brazeau et al. (2000), and Andrasko (2001b, 2003).

When an ink is applied to a substrate and begins "drying," it undergoes numerous changes due to solvent evaporation, oxidation, polymerization, and/or photodecomposition. These processes can change the chemical composition of the ink on the substrate. Given the diversity of inks, substrates, and environmental exposure factors, the complexity of determining what process or processes directly affect the rate at which the ink dries is extremely difficult. The real challenge lies in finding a suitable analytical procedure to isolate the appropriate component for analysis. Brunelle and Crawford (2003) provide a comprehensive review of various methodologies that have been attempted; however, to date, no single ink-dating procedure that has been found to be reliable for all types of inks.

The use of GC–MS for volatile analysis of ballpoint inks in an attempt to determine the age of inks on paper has been studied and reviewed in the literature for more than a decade. Much of the research on volatiles has focused on the analysis of 2-phenoxyethanol (2-PE; also referred to as ethylene glycol monophenyl ether, 1-hydroxy-2-phenoxyethane, β-hydroxyethylphenyl ether, Dowanol EP, and Phenyl Cellosolve). 2-PE is a volatile organic compound commonly used as a principal vehicle in the majority of ballpoint writing inks. Furthermore, it is used in many ink formulations because it is stable in the presence of acids and alkalis. It is nonhygrosopic, nonhazardous, economical, and especially good at dissolving resins and nigrosine (a common solvent black dye used in writing ink formulations). In 1985, Stewart conducted a preliminary study on ballpoint ink aging and the relationship to volatile organic compounds. Beshanishvily et al. (1990) were the first to discuss the identification of 2-PE as it relates to the aging of inks. Since then, Aginsky (1993b) reported that "... significant aging [takes] place over a period of about three months. After this period until the age of fifteen years the extent of the extraction of the volatile component (phenoxyethanol) from the ink entries has been kept at a level of about 20%." Aginsky also describes the ink drying process and surmises that volatile components stop evaporating from a dried sample of ink unless they are freed by heating or a solvent extraction. In 2008, Bügler et al. published a procedure for assessing the age of some ballpoint pen inks using GC–MS coupled with thermal desorption to evaluate the quantity of 2-PE.

Ideally, *ink tags* would be the most reliable method for the dating of inks. Tags can be added to formulations in the form of fluorescent compounds or rare earth elements, and were incorporated into some formulations beginning about 1970 until 1994. Factors have precluded some ink manufacturers from participating in such a program, including, but not limited to, insufficient resources, low priority, and/or disagreement about the type of tag utilized. This is not to say that a widespread tagging agenda is unachievable. On the contrary, efforts continue to convince ink companies to add tags to their formulations. As recently as November 2002, a major ink manufacturer began adding tags to its ink in collaboration with the U.S. Secret Service.

6.4 OFFICE MACHINE SYSTEMS

For the purpose of this discussion, we have chosen to limit the discussion of office machine systems to those that utilize inkjet or toner technology because they are the most commonly encountered printing and copying devices accessible to the general public. Although documents produced via typewriting and dot matrix are not as ubiquitous as some of the other office machine systems mentioned, this does not preclude their importance in some investigations. Occasionally, physical print defects from typewriters, dot matrix, and thermal printers may arise in the printed material and should be examined carefully. Individual and class characteristics from these machines can be important identifying features.

6.4.1 Inkjet Ink

Formulating an inkjet ink to create a high-quality image is an obviously complex task. Inkjet inks for small office/home office (SOHO) machines can have very complex formulations to create acceptable images. In fundamental terms, a well-designed inkjet ink should be able to withstand high heat, dry quickly so as not to smear, have nearly photographic quality, and be light- and water-fast. Since inkjet-printed documents comprise a significant portion of all documents created in the SOHO market and ink formulations are complex, manufacturers utilize a great deal of their resources on the research and development of inks.

Inkjet ink poses several differences as compared to writing inks. Writing ink colorants are mixed together in a single ink formulation to provide a single colored line. Color inkjet, however, is a combination of multiple single-color ink formulations, and each ink formulation can contain one or more colorants. Currently, many SOHO inkjet office machine systems comprise four color printers, incorporating cyan, magenta, yellow, and black (CMYK) into a single print head. The colors can come from either a three-reservoir tricolor cartridge or three single-reservoir cartridges. Typically, the black ink formulation is independent of the colors. Small droplets of ink are jetted onto the substrate to produce a broad range of colors. Figure 6.8 shows microscopic images of inkjet printing on plastic.

Inkjet ink manufacturers typically use acid, direct, and reactive dyes in their formulations. Acid dyes (e.g., azo, anthraquinone, triphenylmethane, azine, and xanthene) are water-soluble anionic dyes that have a wide gamut of hues. Direct dyes (e.g., phthalocyanines) have a variety of bright hues, some of which are fluorescent, as well as possessing relatively good light and water fastness properties. Within the past decade, reactive dyes have been added to some formulations because they contain desired chromophores. This class of dyes link very strongly with

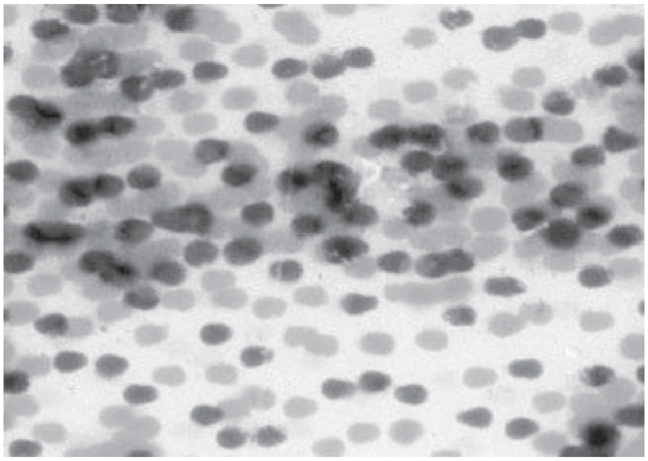

Figure 6.8 Inkjet droplets on a plastic substrate. *(See insert for color representation.)*

cellulosic fibers when heated and they are found in some magentas. Understandably, manufacturers keep their formulations proprietary, which can result in more diverse formulations and thus in greater potential for forensic discrimination. Table 6.5 is a list of various ingredients that are generally used in inkjet ink formulations.

6.4.2 Inkjet Ink Analysis

Many of the same forensic examinations that can be conducted on writing inks may also be completed for inkjet inks. Inkjet inks are typically treated as non-ballpoint pen inks, as they often share similar ink solvents and characteristics. Pagano et al. (2000) presented methodologies for the analysis of jet inks and the different chromatographic systems utilized. Furthermore, LaPorte and Ramotoski (2003) discuss how color inks vary from one manufacturer to the next, making it possible to compare with known standards. LaPorte (2004) also discusses a multifaceted approach to the examination of inkjet printed documents that encompasses physical, optical, and chemical examinations. The U.S. Secret Service maintains the largest known forensic collection of inkjet standards in the world. This collection is composed of print and inkjet ink samples for hundreds of printers and cartridges.

Physical examinations are performed on a printed document to determine the printing process employed. Using at least 10× magnification, inkjet printing appears as a series of irregularly spaced colored dots with some peripheral bleeding. As with writing inks, a FLE is often conducted on inkjet inks. As part of the optical examination, properties of IRR and IRL are noted. Some inkjet inks are formulated using dye-based colorants while others utilize pigments, either in combination or as a replacement to dyes. The selection of colorant (pigment or dye) affects the IRR and IRL properties that are exhibited by the ink. The FDE should exercise caution when evaluating fluorescence and/or absorbance characteristics since inkjet inks may be quenched or enhanced on different types of paper. Optical examinations

TABLE 6.5 Some Common Ingredients That May Be Found in an Inkjet Formulation, Their Purpose, and Approximate Percentage of the Total Weight

Inkjet Ingredient	Purpose	Composition (wt%)
Colorant	Provides the desired color and aesthetic properties	2–8
Solvent	Vehicle that dissolves or suspends the colorant (e.g., water, alcohol)	35–80
Humectant	Inhibits evaporation (e.g., diethylene glycol)	10–30
Surfactant	Lowers the surface tension of the ink to promote wetting (e.g., Surfynol 465)	0.1–2.0
Penetrant	Allows ink to penetrate fibers and promote ambient drying (e.g., pentane-1,5-diol)	1–5
Viscosity modifier	Added to increase viscosity and can be a humectant (e.g., long-chain glycols)	1–3
Dye solubilizer	Helps maintain dye solubility in the solvent because of evaporation in the nozzle (e.g., *n*-methylpyrrolidone)	2–5
Dispersant	Maintains pigment suspensions [e.g., Derussol carbon black (Degussa)]	3–8
Fixative	Helps maintain image permanence and smear resistance (e.g., water-soluble latex)	1–3
pH buffer	Maintains overall pH to promote colorant stability in the vehicle [e.g., Trizma base (Aldrich Chemical)]	0.1–1.0
Chelation agent	Used to complex metal ions (e.g., ethylenediaminetetraacetic acid)	0.1–0.5
Biocide	Kills bacterial growth in aqueous inks [e.g., Dowicil (Dow Chemical)]	0.1–0.3
UV blocker	Increase lightfastness [e.g., Tinuvin 171 (Ciba-Geigy)]	1–5

can sometimes demonstrate that two or more different inkjet formulations were used to prepare a document(s), but this does not necessarily mean that multiple printers were utilized. It is possible for a single printer to use more that one type of inkjet cartridge or for a cartridge to be refilled with a different ink. Some companies, other than the original manufacturer of an office machine, produce compatible cartridges and ink for aftermarket use that are chemically different from those of the original equipment manufacturer (OEM).

Forensic analysis of inkjet inks centers around TLC. In addition to the physical and optical examinations, chemical analysis is an efficient analytical method for classifying inkjet inks. It is important to note that inkjet individualization based on colorant combinations is extremely difficult, due to the interchangeability of inkjet cartridges. Manufacturers often utilize the same dye combinations through several designs of ink cartridges. Given the complexity of the ink, gross changes in formulation are rare. Investigative research on inkjet inks has shown that plastic-backed

TLC plates seem to be as, if not more, efficient at separating the colorants over a fixed distance. Therefore, most TLC comparisons of inkjet inks are completed on plastic-backed TLC plates. It is critical that TLC plates be visualized with an UV source, minimally short wave (254 nm) and long wave (366 nm), since some inkjet inks incorporate proprietary rhodamine dyes.

TLC is only one of several chemical techniques used for examining inkjet inks. However, there has been limited research focusing on other methods of instrumental analyses. Mazella (1999) used diode array MSP to measure the cyan and magenta components in the range 380 to 760 nm on 70 different brands and models of inkjet printers. Doherty (1998) performed a battery of chemical tests, including tests for waterfastness, solubility, and TLC, on documents produced with various inkjet printers. To date, this has been the most comprehensive study on the examination and classification of office machine systems that utilize inkjet technology.

6.4.3 Toner Printing

Laser printers and photocopiers commonly utilize a resinous, powderlike material referred to as *toner*, which is fused to the paper using heat and/or pressure. Toners consist primarily of organic resins, polymers, and colorants such as carbon black, nigrosines, phthalocyanines, azo-pigments, and quinacrodone (Pagano, 2000). Although manufacturers have a selection of resins to choose from, styrene acrylate copolymer and polyester resins are frequently used. Liquid toners, which are often formulated with a hydrocarbon carrier solvent to deliver extremely fine particles to the substrate surface before the liquid evaporates, can also be utilized. Particle-size ranges are approximately 1 to 3 μm for liquid toner and from 6 to 18 μm for dry toner. The size of the toner particle obviously affects the maximum print resolution.

Laser printers and photocopiers function similarly, such that each contains a rotating drum with a photoconductive coating. The surface of the drum is illuminated by a light source controlled by a series of signals based on the image, causing the exposed areas to be charged. As the drum rotates, oppositely charged toner is deposited, thus adhering to the illuminated areas that have been charged with a positive potential. When the paper passes through the machine, an electrical current imparts a static electrical charge onto the paper. The paper then passes through a series of rollers and comes in contact with the rolling drum, causing the toner to be transferred to the paper, resulting in the creation of text or image(s). Since the nonimage areas of the drum are not charged, they do not attract oppositely charged toner and result in white areas on the paper. Afterward, heat and/or pressure is applied to fuse the toner permanently. For a multicolored process, the same steps are repeated for each color component (i.e., cyan, magenta, and yellow).

Two different types of toner are currently produced: conventional toner and chemically prepared toners (CPTs) (Figure 6.9). Conventional toners are created by mixing the colorant(s) into a large batch of polymer. The fine toner powder is then broken away from the large homogeneous mass. CPTs, on the other hand, are grown chemically from small particles. CPTs have gained favor in recent years

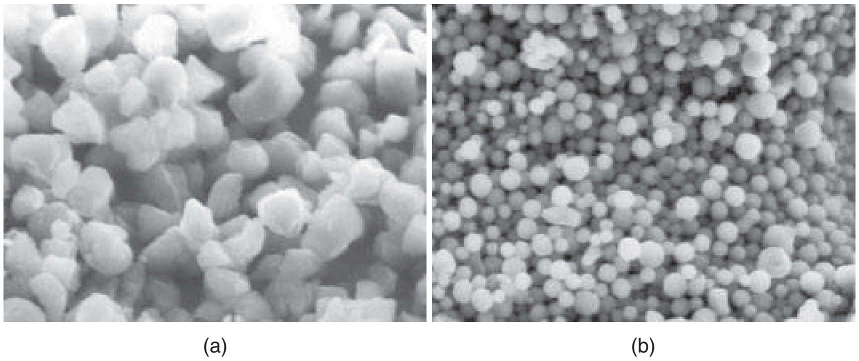

Figure 6.9 (a) Conventional toner. (b) Chemically prepared toner. *(See insert for color representation.)*

due to their more uniform particle-size distributions and shapes. The CPT process allows smaller toner particles to be generated, affording CPTs higher resolutions than conventional toners.

Black toners can be differentiated further based on whether they are mono or dual component. Monocomponent toners incorporate the colorant and the resin into the same large particle. To do so, Fe_xO_y is often used, giving the monocomponent toner a magnetic property. Dual-component toners are typically composed of smaller toner particles and larger developer particles. Although some dual-component particles have magnetic properties, they are not as strong as monocomponent toners. Since dual-component toners have smaller toner particle sizes, they provide better resolution.

6.4.4 Toner Analysis

The analytical scheme for analyzing toners will depend on whether an extensive library is available or if the request is simply to compare a questioned document with other documents or a known printing device. At this time, we are not aware of an extensive collection of black toners that are used for law enforcement purposes, but this does not preclude the possibility that a laboratory has its own internal collection. Various instrumental approaches can be used to analyze toners, and potentially to create a collection of standards, such as pyrolysis GC–MS (Munson, 1989), FT–IR (Andrasko, 1994), x-ray fluorescence (XRF), and SEM–energy-dispersive spectroscopy (SEM–EDS). TLC can also be used to analyze toners and may prove beneficial when conducting comparisons between multiple documents or suspect office machines. Toners can be extracted with chloroform to characterize extractable colorants from the CMY and black components. Given that a number of black toners are carbon black, the forensic value of TLC analysis of black toner can be minimal; however, there are other dyes used in combination with carbon black that may be visualized on a TLC plate. It is highly recommended to visualize developed TLC plates using an UV source. Most important, the results can be used to corroborate other tests and findings.

TLC can be followed with additional methods, but it is critical to understand that the analytical procedures described are not sufficient to individualize a specific printer or copier. It is possible that a manufacturer, or even model, may be distinguished; however, the results are only good enough for discrimination at the class level. XRF analysis can be useful in providing the elemental composition of the toner. The presence and percentage concentration of iron is often a distinguishing factor.

Determining that a toner is iron based (monocomponent) can also be of forensic value, and an excellent area of further research is investigating if quantitative elemental analysis (e.g., Fe) can be used to discriminate toners from different manufacturers or models from the same manufacturer. Scanning electron microscopy (SEM) is also a useful analytical technique, as it provides the high magnification necessary for examining fusing characteristics. Additionally, the method of production used in making the toner can be surmised based on quality images of particle morphology. Given that CPT toners have seen significant use only in the past decade or so, determining that a toner is a CPT is an excellent method for roughly dating a document.

Brandi et al. (1997) offer an extensive discussion and study of toner analysis using diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS), SEM–EDS, and pyrolysis gas chromatography (PyGC). They concluded that DRIFTS was highly discriminating, sensitive, and reproducible after 55 different toners were classified into 15 distinct classes. SEM–EDX was also used to analyze the inorganic contents of 50 toners, subsequently placed into 22 separate classes. The best approach was the combination of the aforementioned techniques, resulting in the characterization of 30 distinct groups of toners. Finally, the authors concluded that "PyGC, with the equipment available, did not appear to provide any greater differentiating power than infrared spectroscopy, gave poor reproducibility, and was a time consuming technique." However, they do acknowledge that a better pyrolysis technique and the use of a mass spectrometer would result in a more discriminating method.

CONCLUSION

The chemical analysis of questioned documments has been relied upon in forensic laboratories for over 50 years. TLC has proven to be a reliable method for the discrimination and identification of some writing and printing inks, especially when used in conjunction with a reference collection. GC–MS is an excellent supplemental procedure and may provide additional information about other components, but the interpretation of the chromatographic and spectral data must be approached with great caution. Inks undergo a drying process that can take place over several months or, possibly, years. Moreover, inks have variable compositions, making the drying processes of different inks nonuniform. This property, in combination with the unknown ink–paper interactions when combined with different substrates and not knowing the storage conditions of a questioned document, result in significant challenges when attempting to characterize and date inks using GC–MS. HPLC and FT–IR can provide additional molecular information, but there has been limited research in the areas of dating inks and characterizing non-ballpoint and printing inks via these methods. Office machines that utilize inkjet and toner technology are ubiquitous and commonly the focus of investigations involving a multitude of crimes. The approach to chemical analyzing inkjets and toners is similar to procedures used for writing inks such that physical, microscopic, and TLC examinations should be a minimum. The gamut of instrumental approaches that follow can vary depending on previous results and the extent of discrimination needed to formulate a reliable conclusion, but it must be recognized that in nearly all cases, the data obtained will not support a conclusion that identifies a particular writing instrument or printing device. Although the information that can be attained from printing and writing inks is based on reliable and generally acceptable analytical methods, it is critical that any person conducting the tests described in this chapter have sufficient training in the physical and microscopic examination of documents. The results from the physical exams help determine how to proceed with chemical methods of analysis, but more important, the combination of all results is necessary to draw relevant and useful conclusions.

REFERENCES

Aginsky, V. N. (1993a). Forensic examination of “slightly soluble” ink pigments using thin-layer chromatography. *J. Forensic Sci.,* **38**:1131–1133.

Aginsky, V. N. (1993b). Some new ideas for dating ballpoint inks: a feasibility study. *J. Forensic Sci.,* **38(5)**:134–150.

Aginsky, V. N. (1994). Determination of the age of ballpoint ink by gas and densitometric thin-layer chromatography. *J. Chromatogr. A*, **678**:117–125.

Aginsky, V. N. (1996). Dating and characterizing writing, stamp pad and jet printer inks by gas chromatography/mass spectrometry. *Int. J. Forensic Doc. Examiners,* **2(2)**:103–115.

Aginsky, V. N. (1998). Measuring ink extractability as a function of age: why the relative aging approach is unreliable and why it is more correct to measure ink volatile components than dyes. *Int. J. Forensic Doc. Examiners*, **4(3)**:214–230.

Andermann, T., and Neri, R. (1998). Solvent extraction techniques: possibilities for dating ballpoint pen inks. *Int. J. Forensic Doc. Examiners*, **4(3)**:231–239.

Andrasko, J. (1994). A simple method for sampling photocopy toners for examination by microreflectance Fourier transform infrared spectrophotometry. *J. Forensic Sci.,* **39(1)**:226–230.

Andrasko, J. (2001a). HPLC analysis of ballpoint pen inks stored at different light conditions. *J. Forensic Sci.,* **46**:21–30.

Andrasko, J. (2001b). Changes in composition of ballpoint pen inks on ageing in darkness. *J. Forensic Sci.,* **47(2)**:324–327.

Andrasko, J. (2003). A simple method for distinguishing between fresh and old ballpoint pen ink entries. *Proceedings of the 3rd Annual European Academy of Forensic Science Meeting*, Istanbul, Turkey, Sept. 22–27.

CHAPTER 7

The Role of Vibrational Spectroscopy in Forensic Chemistry

ALI KOÇAK

Department of Sciences, John Jay College of Criminal Justice, The City University
of New York, New York

Summary Spectroscopy is the study of the interaction of electromagnetic radiation with matter to determine the molecular structure of a solid sample or one dissolved in a specific solvent. This interaction depends on the intrinsic properties of the sample material and can be classified by the energy of the probing electromagnetic radiation. Energy can be in the form of ultraviolet, visible, or infrared light, as well as other forms of energy. One can detect transmitted, absorbed, and/or reflected light. Due to the wide scope of samples typically found in a forensic investigation, a variety of spectroscopic techniques may be needed. Infrared spectroscopy is a good technique to use to identify fibers such as acrylics, nylons, or polyesters, or paints or alkyds, acrylics, or nitrocellulose. The size of the sample may require the use of microscopic infrared spectroscopy, and the nature of the sample may indicate the use of external reflection spectroscopy or attenuated total reflectance spectroscopy. In the chapter we review these techniques as well as related methods of sample identification.

7.1 Introduction to vibrational spectroscopy 252
7.2 Infrared spectroscopy 253
7.3 Infrared sampling techniques 255
  7.3.1 Transmission spectroscopy 255
  7.3.2 External reflection spectroscopy 255
  7.3.3 Attenuated total reflectance 256
  7.3.4 Diffuse reflectance spectroscopy 258
  7.3.5 Infrared microspectroscopy 259
7.4 Raman spectroscopy 260
7.5 Raman spectroscopic techniques 262

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

7.5.1 Surface-enhanced Raman spectroscopy 262
7.5.2 Resonance Raman scattering 263
7.5.3 Coherent anti-Stokes Raman spectroscopy 263
7.5.4 Confocal Raman spectroscopy 263
7.6 Applications of vibrational spectroscopy in forensic analysis 264
References 265

7.1 INTRODUCTION TO VIBRATIONAL SPECTROSCOPY

Molecular spectroscopy is the study of interaction of electromagnetic radiation with matter to elucidate information on molecular structure and dynamics, the environment of the sample molecules and their state of association, interactions with solvents, and many other topics (Graybeal, 1988; Colthup et al., 1990; Schrader, 1995). This interaction depends on the intrinsic properties of the material, classified by the frequency range (energy) of the probing electromagnetic radiation, such as ultraviolet–visible (UV–Vis), infrared (Kocak et al., 2006).

The intrinsic properties of a material that govern its interaction with electromagnetic radiation are described by its complex refractive index (Kocak et al., 2006). Usually represented as having a real part, the *refractive index*, and an imaginary part, the *absorption index*. The complex refractive index of a material is a function that describes the response of the material to the electromagnetic radiation applied. The refractive and absorption indices are both functions of the frequency of the electromagnetic radiation applied (Kocak et al., 2006). Different spectroscopic probes are used to observe and measure the interactions between material and the radiation applied. These measured observables (e.g., transmittance, reflectance, emittance), which are in principle expressible in terms of the intrinsic properties of the material, the molecular composition of the substance, and the wavelength of the illuminating radiation, will determine the type and nature of the spectroscopic technique selected. Due to the wide scope of samples under investigation, a variety of spectroscopic techniques are needed. Selection of the appropriate technique will minimize the time needed for analysis and investigation (Diem, 1993; Schrader, 1995; Mirabella, 1998).

Molecules consist of atoms that have a certain mass, which are connected by elastic covalent bonds. As a result, they can perform periodic motions and continual vibrations relative to each other. The energies associated with these vibrations are quantized; within a molecule, only specific vibrational energy levels are allowed (Colthup et al., 1990; Levine, 1991). For a molecule to absorb radiation, the frequency of the illuminating electromagnetic radiation must match the natural frequency of a bond vibration (Colthup et al., 1990). This phenomenon is similar to the example of a child on a swing: energy is transferred to the swing only if the frequency and the phase of the energy delivered match the frequency of swing movement.

This field of spectroscopy in which the vibration of molecules is involved is called *vibrational spectroscopy* (Colthup et al., 1990; Diem, 1993; Schrader, 1995). Vibrational spectroscopy involves different methods, the most important of which in observing the vibration of molecules are infrared and Raman spectroscopy. Depending on the nature of the vibration, which is determined by the symmetry of the molecule (Cotton, 1990); vibrations may be active or forbidden in the infrared or Raman spectrum (Cotton, 1990; Levine, 1991; Diem, 1993).

The infrared and Raman spectra of two molecules are different depending on whether these molecules have, for whatever reason, different constitutions, isotopic distributions, configurations or conformations, or environments (Diem, 1993; Schrader, 1995). Vibrational spectroscopy detects the characteristic vibrational frequencies of molecules, which provide detailed information about chemical bonding and hence the molecular structure. Consequently, its most common application is in the analysis and identification of small particles. As such, the technique lends itself easily to the study of forensic materials such as controlled drug substances, automotive paint chips, inks, fragments of polymers, textile fibers, papers, inorganic compounds, adhesives, and explosives. In this chapter we describe different sampling techniques and methods involved in both infrared and Raman spectroscopy, along with their forensic applications.

7.2 INFRARED SPECTROSCOPY

Infrared spectroscopy is in principle based on arrangements similar to those of any other spectroscopy: a radiation source, such as globar or another infrared-emitting source, is directed at a sample (Kocak, 1998; Griffiths and de Haseth, 2007). This infrared radiation, which is in the mid-infrared-region range (approximately (400 to 4000 cm^{-1}), interacts with a molecule and its chemical bonds (see Figure 7.1). Most organic functional groups have vibrational and rotational frequencies in the mid infrared region. After interaction with the sample, the infrared radiation is collected by an infrared-sensitive detector, such as a deuterated triglycine sulfate (DTGS) or mercury cadmium telluride (MCT) detector (Kocak, 1998; Griffiths and de Haseth, 2007) and in infrared imaging, a focal-plane array (FPA) detector (Lewis and Levin, 1995).

In classical dispersive infrared instrumentation, a Czerny Turner monochrometer (Ingle and Crouch, 1988; Strobel and Heineman, 1989; Skoog et al., 2007) is used to disperse the infrared light into a single wavelength to plot the spectrum of intensity versus wavelength or wavenumber (Ingle and Crouch, 1988). In modern Fourier transform instrumentation, a Michelson interferometer is used. Based on the development of interferometry, Michelson developed the interferometer in 1880 to study the speed of light (Kocak, 1992; Griffiths and de Haseth, 2007). Using the Michelson interferometer, an interferogram is created by the interference of two or more radiation beams (Ingle and Crouch, 1988; Kocak, 1998; Griffiths and de Haseth, 2007). As a result, all frequencies are detected at once in the time or distance domain. The two domains of time or distance and frequency

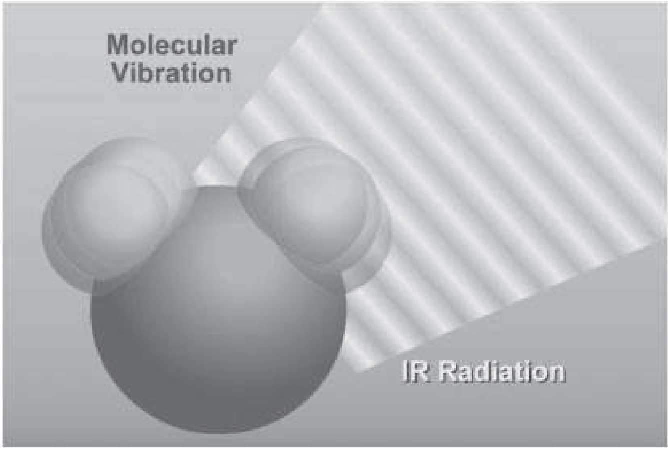

Figure 7.1 Basic principles of IR spectroscopy. *(See insert for color representation.)*

contain the same essential information, and they are interconvertable by fast Fourier transformation (Kocak, 1998; Griffiths and de Haseth, 2007). The spectrum is then plotted as intensity versus wavelength or wavenumber.

Infrared spectroscopy is an important technique for the identification of a class of compound groups or an individual compound present within a specimen. For example, it can be used to identify fibers as acrylics, nylons, or polyesters, or paints as alkyds, acrylics, nitrocellulose, and so on (Kirkbride, 2000). It has a great advantage over other techniques in its ability to obtain structural information that might be difficult to elicit using other techniques. Spectral data representing absorption at specified wavelengths of infrared light characterize chemical and structural groups within the sample, thus providing a molecular "fingerprint" and a tool for sensitive and fast sample composition determination. This allows the technique to be used in the characterization of stereoisomers. For example, it can be a very efficient technique to discriminate between positional isomers of aromatic substances such as various isomers of dimethoxyamphetamine or trimethyl benzene, as well as between other stereoisomers: for example, the isomeric group cocaine, or the group butyl nitrate, isobutyl nitrate, or isoheroin and heroin (Kirkbride, 2000). In drug determination, infrared spectroscopy can be an important tool to distinguish between a free organic and a salt because they have different spectral properties. The N—H stretch in an amine salt is found between 2500 and 3000 cm^{-1}, whereas the N—H stretch in its free base is above 3000 cm^{-1}. Therefore, the spectral properties for drug free bases and their salts are very different within the frequency range of N—H stretch (Kirkbride, 2000; Alzate et al., 2006).

The limit of detection for this technique is usually in the microgram range; this limit may be lowered by using microscope infrared techniques (see below). The technique is relatively rapid and requires minimal sample preparation, which makes it economical and time efficient.

7.3 INFRARED SAMPLING TECHNIQUES

7.3.1 Transmission Spectroscopy

The term *transmission spectroscopy* is used to denote the qualitative and quantitative measurement of the transmittance or absorbance of a material as a function of the wavelength or wavenumber (Ingle and Cronch, 1988; Mirabella, 1998; Griffiths and de Haseth, 2007). The technique involves passing a parallel beam of infrared light through a sample placed in the light beam of the spectrometer, and the *transmittance* of the radiation is measured (see Figure 7.2). This relationship between the incident electromagnetic beam and the pathlength and the concentration of absorbing media is described by Lambert–Beer's law (Swinehart, 1962):

$$A = -\log \frac{I}{I_0} = abc$$

where A is the absorbance, I the intensity of the light, a the absorptivity or extinction coefficient, b the length of the beam in the absorbing medium, and c the concentration of the absorbing species. The resulting spectrum depends on the pathlength or sample thickness, the absorption coefficient of the sample, the reflectivity of the sample, the angle of incidence, the polarization of the incident radiation, and, for particulate matter, on particle size and orientation.

Transmission spectroscopy was the only method in general use until the early 1960s. Thus, large libraries of transmission spectra are readily available. These are frequently used for comparison of spectra obtained by other techniques, even though the frequency response differs between techniques, depending on factors described below. This technique can be used for solid, liquid, and gas sampling. However, to produce useful spectra, in many cases the analyst must engage in a careful and sometimes lengthy sample preparation. The choice of the sample compartment, or cell, is ofen critical. This can cause variations in peak shapes and intensities; for example, a small bubble in a liquid cell will distort intensity ratios (Harrick Scientific Corp., 1987; Mirabella, 1998).

7.3.2 External Reflection Spectroscopy

External reflection spectroscopy is the application of infrared spectroscopy on the characterization of the light reflected from a smooth surface, such as metals, semiconductors, and liquids (Harrick Scientific Corp., 1987; Mirabella, 1998; Humecki,

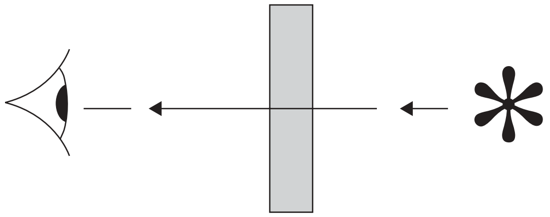

Figure 7.2 Transmitted light as a function of the material and wavelength.

1995). This technique can analyze two classes of samples. The first consists of a highly reflecting substrate (usually, a metal) upon which a material has been deposited (films, contaminants, etc.). The sample beam is reflected from the metal and interacts with film. Since the metal is largely noninteracting, the result is actually an absorption spectrum of the film. This application is sometimes referred to as *reflection–absorption* or *double transmission* (Schrader, 1995; Mirabella, 1998; Griffiths and de Haseth, 2007). A second type of external reflection study is referred to as *specular reflection*, which is the front-surface reflection from the exterior surface of a material. In specular reflection the angle of reflection equals the angle of incidence. These angles are measured from the normal to the reflecting surface. In practice, most external reflection experiments involve a combination of reflection–absorption and specular reflection components from the sample. Since spectral features will differ in the two, it is important to consider these possibilities when interpreting external reflection spectra.

External reflection spectra depend on polarization, angle of incidence, and the substrate's optical properties (see Figure 7.3) (Kortüm, 1969; Harrick Scientific Corp., 1987). The sensitivity of reflected light increases with p-polarized incident light; the phase shift of the reflected light varies with angle of incidence, and hence at large angles of incidence with p-polarized light an optimum sensitivity is obtained. For s-polarized light, the electric field at all angles of incidence has very small values at the interface, resulting in less sensitivity (Harrick Scientific Corp., 1987).

To convert reflectance spectra into absorption-like proportional spectra a Kramers–Kronig transformation (Kocak et al., 2006) should be employed, which is easy to accomplish nowadays, owing to readily available software and computerized spectrometers. Because reflection methods require little or no sample preparation, they are used as frequently as traditional methods. It is a nondestructive method for the measurement of surfaces and coatings without sample preparation. In forensic science it is useful to analyze and distinguish illicit drugs such as cocaine. Koulis et al. (2001) collected and compared the external reflection spectra of cocaine salt and cocaine base.

7.3.3 Attenuated Total Reflectance

The use of internal reflection in infrared (IR) spectroscopy was first suggested by Fahrenfort and Harrick independently in 1959 (Harrick, 1987; Mirabella, 1998). In

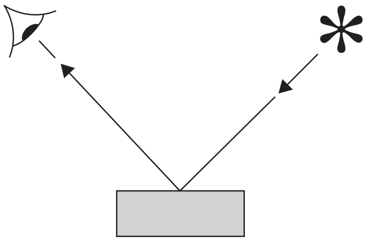

Figure 7.3 External reflection.

the internal reflection mode, an infrared transmissive material with a higher index of refraction than the sample is brought into contact with the sample. The higher-refractive-index material is defined as the internal reflection element (IRE). Typical IREs are ZnSe (refractive index: 2.4), Ge (refractive index: 4.0) and diamond (refractive index: 2.4). In most forensic applications, diamond internal reflection elements are used since they are more durable.

The radiation from the spectrometer's source is directed onto the interface between the sample and the IRE at an incident angle such that the radiation is totally internally reflected (see Figure 7.4). This process causes what is referred to as an *evanescent wave* to be produced a few micrometers beyond the surface of the IRE (Harrick, 1987; Schrader, 1995; Griffiths and de Haseth, 2007). The evanescent wave is absorbed and attenuated by the sample at the characteristic wavelength corresponding to an IR absorption band, producing an attenuated total reflectance (ATR) spectrum. Because the magnitude (the depth of penetration) of the evanescent wave is small, only a few micrometers, ATR is a surface analytical technique and particularly useful for strongly absorbing or thick samples. Several factors can affect the depth of penetration, such as the refractive index of the sample and the crystal, the angle at which the radiation is brought into the crystal, and the wavelength of the radiation (Harrick, 1987; Harrick Scientific Corp., 1987; Griffiths and de Haseth, 2007). Data obtained from ATR measurements are different from the data obtained in transmission measurements. Therefore, they should be compared only for general qualitative purposes. It is worth noting however, that these differences between the two techniques are caused by optical differences rather than compositional ones.

In contrast to external reflection spectroscopy, which requires smooth surface properties, ATR spectroscopy can be used for any surface brought into intimate contact with the internal reflection element. The technique has found a large scope of applications for the analysis of a variety of sample types in a wide spectral range.

For forensic applications, ATR can be utilized in the analysis of pressure-sensitive adhesive tapes (Merrill, 2000). Duct tapes are used in binding victims of violent crimes, and other tapes can be used in packing drugs and illegal materials. Using ATR spectroscopy the major chemical components of these can be determined and leads to the characterization of tapes submitted as evidence to determine if two samples are similar or distinguishable. The brand and the origin

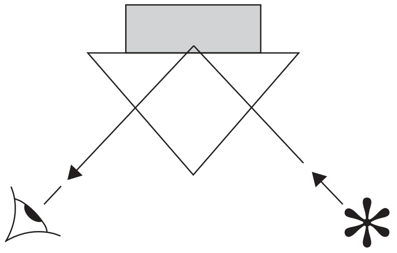

Figure 7.4 Attenuated total reflectance.

of these tapes can also sometimes be determined. It is worthy to note that ATR is a nondestructive technique. However, it may be semidestructive in some cases because the pressure applied to establish a good contact between the sample and the crystal may crush the sample.

7.3.4 Diffuse Reflectance Spectroscopy

Diffuse reflectance is the radiation that penetrates a specimen to a shallow depth and reemerges after scattering off within the sample. It is based not only on reflection and refraction but also on diffraction. These emerging photons from the sample will be collected and analyzed using an infrared detector (see Figure 7.5). Since scattering of the radiation within the sample is an essential phenomenon of the diffuse reflectance technique, it is suitable for inhomogeneous samples that exist as powders or can be converted to powder form. This technique permits the sampling of a solid that has been properly reduced in size, either neat or mixed (~ 10%) with nonabsorbing powder such as KBr or KCl. The process requires grinding and diluting the sample with KBr or KCl to ensure deeper penetration of the incident light into the sample (Harrick Scientific Corp., 1987; Schrader, 1995; Mirabella, 1998).

The diffuse reflectance technique offers the advantage of having a wide solid sampling range; this in particular is useful when a very wide range of solid samples need to be analyzed, as in forensic laboratories. Suzuki and Gresham (Suzuki et al., 1986b) demonstrated a broad study for the use of the diffuse reflectance technique in forensic analysis. They analyzed the antidepressant drug Alprazolam and obtained the spectral results at two extreme concentrations with a lower limit of 20 μg. In general, the diffuse reflectance technique offers several advantages, among them minimal sample preparation, a high degree of sensitivity (less than 100 ng in some cases), and its suitability for very weakly absorbing samples.

Kubelka and Munk (Suzuki et al., 1986a) were first to describe the effect on paint layers of the pigment particles or other scattering inclusions. This model leads to the well-known Kubilka–Munk equation, which is used for the conversion of the measured reflectance spectrum to an absorption-like proportional representation. Almost all modern Fourier transfer (FT)–IR instrument software packages now contain this conversion program.

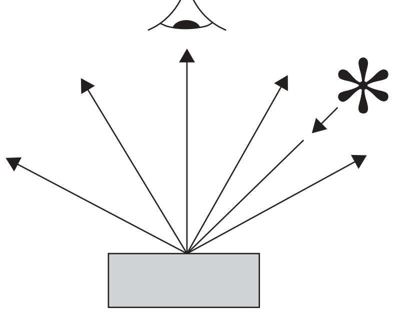

Figure 7.5 Diffuse reflectance.

7.3.5 Infrared Microspectroscopy

The conventional light microscope is an important analytical instrument that is used frequently by forensic scientists, and they are familiar with the technique. It is used in the examination of criminal evidence, including illicit drug crystals, paints, hair and textile fibers, bullets, and documents. These microscopic images are powerful data for the legal evidence, and people without technical backgrounds may have some understanding of the interpretation of the data.

The difficulty that may occur is in distinguishing data obtained from a visual microscope: when, for example, two colored fibers may appear similar under microscopic analysis but are actually stained with two different dyes; or two textile fibers may appear similar under visual microscopical analysis but differ in their molecular composition. A cross section of a polymer may show a uniform microstructure but may have gone through photooxidative degradation.

Combining a conventional light microscope with infrared spectroscopy techniques enables an analyst to obtain an infrared spectrum, or “molecular fingerprint,” of a sample and overcome the difficulties of discriminating the molecular differences mentioned above (Katon and Sommer, 1992; Griffiths and de Haseth, 2007). Using infrared microspectroscopy, it is easily possible to obtain high-quality infrared spectra for chemical composition of an unknown substance as small as about micrometers in area. This enables the analyst to obtain the spectral mapping of the specimen and identify the area within. Spectral mapping involves collecting spectra within a region of a sample while maintaining the spatial resolution inherent in infrared microscopy. Spectra collected at a series of positions can be located spatially within the sample. By mapping these data with respect to a particular band of infrared light, the spatial distribution of the specific chemical components in the sample can be obtained.

Application of IR spectral mapping can readily define the presence and distribution of components of a nonhomogeneous specimen (e.g., drug distribution in human hair). This area of research has been the subject of considerable work, and many interesting results have been obtained (Kalasinsky, 1993). Microscopic probing of small areas within longitudinally microtomed hair sections provides a profile of deposition of a drug along a growth line, and thus individual hairs may reveal the hydrophobic or hydrophilic characteristics of a substance.

Infrared microspectroscopy can record specular, diffuse, internal, or ATR spectra in addition to conventional transmission techniques. It is only the sample itself that determines the choice of the type of analysis technique. Infrared microscopy can be employed in the analysis of small samples of paint, as from motor vehicle accidents; in the case of hit-and-run offenses, for example, one may find small paint chips on the clothes of the victim. A paint chip is embedded and a thin section is produced on a microtome and analyzed. This paint chip analysis can lead to identifying the car type, model, and production year by searching the car paint chip library. Other forensic examples could be single textile fibers as evidence of contact and concentrated extracts of low-dosage drugs (Krishnan and Hill, 1990).

7.4 RAMAN SPECTROSCOPY

Like infrared spectroscopy, Raman spectroscopy probes the vibrational states of complex molecules. However, in Raman spectroscopy the excitation mechanism is different from that of infrared spectroscopy. Raman is concerned with the scattering of radiation by the sample rather than absorption of radiation (Herzberg, 1945; Gerrard, 1994; Schrader, 1995). In a Raman spectrometer the sample is irradiated with an intense source of a monochromatic (laser beam) light with much more energy than that required to cause a transition between vibrational energy states. Since the energy of the illuminating radiation does not correspond to a molecular energy level, no absorption transition will occur. This type of interaction between the illuminating light and the molecule, where no absorption takes place, is called *scattering* (see Figure 7.6). This type of scattered light, which produces photons with the same energy, or color, as exciting radiation is called *Rayleigh scattering* (Gerrard, 1994; Schrader, 1995).

In Rayleigh scattering there will be an elastic collision between the incident photon and the molecule. Since as a result of this type of collision the rotational and vibrational energy of the molecule is unchanged, the energy and therefore the frequency of the scattered photon is the same as that of the incident photon. This appears as the strongest component of the scattered radiation. In 1928, C. V. Raman described another type of scattering, now known as the *Raman effect*. Raman scattering, however, occurs when the molecule does not return to the same energy level from which it originated but ends up in the excited state. As a result, the scattered photon no longer has the same energy or frequency as that of the exciting photon. The Raman effect can be described as an inelastic collision between the incident photon and the molecule where, as a result of the collision, the vibrational and rotational energy of the molecule is changed by an amount ΔE, where ΔE is the difference between incident and scattered radiation. This change will cause a small

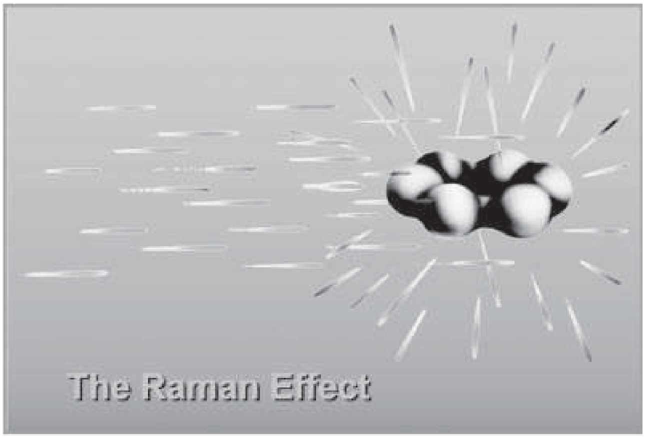

Figure 7.6 The Raman effect: Illuminating light is not absorbed but is scatterred. *(See insert for color representation.)*

fraction of the scattered radiation to exhibit shifted frequencies that correspond to the sample's vibrational transitions. Lines shifted to energies lower than the source are produced by ground-state molecules, while lines at higher frequency are due to molecules in excited vibrational states. These lines, the result of inelastic collision of the light with the sample, are called *Stokes* and *anti-Stokes bands*, respectively. Since molecules of the sample are in the ground state, at room temperature less energy from the source radiation is more favorable than energy gain, and the Stokes band is more intense (see Figure 7.7) (Herzberg, 1945; Diem, 1993; Gerrard, 1994; Schrader, 1995).

The strength of the Raman signal is weak and is inversely proportional to the fourth power of the wavelength of the incident radiation. Therefore, in Raman it is desirable to utilize radiation of relatively short wavelength in the visible region. This may induce fluorescence in the specimen. Fluorescence is likely to block the weak Raman signal and is highly undesirable. A red He–Ne laser (at 632 nm), near-infrared (NIR) diode laser (at about 785 nm), and an Nd:YAG laser (at 1064 nm) can be employed to somewhat reduce this problem and to inhibit fluorescence effect (Turrell and Corset, 1996).

Although both Raman and IR spectroscopy methods probe vibrational states of a specimen, they are both used for qualitative analysis and for molecular structure elucidations, the mechanism of origin of the two techniques is completely different and the two do not produce identical spectra. Infrared absorption can be detected

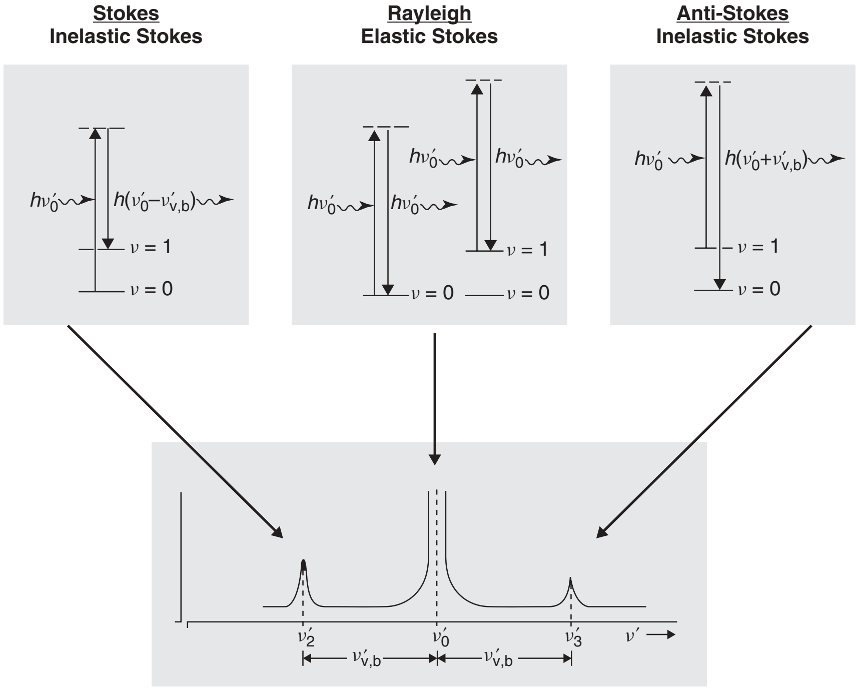

Figure 7.7 The Raman principle. *(See insert for color representation.)*

if the dipole moment in a molecule is changed during the normal vibration. The intensity of an infrared absorption band depends on the change in the dipole moment during this vibration (Colthup et al., 1990).

A Raman active vibration, on the other hand, can be detected if the polarizibility in a molecule is changed during the normal vibration (Colthup et al., 1990; Schrader, 1995). The polarizibility is a measure of how readily the electrons within the molecule are displaced by the electric field of the incident light. The intensity of a Raman band depends on the change of polarizibility during the vibration (Colthup et al., 1990; Schrader, 1995). Therefore, infrared spectroscopy is a very useful probe for those molecules that contain functional groups such as C—O, C—N, N—H, and O—H. Infrared spectroscopy provides little information on nonpolar bonds such as C—C, C═C, and N—N. Raman spectroscopy, on the other hand, can provide useful information in relation to polarizibility of the bond. Nonpolar bonds, particularly double and triple bonds, yield strong Raman bands. Therefore, Raman is considered to be more useful in fiber identification and comparison, as most fibers are composed of long chains of C—C bonds or aromatic rings. Also, Raman techniques are useful for many dyes and pigments, such as aromatic rings or azo linkage.

Raman spectroscopy has several advantages in the analysis, which can be difficult with IR spectroscopy, particularly when dealing with samples of forensic evidence where the chance of losing or destroying the sample is critical. This is particularly important when analyzing fiber samples mounted on a glass slide for visual microscope analysis. Light microscopy analysis requires embedding the sample on a glass slide. Because glass absorbs strongly in the infrared, the fiber needs to be removed from the slide and cleaned prior to IR analysis. This may involve a high risk of losing a evidentiary sample. In Raman spectroscopy, however, one can analyze the sample directly without removing it from the slide (Bartick, 2002).

Raman spectroscopy has long been utilized in forensic research laboratories. However, recent advanced developments in instrumentation have brought about its use in forensic laboratories. The instrumentation has advanced significantly over the last 10 years compared to what it was 20 to 30 years ago. The spectrum can be acquired more rapidly and the technique has become more sensitive and easier to use. During the last few years, Raman spectroscopic imaging has become one of the most popular topics of spectroscopy.

7.5 RAMAN SPECTROSCOPIC TECHNIQUES

7.5.1 Surface-Enhanced Raman Spectroscopy

Surface-enhanced Raman spectroscopy (SERS) is a Raman spectroscopic technique that provides greatly enhanced Raman signal (typically, 10^3- to 10^6-fold) from Raman-active analyte molecules that have been adsorbed onto colloidal metallic surfaces (e.g., on colloidal Ag prepared with citrate), in the particle-size range 25 to 500 nm (Diem, 1993; Schrader, 1995; Campion and Kambhampati, 1998). The importance of SERS is its surface selectivity and high sensitivity. Conventional

Raman spectroscopy is insensitive for surface studies because the photons of the incident laser light simply propagate through the bulk, and the signal from the bulk overwhelms any Raman signal from the analytes at the surface. SERS selectivity at the surface overcomes this problem.

Surface-enhanced Raman scattering arises from two mechanisms: electromagnetic and chemical enhancement. Usually, the electromagnetic effect is dominant, the chemical effect contributing enhancement only in the range of an order or two of magnitude, but this depends on the surface features of SERS-activated substrate (Campion and Kambhampati, 1998). Electromagnetic enhancement is dependent on the presence of the metal surface's roughness features, while chemical enhancement involves changes to the adsorbate electronic states due to chemisorption of the analyte. The magnitude of the electromagnetic enhancement also depends on the intrinsic characteristics of the metal: Ag, Cu, Au, Pt, and so on. The magnitude of electromagnetic enhancement may vary for each of them. The potential used (i.e., the voltage) will also have an effect on both the electromagnetic and chemical mechanisms of enhancement.

7.5.2 Resonance Raman Scattering

Resonance Raman enhancement occurs if the energy of the photon of the exciting laser beam is close to the energy of an electronic transition of the sample. Intensity enhancement can occur by a factor of 10^2 to 10^6 compared with normal Raman scattering (Diem, 1993). Resonance Raman is used in forensic chemistry to achieve detection limits of 10^{-6} to 10^{-8} M.

7.5.3 Coherent anti-Stokes Raman Spectroscopy

In coherent anti-Stokes Raman spectroscopy (CARS), the sample is illuminated by two lasers, one of them with a fixed wavelength, usually referred to as the *pump laser*, and a second tunable laser beam, referred to as the *Stokes frequency*. The pump beam excites the molecule to the first virtual Raman level. The second laser, a tunable probe beam, stimulates the Stokes transition. The pump then again excites the molecule to a higher virtual level, but starting from the higher vibrational state. Finally, the molecule decays back to the ground state, emitting an anti-Stokes photon (Diem, 1993). CARS produces the same Raman signals as those of conventional Raman spectroscopy, but at levels four to five orders of magnitude greater.

7.5.4 Confocal Raman Spectroscopy

Confocal Raman extracts information about the interior of a sample, pinpointing the exact location of an element embedded inside a sample (Everall, 2004a, 2004b). Laser radiation is focused on a spot with the help of a microscope objective. Advantages of confocal Raman are (1) that it extracts information from a small point within the interior of a large sample, (2) that there is better axial and lateral resolution than with a conventional microscope, (3) that stray light is rejected, which avoids interference and eliminates fake peaks, and (4) superior rejection of fluorescence.

7.6 APPLICATIONS OF VIBRATIONAL SPECTROSCOPY IN FORENSIC ANALYSIS

Vibrational spectroscopy has long played a significant role in both research and applications of forensic science. Indeed, the technique has become one of the important tools in many fields of forensic science. A wide range of applications to physical evidence involved in crime investigations have been developed, including examining everything from drugs and paints, fibers, explosives, polymers, and inks to documents. Vibrational spectra has also been an effective tool for analyzing samples collected from crime scenes, such as blood, fabrics, and soil particles.

De Gelbder et al. (2005) utilized the Raman spectroscopy technique for the analysis of automotive paints to determine the origin of the paint and hence the type of automobile. Determining the origin of the paint is important forensic evidence in many car accidents, particularly in hit-and-run cases.

Several researchers have used vibrational spectroscopic techniques to analyze other types of paints, including white household paints (Bell et al., 2005), which are considered important forensic evidence. However, unlike automotive paints, which can be distinguished by color or type, many household paints cannot be distinguished based on color because they are white. Therefore, the composition of the paint needs to be discriminated, and Raman spectroscopy is a useful tool for this purpose. Bell et al. (2005) have demonstrated that Raman spectroscopy has a considerable potential for the analysis of white paints.

Vibrational spectroscopy is an effective tool for an analysis of evidence samples for the presence of explosive materials, and many vibrational spectroscopy techniques have been used for this purpose (Bartick, 2002). Bartick and Mount (2001) demonstrated the use of Raman and infrared spectroscopy in the analysis of suspect explosive components submitted as case evidence (Bartick and Mount, 2001). Samples of evidence were taken from the home of a suspect who was under suspicion for producing bombs. The evidence included all the components needed to prepare an extremely unstable explosive material, hexamethylenetriperoxideamine (HMTD). One of the major components found was a bottle labeled Welloxide, a liquid stabilizer developer. Welloxide is a hair-coloring developer that contains hydrogen peroxide. Hydrogen peroxide is one of the components used to produce HMTD. Batrick at el. analyzed Welloxide using IR and Raman spectroscopic methods to determine if there was sufficient H_2O_2 to produce HMTD. As a result, they concluded that there was sufficient H_2O_2 in the suspect's possession to produce a bomb. Raman spectroscopy, however, was their method of choice, which showed the distinct presence of hydrogen peroxide's O—O stretching peak at 876 cm^{-1}document. The infrared spectrum, on the other hand, was not observable until the water spectrum was subtracted, and the peak of interest was still weak.

Characterization of drug-packing plastic bags is important forensic evidence in the courtroom to trace the origin of a package. Causin et al. (2006) have utilized the ATR infrared spectroscopic technique to differentiate and discriminate apparently similar plastic shopping bags that were used to pack individual doses. Their study showed that using ATR spectroscopy in combination with other techniques is a useful technique to discriminate mass-produced plastic bags and to trace the source of illicit drug doses with significant results.

ATR is also an important technique in the analysis of pressure-sensitive tapes. Merrill (2000) utilized the ATR technique with single-reflection diamond IRE to analyze four different types of pressure-sensitive adhesive tapes, including duct tapes, electric tape, packing tape, and office tape.

Maynard et al. (2001) used the diffuse reflectance IR spectroscopy technique to study personal lubricants found in sexual assault cases. Using extracts from the lubricant specimen, they were able to define classes and subclasses of these lubricants. The samples used were categorized in groups of three bases: siloxane-, poly(ethylene glycol)-, and glycerin-based lubricants. The infrared results indicated that the base component of each lubricant does not change over the course of three months, and it is possible to identify lubricant traces at least three months after preparing a sample for analysis.

Kimberley et al. (2004) demonstrated the possible use of Raman microspectroscopy to determine the presence of drugs on U.S. dollar bills, which requires minimum sample handling and preserves the sample for further analysis. They were able to measure Raman spectra and identify individual crystals on the currency in a heterogeneous mixture of three compounds: benzocaine, isoxsuprine, and norephedrine.

Documents involving fraud and threatening letters produced on printers and copy machines have been an important concern for forensic investigators. Tracing the model and the origin of the printer or the copy machine is an important hint for crime scene investigators. Merrill et al. (2003) were able to track the origin of a packaged bomb with an address label mailed to a victim's address. Investigators suspected that the bomb was mailed by an employee at a company that had 400 copiers of 61 different models. They requested 61 documents copied from the available 61 models, and they prepared the sample by heating the back of the paper onto aluminum foil attached to a glass microscope slide. Spectra were obtained using microscopic reflection–absorption IR spectroscopy, where the IR beam reflected from the aluminum foil to the detector. Analyzing the documents and comparing the spectra to the ink obtained from the address label allowed the investigators to trace the model of the copier and the suspect.

REFERENCES

Alzate, L. F., et al. (2006). The vibrational spectroscopic signature of TNT in clay minerals. *Vibrat. Spectrosc*., **42**:357–368.

Bartick, E. G. (2002). Application of vibrational spectroscopy in criminal forensic analysis. In: *Handbook of Vibrational Spectroscopy*. Hoboken, NJ: Wiley.

CHAPTER 8

Forensic Serology

RICHARD LI

Department of Sciences, John Jay College of Criminal Justice, The City University
of New York, New York

Summary Forensic serology is an important component of modern forensic science. The primary activity of forensic serologists is the identification of bodily fluids. Bodily fluid stains are commonly associated with violent criminal cases. Proving the presence of bodily fluids can confirm alleged violent acts for an investigation.

8.1 Introduction 270
8.2 Identification of blood 271
  8.2.1 Oxidation–reduction reactions 272
  8.2.2 Microcrystal assays 275
  8.2.3 Other assays for blood identification 275
8.3 Species identification 278
  8.3.1 Immunochromatographic assays 278
  8.3.2 Ouchterlony assay 280
  8.3.3 Crossed-over immunoelectrophoresis 281
8.4 Identification of semen 282
  8.4.1 Visual examination 282
  8.4.2 Acid phosphatase assays 283
  8.4.3 Microscopic examination of spermatozoa 284
  8.4.4 Immunochromatographic assays 285
  8.4.5 RNA-based assays 286
8.5 Identification of saliva 286
  8.5.1 Visual and microscopic examination 287
  8.5.2 Identification of amylase 287
  8.5.3 RNA-based assays 289
References 289

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

8.1 INTRODUCTION

Forensic serology is an important component of modern forensic science. The primary activity of forensic serologists is the identification of bodily fluids. It focuses on identifying the presence of blood, semen, saliva, or other bodily fluids in a questioned sample (DeForest et al., 1983; Gaensslen, 1983; Lee et al., 2001; Shaler, 2002; Greenfield and Sloan, 2005; Jones, 2005; Kobilinsky et al., 2005; Saferstein, 2007; Li, 2008). Bodily fluid stains are commonly associated with violent criminal cases. For example, the identification of blood evidence is often necessary for cases involving homicide, aggravated assault, sexual assault, and burglary. Proving the presence of blood can confirm alleged violent acts for an investigation. The identification of semen and saliva is especially important for the investigation of sexual assault cases. For example, the presence of a suspect's semen stains on a victim's clothing can confirm an alleged sexual activity, and the presence of a suspect's saliva stains on a victim's sex organ can confirm alleged oral copulation. In this chapter we focus on the identification of blood, semen, and saliva evidence.

In certain specific circumstances, the identification of other bodily fluids, although less frequently encountered than blood, semen, and saliva evidence, can have probative value for criminal investigations. The identification of urine can be important for an investigation of alleged assault with urination. The identification of vaginal secretions, menstrual blood, and fecal materials can be important for the investigation of sexual assault cases. For example, the identification of vaginal secretions may be important for the investigation of an alleged vaginal rape with a foreign object. Determining whether bloodstains are attributed to vaginal trauma or menstrual bleeding may be necessary in cases of alleged rape. The identification of fecal samples can be important for cases involving anal rape.

Forensic serology can be considered the process of examining and identifying biological evidence, which occurs prior to the individualization of the biological evidence. *Individualization* of biological evidence is used to determine whether or not a bodily fluid sample has come from a particular person. Today, individualization can be achieved by using forensic DNA analysis. However, the identification of bodily fluid cannot be omitted or replaced by forensic DNA analysis. For example, if a stain from a victim's clothing was processed with forensic DNA analysis for a sexual assault case and the DNA profile of the stain was found to match an alleged suspect, then based on the DNA profiling results alone, one may conclude that the suspect's DNA was found on the victim's clothing, thus, establishing a link between the suspect and the victim. However, if forensic serology testing identified the biological stain as a semen stain, one can conclude that semen was found on the clothing taken from the victim and that the suspect is the source of DNA from the semen stain. Thus, it appears that a sexual act may have occurred.

The identification of bodily fluid can be carried out using presumptive and confirmatory assays to test that the sample is the bodily fluid in question. The advantages of *presumptive assays* are that these assays are sensitive, rapid, and simple. A positive reaction of a presumptive assay indicates the possibility of the presence of the bodily fluid in question. However, presumptive assays are not very specific. Therefore, they should not be considered to be conclusive to the presence of a type of bodily fluid. In contrast, a negative assay suggests that the questioned bodily fluid is absent. Thus, they can be used as a screening method and narrow down biological stains prior to other types of analyses, such as forensic DNA testing. Moreover, these assays can be used as a search method to locate biological stains at the crime scene. Additional assays, such as confirmatory assays, should be conducted afterward if necessary. *Confirmatory assays* are more specific for the bodily fluid in question. These assays are employed to identify bodily fluids with higher certainty than presumptive assays. However, confirmatory assays are more time consuming than presumptive assays.

8.2 IDENTIFICATION OF BLOOD

The identification of blood can be performed using presumptive and confirmative assays (Figure 8.1). The most commonly utilized presumptive assays among forensic laboratories are *oxidation–reduction reaction* assays. The presence of blood can then be confirmed further using confirmatory assays. The most commonly utilized confirmatory assays are *microcrystal assays*.

Currently, since time and budget constraints often exist, confirmatory assays for blood identification are infrequently performed. A reddish-brown stain identified through visual examination is usually tested by using presumptive assays. If the presumptive assay of the alleged blood stain is positive, the stain is then analyzed further by forensic DNA analysis. This approach can only derive a conclusion that the results *indicate* the presence of blood. Therefore, if sufficient amounts of biological materials are available, confirmatory assays should be performed if possible. Thus, if the confirmatory assay is positive, one can conclude that *blood was identified* from the evidence. Additional tests to determine if blood evidence originated from a human or animal source can be performed as necessary and are discussed in Section 8.3.

The blood volume of a normal human is approximately 8% of the body weight. The fluid portion of the blood is called *plasma*. The cellular portion of the blood,

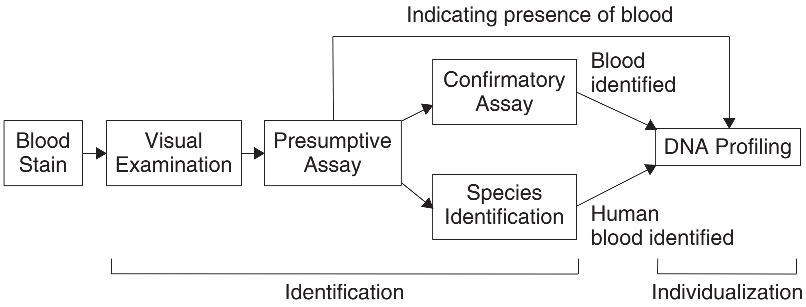

Figure 8.1 Example of the protocol for processing blood samples.

which is suspended in the plasma, consists of *erythrocytes* (also called *red blood cells*), *leucocytes* (also called *white blood cells*), and *thrombocytes* (also called *platelets*). Most presumptive and confirmatory assays of blood identification are based on the detection of hemoglobin. *Hemoglobin* is the protein responsible for the transportation of oxygen and is located in the erythrocytes. The majority of adult human hemoglobin consists of four subunits, two α and two β subunits (designated as $\alpha_2\beta_2$). Each hemoglobin subunit contains a *heme* molecule (also called *ferroprotoporphyrin*). The heme molecule consists of a protoporphyrin IX and a *ferrous* (Fe^{2+}) iron atom (Figure 8.2).

8.2.1 Oxidation–Reduction Reactions

This type of presumptive assay for blood identification is based on the biochemical properties of the heme molecule, which can catalyze an oxidation–reduction reaction. In an oxidation–reduction reaction, changes occur in the oxidation state of chemicals. For example, during the reaction, an oxidant becomes reduced and a reductant becomes oxidized. In the presumptive assays for blood identification, oxidation often coincides with the loss of hydrogen, and reduction often coincides with the gain of hydrogen. An example of an oxidation–reduction reaction for blood identification is shown in Figure 8.3(a).

In an oxidation–reduction reaction for blood identification assays, the most commonly used *reductants* are phenolphthalein, leucomalachite green, tetramethylbenzidine, and luminol (Lee, 1982; Gaensslen, 1983; Sutton, 1999; Greenfield and Sloan, 2005; Laux, 2005; Marie, 2008). Additionally, hydrogen peroxide is usually employed as an *oxidant*. In the presence of heme, a colorless reductant is oxidized, forming a product with color or chemiluminescence. Thus, a positive reaction indicates the possible presence of blood.

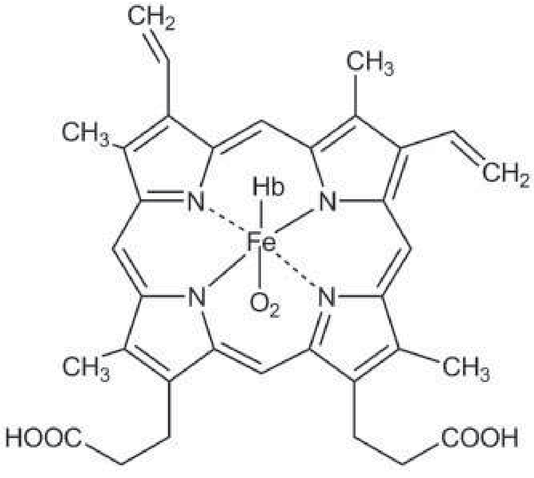

Figure 8.2 Chemical structure of heme. A heme molecule, also called *ferroprotoporphyrin*, consists of a protoporphyrin IX and a ferrous (Fe^{2+}) iron atom.

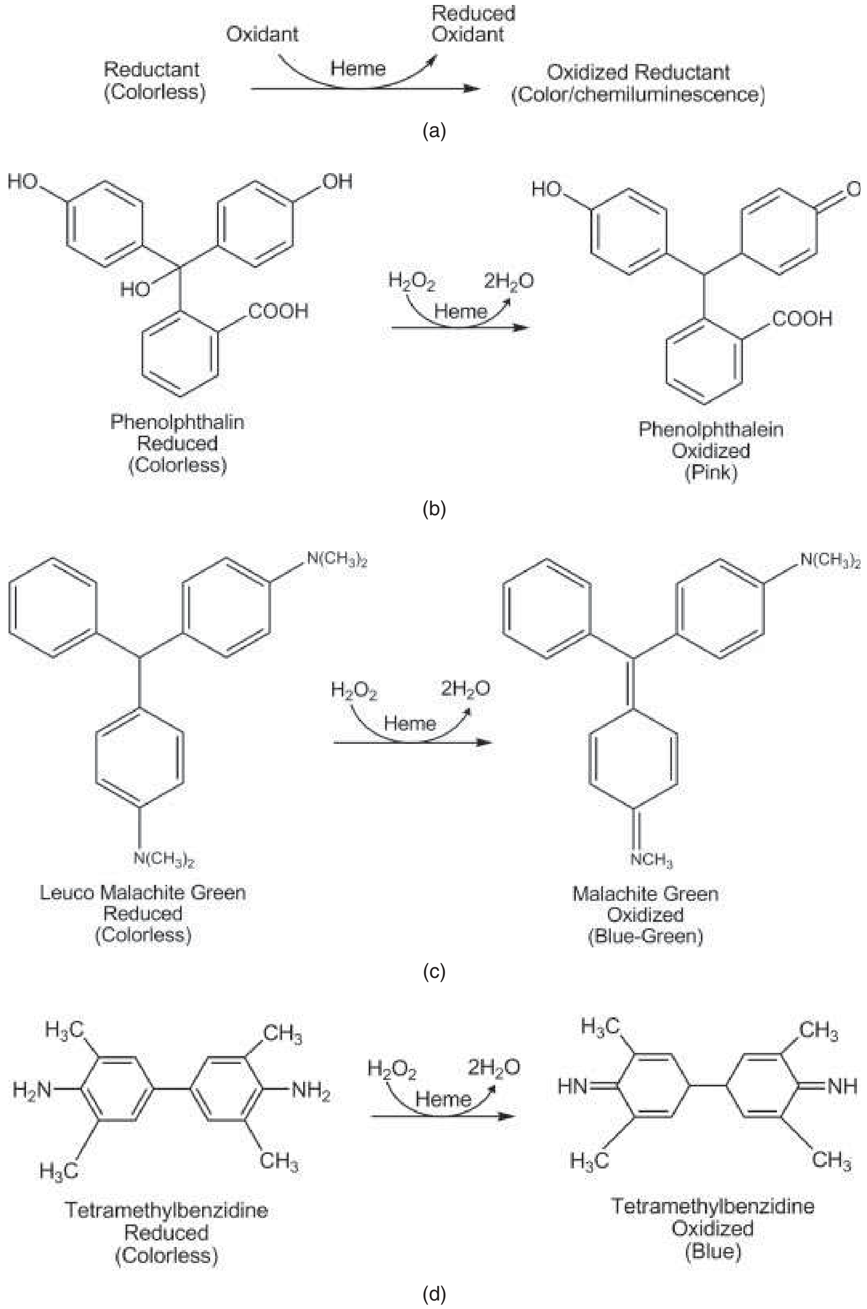
Oxidant
Reduced Oxidant
Reductant (Colorless)
Heme
Oxidized Reductant (Color/chemiluminescence)
(a)
HO
OH
HO
COOH
H_2O_2
$2H_2O$
Heme
HO
O
COOH
Phenolphthalin Reduced (Colorless)
Phenolphthalein Oxidized (Pink)
(b)
$N(CH_3)_2$
H_2O_2
$2H_2O$
Heme
$N(CH_3)_2$
$N(CH_3)_2$
NCH_3
Leuco Malachite Green Reduced (Colorless)
Malachite Green Oxidized (Blue-Green)
(c)
H_3C
CH_3
H_2N
NH_2
H_3C
CH_3
H_2O_2
$2H_2O$
Heme
H_3C
CH_3
HN
NH
H_3C
CH_3
Tetramethylbenzidine Reduced (Colorless)
Tetramethylbenzidine Oxidized (Blue)
(d)

Figure 8.3 Oxidation–reduction reaction and presumptive assays for blood identification: (a) oxidation–reduction reaction as a basis for presumptive assays for blood identification; chemical reactions of (b) Kastle–Meyer assay, (c) leucomalachite green assay, (d) tetramethylbenzidine assay, and (e) luminol assay.

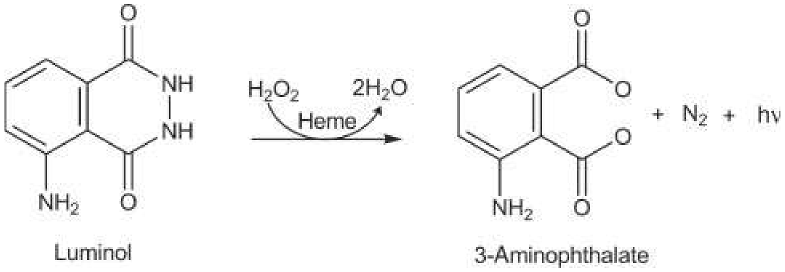

(e)

Figure 8.3 (*Continued*)

For example, the *Kastle–Meyer assay* utilizes colorless phenolphthalin as a reductant. In an oxidation–reduction reaction catalyzed by heme, phenolphthalin can be oxidized to phenolphthalein. Phenolphthalein shows a pink color in alkaline conditions [Figure 8.3(b)]. The *Leucomalachite green assay* utilizes the leuco base form of malachite green, which is colorless. It can be oxidized in the presence of heme to produce malachite green, which shows a green color [Figure 8.3(c)]. The reaction is carried out under acid conditions. In the *tetramethylbenzidine assay*, the oxidation of tetramethylbenzidine can be catalyzed by heme to produce a green color under acidic conditions [Figure 8.3(d)]. The Hemastix assay (Miles Laboratories) is a portable tetramethylbenzidine assay kit that can be used in laboratories and at crime scenes. In a *luminol assay*, oxidation of luminol (3-aminophthalhydrazide), catalyzed by heme, emits a chemiluminescent light [Figure 8.3(e)]. Thus, the assay can be performed in the dark. Luminol reagents can be sprayed at crime scenes to search for alleged bloodstains. A positive reaction not only locates the blood but also detects the patterns, such as footprints, fingerprints, and splatter patterns. This method can help to pinpoint the location of even small traces of blood. Additionally, it is useful for crime scenes that have been cleaned up and where the blood evidence has been tampered with, leaving no visible stains.

The assays discussed above are very sensitive and able to detect blood samples with up to a 10^{-5}-fold dilution. However, they are not blood specific and possibly lead to false-positive results. For example, a false-positive reaction can be caused by the presence of chemicals that are strong oxidants (such as certain metal salts), even in the absence of heme. Plant peroxidases can also catalyze the oxidation reaction in the absence of heme. Thus, peroxidase-containing plants (such as horseradish) can cause a false-positive result. The presumptive assays can be carried out using the swabbing method. A small amount of the questioned stain is transferred onto a moistened cotton swab. Additionally, a substance control (swabbing a nonstained area adjacent to the stained area) can be prepared. The test is conducted in two steps. First, a drop of the testing reagent (reductant) is added onto the swab with the blood sample. If color develops, it is an indication of a false-positive reaction. If no color develops, the assay can be continued by adding a drop of oxidant reagent (e.g., 3% hydrogen peroxide) onto the swab. The presence of blood results in an immediate development of color, while the absence of blood results in no development of color.

8.2.2 Microcrystal Assays

In microcrystal assays, blood crusts from a bloodstain are treated chemically to convert native heme to heme derivatives. These heme derivatives can form crystals with distinctive morphologies. The crystal morphology of heme derivatives can be examined by using microscopic observation. As a result, the presence of the crystals of heme derivatives confirms the presence of blood. One disadvantage of microcrystal assays is that these assays are less sensitive than the presumptive assays for blood identification. The most commonly utilized microcrystal assays are Takayama and Teichmann crystal assays (Lee, 1982; Gaensslen, 1983). Under normal physiological conditions, the ferrous ion of the heme can form six bonds: four bonds with the nitrogen of protoporphyrin IX, one bond with oxygen, and one bond with a hemoglobin subunit. In the *Takayama crystal assay*, blood crusts are treated in an alkaline condition with pyridine and a reducing sugar, such as glucose, that is capable of reducing other chemicals. Following the treatment, pink-colored pyridine ferroprotophorphyrin crystals form and can be observed under a microscope [Figure 8.4(a)]. In the *Teichmann crystal assay*, blood crusts are treated with glacial acetic acid and salts with heating. Brown-colored prismatic crystals of ferriporphyrin chloride form. Ferriporphyrin chloride is a heme derivative in which the iron is in the ferric (Fe^{3+}) state [Figure 8.4(b)].

8.2.3 Other Assays for Blood Identification

Other confirmatory assays for detecting the presence of hemoglobin are available to confirm the presence of blood. For example, human hemoglobin can be detected by using immunological methods with antihuman hemoglobin antibodies. This type of assays is discussed in Section 8.3. Additionally, hemoglobin can be identified using chromatographic and electrophoretic methods based on the mobility characteristics of hemoglobin, and spectrophotometric methods based on the characteristic light spectrum absorbed by hemoglobin.

Recently, RNA-based assays have been developed and can potentially be utilized for the identification of blood (Juusola and Ballantyne, 2003; Nussbaumer et al., 2006). These assays can detect specific types of *messenger RNA* (mRNA) that are only present in erythrocytes. Messenger RNA is transcribed from a gene and is then translated to produce a protein. The erythrocyte-specific genes utilized for blood identification are *HBA1, PBGD*, and *SPTB* (Table 8.1). *HBA1* encodes for the human hemoglobin α1 subunit, which is abundant in erythrocytes. *PBGD* encodes for porphobilinogen deaminase, an erythrocyte-specific isoenzyme of the heme biosynthesis pathway. *SPTB* encodes β-spectrin, which is a subunit of the major protein of the erythrocyte membrane skeleton. These assays are carried out using the *reverse transcription polymerase chain reaction* (RT-PCR) technique to detect erythrocyte-specific mRNA for blood identification. The RT-PCR usually

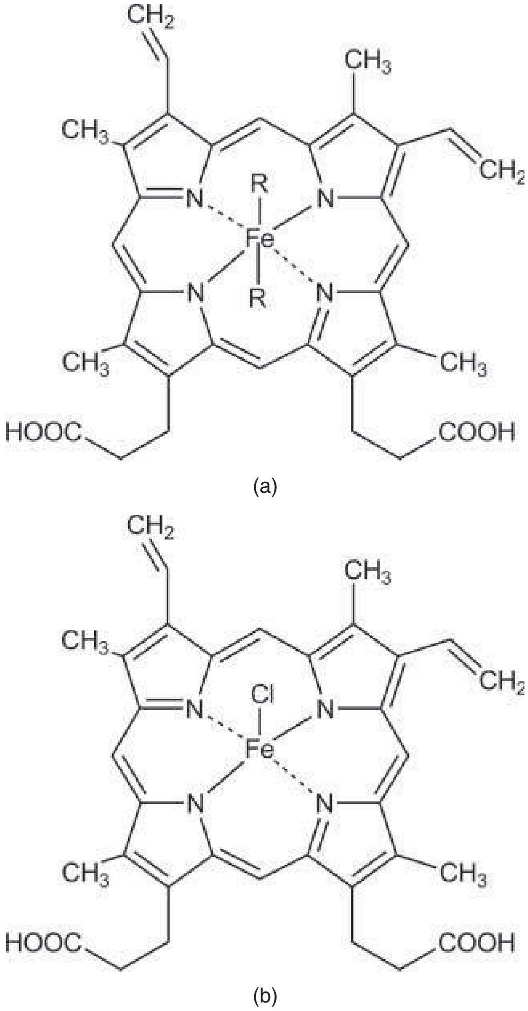

Figure 8.4 Chemical structures of heme derivatives: (a) pyridine ferroprotoporphyrin (R = pyridine) and (b) ferriprotoporphyrin chloride.

consists of two steps (Figure 8.5). In the first step, an mRNA strand is used as a template for the synthesis, carried out by reverse transcriptase, of a DNA strand of complementary sequence, which is referred to as *complementary DNA* (cDNA). In the second step, the resulting cDNA is amplified and the amplified products can then be detected. The advantages of RNA-based assays are that this technique is sensitive and can be used to detect very small quantities of mRNA. Additionally, this technique is specific and can detect erythrocyte-specific mRNA. Moreover, the technique is adaptable to automation. The disadvantage of this technique is that the mRNA present in the sample can be degraded by endogenous ribonucleases.

TABLE 8.1 Use of RNA-Based Assays for Bodily Fluid Identification[a]

Bodily Fluid	Gene Symbol	Gene Product	Description
Blood	*HBA1*	Hemoglobin α1	Hemoglobinα1 subunit abundant in erythrocytes; transports oxygen
	PBGD	Porphobilinogen deaminase	Erythrocyte-specific isoenzyme of the heme biosynthesis pathway
	SPTB	β-Spectrin	Subunit of the major protein of the erythrocyte membrane skeleton
Semen	*KLK3*	Kallikrein 3	Also called prostate-specific antigen (PSA)
	PRM1	Protamine 1	DNA-binding proteins involved in condensation of chromatin of spermatozoa
	PRM2	Protamine 2	DNA-binding proteins involved in condensation of chromatin of spermatozoa
Saliva	*HTN3*	Histatin 3	Histidine-rich protein involved in the nonimmune host defense in the oral cavity
	STATH	Statherin	Inhibitor of the precipitation of calcium phosphate salts in the oral cavity

Source: Adapted from Li (2008).

[a] Tissue-specific genes for identification of bodily fluid are shown.

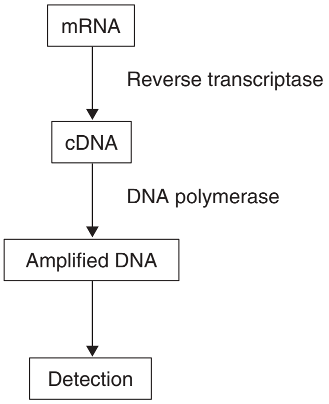

Figure 8.5 RT-PCR technique for bodily fluid identification. The synthesis of cDNA from mRNA can be carried out using reverse transcriptase, and the resulting cDNA is amplified using DNA polymerase.

However, the degree of mRNA degradation, measured by the RT-PCR technique, can potentially be used for determining the age of a bloodstain (Anderson et al., 2005). It is known that mRNA in a blood stain degrades over time after the blood stain is deposited. Thus, the age of a blood stain deposited at a crime scene can be estimated by measuring the detection level of mRNA. This estimation can provide investigative leads, such as when the crime occurred and the length of the postmortem interval.

8.3 SPECIES IDENTIFICATION

Once a stain is identified as blood, the determination of the origin of a blood stain is critical in forensic casework. The blood sample is tested to determine if it is of human or nonhuman origin. Thus, species identification assays can be useful as a screening method to exclude or eliminate nonhuman stains that are not related to the investigation. Thus, if the blood stain is nonhuman, it is not necessary to analyze the stain further. However, if the case involves animals, it may be necessary to identify the species of the blood sample. Currently, species identification is performed infrequently in forensic laboratories. A blood stain is usually processed for DNA profiling analysis. The DNA isolated is quantified using techniques that detect only higher-primate DNA. Thus, the detection of the DNA can indicate that the donor of a blood stain is of human origin, based on the assumption that the crimes involving primate blood are extremely rare in the United States.

The majority of assays for species identification use antibody-based techniques. *Antibodies* are capable of binding to *antigens* (substances that trigger the generation of antibodies). For example, antihuman antibodies are antibodies bound to human antigens and can be used to determine if the sample is a human blood sample. To produce antihuman antibodies, human antigens are usually introduced into a host animal. The host animal generates specific antibodies against the human antigens. The blood is then removed from the host animal to prepare antibodies. The resulting antibody is a *polyclonal* antihuman antibody containing a mixture of antibodies against various human antigens. Additionally, purified human proteins can be used to generate *monoclonal* antibodies. Monoclonal antibodies are specific, homogeneous, and react with a single determinant site of the antigen. The antihuman antibodies utilized should not cross-react with commonly encountered animals but usually have cross-reactivity with higher primates. However, this is not a great concern, due to the fact that crimes involving primates are very rare. An antibody against animal antigens can also be produced using a similar method for determining the animal species in question.

One commonly employed assay is the immunochromatographic assay (Table 8.2). Additional assays that can be utilized are the Ouchterlony assay and the crossed-over immunoelectrophoresis assay, which relies on the binding of an antigen to an antibody causing the formation of a visible precipitate.

8.3.1 Immunochromatographic Assays

A schematic immunochromatographic membrane device is shown in Figure 8.6. The mechanism of the immunochromatographic assays is illustrated in Figure 8.7. The samples can be prepared by cutting out a portion of a stain or scraping off stains from a surface. The sample is then extracted with a small volume of buffer. The extracted sample is loaded into the sample well of the immunochromatographic device. The loaded sample diffuses across the nitrocellulose membrane. The presence of antigens in the sample results in a colored line at the test zone. The immunochromatographic device also utilizes a control zone to ensure that the

TABLE 8.2 Commonly Used Immunochromatographic Assays for Forensic Application

Forensic Application	Assay	Antigen
Blood identification, species identification	ABAcard HemaTrace (Abacus Diagnostics)	Hemoglobin
	RSID-Blood (Independent Forensics)	Glycophorin A
Semen identification	ABAcard p30 (Abacus Diagnostics)	Prostate-specific antigen
	RSID-Semen (Independent Forensics)	Semenogelin
Saliva identification	RSID-Saliva (Independent Forensics)	Human salivary α-amylase

Source: Adapted from Li (2008).

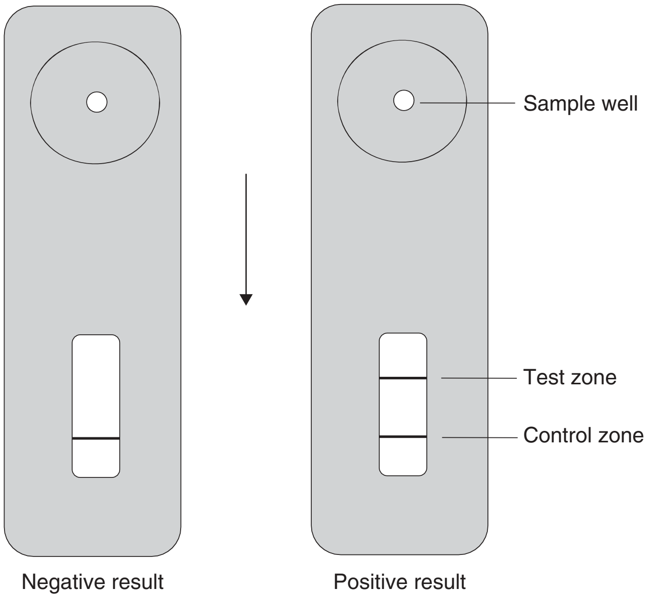

Figure 8.6 Immunochromatographic device. Examples of negative and positive results are shown.

sample has diffused properly along the test strip. Thus, the presence of antigens results in a line at the control zone as well. The assay is considered valid only if the line in the control zone is observed. In contrast, the absence of antigens results in a line in the control zone only. However, a false-negative result may occur in a sample consisting of a very high concentration of antigens. This false-negative result is known as the *high-dose hook effect*. Nevertheless, immunochromatographic assays are rapid, specific, and sensitive and can be used in both laboratory and field tests for species identification.

Most of these assays are based on the detection of human erythrocyte proteins (Lee, 1982; Gaensslen, 1983; Greenfield and Sloan, 2005). For example, purified

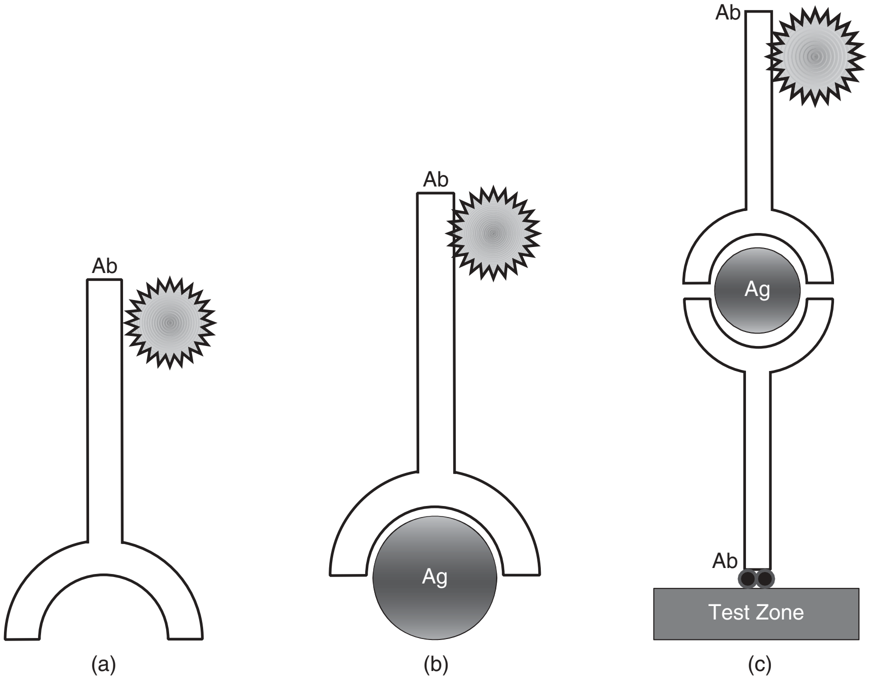

Figure 8.7 Immunochromatographic assay. (a) Dye-labeled antibody (Ab) is placed in the sample well. (b) Sample containing antigen (Ag) is loaded into the sample well and binds to the labeled antibody. This antigen–antibody complex diffuses along the immunochromatographic membrane device. (c) The antigen–antibody complex is captured by an immobilized antibody at the test zone, which forms a colored line for observation.

human hemoglobin can be used to generate antihuman hemoglobin antibodies that detect human hemoglobin. Commercially produced immunochromatographic kits are available, such as the ABAcard HemaTrace kit (Abacus Diagnostics, California). The range of normal blood hemoglobin concentration is 121 to 165 mg/mL. The ABAcard HemaTrace kit can detect very low levels of hemoglobin (as low as 0.07 μg/mL), and specificity studies have shown that the kit is specific for higher primates, including humans. Another immunochromatographic assay utilizes antibodies, recognizing human *glycophorin A* (GPA). Glycophorin A is a human erythrocyte membrane antigen. The RSID-Blood kit (Independent Forensics, Hillside, IL), a commercially produced glycophorin A immunochromatographic kit, is available. The sensitivity of the RSID-Blood kit can be as low as 100 nL of human blood. Species specificity studies have shown that there is no cross reactivity with various animal species, including nonhuman primates.

8.3.2 Ouchterlony Assay

This assay can be performed in an agarose gel supported by a glass slide or polyester film (Lee, 1982; Greenfield and Sloan, 2005). Wells are created by punching holes in the gel layer at desired locations. For example, a pattern with six wells surrounding a center well may be used. An antibody (such as the antihuman antibody) is loaded into the central well with the stain extracts and controls are loaded into the surrounding wells. The diffusion of the antigen and the antibody from the wells is allowed to occur during incubation. In a positive reaction, a line of precipitate forms between each antigen well and the antibody well (Figure 8.8). The gel can then be stained with a dye to enhance the visibility of the line of precipitate. This assay can also analyze more than one antigen in the same assay to determine if the samples in question are from the same or different species of origin.

8.3.3 Crossed-Over Immunoelectrophoresis

Crossed-over immunoelectrophoresis is also known as counterimmunoelectrophoresis (Lee, 1982; Greenfield and Sloan, 2005). In this method, an agarose gel is prepared which contains sample wells opposing each other (Figure 8.9). The antibody and questioned samples are loaded into the wells by pairs. The antihuman antibody is loaded into the wells proximal to the anode, and questioned samples are loaded into the wells proximal to the cathode. The gel electrophoresis is then carried out. During the electrophoresis, the antigen present in a question sample, carrying a negative charge, migrates toward the anode. In contrast, the antibody migrates toward the cathode as a result of flow of fluid. In a positive reaction, a precipitate line forms between the opposing antigen and antibody wells.

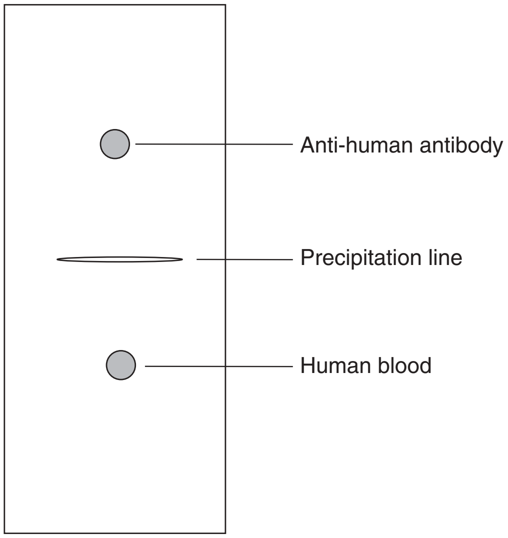

Figure 8.8 Result of an Ouchterlony assay. A precipitate line is observed between a human blood sample and an antihuman antibody.

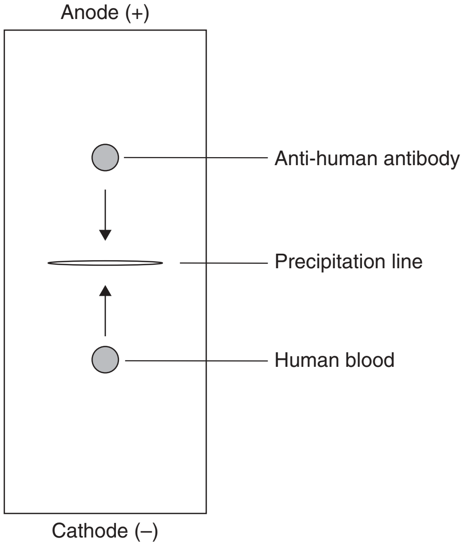

Figure 8.9 Result of a crossover electrophoresis. A precipitate line is observed between a human blood sample and an antihuman antibody.

8.4 IDENTIFICATION OF SEMEN

Normal semen contains seminal fluid and sperm cells (also called *spermatozoa).* The seminal fluid is a mixture of secretions from male accessory glands largely from the seminal vesicles and prostate. Spermatozoa count ranges from 10^7 to 10^8 (spermatozoa/mL of semen) among healthy males. Abnormality in spermatozoa count can be observed among males. For example, *oligospermia* refers to the condition of a male with an abnormally low spermatozoa count. *Aspermia* refers to the condition of a male with no spermatozoa. However, the secretion of the seminal fluid is not affected among these males. A *vasectomy* is a surgical procedure that blocks spermatozoa from reaching the distal portions of the male reproductive tract. A vasectomized male can still produce seminal fluid.

To search for seminal stains on evidence, visual examination of evidence can be facilitated by utilizing alternate light sources. The presence of semen can then be indicated by presumptive assays that detect enzymes present in seminal fluid. The presence of semen can be confirmed by performing confirmatory assays. The most commonly used confirmatory assays are microscopic examination of spermatozoa and immunochromatographic assays that utilize antibody-detecting antigens present in seminal fluid. Recently, RNA-based assays have been developed and are potentially useful for semen identification.

8.4.1 Visual Examination

The visual examination of semen stains can be facilitated by using alternate light sources (Jones, 2005). Semen stains fluoresce when irradiated with alternate light sources. For example, excitation wavelengths from 450 to 495 nm can be used. Fluorescence emitted from a seminal stain can be visualized with colored goggles. Thus, a fluoresced stain indicates the presence of semen and can be tested further using acid phosphatase assays.

8.4.2 Acid Phosphatase Assays

Semen contains high concentrations of *acid phosphatase*, which is secreted primarily from the prostate gland. Thus, acid phosphatase is a useful marker for the identification of semen (Gaensslen, 1983; Greenfield and Sloan, 2005; Jones, 2005). The most commonly used acid phosphatase assays in forensic laboratories are *colorimetric* and *fluororimetric* assays. These assays are based on the principle that acid phosphatase can catalyze a hydrolysis reaction to remove a phosphate group from a substrate.

The acid phosphatase is water soluble and can be transferred from a stain by rubbing the questioned stain area with a moistened cotton swab. The swab can then be used for acid phosphatase assays. The most commonly used colorimetric assay for forensic applications is the use of α-*naphthyl phosphate* as the substrate [Figure 8.10(a)]. The substrate reagent is added to the swab prepared as described above. In the presence of acid phosphatase, the hydrolysis of α-naphthyl phosphate

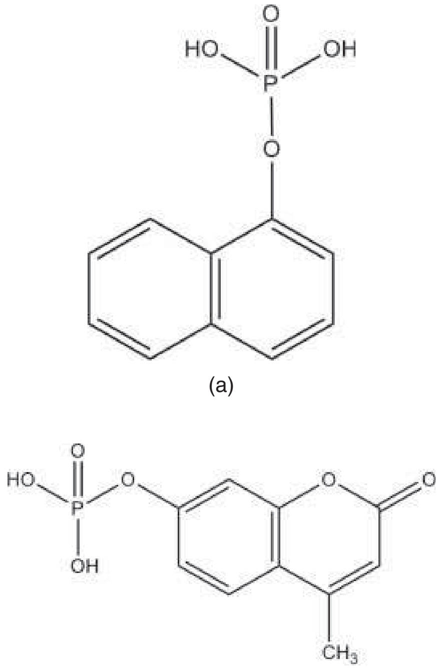

(a)

(b)

Figure 8.10 Chemical structures of acid phosphatase substrates: (a) α-naphthyl phosphate and (b) 4-methylumbelliferone phosphate.

occurs. By subsequently adding brentamine Fast Blue B solution, a purple precipitate forms at the sites of acid phosphatase activity. The immediate color change to purple indicates the presence of semen.

The fluororimetric assay of acid phosphatase detection is highly sensitive and is generally used for mapping semen stains located on evidence. A piece of moistened filter paper is used to transfer the acid phosphatase from the evidence. The moistened filter paper is overlaid in close contact with the evidence, such as a garment. The filter paper is then lifted from the evidence. Background fluorescence on the filter paper is examined in a dark room with long-wave ultraviolet (UV) light. Any background fluorescence should be marked. The filter paper can then be sprayed with the *4-methylumbelliferone phosphate* reagent [Figure 8.10(b)]. In the presence of acid phosphatase, the hydrolysis of the phosphate residue on 4-methylumbelliferone phosphate occurs, which immediately generates fluorescence under UV light. Semen stains can be visualized as fluorescent areas on the filter paper.

The presumptive assays described above cannot distinguish completely between prostatic acid phosphatase and non-prostatic acid phosphatase *isoenzymes* (multiple forms of acid phosphatase). For example, interference due to contamination with acid phosphatase present in vaginal secretions can occur in specimens collected from victims. Thus, the presence of semen should be confirmed by the performance of confirmatory assays.

8.4.3 Microscopic Examination of Spermatozoa

To prove the presence of semen, light microscopic identification of spermatozoa can be carried out (Jones, 2005). The morphology of human spermatozoa can be characterized as the head, the middle piece, and the tail structures (Figure 8.11). The head contains a nucleus and an acrosomal cap (a membranous compartment at the tip of the head). The head is attached to the midpiece, where the mitochondria are located. The midpiece is attached to the tail, which is a flagellum responsible for the motility of a spermatozoon.

To prepare a sample for microscopic examination, it is necessary to transfer spermatozoa from a questioned stain to a microscope slide. A small portion of a stain is cut and extracted with water followed by gentle vortexing. This suspension containing spermatozoa is then placed onto a slide and is air-dried with low heat. To facilitate the microscopic observation of spermatozoa, a histological staining method, using the *Christmas tree stain*, is utilized to stain the spermatozoa held on the slide. This staining technique utilizes a combination of two dyes. Nuclear fast red is a red dye that stains the nucleus and acrosomal cap of the spermatozoa. Picroindigocarmine is a green dye that stains the neck and tail portions of the sperm. As a result, the nucleus and acrosomal cap are red, while the sperm tails and midpiece are green. In addition to spermatozoa, a victim's epithelial cells can often be observed if the sample is collected from the victim. The epithelial cells are stained blue-green with red nuclei. Today, the spermatozoa from such a slide can be separated from the victim's epithelial cells by utilizing the *laser capture*

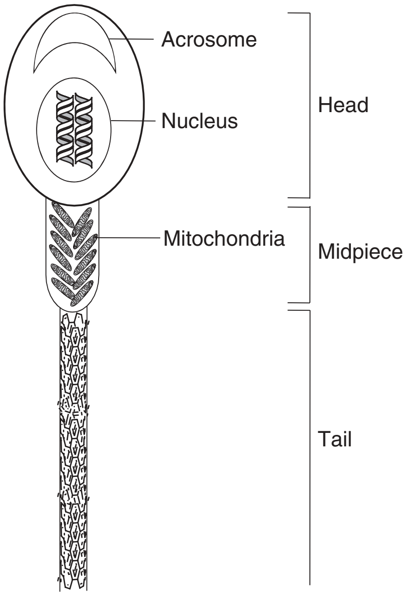

Figure 8.11 Structure of a spermatozoon.

microdissection technique. After the spermatozoa are located on the slide, the head portion of the spermatozoa can be captured. The DNA of the spermatozoa can then be isolated for forensic DNA analysis.

8.4.4 Immunochromatographic Assays

These assays (Jones, 2005) utilize antibodies specific for proteins present in semen such as the *prostate-specific antigen* (PSA) and *seminal vesicle–specific antigens*. PSA (also known as P30 and Kallikrein 3) is a serine protease with a molecular weight of 30 kDa. PSA is produced in the prostate gland and secreted into the seminal fluid. Low levels of PSA can also be found in blood, fecal material, milk, sweat, and urine. Therefore, it can be a concern when a questioned sample is contaminated with those biological materials. Commercially produced immunochromatographic kits are available, such as the ABAcard p30 (Abacus Diagnostics, California). In this device, antihuman PSA antibodies are utilized. The assay is carried out by loading an extracted sample into the sample well. The presence of human PSA results in a pink line at both the test zone and the control zone, while the absence of human PSA results in a pink line in the control zone only. However, the high-dose hook effect, which may cause false-negative results, as described previously, occurs when high quantities of seminal fluid are tested.

Seminal vesicle–specific antigens (also known as *semenogelins*) are the major seminal vesicle–secreted proteins in semen and form a coagulum upon ejaculation. In humans, semenogelin I (SgI) and semenogelin II (SgII) are two major forms of semenogelins. They are present primarily in tissues of the male reproduction system and secreted into seminal fluid. The seminal semenogelin concentration is much higher than seminal PSA concentration. Therefore, the semenogelin test can achieve higher sensitivity than the PSA test. Additionally, semenogelin I and II are present in low amounts in several other tissue types, such as skeletal muscle, kidney, colon, and trachea tissues. However, this is not a great concern since these tissue samples are usually not analyzed for semen detection.

Commercially produced immunochromatographic kits are available, such as the RSID-Semen kit (Independent Forensics, Hillside, IL), in which antisemenogelin antibodies are utilized. The assay is carried out by loading an extracted sample into the sample well. The presence of semenogelins results in a pink line at both the test zone and control zone, while the absence of semenogelins results in a pink line in the control zone only. The RSID-Semen kit can detect seminal fluid with 5×10^4-fold dilution. However, the high-dose hook effect occurs when high amounts of seminal fluid are tested.

8.4.5 RNA-Based Assays

Recently, RNA-based assays have been developed for the identification of semen (Juusola and Ballantyne, 2003; Nussbaumer et al., 2006). These assays can detect specific mRNA that is present exclusively in the spermatozoa and the male accessory glands. For example, the tissue- or cell-specific genes utilized for seminal fluid identification are *KLK3, PRM1*, and *PRM2*. *KLK3* encodes for human PSA. *PRM1* and *PRM2* encode for protamine 1 and protamine 2, respectively. Protamines are DNA-binding proteins involved in the condensation of chromatin in spermatozoa.

8.5 IDENTIFICATION OF SALIVA

The human salivary glands produce 1.0 to 1.5 L of saliva a day. Saliva is a fluid composed mostly of water, but also contains small amounts of electrolytes, proteins, and enzymes. Salivary amylase is one of the enzymes that can be found in saliva. The biochemical function of the amylase is to cleave polysaccharides, such as starch, which is a major component of one's diet. Thus, the identification of the amylase indicates the presence of saliva. Two major amylase isoenzymes (multiple forms of amylase) can be found in humans: human *salivary α-amylase* (HSA) and human *pancreatic α-amylase* (HPA). HSA is produced in the salivary glands and HPA is produced in the pancreas. In addition to saliva, a low level of amylase activity is found in other bodily fluids, such as blood, milk, perspiration, semen, tears, and vaginal secretions.

Visual examinations using alternate light sources can facilitate the search for saliva stains (Jones, 2005). The identification of saliva is carried out by detecting amylase in a questioned sample (Gaensslen, 1983; Greenfield and Sloan, 2005; Jones, 2005). These assays can be divided into presumptive and confirmatory assays. Presumptive assays measure the enzymatic activity of amylase. These assays are not HSA specific and cannot distinguish HSA and its isoenzymes. They are not conclusive as to the specific presence of saliva in a sample. False positives can occur due to contamination by HPA and amylases from plant and bacterial sources. The confirmatory assays, including direct detection of HSA proteins, are more specific than the presumptive assays. Additionally, RNA-based assays (Juusola and Ballantyne, 2003) can potentially be useful as a confirmatory assay for saliva identification.

8.5.1 Visual and Microscopic Examination

The search for saliva stains can be facilitated by using alternate light sources. Saliva stains fluoresce when irradiated with alternate light sources. For example, a 470-nm excitation wavelength can be used. Fluorescence emitted from a saliva stain can be visualized with colored goggles. However, the intensity of fluorescence emitted from saliva stains is less than that of seminal stains. Furthermore, a microscopic examination can be carried out to identify buccal epithelial cells from the stain. The presence of buccal epithelial cells indicates a saliva stain.

8.5.2 Identification of Amylase

Starch–Iodine Assay This assay is based on the principle that iodine (I_2) reacts with starch and develops a dark blue color. In the presence of amylase, starch is cleaved to monosaccharides or disaccharides. The monosaccharides or disaccharides do not react with iodine to develop color. The assay can be carried out in a starch-containing agarose gel with sample wells. The questioned sample is then loaded into the sample well. The gel is incubated and then stained with an iodine solution. The starch-containing gel is stained blue. If amylase is present in the sample, it diffuses out from the sample well and cleaves starch in the gel as it diffuses. A clear area around the sample well indicates amylase activity. The size of the clear area is proportional to the amount of amylase present in the sample. Therefore, the amount of amylase present in a questioned sample can be estimated by comparing it to a standard curve (Figure 8.12).

Phadebas Assay Phadebas (Pharmacia) reagent is an insoluble blue dye–labeled amylase substrate commonly used in forensic laboratories. In the presence of amylase, the dye motif of substrates can be cleaved. The cleaved dye motif is soluble in water and develops a blue color. The amylase reaction can be terminated at an alkaline pH by adding a sodium hydroxide solution. The degree of coloration can be measured (as optical density) at 620 nm using a spectrophotometer and is proportional to the amount of amylase. Then the number of amylase units present in a sample can be determined by comparing the degree of coloration to a standard curve.

Amylase tests can also be used to locate possible saliva stains on evidence through the process of amylase mapping. To perform this process, a piece of filter paper containing Phadebas reagent is placed over the evidence to be tested. The

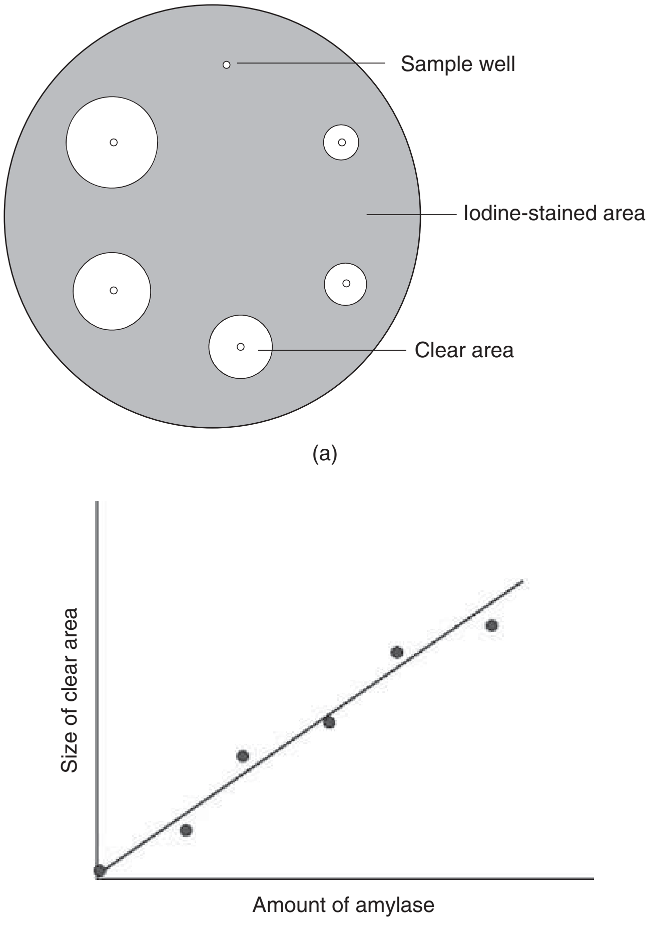

Figure 8.12 Starch–iodine assay for identification of amylase. (a) A series of known concentrations of amylase standards are applied. (b) The size of the clear area resulting from amylase activity is plotted and a liner standard curve (in log scale) can be constructed. The standard curve can be used to determine the amount of amylase present in a questioned sample.

paper is dampened slightly by spraying it with distilled water. The amylase is water soluble and can be transferred from evidence to the moistened filter paper. The amylase transferred to the filter paper reacts with the Phadebas reagent and develops a blue color which corresponds to the site of the saliva stain.

Immunochromatographic Assays Commercially produced immunochromatographic kits are available, such as the RSID-Saliva kit (Independent Forensics). With this device, anti-HSA antibodies are utilized. The assay is carried out by loading an extracted sample into the sample well. The presence of HSA results in a pink line at both the test zone and the control zone, while the absence of HSA results in a pink line in the control zone only. The sensitivity of the RSID-Saliva kit can be as low as 1 μL of saliva.

8.5.3 RNA-Based Assays

Recently, RNA-based assays have been developed for the identification of saliva. These assays can detect specific mRNA that is present exclusively in cells of the oral cavity. Specifically, the tissue- or cell-specific genes utilized for saliva identification are *HTN3* and *STATH*. *HTN3* encodes histatin 3, which is a protein involved in the nonimmune host defense in the oral cavity. *STATH* encodes statherin, which is a salivary protein inhibiting the potentially harmful precipitation of calcium phosphate salts in the salivary glands and the oral cavity.

REFERENCES

Anderson, S., Howard, B., Hobbs, G. R., and Bishop, C. P. (2005). A method for determining the age of a bloodstain. *Forensic Sci. Int.,* **148**:37–45.

DeForest, P., Gaensslen, R., and Lee, H. C. (1983). *Forensic Science: An Introduction to Criminalistics*. New York: McGraw-Hill.

Gaensslen, R. E. (1983). *Sourcebook in Forensic Serology, Immunology, and Biochemistry*. Washington, DC: U.S. Government Printing Office.

Greenfield, A., and Sloan, M. (2005). Identification of biological fluids and stains. In: James, S., and Nordby, J. (Eds.), *Forensic Science: An Introduction to Scientific and Investigative Techniques*, 2nd ed. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Jones, E. (2005). The identification of semen and other body fluids. In: Saferstein, R. (Ed.), *Forensic Science Handbook*, Vol. 2. Englewood Cliffs, NJ: Pearson Prentice Hall.

Juusola, J., and Ballantyne, J. (2003). Messenger RNA profiling: a prototype method to supplant conventional methods for body fluid identification. *Forensic Sci. Int.,* **135**:85–96.

Kobilinsky, L., Liotti, T., and Oeser-Sweat, J. (2005). *DNA: Forensic and Legal Applications*. Hoboken, NJ: Wiley.

Laux, D. L. (2005). The detection of blood using luminol. In: James, S., Kish, P., and Sutton, T. (Eds.), *Principles of Bloodstain Pattern Analysis: Theory and Practice*. Boca Raton, FL: CRC Press, Francis & Talor Group, pp. 369–389.

Lee, H. C. (1982). Identification and grouping of bloodstains. In: Saferstein, R. (Ed.), *Forensic Science Handbook*. Englewood Cliffs, NJ: Prentice Hall.

Lee, H. C., Palmbach, T., and Miller, M. T. (2001). *Henry Lee's Crime Scene Handbook*. San Diego, CA: Academic Press.

Li, R. (2008). *Forensic Biology*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Marie, C. (2008). Presumptive testing and enhancement of blood. In: Bevel, T., and Gardner, R. (Eds.), *Bloodstain Pattern Analysis*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

CHAPTER 9

Forensic DNA Analysis

HENRIETTA MARGOLIS NUNNO

Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Summary In this chapter we describe how DNA became a valuable forensic tool in identifying the source of physical evidence left at a crime scene. The use of restriction fragment length polymorphism analysis in the mid-1980s was replaced by polymerase chain reaction (PCR) methods, which were more sensitive, requiring far less high-molecular-weight DNA; used less hazardous materials; and were more rapid and more economical. PCR-short tandem repeat–based genetic profile typing methods have improved in sensitivity over the past 20 years and have become a basic tool in the crime lab. Where nuclear DNA is insufficient to generate a full genetic profile, mitochondrial DNA can be used to provide identifying information. The chapter also covers low-copy-number procedures and the typing of single-nucleotide polymorphisms within the human genome.

9.1 Introduction 292
    9.1.1 Background on DNA typing 292
    9.1.2 DNA structure 294
    9.1.3 Nuclear and mitochondrial DNA organization 295
9.2 Methodology 296
    9.2.1 Sample collection and DNA extraction 296
    9.2.2 DNA quantification 297
    9.2.3 Polymerase chain reaction 298
    9.2.4 Short tandem repeats 298
    9.2.5 PCR of STRs 300
    9.2.6 Separation and sizing of STR alleles 301
    9.2.7 Combined DNA index system (CODIS) database 305
    9.2.8 Frequency and probability 306
9.3 Problems encountered in STR analysis 307
    9.3.1 Low-copy-number DNA 307

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

9.3.2 Degraded DNA and reduced-size (mini) STR primer sets 308
9.3.3 PCR inhibition 310
9.3.4 Interpretation of mixtures of DNA 310
9.3.5 Null alleles and allele dropout 311
9.3.6 Factors causing extra peaks in results observed 312
9.3.7 Stutter product peaks 312
9.3.8 Nontemplate addition (incomplete adenylation) 313
9.3.9 Technological artifacts 313
9.3.10 Single-nucleotide polymorphism analysis of autosomal DNA SNPs 313
9.3.11 Methods used for SNP analysis 314
9.3.12 Mitochondrial DNA analysis 315

9.4 Methodology for mtDNA analysis 316
9.4.1 Preparation of samples 316
9.4.2 MtDNA sequencing methods 316
9.4.3 Reference sequences 317
9.4.4 Screening assays for mtDNA 318
9.4.5 Interpretation of mtDNA sequencing results 319
9.4.6 Statistics: the meaning of a match for mtDNA 320
9.4.7 Heteroplasmy 320
9.4.8 The future of DNA analysis 321

References 322

9.1 INTRODUCTION

9.1.1 Background on DNA Typing

The structure of DNA was first elucidated by Watson and Crick in 1953. It was shown to take the form of a double helix made by two parallel strands, each of which is composed of four different nucleotides (Figure 9.1) linked together in a specific sequence (Watson and Crick, 1953). While the sequence of the DNA nucleotides is virtually identical in every cell in a person's body, different people differ in portions of their DNA sequences, and each sequence difference is known as a *polymorphism*. In addition, certain sequences of nucleotides are repeated numerous times, frequently one after another, and because people may differ in the number of times that each of these tandem sequence repeats is repeated, there are even more polymorphisms between individuals. It is the examination of such a DNA sequence or repeat number (length) polymorphism that is the basis of DNA identification.

DNA profiling or *typing* (originally called *DNA fingerprinting*) was first used for forensic purposes in the 1980s. The methodology, first described by Alec Jeffreys, used the differences in length of DNA regions created by variations in numbers of repeated sequences to distinguish between individuals (Jeffreys et al., 1985). The repeated regions of DNA that Jeffreys examined are called *variable number tandem repeats* (VNTRs) and the DNA profiling technique that he developed is called

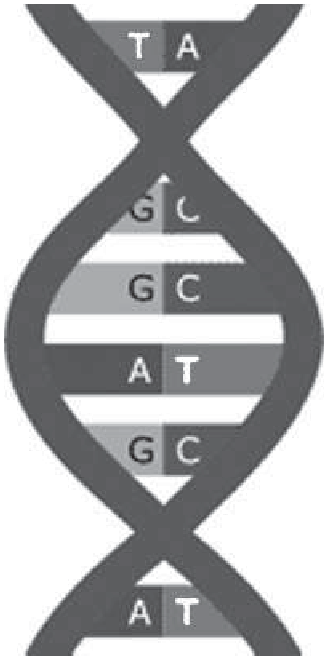

Figure 9.1 DNA structure. The two parallel strands of DNA take the form of a double helix, with each strand composed of four different nucleotides. A is adenine, T is thymine, G is guanine, and C is cytosine.

restriction fragment length polymorphism (RFLP). RFLP uses restriction enzymes, which cut the DNA at specific sequences of bases called *restriction endonuclease recognition sites*, to chop up large pieces of double-stranded DNA in a specific manner. RFLP identifies people by the pattern their DNA segments make when they are separated by size and then hybridized with complementary DNA probes. Because different people may differ in the particular endonuclease recognition sites that are present in their DNA and may also differ in the number of repeats that are present between their recognition sites, the length and therefore the pattern of the various DNA fragments produced by the restriction enzyme digestion will differ from person to person when separated by size.

After the polymerase chain reaction (PCR) was developed in the mid-1980s (Saiki et al., 1985; Mullis, 1990), DNA typing methods began incorporating PCR technology and RFLP was phased out for forensic DNA identification. Because PCR-based methods increase the copy number of DNA fragments to the millions or billions, they greatly reduce the amount of crime scene DNA necessary for DNA identification. In addition, because PCR-based methods are used to amplify relatively small DNA fragments, there is no need for the undegraded DNA required for RFLP. Although the first PCR-based tests used for human identification did not have a high power of discrimination, over the years DNA typing technology has evolved into the current PCR-based methods, which examine short tandem repeat (STR) loci and have powers of discrimination of one in more than a trillion.

9.1.2 DNA Structure

In the Watson and Crick model of DNA structure, each of the two parallel strands forming the double helix is made of linked nucleotides. Each nucleotide contains a phosphate group, a five-carbon sugar (deoxyribose), and one of four nitrogen-containing bases: adenine, thymine, guanine, or cytosine, referred to as A,T, G, and C, respectively. A and G are double-ringed purines and T and C are single-ringed pyrimidines (Figure 9.2). The nucleotides in DNA are linked by bonds (phosphodiester) between the 3′-hydroxyl and 5′-hydroxyl groups of the deoxyribose sugars and the phosphate groups, forming a sugar phosphate backbone on the outside of the double helix [Figure 9.3(a)]. The bases project into the inside and each strand is connected to what is called its *complementary strand* by hydrogen bonding between specific bases. In this complementary base pairing, adenine always pairs with thymine (by forming two hydrogen bonds) and guanine always pairs with cytosine (by forming three hydrogen bonds) (Figure 9.3). There are about 10 nucleotide pairs for each turn of the helix. Polynucleotides are read from the 5′ end to the 3′ end; since the two polynucleotide chains of the helix run in opposite 5′-to-3′ directions, they are said to be *antiparallel* [Figure 9.3(b)].

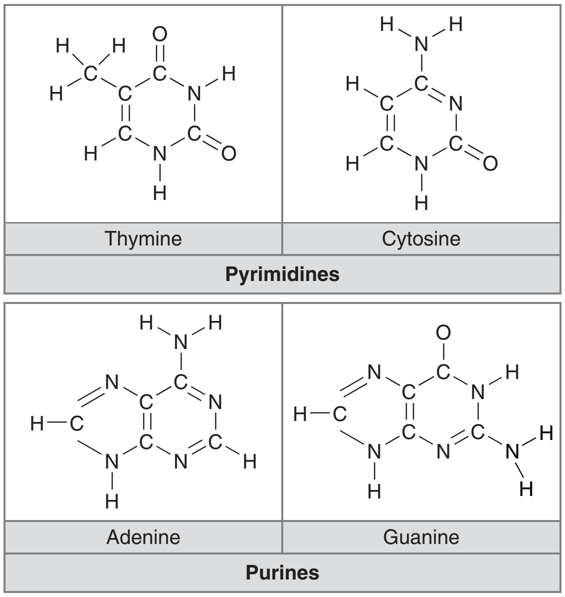

Figure 9.2 Pyrimidines and purines in DNA. The single-ring pyrimidines, thymine (T) and cytosine (C), are shown in the upper panel, and the double-ring purines, adenine (A) and guanine (G), are shown in the lower panel.

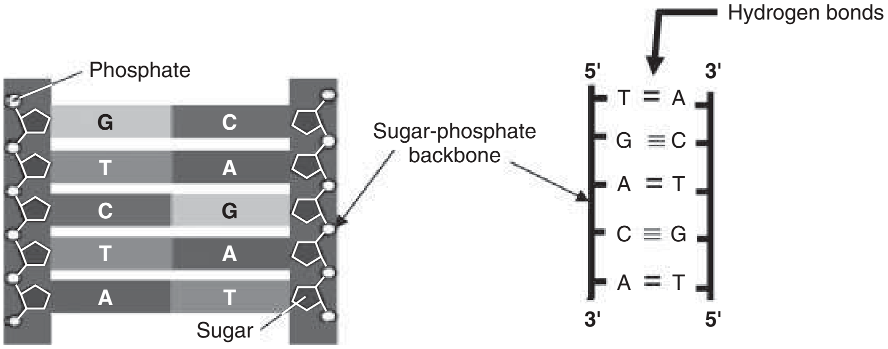

Figure 9.3 Complementary base pairing in DNA structure. (a) A sugar phosphate backbone is on the outside and the four different bases are on the inside of the DNA molecule (left panel). (b) There is base pairing by hydrogen bonding between complementary bases A and T (with two hydrogen bonds) and G and C (with three hydrogen bonds) and both strands run from 5′ to 3′ in opposite directions.

9.1.3 Nuclear and Mitochondrial DNA Organization

Within most human cells DNA is present in both the nucleus and the cytoplasm; cytoplasmic DNA is found in energy-producing organelles called *mitochondria*. Although the DNA found in these two different locations has the same basic double-stranded structure, many differences exist as well.

Nuclear DNA molecules form linear chromosomes and mitochondrial DNA (mtDNA) molecules form circular chromosomes. Whereas nuclear chromosomes are present as pairs (homologs), with one chromosome of each pair received from each parent, mtDNA is inherited uniparentally only through the maternal line. While the nucleus of a typical human cell contains 46 chromosomes made up of 22 pairs of nonsex chromosomes (autosomes) and two sex chromosomes, the total number of mitochondria in a cell varies greatly with the type of cell and its stage of development. Although a set of 23 nuclear chromosomes contains about 3.1 billion nucleotides, the total number of nucleotides in a single mitochondrial chromosome is only 16,569.

Both nuclear and mitochondrial chromosomes contain coding sequences (nucleotides that make up the genes) and noncoding nucleotide sequences whose function is largely unknown. But while the coding and noncoding sequences are distributed intermittently along the length of each nuclear chromosome, in the mitochondrial genome the coding and noncoding areas are separate from each other. The noncoding portion of mtDNA is known as the *control region* (also known as the *D-loop* or *displacement loop*) and it contains about 1100 base pairs (bp), which encompass the two polymorphic hypervariable regions (HV1 and HV2) that are generally used for forensic mtDNA sequencing analysis (Figure 9.4). HV1 occupies positions 16,024 to 16,365, and HV2 occupies positions 73 to 340 in the control region of the mitochondrial chromosome.

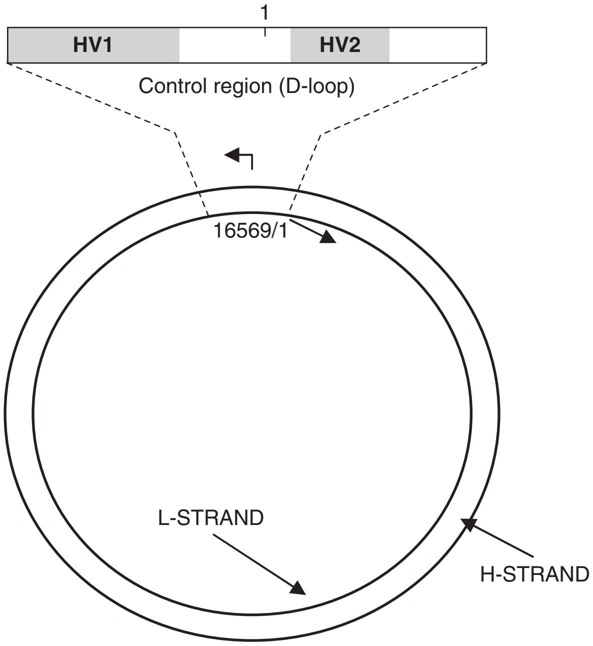

Figure 9.4 Mitochondrial DNA: a schematic representation of the 16,569 bp of the circular mitochondrial DNA genome. Both heavy (H-, outer circle) and light (L-, inner circle) strands are indicated, as are the positions of HV1 (composed of bp 16,024 to 16,365) and HV2 (composed of bp 73 to 340), which are the regions used in forensic mtDNA sequencing.

In addition to differences in organization and number, there are differences in the way that nuclear and mitochondrial DNA molecules are transferred from one generation to the next. Nuclear chromosomes, which are inherited from both parents, undergo genetic recombination, whereas mitochondrial chromosomes, which are inherited from only one parent, do not.

9.2 METHODOLOGY

9.2.1 Sample Collection and DNA Extraction

The isolation of genomic DNA from crime scene evidence samples is complicated by variations in both the quality and quantity of the material that is to be examined. Samples used for DNA analysis may be taken from varied materials, such as fabric, contact swabs, or cigarette butts, and the DNA in these samples may originate from various sources (e.g., blood, saliva, semen, skin). To serve as evidence in court, the biological material must be collected carefully and a chain of custody must be established. In addition, reference DNA samples are necessary to provide material for comparison to crime scene evidence; these can come from convicted offenders, suspects, and/or people who may have left their DNA at the scene merely because they happened to be in the area (elimination specimens).

As methodologies improve, crime scene evidence samples containing less and less DNA (such as those from single fingerprints) are being analyzed. When samples contain very little DNA, purification methods must yield the maximum amount of DNA in the smallest volume possible. Samples containing very small amounts of DNA (usually <100 pg) fall into a category known as low-copy-number DNA (Gill et al., 2000).

The first step in DNA analysis, DNA extraction, involves isolation of the DNA and its separation from other cell components. (Where applicable, cells may first be separated from their substrate swab or fabric by agitation in a minimal amount of liquid.) There are various possible DNA extraction methods, and when dealing with crime scene samples, the type of evidence and the amount of DNA it contains will help determine the extraction method used. Common forensic DNA extraction methods include the use of Chelex beads, solid-phase commercial kits, or organic extraction. [Various methods for manual sample preparation and extraction are described in detail by Butler (2005, Chap. 3).] Whatever extraction method is used, care must be taken to avoid contamination between the samples as well as contamination by extraneous DNA.

For many years most crime labs used the Chelex method for DNA extraction. Chelex is an ion-exchange resin that protects DNA by binding to magnesium ions, thereby inhibiting the function of the magnesium, which requires enzymes that destroy DNA. Cells are first broken open and their DNA released by boiling the cellular material in the presence of Chelex beads. After the centrifugation step that follows, the Chelex and all cellular components except the DNA form a tight pellet at the bottom of the tube. The supernatant liquid containing the extracted DNA is then transferred to a new tube and usually stored frozen at -20°C or -80°C until it is used for analysis.

There is currently a backlog of samples from convicted offenders that must be analyzed and entered into the national database. To ease this backlog and to analyze reference and elimination samples, many laboratories now use commercial kits and/or robotic instruments to prepare these samples, which have abundant DNA and are of a similar type. In addition, new methods and equipment that greatly improve DNA recovery are now available for sample collection, extraction, and quantification, including individualized collection and extraction kits and robotic extraction methodologies (see Brettell et al., 2005, 2007, 2009).

9.2.2 DNA Quantification

Because methods for DNA analysis usually require that the DNA be within a specific concentration range, the amount of human DNA present in an extracted sample must be measured before analysis. In most of the methods used for determining DNA concentration, results for extracted DNA are compared to those for calibrated standards that contain known quantities of DNA. Recent DNA quantification methods have moved from slot blot, to real-time (rt) or quantitative (q)

PCR-based methodology; rtPCR-based methods are more exact than blot-based procedures and can quantify samples with much lower DNA concentrations. Because rtPCR measures the change in fluorescence due to the increase in amplified product that occurs during the initial PCR cycles, it not only quantifies the DNA but indicates how much amplifiable DNA is present in the sample (Butler, 2005, Chap. 4). Kits such as the Quantifiler Human DNA Quantification kit from Applied Biosystems, Inc. are now being used to determine the amount of human DNA present in a sample by rtPCR. In addition, a Quantifiler Duo Quantification kit (Applied Biosystems), which simultaneously quantifies total human and human male DNA, has recently been validated (Barbisin et al. 2009).

9.2.3 Polymerase Chain Reaction

PCR is a process that amplifies DNA and makes millions to billions of copies of a specific DNA sequence through the use of primers. Primers are short segments of DNA (usually 18 to 30 nucleotides in length) that specify which region of the DNA will be copied. For each sequence amplified, two different primers (a forward and a reverse) are used; each primer attaches to a complementary region on one of the two separated DNA strands and designates the beginning or end of the DNA segment to be copied (Figure 9.5). By constructing primers with the appropriate nucleotide sequences, any specific segment of DNA may be amplified. In addition to the primer and the template DNA that is to be copied, a sample undergoing PCR must contain a special type of DNA polymerase which has its optimum function at high temperatures (e.g., *Taq* polymerase, which comes from bacteria native to hot springs), the four building blocks of DNA in a deoxynucleotide triphosphate (dNTP) form (i.e., dATP, dTTP, dGTP, and dCTP), a buffer solution, and salts (e.g., $MgCl_2$) necessary for the DNA polymerase to function.

The PCR process is similar to cellular DNA replication, but with PCR it is temperature changes that accomplish amplification of the DNA. A PCR cycle (see Figure 9.5) consists of three steps. Each cycle starts with a denaturation step where the double-stranded DNA sequence to be copied is separated into two strands by heating to a temperature of about 94°C. The temperature is lowered to about 60°C in the next step, annealing, and the two primers bind by complementary base pairing to each of the single-stranded segments of the target DNA. In the last step, extension, the temperature is raised to about 72°C and the DNA polymerase replicates the DNA by adding the appropriate nucleotide bases (dNTPs) through complementary base pairing. By doubling the amount of DNA with each cycle, the total amount increases exponentially as PCR progresses. The three-step PCR cycle is usually repeated 28 to 30 times for standard forensic DNA analysis.

9.2.4 Short Tandem Repeats

Currently, forensic DNA identification is usually done by PCR-based analysis of short tandem repeat (STR) loci. With PCR of STRs, only small quantities of biological material are necessary and the small PCR product size (about 100 to 500 bp) allows even partially degraded DNA to give full DNA profiles.

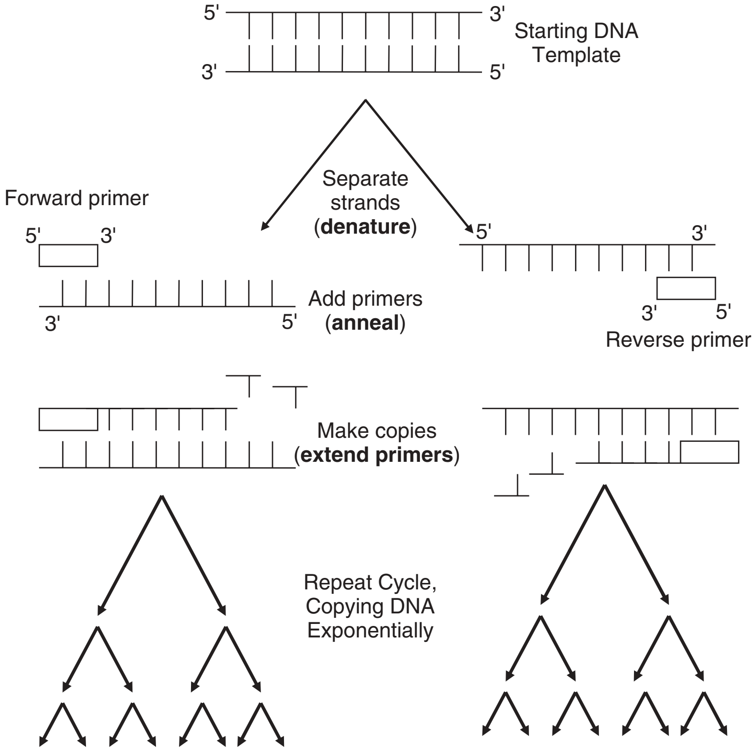
5'
3'
Starting DNA Template
3'
5'
Separate strands (**denature**)
Forward primer
5'
3'
5'
3'
Add primers (**anneal**)
3'
5'
3'
5'
Reverse primer
Make copies (**extend primers**)
Repeat Cycle, Copying DNA Exponentially

Figure 9.5 DNA amplification by polymerase chain reaction (PCR). A PCR cycle consists of the three steps depicted above, which involve changes in the temperature of the sample. In the first step, denaturation, the two DNA template strands are separated by heat. In the next step, annealing, the sample is cooled to a temperature appropriate for the two primers (forward and reverse) to bind. In the last step, extension, the temperature of the sample is raised to the optimal temperature for the DNA polymerase to extend the primers and produce a copy of each DNA template strand. In each cycle, the number of DNA molecules (containing the sequence between the two PCR primers) doubles and the number of copies increases geometrically. [From Butler (2005), copyright ©2005 Elsevier Academic Press, Fig. 4.2.]

STRs (also known as *microsatellite DNA*) are loci that contain units of short sequences of DNA (two to seven, usually four nucleotides long) that are tandemly repeated (Figure 9.6) anywhere from about a half dozen to several dozen times (see Butler, 2005, Chap. 5). There is a great deal of variation in the human population in the number of times the units of four bases are repeated in an STR, and this length variability is used to distinguish between DNA profiles. Polymorphic STR sequences, which vary throughout the population in the number of repeat units displayed, can be found in all parts of the genome, including introns, exons, and flanking portions of coding areas of the DNA as well as in extragenic areas (Economou et al., 1990; Edwards et al., 1991; Beckman and Weber, 1992). There

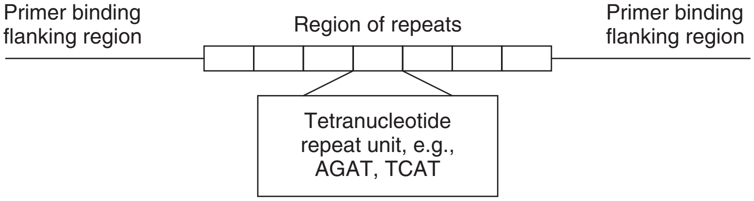
Primer binding flanking region
Region of repeats
Primer binding flanking region
Tetranucleotide repeat unit, e.g., AGAT, TCAT

Figure 9.6 Schematic representation of a short tandem repeat (STR) locus. Most STR loci used for forensic DNA analysis contain units of four bases (tetranucleotides) which are repeated one after another. The repeat portion of the locus is surrounded by flanking regions that contain the DNA segments to which the primers bind. One of the primers (usually the forward one) is labeled with a fluorescent dye of a particular color so that the alleles of the STR being amplified can be identified when its size overlaps with that of a different STR.

are probably more than 1 million STRs in the human genome, and among the more than 20,000 tetranucleotide STR loci that have been characterized (Butler, 2006), a set of 13 core loci have been selected for use in forensic DNA profiling in the United States. These 13 STRs are required for a full DNA profile to be uploaded into the FBI's Combined DNA Index System (CODIS) database. The existence of these core loci allows equivalent genetic information to be shared by national, and in many cases international, databases. The 13 core STR loci are used for various forms of human identity testing, including DNA databasing, forensic casework, missing persons/mass disaster victim identification, or parentage testing.

For each of the 13 core STR loci, a person inherits one maternal allele and one paternal allele. Because these loci are either located on different chromosomes or far enough apart on the same chromosome, the 13 core STR loci assort independently of each other during cell division. This allows their population frequencies to be multiplied by each other to obtain a high power of discrimination.

In addition to their presence on autosomal chromosomes, STRs are also found on the sex chromosomes. Y-chromosome STRs, which are present only in males, have been studied for many years as aids in forensic DNA analysis, and commercial kits are now available for their detection. Since Y chromosomes are inherited uniparentally and only one Y chromosome is present in a male individual, Y-STRs do not have the power of discrimination of core autosomal STRs. However, Y-STRs are frequently used to gain additional information when full autosomal STR profiles are lacking and to help distinguish DNA profiles in samples containing both male and female DNA and/or mixtures of DNA from more than one male. STRs can also be found on X chromosomes and, although not used for forensic DNA typing, many of these have been identified and characterized (Gomes et al., 2009; also see Brettell et al., 2005, 2007, 2009).

9.2.5 PCR of STRs

The primers used for PCR-based STR analysis are complementary to the regions that lay outside (flank) the area of the tandem repeats (Figure 9.6), and one of the two PCR primers is labeled with a fluorescent dye. PCR amplification is usually performed on multiple STR loci simultaneously (i.e., by multiplexing). Different fluorescent dyes are used to label the loci that overlap in size so that these STRs can also be distinguished by color. Simultaneous amplification of many loci enables a high power of discrimination to be attained in a single test without the use of too much DNA (e.g., only 1 ng or less of starting material is necessary). Commercial kits containing fluorescent-labeled primers are available for both autosomal STR and Y-STR analysis.

Numerous multiplex STR test kits are available for DNA typing. Newer multiplex kits increase the number of genetic loci that may be amplified and analyzed simultaneously. To analyze all 13 CODIS loci plus the sex-typing locus, amelogenin, many laboratories use a combination of two commercial kits, either the AmpF*l*STR Cofiler and AmpF*l*STR Profiler Plus, or the GenePrint PowerPlex 1.1 and GenePrint PowerPlex 2.1. Some labs use a newer multiplex kit (i.e., the AmpF*l*STR Identifiler or the PowerPlex 16 System), which can amplify the 13 core STR loci plus amelogenin and two extra STR loci in a single reaction. AmpF*l*STR kits are manufactured by Applied Biosystems Inc. (ABI) and PowerPlex kits by Promega Corp. Four different fluorescent dyes are used in the ABI Profiler and Cofiler kits and in all the Promega kits, while the ABI Identifiler kit uses five dyes.

Commercial kits for STR analysis come with solutions of labeled primers whose relative concentrations are appropriate for multiplexing and a “master mix” in which buffers, dNTPs, and necessary salts are already included. The ease and consistency achieved by these kits in generating STR profiles on the same set of core STR loci make national sharing of criminal DNA profiles possible. In addition, STR kits supply fragment sizing standards as well as allelic ladders containing common STR alleles that have been previously sequenced and characterized for the number of repeat units. The sizing standards are used to determine PCR product size, while the allelic ladders calibrate fragment size in relation to STR repeat number for genotyping purposes.

9.2.6 Separation and Sizing of STR Alleles

Electrophoresis exposes DNA to an electrical field that is used to separate DNA fragments. Because of its negative charge in solution, DNA always moves from a negative to a positive electrode when current is applied. Two types of electrophoresis, slab gel and capillary electrophoresis (CE), have been used for forensic DNA analysis. In slab gel electrophoresis, the separation method that was used for RFLP analysis and for early STR analysis, the DNA samples are placed in separate lanes (depressions) on a gel covered with a buffer solution. When current is applied, shorter DNA pieces move faster than longer pieces and the DNA fragments in each lane are separated by size. Bands of DNA are formed by fragments of the same size migrating together. Sample bands are compared to those of DNA fragment sizing standards (and/or a ladder of common STR alleles) and the length of the DNA in each band is estimated by matching the distance it traveled to a ladder standard that traveled a similar distance. The DNA bands formed in RFLP were detected by hybridization to labeled probes, while STR gel detection first involved silver staining of bands. More recent methodologies for STR analysis by either gel electrophoresis or CE use detection of PCR primers labeled with different color fluorescent dyes.

In forensic DNA labs today, CE is the method generally used for STR profiling. The separation, detection, and analysis of STR fragments of different sizes is accomplished using CE instruments such as the single-capillary ABI 310 or the 16-capillary ABI 3100 Genetic Analyzers made by Applied Biosystems Inc. (Butler et al., 2004); laboratories with very high throughput may use the ABI 3700 Genetic Analyzer, which has 96 capillaries that operate simultaneously.

CE is fairly similar to gel electrophoresis, except that the separation of fragments takes place in a long narrow glass tube or capillary that is filled with a polymer solution. Samples containing PCR amplified, fluorescently labeled STRs are first added to the capillary and when the current is applied the smaller fragments move faster in the polymer than the larger ones. As they pass through the capillary, the STR fragments are exposed to a laser beam that causes the fluorescent dyes in their primers to emit light of specific wavelengths. Sizing standard fragments labeled with a different dye color are also added to each sample tube. This allows a comparison of test sample fragment movement (measured as migration time) with

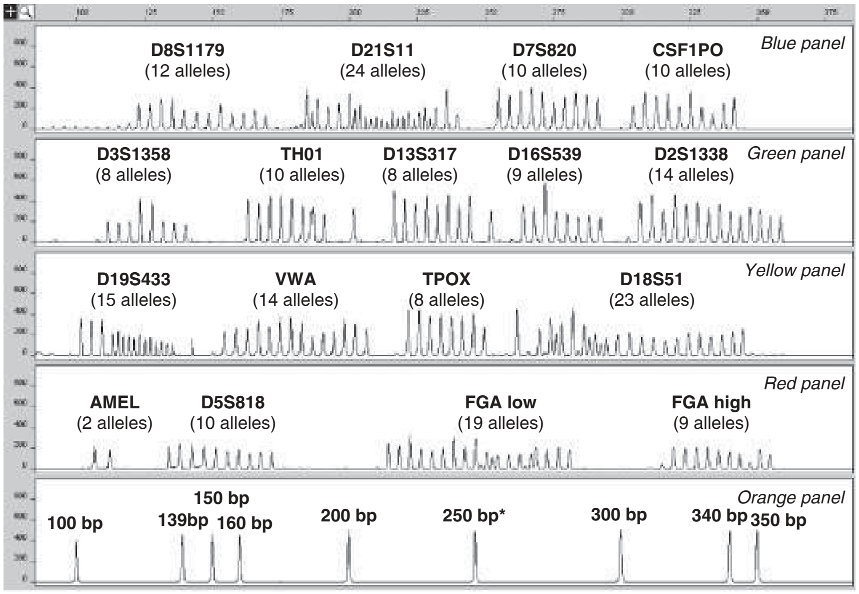

Figure 9.7 Allelic ladders for the AmpF*l*STR Identifiler kit (Applied Biosystems). This allelic ladder, which includes a total of 205 alleles, is used for genotyping a multiplex PCR reaction involving 15 STRs and the sex-typing test for amelogenin (AMEL). The kit uses five dyes, the colors of which are shown in the upper right portion of each panel. The upper four panels show possible alleles for each locus, and the bottom panel shows the peaks produced by the internal DNA sizing standard that is used with this kit. [From Butler (2005), copyright ©2005 Elsevier Academic Press, Fig. 5.6.]

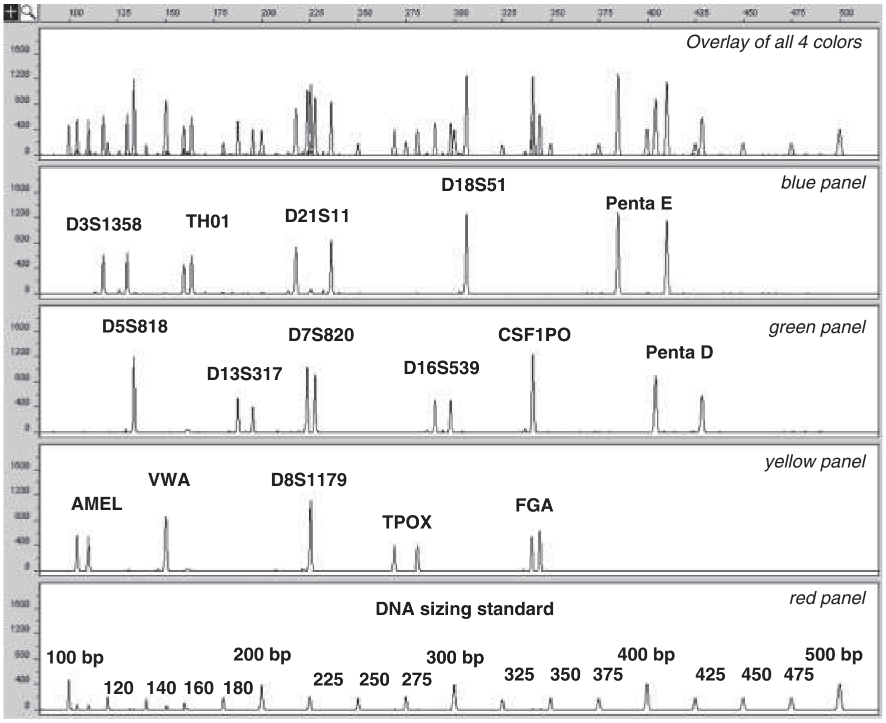

Figure 9.8 Electropherogram showing results for a 1-ng genomic DNA sample produced using the PowerPlex 16 (Promega) kit. This multiplex kit uses four dyes, the colors of which are shown in the upper right portion of each panel, for the labeling of 15 STRs and the amelogenin (AMEL) sex-typing locus. The top panel is an overlay of all four lower panels, including the red lower panel for the internal DNA sizing standard. [From Butler (2005), copyright ©2005 Elsevier Academic Press, Fig. 5.5.]

that of the sizing fragments. A fluorescently labeled allelic ladder sample made up of most of the known alleles for a particular locus is present in the same run and serves as the standard for naming the genotype of the alleles. The allelic ladder for the widely used AmpF*l*STR Identifiler kit (Applied Biosystems) is shown in Figure 9.7. It contains 205 of the alleles present in the 16 loci that are amplified together. These loci represent 15 STRs plus an amelogenin sex-typing assay.

CE data are presented in the form of an electropherogram (e.g., Figure 9.8) which shows a series of peaks, each of which represents the relative fluorescence of an amplified fragment (allele) plotted as relative fluorescence units (rfu) versus migration time (which is related to size). [Figure 9.8 is an electropheragram of results obtained using the PowerPlex 16 System (Promega), which also amplifies the 13 core loci, two extra STRs, and amelogenin in a single reaction.] In addition, CE instrument software calculates the size and genotype of the various alleles in the electropherograms. A peak is identified as a particular allele based on its fluorescent color emission and its length in base pairs, which should match that

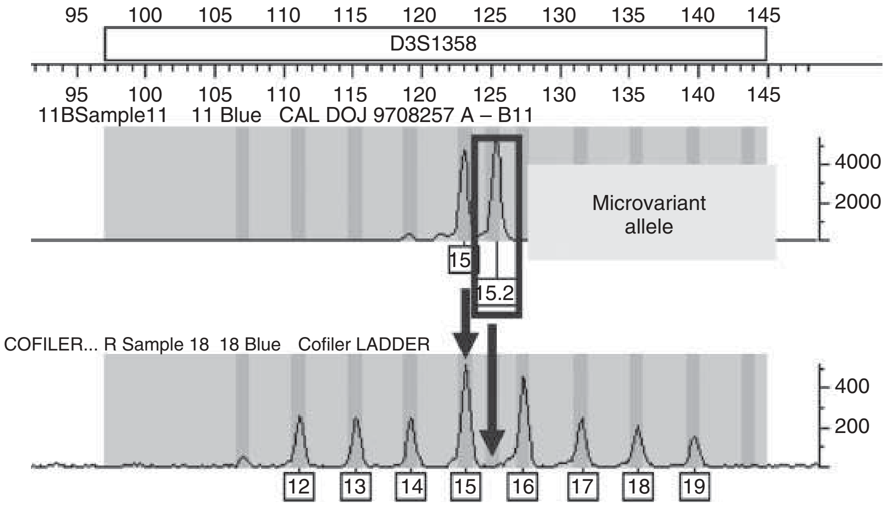

Figure 9.9 Detection of a microvariant allele for D3S1358. When genotyping STRs, alleles are named by a comparison of sample peaks to those of allelic ladders. The sample in the top panel is compared to the allelic ladder for the D3S1358 locus shown in the bottom panel. (Peaks are labeled with the allele number by use of an internal sizing standard.) The allele labeled 15.2 in the top panel does not match any of the ladder alleles. It is a microvariant whose size is two bases larger than allele 15 in the ladder. (www.cstl.nist.gov/STRbase)

of a fragment in the allelic ladder. Alleles that do not match ladder standards are known as *off-ladder* or *microvariant alleles*; these are usually far less common than those in the allelic ladder and therefore may be more statistically significant for human identification. Standard STR allele genotypes are named for the number of repeats they contain. Microvariant alleles contain an incomplete repeat (fewer than four bases) in addition to complete repeats and therefore fall between two allelic ladder standards; they are named by giving the number of full repeats followed by a decimal and then the number of bases in the incomplete repeat (see Figure 9.9).

Like any other experimental, scientific procedure, DNA analysis, requires the inclusion of controls. Negative controls are samples that do not contain any DNA but are subjected to all the same procedures as the samples being analyzed. (Negative controls monitor contamination starting from amplification, whereas reagent blanks, which are frequently used in mitochondrial DNA analysis, are negative controls that monitor contamination starting from extraction.) When no contamination is present, these controls will be negative for DNA. Substrate controls may be run if applicable, and these are specimens taken from an area adjacent to the area (stain) of interest to ensure that the DNA found in the sample did not come from its underlying fabric. Unlike negative controls, a substrate control may contain DNA which may or may not be relevant to the case. To indicate that the test is being done correctly, a known positive control must be included with every analysis and must produce the expected alleles. If negative controls are shown to contain DNA, or if a positive control does not produce the alleles expected, the test is considered inconclusive and the analysis of all samples must be repeated.

DNA fragments of different sizes, whether they come from a heterozygous STR locus or from different STR loci, have different migration times and are therefore distinguishable by CE. For identification purposes, discrimination power increases as more STR loci are studied, and as stated above, at least 13 STRs are required for a DNA profile to be uploaded into the CODIS offenders database. By amplifying multiple STR loci in the same PCR reaction and running them together in the same capillary, evidence is conserved and time and effort expended for DNA analysis are decreased. Many of the alleles of the 13 core STR loci overlap in size, and because of this, primers for each of these overlapping loci must be labeled with different fluorescent dyes so that they may be distinguished from one another.

9.2.7 Combined DNA Index System (CODIS) Database

The 13 core loci that are currently the basis of the U.S. CODIS database were selected in November 1997. The national CODIS database or national DNA index system (NDIS) contains profiles that have been uploaded from various state DNA testing laboratories through their respective systems. A laboratory that wishes to enter data into the CODIS database must meet certain standards to ensure the reliability of its test results.

The national database includes the following 13 core STR loci: D3S1358, D16S539, THO1, TPOX, CSF1PO, D7S820, VWA, FGA, D8S1179, D21S11, D18S51, D5S818, and D13S317. All 13 of these loci are highly polymorphic, and 11 of them are located on different chromosomes. Although CSF1PO and D5S818 are on the same chromosome, they are far enough from each other to assort independently during cell division. None of the core loci are located in coding regions, so they have no functional significance.

Several databases make up CODIS, including a convicted offender database, an evidentiary database, and other specialized databases. DNA obtained from crime scene evidence is included in the evidentiary database. Offenses that require submission of a DNA specimen to the convicted offender database vary from state to state, and the data uploaded to CODIS must meet certain standards. State and local data that differ from those meeting CODIS standards can be used on a local level only. Although most states originally included only those offenders who had been convicted, and usually only those convicted of violent felonies, some states either now include or will soon include profiles of all persons arrested for any serious criminal activity.

Most European countries (including the United Kingdom) uses 10 core STR loci in their DNA databases, and these include eight of the CODIS loci: VWA, FGA, THO1, D8S1179, D21S11, D18S51, D3S1358, and D16S539 (Butler, 2006). The two remaining European core STR loci (D2S1338 and D19S433) can be amplified with the AmpF*l*STR Identifiler kit, which amplifies a total of 15 STR loci in one reaction (Collins et al., 2004).

9.2.8 Frequency and Probability

STR analysis may demonstrate that the genetic profile from the crime scene sample differs from that of the suspect and thus excludes or eliminates the suspect as a potential donor of the evidence. However, if the genetic profiles of the suspect and the crime scene sample are the same, the suspect is included. Thus, DNA evidence can either link a suspect to a crime scene or prove that a person accused incorrectly is innocent of a crime. As of August 2009, 273 prisoners have been exonerated from crimes they did not commit through the use of DNA evidence.

If there is a DNA match, the meaning of that match must be determined. The probability of finding an identical genetic profile in a particular population depends on the frequency of the profile in that population. Because the United States is populated by people from all over the world, it is sometimes difficult to establish the population or populations of origin for a person. The FBI lists the following five large ethnic groups in their CODIS database: (1) African-American, (2) Asian-American, (3) European-American, (4) Southeastern-Hispanic, and (5) Southwestern-Hispanic.

Many data are available giving the frequencies of the core STR alleles in various populations groups, and Table 9.1 shows the frequencies of the most common alleles of some CODIS STR loci in a Caucasian (European-American) population using data from a study conducted by Butler et al. (2003a).

To calculate the probable or expected frequency of a heterozygote in a population, one would multiply the individual allele frequencies by each other and then multiply that product by 2. Since there are two different ways that a heterozygote can be formed (i.e., either one of the two alleles could have come from either the sperm or the egg), the product of the allele frequencies must be doubled. Homozygote probability is calculated differently since only one set

TABLE 9.1 STR Allele Frequencies, U.S. Population Data[a]

Locus	Allele	Caucasian ($n = 302$)	African-American ($n = 258$)	Hispanic ($n = 140$)
CSF1PO	10	0.21689	0.25681	0.23214
	11	0.30132	0.24903	0.29286
	12	**0.36093**	**0.29767**	**0.35714**
TPOX	8	**0.53477**	**0.37209**	**0.47143**
	9	0.11921	0.17829	0.10357
	11	0.24338	0.21899	0.27500
FGA	21	0.18543	0.11628	**0.16786**
	22	**0.21854**	**0.19574**	0.15000
	23	0.13411	0.17054	0.13571
VWA	16	0.20033	**0.24806**	**0.26429**
	17	**0.28146**	0.24225	0.21786

Source: Data from Butler et al. (2003a).

[a] Frequencies for some of the most frequent alleles are given for four CODIS loci. Alleles with highest frequencies are indicated in bold.

of circumstances (i.e., where both the sperm and the egg carry the same allele) leads to homozygote formation. Thus, the frequency of the homozygous allele is multiplied by itself or squared. As an example, in the Caucasian population in Table 9.1, the probability of a person being a 12,11 heterozygote for the CSF1PO locus would be 0.36093 × 0.30132 × 2 (or 0.10876), while the probability that they are an 8,8 homozygote for the TPOX locus would be 0.53477 × 0.53477 (or 0.28598).

To determine the probability of a "random match" for a genetic profile which contains, for example, two of the 13 STRs, the frequency calculated for each of the two genotypes is multiplied by the other because the two loci are inherited independently. Because the chance of multiple genotypes occurring together in the same person is the product of their individual probabilities, calculations of this type follow what is known as the *product rule*. Thus, the frequency of a person possessing the genotype for CSF1P0 and TPOX described in the paragraph above would be 0.10876 × 0.28598, or 0.03110. As a genotype for an additional STR is added to a profile, its frequency is multiplied by the product of the frequencies of the other STRs examined.

For the European-American population, the probability of a random match for the genotype that includes all of the most frequent alleles of the 13 CODIS STRs greatly exceeds the entire population on Earth. Thus, one can say with virtual certainty that this genetic profile would be unique on this planet. In addition, all other profiles for the 13 core STRs in the various population groups provide virtual certainty that a particular genetic profile is unique. (There are various web sites that will calculate the random match probability for any core STR profile; see Butler, 2006.)

9.3 PROBLEMS ENCOUNTERED IN STR ANALYSIS

9.3.1 Low-Copy-Number DNA

Blood or semen stains recovered from violent crime scenes usually contain more than enough DNA to perform traditional STR analysis. To obtain amplification of all of the STR alleles present, most commercial STR kits require that starting samples contain from 0.5 ng (500 pg) to 2 ng of DNA. If full profiles could be obtained for samples with much less DNA, a greater range of evidence samples could be analyzed and many more nonviolent crime investigations might be solved by DNA profiling. More and more crime labs are now performing DNA analysis on crime scene samples such as contact swabs and fingerprints which contain very little DNA.

DNA that is present in a very small amount (i.e., less than 100 pg) is known as *low-copy-number* (LCN) *DNA* (Gill et al., 2000; Gill, 2001). Since there are about 6 pg of DNA in a typical human diploid cell, a LCN sample would contain fewer than 17 cells. To analyze samples containing LCN amounts of DNA, the sensitivity of the assay must be increased. This is usually done by increasing to about 34 cycles the number of PCR cycles from the 28 or 30 that are suggested for most commercial STR kits. Since increased sensitivity also increases the potential for analyzing contaminating DNA, extra care should be taken with LCN DNA analysis.

There are also numerous artifacts that frequently arise with LCN DNA analysis that may interfere with interpretation of the DNA profile (Gill et al., 2000). Because such a small quantity of DNA is present in a LCN sample, less than a full complement of the genome may be available for amplification in each PCR reaction; since it is only the loci that are amplified in the first rounds of PCR that will probably show up in the final DNA profile, repeated amplifications of the same LCN sample may give different results. This phenomenon, known as *stochastic variation*, occurs frequently with LCN DNA analysis. Sometimes only one of the two alleles of a heterozygous STR locus will be amplified, and the locus will mistakenly appear in the profile as homozygous (allele dropout). Other times, an entire STR locus may fail to amplify (locus dropout), or extra alleles may appear (allele drop-in). Because of the stochastic variation of these amplification artifacts, PCR of LCN DNA samples is usually done three times, and the only alleles that are reported are those that are present in at least two of the three profiles produced from the same sample (Gill et al., 2000).

Additional complications occur with LCN DNA because the smaller amount of input DNA and the increased sensitivity of the PCR amplification make contamination artifacts such as allele drop-in much more likely. Thus, ultraclean laboratories and DNA-free reagents must be used for LCN DNA analysis. In addition, with LCN DNA it is possible that all the DNA in the original sample was not connected to the crime under investigation; the DNA may have come into the evidence sample by secondary transfer: that is, from people with whom the offender had contact before touching the evidence, or from other people who may have touched the item before the evidence was collected. Thus, LCN DNA analysis results must be interpreted with extreme caution.

Many crime labs in the United States are now analyzing material with LCN amounts of DNA for leads in burglary cases. In England, where LCN DNA analysis was pioneered, forensically relevant DNA profiles have been generated successfully from items such as discarded tools, matchsticks, and weapon handles. Increased use of LCN DNA analysis by crime labs has led to the development of various testing and interpretation protocols (e.g., Caragine et al., 2009; Gill et al., 2009), and there is still controversy surrounding the use of LCN DNA typing techniques (Buckleton, 2009).

9.3.2 Degraded DNA and Reduced-Size (Mini) STR Primer Sets

DNA degradation refers to the breakdown of relatively large fragments of DNA into smaller fragments, and there are various processes (e.g., physical, oxidative, or biological) that may contribute to this degradation. As fragments get smaller in size, the target DNA sequences that are used for the PCR amplification reaction are also broken down and the larger alleles of a locus may decrease in peak height in the electropherogram. If the larger allele for a heterozygous STR is missing,

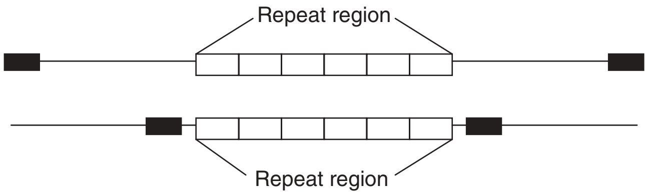

Figure 9.10 Schematic of mini-STR primer positions. Mini STRs are created when primer binding sites (black rectangles) are moved closer to the repeat region in order to shorten the length of the amplified product. The upper portion of the figure shows possible primer positions for a hypothetical traditional STR, while the lower portion shows possible primer positions for a miniSTR with the same number of repeats.

the locus would appear erroneously to be homozygous, and with larger STRs, the entire locus could be missing from the profile. In order to have complete PCR amplification, the full target sequence must be present in the regions where the primers anneal as well as in the portion of DNA between the primer binding sites (Hoss et al., 1996).

Frequently, when the DNA in a forensic sample is present in fragmented form, a full DNA profile cannot be obtained using standard methodology. This problem has been addressed by the development of reduced-size STR primer sets known as *miniplexes*, in which the primer sequences have been moved as close as possible to the repeat region (Figure 9.10) so that smaller STR fragments can be formed (Wiegand and Kleiber, 2001; Butler et al., 2003b). Reduced-size primer sets were first developed for the analysis of STRs by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry, a technique that required shorter DNA fragments for accurate sizing without ladder standards (Butler et al., 1998). Primers for reduced-size STRs also held promise as being better for analyzing degraded DNA and they were used, along with other methods of DNA analysis, to aid in the identification of victims of the World Trade Center disaster. Recent studies comparing mini-STRs with standard STRs and single-nucleotide polymorphism (SNP) assays (see below) found that miniSTRs were usually most effective for the analysis of degraded DNA (Dixon et al., 2006).

Mini-STRs are becoming an important tool for the analysis of degraded DNA in forensic casework samples. Mini primer sets which produce PCR products ranging in size from 50 to 150 bp have been developed for 26 non-CODIS STR markers (Coble and Butler, 2005; Opel et al., 2007; Hill et al., 2008), and a commercial kit for mini-STR amplification of CODIS loci, the AmpFlSTR MiniFiler PCR Amplification kit, is now available from ABI. This kit contains reagents for the mini-STR amplification of the eight largest CODIS loci, which would be the first to drop out in a degraded DNA sample. The MiniFiler kit has recently been validated according to FBI/National Standards and SWGDAM guidelines (SWGDAM, 2000) and was shown to have significant performance improvements over standard STR kits in models of DNA degradation and PCR inhibition (Mulero et al., 2008).

9.3.3 PCR Inhibition

Sometimes substances that interfere with the PCR reaction are present in a forensic sample and coextract with the DNA; the presence of these inhibitors may prevent amplification of some loci or even all the loci. PCR inhibitors may interfere with either the DNA itself or with the polymerase enzyme. Such inhibitors may be present in crime scene DNA samples that were deposited on wood, soil, fabrics, and so on. In addition, hemoglobin in blood, melanin in hair, and certain textile dyes in fabrics, such as that from blue jeans, are inhibitors that coextract with DNA and are frequently present in forensic casework samples (see Besseti, 2007).

The following methods are among those that can be used to decrease PCR inhibition: the DNA sample, and therefore the template, may be diluted so that the inhibitors are also diluted; extra DNA polymerase may be added to overcome inhibition, or bovine serum albumen or betaine may be added to minimize or prevent PCR inhibition; a filtration step may be performed prior to PCR to separate the extracted DNA from the inhibitor [see Raadstom et al. (2004) for a review].

Larger STR loci are more sensitive than smaller loci to PCR inhibition. Because of this, resulting profiles from samples containing inhibitors frequently resemble those of degraded DNA, where larger loci are likely to drop out before smaller ones do. Mini-STR primers, which produce smaller amplified fragments and work well with degraded DNA, may therefore be useful for obtaining full-STR profiles when PCR inhibitors are present.

9.3.4 Interpretation of Mixtures of DNA

If an evidence sample contains a mixture of DNA from more than one person, the profile observed may be confusing and complicated. Mixtures may contain DNA from multiple victims, multiple assailants, victim(s) and assailant(s), or from a single person plus contaminating DNA. If more than two peaks are found for a particular STR locus, the sample is probably a mixture. However, there are cases where three alleles for a single locus (triallelic pattern) may actually be present in a single individual (Butler, 2006), and there are also biological and technological artifacts (described below) that can produce extra peaks. Thus, one must first examine the overall genetic profile in order to determine whether a mixture is present.

Even a two-peak locus that appears to be heterozygous could, in fact, be a mixture from (1) two contributors each of whom is homozygous for a different allele, (2) two contributors both of whom are heterozygous for the same alleles at this locus, or (3) a combination of one contributor who is homozygous and a second who is heterozygous for that locus. If equal amounts of DNA from each person are present, then in the third case one of the two alleles will produce a peak that is three times higher than the other. If the mixture comes from two homozygotes, relative peak heights for the two alleles will depend on how much DNA from each person was present in the mixture.

Since results obtained when DNA mixture components are combined in different ratios are confusing and contain variable peak heights that require further interpretation, it has been suggested that manual interpretation of mixtures not be attempted unless the mixture is the only evidence sample available (Butler 2005, Chap. 7). To aid in mixture interpretation, computer programs are now available to help determine mixture components and ratio; ABI is currently developing GeneMapper ID-X Software Mixture Analysis Tool software for this purpose.

9.3.5 Null Alleles and Allele Dropout

Even with a 1-ng sample of undegraded DNA, there are times when one allele from a person with a heterozygous genotype is not amplified during PCR and the person appears erroneously to be homozygous for that locus. Allele dropout such as this usually occurs because of mutations (sequence polymorphisms) in the flanking region of an STR locus; a mutation in the primer-binding site may prevent one of the primers from annealing, or a mutation close to the primer binding site may block the extension of new DNA during amplification. The presence of alleles with such mutations, known as *null alleles*, can cause errors in the interpretation of STR profiles. However, because the flanking sequence around the tandem repeats is fairly stable and usually not prone to sequence polymorphism, null alleles are rather rare. If a null allele is present and different primers are used for different amplifications, it is possible to obtain a heterozygous genotype with one primer set and an apparent homozygote with another. Since different multiplex kits generally use different primers for the same locus, mutations in or near one of the primer binding sites could produce apparently different genotypes with the different kits. Primer concordance studies have been conducted to examine this phenomenon and, in a comparison of the PowerPlex 16 kit to the Profiler Plus and COfiler kits, allele dropout due to a primer mismatch was found in 22 of 2000 samples in seven of the 13 core STR loci (Budowle and Sprecher, 2001). In addition, because the new MiniFiler STR kit uses new primers with different primer binding sites, it was recently tested for the concordance of its results with those for primers from conventional STR kits (Hill et al., 2007).

Since two identical samples amplified using the same kits will always produce the same genotype results, all samples from the same case should, wherever possible, be tested with the same multiplex primers. However, because DNA database samples (e.g., those of convicted offenders) may have been obtained using a different kit containing primers which differ from those used on the evidence sample, in such a case a database search for an exact match would overlook the true perpetrator if a null allele were present for a primer in one kit but not in the other. To overcome this problem, it has been suggested that fewer stringent match criteria be used in database searches, such as a match of 25 rather than 26 of 26 alleles (Butler, 2005, Chap. 6). In addition, if a primer binds to a region where a polymorphism that produces a null allele is known to occur in some people, the null allele problem may be solved by including “degenerate primers.” These are multiple primers for the same locus, each of which will bind to one of the sequence polymorphisms that may be present for that primer binding site. The AmpFlSTR kits contain degenerate primers for the D16S539 locus, which was found to have sequence polymorphism for its primer binding site; degenerate primers for VWA and D8S1179 are also present in some commercial kits (see Butler, 2005, Chap. 6).

9.3.6 Factors Causing Extra Peaks in Results Observed

Electropherograms should always be examined manually to determine that all peaks called are real and not due to artifacts produced during fragment amplification, separation, or detection. Stutter and incomplete adenine addition are two biological artifacts that can cause extra peaks to appear on an electropherograms while pull-up (bleed-through), dye bobs, and spikes are among the technological artifacts that can produce extra peaks (see Figure 9.11 and below).

9.3.7 Stutter Product Peaks

Additional small peaks, several bases shorter than each STR allele peak, are usually present on STR electropherograms. Called *stutter product peaks*, these are formed during PCR when the STR loci are copied by the DNA polymerase. Stutter peaks have been shown to contain one repeat unit less than the true allele peak (Walsh et al., 1996) and they occur because of slipped-strand mispairing during primer extension, which causes either the primer or the template strand to slip, so that one repeat forms a loop that is not base-paired (Hauge and Litt, 1993). (Stutter peaks that are one repeat unit larger than the true allele can occur but are rarely seen.) Stutter peaks can affect DNA profile interpretation, and when mixtures of DNA

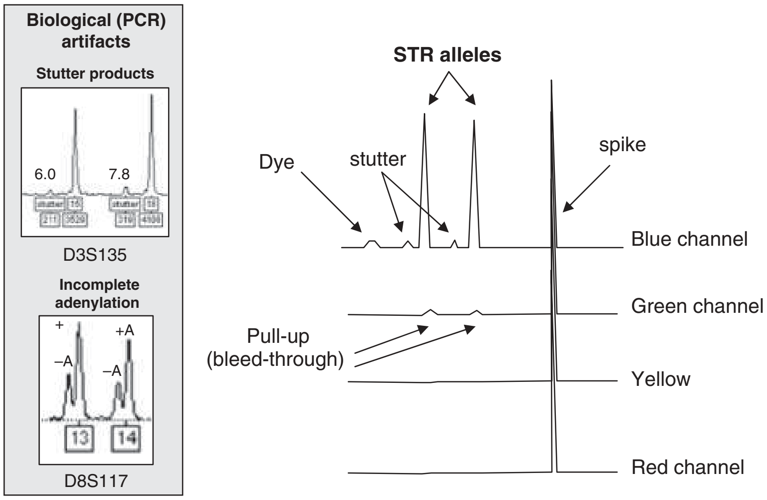

Figure 9.11 Various biology- (left panel) and technology-related artifact peaks that can be observed with STR typing. [From Butler (2005) copyright ©2005 Elsevier Academic Press, Fig. 15.4.]

from two or more sources are present, it can be difficult to determine whether a small peak is a real allele from a minor contributor or a stutter product peak.

Although stutter peak height is proportionally different for each of the different STRs, it is almost always less than 15% of that of the actual allele. Stutter height usually ranges from about 2 to 14% of the adjacent true peak, and the genotyping software does not call peaks in this peak height range.

9.3.8 Nontemplate Addition (Incomplete Adenylation)

The *Taq* polymerase used in PCR frequently adds an extra adenine (A) nucleotide to the 3′ end of a PCR product as it is copying the template strand. This nontemplate addition can lead to the formation of a PCR product that is one base longer than the template DNA. Thus, a peak for an STR may contain two fragments differing in size by only a single adenine base, and this can lead to the formation of electropherograms containing split peaks, broad peaks, or peaks with shoulders. Since sharper peaks increase the accuracy of allele calls by genotyping software, it is preferential to have all peaks either +A or −A rather than a mixture of both. The production of PCR products that all contain an additional adenine can be accomplished by adding an incubation step at 60 or 72°C after the PCR cycles are completed (Kimpton et al., 1993). A final extension step is therefore included in most PCR protocols so that there is enough time for the DNA polymerase to fully adenylate all PCR products. To ensure that accurate calls are made, allelic ladder fragments are also adenylated completely.

9.3.9 Technological Artifacts

Pull-up, also known as *bleed-through* or *matrix failure*, may appear on an electropherogram as a low peak in the same position as a very high allelic peak of a different color. Pull-up is caused by spectral overlap and occurs if the linear range of detection is exceeded due to sample overloading; a peak of a different color pulls up or bleeds through in the same position.

Spikes are sharp peaks that appear as vertical lines. They may be caused by air bubbles in the capillary or by voltage spikes and are easy to distinguish from true peaks since they occur at the same position in all colors.

Dye blobs form peaks that are broader and lower than allele peaks. They are caused by fluorescent dyes coming off their primers and moving through the capillary independently. Dye blobs will possess the spectrum of one of the primer labeling dyes.

9.3.10 Single-Nucleotide Polymorphism Analysis of Autosomal DNA SNPs

A single-nucleotide polymorphism (SNP) is a single base difference between individuals at a particular point in the genome. The human genome contains millions of SNPs (Stoneking 2001), and many have been used for years as markers to track genetic diseases. In addition, SNPs are frequently used together with STR analysis for DNA identification, especially with degraded DNA samples where a complete STR profile cannot be obtained. Both SNP and mini-STR analysis (along with mitochondrial DNA sequencing) were among the methods used in the identification of victims of the World Trade Center disaster (Biesecker et al. 2005).

Whereas STR analysis involves differences in the numbers of multiples of four base repeats, a SNP involves only one differing nucleotide; thus, PCR products from SNPs can be much smaller than those of STRs, making them appropriate for analyzing degraded DNA. However, SNPs are not as informative as STRs since usually only two different nucleotides will be present for a particular SNP, whereas with STRs more than five variations in the number of repeated tetranucleotides can be found at most loci. Thus, to achieve the same random match probability provided by the 13 CODIS STRs, many more SNP markers would have to be analyzed. It has been estimated that 50 to 100 SNPs would probably have to be analyzed to have the same discrimination power as 10 to 16 STRs (Chakraborty et al., 1999; Gill, 2001; Gill et al., 2004).

Because SNPs have lower mutation rates than those of STRs, they are more likely to become fixed in a population and therefore may be potentially useful in predicting phenotypes such as the ethnic origin of an offender. Most of the work on such phenotypic SNPs has, thus far, been related to pigmentation, and certain SNPs have been found to be linked to red hair and fair skin (see Budowle and van Daal, 2008).

9.3.11 Methods Used for SNP Analysis

SNPs are frequently analyzed by *minisequencing*, sometimes called *SNaPshot*, a variation of the typical DNA sequencing method of Sanger et al. (1977) that is described in more detail below for mitochondrial DNA (mtDNA) analysis. In minisequencing the SNP and the DNA region immediately surrounding it are amplified, and this PCR product is then used in a PCR reaction where only one new nucleotide, the one that defines the SNP, is added after the primer binds. In this second PCR, a special type of nucleotide [a dideoxynucleotide triphospate (ddNTP)] is used to stop amplification immediately, and each of the four different ddNTPs is labeled with a different fluorescent dye so that the nucleotide present in the SNP can be identified. In addition, SNP analysis can be performed as a multiplex so that many SNPs can be identified simultaneously. This is done by adding a varying number of extra nucleotides to the 5′ end of the primer (usually, poly T) so that each primer, and therefore each PCR product, differs from the others by several nucleotides, and fragments can then be separated by size (Tully et al., 1996; Vallone et al., 2004).

Other assays used to detect SNPs are variations of techniques used in DNA microarrays (described below in more detail for mtDNA). These arrays consist of many single-stranded DNA probes (fragments representing possible SNPs) attached to a matrix or membrane in a known position in a tightly spaced grid (microarray) or in a linear fashion (linear arrays). The regions of interest are labeled during amplification of the DNA sample (many SNPs can be amplified at once), and the labeled PCR product is then hybridized to the array. Only DNA that is complementary to the various probes will bind to the array and after developing the label, be detectable after washing. By examining which positions on the array are labeled, the sequences (or SNPs) present in the original DNA sample can be determined.

9.3.12 Mitochondrial DNA Analysis

DNA evidence samples may be too limited in amount or too degraded for conventional nuclear DNA analysis or even mini-STR analysis. MtDNA analysis is frequently used to determine the origin of this type of sample. Before miniplex STR primers were developed, mtDNA analysis was the only way to generate any identity information from degraded DNA samples, and it is still used for analysis when there is not enough nuclear DNA or when the nuclear DNA is so degraded that mini-STR analysis is impossible.

Because mitochondrial DNA is inherited from only one parent and therefore does not undergo genetic recombination, mtDNA analysis provides much less genetic identity information than does analysis of nuclear DNA. However, because so many more copies of mtDNA are present in a cell, its analysis can frequently provide information when nuclear DNA cannot. While each cell contains only two copies of each nuclear DNA chromosome (one maternal and one paternal), mtDNA is present in hundreds to thousands of copies per cell. The average cell contains about four or five (with a range of 1 to 15) copies of mtDNA in each mitochondrion, with an estimated average of 500 mtDNA molecules per cell (Satoh and Kuroiwa, 1991).

In addition to the higher number of copies present, other factors contribute to the survival of mtDNA under conditions where nuclear DNA may be highly degraded. The circular nature of the mitochondrial chromosome protects the DNA from enzyme degradation, as does its encapsulation in a double-walled structure within the cell (Butler 2005, Chap. 10). These factors have enabled the success of mtDNA analysis in determining the origin of bones, teeth, and ancient remains.

MtDNA is inherited maternally; thus, all children inherit their mtDNA sequence from their mother and the mtDNA of a male is not passed along. Women have the same mtDNA profile as their siblings, male and female, their children, their grandchildren, and all their maternal relatives. Because the mitochondrial genome does not undergo the type of recombination events that occur in nuclear DNA, unless a mutation occurs, mtDNA is transmitted unchanged in sequence from generation to generation.

Mitochondrial chromosomes are closed circles of double-stranded DNA that contain 16,569 base pairs (bp) numbered in a clockwise direction (Figure 9.4). There is an asymmetric distribution of guanines, the heaviest of the four DNA nucleotides, on the two complementary strands of mtDNA. The strand with the most guanines is known as the heavy (H-) strand and the other as the light (L-) strand. The 16,569 bp of the mitochondrial genome include 37 genes that code for products used in mitochondrial structure and function (exons). There are no introns and very few nucleotides between the coding regions.

The mitochondrial genome also contains a noncoding control region that is 1122 bp in length and contains the origin of replication for one of the two mtDNA strands. The replication of mtDNA begins in the H-strand of the control region. There are two segments of the control region DNA that are highly polymorphic, and these are called *hypervariable* (HV1 and HV2) *regions* (see Figure 4). HV1 is 342 bp long and HV2 is 268 bp long. For forensic mtDNA analysis, all 610 of the HV region bases are sequenced. Unrelated individuals have been shown to vary about 1 to 2% in their hypervariable control region DNA, meaning that 7 to 10 nucleotides of these 610 bases are different (Budowle et al., 1999).

9.4 METHODOLOGY FOR mtDNA ANALYSIS

9.4.1 Preparation of Samples

Because it is present in higher copy number and because its analysis requires more PCR cycles, mtDNA is more sensitive to contamination than nuclear DNA. Thus, care should be taken to avoid contamination by other mtDNA, and the extraction of mtDNA should be performed in a very clean laboratory environment. In addition, the analysis of reference samples should be done after all evidence samples have been processed completely. Reagent blanks and negative controls should also be run to monitor levels of exogenous DNA that may be present in reagents and/or the laboratory environment.

There is usually very little DNA present in the materials that require mtDNA analysis in forensic cases (e.g., hair, teeth, and bones). In the past, mtDNA was frequently quantified indirectly by assuming a fixed ratio between nuclear and mtDNA and taking a fraction of the nuclear DNA; however, there are now methods for the direct measurement of mtDNA using rtPCR (e.g., Meissner et al., 2000; Andreasson et al., 2002). In addition, because the type of sample used for mtDNA analysis frequently contains PCR inhibitors that can be coextracted with the DNA, purification of mtDNA is usually more extensive than that needed for nuclear DNA.

9.4.2 MtDNA Sequencing Methods

The FBI began to examine mtDNA sequences for identification purposes in June of 1996. For forensic analysis, dideoxynucleotide (ddNTP) sequencing (Sanger et al., 1977) is the method generally used to determine the order of the nucleotides in HV1 and HV2, the hypervariable areas of the control region. However, because of the extensive amount of time and labor required for sequencing, more rapid screening methods (described below) are frequently used first, to exclude samples that do not match.

MtDNA sequencing uses the same CE instruments that are used for STR analysis; however somewhat different PCR and CE analysis strategies are used. First, the DNA of each hypervariable region is amplified, usually using 34 to 38 cycles, many more than the 28 to 30 cycles commonly used for nuclear DNA amplification. (Since PCR inhibitors such as melanin can be present in hair samples for mtDNA analysis, extra *Taq* polymerase is sometimes added to overcome this inhibition.) The PCR product for each region is then individually amplified in another PCR reaction. In this PCR, self-terminating nucleotides that stop DNA replication (ddNTPs) are present in addition to the usual dNTP building blocks (Sanger et al., 1977). In the methods currently used with CE sequencing (see Figure 9.12), the four ddNTPs are each labeled with a different color fluorescent dye. When a ddNTP is added to a segment of DNA instead of a dNTP, the extension of the DNA stops immediately and no new nucleotides are added. Because both dNTPs and ddNTPs are present in the PCR reaction, different amplification products are terminated at different points on the DNA template, thereby forming a mixture that contains a series of DNA fragments, each differing by one base pair in length. CE is then used to separate the PCR fragments by length, and because each fragment contains the label of the last base added (in the form of a fluorescent ddNTP), the sequence of bases is obtained for the entire DNA region examined.

9.4.3 Reference Sequences

Following sequencing, the mtDNA sequence of each questioned (evidence) and known (e.g., suspect) sample is compared to a reference sequence and the differences found at specific sites are noted. Human mtDNA was first sequenced in 1981

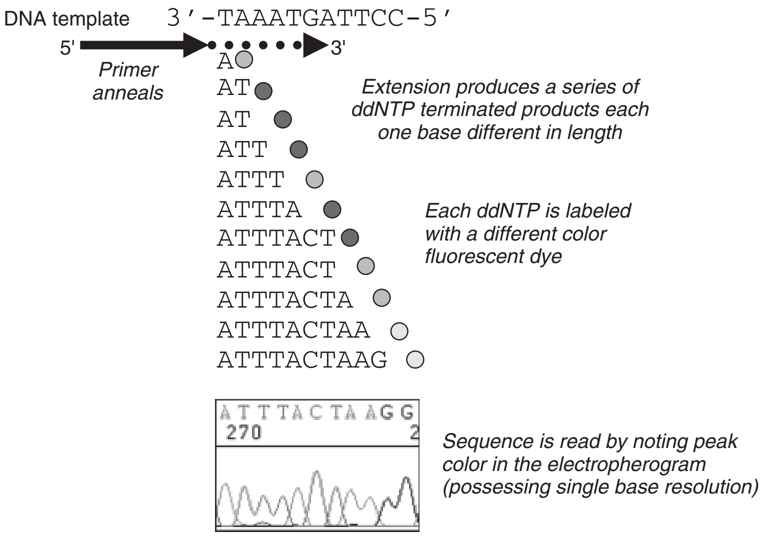

Extension produces a series of ddNTP terminated products each one base different in length

Each ddNTP is labeled with a different color fluorescent dye

Sequence is read by noting peak color in the electropherogram (possessing single base resolution)

Figure 9.12 DNA sequencing using fluorescent dideoxynucleotide triphosphates (ddNTPs). A primer that anneals next to the specific DNA region to be sequenced is extended by DNA polymerase. Since each sample contains a mixture of dNTPs and ddNTPs for each of the possible nucleotides, some of the extension products continue to be extended, while others are stopped by incorporation of a ddNTP. Each ddNTP is labeled with a different dye, and this allows each extension product to be distinguished by color. The separation of extension products by size allows the DNA sequence to be deduced. [From Butler (2005), copyright ©2005 Elsevier Academic Press, Fig. 10.5.]

(Anderson et al., 1981), and this sequence, known as the Anderson or Cambridge Reference Sequence (CRS), has been used as the reference sequence for forensic mtDNA analysis. Sequencing results are reported in terms of variation from the control region on the L-strand of the CRS. Only the bases that deviate from the reference sequence are reported, and all others are assumed to be the same. For example, a sample differing from the CRS because it has a G instead of an A at position 1620 in the HV1 region would be listed as 1620 G in the sequence report. In the absence of mutation, mtDNA coming from the same person or from a person with the same maternal lineage is expected to have the same DNA sequence and therefore the same differences from the reference sequence. Software packages such as Sequencher (Gene Codes Corp., Michigan) are available to align and compare sequences to the reference sequence and to each other.

The same material that was used to generate the CRS was sequenced again in 1999. Although 11 differences in nucleotides were reported, none of these differences were in the control region (Andrews et al., 1999). Thus, for the HV1 and HV2 regions that are used for forensic mtDNA analysis and reporting, the revised CRS (rCRS) does not differ from the CRS. One difference that was found upon resequencing was the loss of a cytosine at position 3106, making the rCRS standard now used one base shorter than the CRS (16,568 vs.1659 bp). To keep base numbers consistent with those reported previously, a deletion is reported for the rCRS at position 3107. Thus, all numbering is the same as in the original sequence, and each base is assigned a number from 1 to 16,569.

9.4.4 Screening Assays for mtDNA

The regions of the mtDNA genome that are sequenced are those that show the most variability between individuals: the hypervariable control regions, HV1 and HV2. Rapid screening techniques for mtDNA have been developed to determine quickly which samples can be excluded and therefore do not have to be sequenced. Although the variations in mtDNA are scattered throughout HV1 and HV2, there are certain hypervariable sites and regions (hotspots) where most of the variation occurs (Stoneking, 2000), and rapid screening methods concentrate primarily on these hotspots. Mitochondrial SNPs have been examined in these areas (Kline et al., 2005) and minisequencing methods (Tully et al., 1996) similar to that described above for nuclear chromosomal SNPs (e.g., SNaPshot technology) is sometimes used for this purpose (Quintans et al., 2004). In addition, many popular screening methods make use of sequence-specific oligonucleotide (SSO) probes (Stoneking et al., 1991), and these have been used in reverse dot blot and linear array assays (Comas et al., 1999; Gabriel et al., 2003). Rather than sequencing the entire hypervariable regions, SSO probes examine the most polymorphic sites in HV1 and HV2 through hybridization of PCR products to oligonucleotide probes designed to base-pair with the different sequence variants.

Roche Applied Sciences has developed the Linear Array Mitochondrial DNA HVI/HVII Region-Sequence Typing kit. This assay uses SSO probes to distinguish 19 polymorphic sites in 10 different locations within the control region, and it

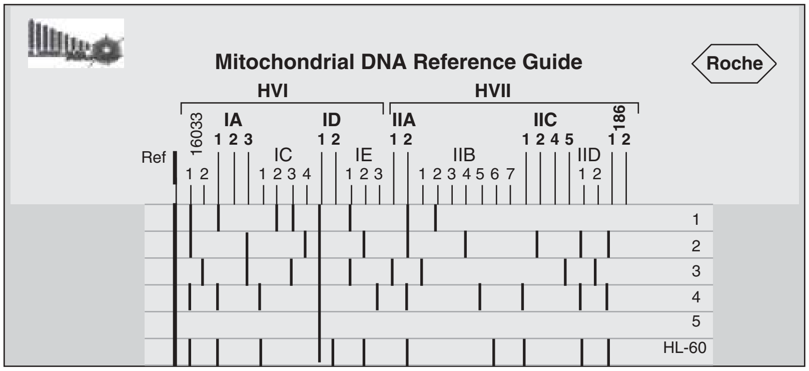

Figure 9.13 Roche linear array mitochondrial DNA HVI/HVII region-sequence typing kit panel results. Samples 1 to 4 are quality control DNA samples; sample 5 is a negative control; sample HL-60 is positive control DNA [HL-60 (NIST) SRM 2392.1]. (Reprinted with permission from Roche Applied Science.)

can do so in only 6 to 8 hours. The primer mix that comes with the kit contains four biotin-labeled primers that amplify the HV1 and HV2 regions simultaneously. Typing of the DNA is accomplished by hybridizing the labeled PCR products to panels containing 33 DNA probes which are striped in 31 lines, and then performing a reaction that produces a color product where fragments containing the biotin-labeled primers have hybridized (see Figure 9.13). In all regions examined, probe 1 is always the Anderson sequence (CRS). This kit has been used successfully for the examination of forensic mtDNA samples (Divne et al., 2005).

9.4.5 Interpretation of mtDNA Sequencing Results

There are FBI guidelines for interpreting mitochondrial sequence comparisons. Known and questioned samples are compared to the rCRS and a comparison of the two samples can result in an exclusion, a failure to exclude, or an inconclusive finding. Following are the guidelines recommended by the Scientific Working Group on DNA Analysis Methods for analyzing mitochondrial sequence data (SWGDAM, 2003).

Exclusion:

> If there are two or more differences in the base sequence of the known and the questioned sequences, one can conclude that there is no common origin between the two specimens and there is an *exclusion*.

Failure to exclude:

a. If there is no difference in sequence between the questioned and the known samples, one *cannot exclude a common origin*.
b. If there is heteroplasmy (described below) at the same site for both the known and questioned samples and there are identical bases at every other position, one *cannot exclude a common origin*.

c. If there is heteroplasmy (described below) at one position in one sample but not in the other, and every base at the other positions is identical, one *cannot exclude a common origin*.

d. If there is an ambiguous base (usually denoted as "N") at a specific position in either or both samples and identical bases at every other position, one *cannot exclude a common origin* for the two samples.

Inconclusive:

If there is only one base difference between the known and the questioned sequence, this result is *inconclusive*.

9.4.6 Statistics: The Meaning of a Match for mtDNA

When mtDNA analysis shows a "failure to exclude," a statistical calculation is required to find the meaning of a match. Unlike the 13 core nuclear STR loci whose individual frequencies can be multiplied because they assort independent of each other during cell division, all sites of the mitochondrial genome, and therefore all mtDNA polymorphisms, are linked on the single mitochondrial chromosome. The method currently used to describe the frequency of mtDNA sequences is the counting method; for example, if after the examination of 4000 samples a particular sequence appears for the first time, the frequency of this genetic profile is said to be 1 in 4001. If the sequence had been found once before, the frequency would be indicated as 2 in 4001. Although statistics such as these may not accurately reflect the actual frequency of any sequence because the population studied may be relatively small in size, the counting method is still the primary method used in forensic cases. Another way to express the frequency of mtDNA sequences is by providing 95% confidence limits and providing the upper and lower limits (the highest and lowest frequency estimates) to indicate the range of possible values for the actual frequency of the profile in the relevant population.

A major problem with mtDNA typing occurs because many sequences fall into categories called "most common types." These are particular sequences for HV1 and HV2 that are present in a disproportionately large portion of the population studied. As an example, an analysis of data for more than 1600 persons found that although the sequences for more than 50% of them were unique for the database, the most common type, which has the same sequence as the rCRS, was present 7.1% of the time. In addition, more than 20% of this Caucasian data set was made up of only 18 mtDNA types (Coble et al., 2004). There are numerous databases containing collections of mitochondrial sequence information (Butler, 2005, Chap. 10; Parson and Dür, 2007), and the FBI has compiled published mtDNA sequence data into the mtDNA population database (CODIS).

9.4.7 Heteroplasmy

Heteroplasmy is defined as the presence in a single person of more than one mitochondrial DNA type at a particular base position in the mtDNA sequence. Heteroplasmy can occur in sequence and/or length. Sequence heteroplasmy is detected when two nucleotides appear at a single site as overlapping peaks in a sequencing electropherogram. Length heteroplasmy occurs most often in the hypervariable regions in areas where there are long stretches of cytosines (C-stretches), that is, in positions 16,184 to 16,193 in HV1 and in positions 303 to 310 in HV2, and there are differing numbers of C's within the same sample.

Heteroplasmy can occur in the known and/or the questioned sample. If heteroplasmy at a specific position is found in both samples, this might help indicate an inclusion because heteroplasmy at a specific position can be an important identifying characteristic.

Heteroplasmy has been shown to differ from tissue to tissue in the same person (e.g., hair and blood may show different levels and types of heteroplasmy) (Salas et al., 2001). There may also be heteroplasmy differences in mtDNA populations between a person's other cells, and also differences within a single cell or within one mitochondrion (Butler, 2005, Chap. 6). All persons are thought to be heteroplasmic at some level, even if it is below the level of detection (Comas et al., 1995; Tully et al., 2000). High rates of heteroplasmy are not surprising since the mutation rate in mtDNA is much higher than the rate in nuclear genes; some regions of the mitochondrial genome have been shown to evolve at 6 to 17 times the rate of single-copy nuclear genes (Tully, 1999). In addition, although there is only one copy of a nuclear gene present within a human egg, that egg may contain 100,000 copies of the mitochondrial genome (Chen et al., 1995). Thus, multiple mutations and a good deal of heteroplasmy, some of which may be undetectable, could be present in a person.

9.4.8 The Future of DNA Analysis

As of June 2010, the National DNA Index (NDIS) portion of the FBI CODIS database contained almost 8.5 million profiles from convicted offenders and more than 324,318 forensic (evidentiary) profiles (see http://www.fbi.gov/hq/lab/codis/clickmap.htm for updates). Since STR analyses form the bulk of the profiles already present in CODIS as well as in other DNA databases throughout the world, changing to another type of DNA profiling would probably make current data unusable without reanalysis. Thus, STR typing is expected to continue as the preferred DNA profiling method for many years to come (Gill et al., 2004; Butler et al., 2007). However, since profiles from reduced-size core STRs are fully compatible with existing database information and because mini-STRs can broaden the range of degraded samples that can be analyzed, mini-STR analysis may eventually replace STR typing analysis with currently used primers (Gill et al., 2004). In addition, DNA markers such as SNPs will probably continue to be used along with STRs in order to obtain additional information about particular samples. In addition to the possibility of providing phenotype information (Budowle and van Daal 2008), SNPs are helpful in mtDNA testing and Y-SNPs can serve as lineage markers for Y-chromosome analysis (Butler et al., 2007).

Forensic DNA analysis is a rapidly progressing field, and new instruments are constantly being developed. Most of the advances currently being made in DNA analysis instrumentation are related to increasing the number of samples that can be processed at once (throughput), decreasing processing time and/or decreasing the size of the instruments that are used. To help speed up the analysis of reference samples, robotic instruments that automate DNA extraction and processing are being developed, and many are already in use in high-throughput laboratories. In addition, there is always a need for new user-friendly data analysis software and "expert" software to lessen the manual input required for accurate data interpretation.

The miniaturization of instruments used for forensic DNA analysis is a popular area of current research. Miniature thermal cyclers and small microchip capillary electrophoresis devices are being developed that greatly shorten analysis time and reduce the consumption of sample and of expensive reagents (Yeung et al., 2006; Greenspoon et al., 2007; Liu et al., 2007). A microfabricated capillary array electrophoresis device was recently tested and found to be appropriate for separating STRs (Greenspoon et al., 2008). The ultimate aim of equipment miniaturization is to produce instruments capable of integrating sample preparation, PCR, and DNA analysis which are also small enough and portable enough to be used at the scene of a crime or mass disaster. A prototype of such an instrument was recently developed and tested at a mock crime scene (Liu et al., 2008).

REFERENCES

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., and Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, **290**:457–465.

Andreasson, H., Gyllensten, U., and Allen, M. (2002). Real-time DNA quantification of nuclear and mitochondrial DNA in forensic analysis. *Biotechniques*, **33**:402–404, 407–411.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**:147.

Barbisin, M., Fang, R., O'Shea, C. E., Calandro, L. M., Furtado, M. R., and Shewale, J. G. (2009). Developmental validation of the Quantifiler Duo DNA Quantification kit for simultaneous quantification of total human and human male DNA and detection of PCR inhibitors in biological samples. *J. Forensic Sci.*, **54**:305–319.

Beckman, J. S., and Weber, J. L. (1992). Survey of human and rat microsatellites. *Genomics*, **12**:627–631.

Besseti, J. (2007). An introduction to PCR inhibitors. *Profiles DNA*, **10(1)**:9–10.

Biesecker, L. G., Bailey-Wilson, J. E., Ballantyne, J., Baum, H., Bieber, F. R., Brenner, C., Budowle, B., Butler, J. M., et al. (2005). Epidemiology. DNA identifications after the 9/11 World Trade Center attack. *Science*, **310**:1122–1123.

Brettell, T. A., Butler, J. M., and Saferstein, R. (2005). Forensic science. *Anal. Chem.*, **77**:3839–3860.

Brettell, T. A., Butler, J. M., and Almirall, J. R. (2007). Forensic science. *Anal. Chem.*, **79**:4365–4384.

CHAPTER 10

Current and Future Uses of DNA Microarrays in Forensic Science

NATHAN H. LENTS

Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Summary DNA microarrays have revolutionized basic research in molecular and cellular biology, biochemistry, and genetics. Through hybridization of labeled probes, this high-throughput technology allows the screening of tens or even hundreds of thousands of data points in a single run. The technology is most advanced with nucleic acids, but protein and antibody microarrays are coming of age as well. Because of the unique ability to screen for large numbers of molecules, such as DNA sequences, at once, the potential utility to forensic investigations is tremendous. Indeed, progress has been made demonstrating that microarrays are powerful tools of use to the forensic laboratory. As the technology matures and associated costs come down, the day that microarray analysis becomes a routine part of the forensic toolkit draws nearer.

10.1 Introduction 328
10.2 What is a DNA microarray? 328
  10.2.1 cDNA microarray 329
  10.2.2 Other types of DNA arrays 330
  10.2.3 The birth of “-omics” 331
10.3 DNA microarrays in toxicogenomics 332
  10.3.1 Sharing information 333
  10.3.2 Forensic application 333
10.4 Detection of microorganisms using microarrays 334
  10.4.1 Historical perspective 334
  10.4.2 DNA fingerprinting 335
  10.4.3 DNA fingerprinting by microarrays 336
  10.4.4 DNA sequence-based detection 337
  10.4.5 Where DNA microarrays come in 337

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

10.4.6 Looking forward: genetic virulence signatures 338
10.5 Probing human genomes by DNA microarrays 340
10.5.1 STR analysis 340
10.5.2 SNP analysis 343
10.5.3 Exploring an unknown genome? 344
Conclusion 345
References 345

10.1 INTRODUCTION

Perhaps once each decade, a technique or technology emerges that revolutionizes scientific research in a way that makes scientists wonder what we ever did without it. In the area of molecular biology, restriction enzymes and molecular cloning, and the polymerase chain reaction, come to mind as examples. More recently, the development of DNA microarray technology is just such an event, allowing molecular research to advance with a rapidity and precision unimaginable just a few years ago (Schena et al., 1995; Shoemaker et al., 2001). Indeed, like polymerase chain reaction (PCR) before it, microarrays have quickly become a seemingly irreplaceable implement in the arsenal of nearly all areas of molecular and genetic research. And as research continues, microarrays are now poised to have a similar impact on molecular forensic investigations.

10.2 WHAT IS A DNA MICROARRAY?

Simply put, DNA microarrays are small pieces of glass, usually the size of a standard microscope slide, onto which tiny spots of DNA have been "printed." The DNA spots on a given microarray, or *chip*, could number in the hundreds, thousands, or even *hundreds of thousands*, allowing a user to probe a given sample quantitatively for the presence of countless specific DNA sequences simultaneously (Schena et al., 1998). In basic research, myriad uses of the DNA microarray platform have evolved, and microarrays printed with other macromolecules, proteins and antibodies especially, have been developed as well (but are not discussed here) (Schena et al., 1998; Hoheisel, 2006; Hughes et al. 2006; Wu and Dewey, 2006; Hall et al., 2007; Werner, 2007; Klenkar and Liedberg, 2008). The information gained in a single cDNA microarray, for example, is equivalent to that of many thousands of gene-specific Northern blots or reverse transcription (RT)-PCR analyses (Allison et al., 2006; Ness, 2006). Thus, by the sheer volume of data obtained in a single run, it is no surprise that microarray technology has had such a revolutionary effect on basic research (Shoemaker et al., 2001). Table 10.1 lists the five major types of DNA microarray and their most common applications.

The use of microarrays in forensics is largely still at the developmental stage, but as research into this area continues, new and powerful possibilities are rapidly

TABLE 10.1 Common Types of DNA Microarrays

Type of Microarray	Printed Spots	Specific Use
cDNA microarray	18 to 30-mers, complementary to coding open reading frames	(a) Analysis of mRNA/miRNA transcription (expression profiling) (b) mRNA splicing analysis
Promoter microarray	PCR products or oligos from known gene promoters	Transcription factor location analysis
Entire chromosome tiling	Varies widely; high resolution: 15–25 bp spanning every ~35 bp of genomic sequence	(a) Transcription factor location analysis (b) Analysis of transcription
Comparative genomic hybridization	60-mers from throughout the genome (coding and noncoding)	Detection of gene deletion or amplification (e.g., during cancer)
SNP microarrays	25-mers	Detection of SNP alleles

emerging. In this chapter we first summarize the basics of microarray technology and the well-established uses in basic research. Then we detail some of the forensic applications of microarrays that are currently under development. Along the way we provide insight into future possibilities for the use of microarrays in forensic science.

10.2.1 cDNA Microarray

The most expansive use of DNA microarrays in scientific laboratories around the world is in the technique of *expression profiling* (Mandruzzato, 2007). Microarrays for this purpose are spotted with DNA molecules, usually 30 to 50-mers, representing complementary DNA sequences from within the coding regions of known or predicted genes (Kreil et al., 2006; Larsson et al. 2006). This platform has been developed so extensively that current high-end cDNA chips bear millions of spots, or *probes*, representing every currently known and predicted coding region in the human genome with multiple spots each. Entire or nearly entire genome coverage has also been achieved for mouse, rat, chicken, sheep, dog, cat, cow, macaque, *Xenopus* (frog), zebrafish, *Drosophila melanogaster, Caenorhabditis elegans*, maize, wheat, rice, barley, grape, cotton, tomato, soybean, sugarcane, *Arabidopsis thaliana*, plasmodium, several species of yeast, *Escherichia coli, Staphylococcus aureus*, and many other microorganisms, plants, animals, and fungi (Affymetrix web site).

cDNA microarrays allow researchers to take a “snapshot” of the gene expression profile, at the level of messenger RNA (mRNA), of a given cell population, as shown in Figure 10.1. Obviously, the utility of this approach is best found in the comparison of two populations: for example, cancer cells vs. tissue-matched normal cells, pre- and postviral infection, or pre- and postexposure to a drug or poison. In this way, the alterations of gene expression that occur as a result of such

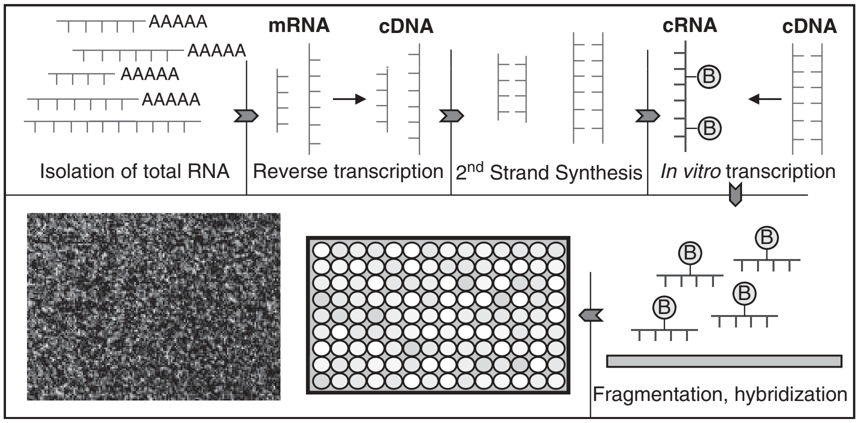

Figure 10.1 Expression profiling by cDNA microarray. Total RNA is isolated from living cells, followed by conversion of mRNA into single-stranded cDNA by reverse transcription using poly(dT) primers. Next, the cDNA is made double-stranded to serve as a template for in vitro transcription using biotinylated NTPs. The cRNA is then fragmented and hybridized onto the microarray platform. cRNA is visualized using a fluorophore-confugated streptavidin moiety and the microarrays are scanned at high resolution to allow quantitation of the fluorescent signal.

treatment conditions can be observed. Following repeated experiments in various settings, contexts, and conditions, a profile of gene expression events emerges. Furthermore, genes for micro-RNAs as well as many known and predicted mRNA splice variants can be discriminated using this same type of chip, and the complexity of the expression profiling data gleaned from microarrays is steadily increasing (Cuperlovic-Culf et al., 2006; Hughes et al., 2006; Kreil et al., 2006; Valencia-Sanchez et al., 2006). It might even be said that our ability to acquire vast amounts of exceedingly complex data is far outpacing our ability to interpret the data.

10.2.2 Other Types of DNA Arrays

Expression profiling is but one application of DNA microarray technology, concerned exclusively with coding regions of the genome. Promoter microarrays are another, displaying DNA sequences representative of gene promoter regions, that is, transcription start sights and proximal sequences of 1 to 10 kb (Ren and Dynlacht, 2004; Blais et al., 2005). Obviously, these types of arrays are not useful for expression studies but, rather, to probe exactly which genes a given transcription factor is physically associated with, at a certain point in time, through a technique called *ChIP-on-chip* or *location analysis* (Buck and Lieb, 2004; Blais et al., 2005; Lee et al., 2006). Other microarrays harbor sequences spanning an entire human chromosome, usually called *tiled arrays* due to the printing pattern of short sequence intervals, or *tiles* (Ishkanian et al., 2004; Graf et al., 2007). These arrays have several uses, including the comprehensive mapping of RNA transcripts and splice variants and chromosome-wide (and eventually true genome-wide) location analysis for transcription factors (Ji and Wang, 2005; Carninci, 2006; Huber et al., 2006). One dramatic example, research using these entire-chromosome arrays, has unexpectedly demonstrated that much more of the "noncoding" regions of the human genome are transcribed than thought previously (Bertone et al., 2004; Johnson et al., 2005). The last major type of DNA array is the comparative genomic hybridization (CGH) array (du Manoir et al., 1993; Kallioniemi et al., 1993). These types of arrays contain DNA sequences, both coding and noncoding, from throughout the genome and are used to detect changes in DNA copy number (gene deletion or gene amplification) that are frequently seen in human cancer and certain genetic diseases (Houldsworth and Chaganti, 1994; Yang et al., 2006; Stjernqvist et al., 2007).

These gene expression arrays, promoter arrays, tiling arrays, and CGH arrays are the most dominantly used microarray platforms in basic molecular research, including toxicogenomics (discussed below) (Ness, 2006). But for most forensic applications, the scrutiny at hand is of small discrete *differences* in closely related DNA sequences that differentiate one individual from another or one strain of bacteria from a close relative. Neither cDNA nor tiling arrays are appropriate for, or capable of, analyzing these subtle DNA differences. Thus, several varieties of microarrays for exactly this purpose have been developed, such as single-nucleotide polymorphism (SNP) arrays and resequencing arrays, which are discussed below. Because they are simply glass slides with spots of DNA, microarrays are easily customized to the application at hand, and countless clever variations have been documented (Mendoza et al., 1999; Lorenz et al., 2003; Dondeti et al., 2004; Thomson et al., 2004; Newton et al., 2005; McGrath et al., 2007). As development costs continually decrease, some laboratories even routinely print their own custom microarrays specific for their area of research and the experimental questions of interest (Ren et al., 2002; Cam et al., 2004; Balciunaite et al., 2005; Blais and Dynlacht, 2005a,b; Acosta-Alvear et al., 2007).

10.2.3 The Birth of "-Omics"

The complete sequencing of the human genome (and that of many other species) has affected nearly every field of biological research (Lander et al., 2001; Venter et al., 2001). Microarray technology is one of the major applications in which these enormous volumes of sequence data are put to direct use (Schena et al., 1998; Shoemaker et al., 2001; Hoheisel, 2006). Indeed, this sequence information, hand-in-hand with the required bioinformatics, was a necessary prerequisite for the development of microarray technology (Schena et al., 1995, 1998; Heller et al., 1997). With this new ability to perform such global genome-wide studies, entirely new scientific disciplines have materialized, and indeed, a new vocabulary has emerged. This is most recognizable by the addition of the suffix "-ome" or "-omics" onto various terms which implies that a global "in total" approach is being applied: for example, genome/genomics, proteome/proteomics, transcriptome, kinome, lipidome, metallome, spliceome, phosphoylome, and glycome (Palsson, 2002; Bishop, 2003, Chaussabel, 2004; Figeys, 2004). Obviously, these new "disciplines" are simply technological approaches to research within existing scientific fields, which are represented by their respective prefixes: for example, metabolomics and toxicogenomics (Aardema and MacGregor, 2002; Dettmer and Hammock, 2004). A new era of molecular research is clearly upon us, the -omics era, and microarrays are at the forefront of this information wave (Chauard et al., 2002; Abu-Issa and Kirby, 2004; Kiechle et al., 2004). In the next few pages we detail how current and future research will bring the -omics wave into the realm of forensic science through the use of DNA microarrays.

10.3 DNA MICROARRAYS IN TOXICOGENOMICS

No area of research has experienced a more profound impact due to advances in microarray technology than the field of genomics (Schena et al., 1998; Bueno-Filho et al., 2006). Stated simply, genomics is the study of the structure and function of DNA and DNA sequences on the genome scale. The field of genomics and, similarly, microarray technology, is intimately connected to the success of the mass sequencing of the human genome (and that of other species) (Schena et al., 1998; Belacel et al., 2006; Lerman et al., 2007). Toxicogenomics is a subdiscipline of genomics that is concerned with deciphering the impact of exogenous molecules such as drugs, poisons, and xenobiotic compounds on gene expression (Lovett, 2000; Aardema and MacGregor, 2002; Gershon, 2002; Hamadeh et al., 2002; Gatzidou et al., 2007). As this name implies, this field was the result of a merger between traditional toxicology research and functional genomics: microarrays are the platform that made this merger possible (Neumann and Galvez, 2002; Pennie, 2002; Ulrich and Friend, 2002; Irwin et al., 2004).

Mechanistic toxicology is the attempt to decipher why and how a drug or poison does what it does. In this way, toxicology is closely related to pharmacology in the search for a "mechanism of action." Predictive toxicology attempts to use this physiological information to make predictions about the toxicity of a related or novel molecule. Toxico*genomics* has greatly accelerated both approaches and deepened our understanding of the biological perturbations caused by chemical compounds (Cunningham, 2006; Fielden and Kolaja, 2006). Drugs and poisons are now grouped in increasingly narrow categories based on common induction of discrete cellular gene expression changes (Oberemm et al., 2005). By aligning physiological response to chemical structures, reliable predictions can be made regarding the toxic effects of a previously untested compound (Raghavan et al., 2005; Fielden and Kolaja, 2006; Martin et al., 2006).

Forensic toxicology is often concerned with the converse situation: the identification of a poisoning agent (which itself may be difficult or impossible to detect), based on the biological effects of that agent. This is often a tenuous process, but a more complete picture of toxin-induced gene expression responses, achieved by microarrays, gives forensic researchers much more to work with in the hunt for specific telltale biomarkers (Lettieri, 2006). Indeed, significant progress has been made in the elucidation of a toxicological "fingerprint" for several drugs and toxins, including arsenic, chromium, nickel, cadmium, methapyriline, clofibrate, hexachlorobenzene, doxorubicin, 5-fluorouracil, and acetaminophen [reviewed by Lettieri (2006)]. The dominant type of microarray technology employed in toxicogenomics is the cDNA microarray for expression profiling (Figure 10.1), although other types of microarrays, such as splicing arrays and CGH arrays, have revealed important pharmacological effects as well (Lee et al., 2005).

10.3.1 Sharing Information

For the voluminous amounts of data gleaned from toxicogenomic studies to be truly useful to the community at large, there must be in place a rigorous common method of analyzing and sharing data. There are several national and international efforts to do just that. The first critical milestone in the creation of public databases for genomic data was the establishment of the MIAME (minimum information about a microarray experiment) guidelines, developed by the European Bioinformatics Institute (EBI) (Brazma et al., 2001). These guidelines provide a common language through which researchers around the world can understand each other's experimental data. Following these guidelines, several publicly accessible repositories for microarray data have been created, the largest being the Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett and Edgar, 2006). Within the field of toxicology, the National Center for Toxicology (NCT) within the National Institute of Environmental Health Science (NIEHS) is developing the Chemical Effects in Biological Systems (CEBS) knowledge base (Waters et al., 2003). CEBS, and its counterpart at the EBI, tox-MIAMExpress, aim to compile and integrate vast amounts of genomic data and thus build a complete picture of the complex biological response to toxicological agents and provide that information to the public (Brazma et al., 2003, 2006).

10.3.2 Forensic Application

Our ability to apply what is learned in experimental toxicogenomics to forensic toxicology rests not only on our understanding of the biological response to a given chemical agent, but also on our ability to detect that response in biological tissues, which often present themselves as evidence in very trace amounts or in degraded condition (Nygaard and Hovig, 2006). This presents a serious problem for gene expression analysis, as mRNA is arguably the most labile biological macromolecule of all. But this challenge is certainly not insurmountable and does not apply whatsoever to freshly harvested samples, as is also often possible in forensic practice. But even when access to fresh tissue sample is not possible, conceivable breakthroughs in microarray technology could one day allow recreation of gene expression information. For example, research into chemical modifications of DNA histones has revealed that gene expression is heavily influenced by chromatin architecture, and the precise structure–function relationships of the "histone code" are being teased out methodically (Jenuwein et al., 2001; Turner, 2002).

It seems increasingly likely that scientists will one day be able to infer with relative certainty the level of active transcription of a given gene simply by examining the chromatin structure of the gene promoter (Gilmore and Washburn, 2007). This is relevant to forensic investigations because chromatin is formed by DNA–protein interactions and is far more stable than mRNA. Further, genome-wide scanning of chromatin structure by microarray is already possible through the technology of modified histone-specific ChIP-on-chip (or location analysis) (Roh et al., 2004; Huebert et al., 2006). Regardless of the approach taken to scrutinize gene expression, the unparalleled multiplex capacity of microarrays will be absolutely essential for detecting these residual biological "fingerprints" of drug exposure in a tissue because it will probably involve the altered expression of hundreds of genes. Examining a small number of genes would be far too ambiguous for forensic casework. Without microarrays, the immense labor of examining the expression of hundreds of genes manually makes it difficult to envision the reasonable forensic utility of establishing gene expression signatures in response to drug or toxin exposure.

10.4 DETECTION OF MICROORGANISMS USING MICROARRAYS

DNA microarrays are powerful tools for the rapid detection and identification of microorganisms, and there are countless scenarios in which the identification of a potentially infectious agent is of paramount importance. Indeed, the improved detection of biothreats tropic to humans, livestock, or agricultural plants is a major focus of research efforts at the Centers for Disease Control, the Food and Drug Administration, the National Institutes of Health, the Department of Health and Human Services, and countless nongovernmental research centers. Beyond malicious and bioterrorism-related scenarios, rapid pathogen identification has enormous lifesaving potential in the area of infectious disease diagnostics. Furthermore, and particularly relevant to the current text, precise recognition of microorganisms can be crucial to the investigation of criminal activity. Examples include the detection of microorganisms in deliberate or accidental biopoisoning, and the analysis of bacterial flora to scrutinize or match samples of soil, water, and biological specimens or fluids.

Microarrays are particularly well suited for these purposes for at least three reasons. First, the ability to multiplex with microarrays (i.e., screen for multiple targets simultaneously) far exceeds that of any other method. Second, because there is no need for time-consuming bioassays or decision-tree iterative protocols, microarrays are usually far faster than most currently dominant strategies. Finally, and most powerfully, microarray platforms can be designed intelligently such that little or no prior knowledge of the nature of the microorganism being detected is necessary. All of these advantages have already been demonstrated, as we discuss below. Although microarray technology is not yet commonplace in the forensic analysis of microorganisms, the rapid development and declining costs of the technology will probably lead to increased utilization.

10.4.1 Historical Perspective

Since the invention of the microscope, classifying the various types of microscopic life has challenged and fascinated microbiologists. But the association of various bacteria and fungi with illness brought these classification and identification efforts into the realm of public health. (Because they are even smaller than the maximum resolving power of light microscopes, viruses were not discovered until centuries later.) The Gram stain, which distinguishes bacteria into two classes based on cell wall characteristics, was among the first biochemical techniques in the classification of bacteria and is still in use today. Microscopic size and physical appearance (circular, rod, or corkscrew shape); colony size, shape, and color; growth rate; nutritional requirements; toxicity to laboratory animals; and even traits such as odor once helped scientists tell one species of bacteria from another. Needless to say, all of these methods are laborious and none distinguishes closely related strains or species.

More recently, serological methods have largely supplanted these descriptive techniques. Briefly, *serotyping* is the classification of bacteria based on the presence of certain “markers,” usually receptor proteins, that can be detected easily using antibody-based procedures such as ELISA (enzyme-linked immunosorbant assay) (Yolken, 1978). For some pathogenic bacteria, detection assays are designed toward the very receptor that comprises the pathogenic component (Ashkenazi and Cleary, 1989; Munoz et al., 1990; Hearty et al., 2006). This important feature affords these assays a reasonable chance of success in detecting most virulent strains (of that particular species), even, potentially, novel ones. For this reason, along with the rapidity, ease, and negligible expense, serotyping is likely to remain a dominant means of identifying and classifying bacteria (Cebula et al., 2005a).

10.4.2 DNA Fingerprinting

In an effort to distinguish closely related strains of microorganisms, microbiologists and forensic specialists are working to establish a working library of strain-specific genetic fingerprints or *signatures* (Pannucci et al., 2004). These are akin to the DNA fingerprinting that is conducted for identity confirmation of humans using short tandem repeat (STR) or SNP analysis. Knowing the unique genetic features harbored by various strains of pathogenic and benign bacteria, viruses, and fungi will allow investigators to quickly screen biological samples for the presence of these known agents and discriminate between harmless flora and their close virulent relatives.

Several approaches have been taken in an effort to establish DNA fingerprints for individual bacterial strains. RFLP (restriction fragment length polymorphism) analysis of bacterial rRNA genes is one example but fails to distinguish closely related strains or even species (Brindley et al., 1991; Pebay et al., 1992; Alonso et al., 1993; Sriprakash and Gardiner, 1997). RFLP analysis of the entire bacterial chromosome, followed by PFGE (pulse-field gel electrophoresis) is a more dominant technique in bacterial DNA fingerprinting (Linhardt et al., 1992; Gundisch et al., 1993). However, the very large DNA fragments generated by this technique (2 to 5 Mbp) cannot be resolved with great precision. Thus, strain-specific mutations and even larger genomic rearrangements (due to recombination events) could result in no detectable difference in the RFLP fingerprint. In that case, two significantly divergent strains could be indistinguishable by RFLP. Conversely, an otherwise insignificant or even silent mutation that happens to reside within a restriction site could result in a markedly different RFLP fingerprint with little or no phenotypic difference. This slight genetic difference hardly gives rise to a distinct bacterial strain. The point made by both cases is that RFLP analysis simply compares relative sizes of random chunks of the genome. Thus, while perfectly appropriate for identification of humans, RFLP analysis is somewhat clumsy for the identification of rapidly evolving and mutating microorganisms. DNA sequence–based methods would be inherently more discriminating and yield far more information about discrete genetic differences between closely related microbiota.

More recently, amplification-based approaches have emerged that track amplification fragment length polymorphisms (AFLPs) (Janssen et al., 1996; Lin et al., 1996). While some employ random priming and others target specific genomic regions, the readout is again a "fingerprint" of random fragment sizes (Vaneechoutte, 1996). These techniques have proven to be sufficiently discriminating to differentiate closely related strains of bacteria and races of fungi (Keim et al., 1997; Kis-Papo et al., 2003; Hynes et al., 2006). But once again, because the objects of these analyses are (1) chunks of DNA (bands on a gel), not sequences, and (2) totally unrelated to any phenotypic difference, such as pathogenicity, these techniques offer very little genetic insight into the nature of the microorganism of interest. In this new era of high-throughput entire-genome sequencing, sequence-independent techniques for genetic characterization of organisms seem blunt indeed.

10.4.3 DNA Fingerprinting by Microarrays

Microarray technology has now been employed in an effort to establish genetic fingerprints for microorganisms (Willse et al., 2005; Doran et al., 2007). One such example involves the amplification of established repetitive extragenic palindromic (REP) sequences, followed by hybridization onto a small microarray harboring about 200 random nonamers (Willse et al., 2004). This approach was successful in clearly distinguishing 25 closely related strains of *Salmonella enterica* (Willse et al., 2004). A microarray-based genetic fingerprinting system offers distinct advantages over electromobility-based systems. First, the increased complexity of the fingerprint allows much greater precision and statistical confidence, important features for any forensic technology that must withstand scrutiny in a court of law. Second, microarrays afford the potential to screen for hundreds of strains of hundreds of species simultaneously. Third, fingerprints from previously unobserved isolates could be compared to those of known strains and some knowledge of genetic heritage could be inferred. Indeed, bioinformatic methods of designing robust and precise DNA fingerprinting arrays for pathogen detection are maturing (Satya et al., 2008). However, similar to other fingerprinting methods, while microarray fingerprints of repetitive sequences could provide clues into the nature of known or novel microbes, important phenotypic characteristics, such as virulence, could not be predicted with confidence.

10.4.4 DNA Sequence-Based Detection

Single-nucleotide polymorphisms (SNPs) represent the smallest and most discrete genetic difference between two living organisms. These single-base-pair differences can result in a slightly different protein product (a nonsynonymous SNP) or not (a synonymous SNP), based on the degeneracy of the genetic code. Regardless, the comparison of SNPs between closely related strains of bacteria, viruses, or fungi allows a much more thorough appreciation of genetic differences and facilitates the recapitulation of evolutionary lineage (Bochner, 2003). The importance of this cannot be overstated. Consider the hypothetical example of the emergence of a novel strain of *Salmonella enterica* in a poultry packing plant; considering its unique RFLP or AFLP pattern would provide little information other than that this strain is *not* one that has been observed previously. In contrast, a SNP analysis would reveal which known strains of *Salmonella* are the closest genetic relatives to this new strain, and thus a quick and reasonably accurate prediction could be made as to whether or not this new strain is a health hazard. (Prediction of virulence is discussed further in the next section.)

The painstaking process of using SNPs derived from sequencing data to classify strains of bacteria into groups based on hereditary relationships involves statistical analyses, cladisitics, and computer-assisted hierarchical clustering [see Cebula et al. (2005a,b) for two excellent reviews]. Nevertheless, to date, SNP analyses have been used to resolve and identify strains of *Bacillus anthracis, Mycobacterium tuberculosis, Helicobacter pylori, Escherichia coli, Shigella dysenteriae*, and *Salmonella enterica* (Levy et al., 2004; Pearson et al., 2004; Cebula et al., 2005b; Filliol et al., 2006; Van Ert et al., 2007). The most famous example is that of the SNP analysis used to confirm that the multiple *B. anthracis* samples of the fall 2001 anthrax terrorist attack were of the same genetic strain (Hoffmaster et al., 2002).

10.4.5 Where DNA Microarrays Come In

Microarrays afford the possibility of rapid and thorough SNP analyses without having to perform actual sequencing reactions. One such technique, known as *resequencing*, first involves the amplification (and fluorescent labeling) of several specific regions of interest (i.e., those with well-established strain-specific SNPs) from a sample containing an unknown bacterial strain (Hacia, 1999). Target amplicon size is kept relatively small, about 200 bp, and amplicons are denatured and fragmented prior to hybridization. The probe spots on the microarray are small oligonucleotides around 25 bp in length containing a variable SNP at the central position. Each SNP of interest comprises eight spots on the chip: all four possible variations of the SNP, in both DNA strand orientations. Thus, for each biological sample, the two brightest spots will correspond to the SNP version harbored by that particular bacterium, virus, and so on. Of course, the beauty of the microarray is that an essentially unlimited number of SNPs could be analyzed in a single scan (Zwick et al., 2005; Malanoski et al., 2006; Wang et al., 2006). Further, selections of SNPs from many hundreds or even thousands of bacterial species could be printed on a single array. Thus, one could easily envision a single microarray that provides coverage for all known enteric bacteria, for example. It is worth noting that because the resequencing technique involves an amplification step with a fairly small amplicon, only minute amounts of starting material are required.

An approach similar to resequencing, called *minisequencing*, represents another microarray-based method of SNP analysis (Huber et al., 2001; Lindroos et al., 2002). However, the small amplicons containing the SNPs of interest are not fluorescently labeled. Further, the 25-bp probes on the microarray do not harbor the SNP itself. Rather, they are the 25 bp of sequence immediately $5'$ of the SNP base. Either prior to hybridization (in-solution minisequencing), or following hybridization (in situ minisequencing), of the targets to the probes, DNA polymerization reactions are performed to extend the probes, but only chain-terminating di-deoxynucleotides (fluorescently labeled) are provided. Thus, the DNA polymerase can only add one base to the probes before chain termination, but this $n + 1$ position will correspond to the SNP base with a coded fluorescent label for each of the four possible bases, as revealed by confocal fluorescent scanning of the chip (see Figure 10.2). The minisequencing approach has two advantages over resequencing. First, it only requires two spots for each SNP rather than eight. (Again, both DNA strands are analyzed as an internal duplicate sample.) Second and more important, the data scan of different color fluorescent spots is unambiguous and far more straightforward to analyze and interpret than the single-color scanning of resequencing. However, the additional chemistry and manipulation of the polymerization step adds cost, labor, and experimental variability, important considerations in the search for a highly automated high-throughput platform for forensic application.

10.4.6 Looking Forward: Genetic Virulence Signatures

Bioinformatics researchers are currently mining the ever-increasing volumes of complete genome sequences from various strains of bacteria in search of “virulence signatures” (Pannucci et al., 2004). The rationale for this search is twofold. First, if a given DNA motif could be established as “necessary” for virulence (and thus present in all pathogenic strains, absent in all innocuous strains), this would represent a potentially momentous breakthrough for the basic scientists investigating mechanisms of microbial pathogenesis. Second, and more germane to the forensics realm, knowing the “virulence signature” of pathogenic strains of *E. coli*, for example, would give investigators a robust tool in detecting the appearance of those pathogens, even, at least in theory, those that have not yet been discovered or are the product of malevolent bioengineering.

It goes without saying that the “training set” of sequence data regarding these virulence signatures would need to be monitored and updated continually as more strains, of both the pathogenic and benign variety, emerge. Experimental validation of the pathogenic mechanisms of virulence-associated DNA sequences should be sought in as timely a manner as possible because of the potential to “overfit” the sequence data, honing in on sequences that are merely correlative and thus not in any way connected with pathogenesis. Examples of “red herring” sequences could

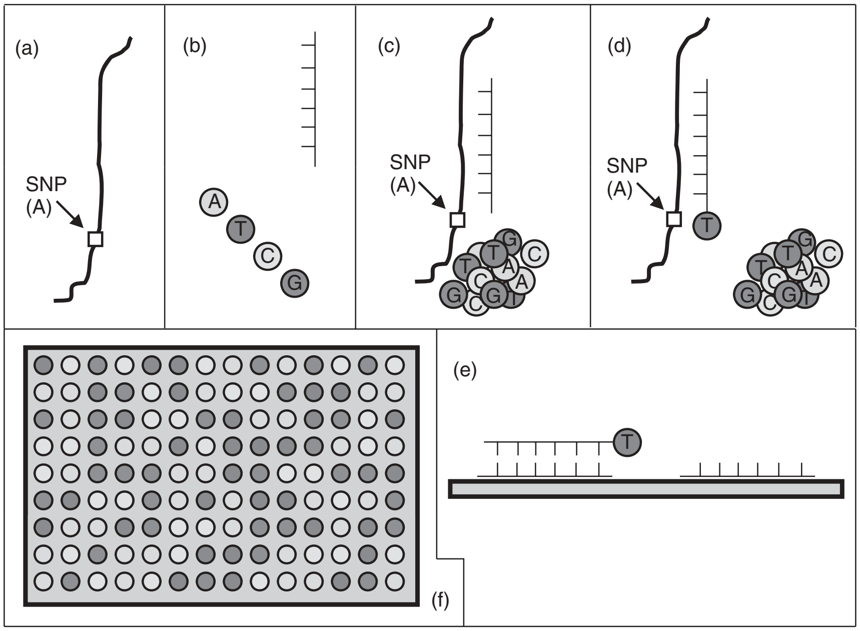

Figure 10.2 SNP analysis by in-solution minisequencing using a four-color microarray. (A) First, the genomic regions containing the target SNPs are amplified by small-amplicon PCR. (B) Then minisequencing reactions are carried out in solution using only dideoxynucleotides (fluorescently labeled). (C) The sequencing primers hybridize with the 3′ end oriented one base upstream of the SNPs. (D) The sequencing primers are extended by one labeled base, complementary to each SNP. (E) The mixture is denatured and hybridized to a microarray printed with DNA sequences containing the SNPs. (F) The microarray containing thousands of SNPs is scanned and each spot is called based on fluorescence.

lead to potentially disastrous false-negative readouts during the screening of an unknown. This potential pitfall can only be overcome by experimental rigor and the diligent reporting and sharing of sequence data.

Although traditional sequencing efforts and "shotgun sequencing" had previously been at the forefront of the search for bacterial and viral genetic fingerprints and virulence signatures, the emergence of relatively low-cost whole-genome sequencing is poised to accelerate this research dramatically (Lorenz, 2002; Read et al., 2003; Pannucci, et al., 2004; Dorrell et al., 2005; Afset et al., 2006; Burrack and Higgins, 2007). Comparative genomics and bioinformatics, followed by experimental validation, will probably illuminate the key genomic elements that confer the pathogenic phenotype(s) for each organism. Indeed, in a spectacularly successful example, a detailed virulence signature for *Staphylococcus aureus* has been established, corresponding to genetic elements in 84 different genes, including those for antibioitic resistance (Spence et al., 2008). Such information is immediately applicable toward building a robust and extremely sensitive microarray that could be used to detect the presence of a virulent form of *S. aureus* with relative certainty.

Once established and validated, virulence signatures for multiple pathogens (each in tens or hundreds of naturally occurring degenerate permutations) can be spotted onto DNA microarrays. Because the spot density capability of DNA chips is so high and increasing (currently >1,000,000 unique probes per chip is possible), it is conceivable that a single chip could harbor the virulence signatures of thousands of different pathogens, allowing broad but precise screening of biological samples in a single pass.

Although the use of microarrays in the forensic identification of microorganisms is still in its infancy, the enormous advantages justify further development of this technology. As costs continue to decline and access increases, it is not inconceivable that microarrays will, one day soon, become indispensables tools for the rapid and reliable identification of microbial strains.

10.5 PROBING HUMAN GENOMES BY DNA MICROARRAYS

The identification of human beings by "DNA fingerprint" has, in a relatively short time, revolutionized forensic science and criminal proceedings (Higuchi et al., 1988; Kobilinsky and Levine, 1988; Lee et al., 1994). This revolution cannot possibly be overstated. However, recent advances in molecular biology and genomics have made possible even further development of forensic DNA technology. Despite this, the extent of DNA technology in most forensic laboratories is currently limited to "DNA fingerprinting" for matching an unknown biological sample to that of a known person. In that regard, DNA microarrays offer several advantages over traditional STR analysis, including adding significant automation, quantitative analysis, and statistical power, as well as further extending the technique to even smaller and more degraded DNA samples than is currently possible. In addition, looking beyond, DNA microarrays may one day allow forensic investigators to transform an unknown and therefore unhelpful DNA sample into a crucial clue for investigating crime.

10.5.1 STR Analysis

Short tandem repeat (STR) analysis is the current technique of choice for identifying unique individual markers of a given DNA sample, and in most cases provides such immense discriminating power that the U.S. and British judicial systems fully accept STR results into a court of law, provided that the technical execution is within established protocols. This technique exploits the polymorphic nature of some repetitive sequences within the human genome, called *microsatellite DNA*. These regions consist of a simple DNA sequence, 3 to 7 bp in length, that is repeated anywhere from 3 to 15 times. [There exists examples of both shorter and longer repeats (satellite or minisatellite DNA) and regions where many hundreds or thousands of repeats exist in tandem, but these are not typically part of forensic STR analysis.] Biallelic in nature (one each on the paternal and maternal chromosome), these microsatellites, or STRs, are highly polymorphic and stably inherited. That is, a given person will have two alleles for a given STR (e.g., 12 and 8 repeats on each of a pair of homologous chromosomes, one inherited from her biological mother, the other from her father), and one will be passed on to each of her children at random. Thus, two nonrelated individuals will have a very low probability of having the same two alleles for this given STR. Even siblings have only a 25% chance of inheriting the same two numbers of repeats for each and every STR. When examining multiple STRs, the likelihood that two different individuals will harbor the same alleles drops dramatically with each additional STR included in the analysis. Accordingly, the standard U.S. Department of Justice's CODIS protocol calls for an examination of 13 established STR loci, which places the statistical odds of two nonidentical individuals having the same two alleles for all 13 loci at less than 2×10^{-15}, a number whose reciprocal is more than 10,000 times more than the number of human persons who have ever walked the Earth (Lee et al., 1994; McEwen, 1995; Budowle et al., 1999; Ruitberg et al., 2001). (Government officials are currently deliberating an increase in the number of CODIS STR loci to 20.)

Clearly, the discovery, development, and careful propagation of STR analysis represented an unparalleled triumph in the application of molecular biology to criminal justice, and it may be tempting to think that current abilities of DNA analysis already meet or exceed their usefulness, with no further improvement necessary. However, a critical examination of current techniques in STR analysis reveals several compelling reasons to pursue further development in this area. First, the approach of STR analysis itself could benefit with increased automation, multiplexing, and high-throughput capabilities, as current protocols call for extensive sample handling and manipulation and significant technician labor. In addition to the costs of time and labor, these manipulations bring opportunities for experimental error which could potentially jeopardize the trustworthiness of a given DNA match. Although the PCR amplifications can now be multiplexed routinely and efficiently (reducing labor and handling), traditional STR analysis still requires forensic researchers to electro-resolve the PCR products and make judgments regarding relative mobility rates of these DNA fragments, as compared to known DNA standards. The result is a DNA "fingerprint" that is somewhat qualitative, rather than strictly quantitative, as is most desirable when objectivity is of such supreme importance.

Without altering the basic premise, small DNA microarrays can be used in the execution of traditional STR analysis. A major challenge in any microarray-based, and thus hybridization-based, approach to forensic STR analysis is the detection of subtle differences in sequence repeat number. For example, while discriminating between 3 and 15 repeats would be relatively straightforward, it is exceedingly difficult to discriminate between 9 and 10 repeats using differential hybridization onto a standard microarray with the necessary degree confidence. Accordingly, researchers have devised two creative solutions to this obstacle. First, electronically active microarrays can be utilized to establish and fine-tune an extremely high stringency hybridization environment (Radtkey et al., 2000). In this way, even the subtlest differences in hybridization energies can be exaggerated and the results are clear spot calls on the microarray, indicating the exact repeat number for each

STR with precision. For this reason, the technology of electronic microarrays has found its way into several microarray platforms (including forensic SNP analysis, discussed below) and development continues (Heller, 2002).

More recently, a second approach to array-based STR analysis has been reported that involves the use of standard glass microarrays and a sequence of enzymatic steps (Kemp et al., 2005). In this somewhat complicated method, called variable-length probe array (VLPA), the microarray spots harbor single-stranded probes containing STRs of various lengths (repeat numbers) bound to the microarray surface at their 5′ end. Following amplification and 5′ fluorescent labeling of STR regions in the target DNA sample, the labeled amplicons are hybridized to the microarray. Obviously, the targets will hybridize to probes with both more and fewer numbers of STR repeats than those of the target itself. To ensure proper alignment at each STR set, complementary *clamp sequences* are included at the 5′ end of the probe DNAs and the 3′ end of the target sequences. Thus, at probe spots containing *fewer* repeats than the hybridized target sequence, the 5′ end of the target DNA (which contains the label) will overhang. At probe spots containing *more* repeats than the target DNA, the spotted 3′ end of the probe DNA will overhang the target DNA. Next, the slides are treated with S1 endonuclease, which digests single-stranded DNA. Thus, all overhangs are removed and, importantly, the fluorescent label will be released from any target DNA that is bound to a probe spot containing an STR sequence that is shorter (fewer repeats) than the target DNA. Conversely, the fluorescent label will remain on all target DNAs that are bound at spots containing probes of equal or greater length, despite the digestion of the 3′ end of probe DNAs that are longer than the target DNA. For example, if a target DNA sample contains 11 repeats of a given STR, fluorescent signals of approximately equal intensities would be detected at probe spots harboring 11, 12, 13, 14,... repeats, but no signal whatsoever would be detected at probe spots harboring 10, 9, 8, 7,... repeats. Although it is not yet clear how this system can delineate heterozygous alleles, a major advantage of this approach is that the spot calls are unambiguous (signal or no signal) and thus easy to automate in a high-throughput manner. Creative use of the clamp sequence could also allow a great deal of multiplexing of many samples simultaneously.

A third microarray-based approach toward STR analysis that has been reported involves the use of the branch migration capabilities of double-stranded DNA. This assay begins similarly to VLPA with hybridization of labeled target DNA to probes of varying length using a clamp sequence to ensure alignment. Next, a succession of two competing probes in opposite orientations initiate branch migration to displace probe–target interactions that are not perfect matches for the number of tandem repeats. Thus, array scanning will elicit signal from only one spot per STR allele, and heterozygosity can be detected and delineated using an algorithm specially designed for this system (Pourmand et al., 2007). In summary, several options for array-based STR analysis have been developed, each with distinct advantages and caveats. The major advantage of any microarray-based platform is the enormous potential for rapid high-throughput screening of thousands of samples simultaneously. And as microarray costs continue to decline, the feasibility of their routine use in forensic applications increases.

10.5.2 SNP Analysis

Although the power of STRs in forensic investigations is truly remarkable, there are limitations to STR analysis that warrant continued research in alternative technologies. Specifically, because each STR locus is first amplified in a PCR reaction, there is a lower limit for required template, beneath which amplification is too inefficient and unreliable for forensic purposes. This minimum sample size is indeed minute, allowing complete DNA profiles to be collected from scant physical evidence such as one intact hair follicle. However, biological material that serves as a template in a PCR reaction is consumed in the process, and when the amount of sample is limiting, as is often true in forensic investigations, there is little or no margin for error. Currently, the lower limit of input DNA to achieve a reliable CODIS (FBI's Combined DNA Index System) DNA profile is between 100 and 200 pg of pristine human genomic DNA (Lorente et al., 1997; Krenke et al., 2002; Leclair et al., 2003). With 6 pg of DNA in each human nucleus, this means that in the best of circumstances and with flawless technique, a biological sample must contain at least 15 to 30 intact cells with undamaged DNA. Although this is certainly an impressive level of detection, the collection of trace forensic evidence often falls below this standard in quantity or quality of DNA. Following the death of a human cell, DNA begins the steady process of decay, and trace DNA specimens from hair or dead skin cells collected at a crime scene are often not of adequate quality for any type of STR analysis.

Single-nucleotide polymorphism (SNP) analysis provides an alternative to STR analysis with much lower requirements of quantity and quality of input DNA, thus allowing further reduction of the threshold requirements of biological material from which a quality DNA profile can be extracted (Walsh, 2004; Nygaard and Havig, 2006). This increased sensitivity arises due to the amplification of significantly smaller amplicons. Unavoidably, forensic STR analysis involves the amplification of repeat sequences as large as 400 bp, and there is a direct relationship between the size of an amplicon and the amount of intact template DNA required for efficient amplification. Because SNPs are a single base in length, amplicons can be designed that are less than 50 bp in length. Thus, SNP analysis is far more robust than STR analysis with trace and/or degraded DNA samples. However, because there are only four possible alleles for each SNP (A, T, C, or G), rather than 7 to 22 possible alleles for each CODIS STR, a far greater panel of SNPs must be examined in order to achieve discriminating power similar to that of forensic STR analysis. For the level of certainty required of a criminal prosecution, at least 50 SNPs would be needed for a positive identification, and a panel of about 100 SNPs would probably be appropriate for a universal forensic protocol. It is this large number of individual loci that need to be examined that has probably discouraged more widespread application of SNP analysis in forensic DNA matching. However, microarrays circumvent this challenge completely, and hundreds or even thousands of SNPs could be analyzed simultaneously with relative ease. Further, clamping sequences or a similar mechanism could facilitate complete SNP screening for multiple (even hundreds) of different biological samples on a single chip.

Human SNP analysis by microarray is already a well-established technology, although application is limited primarily to basic research and clinical diagnosis of genetic diseases. Further, the current dbSNP database at the National Institutes of Health contains millions of documented SNPs in the human population (NCBI web site, http://www.ncbi.nlm.nih.gov/projects/SNP/). Thus, it will not require great strides to bring the technology of SNP microarrays to the forensic realm. Indeed, electronically active arrays have already proven their utility in forensic applications (Borsting et al., 2004, 2005; Huang et al., 2004; Tsang et al., 2004). Exactly which cadre of SNPs will perform best for forensic purposes remains to be seen, but among the most important considerations are high variability of alleles in the human population and ease of amplification in a large multiplex PCR. The latter factor is determined by the genomic "neighborhood" of each SNP and must be tested empirically. A microarray-based strategy for detecting SNPs involves either resequencing or minisequencing (see Section 10.4.4 and Figure 10.2). Although to date, current reports of forensic SNP analysis by microarray have utilized in-solution minisequencing prior to hybridization, resequencing microarrays could prove just as powerful and would involve less labor and manipulation (Allen and Divne 2005; Divne and Allen, 2005).

10.5.3 Exploring an Unknown Genome?

Consider the example, as happens often, when a DNA sample collected from a crime scene does not match with a known DNA profile in any database or from a possible suspect. Currently, the unknown DNA sample would then be considered a "dead end," useful only for screening additional suspect(s), if they are somehow identified by further criminal investigation. However, the DNA sample itself is an entire genome, containing much more information than the simple DNA fingerprints that are used to match samples. Within the DNA sequence of a genome is all the genetic information necessary to build that person, and *only* that person. Thus, a DNA sample with no known matches could be far from a dead-end. It could potentially tell us much about the perpetrator him- or herself, including such genetic traits as eye color, hair color, race and ethnic origin, approximate height/weight/build, and genetic diseases or predispositions. When investigating a crime, such information could be extremely valuable for narrowing down overwhelmingly large pools of suspects into just a few people or one person. Focusing on a smaller suspect pool would greatly reduce the disturbance of innocent people, as well increase the likelihood that detectives could discover additional evidence implicating the guilty party. Further, if a suspect pool was narrowed to a sufficiently small number of people, a court of law could then compel the collection of DNA samples to facilitate a comprehensive match.

Forensic research into the precise genetic basis for physical traits such as race/ethnic origin and eye color is already proceeding rapidly (Frudakis et al., 2003; Rowold and Herrera, 2003; Sturm and Frudakis, 2004; Amorim and Pereira, 2005; Ray et al., 2005; Sobrino and Carracedo, 2005; Wetton et al., 2005; Lao et al., 2006; Kayser et al., 2008). However, information of this type will be most valuable to forensic science when many traits are analyzed at once; for example, "the DNA sample left at the crime scene was from a woman of mixed Central European and East Asian descent, with green eyes, blonde natural hair color, medium build, projected height between 5′6″ and 5′9″ and strong predisposition for asthma." Even accepting the caveat that many of these characteristics will fluctuate and are influenced by environment in combination with genetics, it is easy to imagine the power of such information when investigating a crime. However, this area of forensic research is still in its infancy, awaiting rigorous research, and must be accompanied by careful consideration of the social and ethical concerns that are raised by such an approach. Notwithstanding, there exists no technological or scientific barrier to the development of this approach.

CONCLUSION

The genomics revolution, led by microarray technology, has already informed various subdisciplines in forensic science, most notably toxicogenomics and SNP analysis, and further microarray-catalyzed innovation is surely afoot. One of the hallmark features of the genomics era is the staggering scale and speed with which research and technology development is now progressing. Thus, the wait might not be as long as one would think. It can now be said that the revolutionary potential of microarray technology rivals that of PCR. However, the potential application to forensics is much broader than just DNA fingerprinting and human ID. It will be interesting to observe how quickly potential becomes reality and innovative research becomes the gold standard. A currently limiting factor, both in terms of research and in application, is the seemingly exorbitant cost of microarray printing, fluidics, and high-resolution scanners. However, these costs are steadily decreasing and the ability to strip and reuse microarrays would dramatically slash research costs (although it would not be appropriate for casework). Further, the need to purchase and maintain such microarray instrumentation, including array printers, fluidics stations, and even scanners, is being alleviated by a more "full service" approach, for both in-stock and custom arrays, from the major microarray companies. As declining costs bring microarray technology within reach for more research laboratories, increased numbers of thorough forensic validation studies can be performed and published. Hopefully, this will persuade forensic practitioners that the technology is not just powerful but *reliable* as a bona fide tool for forensic casework.

REFERENCES

Aardema, M. J., and MacGregor, J. T. (2002). Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies. *Mutat. Res.*, **499**:13–25.

CHAPTER 11

Date-Rape Drugs with Emphasis on GHB

STANLEY M. PARSONS

Department of Chemistry and Biochemistry, Program in Biomolecular Science and
Engineering, Neuroscience Research Institute, University of California,
Santa Barbara, California

Summary The date-rape drugs γ-hydroxybutyrate (GHB), 3,4-methylenedioxymethamphetamine, flunitrazepam, and ketamine are discussed with an emphasis on GHB. Recreational, predatory, and lethal doses, why polydosing is dangerous, metabolism including membrane transport, and diagnostic metabolites are covered. Similarities to and differences from the effects and metabolism of ethanol are also discussed. The advantages of field tests to detect date-rape drugs, and limitations of antibodies and advantages of enzymes for field tests, are described. Development of a rapid enzymatic test for GHB is described.

11.1 Introduction 357
11.2 Molecular mechanisms of action 357
  11.2.1 Receptors and transporters 357
  11.2.2 Real GHB receptors 359
11.3 Societal context of date-rape agents 361
  11.3.1 Acute effects of date-rape agents on cognition and behavior 361
  11.3.2 Medicinal uses of date-rape drugs 361
  11.3.3 Self-abuse 362
  11.3.4 Date rape, death, and regulation 363
11.4 Metabolism fundamentals 363
  11.4.1 Complexity in unraveling metabolism of GHB-related compounds 363
  11.4.2 Isozymes in GHB-related metabolism 364
  11.4.3 Subcellular compartmentalization of enzymes, transporters, and substrates 364
  11.4.4 Dynamics and equilibria for enzymes and transporters 365
  11.4.5 Thermodynamics-based analysis of metabolic flux 366

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

- 11.4.6 Metabolism of endogenous GHB versus ingested GHB and prodrugs 367
- 11.4.7 Directionality of in vivo and in vitro enzymatic activity 367
- 11.4.8 Transporters and enzymes mediating GHB-related metabolism 367

11.5 Biosynthesis of endogenous GHB 368
- 11.5.1 First step for GHB biosynthesis in the known pathway 368
- 11.5.2 Second step for GHB biosynthesis in the known pathway 368
- 11.5.3 Third step for GHB biosynthesis in the known pathway 371
- 11.5.4 Which step in GHB biosynthesis is rate limiting? 373
- 11.5.5 Are there other biosynthetic pathways to endogenous GHB? 374

11.6 Absorption and distribution of ingested GHB 376
- 11.6.1 Gastrointestinal tract 376
- 11.6.2 Blood 377

11.7 Initial catabolism of GHB 377
- 11.7.1 Transport into mitochondria 377
- 11.7.2 Iron-dependent alcohol dehydrogenase ADHFe1 377
- 11.7.3 Poorly characterized catabolism of GHB 379

11.8 Chemistry of GHB and related metabolites not requiring enzymes 380

11.9 Experimental equilibrium constants for redox reactions of GHB 380

11.10 Estimated equilibrium constants for redox reactions of GHB in vivo 381

11.11 Different perspectives on turnover of endogenous GHB are consistent 384

11.12 Disposition of succinic semialdehyde 385

11.13 Conversion of prodrugs to GHB and related metabolites 386
- 11.13.1 γ-Butyrolactone 386
- 11.13.2 1,4-Butanediol 387

11.14 Subcellular compartmentalization of GHB-related compounds 388

11.15 Comparative catabolism of ethanol, 1,4-butanediol, fatty acids, and GHB 389

11.16 Catabolism of MDMA, flunitrazepam, and ketamine 390

11.17 Detection of date-rape drugs 390
- 11.17.1 Compounds diagnostic for dosing by synthetic date-rape drugs 390
- 11.17.2 Compounds diagnostic for dosing by GHB 390
- 11.17.3 Gold-standard testing 391
- 11.17.4 Many applications for reliable field tests 392
- 11.17.5 Hospital emergency department example 392
- 11.17.6 Preparation of a sample for delayed analysis 393
- 11.17.7 Time window available to detect dosing 393
- 11.17.8 Extending the time window 394

11.18 Special circumstances of GHB 395
- 11.18.1 Industrial connection 395
- 11.18.2 Enzymes acting on GHB in bacteria, yeast, and plants 395
- 11.18.3 Possible accidental intoxication by GHB in the future 395

11.19 Considerations during development of field tests 396
- 11.19.1 Shortcomings of antibody-based screens for simple analytes 396
- 11.19.2 Advantages of enzyme-based screens for simple natural analytes 397

11.20 Development of an enzymatic test for GHB 399
- 11.20.1 Sensitivity required for the hospital emergency department 399

11.20.2 Choice of enzyme 399
11.20.3 Reliable field test for GHB 400
Conclusion 402
Notes 404
References 406

11.1 INTRODUCTION

Prominent examples of *date-rape drugs* are the natural compound γ-hydroxybutanoic acid (GHB) and the synthetic compounds 3,4-methylenedioxymethamphetamine (MDMA), 5-(2-fluorophenyl)-1,3-dihydro-1-methyl-7-nitro-2H-1,4-benzodiazepin-2-one (flunitrazepam), and 2-(2-chlorophenyl)-2-(methylamino)cyclohexanone (ketamine) (Figure 11.1; Scott-Ham and Burton, 2005; Saint-Martin et al., 2006; Abanades et al., 2007a; Juhascik et al., 2007). These compounds are grouped in ignominy because they can be administered surreptitiously in the relatively small amounts required to make a person compliant with sexual assault. Ethanol has similar effects, but generally it is not called a date-rape drug because it is legal and knowingly consumed. The term *agent* as used here refers to ethanol and date-rape drugs, whereas the term *drug* is restricted to date-rape drugs.

In this chapter we discuss primarily date-rape drugs.[1] We also touch on ethanol just enough to point out some similarities, differences, and interactions between ethanol and date-rape drugs. As GHB is a research topic of special interest to the author, it is emphasized throughout. Much important progress has been made recently in understanding the physiology, biochemistry, cell biology, and molecular biology relevant to these agents. Important progress has also been made in developing better methods to detect dosing. The goals of this chapter are to unify the biological and chemical perspectives of date-rape drugs and to improve our ability to detect dosing. The wide range of subjects covered has required the use of footnotes to explain specialized terminology. References are numerous, but to keep the number of them reasonable, they are not comprehensive (apologies to authors not cited). A number of recent excellent reviews of individual date-rape drugs and forensic methods to detect them may be consulted for more citations: Maitre et al., 2000; Palmer, 2004; Wong et al., 2004a,b; Britt and McCance-Katz, 2005; Gonzalez and Nutt, 2005; Drasbek et al., 2006; Pardi and Black, 2006; Snead and Gibson, 2006; Wedin et al., 2006; Wolff and Winstock, 2006; Colado et al., 2007; Kintz, 2007; Seamans, 2008; Sinner and Graf, 2008.

11.2 MOLECULAR MECHANISMS OF ACTION

11.2.1 Receptors and Transporters

To cause cognitive and behavioral effects, psychotropic agents must interact with neuroreceptors or neurotransporters in the central nervous system. This

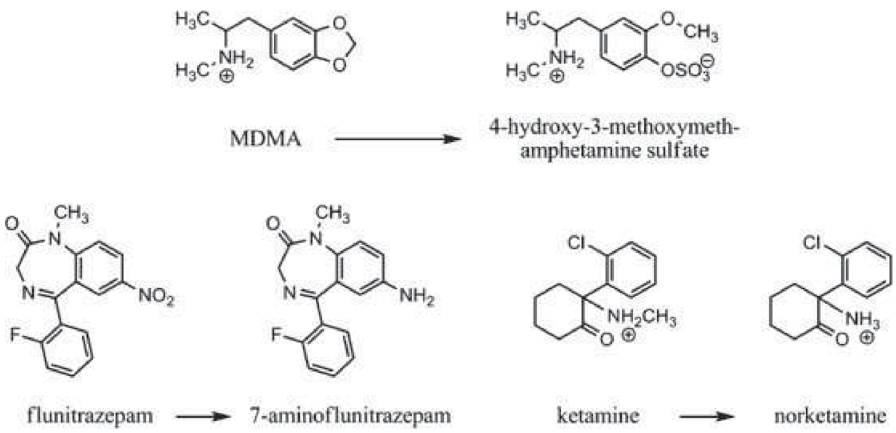

Figure 11.1 Structures of some synthetic date-rape drugs and their major metabolites found in human urine.

requirement means that the agents must pass through the blood–brain barrier to come into contact with neurons. A neuroreceptor generally spans the cytoplasmic membrane of a neuron, and the active site faces the extracellular space so that it can bind to a neurotransmitter coming from there. A neurotransporter can span either the cytoplasmic membrane or the membrane of a synaptic vesicle[2] inside a neuron.

The blood–brain barrier is breached by nonspecific diffusion for ethanol, flunitrazepam, and ketamine (Adachi et al., 2005; Cuadrado et al., 2007). It is breached by specific transport for MDMA, metabolites of MDMA, and GHB (Drewes et al., 2001; Bhattacharya and Boje, 2004; Escobedo et al., 2005). Thus, access by the latter compounds to brain potentially is saturable and inhibitable. More will be said later about transport versus nonspecific diffusion of GHB and related molecules. Some date-rape agents have actions in the periphery, but such actions won't be discussed much, except for metabolism.

Once they have reached neurons in the brain, the agents have many effects. Recent research indicates that ethanol acts primarily by means of a limited number of molecular mechanisms (Vengeliene et al., 2008). Prominent among them is a three-way, synergistic interaction of ethanol, neurosteroids[3], and the inhibitory neurotransmitter γ-aminobutanoic acid (GABA) on separate sites in the ionotropic $GABA_A$ receptor. Together these agents promote opening of the chloride channel in the receptor at concentrations of GABA lower than otherwise required (Herd et al., 2007; Mody et al., 2007; Mody, 2008).[4] Open chloride channels make most neurons resistant to electrical stimulation. This three-way synergy is especially effective at a subtype of the $GABA_A$ receptor located *outside* of synapses, where extracellular GABA concentrations are low but do not fluctuate much. As $GABA_A$ receptors are distributed throughout the brain, the result of ethanol consumption is widespread tonic (i.e., persistent) neural depression.

GHB is present naturally in brain at concentrations that depend on the region (Doherty et al., 1978; Vayer et al., 1988). Values range up to about 10 μM, averaged across tissue compartments. At the >100-fold-higher concentrations reached in sexual assault and self-abuse situations, GHB is a partial agonist for $GABA_B$, the metabotropic receptor (Gobaille et al., 1999; Lingenhoehl et al., 1999; Wong et al., 2004a,b).[5] Activation of $GABA_B$ receptor inhibits the release of neurotransmitters from nerve terminals by several intracellular second-messenger mechanisms (Ulrich and Bettler, 2007). GHB does not produce the full range of its behavioral effects in mice that have one of the two subunit types in $GABA_B$ receptors deleted from the genome (Bettler and Braeuner-Osborne, 2004; Gassmann and Bettler, 2007). The observation confirms involvement of $GABA_B$ receptors in GHB effects in vivo.

GHB increases neurosteroid levels three- to fivefold when it activates $GABA_B$ receptors (Barbaccia et al., 2002; Biggio et al., 2007). Increased neurosteroid levels in turn potentiate opening of chloride channels in $GABA_A$ receptors. Thus, high doses of GHB directly or indirectly activate both major classes of GABA receptor. These actions produce molecular, cellular, and behavioral responses partially overlapping those produced by ethanol and GABAergic drugs[6] such as flunitrazepam, another date-rape drug discussed in this chapter (see below; Baker et al., 2008; Helms et al., 2008; Carter et al., 2009).

Activation of GABA receptors by GHB is entangled with the possibility that a small fraction of ingested GHB or GHB prodrug is converted to GABA itself by reversal of the biosynthetic pathway from GABA to GHB (Figure 11.2; Sections 11.5.2 and 11.5.3) (Collier and De Feudis, 1970; Vayer et al., 1985a; Hechler et al., 1997; Gobaille et al., 1999). Such reversal is thermodynamically feasible. GABA derived from GHB potentially could act on all GABA receptors, including recently discovered $GABA_A$-ρ (also known as $GABA_C$; Schmidt, 2008). However, reversal is controversial (Mohler et al., 1976; Snead et al., 1989; Crunelli et al., 2006; Ren and Mody, 2006). It might not occur when membrane barriers are intact, and the issue is unresolved. Action of GHB on recently identified GHB receptors is described in Section 11.2.2.

Other date-rape drugs have different molecular mechanisms of action. Flunitrazepam binds to an allosteric site in both synaptic and nonsynaptic $GABA_A$ receptors, where it potentiates GABA action to depress the central nervous system (Brenneisen and Raymond, 2001). MDMA (and/or an MDMA metabolite) interacts with several types of transporter to cause release of the neurotransmitter 5-hydroxytryptamine (also known as serotonin) and the neurotransmitter-hormone oxytocin.[7] In addition, extensive use of MDMA can cause permanent neuronal damage, probably by promoting the formation of reactive oxygen species such as hydrogen peroxide and hydroxyl radical (de la Torre et al., 2004)! Ketamine is an antagonist at the NMDA type of receptor for the excitatory neurotransmitter glutamate (Raeder, 2003).[8]

11.2.2 Real GHB Receptors

The mechanisms described above do not involve a receptor selective for GHB. Two forms of the human GHB receptor (GHBh1 and C12K32) have been cloned,

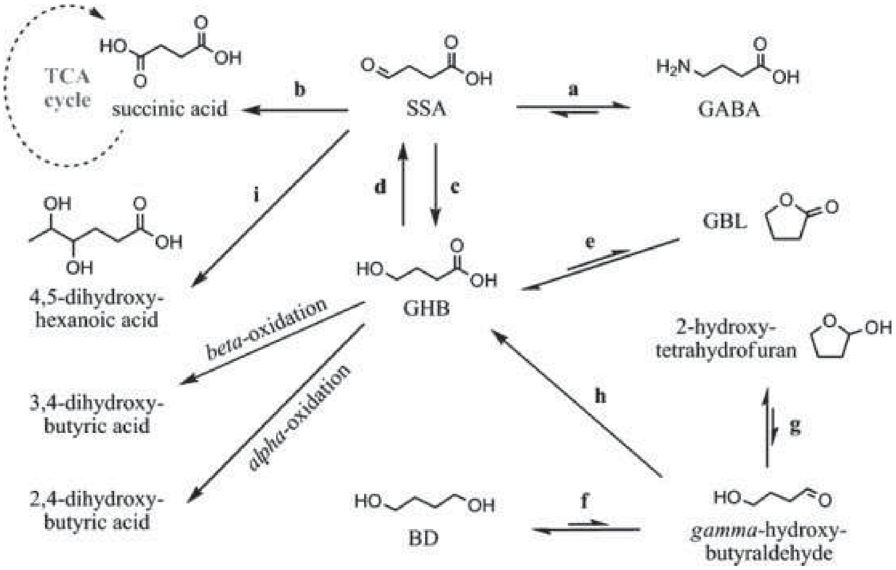
TCA cycle
succinic acid
b
SSA
a
GABA
i
d
c
4,5-dihydroxy-hexanoic acid
GBL
e
GHB
beta-oxidation
alpha-oxidation
2-hydroxy-tetrahydrofuran
3,4-dihydroxy-butyric acid
h
g
2,4-dihydroxy-butyric acid
BD
f
gamma-hydroxy-butyraldehyde

Figure 11.2 Known metabolism of GHB and its prodrugs. Chemical structures are shown above or next to the names or abbreviations. Reversible steps between metabolites are shown with paired half-head arrows and irreversible steps with a single arrow. An arrow longer than a paired arrow indicates the favored direction. Enzymes catalyzing the reactions alphabetically labeled are listed in Table 11.2. Reaction a: GABA transaminase makes SSA in mitochondrial matrix. Reaction b: ALDH5A1 in mitochondrial matrix oxidizes SSA to succinic acid that enters the tricarboxylic acid cycle. Reaction c: AKR1A1 and AKR7A2 reduce SSA in cytoplasm. This step requires transport of SSA from mitochondrial matrix (Figure 11.3). Reaction d: ADHFe1 oxidizes GHB in mitochondrial matrix after transport of GHB into matrix. The reaction probably is irreversible because reaction b is very fast. Untested is the possibility that irreversible ω-oxidation of GHB by cytochrome P450 and catalase also occurs (Section 11.15). Reaction e: γ-Lactonase 1 in blood and liver hydrolyzes GBL. Reaction f: One or more unidentified ADHs oxidize BD in the periphery. A different, irreversible oxidative mechanism that might involve cytochrome P450 and catalase occurs in brain (Section 11.15). Reaction g: Cyclization is spontaneous. Reaction h: Unidentified ALDH in the periphery. A different oxidative mechanism also might exist by analogy to the metabolism of ethanol (Section 11.15). Reaction i: An intermediate in the pyruvate decarboxylase reaction in mitochondria probably forms 5-keto-4-hydroxyhexanoic acid that then is reduced. α- and β-oxidation of GHB also occur, possibly in mitochondria, peroxisomes, and smooth endoplasmic reticulum (Section 11.15).

expressed, and characterized preliminarily (Kemmel et al., 2006; Andriamampandry et al., 2007). Both forms bind to physiologically relevant, sub-μM concentrations of GHB to regulate intracellular second messengers. The receptors also bind to the specific antagonist NCS-382, but not to GABA. GHBh1 expressed in a heterologous host quickly desensitizes after exposure to a saturating concentration of

GHB, whereas C12K32 does not.[9] A putative GHB receptor, which is of uncertain significance because it has aberrant pharmacological properties, has also been cloned from rat (Andriamampandry et al., 2003).

The cloning of the human receptors, combined with prior discoveries of neurotransmitter-like properties of GHB, such as its storage by synaptic vesicles, establishes GHB as a classical neurotransmitter (Kemmel et al., 1998; Muller et al., 2002). The full significance of neuro-signaling by endogenous GHB is not known (Ticku and Mehta, 2008). The terms *de novo* and *endogenous* are used here as synonyms. They mean not arising from ingested GHB or GHB prodrug. Whether GHB receptors in vivo desensitize during dosing with exogenous GHB, and what the physiological consequences of such desensitization would be, are not known. Characterization of these properties is just beginning (Coune et al., 2010).

11.3 SOCIETAL CONTEXT OF DATE-RAPE AGENTS

11.3.1 Acute Effects of Date-Rape Agents on Cognition and Behavior

Ethanol reduces social inhibition, increases suggestibility, and at large doses induces sedation and amnesia (Vengeliene et al., 2008). GHB increases libido and suggestibility, causes euphoria, relaxation, and reduction of social inhibitions, and at large doses induces sedation and amnesia (Drasbek et al., 2006; Kueh et al., 2008). Flunitrazepam induces hypnosis, sedation, anxiolysis, and at large doses, amnesia (Lane et al., 2008). MDMA causes mild euphoria and stimulates feelings of love, empathy, and connection to other people (Colado et al., 2007). Ketamine produces analgesia, anesthesia, and hallucinations (Rowland, 2005). Despite differences in the molecular mechanisms of action and required doses (Sections 11.2.1, 11.2.2, and 11.3.3), the agents are similar to each other in producing mental states that make victims compliant with sexual assault (Elliott, 2008; Lawyer et al., 2010).

11.3.2 Medicinal Uses of Date-Rape Drugs

Date-rape drugs have been used to treat clinical indications. For example, GHB is used for sleep disorders and alcohol-withdrawal syndrome (Agabio and Gessa, 2002; Johnston et al., 2003; Pardi and Black, 2006; Robinson and Keating, 2007; Castelli, 2008; Caputo et al., 2009). GHB might also be useful in cancer treatment and protection from or functional recovery after focal cerebral ischemia (Sadasivan et al., 2006; Emi et al., 2007; Gao et al., 2008). Flunitrazepam has been used for insomnia and anxiety (Matsuo and Morita, 2007), and ketamine has been used for pain (Okon, 2007; Morton, 2008). MDMA is being tested for treatment of posttraumatic stress disorder (Parrott, 2007; Vastag, 2010). Accordingly, the compounds are manufactured by reputable pharmaceutical companies and are available through prescription. Unfortunately, they also are available from illicit sources.

11.3.3 Self-Abuse

Date-rape agents frequently are ingested voluntarily without medical supervision (http://www.drugabuse.gov/Infofacts/clubdrugs.html). Some people engage in this behavior because they perceive the cognition- and behavior-altering properties of the agents to be compatible with recreation (Camacho et al., 2005; Griffiths and Johnson, 2005; Abanades et al., 2007a; Kim et al., 2007b; Horowitz et al., 2008; Lee and Levounis, 2008). Common oral doses for recreational use by people who have not developed tolerance are given in Table 11.1. Effective doses differ over a 50,000-fold range for the various agents!

In addition to having cognitive and behavioral effects, GHB stimulates release of human growth hormone (Volpi et al., 2000; Gonzalez and Nutt, 2005). Some bodybuilders consume it in huge amounts (18 to 250 g/day divided into multiple doses) in an effort to increase muscle mass. The practice results in tolerance and addiction, and severe withdrawal symptoms occur when addicted individuals attempt to discontinue use (Bennett et al., 2007; LeTourneau et al., 2008; Wojtowicz et al., 2008). In rats, the addiction decreases membrane fluidity and increases GABA reuptake from extracellular space in brain (Bhattacharya et al., 2006). GHB greatly alters the levels of some brain proteins (van Nieuwenhuijzen et al., 2010). It increases the acetylation of histone H3, thereby causing epigenetic modification of gene expression (Klein et al., 2009). The latter observations demonstrate physiological adaptation consistent with addiction.

TABLE 11.1 Effective Oral Doses for Human Beings[a]

Compound	LD_{50} (mg/kg)	Recreation (mg/kg)	$\frac{LD_{50}}{\text{Recreation}}$
Ethanol	5000[b]	500[c]	10
GHB	400	50	8
GBL/BD[d]	200	25	8
MDMA	30	2	16
Flunitrazepam	0.4	0.01	30[e]
Ketamine	40	1	>38

Source: Data from Miller (1979), Wang and Bai (1998), Gable (2004), Carai et al. (2006), and http://toxnet.nlm.nih.gov/. Effective doses for rodents are higher.

[a] The numbers apply to normal, healthy, and otherwise undrugged human adults of either gender who have not developed tolerance.

[b] This amount corresponds to about 0.6% ethanol in physiological fluids.

[c] This amount corresponds to consumption of about one-half bottle of wine.

[d] GBL and 1,4-butane diol (BD) are quickly converted to GHB in the body (Sections 11.13.1 and 11.13.2). Effective doses are lower than for GHB because the compounds are absorbed faster (Table 11.5). Safety factors are similar, however.

[e] Rounding errors create the apparent discrepancy relative to the ratio computed from numbers on the left.

11.3.4 Date Rape, Death, and Regulation

In chemical assault for the purpose of date rape, the perpetrator attempts to deliver a high recreational dose that incapacitates the victim. However, date-rape agents can be lethal. The estimated oral LD_{50} values for human beings who have not developed tolerance are given in Table 11.1. The ratio of the LD_{50} to the recreational dose (i.e., the safety factor) is also given for each agent. Some agents have small safety factors, particularly ethanol, GHB, and its prodrugs γ-butyrolactone (GBL) and 1,4-butanediol (BD) (Sections 11.13.1 and 11.13.2). Because different agents have mostly different molecular targets that can interact with each other physiologically, these agents have the potential to synergize and push the central nervous system into an extremely depressed state. Deaths occur at doses of combined agents that are difficult to predict and often much lower than those required for single agents (Jonasson et al., 2000; Lalonde and Wallage, 2004; Mazarr-Proo and Kerrigan, 2005; Jones et al., 2007; Verschraagen et al., 2007; Hovda et al., 2008; Knudsen et al., 2008; Akins et al., 2009). To impede drug-assisted date-rape and possible death, most national governments restrict access to date-rape drugs. Proscription has created a need for reliable analytical procedures to detect the compounds and their metabolites.

11.4 METABOLISM FUNDAMENTALS

11.4.1 Complexity in Unraveling Metabolism of GHB-Related Compounds

Because it is a neurotransmitter, GHB is degraded (catabolized) and biosynthesized (anabolized) at more-or-less constant low rates in mammals. Conceptually, such turnover can be divided into two phases, one starting with and one ending with GHB. In contrast, when GHB is ingested, catabolism starts with a very high concentration and proceeds at a high rate. The cascade of steps that ensues can be conceptualized as parallel to that followed in the catabolism of endogenous GHB. However, is it precisely the same or partially different when much higher concentrations are involved? When prodrug GBL or BD is ingested, rapid metabolic activation generates a higher concentration of GHB. Activation can be conceptualized as being parallel to the de novo biosynthesis of GHB, except that the steps are different. Overall, a full understanding of the actions of GHB and GHB-related compounds in vivo requires descriptions of at least four distinguishable phases of metabolism. The intellectual and experimental overlay of these phases onto each other creates much more complexity than exists for other date-rape drugs, each of which undergoes only one catabolic phase.

To provide the reader with the concepts needed to resolve uncertainties encountered in GHB metabolism, the rest of Section 11.4 covers some relevant fundamentals. In addition to its utility in defining metabolic pathways with certainty, information at the substantial level of detail presented in this chapter is important because it will probably reveal new ways to detect dosing and to mitigate the adverse effects of an overdose. Indeed, the potential payoff motivates the comprehensive approach taken.

11.4.2 Isozymes in GHB-Related Metabolism

Before the tools of recombinant DNA, cell biology, and genomic sequences were available, research on GHB metabolism was extraordinarily difficult. In part, this occurred because of the existence of *isozymes*, which catalyze the same type of reaction on the same type of chemical functional group in different substrates. For example, isozymic alcohol dehydrogenases catalyze the oxidation of alcohols (GHB is an alcohol) to carbonyl compounds. Isozymes are encoded by different genes.[10] Moreover, enzymes catalyze reversible reactions in both directions. In the case of biosynthesis and catabolism of GHB, it can be unclear in which direction a reversible reaction proceeds in vivo.

Isozymes often resemble each other in molecular weight, net charge, and other physicochemical properties due to evolutionary kinship. Isolation of an enzymatic activity from a natural source such as brain, kidney, or liver can produce a mixture of isozymes. Mixtures are especially likely when isozymes form heterooligomers with each other. Unfortunately, some in vitro data are compromised by lack of convincing isozymic identity and/or homogeneity. Many of the isozymes discussed in the chapter now have been expressed in and isolated from a heterologous host such as the bacterium *Escherichia coli*, which is methodology that produces a single isozyme. Where the issue is critical, we cite only references not suffering from isozyme ambiguity.

Genomic and mRNA sequences reveal that more forms of many enzymes exist than formerly realized, including the enzymes that catalyze reduction and oxidation (redox) reactions in GHB-related pathways. There are nine functional human genes for NAD^+-dependent alcohol dehydrogenases (ADHs), 14 for NADPH-dependent aldo-keto reductases (AKRs), and 23 for $NAD(P)^+$-dependent aldehyde dehydrogenases (ALDHs) (http://www.ihop-net.org/; http://www.med.upenn.edu/akr; Hyndman et al., 2003; Hoffmann and Valencia, 2004; Vasiliou and Nebert, 2005; Jin and Penning, 2007). ADHs generally bind NAD^+ in preference to $NADP^+$, and thus they catalyze oxidation of alcohols using primarily NAD^+ rather than $NADP^+$ in vivo. AKRs generally bind NADPH in preference to NADH, and thus they catalyze reduction of aldehydes and ketones primarily using NADPH rather than NADH in vivo. In contrast, many aldehyde dehydrogenases catalyze oxidation of aldehydes to carboxylic acids using either NAD^+ or $NADP^+$.[11]

11.4.3 Subcellular Compartmentalization of Enzymes, Transporters, and Substrates

An isozyme can be located in the same or different cellular and subcellular compartments as another isozyme of the same superfamily. It can have multiple natural substrates, some better than others. Also, a metabolite of interest might be a good substrate for several isozymes when characterized in vitro. When complexity in substrate–isozyme relationships exists in vitro, one cannot choose on the basis of substrate selectivity alone which isozyme functions in a particular pathway in vivo.[12] Rather, the choice can require knowledge of organ, cellular, and subcellular locations for both substrates and isozymes. Also, the levels of expression and cofactors, and the values for Michaelis–Menten parameters, can be important.[13]

If a pathway flows through a noncytosolic compartment such as the mitochondrion, a candidate isozyme and its substrate must be in that compartment. Delivery of a protein to the compartment requires an appropriate intracellular trafficking signal. If the signal is coded in the amino acid sequence, which it usually is, it is coded in the mRNA sequence also. Delivery of the substrate is spontaneous if it is hydrophobic enough to pass through membranes rapidly by nonspecific diffusion. However, if a substrate is highly polar or ionic (e.g., GHB; Section 11.14), delivery requires an appropriate transporter in the compartmental membrane. Isoporters also exist.[14]

11.4.4 Dynamics and Equilibria for Enzymes and Transporters

A transporter and an enzyme are similar to each other in that both are catalysts and both exhibit Michaelis–Menten constants for reaction in the forward and reverse directions. However, they can respond to substrate analogs very differently. For an enzyme, a binding analog of a substrate is always an inhibitor with respect to the substrate, as it blocks binding by substrate. For a transporter, membrane sidedness opens an additional possibility (Figure 11.3). Suppose that the rate at which the empty binding site for substrate reorients across the membrane is slower than the rate for the filled binding site. Such behavior is common. The steady-state rate for transport is limited by the rate of the slowest step, so it is limited in this case by reorientation of the empty binding site to begin a new transport cycle. Binding of an analog to the empty binding site facing the *trans* side of the membrane can speed the steady-state rate for transport of substrate by creating fast exchange between the substrate and the analog. Alternatively, if an analog binds but is not a substrate, it will slow the rate of substrate transport.

A proposed reaction in a metabolic pathway, whether catalyzed by an enzyme or a transporter, cannot have a very unfavorable equilibrium constant in order to function. The constant is important in part because of the Haldane relationship, which for a saturable, single-substrate single-product reaction is $K_{\mathrm{eq}} = V_{\mathrm{maxforward}} K_{\mathrm{Mproduct}} / V_{\mathrm{maxreverse}} K_{\mathrm{Msubstrate}}$ (Cleland, 1982). The symbols have the conventional meanings defined in Michaelis–Menten kinetics. The Haldane relationship is valid even for crude preparations of enzyme or transporter. When $K_{\mathrm{eq}} < 1, V_{\mathrm{maxforward}}$ or K_{Mproduct} (or both terms) is small relative to the corresponding term in the denominator. When $V_{\mathrm{maxforward}}$ is small, the reaction is intrinsically slow. When K_{Mproduct} is small (which means tight binding), product inhibition can prevent the enzyme or transporter from combining with more substrate. In either case, the reaction proceeds slowly, or not at all, in the direction written. This conclusion is important when analyzing incompletely characterized enzymes and transporters, as the K_{eq} value often can be estimated even when Michaelis–Menten values are unknown.

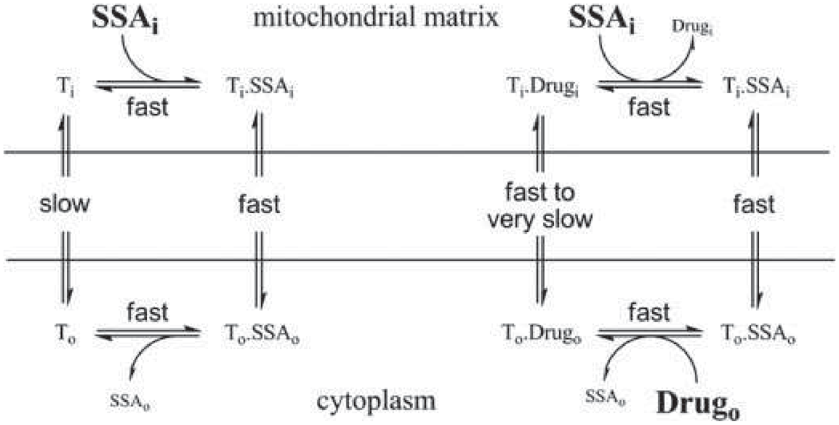

Figure 11.3 *Trans*-stimulation and inhibition of succinic semialdehyde efflux from mitochondria. The transporter for mitochondrial efflux of SSA is unknown, but it might be a member of the MCT family, which as a group has low substrate selectivity (Section 11.6.1). A hypothetical transport cycle starts with an empty binding site for substrate facing mitochondrial matrix (T_i in the cycle on the left) or a drug-loaded binding site facing mitochondrial matrix ($T_i \cdot Drug_i$ in cycle on right). The cycle on the left is rate limited by slow reorientation of an empty transporter to generate T_i. The cycle on the right runs faster than the cycle on the left if a drug molecule in cytoplasm ($Drug_o$) binds to T_o, and reorientation of the $T_o \cdot Drug_o$ complex and exchange to $T_i \cdot SSA_i$ are fast. The cycle would not run at all if a drug molecule binds and stops reorientation of the transporter. Both transport cycles are driven clockwise by net flow of substrates thermodynamically downhill.

11.4.5 Thermodynamics-Based Analysis of Metabolic Flux

But what minimal value for K_{eq} makes a reaction favorable enough for metabolism? Thermodynamics-based analysis of metabolic networks incorporating mass and flux balances provides insights (Beard and Qian, 2005; Henry et al., 2007). The concentrations of substrates and products common to many steps in metabolism, such as [H^+], [NAD^+], [$NADP^+$], [NADH], and [NADPH], are set by the network as a whole. Homeostasis often maintains them constant even in the face of large metabolic stress. If this condition applies, relevant concentrations can be substituted into the overall expression for K_{eq}. Substitution produces a modified K_{eq} value equal to the concentration of reaction-specific product divided by the concentration of reaction-specific substrate that would be present at equilibrium. If the modified K_{eq} is > 1, robust flux in the forward direction occurs because the reaction seeks equilibrium for which $[\text{product}]/[\text{reactant}] > 1$. If the modified K_{eq} value is < 1, flux in the reverse direction would occur *if* there is a source of reactant (which is the product of forward-direction flux) from the network. The principles apply whether reactions in vivo reach equilibrium or are in steady state.

A cell can often overcome a single modified $K_{eq} < 1$ in a multistep pathway in order to produce flux in the forward direction. It does so by adjusting the steady-state concentrations of the reaction-specific substrate and product to make the free-energy change at that step negative (think LeChâtelier's principle). However, when modified $K_{eq} \ll 1$, mass and flux constraints in the network preclude adjustments sufficient to obtain forward-direction flux (Feist et al., 2007). In this chapter, modified $K_{eq} \leq 10^{-4}$ is assumed to be too unfavorable to support normal in vivo metabolism in the direction written. This value is a low threshold requirement, as it is smaller than the modified K_{eq} values for all dehydrogenase and reductase reactions of the tricarboxylic acid (TCA) cycle and glycolysis. In a similar but inverse manner, modified $K_{eq} \geq 10^4$ will be taken as irreversible in vivo.

11.4.6 Metabolism of Endogenous GHB versus Ingested GHB and Prodrugs

Metabolism of endogenous GHB occurs within a metabolic network, and it is subject to the modified K_{eq} threshold and mass and flux balances outlined in Section 4.5. In contrast, because concentrations of ingested GHB (and its prodrugs) are not limited by anabolism, the catabolism of ingested GHB is not subject to the network restrictions. A large in vivo concentration of exogenous GHB might overcome high K_M and subthreshold values for modified K_{eq} at steps that endogenous GHB cannot overcome. Thus, catabolism of ingested GHB and prodrugs might follow multiple pathways (Sections 11.7.3 and 11.15), whereas catabolism of endogenous, low-concentration GHB probably follows mostly a single high-affinity pathway.

11.4.7 Directionality of in Vivo and in Vitro Enzymatic Activity

Reversible reactions producing no significant flux in the forward direction in vivo often can be made to run in that direction in vitro. This phenomenon occurs because constraints on [H^+], [oxidant], [reductant], and pathway-specific [reactant] and [product] are much less restrictive in vitro than they are in vivo. Typically, the initial velocity for an enzymatic reaction is characterized in nearly saturating concentrations of substrates, the absence of products, and optimal nonphysiological pH. Demonstration of an enzymatic reaction under such conditions does not mean that the reaction functions in vivo. On the other hand, barring an unknown activation requirement, an enzyme that does not catalyze a reaction of interest under otherwise optimized conditions in vitro is unlikely to function in vivo.

11.4.8 Transporters and Enzymes Mediating GHB-Related Metabolism

Transporters carrying GHB across plasma membranes of human beings for absorption by the body and metabolism by organs are listed in Table 11.2. Enzymes mediating GHB-related metabolism in human beings are listed in Table 11.3. Some of the listings are certain and some are speculative, as explained in the text. The table includes accession numbers for annotated amino acid sequences at the National Center for Biotechnology Information (NCBI). It also includes

TABLE 11.2 Some Monocarboxylate Transporters for GHB in Cytoplasmic Membranes[a]

Abbreviation	Synonym	NCBI Accession
MCT1[b]	SLC16A1	NP_003042
MCT2[b]	SLC16A7	O60669
MCT4[b]	SLC16A4	NP_004687
SMCT1[c]	SLC5A8[d]	NP_666018
SMCT2[c]	SLC5A12	Q1EHB4

[a] All of these transporters have low substrate affinity and selectivity.
[b] H^+-dependent monocarboxylate transporters 1, 2, and 4.
[c] Sodium-coupled monocarboxylate transporters 1 and 2.
[d] Also known as the sodium iodide–related cotransporter.

identification numbers for three-dimensional atomic-resolution structures in the Protein Data Bank (PDB) when they are known (http://www.ncbi.nlm.nih.gov/; http://www.rcsb.org/pdb/home/home.do). Inspection of Tables 11.2 and 11.3 confirms that impressive progress in knowledge of the molecular and cellular biology of GHB-related pathways has been made, but much more remains to be done. The structures of known and probable chemical intermediates in the pathways are shown in Figure 11.2.

11.5 BIOSYNTHESIS OF ENDOGENOUS GHB

11.5.1 First Step for GHB Biosynthesis in the Known Pathway

A well-documented de novo pathway for the biosynthesis of GHB in brain originates primarily with glutamate. The pathway is shown in Figure 11.2. Glutamate decarboxylases 1 and 2 yield GABA in mitochondrial matrix. The enzymes are isozymes of each other, and the relative frequency of occurrence of cDNA clones reveals that the mRNAs are highly to well expressed in all human tissues (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/). Minor amounts of GABA also arise by deamination of the polyamines spermine, ornithine, and putresine (Tillakaratne et al., 1995).

11.5.2 Second Step for GHB Biosynthesis in the Known Pathway

GABA transaminase reacts GABA with 2-oxoglutarate to produce succinic semialdehyde (SSA) plus glutamate in mitochondrial matrix (Figure 11.2). The reaction transfers the pro-(S)-hydrogen from GABA (Santaniello et al., 1978). Only one gene exists for human GABA transaminase. The mature polypeptide is 472 amino acids long, M_r 53.3 kDa, and a soluble homodimer (Schousboe et al., 1980; Kugler, 1993; Rao and Murthy, 1993). The relative frequency of cDNA clones has revealed that the mRNA is expressed at high to moderate levels in essentially all human tissues (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/).[15] GABA

TABLE 11.3 Enzymes Mediating GHB-Related Metabolism[a]

Common Name	Isozyme	Probable Compartment[b]	Substrate	Product	K_M (mM)	NCBI Accession	Isoform[c]	PDB
GABA transaminase	Only one	Mitochondria	GABA	Succinic semialdehyde (SSA)	0.4	P80404	—	1OHW[d]
Low K_M SSA reductase[e]	AKR7A2	Cytoplasm[f]	SSA	GHB	0.020	O43488	AAH07352, AAH10852	2BP1
Aldehyde reductase	AKR1A1	Cytoplasm	SSA	GHB	0.231	P14550	—	2ALR
2-Oxoglutarate-dependent GHB transhydrogenase[g]	ADHFe1	Mitochondria	GHB	SSA	0.07	NP_653251		
SSA dehydrogenase	ALDH5A1	Mitochondria	SSA	Succinate	0.006	P51649	NP_733936, NP_001071	2 W8N, 2 W8O[h]
α-Oxidation (cytochrome P450)	–	Endoplasmic reticulum/ peroxisomes	GHB	2,4-Dihydroxy-butanoate	1.4 ?	–	–	–
β-Oxidation	Enzymes of fatty acid oxidation	Endoplasmic reticulum/ peroxisomes	GHB	3,4-Dihydroxy-butanoate	–	–	–	–
ω-Oxidation (cytochrome P450)	–	Endoplasmic reticulum	GHB	SSA	1.4 ?	–	–	–
ω-Oxidation (catalase)	Only one	Peroxisomes	GHB	SSA	–	P04040	—	1DGB
γ-Lactonase[i]	1	Blood and liver	γ-Butyrolactone (GBL)	GHB	10	NP_000437	P27169	1V04[j]
β-Oxidation (cytochrome P450)	–	Endoplasmic reticulum	GBL	3-Hydroxy-GBL	–	–	–	–

(*continued*)

TABLE 11.3 (*Continued*)

Common Name	Isozyme	Probable Compartment[b]	Substrate	Product	K_M (mM)	NCBI Accession	Isoform[c]	PDB
γ-oxidation (cytochrome P450)	–	Endoplasmic reticulum	GBL	4-Hydroxy-GBL → SSA	–	–	–	–
Unidentified NAD$^+$-dependent ADH	–	–	BD	γ-Hydroxy-butyraldehyde[k]	–	–	–	1HSZ, 1U3V, 1U3U[l]
Cytochrome P450	2E1	Endoplasmic reticulum	BD	γ-Hydroxy-butyraldehyde[k]	–	–	–	–
Unidentified ALDH	–	–	γ-Hydroxy-butyraldehyde[m]	GHB	–	–	–	–
NAD$^+$-dependent GHB dehydrogenase	Only one	Bacterial cytoplasm	GHB	SSA	2	AAT18823		

---

[a] The enzymes correlate with some of the pathways in Figure 11.2.

[b] In human beings, except for bacterial NAD$^+$-dependent GHB dehydrogenase.

[c] NCBI accession for mRNA sequences of alternatively spliced variants. Other polymorphisms are not considered.

[d] Pig enzyme.

[e] Also known as aflatoxin B1 aldehyde reductase.

[f] A few reports support mitochondrial and Golgi localization.

[g] Also known as hydroxyacid-oxoacid transhydrogenase (HOT) and ADH8. However, the term ADH8 is also used for retinol dehydrogenase. ADHFe1 uses added 2-oxoglutarate instead of added NAD$^+$ as the oxidant. No operative NAD$^+$-dependent GHB dehydrogenase has been identified in human beings.

[h] Oxidized and reduced forms.

[i] The enzyme is also known as paraoxonase.

[j] Chimeric human, rabbit, mouse, and rat γ-lactonase enzyme (NCBI accession AAR95986).

[k] Presumed but based on indirect evidence.

[l] These are structures of ADH1B, which is only very weakly implicated.

[m] Assumed reactant on basis of indirect evidence.

and 2-oxoglutarate (a TCA cycle intermediate) are present in essentially all organs (Tanaka, 1985; Zambonin et al., 1991; Abe et al., 1998). Thus, SSA is probably produced in all organs, although this speculation has not been tested. Because SSA is the immediate precursor to GHB (below), it would be of interest to determine SSA levels in all organs of the mammalian body now that a quantitative analytical method has been developed (Struys et al., 2005a).

The overall equilibrium constant from GABA to SSA is 0.04 (Van Bemmelen et al., 1985). As [2-oxoglutarate]/[glutamate] in mitochondrial matrix is ∼15, the modified K_{eq} value is ∼0.6 (Wiesner et al., 1988). Although modified $K_{\mathrm{eq}} < 1$, the reaction can readily be made to run robustly in the forward direction by lowering the SSA concentration. Lowering indeed occurs by irreversible and rapid conversion of SSA to succinate by SSA dehydrogenase (isozyme ALDH5A1, Section 11.12).

Other than from catabolism of ingested GHB or GHB prodrug, the only significant source of SSA known in mammals is GABA by means of the transaminase reaction. A very minor source is catabolism of vitamin B_6 (http://www.genome.ad.jp/kegg/pathway.html). Another theoretically possible source is 2-oxoglutarate, which in turn is derived from many metabolites. The enzyme 2-oxoglutarate decarboxylase is part of the 2-oxoglutarate dehydrogenase complex in the TCA cycle. It makes SSA bind covalently to thiamine pyrophosphate, which then transfers the SSA to dihydrolipoyl acyltransferase, also in the dehydrogenase complex. The transfer reaction is a form of substrate channeling in which a newly made product molecule is passed directly into the active site of another enzyme without dissociation of free product into solution. There has been no report of the SSA intermediate escaping from the 2-oxoglutarate dehydrogenase complex, although the possibility appears not to have been tested at the level of accuracy required to detect the small amount of free SSA used for the production of endogenous GHB (Section 11.5.3). Human beings and other mammals contain a gene and mRNA coding for 2-oxoglutarate decarboxylase-like hypothetical mitochondrial protein (i.e., OGDHL) of uncertain function that could possibly release SSA (Bunik and Degtyarev, 2008). Formation of free SSA by 2-oxoglutarate decarboxylase not part of a 2-oxoglutarate dehydrogenase complex is a robust reaction in some microbial species.

11.5.3 Third Step for GHB Biosynthesis in the Known Pathway

In brain, a very small fraction (∼0.1%) of the SSA made from GABA is reduced to GHB (Gold and Roth, 1977; Rumigny et al., 1981). The remaining 99.9% is oxidized to succinate (Section 11.12). The only enzymes known to be capable of reducing aldehydes to alcohols in mammals are AKRs acting in their forward (reductive) direction and ADHs acting in their reverse (reductive) direction (http://www.genome.ad.jp/kegg/pathway.html). Among human AKRs tested, only AKR1A1 and AKR7A2 reduce SSA at a significant rate in vitro (Cash et al., 1979). Among tested mammalian ADHs, only ADHFe1 reduces SSA at a significant rate in vitro (Section 11.7.2). Reduction of SSA by an enzyme that prefers NADH over NADPH has not been demonstrated to exist in mammals. Nevertheless, it can be inferred potentially to exist from a single observation of apparent NAD^+-dependent oxidation of GHB by a mammalian extract (Lyon et al., 2007). For reasons discussed in Sections 11.7.2, 11.7.3, and 11.10, however, only AKR1A1 and AKR7A2 currently are realistic candidates for GHB biosynthesis.

The AKR1A1 polypeptide is 325 amino acids long, M_r 36.6 kDa, and a soluble monomer. The atomic-resolution structure has been determined (El-Kabbani et al., 1994). The AKR7A2 polypeptide is 359 amino acids long, M_r39.6 kDa, and a soluble dimer of 79 kDa. It also is known as aflatoxin B1 aldehyde reductase (AFAR; Ireland et al., 1998; Schaller et al., 1999).[16] The atomic-resolution structures of AKR7A2 and the mouse ortholog (Akr7a5) have been determined (Debreczeni et al., 2005; Zhu et al., 2006). AKR1A1 has a k_{cat} value for SSA fivefold higher than that of AKR7A2, but AKR7A2 has a K_{M} value for SSA 11-fold lower (20 μM) than that of AKR1A1 (230 μM) (O'Connor et al., 1999).[17] Thus, the catalytic efficacies ($k_{\text{cat}}/K_{\text{M}}$) are very similar to each other. However, AKR7A2 is twofold faster at low concentrations of SSA, and AKR1A1 is fivefold faster at high concentrations. The analysis does not take into account different levels of expression of the enzymes in vivo.

Which of the isozymes is likely to dominate GHB biosynthesis? To begin the analysis, consider the distribution of endogenous GHB. It is found in brain, heart, kidney, liver, lung, skeletal muscle, and brown fat of human beings and rodents (Nelson et al., 1981; Maitre, 1997; Moriya and Hashimoto, 2005; Richard et al., 2005; Zacharis et al., 2008). GHB is not restricted to the nervous system, but its nonnervous role is not understood. In brain, GHB is located primarily in cytosol and synaptic vesicles, as indicated by subcellular fractionation of homogenates (Snead, 1987; Muller et al., 2002). Brown fat contains an especially high concentration (37 nmol/g), which suggests the possibility that GHB is catabolized slowly in this tissue. A question regarding detection of dosing presents itself. Does the brown fat of dosed human adults take up ingested GHB (or converted prodrug) and retain it longer than other tissues do? If so, is biopsy of brown fat an acceptable forensics technique for living human beings (van Marken Lichtenbelt et al., 2009)?

Only trace levels of endogenous GHB occur in blood and cerebrospinal fluid. The very low levels mean that transfer of GHB from organs that make it to organs that do not is unlikely. Such transfer would require high-affinity ($K_{\text{M}} \leq 1$ μM) uptake of GHB with high selectivity to prevent competitive inhibition of uptake by other common monocarboxylic acids present in blood and cerebrospinal fluid. However, a transporter having these properties has not been described (Section 11.6.1). Tissues that contain endogenous GHB thus make it. They must contain both the biosynthetic enzymes and the initial, pathway-specific substrate, which in the only well-documented de novo pathway to GHB is SSA generated from mitochondrial GABA. However, if another significant pathway to GHB exists in human beings, it is possible that an intermediate of that pathway could be passed from one organ to another through blood or cerebrospinal fluid (below).

Immunostaining for human AKR1A1 and AKR7A2 proteins and their rodent orthologs indicates that AKR7A2 and its orthologs are present in essentially all tissues but that AKR1A1 and its orthologs are not (Ireland et al., 1998; O'Connor et al., 1999; Kelly et al., 2000; Grant et al., 2001; Hedberg et al., 2001; Picklo et al., 2001a,b; Hinshelwood et al., 2002, 2003; Fung et al., 2004; Cui et al., 2009).[18] Assays for mRNA and relative frequencies for cDNA clones coding for AKR1A1 and AKR7A2 reveal that both mRNAs are present at very high to significant levels in essentially all human tissues (http://genome.ucsc.edu/; http://bioinfo2.weizmann.ac.il/cgi-bin/genenote/home_page.pl; http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/).[19] Why do the cellular distribution maps for mRNA and protein corresponding to AKR1A1 disagree? It is possible that immunostaining of AKR1A1 in some tissues does not occur because the major epitopes are masked (Shi et al., 1995).[20] On the other hand, comparisons between the transcriptome and the proteome of human beings often exhibit poor correlations. The disjunction suggests that control mechanisms responding to different physiological imperatives operate at the transcriptional and proteomic levels of expression (Rogers et al., 2008). More research would be required to determine whether substantially more AKR7A2 than AKR1A1 truly is present in vivo. This situation presumably would mean AKR7A2 rather than AKR1A1 dominates GHB biosynthesis, as in fact has been concluded by some researchers (Rumigny et al., 1981). The results reported in Section 11.5.4 support this possibility. They also indicate, however, that it probably does not matter which of these AKRs dominates in different cell types.

11.5.4 Which Step in GHB Biosynthesis is Rate Limiting?

In cultured human SH-SY5Y neuroblastoma cells, suppression of AKR7A2 expression with siRNA causes about a 90% decrease in AKR7A2 protein and SSA reductase activity, but only a 35% decrease in the amount of GHB (Lyon et al., 2007).[21] The compound zopolrestat, a selective inhibitor of AKR1A1 activity, blocks only about one-half of the 10% residual SSA reductase activity in an extract from suppressed cells, even when the SSA concentration is saturating for AKR1A1. Apparently, SH-SY5Y cells express only a small amount of AKR1A1 ($\leq$5% of the total SSA reductase activity), although they make substantial amounts of GHB whether or not AKR7A2 is fully expressed.[22] The results are consistent with rate-limiting AKR7A2 activity in normal GHB biosynthesis and induction of an unknown AKR1A1-independent pathway when AKR7A2 is suppressed.

However, there is a simpler hypothesis. The results are *also* consistent with neither AKR7A2 nor AKR1A1 activity being rate limiting to GHB biosynthesis. Instead, the rate-limiting step could be presentation of SSA for reduction. In this scenario, about 10% residual SSA reductase activity in SH-SY5Y cells could be sufficient to support 65% of the normal rate of GHB biosynthesis.

To assess the feasibility of this model, subcellular compartmentalization of SSA reduction must be analyzed. AKR7A2 and AKR1A1, and their rodent orthologs, have been shown to be in cytosol or adsorbed to cytosolic surfaces of nuclear, Golgi, and mitochondrial membranes (Claros and Vincens, 1996; Andriamampandry et al., 1998; Witzmann et al., 1998; O'Connor et al., 1999; Picklo et al., 2001a; Kelly et al., 2002, 2004; Keenan et al., 2006; Lyon et al., 2007).[23,24] Moreover, reduction by NADPH is nearly 1000-fold more favorable in cytosol than in mitochondrial matrix (Section 11.10). Both factors argue that SSA is reduced in cytosol. This conclusion requires that SSA made in mitochondria be exported to cytosol before it is reduced to GHB.

Presentation of SSA to AKR1A1 and AKR7A2 could be a complex process. It might require three steps, and none of them has been well characterized.

1. Assuming that the GABA transaminase reaction is the only significant source of SSA (Section 11.5.2), a possible rate-limiting step is escape of newly made SSA from substrate channeling in a complex formed between GABA transaminase and ALDH5A1, as discussed in Section 11.12. Such potential channeling should not be confused with the known channeling that occurs in the glutamate dehydrogenase complex (Section 11.5.2).
2. A definitely required and possibly rate-limiting step is transport of SSA out of mitochondrial matrix. The transport might be performed by the same transporter that takes exogenous GHB into mitochondria for catabolism (Section 11.7.1). Alternatively, it might be performed by an unknown transporter.
3. Another possible rate-limiting step is dehydration of SSA hydrate. The fraction of SSA that hydrates spontaneously to form *gem*-diol during steady-state metabolism in vivo is not known. It could be as much as one-half (Section 11.8). Hydrated and nonhydrated forms of SSA probably are transported out of mitochondrial matrix at different rates, and such discrimination could influence the fraction of the SSA in cytoplasm that is unhydrated and available for reduction.

Overall, because the hypothesis is parsimonious, it seems likely that the rate-limiting step in biosynthesis of endogenous GHB is presentation of reducible SSA to AKR1A1 and AKR7A2 in cytoplasm. However, the slow molecular process within the presentation requirement is unidentified.

11.5.5 Are There Other Biosynthetic Pathways to Endogenous GHB?

Some experimental observations have been interpreted to support the existence of a GABA-independent pathway to endogenous GHB:

1. Substantial amounts of endogenous GHB are found in organs containing $\leq 1\%$ as much GABA as brain has (Drummond and Phillips, 1974; Nelson et al., 1981). The implication is that such organs do not contain enough GABA to explain their GHB contents.
2. Anticonvulsants and inhibitors of glutamate decarboxylase and GABA transaminase affect levels of GABA and GHB in rat brain paradoxically (Snead et al., 1982).
3. Exogenous BD is converted quickly to GHB (Section 11.13.2) by a pathway that does not involve GABA (Snead et al., 1989). Free and chemically bound

forms of BD have been reported to be present in undosed human beings and rats at about one-tenth the amount of GHB, as required if BD were part of a biosynthetic pathway to GHB (Barker et al., 1985; Poldrugo et al., 1985).

4. NAD^+-dependent ADH activity that is able to oxidize exogenous BD increases in SH-SY5Y cells treated with siRNA suppressing AKR7A2, as if ADH activity were up-regulated by a feedback mechanism sensing levels of GHB or AKR7A2 (above).

Although many of them have nagged the field for years, the observations and their interpretations are ambiguous or uncertain. Observation 1 is not compelling because GHB biosynthesis in the "reference" organ brain is very curious. Only 1 part in 1000 of GABA-derived SSA is transported to cytosol to become GHB (Sections 11.5.3 and 11.12). Low levels of GABA characteristic of peripheral organs could support the formation of relatively large amounts of GHB if (1) the presentation of reducible SSA to AKR1A1 and AKR7A2 in cytoplasm is rate limiting, and (2) a larger fraction of mitochondrial SSA is presented for reduction in those organs than in brain. Even a 100-fold increase would decrease mitochondrial flux of SSA to succinate only 10%!

The results of an experiment monitoring metabolism of [2,3,3,4,4−$^2H_5$]glutamic acid to SSA and GHB are informative (Niwa et al., 1983). A high concentration of unlabeled GHB and about one-third additional [$^2H_4$]GHB are present in spent culture medium after incubation of primary kidney cells in a minimum essential medium containing 2 mM DL-[$^2H_5$]glutamic acid. GHB is barely detectable after similar incubation of primary liver cells. Instead, high concentrations of SSA are found. The observation is consistent with the approximately 17-fold greater amount of GHB in intact kidney (22 to 28 nmol/g) compared to intact liver (1.4 to 1.6 nmol/g), even though these organs contain similar amounts of GABA (21 and 15 nmol/g, respectively) (Drummond and Phillips, 1974; Nelson et al., 1981). As there are many metabolic sources of glutamic acid, the presence of excess deuterium atoms in about one-fourth of the GHB indicates that most GHB secreted by kidney cells comes from glutamic acid. Different partitioning ratios at metabolic branch points in the conversion of glutamate to GHB probably account for different GHB/GABA ratios in different organs.

The experiment also observed small amounts of 4-hydroxy-2-butenoic acid in both spent media. The compound does not arise from GHB or glutamic acid, as it did not contain excess deuterium. Brain and kidney contain the compound at levels one-sixth to one-half of the amounts of GHB (Niwa et al., 1982; Vayer et al., 1985b). Possibly, 4-hydroxy-2-butenoic acid can be reduced in vivo to GHB, but the conversion has not been demonstrated. This possibility should be tested.

Observation 2 used drugs and inhibitors thought at the time to act primarily on ion channels and enzymes. However, the analysis did not consider the role of the mitochondrial exporter for SSA. Aminooxyacetic acid, valproic acid, and 3-mercaptopropionic acid are three of the compounds utilized in the study, and they are monocarboxylic acids. As discussed in Section 11.4.3 and illustrated in Figure 11.3, depending on how they interact with the unknown exporter for SSA, monocarboxylic acids could either stimulate or inhibit the rate of SSA export and thus GHB biosynthesis. Unfortunately, there is no information available on this issue for the compounds used. All of the “inhibitors” utilized are now known to be much less selective than once thought (Meldrum and Rogawski, 2007), and their effects on the known pathway for GHB biosynthesis cannot be predicted.

Regarding observation 3, recent mass spectrometric analyses have not found BD in human beings, rats, or mice (Sakurada et al., 2002; Richard et al., 2005; http://www.metabolibrary.ca/). Extracts of plants contain minor amounts of long-chain diesters of BD, so trace BD in mammals might arise from diet, which could change through time (Bergelson, 1969). However, a trace dietary source is unlikely to support a robust alternative pathway to GHB, and such a source certainly cannot do so in cultured SH-SY5Y cells. A report that GBL is present in the brains of undosed rats at about one-tenth the amount of GHB (Snead et al., 1989) also has not been confirmed by more recent literature.

Finally, observation 4 is not compelling either, as no demonstration has been reported that the ADH activity induced in SH-SY5Y cells is selective for BD. Overall, available data do not make a persuasive case that a second pathway for biosynthesis of endogenous GHB exists in human beings.

11.6 ABSORPTION AND DISTRIBUTION OF INGESTED GHB

11.6.1 Gastrointestinal Tract

The first signs of intoxication by GHB can appear in as little as 10 min, which suggests that the stomach absorbs some GHB (Tanaka et al., 2003). Strong behavioral disturbances and peak concentrations of GHB in blood occur between 0.5 and 2 h, which is $\geq$ the time required for gastric emptying (Poggioli et al., 1999; Carter et al., 2006b; Abanades et al., 2007b; Goodwin et al., 2009). Thus, the small intestine absorbs most of each ingested dose of GHB.

Intestinal absorption probably is mediated by H^+-dependent monocarboxylate transporters (MCTs) and sodium-coupled monocarboxylate transporters (SMCTs) located in apical and basolateral membranes of epithelial cells in the stomach and small intestine (Table 11.2; Goddard, 1998; Iwanaga et al., 2006; Ganapathy et al., 2008). SMCTs and MCTs transport a wide range of metabolites involved in energy metabolism. Substrates include lactic acid, pyruvic acid, β-ᴅ-hydroxybutanoic acid, acetoacetic acid, and butanoic acid. The transporters are found throughout the human body, including in the blood–brain barrier and nerve terminals (Benavides et al., 1982; Halestrap and Price, 1999; Tunnicliff, 2002; Enerson and Drewes, 2003; Nishimura and Naito, 2005; Pierre and Pellerin, 2005; Simpson et al., 2007; Wang and Morris, 2007; Morris and Felmlee, 2008; Cui and Morris, 2009). SMCTs and MCTs exhibit moderate to low affinity and low selectivity for substrates. The K_M values for GHB are 30 to 700 μM for different SMCTs and 1 to 10 mM for different MCTs.

11.6.2 Blood

Liver, kidney, lung, heart, brain, brown fat, white fat, skeletal muscle, and pancreas catabolize exogenous GHB (Kleinzeller, 1944; Roth and Giarman, 1966; Khizhnyak, 1976; Lettieri and Fung, 1976; Shumate and Snead, 1979; Struys et al., 2006b; Lenz et al., 2008). Because they express different amounts and mixes of SMCTs and MCTs, different organs probably dominate uptake (and catabolism) of different doses of GHB. However, when fatal dosing occurs with GBL, the postmortem amounts of GHB per gram of tissue are within a factor of 2 of each other for all organs (Lenz et al., 2008).

11.7 INITIAL CATABOLISM OF GHB

11.7.1 Transport into Mitochondria

After GHB is taken into cells, where is it catabolized? Mitochondria isolated from rat liver and kidney convert GHB to SSA, and the rate increases when the mitochondria are disrupted (Gibson and Nyhan, 1989). The observations suggest that (1) metabolizing enzyme(s) is(are) in mitochondrial matrix, and (2) *transport* of GHB into mitochondrial matrix *is capacity limited*. The latter term is synonymous with the existence of rate limitation at high doses of GHB. Mitochondrial uptake probably is mediated by an isoform of the MCT family (Gladden, 2007; Yoshida et al., 2007; Hashimoto and Brooks, 2008; Hashimoto et al., 2008). Comprehensive characterization of the roles that intracellular transporters have in GHB catabolism would be welcomed.

11.7.2 Iron-Dependent Alcohol Dehydrogenase ADHFe1

An atypical alcohol dehydrogenase called ADHFe1 initiates one pathway for mitochondrial catabolism of GHB (Kaufman et al., 1988).[25] Antibody against recombinantly tagged ADHFe1 heterologously expressed in a mammalian cell line stains mitochondria intensely (Kim et al., 2007a).[26] mRNA for ADHFe1 is expressed at very high or significant levels in essentially all human and mouse tissues and in SH-SY5Y cells (Deng et al., 2002; Kardon et al., 2006; Kim et al., 2007a; Lyon et al., 2009; http://www.genecards.org/cgi-bin/carddisp.pl?gene=ADHFE1&search=adhfe1; http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/). Four alternatively spliced mRNAs encoding different putative isoforms of the enzyme (M_r = 50.3, 45.1, 32.5, and 27.4 kDa) have been detected (Deng et al., 2002; http://www.uniprot.org/uniprot/Q8IWW8#Q8IWW8-2). Two of the isoforms encode an N-terminal targeting sequence for import by mitochondria, and two do not, suggesting that some ADHFe1 is in matrix and some in cytosol. The two-compartment hypothesis has not been tested carefully, however, and it could have important consequences for the location of GHB catabolism. Because the dehydrogenase activity characterized requires unique oxidants (not a nicotinamide, *o*-quinone, or flavin cofactor; see below), it can be detected unambiguously. It has been found in homogenates of liver, kidney, heart, brain, skeletal muscle, and lung of rats.

The active site of ADHFe1 probably contains tightly bound iron in the +2 oxidation state [i.e., Fe(II) or Fe^{2+}] (Gibson and Nyhan, 1989; Kaufman and Nelson, 1991; Struys et al., 2004; 2006a; Kardon et al. 2006). However, this statement is an inference based on sequence homology with iron-dependent alcohol dehydrogenases isolated from microorganisms. No chemical analysis of mammalian ADHFe1 for metal ions has been reported, as the enzyme has not been purified to homogeneity. The enzyme prefers to use 2-oxoglutarate as the oxidant, from which stoichiometric D-2-hydroxyglutarate [also known as (*R*)-2-hydroxyglutarate] results. The activity does not require added nicotinamide cofactor. Nevertheless, the sequence contains the NAD(P)-binding motif G-X-G-X-X-G, where G is glycine and X is any amino acid (Wierenga et al., 1985). Presumably, a nicotinamide cofactor is very tightly bound, but again this statement is an inference. Such a cofactor would have to cycle between its oxidized and reduced states as the enzyme binds to and reacts with GHB or 2-oxoglutarate in alternating fashion (Figure 11.4). This kinetic mechanism couples the thermodynamics of GHB oxidation to free [2-oxoglutarate]/[D-2-hydroxyglutarate] and not free $[NAD(P)^+]/[NAD(P)H]$.[27]

ADHFe1 has interesting substrate selectivity, as GHB, D-2-hydroxyglutarate, L-3-hydroxybutanoic acid, and L-lactic acid are substrates for the oxidative reaction. SSA, 2-oxoglutaric acid, 2-oxoadipic acid, oxaloacetic acid, pyruvic acid, and 2-oxobutanoic acid are substrates for the reductive reaction. The substrates are listed here in order of decreasing preference (Kaufman et al., 1988; Struys et al., 2005b). Although the reaction couple between GHB and SSA presumably proceeds best in vitro, it produces no net chemical change. The next best reaction couple is GHB/2-oxoglutarate, which is why we assume here that ADHFe1 in vivo uses 2-oxoglutarate preferentially to oxidize GHB. The reverse reaction between the SSA/D-2-hydroxyglutarate couple can be made to proceed in vitro. However, it is very unlikely to occur in vivo, as the modified equilibrium constant probably is $\ll 1$ (Section 11.10) and the concentration of SSA is low (Section 11.12).

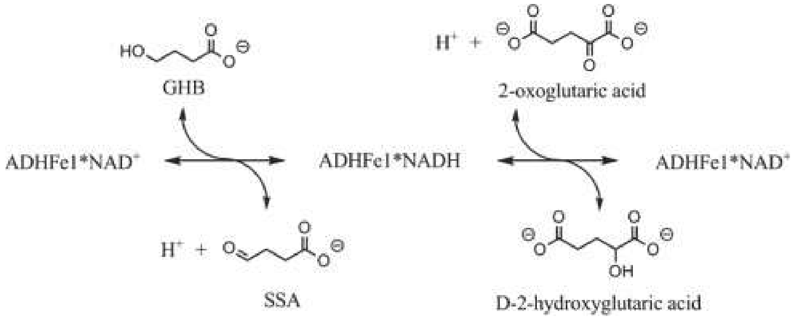

Figure 11.4 ADHFe1-catalyzed reaction. The asterisks indicate very tight complexes that do not release NAD^+ or NADH. Substrate-binding steps are implicit. GHB and 2-oxoglutarate are preferred substrates, but they are not the only substrates.

The known substrates differ in the length of the carbon chain, the terminal or internal position of the redox site, and the number of carboxylate groups. The great diversity in structures suggests that unidentified substrates exist. However, 2-hydroxyisovalerate, 3-hydroxyisovalerate, malate, 3-hydroxyglutarate, L-2-hydroxyglutarate, lactate, glycolate, and 2-hydroxybutanoic acid are not substrates. K_M values for 2-oxoglutaric acid differ up to 70-fold when the enzyme is prepared from liver, fibroblasts, or kidney (Struys et al., 2005b). A smaller variation occurs for SSA (Kaufman et al., 1988). This observation, combined with the existence of alternatively spliced mRNAs (above), raises the possibility that different organs predominantly express isoforms of ADHFe1 that have different substrate selectivities (and possibly cofactor requirement).

11.7.3 Poorly Characterized Catabolism of GHB

Additional pathways for catabolism of GHB *do* exist, but they are poorly characterized. When the GHB load in rodents and human beings is large, some of it is converted to 2,4-dihydroxybutanoic acid (also known as 3-deoxytetronic acid) and 3,4-dihydroxybutanoic acid (also known as 2-deoxytetronic acid) by α- and β-hydroxylations, respectively (Figure 11.2; Walkenstein et al., 1964; Lee, 1977; Jakobs et al., 1981; Niwa et al., 1982; Vayer et al., 1985b; Brown et al., 1987; Vamecq and Poupaert, 1990; Gibson et al., 1998; Shinka et al., 2002, 2003). These metabolites are also derived from sugars and fatty acids, so they are not distinctive for dosing by GHB (http://www.metabolibrary.ca/). Enzymes and subcellular compartments that might mediate α- and β-hydroxylations of GHB are discussed in Section 11.15.

An NAD^+-dependent GHB dehydrogenase activity has been reported in a crude preparation of mitochondria from human SH-SY5Y cells (Lyon et al., 2007). Unfortunately, the activity was not otherwise characterized. Could it be due to an alternatively spliced ADHFe1 that requires NAD^+?

"$NADP^+$-dependent dehydrogenase" was suggested to catalyze oxidation of GHB using tightly coupled reduction of D-glucuronate (Kaufman et al., 1979; Kaufman and Nelson, 1991). The name is confusing, as it names an oxidizing agent rarely used by the ADH superfamily in vivo (Section 11.10). The reaction proposed is similar in concept to obligatorily coupled oxidation of GHB and reduction of 2-oxoglutarate catalyzed by ADHFe1. In vitro, D-glucuronate does stimulate oxidation of GHB under certain conditions. However, the effect apparently is an artifact of optimized conditions that efficiently capture binary $NADP^+$ enzyme or NADPH·enzyme complexes with GHB or D-glucuronate, respectively. Under in vivo–like conditions for substrate concentrations and pH, bound NADPH has time to dissociate and join the pool of free NADPH. Such dissociation couples GHB oxidation to free $[NADP^+]/[NADPH]$ and not free [D-glucuronate]/[L-gulonate]. Ironically, $NADP^+$-dependent dehydrogenase is biosynthetic AKR1A1 discussed in Section 11.5.3. Moreover, the suggested coupling between the oxidation of GHB and the reduction of D-glucuronate has been tested comprehensively in rats and found wanting (Bhattacharya and Boje, 2007). Because critical information was not available when it was proposed, the hypothesis should be gently retired.

11.8 CHEMISTRY OF GHB AND RELATED METABOLITES NOT REQUIRING ENZYMES

A surprisingly large number of spontaneous reactions affects the stability, enzymatic kinetics, equilibrium constants, or subcellular compartmentalization of GHB-related compounds and reactions. GHB cyclizes to GBL with elimination of water, and GBL hydrolyzes to GHB over a period of several days at pH 2 to form an equilibrium mixture consisting of about one-third GHB and two-thirds GBL (Ciolino et al., 2001; Perez-Prior et al., 2005). At pH 7.4, the reactions require months to form an equilibrium mixture containing 1.5% GBL (Fishbein and Bessman, 1966). They are too slow to be of physiological significance, but they can become important during long-term storage of urine and blood for forensic analysis.

At 37°C and neutral pH, SSA covalently hydrates in about 1 min to form an equilibrium mixture consisting of approximately equal amounts of the free aldehyde and *gem*-diol (Le Henaff, 1967; Pocker and Dickerson, 1969). Dehydration occurs at a similar rate. Enzymes catalyzing reduction or oxidation of SSA act only on the free aldehyde (Miller and Pitts, 1967; O'Connor et al., 1999). Thus, dehydration can be the rate-limiting step for reduction or oxidation of SSA (Deetz et al., 1984; Marchitti et al., 2008). Hydrated SSA is likely to form in vivo primarily when newly synthesized SSA must be transported to a different compartment before the next covalent change to SSA occurs. This circumstance probably arises during biosynthesis of endogenous GHB (Section 11.5.4).

In tris(hydroxymethyl)aminomethane (Tris) buffer, SSA rapidly forms the imine (also known as the Schiff base) (Hauptmann and Gabler, 1968; Ogilvie and Whitaker, 1976). In organic solvent and thus possibly in fat droplets of adipocytes, SSA rapidly cyclizes to form an equilibrium mixture consisting of about one-third 4-hydroxybutyrolactone and two-thirds open-chain SSA (Jaeger et al., 2008). The cyclization does not occur in water at pH 7. A cyclic trimer of SSA called 1,3,5-trioxane–2,4,6-tripropanoic acid forms in organic solvent but not in dilute aqueous solution (Carriere, 1921). If 4-hydroxy-GBL is formed by cytochrome P450–mediated hydroxylation of GBL, it will undergo spontaneous ring opening to SSA. The presumed γ-hydroxybutyraldehyde intermediate of BD catabolism cyclizes in water to form the hemiacetal (Section 11.13.2).

11.9 EXPERIMENTAL EQUILIBRIUM CONSTANTS FOR REDOX REACTIONS OF GHB

The reactions considered in this section reversibly interconvert GHB and SSA. Irreversible oxidations (e.g., to form 2,4-dihydroxybutanoic acid, 3,4-dihydroxybutanoic acid, or succinic acid; see Section 11.15) are not considered. A reversible reaction exhibits an equilibrium constant that is independent of the enzyme used to establish the equilibrium or even whether an appropriate enzyme exists. The experimental constant (K_{eq}) obtained from literature sources has been adjusted in this section to pH 7.0 to yield K'_{eq}.[28] Experimental constants were obtained at 25°C and not 37°C, but the temperature effect is small.

A K_{eq} value for the reaction catalyzed by bacterial NAD$^+$-dependent GHB dehydrogenase in the direction toward SSA has been determined (Nirenberg and Jakoby, 1960). Unfortunately, Tris buffer was used, and spontaneous formation of the imine made oxidation of GHB much more favorable than it would be in vivo. Nevertheless, there is a way to estimate a relevant value. As the structures of the alcohol functional groups in GHB and other simple primary alcohols (except methanol) are essentially the same as each other, the K_{eq} values for oxidation to aldehydes by NAD$^+$ must be essentially the same. They are available for ethanol, 3-hydroxypropionate, and 1-butanol in phosphate or pyrophosphate buffer (Goldberg et al., 1993). Adjusted to pH 7.0, the values are $9.97 \times 10^{-5}, 9 \times 10^{-5}$, and 9.1×10^{-5}, respectively, which as expected are essentially the same. Accordingly, the average value (9.5×10^{-5}) is assumed for K'_{eq} in the NAD$^+$-dependent oxidation of GHB. Formation of aldehyde hydrate is ignored, as the reaction occurs to the same extent for all of these aldehydes and is physiological. Also, the reciprocal of K'_{eq} for the NAD$^+$-dependent oxidation of GHB gives K'_{eq} for the NADH-dependent reduction of SSA to GHB, or 1.1×10^4.

Because the reduction potentials for NADP$^+$ and NAD$^+$ are essentially the same, the K'_{eq} values for the NADP-dependent reactions are essentially the same as for the NAD-dependent reactions, namely 9.5×10^{-5} for NADP$^+$-dependent oxidation of GHB and 1.1×10^4 for NADPH-dependent reduction of SSA to GHB at pH 7.0.[29]

K_{eq} for the reaction between GHB and 2-oxoglutarate catalyzed by ADHFe1 has not been determined. However, the sum of the NAD$^+$-dependent GHB dehydrogenase-catalyzed and NAD$^+$-dependent D-2-hydroxyglutarate dehydrogenase-catalyzed reactions in opposite directions is the ADHFe1-catalyzed reaction. K'_{eq} for the half-reaction to form D-2-hydroxyglutarate in phosphate buffer is 6.8×10^4 (Buckel and Miller, 1987). K'_{eq} for the reaction catalyzed by ADHFe1 is the product of the constants for the half-reactions, or $(9.5 \times 10^{-5})(6.8 \times 10^4) = 6.5$. The prime in K'_{eq} is not strictly needed because the equilibrium is independent of pH. For simplicity of nomenclature it will be retained.

K'_{eq} for this reaction can also be calculated using published Michaelis–Menten parameters and the Haldane relationship for a two-substrate, two-product enzyme ($K'_{eq} = V_{\text{maxforward}} K_{\text{Mproduct1}} K_{\text{Mproduct2}} / V_{\text{maxreverse}} K_{\text{Msubstrate1}} K_{\text{Msubstrate2}}$). For ADHFe1 from human liver, the estimate is 7.5, and for ADHFe1 from rat kidney, it is 12 (Kaufman et al., 1988; Struys et al., 2005b). All three of the values calculated are reassuringly similar. The average of 8.7 is taken here for K'_{eq} in the ADHFe1 reaction.

11.10 ESTIMATED EQUILIBRIUM CONSTANTS FOR REDOX REACTIONS OF GHB IN VIVO

This section incorporates into K'_{eq} expressions the pH and [oxidant]/[reductant] values found in vivo. The resulting numbers are modified equilibrium constants that predict $[\text{SSA}]_{\text{eq}}/[\text{GHB}]_{\text{eq}}$ values for cytosolic and mitochondrial compartments. They differ $>10^{10}$-fold, depending on the direction of reaction, type of oxidant or reductant, and the subcellular compartment. The values are instructive in determining which class of redox enzyme in which compartment can function at a particular metabolic step. Modified equilibrium constants do not inform among isozymes utilizing the same [oxidant] and [reductant] reaction couple, however.

In mitochondrial matrix, $\text{pH} = 7.7$, free $[\text{NAD}^+]/[\text{NADH}] \approx 7.7$ (coincidentally the same number as the pH), and free $[\text{NADP}^+]/[\text{NADPH}] \approx 12.$[30] In cytosol, the values are $\text{pH} = 7.4$, free $[\text{NAD}^+]/[\text{NADH}] \approx 1164$, and free $[\text{NADP}^+]/[\text{NADPH}] \approx 0.014$ (Veech et al., 1969; Zhang et al., 2002; Balut et al., 2008). Free $[\text{NAD}^+]/[\text{NADH}]$ and free $[\text{NADP}^+]/[\text{NADPH}]$ exclude enzyme-bound forms, which substantially exceed the free forms and distort many published ratios. The uncertainty in the ratios is about twofold, which in the current discussion is small (Lin and Guarente, 2003). Moreover, dosing of rats with GHB does not affect cytosolic $[\text{NAD}^+]/[\text{NADH}]$ or $[\text{NADP}^+]/[\text{NADPH}]$, although minor alkalosis occurs (MacMillan, 1979). Thus, metabolic homeostasis is nearly maintained. Although extensive analyses of NAD and NADP levels in different organs are not available, mitochondrial redox states in normal brain, heart, kidney, liver, and testis are similar to each other (Mayevsky and Chance, 2007). Published values for pH, free $[\text{NAD}^+]/[\text{NADH}]$, and free $[\text{NADP}^+]/[\text{NADPH}]$ in cytosol and mitochondrial matrix should be applicable to all major organs. Adjustment of K'_{eq} values to incorporate pH, free $[\text{NAD}^+]/[\text{NADH}]$, and free $[\text{NADP}^+]/[\text{NADPH}]$ yields $K_{\text{red}}^{\text{mito}}, K_{\text{red}}^{\text{cyto}}, K_{\text{ox}}^{\text{mito}}$, and $K_{\text{ox}}^{\text{cyto}}$[31] values in the reducing or oxidizing direction in mitochondrial matrix or cytosol (each as indicated) (Table 11.4). No reversible GHB metabolism is known to occur in other subcellular compartments.

Reactions catalyzed by NAD^+-dependent ADHs are addressed first. For oxidation of GHB, $K_{\text{ox}}^{\text{mito}} \approx 3.7 \times 10^{-3}$ and $K_{\text{ox}}^{\text{cyto}} \approx 2.8 \times 10^{-1}$. The reaction is unfavorable in both compartments, but not so unfavorable that similar reactions do not occur. For example, NAD^+ oxidizes malate to oxaloacetate in the mitochondrial matrix ($K_{\text{ox}}^{\text{mito}} \approx 4 \times 10^{-4}$, or about 10-fold smaller than estimated here for GHB) and NAD^+ oxidizes ethanol to acetaldehyde in cytosol ($K_{\text{ox}}^{\text{cyto}} \approx 2.8 \times 10^{-1}$, or equal to that estimated here for GHB). It is clear that NAD^+ *could* oxidize GHB in both cytosol and mitochondrial matrix. Whether it *does* depends on additional factors discussed in Sections 11.11 and 11.15.

The reciprocals of the constants for oxidation of GHB by NAD^+ give the constants for reduction of SSA by NADH. They are $K_{\text{red}}^{\text{mito}} \approx 2.7 \times 10^2$ and $K_{\text{red}}^{\text{cyto}} \approx 3.6 \times 10^0 = 3.6$. The reaction is favorable in both compartments. However, of six human ADH isozymes that have been tested in vitro under conditions even more favorable to reduction than exist in vivo, none catalyzes the reduction of SSA (Deetz et al., 1984). Why? Apparently, none of the isozymes forms a productive complex with SSA. Also, no other well-documented mammalian NADH-dependent ADH reduces SSA. Not only must the modified equilibrium constant be acceptable

TABLE 11.4 Modified Equilibrium Constants for Reactions Interconverting GHB and Succinic Semialdehyde[a]

Reduction Reaction	Enzyme[b]	$K_{\mathrm{red}}^{\mathrm{mito}}$	$K_{\mathrm{red}}^{\mathrm{cyto}}$
$\mathrm{SSA} + \mathrm{NADH} + \mathrm{H}^+ \leftrightarrow \mathrm{GHB} + \mathrm{NAD}^+$	NAD^+-dependent GHB dehydrogenase (reverse)	2.7×10^2	3.6×10^0
$\mathrm{SSA} + \mathrm{NADPH} + \mathrm{H}^+ \leftrightarrow \mathrm{GHB} + \mathrm{NADP}^+$	SSA reductase	1.8×10^2	1.5×10^5
Oxidation Reaction	**Enzyme**	$K_{\mathrm{ox}}^{\mathrm{mito}}$	$K_{\mathrm{ox}}^{\mathrm{cyto}}$
$\mathrm{GHB} + \mathrm{NAD}^+ \leftrightarrow \mathrm{SSA} + \mathrm{NADH} + \mathrm{H}^+$	NAD^+-dependent GHB dehydrogenase	3.7×10^{-3}	2.8×10^{-1}
$\mathrm{GHB} + \mathrm{NADP}^+ \leftrightarrow \mathrm{SSA} + \mathrm{NADPH} + \mathrm{H}^+$	NADP^+-dependent GHB dehydrogenase or SSA reductase (reverse)	5.7×10^{-3}	6.7×10^{-6}
GHB + 2-oxoglutarate ↔ SSA+ D-2-hydroxyglutarate	2-Oxoglutarate-dependent GHB transhydrogenase	> 8.7[c]	-

[a] All equilibrium constants are unitless. The experimentally determined, pH-independent, overall equilibrium constant at 25°C from the literature was adjusted to pH 7.0 (footnote 28). Then the constants were adjusted to mitochondrial and cytoplasmic conditions to give $K_{\mathrm{red}}^{\mathrm{mito}}$, $K_{\mathrm{red}}^{\mathrm{cyto}}$, $K_{\mathrm{ox}}^{\mathrm{mito}}$, and $K_{\mathrm{ox}}^{\mathrm{cyto}}$ in the reducing and oxidizing directions, respectively, by the reductant and oxidant couple indicated (footnote 31). The values apply to most or all organs. Values less than 10^{-4} probably cannot support metabolism in the direction written, and values greater than 10^4 probably produce irreversible reaction (Section 11.4.5).

[b] The enzyme names are phenomenological descriptions and not assigned isozyme names.

[c] This value is a lower boundary based on [2-oxoglutarate]/[D-2-hydroxyglutarate] > 1. The actual value probably is ≥ 870 (Section 11.10).

in vivo, but an isozyme that binds SSA productively must exist for the reaction to occur.

Reactions catalyzed by AKRs are addressed next. For reduction of SSA and other simple aldehydes by NADPH, $K_{\mathrm{red}}^{\mathrm{mito}} \approx 1.8 \times 10^2$ and $K_{\mathrm{red}}^{\mathrm{cyto}} \approx 1.5 \times 10^5$. The reaction is strongly favorable in both compartments. Cytosolic reduction probably is irreversible, as $K_{\mathrm{red}}^{\mathrm{cyto}}$ is well above the threshold of 10^4 for irreversibility (Section 11.4.5)! But is a human AKR known that catalyzes this reaction? Yes, AKR1A1 and AKR7A2 do so (Section 11.5.3).

The reciprocals of $K_{\mathrm{red}}^{\mathrm{mito}}$ and $K_{\mathrm{red}}^{\mathrm{cyto}}$ for AKRs give the constants for oxidation of GHB and other simple alcohols by NADP^+. They are $K_{\mathrm{ox}}^{\mathrm{mito}} \approx 5.7 \times 10^{-3}$ and $K_{\mathrm{ox}}^{\mathrm{cyto}} \approx 6.7 \times 10^{-6}$. The reaction is unfavorable in both compartments. Nevertheless, as $K_{\mathrm{ox}}^{\mathrm{mito}}$ is ≈ 10-fold more favorable than for oxidation of malate (although by NAD^+), oxidation of GHB by NADP^+ in mitochondrial matrix probably *could* occur. Whether it *does* occur depends on additional factors discussed in Section 11.11. In contrast, NADP^+-dependent oxidation of GHB in cytosol is nearly 1000-fold less favorable than in mitochondrial matrix and well below the 10^{-4} threshold for function (Section 11.4.5). Cytosolic NADP^+-dependent GHB dehydrogenase has very little chance of functioning in vivo.

The oxidation of GHB by 2-oxoglutarate catalyzed by ADHFe1 in mitochondrial matrix is addressed next. Free [2-oxoglutarate]/[D-2-hydroxyglutarate] controls $K_{\mathrm{ox}}^{\mathrm{mito}}$. The ratio is not known. Nevertheless, a useful boundary on its value can be set. FAD-dependent (*not* NAD$^+$-dependent) D-2-hydroxyglutarate dehydrogenase is present in human and rat mitochondrial inner membrane facing the matrix (Rzem et al., 2004). Mutations in the gene lead to a human disorder called D-2-hydroxyglutaric aciduria, which presents with up to 30-fold increase in the concentration of D-2-hydroxyglutarate in urine (Struys, 2006; Wickenhagen et al., 2009). The phenotype tells us that FAD-dependent D-2-hydroxyglutarate dehydrogenase controls free [2-oxoglutarate]/[D-2-hydroxyglutarate]. Moreover, the ratio in normal people is > 1, as otherwise a reservoir of 2-oxoglutarate sufficient to produce the observed large increase in [D-2-hydroxyglutarate] when the dehydrogenase is blocked would not exist. The physiological observations are consistent with biochemical determinations that oxidations by FAD-dependent ADHs are ≥ 100-fold more favorable than oxidations by NAD$^+$-dependent ADHs (Goldberg and Alberty, 2008). Because [2-oxoglutarate]/[D-2-hydroxyglutarate] surely is $> 1, K_{\mathrm{ox}}^{\mathrm{mito}}$ for ADHFe1 surely is > 8.7, and it probably is ≥ 870 (Table 11.4). Thus, the reaction of GHB with 2-oxoglutarate catalyzed by ADHFe1 probably is very favorable in vivo. However, it is reversible in vitro (Section 11.7.2).

If isoforms of ADHFe1 exist that prefer an oxidant other than 2-oxoglutarate or are present in cytoplasm, perhaps as a result of alternative splicing of the mRNA (Section 11.7.2), the calculation just presented could change. Reduction of SSA catalyzed by ADHFe1 in vivo has not been proposed, and as already indicated, it is very unlikely to occur due to a low SSA concentration. It will not be addressed.

11.11 DIFFERENT PERSPECTIVES ON TURNOVER OF ENDOGENOUS GHB ARE CONSISTENT

Cytosolic reduction of SSA to GHB catalyzed by NADPH-dependent AKRs, and oxidation of GHB to SSA in mitochondrial matrix catalyzed by ADHFe1, are strongly favored among reactions interconverting SSA and GHB. This conclusion is consistent with most of what is known about substrate selectivity by AKR and ADH isozymes, subcellular distributions of isozymes, and the presence of MCTs in mitochondrial inner membrane. If one hypothesizes that transport of SSA out of mitochondrial matrix is rate limiting for biosynthesis of endogenous GHB, and transport of GHB into mitochondrial matrix is rate limiting for degradation of exogenous but not endogenous GHB, a generally consistent picture emerges. Endogenous GHB is made by excess AKR1A1 and AKR7A2 activity in cytosol. Because its concentration is low, cycling endogenous GHB probably is degraded almost entirely by ADHFe1. In contrast, exogenous GHB probably saturates mitochondrial uptake and/or ADHFe1 activity. Under these conditions, other pathways for catabolism of GHB set in, as evidenced by the formation of products other than SSA. Several possible alternative catabolic pathways are discussed in Section 11.15.

11.12 DISPOSITION OF SUCCINIC SEMIALDEHYDE

Most SSA is irreversibly and rapidly oxidized to succinate by SSA dehydrogenase in mitochondrial matrix (Ryzlak and Pietruszko, 1988).[32] GABA aminotransferase, which in the absence of ingested GHB or GHB prodrug makes most and perhaps all SSA, might form a complex with SSA dehydrogenase (Hearl and Churchich, 1984). Such a complex could allow most newly made SSA to be transferred directly into the active site of SSA dehydrogenase without dissociation into free solution. In this event, escape from channeling would be required for export of GABA-derived SSA from mitochondria (Section 11.5.2). The idea is untested, but such an escape could be the rate-limiting step in GHB biosynthesis and a regulatory control point. More should be learned about the reported complex between GABA aminotransferase and SSA dehydrogenase.

Even though the selectivity of ALDH isozymes for SSA has not been studied systematically, ALDH5A1 can be assigned to SSA dehydrogenase activity in vivo (Marchitti et al., 2008). The primary reason is that rare genetic defects map to the ALDH5A1 gene in human beings, cause GHB aciduria, and block SSA dehydrogenase activity (Knerr et al., 2007; Malaspina et al., 2009). The enzyme has been expressed in *E. coli*, purified to homogeneity, and characterized as a soluble homotetramer of 58 kDa monomers 544 residues long (Kang et al., 2005). It strongly prefers NAD^+ over $NADP^+$, and the K_M value for SSA is $\sim 6\ \mu M$. The protein and mRNA are present in essentially all organs at levels that correlate approximately with those for GABA transaminase, which is consistent with the possible existence of substrate channeling between GABA transaminase and ALDH5A1 (Miller and Pitts, 1967; Chambliss et al., 1995a,b; Tillakaratne et al., 1995; http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/). Two isoforms of ALDH5A1 presumably exist, although they have not been observed directly, as alternative splicing in the coding region of the mRNA occurs (Blasi et al., 2002). Both isoforms are predicted to contain N-terminal mitochondrial targeting sequences (Claros and Vincens, 1996). Succinate resulting from ALDH5A1 activity enters the tricarboxylic acid cycle for final disposition, which includes conversion to α-amino acids and CO_2 (Figure 11.2).

Mice carrying deletion of the orthologous gene (Aldh5a1) have been created as a model for human GHB aciduria (Hogema et al., 2001; Pearl et al., 2009). The block causes levels of GHB and GABA in urine and organ extracts to increase as much as about 20- and threefold, respectively (Gibson et al., 2002; Gupta et al., 2004; Chowdhury et al., 2007). The different percent increases support the notion that different GHB/GABA ratios can arise from causes not related to the existence of a GABA-independent pathway to endogenous GHB (Section 11.5.5). Metabolic profiles of people and mice lacking ALDH5A1 activity are similar to those of genetically normal people who ingest large amounts of GHB (Brown et al., 1987; Gibson et al., 1997; Shinka et al., 2002, 2003; Struys et al., 2006a,b; Sauer et al., 2007).

Human beings and mice lacking SSA dehydrogenase activity produce a small amount of 4,5-dihydroxyhexanoic acid and its γ-lactone [both compounds occur as

(4*R*, 5*R*) and (4*S*, 5*R*) stereoisomers] (Brown et al., 1987; Gibson et al., 2002). The reaction probably occurs by condensation of SSA with the nucleophilic thiamine-bound acetyl intermediate created by pyruvate decarboxylase in the pyruvate dehydrogenase complex (Shaw and Westerfeld, 1968; Brown et al., 1987). Resulting 5-oxo-4-hydroxyhexanoic acid [in a mixture of (4*R*) and (4*S*)] could then be reduced by an AKR at the 5-oxo center. Spontaneous cyclization would produce the γ-lactone. The condensates are not found in baboons acutely dosed with GHB, however, which suggests that they might not be satisfactory markers for acute dosing of people (Struys et al., 2006b). Moreover, the γ-lactone, which is also known as solerole, is found in many foods and beverages, further indicating that these compounds are not satisfactory markers for GHB dosing (Krajewski et al., 1995; Aubert et al., 2003; Rocha et al., 2005).

The atomic-resolution structure of human ALDH5A1 has been obtained (Kim et al., 2009). The enzyme is controlled by redox modulation. It is active when two cysteine residues in the active site are reduced and is inactive when the cysteines are oxidized to the disulfide. Reactive oxygen species such as H_2O_2 cause formation of the disulfide. This phenomenon links GHB catabolism to other oxidative physiology (Mamelak, 2007). $GABA_A$ and NMDA receptors are regulated by redox modulation as well, thereby creating the potential for high-order physiological synergism from abuse of GHB, flunitrazepam, ketamine, and ethanol (Pan et al., 2000; Ikonomidou et al., 2000)! Indeed, GHB recently was shown to increase oxidative stress in vivo by decreasing defenses to oxidative insults (Sgaravatti et al., 2007, 2009).

11.13 CONVERSION OF PRODRUGS TO GHB AND RELATED METABOLITES

11.13.1 γ-Butyrolactone

Exogenous GBL is rapidly converted to GHB in human beings, baboons, cats, and rodents (Roth and Giarman, 1966; Richard et al., 2005; Lenz et al., 2008, Goodwin et al., 2009). It is thus a prodrug that is hydrolyzed by γ-lactonase present primarily in blood and liver (Figure 11.2; Ingels et al., 2000; Teiber et al., 2003; Draganov et al., 2005). There are three γ-lactonase isozymes in human beings, but GBL is hydrolyzed overwhelmingly by γ-lactonase 1 in vitro (Billecke et al., 2000). The enzyme is very nonselective and targets many xenobiotic compounds, as it hydrolyzes phosphotriesters, aryl carboxylate esters, carbonates, and lactones. There is very little γ-lactonase activity in hen, rat, and rabbit brains (Monroy-Noyola et al., 2007). The low activity in brain, combined with the good ability of low-polarity GBL to permeate membranes nonspecifically (Section 11.14) and pass into the periphery (Sgaragli and Zilletti, 1970; Abraham et al., 1995), probably explains why GBL injected into a cerebral ventricle has no behavioral effect in cat, rabbit, and rat (Gessa et al., 1967; Carter et al., 2006a).

GBL is predicted by a computer-based, expert toxicological system to undergo hydroxylations by cytochrome P450 to give 3-hydroxy-GBL and 4-hydroxy-GBL (Rosenkranz, 2001). Although not tested, both of these compounds probably would be hydrolyzed by a γ-lactonase. Also, 4-hydroxy-GBL will undergo spontaneous ring opening to SSA. Hydrolysis of 3-hydroxy-GBL produces 3,4-dihydroxybutanoic acid, which is made as well by β-oxidation of GHB and metabolism of carbohydrates and fatty acids. It would be of interest to determine whether 3,4-dihydroxybutanoic acid arising from GBL and exogenous GHB differs in chirality from 3,4-dihydroxybutanoic acid arising from carbohydrates and fatty acids (Sakata, 1990).

11.13.2 1,4-Butanediol

Similar to GBL, exogenous BD is rapidly converted to GHB in human beings, rodents, and baboons. It thus is a prodrug (Figure 11.2; Maxwell and Roth, 1972; Vree et al., 1978; Poldrugo and Snead, 1984; Snead et al., 1989; Ingels et al., 2000; Richard et al., 2005; Irwin, 2006; Kapadia et al., 2007; Goodwin et al., 2009). Metabolism of BD in the periphery is inhibited by pyrazole and 4-methylpyrazole, which are inhibitors of many Zn^{2+}-dependent ADHs (Bessman and McCabe, 1972; Quang et al., 2004; Carai et al., 2006). Partially purified ADHs, probably containing mixtures of isozymes obtained from human and horse livers, oxidize BD in vitro (Pietruszko et al., 1978). In contrast, metabolism of BD injected into a cerebral ventricle of rat is not inhibited by pyrazole (Poldrugo and Snead, 1986; Snead et al., 1989). Ethanol competitively inhibits BD metabolism in both the periphery and brain. The findings suggest that BD is metabolized by (1) ADH in the periphery, similarly to ethanol, and (2) a non-ADH pathway in brain that ethanol also follows (Gerak et al., 2004; Nishimura and Naito, 2005). As 4-methylpyrazole blocks the behavioral effects of peripherally administered BD in mice, most BD metabolism occurs in the periphery (Carai et al., 2002; Quang et al., 2002). The metabolic competition between BD and ethanol has been simulated numerically, and metabolism of BD has been shown to be capacity limited (Fung et al., 2008). A weak correlation between slower metabolism of BD in human beings and a common allele of ADH1B has been reported (Thai et al., 2007).[33] Nevertheless, no compelling evidence supports assignment of BD dehydrogenase to a particular ADH isozyme. A moderate rate of conversion of BD to GHB in brain, combined with the good ability of low-polarity BD to permeate membranes nonspecifically (Section 11.14) and pass into the periphery (Abraham et al., 1995), probably explains why BD injected into a cerebral ventricle has no behavioral effect in rat (Carter et al., 2006a).

The first step in conversion of BD to GHB, no matter what the enzymatic mechanism, presumably produces γ-hydroxybutyraldehyde. About 86% of this compound spontaneously cyclizes in aqueous solution to the hemiacetal called 2-hydroxytetrahydrofuran (Hurd and Saunders, 1952). However, neither the open-chain nor the hemiacetal form has been identified in human beings or rodents. A search for these compounds, or a distinctive condensate of them (as occurs for SSA, Section 11.12), could be rewarding. Such a metabolite might be more metabolically stable than are BD and GHB and serve as an indicator of dosing after BD has disappeared from blood and urine. The open-chain form of γ-hydroxybutyraldehyde is probably oxidized to GHB by an ALDH, as the ALDH inhibitor disulfiram partially blocks the behavioral effects of BD in mice (Carai et al., 2002). The cyclized hemiacetal form could be oxidized to GBL, but whether this occurs in vivo (e.g., catalyzed by a cytochrome P450 or catalase; see Section 15) is not known.

11.14 SUBCELLULAR COMPARTMENTALIZATION OF GHB-RELATED COMPOUNDS

The logarithm to the base 10 of the distribution coefficient $D\,(\log D)$ measures partitioning of a compound between octanol and water at a given pH and 25°C. At a pH of about 7, the GHB-related compounds GABA, SSA, GHB, 2,4-dihydroxybutanoic acid, 3,4-dihydroxybutanoic acid, and 4,5-dihydroxyhexanoic acid each carry one or two electrical charges, and they have $-\log D$ values of ≥ 2.65 (Table 11.5). In vivo these compounds are subject to compartmentalization controlled by membrane transporters. This restriction in turn means that only certain isozymes can access the compounds.

In contrast, ingested GHB is exposed to intragastric pH between 1 and 4, depending on the person, time of day, and presence or absence of food in the stomach (van Herwaarden et al., 1999). Because its pK_a value is 4.72, GHB is predominantly protonated and uncharged while it is in the stomach (Tao et al., 2008). Protonated GHB has a $-\log D = 0.70$, which is 200-fold less polar than unprotonated GHB is (Table 11.5). This value approaches that for ethanol ($-\log D = 0.19$), which the stomach absorbs nonspecifically (Levitt et al., 1997). Thus, the stomach probably absorbs some ingested GHB nonspecifically. Ingested GHB is not exposed to the acid in the stomach long enough to cyclize to GBL (Sections 11.6.1 and 11.8).

TABLE 11.5 Octanol–Water Partition Coefficients

GHB Related	$-\log D^a$	Prodrug Related	$-\log D^a$
GABA	3.14	GBL	0.76
SSA	2.65	BD	1.02
GHB	3.01, 0.70*	4-Hydroxybutyraldehyde	0.57
2,4-Dihydroxybutanoate	4.18	2-Hydroxytetrahydrofuran^b	0.93
3,4-Dihydroxybutanoate	4.79		
4,5-Dihydroxyhexanoate	3.63		

[a] -Log D values are for pH 7 and 25°C, except for the GHB value marked with *, which applies to pH 1–3 (GHB is mostly protonated and uncharged in this pH range in the stomach). The acids are named to indicate they are charged. Estimates were done by SciFinder Scholar with Advanced Chemistry Development (ACD/Labs) Software V8.14 for Solaris. Larger positive values mean less solubility in octanol, which correlates approximately with less nonspecific permeation through membranes. Compare these values with the values for ethanol (0.19) and acetaldehyde (0.16), which quickly and nonspecifically permeate biological membranes by partitioning into and across the lipid bilayer.

[b] Cyclized 4-hydroxybutyraldehyde in the hemiacetal form.

GBL, BD, and the probable BD metabolites 4-hydroxybutyraldehyde and 2-hydroxytetrahydrofuran are uncharged and have small $-\log D$ values at all physiological values of pH (Table 11.5). These molecules probably permeate membranes nonspecifically (Roth and Giarman, 1966; Lenz et al., 2008). BD has been shown to permeate barrier epithelia resistant to polar but uncharged molecules (Wright and Pietras, 1974). Thus, ingested GBL and BD are absorbed rapidly and efficiently by the stomach (Arena and Fung, 1980; Perez-Prior et al., 2005; Thai et al., 2007; Fung et al., 2008; Hicks and Varner, 2008; Goodwin et al., 2009). Poor compartmentalization is especially relevant to BD, which consequently has access to many, perhaps all, isozymes that could act on it during its presumed two-step conversion to GHB (Section 11.13.2).

11.15 COMPARATIVE CATABOLISM OF ETHANOL, 1,4-BUTANEDIOL, FATTY ACIDS, AND GHB

Most ethanol consumed by human beings is catabolized in the stomach and liver, which use primarily NAD$^+$-dependent cytosolic ADH1B and ADH1C for the initial step to acetaldehyde (Lee et al., 2006). However, these organs also use hydrogen peroxide (H_2O_2)-dependent catalase located in peroxisomal matrix and NADPH- and oxygen (O_2)-dependent cytochrome P450 2E1 located on the cytosolic face of smooth endoplasmic reticulum for the initial step to acetaldehyde. Some ethanol is catabolized by the brain, which uses primarily catalase and cytochrome P450 2E1 (Zimatkin et al., 2006; Hipolito et al., 2007). In the catalase reaction, H_2O_2 is reduced to two water molecules as the alcohol group is oxidized to the aldehyde group. In the cytochrome P450 reaction, one of the O atoms from O_2 attacks C1 of ethanol to form hydrated acetaldehyde, and the other O atom is reduced to H_2O by NADPH. These reactions are irreversible. Human beings have one gene for catalase but ≥ 58 genes for cytochrome P450s. Overall, human beings use *three* types of enzymatic reactions to oxidize ethanol to acetaldehyde. The prominence of each type depends on the organ, and due to different K_M values, probably the dose of ethanol. Does similar parallelism occur in the catabolism of BD and GHB?

Observations presented in Section 11.13.2 strongly suggest that metabolism of BD in the periphery and brain proceeds along the same pathways that ethanol uses. Consistent with this hypothesis, BD is oxidized in vitro to formaldehyde and uncharacterized products by cytochrome P450 in rat liver microsomes (Clejan and Cederbaum, 1992).[34] The BD isomer 2,3-butandiol is oxidized in vitro to 3-hydroxy-2-butanone by catalase (Magner and Klibanov, 1995). Unidentified isozymes of NAD$^+$-dependent ADH also act on BD in vitro (Section 11.13.2). Thus, all three types of enzymes acting on ethanol in human beings also act on BD or BD isomer in vitro.

Similar studies should be carried out on GHB, which can be conceptualized as either a primary alcohol such as ethanol or a small fatty acid. As a primary alcohol, catalase and cytochrome P450 could convert GHB to SSA. As a fatty acid, catalase and cytochrome P450 could convert GHB to 2,4-dihydroxybutanoic acid by α-hydroxylation (Poulos et al., 1993; Alderson et al., 2004; http://pfam.sanger.ac.uk/family?acc=PF00067). Also as a fatty acid, mitochondria and peroxisomes could convert GHB to 3,4-dihydroxybutanoic acid by β-hydroxylation (Visser et al., 2007). In addition, a cytochrome P450 might convert GHB to SSA by ω-oxidation (Orrenius, 1969). As discussed in Section 11.7.3, large amounts of 2,4-dihydroxybutanoic acid and 3,4-dihydroxybutanoic acid, respectively, are formed after dosing by GHB. No in vitro test of whether catalase or cytochrome P450 acts on GHB has been reported. However, GHB competitively inhibits demethylation of *N*-nitrosodimethylamine by rat cytochrome P450 2E1 with an IC_{50} value of 1.4 mM, which implies that it probably is bound to the active site (Wang et al., 1995). The IC_{50} value is similar to the concentrations of GHB reached during dosing.

Overall, there are ample hints that exogenous BD and GHB follow multiple catabolic pathways in human beings. It would be of great interest to characterize them systematically in vitro and in vivo. GBL probably will follow GHB pathways, as it is rapidly hydrolyzed in vivo (Section 11.13.1).

11.16 CATABOLISM OF MDMA, FLUNITRAZEPAM, AND KETAMINE

Because MDMA, flunitrazepam, and ketamine are not natural compounds, there are no biosynthetic pathways for them. The constraints of a metabolic network do not apply (Section 11.4.5). Even the Haldane relationship does not apply, as most of the steps are irreversible and do not have a K_{eq}. For MDMA the major end metabolite is 4-hydroxy-3-methoxymethamphetamine sulfate ester, for flunitrazepam it is 7-aminoflunitrazepam, and for ketamine it is norketamine (Figure 11.1) (Feng et al., 2007).

11.17 DETECTION OF DATE-RAPE DRUGS

11.17.1 Compounds Diagnostic for Dosing by Synthetic Date-Rape Drugs

Because MDMA, flunitrazepam, and ketamine, and their metabolites are not part of natural metabolism, the presence of any one of them in a physiological sample is per se evidence for ingestion of the drug. However, as flunitrazepam and ketamine are used occasionally in medicine, evidence of dosing is not necessarily evidence of drug-assisted sexual assault or self-abuse. The context must be considered in reaching a conclusion.

11.17.2 Compounds Diagnostic for Dosing by GHB

D-2-hydroxyglutaric acid, 4,5-dihydroxyhexanoic acid, and SSA have strong, direct metabolic links to GHB (Figure 11.2; Sections 11.7.2 and 11.12). Nevertheless, none of these metabolites currently is a good marker of dosing, as the endogenous levels in undosed human beings are only partially characterized. Variance in endogenous levels of 2,4-dihydroxybutanoic acid and 3,4-dihydroxybutanoic acid, and of their breakdown products (formic acid, 3-hydroxypropanoic acid, glycolic acid, and acetic acid), is very large because of person-to-person differences in metabolism and diet. Elevated levels provide little evidence for dosing by GHB (Chalmers et al., 1976). GBL and BD are excreted at detectable levels only if they per se have been ingested and not after GHB has been ingested (McCusker et al., 1999; Blanchet et al., 2002; Lora-Tamayo et al., 2003; Wood et al., 2004; Shima et al., 2005).[35]

The amount of GHB present in blood and urine taken from undosed normal people has been comprehensively characterized, and it is always less than 5 and 10 μg/mL, respectively (Elliott, 2003; LeBeau et al., 2006). Persons suffering from uncontrolled diabetes excrete large amounts of the GHB isomers α- and β-hydroxybutanoic acid, but they excrete no excess GHB (Shima et al., 2005). Persons suffering from SSA dehydrogenase deficiency excrete large amounts of the GHB, but the frequency of this disorder is so low that it can be ignored in testing of the public for dosing (Koelker et al., 2008). Even though 96 to 99% of ingested GHB or prodrug typically undergoes catabolism, the current best physiological marker for dosing by these substances is a value of GHB above 5 or 10 μg/mL in serum or urine, respectively (Villain et al., 2003; Yeatman and Reid, 2003; Crookes et al., 2004; Kasprzak et al., 2006; LeBeau et al., 2006). As discussed above for other date-rape drugs, evidence of dosing is not necessarily evidence for drug-assisted sexual assault or self-abuse, as GHB is occasionally used in medicine. The context must be considered.

11.17.3 Gold-Standard Testing

Most regional forensic laboratories and many large clinical laboratories conduct *gold-standard tests* for date-rape drugs and their metabolites. Such tests include, but are not limited to, such techniques as spiking of a urine sample with deuterated drug or metabolite, extraction into organic solvent, chemical derivatization to improve physical properties, high-performance liquid or gas chromatography to reduce background signals, and mass spectrometry to quantitate isotope dilution (Kastrissios et al., 2005; Rendle, 2005; de Oliveira et al., 2007; Kraemer and Paul, 2007; Langman, 2007; Maurer, 2007; Musshoff and Madea, 2007a,b; Pizzolato et al., 2007; Pragst, 2007; Samyn et al., 2007; Smith et al., 2007; Snow, 2007; Tagliaro et al., 2007; de Paoli and Bell, 2008; Gong et al., 2008). Such sophisticated methods can provide superior sensitivity as well as quantitative and unambiguous results that constitute evidence of dosing suitable to systems of justice. However, the certitude that these methods provide is gained at high cost. Expensive instrumentation, highly trained persons to maintain the instruments and perform the tests, relatively slow turnaround, and substantial operating costs are characteristic. For these reasons, gold-standard testing must be strongly justified by circumstances. It is usually not used for routine screening and not available outside high-end laboratories.

11.17.4 Many Applications for Reliable Field Tests

Rapid, reliable, inexpensive, and semiquantitative testing by responsible principals could clarify many situations in which illicit drugs are suspected. Such test results assist in making rapid decisions and taking immediate appropriate action. The principals include staff persons working in hospitals and medical clinics, narcotics enforcement, prisons, and probation departments. The public also has an interest in reliable and easy detection of illicit drugs in bars, schools, and households. These and similar circumstances are termed “field” situations in this chapter to distinguish them from laboratory situations.

An international panel has recommended that at least two uncorrelated test methods be used to document presence of an illicit drug when the evidence is to be presented to a judicial system. The result of a validated field test can qualify for one of the methods if it is supported by the result from a gold standard, category A test (Santos, 2007). An opportunity for synergism exists here, as gold-standard testing can be made to be faster, more efficient, and less expensive by targeting the illicit drug putatively identified in a field test (Cone and Huestis, 2007; Peters, 2007).

11.17.5 Hospital Emergency Department Example

Because their molecular mechanisms of action are generally different, each agent causes some distinctive symptoms after an overdose. However, when agents are co-ingested, as often occurs, mixed symptoms are produced (Ricaurte and McCann, 2005; Wu et al., 2006; Van Sassenbroeck et al., 2007; Munir et al., 2008). A medical doctor can find it difficult to reach a correct diagnosis based on symptoms and oral history alone when dealing with possible polydosing (Bjornaas et al., 2006; Liechti et al., 2006; Devlin and Henry, 2008).

A hypothetical scenario that illustrates the need for both field and gold-standard testing is the following. A disheveled and comatose woman is brought into a hospital emergency department by a friend who does not know what happened. Staff members are aware that recent cases of sexual assault in the city involved date-rape drugs, and they want to test immediately for candidate drugs and ethanol. Rapid discovery of specific agents has many benefits. A correct diagnosis would be more likely, and appropriate treatment could be selected quickly. Expensive and unnecessary tests for head trauma, metabolic disorders, central nervous system infection, and increased intracranial pressure might be avoided. Also, physiological samples could be collected and preserved for gold-standard testing if warranted: for example, when sexual assault was suspected (Zvosec et al., 2007). Field and gold-standard testing are not competitive; they are complementary.

Rapid tests for ethanol are available, of course. Field tests based on antibody recognition of MDMA, ketamine, flunitrazepam, and other date-rape drugs are also available. They require 10 to 20 min to complete (Kunsman et al., 1996; Salamone and Tsai, 2002; Negrusz et al., 2005), and many suppliers are available (http://www.varianinc.com/cgi-bin/nav?products/dat/index&cid=KJOQJOKKFJ; http://www.biocompare.com/jump/2831/Immunoassays.html). Antibody test kits typically are cross-reactive for the major metabolites of a drug. The ketamine kit, for example, picks up both ketamine and norketamine. Cross-reactivity for the metabolites extends the window of time available for detection of dosing, as metabolites can linger in persons longer than the original drug does.

Unfortunately, no reliable field test for GHB is currently available. Purported tests exist, but they are very insensitive, based on unreliable or poorly selective chemistry, or require unrealistically ideal conditions. They should not be used by a responsible principal (Meyers and Almirall, 2004; Grossman et al., 2005; Beynon et al., 2006; Child and Child, 2007; Quest and Horsley, 2007). A sensitive and reliable field test for GHB is described in Section 11.20.3.

11.17.6 Preparation of a Sample for Delayed Analysis

If a physiological fluid is to be tested for a date-rape drug, it should be obtained as soon as possible after a suspected incident, as any drug will be subject to metabolism and excretion (Kastrissios et al., 2005; Drummer, 2007). However, for many reasons that include short-term amnesia caused by the drug, victims often do not report suspected assault until one or more days after an incident. A victim usually will have ingested only a single dose of the drug. The combination of a single dose and significant delay before the report greatly increases the sensitivity required to detect dosing. If testing is to be done days to months after sampling, such as often occurs for gold-standard analysis, the sample must be stored properly to avoid degradation, sequestration, or artifactual generation of analyte (El Mahjoub and Staub, 2000; Clauwaert et al., 2001; Hijazi et al., 2001; Elliott et al., 2004; Holmgren et al., 2004; Berankova et al., 2006; LeBeau et al., 2007; Saude and Sykes, 2007). When a physiological fluid is prepared properly and stored no more than several days in a refrigerator, large errors generally do not occur. However, "proper" depends on both the matrix and analyte. Also, enzymatic treatment to release covalently conjugated analyte can be important for some types of analytes in order to increase sensitivity. The ability of staff members to properly prepare a physiological fluid for gold-standard analysis would benefit from the knowledge gained in a positive field test.

11.17.7 Time Window Available to Detect Dosing

The window of time available to detect dosing depends on the drug. After single effective doses, MDMA, flunitrazepam, and ketamine are routinely detectable in urine for 1.5 to 4 days. Extremely sensitive gold-standard detection methods can extend the time window significantly. Also, metabolites often are detectable for even longer (Verstraete, 2004; Adamowicz and Kala, 2005).

In contrast, single doses of GHB are detectable above the threshold values of 5 and 10 μg/mL in serum or urine, respectively, for as little as 3 h and as long as 12 h in different persons (Section 11.17.2; Palatini et al., 1993; Kavanagh et al., 2001; Borgen et al., 2003; Brenneisen et al., 2004; Haller et al., 2006; Bodson et al., 2008). The time window is quite variable in part because large doses can be detected for *disproportionately* longer time periods, as absorption, metabolism, and elimination of GHB are limited by physiological capacity (Palatini et al., 1993; Scharf et al., 1998; Bhattacharya and Boje, 2007; Odujebe et al., 2007; Jones et al., 2009). Also, metabolism rates are affected by diet and ethanol (Poldrugo et al., 1989; Kaufman and Nelson, 1991; Karch et al., 2001; Borgen et al., 2003; Haller et al., 2006; Thai et al., 2007; Wang et al., 2008). Time courses have been simulated numerically (Bhattacharya and Boje, 2006). In summary, physiological fluids must be sampled in $\leq$ 12 h of possible consumption in order to detect GHB above threshold. Merely increasing the sensitivity of testing does not solve the problem of a limited time window for GHB.

11.17.8 Extending the Time Window

Several relatively new gold-standard technologies extend the time available to detect dosing. One technology utilizes hair, which incorporates a circulating drug when the strands are being formed. In this technique, full-length hairs are obtained from a person, cleaned, and cut into 3-mm segments 4 to 5 weeks after putative dosing. The segments are dissolved, and the solubilized hair is analyzed by tandem mass spectrometry. If the person was dosed, the protocol produces a peak in drug concentration for the hair segment formed when the dosing occurred, and baseline values for hair segments that flank the segment that was formed during dosing. A single GHB dosing of a human being has been detected in this manner (Kintz, 2007).

Another applicable technology is *gene expression profiling* by DNA microarray analysis. Blood mRNAs for the proteins epiregulin (member of the epidermal growth factor family) and PEA-15 (phosphoprotein enriched in astrocytes of 15 kDa) were significantly elevated 48 h after mice were given a single large dose of GHB (Larson et al., 2007). Also, the Fos transcription factor is greatly elevated in rat brain after a high dose of GHB, although samples of brain normally would not be tested in living human beings (van Nieuwenhuijzen et al. 2009). Because Fos is widely distributed, Fos mRNA in the periphery is also probably elevated, although this hypothesis has not been tested. The effects of MDMA and ketamine on mRNA levels in rodent brain have been studied by gene expression profiling (Thiriet et al., 2002; Funada et al., 2005; Lowe et al., 2007). The method has great promise for forensic science. However, there remain the hurdles of (1) identifying suitable sources of mRNA, (2) identifying at least one mRNA that is diagnostic for a target drug, (3) establishing baseline values for mRNAs in the person tested, and (4) reducing statistical variation enough to make results accurate.

Gene expression profiling illustrates one reason why detailed knowledge of metabolic pathways might be valuable to forensic science. The RNAs coding for or regulating the expression of isoceptors, isozymes, and isoporters linked to a target drug are the most likely to respond to dosing. For example, mRNAs for MCTs (Section 6.1) and Fos in white blood cells are available from fresh blood, and one of them might be responsive to dosing by GHB.

11.18 SPECIAL CIRCUMSTANCES OF GHB

11.18.1 Industrial Connection

GHB differs from other date-rape drugs in one more important respect. Because of their rapid and efficient conversion to GHB in the body, GBL and BD have psychotropic effects very similar to those of GHB. Yet because GBL and BD are used by industry for multiple legitimate purposes, they are inexpensive, widely available, and weakly controlled. Chemical mixtures containing $\leq 70\%$ GBL recently were exempted (http://openregs.com/regulations/view/103806/exempt_chemical_mixtures_containing_gamma-butyrolactone). As a result, GBL and BD increasingly are being adopted as alternatives to GHB in drug-assisted sexual assault and self-abuse (Kapadia et al., 2007; Thai et al., 2007; Wood et al., 2008). The rapid test for GHB described below will not detect GBL or BD. However, it will flag their ingestion by detecting excess GHB in serum or urine.

11.18.2 Enzymes Acting on GHB in Bacteria, Yeast, and Plants

Many physiologically stressed bacteria and yeast biosynthesize polymers containing GHB. The polymers are a carbon- and energy-storage material (Moore et al., 2005; Song et al., 2005, Bach et al., 2009).[36] Polymerization requires activation of the GHB carboxyl group, which can be done by butyrate kinase, 4-hydroxybutyrate-CoA transferase, or 4-hydroxybutyryl-CoA synthetase (AMP forming) (Cary et al., 1988; Valentin et al., 2000; Huber et al., 2008). When environmental conditions become favorable to growth, microorganisms hydrolyze stored polymer to mobilize the carbon. Some species of bacteria also consume tetrahydrofuran, which is an environmental pollutant and a common solvent, by oxidizing it to GBL and then hydrolyzing the GBL to GHB (Kohlweyer et al., 2000). In all characterized cases, GHB is brought into central metabolism by an Fe(II)-containing ADH that uses dissociable NAD^+ as the hydride acceptor. These enzymes have sequences similar to human ADHFe1 (Section 11.7.2). Also, some plants use GHB as a redox ballast to survive periods of anoxia, and they make it with SSA reductase (Breitkreuz et al., 2003; Hoover et al., 2007). Only one isozyme is known for each of these enzymes.

11.18.3 Possible Accidental Intoxication by GHB in the Future

The possibility of accidental intoxication by GHB is increasing from an ironic source. Researchers have engineered microorganisms and plants to make large amounts of polymers containing GHB and/or BD monomers. Such polymers can resemble polyethylene and polypropylene in their rheological properties and are termed *bioplastics* (Van der Walle et al., 2001). Bionolle and Ecoflex are commercial examples (Gross and Kalra, 2002).[37]

Although currently minor, the fraction of plastics containing GHB and/or BD monomers could increase dramatically. The possibility arises because "synbio" organisms can make them from renewable low-value inputs such as corn stover and sewage sludge (Zhang and Wang, 1994; Tsunemasa, 1998; Poirier, 2001; Rabetafika et al., 2006). Bioplastics require less input of energy and valuable feedstocks than petroleum-based plastics do (Kurdikar et al., 2000; Efe et al., 2008). Moreover, bioplastics are hydrophilic polyesters that are biodegradable by esterases found in soil-dwelling microorganisms. Some bioplastics are even compostable, meaning that they are degraded completely in only a few months (Tokiwa and Pranamuda, 2001; Deguchi et al., 2002; Abou-Zeid et al., 2004). Composting reduces the volume of waste that must be accommodated by landfills, which are increasingly scarce (Davis and Song, 2006). The combined virtues of renewability and biodegradability might encourage much greater use of bioplastics containing GHB and/or BD in the future (Dornburg et al., 2004; Okino, 2005a,b; Kale et al., 2007; http://www.bioplastics24.com/).

If disposal facilities were to contain large amounts of bioplastics, leachate from them will contain monomers, including GHB and BD. The concentrations in primary leachate will probably be low and nontoxic, assuming that both the bioplastics and drainage meet professional specifications (http://www.astm.org/Standard/index.shtml). However, concentrations could increase to intoxicating levels if leachate were allowed to form puddles that evaporate, as GHB and BD are not very volatile and are chemically stable. Bacteria in the puddles might not efficiently scavenge GHB and BD, due to lack of micronutrients or preferential use of better carbon sources, such as glucose derived from plasticized starch also present in the disposal facility (Greer, 2006). This scenario is most likely to occur around residential compost piles. Dosing of neighborhood wildlife and omnivorous children by decomposing bioplastics might become a significant problem, thanks to green products. Regular environmental monitoring for GHB and BD could become a necessity.

11.19 CONSIDERATIONS DURING DEVELOPMENT OF FIELD TESTS

11.19.1 Shortcomings of Antibody-Based Screens for Simple Analytes

Many field tests for date-rape drugs are based on binding of free drug and its metabolites to an antibody preparation (Singh et al., 2006). Although the method works well for complex synthetic molecules having multiple points of contact between the analyte and the antibody combining site, it often does not work for simple analytes.[38] The reasons are easy to identify. To make a small molecule immunogenic, it must be covalently "conjugated" to a large carrier such as a protein. Conjugation in this case causes a big perturbation in analyte structure. Thus, antibodies often bind to the conjugated analyte better than to the free analyte (Benacerraf and Paul, 1970). Moreover, most antibodies used in forensic science are polyclonal, which means that preparation begins with the blood obtained from an immunized animal. Many different amino acid sequences and three-dimensional structures are present in the antibody combining sites. Polyclonal heterogeneity has adverse consequences on analyte selectivity, as discussed below.

Another type of antibody is monoclonal, which means that the preparation is obtained from an amplified clonal line of an antibody-producing cell. One amino acid sequence and one three-dimensional structure are present in all of the antibodies made by the clonal line. A different clonal line generated against the same analyte will in general secrete an antibody having a different amino acid sequence exhibiting a different three-dimensional structure in the antibody combining site. Thus, cocaine has been found by x-ray crystallography to bind in different conformations and orientations to different monoclonal antibodies (Pozharski et al., 2005). The result implies that an analyte binds to individual members of a polyclonal preparation in many different ways.

It follows that within the diversity of complexes that form with an analyte, a preparation of polyclonal antibodies can also bind many analogs of the analyte. From highest to lowest affinity, the potential complexes will include conjugated analyte, structural relatives of conjugated analyte, free analyte, and metabolites and other analogs of free analyte. Indeed, the ability of a polyclonal preparation to recognize a drug *and* its metabolites depends on antibody heterogeneity. Moreover, if a polyclonal preparation is purified with an immunoaffinity step, which will of necessity use conjugated analyte, the final preparation is biased even more away from binding free analyte and toward binding conjugated analyte (Choi et al., 1999). It is not surprising that complex mixtures of antibody combining sites exposed to complex mixtures of small molecules, *none* of which were the actual immunogen, often produce the worst possible outcome for a small-molecule screen: namely, a high rate of false positives and negatives.

11.19.2 Advantages of Enzyme-Based Screens for Simple Natural Analytes

Enzymes can provide a solution to the shortcomings of antibodies in the detection of simple analytes. First, there is no need to conjugate the analyte covalently. Thus, misdirection toward analyte conjugates and their analogs does not happen. Second, an enzyme has a single amino acid sequence and three-dimensional structure, thereby eliminating binding-site heterogeneity that decreases substrate selectivity. Third, most enzymes have evolved to discriminate against metabolites similar in structure to the substrate desired. Such discrimination is important to minimize crosstalk between pathways (Hartl et al., 1985; Hartl, 1989; Galperin et al., 1998).[39] In contrast, immunization has no mechanism to discriminate against metabolites resembling the conjugated analyte. Fourth, substrate bound in an enzymatic active site must form the near attack conformation (NAC) before it can react (Bruice, 2002). The NAC is accessed by librational dynamics, which are optimized by evolution. The NAC theory is complementary to the better known theory that an enzyme binds to the transition state for a reaction tighter than to the ground state. In both theories, protein dynamics contribute to selectivity in catalysis. In contrast, protein dynamics play no consequential role in an ordinary antibody–analyte complex. In summary, enzymes possess numerous mechanisms for molecular discrimination that antibodies in general and polyclonal antibodies in particular do not have.

Enzyme-based screens also have a time advantage that can be important in field applications. Positive results start to become known within seconds of mixing reagents and a test sample when a high concentration of analyte is present. In contrast, antibody-based screens require some sort of previsualization processing such as lateral flow chromatography or blocking of nonspecific protein adsorption to suppress background before the results can be read. An immunochemical method such as enzyme-linked immunosorbant assay (ELISA), which uses an enzyme to amplify the signal arising from analyte, does not negate the requirement for previsualization processing. In addition, enzymes often have a cost advantage. The complementary DNA sequence for an enzyme is cloned into an expression vector only one time, and the vector is used to transform bacteria only one time. Transformed bacteria can be propagated indefinitely and used to produce the enzyme in a genetically "tagged" form established by the vector. The tag is a polypeptide or protein chosen to allow facile purification of the resulting "fusion protein" by affinity chromatography at moderate cost. In contrast, polyclonal antibodies usually are obtained from immunized animals in a relatively time-consuming and expensive process. Monoclonal antibodies are even more expensive to produce.

The preceding observations suggest rules of thumb for choosing between an antibody and an enzyme for the development of a field test. An enzyme should be considered for a small natural metabolite, as a cognate enzyme *must* exist and an antibody has a high probability of failure. GHB is an example of this type of analyte. To avoid false positives the enzyme must be highly selective for analyte in the samples to be tested. Because of low selectivity, an enzyme that acts on xenobiotics usually would not be appropriate. Also, some readily implemented method must exist to visualize test results outside a laboratory.

An antibody should be considered for a structurally complex or large analyte, as the problem of immunological perturbation by conjugation is lessened. If the analyte is manufactured by synthetic organic chemistry, an enzyme selective for it will not exist in nature. MDMA, flunitrazepam, and ketamine are examples of this type of analyte. Also, even if the analyte is natural, the cloning, expression, and isolation of a cognate enzyme might not be practical, or the enzymatic reaction might not be easy to visualize in the field. An antibody might then be the best recourse. Cocaine is an example of this type of analyte. The pathway for its biosynthesis is not known with certainty, and thus an enzyme possibly having good selectivity for cocaine cannot be identified with certainty. Human carboxylesterase 1 and cytochrome P450 act on cocaine, but they are low-selectivity enzymes targeting many xenobiotic compounds (Ladona et al., 2000; Humphrey and O'Hagan, 2001; Bencharit et al., 2003).

Of course, no rule of thumb is infallible. Some analytes are large natural products, and new tools such as diabodies, aptamers, and directed evolution are being developed continually in the search for new types of biosensors (Fickert et al., 2006; Poyot et al., 2007).

11.20 DEVELOPMENT OF AN ENZYMATIC TEST FOR GHB

11.20.1 Sensitivity Required for the Hospital Emergency Department

Patients presenting to a hospital emergency department because of possible drug overdose are in crisis (Chin et al., 1998). Only moderate sensitivity is required to detect a drug if it is a major cause of the distress, as the amount in the patient's system *at that moment* must be large. In the case of GHB, levels of 30 to 600 μg GHB/mL of serum and 400 to 6000 μg GHB/mL of urine have been found by gold-standard analyses of samples that have been prepared and stored correctly (Louagie et al., 1997; Sporer et al., 2003; Couper et al., 2004; Elliott, 2004; Van Sassenbroeck et al., 2007). These levels are well within the capabilities of typical enzyme-based tests. *In the potentially life-and-death context of a hospital emergency department, a rapid test with good sensitivity is better than a slow test with superior sensitivity*.

11.20.2 Choice of Enzyme

As footnoted above, attempts to develop a field test based on antibody recognition of GHB have failed. An enzyme-based test is an attractive alternative. Three enzymes that act on GHB have been characterized well enough to consider. They are vertebrate NADPH-dependent SSA reductase (AKR1A1 or AKR7A2), vertebrate 2-oxoglutarate–dependent ADHFe1, and bacterial Fe(II)-containing NAD$^+$-dependent GHB dehydrogenase (Sections 11.5.3, 11.7.2, and 11.18.2). To use a reductase, the reaction would have to be run in the reverse direction using NADP$^+$ as oxidant.

What considerations can help us decide which of these enzymes might be suitable? The oxidation catalyzed by ADHFe1 is thermodynamically favorable, but ADHFe1 has only moderate selectivity for GHB among common metabolites. It would be susceptible to false-positive results when testing physiological fluids and other complex matrices. This possibility is decisively unfavorable, leaving vertebrate SSA reductase and bacterial ADH to consider. The overall equilibrium constants are the same, providing no thermodynamic basis to choose between them (Section 11.9). A search of online databases containing DNA and protein sequences reveals many putative SSA reductases and GHB dehydrogenases, as deduced by sequence homology (http://www.ncbi.nlm.nih.gov/; http://www.uniprot.org/). However, sequence homology is much better at predicting the chemistry carried out than the preferred substrate (Hulo et al., 2006). Unless substrate selectivity has been studied by direct biochemical methods, one cannot be sure which substrate is preferred.

A search of online databases for enzymatic substrates, followed by thorough reading of the primary literature, reveals that only a few SSA reductases and GHB dehydrogenases have been characterized biochemically. Examples of these databases are BRENDA (http://www.brenda-enzymes.info/) and KEGG (http://www.genome.ad.jp/kegg/pathway.html). Abstracts of primary literature are freely accessible in PubMed (http://www.ncbi.nlm.nih.gov/) or by subscription to SciFinder Scholar (http://www.cas.org/products/sfacad/index.html). Most SSA reductases have high affinity for SSA and low affinity for GHB, which promotes fast and complete reduction of moderate concentrations of SSA but slow and incomplete oxidation of moderate concentrations of GHB, a decidedly unfavorable property (Keenan et al., 2006). This characteristic narrows the choice of enzyme to bacterial GHB dehydrogenase. Moreover, because a single isozyme is likely, and cloning and expression are usually easier, a bacterial enzyme is preferred to an enzyme from a eukaryotic organism such as an animal, plant, or yeast. The air-stable Fe(II)-containing NAD^+-dependent GHB dehydrogenase from the bacterium *Ralstonia eutropha* meets all criteria (Valentin et al., 1995). The properties of this family of enzymes have been summarized (http://pfam.sanger.ac.uk/family?acc=PF00465). GHB dehydrogenase from the bacterium *Clostridium kluyveri* has also been characterized biochemically. However, it is unacceptable because it is rapidly inactivated by oxygen, which would be a major liability in a field test (Wolff and Kenealy, 1995).

The GHB dehydrogenase reaction is not visible to the unaided eye. This circumstance means that a secondary reaction must be coupled to the reaction to visualize the results of the test. Fortunately, the electron pair stored in NADH can be used by a well-known “turnkey” reaction that reduces a large number of colorless prodyes to intensely colored products (Berridge et al., 2005). All required reagents are commercially available. Such a coupled secondary reaction also helps solve another problem. Oxidation of GHB with NAD^+ is significantly unfavorable at neutral pH (Section 9). Unless something is done to compensate, only a fraction of the GHB in a test sample will react, and it will do so slowly because of the asymptotic approach to equilibrium. Both properties are decidedly undesirable. However, the turnkey reactions are so favorable that the resulting two-step reaction is favorable overall.

11.20.3 Reliable Field Test for GHB

To develop the test, the gene for GHB dehydrogenase in *R. eutropha* was copied from total genomic DNA using appropriate primers and polymerase chain reaction. The resulting DNA was purified, end-trimmed with restriction enzymes, and fused to the gene for the affinity-tag glutathione *S*-transferase, which is under the control of a lactose promoter in the commercial expression vector pGEX-2T. *E. coli* was transformed with the recombinant vector to produce and isolate large quantities of fusion protein that links glutathione *S*-transferase to GHB dehydrogenase. About 150 mg of purified fusion protein typically is obtained from 1 L of bacterial culture in a fast procedure utilizing an affinity matrix containing covalently bound glutathione (Bravo et al., 2004).

A solution assay at pH 8.5 in 2-amino-2-methyl-1,3-propanediol buffer was devised for quantitative research using a laboratory spectrometer. The higher pH value decreases the concentration of product H^+ and makes the reaction more favorable (Fig. 5). The buffer is Tris-like, resulting in formation of the imine of SSA as it is formed. Formation of imine (Section 11.8) pulls GHB oxidation thermodynamically and prevents product inhibition due to SSA buildup. The NADH produced by the GHB dehydrogenase reaction is used to reduce (with the

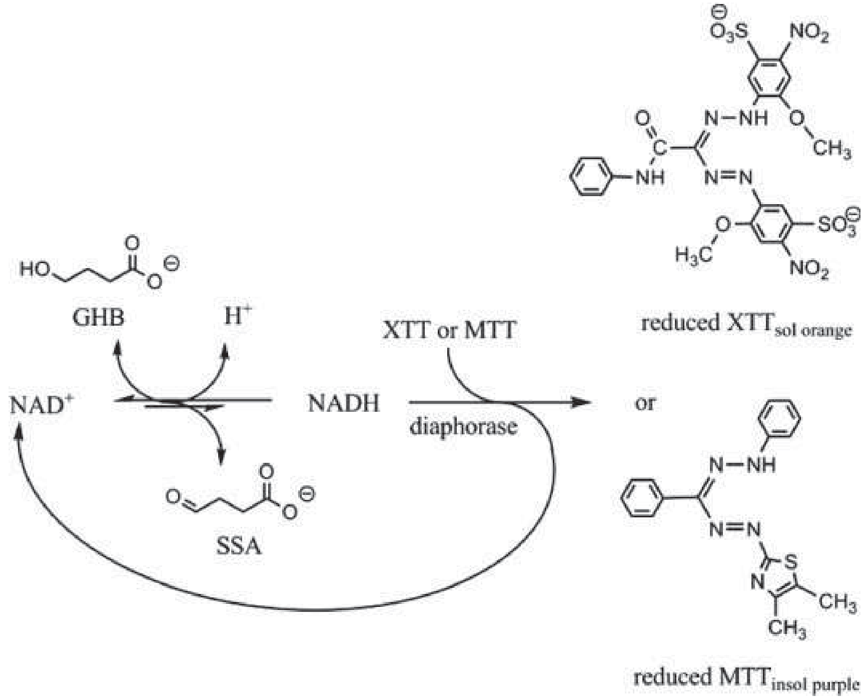

Figure 11.5 Coupled assay for GHB. NAD$^+$ and GHB dehydrogenase from *Ralstonia eutropha* oxidize GHB to SSA in the first step. Because it is otherwise unfavorable, the reaction is run under pulling conditions. A Tris-like buffer (2-amino-2-methyl-1,3-propanediol) that also forms the imine of SSA is utilized at pH 8.5. In addition, diaphorase uses NADH made in the first step to reduce nearly colorless XTT to a soluble orange dye for a laboratory assay using a spectrometer or MTT to an insoluble purple dye deposited on a dipstick for a field assay. Different prodyes that yield different colors and solubility are available.

aid of the electron transfer enzyme diaphorase) a prodye called XTT [sodium 3,3-[(phenylamino)carbonyl]-3,4-tetrazolium-bis(4-methoxy-6-nitro)benzenesulfonic acid hydrate] to form a soluble orange dye (λ_{max} 450 nm) and regenerate NAD$^+$ (Figure 11.5). This reaction also pulls GHB oxidation thermodynamically and removes product inhibition due to NADH buildup. Thus, the assay implements five strategies to make it favorable, faster, and visible.

Michaelis–Menten parameters were determined in 1 mM NAD$^+$. The K_M value for GHB is 2.19 ± 0.13 mM and the V_{max} value is 0.049 ± 0.001 absorbance/min per microgram of fusion protein. The lowest concentration of GHB reported in the urine of emergency department patients (~400 μg/mL) corresponds to about 4 mM. The K_M value is well matched to this concentration, thus providing a rapid and sensitive detection of GHB.

Ethanol is a weakly binding alternative substrate of the enzyme, having an estimated K_M value of 413 ± 14 mM, or 200 times higher than for GHB. The affinity is low enough that ethanol concentrations realistically found in physiological fluids do not produce false-positive results. Recent research has shown that other simple alcohols, such as methanol, 1-propanol, 2-propanol, 1-butanol, and 1-pentanol, are significantly worse substrates than ethanol is (Vinckier et al., 2011).

Multiple studies agree that GHB arising from ingested GHB or GHB prodrug is concentrated about 10-fold in urine relative to blood (Liu et al., 2006). Thus, urine rather than serum is preferred for dose testing. If allowed to react for about 2 h in 1 mL of coupled assay buffer, all of the GHB in a small sample of spiked normal human urine (10 μL) is reacted. The amount of orange color produced is linearly dependent on the amount of GHB. No substances in urine contribute significantly to color development, even after 2 h of reaction. This variation of the test, called an *endpoint assay*, has a limit of detection of 37 μg GHB/mL urine (total A_{450} to background $A_{450} = 2$ to 1). The same procedure using urine containing 0.47% (w/v) ethanol gives the same results, confirming that a realistic concentration of ethanol in urine does not interfere with the endpoint assay.

The solution assay uses a spectrophotometer and is not as fast or convenient as a dipstick assay that can be used in the field. To develop a prototype for a dipstick, formation of an insoluble purple dye by reduction of the soluble prodye MTT (3-[4,5-dimethylthiazol-2-yl]-2,5-diphenyltetrazolium bromide) was used (Figure 11.5). Small samples (10 μL) of water, normal human urine, or normal human urine containing 0.63% (w/v) ethanol and having different concentrations of spiked GHB were spotted onto paper, after which assay reagent (10 μL) was spotted. The test is developed for no longer than 2 min to avoid false-positive results. No significant color develops in the absence of GHB, and strong purple color develops within a few seconds for samples containing concentrations of GHB typically present in urine during an overdose. Color development is fast because both the test sample and reagents are diluted very little. The three tested matrices gave the same results, demonstrating that neither urine nor a physiologically high concentration of ethanol interfere (Figure 11.6). The precision around the cutoff was also assessed. Of 20 samples for each condition, none were positive in the absence of GHB, 25% were positive (and 75% were negative) at 20 μg GHB/mL, 75% were positive at 50 μg GHB/mL, and all were positive at 100 μg GHB/mL and more. Thus, a concentration of GHB four times lower than the lowest concentration found in the urine of patients suffering from GHB overdose is detected every time. A dipstick test strip not requiring spotting of reagents before use is under development for long shelf life and use in hospital emergency departments (Harper et al., 2008).

CONCLUSION

Impressive progress in understanding the fundamental molecular and cellular biology of date-rape drugs has been made in recent years. Nevertheless, many important questions remain. The roles of RNA and regulation of enzymatic activity and gene expression have been studied only a little. For example, does protein phosphorylation regulate the metabolism of date-rape drugs? Use of cultured human cells

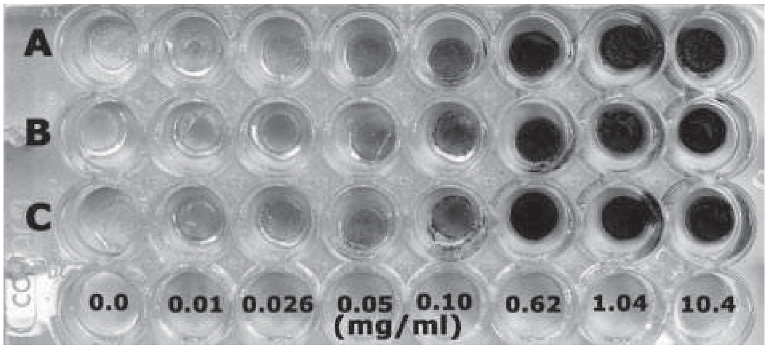

Figure 11.6 Prototype of a dipstick assay for GHB. Ten-microliter samples containing the concentrations of GHB indicated were spotted onto white-paper circles lying in wells of a microtiter plate. Ten microliters of assay reagent containing buffer, GHB dehydrogenase, NAD$^+$, diaphorase, and MTT prodye was spotted onto a circle of filter paper in each well, and color was allowed to develop for 2 min at 23° C. Row A, GHB in water; row B, GHB in normal human urine; row C, GHB in normal human urine containing 0.63% (w/v) ethanol. [From Bravo et al. (2004); used with permission.] *(See insert for color representation.)*

and genetically modified mice should be especially effective tools to answer these questions. In the case of GHB and its prodrugs, the cellular, subcellular, and enzymatic mechanisms for catabolism are not fully understood. Even when a particular isozyme or isoporter has been implicated in GHB catabolism (e.g., mitochondrial ADHFe1), we do not know whether it dominates other isozymes or isoporters under all relevant conditions in all relevant cell types, and whether it is rate limiting to the pathway. Neuronal signaling by GHB under normal and overdose conditions is not understood. Thus, potential consequences of GHB overdose that manifest in subtle, delayed, or synergistic damage have not been identified.

Impressive progress in the forensic science of date-rape drugs has also been made in recent years. Especially sensitive analytical methods have been developed. A surrogate for drugs and their metabolites in the form of perturbed mRNA levels has been discovered. The window of time available to detect dosing has been greatly extended. In the case of GHB, new metabolites have been found that are potentially diagnostic, and more probably remain to be found. An enzyme-based test for GHB has been developed that fills a gap in field tests for date-rape drugs. Moreover, the demand for forensic expertise in date-rape drugs will probably increase in the future due to (1) the possible need for comparison of two uncorrelated test methods in judicial proceedings, (2) development of new tests based on increased knowledge of biology, and (3) greater need for environmental testing. It is a rewarding and exciting time to be working in forensic science!

Acknowledgments

The author thanks fellow researchers David Harris and Dawn Bravo for their insights and hard work, Lawrence Kobilinsky for his invitation to contribute to this book, Karl Sporer, Elizabeth Ellis, Toshimitsu Niwa, and Aaron Rowe for comments on the manuscript draft, Zachary Golden for alerting me to new statuatory law, Don Aue for ab initio calculations of stabilities for GHB metabolites in water, Tetsuya Kawamura for translation services, and Leslie D. Odell for extraordinary editorial assistance.

NOTES

1. GHB also is known as Xyrem and "grievous bodily harm"; MDMA as ecstasy; flunitrazepam as Rohypnol and roofies; and ketamine as Ketanest and special K.
2. Synaptic vesicles are small and spherical or ellipsoidal and reside inside nerve terminals. They store concentrated neurotransmitter for evoked release as a quantum.
3. The central and peripheral nervous systems can synthesize neurosteroids at higher levels in females than in males. This difference might help explain why women often are more susceptible to ethanol at equivalent body burdens. An example of a neurosteroid is allopregnanolone.
4. Subtypes of GABA_A receptor located inside synapses are not as responsive to ethanol as the subtype located outside synapses, which contains the δ-subunit.
5. The GABA_B receptor modulates intracellular second messengers. An agonist activates a receptor. In this case, a high concentration of GHB activates the GABA_B receptor similarly to a low concentration of GABA.
6. GABAergic means GABA-like or GABA-potentiating.
7. All neurotransmitters are agonists, even if transported instead of secreted out of a neuron.
8. An antagonist does not activate a receptor when it binds, but it inhibits binding of agonist, thus preventing activation by agonist.
9. Desensitization means that the receptor becomes unresponsive to agonist, even though no antagonist is present. Many different molecular and cellular mechanisms can mediate desensitization by receptors.
10. Isozymes (also known as *paralogs*) usually, but not always, have partially conserved sequences. The following examples illustrate nomenclature for the dehydrogenases and reductases discussed here. AKR1A1 is listing 1 (the last "1") in family 1 (the first "1", having $\geq$ 40% sequence identity among members) and subfamily A ($\geq$60% sequence identity among members) of the AKR superfamily. This enzyme is also commonly called aldehyde reductase, as it reduces a wide range of aldehydes. Similar nomenclature is used for ADHs and ALDHs by replacing "AKR" with "ADH" or "ALDH," as appropriate. AKR1A1 has no necessary relationship to ADH1A1. An isozyme could have several isoforms that arise from alternative splicing of the pre-mRNA.
11. This occurs because oxidation of an aldehyde to a carboxylate is irreversible. The reduction potential of an NAD^+/NADH or NADP^+/NADPH couple is not very important in an irreversible oxidation. A few mammalian alcohol dehydrogenases use non-nicotinamide redox cofactors such as flavins.
12. For example, SSA is a good substrate for AKR1A1 *and* AKR7A2 (Section 11.5.3). AKR1A1 also prefers the natural substrates 4-hydroxynonenal, methylglyoxal, 16-oxoestrone, D-glyceraldehyde, and D-glucuronic acid. AKR7A2 also prefers the natural substrates aflatoxin B1 aldehyde and 16-oxoestrone.

13. When two enzymes act on the same substrate, one of them can dominate at a low concentration of substrate and the other one at a high concentration if the Michaelis–Menten parameters for the two enzymes (and potentially also for an intervening transporter) have appropriate relationships. This possibility is relevant for GHB and its metabolites, as concentrations in vivo change > 100-fold between normal and dosed conditions.
14. Similar to isozymes, isoporters are transporters coded by different genes, and they often exhibit conservation of sequence and transport mechanism. Different isoporters usually have different preferred substrates.
15. However, mRNA levels often do not correlate quantitatively with protein levels, due in part to regulation of translation and differences in protein stability.
16. Amazingly, both SSA and aflatoxin B1 are preferred substrates for AKR7A2/AFAR1, even though the structures of SSA and aflatoxin B1 differ greatly. The observation warns us that in the absence of experimental evidence, one cannot eliminate the possibility that an isozyme having a preferred substrate very different in structure from SSA nevertheless also prefers SSA.
17. AKR7A2 sometimes is called the low-K_M SSA reductase.
18. An ortholog is the same enzyme in a different species. It will have a somewhat different sequence, due to species divergence, but its functional properties are usually very similar.
19. DNA microarray analysis determines the amount of specific mRNAs in total mRNA isolated from a tissue. Single-stranded synthetic DNA is anchored to thousands of microscopic spots in an indexed two-dimensional pattern on the surface of a “chip.” The DNA on each spot is complementary to a different mRNA. A preparation of total mRNA (or its DNA complement) is hybridized to the chip-bound DNA, after which each spot optically reveals the amount of hybridization.
20. Epitope masking can occur when tissue fixatives react with and destroy an epitope or when the epitope is dependent on native conformation destroyed by denaturing electrophoresis, as occurs in Western blots.
21. Small interfering RNA (siRNA) blocks effective expression of a specific gene by inducing cellular degradation of the mRNA transcribed from that gene.
22. A low level of AKR1A1 in SH-SY5Y cells is not necessarily indicative of the amounts in vivo, as transformed cells are disregulated.
23. Golgi is a subcellular organelle that among other tasks forms synaptic vesicles that secrete neurotransmitters from nerve terminals.
24. Isoforms can have different amino acid sequences even though they arise from the same gene. This often occurs by alternative splicing of the mRNA. The total number of isoforms for ADHs, AKRs, and ALDHs thus exceeds the 46 genes for these enzymes by an unknown margin.
25. ADHFe1 is also known as HOT (hydroxyacid-oxoacid transhydrogenase) and ADH8. However, the latter designation has been used by some authors for retinol dehydrogenase, which is a completely different enzyme from ADHFe1.
26. Genetic tagging attaches a peptide to a protein to make a fusion protein that is to be expressed heterologously. In this case, an epitope tag recognized by a commercially available antibody was attached to ADHFe1.
27. Square brackets around a chemical name indicate the activity of the chemical species, which we approximate as molar concentration.

28. The ratio of the multiplicative concentrations of all chemical products divided by the multiplicative concentrations of all chemical reactants present at equilibrium is equal to K_{eq}. Thus, for ADH-catalyzed oxidation of an alcohol, $K_{\text{eq}} = [\text{aldehyde}]_{\text{eq}}[\text{NADH}]_{\text{eq}}[\text{H}^+]_{\text{eq}}/([\text{alcohol}]_{\text{eq}}[\text{NAD}^+]_{\text{eq}})$, which has net units of molar. The apparent value of K_{eq} for the reaction at pH 7.0 is given by $K'_{\text{eq}} = K_{\text{eq}}/[10^{-7.0}\text{M}]$, which is unitless. Similar equations apply to other reactions considered in this chapter.
29. A reported equilibrium constant of $K'_{\text{eq}} = 5.8$ for reduction of SSA must be in error (Hearl and Churchich, 1985).
30. At the pH and $[\text{NAD}^+]/[\text{NADH}]$ prevailing in mitochondria, $[\text{aldehyde}]_{\text{eq}}/[\text{alcohol}]_{\text{eq}} = K_{\text{ox}}^{\text{mito}} = K'_{\text{eq}}([10^{-7.0}\text{M}]/[10^{-7.7}\text{M}])([\text{NAD}^+]_{\text{mito}}/[\text{NADH}]_{\text{mito}}) = (9.5 \times 10^{-5})(10^{0.7})(7.7) = 3.7 \times 10^{-3}$. The free-energy change for the reaction is given by $\Delta G_{\text{ox}}^{\text{mito}} = -2.3RT \log K_{\text{ox}}^{\text{mito}}$, where the symbols have their usual meanings. Analogous expressions apply to NADP^+-dependent reactions, reverse reactions, and cytoplasm.
31. $K_{\text{ox}}^{\text{cyto}}$, equilibrium constant for oxidation of GHB modified for the pH and ratio of oxidant to reductant prevailing in cytosol; $K_{\text{red}}^{\text{cyto}}$, equilibrium constant for reduction of SSA modified for the pH and ratio of reductant to oxidant prevailing in cytosol; $K_{\text{ox}}^{\text{mito}}$, equilibrium constant for oxidation of GHB modified for the pH and ratio of oxidant to reductant prevailing in mitochondrial matrix; $K_{\text{red}}^{\text{mito}}$, equilibrium constant for reduction of SSA modified for the pH and ratio of reductant to oxidant prevailing in mitochondrial matrix.
32. Oxidation of an aldehyde to a carboxylic acid is irreversible in part because three driving forces are present that are not present in the oxidation of alcohols: neutralization of the carboxylic acid proton, electronic delocalization of the negatively charged carboxylate, and strong hydration of the carboxylate.
33. An allele is a mutant (or wild-type) form of a gene.
34. Microsomes are vesicles created from shearing of endoplasmic reticulum by homogenization of an organ such as liver, which is rich in endoplasmic reticulum. They are rich in P450 isozymes.
35. From this point onward in the chapter, any time that consequences of dosing by GHB are discussed, it should be assumed that similar, and perhaps even more extreme, consequences attach to dosing by GBL or BD.
36. Such species include *Ralstonia eutropha, Azotobacter vinelandii, Escherichia coli, Clostridium kluyveri*, and *Pseudomonas putida*. Poly(GHB) is also known as poly(γ-butyrolactone).
37. The trademark Bionolle is registered to Showa Highpolymer Co., Ltd. The chemical name is poly(butylene succinate). Butylene refers to 1,4-butanediol diester. Ecoflex is registered to BASF Group. The chemical name is poly(butylene adipate).
38. Several companies have attempted to raise antibodies that are selective for GHB, which is a small molecule in the immunogenic context, but they failed (personal communications to SMP).
39. Enzymes that act on xenobiotics are exceptions.

REFERENCES

Abanades, S., Farre, M., Barral, D., Torrens, M., Closas, N., Langohr, K., Pastor, A., and de la Torre, R. (2007a). Relative abuse liability of gamma-hydroxybutyric acid, flunitrazepam, and ethanol in club drug users. *J. Clin. Psychopharmacol*., **27**:625–638.

CHAPTER 12

Forensic and Clinical Issues in Alcohol Analysis

RICHARD STRIPP

Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

Summary Ethanol is a clear volatile liquid that is soluble in water and has a characteristic taste. Ethanol is a central nervous system (CNS) depressant and causes most of its effects on the body by depressing brain function. CNS depression correlates directly to the concentration of alcohol in the blood. The estimation of blood alcohol concentration for a person is based on important parameters such as body weight, ethanol concentration of the beverage consumed and number of such beverages consumed, and length of time and pattern of the drinking. Because men and women have different body water amounts (men average 68% and women 55%), there are differences between the ethanol concentration achieved in men and women of similar weight for the same amount of alcohol. Various methods are described that can help to determine blood alcohol concentration in the field as well as in the laboratory.

12.1 Introduction 436
12.2 Blood alcohol concentration 437
12.3 Alcohol impairment and driving skills 441
12.4 Field sobriety tests 443
12.5 Blood alcohol measurements 444
  12.5.1 Enzymatic methods 444
  12.5.2 Headspace gas chromatography 445
  12.5.3 Breath alcohol testing 446
  12.5.4 Breath alcohol instrumentation 447
  12.5.5 Conversion from BrAC to BAC 449
  12.5.6 Urine and saliva 450
  12.5.7 Ethyl glucuronide 450

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

12.5.8 Postmortem determination of alcohol 451
12.5.9 Quality assurance of alcohol testing 452
References 453

12.1 INTRODUCTION

Alcohol (Ethanol: CH_3CH_2OH) Disposition

Half-life: plasma half-life, dose dependent
Volume of distribution: about 0.6 L/kg
Distribution in blood: plasma/whole blood ratio = 1.2
Saliva: plasma/saliva ratio is about 0.93
Protein binding: in plasma, not significantly bound

Ethyl alcohol has been consumed by humans for millennia and is probably one of the oldest and most commonly consumed drugs in most societies. A very high percentage of the adult population regularly consumes some form of alcohol to varying degrees. Approximately 10% use alcohol to an extent that may result in the manifestation of adverse heath effects. People who regularly consume six or more drinks per day are known to have higher mortality rates, and alcohol contributes to deaths in a variety of ways, including violence, accidents, and other drug-related deaths. Acute ethanol intoxication and its effect is the single most commonly encountered subject in forensic toxicology today. The blood alcohol concentration is commonly determined in medicolegal and forensic cases, and it is the ethanol in blood that is correlated with impairment of human performance. Ethanol is a product of fermentation by the action that yeasts on sugars found in fruit and grains. The content of alcoholic beverages varies for different types of drinks (see Table 12.1).

Alcohols are a class of organic chemicals that are characterized by the presence of a hydroxyl group (OH) on the hydrocarbon chain of the molecule (Figure 12.1). They all share common properties and ethanol is one of many alcohols. Other commonly encountered alcohols are methanol and isopropanol. Ethanol is a clear volatile liquid that is soluble in water and has a characteristic taste. As a drug, ethanol falls into the category of a *central nervous system* (CNS) *depressant* (Garriott, 2003). This means that the predominant effects of alcohol are to depress brain function. CNS depression is directly correlated to the concentration of alcohol in the blood (BAC). In a progressive manner, as the BAC rises, so does the degree of CNS depression. The most common reported units for BAC (%) are expressed as grams of ethanol per 100 mL of blood. Therefore, a BAC of 0.05 g/dL is equivalent to 0.05%. The American Medical Association has defined the BAC where impairment begins to be 0.04%, but for some functions it may be much lower. In most states in the United States a person is legally intoxicated at a BAC of 0.08% and is considered to be impaired at 0.05%. The consequences of impairment are

TABLE 12.1 Alcohol Content of Various Beverages

Beverage	Alcohol Content (%)
Beers (lager)	3.2–4.0
Ales	4.5
Porter	6.0
Stout	6.0–8.0
Malt liquor	3.2–7.0
Sake	14.0–16.0
Table wines	7.1–14.0
Sparkling wines	8.0–14.0
Fortified wines	14.0–24.0
Brandies	40.0–43.0
Whiskies	40.0–75.0
Vodkas	40.0–50.0
Gin	40.0–48.5
Rum	40.0–95.0
Tequila	45.0–50.5

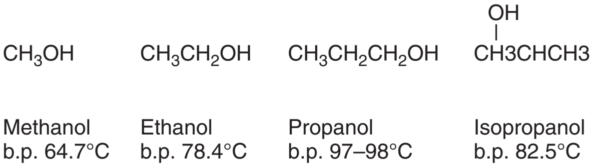

Figure 12.1 Structure of common alcohols.

the major reasons for the intense forensic interest in alcohol consumption. The individual assessment of impairment due to acute alcohol intoxication is of great medical–legal importance and is a complex matter that is influenced by many factors. There is significant variation between individual persons and their alcohol-related effects, but for the most part, predictable effects at various levels of BAC are encountered for a large percentage of the general population (Table 12.2).

It must be emphasized that a specific individual would show signs that can be markedly different for the ranges that are indicated. As already noted, many tolerant individuals may show little sign of physical impairment at the higher end of the table, although significant cognitive, judgmental, and psychological impairment may reasonably be presumed to be present.

12.2 BLOOD ALCOHOL CONCENTRATION

The estimation of blood alcohol concentration must be based on important parameters such as body weight, ETOH concentration of the beverage consumed

TABLE 12.2 Stages of Alcohol Intoxication

BAC (g/100 mL of blood or g/210 L of breath)	Stage	Clinical Symptoms
0.01–0.05	Subclinical	Behavior nearly normal by ordinary observation
0.03–0.12	Euphoria	Mild euphoria, sociability, talkativeness Increased self-confidence; decreased inhibitions Diminution of attention, judgment and control Beginning of sensory-motor impairment Loss of efficiency in finer performance tests
0.09–0.25	Excitement	Emotional instability; loss of critical judgment Impairment of perception, memory and comprehension Decreased sensatory response; increased reaction time Reduced visual acuity; peripheral vision and glare recovery Altered sensory-motor coordination; impaired balance Drowsiness
0.18–0.30	Confusion	Disorientation, mental confusion; dizziness Exaggerated emotional states Disturbances of vision and of perception of color, form, motion and dimensions Increased pain threshold Increased muscular loss in coordination; staggering gait; slurred speech Apathy, lethargy
0.25–0.40	Stupor	General inertia; approaching loss of motor functions Markedly decreased response to stimuli Marked muscular coordination impairment; inability to stand or walk Vomiting; incontinence Impaired consciousness; sleep or stupor
0.35–0.50	Coma	Complete unconsciousness Depressed or abolished reflexes Subnormal body temperature Incontinence Impairment of circulation and respiration Possible death
0.45+	Death	Death from respiratory arrest

(Table 12.3), number of alcoholic beverages consumed, and length and pattern of the drinking.

Other factors that will affect the rate of absorption from the gastrointestinal tract will further affect the final BAC curve (Morgan Jones and Vega, 1972; Karch, 1997; Drummer, 1999; Garriott, 2003). Approximation of the time frame of when the alcohol was consumed is usually of crucial importance (Morgan Jones and Vega, 1972; Karch, 1997). It is also essential to determine if the subject's BAC was on the ascending or descending limb of the BAC curve. The blood alcohol concentration reached after drinking depends not only on the volume of ethanol consumed but

TABLE 12.3 Estimated BAC by Number of Drinks in Relation to Body Weight[a]

Body Weight (lt)	Number of Drinks											
	1	2	3	4	5	6	7	8	9	10	11	12
100	.038	.075	.113	.150	.188	.225	.263	.300	.338	.375	.413	.450
110	.034	.066	.103	.137	.172	.207	.241	.275	.309	.344	.379	.412
120	.031	.063	.094	.125	.156	.188	.219	.250	.281	.313	.344	.375
130	.029	.058	.087	.116	.145	.174	.203	.232	.261	.290	.320	.348
140	.027	.054	.080	.107	.134	.161	.188	.214	.241	.268	.295	.321
150	.025	.050	.075	.100	.125	.151	.176	.201	.226	.251	.276	.301
160	.023	.047	.070	.094	.117	.141	.164	.188	.211	.234	.258	.281
170	.022	.045	.066	.088	.110	.132	.155	.178	.200	.221	.244	.265
180	.021	.042	.063	.083	.104	.125	.146	.167	.188	.208	.229	.250
190	.020	.040	.059	.079	.099	.119	.138	.158	.179	.198	.217	.237
200	.019	.038	.056	.075	.094	.113	.131	.150	.169	.188	.206	.225
210	.018	.036	.053	.071	.090	.107	.125	.143	.161	.179	.197	.215
220	.017	.034	.051	.068	.085	.102	.119	.136	.153	.170	.188	.205
230	.016	.032	.049	.065	.081	.098	.115	.130	.147	.163	.180	.196
240	.016	.031	.047	.063	.078	.094	.109	.125	.141	.156	.172	.188

[a] One standard drink equals: 12 oz regular beer, 5 oz table wine, 1 oz 80 proof liquor (National Highway Traffic Safety Administration (NHTSA)).

also on the speed and pattern of drinking and the rates of absorption, distribution, metabolism, and excretion (ADME) of ethanol, all of which in turn may be affected by a multitude of other factors (Drummer, 1999). Age, gender, race, body mass index, presence of other drugs, beverage type, food, genetics, and health status are a few examples of factors that can alter the ADME of ethanol and hence are important considerations when interpreting BAC levels in a person (Millar et al., 1992; Jones and Jonsson, 1994; Ammon et al., 1996; Karch, 1997; Drummer, 1999; Garriott, 2003). Once ethanol is absorbed it is distributed throughout the body via the blood to other tissues. Ethanol is distributed to the body water, and the greater the water contents of a tissue, the greater the ethanol concentration. The individual factors that determine BAC are body weight and adiposity, and distribution ratio (based on the total body water content) of ethanol. It follows that the concentration of alcohol within a person depends primarily on the amount of alcohol he or she drinks, the person's drinking pattern and total body water, and various other life and environmental issues.

Because men and women have different body water amounts (men average 68% and women 55%), there are differences between the ethanol concentration achieved in men and women of similar weight for the same amount of alcohol consumed. In other words, if a man and a woman of the same weight drink equivalent amounts of ethanol, the female will have a higher BAC on average. Similarly, if two men of different body weights drink the same amount of ethanol, the lighter man will have a higher BAC compared to the heavier man, and so on. In general, the less you weigh, the more you will be affected by a given amount of alcohol, all other factors assumed to be equal for this purpose.

The volume of distribution for ethanol was first expressed by Widmark using a formula that still bears his name (Garriott, 2003):

$$A = \frac{WRC}{0.8}$$

where A is the volume of pure ethanol consumed, W the body weight, R the distribution ratio, C the blood ethanol concentration, and 0.8 the specific gravity of ethanol.

There is a time lag from the period of intake of alcohol until the peak blood alcohol concentration is reached. This is regulated by the rate of alcohol absorption after oral intake. As described above, alcohol absorption is affected by numerous factors. Because of these factors, the time to reach peak blood alcohol concentrations varies greatly and can range between 15 min and 2 h. Hence, if alcohol absorption is not complete, peak alcohol levels will not yet be reached. The BAC is a balance between absorption and elimination of ethanol (Winek and Esposito, 1985).

There can be a threefold variation in ethanol elimination rates among individuals. Gender, age, body weight, and even the time of day can influence the rate of ethanol metabolism. To back-extrapolate from a blood alcohol concentration at one time to a value at an earlier time, one needs to know an accurate value for the ethanol elimination rate, which is difficult in view of the large interindividual variability in ethanol elimination kinetics. Also, to back-extrapolate, one uses linear kinetics, which may not be accurate, especially for the concentration range under consideration. The variability in individual rates of ethanol elimination, the difficulty in knowing exactly when absorption of ethanol was completed and how rates of ethanol elimination can change at different blood ethanol concentrations, cast some uncertainty on such retrograde calculations to an earlier blood ethanol concentration.

Since BAC is a function of the balance between how much is entering the blood versus how much is cleared or removed, the liver, via metabolism, plays a very important role. The liver removes most of the ethanol via oxidation, with lesser amounts excreted in the breath, urine, feces, and saliva or converted to ethyl glucuronide (Figure 12.2). The cytochrome P450 enzymes in liver microsomes catalyze the conversion of ethanol to acetaldehyde and water in the presence of molecular oxygen and NADPH (Figure 12.3). The toxic acetaldehyde product is quickly converted to acetic acid via acetaldehyde dehydrogenase, as shown in Fig. 2. The rate at which ethanol is cleared from the blood varies from one person to another. Heavy drinkers and alcoholics tend to eliminate alcohol much faster than does the inexperienced drinker. In most people the rate of clearance from blood ranges from 0.01 to 0.025% per hour, and is even higher in some very tolerant drinkers, with a mean of about 0.017%. This is very important for the forensic toxicologist.

Assume, for example, that an accident occurs at midnight and the alcohol content of a blood sample taken from the driver at a hospital at 2:00 A.M. (two hours later) was found to be 0.07%. Does this mean that the level was below 0.08% at the

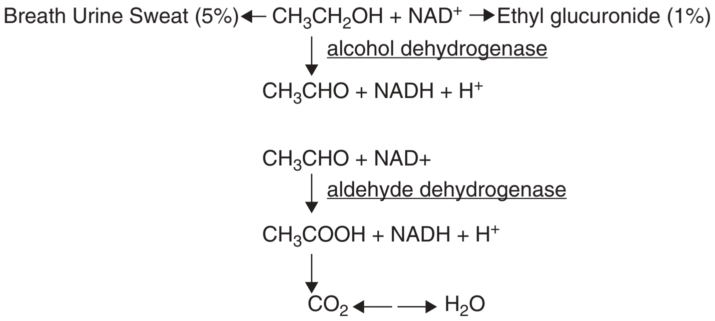

Figure 12.2 General scheme for ethanol oxidation. (a) <10% ethanol excreted in breath, sweat, and urine; (b) ∼90% ethanol removed by oxidation; (c) most of this ethanol oxidation occurs in the liver; (d) ethanol cannot be stored in the liver; (e) no major feedback mechanisms to pace the rate of ethanol metabolism to the physiological conditions of the liver cell.

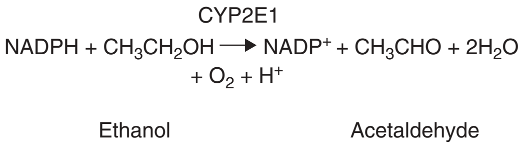

$$\underset{\text{Ethanol}}{NADPH + CH_3CH_2OH + O_2 + H^+} \xrightarrow{\text{CYP2E1}} \underset{\text{Acetaldehyde}}{NADP^+ + CH_3CHO + 2H_2O}$$

Figure 12.3 Microsomal (cytochrome P450) oxidation of ethanol in the liver.

time of the accident? Well . . . it depends. Let's assume that the person stopped drinking at a time prior to the accident, so that all the ethanol was absorbed from the person's gastrointestinal tract at midnight (time lag). We can extrapolate the BAC to the time of the accident as follows: Using the average rate of 0.017 g/dL per hour for two hours, the BAC two hours earlier (12:00 midnight) would yield a BAC of 0.104 g/dL, using the average "burn-off" rate of 0.017 g/dL per hour. Under these circumstances this person would be considered legally intoxicated at the time of the accident. On the other hand, if the person stopped drinking very shortly before the time of the accident, the BAC would still be rising at midnight and would continue to rise and peak sometime between midnight and 2:00 A.M. This clearly complicates the estimation of what the level would have been at midnight, and caution needs to be applied to such assessments. Therefore, a considerable range of BACs may be possible, depending on the drinking pattern, and this must be taken into consideration when interpreting the data.

12.3 ALCOHOL IMPAIRMENT AND DRIVING SKILLS

Driving or operating automobiles or any other vehicles while under the influence of alcohol is dangerous and perhaps the most commonly encountered forensic issues associated with this drug. Legislation around the country has enacted various rules mandating fines and penalties for driving under the influence of alcohol. Alcohol is a central nervous system depressant. The precise mechanism by which alcohol affects cognition and psychomotor control is not known completely, but the fact that it does so is very well documented and is well understood by most members of the public. To appreciate the effects of alcohol on driving skill, the complexity of driving itself must be recognized. Driving is a multifaceted process that requires a combination of hand–eye coordination, psychomotor function, muscle control, and cognitive function. Furthermore, driving is a divided-attention task, and the driver does not have the comfort of focusing exclusively on any one of these components, but rather, must concentrate on all simultaneously (i.e., monitoring a number of sensory inputs and responding appropriately with all the changing circumstances). Alcohol use can affect the implementation of individual tasks, and its effects are therefore more pronounced on more complex tasks, of which driving is a perfect example.

Alcohol causes a slowing of nerve conduction, which results in slower reaction times, difficulty in processing information, and thus impaired performance in divided-attention tasks. People are clearly affected differently at different BAC values for various reasons discussed previously. Sometimes the effects of alcohol are not apparent without specialized testing and must be elicited by tasks that challenge those abilities (see field sobriety tests for impairment). A person can be affected by alcohol to the extent that it can impair driving skills before overt signs of intoxication (e.g., staggering and slurred speech) materialize. The BAC value at which driving ability is affected by alcohol use is a recurrent issue confronting forensic toxicologists, although most toxicologists agree that any person with a BAC above 0.1 g/dL would have impaired driving skills, and those above 0.2 g/dL would show signs of overt intoxication on close inspection, even by a nonexpert.

The pharmacology and toxicology of alcohol differ from person to person. Various behaviors and physiological functions affected by ethanol result in impairment of driving skills, including reaction time, tracking, vigilance or concentrated attention, divided attention, information processing, visual functioning, perception, and psychomotor functions. Detrimental effects on reaction time are known to occur at BAC values in excess of 0.07 g/dL. Bloods levels of ethanol ranging from 0.012 to 0.12 g/dL increase the reaction time more than threefold and BAC values over 0.12 g/dL impair reaction times even further. With more complicated tasks that require more than one attentive task occurring simultaneously (e.g., driving), increased reaction time is even more evident. This situation is known to occur with the impairing effects of ethanol on driving skills. Driving requires that a person control speed, track position, and maintain lane control, and monitor various outside signals visually and react to them, all of which are impaired by increasing BAC. Since alcohol is a central nervous system depressant, it affects a wide range of these functions. (Connors and Maisto, 1980; Linnoila et al., 1980; Moskowitz et al., 1985; Steele and Josephs, 1988; Maylor et al., 1989 Lipscomb and Nathan, 1980). Statistics consistently show that the crash risk is increased significantly when BAC values are above 0.04 g/dL. Several studies with automobile simulators have shown that risk taking is affected at higher levels of alcohol, described as 0.106 g/dL (Mongrain, 1974).

The area most often showing the greatest impairment with alcohol consumption is divided-attention performance. Driving is a task that requires considerable divided attention, and decrements have been shown to occur with BAC as low as 0.02 g/dL and worsens with increasing BAC. The second-fastest rise in impairment occurs with tracking performance, which can occur at or below 0.05 g/dL. Driving is a time-sharing task that requires two activities: compensatory tracking and visual search. Divided-attention tasks are particularly sensitive to the effects of rising BAC, and such impairment may begin to occur at very low levels of BAC (<0.02 g/dL) for some people. There is sufficient evidence to suggest that BAC values of 0.05 g/dL and above are capable of producing impairment of the major components of driving skills, including reaction time, tracking, divided-attention performance, information processing, visual functions, and psychomotor impairment (Nito et al., 1964; Money et al., 1965; Aschan and Gergstedt, 1975; Baloh et al., 1979; Fagan et al., 1987). A major factor that determines the magnitude of such impairing effects in a person is tolerance (see the discussion below). Tolerance to some of the impairment parameters of ethanol can occur, but the extent to which it develops is very difficult to predict (Levine et al., 1975; Lipscomb et al., 1980; Wilson et al., 1984; Niaura et al., 1988). However, it is known that the impairing effects of alcohol on experienced versus naive drinkers is attenuated for some of these tasks in heavy drinkers. Even so, serious effects of alcohol on driving skills occur in all subjects at BAC levels of 0.10 g/dL or higher, and cognitive function will also be affected in these subjects.

12.4 FIELD SOBRIETY TESTS

Field sobriety tests, including the walk and turn, one-leg stand, and horizontal gaze nystagmus, are commonly used in field testing to detect impairment (Aschan, 1958). Fundamentally, these are only screening tests that presumptively detect impairment and must be followed by some other analytical measurement of blood alcohol content directly or via extrapolation. Horizontal gaze nystagmus is more reliable at predicting blood alcohol content than the walk and turn and the one-leg tests. Field sobriety testing is generally applicable to blood alcohol levels of 0.08 g/dL and above. At lower levels more studies are needed. Consequently, at levels lower than 0.08 g/dL, field sobriety tests will not be helpful. Furthermore, the field sobriety test must take into account factors such as prescription medications or altered health status, such as neurological or orthopedic problems.

Pharmacologically what happens when a person starts to drink alcohol? At first, as the BAC begins to rise, there is an apparent stimulatory effect. The person may become more talkative and they may seem more social and have increased self-confidence. This effect is not a stimulatory effect of ethanol but a depression of the inhibitory centers of the brain. As the BAC increases, impairment of judgment, decision making, perception, motor function, and reaction time occurs. The impairment of motor function and muscular coordination starts to affect the ability to walk, speak, and maintain balance. Significant and progressive reduction in mental and physical abilities continues as the BAC rises and the person is visibly intoxicated. In relatively low-tolerant or inexperienced drinkers, a BAC value over 0.30% can result in loss of consciousness and possible death, and with concentrations rising over 0.40%, death becomes increasingly likely. Chronic heavy use leads to a high incidence of brain damage, liver disease, heart disease, stroke, metabolic disorders, and oral, lung, and liver cancers, and abuse during pregnancy increases the risk of fetal malformations and development of fetal alcohol syndrome.

Tolerance is the lessening of the effectiveness of a drug after a period of prolonged or heavy use (Levine et al., 1975; Lipscomb et al., 1980; Wilson et al., 1984; Niaura et al., 1988). Studies have shown that chronic alcohol users can have as much as twice the tolerance for alcohol as an average person. Tolerance develops from two separate mechanisms: dispositional (or metabolic) and functional. *Metabolic tolerance* occurs because the drug is metabolized or inactivated at a faster rate after prolonged use (Lipscomb and Nathan, 1980). Thus, a given dose produces a lower BAC value. *Functional tolerance* is an actual change in a person's sensitivity to the drug at an organ or biochemical system level. With functional tolerance, greater amounts of the drug are required to elicit similar effects with increasing use of the drug. It is important to note that even in heavy users of alcohol, impairment is clearly measurable at the BAC levels that are currently used for traffic law enforcement. Tolerance is not complete for all systems and may be marked in some but not others. However, it is significant to note that even though cognitive impairment is present, many highly tolerant persons are capable of engaging in physical activities and actions (including fighting) at levels that would result in unconsciousness or even death in nontolerant persons.

12.5 BLOOD ALCOHOL MEASUREMENTS

Early blood alcohol testing involved distillation methods based on alcohol from the blood, followed by oxidation and titration. One of the carryovers of these techniques is the use of whole blood as the sample of choice for legal blood alcohol determinations in body fluids. Although popular up until the 1970s, such methods have largely been replaced by enzymatic and gas chromatographic methods (Jones and Pounder, 1997).

12.5.1 Enzymatic Methods

Enzymatic methods are attractive because they can be automated readily and thus allow for high-throughput analysis. Typically, enzymatic ethanol determinations are often done as part of a test panel on plasma or serum for other clinical chemistry parameters. This is often done in a medical laboratory setting and is not considered "forensic" testing. The use of plasma is an important consideration, since the reports issued by hospital laboratories often do not specify the specimen type.

Plasma contains more water than whole blood, and correspondingly more alcohol per milliliter. The accepted practice for converting a plasma or serum alcohol into a whole blood alcohol is to divide by 1.18, although a range of 1.10 to 1.30 gives a 95% confidence interval for the corresponding whole BAC. The fact that there is a range associated with this conversion makes measurement directly in whole blood the method of choice.

Enzymatic methods rely on the enzyme-specific oxidation of ethanol to acetaldehyde with alcohol dehydrogenase. This oxidation requires reduction of the cofactor oxidized nicotinamide–adenine dinucleotide (NAD^+) to NADH, which is accompanied by a change in absorbance that can be monitored spectrophotometrically. Enzymatic methods can be susceptible to interference from other alcohols oxidized by the same enzyme, but the differential kinetics of these reactions lends some specificity to the method. When calibrated using the appropriate standards, conducted in duplicate, and analyzed alongside matrix-matched controls with acceptable results, enzymatic oxidation can be an acceptable forensic method of determining BAC. These precautions are rarely followed in a clinical setting, however, and the use of hospital alcohol results for forensic purposes should be interpreted with this limitation in mind.

12.5.2 Headspace Gas Chromatography

By far the most popular, precise, and accurate method of determining blood alcohol is by gas chromatography (GC). This technique has the advantages of being semiautomatable, extremely reproducible, and compatible with multiple sample types. Currently, both packed column and capillary column procedures are popular and acceptable for forensic blood alcohol analysis. GC assays provide the greatest amount of flexibility and specificity in analyzing for these volatile compounds. Direct injection of biological samples into GC columns has been used in the past as the method of sample introduction. This typically leads to column contamination and decreased performance.

The incorporation of headspace sampling into the method prevents the buildup of nonvolatile contamination at the head of the column and helps to maintain consistent performance and extend column lifetime. Analysis time and resolution are two critical factors when developing a GC assay for ethanol. Analysis time for each sample should be as short as possible, while maintaining baseline resolution for all analytes.

An aliquot of blood (200 μL) in sodium fluoride and potassium oxalate is removed in duplicate to test for ethanol and other volatiles of low relative molecular mass. This is performed by mixing the blood using an automated pipettor or diluter with 2 mL of a 3 M NaCl solution that contains 1.0 g/L of 1-propanol as the internal standard. The mixed sample is dispensed into a 10-mL septum-topped vial and sealed. The vials are loaded into an automated headspace autosampler attached to a GC equipped with flame ionization detection (FID) in duplicate. Each sample is run in combination with a set of contemporaneous calibrators (0.4, 0.8, 1.6, and 3.2 g/L), a blank following the highest standard, and an externally purchased reference control after every tenth sample. The preservatives are important in this analysis. The sodium fluoride is an enzyme poison that will retard the formation or degradation of alcohol by bacteria. This is especially important in postmortem casework in which there is significant opportunity for bacterial contamination after death or at autopsy. It is less significant in blood drawn under sterile conditions from a living person. The potassium oxalate is an anticoagulant, which ensures that the blood remains homogeneous and does not separate into red cells and serum, since the water-enriched serum sample would contain more ethanol than would the whole blood. Since GC–mass spectrometry (GC–MS) is typically not used to test volatiles of low relative molecular mass, the use of complementary GC columns for duplicate testing is highly desirable, since (as in this case) the retention times obtained as a result of the different selectivity of the two columns provide an added dimension in the identification of a compound. Frequent calibration is desirable, but not essential. However, if a laboratory performs quantifications using a historical calibration curve, the reliability of the curve should be checked frequently with appropriate and comprehensive controls. Control samples obtained from and certified by an external vendor allow the laboratory to demonstrate the accuracy of its procedure by comparison with other laboratories.

When blood samples are analyzed for forensic purposes, civil or criminal, a standard procedure is to make duplicate determinations (Gullberg and Jones, 1994; Andreasson and Jones, 1996). The concentration of blood alcohol should be reported with the confidence limits such as 95% or 99% (Goldstein, 1983; Jones and Schubertz, 1989). Therefore, it is crucial that in hospital laboratories, the calibration of standards test by biological specimens along with the unknowns be done and kept on record for forensic evidence (Fraser, 1986). Cases have been thrown out of court due to the absence of the raw standardization data. Analyses done by a clinical toxicology laboratory and a certified forensic lab are dissimilar. Forensic laboratories perform analyses on samples of whole blood, whereas clinical laboratories usually measure plasma or serum ethanol concentration. The water content of plasma and serum is higher than that of whole blood and therefore requires an appropriate conversion (Winek and Carfagna, 1987; Frajola, 1993). Typically, the range is between 15 and 20% for plasma as compared to whole blood, and this conversion factor should be used in forensic work (Rainey, 1993). This difference may be significant in some cases, and should not be missed.

12.5.3 Breath Alcohol Testing

Driving under the influence (DUI or DWI) of alcohol and drugs is one of the most significant issues associated with abuse of these materials. Several testing strategies have been developed and introduced to assess the role of alcohol intoxication in impaired driving and to detect and prosecute offenders. Alcohol is still the drug that contributes most often to impaired driving. The smell of alcohol beverages on the breath has long been recognized as an indication that a person has consumed ethanol. Alcohol is eliminated unchanged in the breath and hence provides a means by which alcohol can be measured in the body. Breath alcohol instruments were developed to provide a fast and easy way of monitoring BAC indirectly. Analysis of a person's expired air provides an indirect way of monitoring volatile substances in the pulmonary blood. Breath has become the biological specimen of choice for measuring alcohol concentration in association with the enforcement and prosecution of drivers suspected of being impaired. Some of the reasons for this preference include:

- Breath collection is generally considered to be noninvasive.
- Rapid and inexpensive analysis allows for timely results, often as a field test.
- Analysis is straightforward and requires minimal operator training.
- Analysis does not require sophistcated laboratory equipment to obtain results.

These factors have combined to establish breath alcohol analysis as the most common forensic analysis employed in drunk-driving enforcement and workplace alcohol testing. Typically applied as a field test by a trained user, the interpretation of results begins with a clear understanding of the purpose of such measurement. Two classes of instrument for breath alcohol analysis can be distinguished, based on whether the results are intended for qualitative screening or as evidentiary testing for prosecution purposes. For example, prearrest screening breath tests clearly have a different purpose from that of the subsequent evidential breath test, which may be used as binding evidence in court. The serious implications associated with a drunk-driving conviction necessitate, at a very minimum, a sound forensic breath test protocol, which includes:

1. Trained and qualified instrument operators
2. At least 15 min of pre-exhalation observation
3. Internal instrument standards to verify calibration
4. Collection of duplicate breath samples
5. Appropriate agreement (i.e., ±10% of the mean) between duplicates
6. External control standards
7. Blank tests performed between all measurments
8. Appropriate printout of all data
9. Error detection systems
10. Instruments evaluated periodically according to an established quality assurance procedure to ensure optimal analytical performance, including accuracy, precision, linearity, and critical-system evaluations (e.g., acetone detection, radio-frequency interference detection, mouth alcohol detection)

12.5.4 Breath Alcohol Instrumentation

Ethanol is detectable in the breath and hence provides a means by which evidentiary testing can be accomplished to provide direct onsite collection and analysis of blood alcohol concentration for forensic purposes, principally for traffic law enforcement. The important characteristic of ethanol that allows for its measurement in the field is that ethanol is volatile. The theory that provides the foundation for such testing is based scientifically on the fact that the distribution of ethanol between the blood and the alveolar air obeys Henry's law. Therefore, at a given temperature, a direct relationship exists between the amount of alcohol (volatile substance) dissolved in blood (liquid) and the amount of alcohol in the vapor above the solution (alveolar air). At the temperature of expired air (34°C), the amount of ethanol in 2100 mL of alveolar air is assumed to be equivalent to the amount of ethanol in 1 mL of venous blood. Various instruments have been developed to take advantage of this ratio (2100 : 1). Various handheld devices are available for roadside testing and prearrest screening. For more controlled quantitative analysis of BrAC for evidential purposes, the instruments are much larger and more sophisticated, and include ways to check calibration, analyze alcohol free air, and produce hardcopies of the results. The best-known and most robust chemical handheld roadside breath alcohol screening tests generally utilize a fuel cell–based technology. These sensors oxidize ethanol to acetaldehyde and in the process produce free electrons, with the electric current generated directly proportional to the amount of alcohol present in the sample. Other alcohols present in the breath, such as methanol and isopropanol, also undergo chemical oxidation, although at different rates. Acetone (e.g., untreated diabetes) exhaled in the breath is not oxidized and will not cause a false positive. However, if acetone levels are exceedingly high, isopropanol can be formed in the body by reduction of acetone. Also for alcohol, enzymatic tests with a color endpoint have been developed for the field measurement of saliva alcohol. Such tests may have some potential value as screening tests in instances where the subject cannot provide breath, such as if injured and immobile.

In 1938, Rolla Harger developed the first commercially available forensic breath alcohol instrument intended for law enforcement applications, called the Drunk-o-meter. To use the Drunk-o-meter, the person being tested blew into a balloon and the air in the balloon was then released into a chemical solution containing potassium permanganate. If there was alcohol in the breath, the chemical solution changed color. The greater the color change, the more alcohol in the breath. The level of alcohol in a person's blood could then be estimated by a simple equation. This was a wet chemistry method based on optical measurement of the color change that occurred in potassium permanganate resulting from the oxidation of alcohol in the breath sample. This was largely replaced by the much more portable Breathalyzer developed in the 1950s. The Breathalyzer also utilized a colorimetric technique following the oxidation of alcohol in a potassium dichromate solution. As concerns began to develop that there was a lack of specificity and accuracy of wet chemical methods, instruments that employed the optical absorption of infrared energy according to Beer's law began to replace these methods for most forensic applications during the 1970s. Several variations of the infrared technology have been applied, including the use of multiple wavelengths to increase the specificity for ethanol. Most evidential breath-testing instruments used today identify and measure the concentration of alcohol by its absorption of infrared energy at wavelengths of 3.4 and/or 9.5 μm, which correspond, respectively, to the C—H and C—O vibration stretching in the ethanol molecules.

During the 1980s, fuel cell technology emerged as a useful analytical method in forensic breath alcohol measurement. Although fuel cells have been employed typically in screening devices, their technology has improved dramatically in recent years. Selectivity has been enhanced in one modern evidential instrument, the Alcotest 7110 (Draeger Instruments, Durango, CO), which now employs both infrared and fuel cell methodologies to quantify ethanol in breath and represents the first dual technology breath alcohol instrument. Employing two different and independent measuring systems, infrared spectroscopy and electrochemical cell technology, it provides the highest level of forensic and legal veracity. Such true dual-sensor instruments are now readily available on the market. By means of these instruments, both the electrochemical sensor and the infrared-optical sensor operating at 9.5-μm-wavelength quantify BrAC.

Modern breath alcohol instruments offer several advantages over their predecessors. The modern instruments are computerized, which greatly reduces the potential for operator error and ensures compliance with specific protocol requirements. In addition, automated instruments enable rapid analysis, error detection, ensured compliance with breath sampling parameters, automated purging of sample chamber and signal processing, regulation of test protocol and sequence, data collection and hardcopy printout of results. Measurement quality control is enhanced, therefore, by allowing only those test results that fully conform with predetermined standards to be printed out and to qualify as evidence in court. Contemporary instruments also allow features of “intelligent measurement” to be incorporated. These features permit the instrument to “make decisions” regarding test performance, run, or analytical properties. Data collection features also allow for the development of control charts and other methods to evaluate instrument and programmed performance over time. Various forms of data analysis can provide relevant information to the courts and enhance confidence in the results. Many features of the emergent analytical technology have been integrated advantageously into forensic breath test instruments.

12.5.5 Conversion from BrAC to BAC

The extrapolation from breath alcohol content to blood alcohol content level at the time of accident or arrest is one of the most important steps in the forensic process. While the conversion process takes into account simple mathematical conversion, it is imperative to understand that in most cases the extrapolation of the amount of blood alcohol concentration to a time other than the time when the specimen of body fluid or breath was taken from the subject is applicable only to the average person in a normal physical condition and may not be relevant otherwise. Current research has shown that there are multiple variables, such as blood breath ratio, absorption rates of alcohol, elimination rates of alcohol, and difference in total body water (Winek and Esposito, 1985; Karch, 1997). The analyst should be aware that the BAC/BrAC ratio of alcohol changes as a function of time after drinking, depending on whether the test was done on the absorptive or postabsorptive phase of the alcohol curve, which my have an important bearing on the interpretation of these results.

12.5.6 Urine and Saliva

Urine has been used in some jurisdictions as a specimen for forensic alcohol testing. It is amenable to analysis by both the enzymatic and chromatographic methods described above. Its major limitations are collection and interpretation. Urine collection, if observed, raises privacy issues, and if not observed, raises issues of chain of custody and the potential for adulteration, substitution or dilution. With respect to interpretation, urinary alcohol concentration is an indirect measure of BAC. Urine is formed in the kidneys as an ultrafiltrate of blood. As with plasma, therefore, a factor must be applied to account for the higher water content. Of more importance, however, is the fact that since urine is formed over time, the urine alcohol concentration is an average corresponding to the blood concentration over the time during which the urine was formed and collected in the bladder. Since this can cover an appreciable amount of time (up to several hours), even a urine sample corrected for water content may not reflect the true BAC at the time of urine collection. To correct for this, for forensic purposes urine is generally collected for analysis about an hour after an initial void. The subject is directed to empty his or her bladder, after which he or she is allowed a period of time (approximately an hour) during which new urine corresponding to their current BAC is formed and collected for testing. Such complications and the associated problems of interpretation mean that urine alcohol determination is not a preferred forensic approach, although it is acceptable in many jurisdictions. Problems with interpretation are generally legislated away with a statutory per se urine alcohol offence.

Saliva or oral fluid is a plasma ultrafiltrate produced through the parotid and other glands in the oral cavity. Drugs, including alcohol, that are present in the blood will also be found in the saliva (Jones, 1981, 1993). Care must be taken when examining oral fluid for alcohol because it can be subject to contamination from drugs and alcohol simply present in the oral cavity. The length of time of the contamination depends on the drug's lipophilicity and water solubility. Mouth alcohol is generally dissipated by absorption and exhalation within 15 min. This is why at least a 15- to 20-min observation and waiting period is stipulated prior to the subject being offered an evidential breath test, and a similar waiting period is essential for a test of salivary alcohol. A number of companies have marketed oral fluid alcohol test kits, which are based on the principle of enzymatic oxidation. These are to be used as screening devices because their quantitative accuracy is variable.

Generally, from an analytical standpoint there is a good correlation between circulating blood alcohol and oral fluid alcohol concentrations, and a properly collected and preserved oral fluid sample can be subjected to an analysis, usually by headspace GC, with good-quality results. Samples are usually collected in a tube with minimal headspace to prevent loss, together with appropriate preservative (enzyme inhibitor) to prevent postcollection alcohol production.

12.5.7 Ethyl Glucuronide

This is a compound formed from a minor metabolic pathway of ethanol (Yegles et al., 2004). Less than 1% of ingested ethanol is conjugated in the liver with glucuronic acid to form a water-soluble compound that is present in blood and excreted in urine. It has also been detected in hair, and this is currently being used in some alcohol and drug monitoring plans clinically (Jurado et al., 2004; Skipper et al., 2004). Ethyl glucuronide is a stable compound and a specific indicator for ethanol ingestion and has been applied to some forensic applications; however, measurement of EG levels in both blood and urine do not bear a constant relationship to BAC (Schmitt et al., 1997; Bergstrom et al., 2003; Jurado et al., 2004; Skipper et al., 2004; Yegles et al., 2004). The demonstration of EG in a specimen only indicates recent (within the last 40 h or even longer) ingestion of ethanol, which limits its use. It is capable of determining whether or not the person had been drinking in the recent past. A serum ethyl glucuronide concentration of greater than 5 mg/L when the BAC is less than 0.10 g/dL has been suggested as representing alcohol misuse (Drummer, 1999).

12.5.8 Postmortem Determination of Alcohol

One of the primary roles of the forensic toxicologist is to determine the role of drugs and alcohol in establishing a cause of death (Pounder and Jones, 2006). Ethanol is a major analyte in postmortem forensic determinations (Turkel and Clifton, 1957; Zumwalt et al., 1982). Commonly, the forensic toxicologist must opine on the role that alcohol may have played in a forensic case in both criminal and civil cases. This includes traumatic injuries such as motor vehicle accidents, alcohol poisoning, criminal behavior and human performance toxicology, and occupational incidents, to name but a few. There are major issues associated with the interpretation of the analytical alcohol results on samples obtained at autopsy. These are mostly due to the lack of homogeneity of blood samples, postmortem alcohol production or loss, and disruption of the gastrointestinal tract following traumatic injury, resulting in the release of alcohol into the surrounding tissues. However, with proper care, ethanol analysis can be accomplished with high confidence on postmortem samples. Postmortem forensic analysis is discussed in detail in Chapter 13, so the topic is only summarized here.

Blood is the usual specimen provided for postmortem ethanol analysis. The site from which postmortem blood is collected is important (Plueckhahn, 1968; Prouty and Anderson, 1987). The preferred site is the femoral vein. Blood from the pericardial sac or chest cavity can have falsely elevated alcohol levels due to postmortem diffusion of alcohol from the stomach (Plueckhahn and Ballard, 1967; Pounder and Smith, 1995). As with antemortem blood, samples should be collected into containers with appropriate preservative (sodium fluoride) and anticoagulant (potassium oxalate), depending on how the analysis is to be performed. Because of ethanol's uniform distribution through the body, other fluids and tissues can be used for analysis as well. Vitreous fluid from within the eye consists of 99% water with small amounts of salts and mucoproteins. It is in an enclosed space and is not normally subject to contamination by microorganisms, and these specimens are easily collected at autopsy. Vitreous is important because due to its anatomical location, it is usually well preserved after death and is generally unaffected by postmortem changes. The high water content of vitreous means that measured ethanol levels should be comparable to blood levels. The ratio of blood to vitreous concentration is between 0.57 and 0.96, with a value of 0.85 the most commonly used best estimate. Blood has a lower water content than vitreous, so the expectation is that the equilibrium blood vitreous alcohol ratio will be less than 1. In cases where the blood/vitreous ratio exceeds 1 it is likely that death occurred while the person was in the absorptive phase and before diffusion equilibrium had been obtained, and this observation may be of forensic significance (Sturner and Coumbis, 1966; Felby and Olson, 1969; Kraut, 1995; Yip, 1995).

Urine collected from the bladder like vitreous is resistant to postmortem contamination by microorganisms. Postmortem alcohol production in urine occurs only when the urine contains large amounts of glucose (diabetes) or there was a urinary tract infection during life (Drummer, 1999). Urine is easily sampled at autopsy provided that the bladder is intact. A urine alcohol measurement does not reflect the alcohol content of urine at any point in time (i.e., time of death). It may be a useful specimen to use to assess possible prior consumption of alcohol. Urine concentrations are generally similar to those of blood but may vary considerably and should not be used to predict BAC without incurring a large error. As long as this is recognized, urinary alcohol estimation can be a useful measurement.

12.5.9 Quality Assurance of Alcohol Testing

Monitoring of the day-to-day performance of the collection, storage, and analytical methodology is a vital component of a proper forensic alcohol analysis (Moffat et al., 2004). The laboratory must have the highest confidence that the quality of the data produced is of utmost precision and accuracy because the results are often used for criminal prosecution or civil litigation. Use of a range of replicate determinations is an important means of controlling the quality assurance of test results. All specimens are carefully inspected when they arrive and the condition is noted. When needed, preservatives and anticoagulant are used to inhibit enzyme activity to prevent post-collection microbiological production of alcohol and blood clotting in the samples, respectively. Proper labeling is essential, with all the required identifying information to be included in a suitable label. The collection container must be properly sealed to prevent unauthorized tampering or loss of sample during transport and storage. For specimens that are stored, appropriate refrigeration (−4°C) of the samples is needed to prevent loss of the volatile alcohol, and the rate of lost alcohol during this time needs to be established. Calibration and measurements of controls and standards must be completed and documented by the measuring laboratory, and copies of the analytical data, including chromatograms should be maintained in a secure storage facility so they are available for scrutiny or produced as evidence. The laboratory performing the alcohol analysis must be properly accredited and must participate in external proficiency testing to establish that all results produced are precise and accurate. Failure to do so greatly reduces the confidence in the data produced and may result in the data being inadmissible in court.

CHAPTER 13

Fundamental Issues of Postmortem Toxicology

DONALD B. HOFFMAN
Department of Sciences, John Jay College of Criminal Justice, The City University of New York, New York

BETH E. ZEDECK
Pediatric Nurse Practitioner, New York

MORRIS S. ZEDECK
(Retired) Department of Sciences, John Jay College of Criminal Justice, New York

Summary The basic principles of forensic postmortem toxicology are presented. In this chapter we discuss the acquisition and usefulness of various specimens, current analytical techniques, and interpretation of findings. Special problems associated with the interpretation of drug levels include the conditions of the specimens and the effects of postmortem redistribution, postmortem drug changes, pharmacogenomics, drug interactions, and embalming fluid.

13.1 Introduction 458
13.2 Tissue and fluid specimens 460
  13.2.1 Blood 460
  13.2.2 Urine 461
  13.2.3 Vitreous humor and cerebrospinal fluid 461
  13.2.4 Gastric contents 462
  13.2.5 Meconium 463
  13.2.6 Brain 464
  13.2.7 Liver and bile 464
  13.2.8 Lung, spleen, kidney, and skin 465
  13.2.9 Muscle 465
  13.2.10 Bone, teeth, nails, and hair 465
  13.2.11 Other materials for analysis 466
13.3 Specimen collection and storage 466

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

13.4 Extraction procedures 467
13.5 Analytical techniques 467
13.6 Interpretation 470
  13.6.1 Postmortem redistribution 470
  13.6.2 Pharmacogenomics 471
  13.6.3 Drug interactions 472
  13.6.4 Drug stability and decomposed tissue 473
  13.6.5 Effects of embalming fluid 474
Conclusion 475
References 476

13.1 INTRODUCTION

One aspect of forensic chemistry is forensic toxicology, the discipline that analyzes biological fluids and tissues for the presence of drugs and poisons, determines their concentration, and interprets the findings. The results may be helpful in determining whether one or more drugs played a role in a person's injury, illness, behavior, or death. Thus, forensic toxicology can be divided into clinical toxicology, human performance toxicology, employment drug testing, and postmortem toxicology. This fascinating and challenging profession draws on the disciplines of pharmacology and of biochemical, analytic, and organic chemistry. It is not our purpose here to present an inclusive detailed treatise on forensic toxicology, as excellent source materials of this nature are available. Methods for the extraction of drugs from various tissues and the analytical procedures used for the detection and measurement of drug concentration have been described in many textbooks and reference books (Chamberlain, 1995; Wong and Sunshine, 1997; Levine, 2006; Smith et al., 2007), as have the effects of drugs (Marquardt et al., 1999; Brunton, et al., 2005; Klaassen, 2008). The reader interested in pursuing these topics in more detail is encouraged to utilize these and all other relevant sources of information.

In this chapter we focus on a special area of forensic toxicology: postmortem toxicology. We wish to present to the forensic scientist an overview of the current key issues involved in the practice of postmortem toxicology. Several earlier articles discussing this particular area of forensic science are cited (Prouty and Anderson, 1990; Hilberg et al., 1999b; Drummer, 2004, 2007; Skopp, 2004; Dolinak, 2005; Levine, 2006; Rodda and Drummer, 2006; Spitz and Spitz, 2006; Watterson et al., 2006; Karch, 2007; Garriott, 2008).

Postmortem forensic toxicology is an essential and integral component of any comprehensive medicolegal investigation into the cause and contributing circumstances of death. As such, it is essential that it be carried out appropriately and efficiently in order to provide the necessary information. An inclusive toxicological investigation encompasses knowledge of case history, specimen selection, choice of extraction procedures, application of multiple analytic techniques, and interpretation of findings.

Postmortem toxicology involves the analysis of tissues and fluids taken from cadavers, either shortly after death or from decomposed or exhumed bodies, and also from formalin-fixed specimens. Each presents special situations that must be taken into consideration to ensure proper interpretation of the data. Drug samples, drug paraphernalia, and prescription items may also be found at the site of death and usually become part of the postmortem toxicological examination.

Postmortem tissues taken from a person who had been confined to a hospital for an extended period may have limited value because of drug clearance occurring over a period of time prior to death. However, the analysis of antemortem blood and urine, when available and especially if taken around the time of hospital admission, may become essential in postmortem protocols when there has been a prolonged period of survival, or when extensive administration of transfusions occurred prior to death.

Knowledge of the case history is very important. The circumstances surrounding the death or the indicated use of specific agents may require the use of non-routine procedures to check for the possible presence of certain classes of drugs or poisons that are not necessarily included in the routine protocols. Examples of such substances are digoxin and other cardiac glycosides, insulin, quaternary ammonium muscle relaxants, water-soluble or very highly polar compounds [e.g., γ-hydroxybutyrate (GHB)], and salts of heavy metals. Such substances have been encountered in homicidal or other criminal poisonings, and some following therapeutic treatments. There must be knowledge as to appropriate specimen selection and preservation. Certain specimens may be particularly appropriate for a particular drug or where the condition of the body is a factor. Issues such as postmortem redistribution of drug from its site at the time of death (discussed below) and postmortem drug stability or transformation play key roles both in the choice of specimen as well as in the subsequent interpretation of findings (Drummer, 2004; Skopp, 2004). Cases involving decomposed, dismembered, embalmed and exhumed bodies require special attention in order to maximize the utilization of such specimens as are available at autopsy.

Once an analyte has been detected in bodily fluids and tissues, questions arise. Is the concentration of analyte reflective of its concentration at the time of death, or is it higher or lower than at the time of death? Conditions that can alter drug concentration include changes resulting from human enzymes, bacterial enzymes, pH, temperature, and hydrolysis. Do the levels of drug found represent an overdose, intentional or otherwise, or a therapeutic level? Are tolerance and/or pharmacogenomics issues? Information as to the interval of time between death and the finding of the body, or information concerning therapeutic treatments of the person prior to death, is often not available. An additional factor to be considered when interpreting drug levels is the situation when a decedent had received significant amounts of replacement fluids prior to death. This is not, however, a simple "dilution" effect since the infusion of fluid results in efforts to reestablish new equilibrium between tissues and blood. Thus, the forensic toxicologist must consider a host of factors in reaching a decision as to the manner and cause of death. A primary objective of this chapter is to discuss those factors that can make it difficult to interpret postmortem toxicological findings in deciding whether or not a drug played a role in causing a person's death.

13.2 TISSUE AND FLUID SPECIMENS

Postmortem specimens taken at autopsy may include blood, urine, bile, cerebrospinal fluid, vitreous humor, gastric contents, meconium, various organs, bone, teeth, hair, and nails. If an autopsy is not performed, the postmortem specimens collected usually include peripheral blood, urine, and vitreous humor.

13.2.1 Blood

It needs to be emphasized that postmortem blood, even if obtained quickly after death, has already been affected by processes that occur at the time of death, such as loss of cell membrane potential and the initiation of hemolysis. Other postmortem processes begin early on and generally increase with increasing postmortem interval. Hence, any sample of postmortem blood is not physiologically or toxicologically equivalent to a clinical sample. Proper selection of the sites for blood sampling is crucial because of the phenomenon of postmortem redistribution. This is defined as the postmortem movement of drug away from its location at the time of death to another site in the body. Several basic mechanisms are believed to be involved (Pélissier-Alicot et al., 2003).

One mechanism is diffusion of drug from inside cells into the surrounding blood. This starts with loss of integrity of the cell membrane, which occurs at the moment of death and generally results in unregulated flow of substances, both endogenous and exogenous, out of cells. Because of this mechanism, postmortem blood is not used for the determination of electrolytes and glucose.

A second mechanism is release of drug from certain tissue depots and the subsequent formation of drug concentration gradients directed toward blood in the heart chambers. Of particular importance is antemortem accumulation of drug in lung tissue, which can act as a depot for many drugs, especially those that are lipid soluble. The postmortem release of drugs from lung depots results in the movement of drug via the pulmonary veins into the heart chambers. Such postmortem diffusion generally occurs with those drugs that have high volumes of distribution ($\geq$3 L/kg) and attain high lung levels; however, there is no hard-and-fast rule for this (Hilberg et al., 1999a). The release of drug from tissue depots is believed to involve the disruption of nonspecific drug–macromolecular binding as a result of pH changes and denaturation. Those drugs that concentrate in heart tissue will be released directly into heart blood, and sampling from the chambers would give an incorrect concentration of drug in blood at the time of death. It is known, however, that highly lipophilic drugs may move from blood into adipose tissue.

A third mechanism is the diffusion of drug away from the gastrointestinal and hepatic systems, which ultimately enters heart blood. This mechanism is particularly relevant when death occurs in the nonequilibrium phase; that is, significant amounts of drug are still present in the stomach, intestines, and in the portal vein–hepatic system. Thus, for example, ethanol from the gastrointestinal tract can redistribute to the pericardial and also to the pleural fluids (Plueckhahn, 1967). Thus, it is essential that when blood is taken for analysis, a sample be drawn from peripheral sites such as the femoral veins (or subclavian veins), as quantitation on femoral blood is considered to be the most reliable measure of the antemortem concentration just prior to death (Dalpe-Scott et al., 1995; Hilberg et al., 1999b). However, whenever possible, the results derived from femoral blood should be compared with results from other tissues and fluids.

The production of gases during decomposition and the resulting pressures can result in the mixing of blood from different sites (Prouty and Anderson, 1990). A sample of femoral blood may contain blood from the inferior vena cava, and cardiac blood may contain blood from the vena cava, subclavian, pulmonary, and aortic vessels. Other factors that make interpretation of postmortem blood data difficult include the degree of hemolysis, or any other changes in the blood matrix. Therefore, unlike clinical toxicology, where blood is separated into plasma or serum, whole blood is almost always analyzed in postmortem toxicology. Analysis of blood clots, especially after trauma such as head injury, may be useful for determining drug exposure prior to death, especially in cases of prolonged survival, during which time the circulating blood may have been cleared of drug. The concentration of drug in a blood clot may reflect the concentration in blood at the time of clotting, assuming that formation occurred at the time of injury. But care in quantitative interpretation is important since the consistency of the clot differs from that of circulating blood, because formation of the clot cannot always be assumed to be instantaneous (where there is a survival time before death) and because clots themselves may partially break down and release drug.

13.2.2 Urine

Unless death is so rapid as to preclude the drug from being excreted into the bladder, urine is useful for initial screening for the different classes of drugs and for their metabolites. Finding drugs in urine indicates prior usage but cannot itself be used to indicate any level of impairment or the concentration of drug in blood prior to death. Urine analysis presents fewer problems than blood analysis; for example, urine does not readily support bacterial growth.

13.2.3 Vitreous Humor and Cerebrospinal Fluid

Analysis of vitreous humor serves several useful roles. It must be emphasized, however, that vitreous humor is useful only when bodily decomposition has not set in. There is some evidence that determination of potassium levels in vitreous humor may be useful in some cases for approximating the postmortem interval (PMI). Potassium levels rise as the interval between death and collection of a sample increases; the correlation, however, is not always accurate (Stephens and Richards, 1987). More recent studies (Muñoz et al., 2001, 2002; Madea and Rödig, 2006) have demonstrated that certain statistical treatments of the data observed provide a somewhat more accurate estimation of the PMI. Nonetheless, the relationship between the concentration of vitreous potassium and PMI remains a controversial issue.

Vitreous humor has been used to analyze for the presence of many drugs and alcohol (Hepler and Isenschmid, 1998; Jones, 1998). Vitreous humor contains very little protein compared to blood and tissues. Only free drug in blood enters the vitreous humor, and since most drugs exhibit at least some significant degree of binding to proteins and other macromolecules in blood, vitreous drug levels are lower than those in femoral blood. Vitreous humor is a protected site; usually, there is little bacterial contamination and therefore little fermentation. Thus, it may be useful for determining whether the presence of alcohol in tissues was due to consumption prior to death or formed as a result of postmortem fermentation. Vitreous humor is a good backup sample for blood ethanol. The levels of alcohol in vitreous humor pre- and post-embalming are in fairly good agreement (Scott et al., 1974; Coe, 1976). Since the pre- and post-embalming samples in these studies were taken within a relatively short interval of each other, one cannot draw conclusions about the effects of embalming fluid over long periods of time. However, caution must be exercised whenever inferring a blood alcohol level from vitreous humor. The vitreous/blood (femoral) ratio of alcohol can be a very useful indicator of whether death occurred in the absorptive or post-absorptive state. At equilibrium (post-absorptive state), the ratio of vitreous alcohol to blood alcohol averages about 1.10 to 1.20 (Garriott, 2008). But if death occurs during the absorptive state (rising blood alcohol level), this ratio will be lower, possibly even less than 1.00. Postmortem redistribution from ocular tissue into vitreous humor may also occur.

Vitreous humor can be very useful for the determination of such antemortem conditions as hypo- or hypernatremia, dehydration or uremic conditions, and hyperglycemia. Vitreous glucose is known to decline, often erratically, with increasing postmortem interval, thus making determination of hypoglycemia generally impractical, if not impossible. However, under certain very limited conditions, a precipitously low level found very shortly after death may be indicative of insulin overdose (Coe, 1972, 1974).

Many drugs are known to enter the cerebrospinal fluid (CSF). Absent blood, CSF has been found useful for the qualitative determination of drugs. Analysis of CSF may reveal recent use of heroin, as 6-monoacetylmorphine has a longer half-life in this specimen than that of other fluids (Jenkins and Lavins, 1998; Engelhart and Jenkins, 2007).

13.2.4 Gastric Contents

The entire gastric contents should be taken for analysis and the total weight should be recorded. Any undigested tablets or other medicinals should be analyzed separately. The total amount of drug present in gastric contents, including any undigested material, is of forensic importance because it may be useful in interpreting the manner of death (i.e., it may be indicative of suicidal overdose). Since gastric content is normally acidic, detecting trace amounts of basic drugs in stomach contents may be due to back diffusion into the stomach as well as representing residual drug following ingestion.

13.2.5 Meconium

Exposure to maternal substance abuse during gestation is associated with many adverse social and significant perinatal complications, including a high incidence of stillbirths, premature rupture of the embryonic sac membranes, maternal hemorrhage, spontaneous abortion, fetal distress, and increased rate of mortality and morbidity (developmental and cognitive impairment) (Ostrea et al. 1989; Moriya et al., 1994; Moore and Negrusz, 1995; Ostrea, 2001). Previously, the history of in utero exposure to drugs was based on urinalysis and on maternal self-reporting, which may be unreliable, due to patient denial about addiction or fear of the consequences (Ostrea et al., 1989; Coles et al. 2005). In stillbirth infants and in infants that die after birth, meconium may be the specimen of choice for drugs of abuse testing because of the difficulty in obtaining urine specimens, and the use of organs may have limited value if the mother had stopped using drugs several days before delivery (Moriya et al., 1994).

Meconium, the first fecal matter of the neonate, which is normally expelled from the intestine during the first 24 to 48 h after birth, is a thick greenish-black material composed of epithelial cells, digestive tract secretions, and residue of swallowed amniotic fluid (Ostrea et al., 1992; Moore and Negrusz, 1995). Drugs accumulate in meconium either by direct deposition from bile or through swallowing of amniotic fluid (Ostrea et al., 1992; Moore and Negrusz, 1995; ElSohly et al., 1999b; Ortega García et al., 2006).

Meconium starts to accumulate in the intestines between weeks 12 and 16 of gestation and can provide a long history of in utero drug exposure as well as indicate during which period of gestation drug use took place (Callahan et al., 1992; Ostrea et al., 1994; Ortega García et al., 2006). Unlike urine, blood, and gastric contents, which allow the detection of drug exposure for 2 to 3 days, meconium extends this window to about 20 weeks (Ostrea et al., 1989; Abusada et al., 1993; Moore and Negrusz, 1995). Lewis et al. (1995) reported that meconium analysis detected threefold more cocaine-exposed babies than urinalysis did.

Meconium is readily available, and only a small sample is required for complete analysis, including confirmation analysis (Ortega García et al., 2006). Meconium testing does have some limitations. It is often an unfamiliar specimen in the clinical laboratory, being a sticky material that is more difficult to work with than urine and requires a thorough preliminary cleanup procedure prior to any analytical assay.

Solid-phase extraction (SPE) has been used to isolate drugs and metabolites from meconium, and both immunoassay and chromatographic techniques have been used for the screening of drugs in meconium. Preliminary screening results must be confirmed by gas chromatography–mass spectrometry (GC–MS). Coles et al. (2007) compared GC–MS with LC–MS–MS (liquid chromatography–tandem mass spectrometry) and found that LC–MS–MS has increased specificity and decreased interference rate, reporting fewer false-positive results. Marin et al.

(2008) developed a LC–MS–MS procedure that can detect and quantitate many benzodiazepines in meconium.

Various drugs and their metabolites that are found in neonatal or maternal urine have also been detected in meconium. Samples containing cocaine and/or cocaethylene also contained benzoylecgonine and *m*-hydroxybenzoylecgonine. The presence of cocaethylene in meconium suggests that both cocaine and alcohol were used during pregnancy (Callahan et al., 1992; Abusada et al., 1993; Steele et al., 1993; Moriya et al., 1994; ElSohly et al., 1999a). One study of postmortem analysis of three fetuses that had been exposed to cocaine showed the presence of cocaine in the meconium of a 17-week-old fetus, suggesting that fetal exposure can be determined early in gestation and in the very premature fetus. At this time, meconium contained a higher concentration of drugs than did neonatal urine (Ostrea et al., 1994).

Moore and Negrusz (1995) and Coles et al. (2005) found that some samples of meconium contained both 11-nor-Δ^9-tetrahydrocannabinol-9-carboxylic acid (THC-COOH) and 11-hydroxy-Δ^9-THC (11-OH-THC). Using trimethylsilyl derivatization and GC–MS, hydrocodone and hydromorphone were detected in meconium along with codeine and morphine (Moore et al., 1995). The level of codeine, hydrocodone, and hydromorphone increased dramatically following acid hydrolysis, while the morphine level also increased but not as significantly.

13.2.6 Brain

The brain is a good repository for lipophilic drugs such as anesthetics and hydrocarbons, whereas many polar compounds are not able to pass the blood–brain barrier in significant amounts. Clearly, psychotropic drugs enter the brain easily and their levels in brain are useful for determining the manner of death. Brain is also used as a backup specimen for blood alcohol. At equilibrium, the brain alcohol/blood alcohol ratio is about 0.85, but this number may be lower if death occurs during the period of a rising blood alcohol. Brain cocaine levels are useful for evaluating cocaine-related deaths (Bertol et al., 2008). From a few minutes to about 3 h after exposure, the brain/blood ratio for cocaine is approximately 4 to 10 (Spiehler and Reed, 1985). The ratio of brain to blood cocaine in cases of death within 3 to 6 h of exposure is approximately 0.4 to 0.8. These ratios may be different, however, in persons who had used cocaine chronically.

13.2.7 Liver and Bile

The liver concentrates most parent drugs and also metabolites, ethanol being the notable exception. There is an extensive body of data on the range of hepatic drug levels in therapeutic, toxic, and lethal cases. Bile is a good source for many drugs and their metabolites, in particular opioids and benzodiazepines, and because of its storage in the gallbladder, it may be the only specimen where residual drugs or metabolites may be detected when absent from other tissues. Bile may be useful in those cases where a prolonged interval elapsed between the last intake of drug and death. However, bile may contain high concentration of drug in cases of acute overdose where death occurs while large amounts of drug are still in the gastrointestinal tract and in the portal system.

13.2.8 Lung, Spleen, Kidney, and Skin

The analysis of lung tissue is useful in cases involving volatile substances such as gases (other than carbon monoxide), anesthetics, and solvents. Because spleen concentrates red blood cells, it may be useful as a last resort specimen for determining cyanide- or carbon monoxide–bound hemoglobin. Since cyanide derived from antemortem toxic exposure may be lost rapidly with advanced decomposition, spleen may be the only organ where toxicologically significant levels of cyanide may still be found. Where a routine blood sample is not available or is not considered usable, spleen may be used cautiously to estimate antemortem exposure to carbon monoxide. Analysis of kidney tissue is useful for detecting heavy metals. Analysis of unburned skin may be useful for detecting use of accelerants in arson-related cases.

13.2.9 Muscle

Muscle tissue is not a routine specimen, but is particularly useful in cases of advanced decomposition because skeletal muscle, being more slowly affected by decomposition, may still be present when little or no other soft tissue is present. Great care must be exercised when interpreting the findings quantitatively. For skeletal muscle, there is little postmortem toxicology data from which one can draw conclusions relating drug level and cause of death. The studies that have been done on fresh skeletal muscle indicate that the data are best used qualitatively. Of interest is the fact that cocaine, which is unstable in blood, has been detected in numerous cases in decomposed but dry skeletal muscle.

13.2.10 Bone, Teeth, Nails, and Hair

It is known that many drugs and heavy metals become incorporated in bone, teeth, nails, and hair. These tissues will not necessarily, however, contain drugs or toxic substances if death occurred rapidly after intake or administration, but are good indicators of chronic dosing. They are usually not analyzed postmortem except for exhumed, badly decomposed, or skeletonized bodies. Teeth are useful for analysis in bodies that have been badly burned. There is an extensive body of data on the incorporation of drugs and heavy metals into hair, which is used primarily in clinical toxicology. Hair can be cut into segments and analyzed individually to determine a time pattern of drug usage. Since hair grows at a rate of approximately 1 cm per month, one can approximate when drug usage was discontinued. Hair may reflect a single acute dose, depending on the drug and dose involved, provided that sufficient time has elapsed to allow entry of drug into the hair. In this regard the analysis of hair follicles may be particularly useful. Hair analysis requires careful and proper preliminary washing steps to remove as much as possible of any substances present as a result of external contamination.

13.2.11 Other Materials for Analysis

1. *Injection site*. If an injection site can be located, analysis may reveal useful qualitative and quantitative data. Since a drug may have come to what appears to be an injection site via the circulation, other skin samples should be taken for comparison.
2. *Insect larvae/maggots*. Larvae and maggots feed on decomposing tissue and will incorporate into their systems drugs that are in the tissues they consume. Collection of these insect specimens from decomposed or skeletonized bodies may provide much useful qualitative information.
3. *Nasal swabs*. Results from swabbing of the nasal cavities must be interpreted cautiously. Drugs are normally secreted into nasal mucosa. Finding small amounts of drug on a swab does not necessarily mean that the person snorted drugs prior to death. The amount of drug detected will help determine if snorting is a likely explanation.

13.3 SPECIMEN COLLECTION AND STORAGE

For all samples, proper collection of the sample in adequate amounts, and proper use of preservatives and storage containers (e.g., glass, plastic, nylon, foil) along with proper storage temperature are essential (Skopp, 2004). It is recommended that glass containers be used for liquids and glass or plastic containers for tissue and gastric contents. However, it must be noted that the use of certain types of plastic containers for storage can result in the loss of THC–COOH (Stout et al., 2000).

Blood should be collected in tubes containing at least 1 to 2% sodium fluoride as preservative and stored at 0 to 4°C. This will inhibit bacterial growth and enzyme activity, which can either produce alcohol or reduce the concentration of alcohol in alcohol-containing specimens. Without preservative, bacterial action at room temperature is likely, and alcohol production by bacteria can reach concentrations up to 0.10 to 0.15 g/dL. Sodium fluoride will also inhibit cholinesterase activity, which can metabolize cocaine and other ester-containing substances. Blood specimens held for long periods can be frozen. Tissues may be stored for short periods at 0 to 4°C. Most drugs are stable at this temperature range, but some drugs (e.g., cocaine) are labile. Tissues may also be frozen. However, the freezing and thawing of tissues before processing is not recommended, as this can rupture cells and cause a loss of drug-containing fluid. Where short- or long-term freezing is utilized, frozen tissues should be rapidly cut and weighed. Vitreous humor and urine can be stored at 0 to 4°C or frozen. Meconium stored at room temperature for 24 h shows a loss of drug, in particular cocaine and cannabinoids. Meconium can be stored at −15°C for nine months without loss of drug concentration (Ostrea, 2001). Samples to be analyzed for volatile substances other than alcohol must be stored frozen and, when appropriate, in hermetically sealed containers. If volatile substances are suspected in a death, the pathologist should submit duplicate samples so that analysis can be performed for nonvolatile drugs as well as for volatile compounds. It should be emphasized that blood and other fluids should be collected in such a manner as to minimize the headspace area in the collection vessel. All other biological-containing evidence should be sealed and stored frozen.

13.4 EXTRACTION PROCEDURES

Choice of the extraction procedures will determine the recovery of analytes from the tissue matrix, and improper selection may result in failure to recover adequately or even to detect one or more agents. The use of liquid–liquid extraction procedures requires careful selection of solvent and pH in relation to the analytes targeted. The polarity of the extracting solvent, or solvent mixture, should be similar to the polarity range of the target analytes. The pH used for the extraction of acidic and basic compounds should be 2 units below and above, respectively, the pK_a value of the analyte. Neutral drugs are not dependent on pH. However, when performing a general unknown analysis, the selection of solvent and pH should be done so as to obtain satisfactory recovery of a wide range of possible analytes, keeping in mind such factors as pH-related drug stability, drugs such as amphetamines that require very high basic pH, and the requirements of amphoteric compounds. The use of solid-phase extraction (SPE) generally requires protein-free extracts. SPE has a high versatility, owing to the broad range of columns available: polar, nonpolar, ion-exchange, and mixed-mode. Columns typically used for general unknown procedures consist of a combination of nonpolar and cation-exchange components. SPE can be very useful for the isolation of highly polar or water-soluble drugs and metabolites. It should also be noted that the condition or state of the biological matrix being extracted (i.e., “fresh”, decomposed, or embalmed) can affect the amount recovered (see below), and this needs to be taken into account when interpreting findings (Skopp, 2004). Experience has shown that many different types and classes of drugs and toxic metals can be recovered from exhumed bodies even after long periods of interment (Skopp, 2004). Great caution must be exercised, however, in the interpretation of all such findings.

13.5 ANALYTICAL TECHNIQUES

Knowledge of the basic theory, range of applicability to forensic samples, and the analytic limitations of all techniques currently utilized (i.e., immunoassay, chromatography, spectrometry, and spectroscopy) is essential. Immunoassays and chromatographic procedures are screening techniques, and all presumptive positives must be confirmed. The standard technique for confirmation is one or more of the modes of mass spectrometry, usually coupled with gas chromatography (GC) or liquid chromatography (LC).

It is beyond the scope of this chapter to discuss in detail the principles and applications of these analytic technologies. However, the following brief comments are relevant. There are ongoing advances in both GC and high-performance liquid chromatographic (HPLC) technology, resulting in the availability of more versatile and stable columns providing higher resolution and greater ranges of class retentivity.

The use of HPLC is gaining increasing popularity for several reasons. The development of mixed-mode or multimechanism columns for HPLC allows for the retention of drugs over a broad range of polarity. These columns incorporate both nonpolar and polar moieties. An analyst can choose from a multitude of normal, reversed-phase, and mixed-mode types of columns and, accordingly, tailor a separation. Compounds that do not chromatograph well on routinely used low- to intermediate-polarity GC columns because of high polarity or size or thermal instability will as a rule be effectively and conveniently separated by LC systems. In many cases this eliminates the need for derivatization required for GC. LC is essential for quaternary ammonium muscle relaxants such as tubocurarine and pancuronium. It has long been used for the acidic barbiturates (commonly analyzed along with neutral drugs as the acid/neutral fraction), and it has many advantages for the analysis of benzodiazepines (Levine, 2006; Marin et al., 2008). It is being applied increasingly to the routine analysis of basic drugs. Although still under development, the appearance of ultrafast or high speed HPLC does promise to increase the efficiency of obtaining effective and reliable separations with a very short run time. This involves the use of shorter columns, higher linear velocities, smaller porous particle size (ranging from less than 2 to 2.2 μm in size), and significantly higher operating pressures. It also generally requires cleaner sample extracts. LC is now commonly coupled to mass spectrometry by means of electrospray ionization or atmospheric pressure chemical ionization interfaces, and this is being used very effectively.

GC is, of course, still used very widely as a screening technique for basic drugs. Headspace GC is the standard method for routine analysis of the "volatile" fraction (i.e., ethanol, methanol, isopropanol, and acetone). A wide variety of columns of low, intermediate, and high polarity is available. As with LC, there is now available fast-track GC, which promises to bring effective separations in much shorter time. Fast GC typically involves the use of very rapid column heating systems, which provide much faster increases in temperature, higher carrier gas pressures, and very fast signal acquisition systems. Column length will vary, but 5-m columns are frequently used. Derivatization may be necessary for the following reasons: improvement of chromatographic characteristics, resolution of coeluting compounds, elimination of thermal instability, and in some cases, as with amphetamines, to produce a derivative with a much more characteristic mass spectral fragmentation pattern than is found with the underivatized compound. Where derivatization is necessary, this involves the replacement of "active" hydrogen with an alkyl, acyl (especially polyfluorinated acyls), or silyl group. A particular advantage to the use of polyfluorinated derivatives is that it allows for highly sensitive detection using an electron capture detector. One common application of this procedure is the gas chromatography–mass spectrometry analysis of cannabinoids.

There have been very significant developments in mass spectrometry technology which have dramatically increased its power and applicability for confirmation. The use of tandem MS, ramped fragmentation (smart fragmentation), simultaneous acquisition of both scan and select ion monitoring data, and the development of hybrid mass spectrometers have all expanded and improved the confirmatory process. Very fast mass scanning systems are available and required for use with fast chromatographic screening techniques. The use of current time-of-flight mass spectrometers can provide extremely accurate mass detection to differentiate compounds with similar masses.

Although generally no longer done routinely as part of a general unknown, the analysis of metals may be indicated based on the circumstances or suspicions associated with a particular case. Homicidal poisoning with certain metal salts (i.e., thallium or arsenic) is still encountered. The analysis of metals starts with digestion of the biological matrix. Wet ashing may involve refluxing the sample with a mixture of strong inorganic acids (i.e., a mixture of nitric, sulfuric, and perchloric acids), or for more volatile metals may involve just incubation with the acid solution. Dry ashing, which involves heating in a muffle furnace at elevated temperature to produce an ash residue, which is then treated with 5% nitric acid to dissolve the metals, cannot be used for mercury and other volatile metals and is used for selected metals only as appropriate. In general, mercury requires the use of cold vapor techniques. The state-of-the-art instrumental technique is inductively coupled plasma mass spectrometry, which provides very high sensitivity over the mass range of metals of forensic interest. However, graphite furnace atomic absorption spectroscopy (AAS) is still very widely used, and cold vapor AAS is the preferred technique for mercury.

The analysis of carboxyhemoglobin and the other hemoglobin species is today typically and conveniently achieved with the use of CO-oximeters, which involve automatic spectrophotometric determinations (Lee et al., 2002, 2003; Lewis et al., 2004). Thermocoagulated, putrified, and contaminated blood samples can pose potential problems. In such cases, sample pretreatment before oximetry has been found effective. Alternatively, the use of headspace GC on carbon monoxide released from the sample after treatment with liberating agents or the application of derivative spectroscopy to the sample itself can provide acceptable results (Perrigo and Joynt, 1989).

Cyanide may be determined conveniently by headspace GC using a nitrogen–phosphorus detector after liberation from the biological matrix under acidic conditions. Spectrophotometric and fluorometric procedures are also available (Felscher and Wulfmeyer, 1998; Gambaro et al., 2007).

It is essential that there be maintenance of all the accepted standards for quality assurance and quality control throughout the entire testing process from sample receipt and accession through all the analytic operations and continuing to sample storage, release, and disposal. This includes the generation and maintenance of all reports necessary to document every aspect of the processes. Proper maintenance logs must also be kept for all analytic instrumentation involved. If it should be necessary to convey test results verbally to an authorized party, then the results and any specific information given should be recorded appropriately to avoid any future misunderstandings.

13.6 INTERPRETATION

The interpretation of postmortem findings is becoming increasingly complex. This is due to such factors as postmortem redistribution, significant differences in individual rates of metabolism arising from genetically based polymorphic variations in metabolic enzymes, the type and condition of the specimen, drug interactions arising from the concurrent use of multiple drugs, and a particular person's tolerance for a specific drug or class of drugs and cross tolerance.

13.6.1 Postmortem Redistribution

The mechanisms and consequences of postmortem redistribution were discussed earlier. The effects of postmortem redistribution increase with increasing postmortem interval, but the mechanism involving unregulated flow of drug out of cells becomes significant at the moment at which death occurs. In general, it has been observed that cardiac blood has a higher concentration of drug than that of peripheral blood, although the magnitude of the change depends on the physiochemical properties of the drug, its volume of distribution, whether death occurred under equilibrium or nonequilibrium conditions, the organ concentrations, and the length of the postmortem interval. Variations in the cardiac/peripheral ratio can be very significant, even for a given drug (Dalpe-Scott et al., 1995; Anderson and Muto, 2000). The general importance of always including a sample of peripheral blood has been noted. If peripheral blood is not available, the interpretation of findings based only on cardiac blood is frequently perilous, and, in such situations tissue concentrations, especially liver, are necessary to minimize the possibility of an erroneous conclusion. Postmortem redistribution may also occur between organs, as high levels of drug from liver, lung, stomach, and heart are released to neighboring sites (Moriya and Hashimoto, 1999). Where there has been traumatic injury to the organs (i.e., laceration or rupture of the stomach, intestines, liver, etc.), there is a high potential for contamination from drug that may be present in the affected organ and now spills out into adjacent areas. Sampling of multiple sites and testing of different specimens may be necessary to resolve such issues.

In cases involving alcohol when blood is not available postmortem, blood alcohol levels can be estimated from correlation data derived from studies comparing alcohol levels in blood with those in other postmortem tissues. The state of alcohol absorption (pre- or post-absorptive) can be determined by analyzing the stomach contents for alcohol, and the results are then helpful in deciding which tissue to use for the correlation (Backer et al., 1980).

Because of the complex processes involved in postmortem redistribution, calculation from a single postmortem blood sample of a dose of drug taken by the deceased prior to death is unwise, and it is especially dangerous when using a single cardiac sample (Prouty and Anderson, 1990; Hilberg et al., 1999b). Back calculation of a dose from even a femoral blood concentration using a clinically derived plasma-based steady-state volume of distribution is strongly discouraged, as it can lead to an erroneous conclusion. The volume of distribution itself for a given drug can show wide interpersonal variation, and its use postmortem is contraindicated. Conversion of postmortem blood levels to time-of-death plasma or serum levels cannot be carried out accurately or reliably. Furthermore, in many cases (i.e., where an intentional or accidental overdose is involved), it cannot be assumed that a steady-state distribution existed at the time of death. The experienced toxicologist may, however, using data from different specimens, cautiously estimate what approximate minimum dose was taken or give a reasonable opinion as to whether "therapeutic" or overdose quantities were taken or administered.

13.6.2 Pharmacogenomics

The interpretation of postmortem drug levels as they relate to cause of death requires careful assessment of all the particulars of a case. Toxicologic analysis should consider genetic influences on the concentration of drug found. The field of pharmacogenomics attempts to explain a person's metabolic capabilities for drug metabolism in relation to the genotype. Poisoning, intentional or otherwise, may be the result of genetically influenced altered metabolism. A person's genetic makeup may also influence the transport of drugs across membranes and the interaction of drugs with receptors. All of these effects will play a role in attaining a concentration of drug in blood as well as the person's response to a given drug (Kupiec et al., 2006). The use of pharmacogenomics for understanding drug reactions and drug-related deaths has been termed *molecular autopsy* (Jannetto et al., 2002). A person's genetic makeup has already entered the clinical area and will be utilized increasingly in the future prior to prescribing medication to better maintain therapeutic levels.

There are many metabolic pathways for the metabolism of drugs and other xenobiotics. Basically, they are divided into phase I and phase II pathways. Phase I is responsible for oxidation, reduction, and hydrolysis reactions, preparing the metabolite for conjugation reactions by phase II enzymes. The final product is usually more water soluble than the parent drug and can easily be eliminated. Most phase I reactions are carried out by the cytochrome P450 (CYP450) monooxygenase enzyme system, located primarily in the liver but also less extensively in other tissues (Omiecinski et al. 1999). These heme-containing proteins are located in the smooth endoplasmic reticulum, and many of the isoenzymes exhibit genetic polymorphism. A list of the various isoenzymes and their substrates, along with inhibitors and inducers of the isoenzymes, is available (Flockhart, 2007). The CYP450 system may play a role in explaining drug-induced deaths by two different mechanisms.

First, based on a person's genomic makeup, one can be a "slow" or a "fast" metabolizer. Another classification of variants is ultrarapid, extensive, intermediate, or poor metabolizer. Those with the slow polymorphic form can accumulate high levels of drug that could result in death, even from therapeutic doses. About 25 to

30% of all xenobiotics are metabolized by one isoenzyme, CYP2D6 (debrisoquine hydroxylase); 5 to 10% of Caucasians have low levels of CYP2D6 or lack it completely, and doses of medications metabolized by this isoenzyme must be reduced (Kupiec et al., 2006; Roden et al., 2006). On the other hand, persons with multiple copies of the gene for CYP2D6 must be given higher doses to maintain therapeutic blood levels. The following examples will serve to illustrate the importance of such genetic variations.

Oxycodone, the opioid prescribed for relief of pain, is metabolized by CYP2D6. Three mutations of the polymorphic gene encoding for this isoenzyme have been identified and may play a role in oxycodone toxicity (Jannetto et al., 2002). Methadone, another opioid, is metabolized by CYP1A2, CYP3A4, and CYP2D6. Again, genetic variants of CYP2D6 showed a trend, although it is not statistically significant, toward poor metabolism of methadone, and higher levels were found in these people than in nonvariant controls (Wong et al., 2003). Fentanyl, another opioid used for relief of pain and as an anesthetic during surgery, is metabolized by CYP3A4 and CYP3A5. Variant alleles of the genes encoding these proteins are known and again may play a role in fentanyl toxicity (Jin et al., 2005). The response to the anticoagulant warfarin (coumadin) is dependent on cytochrome metabolism. Approximately 6% of the variation in response to warfarin is due to variation in the gene encoding CYP2C9, and 27% of the variation in response is due to variation in the gene VKORC1 encoding the warfarin target, vitamin K epoxide reductase (Berg, 2007). The metabolic rates of amphetamine and of methamphetamine are dependent on CYP2D6 activity, and genetic polymorphism for ring hydroxylation of amphetamine has been reported (Miranda-G et al., 2007).

Second, since different drugs can be metabolized by the same enzyme, taking two such drugs at the same time may lead to competition, with one drug being less metabolized than would occur had it been taken alone. This could result in an increase in drug concentration, leading to toxicity and even death. Such a result could also occur if one drug inhibits its own or another metabolic enzyme involved with other drugs. Such effects have been reported with the antidepressant selective serotonin reuptake inhibitors (SSRIs) that are both metabolized by CYP2D6 and inhibit CYP2D6 (Goeringer et al., 2000). The inhibitory effect of ethanol on the metabolism of certain other drugs, e.g. barbiturates when both are taken concurrently is well known. The inhibition of CYP3A4 by the antifungal agent ketoconazole results in terfenadine-induced cardiac toxicity. High concentrations of parent drug may therefore be due to a low rate of metabolism, to metabolic competition by other drugs, or to inhibition of the enzyme by one drug, allowing the second drug to increase in concentration. It must be remembered that frequently more than one CYP450 enzyme is involved in the metabolism of a given drug, and inhibition or induction of a nonprimary pathway may be significant.

13.6.3 Drug Interactions

Polydrug use can result in additive, synergistic, potentiating, inhibitory, or antagonistic interactions, and may involve inhibition or induction of metabolic enzymes by one agent affecting the clearance and effects of others. Some examples were noted earlier. Other common examples are the frequently encountered difficulties in interpretation of methadone and other drug abuse–related deaths (Milroy and Forrest, 2000; Corkery et al., 2004; Jönsson et al., 2007). The widespread use of all types of over-the-counter herbal products, domestic and imported, is an additional, potential complicating factor. Many of these products are known to contain components that interact with prescribed medications or may contain unlisted drugs or toxic substances (Ang-Lee et al., 2001; Saper et al., 2004; Hu et al., 2005; Dasgupta and Bernard, 2006; Woo, 2008). For example, some imported products may contain toxic licit medications which are not listed on the label. Other products have been shown to contain toxic metals. Many of these components may not be routinely detectable in standard analytical protocols. It is hoped that the more stringent regulations issued by the U.S. Food and Drug Administration in June 2007 concerning the importation of these products will reduce this danger, but toxicological vigilance is crucial (U.S. Food and Drug Administration, 2007).

13.6.4 Drug Stability and Decomposed Tissue

The postmortem chemical stability (or the lack thereof) of a drug or toxin or of the metabolites is a function of the combined effects of the chemical's structure, the length of the postmortem interval, the conditions of temperature, bacteria, oxygen, water, pH, and so on, to which the body has been subjected, and the particular matrix containing the agent. The reader is referred to the references cited for a comprehensive discussion of all these issues as they pertain to many different drugs (Grellner and Glenewinkel, 1997; Drummer, 2004, 2007; Skopp, 2004).

The interpretation of toxicological findings in decomposed, exhumed, or embalmed (discussed below) bodies requires special vigilance. Decomposed bodies reflect the processes of autolysis and putrefaction. Autolysis, digestion by natural enzymes, occurs more readily in tissues with a high enzyme content, such as the pancreas and stomach. Tissues with a low content of digestive enzymes, such as heart and liver, are less readily autolyzed. Autolysis leads to liquefaction of the tissues, which can then mix with blood and result in erroneous conclusions. Putrefaction, destruction of tissue by bacteria, depends on the bacteria that are present, the temperature and oxygen content of the environment in which the body is located, and the substrates available for bacterial action. The intestine is particularly susceptible to putrefaction and bacterial fermentation, and is not normally used for postmortem analysis.

Ethanol can be produced because of fermentation in decomposing bodies, but can also decrease from initial antemortem levels due to chemical, microbial, and evaporative processes occurring with decomposition (Kugelberg and Jones, 2007). Postmortem changes are known to cause a rapid loss of the high levels of cyanide associated with fatal poisoning, but cyanide is also subject to varying postmortem production, particularly in deteriorating bodies (Skopp, 2004). Artifactual production of cyanide has also been found to occur in some blood samples from fire death cases, and extreme caution must be exercised in the interpretation of such levels. Cyanide can be a significant component of some toxic fire gases that cause or contribute to death, but in such instances it must be reasonably shown that it derived from pyrolysis of nitrogenous material (e.g., nitrogen containing plastics) present in the fire environment.

Toxicologically significant concentrations of γ-hydroxybutyrate (GHB) may also be produced endogenously in postmortem fluids (Elliott, 2001, 2004; Moriya and Hashimoto, 2005; Beránková et al., 2006). GHB is a sedative-hypnotic agent not currently approved for routine use in the United States. It is, however, available from different sources, and is encountered both as a drug of abuse and as an agent in drug-facilitated sexual assault. Clearly, very serious misinterpretations can occur unless the potential ranges of postmortem production are taken into account.

The finding of a low or negative drug level in decomposed samples cannot be assumed to mean that no antemortem significance was involved. However, the presence in relatively intact or preserved exhumed specimens of significantly high levels of drug or toxin may be cautiously interpreted as being consistent with antemortem overdose, whatever was the circumstance of administration. In some cases where there are suspicions of the involvement of criminal poisoning, resulting in analysis of exhumed remains, the qualitative detection of a poison or of an agent not known to have been used by or prescribed to the deceased may be deemed sufficient to corroborate such allegations. This assumes, of course, that the idea may reasonably be eliminated that the source of any such agent was external contamination of exposed, buried remains or any of the other processes that may produce analytes, as discussed above. When dealing with buried but exposed remains or those in deteriorated coffins (often containing groundwater, soil, and other debris), appropriate controls from the areas in proximity to the remains should always be taken by the forensic investigator. An important example is when such remains are to be examined for metal content, where it is essential that control samples of soil and groundwater be tested.

13.6.5 Effects of Embalming Fluid

The effects of embalming fluid, which contains formaldehyde and small amounts of formic acid, on different classes and types of drugs is varied and complex (Rohrig, 1998; Cingolani et al., 2001–2005; Skopp, 2004). Many basic drugs (i.e., primary and secondary amines) are known to undergo varying degrees of methylation when the tissue containing them has been treated with formalin-containing solutions. Methylation will result in variable loss of the original compound as well as the appearance of a new methylated compound. In some cases the methylated compound may be the nonprescribed parent drug of the actual antemortem drug that has the demethylated structure. Primary amine drugs can be converted to secondary amines, and drugs containing a secondary amine can be methylated by formaldehyde to form *N*-methyl derivatives. This has been reported for the conversion of nortriptyline to amitriptyline, desipramine to imipramine, and fenfluramine to *N*-methylfenfluramine (Gannett et al., 2001). Sertraline was converted to *N*-methylsertraline (Suma and Sai Prakash, 2006). Methamphetamine converts to *N*,*N*-dimethylamphetamine. It may also be noted here that the analysis of formalin-fixed tissue should always be accompanied by the appropriate use of formalin-spiked calibrators. Caution must be exercised in general when interpreting findings from any formalin-fixed tissue matrix as opposed to data derived from "fresh" tissue. These chemical conversions may make it difficult to determine the original concentration of the parent drug. Nonetheless, a wide variety of drugs and toxins have been detected in exhumed bodies, and this is often a crucial factor in the investigation. Exhumed bodies often present a combination of problems, due to decomposition and the effects of embalming fluid. Occasionally, the toxicologist must analyze stored formalin-fixed specimens. One must consider that drugs from the fixed tissue will diffuse into the formalin, thereby decreasing the tissue concentration of drug.

CONCLUSION

The concentration of drug found at the time of sampling from postmortem specimens is the result of many factors, some relating to the host, some to postmortem changes, and some to the physicochemical properties of the analyte, and may not represent the concentration of drug at the time of death. Thus, a negative finding does not always mean that drugs were not involved; and a finding of high levels of drug does not always mean that an overdose occurred. The decedent may have been tolerant of the drug in question and may have required or been using higher doses. Analysis and comparison of drug levels in both blood and tissue (e.g., liver or brain) may be necessary and should be done whenever possible for accurate interpretation of quantitative findings. Finding therapeutic levels of drug may be the result of drug loss due to the factors listed above. Where postmortem exposure to formalin-containing solutions has occurred, finding parent drug may be the result of metabolites having been converted back to the parent drug. Estimating the dose of drug consumed from postmortem data is often unreliable, as discussed above. Also, finding several drugs raises the issue of additive or synergistic effects or drug interactions in causing death. Clearly, the longer the delay between the time of death and the taking of samples, the less accurate may be the interpretation of the role the drugs played in causing the death.

In summary, the forensic toxicologist is presented with a broad spectrum of case findings, ranging from those that point readily toward a particular opinion concerning whether or not one or more substances had a role in the cause of death, to those where the complexity or circumstances do not allow for a definitive interpretation or conclusion. In the latter case, the toxicologist must always resist pressure from those seeking to obtain an opinion not supported by the facts. Where there is uncertainty in interpretation, it is the obligation of the toxicologist to make known to the inquiring party the scientific basis for the uncertainty and all possible conclusions. The toxicologist must never forget that the role of the forensic scientist is first to obtain the most reliable data possible and then, accurately and impartially, interpret and educate the participants in the medicolegal process.

CHAPTER 14

Entomotoxicology: Drugs, Toxins, and Insects

JASON H. BYRD
University of Florida, College of Medicine, Gainesville, Florida

MICHELLE R. PEACE
Department of Forensic Science, Virginia Commonwealth University, Richmond, Virginia

Summary Forensic entomology is gaining widespread acceptance within the forensic sciences as one method to estimate a portion of the postmortem interval by utilizing the time of insect colonization of the body, also known as the period of insect activity. Additionally, insect evidence can be utilized as alternative toxicology samples in cases where no other viable specimens exist. This subfield, known as entomotoxicology, can provide useful qualitative information to investigators as to the presence of drugs in the tissues at the time of larval feeding. The presence of drugs can also alter the developmental period of the insects and should always be taken into consideration by the forensic entomologist. In this chapter we examine the relationship between toxicology and forensic entomology.

14.1 Introduction 484
14.2 The fly and forensic science 484
  14.2.1 History of forensic entomology, toxicology, and the rise of entomotoxicology 485
  14.2.2 Drugs and the fly life cycle 488
  14.2.3 Why use insects as a toxicological specimen? 490
  14.2.4 Drug extraction methods 492
  14.2.5 Qualitative versus quantitative 493
  14.2.6 Changes in insect development: toxins and drugs 494
  14.2.7 The future of entomotoxicology 494
References 495

Forensic Chemistry Handbook, First Edition. Edited by Lawrence Kobilinsky.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

14.1 INTRODUCTION

The use of entomological evidence in a legal investigation can provide extraordinarily valuable information that may otherwise be overlooked. Most often, the utility of an insect in a forensic investigation is to assist investigators in establishing the postmortem interval. More specifically, the period of the postmortem interval is often termed the *period of insect activity* (PIA) or the *time of colonization* (TOC) and can be thought of as the minimum postmortem interval. It is often expressed as the minimum time period necessary to allow the insects collected to develop to the life stage in which they were discovered.

Several different approaches exist in making this determination; they range from insect succession to estimations based on larval length, larval weight, and the accumulated degree hour (or degree day) methodology. Each of these techniques has advantages and disadvantages. One of the most common issues in forensic entomology is that the investigator is not able to determine exactly what the climate conditions were at the death scene during the entire postmortem period. Since insect growth and development are heavily influenced by ambient temperatures, the difficulty in accounting for the conditions at the scene prior to the discovery introduces some uncertainty in application. Thus, the forensic entomologist must account for this variability when providing estimations of the time since colonization. If such variability is addressed properly, forensic entomology can be applied with an impressive degree of accuracy and precision.

Generally, this variability is accounted for by the extrapolation of meteorological recording station data from nearby locations. However, other factors can play a role in the influence and alteration of larval insect growth. One of these factors, often unaccounted for by the forensic entomologist, is the presence of various metabolites of illicit drugs and poisons within the insect that can alter its expected time of development. Such chemical compounds, if present in the host tissue and consumed by the larvae, may alter insect development by hours or days.

The term *Entomotoxicology* was first used in 1994 to describe the interface of entomology, forensic science, and toxicology, and we are only now beginning to understand how various drugs and toxins interact with a developing insect to alter the rate of development of various life stages. In its relatively short history, the discipline of entomotoxicology has been very broadly applied. The deposition of metals within the larvae has assisted in pinpointing human remains to a possible point of origin, or to corroborate or refute suspect and witness statements as to the activities of the victim prior to death. However, the most common utilization has been the detection of various chemical compounds from the larvae long after suitable soft tissues, blood, and urine from the body have decomposed or been consumed by the insects present (Nolte et al., 1992) (Table 14.1).

14.2 THE FLY AND FORENSIC SCIENCE

For many millennia, humans have been aware of, known, and understood the role of flies in the ecosystem. The first mention of the flies most commonly used in

TABLE 14.1 Drug Compounds Recovered from Calliphoridae and Sarcophagidae Larvae

Calliphoridae	Sarcophagidae
Alimenmazine	Amitrptyline
Amitrptyline	Barbiturates
Arsenic	Cocaine (benzoylecgonine)
Barbiturates	Heroin
Bromazepam	Malathion
Clomipramine	Methamphetamines
Cocaine (benzoylecgonine)	Mercury
Laudanum	phencyclidine
Levomepromazine	3,4-Methylenedioxymethamphetamine
Mercury	
Morphine	
Oxazepam	
Phenobarbital	
Triazolam	

present-day criminal investigations was in the fourteenth tablet of Harra-Hubulla, a list of wild animals from the time of Hammurabi 3600 years ago. The Egyptians also keenly understood the metamorphosis of flies, as indicated by the Papyrus Gizeh No. 18026:4:14 (Cairo, Egyptian Museum CG 58009), found in the mouth of a mummy, which bore the inscription: “The maggots will not turn into flies within you (Golénischeff 1927).”

The blowflies (Calliphoridae), flesh flies (Sarcophagidae), and house flies (Muscidae) are the most forensically important flies. These flies are most often the first insects to find and colonize remains. The blowfly can travel up to 20 km a day in search of an appropriate food source. A female can lay several batches of about 250 eggs each, depending on her size and health. Once a female has begun oviposition, other females will also lay in the same area, resulting in a mass of several thousand eggs. The eggs usually hatch within 24 hours and immature larvae penetrate the interior of the remains to avoid desiccation and predation. As the larvae feed and mature, they pass through three stadia: first, second, and third instars. In doing so, their feeding mass can reach temperatures of 50°C. Once the larvae stop feeding, they wander from the food source to find a suitable place to pupate, during which the soft-bodied larvae become shorter, wider, rigid, and dark. Once metamorphosis is complete, the adult flies will emerge from the casing, inflate their wings, and become sclerotized (Figure 14.1). (Kamal, 1958; Greenberg, 1991).

14.2.1 History of Forensic Entomology, Toxicology, and the Rise of Entomotoxicology

The *Washing Away of Wrongs*, a thirteenth-century Chinese manual on forensic and legal medicine, considered the oldest in existence from any civilization, detailed

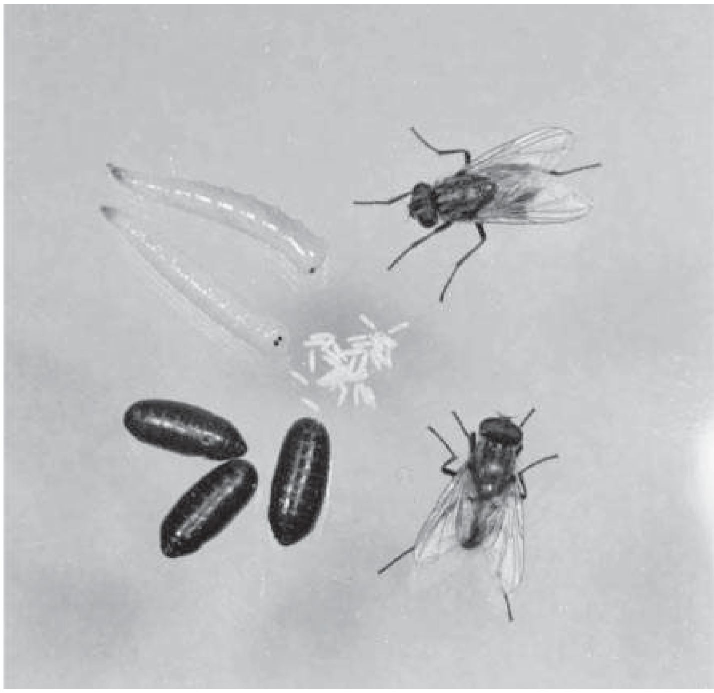

Figure 14.1 Life cycle of *Musca domestica*, showing complete metamorphosis. These life stages are typical of all Muscidae and Calliphoridae. (Photo courtesy of Clemson University; U.S. Department of Agriculture Cooperative Extension Slide Series.) *(See insert for color representation.)*

the death investigation of a man with a wound apparently inflicted by a sickle. The investigator gathered the local farmers with their sickles. In the hot afternoon sun, flies gathered on one sickle, indicating blood residue and incriminating the owner of that sickle. After interrogation and in light of the evidence, the owner of the sickle “knocked his head on the ground and confessed” to the murder (McKnight, 1981). In the late nineteenth century, scientists began to study and characterize the activities of arthropods and their application to death investigations. The first observation of insects on human remains in which the death was attributed to arsenic poisoning was in 1890. F. M. Webster noted that *Conicera* sp. were noted to have developed on remains buried for over two years. Analysis of the stomach contents proved that the poisoning was via arsenic. He noted that it was not surprising that these flies could develop on tissues laced with arsenic because “they are doubtless tenacious of life.” Unfortunately, it is not apparent that he foresaw the forensic science application of their presence (Webster, 1890). In 1894, the French entomologist Mégnin pioneered the use of forensic entomology by painstakingly outlining and describing the eight successive waves of arthropod invasions on a corpse (Nuorteva, 1977; Keh, 1985; Smith, 1986). Those eight stages are still commonly recognized today.

Nonetheless, major issues still arose for forensic entomologists. Wide variability in geographical, topological, and environmental conditions led to even more widely variable insect species and behaviors for any given situation. These variable conditions made a well-characterized foundation of workable information difficult to attain. Differences between the entomologists' experimental conditions and actual forensic cases also compounded the problem of making useful correlations. Ultimately, a more thorough understanding of the basic principles as well as the limitations of using insects in forensic investigations was the result of the corroborative work of pathologists, biologists, detectives, and entomologists. These collaborative works had a significant impact in a variety of cases:

- Estimation of postmortem interval (PMI) by determining the life-cycle stage of the insects present on a body in conjunction with the stage of faunal procession (Bergeret, 1955; Nuorteva, 1977; Hall et al., 1986; Smith, 1986; Catts and Goff, 1992; Catts, 1992; Hall and Doisy, 1993).
- Estimation of PMI from the degree of insect colonization resulting from the degree of body exposure (totally exposed, partially covered, or completely buried) (Nuorteva, 1977; Smith, 1986).
- Acquittal of a person based on larval migration data (Nuorteva, 1977).
- Determination that a body had been moved from the original crime scene by identifying insects on that body not indigenous to the discovery area, thereby determining the area from which it was moved (Benecke, 1998).
- Linkage of suspects to the scene of a crime due to insect bites found on their bodies from insects specific to the vicinity where the body was discovered (Prichard et al., 1986).
- Determination of child abuse, neglect of the elderly, or disease or infection state from insect colonization prior to death (Goff, 1991; Benecke, 1998; Benecke and Lessig, 2001).
- Individualization of human DNA from insect gut contents (Lord et al., 1998).

Interestingly, Mathiew Orfila, a pathologist who is often credited as the father of modern toxicology, can potentially be credited with the first systematic observation of insect succession in a human cadaver (Greenberg, 1991). He documented 30 arthropods from a corpse and noted that live insects were recovered from bodies that were exhumed for several years. Bergeret (1955) is credited with the first case in which entomological evidence was utilized to create a time line. However, it turns out his assessment of the entomological evidence may have been incorrect; the period of insect activity could have been much shorter. Citations for the application of entomology to toxicological analysis did not begin to appear until the early 1970s. Nuorteva described a case in which the mercury concentration in larvae feeding on a corpse was used to determine if the deceased had lived in an area with high mercury pollution (Nuorteva, 1977). Subsequently, the deposition of copper, iron, and zinc and their effects on adult flies and larvae was demonstrated, and the bioaccumulation and effect of mercury on larvae was fully characterized (Sohal and Lamb, 1977; Nuorteva and Nuorteva, 1982).

The first examination of fly larvae from human remains for qualitative identification of drugs was proposed by Beyer et al. in 1980. In this case, Calliphoridae larvae were used to assist investigators in proving a suicide via an overdose of pentobarbital (Beyer et al., 1980). In this case fly larvae were used to determine the source of poisoning in a severely decomposed body that had no other source for a toxicological specimen (Beyer et al., 1980). Gunatilake and Goff (1989) utilized larvae collected from a body to aid in proving that a case involved poisoning by ingestion of malathion. This chemical was detected in the larvae as well as from body fat and gastric contents.

In 1990, Pascal Kintz published a landmark study from a 1987 case illustrating a "new" toxicological method of investigation which demonstrated the usefulness of entomological evidence. In this case example, larval concentrations of five drugs (triazolam, oxazepam, phenobarbital, alimenmazine, and clomipramine) were compared to concentrations in human heart, liver, lung, spleen, and kidney tissue as well as bile. In all cases, the drugs known to be in the human tissue were recovered form the larvae. One year later, in 1988, Kintz reported a case in which a severely decomposed corpse was found and insects were utilized as an alternative toxicological specimen because no other suitable tissue was available (Kintz et al., 1990).

Introna et al. (1990) collected livers from 40 decedents known to be positive for opiates (morphine) and colonized the tissues with eggs of *Calliphora vicina*. Upon analysis of the larvae and tissue via radioimmunoassay, the opiate concentration in the larvae correlated with that of the tissues on which they fed. Subsequently, investigations in entomology and toxicology illustrated the two important areas in forensic entomotoxicology: the effects of drugs on insect life cycles and the use of insects as an alternative toxicological sample (Goff et al., 1989, 1994).

The potential growth of entomotoxicology in casework can be moderately predicted based on the statistics that illustrate the increase in the abuse of illicit and prescription drugs in the past decade. The past decade has seen an increase in the abuse of illicit and prescription drugs. In a three-year time span alone (2000–2002), use of illicit drugs has increased from 6.3% to 8.3% in persons older than 12 years (NHS Survey, 2002). In 2000, 43% of emergency room visits were due to the misuse of prescription drugs (SAMHA, 2003). Not only have drug-related emergency room episodes and subsequent deaths increased (SAMHA 2003; DAWN, 2003), but a clear correlation exists among crime, violence, and drug usage. Of the approximate 14,000 homicides in the United States in 1998, 4.8% of them were murders committed during a narcotics felony (ONDCP, 2000). In the end, with this growth in potential must come research to support widespread acceptance and understanding of the expansive scope of applications and utility.

14.2.2 Drugs and the Fly Life Cycle

Sometimes, entomological evidence is used in death investigations as a tool to better understand the circumstances leading to death. Insects can be used to date a crime scene or estimate a PMI by applying the known developmental rates of insects and the successional patterns of different insect species on the remains (Smith, 1986;

Goff et al. 1988; Catts and Goff, 1992; Goff, 1993; Catts, 1992; Goff and Lord, 1994). As mentioned previously, the PMI estimation is an approximation of the time of colonization of the cadaver by insects. The succession of insects on carrion can be complicated by many geographical and environmental conditions, such as plant cover, water covering the body, inclement weather, and/or temperature (Nuorteva, 1977; Smith, 1986; Goff et al., 1988; Catts and Goff, 1992; Schoenly et al., 1992, 1996; Goff, 1993; Hall and Doisy, 1993; Catts, 1992; LaMotte and Wells, 2000). The time required for an insect to reach any developmental stage at certain environmental conditions has been well characterized, particularly for forensically important insects (Kamal, 1958; Greenberg, 1990a,b; Goff, 1991; Greenberg, 1991; Byrd and Butler, 1997, 1998; Anderson, 2000; Byrd and Allen, 2001). The time of colonization can be estimated using either of these techniques singularly or in combination. However, it is most common to use the stage of development in conjunction with meteorological data. This technique is most frequently referred to as the *accumulated degree method*. This method compares the temperature with the time of development and is often used to estimate the time of colonization and subsequent minimum postmortem interval.

Additionally, the stage of development or age can be determined by measuring the insect larvae's total length, crop length, or weight. Weight and crop length are the least useful estimates of age because broad variations result from the amount of food available, changes in the physiology of the developing larvae, and the lack of a well-defined informational database (Greenberg 1990a, 1991; Wells and LaMotte, 1995). The methods that utilize larval length and/or accumulated degree hours (which can incorporate larval length) are the most commonly used, particularly since an appreciable database has been developed (Greenberg, 1991; Goff, 1993; Byrd and Allen, 2001; Byrd and Castner, 2001).

Before Sohal and Nuorteva ushered in studies to investigate the effects of drugs on insects, Introna reported a solitary investigation in Japan in 1958 which found that flies were attracted to carrion differently depending on the poison ingested (Introna et al., 2001). Since then, Goff and Bourel have made strides in understanding the deposition and affects of drugs and toxins on forensically important insects. In 1989, Goff worked a case with Gunatilake in which a man expired from malathion intoxication. The man had been missing eight days, but the entomology-based PMI estimation was five days. Goff and Gunatilake concluded that the malathion had shortened the duration of the larval stage and that the pesticide retarded the succession of insects on the body (Gunatilake and Goff, 1989). Goff has also explored the effects of amitriptyline, cocaine, heroin, phencyclidine, methamphetamine, and 3,4-methylenedioxymethamphetamine (MDMA) on fly larvae. He found that amitriptyline had no effect on the rate of development of *Parasarcophaga ruficornis* larvae but prolonged the migratory stage. Amitriptyline also caused the body lengths and weights to be greater than those of the control larvae (Goff et al., 1993). Cocaine shortened total developmental times for *Boettcherisca peregrine* larvae and puparia. He found that heroin accelerated the growth rate of the *B. peregrine* larvae and that phencyclidine did not affect the rate of development for *P. ruficornis* larvae but shortened the migratory phase. Finally, Goff concluded that methamphetamine and MDMA generally shortened the duration of the feeding larval stage for *P. ruficornis*, and methamphetamine caused the larvae to grow significantly smaller than the control (Goff et al., 1989, 1992, 1994, 1997). Interestingly, Bourel found that morphine, a drug structurally similar to heroin, slowed the developmental rate of larvae for *Lucilia sericata* (Bourel et al., 1999). Ultimately, these results indicated three potentially significant points: (1) that drugs can affect developmental rates, contrary to expectations; (2) that the same drugs may affect different fly species differently; and (3) that PMI estimations could potentially be miscalculated if drug effects are not taken into account.

14.2.3 Why Use Insects as a Toxicological Specimen?

Invariably, some death investigations involve human remains that are severely decomposed or need to be exhumed after embalming and burial. Typical postmortem specimens for a toxicological analysis include brain, liver, kidney, heart blood, peripheral blood, vitreous humor, bile, urine, and gastric contents (Poklis et al., 1998; Levine, 1999). The tissue alteration and breakdown that results during the process of decomposition can make some toxicological testing methods more challenging. In some cases, the tissue samples are rendered unsuitable for analysis, or they are simply no longer present in the body. The most common problem in warm and humid climates is that no soft tissue remains on which to perform a toxicological analysis (Levine et al., 2000). Additionally, when a body is embalmed, the tissues are fixed, and blood and urine are typically no longer available. Toxicological analyses on embalmed tissue can and will result in drug concentrations that are difficult to interpret (Hanzlick, 1994; Alunni-Perret et al., 2003). Therefore, in these extreme cases, alternative specimens often need to be examined.

Properly preserved or freshly collected larvae can be treated as any other tissue for toxicological analysis (Beyer et al., 1980; Poklis et al., 1998; Levine et al., 2000). Two important premises underscore the utility of entomotoxicology.

1. When larvae feed, they shred the food source with mouth hooks and deposit the meal into a crop, which generally "reflects the adaptive imperative to eat now and digest later" (Greenberg, 1991).
2. Any drug detected in larvae could only have come from the body on which it was feeding (Beyer et al., 1980).

Beyer's case was a landmark in forensic entomotoxicology as the first case that used larvae as the sole source for a toxicological analysis. Toxicological analyses were performed on larvae found on a skeletonized body after the badly decomposed soft tissue was deemed unsuitable for analyses. Phenobarbital was identified in the larvae using thin-layer chromatography and gas chromatography–mass spectrometry (Beyer et al., 1980). In 1987, Kintz also used larvae as a toxicological specimen in a case of completely putrified remains. Despite the severe decomposition, organs were identified and analyzed for drugs. The concentrations of drugs (triazolam, oxazepam, phenobarbital, alimemazine, and clomipramine) in the larvae were compared to the concentrations in the body's bile, heart, liver, lung, and spleen. Kintz stated that they could not establish a correlation between the larval drug concentrations and organ concentrations. However, he concluded that "the application of the toxicological investigations in maggots will surely increase" and that the performance of these analyses on living material "is always more suitable" (Kintz et al., 1990). In 1989, larvae were used to detect the organophosphate suspected as the cause of death of a severely decomposed man (Gunatilake and Goff, 1989). In 1990, Introna reported results from the first systematic study in entomotoxicology in which he found a direct and significant correlation between the concentration of morphine in the food source and the larvae (Introna et al., 1990). In 1992, Nolte investigated a case in which he collected fly larvae and skeletal muscles from an almost completely skeletonized corpse for toxicological analysis. This was the first reported case in which cocaine and benzoylecgonine were recovered from fly larvae. Nolte concluded that the larvae were better specimens for toxicological analyses because the skeletal muscle contained tissue decomposition products which interfered with the analyses (Nolte et al., 1992). It is important to understand that the half-life of cocaine in human tissue is 0.7 to 1.5 h, depending on the route of administration (Baselt, 2004). Therefore, Nolte concluded that the detection of cocaine and its metabolites in larval tissue suggests the use of cocaine in the immediate hours before death. However, while finding that both the parent drug and its metabolite in the larvae could indicate drug usage immediately prior to death, other contributing factors, such as route of administration, dosage, and postmortem redistribution, will impinge on this conclusion, making it difficult or impossible to interpret the findings of drugs in larvae. During Goff's studies to better understand the effects of drugs on the insect life cycle, he found a direct correlation between the concentration of the drugs heroin, methamphetamine, and MDMA in the food source and the larvae. On the other hand, he observed no correlation between the two with phencyclidine (Goff et al., 1991, 1992, 1994, 1997).

Since the late 1990s, several attempts have been made to understand the implications of larval drug concentrations. Wilson and Sadler found that the larvae eliminated amitriptyline, nortriptyline, trimipramine, and trazodone and that the ratios of larvae to food source drug concentrations were widely variable (Wilson et al., 1993; Sadler et al., 1995, 1997a). Sadler also investigated barbiturates and other analgesics with variable results. He found that acetaminophen was eliminated from the larvae and that the ratios of larvae to food source drug concentrations were also widely variable for the barbiturates amphetamine and sodium salicylate. Ultimately, he concluded that his solid-phase drug extraction procedure was not rigorous enough (Sadler et al., 1997b). Most recently, Hédouin performed a whole animal study and found a correlation between mature larval morphine concentrations and the tissues (Hédouin et al., 1999, 2001).

A case in Maryland reported in 2000 underscores the growing importance of more thoroughly understanding the utility of fly larvae as an alternative toxicological specimen. In this case, larvae and the only soft tissue (a portion of calf muscle) from mostly skeletonized remains were analyzed for drugs. Secobarbital was found in the larvae but not in the skeletal muscle. If a large amount of drug was ingested which led immediately to death, the drug had insufficient time to distribute to the calf muscle in any appreciable concentration. Second, the presence of decompositional fluids in the muscle potentially complicated the analysis. Had the larvae not been available for analysis, the prescription bottle found near the deceased would have only served as circumstantial evidence (Levine et al., 2000).

14.2.4 Drug Extraction Methods

If using entomological evidence as a toxicological specimen, larvae and pupae can be treated similarly to human specimens. Standard operating procedures for drug extractions from human tissue and fluid in a toxicology laboratory are a reasonable starting point for elucidating an optimum drug extraction procedure for what can be considered a fairly novel and certainly different biological matrix, insects. However, an extraction procedure for any given drug will probably have to be modified to compensate for the entomological matrix. For example, Sadler admits that he applied methods routinely used in his laboratory to extract barbiturates from urine to extract the same from larvae, but conceded that his inconclusive results were due to needing a more rigorous drug extraction technique.

When extracting drugs from larvae, the first consideration should be the large fat body of the larvae (Downer, 1981; Gullan and Cranston, 1994; Elzinga, 1999). Lipophilic drugs can potentially be sequestered in this fat body, requiring extraction solvents that are more lipophilic and back-extractions to adequately separate the matrix from the drug. Additionally, in order to actually extract the drugs from the larval matrix (the fat body, the hemocoel, and hemolymph), the pK_a of the drug must be considered carefully using the Henderson−Hasselbach equation for acidic drugs,

$$\mathrm{pH} = \mathrm{p}K_a + \log \frac{[\mathrm{B}]}{[\mathrm{BH}^+]}$$

and for basic drugs (Levine, 1999),

$$\mathrm{pH} = \mathrm{p}K_a + \log \frac{[\mathrm{A}^-]}{[\mathrm{HA}]}$$

Assuming that a drug will be mostly ionized in an aqueous-based biological fluid, the initial pH of the aqueous fluid should be adjusted so that the drug is no longer ionized. Once the drug is deionized, it can be sequestered into a hydrophobic solvent, thereby effectively extracting it from the biological matrix. Some cellular components in the matrix may also become sequestered in the hydrophobic extraction solvent (particularly the fat body), requiring a back extraction or a “cleanup” extraction. The cleanup extraction can be executed by changing the pH to drive the drug back into an aqueous layer, leaving hydrophobic cellular components behind in the solvent. Since the drugs need to be in an organic solvent for instrumental analysis, a final pH adjustment will deionize the drugs, allowing their final sequestering in a hydrophobic solvent.

14.2.5 Qualitative Versus Quantitative

Several studies have demonstrated the correlation between the dose and the concentration of drug within the larval body. However, researchers have differed greatly in their interpretation of the meaning of this correlation. In application, the quantitative evaluation of drug concentration in larvae in actual forensic casework is quite limited. One of the primary reasons for this limitation is because of the difficulty in determining exactly where the larvae have fed, since the absorption and distribution of a drug will vary among skeletal muscle and organ tissues. Thus, the location of larval feeding is of critical importance. However, due to the larval movement on the body while feeding, it is difficult (if not impossible) to determine on which tissues the larvae have fed. Although it is possible for an investigator to document the location of larvae as they are collected, it is not known which tissue or tissues were utilized as a food source prior to collection.

Although the rates of drug deposition in the larval body are not known for most species, most commonly, homogenates of the entire larval body are used. It is not known if the drugs and their metabolites are localized in the fat bodies, puparium, or chitinous exoskeleton of the adult (Nolte, 1992). However, much of the published research on entomotoxicology indicates that ingested chemicals can be recovered from the soft tissues of the larvae, pupae, and chitinous exterior of the pupal case. As a result, initial studies on the application of entomotoxicology to casework were promising. The window of opportunity for the detection of chemical compounds in human remains could be extended for days or even weeks by utilizing larvae and puparia. If the chitinous pupa stage had formed, chemicals ingested by the feeding larvae could be detected in the puparial casings years after the insect itself had completed development. In one particular death investigation case, the authors detected the metabolites of cocaine from the puparium two years after the death of the victim.

However, the quantitative use of insects as toxicological specimens in actual casework is problematic. Tracqui et al. (2004) analyzed 29 cases in which organic compounds were known to be present. These compounds were detected in both the human tissue and the insect body. However, no correlation between the drug concentration in the human versus larval samples were noted. Additionally, it was found that interlarvae and intersite variations in drug concentrations were not reproduced between samples. Therefore, arthropod larvae are, at best, most appropriate for qualitative analyses in a toxicological investigation and unreliable and unsuitable as a toxicological sample used to interpret the episode of intoxication prior to death based on a quantitative analysis.

The method of calculating the accumulated degree hours to estimate the time since colonization is as follows (Byrd and Castner, 2001):

1. Accumulated degree hours are calculated for the crime scene using temperature data gathered for a specified period of time prior to the collection of larval evidence.

$$\text{accumulated degree hours} = \text{ADH} = (T - B)(t)$$

where T is the mean temperature (°C)† B the minimum threshold development temperature and t the time in hours.

2. Larval evidence is measured (mm) and compared to species developmental charts (larval length vs. time, or instar molt vs. time) generated at specific temperatures to obtain the ADH required for that size.
3. The ADH required, which is given in the developmental chart, is compared to the ADH calculated at the scene to age the larvae and "date" the scene.

14.2.6 Changes in Insect Development: Toxins and Drugs

Nourteva and Nourteva (1982) were the first to demonstrate a negative effect on larvae from toxic compounds with the larval food source. However, Goff et al. (1989) was the first to show an altered developmental rate for larvae feeding on tissues containing illicit drugs. Goff's study was the first to show that cocaine, metabolized in tissue and ingested by larvae, would shorten the developmental time expected. Subsequent studies by Goff et al. (1993, 1994) have demonstrated significant differences in rates of development for *P. ruficornis* and *B. peregrina* (Diptera: Sarcophagidae). These developmental changes potentially can be of critical importance when estimating the time since colonization. Few studies have been conducted on this effect for the Calliphoridae, which are typically the predominate flies on human remains in the early postmortem period.

With this said, the alteration of insect development by drugs and toxins is problematic on several levels. Often, the forensic entomologist is not aware of the postmortem toxicology results and does not know to take drug-induced alterations into consideration when developing a window for time of colonization. Second, a thorough investigation of all major drugs has not been undertaken with every species of forensically important fly. Therefore, the entomologist has no foundation on which to make adjustments in the time of colonization range. Most important, while some studies have suggested that the developmental rate of some larvae is affected by some drugs, the interpretation of the data must be considered carefully. It is possible that alteration in the life cycle due to drugs or toxins on board is well within the standard error in life-cycle development of drug-free larvae.

14.2.7 The Future of Entomotoxicology

Considered generally, the published literature is only a preliminary treatise of the interface between entomology, toxicology, and medicolegal death investigation. More research is needed on drug accumulation and retention in insect physiology. However, it is clear that the use of entomotoxicology as a quantitative measure is not yet suitable for application in forensic casework. Such a circumstance is unfortunate because the technical application of the premise works well, as the amount of drug compound within the larval body can be determined with great accuracy and precision. Using larvae as a toxicological specimen is readily put into practice because the analysis takes no specialized equipment beyond what is necessary for a standard toxicology analysis for human or animal tissue. Therefore, most any postmortem forensic toxicology laboratory possesses the equipment necessary for the testing. From the technical viewpoint, entomotoxicology looks promising.

---

† Collected from a nearby weather station. If provided, hourly temperatures may be gathered to calculate ADH and provide a defined window for age assessment of the larvae.

However, it is the biology and behavior of the insect itself that may limit its application in forensic investigations. Entomologists must first conduct fundamental research on larval feeding and aggregation behavior. Currently, entomologists do not fully understand larval movement and distribution within the maggot mass, which often forms during the actively feeding third instar. After simple visual observation of the larval feeding and aggregation behavior, it is obvious that the larvae are in constant motion. Individual larvae seem to spend time in the center as well as the periphery of the mass. However, the movement of a single larva within the mass has not been tracked with great accuracy, and no studies have been published on the positional effect on the intensity of larval feeding. With standard entomological procedures currently in practice, it is unknown on what tissues the collected larvae have fed.

The absorption, distribution, metabolism, and elimination of drugs are well documented for most drugs and provides for the understanding of postmortem tissue distributions. However, since larval feeding locations are unknown in applied casework, the resulting larval concentration proves meaningless under current collection and analytical protocols. Insect larvae do accumulate drugs and toxins within bodies as a result of their feeding. Therefore, the qualitative analysis of these compounds may provide valuable information to forensic investigators.

Forensic entomologists are able to provide vast amounts of information to other forensic investigators based on the comparatively limited nature of currently published data sets. Biological variation may lead to cautious, conservative, and “incomplete” answers to many commonly posed questions from investigators. A tremendous amount of research must be conducted to better understand the patterns and implications of natural variables, particularly as they affect forensic investigations in the entire scope of forensic entomology, but specifically with regard to toxicology. As the knowledge base expands and more becomes known about larval behavior, entomotoxicology may one day become a valuable quantitative tool.

REFERENCES

Alunni-Perret, V., Kintz, P., Ludes, B., Ohayon, P., and Quatrehomme, G. (2003). Determination of heroin after embalmment. *Forensic Sci. Int.*, **134(1)**:36–39.

Anderson, G. S. (2000). Minimum and maximum development rates of some forensically important Calliphoridae (Diptera). *J. Forensic Sci.*, **45(4)**:824–832.

Baselt, R.C. (2004). *Disposition of Toxic Drugs and Chemicals*. 7th edition, Foster City, CA: Biomedical Publications.

Benecke, M. (1998). Six forensic entomology cases: description and commentary. *J. Forensic Sci.*, **43**:797–805.

DIY – MAKE YOUR OWN ORGONITE

Making orgonite is easy!

This is the most important thing I have to tell you in this chapter.

A lot of people on the internet and elsewhere are spending a lot of time and money to make you believe it's otherwise.

Orgonite always works, even if you think you got it all wrong. Also the designs of the basic orgonite tools are very error tolerant. When you read about the mix having to be 50/50, that just means that you pour a bucket full of resin into a bucket full of aluminium filings. People have asked me the weirdest question about how to determine the exact volume of the filings since they come in this woolly state and….

It's much easier than that!

Making beautiful orgonite takes more practice and is not that easy, but as much as we all love to make things beautiful, a rough and dirty TB will neutralise a cell phone tower (death force transmitter) just as well as a high gloss one with neatly arranged gemstone inlays. At first you will need to source resin. Most of us work with simple polyester resin. This comes in different qualities. Preferable is the "clear casting" quality that is also used to embed butterflies and dried flowers into nice clear blocks, often used as key rings or for other decorative purposes. We buy this from industrial wholesalers who cater for the boat building and swimming pool manufacturers. If you have no idea where to get resin in your town, phone the guys who work with fibreglass and ask them where they buy their stuff. Polyester resin stinks a bit when it's curing because of the chemical reaction and release of solvents, so you should always work outside or in a very well ventilated room.

If you plan to do this regularly, get yourself a protective mask (against organic fumes like for spray painters) or build a fan into your room that guarantees at least a 40 x per hour air change.

The resin needs to be mixed with a catalyst that will bring about the chemical reaction that makes the resin cure into a crystalline clear substance. Please follow the instructions of the manufacturer regarding percentage of catalyst to be used etc.

Many people have tried other substances as a matrix to suspend the metal filings in, such as beeswax, sugar, tree resins and epoxy resins.

The epoxy and tree resins (when they are treated in a way that they can be cured to a non sticky crystalline matrix) seem to work ok, but we haven't tried it. We have learned to handle the polyester well and find it works so fantastic that we need not look for alternatives. Even sugar seems to work but is of course not exactly weatherproof for outside application.

Beeswax has been a great disappointment even though it's still periodically brought up by people who don’t like the notion of using a “chemical” material to restore the natural energy balance.

See it in a different way: Isn’t it fun, that we can defeat the plans for global control by the self proclaimed elite, using exactly the addictive substances that are most pivotal in their global enslavement policies: Mineral oil (resin) and sugar.

Next you will need lots of aluminium shavings. (Some call them filings) You can get these at machine shops that work with aluminium a lot such as window manufacturers sign makers etc...

A lot of people who have read Wilhelm Reich’s works are very prejudiced against aluminium because of some observations that Reich made in connection with his orgone accumulators. He spoke out strongly against aluminium use in accumulators, but nobody in our network has so far found any evidence that this would hold true for orgonite.

Don has also built many classical Reich-style accumulators with aluminium and also found no adverse effects. Could it be that another factor, as yet unknown, created the negative results, Reich was reporting?

Last but not least you will need crystals. Clear quartz crystals that is. But they can also be milky or otherwise of “inferior quality”.

We have even used white and completely opaque quartz chips we found on a mountain near the track we were driving on, to add into some improvised TBs that we had to make when we ran out of ammo in Namibia in September 2004.

They worked well.

For TBs, any small crumbs of crystal breakage will suffice. For HHGs use single terminated crystals, that is crystals with one naturally formed tip.

For CBs one should preferably use double terminated crystals, which means they have a natural tip in both directions. Even when multiple small tips point all in one direction we still speak of single terminated.

The easiest to start with: making TBs

TBs or Tower Busters are the easiest to make. That’s why I strongly recommend you start with them. You will get a feeling for working with resin that way which will come in handy when you start making the more sophisticated pieces. We make our TBs in standard muffin forms, but you can use anything that will hold approximately 100-150 ml (3-5 oz.) of resin and shavings. Plastic or paper cups are also popular.

If you want to re-use the mould, we strongly recommend that you wax it before pouring. We use “silicone release wax” which we get from the place that sells the resin and catalyst. Baking fat will work as well.

We use aluminium filings. (or sometimes they’re also called shavings) Other metals work as well.

We get best results when the shavings are not to big or small and sufficiently loose so that the resin can run in between them.

Waxing the form

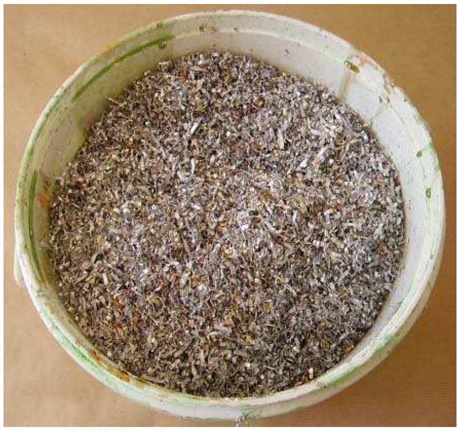

Shavings the way we like them

Apart from the shavings and resin, the mandatory ingredient is quartz.

You don't have to use A-grade tips. Any quartz breakage even of a milky or opaque quality will do. We have used white quartz pebbles collected from the road with great success on our Namibia Exhibition when we were out of TBs at some point.

We like to add some low grade amethyst because we get it cheaply and it makes them nicer. Also we like to add a few crumbs of black tourmaline. If available, powdered pyrite makes any orgonite much stronger in sheer power.

But again: simplicity is the rule. You do not need any crystals other than quartz in order to achieve stunning results.

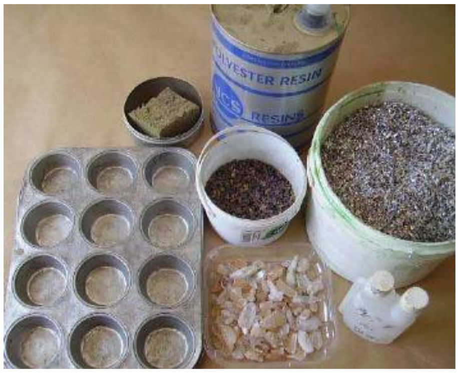

Typical ingredients the way we make our TBs

The best resin for orgonite is polyester resin of the so called “clear casting” quality. It's the same that is used to cast Insects or leaves into decorative key ring pendants and the like. Make sure you get to the industrial suppliers.

Putting the crystals in

Otherwise you pay a fortune for small quantities.

If you don't know where to find those suppliers, talk to the guy who work with fibreglass. (Swimming pool and boat builders for example) The will know.

We start with a thin layer of filings and place the crystals on top of this little nest. That way they will end up being approximately at the centre of the finished TB.

Top layer

We top up to just under the rim. If you want to add some crushed pyrite or tourmaline, you can sprinkle it over the top layer in the end.

Now we can start the pouring procedure. We need the resin, and a suitable quantity of catalyst. The Resin suppliers normally sell special measuring bottles for the catalyst and we recommend that you get one if it's not supplied for free. We also recommend that you use industrial strength rubber gloves for the larger works while surgical latex gloves are good for finer work. If you plan to do this more often you should get yourself a mask with filter against organic solvent vapours. You don't really need that if you can work outside.

POLYESTER RESIN
NCS RESINS

What you need for pouring

Wear a mask if you frequently do this inside. Make sure rooms are well ventilated.

Use bucket with measurements

Pour sufficient quantity of resin into a bucket with measurements.
For one standard muffin tray we use approximately 1.1-1.2 litres.

Measure needed quantity of catalyst

With most makes of resin the standard mix is 2% catalyst at 20°C. But beware: temperature and size of the piece you want to pour influence the need for catalyst. The hotter it gets or the bigger the piece, the least catalyst you need. Refer to resin manufacturers manual for exact quantities.

Pour catalyst into resin

Add pigments if you want colour

Stir thoroughly!

Pour into small jug with snout for easier pouring

And start pouring

Since the resin takes a while to sink in you will have to pour in several passes. It is always good to have a few more objects prepared by the side that can take any surplus resin in case you may have miscalculated the quantity.

You have (depending on the make of the resin and room temperature etc.) 15-25 minutes to do this. That is normally not a problem. If you feel that the resin is becoming like jelly before you are finished, quickly mix the remaining resin with metal filings to make at least some kind of usable orgonite. Keeping your bucket and jugs clean helps to avoid premature gelling.

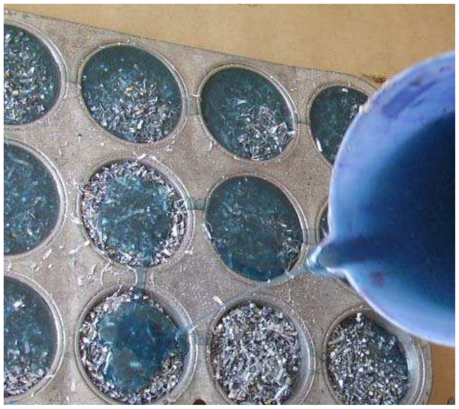

Continue until muffin pan is filled to the rim

When the process is finished, you have to wait a while. The resin will nor start its chemical reaction during which it will get quite hot. If you have used the right amount of catalyst, the TBs should be easily removed after curing by just turning the pan upside down. A few gentle taps with a rubber hammer are sometimes needed.

If the mix has not cured properly (too cold or not enough catalyst) it can be put in direct sunlight to give it some after-curing.

The next step: making HHGs

For making HHGs you want a conical mould. We are using ordinary household funnels because the ones we get have a nice proportion with the base as well as the height being 100mm. But many other forms have been successfully tried: from conical paper party hats to cocktail glasses.

You have to close the nozzle in some elegant way. We used to do that with window putty which could be shaped any way we wanted but are now making a small cone of masking tape that we insert with the sticky side facing outward.

Apart from the filings you need 5 single terminated crystals of 30-40mm length. Additional healing stones are optional. We use crushed black tourmaline and a few of our low grade amethyst pebbles to beautify the energy emanating from the HHG.

Household funnels

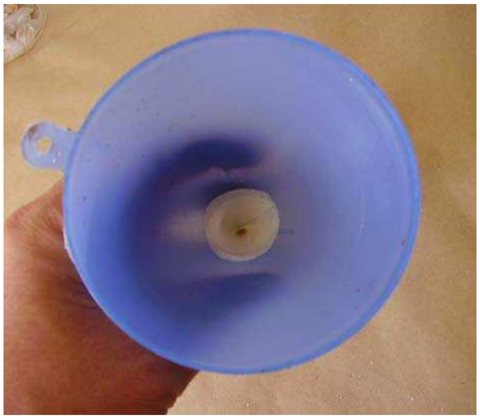

Nozzle closed with masking tape

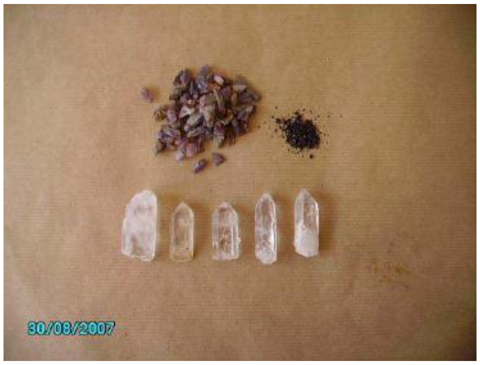

Our ingredients: 5 mandatory crystals, optional amethyst and tourmaline.

We start by putting a few filings into the top in order to make a nest for the top crystal. We want all crystals to be embedded in the orgonite. See the following picture sequence for the further procedure.

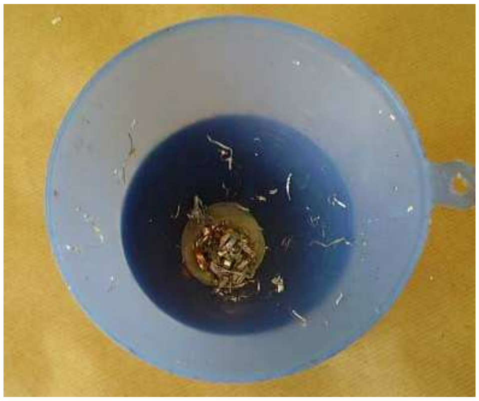

Make a nest for the top crystal

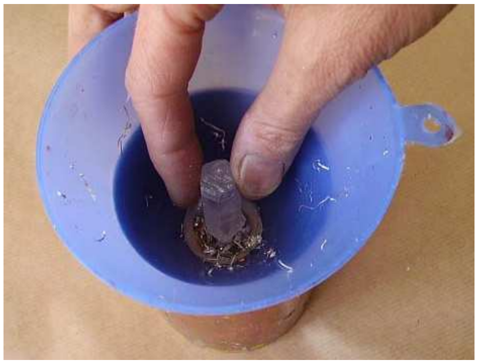

Place top crystal into nest with tip facing downward (up in finished HHG)

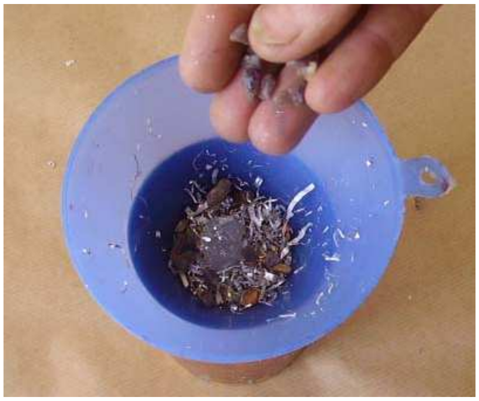

Add more filings and some of the optional additional stones if you wish

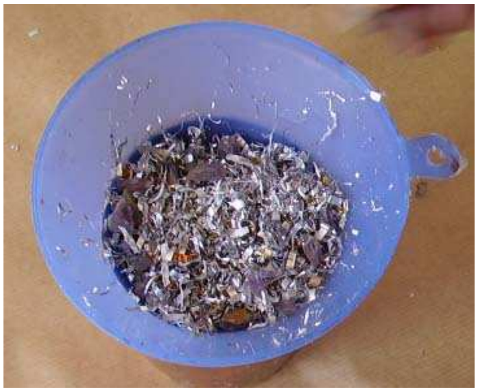

Add more stones as you build up in layers (optional)

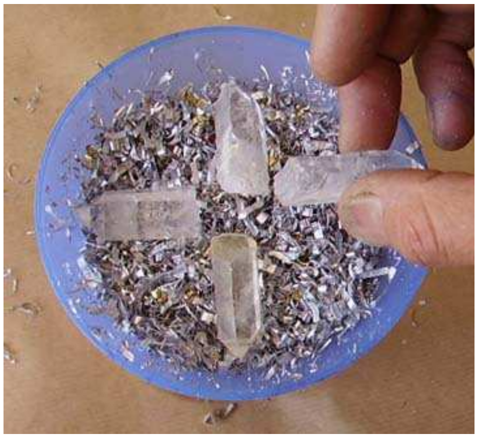

Place the bottom crystals in a cross formation. Leave enough space for the crystals to be covered with a thin layer of filings

Cover crystals

Until they are invisible

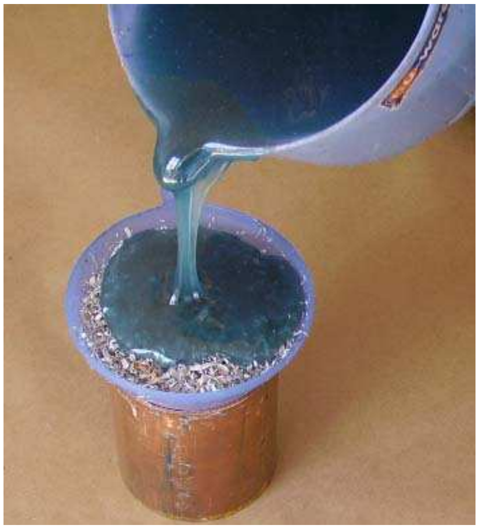

Start pouring

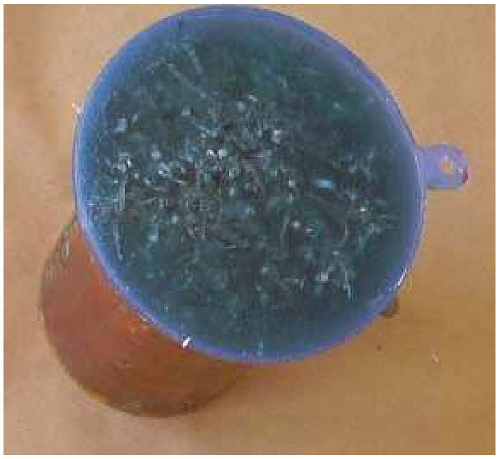

Repeat several times until no more air bubbles come up and funnel is filled to the rim.

This is how the finished product comes out of the mould.

Now you only need to break the rough edge with a file now (or a sanding disk) and sand it until smooth. That's only if you want to make it into a nice piece of art of course. The full power is there without the cosmetics.

Of course any other conical mould will do the job. Cocktail glasses, party hats – anything will do.

There are other variations to the theme such as the ones that Done makes nowadays. They only have a crystal in the tip and a clockwise spiral (from the bottom up) is replacing the 4 bottom crystals.

Finally: Build your own CB

Again, this is only one way to make a functioning orgonite CB. Following these instructions will give you enough of a feel for the matter to deviate from the procedure wherever you feel it would lead to better results.

The standard pipe length according to the inventor of this device, is 6' or approximately 1800mm.

The recommended pipe diametre for a standard CB is 28mm which is almost but not entirely identical to the US 1-1/4"

The crystals may already be fitted inside the short pipe ends. We mostly insert them afterwards and fix them with a bit of resin.

In Don's original instruction they are wrapped into a short piece of garden hose and glued into the bottom end of the pipe.

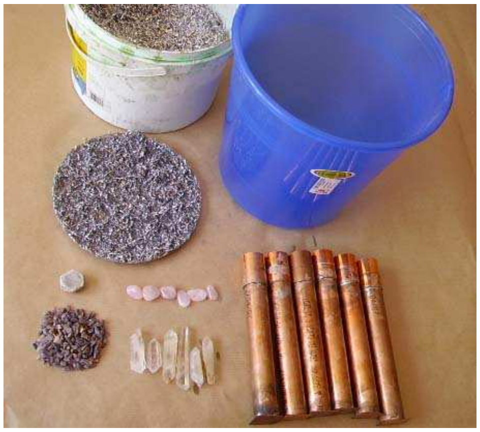

That's what goes into the bucket

The short pipe ends should be closed with a stop fitting (standard from plumbing suppliers) or a copper plate soldered on. (See picture below) It is intended that the pipes form a resonant cavity, so the bottom should not be open. One could use thin foils to close it.

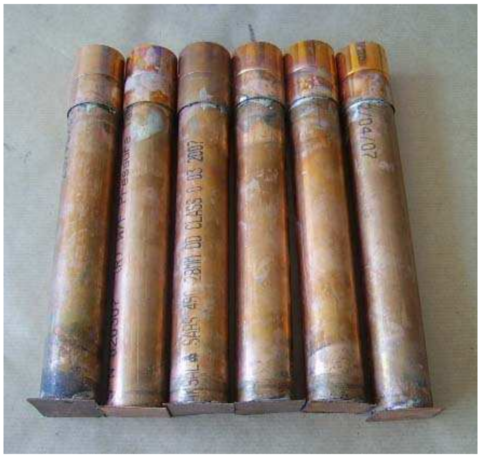

Short pipe ends with bottom plates and connectors soldered on

We use 3 of our standard spacer disks to connect the long pipe ends to a stable rig. We have found that better than inserting special spacers into the bucket as in Don's original instruction.

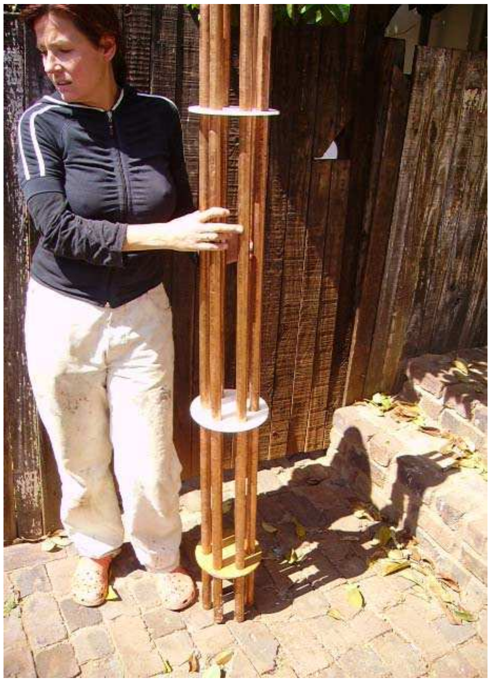

Long pipe ends forming a stable rig

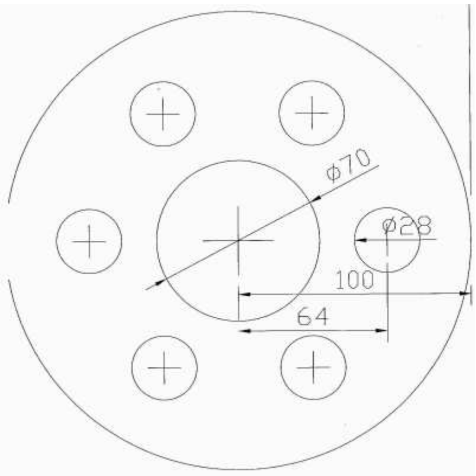

This is how we make our disks. All dimensions are in mm.

Wax bucket with silicone release agent.

Stick short pipe ends onto the rig

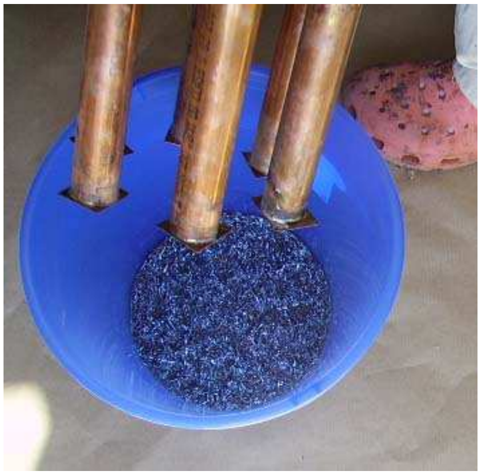

Insert into bucket

You may have noticed that we have already prepared a bottom plate in a previous pouring session. We don't want the pipes to go through to the very bottom but rather want them embedded in orgonite. We make that plate approximately 20mm thick.

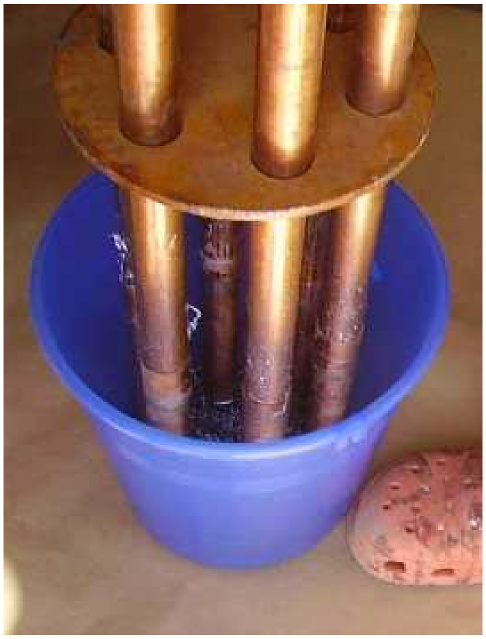

Place the rig upright

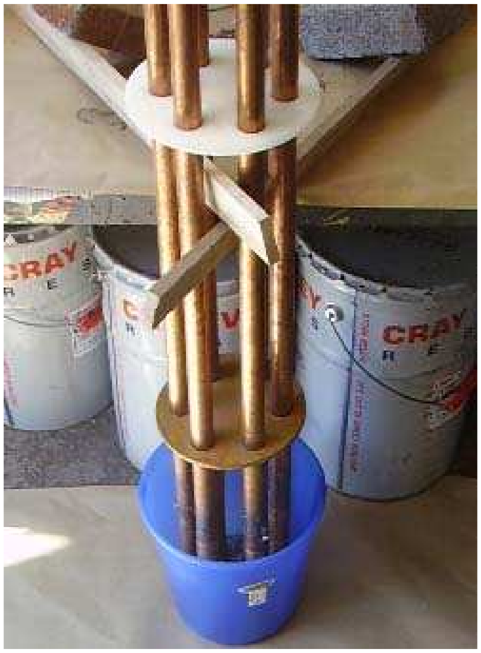

Apply some preliminary bracing

We always put it up next to a table. After applying some bracing, we start to make sure that it's really standing upright and the distance to the border of the bucket is equal on all sides. It helps when the floor is level of course. Otherwise you should use a piece of wood levelled with some wedges as floor. By looking at the pipes from various directions with one squinted eye, you can also ensure that they are parallel and not "drilled".

Laozu's vortex buster is something else. There you want the pipes "drilled" in a spiralling motion. So don't be afraid if you don't get them totally straight. Your CB will work nevertheless.

It is amazing how error tolerant these designs are. As long as you end up with 6 copper tubes with some sort of crystal in the closed bottom, sticking in a bucketful of resin-metal mix in a halfway circular arrangement, you will have a powerful working cloud buster at hand.

Make sure the pipes are level

Now you can start filling up with filings

We fill up to about 2/3 of the intended height. This allows us to pour a substantial amount of resin in one go which will be pushed into the filings by it's own gravity. We put in some "misfired" TBs or other orgonite that hasn't passed the quality test. That way we can prevent cracks even when pouring in one go. Some people still advise to pour in layers. We don't really like it because it takes very long and resin always runs behind the finished parts, resulting in an uneven and messy surface.

Pouring

It is advisable to pour only from one side so that air can still escape on the other side. It helps to have a translucent bucket because then you can see how far the resin has penetrated.

Depending on the texture of your metal filings, you may also premix the orgonite and pour the mix of filings and liquid resin. That way you can prevent air bubbles much easier, but it is also more messy.

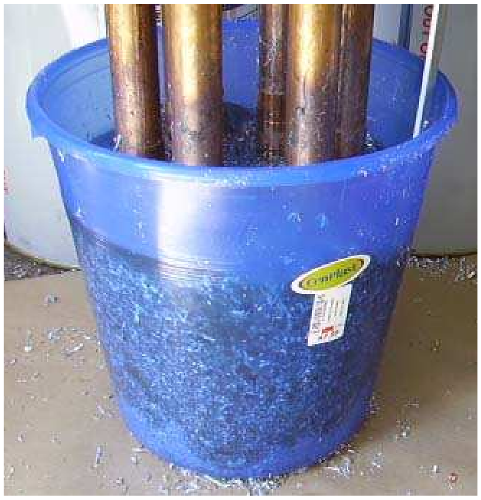

Resin almost down to the bottom

Once the resin has saturated all the filings, we fill up with filings to 1cm under desired height which is in the middle of the connectors. We do that quickly, so we can still use the same resin “wet in wet”.

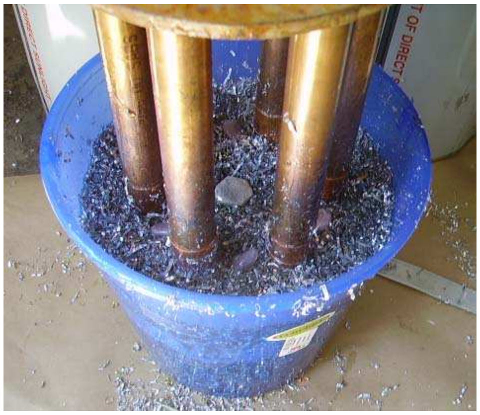

Fill up and pour to centre of connectors

It is important to do that in one session with the same resin. Otherwise you’ll have to wait about 2 hours and get the problems described previously when pouring in layers.

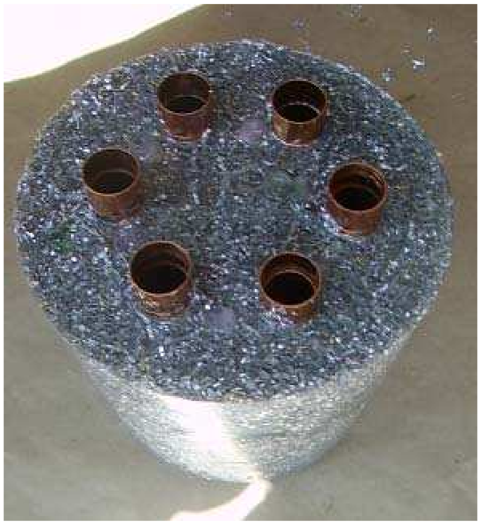

This is how the raw CB base will look like

In most cases you will find that the surface of the raw base might need a final “glazing” coat for aesthetic purposes.

The easiest way to do that is by just taping a rim with masking tape. Small irregularities in the tape will be filed and sanded away afterward when you break the sharp edge.

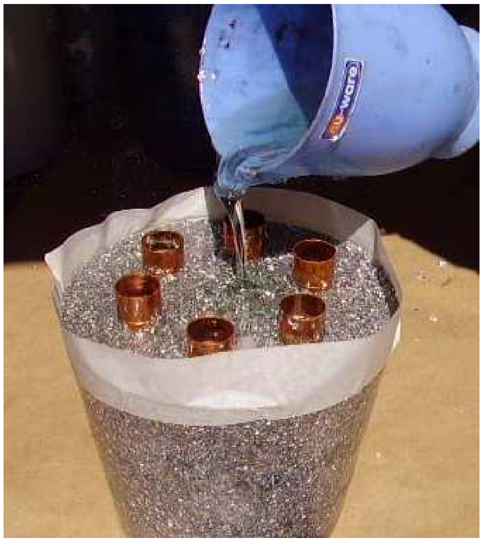

Pouring a smooth top coating with tape

This is how our finished cloudbusters look like.

Yours may look slightly different.
It is very difficult to build a “bad” one. In Uganda we built one with the only copper tubing we could get and that was 22 mm on a roll. We had to straighten it by hand on a concrete floor, so it was never straight and we had no end caps, only chocolate paper to seal the pipes at the bottom..
We also had no pure quartz crystals but only amethyst with some quartz veins. Despite it’s irregular shape it’s influence was perceptible for hundreds of kilometres around it’s location in Kampala.

SP-Crystals and Mobius coils

Text based on Ryan McGinty’s Article on www.ethericwarriors.com with permission of the author.
Photos courtesy of Ryan Mc Ginty

The Items you will need:

1 cordless drill.1 spool of 30 gauge aluminium wire from Radioshack, comes in red, blue and white in 50ft. Do not get the magnet wire that is copper. The copper is very stiff while wrapping also the enamel insulator easily scrapes off causing minute shortages. 1 crystal.

First select a length of wire that will be appropriate for the size of crystal. Very small crystals will need only three feet of wire, larger ones with a diameter of 1 inch or larger will need 20-25 feet or more, the hole spool for giant crystals.

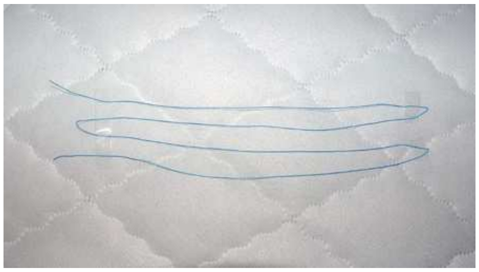

Fold the wire in half, then fold that in half again. Basically folding the wire into quarters.

Now take the end where both ends are bent, place them into the locking chuck of your cordless drill. The easiest way to keep them together evenly in the drill is by starting a small twist to hold them together.

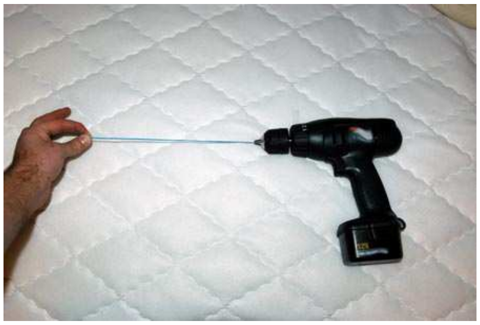

Set the drill direction so that it will remove a screw. (Lefty loosie) With your other hand, hold onto the end with the two leads. Make sure you keep good tension while the drill is rotating. This will minimize irregularities. If the wire is longer than your arm span, have someone help hold it or clamp it to a sturdy stand.

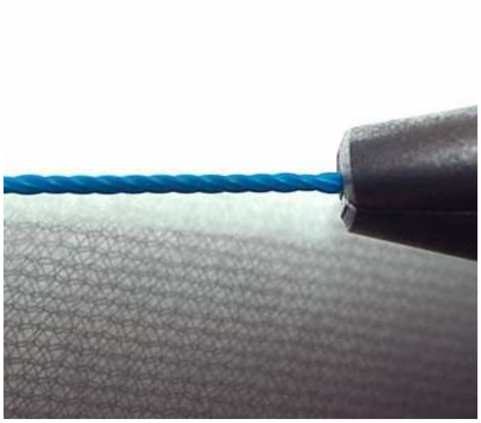

You'll know the wire is correctly wound when the thread angles are at 45 degrees.

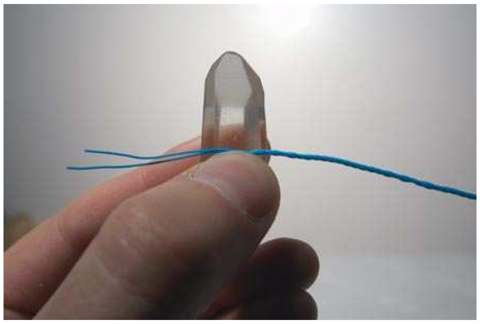

Hold your crystal and wire like so. You will be doing a counter clock wire wrap that works upward.

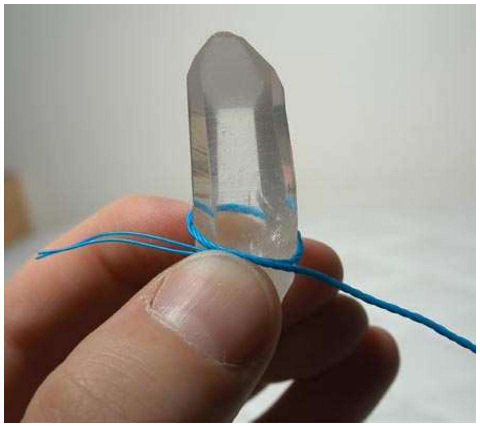

Here is how you start your first knot.

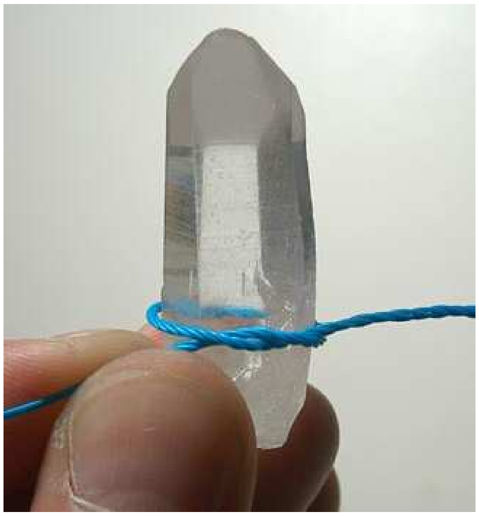

Notice that the knot goes over, down, up-around. It's basically a square knot. Keep this knot a little loose because you'll need the space for the remaining knots.

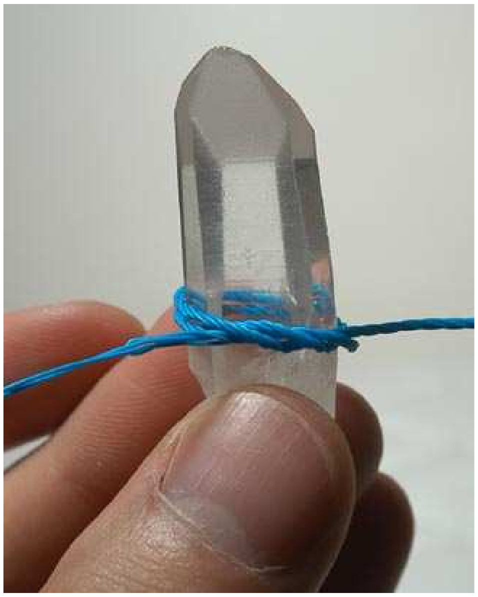

Now repeat the process. You'll notice the wire working upward. While the knot keeps getting larger. Start pulling the wire tight.

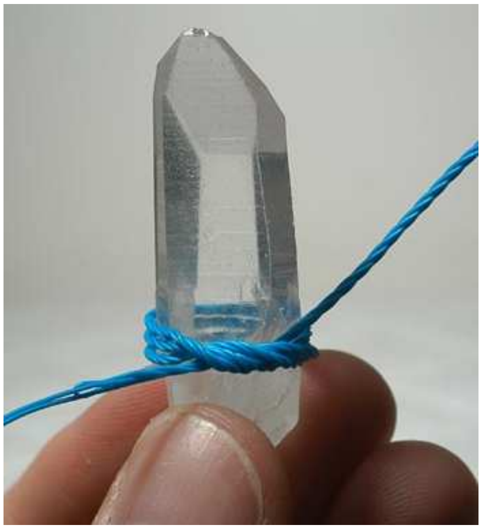

Here is how it looks going around a third time. Each time you start a new knot you start just before the last one. Remember to keep the wraps tight. This will ensure you don't have any wires overlapping on the back.

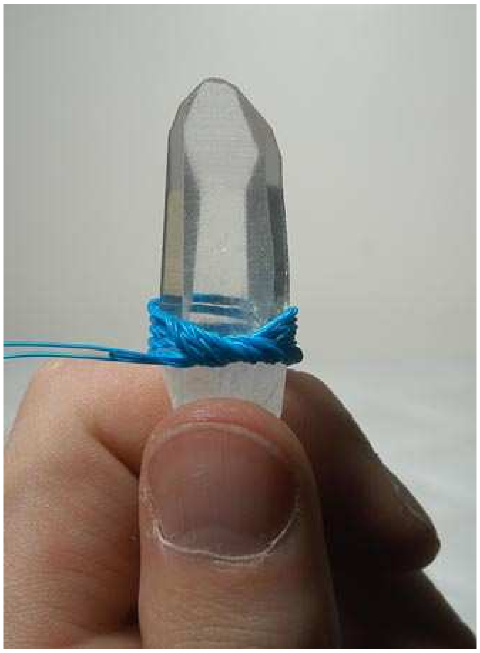

Here is the finished result. You'll notice that the knots start right after the last one begun. Use glue to hold wires in place, a hot glue gun or goop works great. Strip the ends and you got your self an SP! 8-) If you notice the knots getting too difficult to overlap then start another row the same way as the first. Pull the excess just a little above the first Mobius wrap, wrap the wire around CCW then do the knot and keep repeating as many times as needed. It's basically a simple pattern that keeps repeating.

IV. FURTHER READING

The following books and websites can help to understand the perspective from which this book is written and further the depth of your enquiry into some subjects marginally touched here. This does not mean that the author identifies with or endorses all particular views expressed in these sources.

Books

By Dr. Wilhelm Reich:
Character analysis
Mass psychology of fascism
The function of the orgasm
The discovery of the orgone
The cancer biopathy
Ether, god and devil
Cosmic superimposition
Contact with outer space

About Dr. Wilhelm Reich:
Wilhelm Reich – the evolution of his work - David A. Boadella
Der Traumvater - Peter Reich

Orgone and related:
The Life Etheric with Carol Croft – Don Croft
Die Wiederentdeckung des Lebendigen – Prof. Bernd Senf

Science and secret technology
The Tao of physics – Fritjof Capra
PSI – Ostrander, Schroeder
The world weather guide – Pearce, Smith, 1993
Lost science – Gerry Vassilatos
Scalar waves - Prof. Konstantin Meyl
Wholeness and the implicate order – David Bohm
The holographic universe _ Michael Talbot
The self aware Universe – Amit Goswami Ph.D.
The coming Energy revolution, by Jeanne Manning
HAARP the ultimate weapon of the conspiracy – Jerry E. Smith
Angels don’t play this HAARP – Nick Begich
Underwater and underground bases – Richard Sauder Ph. D.
Death in the air – Dr. Leonard G. Horowitz

Reich of the Black Sun (Nazi secret weapons) – Josph P Farell
Fire from Ice (cold fusion) – Eugene F. Mallove

Natural power lines and power spots
Points of cosmic energy – Blanche Merz
The new view over Atlantis – John Michell

Mythology and the origins of mankind
Indaba my children – Credo Mutwa
The 12th Planet – Zecharia Sitchin

Alternative history:
Rule by Secrecy – Jim Marrs
None dare call it conspiracy – Gary Allen
And the truth shall set you free – David Icke
The biggest secret – David Icke
Secrets of the Federal Reserve – Eustace Mullins
Liquid conspiracy – George Piccard
Wall Street and the rise of Hitler – Jeremy F. Sutton
Die CIA und der 11 September – Andreas von Bülow
The conspirators – Secrets of an Iran Contra insider – Al Martin

Medicine
Inventing the AIDS virus – Peter H. Duesberg
Emerging viruses – AIDS and Ebola – Dr. Leonard G Horowitz
The cure for all diseases – Huda R. Clark

Websites (URLs were accurate at time of writing)

On Orgonite and related issues:

www.ethericwarriors.com
www.orgonise-africa.net
www.cb-forum.com
www.worldwithoutparasites.com

Conspiracy research and suppressed news:
(read with discernment)

www.educate-yourself.org
www.whatreallyhappened.com
www.prisonplanet.com
www.chemtrailcentral.com

Text References

---

[1] David Boadella, "Wilhelm Reich. The Evolution of his Work"
[2] Dr. Wilhelm Reich, *Character Analysis*
[3] Dr. Wilhelm Reich, The Function of the Orgasm
[4] See for example Jim Marrs, "Rule by Secrecy" or Gary Allen, "None dare call it Conspiracy" or David Icke, "And the Truth shall set you free"
[5] A good book about these rearguard tactics is "The Liquid Conspiracy" by George Piccard, Adventures Unlimited Press
[6] Zacharia Sitchin, "The 12ths Planet" and other books from his "Earth Chronicles" cycle.
[7] The Bible, authorised King James Version, World Bible Publishers, Iowa, USA
[8] Read some of the popular conspiracy primers like Jim Marrs, "Rule by Secrecy", David Icke, "And the truth shall set you free" and others for the structure and history of the secret society network and its pervasive influence on our society past and present.
[9] See James E. Ewart, "MONEY", Principia Publishing, Seattle1998
[10] G. Edward Griffin, "The Creature from Jekyll Island – A Second Look at the Federal Reserve"
[11] Eustace Mullins, "The Secrets of the Federal Reserve" (Out of print)
[12] Check www.thematrix.com or go to the next video store to get the movie.
The sequels Matrix II and III don't have that quality of being "hidden documentaries" of what's really going on. They rather distract from the message of the original movie.
[13] See "Wall Street and the Rise of Hitler" by Jeremy Sutton, Buccaneer books, New York 1976
[14] See Herman and Chomsky "Manufacturing Consent" for a scholarly analysis of the mechanisms used to manipulate "open societies".
[15] See Cathy O'Brian with Mark Phillips "Tranceformation of America" for a compelling account of a surviving Government Mind Control Victim.
[16] See www.centerpointe.com for more details
[17] See http://www.whale.to/a/electrical_h.html for more links to resources on electronic mind control.
[18] See Dr. Leonard G. Horowitz, "Death in the Air", Tetrahedron Publishing Group 2001
[19] See for example www.carnicom.com for a knowledgeable presentation of the chemtrail programme yet with a fear inducing debilitating twist to it. This is typical for the way in which the Illuminati run their damage control operations once certain aspects of the conspiracy become obvious to a large number of people.
[20] Legend also has it that large parts of the black order that constituted the occult inner core of the 3rd Reich and it's technical elite were evacuated to a large base on Antarctica and several secret bases in South America and high polar regions of Canada, Alaska and Greenland, where they allegedly continue to operate.
[21] Gerry Vassilatos "Lost Science", Adventure Unlimited Press
[22] S.D. Kirlian and V. Kirlian, "Photography and Visual Observation by Means of High-Frequency Currents*, (Russian) Journal of Scientific and Applied Photography,* Vol. 6 No. 6.
[23] Sheila Ostrander and Lynn Schroeder, *Psychic Discoveries Behind the Iron Curtain* (Englewood Cliffs, New Jersey, Prentice-Hall, 1970)
[24] www.ethericwarriors.com is the most authentic website for orgonite application and etheric warfare against the New World Order with many onward links and the more than 90 episodes of "The Adventures of Don and Carol Croft". The site features a forum with invited members who are all active gifters.
[25] See www.donebydooney.com
[26] See www.ethericwarriors.com or www.cb-forum.com (in German)
[27] "Indaba, My Children" by Credo Vuzamazulu Mutwa, Payback Press, Great Britain 1998, first published 1964 in South Africa by Blue Crane Books
[28] "Song of The Stars, the Lore of a Zulu Shaman" by Credo Vuzamazulu Mutwa 1996, Station Hill

---

Openings, Barrytown Ltd.

[29] “Profiles Of Healing – Vuzamazulu Credo Mutwa: Zulu High Zanusi”, Texts by Credo V. Mutwa, edited by Bradford Keeney, Ringing Rocks Press 2001

[30] “The Reptilian Agenda”, Credo Mutwa interviewed by David Icke, published by Bridge if Love Foundation and available through www.davidicke.com

[31] See page 274, “The Social System of the Zulus” by Eileen Jensen Krige, Shuter &Shooter, Pietermaritzburg 1950, first published in 1936

[32] See “Hearing Visions – Seeing Voices” by Mmatshilo Motsei, Cape Town 2004

[33] See http://www.geocities.com/saufor/

[34] Völkermord an den Herero?:Widerlegung einer Lüge by Claus Nordbruch, Grabert Verlag, Tübingen, Germany 2004

[35] See for example: Michael Gienger „Healing Crystals: The A-Z guide to 430 Gemstones” or “Crystal Power, Crystal Healing: The complete Handbook”

The Orgone Revolution

During the last few years, a small but growing group of people worldwide has started a peaceful revolution: planet-wide environmental healing with orgone energy.

Based on easy-to-replicate inventions of US inventor and researcher Don Croft, a network of individuals has started to take active responsibility for their environment.

In *Operation Paradise* Georg Ritschl, a German architect who lives in South Africa, provides an introduction to the subject of orgone energy, including a history of the concept through time and its applications.

Showcasing his own experience based on work in more than 10 African countries, Ritschl demonstrates how orgone energy tools can neutralise the negative influence of electromagnetic pollution, weather manipulation based on HAARP-related technologies and the ubiquitous chemtrails, to create paradise when applied in sufficient quantities at critical places.

The book includes basic DIY instructions for 4 important orgone tools:

- **The Cloudbuster**
- **The Tower Buster**
- **The Holy Hand Grenade**
- **The SP-crystal**

This book is essential reading for anyone who is interested in holistic healing and the convergence of advanced science with the mystical world view.

It is also a book for all those who are worried about the tendencies of a small elite to create a worldwide Orwellian Superstate by means of secret manipulation of world events, centralisation, and secret chemical and electronic warfare against humankind.

It is time to reclaim planet earth for humanity. The compelling message of this book is:

Together we can create Paradise!

OPERATION PARADISE
b/w edition

Copyright: Georg Ritschl 2007

www.orgoniseafrica.com

All rights reserved

ISBN 978-1-84799-369-4
9 781847 993694

An Introduction to Orgonite^{TM}

third edition © 2004 Jon Logan

An overview of Orgone Matrix Material, it's manufacture and application.

Email – wizzer@littlemountainsmudge.com
Phone – 250-446-2032
Mail – PO Box 139, Westbridge BC V0H2B0
Url – www.littlemountainsmudge.com

Introduction to Orgonite^{TM} third edition copyright 2004 Jon Logan

What is orgonite?

The word "Orgonite" refers to a new class of materials which are composed of a mixture of metallic particles suspended in an organic medium. The term "Orgonite" was coined by Karl Welz, the man who is generally credited with inventing it. Though initially intended by Welz to refer only to his commercial product, the name stuck- and has come into usage by most people familiar with the material. So... I suppose in the mind of Karl Welz and few other folks, orgonite means Welz's individual recipe, but to most others it has come to mean any material produced using the same basic conceptual approach. In any case, it refers to an orgone matrix material, as opposed to a layered orgone accumulator. To a degree, a piece of orgonite will act as an orgone accumulator in that it will tend to attract orgone from the area around it, but with an added dimension. The added dimension is that it will process the orgone energy as it interacts with it, whereas an orgone accumulator more simply condenses orgone energy.

There are many variations on the basic concept of suspending the metal within the organic element (as opposed to using thin layers of organic and metallic substances like in an orgone accumulator), and many other substances which can be added to the mixture to make it more potent, or to specialize it for different applications. Most of the people producing orgonite wind up developing their own exact recipes, I use several different recipes for slightly different applications. At the end of the day, the basic, fundamental ingredients of orgonite are metal particles and an organic medium to suspend them in, usually resin.

A wide range of metaphysical disciplines all work with the same energy, know it by various names ... (orgone, odic force, reiki energy, prana, chi, etheric energy, bio energy, bio magnetic energy, huna energy, animal magnetism, soft electrons, mass free electricity, radiesthetic color, astral energy, 4th dimensional energy, emotional body energy, life force energy, and no doubt a bunch of other names I have never even heard of) ... and so far orgone matrix material has proven itself a to be a readily adaptable and highly functional conceptual approach for work in these areas. Orgonite, (by any name) both releases orgone energy when stimulated by other forms of energy, and has the capacity to change orgone from one state to another. It is especially suited to taking the energy from an unhealthy state into a healthy, balanced state, and/or adding a selected property to the orgone energy which it processes. This makes it useful in a wide range of applications, from alternative health through to radionics and other traditional metaphysics.

Organic substances attract ORgone energy and soak it up like a sponge, while metallic substances tend to repel it. Basically, ORgonite pulls in the ORgone energy, and while the energy is inside the device, The metal particles and organic resin both push and pull on it in all directions at the same time. This puts the energy in a scrambled, chaotic state. When the energy comes back out of the device, it collapses back into an organized, defined state .As a side effect of being scrambled and put back together, it comes out clean! DOR is converted back into OR as it passes through the orgonite. (DOR=Deadly ORgone Radiation... Harmful ORgone. OR=ORgone Energy, the Healthy kind)

Orgone matrix material, or orgonite, has not only the property of changing OR (Orgone) from one form to another, but also the property of ‘generating’ OR when stimulated by other kinds of energy... like light, sound, physical motion, or electricity. When using a matrix of randomly arranged particles instead of orderly layers of the two substances (metal and organic) you get a different effect. Some benefits of using a matrix are:

A: Increased efficiency (higher proportionate orgone potential to mass) compared to a traditional ORAC.

B: Changes DOR (harmful ORgone) to OR (Healthy ORgone)... When orgone passes/percolates through the matrix, it is in effect reduced to its component polarities of + (positive/masculine/hot), 0 (zero point/omni potential/infinite), and - (negative/feminine/cool). It is then recombined, and re-emitted. The act of being broken down into its positive and negative charges and recombined basically flashes it across the 0 state and has the effect of restoring the orgone to a ‘base’ state, i.e.. it is no longer ‘healthy’ or ‘unhealthy’ orgone, it is no longer orgone of any specific ‘color’, rather it comes out of the matrix material as more or less undifferentiated orgone. That which makes the orgone ‘colored’ or differentiated is (oversimplified a tad) basically encoded in the proportions of +/-, (+/-)0, and in oscillation rate between + and -. When the orgone exits the matrix material, the conditions which 'force' it to exist in an undefined state no longer exist, and so it collapses back into a defined state, and is 'colored' to varying degree by the exact composition of the matrix material and any minerals/other substances included in the matrix.

C: Orgonite does not actually generate orgone, in the truest sense of the word, but I suppose that is a technicality. Orgonite, when excited with other forms of energy, both converts a portion of the energy used to excite it into orgone energy, and draws additional orgone energy from the Aether (the cosmic omni potential 'pool' of orgone energy that has yet to be 'put to work' in 3D terms), and puts it to work as orgone manifest in 3D terms. Thus for practical intents and purposes, orgonite 'generates' ORgone on demand when excited by scalar waves, magnetic fields, heat, sound, light, electrical energy, kinetic energy, etc.

HHGs and TBs

The type of orgone device which is commonly called a 'Holy hand grenade' is a medium-sized device which attracts harmful ORgone energy, converts it to healthy ORgone energy, and emits it again as a flow of healthy orgone energy. This is different from other orgone devices used up until this point, which had the ability to block or collect harmful energy, but not convert it to healthy energy. The device gets it's name from an old Monty Python movie ;)... When placed in an area, buried, or placed in a body of water, the ORgonite Holy hand grenade will act to continuously clean the ORgone energy within a surprisingly large radius (average 3/4 mile radius from the device). A TB (Towerbuster) does basically the same thing on a little bit smaller scale. The effective range of a towerbuster is generally somewhere around 1/4 mile.

A basic TB is a small (3 to 8 ounce volume), usually cylindrical or conical mass of orgonite with a quartz crystal embedded in it. Generally the xtal is on the verticals axis if the device is tall, and on the horizontal axis if it disk-shaped. A basic HHG is a little larger (usually 8 to 32 ounce volume), usually conical or pyramidal mass of orgonite with 5 double terminate quartz xtals embedded in it, one of them on the vertical axis and four of them on the horizontal axis. The vertical xtal is near the small end of the device, and the horizontal xtals are arranged radially, near the wide end of the device. Don Croft invented these two devices to counteract the adverse physical and mental health effects of HAARP, GWEN, ELF, SCALAR and HPRF transmission towers. In addition to serving well for this purpose, the HHG and TB devices are highly effective multipurpose energy cleaners. Think of them as air freshener for your aura.

To use one of these devices, simply place the device within an area and leave it alone. If possible, it increases the effectiveness to put them into a body of water or bury them in the earth. It is not necessary to bury them or put them in water, and most folks who use them have a few in their homes, to provide a constant cleaning action for the orgone envelope in their living space. With the HHGs, it helps a bit to align the device with the magnetic field of the earth, so that one of the xtals in the base of the unit points north. People who are familiar with natural energy vortices, Curry and Hartman lines, Ley lines, and so forth have been using these devices to restore many of the earth's energy centers to healthy state. This process is generally referred to as ‘gifting’ or ‘towerbusting’. As should be no surprise to anyone familiar with orgone (by any name) , many of us have noticed beneficial changes in the behavior of people within the effective range of such devices, whether they are aware of its presence or not. They are, in effect- getting the same sort of emotional healing and balancing (and in some cases the commensurate physical benefits) that one would expect from more traditional methods of life energy therapy. For several years I used xtals for basic pain relief and first aid in energy work on myself, and now I use something which is basically a fancy TB.

These devices can also be used as meditation aids, used by holding it between the hands while meditating, or by sitting near it while meditating. They can also be carried on the person in order to afford a constant cleaning action for your personal energy fields, and provide shielding against various forms of energy assault, ranging from psychic attack to EM pollution. And, while the devices can be used as intent amplifiers /radionics boosters, IMO there are more specialized orgonite devices which perform that task better. I have noticed that suspending water over the top of an HHG does charge it with orgone energy, and I regularly use this kind of charged water as a general tonic, I find that it has beneficial effects both physically and energetically (vitality, clarity of thought, emotional balance).

One thing which the HHGs and TBs seem to be good at is serving as a basic life energy therapy device, to speed healing and, in some cases, relieve pain. I of course am not a doctor, and am not giving you medical advice, and that applies to everything on this site, bla bla bla. There are several folks who find a TB or a slightly modified TB to be very useful (when held in the practitioner's hand or over the problematic body area on the person being treated) in Reiki, Reflexology, Massage Therapy, healing hands, and so on and so forth (these are all systems which effect healing by working with the human body's orgone energy).

Unlike orgone accumulators, These devices do not have to be carefully protected from EM pollution and DOR. They convert DOR to OR, so I think it is a good idea for most everyone to make at least one or two and put them in their homes, and enjoy the benefits of a healthy life energy environment where they live. As with some of the other similar devices which are already in use, these devices have applications in agriculture and alternative health, in addition to their primary function as an 'energy cleaner'- a device which processes life energy and removes harmful properties in so doing. Every day, we all take a bath in more energy pollution than most folks care to think about, ranging from huge uses of metaphysics with some very nasty intent behind it, to the modern phenomenon of technological assault on natural weather patterns (HAARP), The human mind and body (GWEN), and the harmful effects that various more mundane electrical systems have on life energy. These devices have been shown time and again to have a beneficial effect on the life energy of a given area, being adapted to many uses beyond the original purpose for them, and for which they are still heavily and successfully used, which is to restore natural weather patterns to the environment around us by mitigating or nullifying the effects of weather control technology.

I have seen no evidence that HHGs or TBs will become saturated with DOR from prolonged exposure, but I have seen that such devices have a processing capacity, and that capacity may be exceeded. I have been able to deal with that in most cases by simply using more devices. The capacity of the device is determined by the mass of the device, (larger mass of orgonite = more capacity) the way the internal components are assembled, and the composition of the orgonite itself. I have not found it necessary to move the HHGs and TBs around in order to keep them working, but I have noticed that doing so excites them and 'stirs things up' quite a bit. The exception to this is when you are using them to create a perimeter or clean field in or around your home or another area, in which case you ought to install them by placing them, rotating them clockwise until it 'feels right'... and then try to avoid moving them. Once moved, it takes some time for the fields to re-establish themselves. Where several devices are linked together to create a shield or ley line, moving one causes the whole thing to wobble. Stimulating one of them also stimulates the whole of the field in which the linked devices are included. If you have an SP, or a wand, or even a favorite xtal you use for a beamer you can go around the perimeter clockwise and touch each of the TBs with the tip of your device as you go around. This assists in linking the devices together.

The reason they work better when buried is that the orgonite works a bit more efficiently when in contact with the earth, since the orgone 'resistance' of air is slightly higher than the orgone 'resistance' of earth. Therefore, when the energy flows through the HHG or TB, it can flow a little more quickly if it is buried, because it has a huge potential (the earth or the body of water) to which it is grounded. Even if you do not bury them, just tossing TBs into a problem area works fine, burying them is a boost, and not absolutely necessary. Good thing, because there are many places where burying them is impractical. If possible, toss them where they will land in water (best) or on earth, as opposed to pavement. Drop one in your annoying neighbor's bushes hahaha ...

Orgone Accumulators, Orgone generators and Orgonite

The basic difference between orgone accumulators (ORACs) and orgonite is that orgonite processes orgone energy, whereas ORACs condense orgone energy. Some devices employ orgonites processing ability to 'generate' orgone from other forms of energy, and some devices employ orgonites processing ability to change orgone from one state to another. Examples of this latter use are changing orgone that is in an unhealthy state into a orgone that is in a healthy state, or converting orgone into the state which is most suitable for a given purpose at a given time (say, for therapy or radionics, etc.).

An orgone generator is a device that both collects orgone, and translates other forms of energy into orgone energy. It is possible, using orgonite, to convert heat, light, sound, scalar, kinetic, electrical and magnetic energies into orgone energy. Some people call orgone cleaners 'orgone generators', and in fact the man who invented orgonite has patented the term orgone generator, so indeed there are people who would define these terms differently than I do. However, I am trying to make the point that there are different functions involved, and as long as you are aware of them I suppose it does not matter what you call the, except for the sake of convenience in establishing a cohesive terminology with respect to orgonomy. Orgonomy, by the way, is a name for the study and application of orgone energy.

(Sigh), now... Orgonite does not actually generate orgone, in the truest sense of the word, but I suppose that is a technicality. Orgonite, when excited with other forms of energy, both converts a portion of the energy used to excite it into orgone energy, and draws additional orgone energy from the Aether (the cosmic omni potential 'pool' of orgone energy that has yet to be 'put to work' in 3D terms), and puts it to work as orgone manifest in 3D terms. Thus for practical intents and purposes, orgonite 'generates' ORgone on demand when excited by scalar waves, magnetic fields, heat, sound, light, electrical energy, kinetic energy, etc.

An orgone cleaner is a device that allows orgone to pass through it, and in the passing revitalizes the orgone energy. In addition to revitalizing the orgone, cleaner devices which employ orgonite or other similar conceptual approaches selectively remove and add qualities form the orgone. Again, there will be more articles at some point in the future which go into this matter further, but for now bear in mind that a cleaner is different from both a generator and an accumulator. HHGs, TBs, and Chembusters are examples of orgone cleaners. Also, it bears mentioning that orgonite-based orgone cleaners are not filters, though if overloaded of insufficiently grounded they will still function as filters. A filter picks the bad stuff out and collects it, and you have to periodically clean or replace the filter. Some folks have found that when their HHGs are not grounded well enough, the they have to periodically refresh them by sitting them on the earth or in running water for a few hours. But, if the HHG is in contact with earth or moving water, or connected electrically to the house electrical ground (I'll be producing a kit for this so you can modify your HHG easily at home... still in the testing phase, looks promising so far... and of course I'll post instruction so you can do it even if you do not want to buy it from me hahaha), then the HHG actually converts the bad stuff into good stuff, as opposed to just filtering it out. That is the chief benefit of orgonite-based orgone cleaners, and that is what they are designed to do, and that is what they do when they are constructed and installed properly.

An orgone accumulator is rather like a big capacitor. It is a device, usually in the form of a closed container, which collects the orgone energy from the surrounding area and stores it inside the container. Like a capacitor, an ORAC will collect as much orgone 'pressure' as it can hold, and then it will release some of the orgone so that more can come in. Therefore, it runs cyclically, if often the cycles may be days and months apart, instead of the fractions of a second which we call Hertz when considering the cycle duration of electrical capacitors. Unlike orgonite, an ORAC does not process the orgone, so it can collect a store of orgone on the inside of it which is either good for you or bad for you, depending on the condition of the orgone energy fields in which it is located. It will just collect whatever kind of orgone energy is around it. It is a good idea to have an HHG or TB inside your ORAC if you plan to build one, since most of our electrical grid produces a kind of orgone energy that is unhealthy, and ORACs operated within the EM fields produces by electrical appliances or high tension power lines tend to become saturated with this unhealthy orgone. Thanks to the higher levels of masonry for the decision that 60 Hz is the frequency of alternating current we use in North America. In terms of pulse rate, orgone is not in a healthy state when it pulses at 60 cycles per second (orgone energy will often pulse in sympathy to an electrical current which passes through it. 60 cycle electricity passes through the earth's orgone fields and puts them into a state which is more or less inimical to most living organisms). Of the places in your home to place an HHG, near your electrical mains (don't shock yourself and don't sue me. i mean to put it near the mains, not electrically connected to them) or fuse box is a likely candidate.

An ORAC is able to collect the energy because the sides of the container are composed of alternate layers of metallic and organic material. This arrangement of layers produces something like an orgone 'diode'. It is easier for the orgone to travel one direction through the lamination of metallic and organic layers, and much harder for it to travel in the other direction. So, like a diode, the layers move orgone in one direction. Build a box out of this material, and it will collect and condense orgone on the inside, since it can get into the box much more easily than it can get out. The ORAC was invented by Wilhelm Reich about 70 years ago.

Another thing to bear in mind is that TBs, HHGs, and chembusters will, to some degree, respond to the energy environment around them, becoming more active when there is a lot of DOR or unbalanced, harmful life energy around them, and less active (unless excited) when there is little problematic energy around them. They also respond to large potentials of healthy life energy when they are near them, and will often appear to sit inactive when the energy around them is relatively clean and undisturbed. If you are using an HHG to charge water, for example, it is a good idea to use sound or light to stimulate it, so that you can make sure it is active. Assuming you have the proportions of metal and resin correct, and the particle size of the metal is within working parameters, It is not possible for a TB or HHG to become saturated with DOR the same way as an ORAC because the orgonite has a quality of processing the orgone from one state to another, whereas an ORAC does not. I and several other use TBs to keep our ORACS clean, and (with a TB in it) I have been using a small cylindrical ORAC less than two miles from a large power line without any problems, this for years now (a/o mar 2004).

Chembusters and Cloudbusters

There is a lot of controversy on this subject now, and to be honest I consider it important to distinguish between a Croft-style Chembuster and a Reich-style Cloudbuster. After having had a chance to experiment with both devices, I find that they are indeed different in function. A Cloudbuster is a device invented by Wilhelm Reich about seventy years ago, and it uses orgone energy to influence weather patterns. It is basically an array of parallel pipes which are in turn connected by means of rubber or other organic hoses to either an ORAC or a body of moving water. It relies on the circulation of orgone energy within the water bodies of the earth to clean the orgone by virtue of the action of moving water over stone, this action takes place in riverbeds especially. A Chembuster (or Croft-style Cloudbuster) is a slightly different device, which is based on the Cloudbuster in that it uses a parallel array of pipes to channel orgone energy, in order to affect the weather patterns in the area around it. The main difference between a Cloudbuster and a Chembuster are that a Chembuster has a ballast of orgonite, while a Cloudbuster is grounded directly from the pipes to the earth, or water. And since orgonite processes orgone energy instead of simply collecting it or moving it around a Chembuster cleans the orgone as it passes through. A cloudbuster does not.

Although both Cloudbusters and Chembusters generally pull orgone in the form of DOR down from the atmosphere and channel it into the Earth, it is possible for both devices to work in the opposite direction and send orgone form the Earth into the atmosphere. In both cases, it is the changing orgone fields which affect weather patterns. I do not approve of novices using Reich-style Cloudbusters, but I do approve of just about anyone using Chembusters. This is because a Chembuster is a great deal more user-friendly than a Cloudbuster. A Cloudbuster, used carelessly, can cause tornadoes, imbalances in the earth's orgone fields, and torrential rains or severe droughts. It is nonetheless a very useful device, but one which requires knowledge both of meteorology and orgonomy in order to operate beneficially. The Chembuster, on the other hand, functions more like a 'safety release valve' between the earth and atmosphere, so that it becomes more active when there is a large unbalance in the ambient orgone fields and less active when the energy around it is in a balanced and healthy state.

I have some other articles coming which go into this matter further, but for now suffice it to say that you can cause problems with a Cloudbuster if you do not know what you are doing, but this is not true of a Chembuster, and that is because a Chembuster uses orgonite and quartz crystals to clean the orgone as it passes through, whereas a Cloudbuster does not. It is possible to further enhance the function of a Chembuster by grounding it to the earth. After experimenting with it for some time, I generally ground the orgonite base of my Chembuster instead of grounding the pipes, because if you ground the pipes, then the device will appear to work more rapidly, but what it is actually doing is dumping the excess
DOR into the ground rather than cleaning it. In my opinion, grounding the copper pipes directly is no better than a Reich style cloudbuster, because it is just dumping the DOR into the ground. If you are doing this, then you ought to add some HHGs and TBs or equivalent right near the ground wire where it enters the ground. Also, a wet ground is better than a dry ground. It is better to ground the Chembuster to moving water or wet earth than to dry earth. If you are grounding it to dry earth, you should use a long wire to maximize contact with the earth. You can either attach a length of copper wire to the orgonite, or just cut the bottom out of your bucket and let the orgonite touch the earth.

Orgonite Density

Not all orgonite is the same. The size of the particles used makes a lot of difference to the finished product. Below is a short description of some of the different grades of orgonite with which I have worked. Orgonite can be used for different devices, and the intended function of the device determines what kind of orgonite is best to use. Different materials offer differing amounts of resistance to the orgone as it flows through it, and the different grades of orgonite have sequentially higher 'orgone resistance' values.

LD - Low Density Orgonite

Low Density orgonite offers very little resistance to the orgone as it flows through the device. This means that the device is not very suited to cleaning the orgone, it will not convert DOR to OR with much efficiency. But, since it attracts the orgone and allows it to flow through it very easily, low density orgonite is useful for sending and receiving orgone energy, and for collecting and moving orgone energy from one place to another. One place I often use LD orgonite is to make a channel or layer of LD orgonite within an orgone device, which acts like a wire to conduct the orgone along a chosen path inside the device. LD orgonite also works well as an orgone 'antenna' to send and receive orgone energy within a network of devices. LD orgonite does not work well on its own for making TBs and HHGs, but a thin layer of LD or VLD (Very Low Density) orgonite at the base of an HHG or TB assists the device in grounding to the earth more efficiently. LD orgonite will release some orgone when excited, but not very much.

MD - Medium Density orgonite

MD orgonite offers enough orgone resistance to foster the cleaning action for which orgonite is well known, and works as a material from which to make TBs and HHGs. MD orgonite has a dense enough matrix to convert DOR to OR, but does not work very efficiently in an orgone-on-demand generator. It offers low enough orgone resistance to function well in a passive mode, and high enough orgone resistance to still afford the 'scrubbing' action which is desirable. MD orgonite will release some orgone energy when excited by external energy sources, but like LD orgonite, still not very much. The orgone resistance is low enough in MD orgonite that much of the energy used to excite it either passes through it or is attenuated. It is more suited to use in devices which are meant to operate in a passive mode. TBs, HHGs and Chembusters are generally meant to operate in a passive mode. While MD orgonite may not be the absolute best thing from which to make TBs and HHGs, it will work nonetheless.

MHD - Medium High Density orgonite

MHD orgonite is imo the most efficient type of orgonite from which to make TBs, HHGs, Chembusters and other devices which operate in a passive mode. MHD orgonite offers just low enough orgone resistance to operate in a passive mode, but offers high enough orgone resistance to maximize the 'scrubbing' action while still in a passive mode. MHD orgonite will convert DOR to OR efficiently without being excited by external energy sources (other than a potential of DOR to convert into OR). Because of its higher density, MHD orgonite will also release orgone when excited, and it will do so with more efficiency than MD orgonite. MHD orgonite works well to construct devices which are intended to operate in both passive and active modes. The spaces between the larger particles are filled with smaller particles. This is also the grade of orgonite that I make my HHGs and TBs from now. See the 3 small pics below? MHD orgonite can be gotten by either using smaller particles, as in the two cases on the left and center, or by adding superfine particles to the resin and then using larger metal particles which would ordinarily produce MD orgonite. Either way will produce MHD orgonite. I use about 2 teaspoons of mineral 'flour' to a liter of resin.

HD - High Density orgonite

HD orgonite offers a much higher orgone resistance than MD or MHD orgonite. The scrubbing action provided by HD orgonite is plentiful, provided it is excited in some way. TBs and HHGs made from HD orgonite need to be placed in an area with extreme inimical energy problems if they are to do much. They also benefit from being electrically grounded, though I suppose it is not absolutely necessary. TBs and HHGs made from HD orgonite can be a bit smaller than ones made from MHD orgonite, and still do the same job, provided they are excited in some way, either by external energy sources or by a LARGE potential of DOR to convert into OR. One place in which TBs and HHGs made from HD orgonite work well is on or near household electrical appliances. To make a long story short, HD orgonite processes orgone more efficiently than MHD orgonite, but you need to 'push on it' in order for the orgone to go through it very fast at all. There are few if any large particles, and that there is not very much space between them. HD orgonite does not work very well for devices which are meant to operate in a passive mode, but it does work well for devices which are meant to operate in an active, or 'powered' mode. HD orgonite works well to make orgone-on-demand generators which convert other forms of energy such as heat, light, sound, EM or scalar waves into orgone. HD orgonite works well to make intent amplifiers. HD orgonite can also be used inside an orgone device made from lower density orgonite to control the pattern of energy flow by offering more resistance at select places inside the device, acting like an 'orgone pressure valve' so that orgone will not pass through it until it has reached a sufficient intensity. HD orgonite is the highest density at which orgonite will work both in the passive and active modes, but is more suited to the active or 'powered' mode.

XHD - Extra High Density orgonite

XHD orgonite does not work very well in a passive mode, and does not appear to be doing much of anything until it is excited by some form of external energy. When excited by heat, light, physical motion, sound, EM fields, electrical current, scalar waves, etc., XHD orgonite is VERY efficient at producing orgone energy. XHD orgonite has uses in the internal composition of orgone devices built from lower densities of orgonite, just like HD orgonite, but is especially suited to making orgone-on-demand generators which are designed to operate in an active or 'powered' mode. Note the particle size, all of the have gone through a 1 mm sieve, the largest are about 1.5 x 1 mm, and the majority of the particles are in the form of a fine powder like flour. When making HD or XHD orgonite, you can add the organic part of the orgonite in the form of small particles, and just use enough resin to hold the mass together. While XHD orgonite works very well for making devices which operate in an active mode, it does not work very well at all for devices which operate in a passive mode, except as a small part of the finished device.

Different materials for making orgonite

While many people now use polyester resin and metal particles from a machine shop, there are many different materials from which orgonite can be made. In my experience, not all of these combinations produce a very durable substance, but in some cases produce more orgone than the basic polyester resin and metal filing mixture

Cheap orgonite from waste paint

An extremely cost-effective and colorful variety of orgonite with which I have been experimenting. The orgonite is not especially durable, and needs to be kept indoors or recast in plastic resin to seal it from the elements. It is made from old latex paint (often free at recycling depots) and commercially available waste paint hardener (about $3.00 CDN for enough to do 2/3 of a gallon). The waste paint hardener is in the form of a small packet of crystals which, when mixed into the paint, cause it to solidify. The metal has to be mixed into the paint prior to adding the hardener, since it turns stiff almost immediately. The mixture then takes several weeks to dry fully in a large casting like the 3.5" blue TB (with kyanite from revted ;) shown. This can be quickened by making the mixture and spreading it out into a thin sheet, allowing it to dry, and crumbling it into small pieces like the two shown on the right of the picture. The small pieces can then be combined in different colors and recast with just enough resin to hold them together. Doing it this way, you can reduce the cost of polyester resin by 50 to 70 percent, but it takes a lot longer.

Another orgonite recipe is made from polyester autobody filler, mixed with metal shavings and a little aluminum roof patch paint. You have to work quickly, and mix in the metal and paint before adding the catalyst to the autobody filler. While this kind of orgonite is not ideal for TBs and HHGs, it has the advantage of not leaking out of small holes in the molds, and therefore is useful for some hard-to-cast shapes, as it can be forced into the mold in small amounts with a spoon or spatula.

A few notes about using fine powders in orgonite

You don't need to have NASA grade powdered quartz or ultra fine metal particles. You can use them, but you can also make your own. I make most of my own. One source of copper powder that is easy to get is copper spray paint. Get the cheap kind that rubs off on your fingers after its dry. It is basically metal flakes, solvent, and a little oil. When the solvent evaporates, you are left with fine metal flakes and a little oil. Just spray it into the resin and mix thoroughly. Of course, you have to mix the fine powders into the resin before you pour. Also, if you are using any of the following powders (and probably a few I don't know about)...

copper

paramagnetic earth with high iron content

rust (iron oxide)

white polyester pigment (titanium oxide)

decomposed chalcopyrite (oxidized copper and iron and sulphur)

decomposed iron pyrite (oxidized iron and sulphur)

...then you have to add a bit more catalyst than you normally would, since the copper and paramagnetic earth (some kinds, anyway) slow down the curing (of POLYESTER resin) quite a lot.

There is no answer the question 'which is better, powder or shavings' because the two give different effects. That's kind of like asking whether apples or oranges are better. In short, for an orgone generator (which technically an HHG is NOT) powder is better. For an orgone FILTER/CLEANER (which is what an HHG is) powder can be used as an additive, and you will get a bit more out of it.

If you are trying to make HD orgonite at home, and have problems with the powder settling out, then you can do two things: 1... use less resin! 2... add the organic to the mix in a powdered form and then just use enough resin to stick it all together. To make the higher densities of orgonite, you have to actually mix the ingredients, you cannot just pour the resin over them. Mix all the dry stuff up in a vessel, and then add it to catalyzed resin. Use enough powder that you get a slurry, like oatmeal. Just use a little resin. However, for hhgs and tbs, unless you are gonna put them in a really bad spot or make sure to ground them by burying or tossing in water... you don't want it to be ALL powders. HD max, not XHD. Also, you can make the TBs a bit smaller if they are made from HD than the minimum size they would need to be if they were made from MD. A CB made with HD is a bit different because the pipes create a bit of flow, so there is something to push (or pull) the orgone through the orgonite. Depending on how you have it set up and configured, it may work really well, or not very well at all. Even so, it is possible to get the density too high and 'plug up' the device. I have had that happen before. If you can figure out a way to stimulate it enough, then it will work. One thing you can do is increase the diameter of the pipes a bit to compensate for the density of the ballast. The push or pull provided by the pipes of a CB can act as stimulation for the orgonite if it is of sufficient intensity.

There is one density rating that belongs to the orgonite. Then there is another density rating that belongs to the device as a whole. Ideally, the orgonite density of a TB should be as high as you can get it before it starts to 'plug up'. If you make it too dense, and expect it to work in a passive role, then you will get all discouraged perhaps and say 'it doesn't work'... but that will be because you are basically trying to use a pressure release valve set at 500 psi to relieve 20 psi of peak pressure, understand? Orgonite made with fine particles (high density) works just fine, as long as you use it right. ;)

Anyway, for a TB, the orgonite density should be as high as you can get it and still work, but the overall device density should be a bit lower, so that it will allow energy to flow through it relatively easily. A simple way to say this is the more orgone resistance the device has the better it cleans, but the less orgone resistance the device has, easier energy can flow through it. So you have to find the 'sweet spot' in that range. Imo, for things like TBs and HHGs (as opposed to radionics outputs, pulsers, powerwands, etc) that 'sweet spot' is what I call 'MHD' which stands for 'medium high density'. I also make HD (high density) towerbusters, and they work great... as long as they are deployed CORRECTLY which means being either:

A... right on top (metaphorically speaking) of the tower, or on top of your monitor, or TV, or something else which makes a lot of EM noise...

or B... well buried and spread out a bit in the pattern of gifting.

For HHGs and TBs, what most of us do is to use fine powders as an additive to larger metal particles. Doing this enables us to adjust the density of the orgonite to hit that 'sweet spot' and/or add other energy signatures to the output of the device. So, some people say 'powders are garbage, they don't work' because they try to make TBs with only powders, the density is too high, and (no surprise) it doesn't work. And others say 'well, Welz uses powders so they must be good'. Well... Welz doesn't make devices that have to function in a passive sense, or at least that's not what he's known for. He is not known for making HHGs and TBs. Welz makes devices that use EM (Electro-Magnetic) stimulation to translate EM into orgone, which he then integrates into a radionics machine's output stage. That's why he uses HD/XHD orgonite in his devices. And that's why in devices like my pulsers, ring generators, and so forth, I also use higher densities.

Ingredients for orgonite

Here are some of the ingredients I have worked with in orgonite, and a brief description of their properties according to my opinion and observations. All pronunciations of usefulness for a given purpose are my own views and based on my own perceptions. I am not giving you medical advice. I am not a medical doctor, but there's a good chance you may be ;) Experiment and form your own opinions. I will do my best to update this page from time to time. By no means are all possible ingredients listed here, nor are all the possible effects of the ingredients listed here. This list is my opinion, and is a work in progress. There are many different ingredients being used in orgonite. There are many different recipes of orgonite being used. Some perform better than others all around, and some perform better for a single purpose. In terms of energy work and metaphysics, whatever minerals are included in the orgonite will, to some degree, impart their energy qualities to the orgonite. The same holds true so far for the energy effects of hertzian frequencies, sounds, colors of light and energy impressions via reiki work or other traditional metaphysical discipline in application. The rate of success with this effect, the orgone signature programmed into the orgonite while it cures, seems to vary. In other words, say you have a little piece of turquoise that you like to use for body work. Or perhaps it is rose quartz. I dunno, it's up to you. Say this little piece of turquoise you have is 1 cc. 1 cubic centimeter. Say you cast it inside of 10 cc of orgonite. Well, now it's like you have between 1.5 and 7.5 cc of turquoise to work with. Suppose you are really finicky and you match the recipe of the orgonite to the turquoise, use minerals and metals that the turquoise likes in the orgonite, and maybe charge the orgonite with the energy signature of turquoise while it cures. Now, it's like you have 11 or 12 cc of turquoise to work with, all from that little 1 cc chunk of turquoise cast inside the orgonite. There are people in the world who know a lot more about the specific effects of various frequencies and minerals than I do, and in fact there is really a substantial body of information on the web dealing with the specific energy effects of minerals, colors, and frequencies. Point being, most or all of this knowledge is transferable to working with orgonite, and orgonite is complimentary to these already well established means of working with energy (sounds, colors, frequencies, minerals, etc.).

NOTE - Primary indicates that it can be the main ingredient of it's type, and secondary indicates that it can be used as an addition to the orgonite, but should not imo be used as the main ingredient of it's type.. For example, a TB would have to have at least 1 primary ingredient each (or equivalent not listed in the tables below) from the minerals, metals, and organics tables below. Coloration refers to the 'colors' of orgone that a given substance is harmonic to. I am using the same old 7 spectral colors system with which so many of us are familiar.

Minerals for orgonite:

Mineral	Coloration	Primary	Secondary	Effect
Quartz Crystal	All	✓	✓	Focuses and directs energy within the orgonite. Holds a program. The basic standby in TBs and HHGs, pulsers, succor punches, chembusters, radionics machines and other orgone devices. Is like the master crystal which will respond to some degree to virtually any orgone energy signature, regardless of color or frequency.
Quartz Powder	All		✓	Enhances overall function of the orgonite. Increases the output and responsiveness of the orgonite. Can be used to adjust the density of the orgonite.
Selenite Crystal	White, Blue, Cream		✓	Provides some coloration. Useful for healing. Useful for cleansing and invigorating.. Holds a program. Similar effects to kyanite but a bit more gentle.
Selenite Powder	White, Blue, Cream		✓	Enhances overall function of the orgonite. Increases the output and responsiveness of the orgonite. Provides some coloration and is useful for healing and cleansing. Can be used to adjust the density of the orgonite.
Kyanite Crystal	Violet, Blue, Silver, White	✓	✓	Focuses and directs energy within the orgonite. A workable substitute for quartz in TBs and HHGs, with the condition that it significantly colors the output into the blue and violet range. Produces a surprisingly sharp energy considering it is mostly in the blue range. Useful for cleansing. Similar to selenite bit more sharp and stronger cleansing influence. Holds a program sometimes, but program parameters are limited to those harmonic to the coloration of the mineral.
Kyanite Powder	Violet, Blue, Silver, White		✓	Enhances overall function of the orgonite. Increases the output and responsiveness of the orgonite. Can be used to adjust the density of the orgonite. Retains most of the properties of kyanite xtals.
Citrine Crystal	Gold, White, Yellow, Red	✓	✓	Focuses and directs energy within the orgonite. A workable substitute for quartz in TBs and HHGs, with the condition that it significantly colors the output into golds, yellow and some reds, depending on how darkly colored the specimen is. Holds a program sometimes, but program parameters are limited to those harmonic to the coloration of the mineral. Useful for cleansing and invigorating.

Minerals cont.

Citrine Powder	Gold, Yellow, Red		✔	Enhances overall function of the orgonite. Increases the output and responsiveness of the orgonite. Can be used to adjust the density of the orgonite. Provides some coloration to the orgonite. Gives a nice warm feel to the energy.
Amethyst Crystal	Violet, Blue, White, Gold	✔	✔	Focuses and directs energy within the orgonite. A workable substitute for quartz in TBs and HHGs, with the condition that it significantly colors the output into the violet and gold range. Useful for healing, and for divining, and for cleansing. Holds a program sometimes, but program parameters are limited to those harmonic to the coloration of the mineral. Useful for divining. Useful also for self-searching.
Amethyst Powder	Violet, Blue, Gold		✔	Enhances overall function of the orgonite. Increases the output and responsiveness of the orgonite. Can be used to adjust the density of the orgonite. Imparts some of the effects of amethyst xtals to the orgonite.
Rose Quartz Crystal	Pink, White, Red, Violet	✔	✔	Focuses and directs energy within the orgonite. A workable substitute for quartz in TBs and HHGs, with the condition that it significantly colors the output into the red range, but a light red, more like pink. Soothing and healing energy, surprisingly mellow considering it is mostly in the red range. Holds a program sometimes, but program parameters are limited to those harmonic to the coloration of the mineral. Invigorating.
Rose Quartz Powder	Pink, Red, Light Blue		✔	Enhances overall function of the orgonite. Increases the output and responsiveness of the orgonite. Can be used to adjust the density of the orgonite. Gives a people-friendly, smooth feel to the energy. Invigorating.
Smoky Quartz	Gold, Brown, Green, White	✔	✔	Focuses and directs energy within the orgonite. A workable substitute for quartz in TBs and HHGs, with the condition that it significantly colors the output into the reds and browns and golds. Strongly masculine energy. Useful for diving and protecting.
Turquoise	Blue		✔	Soothing and healing, useful for pain reduction and physical healing. Strong feminine energy, good for healing and regrowth on many levels. Some shielding action but imo more useful for stimulating recovery from injuries physical or energetic. Strong synergy with blood coral and hematite. Increases orgonite's connection to water.
Hematite	Black, White		✔	Strong connection to the earth imparts a 'stability' to the energy from the orgonite. Increases orgonite's capacity to drain away harmful energy by increasing device's mean energy flow capacity. Use as a grounding stone in place of / in addition to electrically grounding the orgonite to earth. Strong synergy with many minerals, especially imo turquoise and blood coral. Somewhat invigorating.
Lapis Lazuli	Gold, Blue, Red, Violet		✔	Very high energy output from this stone for a variety of possible uses. Too much to go into here. Potent and worth reading upon elsewhere. Invigorating. Manifesting.
Garnet	Red		✔	Provides coloration into the red range. Synergy with amethyst.
Fluorite	Blue, Green, Violet		✔	Provides coloration into the green range, and sometimes into the blue or violet range. Rounds the feel of the energy. Assist orgonite with absorbing program while curing. Useful for healing. Does not hold and execute program terribly well, but stores an impression rather well. Useful as a means of transferring energy signatures from one device to another or storing them for future use. Increases orgonite's connection to water.

Minerals cont.

Mica	Green, White		✓	Enhances overall function of the orgonite. Increases the output and responsiveness of the orgonite. Can be used to adjust the density of the orgonite. Sharpens the feel of the energy from the orgonite. Can be used as an organic. Provides coloration into the high green range. An energy "astringent".
Paramagnetic Earth Chunk	Varies, Generally includes Red and up the spectrum to a variable height.		✓	Strongly increases orgonite's mean output. Provides some grounding action similar to hematite, pyrite and chalcopyrite. Increases orgonite's responsiveness to electromagnetic stimulation. Sometimes invigorating, sometimes soothing. Depends on actual earth being used.
Azurite	Blue, White, Violet		✓	Provides strong coloration into the blue and violet ranges. Similar to kyanite in that it is a sharp blue, not a soft blue. Strong synergy with malachite.
Malachite	Green		✓	Provides coloration into the green range. Strong synergy with azurite. A relatively sharp kind of green energy, useful for cleansing and invigorating.
Agate	Yellow, Brown, Gold, White	✓	✓	Focuses and directs energy within the orgonite. substitute for Quartz in a pinch, but tends to flow more slowly than if quartz were used. Strong connection to earth and fire, useful for some protective/shielding influences.
Pearls	Gold, Silver, Cream		✓	Imparts a lovely smooth 'creamy' feel to the energy hahaha. Useful for cleansing and invigorating. Increases orgonite connection to water somewhat. Helps to bridge between fire and water.
Bloodstone	Red, Green		✓	Very masculine energy. Coloration into both green and red ranges. Useful for soul- search and overcoming obstacles. Tends to be invigorating, perhaps too much so for some.
Glass	All	✓	✓	Focuses and directs energy within the orgonite. A poor substitute for quartz xtal, but useful as a means of controlling energy flow within a device, especially if faceted. Tends to give the energy a bit of a rounding/mellowing effect, and gives some coloration of the orgone if it is colored glass.
Quartzite	White, All	✓	✓	Focuses and directs energy within the orgonite. A slightly Better substitute for quartz Than Cut Glass, But still not as good as quartz. Recommend using at least 3X the mass you would if you were using Quartz Xtal.
Quartzite Powder	White, All		✓	Enhances overall function of the orgonite. Increases the output and responsiveness of the orgonite. Can be used to adjust the density of the orgonite.
Blood Coral	Red		✓	Provides coloration into the red range. Strong synergy with turquoise and hematite. A gentle masculine to balance the strong feminine of turquoise. Increases orgonite connection to water somewhat.
Tourmaline Red	Red, Blue		✓	Provides coloration into the red range. Gentler than garnet.
Tourmaline Green	Green, Blue		✓	Provides coloration into the green range. Gentle and soothing energy.
Tiger Eye	Brown, Gold, Yellow, Red		✓	Increases orgonite's capacity to drain away harmful energy by increasing device's mean energy flow capacity. Use as a grounding stone in place of / in addition to electrically grounding the orgonite to earth. Useful for manifestation and repelling unwanted influences. Useful also for self-searching.
Pyrite Chunk	Dark Red, Gold, White, Black		✓	Increases orgonite's capacity to drain away harmful energy by increasing device's mean energy flow capacity. Use as a grounding stone in place of / in addition to electrically grounding the orgonite to earth.

Metals for orgonite:

Metal	Coloration	Primary	Secondary	Effect
Gold	Gold, White, All	✓	✓	High frequencies. Healing, and lifting of mind upwards. Head in the clouds. Useful for divining, soul-searching, overcoming obstacles, and of course purification. Increases orgonite's response to intent stimulation.
Copper	All, Blue, Red	✓	✓	Very wide response range of frequencies. Boosts flow of orgone through device by increasing dynamic action within orgonite matrix. Gives a smooth feel to the energy. Strong synergy with steel or iron.
Brass	Gold, Silver, White, Blue	✓	✓	High and subtle effects. Quite potent, but may be difficult for some to feel. Good for clearing. Not so good for grounding. Increases orgonite's response to intent stimulation. Boosts flow of orgone through device by increasing dynamic action within orgonite matrix.
Cobalt (in blue glass)	Blue, Silver, White		✓	Provides strong coloration into the blue range. Increases orgonite's connection to water. Boosts flow of orgone through device by increasing dynamic action within orgonite matrix.
Bronze	Red, Gold, Blue, Red	✓	✓	Similar to brass but a bit lower on the scale. Increases orgonite's response to intent stimulation. A bit more grounding influence than brass.
Aluminum	All, White, Blue	✓	✓	Similar to copper but a second to copper in terms of the feel of the energy. A wide range of frequency response and coloration. Strong synergy with steel or iron.
Titanium	White, Silver, Gold		✓	Very high frequencies, and relatively sharp energy. Good for dissolving blockages / clearing / energy astringent. Boosts flow of orgone through device by increasing dynamic action within orgonite matrix.
Steel	Blue, Red, Black	✓	✓	Provides stability to the energy signature of the orgonite. A relatively low response range. Provides some grounding influence. Increases orgonite's responsiveness to EM stimulation.
Iron	Red, Blue, Black	✓	✓	Provides stability to the energy signature of the orgonite. A very low response range. Provides more grounding influence than steel. grounding influence. Increases orgonite's responsiveness to EM stimulation.
Bismuth	Blue, Brown		✓	Strongly increases orgonite's response to magnetic stimulation. Some grounding influence. Boosts flow of orgone through device by increasing dynamic action within orgonite matrix. Increases orgonite's responsiveness to intent stimulation.
Lead	Black, Silver		✓	Very low response range. Provides some stability to energy signature of orgonite. Some synergy with copper and gold.
Zinc	Silver, Blue		✓	Increases synergy between magnetic and non magnetic metals. Strong synergy with copper. Strong catalyst in energy terms, good for manifestation. Boosts flow of orgone through device by increasing dynamic action within orgonite matrix.
Iron Pyrite	Dark Red, Gold, White, Black	✓	✓	Provides a grounding, stable note to the energy form the orgonite. Gentler and a bit rounder feel to the energy than steel or iron.
Chalcopyrite	Gold, Red, Blue, White	✓	✓	Provides some grounding and some uplifting action to the energy form the orgonite. Since it is a combination of iron and copper, it possesses the synergy between these two metals in its natural form. Gives a pleasant smooth and potent energy if used alone, mellows the output and increases the response range of the orgonite if used as an additive.

Metals cont.

Paramagnetic Earth Powder	Varies, Generally includes Red and up the spectrum to a variable height.		✓	Strongly increases both the orgonite's connection to the earth, and the orgonite's responsiveness to EM stimulation. Increases the orgonite's responsiveness to intent stimulation in some cases. May or may not carry it's own potent energy signature to impart to the orgonite. Varies greatly from place to place. Boosts flow of orgone through device by increasing dynamic action within orgonite matrix.

Organics for orgonite:

Organic	Primary	Secondary	Comments
Pitch	✓	✓	Wonderful smooth feeling energy. A bit nicer than polyester resin, but less durable. Better imo as a primary than a secondary. I make a fair amount of stuff out of pitch. Will melt in the sun so must be kept indoors or buried. Is ideal for water gifting. Does not dissolve in water. Absorbs a charge well during curing. Strong connection to earth and water.
Beeswax	✓	✓	A workable if not ideal organic primary. Energy output is less than pitch or polyester, and device must be contained somehow. Works better when cold.
Paraffin Wax		✓	Much less functional in orgonite as primary than natural wax. Dunno why.
Carnuba Wax		✓	A little mixed into the pitch gives it a bit 'brighter' feel.
Latex Paint with hardener added		✓	Gives a warm, comfortable, people-friendly feel to the energy when mixed with polyester based orgonite. Often provides coloration according to the color of the paint, but not always. Depends on the chemistry of the paint. With waste paint hardener added, it is a slightly better primary than wax, but not much. Excels as a secondary with polyester., Useful as a means of transferring energy signature form one botch of orgonite to another, for lightening the weight of the orgonite, or for adjusting the density of the orgonite. Absorbs a charge well during curing.
Flour		✓	Gives a warm, comfortable, people-friendly feel to the energy when mixed with polyester based orgonite. Can be used to adjust the density of the orgonite.
Acrylic Plastic	✓		A somewhat smoother feeling yet much more expensive alternative to polyester resin.
Polyester resin	✓		The basic standby against which I measure other organics. Provides some coloration according to the chemistry of the resin. Absorbs a charge well during curing.
Vegetative Matter		✓	Imparts to varying degrees the energy signature of the plants used to the orgonite. Useful for reproducing in orgonite the effects of various herbs. Recommend using well dried, evenly and finely shredded vegetative matter.
Hair		✓	Can be used to permanently link a body of orgonite to an individual. Caution advised. Useful for custom work or tailoring a device to a specific user.
Blood		✓	Can be used to permanently link a body of orgonite to an individual. Caution advised. Useful for custom work or tailoring a device to a specific user.
Dirt		✓	In small amounts can be used to adjust the density. Increases the orgonite's connection to earth.
Incense	✓	✓	Can be used to impart the energy signature of a given incense to the orgonite, if not always the smell.
Sugar	✓		Have not tried this myself yet, but some are reporting success with sugar (hard candy) as an alternative to polyester, similar to pitch.
Honey	✓	✓	Adds a soft, high frequency overtone to the orgonite.
Carbon		✓	Adds a strong grounding influence to the orgonite.

Recipes for orgonite:

Here are some different kinds of orgonite I have been playing with. All pronunciations of usefulness for a given purpose are my own views and based on my own perceptions. I am not giving you medical advice. I am not a medical doctor, but there's a good chance you may be ;) Experiment and form your own opinions. You are responsible for your own safety. *Note - "resin" in recipes refers to catalyzed resin, if applicable. In some recipes, extra catalyst is needed, and given as a percentage value above the regular mix ration. Eg. catalyst +10% means to use the normal amount of catalyst, plus an extra 10% on top of that. The normal resin / catalyst ration for your resin is generally marked on the side of the container you bought your resin in. Particle sizes are given as follows : fine = 0 - 1mm dia., small = 1 - 5 mm dia., large = 5 - 15 mm dia. Proportions of ingredients are approximate.*

Basic LD

Ingredient	Amount in parts	Process	Comments
Copper - Fine	.05 - .1	Mix the catalyst into resin. Add all other ingredients, stirring continually and pour into mold.	Useful for creating a path of lower resistance inside of a device, in order to control the flow of energy. On a more basic note, all HHGs and TBS benefit from a thin layer of something like this at their base. Also works nice as cosmetic second coat for orgonite pieces, and has a bit more output than if just resin were used for the second coat. Also some applications in linking modules for radionics devices.
Quartz Powder - Fine	.05 - .1		
Resin	1		
Catalyst	+50%		

Basic MHD

Ingredient	Amount in parts	Process	Comments
Aluminum - Large	1.5	Mix dry steel and aluminum together. Place mixture into molds. Add xtals / other internal components as applicable. Mix catalyst and resin. Add Quartz Powder to catalyzed resin and stir thoroughly. Pour resin Over metal particles and allow to seep down. Repeat layers as necessary.	A good all-purpose orgonite recipe. Other items can be added as desired.
Steel - Small	1		
Quartz Powder - Fine	.01 - .25		
Resin	To fill molds		

Basic HD

Ingredient	Amount in parts	Process	Comments
Steel - Small	1	Mix all dry ingredients together thoroughly. Add catalyzed resin a bit at a time until there is enough resin to just cover the surface of the particles. The mixture should have the consistency of thick oatmeal, and have just enough resin that all the particles are covered. Allow it to sit for a few minute. If there is more than 1/8" (2mm) of resin covering the surface of the particles, add more steel until the mixture is an even thick slurry like oatmeal. Pour into molds. Stir constantly, including while pouring.	A good basic recipe for use in either heavy-duty TBs and HHGs, or in actively stimulated orgone generators. Different primary and secondary ingredients can be used as desired. .
Secondary metal - Fine (copper or pyrite or bronze suggested)	.1 - .25		
Resin	.75 - 1.15		
Catalyst (if using copper or paramagnetic earth)	+50%		
Secondary Mineral - Fine	.1 - .25		
Secondary Organic - Small	.1 - .25		

Basic XHD

Ingredient	Amount in parts	Process	Comments
Primary Metal - Fine	1	The particles must be fine enough to suspend in the resin. One way to deal with it if this is not the case is to catalyze the resin, wait until it gels, and quickly mix in the other ingredients. Another is to use a thickening agent. In some cases, the thickening agent and the secondary organic can be the same. Premix all dry ingredients thoroughly, and the add the dry ingredients to the catalyzed resin, stirring continually. Add the dry ingredients a bit at a time. You should get a creamy thick liquid that pours slowly into the molds and still is liquid enough to smooth itself off at the surface. If the particles are settling out, there is too much resin. Usually, resin to dry ingredients should be about 1 to 1, sometimes a little more or less depending on the absorbency of the secondary organic.	A useful recipe for making orgone generators designed to be driven by outside power sources like zappers, magnetic pulsers, radionics machines and so forth. Also useful in making small orgonite devices to be carried in a pocket and handled a lot.
Secondary Metal - Fine	1		
Resin	3 - 4		
Secondary Organic - Fine	1.3		
Quartz Powder - Fine	.4		
Secondary Mineral - Fine	.3		
Catalyst if using copper	+ 30% - 50%		

Basic EHD

Ingredient	Amount in parts	Process	Comments
Primary Metal - Small	1	Mix the catalyst into the resin. Add the fine ingredients to the resin. If it is not thick enough at this point to suspend the primary metal particles, add a bit more flour until it is. Mix all the small ingredients into the resin, a bit at a time. It should come out looking like a thick slurry, with bits of metal visible in it. The metal should not all pile up on the bottom. Internal components of devices (like crystals) need to be supported in some way during curing since this orgonite will most likely not hold them in place during curing like MHD will.	It may take a bit of messing around to get this recipe working right, but it has (especially when made with wheat flour) a gentle, people friendly feel to it. This orgonite is good for many things, since it basically behaves like MHD. EHD stands for expanded high density. Good for devices which are meant to be used indoors in living space, and / or handled by people a lot. Useful for both passive and active roles.
Secondary Metal - Fine	.10		
Resin	.75 - 1		
Quartz Powder - Fine	.10		
Secondary Mineral - Small	.25 - .5		
Secondary Organic - Fine (wheat flour suggested)	.75 - 1		

Wood Sprite™

Ingredient	Amount in parts	Process	Comments
Pitch Resin	2.5 - 3	Melt the pitch slowly over a heat source with no open flame. The fumes from melting pitch may be flammable. Don't burn yourself, wear gloves. Once the pitch is melted, Premix all the dry ingredients except for the citrine. Add the dry ingredients to the pitch and stir well. Then add the citrine. Spoon carefully into molds and allow to cool. You generally have a few minutes only before the mixture sets up once removed from the heat. Charge with either sunlight or a charged water bath (or both) while cooling.	A lot of work, but well worth it. Strong connection to the earth, have gotten many obvious reactions from animals when gifting this kind of orgonite to natural energy centers. Also a nice kind to have indoors if protected form hot sun. Mixture will melt if exposed to high heat.
Chalcopyrite or Pyrite - Small	.5		
Aluminum - Large	1.5		
Steel - Small	1		
Paramagnetic Earth - Small or Large	.5		
Dried Crumbled Sage or Chaparral - Small	.3 - .5		
Citrine - Fine	.01		

White Light

Ingredient	Amount in parts	Process	Comments
Aluminum - Large	1	Mix the catalyst Into the resin along with the pigment. The resin should come out bright white and opaque. Place the Large metal in the molds and sprinkle a little of the small metal over it. Place xtals if applicable and pour in enough resin to cover the metal. Charge during curing with 741 Hz.	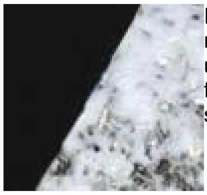 Produces a kind of MHD which is useful for repelling unwanted entities / energies. Also useful for clearing blockages. Energy tends to feel uplifting (as opposed to grounding) and somewhat astringent.
Titanium Oxide White Pigment	Enough to make resin opaque		
Steel or Iron - Small	.25		
Copper or Chalcopyrite - Small	.25		
Resin	To fill molds		
Catalyst	+50%		

Chocolate Shine™

Ingredient	Amount in parts	Process	Comments
Steel - Small or Large	1	Mix all the fine ingredients together. Mix catalyst into resin and add fine ingredients. If the steel particles are large, then place them in the mold and pour the resin mixture over them stirring as you pour. If the steel particles are small, add them to the resin mixture and stir constantly as you pour the mixture into the molds. Charge with 15Hz and / or Schumann frequencies while curing.	Cosmetically appealing and highly EM responsive HD or MHD orgonite. Good for TBS and HHGS as well as generators. Has a warm feel to the energy. Especially suited for work on Geopathic lines, ley lines, Curry / Hartman lines etc. Provided the paramagnetic earth of is sufficient quality, this general recipe has been proven to reduce EMF by conventional testing.
Brass or Copper - Fine	.3		
Paramagnetic earth - Fine	.3		
Citrine Powder - Fine	.05		
Quartz Powder - Fine	.1		
Resin	1.25 - 1.3		
Catalyst	+100%		

Ivory

Ingredient	Amount in parts	Process	Comments
Copper - Fine	.05	Mix all the dry ingredients together. Add the pigment to the resin along with the catalyst. Sift the dry ingredients into the resin, mixing continually, until you have an event consistency. You should have a slurry with the consistency of oatmeal. Pour into molds and allow to set. This recipe will usually support the internal components at least somewhat, but is still thin enough to pour into complex shapes. Charge with both high and low freqs while curing.	Produces HD orgonite that tends to exude a milky white color of orgone. Useful for applications in reiki healing hands radionic therapy body work etc. Useful for removing blockages and stimulating energy circulation. Orgonite tends to have affinity for human body and sunlight. Useful for active roles.
Titanium Oxide Pigment	Enough to make resin opaque		
Paramagnetic Earth - Fine	.05		
Copper or Chalcopyrite - Small	1		
Iron or Steel - Small	1		
Selenite - Small	.3		
Resin	2.5		
Catalyst	+100%		

Water Baby™

Ingredient	Amount in parts	Process	Comments
Blue latex paint with or without charged water added. Paint must be dried out and crumbled.	.3 - .6	The paint is a separate preparation that has to be done ahead of time. Mix 10 drops per liter blue food dye with paint. Then mix the kyanite powder into the paint. Then mix at least .01 part charged water to 1 part paint and dye mixture. Add hardener to paint and / or spread paint out on Wax paper to dry. The paint has to be dry enough you can crumble it into pieces. It does not have to be fully dry as you will be encasing it in resin. Once the paint has dried enough, crumble it into pieces about 5-15mm dia. (1/4" to 1/2"). Mix the blue paint crumbs and the metal (copper) together. Add catalyst to resin and resin. Place metal mixture in molds and pour resin over the metal. Use turquoise or blue glass as the secondary mineral. Charge with the sound of running water while curing, and if possible with moonlight or other silvery blue light. In general, charge in with cool energy as opposed to hot energy.	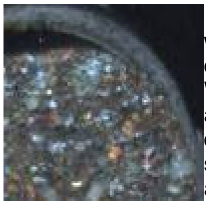 Produces MHD or HD with a strong affinity for water. Also can repeat the energy signature of charged water which was added to the paint. Waste paint hardener or other granular, absorbent material is useful to speed the drying out process. Generally the paint takes several weeks to fully dry, so I set a bunch on at time. This kind of orgonite is pretty to look at if it comes out right, and it is well suited to water gifting. It also has the quality of not containing any ferromagnetic materials, which is useful for some applications. It is a bit less responsive to EM stimulation than other kinds of orgonite, but it works very well for charging water when placed in proximity to a glass vessel of water.
Chalcopyrite or copper - Small or Large	1		
Blue food dye.	10 drops per L of paint.		
Kyanite Powder	.01		
Turquoise - Large	At least 1 chunk per unit.		
Resin	To fill molds		

Orgonite and the curing process

In other articles, i have mentioned a few times that charging the orgonite while the resin is curing has a large effect. Here, i will go into that in a bit more detail. All information given here is based on my own experiments / perceptions, and/or things which other researchers / orgonite users / colleagues have confirmed. I am not a medical doctor, and you are responsible for your own discretion.

To start out with, not all organic components are not equally responsive to charging while the resin cures. Polyester resin, acrylic resin and pitch all demonstrate this property in significant degrees. The waxes seem a lot less responsive in this area, something which i attribute to the formation of crystalline structures within the resin while solidification takes place. Polyester resin, for example, is well know in conventional scientific circles to have significant electromagnetic attenuation properties. That means that it absorbs electromagnetic energy. Since energy is never created or destroyed and only changes form, the energy soaked up by the resin has to go somewhere, doesn't it?

After experimenting for a couple of years, i can comfortably say that the way orgonite is treated during construction creates **at least a 70% difference** in the finished product. In other words, by paying some attention to the thoughts, emotions, energies, ect to which the orgonite is exposed when you make it, you can generally get at least 70% more functionality than you would if you did not pay attention to these factors.

Now, i would like to point out that the beauty of orgonite is that you can get that 30% functionality, regardless of how you make it. It is not NECESSARY to charge the resin while it cures, it is an improvement. And in many cases, 30% functionality is more than enough to make a significant difference. But if the orgonite can be made yet more potent, why not do so? If you can optimize the orgonite to work with a specific ailment, or to do a chosen task, then it only seems prudent to do so.

There are, imo, 2 main ways in which the orgonite absorbs a charge / program while it is curing. One way is through the person making it, and the way that they influence the energy while handling the materials. Another way is to expose the orgonite to energies generated by various metaphysical tools while it cures. Obviously, both of these approaches encompass many different techniques. Both of these approaches can be used either alone, or in combination with each other. Myself, i generally use a combination of the two.

It is also possible for the orgonite to turn out with either a sympathetic or a counteractive effect. The frequencies used in the following description are just examples. So, if you expose orgonite to a hertzian frequency of say 15kHz, generated by a device of some kind, the orgonite will become more responsive to that 15kHz frequency. The fact that the orgonite is now more responsive to 15kHz can be applied in one of two primary ways, depending on what thoughts and intent you expose it to, or what other frequencies are present. If you expose the resin to 15kHz and leave it at that, the finished orgonite will tend to amplify the 15kHz frequency effects when stimulated by it. It will tend to produce that frequency effect as part of it's output, converting other energies it is stimulated with INTO the frequency effect of 15kHz. The same applies if you expose the resin to both a weak 10Hz signal and a strong 15kHz signal. It will tend to 'capture' the act of 10Hz being drowned out by 15kHz, and have that function programmed into it. Doing it that way, you are getting a sympathetic effect towards 15kHz.

But suppose you flip it around the other way. Suppose you expose the resin to a 15kHz signal, and at the same time give it a clear mental impression that you want to dampen out and get rid of 15kHz. If you do it that way, you will get a counteractive effect, in proportion to the clarity of the mental impression you give the resin. It will tend to respond well to 15kHz, but instead of amplifying 15kHz, it will dampen it out, and convert 15kHz into other energy 'colors' or frequencies. The same applies if you were to expose the orgonite to a strong 10Hz signal and a weak 15kHz signal. It would then capture the act of 15kHz being drowned out by 10Hz, and have that function as a 'program'.

In reality, the way that different frequencies interact with each other has arguably a bigger effect that the volume of each signal. There are frequencies that are harmonic to 15kHz, and they tend to boost it. There are frequencies that are discordant to 15kHz, and they tend to dampen it. So, both the relative strength of each signal and the effects of signals on each other produce an effect.

By default, orgonite seems to produce a sympathetic effect to that which excites it during curing. In other words, if you want to get a counteractive effect, you have to 'tell' it you want a counteractive effect. This, doubtless, is why some towerbusters have a greater effect than others. Many people are programming the orgonite with a counteractive effect to the tower's signals, whether or not they are consciously aware of it. Many other people are not, and so the orgonite is running at varying degrees of efficiency. Rather than criticizing, it is my hope that this will help people all around to improve their orgonite manufacture.

People who are already doing reiki, visualization, or other traditional metaphysical disciplines will find that whatever energy they put into the mixture will, to some degree at least, become a permanent part of the finished orgonite's energy signature. This applies, of course, whether or not a person is consciously using metaphysical discipline of some variety. Even if a person is not consciously directing energy into the orgonite while it cures, their thoughts and emotions, the state of their mind, the things they are thinking about while working with the materials... these all have some effect on the finished product. It appears to me (so far) that consciously directing your energy into the orgonite often produces a greater change in the finished product than not. The exception to this would be states of great emotional intensity. If you are in a state of emotional excitement, then a part of that emotional energy will be picked up and integrated into the orgonite.

So, if you are a person who uses prayers, affirmations, declarations, visualization, intent or the like, then you can program your orgonite with these techniques. One thing to bear in mind if using minerals other than quartz crystals is the kinds of jobs that various minerals are suited to doing. For example, if you are making orgonite and programming it to get rid of toxins in the body, then you could make that process either more efficient or less efficient, depending on what minerals you include. If you include minerals that have the effect of stimulating the body to release toxins, then the mineral and the program are harmonic. If you use a mineral that stimulates the body to absorb energy but program the orgonite with a purgative effect, then the result will generally be a bit less effective. In short, the orgone output spectrum of the finished orgonite is defined by a COMBINATION of the ingredients and the programming, so it pays to give a little thought to how you put those 2 factors together.

So what if you are making orgonite and you inadvertently put a bunch of nasty thoughts / emotions into it? Well, they WILL affect it somewhat, but not enough to reduce its effectiveness beyond the 30% line, in my experience. They will make it work less well, but not stop it from working entirely imo. One thing i have discovered is that say, for example, you think 'issue X is horrible'. well, you don't want the orgonite to enhance the horrible nature of issue X, obviously. But once you have thought that thought, it is difficult if not impossible to take it back. So give the orgonite a counteractive program with respect to issue X. Immediately think another thought, namely, that you wish to direct the healing / clearing properties of the orgonite toward issue X, and work on the problem. This produces orgonite that works specifically (counteractively) on issue X. Not a bad recovery from a little slip of the mind. Practitioners of radionics will likely be familiar with this concept, point being the concept is of course transferable to orgonite manufacture.

Another example. Some time ago i made some orgonite TBs which had a few flies fall into the resin while it was curing. I didn't think it that important, though maybe i felt a little sorry for the flies. The orgonite came out with a sympathetic effect towards houseflies. It was amazing to watch how they just couldn't leave it alone. They would be buzzing around the light bulb, and kept stopping and landing on the TBs, which i must admit i found rather annoying. But at the same time, others have reported success using TBs to get rid of problematic insect infestations, by including a sample of the insect in the orgonite. Whether they knew it or not, and i must assume they did, they included the intent that they wanted a counteractive effect. And they saw that the insect infestation was reduced. The difference between my TBs attracting the insects cast inside them, and their TBs repelling the insects cast inside them was simply the difference between a counteractive effect and a sympathetic effect. Again, those familiar with metaphysics will probably find this familiar territory.

Bearing that in mind, i am sure you will understand why programming the resin with the intent that you want to counteract the inimical health effects of the HAARP and GWEN towers is worthwhile. Also perhaps this sheds some light on why there exists such variance in the observable effects of TBs. Without the intent to get a counteractive effect (since if you live in area with towers the resin is exposed to the signal from the towers while it cures) you will get the cleansing effect in the 30% range. But if you add the simple thought that you want the orgonite to have a counteractive effect, you will get much more efficiency, up to several hundred percent improvement.

Rather than going into detail about how to do visualization or focus your intent, i will just say that your emotions are a big part of it. For example, if you are making orgonite to help your sore back, then you want to make sure that the orgonite is programmed by imagining a healthy back, free from pain. Or perhaps imagining a sore back turning into a comfortable, pain-free back. You do not want to program the orgonite with you wondering if it will work. That saps efficiency. So if you have doubts, well that just means you are human. We can not always instantly finish up with our doubts or uncertainty. But we can put them out of our minds for a few minutes. And in those few minutes when we have put our doubts out of our mind and have stopped thinking about them, that is the time to do your intent or programming or visualization. For posterity, the same thing applies with prayers, affirmations, declarations, etc., with or without orgonite.

Ideally, you want to charge the orgonite for the duration of time it takes to cure. This may or may not be possible. I have found so far that with the mental impressions, it is best to do it at least 3 times. Since polyester resin actually continues to cure for a long time after it has solidified, you generally have about a 3 day window of opportunity to charge the resin, though the initial period when it is solidifying appears to be the most critical. Once the plastic has cured, you cannot reprogram the resin, but you can reprogram the crystals embedded within it. You can also direct the energy that the orgonite is producing to a given task with your intent. Directing the output from the orgonite is different than charging the resin. Using your orgonite's energy output is giving the output a job to do. Charging the resin while it cures is defining what the output will be.

When using a machine to program the orgonite, depending on how strong the machine is, a ten to 15 minute period of exposure is often enough to produce an effect, but where possible i often give it up to 72 hours of exposure.

If you have a radionics machine, succor punch, pulser, wand, or some other kind of intent amplifier / orgone generator, you can use that to either charge the orgonite or enhance the mental impression (intent) that you give the orgonite. You can also set your favorite xtal (if you work with xtals) near it while it cures and use that to charge the resin. Varying methods of charging give varying degrees of success. Any kind of attention to the charging process will produce a more effect than none, i find.

One way you can charge the resin is to devise some means of connecting a quartz crystal wrapped with a mobius coil to the output of your pc sound card. I have a bit of info about one such device on the ring generator page (http://www.littlemountainsmudge.com/ringgenerator.htm). Even a smaller device such as a succor punch, connected to the output of a pc sound card, provides a very versatile apparatus for charging orgonite. If you are unable to set up something like that, then even using a tone generator and exposing the orgonite to the sound produced will charge the resin. I am amazed at orgonite's ability to accept a program and repeat it almost indefinitely.

Roberto Forcen, at his site Voicesync (http://www.voicesync.org/), has some very impressive sound analysis and sound generation software. One of his programs that i have worked with a fair amount is called mineral sounds. In his own words, it 'Generates resonant three octave compound tones from a list of more than 3000 minerals. Tones are generated using powder X-ray Diffraction parameters.' What i can say is that when i fist started playing with it, i was very impressed. When i run the sounds from this program through my big mobius-driven orgone generators, the energy produced is very similar in 'feel' and in 'color' to the energy from the minerals themselves. Aside from using it to program orgonite with the effects of a mineral i would like to have, but don't have on hand (not quite the same as having the actual mineral in the orgonite but close if you let it run several hours), i use it to compliment the ingredients of the orgonite. For example, if the orgonite contains steel, copper and quartz, i will open multiple copies of this program and charge the orgonite with the sounds of iron, copper, quartz, and gold. This gives the orgonite a lot more power when it is finished.

Sunlight is another wonderful thing to expose the resin to while it cures, and so is moonlight. These two methods of charging the orgonite are already in widespread use. Especially with exposing the resin to sunlight, you will often notice a difference in the visible color of the resin when it has cured, indication that you are in fact doing something and this is not all hocus pocus... ;)

Astrology also has a certain amount of influence on the orgonite, though not one with which i often work. Using a large orgone generator to charge it can drown out some of the astrological effects, it would appear. One batch i did during the harmonic concordance came out very nice, and it appears to have captured some of the intense catalytic energy of the harmonic concordance.

Another way you can do it is to pour the orgonite outdoors, preferably in contact with the earth, in a clean vortex or ley line. This way tends to both stimulate the vortex or ley line and program the orgonite with a permanent link to the area in which it was poured. Alternately, you can add some dirt or wood chips from the vortex to the orgonite and pour it elsewhere. I have used this technique to link together a few different medicine wheels in the area. Possibly, it could be used to create and/or divert energy lines in the earth's body, leading to or from the vortex.

While i was proofreading this article, i phoned my friend of several years Chris Gozdzik from diviningmind.com (http://www.diviningmind.com/), and asked him if he thought i had left out anything glaringly obvious. He reminded me about linking, which is another way that you can do several things with orgonite, including charging it while it cures. Basically, if you link the orgonite to the energy from an existing large source of energy, either by a sample of the source or clear visualization, then the orgonite will not only pick up the energy from the existing source as a program, but will also have a permanent link to the energy source. Thanks, Chris :)

And 2 final notes- If you are building a ring generator, power radionics machine, or other large power-driven orgone generator, first decide what frequencies you plan to use to drive it with. At least pick out a few of the ones you plan to use a lot. Then, charge the orgonite for the generator with the frequencies it will be running at. This increases power in the finished generator. Example, the pulsers run at 5075Hz and 32kHz, so whatever other programming i give them during curing, i give them a good zap of these two frequencies also.

Many of us have noted that orgonite, once cast, sometimes takes a certain amount of time 'start working'. Charging during curing dramatically reduces the amount of time that it takes a device to 'start working'.

Well, that about takes up the time i have today, but here are a few of my favorite hertzian frequencies for charging orgonite, and some the effects i have observed with them.

741Hz (sine wave) - This is one of the solfeggio tones, and is quite potent imo. When i first started using it i immediately felt a buildup of energy and subsequent release in the solar plexus and heart chakras. It may be a bit intense for some people at first, but they should rapidly acclimatize. It appears to have the effect it is reported to by others, that of assisting with finding the solution to a problem. I use it a fair amount these days, and thanks to whomever it was that posted about it on the yahoo cloudbusters (http://groups.yahoo.com/group/cloudbusters/) forum some time ago.

528Hz (sine wave) - This is another one of the solfeggio tones, and it has proven useful for me in manifestation.

5075Hz (square wave) - A frequency i chose myself for use in the driver circuits of the pulsers. It tends to have a grounding, stabilizing effect. I use it as a multipurpose frequency to drive generators, and to help me calm down when i get cranky ;)

32kHz (square wave) - Another frequency i chose for use in the pulsers, tends to have the opposite effect than 5075Hz, but meshes well with 5075Hz. I use it for removing blockages in body work, for shielding from energy attacks, and for an aid in the mental 'reach upwards' involved in meditation and remote viewing. When set to a frequency higher than 20 kHz, NCH toner produces a lower harmonic of target frequency. Since square waves especially produce many harmonics ranging up and down the scale, imo you are still getting this frequency (or one close to it) as a harmonic when you use a pc sound card connected to a coil and xtal.

15Hz (square wave) - The frequency many people use to drive their succor punches (http://www.educate-yourself.org/). It tends to have a solid, relatively stable effect. Of the low frequencies, it is one commonly used today. It is potent, though imo benefits from the company of a higher frequency. Useful for shielding from energy attacks. Also, if you have a 15Hz driven SP, then charging your orgonite with 15Hz will increase the responsiveness of your orgonite to your SP.

32Hz/31.7Hz (square or sine wave) - A frequency i first heard of from (i believe) Bruce Stenulson (http://www.stenulson.net/energy/doras.htm). It has proven very useful as an aid in programming quartz crystals. Because of it's tendency to carry programming information / intent from the user to a quartz crystal, it is useful imo for intent amplification, and as a general purpose frequency for driving mobius-powered orgone generators.

68.635kHz (square or sine wave) - This one of the Hodowanec frequencies. Now, since the upper end of the sound card is well below this frequency, i am not actually getting this frequency when i put it into the NCH toner (http://www.nch.com.au/tonegen/) program that i use as a signal generator. When set to a frequency higher than 20 kHz, NCH toner produces a lower harmonic of target frequency. Since square waves especially produce many harmonics ranging up and down the scale, imo you are still getting this frequency (or one close to it) as a harmonic. When run through the nch toner, and a mobius/quartz driven orgone generator, it produces an emission of orgone that is almost all in the violet range. Violet orgone is both soothing and cleansing, good for soul-searching as well as healing.

EHD Orgonite – Another step

As mentioned earlier, one can produce MHD or MD orgonite either by allowing the particle size to determine density (as in using larger, curly metal shavings, such as one would find in the average TB) or by adding the organic to the mixture in a particulate form. In fact, when making HD or higher densities, you basically have to add the organic in a particle form, along with the metal, since there is not enough space between the metal particles (when they are really small) to produce a ratio close to 50/50 between the metal and organic ingredients. Some time ago I began going a step further and using metal (fine and superfine particles) that would ordinarily produce HD and adding enough organic to bring the density of the orgone matrix material down to MHD or MD. The material produced this way has some very interesting differences from the standard TB / HHG mixture of curly shavings and polyester resin.

It appears that the matrix thus derived has a lot more capacity to store orgone than regular MHD. A given mass of EHD (Expanded High Density) orgone matrix material will have a slower rate at which the orgone flows through it than the same mass of regular MHD, and it will have a higher overall orgone field strength than the same mass of regular MHD, even though the density and the resistance afforded to the orgone as it flows through is about the same. I attribute the slower flow rate of EHD not to a difference in resistance from MHD, but to the fact that adding the organic element in both liquid and particle form increases the overall energy storage capacity of the matrix somehow. This effect is observed much more when superfine metal particles are used along with fine metal and organic particles, as opposed to using fine metal and organic particles alone.

After experimenting with several different organic particles, I have settled on flour as preferred material for the time being. One could use wheat flour, wood flour or other finely powdered absorbent organic substance. The particles should be absorbent enough that they combine with the polyester resin to produce a thick liquid the consistence of stiff pancake batter. The organic powder is mixed into the resin along with catalyst initially. The metal is added last, beginning with the superfine particles and finishing with the fine particles. For a readily available source of superfine metal particles, I suggest cheap metal spray paint. Just add extra catalyst and spray the paint directly on the surface of your resin in the mixing container. Another source of superfine metal particles are the bronzing powders and other metal powders sold by crafts stores for plating.

I have noticed that especially when the organic material is a fibrous material, the orgone matrix material both takes much longer to absorb a charge during curing, and absorbs much more of a charge than regular MHD. It absorbs a program, intent or energy signature with much higher fidelity than regular MHD. It also takes longer to accumulate enough energy to begin working. Regular MHD that I make is usually finished the curing process and has ‘come online’ by about 3 days after casting. EHD seems to take until about 5 to 7 days after casting to be finished the curing process and ‘powered up’.

EHD is also efficient for use in power driven orgone generating devices such as powerwands and ring generators and pulsers. It also affords the cleaning action of converting unhealthy orgone to healthy orgone, but the charging and curing process is much more determinative of the end result, in terms of cleaning capacity, than with MHD. In General, since it takes so long to store a charge and stores such a high potential of orgone, the energy it is subjected to while curing are much more an influence on the finished product than with regular MHD. I would not recommend making EHD orgonite without working in a clean energy environment. Provided you can supply a clean environment or preferably a clean and charged environment in bioenergy terms, EHD will produce TBS that have a higher output that regular TBs and are more ‘ergonomic’, they feel a bit more alive in the hand. If you make TBs from EHD but do not have a clean energy environment to cast them in, then you will get TBs that are not very efficient at doing anything until you pick them up and play with them, at which point you will still likely be able to feel a strong surge of energy when you hold them in your hand.

For personal orgone devices meant to be held in the hands or carried in a pocket, I now prefer EHD to the other recipes I formerly used.

EHD lends itself not so well to making conventional towerbusters which are designed to be high flow filter cleaners. It lends itself very well to making power driven generators or to making TBs and HHGs for indoor use and human contact. It is like taking a snapshot of an energy field (the field in which it cures), and then it will sit there and broadcast an image of that energy field, continuously and with respectable intensity. It will covert other forms of energy into the energy signature (that snapshot) that it has as it’s output. So, if you can handle the programming and charging end of it, it will turn out much more potent per cubic inch than regular MHD. Instead of producing devices which are simple filter/cleaner devices, making TBs with EHD produce SELECTIVE filter devices, in that they selectively convert orgone into the orgone signature they have as their output. That’s why the charging / curing process is so much more critical with EHD.

Though it produces an orgone energy field of higher intensity that MHD, EHD has an output that is much less chaotic than MHD, and the chaotic energy fields are much more contained within the physical matrix than with regular MHD. With regular MHD, the wave fronts of orgone energy collide within the device as they are reflected by curved surfaces which are relatively close together. The waves of orgone energy inside EHD orgone matrix material react less directly with each other, especially the waves reflected by the more widely spaced fine particles. It’s like instead of the waves smacking into each other up close, the matrix has to build enough pressure until the energy ARCS across between the more widely spaced fine particles. All the while, the superfine particles afford the waves the opportunity for up-close collisions but to a lesser degree, providing a stable basis for the matrix, if that makes any sense to you. For this reason, EHD made without superfine and fine particles will not work very well at all.

The mixture should be thick enough that the fine metal particles (say between .5 and 2 mm dia.) are fully suspended in the resin and do not settle out at all. The mixture should be just thin enough that it smoothes itself off at the surface when vibrated, but it should be thick enough that you have to vibrate it to get it to smooth over on the surface. You should be able to count 20 to 50 of the suspended fine metal particles in a square inch of surface area. There should be enough of the superfine metal particles that when wet, the mixture appears to be slight shiny and metallic. When the mixture cures, the surface will become dusty and opaque but a bit of sanding and a clear coat takes care of that. The mixture also takes longer to chemically cure than regular MHD, especially if copper or iron is the superfine metal. Extra time should be allowed for curing. EHD to be exposed to the elements should be protected with a clear coat of spray lacquer or poly resin.

So, in short, as long as it is programmed by a competent energy worker, it works really good for TBs and HHGs designed for human contact as opposed to agricultural and other nature gifting or tower busting. It also works pretty well for generators which will be excited in some active way while in use. And, it absorbs a charge of higher potential and with higher fidelity during the curing process.

Simple HHG instructions Item:	Quantity per HHG:
- Double terminate Quartz Crystal	- 5 (or 1 DT and 4 STs)
- Polyester Resin & enough Catalyst to cure it	- About 6 Ounces
- Metal particles	- About 5 Ounces of metal particles
- Suitable container to use for a mold	- 1 (min. 4" dia. plastic funnel)
- Small Jar or Can to sit the funnel in (hold it upright)	- 1

Note: While there are many other suitable vessels which could serve as mold for this device, I have used a plastic funnel for this example as it is readily available in most places and relatively durable. In any case, the same technique of construction applies regardless of the container used for a mold, which can be in the shape of a pyramid, cone, cylinder or hemisphere, so long as it has a volume at least great that of a as a cone 4" wide x 4" tall. Presently, pyramids and cones are the shape this device is most often built in.

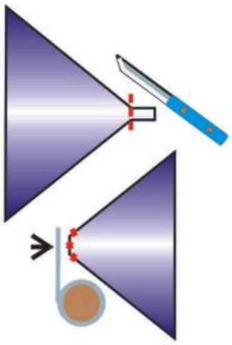

1. - Cut the tip off of the funnel and tape over the hole at the small end of the funnel with aluminum tape (preferable) or parcel tape. Electrical tape, duct tape & masking tape do not work as well for this as aluminum tape. In some way, plug the hole so that the resin does not leak out.

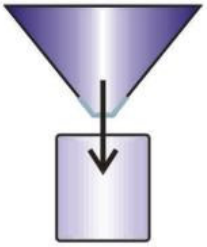

2. - Sit the funnel in the jar so that it is upright and more or less level.

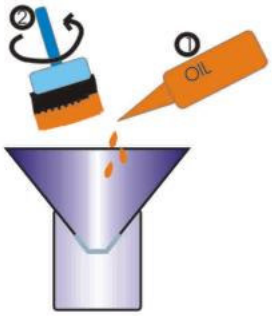

3. - Coat the inside of the funnel with a thin layer of vegetable oil, to keep the resin from sticking to the mold (funnel). Wipe away the excess oil with a brush or clean cloth.

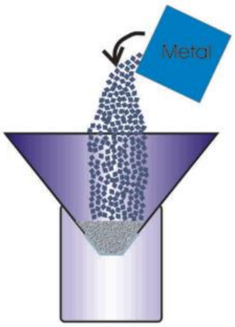

4. - Put about 1" of metal particles in the bottom of the mold.

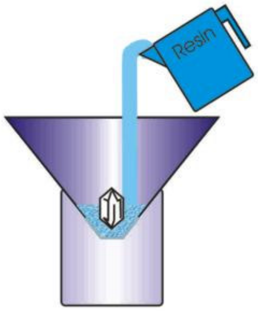

5. - Place one of the Quartz crystals in the tip of the mold, so that it is upright as shown. This one really should be a double terminate, but if you only have a single terminate, put it with the point towards the tip of the funnel. If it is a double terminate, the best tip of the crystal should be pointing towards the tip of the funnel, where you covered the hole with tape.

Mix enough resin and catalyst (be sure to mix well, min. 1 minute of mixing) to come up level with the surface of the metal filings. Pour the resin in and let it run down into the metal filings. Gently shake or tap the mold to release bubbles.

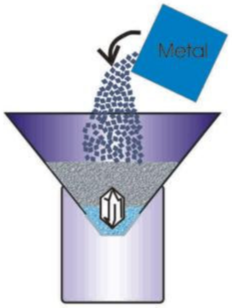

6. - Put in some more metal particles, enough to fill the mold about 1/2 to 2/3 full.

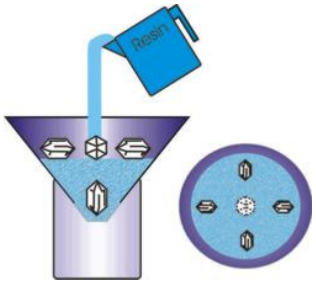

7. - Place the other 4 Quartz crystals on top of the metal filings as shown. They should be arranged like the spokes of a wheel. The better points of the crystals should be pointing towards the outside of the mold. If you have to, you can use single terminate crystals here, and put them with the points either facing in or out (facing out is preferable). Mix some more resin and catalyst, and pour in enough to come up level with the surface of the metal filings again. Gently tap or shake the mold to release bubbles, and let the resin run down into the metal filings.

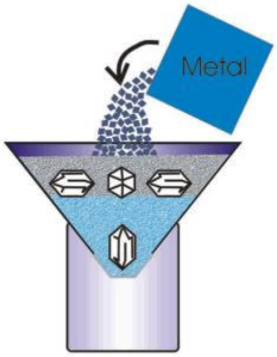

8. - Add some more metal filings, enough to almost (but not quite) fill the mold. Leave about 1/4 inch of empty space at the top of the mold as shown.

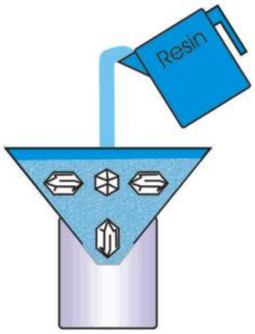

9. - Now, pour in enough resin to finish filling the mold. Let the resin run down into the metal filings, adding a little at a time. Use enough resin to completely cover the metal filings, so that there is a thin layer at the top of the mold which is only resin, no metal. The layer of resin only should be about 1/4" thick. Take a moment to adjust the molds so that they are more or less level at this point.

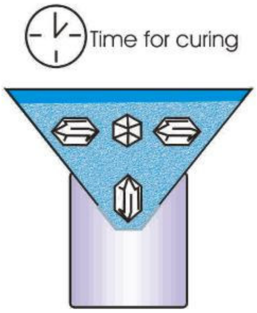

10. - Allow the resin to cure. Keep it warm if possible, cure in a well-ventilated area, and if possible, expose the resin to sunlight while it is curing. If you have a tone generator or other sonotherapy (healing using sound) device, then you can expose the resin to that as well while it cures.

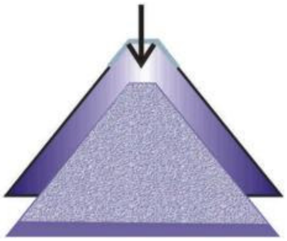

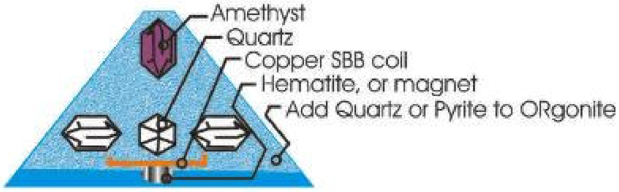

11. - Remove the molds from the jars, turn them upside down, and give them a smack to release the brand new HHG you just made. If the mold sticks, you can either use a thin bladed knife to push in between the mold and the resin, or you can try spraying penetrating oil around the edges where the mold meets the resin. Dropping it on the floor sometimes works, but sometimes it breaks the resin. There are many ways to modify an HHG for enhanced effects, but the basic model shown above will work just fine the way it is. This is one modification which is easy to perform when you are casting the Hhgs. Use an Amethyst point instead of a DT Quartz crystal for the vertical crystal... add a copper SBB coil to the base of the unit, just at the edge of the

metal filings as shown... add a small piece of hematite, pyrite, or a small permanent magnet to the SBB coil as shown (generally I put the N pole of the magnet facing toward the tip of the HHG)... and mix some crushed quartz, crushed Amethyst, or crushed Pyrite in with the metal particles before you cast it.

Simple TB Instructions Item:	Quantity per Tower Buster:
- Quartz Crystal	- 1
- Polyester Resin & enough Catalyst to cure it	- About 3 Ounces
- Metal particles	- About 3 Ounces of metal particles
- Suitable container to use for a mold	- 1 (min. volume about 4 ounces)

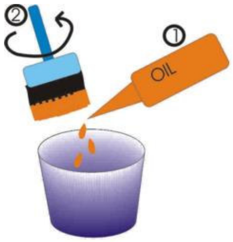

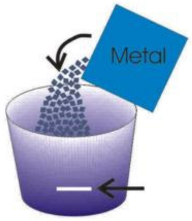

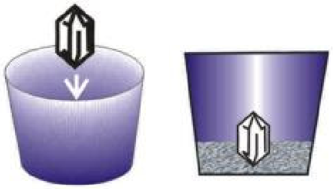

1 - Line the containers you will use for molds with a light coat of vegetable oil. This prevents the resin from sticking. Wipe the excess oil off with a brush or clean cloth. cover the surface of the mold evenly.

2 - Put about 1" of metal particles in the bottom of each mold.

3. - Place the quartz crystal in each mold, using the metal filings to hold it in position. put the crystal in the center of the mold, and if the crystal has a good point on it, most folks put the point towards what will be the top of the unit.

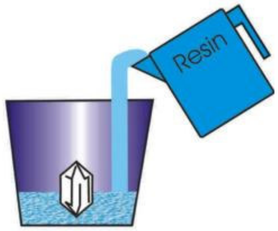

4. - Mix enough resin and catalyst to fill each mold level with the metal filings you have placed in the bottom. Pour enough resin into each mold to come up level with the top of the metal filings. Make sure to mix the resin and catalyst well (mix for about 1 full minute) before you pour it into the molds. Allow the resin to run down into the metal filings. Tap the mold lightly or stir carefully to remove bubbles.

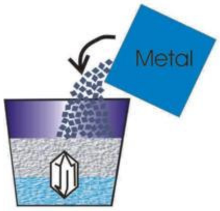

5. - Add another inch or so of metal filings to each container, enough to cover the crystal.

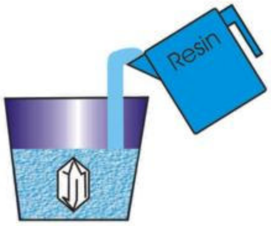

6. - Mix and add enough resin to cover the metal filings you have placed in the molds. It will not harm the function of the device if some metal filings stick out the bottom, but it is a good idea to use enough resin to fully encase the metal filings. Allow the resin to run down into the metal filings. Tap the mold lightly or stir carefully to remove bubbles.

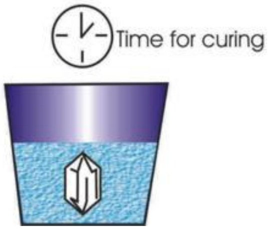

7. - Allow the resin to cure. Keep it warm if possible, cure in a well-ventilated area, and if possible, expose the resin to sunlight while it is curing. If you have a tone generator or other sonotherapy (healing using sound) device, then you can expose the resin to that as well while it cures.

8. - Turn the molds upside down, and give them a smack to release the brand new TB you just made. If the mold sticks, you can either use a thin bladed knife to push in between the mold and the resin, or you can just tear the paper away. Dropping it on the floor sometimes works, but sometimes it breaks the resin. It will not harm the function of the device to leave the paper mold on, but it may be more aesthetically appealing to remove it.

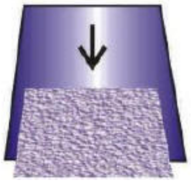

Mobius coils and quartz crystals

Many people have also lately begun using devices in conjunction with orgonite or not, which employ a mobius coil and quartz crystal. The coil can either have it’s leads simply shorted together, or, more efficiently, be connected to a signal generator circuit or radionics machine. Doing this is like turbo charging your crystal.

One name this kind of device has become commonly known by as a succor punch, this is the name don Croft gave to it when he made it public domain. Since then, many variants of this device have been adapted for use in general metaphysics as an intent amplifier or powered wand. There are several different ways to make a mobius coil, the basic concept is that you want to have magnetic fields canceling each other out by intersecting at 90 degrees. This is accomplished by winding the coil in such a way that the wires cross each other at or close to 90 degrees. Then, when a current is put through the coil, either by induction of by direct stimulation with a signal generator (a common circuit used for these coils is the Beck or Clark or Croft style ‘Zapper’ devices invented for bioelectrification) the magnetic fields created by the wires cross at 90 degrees and cancel each other out. When the magnetic fields cancel each other out, they are converted into scalar energy. High voltages are not required in order to produce orgone emissions form the crystal, and could be dangerous. Low voltages are commonly used for this type of device.

The scalar energy then excites the crystal and the crystal produces much more orgone energy output that it would without the coil. The frequency used to stimulate such a device is often determinative of the exact effects, but any signal put through it will produce orgone energy output form the crystal as well as RF emissions. This basic device is at the core of a new generation of rife-style devices, since a pulsed orgone/scalar wave is more efficient at stimulating the human body that either radio waves or ultrasonics. Audio signal will also be translated into orgone/scalar waves when put through this type of coil, and people who use sound therapy will find that inserting this device into their apparatus markedly increases the effects. On the following pages, I have included the instructions for making the style of mobius coil that I use most.

With regard to using an audio signal instead of a simple hertzian frequency to drive this kind of coil, it appears that the underlying emotional current of the audio signal is translated into a pulser orgone / scalar emission. This is quite useful for things like sound therapy and manifestation. However, one should exercise care in the choice of audio signal used, I would stay with the classical music for example instead of heavy metal hahaha. In contrast to a TB or HHG, devices like this can do great good or great harm, depending especially on what kind of intent one is projecting into the device (which intent will be picked up and amplified by the crystal) and what frequencies are used to drive the device. Some frequencies that I know are safe to use are (in Hz) 15, 32, 741, 5075, 5100, 15000, 32000, and 68635. This device is at the core of the large orgone generators I use to charge and program orgonite while the resin is curing. There is a lot of information available on the web about the biological effects of different hertzian frequencies, especially with the Royal Raymond Rife crowd. A device like this will afford a low cost, low – tech, but highly functional output for a Rife style treatment system, without requiring the use of high voltage, plasma tubes, or expensive function generators, since it can be driven by the audio output of a home pc, substituting software for the function generator generally used to drive Rife machines.

As for the waveforms being used to drive the coils, bear in mind that using a frequency to drive a mobius coil has slightly different requirements than just using the properties of a sound. There is a reason that most Zapper circuits used to drive these coil employ a square wave.

When using a frequency for driving scalar producing coils, most of us have generally gotten our minds around the concept that we want to have magnetic fields canceling each other out, so they can be converted into scalar potentials and scalar waves. Well, another concept which applies is magnetic flux. The more rapid the magnetic flux, the more scalar energy you are going to coax out of your mobius coil. Magnetic flux is what you call it when you have a magnetic field that is changing.

Now in a waveform, when you look at that little wavy line, think of it as a line traced by something bobbing up and down, perhaps a pencil if you prefer. And the magnetic field is changing as the pencil goes up and down, understand? As the wave passes by you, the pencil traces it's path and bobs up and down in time with the highs and lows of the waves. How high the pencil goes up and how low the pencil goes down is called amplitude. How long the waves are is called frequency, as in 'how frequently does a wave pass by you' in a given time. The base unit of time used to measure this is the second, and a Hertz (Hz) means "waves per second".

so... a sine wave, which has a smooth and equal curve... work great for some applications where you are just using the sound. It also works to generate scalar waves in a mobius coil, but not as efficiently as a square or sawtooth wave. One difference is that a square wave produces more harmonics in the circuit than a sine wave does. In some cases a sine wave is therefore less efficient in terms of input power to output power, but it is more pure and produces a more ‘pure tone’ which is advantageous in some cases. The magnetic flux is not very abrupt with a sine wave.

For driving mobius coils, a square wave works best of the ones I have experimented with. See, in a square wave, the pencil does not go smoothly up and down (like it does in the sine wave)... it sits still for a little while, and then it jerks up, and then it sits still again for a while and then it jerks down.

And therefore the magnetic field changes very suddenly, that is to say the magnetic flux is much more abrupt in a square wave than in a sine wave. A sawtooth wave, or a triangle wave... is kind of in the middle. (between square waves and sine waves). A sawtooth wave is far better to use in a mobius coil than a sine wave, from my testing so far. The sine wave produces energy that is much more mellow. You have to crank the power up quite a bit to get similar effects to a square wave.

Triangle waves also will work, but a sawtooth wave seems to work a bit better than a triangle wave. Also, if possible, the leading edge of the sawtooth should not be too spiked. I suppose that phi in relation to the amplitude of the wave would be a good height for the spike at the leading edge of the sawtooth wave, but I cannot say for sure at this time. I do know that imo judging from the amount of energy they produce when put through the same device, square waves are best, followed by sawtooth, followed by triangles, followed by sine waves

Tachyon braking, as it applies to scalar wave generation, is basically a fancy way to say that when you are generating scalar waves in a coil, there is also something else happening. See, there is in the aether an infinite potential of scalar energy which already exists (as opposed to the scalar energy you are 'generating' by converting em fields into scalar energy). When you take magnetic energy and convert it to scalar energy, you also tap into that aether potential and draw on it a little, adding the scalar energy from the aether to the scalar energy you have just produced with your coil.

This in fact is the basic working principle of the majority of the 'free energy' devices out there. That is how you appear to, and in fact do... get more energy out that you put in, because you are tapping on the aether and drawing on it. The 'tachyons' (little bits of energy, or particles in a high energy state, depending on the way you want to look at them) are subject to a process called tachyon braking. They 'brake' (slow down) and change state into something else. From a metaphysics perspective, the energy reduces it's vibratory rate and therefore manifests as another, lower form of energy, which appears to pop out of nowhere. Rampa, Cayce, and a jillion others refer to this in metaphysics teaching, the slowing down of energy to produce effects..

So then if you are trying to get electricity out of it, you want the tachyons to pop out of nowhere as electrical potential. But, if you are only taking the process as far as generating scalar waves to excite quartz or orgonite, then the tachyon breaking just makes your mobius coil appear to be more powerful than it actually is hahaha because the scalar energy produced by tachyon breaking adds to the scalar energy produced by em fields canceling each other out in the coil.

The more abrupt the magnetic flux is, the more scalar energy you draw from the aether and add to the scalar energy you are generating with your coil.

First - Make yourself a "mobius Cable" to wind the coil from. While you can wind a mobius coil from single strands of wire, it certainly seems to be a lot more potent when you use a cable made in the manner described here to wind the coil from. Take a length of wire, and double it back on itself twice as shown to the right. Pull a little slack out at the ends of the wire, this will be the leads of the coil when it is finished. You should leave yourself at least 2" for leads, and it is a good idea to give yourself 6" or so, you can always trim the leads to the required length when the coil is finished. It is much easier to use a drill to twist the wires than doing it by hand. Run the drill in reverse and you will get a clockwise twist to the cable. I recommend that you wind the coil itself clockwise also.

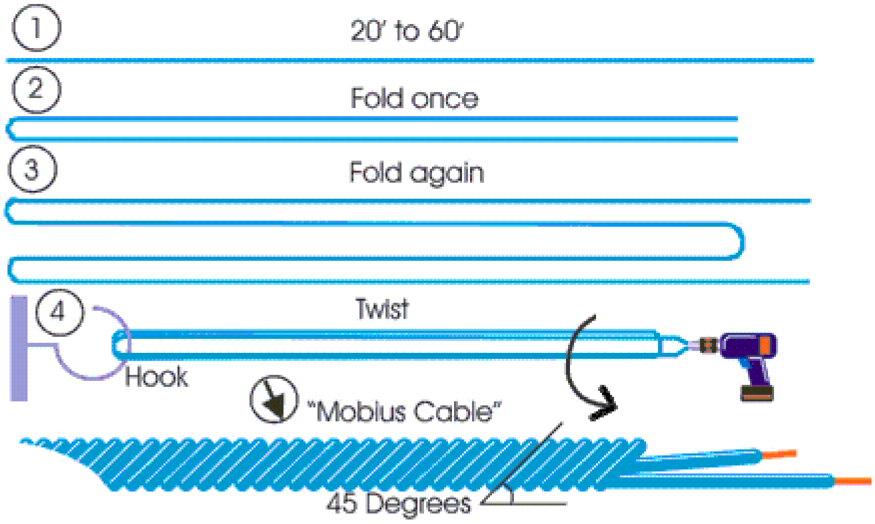

1. - Starting with the end of the cable which does not have the leads, make a circle in the clockwise direction about the size you want your finished coil to be. You can wind the coil around a core (xtal for your SP) or you can wind the coil by itself if you are using stiffer wire.

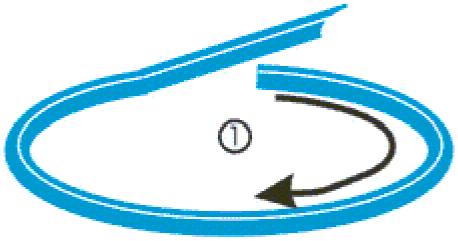

2. - When you complete the first wrap, feed the wire through the center of the circle so it wraps around itself in the clockwise direction as shown. Use a little glue (hot melt or silicone preferable) to hold the wire in place where it crosses over itself.

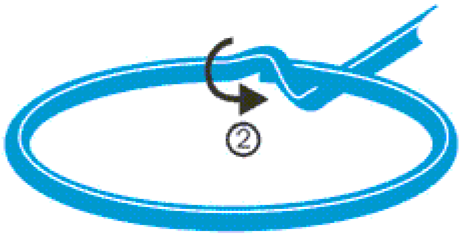

3. - Start winding the wire around the circle in the clockwise direction again.

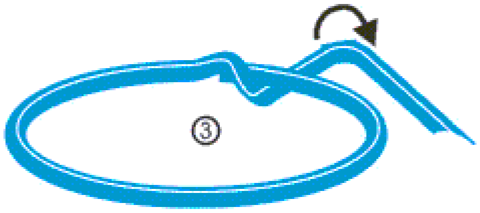

4. - Go around the circle about one third of the way, and wrap the cable around itself again, just like in step 3.

Go around the circle another third of the way, and do the same thing again. You should have 3 wraps around the cable for each time you go around the circle.

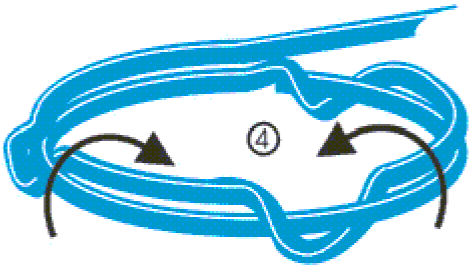

5. - Continue this way, repeating step 4, until you have used up all of the cable. As shown to the right, stay on the same side of the previous wrap with each new revolution.

The 'knots' will run together.

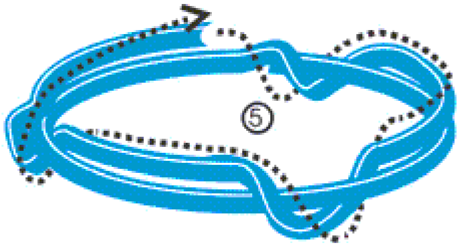

When you are finished, use a little glue to hold the end of the cable in place.

The coil should look like this when you are finished.

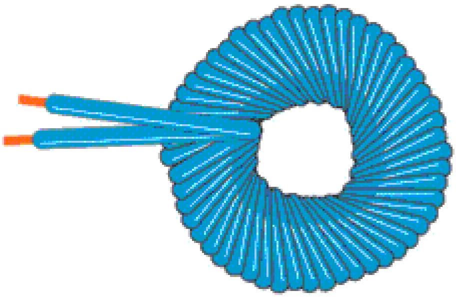

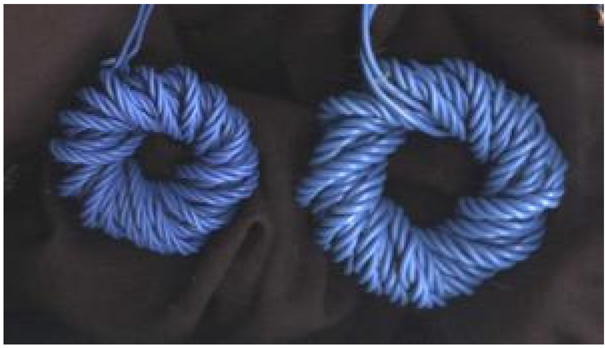

Hopefully, the pictures will be of more use to you than the words, as it is really quite simple to do, just a little cumbersome to try and explain in words. Basically, you just keep wrapping the cable around itself (clockwise) as you go around the circle. With a little practice, you will find that the windings form a pattern, and if you make a mistake it will be obvious as it does not fit the pattern. This coil tends to hold its shape better than the single knot style, and I generally just wind it by itself, not on a core. You should measure the diameter of the object you wish to place in the coil as a core when it is finished, and start with a circle a little larger than the diameter of the intended core.

I recommend using a Quartz, 'Moquis Marble', or Kyanite core for mobius coils, as the energy generated by a mobius coil (scalar waves) can be biologically disruptive when in its raw state. This information is primarily intended for those who wish to use mobius coils as a means of exciting Quartz crystals or ORgonite, and you are responsible for your own safety. By making this coil you agree not to hold me responsible for any damages your experiments may cause to persons or property. Mobius coils generate scalar waves. Scalar waves can interfere with and/or damage electronics when high voltage is put through them. For the purposes of ORgone research, low voltages are sufficient to drive mobius coils.

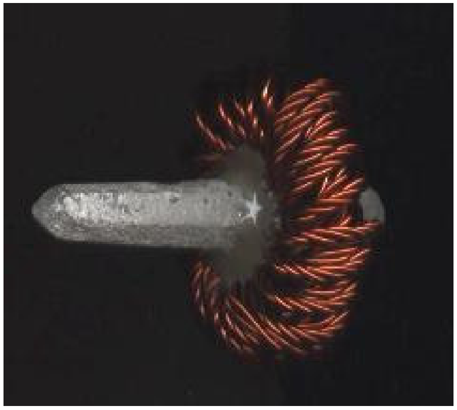

High performance – simple construction orgone generator – the Ring Generator

The design illustrated here is crude in execution and simple in construction. That's the point. My later versions of this device are more evolved than the one illustrated here, but I want to emphasize clearly that a device like this will work really well if you put it together right. This was my main generator for some time. I have posted the design here so that anyone who wants can have heavy-duty orgone generation capability without having to pay for a professionally built unit. It has been my experiences that the do it yourselfers will buy what they want to buy and build what they want to build anyways. A professionally built unit has many refinements over this crude design, but if you wanna do it yourself, here is a way to get a kick-ass if unrefined machine from stuff you probably already have... for about 50 bucks US, not counting the cost of the xtal. Interested?

This is a generator design I have been using for some time now. The design given here is highly adaptable, and can be modified to suit your needs. The pics here are from the first one that I built, and I used it for about 8 months for various things, with great success. The versions of this design I am using now are larger and more refined, but the pics shown here are the actual device which I used as my main generator for several months. In fact, the device is simple enough that it may be somewhat anticlimactic for you when you see how it's made.

The main things that I used this device, and continue to use devices based on this design... for, are as follows:

1- Heavy duty radionics applications, I.E. using the generator to send energy to persons or other entities through the Aether. Have done healing work, remote influence of persons and events, and in one case used this exact device to cut through the spells of a 'nameless' but well-known metaphysics practitioner like a hot knife through butter. Do not ask me for further details about who and when and why, because I will not provide them. It is immaterial. What matters is that the device works. I use the device for some of the radionics undertakings that require a lot of juice but not so much finesse. Finesse comes with refinement of the design and user skill.

2- Heavy duty orgone generation requirements, I.E. weather manipulation experiments, charging of water or other material, or psionics / intent amplification. You will have to figure out how to do that yourself, which if you are familiar with metaphysics should not be too difficult for you. The only input to it from you is by your intent, but you can reach out with your mind and direct the energy from this device same as with any other orgone generator. The frequency used to drive the device makes a lot of difference, and the use of select frequencies or audio signals to achieve desired results is a main aspect of using this device.

3- Charging and programming orgonite while the resin is curing, or cleaning, charging and programming quartz crystals. This use is the main reason why I am posting this design instead of limiting myself to offering much larger and more refined versions of it for sale. In due time I will come out with a production model of this device, but in the meantime it can radically improve the quality of orgonite made by the home orgone crafter. The production model of this device will contain much more orgonite that the one shown here. Pulsers can be used to charge resin while it cures as well. This device is a lot more powerful than a pulser, but less user friendly. Not everyone is going to be making enough orgonite to get their money's worth out of a unit like this when I decide to start selling the, but most people are definitely going to get their 50 bucks worth of use out of a homebuilt unit like this if they make much orgonite. Again, Pulsers, Powerwands and SPs are PORTABLE and therein lies one of their main beauties. This device is not portable, but rather requires both a stationary setup and the input of electrical power, and well as a home computer and an old stereo amplifier to drive it. This device offers more flexibility as well in that you can easily change out the core xtal for pretty much anything you want. Some materials may not be especially healthy for you when you expose them to scalar waves, and you are responsible for your own decision to make and use this device. Depending on how much voltage you put through it you may or may not harm electronics in the area. I have found that relatively low voltages on the RingGen suffice.

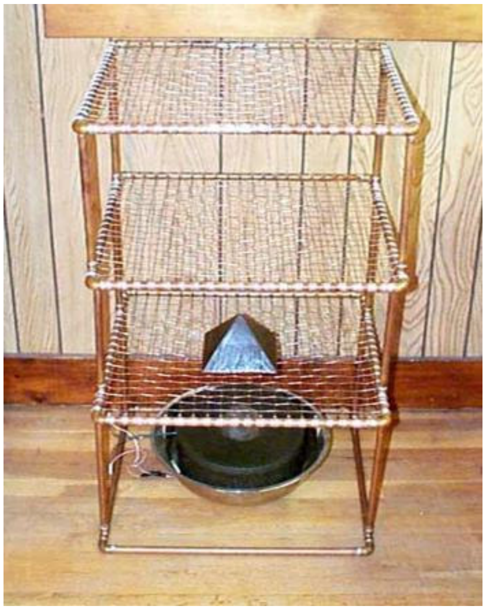

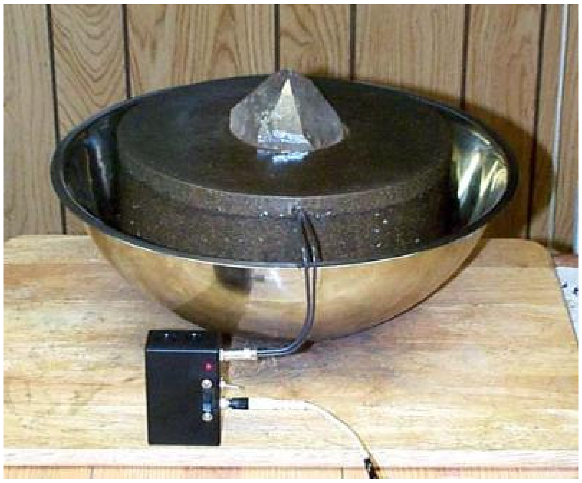

Some shots of a Ring Generator built by a man named David in New Jersey.

This is what you need to construct this device:

- 30 feet minimum of 18 to 15 gage copper wire.

- One large quartz crystal. There is a picture to the right of the xtal I used for the device shown. It does not have to be a nice DT like the one I used, but it should be about that size. In a pinch, you can also use several; smaller xtals together to get about the same amount of xtal mass. I got this xtal from ZSL, a generous gift on his part. He got it from the crystal man (in Canada). Another good place to get xtals is from revted (in America).

- 6 Plastic binder straps.

- Some aluminum tape.

- Enough stuff to make at least 1 liter of orgonite.

- The rings from 2 mason jar lids.

- A small shelving unit made from metal mesh of some kind. I used an old LP record rack, flipped upside down. You could make one from pieces of window screen stretched across a frame, or see what you can find in your garage or the second hand store. It should be a shelf that has the individual shelves made out of a metal GRID or SLATS, so that there is an open mesh, not solid pieces of metal. On the next one I make, I will probably use copper screens for the mesh. I have not tried using aluminum for the shelves, but I think steel or copper would be better for the shelves. Steel works well, I can attest to that.

- A large diameter hemispherical piece of metal to use as a reflector. I used the canister lid of an old steam cleaning rental unit I happened to have, you could use a large wok lid, or a large metal mixing bowl. It does not have to be a perfect parabolic curve, and aluminum will work for the reflector. Again, on the next one I make I will probably try to use copper. I imagine I will have to form a piece of copper sheet into a reflector. The reflector I am using right now is about 24" diameter, I would recommend it be at least 18" diameter. The metal should be bare on the inside of the reflector, it can be covered with paint or enamel on the outside.

- A long section of speaker wire (2 wires) to connect the device to your PC.

- A household stereo, boom box, or any other audio equipment that has an amplifier built into it, and also has a line in or auxiliary input jack so you can route the signal from your PC's sound card through it. I use an ancient 8-track player that has a 50 watt amplifier in it.

- Adapters to connect you PC sound output to the line in of the amplifier.

- Tone generator software for your PC. You can download it off the web, and there are some links provided on the links page of this site where you can find it. Roberto Forcen's stuff is amazing, as soon as I can afford it I will be sending him something for his efforts, in the meantime I can at least give him a plug, his site is called voicesync. He has wonderful donaware for sound generation and analysis. Also, NCH toner is a very adaptable and functional tone generator for your PC with a range from 1 to about 20,000 Hz with control over the waveform. You can also run multiple copies of the program and generate multiple tones at the same time.

Building it... Step 1 -

Build yourself a large mobius coil. Make it big enough you can fit a large xtal in the core, about 2.5" to 3" diameter. It does not have to be a fancy xtal. It can be single point. On the old device shown here I used a single knot mobius coil, but on my new one I am using the continuous knot style. Secure the coil with plastic binders to prevent it from losing it's shape. The coil shown here is from 15 gage magnet wire, and while I recommend using at least 18 gage, it does not have to be magnet wire. magnet wire just works a little better. It isn't necessary to use magnet wire.

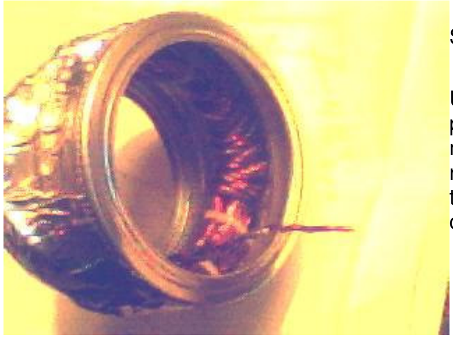

Step 2 -

Using aluminum tape, connect 2 mason jar lid rings to the coil as shown. Work the tape firmly into place, and use at least one full wrap of tape all the way around,. This is not only for the effect of a metal shroud on the coil, it is to keep the orgonite from leaking into the coil. Do your best to line the rings up evenly with each other, so that the ring on one end is parallel to and plumb with the ring on the other end. Pull the leads from the coil out through the center, or if you want to be more precise you can form a little hole in the tape to pass the leads through, and fill the hole with glue so it doesn't leak.

Step 3 -

Find a suitable mold that is cylindrical, and deep enough to come up to the top of the coil assembly you have just built. It should be at least deep enough that you can cast orgonite up to the top of the coil assembly, and wide enough that there is at least one inch of orgonite all the way around the outside of the coil. On the generator I use now there is a lot more orgonite than on the old one shown here. The old one shown here worked well enough to impress me several times with its effectiveness. Coat the mold with release agent of some form, regular vegetable oil or WD40 will both work fine. If you are using orgonite made from polyester autobody filler, you do not have to seal the bottom of the coil assembly to the mold. If you are using orgonite made from resin, you will have to seal the bottom of the coil assembly to the mold so resin does not leak under the coil assembly into the cavity at the center. You want to cast a ring of orgonite with a hollow center where the coil is. A little hot melt glue works well for this (sealing it), since it will keep the resin from leaking into the hole, and is easily removable. It also sets fast, so you don't have to wait for it to dry like you would silicone or goop glue.

Step 4 -

Cast it with MHD, HD or XHD orgonite. Allow the orgonite to cure, and remove the coil cast in orgonite from the mold. You should now have a hollow cylinder of orgonite with a coil in the center of it. Place the cylinder with the coil in the center of the reflector as shown below. You can fasten it with a little glue if you want. Then place the reflector on the lowest shelf of the metal shelving unit as shown. The quartz xtal goes in the hole inside the coil, as I imagine ought to be obvious. You can also easily remove and change xtals this way. I have a large cluster of kyanite which I use inside the coil too, gives a great energy. Let your imagination be your guide. Experiment with using different minerals and other things inside the coil, but be aware that not everything you put in there will work as well as a quartz xtal. A safer way to experiment is to ADD things in there with the quartz xtal, and see what kind of orgone they produce before you try them alone. Now you're ready to hook it up. If you want to, you can line the metal shelf up with the compass directions (I.E. so that one side of it faces north) but this is not necessary, it just makes it work a bit more efficiently). The metal shelves and reflector should be insulated from each other electrically, or at least that is the way I find it works best so far. The metal rings on the ends of the cylinder do not have to be electrically insulated from the metal of the reflector.

Not only do the metal screens seem to act like an broadcast antenna / amplifier, they also work really slick for making orgonite. Later on, when making orgonite, fire up the generator and put your molds on them while the resin is curing. The metal in the shelves pulses in rhythm to the scalar energy coming off the coil, and although I guess I still haven't figured out exactly why, they emit a strong stream of orgone as well. Further refinements to the design would involve, among other things, the proportions of the metal shelves in relation to the rest of the assembly.

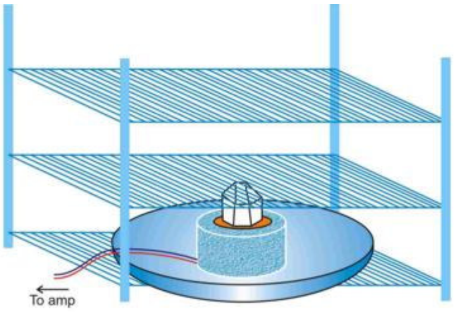

If you are going to use the device for radionics, then use the metal shelves as your witness well. You can also use the cavity inside the coil as witness well, but the energy sent to target that way may not be beneficial unless there is a quartz xtal in the coil along with it. If you just want to generate a big cloud of healthy orgone, then drop a large TB inside the coil and fire it up. If you have a wand or orgone beamer of some form, you can put that inside the coil and supercharge it.

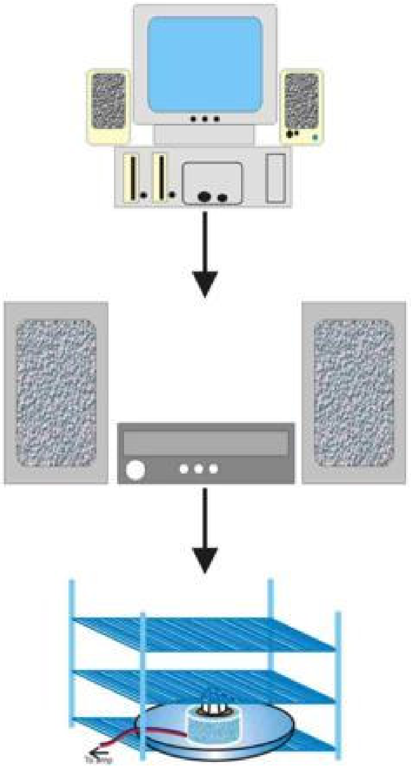

This is basically how you hook it up. You want to take the signal from your PC's sound card, route it through your amplifier, and then run it through the coil in the device. There are several ways you can do this, and no doubt those of you familiar with electronics will see that there are better ways than the method I have illustrated. I chose this method for this page because it is simple. Just use the adapter to connect the sound output from your PC to the line in or aux input of the amplifier. Then take one of the wire pairs leading to a speaker from the stereo, and cut ONE of the wires leading to the speaker. Leave the other wire intact. Preferably, use the positive wire. Then, connect the leads of the coil in the device to either end of the speaker wire which you have cut. This has the downside that you have to listen to the sounds being generated, but it PROTECTS YOUR AMPLIFIER from blowing. Mobius coils have very low resistance, and if you connect it directly to the amplifier output without having enough resistance you may damage your amp. The resistance in the method of connection shown is provided by the speaker. The current comes from the PC, is amplified by the amp, then goes through the coil. After it leaves the coil it goes through the speaker and back into the circuit.

Alternately, you could add a resistor to the mobius coil to give it enough resistance so that the amp would not perceive it as a short. In order to do this, you would need to find out what resistance your speakers are. Usually the speakers are marked (inside the speaker box on the back of the speaker magnet) with a little number beside an ohm symbol. This is the resistance. An ohm symbol looks like a little "O" with the bottom cut off and two little tails (Greek letter omega).

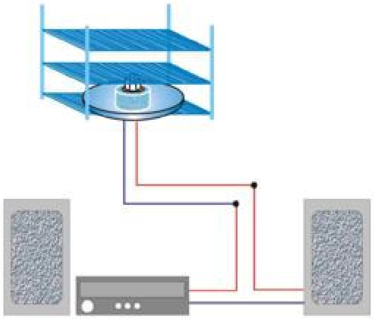

Again, alternately, you could connect the coil directly to your pc sound card by adding 8 ohm resistance to the mobius coil, as I gather 8 ohm is the standard resistance for small speakers like the ones used in PCs. In fact, I may be offering this (the ability to plug it directly into your PC sound output) as an option on the Pro Model Pulsers at some point in the future. For devices the size of the one shown on this page, I think it is better to have the extra power provided by the amp. If you use a device like this to charge your orgonite while it cures, you can increase the potency of your orgonite significantly. Some suggested frequencies: 15 Hz, 32 Hz, 5075 Hz. Obviously, one could build a more refined device based on this general concept, but hey, I can't give away all of my trade secrets ;). Besides, lots of folks won't mind that it looks hokey as long as it works

If you have access to the internet, you can check for the latest edition of this document at www.littlemountainsmudge.com in the info section of the site.

Legal notice:

Introduction to Orgonite™ copyright 2003 Jon Logan
Introduction to Orgonite™ second edition copyright 2004 Jon Logan

“Introduction to Orgonite™” and/or Introduction to Ergonite™” are compilations of articles written by Jon Logan, concerning Orgone Matrix Material, it’s manufacture and application. These compilations will be released in updated form as time progresses.

Orgonite ™ is a trademark belonging to Karl Welz, the inventor of orgonite, and is used on littlemountainsmudge.com and in this document for educational purposes

Wizzer's Workshop ™, Ergonite ™, Welzite ™, Wood Sprite ™, Water Baby ™, Chocolate Shine ™, O-Matrix Material ™ and all text / images on littlemountainsmudge.com, unless clearly otherwise stated, are copyright 2003 Jon Logan, all rights reserved.

Limited authorization to reproduce: This material may be reproduced or reprinted freely, in whole or in part, in printed paper or digital format for educational purposes. Additionally, the trademarks Ergonite ™ and Welzite ™ may be used freely by anyone for private or commercial purposes, so long as they are specifically used to refer to Orgone Matrix Material, a mixture of metal particles suspended in organic resin.

Owner of copyright for text and images on littlemountainsmudge.com and trademarks used on littlemountainsmudge.com reserves the right to refuse, suspend or revoke limited authorization to reproduce this material if the material is reproduced inaccurately, or if is sold at a cost in excess of it's production without prior written consent from the owner, or if printed material from this site is used commercially without stating in the printed copy that the material is copyright 2003 Jon Logan.

In reference to “Introduction to Orgonite™, **this document may be reproduced freely by anyone for private or commercial purposes without prior written consent so long as the following conditions are met:**

3 The document must be reproduced in it’s entirety and not altered in any way.
4 You may not charge a fee for distributing or reproducing this document in digital format.
5 You may charge a fee for distributing hard copies of this document, of not more than 20 Dollars CDN or 17 Dollars US, whichever is greater.

Email – wizzer@littlemountainsmudge.com
Phone – 250-446-2032
Mail – PO Box 139, Westbridge BC V0H2B0
Url – www.littlemountainsmudge.com
Introduction to Orgonite™ third edition ©2004 Jon Logan

Cloudbusting

montalk 20/11/98
all rights reserved

Part I: Microbusting An offshoot of orgonomy, weather engineering offers both electrifying experiences and burdens of responsibility for the cloudbuster operator. Though theories as to how or why it works are abundant, research by such notables as Trevor James Constable, James DeMeo, Robert McCullough, and, of course, Wilhelm Reich) has unequivocably shown that methods employed in cloudbusting *do* influence weather patterns. Further investigations by backyard experimenters as myself only confirm this fact.

My own weather operations started in May of 1996 and ended in September of 1997, and although seventeen months may seem insignificant to the years of research conducted by others, the months were sufficient to allow an assemblage of theories and protocols for the production of desired atmospheric precipitation.

Traditional practices in cloudbusting centered around the concept of *orgone* energy, claimed by its discoverer, Wilhelm Reich, to be a negentropic life force affecting physical, biological, and psychological systems depending on its charge density, rate of flow, and excitability. The various forms of orgone, such as DOR (deadly orgone), and ORANUR (orgone excited by radioactive energy), all have specific effects upon these systems, which are described in the following chart:

	physical	biological	psychological	atmospheric
Orgone	Negative ionic, affinity for water, negentropic, blue glow, life-positive, alters alpha decay of radioactive material.	Abiogenesis of bions, tingling warmth on skin, enhances healthy growth of life, stimulates cilia movement.	Free flow of emotions, sense of well being, vitality, rapidly discharges during orgasm.	In dense concentrations accompanied by cumulous clouds and/or precipitation, in low densities by aridity, adds clarity to blueness of sky.
DOR	Positive ionic, signifies lack or stagnation of orgone, created by TV's, microwaves, and other	Destimulates cilia, slows intake of oxygen, dehydrates the body,	Creates fatigue, blocks emotional release, deppression.	Black clouds without rain, desert-like arid stagnation of climate, stagnant

	electromagnetic sources.	causes severe thirst.		weather patterns, haze
Oranur	Densely positive ionic, orgone dangerously excited by alpha radiation (as found in smoke detectors), decomposes minerals into brown melanor.	Along with symptoms of radiation sickness, clears life-forms within its field of effect.	In buildings, creates an atmosphere of aggression, hysteria, claustrophobia, and uneasiness.	Shred-like decomposition of cloud edges, cobalt-blue sky.

All three forms of orgone have specific atmospheric effects due to its intense affinity for water; strong concentrations of "good" orgone are accompanied by water, and low densities by arridity. A cloud, for example, is a pocket of dense orgone energy, but the surrounding blue sky contains a lower concentration. According to classic cloudbuster theory, in order for a cloud to "bust," its orgone content must be "drawn" away by some means.

It was by accident that such a means was found. Wilhelm Reich and his associates observed that pipes pointed at a lake before them influenced the shape of waves upon the water. Years later, Reich capitalized upon this effect by relieving what was then known as the DOR Emergency, during which the inhabitants of his laboratory in Rangeley, Maine, became victims to "DOR sickness." He sought ways to rid the area of this deadly orgone which had not only sickened his co-workers, but also caused the static accumulation of dark, drought-like clouds in the skies above. What soon evolved was a contraption composed of twelve copper pipes, each 1.5 - 2" diameter and 10' long, whose bottom ends were connected in parallel to BX electrical cables which were dropped in a well to siphon the DOR from the sky toward the water within the well. After ridding the region of DOR, the cloudbuster was found to be useful for either producing or preventing rain in the region. For several years, Reich offered his services in preventing forecasted rainstorms for special events such as parades and other special events, as well as breaking several droughts beflicting the region.

The theory behind cloudbusting involved a monumental principle: orgone flows from low potential to high. In practice, this meant that to draw orgone from a cloud into a collection of pipes, the orgone potential of the pipes had to be made higher than that of the cloud. This was the purpose of grounding the pipes in water, prefferably running water. Since much orgone was already present in the water contacting the pipes, much more so than within the cloud, when the pipes were aimed at the center of the cloud, the orgone by means of the above principle naturally flowed from cloud to pipe and into the water to be flushed away.

Reich made an analogy between a lightning rods and cloudbusters: both discharge orgone from the sky (lightning was believed to be a super-dense bolt of electricity and orgone), except whereas a lightning rod does so in the fraction of a second, the cloudbuster extends its influence over tens of minutes -- an hour is the usual draw duration. Though lightning rods channel much greater magnitudes of energy than cloudbusters, it is not difficult to see why a cloudbuster produces a more apparent change in shape of a cloud than lightning rods do. It is not the *magnitude* of energy involved, but rather the *rate* of energy flow. With the cloudbuster's slower rate of orgone siphoning, the chaotically evolving cloud has a longer time to respond to the influence of a cloudbuster operator.

The rate of draw is dependent upon rate of water flow within the pipe, number of pipes, and the skill at which the cloudbuster is operated. The length-to-diameter ratio determines the range of influence of the cloudbuster; the larger the ratio, the longer the range. Requirements of fast-running water hint at the mechanism by which cloudbusters function. To fully understand how metal tubes pointed at the sky can change it, one must realize that orgone is both an *ionic* and *conscious* phenomena. The ionic interactions occuring within the cloud at a molecular level explain, for the most part, how and why cloudbusting is possible. As mentioned, orgone bears characteristics uncannily similar to electrical phenomena, specifically that of ionized air. Orgone and ionized air are compared as follows:

Orgone	Ions
Highly charged orgone accumulators placed in the dark are seen to emit a blue glow.	Objects ionized by high voltage electricity give off a blue corona.
A healthy dose of orgone "freshens" the air, creates an energized sense of well being.	Negative ion generators also freshen the air, stimulate cilia movement within air passages of subjects breathing negative ions -- leading to a better filtration of dirty air and increased oxygen utilization within the body.
DOR causes thirst, fatigue, irritability, and blank stares	Positive ions (lack of negative ions) decrease cilia movement, leading to throat irritations and thirst caused by lowered production of mucous and resulting increase of foreign particulates entering the lung, and oxygen utilization is lowered, also creating fatigue.

among its victims.	
Orgone is neutralized or absorbed by water.	Static electricity is also neutralized of absorbed by water.
Cloudbusters are like lightning rods, but draw orgone at a slower rate.	Lightning rods channel bolts of electricity, cloudbusters channel ions slowly.
Orgone can alter the decay rate of radioactive materials.	Ions are known to affect radioactive decay.

Quite clearly, orgone not only *resembles* electricity, but could actually *be* electricity with one exception: several behaviors of orgone cannot be fully explained by the physics of ionization alone, which leads to my hypothesis that the unnaccounted factor is of psychic origin, in other words, it is a *radionic/conscious* factor.

First, however, let's analyze the ionic aspects of orgone before we return to the radionic hypothesis in weather engineering. If we visualize the interaction between a cloudbuster and its target as being not one of orgone "drawn" into the pipe, but rather ions from the pipe being *released* into the atmosphere toward that target, then the generation of the ions and their resulting interaction with a cloud's water droplets can elegantly be explained.

One requirement when cloudbusting is fast running water, as mentioned above. The simplest of this type is a copper pipe with a spray nozzled hose inserted into its lower end, as shown in this diagram:

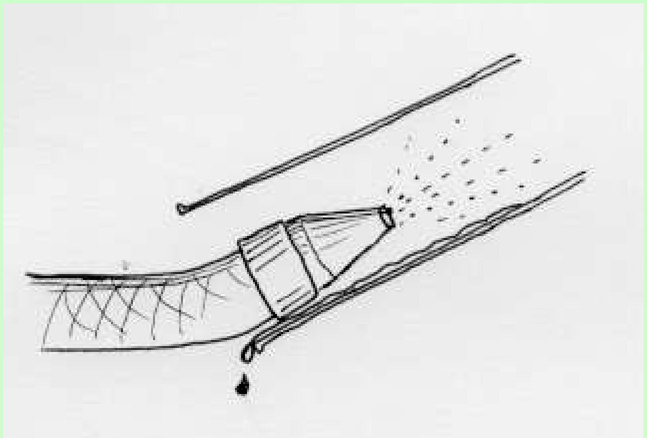

To allow full emersion of the pipe's end into water, a tub can be used:

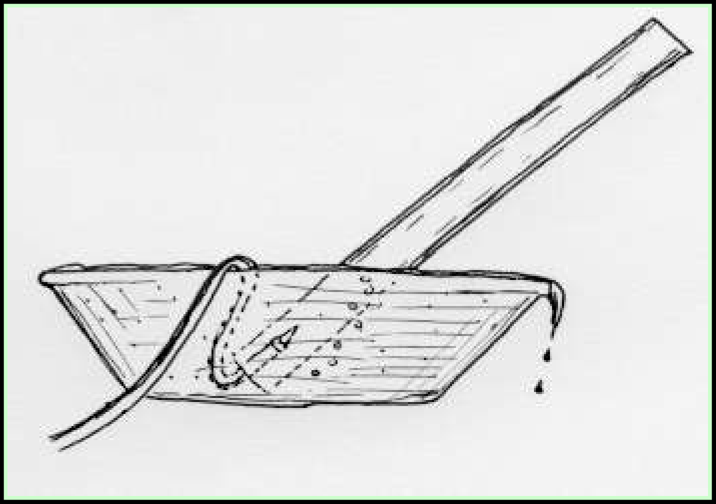

More advanced and efficient configurations involve atomizing the water into a fine mist before injecting them into the tube. A make-shift version of this type utilizes a motorized air humidifier commonly found at Wal-Mart for $15 during the winter season. A flexible, expandable hose is connected from the humidifier to the pipe:

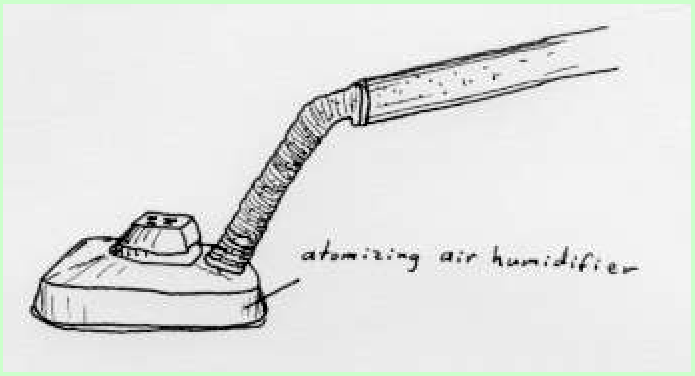

At the molecular level, neutral streams of water are atomized by the nozzle or humidifier into small droplets light enough to be suspended in air. These droplets, by a well known phenomena, are ionized; when water is split into smaller globules, charge imbalances occur. Often, more negative ions are produced than positive ones as in a rainstorm, where raindrops splashing on the ground shoot off negatively charged droplets (in rare instances, such raindrops have been seen to discharge sparks of electricity upon contact with the ground). Within the cloudbuster's pipe, much of the positive ions are grounded, creating in effect a filter mechanism by which negatively ionized water droplets are produced and forced from the muzzle of a cloudbuster. This explains why fast-running water is a recommendation: the faster the water, the larger the spray, and the more ionized droplets are created.

Ions leave the cloudbuster to dissipate into the atmosphere and interact with an aimed target, the cloud. Precision to which these ions are capable of being aimed shall be discussed below, for it is radionic fundamentals that explain the precision. Contrary to expectations the ions do not simply drift away like smoke from a pipe, but coherently reach their target in a conical, if not beam-like projection. Once the ions reach a water droplet in the cloud, the largely neutral droplet (which is really a collection of negative droplets around a positive condensation nucleus, the total charge being zero), is split into smaller negatively charged droplets now small enough to be classified as *water vapor*, which is invisible to the naked eye except for its filtering effect of the sun's rays causing our sky's blue hue. A cloud can therefore be made to "disappear" when its water droplets have been ionized by the cloudbuster's ions into invisible water vapor. Conversely, a section of blue sky, occupied mostly by uniformly charged positive water vapor ions, can be made to coalesce into visible water droplets through neutralization by the interaction with the cloudbuster's negative ions:

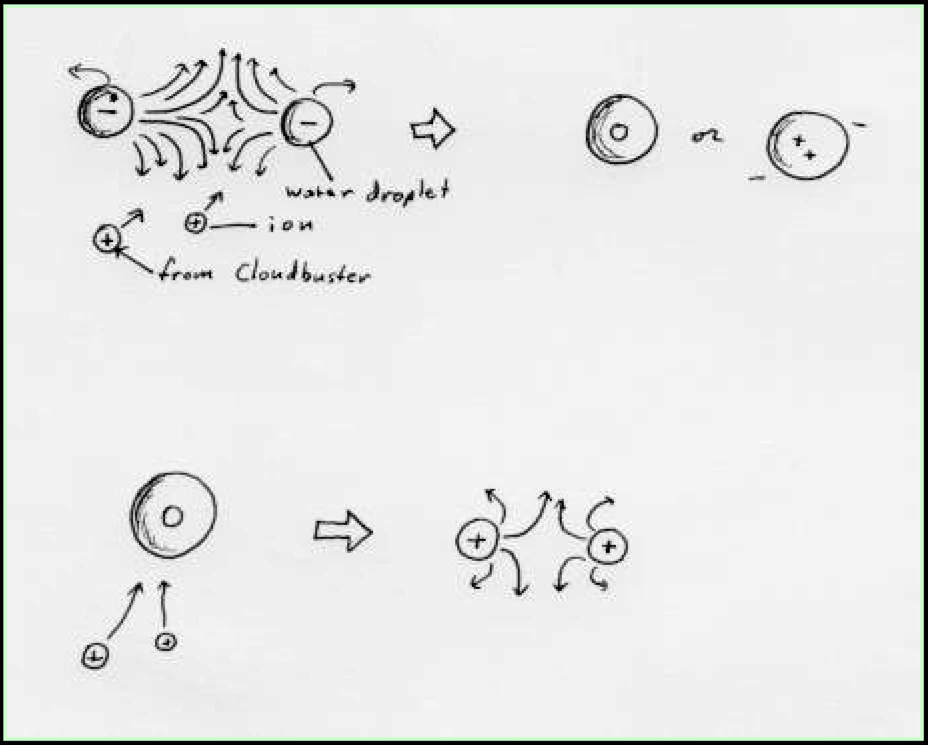

Because opposites attract and similars repel (Law of Attraction and Repulsion of Charges), large water droplets can either be made to split into two or more repelling water vapor molecules, or ionized water vapor molecules can be neutralized and caused to coalesce into a cloud.

Therefore, to destroy a cloud, the cloudbuster is aimed at the cloud, and the ions dissociate it:

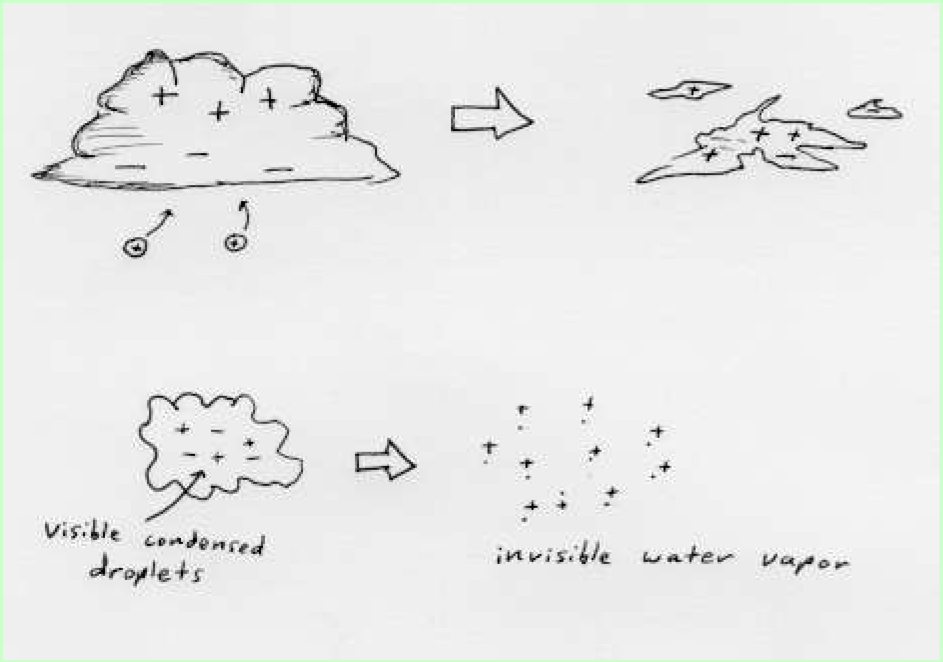

This is in complete agreement with classic cloudbusting protocol.

Now that ions are known to be partially responsible for weather modifying effects, one can see why waterless cloudbuster units are possible. Trevor James Constable, whose work is detailed in the book Loom of the Future (one I recommend you read for its philosophical insights and profuse number of photographs), mentions waterless units called "blue-bellies", which are in essence orgone accumulators wrapped around the lower half of PVC pipe and bound with duct tape. For those who do not know what an "orgone accumulator" is, visit the PORE webpage. Orgone accumulators are composed of layers of organic and inorganic material, most commonly cotton and steel wool in alternating layers. Such a configuration of materials is said to become charged with orgone, and radiate the orgone to objects near the accumulator. Any physics student can see that alternating layers of metal and insulator forms a capacitor. By virtue of its porous and fluffy texture, such accumulators gather electrical energy and emit a constant field of charged air ions. The source of this charge is gravity, known by Townsend Brown and George Hodowanec to produce self-charge on a capacitor. When an accumulator is wrapped around a pipe, plastic or metal as in the orgone "Volcano" model below,

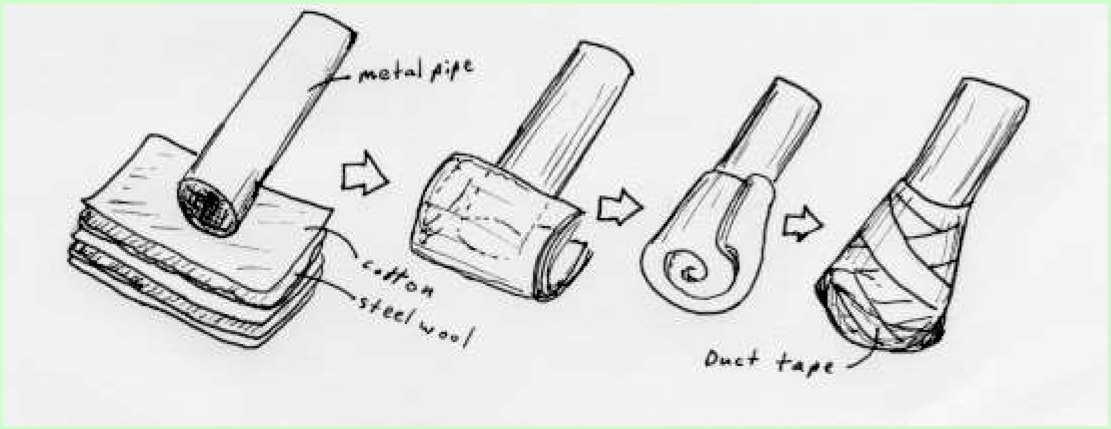

the pipe will emit ions just as a regular cloudbuster, except without the need for water.

This principle can thus be applied through a high voltage ion generator inserted into the end of a metal or plastic pipe and aimed at the sky. This method allows, in effect, a power-cloudbuster to be constructed which needs only to be pointed, set, plugged in, and monitored.

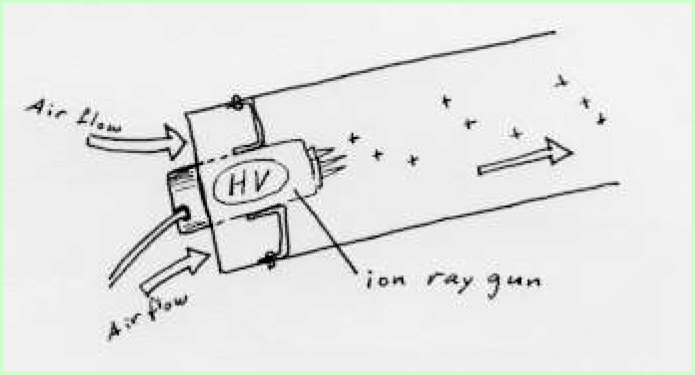

However, a question arises as to how so few ions can impact the development of thunderheads (cumulo-nimbus clouds hundreds of yards in height). The answer to this question is easily apparent and stems from the same reason why, under Chaos theory, the air currents generated by a butterfly flapping its wings could eventually lead to the formation of a hurricane. Cloud development, and weather, for that matter, is extremely chaotic and volatile to the smallest factors. A cloudbuster's ion can act as a a crucial trigger in causing millions of other water vapor molecules to coalesce in a chain reaction, much like cloud seeding utilizing dry ice dropped from airplanes into clouds also create a chain reaction.

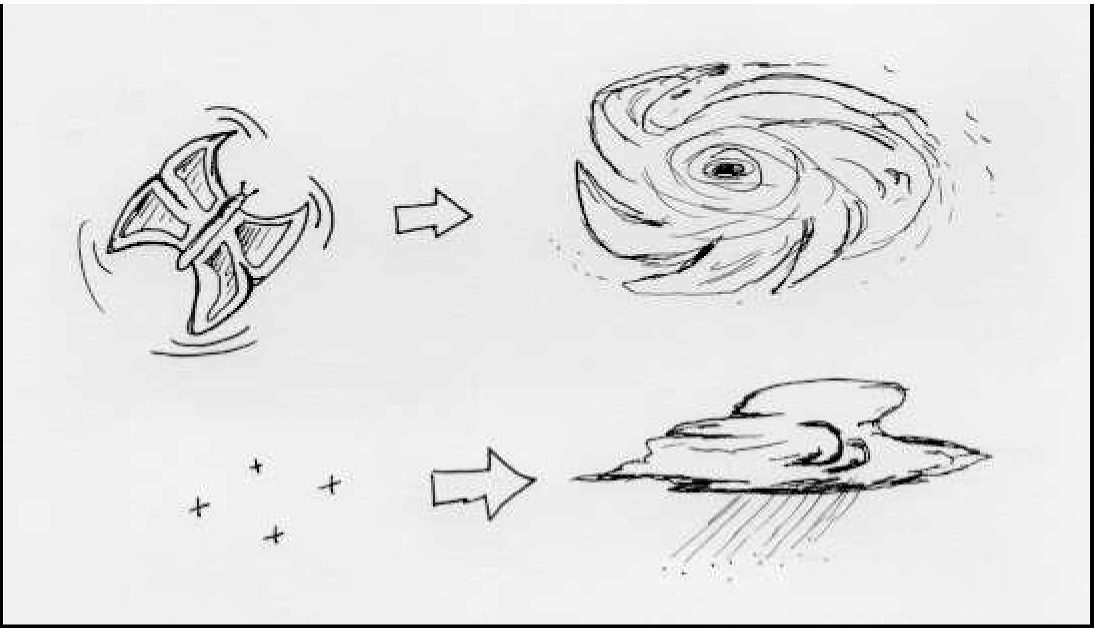

Unlike the butterfly effect, however, where the trigger's effect is chaotic, the effects of cloudbusting are controlled by radionic means through the mind of the cloudbuster operator. This is an example of consciousness affecting chaos. The more chaotic a system is, the more easily consciousness can both saturate and influence it. A cloud, or the weather itself, is quite chaotic. Just as a sculptor shapes a figure with his hands, so does a cloudbuster operator shape a cloud with his mind into a state more conducive to formation or dissipation, but not without the help of his tools, of course.

Cloudbusters are no different. Ions do, by evidence of their electrical properties, interact with water droplets in clouds, but it is the consciousness factor that tips the scale of the "butterfly effect" one way or the other, towards a thunderhead with a straight column, or a curved one, for example. Since the mind has a special connection to chaotic systems due to it being intelligent and resultingly unpredictable, the cloudbuster operator's mind can influence the otherwise unpredictable development of clouds *without* a cloudbuster, but unless he were extremely competent in parapsychological feats, his efforts would be in vain without the help of such a device. It is only through the interface of mind and machine that the average person can achieve the spectacular results of cloudbusting.

Orgone is thus a combination of mind and ion, whose effects range from wholly ionic (blue corona) to entirely psychological (emotional armoring).

Neither pure orgonomy nor pure present day science can comprehensively describe or explain the process of cloudbusting,

but a combination of orgonomy, radionics, and science can. Once again, it is consilience through which truth is revealed.

Part II: Macrobusting

www.montalk.net

THE ORGONE ACCUMULATOR HANDBOOK

Construction Plans, Experimental Use and Protection Against Toxic Energy

by James DeMeo, Ph.D.
With a Foreword by Eva Reich, M.D.

The Orgone Accumulator Handbook

by James DeMeo, Ph.D., with a foreword by Eva Reich, M.D., and an appendix examining new experimental evidence demonstrating the effectiveness of the orgone energy accumulator.

Learn how to concentrate and work with *orgone* (life) energy using simple, readily-available materials. A limitless supply is freely available from the atmosphere. These experimental devices have been shown to stimulate plant growth and human physiology in controlled studies.

The orgone accumulator is based on discoveries by the late **Dr. Wilhelm Reich**, whose work remains vital and of wide interest, despite unprecedented hostility to his scientific research. In a court decision unique in American history, his books were ordered *burned* by the U.S. Federal government in 1957. Dr. Reich reported medical benefits for his patients who used the accumulators.

In this edition Dr. DeMeo updates information on the construction of orgone energy blankets, shooters, box-type accumulators, bion packs and "draw buckets" for atmospheric cleansing, with many instructional diagrams and photographs, as well as a complete bibliography and resource guide for research.

About the Author: James DeMeo has been researching the topic of the life energy, and specifically the works of Dr. Wilhelm Reich since 1970. He has authored or edited dozens of published articles, books and research compendiums (including *Saharasia, Heretic's Notebook, On Wilhelm Reich & Orgonomy*, and the German-language *Nach Reich*) which subjected Reich's ideas to rigorous testing. DeMeo formally studied Environmental Science and Geography at Florida International Univ., and the Univ. of Kansas, where he earned his Ph.D. He served on the Faculty of Geography at Illinois State Univ. and the Univ. of Miami, and founded the Orgone Biophysical Research Laboratory in 1978.

ISBN 0-9621855-0-7
51295
9 780962 185502

List Price: $12.95

Books by James DeMeo

* *Saharasia: The 4000 BCE Origins of Child Abuse, Sex-Repression, Warfare and Social Violence In the Deserts of the Old World*

* (Editor) *Heretic's Notebook: Emotions, Protocells, Ether-Drift and Cosmic Life-Energy, with New Research Supporting Wilhelm Reich*

* *The Orgone Accumulator Handbook: Construction Plans, Experimental Use, and Protection Against Toxic Energy (Second revised edition)*

* (Editor) *On Wilhelm Reich and Orgonomy*

* (Co-Editor with Bernd Senf) *Nach Reich: Neue Forschungen zur Orgonomie: Sexualökonomie, Die Entdeckung der Orgonenergie*

* *Bibliography on Orgone Biophysics: 1934 to 1986*

THE ORGONE ACCUMULATOR HANDBOOK

Construction Plans, Experimental Use and Protection Against Toxic Energy

by

James DeMeo, Ph.D.

Natural Energy Works
Greensprings, PO Box 1148
Ashland, Oregon 97520 USA

Published by

Natural Energy Works
Greensprings, PO Box 1148
Ashland, Oregon 97520
United States of America

Tel/Fax: 541-552-0118 Email: demeo@mind.net

The ORGONE ACCUMULATOR HANDBOOK, Copyright © 1989, 1999 by V. James DeMeo. All Rights Reserved. Printed in the United States of America. No part of this book may be used or reproduced in any manner whatsoever without written permission, except in the case of brief quotations embodied in critical articles or reviews. For information, send inquiries to Natural Energy Works.

ISBN 0-9621855-0-7

Second Revised Edition, with Postscript 1999

7 8 9 10 11 12 13 14 15

Library of Congress Catalog Card No. 89-90975

Cover: NASA photo of Apollo 12 astronaut walking on the surface of the moon (see Life Magazine, 12 Dec. 1969). His body orgone energy field is softly glowing a blue color in the lunar vacuum, possibly due to excitation by his high-frequency radio communications equipment. This blue coloration of the energy field in the photo, which has been seen in a few other lunar astronaut images, has been systematically ignored or explained away as an effect of "lunar dust" or "water vapor". In fact, it is a visible expression of the human orgone (life) energy field.

Acknowledgements

This book is the product of many years of study of the prior works of Wilhelm Reich, and other dedicated scientists, without whose research findings it would not have been possible. The reader will find their names and research papers listed in the reference section of this book. I have corresponded with, and learned from many of these researchers over the years. In particular, I thank Eva Reich and Jutta Espanca for their constructive criticisms of the first edition of this Handbook. Additionally, I thank my own mentors, Robert Morris, Robert Nunley, and Richard Blasband, from each of whom I learned different things about life energy. I thank Theirrie Cook and Don Bill, loyal friends who have helped with the progress of my research in many ways, and also Jim Martin for the initial ideas and encouragement to revise my original small Handbook into the present edition. Jim also prepared the typesetting and most of the graphics in this edition, providing many helpful points along the way. Thanks to Stephan Müschenich and Rainer Gebauer for the courage they showed in doing their double-blind, controlled study of the human physiological response to accumulator treatment, at the University of Marburg, and for the Appendix document. Thanks also to the various researchers and doctors in Germany, who today are openly working with the accumulator in a manner presently impossible in the USA. In particular, I thank Heiko Lassek, Dorothea Opferman-Fuckert and Manfred Fuckert from whom I learned about the possiblities and limitations of physical orgone therapy. Also thanks to Vince Wiberg, for the simple and inexpensive methods for detecting electromagnetic disturbances, presented here. And finally, thank you Wilhelm Reich, for the discovery of the orgone energy and the orgone accumulator.

"We consider the discovery of the orgone energy among the greatest events in human history."– from a letter to the *American Medical Association*, signed by 17 medical doctors in 1949.

"THE ORGONE ACCUMULATOR IS THE MOST IMPORTANT SINGLE DISCOVERY IN THE HISTORY OF MEDICINE, BAR NONE" – Theodore P. Wolfe, M.D., from *Emotional Plague Versus Orgone Biophysics*, 1948.

"It is justifiable that the discovery of orgone energy and its medical applications by means of the orgone accumulator, the orgone shooter, bionous earth, and orgone water have opened up an abundance of new and, it appears, amazingly good prospects." – Wilhelm Reich, M.D. from *The Cancer Biopathy (Discovery of the Orgone, Volume 2)*, 1948.

"What would you say of the leading philosophers here to whom I have offered a thousand times of my own accord to show my studies, but who with the lazy obstinacy of a serpent who has eaten his full have never consented to look at the planets, the moon, or telescope? To such people philosophy is a kind of book... where the truth is to be sought not in the Universe or in nature, but (I use their own words) by comparing texts." – Galileo Galilei, Italian astronomer of the 1600s who proved that the Earth moved in the heavens, shortly before being prosecuted and tortured by the Catholic Church.

"...the orgone energy does not exist."– Judge John D. Clifford, from a 1954 U.S. court ruling in which all of Dr. Reich's books and research journals were banned, and ordered burned in incinerators; Reich was sent to a Federal Penitentiary, where he died.

CONTENTS

Page

1. Foreword, by Eva Reich, M.D. 1
2. Author's Preface 2

Part I:
The Biophysics of the Orgone Energy

3. What Is the Orgone Energy? 11
4. Wilhelm Reich's Discovery of the Orgone Energy and Invention of the Orgone Accumulator 13
5. Objective Demonstration of the Orgone Energy 24
6. Discovery of an Unusual Energy by Other Scientists 30

Part II:
The Safe and Effective Use of
Orgone Accumulating Devices

7. General Principles for Construction and Experimental Use of the Orgone Accumulator....... 42
8. Warnings and Dangers: Oranur and Dor 48
9. Cleaning Up Your Bioenergetic Environment 63
10. Draw Buckets, Bion Waters, Bion Packs 72
11. Physiological and Biomedical Effects 78
12. Personal Observations with the Accumulator 86
13. Some Simple and Not-So-Simple Experiments with the Orgone Accumulator 93
14. Questions and Answers 104

Page

Part III: Construction Plans for Orgone Accumulating Devices

15. Construction of a 2-Ply Orgone Blanket 108
16. Construction of a 5-Ply Garden Seed Charger 114
17. Construction of a 10-Ply Orgone Accumulator 118
18. Construction of an Orgone Shooter Funnel 124
19. Construction of an Orgone Shooter Wand 125
20. Construction of a 3-Ply Human-Sized Orgone Energy Accumulator 126

Selected References 135

Sources of Additional Information:

On Wilhelm Reich and Orgonomy 145
On Alternative Health Clinics and Institutes 147

Appendix: "The Psycho-Physiological Effects of the Reich Orgone Energy Accumulator", by Stefan Müschenich and Rainer Gebauer 148

Postscript 1999 157

CONTENTS

PAGE

Part III

Contributions [illegible]

[illegible]

[illegible]

[illegible]

1. Foreword

At last, thirty two years after the 1957 death of Wilhelm Reich, human beings can begin to study orgonomy like any other body of knowledge, helped by the *Orgone Accumulator Handbook*. This concise and informative book contains in a nutshell a condensed, clear account of the discovery, made usable by all who are interested in the cosmic life energy. Herein is printed: the scientific definition of the orgone energy; the history of how the steps of observation, experimentation, and theoretical insight led Reich to practical applications; the principles for construction and experimental uses of the orgone energy accumulator, with detailed suggestions for needed material, layering, and dimensions; and finally a very useful reference list. Professor J. DeMeo shows his thorough knowledge of the subject, which as yet is banned and omitted from the 20th Century academic curriculum, except for a few pioneering lecture courses (in New York and West Berlin).

Wilhelm Reich said that even though the life energy had been known for thousands of years, he managed to make it concretely usable, and that the era of its applications has just started. However, this *Handbook* is the first printed material in recent years on specifically how to concentrate the energy from the Earth's atmosphere. It is usable for a laboratory course on the subject of cosmic life energy. This material could be comprehended by intelligent high school or college students. It answers my almost fifty year old hope, for the inclusion of life energy facts in the body of knowledge that all educated people on Earth should learn in their schooling. Thank you James DeMeo.

Eva Reich, M.D.
West Berlin, March 1989

2. Author's Preface

When I was 12 years old, a favorite uncle of mine died a suffering death from lung cancer. The doctors had removed one of his lungs, and for a few months he lingered, being incapable of talking or moving very much, and in a great deal of pain. My aunts would not allow the children to see him in such sorry shape, except for one time, when he was dressed up for the entire family, which had gathered to quietly say good-bye. I was very sad when he died. When I was 15, my mother was diagnosed as having breast cancer. I was by her bed in the hospital when she recovered from surgery, and when she was told that her breast had been amputated in a radical mastectomy. I will never forget the look on her face. She survived the surgery, but the emotional resignation which she carried, and which preceded her cancer by several decades, was never diagnosed or discussed. Friends of our family had urged us to look into some alternative treatments for cancer, but everybody believed that the doctors in the hospital knew best. Listed as a "survivor" in the cancer statistics, my mother progressively declined after the surgery, and died about eight years later, having refused to undergo additional surgery.

My experience with relatives dying of cancer is not unusual, as degenerative disease is now at epidemic levels. The statistics today demonstrate that the "war on cancer" has been lost, and that in spite of all the radical surgeries, drugs and radiation treatments, patients today survive no longer or more frequently than they did in the 1950s. Indeed, degenerative disorders have today spread into youthful age groups and populations where they once were rare. Scientific evidence does not exist to support the assertion that surgery, radiation, and chemotherapy are effective forms of treatment for cancer, and traditional medicine today gives hardly more than lip service to preventative considerations. These troubling facts become all the more disturbing when one begins to study the various alternative, non-invasive and non-toxic cancer therapies. Dismissed for decades as "quackery" by organized medicine, most of these therapies appear to be reasonably or even remarkably effective. Their advocates and practitioners have often taken great risks to bring what they believe are safe and effective treatments to sick people. The organized medical community, with financial links to the pharmaceutical industry, has not cared to seriously look into these techniques. Instead, the techniques have been unwarrantedly attacked, and pseudo investigations have been launched, with predictable outcomes: clinics have been shut by brute police force, through court orders; medical records and research protocols have been seized, and jail sentences have been handed out. Books have also been burned. In this context, a great fraud has been perpetrated upon the American people, and upon our courts and legal system, by the larger organized medical associations, and related government bureaucracies.

In this short *Handbook*, I cannot give a history of these antiscientific and unethical abuses, but a few articles and books on the question are listed in the reference section. Clearly, a major reason for the impotence of modern medicine in dealing with degenerative diseases lies in the fact that the organized medical community has used police-state tactics to suppress important new findings, and the unorthodox practitioner, irrespective of any scientific evidence that exists. In fact, the most well documented and effective unorthodox therapies have been the most hotly attacked. Many, many promising therapies have been suppressed over the years, through an emotional collusion, and some outright conspiracy for economic motives, between medical interests, drug companies, plague journalists, prejudiced academics, and puritanical government bureaucrats and judges. In this process, the truth has been trampled badly, and the methods of science, discarded.

The most clear and telling example of how these social forces combine to kill a new discovery, and its discoverer, is the case of Dr. Wilhelm Reich and his orgone energy accumulator. Reich had been one of Freud's younger co-workers, and was a prime mover in the early psychoanalytic movement in Vienna and Berlin. However, his ideas were more revolutionary than those of the older psychoanalysts. He forcefully argued that human misery and mental illness were partly the result of social conditions, which could be changed to prevent neurosis. He urged reform in laws regarding the abusive treatment of children, and in the economic plight of women. He formed a viable social and political movement in pre-Hitler Germany, which urged reforms in laws restricting divorce and access to contraception, which were both then illegal. He championed the rights of young unmarried people to a healthy sexual life, and a loosening of the father's power and dominance over the family. He wrote extensively about these matters in the 1920s and 1930s, and exposed the roots of the Nazi movement in the obedience-demanding, patriarchal, sex-negative German family structure. For his writings on genitality, and his anti-fascist activities, Reich was eventually expelled from Freud's inner circle, and also from the International Psychoanalytic Association. German psychoanalysis was at that time leaning towards appeasement of the Nazis, and some analysts, such as Carl Jung, even became spokespersons or apologists for National Socialism. Reich was eventually placed on both Hitler's and Stalin's death lists in the 1930s, and had to flee to Scandinavia, and later from there to the United States. His writings were condemned to flames in both Germany and Russia.

By the time Reich arrived in the United States, in 1939, he had made a number of major scientific discoveries, and quickly attracted a group of young, enthusiastic scientists and doctors to assist with his work. The American period of his research, which lasted until his death in 1957, was particularly productive, in spite of the later court actions against him. It was during this period that Reich experimentally clarified, and made practical use of the biological and atmospheric life energy, which he called the orgone energy. The orgone energy, Reich observed, was a real, physical energy, which radiated from microbes, animals, humans, and from inorganic materials as well. It charged living creatures, yet also existed in the atmosphere in a free form. It could be accumulated in special enclosures, where it could be observed, felt, and measured.

Reich's orgone energy experiments attracted the hostile criticisms of many in the medical community, and a smear campaign in the press triggered an investigation by the U.S. Food and Drug Administration (FDA). In its efforts to stop Reich's work, the FDA did not seek to responsibly or accurately reproduce his experimental findings. Instead, FDA bureaucrats relied upon gossip and rumor, ignored available published evidence, and secured the testimony of "expert" witnesses who had no real familiarity with or interest in the scientific facts involved. Many breaches of the scientific method, and of legal, moral, and scientific ethics occurred, as government officials, and a number of psychiatrists, analysts, and physicists, sought to put an end to his work.

We also know a lot more today about the inner workings of the FDA regarding the Reich case, given that several scholarly reviews of FDA files, using the *Freedom of Information Act*, have taken place. These are cited in the reference section. Reich originally offered to cooperate with the FDA investigators regarding the orgone energy. But they refused his offer, as they were hot to "get Reich" on whatever charges they could, and had been urged in this direction by various highly-placed individuals within the medical communities. The *Bulletin of the Menninger Clinic*, and the *Journal of the American Medical Association*, for example, participated in the spreading of false gossip and rumors regarding Reich's work. Reich was aware of these unethical actions, and a number of his co-workers had been professionally hurt by the lies, and by the actions of the FDA. At no time did any governmental or private scientific or medical organization attempt to seriously review or replicate his experiments.

These kinds of attacks made Reich understandably furious, such that when the FDA finally sought a Complaint for Injunction against his activities, he refused to appear in court, to act, as he put it, "as a 'defendant' in matters of basic natural scientific research". Instead, he wrote a compelling *Response* ("Motion to Dismiss") to the judge. Reich's refusal to grant authority to the courts regarding the validity of his orgone research prompted a legal over-reaction, the likes of which we usually associate with despotic societies. In a judicial ruling that is, to the best of my knowledge, unique in American history, the FDA sought and obtained a Federal Court Decree of Injunction, which ruled that the orgone energy "does not exist". The court ordered a ban on the interstate shipment of books bearing the word "orgone". This included books into which the forbidden word appeared only in the preface, or introductory remarks. Additionally, all books and research journals discussing the orgone energy in detail were *ordered destroyed*, and devices using the energy, dismantled or destroyed. (Case #1056, March 19, 1954, US District Court, Portland, Maine, Judge John D. Clifford, Jr.)

"BANNED, until expunged of all references to the orgone energy:
- *The Discovery of the Orgone*
  - *Vol. I, The Function of the Orgasm*
  - *Vol. II, The Cancer Biopathy*
- *The Sexual Revolution*
- *Ether, God and Devil*
- *Cosmic Superimposition*
- *Listen, Little Man*
- *The Mass Psychology of Fascism*
- *Character Analysis*
- *The Murder of Christ*
- *People in Trouble*

BANNED and ORDERED DESTROYED:
- *The Orgone Energy Accumulator, Its Scientific and Medical Use*
- *The Oranur Experiment*
- *The Orgone Energy Bulletin*
- *The Orgone Energy Emergency Bulletin*
- *International Journal of Sex-Economy and Orgone Research*
- *Internationale Zeitschrift fur Orgonomie*
- *Annals of the Orgone Institute*"

And so, in the late 1950s and early 1960s, Reich's books and research journals, even those which were "only" banned, were periodically seized by FDA agents and Federal Marshals, and burnt in incinerators in Maine and New York. No scientific or professional organizations, journalists or writer's unions publicly objected to the book burnings, or acted to help Reich, whose laboratory headquarters was invaded by FDA agents, who destroyed instruments with axes. In addition to the above actions, the court ordered Reich to cease "disseminating information" on the orgone energy, effectively censoring his writing and speaking on the subject. Several years later, Reich was charged with Contempt of Court when his assistant committed a technical violation of the Injunction, at a time when Reich was more than a thousand miles away, engaged in field work in the deserts of Arizona. Though he appealed all the way to the Supreme Court, Reich lost the case, and was incarcerated in Lewisburg Federal Penitentiary, where he died in 1957. His death in prison occurred two weeks prior to his parole date, at a time when he was happily anticipating his freedom, and a life in Switzerland with his new wife.

Whatever we may think of Reich's response to the court challenge, the principles upon which he stood were very important, and date back at least to Galileo's trial by fire with the Catholic Church. The lesson from Galileo's time was that no Court, Tribunal, or religious or scientific organization on Earth has the capacity to say, on the basis of textual comparisons or divine revelation, just what is or is not Natural Law. The results of an experiment cannot be judged by those who have never reproduced it, and the unresearched opinions of scientists are no better than the unresearched opinions of anyone else, be they members of the American Medical Association, the National Academy of Sciences, or the same Country Club attended by the President. Galileo urged his critics to "look into the telescope", to verify his observations in a most direct and simple way. But they refused to do so on moral principle, and derisively mocked him. Reich's critics have taken the same approach, in their adamant refusal to reproduce his experiments, and in most cases, to even review the published evidence. Today, over 30 years since Reich's death, his most vocal critics still take the same anti-scientific approach, and condemn what they have not personally read or investigated.

Based upon these historical facts, it is clear that **the FDA, and indeed, all courts, academic bodies, and governmental agencies of all kinds, have forever relinquished any right to say anything about what the average citizen may or may not do with respect to the orgone energy accumulator**. The discovery of the orgone is in much safer hands among the average citizen, than in the hands of the various politicians, academies of science, and medical organizations. This *Handbook* is therefore not primarily directed towards an academic or medical audience. Instead, the case of Dr. Wilhelm Reich and the orgone energy accumulator is taken directly to the general public. Like the sunlight, the air and water, the orgone energy is a part of nature, existing everywhere, and ought to be available to everyone, free of restrictive regulation and control. As an invention, the orgone accumulator is also now in the public domain, unpatentable, and cannot be dominated by any single individual or corporation. It is also still *perfectly legal* for citizens to build, own, and use orgone accumulators.

Of course, with this right comes a great deal of responsibility, as the proper use and maintenance of an accumulator makes both social and environmental demands upon its owner. The cosmic orgone energy ocean is, like our air, food, and water, fast becoming contaminated and poisoned, and swift action by concerned people is required if life on the planet is to continue. This *Handbook* will give a basic overview of the orgone energy, the accumulator, and the construction and safe use of orgone accumulating devices. For the more precise scientific details and data, the reader is encouraged to obtain and review the published materials listed in the Reference and Information sections.

Three years after Reich's death, the trustees of his estate arranged for republication of his works. Today, most of his books have been republished, or are available in libraries, and through other sources. A group of Reich's co-workers also founded the American College of Orgonomy, and by 1967 had published the first issue of the *Journal of Orgonomy*. Since that time, interest in Reich's works has gradually increased, and many new studies verifying his findings on the orgone energy, and the accumulator, have taken place. College courses focusing upon Reich's life and works now exist, and he has been the subject of many reviews, biographies, and films. A younger generation of scientists and health practitioners is rediscovering Reich, and new journals devoted to orgone research or bioenergetic functions have appeared in the USA, Germany, France, Italy, England, Canada, and Japan. One major new study on the accumulator from Germany is summarized in the Appendix.

The effort to kill the discovery of the orgone has failed.

irritated protoplasm. It can be registered on specially adapted Geiger counters. The orgone also is the *medium* through which electromagnetic disturbances are transmitted, much in the manner of the older concept of *aether*, though it is not itself electromagnetic in nature.

Streamings of orgone energy within the Earth's atmosphere affect changes in air circulation patterns; atmospheric orgone functions underlie the buildup of storm potentials, and influence air temperature, pressure, and humidity. Cosmic orgone energy functions also appear to be at work in space, affecting gravitational and solar phenomena. Still, the mass-free orgone energy is not any one of these physicomechanical factors, or even the sum of them. The properties of the orgone energy derive more from life itself, much in the manner of the older concept of a *vital force*, or *élan vital*; unlike those older concepts, however, the orgone also has been found to exist in a mass-free form, in the atmosphere and in space. It is primary, primordial cosmic *life energy*, while all other forms of energy are secondary in nature.

In the living world, orgone energy functions underlie major life processes; pulsation, streaming, and charge of the biological orgone determines the movements, actions, and behavior of protoplasm and tissues, as well as the strength of "bioelectrical" phenomena. Emotion is the ebb and flow, the charge and discharge of the orgone within the membrane of an organism, just as weather is the ebb and flow, the charge and discharge of the orgone in the atmosphere. Both organism and weather respond to the prevailing character and state of the life energy. Orgone energy functions appear across the whole of creation, in microbes, animals, stormclouds, hurricanes, and galaxies. Orgone energy not only charges and animates the natural world; we are immersed in a sea of it, much as a fish is immersed in water. More, it is the medium which communicates emotion and perception, through which we are connected to the cosmos, and made kin to all that is living.

4. Wilhelm Reich's Discovery of the Orgone Energy, and Invention of the Orgone Accumulator

Reich's initial work on the question of a biological energy began in the 1920's, when he was a student of Sigmund Freud, the originator of psychoanalysis. Freud's early theories on human behavior discussed in metaphorical terms the energy of the drives, which he termed the *libido*. While Freud and most other analysts eventually ceased using this term, Reich found it to be a very useful concept, and he continued to seek evidence for this force, which governed human emotion, behavior, and sexuality.

Reich's extensive clinical work led to the observation of *vegetative streamings* or *currents* of emotional energy in the body, which occurred in healthy individuals during states of great relaxation, as following a strong release of emotion, or after a very gratifying genital orgasm. The free, uninhibited expression of emotion, and natural sexual excitation and gratification during orgasm were identified by Reich as expressions of unimpeded energetic movement in the body. When the individual experienced great pain, as from childhood traumas, when the emotions were rigidly suppressed and held back ("big boys don't cry", "nice girls don't get angry"), or when chronic sexual stasis and starvation was experienced, the entire nervous system and musculature participated in the process of emotional suppression, or warding off of feeling. This "holding back" of feeling was also accompanied by a greater or lesser anxious retreat from pleasurable, or even potentially pleasurable situations, which would otherwise stir up suppressed and unpleasant feelings. Reich observed that when this kind of response to feeling and pleasure became chronic, so too did the individual experience a chronic stiffening and desensitization of the body, along with a reduction in respiration and contactfulness.

This chronic neuromuscular *armoring*, as Reich called it, was not a natural condition, though it had a certain rational survival value for situations of pain and trauma. When the armoring became chronic, however, as a *way of life*, it would thwart the individual's natural biological functioning, and affect their behavior even in circumstances where pain or trauma was not likely. The armoring effectively perpetuated the individual's pleasure-avoiding behaviors and emotion-censoring attitudes. Deep seated fears, and pressures to conform to the prevalent armored form of social life, usually prevented the individual from moving towards emotional health, or taking effective steps to change their situation. The bulk of Reich's early writings focused upon these social, sexual, and emotional concerns.

Reich also argued that the heterosexual genital orgasm played a central regulatory role in the energy economy of the individual, as a means to periodically discharge accumulated bioenergetic tension. The more intense the orgastic discharge of accumulated bioenergy, the more gratified, relaxed and pleasurably expansive one felt afterward. When sexual urges and other emotions were chronically frustrated, dammed-up, and repressed, however, great internal tension could build up to a bursting point, where neurotic symptoms or sadistic urges would appear. Reich developed therapeutic techniques for releasing dammed-up emotional energy within his patients, techniques which led to the release of long-buried feelings, and to a greater capacity for pleasure in life, particularly genital pleasure. As his patients became more healthy sexually, and as they reported an increase in genital gratification, he observed that their neurotic symptoms disappeared, as the quantity of dammed-up emotion and sexual tension was reduced. Some of Reich's early contributions to psychoanalytic theory and technique were at first welcomed. But later on, as he increasingly focused upon the consequences of child abuse and sexual repression, the more conservative analysts rejected and attacked him.

Reich eventually left psychoanalysis altogether, and he articulated his work under the new term, *Sex-Economy*.

Reich's early observations regarding human behavior, emotions, the orgasm, and vegetative streaming sensations strongly suggested a real, tangible nature to the emotional energy. He later used sensitive millivoltmeters to confirm this point of view, and to quantify bioelectrical energy currents, and their emotional correlates. However, he was convinced that the very low levels of observed bioelectrical activity could not fully explain the powerful energy forces observed in human behavior. This was particularly so regarding chronic immobilizing psychic disturbances in catatonic and other completely withdrawn mental patients. When their emotions were finally broken lose, these patients would experience a tremendous outpouring of sadness or rage. Afterward, they would also experience a dramatic relaxation of musculature, a spontaneous deepening of respiration, and a return to more contactful lucidity. In these cases, the patient's emotional energy was held down and bound up, until finally set free in the clinical setting. These observations of energy bound, and energy released, were reinforced by parallel observations regarding the discharge function of the orgasm. Based upon these kinds of observations, the question of exactly how and from where the organism acquired its emotional energy, and its exact nature, became increasingly important.

It was at this point in his research that Reich was forced to flee Germany for Scandinavia, following Hitler's rise to power. In Norway, Reich sought to find a way to confirm his model of human functioning. Pleasure, he observed, was identified by an increasing bioelectrical charge at the skin surface, while anxiety was accompanied by a loss of this same peripheral bioelectrical charge. Persons with a deep respiration and a relaxed posture would, he observed, regularly give stronger readings at the millivoltmeter than contracted, anxious, highly armored persons, who had a life-history of trauma, abuse, repressed emotion, and ungratified sexuality. As a child grew to adulthood, and became habituated or conditioned towards pleasure-seeking, or pleasure-avoiding (pain seeking) behaviors, so too would their skin charge, and other physiological measures, reflect a corresponding high or low energy charge. This movement of the organism, and their energy charge, in a direction "toward" or "away" from the world, he argued, was the result of one's life history. Life naturally moved towards pleasure, but retreated, and shrunk from pain. Chronic painful experience would eventually armor the organism, and make it difficult for them to reach out, towards the painful world. From this central set of observations, he postulated that a similar process could be duplicated and observed in lower organisms, such as the snail, earthworm, or even microscopic ameba.

Reich noted that the ameba had no "nervous system", or "brain", as with the higher animal, yet it expanded towards or contracted away from its environment in a manner similar to the higher animals. He believed that many of the functions attributed to the brain were really functions of whole-body processes, involving the participation of the autonomic nervous system, but primarily being the result of the energetic forces he had documented in a clinical and laboratory setting. These currents of biological energy, he argued, functioned the same in all living creatures, and he sought to test the idea by making millivoltmeter measurements of ameba during states of expansion and contraction. Reich went to the University of Oslo Microbiological Institute, and asked to obtain a culture of ameba. He was told that these kinds of simple organisms were never kept on hand in stored cultures, because they could be cultured directly from a moss or grass infusion. Reich was fully aware of the air germ theory, but was surprised to hear this, as the theory had not at that time been used to explain the genesis of more complex microbes, such as ameba and paramecium. These more complex microbes cannot be cultured directly from the air, for example.

Reich made the moss and grass infusions, but also made extended and careful microscopical observation of the process whereby the ameba developed. He did not see spores on the grass blades, swelling up to become new ameba. Instead, he observed that the moss and grass itself would disintegrate and break down into small blue-green vesicles. The tiny vesicles would, over a period of several days, develop and clump together, after which a new membrane would form around the clump; the clump of vesicles would roll and pulsate inside the membrane for a period, and eventually the whole thing would move away on its own, having *turned into a new ameba*. Moreover, Reich observed that a number of materials, both organic and inorganic, would, when allowed to disintegrate and swell in a sterile nutrient solution, form the tiny blue-green vesicles. These observations were greeted by the university microbiologists with skepticism, and Reich developed a series of stringent control tests to answer their objections, and to more clearly demonstrate the observable process. These control procedures involved lengthy autoclaving of nutrient solutions, and heating over flame, to incandescence, of the materials placed in the sterile nutrient medium. His control procedures and observations on this question were repeated and confirmed by other scientists of the day, and were presented to the French Academy of Science in 1938. But this did little to satisfy his critics, who shamelessly refused to reproduce the experiments, while simultaneously attacking him in the Norwegian newspapers.

Reich used very high magnifications, around 3500 to 4500 power, but not the usual microbiological stains or procedures which kill the life in the specimen. These facts made Reich's preparations very different from those of the average microbiologist, who to this day still kill and stain their preparations with a religious fervor, and see little value to observing living microbes in the light microscope above 1000 power. Standard electron microscope images, for example, can not be made of living specimens.

Reich gave a new name to the unusual microscopic vesicle he had discovered: the *bion*. Bions of similar size, shape, and motility would appear in the light microscope when various materials were subject to a process of slow swelling and disintegration, or when substances were heated to incandescence, and then immersed into sterile nutrient solutions.

Boiling, autoclaving, or heating samples to glowing incandescence would not eliminate the bions from the cultures, but could actually liberate them in greater numbers. Reich also studied the process of disintegration and decay of foodstuffs in the microscope, and noted that similar bionous processes were at work. The bions exhibited a *bluish* coloration, and radiant energy effects were likewise observed. It was during these microscopic observations of the bions that Reich first discovered the orgone radiation, and later, the principle of the orgone energy accumulator.

Like his findings on human behavior, Reich's bion experiments are far too intricate and important to be fully reviewed here, but it can be noted that they have been widely replicated by various scientists around the world. Classical microbiology of today has made confirming discoveries of very similar small vesicles, though Reich's priority has yet to be acknowledged. His findings on the bions also resolved two parallel riddles, the origins of protozoa from disintegrated dead plant tissue in the natural environment, and the origins of protozoan *cancer cells* from the energetically (emotionally) deadened tissues of the human body. Reich observed similar processes at work in both dead grass and deadened animal tissue: disintegration into bions, followed by a spontaneous reorganization of bions into protozoan forms. In both cases, of soil or tissues, Reich argued that the process was initiated by a *loss of life energy charge* of the tissues, followed by putrefaction and disintegration.

One special bion preparation, made from pulverized beach sand heated to incandescence and immersed into a sterile nutrient broth, yielded a powerful radiant energy phenomena. Lab workers developed conjunctivitis if they observed the preparations too long, while a skin inflammation could be developed by placing the bion solution close to the skin for a period. Working for extended hours in the laboratory, Reich developed a dark tan through his clothing, in the middle of winter. The radiation imparted a magnetic charge to nearby iron or steel implements, and a static charge to nearby insulators, such as rubber gloves. Film stored in nearby metal laboratory cabinets spontaneously fogged. He noted that whatever this bion radiation was, it was rapidly attracted to metals, but just as rapidly reflected away, or dissipated into the surrounding air. Organic materials, however, absorbed this radiation and held onto it. Attempts to identify the new radiation using traditional nuclear or electromagnetic radiation detectors failed.

Reich also noted that the air in rooms containing the special bion cultures would feel "heavy" or charged. When observed at night, in full darkness, the air would visibly scintillate and glow with a pulsing energy. He attempted to capture the energy radiating from his bion cultures inside a special cubical enclosure lined with sheet metal, which he felt would reflect and trap the radiations inside. As expected, the special metal-lined enclosure trapped and amplified the effects of the bion radiation. However, to his amazement, he found that the radiation was also present in the experimental enclosure *even when the bion cultures were removed*. In fact, there was nothing which could be done to make the observed radiation "go away". The special metal-lined enclosure appeared to pull the same form of radiation from the air which previously had been observed coming from the bion cultures.

Reich eventually became convinced that the special enclosures were capturing a free atmospheric form of the same energy that he also observed coming from living organisms. He called the newly discovered energy the *orgone*, and he developed ways to amplify the energy accumulating affects of the enclosure, mainly through multiple layering of the metallic and organic materials. No electricity, magnetism, electromagnetism, or nuclear radiations were employed in these accumulating structures, which were entirely passive in design. The special enclosures were thereafter called *orgone energy accumulators*.

The full sweep of Dr. Reich's clinical findings, his experiments with bioelectricality, the bions, on biogenesis and the origins of the cancer cell, and his discovery of the orgone energy and the orgone energy accumulator, cannot be given here, but a few points are summarized. The orgone accumulator was found to have specific life-positive effects upon plants and animals exposed to the concentrated life force inside it. A host of quantifiable effects upon the physical properties of the air, or other materials charged up inside the accumulators, were also discovered and documented. Reich and his co-workers published a host of research articles on the orgone energy accumulator, its unusual physical properties, and its life-positive biomedical effects. These effects have been repeatedly confirmed, and a research tradition in orgone biophysics continues to this day. Briefly, we may identify some of the known properties of the orgone energy, and effects of the orgone energy accumulator.

Properties of the Orgone Energy:

A) Ubiquitous, fills all space.
B) Mass-free; cosmic, primordial in nature.
C) Penetrates all matter, but at different rates of speed.
D) Spontaneously pulsates, expands and contracts, and flows with a spinning wave characteristic.
E) Directly observable and measurable.
F) Negatively entropic.
G) Strong mutual affinity and attraction to/by water.
H) Accumulated naturally by living organisms through foods, water, breath, and through the skin.
I) Mutual excitation and attraction of separate orgone energy streams, or of separate systems charged with orgone.
J) Excitability via secondary energies (nuclear, electromagnetism, electrical sparking, friction).

Physical effects of a strong orgone charge:

K) Slightly higher air temperature as compared to surroundings.
L) Higher electrostatic potential, with a slower electroscopical discharge rate as compared to surroundings.
M) Higher humidity and lower water evaporation rates as compared to surroundings.
N) Squelching of ionization effects inside gas-filled ionization Geiger-Muller tubes.
O) Development of ionization effects inside non-ionizable vacuum tubes (0.5 micron pressure or lower).
P) Ability to impede and absorb electromagnetism.

Biological effects of a strong orgone charge:

Q) General vagotonic, expansive effect on entire system.
R) Sensations of tingling and warmth at skin surface.
S) Increased core and skin temperature, flushing.
T) Moderation of blood pressure and pulse rate.
U) Increased peristalsis, deeper respiration.
V) Increased germination, budding, flowering and fruiting of plants.
W) Increased rates of tissue growth, repair, and wound healing, as determined through animal studies and human clinical trials.
X) Increased field strength, charge, integrity of tissues and immunity.
Y) Greater energy level, activity, and liveliness.

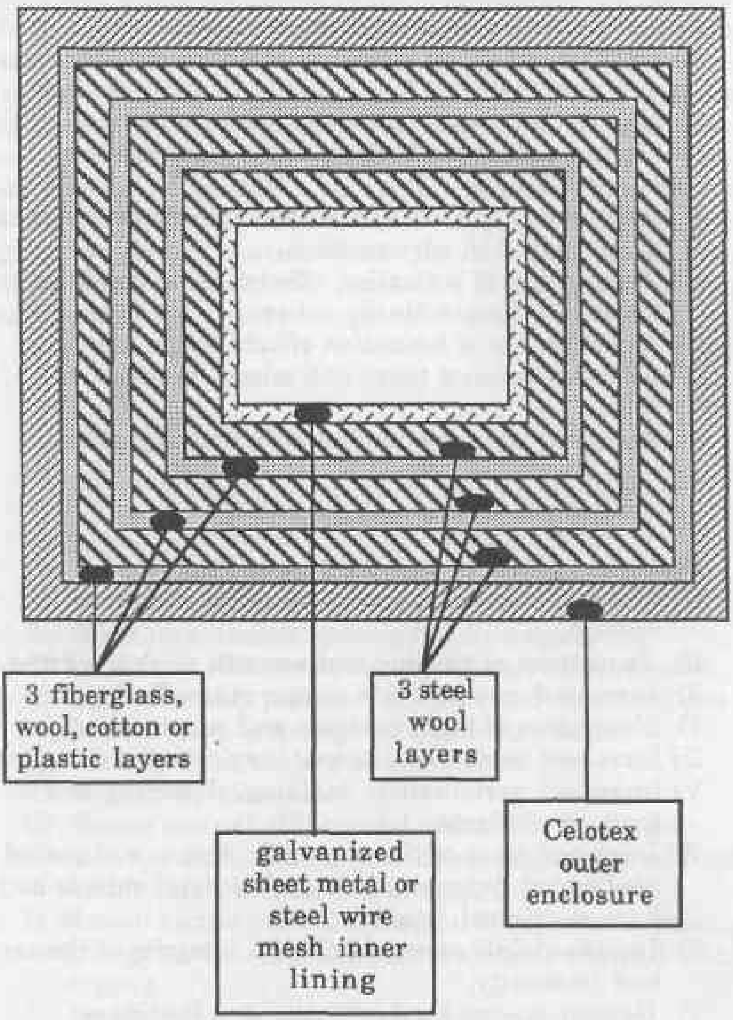
3 fiberglass, wool, cotton or plastic layers
3 steel wool layers
galvanized sheet metal or steel wire mesh inner lining
Celotex outer enclosure

Simplified Diagram of the Orgone Energy Accumulator

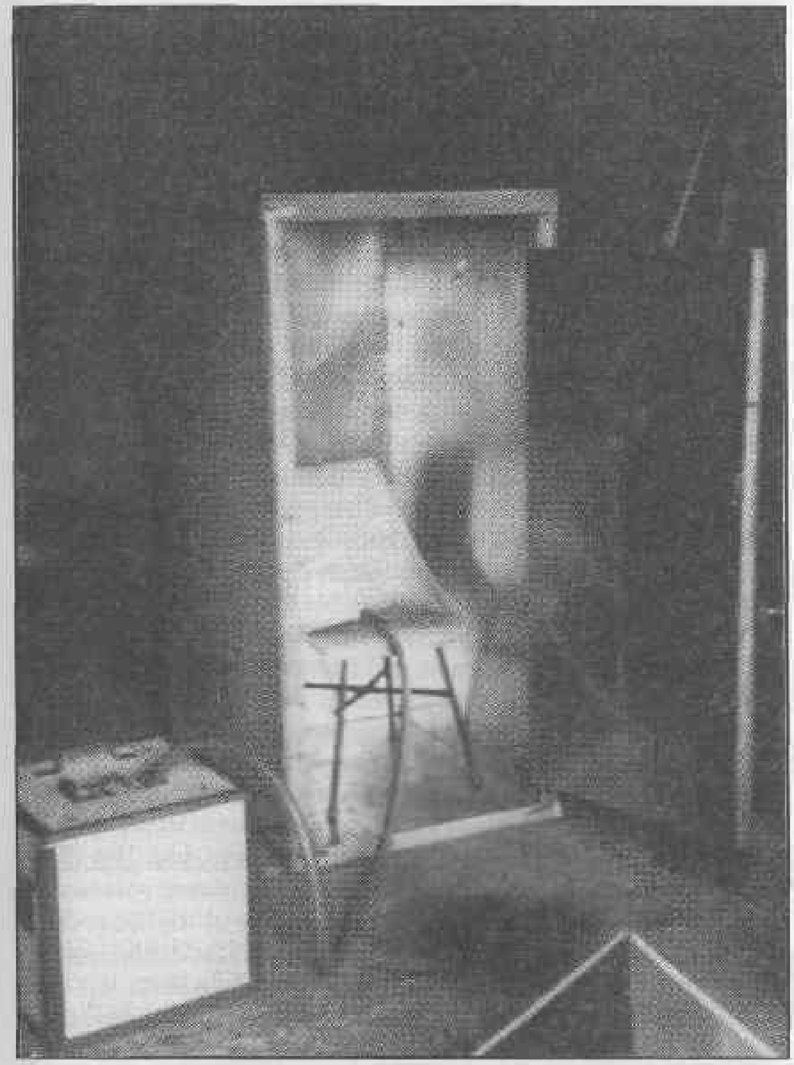

A three-ply accumulator in the author's laboratory. A ten-ply charger, with attached funnel shooter, is at the lower left. This charger normally sits inside the larger accumulator, under a wooden bench inside the accumulator.

5. Objective Demonstration of the Orgone Energy

A number of techniques have been developed over the years to document, measure, and objectify the orgone energy. These techniques are briefly listed, but the interested reader is directed to the references for more specific details.

A) Bioelectric Fields: Reich identified various bioelectrical phenomenon which he felt demonstrated a more powerful energy current at work in the body. The small millivolt currents of "bioelectricality", he argued, were only a small portion of this stronger energy current in the body, which he identified as being both emotional and sexual in nature, and which was later objectively identified as the orgone energy.

B) Radiant Effects from Bion Cultures: Special bion cultures derived from beach sand emitted a powerful radiation which could be felt and seen in darkrooms. This radiation did not register on instruments for detecting nuclear or electromagnetic energies. Additionally, the radiation could fog film, impart a static charge to insulators, and a magnetic charge to steel laboratory implements.

C) Darkroom and Atmospheric Observations, the Orgonoscope: Reich also observed and categorized various observable phenomena which could be seen by the dark-adjusted eye in the open air of darkrooms. Scintillating foglike forms, and dancing, luminescent pin-points of light were observed, and numerous techniques were developed which demonstrated their real, objective nature. One of these techniques involved the development of a new instrument, the *orgonoscope*, which used hollow tubes, lenses, and a fluorescent screen for magnifying the various subjective light phenomena. Large room-sized orgone accumulators were also constructed, and observations made in these amplified and clarified many of the effects. A special corpuscular *orgone unit* was identified, whose lawful behavior changed according to cosmic and meteorological factors. These macroscopic particles were also observed in the daytime sky with the unaided eye, as a generally common phenomenon, visible to most people once they were pointed out. The Earth was observed to possess its own orgone energy envelope, or energy field, much as did individual living creatures.

D) X-Ray Photographs: Reich observed that the x-ray "ghost" phenomenon (spontaneous, unexplained fogging of x-ray films) could be explained as an effect of the orgone radiation, or life energy. He published several photos where the ghosts were purposefully created by excitation of orgone energy within the field of the x-ray machine.

E) Visible Light Photographs: Reich observed that his special radiating bion cultures would fog film stored in nearby metal cabinets. Culture dishes of radiating bions set directly upon the film would also render an image of the culture dish and its contents. More recently, Thelma Moss of UCLA has shown that life energy field photos can be made without electrical stimulation (as with Kirlian techniques), by energy field enhancement; living objects placed directly on a film for a few days within a darkened orgone accumulator will, under the proper conditions, render an image.

F) The Orgone Energy Field Meter: Reich developed this device to measure the strength of energy fields. Using a Tesla coil and special accumulator-like metal plates, the device could quantify the differences in energy level between people or objects.

G) The Orgone Energy Pulsation Demonstrator: Reich demonstrated that the energy field pulsations of a large metal sphere were capable of setting into motion a smaller metallic/organic pendulum suspended nearby.

H) The Accumulator Temperature Differential (To-T): An accumulator will spontaneously develop a slightly higher temperature than either its surroundings or a control enclosure, on those sunny and clear days when the orgone charge at the Earth's surface is strong. The effect vanishes during stormy, rainy weather, when the orgone charge at the

Earth's surface is weak (but strong in the Atmosphere). The results of this temperature experiment, which has been replicated many times, demonstrates that the orgone energy functions in opposition to the second law of thermodynamics.

I) The Accumulator Electrostatic Effects: An electroscope kept inside an orgone accumulator will dissipate its charge more slowly than an identical one kept in the open air, or inside of a control enclosure. A partly charged, or uncharged static electroscope which is kept inside an accumulator will sometimes spontaneously charge itself up. As with the temperature differential effect, the electrostatic effects vanished during rainy or overcast weather, when the orgone charge at the Earth's surface is weak.

J) The Accumulator Ionization Suppression/Amplification Effect: Geiger-Muller tubes and counters charged up inside a very strong accumulator for several weeks or months tend to go "dead" for a period, but may eventually yield an erratic count rate for background. Special vacuum tubes, which Reich had constructed and called vacor tubes (which mimic the design of the Geiger-Muller tube but are evacuated well below the level at which ionization could occur), will initially not yield any counts when hooked to a radiation detector. After charging up inside a very strong accumulator for weeks or months, however, these same vacor tubes will begin to yield very high counts per minute for background, even at very low exciting voltages. The results of this experiment run counter to the classical interpretation of the ionization effect inside the Geiger-Muller tube, and hence, to the classical particulate interpretation of radioactive decay.

K) The Accumulator Humidity/Water Evaporation Effect (EVo-EV): More recent studies have suggested that an accumulator tends to attract a slightly higher humidity into itself, and suppress the evaporation of water from an open vessel inside. As with other accumulator phenomena, this effect also diminishes or vanishes during rainy weather.

L) Atmospheric Energetic Pulsation, and the Reversed Orgonotic Potential: Based upon observations of the thermal, electroscopical, and ionization characteristics of the orgone accumulator, Reich identified a patterned and lawful energetic cycle at work within the atmosphere and energy field of the Earth. These observations likewise led to the identification of a reversed potential at work in the orgone energy, running counter to thermodynamic principles, which explained why natural orgonotic systems (organisms, weather systems, planets) maintained a higher concentration of energy than their surrounding environments. The stronger of two orgonotic systems will drain energy from the weaker system, and increase its own potential or charge, until the weaker system is drained, or some maximum capacity level is achieved. Discharge may occur thereafter. Under sunny, clear weather, the orgone charge at the Earth's surface is quite strong, and in a state of expansion, preventing any significant growth of clouds. Where overcast or stormy conditions prevail, the orgone charge at the Earth's surface is weak, while the charge in the atmosphere is strong and in a state of general contraction. This loss of charge at the Earth's surface during rainy weather slows down the activity of living creatures, and the accumulator will not function well at those times.

M) The Millivoltmeter: Virtually all objects and organisms within a given environment, to include the air, water, and the Earth itself, have a charge that will increase and decrease in a cyclical or pulsatory manner, timed to cosmic and meteorological factors. In living creatures, high potentials produce more active physical and emotional periods, while low potentials signal less active periods. In Nature, high atmospheric potentials signal cloudy periods with stronger storms, while high Earth potentials signal cloud-free conditions. The tiny currents are excellent predictors of powerful biological or environmental processes, but are themselves too slight and weak to be the causative agent. Reich, and other researchers who have examined these small voltage potentials (such as H.S. Burr), viewed them as indicative of a more powerful, ubiquitous natural phenomena which energetically linked together the Sun, Moon, Earth, weather systems, and all living creatures.

N) Plant Growth Enhancement Studies: Seeds and plants correctly charged up inside an accumulator demonstrate higher growth rates and yields of fruit. This is one of the more telling, and widely replicated experiments with the orgone accumulator. In my own trials, I have seen a six-fold increase in the length of mung bean sprouts inside a strong accumulator, as compared to a control group of sprouts. Germination rates, growth rates, budding, flowering and fruiting can be increased by charging seeds or growing plants directly inside the accumulator. The seeds can be sprouted directly inside the accumulator, or charged up for a few hours, days, or weeks prior to planting. Growth enhancement effects may also occur when the water alone is charged, and then externally given to the plants.

O) Non-Human Animal Studies: Controlled studies on the effects of orgone radiation from an accumulator on cancer mice and wounded mice have been performed. These studies generally confirm Reich's earlier arguments that tissues with a stronger energetic charge will heal more quickly, and develop tumors more slowly, or not at all, as compared to energetically weakened tissues. These findings invalidate many aspects of the DNA theory of cellular differentiation, which appears to be more directly under the structuring influence of the organism's own life energy field.

P) Human Studies: Aside from those clinical trials performed by Reich and his co-workers in the 1940s and 1950s, very little work has been done in the USA regarding the bioeffects of the accumulator on humans. All research into these questions was halted by medical police actions in the 1950s. Recent studies from Germany, however, have confirmed such bioeffects. In general, a person sitting inside an accumulator will feel a variety of warm, glowing, or sometimes tingling sensations at the skin surface; their body core temperature will rise and skin flush, while blood pressure and pulse rate will trend towards moderate levels, being neither too high or low. When properly used, it has a distinct vagotonic, enlivening effect. The chapter on

"Physiological and Biomedical Effects", as well as the Appendix article, will provide details on these questions.

6. Discovery of an Unusual Energy by Other Scientists

Studies by various natural scientists have, over the years, demonstrated energetic principles at work in the natural world which are similar to the orgone energy. Early Chinese medicine acknowledged the existence of such a force, called *Chi*, and the traditional method of acupuncture is based upon the existence of such an energy principle within the human body. Acupuncture points do not correspond directly with nerve endings, and the most able acupuncturists do not rely upon the Western models of physiology to explain its effects. Given the absence of a vital energy principle, Western medicine cannot explain acupuncture, and has resisted its adoption in the United States. Acupuncture furthermore works on animals, invalidating invocation of the placebo effect; ancient texts from India have also referred to the life energy, called *Prana*, and provide maps of *Nila points* (similar to acupuncture points) on elephants. The texts from ancient China and India speak about an energy that is taken in through the breath, and flows through the body along the various meridians. Health is constituted by a free, unimpeded flow of this energy, while sickness occurs when the flow of vital energy is blocked. This is very similar to Reich's ideas on the orgone energy, though the Asian sources say little about the free expression of emotion; they also often advocated a conscious control of emotions and sexual feeling (*orgasm avoidance)*. In contrast, Reich demonstrated that such chronic restraint or self-control was the reason why the life energy became blocked or dammed up in the first place.

In the Western tradition, the vitalists of the eighteenth and nineteenth centuries also discussed the existence of a biological energy or life force, which was called *animal magnetism*, the *odic force, psychic force, élan vital,* and so on.

Indeed, Mesmer spoke of animal magnetism as an atmospheric fluid which surrounded, charged, and animated living creatures, and could be projected across a distance by a therapist. Mesmer was a teacher of Charcot, who was, in turn, the teacher of Freud, who was one of Reich's early mentors. Reich also studied with other vitalists, such as Kammerer and Bergson, and the vitalist tradition has persisted as a quietly spoken minority viewpoint in biology. Besides Reich, the more recent advocates of a vital or dynamic energy principle in nature included the late Harold S. Burr of Yale University. Burr argued for the existence of a powerful *electrodynamic field* at work in nature, affecting both weather and living creatures. The biologist Rupert Sheldrake has similarly developed a theory on *morphogenetic fields* which also takes from this tradition. Like Burr's work, Sheldrake's theory provides a dynamic, energetic explanation for inheritance, making the biochemical DNA theory unnecessary. Most recently, editors for the academic publication *New Scientist* called Sheldrake's book the "best candidate for burning" they had seen in a while.

The surgeon Robert O. Becker developed these prior principles to a most amazing state of advancement. His early research led to the development of a class of devices for the electrical stimulation of bone healing and pain relief. His later work took these principles, and developed them to the point that he could artificially stimulate the *regenerative growth of amputated limbs of laboratory mice*, in a manner similar to the way in which a salamander or spider will regrow a lost limb. This kind of regrowth is by nature limited only to less complex creatures, and does not generally exist among mammals, such as mice, rabbits, and humans. Regrowth of an amputated limb had never been previously demonstrated in a mouse, or any mammal for that matter. Becker's work was a severe blow for both the biochemical DNA theory of cellular regulation, and the theory that the bioelectrical field of a creature was just a meaningless "byproduct" of chemical metabolism, like the electric field surrounding a running automobile engine. His work proved that the energy field of the animal was a primary determinant of growth and repair, as was the case with Reich's work. Becker was preparing to replicate the limb regrowth experiments on humans, when the biomedical community reacted with severe outrage against him, pulling dirty tricks of all sorts to have his research funding cancelled, and his laboratory shut down.

Another vitalist of our era is Bjorn Nordenstrom, director of the Karolinska Radiological Institute in Sweden. Nordenstrom, like Reich, made a study of the x-ray "ghost" phenomena, which is an unusual spontaneous fogging of x-ray films. It appears as a wispy, smoke-like or blob-like form on the x-ray images of patients, and sometimes can be seen on the monitors of airport baggage x-ray equipment. It cannot be predicted, and most radiologists consider it to be a nuisance. However, Nordenstrom studied it, and observed distinct patterns correlated to his patient's bioelectrical fields. Like Reich, he also discovered and measured currents of bioelectricality in the body. His meticulous research was summarized in a book titled *Biologically Closed Electric Circuits: Clinical, Experimental, and Theoretical Evidence for an Additional Circulatory System*. After being heavily advertised in medical journals in the USA, it sold less than 200 copies, evidencing a contempt among mainstream medical doctors for any new findings that would support the principle of a life energy, even one of a purely bioelectrical nature. Unable to find support for his work in the West, Nordenstrom recently went to China to pursue his clinical research.

Other biological scientists have inferred the existence of such a vital energy principle, on the basis of their experimental work. When they provide good confirming evidence, they are hotly attacked. The French scientist Louis Kervran, for example, spent years developing very elegant and simple experiments demonstrating that the basic elements of chemistry were being *transmuted* by living creatures. Chickens fed a diet free of calcium, for example, would not lay mushy or fragile eggs, unless dietary silica was restricted. With a restricted silica intake, however, they laid mushy and fragile eggs, and it did not matter how much calcium they ate. Likewise, laboratory mice would heal broken bones very quickly when fed a diet high in organic silica, but not so fast when silica was minimized and only calcium was provided. These experiments strongly suggested that dietary silica was being transmuted into calcium in the bodies of animals. Kervran also experimentally demonstrated other likely transmutations, and other scientists in Europe and Japan confirmed his findings. He eventually came to the conclusion that there had to be some unknown form of powerful biological energy at work to drive the transmutations. But when he wrote to a prominent American scientist for assistance in obtaining equipment for an important experiment, he was impolitely told to go "read an introductory textbook on biology". In the United States, Kervran is better known among homeopathic doctors and organic farmers than by university professors. However, if Kervran is right–and the experimental evidence suggests that he is–then the textbooks of biochemistry will need to be rewritten. As Kervran pointed out, biology and biochemistry are two entirely different disciplines and should not be confused. Biology is concerned with observable fact, while biochemistry attempts to explain observed facts by a chemical theory which assumes elemental constancy. And it is with this basic assumption that part of the error lies.

Another French scientist, Jacques Benveniste, actually demonstrated such an energy principle at work in homeopathic dilutions. His experimental work was successfully replicated by independent laboratories in other countries, to satisfy his obstinate critics. But that was not good enough. For making this offending discovery, which lent some support to homeopathic physicians (who are often prosecuted and jailed in the United States), the science journal *Nature* dispatched a "hit squad" of fraud investigators, debunking magicians, and skeptical editors to his laboratory, under the guise of "evaluating" his laboratory procedures. The *Nature* science cops made a mess of Benveniste's lab, distracting laboratory workers, performing slight-of-hand tricks, and shouting, before finally being told to leave. *Nature* subsequently tried to smear Benveniste in their editorials, but did not factually refute his work through replication of experiments. Such is the warp and weave of traditional academic science in the 1980s.

In the atmospheric sciences, the tradition of dynamic energy forces which affect entire regions was preserved for a period by older weather forecasters, who used streamline, rather than frontal theory to predict weather. Streamline analysis more coherently focused upon streaming movements of air, or jet streams as they are called today. For example, when you look at the dynamic images of clouds, as seen from a satellite in space, you do not see "fronts". But you do see *streaming movements of clouds*. Reich independently discovered the basic configurations of these streams, years before the first weather satellites were launched. Likewise, the older atmospheric scientists often argued for a great interconnectivity in the atmosphere. Charles G. Abbot, head of the Smithsonian Astrophysical Observatory in the 1950s, used related energetic concepts to predict the weather months into the future. But he was ignored and ridiculed for his findings, in spite of their uncanny accuracy. Irving Langmuir, one of the originators of cloudseeding techniques, once objectively demonstrated that cloudseeding in New Mexico would trigger rain storms all the way into Ohio, and he warned his fellow workers about this danger. The cloudseeders of today, funded by millions in Federal dollars, act as if Langmuir's work never took place, and refuse to replicated his simple experiment. They deny the existence of long distance effects from cloudseeding, knowing that if such effects became public knowledge, they would be forced to stop.

Among the physical scientists, the idea of an energy in space was embodied in the concept of an *aether,* which dates back hundreds of years. The theologian/physicist Isaac Newton forcefully argued that this aether *had to be static*, in order to prevent it from directly participating in the movement and ordering of the heavens. That role, Newton argued, belonged only to the anthropomorphic God (who at that time was demanding that unbelievers be ruthlessly tortured and burned at the stake). And, over the years, a dead, unmoving aether has never been detected. However, an *aether with more* *dynamic properties* was objectively demonstrated by the physicist Dayton Miller. Miller also explained why prior attempts to measure the aether had failed. First, he observed that the aether is *entrained* at the Earth's surface, and moves faster at higher altitudes than lower altitudes. Prior attempts to measure its movement had taken place only at lower altitudes, or in heavy stone buildings or basement locations. Second, Miller's aether was *reflected by metals*, and prior attempts to measure it used instruments with the critical parts housed inside metal enclosures. Miller found that by doing the crucial aether-drift experiments on a mountain top, inside a flimsy building without metals or dense window materials, that it was readily detectable and measurable. He made over 200,000 separate measurements, over the course of 30 years of investigation. Contrast this to the famous Michelson-Morley experiment, which involved a grand total of six hours of actual measurement time, made over four days in 1887. The Michelson-Morley experiment is widely misquoted as having completely failed in the detection of the aether. It was a hinge-point in the sciences, after which the idea of the aether was given up entirely for the "empty space" theories of relativity and quantum dynamics.

Miller's extensive work on the aether question was never rebutted when he was alive, but his research was contemptuously compared with "searching for perpetual motion". After his death, the adherents of the empty space theory breathed a heavy sigh of relief. Today, every physics textbook starts out with the falsehood that "the aether was never measured or demonstrated". It should be pointed out that the theories of relativity and quantum dynamics, plus the expanding universe and "big bang" theories, are utterly shattered by the discovery of an energy in space, and many physicists, who cling to their theories religiously, simply refuse to look at this kind of evidence. Worse, the discipline of physics has become a military-oriented growth industry, with multi-billion dollar funding for the sustenance of nuclear power and bomb-making technology, for fusion plants, particle accelerators, and "star wars" experiments. This kind of research has not brought forth any real benefits or fruits for humankind, but has become a self-perpetuating "research industry", a multi-billion dollar edifice which, like the medico-pharmaceutical industry, is threatened to the core by these discoveries of a primary, cosmic life energy. The physics community has unfortunately reacted to these new findings with the same arrogance and viciousness that characterizes the medical communities' reaction to the life energy. Einstein's followers, for example, have recently been accused, in print, of very nasty stab-in-the-back tactics of censorship and suppression. An entire new journal, *Scientific Ethics*, for a short period at least began to expose the whole stinking mess.

Of great interest for Reich's work is that Miller's dynamic aether was *more active at higher altitudes*, and *reflected by metals*. The capacity to be reflected by metals, with a more active state at higher altitudes, are basic properties of the orgone energy, as independently discovered by Reich. The orgone also satisfies many other of the basic properties and functions of an aether, being ubiquitous and mass-free, and by providing a medium for the transmission of electromagnetic excitation. However, the orgone also spontaneously pulses, superimposes, and directly participates in the creation of both matter and life. But even without using the taboo word "aether", or the more offending word "orgone", another group of physicists have detected or inferred the existence of dynamic energy currents at work in deep space. For example, the American astrophysicist Halton Arp made so many photographs of energy/matter bridges between deep-space objects, where those energy/matter bridges should not have been there, that he was actually banned from using the big American telescopes. His simple photos demolished the theories of empty space, the expanding universe, and the "big bang" with a single shutter click. So great was the hatred against his work that he ultimately had to go to Germany to continue with his research. Hannes Alfven, another famous physicist, also deeply offended his contemporaries by suggesting, like Reich, that space was filled with streaming currents of plasmatic energy. The space scientists to this day refuse to send satellite probes where he says they should, as to do so might confirm that space is energetically rich. In fact, physics of today is in a state of turmoil, and is desperately trying to explain away the newer evidence for an energy in space, to preserve the big bang theory, relativity, quantum dynamics, and the billions of dollars of research money that supports the "empty space" religion, and its institutional priesthood.

Few of the above ideas, nor the findings on sunspot-weather correlations, are given much funding or investigation today. Science journals still routinely carry the false statement that "no mechanism" has been found for solar-terrestrial correlations, just as physics textbooks carry the falsehood that "the aether had never been detected". And it is true that these relationships cannot be true, nor do they make any sense, from the standpoint of the "empty space" theories of physics. They require a medium in the atmosphere and in space, through which excitations and influences can pass, independent from thermal or pressure phenomena, a force which propagates in the atmosphere faster than air currents, and which can likewise quickly propagate influences across the depths of space. Again, Reich's orgone energy fits such a description.

Other research has been done to show that living creatures, and the physical chemistry of water, are sensitive to weather or cosmic factors in a manner than cannot be explained according to simple mechanical phenomena, such as light, temperature, humidity, or pressure. Frank Brown, of Northwestern University, spent decades demonstrating that the biological clocks of various living creatures were sensitive to lunar cycles and other cosmic forces. Nobody could refute him when he was alive, but today, after his death, his findings are widely ignored. Likewise the works of the Italian chemist, Giorgio Piccardi, who demonstrated that the physical chemistry of water was changed by magnetism, sunspots, and other cosmic phenomena. His work helped to fuel an interest in the magnetic treatment of water in Europe, leading to new methods for reducing scale deposits in household plumbing, and in industrial boilers. Magnetism, correctly applied, can alter the solubility characteristics of water, allowing dissolved substances to remain in solution at concentrations higher than normal for a given temperature. In the USA, these findings have been greeted with derision, as every physics textbook says magnetism has no effect upon water. Also, almost every chemical laboratory uses magnetic stirring devices to mix their chemical solutions, instead of the "old-fashioned" hand-operated glass stirring rods; these magnetic stirring devices would, if Piccardi is correct (and he is) alter the chemistry, precipitate quantity, and titration curves for every chemical reaction exposed to them. And so, the new findings are ignored in the USA, while abroad, new products based upon the discovery are entering the marketplace. Simple magnetic water treatment systems for the home are now common in Europe, replacing in many cases the ion-exchange water softeners, with their bags and bags of salt. In the USA, meanwhile, the water softener industry, in collusion with dogmatic academics and politicians, has managed to have laws passed in a few states to forbid the sale of magnetic water treatment devices.

Piccardi's work extends beyond the issue of simple magnetic treatment of water, however. At one point he attempted to isolate an unknown cosmic energy which was affecting his chemical experiments, in a manner similar to strong magnetism. In order to block out the unknown radiation, which was correlated to sunspots, he constructed an electromagnetic shield around his experiments, in the form of an Earth-grounded metal box enclosure. Then, in order to stabilize the temperature inside the metal box, he placed a layer of wool around the outside. To his amazement, the metal box did not extinguish the cosmic phenomena, but amplified it. He and his co-workers spent decades performing chemical experiments inside similar enclosures, which mirror the construction of Reich's orgone energy accumulator. This independent corroboration of the orgone accumulator principle by Piccardi was also confirmed, though in a less direct manner, by the biologist Brown. Brown observed that hermetically sealed *metal enclosures*, with a constant pressure, temperature, light and humidity inside, would not extinguish cosmic influences upon biological clocks, but would instead allow them to be more clearly observed, or even add an unusual dimension to their behavior. For example, inside the metal box, the metabolism of potatoes followed a cycle that correlated with lunar, solar, and galactic parameters. Potato metabolism additionally demonstrated a correlation to local weather; *not the weather today, but the weather two days into the future!* In the enclosure, the energized potato would respond to external energetic factors in the environment which were also determinants of future weather events.

The above are just a few of the kinds of evidence that exist for an energetic principle similar, or identical to the orgone energy. In many cases, these researchers had not known of Reich's work. In a few, they hated Reich's guts, and would hardly tolerate mention of his name by their students! And yet, the facts speak powerfully for a corroboration of Reich's orgone energy. It must be stated, however, that Reich's discovery on the orgone energy is far more inclusive, comprehensive, and tangible than any of the above concepts. In addition to having been quantified, photographed, and measured, the orgone can be seen, felt, and, as noted in this book, accumulated within special experimental enclosures.

An additional word must also be given regarding the response of the scientific and academic communities to these new discoveries. The reader will note that most, if not all, of the above researchers were hotly attacked, or isolated and ignored for their findings, irrespective of their credentials, reputations, or the amount of evidence they provided. This emotional reaction, of running away from or attacking disturbing new ideas, was explained by Reich as being the results of a specific emotional disorder, which he called the *emotional plague*. Virtually all scientists who have verified portions of this common natural energy principle have been assaulted by special *emotional plague characters*, who build their reputations not upon work or research, but upon political power, and the number of scalps they have taken. Gossip, slander, political tactics, the sneak attack, and even manipulation of the courts and police are standard tactics of the plague. Their secret goal, like the Grand Inquisitors of the Church, is to kill disturbing new findings, and the men and women who make them. The history of science is filled with evidence for this kind of behavior. The reader is encouraged to read Reich's discussion on the emotional plague, in *Character Analysis (3rd Ed.), People in Trouble*, and *The Murder of Christ*, as it still constitutes the major obstacle in the way of human social progress, and scientific research.

Part II: The Safe and Effective Use of Orgone Accumulating Devices

7. General Principles for Construction and Experimental Use of the Orgone Energy Accumulator

A) The interior surface of all accumulators must be composed of bare metal. Paints, varnishes or coatings on the metal will interfere with the accumulating effect, though zinc galvanizing does not.

B) The outer, exterior surface of all accumulators must be composed of an orgone-absorbing, generally organic, non-metallic substance.

C) Metals and non-metallic materials may be alternated in multiple layers within the walls of the accumulator for stronger energy accumulation. The more layers, the more powerful the accumulator, though one does not simply double the strength by doubling the layers. A three ply accumulator will have about 70% of the strength of a ten ply accumulator (one "ply" consists of a layer of metal plus a layer of non-metal). Accumulators of different sizes may also be nested one inside another, to develop an even stronger charge. Points A and B above must be strictly followed, however. In multiple ply accumulators, you can double-up the final outer organic, non-metal layer, and the innermost metal layer, for additional energy accumulating capacity.

D) A major common error made by some who reproduce Reich's orgone accumulator experiments is the use of inappropriate accumulator materials. For accumulators used on living systems, and particularly for human use, copper, aluminum, and other non-ferrous materials must be completely avoided as they yield *toxic effects*. Similarly, certain types of polyurethane foams, rigid or soft, do not have a good effect upon the living system when used in an accumulator. Any type of material impregnated with formaldehyde, or made with other highly toxic glues or resins should not be used.

Good non-metals
wool, raw cotton
acrylic, styrene plastic
celotex (soundboard)
cork sheeting
glasswool, fibreglass
rock wool
bees wax, candle wax
soil, water

Poor or toxic non-metals
wood or plywood
urethane or polyurethane
pressboard (very hard)
organic materials containing formaldehyde, asbestos, or other toxic chemicals

Good metals
steel or iron, sheet, screen
galvanized steel
steel wool
stainless steel
steel/tin can alloy

Poor or toxic metals
aluminum, sheet or screen
lead
copper

E) Some individuals have experimented with accumulators composed of buried metal boxes, surrounded with rich dark soil, free of pesticides and herbicides. The larger of these kinds of accumulators give the appearance of a root cellar or "burial mound". Some authors familiar with ancient archaeological sites have even speculated that the life energy principles were known and used by ancient peoples. Certain ancient mounds and structures have a layered characteristic, using clay soils or stone of high iron content, covered over with other layers of organic-rich soils or peat.

F) An exceptionally powerful accumulator can be made by using bees wax or other dielectrical materials for the outer, non-metallic layers. These materials may be quite expensive for a larger accumulator, and are also fragile. If you use a fragile or crumbly material for the outer nonmetal layer, you can coat the outer surface with clear shellac. This has been tried by many people, and does not appear to interfere with the accumulation or life enhancing qualities of the energy. Never use shellac on the interior surfaces, however.

G) Experiments have demonstrated that the shape of the accumulator is a factor of lesser importance than its material composition. However, accumulators made in the shapes of cones, pyramids, or tetrahedrons have yielded occasional unexplainable life-negative effects. Unless one is testing for such effects, accumulators should be constructed in rectangular, cubical, or cylindrical shapes. These have given the best results, and are also easier to construct. One anecdote here: in 1980, the author was in Egypt, and went into the Great Pyramid of Cheops. While inside, I was struck down by a most intense choking, and could not get a breath. The feeling was relieved by emptying my canteen of water over my head and chest. Later, I heard reports of whole groups of tourists being similarly stricken, to the point that some people had fainted away, and had to be resuscitated outside. I cannot say if this is an effect of poor ventilation or not, but in my own case, I was the only one out of 8 people in a tour group to be badly affected. Given my observations of stunted and killed seedlings within conical and pyramidal accumulators, it seems possible to me that these effects are the result of a toxic accumulation or overcharge effect. More work needs to be done to clarify these factors related to shape, as well as accumulator use in energetically stagnated environments, such as deserts. See the chapter on "Warnings and Dangers" for more details.

H) Corners of accumulators do not have to be precisely constructed, nor do layers have to be air tight or precisely fitted, though one certainly wishes to have as neat and clean of construction as possible. In some cases, I have seen metal boxes loosely wrapped with layered steel wool and cotton, felt, or wool. Also, some have used tin cans, as used for food preservation, wrapped with plastic, and then placed inside another larger can which was, in turn, wrapped with more plastic. These tin-cans were nested inside each other to make reasonably effective four or five ply accumulators, for seed charging or other purposes. They do not look especially neat or "scientific", but they do function.

I) Accumulators should be kept where fresh air can circulate. The door or lid to the accumulator should also be kept partly open when not in use. Its interior may be kept fresh and sparkling by keeping an open basin of water sitting inside when not in use. Periodically wipe the interior and exterior clean with a damp cloth.

J) Larger accumulators used by humans or farm animals are best kept outdoors under a sheltered area, where rain will not fall on it. Good air circulation and sunlight will assist with the accumulation affect. The best location for accumulator research would be inside a large wooden barn in the countryside, away from all kinds of electrical transmission lines, electromagnetic devices, and nuclear facilities. This finding on the best environment for the life energy is in full agreement with more recent findings on *house ecology*, wherein a constructed habitat is critically reviewed for toxic effects upon its inhabitants. See the "Warnings and Dangers" chapter for more details.

K) The accumulator will not develop a strong charge during wet, rainy weather. On such days, the orgone charge at the Earth's surface is very low, most of it being taken up into the storm clouds overhead or at a distance. The strongest orgone charge is found in the accumulator on clear, sunny days, when the orgone charge at the Earth's surface is also quite strong.

L) Orgone accumulators used at higher altitudes tend to yield stronger charges than at lower altitudes; lower latitudes may yield stronger charges than higher latitudes; lower humidity atmospheres tend to yield stronger charges than higher humidity atmospheres. Periods with many sunspots and solar flares coincide with periods of stronger orgone charge, as compared to periods with few sunspots and flares. Alignments between the Earth, Sun and Moon, during full and new moon periods, appears to yield a stronger, more excited charge in the atmosphere, and within the accumulator.

M) If you run a controlled experiment with the accumulator, do not place any relevant instruments immediately adjacent to it. Remember that the accumulator has an energy field, and will partially influence nearby objects in a manner similar to those kept inside it. The electrical or electromagnetic fields of various instruments might also disturb or otherwise affect an accumulator, making this caution doubly important for the research scientist.

N) Do not use any household electrical appliances connected to a wall plug inside or near to the accumulator. Neither should portable computers, TV sets, or other radiating devices be used. These will disturb the energy inside. The interior metal walls also conduct electricity, and there may be a danger of electrical shock. For human-sized accumulators, use a battery powered reading lamp if light is desired, or place a strong lamp just outside the door opening. Many people use such a light to read a book while sitting inside. Radio *receivers* do not appear to have a negative effect if used in the room, but the effects of "walkman" type headsets inside the accumulator are unknown.

O) For experimental accumulators, realize that any organic or moisture-bearing materials which are placed inside will absorb the orgone charge. Do not unnecessarily store or bring items inside an accumulator.

P) For human sized accumulators, one wishes the exterior walls to be no more than 2 to 4 inches from the skin surface. When sitting inside, it is best to partially or completely disrobe, as heavy clothing will interfere with absorption of the orgone radiation. A wooden chair or bench may be used as dry lumber is a relatively poor absorber of the orgone. Metal chairs are also OK, but may be uncomfortably cold to sit on.

Q) NOTE: A too-frequent or too-long use of the accumulator may lead to symptoms of overcharge, such as pressure in the head, slight nausea, general ill-feelings or dizziness. In such a case, leave the accumulator immediately and rest in the fresh air for a moment. Such symptoms will go away in a few minutes. However, Reich warned persons with a history of overcharged biopathies to use the accumulator only with caution, and then only for shorter periods. These overcharged biopathies include: hypertension, decompensated heart diseases, brain tumors, arteriosclerosis, glaucoma, epilepsy, heavy obesity, apoplexia, skin inflammations, and

conjunctivitis.

R) The question of "how much is enough" is related to one's own energy level, and is primarily a subjective determination, different for each individual. No one ever tells you how much water to drink to quench a thirst. You simply drink until you have the feeling of "having enough". The same is true regarding the use of the accumulator. When you have the feeling of "enough", then get out. With most people, this will be sometime after they have reached the point where their own energy field is gently *luminating*, or glowing with a warm excitement at their skin surface, and after sweating has commenced. If you are unsure about these kinds of feelings, be patient, as with some people, it may require many sessions before they can really feel the energetic effects. A good rule of thumb is to limit the sitting period to no longer than about 30 to 45 minutes. It may be used more than once a day, however. One should not attempt to "nap" inside for prolonged periods. Additional information on these bioeffects is given in the chapter on "Physiological and Biomedical Effects".

S) The qualitative state of the orgone, as well as its absolute charge, are constantly varying at any given location on the Earth's surface. Weather cycles cause the accumulator to vary its charge, and toxic environmental conditions (oranur and dor, see below) may periodically or chronically contaminate the accumulator, making its use potentially unsafe. Experimental use of the accumulator therefore demands that one learn about weather cycles and other environmental factors.

8. Warnings and Dangers: The Effects of Oranur and Dor

A common problem with use of the accumulator is the difficulty in finding an energetically clean environment in which to use it. The orgone energy in the atmosphere is very sensitive to certain kinds of disturbance and agitation. Much in the manner of living protoplasm, orgone energy can be *excited* or *irritated*, and certain environmental influences can drive it towards a toxic condition. If the energetic atmosphere in your home or neighborhood has been made toxic in this manner, use of an accumulator is disadvised, or advised only with great caution, as it will be very difficult to accumulate anything except a toxic charge.

For example, **orgone accumulators, particularly those intended for biological experiments or human use, should never be used in rooms with the following orgone-irritating devices:**

fluorescent lights
television sets
computer or microcomputer
other cathode ray tube devices
microwave ovens, or eddy-current stoves
electric blankets (even if only plugged in, and off)
diathermy, x-ray machines
sparking electric motors
induction devices or coils
other electromagnetic devices
ionization-type radioactive smoke detectors
clocks, wristwatches, or other devices containing radioactive, glow-in-the-dark materials (phospholuminescent materials, which work on the principle of absorbed visible light, are OK)
other radioactive materials, or strong chemical fumes

Orgone accumulators should not be used even in the same buildings where the more powerful of the above kinds of devices (such as x-ray machines) are used, or were recently used. Experiments by Reich, and more recently by others in large German hospitals, have demonstrated that x-ray equipment will destroy the life-enhancing effects of the orgone radiation. Additionally, there is a *persistence effect*, wherein toxic energetic conditions remain for a time after the irritating devices are shut down and removed from a room or building. **Orgone accumulators should likewise not be used in the immediate neighborhoods or vicinities where the following facilities are located:**

airport radar systems
cellular telephone or microwave relay towers
very high-tension power lines
AM, FM, or TV broadcast towers
nuclear power plants, storage facilities,
  or nuclear waste dumps
military installations with nuclear bomb storage
past or present nuclear bomb testing areas

Reich and others associated with him made warnings about these devices in the 1940s and 1950s, but only today do we see epidemiological studies corroborating their life-negative effects. Part of this problem has hinged upon the difficulty that, by simply demonstrating a correlation between two events, you do not prove causality. One has to show or demonstrate just what the mechanism is, and objectively demonstrate each step between the two correlated events, before cause and effect are proven. This is in most cases a very wise policy, but it is very unevenly applied in the world of the sciences. Orthodox theorems are rarely subjected to valid critical review based upon their failures to meet this strict criteria (eg, "bad genes", "viruses", etc.), while unorthodox theories are denied funding or simply thrown out or repressed for whatever weaknesses they may have. Industrial polluters may also raise this issue to avoid taking responsibility for the environmental damage they have done.

Regarding energetic questions, according to the best calculations of the physical scientists, low-level radiation *ought not* to have a deleterious effect upon the living system. The energy present in the low-level radiation, *as detected with conventional radiation detection instruments*, is simply not sufficient to do significant damage. And yet, the damage occurs. I emphasize the concern about "conventional radiation detection instruments", because a major fallacy of physics is that if an instrument does not measure an environmental disturbance, then no disturbance has occurred. The error here lies in the false assumption that their energy detection instruments must be detecting 100% of any disturbance. This unprovable assumption is, of course, challenged by biological or epidemiological evidence which demonstrates that an effect does exist. There is furthermore a great distrust of the body in the modern sciences, in that average people who are made sick by our modern energy radiating devices often are not believed, or are viewed suspiciously.

It is precisely here that Reich's findings on the orgone energy provide a clarification, as the life energy (and disturbances within it) *cannot be detected with ordinary nuclear or electromagnetic sensing instruments*. One has to modify the instrument to pick it up, or use entirely different methods. The orgone is also an interconnecting energy continuum, which provides a connection between the offending facility or appliance (nuclear plant, microwave tower, fluorescent light, TV set) and the living creature that is affected. As the local orgone energy field of the Earth, or the energy field of a home is badly disturbed and agitated by these devices, so too does the orgone energy field of a person in that environment become disturbed.

Modern physics partly acknowledges these connections, in that all nuclear bombs, nuclear reactors, and related facilities are said to radiate unshieldable, undetectable, theoretical *neutrinos* in tremendous quantities. These neutrinos race out from the facilities, penetrating all forms of ra-be done only on a very crisp, clear, and sparkling day, when the orgone charge at the Earth's surface, and in the accumulator, is quite strong and lively. Otherwise, charging may have to take place for slightly longer periods. Also note that seeds can be overcharged; attempts at charging seeds for 30 days or more often result in little difference between controls and charged seedlings, or even stunted growth.

E) Orgone Charging of Potted Plants: This can be done by charging up seeds prior to planting, as discussed above, or by charging up soil or water before using. One can also make a "wrap around" accumulator, using a metal can with the ends removed, and plastic and steel wool layers wrapped around the outside. Be sure that the final outside layer of plastic is fairly thick, and do not use aluminum materials. Leave the steel wool in a fluffy condition; don't compress it.

F) Home Seed Sprouting Experiments: The life-positive effects of the accumulator may also be observed in the way it enhances seed sprouting. Build an accumulator to house your seed sprouting apparatus. Store one sprout container in a darkened area away from the orgone accumulator, and another inside the darkened orgone accumulator. Make sure the temperature, ventilation, and light exposure of the two groups is identical and, once again, keep both groups away from oranur-producing devices. Measure the quantity of seeds going into each container, and make sure the quantity of water in each is about the same. Observe and record any subsequent differences in growth and taste. The accumulator group should have a greater growth and yield.

G) Laboratory Seed Sprouting Experiments: Obtain two small shallow glass dishes with a flat bottom, or two shallow glass laboratory culture dishes, about 4" diameter with a 1" lip. Into each dish place around 20 or 30 dry mung beans, to form a single layer of beans on the bottom of the dish. Add a measured quantity of water to each dish, which only halfway covers up the beans. The tops of the beans should remain exposed to the air, while the bottoms are wet. Place one dish of beans into a small but strong orgone accumulator, and the other dish into a control wood or cardboard enclosure of similar dimensions, but with no metals. Cover both the accumulator and control enclosure with a layer of black plastic, to seal them against any intruding light. Place the enclosures in a well-ventilated area, of equal temperature, and out of the direct sunlight. The enclosures should be in nearly identical environments with respect to light and temperature, but not placed closer to each other than about a meter. Again, no oranur-producing devices should be nearby. Each day, open the enclosures up and add as much water as is needed to keep the bean dishes wet to approximately the same height as originally was necessary to cover the beans halfway. If one dish begins to grow more quickly it will require more water, and this should be provided on demand. After one of the dishes of sprouts has reached about 4" high, record your observations of germination rate, length or weight of sprouts, general appearance, and other characteristics. Contrast the two groups. The accumulator group should have a greater growth and germination rate.

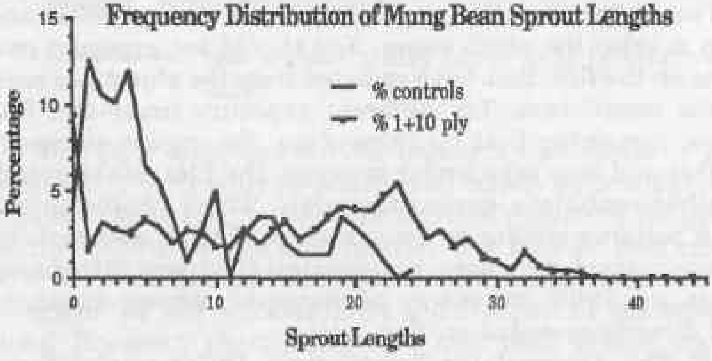

Contrast of mung bean sprout lengths between a group sprouted inside an orgone accumulator, and a group sprouted inside a control enclosure. (from DeMeo, J.: "Seed Sprouting Inside the Orgone Accumulator", *J. Orgonomy*, 12:253-258, 1978)

H) Passive Life Energy Photographs: Orgone energy field photographs can be made in a manner similar to that used to make *Kirlian* electrophotos, except that no electricity is needed. The techniques given here were developed by Thelma Moss, who pioneered the techniques of energy field electrophotography. From Dr. Moss' descriptions, you will need either a darkroom and photo lab, or access to a camera shop which can develop individual sheets of photo print paper. First, obtain some light-sensitive color or black and white print paper, making sure you do not open the package except in complete darkness. Obtain also a separate, empty, light-tight film box in which you can run your experiments. While in the dark, remove a piece of film paper and place it on the bottom of the empty film box. On top of the sheet of film, place a flower bud, leaf, piece of fruit, sprouting seedling, magnet, or other energized, living, or once-living object. Close the box and wrap it completely inside an orgone energy blanket, or place it inside a strong orgone accumulator. Make sure the film box is light-tight; cover it and the blanket with black plastic, or otherwise insure that no outside light will reach the film. Wait a day or two, or up to a week, and then develop the photo paper. You should see exposure patterns on the film that have radiated from the object you used in the experiment. Try different exposure times and film types; remember that on rainy days, the orgone charge is weaker and may take longer to expose the film. Make note of weather conditions during the trials. These photos should yield patterns similar to those seen on Kirlian electrophotographs, except that here, no electrical field was introduced. These are 100% *life energy photographs,* orgone-enhanced and directly recorded on film.

I) The Accumulator Temperature Differential Effect: Reich demonstrated that the warm glow felt inside the accumulator possessed an objective aspect which could be measured with a sensitive thermometer. An air-tight orgone accumulator will spontaneously warm up the air inside itself by a few tenths of a degree, up to several degrees. This temperature increase will make the accumulator interior slightly warmer than the surrounding air temperature, or

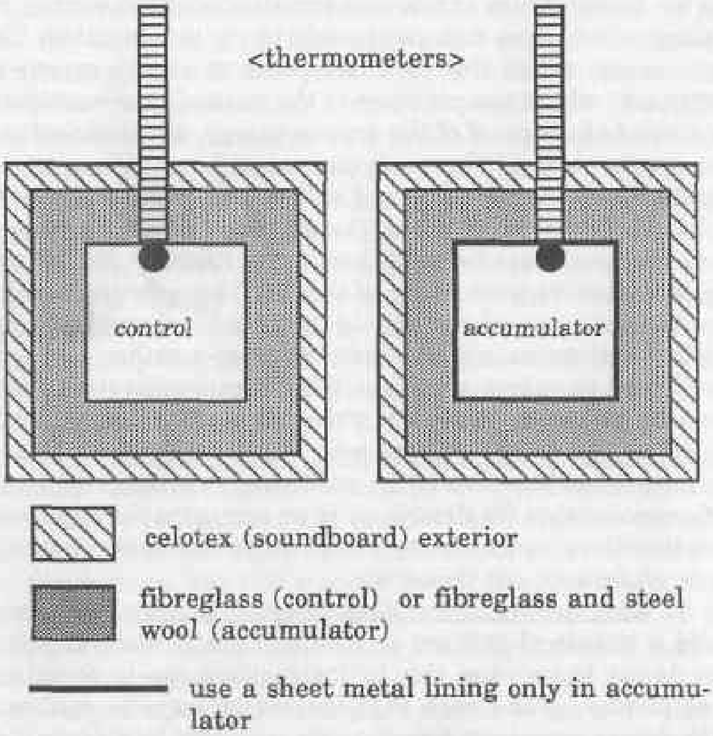
<thermometers>
control
accumulator
celotex (soundboard) exterior
fibreglass (control) or fibreglass and steel wool (accumulator)
use a sheet metal lining only in accumulator

The To-T Experiment: The orgone accumulator (right) and control (left) must be matched as closely as possible with respect to size, ventilation, and thermal dynamics for quantitative orgone energy experiments. The interior and exterior of the control enclosure should be the same size and dimensions as the accumulator, with identical lid dimensions. However, sheet metal and steel wool should not be used in the control enclosure. While metals are used in construction of the accumulator, they are strictly excluded from construction of the control. In the control, additional plastic or fibreglass layering is used in the place of steel wool. The thermal resistance *and* heat capacity *must be as closely matched as possible. A similar set of environmental conditions must also be maintained for the two enclosures.*

the air temperature inside a thermally-balanced control enclosure which does not use metals in its construction. This experiment, called the To-T (temperature in the orgone accumulator minus temperature of the control) was considered by Reich to be a proof of the orgone energy, and a violation of the second law of thermodynamics. Albert Einstein once replicated the experiment and called it a "bomb in physics"; a fascinating booklet titled *The Einstein Affair* documents the correspondence between Reich and Einstein on the matter. A definitive evaluation of the To-T experiment requires the construction of thermally balanced accumulator and control enclosures, careful monitoring of weather and environmental temperatures, sensitive thermometers capable of recording down to tenths of a degree, and prolonged systematic measurements. Those who wish to reproduce this experiment should consult the published reports given in the reference section for details. It is an area ripe for innovative investigation, and I strongly encourage the experimentalist to carefully seek out this effect.

J) The Accumulator Electrostatic Effects: Obtain or build a simple aluminum or gold leaf static electroscope. If you do not know what this is, instructions can be found in a good library. Make sure that the electroscope is calibrated with degree markings, from 0 to 90, such that its degree of deflection can be accurately measured. By running a plastic rod or comb through your dry hair, you can gather a significant static electrical charge and transfer it to the electroscope. Using a stopwatch, or a watch with a second hand, determine how long it takes for the electroscope to slowly lose its charge into the air, through a predetermined deflection angle. For example, you want to know how long it will take for the electroscope to discharge from a 50 to a 30 degree angle. You therefore should charge the electroscope up to a deflection angle *greater* than 50 degrees, waiting until it discharges to the 50 degree mark. Once this happens, you can count the number of elapsed seconds that pass until it reaches 30 degrees. The time elapsed is the *electroscopical discharge rate*. On sunny days, the discharge rate will be quite slow, while on rainy days, the discharge rate will be very quick, so quick that you might not even be able to measure it. If you time the electroscopical discharge rate inside an orgone accumulator, you will find that *it takes a longer time to discharge in the accumulator than in the open air*. The difference between the discharge rate inside the accumulator, and the discharge rate in the open air, is called the electroscopical discharge rate differential. This differential will be large on clear, sunny days, and minimal or zero on rainy overcast days. On rare occasions, an electroscope that is only weakly charged, or even completely discharged, may--if allowed to soak inside an orgone accumulator--spontaneously charge itself up to a higher level. All of these effects will vanish on rainy, overcast days. For more details, see the citations in the "Reference" section.

K) The Accumulator Evaporation Suppression Effect: This experiment requires a sensitive weighing scale that can measure to fractions of a gram. It also requires an accumulator and a thermally-balanced control enclosure of similar dimensions. For this control, do not use water-absorbent materials in the interior; instead, line the interior of the control with a non-metallic waterproofing material, such as plastic, enamel or varnish. Obtain and weigh two small, identically shaped and sized glass dishes of about 4" diameter and 1" high. Weight the dishes when empty, clean and dry. Next, add identical quantities of water to each vessel, filling them about halfway, and weigh again, calculating via subtraction the weight of water in each dish. Place one vessel of water inside the orgone accumulator, elevated on a small wood block, such that the bottom of the dish does not come into direct contact with the metal interior of the accumulator. The lid of the accumulator should be shut, but fixed open with a crack such that air may circulate. It should not be placed in a windy or sunlit area, however. Place the second vessel of water inside the control enclosure in a similar manner, on an identical wood block, and also prop its lid open. Place it in a location at least a meter away from the orgone accumulator, but with similar light, temperature, and wind characteristics. You may wish to drape a piece of black plastic over both the accumulator and control, in order to

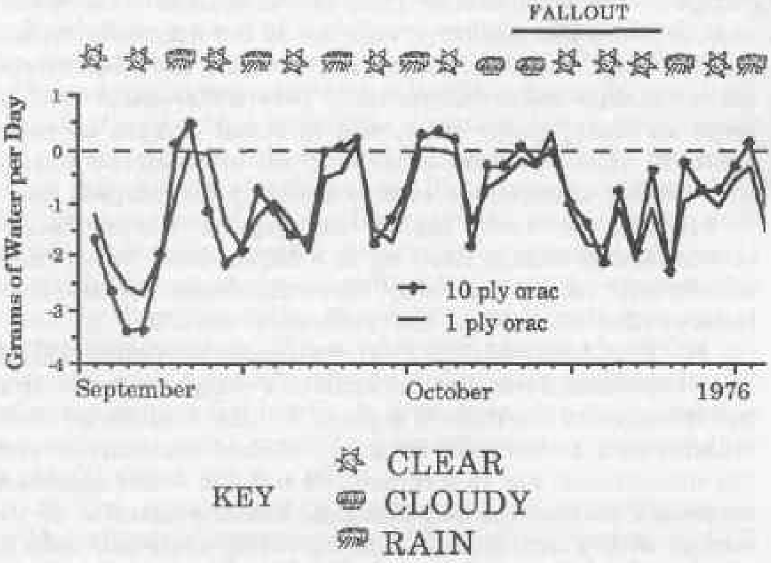

Water Evaporation Curve, EVo-EV: The curves identify the amount of water evaporated each day from a one cubic foot, ten ply orgone accumulator, and a similar one ply accumulator, minus the amount evaporated from a control enclosure. As the accumulator builds up its charge on clear, sunny days, it suppresses the evaporation of water inside itself. The control enclosure, however continues to evaporate moisture at a relatively high rate, up to several grams per day more than the accumulator. On wet, stormy days, the effect vanishes as orgone charge is lost at the Earth's surface, being taken up into the stormclouds. Disturbances in the regularity of the curve may also occur, such as (above) when radioactive fallout arrived at the laboratory site causing the accumulator to go temporarily "dead". (from DeMeo, J.: "Water Evaporation Inside the Orgone Accumulator", J. *Orgonomy*, 14:171-175, 1980) control for slight differences in light. Wait exactly 24 hours and remove the dishes of water, being careful not to spill any. Carefully weigh the dishes and compute the evaporative loss for the 24 hour period. Make this measurement once per day, preferably in the late evening, such that you can determine the amount of water evaporated each day from each container. You should find that the control enclosure evaporates significantly more moisture on clear, sunny days, when the accumulator suppresses water evaporation. On rainy days, when the accumulator no longer builds up a charge, the evaporation in the accumulator and control will be nearly identical. Subtract the quantity of water evaporated in the orgone accumulator from the quantity evaporated in the control for each 24 hour period. This quantity, called EVo-EV, will reveal the changing quantity of orgone energy charge in the local atmosphere, and in the accumulator. The evaporative values on any one day are less interesting than the dynamic manner in which the evaporation differential increases and decreases, according to the orgone energy charge at the Earth's surface.

14. Questions and Answers

- *Q: If the orgone energy really exists, why don't we hear about it from scientists working in the universities?*

A: Scientists working in the universities and research institutions have engaged in verifying research on the bions, orgone accumulator, and cloudbuster, and on the bioelectrical aspects of life, which Reich also pioneered. For example, Dr. James DeMeo, author of this *Handbook*, did research on the weather-related aspects of Reich's discoveries while he was a graduate student and Instructor at the University of Kansas. He continued with that research while a member of the faculty at Illinois State University and the University of Miami. Müschenich and Gebauer, of the University of Marburg in West Germany, recently completed a double-blind, controlled study on the physiological effects of the orgone accumulator on humans. Other scholars with a research or historical interest in the works of Wilhelm Reich have held positions at Harvard University, Temple University, the State University of New York, York University, McGill University, and elsewhere. Other research scientists have repeatedly verified and confirmed Reich's findings. Workshops and courses devoted to his works are now held at a few colleges and universities in the USA. Nonetheless, the history of science repeatedly shows that large institutions do not easily accommodate innovative research which may force radical changes in the major theories of science.

- *Q: Can an accumulator be used during wet or cloudy weather?*

A: Use of an accumulator during wet weather conditions will not be harmful, but it will be less effective, as the charge is significantly lower or absent at those times. It is best to use it during clear and sunny weather, when the atmospheric orgone energy continuum is strong and expansive, and the charge at the Earth's surface is greater.

- *Q: These accumulating devices are quite simple to construct. Aren't a lot of them accidentally constructed?*

A: A lot of "accumulators" are being constructed, without knowledge of the people involved. Every mobile home or house with a metal skin or side paneling will accumulate a charge, and a toxic one if aluminum is used. Oranur, and other toxic effects appear to readily occur in such houses, which are filled with all the modern orgone disturbing electromagnetic appliances, such as TV sets, microwave ovens, fluorescent lights, and so forth. Epidemiological studies that would address these general observations have never been performed.

- *Q: I have an old styrofoam beverage cooler. Can I line it with aluminum foil and make an accumulator?*

A: You can try this, but don't expect any firm results unless you comprehend and take into consideration virtually *all* of the procedures and warnings given in this *Handbook*. Styrofoam and aluminum are life-negative accumulating materials. If you run a biological experiment, you may wind up demonstrating only a life negative effect. For the scientist interested in orgone energy, these considerations are even more crucial and cannot be ignored.

- *Q: My accumulator gave very good charge the first months when it was in use, but now does not yield a good charge any more. Why is this?*

A: It is likely that the accumulator has been contaminated with dor. Some researchers have noted this effect, where the accumulator goes temporarily "dead", and hence keep their accumulators outdoors, sheltered from rain, but in the fresh air, with the lids or doors open so that air can circulate inside quite freely. You may be able to refresh a "dead" accumulator by wiping it inside and out with a damp cloth, every day for a week or so. Also, keep a bowl of water, or a draw bucket with draw tubes, inside it when it is not in use. Change the water in this bucket every day. Also, be certain that the accumulator is not near any of the oranur-producing devices previously identified, and that your neighborhood is as oranur-free as possible. The accumulator can also be sun-charged, by allowing it to sit in the direct sunlight for a few days. These steps should eliminate any dorish tendencies, and "rekindle" the charge.

- *Q: I have heard that sitting inside an accumulator will make a person more sexually potent. Is this true?*

A: This is mostly a false rumor propagated by Reich's enemies, who wrote smear articles in the 1940s and 1950s that branded Reich a lunatic, the accumulator a "sex box", and put false words into his mouth about the ability of the accumulator to restore lost sexual potency. However, Reich never made such a claim; in fact, he continually stressed the emotional and psychological foundations of sexual dysfunction, which could not be touched by treatment in an accumulator.

- *Q: Is the orgone energy accumulator legal? Can I get in trouble with the law for building or using one?*

A: There is no law against orgone energy, or the orgone energy accumulator. You may build, own, and use the blanket or accumulator in your own home, or elsewhere, as you choose. It can also be legally used for the self-treatment of any health-related condition, just as you can make very beneficial soups, purchase vitamins, or take soaking baths, without asking a doctor or the police. Understand, however, that forces within the medical community, pharmaceutical industry, and government are hard at work to make it illegal for you to do these things. If you are concerned about protecting your health freedoms, you should join forces with those social organizations that are working to preserve or extend those freedoms. *The price of liberty is eternal vigilance!*

Part III: Construction Plans for Orgone Accumulating Devices

15. Construction of a 2-Ply Orgone Energy Blanket

The orgone energy blanket is the most simple to construct of all the orgone charging and accumulating devices. They can be made at any size, and can be easily transported. You can use small ones while resting, or place large ones under and over a person who is immobilized in a bed. Like a standard accumulator, orgone blankets are not meant to be used for any prolonged period, though one can rest or nap with one if there is a need. In my experience, people will, even when sleeping, push off an uncomfortable orgone blanket just as they would a regular blanket. The following steps tell how to make an orgone energy blanket with final dimensions of 2' by 2'.

A) Obtain enough 100% wool fabric, or acrylic felt, to make three 2' x 2' squares. Also obtain several packages of very fine ("000" or "0000") steel wool pads.

B) Lay a 2' x 2' piece of fabric on a flat surface. Cover the exposed top surface of this fabric with a layer of steel wool, from unwrapped steel wool pads. Spread the steel wool out, so that it is not too thick. You should be able to see parts of the underlying fabric here and there.

C) On top of the steel wool, lay another 2' x 2' piece of fabric.

D) Cover the exposed top surface of this second fabric piece with another layer of steel wool.

E) Finish with another 2' x 2' piece of fabric, placed on top of the last steel wool layer. You now should have

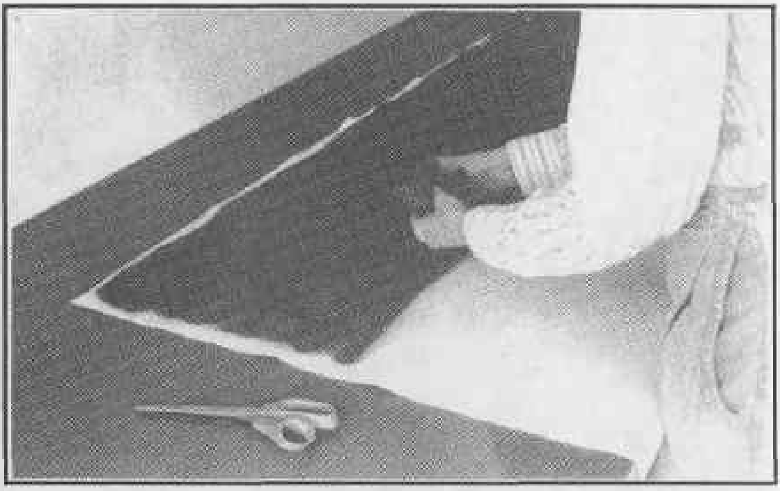

Spread steel wool loosely on section of cloth. Trim edges as shown.

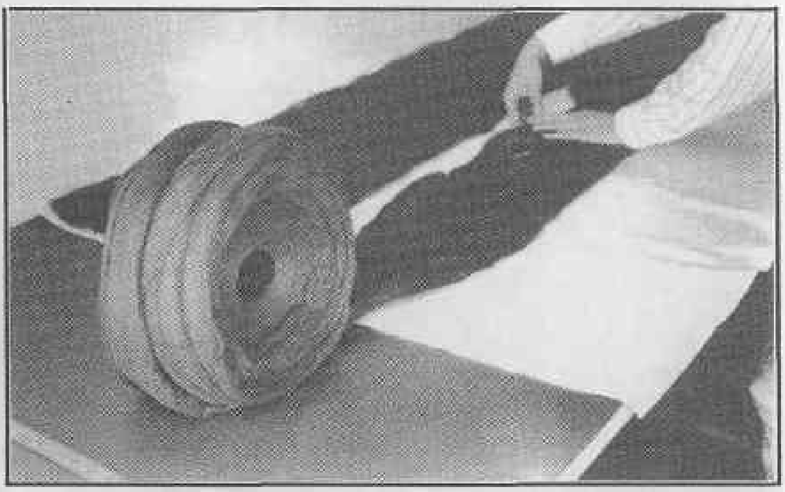

Rolls of steel wool can be purchased from manufacturers, or common steel wool pads, of "000" or "0000" grade, can be obtained at most hardware or paint stores.

Particle masks should be worn during construction to avoid breathing fine steel dust.

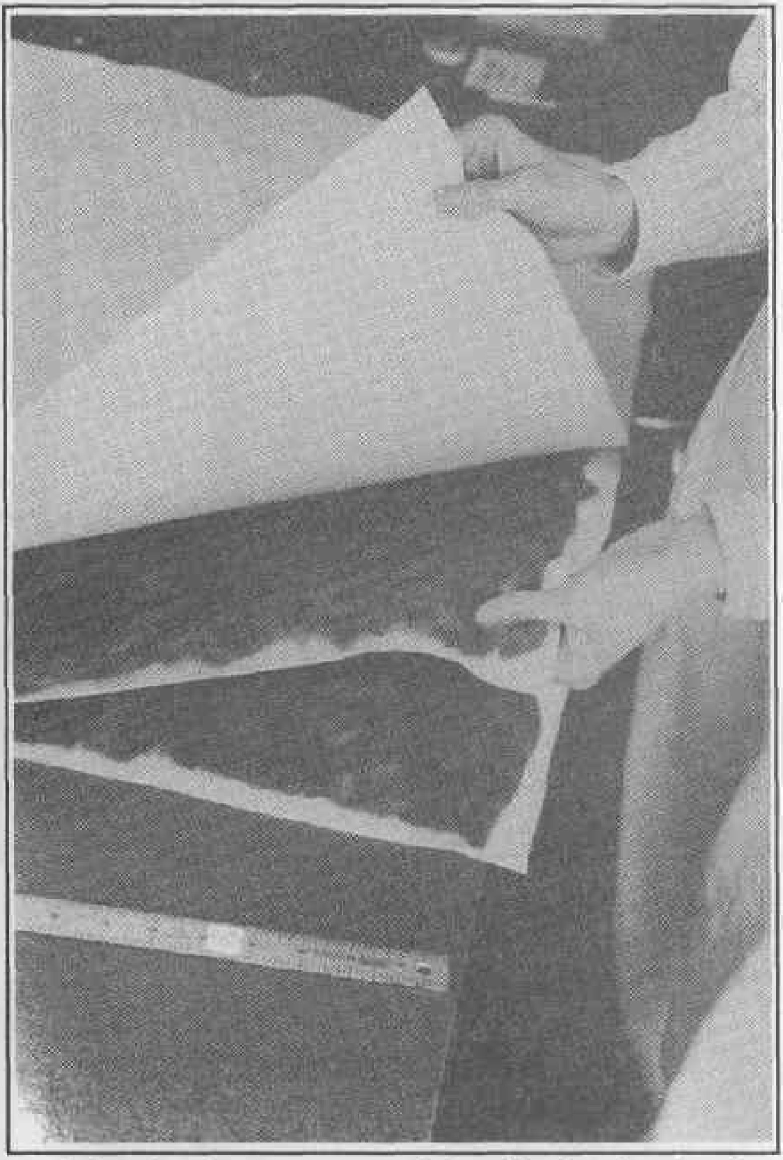

A functional orgone accumulating blanket showing the alternating layers of steel wool and cloth.

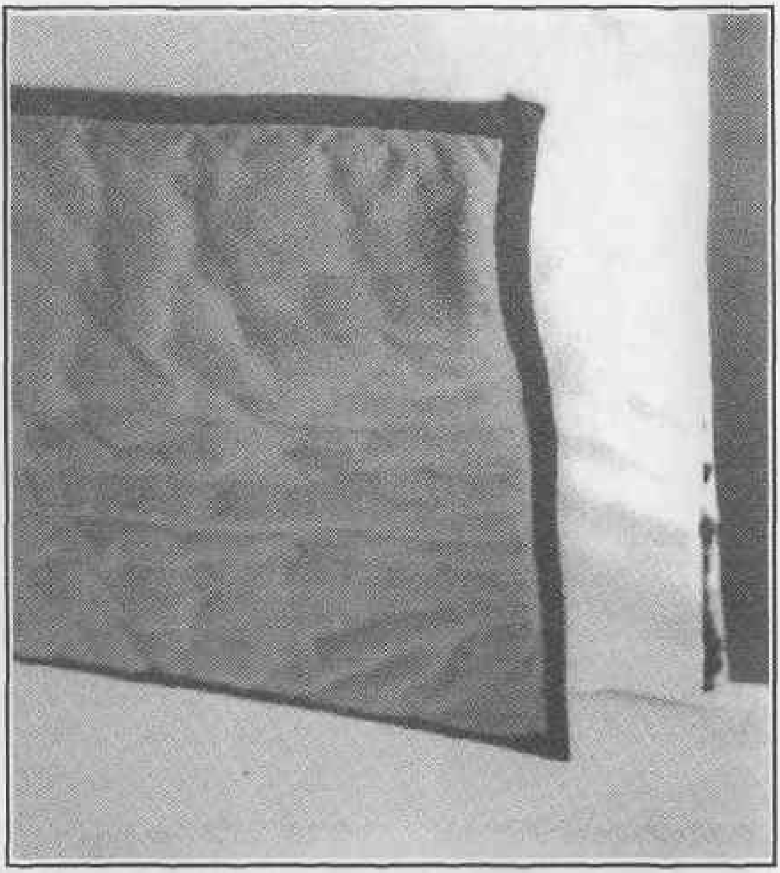

Finish off your blanket with a trim to keep everything together. Several quilt-stitches should be added to keep the insides from shifting around.

three pieces of fabric, with two layers of steel wool sandwiched in between.

F) Trim, stitch, and finish the borders, according to your own tastes and sewing skills.

G) Keep and use the blanket in an environment similar to that advised for a regular accumulator, away from any TV sets, microwave ovens, fluorescent lights, or other electromagnetic or radioactive gadgets. *Never use an orgone blanket with an electric blanket.* It can be stored on a hangar in the open air, or even inside a larger accumulator for greater charging.

H) Do not ever wash or dry clean your orgone blanket, as the steel wool will rust! Spot clean only with a *slightly* damp sponge.

I) Reich once made very heavy orgone blankets, composed of galvanized steel wire mesh, and alternating layers of wool and steel wool. While these work quite well, I find them to be uncomfortable and difficult to use. They do not appear to be any more effective than the simple design given here.

16. Construction of a 5-Ply Garden Seed Charger; The "Coffee Can" Accumulator

One can make a simple garden seed charger from a cleaned food container can or coffee can, using additional steel wool and fabric materials.

A) Empty a large coffee can or other steel or steel/tin food container (no aluminum!), clean it, remove all labels, and dry thoroughly. Make sure you save the cut-out metal lid, or make a replacement from another can, or from galvanized steel sheet metal. Use a can large enough to hold all the seeds you will be charging.

B) Obtain several yards of good 100% wool, or acrylic felt fabric. You will need enough fabric to go around the can about 5 times, plus enough for 5 round top pieces, and 5 similar bottom pieces.

C) Purchase several packages of very fine grade ("000" or "0000") steel wool. You will need enough unwrapped steel wool to cover an area equal to that of the fabric. Again, unwrap the steel wool pads as you need them, and spread it out.

D) Cut the fabric into a very long strip which is as wide as the can is tall. The length of this long strip should be about 6 times the circumference of the can. As you may not have a single strip of fabric this long. You can splice several pieces together.

E) Lay the long fabric strip out flat, and spread a thin layer of steel wool on top of it. Lay the empty can on one end of the steel wool/fabric strip and roll the can up inside the strip. Stop when the strip has been wrapped around the can about five times or more. Add a final layer of fabric to the outside, and stitch or tape it in place so that it will not unravel.

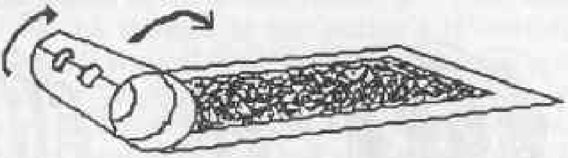

F) Measure the diameter across the top of the can, including the wrappings of fabric and steel wool. Cut out 10 circles of fabric material of this same diameter, 5 for the top, 5 for the bottom.

G) Sandwich steel wool between the fabric circles such that you have 4 layers of steel wool between 5 layers of fabric. Make two of these fabric/steel wool sandwiches, one of which will be used to cover the bottom of the can, and the other for the top.

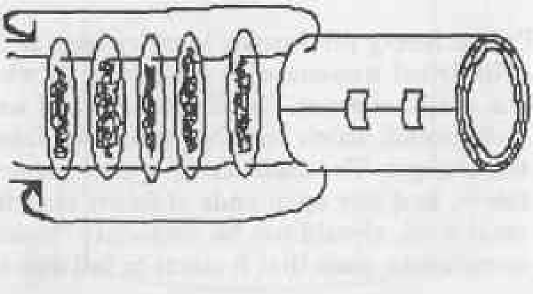

H) Take the metal disk cut from the top of the can and file smooth any jagged edges. Punch two small nail holes near the center, about 1/4" apart. Using a heavy upholstery or knitting needle, thread some heavy twine, wire, or yarn through the center of one of the steel wool/fabric sandwiches, and bring the twine through the holes in the metal can lid. Secure the can lid to the center of the steel wool/fabric sandwich. The steel wool/fabric sandwich should be about 2" wider in diameter than the metal can lid.

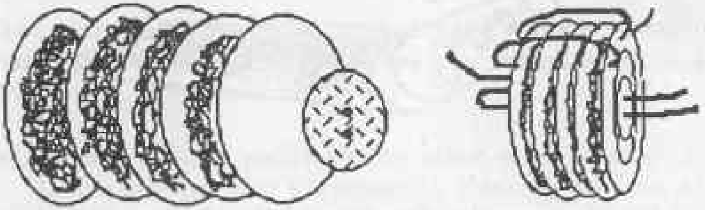

I) Using heavy thread, loosely stitch together the edges of the top steel wool/fabric sandwich (the one sewn to the metal can lid). Also loosely stitch the edges of the bottom steel wool/fabric sandwich together, and sew it to the bottom edge of the steel wool/fabric strip, wound around the can. Except for the top opening, the metal can should now be encased in the steel wool/fabric material.

J) Find a heavy pillowcase, laundry bag, or other larger cylindrical non-metallic container in which to keep the entire charger. Or, if you are good with a needle and thread, stitch together your own fabric cover for the charger. The main thing is that the outer layer of fabric, and any open ends of fabric showing pieces of steel wool, should not be subject to "knocking about" or moisture, such that it starts to fall apart or rust.

K) Review the section on seed charging, in the "Simple Experiments" chapter for instructions and additional ideas on the use of your charger. Or, as an alternative to constructing this accumulator, you could store your seeds inside a large metal cookie tin which would then be wrapped up inside the multiple folds of a very large orgone blanket, or be stashed inside one of the other, larger accumulators. Just realize that the greater the number of plys, and the greater the absolute quantity of materials going into the construction of the accumulator "pile", the stronger will be the charge. At the author's laboratory, for instance, the small 5- ply coffee can charger is stored inside the 10-ply, one-cubic foot accumulator, which in turn is stored inside the 3-ply large accumulator. This is a total of 18 plys, and yields a charge which is readily sensible.

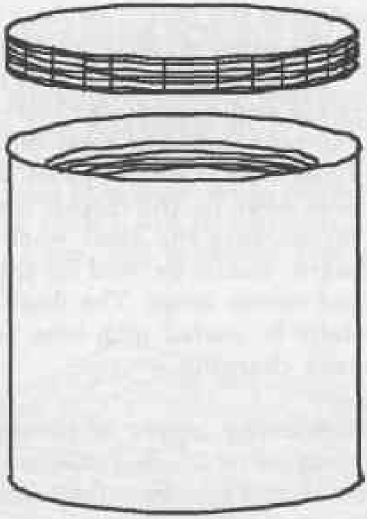

17. Construction of a 10-Ply Orgone Energy Accumulator

A very powerful, one-cubic foot, 10-ply accumulator can be made following the instructions below.

A) Cut six galvanized steel, 27 gauge sheet metal squares, measuring 1' x 1'. Use heavy tape, on the outer metal walls only, to construct a metal cube. Leave the top of the cube open, and do not tape it into place. The interior of the cube should remain bare metal, without any tape showing.

B) Use very fine ("000" or "0000") steel wool, and heavy clear plastic acrylic carpet protector material for construction of the plys. The clear plastic carpet protector is the same material used in model homes to protect carpets from wear, and is often sold on rolls in hardware and department stores. It is not cheap, but works very well. The clear plastic carpet runner should have rows of small tips on the side which normally faces down on the carpet; these tips work very nicely for holding the steel wool in place. Celotex (soundboard) should be used for the final outer layer, with wood corner strips. The final outer celotex may additionally be coated with bees wax and/or shellac to increase charging.

C) Ten alternating layers of plastic and steel wool roughly measure 2" in thickness. This being the case, you should construct the outer celotex casing in the shape of a cube, the interior dimensions of which will be 16" x 16" x 16". Cut six celotex panels with the following dimensions:

Celotex Panels:

top: 17" x 17"
bottom: 17" x 17"
2 sides: 17" x 16"
2 sides: 16" x 16"

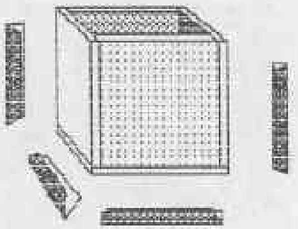

D) Use small nails and glue to fasten five of the six celotex panels together, to make a cube shape. Again, as with the metal box, do not attach the top. Add extra glue to the edges of the assembled celotex box, and allow it to dry before proceeding.

E) Using a mitre box, cut wood corner strips for the outside edges of the celotex box. Nail and glue these wood corner strips to the celotex box for added strength.

F) Cut 20 square pieces of plastic carpet runner, 16" x 16". Put ten of these squares aside for later use. Lay the other ten squares one at a time inside the bottom of the celotex box, tips facing up. In between each plastic square place a layer of steel wool, taken from unwrapped steel wool pads. When finished, the top of the last plastic square will face upward in the bottom of the celotex box, and it should also be covered with steel wool.

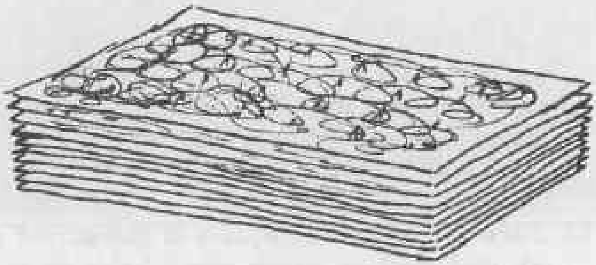

G) Place the galvanized steel sheet metal cube inside the celotex box, on top of the ten plys of plastic and steel wool. If you have constructed the celotex box correctly, the top of the metal cube should be about 2" below the top of the celotex box, and a space of about 2" should exist between the sides of the metal cube and the interior sides of the celotex box.

H) Cut 20 pieces of plastic, 12" x 16", and 20 more pieces, 12" x 12". These will be used to fill the side spaces between the metal cube and the celotex. Layer each plastic piece with steel wool, and stack into bunches of ten layers each. Do this on a flat surface before attempting to place them into a vertical position, in between the celotex and metal boxes.

I) Place the two stacks of 16" x 12" plastic/steel wool between the celotex and metal cube, on opposite sides of the metal cube. An outer layer of plastic should lay against the inner wall of the celotex box, while an inner layer of steel wool should rest against the outside of the metal box. The upper edge of the plastic should be nearly flush with the upper edge of the metal cube, both of which should remain about 2" below the upper edge of the celotex box.

J) Place the two stacks of 12" x 12" plastic/steel wool in the two remaining spaces between the celotex and metal cube, as given in the previous step.

K) Take the 10 remaining pieces of 16" x 16" plastic carpet runner and layer them with steel wool. Stack them up and set them aside. Unlike prior stacks, however, do not finish the final plastic layer with steel wool.

L) Take the remaining square of galvanized sheet metal, and drill or punch small holes in each corner, about 1/2" from each corner. The holes should be

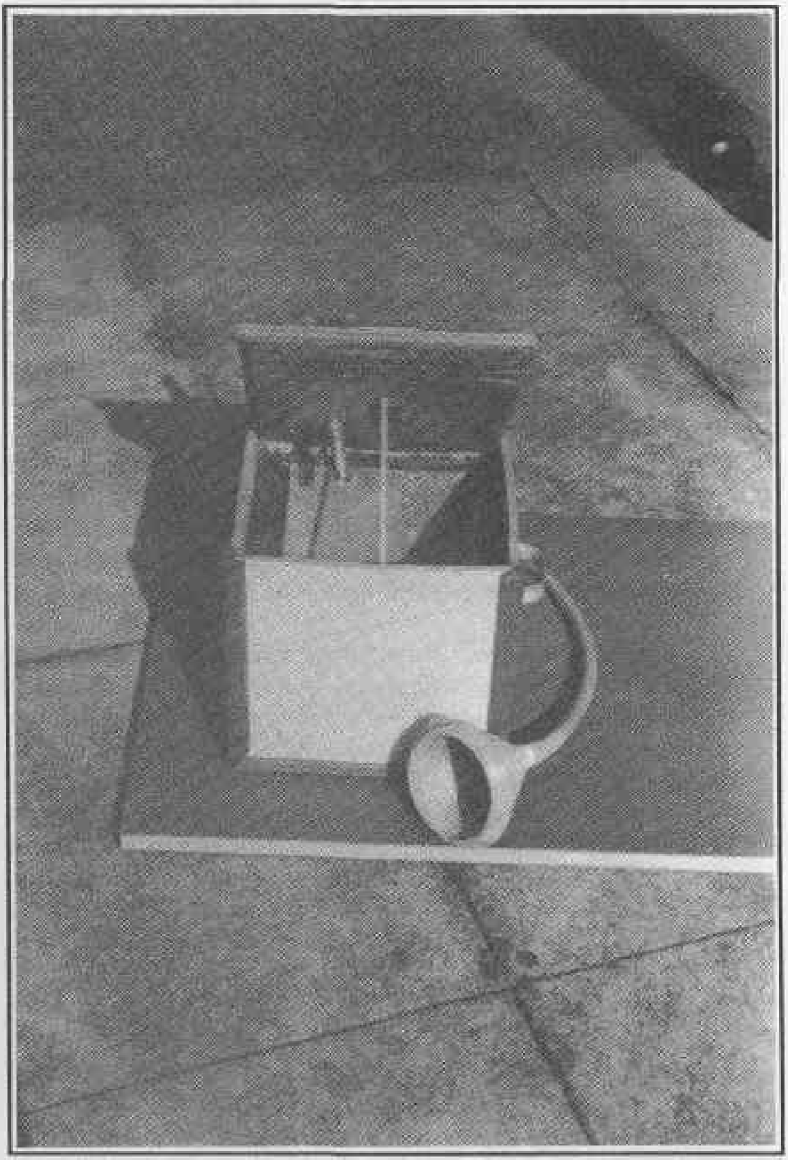

A one-cubic foot, 10-ply accumulator with attached funnel.

large enough to accommodate a long but narrow screw.

M) Find a rough wood or carpeted surface to work on. Lay the stack of plastic/steel wool squares on the last piece of celotex. Center it on the celotex; there should be about 1/2" of celotex appearing all around the stack. Now place the sheet metal square (with the holes) on top of the plastic/steel-wool/celotex stack, and center it also. There should be about 2" of plastic extending beyond the edge of the metal square on each side. Use some masking tape to temporarily hold the celotex/plastic/steel stack together.

N) Using an ice pick, carefully make four vertical holes through the plastic/steel wool stack, and through the celotex, using the corner holes in the sheet metal square as a guide. Do not use a drill, as the steel wool will spindle around the drill bit.

O) Using four skinny bolts, along with nuts and LARGE washers, secure the metal square and plastic/steel wool stack against the celotex square. Use a bolt no longer than necessary, such that the bolt ends will not protrude significantly. When completed, this whole lid assembly should fit snugly on top of the celotex box. The metal plate attached to the lid should closely, but not perfectly, align with the metal cube interior. With the lid in place, only bare metal should face to the interior of the accumulator.

P) For handles, first firmly glue a flat, wide and long wood strip to two outer sides of the celotex cube, near to the top. When completely dry, screw wood or metal handles to these wood strips. A handle should also be attached to the upper, outer side of the lid assembly, using a similar support mechanism. The celotex is simply too lightweight to accept screws or nails alone. You may likewise install a hinge between the lid assembly and the box, or coaster wheels to the bottom, but these are not necessary.

Q) For additional charging strength, the outer celotex walls may be coated with melted bees wax or paraffin (not really necessary), but in any case should be given several coats of protective natural shellac.

R) For additional charging strength, store this cubical accumulator in the bottom part of a larger, human-sized accumulator, under the bench that you normally sit upon. Be sure to maintain a clean and uncontaminated environment for your accumulator, as per the points given in the previous chapters. Prop the lid open when not in use, and store in a clean, dry place, without electromagnetic or nuclear contamination.

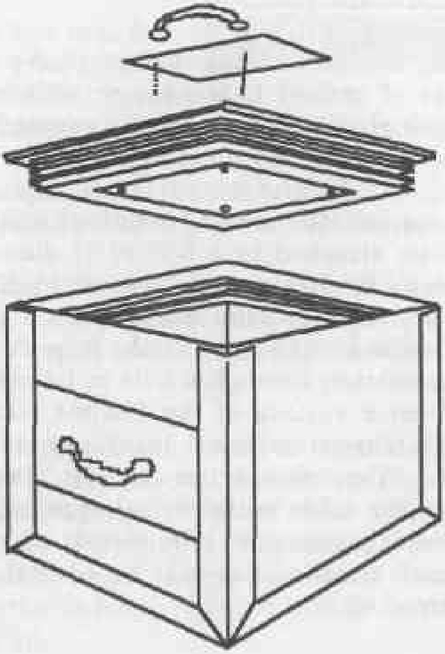

18. Construction of an Orgone Shooter Funnel

The shooter funnel is similar to other accumulating devices, but has an open face allowing for external irradiation of objects. It is often connected to a larger box accumulator, but this is not absolutely necessary.

A) Obtain a galvanized steel funnel with about 6" diameter at the large end, from a hardware, farm supply, or auto supply store. Auto supply stores sometimes sell these with a flexible metal hose already attached, for dispensing oil into a car engine, and this may help in the later attachment to a box accumulator.

B) Coat the outer metal surface of the funnel with a layer of melted bees-wax, or with strips of black plastic electrical tape, leaving exposed the bare metal surface of the funnel interior.

C) If desired, the small "draining" end of the funnel can be attached to a 3/4" or 1" diameter length of hollow, flexible metal "greenfield BX" cable (use galvanized steel, *not aluminum*). The other end of the cable is then run into the interior of a small box accumulator, through a hole in its side or lid. Wrap the outer surface of the flexible metal cable with black-plastic electrical insulation tape. (See page 121.) Your shooter funnel will then draw orgone down the cable to the funnel opening, increasing its radiating strength. Or, simply store the shooter funnel inside a box-type accumulator, to keep it charged up.

19. Construction of an Orgone Shooter Wand

The shooter wand is a very simple means for demonstrating the subjective sensations of the orgone radiation, and also for irradiation of orgone energy into body cavities. Simply store the completed shooter wand inside a well maintained accumulator, and remove it for use when necessary. Under relaxed conditions, most people can hold this wand in the hand, or place it on the solar plexus or upper lip, and readily feel the soft radiant glow of the orgone energy.

A) Obtain a pyrex test tube of about 3/4" to 1" diameter, and 6" to 9" long, from a laboratory or medical supply company.

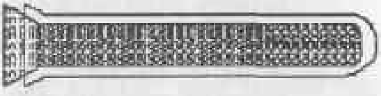

B) Fill the test tube full of fine grade ("000" or "0000") steel wool. Compress to a reasonable firmness.

C) Seal the open end of the test tube with a rubber stopper, and tape it shut with plastic electrical tape.

D) Place the shooter wand inside a small orgone accumulator for a period of several days or weeks before using. Keep it stored inside the accumulator in between uses.

E) If you use the shooter wand to irradiate the throat or other body cavities, or if the pyrex glass otherwise becomes soiled, wipe the glass clean with isopropyl alcohol prior to putting it back inside the accumulator. The wand should always be alcohol cleaned and air dried prior to being replaced into the accumulator for charging.

20. Construction of a Large 3-Ply Orgone Energy Accumulator

This accumulator is large enough to sit in, and is composed of 6 large rectangular panels. Each panel is made from a wood frame, galvanized steel sheet metal (27 gauge), steel wool, fibreglass (*without* aluminum backing), and celotex (soundboard). One side of each wood frame is faced with galvanized steel, the other with celotex, and the three alternating layers of fibreglass and steel wool are sandwiched in between.

A) First compute the size of the panels for an accumulator that will fit your personal needs, adding the necessary dimensions for overlap between the various panels. The side panels and back panel should physically sit on the edges of the bottom panel. The back panel should snuggle between the two side panels. The top panel should overlap and rest upon both side panels and the back panel, covering them. The door panel should, like the back panel, snuggle between the side panels when it is shut. I believe this arrangement is the simplest possible, and most efficient to build. Dimensions are given below for accumulators to accommodate people of varying sitting heights, but of average weight. As the distance of the body surface from the metal walls increases, there will be a reduction in effectiveness of the accumulator. The dimensions of the accumulator should be carefully selected to meet your needs. An additional 1/2" clearance is provided in the width dimension (1/4" each side), such that the door will open and shut freely.

Panel Dimensions:	*Large Size*	*Medium Size*	*Small Size*
Top:	29.5" x 35"	26.5" x 32"	23.5" x 28"
Bottom:	29.5" x 35"	26.5" x 32"	23.5" x 28"
Left side:	35" x 58"	32" x 54"	28" x 50"
Right side:	35" x 58"	32" x 54"	28" x 50"
Back:	25" x 58"	22" x 54"	19" x 50"
Door:	25" x 52"	22" x 48"	19" x 44"

Interior Dimensions:			
Height:	58"	54"	50"
Width:	25"	22"	19"
Depth:	31"	28"	24"

You will need:
Height = height sitting erect in a chair + about 3"
Width = shoulder width + about 4" (2" each side)
Breadth = sitting, knee to back distance + about 3"

B) Make wood frames of 3/4" x 1-1/2" pine (called "one by two" in lumber yards), like picture frames, such that the outer edges conform to the computed dimensions for your accumulator panels. Nail and glue all joints.

C) Arrange the open wood frames together, as they will be when the accumulator is complete, to make sure that all dimensions have been properly calculated and cut. If there is an error in your calculations, now is the time to find out, before the more expensive celotex and galvanized sheet metal have been cut.

D) Cut the celotex panels to size, from 1/2" or 3/4" thick celotex sheet, to cover fully one side of each wood frame. Nail and glue the celotex panels to each wood frame. Use 1/4" plywood instead of celotex for the bottom panel (and only the bottom panel).

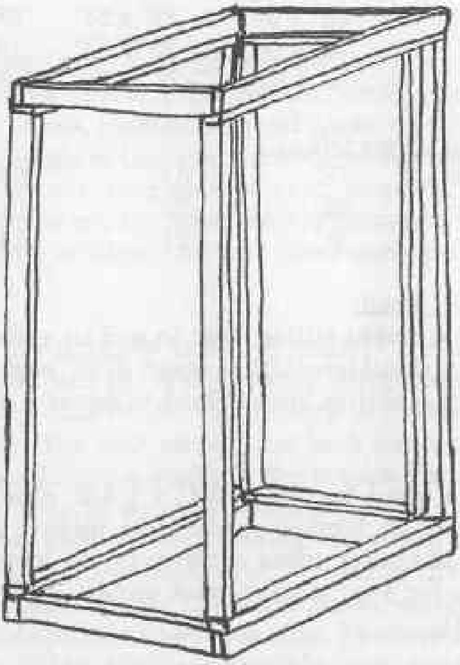

Assembled Wood Frames for Accumulator Panels

E) Cut batts of 1/4" thick fibreglass material to size and place a layer inside each of the open panel frames. Use gloves and a mask to protect yourself. Do not compress. Avoid lumps and holes. You can use wool, cotton, acrylic felt, or rock wool instead, if you wish, but for a large accumulator like this, the costs will be higher, and the accumulation will not be significantly stronger. These other materials may yield a different "feel" to the orgone charge, and if that is important to you, the cost may be justified.

F) Unroll very fine ("000" or "0000") steel wool pads and place a layer inside each of the open panel frames, on top of the fibreglass. Leave it fluffy, at around the same thickness as when unrolled, in as uniform a layer as possible. Some steel wool comes in large rolls which, if you can get it, will speed up construction for large accumulators.

G) Repeat steps E and F, placing a new layer of fibreglass on top of the previous steel wool layer, and another layer of steel wool on top of that.

H) Again repeat steps E and F, placing another new layer of fibreglass on top of the previous steel wool layer, and another layer of steel wool on top of that.

I) You should now have three alternating layers of fibreglass and steel wool contained in the open frame of each panel. The final layer facing you should be composed of steel wool. The panels should also be filling up, and may need to be slightly compressed before the final layer of galvanized steel sheet metal can be added. If you have used some other type of non-metallic material than fibreglass, and if the material lies loosely in the frame, you may have problems with slumping of the plys when the panels are finally sealed and placed in upright position. If this is the case, now is the time to do something to pre-

End Cross-Section of Side, Top, Back and Door Accumulator Panels

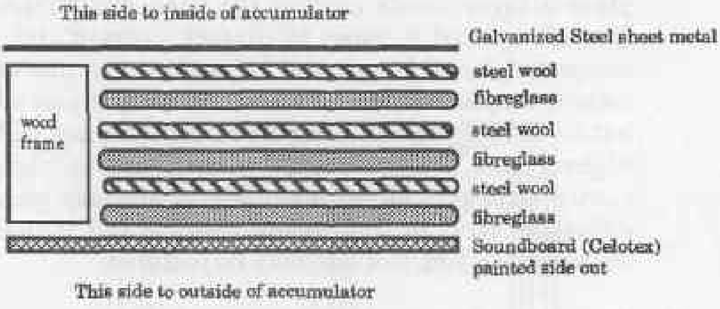
This side to inside of accumulator
Galvanized Steel sheet metal
steel wool
fibreglass
wood frame
steel wool
fibreglass
steel wool
fibreglass
Soundboard (Celotex) painted side out
This side to outside of accumulator

End Cross-Section of Bottom Accumulator Panel

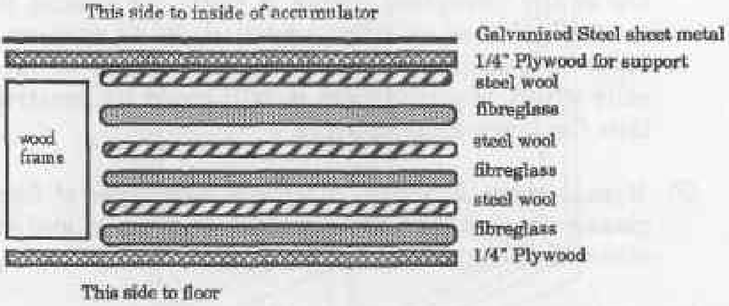
This side to inside of accumulator
Galvanized Steel sheet metal
1/4" Plywood for support
steel wool
fibreglass
wood frame
steel wool
fibreglass
steel wool
fibreglass
1/4" Plywood
This side to floor

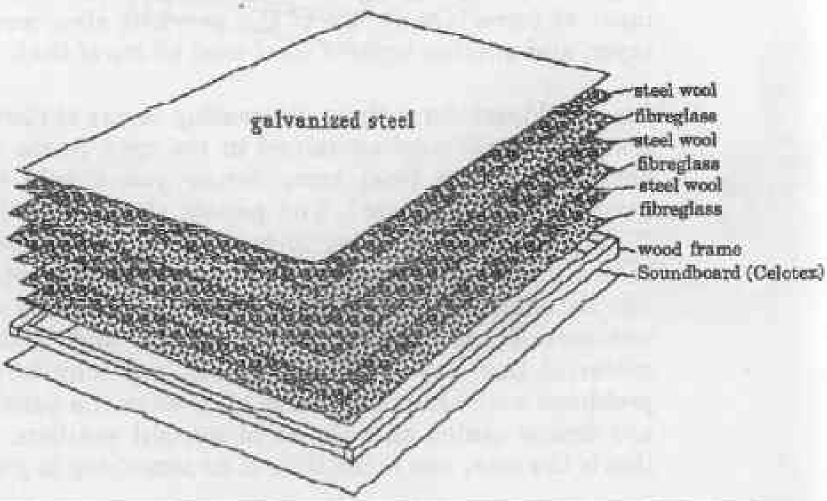
galvanized steel
steel wool
fibreglass
steel wool
fibreglass
steel wool
fibreglass
wood frame
Soundboard (Celotex)

Front View, full-size 3-ply accumulator

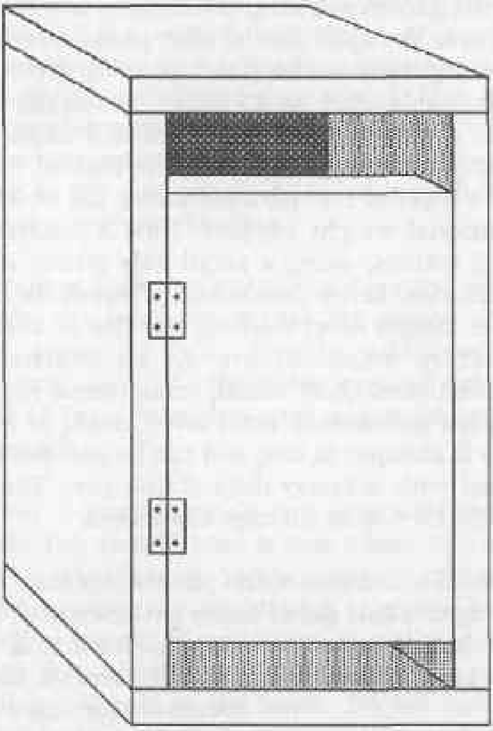

For Warmer, Tropical Climates: *Use door with 3" convection gaps at top and bottom, as per measurements on p.127. (Photo on p.23.) If needed, tack screen strips to door top and bottom, across gaps, to keep out insects.*

For Cooler or Cold Climates: *Use full height door with small window. (Photo on back cover.) Before cutting door materials, adjust door measurements (p.127) to only 1/4" less than full height of side panels. Before assembling door, cut matching 6" square openings in door sheet metal and celotex, centered at face height, then frame opening with 1x2" wood strips. Complete assembly as per instructions.*

Use hinges with removable hinge pins for easier assembly and take-down. Secure inside of door with hook and eye.

vent slumping.

J) Cut the galvanized steel sheet metal to size such that it fits over the open part of each panel, overlapping and covering the wood frames. Use the lightest sheet metal available, such as 27 gauge, which can be cut with hand shears for trimming, and still add to structural strength of the panels. For the bottom (floor) panel, add a layer of 1/4" plywood under the sheet metal, for additional weight support. Nail it securely onto the wood frames, using a small hole punch if necessary. Small steel tacks should easily penetrate through the sheet metal. After nailing, use file or shears to trim all sharp metal corners. As an alternative to galvanized steel sheet metal, some people have effectively used galvanized steel *wire mesh*, or *wire screen*. This is cheaper to use, and can be secured to the wood frame with a heavy duty staple gun. The steel wool should be visible through the screen.

K) Assemble and secure the panels together. Start by securing one side panel to the bottom panel, by using an "L" shaped metal brace at the front and rear of the side panel, near the bottom. Use screws, such that the accumulator can be taken apart later for easy moving. Secure the other side panel in a similar fashion, followed by the back panel. The back panel must be secured independently by small metal braces, placed between the wood frames of the bottom and back. Add the top panel, and secure it to the sides and back in a similar manner. The accumulator should now be quite sturdy, and is almost complete.

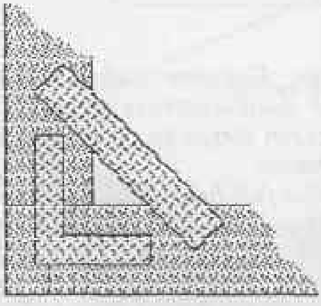

L) Carefully mark and drill holes for door hinges and attach them to the door panel and side panel. Make sure you center the door in a way that leaves equal space at the top and bottom, either for convection gaps (for the vented door style), or as clearance to allow the door to freely swing open and shut (for the full-height windowed door style). Use hinges with removable pins so that you can easily take the door off when moving the accumulator. After fixing the door to the side panel, it should open fully and close snugly, without binding.

M) Attach a hook and eye to the door and side panel opposite the hinge, such that the person sitting inside can secure it shut. Finally, add several coats of natural shellac to the outer exposed celetox, to protect it from humidity and add to its accumulating strength.

N) Your accumulator is now complete, except for a seat. You should have a seat which will allow you to place other accumulator chargers underneath. For this purpose, you might wish to specially construct a wooden bench. Wood is a good material to use, as it does not absorb the orgone energy significantly, and is not cold to the touch. Do not use woods that have been soaked in preservative or formaldehyde, however. Metal chairs are OK, but will be quite cold to sit on unless covered with a light fabric.

O) You may also wish to construct a *chest board,* or *orgone pillow*, for use inside the accumulator. As you sit inside, notice that there is a great distance from your chest to the front metal wall. This large distance inhibits orgone irradiation to your chest. An additional, small accumulator panel, similar to those used for the wall panels, could be constructed for use inside the larger box, to bring the radiation closer to the chest. However, a simpler way is to use a bundle of cotton, wool, or acrylic felt, rolled into a large pillow shape with equal layers of steel wool. The final outer layer should be composed of steel wool, and the entire bundle is then placed inside a thin cotton pillowcase. It should be large enough that it fits very snugly into the pillowcase. By holding this orgone pillow close to the chest, while you are in the accumulator, it will irradiate those frontal areas of your body that are not so well irradiated by the accumulator walls. Leave the orgone pillow inside the accumulator when not in use, to keep it charged up. You can also use this pillow outside of the accumulator, in a manner similar to the orgone blanket, with equally good results.

P) Do not connect electrical appliances to the accumulator. Follow the cautions given in the previous chapters. You can read a book while inside the accumulator, but should use either a strong *external* light source (to shine a beam of incandescent light into the accumulator), or you can use a battery-powered reading light inside. Again: *no fluorescent lights, television sets, heating blankets or pads, or other electrical or electromagnetic devices!!*

poor materials for orgone accumulator construction remain unchanged and time-tested.

Cautionary Note on New "Wireless" Electromagnetic Gadgets

On the subject of safe use of the accumulator (see Part II), we may add the following: Over the last decade, there has been an increased use of new "high tech" devices for telecommunications purposes which actively emit potentially toxic low-levels of electromagnetic radiation: cellular telephones and cell-phone relay systems (often located in populated areas) emit microwave radiations at potentially toxic levels, as well as new remote-control apparatus which employ *active far infrared (close to microwave) frequencies*, such as inventory scanners and various kinds of "wireless" technology. In all cases, orgone accumulators should not be located near to any of these devices, nor should any human being wish to live, work or sleep near them either! New electromagnetic devices should be viewed as contrary for safe use of orgone devices unless proven otherwise. On the other hand, various *passive* infra-red sensing devices (such as automated outdoor security lights) pose no problems.

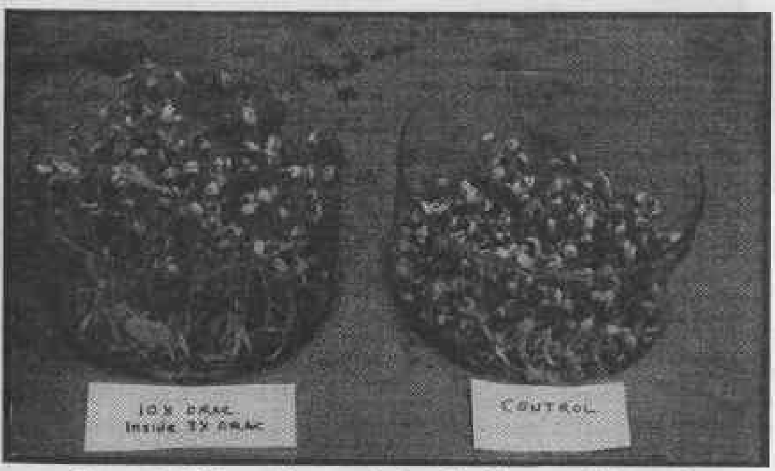

Sprouted Mung Beans: Left side, Orgone-Charged. Right side, Controls. (Experimental protocol p.96-97)

New Publications and Controlled Studies on the Human Bio-Effects of the Orgone Energy Accumulator

In recent years, mainstream publishers in Germany have expressed a growing interest in the works of Wilhelm Reich, and on the subject of his research findings (an interest not currently shared by any mainstream American publisher). Some examples of this interest are given below.

Appendix I of this *Handbook* (page 148-155) presents an extended abstract describing the first double-blind controlled study of human physiological response to the orgone accumulator, as undertaken by Stefan Müschenich and Rainer Gebauer. Their study was completed in 1986 at the University of Marburg in Germany. Dr. Müschenich subsequently wrote a major book on this same subject (in the German language):

Stefan Müschenich, *Der Gesundheitsbegriff im Werk des Arztes Wilhelm Reich (The Concept of Health in the Works of Dr. Wilhelm Reich)*, Doktorarbeit am Fachbereich Humanmedizin der Philipps-Universität Marburg, Verlag Görich & Weiershäuser, Marburg 1995, 425 pages. ISBN 3-922906-54-0.

In more recent years, another similar double-blind and controlled study was undertaken at the University of Vienna by Günter Hebenstreit, producing similar positive results. The Hebenstreit study, like that of Müschenich and Gebauer, used an orgone accumulator, a control dummy-box that looked like an accumulator, and a relaxation-chair in which test subjects were evaluated both before and after exposure to either the accumulator or control enclosure. It produced positive results, I am informed. Unfortunately, we do not have any printed protocols, but can give the exact citation for those who wish to obtain and read the original German-language publication:

Günter Hebenstreit, *Der Orgonakkumulator Nach Wilhelm Reich. Eine Experimentelle Untersuchung zur Spannungs-Ladungs-Formel*, Diplomarbeit zur Erlangung des Magistergrades der Philosophie an der Grung- und Integrativwissenschaftlichen Fakultät der Universität Wien, 1995.

Another German-language title of significance was published in 1998, containing the contributed research papers of

40 different scientific-professional authors on the subject of Wilhelm Reich's social and biophysical findings, to include many papers on the subject of bions, orgone energy, and the orgone energy accumulator:

James DeMeo & Bernd Senf, Editors, *Nach Reich: Neue Forschungen zur Orgonomie: Sexualökonomie, Die Entdeckung Der Orgonenergie (After Reich: New Research in Orgonomy: Sex-Economy, Discovery of the Orgone Energy)*, Zweitausendeins Verlag, Frankfurt, 1998. 898 pp. ISBN 3-86150-239-9.

An Orgone Energy Darkroom (room-sized accumulator) at the Orgone Biophysical Research Lab's high-altitude Greensprings Center near Ashland, Oregon. This structure is made in the form of a large accumulator, with a metal-lined interior and under-layering of fiberglass. Inside are several human-sized orgone accumulators and experimental chargers. The strength of the atmospheric orgone energy charge inside such an orgone room is dramatically increased. The orgone energy can be more easily felt and observed with the eye, and registers more clearly in experiments and with measuring apparatus.

About the Orgone Biophysical Research Lab (OBRL) and Greensprings Center

The Orgone Biophysical Research Lab is a non-profit science research and educational foundation, established in 1978. Over the years, OBRL has supported various laboratory and field research projects, educational lectures and seminars in both the USA and overseas, focused upon the sex-economic and orgone biophysical discoveries of the late, great natural scientist, Wilhelm Reich. OBRL also publishes a journal, *Pulse of the Planet*, on an irregular basis. The founder and director of OBRL, James DeMeo, has been investigating and extending Dr. Reich's original findings in both the social and natural sciences since 1971. Early in 1995, the *Greensprings Center* was established in a vibrant forested region of rural Oregon, to provide a better foundation for continued public outreach and in-depth study of orgone energy functions in nature. The *Center* also has the West Coast's only *Orgone Energy Darkroom*, for experiments requiring a higher energy charge, and for direct visual observation of orgone energy functions. New books (described at right) are also occasionally published, and public educational programs, the *Greensprings Seminars*, are held each summer. Accomplished scholars in the field of orgonomy are invited each year to share their knowledge and experience with interested students and professionals from around the world. Yearly seminars include: *Bions, Biogenesis and the Reich Blood Test, The Orgone Energy Accumulator*, and *Guided Independent Study*. Contact OBRL for an informative brochure and publications catalog, and to get on the international mailing list for seminar announcements.

Orgone Biophysical Research Lab
Greensprings Center, PO Box 1148
Ashland, Oregon 97520 USA
Tel/Fax: 541-552-0118 Email: demeo@mind.net
OBRL's Global Internet Web Site:
http://www.orgonelab.org

Additional Publications Available

(Contact OBRL for ordering information)

Heretic's Notebook:

Heretic's Notebook (Pulse of the Planet #5): Emotions, Protocells, Ether-Drift and Cosmic Life-Energy, with New Research Supporting Wilhelm Reich, Edited by James DeMeo

Contains 28 insightful essays and research articles by 17 different authors, on natural childbirth, sexuality, archaeology of early human violence, Reich's orgonomic functionalism, exposés on Reich's detractors, Giordano Bruno's work, bion-biogenesis research, Dayton Miller's ether-drift discoveries, emotional effects in REG (psychokinesis) experiments, new detector for orgone energy, dowsing research, cloudbusting desert-greening experiments in Africa, plant growth stimulation in the orgone accumulator, the orgone energy motor and "free energy", plus UFO research, book reviews, and much more, with color cover photos, text- photos and illustrations.
272 pages, Large format $24 Softcover

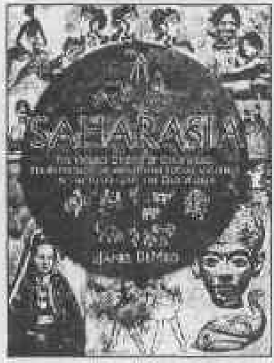

SAHARASIA: The 4000 BCE Origins of Child-Abuse, Sex-Repression, Warfare and Social Violence, In the Deserts of the Old World,
by James DeMeo

Dr. DeMeo's *magnum opus* on the origins of human violence and biophysical armoring, the first geographical, cross-cultural study of human behavior around the world, using Wilhelm Reich's sex-economic discoveries as a basic starting point, presenting world

maps of different behaviors and social institutions. Source-regions (Arabia and Central Asia) for patriarchal authoritarian culture were identified, and migratory-diffusion patterns were traced, back in time, to pinpoint where and how the human tragedy began. Solves the riddle of the origins of human violence and armoring. A breakthrough in the scientific study of human sexuality, psychology and anthropology, and must-reading for every parent, student, professor and clinical worker in the field of human health and behavior. 454 pages, with over 100 maps, photos, and illustrations. Large format with vivid full-color cover, extensive bibliography and index. Softcover $34
Author-Signed Limited Hardbound First Edition $90

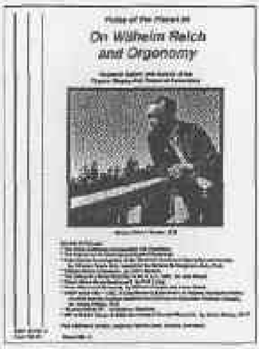

On Wilhelm Reich and Orgonomy (Pulse of the Planet #4)

Edited by James DeMeo

Contains Reich's milestone articles on psychic and somatic (mind-body) processes, and on the bioelectrical aspects of human emotion and sexuality, with articles by R.D. Laing on Reich, and a discussion on Reich's work in Denmark when he fled Nazi terror in Germany and was also expelled from the International Psychoanalytic Association. Other papers discuss: Reich's research on biogenesis and discovery of the microscopical *bion*; the discovery of orgone (life) energy, and the orgone energy accumulator. Also featured are articles about the Food & Drug Administration's attack upon Reich, and their present-day war against the natural health movement; the deadly effects from nuclear power plants, and an illuminating scientific challenge to the HIV theory of AIDS — plus other reports on current life energy research, weather anomalies from nuclear bomb tests, a cloudbusting desert-greening experiment in Israel, provocative book reviews, and more! 176 pp. Large format $20 Softcover.

PLANS

FOR
CONSTRUCTING

AN

ORGONE
ENERGY
ACCUMULATOR

Based on information originally distributed by the Wilhelm Reich Foundation

Notes Regarding the Use of the Orgone Accumulator

The Orgone Accumulator is built so as to collect the natural energy of the atmosphere (Orgone Energy) and to make this energy about five times more concentrated inside the accumulator than in the outside air. Since orgone is the same energy that is in the bodies of all living organisms, the accumulator has the effect of increasing the body energy of the user--increasing the natural charge of the user's tissue and blood--thus strengthening the user's resistance to disease.

Orgone is not electricity, and the orgone accumulator has no electrical connections. WARNING: DO NOT ALLOW EXPOSED ELECTRICAL WIRES TO COME IN CONTACT WITH THE ACCUMULATOR!

The accumulator should be used once or twice daily as follows: the user should sit inside the accumulator lightly clothed or without clothing until the user becomes quite warm, or for roughly 15 to 45 minutes. The effect of the orgone will be best if the user relaxes, breathing out deeply but not forcefully. Slight dizziness or heaviness in the head is a signal to come out. Fresh air will eliminate such effects immediately.

In humid weather, the normal time of use should be lengthened, as there is then less orgone in the atmosphere. Air the room in which the accumulator is kept every day, for the increased orgone concentration tends to make the air in the room heavy.

The funnel end from the shooter box-seat is used for local application of orgone. This unattached end should be held close to, but not quite touching, the region in question for between two and ten minutes, depending on the sensitivity of the region. If it is used on the eyes, they should be kept closed. A slight burning sensation in the region that is being treated is a signal that the body has had enough for the time being. In general, the tube should only be used until the area becomes warm. The shooter box-seat tube can be used several times a day.

© Copyright 1991 by Mary Boyd Higgins as Trustee of the Wilhelm Reich Infant Trust Fund

Questions Regarding the Use of the Orgone Accumulator

Q. Does it lessen the effect of the accumulator if any part of the body comes in direct contact with the metal walls?
A. It is better if no part of the body touches the metal walls. However this cannot always be avoided and it does not lessen the effect to any considerable degree.

Q. If two or three people use an accumulator rapidly in succession, is the orgone "used up"; does the last one to use it get energy?
A. The effect of the orgone accumulator consists of a mutual excitation of the orgone energy which is concentrated within the accumulator and the body-orgone energy of the user. It can be assumed that the successive use of the accumulator by several persons will not lessen the efficacy.

Q. Should you spend more or less time in the accumulator, depending on the humidity?
A. Since the orgone energy concentration in the atmosphere diminishes with high humidity, it is advisable to prolong the use of the accumulator on very humid days.

Q. Does the funnel for the pipe of the tube box have to be metal?
A. Yes, since plastic or other non-metallic material will tend to "hold" orgone energy, while the metal will give it off immediately.

Q. Can you use the tube box inside a larger accumulator for reasons of more intensity, or is it more advisable to use it outside?
A. The concentration of orgone energy in the tube box will be stronger if it is used within the larger accumulator.

Q. Is it all right to take objects into the accumulator while using it, such as books, magazines, notebook, pen, pillow?
A. There is no objection to taking reading or writing material into the accumulator or knitting or needlework, etc. We would advise against taking pillows into the accumulator because they will soak up orgone energy and thus divert it from the body.

Q. Is it all right to paint the accumulator to match walls?
A. As long as the paint does not contain lead or other metals, the accumulator can be painted in any color. It may also be covered with wallpaper.

Q. If one makes one's accumulator light-tight and then sits inside it, is it possible to see the orgone energy light phenomena that are described in The Cancer Biopathy?

A. No; bodily evaporation under these conditions creates too much humidity and the orgone phenomena cannot be seen when the humidity is high.

Q. What is the difference in strength between a one layer and three layer accumulator?

A. The one layer accumulator, has an orgone concentration, measured electroscopically, of about 3 times that of the surrounding atmosphere. The three layer accumulator, has an orgone concentration of about 4 to 5 times that of the surrounding atmosphere. However, the concentration of the orgone energy does not increase proportionally with the numbers of layers. Thus a tenfold accumulator will have a concentration of approximately 6 times that of the surroundings.

Q. My accumulator does not fit tightly together. Does this in any way lessen the effectiveness?

A. Occasionally when an accumulator is dismantled and put together again, the walls do not close tightly. These cracks, even if they are as large as one inch, do not in any way affect the strength of the accumulator. In fact, experimental accumulators have been built with openings at the top and bottom for circulation purposes.

Q. Can paper, felt or material other than fiberglass be used for the non-metallic layers of the accumulator?

A. Experiments have been made with other materials. However, all these materials are liable to absorb humidity, whereas the fiberglass will not absorb any humidity. Too strong an absorption of humidity would definitely lower the strength of the accumulator.

Q. If you go into the accumulator with a fever, is the fever increased?

A. In most cases, experience shows that the fever will be slightly increased while the person uses the accumulator, but that the temperature will drop below the previous level after the use of the accumulator.

Q. Is there any special time of day when the accumulator is more effective than at other times?

A. Electroscopic measurements show the greatest concentration of orgone energy between 12 noon and 2 p.m. However, these variations in a given day are so small that it will not make any significant difference at what time of the day the accumulator is used. It can easily be regulated by timing.

Construction of a Three-Fold Orgone Energy Accumulator

1. General information

a. The accumulator is made of six panels which are to be screwed together. All panels except the bottom are constructed in the same manner, and differ only in dimensions. Each panel consists of an inner surface of galvanized sheetmetal and an outer non-metallic surface of celotex which encloses a braced wood frame and alternate layers of fiberglass and steelwool.

b. The materials specified may, if necessary, be replaced by other materials: 1/4" plexiglass or 1/4" upsonite along with a 1/4" fiberglass layer may be used in place of the 1/2" celotex. If substitutions are made, some adjustments in dimensions of the frames may be necessary.

c. Consult the accompanying drawings and tables for dimensions, construction details, etc.

2. Construct the frames

a. Cut the 1 1/8" x 1 1/2" clear pine to the specified lengths (see fig. 1). Cut a 1/2" x 1/2" rabbet into one side of each piece (see fig. 2). The bottom panel frame requires two 1/4" x 1/2" rabbets (see fig. 3). Mitre the corners and join with carpenters' glue and 8-penny finish nails.

b. Brace each frame with two pieces of 1 1/8" x 1" clear pine (see fig. 4). Join with two 2 1/2" #8 drywall screws. The top and bottom frame do not need bracing.

3. Attach the outer surface

a. Cut the celotex to fit inside the rabbets of each frame except for the bottom frame. Fasten in place with 1" galvanized wire nails.

---

1. For a five-fold accumulator, increase the dimensions of the frames to allow for four more 1/4" layers, two layers of fiberglass and two layers of steelwool.

2. The 1/2" celotex (which has one side coated with white primer paint) may be known as: white construction board, soundboard, etc. It is basically a composite of wood fibers. Flakeboard, chipboard, etc. are not the same product and are too dense. There is a styrofoam product with the trademark name Celotex; do not use this.

3. fiberglass building insulation

4. What is referred to as 5/4" stock by the building industry actually measures 1 1/8 inch.

b. Cut two pieces of 1/4" finish grade plywood to fit inside the rabbets of the bottom frame. Fasten one piece in place with 1" galvanized wire nails.

4. Place the steelwool and fiberglass in the panels

a. Place a layer of steelwool upon the celotex inside one of the frames. Steelwool pads (oo fine or finer) when unrolled are the correct thickness. Make the layer as uniform as possible; leave the steel wool fluffy.

b. Next place a layer of fiberglass about 1/4" thick upon the steelwool layer. Avoid lumps and holes. Do not compress the fiberglass. The standard 3 1/2" x 15" fiberglass insulation comes in a roll with either a paper or foil backing. Peel off the paper or foil. This type of fiberglass is difficult to peel evenly. Try to peel it carefully into roughly 1/4" thick pieces. Too much unevenness in the layering will cause the sides of the accumulator to bulge. (There is a fiberglass insulation that peels more uniformly but might be difficult to purchase. It is a Manville product called 1/2" duct liner. It comes in a roll 48" wide by 100 feet long and may be available at your local heating and airconditioning shop.)

c. In a similar manner place the remaining alternate layers of steelwool and fiberglass in position (see fig 2).

d. Place the steelwool and fiberglass layers in the other panels.

e. The bottom panel has a different number of layers. Start with a layer of fiberglass upon the 1/4" plywood inside the frame (see fig. 3).

5. Attach the inner surfaces

a. Cut the sheetmetal slightly smaller than the frames. You might need to have the 26 or 28 gage galvanized sheetmetal cut at a sheetmetal shop. It can be cut by hand, but the results may be too rough. File the edges smooth and round the corners. Cut a 9" x 12" hole for the window in one of the 53" x 24" sheets. Figure 4 will give the exact location.

b. Drill 1/8" holes at 1" intervals around the edge of the sheetmetal. They should be 1/4" from the edge.

c. Nail the pieces of sheetmetal to the frames with 3/4" galvanized wire nails. Nail the remaining piece of 1/4" plywood to the bottom frame before nailing the final piece of sheetmetal over it. (See fig. 3)

---

5. Wear protective clothing and a dust mask when handling the steelwool and fiberglass.

6. Steelwool in long rolls can be purchased from an industrial supplier.

6. Attach supports to the bottom panel

a. Cut two pieces of 1 1/8" x 1 1/2" clear pine 25" long.

b. Screw them into the outer surface of the panel. (See fig 3.)

c. Nail two furniture slides to the undersurface of each support.

7. Assemble the accumulator

Accuracy is necessary in fastening the sides, top, bottom and door of the accumulator. You will want to be able to take apart and put back together your accumulator in the event that you need to transport it. It is very bulky when assembled.

a. Pre-drill holes for the 2 1/2" #10 flathead philips screws with a 3/16" drill bit. So that the screw heads will countersink, drill a 1/2" hole 1/4" deep. Figure 5 gives the locations of these holes.

b. Lay the two side panels upright on their long sides. Place the bottom panel upright against the bottom ends of the two side panels and screw together. Then place the top panel upright against the top ends of the two side panels and screw together. Place the back panel onto the side panels between the overlaps of the top and bottom panels and screw together (figure 5).

c. Turn the box over onto its back and place the door on the side panels. Make sure that the door is flush with the top of the box (this will leave a one-inch gap at the bottom to allow air circulation). Mount the hinges to the side panel and to the door (fig. 6). You can hang the door to open either left or right.

d. Now stand the accumulator up.

e. You will need a hook and eye to hold the door closed while you sit in the accumulator. Place it where most convenient.

8. Assemble the seat for the accumulator

You have two options. The simplest is to suspend a piece of 3/4" x 14" x 22" wood on two supports fastened to the lower support braces of the two side panels. The other option is to turn the seat into a combination shooter box and chest panel.

Construct the shooter box-seat

a. Construct two panels in a manner similar to the other panels as explained in sections 2 through 6. 1/4" finish grade plywood is needed for the top and front of the seat panels (fig. 7).

b. Prepare supports to hold the shooter box-seat panels. Cut four pieces of 1 1/8" x 1 1/2" pine 12 1/2" long. Fasten two 12 1/2" pieces to the lower supports in the side panels with 2" #8 drywall screws. Then fasten the other two pieces vertically between the horizontal support and the floor of the accumulator with 2" screws. Fasten one 14" x 21 3/4" panel to the horizontal cleat A with 2 1/2" drywall screws. The other 14" x 21 3/4" panel will rest against the vertical cleats B (fig. 8). Do not fasten this panel. Attach a cabinet pull (metal or ceramic) to the front of this panel so you can easily remove it to use as a chest panel (fig. 9).

9. Construct a shooter funnel for the box seat

a. Take a four-foot length of 3/8" steel Greenfield casing. (Greenfield casing can be purchased from an electrical supply store. If you can't find it, you can substitute #12 BX electrical cable, but remove the wiring from the flexible metal casing before using.)

b. Insert the spout of a 3" diameter galvanized sheetmetal funnel into one end of the flexible metal casing. Wrap carefully the outside of both the flexible casing and the funnel with plastic electricians tape, so that no metal is showing. Drill a 5/8" hole through the top back corner of the accumulator box seat's top panel. Push the flexible tubing into the space below.

10. Optional

In order to draw stagnant orgone energy from the accumulator and shooter box seat:

a. You will need approximately three 48" lengths of 3/8" Greenfield casing. Use more or less than three depending on individual circumstances.

b. Drill three 5/8" holes (in a vertical row) through the wood frame of one of the side panels. Make sure that one hole will lead into the shooter box seat. Insert the three casings into the three holes. Make sure that the casings don't protrude inside of the accumulator.

---

7. Greenfield casing is a convenient hollow flexible metal tube. IT IS NOT an electrical connection and should NOT be used as one.

c. Place the loose ends of the three casings into a pail of clean water. You need to change the water every week.

#######

The principle of the accumulator box can be adapted to different shapes for specific needs. An orgone blanket, or a collar for the neck, or a large funnel for the breast can be made as long as the principle of alternating layers of non-metallic and metallic materials is maintained.

---

8. The inner surface should be made of a flexible metal screen instead of galvanized sheetmetal. Screen made from galvanized steel or stainless steel is good. Screen made from copper, aluminum and other metals that are not body-own are NOT recommended.

MATERIALS LIST

ITEM	APPROXIMATE QUANTITY
1/2" celotex	four 4' x 8' sheets
galvanized sheetmetal (26 or 28 gage)	2 pieces 53" x 26 1/2"
" " " "	2 pieces 53" x 24"
" " " "	2 pieces 28" x 24"
fiberglass insulation	3 1/2" x 15"(1 roll or 40 linear feet)
steelwool (oo fine or finer)	200 pads or 10 lbs.
1/4" plywood (finish grade)	one 4' x 8' sheet
clear white pine: 1 1/8" x 1 1/2"	80 linear feet
" " " 1 1/8 x 1"	20 linear feet
2 1/2" #10 philips flathead wood screws	30 pcs.
2 1/2 #8 drywall screws	30 pcs.
2" #8 drywall screws	10 pcs.
3" non-mortise or surface mounted hinges	one pair
hook and eye	one 2"
nail-on furniture glides	4 pcs.
nails: 3/4" galvanized wire nails	3 oz.
1" galvanized wire nails	3 oz.
6-penny finish nails	2 lbs.
carpenters' glue	one small bottle
For the shooter box seat panels:	
galvanized sheet metal (26 or 28 gage)	2 pieces 13" x 20 3/4"
clear white pine: 1 1/8" x 1 1/2"	16 linear feet
3" diameter galvanized sheetmetal funnel	one
3/8" Greenfield flexible metal casing	16 feet
electricians' tape (plastic tape)	one roll

(FIG. 1) TABLE OF DIMENSIONS

PANEL	LENGTH	WIDTH
TOP	29"	25"
BOTTOM	29"	25"
SIDES (2)	54"	27 1/2"
BACK	54"	25"
DOOR	54"	25"
SEATS (2)	14"	21 3/4"

(FIG. 2) DETAIL OF LAYERING OF SIDES (2) TOP, BACK AND DOOR PANELS

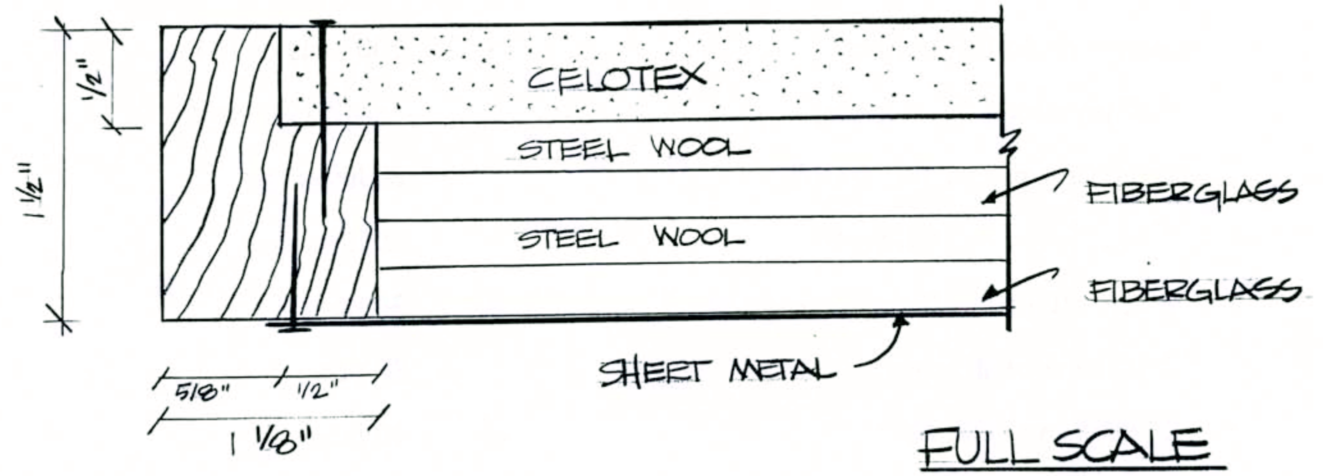
1/2"
1 1/2"
CELOTEX
STEEL WOOL
FIBERGLASS
STEEL WOOL
FIBERGLASS
SHEET METAL
5/8"
1/2"
1 1/8"

FULL SCALE

(FIG. 3) DETAIL OF LAYERING - BOTTOM PANEL

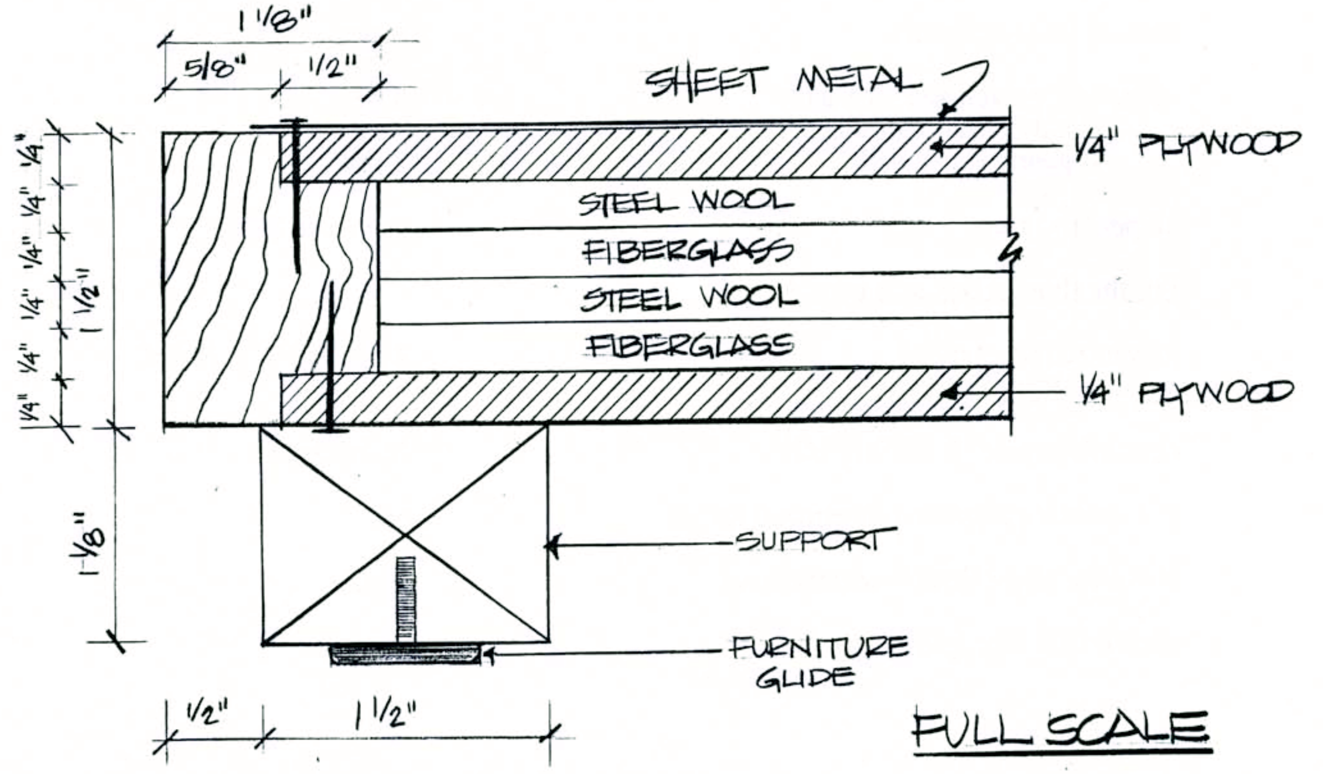
1 1/8"
5/8"
1/2"
SHEET METAL
1/4" PLYWOOD
STEEL WOOL
FIBERGLASS
STEEL WOOL
FIBERGLASS
1/4" PLYWOOD
1/4"
1/4"
1/4"
1/4"
1/4"
1/4"
1 1/2"
1-1/8"
SUPPORT
FURNITURE GLIDE
1/2"
1 1/2"

FULL SCALE

(FIG. 4) FRAMING DETAILS - SIDES (2), BACK, FRONT (DOOR), BOTTOM, TOP & SHOOTER BOX - SEAT (2)

SIDE PANEL FRAME (2)

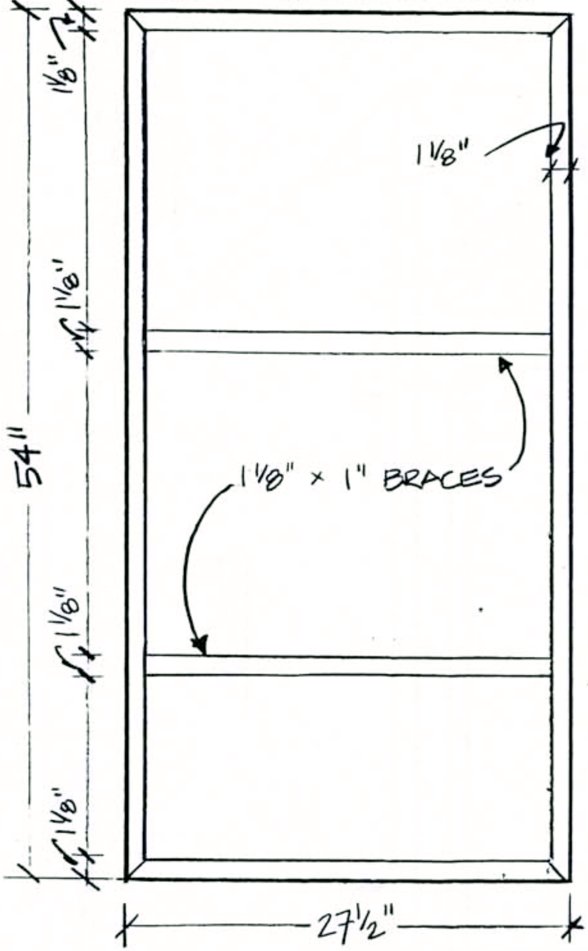
54"
1⅛"
1⅛"
1⅛"
1⅛"
1⅛"
1⅛" × 1" BRACES
27½"

BACK PANEL FRAME

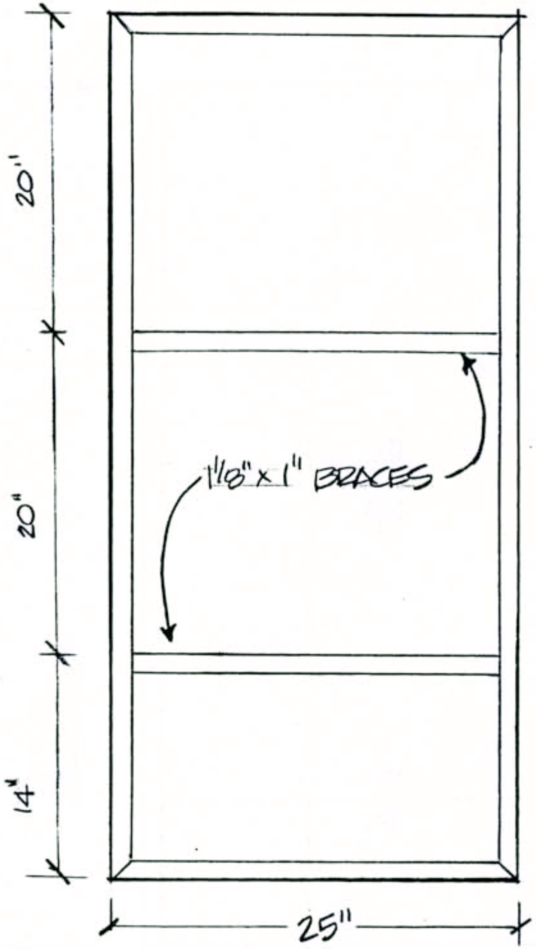
20"
20"
14"
1⅛" × 1" BRACES
25"

FRONT (DOOR) PANEL FRAME

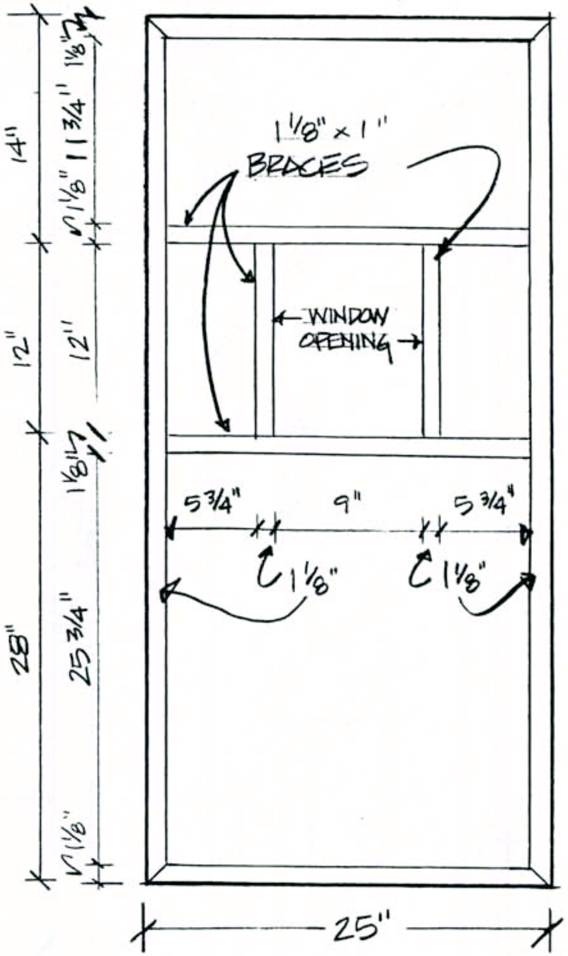
14"
12"
28"
1⅛"
11 3/4"
1⅛"
12"
1⅛"
25 3/4"
1⅛"
1⅛" × 1"
BRACES
WINDOW
OPENING
5 3/4"
9"
5 3/4"
1⅛"
1⅛"
25"

SIDES (2), BACK & FRONT (DOOR) PANEL FRAMES TO BE 'FASHIONED' FROM 1⅛" × 1½" PINE WITH ½" × ½" RABBET (SEE FIG. 2)

BRACES TO BE 1⅛" × 1" PINE.

SCALE: 1" = 1'-0"

NOTE: TOP & BOTTOM FRAMES HAVE SAME WIDTH, LENGTH AND THICKNESS.

BOTTOM FRAME CONSTRUCTED FROM 1⅛" × 1½" PINE WITH A ¼" × ½" RABBET ON TWO SIDES (SEE FIG. 3)

TOP FRAME CONSTRUCTED FROM 1⅛" × 1½" PINE WITH A ½" RABBET ON ONE SIDE. (SEE FIG. 2)

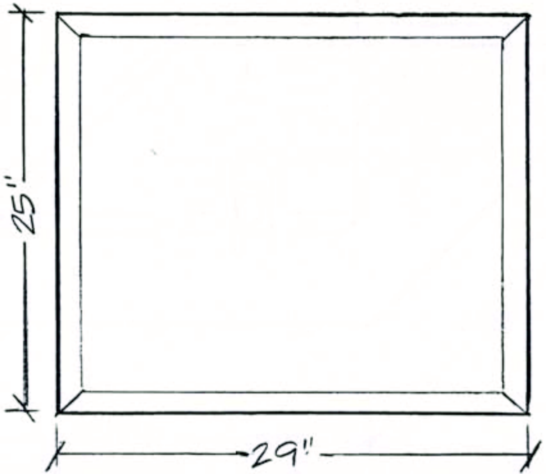

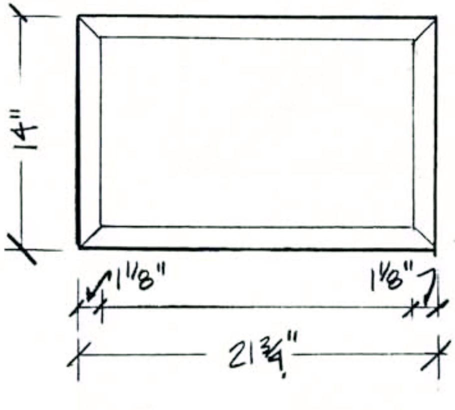

SHOOTER BOX SEAT (2)

FRAME TO BE CONSTRUCTED FROM 1⅛" × 1½" PINE WITH A ¼" × ½" RABBET (SEE FIG. 7)

(FIG. 5)
ASSEMBLING THE PANELS

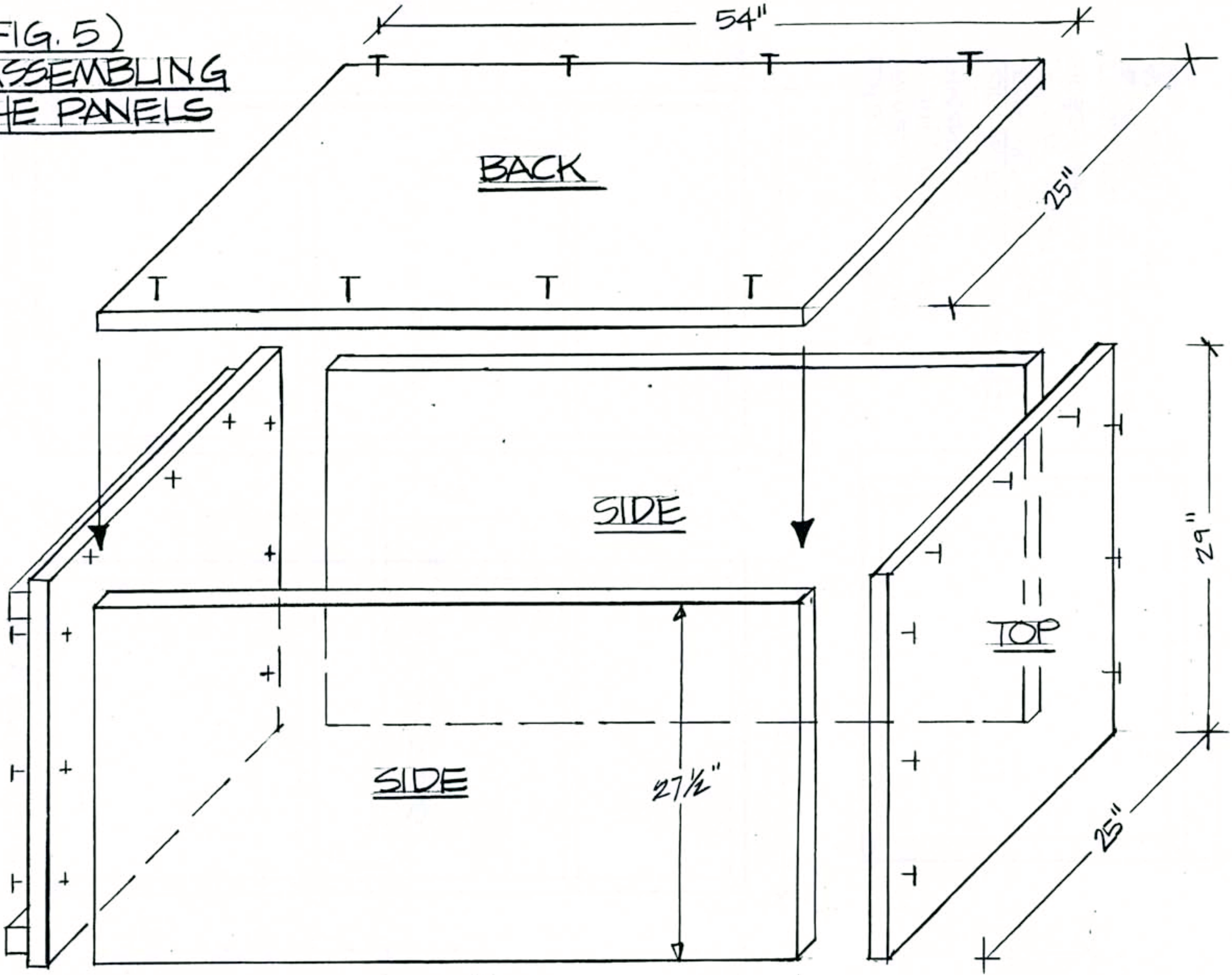
54"
BACK
25"
SIDE
29"
TOP
SIDE
27½"
25"
54"

T=2½" SCREWS

(FIG. 6)
ATTACH DOOR

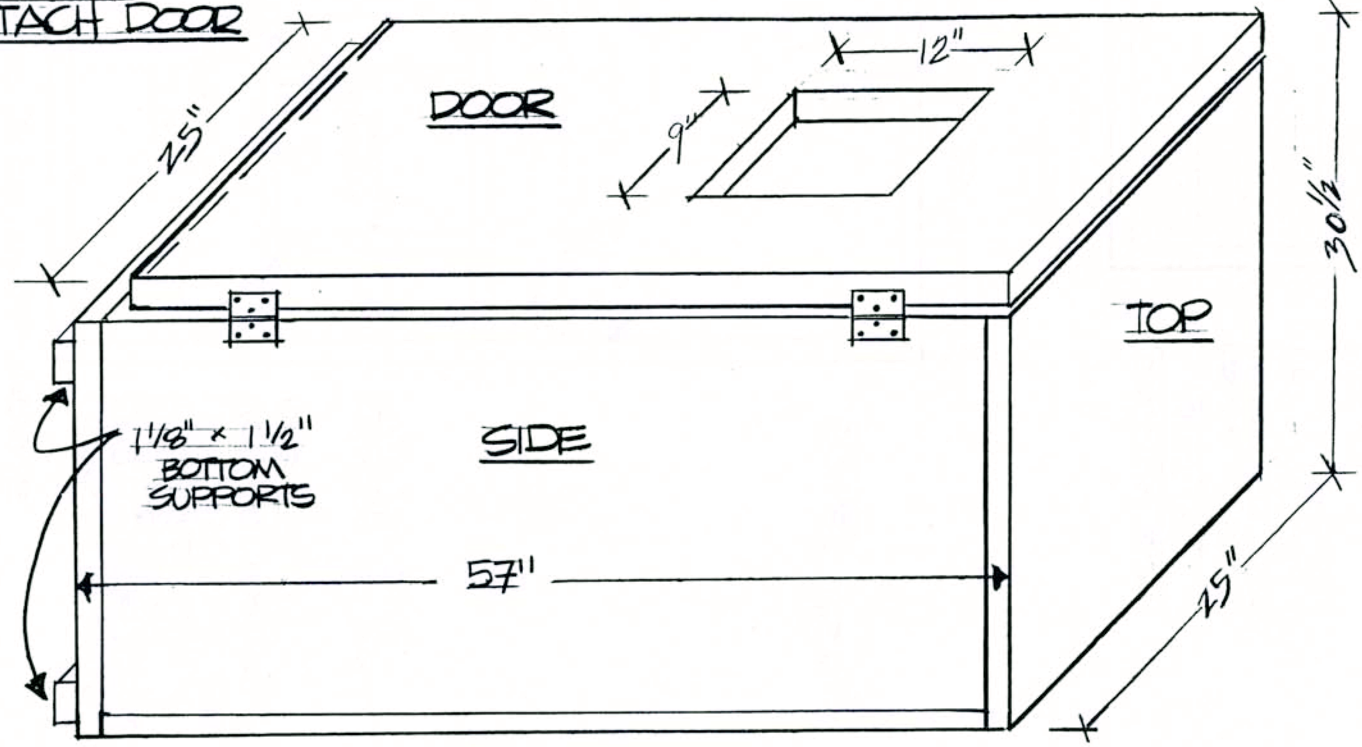
12"
DOOR
9"
25"
30½"
TOP
1⅛" × 1½"
BOTTOM SUPPORTS
SIDE
57"
25"

(FIG. 7) DETAIL FOR SHOOTER BOX - SEAT PANELS

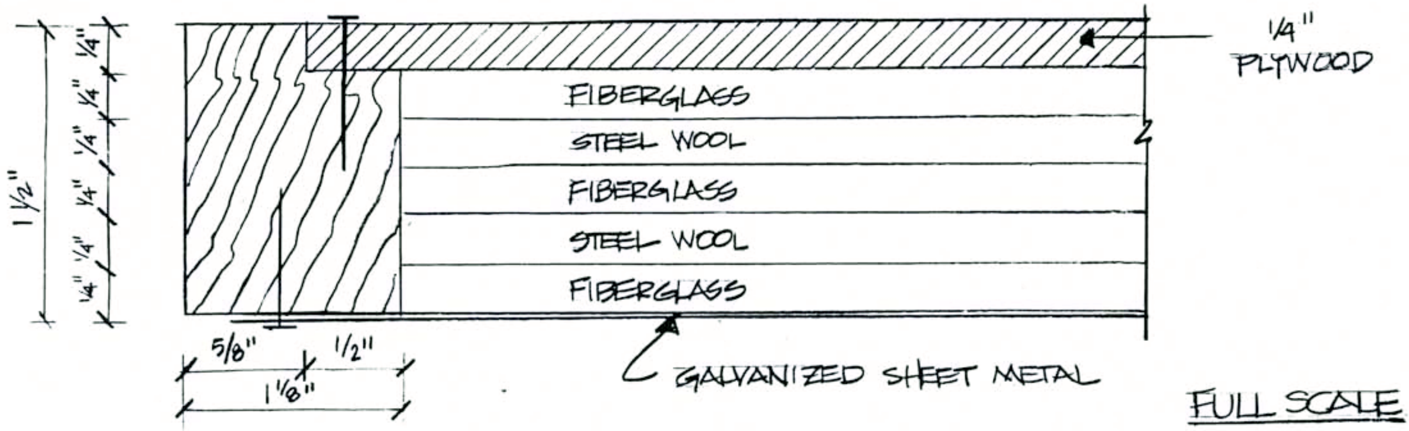

1/4" PLYWOOD
FIBERGLASS
STEEL WOOL
FIBERGLASS
STEEL WOOL
FIBERGLASS
GALVANIZED SHEET METAL
1 1/2"
1/4"
1/4"
1/4"
1/4"
1/4"
1/4"
5/8"
1/2"
1 1/8"
FULL SCALE

(FIG. 8) ASSEMBLING THE SHOOTER BOX - SEAT

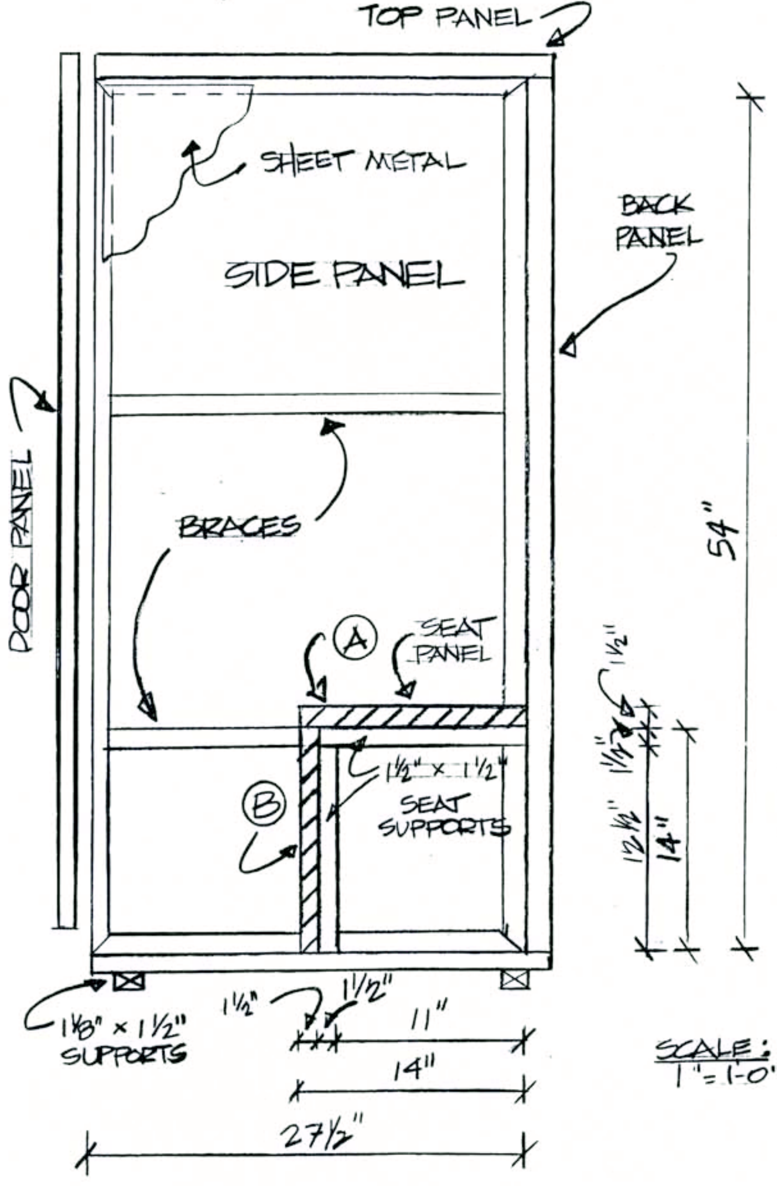

TOP PANEL
SHEET METAL
SIDE PANEL
BACK PANEL
DOOR PANEL
BRACES
Ⓐ
SEAT PANEL
Ⓑ
1 1/2" x 1 1/2" SEAT SUPPORTS
1 1/8" x 1 1/2" SUPPORTS
1 1/2"
1 1/2"
11"
14"
27 1/2"
1 1/2"
1 1/2"
12 1/2"
14"
54"
SCALE: 1" = 1'-0"

(FIG 9) HOW TO USE CHEST PANEL

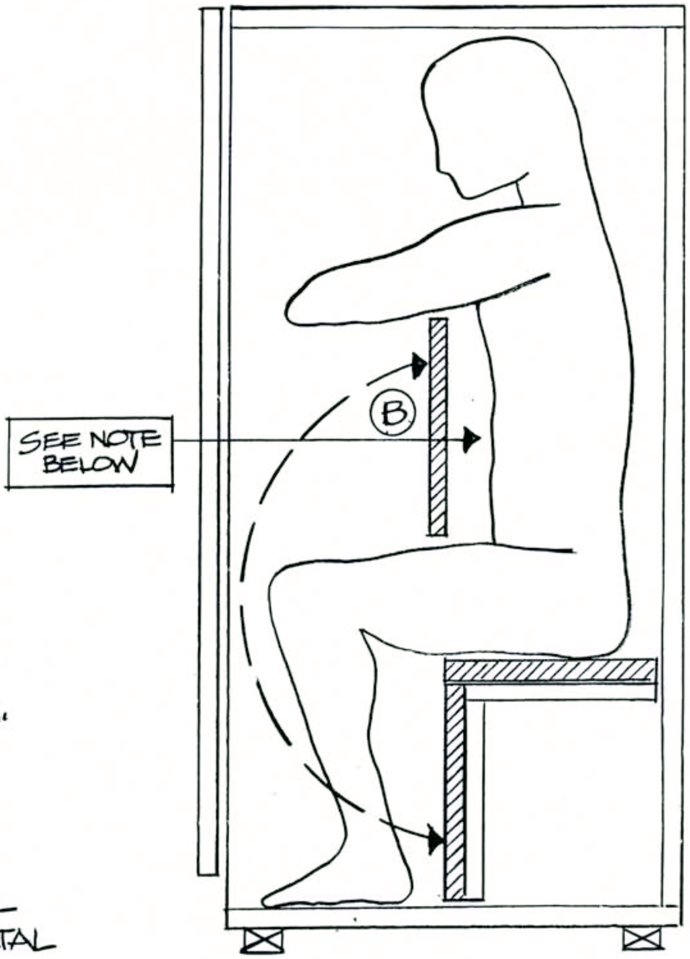

SEE NOTE BELOW
Ⓑ

SCALE: 1" = 1'-0"

NOTE: ALLOW SPACE BETWEEN PANEL AND CHEST OF USER. SHEET METAL SIDE SHOULD FACE CHEST.

ENCYCLOPEDIA
OF
MODERN OPTICS

EDITED BY
B.D. GUENTHER
A. MILLER
L. BAYVEL
J.E. MIDWINTER

VOLUME ONE

EDITOR-IN-CHIEF

Robert D. Guenther
Duke University
Durham, NC, USA

EDITORS

Duncan G. Steel
University of Michigan
Ann Arbor, MI, USA

Leopold Bayvel
London, UK

CONSULTING EDITORS

John E. Midwinter
University College London, London, UK

Alan Miller
University of St Andrews, Fife, UK

EDITORIAL ADVISORY BOARD

R Alferness
Lucent Technologies Bell Laboratories
Holmdel, NJ, USA

D J Brady
Duke University
Durham, NC, USA

J Caulfield
Diversified Research Corporation
Cornersville, TN, USA

R L Donofrio
Display Device Consultants LLC
Ann Arbor, MI, USA

J Dudley
Université de Franche-Comté
Besançon, France

J G Eden
University of Illinois
Urbana, IL, USA

H O Everitt
U.S. Army Research Office
Research Triangle Park, NC, USA

E L Hecht
Adelphi University
Garden City, NY, USA

H S Hinton
Utah State University
Logan, UT, USA

T N Hurst
Hewlett-Packard Laboratories
Tucson, AZ, USA

D Killinger
University of South Florida
Tampa, FL, USA

T A King
University of Manchester
UK

E Leith
University of Michigan
Ann Arbor, MI, USA

A Macleod
Thin Film Center, Inc.
Tucson, AZ, USA

P F McManamon
Air Force Research Laboratory
Wright-Paterson AFB, OH, USA

J N Mait
U.S. Army Research Laboratory
Adelphi, MD, USA

D Phillips
Imperial College
London, UK

C R Pidgeon
Heriot-Watt University
Edinburgh, UK

R Pike
King's College London
UK

E Van Stryland
University of Central Florida
Orlando, FL, USA

Preface

We live in a world powered by light, but much of the understanding of light was developed around the time of the French Revolution. Before the 1950s, optics technology was viewed as well established and there was little expectation of growth either in scientific understanding or in technological exploitation. The end of the Second World War brought about huge growth in scientific exploration, and the field of optics benefited from that growth. The key event was the discovery of methods for producing a source of coherent radiation, with the key milestone being the demonstration of the first laser by Ted Maiman in 1960. Other lasers, nonlinear optical phenomena, and technologies such as holography and optical signal processing, followed in the early 1960s. In the 1970s the foundations of fiber optical communications were laid, with the development of low-loss glass fibers and sources that operated in the wavelength region of low loss. The 1980s saw the most significant technological accomplishment: the development of efficient optical systems and resulting useful devices. Now, some forty years after the demonstration of the first coherent light source, we find that optics has become the enabling technology in areas such as:

- information technology and telecommunications;
- health care and the life sciences;
- sensing, lighting, and energy;
- manufacturing;
- national defense;
- manufacturing of precision optical components and systems; and
- optics research and education.

We find ourselves depending on CDs for data storage, on digital cameras and printers to produce our family photographs, on high speed internet connections based on optical fibers, on optical based DNA sequencing systems; our physicians are making use of new therapies and diagnostic techniques founded on optics.

To contribute to such a wide range of applications requires a truly multidisciplinary effort drawing together knowledge spanning many of the traditional academic boundaries. To exploit the accomplishments of the past forty years and to enable a revolution in world fiber-optic communications, new modalities in the practice of medicine, a more effective national defense, exploration of the frontiers of science, and much more, a resource to provide access to the foundations of this field is needed. The purpose of this Encyclopedia is to provide a resource for introducing optical fundamentals and technologies to the general technical audience for whom optics is a key capability in exploring their field of interest.

Some 25 internationally recognized scientists and engineers served as editors. They helped in selecting the topical coverage and choosing the over 260 authors who prepared the individual articles. The authors form an international group who are expert in their discipline and come from every part of the technological community spanning academia, government and industry. The editors and authors of this Encyclopedia hope that the reader finds in these pages the information needed to provide guidance in exploring and utilizing optics.

As Editor-in-Chief I would like to thank all of the topical editors, authors and the staff of Elsevier for each of their contributions. Special thanks should go Dr Martin Ruck of Elsevier who provided not only organizational skills but also technological knowledge which allowed all of the numerous loose ends to be tied.

B D Guenther
Editor-in-Chief

ALL-OPTICAL SIGNAL REGENERATION

O Leclerc, Alcatel Research & Innovation,
Marcoussis, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The breakthrough of optical amplification, combined with the techniques of wavelength division multiplexing (WDM) and dispersion management, have made it possible to exploit a sizeable fraction of the optical fiber bandwidth (several terahertz). Systems based on 10 Gbit/s per channel bit-rate and showing capacities of several terabit/s, with transmission capabilities of hundreds or even thousands of kilometers, have reached the commercial area.

While greater capacities and spectral efficiencies are likely to be reached with current technologies, there is potential economic interest in reducing the number of wavelength channels by increasing the channel rate (e.g., 40 Gbit/s). However, such fourfold increase in the channel bit-rate clearly results in a significant increase in propagation impairments, stemming from the combined effects of noise accumulation, fiber dispersion, fiber nonlinearities, and inter-channel interactions and contributing to two main forms of signal degradation. The first one is related to the amplitude domain; power levels of marks and spaces can suffer from random deviations arising from interaction between signal and amplified spontaneous emission (ASE) noise or with signals from other channels through cross-phase modulation (XPM) from distortions induced by chromatic dispersion. The second type of signal degradations occurs in the time domain; time position of pulses can also suffer from random deviations arising from interactions between signal and ASE noise through fiber dispersion. Preservation of high power contrast between '1' and '0', and of both amplitude fluctuations and timing jitter below some acceptable levels, are mandatory for high transmission quality, evaluated through bit-error-rate (BER) measurements or estimated by Q-factors. Moreover, in future optical networks, it appears mandatory to ensure similar but high optical signal quality at the output of whatever nodes in the networks, as to enable successful transmission of the data over arbitrary distance.

Among possible solutions to overcome such systems limitations is the implementation of Optical Signal Regeneration, either in-line for long-haul transmission applications or at the output of network nodes. Such Signal Regeneration performs, or should be able to perform, three basic signal-processing functions that are Re-amplifying, Re-shaping, and Re-timing, hence the generic acronym '3R' (Figure 1). When Re-timing is absent, one usually refers to the regenerator as '2R' device, which has only re-amplifying and re-shaping capabilities. Thus, full 3R regeneration with retiming capability requires clock extraction.

Given system impairments after some transmission distance, two solutions remain for extending the actual reach of an optical transmission system or the scalability of an optical network. The first consists in segmenting the system into independent trunks, with full electronic repeater/transceivers at interfaces (we shall refer to this as 'opto-electronic regeneration' or O/E Regeneration forthwith). The second solution, i.e., all-optical Regeneration, is not the optical version of the first which would have higher bandwidth capability but still performs the same signal-restoring functions with far reduced complexity. At this point, it should be noted that Optical 3R techniques are not necessarily void of any electronic functions (e.g., when using electronic clock recovery and O/E modulation), but the main feature is that these electronic functions are narrowband (as opposed to broadband in the case of electronic regeneration).

Some key issues have to be considered when comparing such Signal Regeneration approaches. The first is that today's and future optical transmission systems or/and networks are WDM networks.

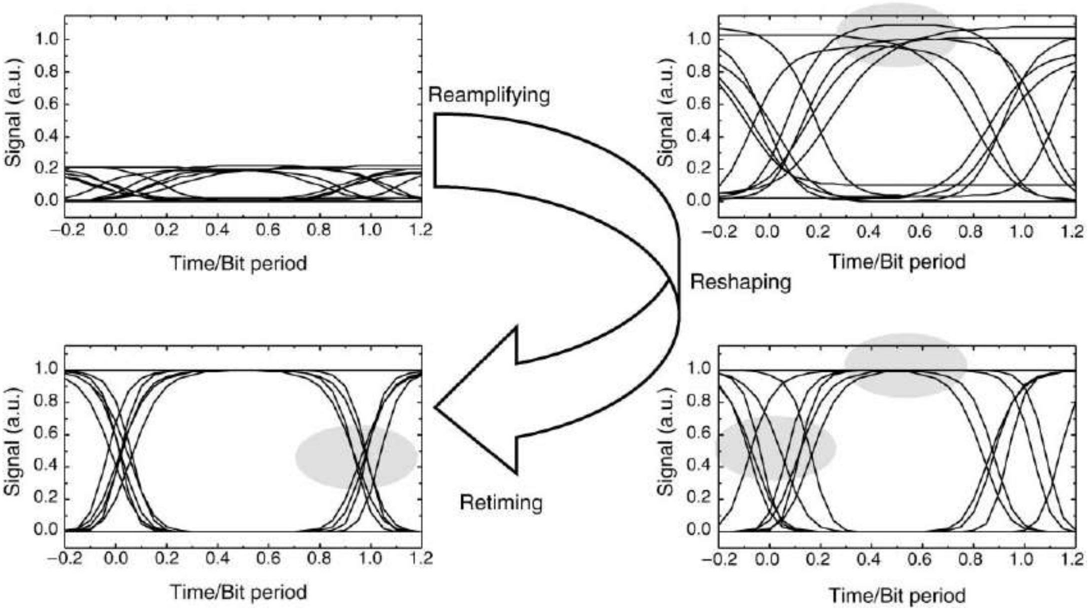

Figure 1 Principle of 3R regeneration, as applied to NRZ signals. (a) Re-Amplifying; (b) Re-Shaping; (c) Re-Timing. NB: Such eye diagrams can be either optical or electrical eye diagrams.

Under this condition, the WDM compatibility – or the fact that any Regeneration solution can simultaneously process several WDM channels – represents a key advantage. The maturity of the technology – either purely optical or opto-electronic – also plays an important role in the potential (pre-)development of such solutions. But the main parameter that will decide the actual technology (and also technique) relies on the tradeoff between actual performance of the regeneration solutions and their costs (device and implementation), depending on the targeted applications (long-haul system, medium haul transport, wide area optical network, etc.).

In this article, we review the current alternatives for all-optical Signal Regeneration, considering both theoretical and experimental performance and practical implementation issues. Key advantages and possible drawbacks of each solutions are discussed, to sketch the picture in this field. However, first we must focus on some generalities about Signal Regeneration and the way to define (and qualify) such regenerator performance. In a second part, we will detail the currently-investigated optical solutions for Signal Regeneration with a specific highlight for semiconductor-based solutions using either semiconductor optical amplifiers (SOA) technology or newly-developed saturable absorbers. Optical Regeneration techniques based on synchronous modulation will also be discussed in a third section. The conclusion will summarize the key features of each solution, so as to underline the demanding challenge optical components are facing in this application.

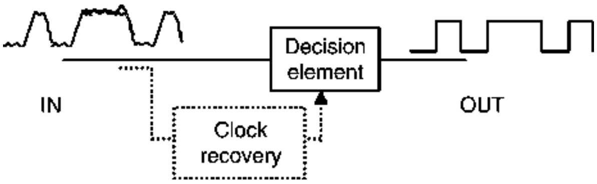

Figure 2 Generic structure of Signal 2R/3R Regenerator based on Decision Element (2R) and Decision Element and Clock Recovery (3R).

Generalities on Signal Regeneration

Principles

In the general case, Signal Regeneration is performed using a decision element exhibiting a nonlinear transfer function. Provided with a threshold level and when associated with an amplifier, such an element then performs the actual Re-shaping of the incoming data (either in the electrical or optical domain) and completes a 2R Signal Regenerator. Figure 2 shows the generic structure of such a Signal Regenerator in the general case as applied to non return to zero (NRZ) data. A clock recovery block can be added (dotted lines) to provide the decision element with time reference and hence perform the third R (Re-timing) of full Signal 3R Regeneration. At this point, it should be mentioned that the decision element can operate either on electrical signals (standard electrical DFF) provided that optical → electrical and electrical → optical signal conversion stages are added or directed onto optical signals using the different techniques described below. The clock signal can be of an electrical nature, as for electrical decision element in O/E regenerator – or either an electrical or a purely optical signal in all-optical regenerators.

Prior to reviewing and describing the various current technology alternatives for such Optical Signal Regeneration, the issue of the actual characterization of regenerator performance needs to be explained and clarified. As previously mentioned, the core element of any Signal Regenerator is the decision element showing a nonlinear transfer function that can be of varying steepness. As will be seen in Figure 3, the actual regenerative performance of the regenerator will indeed depend upon the degree of nonlinearity of the decision element transfer function.

Figure 3 shows the principle of operation of a regenerator incorporating a decision element with two steepnesses of the nonlinear transfer function. In any case, the '1' and '0' symbols amplitude probability densities (PD) are squeezed after passing through the decision element. However, depending upon the addition of a clock reference for triggering the decision element, the symbol arrival time PD will be also squeezed (clocked decision = 3R regeneration) or enlarged (no clock reference = 2R regeneration) resulting in conversion of amplitude fluctuations to time position fluctuations.

As for system performance – expressed through BER – regenerative capabilities of any regenerator simultaneously depend upon both the output amplitude and arrival time PD of the '1' and '0' symbols. In the unusual case of 2R regeneration (no clocked decision), a tradeoff has then to be derived, considering both the reduction of amplitude PD and the enlarged arrival time PD induced by the regenerator, to ensure sufficient signal improvement. In Figure 3a, we consider a step function for the transfer function of the decision element. In this case, amplitude PD are squeezed to Dirac PD after the decision element, and depending upon addition or not of a clock reference, arrival time PD is reduced (3R) or dramatically enlarged (2R). In Figure 3b, the decision element exhibits a moderately nonlinear transfer function. This results in an asymmetric and less-pronounced squeezing of the amplitude PD compared to the previous case, but in turn results in a significantly less enlarged arrival time PD when no clock reference is added (2R regeneration). Comparison of these two decision element of different nonlinear transfer function indicates that for 3R regeneration applications, the more nonlinear the transfer function of the decision element the better performance, the ideal case being the step function. In the case of 2R regeneration applications, a tradeoff between the actual reduction of the amplitude PD and enlargement of timing PD is to be derived and clearly depends upon the actual shape of the nonlinear transfer function of the decision element.

Qualification of Signal Regenerator Performance

To further illustrate the impact of the actual shape of the nonlinear transfer function of the decision

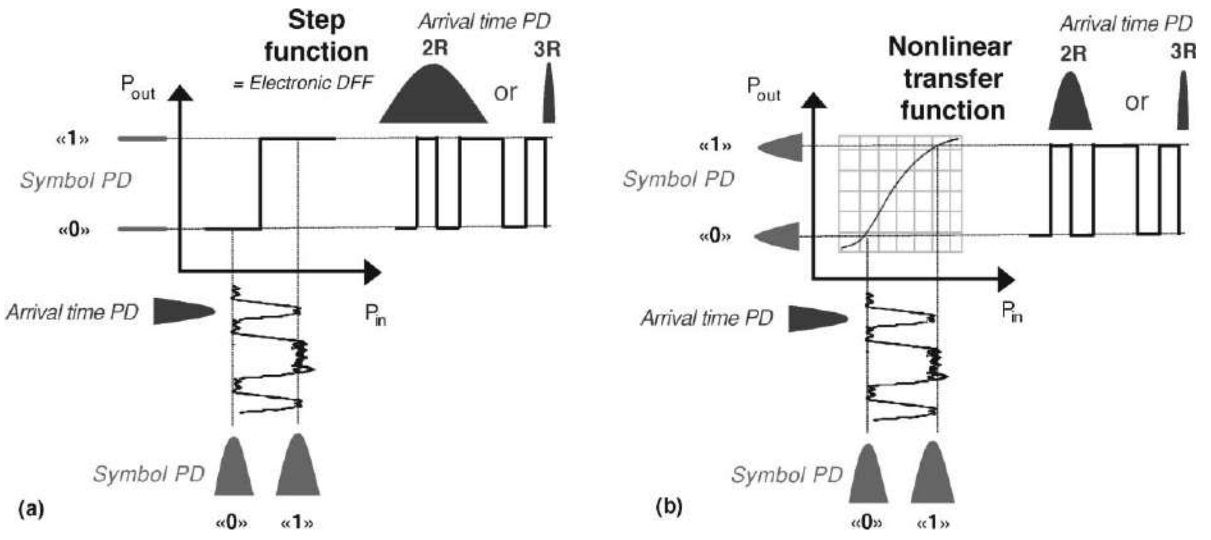

Figure 3 Signal Regeneration process using Nonlinear Gates. (a) Step transfer function (= Electronic DFF); (b) 'moderately' nonlinear transfer function. As an illustration of the Regenerator operation '1' and '0' symbols amplitude probability density (PD) and arrival time probability density (PD) are shown in light gray and dark gray, respectively.

element in 3R application, the theoretical evolution of BER with number of concatenated regenerators have been plotted for regenerators having different nonlinear responses. Figure 4 shows the numerically calculated evolution of the BER of a 10 Gbit/s NRZ signal with fixed optical signal-to-noise ratio (OSNR) at the input of the 3R regenerator, as a function of the cascaded regenerator incorporating nonlinear gates with nonlinear transfer function of different depths. From Figure 4, can be seen different behaviors depending on the nonlinear function shape. As previously stated, the best regeneration performance is obtained with an ideal step function (case a), which is actually the case for O/E regenerator using electronic decision flip-flop (DFF). In that case, BER linearly increase (i.e., more errors) in the cascade. Conversely, when nonlinearity is reduced (cases (b) and (c)), both BER and noise accumulate, until the concatenation of nonlinear functions reach some steady-state pattern, from which BER linearly increases. Concatenation of nonlinear devices thus magnifies shape differences in their nonlinear response, and hence their regenerative capabilities.

Moreover, as can be seen in Figure 4, all curves standing for different regeneration efficiencies pass through a common point defined after the first device. This clearly indicates that it is not possible to qualify the regenerative capability of any regenerator when considering the output signal after only one regenerator. Indeed, the BER is the same for either a 3R regenerator or a mere amplifier if only measured after a single element. This originates from the initial overlap between noise distributions associated with marks and spaces, that cannot be suppressed but only minimized by a single decision element through threshold adjustment.

As a general result, the actual characterization of the regenerative performance of any regenerator should *in fine* be conducted considering a cascade of regenerators. In practice this can easily be done with the experimental implementation of the regenerator under study in a recirculating loop. Moreover, such an investigation tool will also enable access to the regenerator performance with respect to the transmission capabilities of the regenerated signal, which should not be overlooked.

Let us now consider the physical implementation of such all-optical Signal Regenerators, along with the key features offered by the different alternatives.

All-Optical 2R/3R Regeneration Using Optical Nonlinear Gates

Prior to describing the different solutions for all-optical in-line Optical Signal Regeneration, it should be mentioned that since the polarization states of the optical data signals cannot be preserved during propagation, it is required that the regenerator exhibits an extremely low polarization sensitivity. This clearly translates to a careful optimization of all the different optical components making up the 2R/3R regenerator. It should be noted that this also applies to the O/E solutions but is of limited impact, since only the photodiode has to be polarization insensitive.

Figure 5 illustrates the generic principle of operation of an all-optical 3R Regenerator using optical nonlinear gates. Contrary to what occurs in O/E, regenerator where the extracted clock signal drives the electrical decision element, the incoming and distorted optical data signal triggers the nonlinear gate, hence generating a switching window which is

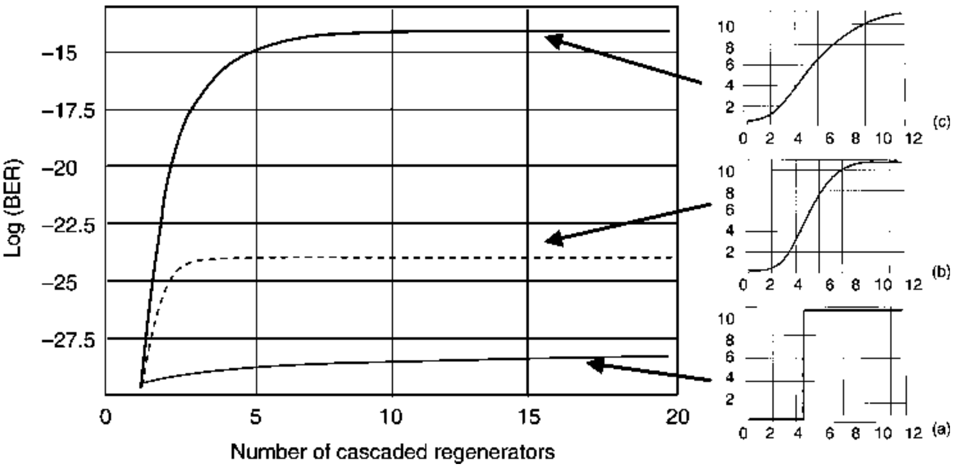

Figure 4 Evolution of the BER with concatenated regenerators for nonlinear gates with nonlinear transfer function of decreasing depths from case (a)–(c) (step function).

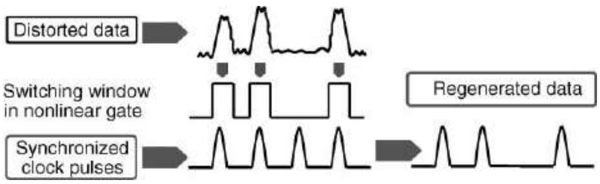

Figure 5 Principle of operation of an all-Optical Signal 3R Regenerator using nonlinear gates.

applied to a newly generated optical clock signal so as to reproduce the initial data stream on the new optical carrier.

In the case of 2R Optical Signal Regeneration, a continuous wave (CW) signal is substituted for the synchronized optical clock pulses. As previously mentioned, the actual regeneration performance of the 2R/3R devices will mainly depend upon the nonlinearity of the transfer function of the decision element but in 3R applications the quality of the optical clock pulses has also to be considered. In the following, we describe current solutions to explain the two main building blocks of all-optical Signal Regenerators: the decision element (i.e., nonlinear gate) and the clock recovery (CR) elements.

Optical Decision Element

In the physical domain, optical decision elements with ideal step response – as for electrical DFF – do not exist. Different nonlinear optical transfer functions, approaching more or less the ideal case, can be realized in various media such as fiber, SOA, electro-absorption modulators (EAM), and lasers. Generally, as described below, the actual response (hence the regenerative properties of the device) of such optical gates directly depends upon the incoming signal instantaneous power. Under these conditions, it appears essential to add an adaptation stage so as to reduce intensity fluctuations (as caused by propagation or crossing routing/switching node) and provide the decision element with fixed power conditions. In practice, this results in the addition of a control circuitry (either optical or electrical) in the Re-amplification block, whose complexity directly depends on actual system environment (ultra-fast power equalization for packet-switching applications and compensation of slow power fluctuations in transmission applications).

As previously described the decision gate performs Re-shaping (and Re-timing when clock pulses are added) of the incoming distorted optical signal, and represent the regenerator's core element. Ideally, it should also act as a transmitter with characteristics ensuring the actual propagation of the regenerated data stream. In that respect, the chirp possibly induced by the optical decision gate onto the regenerated signal – and the initial quality of the optical clock pulses in 3R applications – should be carefully considered (ideally by means of loop transmission) as to adequately match line transmission requirements.

Different solutions for the actual realization of the optical decision element have been proposed and extensively investigated using, for example, cross gain modulation in semiconductor optical amplifier (SOA) devices but the most promising and flexible devices probably are interferometers, for which, descriptions of the generic principle of operation follows. Consider a CW signal (probe) at λ_2 wavelength injected into an optical interferometer, in which one arm incorporates a nonlinear medium in which an input signal carried by λ_1 wavelength (command) is, in turn, injected. Such a signal, at λ_1 wavelength, induces a phase shift through cross-phase modulation (XPM) in this arm of the interferometer, the amount depending upon power level $P_{\mathrm{in},\lambda1}$. In turn, such phase modulation (PM) induces amplitude modulation (AM) on the signal at λ_2 wavelength when recombined at the output of the interferometer and translates the information carried by wavelength λ_1 onto λ_2. Under these conditions, such optical gates clearly act as wavelength converters (it should be mentioned that Wavelength Conversion is not necessarily equivalent to Regeneration; i.e., a linear transfer function performs suitable Wavelength Conversion but by no means Signal Regeneration).

Optical interferometers can be classified according to the nature of the nonlinearity exploited to achieve a π phase shift. In the case of fiber-based devices, such as the nonlinear optical loop mirror (NOLM), the phase shift is induced through the Kerr effect in an optical fiber. The key advantage of fiber-based devices such as NOLM lies in the near-instantaneous (fs) response of the Kerr nonlinearity, making them very attractive for ultra-high bit-rate operation ($\geq$160 Gbit/s). Polarization-insensitive NOLMs have been realized, although with the same drawbacks concerning integrability. With recent developments in highly nonlinear (HNL) fibers, however, the required NOLM fiber length could be significantly reduced, hence dramatically reducing environmental instability.

A second type of device is the integrated SOA-based Mach-Zehnder interferometers (MZI). In MZIs, the phase shift is due to the effect of photo-induced carrier depletion in the gain saturation regime of one of the SOAs. The control and probe can be launched in counter- or co-directional ways. In the first case, no optical filter is required at the output of the device for rejecting the signal at λ_1 wavelength but operation of the MZI is limited by its speed. At this point, one should mention that the photo-induced modulation effects in SOAs are intrinsically limited in speed by the gain recovery time, which is a function of the carrier lifetime and the injection current. An approach referred to as differential operation mode (DOM) and illustrated on Figure 6, which takes advantage of the MZI's interferometric properties, makes it possible to artificially increase the operation speed of such 'slow' devices up to 40 Gbit/s.

As discussed earlier, the nonlinear response is a key parameter for regeneration efficiency. Combining two interferometers is a straightforward means to improve the nonlinearity of the decision element transfer function, and hence regeneration efficiency. This approach was validated at 40 Gbit/s using a cascade of two SOA-MZI, (see Figure 7 (**left**)). Such a scheme offers the advantage of restoring data polarity and wavelength, hence making the regenerator inherently transparent. Finally, the second conversion stage can be used as an adaptation interface to the transmission link achieved through chirp tuning in this second device.

Such an Optical 3R Regenerator was upgraded to 40 Gbit/s, using DOM in both SOA-MZIs with validation in a 40 Gbit/s loop RZ transmission. The 40 Gbit/s eye diagram monitored at the regenerator output after 1, 10, and 100 circulations are shown in Figure 7 (**right**) and remain unaltered by distance. With this all-optical regenerator structure, the minimum OSNR tolerated by the regenerator (1 dB sensitivity penalty at 10^{-10} BER) was found to be as low as 25 dB/0.1 nm. Such results clearly illustrate the high performance of this SOA-based regenerator structure for 40 Gbit/s optical data signals.

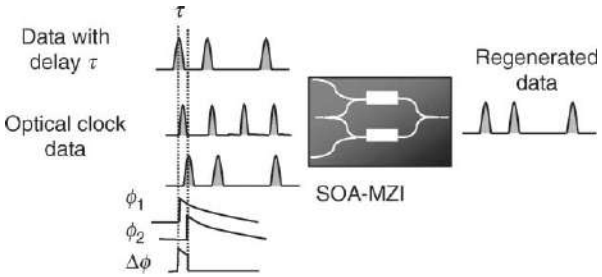

Figure 6 Schematic and principle of operation of SOA-MZI in differential mode.

Such a complex mode of operation for addressing 40 Git/s bit-rates will probably be discarded when we consider the recent demonstration of standard-mode wavelength conversion at 40 Gbit/s, which uses a newly-designed active-passive SOA-MZI incorporating evanescent-coupling SOAs. The device architecture is flexible in the number of SOAs, thus enabling easier operation optimization and reduced power consumption, leading to simplified architectures and operation for 40 Gbit/s optical 3R regeneration.

Based on the same concept of wavelength conversion for Optical 3R Regeneration, it should be noted many devices have been proposed and experimentally validated as wavelength converters at rates up to 84 Gbit/s, but with cascadability issues still to be demonstrated to assess their actual regenerative properties.

Optical Clock Recovery (CR)

Next to the decision, the CR is a second key function in 3R regenerators. One possible approach for CR uses electronics while another only uses optics. The former goes with OE conversion by means of a photodiode and subsequent EO conversion through a modulator. This conversion becomes more complex and power-hungry as the data-rate increases. It is clear that the maturity of electronics gives a current advantage to this approach. But considering the pros and cons of electronic CR for cost-effective implementation, the all-optical approach seems more promising, since full regenerator integration is potentially possible with reduced power consumption. In this view, we shall focus here on the optical approach and more specifically on the self-pulsating effect in three-sections distributed feedback (DFB) lasers or more recently in distributed Bragg reflector

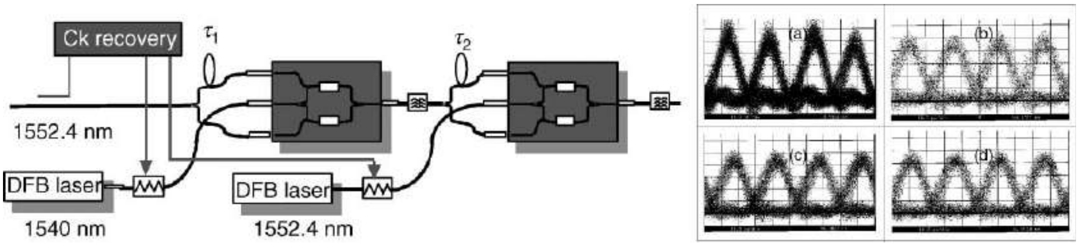

Figure 7 Optimized structure of a 40 Gbit/s SOA-based 3R regenerator. 40 Gbit/s eye diagram evolution: (a) B-to-B; (b) 1 lap; (c) 10 laps; (d) 100 laps.

(DBR) lasers. Recent experimental results illustrate the potentials of such devices for high bit rates (up to 160 Gbit/s), broad dynamic range, broad frequency tuning, polarization insensitivity, and relatively short locking times (1 ns). This last feature makes these devices good candidates for operation in asynchronous-packet regimes.

Optical Regeneration by Saturable Absorbers

We next consider saturable absorbers (SA) as nonlinear elements for optical regeneration. Figure 8 (**left**) shows a typical SA transfer function and illustrates the principle of operation. When illuminated with an optical signal with peak power below some threshold (P_{sat}), the photonic absorption of the SA is high and the device is opaque to the signal (low transmittance). Above P_{sat}, the SA transmittance rapidly increases and asymptotically saturates to transparency (passive loss being overlooked). Such a nonlinear transfer function only applies to 2R optical regeneration.

Different technologies for implementing SAs are available, but the most promising approach uses semiconductors. In this case, SA relies upon the control of carrier dynamics through the material's recombination centers. Parameters such as on–off contrast (ratio of transmittance at high and low incident powers), recovery time (1/e) and saturation energy, are key to device optimization. In the following, we consider a newly-developed ion-irradiated MQW-based device incorporated in a micro-cavity and shown on Figure 8 (**right**). The device operates as a reflection-mode vertical cavity, providing both a high on/off extinction ratio by canceling reflection at low intensity and a low saturation energy of 2 pJ. It is also intrinsically polarization-insensitive. Heavy ion-irradiation of the SA ensures recovery times (at 1/*e*) shorter than 5 ps (hence compatible with bit-rate above 40 Gbit/s), while maintaining a dynamic contrast in excess of 2.5 dB at 40 GHz repetition rate.

The regenerative properties of SA make it possible to reduce cumulated amplified spontaneous emission (ASE) in the '0' bits, resulting in a higher contrast between mark and space, hence increasing system performance. Yet SAs do not suppress intensity noise in the marks, which makes the regenerator incomplete. A solution for this noise suppression is optical filtering with nonlinear (soliton) pulses. The principle is as follows. In absence of chirp, the soliton temporal width scales in the same way as the reciprocal of its spectral width (Fourier-transform limit) times its intensity (fundamental soliton relation). Thus, an increase in pulse intensity corresponds to both time narrowing and spectral broadening. Conversely, a decrease in pulse intensity corresponds to time broadening and spectral narrowing. Thus, the filter causes higher loss when intensity increases, and lower loss when intensity decreases. The filter thus acts as an automatic power control (APC) in feed-forward mode, which causes power stabilization. The resulting 2R regenerator (composed by the SA and the optical filter) is fully passive, which is of high interest for submarine systems where the power consumption must be minimal, but it does not include any control in the time domain (no Re-timing).

System demonstrations of such passive SA-based Optical Regeneration have been reported with a 20 Gbit/s single-channel loop experiment. Implementation of the SA-based 2R Regenerator with 160 km-loop periodicity made it possible to double the error-free distance ($Q = 15.6$ dB or 10^{-9} BER) of a 20 Gbit/s RZ signal. So as to extend the capability of passive 2R regeneration to 40 Gbit/s systems, an improved configuration was derived from numerical optimization and experimentally demonstrated in a 40 Gbit/s WDM-like, dispersion-managed loop transmission, showing more than a fourfold increase in the WDM transmission distance at 10^{-4} BER (1650 km without the SA-based regenerator and 7600 km when implementing the 2R regenerator with 240 km periodicity).

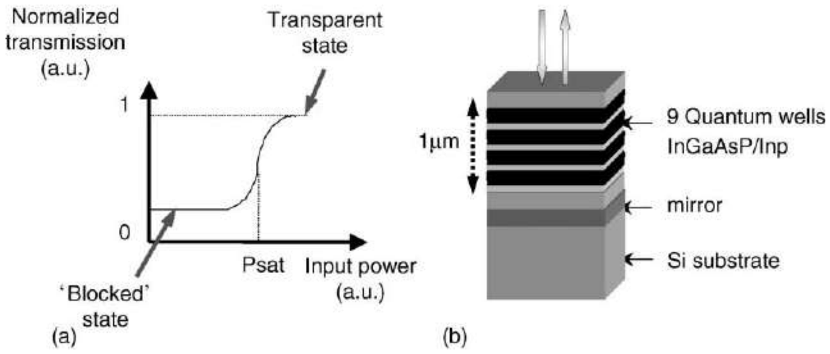

Figure 8 (a) Saturable Absorber (SA) ideal transfer function. (b) Structure of Multi-Quantum Well SA.

Such a result illustrates the potential high interest of such passive optical 2R regeneration in long-haul transmission (typically in noise-limited systems) since the implementation of SA increases the system's robustness to OSNR degradation without any extra power consumption. Reducing both saturation energy and insertion loss along with increasing dynamic contrast represent key future device improvements. Regeneration of WDM signals from the same device, such as one SA chip with multiple fibers implemented between Mux/DMux stages, should also be thoroughly investigated. In this respect, SA wavelength selectivity in quantum dots could possibly be advantageously exploited.

Synchronous Modulation Technique

All-optical 3R regeneration can be also achieved through in-line synchronous modulation (SM) associated with narrowband filtering (NF). Figure 9 shows the basic layout of such an Optical 3R Regenerator. It is composed of an optical filter followed by an Intensity and Phase Modulator (IM/PM) driven by a recovered clock. Periodic insertion of SM-based modulators along the transmission link provides efficient jitter reduction and asymptotically controls ASE noise level, resulting in virtually unlimited transmission distances. Re-shaping and Re-timing provided by IM/PM intrinsically requires nonlinear (soliton) propagation in the trunk fiber following the SM block. Therefore, one can refer to the approach as distributed optical regeneration. This contrasts with lumped regeneration, where 3R is completed within the regenerator (see above with Optical Regeneration using nonlinear gates), and is independent of line transmission characteristics.

However, when using a new approach referred to 'black box' optical regeneration (BBOR), it is possible to make the SM regeneration function and transmission work independently in such a way that any type of RZ signals (soliton or non-soliton) can be transmitted through the system. The BBOR technique includes an adaptation stage for incoming RZ pulses in the SM-based regenerator, which ensures high regeneration efficiency regardless of RZ signal format (linear RZ, DM-soliton, C-RZ, etc.). This is achieved using a local and periodic soliton conversion of RZ pulses by means of launching an adequate power into some length of fiber with anomalous dispersion. The actual experimental demonstration of the BBOR approach and its superiority over the 'classical' SM-based scheme for DM transmission was experimentally investigated in 40 Gbit/s DM loop transmission. Under these conditions, one can then independently exploit dispersion management (DM) techniques for increasing spectral efficiency in long-haul transmission, while ensuring high transmission quality through BBOR.

One of the key properties of the SM-based all-optical Regeneration technique relies on its WDM compatibility. The first (Figure 10, **left**) and straightforward solution to apply Signal Regeneration to WDM channels amounts to allocating a regenerator to each WDM channel. The second consists in sharing a single modulator, thus processing the WDM channels at once in serial fashion. This approach requires WDM synchronicity, meaning that all bits be synchronous with the modulation, that can be achieved either by use of appropriate time-delay lines located within a DMux/Mux apparatus (Figure 10, **upper right**), or by making the WDM channels inherently time-coincident at specific regenerator locations (Figure 10, **bottom right**). Clearly, the serial regeneration scheme is far simpler and cost-effective than the parallel version; however, optimized re-synchronization schemes still remain to be

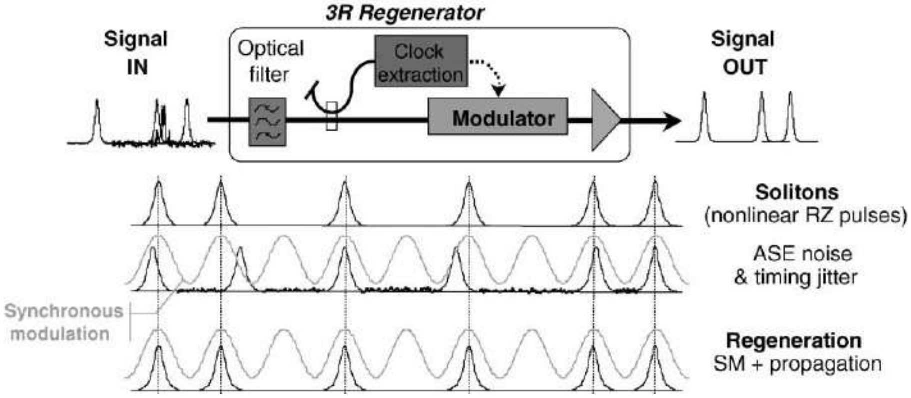

Figure 9 Basic layout of the all-optical Regenerator by Synchronous Modulation and Narrowband Filtering and illustration of the principle of operation.

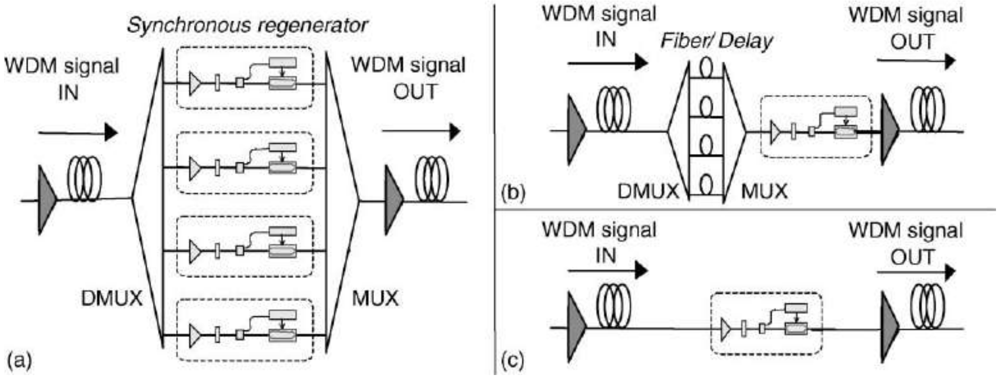

Figure 10 Basic implementation schemes for WDM all-Optical Regeneration. (a) parallel asynchronous; (b) serial re-synchronized; (c) serial self-synchronized.

developed for realistic applications. Experimental demonstration of this concept was assessed by means of a 4 × 40 Gbit/s dispersion-managed transmission over 10 000 km (BER $< 5.10^{-8}$) in which a single modulator was used for the simultaneous regeneration of the 4 WDM channels.

Considering next all-optical regeneration schemes with ultra-high speed potential, a compact and loss-free 40 Gbit/s Synchronous Modulator, based on optically-controlled SOA-MZI, was proposed and loop-demonstrated at 40 Gbit/s with an error-free transmission distance in excess of 10 000 km. Moreover, potential ultra-high operation of this improved BBOR scheme was recently experimentally demonstrated by means of a 80 GHz clock conversion with appropriate characteristics through the SOA-MZI. One should finally mention all fiber-based devices such as NOLM and NALM for addressing ultra-high speed SM-based Optical Regeneration, although no successful experimental demonstrations have been reported so far in this field.

Conclusion

In summary, optical solutions for Signal Regeneration present many key advantages. These are the only advantages to date to possibly ensure WDM compatibility of the regeneration function (mostly 2R related). Such optical devices clearly exhibits the best 2R regeneration performance (wrt to O/E solutions) as a result of the moderately nonlinear transfer function (which in turn can be considered as a drawback in 3R applications), but the optimum configuration is still to be clearly derived and identified depending upon the system application. Optics also allow to foresee and possibly target ultrafast applications above 40G, for signal regeneration if needed. Among the current drawbacks, one should mention the relative lack of wavelength/formats flexibility of these solutions (compared to O/E solutions). It is complex or difficult to restore the input wavelength or address any C-band wavelength at the output of the device or to successfully regenerate modulation formats other than RZ. In that respect, investigations should be conducted to derive new optical solutions capable of processing more advanced modulation formats at 40G. Finally, the fact that the nonlinear transfer function of the optical is in general triggered by the input signal instantaneous power also turns out to be a drawback since it requires control circuitry. The issue of cost (footprint, power consumption, etc.) of these solutions, compared to O/E ones, is still open. In this respect, purely optical solutions incorporating all-optical clock recovery, the performance of which is still to be technically assessed, are of high interest for reducing costs. Complete integration of an all-optical 2R/3R regenerator or such parallel regenerators onto a single semiconductor chip should also contribute to make all-optical solutions cost-attractive even though acceptable performance of such fully integrated devices is still to be demonstrated.

From today's status concerning the two alternative approaches for in-line regeneration (O/E or all-optical), it is safe to say that the choice between either solution will be primarily dictated by both engineering and economical considerations. It will result from a tradeoff between overall system performance, system complexity and reliability, availability, time-to-market, and rapid returns from the technology investment.

See also

Interferometry: Overview. **Optical Amplifiers:** Semiconductor Optical Amplifiers. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Bigo S, Leclerc O and Desurvire E (1997) All-optical fiber signal processing for solitons communications. *IEEE Journal of Selected Topics on Quantum Electronics* 3(5): 1208–1223.

Dagens B, Labrousse A, Fabre S, *et al*. (2002) New modular SOA-based active–passive integrated Mach–Zehnder interferometer and first standard mode 40 Gb/s all-optical wavelength conversion on the C-band. paper PD 3.1, Proceedings of ECOCOL 02, Copenhague.

Durhuus T, Mikkelsen B, Joergensen C, Danielsen SL and Stubkjaer KE (1996) All-optical wavelength conversion by semiconductor optical amplifiers. *Journal of Lightwave Technology* 14(6): 942–954.

Lavigne B, Guerber P, Brindel P, Balmefrezol E and Dagens B (2001) Cascade of 100 optical 3R regenerators at 40 Gbit/s based on all-active Mach–Zehnder interferometers. paper We.F.2.6, Proceedings of ECOC 2001, Amsterdam.

Leclerc O, Lavigne B, Chiaroni D and Desurvire E (2002) All-optical regeneration: Principles and WDM implementation. In: Kaminow IP and Li T (eds) *Optical Fiber Technologies IVA*, chap. 15, pp. 732–784. San Diego: Academic Press.

Leuthold J, Raybon G, Su Y, *et al*. (2002) 40 Gbit/s transmission and cascaded all-optical wavelength conversion over 1,000,000 km. *Electronic Letters* 38(16): 890–892.

Öhlén P and Berglind E (1997) Noise accumulation and BER estimates in concatenated nonlinear optoelectronic repeaters. *IEEE Photon Technol. Letters* 9(7): 1011–1013.

Otani T, Suzuki M and Yamamoto S (2001) 40 Gbit/s optical 3R regenerator for all-optical networks. We.F.2.1, Proceedings of ECOC 2001, Amsterdam.

Oudar J-L, Aubin G, Mangeney J, *et al*. (2004) Ultra-fast quantum-well saturable absorber devices and their application to all-optical regeneration of telecommunication optical signals. *Annales des Télécommunications* 58(11–12): 1667–1707.

Sartorius B, Mohrle M, Reichenbacher S, *et al*. (1997) Dispersive self-Q-switching in self-pulsating DFB lasers. *IEEE Journal of Quantum Electronics* 33(2): 211–218.

Ueno Y, Nakamura S, Sasaki J, *et al*. (2001) Ultrahigh-speed all-optical data regeneration and wavelength conversion for OTDM systems. Th.F.2.1, Proceedings of ECOC 2001, Amsterdam.

B

BABINET'S PRINCIPLE

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

In the development of the theory of diffraction, the diffraction field is due to a surface integral, the Fresnel integral, but no restrictions are imposed on the choice of the surface over which the integration must be performed. This fact leads to a very useful property called Babinet's principle, first stated by Jacques Babinet (1794–1872) in 1837 for scalar waves. We will discuss only the scalar Babinet's principle; discussion of the rigorous vector formulation can be found in a number of books on electromagnetic theory.

To introduce Babinet's principle, we label a plane, separating the source, S, and observation point, P, as Σ. If no obstructions are present, a surface integration over Σ yields the light distribution at P. If we place an opaque obstruction in this plane with a clear aperture, Σ_1, then the field at P is given by integrating over only Σ_1; contributions from Σ, outside of Σ_1, are zero since the obstruction is opaque.

We may define an aperture, Σ_2, as complementary to Σ_1 if the obstruction is constructed by replacing the transparent regions of Σ, i.e., Σ_1, by opaque surfaces and the opaque regions of Σ by clear apertures. Figure 1 shows two complementary obstructions, where the shaded region indicates an opaque surface.

The surface integral over Σ generates the field, E, in the absence of any obstruction. If obstruction Σ_1 is present then the diffraction field is E_1, obtained by integrating over Σ_1. According to Babinet's principle, the diffraction field due to obstruction Σ_2, must be

$$\tilde{E}_2 = \tilde{E} - \tilde{E}_1$$

We will look at examples of the application of the principle for both Fraunhofer and Fresnel diffraction.

Fraunhofer Diffraction

The electric field due to Fraunhofer diffraction is given by

$$E_P(\omega_x, \omega_y) = \frac{i\alpha e^{-ikR_0}}{\lambda R_0} \iint_{\Sigma} f(x, y) e^{-i(\omega_x x + \omega_y y)} \, dx \, dy$$

where α is the amplitude of the incident plane wave and R_0 is the distance from the obstruction to the observation point. We have defined the spatial frequencies, ω_x and ω_y by the equations

$$\omega_x = kL = -\frac{2\pi\xi}{\lambda R_0} \qquad \omega_y = kM = -\frac{2\pi\eta}{\lambda R_0}$$

This surface integral is an approximation of the Hygens–Fresnel integral and can be identified as a Fourier transform of the aperture transmission function, $f(x, y)$ in two dimensions. For our discussion we will consider only one dimension and ignore the parameters in front of the integral.

Assume that $a(x)$ is the amplitude transmission of an aperture in an obstruction, $b(x)$ is the amplitude transmission function of the complementary obstruction and g is the amplitude of the wave when no

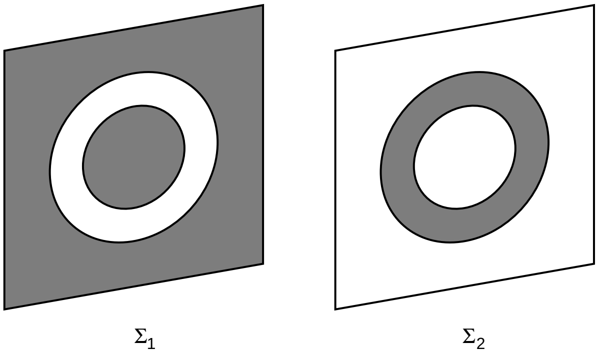

Figure 1 An aperture Σ_1 shown by the unshaded region and its complement Σ_2. Reproduced with permission from Guenther R (1990) *Modern Optics*. New York: Wiley. Copyright © 1990 John Wiley & Sons.

obstruction is present. The Fourier transforms of the one-dimensional amplitude transmission functions of the two apertures are equal to the Fraunhofer diffraction patterns that will be generated by the two apertures

$$A(k) = \int_{-\infty}^{\infty} a(x)\mathrm{e}^{-ikx}\,\mathrm{d}x$$

$$B(k) = \int_{-\infty}^{\infty} b(x)\mathrm{e}^{-ikx}\,\mathrm{d}x$$

With no aperture present, the far field amplitude is

$$G(k) = g\int_{-\infty}^{\infty} \mathrm{e}^{-ikx}\,\mathrm{d}x$$

Babinet's principle states that

$$B(x) = G - A(x)$$

must be the Fraunhofer diffraction field of the complementary aperture. We may rewrite this equation for the diffraction field as

$$B(k) = G\delta(k) + A(k)\mathrm{e}^{\mathrm{i}\pi} \qquad [1]$$

The first term on the right of the Fraunhofer diffraction field for the complementary obstruction [1] is located at the origin of the observation plane and is proportional to the amplitude of the unobstructed wave. The second term in the equation for the Fraunhofer diffraction pattern [1] is identical to the Fraunhofer diffraction pattern of the original aperture, except for a constant phase. Thus the Fraunhofer diffraction from the two complementary apertures are equal, except for a constant phase and the bias term at the origin. Physically this means that the diffraction intensity distributions of complementary apertures will be identical but their brightness will differ!

Fresnel Diffraction

We can calculate the Fresnel diffraction from an opaque disk by applying Babinet's principle to the diffraction pattern calculated for a circular aperture.

We assume that a circular aperture of radius a is illuminated by a point source a distance Z' from the aperture. We observe the transmitted wave at the point P, located a distance Z from the aperture (see Figure 2).

The Fresnel diffraction integral is

$$\tilde{E}_{\mathrm{P}} = \frac{\mathrm{i}\alpha}{\lambda\rho D}\mathrm{e}^{-\mathrm{i}kD} \times \iint f(x,y)\mathrm{e}^{-\frac{ik}{2\rho}\left[(x-x_0)^2+(y-y_0)^2\right]}\,\mathrm{d}x\,\mathrm{d}y \qquad [2]$$

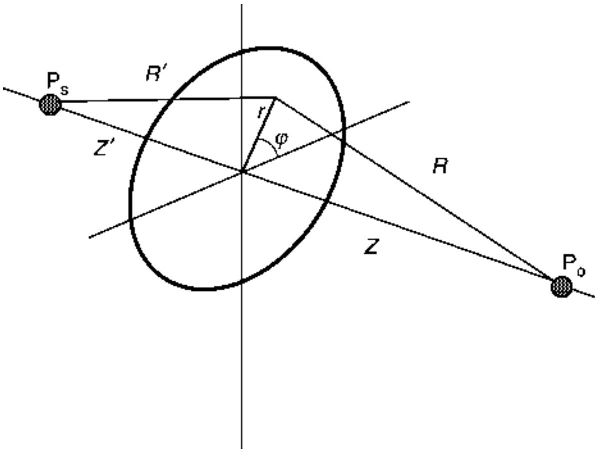

Figure 2 Geometry for the analysis of Fresnel diffraction of a circular aperture. Reproduced with permission from Guenther R (1990) *Modern Optics*. New York: Wiley. Copyright © 1990 John Wiley & Sons.

where the distance between the source and observation point is

$$D = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z+Z')} \qquad \frac{1}{\rho} = \frac{1}{Z} + \frac{1}{Z'}$$

and the coordinates of the stationary point are

$$x_0 = \frac{Z'\xi + Zx_s}{Z+Z'} \qquad y_0 = \frac{Z'\eta + Zy_s}{Z+Z'}$$

The parameters of the integral are simplified by the geometry selected for Figure 2. Both the observation point and source are located on the axis of symmetry of the aperture which lies along the z-axis of a cylindrical coordinate system thus

$$x_0 = y_0 = \xi = \eta = 0 \qquad \text{and} \qquad D = Z' + Z$$

The approximation for the distance from the source and observation point to the aperture yields

$$R + R' \approx Z + Z' + \frac{r^2}{2\rho} = D + \frac{r^2}{2\rho}$$

where $r^2 = x^2 + y^2$. The cylindrical coordinate version of [2] is then

$$\tilde{E}_{\mathrm{P}} = \frac{\mathrm{i}\alpha}{\lambda\rho D}\mathrm{e}^{-\mathrm{i}kD}\int_0^a\int_0^{2\pi} f(r,\varphi)\mathrm{e}^{-\mathrm{i}\frac{kr^2}{2\rho}}\,r\,\mathrm{d}r\,\mathrm{d}\phi \qquad [3]$$

We assume that the transmission function of the aperture is a constant, $f(r,\phi) = 1$, to make the integral as simple as possible. After performing the integration over the angle ϕ, [3] can be rewritten as

$$\tilde{E}_{\mathrm{Pa}} = \frac{\mathrm{i}\alpha\pi}{D}\mathrm{e}^{-\mathrm{i}kD}\int_0^{\frac{a^2}{\rho\lambda}} \mathrm{e}^{-\mathrm{i}\pi r^2/\rho\lambda}\,\mathrm{d}\left(\frac{r^2}{\rho\lambda}\right) \qquad [4]$$

Performing the integration of [4], results in the Fresnel diffraction amplitude

$$\tilde{E}_{\mathrm{Pa}} = \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD}\left[1 - \mathrm{e}^{-\mathrm{i}\pi a^2/\rho\lambda}\right] \qquad [5]$$

The intensity of the diffraction field is

$$\begin{aligned} I_{\mathrm{Pa}} &= 2I_0\left(1 - \cos\frac{\pi a^2}{\rho\lambda}\right) \\ &= 4I_0\sin^2\frac{\pi a^2}{2\rho\lambda} \qquad [6] \end{aligned}$$

[6] predicts a sinusoidal variation in intensity as the observation point moves toward the aperture.

Babinet's principle can now be used to evaluate Fresnel diffraction by an opaque disk, the same size as the circular aperture, i.e., of radius a. The field for the disk is obtained by subtracting the field, diffracted by a circular aperture, from the field of an unobstructed spherical wave, E_{P}:

$$\begin{aligned} \tilde{E}_{\mathrm{Pd}} &= \tilde{E}_{\mathrm{P}} - \tilde{E}_{\mathrm{Pa}} \\ \tilde{E}_{\mathrm{Pd}} &= \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD} - \tilde{E}_{\mathrm{Pa}} \\ &= \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD}\mathrm{e}^{-\mathrm{i}\pi a^2/\rho\lambda} \qquad [7] \end{aligned}$$

The intensity is

$$I_{\mathrm{Pd}} = I_0$$

The result of the application of Babinet's principle is the conclusion that, at the center of the geometrical shadow cast by the disk, there is a bright spot with the same intensity as would exist if no disk were present. This is Poisson's spot. The intensity of this spot is independent of the choice of the observation point along the z-axis.

See also

Diffraction: Fraunhofer Diffraction; Fresnel Diffraction.

Further Reading

Abramowitz M and Stegun IA (1964) *Handbook of Mathematical Functions*. NBS Applied Mathematics Series-55 Washington, DC: National Bureau of Standards.

Born M and Wolf E (1959) *Principles of Optics*. New York: Pergamon Press.

Guenther R (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Rossi B (1957) *Optics*. Reading, MA: Addison-Wesley.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

C

CHAOS IN NONLINEAR OPTICS

R G Harrison and W Lu, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Instabilities in laser emission, notably in the form of spontaneous coherent pulsations, have been observed almost since the first demonstration of laser action. Indeed, the first laser operated in 1960 by Maiman produced noisy spiked output even under conditions of quasi-steady excitation and provided the early impetus for studies of such effects. Subsequent theoretical efforts towards understanding these phenomena occurred up to the 1980s at a modest level, due in part to the wide scope of alternative areas of fertile investigation provided by lasers during this period. However, since the 1990s, there has been a major resurgence of interest in this area. This has been due to profound mathematical discoveries over this period which have revolutionized our understanding of classical dynamical systems. It is now clear that many systems containing some form of nonlinearity are dynamically unstable and even chaotic and that such behavior is deterministic. Further, the discovery that chaos evolves through particular routes with well-defined scenarios and that such routes are universal, has stimulated experimentalists to search for physical systems that exhibit these properties. These phenomena have now been observed in areas as diverse as fluid flow, chemical reactions, population ecology, and superconductor devices. The laser, perhaps the most recent newcomer to the field, is a paradigm for such investigation, owing to its simplicity both in construction and in the mathematics that describe it, and already a wide range of these phenomena have been observed. Along with a proliferation of new observations, many of the instabilities known to occur in these systems are now being reinterpreted in lieu of our new insight in this area. From this, fascinating new concepts of control and synchronization of chaos have emerged and spawned new fields of applications, not least in secure communication. In this article we introduce the general principles of dynamical instabilities, chaos, and control in lasers. For in-depth study, the reader is referred to texts and review articles cited in the Further Reading section at the end of this article.

Laser Physics

Our conventional understanding of laser physics is concerned with how cooperative order evolves from randomness. This transition is explained by first considering the ensemble of lasing atoms in thermal equilibrium. Every atom executes a small motion, giving rise to an oscillating electric dipole, described by linear dynamics. For larger motions, the atomic dipoles interfere with each other's motion and beyond a particular threshold, the motion becomes 'co-operative' or ordered over a long range. The role of the electromagnetic field in the laser cavity in ordering the induced atomic dipoles, is collectively described as giving rise to a macroscopic polarization, the magnitude of which depends on the number of dipoles (excited atoms). The dynamical interplay between the cavity field amplitude (E) as one variable and the atomic material variables of polarization (P) and population (D) of excited atoms, all of which are in general complex quantities, provide a full description of lasing action. For the simplest laser, a single mode two-level system with a homogeneously broadened gain lasing at resonance, these reduce to just three real variables, the Maxwell–Bloch equations:

$$E = -\kappa E + \kappa P \qquad [1]$$

$$P = \gamma_{\perp} ED - \gamma_{\perp} P \qquad [2]$$

$$D = \gamma_{II}(\lambda + 1) - \gamma_{II} D - \gamma_{II} \lambda EP \qquad [3]$$

where κ is the cavity decay rate, $\gamma_\perp$ is the decay rate of atomic polarization, γ_{II} is the decay rate of population inversion, and λ the pumping parameter.

Described in this way the laser acts as a single oscillator in much the same way as a mechanical nonlinear oscillator, where laser output represents damping of the oscillator and excitation of the atoms is the driving mechanism to sustain lasing (oscillation). For ordered or coherent emission, it is necessary for one or both of the material variables (which are responsible for further lasing emission) to respond sufficiently fast to ensure a phase correlation with the existing cavity field. This is readily obtained in many typical lasers with output mirrors of relatively high reflectivity, since the field amplitude within the laser cavity will then vary slowly compared to the fast material variables which may then be considered through their equilibrium values. This situation, commonly referred to as adiabatic elimination of fast variables, reduces the dynamics of lasing action to that of one (or two) variables, the field (and population), all others being forced to adapt constantly to the slowly varying field variable. Our familiar understanding of DC laser action presupposes such conditions to hold. However when, for example, the level of excitation of the atoms is increased to beyond a certain value, i.e., the second laser threshold, all three dynamical variables may have to be considered, which satisfies the minimum requirement of three degrees of freedom (variables) for a system to display chaos. So this simple laser is capable of chaotic motion, for which the emission is aperiodic in time and has a broadband spectrum. Prediction of such behavior was initially made by Haken in 1975 through establishing the mathematical equivalence of the Maxwell–Bloch equations to those derived earlier by Lorentz describing chaotic motion in fluids.

Nonlinear Dynamics and Chaos in Lasers

In general there is no general analytical approach for solving nonlinear systems such as the Maxwell–Bloch equations. Instead solutions are obtained by numerical means and analyzed through geometric methods originally developed by Poincaré as early as 1892. This involves the study of topological structures of the dynamical trajectories in phase space of the variables describing the system. If an initial condition of a dissipative nonlinear dynamical system, such as a laser, is allowed to evolve for a long time, the system, after all the transients have died out, will eventually approach a restricted region of the phase space called an attractor. A dynamical system can have more than one attractor, in which case different initial conditions lead to different types of long-time behavior. The simplest attractor in phase space is a fixed point which is a solution with just one set of values for its dynamical variables; the nonlinear system is attracted towards this point and stays there, giving a DC solution. For other control conditions the system may end up making a periodic motion; the attractor of this motion is called a limit cycle. However, when the operating conditions exceed a certain critical value, the periodic motion of the system breaks down into a more complex dynamical trajectory, which never repeats. This motion represents a third kind of attractor in phase space called a chaotic or strange attractor.

Figure 1 shows a sequence of trajectories of the Lorentz equations in the phase space of (x, y, z) on increasing the value of one of the control parameters; this corresponds to (E, P, D) in the Maxwell–Block equations on increasing the laser pump. For a control setting near zero, all trajectories approach stable equilibrium at the origin, the topological structure of the basin of attraction being hyperbolic about zero (Figure 1a). For the laser equations, this corresponds to operation below lasing threshold for which the magnitude of the control parameter (laser gain) is insufficient to produce lasing. As the control parameter is increased, the basin lifts at its zero point to create an unstable fixed point here but with now two additional fixed points located in the newly formed troughs on either side of the zero point, which is now a saddle point. This is illustrated in Figure 1b where the system eventually settles to one or other of the two new and symmetrical fixed points, depending on the initial conditions. This corresponds to DC or constant lasing as defined by the parameter values of one or other of these points which are indistinguishable. Pictorially, think of a conventional saddle shape comprising a hyperbolic and inverted hyperbolic form in mutually perpendicular planes and connected tangentially at the origin. With the curve of the inverted hyperbola turned up at its extremity and filling in the volume with similar profiles which allow the two hyperbolic curves to merge into a topological volume, one sees that a ball placed at the origin is constrained to move most readily down either side of the inverted hyperbolas into one or other of the troughs formed by this volume. For the Lorentz system, chaotic behavior occurs at larger values of the control parameter when all three fixed points become saddles. Since none of the equilibrium points are now attracting, the behavior of the system cannot be a steady motion. Although perhaps difficult to visualize topologically, it is then possible to find a region in this surface enclosing all three points and

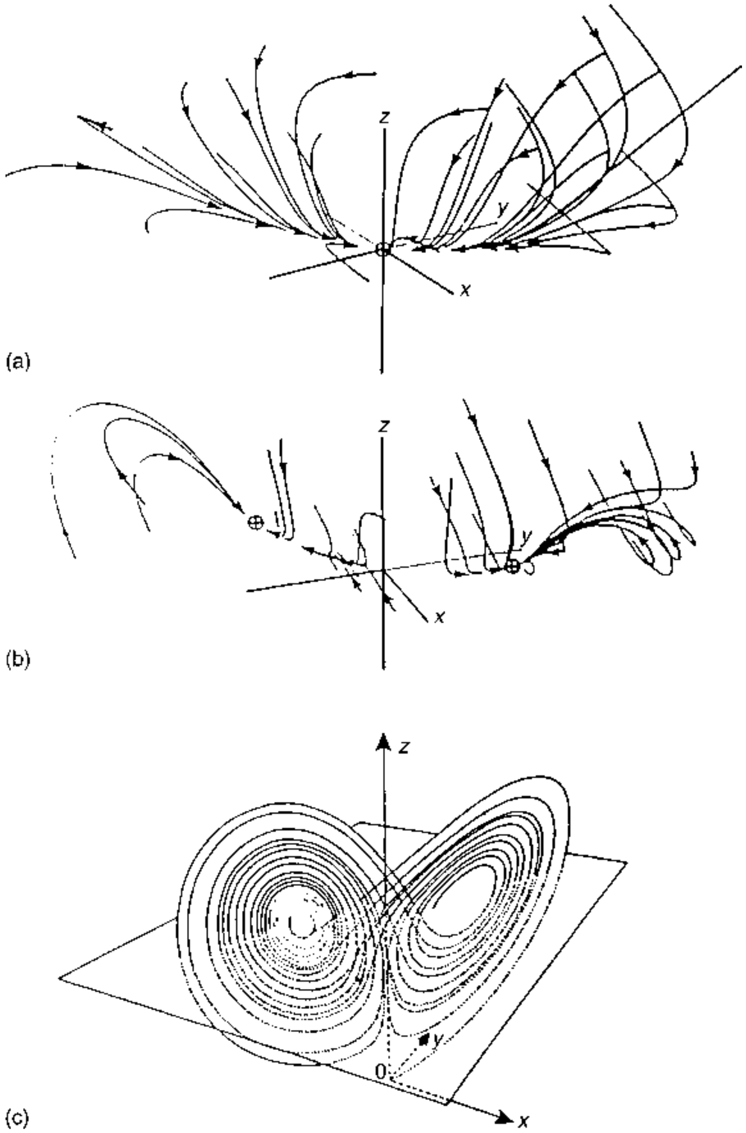

Figure 1 Trajectories in the three-dimensional phase space of the Lorentz attractor on increasing one of the control parameters. (Reproduced with permission from Thompson JMT and Stewart HB (1987) *Nonlinear Dynamics and Chaos*. New York: John Wiley.)

large enough so that no trajectories leave the region. Thus, all initial conditions outside the region evolve into the region and remain inside from then on. A corresponding chaotic trajectory is shown in Figure 1c. A point outwardly spirals from the proximity of one of the new saddle points until the motion brings it under the influence of the symmetrically placed saddle, the trajectory then being towards the center of this region from where outward spiraling again occurs. The spiraling out and switching over continues forever though the trajectory never intersects. In this case, arbitrarily close initial conditions lead to trajectories, which, after a sufficiently long time, diverge widely. Since a truly exact assignment of the initial condition is never possible, even numerically, a solution comprising several such trajectories therefore evolves and, as a consequence, long-term predictability is impossible. This is in marked contrast to that of the fixed point and limit cycle attractors, which settle down to the same solutions. A recording of one of these variables in time, say for a laser the output signal amplitude (in practice recorded as signal intensity, proportional to the square of the field amplitude), then gives oscillatory emission of increasing intensity with sudden discontinuities (resulting from flipping from one saddle to the other in Figure 1c) as expected.

While the Lorentz/Haken model is attractive for its relative simplicity, many practical lasers cannot be reduced to this description. Nevertheless, the

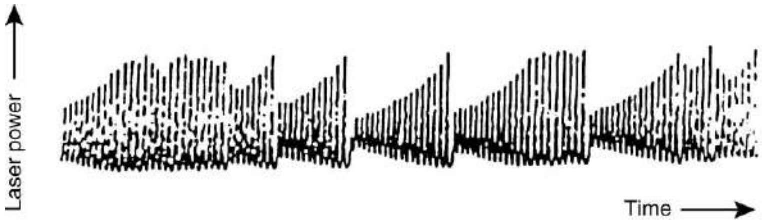

Figure 2 Lorentz-type chaos in the NH_3 laser (emitting at 81 μm optically pumped by a N_2O laser.

underlying topology of the trajectory of solutions in phase space for these systems is often found to be relatively simple and quite similar, as for three-level models descriptive of optically pumped lasers. An experimental example of such behavior for a single mode far infared molecular laser is shown in Figure 2. In other lasers, for which there is adiabatic elimination of one or other of the fast variables, chaotic behavior is precluded if, as a consequence, the number of variables of the system is reduced to less than three. Indeed this is the case for many practical laser systems. For these systems the addition of an independent external control parameter to the system, such as cavity length or loss modulation, have been extensively used as a means to provide the extra degrees of freedom. Examples include gas lasers with saturable absorbers, solid state lasers with loss modulation, and semiconductor lasers with external cavities, to name but a few. In contrast, for multimode rather than single mode lasers, intrinsic modulation of inversion (or photon flux) by multimode parametric interaction ensures the additional degrees of freedom. Furthermore, when the field is detuned from gain center, the dynamical variables E, P, D are complex, providing five equations for single mode systems, which is more than sufficient to yield deterministic chaos for suitable parameter values. Also, of significance is the remarkably low threshold found for the generation of instabilities and chaos in single-mode laser systems in which the gain is inhomogeneously broadened, an example being the familiar He–Ne laser.

Erratic and aperiodic temporal behavior of any of the system's variables implies a corresponding continuous spectrum for its Fourier transform, which is thus a further signature of chaotic motion. Time series, power spectra, and routes to chaos collectively provide evidence of deterministic behavior. Of the wide range of possible routes to chaos, three have emerged as particularly common and are frequently observed in lasers. These are period doubling, intermittency, and two-frequency scenarios. In the first, a solution, which is initially stable is found to oscillate, the period of which successively doubles at distinct values of the control parameter. This continues until the number of fixed points becomes infinite at a finite parameter value, where the variation in time of the solution becomes irregular. For the intermittency route, a signal that behaves regularly in time becomes interrupted by statistically distributed periods of irregular motion, the average number of which increases with the external control parameter until the condition becomes chaotic. The two-frequency route is more readily identified with early concepts of turbulence, considered to be the limit of an infinite sequence of instabilities (Hopf bifurcation) evolving from an initial stable solution each of which creates a new basic frequency. It is now known that only two or three instabilities (frequencies) are sufficient for the subsequent generation of chaotic motion.

Applications of Chaos in Lasers

It has been accepted as axiomatic since the discovery of chaos that chaotic motion is in general neither predictable nor controllable. It is unpredictable because a small disturbance will produce exponentially growing perturbation of the motion. It is uncontrollable because small disturbances lead to other forms of chaotic motion and not to any other stable and predictable alternative. It is, however, this very sensitivity to small perturbations that has been more recently used to stabilize and control chaos, essentially using chaos to control itself. Among the many methods proposed in the late 1980s, a feedback control approach was proposed by Ott, Grebogi, and Yorke (OGY) in which tiny feedback was used to stabilize unstable periodic orbits or fixed points of chaotic attractors. This control strategy can be best understood by a schematic of the OGY control algorithm for stabilizing a saddle point P^*, as shown in Figure 3. Curved trajectories follow a stable (unstable) manifold towards (away from) the saddle point. Without perturbation of a parameter, the starting state, s_1, would evolve to the state s_2.

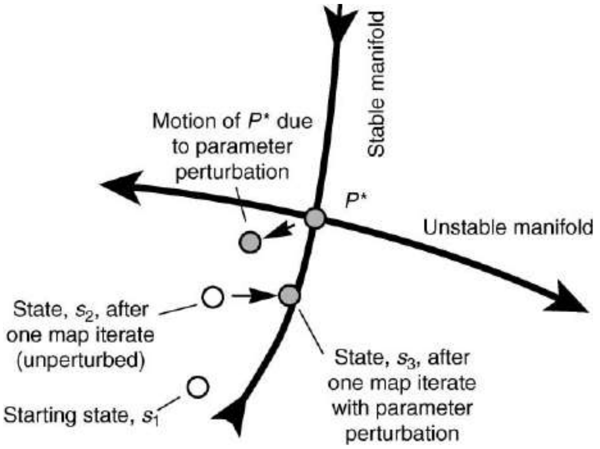
Stable manifold
Motion of P^* due to parameter perturbation
P^*
Unstable manifold
State, s_2, after one map iterate (unperturbed)
State, s_3, after one map iterate with parameter perturbation
Starting state, s_1

Figure 3 The OGY control strategy for stabilizing an orbit in a chaotic attractor.

The effect of changing a parameter of the system is depicted as shifting states near P^* along the solid black arrows, whereas the combination of the unperturbed trajectory and the effect of the perturbation is to bring the state to a point, s_3, on the stable manifold. Once on the stable manifold, the trajectory naturally tends towards the desired point. This algorithm has been successfully applied in numerous experimental systems without *a priori* modeling of these systems, examples being in cardiology, electronics, and lasers. Based on the concept of feedback control, various other approaches have been developed where emphasis has been given to algorithms which are more readily implemented in practical systems, in particular that utilizing occasional proportional feedback (OPF). These pioneering studies have since inspired prolific activities, both in theory and experiment, of control of chaos across many disciplines, opening up possibilities of utilizing chaos in many diverse systems. In optics, Roy *et al.* first demonstrated dynamical control of an autonomously chaotic and high-dimensional laser system on microsecond time-scales. The laser, a diode pumped solid-state Nd:YAIG system with an intracavity KTP doubling crystal, exhibited chaotic behavior from coupling of the longitudinal modes through nonlinear sum-frequency generation. To control the system, the total laser output was sampled within a window of selected offset and width. A signal proportional the deviation of the sampled intensity from the center of the window was generated and applied to perturb the driving current of the diode laser. The sampling frequency is related to the relaxation oscillation frequency of the system. This control signal repeatedly attempts to bring the system closer to a periodic unstable orbit that is embedded in the chaotic attractor, resulting in a realization of the periodic orbit. By adjusting the frequency of the feedback signal they observed the basic (relaxation oscillation) and many higher-order periodic waveforms of the laser intensity. In a subsequent experiment, Gills *et al.* showed that an unstable steady state of this laser could also be stabilized by the OPF control technique and tracked with excellent stability to higher output power with increase of the pump excitation.

Another area to receive considerable attention in the last few years is that of synchronization of chaos. It may be expected that chaotic systems will defy synchronization because two such identical systems started at nearly the same initial conditions have dynamical trajectories that quickly becomes uncorrelated. However, it has now been widely demonstrated that when these systems are linked, their chaotic trajectories converge to be the same and remain in step with each other. Further, such synchronization is found to be structurally stable and does not require the systems to be precisely identical. Not surprisingly, these findings have attracted interest from the telecommunication community; the natural masking of information by chaotic fluctuations offering a means to a certain degree of security. In this new approach, a chaotic carrier of information can be considered as a generalization of the more traditional sinusoidal carrier. In communication systems that use chaotic waveforms, information can be recovered from the carrier using a receiver synchronized, or tuned, to the dynamics of the transmitter. Optical systems are particularly attractive since they display fast dynamics, offering the possibility of communication at bandwidths of a hundred megahertz or higher. Van Wiggeren and Roy first demonstrated data transmission rates of 10 Mbites per second using erbium-doped fiber ring lasers. These lasers are particularly well-suited for communication purposes since their lasing wavelength is close to the minimum-loss wavelength in optical fiber. Figure 3 shows a schematic of this system. The tiny message, in order of one thousandth of the carrier intensity, is decoded in the transmitter. The receiver tends to synchronize its behavior to the chaotic part of the transmitted wave (but not the message). Subtracting the waveform created in the receiver from the transmitted signal recovers the tiny message. In principle, it is possible to communicate information at ultrahigh data rate with the use of this scheme because the chaotic dynamics in the ring laser has a very large spectral width. In a later experiment they showed that a receiver can recover information at 126 Mbits/sec from the chaotic carrier (Figure 4).

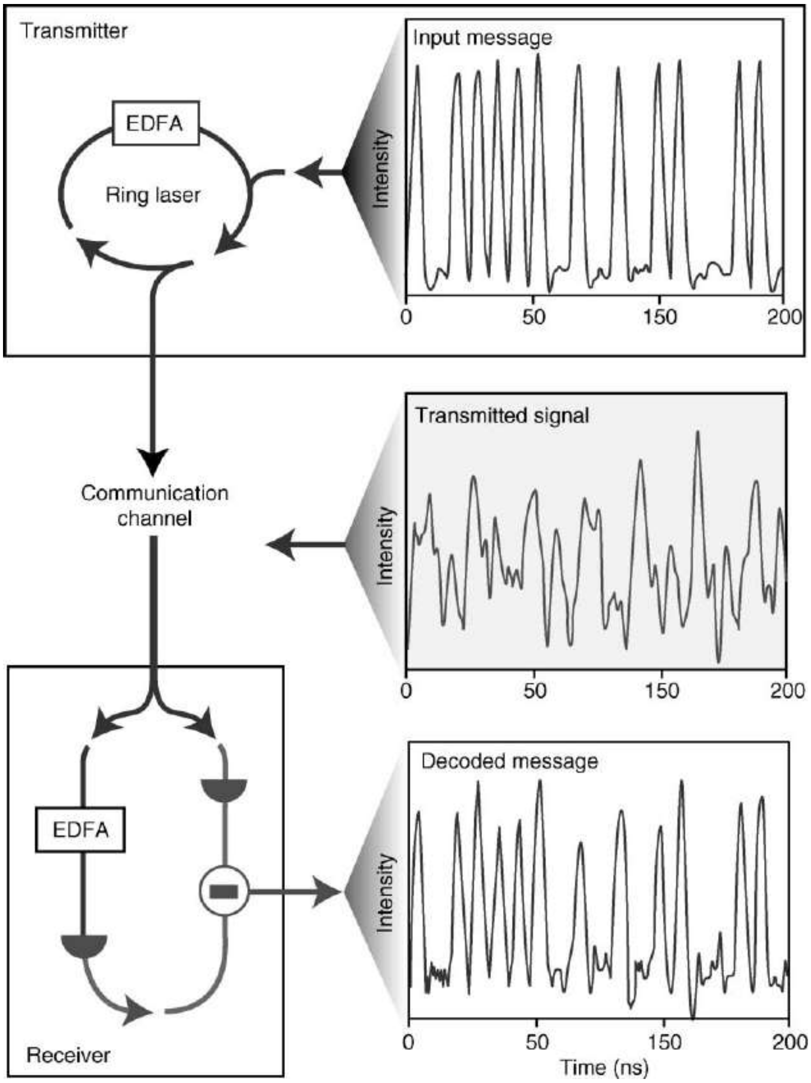
Transmitter
EDFA
Ring laser
Input message
Intensity
0
50
150
200
Communication channel
Transmitted signal
Intensity
0
50
150
200
EDFA
Receiver
Decoded message
Intensity
0
50
150
200
Time (ns)

Figure 4 Schematic of communication with chaotic erbium-doped fiber amplifiers (EDFAs). Injecting a message into the transmitter laser folds the data into the chaotic frequency fluctuations. The receiver reverses this process, thereby recovering a high-fidelity copy of the message. (Reproduced with permission from Gauthier DJ (1998) Chaos has come again. *Nature* 279: 1156–1157.)

See also

Fourier Optics. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Polarization:** Introduction.

Further Reading

Abraham NB, Lugiato LA and Narducci LM (eds) (1985) Instabilities in active optical media. *Journal of the Optical Society of America B* 2: 1–272.

Arecchi FT and Harrison RG (1987) *Instabilities and Chaos in Quantum Optics*. Synergetics Series 34, 1–253. Berlin: Springer Verlag.

Arecchi FT and Harrison RG (eds) (1994) *Selected Papers on Optical Chaos*. SPIE Milestone Series, vol. MS 75, SPIE Optical Engineering Press.

Gauthier DJ (1998) Chaos has come again. *Nature* 279: 1156–1157.

Gills Z, Iwata C and Roy R (1992) Tracking unstable steady states: extending the stability regime of a multimode laser system. *Physical Review Letters* 69: 3169–3172.

Haken H (1985) Light. In: *Laser Light Dynamics*, vol. 2. Amsterdam: North-Holland.

Harrison RG (1988) Dynamical instabilities and chaos in lasers. *Contemporary Physics* 29: 341–371.

Harrison RG and Biswas DJ (1985) Pulsating instabilities and chaos in lasers. *Progress in Quantum Electronics* 10: 147–228.

Hunt ER (1991) Stabilizing high-period orbits in a chaotic system – the diode resonator. *Physical Review Letters* 67: 1953–1955.

Ott E, Grebogi C and Yorke JA (1990) Controlling chaos. *Physical Review Letters* 64: 1196–1199.

Pecore LM and Carrol TL (1990) Synchronization of chaotic systems. *Physical Review Letters* 64: 821–824.

Roy R, *et al.* (1992) Dynamical control of a chaotic laser: experimental stabilization of a globally coupled system. *Physical Review Letters* 68: 1259–1262.

VanWiggeren GD and Roy R (1998) Communication with chaotic lasers. *Science* 279: 1198–1200.

VanWiggeren GD and Roy R (1998) Optical communication with chaotic waveforms. *Physical Review Letters* 279: 1198–1200.

CHEMICAL APPLICATIONS OF LASERS

Contents

Detection of Single Molecules in Liquids
Diffuse-Reflectance Laser Flash Photolysis
Laser Manipulation in Polymer Science
Nonlinear Spectroscopies
Photodynamic Therapy of Cancer
Pump and Probe Studies of Femtosecond Kinetics
Time-Correlated Single-Photon Counting
Transient Holographic Grating Techniques in Chemical Dynamics

Detection of Single Molecules in Liquids

A J de Mello, J B Edel and E K Hill, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A significant challenge facing experimentalists in the physical and biological sciences is the detection and identification of single molecules. The ability to perform such sensitive and selective measurements is extremely valuable in applications such as DNA analysis, immunoassays, environmental monitoring, and forensics, where small sample volumes and low analyte concentrations are the norm. More generally, most experimental observations of physical systems provide a measurement of ensemble averages, and yield information only on mean properties. In contrast, single molecule measurements permit observation of the interactions and behavior of a heterogeneous population in real time.

Over the past few years a number of techniques with sufficient sensitivity have been developed to detect single molecules. Scanning probe microscopies (most notably scanning tunneling microscopy and atomic force microscopy) have been used to great effect in the analysis of surface bound species, but for the detection of single molecules in liquids, optical methods incorporating the measurement of absorption or emission processes, have proved most successful.

The Absorption–Emission Cycle

The key concept underlying most emissive approaches to single molecule detection is that a single molecule can be cycled repeatedly between its ground state and an excited electronic state to yield multiple photons. The process can be understood by reference to Figure 1. Fluorescence emission in the condensed phase can be described using a four-step cycle. Excitation from a ground electronic state to an excited state is followed by rapid (internal) vibrational relaxation. Subsequently, radiative decay to the ground state is observed as fluorescence emission and is governed by the excited state lifetime. The final stage is internal relaxation back to the original ground state. Under saturating illumination, the rate-limiting step for this cycle is governed by the fluorescence lifetime (τ_f), which is typically of the order of a few nanoseconds. If a single molecule diffuses through an illuminated zone (e.g., the focus of a laser beam) it may reside in that

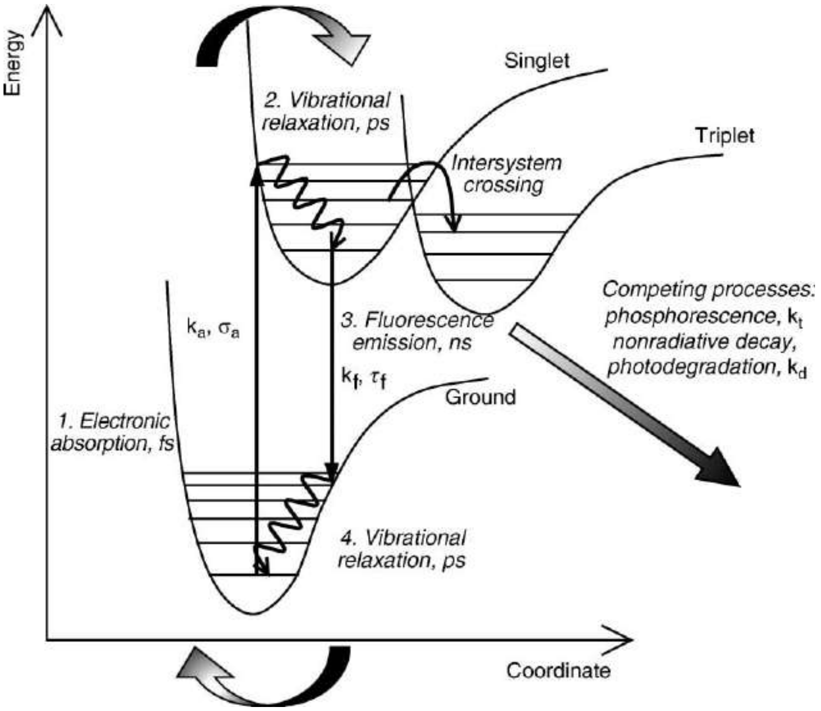
Energy
2. Vibrational relaxation, ps
Singlet
Triplet
Intersystem crossing
Competing processes: phosphorescence, k_t nonradiative decay, photodegradation, k_d
k_a, σ_a
3. Fluorescence emission, ns
k_f, τ_f
Ground
1. Electronic absorption, fs
4. Vibrational relaxation, ps
Coordinate

Figure 1 Schematic illustration of the molecular absorption–emission cycle and timescales for the component processes. Competing processes that may reduce the ultimate photon yield are also shown.

region for several milliseconds. The rapid photon absorption-emission cycle may therefore be repeated many times during the residence period, resulting in a burst of fluorescence photons as the molecule transits the beam. The burst size is limited theoretically by the ratio of the beam transit time (τ_r) and the fluorescence lifetime:

$$N_{\text{photons}} = \frac{\tau_r}{\tau_f} \qquad [1]$$

For typical values of τ_r (5 ms) and τ_f (5 ns) up to one million photons may be emitted by the single molecule. In practice, photobleaching and photodegradation processes limit this yield to about 10^5 photons. Furthermore, advances in optical collection and detection technologies enable registration of about 1–5% of all photons emitted. This results in a fluorescence burst signature of up to a few thousand photons or photoelectrons.

Successful single molecule detection (SMD) depends critically on the optimization of the fluorescence burst size and the reduction of background interference from the bulk solvent and impurities. Specifically a molecule is well suited for SMD if it is efficiently excited by an optical source (i.e., possesses a large molecular absorption cross-section at the wavelength of interest), has a high fluorescence quantum efficiency (favoring radiative deactivation of the excited state), has a short fluorescence lifetime, and is exceptionally photostable.

Ionic dyes are often well suited to SMD as fluorescence quantum efficiencies can be close to unity and fluorescence lifetimes below 10 nanoseconds. For example, xanthene dyes such as Rhodamine 6G and tetramethyl-rhodamine isothiocyanate are commonly used in SMD studies. However, other highly fluorescent dyes such as fluorescein are unsuitable for such applications due to unacceptably high photodegradation rate coefficients. Furthermore, some solvent systems may enhance nonradiative processes, such as intersystem crossing, and yield significant reduction in the photon output. Structures of three common dyes suitable for SMD are shown in Figure 2.

Signal vs Background

The primary challenge in SMD is to ensure sufficient reduction in background levels to enable discrimination between signal and noise. As an example, in a 1 nM aqueous dye solution each solute molecule occupies a volume of approximately 1 fL. However, this same volume also contains in excess of 10^{10} solvent molecules. Despite the relatively small scattering cross-section for an individual water molecule ($\sim 10^{-28}$ cm^2 at 488 nm), the cumulative scattering

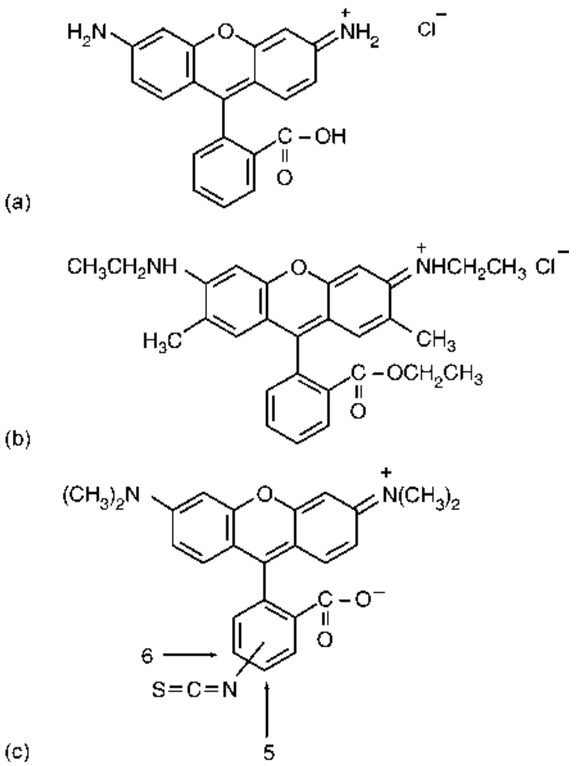

Figure 2 Structures of common dye molecules suitable for SMD in solution: (a) 3,6-diamino-9-(2-carboxyphenyl)-chloride (rhodamine 110); (b) 9-(2-(ethoxycarbonyl)phenyl)-3,6-bis (ethylamino)-2,7-dimethyl chloride (rhodamine 6G); (c) 9-(2-carboxy-isothiocyanatophenyl)-3,6-bis(dimethylamino)-inner salt (tetra methylrhodamine-5-(and-6)-isothiocyanate).

signal from the solvent may swamp the desired fluorescence signal. The principal method of reducing the solvent background is to minimize the optical detection volume: the signal from a single molecule is independent of probe volume dimensions, but the background scales proportionally with the size of the detection region. Although there are several experimental approaches to SMD in solution, several factors hold common:

1. Tiny detection volumes (10^{-12}–10^{-15} L) are used to reduce background signals.
2. A low analyte concentration combined with the small observation volume, ensures that less than one analyte molecule is present in the probe volume on average.
3. High-efficiency photon collection (optics) and detection maximize the proportion of the isotropic fluorescence burst that is registered.
4. Background reduction methods are employed to improve signal-to-noise ratios. These include: optical rejection of Raman and Rayleigh scatter, time-gated discrimination between prompt scatter and delayed emission, and photobleaching of the solvent immediately before detection.

The minute volumes within which single molecules are detected can be generated in a variety of ways. Picoliter volumes can be defined by mutually orthogonal excitation and detection optics focused in a flowing stream. Much smaller, femtoliter probe volumes are generated using confocal microscopes. At this level, background emission is significantly reduced and high signal-to-noise ratios can be achieved. Confocal detection techniques are versatile and have been widely adopted for SMD in freely diffusing systems. Consequently, confocal methods will be discussed in detail in this article. The other general approach to performing SMD in solution involves the physical restriction of single molecules within defined volumes. Of particular note are techniques where single molecules are confined within a stream of levitated microdroplets. Droplet volumes are typically less than 1 fL and imaging of the entire microdroplet enables single molecule fluorescence to be contrasted against droplet 'blanks' with good signal-to-noise. Furthermore, since molecules are confined within discrete volumes, the technique can be utilized for high-efficiency molecular counting applications. More recently, spatial confinement of molecules in capillaries and microfabricated channels (with submicron dimensions) has been used to create probe volumes between 1 fL and 1 pL, and immobilized molecules on surfaces have been individually probed using wide-field microscopy with epi-illumination or evanescent wave excitation.

Single Molecule Detection using Confocal Microscopy

As previously stated, the confocal fluorescence microscope is an adaptable and versatile tool for SMD. In its simplest form, a confocal microscope is one in which a point light source, a point focus in the object plane, and a pinhole detector are all confocal with each other. This optical superposition generates superior imaging properties and permits definition of ultra-small probe volumes. The concepts behind a basic confocal microscope and its use in SMD are schematically illustrated in Figure 3. Coherent light (typically from a laser and tuned to an optical transition of the molecule under investigation) behaves as a point light source and is focused into a sample chamber using a high-numerical aperture objective lens. As a single molecule traverses the laser beam it is continuously cycled between the ground and an excited electronic state, emitting a burst of fluorescence photons. Fluorescence emission is isotropic (spontaneous emission), so photons are emitted in all directions (4π steradians).

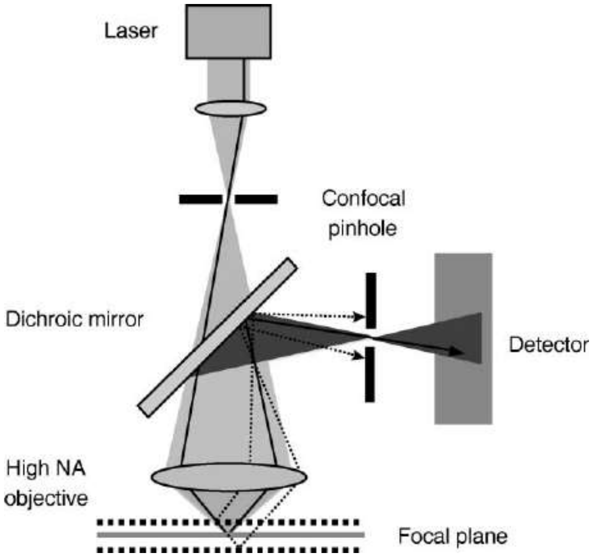

Figure 3 Principle of confocal detection. A confocal pinhole only selects light that emanates from the focal region. Dashed lines indicate paths of light sampled above and below the focal plane that are rejected by the pinhole. The solid ray derives from the focal point, and is transmitted through the pinhole to a detector.

Consequently, the high numerical aperture is used to collect as large a fraction of photons emitted from the focal plane as possible. Designing the objective to be used with an immersion medium, such as oil, glycerin, or water, can dramatically increase the objective numerical aperture, and thus the number of collected photons. Light collected by the objective is then transmitted towards a dichroic beam splitter. In the example shown, fluorescence photons (of lower energy) are reflected towards the confocal detector pinhole, whereas scattered radiation (of higher energy) is transmitted through the dichroic towards the light source. Creation of a precise optical probe volume is effected through the definition of the confocal pinhole. The detector is positioned such that only photons that pass through the pinhole are detected. Consequently, light emanating from the focal plane in the sample is transmitted through the pinhole and detected, whereas light not deriving from the focal plane is rejected by the aperture, and therefore not detected (Figure 3).

To ensure that the maximum number of photons are detected by the system, high efficiency detectors must be used. Photomultiplier tubes (the most common detectors for light-sensing applications) are robust and versatile but have poor detection efficiencies (approximately 5% of all photons that fall on the photocathode yield an electrical signal). Consequently, the most useful detectors for SMD (or low light level) applications are single photon-counting avalanche photodiodes (SPADs). A SPAD is essentially a p–n junction reverse biased above the breakdown voltage, that sustains an avalanche current when triggered by a photon-generated carrier. Detection efficiencies for typical SPADs are normally between 60–70% and are thus ideal for SMD in solution. An approximation of the overall detection efficiency of a confocal system for SMD can be made using eqn [2], which incorporates an estimation of photon losses at all stages of the collection/detection process. Typical transmission efficiencies for each step are also shown.

overall detection efficiency		objective collection efficiency		dichroic transmission coefficient		additional optical losses		detector efficiency
	$\approx$		$\times$		$\times$		$\times$	
0.06		0.24		0.9		0.5		0.6

[2]

Optical Probe Volumes

The precise nature of the probe volume is determined by the image of the pinhole in the sample and the spherical aberration of the microscope objective. Routinely, confocal probe volumes are approximated as resembling a cylinder with a radius defined by the diffraction-limited waist of a Gaussian beam. This approximation is useful when the incident beam is narrow or not tightly focused. However, when the radius of the incident beam is large, the corresponding diffraction limited focus is narrowed, and the probe volume more closely resembles a pair of truncated cones. Figure 4a illustrates the dependence of the curvature of the $1/e^2$ intensity contour on the collimated beam radius.

Consequently, it is clear that a simple cylindrical approximation for the probe volume breaks down for wide, tightly focused beams. If a broad incident beam (diameter >1.5 mm) is used, a large noncylindrical contribution to the probe volume is anticipated and a more appropriate model is required. An alternative and more accurate model for the confocal probe volume considers the Gaussian profile of the focused beam. The $1/e^2$ intensity contour radius of a Gaussian waveform with wavelength λ, at some distance z from the beam waist radius w_0, is given by eqn [3]:

$$w(z) = w_0\sqrt{1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2} \qquad [3]$$

In this case, the probe volume V is given by the volume of rotation of $w(z)$ around the z-axis between Z' and $-Z'$. The volume of rotation can therefore be

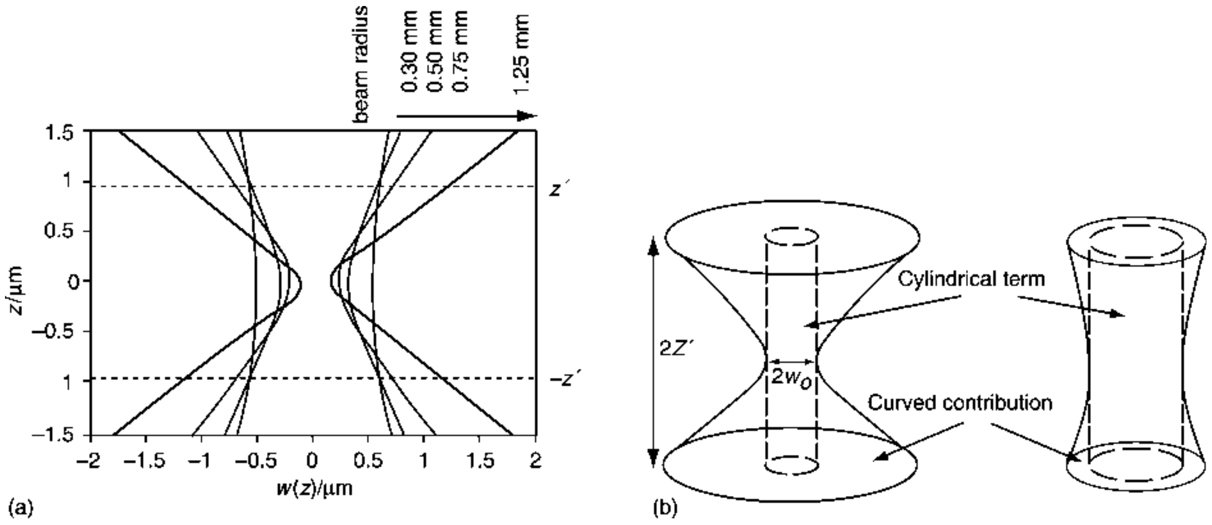

Figure 4 (a) $1/e^2$ Gaussian intensity contours plotted for a series of laser beam radii ($\lambda = 488$ nm, $f = 1.6$ mm, $n = 1.52$). (b) Cylindrical and curved components of the Gaussian probe volume. The curved contribution is more significant for larger beam radii and correspondingly tight beam waists.

simply defined according to,

$$V = \int_{-Z'}^{Z'} \pi w(z)^2 dz \qquad [4]$$

Solution of eqn [4] yields

$$V = 2\pi w_0^2 Z' + \frac{2\lambda^2}{3\pi w_0^2} Z'^3 \qquad [5]$$

The Gaussian volume expression contains two terms. The first term, $2\pi w_0^2 Z'$, corresponds to a central cylindrical volume; the second term has a more complex form that describes the extra curved volume (Figure 4b). The diffraction-limited beam waist radius w_0 can be defined in terms of the focusing objective focal length f, the refractive index n, and the collimated beam radius R according to

$$w_0 = \frac{\lambda f}{n\pi R} \qquad [6]$$

Substitution in eqn [5] yields,

$$V = 2\pi\left(\frac{\lambda f}{n\pi R}\right)^2 Z' + \frac{2\lambda^2}{3\pi}\left(\frac{n\pi R}{\lambda f}\right)^2 Z'^3$$

$$= \frac{2\lambda^2 f^2}{\pi n^2 R^2} Z' + \frac{2\pi n^2 R^2}{3 f^2} Z'^3 \qquad [7]$$

The volume is now expressed in terms of identifiable experimental variables and constants. Once again, the first term may be correlated with the cylindrical contribution to the volume, and the second term is the additional volume due to the curved contour. It is clear from Figure 4a that, for a given focal length, the noncylindrical contribution to the probe volume increases with incident beam diameter, when the diffraction limited focus is correspondingly sharp and narrow. Furthermore, it can also be seen that the second term in eqn [7] is inversely proportional to f^2, and thus the extent to which the probe volume is underestimated by the cylindrical approximation increases with decreasing focal length. This fact is significant when performing confocal measurements, since high numerical aperture objectives with short focal lengths are typical. Some realistic experimental parameters give an indication of typical dimensions for the probe volume in confocal SMD systems. For example, if $\lambda = 488$ nm, $f = 1.6$ mm, $Z' = 1.0$ μm and $n = 1.52$, a minimum optical probe volume of 1.1 fL is achievable with a collimated beam diameter of 1.1 mm.

Intensity Fluctuations: Photon Burst Statistics

When sampling a small volume within a system that may freely exchange particles with a large surrounding analyte bath, a Poisson distribution of particles is predicted. A Poisson distribution is a discrete series that is defined by a single parameter μ equating to the mean and variance of the distribution:

$$P(n = x) = \frac{\mu^x e^{-\mu}}{x!} \qquad [8]$$

Common Poisson processes include radioactive disintegration, random walks and Brownian motion. Although particle number fluctuations in the excitation volume are Poissonian in nature, the corresponding fluorescence intensity modulation induces a stronger correlation between photon counts. For a single molecular species the model is described by two parameters: an intrinsic molecular brightness and the average occupancy of the observation volume. A super-Poissonian distribution has a width or variance that is greater than its mean; in a Poisson distribution the mean value and the variance are equal. The fractional deviation Q is defined as the scaled difference between the variance and the expectation value of the photon counts, and gives a measure of the broadening of the photon counting histogram (PCH). Q is directly proportional to the molecular brightness factor ε and the shape factor γ of the optical point spread function. (γ is constant for a given experimental geometry.)

$$Q = \frac{\langle \Delta n^2 \rangle - \langle n \rangle}{\langle n \rangle} = \gamma\varepsilon \qquad [9]$$

A pure Poisson distribution has $Q = 0$; for super-Poissonian statistics $Q > 0$. Deviation from the Poisson function is maximized at low number density and high molecular brightness.

In a typical SMD experiment raw data are generally collected with a multichannel scalar and photons are registered in binned intervals. Figure 5 illustrates typical photon burst scans demonstrating the detection of single molecules (R-phycoerythrin) in solution. Fluorescence photon bursts, due to single molecule events, are clearly distinguished above a low background baseline (top panel) of less than 5 counts per channel in the raw data. It is noticeable that bursts vary in both height and size. This is in part due to the range of possible molecular trajectories through the probe volume, photobleaching kinetics, and the nonuniform illumination intensity in the probe region. In addition, it can be seen that the burst frequency decreases as bulk solution concentration is reduced. This effect is expected since the properties of any given single-molecule event are determined by molecular parameters alone (e.g., photophysical and diffusion constants) and concentration merely controls the frequency/number of events.

Although many fluorescence bursts are clearly distinguishable from the background, it is necessary to set a count threshold for peak discrimination in order to correctly identify fluorescence bursts above the background. A photocount distribution can be used as the starting point for determining an appropriate threshold for a given data set. The overlap between signal and background photocount distributions affects the efficiency of molecular detection. Figure 6 shows typical signal and background photocount probability distributions, with a threshold set at approximately 2 photocounts per channel. The probability of spurious (or false)

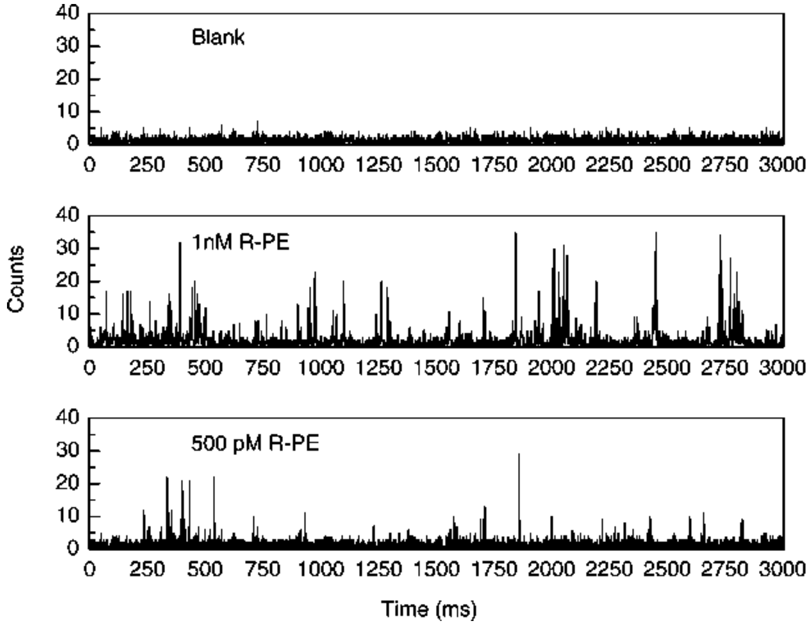

Figure 5 Photon burst scans originating from 1 nM and 500 pM R-phycoerythrin buffered solutions. Sample is contained within a 50 μm square fused silica capillary. Laser illumination = 5 μW, channel width = 1 ms. The top panel shows a similar burst scan originating from a deionized water sample measured under identical conditions.

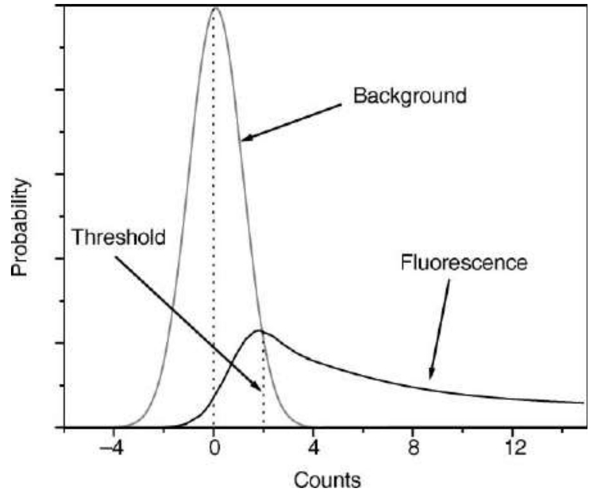

Figure 6 Simulated fluorescence and background photocount probability distributions. The vertical dashed line at 2 counts represents an arbitrarily defined threshold value for peak determination.

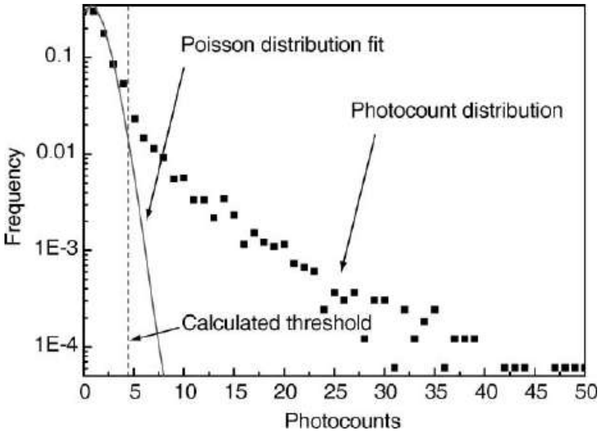

Figure 7 A photon counting histogram generated from a 16 second photon burst scan originating from a 10 μg/mL solution of 1000 nm fluorescent microbeads. The dotted curve shows a least-squares fit of early channels to a Poisson distribution, and the dashed vertical line marks the peak threshold (defined as $\mu + 3\sqrt{\mu} = 4.47$ counts).

detection resulting from statistical fluctuations in the background can be quantified by the area under the ‘background’ curve at photocount values above the threshold. Similarly, the probability that ‘true’ single molecule events are neglected can be estimated by the area under the ‘fluorescence’ curve at photocount values below the threshold. Choice of a high threshold value will ensure a negligible probability of calling a false positive, but will also exclude a number of true single molecule events that lie below the threshold value. Conversely, a low threshold value will generate an unacceptably high number of false positives. Consequently, choice of an appropriate threshold is key in efficient SMD.

Since the background shot noise is expected to exhibit Poisson statistics, the early part of the photocount distribution (i.e., the portion that is dominated by low, background counts) can be modeled with a Poisson distribution, to set a statistical limit for the threshold. Photon counting events above this threshold can be defined as photon bursts associated with the presence of single molecules. In an analogy with Gaussian systems the selected peak discrimination threshold can be defined as three standard deviations from the mean count rate:

$$n_{\text{threshold}} = \mu + 3\sqrt{\mu} \qquad [10]$$

Adoption of a threshold that lies 3 standard deviations above the mean yields a confidence limit that is typically greater than 99%. Figure 7 illustrates a sample photocount distribution, a least-squares fit to an appropriate Poisson distribution, and the calculated threshold that results. Once the threshold has been calculated, its value is subtracted from all channel counts and a peak search utility used to identify burst peaks in the resulting data set.

Data Filtering

As stated, the primary challenge in detecting single molecules in solution is not the maximization of the detected signal, but the maximization of the signal-to-noise ratio (or the reduction of background interferences). Improving the signal-to-noise ratio in such experiments is important, as background levels can often be extremely high.

Several approaches have been used to smooth SMD data with a view to improving signal-to-noise ratios. However, the efficacy of these methods is highly dependent on the quality of the raw data obtained in experiment. As examples, three common methods are briefly discussed. The first method involves the use of a weighted quadratic sum (WQS) smoothing filter. The WQS function creates a weighted sum of adjacent terms according to

$$\tilde{n}_{k,WQS} = \sum_{j=0}^{m-1} w_j (n_{k+j})^2 \qquad [11]$$

The range of summation m is the same order as the burst width, and the weighting factors w_j are chosen

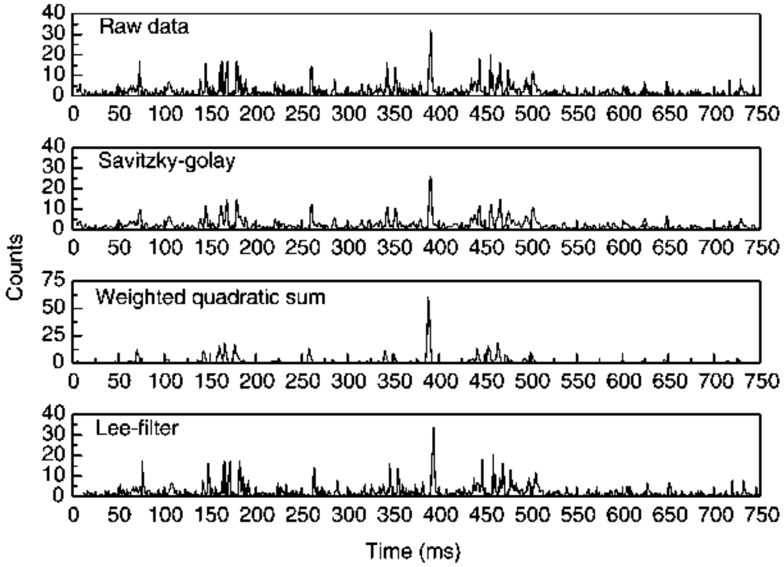

Figure 8 Effect of various smoothing filters on a 750 ms photon burst scan originating from a 1 nM R-phycoerythrin buffered solution. Raw data are shown in the top panel.

to best discriminate between single-molecule signals and random background fluctuations. This method proves most useful for noisy systems, in which the raw signal is weak. There is a practical drawback, in that peak positions are shifted by the smoothing function, and subsequent burst analysis is therefore hampered.

Another popular smoothing filter is the Lee-filtering algorithm. The Lee filter preferentially smoothes background photon shot noise and is defined according to

$$\tilde{n}_k = \bar{n}_k + (n_k - \bar{n}_k)\frac{\sigma_k^2}{\sigma_k^2 + \sigma_0^2} \qquad [12]$$

where the running mean (n_k) and running variance (σ_k^2) are defined by

$$\bar{n}_k = \frac{1}{(2m+1)}\sum_{j=-m}^{m} n_{k+j} \quad m < k \leq N - m \qquad [13]$$

$$\sigma_k^2 = \frac{1}{(2m+1)}\sum_{j=-m}^{m} (n_{k+j} - \bar{n}_{k+j})^2 \quad 2m < k \leq N - 2m \qquad [14]$$

for a filter $(2m+1)$ channels wide. Here, n_k is the number of detected photons stored in a channel k, σ_0 is a constant filter parameter, and N is the total number of data points.

A final smoothing technique worth mentioning is the Savitsky-Golay smoothing filter. This filter uses a least-squares method to fit an underlying polynomial function (typically a quadratic or quartic function) within a moving window. This approach works well for smooth line profiles of a similar width to the filter window and tends to preserve features such as peak height, width and position, which may be lost by simple adjacent averaging techniques. Figure 8 shows the effects of using each approach to improve signal-to-noise for raw burst data.

Photon Burst Statistics

A valuable quantitative analysis method for analysis of fluorescence bursts utilizes the analysis of Poisson statistics. Burst interval distributions are predicted to follow a Poissonian model, in which peak separation frequencies adopt an exponential form. The probability of a single molecule (or particle) event occurring after an interval Δt is given by eqn [15]:

$$N(\Delta t) = \lambda \exp(-\beta t) \qquad [15]$$

where λ is a proportionality constant and β is a characteristic frequency at which single molecule events occur. The recurrence time τ_R can then be simply defined as

$$\tau_R = \frac{1}{\beta} \qquad [16]$$

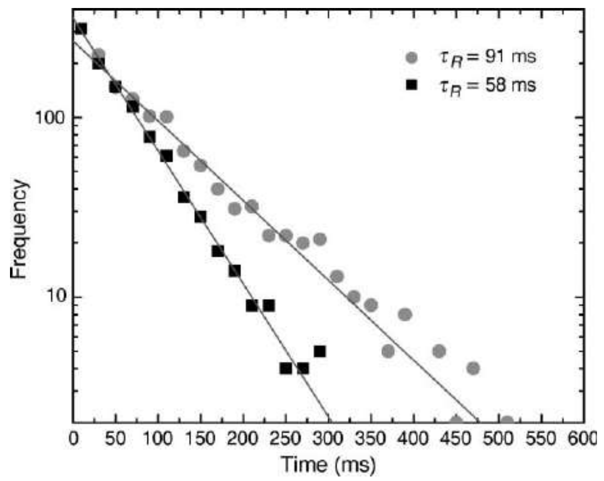

Figure 9 Burst interval distribution analysis of photon burst scans. Data originate from 1 μm fluorescent beads moving through 150 μm wide microchannels at flow rates of 200 nL min^{-1} (circles) and 1000 nL min^{-1} (squares). Least squares fits to a single exponential function are shown by the solid lines.

Equation [6] simply states that longer intervals between photon bursts are less probable than shorter intervals. Furthermore, the recurrence time reflects a combination of factors that control mobility, probe volume occupancy, or other parameters in the single molecule regime. Consequently, it is expected that τ_R should be inversely proportional to concentration, flow rate or solvent viscosity in a range of systems. Figure 9 shows an example of frequency $N(\Delta t)$ versus time plots for two identical particle systems moving at different velocities through the probe volume. A least-squares fit to a single exponential function yields values of $\tau_R = 91$ ms for a volumetric flow rate of 200 nL/min and $\tau_R = 58$ ms for a volumetric flow rate of 1000 nL/min.

Temporal Fluctuations: Autocorrelation Analysis

Autocorrelation analysis is an extremely sensitive method for detecting the presence of fluorescence bursts in single molecule experiments. This approach essentially measures the average of a fluctuating signal as opposed to the mean spectral intensity. As previously discussed, the number of molecules contained within a probe volume at any given time is governed by Poisson statistics. Consequently, the root mean square fluctuation can be defined according to eqn [17],

$$\frac{\sqrt{\langle(\delta N)^2\rangle}}{\langle N\rangle} = \frac{\sqrt{\langle(N-\langle N\rangle)^2\rangle}}{\langle N\rangle} = \frac{1}{\sqrt{\langle N\rangle}} \qquad [17]$$

where N is the number of molecules. It is observed that the relative fluctuation diminishes as the number of particles measured is increased. Hence, it is important to minimize the number of molecules present in the probe volume. It should be noted, however, that if there are too few molecules in the solution there will be long dark periods were no single molecule bursts are observed. If the probe volume is bathed in radiation of constant intensity, fluctuation of the resulting fluorescence signal can simply be defined as deviations from the temporal signal average:

$$\langle F(t)\rangle = \frac{1}{T}\int_0^T F(t)\mathrm{d}t \qquad [18]$$

Here, t is defined as the total measurement time, $F(t)$ is the fluorescence signal at time t, and $\langle F(t)\rangle$ is the temporal signal average. Fluctuations in the fluorescence intensity, $\delta F(t)(\delta F(t) = F(t) - \langle F(t)\rangle)$, with time t, about an equilibrium value $\langle F\rangle$, can be statistically investigated by calculating the normalized autocorrelation function, $G(\tau)$, where

$$G(\tau) = \frac{\langle F(t+\tau)F(t)\rangle}{\langle F\rangle^2} = \frac{\langle \delta F(t+\tau)\delta F(t)\rangle}{\langle F\rangle^2} + 1 \qquad [19]$$

In dedicated fluorescence correlation spectroscopy experiments, the autocorrelation curve is usually generated in real time in a high-speed digital correlator. Post data acquisition calculation is also possible using the following expression

$$G(\tau) = \sum_{t=0}^{N-1} g(t)g(t+\tau) \qquad [20]$$

Here $g(t)$ is the total number of counts during the time interval $(t, t+\Delta t)$, $g(t+\tau)$ is the number of counts detected in an interval of Δt at a later time $t+\tau$, and N is the total number of time intervals in the dataset. In a diffusion controlled system with a single fluorescent molecule that is irradiated with a three dimensional Gaussian intensity profile, the autocorrelation curve is governed by the mean probe volume occupancy N and the characteristic diffusion time (τ_D). The laser beam waist radius ω and the probe depth $2z$ describe the Gaussian profile:

$$G(\tau) = 1 + \frac{1}{N}\left(1 + \frac{\tau}{\tau_d}\right)^{-1}\left(1 + \left(\frac{\omega}{z}\right)^2 \frac{\tau}{\tau_d}\right)^{-1/2} \qquad [21]$$

The diffusion time is a characteristic molecular residence time in the probe volume and inversely

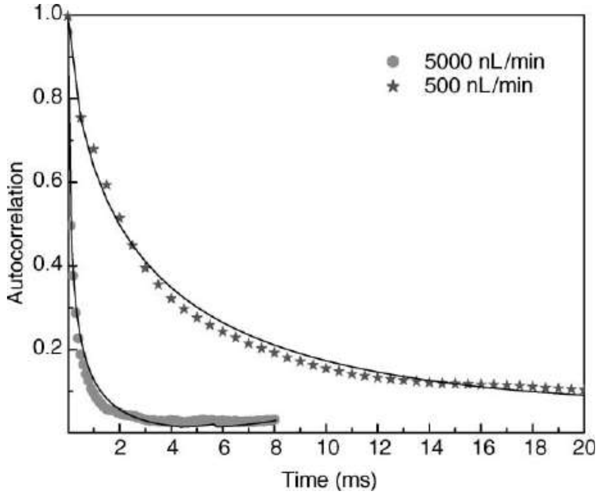

Figure 10 Autocorrelation analysis of photon burst scans of 1 μm fluorescent beads moving through 150 μm wide microchannels at flow rates of 500 nL min^{-1} (stars) and 1000 nL min^{-1} (circles). Solid lines represent fits to the data according to eqn [23].

related to the translational diffusion coefficient for the molecule:

$$D = \frac{w^2}{4\tau_D} \qquad [22]$$

In a flowing system, the autocorrelation function depends on the average flow time through the probe volume τ_{flow}. A theoretical fit to the function can be described according to

$$G(\tau) = 1 + \frac{1}{N} A \exp\left\{\left(\frac{\tau}{\tau_{\text{flow}}}\right)^2 A\right\}$$
$$A = \left(1 + \frac{\tau}{\tau_d}\right)^{-1}\left(1 + \left(\frac{\omega}{z}\right)^2 \frac{\tau}{\tau_d}\right) \qquad [23]$$

where N is the mean probe volume occupancy; the flow velocity v can then be extracted from the characteristic flow time according to

$$v = \frac{w}{\tau_{\text{flow}}} \qquad [24]$$

It should be noted that in the case that directed flow is negligible or defined to be zero, the autocorrelation function simplifies to eqn [21].

Figure 10 illustrates experimentally determined autocorrelation curves for two identical particle systems moving at different velocities through the probe volume. As particle flow velocity is increased, the width of the autocorrelation curves is seen to narrow as a result of the reduced residence time in the probe volume. A plot of the reciprocal of the full width half maximum of the autocorrelation curve as a function of volumetric flow rate is linear, and provides a simple way of calculating particle/molecule velocities within flowing systems.

Applications

The basic tools and methods outlined in this chapter have been used to perform SMD in a variety of chemically and biologically relevant systems, and indeed there is a large body of work describing the motion, conformational dynamics and interactions of individual molecules (see Further Reading). A primary application area has been in the field of DNA analysis, where SMD methods have been used in DNA fragment sizing, single-molecule DNA sequencing, high-throughput DNA screening, single-molecule immunoassays, and DNA sequence analysis. SMD methods have also proved highly useful in studying protein structure, protein folding, protein-molecule interactions, and enzyme activity.

More generally, SMD methods may prove to be highly important as a diagnostic tool in systems where an abundance of similar molecules masks the presence of distinct molecular anomalies that are markers in the early stages of disease or cancer.

See also

Microscopy: Confocal Microscopy.

Further Reading

Ambrose WP, Goodwin PM, Jett JH, Van Orden A, Werner JH and Keller RA (1999) Single molecule fluorescence spectroscopy at ambient temperature. *Chem. Rev.* 99: 2929–2956.

Barnes MD, Whitten WB and Ramsey JM (1995) Detecting single molecules in liquids. *Anal. Chem.* 67: 418A–423A.

Basche T, Orrit M and Rigler R (2002) *Single Molecule Spectroscopy in Physics, Chemistry, and Biology.* New York: Springer-Verlag.

Zander CJ, Enderlein J and Keller RA (2002) *Single Molecule Detection in Solution: Methods and Applications*. New York: John Wiley & Sons.

Diffuse-Reflectance Laser Flash Photolysis

D R Worrall and S L Williams, Loughborough University, Loughborough, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Flash photolysis was developed as a technique to study short-lived intermediates in photoinduced reactions by George Porter (later Lord Porter of Luddenham) and Ronald Norrish in 1949, drawing on Porter's wartime experience with radar techniques. Such was the impact of this development that it earned the Nobel prize jointly for Porter, Norrish, and Eigen in 1967. In this article we will describe the application of flash photolysis to opaque, scattering samples, detailing how light propagation in such samples can be treated theoretically, and will discuss methods by which the data obtained from diffuse reflectance flash photolysis experiments may be analyzed.

The technique of flash photolysis was originally based on using an intense flash of light (the photolysis flash) from a xenon tube to excite the sample, followed a certain time delay later by a spectroscopic flash of lower intensity from a second flash tube, the light from the latter being detected using a photographic plate. The photolysis flash is of sufficient intensity to produce a large population of intermediates (radicals, ions, excited states, isomers) in the sample, which then absorb light from the spectroscopic flash depending on the concentration of intermediates according to the Beer–Lambert law:

$$A = \varepsilon c l \qquad [1]$$

where A is the sample absorbance, ε the molar absorption coefficient, c the concentration and l the pathlength. The absorbance is related to the incident and transmitted intensities as

$$A = \log_{10} I_0 / I \qquad [2]$$

With I_0 the incident and I the transmitted intensities, there is therefore an exponential fall-off of intensity with pathlength for a homogeneous absorber. Hence, by monitoring the evolution of the absorption spectra, the changes in concentration of the photoproduced intermediates and hence the kinetics of the processes in which they are involved are elucidated.

Flash photolysis has evolved subsequently to make use of laser sources and sophisticated electronic detection apparatus to push the limits of time resolution to the femtosecond regime. Indeed, recently the Nobel prize for chemistry was awarded to Ahmed Zewail for his work with ultrafast pump-probe techniques. However, in its conventional geometry, flash photolysis is limited to transparent samples, since it is necessary to be able to probe the excited species by monitoring the absorption spectra.

Many biological systems and industrially important samples are opaque or highly scattering, and hence the attenuation of light through the sample is no longer described by the Beer–Lambert law. In 1981, Frank Wilkinson and Rudolph Kessler had the idea of using diffusely reflected light to interrogate the changes in concentration within a scattering sample subjected to a high-intensity excitation pulse. When photons enter a sample, they may be absorbed or scattered. Those which are scattered may re-emerge from the irradiated surface as diffusely reflected light. The intensity of diffusely reflected light emerging from the surface at a particular wavelength is a unique function of the ratio of scattering to absorption. The more scattering events occurring before absorption, the more likely the photon is to escape from the sample as diffusely reflected light. Hence the probability of escape decreases as the absorption probability increases, and the diffusely reflected light is deficient in those wavelengths where the absorption is strongest, i.e., the ratio of the incident to absorbed light intensity at a given wavelength is related to the absorption of the sample at that wavelength.

Kubelka–Munk Theory of Reflectance

The theory which describes the relationship between incident and scattered light intensity, absorption and scatter, and concentration which is widely applied in this context, is the Kubelka–Munk theory of reflectance. The theory was originally developed to describe the reflectance characteristics of paint films but it works quite well for many samples containing a homogeneous distribution of scatterers and absorbers. The limiting assumption in this theory is that the scatterers from which the scattering layer is composed are very much smaller than the total layer thickness. Additionally, the layer should be optically thick such that all of the light entering the layer should be either absorbed or reflected, with a negligible fraction transmitted.

For a layer of thickness X diffusely irradiated with monochromatic light, the diagram shown in Figure 1

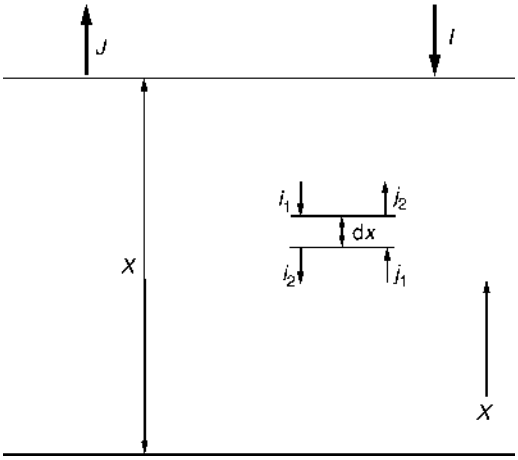

Figure 1 Schematic of counterpropagating fluxes in a diffusely irradiated scattering medium.

can be constructed, with I the incident flux and J the diffusely reflected flux, and i and j the fluxes traveling upwards and downwards through an infinitesimally small thickness element dx. Two further parameters may be defined which are characteristic of the medium.

K The absorption coefficient. Expresses the attenuation of light due to absorption per unit thickness.
S The scattering coefficient. Expresses the attenuation of light due to scattering per unit thickness.

Both of these parameters can be thought of as arising due to the particle (or chromophore) acting as a sphere of characteristic size, which casts a shadow either due to the prevention of on-axis transmission or due to absorption. The scattering or absorption coefficient then depends on the effective size of this sphere, and the number density of spheres in the medium. In each case, the probability P of a photon being transmitted through a particular thickness X of a medium is related exponentially to the absorption or scattering coefficient:

$$P = \exp(-KX) \tag{3}$$

$$P = \exp(-SX) \tag{4}$$

The scattering coefficient S depends on the refractive index difference between the particle and the dispersion medium. The scattering coefficient is also dependent upon the particle size, and shows an inverse correlation, i.e., the scattering coefficient increases as the particle size decreases. This effect is a function of the size of the particle relative to the wavelength of light impinging on it, with the scattering coefficient increasing as the wavelength decreases. However, this change with wavelength is small provided the particle size is large relative to the wavelength of the light.

The effect of the material in the element dx on the counterpropagating fluxes i and j depends on the absorption and scattering coefficients. Both i and j will be attenuated by both absorption and scattering:

$$i_2 = i_1 - i_1(S + K)\mathrm{d}x \tag{5}$$

$$j_2 = j_1 - j_1(S + K)\mathrm{d}x \tag{6}$$

Both i and j are reinforced by backscattering from the other flux:

$$i_2 = i_1 + j_1 S\mathrm{d}x \tag{7}$$

$$j_2 = j_1 + i_1 S\mathrm{d}x \tag{8}$$

The net effect of this in the attenuation of the flux propagating into the sample (i) and the flux backscattered from the sample (j) can be expressed as the following differential equations:

$$\mathrm{d}i = (S + K)i_1\mathrm{d}x - j_1 S\mathrm{d}x \tag{9}$$

$$\mathrm{d}j = -(S + K)j_1\mathrm{d}x + i_1 S\mathrm{d}x \tag{10}$$

For a sample of infinite thickness, these equations can be solved to give an analytical solution for the observed reflectance of the sample in terms of the absorption and scattering coefficients:

$$\frac{K}{S} = \frac{(1 - R_\infty)^2}{2R_\infty} \tag{11}$$

where R_∞ is the reflectance from a sample of such thickness that an increase in sample thickness has no effect on the observed reflectance. The absorption coefficient K is dependent upon the concentration of absorbers in the sample through

$$K = 2\varepsilon c \tag{12}$$

with ε the naperian absorption coefficient, c the concentration of absorbers and the factor 2 which is the geometrical factor for an isotropic scatterer. For multiple absorbers, K is simply the linear sum of absorption coefficients and concentrations. Hence for a diffusely scattering medium, an expression analogous to the Beer–Lambert law can be derived to relate concentration to a physically measurable parameter, in this case the sample reflectance. The ratio K/S is usually referred to as the Kubelka–Munk remission function, and is the parameter usually quoted in this context. It is important to note that the relationship with concentration is only valid for a homogeneous distribution of absorbers within a sample.

Transient Concentration Profiles

As has been discussed in the previous section, a light flux propagating through a scattering medium is attenuated by both scattering and absorption events, whilst in a nonturbid medium attenuation is by absorption only. Hence even in a sample with a very low concentration of absorbers, the light flux is rapidly attenuated as it penetrates into the sample by scattering events. This leads to a significant flux gradient through the sample, resulting in a reduction in the concentration of photo-induced species as penetration depth into the sample increases. There are three distinct concentration profiles which may be identified within a scattering sample which are dependent on the scattering and absorption coefficients. These concentration profiles are interrogated by a beam of analyzing light, and hence an understanding of the effect of these differing profiles on the diffusely reflected intensity is vital in interpreting transient absorption data. The transient depth profiles are illustrated in Figure 2.

Kubelka–Munk Plug

This occurs when the photolysis flash, i.e., laser fluence is high and the absorption coefficient is low at the laser wavelength. If we assume a simple photophysical model involving simply the ground state S_0, the first excited singlet state S_1 and first excited triplet state T_1, and we assume that either the quantum yield of triplet state production is high or the S_1 lifetime is very much shorter than the laser pulse allowing re-population of the ground state, then it is possible at high enough fluence to completely convert all of the S_0 states to T_1 states within the sample. Provided the fluence is high enough, this complete conversion will penetrate some way into the sample (Figure 2). Under circumstances where the T_1 state absorbs strongly at some wavelength other than the photolysis wavelength, probing at this wavelength will result in the probe beam being attenuated significantly within a short distance of the surface of the sample, and thus will only interrogate regions where there is a homogeneous excited state concentration. Under these circumstances the reflectance of the sample as a function of the concentration of excited states is described by the Kubelka–Munk equation, and the change in remission function can be used to probe the change in excited state concentration.

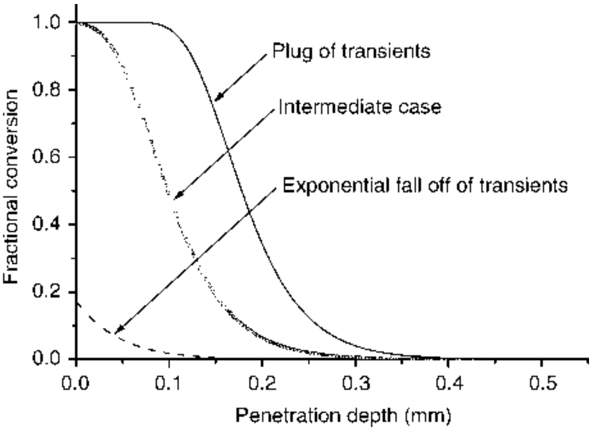

Figure 2 Transient concentration profiles following pulsed laser excitation of a scattering and absorbing sample.

Exponential Fall-Off of Concentration

This occurs when the laser fluence is low and the absorption coefficient at the laser wavelength high. Again considering the simple photophysical model, most of the laser flux will be absorbed by the ground state in the first few layers of sample, and little will penetrate deeply into the sample. In the limiting case this results in an exponential fall-off of transient concentration with penetration depth into the sample (Figure 2). Here the distribution of absorbers is not random and the limiting Kubelka–Munk equation is no longer applicable since the mean absorption coefficient varies with sample penetration depth. Lin and Kan solved eqs [9] and [10] with the absorption coefficient K varying exponentially with penetration depth and showed that the series solution converges for changes in reflectance of less than 10% such that the reflectance change is a linear function of the number of absorbing species.

Intermediate Case

Between the two extremes described above is a case where significant transient conversion takes place at the front surface, but with little penetration into the sample. This can occur, for example, with high laser fluences and large ground state absorption coefficients at the laser wavelength. Under these circumstances, illustrated in Figure 2, the analyzing light interrogates not only the transient concentration profile, but also a significant amount of the analyzing light may penetrate through the transient layer into the unconverted sample behind, if the transient absorption at the analyzing wavelength is low. This creates a more complex problem for analysis since effectively the sample is irradiated from both front and back faces, with consequent effects on the diffusely reflected intensity. It is possible to numerically model the reflectance change as a function of transient concentration under these circumstances but a precise knowledge of the absorption and scattering coefficients is required. Under most circumstances, this case is avoided and diffuse reflectance flash photolysis experiments are arranged such that one of the two limiting cases above, generally the exponential fall-off (usually achieved by attenuation of the laser beam), prevails in a particular experiment.

Sample Geometry

In the case of conventional nanosecond laser flash photolysis, it is generally the case that right-angled geometry is employed for the photolysis and probe beams (Figure 3a). This geometry has a number of advantages over alternatives. The photolysis beam and analyzing beam are spatially well separated, such that the analyzing beam intensity is largely unaffected by scattered photolysis light. Also, the fluorescence generated by the sample will be emitted in all directions, whilst the analyzing beam is usually collimated allowing spatial separation of fluorescence and analyzing light; sometimes an iris is used to aid this spatial discrimination. This geometry is appropriate for quite large beam diameters and fluences; where smaller beams are used collinear geometry may be more appropriate in order to achieve long interaction pathlengths.

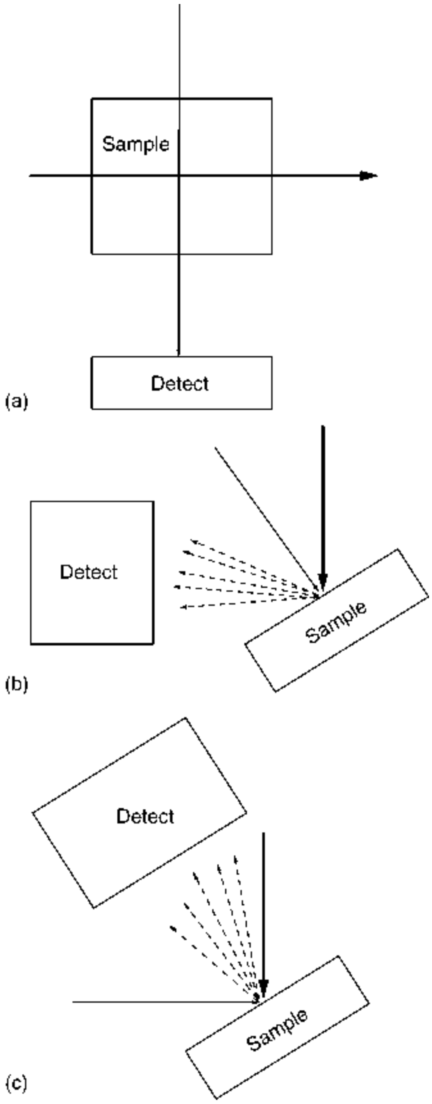

Figure 3 Sample geometries for (a) conventional and (b,c) diffuse reflectance flash photolysis. ——→, photolysis beam; ——→, analyzing beam.

In the case of nanosecond diffuse reflectance flash photolysis, the geometry required is quite different (Figure 3b,c). Here, the photolysis beam and analyzing beam must be incident on the same sample surface, and the diffusely reflected analyzing light is collected from the same surface. The geometry is often as shown in Figure 3b, where the analyzing light is incident almost perpendicularly on the sample surface, with the photolysis beam incident at an angle such that the specular reflection of the photolysis beam passes between detector and analyzing beam (not shown). Alternatively, the geometry shown in Figure 3c may be employed, where diffusely reflected light is detected emerging perpendicular to the sample surface. In both cases, the geometry is chosen such that specular (mirror) reflection of either exciting or analyzing light from the sample is not detected, since specular reflection is light which has not penetrated the sample and therefore contains no information regarding the concentrations of species present. A requirement, as in conventional flash photolysis, is that the analyzing beam probes only those areas which are excited by the photolysis beam, requiring the latter to be larger than the former. The nature of the scattering described previously means that the required geometry does not give spatial discrimination at the detector between photolysis and analyzing light, and fluorescence. This is since both analyzing and photolysis light undergo scattering and absorption processes (although with wavelength-dependent absorption and scattering coefficients) and emerge with the same spatial profiles. Fluorescence, principally that stimulated by the photolysis beam since this is of greatest intensity, originates from within the sample but again undergoes absorption and scattering and emerges with the same spatial distribution as the exciting light. In diffuse reflectance flash photolysis, separation of the analyzing light from the excitation or fluorescence must be achieved using spectral (filters and/or monochromators) or temporal, rather than spatial discrimination. Time-gated charge coupled devices

(CCD) or photodiode array detectors can be effectively used in nanosecond laser flash photolysis to exclude excitation light and fluorescence, since these occur on time-scales usually much shorter than the transient absorption of species of interest. The usual analyzing light source used in nanosecond diffuse reflectance flash photolysis is a xenon arc lamp, due to its good spectral coverage and high intensity. High probe intensity is particularly important in diffuse reflectance flash photolysis since the scattered intensities are often low, and the scattered light emerges in a large solid angle, only part of which can be effectively collected and detected. Also, high intensities allow for the light from the analyzing source to dominate over fluorescence if the latter is relatively weak.

In femtosecond diffuse reflectance laser flash photolysis, sample geometry considerations are also important. Such experiments are performed using the pump-probe technique, with the probe often being a femtosecond white-light continuum. The sample geometry employed is illustrated in Figure 4.

Here the pump and probe are almost colinear, and are incident on the same sample area; again, the requirement is for the pumped area to be larger than the probed area. Diffusely reflected light is then collected and analyzed, time resolution being achieved by varying the delay between pump and probe beams. It should be noted that the temporal resolution is worse than in conventional pump-probe techniques. In conventional femtosecond flash photolysis, the time resolution is generally limited by the widths of pump and probe pulses; in diffuse reflectance mode, the pulses undergo numerous refractions, reflections, and diffractions such that the pulse is temporally broadened during its transit through the material. The extent of this broadening is a sensitive function of the optical properties of the individual sample.

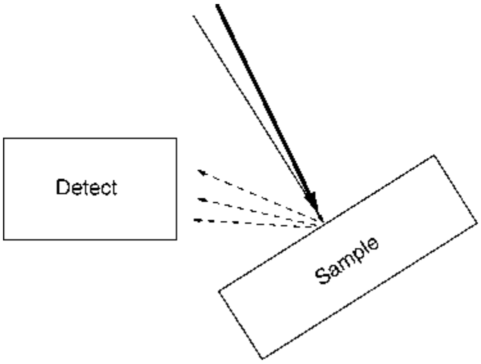

Figure 4 Sample geometry for femtosecond diffuse reflectance flash photolysis. ⟶, photolysis beam; →, analyzing beam.

Kinetic Analysis

Kinetic analysis, and of course time-resolved spectroscopic analysis, require a quantitative treatment of the concentration changes within a sample following an excitation pulse as a function of time. When studying transient absorption phenomena in opaque samples, it is usual to define the reflectance change in terms of the sample reflectance before and after excitation, such that in spectrally resolved data a transient difference spectrum rather than an absolute reflectance spectrum of the transient species is obtained. The latter can, however, be reconstructed from a knowledge of the absorption coefficients and concentration of the species involved. It is possible to define the reflectance change as

$$\frac{\Delta J_t}{J_0} = 1 - \frac{R_t}{R_0} \qquad [13]$$

where R_0 and R_t represent the intensity of probe light diffusely reflected from the sample before excitation and at a time t after excitation, respectively. Frequently reflectance change is expressed as '% absorption', which is defined in eqn [14].

$$\%\ \text{absorption} = 100 \times \left(1 - \frac{R_t}{R_0}\right) \qquad [14]$$

These parameters are often used as being proportional to transient concentration, subject to satisfying the criteria for an exponential fall-off of transient concentration with penetration depth as discussed previously, and are used to replace transient concentration in kinetic equations used in data analysis.

It is generally the case that the samples studied using diffuse reflectance laser flash photolysis have either some degree of heterogeneity, for example paper, microcrystalline cellulose, silica gel or alumina, or have well-defined porous structures such as zeolites. Molecules adsorbed to these supports may be present on the surface, within micro- or mesopores, or intimately included within the structure. Hence each molecule may experience its own unique environment, and this will obviously influence its observed photophysics. It is therefore the case that in these systems, even very simple photo-induced reactions such as unimolecular photoisomerizations do not follow first-order kinetics; rather, a distribution of rates is observed which reflect the differing environments experienced by the molecules, and hence the molecules act as probes of this heterogeneity.

There are a number of approaches to kinetic analysis in opaque, often heterogeneous systems, as described below.

Rate Constant Distributions

Here the sample heterogeneity is treated as producing a series of micro-environments, and the reaction studied will have its own unique rate constant in each of these environments. The width of the observed rate constant distribution therefore describes the number of these possible environments, with the distribution amplitude at any particular rate constant, reflecting the probability of the molecule existing in the corresponding environment.

Exponential series lifetime distribution analysis is the simplest approach to distributed lifetime analysis. This analysis has the inherent advantage that there are no presuppositions regarding the kinetic model describing the data. Rather, a large number of exponentials with fixed rate constants and with amplitudes as adjustable parameters are applied to the data, and a least-squares procedure used to optimize the amplitude of each exponential. Generally there is a relationship between the rate constants, they being equally spaced on either a linear or logarithmic scale. Physically this can be interpreted as a large number of first-order or pseudo-first-order (in the case of bimolecular reactions) rates arising from the sample heterogeneity. Hence the rate constant distribution emerges naturally from the fitting procedure, with no preimposed constraints. Since there are a large number of adjustable parameters, these models are applied most successfully to data with good signal-to-noise.

Alternative approaches involve imposing a distribution shape onto a set of exponentials, and optimizing this distribution to the data. This approach has the advantage that by assuming a rate constant distribution, the relative amplitudes of the exponentials are fixed and hence the number of fitting parameters greatly reduced. One of the more successful models applied in this context is that developed by Albery *et al.* In the development of their model it is assumed that, for the reaction in question, the free energy of activation, $\Delta G^{\neq}$, is distributed normally around a mean value $\Delta\bar{G}^{\neq}$ according to eqn [15]:

$$\Delta G^{\neq} = \Delta\bar{G}^{\neq} - \gamma x RT \qquad [15]$$

with γ being the width of the distribution for $\infty \geq x \geq -\infty$. This assumed normal distribution of the free energy of activation leads to a log-normal distribution of the decay rate constants distributed around some average rate constant $\bar{k}$. The dispersion in the first-order rate constant is then

$$\ln(k) = \ln(\bar{k}) + \gamma x \qquad [16]$$

The equation used to describe the data is given as:

$$\frac{c}{c_0} = \frac{\int_{-\infty}^{\infty} \exp(-t^2)\exp\Big[-\bar{k}t\exp(\gamma t)\Big]\mathrm{d}t}{\int_{-\infty}^{\infty} \exp(-t^2)\mathrm{d}t} \qquad [17]$$

with c and c_0 being concentration at time $t = t$ and $t = 0$ relative to the excitation flash, respectively. Where reflectance change is proportional to transient concentration, reflectances can be used directly in place of concentrations. This equation can be solved by making the following substitutions:

$$\int_{-\infty}^{\infty} \exp(-t^2)\mathrm{d}t = \pi^{1/2} \qquad [18]$$

$$\int_{-\infty}^{\infty} \exp(-t^2)\exp\Big[-\bar{k}t\exp(\gamma t)\Big]\mathrm{d}t = \int_0^1 \lambda^{-1}\exp\Big(-(\ln\lambda^2)\Big)\Big(\exp(-\bar{k}t\lambda^{\gamma})+\exp(-\bar{k}t\lambda^{-\gamma})\Big)\mathrm{d}\lambda \qquad [19]$$

Hence here there are only two fitting parameters, namely the width of the distribution γ and the distribution center, $\bar{k}$. Note that for $\gamma = 0$, eqn [17] reduces to a single exponential function.

A further model which constrains the rate constant distribution, which has been successfully applied to describe the rates of photo-induced processes on silica gel surfaces, is a Lévy stable distribution of rate constants, described as:

$$\mathrm{P}(\bar{k}, \alpha, \gamma) = \frac{1}{\pi}\int_0^{\infty} \exp(-\gamma q^{\alpha})\cos(\bar{k}\mathrm{q})\mathrm{dq} \qquad [20]$$

where α is the characteristic power law exponent ($0 < \alpha \leq 2$) and γ is the distribution width ($\gamma > 0$). Special cases of the Lévy stable distribution occur for $\alpha = 1$ and $\alpha = 2$, where the distribution shape becomes Lorentzian and Gaussian, respectively. The Lévy distribution gives an increasing weighting to the tails of the distribution as the distribution width decreases, and can be described as a random walk consisting of long jumps followed by several short walks. It has been shown that this type of foraging behavior is more efficient at seeking out randomly distributed targets than a simple random walk. The Lévy stable distribution has three adjustable parameters which allows greater flexibility in the distribution of rate constants than does the Gaussian model, but still constrains the distribution to be symmetrical about some mean value.

Physical Models

A number of these have been applied to very specific data sets, where parameters of the systems are accurately known. These include random walk models in zeolites, and time-dependent fractal-dimensional rate constants to describe kinetics on silica gel surfaces. The advantage of these methods over rate constant distributions is that since they are based on the physical description of the system, they yield physically meaningful parameters such as diffusion coefficients from analysis of kinetic data. However, for many systems the parameters are known in insufficient detail for accurate models to be developed.

Examples

An example of bimolecular quenching data is shown in Figure 5. Here, anthracene at a concentration of 1 μmol g^{-1} is co-adsorbed to silica gel from acetonitrile solution together with azulene at a concentration of 0.8 μmol g^{-1}. Laser excitation at 355 nm from an Nd:YAG laser produces the excited triplet state of the anthracene, which undergoes energy transfer to the co-adsorbed azulene molecule as a result of the rapid diffusion of the latter.

The data shown in Figure 5 are recorded monitoring at 420 nm, and the laser energy (approximately 5 mJ per pulse) is such that an exponential fall-off of transient concentration with penetration depth is expected such that reflectance change is proportional to transient concentration (see section on Transient Concentration Profiles above). The data are shown on a logarithmic time axis for clarity. The fitted line is obtained by applying the model of Albery *et al.* (see section on Rate Constant Distributions above) with fitting parameters $\bar{k} = 1.01 \times 10^4\ s^{-1}$ and $\gamma = 0.73$.

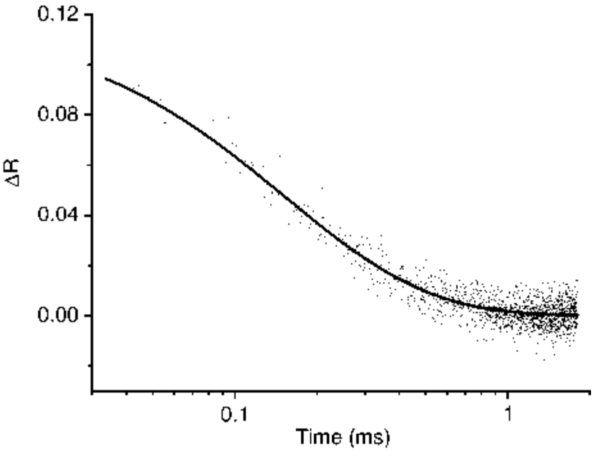

Figure 5 Transient absorption decay for anthracene (1 μmol g^{-1}) co-adsorbed with azulene (0.8 μmol g^{-1}) to silica gel monitored at 420 nm. Fitted using the model of Albery *et al.*

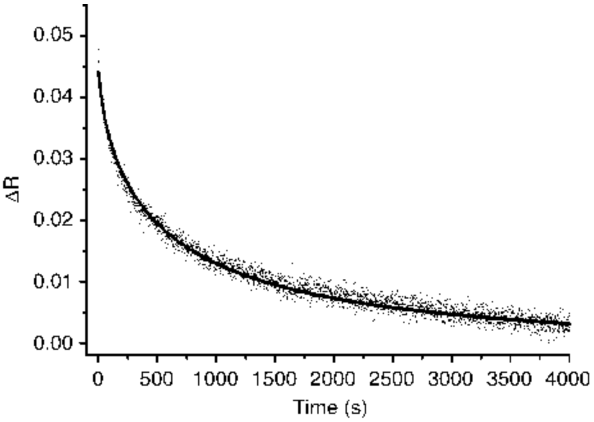

Figure 6 Transient absorption decay for the naphthalene (1 μmol g^{-1}) radical cation monitored at 680 nm on silica gel. Fitted using a Lévy stable distribution.

Where ion–electron recombination is concerned, the Albery model often fails to adequately describe the data obtained since it does not allow sufficient small rate constants relative to the value of $\bar{k}$, given the constraints of the Gaussian distribution. The Lévy stable distribution is ideal in this application since it allows greater flexibility in the shape of the distribution.

Figure 6 shows example data for naphthalene adsorbed to silica gel (1 μmol g^{-1}). The laser pulse energy at 266 nm (approximately 40 mJ per pulse) is such that photo-ionization of the naphthalene occurs producing the radical cation. The subsequent decay of the radical cation via ion–electron recombination can be monitored at 680 nm. Note that decay is observed on a time-scale of several thousand seconds. The fitted line is according to a Lévy stable distribution with parameters $\bar{k} = 8.2 \times 10^{-4}\ s^{-1}$, $\gamma = 0.5$ and $\alpha = 1.7$ (see section on Rate Constant Distributions above). This model allows the shape of the distribution to deviate from a Gaussian, and can be more successful than the model of Albery *et al.* in modeling the complex kinetics which arise on surfaces such as silica gel. Note that where the model of Albery *et al.* can successfully model the data, the data can also be described by a Lévy stable distribution with $\alpha = 2$.

List of Units and Nomenclature

A	Sample absorbance
c	Concentration
I	Incident flux
J	Diffusely reflected flux

k	Rate constant
$\bar{k}$	Mean rate constant
K	Absorption coefficient (absorption per unit thickness) [cm^{-1}]
l	Pathlength [cm]
P	Probability of transmission of a photon through a sample of defined thickness
R_∞	Reflectance of an infinitely thick sample
S	Scattering coefficient (scattering per unit thickness) [cm^{-1}]
t	Time
X	Thickness of sample layer [cm]
α	Characteristic Lévy power law exponent
γ	Width of rate constant distribution
$\Delta G^{\neq}$	Free energy of activation [kJ mol^{-1}]
$\Delta\bar{G}^{\neq}$	Mean free energy of activation [kJ mol^{-1}]
ε	Molar decadic or naperian absorption coefficient

See also

Chemical Applications of Lasers: Pump and Probe Studies of Femtosecond Kinetics. **Optical Materials**: Measurement of Optical Properties of Solids. **Scattering**: Scattering from Surfaces and Thin Films; Scattering Theory.

Further Reading

Albery WJ, Bartlett PN, Wilde CP and Darwent JR (1985) A general-model for dispersed kinetics in heterogeneous systems. *Journal of the American Chemical Society* 107: 1854–1858.

Anpo M (ed.) (1996) Surface photochemistry. *Wiley Series in Photoscience and Photoengineering*, vol. 1. Chichester, UK: John Wiley and Sons.

Anpo M and Matsuura T (eds) (1989) Photochemistry on solid surfaces. *Elsevier Series in Studies in Surface Science and Catalysis*, vol. 47. Amsterdam, The Netherlands: Elsevier Science Publishers B.V.

Asahi T, Furube A, Fukumura H, Ichikawa M and Masuhara H (1998) Development of a femtosecond diffuse reflectance spectroscopic system, evaluation of its temporal resolution, and applications to organic powder systems. *Review of Scientific Instruments* 69: 361–371.

Bertoin J (1998) *Lévy Processes*. Cambridge, UK: Cambridge University Press.

Hapke B (1993) *Introduction to the Theory of Reflectance and Emittance Spectroscopy*. Cambridge, UK: Cambridge University Press.

Kamat PV (1993) Photochemistry on nonreactive and reactive (semiconductor) surfaces. *Chemical Reviews* 93: 287–300.

Kan HKA and Lin TP (1970) Calculation of reflectance of a light diffuser with non-uniform absorption. *Journal of the Optical Society of America* 60: 1252–1256.

Kessler RW and Wilkinson F (1981) Diffuse reflectance triplet–triplet absorption spectroscopy of aromatic hydrocarbons chemisorbed on γ-alumina. *Journal of the Chemical Society – Faraday Transactions I* 77: 309–320.

Kortum G (1969) *Reflectance Spectroscopy*. Berlin: Springer-Verlag.

Kubelka P and Munk F (1931) Ein beitrag zur optik der farbanstriche. *zeitschrift fuer Technische Physik* 12: 593–601.

Ramamurthy V (ed.) (1991) *Photochemistry in Organised & Constrained Media*. New York: VCH Publishers.

Thomas JK (1993) Physical aspects of photochemistry and radiation chemistry of molecules adsorbed on SiO_2, γ-Al_2O_3, zeolites and clays. *Chemical Reviews* 93: 301–320.

Thomas JK and Ellison EH (2001) Various aspects of the constraints imposed on the photochemistry of systems in porous silica. *Advances in Colloid and Interface Science* 89–90: 195–238.

Wendlandt WW and Hecht HG (1966) Reflectance Spectroscopy. *Chemical Analysis Series*, vol. 21. New York: Interscience Publishers.

Wilkinson F and Kelly G (1989) Diffuse reflectance flash photolysis. In: Scaiano JC (ed.) *Handbook of Organic Photochemistry*, vol. 1, pp. 293–314. Florida: CRC Press.

Laser Manipulation in Polymer Science

S Ito, Y Hosokawa and H Masuhara, Osaka University, Osaka, Japan

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Laser manipulation is a method for manipulating single particles with a size of less than a few tens of microns in which optical pressure of a focused laser beam is used to trap and control the particles without any mechanical contact. As infrared lasers, as a trapping light source, have become more user-friendly, even operators who are not familiar with lasers and microscopes can perform laser manipulation. Those laser manipulators already on sale have attracted significant attention as a new tool. It is especially interesting to combine this method with nanotechnology and biotechnology, which have progressed rapidly during the 1990s. In this article, the principle and method of laser manipulation is described and then its applications and possibilities in nanotechnology and biotechnology are summarized.

Principle and Method

When light is reflected or refracted at the interface between two media with different refractive indices, the momentum of photons is changed, which leads to the generation of photon force as a reaction to the momentum change. For example, when a laser beam is reflected by a mirror, the momentum of photons is changed by Δp (Figure 1a). As a result, a photon force, $\mathbf{F}_{\text{phot}}$, acts on the mirror to deflect it vertically away from the reflected beam. In refraction, the momentum of photons is also changed, so that the photon force acts on the interface, as shown in Figure 1b. Thus light does not only give its energy to materials via absorption but also applies a mechanical force to them. However, when we are exposed to light, such as from a halogen lamp, we are never aware of the photon force because its magnitude is less than an order of pN. However, the photon force of the pN order acquired from a focused laser beam is sufficient to manipulate nm–μm-sized particles in solution, under an optical microscope.

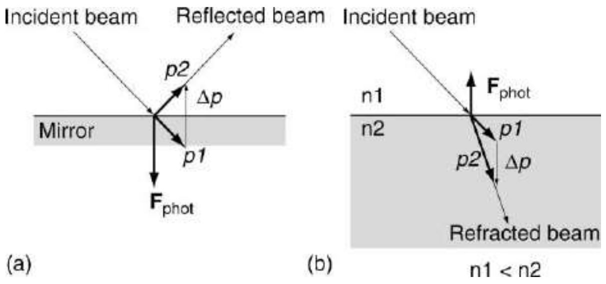

Figure 1 Optical pressure originating from the momentum change of photons.

Principle of Laser Trapping

If the size of a particle is larger than the wavelength of a trapping laser beam (μm-sized particles), the principle of single-beam gradient force optical trapping can be explained in terms of geometrical optics. When a tightly focused laser beam is irradiated onto a transparent dielectric particle, an incident laser beam is refracted at the interface between the particle and medium, as represented by Figure 2a. The propagation direction of the beam is changed, i.e., the momentum of the photon is changed, and consequently, a photon force is generated. As the laser beam leaves the particle and enters the surrounding medium, this refraction causes a photon force to be exerted again on that interface. Summing up the force contributions of all rays, if the refractive index of the particle is higher than that of the medium, a resultant force exerted on the particle is directed toward the focal point as an attractive force. However, reflection at the surface of the particle is negligible if that particle is transparent and the refractive index ratio of the particle to medium is close to unity. Therefore, the incident beam is reflected at two surfaces by a small amount, and as a result, the particle is directed slightly to the propagation direction of the incident light. Where the particle absorbs the incident beam, the photon force is also generated to push it in the propagation direction. The negligible effects of the

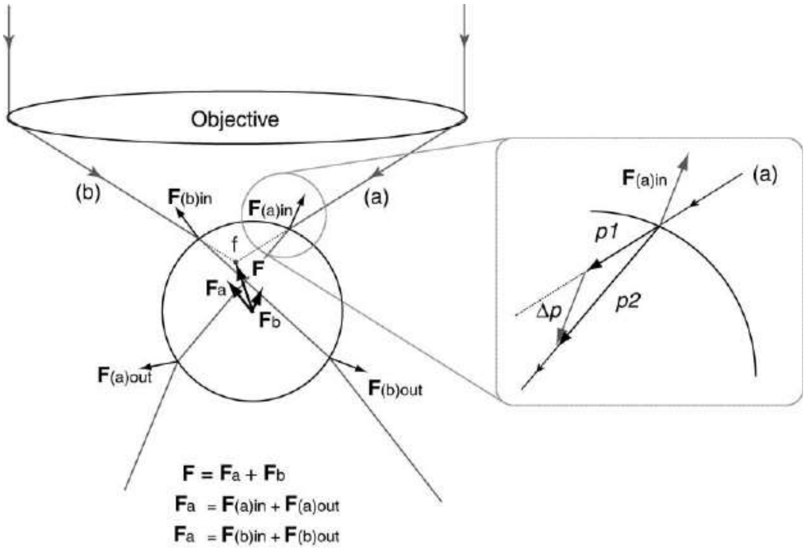

Figure 2 Principle of single-beam gradient force optical trapping explained by ray optics.

reflection and absorption occur in trapping experiment of transparent particles, such as polymer particles, silica microspheres, etc. However, particles with high reflectance and absorption coefficients such as metallic particles, exert a more dominant force at the surface where absorption and reflection occurs. This force is a repulsive force away from the focal point; consequently, metallic microparticles cannot be optically trapped.

If a dielectric particle is much smaller than the wavelength of a trapping light (nm-sized), it can be regarded as a point dipole (Rayleigh approximation) and the photon force ($\mathbf{F}_{\mathrm{phot}}$) acted on it is given by

$$\mathbf{F}_{\mathrm{phot}} = \mathbf{F}_{\mathrm{grad}} + \mathbf{F}_{\mathrm{scat}} \qquad [1]$$

Here, $\mathbf{F}_{\mathrm{grad}}$ and $\mathbf{F}_{\mathrm{scat}}$ are called the gradient force and scattering force, respectively. The scattering force is caused by the scattering of light, and it pushes the particle toward the direction of light propagation. On the other hand, the gradient force is generated when a particle is placed in a heterogeneous electric field of light. If the dielectric constant of the particle is higher than that of the surrounding medium, the gradient force acts on the particle to push it toward the higher intensity region of the beam. In the case of laser trapping of dielectric nanoparticles, the magnitude of gradient force is much larger than the scattering force. Consequently, the particle is trapped at the focal point of the trapping laser beam, where the beam intensity (electric field intensity) is maximum. The photon force can be expressed as follows:

$$\mathbf{F}_{\mathrm{phot}} \cong \mathbf{F}_{\mathrm{grad}} = \frac{1}{2}\varepsilon_{\mathrm{m}}\alpha\nabla|\mathbf{E}|^2 \qquad [2]$$

$$\alpha = 3V\frac{\varepsilon_{\mathrm{p}} - \varepsilon_{\mathrm{m}}}{\varepsilon_{\mathrm{p}} + 2\varepsilon_{\mathrm{m}}} \qquad [3]$$

where, $\mathbf{E}$ is electric field of the light, V is volume of the particle, and ε_{p} and ε_{m} are dielectric constants of the particle and medium. In trapping nanoparticles that absorb the laser beam, such as gold nanoparticles, the complex dielectric constant and attenuation of the electric field in the particle need to be taken into consideration. Although the forces due to light absorption and scattering are both propelling the particle toward the direction of light propagation, the magnitude of gradient force is much larger than these forces. This is in contrast to the μm-sized metallic particle, which cannot be optically trapped.

Laser Manipulation System

An example of a laser micromanipulation system, with a pulsed laser to induce photoreactions, is shown in Figure 3. A linearly polarized laser beam from a CW Nd^{3+}:YAG laser is modulated to the circularly polarized light by a $\lambda/4$ plate and then split into horizontally and vertically polarized beams by a

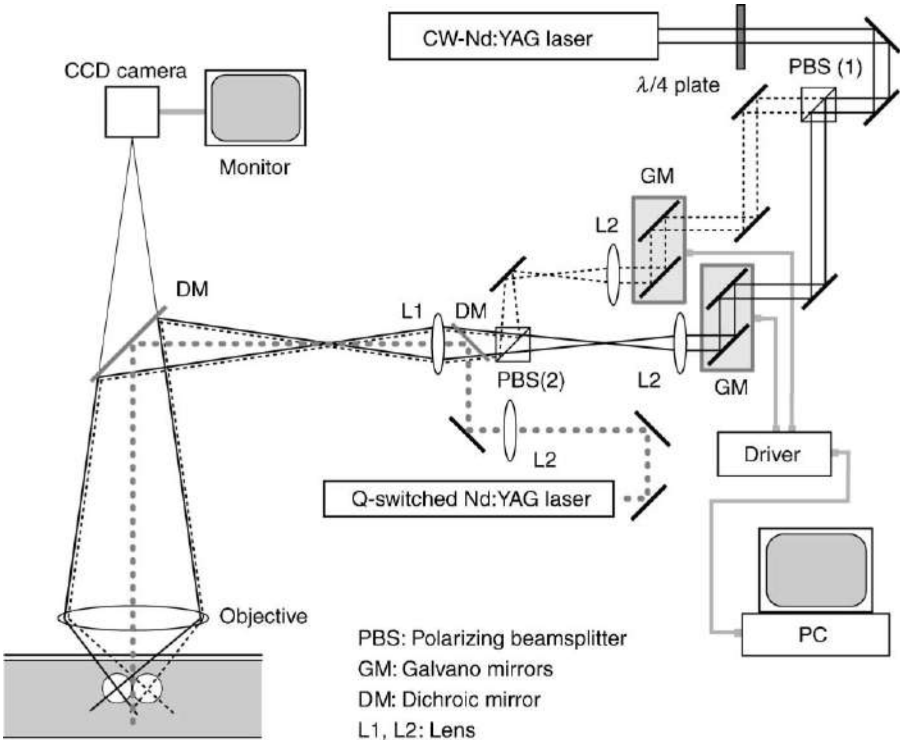

Figure 3 A block diagram of a dual beam laser manipulation-reaction system.

polarizing beamsplitter (PBS1). Both laser beams are combined by another (PBS2), introduced coaxially into an optical microscope, and then focused onto a sample through a microscope objective. The focal spots of both beams in the sample solution can be scanned independently by two sets of computer-controlled galvano mirrors. Even if the beams are spatially overlapping, interference does not take place because of their orthogonal polarization relation. Pulsed lasers, such as a Q-switched Nd^{3+}:YAG laser, are used to induce photoreactions, such as photopolymerization, photothermal reaction, laser ablation, etc.

Laser Manipulation and Patterning of Nanoparticles

Laser manipulation techniques enable us to capture and mobilize fine particles in solution. Most studies using this technique have been conducted on μm-sized objects such as polymer particles, microcrystals, living cells, etc. Because it is difficult to identify individual nanoparticles with an optical microscope, laser manipulation techniques have rarely been applied to nanotechnology and nanoscience. However, the laser trapping technique can be a powerful tool for manipulation of nanoparticles in solution where individual particles can be observed. This is achieved by detection of fluorescence emission from labeled dyes or scattered light. Now, even single molecule fluorescence spectroscopy has been achieved by the use of a highly sensitive photodetector, and a single metallic nanoparticle can be examined by detecting the light scattered from it. There have been several reports on the application of laser manipulation techniques for patterning of nm-sized particles. Here, fixation methods of nanoparticles, using the laser manipulation technique and local photoreactions, are introduced.

Patterning of Polymer Nanoparticles

Patterning of individual polymer nanoparticles onto a substrate can be achieved by using local photopolymerization. The following example shows the strength of this method. Polystyrene nanoparticles of 220 nm with fluorescent dye were dispersed in ethylene glycol solution containing polymerizable vinyl monomer (acrylamide, 31 wt%), crosslinker (*N*,*N*′-methylenebisacrylamide, 2.2 wt%), and radical photoinitiator (Irgacure2959, Ciba Specialty Chemicals, 1.1 wt%). When the sample was irradiated by blue light from a high-pressure mercury lamp, green fluorescence from dye molecules within each nanoparticle was observed. A nanoparticle that entered the region, irradiated by a near-infrared laser beam (1064 nm), was trapped at the focal point, and moved onto the surface of the glass substrate by handling the 3D stage of the microscope. An additional fixation laser beam (355 nm, 0.03 μJ, pulse duration ~6 ns, repetition rate ~5 Hz) was then focused to the same point for ~10 s, which led to the generation of acrylamide gel around the trapped nanoparticle. By repeating the procedure, patterning of single polymer nanoparticles on a glass substrate was achieved, and a fluorescence image of single nanoparticles as a letter 'H', is shown in Figure 4. A magnified atomic force microscope (AFM) image of one of the fixed nanoparticles is also shown, which confirms that only one polymer nanoparticle was contained in the polymerized gel.

Multiple polymer nanoparticles can also be gathered, patterned, and fixed on a glass substrate by scanning both trapping and fixation laser beams with the use of two pairs of galvano mirrors. The optical transmission and fluorescence images of the 'H' patterned nanoparticles on a glass substrate are shown in Figure 5a and b, respectively. The letter 'H' consists of three straight lines of patterned and fixed nanoparticles. The trapping laser beam (180 mW) was scanned at 30 Hz along each line with a length of ~10 μm on a glass substrate for 300 s. Nanoparticles were gathered and patterned along the locus of the focal point on the substrate. Then the fixation laser beam (0.097 μJ) was scanned for another 35 s. As a result, the straight lines of

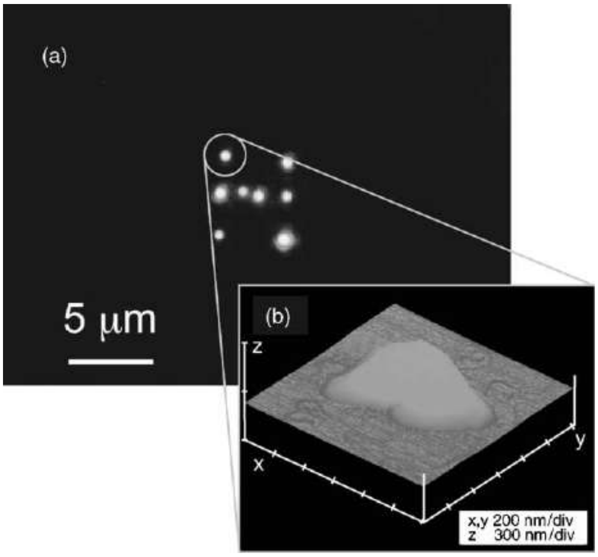

Figure 4 (a) A fluorescence image of spatially patterned individual polymer nanoparticles as the letter 'H' in distilled water. (b) A magnified AFM image of one of the produced acrylamide gels containing only one polymer nanoparticle on a glass substrate in the air.

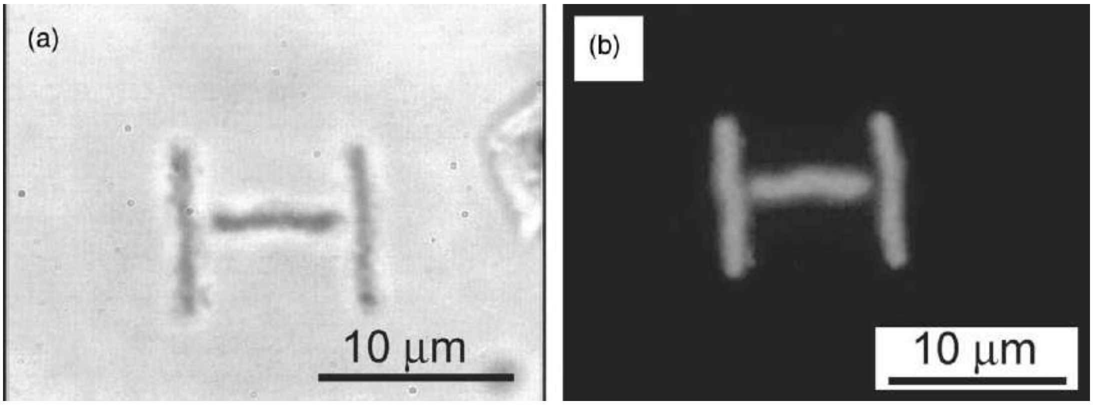

Figure 5 Optical transmission (a) and fluorescence (b) images of patterned and fixed polymer nanoparticles on the glass substrate, as the letter 'H'.

patterned nanoparticles were fixed in generated acrylamide gel on the substrate. Combining several simple fixed patterns, a more complex arrangement can be created with use of the present manipulation and fixation techniques.

Fixation of Individual Gold Nanoparticles

Manipulation and fixation of single metallic nanoparticles in solution has been achieved by means of photo-induced transient melting of nanoparticles. As an example, the following method describes this technique. Gold nanoparticles (diameter ~80 nm) were dispersed in ethylene glycol. In order to identify a single gold nanoparticle, extinction of light from the halogen lamp of the optical microscope by the particle was utilized. Thus, the trapped particle was observed as a black spot in transmission image. The optically trapped single gold nanoparticle was then transferred to a precise position on the surface of glass substrate in the sample solution. The focused 355 nm pulse was additionally irradiated onto the pressed nanoparticle, which led to a transient temperature elevation of the gold nanoparticle to enable its fixation. It was confirmed by AFM observation that, at suitable laser fluence (32–64 mJ cm^{-2}), a gold nanoparticle was fixed on the glass substrate without fragmentation. Repeating the same manipulation and fixation procedure, single gold nanoparticles could be patterned on a glass substrate. Figure 6 shows an AFM image of successive spatial patterning of single 80 nm gold particles.

The significance of the laser manipulation-fixation technique is that we can trap, manipulate, and fix single and/or many nanoparticles in solution at room temperature. We have already demonstrated assembling entangled polymer chains of 10–20 nm mean radius by laser trapping, and formation of a single μm-sized particle. Thus, these various kinds of nanoscopic materials can be well manipulated using these techniques and it is believed that the present nanomanipulation-fixation method will be useful for future nanoscience and nanotechnology.

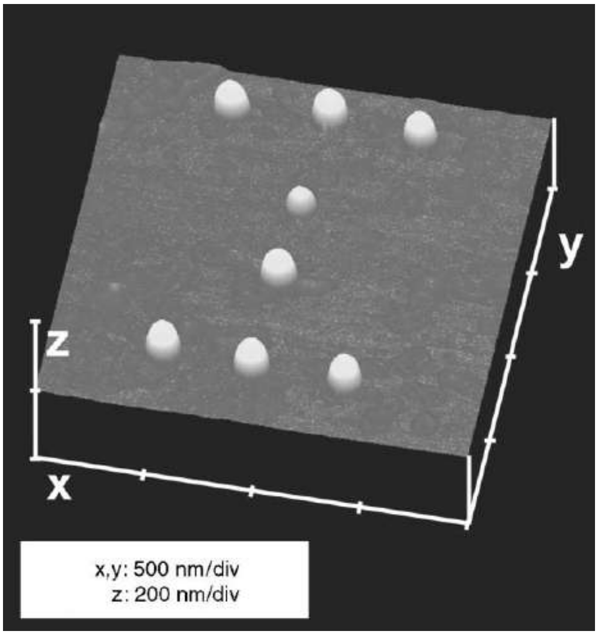

Figure 6 An AFM image of fixed and patterned gold nanoparticles, as the letter I.

Application to Biotechnology

Recent progress in biotechnology, using single cell manipulation by laser and microscope, has been attracting significant attention. In conventional methods, cell manipulation has been performed by mechanical manipulation using microneedles and micropipettes. However, laser manipulation can be applied as microtweezers. In comparison with the former, this form of laser manipulation has the

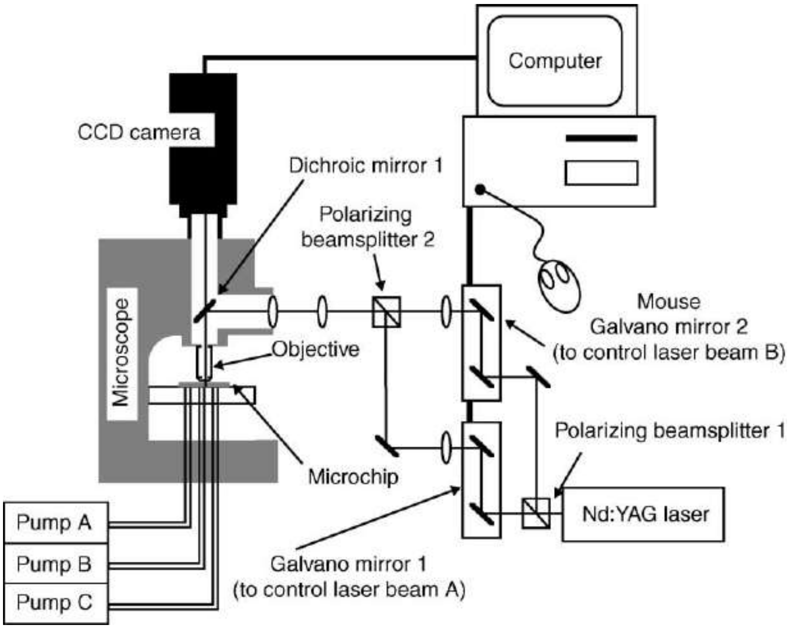
Computer
CCD camera
Dichroic mirror 1
Polarizing beamsplitter 2
Microscope
Objective
Mouse
Galvano mirror 2 (to control laser beam B)
Polarizing beamsplitter 1
Microchip
Nd:YAG laser
Pump A
Pump B
Pump C
Galvano mirror 1 (to control laser beam A)

Figure 7 An experimental setup of dual-beam laser manipulation and microfluidic devices.

advantage of no contact with cells and can perform some complex and characteristic movement, for example, rotation, separation, accumulation, etc. Furthermore, by combining it with a microcutter using laser ablation, cell fusion and extraction/injection of organs from/into the cells can be achieved. In this section, single-cell manipulation achieved by using the laser manipulation, is introduced.

Noncontact Rotation of Cells Using Dual Beam Laser Manipulation

As a special feature of laser-cell manipulation, it is noticeable that multiple laser tweezers can be operated independently without any contact. Here, noncontact rotation of the fission yeast, *Schizosaccharomyces pombe* (h^-), demonstrates this method. The yeast has the shape of an ellipse of length 8 and 3 μm for major and minor axes, and is shown set on a dual-beam laser manipulation system in Figure 7. One trapping beam A was focused at the end of the cell to anchor it, and another beam, B, was at the other end and scanned circularly around beam A by controlling the Galvano mirrors. The rotation of the yeast cell was realized at a frequency of less than 2 Hz, as shown in Figure 8. Such a cell manipulation is impossible by mechanical manipulation and indicates the superior performance of laser manipulation.

Transfer of Cells in Microchannel

A flow cytometry to physically separate and identify specific types of cells from heterogeneous populations

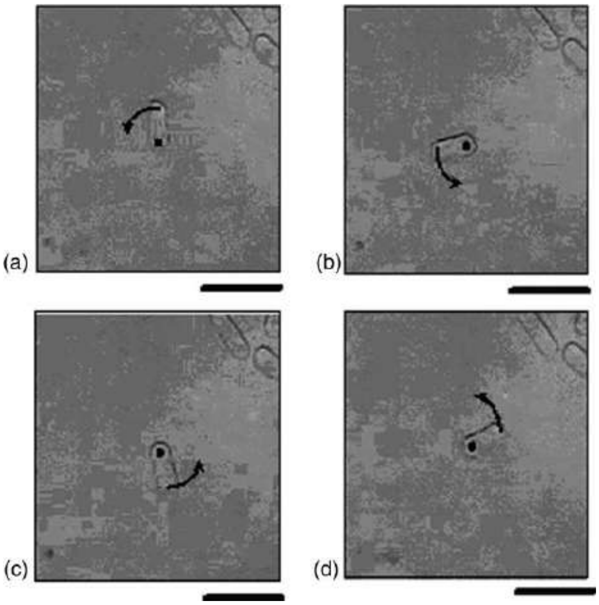

Figure 8 Rotation of fission yeast cells using dual-beam laser manipulation. The bars are 10 μm.

by fluorescence, which is called fluorescence-activated cell sorter (FACS), has attracted significant attention as an important technique in biotechnology. In the separation process, charged single droplets containing single, fluorescence labeled cells are selected by passing them between two high-voltage deflection plates. Since it is a sequential process, which does not use microscopy, flexible and high-purity cell separation is limited. To overcome this problem, a selective cell separation system, combining the laser trapping with the microchannel, is used. The method using microscopy is useful and flexible, because cells are not simply detected by fluorescence but also identified by their shape, size, surface morphology, absorption, etc. However, cell sorting under a microscope is not processed efficiently by laser trapping. A more effective method is described by the following.

The microchannel was prepared with laser microfabrication and photopolymerization, whose top view is shown in Figure 9a, and was then set on the microscope stage. Three syringes were equipped with homemade pumps A, B, and C, and connected with a microchip. There were two wing-like chambers in the microchip, that were connected with pumps A and C. Between chambers, there was a microchannel 100 μm wide, which joined these chambers and crossed a long microchannel connected with the pump B, thus forming a drain. Thickness of these chambers was 100 μm. By controlling these pumps, the left and right chambers were filled by the culture medium including the yeast cells and the neat culture medium, respectively, as shown in Figure 9b.

Individual cells were transferred from left to right chambers by using a single laser beam and handling an electrically movable microscope stage. The trapping laser irradiated a yeast cell in the left chamber and the trapped cell was transferred to the right chamber. A representative demonstration is given in Figure 9c. The position of the trapped yeast cell was controlled by the microscope stage with the velocity of 20 μm/s. The single yeast cells were successfully transferred from left to right chambers at the rate of 26 s. By combining laser trapping with the microchannel, single separation of cells was achieved under a microscope.

Collection and Alignment of Cells

When cell sorting using laser trapping is applied to biotechnology, the time to transfer cells should be short compared to the above-mentioned times. One method is to transfer multiple cells simultaneously by developing dual-beam laser manipulation, which is shown schematically in Figure 10. One trapping beam A was scanned on a line given as an arrow in Figure 10a at a rate faster than motions of cells.

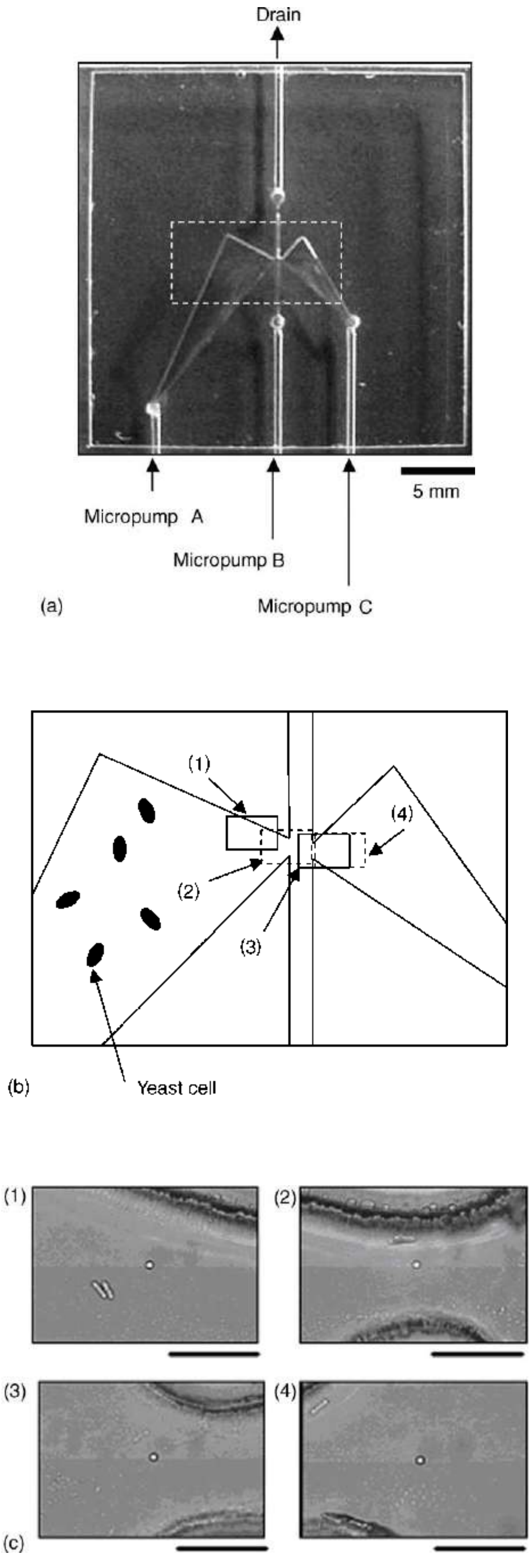

Figure 9 (a) A microchip fabricated by laser polymerization technique and (b) its schematic representation corresponding to dashed area in the microchip. (c) Cell transfer in microchip using laser trapping. The bars are 100 μm. A trapping laser beam is fixed at the center of each picture, by which a particle is trapped and transferred from left to right chambers. The position of each picture is shown in (b).

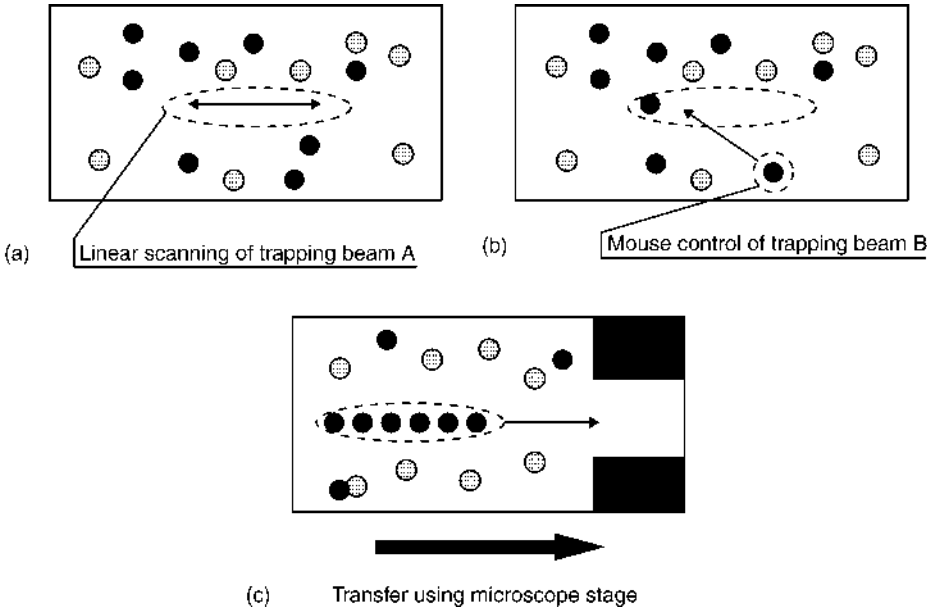
(a)
Linear scanning of trapping beam A
(b)
Mouse control of trapping beam B
(c)
Transfer using microscope stage

Figure 10 A schematic illustration of efficient collection and transfer of cells based by dual-beam laser manipulation.

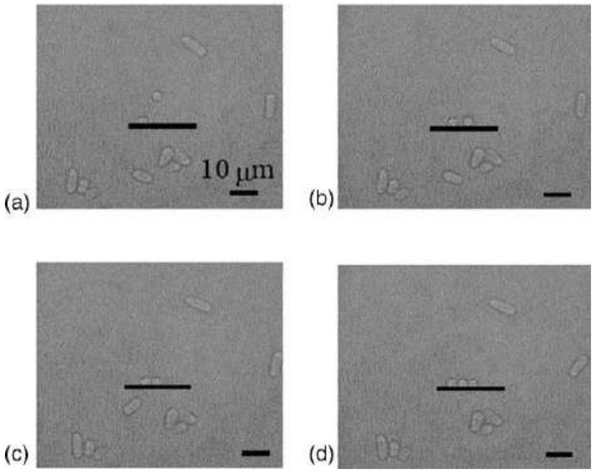

Figure 11 Collection and alignment of fission yeast cells using dual-beam laser manipulation.

The other trapping beam B was used to move individual cells to the line sequentially (Figure 10b). Finally, by moving the microscope stage, many particles trapped at the scanning line can be transferred to another chamber (Figure 10c).

An example is shown in Figure 11. The trapping beam A was scanned on a line with the rate of 15 Hz whose length was 30 μm. The trapping beam B was used to trap a yeast cell and move it to the line, which was controlled by a computer mouse pointer. The cells stored on the line were successfully isolated and with this method three cells can be collected within 15 s. Furthermore, the cells could be transferred with the velocity of 20 μm/s by driving the microscope stage. By improving the presented cell-sorting system, it is expected to realize flexible and high-purity cell separation.

List of Units and Nomenclature

Laser fluence (light energy density per pulse)	[mJ cm^{-2}]

See also

Time-Resolved Fluorescence: Measurements in Polymer Science.

Further Reading

Ashkin A, Dziedzic MJ, Bjorkholm JE and Chu S (1986) Observation of a single-beam gradient force optical trap for dielectric particles. *Optical Letters* 11: 288–290.

Ashkin A, Dziedzic MJ and Yamane T (1987) Optical trapping and manipulation of single cells using infrared laser beams. *Nature* 330: 769–771.

Berns WM, Aist J, Edwards J, *et al.* (1981) Laser microsurgery in cell and developmental biology. *Science* 213: 505–513.

Hoffmann F (1996) Laser microbeams for the manipulation of plant cells and subcellular structures. *Plant Science* 113: 1–11.

Hosokawa Y, Masuhara H, Matsumoto Y and Sato S (2002) Dual-beam laser micromanipulation for sorting biological cells and its device application. *Proceedings of SPIE* 4622: 138–142.

Ito S, Yoshikawa H and Masuhara H (2001) Optical patterning and photochemical fixation of polymer nanoparticles on glass substrates. *Applied Physics Letters* 78: 2566–2568.

Ito S, Yoshikawa H and Masuhara H (2002) Laser manipulation and fixation of single gold nanoparticles in solution at room temperature. *Applied Physics Letters* 80: 482–484.

Kamentsky AL, Melamed RM and Derman H (1965) Spectrophotometer: new instrument for ultrarapid cell analysis. *Science* 150: 630–631.

Kim BH, Kogi O and Kitamura N (1999) Single-microparticle measurements: laser trapping-absorption microscopy under solution-flow conditions. *Analytical Chemistry* 71: 4338–4343.

Masuhara H, De Schryver FC, Kitamura N and Tamai N (eds) (1994) *Microchemistry: Spectroscopy and Chemistry in Small Domains*. Amsterdam: Elsevier-Science.

Sasaki K, Koshioka M, Misawa H, Kitamura N and Masuhara H (1991) Laser-scanning micromanipulation and spatial patterning of fine particles. *Japanese Journal of Applied Physics* 30: L907–L909.

Schütze K, Clement-Segelwald A and Ashkin A (1994) Zona drilling and sperm insertion with combined laser microbeam and optical tweezers. *Fertility and Sterility* 61: 783–786.

Shikano S, Horio K, Ohtsuka Y and Eto Y (1999) Separation of a single cell by red-laser manipulation. *Applied Physics Letters* 75: 2671–2673.

Svoboda K and Block SM (1994) Optical trapping of metallic Rayleigh particles. *Optical Letters* 19: 930–932.

Nonlinear Spectroscopies

S R Meech, University of East Anglia, Norwich, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Chemistry is concerned with the induction and observation of changes in matter, where the changes are to be understood at the molecular level. Spectroscopy is the principal experimental tool for connecting the macroscopic world of matter with the microscopic world of the molecule, and is, therefore, of central importance in chemistry. Since its invention, the laser has greatly expanded the capabilities of the spectroscopist. In linear spectroscopy, the monochromaticity, coherence, high intensity, and high degree of polarization of laser radiation are ideally suited to high-resolution spectroscopic investigations of even the weakest transitions. The same properties allowed, for the first time, the investigation of the nonlinear optical response of a medium to intense radiation. Shortly after the foundations of nonlinear optics were laid, it became apparent that these nonlinear optical signals could be exploited in molecular spectroscopy, and since then a considerable number of nonlinear optical spectroscopies have been developed. This short article is not a comprehensive review of all these methods. Rather, it is a discussion of some of the key areas in the development of the subject, and indicates some current directions in this increasingly diverse area.

Nonlinear Optics for Spectroscopy

The foundations of nonlinear optics are described in detail elsewhere in the encyclopedia, and in some of the texts listed in the Further Reading section at the end of this article. The starting point is usually the nonlinearity in the polarization, P_i, induced in the sample when the applied electric field, E, is large:

$$P_i = \varepsilon_0\left[\chi^{(1)}_{ij}E_j + \frac{1}{2}\chi^{(2)}_{ijk}E_jE_k + \frac{1}{4}\chi^{(3)}_{ijkl}E_jE_kE_l + \cdots\right] \qquad [1]$$

where ε_0 is the vacuum permittivity, $\chi^{(n)}$ is the nth order nonlinear susceptibility, the indices represent directions in space, and the implied summation over repeated indices convention is used. The signal field, resulting from the nonlinear polarization, is calculated by substituting it as the source polarization in Maxwell's equations and converting the resulting field to the observable, which is the optical intensity.

The nonlinear polarization itself arises from the anharmonic motion of electrons under the influence of the oscillating electric field of the radiation. Thus, there is a microscopic analog of eqn [1] for the induced molecular dipole moment, μ_i:

$$\mu_i = [\varepsilon_0\alpha_{ij}d_j + \varepsilon_0^{-2}\beta_{ijk}d_jd_k + \varepsilon_0^{-3}\gamma_{ijkl}d_jd_kd_l + \cdots] \qquad [2]$$

in which α is the polarizability, β the first molecular hyperpolarizability, γ the second, etc. The power series is expressed in terms of the displacement field $\mathbf{d}$ rather than $\mathbf{E}$ to account for local field effects. This somewhat complicates the relationship between the molecular hyperpolarizabilities, for example, γ_{ijkl} and the corresponding macroscopic susceptibility, $\chi^{(3)}_{ijkl}$, but it is nevertheless generally true that a molecule exhibiting a high value of γ_{ijkl} will also yield a large third-order nonlinear polarization. This relationship between the macroscopic and microscopic para meters is the basis for an area of chemistry which has influenced nonlinear optics (rather than the inverse). The synthesis of molecules which have giant molecular hyperpolarizabilities has been an active area, because of their potential applications in lasers and electro-optics technology. Such molecules commonly exhibit a number of properties, including an extended linear π electron conjugation and a large dipole moment, which changes between the ground and excited electronic states. These properties are in accord with the predictions of theory and of quantum chemical calculations of molecular hyperpolarizabilities.

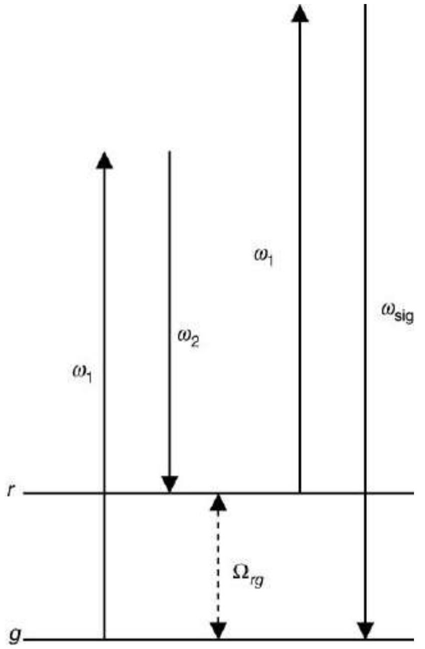

Figure 1 An illustration of the resonance enhancement of the CARS signal, $\omega_{sig} = 2\omega_1 - \omega_2$ when $\omega_1 - \omega_2 = \Omega_{rg}$.

Nonlinear optical spectroscopy in the frequency domain is carried out by measuring the nonlinear signal intensity as a function of the frequency (or frequencies) of the incident radiation. Spectroscopic information is accessible because the molecular hyperpolarizability, and therefore the nonlinear susceptibility, exhibits resonances: the signal is enhanced when one or more of the incident or generated frequencies are resonant with the frequency of a molecular transition. The rich array of nonlinear optical spectroscopies arises in part from the fact that with more input fields there are more accessible resonances than is the case with linear spectroscopy. As an example we consider the third-order susceptibility, $\chi^{(3)}_{ijkl}E_jE_kE_l$, in the practically important case of two incident fields at the same frequency, ω_1, and a third at ω_2. The difference between the two frequencies is close to the frequency of a Raman active vibrational mode, Ω_{rg} (Figure 1). The resulting susceptibility can be calculated to have the form:

$$\chi^{(3)}_{ijkl}(-2\omega_1+\omega_2;\omega_1,\omega_1,-\omega_2) = \frac{N\Delta\rho_{rg}\Omega_{rg}(\alpha^{R}_{ij}\alpha^{R}_{kl}+\alpha^{R}_{ik}\alpha^{R}_{jl})}{12\hbar[\Omega^2_{rg}-(\omega_1-\omega_2)^2+\Gamma^2-2i(\omega_1-\omega_2)\Gamma]} \qquad [3]$$

in which the α^{R} are elements of the Raman susceptibility tensor, Γ is the homogeneous linewidth of the Raman transition, $\Delta\rho_{rg}$ a population difference, and N the number density. A diagram illustrating this process is shown in Figure 1, where the signal field is at the anti-Stokes Raman frequency ($2\omega_1 - \omega_2$). The spectroscopic method which employs this scheme is called coherent anti-Stokes Raman spectroscopy (CARS) and is one of the most widely applied nonlinear optical spectroscopies (see below). Clearly, from eqn [3] we can see that the signal will be enhanced when the difference frequency is resonant with the Raman transition frequency.

With essentially the same experimental geometry there will also be nonlinear signals generated at both the Stokes frequency ($2\omega_2 - \omega_1$) and at the lower of the incident frequencies, ω_2. These signals have distinct phase matching conditions (see below), so they can easily be discriminated from one another, both spatially and energetically. Additional resonance enhancements are possible if either of the individual frequencies is resonant with an electronic transition of the molecule, in which case information on Raman active modes in the excited state is also accessible.

It is worthwhile noting here that there is an equivalent representation of the nth order nonlinear susceptibility tensor $\boldsymbol{\chi}^{(n)}(\omega_{sig}:\omega_1,\ldots,\omega_n)$ as a time domain response function, $\mathrm{R}^{(n)}(\tau_1,\ldots,\tau_n)$. While it is possible to freely transform between them, the frequency domain representation is the more commonly used. However, the response function approach is increasingly applied in time domain nonlinear optical spectroscopy when optical pulses shorter than the homogeneous relaxation time are used. In that case, the time ordering of the incident fields, as well as their frequencies, is of importance in defining an experiment.

An important property of many nonlinear optical spectroscopies is the directional nature of the signal, illustrated in Figure 2. The directional nonlinear

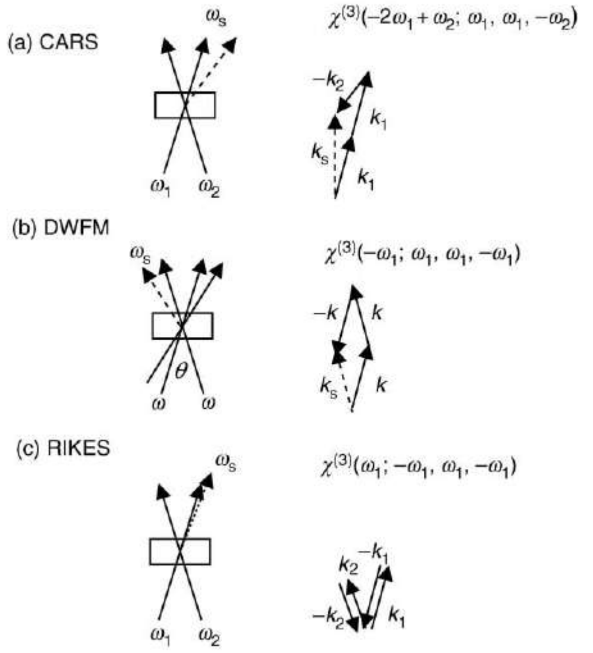

Figure 2 Experimental geometries and corresponding phase matching diagrams for (a) CARS (b) DFWM (c) RIKES. Note that only the RIKES geometry is fully phase matched.

signal arises from the coherent oscillation of induced dipoles in the sample. Constructive interference can lead to a large enhancement of the signal strength. For this to occur the individual induced dipole moments must oscillate in phase – the signal must be phase matched. This requires the incident and the generated frequencies to travel in the sample with the same phase velocity, $k_\omega/\omega = c/n_\omega$ where n_ω is the index of refraction and k_ω is the wave propagation constant at frequency ω. For the simplest case of second-harmonic generation (SHG), in which the incident field oscillates at ω and the generated field at 2ω, phase matching requires $2k_\omega = k_{2\omega}$. The phase-matching condition for the most efficient generation of the second harmonic is when the phase mismatch, $\Delta k = 2k_\omega - k_{2\omega} = 0$, but this is not generally fulfilled due to the dispersion of the medium, $n_\omega < n_{2\omega}$. In the case of SHG, a coherence length L can be defined as the distance traveled before the two waves are 180° out of phase, $L = |\pi/\Delta k|$. For cases in which the input frequencies also have different directions, as is often the case when laser beams at two frequencies are combined (e.g., CARS), the phase-matching condition must be expressed as a vectorial relationship, hence, for CARS, $\Delta \mathbf{k} = \mathbf{k}_{AS} - 2\mathbf{k}_1 + \mathbf{k}_2 \approx 0$, where $\mathbf{k}_{AS}$ is the wavevector of the signal at the anti-Stokes frequency (Figure 2). Thus for known input wavevectors one can easily calculate the expected direction of the output signal, $\mathbf{k}_s$. This is illustrated for a number of important cases in Figure 2.

Nonlinear Optical Spectroscopy in Chemistry

As already noted, there is a vast array of nonlinear optical spectroscopies, so it is clear some means of classification will be required. For coarse graining the order of the nonlinear process is very helpful, and that is the scheme we will follow here.

Second-Order Spectroscopies

Inspection of eqn [1] shows that in gases, isotropic liquids, and solids where the symmetry point group contains a center of inversion, $\boldsymbol{\chi}^{(2)}$ is necessarily zero. This is required to satisfy the symmetry requirement that polarization must change sign when the direction of the field is inverted, yet for a quadratic, or any even order dependence on the field strength, it must remain positive. Thus second-order nonlinearities might not appear very promising for spectroscopy. However, there are two cases in which second-order nonlinear optical phenomena are of very great significance in molecular spectroscopy, harmonic generation and the spectroscopy of interfaces.

Harmonic conversion

Almost every laser spectroscopist will have made use of second-harmonic or sum frequency generation for frequency conversion of laser radiation. Insertion of an oscillating field of frequency ω into eqn [1] yields a second-order polarization oscillating of 2ω. If two different frequencies are input, the second-order polarization oscillates at their sum and difference frequencies. In either case, the second-order polarization acts as a source for the second-harmonic (or sum, or difference frequency) emission, provided $\boldsymbol{\chi}^{(2)}$ is nonzero. The latter can be arranged by selecting a noncentrosymmetric medium for the interaction, the growth of such media being an important area of materials science. Optically robust and transparent materials with large values of $\boldsymbol{\chi}^{(2)}$ are available for the generation of wavelengths shorter than 200 nm to longer than 5 μm. Since such media are birefringent by design, a judicious choice of angle and orientation of the crystal with respect to the input beams allows a degree of control over the refractive indices experienced by each beam. Under the correct phase matching conditions $n_\omega \approx n_{2\omega}$, and very long interaction lengths result, so the efficiency of signal generation is high.

Higher-order harmonic generation in gases is an area of growing importance for spectroscopy in the deep UV and X-ray region. The generation of such short wavelengths depends on the ability of amplified ultrafast solid state lasers to generate extremely high instantaneous intensities. The mechanism is somewhat different to the one outlined above. The intense pulse is injected into a capillary containing an appropriate gas. The high electric field of the laser causes ionization of atoms. The electrons generated begin to oscillate in the applied laser field. The driven recombination of the electrons with the atoms results in the generation of the high harmonic emission. Although the mechanism differs from the SHG case, questions of phase matching are still important. By containing the gas in a corrugated waveguide phase matching is achieved, considerably enhancing the intensity of the high harmonic. Photon energies of hundreds of electronvolts are possible using this technique. The generation of such high energies is not yet routine, but a number of potential applications are already apparent. A powerful coherent source of X-ray and vacuum UV pulses will certainly aid surface analysis techniques such as UV photoemission and X-ray photoelectron spectroscopy. Much excitement is currently being generated by the possibility of using intense ultrashort X-ray pulses to record structural dynamics on an ultrafast timescale.

Surface second-order spectroscopy

At an interface inversion symmetry is absent by definition, so the second-order nonlinear susceptibility is finite. If the two bulk phase media are themselves isotropic, then even a weak second-order signal necessarily arises from the interface. This surface-specific all-optical signal is unique, because it can be used to probe the interface between two condensed phases. This represents a great advantage over every other form of surface spectroscopy. In linear optical spectroscopy, the signal due to species at the interface are usually swamped by contributions from the bulk phase. Other surface-specific signals do exist, but they rely on the scattering of heavy particles (electrons, atoms) and so can only be applied to the solid vacuum interface. For this reason the techniques of surface SHG and SFG are widely applied in interface spectroscopy.

The most widely used method is sum frequency generation (SFG) between temporally overlapped tuneable infrared and fixed frequency visible lasers, to yield a sum frequency signal in the visible region of the spectrum. The principle of the method is shown in Figure 3. The surface nonlinear susceptibility exhibits resonances at vibrational frequencies, which are detected as enhancements in the visible SFG intensity. Although the signal is weak, it is directional and background free, so relatively easily measured by photon counting techniques. Thus, SFG is used to measure the vibrational spectra of essentially any optically accessible interface. There are, however, some limitations on the method. The surface must exhibit a degree of order – if the distribution of adsorbate orientation is isotropic the signal again becomes zero by symmetry. Second, a significant enhancement at the vibrational frequency requires the transition to be both IR and Raman allowed, as suggested by the energy level diagram (Figure 3).

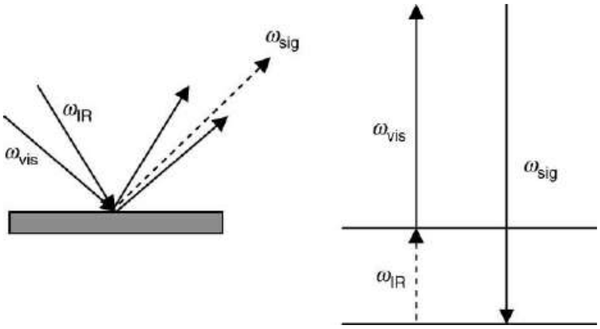

Figure 3 The experimental geometry for SFG, and an illustration of the resonance enhancement of $\omega_{sig} = \omega_{IR} + \omega_{vis}$ at a Raman and IR allowed vibrational transition.

The SHG signal can also be measured as a function of the frequency of the incident laser, to recover the electronic spectrum of the interface. This method has been used, particularly in the study of semiconductor surfaces, but generally the electronic spectra of adsorbates contain less information than their vibrational spectra. However, by measuring the SHG intensity as a function of time, information on adsorbate kinetics is obtained, provided some assumptions connecting the surface susceptibility to the molecular hyperpolarizability are made. Finally, using similar assumptions, it is possible to extract the orientational distribution of the adsorbate, by measuring the SHG intensity as a function of polarization of the input and output beams. For these reasons, SHG has been widely applied to analyze the structure and dynamics of interfaces.

Third-Order Spectroscopies

The third-order coherent Raman spectroscopies were introduced above. One great advantage of these methods over conventional Raman is that the signal is generated in a coherent beam, according to the appropriate phase matching relationship (Figure 2). Thus, the coherent Raman signal can easily be distinguished from background radiation by spatial filtering. This has led to CARS finding widespread application in measuring spectra in (experimentally) hostile environments. CARS spectroscopy has been widely applied to record the vibrational spectra of flames. Such measurements would obviously be very challenging for linear Raman or IR, due to the strong emission from the flame itself. The directional CARS signal in contrast can be spatially filtered, minimizing this problem. CARS has been used both to identify unstable species formed in flames, and to probe the temperature of the flame (e.g., from measured population differences eqn [3]). A second advantage of the technique is that the signal is only generated when the two input beams overlap in space. Thus, small volumes of the sample can be probed. By moving the overlap position around in the sample, spatially resolved information is recovered. Thus, it is possible to map the population of a particular transient species in a flame.

CARS is probably the most widely used of the coherent Raman methods, but it does have some disadvantages, particularly in solution phase studies. In that case the resonant $\boldsymbol{\chi}^{(3)}$ signal (eqn [3]) is accompanied by a nonresonant third-order background. The interference between these two components may result in unusual and difficult to interpret lineshapes. In this case, some other coherent Raman methods are more useful. The phase matching scheme for Raman-induced Kerr effect spectroscopy (RIKES) was shown in Figure 2. The RIKES signal is always phase matched, which leads to a long interaction length. However, the signal is at ω_2 and in the direction of ω_2, which would appear to be a severe disadvantage in terms of detection. Fortunately, if the input polarizations are correctly chosen the signal can be isolated by its polarization. In an important geometry, the signal (ω_2) is isolated by transmission through a polarizer oriented at 45° to a linearly polarized pump (ω_1). The pump is overlapped in the sample with the probe (ω_2) linearly polarized at $-45°$. Thus the probe is blocked by the polarizer, but the signal is transmitted. This geometry may be viewed as pump-induced polarization of the isotropic medium to render it birefringent, thus inducing ellipticity in the transmitted probe, such that the signal leaks through the analyzing polarizer (hence the alternative name optical Kerr effect).

In this geometry, it is possible to greatly enhance signal-to-noise ratios by exploiting interference between the probe beam and the signal. Placing a quarterwave plate in the probe beam with its fast axis aligned with the probe polarization, and reorienting it slightly ($<1°$) yields a slightly elliptically polarized probe before the sample. A fraction of the probe beam, the local oscillator (LO), then also leaks through the analyzing polarizer, temporally and spatially overlapped with the signal. Thus, the signal and LO fields are seen by the detector, which measures the intensity as:

$$\begin{aligned} I(t) &= \frac{nc}{8\pi}|\mathbf{E}_{\mathrm{LO}}(t)+\mathbf{E}_{\mathrm{S}}(t)|^2 \\ &= I_{\mathrm{LO}}(t)+I_{\mathrm{S}}(t)+\frac{nc}{4\pi}\mathrm{Re}[\mathbf{E}_{\mathrm{S}}^{*}(t)\cdot\mathbf{E}_{\mathrm{LO}}(t)] \qquad [4] \end{aligned}$$

where the final term may be very much larger than the original signal, and is linear in $\boldsymbol{\chi}^{(3)}$. This term is usually isolated from the strong I_{LO} by lock-in detection. This method is called optical heterodyne detection (OHD), and generally leads to excellent signal to noise. It can be employed with other coherent signals by artificially adding the LO to the signal, provided great care is taken to ensure a fixed phase relationship between LO and signal. In the RIKES experiment, however, the phase relationship is automatic. The arrangement described yields an out-of-phase LO, and measures the real part of $\boldsymbol{\chi}^{(3)}$, the birefringence. Alternatively, the quarterwave plate is omitted, and the analyzing polarizer is slightly reoriented, to introduce an in-phase LO, which measures the imaginary part of $\boldsymbol{\chi}^{(3)}$, the dichroism. This is particularly useful for absorbing media. The OHD-RIKES method has been applied to measure the spectroscopy of the condensed phase, and has found particularly widespread application in transient studies (below).

Degenerate four wave mixing (DFWM) spectroscopy is a simple and widely used third-order spectroscopic method. As the name implies, only a single frequency is required. The laser beam is split into three, and recombined in the sample, in the geometry shown in Figure 2. The technique is also known as laser-induced grating scattering. The first two beams can be thought of as interfering in the sample to write a spatial grating, with fringe spacing dependent on the angle between them. The signal is then scattered from the third beam in the direction expected for diffraction from that grating. The DFWM experiment has been used to measure electronic spectra in hostile environments, by exploiting resonances with electronic transitions. It has also been popular in the determination of the numerical value of $\boldsymbol{\chi}^{(3)}$, partly because it is an economical technique, requiring only a single laser, but also because different polarization combinations make it possible to access different elements of $\boldsymbol{\chi}^{(3)}$. The technique has also been used in time resolved experiments, where the decay of the grating is monitored by diffraction intensity from a time delayed third pulse.

Two-photon or, more generally, multiphoton excitation has applications in both fundamental spectroscopy and analytical chemistry. Two relevant level

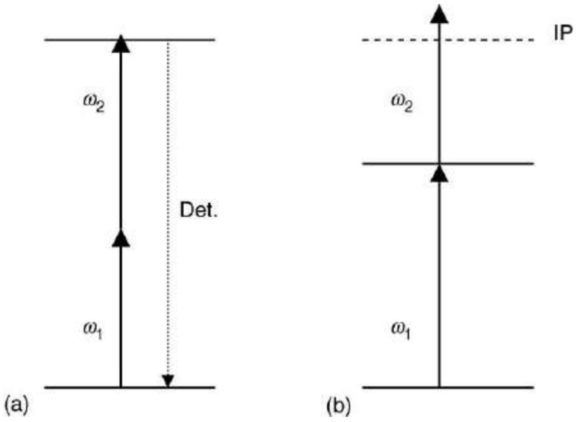

Figure 4 Two cases of resonant two photon absorption. In (a) the excited state is two-photon resonant, and the process is detected by the emission of a photon. In (b) the intermediate state is resonant, and the final energy is above the ionization potential (IP) so that photocurrent or mass detection can be used.

schemes are shown in Figure 4. Some property associated with the final state permits detection of the multiphoton absorption, for example, fluorescence in (a) and photocurrent in (b).

Excitation of two-photon transitions, as in Figure 4a, is useful in spectroscopy because the selection rules are different to those for the corresponding one-photon transition. For example the change in angular momentum quantum number, ΔL, in a two-photon transition is 0, ± 2, so, for example, an atomic S to D transition can be observed. High spectroscopic resolution may be attained using Doppler free two-photon absorption spectroscopy. In this method, the excitation beams are arranged to be counter-propagating, so that the Doppler broadening is cancelled out in transitions where the two excitation photons arise from beams with opposing wavevectors. In this case, the spectroscopic linewidth is governed only by the homogeneous dephasing time.

The level scheme in Figure 4b is also widely used in spectroscopy, but in this case the spectrum of the intermediate state is obtained by monitoring the photocurrent as a function of ω_1. The general technique is known as resonance enhanced multiphoton ionization (REMPI) and yields high-quality spectra of intermediate states which are not detectable by standard methods, such as fluorescence. The sensitivity of the method is high, and it is the basis of a number of analytical applications, often in combination with mass spectrometry.

Ultrafast Time Resolved Spectroscopy

The frequency and linewidth of a Raman transition may be extracted from the CARS measurement, typically by combining two narrow bandwidth pulsed lasers, and tuning one through the resonance while measuring the nonlinear signal intensity. The time resolved analogue requires two pulses, typically of a few picoseconds duration (and therefore a few wavenumbers bandwidth) at ω_1 and ω_2 to be incident on the sample. This pair coherently excites the Raman mode. A third pulse at ω_1 is incident on the sample a time t later, and stimulates the CARS signal at $2\omega_1 - \omega_2$ in the phase-matched direction. The decay rate of the vibrational coherence is measured from the CARS intensity as a function of the delay time. Thus, the frequency of the mode is measured in the frequency domain, but the linewidth is measured in the time domain. If very short pulses are used (such that the pulsewidth is shorter than the inverse frequency of the Raman active mode) the Raman transition is said to be impulsively excited, and the CARS signal scattered by the time delayed pulse reveals an oscillatory response at the frequency of the Raman active mode, superimposed on its decay. Thus, in this case, spectroscopic information is measured exclusively in the time domain. In the case of nonlinear Raman spectroscopy, similar information is available from the frequency and the time domain measurements, and the choice between them is essentially one of experimental convenience. For example, time domain CARS, RIKES, and DFWM spectroscopy have turned out to be particularly powerful routes to extracting low-frequency vibrational and orientational modes of liquids and solutions, thus providing detailed insights into molecular interactions and reaction dynamics in the condensed phase.

Other time domain experiments contain information that is not accessible in the frequency domain. This is particularly true of photon echo methods. The name suggests a close analogy with nuclear magnetic resonance (NMR) spectroscopy, and the (optical) Bloch vector approach may be used to describe both measurements, although the transition frequencies and time-scales involved differ by many orders of magnitude. In the photon echo experiment, two or three ultrafast pulses with carefully controlled interpulse delay times are resonant with an electronic transition of the solute. In the two-pulse echo, the echo signal is emitted in the phase match direction at twice the interpulse delay, and its intensity as a function of time yields the homogeneous dephasing time associated with the transition. In the three-pulse experiment the pulses are separated by two time delays. By measuring the intensity of the stimulated echo as a function of both delay times it is possible to separately determine the dephasing time and the population relaxation time associated with the resonant transition. Such information is not accessible from linear spectroscopy, and can be extracted only with difficulty in the frequency domain. The understanding of photon echo spectroscopy has expanded well beyond the simple description given here, and it now provides unprecedented insights into optical dynamics in solution, and thus informs greatly our understanding of chemistry in the condensed phase. The methods have recently been extended to the infra red, to study vibrational transitions.

Higher-Order and Multidimensional Spectroscopies

The characteristic feature of this family of spectroscopies is the excitation of multiple resonances, which may or may not require measurements at $\chi^{(n)}$ with $n > 3$. Such experiments require multiple frequencies, and may yield weak signals, so they only became experimentally viable upon the availability of stable and reliable solid state lasers and optical parametric generators. Measurements are made in either the time or the frequency domain, but in either case benefit from heterodyne detection.

One of the earliest examples was two-dimensional Raman spectroscopy, where multiple Raman active modes are successively excited by temporally delayed pulse pairs, to yield a fifth-order nonlinear signal. The signal intensity measured as a function of both delay times (corresponding to the two dimensions) allows separation of homogeneous and inhomogeneous contributions to the line shape. This prodigiously difficult $\chi^{(5)}$ experiment has been completed in a few cases, but is plagued by interference from third-order signals.

More widely applicable are multidimensional spectroscopies using infrared pulses or combinations of them with visible pulses. The level scheme for one such experiment is shown in Figure 5 (which is one of many possibilities). From the scheme, one can see that the nonlinear signal in the visible depends on two resonances, so both can be detected. This can be regarded as a multiply resonant nondegenerate four-wave mixing (FWM) experiment. In addition, if the two resonant transitions are coupled, optical excitation of one affects the other. Thus, by measuring the signal as a function of both frequencies, the couplings between transitions are observed. These appear as cross peaks when the intensity is plotted in the two frequency dimensions, very much as with 2D NMR. This technique is already providing novel information on molecular structure and structural dynamics in liquids, solutions, and proteins.

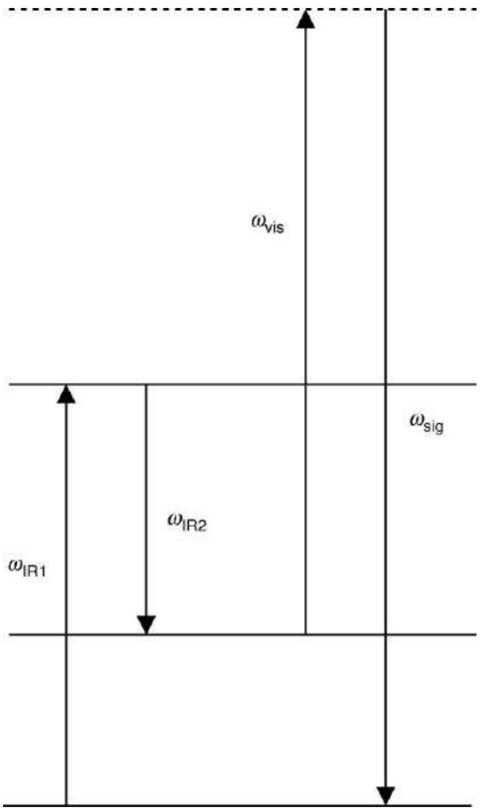

Figure 5 Illustration of multiple resonance enhancements in a FWM geometry, from which 2D spectra may be generated.

Spatially Resolved Spectroscopy

A recent innovation is nonlinear optical microscopy. The nonlinear dependence of signal strength on intensity means that nonlinear processes are localized at the focal point of a lens. When focusing is strong, such as in a microscope objective, spatial localization of the nonlinear signal can be dramatic. This is the basis of the two-photon fluorescence microscopy method, where a high repetition rate source of low-energy ultrafast pulses is focused by a microscope objective into a sample labeled with a fluorescent molecule, which has absorption at half the wavelength of the incident photons (Figure 4). The fluorescence is necessarily localized at the focal point because of its dependence on the square of the incident intensity. By measuring intensity while scanning the position of the focal point in space, a 3D image of the distribution of the fluorophore is constructed. This technique turns out to have a number of advantages over one photon fluorescence microscopy, most notably in terms of ease of implementation, minimization of sample damage, and depth resolution. The technique is widely employed in cell biology.

Stimulated by the success of two-photon microscopy, further nonlinear microscopies have been developed, all relying on the spatial localization of the signal. CARS microscopy has been demonstrated to yield 3D images of the distribution of vibrations in living cells. It would be difficult to recover such data by linear optical microscopy. SHG has been applied in microscopy. In this case, by virtue of the symmetry selection rule referred to above, a 3D image of orientational order is recovered. Both these and other nonlinear signals provide important new information on complex heterogeneous samples, most especially living cells.

See also

Spectroscopy: Nonlinear Laser Spectroscopy; Raman Spectroscopy.

Further Reading

Andrews DL and Allcock P (2002) *Optical Harmonics in Molecular Systems*. Weinheim, Germany: Wiley-VCH.

Andrews DL and Demidov AA (eds) (2002) *An Introduction to Laser Spectroscopy*. New York: Kluwer Academic.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Cheng JX and Xie XS (2004) Coherent anti-Stokes Raman scattering microscopy: Instrumentation, theory, and applications. *Journal of Physical Chemistry B* 108: 827.

de Boeij WP, Pshenichnikov MS and Wiersma DA (1998) Ultrafast solvation dynamics explored by femtosecond photon echo spectroscopies. *Annual Review of Physical Chemistry* 49: 99.

Eisenthal KB (1992) Equilibrium and dynamic processes at interfaces by 2nd harmonic and sum frequency generation. *Annual Review of Physical Chemistry* 43: 627.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Oxford University Press.

Fleming GR and Cho M (1996) Chromophore-solvent dynamics. *Annual Review of Physical Chemistry* 47: 109.

Fourkas JT (2002) Higher-order optical correlation spectroscopy in liquids. *Annual Review of Physical Chemistry* 53: 17.

Hall G and Whitaker BJ (1994) Laser-induced grating spectroscopy. *Journal of Chemistry Society Faraday Transactions* 90: 1.

Heinz TF (1991) Second order nonlinear optical effects at surfaces and interfaces. In: Ponath H-E and Stegeman GI (eds) *Nonlinear Surface Electromagnetic Phenomena*, pp. 353.

Hesselink WH and Wiersma DA (1983) Theory and experimental aspects of photon echoes in molecular solids. In: Hochstrasser RM and Agranovich VM (eds) *Spectroscopy and Excitation Dynamics of Condensed Molecular Systems*, chap. 6. Amsterdam: North Holland.

Levenson MD and Kano S (1988) *Introduction to Nonlinear Laser Spectroscopy*. San Diego, CA: Academic Press.

Meech SR (1993) Kinetic application of surface nonlinear optical signals. In: Lin SH, Fujimura Y, and Villaeys A (eds) *Advances in Multiphoton Processes and Spectroscopy*, vol. 8, pp. 281. Singapore: World Scientific.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. Oxford: Oxford University Press.

Rector KD and Fayer MD (1998) Vibrational echoes: a new approach to condensed-matter vibrational spectroscopy. *International Review of Physical Chemistry* 17: 261.

Richmond GL (2001) Structure and bonding of molecules at aqueous surfaces. *Annual Review of Physical Chemistry* 52: 357.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Smith NA and Meech SR (2002) Optically heterodyne detected optical Kerr effect – Applications in condensed phase dynamics. *International Review of Physical Chemistry* 21: 75.

Tolles WM, Nibler JW, McDonald JR and Harvey AB (1977) Review of theory and application of coherent anti-Stokes Raman spectroscopy (CARS). *Applied Spectroscopy* 31: 253.

Wright JC (2002) Coherent multidimensional vibrational spectroscopy. *International Review of Physical Chemistry* 21: 185.

Zipfel WR, Williams RM and Webb WW (2003) Nonlinear magic multiphoton microscopy in the biosciences. *Nature Biotechnology* 21: 1368.

Zyss J (ed.) (1994) *Molecular Nonlinear Optics*. San Diego: Academic Press.

Photodynamic Therapy of Cancer

A J MacRobert and T Theodossiou, University College London, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

One of the most active areas of photomedical research, in recent years, has been the exploration of the use of light-activated drugs known as photosensitizers. These compounds may be activated using light, usually provided by a laser via an optical fiber which is placed at the site of the target lesion. This treatment is known as photodynamic therapy (PDT) and is being applied to the local destruction of malignant tumors and certain nonmalignancies. Activation of the photosensitizer results in the generation of reactive oxidizing intermediates which are toxic to cells, and this process ultimately leads to tumor destruction. PDT is a relatively low-power, nonthermal, photochemical technique that uses fluence rates not exceeding 200 mW/cm^2 and total light doses or fluences of typically 100 J/cm^2. Generally red or near-infrared light is used, since tissue is relatively transparent at these wavelengths. It is a promising alternative approach to the local destruction of tumors for several reasons. First, it is a minimally invasive treatment since laser light can be delivered with great accuracy to almost any site in the body via thin flexible optical fibers with minimal damage to overlying normal tissues. Second, the nature of PDT damage to tissues is such that healing is safer and more complete than after most other forms of local tissue destruction (e.g., radiotherapy). PDT is also capable of some degree of selectivity for tumors when the sensitizer levels, light doses, and irradiation geometry are carefully controlled. This selectivity is based on the higher sensitizer retention in tumors after administration relative to the adjacent normal tissues in which the tumor arose (generally 3:1 for extracranial tumors but up to 50:1 for brain tumors).

The photosensitizer is administered intravenously to the patient and time allowed (3–96 hours depending on the sensitizer) for it to equilibrate in the body before the light treatment (Figure 1). This time is called the drug–light interval. PDT may also be useful for treating certain nonmalignant conditions, in particular, psoriasis and dysfunctional menorrhagia (a disorder of the uterus), and local treatment of infections; such as genital papillomas, infections in the mouth and the upper gastrointestinal tract. In certain cases, the photosensitizer may be applied directly to the lesions, particularly for treatment of skin tumors, as discussed later. The main side-effect of PDT is skin photosensitivity, owing to retention of the drug in the skin, so patients must avoid exposure to sunlight in particular for a short period following treatment. Retreatment is then possible once the photosensitizer has cleared from the skin since these drugs have little intrinsic toxicity, unlike many conventional chemotherapy agents.

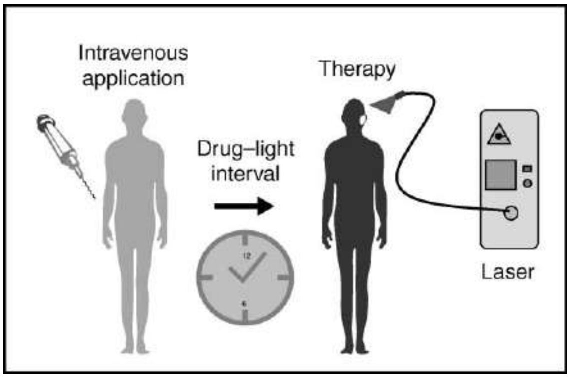

Figure 1 Photodynamic therapy: from sensitization to treatment.

Photoproperties of Photosensitizers

By definition, there are three fundamental requirements for obtaining a photodynamic effect: (a) light of the appropriate wavelength matched to the photosensitizer absorption; (b) a photosensitizer; and (c) molecular oxygen. The ideal photochemical and biological properties of a photosensitizer may be easily summarized, although the assessment of a sensitizer in these terms is not as straightforward as might be supposed because the heterogeneous nature of biological systems can sometimes profoundly affect these properties. Ideally though, a sensitizer should possess the following attributes: (a) red or near infrared light absorption; (b) nontoxic, and with low skin photosensitizing potency; (c) selective retention in tumors relative to normal adjacent tissue; (d) an efficient generator of cytotoxic species, usually singlet oxygen; (e) fluorescence, for visualization; (f) a defined chemical composition, and (g) preferably water soluble. A list of several photosensitizers possessing the majority of the above-mentioned attributes is given in Table 1.

The reasons for these requirements are partly self-evident, but worth amplifying. Strong absorption is desirable in the red and near-infrared spectral region, where tissue transmittance is optimum enabling penetration of the therapeutic light within the tumor (Figure 2). To minimize skin photosensitization by solar radiation, the sensitizer absorption spectrum should ideally consist of a narrow red wavelength band, with little absorption at other wavelengths down to 400 nm, below which solar irradiation falls off steeply. Another advantage of red wavelength irradiation is that the potential mutagenic effects encountered with UV-excited sensitisers (e.g., psoralens) are avoided. Since the object of the treatment is the selective destruction of tumor tissue, leaving surrounding normal tissue undamaged, some degree of selective retention of the dye in tumor tissue is desirable. Unfortunately, the significance of this aspect has been exaggerated in the literature and in many cases treatment selectivity owes more to careful light irradiation geometry. Nevertheless, many normal tissues have the capacity to heal safely following PDT damage.

The key photochemical property of photosensitizers is to mediate production of some active

Table 1 Properties of several photosensitizers

Compound	*λ/nm (ε/M^{-1} cm^{-1})*	*Drug dose/mg kg^{-1}*	*Light dose/J cm^{-2}*	*Diseases treated*
Hematoporphyrin (HpD–Photofrin–Photosan)	628 (3.0×10^3)	1.5–5	75–250	Early stage esophagus, bladder, lung, cervix, stomach, and mouth cancers. Palliative in later stages
ALA (converted to protoporphyrin IX)	635 (5×10^3)	60	50–150	Skin, stomach, colon, bladder, mouth cancers. Esophageal dysplasia. Various nonmalignant conditions
Benzoporphyrin derivative (BpD)	690 (3.5×10^4)	4	150	Age-related macular degeneration (AMD)
Tin etiopurpurin (SnET2–Purlytin)	665 (3.0×10^4)	1.2	150–200	Breast and skin cancers, AMD
Monoaspartyl chlorin e_6 (MACE)	660 (4.0×10^4)	1.0	25–200	Skin cancers
Lutetium texaphyrin (Lu–Tex)	732 (4.2×10^4)	1.0	150	Metastatic brain tumors, breast cancers, atherosclerotic plaques
Aluminum disulfonated phthalocyanine	675 (2.0×10^5)	1.0	50–200	Brain, colon, bladder, and pancreatic cancers. Head and neck cancers in animal studies only
Metatetrahydroxychlorin (mTHPC–temoporphin–Foscan)	652 (3.5×10^4)	0.15	5–20	Head, neck, prostate, pancreas, lung, brain, biliary tract, and mouth cancers. Superior to HpD and ALA in mouth cancers
Palladium pheophorbide (Tookad)	763 (8.6×10^4)	2.0	–	Prostate cancer
Hypericin	590 (3×10^4)	–	–	Psoriasis

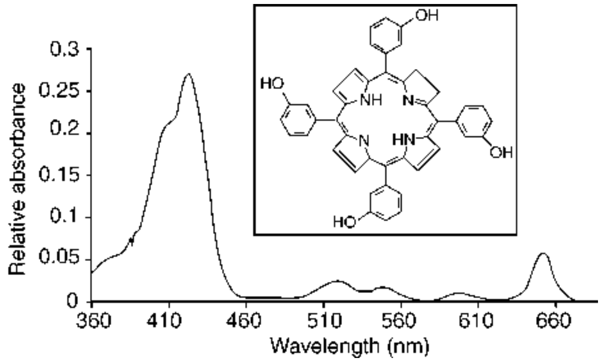

Figure 2 The absorption spectrum of m-THPC. The structure of m-THPC is shown in the inset.

molecule which is cytotoxic, that is, will destroy cells. The first electronically excited state of molecular oxygen, so-called singlet oxygen, fulfills this role very well and may be produced via the interaction of an excited electronic state of the sensitizer with oxygen present in the tissue. Thus, in summary, to achieve effective photosensitization, a sensitizer should exhibit appreciable absorption at red to near-infrared wavelengths and generate cytotoxic species via oxygen-dependent photochemical reactions. The first clinical photosensitizer was hematoporphyrin derivative (HpD), which is derived synthetically from hematoporphyrin by reaction with acetic and sulfuric acids to give a complex mixture of porphyrins. A purified fraction of these (Photofrin) is available commercially, and this has been used most widely in clinical applications to date. Second-generation (photo-sensitizers are now becoming available, including phthalocyanine and chlorin compounds as shown in Table 1.

A new approach to PDT has recently emerged involving the administration of a natural porphyrin precursor, 5-aminolaevulinic acid (ALA), which is metabolized within cells to produce protoporphyrin IX (see Table 1). This porphyrin is known to be a powerful photosensitizing agent but suffers from the drawback of being a poor tumor localizer when used directly. In contrast, administration of ALA induces protoporphyrin biosynthesis, particularly in rapidly proliferating cells, which may then be destroyed using irradiation at 630 nm. Therefore, this new therapy may offer enhanced treatment selectivity with little risk of skin photosensitivity owing to the rapid clearance after 24 h of the protoporphyrin. Investigation of this new approach has already proved successful in clinical treatment of skin tumors using topical application of ALA in a thick emulsion.

Considerable effort is also being expended on exploiting the retention of sensitizers in tumors for diagnostic purposes, although the results are rather mixed to date. The prospects with ALA-induced protoporphyrin IX are, however, more promising, as illustrated in Figure 3 where the fluorescence is selectively confined to the skin tumor (basal cell carcinoma).

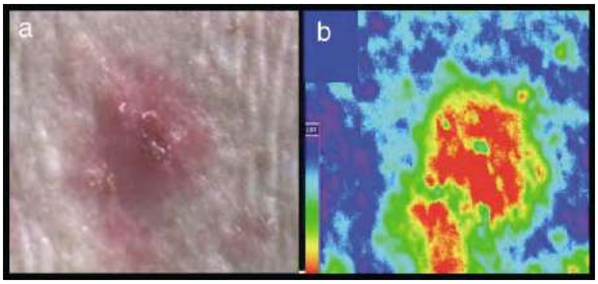

Figure 3 (a) Image of a basal cell carcinoma; (b) fluorescence imaging of the same lesion after ALA sensitization and 405 nm light excitation.

Mechanisms of Photodynamic Therapy

The main principles of the photosensitization mechanism are now well established with the initial step being excitation of the sensitizer from its electronic ground state to the short-lived fluorescent singlet state. The lifetime of the singlet state is generally only a few nanoseconds and the main role of this state in the photosensitization mechanism is to act as a precursor of the metastable triplet state through intersystem crossing. Efficient formation of this metastable state is required because it is the interaction of the triplet state with tissue components that generates cytotoxic species such as singlet oxygen. Thus the triplet state quantum yield (i.e., probability of triplet state formation per photon absorbed) of photosensitizers should ideally approach unity. Interaction of the metastable triplet state (which in de-aerated solutions has a lifetime extending to the millisecond range) with tissue components may proceed via either a type I or II mechanism or a combination (see Figure 4). A type I process can involve hydrogen abstraction from the sensitizer to

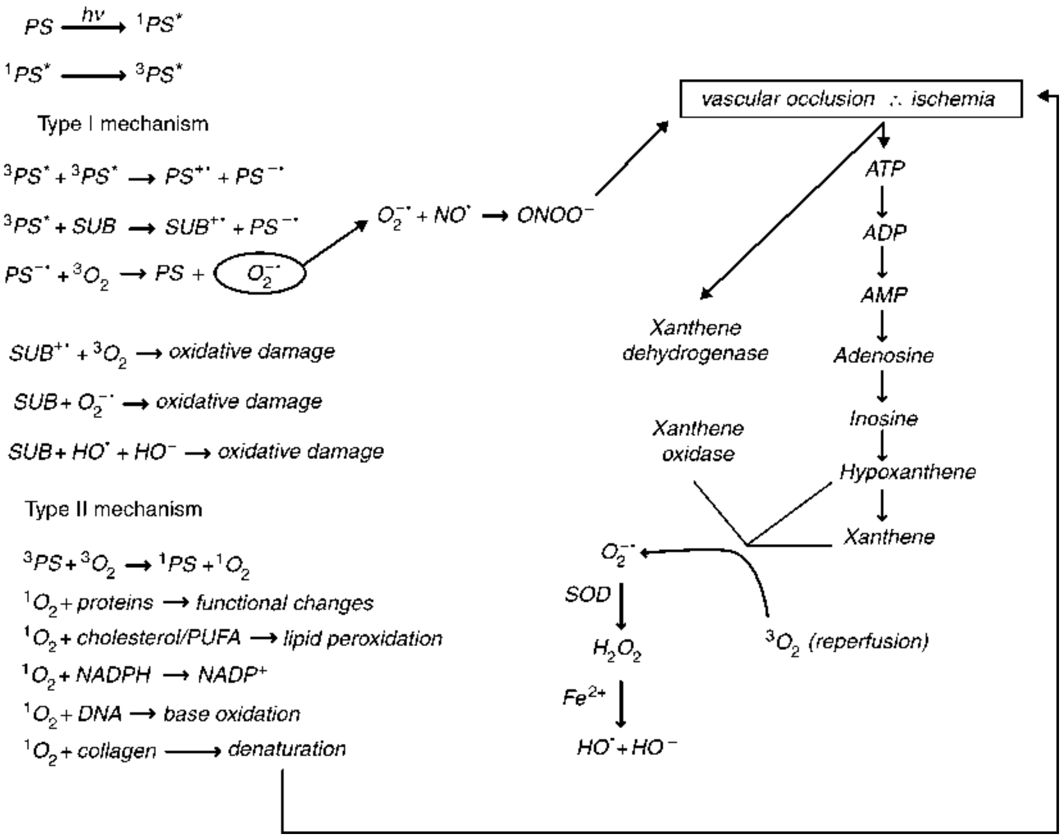
$PS \xrightarrow{h\nu} {}^1PS^*$
${}^1PS^* \longrightarrow {}^3PS^*$
Type I mechanism
${}^3PS^* + {}^3PS^* \longrightarrow PS^{+\cdot} + PS^{-\cdot}$
${}^3PS^* + SUB \longrightarrow SUB^{+\cdot} + PS^{-\cdot}$
$PS^{-\cdot} + {}^3O_2 \longrightarrow PS + O_2^{-\cdot}$
$O_2^{-\cdot} + NO^{\cdot} \longrightarrow ONOO^-$
$SUB^{+\cdot} + {}^3O_2 \longrightarrow$ oxidative damage
$SUB + O_2^{-\cdot} \longrightarrow$ oxidative damage
$SUB + HO^{\cdot} + HO^- \longrightarrow$ oxidative damage
Type II mechanism
${}^3PS + {}^3O_2 \longrightarrow {}^1PS + {}^1O_2$
${}^1O_2$ + proteins ⟶ functional changes
${}^1O_2$ + cholesterol/PUFA ⟶ lipid peroxidation
${}^1O_2 + NADPH \longrightarrow NADP^+$
${}^1O_2$ + DNA ⟶ base oxidation
${}^1O_2$ + collagen ⟶ denaturation
vascular occlusion ∴ ischemia
ATP
ADP
AMP
Adenosine
Inosine
Hypoxanthene
Xanthene
Xanthene dehydrogenase
Xanthene oxidase
$O_2^{-\cdot}$
SOD
H_2O_2
Fe^{2+}
$HO^{\cdot} + HO^-$
${}^3O_2$ (reperfusion)

Figure 4 The photophysical and photochemical mechanisms involved in PDT. PS denotes the photosensitizer, SUB denotes the substrate, either biological, a solvent or another photosensitizer; * denotes the excited state and • denotes a radical.

produce free radicals or electron transfer, resulting in ion formation. The type II mechanism, in contrast, exclusively involves interaction between molecular oxygen and the triplet state to form singlet oxygen which is highly reactive in biological systems. The near-resonant energy transfer from the triplet state to O_2 can be highly efficient and the singlet oxygen yield can approach the triplet state yield provided that the triplet state energy exceeds 94 kJ/mol, the singlet oxygen excitation energy. It is widely accepted that the type II mechanism underlies the oxygen-dependent phototoxicity of sensitizers used for photodynamic therapy. Both proteins and lipids (the main constituents of membranes) are susceptible to photooxidative damage induced via a type II mechanism which generally results in the formation of unstable peroxide species. For example, unsaturated lipids may be oxidized by the 'ene' reaction where the singlet oxygen reacts with a double bond. Other targets include such important biomolecules as cholesterol, certain amino acid residues, collagen, and the coenzyme, NADPH. Another synergistic mode of action involves the occlusion of the microcirculation to tumors and reperfusion injury. Reperfusion injury relates to the formation of xanthene oxidase from xanthene dehydrogenase in anoxic conditions which reacts with xanthene or hypoxanthene products of ATP dephosphorylation to convert the restored oxygen into superoxide anion which directly or via the Fenton reaction causes cellular damage.

An important feature of the type II mechanism, which is sometimes overlooked, is that when the sensitizer transfers electronic energy to O_2 it returns to its ground state. Thus the cytotoxic singlet oxygen species is generated without chemical transformation of the sensitizer which may then absorb another photon and repeat the cycle. Effectively, a singlet photosensitizer molecule is capable of generating many times its own concentration of singlet oxygen, which is clearly a very efficient means of photosensitization provided the oxygen supply is adequate.

Lasers in PDT

Lasers are the most popular light source in PDT since they have several key characteristics that differentiate them from conventional light sources, namely coherence, monochromaticity, and collimated output. The two main practical features that make them so useful in PDT are their monochromaticity and their combination with fiber-optic delivery. The monochromaticity is important since the laser can be tuned to a specific absorption peak of a photosensitizer, thus ensuring that all the energy delivered is utilized for the excitation and photodynamic activation of the photosensitizer. This is not true for a conventional light source (as for example a tungsten lamp) where the output power is divided over several hundred nanometers throughout the UV, visible, and IR regions and only a fraction of the power lies within the absorption band of the photosensitizer. To numerically illustrate this, let us simplify things by representing the lamp output as a square profile and also consider a photosensitizer absorption band about 30 nm full width half maximum. If a laser with a frequency span of about 2 nm and a lamp with the same power but with a frequency span of about 600 nm are used, then the useful portion of it is about 0.05 and if we consider a Gaussian profile for the sensitizer absorption band, the above fraction drops even lower, perhaps even to 0.01. That means that the lamp would achieve 100 times less the excitation rate of the laser, or in other words, we would require a lamp with an output power of 100 times that of the laser to achieve the same rate of excitation and consequently the same time of treatment, provided we use a low enough laser power to remain within the linear regime of the photosensitizer excitation. The other major disadvantage of a lamp source is the lack of collimation which results in low efficiency for fiber-optic delivery.

We now review the different lasers that have found application in PDT. Laser technology has significantly advanced in recent years and there are now a range of options in PDT laser sources for closely matching the absorption profiles of the various photosensitizers. Moreover, since PDT is becoming more widely used clinically, practical considerations, such as portability and turn-key operation are increasingly important. In Figure 5 the optical spectrum is shown from about 380 nm (violet–blue) to 800 nm (near-infrared, NIR). We have superimposed on this spectrum, the light penetration depth of tissue to roughly illustrate how deeply light can penetrate (and consequently activate photosensitizer molecules) into lower-lying tissue layers. The term 'penetration depth' is a measure of the light attenuation by the tissue so the figure of 2 mm corresponds to $1/e$ attenuation of the incident intensity at a tissue depth of 2 mm. It is obvious that light at the blue end of the spectrum has only a superficial effect whereas the more we shift further into the red and NIR regions the penetration depth increases. This is due to two main reasons: absorption and scattering of light by various tissue component molecules. For example, in the blue/green region for the absorption of melanin and hemoglobin is relatively high. But there is a noticeable increase in penetration depth beyond about 600 nm leading to an optimal wavelength region, or

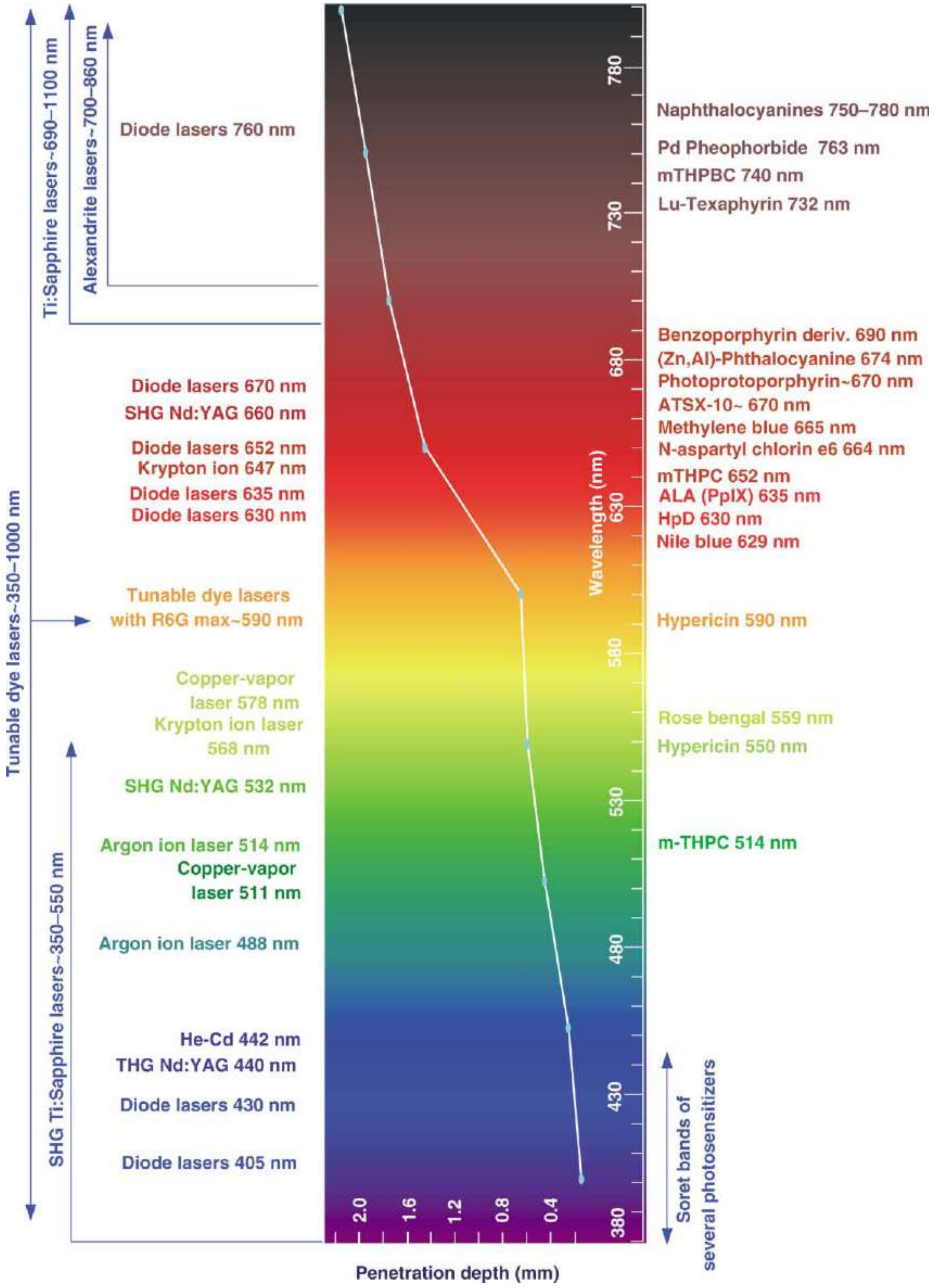

Figure 5 Photosensitizers and laser light sources available in the visible and near-infrared spectral regions.

'therapeutic window' for laser therapy of around 600–1100 nm (Figure 5).

The fundamental wavelength of the Nd:YAG laser lies within this therapeutic window at 1064 nm and this laser is now widely used in thermal laser therapy. The Nd:YAG lasers can be operated either in cw (output power ~ 200 W multimode), long-pulse (~ 500 W average power at 50 Hz), or Q-switched

(50 MW peak power at around 10 ns pulse duration) modes. Nd:YAG is a solid-state laser with a yttrium aluminum garnet crystal doped with about 1% of trivalent Nd ions as the active medium of the laser; and with another transition of the Nd ion, this laser (with the choice of different optics) can also operate at 1320 nm. Although dyes are available which absorb from 800–1100 nm, the generation of singlet oxygen via the type II mechanism is energetically unfavorable because the triplet state energies of these dyes are too low. However, the fundamental frequency of the laser can be doubled (second-harmonic generation, SHG) or tripled (third-harmonic generation, THG) with the use of nonlinear crystals, to upconvert the output radiation to the visible range, thus rendering this laser suitable for PDT: i.e., frequency doubling to 532 nm and tripling to 355 nm. Note that the 532 nm output is suitable for activation of the absorption band of hypericin with a maximum at 550 nm even though it is not tuned to the maximum of this absorption.

In the early days of PDT and other laser therapies, ion lasers were widely used. The argon ion laser uses ionized argon plasma as gain medium, and produces two main wavelengths at 488 nm and 514 nm. The second of the two lies exactly at the maximum of the 514 nm m-THPC absorption. Argon ion lasers are operated in cw mode and usually have output powers in the region of 5–10 W at 514 nm (the most powerful argon ion line). Krypton ion lasers, which emit at 568 or 647 nm, are similar to their argon ion counterparts; however, they utilize ionized krypton plasma as gain medium. The 568 nm output can be used to activate the Rose Bengal peak at 559 nm, whereas the 647 nm most powerful line of krypton ion lasers, has been used for activation of m-THPC (652 nm).

Ideally, for optimal excitation, it is best to exactly tune the laser wavelength to the maxima of photosensitizer absorption bands. For this reason tunable organic dye lasers have been widely used for PDT. In these lasers the gain medium is an organic dye with a high quantum yield of fluorescence. Due to the broad nature of their fluorescence gain profile, tuning elements within the laser cavity can be used for selection of the lasing frequency within the gain band. Tunable dye lasers with the currently available dyes can quite easily cover the range from about 350–1000 nm. However, their operation is not of a 'turn key' nature since they require frequent replacement of their active material either for tuning to a different spectral region or for a better performance when the dye degrades. Tunable dye lasers also need to be pumped by some other light source like an argon ion laser, an excimer laser, a solid state laser (e.g., Nd:YAG 532 or 355 nm), a copper vapor laser, or lamps. Dye lasers have been used for the excitation of HpD at 630 nm, protoporphyrin IX at 635 nm, photoprotoporphyrin at 670 nm, and phthalocyanines around 675 nm. A further example is hypericin which, apart from the band at 550 nm, has a second absorption band at 590 nm. This is the optimum operation wavelength of dye lasers with rhodamine 590, better known as rhodamine 6G.

Dye lasers can operate in either pulsed or cw mode. However, there is a potential disadvantage in using a pulsed laser for PDT which becomes apparent when using lower repetition rates and higher pulse energies (>1 mJ per pulse) laser excitation. High pulse energies can induce saturation or transient bleaching of the sensitizer during the laser pulse and consequently much of the energy supplied is not efficiently absorbed by the sensitizer. It has been suggested that this effect accounts for the lack of photosensitized damage in tissue sensitized with a phthalocyanine and irradiated by a 5 Hz, 25 mJ-per-pulse flash lamp pumped dye laser. However, using a low pulse energy, high repetition rate copper vapor pumped dye laser (see below), the results were indistinguishable from cw irradiation with an Argon ion dye laser.

Another laser that has found clinical application in PDT is the copper vapor laser. In this laser the active medium is copper vapor at high temperature maintained in the tube by a repetitively pulsed discharge current. The copper vapor lasers are pulsed with typical pulse duration of about 50 ns and repetition rates reaching 20 kHz. Copper vapor lasers have been operated at quite high average powers up to about 40 W. The output radiation is produced at two distinct wavelengths, namely 511 and 578 nm both of which have found clinical use. Copper vapor pumped dye lasers have also been widely used for activating red-absorbing photosensitizers, and the analogous gold vapor lasers operating at 628 nm have been used to activate HpD.

A relatively new class of tunable lasers is the solid state Ti:Sapphire laser. The gain medium in this laser is a ~1%Ti doped sapphire (Al_2O_3) crystal and its output may be tuned throughout the 690–1100 nm region, covering many photosensitizer absorption peaks. For example, most of the bacteriochlorin type photosensitizers have their absorption bands in that spectral region, e.g., m-THPBC with an absorption maximum at 740 nm. Also, Texaphyrin sensitizers absorb highly in this spectral region: lutetium texaphyrin has an absorption maximum at 732 nm.

Although tunable dye or solid-state lasers are widely used in laboratory studies for PDT, fixed-wavelength semiconductor diode lasers are gradually supplanting tunable lasers, owing to their practical convenience. It is now generally the case that when a photosensitizer enters clinical trials or enters standard clinical use, laser manufacturers provide dedicated diode lasers with outputs matched to the chosen sensitizer. In this context there are diode lasers available at 630 nm for use with HpD, at 635 nm for use with ALA-induced protoporphyrin IX, 670 nm for ATSX-10 or phthalocyanines, or even in the region of 760 nm, for use with bacteriochlorins. The advantage of these diode lasers is that they are tailormade for use with a particular photosensitizer, they are highly portable, and they have a relatively easy turn-key operation.

So far we have concentrated on red wavelengths but most of the porphyrin and chlorin family of photosensitizers exhibit quite strong absorption bands in the blue region, known as 'Soret' bands. Despite the fact that tissue penetration in this spectral region is minimal, these bands are far more intense than the corresponding red absorption bands of the sensitizer. In this respect it is possible to activate these sensitizers at blue wavelengths, especially in the case of treatment of superficial (e.g., skin) malignant lesions. There are now blue diode lasers (and light-emitting diodes) for the activation of these bands, in particular at 405 nm where the Soret band of protoporphyrin IX lies or at 430 nm for selective excitation of the photoprotoporphyrin species. For very superficial lesions (e.g., in the retina) it may even be possible to use multiphoton excitation provided by femtosecond diode lasers operating at about 800 nm.

Clinical Applications of Lasers in PDT

In the previous section we reviewed the various types of lasers being used for PDT but their combination with fiber-optic delivery is also of key importance for their clinical application. The laser light may be delivered to the target lesion either externally using surface irradiation, or internally within the lesion which is denoted as interstitial irradiation, as depicted in Figure 6. For example, in case of superficial cutaneous lesions the laser irradiation is applied externally (Figure 6a). The area of the lesion is marked prior to treatment to determine the diameter of the laser spot. For a given fluence rate (100–200 W/cm^2) and the area of the spot, the required laser output power is then calculated. A multimode optical fiber, terminated with a microlens to ensure uniform irradiation, is then positioned at a distance from the lesion yielding the desired beam waste, and the surrounding normal tissue is shielded with dark material. However, if the lesion is a deeper-lying solid tumor, interstitial PDT is employed

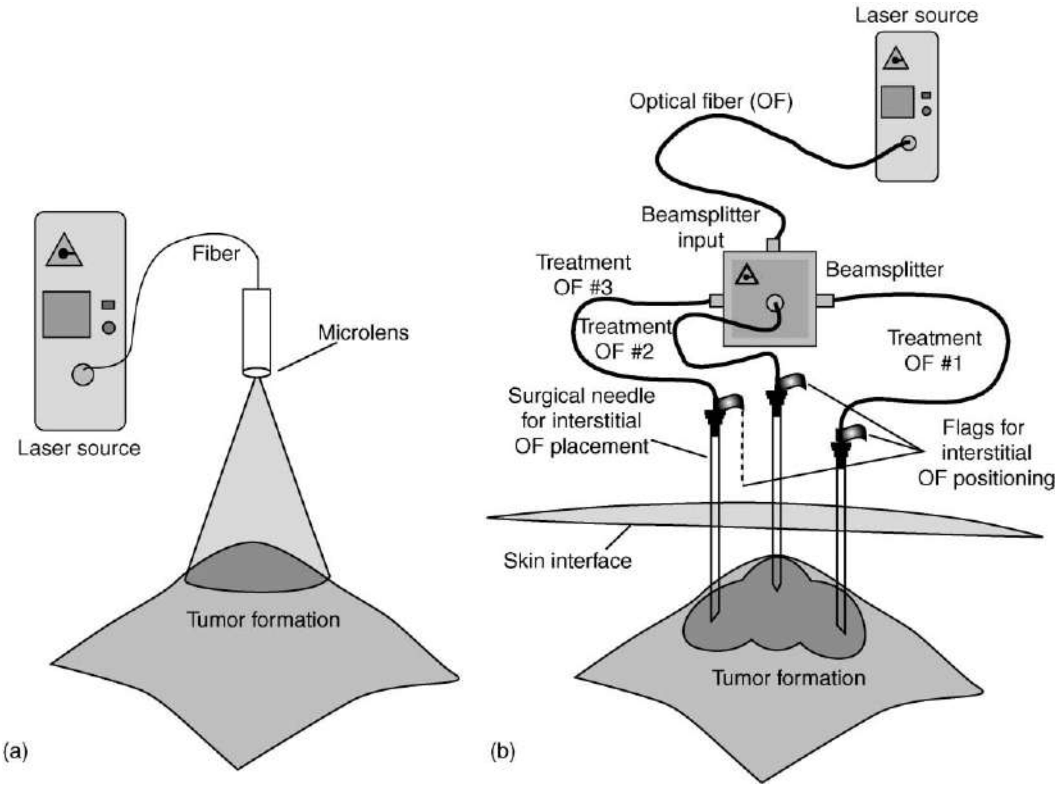

Figure 6 Clinical application of PDT. (a) Surface treatment. (b) Interstitial treatment.

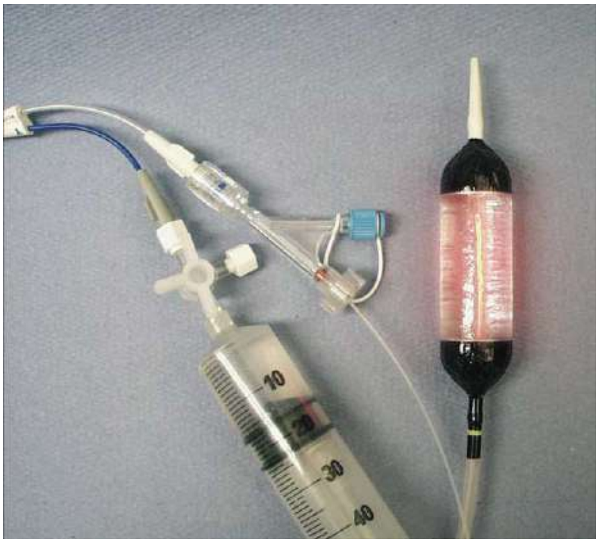

Figure 7 Balloon applicator and diffuser fiber used for PDT in Barrett's esophagus.

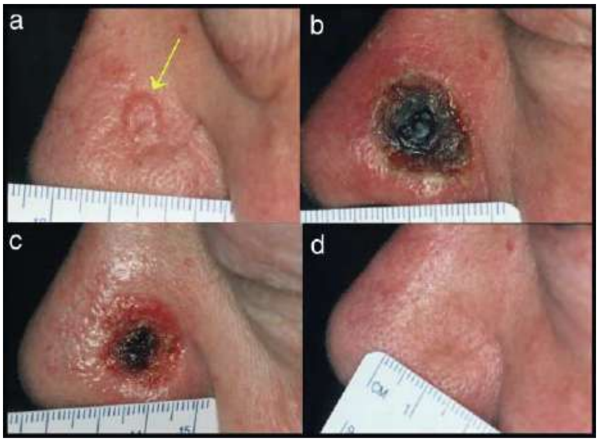

Figure 8 PDT treatment of a nasal basal cell carcinoma with m-THPC (Foscan®): (a) prior to PDT; (b) tissue necrosis a week after treatment; (c) four weeks after treatment; gradual recession of the necrosis and inflammation; (d) healing two months after treatment. Courtesy of Dr Alex Kuebler.

(Figure 6b). Surgical needles are inserted in an equidistant parallel pattern within the tumor, either freehand or under image (MRI/CT) guidance. Bare tip optical fibers are guided through the surgical needles to the lesion and flagged for position. A beamsplitter is used to divide the laser output into 2–4 components so that two to four optical fibers can be simultaneously employed. The laser power is adjusted so that the output fluence of each of the optical fibers is 100–200 mW. Even light distribution in the form of overlapping spheres approximately 1 cm diameter ensures the treatment of an equivalent volume of tissue around each fiber tip. Once each 'station' is likewise treated the fibers are repositioned accordingly, employing a pull-back technique in order for another station to be treated. Once the volume of the tumor is scanned in this manner the treatment is complete. Interstitial light delivery allows PDT to be used in the treatment of large, buried tumors and is particularly suitable for those in which surgery would involve extensive resection.

The treatment of tumors of internal hollow organs is also possible with PDT. The most representative example of this is the treatment of precancerous lesions in the esophagus, known medically as 'Barrett's esophagus'. In this case, a balloon applicator is used to house a special fiber with a light-diffusing tip or 'diffuser' (Figure 7). These diffusers are variable in length and transmit light uniformly along the whole of their length in order to facilitate treatment of circumferential lesions. The balloon applicator is endoscopically inserted into the patient. The balloon is inflated to stretch the esophageal walls and the flagged diffuser fiber is placed in position, central to the balloon optical window. In this way the whole of the lesion may be treated circumferentially at the same time.

Finally, the clinical response to PDT treatment is shown in Figure 8. The treated area becomes necrotic and inflamed within two to four days following PDT. This necrotic tissue usually sloughs off or is mechanically removed. The area eventually heals with minimal scarring (Figure 8).

Future Prospects

A major factor in the future development of PDT will be the application of relatively cheap and portable semiconductor diode lasers. High-power systems consisting of visible diode arrays coupled into multimode fibers with an output of several Watts, have recently become available which will greatly simplify the technical difficulties which have held back PDT. The use of new photosensitizers, in the treatment of certain tumors by PDT, is making steady progress, and although the toxicology testing and clinical evaluation are lengthy and costly processes, the expectation is that these compounds will become more widely available for patient use during the next five years. Regarding the current clinical status of PDT, Photofrin has recently been approved for treatment of lung and esophageal cancers in Europe, USA, and Japan. Clinical trials are in progress for several of the second-generation photosensitizers which offer significant improvements over Photofrin in terms of chemical purity, photoproperties, and skin clearance. With the increasing clinical application of compact visible diode lasers the prospects for photodynamic therapy are therefore encouraging.

See also

Lasers: Dye Lasers; Free Electron Lasers; Metal Vapor Lasers. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Bonnet R (2000) *Chemical Aspects of Photodynamic Therapy*. Amsterdam: Gordon and Breach Science Publishers.

Dolmans DEJGJ, Fukumura D and Jain RK (2003) Photodynamic therapy for cancer. *Nature Reviews (Cancer)* 3: 380–387.

Hopper C (2000) Photodynamic therapy: a clinical reality in the treatment of cancer. *Lancet Oncology* 1: 212–219.

Milgrom L and MacRobert AJ (1998) Light years ahead. *Chemistry in Britain* 34: 45–50.

Rosen GM, Britigan BE, Halpern HJ and Pou S (1999) *Free Radicals Biology and Detection by Spin Trapping*. New York: Oxford University Press.

Svelto O (1989) *Principles of Lasers*, 3rd edn. New York: Plenum Press.

Vo-Dinh T (2003) *Biomedical Photonics Handbook*. Boca Raton, FL: CRC Press.

Pump and Probe Studies of Femtosecond Kinetics

G D Scholes, University of Toronto, Toronto, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Pump-probe spectroscopy is a ubiquitous time-resolved optical spectroscopy that has found application to the study of all manner of ultrafast chemical processes that involve excited states. The principle of the measurement is that a 'pump pulse' – usually an intense, short laser pulse – impulsively excites a sample, thus defining the start time for the ensuing photophysical dynamics. A probe pulse interrogates the system at later times in order to capture 'snap shots' of the state of the system. Many kinds of pump-probe experiments have been devised. The present article will mainly focus on transient absorption measurements.

Femtosecond pump-probe experiments have been employed to reveal the kinetics of excited state chemical processes such as solvation, isomerization reactions, proton transfer, electron transfer, energy transfer, or photochromism. It is becoming more common to use pump-probe techniques to study systems of increasing complexity, such as photosynthetic proteins, photoreceptor proteins, molecular aggregates, nanostructured materials, conjugated polymers, semiconductors, or assemblies for artificial light harvesting. However, the underlying chemical kinetics are not always easily revealed by pump-probe signals.

The principle of the pump-probe measurement is related to the seminal flash-photolysis experiment devised by Norrish and Porter. However, in order to achieve ultrafast time resolution – beyond that attainable by electronic detection – the probe pulse is temporally short and is controlled to arrive at the sample at variable time delays after the pump pulse has excited the sample. The change in intensity of the probe pulse, as it is transmitted through the sample, is monitored at each pump-probe time delay using a 'slow' detector that integrates over the probe pulse duration. One possible xprocess that can diminish the probe transmission is transient absorption. After formation of an excited state S_1 by the pump pulse, the probe pulse can monitor a resonance with a higher excited state $S_1 \rightarrow S_n$ absorption. Figure 1 shows transient absorption spectra, corresponding to excited state absorption at various time delays, that has been probed by a white light continuum probe and monitored by a multichannel CCD detector. The transient spectrum is seen to blue-shift with

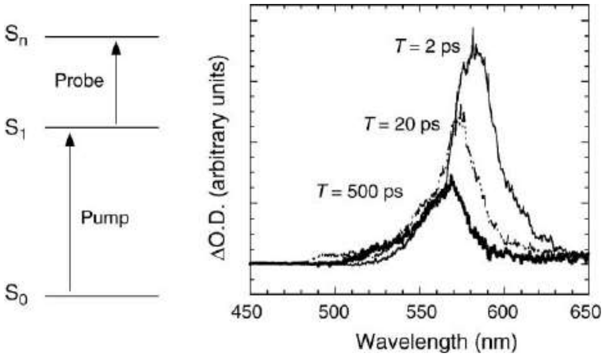

Figure 1 Excited state absorption spectra of 4′-*n*-pentyl-4-cyanoterphenyl in octanol solvent. The pump wavelength is 277 nm (100 nJ, 150 fs, 40 kHz) and the probe is a white light continuum (500 to 700 nm) generated in a sapphire crystal. The pump-probe delay *T* for each spectrum is indicated on the plot.

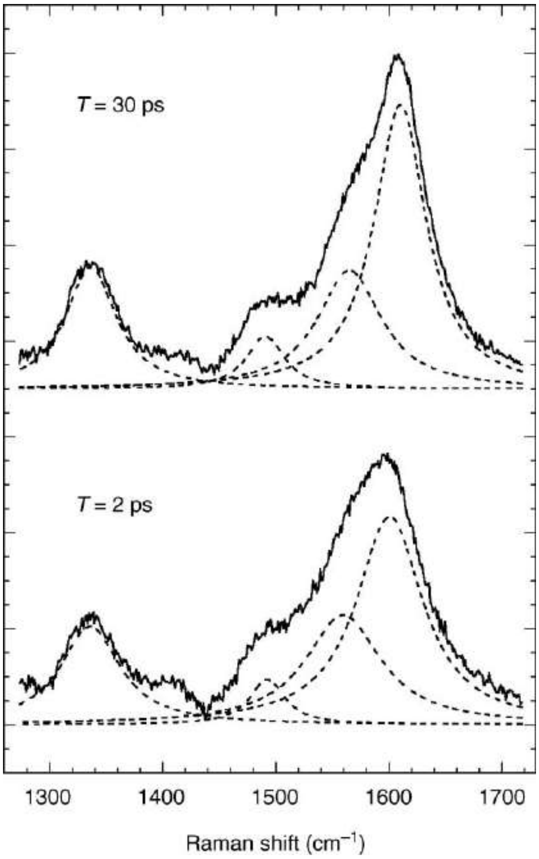

Figure 2 TR^3 spectra of 4′-*n*-pentyl-4-cyanoterphenyl in octanol solvent, recorded using the same set-up as summarized in Figure 1. The dashed lines are deconvolutions of the data to show the Raman bands.

pump-probe delay owing to solvation of the S_1 state of the molecule – the dynamic Stokes' shift. The dynamic Stokes' shift is the shift to lower energy of the emission spectrum as the solvent relaxes in response to the change in dipole moment of the excited electronic state compared to the ground state. The overall intensity of the induced absorption decays according to the lifetime of the excited state.

Although they will not be discussed in detail in the present article, various other pump-probe experiments are possible. For example, when in resonance with this transient absorption, the probe pulse can induce resonance Raman scattering, which reveals the evolution of the vibrational spectrum of the excited state. In this kind of pump-probe experiment we do not monitor the change in probe transmission through the sample. Instead, we monitor the difference between probe-induced resonance Raman scatter from the excited and ground states of the system. The resonance Raman (TR^3) scatter, time-resolved as a function of pump-probe delay, can be detected provided that it is not overwhelmed by fluorescence emission. Typical TR^3 data for different pump-probe delays are shown in Figure 2. The dashed lines show a deconvolution of the spectra to reveal the underlying vibrational bands. These bands are rather broad owing to the spectral width of the 1 ps laser pulses used in this experiment. It is clear that the time-resolution of a TR^3 experiment is limited to the picosecond regime, otherwise most of the vibrational band information is lost. TR^3 and complementary pump-probe infrared transient absorption experiments, utilizing an infrared probe beam, have been useful for observing the evolution of structural information, such as geometry changes, subsequent to photoexcitation.

In the present article we will describe the foundation of experiment and theory necessary to understand resonant ultrafast pump-probe spectroscopy applied to chemical processes in the condensed phase. By resonant, it is meant that the pump and probe pulses have frequencies (energies) resonant with electronic transitions of the molecules being studied. The implications of condensed phase are that the experiment interrogates an ensemble of molecules and the electronic transitions of these molecules are coupled to the random motions and environments of a bath, for example the solvent.

Experimental Measurement of Pump-Probe Data

The experimental setup for a one-color pump-probe experiment (i.e. pump and probe of the same color) is shown in Figure 3. The setup is easily adapted for a two-color experiment. Most of the laser intensity (70%) is transmitted through the first beam splitter so that the pump is more intense than the probe. A second beam splitter is used to split off a small amount of probe light for the reference beam. Note that the arrangement of these beam splitters results in both the pump and probe passing through the same amount of dispersive material on their way to the sample. The pump and probe beams each propagate through half-wave plates in order to control polarization. Usually the probe polarization is set to the magic angle (54.7°) relative to the pump in order to remove polarization bias. The beams are sent towards the sample by way of retroreflectors mounted on $x-y-z$ translation stages. The pump retroreflector is mounted on a precision computer-controlled translation stage such that it can be scanned in the x-direction to arrive at variable delays before and after the probe. The pump and probe beams are aligned using the $y-z$ controls on the retroreflector translation stages together with the pick-off mirrors

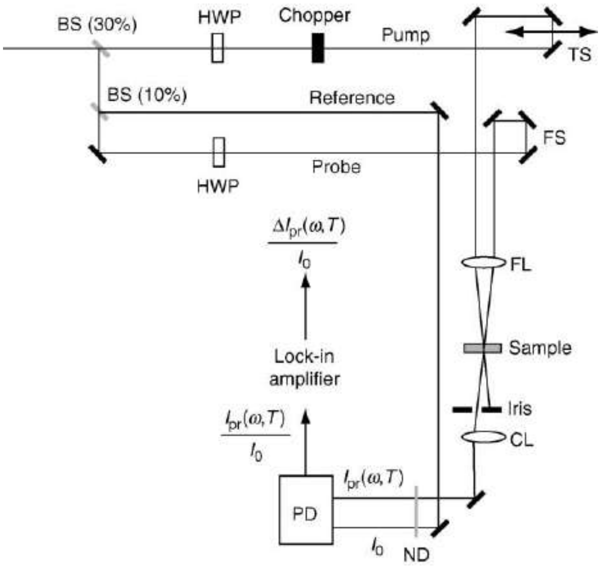

Figure 3 An experimental layout for femtosecond pump-probe spectroscopy. The reference beam ideally passes through the sample (not shown as such in aid of clarity). See text for a description. HWP, half-wave plate; BS, beamsplitter (% reflection); FL, focusing lens; CL, collimating lens; ND, neutral density filter; PD, photodetectors.

such that they travel in parallel towards the sample. The overall setup of the beams is in a box geometry so that pump and probe arms have the same path length.

The pump and probe beams are focused into the sample using a transmissive (i.e. lens) or reflective optic FL. The focal length is typically 20 to 30 cm, providing a small crossing angle to improve phase matching. The pump beam should have a larger spot size at its focal point in the sample than the probe beam, in order to avoid artifacts arising from the wandering of the pump and probe spatial overlap as a function of delay time – a potential problem, particularly for long delays. Good spatial overlap of the beams in the sample is important. It can be checked using a 50 μm pinhole. Fine adjustment of the pump-probe delay, to establish temporal overlap of the pump and probe pulses, can be dictated by autocorrelating the beams in a second harmonic generating crystal mounted at the sample position. At the same time the pulse compression is tweaked to ensure that the pulse dispersion in the experiment has been pre-compensated.

The sample is usually flowed or mounted in a spinning cell if it is liquid, in order to avoid thermal effects or photochemical bleaching. The path length of the sample is typically ≤ 1 mm. The transmitted probe beam is spatially isolated from the pump beam using an iris, is collimated, and then directed onto the photodetector.

The reference beam, which provides the relative reference intensity I_0 should preferably pass through the sample (away from the pump spot). This provides the probe-only transmission – that is, the reference beam intensity is attenuated according to sample ground state optical density. The photodetector outputs are used to ratio the pump-probe arm intensity transmitted through the sample with the reference intensity, $I_{pr}(\omega, T)/I_0$. Since the pump beam is being chopped at the lock-in reference frequency, the lock-in amplifier outputs the pump-induced fractional change in transmission intensity, $\Delta I_{pr}(\omega, T)/I_0$, usually simply written as $\Delta T/T$. When the pump-induced signal I_{pr} is small compared to the reference intensity I_0, then the detected $\Delta I_{pr}(\omega, T)/I_0$ signal is approximately equal to the change in optical density, ΔO.D.:

$$\Delta\text{O.D.} = \log\left(\frac{I_0 - \Delta I_{pr}}{I_0}\right) \approx \frac{-\Delta I_{pr}}{I_0} \equiv \frac{-\Delta T}{T} \qquad [1]$$

Here we have used a Taylor expansion $\log(1 + x) \approx x$ for small x.

The power dependence of the signal in the $\chi^{(3)}$ limit should be linear in both the pump and probe intensities. It is also possible to pump the sample using two-photon absorption, then probe an excited state absorption. This is a $\chi^{(5)}$ experiment, so that the signal depends quadratically on the pump intensity.

What is Measured in a Pump-Probe Experiment?

Origin of the Signal

The pump-probe measurement is a third-order nonlinear spectroscopy. A complete calculation of the signal requires computation of the third-order polarization induced by the interaction of the pump and probe with the material sample, together with an account for the kinetic evolution of populations (e.g., excited state reactant and product states, etc.).

The pump pulse, wavevector $\mathbf{k}_1$, interacts twice with the sample, thereby creating excited state population density $|e\rangle\langle e|$ and a hole-in-the-ground state population density $|g\rangle\langle g|$. These population densities propagate until the probe pulse, with wavevector $\mathbf{k}_2$, interacts with the system to induce a polarization that depends on the state of the system at time T after the pump. This induced polarization $P_s(t)$ is radiated in the $\mathbf{k}_s = \mathbf{k}_1 - \mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_2$ direction. Because the signal is radiated in the probe direction, the probe pulse acts as a local oscillator to heterodyne the signal. The change in probe-pulse spectral intensity after transmission through an

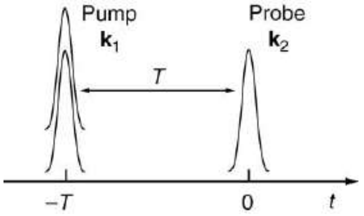

Figure 4 Labeling and time variable for the description of a pump-probe experiment. T defines the pump-probe delay.

optically thin sample of path length l, with concentration of absorbers c and refractive index n, is given by eqn [2]

$$\Delta I_{\mathrm{pr}}(\omega, T) \propto \frac{cl\omega_{\mathrm{pr}}}{n}\mathrm{Im}\left[\int dt E^*_{\mathrm{pr}}(t)P^{(3)}(0,T,t)e^{i\omega t}\right] \qquad [2]$$

In eqn [2] $E_{\mathrm{pr}}(t)$ is the probe pulse electric field, ω is the center frequency of the probe, and * indicates complex conjugate. The time variables correspond to those indicated in Figure 4, and the signal is resolved as a function of frequency by Fourier transformation of $t \to \omega$. Experimentally this is achieved by dispersing the transmitted probe using a spectrograph. The first-time variable, the time interval between the two pump pulse interactions, is set to zero because we are interested here in the limit where the pulse duration is much shorter than T. In this regime – impulsive pump-probe – the measurement is sensitive to formation and decay of excited state species.

The induced polarization is given by eqn [3]

$$P^{(3)}(0,T,t) = \int_0^\infty \mathrm{d}t \int_0^\infty \mathrm{d}T[R^{\mathrm{SE}} - R^{\mathrm{ESA}} + R^{\mathrm{GSR}}] \times E_1^*(t-T)E_1(t-T)E_2(t) \qquad [3]$$

where the response functions that contain information about the lineshape functions and kinetics are given by eqn [4]:

$$R^{\mathrm{SE}}(0,T,t) = |\mu_{eg}|^2|\mu_{eg}|^2\exp(-i\omega_{eg}t) \times \exp[-g^*(t) + 2ig''(T+t) - 2ig''(T)] \times K^{\mathrm{SE}}(T) \qquad [4a]$$

$$R^{\mathrm{ESA}}(0,T,t) = |\mu_{eg}|^2|\mu_{fe}|^2\exp(-i\omega_{fe}t) \times \exp[-g^*(t) - 2ig''(T+t) + 2ig''(T)] \times K^{\mathrm{ESA}}(T) \qquad [4b]$$

$$R^{\mathrm{GSR}}(0,T,t) = |\mu_{eg}|^2|\mu_{eg}|^2\exp(-i\omega_{eg}t)\exp[-g(t)] \times K^{\mathrm{GSR}}(T) \qquad [4c]$$

The overall intensity of each contribution to the signal is scaled by the transition dipole moment that connects the ground and excited state μ_{eg} or excited state and a higher excited state μ_{fe}. The intensity is further influenced by resonance of the probe-pulse spectrum with the transition frequencies ω_{eg} and ω_{fe}. These transition energies have a time-dependence owing to the relaxation of the excited state density in the condensed phase environment – the dynamic Stokes' shift. The details of the dynamic Stokes' shift depend on the bath, and are therefore contained in the imaginary part of the lineshape function $g(t) = g'(t) + ig''(t)$. The evolution of the Stokes' shift with pump-probe delay T is clearly seen in the excited state absorption spectrum shown in Figure 1. The lineshape function contains all the details of the time-scales of fluctuations of the bath and the coupling of the electronic transition of the molecule to these fluctuations. The lineshape function can be cast in the simple form given by eqn [5] using the Brownian oscillator model in the high temperature, high friction limit:

$$g(t) \approx (2\lambda k_B T_B/\hbar\Lambda)[\Lambda t - 1 + \exp(-\Lambda t)] + i(\lambda/\Lambda)[1 - \exp(-\Lambda t)] \qquad [5]$$

The Brownian oscillator model attributes fluctuations of the transition frequencies to coupling between electronic states and an ensemble of low-frequency bath motions. In eqn [5]; λ is the solvent reorganization energy (half the Stokes' shift) and Λ is the modulation frequency of the solvent oscillator (τ_{L}^{-1} for a Debye solvent, where τ_{L} is the longitudinal dielectric relaxation time); k_{B} is the Boltzmann constant, and T_{B} is the temperature.

Information about the kinetics of the evolution of the system initiated by the pump pulse is contained in the terms K^{SE}, K^{ESA}, and K^{GSR}. These terms will be discussed in the next section.

The evolution of the excited state density is mapped onto the response functions R^{SE} and R^{ESA}. R^{SE} is the contribution arising from the probe pulse stimulating emission from the excited state e in the probe direction, which increases the probe intensity. Thus the spectrum of the signal arising from R^{SE} resembles the fluorescence emission of the sample, as shown in Figure 5. The excited state absorption contribution to the signal R^{ESA} depletes the probe intensity, and corresponds to an e → f electronic resonance, such as that shown in Figure 1. Note that the ESA signal contributes to the overall $\Delta I_{\mathrm{pr}}(\omega, T)$ with an opposite sign to the SE and GSR contributions. The ground-state recovery term R^{GSR} arises from depletion of the ground-state population by the pump pulse, thereby increasing the transparency of the sample over the

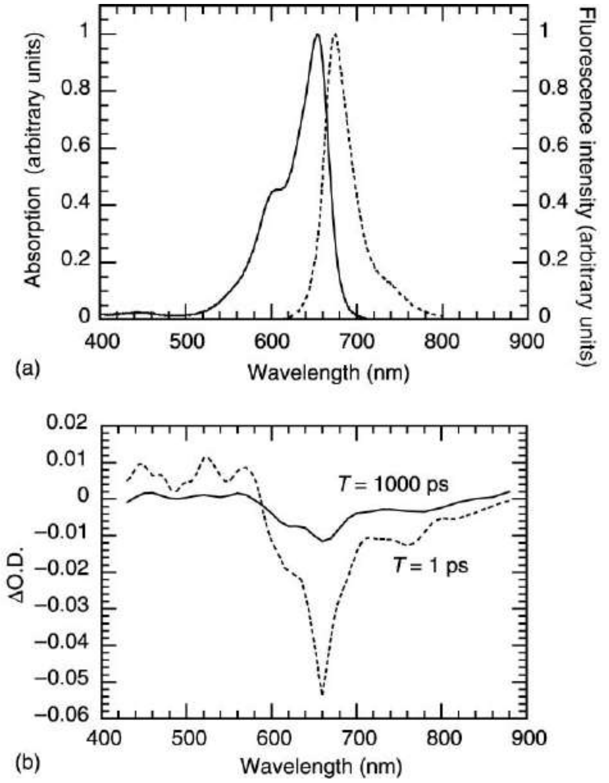

Figure 5 (a) shows the absorption and emission spectrum of a dilute solution of oxazine-1 in water. (b) shows the transient absorption spectrum of this system at various delays after pumping at 600 nm using picosecond pulses. The transient spectrum, ground state recovery and stimulated emission resemble the sum of the absorption and emission spectra.

spectrum of the ground-state absorption (see Figure 5). Thus the transmitted probe intensity is increased relative to that without the pump pulse. The GSR contribution to $\Delta I_{\mathrm{pr}}(\omega, T)$ decreases with T according to the excited state lifetime of the photo-excited species.

The Coherent Spike

To calculate the pump-probe signal over delay times that are of the order of the pulse duration, one must calculate all possible ways that a signal can be generated with respect to all possible time orderings of the two pump interactions and the probe interactions (Liouville space pathways). Such a calculation reveals coherent as well as sequential contributions to the response of the system owing to the entanglement of time orderings that arise from the overlap of pulses of finite duration. The nonclassical, coherent, contributions dominate the signal for the time period during which pump and probe pulses overlap, leading to the coherent spike (or coherent 'artifact') (Figure 6).

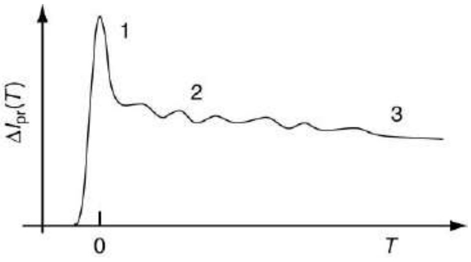

Figure 6 A schematic depiction of a pump-probe signal. The coherent spike is labeled 1. In region 2 we can see quantum beats as a sum of sinusoidal modulations of the signal. The label 3 denotes the long time decay of the signal, as determined by the population dynamics, K^{SE}, K^{GSR}, and K^{ESA}.

Nuclear Wavepackets

The electronic absorption spectrum represents a sum over all possible vibronic transitions, each weighted by a corresponding nuclear overlap factor according to the Franck–Condon principle. It was recognized by Heller and co-workers that in the time-domain representation of the electronic absorption spectrum, the sum over vibrational overlaps is replaced by the propagation of an initially excited nuclear wavepacket on the excited state according to $|i(t)\rangle = \exp(-Ht/\hbar)\,|i(0)\rangle$. Thus the absorption spectrum is written as

$$\sigma_i(\omega) = \frac{2\pi\omega}{3c}\int_{-\infty}^{\infty} \mathrm{d}t \exp[-i(\omega - \omega_{eg}^{i})t - g(t)] \times \langle i(0)\,|i(t)\rangle \qquad [6]$$

for mode i. The Hamiltonian H is determined by the displacement of mode i, Δ_i. A large displacement results in a strong coupling between the vibration and the electronic transition from state g to e. In the frequency domain this is seen as an intense vibronic progression. In the time domain this translates to an oscillation in the time-dependent overlap $\langle i(0)\,|i(t)\rangle$, with frequency ω_i, thereby modulating the intensity of the linear response as a function of t.

It follows that an impulsively excited coherent superposition of vibrational modes, created by an optical pulse that is much shorter than the vibrational period of each mode, propagates on the excited state as a wavepacket. The oscillations of this wavepacket then modulate the intensity of the transmitted probe pulse as a function of T to introduce features into the pump-probe signal known as quantum beats (see Figure 6). In fact, nuclear wavepackets propagate on both the ground and excited states in a pump-probe experiment. The short probe pulse interrogates the evolution of the nuclear wavepackets by projecting them onto the manifold of vibrational wavefunctions in an excited or ground state.

Femtosecond Kinetics

According to eqn [4] the kinetics of the evolution of the system initiated by the pump pulse are governed by the terms K^{SE}, K^{ESA}, and K^{GSR}. Each of these terms describes the kinetics of decay, production, or recovery of excited state and ground state populations. In the simplest conceivable case, each might be a single exponential function. The stimulated emission term K^{SE} contains information about the depopulation of the initially excited state, which for example, may occur through a chemical reaction involving the excited state species or energy transfer from that state. The excited state absorption term K^{ESA} contains information on the formation and decay of all excited state populations that give a transient absorption signal. The ground-state recovery term K^{GSR} contains information on the timescales for return of all population to the ground state, which will occur by radiative and nonradiative processes. Usually the kinetics are assumed to follow a multi-exponential law such that

$$K^m(T, \lambda_{exc}, \omega) = \sum_{j=1}^{N} A_j(\lambda_{exc}, \omega) \exp(-t/\tau_j) \qquad [7]$$

where the amplitude A_j, but not the decay time coefficient τ_j, of each contribution in the coupled N-component system depends on the excitation wavelength λ_{exc} and the signal frequency ω. The index m denotes SE, ESA, or GSR.

In order to extract meaningful physical parameters from analysis of pump-probe data, a kinetic scheme in the form of eqn [7] needs to be constructed, based on a physical model to connect the decay of initially excited state population to formation of a product excited state and subsequent recovery of the ground-state population. An example of such a model is depicted in Figure 7. Here the stimulated emission contribution to the signal is attenuated according to the rate that state 1 goes to 2. The transient absorption appears at that same rate and decays according to the radiationless process 2 to 3. The rate of ground state recovery depends on 1 to 2 to 3 to 5 and possibly pathways involving the isomer 4.

The kinetic scheme is not easily extracted from pump-probe data because contributions from each of K^{SE}, K^{ESA}, and K^{GSR} contribute additively at each detection wavelength ω, as shown in eqns [2]–[4]. Moreover, additional, wavelength-dependent, decay or rise components may appear in the data on ultrafast time-scales owing to the time-dependent Stokes' shift (the $g''(t)$ terms indicated in eqns [4]). A spectral shift can resemble either a decay or rise component. In addition, the very fastest dynamics may be hidden beneath the coherent spike. For these reasons, simple fitting of pump-probe data usually cannot reveal the underlying kinetic model. Instead the data should be simulated according to eqns [3] and [4].

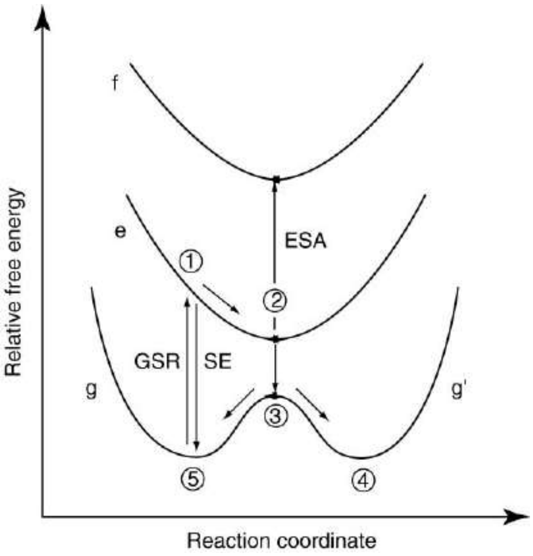

Figure 7 Free energy curves corresponding to a model for an excited state isomerization reaction, with various contributions to a pump-probe measurement indicated. See text for a description. (Courtesy of Jordanides XJ and Fleming GR.)

Alternative strategies involve global analysis of wavelength-resolved data to extract the kinetic evolution of spectral features, known as 'species associated decay spectra'. Global analysis of data effectively reduces the number of adjustable parameters in the fitting procedure and allows one to obtain physically meaningful information by associating species with their spectra. These methods are described in detail by Holzwarth.

List of Units and Nomenclature

ESA	excited state absorption
GSR	ground state recovery
SE	stimulated emission
S_n	nth singlet excited state
TR^3	time-resolved resonance Raman
$\chi^{(3)}$	nth-order nonlinear optical susceptibility

See also

Coherent Lightwave Systems. **Coherent Transients:** Ultrafast Studies of Semiconductors. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. New York: Oxford University Press.

Heller EJ (1981) The semi-classical way to molecular spectroscopy. *Accounts of Chemical Research* 14: 368–375.

Holzwarth AR (1996) Data analysis of time-resolved measurements. In: Amesz J and Hoff AJ (eds) *Biophysical Techniques in Photosynthesis*, pp. 75–92. Dordrecht, Germany: Kluwer.

Ippen EP, Shank CV, Wiensenfeld JM and Migus A (1980) Subpicosecond pulse techniques. *Philosophical Transactions of the Royal Society of London A* 298: 225–232.

Jonas DM and Fleming GR (1995) Vibrationally abrupt pulses in pump-probe spectroscopy. In: El-Sayed MA, Tanaka I and Molin Y (eds) *Ultrafast Processes in Chemistry and Photobiology*, pp. 225–256. Oxford, UK: Blackwell Scientific.

Klimov VI and McBranch DW (1998) Femtosecond high-sensitivity, chirp-free transient absorption spectroscopy using kilohertz lasers. *Optics Letters* 23: 277–279.

Lee SY (1995) Wave-packet model of dynamic dispersed and integrated pump-probe signals in femtosecond transition-state spectroscopy. In: Manz J and Wöste L (eds) *Femtosecond Chemistry*, vol. 1, pp. 273–298. Weinheim, Germany: VCH.

Lin SH, Alden R, Islampour R, Ma H and Villaeys AA (1991) *Density Matrix Method and Femtosecond Processes*. Singapore: World Scientific.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford University Press.

Norrish RGW and Porter G (1949) Chemical reactions produced by very high light intensities. *Nature* 164: 658.

Towrie M, Parker AW, Shaikh W and Matousek P (1998) Tunable picosecond optical parametric generator-amplifier system for time resolved Raman spectroscopy. *Measurement Science and Technology* 9: 816–823.

Yan YJ and Mukamel S (1990) Femtosecond pump-probe spectroscopy of polyatomic molecules in condensed phases. *Physical Reviews A* 41: 6485–6504.

Time-Correlated Single-Photon Counting

A Beeby, University of Durham, Durham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Time-correlated single-photon counting, TCSPC, is a well-established technique for the determination of fluorescence lifetimes and related time-resolved fluorescence properties. Since the 1980s, there have been dramatic changes in the laser sources used for this work, which in turn have opened up the technique to a wide user base.

TCSPC: The Basics

In the TCSPC experiment, the sample is repeatedly excited by a high-repetition rate, low-intensity light source. Fluorescence from the sample is collected and may be passed though a cut-off filter or monochromator to remove scattered excitation light and to select the emission wavelength. The fluorescence is then focused onto a detector, typically a photomultiplier (PMT) or single-photon avalanche diode (SPAD) which detects single emitted photons. The 'raw' signal from the detector fluctuates considerably from event to event. In order to remove any timing error that could be induced by this, the signal is processed first by a constant fraction discriminator (CFD) in order to provide a stable and consistent timing pulse. This signal is used as an input for the time-to-amplitude converter (TAC) which determines the time interval between the excitation of the sample and the emission of the photon. The output from the TAC is a relatively slow pulse whose intensity is proportional to the time interval between the start and stop signals. A pulse height analyzer (PHA) is used to process the output from the TAC and increments a channel in its memory corresponding to a time window during which the photon was detected. As the measurement is repeated many times over, a histogram is constructed in the memory of the PHA, showing the number of photons detected as a function of the time interval. This histogram corresponds to the fluorescence decay.

As outlined below, it is desirable to have a light source with a moderately high repetition rate, typically of the order of MHz. Because it is not practical to start the TAC each time the light source 'fires', it is more usual to operate the TAC in the so-called 'reversed mode', whereby the start signal is derived from the detected photon and the stop is derived from the light source. The electronic circuitry is illustrated in Figure 1. In practice, many researchers still use modular components that are wired together manually, although there is increasingly a tendency towards the use of single PC-based cards which contain all the necessary discriminators, timing circuits, and data acquisition electronics. There are a number of commercial PC cards that contain all the timing electronics, etc., required for TCSPC, for example, Becker and Hickl (Germany).

This approach to obtaining the fluorescence decay profile can only work if the detected photons are 'randomly selected' from the total emission from the sample. In practice, this means that the rate of detection of photons should be of the order of 1% of the excitation rate: that is, only one in a hundred excitation pulses gives rise to a detected photon. This histogram mirrors the fluorescence decay of the sample convolved with the instrument response function (IRF). The IRF is itself a convolution of the temporal response of the optical light pulse, the optical path in the collection optics, and the response of the detector and ancillary electronics. Experimentally, the IRF is usually recorded by placing a scattering material in the spectrometer.

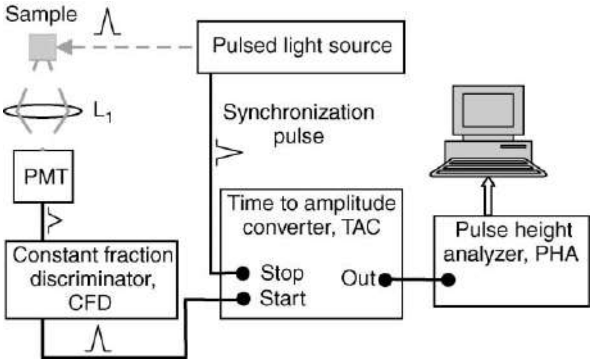
Sample
Pulsed light source
L_1
Synchronization pulse
PMT
Time to amplitude converter, TAC
Pulse height analyzer, PHA
Constant fraction discriminator, CFD
Stop
Start
Out

Figure 1 Schematic diagram of a TSCPC spectrometer. Fluorescence from the sample is collected by L_1, which may also include a monochromator/wavelength selection filter. A synchronization pulse is obtained either from the laser driver electronics directly, or via a photodiode and second constant fraction discriminator. The TAC is shown operating in a 'reversed mode' as is normal with a high repetition rate laser source.

In principle, the decay function of the sample can be extracted by deconvolution of the IRF from the decay function, although in practice it is more common to use the method of iterative re-convolution using a model decay function(s). Thus, the experimentally determined IRF is convolved with a model decay function and parameters within the function are systematically varied to obtain the best fit for a specific model. The most commonly used model assumes the decaying species to follow one or more exponential functions (sum of exponentials) or a distribution of exponentials. Further details regarding the mechanics of data analysis and fitting procedures can be found in the literature. A typical fit is illustrated in Figure 2.

The main aim of this review is to illustrate how recent advances in laser technology have changed the face of TCSPC and to discuss the types of laser light sources that are in use. The easiest way of approaching this topic is to define the criteria for an ideal light source for this experiment, and then to discuss the ways in which new technologies are meeting these criteria.

High Repetition Rate

A typical decay data set may contain something of the order of 0.1–5 million 'counts': it is desirable to

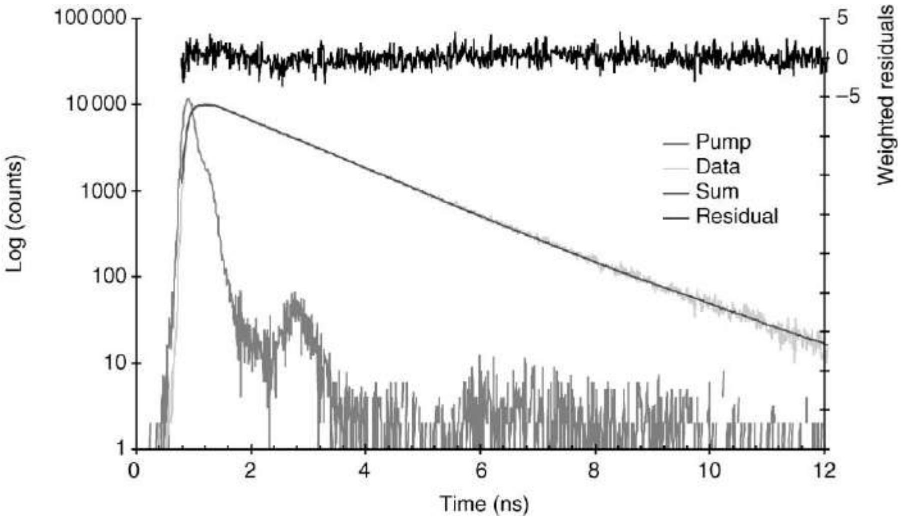

Figure 2 The fluorescence decay of Rhodamine B in water. The sample was excited using a 397 nm pulsed laser diode (1 MHz, < 200 ps FWHM), and the emission was observed at 575 nm. Deconvolution of the IRF (dark gray trace) and observed decay (light gray trace) with a single exponential function gave a lifetime of 1.50 ns, with a $\chi^2 R = 1.07$, Durbin–Watson parameter = 1.78. The fitted curve is shown overlaid in gray, and the weighted residuals are shown offset in black.

acquire such a number of events in order that we can reliably employ statistical parameters to judge the quality of the fitted decay function(s). As stated above, the rate of photon acquisition should be of the order of 1% of the excitation rate in order to obtain reliable, good-quality data. It is clear then that a high repetition rate source is desirable in order to reduce the overall data acquisition time. However, it is important that the repetition rate of the source is not too high, otherwise the data show 'overlap' effects, caused by the fluorescence intensity from one excitation pulse not completely decaying before the next laser pulse arrives at the sample. Under these conditions the observed fluorescence decay is perturbed and extraction of complex kinetic behavior is more difficult. As a general rule of thumb, it is recommended that the inter-pulse separation is greater than five times the lifetime of the longest component of the fluorescence decay. Fluorescence lifetimes can be as long as 100 ns, indicating a maximum repetition rate of 2 MHz, although when investigating fluorophores with a shorter lifetime correspondingly higher frequencies can be selected. Clearly the ability to vary the repetition rate according to the type of sample under investigation is an advantage.

Short Pulse Duration

Typically the minimum lifetime that can be determined by a TCSPC spectrometer is limited by the instrument response function: as a general rule it is possible to extract fluorescence lifetimes of about one tenth of the full width at half maximum (FWHM) of the IRF by reconvolution methods. As discussed above, the IRF is the product of a number of factors, including the optical pulse duration, differences in the optical path traversed by the fluorescence, the response time of the detector, and the temporal response of the associated timing electronics. Ideally the excitation source should be as short as possible, and with modern laser-based TCSPC systems, it is usually the case that the time-response of the detector is the limiting factor. The introduction of high-speed microchannel plate (MCP) photomultipliers have lead to very short IRFs, which can be <50 ps FWHM. When using these detectors is it essential to keep the optical transit-time spread as low as possible: that is, the path taken by light through a monochromator should be independent of wavelength and trajectory taken through the optics. Many research groups advocate the use of a subtractive dispersive double monochromator for this purpose.

Tunability

Early TCSPC experiments were carried out using hydrogen- or nitrogen-filled flashlamps which produced broadband ns-duration pulses with low repetition rates (<100 kHz). Variable wavelength excitation is possible using these sources, but their low intensity and low repetition rates mean that data acquisition periods can be very long. The early 1980s saw the introduction of cavity-dumped ion lasers and synchronously pumped mode-locked cavity-dumped dye lasers, which provided significantly higher repetition rates, typically up to 4 MHz, and greater intensities. However, the TCSPC experiments that could be carried out using these laser-based sources were somewhat limited by the wavelengths that could be generated by the argon ion and synchronously pumped dye lasers, and their second harmonics. The dye lasers were restricted to the use of Rhodamine 6G as the gain medium, which has a limited tuning range (~580–620 nm). It is often desirable to vary the excitation wavelength as part of a photophysical study, either as a systematic study of the effects of excitation energy upon the excited state behavior, or in order to optimize excitation of a particular chromophore in a system. To some extent in the past, the wavelengths available from the laser systems dictated the science that they could be used to address. The mid- to late-1980s saw the introduction of mode-locked continuous wave (CW) Nd:YAG lasers as pump lasers and increased number of dyes that could be synchronously pumped, providing a broader range of excitation wavelengths as well as increased stability. One drawback of these complex and often fickle laser systems was the level of expertise and attention they required to keep their performance at its peak.

Since the 1990s, the introduction of two important new types of laser have had a significant impact on the use of TCSPC as an investigative method. The first is based upon the mode-locked titanium sapphire laser which offers tuneable radiation from the deep UV to the NIR, whilst the second series of laser are based upon (relatively) low-cost, turn-key solid state diode lasers. The developments and merits of these new laser sources, which now dominate the market for TCSPC sources, are discussed below.

Ti-Sapphire

The discovery of the titanium doped sapphire laser has led to a revolution in ultrafast laser sources. Coupled with efficient diode-pumped solid state

Nd:YAG pump lasers, this medium currently dominates the entire ultrafast laser market, from high repetition, low pulse energy systems through to systems generating very high pulse energies at kHz repetition rates.

Ti-sapphire is, in many respects, an ideal laser gain medium: it has a high energy storage capacity, it has a very broad-gain bandwidth, and it has good thermal properties. Following the first demonstration of laser action in a Ti-sapphire crystal, the CW Ti-sapphire laser was quickly commercialized. It proved to be a remarkable new laser medium, providing greater efficiency, with outputs of the order of 1 W. The laser also provided an unprecedented broad tuning range, spanning from just below 700 nm to 1000 nm. The most important breakthrough for time-resolved users came in 1991, when self-mode-locking in the medium was demonstrated. This self-mode-locking, sometimes referred to as Kerr-lens mode-locking, arises from the nonlinear refractive index of the material and is implemented by simply placing an aperture within the cavity at an appropriate point. The broad-gain bandwidth of the medium results in very short pulses: pulse durations of the order of tens of femtoseconds can be routinely achieved with commercial laser systems.

The mode-locked Ti-sapphire laser was quickly seized upon and commercialized. The first generation of systems used multiwatt CW argon ion lasers to pump them, making the systems unwieldy and expensive to run, but in recent years these have been replaced by diode-pumped Nd:YAG lasers, providing an all solid-state, ultrafast laser source. Equally significantly, the current generation of lasers are very efficient and do not require significant electrical power or cooling capacity. There are several commercial suppliers of pulsed Ti-sapphire lasers, including Coherent and Spectra Physics.

A typical mode-locked Ti-sapphire laser provides an 80 MHz train of pulses with a FWHM of ~100 fs, with an average power of up to 1.5 W and are tuneable from 700 nm to almost 1000 nm. The relatively high peak powers associated with these pulses means that second- or even third-harmonic generation is relatively efficient, providing radiation in the ranges 350–480 nm and 240–320 nm. The output powers of the second and third harmonics far exceeds that required for TCSPC experiments, making them an attractive source for the study of fluorescence lifetime measurements. The simple mode-locked Ti-sapphire laser has some drawbacks: the very short pulse duration provides a broad-bandwidth excitation, the repetition rate is often too high with an inter-pulse separation of ~12.5 ns, and the laser does not provide continuous tuning from the deep-UV though to the visible. Incrementally, all of these points have been addressed by the laser manufacturers.

As well as operating as a source of femtosecond pulses, with their associated bandwidths of many tens of nm, some modern commercial Ti-sapphire lasers can, by adjustment of the optical path and components, be operated in a mode that provides picosecond pulses whilst retaining their broad tuning curve. Under these conditions, the bandwidth of the output is greatly reduced, making the laser resemble a synchronously pumped mode-locked dye laser.

As stated above, a very high repetition rate can be undesirable if the system under study has a long fluorescence lifetime due to the fluorescence decay from an earlier pulse not decaying completely before the next excitation pulse. Variation of the repetition of output frequency can be addressed by two means. The first is to use a pulse-picker. This is an opto-accoustic device which is placed outside the laser cavity. An electrical signal is applied to a device, synchronized to the frequency of the mode-locked laser to provide pulses at frequencies from ~4 MHz down to single-shot. The operation of a pulse picker is somewhat inefficient, with less than half of the original pulse energy being obtained in the output, although output power is not usually an issue when using these sources. More importantly, when using pulse-pickers, it is essential to ensure that there is no pre- or post-pulse which can lead to artifacts in the observed fluorescence decays. The second means of reducing the pulse duration is cavity dumping. This is achieved by extending the laser cavity and placing an opto-acoustic device, known as a Bragg cell, at a beam-waist within the cavity. When a signal is applied to the intracavity Bragg cell, synchronized with the passage of a mode-locked pulse though the device, a pulse of light is extracted from the cavity. The great advantage of cavity-dumping is that it builds up energy within the optical cavity and when pulses are extracted these are of a higher pulse energy than the mode-locked laser alone. This in turn leads to more efficient harmonic generation. This addition of the cavity dumper system extends the laser cavity and hence reduces the overall frequency of the mode-locked laser to ~50 MHz, and can provide pulses at 9 MHz down to single shot.

The extension of the tuning range of the Ti-sapphire laser has been facilitated by the introduction of the optical parametric oscillator (OPO). In the OPO, a short wavelength pump beam passes through a nonlinear optical crystal and generates two tuneable

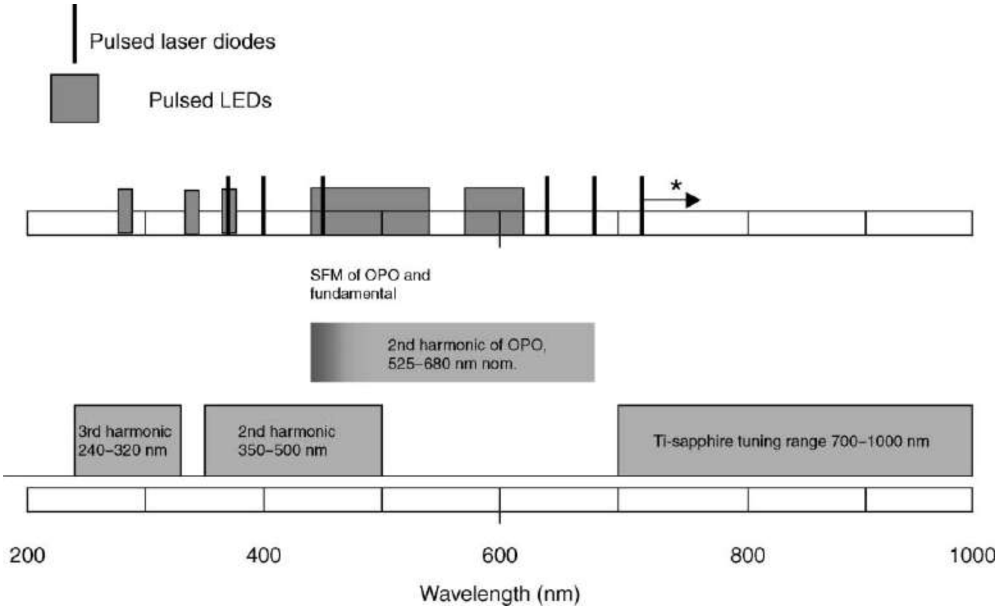

Figure 3 An illustration of the spectral coverage provided by currently available pulsed semiconductor lasers and LEDs, and a modern Ti-sapphire laser and OPO. Note that the OPO's output is frequency doubled, and the wavelengths in the range 525– ~ 440 nm may be obtained by sum frequency mixing of the Ti-sapphire's fundamental output with the output of the OPO. * denotes that there are many long wavelength pulsed laser diodes falling in the range $\lambda > 720$ nm.

beams of longer wavelengths. Synchronous pumping of the OPOs with either a fs or ps Ti-sapphire laser ensures reasonably high efficiencies. The outputs from the OPO may be frequency doubled, with the option of intracavity doubling enhancing the efficiency still further. Using an 800 nm pump wavelength, an OPO provides typical signal and idler tuning ranges of 1050–>1350 nm and 2.1–3.0 μm, respectively. The signal can be doubled to give 525–680 nm, effectively filling the visible spectrum. The average powers obtainable from Ti-sapphire pumped OPOs makes them eminently suitable for TCSPC experiments. A summary of the spectral coverage of the Ti-sapphire laser and OPO, and low-cost solid-state sources are illustrated in Figure 3.

An additional advantage of the Ti-sapphire laser is the possibility of multiphoton excitation. The combination of high average power and very short pulse duration of a mode-locked Ti-sapphire laser means that the peak-powers of the pulses can be very high. Focusing the near-IR output of the laser gives rise to exceedingly high photon densities within a sample which can lead to the simultaneous absorption of two or more photons. This is of particular significance when recording fluorescence lifetimes in conjunction with a microscope, particularly when the microscope is operated in a confocal configuration. Such systems are now being used for a derivative of TCSPC, known as fluorescence lifetime imaging (FLM).

In summary, the Ti-sapphire laser is a very versatile and flexible source that is well suited to the TCSPC experiment. Off-the-shelf systems provide broad spectral coverage and variable repetition rates. Undoubtedly this medium will be the mainstay of TCSPC for some time to come.

Diode Lasers

Low-power diode lasers, with outputs in the range 630–1000 nm, have been commercially available for some time. These devices, which are normally operated as CW sources, can be operated in a pulsed mode, providing a high frequency train of sub-ns pulses. However, the relatively long output wavelengths produced by this first generation of diode lasers severely restricted the range of chromophores that could be excited and this proved to be a limiting factor for their use as TCSPC sources. Attempts to generate shorter wavelengths from these NIR laser diodes by second harmonic generation (SHG), were never realized: the SHG process is extremely inefficient due to the very low peak output power of these lasers. Shorter wavelengths could be obtained by pulsing conventional light-emitting diodes (LEDs), which were commercially available with emission wavelengths covering the visible spectrum down to ~450 nm. However, these had the disadvantage of exhibiting longer pulse durations than those achievable from laser diodes and have much broader emission bandwidths than a laser source.

The introduction of violet and blue diode lasers and UV-LEDs by the Nichia Corporation (Japan) in the late 1990s resulted in a renewed interest in sources utilizing these small, solid-state devices. At least three manufacturers produce ranges of commercial TCSPC light sources based upon pulsed LEDs and laser diodes and can now boast a number of pulsed laser sources in their ranges including 375, 400, 450, 473, 635, 670, and 830 nm. Currently commercial pulsed laser diodes and LEDs are manufactured by Jobin-Yvon IBH (UK), Picoquant (Germany), and Hamamatso (Japan). These small, relatively low-cost devices offer pulse durations of <100 ps and excitation repetition rates of up to tens of MHz making them ideally suited to 'routine' lifetime measurements. Although these diode laser sources are not tunable and do not span the full spectrum, they are complemented by pulsed LEDs which fill the gap in the visible spectrum. Pulsed LEDs generally have a longer pulse duration and broader, lower intensity spectral output, but are useful for some applications. At the time of writing there are no diode laser sources providing wavelengths of <350 nm, although Jobin-Yvon IBH (UK) have recently announced pulsed LEDs operating at 280 nm and 340 nm. As LEDs, and possibly diode lasers, emitting at these shorter wavelengths become more widely available, the prospects of low running costs and ease of use makes these semiconductor devices an extremely attractive option for many TCSPC experiments. Commercial TCSPC spectrometers based around these systems offer a comparatively low-cost, turnkey approach to fluorescence lifetime based experiments, opening up what was a highly specialized field to a much broader range of users.

See also

Lasers: Noble Gas Ion Lasers. **Light Emitting Diodes**. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

Further Reading

Lakowicz JR (ed.) (2001) *Principles of Fluorescence Spectroscopy*. Plenum Press.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*.

Spence DE, Kean PN and Sibbett W (1991) *Optics Letters* 16: 42–44.

Transient Holographic Grating Techniques in Chemical Dynamics

E Vauthey, University of Geneva, Geneva, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Over the past two decades, holographic techniques have proved to be valuable tools for investigating the dynamics of chemical processes. The aim of this chapter is to give an overview of the main applications of these techniques for chemical dynamics. The basic principle underlying the formation and the detection of elementary transient holograms, also called transient gratings, is first presented. This is followed by a brief description of a typical experimental setup. The main applications of these techniques to solve chemical problems are then discussed.

Basic Principle

The basic principle of the transient holographic technique is illustrated in Figure 1. The sample material is excited by two laser pulses at the same wavelength and crossed at an angle $2\theta_{pu}$. If the two pump pulses have the same intensity, I_{pu}, the intensity distribution at the interference region, assuming plane waves, is

$$I(x) = 2I_{pu}\left[1 + \cos\left(\frac{2\pi x}{\Lambda}\right)\right] \qquad [1]$$

where $\Lambda = \lambda_{pu}/(2\sin\theta_{pu})$ is the fringe spacing, λ_{pu} is the pump wavelength, and I_{pu} is the intensity of one pump pulse.

As discussed below, there can be many types of light–matter interactions that lead to a change in the optical properties of the material. For a dielectric material, they result in a spatial modulation of the optical susceptibility and thus of the complex refractive index, $\tilde{n}$. The latter distribution can be described as a Fourier cosine series:

$$\tilde{n}(x) = \tilde{n}_0 + \sum_{m=1}^{\infty} \tilde{n}_m \cos\left(\frac{m2\pi x}{\Lambda}\right) \qquad [2]$$

where $\tilde{n}_0$ is the average value of $\tilde{n}$. In the absence of saturation effects, the spatial modulation of $\tilde{n}$ is harmonic and the Fourier coefficients with $m > 1$ vanish. In this case, the peak to null variation of the complex refractive index, $\Delta\tilde{n}$, is equal to the Fourier coefficient $\tilde{n}_1$. The complex refractive index can be split into its real and imaginary components:

$$\tilde{n} = n + iK \qquad [3]$$

where n is the refractive index and K is the attenuation constant.

The hologram created by the interaction of the crossed pump pulses consists in periodic, one-dimensional spatial modulations of n and K. Such distributions are nothing but phase and amplitude gratings, respectively. A third laser beam at the probe wavelength, λ_{pr}, striking these gratings at Bragg angle, $\theta_B = \arcsin(\lambda_{pr}/2\Lambda)$, will thus be partially diffracted (Figure 1b). The diffraction efficiency, η, depends on the modulation amplitude of the optical properties. In the limit of small diffraction efficiency ($\eta < 0.01$), this relationship is given by

$$\eta = \frac{I_{dif}}{I_{pr}} \cong \left[\left(\frac{\ln 10\Delta A}{4\cos\theta_B}\right)^2 + \left(\frac{\pi d\Delta n}{\lambda_{pr}\cos\theta_B}\right)^2\right] \times \exp\left(-\frac{\ln 10A}{\cos\theta_B}\right) \qquad [4]$$

where I_{dif} and I_{pr} are the diffracted and the probe intensity respectively, d is the sample thickness, and $A = 4\pi dK/(\lambda \ln 10)$ is the average absorbance.

The first and the second terms in the square bracket describe the contribution of the amplitude and phase gratings respectively, and the exponential term accounts for the reabsorption of the diffracted beam by the sample.

The main processes responsible for the variation of the optical properties of an isotropic dielectric material are summarized in Figure 2.

The modulation of the absorbance, ΔA, is essentially due to the photoinduced concentration change,

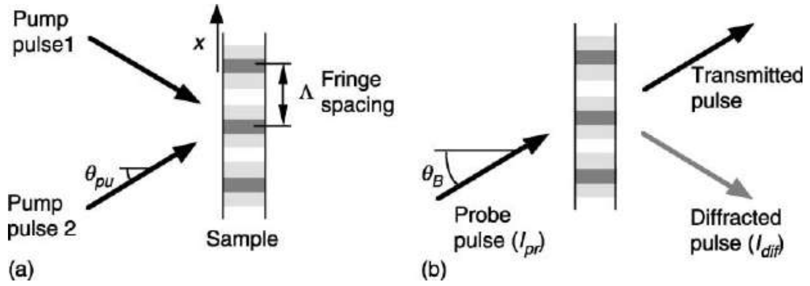
Pump pulse1
x
Λ
Fringe spacing
θ_pu
Pump pulse 2
Sample
(a)
Transmitted pulse
θ_B
Probe pulse (I_pr)
Diffracted pulse (I_dif)
(b)

Figure 1 Principle of the transient grating technique: (a) grating formation (pumping), (b) grating detection (probing).

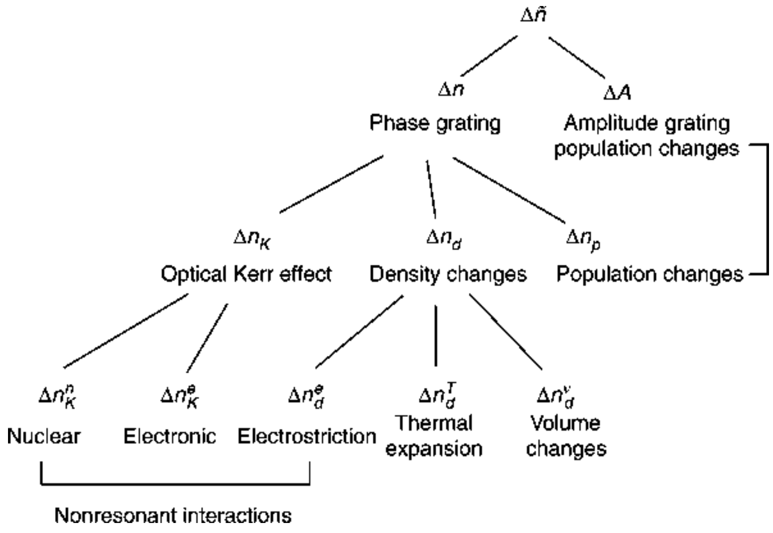
Δñ
Δn
Phase grating
ΔA
Amplitude grating
population changes
Δn_K
Optical Kerr effect
Δn_d
Density changes
Δn_p
Population changes
Δn_K^n
Nuclear
Δn_K^e
Electronic
Δn_d^e
Electrostriction
Δn_d^T
Thermal expansion
Δn_d^v
Volume changes
Nonresonant interactions

Figure 2 Classification of the possible contributions to a transient grating signal.

ΔC, of the different chemical species i (excited state, photoproduct, ...):

$$\Delta A(\lambda_{pr}) = \sum_i \varepsilon_i(\lambda_{pr})\Delta C_i \qquad [5]$$

where ε_i is the absorption coefficient of the species i.

The variation of the refractive index, Δn, has several origins and can be expressed as

$$\Delta n = \Delta n_K + \Delta n_p + \Delta n_d \qquad [6]$$

Δn_K is the variation of refractive index due to the optical Kerr effect (OKE). This nonresonant interaction results in an electronic polarization (electronic OKE) and/or in a nuclear reorientation of the molecules (nuclear OKE) along the direction of the electric field associated with the pump pulses. As a consequence, a transient birefringence is created in the material. This effect is usually discussed within the framework of nonlinear optics in terms of intensity dependent refractive index or third order nonlinear susceptibility. Electronic OKE occurs in any dielectric material under sufficiently high light intensity. On the other hand, nuclear OKE is mostly observed in liquids and gases and depends strongly on the molecular shape.

Δn_p is the change of refractive index related to population changes. Its magnitude and wavelength dependence can be obtained by Kramers–Kronig transformation of $\Delta A(\lambda)$ or $\Delta K(\lambda)$:

$$\Delta n_p(\lambda) = \frac{1}{2\pi^2}\int_0^\infty \frac{\Delta K(\lambda')}{1-(\lambda'/\lambda)^2}\,\mathrm{d}\lambda' \qquad [7]$$

Δn_d is the change of refractive index associated with density changes. Density phase gratings can have essentially three origins:

$$\Delta n_d = \Delta n_d^t + \Delta n_d^v + \Delta n_d^e \qquad [8]$$

Δn_d^t is related to the temperature-induced change of density. If a fraction of the excitation energy is converted into heat, through a nonradiative transition or an exothermic process, the temperature becomes spatially modulated. This results in a variation of density, hence to a modulation of refractive index with amplitude Δn_d^t. Most of the temperature dependence of n originates from the density. The temperature-induced variation of n at constant density is much smaller than Δn_d^t.

Δn_d^v is related to the variation of volume upon population changes. This volume comprises not only the reactant and product molecules but also their environment. For example, in the case of a photodissociation, the volume of the product is larger than that of the reactant and a positive volume change can be expected. This will lead to a decrease of the density and to a negative Δn_d^v.

Finally, Δn_d^e is related to electrostriction in the sample by the electric field of the pump pulses. Like OKE, this is a nonresonant process that also contributes to the intensity dependent refractive index. Electrostriction leads to material compression in the regions of high electric field strength. The periodic compression is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm(2\pi/\Lambda)\vec{i}$, where $\vec{i}$ is the unit vector along the modulation axis. The interference of these acoustic waves leads to a temporal modulation of Δn_d^e at the acoustic frequency v_{ac}, with $2\pi v_{ac} = k_{ac}v_s$, v_s being the speed of sound. As Δn_d^e oscillates between negative and positive values, the diffracted intensity, which is proportional to $(\Delta n_d^e)^2$, shows at temporal oscillation at twice the acoustic frequency. In most cases, Δn_d^e is weak and can be neglected if the pump pulses are within an absorption band of the sample.

The modulation amplitudes of absorbance and refractive index are not constant in time and their temporal behavior depends on various dynamic processes in the sample. The whole point of the transient grating techniques is precisely the measurement of the diffracted intensity as a function of time after excitation to deduce dynamic information on the system.

In the following, we will show that the various processes shown in Figure 2, that give rise to a diffracted signal, can in principle be separated by choosing appropriate experimental parameters, such as the timescale, the probe wavelength, the polarization of the four beams, or the crossing angle.

Experimental Setup

A typical experimental arrangement for pump–probe transient grating measurements is shown in Figure 3. The laser output pulses are split in three parts. Two parts of equal intensity are used as

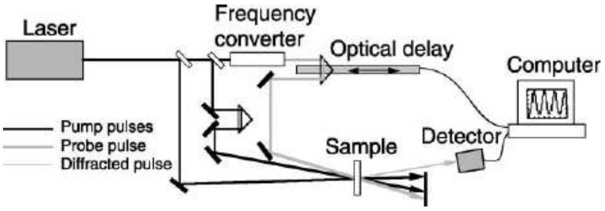

Figure 3 Schematic of a transient grating setup with pump–probe detection.

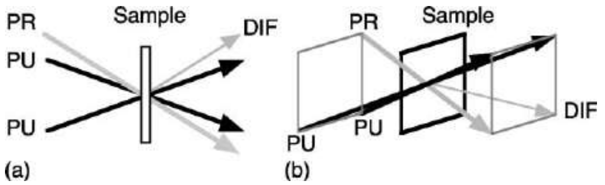

Figure 4 Beam geometry for transient grating: (a) in plane; (b) boxcars.

pump pulses and are crossed in the sample. In order to ensure time coincidence, one pump pulse travels along an adjustable optical delay line. The third part, which is used for probing, can be frequency converted using a dye laser, a nonlinear crystal, a Raman shifter, or white light continuum generation. The probe pulse is sent along a motorized optical delay line before striking the sample at the Bragg angle. There are several possible beam configurations for transient grating and the two most used are illustrated in Figure 4. When the probe and pump pulses are at different wavelengths, they can be in the same plane of incidence as shown in Figure 4a. However, if the pump and probe wavelengths are the same, the folded boxcars geometry shown in Figure 4b has to be used. The transient grating technique is background free and the diffracted signal propagates in a well-defined direction. In a pump–probe experiment, the diffracted signal intensity is measured as a function of the time delay between the pump and probe pulses. Such a setup can be used to probe dynamic processes occurring in timescales going from a few fs to a few ns, the time resolution depending essentially on the duration of the pump and probe pulses. For slower processes, the grating dynamics can be probed in real time with a *cw* laser beam and a fast photodetector.

Applications

The Transient Density Phase Grating Technique

If the grating is probed at a wavelength far from any absorption band, the variation of absorbance, $\Delta A(\lambda_{pr})$, is zero and the corresponding change of refractive index, $\Delta n_p(\lambda_{pr})$, is negligibly small. In this case, eqn [4] simplifies to

$$\eta \cong \left(\frac{\pi d}{\lambda_{pr}\cos\theta_B}\right)^2 \times \Delta n_d^2 \qquad [9]$$

In principle, the diffracted signal may also contain contributions from the optical Kerr effect, Δn_K, but we will assume here that this non-resonant and ultrafast response is negligibly small. The density change can originate from both heat releasing processes and volume differences between the products and reactants, the former contribution usually being much larger than the latter. Even if the heat releasing process is instantaneous, the risetime of the density grating is limited by thermal expansion. This expansion is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm(2\pi/\Lambda)\vec{i}$. One can distinguish two density gratings: 1) a diffusive density grating, which reproduces the spatial distribution of temperature and which decays by thermal diffusion; and 2) an acoustic density grating originating from the standing acoustic wave and whose amplitude oscillates at the acoustic frequency v_{ac}.

The amplitudes of these two gratings are equal but of opposite sign. Consequently, the time dependence of the modulation amplitude of the density phase grating is given by

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V\right)\rho\left(\frac{\partial n}{\partial \rho}\right)R(t) \qquad [10a]$$

with

$$R(t) = 1 - \cos(2\pi v_{ac} t) \times \exp(-\alpha_{ac} v_s t) \qquad [10b]$$

where ρ, β, C_v, and α_{ac} are the density, the volume expansion coefficient, the heat capacity, and the acoustic attenuation constant of the medium, respectively, Q is the amount of heat deposited during the photoinduced process, and ΔV is the corresponding volume change. As the standing acoustic wave oscillates, its corresponding density grating interferes with the diffusive density grating. Therefore, the total modulation amplitude of the density and thus Δn_d exhibits the oscillation at v_{ac}. Figure 5 shows the time profile of the diffracted

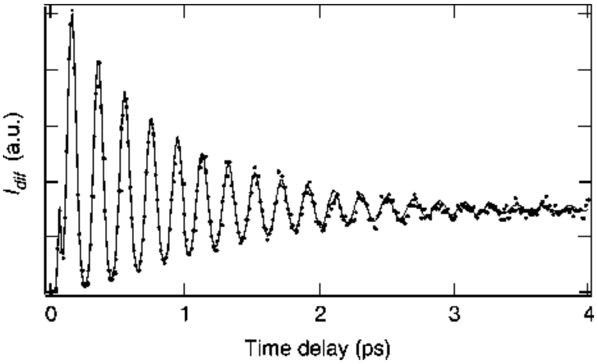

Figure 5 Time profile of the diffracted intensity measured with a solution of malachite green after excitation with two pulses with close to counterpropagating geometry.

intensity measured with a solution of malachite green. After excitation to the S_1 state, this dye relaxes nonradiatively to the ground state in a few ps. For this measurement, the sample solution was excited with two 30 ps laser pulses at 532 nm crossed with an angle close to 180°. The continuous line is the best fit of eqns [9] and [10]. The damping of the oscillation is due to acoustic attenuation. After complete damping, the remaining diffracted signal is due to the diffusive density grating only.

$R(t)$ can be considered as the response function of the sample to a prompt heat release and/or volume change. If these processes are not instantaneous compared to an acoustic period ($\tau_{ac} = \nu_{ac}^{-1}$), the acoustic waves are not created impulsively and the time dependence of Δn_d is

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V\right)\rho\left(\frac{\partial n}{\partial \rho}\right)F(t) \qquad [11a]$$

with

$$F(t) = \int_{-\infty}^{t} R(t - t')\cdot f(t')\mathrm{d}t' \qquad [11b]$$

where $f(t)$ is a normalized function describing the time evolution of the temperature and/or volume change. In many cases, $f(t) = \exp(-k_r t)$, k_r being the rate constant of the process responsible for the change. Figure 6 shows the time profile of the diffracted intensity calculated with eqns [9] and [11] for different values of k_r.

If several processes take place, the total change of refractive index is the sum of the changes due to the individual processes. In this case, Δn_d should be expressed as

$$\Delta n_d(t) = \sum_i \left(\frac{\beta Q_i}{\rho C_v} + \Delta V_i\right)\rho\left(\frac{\partial n}{\partial \rho}\right)F_i(t) \qquad [12]$$

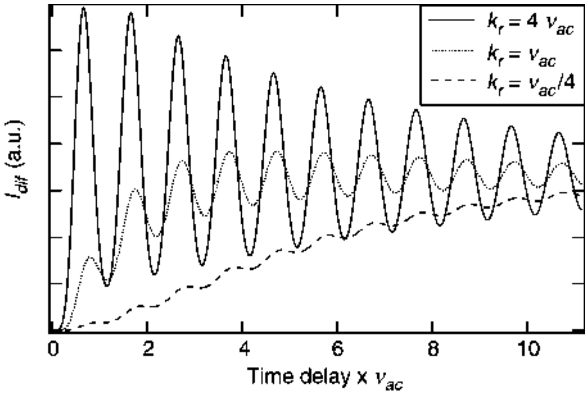

Figure 6 Time profiles of the diffracted intensity calculated using eqns [9] and [11] with various k_r values.

The separation of the thermal and volume contributions to the diffracted signal is problematic. Several approaches have been proposed, the most used being the measurement in a series of solvents with different expansion coefficient β. This method requires that the other solvent properties like refractive index, dielectric constant, or viscosity, are constant or that the energetics of the system investigated does not depend on them. This separation is easier when working with water because its β value vanishes at 4 °C. At this temperature, the density variations are due to the volume changes only.

The above equations describe the growth of the density phase grating. However, this grating is not permanent and decays through diffusive processes. The phase grating originating from thermal expansion decays via thermal diffusion with a rate constant k_{th} given by

$$k_{th} = D_{th}\left(\frac{2\pi}{\Lambda}\right)^2 \qquad [13]$$

where D_{th} is the thermal diffusivity. Table 1 shows k_{th} values in acetonitrile for different crossing angles.

The decay of the phase grating originating from volume changes depends on the dynamics of the population responsible for ΔV (vide infra).

Time resolved optical calorimetry

A major application of the density phase grating in chemistry is the investigation of the energetics of photo-induced processes. The great advantage of this technique over other optical calorimetric methods, like the thermal lens or the photoacoustic spectroscopy, is its superior time resolution. The time constant of the fastest heat releasing process that can be time resolved with this technique is of the order of the acoustic period. The shortest acoustic period is achieved when forming the grating with two counterpropagating pump pulses ($2\theta_{pu} = 180°$). In this case the fringe spacing is

Table 1 Fringe spacing, Λ, acoustic frequency, ν_{ac}, thermal diffusion rate constant, k_{tr}, for various crossing angles of the pump pulses, 2 θ_{pu}, at 355 nm and in acetonitrile

$2\theta_{pu}$	Λ (μm)	ν_{ac}(s^{-1})	k_{th} (s^{-1})
0.5°	40.7	3.2×10^7	4.7×10^3
50°	0.42	3.1×10^9	4.5×10^7
180°	0.13	9.7×10^9	4.7×10^8

$\Lambda = \lambda_{pu}/2n$ and, with UV pump pulses, an acoustic period of the order of 100 ps can be obtained in a typical organic solvent.

For example, Figure 7a shows the energy diagram for a photoinduced electron transfer (ET) reaction between benzophenone (BP) and an electron donor (D) in a polar solvent. Upon excitation at 355 nm, $^1\mathrm{BP}^*$ undergoes intersystem-crossing (ISC) to $^3\mathrm{BP}^*$ with a time constant of about 10 ps. After diffusional encounter with the electron donor, ET takes place and a pair of ions is generated. With this system, the whole energy of the 355 nm photon ($E = 3.49$ eV) is converted into heat. The different heat releasing processes can be differentiated according to their timescale. The vibrational relaxation to $^1\mathrm{BP}^*$ and the ensuing ISC to $^3\mathrm{BP}^*$ induces an ultrafast release of 0.49 eV as heat. With donor concentrations of the order of 0.1 M, the heat deposition process due to the electron transfer is typically in the ns range. Finally, the recombination of the ions produces a heat release in the microsecond timescale. Figure 7b shows the time profile of the diffracted intensity measured after excitation at 355 nm of BP with 0.05 M D in acetonitrile. The 30 ps pump pulses were crossed at 27° ($\tau_{ac} = 565$ ps) and the 10 ps probe pulses were at 590 nm. The oscillatory behavior is due to the ultrafast heat released upon formation of $^3\mathrm{BP}^*$, while the slow dc rise is caused by the heat dissipated upon ET. As the amount of energy released in the ultrafast process is known, the energy released upon ET can be determined by comparing the amplitudes of the fast and slow components of the time profile. The energetics as well as the dynamics of the photoinduced ET process can thus be determined. Figure 7c shows the decay of the diffracted intensity due to the washing out of the grating by thermal diffusion. As both the acoustic frequency and the thermal diffusion depends on the grating wavevector, the experimental time window in which a heat releasing process can be measured, depends mainly on the crossing angle of the pump pulses as shown in Table 1.

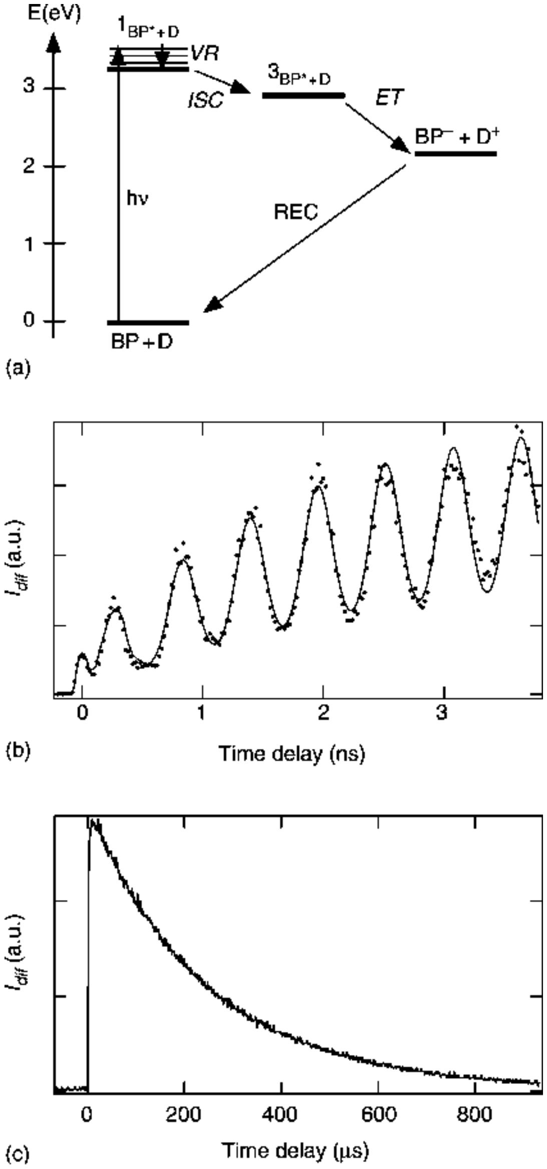

Figure 7 (a) Energy diagram of the states involved in the photoinduced electron transfer reaction between benzophenone (BP) and an electron donor (D) (VR: vibrational relaxation, ISC: intersystem crossing; ET: electron transfer; REC: charge recombination). (b) Time profile of the diffracted intensity measured at 590 nm after excitation at 355 nm of a solution of BP and 0.05 M D. (c) Same as (b) but measured at 1064 nm with a *cw* laser.

Determination of material properties

Another important application of the transient density phase grating technique is the investigation of material properties. Acoustics waves of various frequencies, depending on the pump wavelength and crossing angle, can be generated without physical contact with the sample. Therefore, acoustic properties, such as the speed of sound and the acoustic attenuation of the material, can be easily obtained from the frequency and the damping of the oscillation of the transient grating signal.

Similarly, the optoelastic constant of a material, $\rho \partial n / \partial \rho$, can be determined from the amplitude of the signal (see eqn [11]). This is done by comparing the signal amplitude of the material under investigation with that obtained with a known standard.

Finally, the thermal diffusivity, D_{th}, can be easily obtained from the decay of the thermal density phase grating, such as that shown in Figure 7c. This technique can be used with a large variety of bulk materials as well as films, surfaces and interfaces.

Investigation of Population Dynamics

The dynamic properties of a photogenerated species can be investigated by using a probe wavelength within its absorption or dispersion spectrum. As shown by eqn [4], either ΔA or Δn_p has to be different from zero. For this application, Δn_K and Δn_d should ideally be equal to zero. Unless working with ultrashort pulses (< 1 ps) and a weakly absorbing sample, Δn_K can be neglected. Practically, it is almost impossible to find a sample system for which some fraction of the energy absorbed as light is not released as heat. However, by using a sufficiently small crossing angle of the pump pulses, the formation of the density phase grating, which depends on the acoustic period, can take as much as 30 to 40 ns. In this case, Δn_d is negligible during the first few ns after excitation and the diffracted intensity is due to the population grating only:

$$I_{dif}(t) \propto \sum_i \Delta C_i^2(t) \qquad [14]$$

where ΔC_i is the modulation amplitude of the concentration of every species i whose absorption and/or dispersion spectrum overlaps with the probe wavelength.

The temporal variation of ΔC_i can be due either to the dynamics of the species i in the illuminated grating fringes or to processes taking place between the fringes, such as translational diffusion, excitation transport, and charge diffusion. In liquids, the decay of the population grating by translational diffusion is slow and occurs in the microsecond to millisecond timescale, depending on the fringe spacing. As thermal diffusion is typically hundred times faster, eqn [14] is again valid in this long timescale. Therefore if the population dynamics is very slow, the translational diffusion coefficient of a chemical species can be obtained by measuring the decay of the diffracted intensity as a function of the fringe spacing. This procedure has also been used to determine the temperature of flames. In this case however, the decay of the population grating by translational diffusion occurs typically in the sub-ns timescale.

In the condensed phase, these interfringe processes are of minor importance when measuring the diffracted intensity in the short timescale, i.e., before the formation of the density phase grating. In this case, the transient grating technique is similar to transient absorption, and it thus allows the measurement of population dynamics. However, because holographic detection is background free, it is at least a hundred times more sensitive than transient absorption.

The population gratings are usually probed with monochromatic laser pulses. This procedure is well suited for simple photoinduced processes, such as the decay of an excited state population to the ground state. For example, Figure 8 shows the decay of the diffracted intensity at 532 nm measured after excitation of a cyanine dye at the same wavelength. These dynamics correspond to the ground state recovery of the dye by non-radiative deactivation of the first singlet excited state. Because the diffracted intensity is proportional to the square of the concentration changes (see eqn [14]), its decay time is twice as short as the ground state recovery time. If the reaction is more complex, or if several intermediates are involved, the population grating has to be probed at several wavelengths. Instead of repeating many single wavelength experiments, it is preferable to perform multiplex transient grating. In this case, the grating is probed by white light pulses generated by focusing high intensity fs or ps laser pulses in a dispersive medium. If the crossing angle of the pump pulses is small enough ($< 1°$), the Bragg angle for probing is almost independent on the wavelength. A transient grating spectrum is obtained by dispersing the diffracted signal in a spectrograph. This spectrum consists in the sum of the square of the transient absorption and transient dispersion spectra. Practically, it is very similar to a transient absorption spectrum, but with a much superior signal to noise ratio. Figure 9 shows the transient grating spectra measured at various time delays after excitation of a solution of chloranil (CA) and

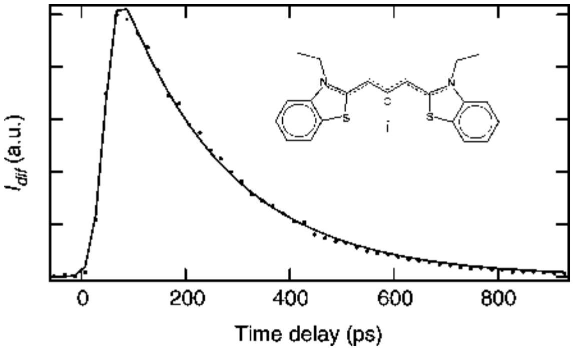

Figure 8 Time profile of the diffracted intensity at 532 nm measured after excitation at the same wavelength of a cyanine dye (inset) in solution and best fit assuming exponential ground state recovery.

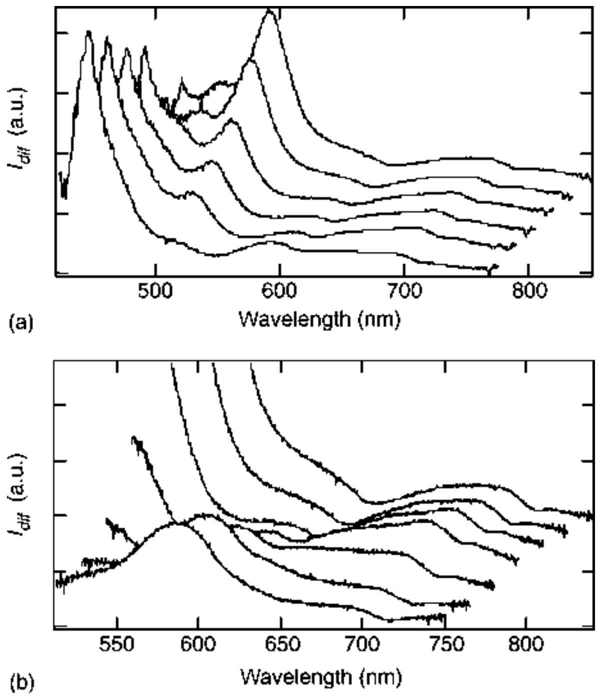

Figure 9 Transient grating spectra obtained at various time delays after excitation at 355 nm of a solution of chloranil and 0.25 M methylnaphthalene (a): from top to bottom: 60, 100, 180, 260, 330 and 600 ps; (b): from top to bottom: 100, 400, 750, 1100, 1500 ps).

methylnaphthalene (MNA) in acetonitrile. The reaction that takes place is

$$^3\mathrm{CA}^* \cdot \mathrm{MNA} \xrightarrow{\text{diffusion}} (^3\mathrm{CA}^* \cdot \mathrm{MNA}) \xrightarrow{\text{ET}}$$
$$^3(\mathrm{CA}^- \cdot \mathrm{MNA}^+) \xrightarrow{\text{MNA}} {}^3(\mathrm{CA}^- \cdot \mathrm{MNA}^+)$$

After its formation upon ultrafast ISC, ^{3}CA*, which absorbs around 510 nm, decays upon ET with MNA to generate CA$^-$(450 nm) and MNA$^+$ (690 nm). The ion pair reacts further with a second MNA molecule to form the dimer cation (580 nm).

The time profile of the diffracted intensity reflects the population dynamics as long as these populations follow first- or pseudo-first order kinetics. Higher order kinetics leads to inharmonic gratings and in this case the time dependence of the diffracted intensity is no longer a direct measure of the population dynamics.

Polarization Selective Transient Grating

In the above applications, the selection between the different contributions to the diffracted signal was essentially made by choosing the probe wavelength and the crossing angle of the pump pulses. However, this approach is not always sufficient. Another important parameter is the polarization of the four waves involved in a transient grating experiment. For example, when measuring population dynamics, the polarization of the probe beam has to be at magic angle (54.7°) relatively to that of the pump pulses. This ensures that the observed time profiles are not distorted by the decay of the orientational anisotropy of the species created by the polarized pump pulses.

The magnitude of ΔA and Δn depends on the pump pulses intensity, and therefore the diffracted intensity can be expressed by using the formalism of nonlinear optics:

$$I_{dif}(t) = C\int_{-\infty}^{+\infty} \mathrm{d}t\, I_{pr}(t - t'') \times \left[\int_{-\infty}^{t} \mathrm{d}t' |R_{ijkl}(t'' - t')| I_{pu}(t')\right]^2 \qquad [15]$$

where C is a constant and R_{ijkl} is an element of the fourth rank tensor **R** describing the nonlinear response of the material to the applied optical fields. In isotropic media, this tensor has only 21 nonzero elements, which are related as follows:

$$R_{1111} = R_{1122} + R_{1212} + R_{1221} \qquad [16]$$

where the subscripts are the Cartesian coordinates. Going from right to left, they design the direction of polarization of the pump, probe, and signal pulses. The remaining elements can be obtained by permutation of these indices ($R_{1122} = R_{1133} = R_{2211} = \ldots$). In a conventional transient grating experiment, the two pump pulses are at the same frequency and are time coincident and therefore their indices can be interchanged. In this case, $R_{1212} = R_{1221}$ and the number of independent tensor elements is further reduced:

$$R_{1111} = R_{1122} + 2R_{1212} \qquad [17]$$

The tensor **R** can be decomposed into four tensors according to the origin of the sample response: population and density changes, electronic and nuclear optical Kerr effects:

$$\mathbf{R} = \mathbf{R}(p) + \mathbf{R}(d) + \mathbf{R}(K, e) + \mathbf{R}(K, n) \qquad [18]$$

As they describe different phenomena, these tensors do not have the same symmetry properties. Therefore, the various contributions to the diffracted intensity can, in some cases, be measured selectively by choosing the appropriate polarization of the four waves.

Table 2 shows the relative amplitude of the most important elements of these four tensors.

The tensor $\mathbf{R}(p)$ depends on the polarization anisotropy of the sample, r, created by excitation with polarized pump pulses:

$$r(t) = \frac{N_{\parallel}(t) - N_{\perp}(t)}{N_{\parallel}(t) + 2N_{\perp}(t)} = \frac{2}{5}\mathrm{P}_2[\cos(\gamma)]\exp(-k_{re}t) \qquad [19]$$

where $N_{\parallel}$ and $N_{\perp}$ are the number of molecules with the transition dipole oriented parallel and perpendicular to the polarization of the probe pulse, respectively, γ is the angle between the transition dipoles involved in the pump and probe processes, P_2 is the second Legendre polynomial, and k_{re} is the rate constant for the reorientation of the transition dipole, for example, by rotational diffusion of the molecule or by energy hopping.

The table shows that $R_{1212}(d) = 0$, i.e., the contribution of the density grating can be eliminated with the set of polarization $(0°, 90°, 0°, 90°)$, the so-called crossed grating geometry. In this geometry, the two pump pulses have orthogonal polarization, the polarization of the probe pulse is parallel to that of one pump pulse and the polarization component of the signal that is orthogonal to the polarization of the probe pulse is measured. In this geometry, $R_{1212}(p)$ is nonzero as long as there is some polarization anisotropy, ($r \neq 0$). In this case, the diffracted intensity is

$$I_{dif}(t) \propto |R_{1212}(p)|^2 \propto [\Delta C(t) \times r(t)]^2 \qquad [20]$$

The crossed grating technique can thus be used to investigate the reorientational dynamics of molecules, through $r(t)$, especially when the dynamics of r is faster than that of ΔC. For example, Figure 10 shows the time profile of the diffracted intensity measured in the crossed grating geometry with rhodamine 6G in ethanol. The decay is due to the reorientation of the molecule by rotational diffusion, the excited lifetime of rhodamine being about 4 ns.

On the other hand, if the decay of r is slower than that of ΔC, the crossed grating geometry can be used to measure the population dynamics without any interference from the density phase grating. For example, Figure 11 shows the time profile of the diffracted intensity after excitation of a suspension of $\mathrm{TiO_2}$ particles in water. The upper profile was measured with the set of polarization $(0°, 0°, 0°, 0°)$ and thus reflects the time dependence of $R_{1111}(p)$ and $R_{1111}(d)$. $R_{1111}(p)$ is due to the trapped electron population, which decays by charge recombination, and $R_{1111}(d)$ is due to the heat dissipated upon both charge separation and recombination. The lower time profile was measured in the crossed grating geometry and thus reflects the time dependence of R_{1212} and in particular that of $R_{1212}(p)$.

Each contribution to the signal can be selectively eliminated by using the set of geometry $(\zeta, 45°, 0°, 0°)$ where the value of the 'magic angle' ζ for each

Table 2 Relative value of the most important elements of the response tensors $\mathbf{R}(i)$ and polarization angle ζ of the signal beam, where the contribution of the corresponding process vanishes for the set of polarization $(\zeta, 45°, 0°, 0°)$.

Process	R_{1111}	R_{1122}	R_{1212}	ζ
Electronic OKE	1	1/3	1/3	−71.6°
Nuclear OKE	1	−1/2	3/4	63.4°
Density	1	1	0	−45°
Population:				
$\gamma = 0°$	1	1/3	1/3	−71.6°
$\gamma = 90°$	1	2	−1/2	−26.6°
No correlation: $r = 0$	1	1	0	−45°

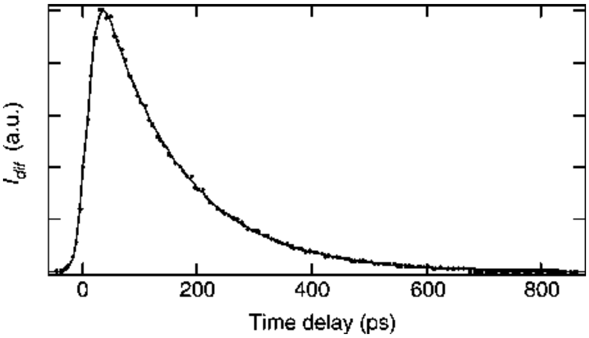

Figure 10 Time profile of the diffracted intensity measured with a solution of rhodamine 6G with crossed grating geometry.

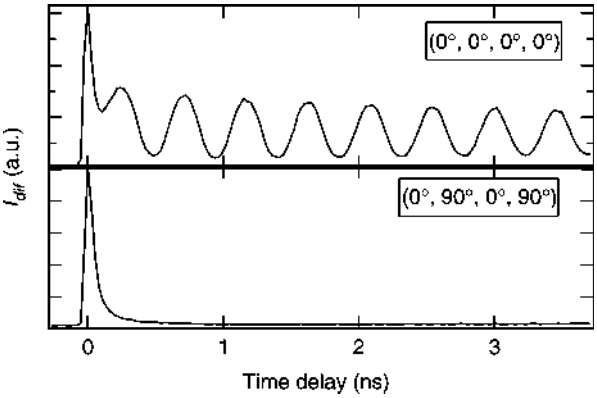

Figure 11 Time profiles of the diffracted intensity after excitation at 355 nm of a suspension of $\mathrm{TiO_2}$ particles in water and using different set of polarization of the four beams.

contribution is listed in Table 2. This approach allows for example the nuclear and electronic contributions to the optical Kerr effect to be measured separately.

Concluding Remarks

The transient grating techniques offer a large variety of applications for investigating the dynamics of chemical processes. We have only discussed the cases where the pump pulses are time coincident and at the same wavelength. Excitation with pump pulses at different wavelengths results to a moving grating. The well-known CARS spectroscopy is such a moving grating technique. Finally, the three pulse photon echo can be considered as a special case of transient grating where the sample is excited by two pump pulses, which are at the same wavelength but are not time coincident.

List of Units and Nomenclature

C	concentration [mol L^{-1}]
C_v	heat capacity [J K^{-1} kg^{-1}]
d	sample thickness [m]
D_{th}	thermal diffusivity [m^2 s^{-1}]
I_{dif}	diffracted intensity [W m^{-2}]
I_{pr}	intensity of the probe pulse [W m^{-2}]
I_{pu}	intensity of a pump pulse [W m^{-2}]
k_{ac}	acoustic wavevector [m^{-1}]
k_r	rate constant of a heat releasing process [s^{-1}]
k_{re}	rate constant of reorientation [s^{-1}]
k_{th}	rate constant of thermal diffusion [s^{-1}]
K	attenuation constant
n	refractive index
$\tilde{n}$	complex refractive index
N	number of molecule per unit volume [m^{-3}]
r	polarization anisotropy
$\mathbf{R}$	fourth rank response tensor [m^2 V^{-2}]
v_s	speed of sound [m s^{-1}]
V	volume [m^3]
α_{ac}	acoustic attenuation constant [m^{-1}]
β	cubic volume expansion coefficient [K^{-1}]
γ	angle between transition dipoles
Λ	fringe spacing [m]
Δn_d	variation of refractive index due to density changes
Δn_d^e	Δn_d due to electrostriction
Δn_K	variation of refractive index due to optical Kerr effect
Δn_d^p	Δn_d due to volume changes
Δn_p	variation of refractive index due to population changes
Δn_d^t	Δn_d due to temperature changes
Δx	peak to null variation of the parameter x
ε	molar decadic absorption coefficient [cm L mol^{-1}]
ζ	angle of polarization of the diffracted signal
η	diffraction efficiency
θ_B	Bragg angle (angle of incidence of the probe pulse)
θ_{pu}	angle of incidence of a pump pulse
λ_{pr}	probe wavelength [m]
λ_{pu}	pump wavelength [m]
ρ	density [kg m^{-3}]
τ_{ac}	acoustic period [s]
υ_{ac}	acoustic frequency [s^{-1}]

See also

Chemical Applications of Lasers: Nonlinear Spectroscopies; Pump and Probe Studies of Femtosecond Kinetics. **Holography, Applications:** Holographic Recording Materials and their Processing. **Holography, Techniques:** Holographic Interferometry. **Materials Characterization Techniques:** $\chi^{(3)}$. **Nonlinear Optics Basics:** Four-Wave Mixing. **Ultrafast Technology:** Femtosecond Chemical Dynamics, Gas Phase; Femtosecond Condensed Phase Spectroscopy – Structural Dynamics.

Further Reading

Bräuchle C and Burland DM (1983) Holographic methods for the investigation of photochemical and photophysical properties of molecules. *Angewandte Chemie International Edition Engl* 22: 582–598.

Eichler HJ, Günter P and Pohl DW (1986) *Laser-Induced Dynamics Gratings*. Berlin: Springer.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford: Oxford University Press.

Fourkas JT and Fayer MD (1992) The transient grating: a holographic window to dynamic processes. *Accounts of Chemical Research* 25: 227–233.

Hall G and Whitaker BJ (1994) Laser-induced grating spectroscopy. *Journal of the Chemistry Society Faraday Trans* 90: 1–16.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy*, revised edn. Boston: Academic Press.

Mukamel S (1995) *Nonlinear Optical Spectroscopy*. Oxford: Oxford University Press.

Rullière C (ed.) (1998) *Femtosecond Laser Pulses*. Berlin: Springer.

Terazima M (1998) Photothermal studies of photophysical and photochemical processes by the transient grating method. *Advances in Photochemistry* 24: 255–338.

cross-sectional area 10^4 times and dispersed the beam across it. The CPA technique makes it possible to use conventional laser amplifiers and to stay below the onset of nonlinear effects.

See also

Diffraction: Diffraction Gratings. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

COHERENCE

Contents

Overview
Coherence and Imaging
Speckle and Coherence

Overview

A Sharma and A K Ghatak, Indian Institute of Technology, New Delhi, India

H C Kandpal, National Physical Laboratory, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Coherence refers to the characteristic of a wave that indicates whether different parts of the wave oscillate in tandem, or in a definite phase relationship. In other words, it refers to the degree of confidence by which one can predict the amplitude and phase of a wave at a point, from the knowledge of these at another point at the same or a different instant of time. Emission from a thermal source, such as a light bulb, is a highly disordered process and the emitted light is incoherent. A well-stabilized laser source, on the other hand, generates light in a highly ordered manner and the emitted light is highly coherent. Incoherent and coherent light represent two extreme cases. While describing the phenomena of physical optics and diffraction theory, light is assumed to be perfectly coherent in both spatial as well as temporal senses, whereas in radiometry it is generally assumed to be incoherent. However, practical light sources and the fields generated by them are in between the two extremes and are termed as partially coherent sources and fields. The degree of order that exists in an optical field produced by a source of any kind may be described in terms of various correlation functions. These correlation functions are the basic theoretical tools for the analyses of statistical properties of partially coherent light fields.

Light fields generated from real physical sources fluctuate randomly to some extent. On a microscopic level quantum mechanical fluctuations produce randomness and on macroscopic level the randomness occurs as a consequence of these microscopic fluctuations, even in free space. In real physical sources, spontaneous emission causes random fluctuations and even in the case of lasers, spontaneous emission cannot be suppressed completely. In addition to spontaneous emission, there are many other processes that give rise to random fluctuations of light fields. Optical coherence theory was developed to describe the random nature of light and it deals with the statistical similarity between light fluctuations at two (or more) space–time points.

As mentioned earlier, in developing the theory of interference or diffraction, light is assumed to be perfectly coherent, or, in other words, it is taken to be monochromatic and sinusoidal for all times. This is, however, an idealization since the wave is obviously generated at some point of time by an atomic or molecular transition. Furthermore, a wavetrain generated by such a transition is of a finite duration, which is related to the finite lifetime of the atomic or molecular levels involved in the transition. Thus, any wave emanating from a source is an ensemble of a large number of such wavetrains of finite duration, say τ_c. A simplified visualization of such an ensemble is shown in Figure 1 where a wave is shown as a series of wavetrains of duration τ_c. It is evident from the figure that the fields at time t and $t + \Delta t$ will have a definite phase relationship if $\Delta t \ll \tau_c$ and will have no or negligible phase relationship when $\Delta t \gg \tau_c$. The time τ_c is known as the coherence time of the

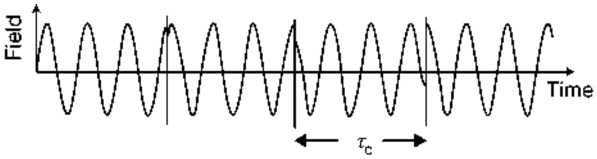

Figure 1 Typical variation of the radiation field with time. The coherence time $\sim \tau_c$.

radiation and the field is said to remain coherent for time $\sim \tau_c$. This property of waves is referred to as the time coherence or the temporal coherence and is related to the spectral purity of the radiation. If one obtains the spectrum of the wave shown in Figure 1 by taking the Fourier transform of the time variation, it would have a width of $\Delta\nu$ around ν_0 which is the frequency of the sinusoidal variation of individual wavetrains. The spectral width $\Delta\nu$ is related to the coherence time as

$$\Delta\nu \sim 1/\tau_c \qquad [1]$$

For thermal sources such as a sodium lamp, $\tau_c \sim 100$ ps, whereas for a laser beam it could be as large as a few milliseconds. A related quantity is the coherence length l_c, which is the distance covered by the wave in time τ_c,

$$l_c = c\tau_c \sim \frac{c}{\Delta\nu} = \frac{\lambda_0^2}{\Delta\lambda} \qquad [2]$$

where λ_0 is the central wavelength ($\lambda_0 = c/\nu_0$) and $\Delta\lambda$ is the wavelength spread corresponding to the spectral width $\Delta\nu$. In a two-beam interference experiment (e.g., Michelson interferometer, Figure 2), the interfering beam derived from the same source will produce good interference fringes if the path difference between the two interfering beams is less than the coherence length of the radiation given out by the source.

It must be added here that for the real fields, generated by innumerable atoms or molecules, the individual wavetrains have different lengths around an average value τ_c. Furthermore, several wavetrains in general are propagating simultaneously, overlapping in space and time, to produce an ensemble whose properties are best understood in a statistical sense as we shall see in later sections.

Another type of coherence associated with the fields is the space coherence or the spatial coherence, which is related to the size of the source of radiation. It is evident that when the source is an ideal point source, the field at any two points (within the coherence length) would have definite phase relationship. However, the field from a thermal source of finite area S can be thought as resultant of the fields from each point on the source. Since each point source would usually be independent of the other, the phase relationship between the fields at any two points would depend on their position and size of the source. It can be shown that two points will have a strong phase relationship if they lie within the solid angle $\Delta\Omega$ from the source (Figure 3) such that

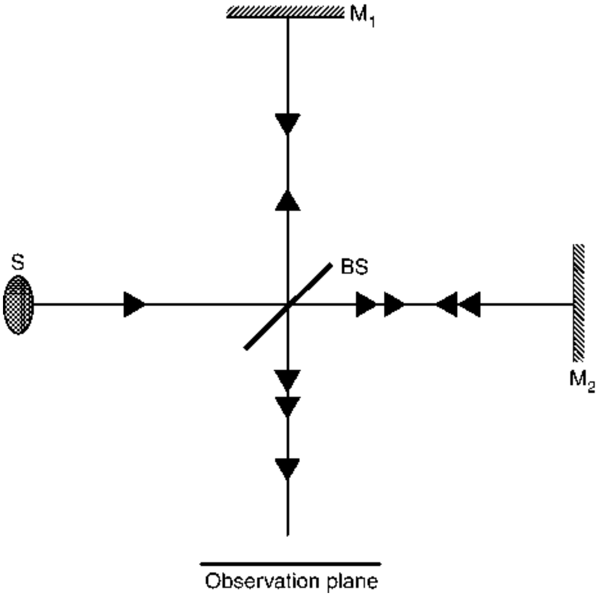

Figure 2 Michelson interferometer to study the temporal coherence of radiation from source S; M_1 and M_2 are mirrors and BS is a 50%–50% beamsplitter.

$$\Delta\Omega \sim \frac{\lambda_0^2}{S} \qquad [3]$$

Thus, on a plane R distance away from the source, one can define an area $A_c = R^2\Delta\Omega$ over which the field remains spatially coherent. This area A_c is called the coherence area of the radiation and its square root is sometimes referred to as the transverse coherence length. It is trivial to show that

$$A_c \sim \frac{\lambda_0^2}{\Delta\Omega_S} \qquad [4]$$

where $\Delta\Omega_S$ is the solid angle subtended by the source on the plane at which the coherence area is defined. In Young's double-hole experiment, if the two pinholes lie within the coherence area of the radiation from the primary source, good contrast in fringes would be observed.

Combining the concepts of coherence length and the coherence area, one can define a coherence volume as $V_c = A_c l_c$. For the wavefield from a thermal source, this volume represents that portion of the space in which the field is coherent and any

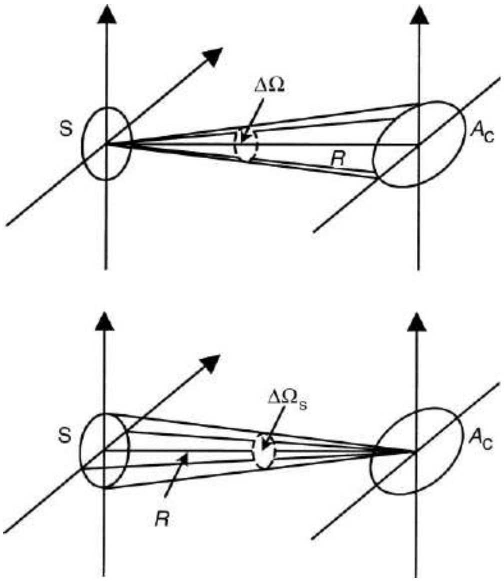

Figure 3 Spatial coherence and coherence area.

interference produced using the radiation from points within this volume will produce fringes of good contrast.

Mathematical Description of Coherence

Analytical Field Representation

Coherence properties associated with fields are best analyzed in terms of complex representation for optical fields. Let the real function $V^{(r)}(\mathbf{r}, t)$ represent the scalar field, which could be one of the transverse Cartesian components of the electric field associated with the electromagnetic wave. One can then define a complex analytical signal $V(\mathbf{r}, t)$ such that

$$\begin{gathered} V^{(r)}(\mathbf{r}, t) = \mathrm{Re}[V(\mathbf{r}, t)], \\ V(\mathbf{r}, t) = \int_0^\infty \tilde{V}^{(r)}(\mathbf{r}, \nu) \mathrm{e}^{-2\pi i \nu t}\, \mathrm{d}\nu \end{gathered} \tag{5}$$

where the spectrum $\tilde{V}^{(r)}(\mathbf{r}, \nu)$ is the Fourier transform of the scalar field $V^{(r)}(\mathbf{r}, t)$ and the spectrum for negative frequencies has been suppressed as it does not give any new information since $\tilde{V}^{(r)^*}(\mathbf{r}, \nu) = \tilde{V}^{(r)}(\mathbf{r}, -\nu)$.

In general, the radiation from a quasi-monochromatic thermal source fluctuates randomly as it is made of a large number of mutually independent contributions from individual atoms or molecules in the source. The field from such a source can be regarded as an ensemble of a large number of randomly different analytical signals such as $V(\mathbf{r}, t)$. In other words, $V(\mathbf{r}, t)$ is a typical member of an ensemble, which is the result of a random process representing the radiation from a quasi-monochromatic source. This process is assumed to be stationary and ergodic so that the ensemble average is equal to the time average of a typical member of the ensemble and that the origin of time is unimportant. Thus, the quantities of interest in the theory of coherence are defined as time averages:

$$\langle f(t) \rangle = \operatorname*{Lim}_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t)\mathrm{d}t \tag{6}$$

Mutual Coherence

In order to define the mutual coherence, the key concept in the theory of coherence, we consider Young's interference experiment, as shown in Figure 4, where the radiation from a broad source of size S is illuminating a screen with two pinholes P_1 and P_2. The light emerging from the two pinholes produces interference, which is observed on another screen at a distance R from the first screen. The field at point P, due to the pinholes, would be $K_1 V(\mathbf{r}_1, t - t_1)$ and $K_2 V(\mathbf{r}_2, t - t_2)$ respectively, where $\mathbf{r}_1$ and $\mathbf{r}_2$ define the positions of P_1 and P_2, t_1 and t_2 are the times taken by the light to travel to P from P_1 and P_2, and K_1 and K_2 are imaginary constants that depend on the geometry and size of the respective pinhole and its distance from the point P. Thus, the resultant field at point P would be given by

$$V(\mathbf{r}, t) = K_1 V(\mathbf{r}_1, t - t_1) + K_2 V(\mathbf{r}_2, t - t_2) \tag{7}$$

Since the optical periods are extremely small as compared to the response time of a detector, the detector placed at point P will record only the time-averaged intensity:

$$I(P) = \langle V^*(\mathbf{r}, t) V(\mathbf{r}, t) \rangle \tag{8}$$

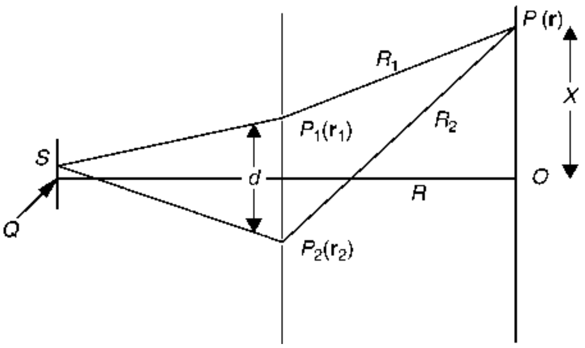

Figure 4 Schematic of Young's double-hole experiment.

where some constants have been ignored. With eqn [7], the intensity at point P would be given by

$$I(P) = I_1 + I_2 + 2\,\mathrm{Re}\left\{K_1 K_2^* \Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2)\right\} \quad [9]$$

where I_1 and I_2 are the intensities at point P, respectively, due to radiations from pinholes P_1 and P_2 independently (defined as $I_i = \langle |K_i V(\mathbf{r}_i, t)|^2 \rangle$) and

$$\begin{aligned}\Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2) &\equiv \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) \\ &= \left\langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau) \right\rangle \quad [10]\end{aligned}$$

is the mutual coherence function of the fields at P_1 and P_2, and it depends on the time difference $\tau = t_2 - t_1$, since the random process is assumed to be stationary. This function is also sometimes denoted by $\Gamma_{12}(\tau)$. The function $\Gamma_{ii}(\tau) \equiv \Gamma(\mathbf{r}_i, \mathbf{r}_i, \tau)$ defines the self-coherence of the light from the pinhole at P_i, and $|K_i|^2 \Gamma_{ii}(0)$ defines the intensity I_i at point P, due to the light from pinhole P_i.

Degree of Coherence and the Visibility of Interference Fringes

One can define a normalized form of the mutual coherence function, namely:

$$\begin{aligned}\gamma_{12}(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) &= \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}} \\ &= \frac{\langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau) \rangle}{\left[\left\langle |V(\mathbf{r}_1, t)|^2 \right\rangle \left\langle |V(\mathbf{r}_2, t)|^2 \right\rangle\right]^{1/2}} \quad [11]\end{aligned}$$

which is called the complex degree of coherence. It can be shown that $0 \le |\gamma_{12}(\tau)| \le 1$. The intensity at point P, given by eqn [9], can now be written as

$$I(P) = I_1 + I_2 + 2\sqrt{I_1 I_2}\,\mathrm{Re}[\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)] \quad [12]$$

Expressing $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ as

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \exp[\mathrm{i}\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi \mathrm{i} \nu_0 \tau] \quad [13]$$

where $\alpha(\vec{r}_1, \vec{r}_2, \tau) = \arg[\gamma(\vec{r}_1, \vec{r}_2, \tau)] + 2\pi\nu_0\tau$, and ν_0 is the mean frequency of the light, the intensity in eqn [12] can be written as

$$\begin{aligned}I(P) = {} & I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \\ & \times \cos[\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi\nu_0\tau] \quad [14]\end{aligned}$$

Now, if we assume that the source is quasi-monochromatic, i.e., its spectral width $\Delta\nu \ll \nu_0$, the quantities $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ and $\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau)$ vary slowly on the observation screen, and the interference fringes are mainly obtained due to the cosine term. Thus, defining the visibility of fringes as $v = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$, we obtain

$$v = \frac{2(I_1 I_2)^{1/2}}{I_1 + I_2} |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \quad [15]$$

which shows that for maximum visibility, the two interfering fields must be completely coherent ($|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$). On the other hand, if the fields are completely incoherent ($|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 0$), no fringes are observed ($I_{\max} = I_{\min}$). The fields are said to be partially coherent when $0 < |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| < 1$. When $I_1 = I_2$, the visibility is the same as $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|$. The relation in eqn [15] shows that in an interference experiment, one can obtain the modulus of the complex degree of coherence by measuring I_1, I_2, and the visibility. Similarly, eqn [14] shows that from the measurement the positions of maxima, one can obtain the phase of the complex degree of coherence, $\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau)$.

Temporal and Spatial Coherence

If the source illuminating the pinholes is a point source of finite spectral width situated at point Q, the fields at point P_1 and P_2 (Figure 4) at any given instant are the same and the mutual coherence function becomes

$$\begin{aligned}\Gamma_{11}(\tau) = \Gamma(\mathbf{r}_1, \mathbf{r}_1, \tau) &= \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_1, t + \tau) \rangle \\ &= \langle V^*(\mathbf{r}_2, t) V(\mathbf{r}_2, t + \tau) \rangle = \Gamma_{22}(\tau) \quad [16]\end{aligned}$$

The self-coherence, $\Gamma_{11}(\tau)$, of the light from pinhole P_1, is a direct measure of the temporal coherence of the source. On the other hand, if the source is of finite size and we observe the interference at point O which corresponds to $\tau = 0$, the mutual coherence function would be

$$\Gamma_{12}(0) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t) \rangle \equiv J_{12} \quad [17]$$

which is called the mutual intensity and is a direct measure of the spatial coherence of the source. In general, however, the function $\Gamma_{12}(\tau)$ measures, for a source of finite size and spectral width, a combination of the temporal and spatial coherence, and in only some limiting cases, the two types of coherence can be separated.

Spectral Representation of Mutual Coherence

One can also analyze the correlation between two fields in the spectral domain. In particular, one can define the cross-spectral density function $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$ which defines the correlation between the amplitudes of the spectral components of frequency ν of the light at the points P_1 and P_2. Thus:

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu)\delta(\nu - \nu') = \langle V^*(\mathbf{r}_1, \nu)V(\mathbf{r}_2, \nu')\rangle \qquad [18]$$

Using the generalized Wiener–Khintchine theorem, the cross-spectral density function can be shown to be the Fourier transform of the mutual coherence function:

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \int_0^\infty W(\mathbf{r}_1, \mathbf{r}_2, \nu)\mathrm{e}^{-2\pi i \nu t}\,\mathrm{d}\nu \qquad [19]$$

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^\infty \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\mathrm{e}^{2\pi i \nu t}\,\mathrm{d}\tau \qquad [20]$$

If the two points P_1 and P_2 coincide (i.e., there is only one pinhole), the cross-spectral density function reduces to the spectral density function of the light, which we denote by $S(\mathbf{r}, \nu)[\equiv W(\mathbf{r}, \mathbf{r}, \nu)]$. Thus, it follows from eqn [20] that the spectral density of light is the inverse Fourier transform of the self-coherence function. This leads to the Fourier transform spectroscopy, as we shall see later. One can also define spectral degree of coherence at frequency ν as

$$\begin{aligned}\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) &= \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{W(\mathbf{r}_1, \mathbf{r}_1, \nu)W(\mathbf{r}_2, \mathbf{r}_2, \nu)}} \\ &= \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{S(\mathbf{r}_1, \nu)S(\mathbf{r}_2, \nu)}}\end{aligned} \qquad [21]$$

It is easy to see that $0 \leq |\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)| \leq 1$. It is also sometimes referred to as the complex degree of coherence at frequency ν. It may be noted here that in the literature the notation $G^{(1,1)}(\nu) \equiv G(\mathbf{r}_1, \mathbf{r}_2, \gamma)$ has also been used for $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$.

Propagation of Coherence

The van Cittert–Zernike Theorem

Perfectly coherent waves propagate through diffraction formulae, which have been discussed in this volume elsewhere. However, incoherent and partially coherent waves would evidently propagate somewhat differently. The earliest treatment of propagation noncoherent light is due to van Cittert, which was later generalized by Zernike to obtain what is now the van Cittert–Zernike theorem. The theorem deals with the correlations developed between fields at two points, which have been generated by a quasi-monochromatic and spatially incoherent planar source. Thus, as shown in Figure 5, we consider a quasi-monochromatic ($\Delta\nu \ll \nu_0$) planar source σ, which has an intensity distribution $I(\mathbf{r}')$ on its plane and is spatially incoherent, i.e., there is no correlation between the fields at any two points on the source.

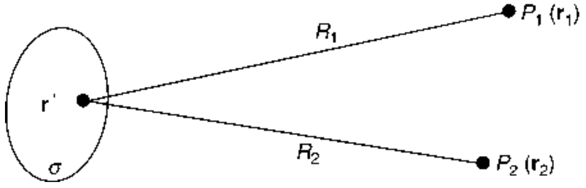

Figure 5 Geometry for the van Cittert–Zernike theorem.

The field, due to this source, would develop finite correlations after propagation, and the theorem states that

$$\begin{aligned}\gamma(\mathbf{r}_1, \mathbf{r}_2, 0) &= \frac{\Gamma_{12}(0)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{12}(0)}} \\ &= \frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}}\int\int_\sigma I(\mathbf{r}')\frac{\mathrm{e}^{2\pi i\nu_0(R_2 - R_1)/c}}{R_1R_2}\mathrm{d}^2r'\end{aligned} \qquad [22]$$

where $R_i = |\mathbf{r}_i - \mathbf{r}'|$ and $I(\mathbf{r}_i)$ is the intensity at point P_i. This relation is similar to the diffraction pattern produced at point P_1 due to a wave, with a spherical wavefront converging towards P_2 and with an amplitude distribution $I(\mathbf{r}')$ when it is diffracted by an aperture of the same shape, size, and the intensity distribution as those of the incoherent source σ. Thus, the theorem shows that a radiation, which was incoherent at the source, becomes partially coherent as it propagates.

Generalized Propagation

The expression $\mathrm{e}^{-2\pi i \nu_0 R_1/c}/R_1$ can be interpreted as the field obtained at P_1 due to a point source located at the point $\mathbf{r}'$ on the planar source. Thus, this expression is simply the point spread function of the homogeneous space between the source and the observation plane containing the points P_1 and P_2. Hopkins generalized this to include any linear optical system characterized by a point spread function $h(\mathbf{r}_j - \mathbf{r}')$ and obtained the formula for the complex degree of coherence of a radiation emerging from an incoherent, quasi-monochromatic planar source after it has propagated through such a linear optical system:

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}}\int\int_\sigma I(\mathbf{r}')h(\mathbf{r}_1 - \mathbf{r}')h^*(\mathbf{r}_2 - \mathbf{r}')\mathrm{d}^2r' \qquad [23]$$

It would thus seem that the correlations propagate in much the same way, as does the field. Indeed, Wolf showed that the mutual correlation function

$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ satisfies the wave equations:

$$\begin{aligned}\nabla_1^2\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)&=\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)\\ \nabla_2^2\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)&=\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)\end{aligned}\qquad[24]$$

where ∇_j^2 is the Laplacian with respect to the point $\mathbf{r}_i$. Here the first of eqns [24], for instance, gives the variation of the mutual coherence function with respect to $\mathbf{r}_1$ and τ for fixed $\mathbf{r}_2$. Further, the variable τ is the time difference defined through the path difference, since $\tau=(R_2-R_1)/c$, and the actual time does not affect the mutual coherence function (as the fields are assumed to be stationary).

Using the relation in eqn [20], one can also obtain from eqn [24], the propagation equations for the cross-spectral density function $W(\mathbf{r}_1,\mathbf{r}_2,\nu)$:

$$\begin{aligned}\nabla_1^2 W(\mathbf{r}_1,\mathbf{r}_2,\nu)+k^2W(\mathbf{r}_1,\mathbf{r}_2,\nu)&=0\\ \nabla_2^2 W(\mathbf{r}_1,\mathbf{r}_2,\nu)+k^2W(\mathbf{r}_1,\mathbf{r}_2,\nu)&=0\end{aligned}\qquad[25]$$

where $k=2\pi\nu/c$.

Thompson and Wolf Experiment

One of the most elegant experiments for studying various aspects of coherence theory was carried out by Thompson and Wolf by making slight modifications in the Young's double-hole experiment. The experimental set-up shown in Figure 6 consists of a quasi-monochromatic broad incoherent source S of diameter $2a$. This was obtained by focusing filtered narrow band light from a mercury lamp (not shown in the figure) on to a hole of size $2a$ in an opaque screen A. A mask consisting of two pinholes, each of diameter $2b$ with their axes separated by a distance d, was placed symmetrically about the optical axis of the experimental setup at plane B between two lenses L_1 and L_2, each having focal length f. The source was at the front focal plane of the lens L_1 and the observations were made at the plane C at the back focus of the lens L_2. The separation d was varied to study the spatial coherence function on the mask plane by measuring the visibility and the phase of the fringes formed in plane C.

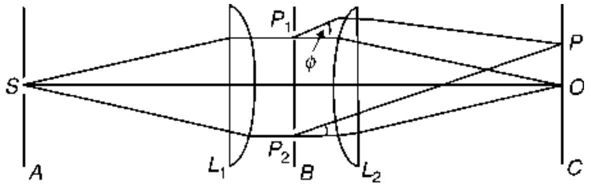

Figure 6 Schematic of the Thompson and Wolf experiment.

Using the van Cittert–Zernike theorem, the complex degree of coherence at the pinholes P_1 and P_2 on the mask after the lens L_1 is

$$\gamma_{12}=|\gamma_{12}|e^{i\beta_{12}}=\frac{2J_1(v)}{v}\ \text{with}\ v=\left(\frac{2\pi\nu ad}{cf}\right)\qquad[26]$$

for symmetrically placed pinholes about the optical axis. Here β_{12} is the phase of the complex degree of coherence and in this special case where the two holes are equidistant from the axis, β_{12} is either zero or π, respectively, for positive or negative values of $2J_1(v)/v$. Let us assume that the intensities at two pinholes P_1 and P_2 are equal. The interference pattern observed at the back focal plane of lens L_2 is due to the superposition of the light diffracted from the pinholes. The beams are partially coherent with degree of coherence given by eqn [26]. Since the pinholes are symmetrically placed, the intensity due to either of the pinholes at a point P at the back focal plane of the lens L_2 is the same and is given by the Fraunhofer formula for diffraction from circular apertures, i.e.:

$$I_1(P)=I_2(P)=\left|\frac{2J_1(u)}{u}\right|^2,\ \text{with}\ u=\frac{2\pi\nu}{c}b\sin\phi\qquad[27]$$

and ϕ is the angle that the diffracted beam makes from normal to the plane of the pinholes. The intensity of the interference pattern produced at the back focal plane of the lens L_2 is:

$$I(P)=2I_1(P)\left[1+\left|\frac{2J_1(v)}{v}\right|^2\cos(\delta+\beta_{12})\right]\qquad[28]$$

where $\delta=\mathrm{d}\sin\phi$ is the phase difference between the two beams reaching P from P_1 and P_2. For the on-axis point O, the quantity δ is zero.

The maximum and minimum values of $I(P)$ are given by

$$I_{\max}(P)=2I_1(P)\left[1+|2J_1(v)/v|^2\right]\qquad[29(a)]$$

$$I_{\max}(P)=2I_1(P)\left[1-|2J_1(v)/v|^2\right]\qquad[29(b)]$$

Figure 7 shows an example of the observed fringe patterns obtained by Thompson (1958) in a subsequent experiment.

Types of Fields

As mentioned above, $\gamma_{12}(\tau)$ is a measure of the correlation of the complex field at any two points P_1

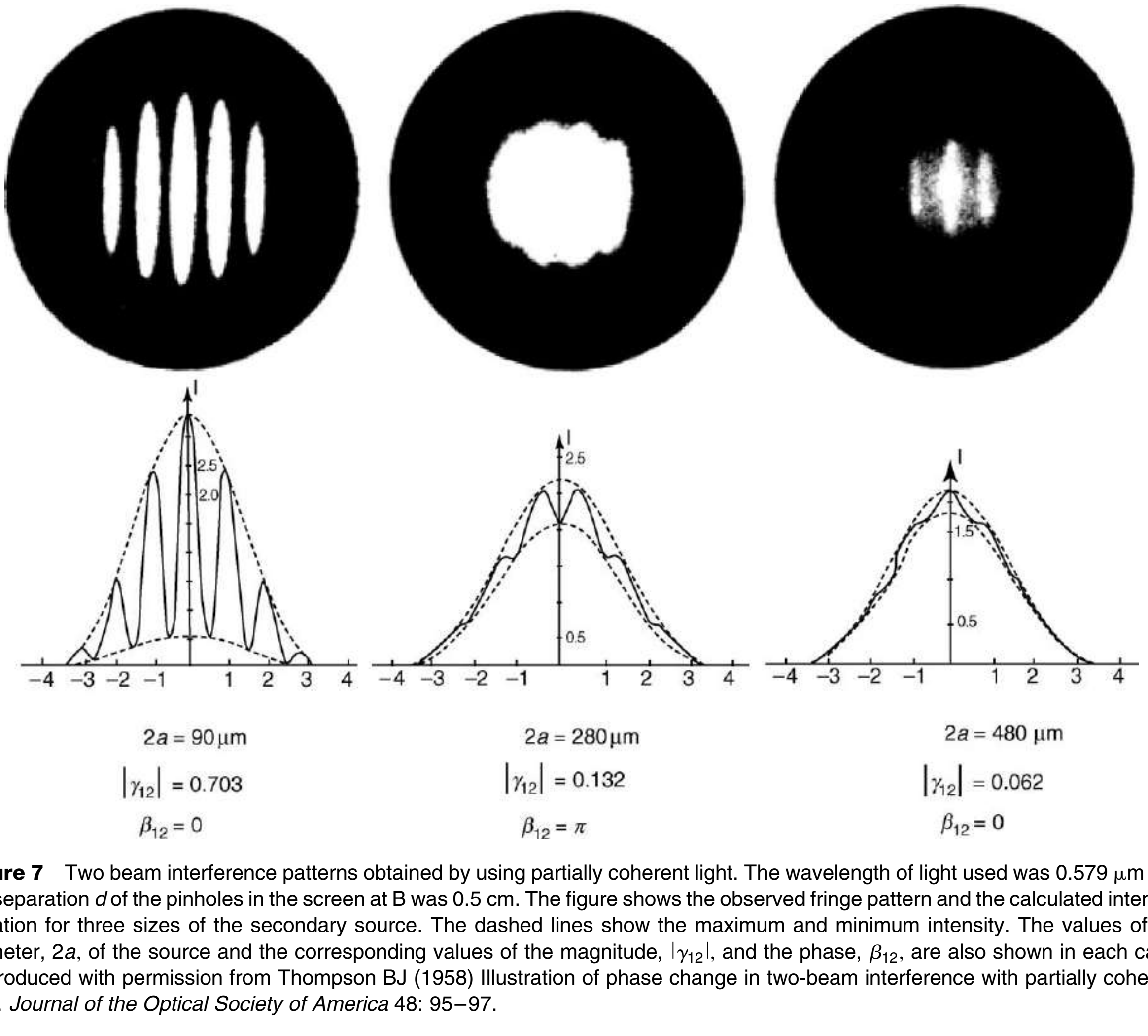

$2a = 90\,\mu\text{m}$ $|\gamma_{12}| = 0.703$ $\beta_{12} = 0$

$2a = 280\,\mu\text{m}$ $|\gamma_{12}| = 0.132$ $\beta_{12} = \pi$

$2a = 480\,\mu\text{m}$ $|\gamma_{12}| = 0.062$ $\beta_{12} = 0$

Figure 7 Two beam interference patterns obtained by using partially coherent light. The wavelength of light used was 0.579 μm and the separation d of the pinholes in the screen at B was 0.5 cm. The figure shows the observed fringe pattern and the calculated intensity variation for three sizes of the secondary source. The dashed lines show the maximum and minimum intensity. The values of the diameter, $2a$, of the source and the corresponding values of the magnitude, $|\gamma_{12}|$, and the phase, β_{12}, are also shown in each case. Reproduced with permission from Thompson BJ (1958) Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48: 95–97.

and P_2 at specific time delay τ. Extending this definition, an optical field may be coherent or incoherent, if $|\gamma_{12}(\tau)| = 1$ or $|\gamma_{12}(\tau)| = 0$, respectively, for all pairs of points in the field and for all time delays. In the following, we consider some specific cases of fields and their properties.

Perfectly Coherent Fields

A field would be termed as perfectly coherent or self-coherent at a fixed point, if it has the property that $|\gamma(\mathbf{R}, \mathbf{R}, \tau)| = 1$ at some specific point $\mathbf{R}$ in the domain of the field for all values of time delay τ. It can also be shown that $\gamma(\mathbf{R}, \mathbf{R}, \tau)$ is periodic in time, i.e.:

$$\gamma(\mathbf{R}, \mathbf{R}, \tau) = \mathrm{e}^{-2\pi \mathrm{i} \nu_0 \tau} \quad \text{for} \quad \nu_0 > 0 \qquad [30]$$

For any other point $\mathbf{r}$ in the domain of the field, $\gamma(\mathbf{R}, \mathbf{r}, \tau)$ and $\gamma(\mathbf{r}, \mathbf{R}, \tau)$ are also unimodular and also periodic in time.

A perfectly coherent optical field at two fixed points has the property that $|\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau)| = 1$ for any two fixed points $\mathbf{R}_1$ and $\mathbf{R}_2$ in the domain of the field for all values of time delay τ. For such a field, it can be shown that $\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \exp[\mathrm{i}(\beta - 2\pi\nu_0\tau)]$ with ν_0 (> 0) and β are real constants. Further, as a corollary, $|\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau)| = 1$ and $|\gamma(\mathbf{R}_2, \mathbf{R}_2, \tau)| = 1$ for all τ, i.e., the field is self-coherent at each of the field points $\mathbf{R}_1$ and $\mathbf{R}_2$, and

$$\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau) = \gamma(\mathbf{R}_2, \mathbf{R}_2, \tau) = \exp(-2\pi \mathrm{i} \nu_0 \tau) \text{ for } \nu_0 > 0$$

It can be shown that for any other point $\mathbf{r}'$ within the field

$$\begin{aligned} &\gamma(\mathbf{r}, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0)\gamma(\mathbf{r}, \mathbf{R}_1, \tau) = \gamma(\mathbf{r}, \mathbf{R}_1, \tau)\mathrm{e}^{\mathrm{i}\beta} \\ &\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0)\exp[-2\pi \mathrm{i}\nu_0\tau] \end{aligned} \qquad [31]$$

A perfectly coherent optical field for all points in the domain of the field has the property that $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$ for all pairs of points $\mathbf{r}_1$ and $\mathbf{r}_2$ in the domain of the field, for all values of time delay τ.

It can then be shown that $\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\exp\{\mathrm{i}[\alpha(\mathbf{r}_1)-\alpha(\mathbf{r}_2)]-2\pi\mathrm{i}\nu_0\tau\}$, where $\alpha(\mathbf{r}')$ is real function of a single position $\mathbf{r}'$ in the domain of the field. Further, the mutual coherence function $\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ for such a field has a factorized periodic form as

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=U(\mathbf{r}_1)U^*(\mathbf{r}_2)\ \exp(-2\pi\mathrm{i}\nu_0\tau) \qquad [32]$$

which means the field is monochromatic with frequency ν_0 and $U(\mathbf{r})$ is any solution of the Helmholtz equation:

$$\nabla^2 U(\mathbf{r})+k_0^2U(\mathbf{r})=0, \quad k_0=2\pi\nu_0/c \qquad [33]$$

The spectral and cross-spectral densities of such fields are represented by Dirac δ-functions with their singularities at some positive frequency ν_0. Such fields, however, can never occur in nature.

Quasi-Monochromatic Fields

Optical fields are found in practice for which spectral spread $\Delta\nu$ is much smaller than the mean frequency $\bar{\nu}$. These are termed as quasi-monochromatic fields. The cross-spectral density $W(\mathbf{r}_1,\mathbf{r}_2,\nu)$ of the quasi-monochromatic fields attains an appreciable value only in the small region $\Delta\nu$, and falls to zero outside this region.

$$W(\mathbf{r}_1,\mathbf{r}_2,\nu)=0, \quad |\nu-\bar{\nu}|>\Delta\nu \ \text{ and } \ \Delta\nu\ll\bar{\nu} \qquad [34]$$

The mutual coherence function $\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ can be written as the Fourier transform of cross-spectral density function as

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\mathrm{e}^{-2\pi\mathrm{i}\bar{\nu}\tau}\int_0^\infty W(\mathbf{r}_1,\mathbf{r}_2,\nu)\mathrm{e}^{-2\pi\mathrm{i}(\nu-\bar{\nu})\tau}\mathrm{d}\nu \qquad [35]$$

If we limit our attention to small τ such that $\Delta\nu|\tau|\ll 1$ holds, the exponential factor inside the integral is approximately equal to unity and eqn [35] reduces to

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\exp(-2\pi\mathrm{i}\bar{\nu}\tau)\int_0^\infty [W(\mathbf{r}_1,\mathbf{r}_2,\nu)]\mathrm{d}\nu \qquad [36]$$

which gives

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\Gamma(\mathbf{r}_1,\mathbf{r}_2,0)\ \exp(-2\pi\mathrm{i}\bar{\nu}\tau) \qquad [37]$$

Equation [37] describes the behavior of $\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ for a limited range of τ values for quasi-monochromatic fields and in this range it behaves as a monochromatic field of frequency $\bar{\nu}$. However, due to the factor $\Gamma(\mathbf{r}_1,\mathbf{r}_2,0)$, the quasi-monochromatic field may be coherent, partially coherent, or even incoherent.

Cross-Spectrally Pure Fields

The complex degree of coherence, if it can be factored into a product of a component dependent on spatial coordinates and a component dependent on time delay, is called reducible. In the case of perfectly coherent light, the complex degree of coherence is reducible, as we have seen above, e.g., in eqn [32], and in the case of quasi-monochromatic fields, this is reducible approximately as shown in eqn [37]. There also exists a very special case of a cross-spectrally pure field for which the complex degree of coherence is reducible. A field is called a cross-spectrally pure field if the normalized spectrum of the superimposed light is equal to the normalized spectrum of the component beams, a concept introduced by Mandel. In the space-frequency domain, the intensity interference law is expressed as the so-called spectral interference law:

$$S(\mathbf{r},\nu)= S^{(1)}(\mathbf{r},\nu)+S^{(2)}(\mathbf{r},\nu) + 2\left[\sqrt{S^{(1)}(\mathbf{r},\nu)}\sqrt{S^{(2)}(\mathbf{r},\nu)}\right]\mathrm{Re}[\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)\mathrm{e}^{-2\pi\mathrm{i}\nu\tau}] \qquad [38]$$

where $\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)$ is the spectral degree of coherence, defined in eqn [21] and τ is the relative time delay that is needed by the light from the pinholes to reach any point on the screen; $S^{(1)}(\mathbf{r},\nu)$ and $S^{(2)}(\mathbf{r},\nu)$ are the spectral densities of the light reaching P from the pinholes P_1 and P_2, respectively (see Figure 4) and it is assumed that the spectral densities of the field at pinholes P_1 and P_2 are the same $[S(\mathbf{r}_1,\nu)=CS(\mathbf{r}_2,\nu)]$. Now, if we consider a point for which the time delay is τ_0, then it can be seen that the last term in eqn [38] would be independent of frequency, provided that

$$\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)\ \exp(-2\pi\mathrm{i}\nu\tau_0)=f(\mathbf{r}_1,\mathbf{r}_2,\tau_0) \qquad [39]$$

where $f(\mathbf{r}_1,\mathbf{r}_2,\tau_0)$ is a function of $\mathbf{r}_1$, $\mathbf{r}_2$ and τ_0 only and the light at this point would have the same spectral density as that at the pinholes. If a region exists around the specified point on the observation plane, such that the spectral distribution of the light in this region is of the same form as the spectral distribution of the light at the pinholes, the light at the pinholes is cross-spectrally pure light.

In terms of the spectral distribution of the light $S(\mathbf{r}_1,\nu)$ at pinhole P_1 and $S(\mathbf{r}_2,\nu)[=CS(\mathbf{r}_1,\nu)]$ at pinhole P_2, the mutual coherence function at the pinholes can be written as

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\sqrt{C}\int S(\mathbf{r}_1,\nu)\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)\ \exp(-2\pi\mathrm{i}\nu\tau)\mathrm{d}\nu \qquad [40]$$

and using eqn [39], we get the very important condition for the field to be cross-spectrally pure, i.e.:

$$\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau_0)\gamma(\mathbf{r}_1,\mathbf{r}_1,\tau-\tau_0) \qquad [41]$$

The complex degree of coherence $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ is thus expressible as the product of two factors: one factor characterizes the spatial coherence at the two pinholes at time delay τ_0 and the other characterizes the temporal coherence at one of the pinholes. Equation [41] is known as the reduction formula for cross-spectrally pure light. It can further be shown that

$$\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) = \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau_0)\ \exp(2\pi i \nu \tau_0) \qquad [42]$$

Thus, the absolute value of the spectral degree of coherence is the same for all frequencies and is equal to the absolute value of the degree of coherence for the point specified by τ_0. It has been shown that cross-spectrally pure light can be generated, for example, by linear filtering of light that emerges from the two pinholes in Young's interference experiments.

Types of Sources

Primary Sources

A primary source is a set of radiating atoms or molecules. In a primary source the randomness comes from true source fluctuations, i.e., from the spatial distributions of fluctuating charges and currents. Such a source gives rise to a fluctuating field. Let $Q(\mathbf{r}, t)$ represent the fluctuating source variable at any point $\mathbf{r}$ at time t, then the field generated by the source is represented by fluctuating field variable $V(\mathbf{r}, t)$. The source is assumed to be localized in some finite domain such that $Q(\mathbf{r}, t) = 0$ at any time $t > 0$ outside the domain. Assuming that field variable $V(\mathbf{r}, t)$ and the source variable $Q(\mathbf{r}, t)$ are scalar quantities, they are related by an inhomogeneous equation as

$$\nabla^2 V(\mathbf{r}, t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} V(\mathbf{r}, t) = -4\pi Q(\mathbf{r}, t) \qquad [43]$$

The mutual coherence functions of the source $\Gamma_Q(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle Q^*(\mathbf{r}_1, t) Q(\mathbf{r}_2, t + \tau)\rangle$ and of the field $\Gamma_V(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau)\rangle$ characterize the statistical similarity of the fluctuating quantities at the points $\mathbf{r}_1$ and $\mathbf{r}_2$. Following eqn [20], one can define, respectively, the cross-spectral density function of the source and the field as

$$W_Q(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma_Q(\mathbf{r}_1, \mathbf{r}_2, \tau) \mathrm{e}^{-2\pi i \nu t}\, \mathrm{d}\tau \qquad [44]$$

$$W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma_V(\mathbf{r}_1, \mathbf{r}_2, \tau) \mathrm{e}^{-2\pi i \nu t}\, \mathrm{d}\tau$$

The cross-spectral density functions of the source and of the field are related as

$$(\nabla_2^2 + k^2)(\nabla_1^2 + k^2) W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = 4\pi^2 W_Q(\mathbf{r}_1, \mathbf{r}_2, \nu) \qquad [45]$$

The solution of eqn [45] is represented as

$$W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_S \int_S W_Q(\mathbf{r}'_1, \mathbf{r}'_2, \nu) \frac{\mathrm{e}^{ik(R_2 - R_1)}}{R_1 R_2} \mathrm{d}^3 r'_1\, \mathrm{d}^3 r'_2 \qquad [46]$$

where $R_1 = |\mathbf{r}_1 - \mathbf{r}'_1|$ and $R_2 = |\mathbf{r}_2 - \mathbf{r}'_2|$ (see Figure 8). Using eqn [46], one can then obtain an expression for the spectrum at a point $(\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r} = r\mathbf{u})$ in the far field $(r \gg r'_1, r'_2)$ of a source as

$$S^\infty(r\mathbf{u}, \nu) = \frac{1}{r^2} \int_S \int_S W_Q(\mathbf{r}'_1, \mathbf{r}'_2, \nu) \mathrm{e}^{-ik\mathbf{u}\cdot(\mathbf{r}'_1 - \mathbf{r}'_2)} \mathrm{d}^3 r'_1\, \mathrm{d}^3 r'_2 \qquad [47]$$

where $\mathbf{u}$ is the unit vector along $\mathbf{r}$. The integral in eqn [47], i.e., the quantity defined by $r^2 S^\infty(r\mathbf{u}, \nu)$, is also defined as the radiant intensity, which represents the rate of energy radiated at the frequency ν from the source per unit solid angle in the direction of $\mathbf{u}$.

Secondary Sources

Sources used in a laboratory are usually secondary planar sources. A secondary source is a field, which arises from the primary source in the region outside the domain of the primary sources. This kind of source is an aperture on an opaque screen illuminated either directly or via an optical system by primary sources. Let $V(\boldsymbol{\rho}, t)$ represent the fluctuating field in a secondary source plane σ at $z = 0$ (Figure 9) and $W_0(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \nu)$ represent its cross-spectral density (the subscript 0 refers to $z = 0$). One can then solve eqn [25] to obtain the propagation of the cross-spectral density from this planar source. For two points P_1 and P_2 located at distances which are large compared to wavelength, the cross-spectral density is

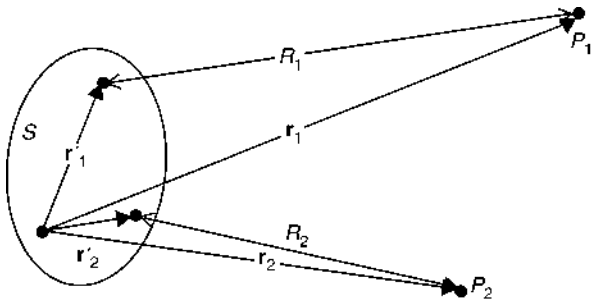

Figure 8 Geometry of a 3-dimensional primary source S and the radiation from it.

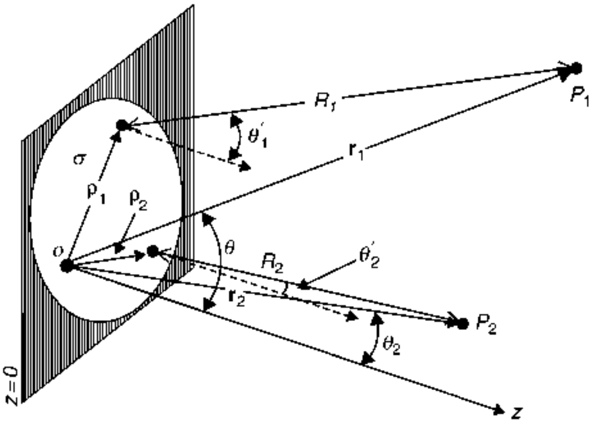

Figure 9 Geometry of a planar source σ and the radiation from it.

then given by

$$W(\mathbf{r}_1,\mathbf{r}_2,\nu)=\left(\frac{k}{2\pi}\right)^2\int_\sigma\int_\sigma W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2,\nu)\times\frac{\mathrm{e}^{ik(R_2-R_1)}}{R_1R_2}\cos\theta'_1\cos\theta'_2\,\mathrm{d}^2\rho_1\,\mathrm{d}^2\rho_2 \quad [48]$$

where $R_j=|\mathbf{r}_j-\boldsymbol{\rho}_j|$, $(j=1,2)$, θ'_1 and θ'_2 are the angles that R_1 and R_2 directions make with the z-axis (Figure 9). Using eqn [48], one can then obtain an expression for the spectrum at a point ($\mathbf{r}_1=\mathbf{r}_2=\mathbf{r}=r\mathbf{u}$) in the far field ($r\gg\rho_1,\rho_2$) of a planar source as

$$S^\infty(r\mathbf{u},\nu)=\frac{k^2}{2\pi^2}\frac{\cos^2\theta}{r^2}\int_\sigma\int_\sigma W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2,\nu)\times\mathrm{e}^{-ik\mathbf{u}_\perp\cdot(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)}\mathrm{d}^2\rho_1\,\mathrm{d}^2\rho_2 \quad [49]$$

where $\mathbf{u}_\perp$ is the projection of the unit vector $\mathbf{u}$ on the plane σ of the source.

Schell-Model Sources

In the framework of coherence theory in space–time domain, two-dimensional planar model sources of this kind were first discussed by Schell. Later, the model was adopted for formulation of coherence theory in space–frequency domain. Schell-model sources are the sources whose degree of spectral coherence $\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)$ (for either primary or secondary source) is stationary in space. It means that $\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)$ depends on $\mathbf{r}_1$ and $\mathbf{r}_2$ only through the difference $\mathbf{r}_2-\mathbf{r}_1$, i.e., of the form:

$$\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)\equiv\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu) \quad [50]$$

for each frequency ν present in the source spectrum. Here A stands for field variables V, in the case of a Schell-model secondary source and for source variables Q, in the case of a Schell-model primary source. The cross-spectral density function of a Schell-model source is of the form:

$$W_A(\mathbf{r}_1,\mathbf{r}_2,\nu)=[S_A(\mathbf{r}_1,\nu)]^{1/2}[S_A(\mathbf{r}_2,\nu)]^{1/2}\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu) \quad [51]$$

where $S_A(\mathbf{r},\nu)$ is the spectral density of the light at a typical point in the primary source or on the plane of a secondary source. Schell model sources do not assume low coherence, and therefore, can be applied to spatially stationary light fields of any state of coherence. The Schell-model of the form given in eqn [51] has been used to represent both three-dimensional primary sources and two-dimensional secondary sources.

Quasi-Homogeneous Sources

Useful models of partially coherent sources that are frequently encountered in nature or in the laboratory are the so-called quasi-homogeneous sources. These are an important sub-class of Schell-model sources. A Schell-model source is called quasi-homogeneous if the intensity of a Schell model source is essentially constant over any coherence area. Under these approximations, the cross-spectral density function for a quasi-homogeneous source is given by

$$\begin{aligned}W_A(\mathbf{r}_1,\mathbf{r}_2,\nu)&=S_A\left[\tfrac{1}{2}(\mathbf{r}_1+\mathbf{r}_2),\nu\right]\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu)\\&=S_A[\mathbf{r},\nu]\mu_A(\mathbf{r}',\nu)\end{aligned} \quad [52]$$

where $\mathbf{r}=(\mathbf{r}_1+\mathbf{r}_2)/2$, and $\mathbf{r}'=\mathbf{r}_2-\mathbf{r}_1$. The subscript A stands for either V or Q for the field variable or a source variable, respectively. It is clear that for a quasi-homogeneous source the spectral density $S_A(\mathbf{r},\nu)$ varies so slowly with position that it is approximately constant over distances across the sources that are of the order of the correlation length Λ, which is a measure of the effective width of $|\mu_A(\mathbf{r}',\nu)|$. Therefore, $S_A(\mathbf{r},\nu)$ is a slowly varying function of $\mathbf{r}$ (Figure 10b) and $|\mu_A(\mathbf{r}',\nu)|$ is a fast varying function of $\mathbf{r}'$ (Figure 10a). In addition, the linear dimensions of the source are large compared with the wavelength of light and with the correlation length Δ (Figure 10c).

Quasi-homogeneous sources are always spatially incoherent in the 'global' sense, because their linear dimensions are large compared with the correlation length. This model is very good for representing two-dimensional secondary sources with sufficiently low coherence such that the intensity does not vary over the coherence area on the input plane. It has also been applied to three-dimensional primary sources,

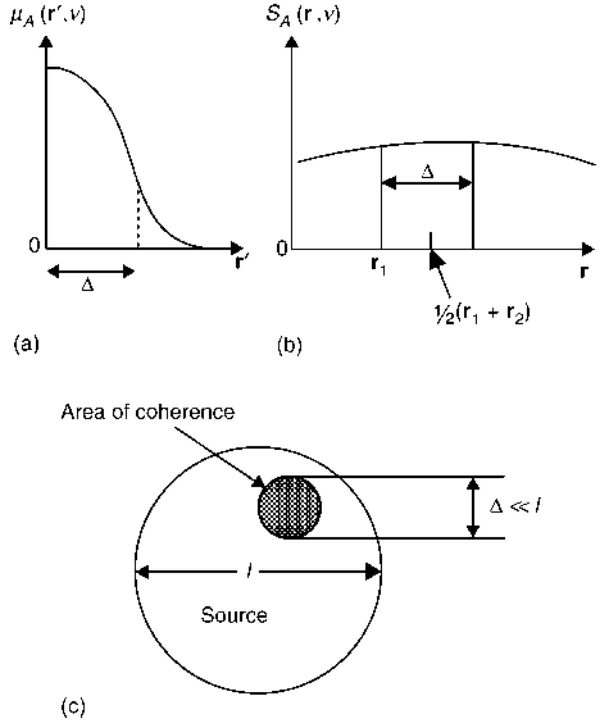

Figure 10 Concept of quasi-homogeneous sources.

three-dimensional scattering potentials, and two-dimensional primary and secondary sources.

Equivalence Theorems

The study of partially coherent sources led to the formulations of a number of equivalence theorems, which show that sources of any state of coherence can produce the same distribution of the radiant intensity as a fully spatially coherent laser source. These theorems provide conditions under which sources of different spatial distribution of spectral density and of different state of coherence will generate fields, which have the same radiant intensity. It has been shown, by taking examples of Gaussian–Schell model sources, that sources of completely different coherence properties and different spectral distributions across the source generate identical distribution of radiant intensity. Experimental verifications of the results of these theorems have also been carried out. For further details on this subject, the reader is referred to Mandel and Wolf (1995).

Correlation-Induced Spectral Changes

It was assumed that spectrum is an intrinsic property of light that does not change as the radiation propagates in free-space, until studies on partially coherent sources and radiations from them in 1980s, revealed that this was true only for specific type of sources. It was discovered on general grounds that the spectrum of light, which originates in an extended source, either a primary source or a secondary source, depends not only on the source spectrum but also on the spatial coherence properties of the source. It was also predicted theoretically by Wolf that the spectrum of light would, in general, be different from the spectrum of the source, and be different at different points in space on propagation in free space.

For a quasi-homogeneous planar secondary source defined by eqn [52], whose normalized spectrum is the same at each source point, one can write the spectral density as

$$S_0(\boldsymbol{\rho},\nu) = I_0(\boldsymbol{\rho})g_0(\nu) \quad \text{with} \quad \int_0^{\infty} g_0(\nu)\mathrm{d}\nu = 1 \qquad [53]$$

where $I_0(\boldsymbol{\rho})$ is the intensity of light at point $\boldsymbol{\rho}$ on the plane of the source, $g_0(\nu)$ is the normalized spectrum of the source and the subscript 0 refers to the quantities of the source plane. Using eqn [49], one can obtain an expression for the far-field spectrum due to this source as

$$S^{\infty}(r\mathbf{u},\nu) = \frac{k^2}{2\pi^2}\frac{\cos^2\theta}{r^2}\int_{\sigma}\int_{\sigma} I_0(\boldsymbol{\rho})g_0(\nu)\mu_0(\boldsymbol{\rho}',\nu) \times \mathrm{e}^{-ik\mathbf{u}_{\perp}\cdot(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)}\,\mathrm{d}^2\rho_1\,\mathrm{d}^2\rho_2 \qquad [54]$$

Noting that $\boldsymbol{\rho} = (\boldsymbol{\rho}_1 + \boldsymbol{\rho}_2)/2$ and $\boldsymbol{\rho}' = \boldsymbol{\rho}_2 - \boldsymbol{\rho}_1$, one can transform the variables of the integration and obtain after some manipulation:

$$S^{(\infty)}(r\mathbf{u},\nu) = \frac{k^2\cos^2\theta}{(2\pi r)^2}\tilde{I}_0\tilde{\mu}_0(k\mathbf{u}_{\perp},\nu)g_0(\nu) \qquad [55]$$

where

$$\tilde{\mu}_0(k\mathbf{u}_{\perp},\nu) = \int_{\sigma}\mu_0(\boldsymbol{\rho}',\nu)\mathrm{e}^{-ik\mathbf{u}_{\perp}\cdot\boldsymbol{\rho}'}\,\mathrm{d}^2\rho' \qquad [56]$$

and

$$\tilde{I}_0 = \int_{\sigma} I_0(\boldsymbol{\rho})\mathrm{d}^2\rho \qquad [57]$$

Equation [55] shows that the spectrum of the field in the far-zone depends on the coherence properties of the source through its spectral degree of coherence $\mu_0(\boldsymbol{\rho}',\nu)$ and on the normalized source spectrum $g_0(\nu)$.

Scaling Law

The reason why coherence-induced spectral changes were not observed until recently is that the usual thermal sources employed in laboratories or commonly encountered in nature have special coherence properties and the spectral degree of coherence has the function form:

$$\mu^{(0)}(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1,\nu)=f[k(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1)] \quad \text{with} \quad k=\frac{2\pi\nu}{c} \qquad [58]$$

which shows that the spectral degree of coherence depends only on the product of the frequency and space coordinates. This formula expresses the so-called scaling law, which was enunciated by Wolf. Commonly used sources satisfy this property. For example, the spectral degree of coherence of Lambertian sources and black-body sources can be shown to be

$$\mu_0(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1,\nu)=\frac{\sin(k|\boldsymbol{\rho}_2-\boldsymbol{\rho}_1|)}{k|\boldsymbol{\rho}_2-\boldsymbol{\rho}_1|} \qquad [59]$$

This expression evidently satisfies the scaling law. If the spectral degree of coherence does not satisfy the scaling law, the normalized far-field spectrum will, in general, vary in different directions in the far-zone and will differ from the source spectrum.

Spectral Changes in Young's Interference Experiment

Spectral changes in Young's interference experiment with broadband light are not as well understood as in experiments with quasi-monochromatic, probably because in such experiments no interference fringes are formed. However, if one were to analyze the spectrum of the light in the region of superposition, one would observe changes in the spectrum of light in the region of superposition in the form of a shift in the spectrum for narrowband spectrum and spectral modulation for broadband light. One can readily derive an expression for the spectrum of light in the region of superposition. Let $S^{(1)}(P,\nu)$ be the spectral density of the light at P which would be obtained if the small aperture at P_1 alone was open; $S^{(2)}(P,\nu)$ has a similar meaning if only the aperture at P_2 was open (Figure 4).

Let us assume, as is usually the case, that $S^{(2)}(P,\nu)\approx S^{(1)}(P,\nu)$ and let d be the distance between the two pinholes. Consider the spectral density at the point P, at distance x from the axis of symmetry in an observation plane located at distance of R from the plane containing the pinholes. Assuming that $x/R \ll 1$, one can make the approximation $R_2-R_1\approx xd/R$. The spectral interference law (eqn [38]) can then be written as

$$S(P,\nu)\approx 2S^{(1)}(P,\nu)\{1+|\mu(P_1,P_2,\nu)| \times\cos[\beta(P_1,P_2,\nu)+2\pi\nu xd/cR]\} \qquad [60]$$

where $\beta(P_1,P_2,\nu)$ denotes the phase of the spectral degree of coherence. Equation [60] implies the two results:

(i) at any fixed frequency ν, the spectral density varies sinusoidally with the distance x of the point from the axis, with the amplitude and the phase of the variation depending on the (generally complex) spectral degree of coherence $\mu(P_1,P_2,\nu)$; and

(ii) at any fixed point P in the observation plane the spectrum $S(P,\nu)$ will, in general, differ from the spectrum $S^{(1)}(P,\nu)$, the change also depending on the spectral degree of coherence $\mu(P_1,P_2,\nu)$ of the light at the two pinholes.

Experimental Confirmations

Experimental tests of the theoretical prediction of spectral invariance and noninvariance due to correlation of fluctuations across the source were performed just after the theoretical predictions. Figure 11 shows results of one such experiment in which spectrum changes in the Young's experiment were studied. Several other experiments also reported confirmation of the source correlation-dependent spectral changes. One of the important applications of these observations has been to explain the discrepancies in the maintenance of the spectroradiometric scales by national laboratories in different countries. These studies also have potential application in determining experimentally the spectral degree of coherence of partially coherent fields. The knowledge of spectral degree of coherence is often important in remote sensing, e.g., for determining angular diameters of stars.

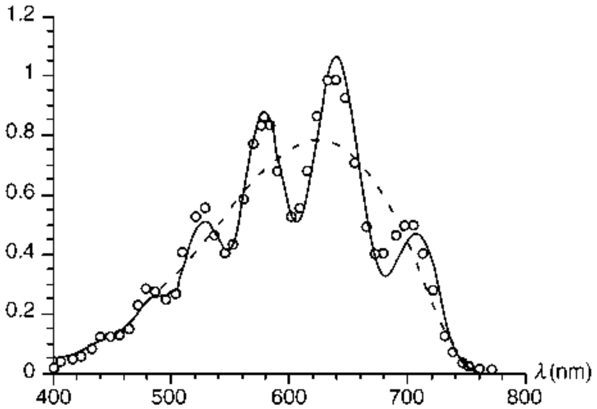

Figure 11 Correlation-induced changes in spectrum in Young's interference. Dashed line represents the spectrum when only one of the slits is open, the continuous curve shows the spectrum when both the slits are open and the circles are the measured values in the latter case. Reproduced with permission from Santarsiero M and Gori F (1992) Spectral changes in Young interference pattern. *Physics Letters* 167: 123–128.

Applications of Optical Coherence

Stellar Interferometry

The Michelson stellar interferometer, named after Albert Michelson, was used to determine the angular diameters of stars and also the intensity distribution across the star. The method was devised by Michelson without using any concept of coherence, although subsequently the full theory of the method was developed on the basis of propagation of correlations. A schematic of the experiment is shown in Figure 12. The interferometer is mounted in front of a telescope, a reflecting telescope in this case. The light from a star is reflected from mirrors M_1 and M_2 and is directed towards the primary mirror (or the objective lens) of the telescope. The two beams thus collected superpose in the focal plane F of telescope where an image crossed with fringes is formed. The outer mirrors M_1 and M_2 can be moved along the axis defined as $M_1M_3M_4M_2$ while the inner mirrors M_3 and M_4 remain fixed. The fringe spacing depends on the position of mirrors M_3 and M_4 and hence is fixed, while the visibility of the fringes depends on the separation of the mirrors M_1 and M_2 and hence, can be varied. Michelson showed that from the measurement of the variation of the visibility with the separation of the two mirrors, one could obtain information about the intensity distribution of the stars, which are rotationally symmetric. He also showed that if the stellar disk is circular and uniform, the visibility curve as a function of the separation d of the mirrors M_1 and M_2 will have zeros for certain values of d, and that the smallest of these d values for which zero occurs is given by $d_0 = 0.61\lambda_a/\alpha$, where α is the semi-angular diameter of the star and λ_a is the mean wavelength of the filtered quasi-monochromatic light from the star. Angular diameters of several stars down to 0.02 second of an arc were determined.

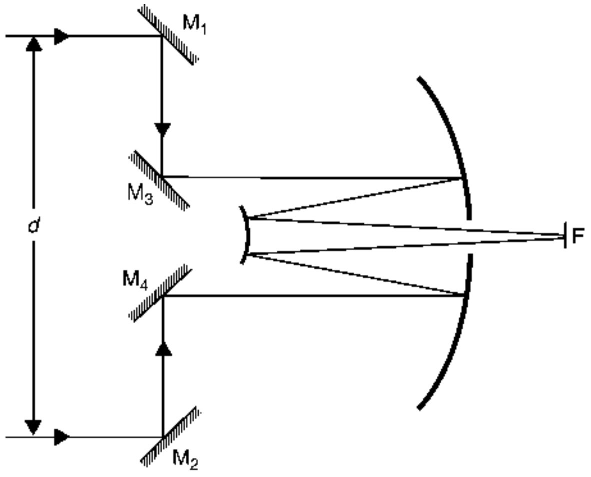

Figure 12 Schematic of the Michelson stellar interferometer.

From the standpoint of second-order coherence theory the principles of the method can be readily understood. The star is considered an incoherent source and according to the van Cittert–Zernike theorem, the light reaching the outer mirrors M_1 and M_2 of the interferometer will be partially coherent. This coherence would depend on the size of and the intensity distribution across the star. Let (x_1, y_1) and (x_2, y_2) be the coordinates of the positions of the mirrors M_1 and M_2, respectively, and (ξ, η) the coordinates of a point on the surface plane of the star which is assumed to be at a very large (astronomical) distance R from the mirrors. The complex degree of coherence at the mirrors would then be given by eqn [22] which can now be written as

$$\gamma(\Delta x, \Delta y, 0) = \frac{\int_\sigma I(u,v)\mathrm{e}^{-ik_a(u\Delta x + v\Delta y)}\,\mathrm{d}u\,\mathrm{d}v}{\int_\sigma I(u,v)\mathrm{d}u\,\mathrm{d}v} \qquad [61]$$

where $I(u,v)$ is the intensity distribution across the star disk σ as a function of the angular coordinates $u = \xi/R$, $v = \eta/R$, $\Delta x = x_1 - x_2$, $\Delta y = y_1 - y_2$, and $k_a = 2\pi/\lambda_a$, λ_a being the mean wavelength of the light from the star. Equation [61] shows that the equal-time ($\tau = 0$) complex degree of coherence of the light incident at the outer mirrors of the interferometer is the normalized Fourier-transform of the intensity distribution across the stellar disk. Further, eqn [15] shows that the visibility of the interference fringes is the absolute value of γ, if the intensity of the two interfering beams is equal, as in the present case. The phase of γ can be determined by the position of the intensity maxima of the fringe pattern (eqn [14]). If one is interested in determining only the angular size of the star and the star is assumed to be a circularly symmetric disk of angular diameter 2α and of uniform intensity [$I(u,v)$ is constant across the disk], then eqn [61] reduces to

$$\gamma(\Delta x,\Delta y) = \frac{2J_1(v)}{v}, \quad v = \frac{2\pi\alpha}{\lambda_a}d, \quad d = \sqrt{(\Delta x)^2 + (\Delta y)^2} \qquad [62]$$

The smallest separation of the mirrors for which the visibility γ vanishes corresponds to $v = 3.832$, i.e., $d_0 = 0.61\lambda_a/\alpha$, which is in agreement with Michelson's result.

Interference Spectroscopy

Another contribution of Michelson, which was subsequently identified as an application of the coherence theory, was the use of his interferometer (Figure 2) to determine the energy distribution in the spectral lines. The method he developed is capable to resolving spectral lines that are too narrow to be analyzed by the spectrometer. The visibility of the interference fringes depends on the energy distribution in the spectrum of the light and its measurement can give information about the spectral lines. In particular, if the energy distribution in a spectral line is symmetric about some frequency ν_0, its profile is simply the Fourier transform of the visibility variation as a function of the path difference between the two interfering beams. This method is the basis of interference spectroscopy or the Fourier transform spectroscopy.

Within the framework of second-order coherence theory, if the mean intensity of the two beams in a Michelson's interferometer is the same, then the visibility of the interference fringes in the observation plane is related to the complex degree of coherence of light at a point at the beamsplitter where the two beams superimpose (eqn [15]). The two quantities are related as $v(\tau) = |\gamma(\tau)|$ where $\gamma(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_1, \tau)$ is the complex degree of self-coherence at the point $\mathbf{r}_1$ on the beamsplitter. Following eqn [19], $\gamma(\tau)$ can be represented by a Fourier integral as

$$\gamma(\tau) = \int_0^\infty s(\nu)\exp(-\mathrm{i}2\pi\nu\tau)\mathrm{d}\nu \qquad [63]$$

where $s(\nu)$ is the normalized spectral density of the light defined as $s(\nu) = S(\nu)/\int_0^\infty S(\nu)\mathrm{d}\nu$ and $S(\nu) \equiv S(\mathbf{r}_1, \nu) = W(\mathbf{r}_1, \mathbf{r}_1, \nu)$ is the spectral density of the beam at the point $\mathbf{r}_1$. The method is usually applied to very narrow spectral lines for which one can assume that the peak that occurs at ν_0, and $\gamma(\tau)$ can be represented as

$$\begin{aligned} \gamma(\tau) &= \bar{\gamma}(\tau)\exp(-2\pi \mathrm{i}\nu_0\tau) \quad \text{with} \\ \bar{\gamma}(\tau) &= \int_{-\infty}^{\infty} \bar{s}(\mu)\exp(-\mathrm{i}2\pi\mu\tau)\mathrm{d}\mu \end{aligned} \qquad [64]$$

where $\bar{s}(\mu)$ is the shifted spectrum such that

$$\begin{aligned} \bar{s}(\mu) &= s(\nu_0 + \mu) & \mu &\geq -\nu_0 \\ &= 0 & \mu &< -\nu_0 \end{aligned}$$

From the above one readily gets

$$v(\tau) = |\gamma(\tau)| = \left| \int_{-\infty}^{\infty} \bar{s}(\mu)\, \exp(-\mathrm{i}2\pi\mu\tau)\mathrm{d}\mu \right| \qquad [65]$$

If the spectrum is symmetric about ν_0, then $v(\tau)$ would be an even function of τ and the Fourier inversion would give:

$$\bar{s}(\mu) = s(\nu_0 + \mu) = 2\int_0^\infty v(\tau)\cos(2\pi\mu\tau)\mathrm{d}\tau \qquad [66]$$

which can be used to calculate the spectral energy distribution for a symmetric spectrum about ν_0 from the visibility curve. However, for an asymmetric spectral distribution, the visibility and the phase of the complex degree of coherence must be determined as the Fourier transform of the shifted spectrum is no longer real everywhere.

Higher-Order Coherence

So far we have considered correlations of the fluctuating field variables at two space–time $(\mathbf{r}, t)$ points, as defined in eqn [10]. These are termed as second-order correlations. One can extend the concept of correlations to more than two space–time points, which will involve higher-order correlations. For example, one can define the space–time cross correlation function of order (M, N) of the random field $V(\mathbf{r}, t)$, represented by $\Gamma^{(M,N)}$, as an ensemble average of the product of the field $V(\mathbf{r}, t)$ values at N space–time points and $V^*(\mathbf{r}, t)$ at other M points. In this notation, the mutual coherence function as defined in eqn [10] would now be $\Gamma^{(1,1)}$. Among higher-order correlations, the one with $M = N = 2$, is of practical significance and is called the fourth-order correlation function, $\Gamma^{(2,2)}(\mathbf{r}_1, t_1, \mathbf{r}_2, t_2; \mathbf{r}_3, t_3, \mathbf{r}_4, t_4)$. The theory of Gaussian random variables tells us that any higher correlation can be written in terms of second-order correlations over all permutations of pairs of points. In addition, if we assume that $(\mathbf{r}_3, t_3) = (\mathbf{r}_1, t_1)$ and $(\mathbf{r}_4, t_4) = (\mathbf{r}_2, t_2)$, and that the field is stationary, then $\Gamma^{(2,2)}$ is called the intensity–intensity correlation and is given as

$$\begin{aligned} &\Gamma^{(2,2)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) \\ &= \langle V(\mathbf{r}_1, t_1)V(\mathbf{r}_2, t_2)V^*(\mathbf{r}_1, t_1)V^*(\mathbf{r}_2, t_2)\rangle \\ &= \langle I(\mathbf{r}_1, t_1)I(\mathbf{r}_2, t_2)\rangle \\ &= \langle I(\mathbf{r}_1, t_1)\rangle\langle I(\mathbf{r}_2, t_2)\rangle(1 + |\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)|^2) \end{aligned} \qquad [67]$$

where

$$\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) = \frac{\Gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)}{[\langle I(\mathbf{r}_1, t_1)\rangle]^{1/2}[\langle I(\mathbf{r}_2, t_2)\rangle]^{1/2}} \qquad [68]$$

We now define fluctuations in intensity at $(\mathbf{r}_j, t_j)$ as

$$\Delta I_j = I(\mathbf{r}_j, t_j) - \langle I(\mathbf{r}_j, t_j)\rangle$$

and then the correlation of intensity fluctuations becomes

$$\begin{aligned}\langle \Delta I_1 \Delta I_2 \rangle &= \langle I(\mathbf{r}_1,t_1) I(\mathbf{r}_2,t_2) \rangle - \langle I(\mathbf{r}_1,t_1)\rangle \langle I(\mathbf{r}_2,t_2)\rangle \\ &= \langle I(\mathbf{r}_1,t_1)\rangle \langle I(\mathbf{r}_2,t_2)\rangle |\gamma^{(1,1)}(\mathbf{r}_1,\mathbf{r}_2,t_2-t_1)|^2 \end{aligned} \quad [69]$$

where we have used eqn [67]. Equation [69] forms the basis for intensity interferometry.

Hanbury-Brown and Twiss Experiment

In this landmark experiment conducted both on the laboratory scale and astronomical scale, Hanbury-Brown and Twiss demonstrated the existence of intensity–intensity correlations in terms of the correlations in the photocurrents in the two detectors and thus measured the squared modulus of complex degree of coherence. In the laboratory experiment, the arc of a mercury lamp was focused onto a circular hole to produce a secondary source. The light from this source was then divided equally into two parts through a beamsplitter. Each part was, respectively, detected by two photomultiplier tubes, which had identical square apertures in front. One of the tubes could be translated normally to the direction of propagation of light and was so positioned that its image through the splitter could be made to coincide with the other tube. Thus, by suitable translation a measured separation d between the two square apertures could be introduced. The output currents from the photomultiplier tubes were taken by cables of equal length to a correlator. In the path of each cable a high-pass filter was inserted, so that only the current fluctuations could be transmitted to the correlator. Thus, the normalized correlations between the two current fluctuations:

$$C(d) = \frac{\langle \Delta J_1(t) \Delta J_2(t) \rangle}{\langle [\Delta J_1(t)]^2 \rangle^{1/2} \langle [\Delta J_2(t)]^2 \rangle^{1/2}} \quad [70]$$

as a function of detector separation d could be measured. Now when the detector response time is much larger than the time-scale of the fluctuations in intensity, then it can be shown that the correlations in the fluctuations of the photocurrent are proportional to the correlations of the fluctuations in intensity of the light being detected. Thus, we would have

$$C(d) \approx \delta |\gamma^{(1,1)}(\mathbf{r}_1,\mathbf{r}_2,0)|^2 \quad [71]$$

where δ is the average number of photocounts of the light of one polarization during the time-scale of the fluctuations (for general thermal sources, this is much less than one). Equation [71] represents the Hanbury-Brown–Twiss effect.

Stellar Intensity Interferometry

Michelson stellar interferometry can resolve stars which have angular sizes of the order of 0.01″, since for smaller stars, the separation between the primary mirrors runs into several meters and maintaining stability of mirrors such that the optical paths do not change, even by fraction of a wavelength, is extremely difficult. The atmospheric turbulence further adds to this problem and obtaining stable fringe pattern becomes next to impossible for very small stars. Hanbury-Brown and Twiss applied the intensity interferometry based on their photoelectric correlation technique for determining the angular sizes of such stars. Two separate parabolic mirrors collected light from the star and the output of the photodetectors placed at the focus of each mirror was sent to a correlator. The cable lengths were made unequal so as to compensate for the time difference of the light arrival at the two mirrors. The normalized correlation of the fluctuations of the photocurrents was determined as described above. This would give the variation of the modulus-squared degree of coherence as a function of the mirror separation d from which the angular size of the stars can be estimated. The advantage of the stellar intensity interferometer over the stellar (amplitude) interferometer is that the light need not interfere as in the latter, since the photodetectors are mounted directly at the focus of the primary mirrors of the telescope. Thus, the constraint on the large path difference between the two beams is removed and large values of d can now be used. Moreover, the atmosphere turbulence and the mirror movements have very small effect. Stellar angular diameters as small as 0.0004″ of arc with resolution of 0.00003″ could be measured by such interferometers.

See also

Coherence: Coherence and Imaging. **Coherent Lightwave Systems**. **Coherent Transients:** Coherent Transient Spectroscopy in Atomic and Molecular Vapours. **Coherent Control:** Applications in Semiconductors; Experimental; Theory. **Interferometry:** Overview. **Information Processing:** Coherent Analogue Optical Processors. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Beran MJ and Parrent GB (1964) *Theory of Partial Coherence*. Englewood Cliffs, NJ: Prentice-Hall.

Born M and Wolf E (1999) *Principles of Optics*. New York: Pergamon Press.

Carter WH (1996) Coherence theory. In: Bass M (ed.) *Hand Book of Optics*. New York: McGraw-Hill.

Davenport WB and Root WL (1960) *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill.

Goodman JW (1985) *Statistical Optics*. Chichester: Wiley.

Hanbury-Brown R and Twiss RQ (1957) Interferometry of intensity fluctuations in light: I basic theory: the correlation between photons in coherent beams of radiation. *Proceedings of the Royal Society* 242: 300–324.

Hanbury-Brown R and Twiss RQ (1957) Interferometry of intensity fluctuations in light: II an experimental test of the theory for partially coherent light. *Proceedings of the Royal Society* 243: 291–319.

Kandpal HC, Vaishya JS and Joshi KC (1994) Correlation-induced spectral shifts in optical measurements. *Optical Engineering* 33: 1996–2012.

Mandel L and Wolf E (1965) Coherence properties of optical fields. *Review of Modern Physics* 37: 231–287.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Marathay AS (1982) *Elements of Optical Coherence*. New York: John Wiley.

Perina J (1985) *Coherence of Light*. Boston, MA: M.A. Reidel.

Santarsiero M and Gori F (1992) Spectral changes in Young interference pattern. *Physics Letters* 167: 123–128.

Schell AC (1967) A technique for the determination of the radiation pattern of a partially coherent aperture. *IEEE Transactions on Antennas and Propagation* AP-15: 187.

Thompson BJ (1958) Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48: 95–97.

Thompson BJ and Wolf E (1957) Two beam interference with partially coherent light. *Journal of the Optical Society of America* 47: 895–902.

Wolf E and James DFV (1996) Correlation-induced spectral changes. *Report on Progress in Physics* 59: 771–818.

Coherence and Imaging

J van der Gracht, HoloSpex, Inc., Columbia, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical imaging systems are strongly affected by the coherence of the light that illuminates the object of interest. In many cases, the light is approximately coherent or incoherent. These approximations lead to simple mathematical models for the image formation process and allow straightforward analysis and design of such systems. When the light is partially coherent, the mathematical models are more complicated and system analysis and design is more difficult. Partially coherent illumination is often used in microscopy, machine vision, and optical lithography. The intent of this article is to provide the reader with a basic understanding of the effects of coherence on imaging. The information should enable the reader to recognize when coherence effects are present in an imaging system and give insight into when coherence can be modified to improve imaging performance. The material relies mostly on concepts drawn from the Fourier optics perspective of imaging and a rigorous coherence theory treatment is avoided. We encourage the reader to consult the Further Reading list at the end of this article, for more complete definitions of terms for coherence theory. A number of experimental results highlight the effects of spatial coherence of the illumination on image formation. Discussions of the role of coherence in such key applications are interspersed throughout the article.

Image Formation – Ideal and Optimal

An image is typically defined as the reproduction or likeness of the form of an object. An image that is indistinguishable from the original object is generally considered to be ideal. In a general context, the sound of a voice coming from a loudspeaker can be thought of as the image of the sound coming directly from the original speaker's mouth. In optical imaging, the ideal image replicates the light emanating from the object. Taken to the extreme, the ideal image replicates the light leaving the object in terms of intensity, wavelength, polarization, and even coherence. When the final image is viewed by the eye, the ideal image only needs to replicate the spatial distribution of the light leaving the object in terms of color and relative intensity at each point on the object. (In this article, intensity is defined as optical power per unit area (watts per meter squared).)

A general diagram of a direct view image formation system is shown in Figure 1. The condenser optics gathers light from a primary source and illuminates a transmissive object having a complex wave amplitude transmission of $O(x, y)$. The imaging optics produce an optical image that is viewed directly by the viewer. A growing number of image formation systems now include a solid-state image detector as shown in Figure 2. The raw intensity optical image is converted to an electronic signal that can be digitally processed and then displayed on a monitor. In this system, the spatial intensity distribution emanating from the monitor is the final image. The task of the optical system to the left of the detector is to gather spatial information about the light properties of the object. In fact, the information gathered by the detector includes information about the object, the illumination system, and the image formation optics. If the observer is only interested in the light transmission properties of the object, the effects of the illumination and image formation optics must be well understood. When the light illuminating the object is known to be coherent or incoherent, reasonably simple models for the overall image formation process can be used. More general, partially coherent illumination can produce optically formed images that differ greatly from the intensity pattern leaving the object. Seen from this perspective, partially coherent illumination is an undesirable property that creates nonideal images and complicates the image analysis.

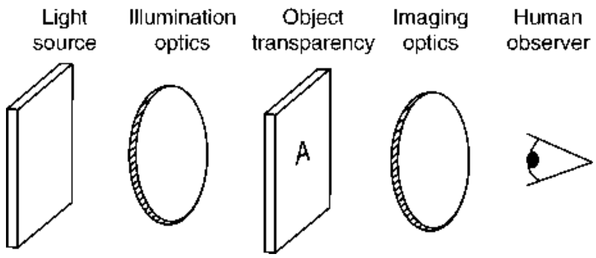

Figure 1 Direct view optical system. Imaging optics conveys the light from an illuminated object directly to the human viewer.

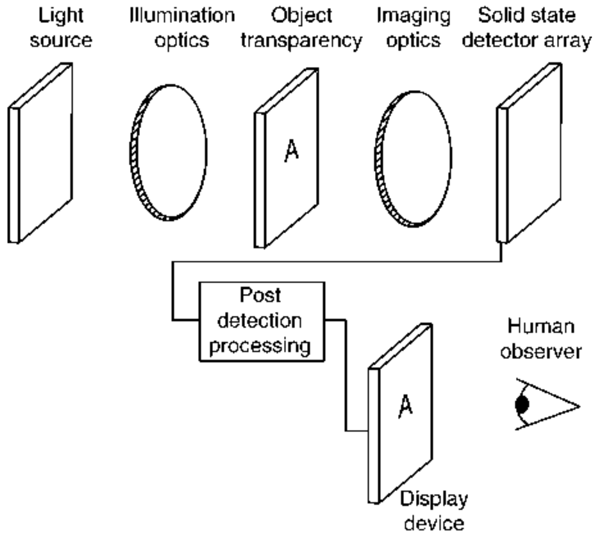

Figure 2 Indirect view optical system. The raw intensity optical image is detected electronically, processed and a final image is presented to the human observer on a display device.

The ideal image as described above is not necessarily the optimum image for a particular task. Consider the case of object recognition when the object has low optical contrast. Using the above definition, an ideal image would mimic the low contrast and render recognition difficult. An image formation system that alters the contrast to improve the recognition task would be better than the so-called ideal image. The image formation system that maximized the appropriate performance metric for a particular recognition task would be considered optimal. In optimizing the indirect imaging system of Figure 2, the designer can adjust the illumination, the imaging optics, and the post-detection processing. In fact, research microscopes often include an adjustment that modifies illumination coherence and often alters the image contrast. Darkfield and phase contrast imaging microscopes usually employ partially coherent illumination combined with pupil modification to view otherwise invisible objects. Seen from this perspective, partially coherent illumination is a desirable property that provides more degrees of freedom to the imaging system designer. Quite often, partially coherent imaging systems provide a compromise between the performance of coherent and incoherent systems.

Elementary Coherence Concepts

Most readers are somewhat familiar with the concept of temporal coherence. In Figure 3, a Michelson interferometer splits light from a point source into two paths and recombines the beams to form interference fringes. The presence of interference fringes indicates that the wave amplitude fluctuations of the two beams are highly correlated so the light adds in wave amplitude. If the optical path difference between the two paths can be made large without reducing the fringe contrast, the light is said to be

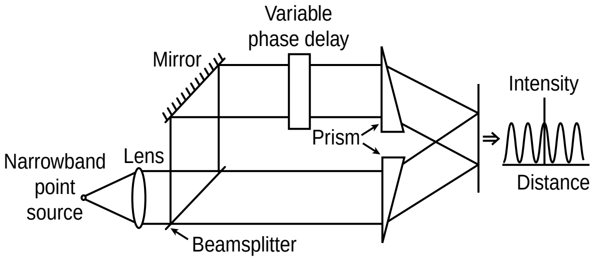

Figure 3 The presence of high contrast fringes in a Michelson interferometer indicates high temporal coherence.

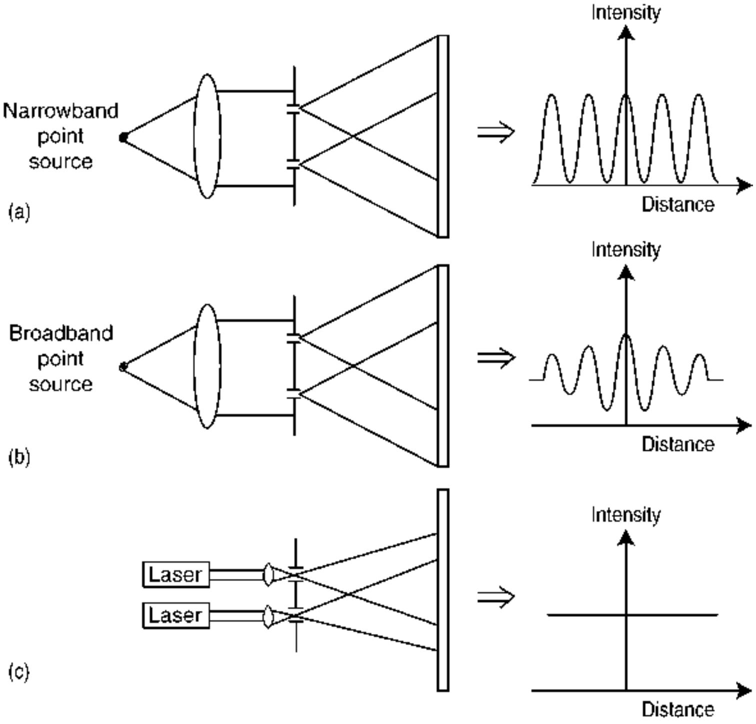

Figure 4 A Young's two pinhole interferometer produces: (a) a uniform high contrast fringe pattern for light that is highly spatially and temporally coherent; (b) high contrast fringes only near the axis for low temporal coherence and high spatial coherence light; and (c) no fringe pattern for high temporal coherence, low spatial coherence light.

highly temporally coherent. Light from very narrowband lasers can maintain coherence over very large optical path differences. Light from broadband light sources requires very small optical path differences to add in wave amplitude.

Spatial coherence is a measure of the ability of two separate points in a field to interfere. The Young's two pinhole experiment in Figure 4 measures the coherence between two points sampled by the pinhole mask. Only a one-dimensional pinhole mask is shown for simplicity. Figure 4a shows an expanded laser beam illuminating two spatially separated pinholes and recombining to form an intensity fringe pattern. The high contrast of the fringe indicates that the wave amplitude of the light from the two pinholes is highly correlated. Figure 4b shows a broadband point source expanded in a similar fashion. The fringe contrast is high near the axis because the optical path difference between the two beams is zero on the axis and relatively low near the axis. For points far from the axes, the fringe pattern disappears because the low temporal coherence from the broadband source results in a loss in correlation of the wave amplitudes.

A final Young's experiment example in Figure 4c shows that highly temporally coherent light can be spatially incoherent. In the figure, the light illuminating the two pinholes comes from two separate and highly temporally coherent lasers that are designed to have the same center frequency. Since the light from the lasers is not synchronized in phase, any fringes that might form for an instant will move rapidly and average to a uniform intensity pattern over the integration time of a typical detector. Since the fringe contrast is zero over a practical integration time, the light at the two pinholes is effectively spatially incoherent.

Two-point Imaging

In a typical partially coherent imaging experiment we need to know how light from two pinholes adds at the optical image plane, as shown in Figure 5. Diffraction and system aberrations cause the image of a single point to spread so that the images of two spatially separated object points overlap in the image plane. The wave amplitude image of a single pinhole is called the coherent point spread function (CPSF). Since the CPSF is compact, two CPSFs will only overlap when the corresponding object points are closely spaced. When the two points are sufficiently close, the relevant optical path differences will be small so full temporal coherence can be assumed. Spatial coherence will be the critical factor in determining how to add the responses to pairs of point images. We assume full temporal coherence in the subsequent analyses and the light is said to be *quasimonochromatic*.

Coherent Two-Point Imaging

Consider imaging a two-point object illuminated by a spatially coherent plane wave produced from a point source, as depicted in Figure 6a. In the figure, the imaging system is assumed to be space-invariant. This means that the spatial distribution of the CPSF is the same regardless of position of the input pinhole. The CPSF in the figure is broad enough such that the image plane point responses overlap for this particular pinhole spacing. Since the wave amplitudes from the two pinholes are correlated, the point responses add in wave amplitude, resulting in a two-point image intensity given by

$$I_{2\mathrm{coh}}(x) = I_0|h(x - x_1)|^2 + I_0|h(x - x_2)|^2 \qquad [1]$$

where $h(x)$ is the normalized CPSF and I_0 is a scaling factor that determines the absolute image intensity value. Since the CPSF has units of wave amplitude, the magnitude squaring operation accounts for the square law response of the image plane detector.

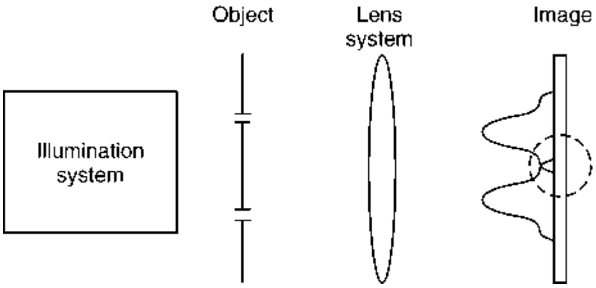

Figure 5 Imaging of two points. The object plane illumination coherence determines how the light adds in the region of overlap of the two image plane points.

More generally, the image intensity for an arbitrary object distribution can be found by breaking the object into a number of points and adding the CPSFs due to each point on the object. The resultant image plane intensity is the spatial convolution of the object amplitude transmittance with the CPSF and is given by

$$I_{\mathrm{coh}}(x) = I_0\left|\int O(\xi)h(\xi - x)\mathrm{d}\xi\right|^2 \qquad [2]$$

where $O(x)$ is the object amplitude transmittance.

Incoherent Two-Point Imaging

Two pinhole imaging with spatially incoherent light is shown in Figure 6b, where a spatially extended blackbody radiator is placed directly behind the pinholes. Once again the imaging optics are assumed to be space-invariant and have the same CPSF as the system in Figure 6a. Since the radiation is originating from two completely different physical points on the source, no correlation is expected between the wave amplitude of the light leaving the pinholes and the light is said to be spatially incoherent. The resultant image intensity corresponding to the two pinhole object with equal amplitude transmittance values is calculated by adding the individual intensity responses to give

$$I_{2\mathrm{inc}}(x) = I_0|h(x - x_1)|^2 + I_0|h(x - x_2)|^2 \qquad [3]$$

For more general object distributions, the image intensity is the convolution of the object intensity transmittance with the incoherent PSF, and is given by

$$I_{\mathrm{inc}}(x) = I_0\int |o(\xi)|^2|h(x - \xi)|^2\mathrm{d}\xi \qquad [4]$$

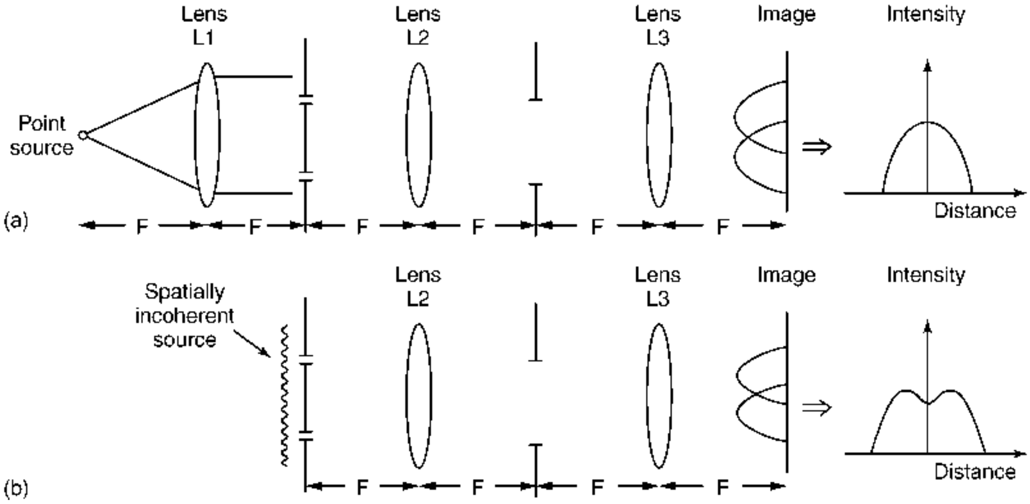

Figure 6 Imaging of two points for (a) spatially coherent illumination and (b) spatially incoherent imaging.

where the intensity transmittance is the squared magnitude of the amplitude transmittance and the incoherent PSF is the squared magnitude of the CPSF.

Partially Coherent Two-Point Imaging

Finally we consider two pinhole imaging illuminated by partially spatially coherent light. Now the two-point responses do not add simply in amplitude or intensity. Rather, the image intensity is given by the more general equation:

$$I_{2\mathrm{pc}}(x) = I_0\big[|h(x-x_1)^2 + |h(x-x_2)|^2 + 2\mathrm{Re}\{\mu(x_1;x_2)h(x-x_1)h^*(x-x_2)\}\big] \qquad [5]$$

where the $*$ denotes complex conjugations and $\mu(x_1;x_2)$ is the normalized form of the mutual intensity of the object illumination evaluated at the object points in question. The mutual intensity function is often denoted as $J_0(x_1;x_2)$ and is a measure of the cross correlation of the wave amplitude distributions leaving the two pinholes. Rather than providing a rigorous definition, we note that the magnitude of $J_0(x_1;x_2)$ corresponds to the fringe contrast that would be produced if the two object pinholes were placed at the input of a Young's interference experiment. The phase is related to the relative spatial shift of the fringe pattern. When the light is uncorrelated, $J_0(x_1;x_2) = 0$ and eqn [5] collapses to the incoherent limit of eqn [4]. When the light is coherent, $J_0(x_1;x_2) = 1$ and eqn [5] reduces to the coherent form of eqn [2].

Image formation for a general object distribution is given by the bilinear equation:

$$I_{\mathrm{pc}}(x) = I_0 \int o(\xi_1)o^*(\xi_2)h(x-\xi_1)h^*(x-\xi_2) \times J_o(\xi_1-\xi_2)\mathrm{d}\xi_1\,\mathrm{d}\xi_2 \qquad [6]$$

Note that, in general, $J_0(x_1;x_2)$ must be evaluated for all pairs of object points. Close examination of this equation reveals that just as a linear system can be evaluated by considering all possible single point responses, a bilinear system requires consideration of all possible pairs of points. This behavior is much more complicated and does not allow the application of the well-developed linear system theory.

Source Distribution and Object Illumination Coherence

According to eqn [6], the mutual intensity of all pairs of points at the object plane must be known, to calculate a partially coherent image. Consider the telecentric Kohler illumination imaging system shown in Figure 7. The primary source is considered to be spatially incoherent and illuminates the object after passing through lens L1 located one focal distance away from the source and one focal distance away from the object. Even though the primary source is spatially incoherent, the illumination at the object plane is partially coherent. The explicit mutual intensity function corresponding to the object plane illumination is given by applying the van Cittert Zernike theorem:

$$J_0(\Delta x) = \int S(\xi)\exp[\mathrm{j}2\pi\Delta x\xi/\lambda F]\mathrm{d}\xi \qquad [7]$$

where $S(x)$ is the intensity distribution of the spatially incoherent primary source, F is the focal length of the lens, λ is the illumination wavelength, and $\Delta x = x_1 - x_2$. The van Cittert Zernike theorem reflects a Fourier transform relationship between the source image intensity distribution and the mutual intensity at the object plane. When the source plane is effectively spatially incoherent, the object plane mutual intensity is only a function of separation distance. For a two-dimensional object, the mutual intensity needs to be characterized for all pairs of unique spacings in x and y. In Figure 7, the lens arrangement ensures that the object plane is located in the far field of the primary source plane. In fact, the van Cittert Zernike theorem applies more generally, even in the Fresnel propagation regime, as long as the standard paraxial approximation for optics is valid.

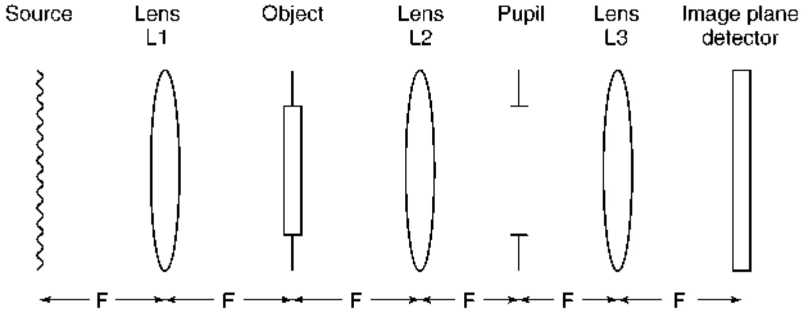

Figure 7 A telecentric Kohler illumination system with a spatially incoherent primary source imaged onto the pupil.

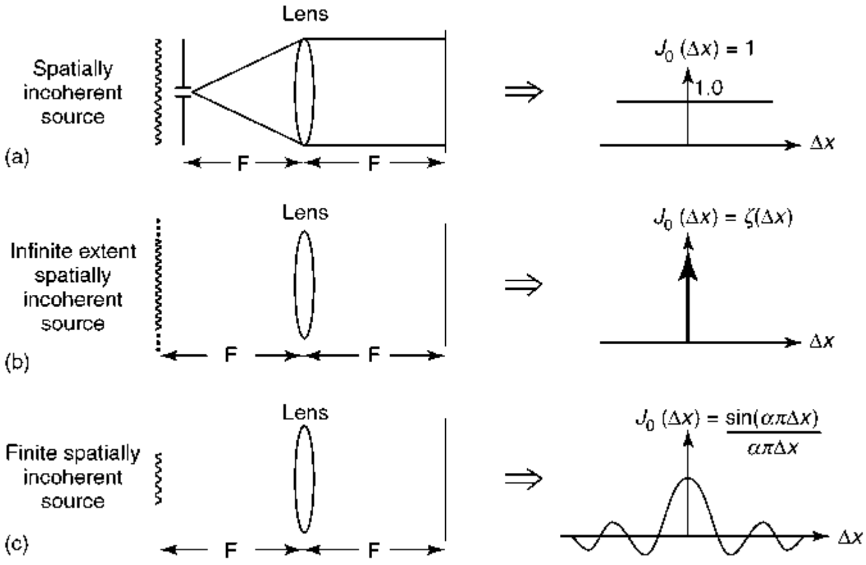
Lens
Spatially incoherent source
(a)
F
F
$J_0(\Delta x) = 1$
1.0
Δx
Lens
Infinite extent spatially incoherent source
(b)
F
F
$J_0(\Delta x) = \zeta(\Delta x)$
Δx
Lens
Finite spatially incoherent source
(c)
F
F
$J_0(\Delta x) = \frac{\sin(\alpha\pi\Delta x)}{\alpha\pi\Delta x}$
Δx

Figure 8 Light propages from the source plane to the object plane and produces (a) coherent illumination from a single source point; (b) incoherent illumination from an infinite extent source; and (c) partially coherent illumination from a finite extent primary source.

Figure 8 shows some simple examples of primary source distributions and the corresponding object plane mutual intensity function. Figure 8a assumes an infinitesimal point source and the corresponding mutual intensity is 1.0 for all possible pairs of object points. Figure 8b assumes an infinite extent primary source and results in a dirac delta function for the mutual intensity function. This means that there is no correlation between any two object points having a separation greater than zero so the object plane illumination is spatially incoherent. In Figure 5c, a finite extent uniform source gives a mutual intensity function of the form $\sin(\alpha\pi\Delta x)/(\alpha\Delta x)$. The finite-sized source corresponds to partially coherent imaging and shows that the response to pairs of points is affected by the spatial distribution of the source in a complicated way. Note that a large primary source corresponds to a large range of angular illumination at the object plane.

Varying the size of the source in the imaging system of Figure 8 will affect the spatial coherence of the illumination and hence the optical image intensity. Many textbook treatments discuss the imaging of two points separated by the Rayleigh resolution criterion which corresponds to the case where the first minimum of one point image coincides with the maximum of the adjacent point image. With a large source that provides effectively incoherent light, the two point image has a modest dip in intensity between the two points, as shown in Figure 9a. Fully coherent illumination of two points separated by the Rayleigh resolution produces a single large spot with no dip in intensity, as shown in Figure 9b. Varying the source size and hence the illumination spatial coherence produces a dip that is less than the incoherent intensity dip. This result is often used to suggest that coherent imaging gives poorer resolution than incoherent imaging. In fact, generalizations about two-point resolution can be misleading.

Recent developments in optical lithography have shown that coherence can be used to effectively increase two-point resolution beyond traditional diffraction limits. In Figure 9c, one of the pinholes in a coherent two-point imaging experiment has been modified with a phase shift corresponding to one half of a wavelength of the illuminating light. The two images add in wave amplitude and the phase shift creates a distinct null at the image plane and effectively enhances the two-point resolution. This approach is termed phase screen lithography and has been exploited to produce finer features in lithography by purposely introducing small phase shift masks at the object mask. In practice, the temporal and spatial coherence of the light is engineered to give sufficient coherence to take advantage of the two-point enhancement while maintaining sufficient incoherence to avoid speckle-like artifacts associated with coherent light.

Spatial Frequency Modeling of Imaging

Spatial frequency models of image formation are also useful in understanding how coherence affects image formation. A spatially coherent imaging system has a particularly simple spatial frequency model. In the

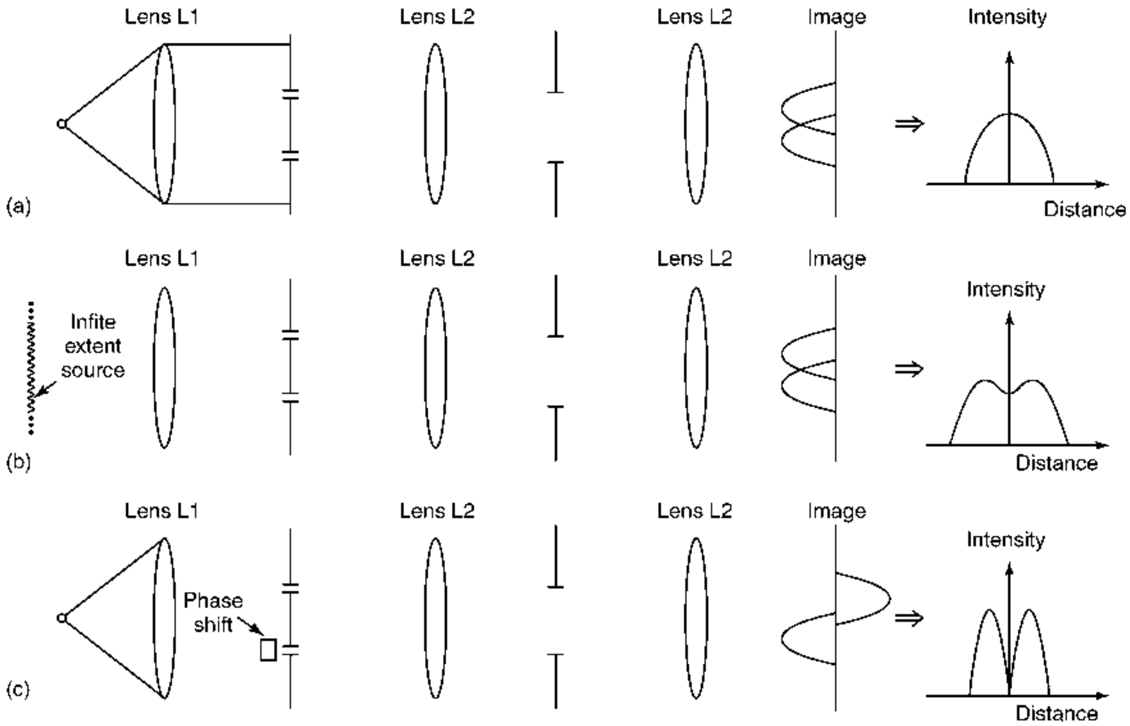

Figure 9 Comparison of coherent and two point imaging of two pinholes separated by the Rayleigh distance. (a) Coherent image cannot resolve the two points. (b) Incoherent image barely resolves the two points. (c) Coherent illumination with a phase shifting plate at one pinhole produces a null between the two image points.

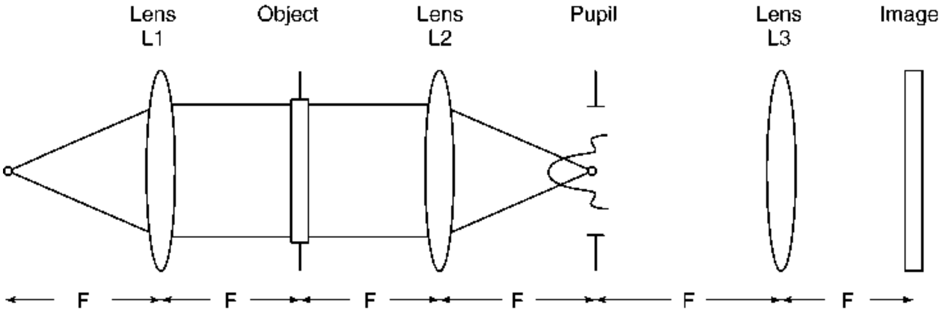

Figure 10 The object Fourier transform is filtered by the pupil function in a coherent imaging system.

spatially coherent imaging system shown in Figure 10, the on-axis plane wave illumination projects the spatial Fourier transform of the object, or Fraunhofer pattern, onto the pupil plane of the imaging system. The pupil acts as a frequency domain filter that can be modified to perform spatial frequency filtering. The complex transmittance of the pupil is the wave amplitude spatial frequency transfer function of the imaging system. It follows that the image plane wave amplitude frequency spectrum, $\tilde{U}_c(f_x)$, is given by

$$\tilde{U}_c(f_x) = \tilde{O}(f_x)H(f_x) \qquad [8]$$

where $\tilde{O}(f_x)$ is the spatial Fourier transform of the object amplitude transmittance function and $H(f_x)$ is proportional to the pupil plane amplitude transmittance. Equation [8] is the frequency domain version of the convolution representation given by eqn [2], but does not account for the squared magnitude response of the detector.

In the previous section, we learned that an infinite extent source is necessary to achieve fully incoherent imaging for the imaging system of Figure 7. As a thought experiment, one can start with a single point source on the axis and keep adding mutually incoherent source points to build up from a coherent imaging system to an incoherent imaging system. Since the individual source points are assumed to be incoherent with each other, the images from each point source can be added in intensity. In Figure 11,

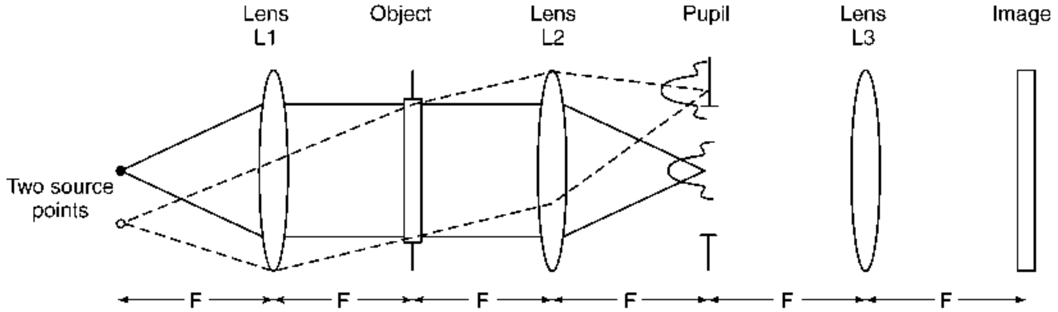

Figure 11 A second source point projects a second object Fourier transform onto the pupil plane. The images produced by the two source points add in intensity at the image plane.

only two point sources are shown. Each point source projects an object Fraunhofer pattern onto the pupil plane. The centered point source will result in the familiar coherent image. The off-axis point source projects a displaced object Fraunhofer pattern that is also filtered by the pupil plane before forming an image of the object. The image formed from this second point source can be modeled with the same coherent imaging model with a shifted pupil plane filter. Since the light from the two source points is uncorrelated, the final image is calculated by adding the intensities of the two coherently formed images.

The two source point model of Figure 11 can be generalized to an arbitrary number of source points. The final image is an intensity superposition of a number of coherent images. This suggests that partially coherent imaging systems behave as a number of redundant coherent imaging systems, each having a slightly different amplitude spatial frequency transfer function due to the relative shift of the pupil filter with respect to the object Fraunhofer pattern. As the number of point sources is increased to infinity, the primary source approaches an infinite extent spatially incoherent source. In practice, the source need not be infinite. When the source is large enough to effectively produce linear-in-intensity imaging, the imaging system is effectively spatially incoherent. The corresponding image intensity in the spatial frequency domain, $\tilde{I}_{\text{inc}}(f_x)$, is given by

$$\tilde{I}_{\text{inc}}(f_x) = \tilde{I}_{\text{obj}}(f_x)\text{OTF}(f_x) \qquad [9]$$

where $\tilde{I}_{\text{obj}}(f_x)$ is the spatial Fourier transform of the object intensity transmittance and $\text{OTF}(f_x)$ is the incoherent optical transfer function which is proportional to the spatial autocorrelation of the pupil function.

Many texts include detailed discussions comparing the coherent transfer function and the OTF and attempt to make comparisons about the relative performance of coherent and incoherent systems. These comparisons are often misleading since one transfer function describes the wave amplitude spatial frequency transfer function and the other describes the intensity spatial frequency transfer function. Here we note that incoherent imaging systems do indeed allow higher object amplitude spatial frequencies to participate in the image formation process. This argument is often used to support the claim that incoherent systems have higher resolution. However, both systems have the same image intensity spatial frequency cutoff. Furthermore, the nature of the coherent transfer function tends to produce high contrast images that are typically interpreted as higher-resolution images than their incoherent counterparts. Perhaps the real conclusion is that the term resolution is not well defined and direct comparisons between coherent and incoherent imaging must be treated carefully.

The frequency domain treatment for partially coherent imaging of two-dimensional objects involves a four-dimensional spatial frequency transfer function that is sometimes called the bilinear transfer function or the transmission cross-coefficient model. This model describes how constituent object wave amplitude spatial frequencies interact to form image plane intensity frequencies. The utility of this approach to analysis is limited for someone new to the field, but is often used in numerical simulations of partially coherent imaging systems used in optical lithography.

Experimental Examples of Important Coherence Imaging Phenomena

Perhaps the best way to gain an understanding of coherence phenomena in imaging is to examine experimental results. In the following section we use experimental data to see how coherence affects edge response, noise immunity, and depth of field. Several experimental configurations were used to collect the image data, but all of the systems can be represented generically by the Kohler illumination system shown in Figure 12. Kohler illumination is often employed in many projection illumination systems. In order to obtain Kohler illumination, the primary spatially incoherent source is imaged onto the pupil plane of the imaging portion of the system and the object is placed at the pupil plane of the condenser optics.

Primary Source Generation

Figure 13 shows a highly coherent illumination system produced by focusing a laser beam to a point and imaging the focused spot onto the pupil of the imaging system to produce spatially coherent illumination. The use of a laser produces highly temporally coherent light.

Two methods were used to create an extended spatially incoherent primary source with control over the spatial intensity distribution. Figure 14a shows a collimated laser beam with a 633 nm center wavelength illuminating a rotating diffuser. A photographically produced mask defines the spatial shape of the primary source. The laser provides highly temporally coherent light and the diffuser destroys the spatial coherence of the light. Consider the thought experiment of two pinholes placed immediately to the right of the moving diffuser. Without the diffuser, the two wave amplitudes would be highly correlated. Assuming that the diffuser can be modeled as a spatially random phase plate, a fixed diffuser would only introduce a fixed phase difference between the amplitudes leaving the two pinholes and would not destroy the coherence. When the diffuser is rotated, the light from each pinhole encounters a different phase modulation that is changing over time. This random modulation destroys the effective correlation between the wave amplitude of the light leaving the two pinholes provided that the rotation speed is sufficiently fast.

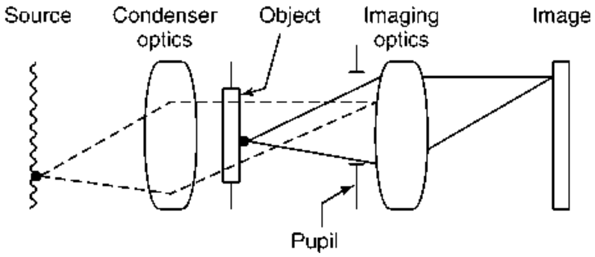

Figure 12 General representation of the Kohler illumination imaging systems used in the experimental result section. The primary incoherent source is imaged onto the imaging system pupil plane and the object resides in the pupil of the condenser optics.

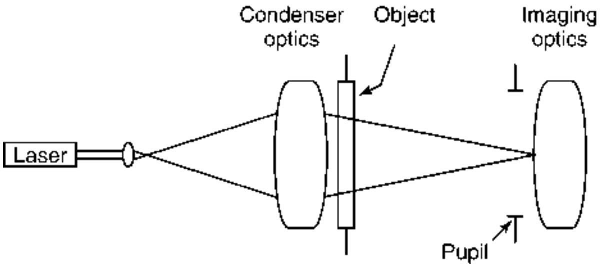

Figure 13 Highly spatially and temporally coherent illumination produced by imaging a focused laser beam into the imaging system pupil.

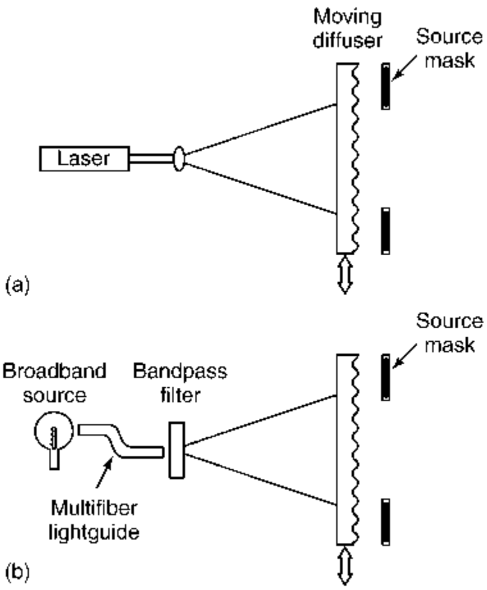

Figure 14 Two methods for generating a spatially incoherent primary source. (a) An expanded laser beam passes through a moving diffuser followed by a mask to define the extent and shape of the source. (b) A broadband source exits a multifiber lightguide and propagates to a moving diffuser followed by a source mask.

The moving diffuser method is light inefficient but is a practical way of exploring coherence in imaging in a laboratory environment. The choice of the diffuser is critical. The diffuser should spread light out uniformly over an angular subtense that overfills the object of interest. Many commercially available diffusers tend to pass too much light in the straight through direction. Engineered diffusers can be purchased to produce an optimally diffused beam. When quick and inexpensive results are required, thin plastic sheets used in day-to-day packaging often serve as excellent diffusers. When the diffusion angles are not high enough, a number of these plastic sheets can be layered on top of each other to achieve the appropriate angular spread.

The second method for producing a spatially incoherent source is shown in Figure 14b. Broadband light is delivered through a large multifiber lightguide and illuminates a moving diffuser. The main purpose of the diffuser is to ensure that the entire object is illuminated uniformly. The low temporal coherence of the source and the long propagation distances are usually enough to destroy any spatial coherence at the plane of the diffuser. The rotation further ensures that no residual spatial coherence exists at the primary source plane. A chromatic filter centered around 600 nm, with a spectral width of approximately 100 nm, is shown in the light path. The filter was used to minimize the effects of chromatic aberrations in the system. The wide spectral width certainly qualifies as a broadband source in relation to a laser. High-contrast photographically defined apertures determine the spatial intensity distribution at the plane of the primary source. In the following results, the primary source distributions were restricted to circular sources of varying sizes as well as annular sources. These shapes are representative of most of the sources employed in microscopy, machine vision, and optical lithography.

Noise Immunity

Coherent imaging systems are notorious for introducing speckle-like noise artifacts at the image. Dust and optically rough surfaces within an optical system result in a complicated textured image plane intensity distribution that is often called speckle noise. We refer to such effects as coherent artifact noise. Figure 15a shows the image of a standard binary target as imaged by a benchtop imaging system, with the highly coherent illumination method shown in Figure 13. Some care was taken to clean individual lenses and optical surfaces within the system and no dust was purposely introduced. The image is corrupted by a complicated texture that is due in

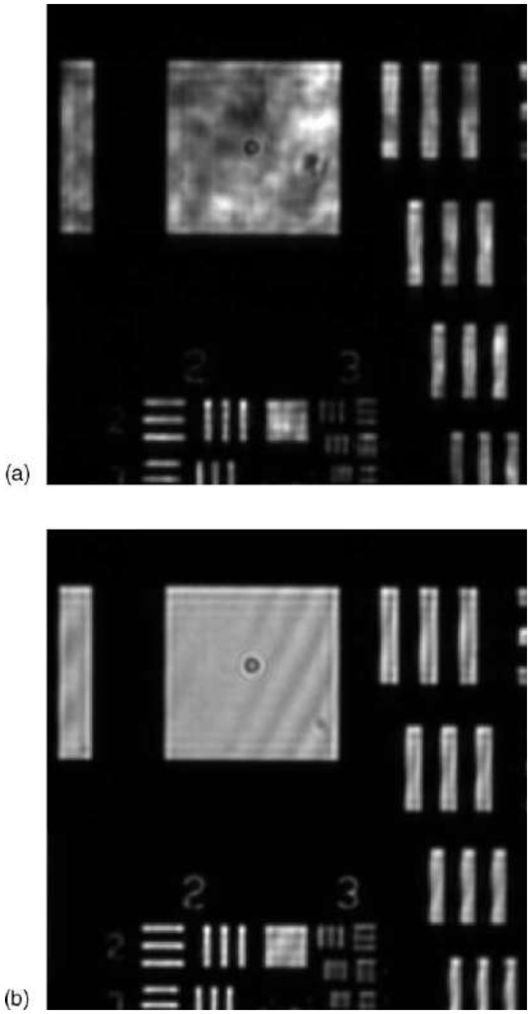

Figure 15 High temporal coherence imaging with disk sources corresponding to (a) extremely high spatial coherence ($K = 0.05$) and (b) slightly reduced but high spatial coherence ($K = 0.1$).

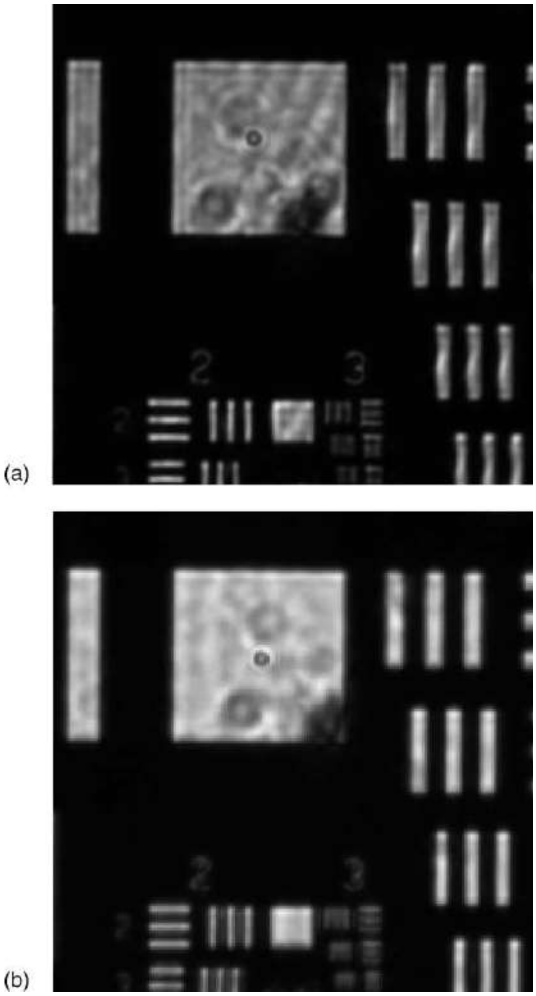

Figure 16 High spatial coherence disk illumination ($K = 0.1$) imaging through a dust covered surface with (a) narrowband laser illumination and (b) with broadband illumination.

part to imperfections in the laser beam itself and in part to unwanted dust and optically rough surfaces. Certainly, better imaging performance can be obtained with more attention to surface cleanliness and laser beam filtering, but the result shows that it can be difficult to produce high-quality coherent images in a laboratory setting.

Lower temporal coherence and lower spatial coherence will reduce the effect of speckle noise. The image of Figure 15b was obtained with laser illumination according to Figure 14a with a source mask corresponding to a source to pupil diameter ratio of $K = 0.1$. The object plane illumination is still highly temporally coherent and the spatial coherence has been reduced but is still very high. The artifact noise has been nearly eliminated by the modest reduction of spatial coherence. The presence of diagonal fringes in some regions is a result of multiple reflections produced by the cover glass in front of the CCD detector. The reflections produce a weak secondary image that is slightly displaced from the main image and the high temporal coherence allows the two images to interfere.

The noise performance was intentionally perturbed in the image of Figure 16a by inserting a slide with a modest sprinkling of dust at an optical surface in between the object and the pupil plane. The illumination conditions are the same as for the image of Figure 15b. The introduction of the dust has further degraded the image. The image of Figure 16b maintains the same spatial coherence ($K = 0.1$) but employs the broadband source. The lower temporal coherence eliminates the unwanted diagonal fringes but the speckle noise produced by the dust pattern does not improve significantly relative to the laser illumination $K = 0.1$ system. Figures 17a–c show that increasing the source size (and hence the range of angular illumination) reduces

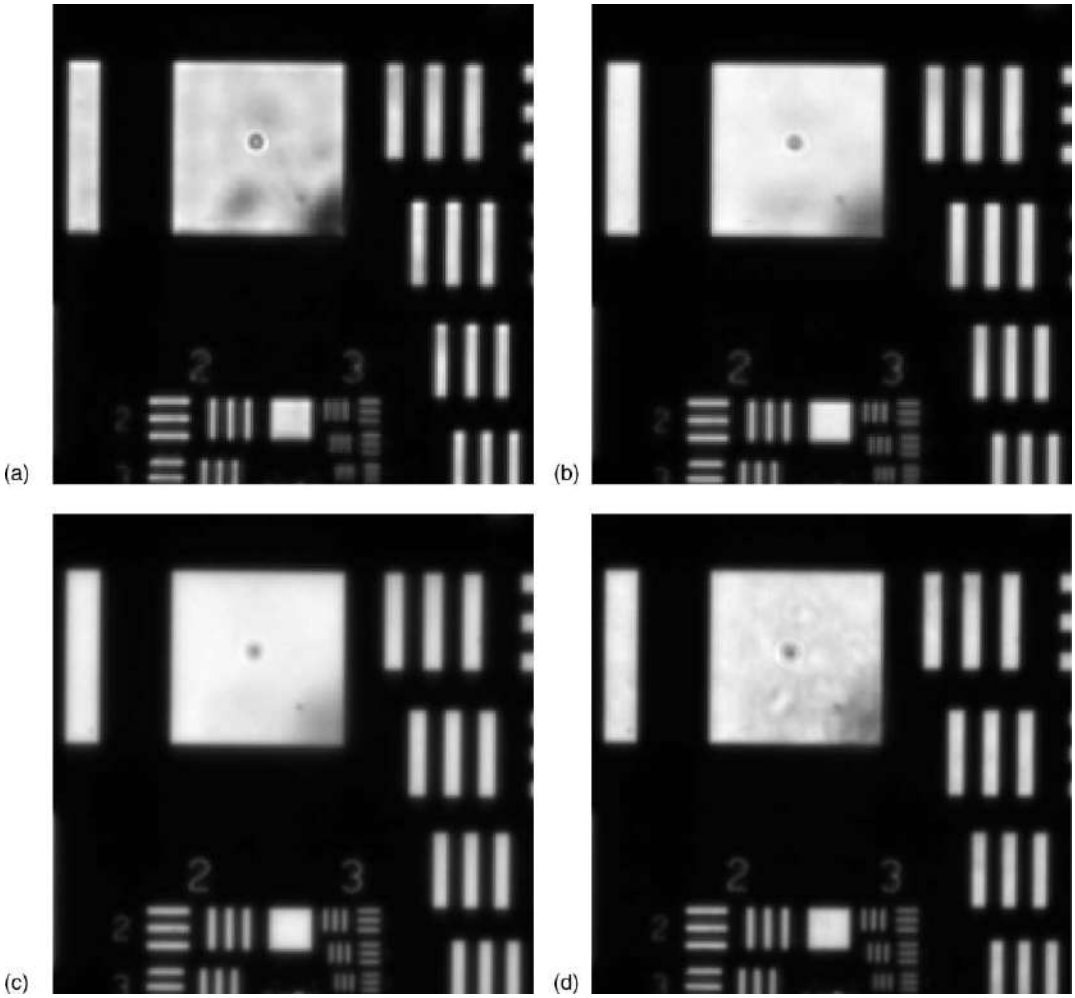

Figure 17 Low temporal coherence imaging through a dust-covered surface with various source sizes and shapes: (a) disk source with $K = 0.3$; (b) disk source with $K = 0.7$; (c) disk source with $K = 2.0$; and (d) a thin annular source with an outer diameter corresponding to $K = 0.5$ and inner diameter corresponding to $K = 0.45$.

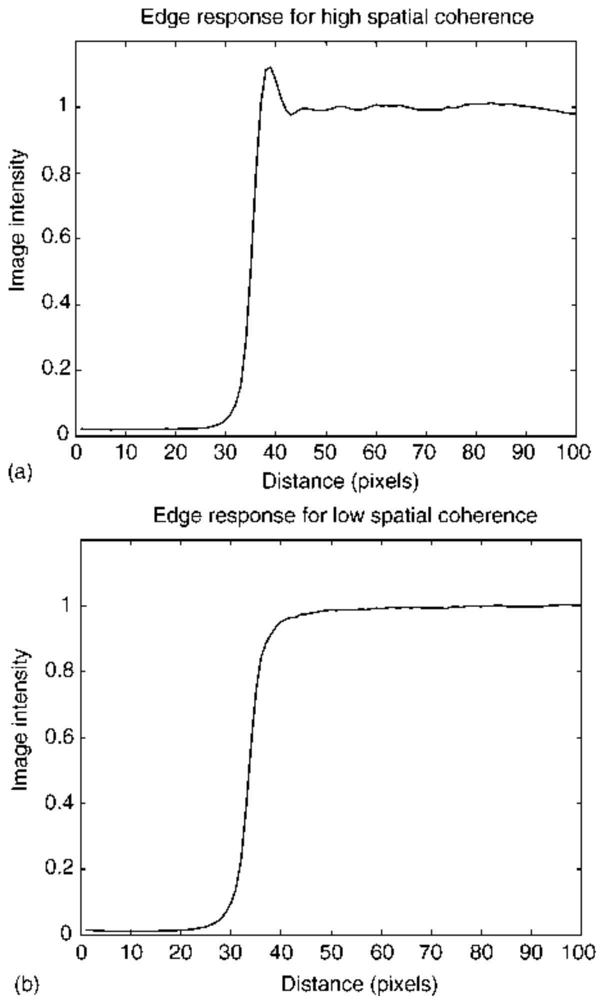

Figure 18 Slices of edge intensity images of high contrast edges taken by a laboratory system with low temporal coherence and (a) high spatial coherence (disk source with $K = 0.1$) and (b) effectively incoherent (disk source with $K = 2$).

the spatial coherence and virtually eliminates the unwanted speckle pattern. Finally, Figure 17d shows that an annular source can also provide some noise immunity. The amount of noise immunity is related to the total area of the source rather than the outer diameter of annular source. Seen from the spatial frequency model, each source point produces a noise-corrupted coherent image. Since the effect of the noise is different for each coherent image, the image plane noise averages out as the images add incoherently. This perspective suggests that an extended source provides redundancy in the transfer of object information to the image plane.

Edge Response

Spatial coherence has a strong impact on the images of edges. The sharp spatial frequency cutoff of spatially coherent imaging systems creates oscillations at the images of an edge. Figure 18 shows slices of experimentally gathered edge imagery as a function of the ratio of source diameter to pupil diameter. Higher coherence systems produce sharper edges, but tend to have overshoots. The sharper edges contribute to the sense that high-coherence systems produce higher-resolution images. One advantage to an incoherent imaging system is that the exact location of the edge corresponds to the 50% intensity location. The exact location of an edge in a partially coherent imaging system is not as easily determined. The presence or lack of edge ringing can be used to assess whether a given imaging system can be modeled as spatially incoherent or partially coherent.

Depth of Field

Coherent imaging systems exhibit an apparent increase in depth-of-field compared to spatially incoherent systems. Figure 19 shows spatially incoherent imagery for four different focus positions. Figure 20 shows imagery with the same amount of defocus produced with highly spatially coherent light. Finally, Figure 21 gives an example of how defocused imagery depends on the illumination coherence. The images of a spoke target were all gathered with a fixed amount of defocus and the source size was varied to control the illumination coherence. The images of Figure 21 differ greatly, even though the CPSF was the same for all the cases.

Digital Post-detection Processing and Partial Coherence

The model of Figure 2 suggests that post-detection image processing can be considered as part of the image formation system. Such a general view can result in imaging that might be otherwise unobtainable by classical means. In fact, microscopists routinely use complicated deblurring methods to reconstruct out-of-focus imagery and build up three-dimensional images from multiple image slices.

The coherence of the illumination should be considered when undertaking such image restoration. Rigorous restoration of partially coherent imagery is computationally intensive and requires precise knowledge of the coherence. In practice, a linear-in-intensity model for the image formation is almost always used in developing image restoration algorithms. Even nonlinear restoration algorithms have built-in linear assumptions about the image formation models which imply spatially incoherent illumination.

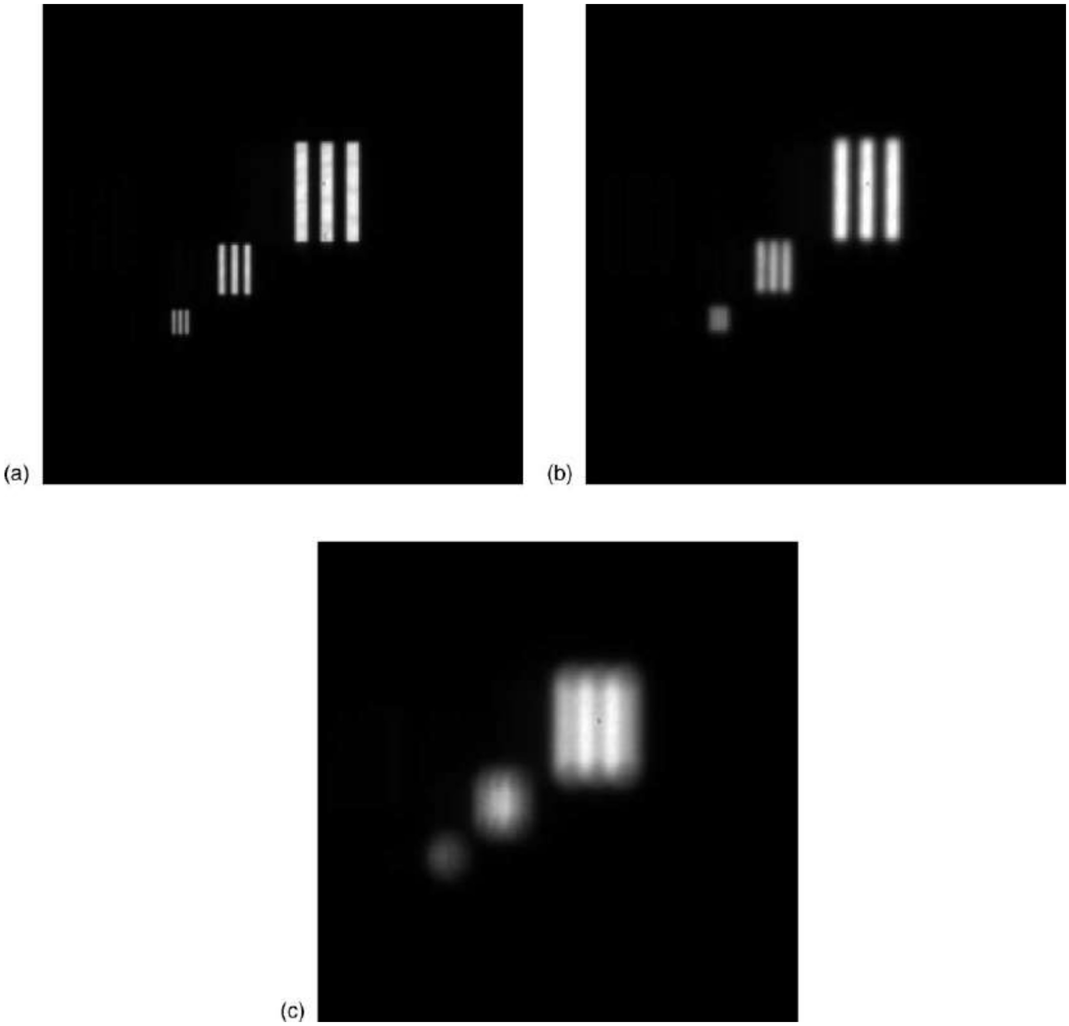

Figure 19 Spatially incoherent imaging for (a) best focus, (b) moderate misfocus, and (c) high misfocus.

The images of Figures 22a and b are digitally restored versions of the images of Figure 19b and Figure 20b. The point spread function was directly measured and used to create a linear restoration filter that presumed spatially incoherent illumination. The restored image in Figure 22a is faithful since the assumption of spatially incoherent light was reasonable. The restored image of Figure 22b suffers from a loss in fidelity since the actual illumination coherence was relatively high. The image is visually pleasing and correctly conveys the presence of three bar targets. However, the width of the bars is not faithful to the original target which had spacings equal to the widths of the bars.

As discussed earlier, the ideal image is not always the optimal image for a given task. In general, restoration of partially coherent imaging with an implicit spatially incoherent imaging model will produce visually pleasing images that are not necessarily faithful to the object plane intensity. While they are not faithful, they often preserve and even enhance edge information and the overall image may appear sharper than the incoherent image. When end task performance is improved, the image may be considered to be more optimal than the ideal image. It is important to keep in mind that some of the spatial information may be misleading and a more complete understanding of the coherence may be needed for precise image plane measurements. This warning is relevant in microscopy where the user is often encouraged to increase image contrast by reducing illumination spatial coherence. Annular sources can produce highly complicated spatial coherence functions that will strongly impact the restoration of such images.

Summary and Discussion

The coherence of the illumination at the object plane is important in understanding image formation.

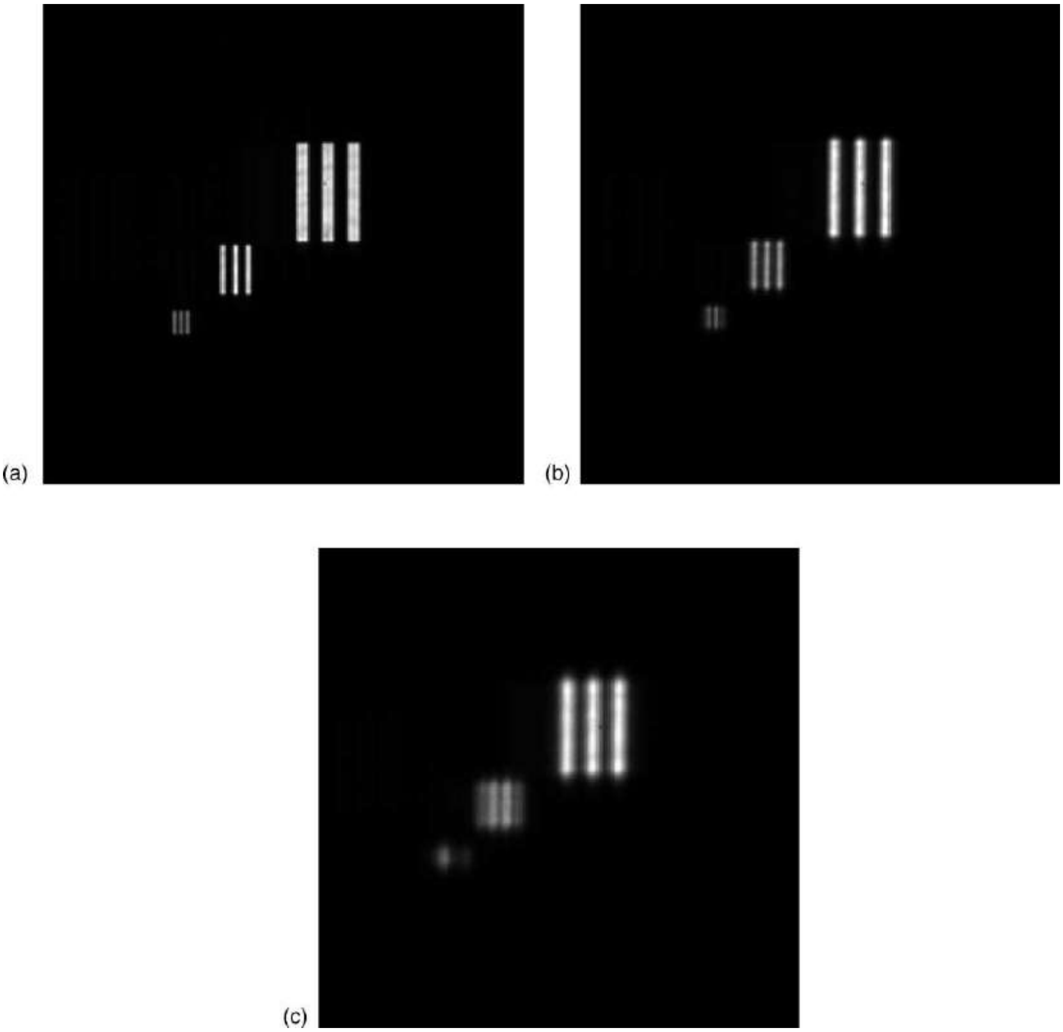

Figure 20 High spatial coherence disk source ($K = 0.2$) imaging for (a) best focus, (b) moderate misfocus, and (c) high misfocus.

Highly coherent imaging systems produce high-contrast images with high depth of field and provide the opportunity for sophisticated manipulation of the image with frequency plane filtering. Darkfield imaging and phase contrast imaging are examples of frequency plane filtering. Unfortunately, coherent systems are sensitive to optical noise and are generally avoided in practical system design. Relaxing the temporal coherence of the illumination can provide some improvement, but reducing the spatial coherence is more powerful in combating noise artifacts. Research microscopes, machine vision systems, and optical lithography systems are the most prominent examples of partially coherent imaging systems. These systems typically employ adjustable spatially incoherent extended sources in the shapes of disks or annuli. The exact shape and size of the primary source, shapes the angular extent of the illumination at the object plane and determines the spatial coherence at the object plane. Spatial coherence effects can be significant, even for broadband light. Control over the object plane spatial coherence allows the designer to find a tradeoff between the various strengths and weaknesses of coherent and incoherent imaging systems.

As more imaging systems employ post-detection processing, there is an opportunity to design fundamentally different systems that effectively split the image formation process into a physical portion and a post-detection portion. The simple example of image deblurring cited in this article shows that object plane coherence can affect the nature of the digitally restored image. The final image can be best understood when the illumination coherence effects are well understood. The spatially incoherent case results in the most straightforward model for image interpretation, but is not always the best choice since the coherence can often be manipulated to increase the contrast and hence the amount of useful information in the raw image. A completely general approach to

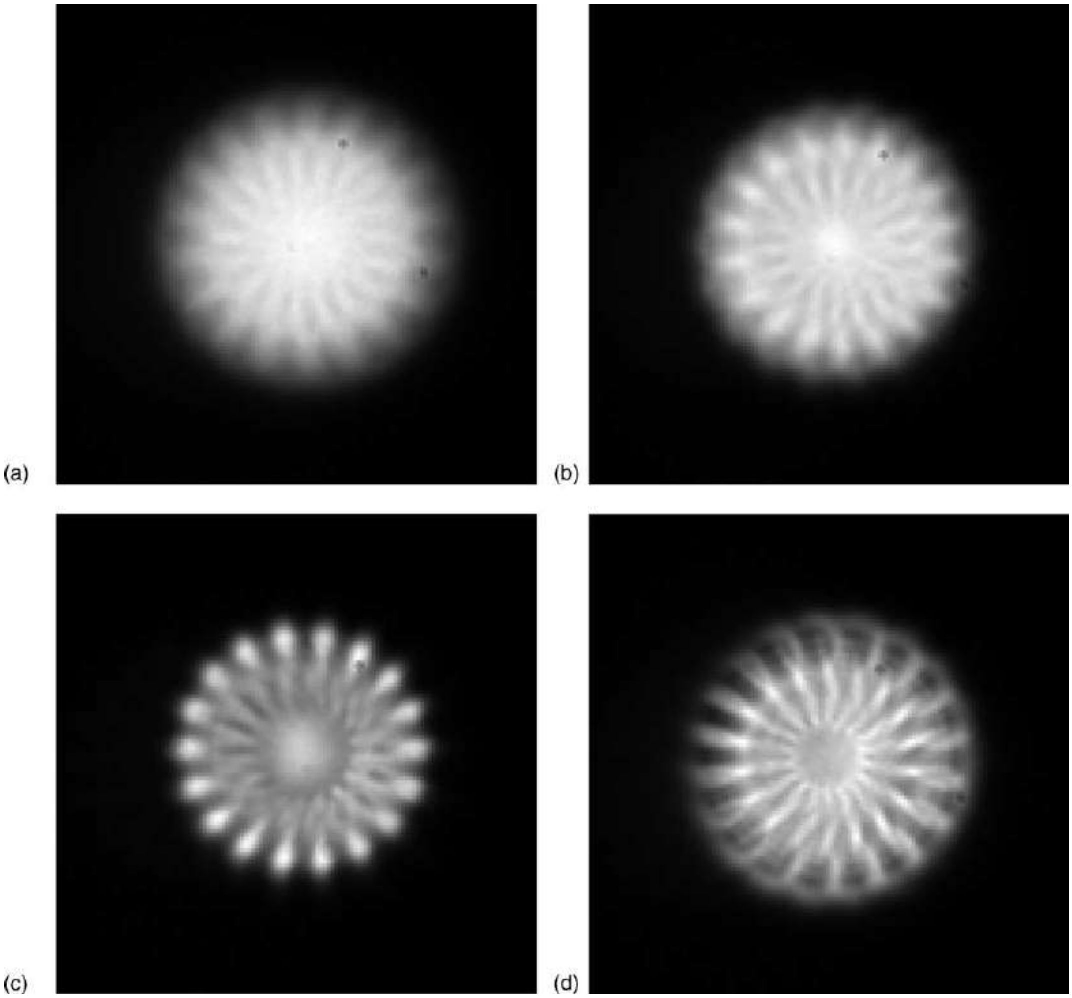

Figure 21 Imaging with a fixed amount of misfocus and varying object spatial coherence produced by: (a) highly incoherent disk source illumination ($K = 2$); (b) moderate coherence disk source ($K = 0.5$); (c) high coherence disk source ($K = 0.2$); and (d) annular source with outer diameter corresponding to $K = 0.5$ and inner diameter corresponding to $K = 0.45$.

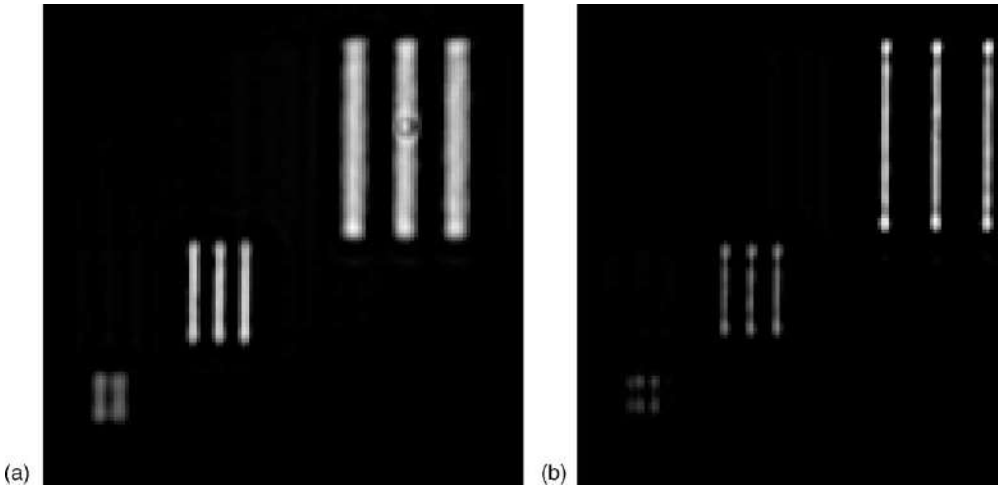

Figure 22 Digital restoration of blurred imagery with an inherent assumption of linear-in-intensity imaging for (a) highly spatially incoherent imaging ($K = 2$) and (b) high spatial coherence ($K = 0.2$).

imaging would treat the coherence of the source, the imaging optics, and the post-detection restoration all as free variables that can be manipulated to produce the optimal imaging system for a given task.

See also

Coherent Lightwave Systems. **Information Processing:** Coherent Analogue Optical Processors. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Bartelt H, Case SK and Hauck R (1982) Incoherent Optical Processing. In: Stark H (ed.) *Applications of Optical Fourier Transforms*, pp. 499–535. London: Academic Press.

Born M and Wolf E (1985) *Principles of Optics*, 6th edn. (corrected), pp. 491–554. Oxford, UK: Pergamon Press.

Bracewell RN (1978) *The Fourier Transform and its Applications*. New York: McGraw Hill.

Gonzales RC and Woods EW (1992) *Digital Image Processing*, 3rd edn., pp. 253–304. Reading, MA: Addison-Wesley.

Goodman JW (1985) *Statistical Optics*, 1st edn. New York: John Wiley & Sons.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw Hill.

Hecht E and Zajac A (1979) *Optics*, 1st edn., pp. 397–435, Reading, MA: Addison-Wesley.

Inoue S and Spring KR (1997) *Video Microscopy*, 2nd edn. New York: Plenum.

Reynolds GO, DeVelis JB, Parrent GB and Thompson JB (1989) *Physical Optics Notebook: Tutorials in Fourier Optics*, 1st edn. New York: SPIE, Bellingham and AIP.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*, pp. 343–381. New York: John Wiley & Sons.

Speckle and Coherence

G Häusler, University of Erlangen-Nürnberg, Erlangen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Opticists are aware that the amount of coherence plays a significant role in imaging systems: laser speckles are known to add significant noise to the image, as well as parasitic interferences from dusty lenses. Optical systems are often called coherent, if a laser is used (right), and incoherent if other sources come into play (wrong). Many users of optical systems are unaware that it is not the high temporal coherence but the spatial coherence that commonly afflicts the image quality, and that this parasitic spatial coherence is ubiquitous, even though not obvious. Coherent artifacts can occur without the use of lasers, although speckle noise is more prominent with lasers. Even opticists sometimes underestimate the damage that residual coherent noise can cause, and as laser oriented sensor funding programs are 'en vogue', nonexperts are disappointed if some metrology device does not include a laser.

This encyclopedia addresses the many uses of lasers. In this article, we will discuss the costs of coherence. The commonly pretended incoherent approach of everyday optics may lead to significant quantitative measuring errors of illumination or reflectivity, 3d shape, distance or size. Spatial coherence is the dominant source of noise. We will give some rules of thumb to estimate these errors and a few tricks to reduce coherent noise. These rules will help to minimize coherent noise, however, it turns out that as spatial coherence is ubiquitous, there are only limited options available to clear it. One of the options to build good sensors that measure shape, reflectivity, etc. is avoiding the use of lasers!

To become familiar with some basics of the theory of coherence we refer the reader the Further Reading section at the end of this article.

Coherence can be boon or disaster for the opticist, as is explained in other articles of this encyclopedia about interferometry, diffraction, and holography. A specific topic is information acquisition from coherently scattered light. An enlightening example is where speckles in white light interferometry at rough surfaces and in speckle interferometry are exploited. We will briefly discuss white light interferometry at rough surfaces in the section on speckles as carriers of information below.

Practical Coherence Theory

A major issue of this chapter will be corrupting properties of coherence in the daily life of an optical metrologist. We will demonstrate that 'speckle' noise is ever present, and essentially unavoidable, in the images of (diffusively reflecting) objects. Its influence on the quality of optical measurements leads to a lower limit of the physical measuring uncertainty.

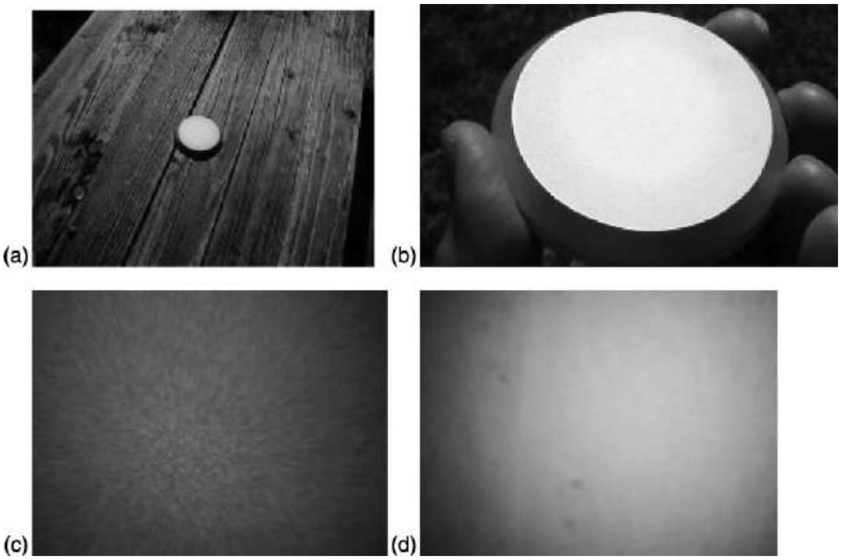

Figure 1 Ground glass in sunlight: (a) the image of a ground glass, illuminated by the sun, observed with small aperture (from large distance), displays no visible speckles; (b) a medium observation aperture (smaller distance) displays weak visible speckles; (c) observation with very high aperture displays strong contrast speckles; and (d) the image of the ground glass on a cloudy day does not display speckles, due to the very big aperture of illumination. Reproduced with permission from Häusler G (1999) Optical sensors and algorithms for reverse engineering. In: Donah S (ed.) Proceedings of the 3rd Topical Meeting on Optoelectronic Distance/Displacement Measurements and Applications. IEEE-LEOS.

This 'coherent noise' limit is, not surprisingly, identical to Heisenberg's limit. We will start by summarizing the important results, and will give some simple explanations later.

Some Basic Observations

(i) The major source of noise in optical measurements is spatial coherence, the temporal coherence status, generally, does not play a significant role.
(ii) An optically rough surface displays subjective speckles with contrast $C = 1$, if the observation aperture $\sin u_o$ is larger than the illumination aperture $\sin u_q$.
(iii) The speckle contrast is $C \sim \sin u_q / \sin u_o$, if the observation aperture is smaller than the illumination aperture.

Figure 1 illustrates the situation by a simple experiment.

The results of Figure 1 can be explained simply by summarizing some 80 pages of coherence theory in a nutshell, by useful simplifications and approximations (Figure 2).

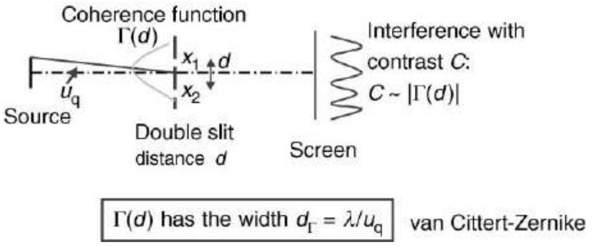

Figure 2 Basic experiment for spatial coherence. An extended source, such as the sun or some incandescent lamp illuminates a double slit, from a large distance. On a screen behind the double slit we can observe 'double slit interference', if the waves coming from x_1 and x_2 display coherence. The contrast of the interference fringes is given by the magnitude of the coherence function Γ (neglecting some normalization factor). Γ is a function of the two slit locations x_1 and x_2.

If the source is at an infinite distance, the coherence function no longer depends on the two variables but just on the slit distance $d = |x_1 - x_2|$. In this case, the coherence function $\Gamma(d)$ can be easily calculated as the Fourier transform of the spatial intensity distribution of the source (van Cittert-Zernike theorem). Let us assume a one dimensional source with an angular size of $2u_q$, where u_q is commonly called the aperture of the illumination. In this case, the coherence function will be:

$$\Gamma(d) \sim \sin c(2u_q d / \lambda) \qquad [1]$$

The width of the coherence function which gives the size of the coherently illuminated area (coherence area) can be approximated from eqn [1]:

$$d_\Gamma = \lambda / u_q \qquad [2]$$

Equation [1] includes some approximations: for circular sources such as the sun, we get an additional factor 1.22 for the width of the coherence function – the 'sinc-function' has to be replaced by the Bessel function $J_1(r)/(r)$. There is another approximation: we should replace the aperture angle u_q by $\sin u_q$, for larger angles. We choose the wavelength of the maximum sensitivity of the eye, at $\lambda = 0.55\ \mu\text{m}$, which is the wavelength of the maximum emission of the sun also.

In conclusion, if two points at the object are closer than d_Γ these points will be illuminated with a high degree of coherence. The light waves, scattered from these object-locations can display interference contrast. Specifically, they may cancel each other out, causing a low signal-to-noise ratio of only 1:1. This does not happen if the points have a distance larger than d_Γ.

From Figure 3 we learn that the width of the spatial coherence function from sunlight illumination at the earth's surface is ~110 μm. (Stars more distant than the sun appear smaller, and hence have a wider coherence function, at the earth's surface. Michelson was able to measure the angular size of some close stars, by measuring the width of this coherence function, which was about $d_\Gamma \sim 10$ m.)

Figure 4 again displays an easy to perform experiment. We can see speckles at sunlight illumination. We can observe speckles as well in shops because they often use small halogen spot illumination.

So far we have only discussed the first stage (from the source to the object). We still have to discuss the way from the object to the retina (or CCD-chip of a video camera). This is illustrated by Figure 5.

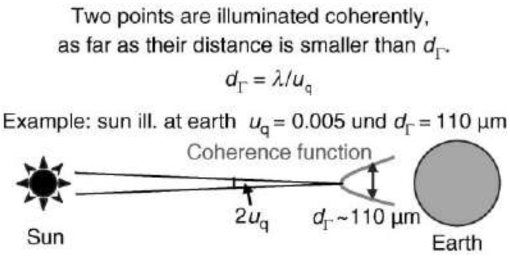
Two points are illuminated coherently,
as far as their distance is smaller than d_Γ.
$d_\Gamma = \lambda/u_q$
Example: sun ill. at earth $u_q = 0.005$ und $d_\Gamma = 110$ µm
Coherence function
$2u_q$
$d_\Gamma \sim 110$ µm
Sun
Earth

Figure 3 Coherence from sunlight illumination. With the van Cittert–Zernike theorem and an approximate illumination aperture of the sun $u_q^{\text{sun}} \sim 0.005$ (0.25°), we get a width of the coherence function at the Earth's surface of about $d_\Gamma \sim 110$ μm.

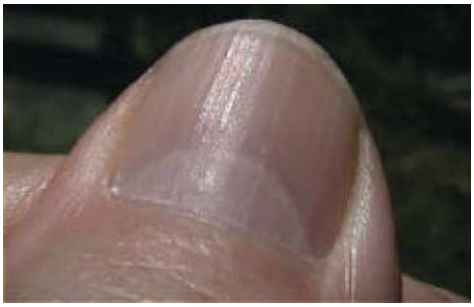

Figure 4 Finger nail in sunlight, with speckles.

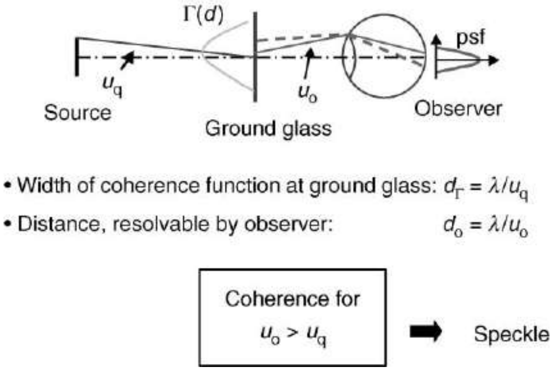
$\Gamma(d)$
psf
u_q
u_o
Source
Ground glass
Observer
• Width of coherence function at ground glass: $d_\Gamma = \lambda/u_q$
• Distance, resolvable by observer: $d_o = \lambda/u_o$
Coherence for $u_o > u_q$
Speckle

Figure 5 Coherence in a nutshell: spatial coherence and observer.

As an object, we choose some diffusely reflecting or transmitting surface (such as a ground glass). The coherence function at the ground glass has again a width of d_Γ. What happens if we image this ground glass onto our retina? Let us assume our eye to be diffraction limited (which, in fact, it is at sunlight conditions, where the eye pupil has only a diameter $\Phi_{\text{pupil}} = 2.5$ mm diameter or less). Then a point x_1 at the ground glass will cause a diffraction spot at the retina, with a size of:

$$d'_{\text{diffr}} = \lambda/u'_o = 8\ \mu\text{m} \qquad [3]$$

The image sided aperture of observation u'_0 of the eye is calculated from Φ_{pupil} and from its focal length $f_{\text{eye}} = 18$ mm as $u'_o = \Phi_{\text{pupil}}/2f_{\text{eye}} = 0.07$. If we project the diffraction spot at the retina back onto the object, we get its size d from:

$$d_{\text{diffr}} = \lambda/u_o = \lambda/(\Phi_{\text{pupil}}/2z_o) \qquad [4]$$

with z_o = observation distance of the object from the eye. We call $u_o = \Phi_{\text{pupil}}/2z_o$ the object sided observation aperture. Let us calculate the laterally resolvable distance d_{diffr} at the object, with a distance $z_o = 250$ mm, which is called the 'comfortable viewing distance':

$$d_{\text{diffr}}(z_o = 250\ \text{mm}) = 110\ \mu\text{m} \qquad [5]$$

After these preparations we come to the crucial issue: how does the image at the retina appear if we cannot resolve distances at an object smaller than d_Γ, and how does it appear if the resolution of the eye is sufficient to resolve distances smaller than d_Γ?

Let us start with the first assumption: $d_{\text{diffr}} \gg d_\Gamma$. Now the images of points at the object, over an area of the diffraction spot are more or less incoherently averaged at the retina, so we will see little interference contrast or 'speckles'. From eqn [4] we see that at this incoherent averaging starts for distances larger than 250 mm, which is the comfortable viewing distance. Adults cannot see objects from much closer distances. We generally do not see speckles, not just at the limit of maximum accommodation. For larger distances, if we take $z_o = 2500$ mm, the resolvable distance at the object will be 1.1 mm, which is 10 times larger than the diameter of the coherence area. Averaging over such a large area will drop the interference contrast by a factor of 10. Note that such a small interference contrast might not be visible, but it is not at zero!

Coming to the second assumption we can laterally resolve distances smaller than d_Γ. In order to understand this, we first have to learn what is an 'optically rough' object surface. Figure 6 illustrates the problem.

A surface commonly is called 'rough' if the local height variations are larger than λ. However, a surface appears only rough, if the height variation within the distance that can be resolved by the observer, is larger than λ (reflected light will travel the distance twice, hence a roughness of $\lambda/2$ will be sufficient). Then the scattered different phasors, or 'Huygens elementary waves':

$$u_s \sim \exp[i2kz(x, y)] \qquad [6]$$

scattered from the object at x, y may have big phase differences so that destructive interferences (and speckle) can occur in the area of the diffraction image. So the attribute 'rough' depends on the object as well as on the observation aperture. With a high observation aperture (microscope), the diffraction image is small and within that area the phase differences might be small as well. So a ground glass may then look locally like a mirror, while with a small observation aperture it appears 'rough'.

With the 'rough' observation mode assumption, we can summarize what we have assumed: with resolving distances smaller than the coherence width d_Γ we will see high interference contrast; in fact we see a speckle contrast of $C = 1$, if the roughness of the object is smaller than the coherence length of the source. This is the case for most metal surfaces, for ground glasses, or for worked plastic surfaces. The assumption is not true for 'translucent' surfaces, such as skin, paper, or wood. This will be discussed below.

The observation that coherent imaging is achieved, if we can resolve distances smaller than the coherence width d_Γ, is identical to the simple rule that fully coherent imaging occurs if the observation aperture u_o is larger than the illumination aperture u_q.

As mentioned, we will incoherently average in the image plane, over some object area, determined by the size of the diffraction spot. According to the rules of speckle-averaging, the speckle contrast C decreases with the inverse square root of the number N of incoherently averaged speckle patterns. This number N is equal to the ratio of the area A_{diff} of the diffraction spot divided by the coherence area $A_\Gamma = d_\Gamma^2$. So we obtain for the speckle contrast C:

$$C = 1 \quad \text{for } u_q < u_o \qquad [7a]$$

$$C = 1/\sqrt{(N)} = u_o/u_q, \quad \text{for } u_q >= u_o \qquad [7b]$$

We summarize the results in Figure 7.

Equation [7b] has an interesting consequence: we never get rid of speckle noise, even for large illumination apertures and small observation apertures. In many practical instruments, such as a slide projector, the illumination 'aperture stop' cannot be greater than the imaging lens. Fortunately, the observer's eye commonly has a pupil smaller than that of the projector, and/or looks from a distance at the projection screen. Laser projection devices however cause strong and disturbing speckle effects for the user and significant effort is invested to cope with this effect.

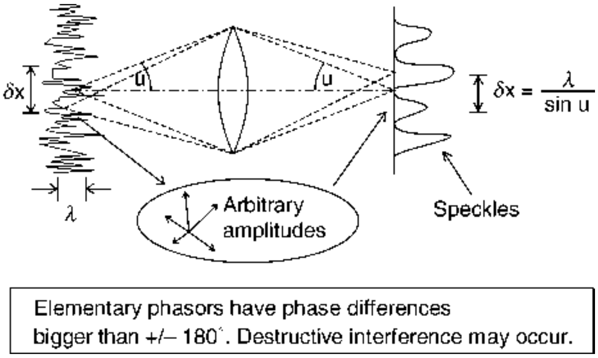
δx
u
u
δx = λ / sin u
λ
Arbitrary amplitudes
Speckles
Elementary phasors have phase differences bigger than +/– 180°. Destructive interference may occur.

Figure 6 What is a rough surface?

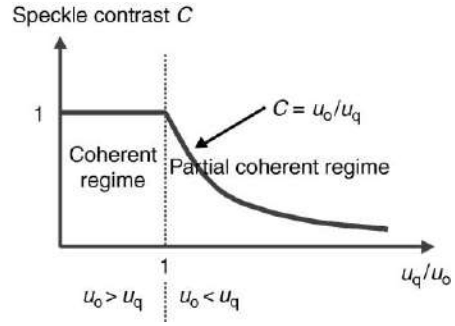

Figure 7 Coherence in a nutshell: what are the conditions for incoherent and coherent imaging?

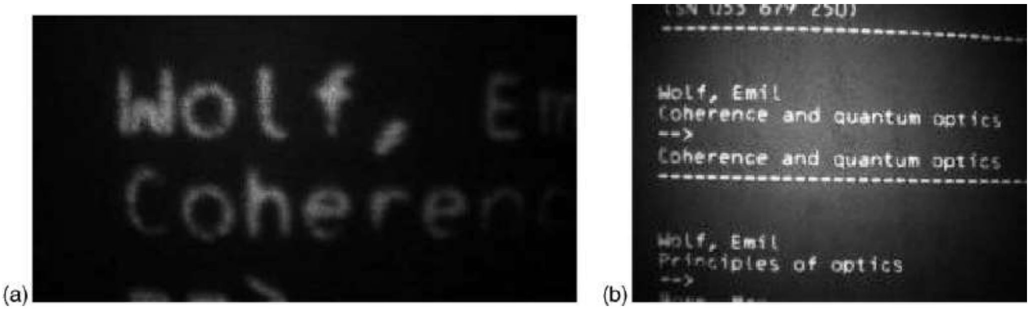

Figure 8 Microfiche projection. For high observation aperture (a) strong speckles occur, and for small observation aperture (b), the speckle contrast is low.

In Figure 8, an image is depicted from a microfiche reading projector. Figure 8a displays the close-up look (high observation aperture) and Figure 8b depicts a more distant view (small observation aperture). The amount of noise is much less in the second case, as it should be according to the rules of eqn [7].

Let us finish this section with some speculation. How would a fly see the sunny world, with a human type of eye? With high observation aperture and at short distance, the world is full of speckle noise. Fortunately, nature invented the facet-eye, for insects, as depicted in Figure 9.

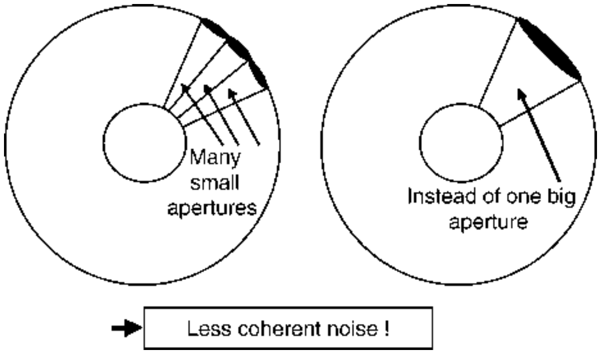

Figure 9 Facet eye and 'human' eye. The facet eye consists of many low aperture lenses, in contrast to the human eye.

Speckle Limits of Metrology

The consequences of the effects discussed above are often underestimated. We should be suspicious, if some effect in nature that disturbs us – like our coherent noise – is ubiquitous, that there might be some deep underlying principle that does not allow us to know everything about the object under observation. Indeed, it turns out that Heisenberg's uncertainty principle is strongly connected with coherent noise (Figure 10).

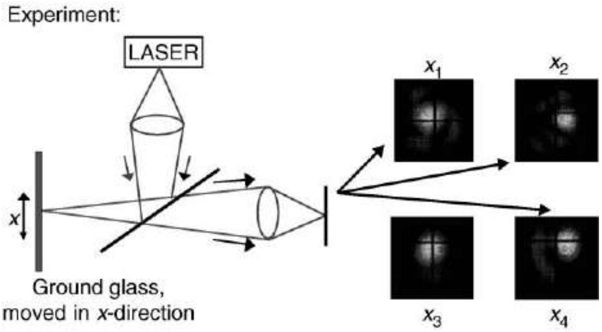

Figure 10 Coherence makes it impossible to localize objects with very high accuracy. Laser spots projected onto a ground glass cannot be imaged without some uncertainty of the lateral position. The four pictures of the spot images, taken with different lateral positions of the ground glass, display the cross bar (true position) and the apparent position of the spot images. The position uncertainty is equal to the uncertainty calculated from Heisenberg's uncertainty principle.

We can see this from the following experiment: a laser spot is projected onto a ground glass, and imaged with high magnification by a video camera, with an aperture sin u_o. The ground glass is macroscopically planar. When the ground glass is laterally shifted, we find that the observed spot is 'dancing' at the video target. Its observed position is not constant, although its projected position is. It turns out that the standard deviation of the observed position is equal to the uncertainty calculated from the aperture by Heisenberg's principle.

We can calculate the limit for the distance measuring uncertainty, from speckle theory, as well from Heisenberg's principle (within some factor of order 1):

$$\delta z \cdot \delta p_z \gg h/4\pi \qquad [8]$$

where δp_z is the uncertainty of the photon impulse h/λ in the z-direction (along the optical axis of the measuring system) and h is Planck's constant. For a small measurement uncertainty of the distance, we should allow a big uncertainty δp_z. This can be achieved by a large aperture of the observation system, giving the photons a wide range of possible directions to the lens. We can also allow different wavelengths, and come to white light interferometry, see section on speckles as carriers of information below. The result is – not surprising – the same as Rayleigh scattering found for the depth of field:

$$\delta z = \lambda / \sin^2 u_o \qquad [9]$$

Coherent noise is the source of the fundamental limit of the distance measuring uncertainty δz of triangulation based sensors.

$$\delta z = C \cdot \lambda / (\sin u_o \sin \theta) \qquad [10]$$

where C is the speckle contrast, λ is the wavelength, $\sin u_o$ is the aperture of observation, and θ is the angle of triangulation (between the direction of the projection and the direction of observation). For a commercially available laser triangulation sensor, (see Figure 11), with $\sin u_o \sim 0.01, C = 1$ and $\theta = 30°$, the measuring uncertainty δz will be larger than 100 µm. We may add, that for $\sin u_o = \sin \theta$, which is valid for auto-focus sensors (such as the confocal scanning microscope), eqn [10] degenerates to the well known Rayleigh depth of field (eqn [9]).

The above results are remarkable in that we cannot know the accurate position of the projected spot or about the position of an intrinsic reflectivity feature; we also cannot know the accurate local reflectivity of a coherently illuminated object. A further consequence is that we do not know the accurate shape of an object in 3D-space. We can calculate this 'physical measuring uncertainty' from the considerations above.

These consequences hold for optically rough surfaces, and for measuring principles that exploit the local intensity of some image, such as with all triangulation type of sensors. The consequences of these considerations are depressing: triangulation with a strong laser is not better than triangulation with only one single photon. The deep reason is, that all coherent photons stem from the same quantum mechanical phase cell and are indistinguishable. Hence, many photons do not supply more information than one single photon.

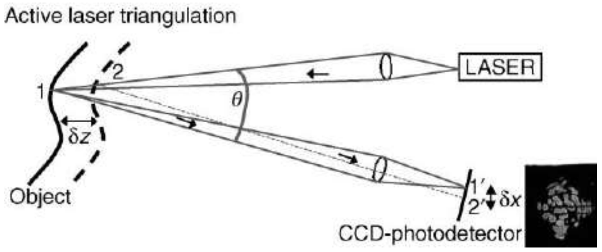

Figure 11 Principle of laser triangulation. The distance of the projected spot is calculated from the location of the speckled spot image and from the angle of triangulation. The speckle noise introduces an ultimate limit of the achievable measuring uncertainty. Reproduced with permission from *Physikalische Plätter* (May 1997): 419.

Why Are We Not Aware of Coherence

Generally, coherent noise is not well visible, even in sophisticated technical systems the visibility might be low. There are two main reasons; first, the observation aperture might be much smaller than the aperture of illumination. This is often true, for large distance observation, even with small apertures of illumination. The second reason holds for technical systems, if the observation is implemented via pixelized video targets. If the pixel size is much larger than a single speckle, which is commonly so, then, by averaging over many speckles, the noise is greatly reduced. However, we have to take into account that we pay for this by loss of lateral resolution $1/\delta x$. We can formulate another uncertainty relation:

$$\delta x > \lambda / (C \cdot \sin u_o) \qquad [11]$$

which says that if we want to reduce the speckle contrast C (noise), by lateral averaging, we can do this but lose lateral resolution $1/\delta x$.

Can We Overcome Coherence Limits?

Since the daily interaction of light with matter is coherent scattering, we can overcome the limit of eqn [10] only by looking for measuring principles that are not based on the exploitation of local reflectivity, i.e., on 'conventional imaging'. Concerning optical 3D-measurements, the principle of triangulation uses the position of some image detail, for the calculation of the shape of objects, and has to cope with coherent noise.

Are there different mechanisms of photon–matter interaction, with noncoherent scattering? Fluorescence and thermal excitation are incoherent mechanisms. It can be shown (see Figure 12) that triangulation utilizing fluorescent light, displays much less noise than given by eqn [10]. This is exploited in fluorescent confocal microscopy.

Thermally excited matter emits perfectly incoherent radiation as well. We use this incoherence to measure the material wear in laser processing (see Figure 13), with, again, much better accuracy than given by eqns [9,10].

The sensor is based on triangulation. Nevertheless, by its virtually zero coherent noise, it allows a measuring uncertainty which is limited only by camera noise and other technical imperfections. The uncertainty of the depth measurement through the narrow aperture of the laser nozzle is only a few

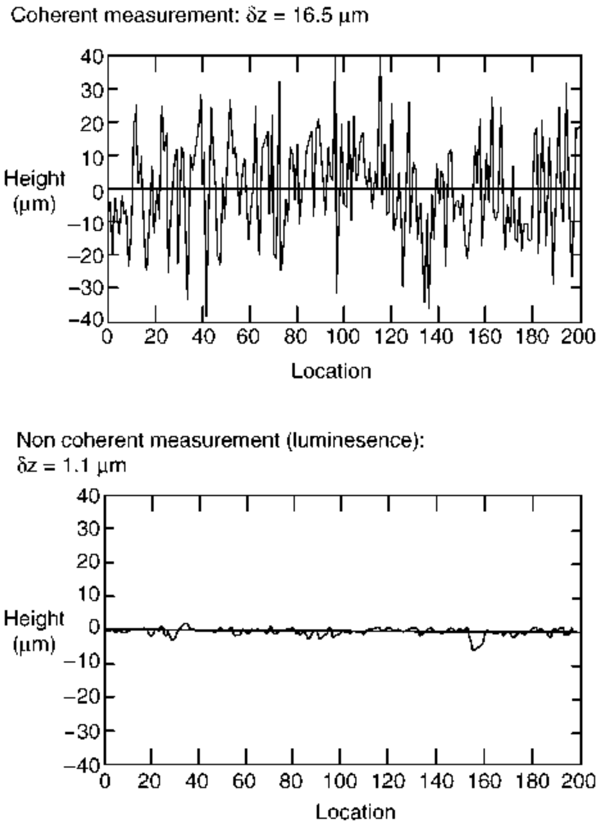

Figure 12 Reduction of speckle noise by triangulation with fluorescent light. A flat metal surface is measured by laser triangulation. Top: surface measured by laser triangulation with full speckle contrast. Bottom: the surface is covered with a very thin fluorescent layer, and illuminated by the same laser, however, the triangulation is done after suppressing the scattered laser light, and utilizing the (incoherent!) fluorescent light. This experiment proves that it is the coherent noise that causes the measuring uncertainty.

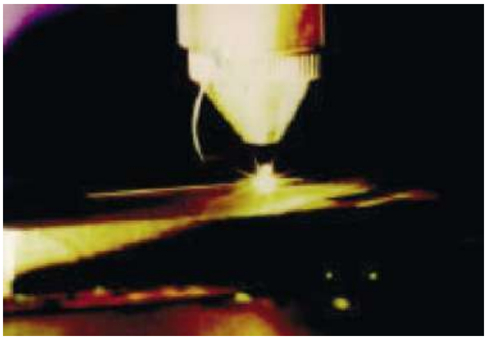

Figure 13 A 100 W laser generates a plasma that emits perfectly incoherent radiation at the object surface. The emitted light can be used to measure on-line the distance of the object surface. A triangulation sensor, exploiting the incoherent plasma emission, controls surface ablation with an accuracy of only a few microns. Reproduced with permission from F & M, *Feinwerktechnik Mikrotechnik Masstechnik* (1995) Issue 9.

microns, beating the limitation of eqn [10] by a factor of about 5, even in the presence of a turbulent plasma spot with a temperature of 3000 K.

Broadband Illumination

Speckle noise can hardly be avoided by broadband illumination. The speckle contrast C is related to the surface roughness σ_z and the coherence length l_c by eqn [12]:

$$C^2 \sim l_c/2\sigma_z \quad \text{for} \quad \sigma_z \gg l_c \qquad [12]$$

So, only for very rough surfaces, and white light illumination, C can be reduced. However, the majority of technical objects is smooth, with a roughness of a few micrometers, hence speckles can be observed even with white light illumination ($l_c \sim 3$ μm). This situation is different for 'translucent' objects such as skin, paper, or some plastic material.

Speckles as a Carrier of Information

So far we discussed the problems of speckles when measuring rough surfaces. We can take advantage of speckles, if we do not stick to triangulation as the basic mechanism of distance measurement. We may utilize the fact that although the phase within each speckle has a random value, this phase is constant over some area. This observation can be used to build an interferometer that works even at rough surfaces. So, with proper choice of the illumination aperture and observation aperture, and with a surface roughness less than the coherence length, we can measure distances by the localization of the temporal coherence function. This function can easily be measured by moving the object under test along the optical axis and measuring the interference intensity (correlogram) at each pixel of the camera. The signal generating mechanism of this 'white light interferometer' distance measurement is not triangulation but 'time of flight'. Here we do not suffer from the limitations of eqn [10]. The limits of white light rough surface interferometry are discussed in the Further Reading section at the end of this article.

The 'coherence radar' and the correlograms of different speckles are depicted in Figures 14 and 15.

The physical mechanism of the signal formation in white light interferometry at rough surfaces is different from that of white light interferometry at smooth surfaces. That is why the method is sometimes referred as 'coherence radar'. We will briefly summarize some advantageous, and – probably unexpected – features of the coherence radar.

We call the real shape of the object, or the surface profile, $z(x, y)$, where z is the local distance at the position (x, y). We give the measured data an index 'm^2':

(i) The physical measuring uncertainty of the measured data is independent from the observation aperture. This is quite a remarkable property, as it enables us to take accurate measurements at the bottom of deep boreholes.
(ii) The standard deviation σ_z of the object can be calculated from the measured data z_m: $\sigma_z = <|z_m|>$. According to German standards, σ_z corresponds to the roughness measure R_q.

This measure R_q was calculated from measurements at different roughness gauges, with different observation apertures. Two measurements are depicted in Figure 16.

The experiments shown in Figure 16 display the correct roughness measure, even if the higher frequencies of the object spectrum $Z(\nu, \mu)$ (Figure 17), are not optically resolved. The solution of this paradox might be explained as follows.

From the coherently illuminated object each point scatters a spherical wave. The waves scattered from different object points have, in general, a different phase. At the image plane, we see the laterally

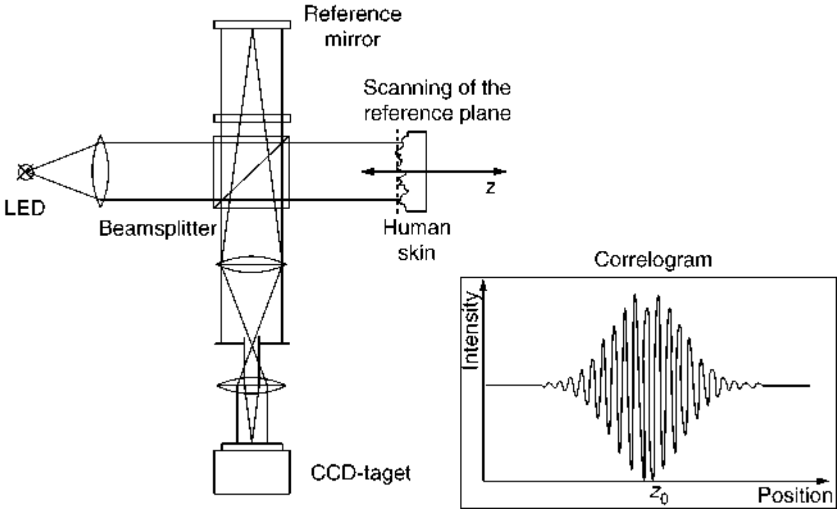

Figure 14 'Coherence radar', white light interferometry at rough surfaces. The surface under test is illuminated with high spatial coherence and low temporal coherence. We acquire the temporal coherence function (correlogram) within each speckle, by scanning the object under test in depth.

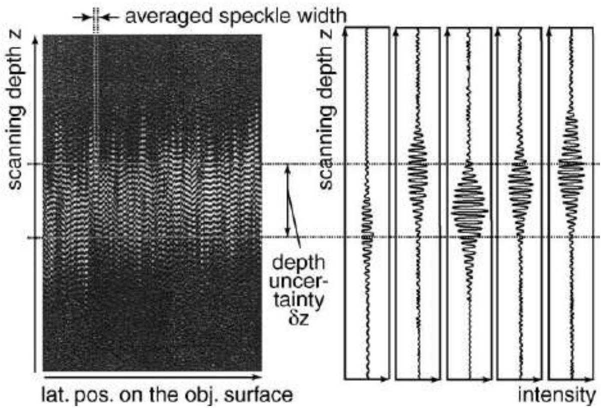

Figure 15 White light speckles in the $x-z$ plane. The left side displays the acquired correlograms, for a couple of speckles, the right side shows the graphs of some of the correlograms. We see that the correlograms are not located at a constant distance, but display some 'distance uncertainty'. This uncertainty does not originate from the instrument, but from the roughness of the object.

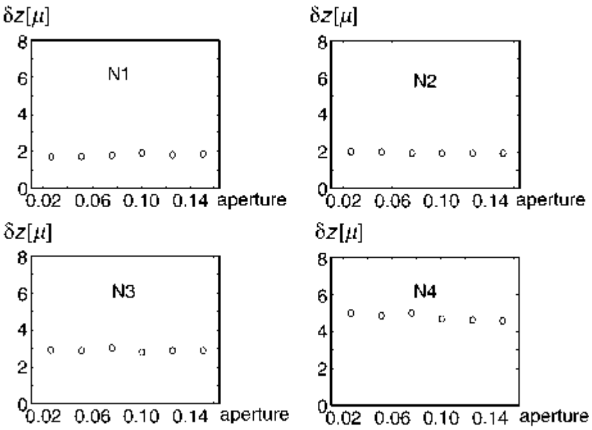

Figure 16 Optical measurement of the roughness beyond the Abbe resolution limit. The roughness is measured for 4 roughness standards N1–N4, with different observation apertures. The roughness is measured correctly, although the microtopology is not resolved by the observing optics.

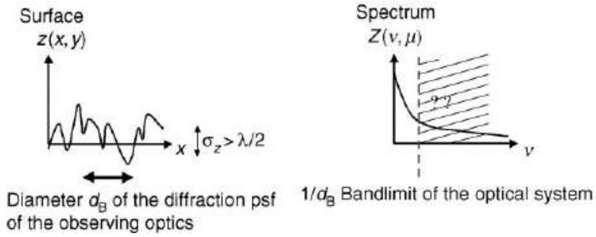
Surface
$z(x, y)$
x
$\sigma_z > \lambda/2$
Diameter d_B of the diffraction psf of the observing optics
Spectrum
$Z(\nu, \mu)$
??
ν
$1/d_B$ Bandlimit of the optical system

Figure 17 Signal generation beyond the Abbe limit. The microtopology of the object $z(x, y)$ is not resolved, due to the large diffraction image with diameter d_B. Yet we get information from beyond the bandlimit $1/d_B$ of the observing optics.

averaged complex amplitude $<u>$, according to eqn [6]. The averaging is due to the diffraction limited resolution.

Equation [6] reveals that the (spatially) averaged field amplitude $<u>$ is a nonmonotonic, nonlinear function of the surface profile $z(x, y)$, if the surface is rough (i.e., $z \gg \lambda$). We do not average over the profile z but over the complex amplitude. As a consequence of this nonlinearity, the limited observation aperture does not only collect the spatial frequencies ν, μ of the spatial Fourier spectrum $Z(\nu, \mu)$ within the bandlimit $1/d_B$, but also acquires information beyond this bandlimit. The reason is the 'down conversion' of higher spatial frequencies by nonlinear mixing. Simulations and experiments confirm the consequences.

Thus we can evaluate the surface roughness, even without laterally resolving the microtopology. However, there is a principal uncertainty on the measured data. We do not see the true surface but a surface with a 'noise' equivalent to the roughness of the real surface.

Nonlinear nonmonotonic nonlinearities in a sequence of operations may cause 'chaotic' behavior, in other words, small parameter changes such as vibrations, humidity on the surface, etc. may cause large differences of the outcome. In fact, we observe a significant variation of the measured data $z_m(x, y)$ within a sequence of repeated measurements. Our conjecture is that the complex signal formation may be involved in this irregular behavior. The hypothesis may be supported by the observation that much better repeatability can be achieved at specular surfaces, where eqn [6] degenerates to a linear averaging over the surface profile:

$$<u> \sim 1 + ik <z(x, y)>, \quad \text{for } z \ll \lambda \qquad [13]$$

Summary

Spatial coherence is ubiquitous and unavoidable. Spatial coherence disturbs most optical measurements, where optically rough surfaces exist. Then, coherence gives the ultimate limit of the achievable measuring uncertainty. There are simple rules to estimate the measuring uncertainty, by calculating the width of the coherence function and the resolution of the observation system. Spatial coherence and Heisenberg's uncertainty principle lead to the same results of measuring uncertainty.

On the other hand – the formation of signals from coherent light that is scattered at rough surfaces is quite complex and a strongly nonlinear process, which sometimes might encode information which is not otherwise available. One example appears to be the measurement of the roughness with white light interferometry – which is possible, even if the microtopology of the surface is not optically resolved.

See also

Coherence: Overview; Coherence and Imaging.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Brooker G (2003) *Modern Classical Optics*. Oxford: Oxford University Press.

Dainty JC (ed.) (1984) *Laser Speckle and Related Phenomena*. Berlin, Heidelberg, New York: Springer Verlag.

Dorsch R, Herrmann J and Häusler G (1994) Laser triangulation: fundamental uncertainty in distance measurement. *Applied Optics* 33: 1306–1314.

Dresel T, Venzke H and Häusler G (1992) 3D-sensing of rough surfaces by coherence radar. *Applied Optics* 31: 919–925.

Ettl P (2002) *Über die Signalenstehung bei Weisslichtinterferometrie*. PhD Dissertation, University of Erlangen-Nürnberg.

Häusler G, Ettl P, Schenk M, Bohn G and Laszlo I (1999) Limits of optical range sensors and how exploit them. In: Asakura T (ed.) *Trends in Optics and Phototonics*. Ico IV, Springer Series in Optical Sciences, vol. 74, pp. 328–342. Berlin, Heidelberg, New York: Springer Verlag.

Häusler G and Herrmann J (1992) Physical limits of 3D-sensing. Proceedings of the SPIE Conf. 1822. In: Svetkov D (ed.) *Optics, Illumination, and Image Sensing for Machine Vision VII*, pp. 150–158. Boston, MA: SPIE.

Klein M and Furtak TE (1986) *Optics*. New York: John Wiley & Sons.

Klinger P, Spellenberg B, Herrmann JM and Häusler G (2001) *In Process 3D-Sensing for Laser Material Processing*. Proceedings of Third Int. Conf. on 3-D Digital Imaging and Modeling, Quebec City, Canada. Los Alamitos: IEEE Computer Society, 38–41.

Pedrotti F and Pedrotti L (1993) *Introduction to Optics*, 2nd edn. New Jersey: Prentice Hall.
Pérez J-Ph (1996) *Optik, Spektrum Akademischer*. Heidelberg: Verlag.
Sharma A, Kandpal HC and Ghatak AK (2004) *Optical Coherence*. See this Encyclopedia.
van de Gracht (2004) *Coherence and Imaging*. See this Encyclopedia.
Young M (1993) *Optics and Lasers, Including Fibers and Optical Wave Guides*, 4th revised edn. Berlin, Heidelberg: Springer Verlag.

COHERENT CONTROL

Contents

Theory
Experimental
Applications in Semiconductors

Theory

H Rabitz, Princeton University, Princeton, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the inception of quantum mechanics almost a century ago, a prime activity has been the observation of quantum phenomena in virtually all areas of chemistry and physics. However, the natural evolution of science leads to the desire to go beyond passive observation to active manipulation of quantum mechanical processes. Achieving control over quantum phenomena could be viewed as engineering at the atomic scale guided by the principles of quantum mechanics, for the alteration of system properties or dynamic behavior. From this perspective, the construction of quantum mechanically operating solid-state devices through selective material growth would fall into this category. The focus of this article is principally on the manipulation of quantum phenomena through tailored laser pulses. The suggestion of using coherent radiation for the active alteration of microworld processes may be traced to the early 1960s, almost immediately after the discovery of lasers. Since then, the subject has grown enormously to encompass the manipulation of (1) chemical reactions, (2) quantum electron transport in semiconductors, (3) excitons in solids, (4) quantum information systems, (5) atom lasers, and (6) high harmonic generation, amongst other topics. Perhaps the most significant use of these techniques may be their provision of refined tools to ultimately better understand the basic physical interactions operative at the atomic scale.

Regardless of the particular application of laser control over quantum phenomena, there is one basic operating principle involved: active manipulation of constructive and destructive quantum wave interferences. This process is depicted in Figure 1, showing the evolution of a quantum system from the initial state $|\psi_i\rangle$ to the desired final state $|\psi_f\rangle$ along three of possibly many interfering pathways. In general, there may be many possible final accessible states $|\psi_f\rangle$, and often the goal is to achieve a high amplitude in one of these states and low amplitude in all the others. The target state might actually be a superposition of states, and an analogous picture to

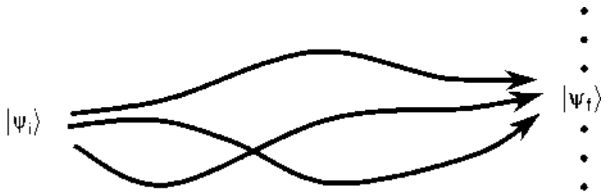

Figure 1 The evolution of a quantum system under laser control from the initial state $|\psi_i\rangle$ to the final state $|\psi_f\rangle$. The latter state is chosen to have desirable physical properties, and three of possibly many pathways between the states are depicted. Successful control of the process $|\psi_i\rangle \to |\psi_f\rangle$ by a tailored laser pulse generally requires creating constructive quantum wave interferences in the state $|\psi_f\rangle$ from many pathways, and destructive interferences in all other accessible final states $|\psi_{f'}\rangle \neq |\psi_f\rangle$.

that in Figure 1 would also apply to steering about the density matrix. The process depicted in Figure 1 may be thought of as a microscale analog of the classic double-slit experiment for light waves. As the goal is often high-finesse focusing into a particular target quantum state, success may call for the manipulation of many quantum pathways (i.e., the notion of a 'many-slit' experiment). Most applications are concerned with achieving a desirable outcome for the expectation value $\langle\psi_f|O|\psi_f\rangle$ associated with some observable Hermitian operator O.

The practical realization of the task above becomes a control problem when the system is expressed through its Hamiltonian $H = H_0 + V_c$, where H_0 is the free Hamiltonian describing the dynamics without explicit control; it is assumed that the free evolutionary dynamics under H_0 will not satisfactorily achieve the physical objective. Thus, a laboratory-accessible control term V_c is introduced in the Hamiltonian to achieve the desired manipulation. Even just considering radiative interactions, the form of V_c could be quite diverse depending on the nature of the system (e.g., nuclear spins, electrons, atoms, molecules, etc.) and the intensity of the radiation field. Many problems may be treated through an electric dipole interaction $V_c = -\boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)$ where $\boldsymbol{\mu}$ is a system dipole moment and $\boldsymbol{\varepsilon}(t)$ is the laser electric field. Where appropriate, this article will consider the interaction in this form, but other suitable radiative control interactions may be equally well treated. Thus, the laser field $\boldsymbol{\varepsilon}(t)$ as a function of time (or frequency) is at our disposal for attempted manipulation of quantum systems.

Before considering any practical issues associated with the identification of shaped laser control fields, a fundamental question concerns whether it is, in principle, possible to steer about any particular quantum system from an arbitrary initial state $|\psi_i\rangle$ to an arbitrary final state $|\psi_f\rangle$. Questions of this sort are addressed by a controllability analysis of Schrödinger's equation

$$i\hbar\frac{\partial|\psi(t)\rangle}{\partial t} = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]|\psi(t)\rangle, \quad |\psi(0)\rangle = |\psi_i\rangle \qquad [1]$$

Controllability concerns whether, in principle, some field $\boldsymbol{\varepsilon}(t)$ exists, such that the quantum system described by $H_0 = H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)$ permits arbitrary degrees of control. For finite-dimensional quantum systems (i.e., those described by evolution amongst a discrete set of quantum states), the formal tools for such a controllability analysis exist both for evaluating the controllability of the wave function, as well as the more general time evolution operator $U(t)$, which satisfies

$$i\hbar\frac{\partial U}{\partial t} = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]U, \quad U(0) = 1 \qquad [2]$$

Analyses of this type can be quite insightful, but they require detailed knowledge about H_0 and $\boldsymbol{\mu}$. Cases involving quantum information science applications are perhaps the most demanding with regard to achieving total control. Most other physical applications would likely accept much more modest levels of control and still be categorized as excellent achievements.

Theoretical tools and concepts have a number of roles in considering the control of quantum systems, including (1) an exploration of physical/chemical phenomena under active control, (2) the design of viable control fields, (3) the development of algorithms to actively guide laboratory control experiments towards achieving their dynamical objectives, and (4) the introduction of special algorithms to reveal the physical mechanisms operative in the control of quantum phenomena. Activities (1)–(3) have thus far been the primary focus of theoretical studies in this area, and it is anticipated that item (4) will grow in importance in the future.

A few comments on the history of laser control over quantum systems are relevant, as they speak to the special nature of the currently employed successful closed-loop quantum control experiments. Starting in the 1960s and spanning roughly 20 years, it was thought that the design of lasers to manipulate molecular motion could be achieved by the application of simple physical logic and intuition. In particular, the thinking at the time focused on using cw laser fields resonant with one or more local modes of the molecule, as state or energy localization was believed to be the key to successful control. Since quantum dynamics phenomena typically occur on ultrafast time-scales, possibly involving spatially dispersed wave packets, expecting to achieve high quality control with a light source operating at one or two resonant frequencies is generally wishful thinking. Quantum dynamics phenomena occur in a multifrequency domain, and controls with a few frequencies will not suffice. Over approximately the last decade, it became clear that successful control often calls for manipulating multiple interfering quantum pathways (cf., Figure 1). In turn, this recognition led to the need for broad-bandwidth laser sources (i.e., tailored laser pulses). Fortunately, the necessary laser pulse-shaping technologies have become available, and these sources continue to expand into new frequency ranges with enhanced bandwidth capabilities. Many chemical and physical applications of control over quantum phenomena involve performing a significant amount of work (i.e., quantum mechanical action), and sufficient laser field intensity is required. A commonly quoted adage is 'no field, no yield', and at the other extreme, there was much speculation that operating nonlinearly at high field intensities would lead to a loss of control because the quantum system would effectively act as an amplifier of even weak laser field noise. Fortunately, the latter outcome has not occurred, as is evident from a number of successful high field laser control experiments.

Quantum control may be expressed as an inverse problem with a prescribed chemical or physical objective, and the task being the discovery of a suitable control field $\boldsymbol{\varepsilon}(t)$ to meet the objective. Normally, Schrödinger's equation [1] is viewed as linear, and this perspective is valid if the Hamiltonian is known a priori, leaving the equation to be solved for the wave function. However, in the case of quantum control, neither the wave function nor the control field is known a priori, and the two enter bilinearly on the right-hand side of eqn [1]. Thus, quantum control is mathematically a nonlinear inverse problem. As such, one can anticipate possibly diverse behavior in the process of seeking successful controls, as well as in the ensuing quantum dynamical control behavior. The control must take into account the evolving system dynamics, thereby resulting in the control field being a function(al) of the current state of the system $\boldsymbol{\varepsilon}(\psi)$ in the case of quantum control design, or dependent on the system observations $\boldsymbol{\varepsilon}(\langle O\rangle)$ in the case of laboratory closed-loop field discovery efforts where $\langle O\rangle = \langle\psi|O|\psi\rangle$. Furthermore, the control field at the current time t will depend in some implicit fashion on the future state of the evolving system at the final target time T. It is evident that control field design and laboratory discovery present highly nontrivial tasks.

Although the emphasis in this review is on the theoretical aspects of quantum control theory, the subject exists for its laboratory realizations, and those realizations, in turn, intimately depend on the capabilities of theoretical and algorithmic techniques for their successful implementations. Accordingly, theoretical laser control field design techniques will be discussed, along with algorithmic aspects of current laboratory practice. This material respectively will focus on optimal control theory (OCT) and optimal control experiments (OCEs), as seeking optimality provides the best means to achieve any posed objective. Finally, the last part of the article will present some general conclusions on the state of the quantum control field.

Quantum Optimal Control Theory for Designing Laser Fields

When considering control in any domain of application, a reasonable approach is to computationally design the control for subsequent implementation in the laboratory. Quantum mechanical laser field design has taken on a number of realizations. At one limit is the application of simple intuition for design purposes, and in some special cases (e.g., the weak field perturbation theory regime), this approach may be applicable. Physical insights will always play a central role in laser field design, but to be especially useful, they need to be channeled into the proper mathematical framework. Many of the interesting applications operate in the strong-field nonperturbative regime. In this domain, serious questions arise regarding whether sufficient information is available about the Hamiltonian to execute reliable designs. Regardless of whether the Hamiltonian is known accurately or is an acknowledged model, the theoretical study of quantum control can provide physical insight into the phenomena involved, as well as possibly yielding trial laser pulses for further refinement in the laboratory.

Achieving control over quantum mechanical phenomena often involves a balance of competing dynamical processes. For example, in the case of aiming to create a particular excitation in a molecule or material, there will always be the concomitant need to minimize other unwanted excitations. Often, there are also limitations on the form, intensity, or other characteristics of the laser controls that must be adhered to. Another goal is for the control outcome to be as robust as possible to the laser field fluctuations and Hamiltonian uncertainties. Overriding all of these issues is merely the desire to achieve the best possible physical result for the posed quantum mechanical objective. In summary, all of these desired extrema conditions translate over to posing the design of control laser fields as an optimization problem. Thus, optimal control theory forms a basic foundation for quantum control field design.

Control field design starts by specification of the Hamiltonian components H_0 and $\boldsymbol{\mu}$ in eqn [1], with the goal of finding the best control electric field $\boldsymbol{\varepsilon}(t)$ to balance the competing objectives. Schrödinger's equation must be solved as part of this process, but this effort is not merely a forward propagation task, as the control field is not known a priori. As an optimal design is the goal, a cost functional $J = J(\text{objectives, penalties, } \boldsymbol{\varepsilon}(t))$ is prescribed, which contains the information on the physical objectives, competing penalties, and any costs or constraints associated with the structure of the laser field. The functional J could contain many terms if the competing physical goals are highly complex. As a specific simple illustration, consider the common goal of steering the system to achieve an expectation value $\langle\psi(T)|O|\psi(T)\rangle$ as close as possible to the target value O_{target} associated with the observable operator O at time T. An additional cost is imposed to minimize the laser fluence. These criteria may be embodied in a cost functional of the form

$$J = \left(\langle\psi(T)|O|\psi(T)\rangle - O_{\text{target}}\right)^2 + \omega\int_0^T \varepsilon^2(t)\,\mathrm{d}t - 2\Im\int_0^T \mathrm{d}t\left\langle\lambda(t)|\mathrm{i}\hbar\frac{\partial}{\partial t} - H_0 + \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)|\psi(t)\right\rangle \qquad [3]$$

Here, the parameter $\omega \geq 0$ is a weight to balance the significance of the fluence term relative to achieving the target expectation value, and $|\lambda(t)\rangle$ serves as a Lagrange multiplier, assuring that Schrödinger's equation is satisfied during the variational minimization of J with respect to the control field. Carrying out the latter minimization will lead to eqn [1], along with the additional relations

$$\mathrm{i}\hbar\frac{\partial}{\partial t}|\lambda(t)\rangle = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]|\lambda(t)\rangle, \quad |\lambda(T)\rangle = 2\sigma O|\lambda(T)\rangle \qquad [4]$$

$$\boldsymbol{\varepsilon}(t) = \frac{1}{\omega}\Im\langle\lambda(t)|\boldsymbol{\mu}|\psi(t)\rangle \qquad [5]$$

$$\sigma = \langle\psi(T)|O|\psi(T)\rangle - O_{\text{target}} \qquad [6]$$

Equations [1] and [4]–[6] embody the OCT design equations that must be solved to yield the optimal field $\boldsymbol{\varepsilon}(t)$ given in eqn [5]. These design equations have an unusual mathematical structure within quantum mechanics. First, insertion of the field expression in eqn [5] into eqns [1] and [4] produces two cubically nonlinear coupled Schrödinger-like equations. These equations are in the same family as the standard nonlinear Schrödinger equation, and as such, one may expect that unusual dynamical behavior could arise. Although eqn [4] for $|\lambda(t)\rangle$ is identical in form to the Schrödinger equation [1], only $|\psi(t)\rangle$ is the true wavefunction, with $|\lambda(t)\rangle$ serving to guide the controlled quantum evolution towards the physical objective. Importantly, eqn [1] is an initial value problem, while eqn [4] is a final value problem. Thus, the two equations together form a two-point boundary value problem in time, which is an inherent feature of temporal engineering optimal control as well. The boundary value nature of these equations typically leads to the existence of multiple solutions, where σ in eqn [6] plays the role of a discrete eigenvalue, specifying the quality of the particular achieved control field design. The fact that there are generally multiple solutions to the laser design equations can be attractive from a physical perspective, as the designs may be sorted through to identify those being most attractive for laboratory implementation.

The overall mathematical structure of eqns [1] and [4]–[6] can be melded together into a single design equation

$$\mathrm{i}\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(\psi,\sigma,t)]|\psi(t)\rangle, \quad |\psi(0)\rangle = |\psi_{\mathrm{i}}\rangle \qquad [7]$$

The structure of this equation embodies the comments in the introduction that control field design is inherently a nonlinear process. The main source of complexity arising in eqn [7] is through the control field $\boldsymbol{\varepsilon}(\psi,\sigma,t)$ depending not only on the wave function at the present time t, but also on the future value of the wavefunction at the target time T contained in σ. This structure again reflects the two-point boundary value nature of the control equations.

Given the typical multiplicity of solutions to eqns [1] and [4]–[6], or equivalently, eqn [7], it is attractive to consider approximations to these equations (except perhaps for the inviolate Schrödinger equation [1], and much work continues to be done along these lines. One case involves what is referred to as tracking control, whereby a path is specified for $\langle\psi(t)|O|\psi(t)\rangle$ evolving from $t = 0$ out to $t = T$. Tracking control eliminates σ to produce a field of the form $\boldsymbol{\varepsilon}(\psi,t)$, thereby permitting the design equations to be explicitly integrated as a forward-marching problem toward the target; the tracking equations are still nonlinear with respect to the evolving state. Many variations on these concepts can be envisioned, and other approximations may also be introduced to deal with special circumstances. One technique appropriate for at least few-level systems is stimulated Raman adiabatic passage (STIRAP), which seeks robust adiabatic passage from an initial state to a particular final state, typically with nearly total destructive interference occurring in the intermediate states. This design technique can have special attractive robustness characteristics with respect to field errors. STIRAP, tracking, and various perturbation theory-based control design techniques can be expressed as special cases of OCT, with suitable cost functionals and constraints. Further approximation methods will surely be developed, with the rigorous OCT concepts forming the general foundation for control field design.

As the OCT equations are inherently nonlinear, their solution typically requires numerical iteration, and a variety of procedures may be utilized for this purpose. Almost all of the methods employ local search techniques (e.g., gradient methods), and some also show monotonic convergence with respect to each new iteration step. Local techniques typically evolve to the nearest solution in the space of possible control fields. Such optimizations can be numerically efficient, although the quality of the attained result can depend on the initial trial for the control field and the details of the algorithm involved. In contrast, global search techniques, such as simulated annealing or genetic algorithms, can search more broadly for the best solution in the control field space. These more expansive searches are attractive, but at the price of typically requiring more intensive computations. Effort has also gone into seeking global or semiglobal input → output maps relating the electric field $\boldsymbol{\varepsilon}(t)$ structure to the observable $O[\boldsymbol{\varepsilon}(t)]$. Such maps may be learned, hopefully, from a modest number of computations with a selected set of fields, and then utilized as high-speed interpolators over the control field space to permit the efficient use of global search algorithms to attain the best control solution possible. At the foundation of all OCT design procedures is the need to solve the Schrödinger equation. Thus, computational technology to improve this basic task is of fundamental importance for designing laser fields.

Many computations have been carried out with OCT to attain control field designs for manipulating a host of phenomena, including rotational, vibrational, electronic, and reactive dynamics of molecules, as well as electron motion in semiconductors. Every system has its own rich details, which in turn, are compounded by the fact that multiple control designs will typically exist in many applications producing comparable physical outcomes. Collectively, these control design studies confirm the manipulation of constructive and destructive quantum wave interferences as the general mechanism for achieving successful control over quantum phenomena. This conclusion may be expressed as the following principle:

> **Control field–system cooperativity principle**: Successful quantum control requires that the field must have the proper structure to take full advantage of all of the dynamical opportunities offered by the system, to best satisfy the physical objective.

This simple statement of system–field cooperativity embodies the richness, as well as the complexity, of seeking control over quantum phenomena, and also speaks to why simple intuition alone has not proved to be a generally viable design technique. Quantum systems can often exhibit highly complex dynamical behavior, including broad dispersion of the wave packet over spatial domains or multitudes of quantum states. Handling such complexity can require fields with subtle structure to interact with the quantum system in a global fashion to manage all of the motions involved. Thus, we may expect that successful control pulses will often have broad bandwidth, including amplitude and phase modulation. Until relatively recently, laser sources with these characteristics were not available, but the technology is now in hand and rapidly evolving (see the discussion later).

The cooperativity principle above is of fundamental importance in the control of all quantum phenomena, and a simple illustration of this principle is shown in Figure 2 for the control of wave packet motion on an excited state of the NO molecule. The target for the control is a narrow wave packet located over the well of the excited B state. The optimal control field consists of two coordinated pulses, at early and late times, with both features having internal structure. The figure indicates that the ensuing excited state wave packet has two components, one of which actually passes through the target region during the initial evolution, only to return again and meet the second component at just the right place and time, to achieve the target objective as best as possible. Similar cooperativity interpretations can be found in virtually all implementations with OCT.

The main reason for performing quantum field design is to ultimately implement the designs in the laboratory. The design procedures must face laboratory realities, which include the fact that most Hamiltonians are not known to high accuracy (especially for polyatomic molecules and complex solid-state structures), and secondly, a variety of laboratory field imperfections may unwittingly be present. Notwithstanding these comments, OCT has been fundamental to the development of quantum control, including laying out the logic for how to perform the analogous optimal control experiments. Design implementations and further OCT development will continue to play a basic role in the quantum control field. At present, perhaps the most important contribution of OCT has been to (1) highlight the basic control cooperativity principle above, and (2) provide the basis for developing algorithms to successfully guide optimal control experiments, as discussed below.

Algorithms for Implementing Optimal Control Experiments

The ultimate purpose of considering control theory within quantum mechanics is to take the matter into

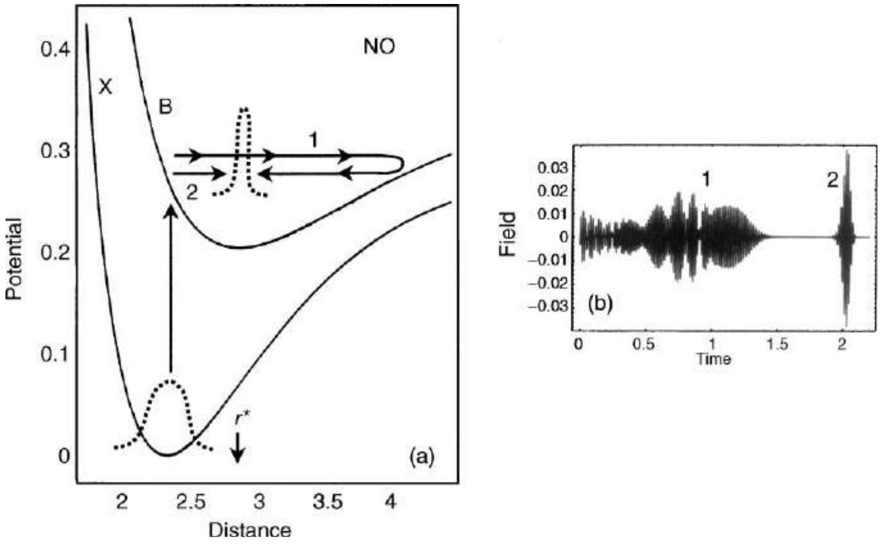

Figure 2 Control of wave packet evolution on the electronically excited B state of NO, with the goal of creating a narrow final packet over the excited state well. (a) indicates that the ground-state packet is brought up in two pieces using the coordinated dual pulse in (b). The piece of the packet from the first pulse actually passes through the target region, then bounces off the right side of the potential, to finally meet the second piece of the packet at just the right time over the target location r^* for successful control. This behavior is an illustration of the control field–system cooperativity principle stated in the text.

the laboratory for implementation and exploitation of its capabilities. Ideally, theoretical control field designs may be attained using the techniques discussed above, followed by the achievement of successful control upon execution of the designs in the laboratory. This appealing approach is burdened with three difficulties: (1) Hamiltonians are often imprecisely known, (2) accurately solving the design equations can be a significant task, and (3) realization of any given design will likely be imperfect due to laboratory noise or other unaccounted-for systematic errors. Perhaps the most serious of these difficulties is point (1), especially considering that the best quality control will be achieved by maximally drawing on subtle constructive and destructive quantum wave interference effects. Exploiting such subtleties will generally require high-quality control designs that, in turn, depend on having reliable Hamiltonians. Although various designs have been carried out seeking robustness with respect to Hamiltonian uncertainties, the issue in point (1) should remain of significance in the foreseeable future, especially for the most complex (and often, the most interesting!) chemical/physical applications. Mitigating this serious problem is the ability to create shaped laser pulse controls and apply them to a quantum system, followed by a probe of their effects at an unprecedented rate of thousands or more independent trials per minute. This unique capability led to the suggestion of partially, if not totally, sidestepping the design process by performing closed-loop experiments to let the quantum system teach the laser how to achieve its control in the laboratory. Figure 3 schematically shows this closed loop process, drawing on the following logic:

1. **The molecular view**. Although there may be theoretical uncertainty about the system Hamiltonian, the actual chemical/physical system under study ‘knows’ its own Hamiltonian precisely! This knowledge would also include any unusual exigencies, perhaps associated with structural or other defects in the particular sample. Furthermore, upon exposure to a control field, the system ‘solves’ its own Schrödinger equation impeccably accurately and as fast as possible in real time. Considering these points, the aim is to replace the offline arduous digital computer control field design effort with the actual quantum system under study acting as a precise analog computer, solving the true equations of motion.
2. **Control laser technology**. Pulse shaping under full computer control may be carried out using even hundreds of discrete elements in the frequency domain controlling the phase and amplitude structure of the pulse. This technology is readily available and expanding in terms of pulse center frequency flexibility and bandwidth capabilities.

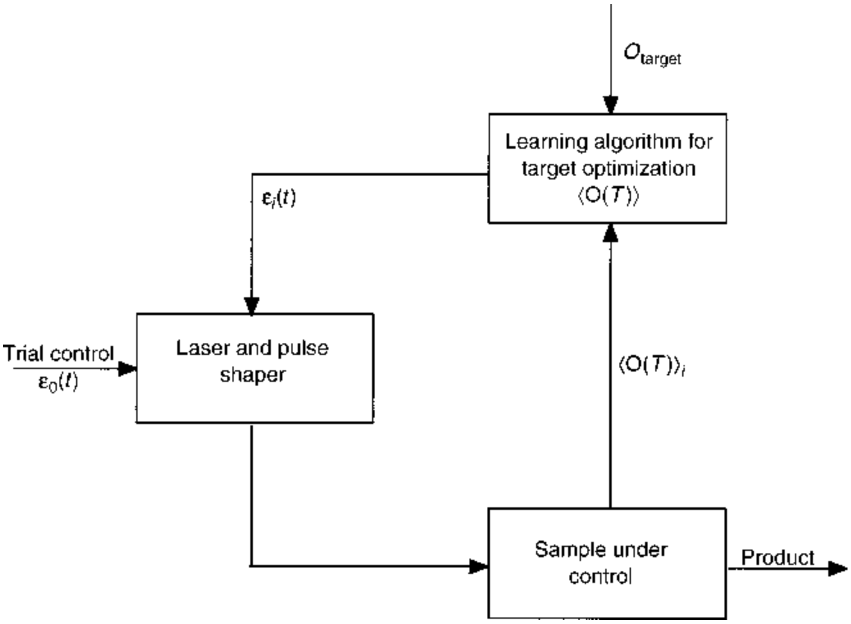

Figure 3 A schematic of the closed-loop concept for allowing the quantum system to teach the laser how to achieve its control. The actual quantum system under control is in a loop with a laser and pulse shaper all slaved together with a pattern recognition algorithm to guide the excursions around the loop. The success of this concept relies on the ability to perform very large numbers of control experiments in a short period of time. In principle, no knowledge of the system Hamiltonian is required to steer the system to the desired final objective, although a good trial design $\varepsilon_0(t)$ may accelerate the process. The ith cycle around the loop attempts to find a better control field $\varepsilon_i(t)$ such that the system response $\langle O(T)\rangle_i$ for the observable operator O comes closer to the desired value O_{target}.

3. **Quantum mechanical objectives.** Many chemical/physical objectives may be simply expressed as the desire to steer a quantum mechanical flux out one clearly defined channel versus another.
4. **Detection of the control action**. Detection of the control outcome can often be carried out by a second laser pulse or any other suitable high duty cycle detection means, such as laser-induced mass spectrometry. The typical circumstance in point (3) implies that little, if any, time-consuming offline data analysis will be necessary beyond that of simple signal averaging.
5. **Fast learning algorithms**. All of the four points above may be slaved together using pattern recognition learning algorithms to identify those control fields which are producing better results, and bias the next sequence of experiments in their favor for further improvement. Although many algorithms may be employed for this purpose, the global search capabilities of genetic algorithms are quite attractive, as they may take full advantage of the high throughput nature of the experiments.

In linking together all of these components into a closed-loop OCE, it is important that no single operational step significantly lags behind any other, for efficiency reasons. Fortunately, this economy is coincident with all the tasks and technologies involved. For achieving control alone, there is no requirement that the actual laser pulse structure be identified on each cycle of the loop. Rather, the laser control 'knob' settings are adequate information for the learning algorithm, as the only criterion is to suitably adjust the knobs to achieve an acceptable value for the chemical/physical objective. The learning algorithm in point (5) operates with a cost functional J, analogous to that used for computational design of fields, but now the cost functional can only depend on those quantities directly observable in the laboratory (i.e., minimally, the current achieved target expectation value). In cases where a physical understanding is sought about the control mechanism, the laser field structure must also be measured. This additional information is typically also not a burden to attain, as it often is only required for the single best control field at the end of the closed-loop learning experiments.

The OCE process in Figure 3 and steps (1)–(5) above constitute the laboratory process for achieving optimal quantum control, and exactly the same reasons put forth for OCT motivate the desire for attaining optimal performance: principally, seeking the best control result that can possibly be obtained. Seeking optimal performance also ensures some degree of inherent robustness to field noise, as fleeting observations would not likely survive the signal averaging carried out in the laboratory. Additionally, robustness as a specific criterion may also be included in the laboratory cost functional in item (5). When a detailed physical interpretation of the controlled dynamics is desired, it is essential to remove any extraneous control field features that have little impact on the system manipulations. Without the latter clean-up carried out during the experiments, the final field may be contaminated by structures that have little physical impact. Although quantum dynamics typically occurs on femtosecond or picosecond time-scales, the loop excursions in Figure 3 need not be carried out on the latter time-scales. Rather, the loop may be traversed as rapidly as convenient, consistent with the capabilities of the apparatus. A new system sample is introduced on each cycle of the loop. This process is referred learning control, to distinguish it from real-time feedback control.

The closed-loop quantum learning control algorithm can be self-starting, requiring no prior control field designs, under favorable circumstances. Virtually all of the current experiments were carried out in this fashion. It remains to be seen if this procedure of ‘going in blind’ will be generally applicable in highly complex situations where the initial state is far from the target state. Learning control can only proceed if at least a minimal signal is observed in the target state. Presently, this requirement has been met by drawing on the overwhelmingly large number of exploratory experiments that can be carried out, even in a brief few minutes, under full computer control. However, in some situations, the performance of intermediate observations along the way to the target goal may be necessary in order to at least partially guide the quantum dynamics towards the ultimate objective. Beneficial use could also be made from prior theoretical laser control designs $\boldsymbol{\varepsilon}_0(t)$ capable of at least yielding a minimal target signal.

The OCE closed-loop procedure was based on the growing number of successful theoretical OCT design calculations, even with all of their foibles, especially including less than perfect Hamiltonians. The OCT and OCE processes are analogous, with their primary distinction involving precisely what appears in their corresponding cost functionals. Theoretical guidance can aid in identifying the appropriate cost functionals and learning algorithms for OCE, especially for addressing the needs associated with attaining a physical understanding about the mechanisms governing controlled quantum dynamics phenomena.

A central issue is the rate that learning control occurs upon excursions around the loop in Figure 3. A number of factors control this rate of learning, including the nature of the physical system, the choice of objective, the presence of field uncertainties and measurement errors, the number of control variables, and the capabilities of the learning algorithm employed. Achieving control is a matter of discrimination, and some of these factors, whether inherent to the system or introduced by choice, may work against attaining good-quality discrimination. At this juncture, little is known quantitatively about the limitations associated with any of these factors.

The latter issues have not prevented the successful performance of a broad variety of quantum control experiments, and at present the ability to carry out massive numbers of closed-loop excursions has overcome any evident difficulties. The number of examples of successful closed-loop quantum control is rapidly growing, with illustrations involving the manipulation of laser dye fluorescence, chemical reactivity, high harmonic generation, semiconductor optical switching, fiber optic pulse transmission, and dynamical discrimination of similar chemical species, amongst others. Perhaps the most interesting cases are those carried out at high field intensities, where prior speculation suggested that such experiments would fail due to even modest laser field noise being amplified by the quantum dynamics. Fortunately, this outcome did not occur, and theoretical studies have indicated that surprising degrees of robustness to field noise may accompany the manipulation of quantum dynamics phenomena. Although the presence of field noise may diminish the beneficial influence of constructive and destructive interferences, evidence shows that field noise does not appear to kill the actual control process, at least in terms of obtaining reasonable values for the control objectives. One illustration of control in the high-field regime is shown in Figure 4, demonstrating the learning process for the dissociative rearrangement of acetophenone to form toluene. Interestingly, this case involves breaking two bonds, with the formation of a third, indicating that complex dynamics can be managed with suitably tailored laser pulses.

Operating in the strong-field regime, especially for chemical manipulations, also has the important benefit of producing a generic laser tool to control a broad variety of systems by overcoming the long-standing problem of having sufficient bandwidth. For example, the result in Figure 4 uses a Ti:sapphire laser in the near-infrared regime. Starting with a bandwidth-limited pulse of intensity near

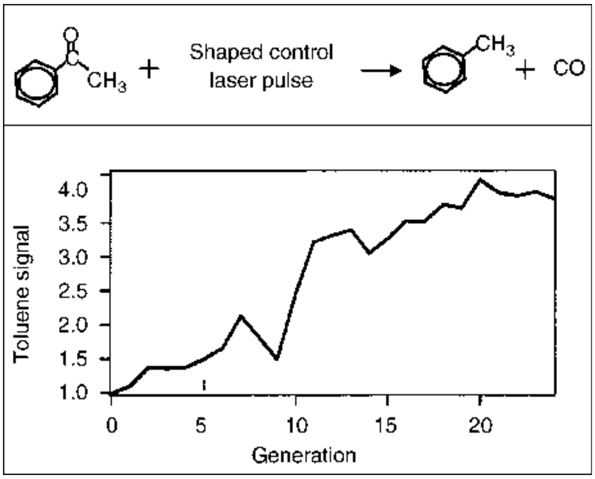

Figure 4 An illustration of dissociative rearrangement achieved by closed-loop learning control in Figure 3. The optimally deduced laser pulse broke two bonds in the parent acetophenone molecule and formed a new one to yield the toluene product. The laboratory learning curve is shown for the toluene product signal as a function of the generations in the genetic algorithm guiding the experiments. The fluctuations in the learning curve for optimizing the toluene yield corresponds to the algorithm searching for the optimal yield as the experiments proceed.

$\sim 10^{14}$ W cm^{-2}, in this regime, the dynamic power broadening can easily be on the order of ~ 1 eV or more, thereby effectively converting the otherwise discrete spectrum of the molecule into an effective continuum for ready multiphoton matching by the control laser pulse. Operating under these physical conditions is very attractive, as the apparatus is generic, permitting a single shaped-pulse laser source (e.g., a Ti:sapphire system with phase and amplitude modulation) to be utilized with virtually any chemical system where manipulations are desired. Although every desired physical/chemical goal may not be satisfactorily met (i.e., the goals must be physically attainable!), the means are now available to explore large numbers of systems. Operating under closed loop in the strong-field tailored-pulse regime eliminates the prior serious limitation of first finding a physical system to meet the laser capabilities; now, the structure of the laser pulse can be shaped to meet the molecule's characteristics. Strong field operations may be attractive for these reasons, but other broadband laser sources, possibly working at weaker field intensities, might be essential in some applications (e.g., controlling electron dynamics in semiconductors, where material damage is to be avoided). It is anticipated that additional broadband laser sources will become available for these purposes.

The coming years should be a period of rapid expansion, including a thorough exploration of closed-loop quantum control capabilities. As with the use of any laboratory tool, certain applications may be more amenable than others to attaining successful control. The particular physical/chemical questions need to be well posed, and controls need to have sufficient flexibility to meet the objectives. The experiments ahead should be able to reveal the degree to which drawing on optimality in OCE, combined with the performance of massive numbers of experiments can lead to broad-scale successful control of quantum phenomena. One issue of concern is the richness associated with the large numbers of phase and amplitude control knobs that may be adjusted in the laboratory. Some experiments have already operated with hundreds of knobs, while others have restricted their number in a variety of ways, to simplify the search process. Additional technological and algorithmic advances may be required to manage the high-dimensional control space searches. Fortunately, for typical applications, the search does not reduce to seeking a needle in a haystack, as generally there are multiple control solutions, possibly all of very good quality.

As a final comment on OCE, it is useful to appreciate the subtle distinction of the procedure from closed-loop feedback experiments. This distinction is illustrated in Figure 5, pointing to three types of closed-loop quantum control laboratory experiments. The OCE procedure in Figure 5a produces a generic laser tool capable of controlling a broad variety of systems with an emphasis in the figure placed on the point that each cycle around the loop starts with a new sample for control. The replacement of the sample on each cycle eliminates a number of difficulties, principally including concerns about sample damage, avoidance of operating the loop at the ultrafast speed of the quantum mechanical processes, and elimination of the effect that 'to observe is to disturb' in quantum mechanics. Thus, learning control provides a generic practical procedure, regardless of the nature of the quantum system. Figure 5b has the same structure as that of Figure 5a, except the loop is now closed around the same single quantum system, which is followed throughout its evolution. All of the issues mentioned above that are circumvented by operating through laboratory learning control in Figure 5a must now be directly faced in the setup of Figure 5b. The procedure in Figure 5b will likely only be applicable in special circumstances, at least for the reason that many quantum systems operate on time-scales far too fast for opto-electronics and computers to keep up with. However, there are certain quantum mechanical processes that are sufficiently slow to meet the criteria. Furthermore, a period of free evolutionary

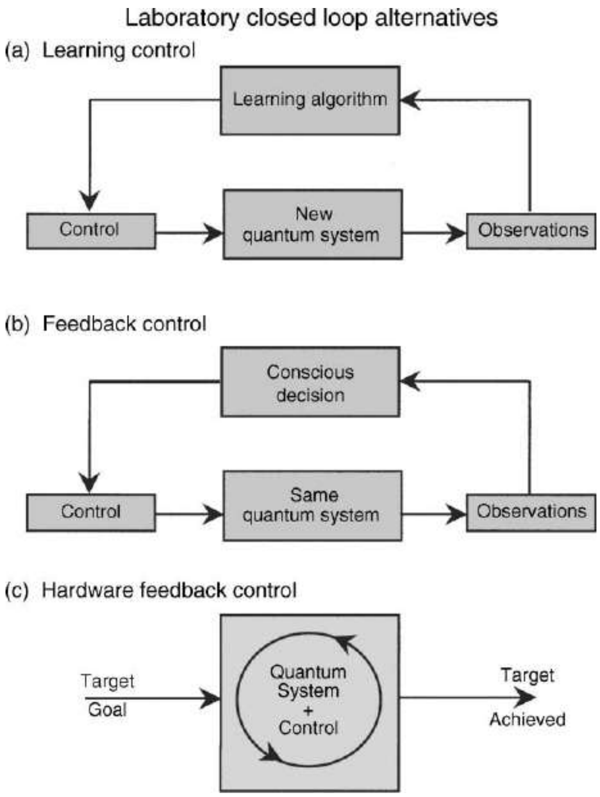

Figure 5 Three possible closed-loop formulations for attaining laser control over quantum phenomena. (a) is a learning control process where a new system is introduced on each excursion around the loop. (b) utilizes feedback control with the loop being traversed with the same sample, generally calling for an accurate model of the system to adjust the controls on each loop excursion. (c) is, at present, a dream machine where the laser and the sample under control are one unit operating without external algorithmic guidance. The closed-loop learning control procedure in (a) appears to form the most practical means to achieve laser control over quantum phenomena, especially for complex systems.

quantum dynamics may be permitted to occur between one control pulse and the next, to make the time issue manageable. While the learning control process in Figures 3 and 5a can be performed model-free, the feedback algorithm in Figure 5b generally must operate with a sufficiently reliable system Hamiltonian to carry out fast design corrections to the control field, based on the previous control outcome. These are very severe demands, but they may be met under certain circumstances. One reason for considering closed-loop feedback control in Figure 5b is to explore the basic physical issue of quantum mechanical limitations inherent in the statement ‘to observe is to disturb’. It is suggestive that control over many types of quantum phenomena may fall into the semiclassical regime, lying somewhere between classical engineering behavior and the hard limitations of quantum mechanics. Experiments of the type in Figure 5b will be most interesting for exploring this matter. Finally, Figure 5c introduces a *gedanken* experiment in the sense that the closed-loop process is literally built into the hardware. That is, the laser control and quantum system act as a single functioning unit operating in a stable fashion, so as to automatically steer the system to the desired target. This process may involve engineering the quantum mechanical system prior to control in cases where that freedom exists, as well as engineering the laser components involved. The meaning of such a device in Figure 5c can be understood by considering an analogy with airplane flight, where the aircraft is constructed to have an inherent degree of aerodynamic stability and will essentially fly (glide) on its own accord when pushed forward. It is an open question whether closing the loop in the hardware can be attained for quantum control, and an exploration of this concept may require additional laser technologies.

Conclusions

This article presented an overview of theoretical concepts and algorithmic considerations associated with the control of quantum phenomena. Theory has played a central role in this area by revealing the fundamental principles underlying quantum control, as well as by providing algorithms for designing controls and guiding experiments to discover successful controls in the laboratory. The challenge of controlling quantum phenomena, in one sense, is an old subject, going back at least 40 years. However, roughly the first 30 of these years would best be described as a period of frustration, due to a lack of full understanding of the principles involved and the nature of the lasers needed to achieve success. Thus, from another perspective, the subject is quite young, with perhaps the most notable development being the introduction of closed-loop laboratory learning procedures. At the time of writing this article, these procedures are just beginning to be explored for their full capabilities. To appreciate the young nature of this subject, it is useful to note that the analogous engineering control disciplines presently occupy many thousands of engineers worldwide, both theoreticians and practitioners, and have done so for many years. Yet, engineering control is far from considered a mature subject. Armed now with the basic concepts and proper laboratory tools, one may anticipate a thorough exploration of control over quantum phenomena, including its many possible applications.

Acknowledgments

Support for this work has come from the National Science Foundation and the Department of Defense.

Further Reading

Brixner T, Damrauer NH and Gerber G (2001) Femtosecond quantum control. *Advances in Atomic, Molecular, and Optical Physics* 44: 1–54.

Rabitz H and Zhu W (2000) Optimal control of molecular motion: Design, implementation and inversion. *Accounts of Chemical Research* 33: 572–578.

Rabitz H, de Vivie-Riedle R, Motzkus M and Kompa K (2000) Whither the future of controlling quantum phenomena? *Science* 288: 824–828.

Rice SA and Zhao M (2000) Optical Control of Molecular Processes. NewYork: John Wiley.

Experimental

R J Levis, Temple University, Philadelphia, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The experimental control of quantum phenomena in atoms, molecules, and condensed phase materials has been a long-standing challenge in the physical sciences. For example, the idea that a laser beam could be used to selectively cleave a chemical bond has been pursued since the demonstration of the laser in the early 1960s. However, it was not until very recent times that one could realize selective bond cleavage. One key to the rapid acceleration of such experiments is the capability of manipulating the relative phases between two or more paths leading to the same final state. This is done in practice by controlling the phase of two (or two thousand!) laser frequencies coupling the initial state to some desired final state. The present interest in phase-related control stems in part from research focusing on quantum information sciences, controlling chemical reactivity, and developing new optical technologies, such as in biophotonics for imaging cellular materials. Ultimately, one would like to identify coherent control applications having impact similar to linear regime applications like compact disk reading (and writing), integrated circuit fabrication, and photodynamic therapy. Note that in each of these the phase of the electromagnetic field is not an important parameter and only the brightness of the laser at a particular frequency is used. While phase is of negligible importance in any one photon process, for excitations involving two or more photons, phase not only becomes a useful characteristic, it can modulate the yield of a desired process between zero and one hundred percent.

The use of optical phase to manipulate atomic and molecular systems rose to the forefront of laser science in the 1980s. Prior to this there was considerable effort directed toward using the intensity and high spectral resolution features of the laser for various applications. For example, the high spectral resolution character was employed in purifying nuclear isotopes. Highly intense beams were also used in the unsuccessful attempt to pump enough energy into a single vibrational resonance to selectively dissociate a molecule. The current revolution in coherent control has been driven in part by the realization that controlling energy deposition into a single degree of freedom (such as selectively breaking a bond in a polyatomic molecule) can not be accomplished by simply driving a resonance with an ever more intense laser source. In the case of laser-induced photo dissociation, such excitation ultimately results in statistical dissociation throughout the molecule due to vibrational mode coupling. In complex systems, controlling energy deposition into a desired quantum state requires a more sophisticated approach and one scheme involves coherent control.

The two distinct regimes for excitation in coherent control, the time and frequency domain experiments, are formally equivalent. While the experimental approach and level of success for each domain is vastly different, both rely on precisely specifying the phase relations for a series of excitation frequencies. The basic idea behind phase control is simple. One emulates the interference effects that can be easily visualized in a water tank experiment with a wave propagating through two slits. In the coherent control experiment, two or more indistinguishable pathways must be excited that connect one quantum state to another. In either water waves, or coherent control, the probability for observing constructive or destructive interference at a target state depends on the phase difference between the paths connecting the two locations in the water tank, or the initial and final states in the quantum system. In the case of coherent control, the final state may have some technological value, for instance, a desirable chemical reaction product or an intermediate in the implementation of quantum computation.

Frequency domain methods concentrate on interfering two excitation routes in a system by controlling the phase difference between two distinct frequencies coupling an initial state to a final state. Perhaps the first mention of the concept of phase control in the optical regime can be traced to investigations in Russia in the early 1960s, where the interference between a one- and three-photon path was theoretically proposed as a means to modulate a two-photon absorption. This two-path interference idea lay dormant until the reality of controlling the relative phase of two distinct frequencies was achieved in 1990 using gas phase pressure modulation for frequency domain experiments. The idea was theoretically extended to the control of nuclear motion in 1988. The first experimental example of two-color phase control over nuclear motion was demonstrated in 1991. It is interesting that the first demonstrations of coherent control did not, however, originate in the frequency domain approach, but came from time domain investigations in the mid-1980s. This was work that grew directly out of the chemical dynamics community and will be described below.

While frequency domain methods were developed exclusively for the observation of quantum interference, the experimental implementation of time-domain methods was first developed for observing nuclear motion in real time. In the time domain, a superposition of multiple states is prepared in a system using a short duration (~ 50 fs) laser pulse. The superposition state will subsequently evolve according to the phase relationships of the prepared states and the Hamiltonian of the system. At some later time the evolution must be probed using a second, short-time-duration laser pulse. The time-domain methods for coherent control were first proposed in 1985 and were based on wavepacket propagation methods developed in the 1970s. The first experiments were reported in the mid-1980s.

To determine whether the time or frequency domain implementation is more applicable for a particular control context, one should compare the relevant energy level spacing of the quantum system to the bandwidth of the excitation laser. For instance, control of atomic systems is more suited to frequency domain methods because the characteristic energy level spacing in an atom (several electron volts) is large compared to the bandwidth of femtosecond laser pulses (millielectron volts). In this case, preparation of a superposition of two or more states is impractical with a single laser pulse at the present time, dictating that the control scheme must employ the initial state and a single excited eigenstate. In practice, quasi-CW nanosecond duration laser pulses are employed for such experiments because current ultrafast (fs duration) laser sources cannot typically prepare a superposition of electronic states. One instructive exception involves the excitation of Rydberg states in atoms. Here the electronic energy level spacing can be small enough that the bandwidth of a femtosecond excitation laser may span several electronic states and thus can create a superposition state. One should also note that the next generation of laser sources in the attosecond regime may permit the preparation of a wavepacket from low-lying electronic states in an atom. In the case of molecules, avoiding a superposition state is nearly impossible. Such experiments employ pico- to femtosecond duration laser pulses, having a bandwidth sufficient to excite many vibrational levels, either in the ground or excited electronic state manifold of a molecule. To complete the time-dependent measurement, the propagation of the wavepacket to the final state of interest must be observed (probed) using a second ultrafast laser pulse that can produce fluorescence, ionization, or stimulated emission in a time-resolved manner. The multiple paths that link the initial and final state via the pump and the probe pulses are the key to the equivalence of the time and frequency domain methods.

There is another useful way to classify experiments that have been performed recently in coherent control, that is into either open-loop or closed-loop techniques. Open-loop signifies that a calculation may be performed to specify the required pulse shape for control before the experiment is attempted. All of the frequency-based, and most of the time-based experiments fall into this category. Closed-loop, on the other hand, signifies that the results from experimental measurements must be used to assist in determining the correct parameters for the next experiment, this pioneering approach was suggested in 1992. It is interesting that in the context of control experiments performed to date, closed-loop experiments represent the only route to determining the optimal laser pulse shape for controlling even moderately complex systems. In this approach, no input from theoretical models is required for coherent control of complex systems. This is unlike typical control engineering environments, where closed-loop signifies the use of experiments to refine parameters used in a theoretical model.

There is one additional distinction that is of value for understanding the state of coherent control at the present time. Experiments may be classified into systems having Hamiltonians that allow calculation of frequencies required for the desired interference patterns (thus enabling open-loop experiments) and those having ill-defined Hamiltonians (requiring closed-loop experiments). The open-loop experiments have demonstrated the utility of various control schemes, but are not amenable to systems having even moderate complexity. The closed-loop experiments are capable of controlling systems of moderate complexity but are not amenable to precalculation of the required control fields. Systems requiring closed-loop methods include propagation of short pulses in an optical fiber, control of high harmonic generation for soft X-ray generation, control of chemical reactions and control of biological systems. In the experimental regime, the value of learning control for dealing with complex systems has been well documented.

The remainder of this article represents an overview of the experimental implementation of coherent control. The earliest phase control experiments involved time-dependent probing of molecular wavepackets, and were followed by two-path interference in an atomic system. The most recent, and general, implementation of control involves tailoring a time-dependent electromagnetic field for a desired objective using closed-loop techniques. Since the use of tailored laser pulses appears to be rather general, a more complete description of the experiment is provided.

Time-Dependent Methods

The first experiments demonstrating the possibility of coherent control were performed to detect nuclear motion in molecules in real time. These investigations were the experimental realization of the pump-dump or pulse-timing control methods as originally described in 1985. The experimental applications have expanded to numerous systems including diatomic and triatomic molecules, small clusters, and even biological molecules. For such experiments, a vibrational coherence is prepared by exciting a superposition of states. These measurements implicitly used phase control both in the generation of the ultrashort pulses and in the coherent probing of the superposition state by varying the time delay between the pump and probe pulses, in effect modulating the outcome of a quantum transition. Since the earliest optical experiment in the mid-1980s, there have been many hundreds of such experiments reported in the literature. The motivation for the very first experiments was not explicitly phase control, but rather the observation of nuclear motion of molecules in real time. In such later measurements, oscillations in the superposition states were observable for many cycles, up to many tens of picoseconds in favorable cases, i.e., diatomic molecules in the gas phase. The loss of signal in the coherent oscillations is ultimately due to passage of the superposition state in to surrounding bath states. This may be due to collision with another molecule, dissociation of the system or redistribution of the excitation energy into other modes of the system not originally excited in the superposition. It is notable that coherence can be maintained in solution phase systems for tens of picoseconds and for the case of resonances having weak coupling to other modes, coherence can even be modulated in large biological systems on the picosecond time scale.

Decoherence represents loss of phase information about a system. In this context, phase information represents our detailed knowledge of the electronic and nuclear coordinates of the system. In the case of vibrational coherence, decoherence may be represented by intramolecular vibrational energy transfer to other modes of the molecule. This can be thought of as the propensity of vibrational modes of a molecule to couple together in a manner that randomizes deposited energy throughout the molecule. Decoherence in a condensed phase system includes transfer of energy to the solvent modes surrounding the system of interest. For an electronically excited molecule or atom, decoherence can involve spontaneous emission of radiation (fluorescence), or dissociation in the case of molecules. Certain excited states may have high probability for fluorescence and would represent a significant decoherence pathway and an obstacle for coherent control. These states may be called 'lossy' and are often used in optical pumping schemes as well as for stimulated emission pumping. An interesting question arises as to whether one can employ such states in a pathway leading to a desired final state, or must such lossy states be avoided at all costs in coherent control. This question has been answered to some degree by the method of rapid adiabatic passage. In this experiment, an initial state is coupled to a final state by way of a lossy state. The coupling is a coherent process and the preparation involves dressed states that are eigenstates of the system. In this coherent superposition the lossy state is employed, but no population is allowed to build up and thus decoherence is circumvented.

Two-Path Interference Methods

Interference methods form perhaps the most intuitive example of coherent control. The first experimental demonstration was reported in 1990 and involved the modulation of ionization probability in Hg by interfering a one-photon and three-photon excitation pathway to an excited electronic state (followed by two-photon ionization). The long delay between the introduction of the two-path interference concept in 1960 and the demonstration in 1990 was the difficulty in controlling the relative phase of the one- and three-photon paths. The solution involved generating the third harmonic of the fundamental in a nonlinear crystal, and copropagating the two beams through a low-density gas. Changing the pressure of the gas changes the optical phase retardance at different rates for different frequencies, thus allowing relative phase control between two colors. The ionization yield of Hg was clearly modulated as a function of pressure. Such schemes have been extended to diatomic molecular systems and the concept of determining molecular phase has been investigated. While changing the phase of more than two frequencies has not been demonstrated using pressure tuning (because of experimental difficulties), a more flexible method has been developed to alter relative phase of up to 1000 frequency bands, as will be described next.

Closed-Loop Control Methods

Perhaps the most exciting new aspect of control in recent years concerns the prospect for performing closed-loop experiments. In this approach a more complex level of control in matter is achieved in comparison to the two-path interference and pump-probe control methods. In the closed-loop method, an arbitrarily complex time-dependent electromagnetic field is prepared that may have multiple subpulses and up to thousands of interfering frequency bands. The realization of closed-loop control over complex processes relies on the confluence of three technologies: (i) production of large bandwidth ultrafast laser pulses; (ii) spatial modulation methods for manipulating the relative phases and amplitudes of component frequencies of the ultrafast laser pulse; and (iii) closed-loop learning control methods for sorting through the vast number of pulse shapes available as potential control fields. In the closed-loop method, the result of a laser molecule interaction is used to tailor an optimized pulse.

In almost all of the experimental systems used for closed-loop control today, Kerr lens mode-locking in the Ti:sapphire crystal is used to phase lock the wide bandwidth of available frequencies (750–950 nm) to create the ultrafast pulse. In this phenomenon, the intensity variation in the short laser pulse creates a transient lens in the Ti:sapphire lasing medium that discriminates between free lasing and mode-locked pulses. As a result, all of the population inversion available in the lasing medium may be coerced into enhancing the ultrashort traveling wave in the cavity in this way. The wide bandwidth available may then be amplified and tailored into a shaped electromagnetic field. Spatial light modulation is one such method for modulating the relative phases and amplitudes of the component frequencies in the ultrafast laser pulse to tailor the shaped laser pulse. Pulse shaping first involves dispersing the phase-locked frequencies on an optical grating, the dispersed radiation is then collimated using a cylindrical lens and the individual frequencies are focused to a line forming the Fourier plane. At the Fourier plane, the time-dependent laser pulse is transformed into a series of phase-locked continuous wave (CW) laser frequencies, and thus Fourier transformed from time to frequency space. The relative phases and amplitudes may be modulated in the Fourier plane using an array of liquid crystal pixels. After altering the spectral phase and amplitude profile, the frequencies are recombined on a second grating to form the shaped laser pulse. With one degree of phase control and 10 pixels there are an astronomic number (360^{10}) of pulse shapes that may be generated in this manner. (At the present time, pulse shapers have up to 2000 independent elements.) Evolutionary algorithms are employed to manage the available phase space. Realizing that the pulse shape is nothing more than the summation of a series of sine waves, each having an associated phase and amplitude, we find that a certain pulse shape can be represented by a genome consisting of an array of these frequency-dependent phases and amplitudes. This immediately suggests that the methods of evolutionary search strategies may be useful for determining the optimal pulse shape for a desired photoexcitation process.

The method of closed-loop optimal control for experiments was first proposed in 1992 and the first experiments were reported in 1997 regarding optimization of the efficiency of a laser dye molecule in solution. Since that time a number of experiments have been performed in both the weak and strong field regime. In the weak field regime, the intensity of the laser does not alter the field free structure of the excited states of the system under investigation. In the strong field, the field free states of the system are altered by the electric field of the laser. Whether one is in the weak or strong field regime depends on parameters including the characteristic level spacing of the states of the system to be controlled and the intensity of the laser coupling into those states. Examples of weak field closed-loop control include optimization of laser dye efficiency, compression of an ultrashort laser pulse, dissociation of alkali clusters, optimization of coherent anti-Stokes

Raman excitation, and optical pulse propagation in fibers. In the strong field, the laser field is used to manipulate the excited states of the molecule as well as induce population transfer among these states. In this sense, the laser is both creating and exciting resonances, in effect allowing universal excitation with a limited frequency bandwidth of 750–850 nm. For chemical applications, it is worth noting that most molecules do not absorb in this wavelength region in the weak field. Strong field control has been used for controlling bond dissociation, chemical rearrangement, photonic reagents, X-ray generation, molecular centrifuges and the manipulation of mass spectral fragmentation intensities. In the latter case, a new sensing technology has emerged wherein each individual pulse shape represents an independent sensor for a molecule. Given the fact that thousands of pulse shapes can be tested per minute, this represents a new paradigm for molecular analysis.

At the time of this writing, the closed-loop method for adaptively tailoring control fields, has far outstripped our theoretical understanding of the process (particularly in the strong field regime). Adaptive control is presently an active field of investigation, encompassing the fields of physics, engineering, optics, and chemistry. In the coming years, there will be as much work done on the mechanism of control as in the application of the methods. Due to small energy level spacings and complexity of the systems, we anticipate the majority of applications to be in the chemical sciences.

See also

Coherence: Coherence and Imaging; Overview. **Interferometry:** Overview. **Microscopy:** Interference Microscopy.

Further Reading

Brumer PW and Shapiro M (2003) *Principles of the Quantum Control of Molecular Processes*. New York: Wiley Interscience.

Chen C and Elliott DS (1990) Measurement of optical phase variations using interfering multiphoton ionization processes. *Physics Review Letters* 65: 1737.

Levis RJ and Rabitz H (2002) Closing the loop on bond selective chemistry using tailored strong field laser pulses. *Journal of Physical Chemistry A* 106: 6427–6444.

Levis RJ, Menkir GM and Rabitz H (2001) Selective bond dissociation and rearrangement with optimally tailored, strong-field laser pulses. *Science* 292(5517): 709–713.

Rice SA and Zhao M (2000) *Optical Control of Molecular Dynamics*. New York: Wiley Interscience.

Stolow A (2003) Femtosecond time-resolved photoelectron spectroscopy of polyatomic molecules. *Annual Review of Physical Chemistry* 54: 89–119.

Zewail AH, Casati G, Rice SA, *et al.* (1997) Femtochemistry: Chemical reaction dynamics and their control. *Chemical Reactions and Their Control on the Femtosecond Time Scale, Solvay Conference on Chemistry* 101: 3–45.

Applications in Semiconductors

H M van Driel and J E Sipe, University of Toronto, Toronto, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interference phenomena are well-known in classical optics. In the early 19th century, Thomas Young showed conclusively that light has wave properties, with the first interference experiment ever performed. He passed quasi-monochromatic light from a single source through a pair of double slits using the configuration of Figure 1a and observed an interference pattern consisting of a series of bright and dark fringes on a distant screen. The intensity distribution can be explained only if it is assumed that light has wave or phase properties. In modern terminology, if E_1 and E_2 are the complex electric fields of the two light beams arriving at the screen, by the superposition principle of field addition the intensity at a particular point on the screen can be written as

$$I \propto |E_1 + E_2|^2 = |E_1|^2 + |E_2|^2 + |E_1||E_2|\cos(\phi_1 - \phi_2) \qquad [1]$$

where $\phi_1 - \phi_2$ is the phase difference of the beams arriving at the screen. While this simple experiment was used originally to demonstrate the wave properties of light, it has subsequently been used for many other purposes, such as measuring the wavelength of light. However, for our purposes, here the Young's double slit apparatus can also be viewed as a device that redistributes light, or

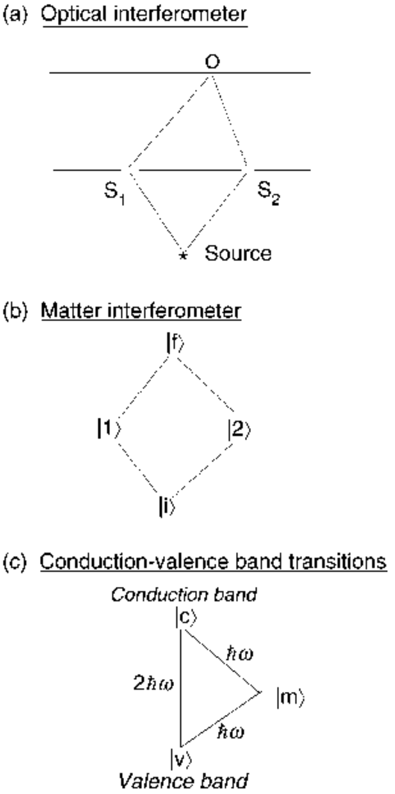
(a) Optical interferometer
O
S1
S2
Source
(b) Matter interferometer
|f⟩
|1⟩
|2⟩
|i⟩
(c) Conduction-valence band transitions
Conduction band
|c⟩
ħω
2ħω
|m⟩
ħω
|v⟩
Valence band

Figure 1 (a) Interference effects in the Young's double slit experiment; (b) illustration of the general concept of coherence control via multiple quantum mechanical pathways; (c) interference of single- and two-photon transitions connecting the same valence and conduction band states in a semiconductor.

controls its intensity at a particular location. For example, let's consider one of the slits to be a source, with the other slit taken to be a gate, with both capable of letting through the same amount of light. If the gate is closed, the distant screen is nearly uniformly illuminated. But if the gate is open, at a particular point on the distant screen there might be zero intensity or as much as four times the intensity emerging from one slit because of interference effects, with all the light merely being spatially redistributed. In this sense the double slit system can be viewed as a device to redistribute the incident light intensity. The key to all this is the superposition principle and the properties of optical phase.

The superposition principle and interference effects also lie at the heart of quantum mechanics. For example, let's now consider a system under the influence of a perturbation with an associated Hamiltonian that has phase properties. In general the system can evolve from one quantum state $|i\rangle$ to another $|f\rangle$ via multiple pathways involving intermediate states $|m\rangle$. Because of the phased perturbation, interference between those pathways can influence the system's final state. If a_m is the (complex) amplitude associated with a transition from the initial to the final state, via intermediate virtual state $|m\rangle$, then for all possible intermediate states the overall transition probability, W, can be written as

$$W = \left| \sum_m a_m \right|^2 \qquad [2]$$

In the case of only two pathways, as illustrated in Figure 1b, W becomes

$$\begin{aligned} W &= |a_1 + a_2|^2 \\ &= |a_1|^2 + |a_2|^2 + |a_1||a_2| \cos(\phi_1 - \phi_2) \end{aligned} \qquad [3]$$

where ϕ_1 and ϕ_2 are now the phases of the two transition amplitudes. While this expression strongly resembles that of eqn [1], here ϕ_1 and ϕ_2 are influenced by the phase properties of the perturbation Hamiltonian that governs the transition process; these phase factors arise, for example, if light fields constitute the perturbation. But overall it is clear that the change in the state of the system is affected by both the amplitude and phase properties of the perturbation. Analogous with the classical interferometer, equality of the transition amplitudes leads to maximum contrast in the transition rate. Although it has been known since the early days of quantum mechanics that the phase of a perturbation can influence the evolution of a system, generally phase has only been discussed in a passive role. Only in recent years has phase been used as a control parameter for a system, on the same level as the strength of the perturbation itself.

Coherence Control

Coherence control, or quantum control as it is sometimes called, refers to the active process through which one can use phase-dependent perturbations originating with, for example, coherent light waves, to control one or more properties of a quantum system, such as state population, momentum, or spin. This more general perspective of interference leads to a picture of interference of matter waves, rather than simply light waves, and one can now speak of an effective 'matter interferometer'. Laser beams, in particular, are in a unique position to play a role in such processes, since they offer a macroscopic 'phase handle' with which to create such effects. This was recognized by the community of atomic and molecular scientists who emphasized the active manifestations of quantum interference effects, and proposed that branching ratios in photochemical reactions might be controlled through laser-induced interference processes. The use of phase as a control parameter also represents a novel horizon of applications for the laser since most previous applications had involved only amplitude (intensity).

Of course, just as the 'visibility' of the screen pattern for a classical Young's double slit interferometer is governed by the coherence properties of the two slit source, the evolution of a quantum system reflects the coherence properties of the perturbations, and the degree to which one can exercise phase control via external perturbations is also influenced by the interaction of the system of interest with a reservoir – essentially any other degrees of freedom in the system – which can cause decoherence and reduce the effectiveness of the 'matter interferometer'. Since the influence of a reservoir will generally increase with the complexity of a system, one might think that coherence control can only be effectively achieved in simple atomic and molecular systems. Indeed, some of the earliest suggestions for coherence control of a system involved using interference between single and three photon absorption processes, connecting the same initial and final states with photons of frequency 3ω and ω, respectively, and controlling the population of excited states in atomic or diatomic systems. However, it has been difficult to extend this 'two color' paradigm to more complex molecular systems. Since the reservoir leads to decoherence of the system overall, shorter and shorter pulses must be used to overcome decoherence effects. However, short pulses possess a large bandwidth with the result that, for complex systems, selectivity of the final state can be lost. One must therefore consider using interference between multiple pathways in order to control the system, as dictated by the details of the unperturbed Hamiltonian of the system.

For a polyatomic molecule or a solid, the complete Hamiltonian is virtually impossible to determine exactly and it is therefore equally difficult to prescribe the optimal spectral (amplitude and phase) content of the pulses that should be used for control purposes. An alternative approach has therefore emerged, in which one foregoes knowledge of the eigenstates of the Hamiltonian and details of the different possible interfering transition paths in order to achieve control of the end state of a system. This branch of coherence control has come to be known as optimal control. In optimal control one employs an optical source for which one can (ideally) have complete control over the spectral and phase properties. One then uses a feedback process in which experiments are carried out, the effectiveness of achieving a certain result is determined, and the pulse characteristics are then altered to obtain a new result. A key component is the use of an algorithm to select the new pulse properties as part of the feedback system. In this approach the molecule teaches the external control system what it 'requires' for a certain result to be optimally achieved. While the 'best' pulse properties may not directly reveal details of the multiple interference process required to achieve the optimal result, this technique can nonetheless be used to gain some insight into the properties of the unperturbed Hamiltonian and, regardless of such understanding, achieve a desirable result. It has been used to control chemical reaction rates involving several polyatomic molecules, with considerable enhancement in achieving a certain product relative to what can be done using simple thermodynamics.

Coherence Control in Semiconductors

Since coherence control of chemical reactions involving large molecules has represented a significant challenge, it was generally felt that ultrafast decoherence processes would also make control in solids, in general, and semiconductors, in particular, difficult. Early efforts therefore focused on atomic-like situations in which the electrons are bound and associated with discrete states and long coherence times. In semiconductors, the obvious choice is the excitonic system, defect states, or discrete states offered by quantum wells. Population control of excitons and directional ionization of electrons from quantum wells has been clearly demonstrated, similar to related population control of and directional ionization from atoms. Other manifestations of coherence control of semiconductors include control of electron–phonon interactions and intersub-band transitions in quantum wells.

Among the different types of coherence control in semiconductors is the remarkable result that it is possible to coherently control the properties of free electrons associated with continuum states. Although optimal control might be used for some of these processes, we have achieved clear illustrations of control phenomena based on the use of harmonically related beams and interference of single- and two-photon absorption processes connecting the continuum valence and conduction band states as generically illustrated in Figure 1c. Details of the various processes observed can be found elsewhere. Of course, the momentum relaxation time of electrons or holes in continuum states is typically of the order of 100 fs at room temperature, but this time is sufficiently long that it can permit phase-controlled processes. Indeed, with respect to conventional carrier transport, this 'long time' lapse is responsible for the typically high electrical mobilities in crystalline semiconductors such as Si and GaAs. In essence, a crystalline semiconductor with translational symmetry has crystal momentum as a good quantum number, and selection rules for scattering prevent the momentum relaxation time from being prohibitively small. Since electrons or holes in any continuum state can participate in such control processes, one need not be concerned about pulsewidth or bandwidth, unless the pulses were to be so short that carriers of both positive and negative effective mass were generated within one band.

Coherent Control of Electrical Current Using Two Color Beams

We now illustrate the basic principles that describe how the interference process involving valence and conduction band states can be used to control properties of a bulk semiconductor. We do so in the case of using one or two beams to generate and control carrier population, electrical current, and spin current. In pointing out the underlying ideas behind coherence control of semiconductors, we will skip or suppress many of the mathematical details which are required for a full understanding but which might obscure the essential physics. In particular we consider how phase-related optical beams with frequencies ω and 2ω interact with a direct gap semiconductor such that $\hbar\omega < E_g < 2\hbar\omega$ where E_g is the electronic bandgap. For simplicity we consider exciting electrons from a single valence band via single-photon absorption at 2ω and two-photon absorption at ω. As is well known, within an independent particle approximation, the states of electrons and holes in semiconductors can be labeled by their vector crystal momentum $\mathbf{k}$; the energy of states near the conduction or valence bandedges varies quadratically with $|\mathbf{k}|$. For one-photon absorption the transition amplitude can be derived using a perturbation Hamiltonian of the form $H = e/mc\, \mathbf{A}^{2\omega} \cdot \mathbf{p}$ where $\mathbf{A}^{2\omega}$ is the vector potential associated with the light field and $\mathbf{p}$ is the momentum operator. The transition amplitude is therefore of the form

$$\mathrm{a}^{2\omega} \propto E^{2\omega} e^{i\phi_{2\omega}} p_{cv} \qquad [4]$$

where p_{cv} is the interband matrix element of $\mathbf{p}$ along the field ($E^{2\omega}$) direction; for illustration purposes the field is taken to be linearly polarized. The overall transition rate between two particular states of the same $\mathbf{k}$ can be expressed as $W_1 \propto \mathrm{a}^{2\omega}(\mathrm{a}^{2\omega})^* \propto I^{2\omega}|p_{cv}|^2$ where $I^{2\omega}$ is the intensity of the beam. This rate is independent of the phase of the light beam as well as the sign of $\mathbf{k}$. Hence the absorption of light via single-photon transitions generally populates states of equal and opposite momentum with equal probability or, equivalently, establishes a standing electron wave with zero crystal momentum. This is not surprising since photons possess very little momentum and, in the approximation of a uniform electric field, do not give any momentum to an excited electron. The particular states that are excited depends on the light polarization and crystal orientation. However, the main point is that while single-photon absorption can lead to anisotropic filling of electron states, the distribution in momentum space is not polar (dependent on sign of $\mathbf{k}$). Similar considerations apply to the excited holes, but to avoid repetition we will focus on the electrons only.

For two-photon absorption involving the ω photons and connecting states similar to those connected with single-photon absorption, one must employ the 2nd order perturbation theory using the Hamiltonian $H = e/mc\mathbf{A}^{\omega} \cdot \mathbf{p}$. For two-photon absorption there is a transition from the valence band to an (energy nonallowed) intermediate state followed by a transition from the intermediate state to the final state. To determine the total transition rate one must sum over all possible intermediate states. For semiconductors, by far the dominant intermediate state is the final state itself, so that the associated transition amplitude has the form

$$\mathrm{a}^{\omega} \propto (E^{\omega} e^{i\phi_{\omega}} p_{cv})(E^{2\omega} e^{i\phi_{\omega}} p_{cc}) \propto (E^{\omega})^2 e^{2i\phi_{\omega}} p_{cv} \hbar k \qquad [5]$$

where the matrix element p_{cc} is simply the momentum of the conduction band state ($\hbar k$) along the direction of the field. Note that unlike an atomic system, p_{cc} is nonzero, since Bloch states in a crystalline solid do not possess inversion symmetry. If two-photon absorption acts alone, the overall transition rate between two particular $\mathbf{k}$ states would be $W_2 \propto (I^{\omega})^2 |p_{cv}|^2 k^2$. As with single-photon absorption, because this transition rate is independent of the sign of k, two-photon absorption leads to production of electrons with no net momentum.

When both single- and two-photon transitions are present simultaneously, the transition amplitude is the sum of the transition amplitudes expressed in eqn [3]. The overall transition rate is then found using eqn [1] and yields $W = W_1 + W_2 + Int$ where the interference term Int is given by

$$Int \propto E^{2\omega} E^{\omega} E^{\omega} \sin(\phi_{2\omega} - 2\phi_{\omega}) k \qquad [6]$$

Note, however, that the interference effect depends on the sign of k and hence can be constructive for one part of the conduction band but destructive for other parts of that band, depending on the value of the relative phase, $\Delta\phi = \phi_{2\omega} - 2\phi_\omega$. In principle one can largely eliminate transitions with $+\mathbf{k}$ and enhance those with $-\mathbf{k}$. This effectively generates a net momentum for electrons or holes and hence, at least temporarily, leads to an electrical current in the absence of any external bias. The net momentum of the carriers, which is absent in the individual processes, must come from the lattice. Because the carriers are created with net momentum during the pulses one has a form of current injection that can be written as

$$dJ/dt \propto E^{2\omega}E^{\omega}E^{\omega}\sin(\Delta\phi) \qquad [7]$$

where J is the current density. This type of current injection is allowed in both centrosymmetric and noncentrosymmetric materials. The physics and concepts behind the quantum interference leading to this form of current injection are analogous with the Young's double slit experiment. Note as well that if this 'interferometer' is balanced (single- and two-photon transition rates are similar), then it is possible to control the overall transition rate with great contrast. Roughly speaking, one balances the two arms of the 'effective interferometer' involved in the interference process, one 'arm' corresponding to the one-photon process and the other to the two-photon process. As an example, let's consider the excitation of GaAs, which has a room-temperature bandgap of 1.42 eV (equivalent to 870 nm). For excitation of GaAs using 1550 and 775 nm light under balanced conditions, the electron cloud is injected with a speed close to 500 km s^{-1}. Under 'balanced' conditions, nearly all the electrons are moving in the same direction. Assuming that the irradiance of the 1550 nm beam is 100 MW cm^{-2}, while that of the second harmonic beam is only 15 kW cm^{-2} (satisfying the 'balance' condition), with Gaussian pulse widths of 100 fs, one obtains a surprisingly large peak current of about 1 kA cm^{-2} for a carrier density of only 10^{14} cm^{-3} if scattering effects are ignored. When scattering is taken into account, the value of the peak current is reduced and the transient current decays on a time-scale of the momentum relaxation time. Figure 2a shows an experimental setup which can be used to demonstrate coherence control of electrical current using the above parameters. Figure 2a shows the region between a pair of electrodes on GaAs being illuminated by a train of harmonically related pulses. Figure 2b illustrates how the steady-state experimental voltage across the capacitor changes as the phase parameter $\Delta\phi$ is varied. Transient electrical currents, generated though incident femtosecond optical pulses, have also been detected through the emission of the associated Terahertz radiation.

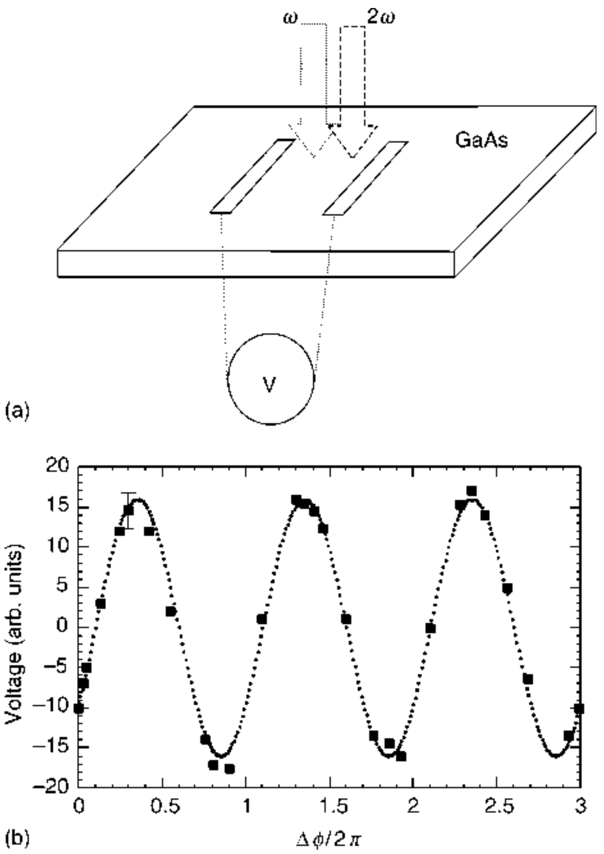

Figure 2 (a) Experimental setup to measure steady-state voltage (V) across a pair of electrodes on GaAs with the intervening region illuminated by phased radiation at ω and 2ω; (b) induced voltage across a pair of electrodes on a GaAs semiconductor as a function of the phase parameter associated with two harmonically related incident beams.

The current injection via coherence control and conventional cases (i.e., under a DC bias) differs with respect to their evolution. The current injected via coherent control has an onset determined by the rise time of the optical pulses. In the case of normal current production, existing carriers are accelerated by an electric field, and the momentum distribution is never far from isotropic. For a carrier density of 10^{14} cm^{-3} a DC field ~80 kV cm^{-1} is required to produce a current density of 1 kA cm^{-2}. In GaAs, with an electron mobility of 8000 cm^{2} V^{-1} s^{-1} this current would occur about 1/2 ps after the field is 'instantaneously' turned on. This illustrates that the coherently controlled phenomenon efficiently and quickly produces a larger current than can be achieved with the redirecting of statistically distributed electrons.

A more detailed analysis must take into account the actual light polarization, crystal symmetry, crystal face, and orientation relative to the optical polarization. For given optical intensities of the two beams, the maximum electrical current injection in the case of GaAs occurs for linearly polarized beams both oriented along the (111) or equivalent direction. However, large currents can also be observed with the light beams polarized along other high symmetry directions.

Coherent Control of Carrier Density, Spin Population, and Spin Current Using Two Color Beams

The processes described above do not exhaust the coherent control effects that can be observed in bulk semiconductors using harmonic beams. Indeed, for noncentrosymmetric materials, certain light polarizations and crystal orientations allow one to coherently control the total carrier generation rate with or without generating an electrical current. In the case of the cubic material GaAs, provided the ω beam has electric field components along two of the three principal crystal axes, with the 2ω beam having a component along the third direction, one can coherently control the total carrier density. However, the overall degree of control here is determined by how 'noncentrosymmetric' the material is.

The spin degrees of freedom of a semiconductor can also be controlled using two color beams. Due to the spin–orbit interaction, the upper valence bands of a typical semiconductor have certain spin characteristics. To date, optical manipulation of electron spin has been largely based on the fact that partially spin-polarized carriers can be injected in a semiconductor via one-photon absorption of circularly polarized light from these upper valence bands. In such carrier injection – where in fact two-photon absorption could be used as well – spins with no net velocity are injected, and then are typically dragged by a bias voltage to produce a spin-polarized current. However, given the protocols discussed above it should not come as a surprise that the two color coherence scheme, when used with certain light polarizations, can coherently control the spin polarization, making it dependent on $\Delta\phi$. Furthermore, for certain polarization combinations it is also possible to generate a spin current with or without an electrical current. Given the excitement surrounding the field of spintronics, where the goal is the use of the spin degree of freedom for data storage and processing, the control of quantities involving the intrinsic angular momentum of the electron is of particular interest.

Various polarization and crystal geometries can be examined for generating spin currents with or without electrical current. For example, with reference to Figure 3a, for both beams propagating in the **z** (*normal*) direction of a crystal and possessing the same circular polarization, an electrical current can be injected in the (xy) plane, at an angle from the crystallographic **x** direction dependent on the relative phase parameter, $\Delta\phi$; this current is spin-polarized in the **z** direction. As well, the injected carriers have a ±**z** component of their velocity, as many with one component as with the other; but those going in one direction are preferentially spin-polarized in one direction in the (xy) plane, while those in the other direction are preferentially spin-polarized in the opposite direction. This is an example of a spin current in the absence of an electrical current.

Such a pure spin current that is perhaps more striking is observable with the two beams cross linearly polarized, for example with the fundamental beam in the **x** direction and the second harmonic beam in the **y** direction as shown in Figure 3b. Then there is no net spin injection; the average spin in any

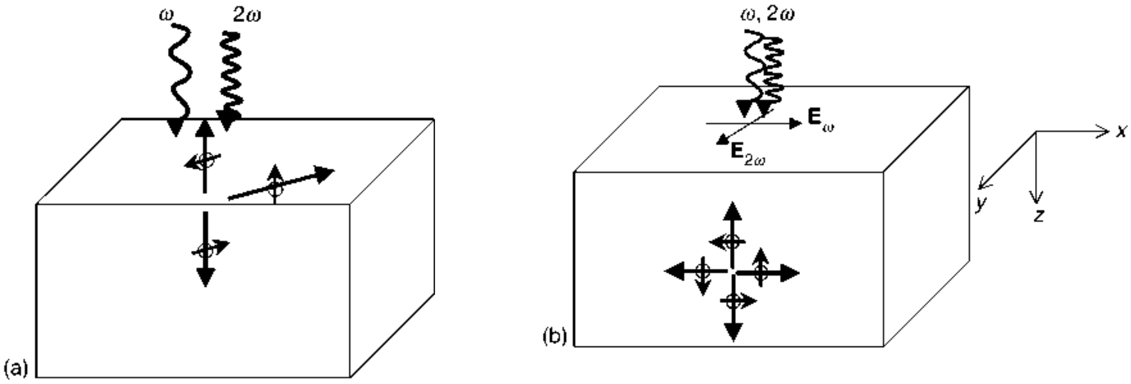

Figure 3 (a) Excitation of a semiconductor by co-circularly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the $x-y$ plane in a direction dependent on $\Delta\phi$ while a pure spin current is generated in the beam propagation (z) direction. (b) Excitation of a semiconductor by orthogonally, linearly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the direction of the fundamental beam polarization as well as along the beam propagation (z) direction.

direction is zero. And for a vanishing phase parameter $\Delta\phi$ there is no net electrical current injection. Yet, for example, the electrons injected with a $+\mathbf{x}$ velocity component will have one spin polarization with respect to the $\mathbf{z}$ direction, while those injected with a $-\mathbf{x}$ component to their velocity will have the opposite spin polarization with respect to the $\mathbf{z}$ direction.

The examples given above are the spin current analog of the two-color electrical current injection discussed above. There should also be the possibility of injecting a spin current with a single beam into crystals lacking center-of-inversion symmetry.

The simple analysis presented above relies on simple quantum mechanical ideas and calculations using essentially nothing more than Fermi's Golden Rule. Yet the process involves a mixing of two frequencies, and can therefore be thought of as a nonlinear optical effect. Indeed, one of the key ideas to emerge from the theoretical study of such phenomena is that an interpretation of coherence control effects alternate to that provided by the simple quantum interference picture that is provided by the usual susceptibilities of nonlinear optics. These susceptibilities are, of course, based on quantum mechanics, but the macroscopic viewpoint allows for the identification and classification of the effects in terms of 2nd order nonlinear optical effects, 3rd order optical effects, etc. Indeed, one can generalize many of the processes we have discussed above to a hierarchy of frequency mixing effects or high-order nonlinear processes involving multiple beams with frequency $n\omega, m\omega, p\omega\ldots$ with n, m, p being integers. Many of theses schemes require high intensity of one or more of the beams, but can occur in a simple semiconductor.

Coherent Control of Electrical Current Using Single Color Beams

It is also possible to generate coherence control effects using beams at a single frequency (ω), if one focuses on the two orthogonal components (e.g., x and y) polarization states and uses a noncentrosymmetric semiconductor of reduced symmetry such as a strained cubic semiconductor or a wurtzite material such as CdS or CdSe. In this case, interference between absorption pathways associated with the orthogonal components can lead to electrical current injection given by

$$dJ/dt \propto E^{\omega}E^{\omega}\sin(\phi_{\omega}^{x} - \phi_{\omega}^{y}) \qquad [8]$$

Since in this process current injection is linear in the beam's intensity, the high intensities necessary for two-photon absorption are not necessary. Nonetheless, the efficacy of this process is limited by the fact that it relies on the breaking of center-of-inversion symmetry of the underlying crystal. It is also clear that the maximum current occurs for circularly polarized light and that right and left circularly polarized light lead to a difference in sign of the current injection. Finally, for this particular single beam scheme, when circularly polarized light is used the electrical current is partially spin polarized.

Conclusions

Through the phase of optical pulses it is possible to control electrical and spin currents, as well as carrier density, in bulk semiconductors on a time-scale that is limited only by the rise time of the optical pulse and the intrinsic response of the semiconductor. A few optically based processes that allow one to achieve these types of control have been illustrated here. Although at one level one can understand these processes in terms of quantum mechanical interference effects, at a macroscopic level one can understand these control phenomena as manifestations of nonlinear optical phenomena. For fundamental as well as applied reasons, our discussion has focused on coherence control effects using the continuum states in bulk semiconductors, although related effects can also occur in quantum dot, quantum well, and superlattice semiconductors. Applications of these control effects will undoubtedly exploit the all-optical nature of the process, including the speed at which the effects can be turned on or off. The turn-off effects, although not discussed extensively here, are related to transport phenomena as well as momentum scattering and related dephasing processes.

See also

Semiconductor Physics: Spin Transport and Relaxation in Semiconductors; Spintronics. **Interferometry**: Overview. **Coherent Control:** Experimental; Theory.

Further Reading

Brumer P (1995) Laser control of chemical reactions. *Scientific American* 272: 56.

Rabitz H, de Vivie-Riedle R, Motzkus M and Kompa K (2000) Whither the future of controlling quantum phenomena? *Science* 288: 824.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. New York: Springer.

van Driel HM and Sipe JE (2001) Coherent control of photocurrents in semiconductors. In: Tsen T (ed.) *Ultrafast Phenomena in Semiconductors*, p. 261. New York: Springer Verlag.

COHERENT LIGHTWAVE SYSTEMS

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Conventional optical communication systems are based on the intensity modulation (IM) of an optical carrier by an electrical data signal and direct detection (DD) of the received light. The simplest scheme employs on/off keying (OOK), whereby turning on or off, an optical source transmits a binary 1 or 0. The signal light is transmitted down an optical fiber and at the receiver is detected by a photodetector, as shown in Figure 1. The resulting photocurrent is then processed to determine if a 1 or 0 was received. In the ideal case, where a monochromatic light source is used and where no receiver noise (detector dark current and thermal noise) is present, the probability of error is determined by the quantum nature (shot noise) of the received light. In this case, the number of photons/bit required (sensitivity) at the receiver, to achieve a bit-error-rate (BER) of 10^{-9}, is 10. This is called the quantum-limit. In practical receivers without an optical preamplifier, operating in the 1.3 μm and 1.55 μm optical fiber communication windows, the actual sensitivity is typically 10–30 dB less than the theoretical sensitivity. This is due to receiver noise, in particular thermal noise. This means that the actual sensitivity is between 100 to 10 000 photons/bit. The capacity of IM/DD fiber links can be increased using wavelength division multiplexing (WDM). At the receiver the desired channel is selected using a narrowband optical filter. Some advanced commercial IM/DD systems utilize dense WDM with 50 GHz spacing and single-channel bit rates as high as 40 Gb/s. Indeed, IM/DD systems with terabit (10^{12} bit/s) capacity are now possible.

Compared to IM/DD systems, coherent optical communication systems have greater sensitivity and selectivity. In the context of coherent lightwave systems, the term coherent refers to any technique employing nonlinear mixing between two optical waves on a photodetector, as shown in Figure 2. Typically, one of these is an information-bearing signal and the other is a locally generated wave (local oscillator). The result of this heterodyne process is a modulation of the photodetector photocurrent at a frequency equal to the difference between the signal and local oscillator frequencies. This intermediate frequency (IF) electronic signal contains the information in the form of amplitude, frequency, or phase that was present in the original optical signal. The IF signal is filtered and demodulated to retrieve the transmitted information. An automatic frequency control circuit is required to keep the local-oscillator frequency stable.

It is possible for the information signal to contain a number of subcarriers (typically at microwave frequencies), each of which can be modulated by a separate data channel. It is a simple matter to select the desired channel at the receiver by employing electronic heterodyning and low pass filtering, as shown in Figure 3. In IM/DD systems, channel selection can only be carried out using narrowband optical filters.

Coherent optical communications utilize techniques that were first investigated in radio communications. Most of the basic research work on coherent optical communications was carried out in the 1980s and early 1990s, and was primarily motivated by the need for longer fiber links using no repeaters. Improving receiver sensitivity using coherent techniques made it possible to increase fiber link spans. In the mid-1990s, reliable optical fiber amplifiers became available. This made it possible to construct fiber links using optical amplifiers spaced at appropriate intervals to compensate for fiber and other transmission losses.

The sensitivity of an IM/DD receiver can be greatly improved by the use of an optical preamplifier. Indeed the improvement is so marked that coherent techniques are not a viable alternative to IM/DD in the vast majority of commercial optical communication systems due to their complexity and higher cost. Moreover, because semiconductor lasers with very precise wavelengths and narrowband optical filters can be fabricated, the selectivity advantage of coherent optical communications has become less important. This has meant that the research work currently being carried out on coherent optical communications is very limited, and in particular, very few field trials have been carried out since the early 1990s.

We will review the basic principles underlying coherent optical communications, modulation schemes, detection schemes, and coherent receiver sensitivity. The effects of phase noise and polarization on coherent receiver performance will be outlined. A brief summary of pertinent experimental results will also be presented.

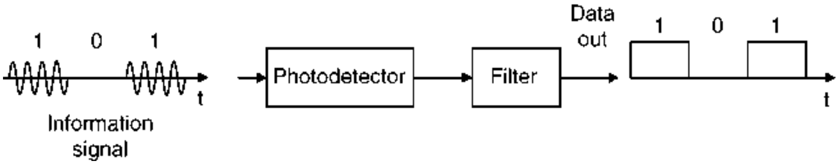
1 0 1
Information signal
t
Photodetector
Filter
Data out
1 0 1
t

Figure 1 IM/DD optical receiver.

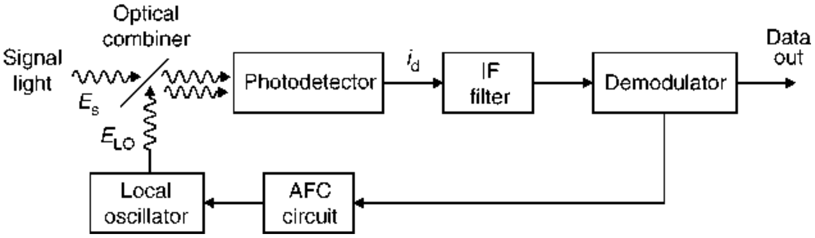
Optical combiner
Signal light
E_s
E_{LO}
Photodetector
i_d
IF filter
Demodulator
Data out
Local oscillator
AFC circuit

Figure 2 Basic coherent optical receiver. AFC: automatic frequency control.

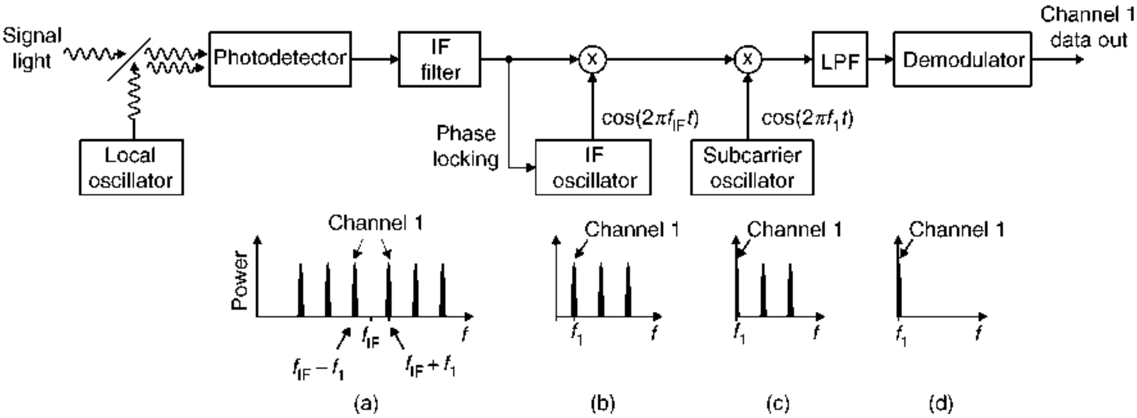
Signal light
Photodetector
IF filter
Local oscillator
Phase locking
IF oscillator
cos(2πf_IF t)
Subcarrier oscillator
cos(2πf_1 t)
LPF
Demodulator
Channel 1 data out
Power
Channel 1
f_{IF}
$f_{IF} - f_1$
$f_{IF} + f_1$
f
f_1
(a)
(b)
(c)
(d)

Figure 3 Coherent detection and processing of a subcarrier modulated signal lightwave. The power spectrum of the detected signal is shown: (a) after IF filtering; (b) after multiplication by the IF oscillator; (c) after multiplication with the subcarrier oscillator (in this case channel 1's carrier frequency); and (d) after low pass filtering prior to demodulation and data detection. LPF: lowpass filter.

Basic Principles

In coherent detection, a low-level optical signal field E_s is combined with a much larger power optical signal field E_{LO} from a local oscillator laser. E_s and E_{LO} can be written as

$$E_s(t) = \sqrt{2P_s}\cos[2\pi f_s t + \phi_s(t)] \qquad [1]$$

$$E_{LO}(t) = \sqrt{2P_{LO}}\cos[2\pi f_{LO} t + \phi_{LO}(t)] \qquad [2]$$

where P_s, f_s, and ϕ_s are the signal power, optical frequency, and phase (including phase noise), respectively. P_{LO}, f_{LO}, and ϕ_{LO} are the equivalent quantities for the local oscillator. For ideal coherent detection, the polarization states of the signal and local oscillator must be equal. If this is the case, the photocurrent is given by

$$i_d(t) = R[E_s(t) + E_{LO}(t)]^2 \qquad [3]$$

where R is the detector responsivity. Because the detector cannot respond to optical frequencies, we have

$$i_d(t) = R\{P_s + P_{LO} + 2\sqrt{P_s P_{LO}} \times \cos[2\pi f_{IF} t + \phi_s(t) - \phi_{LO}(t)]\} \qquad [4]$$

The IF signal is centered at frequency $f_{IF} = f_s - f_{LO}$. In homodyne detection $f_s = f_{LO}$ and the homodyne signal is centered at baseband. In heterodyne detection $f_s \neq f_{LO}$ and the signal generated by the photodetector is centered around f_{IF} (typically three to six times the bit rate). The IF photocurrent is proportional to $\sqrt{P_s}$, rather than P_s as is the case in direct detection, and is effectively amplified by a factor proportional to $\sqrt{P_{LO}}$. If P_{LO} is large enough, the signal power can be raised above the receiver noise, thereby leading to a greater sensitivity than possible with conventional IM/DD receivers using p–i–n or avalanche photodiodes. Theoretically it is possible to reach shot noise limited receiver sensitivity although this is not possible in practice. Homodyne optical receivers usually require less bandwidth and are more sensitive than heterodyne receivers but require that the local oscillator is phase locked to the information signal. This requires an optical phase locked loop (OPLL), which is very difficult to design.

Modulation Schemes

The three principal modulation schemes used in coherent optical communications; amplitude shift keying (ASK), frequency shift keying (FSK) and phase shift keying (PSK), are shown in Figure 4. ASK modulation is essentially the same as the OOK scheme used in IM/DD systems. The optical signal is given by

$$E_s(t) = 2a(t)\sqrt{2P_s}\cos[2\pi f_s t + \phi_s(t)] \qquad [5]$$

where $a(t) = 0$ or 1 for transmission of a 0 or 1, respectively. P_s is the average signal power, assuming that it is equally likely that a 0 or 1 is transmitted. ASK modulation can be achieved by modulating the output of a semiconductor laser using an external modulator. The most common type is based on LiNbO$_3$ waveguides in a Mach–Zehnder interferometer configuration.

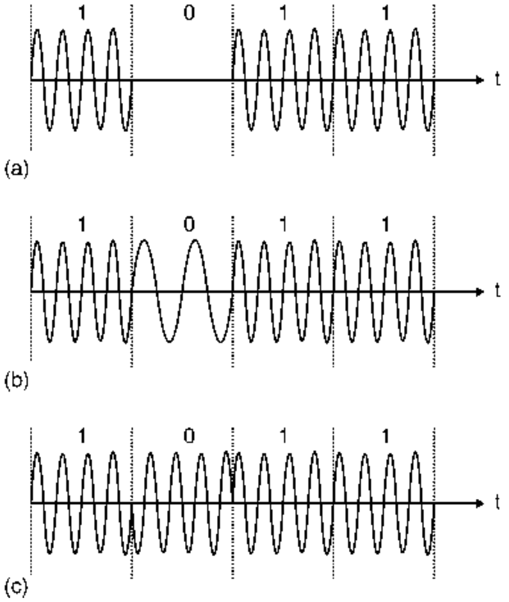

Figure 4 Modulation schemes for coherent optical communications: (a) ASK; (b) FSK; and (c) PSK.

In FSK modulation, the optical frequency of the signal lightwave is changed slightly by the digital data stream. In binary FSK, the optical signal is given by

$$E_s(t) = \sqrt{2P_s}\cos[2\pi(f_s + a(t)\Delta f)t + \phi_s(t)] \qquad [6]$$

where $a(t) = -1$ or 1 for transmission of a 0 or 1, respectively. The choice of the frequency deviation Δf depends on the available bandwidth, the bit rate B_T and the demodulation scheme used in the receiver. The modulation index m is given by

$$m = \frac{2\Delta f}{B_T} \qquad [7]$$

The minimum value of m that produces orthogonality (for independent detection of the 0s and 1s) is 0.5. If $m = 0.5$, the signal is referred to as minimum FSK (MFSK). When $m \geq 2$, the signal is referred to as wide-deviation FSK, and when $m < 2$, as narrow-deviation FSK. Wide-deviation FSK receivers are more resilient to phase noise than narrow-deviation FSK receivers. FSK modulation can be achieved using LiNbO$_3$ external modulators, acoustic optic modulators and distributed feedback semiconductor lasers. The phase of the lightwave does not change between bit transitions.

In PSK modulation, the digital data stream changes the phase of the signal lightwave. In binary PSK the optical signal is given by

$$E_s(t) = \sqrt{2P_s}\cos\left[2\pi f_s t + a(t)\theta + \frac{\pi}{2} + \phi_s(t)\right] \qquad [8]$$

where $a(t) = -1$ or 1 for transmission of a 0 or 1, respectively. The phase shift between a 0 and 1 is $\theta + \pi/2$. Most PSK schemes use $\theta = \pi/2$. PSK modulation can be achieved by using an external phase modulator or a multiple quantum well electro-absorption modulator.

Detection Schemes

There are two classes of demodulation schemes: synchronous and asynchronous. The former exploits the frequency and phase of the carrier signal to perform the detection. The latter only uses the envelope of the carrier to perform the detection. In the following we review the principal demodulation schemes.

PSK Homodyne Detection

In an ideal PSK homodyne receiver, shown in Figure 5, the signal light and local oscillator have identical

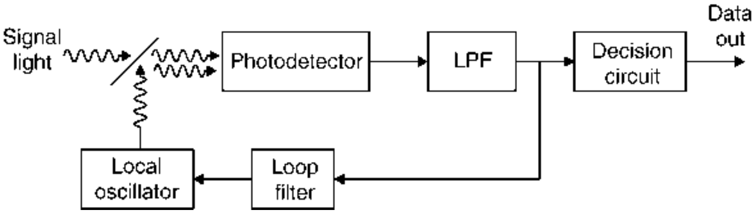
Signal
light
Photodetector
LPF
Decision
circuit
Data
out
Local
oscillator
Loop
filter

Figure 5 PSK homodyne receiver.

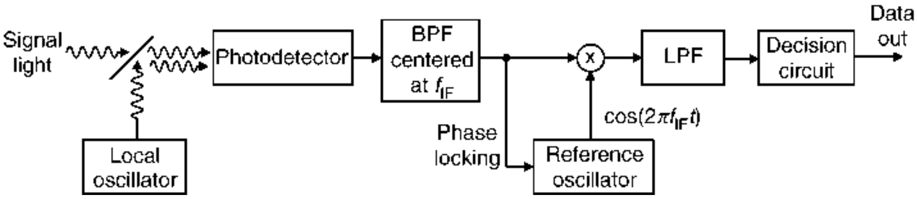
Signal
light
Photodetector
BPF
centered
at f_{IF}
X
LPF
Decision
circuit
Data
out
$\cos(2\pi f_{IF}t)$
Phase
locking
Reference
oscillator
Local
oscillator

Figure 6 PSK synchronous heterodyne receiver. BPF: bandpass filter.

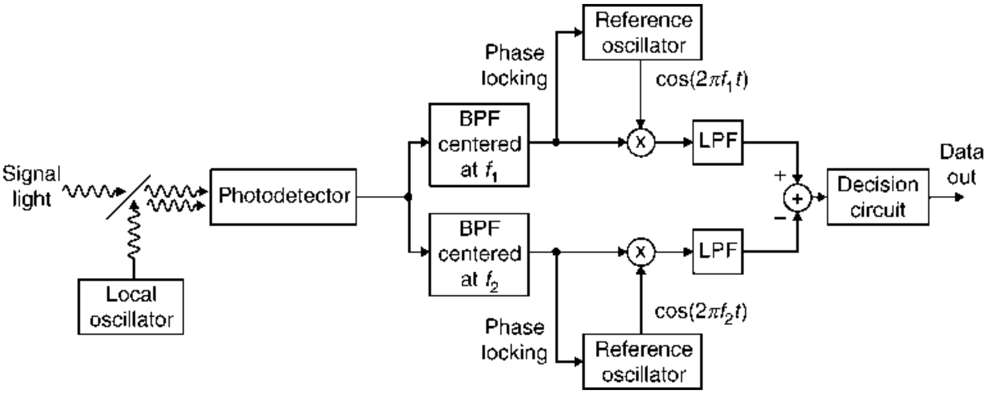
Reference
oscillator
Phase
locking
$\cos(2\pi f_1 t)$
BPF
centered
at f_1
X
LPF
Signal
light
Photodetector
+
+
−
Decision
circuit
Data
out
BPF
centered
at f_2
X
LPF
$\cos(2\pi f_2 t)$
Phase
locking
Reference
oscillator
Local
oscillator

Figure 7 FSK heterodyne synchronous receiver.

optical frequency and phase. If the resulting baseband information signal is positive, then a 1 has been received. If the baseband information signal is negative, then a 0 bit has been received. This scheme has the highest theoretical sensitivity but requires an OPLL and is also highly sensitive to phase noise.

ASK Homodyne Detection

An ideal ASK homodyne receiver is identical to an ideal PSK homodyne receiver. Because the baseband information signal is either zero (0 bit) or nonzero (1 bit), the receiver sensitivity is 3 dB less than the PSK homodyne receiver.

PSK Heterodyne Synchronous Detection

Figure 6 shows an ideal PSK synchronous heterodyne receiver. The IF signal is filtered by a bandpass filter, multiplied by the phase locked reference oscillator to move the signal to baseband, low pass filtered and sent to a decision circuit that decides if the received bit is a 0 or 1. The receiver requires synchronization of the reference oscillator with the IF signal. This is not easy to achieve in practice because of the large amount of semiconductor laser phase noise.

ASK Heterodyne Synchronous Detection

This scheme is basically the same as the PSK heterodyne detection and is similarly difficult to design.

FSK Heterodyne Synchronous Detection

This scheme is shown in Figure 7 for a binary FSK signal consisting of two possible frequencies $f_1 = f_s - \Delta f$ and $f_2 = f_s + \Delta f$. There are two separate branches in which correlation with the two possible IF signals at f_1 and f_2 is performed. The two resulting signals are subtracted from each other and a decision taken on the sign of the difference. Because of the requirement for two electrical phase locked loops, the scheme is not very practical.

ASK Heterodyne Envelope Detection

In this scheme, shown in Figure 8, the IF signal is envelope detected to produce a baseband signal proportional to the original data signal. The output of the envelope detector is relatively insensitive to laser phase noise.

FSK Heterodyne Dual-Filter Detection

In this scheme, shown in Figure 9, two envelope detectors are used to demodulate the two possible IF signals at f_1 and f_2 and a decision taken based on the outputs. The scheme is tolerant to phase noise and can achieve high sensitivity. However, it needs high-bandwidth receiver electronics because a large frequency deviation is required.

FSK Heterodyne Single-Filter Detection

This scheme omits one branch of the FSK heterodyne dual-filter receiver and performs a decision in the same way as for the ASK heterodyne envelope detection. Because half the power of the detected signal is not used, the sensitivity of the scheme is 3 dB worse than the dual-filter scheme.

CPFSK Heterodyne Differential Detection

This scheme, shown in Figure 10, uses the continuous phase (CP) characteristic of an FSK signal, i.e., there is no phase discontinuity at bit transition times. After bandpass filtering, the IF photocurrent is

$$i_{\mathrm{IF}}(t) = A\cos\{2\pi[(f_{\mathrm{IF}} + a(t)\Delta f)]t\} \qquad [9]$$

where $A = R\sqrt{P_s P_{\mathrm{LO}}}$. The output from the delay-line demodulator, after low pass filtering to remove the double IF frequency term, is

$$x(t) = \frac{A^2}{2}\cos\{2\pi[f_{\mathrm{IF}} + a(t)\Delta f]\tau\} \qquad [10]$$

The function of the delay-line demodulator is to act as a frequency discriminator. We require that $x(t)$ is a maximum when $a(t) = 1$ and a minimum when $a(t) = -1$. This is the case if the following relations are satisfied:

$$2\pi(f_{\mathrm{IF}} + \Delta f)\tau = 2\pi k \qquad [11]$$

$$2\pi(f_{\mathrm{IF}} - \Delta f)\tau = (2k-1)\pi, \qquad k \text{ an integer}$$

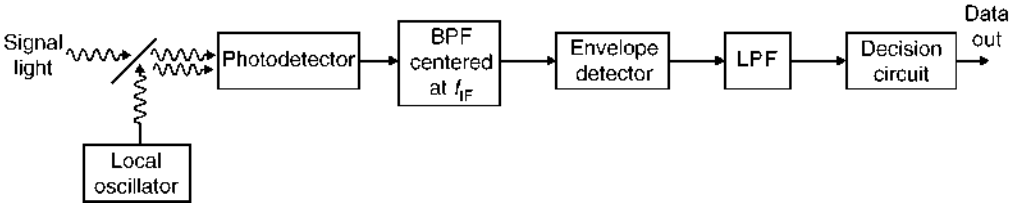
Signal light
Photodetector
BPF centered at f_{IF}
Envelope detector
LPF
Decision circuit
Data out
Local oscillator

Figure 8 ASK heterodyne envelope receiver.

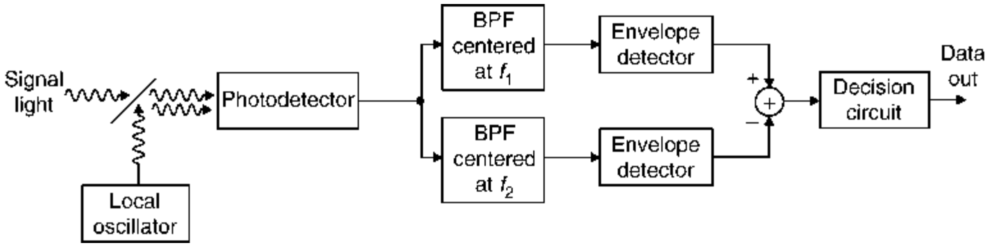
Signal light
Photodetector
BPF centered at f_1
BPF centered at f_2
Envelope detector
Envelope detector
+
−
+
Decision circuit
Data out
Local oscillator

Figure 9 FSK heterodyne dual-filter receiver.

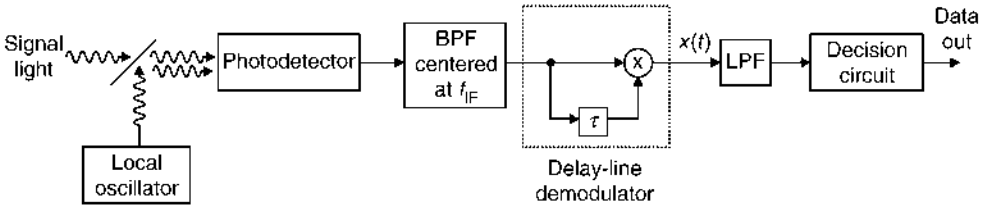
Signal light
Photodetector
BPF centered at f_{IF}
X
τ
Delay-line demodulator
x(t)
LPF
Decision circuit
Data out
Local oscillator

Figure 10 CPFSK heterodyne differential detection receiver.

Hence, we require that

$$\Delta f \tau = \frac{1}{4} \qquad [12]$$

$$f_{\mathrm{IF}}\tau = \left[k - \frac{1}{4}\right], \qquad k \text{ integer}$$

In terms of the modulation index we have

$$\tau = \frac{1}{2mB_{\mathrm{T}}} \qquad [13]$$

The above equation shows that it is possible to decrease the value of τ required by increasing m, thereby reducing the sensitivity of the receiver to phase noise. The minimum value of m possible is 0.5. This scheme can operate with a smaller modulation index compared to the dual-filter and single-filter schemes for a given bit rate, thereby relaxing the bandwidth requirements of the receiver electronics.

DPSK Heterodyne Differential Detection

In differential phase shift key (DPSK) modulation, a 0 is sent by changing the phase of the carrier by π with respect to the previous signal. A 1 is sent by not changing the phase. The receiver configuration is identical to the CPFSK heterodyne differential detection receiver, except that the time delay τ is equal to the inverse of the data rate. Compared to other PSK detection schemes, DPSK differential detection is relatively simple to implement and is relatively immune to phase noise.

Coherent Receiver Sensitivity

The signal-to-noise ratio γ_{c} of an ideal PSK homodyne receiver, in the shot-noise limit, is given by

$$\gamma_{\mathrm{c}} = \frac{2RP_{\mathrm{s}}}{eB_{\mathrm{c}}} \qquad [14]$$

where B_{c} is the receiver bandwidth. The detector responsivity is given by

$$R = \frac{\eta e}{hf_{\mathrm{s}}} \qquad [15]$$

where η is the detector quantum efficiency (≤ 1). The average signal power $P_{\mathrm{s}} = N_{\mathrm{s}}hf_{\mathrm{s}}B_{\mathrm{T}}$, where N_{s} is the average number of photons per bit. If it is assumed that $B_{\mathrm{c}} = B_{\mathrm{T}}$, then

$$\gamma_{\mathrm{c}} = 2\eta N_{\mathrm{s}} \qquad [16]$$

The actual detected number of photons per bit $N_{\mathrm{R}} = \eta N_{\mathrm{s}}$, so

$$\gamma_{\mathrm{c}} = 2N_{\mathrm{R}} \qquad [17]$$

If the noise is assumed to have a Gaussian distribution, then the bit error rate (BER) is given by

$$\mathrm{BER} = \frac{1}{2}\mathit{erfc}\left(\sqrt{2N_{\mathrm{R}}}\right) \qquad [18]$$

where *erfc* is the complementary error function; $\mathit{erfc}(x) \approx \mathrm{e}^{-x^2}/(x\sqrt{\pi})$ for $x > 5$. The receiver sensitivity for a $\mathrm{BER} = 10^{-9}$ is 9 photons/bit.

BER expressions and sensitivity for the demodulation schemes discussed above are listed in Table 1. The sensitivity versus N_{R} is plotted for some of the demodulation schemes in Figure 11. The received signal power required for a given BER is proportional to B_{T}. This dependency is shown in Figure 12 for some of the demodulation schemes.

Table 1 Shot-noise limited BER expressions and sensitivity for coherent demodulation schemes

Modulation scheme	*BER*	*Sensitivity (photons/bit)*
Homodyne		
PSK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{2N_{\mathrm{R}}}\right)$	9
ASK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{N_{\mathrm{R}}}\right)$	18 (peak 36)
Synchronous heterodyne		
PSK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{N_{\mathrm{R}}}\right)$	18
ASK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{\frac{N_{\mathrm{R}}}{2}}\right)$	36 (peak 72)
FSK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{\frac{N_{\mathrm{R}}}{2}}\right)$	36
Asynchronous heterodyne		
ASK (envelope)	$\frac{1}{2}\exp\left(-\frac{N_{\mathrm{R}}}{2}\right)$	40 (peak 80)
FSK (dual-filter)	$\frac{1}{2}\exp\left(-\frac{N_{\mathrm{R}}}{2}\right)$	40
FSK (single-filter)	$\frac{1}{2}\exp\left(-\frac{N_{\mathrm{R}}}{4}\right)$	80
Differential detection		
CPFSK	$\frac{1}{2}\exp(-N_{\mathrm{R}})$	20
DPSK	$\frac{1}{2}\exp(-N_{\mathrm{R}})$	20

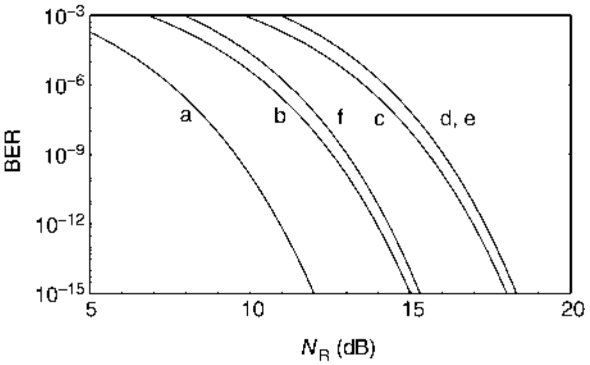

Figure 11 Shot-noise limited BER versus number of received photons/bit for various demodulation schemes: (a) PSK homodyne; (b) synchronous PSK; (c) synchronous FSK; (d) ASK envelope; (e) FSK dual-filter; and (f) CPFSK and DPSK.

Phase Noise

A major influence on the sensitivity possible with practical coherent receivers is laser phase noise, which is responsible for laser linewidth (3 dB bandwidth of the laser power spectrum). The effect of phase noise on system performance depends on the modulation and demodulation scheme used. In general, the effects of phase noise are more severe in homodyne and synchronous heterodyne schemes than in asynchronous schemes. If $\Delta\nu_{\mathrm{IF}}$ is the beat linewidth (without modulation present) between the signal and local oscillator lasers, the normalized beat linewidth $\Delta\nu$ is defined as

$$\Delta\nu = \frac{\Delta\nu_{\mathrm{IF}}}{B_{\mathrm{T}}} \qquad [19]$$

where $\Delta\nu_{\mathrm{IF}}$ is the IF phase noise bandwidth. Semiconductor lasers usually have a Lorentzian spectrum in which case $\Delta\nu_{\mathrm{IF}}$ is equal to the sum of the signal and local-oscillator laser linewidths. The requirements placed on $\Delta\nu$ become less severe as the bit rate increases. Table 2 compares the $\Delta\nu$ requirement for a 1 dB power penalty at a BER of 10^{-9} for some of the principal modulation/demodulation schemes.

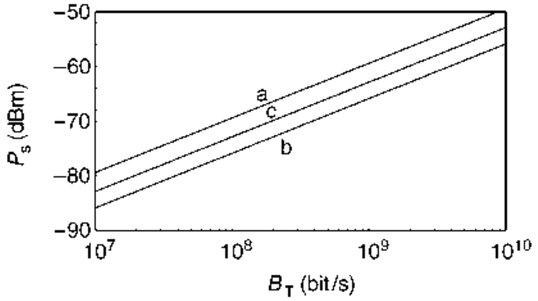

Figure 12 Required signal power versus bit rate for a BER = 10^{-9} for various shot-noise limited demodulation schemes: (a) PSK homodyne; (b) ASK envelope; and (c) CPFSK. The signal wavelength is 1550 nm and the detector is assumed to have unity quantum efficiency.

Table 2 Required normalized beat linewidth at 1 dB power penalty for a BER = 10^{-9} for various modulation/demodulation schemes

	Modulation scheme	*$\Delta\nu$ (%)*
Homodyne	PSK	6×10^{-4} (balanced loop) 0.01 (decision-driven loop)
Synchronous heterodyne	PSK	0.2–0.5
Asynchronous heterodyne	PSK	0.3–0.7
	ASK (envelope)	3
	FSK (dual-filter)	10–20
Differential detection	CPFSK	0.33 ($m = 0.5$) 0.66 ($m = 1.0$)
	DPSK	0.33

PSK Homodyne Detection – Phase Locked Loop Schemes

Practical homodyne receivers require an OPLL to lock the local-oscillator frequency and phase to that of the signal lightwave. There are a number of OPLL schemes possible for PSK homodyne detection, including the balanced PLL, Costas-type PLL and the decision-driven PLL.

The balanced PLL scheme, shown in Figure 13, requires that the PSK signal uses a phase shift of less than π between a 0 and 1, $E_s(t)$ and $E_{\mathrm{LO}}(t)$ can be written as

$$E_s(t) = \sqrt{2P_s}\cos\left[2\pi f_0 t + a(t)\theta + \phi_{\mathrm{N,s}}(t) + \frac{\pi}{2}\right] \qquad [20]$$

$$E_{\mathrm{LO}}(t) = \sqrt{2P_{\mathrm{LO}}}\cos[2\pi f_0 t + \phi_{\mathrm{N,LO}}(t)] \qquad [21]$$

where f_0 is the common signal and local oscillator optical frequency and $\phi_{\mathrm{N,s}}(t)$ and $\phi_{\mathrm{N,LO}}(t)$ are the signal and local oscillator phase noise, respectively. When $\theta \neq \pi/2$, $E_s(t)$ contains an unmodulated carrier component (residual carrier) in quadrature with the information-bearing signal. $E_s(t)$ and $E_{\mathrm{LO}}(t)$ are combined by an optical 180° hybrid. The output optical signals from the π-hybrid are given by

$$E_1(t) = \frac{1}{\sqrt{2}}[E_s(t) + E_{\mathrm{LO}}(t)] \qquad [22]$$

$$E_2(t) = \frac{1}{\sqrt{2}}[E_s(t) - E_{\mathrm{LO}}(t)] \qquad [23]$$

The voltage at the output of the summing circuit is given by

$$\begin{aligned} v_{\mathrm{L}}(t) = 2RR_{\mathrm{L}}\sqrt{P_s P_{\mathrm{LO}}}\{&a(t)\sin\theta\cos[\phi_e(t)] \\ &+ \cos\theta\sin[\phi_e(t)]\} \end{aligned} \qquad [24]$$

where the photodetectors have the same responsivity and load resistance R_{L}. The phase error is given by

$$\phi_e(t) = \phi_{\mathrm{N,s}}(t) - \phi_{\mathrm{N,LO}}(t) - \phi_c(t) \qquad [25]$$

$\phi_c(t)$ is the controlled phase determined by the control signal $v_c(t)$ at the output of the loop filter. v_{L} contains two terms; an information-bearing signal proportional to $\sin\theta$, which is processed by the data detection circuit, and a phase error signal proportional to $\cos\theta$ used by the PLL for locking. The power penalty, due to the residual carrier transmission, is $10\log_{10}(1/\sin^2\theta)$ dB.

This scheme is not practical because the $\Delta\nu$ needed for a power penalty of <1 dB is typically less than 10^{-5}. This would require that $\Delta\nu_{\mathrm{IF}} < 10$ kHz for

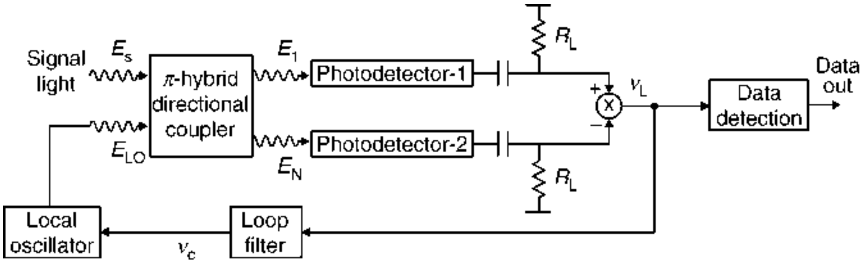
Signal light
E_s
E_1
π-hybrid directional coupler
Photodetector-1
Photodetector-2
E_{LO}
E_N
R_L
R_L
v_L
+
−
Data detection
Data out
Local oscillator
Loop filter
v_c

Figure 13 PSK homodyne balanced PLL receiver.

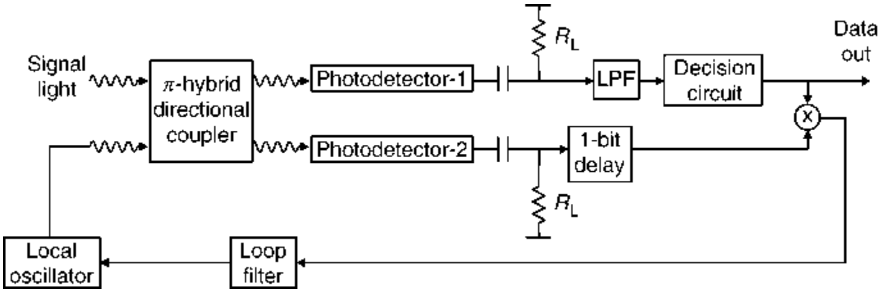
Signal light
π-hybrid directional coupler
Photodetector-1
Photodetector-2
R_L
R_L
LPF
Decision circuit
Data out
1-bit delay
Local oscillator
Loop filter

Figure 14 PSK decision-driven PLL.

a data rate of 1 Gb/s. The narrow signal and local-oscillator linewidths required are only achievable with external cavity semiconductor lasers.

The Costas-type and decision-driven PLLs also utilize π-hybrids and two photodetectors but as they have full suppression of the carrier ($\theta = \pi/2$), nonlinear processing of the detected signals is required to produce a phase error signal that can be used by the PLL for locking. In the decision-driven PLL, shown in Figure 14, the signal from one of the photodetector outputs is sent to a decision circuit and the output of the circuit is multiplied by the signal from the second photodetector. The mixer output is sent back to the local oscillator laser for phase locking. The $\Delta\nu$ required for a power penalty of <1 dB is typically less than 2×10^{-4}. This is superior to both the balanced PLL and Costas-type PLL receivers.

Heterodyne Phase Locking

In practical synchronous heterodyne receivers phase locking is performed in the electronic domain utilizing techniques from radio engineering. In the PSK case, electronic analogs of the OPLL schemes used in homodyne detection can be used. Synchronous heterodyne schemes have better immunity to phase noise than the homodyne schemes.

Asynchronous Systems

Asynchronous receivers do not require optical or electronic phase locking. This means that they are less sensitive to phase noise than synchronous receivers. They do have some sensitivity to phase noise because IF filtering leads to phase-to-amplitude noise conversion.

Differential Detection

Differential detection receivers have sensitivities to phase noise between synchronous and asynchronous receivers because, while they do not require phase locking, they use phase information in the received signal.

Phase-Diversity Receivers

Asynchronous heterodyne receivers are relatively insensitive to phase noise but require a much higher receiver bandwidth for a given bit rate. Homodyne receivers only require a receiver bandwidth equal to the detected signal bandwidth but require an OPLL, which is difficult to implement. The phase diversity receiver, shown in Figure 15, is an asynchronous homodyne scheme, which does not require the use of an OPLL and has a bandwidth approximately equal to synchronous homodyne detection. This is at the expense of increased receiver complexity and reduced sensitivity compared to synchronous homodyne detection. The phase-diversity scheme can be used with ASK, DPSK, and CPFSK modulation. ASK modulation uses an squarer circuit for the demodulator component, while DPSK and CPFSK modulation use a delay line and mixer.

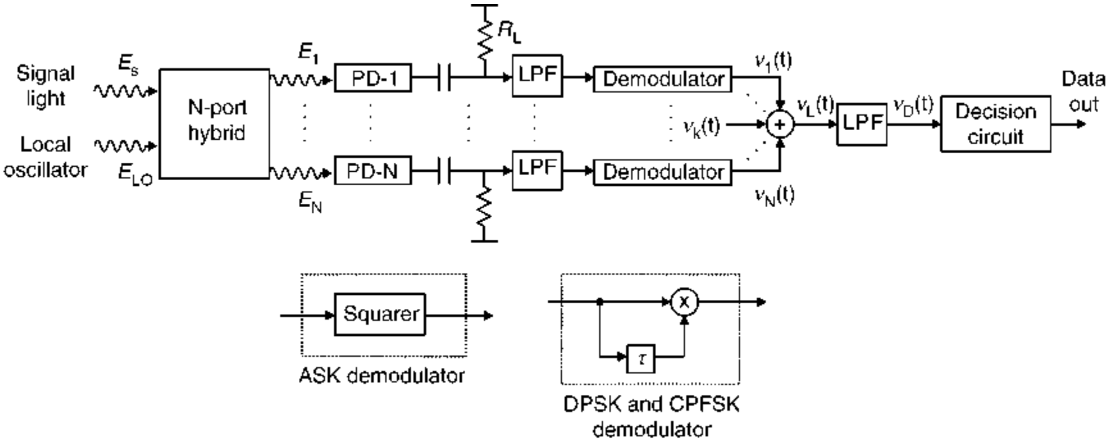

Figure 15 Multiport phase-diversity receiver. PD: photodetector.

In the case of ASK modulation, the outputs of the N-port hybrid can be written as

$$E_k(t) = \frac{1}{\sqrt{N}}\left\{a(t)\sqrt{2P_s}\cos\left[2\pi f_s t + \phi_s(t) + \frac{2\pi k}{N}\right] + \sqrt{2P_{LO}}\cos[2\pi f_{LO} t + \phi_{LO}(t)]\right\},$$
$$k = 1, 2, \ldots, N, \; N \geq 3 \qquad [26]$$

For $N = 2$, the phase term is $\pi/2$ for $k = 2$. The voltage input to the k-th demodulator (squarer), after the dc terms have been removed by the blocking capacitor, is given by

$$v_k(t) = a(t)\frac{RR_L}{N}2\sqrt{P_s P_{LO}}\cos\left[2\pi(f_s - f_{LO}) + \phi_s(t) - \phi_{LO}(t) + \frac{2\pi k}{N}\right] \qquad [27]$$

These voltages are then squared and added to give an output voltage

$$v_L(t) = \alpha\left(\frac{RR_L}{N}\right)^2 4P_s P_{LO} a(t)\sum_{k=1}^{N}\left\{1 + \cos\left[4\pi \times \left(f_s - f_{LO}\right)t + 2\phi_s(t) - 2\phi_{LO}(t) + \frac{4\pi k}{N}\right]\right\} \qquad [28]$$

where α is the squarer parameter. The lowpass filter at the summing circuit output effectively integrates $v_L(t)$ over a bit period. If $2(f_s - f_{LO}) \ll B_T$ and the difference between the signal and local oscillator phase noise is small within a bit period, then the input voltage to the decision circuit is

$$v_D(t) = \alpha\frac{(RR_L)^2}{N}4P_s P_{LO} a(t) \qquad [29]$$

which is proportional to the original data signal. In shot-noise limited operation, the receiver sensitivity is 3 dB worse than the ASK homodyne case. The scheme is very tolerant of laser phase noise. The $\Delta\nu$ for a power penalty of <1 dB is of the same order as for ASK heterodyne envelope detection.

Polarization

The mixing efficiency between the signal and local-oscillator lightwaves is a maximum when their polarization states are identical. In practice the polarization state of the signal arriving at the receiver is unknown and changes slowly with time. This means that the IF photocurrent can change with time and in the worst case, when the polarization states of the signal and local-oscillator are orthogonal, the IF photocurrent will be zero. There are a number of possible solutions to this problem.

1. Polarization maintaining (PM) fiber can be used in the optical link to keep the signal polarization from changing. It is then a simple matter to adjust the local-oscillator polarization to achieve optimum mixing. However, the losses associated with PM fiber are greater than for conventional single-mode (SM) fiber. In addition most installed fiber is SM fiber.
2. An active polarization controller, such as a fiber coil using bending-induced birefringence, can be used to ensure that the local-oscillator polarization tracks the signal polarization.
3. A polarization scrambler can be used to scramble the polarization of the transmitted signal at a speed greater than the bit rate. The sensitivity degradation due to the polarization scrambling is 3 dB compared to a receiver with perfect polarization matching. However, this technique is only feasible at low bit rates.
4. The most general technique is to use a polarization-diversity receiver, as shown in Figure 16. In this scheme, the signal light and local-oscillator

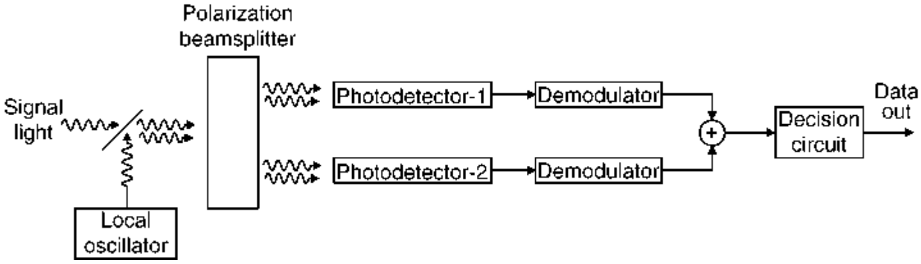

Figure 16 Polarization-diversity receiver.

Table 3 General comparisons between various modulation/demodulation schemes. The sensitivity penalty is relative to ideal homodyne PSK detection

Modulation/demodulation scheme	*Sensitivity penalty (dB)*	*Immunity to phase noise*	*IF bandwidth/bit rate*	*Complexity*
Homodyne PSK	0	Very poor	1	Requires an OPLL
Synchronous heterodyne PSK	3	Poor	3–6	Requires electronic phase locking
Asynchronous heterodyne FSK (dual-filter)	6.5	Excellent	3–6	High bandwidth IF circuits and two envelope detectors required
Differential detection CPFSK	3.5	Moderate	≈2	Relatively simple
DPSK	3.5	Moderate	≈2	Relatively simple

are split into orthogonally polarized lightwaves by a polarization beamsplitter. The orthogonal components of the signal and local-oscillator are separately demodulated and combined. The resulting decision signal is polarization independent. The demodulator used depends on the modulation format.

Comparisons Between the Principal Demodulation Schemes

General comparisons between the demodulation schemes discussed above are given in Table 3. FSK modulation is relatively simple to achieve. FSK demodulators have good receiver sensitivity and low immunity to phase noise. PSK modulation is also easy to achieve and gives good receiver sensitivity. ASK modulation has no real merit compared to FSK or PSK.

In general, homodyne schemes have the best theoretical sensitivity but require optical phase locking and are very sensitive to phase noise. Phase diversity receivers are a good alternative to homodyne receivers, in that they do not require an OPLL. Asynchronous heterodyne receivers have similar sensitivities to synchronous receivers but are much less complex. However, they require an IF bandwidth much greater than the data rate.

Differential detection schemes have good receiver sensitivity, do not require wide bandwidth IF circuits and have moderate immunity to phase noise. The advantage of CPFSK over DPSK is that the immunity to phase noise can be increased by increasing the FSK modulation index.

Table 4 Coherent systems experiments. The penalty is with respect to the theoretical sensitivity

Receiver	*Bit rate (Mb/s)*	*Sensitivity (dBm)*		*Penalty (dB)*
		dBm	*Photons/bit*	
ASK	400	−47.2	365	9.7
FSK (dual-filter)	680	−39.1	1600	16
CPFSK ($m = 0.5$)	1000	−37	1500	18.6
differential detection	4000	−31.3	1445	18.6
DPSK	400	−53.3	45.1	3.5
	2000	−39.0	480	13.8

Polarization diversity receivers can be used with any modulation/demodulation scheme and offer the most general approach to overcoming polarization effects.

System Experiments

Most coherent communication system experiments date from the 1980s and early 1990s and were based on the technology available at that time. This restricted the bit rates possible to typically <4 Gb/s. Some coherent systems experimental results obtained around this time are listed in Table 4. All of the listed experimental trials used lasers (usually external cavity lasers) with negligible phase noise in the 1550 nm region. It can be seen that the measured sensitivities deteriorate at higher bit rates. This is attributable to the nonuniform response and bandwidth limitations of the modulator, photodetector, and demodulator. There is no doubt that the performance could be improved by the use of current optoelectronic technologies.

Conclusion

Coherent optical communication systems offer the potential for increased sensitivity and selectivity compared to conventional IM/DD systems. However, coherent optical receivers are more complex and have more stringent design parameters compared to IM/DD receivers. Moreover, advances in optoelectronic component design and fabrication along with the advent of reliable optical amplifiers has meant that coherent systems are not a viable alternative to IM/DD systems in the vast majority of optical communication systems.

List of Units and Nomenclature

e	Electronic charge
h	Planck constant

See also

Modulators: Modulation and Demodulation of Optical Signals. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems; Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Lightwave Transmitters; Wavelength Division Multiplexing. **Polarization:** Introduction.

Further Reading

Barry JR and Lee EA (1990) Performance of coherent optical receivers. *IEEE Proceedings* 78: 1369–1394.

Cvijetic M (1996) *Coherent and Nonlinear Lightwave Communications*. Boston, MA: Artech House.

Kazovsky LG (1989) Phase- and polarization-diversity coherent optical techniques. *Journal of Lightwave Technology* 7: 279–292.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Boston, MA: Artech House.

Linke RA and Gnauck AH (1988) High-capacity coherent lightwave systems. *Journal of Lightwave Technology* 6: 1750–1769.

Ryu S (1995) *Coherent Lightwave Systems*. Boston, MA: Artech House.

COHERENT TRANSIENTS

Contents

Coherent Transient Spectroscopy in Atomic and Molecular Vapors
Foundations of Coherent Transients in Semiconductors
Ultrafast Studies of Semiconductors

Coherent Transient Spectroscopy in Atomic and Molecular Vapors

P R Berman, University of Michigan, Ann Arbor, MI, USA

R G Brewer, Stanford University, Stanford, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical spectroscopy is a method that allows one to determine atomic and molecular transition frequencies, as well as relaxation rates in atomic and molecular vapors. The transition frequencies serve as ‘fingerprints’ that can be used to identify and classify atoms and molecules. Optical spectroscopic methods fall into two broad categories, continuous wave (cw) or stationary spectroscopy and time-dependent or transient spectroscopy. In cw spectroscopy, the absorption or emission line shapes are measured as a function of the frequency of a probe field. On the basis of theoretical line shape formulas, it is then possible to extract the relevant transition frequencies and relaxation rates from the line shapes. In transient optical spectroscopy, pulsed optical fields are used to create atomic state populations or coherences between atomic states. Following the excitation, the time-evolution of the atoms is monitored, from which transition frequencies and relaxation rates can be obtained.

The earliest coherent transient effects were observed with atomic nuclear spins in about 1950 by H C Torrey, who discovered transient spin nutation, and by E L Hahn who detected spin echoes and free induction decay (FID). The optical analogs of these and other nuclear magnetic resonance (NMR) effects have now been detected using either optical pulses or techniques involving Stark or laser frequency switching.

In this brief introduction to coherent optical transients, several illustrative examples of coherent transient phenomena are reviewed. The optical Bloch equations are derived and coupled to Maxwell's Equations. The Maxwell–Bloch formalism is used to analyze free precession decay, photon echoes, stimulated photon echoes, and optical Ramsey fringes. Experimental results are presented along with the relevant theoretical calculations. In addition, coherent transient phenomena involving quantized matter waves are discussed. Although the examples considered in this chapter are fairly simple, it should be appreciated that sophisticated coherent transient techniques are routinely applied as a probe of complex structures such as liquids and solids or as a probe of single atoms and molecules.

Optical Bloch Equations

Many of the important features of coherent optical transients can be illustrated by considering the interaction of a radiation field with a two-level atom. The lower state $|1\rangle$ has energy $-\hbar\omega_0/2$ and upper state $|2\rangle$ has energy $\hbar\omega_0/2$. For the moment, the atom is assumed to be fixed at $\mathbf{R} = 0$ and all relaxation processes are neglected. The incident electric field is

$$\mathbf{E}(\mathbf{R}=0,t) = \frac{1}{2}\boldsymbol{\varepsilon}[E(t)e^{-i\omega t} + E(t)e^{i\omega t}] \qquad [1]$$

where $E(t)$ is the field amplitude, $\boldsymbol{\varepsilon}$ is the field polarization and ω is the carrier frequency. The time dependence of $E(t)$ allows one to consider pulses having arbitrary shape. A time-dependent phase for the field could have been included, allowing one to consider the effect of arbitrary frequency 'chirps,' but such effects are not included in the present discussion. It is convenient to expand the atomic state wave function in a field interaction representation as

$$|\psi(t)\rangle = c_1(t)e^{i\omega t/2}|1\rangle + c_2(t)e^{-i\omega t/2}|2\rangle \qquad [2]$$

The atom-field interaction potential is

$$V(\mathbf{R},t) = -\boldsymbol{\mu}\cdot\mathbf{E}(\mathbf{R},t) \qquad [3]$$

where $\boldsymbol{\mu}$ is a dipole moment operator. When eqn [2] is substituted into Schrödinger's equation and rapidly varying terms are neglected (rotating-wave approximation), one finds that the state amplitudes evolve according to

$$i\hbar\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t} = \tilde{\mathbf{H}}\mathbf{c}; \qquad \tilde{\mathbf{H}} = (\hbar/2)\begin{pmatrix} -\delta_0 & \Omega_0(t) \\ \Omega_0(t) & \delta_0 \end{pmatrix} \qquad [4]$$

where $\mathbf{c}$ is a vector having components (c_1, c_2)

$$\delta_0 = \omega_0 - \omega \qquad [5]$$

is an atom-field detuning

$$\Omega_0(t) = -\frac{\mu E(t)}{\hbar} = -\frac{\mu}{\hbar}\sqrt{\frac{2S(t)}{\varepsilon_0 c}} \qquad [6]$$

is a Rabi frequency, $\mu = \langle 1|\mu\cdot\varepsilon|2\rangle = \langle 2|\mu\cdot\varepsilon|1\rangle$ is a dipole moment matrix element, ε_0 is the permittivity of free space, and $\mathbf{S}(t)$ is the time-averaged Poynting vector of the field. Equation [4] can be solved numerically for arbitrary pulse envelopes.

Expectation values of physical observables are conveniently expressed in terms of density matrix elements defined by

$$\rho_{ij} = c_i c_j^* \qquad [7]$$

which obey equations of motion

$$\begin{aligned}
\dot\rho_{11} &= -i\Omega_0(t)(\rho_{21} - \rho_{12})/2 \\
\dot\rho_{22} &= i\Omega_0(t)(\rho_{21} - \rho_{12})/2 \\
\dot\rho_{12} &= -i\Omega_0(t)(\rho_{22} - \rho_{11})/2 + i\delta_0\rho_{12} \\
\dot\rho_{21} &= i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - i\delta_0\rho_{21}
\end{aligned} \qquad [8]$$

An alternative set of equations in terms of real variables can be obtained if one defines new parameters

$$\begin{aligned}
u &= \rho_{12} + \rho_{21}; & \rho_{12} &= (u + iv)/2 \\
v &= i(\rho_{21} - \rho_{12}); & \rho_{21} &= (u - iv)/2 \\
w &= \rho_{22} - \rho_{11}; & \rho_{22} &= (m + w)/2 \\
m &= \rho_{11} + \rho_{22}; & \rho_{11} &= (m - w)/2
\end{aligned} \qquad [9]$$

which obey

$$\begin{aligned}
\dot u &= -\delta_0 v \\
\dot v &= \delta_0 u - \Omega_0(t)w \\
\dot w &= \Omega_0(t)v \\
\dot m &= 0
\end{aligned} \qquad [10]$$

The last of these equations reflects the fact that $\rho_{11} + \rho_{22} = 1$, while the first three can be rewritten as

$$\dot{\mathbf{U}} = \mathbf{\Omega}(t) \times \mathbf{U} \qquad [11]$$

where the *Bloch vector* **U** has components (u, v, w) and the *pseudofield vector* $\mathbf{\Omega}(t)$ has components $[\Omega_0(t), 0, \delta_0]$.

Equations [8] or [10] constitute the optical Bloch equations without decay. The vector **U** has unit magnitude and traces out a path on the Bloch sphere as it precesses about the pseudofield vector. The component w is the population difference of the two atomic states, while u and v are related to the quadrature components of the atomic polarization (see below).

Equations [8] and [10] can be generalized to include relaxation. A simplified relaxation scheme that has a wide range of applicability is one in which state 2 decays spontaneously to state 1 with rate γ_2, while the relaxation of the coherence ρ_{12} is characterized by a complex decay parameter Γ_{12}, resulting from phase-changing collisions with a background gas. When such processes are included in eqn [8], they are modified as

$$\dot{\rho}_{11} = -i\Omega_0(t)(\rho_{21} - \rho_{12})/2 + \gamma_2\rho_{22} \qquad [12a]$$

$$\dot{\rho}_{22} = i\Omega_0(t)(\rho_{21} - \rho_{12})/2 - \gamma_2\rho_{22} \qquad [12b]$$

$$\dot{\rho}_{12} = -i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - \gamma\rho_{12} + i\delta\rho_{12} \qquad [12c]$$

$$\dot{\rho}_{21} = i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - \gamma\rho_{21} - i\delta\rho_{21} \qquad [12d]$$

where $\gamma = \gamma_2/2 + \mathrm{Re}(\Gamma_{12})$ and the detuning $\delta = \delta_0 - \mathrm{Im}(\Gamma_{12})$ is modified to include the collisional shift. With the addition of decay, the length of the Bloch vector is no longer conserved. The quantity γ_2 is referred to as the longitudinal relaxation rate. Moreover, one usually refers to $T_1 = \gamma_2^{-1}$ as the longitudinal relaxation time and $T_2 = \gamma^{-1}$ as the transverse relaxation time. In the case of purely radiative broadening, $\gamma_2 = 2\gamma$ and $T_1 = T_2/2$.

Maxwell–Bloch Equations

As a result of atom-field interactions, it is assumed that a polarization is created in the medium of the form

$$\mathbf{P}(\mathbf{R}, t) = \frac{1}{2}\boldsymbol{\varepsilon}\left[P(Z,t)e^{i[kZ-\omega t]} + P^*(Z,t)e^{-i[kZ-\omega t]}\right] \qquad [13]$$

which gives rise to a signal electric field of the form

$$\mathbf{E}_s(\mathbf{R}, t) = \frac{1}{2}\boldsymbol{\varepsilon}\left[E_s(Z,t)e^{i[kZ-\omega t]} + E_s^*(Z,t)e^{-i[kZ-\omega t]}\right] \qquad [14]$$

The z-axis has been chosen in the direction of **k**. It follows from Maxwell's equations that the quasi-steady-state field amplitude is related to the polarization by

$$\frac{\partial E_s(Z,t)}{\partial Z} = \frac{ik}{2\varepsilon_0}P(Z,t) \qquad [15]$$

To arrive at eqn [15], it is necessary to assume that transverse field variations can be neglected, that the *phase matching* condition $k = \omega/c$ is met, and that the complex field amplitudes $P(Z,t)$ and $E_s(Z,t)$ vary slowly in space compared with e^{ikZ} and slowly in time compared with $e^{i\omega t}$.

The polarization $\mathbf{P}(\mathbf{R}, t)$, defined as the average dipole moment per unit volume, is given by the expression

$$\mathbf{P}(\mathbf{R}, t) = \mathcal{N}\Big[\boldsymbol{\mu}_{21}\langle\rho_{12}(Z,t)\rangle e^{-i[kZ-\omega t]} + \boldsymbol{\mu}_{12}\langle\rho_{21}(Z,t)\rangle e^{i[kZ-\omega t]}\Big] \qquad [16]$$

where $\mathcal{N}$ is the atomic density and $\rho_{12}(Z,t)$ and $\rho_{21}(Z,t)$ are single-particle density matrix elements that depend on the atomic properties of the medium. As such, eqn [16] provides the link between Maxwell's equations and the optical Bloch equations. The $\langle\ \rangle$ brackets in eqn [16] indicate that there may be additional averages that must be carried out. For example, in a vapor, there is a distribution of atomic velocities that must be summed over, while, in a solid, there may be an inhomogeneous distribution of atomic frequencies owing to local strains in the media. By combining eqns [9], [13], [15] and [16], and using the fact that $\mu = \mu_{12}\cdot\varepsilon = \mu_{21}\cdot\varepsilon$, one arrives at

$$\begin{aligned}\frac{\partial E_s(Z,t)}{\partial Z} &= \frac{ik\mathcal{N}\mu}{\varepsilon_0}\langle\rho_{21}(Z,t)\rangle \\ &= \frac{ik\mathcal{N}\mu}{2\varepsilon_0}\langle u(Z,t) - iv(Z,t)\rangle \qquad [17]\end{aligned}$$

which, together with eqn [12], constitute the Maxwell–Bloch equations. From eqn [17], it is clearly seen that the coherence $\langle\rho_{21}(Z,t)\rangle$ drives the signal field. If $\langle\rho_{21}(Z,t)\rangle$ is independent of Z, the intensity of radiation exiting a sample of length L is proportional to

$$I(L,t) = \frac{k\mathcal{N}\mu L}{\varepsilon_0}{}^2|\langle\rho_{21}(t)\rangle|^2 \qquad [18]$$

Alternatively, the outgoing field can be heterodyned with a reference field to extract the field amplitude rather than the field intensity.

Coherent Transient Signals

The Maxwell–Bloch equations can be used to describe the evolution of coherent transient signals. In a typical experiment, one applies one or more 'short' radiation pulses to an atomic sample and monitors the radiation field emitted by the atoms following the pulses. A 'short' pulse, is one satisfying

$$|\delta|\tau, ku\tau, \gamma\tau, \gamma_2\tau \ll 1 \qquad [19]$$

where τ is the pulse duration. Conditions [19] allow one to neglect any effects of detuning (including Doppler shifts or other types of inhomogeneous broadening) or relaxation during the pulse's action. The evolution of the atomic density matrix during each pulse is determined solely by the pulse area defined by

$$\theta = \int_{-\infty}^{\infty} \Omega_0(t)\mathrm{d}t \qquad [20]$$

In terms of density matrix elements in a standard interaction representation defined by

$$\begin{aligned} \rho_{12}^I &= \rho_{12}e^{-i[kZ+\delta t]}, & \rho_{21}^I &= \rho_{21}e^{i[kZ+\delta t]}, \\ \rho_{22}^I &= \rho_{22}, & \rho_{11}^I &= \rho_{11} \end{aligned}$$

the change in density matrix elements for an interaction occurring at time T_j is given by

$$\begin{pmatrix} \rho_{11}^I \\ \rho_{22}^I \\ \rho_{12}^I \\ \rho_{21}^I \end{pmatrix}^+ = \frac{1}{2}\begin{pmatrix} 1+\cos\theta_j & 1-\cos\theta_j & i\sin\theta_j e^{i\Phi(T_j)} & -i\sin\theta_j e^{-i\Phi(T_j)} \\ 1-\cos\theta_j & 1+\cos\theta_j & -i\sin\theta_j e^{i\Phi(T_j)} & i\sin\theta_j e^{-i\Phi(T_j)} \\ i\sin\theta_j e^{-i\Phi(T_j)} & -i\sin\theta_j e^{-i\Phi(T_j)} & 1+\cos\theta_j & (1-\cos\theta_j)e^{-2i\Phi(T_j)} \\ -i\sin\theta_j e^{i\Phi(T_j)} & i\sin\theta_j e^{i\Phi(T_j)} & (1-\cos\theta_j)e^{2i\Phi(T_j)} & 1+\cos\theta_j \end{pmatrix}\begin{pmatrix} \rho_{11}^I \\ \rho_{22}^I \\ \rho_{12}^I \\ \rho_{21}^I \end{pmatrix}^- \qquad [21]$$

where $\mp$ superscripts refer to times just before and after the pulse

$$\Phi(T_j) = \mathbf{k}\cdot\mathbf{R}(T_j) + \delta T_j$$

$$\mathbf{R}(T_j) = \mathbf{R} - \mathbf{v}(t - T_j)$$

θ_j is the pulse area, and $\mathbf{v}$ is the atomic velocity. Between pulses, or following the last pulse, the atoms evolve freely in the absence of any applied fields. The density matrix elements evolve as

$$\begin{aligned} \dot{\rho}_{11}^I &= \gamma_2\rho_{22}^I; & \dot{\rho}_{22}^I &= -\gamma_2\rho_{22}^I; \\ \dot{\rho}_{12}^I &= -\gamma\rho_{12}^I; & \dot{\rho}_{21}^I &= -\gamma\rho_{21}^I \end{aligned} \qquad [22]$$

A coherent transient signal is constructed by piecing together field interaction zones and free evolution periods. There are many different types of coherent transient signals. A few illustrative examples are given below.

Free Polarization Decay

Free polarization decay (FPD), also referred to as optical FID, is the optical analog of free induction decay (FID) in NMR. This transient follows preparation of the atomic sample by any of three methods: (1) an optical pulse resonant with the atomic transition; (2) a Stark pulse that switches the transition frequency of a molecule (atom) into resonance with an incident *cw* laser beam; or (3) an electronic pulse that switches the frequency of a *cw* laser into coincidence with the molecular (atomic) transition frequency. In methods (2) and (3), the FPD emission propagates coherently in the forward direction along with the *cw* laser beam, automatically generating a heterodyne beat signal at a detector that can be orders of magnitude larger than the FPD signal. The FPD field intensity itself exiting the sample is equal to

$$\begin{aligned} I(L,t) &= \left(\frac{k\mathcal{N}\mu L}{\varepsilon_0}\right)^2 |\langle\rho_{21}(t)\rangle|^2 \\ &= \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta\right)^2 \left|\int \mathrm{d}\mathbf{v}\,W_0(\mathbf{v})e^{-[\gamma - i(\delta+\mathbf{k}\cdot\mathbf{v})]t}\right|^2 \end{aligned} \qquad [23]$$

For a Maxwellian velocity distribution

$$W_0(\mathbf{v}) = \frac{1}{(\pi u^2)^{3/2}}e^{-(v/u)^2} \qquad [24]$$

where u is the most probable atomic speed, one finds

$$I(L,t) = \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta\right)^2 e^{-2\gamma t}e^{-(kut)^2/2} \qquad [25]$$

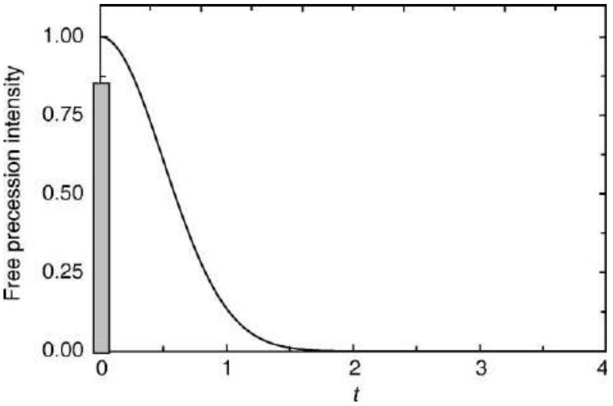

Figure 1 Pulse sequence and signal intensity for free precession decay. In this and subsequent figures, time is measured in units of 2/*ku*.

If $ku \gg \gamma$, the signal decays mainly owing to inhomogeneous broadening, as shown in Figure 1. Owing to different Doppler shifts for atoms having different velocities, the optical dipoles created by the pulse lose their relative phase in a time of order $T_2^* = 2/ku$. The FPD signal can be used to measure T_2^*, which can be viewed as an inhomogeneous, transverse relaxation time. At room temperature ku/γ is typically of order of 100.

When FPD signals are observed using laser frequency switching, method 3 above, the *cw* field can modify the molecular velocity distribution, exciting only those atoms having $\mathbf{k}\cdot\mathbf{v} = -\delta \pm \gamma'$, where γ' is a power broadened homogeneous width. In the linear field regime, the FPD signal again decays on a time scale of order $(ku)^{-1}$, since the velocity distribution is unchanged to first order in the field amplitude. When one considers nonlinear interactions with the field, however, the *cw* field excites a narrow velocity subclass that is no longer subject to inhomogeneous broadening. These atoms give rise to a contribution to the FPD signal that decays with rate $T_2^{-1} = (\gamma + \gamma')$. Thus by using nonlinear atom-field interactions, one can extract homogeneous decay rates in situations where there is large inhomogeneous broadening. The price one pays is that only a small percentage of the atoms (those having velocities for which the applied field is resonant) contribute to the signal.

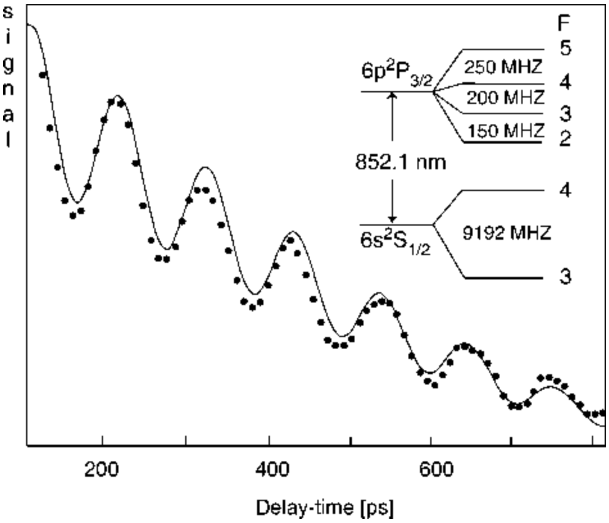

Figure 2 Optical free precession decay in Cs using an excitation pulse having 20 ns duration. Reproduced with permission from Lehmitz H and Harde H (1986) In: Prior Y, Ben-Reuven A and Rosenbluh M (eds) *Methods of Laser Spectroscopy*, pp. 109–112. New York: Plenum.

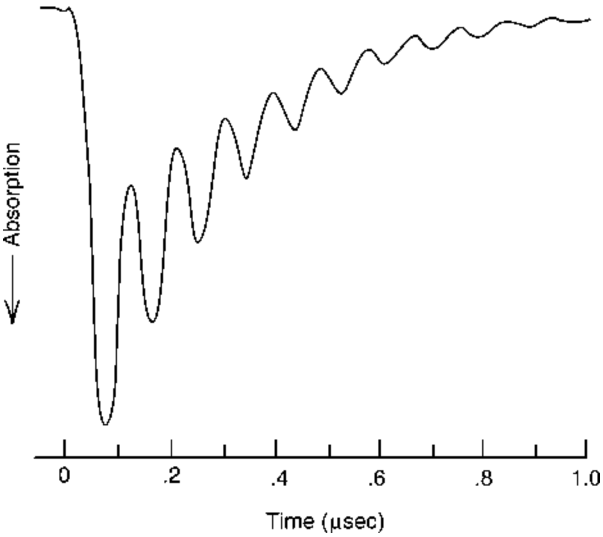

Figure 3 Optical free precession in NH_2D using the Stark switching technique. Reproduced with permission from Brewer RG and Shoemaker RL (1972) Optical free induction decay. *Physical Review A* 6: 2001. Copyright (1972) by the American Physical Society.

A FPD signal obtained on the D_2 transition in Cs is shown in Figure 2, where $T_2^* = 1.4$ ns. Oscillations in the signal originate from ground state hyperfine splitting. Figure 3 represents a FPD signal obtained in NH_2D at 10.6 μm using the method of Stark switching with *cw* state preparation. The oscillations are the heterodyne beat signal while the slowly varying increase in the signal is the result of optical nutation of molecules switched into resonance by the Stark pulse. The FPD signal manifests itself as a reduction of the amplitude of the oscillation with time. This amplitude decays with the (power-broadened) homogeneous decay rate γ'.

Photon Echo

Although the FPD signal produced by short excitation pulses decays in a time of order $(ku)^{-1}$, the coherence of individual atoms decays in a much longer time, T_2.

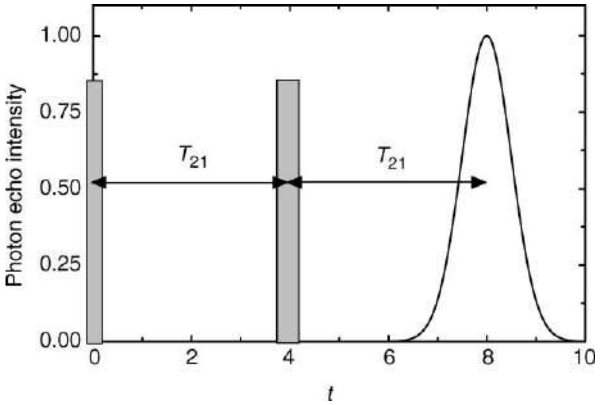

Figure 4 Pulse sequence and signal intensity for the photon echo.

The question arises as to whether it is possible to bring all the atomic dipoles back into phase so they can radiate a coherent signal. The photon echo accomplishes this goal, although it has very little to do with either photons or echoes. The pulse sequence and echo signal are shown in Figure 4. The first pulse creates a phased array of dipoles, which begin to dephase owing to inhomogeneous broadening. At time T_{21} the second pulse provides a nonlinear interaction that couples ρ_{12} and ρ_{21}. As a result of the second pulse, the dipoles begin a rephasing process that is completed at time $t = 2T_{21}$. The 'echo' appears at this time. The echo intensity exiting the sample is

$$\begin{aligned} I(L,t) &= \left(\frac{k\mathcal{N}\mu L}{\varepsilon_0}\right)^2 |\langle\rho_{21}(t)\rangle|^2 \\ &= \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta_1 \sin^2(\theta_2/2)\right)^2 \\ &\quad\times e^{-4\gamma T_{21}} e^{-k^2u^2(t-2T_{21})^2/2} \end{aligned} \qquad [26]$$

If the echo intensity is measured as a function of T_{21}, one can extract the homogeneous decay rate $\gamma = T_{21}^{-1}$. It is interesting to note that the echo intensity near $t = 2T_{21}$ mirrors the FPD intensity immediately following the first pulse. For experimental reasons, it is often convenient to use a different propagation vector for the second pulse. If $\mathbf{k}_1$ and $\mathbf{k}_2$ are propagation vectors of the first and second pulses, the echo signal is in the direction $\mathbf{k} = 2\mathbf{k}_2 - \mathbf{k}_1$, provided $|\mathbf{k}_2 - \mathbf{k}_1|/k_1 \ll 1$.

The echo signal is sensitive only to $\rho_{21}(t)$ and $\rho_{12}(t)$ in the entire time interval of interest. Any interaction that results in a degradation of the dephasing–rephasing process for these density matrix elements leads to a decrease in the echo intensity. As such, echo signals can serve as a probe of transverse relaxation. Transverse relaxation generally falls into two broad categories. First, there are dephasing processes which produce an exponential damping of the coherences and contribute to γ. Second there is *spectral diffusion*; in a vapor, spectral diffusion is produced by velocity-changing collisions that change the effective field frequency seen by the atoms owing to the Doppler effect. Such terms enter the optical Bloch equations as *integral* terms, transforming the equations into differentio-integral equations. In general, the phase-changing and velocity-changing aspects of collisions are entangled; however, if the collisional interaction is *state-independent*, as it is for some molecular transitions, then collisions are purely velocity-changing in nature, leading to an echo that decays exponentially as T_{21}^3 for early times and T_{21} for later times. For electronic transitions, collisions are mainly phase-changing in nature, but there is a velocity-changing contribution that persists in the forward diffractive scattering cone. Photon echo techniques are sufficiently sensitive to measure the relaxation caused by such diffractive scattering.

The photon echo was first observed by NA Kurnit and co-workers, using a pulsed ruby laser (Figure 5). In Stark switching experiments, the photon echo follows two Stark pulses as shown in Figure 6. The effects of velocity-changing collisions on the photon echo signal in a CH_3F molecular vapor is represented in Figure 7. The signal exhibits exponential decay varying as T_{21}^3 for early times and T_{21} for later times. A theoretical interpretation of the observed echo decay permits a determination of the binary collision parameters for CH_3F, namely, the characteristic velocity jump $\Delta u_{\mathrm{rms}} = 200\ \mathrm{cm\ s^{-1}}$ and the cross section of $\sigma = 580\ \text{Å}^2$.

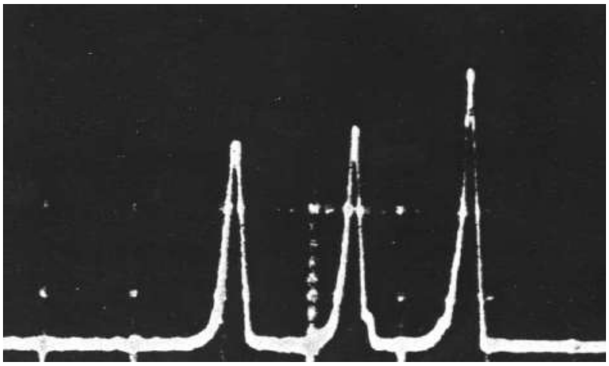

Figure 5 A photon echo signal from ruby. Time increases to the right with a scale of 100 ns/division. The pulse on the right is the echo signal, while the first two pulses are the (attenuated) input pulses. Reprinted with permission from Kurnit NA, Abella ID and Hartmann SR (1964) Observation of a photon echo. *Physical Review Letters* 13: 567. Copyright (1964) by the American Physical Society.

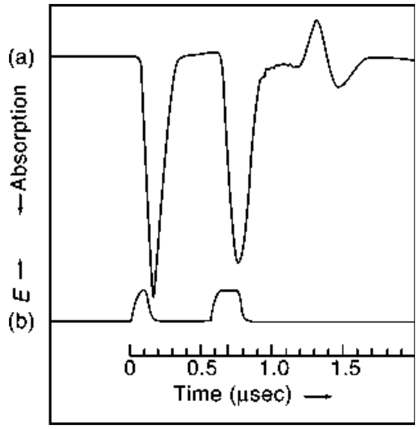

Figure 6 A photon echo signal in $^{13}CH_3F$ using the Stark switching method. The third pulse in trace (a) is the echo heterodyne beat signal. This signal is preceded by 'nutation' transient signals following each of the two Stark pulses shown in trace (b). Reprinted with permission from Brewer RG and Shoemaker RL (1971) Photo echo and optical nutation in molecules. *Physical Review Letters* 27: 631. Copyright (1971) by the American Physical Society.

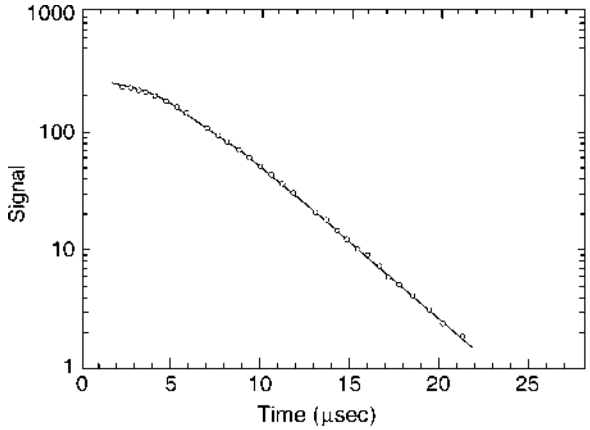

Figure 7 Decay of the photon echo signal in $^{13}CH_3F$ as a function of pulse separation. Reprinted with permission from Schmidt J, Berman PR and Brewer RG (1973) Coherent transient study of velocity-changing collisions. *Physical Review Letters* 31: 1103. Copyright (1973) by the American Physical Society.

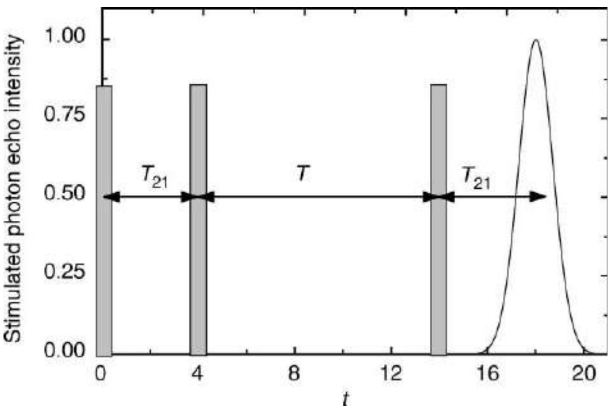

Figure 8 Pulse sequence and signal intensity for the stimulated photon echo.

Stimulated Photon Echo

In contrast to conventional photon echoes, *stimulated photon echoes* can be used to *simultaneously* measure both transverse and longitudinal relaxation times. Stimulated photon echoes have become an important diagnostic probe of relaxation in condensed matter systems as well as atomic vapors. The pulse sequence is shown in Figure 8 and consists of three pulses, having areas $\theta_1, \theta_2, \theta_3$, and equal propagation vectors $\mathbf{k}$. The time interval between the first two pulses is $T_{21} \gg T_2^*$ and pulse 3 occurs at time $t = T_{21} + T$. The maximum echo amplitude is given by

$$I(L,t_e) = \frac{1}{2}\left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta_1 \sin\theta_2 \sin\theta_3\right)^2 \times e^{-2\gamma_2 T} e^{-4\gamma T_{21}} \qquad [27]$$

and occurs when $t_e - (T_{21} + T) = T_{21}$. The optimal pulse sequence consists of three $\pi/2$ pulses.

If slightly different $\mathbf{k}$ vectors are chosen such that $\mathbf{k}_1 = \mathbf{k}_3 \approx \mathbf{k}_2$ the echo intensity varies as

$$e^{-2\gamma_2 T} e^{-4\gamma T_{21}} e^{-|\mathbf{k}_1 - \mathbf{k}_2|^2 u^2 (T+2T_{21})^2/2}$$

By varying the angle between $\mathbf{k}_1$and $\mathbf{k}_2$, one can determine the Doppler width ku. By monitoring the echo signal as a function of $T_{21}(T)$, one obtains information on the transverse (longitudinal) relaxation.

The situation changes dramatically if there is an *additional* ground state level (level 0) to which the excited state can decay. In that case, the system is 'open' since population can leave the 'two-level' subsystem. The total population of the 1–2 state subsystem is no longer conserved, requiring an *additional* decay rate to account for relaxation. Let us suppose that all states decay with rate Γ_t as a result of their finite time in the laser beams. Moreover, let $\Gamma_{2,1}$ and $\Gamma_{2,0}$ be the decay rates of level 2 to levels 1 and 0, respectively. Assuming that $\gamma_2 T \gg 1$, one finds that the factor $e^{-2\gamma_2 T}$ in eqn [27] must be replaced by $(\Gamma_{2,0}/(\Gamma_{2,1} + \Gamma_{2,0}))^2 e^{-2\Gamma_t T}$. If $\Gamma_{2,0} \neq 0$, there is a long-lived component in the ground state population that contributes to the echo signal. One can exploit this feature of open systems to study spectral diffusion or velocity-changing collisions with very high sensitivity. Echoes can occur for time separations T much greater than the excited state lifetime. The creation of such long-lived stimulated photon echoes can be attributed to the fact that, in open systems, a portion of the velocity modulation created by the first two pulses in the ground state remains following spontaneous decay from the excited state.

The stimulated photon echo observed on the D_1 transition in sodium is shown in Figure 9. This is

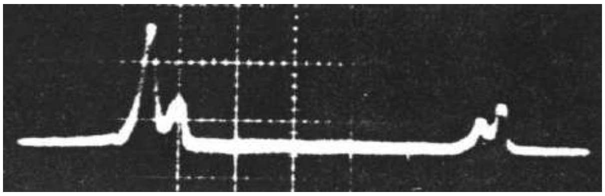

Figure 9 Stimulated photon echo observed on the D_1 transition in sodium. Reprinted with permission from Mossberg T, Flusberg A, Kachru R and Hartmann SR (1979) Total scattering cross section for Na on He measured by stimulated photon echoes. *Physical Review Letters* 42: 1665. Copyright (1979) by the American Physical Society.

an 'open' system, owing to ground state hyperfine structure. For the data shown, T is 17 times the 16 ns lifetime of the excited state. The first three pulses represent light scattered from the three input pulses and the fourth pulse is the echo. The echo appears at $t = T + 2T_{21}$.

Optical Ramsey Fringes

We have seen that coherent transients can be used to measure relaxation rates. With a slight modification of the stimulated photon echo geometry, it is also possible to use coherent transient signals to measure transition frequencies. If one chooses $\mathbf{k}_3$ in the $-\mathbf{k}_1$ direction instead of the $\mathbf{k}_1$ direction, it is possible to generate a phase matched signal in the

$$\mathbf{k} = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3 = -\mathbf{k}_2$$

direction. To simplify matters, it is assumed that $\mathbf{k}_1 \approx \mathbf{k}_2 = -\mathbf{k}$ and terms of order $|\mathbf{k}_1 - \mathbf{k}_2|u \times (T + 2T_{21})$ are neglected.

The averaged density matrix element in the vicinity of the echo is

$$\langle \rho_{21}(t) \rangle = (i/8) \sin \theta_1 \sin \theta_2 \sin \theta_3 e^{-\gamma_2 T} e^{-2\gamma T_{21}} \times e^{i\delta[(t-T_{21}-T)+T_{21}]} e^{-k^2 u^2 |(t-T_{21}-T)-T_{21}|^2/4} \qquad [28]$$

An echo is formed at time $t_e = T + 2T_{21}$, just as for the stimulated photon echo, but there is an additional phase factor, given by $e^{2i\delta T_{21}}$ when $t = t_e$, that is absent for the nearly collinear geometry. This phase factor is the optical analog of the phase factor that is responsible for the generation of Ramsey fringes in the microwave domain. The phase factor can be measured directly by heterodyning the signal field with a reference field or, indirectly, by converting the off-diagonal density matrix element into a population by the addition of a *fourth* pulse in the $\mathbf{k}_3$ direction at time $t = T + 2T_{21}$. In either case, the signal varies as $\cos(2\delta T_{21})$. The central fringe, corresponding to $\delta = 0$ can be isolated using one of two methods. If the experiment is carried out using an atomic beam rather than atoms in a cell, $T_{21} = L/u_0$, will be different for atoms having different u_0 (L = spatial separation of the first two pulses and u_0 is the longitudinal velocity of the atoms). On averaging over a distribution of u_0, the fringe having $\delta = 0$ will have the maximum amplitude. Experiments of this type allow one to measure optical frequencies with accuracy of order T_{21}^{-1}. For experiments using temporally-separated pulses acting on atoms in a cell, it is necessary to take data as a function of δ for several values of T_{21}, and then average the data over T_{21}. It might seem remarkable that narrow pulses having large bandwidths can be used to measure optical frequencies with arbitrary accuracy as T_{21} is increased. Of course, it is the *total* duration of the pulse sequence rather than the duration of a single pulse that is the relevant time parameter. For the optical coherence, T_{21} is the appropriate time parameter (during the interval T, it is population rather than coherence that contributes to the signal). The price one pays by using large T_{21} is a signal that decays as $e^{-4\gamma T_{21}}$.

Figure 10 illustrates an optical Ramsey fringe signal on the 657 nm intercombination line in Ca. Four-field zones were used. The most probable value of T_{21} was about 10^{-5} s for an effusive beam having a most probable longitudinal speed equal to 800 ms^{-1}, giving a central fringe width of order 60 kHz.

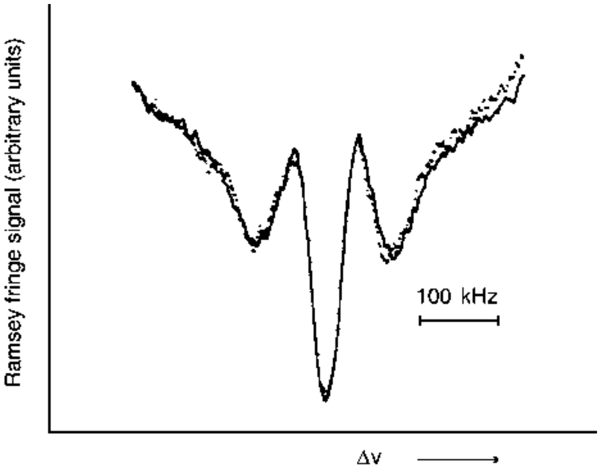

Figure 10 An optical Ramsey fringe signal on the 657 nm intercombination line in Ca. Four field zones were used. The solid and dotted lines represent runs with the directions of the laser field reversed, to investigate phase errors in the signals. Reprinted from Ito N, Ishikawa J and Moringa A (1994) Evaluation of the optical phase shift in a Ca Ramsey fringe stabilized optical frequency standard by means of laser-beam reversal. *Optics Communications* 109: 414, with permission from Elsevier.

Atom Interferometry

Matter wave atom interferometers rely on the wave nature of the center-of-mass motion for their operation. Such interferometers illustrate some interesting features of coherent optical transients not found in NMR. As an example, one can consider the response of a cold atomic ensemble to two, off-resonant standing-wave optical pulses separated in time by T. The electric field amplitude of pulse $j(j = 1, 2)$ is given by $\mathrm{E}_j(\mathbf{Z}, t) = \boldsymbol{\varepsilon} E_j(t)\cos(kZ)\cos(\omega t)$. The net effect of the field is to produce a spatially-modulated, ac Stark or light shift of the ground state energy. As a result, pulse j modifies the ground state amplitude by the phase factor $\exp[i\theta_j \cos(2kZ)]$. In other words, the standing wave field acts as a phase grating for the atoms. The phase factor can be interpreted in terms of momentum exchange between the two traveling wave components of the standing wave field. Owing to the fields' interaction with the atoms, all even integral multiples of $2\hbar k$ can be exchanged by the fields, imparting impulsive momenta of $2n\hbar k$ (n is a positive or negative integer) to the atoms. For an atom having momentum $\mathbf{P}$ and mass M, the frequency change associated with this momentum change is

$$\begin{aligned} E_{P,\mathbf{P}\pm 2n\hbar k\hat{Z}}/\hbar &= \left(\frac{P^2}{2M} - \frac{|\mathbf{P} \pm 2n\hbar k\hat{Z}|^2}{2M}\right)\Big/\hbar \\ &= \mp\frac{2nP_Z k}{M} - \omega_{2nk} \qquad [29] \end{aligned}$$

where

$$\omega_\kappa = \frac{\hbar\kappa^2}{2M}$$

There are two contributions to the frequency eqn [29]. The first part is independent of $\hbar$ and represents a classical Doppler shift, while the second part is proportional to $\hbar$ and represents a quantum, matter-wave effect. The quantum contribution will become important for times of order ω_{2nk}^{-1}. As with the echoes described previously, the Doppler effect results in a dephasing and rephasing of the optical response. However, in contrast to the previous echoes, the ground state density is no longer constant for times $t > \omega_{2nk}^{-1}$, when matter wave effects begin to play a role. At such times the phase grating created by the pulses can be transformed into amplitude gratings of the atomic density.

Long interaction times are possible, limited only by the time the atoms spend in the field. As such, atom interferometers offer exciting new possibilities for precision measurements of rotation rates, fundamental constants such as $\hbar$, gravitational acceleration, and gravitational gradients. Moreover, one can take advantage of the nonlinear atom-field interaction to create atomic density patterns having period $\lambda/2n$ from a standing wave optical field having wavelength λ. In this manner one has the possibility to construct atom interferometers and atom structures that operate on the nano-scale (spacings on the order of tens of nanometers).

Conclusion

Optical coherent transients are now used routinely as probes of atomic vapors, liquids, and condensed matter. Sophisticated techniques have been developed using fast and ultrafast pulses (having duration of 10 to 100 femtoseconds) to probe such systems and obtain the relevant relaxation rates. The manner in which optical coherent transients can be used to extract information on transition frequencies and relaxation rates in atomic and molecular vapors has been reviewed. In atomic vapors, optical coherent transients provide a means for obtaining both longitudinal and transverse relaxation rates. Moreover, echo experiments serve as an extremely sensitive probe of velocity changing collisions. Similar techniques are being rediscovered as an important tool in the analysis of cold atomic vapors and Bose–Einstein condensates. Only the most basic optical coherent transient phenomena have been described in this chapter. More elaborate, multi-pulse excitation schemes are joining established methods in attempts to understand the complex decay dynamics that one encounters in many branches of physics, chemistry and biology.

Acknowledgments

The research of PRB is supported by the National Science Foundation under Grants No. PHY-0098016 and the FOCUS Center Grant, and by the U. S. Army Research Office under Grant No. DAAD19-00-1-0412.

List of Units and Nomenclature

$\mathbf{E}(\mathbf{R}, t)$	Electric field vector
$E(t)$	Electric field amplitude
$E_s(Z, t)$	Complex signal electric field amplitude
$I(L, t)$	Signal intensity exiting the sample
$\mathbf{k}$	Field propagation vector
L	Sample length
$\mathcal{N}$	Atomic density
$\mathbf{P}$	Center-of-mass momentum
$\mathbf{P}(\mathbf{R}, t)$	Polarization vector
$P(Z, t)$	Complex polarization field amplitude
T_{21}, T	Time interval between pulses
u	Most probable atomic speed

(u, v, w)	Elements of Bloch vector
$\mathbf{U}(t)$	Bloch vector
$\mathbf{v}$	Atomic velocity
$W_0(\mathbf{v})$	Atomic velocity distribution
γ	Transverse relaxation rate
γ_2	Excited state decay rate or longitudinal relaxation rate
δ	Atom-field detuning
θ	Pulse area
μ	Dipole moment matrix element
ρ_{ij}^I	Density matrix element in 'normal' interaction representation
$\rho_{ij}(Z, t)$	Density matrix element in a field interaction representation
τ	Pulse durations
ω	Field frequency
ω_k	Recoil frequency
ω_0	Atomic transition frequency
$\Omega(t)$	Generalized Rabi frequency
$\boldsymbol{\Omega}(t)$	Pseudofield vector
$\Omega_0(t)$	Rabi frequency

See also

Interferometry: Gravity Wave Detection; Overview.

Further Reading

Allen L and Eberly J (1975) *Optical Resonance and Two-Level Atoms*, Chapters 2,4,9, New York: Wiley.

Berman PR and Steel DG (2001) *Coherent optical transients*. In: Bass M, Enoch JM, Van Stryland E and Wolf WL (eds) Handbook of Optics, vol. IV. chap. 24 New York: McGraw-Hill. (Material presented in this chapter is a condensed version of this reference.).

Brewer RG (1977) Coherent optical spectroscopy. In: Balian R, Haroche S and Liberman S (eds) *Frontiers in Laser Spectroscopy, Les Houches, Session XXVII*, vol. 1, pp. 341–396. Amsterdam: North Holland.

Brewer RG (1977) Coherent optical spectroscopy. In: Bloembergen N (ed.) *Nonlinear Spectroscopy*, Proc. Int. School Phys., Enrico Fermi, Course 64, Amsterdam: North-Holland.

Brewer RG (1977) *Coherent optical transients*. In *Physics Today*, vol. 30(5), 50–59.

Brewer RG and Hahn EL (1984) Atomic memory. In *Scientific American*, vol. 251(6), 50–57.

Journal of Optical Society America (1986) Special issue on coherent optical spectroscopy, Vol. A3(4).

Levenson MD and Kano SS (1988) *Introduction to Nonlinear Laser Spectroscopy*, Chapter 6. Boston: Academic Press.

Meystre P and Sargent M III (1999) *Elements of Quantum Optics*, 3rd edn. Chapter 12. Berlin: Springer-Verlag.

Shen YR (1984) *The Principles of Nonlinear Optics*, Chapter 17. New York: Wiley.

Shoemaker RL (1978) Coherent transient infrared spectroscopy. In: Steinfeld JI (ed.) *Laser and Coherence Spectroscopy*, pp. 197–371. New York: Plenum.

Zinth W and Kaiser W (1988) *Ultrafast Coherent Spectroscopy*. In: Kaiser W (ed.) *Ultrashort Laser Pulses and Applications*, vol. 60, pp. 235–277. Berlin: Springer-Verlag.

Foundations of Coherent Transients in Semiconductors

T Meier and S W Koch, Philipps-University, Marburg, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Electromagnetic irradiation of matter generates a coherent superposition of the excited quantum states. The attribute 'coherent' relates to the fact that the material excitations originally have a well defined phase dependence which is imposed by the phase of the excitation source, often a laser in the optical regime. Macroscopically, the generated superposition state can be described as an optical polarization which is determined by the transition amplitudes between the participating quantum states.

The optical polarization is a typical non-equilibrium quantity that decays to zero when a system relaxes to its equilibrium state. Coherent effects are therefore only observable in a certain time window after pulsed photo-excitation, or in the presence of a continuous-wave (cw) beam. As coherent transients one usually refers to phenomena that can be observed during or shortly after pulsed laser excitation and critically depend on the presence of the induced optical polarization.

Many materials such as atoms, molecules, metals, insulators, semiconductors including bulk crystals, heterostructures, and surfaces, as well as organic and biological structures are studied using coherent optical spectroscopy. Depending on the particular system, the states participating in the optical transitions, the interactions among them, and the resulting time-scale for the decay of the induced polarization may be very different. As a result, the time window during which coherent effects are observable can be as long as seconds for certain atomic transitions, or it may be as short as a few femtoseconds (10^{-15}s), e.g., for metals, surfaces, or highly excited semiconductors.

Coherent spectroscopy and the analysis of coherent transients has provided valuable information on the nature and dynamics of the optical excitations. Often it is possible to learn about the interaction processes among the photoexcitations and to follow the temporal evolution of higher-order transitions which are only accessible if the system is in a non-equilibrium state.

The conceptually simplest experiment which one may use to observe coherent transients is to time resolve the transmission or reflection induced by a single laser pulse. However, much richer information can be obtained if one excites the system with several pulses. A pulse sequence with well-controlled delay times makes it possible to study the dynamical evolution of the photo-excitations, not only by time resolving the signal, but also by varying the delay. Prominent examples of such experiments are pump-probe measurements, which usually are performed with two incident pulses, or four-wave mixing, for which one may use two or three incident pulses.

Besides the microscopic interaction processes, the outcome of an experiment is determined by the quantum mechanical selection rules for the transitions and by the symmetries of the system under investigation. For example, if one wants to investigate coherent optical properties of surface states one often relies on phenomena, such as second-harmonic or sum-frequency generation, which give no signal in perfect systems with inversion symmetry. Due to the broken translational invariance, such experiments are therefore sensitive only to the dynamics of surface and/or interface excitations.

In this article the basic principles of coherent transients are presented and several examples are presented. The basic theoretical description and its generalization for the case of semiconductors are introduced.

Basic Principles

In the absence of free charges and currents, Maxwell's equations show that the electromagnetic field interacts with matter via the optical polarization. This polarization **P**, or more precisely its second derivative with respect to time $(\partial^2/\partial t^2)\mathbf{P}$, appears as a source term in the wave equation for the electric field **E**. Consequently, if the system is optically thin such that propagation effects within the sample can be ignored and if measurements are performed in the far field region, i.e., at distances exceeding the characteristic optical wavelength λ, the emitted electric field resulting from the polarization is proportional to its second time derivative, $\mathbf{E} \propto (\partial^2/\partial t^2)\mathbf{P}$. Thus the measurement of the emitted field dynamics yields information about the temporal evolution of the optical material polarization.

Microscopically the polarization is determined by the transition amplitudes between the different states of the system. These may be the discrete states of atoms or molecules, or the microscopic valence and conduction band states in a dielectric medium, such as a semiconductor. In any case, the macroscopic polarization **P** is computed by summing over all microscopic transitions p_{cv} via $\mathbf{P} = \sum_{c,v} (\boldsymbol{\mu}_{cv} p_{cv} + c.c.)$, where $\boldsymbol{\mu}_{cv}$ is the dipole matrix element which determines the strength of the transitions between the states v and c, and *c.c.* denotes the complex conjugate. If ε_c and ε_v are the energies of these states, their dynamic quantum mechanical evolution is described by the phase factors $e^{-i\varepsilon_c t/\hbar}$ and $e^{-i\varepsilon_v t/\hbar}$, respectively. Therefore, each p_{cv} is evolving in time according to $e^{-i(\varepsilon_c - \varepsilon_v)t/\hbar}$. Assuming that we start at $t = 0$ with $p_{cv}(t=0) = p_{cv,0}$, which may be induced by a short optical pulse, we have for the optical polarization $\mathbf{P}(t) = \sum_{c,v} (\boldsymbol{\mu}_{cv} p_{cv,0} e^{-i(\varepsilon_c - \varepsilon_v)t/\hbar} + c.c.)$. Thus $\mathbf{P}(t)$ is given by a summation over microscopic transitions which all oscillate with frequencies proportional to the energy differences between the involved states. Hence, the optical polarization is clearly a coherent quantity which is characterized by amplitude and phase. Furthermore, the microscopic contributions to $\mathbf{P}(t)$ add up coherently. Depending on the phase relationships, one may obtain either constructive superposition, interference phenomena like quantum beats, or destructive interference leading to a decay (dephasing) of the macroscopic polarization.

Optical Bloch Equations

The dynamics of photo-excited systems can be conveniently described by a set of equations, the optical Bloch equations, named after Felix Bloch (1905–1983) who first formulated such equations to analyze the spin dynamics in nuclear magnetic resonance. For the simple case of a two-level model, the Bloch equations can be written as

$$i\hbar \frac{\partial}{\partial t} p = \Delta\varepsilon p + \mathbf{E}\cdot\boldsymbol{\mu} I \qquad [1]$$

$$i\hbar \frac{\partial}{\partial t} I = 2\mathbf{E}\cdot\boldsymbol{\mu}(p - p^*) \qquad [2]$$

Here, $\Delta\varepsilon$ is the energy difference and I the inversion, i.e., the occupation difference between upper and lower state. The field **E** couples the polarization to the product of the Rabi energy $\mathbf{E}\cdot\boldsymbol{\mu}$

and the inversion I. In the absence of the driving field, i.e., $\mathbf{E} \equiv 0$, eqn [1] describes the free oscillation of p discussed above.

The inversion is determined by the combined action of the Rabi energy and the transition p. The total occupation N, i.e., the sum of the occupations of the lower and the upper states, remains unchanged during the optical excitation since light does not create or destroy electrons; it only transfers them between different states. If N is initially normalized to 1, then I can vary between -1 and 1. $I = -1$ corresponds to the ground state of the system, where only the lower state is occupied. In the opposite extreme, i.e., for $I = 1$, the occupation of the upper state is 1 and the lower state is completely depleted.

One can show that eqns [1]–[2] contain another conservation law. Introducing $\mathcal{P} = p + p^* = 2Re[p]$ and $\mathcal{J} = i(p - p^*) = -2Im[p]$, which are real quantities and often named polarization and polarization current, respectively, one finds that the relation

$$\mathcal{P}^2 + \mathcal{J}^2 + I^2 = 1 \qquad [3]$$

is fulfilled. Thus the coupled coherent dynamics of the transition and the inversion can be described on a unit sphere, the so-called Bloch sphere. The three terms $\mathcal{P}, \mathcal{J}$, and I can be used to define a three-dimensional Bloch vector $\mathbf{S} = (\mathcal{P}, \mathcal{J}, I)$. With these definitions, one can reformulate eqns [1]–[2] as

$$\frac{\partial}{\partial t}\mathbf{S} = \boldsymbol{\Omega} \times \mathbf{S} \qquad [4]$$

with $\boldsymbol{\Omega} = (-2\boldsymbol{\mu}\cdot\mathbf{E}, 0, \Delta\varepsilon)/\hbar$ and × denoting the vector product.

The structure of eqn [4] is mathematically identical to the equations describing either the angular momentum dynamics in the presence of a torque or the spin dynamics in a magnetic field. Therefore, the vector $\mathbf{S}$ is often called pseudospin. Moreover, many effects which can be observed in magnetic resonance experiments, e.g., free decay, quantum beats, echoes, etc. have their counterparts in photo-excited optical systems.

The vector product on the right-hand side of eqn [4] shows that $\mathbf{S}$ is changed by $\boldsymbol{\Omega}$ in a direction perpendicular to $\mathbf{S}$ and $\boldsymbol{\Omega}$. For vanishing field the torque $\boldsymbol{\Omega}$ has only a z-component. In this case the z-component of $\mathbf{S}$, i.e., the inversion, remains constant, whereas the x- and y-components, i.e., the polarization and the polarization current, oscillate with the frequency $\Delta\varepsilon/\hbar$ on the Bloch sphere around $\boldsymbol{\Omega}$.

Semiconductor Bloch Equations

If one wants to analyze optical excitations in crystallized solids, i.e., in systems which are characterized by a periodic arrangement of atoms, one has to include the continuous dispersion (bandstructure) into the description. In translationally invariant, i.e., ordered systems, this can be done by including the crystal momentum $\hbar\mathbf{k}$ as an index to the quantities which describe the optical excitation of the material, e.g., p and I in eqns [1]–[2]. Hence, one has to consider a separate two-level system for each crystal momentum $\hbar\mathbf{k}$. As long as all other interactions are disregarded, the bandstructure simply introduces a summation over the uncoupled contributions from the different $\mathbf{k}$ states, leading to inhomogeneous broadening due to the presence of a range of transition frequencies $\Delta\varepsilon(\mathbf{k})/\hbar$.

For any realistic description of optical processes in solids, it is essential to go beyond the simple picture of non-interacting states and to treat the interactions among the elementary material excitations, e.g., the Coulomb interaction between the electrons and the coupling to additional degrees of freedom, such as lattice vibrations (electron–phonon interaction) or other bath-like subsystems. If crystals are not ideally periodic, imperfections, which can often be described as a disorder potential, need to be considered as well.

All these effects can be treated by proper extensions of the optical Bloch equations introduced above. For semiconductors the resulting generalized equations are known as semiconductor Bloch equations, where the microscopic interactions are included at a certain level of approximation. For a two-band model of a semiconductor, these equations can be written schematically as

$$i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}} = \Delta\varepsilon_{\mathbf{k}}p_{\mathbf{k}} + \Omega_{\mathbf{k}}(n^c_{\mathbf{k}} - n^v_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}}|_{\mathrm{corr}} \qquad [5]$$

$$i\hbar\frac{\partial}{\partial t}n^c_{\mathbf{k}} = (\Omega^*_{\mathbf{k}}p_{\mathbf{k}} - \Omega_{\mathbf{k}}p^*_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}n^c_{\mathbf{k}}|_{\mathrm{corr}} \qquad [6]$$

$$i\hbar\frac{\partial}{\partial t}n^v_{\mathbf{k}} = -(\Omega^*_{\mathbf{k}}p_{\mathbf{k}} - \Omega_{\mathbf{k}}p^*_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}n^v_{\mathbf{k}}|_{\mathrm{corr}} \qquad [7]$$

Here $p_{\mathbf{k}}$ is the microscopic polarization and $n^c_{\mathbf{k}}$ and $n^v_{\mathbf{k}}$ are the electron populations in the conduction and valence bands (c and v), respectively. Due to the Coulomb interaction and possibly further processes, the transition energy $\Delta\varepsilon_{\mathbf{k}}$ and the Rabi energy $\Omega_{\mathbf{k}}$ both depend on the excitation state of the system, i.e., they are functions of the time-dependent polarizations and populations. This leads, in particular, to a coupling among the excitations for all different values of the crystals momentum $\hbar\mathbf{k}$. Consequently, in the presence of interactions, the optical excitations can no longer be described as independent two-level systems but have to be treated as a coupled many-body system. A prominent and important example in this context is the appearance of strong exciton resonances which, as a consequence of the Coulomb interaction, show up in the absorption spectra of semiconductors energetically below the fundamental bandgap.

The interaction effects lead to significant mathematical complications since they induce couplings between all the different quantum states in a system and introduce an infinite hierarchy of equations for the microscopic correlation functions. The terms given explicitly in eqns [5]–[7] arise in a treatment of the Coulomb interaction on the Hartree-Fock level. Whereas this level is sufficient to describe excitonic resonances, there are many additional effects, e.g., excitation-induced dephasing due to Coulomb scattering and significant contributions from higher-order correlations, like excitonic populations and biexcitonic resonances, which make it necessary to treat many-body correlation effects that are beyond the Hartree-Fock level. These contributions are formally included by the terms denoted as $|_{\text{corr}}$ in eqns [5]–[7]. The systematic truncation of the many-body hierarchy and the analysis of controlled approximations is the basic problem in the microscopic theory of optical processes in condensed matter systems.

Examples

Radiative Decay

The emitted field originating from the non-equilibrium coherent polarization of a photo-excited system can be monitored by measuring the transmission and the reflection as function of time. If only a single isolated transition is excited, the dynamic evolution of the polarization and therefore of the transmission and reflection is governed by radiative decay. This decay is a consequence of the coupling of the optical transition to the light field described by the combination of Maxwell- and Bloch-equations. Radiative decay simply means that the optical polarization is converted into a light field on a characteristic time scale $2T_{\text{rad}}$. Here, T_{rad} is the population decay time, often also denoted as T_1-time. This radiative decay is a fundamental process which limits the time on which coherent effects are observable for any photo-excited system. Due to other mechanisms, however, the non-equilibrium polarization often vanishes considerably faster.

The value of T_{rad} is determined by the dipole matrix element and the frequency of the transition, i.e., $T_{\text{rad}}^{-1} \propto |\boldsymbol{\mu}|^2 \Delta\omega$, with $\Delta\omega = \Delta\varepsilon/\hbar$. The temporal evolution of the polarization and the emitted field are proportional to $e^{-i\Delta\omega t - t/(2T_{\text{rad}})}$. Usually one measures the intensity of a field, i.e., its squared modulus, which evolves as $e^{-t/T_{\text{rad}}}$ and thus simply shows an exponential decay, the so-called *free-induction decay* (see Figure 1). T_{rad} can be very long for transitions with small matrix elements. For semiconductor quantum wells, however, it is in the order of only 10 ps, as the result of the strong light-matter interaction.

The time constant on which the optical polarization decays is often called T_2. In the case where this decay is dominated by radiative processes, we thus have $T_2 = 2T_{\text{rad}}$.

Superradiance

The phenomenon of superradiance can be discussed considering an ensemble of N two-level systems which are localized at certain positions $\mathbf{R}_i$. In this case

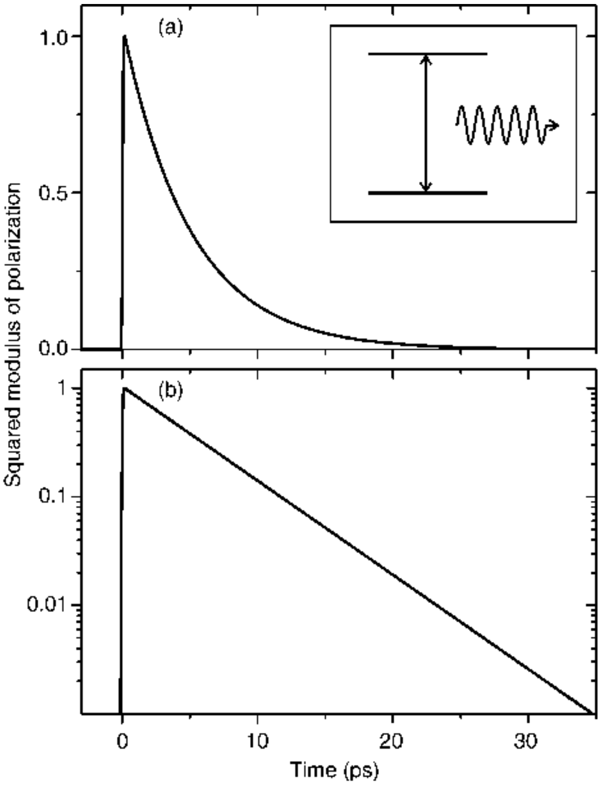

Figure 1 A two-level system is excited by a short optical pulse at $t = 0$. Due to radiative decay (inset) and possibly other dephasing mechanisms, the polarization decays exponentially as function of time with the time constant T_2, the dephasing time. The intensity of the optical field which is emitted as a result of this decay is proportional to the squared modulus of the polarization, which falls off with the time constant $T_2/2$. The decay of the squared modulus of the polarization is shown in (a) on a linear scale and in (b) on a logarithmic scale. The dephasing time was chosen to be $T_2 = 10$ ps which models the radiative decay of the exciton transition of a semiconductor quantum well.

Maxwell's equations introduce a coupling among all these resonances since the field emitted from any specific resonance interacts with all other resonances and interferes with their emitted fields. As a result, this system is characterized by N eigenmodes originating from the radiatively coupled optical resonances.

A very spectacular situation arises if one considers N identical two-level systems regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance, i.e., $\lambda = c/\Delta\omega$, where c is the speed of light in the considered material (see Figure 2a). In this case all emitted fields interfere constructively and the system behaves effectively like a single two-level system with a matrix element increased by $\sqrt{N}$. Consequently, the radiative decay rate is increased by N and the polarization of the coupled system decays N-times faster than that of an isolated system (see Figure 2b). This effect is called superradiance.

It is possible to observe superradiant coupling effects, e.g., in suitably designed semiconductor heterostructures. Figure 2c compares the measured time-resolved reflection from a single quantum well (dashed line) with that of a Bragg structure, i.e., a multiple quantum-well structure which consists of 10 individual wells that are separated by $\lambda/2$ (solid line), where λ is the wavelength of the exciton resonance. For times greater than about 2 ps the direct reflection of the exciting pulse has decayed sufficiently and one is left with the exponential decay of the remaining signal. Figure 2c shows that this decay is much more rapid for the Bragg structure than for the single quantum well, due to superradiance introduced by the radiative coupling among the quantum wells.

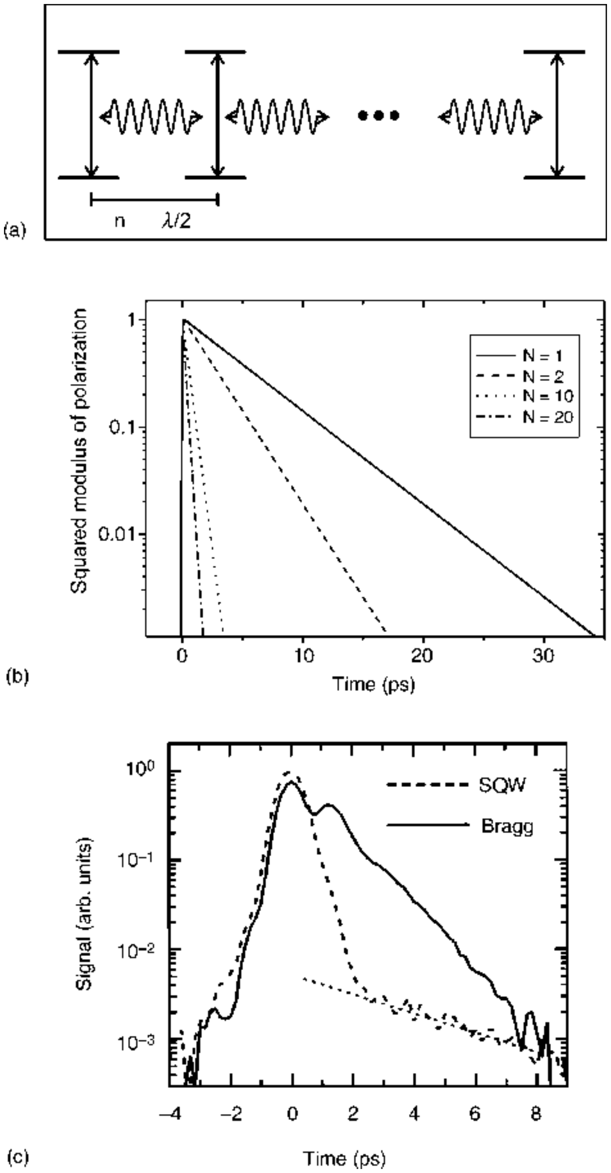

Figure 2 (a) N identical two-level systems are depicted which are regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance. Due to their coupling via Maxwell's equations all fields emitted from the two-level systems interfere constructively and the coupled system behaves effectively like a single two-level system with an optical matrix element increased by $\sqrt{N}$. (b) The temporal decay of the squared modulus of the polarization is shown on a logarithmic scale. The radiative decay rate is proportional to N and thus the polarization of the coupled system decays N-times faster than that of an isolated system. This effect is called superradiance. (c) Measured time-resolved reflection of a semiconductor single quantum well (dashed line) and a $N = 10$ Bragg structure, i.e., a multi quantum well where the individual wells are separated by $\lambda/2$ (solid line), on a logarithmic scale. [Part (c) is reproduced with permission from Haas S, Stroucken T, Hübner M, *et al.* (1998) Intensity dependence of superradiant emission from radiatively coupled excitons in multiple-quantum-well Bragg structures. *Physical Review B* 57: 14860. Copyright (1998) by the American Physical Society].

Destructive Interference

As the next example we now consider a distribution of two-level systems which have slightly different transition frequencies characterized by the distribution function $g(\Delta\omega - \bar{\omega})$ of the transition frequencies which is peaked at the mean value $\bar{\omega}$ and has a spectral width of $\delta\omega$ (see Figure 3a). Ignoring the radiative coupling among the resonances, the optical polarization of the total system evolves after excitation with a short laser pulse at $t = 0$ proportional to $\int d\omega\, g(\Delta\omega - \bar{\omega})e^{-i\Delta\omega t} \propto \tilde{g}(t)e^{-i\bar{\omega}t}$, where $\tilde{g}(t)$ denotes the Fourier transform of the frequency distribution function. $\tilde{g}(t)$ and thus the optical polarization decays on a time-scale which is inversely proportional to the spectral width of the distribution function $\delta\omega$. Thus the destructive interference of many different transition frequencies results in a rapid decay of the polarization (see Figure 3b).

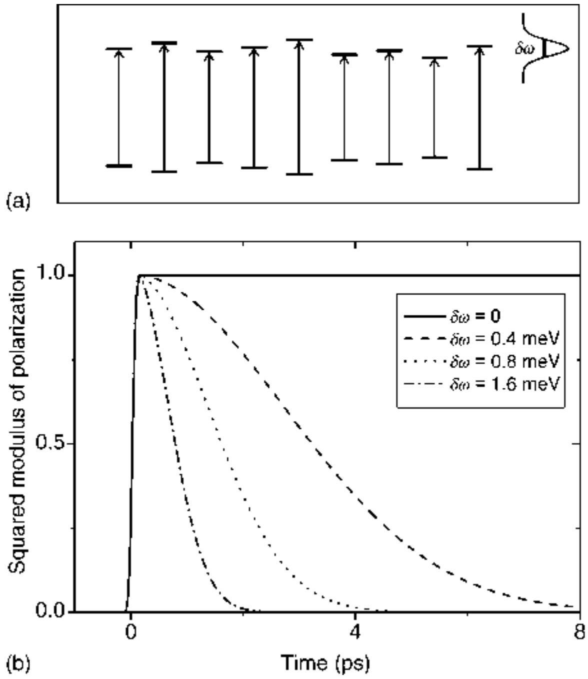

Figure 3 (a) An ensemble of two-level systems is shown where the resonance frequencies are randomly distributed according to a Gaussian distribution function of width $\delta\omega$ around an average value. (b) The temporal dynamics of the squared modulus of the polarization of ensembles of two-level systems after excitation with a short optical pulse at $t = 0$ is shown on a linear scale. Since the dephasing time is set to infinity, i.e., $T_2 \to \infty$, the polarization of an ensemble of identical two-level system ($\delta\omega = 0$) does not decay. However, for a finite width of the distribution, $\delta\omega > 0$, the individual polarizations of the ensemble oscillate with different frequencies and, therefore, due to destructive interference the polarization of the ensemble decays as function of time. Since the Fourier transform of a frequency-domain Gaussian is a Gaussian in the time domain, the dashed, dotted, and dashed dotted line have a Gaussian shape with a temporal width which is inversely proportional to $\delta\omega$.

In the spectral domain this rapid decay shows up as inhomogeneous broadening. Depending on the system under investigation, there are many different sources for such an inhomogeneous broadening. One example is the Doppler broadening in atomic gases or disorder effects such as well-width fluctuations in semiconductor quantum wells or lattice imperfections in crystals.

In the nonlinear optical regime it is, under certain circumstances, possible to reverse the destructive interference of inhomogeneously broadened coherent polarizations. For example, in four-wave mixing, a second pulse may lead to a rephasing of the contributions with different frequencies, which results in the photon echo (see discussion below).

Quantum Beats

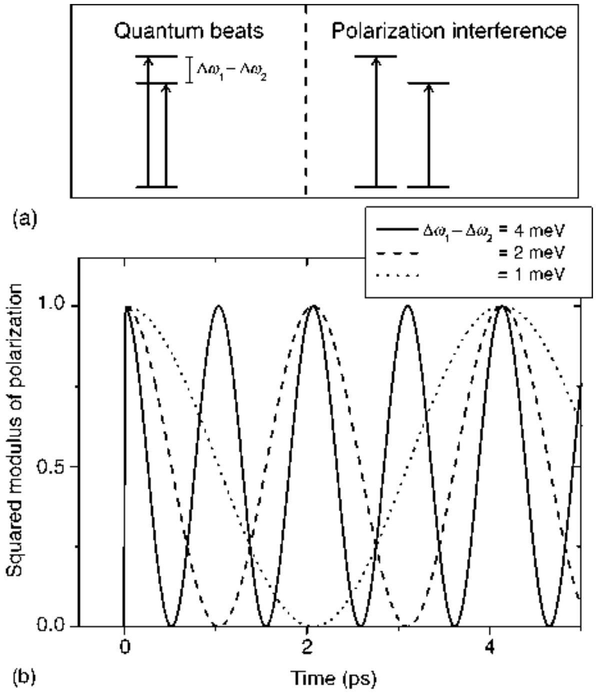

Figure 4 (a) Two optical resonances with frequency difference $\Delta\omega_1 - \Delta\omega_2$ may be realized by either a three-level system or by two uncoupled two-level systems. After impulsive excitation with an optical pulse the linear polarization of both types of systems show a modulation of the squared modulus of the polarization with the time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$. For the intrinsically coupled three-level system these modulations are called quantum beats, whereas for an interference of uncoupled systems they are named polarization interferences. Using nonlinear optical techniques, e.g., four-wave mixing and pump probe, it is possible to distinguish quantum beats and polarization interferences since coupled and uncoupled systems have different optical nonlinearities. (b) The temporal dynamics of the squared modulus of the polarization is shown for systems with two resonances and frequency difference $\Delta\omega_1 - \Delta\omega_2$, neglecting dephasing, i.e., setting $T_2 \to \infty$. After excitation with a short optical pulse at $t = 0$ the squared modulus of the polarization is periodically modulated with a time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$.

The occurrence of quantum beats can be understood most easily in a system where the total optical polarization can be attributed to a finite number of optical transitions. Let us assume for simplicity that all these transitions have the same matrix element. In this case, after excitation with a short laser pulse at $t = 0$ the optical polarization of the total system evolves proportional to $\sum_i e^{-i\Delta\omega_i t}$. The finite number of frequencies results in a temporal modulation with time periods $2\pi/(\Delta\omega_i - \Delta\omega_j)$ of the squared modulus of the polarization which is proportional to the emitted field intensity. For the case of two frequencies the squared modulus of the polarization is proportional to $[1 + \cos((\Delta\omega_1 - \Delta\omega_2)t)]$, i.e., due to the interference of two contributions the polarization varies between a maximum and zero (see Figure 4b).

In the linear optical regime it is impossible to distinguish whether the optical transitions are uncoupled or coupled. As shown in Figure 4, the two uncoupled two-level systems give the same linear polarization as a three-level systems where the two transitions share a common state. It is, however, possible to decide about the nature of the underlying transitions if one performs nonlinear optical spectroscopy. This is due to the fact that the so-called quantum beats, i.e., the temporal modulations of the polarization of an intrinsically coupled system, show a different temporal evolution as the so-called polarization interferences, i.e., the temporal modulations of the polarization of uncoupled systems. The former ones are also much more stable in the presence of inhomogeneous broadening than the latter ones.

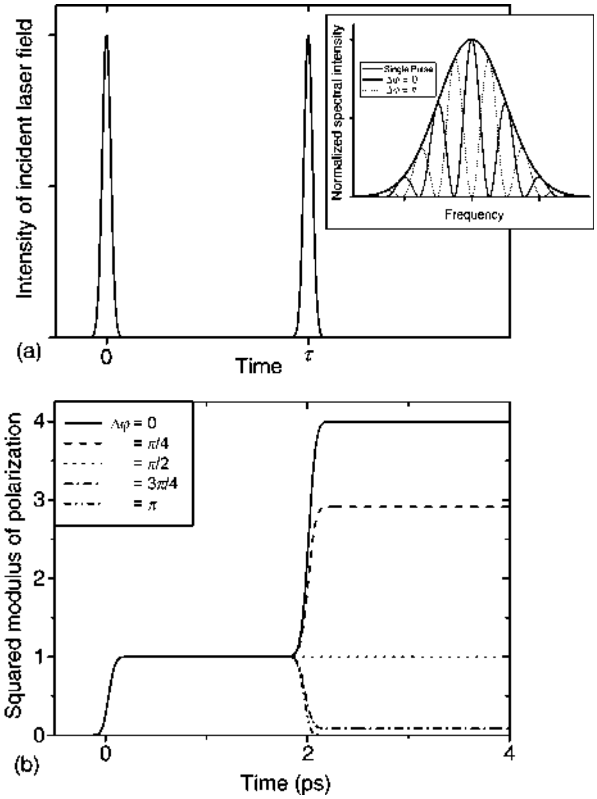

Figure 5 (a) Two Gaussian laser pulses separated by the time delay τ can lead to interferences. As shown by the thick solid line in the inset, the spectral intensity of a single pulse is a Gaussian with a maximum at the central frequency of the pulse. The width of this Gaussian is inversely proportional to the duration of the pulse. The spectral intensity of the field consisting of both pulses shows interference fringes, i.e., is modulated with a spectral period that is inversely proportional to τ. As shown by the thin solid and the dotted lines in the inset, the phase of the interference fringes depends on the phase difference $\Delta\varphi$ between the two pulses. Whereas for $\Delta\varphi = 0$ the spectral intensity has a maximum at the central frequency, it vanishes at this position for $\Delta\varphi = \pi$. (b) The temporal dynamics of the squared modulus of the polarization is shown for a two-level system excited by a pair of short laser pulses neglecting dephasing, i.e., setting $T_2 \rightarrow \infty$. The first pulse excites at $t = 0$ an optical polarization with a squared modulus normalized to 1. The second pulse, which has the same intensity as the first one, also excites an optical polarization at $t = 2$ ps. For $t > 2$ ps the total polarization is given by the sum of the polarizations induced by the two pulses. Due to interference, the squared modulus of the total polarization depends sensitively on the phase difference $\Delta\varphi$ between the two pulses. For constructive interference the squared modulus of the total polarization after the second pulse is four times bigger than that after the first pulse, whereas it vanishes for destructive interference. One may achieve all values in between these extremes by using phase differences $\Delta\varphi$ which are no multiples of π. It is thus shown that the second pulse can be used to coherently control the polarization induced by the first pulse.

In semiconductor heterostructures, quantum-beat spectroscopy has been widely used to investigate the temporal dynamics of excitonic resonances. Also the coupling among different optical resonances has been explored in pump-probe and four-wave-mixing measurements.

Coherent Control

In the coherent regime the polarization induced by a first laser pulse can be modified by a second, delayed pulse. For example, the first short laser pulse incident at $t = 0$ induces a polarization proportional to $e^{-i\Delta\omega t}$ and a second pulse arriving at $t = \tau > 0$ induces a polarization proportional to $e^{-i\varphi}e^{-i\Delta\omega(t-\tau)}$. Thus for $t \geq \tau$ the total polarization is given by $e^{-i\Delta\omega t}(1 + e^{-i\Delta\varphi})$ with $\Delta\varphi = \varphi - \Delta\omega\tau$. If $\Delta\varphi$ is an integer multiple of 2π, the polarizations of both pulses interfere constructively and the amplitude of the resulting polarization is doubled as compared to a single pulse. If, on the other hand, $\Delta\varphi$ is an odd multiple of π, the polarizations of both pulses interfere destructively and the resulting polarization vanishes after the action of both pulses. Thus by

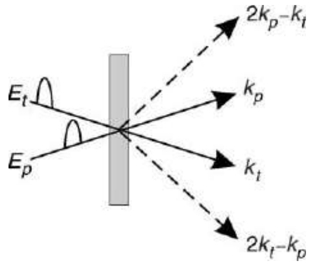

Figure 6 Transient optical nonlinearities can be investigated by exciting the system with two time delayed optical pulses. The pump pulse (E_p) puts the system in an excited state and the test (probe) pulse (E_t) is used to measure its dynamics. The dynamic nonlinear optical response may be measured in the transmitted or reflected directions of the pump and test pulses, i.e., $\pm k_p$ and $\pm k_t$, respectively. In a pump-probe experiment one often investigates the change of the test absorption induced by the pump pulse, by measuring the absorption in the direction k_t with and without E_p. One may also measure the dynamic nonlinear optical response in scattering directions like $2k_p - k_t$ and $2k_t - k_p$ as indicated by the dashed lines. This is what is done in a four-wave-mixing or photon-echo experiment.

varying the phase difference of the two pulses it is possible to coherently enhance or destroy the optical polarization (see Figure 5b).

One can easily understand the coherent control in the frequency domain by considering the overlap between the absorption of the system and the pulse spectrum. For excitation with two pulses that are temporally delayed by τ, the spectrum of the excitation shows interference fringes with a spectral oscillation period that is inversely proportional to τ. The positions of the maxima and the minima of the fringes depend on the phase difference between the two pulses. For constructive interference, the excitation spectrum is at a maximum at the resonance of the system, whereas for destructive interference the excitation spectrum vanishes at the resonance of the system (see Figure 5a).

Coherent control techniques have been applied to molecules and solids to control the dynamical evolution of electronic wavepackets and also the coupling to nuclear degrees of freedom. In this context, it is sometimes possible to steer certain chemical reactions into a preferred direction by using sequences of laser pulses which can be chirped, i.e., have a time-dependent frequency. Furthermore, in semiconductors and heterostructures, coherent control has been used to optically inject electronic currents on an ultrafast time-scale.

Transient Absorption Changes

In a typical pump-probe experiment one excites the system with a pump pulse (E_p) and probes its dynamics with a weak test pulse (E_t) (see Figure 6). With such experiments one often measures the differential absorption $\Delta\alpha(\omega)$ which is defined as

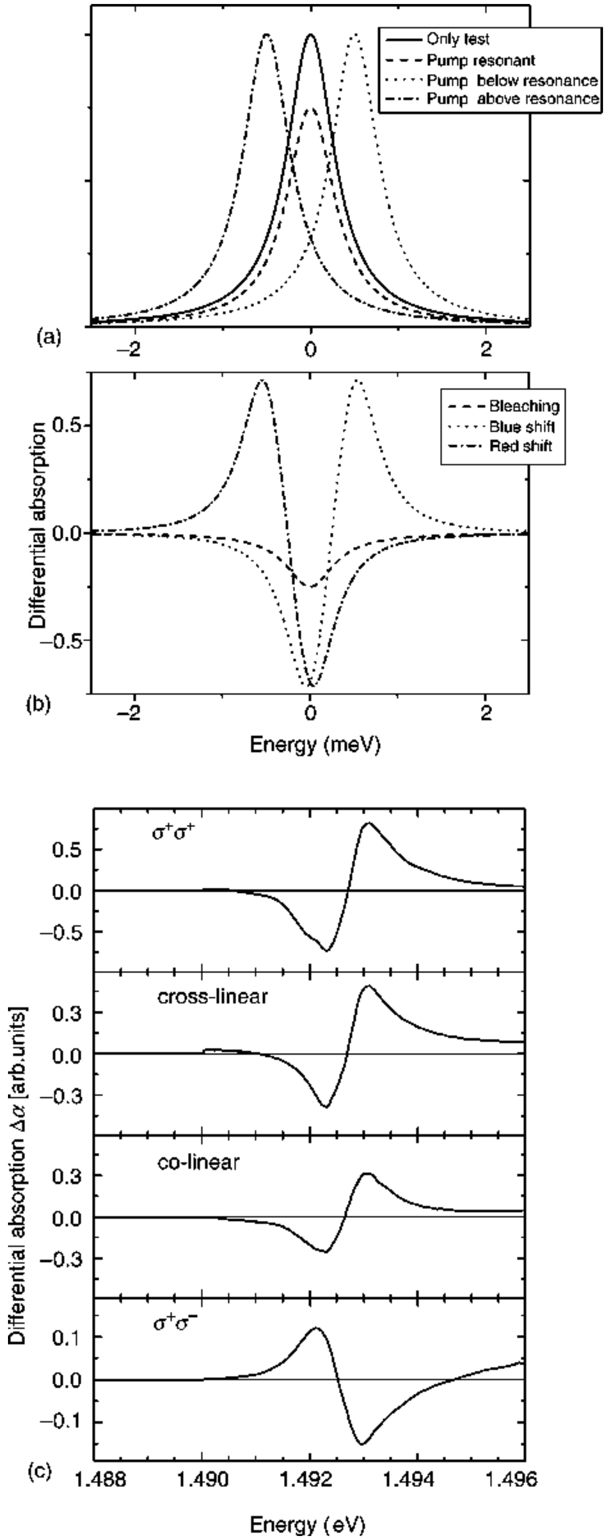

Figure 7 (a) Absorption spectra of a two-level system. The solid line shows the linear optical absorption spectrum as measured by a weak test pulse. It corresponds to a Lorentzian line that is centered at the transition frequency $\Delta\omega$, which has been set to zero. The width of the line is inversely proportional to the dephasing time T_2. In the presence of a pump pulse which resonantly excited the two-level system, the absorption monitored by the test pulse is reduced in amplitude, i.e., bleached, since the pump puts the system in an excited state (dashed line). In the optical Stark effect the frequency of the pump is nonresonant with the transition frequency. If the pump is tuned below (above) the transition frequency, the absorption is shifted to higher (lower) frequencies, i.e., shifted to the blue (red) part of the spectrum, see dotted (dashed-dotted) line. (b) The differential absorption obtained by taking the difference between the absorption in the presence of the pump and the absorption without pump. The dashed line shows the purely negative bleaching obtained using a resonant pump. The dotted and the dashed-dotted lines correspond to the dispersive shape of the blue and red shift obtained when pumping below and above the resonance, respectively. (c) Measured differential absorption spectra of a semiconductor quantum well which is pumped off resonantly 4.5 *meV* below the 1s heavy-hole exciton resonance. The four lines correspond to different polarization directions of the pump and probe pulses as indicated. [Part (c) is reproduced with permission from Sieh C, Meier T, Jahnke F, *et al.* (1999) Coulomb memory signatures in the excitonic optical Stark effect. *Physical Review Letters* 82: 3112. Copyright (1999) by the American Physical Society.]

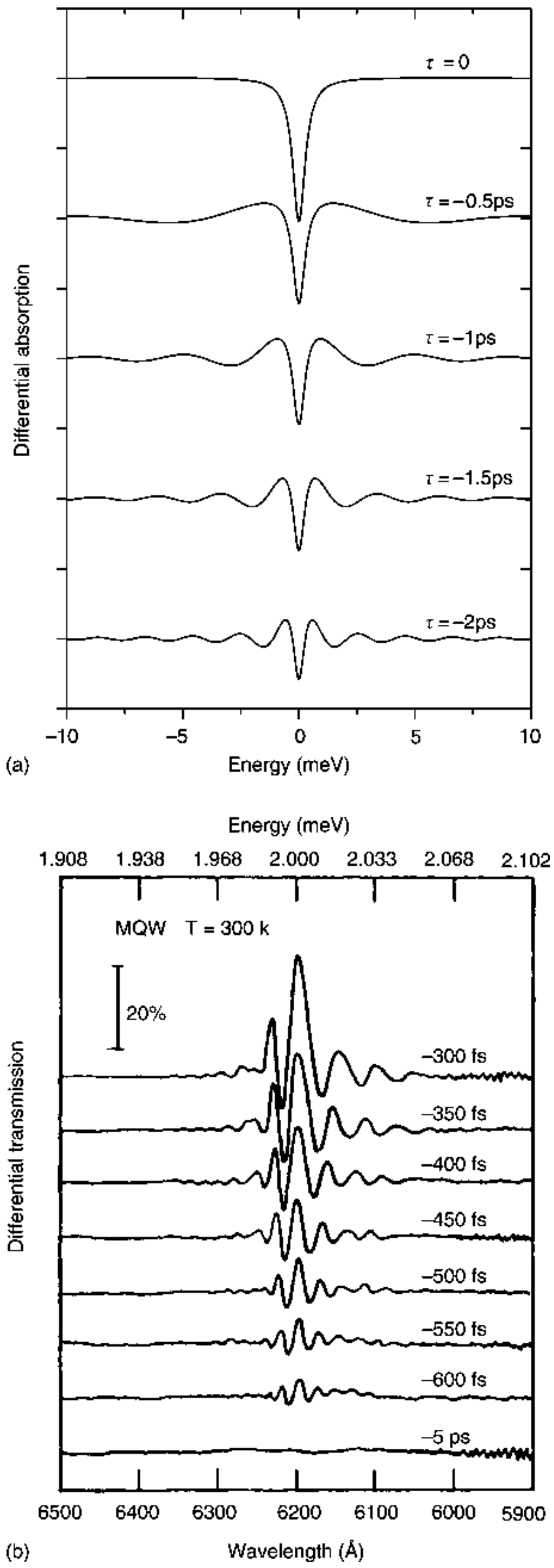

Figure 8 (a) Differential absorption spectra are shown for resonant pumping of a two-level system for various time delays τ between the pump and the test pulse. When the pump precedes the test, i.e., $\tau < 0$, the differential absorption exhibits spectral oscillations with a period which is inversely proportional to the time delay. When τ approaches zero, these spectral oscillations vanish and the differential absorption develops into purely negative bleaching. (b) Measured differential transmission spectra of a multi quantum well for different negative time delays as indicated. [Part (b) is reproduced with permission from Sokoloff JK, Joffre M, Fluegel B, *et al.* (1988) Transient oscillations in the vicinity of excitons and in the band of semiconductors. *Physics Review B* 38: 7615. Copyright (1988) by the American Physical Society.]

the difference between the probe absorption in the presence of the pump $\alpha_{\text{pump on}}(\omega)$ and the probe absorption without the pump $\alpha_{\text{pump off}}(\omega)$.

For resonant pumping and for a situation where the pump precedes the test (positive time delays $\tau > 0$), the absorption change is usually negative in the vicinity of the resonance frequency $\Delta\omega$, indicating the effect of absorption bleaching (see Figure 7a,b). There may be positive contributions spectrally around the original absorption line due to resonance broadening and, at other spectral positions, due to excited state absorption, i.e., optical transitions to energetically higher states which are only possible if the system is in an excited state. The bleaching and the positive contributions are generally present in coherent, and also in incoherent situations, where the polarization vanishes but occupations in excited states are present.

For detuned pumping, the resonance may be shifted by the light field as for example, in the optical Stark effect. Depending on the excitation configuration and the system, this transient shift may be to higher (blue shift) or lower energies (red shift) (see Figure 7a,b). With increasing pump-probe time delay the system gradually returns to its unexcited state and the absorption changes disappear.

As an illustration we show in Figure 7c experimentally measured differential absorption spectra of a semiconductor quantum well which is pumped spectrally below the exciton resonance. For a two-level system one would expect a blue shift of the absorption, i.e., a dispersive shape of the differential absorption with positive contributions above and negative contributions below the resonance frequency. This is indeed observed for most polarization directions of the pump and probe pulses. However, if both pulses are oppositely circularly polarized, the experiment shows a resonance shift in the reverse direction, i.e., a red shift. For the explanation of this behavior one has to consider the optical selection rules of the quantum well system. One finds that the signal should actually vanish for oppositely circularly

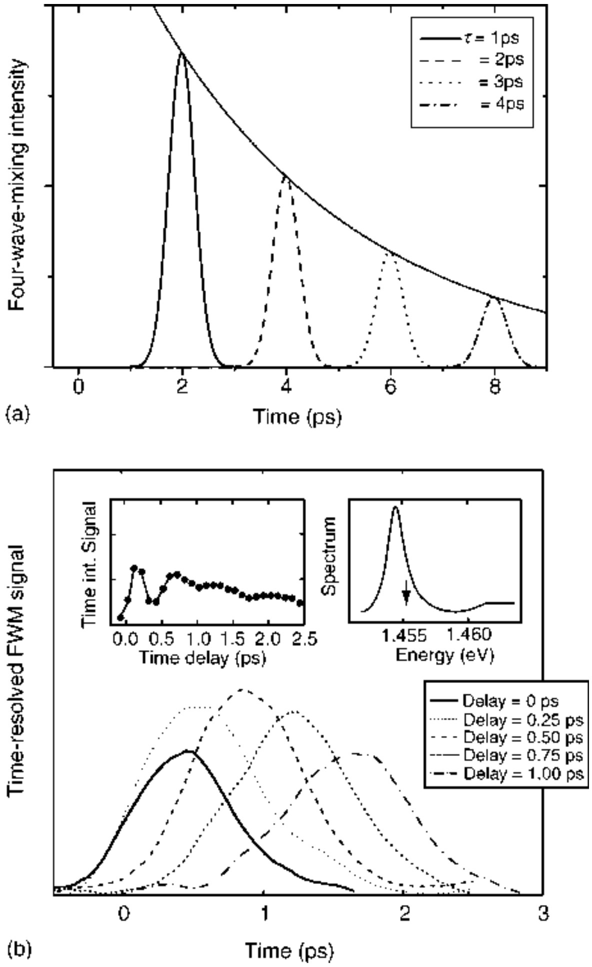

Figure 9 (a) The intensity dynamics as measured in a four-wave-mixing experiment on an ensemble of inhomogeneously broadened two-level systems for various time delays τ of the two incident pulses. The first pulse excites the system at $t = 0$. In the time period $0 < t < \tau$ the individual polarizations of the inhomogeneously broadened system oscillate with their respective frequencies and due to destructive interference the total polarization decays, cp. **Figure 3**. The second pulse arriving at $t = \tau$ leads to a partial rephasing of the individual polarizations. Due to phase conjugation all individual polarizations are in phase at $t = 2\tau$ which results in a macroscopic measurable signal, the photon echo. Due to dephasing processes the magnitude of the photon echo decays with increasing τ, see solid, dashed, dotted and dashed-dotted lines. Describing the dephasing via a T_2 time corresponds to an exponential decay as indicated by the thin dotted line. By measuring the decay of the echo with increasing τ, one thus gets experimental information on the dephasing of the optical polarization. (b) Time-resolved four-wave-mixing signal measured on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure for different time delays as indicated. The origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t = \tau$. The insets show the corresponding time-integrated four-wave-mixing signal and the linear absorption. [Part (b) is reproduced with permission from Jahnke F, Koch M, Meier T, *et al.* (1994) Simultaneous influence of disorder and Coulomb interaction on photon echoes in semiconductors. *Physics Review B* 50: 8114. Copyright (1994) by the American Physical Society.]

polarized pulses, as long as many-body correlations are neglected. Thus in this case the entire signal is due to many-body correlations and it is their dynamics which gives rise to the appearance of the red shift.

Spectral Oscillations

For negative time delays $\tau < 0$, i.e., if the test pulse precedes the pump, the absorption change $\Delta\alpha(\omega)$ is characterized by spectral oscillations around $\Delta\omega$ which vanish as τ approaches zero (see Figure 8a). The spectral period of the oscillations decreases with increasing $|\tau|$. Figure 8b shows measured differential transmission spectra of a multiple quantum-well structure for different negative time delays. As with the differential absorption, also the differential transmission spectra are dominated by spectral oscillations around the exciton resonance whose period and amplitude decrease with increasing $|\tau|$.

The physical origin of the coherent oscillations is the pump-induced perturbation of the free-induction decay of the polarization generated by the probe pulse. This perturbation is delayed by the time τ at which the pump arrives. The temporal shift of the induced polarization changes in the time domain leads to oscillations in the spectral domain, since the Fourier transformation translates a delay in one domain into a phase factor in the conjugate domain.

Photon Echo

In the nonlinear optical regime one may (partially) reverse the destructive interference of a coherent, inhomogeneously broadened polarization. For example, in four-wave mixing, which is often performed with two incident pulses, one measures the emitted field in a background-free scattering direction (see Figure 5). The first short laser pulse excites all transitions at $t = 0$. As a result of the inhomogeneous broadening the polarization decays due to destructive interference (see Figure 3). The second pulse arriving at $t = \tau > 0$ is able to conjugate the phases ($e^{i\varphi} \rightarrow e^{-i\varphi}$) of the individual polarizations of the inhomogeneously broadened system. The subsequent unperturbed dynamical evolution of the polarizations leads to a measurable macroscopic signal at $t = 2\tau$. This photon echo occurs since at this point in time all individual polarizations are in phase and add up constructively (see Figure 9a). Since this rephasing process leading to the echo is only possible as long as the individual polarizations remain coherent, one can analyze the loss of coherence (dephasing) by measuring the decay of the photon echo with increasing time delay.

Echo experiments are frequently performed using more complicated geometries, e.g., more than two laser pulses. Further echo phenomena can also be observed as the result of the coupling of the photoexcitation to nuclear degrees of freedom and also due to spatial motion of the electrons, i.e., electronic currents.

Time-resolved four-wave-mixing signal, measured on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure for different time delays, is presented in Figure 9b. Compared to Figure 9a, the origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t = \tau$. Due to the inhomogeneous broadening the maximum of the signal is shifting to longer times with increasing delay. Further details of the experimental results, in particular, the exact position of the maximum and the width of the signal, cannot be explained on the basis of non-interacting two-level systems, but require the analysis of many-body effects in the presence of inhomogeneous broadening.

See also

Coherent Control: Applications in Semiconductors; Experimental; Theory. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Allen L and Eberly JH (1975) *Optical Resonance and Two-Level Atoms*. New York: Wiley.

Baumert T, Helbing J, Gerber G, *et al.* (1997) Coherent control with femtosecond laser pulses. *Advanced Chemical Physics* 101: 47–82.

Bergmann K, Theuer H and Shore BW (1998) Coherent population transfer among quantum states of atoms and molecules. *Review of Modern Physics* 70: 1003–1025.

Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.

Brewer RG (1977) *Coherent Optical Spectroscopy*. In: Balian, *et al.* (ed.). Course 4 in Les Houches Session XXVII, *Frontiers in Laser Spectroscopy*, vol. 1. Amsterdam: North-Holland Publ.

Cohen-Tannoudji C, Dupont-Roc J and Grynberg G (1989) *Photons and Atoms*. New York: Wiley.

Haug H and Koch SW (2004) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, 4th edn. Singapore: World Scientific.

Koch SW, Peyghambarian N and Lindberg M (1988) Transient and steady-state optical nonlinearities in semiconductors. *Journal of Physics* C 21: 5229–5250.

Macomber JD (1976) *The Dynamics of Spectroscopic Transitions*. New York: Wiley.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford.

Peyghambarian N, Koch SW and Mysyrowicz A (1993) *Introduction to Semiconductor Optics*. Englewood Cliffs, NJ: Prentice-Hall.

Pötz W and Schroeder WA (eds) (1999) *Coherent Control in Atoms, Molecules, and Semiconductors*. Dordrecht: Kluwer.

Schäfer W and Wegener M (2002) *Semiconductor Optics and Transport Phenomena*. Berlin: Springer.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*, 2nd edn. Berlin: Springer.

Yeazell JA and Uzer T (eds) (2000) *The Physics and Chemistry of Wave Packets*. New York: Wiley.

Ultrafast Studies of Semiconductors

J Shah, Arlington, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Linear optical spectroscopy, using the techniques of absorption, transmission, reflection and light scattering, has provided invaluable information about the electronic and vibrational properties of atoms, molecules, and solids. Optical techniques also possess some additional unique strengths: the ability to

1. generate nonequilibrium distributions functions of electrons, holes, excitons, phonons, etc. in solids;
2. determine the distribution functions by optical spectroscopy;
3. determine the nonequilibrium distribution functions on femtosecond timescales;
4. determine femtosecond dynamics of carrier and exciton transport and tunneling; and
5. investigate interactions between various elementary excitations as well as many-body processes in semiconductors.

Researchers have exploited these unique strengths to gain fundamental new insights into nonequilibrium, nonlinear, and transport physics of semiconductors and their nanostructures over the past four decades.

A major focus of these efforts has been devoted to understanding how a semiconductor in thermodynamic equilibrium, excited by an ultrashort optical pulse, returns to the state of thermodynamic equilibrium. Four distinct regimes can be identified:

(a) the coherent regime in which a well-defined phase relationship exists between the excitation created by the optical pulse and the electromagnetic (optical) field creating it;
(b) the nonthermal regime in which the coherence (well-defined phase relationships) has been destroyed by various collision and interference processes but the distribution function of excitations is nonthermal (i.e., cannot be described by a Maxwell–Boltzmann distribution or its quantum (degenerate) counterparts);
(c) the hot carrier regime in which the distributions for various excitations are thermal, but with different characteristic temperatures for different excitations and thermal bath; and
(d) the isothermal regime in which the excitations and the thermal bath are at the same temperature but there is an excess of excitation (e.g., electron–hole pairs) compared to thermodynamic equilibrium.

Various physical processes take the semiconductor from one regime to the next, and provide information not only about the fundamental physics of semiconductors but also about the physics and ultimate performance limits of electronic, optoelectronic, and photonic devices.

In addition to coherent and relaxation dynamics, ultrafast studies of semiconductors provide new insights into tunneling and transport of carriers, and demonstrate novel quantum mechanical phenomena and coherent control in semiconductors.

The techniques used for such studies have also advanced considerably over the last four decades. Ultrafast lasers with pulse widths corresponding to only a few optical cycles (1 optical cycle ≈2.7 fs at 800 nm) have been developed. The response of the semiconductor to an ultrafast pulse, or a multiple of phase-locked pulses, can be investigated by measuring the dynamics of light emitted by the semiconductor using a streak camera (sub-ps time resolution) or a nonlinear technique such as luminescence up-conversion (time resolution determined by the laser pulse width for an appropriate nonlinear crystal). The response of the semiconductor can also be investigated using a number of two or three beam pump-probe techniques such as transmission, reflection, light scattering or four-wave mixing (FWM). Both the amplitude and the phase of the emitted radiation or the probe can be measured by phase-sensitive techniques such as spectral interferometry. The lateral transport of excitation can be investigated by using time-resolved spatial imaging and the vertical transport and tunneling of carriers can be investigated using the technique of 'optical markers'. Electro-optic sampling can be used for transmission/reflection studies or for measuring THz response of semiconductors. More details on these techniques, and indeed many topics discussed in this brief article, can be found in the Further Reading.

Coherent Dynamics

The coherent response of atoms and molecules is generally analyzed for an ensemble of independent (noninteracting) two-level systems. The statistical properties of the ensemble are described in terms of the density matrix operator whose diagonal components relate to population of the eigenstates and off-diagonal components relate to coherence of the superposition state. The time evolution of the density matrix is governed by the Liouville variant of the Schrödinger equation $i\hbar\dot{\rho} = [H, \rho]$ where the system Hamiltonian $H = H_0 + H_{\text{int}} + H_R$ is the sum of the unperturbed, interaction (between the radiation field and the two-level system), and relaxation Hamiltonians, respectively. Using a number of different assumptions and approximations, this equation of motion is transformed into the optical Bloch equations (OBE), which are then used to predict the coherent response of the system, often using iterative procedures, and analyze experimental results.

Semiconductors are considerably more complex. Coulomb interaction profoundly modifies the response of semiconductors, not only in the linear regime (e.g., strong exciton resonance in the absorption spectrum) but also in the coherent and nonlinear regime. The influence of Coulomb interaction may be introduced by renormalizing the electron and hole energies (i.e., introducing excitons), and by renormalizing the field–matter interaction strength by introducing a renormalized Rabi frequency. These changes and the relaxation time approximation for the relaxation Hamiltonian lead to an analog of the optical Bloch equations, the semiconductor Bloch equations which have been very successful in analyzing the coherent response of semiconductors.

In spite of this success, it must be stressed that the relaxation time approximation, and the Boltzmann kinetic approach on which it is based, are not valid under all conditions. The Boltzmann kinetic approach is based on the assumption that the duration of the collision is much shorter than the interval between the collisions; i.e., the collisions are instantaneous. In the non-Markovian regime where these assumptions are not valid, each collision does not strictly conserve the energy and momentum and the quantum kinetic approach becomes more appropriate. This is true for photo-excited semiconductors as well as for the quantum transport regime in semiconductors. Also, semiconductor Bloch equations have been further extended by including four-particle correlations. Finally, the most general approach to the description of the nonequilibrium and coherent response of semiconductors following excitation by an ultrashort laser pulse is based on nonequilibrium Green's functions.

Experimental studies on the coherent response of semiconductors can be broadly divided into three categories:

1. investigation of novel aspects of coherent response not found in other simpler systems such as atoms;
2. investigation of how the initial coherence is lost, to gain insights into dephasing and decoherence processes; and
3. coherent control of processes in semiconductors using phase-locked pulses.

Numerous elegant studies have been reported. This section presents some observations on the trends in this field.

Historically, the initial experiments focused on the decay of the coherent response (time-integrated FWM as a function of delays between the pulses), analyzed them in terms of the independent two-level model, and obtained very useful information about various collision processes and rates (exciton–exciton, exciton–carrier, carrier–carrier, exciton–phonon, etc.). It was soon realized, however, that the noninteracting two-level model is not appropriate for semiconductors because of the strong influence of Coulomb interactions. Elegant techniques were then developed to explore the nature of coherent response of semiconductors. These included investigation of time- and spectrally resolved coherent response, and of both the amplitude and phase of the response to complement intensity measurements. These studies provided fundamental new insights into the nature of semiconductors and many-body processes and interactions in semiconductors. Many of these observations were well explained by the semiconductor Bloch equations. When lasers with <10 fs pulse widths became laboratory tools, the dynamics was explored on a time-scale much shorter than the characteristic phonon oscillation period (≈ 115 fs for GaAs longitudinal optical phonons), the plasma frequency (≈ 150 fs for a carrier density of $5 \times 10^{17}\,\mathrm{cm}^{-3}$), and the typical electron–phonon collision interval (≈ 200 fs in GaAs). Some remarkable features of quantum kinetics, such as reversal of the electron–phonon collision, memory effects, and energy nonconservation, were demonstrated. More recent experiments have demonstrated the influence of four-particle correlations on coherent nonlinear response of semiconductors such as GaAs in the presence of a strong magnetic field.

The ability to generate phase-locked pulses with controllably variable separation between them provides an exciting opportunity to manipulate a variety of excitations and processes within a semiconductor. If the separation between the two phase-locked pulses is adjusted to be less than the dephasing time of the system under investigation, then the amplitudes of the excitations produced by the two phase-locked pulses can interfere either constructively or destructively depending on the relative phase of the carrier waves in the two pulses. The nature of interference changes as the second pulse is delayed over an optical period. A number of elegant experiments have demonstrated coherent control of exciton population, spin orientation, resonant emission, and electrical current. Phase-sensitive detection of the linear or nonlinear emission provides additional insights into different aspects of semiconductor physics.

These experimental and theoretical studies of ultrafast coherent response of semiconductors have led to fundamental new insights into the physics of semiconductors and their coherence properties. Discussion of novel coherent phenomena is given in a subsequent section. Coherent spectroscopy of semiconductors continues to be a vibrant research field.

Incoherent Relaxation Dynamics

The qualitative picture that emerges from investigation of coherent dynamics can be summarized as follows. Consider a direct gap semiconductor excited by an ultrashort laser pulse of duration τ_L (with a spectral width $h\Delta\nu_L$) centered at photon energy $h\nu_L$ larger than the semiconductor bandgap E_g. At the beginning of the pulse the semiconductor does not know the pulse duration so that coherent polarization is created over a spectral region much larger than $h\Delta\nu_L$. If there are no phase-destroying events during the pulse (dephasing rate$1/\tau_D \ll 1/\tau_L$), then a destructive interference destroys the coherent polarization away from $h\nu_L$ with increasing time during the excitation pulse. Thus, the coherent polarization exists only over the spectral region $h\Delta\nu_L$ at the end of the pulse. This coherence will be eventually destroyed by collisions and the semiconductor will be left in an incoherent (off-diagonal elements of the density matrix are 0) nonequilibrium state with peaked electron and hole population distributions whose central energies and energy widths are determined by $h\nu_L$, $h\Delta\nu_L$ and E_g, if the energy and momentum relaxation rates $(1/\tau_E, 1/\tau_M)$ are much smaller than the dephasing rates.

The simplifying assumptions that neatly separate the excitation, dephasing and relaxation regimes are obviously not realistic. A combination of all these (and some other) physical parameters will determine the state of a realistic system at the end of the exciting pulse. However, after times $\sim \tau_D$ following the laser pulse, the semiconductor can be described in terms of electron and hole populations whose distribution functions are not Maxwell–Boltzmann or Fermi–Direct type, but nonthermal in a large majority of cases in which the energy and momentum relaxation rates are much smaller than the dephasing rates. This section discusses the dynamics of this incoherent state.

Extensive theoretical work has been devoted to quantitatively understand these initial athermal distributions and their further temporal evolution. These include Monte Carlo simulations based on the Boltzmann kinetic equation approach (typically appropriate for time-scales longer than ~100 fs, the carrier–phonon interaction time) as well as the quantum kinetic approach (for times typically less than ~100 fs) discussed above. We present below some observations on this vast field of research.

Nonthermal Regime

A large number of physical processes determines the evolution of the nonthermal electron and hole populations generated by the optical pulse. These include electron–electron, hole–hole, and electron–hole collisions, including plasma effects if the density is high enough, intervalley scattering in the conduction band and intervalence band scattering in valence bands, intersub-band scattering in quantum wells, electron–phonon, and hole–phonon scattering processes. The complexity of the problem is evident when one considers that many of these processes occur on the same time-scale. Of these myriad processes, only carrier–phonon interactions and electron–hole recombinations (generally much slower) can alter the total energy in electronic systems. Under typical experimental conditions, the redistribution of energy takes place before substantial transfer of energy to the phonon system. Thus, the nonthermal distributions first become thermal distributions with the same total energy. Investigation of the dynamics of how the nonthermal distribution evolves into a thermal distribution provides valuable information about the nature of various scattering processes (other than carrier–phonon scattering) described above.

This discussion of separating the processes that conserve energy in the electronic system and those that transfer energy to other systems is obviously too simplistic and depends strongly on the nature of the problem one is considering. The challenge for the experimentalist is to devise experiments that can isolate these phenomena so each can be studied separately. If the laser energy is such that $h\nu_L - E_g < \hbar\omega_{LO}$, the optical phonon energy, then majority of the photo-excited electrons and holes do not have sufficient energy to emit an optical phonon, the fastest of the carrier–phonon interaction processes. The initial nonthermal distribution is then modified primarily by processes other than phonon scattering and can be studied experimentally without the strong influence of the carrier–phonon interactions. Such experiments have indeed been performed both in bulk and quantum well semiconductors. These experiments have exhibited spectral hole burning in the pump-probe transmission spectra and thus demonstrated that the initial carrier distributions are indeed nonthermal and evolve to thermal distributions. Such experiments have provided quantitative information about various carrier–carrier scattering rates as a function of carrier density. In addition, experiments with $h\nu_L - E_g > \hbar\omega_{LO}$ and $h\nu_L - E_g >$ intervalley separation have provided valuable information about intravalley as well as intervalley electron–phonon scattering rates. Experiments have been performed in a variety of different semiconductors, including bulk semiconductors and undoped and modulation-doped quantum wells. The latter provide insights into intersub-band scattering processes and quantitative information about the rates.

These measurements have shown that under typical experimental conditions, the initial nonthermal distribution evolves into a thermal distribution in times < or ~1 ps. However, the characteristic temperatures of the thermal distributions can be different for electrons and holes. Furthermore, different phonons may also have different characteristic temperatures, and may even be nonthermal even when the electronic distributions are thermal. This is the hot carrier regime that is discussed in the next subsection.

Hot Carrier Regime

The hot carriers, with characteristic temperatures T_c (T_e and T_h for electrons and holes, respectively), have high energy tails in the distribution functions that extend several kT_c higher than the respective Fermi energies. Some of these carriers have sufficient energy to emit an optical phonon. Since the carrier–optical phonon interaction rates and the phonon energies are rather large, this may be the most effective energy loss mechanism in many cases even though the number of such carriers is rather small. The emitted optical phonons have relatively small wavevectors and nonequilibrium populations larger than expected for the lattice temperature. These optical phonons are often referred to as hot phonons although nonequilibrium phonons may be a better term since they probably have nonthermal distributions. The dynamics of hot phonons will be discussed in the next subsection, but we mention here that in case of a large population of hot phonons, one must consider not only phonon emission but also phonon absorption. Acoustic phonons come into the picture at lower temperatures when the fraction of carriers that can emit optical phonons is extremely small or negligible, and also in the case of quantum wells with sub-band energy separation smaller than optical phonons.

The dynamics of hot carriers have been studied extensively to obtain a deeper understanding of carrier–phonon interactions and to obtain quantitative measures of the carrier–phonon interaction rates. Time resolved luminescence and pump-probe transmission spectroscopy are the primary tools used for such studies. For semiconductors like GaAs, the results indicate that polar optical phonon scattering (longitudinal optical phonons) dominates for electrons whereas polar and nonpolar optical phonon scattering contribute for holes. Electron–hole scattering is sufficiently strong in most cases to maintain a common electron and hole temperature for times $> \sim 1$ ps. Since many of these experiments are performed at relatively high densities, they provide important information about many-body effects such as screening. Comparison of bulk and quasi-two-dimensional semiconductors (quantum wells) has been a subject of considerable interest. The consensus appears to be that, in spite of significant differences in the nature of electronic states and phonons, similar processes with similar scattering rates dominate both types of semiconductors. These studies reveal that the energy loss rates are influenced by many factors such as the Pauli exclusion principle, degenerate (Fermi–Dirac) statistics, hot phonons, and screening and many-body aspects.

Hot carrier distribution can be generated not only by photo-excitation, but also by applying an electric field to a semiconductor. Although the process of creating the hot distributions is different, the processes by which such hot carriers cool to the lattice temperature are the same in two cases. Therefore, understanding obtained through one technique can be applied to the other case. In particular, the physical insights obtained from the optical excitation case can be extremely valuable for many electronic devices that operate at high electric fields, thus in the regime of hot carriers. Another important aspect is that electrons and holes can be investigated separately by using different doping. Furthermore, the energy loss rates can be determined quantitatively because the energy transferred from the electric field to the electrons or holes can be accurately determined by electrical measurements.

Isothermal Regime

Following the processes discussed above, the excitations in the semiconductor reach equilibrium with each other and the thermal bath. Recombination processes then return the semiconductor to its thermodynamic equilibrium. Radiative recombination typically occurs over a longer time-scale but there are some notable exceptions such as radiative recombination of excitons in quantum wells occurring on picosecond time-scales.

Hot Phonons

As discussed above, a large population of nonequilibrium optical phonons is created if a large density of hot carriers has sufficient energy to emit optical phonons. Understanding the dynamics of these nonequilibrium optical phonons is of intrinsic interest.

At low lattice temperatures, the equilibrium population of optical phonons is insignificant. The optical phonons created as a result of photo-excitation occupy a relatively narrow region of wavevector (k) space near $k = 0$ ($\sim 1\%$ of the Brillouin zone). This nonequilibrium phonon distribution can spread over a larger k-space by various processes, or anharmonically decay into multiple large-wavevector acoustic phonons. These acoustic phonons then scatter or decay into small-wavevector, low-energy acoustic phonons that are eventually thermalized.

Pump-probe Raman scattering provides the best means of investigating the dynamics of nonequilibrium optical phonons. Such studies, for bulk and quantum well semiconductors, have provided invaluable information about femtosecond dynamics of phonons. In particular, they have provided the rate at which the nonequilibrium phonons decay. However, this information is for one very small value of the phonon wavevector so the phonon distribution function within the optical phonon branch is not investigated. The large-wavevector acoustic phonons are even less accessible to experimental techniques. Measurements of thermal phonon propagation provide some information on this subject. Finally, the nature of phonons is considerably more complex in quantum wells, and some interesting results have been obtained on the dynamics of these phonons.

Tunneling and Transport Dynamics

Ultrafast studies have also made important contributions to tunneling and transport dynamics in semiconductors. Time-dependent imaging of the luminescing region of a sample excited by an ultrashort pulse provides information about spatial transport of carriers. Such a technique was used, for example, to demonstrate negative absolute mobility of electrons in p-modulation-doped quantum wells. The availability of near-field microscopes enhances the resolution of such techniques to measure lateral transport to the subwavelength case.

A different technique of 'optical markers' has been developed to investigate the dynamics of transport and tunneling in a direction perpendicular to the surface (vertical transport). This technique relies heavily on fabrication of semiconductor nanostructures with desired properties. The basic idea is to fabricate a sample in which different spatial regions have different spectral signatures. Thus if the carriers are created in one spatial region and are transported to another spatial region under the influence of an applied electric field, diffusion, or other processes, the transmission, reflection, and/or luminescence spectra of the sample change dynamically as the carriers are transported.

This technique has provided new insights into the physics of transport and tunneling. Investigation of perpendicular transport in graded-gap superlattices showed remarkable time-dependent changes in the spectra, and analysis of the results provided new insight into transport in a semiconductor of intermediate disorder. Dynamics of carrier capture in quantum wells from the barriers provided information not only about the fundamentals of capture dynamics, but also about how such dynamics affects the performance of semiconductor lasers with quantum well active regions.

The technique of optical markers has been applied successfully to investigate tunneling between the two quantum wells in an a-DQW (asymmetric double quantum well) structure in which two quantum wells of different widths (and hence different energy levels) are separated by a barrier. The separation between the energy levels of the system can be varied by applying an electric field perpendicular to the wells. In the absence of resonance between any energy levels, the wavefunction for a given energy level is localized in one well or the other. In this nonresonant case, it is possible to generate carriers in a selected quantum well by proper optical excitation. Dynamics of transfer of carriers to the other quantum well by tunneling can then be determined by measuring dynamic changes in the spectra. Such measurements have provided much insight into the nature and rates of tunneling, and demonstrated phonon resonances. Resonant tunneling, i.e., tunneling when two electronic levels are in resonance and are split due to strong coupling, has also been investigated extensively for both electrons and holes. One interesting insight obtained from such studies is that tunneling and relaxation must be considered in a unified framework, and not as sequential events, to explain the observations.

Novel Coherent Phenomena

Ultrafast spectroscopy of semiconductors has provided valuable insights into many other areas. We conclude this article by discussing some examples of how such techniques have demonstrated novel physical phenomena.

The first example once again considers an a-DQW biased in such a way that the lowest electron energy level in the wide quantum well is brought into resonance with the first excited electron level in the narrow quantum well. The corresponding hole energy levels are not in resonance. By choosing the optical excitation photon energy appropriately, the hole level is excited only in the wide well and the resonant electronic levels are excited in a linear superposition state such that the electrons at $t = 0$ also occupy only the wide quantum well. Since the two electron eigenstates have slightly different energies, their temporal evolutions are different with the result that the electron wavepacket oscillates back and forth between the two quantum wells in the absence of damping. The period of oscillation is determined by the splitting between the two resonant levels and can be controlled by an external electric field. Coherent oscillations of electronic wavepackets have indeed been experimentally observed by using the coherent nonlinear technique of four-wave mixing. Semiconductor quantum wells provide an excellent flexible system for such investigations.

Another example is the observation of Bloch oscillations in semiconductor superlattices. In 1928 Bloch demonstrated theoretically that an electron wavepacket composed of a superposition of states from a single energy band in a solid undergoes a periodic oscillation in energy and momentum space under certain conditions. An experimental demonstration of Bloch oscillations became possible only recently by applying ultrafast optical techniques to semiconductor superlattices. The large period of a superlattice (compared to the atomic period in a solid) makes it possible to satisfy the assumptions underlying Bloch's prediction. The experimental demonstration of Bloch oscillations was also performed using four-wave mixing techniques. This provides another prime example of the power of ultrafast optical studies.

Summary

Ultrafast spectroscopy of semiconductors and their nanostructures is an exciting field of research that has provided fundamental insights into important physical processes in semiconductors. This article has attempted to convey the breadth of this field and the diversity of physical processes addressed by this field. Many exciting developments in the field have been omitted out of necessity. The author hopes that this brief article inspires the readers to explore some of the Further Reading.

See also

Coherence: Coherence and Imaging. **Coherence Control:** Experimental; Theory. **Coherent Transients:** Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Scattering:** Scattering Theory.

Further Reading

Allen L and Eberly JH (1975) *Optical Resonance and Two-Level Atoms*. New York: Wiley Interscience.

Chemla DS (1999) Chapter 3. In: Garmire E and Kost AR (eds) *Nonlinear Optics in Semiconductors*, pp. 175–256. New York: Academic Press.

Chemla DS and Shah J (2001) Many-body and correlation effects in semiconductors. *Nature* 411: 549–557.

Haug H (2001) Quantum kinetics for femtosecond spectroscopy in semiconductors. In: Willardson RK and Weber ER (eds) *Semiconductors and Semimetals*, vol. 67, pp. 205–229. London: Academic Press.

Haug H and Jauho AP (1996) *Quantum Kinetics in Transport and Optics of Semiconductors*. Berlin: Springer-Verlag.

Haug H and Koch SW (1993) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. Singapore: World Scientific.

Kash JA and Tsang JC (1992) Chapter 3. In: Shank CV and Zakharchenya BP (eds) *Spectroscopy of Nonequilibrium Electrons and Phonons*, pp. 113–168. Amsterdam: Elsevier.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford University Press.

Shah J (1992) *Hot Carriers in Semiconductor Nanostructures: Physics and Applications*. Boston, MA: Academic Press.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. 2nd edn. Heidelberg, Germany: Springer-Verlag.

Steel DG, Wang H, Jiang M, Ferrio KB and Cundiff ST (1994) In: Phillips RT (ed.) *Coherent Optical Interactions in Solids*, pp. 157–180. New York: Plenum Press.

Wegener M and Schaefer W (2002) *Semiconductor Optics and Transport*. Heidelberg, Germany: Springer-Verlag.

Zimmermann R (1988) *Many-Particle Theory of Highly Excited Semiconductors*. Leipzig, Germany: Teubner.

COLOR AND THE WORLD

G N Rao, Adelphi University, Garden City, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The human visual system is capable of detecting a small part of the electromagnetic spectrum which is called light. The visible region has a typical wavelength range of about 380–780 nm. Under optimal conditions the range is 360–830 nm. The most important point is that color is a perception. It is how human beings perceive the visual part of the electromagnetic spectrum. The color of an object perceived depends on the physical characteristics of the radiation illuminating the object, spectral scattering and reflectivity of the object, physiology of the eye, photochemistry of the sensors in the retina, the complex data processing that takes place in the visual system, and the psychology of color perception. In this review, we present an overview of color in nature and the world. Initially, color and wavelength is presented, then the human visual system and color perception. Then we go on to discuss measuring color and colorimetry and then color in nature, followed by a brief account of color in art and holography and the structural colors in nature.

Color and Wavelength (Spectral and Nonspectral Colors)

The ingenious yet simple experiments performed by Isaac Newton in the year 1666 and his ideas presented in his famous treatise 'Opticks', are the foundations of our understanding of color and color perception. Based on his experiments, Newton showed that different regions of the visible spectrum are perceived as different colors and the solar radiation (white light) consists of spectral colors violet, blue, cyan, green, yellow, orange, and red. Thus, natural day light is quite colorful though we perceive it as colorless. Table 1 gives approximate wavelength ranges of the spectral colors. Superposition of all the spectral colors results in a perception of white.

Colors can be produced in a number of ways:

1. *Gas discharges* emit characteristic radiation in the visible region. For example, neon discharge is red in color, argon discharge blue, sodium discharge yellow, helium-neon laser gives red radiation and argon ion laser blue-green radiation. A variety of fluorescent lamps are available and they provide continuous spectral power distribution with characteristic lines of mercury in the blue-green region. Gas-filled sodium and fluorescent lamps offer long life and high efficiency and are popular for indoor and outdoor lighting.
2. *Transition metal* (Cr, Mn, Fe, Co, Ni, Cu) compounds or transition metal impurities are responsible for the colors of the minerals, paints, gems, and pigments. They involve inorganic compounds of transition metal or rare earth ions with unpaired d or f orbitals. About 1% chromium sesquioxide (Cr_2O_3) in colorless aluminum oxide results in beautiful red ruby, a gemstone. The Cr^{3+} impurity in pure beryllium aluminum silicate gives a beautiful emerald green resulting in sapphire, another gemstone. Transition metal impurities are added to glass in the molten state to prepare colored glass. For example, adding Cr^{3+} gives green, Mn^{3+} purple, Fe^{3+} pale yellow, Co^{3+} reddish blue, Ni^{2+} brown, and Cu^{2+} greenish blue color.
3. *Color in Organic Compounds* – Most natural and synthetic dyes and biological (vegetable and animal origin) colorants are complicated organic molecules. They absorb light, exciting molecular orbitals. The dyes are used for food coloration, clothing, photographic, and sensitizer purposes.
4. *Colors of metals, semiconductor materials and color centers* – The colors of these materials are caused by the electronic transitions involving the energy bands.

   Diode lasers emit radiation corresponding to the bandgap of the semiconductor materials used for their fabrication. They are inexpensive, compact, and can be operated using dry cell batteries. They are extensively used in compact disks, barcode readers, and optical communications.

   Light emitting diodes (LED) are also prepared using semiconductor materials and they emit radiation corresponding to the bandgap of the semiconductor materials used for their fabrication. LEDs are available in a variety of colors, red (GaAsP), orange (AlInGaP), yellow (InGaAs), green (CdSSe), blue (CdZnS) and are widely used in clocks, toys, tuners, displays, electronic billboards, and appliances.

   Color centers may be produced by irradiating some alkali halides and glass materials with electromagnetic radiation (gamma rays, X-rays, ultraviolet, and visible radiation) or with charged or uncharged particles (electrons, protons, neutrons). The irradiation produces an electron hole pair and the electron is trapped forming an electron center. The electron or hole, or both, form a color center absorbing part of the visible radiation. The color centers can be reversed by heating to high temperatures.
5. *Optical Phenomenon* lead to spectacular colors in nature and in biological materials. Scattering of light by the atmosphere is responsible for colors of the sky, beautiful sunsets and sunrise, twilight, blue moon, and urban glows. Dispersion and polarization cause rainbows and halos. Interference gives rise to the beautiful colors of thin films on water, foam, bubbles, and some biological colors (butterflies). Diffraction is responsible for the colors of liquid crystals, coronae, color of gems (opal), color of animals (sea mouse), the colors of butterfly (*Morpho rhetenor*), and diffraction gratings.

Table 1 Approximate wavelength ranges of different spectral colors

Color	*Wavelength range (nm)*
Violet	380–435
Blue	435–500
Cyan	500–520
Green	520–565
Yellow	565–597
Orange	597–625
Red	625–780

The Human Visual System and Color Perception

Photoreceptors in the Eye: Rods and Cones

As light enters the eye, the cornea and the eye lens focus the image on the retina. The retina has two kinds of photoreceptors: rods and cones so named because of their shapes. Typically, there are about 7 million cones and 110–125 million rods. Cones are activated during photopic (daylight) conditions and the rods during scotopic (night) conditions. Cones are responsible for color perception whereas rods can perceive only white and gray shades. Cones are predominantly located near the fovea, the central region of the retina. When light falls on the rods and the cones, the photosensitive material in them, namely rhodopsin, is activated, generating an electrical signal which is further processed in the visual system and transmitted to the visual cortex. Considerable data processing takes place within the retina.

Color Perception

Subjective nature of color

Color is a perception and is subjective. There is no way one can define color in an absolute or quantitative means, even though it can be precisely characterized by the spectral power distribution (SPD) of the radiation. For monochromatic light of different wavelengths we see the spectral colors detailed in Table 1. In general, the SPD of light from an object is complex and to predict the exact color of such a spectrum is quite difficult. However, in the following we discuss some simple color schemes.

Additive color mixing

Newton's experiments demonstrated that using the three primary color light sources, blue, green, and red and by varying their relative intensities, one can generate most of the colors by additive color mixing. Additive color mixing is employed in stage lighting, and large screen color television projection systems. If we are concerned only with the hue and ignore saturation and brightness, the superposition of colored light beams of equal brightness on a white screen gives the following simple additive color scheme:

$$\text{Green} + \text{Red} \Rightarrow \text{Yellow}$$

$$\text{Blue} + \text{Red} \Rightarrow \text{Magenta}$$

$$\text{Blue} + \text{Green} \Rightarrow \text{Cyan}$$

$$\text{Blue} + \text{Green} + \text{Red} \Rightarrow \text{White}$$

Subtractive color mixing

When white light passes through, say, a red filter, the radiation corresponding to the red radiation is transmitted and all other parts of the visible spectrum are absorbed. If we insert more red filters the brightness of the light transmitted becomes less though the hue is likely to be more saturated. If you insert red and blue filters together in a white beam, you will see no light transmitted because the red filter blocks the entire spectrum except the radiation corresponding to red color which in turn is blocked by the blue filter.

Subtractive color scheme is also applicable when paints of various colors are mixed together. The subtractive primary colors are magenta, yellow, and cyan. The artists often call them red, yellow, and blue. By appropriately mixing the primary colored paints one can, in principle, generate most of the colors.

How many colors can we differentiate?

If we do a simple experiment beginning from the lowest wavelength of the visible spectrum, say 380 nm, and gradually move to the long wavelength end of the visible spectrum, one can differentiate about 150 spectral colors. For each spectral color there will be a number of colors that vary in brightness and saturation. Additionally, we can have additive mixtures of these spectral colors, which in turn have variations in brightness and saturation. It is estimated that under optimal conditions we can differentiate as many as 7 to10 million different colors and shades.

Temporal Response of the Human Visual System

Positive afterimages

The visual system responds to changes in the stimulation in time. However, the response of the visual system is both delayed and persists for some time after the stimulation is off. Because of the persistence of vision, one would see positive afterimages. The afterimages last about 1/20 s under normal conditions but are shorter at high light conditions. A rapid succession of images show results in a perception of continuous motion which is the basis for movies and TV.

Close your eyes for about five minutes so that any after images of the objects you have seen before are cleared. Open your eyes to see a bright colored object and then close your eyes again; you will see positive afterimages of the object in its original colors. Even though the positive afterimage at the initial stages will be of the same colors as the original, they gradually change color showing that different cones recover at different rates.

Negative afterimages

If you stare at a color picture intensely for about a minute and look at a white background, such as a sheet of white paper, you will see the complementary colors of the picture. The image of the object falls on a particular part of the retina and if we look at a green colored object, the cones that are more sensitive to the mid-wavelength region are activated and are likely to be saturated. If you look at a bright sheet of white paper immediately, the cones that are sensitive to the middle wavelength region are inactivated since they need some time to replenish the active chemical whereas the cones that have maximum sensitivity for the low and high wavelengths are active, therefore you will see the object in complementary colors. Yellow will be seen after looking at blue image, cyan after looking at red image and vice versa.

Spatial Resolution of the Human Visual System

Lateral inhibition (color constancy)

By and large the perceived color of an object is independent of the nature of the illuminating white light source. The color of the object does not change markedly under different white light sources of illumination. For example, we see basically the same color of the object in the daytime in the light of the blue sky or under an incandescent light source even though the spectral power distributions of these sources are different and the intensity of light reaching the eye in both cases is also drastically different. The color constancy is a result of chromatic lateral inhibition. The human visual system has a remarkable way of comparing the signals from the object and manipulating them with the signals generated from the surroundings and picking up the correct color independent of the illuminating source. The color constancy and simultaneous lightness contrast are due to lateral inhibition. Lateral inhibition was exploited by a number of artists to provide certain visual effects that would not have been possible otherwise. For example, Georges Seurat in his famous painting 'La Poseuse en Profil' improved the light intensity contrast in different regions of his painting by employing edge enhancement effects resulting from lateral inhibition. Victor Vasarely's art work 'Arcturus' demonstrates the visual effects of lateral inhibition in different colors.

Spatial assimilation

The color of a region assimilates the color of the neighboring region, thus changing the perception of its color. Closely spaced colors, when viewed from a distance, tend to mix partitively and are not seen as distinct colors but as a uniform color due to additive color mixing. Georges Seurat in his painting 'Sunday Afternoon on the Island of La Grande Jatte', uses thousands of tiny dots of different colors to produce a variety of color sensations, a technique which is known as pointillism. The visual effects of spatial assimilation were clearly demonstrated by Richard Anuszkiewicz in his painting 'All Things Do Live in the Three'.

Color Perception Models

Thomas Young, in 1801, postulated that there are three types of sensors thus proposing trichromacy to explain the three attributes of color perception: hue, saturation, and brightness. Data on microspectrophotometry of excised retinas, reflection densitometry of normal eyes, and psychophysical studies of different observers confirm the existence of three types of cones. As shown in Figure 1, the three types of cones have different responsivities to light. The cones that have maximum responsivity in the short wavelength region are often refereed as S-cones, the cones that have maximum in the intermediate wavelength range as I-cones, and the cones that have maximum in the long wavelength region as L-cones.

The three response curves have considerable overlap. The overlap of the response curves and the complex data processing that takes place in the visual system enable us to differentiate as many as 150 spectral colors and about seven to ten million different colors and shades. An important consequence of the existence of only three types of color sensors is that different spectral power distributions can produce identical stimuli resulting in the perception of the same color. Such spectral power distributions are called metamers.

The neural impulses generated by the cones are processed through a complex cross-linking of bipolar cells, horizontal cells, amacrine cells, and communicated to ganglion cells leading to the optic chiasmas through the optic nerve. Information gathered by about 100 million photoreceptors, after considerable processing, is transmitted by about a million ganglion cells.

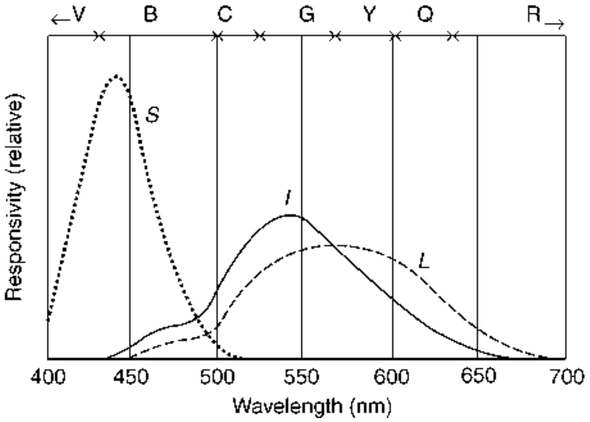

Figure 1 Relative spectral absorption of the three types of cones. *S*, *I*, and *L* respectively stand for the cones that have maximum responsivity for the short, intermediate, and long wavelength regions.

The zone theory of color vision assumes two separate zones where the signals are sequentially processed. In the first zone the signals are generated by the cones and in the second zone, the signals are processed to generate one achromatic and two opponent chromatic signals. Possibly other zones exist where the signals generated in the first and the second zones are further processed and communicated to the visual cortex where they are processed and interpreted based on the temporal and spatial response and comparisons with the information already available in the memory.

Measuring Color and Colorimetry

Since color is a human perception, how do we measure color? Colorimetry is based on the human visual system and how different observers perceive color. Though, the spectral power distribution of light is helpful for characterizing spectral hues, it is very difficult to infer the color from the spectral power distribution.

The three attributes of color perception are hue, saturation, and brightness. Hue is the color such as red, green, and blue. Saturation refers to the purity of the color. A color is said to be more saturated if the whiteness in the color is less and vice versa. Brightness is related to the intensity of light. For the sun at sunrise the hue is red, saturated, and low brightness. About half an hour after sunrise, the hue is yellow, less saturated, and higher brightness. Though a variety of color schemes are available in the literature, in the following we discuss the Munsell system and the C.I.E. diagram only.

Munsell Color System

Albert Munsell, a painter and art teacher, devised a color atlas in 1905 by arranging different colors in an ordered three-dimensional space. He used the three color attributes hue, chroma, and value corresponding to hue, saturation, and brightness. In this system there are 10 basic hues, each of which is divided into 10 equal gradations, resulting in 100 equally spaced hues. Each hue has a chart with a number of small chips arranged in rows and columns. Under daylight illumination, the chips in any column are supposed to have colors of equal chroma and the chips in any row are supposed to be of equal brightness. The brightness increases from the bottom of the chart to the top in steps that are perceptually equal. The saturation of the color increases from the inside edge to the outside edge of the chart in steps that are also perceptually equal. Each chip is identified by hue, value/chroma. In the Munsell scale, black is given a value of 0 and white 10, and the nine grays are uniformly placed in between. Though the Munsell system has gone through many revisions, typically it has a total of about 1450 chips. The Munsell color scheme, though subjective in nature, is widely used because of its simplicity and ease.

The Commission Internationale de l'Echlairage (C.I.E.) Diagram

Even though the visible light reaching the eye has a very complex SPD, color perception is mostly dependent on the signals generated by the three different types of cones. However, we do not know precisely the absorption and sensitivity of the cones and therefore are not in a position to calculate the signal strengths generated by the three different cones. Additionally, the response of the cones is specific to the individual. Therefore, the C.I.E. system is based on the average response of a large sample of normal observers. The C.I.E. system employs three color matching functions $\bar{x}(\lambda), \bar{y}(\lambda)$, and $\bar{z}(\lambda)$ (Figure 2), based on the psychological observations of a large number of standard observers to calculate the 'tristimulus values' X, Y, and Z defined as

$$X = c\int S(\lambda)\bar{x}(\lambda)\mathrm{d}\lambda \qquad [1]$$

$$Y = c\int S(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda \qquad [2]$$

$$Z = c\int S(\lambda)\bar{z}(\lambda)\mathrm{d}\lambda \qquad [3]$$

Here, λ is the wavelength, $S(\lambda)$ the spectral power distribution, and c is the normalization constant. The integration is carried out over the visible region, normally 380 to 780 nm. The system is based on the assumption of additivity and linearity. The sum of the tristimulus values $(X + Y + Z)$ is normalized to 1. We choose $y(\lambda)$ so that the Y tristimulus value is proportional to the luminance which is a quantitative measure of the intensity of light leaving the surface. We further calculate x and y which are called chromaticity coordinates:

$$x = \frac{X}{(X + Y + Z)} \qquad [4]$$

$$y = \frac{Y}{(X + Y + Z)} \qquad [5]$$

The psychological color is then specified by the coordinates (Y, x, y). Relationship between colors is usually displayed by plotting the x and y values on

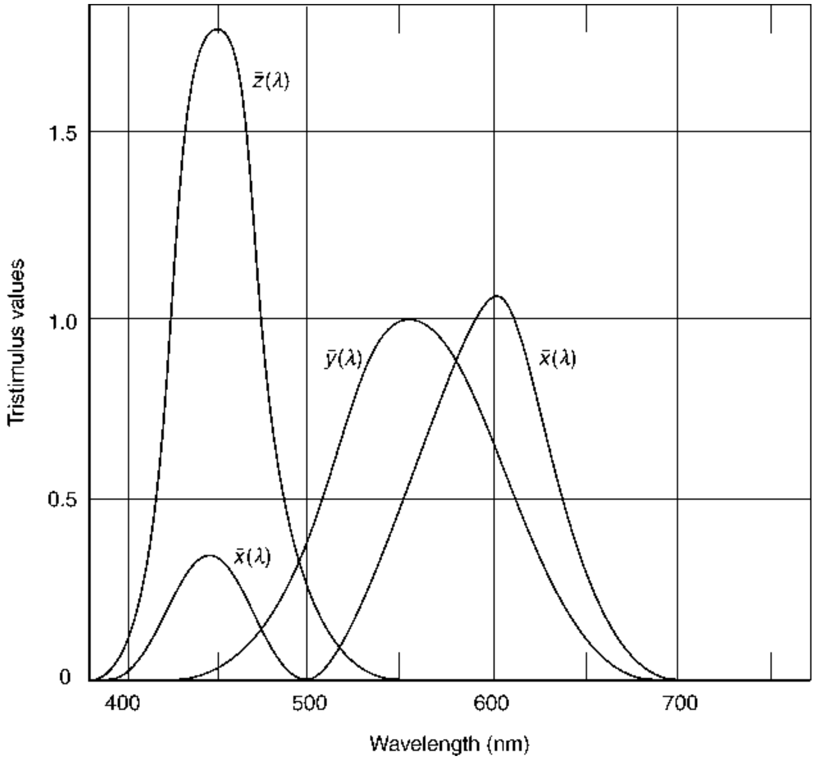

Figure 2 Color-matching functions $\bar{x}$ (λ),$\bar{y}$ (λ), and $\bar{z}$ (λ) of the C.I.E. 1931 standard observer.

a two-dimensional Cartesian Coordinate system with z as an implicit parameter ($z = 1 - x - y$) which is known as the C.I.E. chromaticity diagram (Figure 3). In this plot, the monochromatic hues are on the perimeter of a horseshoe shaped curve and is called spectrum locus. A straight line joining the ends of the spectrum is called the line of purples. Every chromaticity is represented by means of two coordinates (x, y) in the chromaticity diagram. The point corresponding to $x = 1/3$ and $y = 1/3$ (also $z = 1/3$) is the point of 'equal energy' and represents achromatic point. The complementary colors are on the opposite sides of the achromatic point. Complementary wavelength of a color is obtained by drawing a line from the color through the achromatic point to the perimeter on the other side.

Continuum (white) sources

A perfect white source is the one that has constant SPD over the entire visible region. Such a source will not distort the color of the objects they illuminate. A number of sources in general are considered as white sources; they include incandescent lamps, day skylight, a variety of fluorescent lamps, etc. It should be noted that the SPD of these sources is not constant and varies with the source. Typical SPDs of commonly used light sources are given in Figure 4.

Color temperature

The thermal radiation emitted by a black body has a characteristic SPD which depends solely on its temperature. The wavelength λ of the peak of this distribution is inversely proportional to its absolute temperature. As the absolute temperature of the blackbody increases, the peak shifts towards shorter wavelengths (blue region) and also the width of the distribution decreases. Since the peak of the black body radiation is dependent on the absolute temperature, the color of the blackbody can be approximately defined by its temperature.

The color temperature of a body is that temperature at which the SPD of the body best matches with that of the blackbody radiation at that temperature. The concept of color temperature is often used to characterize the sources. Low color temperature sources have a reddish tinge and high color temperature sources tend to be slightly bluish. For example, the color temperature of red looking star Antares is about 5000 K and that of the bluish white looking Sirius is about 11 000 K. One should note that the color temperature of a body is different from its actual temperature. The fluorescent lamps have a

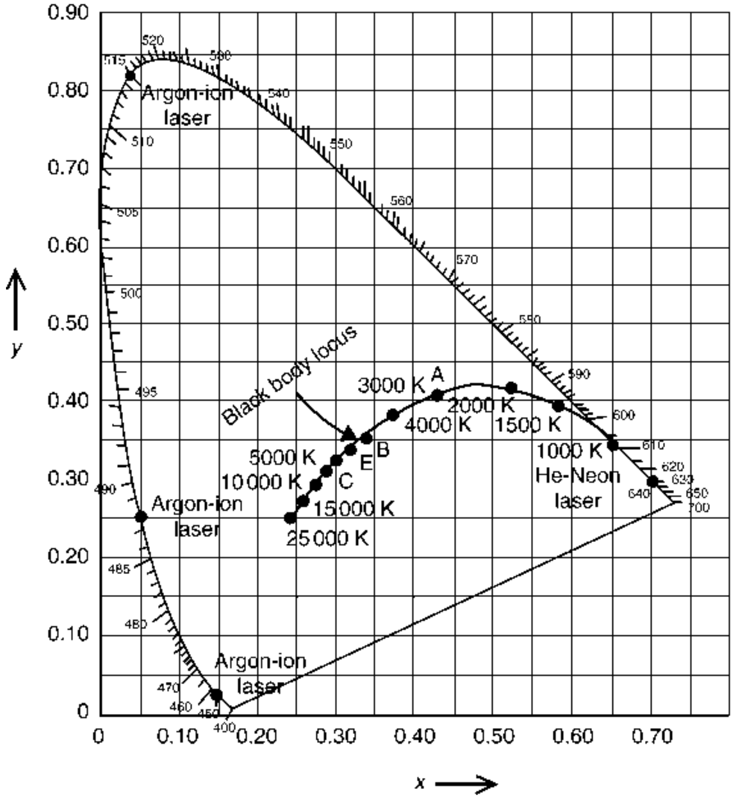

Figure 3 Chromaticity diagram of the C.I.E. 1931 standard observer. The chromaticities of the incandescent sources and their color temperatures are given inside the curve. A, B, and C are the C.I.E. standard illuminants. The chromaticities of the helium-neon laser (red, 633 nm) and the argon-ion laser (green 515 nm, blue 488 nm, and indigo, 458 nm) are also shown.

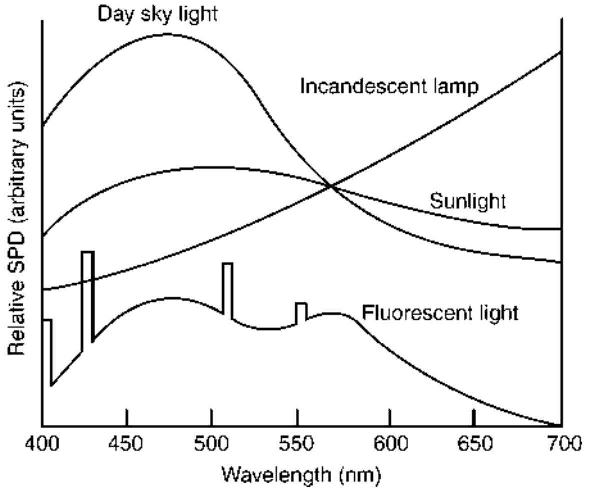

Figure 4 Relative SPDs of commonly used white light sources.

color temperature range of 5000–7500 K, incandescent lamps about 3000 K, direct sunlight ~4870 K and overcast skylight has a color temperature of ~7000 K.

The C.I.E. chromaticity diagram along with the color temperatures of the commonly employed sources, is shown in Figure 3.

Color in the Nature

As the solar radiation passes through the atmosphere, it is partly absorbed and partly scattered by the atmospheric constituents: air, water vapor, and dust particles. Absorption depends on the molecular properties of the air molecules, water vapor, and the dust particles, and is strongly wavelength dependent. Absorbed energy increases the internal energy of the molecule finally manifesting itself as thermal energy. When the sun is at the zenith, the solar radiation passes through one air mass loosing approximately 32% of its power. The SPD of the solar radiation above the atmosphere and at the Earth's surface at air mass 1 is shown in Figure 5. Scattering and absorption of the solar radiation by the atmosphere are responsible for a variety of interesting colorful views in the nature and they are briefly discussed in this section.

Atmospheric Colors

1. *Colors of the sky.* Scattering occurs over most of the visible part of the electromagnetic spectrum and is a function of the frequency of

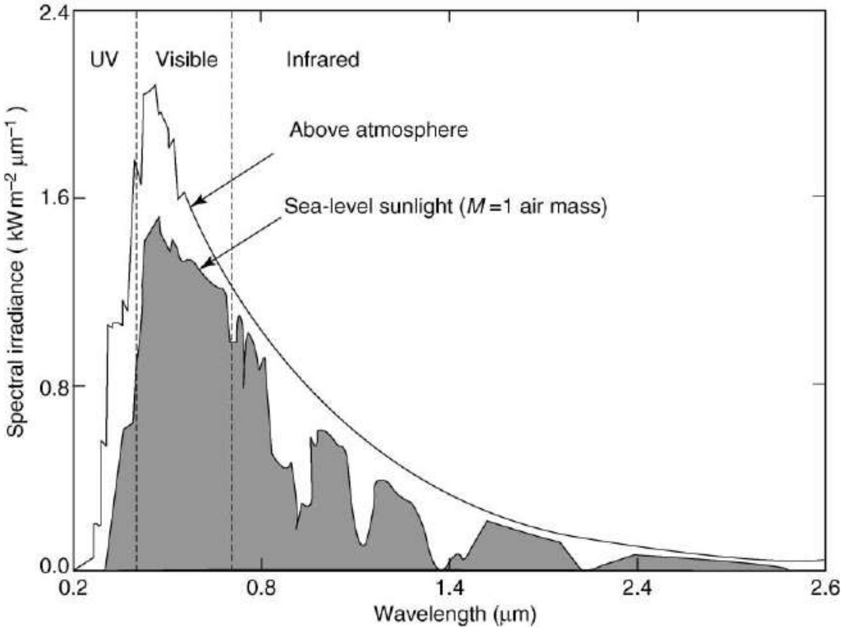

Figure 5 Spectral irradiance (kW m^{-2} μm^{-1}) of the solar radiation measured above the atmosphere and at the Earth's surface at sea level on a clear day and the sun at the zenith.

the radiation. According to Lord Raleigh's theory of scattering, the scattering probability varies inversely as a function of the fourth power of the wavelength of the radiation. If we consider the radiation corresponding to blue (400 nm) and the radiation corresponding to say red (700 nm), the photons corresponding to blue radiation are scattered by about 9.4 times that of photons corresponding to the red radiation. During noontime, when we look at the sky, not directly at the sun, the sky looks unsaturated blue because of the predominant scattering experienced by the high frequency radiation corresponding to blue color. However, it should be noted that the daytime skylight contains all the spectral colors of the solar radiation though the spectral power distribution has changed resulting in an increase in the intensity of the high frequency (blue) radiation. During noontime at the horizon, the sky looks close to white in color with a bluish tinge, because the solar radiation has passed through many air masses. As the thickness of the atmosphere increases, the intensity of the blue radiation decreases and ultimately as the radiation reaches the horizon, all the spectral colors are scattered about the same level resulting in white or bluish white sky.

2. *Polarization of the solar radiation by the atmosphere*. Light scattered by a molecule is polarized. The blue skylight is polarized with a maximum polarization of about 75% to 85% at 90° from the sun.
3. *Distant mountains seem bluish in color.* When an artist wishes to paint a distant mountain in perspective, he/she paints it blue; why? This is called airlight. The low wavelength part of the solar radiation is scattered more prominently, giving it a blue perspective when observed from a distance. However, it should be mentioned that the distant mountains may look any color depending on the scattering and absorption of the solar radiation sometimes resulting in 'Purple Mountain's majesty.'
4. *Sun looks red at the sunrise and sunset.* At sunset as the sun approaches the horizon, the sun changes its color from dazzling white to yellow, orange, bright red, and dull red and the order is reversed at sunrise. At the sunsets and sunrise, the solar radiation travels through many airmasses compared to approximately one airmass when the sun is at the zenith. Because of the large mass of air the radiation has to pass through, first the blue part of the spectrum is removed and the transmitted green to red spectral colors give the appearance of yellow. As this spectrum passes through more airmass, green is scattered most, leaving spectral yellow, orange, and red. As the radiation further passes through air, yellow and orange are also scattered, leaving only red.

5. *Twilight*. After sunset (dusk), or before the sunrise (dawn), the sky is illuminated by light which in turn illuminates the landscape. At sunsets and sunrise the solar radiation passes through a large mass of air. At a solar depression of about 0°, most of the blue is lost by scattering, resulting in a yellow arch and bright sky, and as the solar depression increases to about 10°, the arch will be reddish and the brightness of the sky reduced.
6. *Alpenglow*. Snow-covered mountains glow orange yellow from the low sun and after the sunset the same mountains look purple because of the reflection of the twilight.
7. *Twinkling of stars*. The stars twinkle. The local variations in the density of air are responsible for the twinkling of stars. As the density of air changes, so does the index of refraction and the starlight is continually refracted and dispersed. Twinkling is more prominent when the stars are on the horizon or on cold clear and windy nights.
8. *Mirages*. Mirages are refracted images produced because of the existence of refractive index gradients in the atmosphere. This results in the sight of shimmering water on a highway, inverted view of a car, etc.
9. *Aurora Borealis*. Aurorae are diffused slowly moving lights seen at high latitudes. They are often greenish yellow with different shapes. The lights in the northern hemisphere are called 'Aurora Borealis' and the lights in the southern hemisphere 'Aurora Australis.' They are caused primarily by the high-energy electrons deflected by the Earth's magnetic field towards the Earth's magnetic poles interacting with the atmospheric molecules resulting in molecular excitation and ionization. When the molecules de-excite, the characteristic electromagnetic radiation is commonly emitted. The commonly seen yellow-green light is due to the molecular emission of oxygen at 557 nm.
10. *Urban glows*. The scattered light from cities can be observed from space. The metropolitan Los Angeles can be observed from about 250 km in space. Extensively used sodium lamps with strong spectral yellow and blue-green mercury and fluorescent lights have unique color spectra of their own.

Colors due to Water and Water Droplets

1. *What is the color of pure water?* The transmission coefficient of pure water as a function of wavelength for different thicknesses of water column is shown in Figure 6. Our perception of the color of water changes with the thickness of water. For 1 mm thick water column, all the wavelengths are transmitted, resulting in the perception of white; 1 m thick water column looks unsaturated bluish green. As the thickness of water increases, the bluish green color will get more and more saturated. The color of water perceived by us depends on: (i) light reflected from the water surface; (ii) the light refracted at the water surface and diffused and scattered by the water molecules; and (iii) the refracted light reaches the bottom and gets reflected, which is further diffused and scattered and refracted at the water surface again. What we see is an additive sum of all these components.

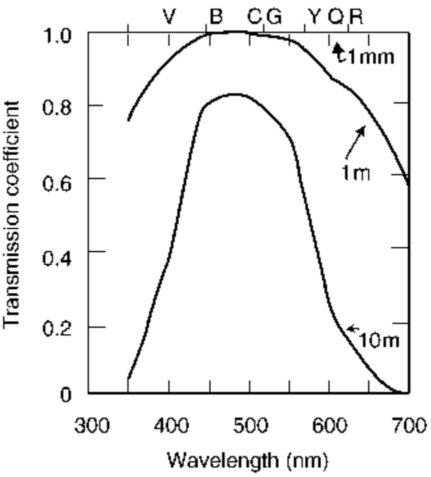

Figure 6 Transmission coefficient of water as a function of wavelength for different thicknesses of water (1 mm, 1 m, and 10 m).

2. *Refraction through water–air interface*. When light travels from an optically denser medium to an optically lighter medium, the light is deviated away from the normal. This is the reason why the apparent depth of a fish located in a pond is less than the true depth and a pencil struck in water seems to bend. If you are a diver when you look at the horizon from under water, the hemisphere seems to be compressed into a circle of diameter 97.2° instead of 180° which is known as 'Optical Manhole.'
3. *Glitter*. When you look at the reflection of the sun or the moon in wavy water, you notice a bright elongated reflected light known as glitter. It has a brilliant reddish orange color when the sun is above the horizon. Since the water is wavy, light reflected by a certain part of the wave at a certain time has the correct angle of reflection to reach your eye, which is known as glint. An ensemble of a large number of glints

coming from different waves and different parts of the waves results in glitter. Since you are looking at the reflected light, it will be polarized.

4. *Sky pools and landpools*. The light coming from a wavy water surface of a lake consists of oval shaped bright colored patterns that continually change in shape. This results from the reflection of skylight by the wavy water. If the waves are not turbulent the reflected light reaching your eye continually changes resulting in bright colored patterns. Landpools are caused by a nearby sloping landscape and the reflected light pattern from the landscape.
5. *Thin films*. Oil floats on water and spreads into thin films because of its low surface tension. When viewed in reflected natural white light, the light reflected from the top surface of the oil and the light reflected from the interfacial surface of oil and water interfere to produce colorful patterns. Brilliant colored patterns are commonly seen on the roads after rain because of gasoline on water. Foam ordinarily consists of bubbles ranging typically 10^{-4} mm to several mm. Even though each bubble produces colorful light due to the interference effects, the light coming out of the foam adds up to white light.
6. *Wet surfaces look darker*. A water spot on a concrete road looks darker. Since concrete is not porous a thin layer of water sticks to the surface. At the water surface part of the light incident is reflected, and a large part is transmitted. The transmitted light is diffusively scattered, part of which goes through total internal reflection and only a small part reaches the eye, making the surface look darker.

   When a fabric is wet it looks darker because of a slightly different mechanism. The surface of the fabric gets coated with a thin layer of water (refractive index = 1.33) which is smaller than the refractive index of the fabric. When the cloth is wet more light is transmitted, and less gets reflected than before. The transmitted light penetrates the fabric and gets scattered, resulting in a darker appearance.
7. *Rainbows*. Water with a refractive index of 1.33 is a dispersive medium. White light consisting of different wavelengths travel at different speeds in the medium and emerge at different directions. If the dispersion is adequate, different spectral colors can be seen by the naked eye. The dew covered lawn on a sunny day displays the spectral colors. If you stand close to the dew drop without blocking the sun's radiation falling on the drop, at some angle one should see the spectral colors. The primary rainbow is produced by the water droplets in the atmosphere after rain. The radius of the primary rainbow is about 42° with a width of about 2°. The color of the primary rainbow would be blue to red from the center of the rainbow. A rainbow is observed at the antisolar point. The brightness of the rainbow depends on the ambient sunlight, the size of the droplets, and whether the sun is low or high in the sky. The secondary rainbow, which is fainter, is often seen outside the primary rainbow at about 51°. The secondary rainbow will have red to blue from the center of the secondary rainbow. The secondary rainbow is approximately half as bright as the primary rainbow. The secondary rainbow is caused by two total internal reflections within the droplet whereas the primary rainbow is the result of one total internal reflection only. Because of the total internal reflection process, light is about 95% tangentially polarized in the primary rainbow.
8. *Heiligenschein*. Heiligenschein or the 'holy light' can be seen around the shadow of your head on a lawn on a sunny morning when dew drops are present. Usually, the dew drops are held by the tiny hairs of the grass at some distance from the blade of the grass. As light falls on the droplet, the spherical drop converges light at its focus and if the grass blade is located at that point, the light gets retroreflected approximately towards the sun resulting in holy light around your shadow.
9. *Coronae*. Colored concentric patchy rings seen around the moon and the sun are called coronae. They are primarily produced by the water droplets in the thin clouds. Even though the water droplets are randomly located, the diffracted light from the droplets has a defined angle and the superposition of the diffraction patterns from different water droplets results in colored ring structures. Coronae can be easily seen around the full moon viewed through thin clouds.
10. *The Glory*. On a foggy day with the sun at your back, one may see bright colored rings in the form of arches of about 5° to 10° at the solar point. The light in the glory is polarized and the ring structure is produced by reflection of the solar radiation by the fog.
11. *Clouds*. Clouds contain suspended water droplets or ice crystals. The size of the droplets ranges from less than a micrometer to about 100 micrometers (10^{-6} m). When the air is supersaturated with water vapor, the water condenses into water droplets, resulting in a definite shape to the cloud. When the sun's radiation falls on

the cloud, the light is scattered by the water droplets. Even though one would expect that the scattered light may have some color distribution, since we look at the scattered light from a very large number of droplets, the summation of all the colors results in white. Some clouds look dark or gray because, either they are in the shadow of other clouds or they are so thick that most of the light is absorbed. During sunset and sunrise, the clouds look bright orange because of the colored light illuminating them. Even though all the clouds are basically white, they may appear colored because of the selective absorption of certain colors in the atmosphere or due to the colored background radiation. Often, you see a silver lining at the edges of a cloud, because at the edges the density of water is low and when the sun illuminates it from the back, the radiation is scattered in the forward direction quite efficiently, resulting in a bright and shiny 'silver lining' to the cloud.

12. *Lightning*. Lightning is caused by an intense electrical discharge due to the high potential differences developed. On a cloudy and rainy day, as the water droplets grow in size, the falling drops get distorted to ellipsoidal shape and get polarized. The top smaller portion becoming negative and the lower larger part becoming positive. As the deformation increases, the drop ultimately breaks and the smaller drops being negatively charged and the larger ones positively charged. The wind currents easily lift the buoyant small drops to higher altitudes and the heavier positively charged droplets fall towards the Earth. The final result is that the upper region of the cloud is negatively charged whereas the lower region of the cloud is positively charged. The portion of the Earth under the cloud becomes negatively charged due to induction. The localized ions in the positively charged cloud produce a strong electric field which is further augmented by the induced negative charges on the Earth below. The strong electric field ionizes the air molecules setting up a current discharge to the earth. This discharge is called stepped leader. Following the stepped leader, the charges in the upper region of the cloud also discharge resulting in what is called return stroke. The sequence of leader and return strokes occur typically about ten times in a second. Since the clouds are charged and they are closer to each other, lightening is more frequent between the oppositely charged clouds than cloud and Earth. It is estimated that the lightning between the clouds is about five times more frequent compared to the lightening events between a cloud and the Earth. In this process, the nitrogen and oxygen molecules in air are excited and ionized. The flash is produced by the de-excitation of these molecular species. Since the temperatures involved are quite high (30 000 K), the flash has a strong unsaturated bluish hue. Some hydrogen lines are also observed in the flash. Hydrogen is produced as a result of the dissociation of water molecules during the discharge process.
13. *Ice and Halos*. Snow reflects most of the light falling with a reflectivity of about 95%. Snowflakes are crystalline and have hexagonal structure. In nature they occur in assorted sizes and shapes. As white light falls on them, light may be simply reflected from their surfaces making them shining white, which are often called glints. The light may also pass through the crystalline material get dispersed and emerge with different colors, which are known as sparkles.

    The halos appear as thin rings around the sun or the moon. They are produced because of the minimum deviation of light in snow crystals. The 22° halo is most common and is often seen a number of times in a year. The halos are circular and have a faint reddish outer edge and a bluish white diffused spot at the center. They often look white because of the superposition of different colors resulting in a more or less white color.

Color in Art and Holography

Sources of Color

There are two major color sources, dyes and paint pigments. Dyes are usually dissolved in a solvent. The dye molecules have selective absorption resulting in selective transmission. Paint pigments are powdered and suspended in oil or acrylic. The paint particles reflect part of the radiation falling on them, selectively absorb part of the radiation and transmit the balance. Lakes are made of translucent materials such as tiny grains of alumina soaked in a dye of appropriate color.

In the case of water colors and printer's inks, part of the light incident is reflected from the surface and the remaining transmitted which is selectively absorbed by the dye. The transmitted light reflected by the paper, passes through the dye again and emerges. Each transmission produces color by subtractive color scheme and the emerging light is mixed partitively by the observer.

In the case of paints, the color depends on the size and concentration of the pigment particles and the medium in which they are suspended. Part of the light incident on the artwork gets reflected at the surface and the remaining selectively absorbed and transmitted which in turn is further reflected and transmitted by other pigment particles each time following a subtractive color scheme. The light reaching the canvas gets reflected and passes through the medium and emerges. Again, all the emerging beams are partitively mixed by the observer.

Holograms may be broadly categorized as transmission and reflection holograms. Both transmission and reflection (volume) holograms are prepared by producing the interference pattern of the reference and object laser beams. Transmission holograms can be viewed in the presence of the reference laser beam that was employed for its preparation. Spectacular three-dimensional colored images are reconstructed by the reflection holograms depending on the angle of incidence of the white light and the direction of observation. Embossed holograms are popular for advertising, security applications (credit cards), display, and banknotes. Some studios prepare three-dimensional holographic portraits that can be viewed in natural light. In the future, three-dimensional videos and even holographic three-dimensional movies may be a reality.

Structural Colors in Nature

Nanoscale Photonic Lattices

Newton observed that iridescence (change in color with the direction of observation) is caused by optical interference. The colors are produced by the interference or the light diffracting characteristics of the microstructures and not by the selective reflection or absorption of light, as in the case of pigments. Photonic band structures consist of periodic arrangement of dielectric materials even though they are transparent, have a bandgap resulting in inhibition of certain wavelengths. Light with frequencies in the gaps between the bands get reflected resulting in the specific color.

The structural colors of light can be seen in the animal world and minerals. For example, opal, a sedimentary gem does not have crystalline structure, but consists of tiny spheres of silica packed together. The diffraction of light through these structures results in spectacular colors determined by the size of the spheres and their periodicity. The spines of the sea mouse Aphrodite are covered with natural photonic structures that produce iridescence resulting in brilliant colors depending on the direction of the incident light and the direction of observation.

In the case of the butterfly, *Morpho rhetenor*, the metallic blue color is produced by the periodic structure of its wings. Its wings have layers of scales, each of about 200 mm long and 80 mm wide and about 1300 thin parallel structures per mm which form a diffraction grating.

The bird feathers viewed through an optical microscope show the structural components which are responsible for their coloration. The structural details of the peacock's eye pattern employing the scanning electron microscopy show photonic structures on peacock barbules. Barbules are the strap like branches on the peacock feathers. The four principal colors on the peacock feathers blue, green, yellow, and brown were identified wth the specific structures. The interference between the light reflected from the front surface to that reflected from the back was shown to be responsible for the brown color in the peacock's eye. The photonic structure of the green barbule has a periodic structure of melanin cylinders with a spacing of about 150 nm whereas the blue barbule has a periodic structure of 140 nm, and yellow barbules have a periodicity of 165 nm.

See also

Dispersion Management. **Holography, Techniques:** Color Holography. **Photon Picture of Light**. **Polarization:** Introduction. **Scattering:** Scattering Theory.

Further Reading

Blau SK (2004) Light as a feather: structural elements give peacock plumes their color. *Physics Today* 57: 18–21.

Color Science, http://www.physics.sfasu.edu/astro/color.html.

Colourware, Colour FAQ, http://www.colourware.co.uk/cpfaq.htm.

CVRL Color & Vision database, http://www.cvrl.org/.

Hecht E (2002) *Optics*, 4th edn. San Francisco, CA: Addison Wesley.

Lynch DK and Livingston W (2001) *Color and Light in Nature*, 2nd edn. Cambrige, UK: Cambridge University Press.

Minnaert MGJ (1993) *Light and Color in the Outdoors*. Heidelbergh: Springer-Verlag.

Nassau K (2000) *The Physics and Chemistry of Color*. New York: John Wiley and Sons.

Poynton CA, Color technology, http://www.Poynton.com/ColorFAQ.html.

Tayeb G, Gralak B and Enoch S (2003) Structural colors in nature and butterfly-wing modeling. *Optics and Photonics News* 14(2): 38–43.

Wyszecki G and Stiles WS (2000) *Color Science*. New York: John Wiley.

D

DETECTION

Contents

Fiber Sensors
Heterodyning
Image Post-Processing and Electronic Distribution
Smart Pixel Arrays

Fiber Sensors

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The growth in the fiber optic communications and optoelectronics industries has led to the development and commercialization of many types of optical components. Fiber sensor devices and systems are a major user of such technologies. This has resulted in the development of commercial fiber sensors that can compete with conventional sensors. Fiber sensors are capable of measuring a wide variety of stimuli including: mechanical (displacement, velocity, acceleration, strain, pressure); temperature; electromagnetic (electric field, magnetic field, current); radiation (X-ray, nuclear); chemical composition; flow and turbulence in liquids; and biomedical. The main advantages of fiber sensors are that they are low cost, compact and lightweight, robust, passive, immune to electromagnetic interference, and highly sensitive.

Fiber sensors can be grouped into two basic classes: intrinsic and extrinsic. In an intrinsic fiber sensor the sensing is carried out by the optical fiber itself. In an extrinsic sensor the fiber is simply used to carry light to and from an external optical device where the sensing takes place. Point fiber sensors are localized to discrete regions; quasi-distributed sensors utilize point sensors at various locations along a fiber and distributed sensors are capable of sensing over the entire length of fiber.

A basic fiber sensor system, as shown in Figure 1, consists of a light source, fiber sensor, optical detector, and signal processing electronics. A measurand causes some change (intensity, phase, polarization, spectrum, etc.) in the light propagating through the sensor. This change is detected and processed to give an output signal proportional to the measurand. A large part of fiber sensor research concerns the development of fiber sensors sensitive to particular measurands and appropriate signal processing techniques. We describe some of the most important fiber sensors and also consider quasi-distributed and distributed sensing, including the use of multiplexing techniques.

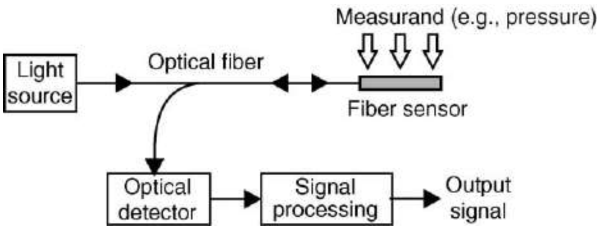

Figure 1 Basic fiber sensor system.

Intensity-Based Fiber Sensors

The intensity modulation (IM) of light is a simple method for optical sensing. There are several mechanisms that can produce a measurand-induced change in the optical intensity propagated by an optical fiber. Perhaps the simplest type of IM fiber sensor is the microbend sensor shown in Figure 2. The sensor consists of two grooved plates between which passes an optical fiber. The upper plate can move in response to pressure. When the fiber is bent sufficiently, light escapes into the fiber cladding and is lost. The greater the pressure on the plates the more loss occurs.

Coupling-based fiber sensors are useful for measurement of displacement or dynamic pressure. Transmission and reflective configurations are possible as shown in Figure 3. The transmission coupling-based sensor consists of two fibers with a small gap between them. The amount of light coupled to the second fiber depends on the fiber acceptance angles and the distance between the fibers. One of the fibers can move in response to vibration or pressure thereby changing the distance between the fibers and hence the coupling loss. The reflection-based sensor operates in a similar fashion, where light is reflected from a flexible diaphragm back into a collecting fiber. The reflected light intensity changes as the diaphragm is flexed. Once the coupling relationship between the input fiber, diaphragm and collecting fiber is known, intensity changes can be related to the applied displacement or pressure.

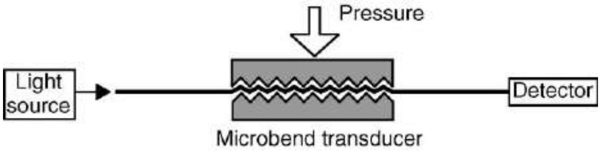

Figure 2 Microbend fiber sensor. The transducer moves in response to pressure and in doing so changes the bending radius of the fiber and thereby the fiber loss.

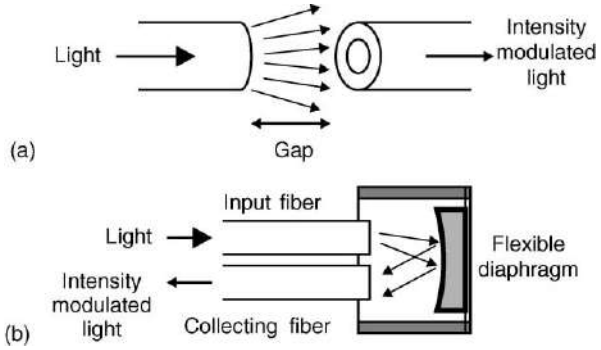

Figure 3 (a) Transmission and (b) reflective coupling-based fiber sensors.

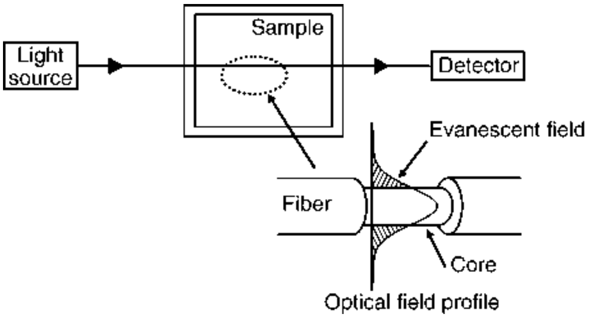

Figure 4 Evanescent wave fiber chemical sensor.

Evanescent wave fiber sensors exploit the fact that some of the energy in the guided mode of an optical fiber penetrates a short distance from the core into the cladding. The penetration of light energy into the cladding is called the evanescent wave. It is possible to design a sensor where energy is absorbed from the evanescent wave in the presence of certain chemicals as shown in Figure 4. This is achieved by stripping the cladding from a section of the fiber and using a light source having a wavelength that can be absorbed by the chemical that is to be detected. The resulting change in light intensity is a measure of the chemical concentration. Measurements can also be performed in a similar fashion by replacing the cladding with a material such as an organic dye whose optical properties can be changed by the chemical under investigation. Evanescent wave fiber sensors have found many applications in the biomedical field, such as blood component meters for detection of cholesterol and uric acid concentrations in blood.

A linear position sensor based on time division multiplexing is shown in Figure 5. It uses a square-wave modulated light source, optical delay loops and an encoded card. The delay loops separate the return signal from the encoded card by a time that is greater than the pulse duration. The encoded return signal can be decoded to determine the card position and velocity.

Interferometric Sensors

Interferometric fiber sensors operate on the principle of interference between two or more light beams to convert phase differences to intensity changes. Common configurations used include Michelson, Mach–Zehnder, and Sagnac interferometers, polarimetric systems, grating and etalon-based interferometers and ring resonators. Interferometric fiber sensors have extremely high sensitivity and are able to resolve path differences of the order of 10^{-6} of the light

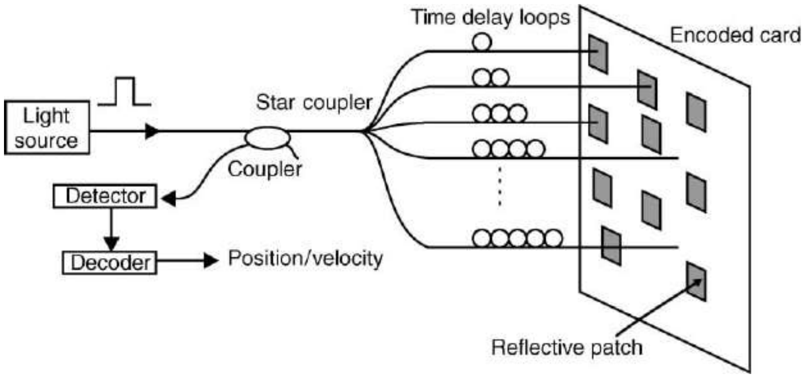
Time delay loops
Encoded card
Light source
Star coupler
Coupler
Detector
Decoder
Position/velocity
Reflective patch

Figure 5 Linear position sensor using time division multiplexing.

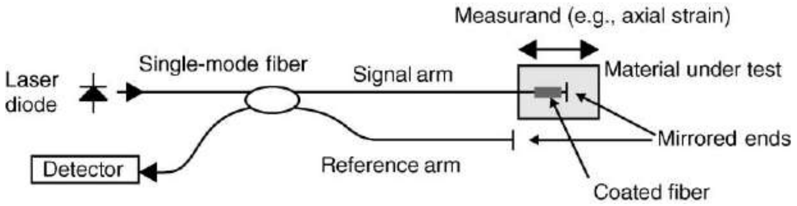
Measurand (e.g., axial strain)
Laser diode
Single-mode fiber
Signal arm
Material under test
Mirrored ends
Detector
Reference arm
Coated fiber

Figure 6 Michelson interferometric strain sensor. Part of the signal arm is embedded in a material, such as concrete (in civil engineering structures) or carbon composite (used in the aerospace industry).

source wavelength. To realize high sensitivity, spectrally pure light sources such as semiconductor lasers must be used. In addition, single mode fibers and components are used to maintain the spatial coherence of the light beams. Factors that affect the performance of interferometric sensors include the optical source phase noise and the polarization states of the interfering light beams.

In two-beam interferometric sensors, such as the Michelson interferometric strain sensor shown in Figure 6, one of the light beams travels through a fiber where its phase can be modulated by the measurand. The two reflected beams recombine on a detector, the output of which is of the form $1 + V \cos \phi$ as shown in Figure 7. ϕ is the relative phase shift between the return beams and V the interferometer visibility. Compensating techniques must be used to ensure that the interferometer sensitivity is maximized. Active techniques involve the insertion of an active device, such as a fiber stretcher, in one of the interferometer arms to control the mean phase difference between the beams. Passive schemes have the advantage in that no active components are required, but usually involve complex signal processing schemes such as active homodyne or synthetic heterodyne demodulation. The sensitivity of the fiber to the measurand can be further improved through the use of specialized coatings.

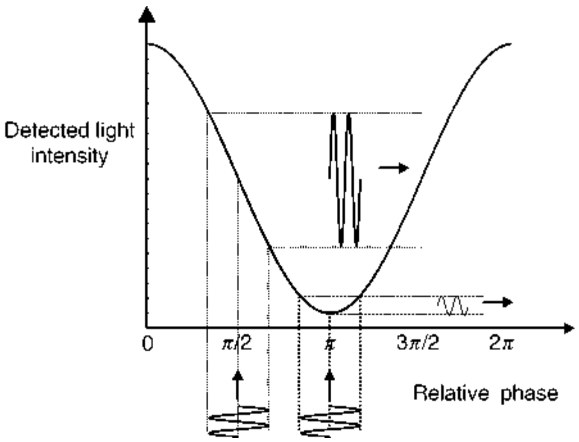

Figure 7 Two-beam interferometer transfer function. The sensitivity is maximized when the mean phase difference between the detected light beams is an odd integer multiple of $\pi/2$. When this is the case the interferometer is said to be operating in quadrature.

Compact interferometric sensors can be constructed using intrinsic or extrinsic Fabry–Perot type configurations, the most common of which are shown in Figure 8. If the Fabry–Perot cavity reflectivities are $\ll 1$, then it can be considered to be a two-beam interferometer.

Polarimetric fiber sensors relate changes induced in the polarization state of light to the measurand.

These sensors usually use high birefringence (Hi-Bi) fiber. The phase difference between the two orthogonally polarized modes of Hi-Bi fiber is given by

$$\phi = \Delta\beta l \qquad [1]$$

where $\Delta\beta$ is the fiber (linear) birefringence and l the fiber length. Both $\Delta\beta$ and l can be changed by the measurand. A polarimetric strain sensor using Hi-Bi fiber embedded in carbon fiber composite, is shown in Figure 9. Light from a linearly polarized He–Ne laser is launched into the Hi-Bi fiber, through a half-waveplate and lens. The waveplate is used to rotate the light plane of polarization so that it is at 45° to the principal axes of the fiber. This ensures that half of the input light power is coupled to each of the fast and slow modes of the Hi-Bi fiber. The output light from the Hi-Bi fiber is passed through a polarization beamsplitter. The beamsplitter separates the light into two orthogonally polarized beams, which are then detected. The output from each of the detectors is given by $V_1 = I_0 \sin^2\phi$ and $V_2 = I_0 \cos^2\phi$, respectively, where I_0 is the input light intensity. The state of polarization (SOP) of the output light from the Hi-Bi fiber is given by

$$\mathrm{SOP} = \frac{V_1 - V_2}{V_1 + V_2} = \cos^2\phi + \sin^2\phi \qquad [2]$$

It is a simple matter to calculate the SOP, from which ϕ can be determined. The form of eqn [2] is very similar to the two-beam interferometer transfer function, shown in Figure 7.

A polarimetric electric current sensor based on the Faraday effect is shown in Figure 10. The Faraday effect provides a rotation of the light's polarization state when a magnetic field is parallel to the optical path in glass. If an optical fiber is closed around a current-carrying conductor, the Faraday rotation is directly proportional to the current. By detecting the polarization rotation of light in the fiber, the current can be measured.

Linearly polarized light is equivalent to a combination of equal intensity right and left circularly polarized components. The linear state of polarization of the polarized laser light rotates in the presence of a magnetic field because the field produces

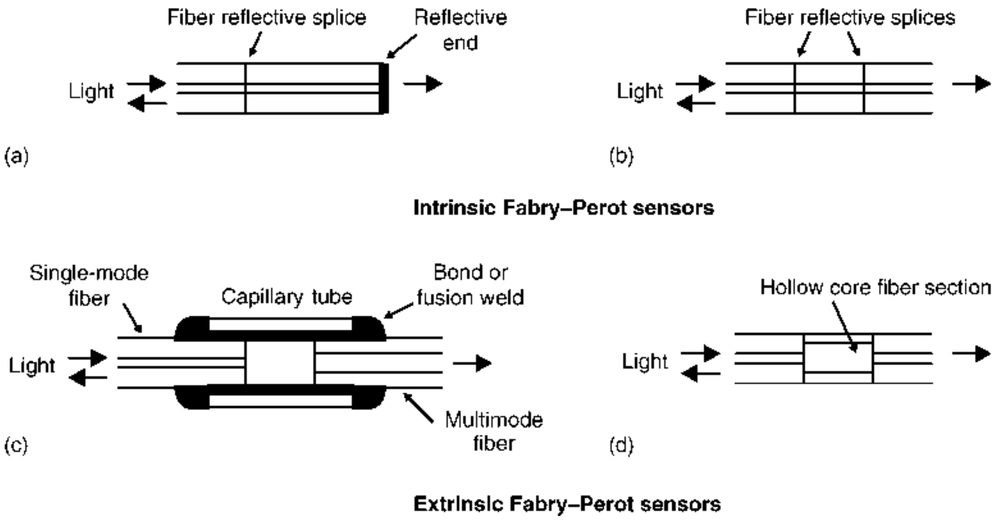
Fiber reflective splice
Reflective end
Light
(a)
Fiber reflective splices
Light
(b)
Intrinsic Fabry–Perot sensors
Single-mode fiber
Capillary tube
Bond or fusion weld
Light
Multimode fiber
(c)
Hollow core fiber section
Light
(d)
Extrinsic Fabry–Perot sensors

Figure 8 Interferometric fiber Fabry–Perot configurations.

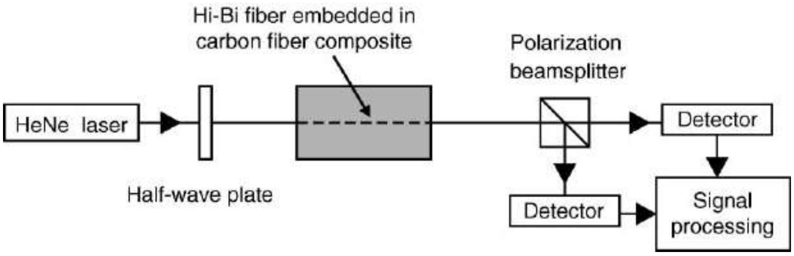
Hi-Bi fiber embedded in carbon fiber composite
Polarization beamsplitter
HeNe laser
Detector
Half-wave plate
Detector
Signal processing

Figure 9 Polarimetric strain sensor system.

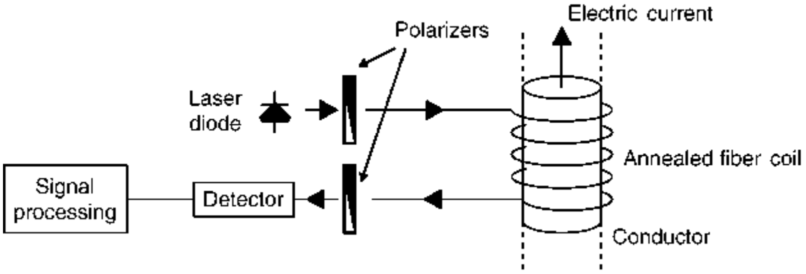

Figure 10 Polarimetric electric current sensor system.

circular birefringence in the fiber. This means that right and left circularly polarized light will travel at different speeds and accumulate a relative phase difference given by

$$\phi = VBl \tag{3}$$

where V is the Verdet constant (rad Tesla^{-1} m^{-1}) of the glass, B the magnetic field flux density (Tesla), and l the length of the fiber exposed to the magnetic field. In the polarimetric current sensor, the polarization rotation is converted to an intensity change by the output polarizer. The output to the detector is proportional to $1 + \sin(2\phi)$, from which ϕ and the current can be determined.

The linear birefringence of conventional silica fiber is much greater than its circular birefringence. To make a practical current sensor, the linear birefringence must be removed from the fiber and this can be achieved by annealing the fiber. Annealing involves raising the temperature of the fiber, during manufacture, to a temperature above the strain point for a short period of time and slowly cooling back to room temperature. This reduces stresses in the glass, which are the principal cause of linear birefringence, and also cause physical and chemical changes to the glass. Waveguide-induced birefringence cannot be removed by annealing, but can be significantly reduced by twisting the fiber.

White light – or more accurately low-coherence – interferometry utilizes broadband sources such as LEDs and multimode lasers in interferometric measurements. Optical path differences (OPDs) are observed through changes in the interferometric fringe pattern. A processing interferometer is required in addition to the sensing interferometer to extract the fringe information.

The processing of white light interferometry signals relies on two principal techniques. The first technique involves scanning the OPD of the processing interferometer to determine regions of optical path balance. The second technique involves determination of the optical spectrum using an optical spectrum analyzer. The resulting fringe pattern fringe spacing is then related to the OPD of the sensing interferometer.

An example of the optical path scanning technique is shown in Figure 11a. The sensing interferometer is designed such that its OPD is much greater than the coherence length of the light source, so at its output no interference fringes are observed. The output of the sensing interferometer is fed to the processing interferometer. The OPD of the processing interferometer is scanned using a piezoelectric fiber stretcher driven by a suitable scanning voltage. As the processing interferometer is scanned, interference fringes are observed at two distinct points as shown in Figure 11b. The first set of fringes occurs when the OPD of the processing interferometer is within the coherence length of the optical source (close to zero). The second set of fringes occurs when the OPDs in the two interferometers are equal. The OPD in the sensing interferometer can be determined by measuring the OPD in the processing interferometer between the two positions of maximum visibility in the output signal from the detector. This, in turn, can be related to OPD changes due to the action of the measurand, in this case strain.

Sagnac interferometric sensors can be used to create highly sensitive gyroscopes that can be used to sense angular velocity (e.g., in aircraft navigation systems). It is based on the principle that the application of force (e.g., centrifugal force) will alter the wavelength of light as it travels around a coil of optical fiber. A basic open-loop fiber gyroscope is shown in Figure 12. The broadband source (e.g., superluminescent diode) is split into two counter propagating light beams traveling in the clockwise and anticlockwise directions. The polarizer is used to ensure the reciprocity of the counter propagating waves. The inherent nonlinear response of the gyroscope can be overcome by using a phase modulator and signal processing techniques. In the ideal case the detector output is proportional to $1 + \cos\phi_s$. The Sagnac phase shift ϕ_s between the two returning beams is given by

$$\phi_s = \frac{8\pi A N \Omega}{c\lambda_0} \qquad [4]$$

A is the area enclosed by a single loop of the fiber, N the number of turns, Ω the component of the angular velocity perpendicular to the plane of the loop, λ_0 the free-space wavelength of the optical source, and c the speed of light in a vacuum. Sensitivities greater than 10^{-8} rad/s and dynamic ranges in excess of 40 dB have been achieved with open-loop fiber gyroscopes. More advanced gyroscopes can greatly improve on this performance.

Fiber Grating Sensors

Many types of fiber gratings can be used in sensing applications including Bragg, long-period, and

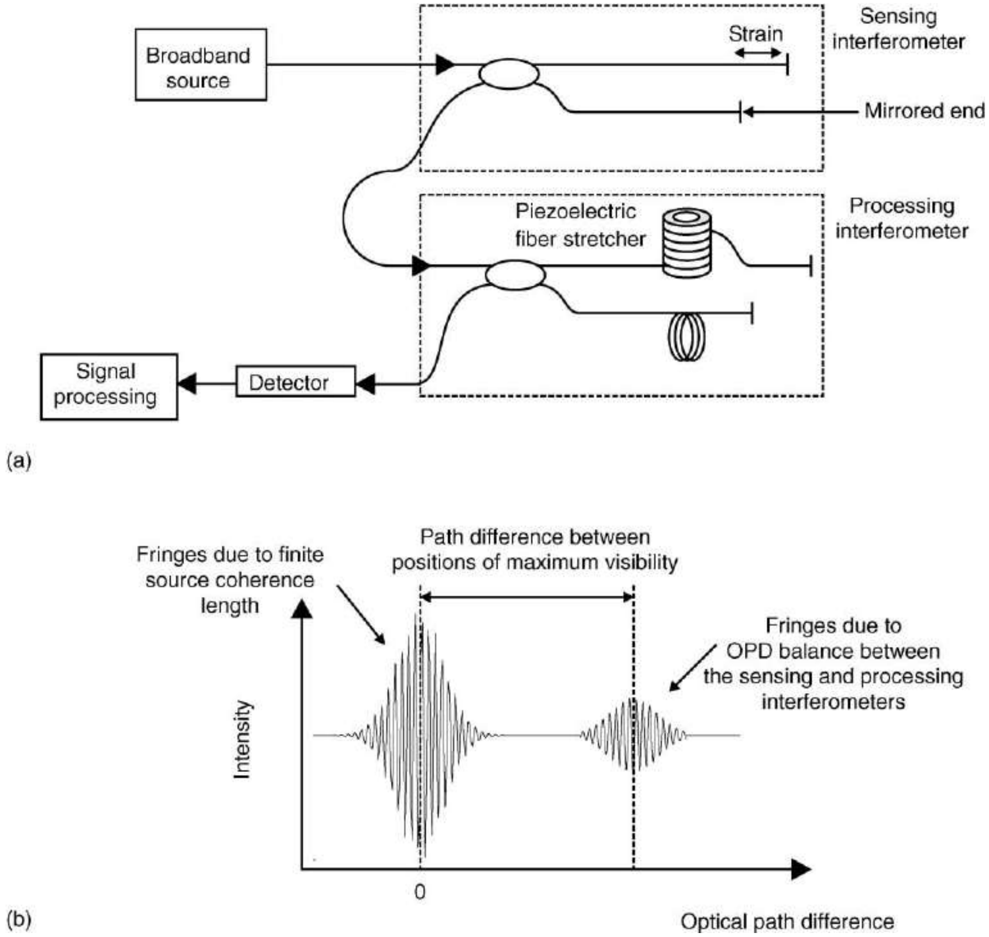
Broadband source
Strain
Sensing interferometer
Mirrored end
Piezoelectric fiber stretcher
Processing interferometer
Signal processing
Detector
(a)
Fringes due to finite source coherence length
Path difference between positions of maximum visibility
Fringes due to OPD balance between the sensing and processing interferometers
Intensity
0
Optical path difference
(b)

Figure 11 Low-coherence interferometric sensor. (a) Schematic diagram. (b) Output of the scanned processing interferometer.

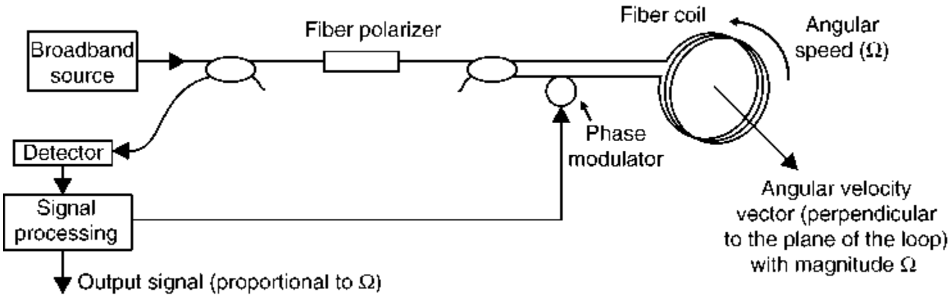
Broadband source
Fiber polarizer
Fiber coil
Angular speed (Ω)
Detector
Phase modulator
Signal processing
Angular velocity vector (perpendicular to the plane of the loop) with magnitude Ω
Output signal (proportional to Ω)

Figure 12 Basic open-loop fiber gyroscope.

chirped gratings. Because fiber gratings are small and have a narrow wavelength response, they can be used for both point and quasi-distributed sensing. They can be embedded in composite materials for smart structure monitoring and also in civil engineering structures such as bridges.

Fiber Bragg gratings (FBGs) are the most popular type of grating sensor. A typical FBG consists of a short section (typically a few mm) of single-mode fiber with a periodic modulation (typically 100s of nm) of the core refractive index. The index modulation causes light in the forward propagating core mode to be coupled into the backward core mode. This causes the FBG to act as a highly wavelength-selective rejection filter. The wavelength of peak reflectivity is the Bragg wavelength $\lambda_B = 2n_{\text{eff}}\Lambda$, where n_{eff} is the effective refractive index of the guided mode in the fiber and Λ the index modulation period. Both n_{eff} and Λ can be changed by an external measurand, resulting in a shift in λ_B.

The basic principle of FBG sensors is the measurement of an induced shift in the wavelength of an optical source due to a measurand, such as strain or temperature. A basic reflective FBG sensor system is shown in Figure 13. A broadband light source is used to interrogate the grating, from which a narrowband slice is reflected. The peak wavelength of the reflected spectrum can be compared to λ_B, from which strain or temperature can be inferred. The shift in the Bragg wavelength $\Delta\lambda_B$ with applied microstrain $\Delta(\mu\varepsilon)$ and change in temperature ΔT, for silica fiber is given approximately by

$$\frac{\Delta\lambda_B}{\lambda_B} \approx 0.78 \times 10^{-6}\Delta(\mu\varepsilon) + 6.67 \times 10^{-6}\Delta T \qquad [5]$$

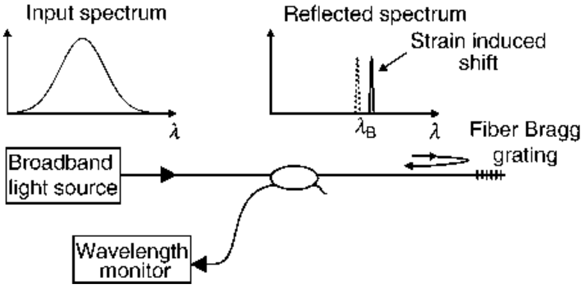

Figure 13 Basic reflective FBG sensor system.

A wavelength resolution of ~1 pm is required (at 1.3 μm) to resolve a temperature change of 0.1 °C or a strain change of 1 $\mu\varepsilon$. The response of the grating to strain can be improved through the use of specialized coatings. Polyimide coatings are commonly used as an efficient strain transducer for gratings embedded in composite materials. Because thermal effects in materials are usually very slow, it is relatively easy to measure dynamic strain (>1 Hz). However, in structural monitoring it can be necessary to distinguish between wavelength shifts due to static strain and those due to temperature. One technique is to use two collocated gratings whose response to strain and temperature is significantly different. In addition to strain and temperature measurement, grating sensors have also been used to measure flow, vibration, electromagnetic fields and chemical effects.

An example of quasi-distributed strain sensing, using a wavelength division multiplexed array of FBGs, is shown in Figure 14. Each FBG in the array has a unique Bragg wavelength. The return light from the FBG array is passed through a tunable narrowband Fabry–Perot filter. As the filter is tuned, the wavelengths returned by the individual FBGs can be analyzed and the strain present at each grating determined.

FBGs can be used as narrowband reflectors for creating fiber laser sensors, capable of measuring temperature, static strain, and very high-resolution dynamic strain. The basic form of an FBG laser sensor system shown in Figure 15 consists of a doped fiber section between two identical FBGs. The doped fiber is optically pumped to provide gain and thereby enable lasing to occur. Single-mode or multi-mode lasing is possible depending on the cavity length. In single-mode operation the FBG laser linewidth can be much smaller than the linewidth of diode lasers. This means that FBG laser sensors have greater sensitivities compared to passive FBG sensors.

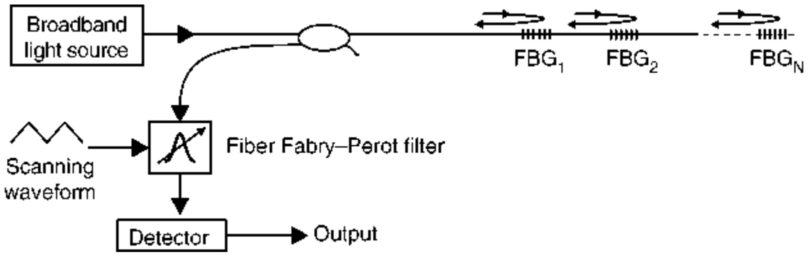

Figure 14 Quasi-distributed strain sensing using a wavelength division multiplexed array of FBGs.

When the cavity is subject to weak dynamic strain, the laser output is frequency modulated. This frequency modulation can be detected by using interferometric techniques. The main advantage of the FBG laser sensor over direct interferometry is that it is possible to obtain comparable strain sensitivities using a much shorter length of fiber.

Long-period fiber gratings (LPFGs) are attracting much interest for use in sensing applications. They are more sensitive to measurands than FBGs and easier to manufacture. A typical LPFG has a length of tens of mm with a grating period of 100s of μm. Its operation is different to an FBG in that coupling occurs between the forward propagating core mode and co-propagating cladding modes. The high attenuation of the cladding modes results in a series of minima occurring in the transmission spectrum of the fiber. This means that the spectral response is strongly influenced by the optical properties of the cladding and surrounding medium. This can be exploited for chemical sensing as shown in Figure 16, where a broadband source is used to interrogate an LPFG. The wavelength shifts of the output spectrum minima can be used to determine the concentration of particular chemicals in the substance surrounding the grating. The longest wavelength attenuation bands are the most sensitive to the refractive index of the substance surrounding the grating. This is because higher order cladding modes extend a greater distance into the external medium. LPFGs can also be used as strain, temperature, refractive index, bend, and load sensors.

Fiber Laser Doppler Velocimeter

Laser Doppler velocimetry (LDV) is a technique used for measuring velocity, especially of fluid flows. A fiber LDV system and associated scattering geometry is shown in Figure 17. In LDV two coherent laser beams intersect in a small measurement volume where they can interfere. The light reflected by a seed particle passing through the measurement volume is modulated at a frequency proportional to the spatial frequency (Doppler difference frequency Δf) of the interference fringes and the component of its velocity

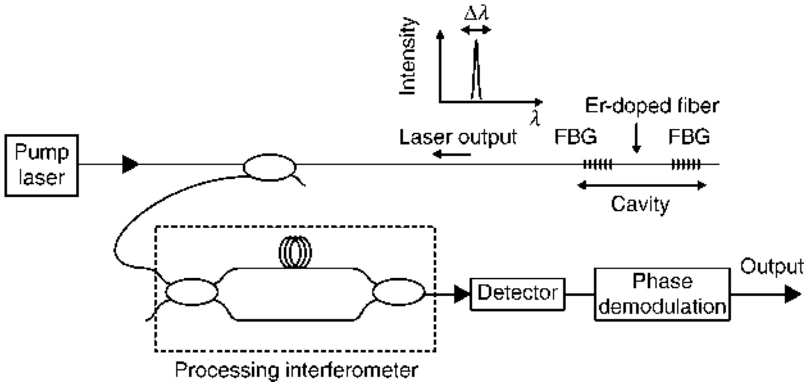
Intensity
$\Delta\lambda$
λ
Er-doped fiber
Laser output
FBG
FBG
Cavity
Pump laser
Detector
Phase demodulation
Output
Processing interferometer

Figure 15 FBG laser sensor with interferometric detection.

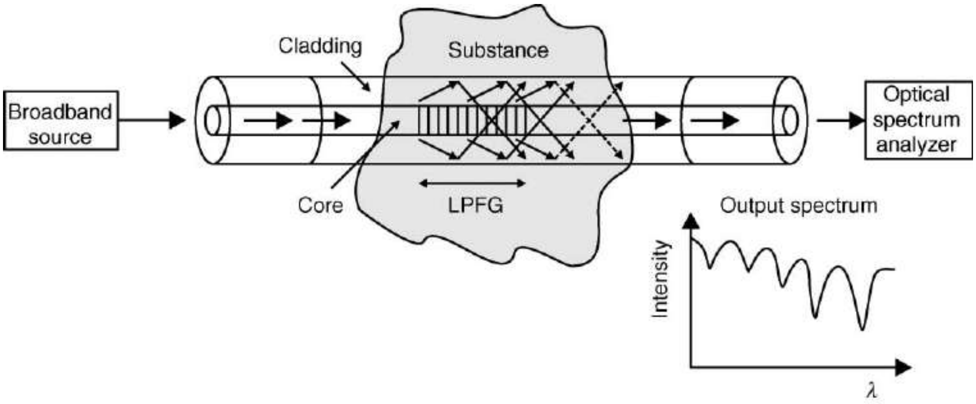
Cladding
Substance
Broadband source
Optical spectrum analyzer
Core
LPFG
Output spectrum
Intensity
λ

Figure 16 Detail of LPFG sensor.

normal to the interference fringes. Δf is given by

$$\Delta f = \frac{2nV \cos \beta}{\lambda_0} \sin(\theta/2) \qquad [6]$$

where n is the fluid refractive index, V the particle velocity and λ_0 the laser free-space wavelength. The output of the detector is processed to extract Δf and therefore $V \cos \beta$. Δf is independent of the scattered light direction so collection of the scattered light using a lens increases the system sensitivity. The form of eqn [6] indicates that the direction of flow cannot be ascertained. The simplest technique to resolve this ambiguity is to apply a frequency shift Δf_s to one of the input beams. This can be achieved by the use of a piezoelectric frequency shifter. The frequency shift causes a phase shift to appear between the two beams. The phase shift increases linearly with time. This results in a fringe pattern of spacing s, which moves with constant velocity $V_f = s\Delta f_s$. In this case the measured Δf will be less than or greater than Δf_s, depending on whether the particle is moving with or against the fringe motion. There will then be an unambiguous detectable range of velocities from zero to V_f.

Luminescence-Based Fiber Sensors

Luminescence-based fiber sensors are usually based on fluorescence or amplified spontaneous emission occurring in rare earth materials. They can be used in many applications such as chemical, humidity, and temperature sensing. It is possible to connect a fiber to luminescent material or to introduce luminescent dopants into the fiber. An example of the latter, used to detect chemical concentration is shown in Figure 18. A laser pulse causes the doped section of the fiber to luminesce at a longer wavelength than the

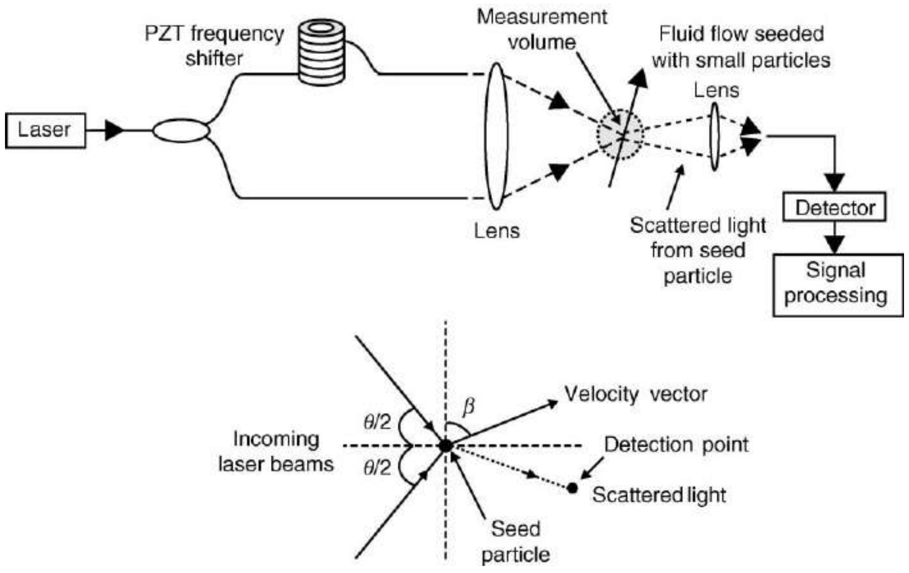
PZT frequency shifter
Measurement volume
Fluid flow seeded with small particles
Lens
Laser
Detector
Lens
Scattered light from seed particle
Signal processing
θ/2
β
Velocity vector
Incoming laser beams
Detection point
θ/2
Scattered light
Seed particle

Figure 17 Fiber Doppler velocimeter and scattering geometry.

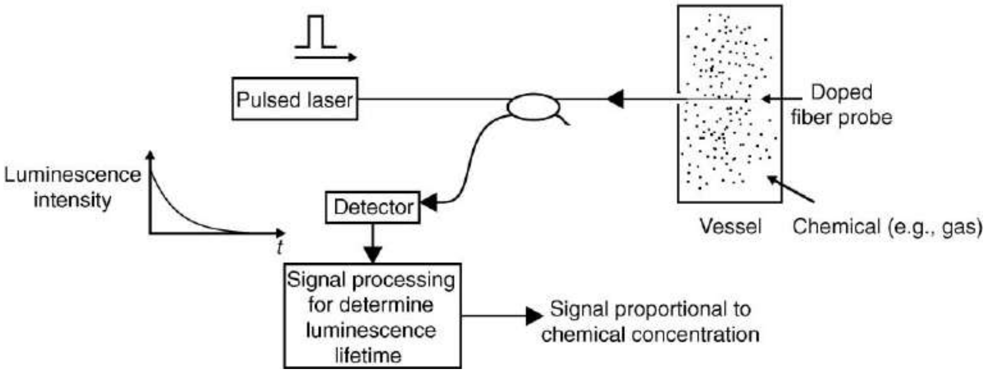
Pulsed laser
Doped fiber probe
Luminescence intensity
Detector
t
Vessel
Chemical (e.g., gas)
Signal processing for determine luminescence lifetime
Signal proportional to chemical concentration

Figure 18 Chemical sensor based on fluorescence in doped optical fiber.

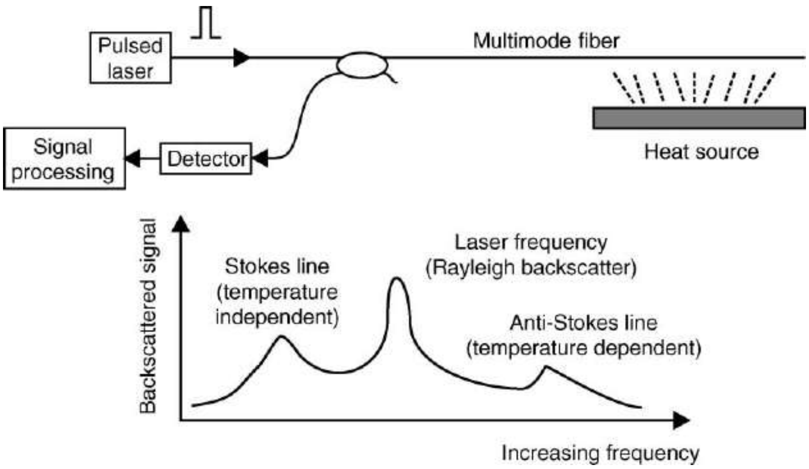
Pulsed laser
Multimode fiber
Heat source
Signal processing
Detector
Backscattered signal
Stokes line (temperature independent)
Laser frequency (Rayleigh backscatter)
Anti-Stokes line (temperature dependent)
Increasing frequency

Figure 19 Anti-Stokes Raman thermometry system.

laser. The luminescence intensity *I(t)* at the detector has an exponential decay profile given by

$$I(t) = I_0 \exp\bigl(-(k_1 + k_2 C)t\bigr) \qquad [7]$$

where I_0 is the initial luminescence intensity, t time, k_1, k_2 constants, and C the chemical concentration. The luminescence time constant $1/(k_1 + k_2 C)$ can be determined by comparing the luminescence intensity at various times after excitation by the laser pulse, from which C can be determined. The use of time division multiplexing allows quasi-distributed measurement of chemical concentration. The use of plastic optical fiber for luminescence-based sensors is attracting much interest.

Distributed Fiber Sensing

We have seen that both point and quasi-distributed sensing are possible using fiber sensors. Distributed sensing can be achieved through the use of linear or nonlinear backscattering or forward scattering techniques. In linear backscattering systems, light backscattered from a pulse propagating in an optical fiber is time resolved and analyzed to obtain the spatial distribution of the measurand field, e.g., polarization optical time domain reflectometry analyzes the polarization state of backscattered light to determine the spatial distribution of electromagnetic fields. Nonlinear backscattering schemes use effects such as Raman or Brillouin scattering. An important example of the former is the anti-Stokes Raman thermometry system shown in Figure 19. A light pulse is transmitted down the sensing fiber. Spontaneous Raman scattering causes Stokes and anti-Stokes photons to be generated along the fiber. Some of these photons travel back along the fiber to a fast detector. The intensity of the Stokes line is temperature independent. The anti-Stokes line intensity is a function of temperature. The ratio of the two intensities provides a very accurate measurement of temperature. The measurement location is determined by timing of the laser pulse.

See also

Environmental Measurements: Laser Detection of Atmospheric Gases. **Fiber Gratings**. **Interferometry:** Overview; Phase Measurement Interferometry; White Light Interferometry. **Optical Materials:** Smart Optical Materials.

Further Reading

Culshaw B and Dakin JP (1988, 1989, 1997) *Optical Fiber Sensors*, vols. I, II, III. Norwood, MA: Artech House.

Grattan KTV and Meggitt BT (1994, 1997, 1999, 1999) *Optical Fiber Sensor Technology*, vols. I, II, III and IV. Boston, CA: Kluwer Academic.

Grattan KTV and Meggitt BT (2000) *Optical Fiber Sensor Technology*. Boston, MA: Kluwer Academic.

Kersey AD, Davis MA, Patrick HJ, *et al.* (1997) Fiber grating sensors. *Journal of Lightwave Technology* 15: 1442–1463.

Krohn DA (1999) *Fiber Optic Sensors: Fundamentals and Applications*, 3rd edn. Research Triangle Park, NC: Instrumentation Society of America.

Lopez-Higuera JM (2001) *Handbook of Optical Fiber Sensing Technology*. New York: Wiley.

Othonos A and Kalli K (1999) *Fiber Bragg Gratings: Fundamentals and Applications in Telecommunications and Sensing*. Norwood, MA: Artech House.

Udd E (1991) *Fiber Optic Sensors: An Introduction for Engineers and Scientists*. New York: Wiley.

Udd E (1995) *Fiber Optic Smart Structures*. New York: Wiley.

Yu FTS and Yin S (2002) *Fiber Optic Sensors*. New York: Marcel Dekker.

Heterodyning

T-C Poon, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Heterodyning, also known as frequency mixing, is a frequency translation process. Heterodyning has its root in radio engineering. The principle of heterodyning was discovered in the late 1910s by radio engineers experimenting with radio vacuum tubes. Russian cellist and electronic engineer Leon Theremin invented the so-called Thereminvox, which was one of the earliest electronic musical instruments that generates an audio signal by combining two different radio signals. American electrical engineer Edwin Armstrong invented the so-called super-heterodyne receiver. The receiver shifts the spectrum of the modulated signal, that is, the frequency contents of the modulated signal, so as to demodulate the audio information from it. Commercial amplitude modulation (AM) broadcast receivers nowadays are super-heterodyne type. After illustrating the basic principle of heterodyning by some simple mathematics, we will discuss some examples on how the principle of heterodyning is used in radio and in optics.

For a simple case, when signals of two different frequencies are heterodyned or mixed, the resulting signal produces two new frequencies, the sum and difference of the two original frequencies. Figure 1 illustrates how signals of two frequencies are mixed or heterodyned to produce two new frequencies, $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$, by simply multiplying the two signals $\cos(\omega_1 t + \theta_1)$ and $\cos(\omega_2 t)$, where ω_1 and ω_2 are radian frequencies of the two signals and θ_1 is the phase angle between the two signals. Note that when the frequencies of the two signals to be heterodyned are the same, i.e., $\omega_1 = \omega_2$, the phase information of $\cos(\omega_1 t + \theta_1)$ can be extracted to get $\cos\theta_1$ if we use an electronic lowpass filter (LPF) to filter out the term $\cos(2\omega_1 t + \theta_1)$. This is shown in Figure 2. Heterodyning is often referred to as homodyning for the mixing of two signals of the same frequency.

For a general case of heterodyning, we can have a signal represented by $s(t)$ with its spectrum $S(\omega)$, which is given by the Fourier transform of $s(t)$, and when it is multiplied by $\cos(\omega_2 t)$, the resulting spectrum is frequency-shifted to new locations in the frequency domain as:

$$\mathcal{F}\{s(t)\cos(\omega_2 t)\} = \frac{1}{2}S(\omega - \omega_2) + \frac{1}{2}S(\omega + \omega_2) \quad [1]$$

where $\mathcal{F}\{s(t)\} = S(\omega)$ and $\mathcal{F}\{\cdot\}$ denotes the Fourier transform of the quantity being bracketed. The spectrum of $s(t)\cos(\omega_2 t)$, along with the spectrum of $s(t)$, is illustrated in Figure 3. It is clear that multiplying $s(t)$ with $\cos(\omega_2 t)$ is a process of heterodyning, as we have translated the spectrum of the signal $s(t)$. This process is known as modulation in communication systems.

In order to appreciate the process of heterodyning let us now consider, for example, heterodyning in radio. In particular, we consider AM. While the frequency content of an audio signal (from 0 to around 3.5 kHz) is suitable to be transmitted over a pair of wires or coaxial cables, it is, however, difficult to be transmitted in air. By impressing the audio information onto a higher frequency, say 550 kHz, as one of the broadcast bands in AM radio, i.e., by

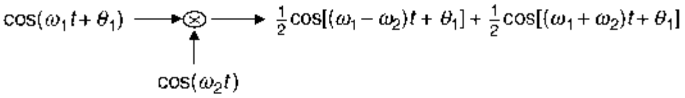
$\cos(\omega_1 t + \theta_1)$
$\frac{1}{2}\cos[(\omega_1 - \omega_2)t + \theta_1] + \frac{1}{2}\cos[(\omega_1 + \omega_2)t + \theta_1]$
$\cos(\omega_2 t)$

Figure 1 Heterodyning of two sinusoidal signals.

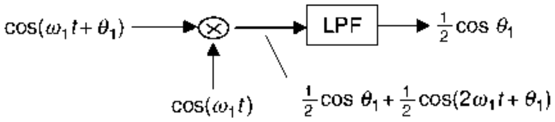
$\cos(\omega_1 t + \theta_1)$
LPF
$\frac{1}{2}\cos\theta_1$
$\cos(\omega_1 t)$
$\frac{1}{2}\cos\theta_1 + \frac{1}{2}\cos(2\omega_1 t + \theta_1)$

Figure 2 Heterodyning becomes homodyning when the two frequencies to be mixed are the same: homodyning allows the extraction of the phase information of the signal.

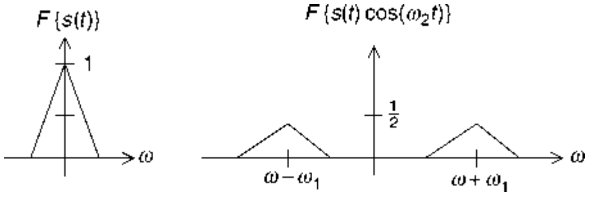

Figure 3 Spectrum of $s(t)$ and $s(t)\cos(\omega_2 t)$. It is clear that spectrum of $s(t)$ has been translated to new locations upon multiplying a sinusoidal signal.

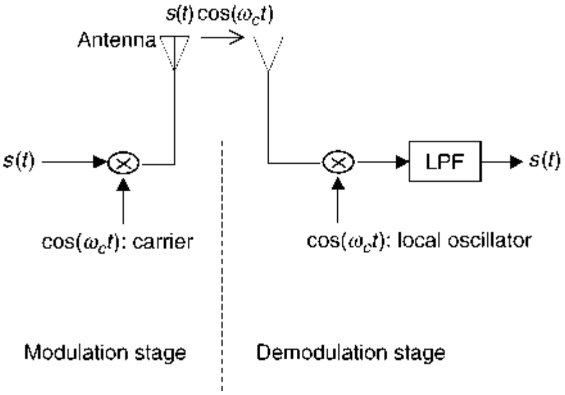

Figure 4 Radio link: heterodyning in the demodulation stage is often known as heterodyne detection.

modulating the audio signal, the resulting modulated signal can now be transmitted using antennas of reasonable dimensions. In order to recover (demodulate) the audio signal from the modulated signal, the modulated signal is first received by an antenna in a radio receiver, multiplied by the so-called local oscillator with the same frequency as the signal used to modulate the audio signal, then followed by a lowpass filter to eventually obtain the audio information back. The situation is illustrated in Figure 4. We shall now describe the process mathematically with reference to Figure 4.

We denote $s(t)$ as the audio signal. After multiplying by $\cos(\omega_c t)$, which is usually called a carrier in radio in the modulation stage, where ω_c is the radian frequency of the carrier. The modulated signal $s(t)\cos(\omega_c t)$. is transmitted via an antenna. When the modulated signal is received in the demodulation stage, the modulated signal is then multiplied by the local oscillator of signal waveform $\cos(\omega_c t)$. The output of the multiplier is given by

$$s(t)\cos(\omega_c t)\cos(\omega_c t) = s(t)\tfrac{1}{2}[1+\cos(2\omega_c t)] \qquad [2]$$

Note that the two signals in the demodulation stage, $s(t)\cos(\omega_c t)$ and $\cos(\omega_c t)$, are heterodyned to produce two new frequencies, the sum $\omega_c + \omega_c = 2\omega_c$ to give the term $\cos(2\omega_c t)$ and the difference $\omega_c - \omega_c = 0$ to give the term $\cos(0) = 1$. Since the two frequencies to be multiplied in the demodulation stage are the same, this is homodyning as explained above. Now by performing lowpass filtering (LPF), we can recover our original audio information $s(t)$. Note that if the frequency of the local oscillator in the demodulation stage within the receiver is higher than the frequency of the carrier used in the modulation stage, heterodyning in the receiver is referred to as superheterodyning, which most receivers nowadays use for amplitude modulation. In general, we have two heterodyning processes in the radio system just described, one in the modulation stage and the other in the demodulation stage. However, it is unusual to speak of heterodyning in the modulation stage, and so we just refer to the process in the demodulation stage as ‘heterodyne detection.’

In summary, heterodyne detection can extract the information from a modulated signal and can also extract the phase information of a signal if homodyning is used. We shall see, in the next section, how optical heterodyne detection is employed.

Optical information is usually carried by coherent light such as a laser. Let $\psi_p(x,y)$ be a complex amplitude of the light field, which may physically represent a component of the electric field. We further assume that the light field is oscillating at temporal frequency ω_0. Therefore, we can write the light field as $\psi_p\exp(i\omega_0 t)$. By taking the real part of $\psi_p\exp(i\omega_0 t)$, i.e., $\mathrm{Re}[\psi_p\exp(i\omega_0 t)]$, we recover the usual physical real quantity. For a simple example, if we let $\psi_p = A\exp(-ik_0 z)$, where A is some constant and k_0 is the wavenumber of the light, $\mathrm{Re}[\psi_p\exp(i\omega_0 t)] = \mathrm{Re}[A\exp(-ik_0 z)\exp(i\omega_0 t)] = A\cos(\omega_0 t - k_0 z)$, which is a plane wave propagating along the positive z direction in free space. To detect light energy, we use a photodetector (PD), as shown in Figure 5. Assuming a plane wave for simplicity, as $\psi_p = A$, we have also taken $z = 0$ at the surface of the photodetector. As the photodetector responds to intensity, i.e., $|\psi_p|^2$, which gives the current, i, as output by spatially integrating the intensity:

$$i \propto \int_S |\psi_p\exp(i\omega_0 t)|^2\mathrm{d}x\mathrm{d}y = A^2 S \qquad [3]$$

where S is the surface area of the photodetector. We can see that the photodetector current is proportional to the intensity, A^2, of the incident light. Hence the output current varies according to the intensity of the

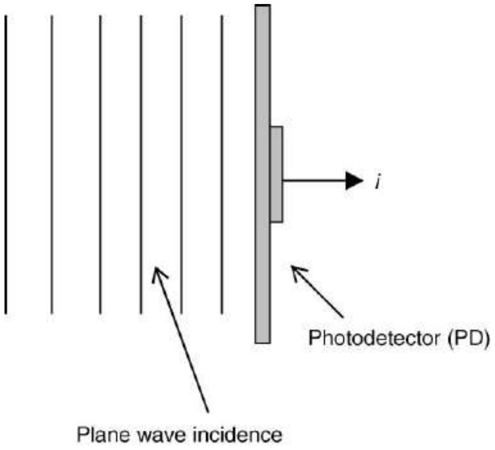

Figure 5 Optical direction detection or optical incoherent detection.

optical signal intensity. This mode of photodetection is called direct detection or incoherent detection in optics.

Let us now consider the heterodyning of two plane waves on the surface of the photodetector. We assume an information-carrying plane wave, also called the signal plane wave, $A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)] \times \exp(-ik_0 x \sin\phi)\}$, and a reference plane wave, $A_r \exp(i\omega_0 t)$, or called a local oscillator in radio. The situation is shown in Figure 6.

Note that the frequency of the signal plane wave is ω_s higher than that of the reference signal, and it is inclined at an angle ϕ with respect to the reference plane wave, which is normal incident to the photodetector. Also, the information content $s(t)$ is in the phase of the signal wave. In the situation shown in Figure 6, we see that the two plane waves are interfered on the surface of the photodetector, giving the total light field ψ_t, given by

$$\psi_t = A_r \exp(i\omega_0 t) + A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)]\} \times \exp(-ik_0 x \sin\phi) \qquad [4]$$

Again, the photodetector responds to intensity, giving the current

$$\begin{aligned} i &\propto \int_S |\psi_t|^2 \mathrm{d}x\mathrm{d}y \\ &= \int_{-a}^{a}\int_{-a}^{a} [A_r^2 + A_s^2 + 2A_r A_s \cos(\omega_s t + s(t) \\ &\quad - k_0 x \sin\phi)]\mathrm{d}x\mathrm{d}y \end{aligned} \qquad [5]$$

where we have assumed that the photodetector has a $2a \times 2a$ square area. The current can be evaluated to be:

$$i(t) \propto 2a(A_r^2 + A_s^2) + 4A_r A_s \frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} \cos(\omega_s t + s(t)) \qquad [6]$$

The current output has two parts: the DC current and the AC current. The AC current at frequency ω_s is commonly known as the heterodyne current. Note that the information content $s(t)$ originally embedded in the phase of the signal plane wave has now been preserved and transferred to the phase of the heterodyne current. The above process is called optical heterodyning. In optical communications, it is often referred to as optical coherent detection. In contrast, if the reference plane wave has not been used for the detection, we have the incoherent detection. The information content carried by the signal plane wave would be lost, as it is evident that for $A_r = 0$, the above equation gives only a DC current at a value proportional to the intensity of the plane wave, A_s^2.

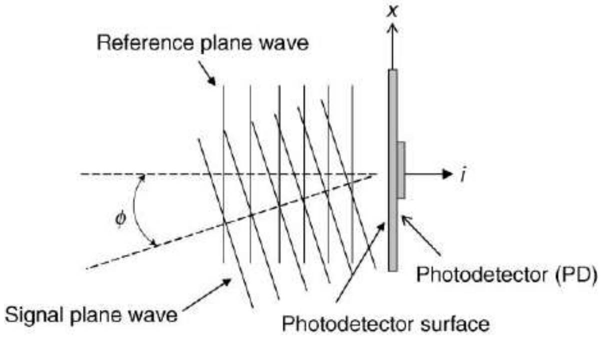

Figure 6 Optical heterodyne detection.

Let us now consider some of the practical issues encountered in coherent detection. Again, the AC part of the current given by the above equation is the heterodyne current $i_{het}(t)$, given by

$$i_{het}(t) \propto A_r A_s \frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} \cos(\omega_s t + s(t)) \qquad [7]$$

We see that since the two plane waves propagate in slightly different directions, the heterodyne current output is degraded by a factor of

$$\frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} = a\mathrm{sinc}(k_0 a \sin\phi)$$

where $\mathrm{sinc}(x) = \sin(x)/x$. For small angles, i.e., $\sin\phi \approx \phi$, the current amplitude falls off as $\mathrm{sinc}(k_0 a\phi)$. Hence, the heterodyne current is at a maximum when the angular separation between the signal plane wave and the reference plane wave is zero, i.e., the two plane waves are propagating exactly in the same direction. The current will go to zero when $k_0 a\phi = \pi$, or $\phi = \lambda_0/2a$, where λ_0 is the wavelength of the laser light. To see how critical it is for the angle ϕ to be aligned in order to have any heterodyne output, we assume the size of the photodetector $2a = 1$ cm and the laser used is red, i.e., $\lambda_0 \approx 0.6$ μm; ϕ is calculated to be about 2.3×10^{-3} degrees. Hence to be able to work with coherent detection, we need to have precise optomechanical mounts for angular rotation.

Finally we describe a famous application of heterodyning in optics – holography. As we have seen, the mixing of two light fields with different temporal frequencies will produce heterodyne current at the output of the photodetector. But we can record the two spatial light fields of the same temporal frequency with photographic films instead of using electrical devices. Photographic films respond to light intensity as well. We discuss holographic recording of a point source object as a simple example. Figure 7 shows a collimated laser split into two plane waves and recombined by the use of two mirrors (M) and two beamsplitters (BS). One plane wave is used to illuminate the pinhole aperture (our point source object), and the other

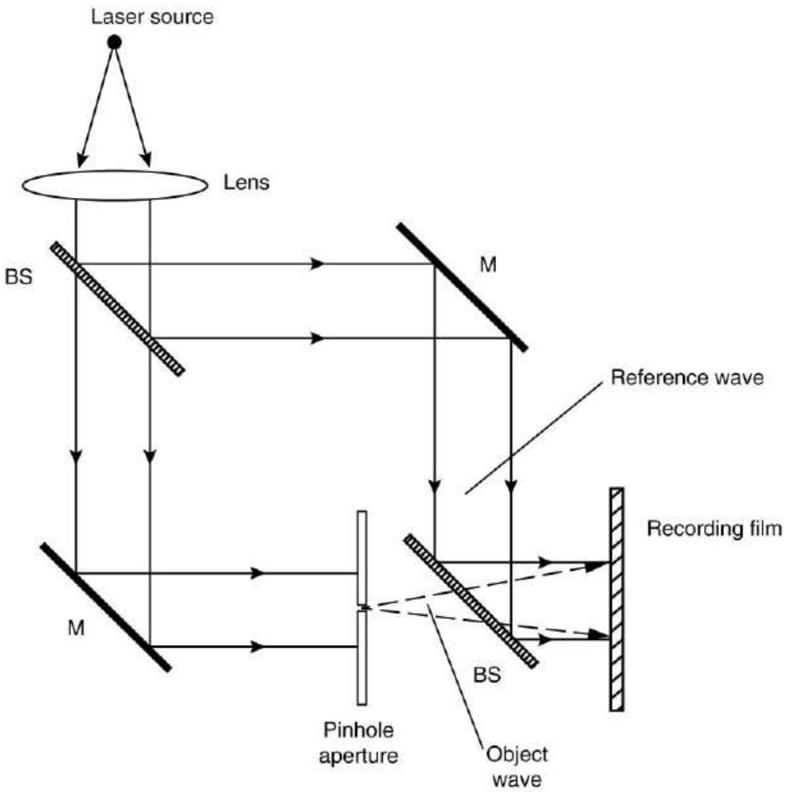

Figure 7 Holographic recording of a point source object.

illuminates the recording film directly. The plane wave that is scattered by the point source object, which is located z_0 away from the film, generates a diverging spherical wave. This diverging wave is known as an object wave in holography. The object wave arising from the point source object on the film is given by, according to Fresnel diffraction:

$$\psi_0 = A_0 \exp(-ik_0 z_0)\ \frac{ik_0}{2\pi z_0} \times \exp[-ik_0(x^2+y)^2/2z_0]\exp(i\omega_0 t) \quad [8]$$

This object wave is a spherical wave, where A_0 is the amplitude of the spherical wave.

The plane wave that directly illuminates the photographic plate is known as a reference wave, ψ_r. For the reference plane wave, we assume that the plane wave has the same phase with the point source object at a distance z_0 away from the film. Its field distribution on the film is, therefore, $\psi_r = A_r \exp(-ik_0 z_0)\exp(i\omega_0 t)$, where A_r is the amplitude of the plane wave. The film now records the interference of the reference wave and the object wave, i.e., what is recorded on the film as a 2D pattern is given by $t(x,y) \propto |\psi_r + \psi_0|^2$. The resulting recorded 2D pattern, $t(x,y)$, is called the hologram of the object. This kind of recording is known as holographic recording, distinct from a photographic recording in which the reference wave does not exist and hence only the object wave is recorded. Now:

$$\begin{aligned} t(x,y) &\propto |\psi_r + \psi_0|^2 \\ &= \Big| A_r \exp(-ik_0 z_0)\exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0) \\ &\quad \times \frac{ik_0}{2\pi z_0}\exp[-ik_0(x^2+y)^2/2z_0]\exp(i\omega_0 t)\Big|^2 \\ &= A + B\sin\left\{\frac{k_0}{2z_0}[(x^2+y)^2/2z_0]\right\} \end{aligned} \quad [9]$$

where A and B are some inessential constants. Note that the result of recording two spatial light fields of the same temporal frequency preserves the phase information (noticeably the depth parameter z_0) of the object wave. This is considered optical homodyning as it is clear from the result shown in Figure 2.

The intensity distribution being recorded on the film, upon being developed, will have transmittance given by the above equation. This expression is called the Fresnel zone plate, which is the hologram of a point source object and we shall call it the point-object hologram. Figure 8 shows the hologram for a particular value of z_0 and k_0. The importance of this phase-preserving recording is that when we illuminate the hologram with a plane wave, called the reconstruction wave, a point object is reconstructed at the same location of the original point source object if we look towards the hologram as shown in Figure 9, that is the point source is reconstructed at a distance z_0 from the hologram as if it were at the same distance away from the recording film. For an arbitrary 3D object, we can imagine the object as a collection of points, and therefore, we can see that we have a collection of Fresnel zone plates on the hologram. Upon reconstruction, such as the illumination of the hologram by a plane wave, the observer would see a 3D virtual object behind a hologram.

As a final example, we discuss a state-of-the-art holographic recording technique called optical scanning holography, which employs the use of optical heterodyning and electronic homodyning to achieve real-time holographic recording without the use of films. We will take the holographic recording of a point object as an example. Suppose we superimpose a plane wave and a spherical wave of different temporal frequencies and use the resulting intensity pattern as an optical scanning beam to raster scan a point object located z_0 away from the point source that generates the spherical wave. The situation is shown in Figure 10. Point C is the point source that generates the spherical wave (shown with dashed lines) on the pinhole object, our point object. This point source, for example, can be generated by a focusing lens. The solid parallel rays represent the plane wave. The interference of the two waves generates a Fresnel zone plate type pattern on the pinhole object:

$$\begin{aligned} &I_s(x, y;\ z_0, t) \\ &= \left| A_r \exp(-ik_0 z_0) \exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0) \frac{ik_0}{2\pi z_0} \right. \\ &\quad \left. \times \exp[-ik_0(x^2 + y)^2/2z_0] \exp[i(\omega_0 + \Omega)t] \right|^2 \\ &= 1 + \left(\frac{1}{\lambda_0 z_0} \right)^2 + \frac{1}{\lambda_0 z_0} \sin\left[\frac{\pi}{\lambda_0 z_0}(x^2 + y^2) - \Omega t \right] \end{aligned} \qquad [10]$$

assuming $A_r = A_0 = 1$ for simplicity. Note that the plane wave is at temporal frequency ω_0, and the spherical wave is at temporal frequency $\omega_0 + \Omega$. The expression $I_s(x, y;\ z_0, t)$ is a temporally modulated Fresnel zone plate and is known as the time-dependent Fresnel zone plate (TDFZP). If we freeze time, say at $t = t_0$, we have the Fresnel zone plate pattern on the pinhole object as shown in Figure 10. However, if we let the time run in the above expression, physically we will have running zones that would be moving away from the center of the pattern. These running zones are the result of optical heterodyning of the plane wave and the spherical wave of different temporal frequencies. It is this TDFZP that is used to raster scan a 3D object to obtain holographic information of the scanned object and such technique is called optical scanning holography.

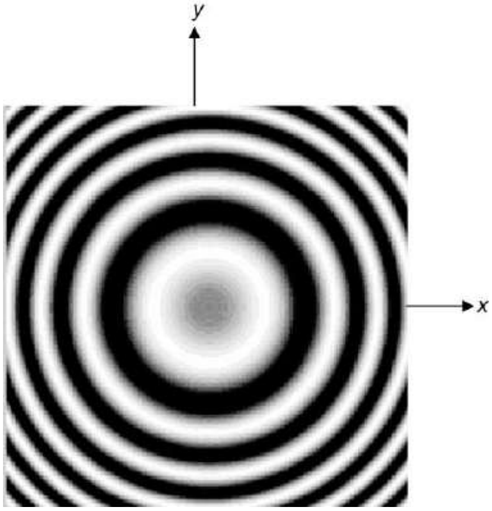

Figure 8 Point-object hologram.

Upon scanning by the TDFZP, the photodetector, captures all the transmitted light and delivers a current, which consists of a heterodyne current at

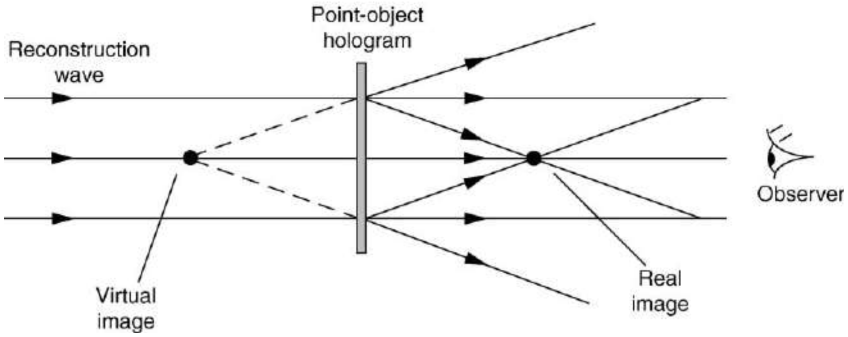

Figure 9 Reconstruction of a point-object hologram.

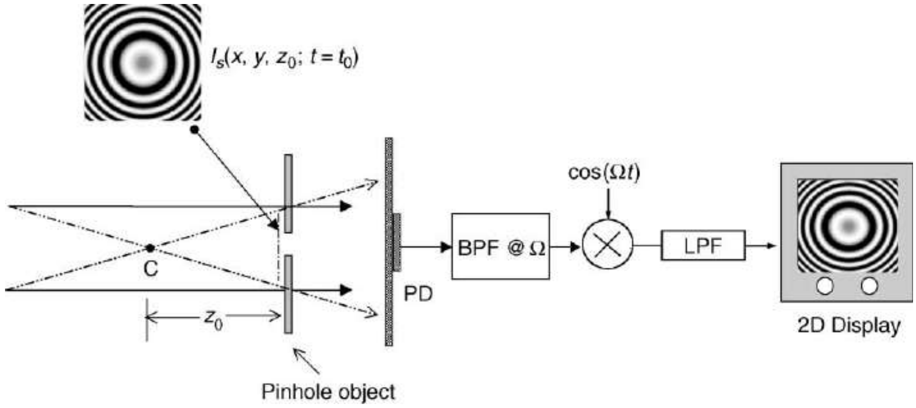

Figure 10 Optical scanning holography (use of optical heterodyning and electronic homodyning to record holographic information upon scanning the object in two dimensions).

frequency Ω. After electronic bandpass filtering (BPF) at Ω, the heterodyne current is homodyned electronically by cos(Ω*t*) and lowpass filtered (LPF) to extract the phase information of the current. When this final scanned and processed current is display in a 2D display, as shown in Figure 10, we have the Fresnel zone plate, which is the hologram of the pinhole object, on the display. We can photograph this 2D display to obtain a transparency and have it illuminated by a plane wave to have the reconstruction as shown in Figure 9, or, since the hologram is now in electronic form, we can store it in a PC and reconstruct it digitally. When holographic reconstruction is performed digitally, we have so-called digital holography.

See also

Diffraction: Fresnel Diffraction. **Holography, Techniques:** Digital Holography.

Further Reading

Banerjee PP and Poon TC (1991) *Principles of Applied Optics*. MA: Irwin.
Palais JC (1988) *Fiber Optic Communications*. New Jersey: Prentice-Hall.
Poon TC and Banerjee PP (2001) *Contemporary Optical Image Processing with MATLAB*. Oxford: Elsevier.
Poon TC and Qi Y (2003) Novel real-time joint-transform correlation using acousto-optic heterodyning. *Applied Optics* 42: 4663–4669.
Poon TC, Wu MH, Shinoda K and Suzuki Y (1996) Optical scanning holography. *Proc. IEEE* 84: 753–764.
Pratt WK (1969) *Laser Communication Systems*. New York: Wiley.
Stremler FG (1990) *Introduction to Communication Systems*. MA: Addison-Wesley.
VanderLugt A (1992) *Optical Signal Processing*. New York: Wiley.

Image Post-Processing and Electronic Distribution

K M Iftekharuddin, The University of Memphis, Memphis, TN, USA

F Ahmed, The Catholic University of America, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The spectacular development of imaging sensors, communication, and internet infrastructure has immensely facilitated the creation, processing, and distribution of electronic digital images. One nontrivial aspect of the digital image is its complete application dependence. The image may be created using a number of different acquisition methods and sensors. It is then processed in different ways, depending upon myriad of digital imaging applications. Consequently, the effectiveness of a useful electronic image processing engine may involve development of robust pre-processing techniques, different types of goal-directed processing such as detection, recognition, identification, classification, tracking, reconstruction and/or registration, as well as post-processing algorithms for different application areas such as multimedia, biomedical, astronomical, defense, consumer, industrial, etc. A notional schematic of this electronic image processing cycle is shown in Figure 1. Note that while the goal-directed techniques may be different, based on problems at hand, the pre- and post-processing techniques and tools are very similar. Thus, most of the common image pre-processing techniques are also applicable for image post-processing purposes.

Electronic image post-processing techniques involve three broad domains such as: (i) spatial; (ii) frequency; and (iii) time-frequency. Each of these domain techniques may be grouped into different types of processing such as: (i) spatial domain technique to include filtering, spatial processing such as histogram modification, morphological processing, texture processing, etc.; (ii) frequency domain technique to include filter; and (iii) time-frequency domain technique to include short-time Fourier transform, wavelet, and Wigner-ville transforms.

On the other hand, it is useful to identify the typical lifecycle of a digital image from the distribution perspective as shown in Figure 2. The image may be created using a number of different acquisition methods and sensors as mentioned above. The processing block next combines all different types of processing, as mentioned in Figure 1. The image may then be transmitted over the internet or it can leave the digital world to go to the print world. It may then go through wear/tear, subsequently being scanned back to digital domain again. After some processing, it may again be distributed over the internet.

Consequently, the image may be copied legally or fraudulently, processed, and used by a user. Further, the threats and vulnerabilities of the internet have also increased proportionately. The rapid development and availability of high-end computers, printers, and scanners has made the tools of counterfeiting easily accessible and more affordable. With these products, one may counterfeit digital images, which are easy to distribute and modify. As a result, copyright ownership poses a formidable challenge to image post-processing and distribution.

In the subsequent sections, we briefly review relevant background in image post-processing and distribution techniques. We also present a few representative application examples of image post-processing and copyright protection in image distribution.

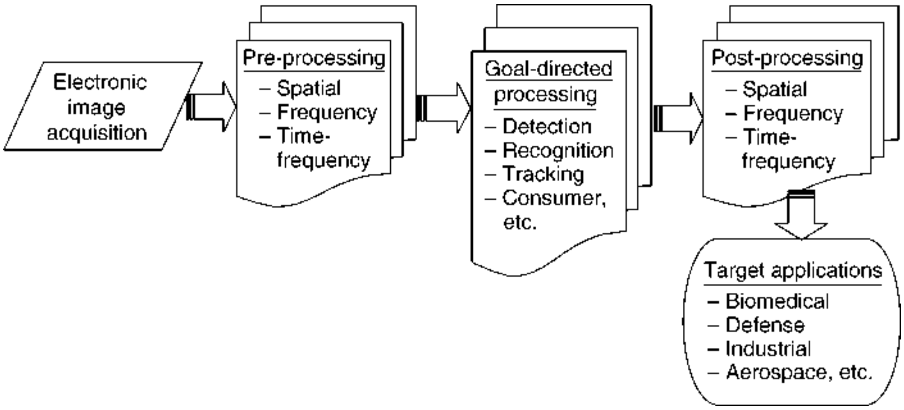
Electronic image acquisition
Pre-processing
– Spatial
– Frequency
– Time-frequency
Goal-directed processing
– Detection
– Recognition
– Tracking
– Consumer, etc.
Post-processing
– Spatial
– Frequency
– Time-frequency
Target applications
– Biomedical
– Defense
– Industrial
– Aerospace, etc.

Figure 1 Notional schematic of electronic image processing.

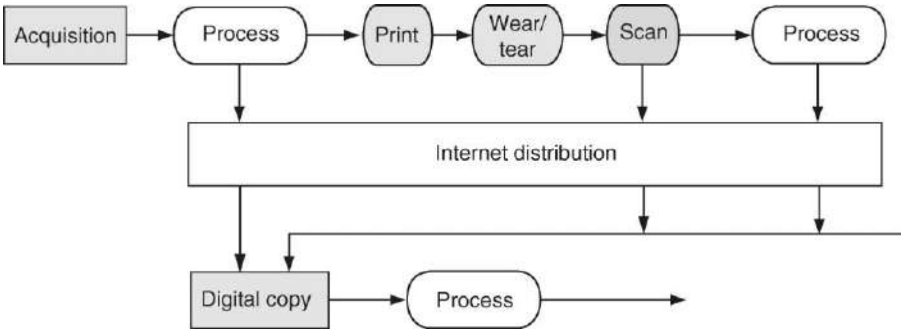
Acquisition
Process
Print
Wear/tear
Scan
Process
Internet distribution
Digital copy
Process

Figure 2 Typical lifecycle of a digital image distribution cycle.

Techniques in Image Post-Processing

Image Formation

We begin by describing the image formation technique briefly. Consider that a three-dimensional (3D) object or scene is imaged onto a 2D imaging plane by means of a recording system such as a camera. If the image formation process is linear, the recorded image may be modeled as the output of the system shown in Figure 3, which is given mathematically by

$$g(x,y) = s\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y;s,t)f(s,t)\mathrm{d}s\,\mathrm{d}t\right\} + n(x,y) \qquad [1]$$

where $g(x,y)$ denotes the recorded output image, and $f(x,y)$ is the ideal electronic image. The function $h(x,y;s,t)$ is the 2D impulse response or the point-spread function (PSF) of the image formation system. The PSF determines the radiant energy distribution in the image plane due to a point source located in the object plane. Usually the PSF is normalized such that

$$\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y;s,t)\mathrm{d}s\,\mathrm{d}t\right\} = 1 \qquad [2]$$

The noise contribution $n(x,y)$ is shown as an additive random process that is statistically uncorrelated with the image. This is a simplification because noises such as film-grain noise and the noise caused by photon statistics, which often corrupt images, are not uncorrelated with the input. This simplification nonetheless leads to reasonable and useful results. One of the primary goals of image pre-processing (and post-processing as well) is to devise techniques that reduce this additive random noise.

The PSF is a function of only the argument differences $x-s$ and $y-t$ for stationary image and object fields. Thus, it is possible to envision an image formation system that is space invariant; hence the PSF is independent of position. In this case, we may rewrite the superposition integral in eqn [1] into a more familiar convolution integral as follows:

$$g(x,y) = \left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x-s,y-t)f(s,t)\mathrm{d}s\,\mathrm{d}t\right\} + n(x,y) \qquad [3]$$

Equivalently, eqn [3] is written as

$$g(x,y) = \{h(x,y)^{*}f(x,y)\} + n(x,y) \qquad [4]$$

where ($*$) is used to denote 2D convolution. Converting from a continuous image $f(x,y)$ to its discrete representation $f(i,j)$ requires the process of sampling. In the ideal sampling system, $f(x,y)$ is multiplied by an ideal 2D impulse train given as

$$f(i,j) = \sum_{m=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty} f(mX_0,nY_0)\,\delta(x-mX_0,y-Y_0) \qquad [5]$$

where X_0 and Y_0 are the sampling distances or intervals, $\delta(^*,^*)$ is the ideal impulse function. Note that the square sampling implies that $X_0 = Y_0$. Sampling with an impulse function corresponds to sampling with an infinitesimally small point. If one is interested in image processing, one should choose a sampling density based upon the classical Nyquist sampling theory. Thus, in discrete domain implementation, eqns [3] and [4] are expressed as

$$g(i,j) = \left\{\sum_{\forall(k,l)} h(i,j;k,l)f(k,l)\right\} + n(i,j),\ 0 \le i,j \le N-1 \qquad [6]$$

and,

$$g(i,j) = s\left\{\sum_{\forall(k,l)} h(i-k,j-l)f(k,l)\right\} + n(i,j) \qquad [7]$$

Thus, we may rewrite eqn [6] in discrete domain as follows:

$$g(i,j) = h(i,j)^{*}f(i,j) + n(i,j) \qquad [8]$$

where ($*$) represents convolution operation. Equation [8] also describes the spatial domain filtering wherein $h(x,y;s,t)$ may be considered as the spatial domain filter mask. The discrete Fourier transform may be used to yield the frequency domain model as follows:

$$G(u,v) = H(u,v)F(u,v) + N(u,v) \qquad [9]$$

where G, H, F, and N are corresponding Fourier transforms and u and v are the dummy variables. In eqn [9], $H(u,v)$ is again the frequency domain filter.

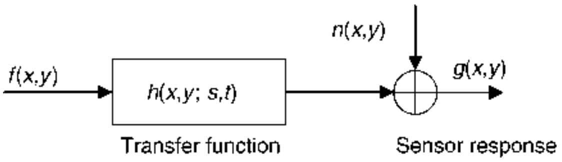

Figure 3 Model for image formation and recording.

We discuss the frequency domain filtering concept further, later in this article.

Spatial Domain Processing

The image pixel domain operations are frequently described as the spatial domain processing. These techniques are important for image pre- as well as post-processing operations. We describe two representative spatial domain processing techniques such as histogram processing and fractal processing for texture analysis below.

Histogram processing

One of the important issues in post-processing is image enhancement. Distortions in image may be introduced due to different noise factors, relative motion between a camera and the object, a camera that is out of focus, or atmospheric turbulence. Noise may be due to thermal measurement errors, recording medium, transmission medium, or digitization process. The histogram processing operation is one of the simplest techniques for enhancing the distorted image for certain type of noises. Histogram equalization is a contrast enhancement technique with the objective to obtain a new enhanced image with a uniform histogram. This can be achieved by using the normalized cumulative histogram as the gray scale mapping function. Histogram modeling is usually introduced using continuous domain functions. Consider that the images of interest contain continuous intensity levels in the interval [0,1] and that the transformation function τ which maps an input image $f(x,y)$ onto an output image $w(x,y)$ is continuous within this interval. We assume that the transfer function, $K_\mathrm{f} = \tau(K_\mathrm{w})$, is single-valued and monotonically increasing such that it is possible to define the inverse law $K_\mathrm{w} = \tau^{-1}(K_\mathrm{f})$. An example of such a transfer function is illustrated in Figure 4. All pixels in the input image, with densities in the interval D_f to $D_\mathrm{f} + \mathrm{d}D_\mathrm{f}$, are mapped using the transfer function $K_\mathrm{f} = \tau(K_\mathrm{w})$, such that they assume an output pixel density value in the interval from D_w to $D_\mathrm{w} + \mathrm{d}D_\mathrm{w}$. Consider that if the histogram h is regarded as a continuous probability density function p, describing the distribution of the intensity levels, then we obtain:

$$p_\mathrm{w}(w) = p_\mathrm{f}(D_\mathrm{f})/\mathrm{d}(D_\mathrm{f}) \qquad [10]$$

where

$$D_\mathrm{f} = \mathrm{d}u(f)/\mathrm{d}f \qquad [11]$$

In the case of histogram equalization, the output probability densities are all an equal fraction of the maximum number of intensity levels in the input image, thus generating a uniform intensity distribution. Sometimes it is desirable to control the shape of the output histogram in order to enhance certain intensity levels in an image. This can be accomplished by the histogram specialization operator which maps a given intensity distribution into a desired distribution using a histogram equalized image as an intermediate stage.

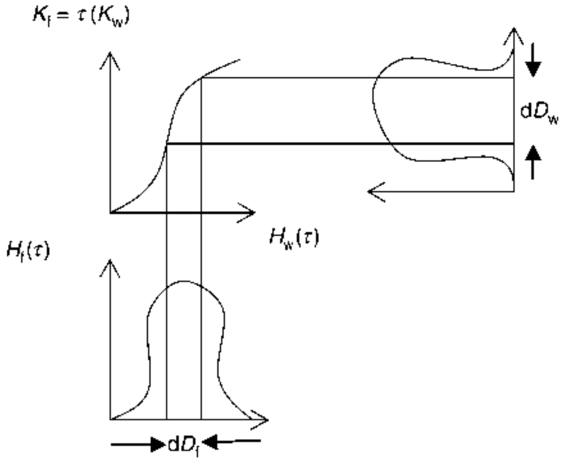

Figure 4 A histogram transformation function.

Fractals for Texture Analysis

The fractal aspect of an image has successfully been exploited for image texture analysis. The fractal concept developed by Mandelbrot, who coined the term fractal from the Latin 'fractus', provides a useful tool to explain a variety of naturally occurring phenomena. A fractal is an irregular geometric object with an infinite nesting of structure at all scales. Fractal objects can be found everywhere in nature such as coastlines, fern trees, snowflakes, clouds, mountains, and bacteria. Some of the most important properties of fractals are self-similarity, chaos, and noninteger fractal dimension (FD). The FD of an object can be correlated with the object properties. Object textures in images are candidates for characterization using fractal analysis because of their highly complex structure. The fractal theory developed by Mandelbrot is based, in part, on the work of mathematicians Hausdorff and Besicovitch. The Hausdorff–Besicovitch dimension, D_H, is defined as:

$$D_\mathrm{H} = \lim_{\varepsilon \to 0^+} \frac{\ln N}{\ln 1/r} \qquad [12]$$

where N is the number of elements of box size r required to form a cover of the object. The FD can be defined as the exponent of the number of self-similar pieces (N) with magnification factor ($1/r$) into which a figure may be broken. The FD is a noninteger value in contrast to objects that lie strictly in Euclidean space. The equation for FD is as follows:

$$\mathrm{FD} = \frac{\ln\text{ (number of self-similar pieces)}}{\ln\text{ (magnification factor)}} = \frac{\ln N}{\ln(1/r)} \qquad [13]$$

Frequency Domain Image Filtering

Ignoring noise contribution, we may obtain the basic filtering expression using eqns [7] and [8] as follows:

$$f(i,j) = \Im^{-1}[G(u,v)/H(u,v)] \qquad [14]$$

where $\Im^{-1}$ is the inverse Fourier transform and $[1/H(u,v)]$ is the inverse frequency domain filter. The key importance of this result is that, instead of performing a convolution as shown in eqn [7] to obtain the output of the system, we obtain the inverse of the Fourier transform of the impulse response $h(x,y;s,t)$ and the output image $g(x,y)$, multiply them and then take the inverse Fourier transform of the product to obtain $f(x,y)$. This is the basis for all the filtering as well as other types of image post-processing such as image reconstruction works. The filter function in eqn [9] may have different forms and shapes based on different types of functions such as uniform, Gaussian, lowpass, highpass, etc. While these are all linear filters, there are examples of nonlinear filters such as median, mean, etc. that are equally applicable for image post-processing.

Time-Frequency Domain Processing

One of the fundamental aspects of early image representation is the notion of scale. An inherent property of objects in the world and details in image is that they only exist as meaningful entities over certain ranges of scale. Consequently, a multiscale representation is of crucial importance if one aims at describing the structure of the world, or more specially the structure of projections of the three-dimensional world onto two-dimensional electronic images. The multiscale image post-processing may be facilitated by the wavelet (WT) computation.

A wavelet is, by definition, an amplitude-varying short waveform with a finite bandwidth. Naturally, wavelets are more effective than the sinusoids of Fourier analysis for matching and reconstructing signal features. In wavelet transformation and inversion, all transient or periodic data features can be detected and reconstructed by stretching or contracting a single wavelet to generate the matching building blocks. Consequently, wavelet analysis provides many flexible and effective ways to post-process images that surpass classical techniques – making it very attractive for data analysis, and modeling. The continuous WT is defined as

$$\Phi(a,b) = |a|^{-1/2}\int_{-W}^{W} x(t)\Psi\left[\frac{t-b}{a}\right]\mathrm{d}t \qquad [15]$$

where a and b are scaling and translation factors and $\Psi(t)$ is the mother wavelet. The WT may be viewed as a multiresolution analysis with shifted and scaled versions of the mother wavelet. Corresponding WT coefficients $c_{m,n}$ are

$$c_{m,n} = |a_0|^{-m/2}\int x(t)\Psi([a_0^{-m}t - nb_0])\mathrm{d}t \qquad [16]$$

The reconstruction of the signal is achieved using the admissibility condition, $\int \Psi(t)\mathrm{d}t = 0$, as follows:

$$x(t) = c\int\int_{a>0}\left[\Phi(a,b)\frac{\Psi a, b(t)}{a}\right]\mathrm{d}a\,\mathrm{d}b \qquad [17]$$

where c is a constant. For compact dyadic wavelets, binary scaling and dyadic translation of the mother wavelet form the basis functions. A dyadic translation is a shift by the amount $n/2^m$ that is an integer multiple of the binary scale factor and of the width of the wavelet as well. The discrete WT (DWT) is the natural choice for digital implementation of WT. The DWT basis and 'sub-band coding' process for discretization of a and b are identified. The iterative bandpass filter that finally leads to the wavelet is given as

$$\Psi m, n(x) = 2^{m/2}\Psi(2^{-m}x - n) \qquad [18]$$

The above eqn [18] forms the basis for 'sub-band' implementation for multiresolution analysis (MRA) of wavelet processing.

Image Post-Processing Examples

Some of the important application areas for electronic image post-processing involve the following:

(i) Biomedical image analysis and object detection;
(ii) Image compression for transmission and distribution;
(iii) Satellite image processing and noise removal;
(iv) IR/SAR/detection and clutter removal; and
(v) Image classification and machine vision.

In this article, we provide examples of some of these image-post-processing applications.

Example I: Spatial Domain Processing: Image Post-Processing in Biomedical Image

The mega voltage X-ray images are the X-ray images taken using high voltage X-radiations. These images are often characterized by low resolution, low contrast, and blur. The importance of mega voltage X-ray images arises in the context of radiation therapy for cancer treatment wherein the mega voltage X-radiation is used as a therapeutic agent. Figure 5 shows such mega voltage X-ray images, also known as an electronic portable digital (EPID) image with a very low level of detail. Figure 5a shows poor dynamic range. In order to improve the contrast of this image, without affecting the structure (i.e., geometry) of the information contained therein, we can apply the histogram equalization operation discussed in eqn [10]. Figure 5b presents the result of the histogram image enhancement post-processing process. The low level of detail in the bony anatomy is easily visible in Figure 5b.

Example II: Frequency Domain Processing: Image Post-Processing in Biomedical Tissue Detection

In medical imaging, texture is a fundamental characteristic that can be exploited to analyze tissues and pathologies in a biomedical image. The image texture is often characterized by the random nature of the objects such as tissues and body structures, noise, imaging modality, imaging limitations, and a variety of measurement parameters affecting the image acquisition process. Extensive research suggests that the fractal analysis may be combined with MRA techniques to post-process the texture content of an image in general and tumor detection in particular. In addition, the stochastic fractal Brownian motion (fBm) is well suited to describe texture and tumor characteristics since fBm offers an elegant integration of fractal and multiresolution analysis.

The fBm is a part of the set of $1/f$ processes, corresponding to a generalization of the ordinary Brownian motion $B_H(s)$. It is a nonstationary zero-mean Gaussian random function, defined as

$$B_H(t) - B_H(s) = \frac{1}{\Gamma(H+0.5)} \times \left\{ \int_{-\infty}^{0} [(t-s)^{H-0.5} - s^{H-0.5}] \mathrm{d}B(s) + \int_{0}^{\infty} (t-s)^{H-0.5} \mathrm{d}B(s) \right\}, \quad B_H(0) = 0 \qquad [19]$$

where the Hurst coefficient H, restricted to $0 < H < 1$, is the parameter that characterizes fBm; t and s correspond to different observation times of the process B_H, and Γ to the Euler's Gamma function. From the frequency domain perspective, it is known that some of the most frequently seen structures in fractal geometry, generally known as $1/f$ processes, show a power spectrum satisfying the power law relationship:

$$S(\omega) \propto \frac{k}{|\omega|^{\gamma}} \qquad [20]$$

where ω corresponds to the spatial frequency, and $\gamma = 2H + 1$. The observations of stationarity and self-similarity of fBm suggest that the use of time-frequency signal decomposition techniques, such as MRA, are well suited for the analysis of fractal signals. Thus, the fBm analysis framework, along with eqn [13], may be successfully exploited to estimate H, and hence, FD as follows:

$$\mathrm{FD} = D_E + 1 - H \qquad [21]$$

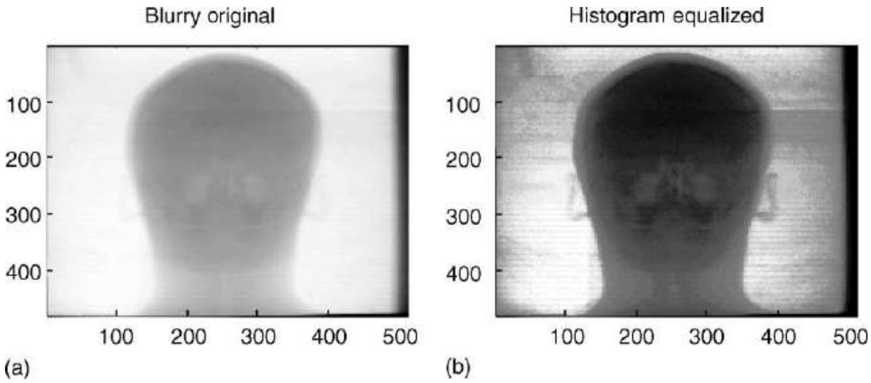

Figure 5 (a) Original EPID image; and (b) Histogram enhanced image.

where D_E is the Euclidean dimension that contains the fBm (i.e., the position of each point of the process is described with the vector $x = (x_1, \ldots, x_E)$).

We exploit the wavelet MRA combined with complementary signal processing techniques such as time-frequency descriptions and parameter estimation of stochastic signals in fBm framework for tumor detection in CT images. We demonstrate an example of image post-processing for tumor segmentation. We test our models to estimate FD in brain MR images as shown in Figures 6a and b, respectively. The images are scanned pixel-by-pixel with a sliding window that defines the subimage for analysis. Pre-processing, such as median filtering, is used to remove noise from the original images.

The results for the spectral analyses of CT image are shown in Figure 7. The results suggest that the FD values do not depend on the mother wavelet used. The window size that offers the best results is the 16 × 16 pixels, while better discrimination between low and high H values is obtained using biorthogonal wavelet. To improve tumor detection performance of the spectral technique, we apply post-processing filters on the results in Figures 8a and b, respectively. In Figure 8, for an example, we show the filtered fractal analysis results corresponding to a 16 × 16 pixels window with Daubechies 4. It is possible to obtain an even better description of the location and shape of the tumor after applying a two-dimensional post-processing filter to the matrices derived from the variance estimation algorithms with biorthogonal 1.5 wavelet and an 8 × 8 pixels window, respectively. The results are shown in Figure 9 and are similar for both averaging and Gaussian filters as well.

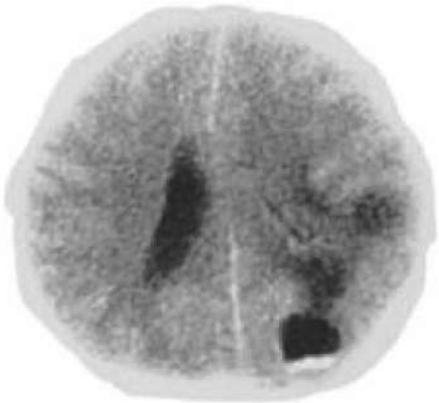

Figure 6 Biomedical images used to test the proposed algorithms.

Example III: Time-Frequency Domain Filtering: Image Post-Processing in Noise Reduction (De-Noising)

This section discusses an example of image post-processing in image noise removal (or de-noising) using a scale-frequency technique such as wavelets. The wavelet transform has been proved to be very successful in reducing noise (de-noising) in an image. We recently propose a new family of wavelets namely Joint Harmonic Wavelet Transform (JHWT) that offers better speckle noise reduction for image post-processing. Unlike wavelets generated by discrete dilation functions, as shown in eqn [18], whose shape cannot be expressed in functional form, JHWT has a simple structure, which can be written, in the

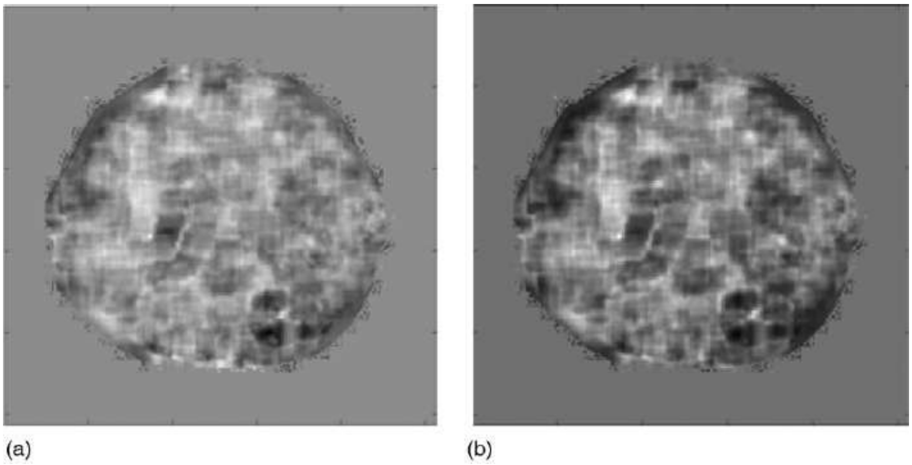

Figure 7 MR-Fractal analysis using spectral method (a) and (b) correspond to a 16 × 16 analyzing window for Daubechies 4 and biorthogonal 1.5 wavelets, respectively.

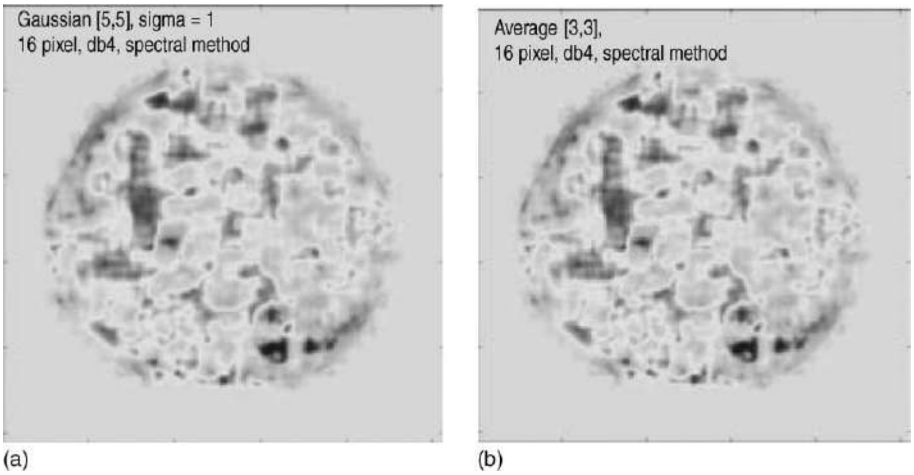
Gaussian [5,5], sigma = 1
16 pixel, db4, spectral method
(a)
Average [3,3],
16 pixel, db4, spectral method
(b)

Figure 8 Post-processing results using: (a) 5 × 5 Gaussian filtering with $\sigma = 1$ and (b) 3 × 3 averaging filter.

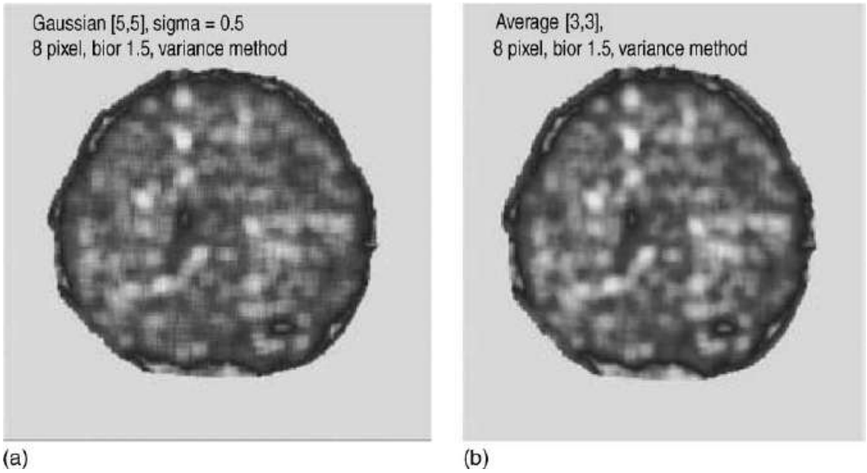
Gaussian [5,5], sigma = 0.5
8 pixel, bior 1.5, variance method
(a)
Average [3,3],
8 pixel, bior 1.5, variance method
(b)

Figure 9 Post-processing on the fractal analysis results of Figure 8 for (a) 5 × 5 Gaussian filtering with $\sigma = 1$, and (b) 3 × 3 averaging filtering.

frequency domain:

$$W(\omega) = \int_{-\infty}^{\infty} w(x)\exp(-\mathrm{i}\omega x)\mathrm{d}x, \\ = \begin{cases} 1 & \text{for } 2\pi \leq \omega < 4\pi \\ 0 & \text{elsewhere} \end{cases} \qquad [22]$$

Now, at level j, we let $g(x) = w(2^j x)$ and shift it by $k/2^j$ to obtain the harmonic wavelet as:

$$g(x - k/2^j) = \{\exp\{\mathrm{i}4\pi(2^j x - k)\} \\ - \exp\{\mathrm{i}2\pi(2^j x - k)\}\}/\mathrm{i}2\pi(2^j x - k) \qquad [23]$$

where integers j and k define the scale and translation of the wavelet $w(x)$, respectively on the x-axis. Now consider a real function $f(t)$ is represented by the sequence f_r of length $(n - 1)$, where $r = 0$ to $2^n - 1$. To obtain JHWT coefficients, we take FFT of f_r such that,

$$|F_{2^j+s}|^2 = (1/n^2)\sum_{k=0}^{n-1} |a_{2^j+k}|^2 \exp(-\mathrm{i}4\pi sk/n), \\ \text{where } r = 0 \text{ to } n-1 \text{ and } s = 0 \text{ to } n-1 \qquad [24]$$

where j denotes the level of the wavelet, k is the translation, m and n are adjacent levels of wavelets with $n = 2^j$, and $s = m - n$. The JHWT in eqn [24] offers all the properties and advantages of standard discrete wavelet transform. Putting $n = 2^j$ and $m = s + n$, eqn [24] yields all the JHWT coefficients from

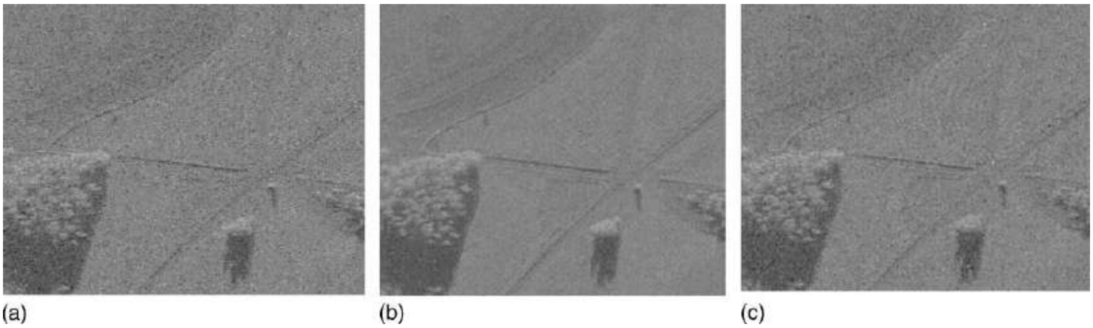

Figure 10 (a) Noisy image; (b) de-noised image with JHWT-filtering; and (c) de-noised image with WP-filtering.

the Fourier coefficients as

$$\sum_{j=0}^{N-1} |F_{2^j+4}|^2 = |a_0|^2 + \sum_{j=0}^{n-2} (1/2^j) \times \sum_{k=0}^{2^j-1} \left(|a_{2^j+k}|^2 + |a_{N-2^j-k}|^2 \right) + (a_{N/2})^2 = (1/N) \sum_{r=0}^{N-1} f_r^2 \qquad [25]$$

Unlike conventional correlation techniques, the perfect octave band limited characteristics of JHWT eliminate the necessity of any multiplication. We show an example of image de-noising in Figure 10 using JHWT.

Image Post-processing, Transmission, and Distribution

One of the areas wherein image post-processing plays a vital role is in image transmission. A significant problem for any digital library is the network bandwidth required to transmit large digital images. The bandwidth consumption depends on the number of images used by a single user, and with the number of users requesting the services of the library. The popularity of database browsers such as Gopher and Mosaic has serious implications for image transmission and distribution. There has been active research in innovative new compression schemes and transmission protocols that reduce bandwidth requirements for many users along with software servers and visualization. Further, compression of digital image and video data plays an important role in reducing the transmission and distribution cost for visual communication.

One of the requirements for effective image transmission and distribution is reduced data size for faster speed. Image compression schemes are often used for image data reduction. Most of the successful image compression schemes exploit subdividing the image into blocks for increased compression. However, this subdivision of image renders a blocky effect on the compressed image. There has been intense research in removing the blocky effect on the compressed image. To achieve high bit rate reduction while maintaining the best possible perceptual quality, post-processing techniques provide one attractive solution. There has been a significant amount of work done in post-processing of compressed image/video. Some of them are developed from the image enhancement point of view while others are from image restoration. In the case of image restoration, the post-processing is considered as a reconstruction problem. Image enhancement methods are performed based on the local contents of the decoded image as discussed above.

The image post-processing algorithms are, generally, very effective in removing blocking artifacts. A straightforward approach to remove blocking artifacts is to apply lowpass filtering to the region, where artifacts occur. Consequently, the filter must be adapted to the block boundaries, i.e., the image partition. Further, if high frequency details such as edges and texture are to be preserved, then lowpass filter coefficients should be selected appropriately in regions where such details occur. Therefore, adaptive image post-processing such as spatial filtering methods may be applied to deal with this problem. In low-bit-rate image and video compression and distribution systems, high frequency transformed coefficients are often lost due to aggressive quantization or uneven error protection schemes. However, with the block-based encoding and transmission methods, the amount of high frequency texture loss is uneven between adjacent blocks. Thus, it is possible to exploit this texture-level correlation between adjacent image blocks to reconstruct the lost texture.

A relevant problem of robust transmission and distribution of compressed electronic images involves unreliable networks. The pre- and post-processing of the image data may play a vital role in addressing this problem. With the rapid growth of mobile wireless communication systems, wireless multimedia has recently undergone enormous development. However, to design an efficient multimedia communication system over wireless channels, there still exist many challenges due to severe wireless channel conditions, special characteristics of the compressed multimedia data, and resource limitation, such as power supply. Since multimedia data, such as an image, contain redundancy, to efficiently utilize limited resources, source compression may be necessary. Further, as mentioned above, the ease of image creation, processing, transmission, and distribution has also magnified the associated counterfeiting problems. We discuss the relevant issues of image distribution in the following sections.

Techniques in Image Distribution

The spectacular development of imaging sensors, communication, and internet infrastructure has also made the counterfeiting tools easily accessible and more affordable. In addition, the threats and vulnerabilities of the internet have also increased manyfold. All of these together pose a formidable challenge to the secure distribution and post-processing of images.

Requirements of Secure Digital Image Distribution

In light of the above-mentioned vulnerabilities of image distribution channels and processes, below are a few requirements of secure image distribution:

- *Data Integrity and Authentication*: Digital image is easy to tamper with. Powerful publicly available image processing software packages such as Adobe Photoshop or PaintShop Pro make digital forgeries a reality. Integrity of the image content and authentication of the source and destination of the image distribution are, therefore, of utmost importance.
- *Digital Forensic Analysis*: Consider a litigation involving a medical image that is used for diagnosis and treatment of a disease, as shown in the post-processing example in Figures 7 or 8 above. Integrity of the image again is vital for the proper forensic analysis.
- *Copyright Protection*: Digital copyright protection has become an essential issue to keep the healthy growth of the internet and to protect the rights of on-line content providers. This is something that a textual notice alone cannot do, because it can so easily be forged. An adversary can steal an image, use an image processing program to replace the existing copyright notice with a different notice and then claim to own the copyright himself. This requirement has resulted in a generalized digital rights management initiative that involves copyright protection technologies, policies, and legal issues.
- *Transaction Tracking/Fingerprinting*: In the continuous (re)distribution of digital images through internet and other digital media (CD, camera, etc.), it is important to track the distribution points to track possible illegal distribution of images.

Next, we will discuss how the digital watermarking technology can help enforce these security requirements.

Example of secure image distribution: digital watermarking

Digital watermarking is the process of embedding a digital code (watermark) into a digital content like image, audio, or video. The embedded information, sometimes called a watermark, is dependent on the security requirements mentioned above. For example, if it is copyright application, the embedded information could be the copyright notice. Figure 11 shows a block diagram of the digital watermarking process.

Consider that we want to embed a message M in an image, f_o, also known as the original image. For the sake of brevity, we drop the indices representing the pixel coordinates. The message is first encoded using source coding and optionally with the help of error correction and detection coding, represented in Figure 11 as $e(M)$ such that

$$W_m = e(M) \qquad [26]$$

The encoded message W_m is then combined with the key-based reference pattern W_k, in addition to the scaling factor α, to result in W_a which is the signal that is actually added to the original image, given as

$$W_a = \alpha(f_o)[W_M \otimes W_K] \qquad [27]$$

Note that the scaling coefficient α may be a constant or image dependent or derived from a model of image perceptual quality. When the scaling factor α is dependent on the original image, we can have informed embedding which may result in more perceptually adaptive watermarking. The operation involving W_m and W_k is the modulation part, which is dependent on the actual watermarking algorithm used. As an example, in a typical spread-spectrum based watermarking, this operation is usually an exclusive NOR (XNOR) operation. The final operation is an additive or multiplicative embedding of W_a with the original image. We show a simple additive watermarking process as

$$f_w = f_o + W_a \qquad [28]$$

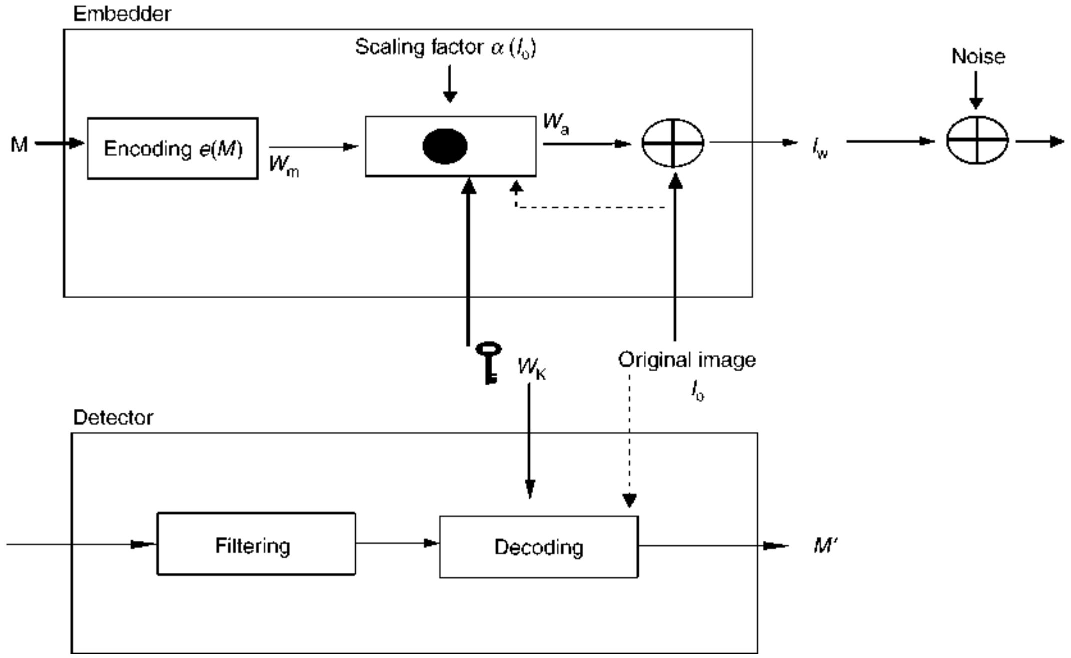

Figure 11 A simple watermarking process: embedder and detector.

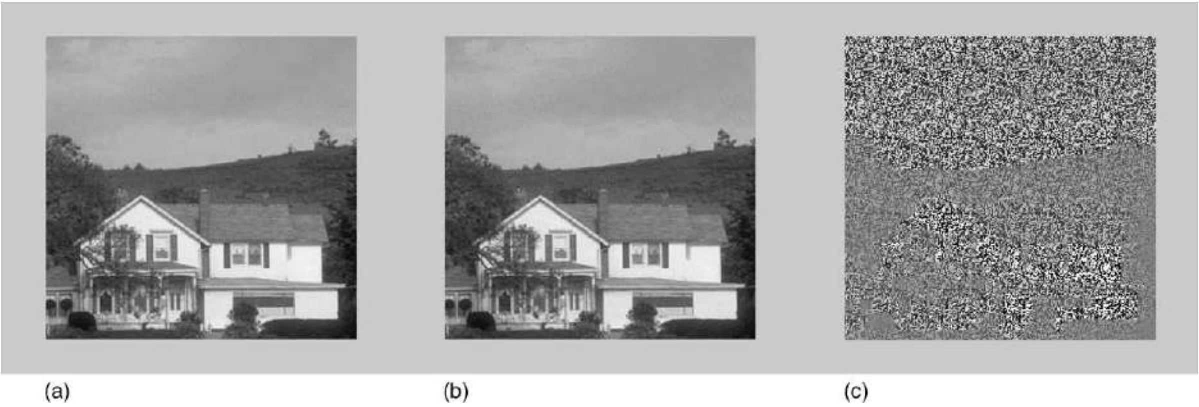

Figure 12 (a) Original image, (b) watermarked image, and (c) the difference image.

The result of this process is the watermarked image f_w which may actually need some clipping and clamping operations to have the image pixel values in the desired range.

The detector, as shown in Figure 11, does the inverse operations of the embedder. If the original image is known, as in the case of nonblind watermarking, it is first subtracted from the watermarked image. Blind watermarking detector, on the other hand, uses some filtering with a goal to estimate and subtract the original image component. The result of filtering is then passed to the decoder, which collects the message bits with the help of the shared key, and finally gets back the decoded message M'.

Figure 12 shows the original image, watermarked image, and the difference image using a spread-spectrum based watermarking algorithm, respectively. The enhanced difference image represents a noise-like image, wherein the scaling coefficient α is clearly dependent on the original image. As evident from Figure 12b, the hidden message is imperceptible as well as the change in the original image. The imperceptibility is one of the primary criteria of digital watermarking.

There are other criteria such as unobtrusiveness, capacity, robustness, security, and detectability. Thus, digital watermarking is a multidimensional problem, which attempts to make a trade-off among a number of different criteria. The inserted watermark should not be obtrusive to the intended use of the original image. Because the watermarking process does not increase the size of the original image, it may be desirable to add as much information as possible. However, generally speaking, the more information one adds, the more severe will be the impact on the perceptual quality of the image. From the detection point of view, watermark should be robust enough to tolerate a range of image post-processing and degradation. In addition, the embedded watermark should also be secure enough so that it may not be removed from the watermarked image. Finally, detectability of the embedded watermark is an important criterion that places some constraints on the embedding algorithm.

In order to fully appreciate the use of watermarking for image distribution, it is instructive to look at different classes of watermark, which is delineated in Figure 13.

For an example, depending on the robustness criterion of watermarking, they can be classified into three categories such as robust, fragile, and semi-fragile. Note that while the robustness of watermark is important, it is not always equally desirable in all applications. Let us take the authentication as an example. If you, as the owner of a digital image, embed some authentication information into the image using watermarking, you might want to see the authentication fail, when some one steals your image. This is an example of nonrobust watermark, where the fragility of the embedded watermark is required to detect any forging done on the digital image.

Now, consider how watermark provides the security requirements for secure image distribution. In particular, we look into the authentication of watermark. Figure 14 shows the use of watermark for image authentication, which has some similarities with the cryptographic authentication using digital signature. However, there are potential benefits to using watermarks in content authentication and verification. Unlike cryptography, watermarking offers both in-storage and in-transit authentication. Watermarking is also faster compared to encryption, which is very important in internet-based image distribution. By comparing a watermark against a known reference, it may be possible to infer not just that an alteration occurred but what, when and where changes have occurred. As shown in Figure 14, first we identify some features from the image and then we compute the digital signature, which is then embedded into the image. The detector takes a test image, computes the signature and extracts the embedded

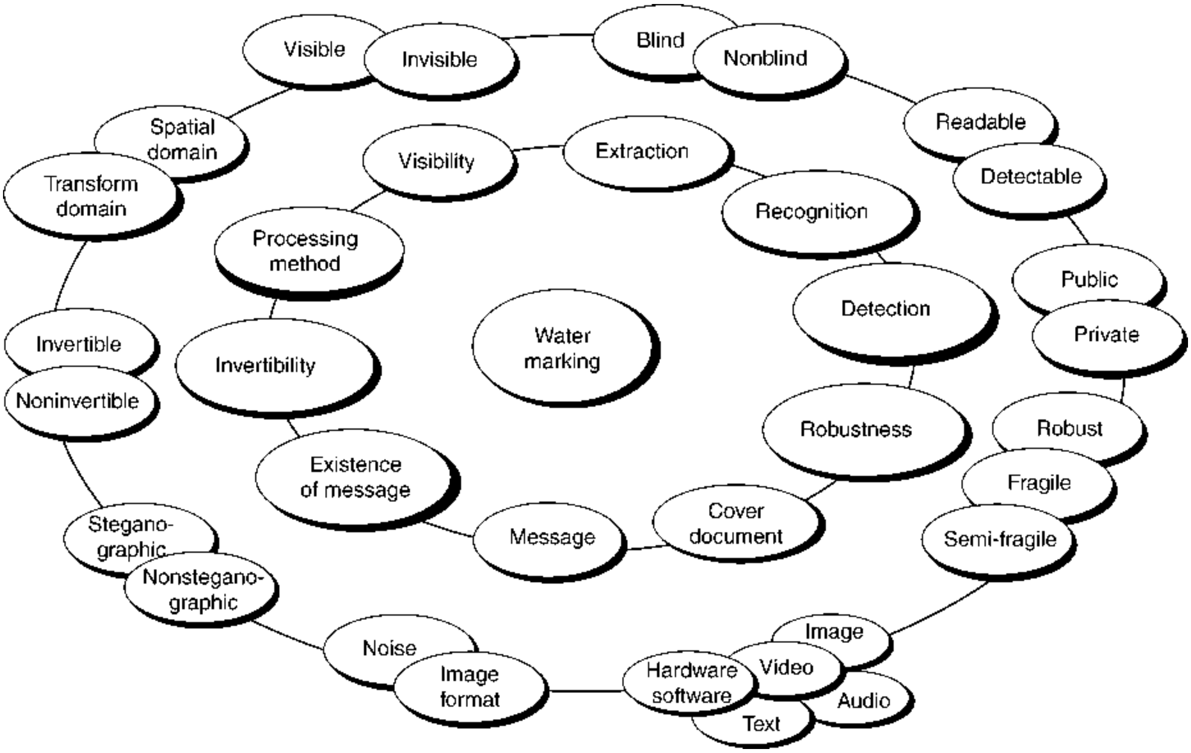
Visible
Invisible
Blind
Nonblind
Spatial domain
Transform domain
Readable
Detectable
Visibility
Extraction
Recognition
Processing method
Public
Private
Detection
Water marking
Invertible
Noninvertible
Invertibility
Robustness
Robust
Fragile
Semi-fragile
Existence of message
Message
Cover document
Stegano-graphic
Nonstegano-graphic
Noise
Image format
Hardware software
Image
Video
Audio
Text

Figure 13 Watermark classifications.

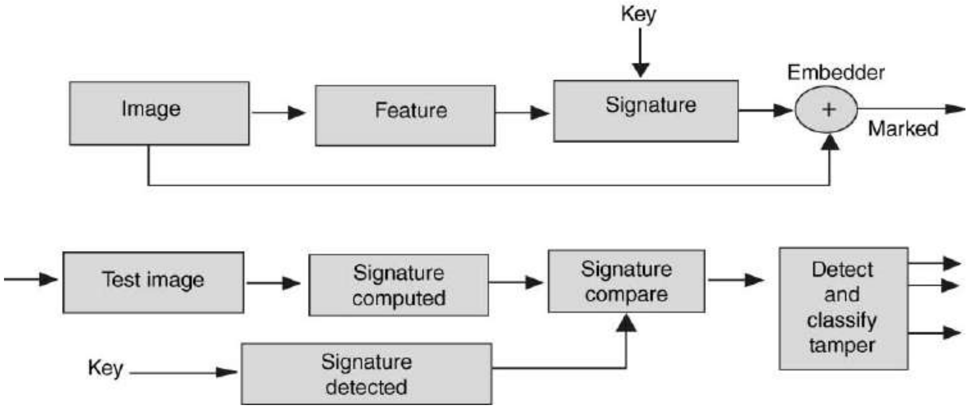

Figure 14 Image authentication using watermark.

signature, and then compares the two. If they match, the image is authenticated. If they do not match, then the image may have gone through some forging/tampering effects and the algorithm may optionally detect the tamper area as well.

One of the limitations of the watermarking approach in secure image distribution is that the technology is not yet widely deployed, and nor is the protocol satisfactorily standardized. Other major limitation is that the watermarking process cannot be made sufficiently robust to arbitrary types of different attacks. In reality, therefore, the hybrid combination of cryptography and watermarking is expected to improve the secure distribution of images across the unreliable internet.

Concluding Remarks

In this chapter, we discussed the principles of image post-processing and distribution. We showed a few representative examples for post-processing and distribution. We also explained the relevant issues in image distribution and present digital watermarking as one of the solutions for secure image distribution.

Acknowledgment

One of the authors (KMI) wishes to acknowledge the partial support of this work through a grant (RG-01-0125) from the Whitaker Foundation.

See also

Imaging: Information Theory in Imaging.

Further Reading

Ahmed F and Moskowitz IS (2004) A correlation-based watermarking method for image authentication applications, accepted for publication in *Optical Engineering*.

Barlaud M (1994) *Wavelets in Image Communication, Advances in Image Communication 5*. Amsterdam: Elsevier.

Chui A (1992) *An Introduction to Wavelets*. San Diego: Academic Press.

Cox NI, Bloom J and Miller M (2001) *Digital Watermarking: Principles & Practice*. Francis, CO: Morgan Kauffman Publishers.

Daubechies I (1990) The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory* 36(5): 961–1005.

Gonzalez RC and Woods RE (1993) *Digital Image Processing*. New Jersey: Addison-Wesley Publishing Company.

Iftekharuddin KM and Parra C (2004) Multiresolution-fractal analysis of medical images. Under Review. *Neuroimaging*, vol. 1. New York: Wiley/IEEE Press, Chapter 4.

Jain AK (1989) *Fundamental of Digital Image Processing*. New Jersey: Prentice-Hall.

Lu N (1997) *Fractal Imaging*. San Diego: Academic Press.

Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(7): 674–693.

Mandelbrot BB (1983) *The Fractal Geometry of Nature*. San Francisco, CA: Freeman.

Schmucker AM and Wolthusen SD (2003) *Techniques and Applications of Digital Watermarking and Content Protection*. Norwood, MA: Artech House.

Serra J (1982) *Image Analysis and Mathematical Morphology*. San Diego: Academic Press.

Stallings W (2003) *Cryptography and Network Security – Principles and Practices*, 3rd edn. New Jersey: Prentice-Hall.

Wu M and Liu B (2002) *Multimedia Data Hiding*. New York: Springer.

Smart Pixel Arrays

P Seitz, Centre Suisse d'Electronique et de Microtechnique SA, Zurich, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

For more than 50 years silicon has been the semiconductor on which the enormous progress of microelectronics is based. Since silicon is also an excellent detector for electromagnetic radiation, in particular in the visible spectrum between 400 and 700 nm, the advances in semiconductor technology have directly influenced the progress of solid-state image sensing. Once the relentlessly shrinking dimensions of transistors have reached a critical minimum limit, it has become economically feasible to co-integrate analog and digital circuitry with each light-sensitive pixel. This capability is at the heart of the technological revolution called smart pixel arrays.

The first section relates how the shrinkage of minimum feature size on silicon chips paralleled the reduction of solid-state pixel geometry, as a function of time. This development ceased once the pixel size reached the optical diffraction limit, which marked the birth of the field of smart pixel arrays. In the second section, the basic functionality and the most important physical limitations of conventional pixels and photosensors are given. The third section deals with opto-electronic functionality that is preferred today, of smart pixel arrays; in particular, the capability of spatio-temporal demodulation is very much in demand. In the fourth section, some of the most useful modular building blocks for smart pixel arrays are presented, together with a few key applications. An outlook on photonic systems-on-chip (SoC) is given in the fifth section, with a discussion of single-chip machine vision systems (so-called seeing chips), organic semiconductors for even higher levels of integration of photonic microsystems, and smart pixel arrays that are sensitive to physical parameters other than electromagnetic radiation.

Silicon Technology for Image Sensing

Long before it was realized that silicon would be the material on which modern semiconductor industry and the information technology revolution is based, in 1940 Russell Ohl discovered that silicon is an excellent detector of electromagnetic radiation. In particular, silicon is an excellent detector for light in the visible and near infrared part of the spectrum, for wavelengths between 400 and 1,000 nm. This is very fortunate because it allows designers of image sensors to profit directly from the amazing progress of silicon technology. Since the 1970s, the minimum feature size on silicon chips showed a reliable exponential reduction of about 10% per year, as illustrated in Figure 1. This is a direct consequence of the observation by Intel's Gordon Moore, sometimes called 'Moore's law', that the number of transistors per integrated circuit doubles about every two years. The evolution of minimum feature size is paralleled by the evolution of minimum pixel size of solid-state image sensor, as illustrated in Figure 2. Once the pixel size reached a record low value of 2.4 μm in the year 1996, no further reduction in pixel size seemed desirable, because the optical diffraction limit was essentially reached.

The combination of technological availability of smaller and smaller electronic components and pixel size that demanded no further reduction, inspired researchers to provide each pixel with an increasing amount of electronic functionality, all of which is available simultaneously for massively parallel information processing. This represents the heart of the smart pixel array revolution; pixels that not only sense the incident electromagnetic radiation but that are capable of pre-processing the acquired information in a relevant fashion.

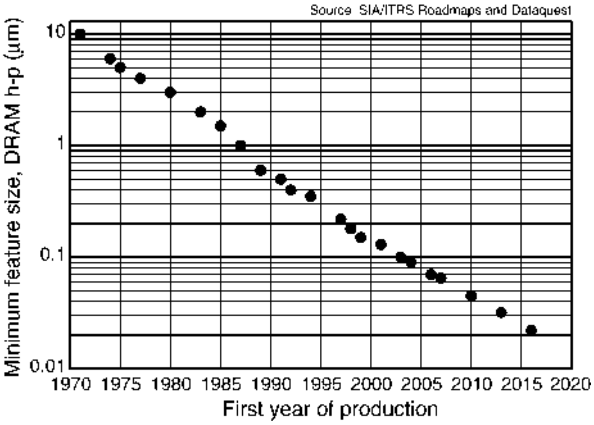

Figure 1 Evolution of minimum feature size (DRAM half-pitch) in state-of-the-art silicon technology, showing an exponential reduction for the past 30 years, which is expected to continue for at least another decade.

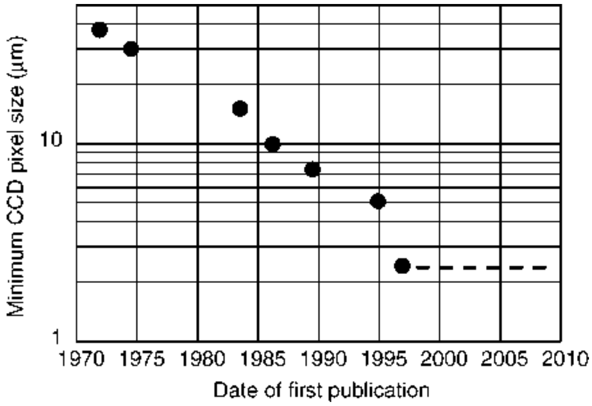

Figure 2 Evolution of minimum pixel pitch in CCD image sensors, tracking the exponential reduction of minimum feature size shown in Figure 1, until a lower limit was reached in 1997, imposed by optical diffraction. Since then, no smaller pixel pitches have been described in the literature; this is indicated graphically with the hatched 'diffraction limit' line.

Basic Functionality and Physical Limitations of Conventional Solid-State Photosensors

The prime task of a pixel is the detection of incident electromagnetic radiation. The detection process consists of six basic steps:

1. The incident photon impinges on the surface of the image sensor, and it has to reach the bulk of the semiconductor.
2. In the bulk of the semiconductor the photon has to interact with the semiconductor, converting its energy into a number of electron–hole pairs.
3. The photogenerated charge carriers must be transported to the sensor surface.
4. At the surface, the photogenerated charge needs to be collected and accumulated in the storage part of the pixel.
5. The accumulated photocharge must be transferred to the input of an electronic circuit capable of detecting this charge.
6. The charge detection circuit converts photocharge into a corresponding voltage, exhibiting a monotonous (but not necessarily linear) transfer function.

Since the interaction length of electromagnetic radiation in silicon is in the tens of micrometers for incident photon energies between 1 eV and 10 keV (corresponding to a wavelength range of about 0.1 to 1,000 nm), silicon is an excellent detector for this large spectral range. This is illustrated for the visible and near infrared spectral range in Figure 3, showing the absolute external quantum efficiency of a pixel realized with an industry standard silicon process. In the visible spectral range, absolute external quantum efficiencies close to 100% can be reached, even with standard silicon processes for the microelectronics industry that were not optimized for photosensing applications.

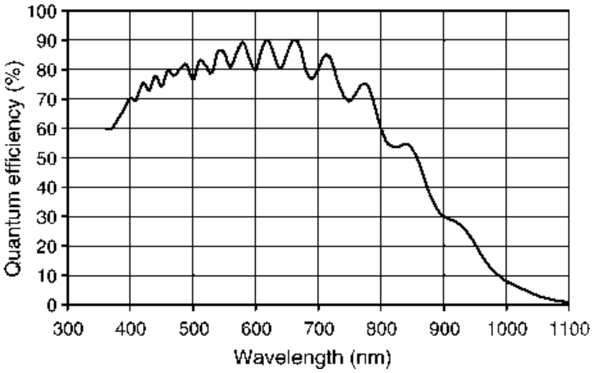

Figure 3 External quantum efficiency of a photosensor realized with an industry standard CMOS process, showing quantum efficiencies above 50% in the whole visible spectral region.

Photocharge transport to the surface of the semiconductor is mainly provided by diffusion, i.e. random thermal motion of the charge carriers that exhibit a velocity of about 10^5 m/s in silicon at room temperature. The mean distance L traveled by a diffusing charge carrier during time t is given by

$$L = \sqrt{Dt} \qquad [1]$$

with the diffusivity D, which has a value of around 40 cm^2/s for electrons in low-doped silicon at room temperature.

Once the photocharge carriers are close to the surface of the semiconductor, they are transported by drift in an electric field E, exhibiting a linear dependence of the transport distance L with time t:

$$L = \mu E t \qquad [2]$$

with the mobility μ which has a value of around 1300 cm^2/Vs for electrons in low-doped silicon at room temperature.

There are two basic principles for creating the electric field with which the photocharge carriers are transported to their storage site within the pixel. A photodiode consists of a semiconductor junction between a p-doped and an n-doped region, as illustrated in Figure 4. Such a photodiode also represents a capacitance on which photocharge can be accumulated and stored. Photodiodes are used as photosensors in one

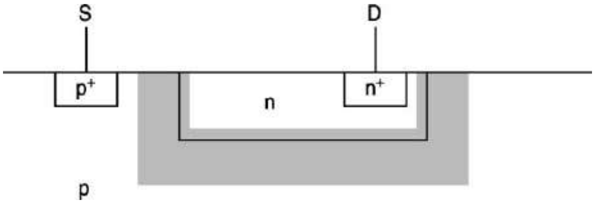

Figure 4 Cross-section of a photodiode, consisting of an n-doped region in a p-doped substrate. The substrate is contacted using a heavily p-doped diffusion S, and the n-region is contacted using a heavily n-doped diffusion D. The space–charge region is illustrated as hatched area at the pn-junction.

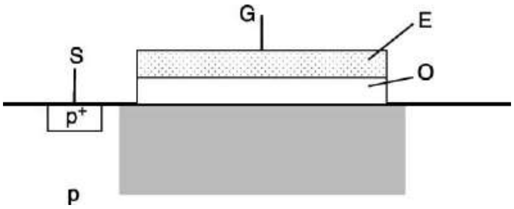

Figure 5 Cross-section of an MOS capacitance, the basic element of a CCD image sensor. On top of a p-substrate a thin transparent insulator O and a transparent electrode E are fabricated. The substrate is contacted with a heavily p-doped diffusion S, and the electrode E is contacted with the gate contact G. The hatched area illustrates the space–charge region under the electrode.

popular type of image sensor, so-called APS (active pixel sensor) or CMOS (complementary metal oxide semiconductor) imagers.

An alternative to a photodiode is an MOS (metal oxide semiconductor) capacitor, consisting of a transparent electrode that covers an insulating layer on top of semiconducting material, as illustrated in Figure 5. An MOS capacitor is the basic element of a charge coupled device (CCD), the other widespread type of image sensor.

The charge transfer between photodiodes and photodetector circuits is based on electrical conduction, occurring essentially with the speed of light. In CCDs, the charge transfer is effectuated with clocks, and excellent charge transfer efficiencies can be observed at clock frequencies of several tens of MHz.

The major limitation in semiconductor photosensors is the last element in the line, i.e. the electronic charge detection circuit. In Figure 6, a so-called source follower is illustrated, which is by far the most frequently used photocharge detection circuit. Its performance is limited by the thermal (Johnson) noise in the channel of the first field effect transistor (FET). This corresponds to a statistical uncertainty with which an amount of photocharge can be measured at the gate of the FET. The input referred charge noise variance s_Q is given approximately by

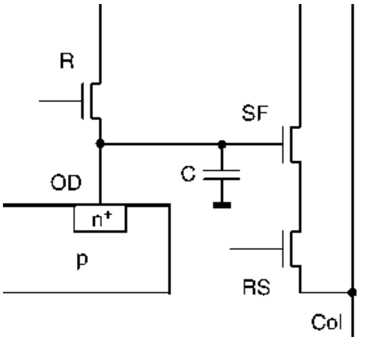

Figure 6 Schematic diagram of a photosensor's output diffusion OD connected to a reset transistor R and a source follower transistor SF, the most widely used charge detection circuit for image sensors and smart pixel arrays. The photocharge is stored on the effective capacitance C at the gate of transistor SF, which is read out using a row select transistor RS, connecting the source follower to the readout column Col.

$$s_Q \cong C^2 \frac{4kTB}{g} \qquad [3]$$

where C is the effective capacitance seen by the gate of the FET, k is Boltzmann's constant, T is the absolute temperature, B is the effective bandwidth at the FET's channel, and g is the transconductance of the FET.

Desired Optoelectronic Functionality in Smart Pixel Arrays

From the previous section it becomes clear that even conventional pixel and photosensors can be improved in several desirable respects:

- The spectral response can be improved in the ultraviolet (50–400 nm), or it can be expanded significantly into the X-ray region (below 0.1 nm) and into the infrared region (above 1,000 nm).
- The speed of the photosensors can be increased by replacing diffusion based photocharge transfer mechanisms with the directed, more efficient drift in electric fields. This is true for the vertical as well as for the lateral charge transport direction.
- The sensitivity of the electronic charge detection process can be improved in two ways. It can be achieved by decreasing the effective detection capacitance C at the input of the charge detection circuit, as can easily be seen from eqn [3]. An alternative is to make use of a physical charge amplification process (i.e. avalanche multiplication in a high electric field), that has low intrinsic noise.

- A derivative improvement that is desired is the increase of the dynamic range D/R of the complete detection process, as defined, for example, by

  $$D/R = 10^{10} \log \frac{Q_{\max}^2}{s_Q} \qquad [4]$$

  where $Q_{\max}$ denotes the maximum (or saturation) charge and s_Q is the total charge noise variance of the electronic detection circuit.

In addition to the above described improvements of conventional pixel functionality, a huge demand exists for the capability to measure the parameters of modulated electromagnetic signals, i.e. to demodulate such signals. This demand is related to the large number of very powerful electronic and opto-electronic measurement principles, capable of improving the sensitivity of the measurement process significantly by employing modulated signals, either in the temporal or in the spatial domain. For this reason, smart pixel arrays are sought that offer this combination of photodetection and demodulation.

Modular Building Blocks for Smart Pixel Arrays

Towards the end of the last century, a large body of literature has been created, addressing various approaches for the realization of one of the above described desired opto-electronic functionalities. The result is a powerful 'toolbox' of modular building blocks for custom smart pixel arrays, with which optimized solutions for optical measurement problems and practical applications can be constructed.

High-Sensitivity Charge Detection

From the noise formula (eqn [3]) two possibilities for the improvement of the sensitivity of image sensors can be deduced: A first approach is the reduction of the capacitance value. An excellent example for this is the invention of the 'double gate' FET, in which the charge to be measured is not placed on a gate electrode at the surface but resides in the semiconductor bulk. In this way the effective capacitance of the pixel can be reduced by at least one order of magnitude compared to conventional pixels. The resulting capacitance of less than 1 fF leads to a readout noise of about one electron at room temperature and for a readout bandwidth of several MHz. Unfortunately the small capacitance allows the storage of only around ten thousand electrons. For such small charge packets, quantum noise is still relevant, leading to significantly degraded image quality, even when the images are taken under favorable lighting conditions.

A second approach to high-sensitivity photosensing exploits the dependence of the charge detection variance on the measurement bandwidth, as indicated in eqn [3]. One method for exploiting this consists of reading out the image sensor at a significantly reduced bandwidth of a few 10 kHz. At the same time, the image sensor is cooled to temperatures of around −100 °C. This becomes necessary because readout at such slow speeds increases the time, to read one complete image, to many minutes, requiring a substantial reduction of the dark current. Instead of working at low readout bandwidths, it is also possible to read out the same information repeatedly and non-destructively at high speed, and to average many measurements of the same amount of photocharge. Both methods have been shown to lead to readout noise levels of about one electron; on the other hand, both approaches require very long times for the acquisition of complete images.

This problem can be resolved by exploiting the massive parallelism that smart pixel arrays offer: low-bandwidth techniques can be employed without losing effective readout speed. Pixels and columns are read out with low-bandwidth amplifiers that are all active simultaneously, and only the final off-chip readout amplifier needs to operate at a large bandwidth. In practice, this approach has been used successfully in single-chip digital cameras with Megapixel resolution, exhibiting a readout noise of only a very few electrons.

A completely different approach is based on the use of a physical amplification mechanism with low intrinsic noise, before detecting the amplified charge packets with a conventional electronic circuit; the avalanche multiplication effect in high electric fields is very well suited to this approach. Each pixel is supplied with its own combination of stabilized avalanche multiplier and conventional charge detection circuit. This approach has been used for the realization of a fully integrated 2D array of single photon detectors that can be fabricated with industry standard CMOS processes. Similar performance is obtained with the Impactron™ series of high-sensitivity CCD image sensors offered by Texas Instruments Inc., making use of a single multistage avalanche multiplier at the output of the image sensor. For both high-sensitivity smart pixel arrays, single-photoelectron performance is reported at room temperature and at readout frequencies of several MHz, as required for video applications.

Photosensing with Improved Dynamic Range

Conventional solid-state image sensors and cameras usually have a very limited dynamic range, as defined in eqn [4]. Typical values are between 40 and 80 dB, rarely exceeding four orders of magnitude. In natural scenes, however, much higher *D/R* values can be encountered. To cope with this, the human eye is capable of spanning a dynamic range of more than 150 dB. This realization has spurred the invention of several methods with which the *D/R* of smart pixel arrays can be increased.

One approach consists of acquiring several images of the same scene with different exposure times, with the subsequent combination into one high-dynamic-range composite picture. *D/R* values exceeding 120 dB have been obtained with such multi-exposure methods. Fast-moving objects, however, will have relocated from one exposure to the next, resulting in false, blurred representations of these objects in the composite picture.

This can be overcome using nonlinear response characteristics similar to the ones found in natural vision. Impinging light produces a photocurrent that is highly linear with the incident light intensity. Because the voltage across the photodiode is logarithmically dependent on the photocurrent, a logarithmic voltage-illumination response is obtained, with a *D/R* exceeding 120 dB. Such logarithmic pixels are well-suited for applications providing high illumination levels but they suffer from excessive noise and slow response speed at low light levels.

This disadvantage led to the development of linear-logarithmic pixels, exhibiting a high-sensitivity linear response at low light levels, and a gradual changeover to logarithmic response once a certain, programmable light level is reached. The design principle is illustrated in Figure 7. An otherwise conventional CMOS/APS pixel is supplied with a device that exhibits logarithmic voltage–current response, such as a diode. As long as the cathode is at a higher voltage than the anode, no current flows across the diode, and the pixel response is linear. Once the photocurrent has lowered the cathode voltage below the anode voltage, the pixel response becomes logarithmic, as described above. The *D/R* of such a linear-logarithmic pixel reaches unprecedented high values; almost 200 dB have already been obtained in practice. Because of the simplicity and robustness of this principle, such linear-logarithmic smart pixel arrays have found their way into commercial sensor and camera products that are available commercially, as illustrated in Figure 8. A circuit board and a light bulb are imaged in linear mode (Figure 8a), showing highly saturated areas, and in linear-logarithmic mode (Figure 8b), in which even the light bulb's filament can clearly be seen.

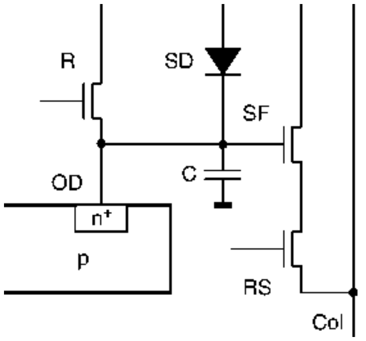

Figure 7 Schematic diagram of a high-dynamic-range pixel, consisting of a conventional APS/CMOS pixel, as illustrated in Figure 6, in combination with a shunt diode SD. The photocurrent lowers the voltage at the gate of transistor SF continuously. Below a certain value of this voltage, the diode starts to become forward biased, leading to a linear-logarithmic illumination response, with a dynamic range that can be as high as 200 dB.

Extension of the Spectral Sensitivity Range

The spectral sensitivity of silicon is essentially restricted to the spectral wavelength range between 0.1 and 1,100 nm. With a suitable choice of materials that cover the silicon sensor's surface, this sensitive wavelength range can be extended. For high-energy radiation, scintillator materials are employed, capable of converting incident high-energy photons into a proportional number of visible photons. Scintillators render smart image sensors particularly sensitive in the UV (300–400 nm), deep UV (below 300 nm), X-ray (0.002–1 nm), and gamma ray (below 0.002 nm) spectral region. Typically, the scintillator materials are fabricated as platelets that are glued on the surface of the produced CMOS/CCD image sensor.

The extension of the sensitive spectral range to the infrared can be achieved with semiconductor materials with smaller energy bandgaps than silicon. Examples include Ge, SiGe, SiC, PtSi, IrSi, InAs, InGaAs, or InGaAsP. Typically, these materials are deposited on the produced CMOS/CCD image sensor in an ultra-high vacuum environment. An alternative is the use of bump-bonding, with which the CMOS/CCD image sensor can be connected to a small-bandgap material chip at each pixel site. With these technologies, silicon-based smart pixel arrays become sensitive to wavelengths of several microns.

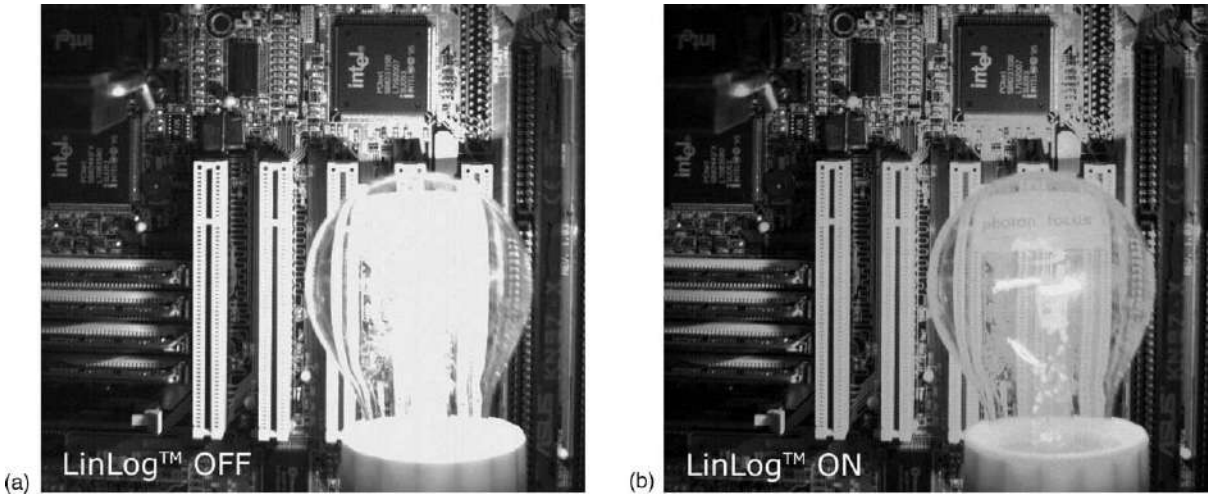

Figure 8 Images of a scene with high dynamic range, taken with an image sensor exhibiting linear illumination response (a) and linear-logarithmic illumination response (b). A similar pixel architecture as illustrated in Figure 7 is employed. (Courtesy of Photonfocus AG, Lachen, Switzerland.)

High-Speed Image Sensing

Increasing the speed of the photosensing process necessitates improvements in two key areas. Photocharges need to be collected and transported faster, and the acquired information must be read out faster.

A comparison of eqns [1] and [2] immediately reveals that the replacement of diffusion by drift transport provides the desired speed improvement. In the vertical direction, the electric field of photodiodes or MOS capacitances reaches more deeply into the semiconductor bulk if lower doping levels of the substrate are used or the bias voltage is increased. Improving the lateral transport speed is more difficult. A highly successful approach in so-called drift-field pixels is illustrated in Figure 9. On top of a thin (a few tens of nanometers) transparent insulator, a highly resistive layer – often of polysilicon – is placed. When a voltage difference is applied at the two ends, a small current flows across the resistive layer, inducing a linear potential variation at the semiconductor surface. A simple calculation shows that very small transit times for photocharges can be obtained in this way. Assuming a (quite large) pixel size of 25 μm and a voltage difference of 5 V, the maximum transit time of photoelectrons that have to move from one end of the pixel to the other, is less than 1 ns. This is more than two orders of magnitude faster than can be achieved with diffusion transport.

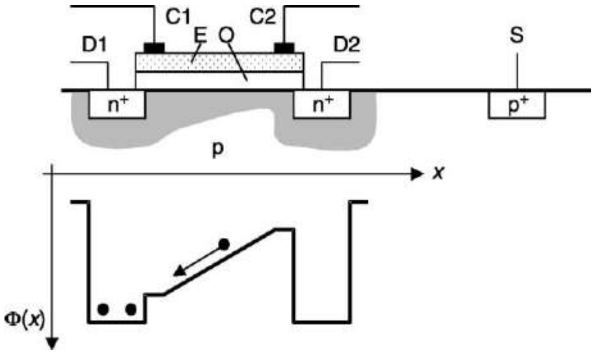

Figure 9 Cross-section of a drift field pixel, exploiting the lateral drift field at the semiconductor–oxide interface. The corresponding sloped potential distribution $\Phi(x)$ is created with a highly resistive transparent electrode E on top of a thin transparent insulator O; this conductor has two contacts C1 and C2 at its ends, to which a voltage difference is applied. Depending on the sign of the slope, the photoelectrons are collected either in diffusion D1 or D2. The hatched area indicates the space-charge region.

Smart pixel arrays provide a simple solution to the problem of speeding up the readout of the acquired picture information; a large degree of parallelism. This is illustrated with the architecture of a split-field high-speed image sensor, shown schematically in Figure 10. The photosensitive area is split into two parts, so that the column amplifiers of one half have time to settle to their proper voltages, while the columns of the other half are being read out. No time is lost, therefore, when another row of the image is selected for readout. A large number of output amplifiers are provided for highly parallel readout of the image data, each with a typical readout rate of 40–100 MHz. In this way, sustainable pixel rates of several GHz can be achieved in practice.

Static Spatial Demodulation

The spatially varying geometry of the photosensitive part of a pixel can be used for the static spatial

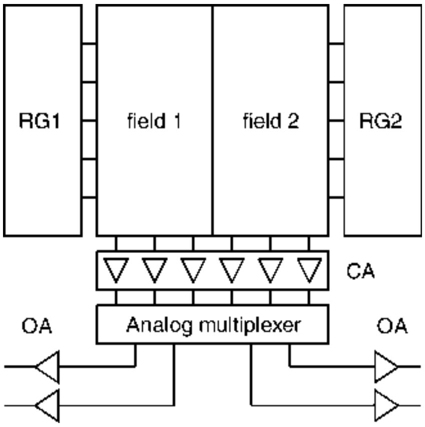

Figure 10 Block diagram of a high-speed image sensor, making use of the large amount of parallelism possible in smart pixel arrays. The light-sensitive area is split into two fields whose rows can be addressed independently with two separate row address generators RG. In this way, one row of one field can be selected and the corresponding column amplifiers CA can reach equilibrium, while the columns of the other field are being read out using several parallel output amplifiers OA.

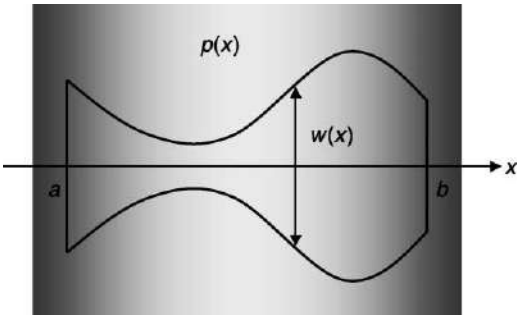

Figure 11 Illustration of a one-dimensional intensity distribution $p(x)$ of light which is incident on a shaped photodiode with position-dependent width $w(x)$, between the two boundaries a and b.

demodulation of an impinging light field. When a one-dimensional intensity distribution $p(x)$ is incident on a photodiode with spatially varying width $w(x)$ between the limits a and b, as illustrated in Figure 11, an output signal r is produced:

$$r = \int_a^b p(x)w(x)\mathrm{d}x \qquad [5]$$

By shaping the photodiode appropriately, any type of parametrized linear transform can be realized, as illustrated by the photomicrograph in Figure 12. In this practical example, four photodiodes are shaped sinusoidally and cosinusoidally, weighted with a slowly varying windowing function. Their signals are used for determining the Fourier coefficients (the sine and cosine transform) of an incident light distribution for a fixed spatial frequency. This is employed, for example, in cost-effective optical encoders, capable of measuring linear displacement with an accuracy below 50 nm.

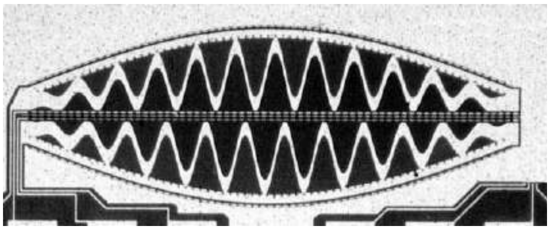

Figure 12 Photomicrograph of two photodiode pairs that are shaped as spatial sine and cosine functions, weighted with a slowly varying windowing function. The four photosignals can be employed to determine the phase and amplitude of a one-dimensional intensity signal that has the same spatial frequency as the sine and cosine. In practice, the accuracy of the phase determination is as high as 0.1% of the spatial wavelength. The width of the photomicrograph corresponds to about 1 mm in reality.

Dynamic Spatial Convolution

Many image-processing tasks require the convolution of the pixel data $P_{i,j}$ at the position (i, j) in an image with a two-dimensional convolution kernel $w_{k,l}$:

$$R_{k,l} = \sum_{i=-N}^{N} \sum_{j=-N}^{N} P_{i,j} w_{k-i,l-j} \qquad [6]$$

Instead of employing a purely geometrical solution, as described in the previous section, an alternate approach is found when interpreting the convolution eqn [6] in the following way. The resulting image R is obtained by displacing the original image P laterally by an amount (k,l), by multiplying the shifted pixel values with the corresponding weight $w_{k,l}$, and by adding up all these products.

The lateral displacement is realized with a CCD structure, implementing the 2D lateral movement of the acquired image charge packets, the multiplications with the convolution kernel values are realized by using various exposure times for each lateral shift position, proportional to the corresponding weight value, and the summation is just the natural charge accumulation of CCD structures. The problem of negative weights is solved by separately determining the positive and negative parts of the weight function, each stored separately in the pixel, and by calculating the difference of the two values during readout, either on-chip or off-chip. In this way, such convolution image sensors can be used for the determination of complete, programmable, real-time image convolutions, with arbitrary kernels that are fully under the dynamic control of digital logic or microprocessors. Practical experience with such convolution imagers shows that the convolution results exhibit inaccuracies of less than 1%.

Temporal Demodulation

In many relevant optical measurement techniques, temporally modulated light fields are employed. For this reason, smart pixel arrays are desired, providing synchronous demodulation of these signals. An electromagnetic signal $S(t)$ that is harmonic as a function of time with given frequency ω, having unknown modulation amplitude A, offset B and phase delay φ, is described by

$$S(t) = A \sin(\omega t + \varphi) + B \qquad [7]$$

The three unknowns A, B and φ can be measured if three or more sampling points are sampled synchronously per period in so-called ‘lock-in’ pixels. An example of a lock-in pixel employing the drift-field transport principle with four accumulation taps, is illustrated in Figure 13. Demodulation frequencies of several tens of MHz have already been demonstrated in practice.

The most important application of such synchronous demodulation pixels is in optical time-of-flight range cameras. Such a camera incorporates a light source, usually an array of infrared or visible LEDs, that is modulated at a few tens of MHz. The modulated light is reflected diffusely by the various objects, and the camera’s lens images it onto a lock-in image sensor. Each of its pixels carries out synchronous sampling/accumulation of photocharges. At the end of the exposure time, the accumulated photocharge samples are read out, and the three modulation parameters of the reflected light are determined for each pixel. The phase shift is a direct measure of the local distance. It has been shown that for a wide range of operating conditions, the distance resolution of such optical range cameras is already limited just by the photo shot noise. A typical result of such a depth image is shown in Figure 14.

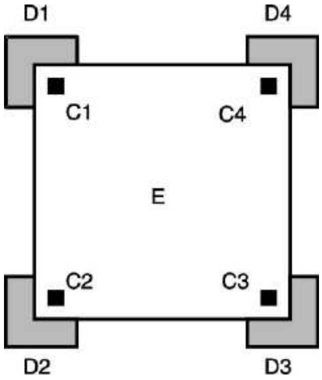

Figure 13 Schematic top view of a lock-in pixel based on the lateral drift-field approach illustrated in Figure 9. A square electrode is fabricated with a highly resistive transparent material on top of a thin insulator. The electrode contacts C1 to C4 are employed to create a ‘reeling’ potential distribution that rotates with the modulation frequency of the incident light that requires synchronous demodulation, so that the photocharges are collected at diffusions D1 to D4, respectively, at the appropriate time during each period.

Photonic Systems-on-chip

The microelectronics revolution is based on the key notion of integration. More functionality integrated monolithically on the same chip leads to faster, smaller, better performing, and cheaper products. The same is true for smart pixel arrays, for which progress through integration is attained in several respects.

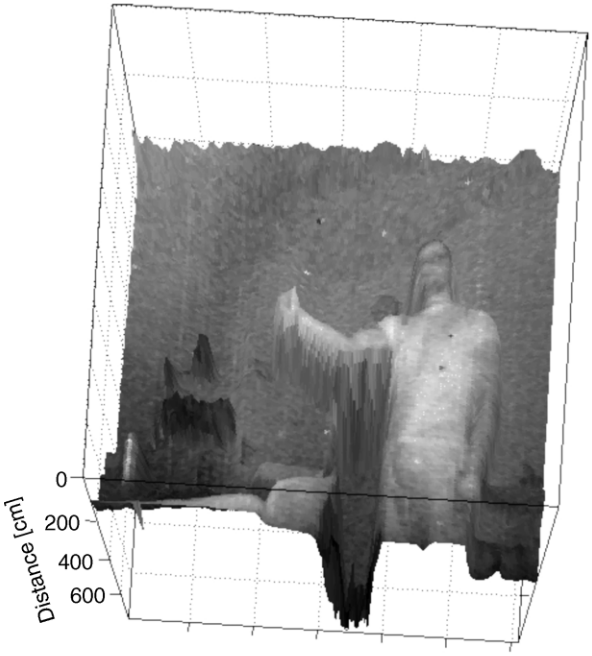

Figure 14 Example of a distance picture (a person) out of a 3D range image sequence, taken with an optical time-of-flight distance camera making use of lock-in pixels. Depending on the available light level, the background illumination and the modulation frequency, such solid-state range cameras can have a resolution of a few mm.

Seeing Chips

Already, in the mid-1980s, researchers were excited about the possibility of integrating local signal processing capabilities in each pixel, implementing some functions of biological vision systems. It was envisioned that an increasing amount of bio-inspired functionality would lead to smart image sensors, exhibiting a primitive sense of vision. The idea of a 'seeing chip', with the capability to interpret a scene, was born.

Unfortunately, even today, our understanding of the enormously successful biological vision systems has not yet reached a mature enough state allowing us to implement signal processing functions that approximate any 'image understanding'. Published algorithms are still incapable of interpreting a scene and 'perceiving' its contents. Obviously, no set of vision algorithms is yet known with which seeing chips could be realized. Although progress is slower than envisioned, single-chip machine vision systems of growing complexity and increasing functionality are being developed.

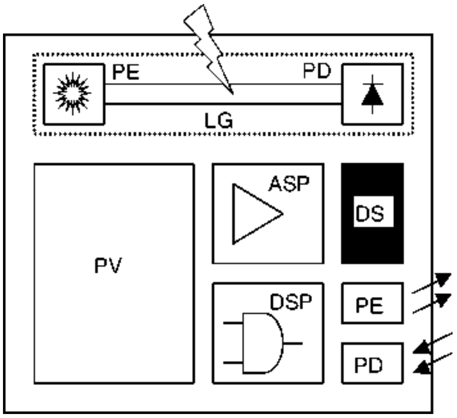

Figure 15 Schematic block diagram of a monolithically integrated photonic microsystem, that is expected to be producable in the near future with organic semiconductors. All optoelectronic devices are fabricated on the same planar surface: photovoltaic power supplies (PV); analog and digital signal processing (ASP and DSP); photodetectors (PD); photoemitters (PE) – used for bidirectional communication or as parts of a photonic detector subsystem – a light guiding structure (LG) that can be sensitized for specific sensing applications, and a display subsystem (DS).

Organic Semiconductors for Monolithic Photonic Microsystems

The realization that organic materials with conjugated double bonds, in particular noncrystalline polymers, are direct bandgap semiconductors, has recently led to a flurry of R&D activities in the domain of organic photonics. Organic light-emitting diodes (OLEDs), color displays, lighting panels, photodiodes, image sensors, photovoltaic cells, RF transponders, and tags, and several other organic photonic devices have not only been demonstrated but many are already commercially available. The special appeal of organic semiconductors stems from their potential for monolithic, low-cost integration of complete, large-area photonic microsystems, even on thin, flexible substrates. In Figure 15, such a monolithic photonic microsystem is illustrated schematically, combining a photovoltaic power supply, a photonic detector subsystem, analog and digital signal processing, as well as bidirectional optical communications with the outside world, possibly including also an on-chip display. The photonic detector subsystem typically consists of a light source, a light guiding structure that is sensitized – for example chemically – and a photodetector/image sensor subsection. In such a monolithic photonic microsystem, the term 'smart pixel array' obtains a much broader meaning, opening up a huge number of practical applications in essentially all domains of our daily and professional lives. This domain of organic semiconductors for integrated photonic microsystems is still largely unexplored, and it is offering exciting opportunities for research and product development.

Smart Pixel Arrays Beyond Electromagnetic Field Imaging

The sensing capabilities of smart pixel arrays are not restricted to electromagnetic radiation; this technology can also be used for the acquisition of other two- or three-dimensional physical fields. As an example, electrostatic imaging can be used for the temporally and spatially resolved acquisition of surface potential distributions in living cells. Electrostatic potential values of several hundred mV can be acquired with a precision of 1 mV and a temporal resolution below 1 millisecond. Other acquisition modalities include magnetostatic field imaging, for example by employing the Hall effect, or imaging of local electrostatic charge or current density distributions.

Summary

The relentless advances of semiconductor technology make it possible to integrate an increasing amount of photosensitive, analog, and digital functionality in the same smart pixel and on the same smart pixel array. Unusual photosensor devices can be fabricated with commercially available semiconductor processes, implementing innovative designs and layouts, as well as novel electronic circuits, with which the physical limits of photosensing and photocharge processing may be reached. Novel opto-electronic microsystems can be developed, based on application-specific smart pixel arrays, either by realizing known optical measurement techniques more efficiently or by implementing new techniques, whose realization requires the availability of nonconventional image sensing functionality.

Already, smart pixel arrays are opening up a large number of new opportunities in photonics research and product development. The continuously increasing levels of integration lead to new forms of system-on-chip solutions for a wide range of practical applications. Smart pixel arrays can perform at the basic limits imposed by physics, and they have become indispensable tools in almost any type of research laboratory and any field of science and engineering.

List of Units and Nomenclature

Bandwidth	B	s^{-1}
Boltzmann's constant	k	J/K
Capacitance	C	F
Charge	Q	C
Charge carrier mobility	μ	$m^2/(V\ s)$
Charge noise variance	s_Q	C^2
Diffusivity	D	m^2/s
Distance	L	m
Dynamic range	D/R	dB
Field effect transistor transconductance	g	A/V
Integrated power measured with shaped photodiode	r	W
Intensity of electromagnetic wave	S, A, B	W/m^2
One-dimensional light intensity distribution	$p(x)$	W/m^2
One-dimensional width distribution	$w(x)$	m
Phase delay	φ	–
Pixel value at position (i,j) in an image	$P_{i,j}; R_{i,j}$	–
Temperature	T	K
Temporal frequency	ω	s^{-1}
Time	t	s
Two-dimensional convolution kernel	$w_{k,l}$	–

See also

Electromagnetically Induced Transparency. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides. **Semiconductor Materials:** Group IV Semiconductors, Si/SiGe.

Further Reading

Forrest SR (2000) Active optoelectronics using thin-film organic semiconductors. *IEEE Journal on Selected Topics in Quantum Electronics* 6: 1072–1083.

Fossum ER (1997) CMOS image sensors: electronic camera-on-a-chip. *IEEE Transactions on Electron Devices* 44: 1689–1698.

Hosticka BJ, Seitz P and Simoni A (2004) Optical time-of-flight range imaging for solid-state 3D vision. In: Grimes G and Dickey EC (eds) *Encyclopedia of Sensors*. Stevenson Ranch, CA: American Scientific Publishers (in print).

Koch C (1989) Seeing chips: analog VLSI circuits for computer vision. *Neural Computation* 1: 184–200.

Lutz G (1999) *Semiconductor Radiation Detectors*. Berlin: Springer.

Mead C (1989) *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley and Sons.

Seitz P (1999) Solid-state image sensing. In: Jaehne B (ed.) *Handbook of Computer Vision and Applications*. London: Academic Press.

Seitz P and Blanc N (2003) CCD and APS/CMOS technology for smart pixels and image sensors. *Proceedings of SPIE* 5251 (in print).

Sze SM (1994) *Semiconductor Sensors*. New York: John Wiley and Sons.

Theuwissen AJP (1995) *Solid State Imaging with Charge Coupled Devices*. Dordrecht, Germany: Kluwer.

DIFFRACTION

Contents

Diffraction Gratings
Fraunhofer Diffraction
Fresnel Diffraction

Diffraction Gratings

J Turunen and T Vallius, University of Joensuu, Joensuu, Finland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The classic textbook example of a diffraction grating is a periodic arrangement of parallel opaque wires: if a plane wave is incident upon such a structure, it is divided into a multitude of plane waves propagating in well-defined directions. In particular, the dramatic wavelength-dependence of these directions has attracted great interest ever since the potential of diffraction gratings in spectroscopy was realized in the late 1700s. Today spectroscopy is only one, though important, application of diffraction gratings. Gratings of many different forms belong to the basic building blocks in modern wave-optics-based design of optical elements and systems.

Grating Equations

Some examples of the wide variety of possible grating structures are illustrated in Figure 1, which shows the profiles of some commonly encountered diffraction gratings that are invariant in the y direction. These are known as linear gratings. Each grating may be of reflection type (the refractive index n_{III} is complex) or of transmission type (n_{III} is real). In all cases n_1 and n_2 can both be real or one of them can be complex. Here we assume that the incident field is a plane wave with wave vector $\boldsymbol{k}$ in the x–z plane; more general illumination waves can be represented as superpositions of plane waves.

Figure 1a illustrates a binary grating: in the modulated region $0 < z < H$ the refractive index alternates between n_1 and n_2, changing from one value to another at certain transition points within the period d. The structure in Figure 1b represents a sinusoidal surface-relief grating as an example of smooth grating profiles. Finally, Figure 1c shows a triangular grating profile.

The periodicity of a grating, illuminated by a plane wave, usually leads to the appearance of a set of reflected and transmitted plane waves (diffraction orders of the grating) that propagate in directions determined by the grating period d, the illumination wavelength λ, and the refractive indices n_{I} and n_{III} (see Figure 2a). These directions are obtained from the transmission grating equation:

$$n_{\mathrm{III}} \sin \theta_m = n_{\mathrm{I}} \sin \theta + m\lambda/d \qquad [1]$$

where m is the index of the diffraction order and θ_m is its propagation angle. For reflected orders (superscript r) one simply replaces n_{III} by n_{I} in eqn [1].

Obviously only a limited number of orders satisfy the condition $|\theta_m| < \pi/2$ and thereby represent conventional propagating plane waves. Higher, so-called evanescent orders are inhomogeneous waves that propagate along the grating surface and decay exponentially in the z direction. However, they cannot be ignored in grating analysis, as we shall see.

In the case of a biperiodic grating with a period $d_x \times d_y$ the propagation direction of the incident beam is defined by two angles θ and ϕ illustrated in Figure 2b. The grating equations now read as:

$$n_{\mathrm{III}} \sin \theta_{mn} \cos \phi_{mn} = n_{\mathrm{I}} \sin \theta \cos \phi + m\, \lambda/d_x \qquad [2]$$

$$n_{\mathrm{III}} \sin \theta_{mn} \sin \phi_{mn} = n_{\mathrm{I}} \sin \theta \sin \phi + n\lambda/d_y, \qquad [3]$$

where θ_{mn} and ϕ_{mn} define the propagation directions of the diffraction order with indices m and n. For reflected orders n_{III} is again replaced by n_{I}. These expressions apply also to the special case of linear gratings (limit $d_y \to \infty$) illuminated by a plane

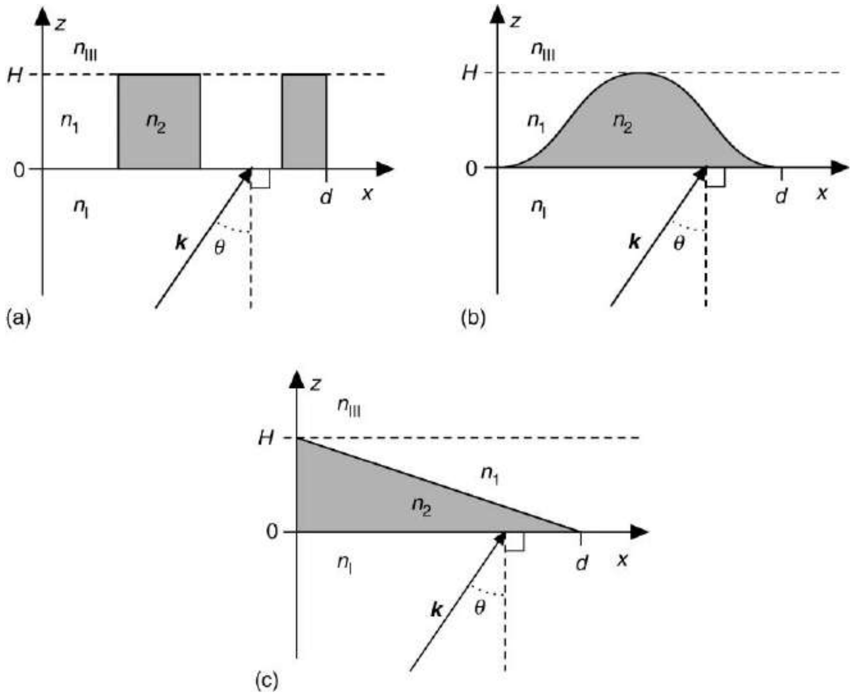

Figure 1 Examples of linear gratings. (a) Binary grating; (b) sinusoidal grating; (c) triangular grating. Here *d* is the grating period, *H* is the thickness of the modulated region, θ is the angle of incidence, and *n* denotes refractive index.

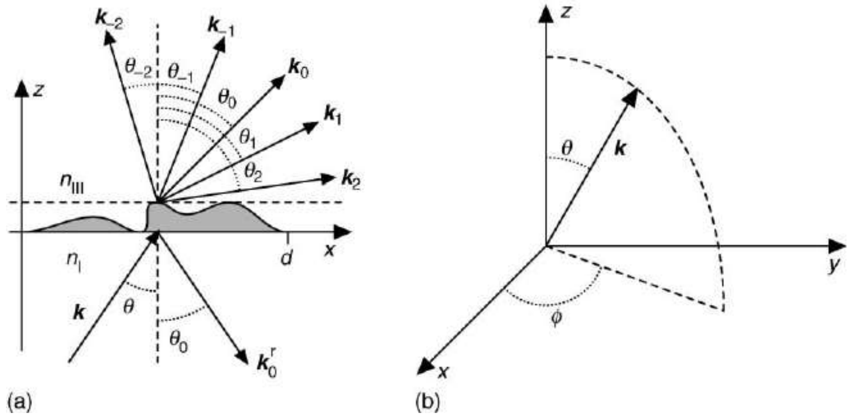

Figure 2 Diffraction by gratings. (a) Linear grating; (b) definitions of direction angles in the case of a biperiodic grating.

wave not propagating in the x–z plane (conical incidence).

Some two-dimensionally periodic gratings are illustrated in Figure 3. The structure illustrated in Figure 3a is an extension of a binary grating to the biperiodic geometry: here a single two-dimensional feature is placed within the grating period but, of course, more features with different outlines could be included as well. Moreover, each period could contain a continuous 'mountain-range' or consist of a 'New-York'-style pixel structure. Figure 3b illustrates a structure consisting of an array of rectangular-shaped holes pierced in a flat screen, while Figure 3c defines an array of isolated particles on a flat substrate. If the grid or the particles are metallic, surrounded by dielectrics from all sides, one speaks of inductive and capacitive structures, respectively: in the former case relatively free-flowing currents are induced in the metallic grid structure, but in the latter case the isolated metal particles become polarized (and therefore capacitive) when illuminated by an electromagnetic field.

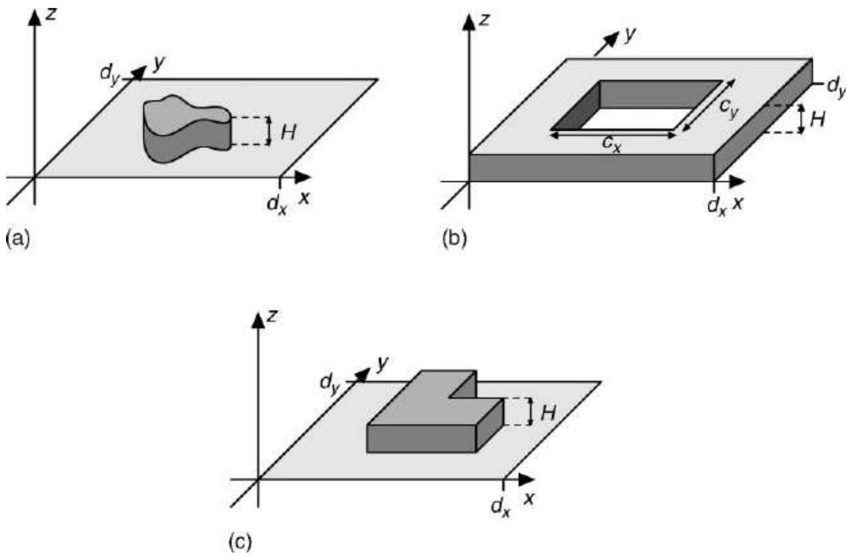

Figure 3 Examples of biperiodic gratings. (a) General binary grating; (b) inductive structure; (c) capacitive structure.

Grating Theory

Gratings equations provide exact knowledge of the propagation directions of the diffraction orders, but they give no information about the distribution of light among different orders. To obtain such information, one must in general solve Maxwell's equations exactly (rigorous diffraction theory), but within certain limits simpler methods are available. In all cases, however, the transmitted and reflected fields are represented as superpositions of plane waves that propagate in the directions allowed by the grating equations (these are known as Rayleigh expansions).

The most popular method for rigorous grating analysis is the Fourier Modal Method (FMM), though many alternative analysis methods exist. To understand the principle of FMM, consider first binary gratings. The field inside the modulated region $0 < z < H$ is represented as a superposition of waveguide modes, determined by imagining that the modulated region would extend from $x = -\infty$ to $x = \infty$. With given transition points the electric permittivity, its inverse, and the electric field inside the grating, can be represented in the form of Fourier series and one ends up with a matrix eigenvalue problem with finite dimensions when the matrices are truncated. A numerical solution of this problem provides a set of orthogonal modes that can propagate in the structure with certain propagation constants. The relative weights of the modes are then determined by applying the electromagnetic boundary conditions at the planes $z = 0$ and $z = H$ to match the modal expansions inside the grating with the plane-wave (Rayleigh) expansions of the fields in the homogeneous materials on both sides of the grating. This matching procedure, which is numerically implemented by solving a set of simultaneous linear equations, yields the complex amplitudes of the electric and magnetic field components associated with the reflected and transmitted diffraction orders of the grating.

Of main interest are the diffraction efficiencies, η_m (or η_{mn} for biperiodic gratings) of the orders, which tell the distribution of energy between different propagating orders. In electromagnetic theory the diffraction efficiency is defined as the z-component of the time-averaged Poynting vector associated with the particular order, divided by that of the incident plane wave.

Gratings with z-dependent surface profiles can be handled similarly if they are 'sliced' into a sufficient number of layers, in each of which the grating is modeled as a binary structure. The electromagnetic boundary conditions are applied at each slice boundary and the approach models the real continuous structure with arbitrary degree of accuracy when the number of the slices is increased; in practice it is usually sufficient to use approximately 20–30 slices for sufficiently accurate modeling of any continuous grating profile by FMM.

The main problem with rigorous techniques such as FMM is bad scaling. If, in the case of 1D modulated (or y-invariant) gratings, the period is increased by a factor of two, the computation time required to solve the eigenvalue problem is increased by a factor of around eight. Thus we have a d^3 type scaling law. The situation is much worse for biperiodic gratings, for which the exponent is around six. Thus no foreseeable developments in computer technology will solve the problem, and approximate analysis methods have to be sought. A multitude of such methods are under active development, but in this context we discuss the simplest example only.

If $d \gg \lambda$ and the grating profile does not contain wavelength-scale features, one can describe the effects of a grating on the incident field merely by considering it as a complex-amplitude-modulating object, which imposes a position-dependent phase shift and amplitude transmission (or reflection) coefficient on the incident field. This model is known as the thin-element approximation (TEA).

At normal incidence the plane-wave expansion of any scalar field component U immediately behind the grating (at $z = H$) takes the form:

$$U(x, y, H) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{mn} \times \exp[\mathrm{i}2\pi(mx/d_x + ny/d_y)] \qquad [4]$$

where

$$T_{mn} = \frac{1}{d_x d_y} \int_0^{d_x} \int_0^{d_y} U(x, y, H) \times \exp[-\mathrm{i}2\pi(mx/d_x + ny/d_y)]\mathrm{d}x\,\mathrm{d}y \qquad [5]$$

are the complex amplitudes of the transmitted diffraction orders. In TEA the field at $z = H$ may be connected to the field at $z = 0$ (a unit-amplitude plane wave is assumed here) by optical path integration along a straight line though the modulated region of the grating:

$$U(x, y, H) = \int_0^H \exp\left[\mathrm{i}\frac{2\pi}{\lambda}\hat{n}(x, y, z)\right]\mathrm{d}z \qquad [6]$$

where $\hat{n}(x, y, z) = n(x, y, z) + \mathrm{i}\kappa(x, y, z)$ is the complex refractive-index distribution in the region $0 < z < H$. With TEA the diffraction efficiencies can be calculated from a simple formula:

$$\eta_{mn} = |T_{mn}|^2 \qquad [7]$$

which, however, ignores Fresnel reflections.

Equations [5]–[7] allow one to obtain closed-form expressions for the diffraction efficiencies η_{mn} in many grating geometries. For example, a y-invariant triangular profile with $H = \lambda/|n_\mathrm{I} - n_\mathrm{III}|$ gives $\eta_{-1} = 100\%$ if $n_1 = n_\mathrm{III}$, $n_2 = n_\mathrm{I}$, and $\theta = 0$. For a Q-level stair-step-quantized version of this grating we obtain:

$$\eta_m = \sin c^2(1/Q) \qquad [8]$$

where $\sin cx = \sin(\pi x)/(\pi x)$. It should be noted, however, that these results are valid only for gratings with $d \gg \lambda$, even if they are corrected by Fresnel transmission losses at the interface between the two media.

Fabrication of Gratings

Gratings can be manufactured in a number of different ways, perhaps the simplest being to plot an equidistant array of straight black and white lines and to reduce the scale of the pattern photographically. This method can be readily adapted to the generation of more complex grating structures and diffractive elements, but it is not suitable for the production of high-quality gratings required in, for example, spectroscopy.

Ruling engines represent the traditional way of making gratings with periods of the order of the wavelength of light, especially for spectroscopic applications. These are machines equipped with a triangularly shaped diamond ruling edge and an interferometrically controlled two-dimensional mechanical translation stage that moves the substrate with respect to the ruling edge. Triangular-profile gratings of high quality can be produced with such machines, but the fabrication process is slow and the cost of the manufactured gratings is therefore high. In addition, even with real-time interferometric control of the ruling-edge position, it is impossible to avoid small local errors in the grating geometry. Such errors are particularly harmful if they are periodic since, according to the grating equation, these errors will generate spurious spectral lines known as ghosts. Such lines generated by the grating ‘super-period’ are weak but they may be confused with real spectral lines in the source spectrum. If the ruling errors are random, they result in ‘pedestal-type’ background noise. While such errors do not introduce ghosts, they nevertheless reduce the visibility of weak original spectral lines.

The introduction of lasers in the early 1960s made it possible to form stable and high-contrast interference patterns over a sufficient time-scale to allow the exposure of photographic films and other photosensitive materials such as polymeric photoresists. This interferometric technique offers a direct method for the fabrication of large diffraction gratings without appreciable periodic errors. With this technique it is possible to make highly regular gratings for spectroscopy, but it does not permit easy generation of arbitrarily shaped grating structures. The two intersecting plane waves can also be generated with a linear grating using diffraction orders $m = +1$ and $m = -1$, suppressing the zero order by appropriate design of the grating profile and choosing the period in such a way that all higher transmitted orders are evanescent. This so-called phase mask technique is suitable, for example, for the exposure of Bragg gratings in fibers.

To generate arbitrary gratings one usually employs lithographic techniques. Some of the possible processes are illustrated in Figure 4. The exposure can be performed with scanning photon, electron, or ion beams, etc. The process starts with substrate preparation, i.e., coating it with a beam-sensitive layer known as resist, and in the case of electron beam exposure with a thin conductive layer to prevent charging. Following the exposure the resist is developed chemically, and in this process either the exposed or nonexposed parts of the resist are dissolved (depending on whether the resist is of positive or negative type), leading to a surface-relief profile. Several alternative fabrication steps may follow the development, depending on the desired grating type. If a high-contrast resist is used, a slightly 'undercut' profile as illustrated in Figure 4c is obtained. This can be coated with a metal layer and the resist can be 'lifted off' chemically. Thus a binary metallic grating is formed. To produce a binary dielectric profile one bombards the substrate with ions either purely mechanically (ion beam milling) or with chemical assistance (reactive ion etching), and finally the residual metal is dissolved.

Alternatively, the particle beam dose can be varied with position to obtain an analog profile in the resist after development. This resist profile can be transformed into the substrate by proportional etching, which leads to an analog dielectric surface-relief structure (see Figure 4f). This profile can be made

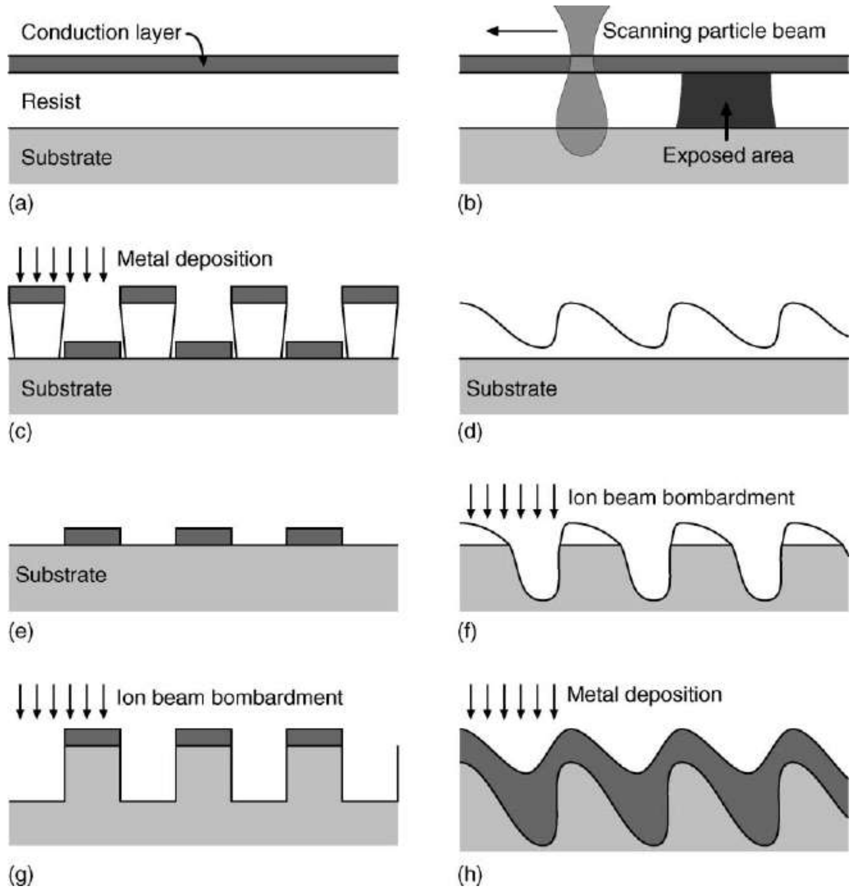

Figure 4 Lithographic processes for grating fabrication. (a) Substrate after preparation for electron beam exposure; (b) exposure by a scanning electron beam; (c) slightly undercut profile in hight-contrast resist after development, under metal film evaporation; (d) analog resist profile after variable-dose exposure and resist development; (e) finished binary metallic grating after resist lift-off; (f) transfer of the analog surface-relief profile into the substrate by proportional etching; (g) etching of a binary profile into the substrate with a metal mask; (h) growth of a metal film on an analog surface profile.

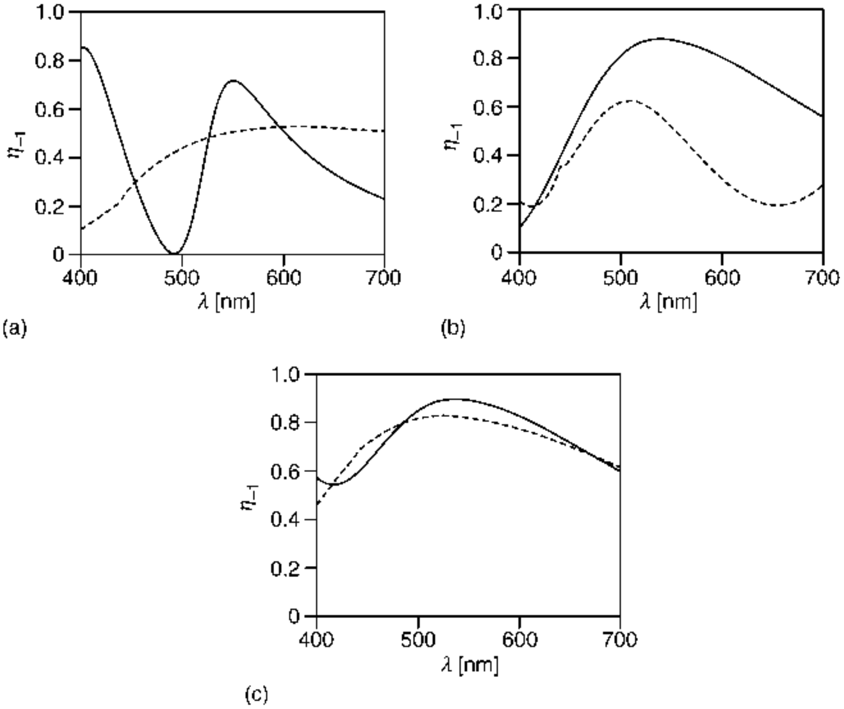

Figure 5 Spectral diffraction efficiency curves of spectroscopic gratings in TE polarization. (a) Binary; (b) sinusoidal; and (c) triangular reflection gratings illuminated at Bragg angle. Solid lines: $d = \lambda_0$. Dashed lines: $d = 2\lambda_0$.

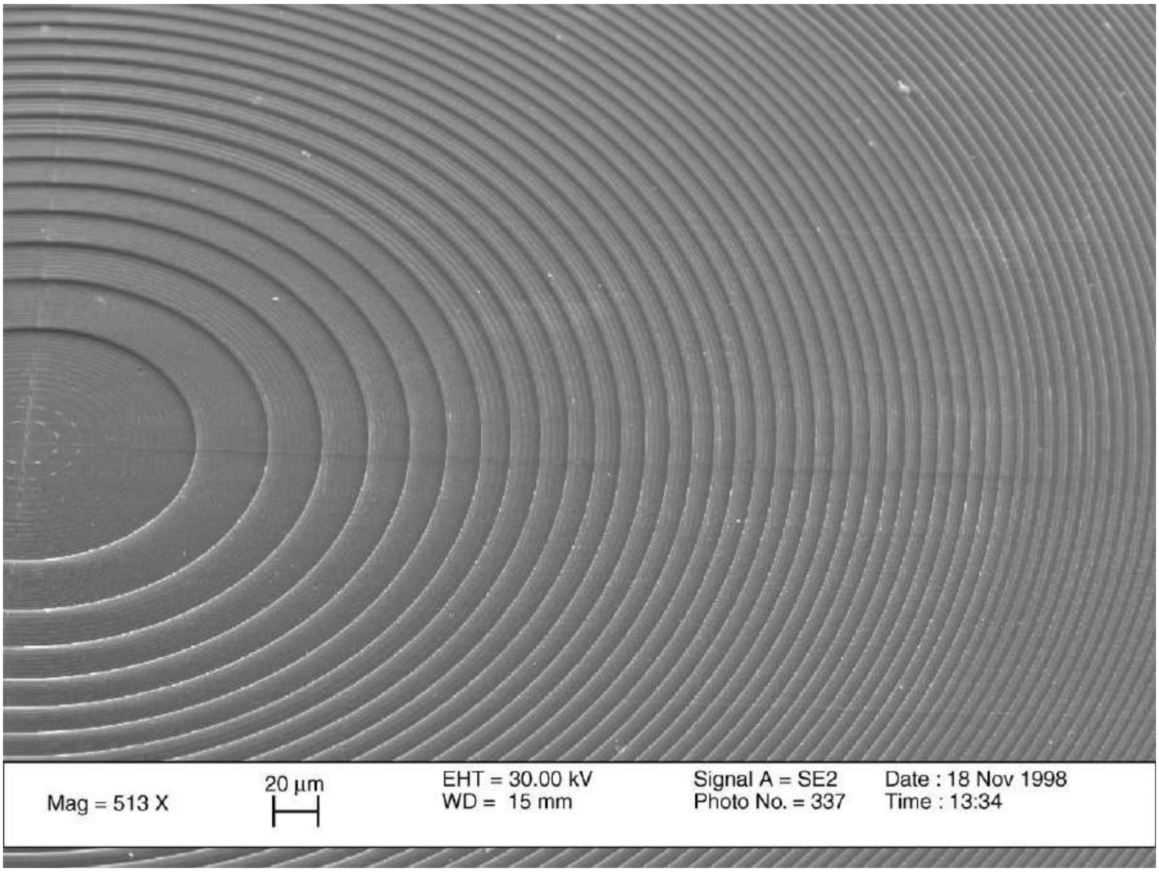
Mag = 513 X
20 µm
EHT = 30.00 kV
WD = 15 mm
Signal A = SE2
Photo No. = 337
Date : 18 Nov 1998
Time : 13:34

Figure 6 A scanning electron micrograph of a part of a diffractive lens.

reflective by depositing a thin metal film on top of it. If large-scale fabrication of identical gratings is the goal, a thick (~100 μm) metal layer (for example nickel) can be grown on the surface by electroplating. After separation from the substrate, the electroplated metal film, which is a negative of the original profile, can be used as a stamp in replicating the original profile in plastics using methods such as hot embossing, ultraviolet radiation curing, and injection molding. In particular, the latter method provides the key to high-volume and low-cost production of various types of diffraction gratings and other grating-based diffractive structures in the industrial scale.

Some Grating Applications

We now proceed to describe a collection of representative grating applications. More exhaustive coverage of grating applications can be found in the Further Reading section at the end of this article.

Spectroscopic Gratings

The purpose of a spectroscopic grating is to divide the incident light into a spectrum, i.e., to direct different frequency components of the incident beam into different directions given by the grating (eqn [1]). When the grating period is reduced, the angular dispersion $\partial\theta_m/\partial\lambda$, and therefore the resolution of the spectrograph, increases.

In the numerical examples presented in Figure 5 we assume a reflection grating in a Littrow mount, i.e., $\theta_m^r = -\theta$. This mount is common in spectroscopy; it obviously requires that the grating is rotated when the spectrum is scanned. We see from eqn [1] that in the Littrow mount:

$$\sin\theta = \frac{m\lambda}{2d} \qquad [9]$$

By inserting this result back into eqn [1] and differentiating we obtain a quantitative measure for dispersion in the Littrow mount:

$$\frac{\partial\theta_m}{\partial\lambda} = -\frac{m}{2d\cos\theta} \qquad [10]$$

We consider TE polarization (electric field is linearly polarized in the y-direction) and order $m = -1$, optimize H to give maximum efficiency at $\lambda_0 = 550$ nm, and plot the diffraction efficiency η_{-1} over the visible wavelength range for two values of the grating period, $d = \lambda_0$ and $d = 2\lambda_0$. Rigorous diffraction theory is used with $n_I = n_2 = 1$, and the metal is assumed to be aluminum (n_1 and n_{III} are complex-valued).

In Figure 5a binary gratings with one groove (width $d/2$) per period is considered. Parametric optimization gives $H = 422$ nm if $d = 1\lambda$, and $H = 168$ nm if $d = 2\lambda$. Figure 5b illustrates the results for sinusoidal gratings with $H = 345$ nm if $d = 1\lambda$, and

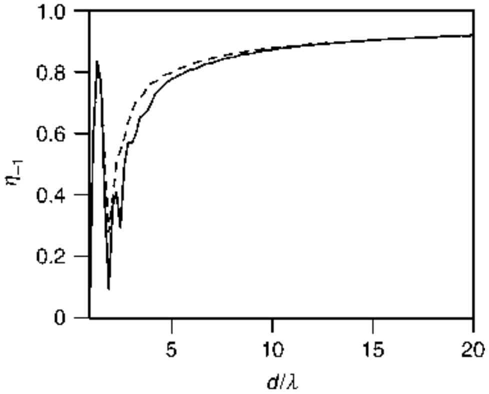

Figure 7 Diffraction efficiency versus wavelength for triangular transmission gratings illuminated at normal incidence. Solid lines: TE polarization. Dashed lines: TM polarization.

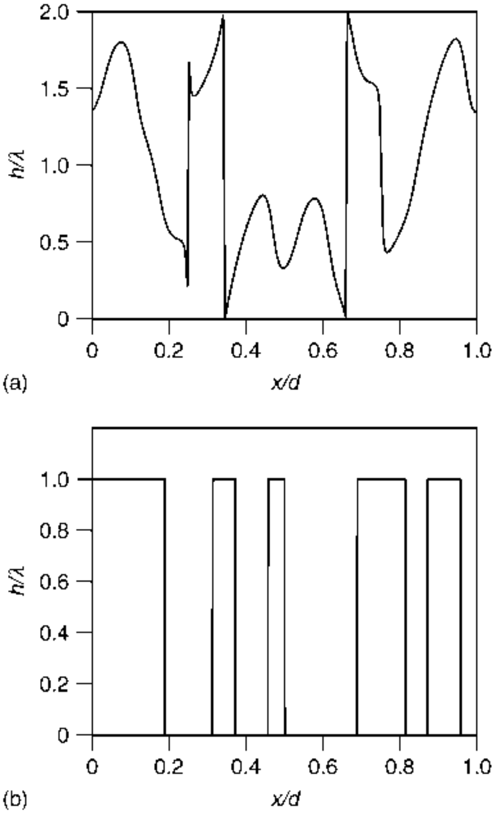

Figure 8 Surface-relief profiles of (a) continuous and (b) binary 1 → 8 beamsplitter gratings.

$H = 520$ nm if $d = 2\lambda$. Finally, triangular gratings with $H = 431$ nm if $d = 1\lambda$, and $H = 316$ nm if $d = 2\lambda$, are considered. The triangular grating profile is seen to be the best choice in general.

Diffractive Lenses

Diffractive lenses are gratings with locally varying period and groove orientation. Figure 6 illustrates the central part of a diffractive microlens fabricated by lithographic methods. It is seen to consist of concentric zones with radially decreasing width.

The phase transmission function of an ideal diffractive focusing lens can be obtained by a simple optical path calculation, and is given by

$$\phi(r) = \frac{2\pi}{\lambda_0}\left(f - \sqrt{f^2 + r^2}\right) \approx -\frac{\pi}{\lambda_0 f} r^2 \qquad [11]$$

where f is the focal length, r is the radial distance from the optical axis, and the approximate expression is valid in the paraxial region $r \ll f$. Here λ_0 is the design wavelength, which determines the modulation depth of the locally blazed grating profile:

$$H = \frac{\lambda_0}{n(\lambda_0) - 1} \qquad [12]$$

and $n(\lambda_0)$ is the refractive index of the dielectric material at $\lambda = \lambda_0$.

The zone boundaries r_n are defined by the condition $\phi(r_n) = -2\pi n$, and therefore we obtain from eqn [11]:

$$r_n = \sqrt{2nf\lambda_0 + (n\lambda_0)^2} \approx \sqrt{2nf\lambda_0} \qquad [13]$$

Obviously, when the numerical aperture of the lens is large, the zone width in the outer regions of the lens is reduced to values of the order of λ. Figure 7 illustrates the local diffraction efficiency in different part of a diffractive lens: we use the order $m = -1$ of a locally triangular diffraction grating of the

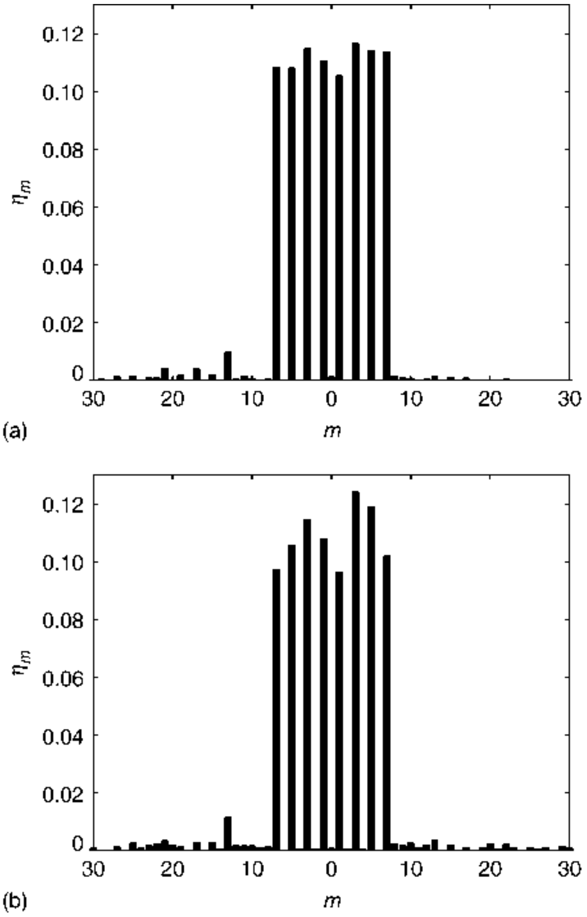

Figure 9 Diffraction pattern produced by a continuous-profile beamsplitter grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

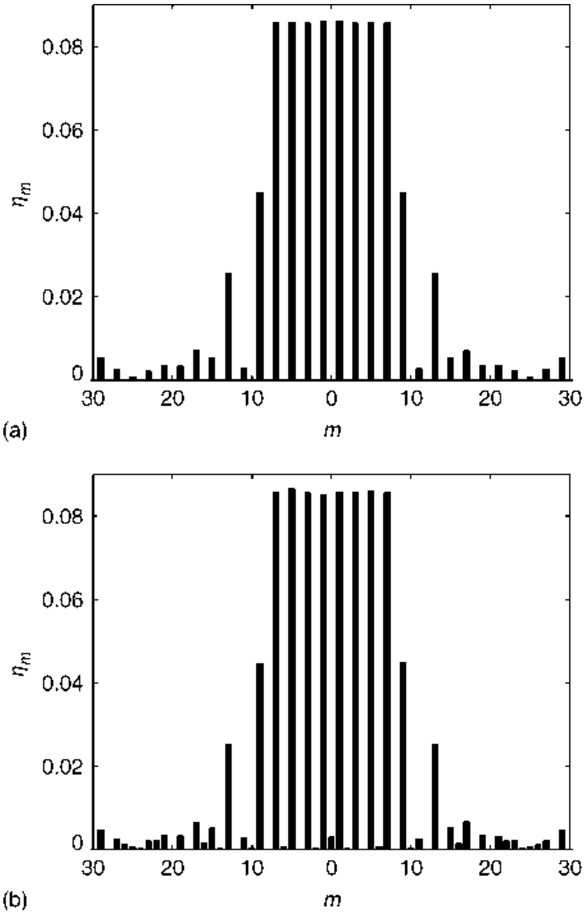

Figure 10 Diffraction pattern produced by a binary beamsplitter grating grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

type shown in Figure 1c with $\theta = 0$, assuming that $n_{\mathrm{I}} = n_2 = 1.5$ (fused silica) and $n_1 = n_{\mathrm{III}} = 1$. The efficiency η_{-1} is calculated as a function of λ using rigorous diffraction theory for both TE and TM polarization (the electric vector points in the groove direction in the TE case and is perpendicular to it in the TM case). It is seen that the local efficiency of the lens becomes poor in regions where the local grating period is small, and thus diffractive lenses operate best in the paraxial region (it is, however, possible to obtain improved nonparaxial performance by optimizing the local grating profile).

A factor that limits the use of diffractive lenses with polychromatic light is their large chromatic aberration: in the paraxial limit the focal length depends on the wavelength according to the formula:

$$f(\lambda) = \frac{\lambda_0}{\lambda} f \qquad [14]$$

However, since the dispersion of a grating has an opposite sign compared to the diffraction of a prism, a weak positive diffractive lens can be combined with a relatively strong positive refractive lens to form a single-element achromatic lens. Such a hybrid lens is substantially thinner and lighter than a conventional all-refractive achromatic lens made of crown and flint glasses.

Beam Splitter Gratings

Multiple beamsplitters, also known as array illuminators, are gratings with sophisticated periodic structure that are capable of transforming an incident plane wave into a set of diffraction orders with a specified distribution of diffraction efficiencies. Most often one wishes to obtain a one- or two-dimensional array of diffracted plane waves with equal efficiency. This goal can be achieved by appropriate design of the grating profile, which may be of binary, multilevel, or continuous form.

Figure 8 illustrates binary and continuous grating profiles capable of dividing one incident plane wave into eight diffracted plane waves (central odd-numbered orders of the grating) with equal efficiency in view of TEA. Figure 9 shows the distribution of energy among the different diffraction orders in and around the array produced by the continuous profile, evaluated by rigorous theory for two different grating periods, while Figure 10 gives corresponding results for the binary grating.

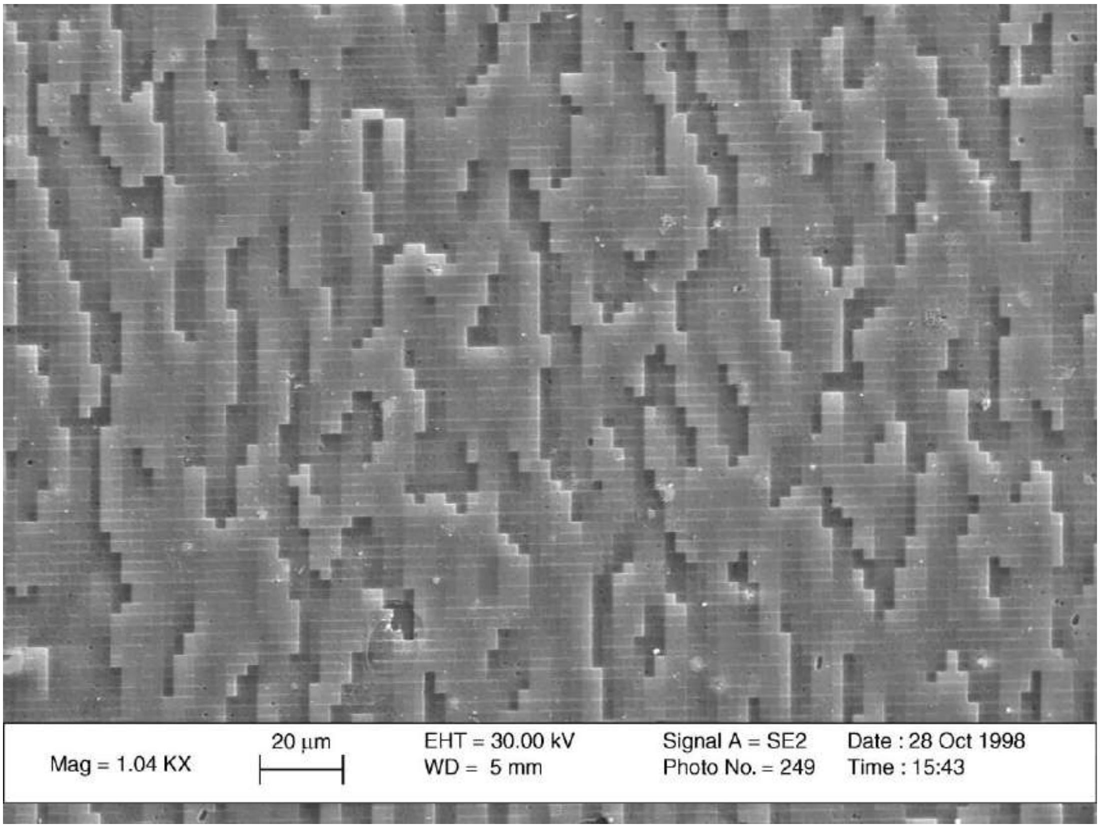

Figure 11 Scanning electron micrograph of a sophisticated beamsplitter grating.

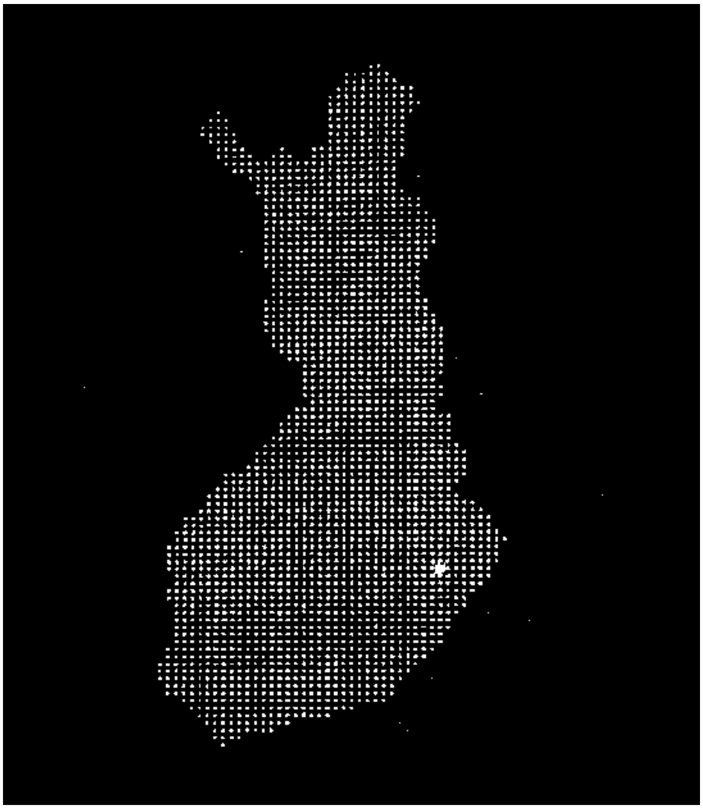

Figure 12 The far-field diffraction pattern produced by the grating in Figure 11.

The fraction of incident energy that ends up in the desired eight orders is 72% for the binary grating and 96% for the continuous grating according to TEA, making the latter more attractive from a theoretical viewpoint. However, the continuous profile is more difficult to fabricate accurately than the binary one, especially for small values of the ratio d/λ. Moreover, the array uniformity produced by the continuous profile degrades faster than that of the binary profile when the period is reduced.

An example of the structure of a biperiodic grating that is capable of forming a large, shaped array of diffraction orders with equal efficiency is presented in Figure 11. Figure 12 shows the far-field diffraction pattern produced by this element: the shape of Finland with the location of the town of Joensuu highlighted by placing the zero diffraction order on the appropriate position on the map (the apparent nonuniformity in the array is mostly due to the finite resolution of the CCD detector).

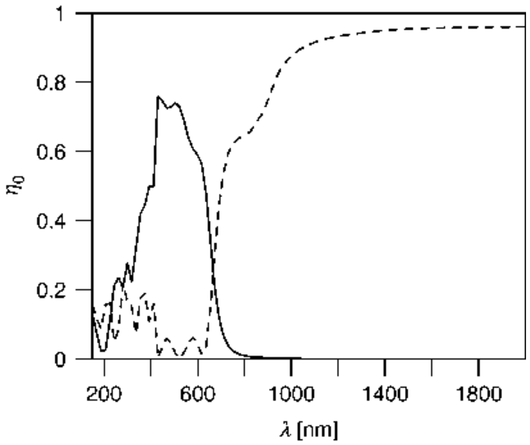

Figure 13 Zero order spectral transmission (solid line) and reflection (dashed line) of an aluminium inductive grid filter with 400 nm period.

Inductive Grid Filters

As a final example we consider a scaled-down version of a device well known from the door of a microwave oven: one can see in through the holes pierced in the metallic screen in the oven door, but the microwaves are trapped inside because their wavelengths are large in comparison with the dimensions of the holes. Another example of these inductive grid filters is found in large radio telescope mirrors for centimeter-scale wavelengths, which are wide-grid structures rather than smoothly curved metal surfaces.

Figure 13 illustrates the spectral transmission and reflection curves of an inductive grid filter of the type illustrated in Figure 3b. We assume that $d_x = d_y = d = 400$ nm, $c_x = c_y = c = 310$ nm, $H = 500$ nm, and that the grid is a self-supporting structure ($n_{\mathrm{I}} = n_{\mathrm{III}} = 1$) made of aluminum. Only the zero reflected and transmitted orders propagate when $\lambda > d$, but other orders (with rather low efficiency) appear for smaller wavelengths. We see that the filter reflects infrared radiation effectively, but transmits a substantial part of radiation at visible and shorter wavelength regions.

List of Units and Nomenclature

Complex field amplitude	U
Diffraction angles [rad]	θ, ϕ
Diffraction efficiency	η
Focal length [m]	f, f_0
Grating period [m]	d
Grating thickness [m]	H
Number of quantization levels	Q
Phase function [rad]	$\phi(r)$
Refractive index	n
Wavelength [m]	λ, λ_0
Wave vector [m^{-1}]	$\mathbf{k}$

See also

Fiber Gratings.

Further Reading

Chandezon J, Maystre D and Raoult G (1980) A new theoretical method for diffraction gratings and its numerical application. *Journal of Optics* 11: 235–241.

Gaylord TK and Moharam MG (1985) Analysis and applications of optical diffraction by gratings. *Proceedings of IEEE* 73: 894–937.

Herzig HP (ed.) (1997) *Micro-optics: Elements, Systems and Applications*. London: Taylor & Francis.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Li L (2001) Mathematical reflections on the fourier modal method in grating theory. In: Bao G, Cowsar L and Masters W (eds) *Mathematical Modelling in Optical Science*, pp. 111–139. Philadelphia: SIAM.

Li L, Chandezon J, Granet G and Plumey J-P (1999) Rigorous and efficient grating-analysis method made easy for optical engineers. *Applied Optics* 38: 304–313.

Petit R (ed.) (1980) *Electromagnetic Theory of Gratings*. Berlin: Springer-Verlag.

Rayleigh JWS (1907) On the dynamical theory of gratings. *Proceeding of the Royal Society A* 79: 399.

Solymar L and Cooke DJ (1981) *Volume Holography and Volume Gratings*. London: Academic Press.

Turunen J, Kuittinen M and Wyrowski F (2000) Diffractive optics: electromagnetic approach. In: Wolf E (ed.) *Progress in Optics XL*. Amsterdam: Elsevier.

Turunen J and Wyrowski F (eds) (1997) *Industrial and Commercial Applications of Diffractive Optics*. Berlin: Akademie Verlag.

van Renesse RL (1993) *Optical Document Security*. Boston, MA: Artech House.

Fraunhofer Diffraction

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Geometrical optics does a reasonable job of describing the propagation of light in free space unless that light encounters an obstacle. Geometrical optics predicts that the obstacle would cast a sharp shadow. On one side of the shadow's edge would be a bright uniform light distribution, due to the incident light, and on the other side there would be darkness. Close examination of an actual shadow edge reveals, however, dark fringes in the bright region and bright fringes in the dark region. This departure from the predictions of geometrical optics was given the name diffraction by Grimaldi.

There is no substantive difference between diffraction and interference. The separation between the two subjects is historical in origin and is retained for pedagogical reasons. Since diffraction and interference are actually the same physical process, we expect the observation of diffraction to be a strong function of the coherence of the illumination. With incoherent sources, geometrical optics is usually all that is needed to predict the performance of an optical system and diffraction can be ignored except at dimensions on the scale of a wavelength. With light sources having a large degree of spatial coherence, it is impossible to neglect diffraction in any analysis.

Even though diffraction produced by most common light sources and objects is a small effect, it is possible to observe diffraction without special equipment. By viewing a light source through a slit formed by two fingers that are nearly touching, fringes can be observed. Diffraction is used by photographers and television cameramen to provide artistic highlights in an image containing small light sources. They accomplish this by placing a screen in front of the camera lens. Light from point sources in the field of view is diffracted by the screen and produces ‘stars’ in the image. A third place where diffraction effects are easily observed is when a mercury or sodium streetlight, a few hundred meters away, is viewed through a sheer curtain.

Fresnel originally developed an approximate scalar theory based on Huygens’ principle which states that:

> Each point on a wavefront can be treated as a source of a spherical wavelet called a secondary wavelet or a Huygens’ wavelet. The envelope of these wavelets, at some later time, is constructed by finding the tangent to the wavelets. The envelope is assumed to be the new position of the wavefront.

Rather than simply using the wavelets to construct an envelope, Fresnel’s scalar theory assumes that the Huygens’ wavelets interfere to produce a new wavefront (*see* **Diffraction:** Fresnel Diffraction).

A rigorous diffraction theory is based on Maxwell’s equations and uses the boundary conditions associated with the obstacle to calculate a field scattered by the obstacle. The origins of this scattered field are currents induced in the obstacle by the incident field. The scattered field is allowed to interfere with the incident field to produce a resultant diffracted field. The application of the rigorous theory is very difficult and for most problems an approximate scalar theory developed by Kirchhoff and Sommerfeld is used.

Kirchhoff Theory of Diffraction

Kirchhoff’s theory was based on the elastic theory of light but can be reformulated into a vector theory. Here we will limit our discussion to the scalar formulation. Given an incident wave, φ, we wish to calculate the optical wave at point P_0, in Figure 1, located at $\mathbf{r}_0$, in terms of the wave’s value on a surface, S, that we construct about the observation point.

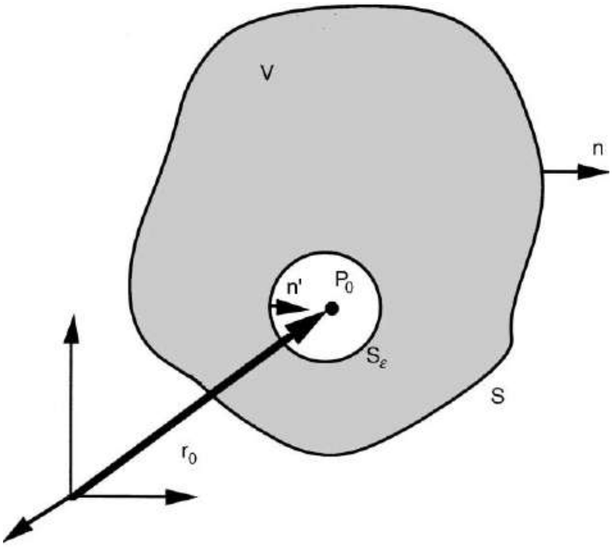

Figure 1 Region of integration for solution of Kirchhoff’s diffraction integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

To obtain a solution to the wave equation at the point, P_0, we select as a Green’s function the one proposed by Kirchhoff (Box 1), a unit amplitude spherical wave, expanding about the point at $\mathbf{r}_0$, denoted by Ψ, and then apply Green’s theorem (see Box 2). We will run into boundary condition problems as we proceed with the derivation of the solution. Sommerfeld removed the problems by assuming a slightly more complex Green’s function (see Box 1). We will limit our discussion to the Kirchhoff Green’s function.

Green’s theorem requires that there be no sources (singularities) inside the surface S but our Green’s function [1] is based on a source at r_0. We eliminate the source by constructing a small spherical surface S_ε, of radius ε, about $\mathbf{r}_0$, excluding the singularity at $\mathbf{r}_0$ from the volume of interest, shown in gray in Figure 1. The surface integration that must be performed in Green’s theorem is over the surface $S' = S + S_\varepsilon$.

Within the volume enclosed by S', the Green’s function, Ψ, and the incident wave, φ, satisfy the scalar Helmholtz equation so that the volume integral can be written as

$$\iiint_V (\Psi\nabla^2\varphi - \varphi\nabla^2\Psi)\mathrm{d}v = -\iiint_V (\Psi\varphi k^2 - \varphi\Psi k^2)\mathrm{d}v \qquad [7]$$

The right side of [7] is identically equal to zero. This fact allows us to use [6] to produce a simplified statement of Green’s theorem:

$$\iint_{S'} \left(\Psi\frac{\partial\varphi}{\partial n} - \varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s = 0 \qquad [8]$$

Box 1

Green's Function

1. **Kirchhoff Green's function**: Assume that the wave, Ψ, is due to a single point source at $\mathbf{r}_0$. At an observation point, $\mathbf{r}_1$, the Green's function is given by

$$\Psi(\mathbf{r}_1) = \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}, \qquad [1]$$

where $r_{01} = |\mathbf{r}_{01}| = |\mathbf{r}_1 - \mathbf{r}_0|$ is the distance from the source of the Green's function, at $\mathbf{r}_0$, to the observation position, at $\mathbf{r}_1$.

2. **Sommerfeld Green's function**: Assume that there are two point sources, one at P_0 and one at P'_0 (see Figure 2 where the source, P'_0, is positioned to be the mirror image of the source at P_0 on the opposite side of the screen, thus

$$r_{01} = r'_{01} \qquad \cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) = -\cos(\hat{\mathbf{n}}, \mathbf{r}'_{01}) \qquad [2]$$

A Green's function that could be used is

$$\Psi' = \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} - \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}'_{01}}}{r'_{01}} \qquad [3a]$$

or

$$\Psi' = \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} + \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}'_{01}}}{r'_{01}} \qquad [3b]$$

Box 2

Green's Theorem

To calculate the complex amplitude, $\tilde{E}(\mathbf{r})$, of a wave at an observation point defined by the vector $\mathbf{r}$, we need to use a mathematical relation known as Green's Theorem. Green's theorem states that, if φ and Ψ are two scalar functions that are well behaved, then

$$\int\!\!\int_S (\varphi\nabla\Psi\cdot\hat{\mathbf{n}} - \Psi\nabla\varphi\cdot\hat{\mathbf{n}})\mathrm{d}s = \int\!\!\int\!\!\int_V (\varphi\nabla^2\Psi - \Psi\nabla^2\varphi)\mathrm{d}v \qquad [4]$$

The vector identities

$$\nabla\Psi\cdot\hat{\mathbf{n}} = \frac{\partial\Psi}{\partial n} \qquad \nabla\varphi\cdot\hat{\mathbf{n}} = \frac{\partial\varphi}{\partial n} \qquad [5]$$

allow Green's theorem to be written as

$$\int\!\!\int_S \left[\varphi\frac{\partial\Psi}{\partial n} - \Psi\frac{\partial\varphi}{\partial n}\right]\mathrm{d}s = \int\!\!\int\!\!\int_V [\varphi\nabla^2\Psi - \Psi\nabla^2\varphi]\mathrm{d}v \qquad [6]$$

This equation is the prime foundation of scalar diffraction theory but only the proper choice of Ψ, φ and the surface, S, allows direct application to the diffraction problem.

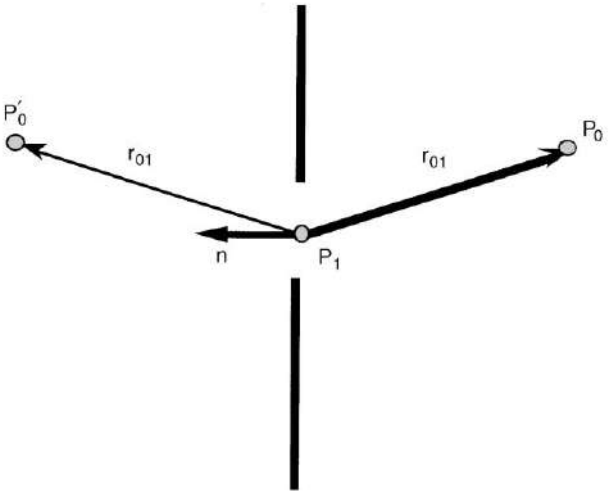

Figure 2 Geometry for the Sommerfeld Green's function. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Because the integral over the combined surfaces, S and S_ε, is equal to zero, the integral over the surface S must equal the negative of the integral over the surface, S_ε:

$$-\iint_{S_\varepsilon}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)ds=\iint_{S}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)ds \qquad [9]$$

To perform the surface integrals, we must evaluate the Green's function on the surface, S'. We therefore select $\mathbf{r}_1$, the vector defining the observation point, to be on either of the two surfaces that make up S'. The derivative, $\partial\Psi/\partial n$, to be evaluated is

$$\frac{\partial\Psi}{\partial n}=\frac{\partial\Psi}{\partial r}\frac{\partial r}{\partial n}$$

$$=\left(-ik\frac{e^{-ik\cdot r_{01}}}{r_{01}}-\frac{e^{-ik\cdot r_{01}}}{r_{01}^2}\right)\cos(n,\mathbf{r}_{01}) \qquad [10]$$

where $\cos(\hat{\mathbf{n}},\mathbf{r}_{01})$ is the cosine of the angle between the outward normal $\hat{\mathbf{n}}$ and the vector $\mathbf{r}_{01}$, the vector between points P_0 and P_1. (Note from Figure 1 that the outward normal on S_ε is inward, toward P_0, while the normal on S is outward, away from P_0.)

For example if $\mathbf{r}_1$ were on S_ε, then

$$\cos(\mathbf{n},\mathbf{r}_{01})=-1, \qquad [11]$$

$$\frac{\partial\Psi(\mathbf{r}_1)}{\partial n}=\left[\frac{1}{\varepsilon}+ik\right]\frac{e^{-ik\varepsilon}}{\varepsilon}, \qquad [12]$$

where ε is the radius of the small sphere we constructed around the point at P_0.

We now use [10] to rewrite the two integrals in [9]. The integral over the surface S_ε is

$$\iint_{S_\varepsilon}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)ds$$

$$=\iint_{S_\varepsilon}\left[\frac{\partial\varphi}{\partial n}\frac{e^{-ik\varepsilon}}{\varepsilon}-\varphi\frac{e^{-ik\varepsilon}}{\varepsilon}\left(\frac{1}{\varepsilon}+ik\right)\right]$$

$$\times\varepsilon^2\sin\theta\,d\theta\,d\phi \qquad [13]$$

while the integral over the surface S is

$$\iint_{S}\left[\frac{e^{-ik\cdot r_{01}}}{r_{01}}\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial}{\partial n}\left(\frac{e^{-ik\cdot r_{01}}}{r_{01}}\right)\right]r^2\sin\theta\,d\theta\,d\phi \qquad [14]$$

The omitted volume contained within the surface S_ε is allowed to shrink to zero by taking the limit as $\varepsilon\to 0$. Equation [14] will not be affected by taking the limit. The first and last terms of the right side of [13] go to zero as $\varepsilon\to 0$ because they contain $\varepsilon e^{-ik\varepsilon}$. The second term in [13] contains $e^{-ik\varepsilon}$ which goes to 1 as $\varepsilon\to 0$. Therefore, in the limit as $\varepsilon\to 0$, [9] becomes

$$\varphi(\mathbf{r}_0)=\frac{1}{4\pi}\iint_{S}\left[\frac{e^{-ik\cdot r_{01}}}{r_{01}}\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial}{\partial n}\left(\frac{e^{-ik\cdot r_{01}}}{r_{01}}\right)\right]ds \qquad [15]$$

which is sometimes called the integral theorem of Kirchhoff. By carefully selecting the surface of integration in [15], we can calculate the diffraction field observed at P_0 produced by an aperture in an infinite opaque screen.

Assume that a source at P_2 in Figure 3 produces a spherical wave that is incident on an infinite, opaque screen from the left. To find the field at P_0 in Figure 3, we apply [15] to the surface $S_1+S_2+\Sigma$, where S_1 is a plane surface adjacent to the screen, Σ is that portion of S_1 in the aperture and S_2 is a large spherical surface, of radius R, centered on P_0, see Figure 3. The first question to address is how to calculate the contribution from S_2 or better yet how do we show we can ignore contributions from S_2. As R increases, Ψ and φ will decrease as $1/R$. However, the area of integration increases as R^2 so the $1/R$ fall-off is not a sufficient reason for neglecting the contribution of the integration over S_2.

On the surface S_2, the Green's function and its derivative are

$$\Psi=\frac{e^{-ikR}}{R} \qquad [16]$$

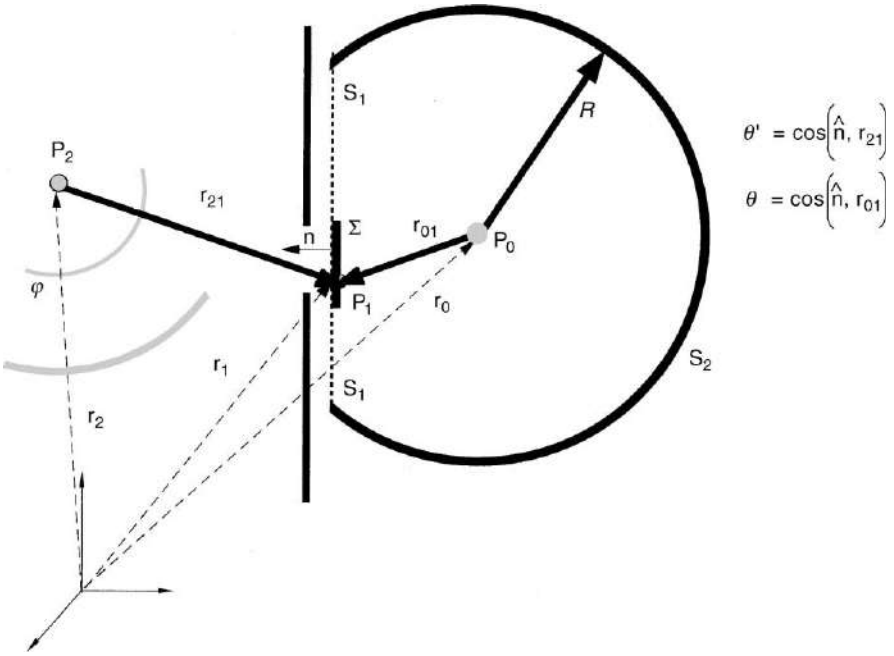

Figure 3 Integration surface for diffraction calculation. P_2 is the source and P_0 is the point where the field is to be calculated. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$\frac{\partial \Psi}{\partial n} = \left(-\mathrm{i}k - \frac{1}{R}\right)\frac{\mathrm{e}^{-\mathrm{i}kR}}{R} \approx -\mathrm{i}k\Psi \qquad [17]$$

where the approximate equality is for large R. Thus, the integral over S_2 for large R is

$$\iint_{S_2}\left[\Psi\frac{\partial\varphi}{\partial n} + \varphi(\mathrm{i}k\Psi)\right]\mathrm{d}s$$

$$= \iint_{S_2}\Psi\left(\frac{\partial\varphi}{\partial n} + \mathrm{i}k\varphi\right)R^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \qquad [18]$$

The quantity $R\Psi = \mathrm{e}^{-\mathbf{ik\cdot R}}$ is finite as $R \to \infty$ so for the integral to vanish, we must have

$$\lim_{R\to\infty} R\left(\frac{\partial\varphi}{\partial n} + \mathrm{i}k\varphi\right) = 0 \qquad [19]$$

This requirement is called the Sommerfeld radiation condition and is satisfied if $\varphi \to 0$ as fast as $1/R$ (for example, a spherical wave). Since the illumination of the screen will be a spherical wave, or at least a linear combination of spherical waves, we should expect the integral over S_2 to make no contribution (it is exactly zero).

Another way to insure that surface S_2 makes no contribution is to assume that the light turns on at time t_0. At a later time, t, when we desire to know the field at P_0, the radius of S_2 is $R > c\,(t - t_0)$ which physically means that the light has not had time to reach S_2. In this situation, there can be no contribution from S_2. This is not a perfect solution because when the wave is of finite duration it can no longer be considered monochromatic.

We should expect the major contribution in the integral over S_1 to come from the portion of the surface in the aperture, Σ. We make the following assumptions about the incident wave, φ (these assumptions are known as St. Venaut's hypothesis or Kirchhoff's boundary conditions)

1. We assume that, in the aperture, φ and $\partial\varphi/\partial n$ have the values they would have if the screen were not in place.
2. On the portion of S_1 not in the aperture, φ and $\partial\varphi/\partial n$ are identically zero.

The Kirchhoff boundary conditions allow the screen to be neglected, reducing the problem to an integration only over the aperture. It is surprising that the assumptions about the contribution of the surface S_2 and the screen yield very accurate results (as long as polarization is not important, the aperture is large with respect to the wavelength, and we don't get too close to the aperture).

Mathematically the Kirchhoff boundary conditions are incorrect. The two Kirchhoff boundary conditions imply that the field is zero everywhere behind the screen, except in the aperture, which makes Ψ and $\partial\Psi/\partial n$ discontinuous on the boundary of the aperture. Another problem with the boundary conditions is that Ψ and $\partial\Psi/\partial n$ are known only if the problem has already been solved.

The Sommerfeld Green's functions will remove the inconsistencies of the Kirchhoff theory. If we use the Green's function [3a] then Ψ vanishes in the aperture while $\partial\Psi/\partial n$ can assume the value required by Kirchhoff's boundary conditions. If we use the Green's functions [3b] then the converse holds, i.e. $\partial\Psi/\partial n$ is zero in the aperture. This improved Green's function makes the problem harder with little gain in accuracy so we will retain the Kirchhoff formalism.

As a result of the above discussion, the surface integral [15] is only performed over the surface, Σ, in the aperture. The evaluation of the Green's function on this surface can be simplified by noting that normally $r_{01} \gg \lambda$, i.e., $k \gg 1/r_{01}$, thus on Σ,

$$\begin{aligned}\frac{\partial\Psi}{\partial n} &= \cos(\mathbf{n}, \mathbf{r}_{01})\left(-ik - \frac{1}{r_{01}}\right)\frac{e^{-ik\cdot r_{01}}}{r_{01}} \\ &\approx -ik\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})\frac{e^{-ik\cdot r_{01}}}{r_{01}}\end{aligned} \qquad [20]$$

Substituting this approximate evaluation of the derivative into [15] yields

$$\varphi(\mathbf{r}_0) = \frac{1}{4\pi}\iint_{\Sigma} \frac{e^{-ik\cdot r_{01}}}{r_{01}}\left[\frac{\partial\varphi}{\partial n} + ik\varphi\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})\right]ds \qquad [21]$$

The source of the incident wave is a point source located at P_2, with a position vector, $\mathbf{r}_2$, measured with respect to the coordinate system and a distance $|\mathbf{r}_{21}|$ away from P_1, a point in the aperture (see Figure 3). The incident wave is therefore a spherical wave of the form

$$\varphi(\mathbf{r}_{21}) = A\frac{e^{-ik\cdot r_{21}}}{r_{21}} \qquad [22]$$

which fills the aperture. Here also we will assume that $r_{21} \gg \lambda$ so that the derivative of the incident wave assumes the same form as [20]. Then [21] can be written

$$\begin{aligned}\varphi(\mathbf{r}_0) = \frac{iA}{\lambda}\iint_{\Sigma} &\frac{e^{-ik\cdot(r_{21}+r_{01})}}{r_{21}r_{01}} \\ &\times\left[\frac{\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})}{2}\right]ds\end{aligned} \qquad [23]$$

This relationship is called the Fresnel–Kirchhoff diffraction formula. It is symmetric with respect to r_{01} and r_{21}, making the problem identical when the source and measurement point are interchanged.

A physical understanding of [23] can be obtained if it is rewritten as

$$\varphi(\mathbf{r}_0) = \iint_{\Sigma} \Phi(\mathbf{r}_{21})\frac{e^{-ik\cdot r_{01}}}{r_{01}}ds \qquad [24]$$

The field at P_0 is due to the sum of an infinite number of secondary Huygens sources in the aperture Σ. The secondary sources are point sources radiating spherical waves of the form

$$\Phi(\mathbf{r}_{21})\frac{e^{-ik\cdot r_{01}}}{r_{01}} \qquad [25]$$

with amplitude $\Phi(r_{21})$, defined by

$$\Phi(\mathbf{r}_{21}) = \frac{i}{\lambda}\left[A\frac{e^{-ik\cdot r_{21}}}{r_{21}}\right]\left[\frac{\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})}{2}\right] \qquad [26]$$

The imaginary constant, i, causes the wavelets from each of these secondary sources to be phase shifted with respect to the incident wave. The obliquity factor, in the amplitude [26]

$$\tfrac{1}{2}[\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})] \qquad [27]$$

causes the secondary sources to have a forward directed radiation pattern.

If we had used the Sommerfeld Green's function, the only modification to our result would be a change in the obliquity factor to $\cos(\mathbf{n}, \mathbf{r}_{01})$. In our discussions below we will assume that the angles are all small, allowing the obliquity factor to be replaced by 1 so that in the remaining discussion the choice of Green's function has no impact.

We will assume the source of light is at infinity, $z_1 = \infty$ in Figure 3, so that the aperture is illuminated by a plane wave, traveling parallel to the z-axis. With this assumption

$$\theta' = \cos(n, r_{21}) \approx -1 \qquad [28]$$

We will also make the paraxial approximation which assumes that the viewing position is close to the z-axis, leading to

$$\theta = \cos(\mathbf{n}, \mathbf{r}_{01}) \approx 1 \qquad [29]$$

With these assumptions eqn [24] becomes identical to the Huygens–Fresnel integral discussed in the section on Fresnel diffraction:

$$\tilde{E}(\mathbf{r}_0) = \frac{i}{\lambda}\iint_{\Sigma} \frac{\tilde{E}_i(\mathbf{r})}{R}e^{-ik\cdot R}ds \qquad [30]$$

The modern interpretation of the Fresnel–Kirchhoff (or now in the small-angle approximation, the Fresnel–Huygens) integral is to view it as a convolution integral. By considering free space to be a linear system, the result of propagation can be calculated by convolving the incident (input) wave with the impulse response of free space (our Green's function),

$$\frac{\mathrm{i}}{\lambda}\,\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}}}{R} \qquad [31]$$

The job of calculating diffraction has only begun with the derivation of the Fresnel–Kirchhoff integral. In general, an analytic expression for the integral cannot be found because of the difficulty of performing the integration over R. There are two approximations that will allow us to obtain analytic expressions of the Fresnel–Kirchhoff integral. In both approximations, all dimensions are assumed to be large with respect to the wavelength. In one approximation, the viewing position is assumed to be far from the obstruction; the resulting diffraction is called Fraunhofer diffraction and will be discussed here. The second approximation, which leads to Fresnel diffraction, assumes that the observation point is nearer the obstruction, to be quantified below. This approximation, which is the more difficult mathematically, is discussed in the article on **Diffraction:** Fresnel Diffraction.

Fraunhofer Approximation

In Fraunhofer diffraction, we require that the source of light and the observation point, P_0, be far from the aperture so that the incident and diffracted wave can be approximated by plane waves. A consequence of this requirement is that the entire waveform passing through the aperture contributes to the observed diffraction.

The geometry to be used in this derivation is shown in Figure 4. The wave incident normal to the aperture, Σ, is a plane wave and the objective of the calculation is to find the departure of the transmitted wave from its geometrical optical path. The calculation will provide the light distribution, transmitted by the aperture, as a function of the angle the light is deflected from the incident direction. We assume that diffraction makes only a small perturbation on the predictions of geometrical optics. The deflection angles encountered in this derivation are, therefore, small, as assumed above, and we will be able to use the paraxial approximation.

The distance from a point P, in the aperture, to the observation point P_0, of Figure 4, is

$$R^2 = (x-\xi)^2 + (y-\eta)^2 + Z^2 \qquad [32]$$

From Figure 4 we see R_0 is the distance from the center of the screen to the observation point, P_0,

$$R_0^2 = \xi^2 + \eta^2 + Z^2 \qquad [33]$$

The difference between these two vectors is

$$\begin{aligned} R_0^2 - R^2 &= (R_0 - R)(R_0 + R) \\ &= \xi^2 + \eta^2 + Z^2 - Z^2 - (x^2 - 2x\xi + \xi^2) \\ &\quad - (y^2 - 2y\eta + \eta^2) \\ &= 2(x\xi + y\eta) - (x^2 + y^2) \end{aligned} \qquad [34]$$

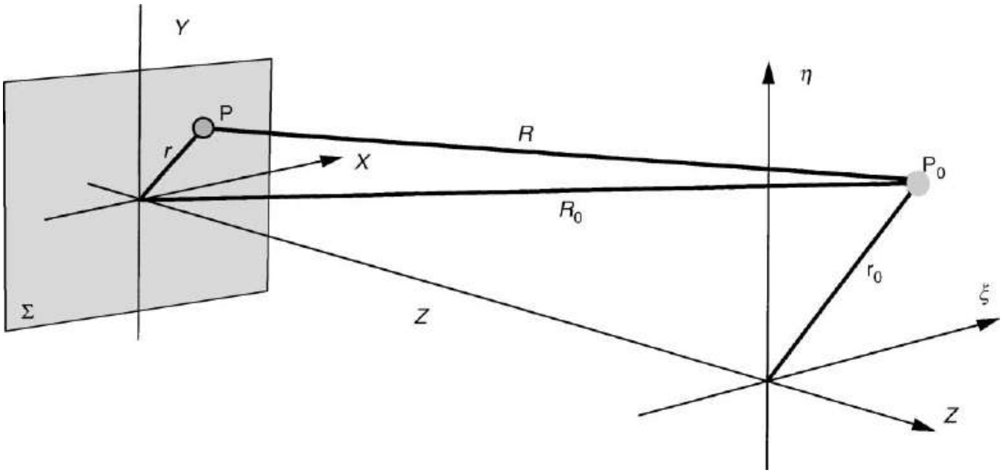

Figure 4 Geometry for Fraunhofer diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The equation for the position of point P in the aperture can be written in terms of [34]:

$$r = R_0 - R = \frac{R_0^2 - R^2}{R_0 + R}, \\ = [2(x\xi + y\eta) - (x^2 + y^2)]\frac{1}{R_0 + R} \qquad [35]$$

The reciprocal of $(R_0 + R)$ can be written as

$$\frac{1}{R_0 + R} = \frac{1}{2R_0 + R - R_0} = \frac{1}{2R_0}\left(1 + \frac{R - R_0}{2R_0}\right)^{-1} \qquad [36]$$

Now using [36]

$$R_0 - R = \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right]\left[1 - \frac{R_0 - R}{2R_0}\right]^{-1} \\ = \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right]\left[1 - \frac{r}{2R_0}\right]^{-1} \qquad [37]$$

If the diffraction integral [30] is to have a finite (nonzero) value, then

$$k|R_0 - R| \ll kR_0 \qquad [38]$$

This requirement insures that all the Huygens's wavelets, produced over the aperture from the center out to the position r, will have similar phases and will interfere to produce a nonzero amplitude at P_0. This requirement results in

$$\frac{1}{1 - \dfrac{r}{2R_0}} \approx 1 \qquad [39]$$

From this equation we are tempted to state that the aperture dimension must be small relative to the observation distance. However, for the exponent to provide a finite contribution in the integral, it is kr that is small, not the aperture dimension, r.

Using the approximation for $(R_0 - R)$, we obtain the diffraction integral

$$\tilde{E}_P = \frac{\mathrm{i}\alpha\, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0} \iint_{\Sigma} f(x, y)\mathrm{e}^{+\mathrm{i}k\left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right)}\,\mathrm{d}x\,\mathrm{d}y \qquad [40]$$

The change in the amplitude of the wave due to the change in R, as we move across the aperture, is neglected, allowing R in the denominator of the Huygens–Fresnel integral to be replaced by R_0 and moved outside of the integral. We have introduced the complex transmission function, $f(x, y)$, of the aperture to allow a very general aperture to be treated. If the aperture function described the variation in absorption of the aperture as a function of position, as would be produced by a photographic negative, then $f(x, y)$ would be a real function. If the aperture function described the variation in transmission of a biological sample, it might be entirely imaginary.

The argument of the exponent in [40] is

$$\mathrm{i}k\left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right) \\ = \mathrm{i}2\pi\left(\frac{x\xi + y\eta}{\lambda R_0} - \frac{x^2 + y^2}{2\lambda R_0}\right) \qquad [41]$$

If the observation point, P_0, is far from the screen, we can neglect the second term and treat the phase variation across the aperture as a linear function of position. This is equivalent to assuming that the diffracted wave is a collection of plane waves. Mathematically, the second term in [41] can be neglected if

$$\frac{x^2 + y^2}{2\lambda R_0} \ll 1 \qquad [42]$$

This is called the far-field approximation and the theory yields Fraunhofer diffraction. If the quadratic term of [41] is on the order of 2π then the fraction is

$$\frac{x^2 + y^2}{2\lambda R_0} \approx \Theta(1) \qquad [43]$$

and we must retain the quadratic term and the theory yields Fresnel diffraction.

We define spatial frequencies in the x and y direction by

$$\omega_x = \frac{2\pi}{\lambda}\sin\theta_x = -\frac{2\pi\xi}{\lambda R_0} \\ \omega_y = \frac{2\pi}{\lambda}\sin\theta_y = -\frac{2\pi\eta}{\lambda R_0} \qquad [44]$$

We define the spatial frequencies with a negative sign to allow equations involving spatial frequencies to have the same form as those involving temporal frequencies. The negative sign is required because ωt and $\mathbf{k}\cdot\mathbf{r}$ appear in the phase of the wave with opposite signs and we want the Fourier transform in space and in time to have the same form.

With the variables defined in [44], the integral becomes

$$\tilde{E}_P(\omega_x, \omega_y) = \frac{\mathrm{i}\alpha\, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0} \iint_{\Sigma} f(x, y)\mathrm{e}^{-\mathrm{i}(\omega_x x + \omega_y y)}\,\mathrm{d}x\,\mathrm{d}y \qquad [45]$$

The result of our derivation is that the Fraunhofer diffraction field, $\tilde{E}_P$, can be calculated by performing the two-dimensional Fourier transform of the aperture's transmission function.

Definition of Fourier Transform

In this discussion the function $F(\omega)$ is defined as the *Fourier transform* of $f(t)$:

$$\mathcal{F}\{f(t)\} \equiv F(\omega) \equiv \int_{-\infty}^{\infty} f(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau \qquad [46]$$

The transformation from a temporal to a frequency representation given by [46] does not destroy information; thus, the inverse transform can also be defined

$$\mathcal{F}^{-1}\{F(\omega)\} = f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega \qquad [47]$$

By assuming that the illumination of an aperture is a plane wave and that the observation position is in the far field, the diffraction of an aperture is found to be given by the Fourier transform of the function describing the amplitude transmission of the aperture. The amplitude transmission of the aperture, $f(x, y)$, may thus be interpreted as the superposition of mutually coherent plane waves.

Diffraction by a Rectangular Aperture

We will now use Fourier transform theory to calculate the Fraunhofer diffraction pattern from a rectangular slit and point out the reciprocal relationship between the size of the diffraction pattern and the size of the aperture.

Consider a rectangular aperture with a transmission function given by

$$f(x, y) = \begin{cases} 1 & |x| \le x_0, |y| \le y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \qquad [48]$$

Because the aperture is two-dimensional, we need to apply a two-dimensional Fourier transform but it is not difficult because the amplitude transmission function is separable, in x and y. The diffraction amplitude distribution from the rectangular slit is simply one-dimensional transforms carried out individually for the x and y dimensions:

$$\tilde{E}_P = \frac{\mathrm{i}\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\int_{-x_0}^{x_0} f(x)\mathrm{e}^{-\mathrm{i}\omega_x x}\,\mathrm{d}x \int_{-y_0}^{y_0} f(y)\mathrm{e}^{-\mathrm{i}\omega_y y}\,\mathrm{d}y \qquad [49]$$

Since both $f(x)$ and $f(y)$ are defined as symmetric functions, we need only calculate the cosine transforms to obtain the diffracted field:

To calculate the Fourier transform defined by [46] we can rewrite the transform as

$$F(\omega) = \int_{-\infty}^{\infty} f(\tau)\cos\omega\tau\,\mathrm{d}\tau - \mathrm{i}\int_{-\infty}^{\infty} f(\tau)\sin\omega\tau\,\mathrm{d}\tau \qquad [50]$$

If $f(\tau)$ is a real, even function then the Fourier transform can be obtained by simply calculating the cosine transform:

$$\int_{-\infty}^{\infty} f(\tau)\cos\omega\tau\,\mathrm{d}\tau \qquad [51]$$

$$\tilde{E}_P = \mathrm{i}\frac{4x_0y_0\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\frac{\sin\omega_x x_0}{\omega_x x_0}\frac{\sin\omega_y y_0}{\omega_y y_0} \qquad [52]$$

The intensity distribution of the Fraunhofer diffraction produced by the rectangular aperture is

$$I_P = I_0\frac{\sin^2\omega_x x_0}{(\omega_x x_0)^2}\frac{\sin^2\omega_y y_0}{(\omega_y y_0)^2} \qquad [53]$$

The maximum intensities in the x and y directions occur at

$$\omega_x x_0 = \omega_y y_0 = 0 \qquad [54]$$

The area of this rectangular aperture is defined as $A = 4x_0y_0$, resulting in an expression for the maximum intensity of

$$I_0 = \left[\frac{\mathrm{i}4x_0y_0\alpha\,\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\right]^2 = \frac{A^2\alpha^2}{\lambda^2R_0^2} \qquad [55]$$

The minima of [53] occur when $\omega_x x_0 = n\pi$ or when $\omega_y y_0 = m\pi$. The location of the zeros can be specified as a dimension in the observation plane or, using the paraxial approximation, in terms of the angle defined by [44]

$$\begin{aligned} \sin\theta_x \approx \theta_x \approx \frac{\xi}{R_0} = \frac{n\lambda}{2x_0} \\ \sin\theta_y \approx \theta_y \approx \frac{\eta}{R_0} = \frac{m\lambda}{2y_0} \end{aligned} \qquad [56]$$

The dimensions of the diffraction pattern are characterized by the location of the first zero, i.e., when $n = m = 1$, and are given by the observation plane coordinates, ξ and η. The dimensions of the diffraction pattern are inversely proportional to the dimensions of the aperture. As the aperture dimension expands, the width of the diffraction pattern decreases until, in the limit of an infinitely wide aperture, the diffraction pattern becomes a delta function.

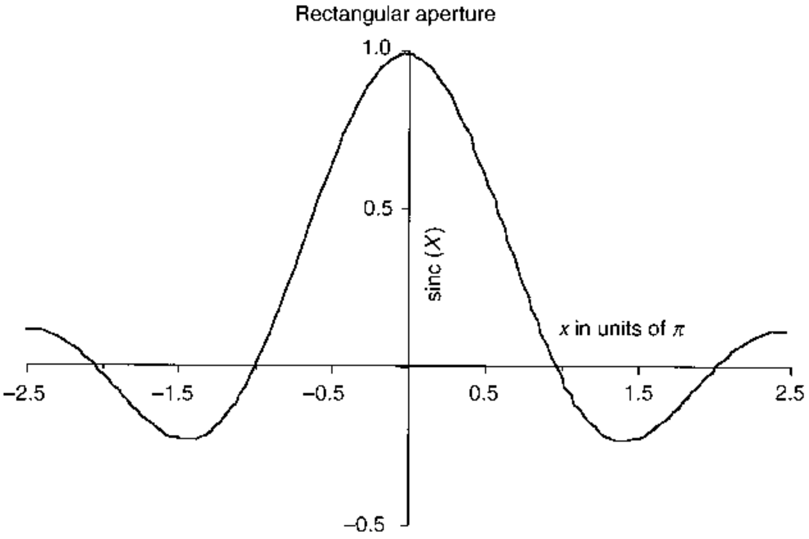

Figure 5 The amplitude of the wave diffracted by a rectangular slit. This sinc function describes the light wave's amplitude that would exist in both the *x* and *y* directions. The coordinate would have to be scaled by the dimension of the slit. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Figure 5 is a theoretical plot of the amplitude of the diffraction wave from a rectangular slit.

Diffraction from a Circular Aperture

To obtain the diffraction pattern from a circular aperture we first convert from the rectangular coordinate system we have used to a cylindrical coordinate system. The new cylindrical geometry is shown in Figure 6. At the aperture plane

$$\begin{aligned} x &= s\cdot\cos\varphi & y &= s\cdot\sin\varphi \\ f(x,y) &= f(s,\varphi) & \mathrm{d}x\,\mathrm{d}y &= s\,\mathrm{d}s\,\mathrm{d}\varphi \end{aligned} \qquad [57]$$

At the observation plane

$$\xi = \rho\cdot\cos\theta \qquad \eta = \rho\cdot\sin\theta \qquad [58]$$

In the new, cylindrical, coordinate system at the observation plane, the spatial frequencies are written as

$$\begin{aligned} \omega_x &= -\frac{k\xi}{R_0} = -\frac{k\rho}{R_0}\cos\theta \\ \omega_y &= -\frac{k\eta}{R_0} = -\frac{k\rho}{R_0}\sin\theta \end{aligned} \qquad [59]$$

From Figure 6 we see that the observation point, P, can be defined in terms of the angle ψ, where

$$\sin\psi = \frac{\rho}{R_0} \qquad [60]$$

This allows an angular representation of the size of the diffraction pattern if it is desired.

Using [57] and [59], we may write

$$\begin{aligned} \omega_x x + \omega_y y &= -\frac{ks\rho}{R_0}(\cos\theta\cos\varphi + \sin\theta\sin\varphi), \\ &= -\frac{ks\rho}{R_0}\cos(\theta-\varphi) \end{aligned} \qquad [61]$$

The Huygens–Fresnel integral for Fraunhofer diffraction can now be written in terms of cylindrical coordinates:

$$\tilde{E}_P = \frac{\mathrm{i}\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}k\cdot\mathbf{R}_0}\int_0^{\frac{a}{2}}\int_0^{2\pi} f(s,\varphi)\mathrm{e}^{-\mathrm{i}k\frac{s\rho}{R_0}\cos(\theta-\varphi)}\, s\,\mathrm{d}s\,\mathrm{d}\varphi \qquad [62]$$

We will use this equation to calculate the diffraction amplitude from a clear aperture of diameter *a*, defined by the equation

$$f(s,\varphi) = \begin{cases} 1 & s \le \frac{a}{2},\ \text{all } \varphi \\ 0 & s > \frac{a}{2} \end{cases} \qquad [63]$$

The symmetry of this problem is such that

$$f(s,\varphi) = f(s)g(\varphi) = f(s) \qquad [64]$$

$$\mathcal{F}\{f(s,\phi)\} = \mathcal{F}\{f(s)\} = F(\rho) \qquad [65]$$

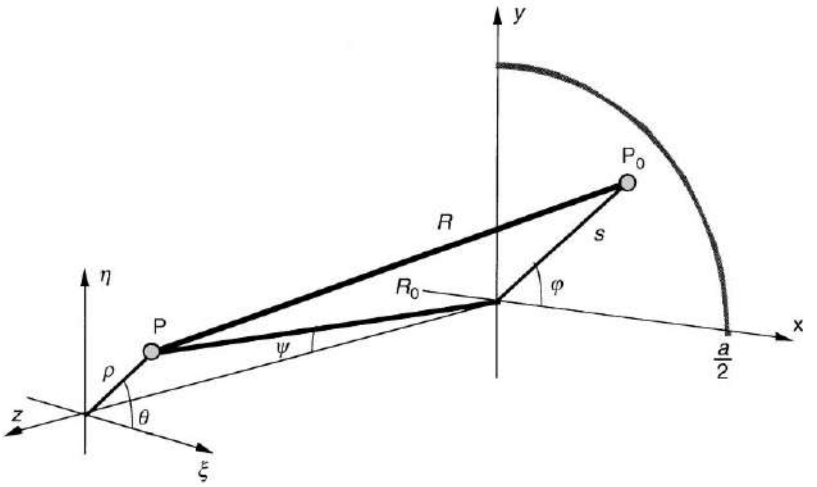

Figure 6 Geometry for calculation of Fraunhofer diffraction from a circular aperture. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$F(\rho, \theta) = \int_0^\infty f(s)s\,\mathrm{d}s \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}\rho\cdot s\cos(\varphi-\theta)}\,\mathrm{d}\varphi \qquad [66]$$

$$F(\rho) = \int_0^\infty f(s)\,s\,\mathrm{d}s \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}\rho\cdot s\cos\varphi}\,\mathrm{d}\varphi \qquad [67]$$

The second integral in [67] belongs to a class of functions called the Bessel function defined by the integral

$$\mathrm{J}_n(s\rho) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{\mathrm{i}[s\rho\sin\varphi - n\varphi]}\mathrm{d}\varphi \qquad [68]$$

In [67] the integral corresponds to the $n = 0$, zero-order, Bessel function. Using this definition, we can rewrite [67] as

$$F(\rho) = \int_0^\infty f(s)\mathrm{J}_0(s\rho)s\,\mathrm{d}s \qquad [69]$$

This transform is called the Fourier–Bessel transform or the Hankel zero-order transform.

Using these definitions, the transform of the aperture function, $f(s)$, is

$$F(\rho) = \int_0^{\frac{a}{2}} \mathrm{J}_0(s\rho)s\,\mathrm{d}s \qquad [70]$$

We use the identity

$$x\mathrm{J}_1(x) = \int_0^x \chi\mathrm{J}_0(\chi)\,\mathrm{d}\chi \qquad [71]$$

to obtain

$$F(\rho) = \frac{a}{2\rho}\mathrm{J}_1\left(\frac{\rho a}{2}\right) \qquad [72]$$

We can now write the amplitude of the diffracted wave as

$$\tilde{E}_P = \frac{\mathrm{i}\alpha}{\lambda}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\left[\frac{\pi a}{k\rho}\mathrm{J}_1\left(\frac{ka\rho}{2R_0}\right)\right] \qquad [73]$$

A plot of the function contained within the bracket of [73] is given in Figure 7. If we define

$$u = \frac{ka\rho}{2R_0} \qquad [74]$$

then the spatial distribution of intensity in the diffraction pattern can be written in a form known as the Airy formula,

$$I = I_0\left[\frac{2\mathrm{J}_1(u)}{u}\right]^2 \qquad [75]$$

where we have defined

$$I_0 = \left(\frac{\alpha A}{\lambda R_0}\right)^2 \qquad [76]$$

with A representing the area of the aperture, $A = \pi(\frac{a}{2})^2$.

The intensity pattern described by [75] is called the Airy pattern. The intensity at $u = 0$ is the same as was obtained for a rectangular aperture of the same area, because, in the limit,

$$\lim_{u\to 0}\left[\frac{2\mathrm{J}_1(u)}{u}\right] = 1 \qquad [77]$$

For the Airy pattern, 84% of the total area is contained in the disk between the first zeros of [75]. Those zeros occur at $u = \pm 1.22\pi$, which corresponds

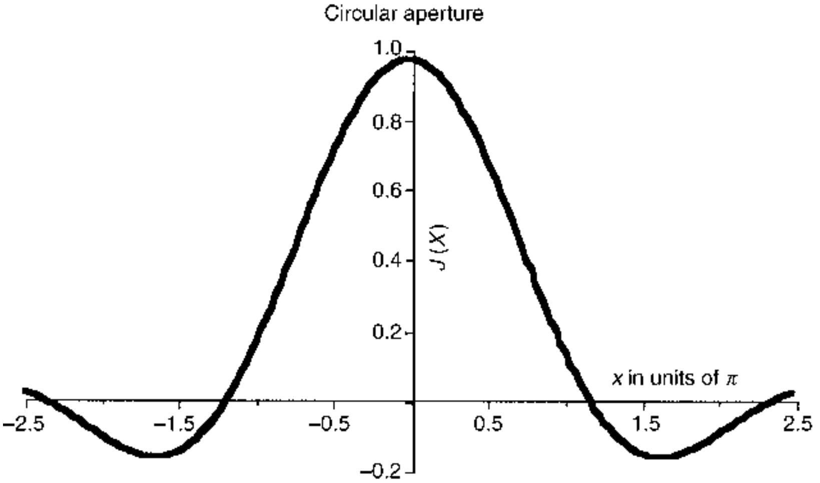

Figure 7 Amplitude of a wave diffracted by a circular aperture. The observed light distribution is constructed by rotating this Bessel function around the optical axis. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

to a radius in the observation plane of

$$\rho = \frac{1.22(\lambda R_0)}{a} \qquad [78]$$

91% of the light intensity is contained within the circle bounded by the second minimum at $u = 2.233\pi$. The intensities in the secondary maxima of the diffraction pattern of a rectangular aperture [53] are much larger than the intensities in the secondary maxima of the Airy pattern of a circular aperture of equal area. The peak intensities, relative to the central maximum, of the first three secondary maxima of a rectangular aperture are 4.7%, 1.6%, and 0.8%, respectively. For a circular aperture, the same quantities are 1.7%, 0.04%, and 0.02%, respectively.

Array Theorem

There is an elegant mathematical technique for handling diffraction from multiple apertures, called the array theorem. The theorem is based on the fact that Fraunhofer diffraction is given by the Fourier transform of the aperture function and utilizes the convolution integral.

The convolution of the functions $a(t)$ and $b(t)$ is defined as

$$g(t) = a(t) \otimes b(t) = \int_{-\infty}^{\infty} a(t)b(\tau - t)\,\mathrm{d}t \qquad [79]$$

The Fourier transform of a convolution of two functions is the product of the Fourier transforms of the individual functions:

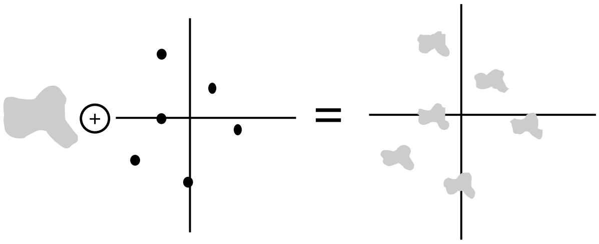

Figure 8 The convolution of an aperture with an array of delta functions will produce an array of identical apertures, each located at the position of one of the delta functions. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$\mathcal{F}\{a(t) \otimes b(t)\} = \mathcal{F}\left\{\int_{-\infty}^{\infty} a(t)b(\tau - t)\,\mathrm{d}t\right\} = A(\omega)B(\omega) \qquad [80]$$

We will demonstrate the array theorem for one dimension, where the functions represent slit apertures. The results can be extended to two dimensions in a straightforward way.

Assume that we have a collection of identical apertures, shown on the right of Figure 8. If one of the apertures is located at the origin of the aperture plane, its transmission function is $\psi(x)$. The transmission function of an aperture located at a point, x_n, can be written in terms of a generalized aperture function, $\psi(x - \alpha)$, by the use of the sifting property of the delta function

$$\psi(x - x_n) = \int \psi(x - \alpha)\delta(\alpha - x_n)\,\mathrm{d}\alpha \qquad [81]$$

The aperture transmission function representing an array of apertures will be the sum of the distributions of the individual apertures, represented graphically in Figure 8 and mathematically by the summation

$$\Psi(x) = \sum_{n=1}^{N} \psi(x - x_n) \qquad [82]$$

The Fraunhofer diffraction from this array is $\Phi(\omega_x)$, the Fourier transform of $\Psi(x)$,

$$\Phi(\omega_x) = \int_{-\infty}^{\infty} \Psi(x) \mathrm{e}^{-\mathrm{i}\omega_x x} \, \mathrm{d}x \qquad [83]$$

which can be rewritten as

$$\Phi(\omega_x) = \sum_{n=1}^{N} \int_{-\infty}^{\infty} \psi(x - x_n) \mathrm{e}^{-\mathrm{i}\omega_x x} \, \mathrm{d}x \qquad [84]$$

We now make use of the fact that $\psi(x - x_n)$ can be expressed in terms of a convolution integral. The Fourier transform of $\psi(x - x_n)$ is, from the convolution theorem [80], the product of the Fourier transforms of the individual functions that make up the convolution:

$$\begin{aligned}\Phi(\omega_x) &= \sum_{n=1}^{N} \mathcal{F}\{\psi(x-\alpha)\}\mathcal{F}\{\delta(\alpha - x_n)\} \\ &= \mathcal{F}\{\psi(x-\alpha)\} \sum_{n=1}^{N} \mathcal{F}\{\delta(\alpha - x_n)\} \\ &= \mathcal{F}\{\psi(x-\alpha)\}\mathcal{F}\left\{ \sum_{n=1}^{N} \delta(\alpha - x_n) \right\} \qquad [85]\end{aligned}$$

The first transform in [85] is the diffraction pattern of the generalized aperture function and the second transform is the diffraction pattern produced by a set of point sources with the same spatial distribution as the array of identical apertures. We will call this second transform the array function.

To summarize, the array theorem states that the diffraction pattern of an array of similar apertures is given by the product of the diffraction pattern from a single aperture and the diffraction (or interference) pattern of an identically distributed array of point sources.

N Rectangular Slits

An array of N identical apertures is called a diffraction grating in optics. The Fraunhofer diffraction patterns produced by such an array have two important properties; a number of very narrow beams are produced by the array and the beam positions are a function of the wavelength of illumination of the apertures and the relative phase of the waves radiated by each of the apertures.

Because of these properties, arrays of diffracting apertures have been used in a number of applications.

- At radio frequencies, arrays of dipole antennas are used to both radiate and receive signals in both radar and radio astronomy systems. One advantage offered by diffracting arrays at radio frequencies is that the beam produced by the array can be electrically steered by adjusting the relative phase of the individual dipoles.
- An optical realization of a two-element array of radiators is Young's two-slit experiment and a realization of a two-element array of receivers is Michelson's stellar interferometer. Two optical implementations of arrays, containing more diffracting elements, are diffraction gratings and holograms. In nature, periodic arrays of diffracting elements are the origin of the colors observed on some invertebrates.
- Many solids are naturally arranged in three-dimensional arrays of atoms or molecules that act as diffraction gratings when illuminated by X-ray wavelengths. The resulting Fraunhofer diffraction patterns are used to analyze the ordered structure of the solids.

The array theorem can be used to calculate the diffraction pattern from N rectangular slits, each of width a and separation d. The aperture function of a single slit is equal to

$$\psi(x, y) = \begin{cases} 1 & |x| \le \frac{a}{2}, |y| \le y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \qquad [86]$$

The Fraunhofer diffraction pattern of this aperture has already been calculated and is given by [52]:

$$\mathcal{F}\{\psi(x, y)\} = K \frac{\sin \alpha}{\alpha} \qquad [87]$$

where the constant K and the variable α are

$$\begin{aligned} K &= \frac{\mathrm{i}2ay_0\alpha}{\lambda R_0} \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0} \frac{\sin \omega_y y_0}{\omega_y y_0} \\ \alpha &= \frac{ka}{2} \sin \theta_x \end{aligned} \qquad [88]$$

The array function is

$$A(x) = \sum_{n=1}^{N} \delta(x - x_n) \quad \text{where } x_n = (n-1)d \qquad [89]$$

and its Fourier transform is given by

$$\mathcal{F}\{A(x)\} = \frac{\sin N\beta}{\sin \beta} \qquad [90]$$

$$\beta = \frac{kd}{2} \sin \theta_x$$

The Fraunhofer diffraction pattern's intensity distribution in the x-direction is thus given by

$$I_\theta = I_0 \underbrace{\frac{\sin^2\alpha}{\alpha^2}}_{\text{shape factor}} \underbrace{\frac{\sin^2 N\beta}{\sin^2\beta}}_{\text{grating factor}} \qquad [91]$$

We have combined the variation in intensity in the y-direction into the constant I_0 because we assume that the intensity variation in the x-direction will be measured at a constant value of y.

A physical interpretation of [91] views the first factor as arising from the diffraction associated with a single slit; it is called the shape factor. The second factor arises from the interference between light from different slits; it is called the grating factor. The fine detail in the spatial light distribution of the diffraction pattern is described by the grating factor and arises from the coarse detail in the diffracting object. The coarse, overall light distribution in the diffraction pattern is described by the shape factor and arises from the fine detail in the diffracting object.

Young's Double Slit

The array theorem makes the analysis of Young's two-slit experiment a trivial exercise. This application of the array theorem will demonstrate that the interference between two slits arises naturally from an application of diffraction theory. The result of this analysis will support a previous assertion that interference describes the same physical process as diffraction and the division of the two subjects is an arbitrary one.

The intensity of the diffraction pattern from two slits is obtained from [91] by setting $N = 2$:

$$I_\theta = I_0 \frac{\sin^2\alpha}{\alpha^2} \cos^2\beta \qquad [92]$$

The sinc function describes the energy distribution of the overall diffraction pattern, while the cosine function describes the energy distribution created by interference between the light waves from the two slits. Physically, α is a measure of the phase difference between points in one slit and β is a measure of the phase difference between similar points in the two slits. Zeros in the diffraction intensity occur whenever $\alpha = n\pi$ or whenever $\beta = \frac{1}{2}(2n+1)\pi$. Figure 9 shows the interference maxima, from the grating factor, under the central diffraction peak, described by the shape factor. The number of interference maxima contained under the central maximum is given by

$$\frac{2d}{a} - 1 \qquad [93]$$

In Figure 9 three different slit spacings are shown with the ratio d/a equal to 3, 6, and 9, respectively.

The Diffraction Grating

In this section we will use the array theorem to derive the diffraction intensity distribution of a large number of identical apertures. We will discover that the positions of the principal maxima are a function of the illuminating wavelength. This functional relationship has led to the application of a diffraction grating to wavelength measurements.

The diffraction grating normally used for wavelength measurements is not a large number of diffracting apertures but rather a large number of reflecting grooves cut in a surface such as gold or aluminum. The theory to be derived also applies to these reflection gratings but a modification must be made to the theory. The shape of the grooves in the reflection grating can be used to control the fraction of light diffracted into a principal maximum and we will examine how to take this into account. A grating whose groove shape has been controlled to enhance the energy contained in a particular principal maximum is called a blazed grating. The use of special groove shapes is equivalent to the modification of the phase of individual elements of an antenna array at radio frequencies.

The construction of a diffraction grating for use in an optical device for measuring wavelength was first suggested by David Rittenhouse, an American astronomer, in 1785, but the idea was ignored until Fraunhofer reinvented the concept in 1819. Fraunhofer's first gratings were fine wires spaced by wrapping the wires in the threads of two parallel screws. He later made gratings by cutting (*ruling*) grooves in gold films deposited on the surface of glass. H.A. Rowland made a number of well designed ruling machines, which made possible the production of large-area gratings. Following a suggestion by Lord Rayleigh, Robert Williams Wood (1868–1955) developed the capability to control the shape of the individual grooves.

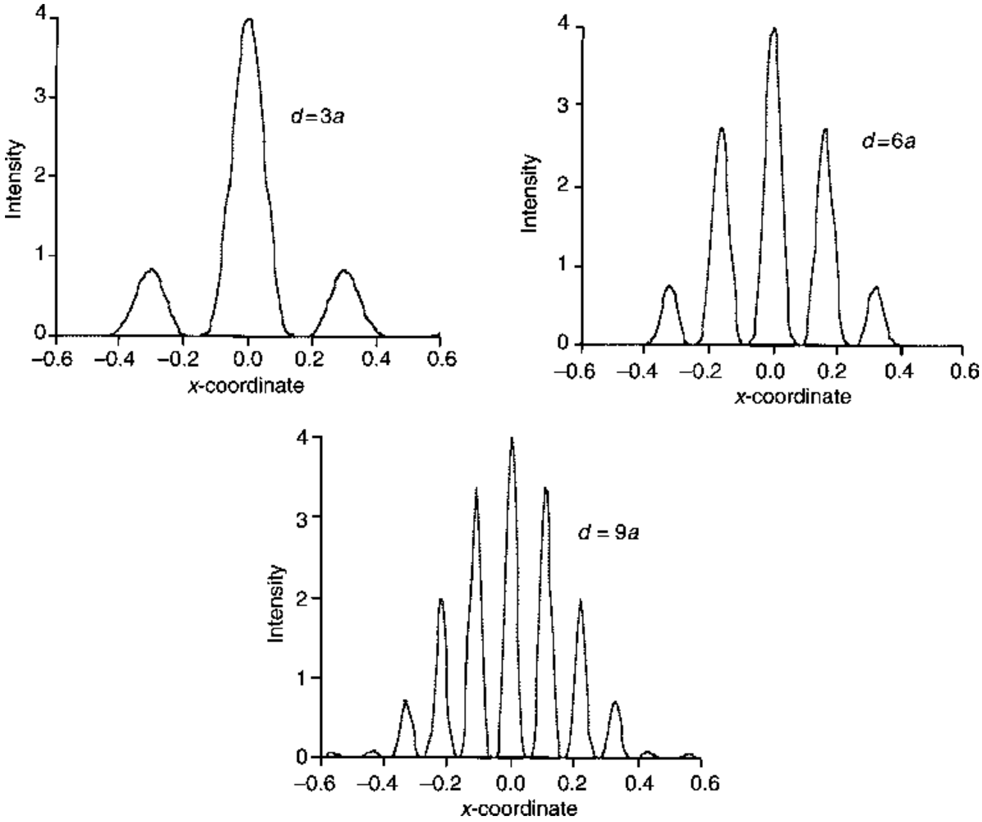

Figure 9 The number of interference fringes beneath the main diffraction peak of a Young's two-slit experiment with rectangular apertures. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If N is allowed to assume values much larger than 2, the appearance of the interference fringes, predicted by the grating factor, changes from a simple sinusoidal variation to a set of narrow maxima, called principal maxima, surrounded by much smaller, secondary maxima. To calculate the diffraction pattern we must evaluate [91] when N is a large number.

Whenever $N\beta = m\pi$, $m = 0, 1, 2, \ldots$, the numerator of the second factor in [91] will be zero, leading to an intensity that is zero, $I_\theta = 0$. The denominator of the second factor in [91] is zero when $\beta = l\pi$, $l = 0, 1, 2, \ldots$ If both conditions occur at the same time, the ratio of m and N is equal to an integer, $m/N = l$, and instead of zero intensity, we have an indeterminate value for the intensity, $I_\theta = 0/0$. To find the actual value for the intensity, we must apply L'Hospital's rule

$$\lim_{\beta \to l\pi} \frac{\sin N\beta}{\sin \beta} = \lim_{\beta \to l\pi} \frac{N \cos N\beta}{\cos \beta} = N \qquad [94]$$

L'Hospital's rule predicts that whenever

$$\beta = l\pi = \frac{m}{N}\pi \qquad [95]$$

where l is an integer, a principal maximum in the intensity will occur with a value given by

$$I_{\theta P} = N^2 I_0 \frac{\sin^2 \alpha}{\alpha^2} \qquad [96]$$

Secondary maxima, much weaker than the principal maxima, occur when

$$\beta = \left(\frac{2m+1}{2N}\right)\pi \qquad m = 1, 2, \ldots \qquad [97]$$

(When $m = 0$, m/N is an integer, thus the first value that m can have in [97] is $m = 1$.) The intensity of each secondary maximum is given by

$$\begin{aligned} I_{\theta S} &= I_0 \frac{\sin^2 \alpha}{\alpha^2} \frac{\sin^2 N\beta}{\sin^2 \beta} \\ &= I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{\sin^2\left(\frac{2m+1}{2}\right)\pi}{\sin^2\left(\frac{2m+1}{2N}\right)\pi} \right] \\ &= I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{1}{\sin\left(\frac{2m+1}{2N}\right)\pi} \right]^2 \end{aligned} \qquad [98]$$

The quantity $(2m+1)/2N$ is a small number for large values of N, allowing the small angle approximation to be made:

$$I_{\theta S} \approx I_0 \frac{\sin^2\alpha}{\alpha^2}\left[\frac{2N}{\pi(2m+1)}\right]^2 \qquad [99]$$

The ratio of the intensity of a secondary maximum and a principal maximum is given by

$$\frac{I_{\theta S}}{I_{\theta P}} = \left[\frac{2}{\pi(2m+1)}\right]^2 \qquad [100]$$

The strongest secondary maximum occurs for $m = 1$ and, for large N, has an intensity that is about 4.5% of the intensity of the neighboring principal maximum.

The positions of principal maxima occur at angles specified by the grating formula.

$$\beta = \frac{kd\sin\theta}{2} = l\pi = \frac{m}{N}\pi \qquad [101]$$

The angular position of the principal maxima (the diffracted angle θ_d) is given by the Bragg equation

$$\sin\theta_d = \frac{l\lambda}{d} \qquad [102]$$

where l is called the grating order. This simple relationship between the angle and the wavelength can be used to construct a device to measure wavelength, the grating spectrometer.

The model we have used to obtain this result is based on a periodic array of identical apertures. The transmission function of this array is a periodic square wave. If, for the moment, we treat the grating as infinite in size, we discover that the principal maxima in the diffraction pattern correspond to the terms of the Fourier series describing a square wave.

The zero order, $m = 0$, corresponds to the temporal average of the periodic square wave and has an intensity proportional to the spatially averaged transmission of the grating. Because of its equivalence to the temporal average of a time-varying signal, the zero-order principal maximum is often called the dc term.

The first-order principal maximum corresponds to the fundamental spatial frequency of the grating and the higher orders correspond to the harmonics of this frequency.

The dc term provides no information about the wavelength of the illumination. Information about the wavelength of the illuminating light can only be obtained by measuring the angular position of the first or higher order principal maxima.

Grating Spectrometer

The curves shown in Figure 10 display the separation of the principal maxima as a function of $\sin\theta_d$. The angular separation of principal maxima can be converted to a linear dimension by assuming a distance, R, from the graing to the observation plane. In the lower right-hand curve of Figure 10, a distance of 2 meters was assumed. Grating spectrometers are classified by the size in meters of R used in their design. The larger R, the easier it is to resolve wavelength differences. For example, a 1 meter spectrometer is a higher-resolution instrument than a 1/4 meter spectrometer.

The fact that the grating is finite in size causes each of the orders to have an angular width that limits the resolution with which the illuminating wavelength can be measured. To calculate the resolving power of the grating, we first determine the angular width of a principal maximum. This is accomplished by measuring the angular change, of the principal maximum's position, when β changes from $\beta = l\pi = m\pi/N$ to $\beta = (m+1)\pi/N$, i.e., $\Delta\beta = \pi/N$. Using the definition of β

$$\beta = \frac{kd\sin\theta_d}{2} \qquad [103]$$

$$\Delta\beta = \frac{\pi d\cos\theta_d\Delta\theta}{\lambda}$$

and the angular width is

$$\Delta\theta = \frac{\lambda}{Nd\cos\theta_d} \qquad [104]$$

The derivative of the grating formula gives

$$\Delta\lambda = \frac{d}{l}\cos\theta_d\Delta\theta \qquad [105]$$

$$\Delta\theta = \frac{l\Delta\lambda}{d\cos\theta_d}$$

Equating this spread in angle to eqn [104] yields

$$\frac{l\Delta\lambda}{d\cos\theta} = \frac{\lambda}{Nd\cos\theta} \qquad [106]$$

The resolving power of a grating is therefore

$$\frac{\lambda}{\Delta\lambda} = Nl \qquad [107]$$

The improvement of resolving power with N can be seen in Figure 10. A grating 2 inches wide and containing 15 000 grooves per inch would have a resolving power in second order ($l = 2$) of 6×10^4. At a wavelength of 600 nm this grating could resolve two waves, differing in wavelength by 0.01 nm.

The diffraction grating is limited by overlapping orders. If two wavelengths, λ and $\lambda+\Delta\lambda$, have successive orders that are coincident, then

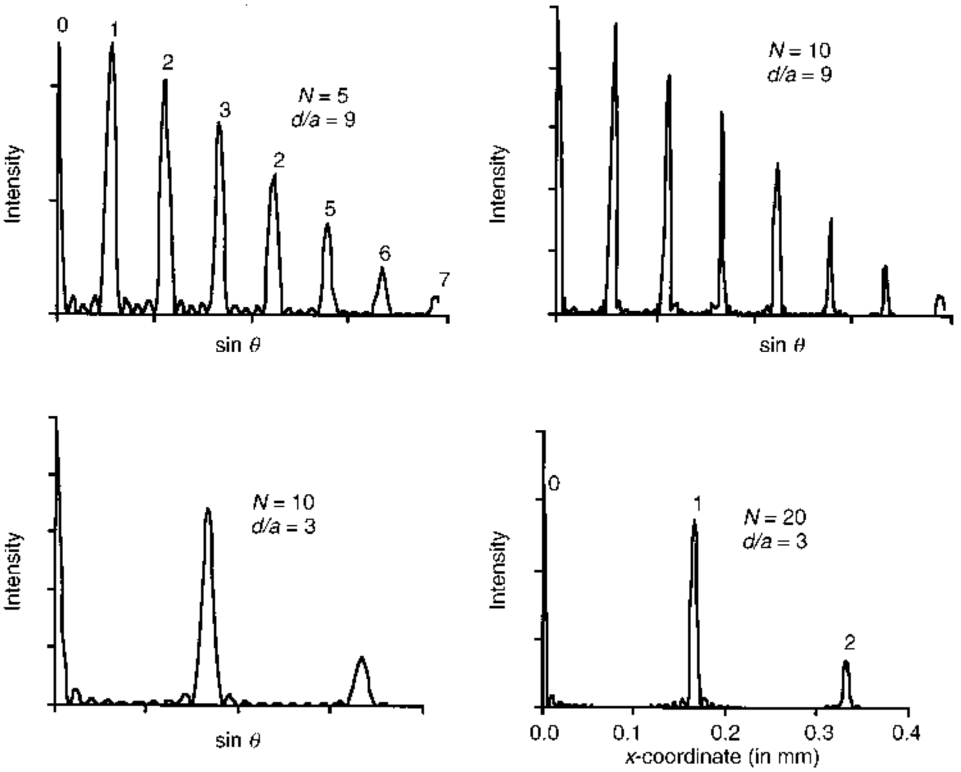

Figure 10 Decrease in the width of the principal maxima of a transmission grating with an increasing number of slits. The various principal maxima are called orders, numbering from zero, at the origin, out to as large as seven in this example. Also shown is the effect of different ratios, *d*/*a*, on the number of visible orders. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

from eqn [102]

$$(l+1)\lambda = l(\lambda + \Delta\lambda) \qquad [108]$$

The minimum wavelength difference for which this occurs is defined as the free spectral range of the diffraction grating

$$(\Delta\lambda)_{\mathrm{SR}} = \frac{\lambda}{l} \qquad [109]$$

Blazed Gratings

We have been discussing amplitude transmission gratings. Amplitude transmission gratings have little practical use because they waste light. The light loss is from a number of sources.

1. Light is diffracted simultaneously into both positive and negative orders (the positive and negative frequencies of the Fourier transform). The negative diffraction orders contain redundant information and waste light.
2. In an amplitude transmission grating, light is thrown away because of the opaque portions of the slit array.
3. The width of an individual aperture leads to a shape factor for a rectangular aperture of $\mathrm{sinc}^2\alpha = (\sin^2\alpha)/\alpha^2$, which modulates the grating factor and causes the amplitude of the orders to rapidly decrease. This can be observed in **Figure 10**, where the second order is very weak. Because of the loss in intensity at higher orders, only the first few orders ($l = 1, 2$ or 3) are useful. The shape factor also causes a decrease in diffracted light intensity with increasing wavelength for higher-order principal maxima.
4. The location of the maximum in the diffracted light, i.e., the angular position for which the shape factor is a maximum, coincides with the location of the principal maximum due to the zero-order interference. This zero-order maximum is independent of wavelength and is not of much use.

One solution to the problems created by transmission gratings would be the use of a grating that modified only the phase of the transmitted wave. Such gratings would operate using the same physical processes as a microwave phased array antenna, where the location of the shape factor's maximum is controlled by adding a constant phase shift to each antenna element. The construction of an optical transmission phase grating with a uniform phase variation across the aperture of the grating is very difficult. For this reason, a second approach, based on the use of reflection gratings, is the practical solution to the problems listed above.

By tilting the reflecting surface of each groove of a reflection grating, Figure 11, the position of the shape factor's maximum can be controlled. Problems 1, 3, and 4 are eliminated because the shape factor maximum is moved from the optical axis out to some angle with respect to the axis. The use of reflection gratings also removes Problem 2 because all of the incident light is reflected by the grating.

Robert Wood, in 1910, developed the technique of producing grooves of a desired shape in a reflective grating by shaping the diamond tool used to cut the grooves. Gratings, with grooves shaped to enhance their performance at a particular wavelength, are said to be blazed for that wavelength. The physical properties on which blazed gratings are based can be understood by using Figure 11. The groove faces can be treated as an array of mirror surfaces. The normal to each of the groove faces makes an angle θ_B with the normal to the grating surface. We can measure the angle of incidence and the angle of diffraction with respect to the grating normal or with respect to the groove normal, as shown in Figure 11. From Figure 11, we can write a relationship between the angles

$$\theta_i = \varphi_i - \theta_B \qquad -\theta_d = -\varphi_d + \theta_B \tag{110}$$

(The sign convention used defines positive angles as those measured in a counterclockwise rotation from the normal to the surface. Therefore, θ_B is a negative angle.) The blaze angle provides an extra degree of freedom that will allow independent adjustment of the angular location of the principal maxima of the grating factor and the zero-order, single-aperture, diffraction maximum. To see how this is accomplished, we must determine, first, the effect of off-axis illumination of a diffraction grating.

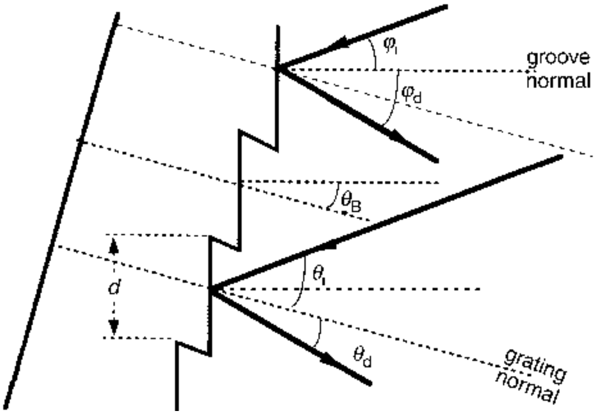

Figure 11 Geometry for a blazed reflection grating. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Off-axis illumination is easy to incorporate into the equation for the diffraction intensity from an array. To include the effect of an off-axis source, the phase of the illuminating wave is modified by changing the incident illumination from a plane wave of amplitude, E, traveling parallel to the optical axis, to a plane wave with the same amplitude, traveling at an angle θ_i to the optical axis: $\tilde{E}\, e^{-ikx \sin \theta_i}$. (Because we are interested only in the effects in a plane normal to the direction of propagation, we ignore the phase associated with propagation along the z-direction, $kz \cos \theta_i$.) The off-axis illumination results in a modification of the parameter for single-aperture diffraction from

$$\alpha = \frac{ka}{2} \sin \theta_d \tag{111}$$

to

$$\alpha = \frac{ka}{2} (\sin \theta_i + \sin \theta_d) \tag{112}$$

and for multiple aperture interference from

$$\beta = \frac{kd}{2} \sin \theta_d \tag{113}$$

to

$$\beta = \frac{kd}{2} (\sin \theta_i + \sin \theta_d) \tag{114}$$

The zero-order, single-aperture, diffraction peak occurs when $\alpha = 0$. If we measure the angles with respect to the groove face, this occurs when

$$\alpha = \frac{ka}{2} (\sin \varphi_i + \sin \varphi_d) = 0 \tag{115}$$

The angles are therefore related by

$$\sin \varphi_i = -\sin \varphi_d \tag{116}$$

$$\varphi_i = -\varphi_d$$

We see that the single-aperture, diffraction maximum (the shape factor's maximum) occurs at the same angle that reflection from the groove faces occurs. We can write this result in terms of the angles measured with respect to the grating normal

$$\theta_i = -(\theta_d + 2\theta_B) \tag{117}$$

The blaze condition requires the single aperture diffraction maximum to occur at the lth principal maximum, for wavelength λ_B. At that position

$$l\pi = \frac{2\pi d}{\lambda_B}(\sin\theta_i + \sin\theta_d) \qquad [118]$$

$$l\lambda_B = 2d\sin\tfrac{1}{2}(\theta_i + \theta_d)\cos\tfrac{1}{2}(\theta_i - \theta_d)$$

For the special geometrical configuration called the Littrow condition, where $\theta_i = \theta_d$, we find that [117] leads to the equation

$$l\lambda_B = 2d\sin\theta_B \qquad [119]$$

By adjusting the blaze angle, the single-aperture diffraction peak can be positioned on any order of the interference pattern. Typical blaze angles are between 15° and 30° but gratings are made with larger blaze angles.

See also

Diffraction: Fresnel Diffraction.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*. 3rd edn. Reading, MA: Addison-Wesley.

Jackson JD (1962) *Classical Electrodynamics*. New York: Wiley.

Klein MV (1970) *Optics*. New York: Wiley.

O'Neill EL (1963) *Introduction to Statistical Optics*. Reading, MA: Addison-Wesley.

Rossi B (1957) *Optics*. Reading, MA: Addison-Wesley.

Rubinowicz A (1984) Tahe Miyamoto–Wolf diffraction wave. In: Wolf E (ed.) *Progress in Optics*, vol. IV, p. 201. Amsterdam: North-Holland.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

Stone JM (1963) *Radiation and Optics*. New York: McGraw-Hill.

Fresnel Diffraction

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Fresnel was a civil engineer who pursued optics as a hobby, after a long day of road construction. Based on his own very high-quality observations of diffraction, Fresnel used the wave propagation concept developed by Christiaan Huygens (1629–1695) to develop a theoretical explanation of diffraction. We will use a descriptive approach to obtain the Huygens–Fresnel integral by assuming an aperture can be described by N pinholes which act as sources for Huygens' wavelets. The interference between these sources will lead to the Huygens–Fresnel integral for diffraction.

Huygens' principle views wavefronts as the product of wavelets from various luminous points acting together. To apply Huygens' principle to the propagation of light through an aperture of arbitrary shape, we need to develop a mathematical description of the field from an array of Huygens' sources filling the aperture. We will begin by obtaining the field from a pinhole that is illuminated by a plane wave,

$$\tilde{\mathrm{E}}_i(\mathbf{r}, t) = \tilde{\mathrm{E}}_i(\mathbf{r})e^{i\omega t}$$

Following the lead of Fresnel, we will use theory of interference to combine the fields from two pinholes and then generalize to N pinholes. Finally by letting the areas of each pinhole approach an infinitesimal value, we will construct an arbitrary aperture of infinitesimal pinholes. The result will be the Huygens–Fresnel integral.

We know that the wave obtained after propagation through an aperture must be a solution of the wave equation,

$$\nabla^2\tilde{\mathrm{E}} = \mu\varepsilon\frac{\partial^2\tilde{\mathrm{E}}}{\partial t^2}$$

We will be interested only in the spatial variation of the wave so we need only look for solutions of the Helmholtz equation

$$(\nabla^2 + k^2)\tilde{\mathrm{E}} = 0$$

The problem is further simplified by replacing this vector equation with a scalar equation,

$$(\nabla^2 + k^2)\tilde{E}(x, y, z) = 0$$

This replacement is proper for those cases where $\mathbf{n}E(x,y,z)$ [where $\mathbf{n}$ is a unit vector] is a solution of the vector Helmholtz equation. In general, we cannot substitute $\mathbf{n}E$ for the electric field $\mathbf{E}$ because of Maxwell's equation

$$\nabla\cdot\mathbf{E} = 0$$

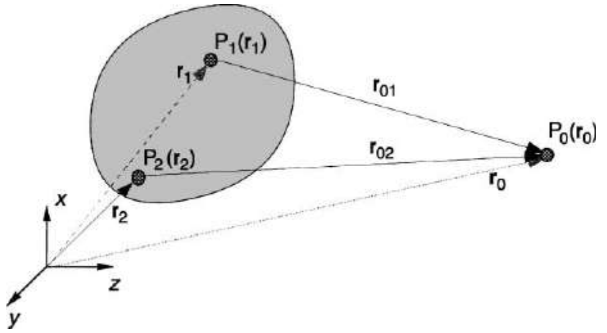

Figure 1 Geometry for application of Huygens' principle to two pinholes. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

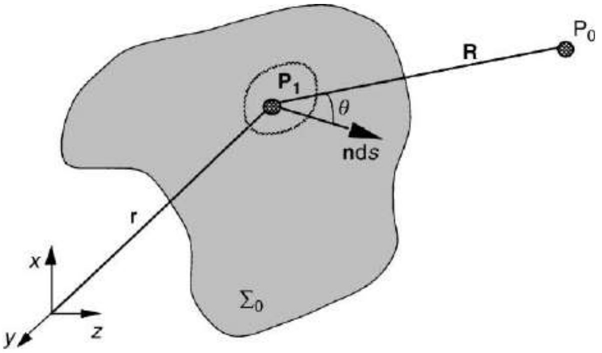

Figure 2 The geometry for calculating the field at P_0 using [9]. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Rather than working with the magnitude of the electric field, we should use the scalar amplitude of the vector potential. We will neglect this complication and assume the scalar, E, is a single component of the vector field, $\mathbf{E}$. A complete solution would involve a scalar solution for each component of $\mathbf{E}$.

The pinhole is illuminated by a plane wave, and the wave that leaves the pinhole will be a spherical wave which is written in complex notation as

$$\tilde{E}(\mathbf{r})e^{i\omega t} = A\frac{e^{-i\delta}e^{-ik\cdot r}}{r}e^{i\omega t}$$

The complex amplitude

$$\tilde{E}(\mathbf{r}) = A\frac{e^{-i\delta}e^{-ik\cdot r}}{r} \qquad [1]$$

is a solution of the Helmholtz equation. The field at P_0, in Figure 1, from two pinholes: one at P_1, located a distance $\mathbf{r}_{01} = |\mathbf{r}_0 - \mathbf{r}_1|$ from P_0, and one at P_2, located a distance $\mathbf{r}_{02} = |\mathbf{r}_0 - \mathbf{r}_2|$ from P_0, is a generalization of Young's interference. The complex amplitude is

$$\tilde{E}(\mathbf{r}_0) = \frac{A_1}{r_{01}}e^{-ik\cdot r_{01}} + \frac{A_2}{r_{02}}e^{-ik\cdot r_{02}} \qquad [2]$$

We have incorporated the phase δ_1 and δ_2 into the constants A_1 and A_2 to simplify the equations.

The light emitted from the pinholes is due to a wave, E_i, incident onto the screen from the left. The units of E_i are per unit area so to obtain the amount of light passing through the pinholes, we must multiply E_i by the areas of the pinholes. If $\Delta\sigma_1$ and $\Delta\sigma_2$ are the areas of the two pinholes, respectively, then

$$A_1 \propto \tilde{E}_i(\mathbf{r}_1)\Delta\sigma_1 \qquad A_2 \propto \tilde{E}_i(\mathbf{r}_2)\Delta\sigma_2$$

$$\tilde{E}(\mathbf{r}_0) = C_1\frac{\tilde{E}_i(\mathbf{r}_1)}{r_{01}}e^{-ik\cdot r_{01}}\Delta\sigma_1 + C_2\frac{\tilde{E}_i(\mathbf{r}_2)}{r_{02}}e^{-ik\cdot r_{02}}\Delta\sigma_2 \qquad [3]$$

where C_i is the constant of proportionality. The constant will depend on the angle that $\mathbf{r}_{0i}$ makes with the normal to $\Delta\sigma_i$. This geometrical dependence arises from the fact that the apparent area of the pinhole decreases as the observation angle approaches 90°.

We can generalize [3] to N pinholes

$$\tilde{E}(\mathbf{r}_0) = \sum_{j=1}^{N} C_j\frac{\tilde{E}_i(\mathbf{r}_j)}{r_{0j}}e^{-ik\cdot r_{0j}}\Delta\sigma_j \qquad [4]$$

The pinhole's diameter is assumed to be small compared to the distances to the viewing position but large compared to a wavelength. In the limit as $\Delta\sigma_j$ goes to zero, the pinhole becomes a Huygens source. By letting N become large, we can fill the aperture with these infinitesimal Huygens' sources and convert the summation to an integral.

It is in this way that we obtain the complex amplitude, at the point P_0, see Figure 2, from a wave exiting an aperture, Σ_0, by integrating over the area of the aperture. We replace $\mathbf{r}_{0j}$, in [4], by $\mathbf{R}$, the position of P_1, the infinitesimal Huygens' source of area, ds, measured with respect to the observation point, P_0. We also replace $\mathbf{r}_j$ by $\mathbf{r}$, the position of the infinitesimal area, ds, with respect to the origin of the coordinate system. The discrete summation [4] becomes the integral

$$\tilde{E}(\mathbf{r}_0) = \iint_A C(\mathbf{r})\frac{\tilde{E}_i(\mathbf{r})}{R}e^{-ik\cdot R}ds \qquad [5]$$

This is the Fresnel integral. The variable $C(\mathbf{r})$ depends upon θ, the angle between $\mathbf{n}$, the unit vector normal to the aperture, and $\mathbf{R}$, shown in Figure 2. We now need to determine how to treat $C(\mathbf{r})$.

The Obliquity Factor

When using Huygens' principle, a problem arises with the spherical wavelet produced by each

Huygens' source. Part of the spherical wavelet propagates backward and results in an envelope propagating toward the light source, but such a wave is not observed in nature. Huygens neglected this problem. Fresnel required the existence of an obliquity factor to cancel the backward wavelet but was unable to derive its functional form. When Kirchhoff placed the theory of diffraction on a firm mathematical foundation, the obliquity factor was generated quite naturally in the derivation. In this intuitive derivation of the Huygens–Fresnel integral, we will present an argument for treating the obliquity factor as a constant at optical wavelengths. We will then derive the constant value it must be assigned.

The obliquity factor, $C(\mathbf{r})$, in [5], causes the amplitude per unit area of the transmitted light to decrease as the viewing angle increases. This is a result of a decrease in the effective aperture area with viewing angle. When Kirchhoff applied Green's theorem to the scalar diffraction problem, he found the obliquity factor to have an angular dependence given by

$$\frac{\cos(\hat{\mathbf{n}},\mathbf{R}) - \cos(\hat{\mathbf{n}},\mathbf{r}_s)}{2} \qquad [6]$$

which includes the geometrical effect of the incident wave arriving at the aperture, at an angle with respect to the normal to the aperture from a source located at a position $\mathbf{r}_s$, with respect to the aperture. If the source is at infinity, then the incident wave can be treated as a plane wave, incident normal to the aperture, and we may simplify the angular dependence of the obliquity factor to

$$\frac{1+\cos\theta}{2}$$

where $\theta \rightarrow (\hat{\mathbf{n}},\mathbf{R})$ is the angle between the normal to the aperture and the vector $\mathbf{R}$. This is the configuration shown in Figure 2. The obliquity factor provides an explanation of why it is possible to ignore the backward-propagating wave that occurs in application of Huygens' principle. For the backward wave, $\theta = \pi$, and the obliquity factor is zero.

The obliquity factor increases the difficulty of working with the Fresnel integral and it is to our benefit to be able to treat it as a constant. We can neglect the angular contribution of the obliquity factor by making an assumption about the resolving power of an optical system operating at visible wavelengths.

Assume we are attempting to resolve two stars that produce plane waves at the aperture of a telescope with an aperture diameter, a. The wavefronts from the two stars make an angle ψ with respect to each other, Figure 3:

$$\tan\psi \approx \psi = \frac{\Delta x}{a}$$

The smallest angle, ψ, that can be measured is determined by the smallest length, Δx, that can be measured. We know we can measure a fraction of wavelength with an interferometer but, without an interferometer, we can measure a length no smaller than λ, leading to the assumption that $\Delta x \geq \lambda$. This reasoning leads to the assumption that the smallest angle we can measure is

$$\psi \geq \frac{\lambda}{a} \qquad [7]$$

The resolution limit established by the above reasoning places a limit on the minimum separation that can be produced at the back focal plane of the telescope. The minimum distance on the focal plane between the images of star 1 and 2 is given by

$$d = f\psi \qquad [8]$$

From [7]

$$d = \frac{\lambda f}{a}$$

The resolution limit of the telescope can also be expressed in terms of the cone angle produced when the incident plane wave is focused on the back focal plane of the lens. From the geometry of Figure 3, the cone angle is given by

$$\tan\theta = \frac{a}{2f}$$

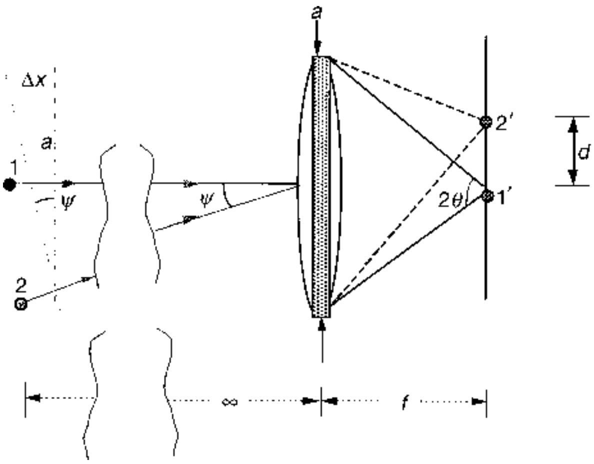

Figure 3 Telescope resolution. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The separation between the stars 1 and 2 at the back focal plane of the lens in Figure 3 is thus

$$d = \frac{\lambda}{2\tan\theta} \qquad [9]$$

If we assume that the minimum d is 3λ (this is four times the resolution of a typical photographic film at visible wavelengths), then the largest obliquity angle that should be encountered in a visible imaging system is

$$\tan\theta = \frac{\lambda}{2d} = \frac{\lambda}{6\lambda} = \tfrac{1}{6} = 0.167$$

$$\theta = 9.5° \approx 10°$$

yielding a value for $\cos\theta = 0.985$ and $(1 + \cos\theta)/2 = 0.992$. The obliquity factor has only changed by 0.8% over the angular variation of 0° to 10°. The obliquity factor as a function of angle is shown in Figure 4 for an incident plane wave. The obliquity factor undergoes a 10% change when θ varies from 0° to about 40°; therefore, the obliquity factor can safely be treated as a constant in any optical system that involves angles less than 40°.

While we have shown that it is safe to ignore the variability of C, we still have not assigned a value to the obliquity factor. To find the proper value for C, we will compare the result obtained using [5] with the result predicted by using geometric optics.

We illuminate the aperture Σ_0 in Figure 5 with a plane wave of amplitude α, traveling parallel to the z-axis. Geometrical optics (equivalently the propagation of an infinite plane wave) predicts a field at P_0, on the z-axis, a distance z_0 from the aperture, given by

$$\tilde{E}_{\text{geom}} = \alpha e^{-ikz_0} \qquad [10]$$

The area of the infinitesimal at P_1 (the Huygens' source) is

$$ds = r\,dr\,d\phi$$

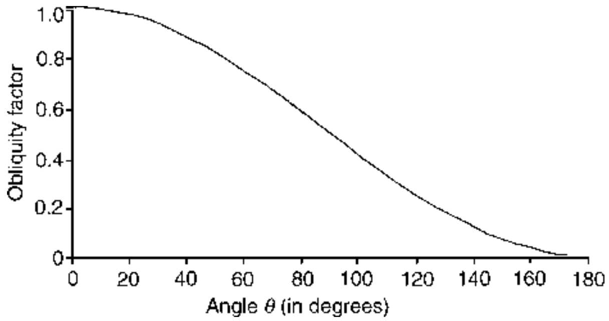

Figure 4 The obliquity factor as a function of the angle defined in eqn [6] for an incident plane wave. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

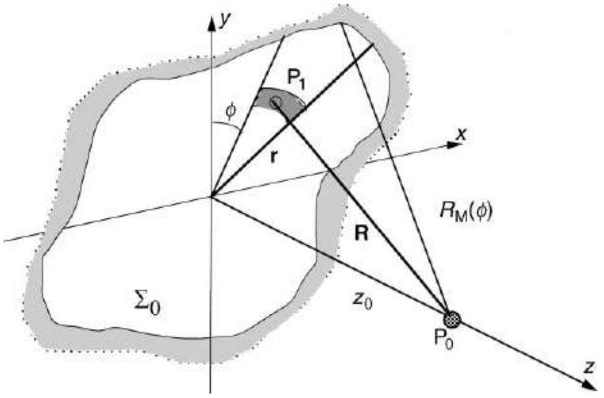

Figure 5 Geometry for evaluating the constant in the Fresnel integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Based on our previous argument, the obliquity factor can be treated as a constant, C, that can be removed from under the integral. The incident wave is a plane wave whose value at $z = 0$ is $E(r) = \alpha$. Using these parameters, the Fresnel integral [5] can be written as

$$\tilde{E}(z_0) = C\alpha \int\!\!\int \frac{e^{-ik\cdot R}}{R}\, r\,dr\,d\phi \qquad [11]$$

From the geometry in Figure 5, the distance from the Huygens source to the observation point is

$$z_0^2 + r^2 = R^2$$

where z_0 is a constant equal to the distance from the observation plane to the aperture plane. The variable of integration can be written in terms of R:

$$r\,dr = R\,dR$$

The limits of integration over the aperture extend from $R = z_0$ to the maximum value of R, $R = R_M(\phi)$.

$$\tilde{E}(z_0) = C\alpha \int_0^{2\pi} \int_{z_0}^{R_M(\phi)} e^{-ikR}\,dR\,d\phi \qquad [12]$$

The integration over R can now be carried out to yield

$$\tilde{E}(z_0) = \underbrace{\frac{C\alpha}{ik} e^{-ikz_0} \int_0^{2\pi} d\phi}_{\text{Geometrical Optics}} - \underbrace{\frac{C\alpha}{ik} \int_0^{2\pi} e^{-ikR_M(\phi)}\,d\phi}_{\text{Diffraction}} \qquad [13]$$

The first integration in [13] is easy to perform and contains the amplitude at the observation point due to geometric optics. The second term may be interpreted as interference of the waves diffracted by the boundary of the aperture. An equivalent statement is that the second term is the interference of the waves scattered from the aperture's boundary, an interpretation of diffraction that was first suggested by Young.

The aperture is irregular in shape, at least on the scale of a wavelength, thus in general, kR_M will vary over many multiples of 2π as we integrate around the aperture. For this reason, we should be able to ignore the second integral in [13] if we confine our attention to the light distribution on the z-axis. After neglecting the second term, we are left with only the geometrical optics component of [13]:

$$\tilde{E}(z_0) = \frac{2\pi C}{\mathrm{i}k}\alpha \mathrm{e}^{-\mathrm{i}kz_0} \qquad [14]$$

For [14] to agree with the prediction of geometric optics [10] the constant C must be equal to

$$C = \frac{\mathrm{i}k}{2\pi} = \frac{\mathrm{i}}{\lambda} \qquad [15]$$

The Huygens–Fresnel integral can be written, using the value for C just derived,

$$\tilde{E}(\mathbf{r}_0) = \frac{\mathrm{i}}{\lambda}\int\!\!\int_{\Sigma} \frac{\tilde{E}_{\mathrm{i}}(\mathbf{r})}{R}\,\mathrm{e}^{-\mathrm{ik\cdot R}}\,\mathrm{d}s \qquad [16]$$

The job of calculating diffraction has only begun with the derivation of the Huygens–Fresnel integral [16]. Rigorous solutions exist for only a few idealized obstructions. To allow discussion of the general properties of diffraction, it is necessary to use approximate solutions. To determine what type approximations we can make let's look at the light propagation path shown in Figure 6.

For the second term in [13] to contribute to the field at point P_0, in Figure 6, the phase of the exponent that we neglected in [13] must not vary over 2π when the integration is performed over the aperture, i.e.,

$$\Delta = \mathbf{k}\cdot(\mathbf{R}_1 - \mathbf{R}'_1 + \mathbf{R}_2 - \mathbf{R}'_2) < 2\pi$$

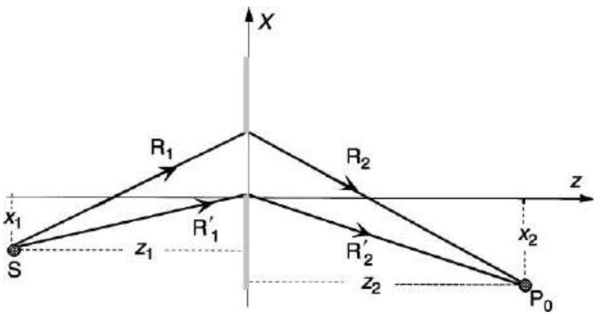

Figure 6 One-dimensional slit used to establish Fresnel approximations. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

From the geometry of Figure 6, the two paths from the source, S, to the observation point, P_0, are

$$R_1 + R_2 = \sqrt{z_1^2 + (x_1 + b)^2} + \sqrt{z_2^2 + (x_2 + b)^2}$$

$$R'_1 + R'_2 = \sqrt{z_1^2 + x_1^2} + \sqrt{z_2^2 + x_2^2}$$

By assuming that the aperture, of width b, is small compared to the distances z_1 and z_2, the difference between these distances, Δ, can be rewritten as a binomial expansion of a square root

$$\sqrt{1+b} = 1 + \frac{b}{2} - \frac{b^2}{8} + \dots \qquad [17]$$

thus

$$\Delta \approx k\left(\frac{x_1}{z_1} + \frac{x_2}{z_2}\right)b + \frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)b^2 + \dots \qquad [18]$$

If we assume that the second term of this expansion is small, i.e.,

$$\frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)b^2 \ll 2\pi$$

$$\frac{b^2}{2}\,\frac{1}{z_1} + \frac{1}{z_2} < \lambda$$

we see that the phase of the wavefront in the aperture is assumed to have a quadratic dependence upon aperture coordinates. Diffraction predicted by this assumption is called Fresnel diffraction.

To discover the details of the approximation consider Figure 7. In the geometry of Figure 7, the Fresnel integral [16] becomes

$$\tilde{E}_{\mathrm{P}_0} = \frac{\mathrm{i}\alpha}{\lambda}\int\!\!\int_{\Sigma} f(x,y)\frac{\mathrm{e}^{-\mathrm{ik\cdot(R+R')}}}{RR'}\,\mathrm{d}x\,\mathrm{d}y \qquad [19]$$

As we mentioned in our discussion of [13] the integral that adds diffractive effects to the geometrical optics field is nonzero only when the phase of the integrand is stationary. For Fresnel diffraction, we can insure that the phase is nearly constant if R does not differ appreciably from Z, or R' from Z'. This is equivalent to stating that only the wave in the aperture, around a point in Figure 7 labeled **S**, called the stationary point, will contribute to E_{p}. The stationary point, **S**, is the point in the aperture plane where the line, connecting the source and observation positions, intersects the plane. Physically, only light propagating over paths nearly equal to the path predicted by geometrical optics (obtained from Fermat's principle) will contribute to E_{P_0}.

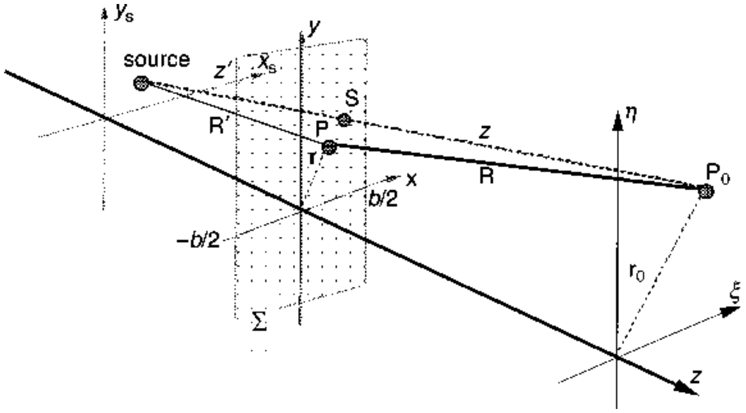

Figure 7 Geometry for Fresnel diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The geometry of Figure 7 allows the distances R and R' to be written as

$$R^2 = (x - \xi)^2 + (y - \eta)^2 + Z^2 \qquad [20]$$

$$R'2 = (x_s - x)^2 + (y_s - y)^2 + Z'2$$

To solve these expressions for R and R', we apply the binomial expansion [17] and retain the first two terms:

$$R + R' \approx Z + Z' + \left[\frac{(x-\xi)^2}{2Z} + \frac{(x_s - x)^2}{2Z'} \right] + \left[\frac{(y-\eta)^2}{2Z} + \frac{(y_s - y)^2}{2Z'} \right] \qquad [21]$$

In the denominator of [19], R is replaced by Z and R' by Z'. (By making this replacement, we are implicitly assuming that the amplitudes of the spherical waves are not modified by the differences in propagation distances over the area of integration. This is a reasonable approximation because the two propagation distances are within a few hundred wavelengths of each other). In the exponent, we must use [21], because the phase changes by a large fraction of a wavelength, as we move from point to point in the aperture.

If the wave, incident on the aperture, were a plane wave we can assume $R' = \infty$ and [19] becomes

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda} \frac{e^{-ikZ}}{Z} \int\int_{\Sigma} f(x, y) e^{-\frac{ik}{2Z}[(x-\xi)^2 + (y-\eta)^2]} dx\, dy \qquad [22]$$

The physical interpretation of [22] states that when a plane wave illuminates the obstruction, the field at point P_0 is a spherical wave, originating at the aperture, a distance Z away from P_0:

$$\frac{e^{-ikZ}}{Z}$$

The amplitude and phase of this spherical wave are modified by the integral containing a quadratic phase dependent on the obstruction's spatial coordinates.

By defining three new parameters,

$$\rho = \frac{ZZ'}{Z + Z'} \quad \text{or} \quad \frac{1}{\rho} = \frac{1}{Z} + \frac{1}{Z'} \qquad [23a]$$

$$x_0 = \frac{Z'\xi + Zx_s}{Z + Z'} \qquad [23b]$$

$$y_0 = \frac{Z'\eta + Zy_s}{Z + Z'} \qquad [23c]$$

the more general expression for Fresnel diffraction of a spherical wave can be placed in the same format as [22]. The newly defined parameters x_0 and y_0 are the coordinates, in the aperture plane, of the stationary point, **S**. The parameters in [23] can be used to express, after some manipulation, the spatial dependence of the phase, in the integrand of [19]

$$R + R' = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z + Z')} + \left[\frac{(x - x_0)^2 + (y - y_0)^2}{2\rho} \right]$$

A further simplification to the Fresnel diffraction integral can be made by obtaining an expression for the distance, D, between the source and the observation point:

$$D = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z + Z')} \qquad [24]$$

We use this definition of D to also write

$$\frac{1}{ZZ'} = \frac{Z+Z'}{ZZ'}\frac{1}{Z+Z'} \approx \frac{1}{\rho D}$$

Using the parameters we have just defined, we may rewrite [19] as

$$\tilde{E}_{\mathrm{P}_0} = \frac{\mathrm{i}\alpha}{\lambda\rho D}\mathrm{e}^{-\mathrm{i}kD}\iint f(x,y)\mathrm{e}^{-\frac{\mathrm{i}k}{2\rho}\left[(x-x_0)^2+(y-y_0)^2\right]}\mathrm{d}x\,\mathrm{d}y \tag{25}$$

With the use of the variables defined in [23], the physical significance of the general expression of the Huygens–Fresnel integral can be understood in an equivalent fashion as was [22]. At point P_0, a spherical wave

$$\frac{\mathrm{e}^{-\mathrm{i}kD}}{D}$$

originating at the source, a distance D away, is observed, if no obstruction are present. Because of the obstruction, the amplitude and phase of the spherical wave are modified by the integral in [25]. This correction to the predictions of geometrical optics is called Fresnel diffraction.

Mathematically, [25] is an application of the method of stationary phase, a technique developed in 1887 by Lord Kelvin to calculate the form of a boat's wake. The integration is nonzero only in the region of the critical point we have labeled **S**. Physically, the light distribution, at the observation point, is due to wavelets from the region around **S**. The phase variations of light, coming from other regions in the aperture, are so rapid that the value of the integral over those spatial coordinates is zero. The calculation of the integral for Fresnel diffraction is complicated because, if the observation point is moved, a new integration around a different stationary point in the aperture must be performed.

Rectangular Apertures

If the aperture function, $f(x,y)$, is separable in the spatial coordinates of the aperture, then we can rewrite [25] as

$$\tilde{E}_{\mathrm{P}_0} = \frac{\mathrm{i}\alpha}{\lambda\rho D}\mathrm{e}^{-\mathrm{i}kD}\int_{-\infty}^{\infty} f(x)\mathrm{e}^{-\mathrm{i}g(x)}dx\int_{-\infty}^{\infty} f(y)\mathrm{e}^{-\mathrm{i}g(y)}\mathrm{d}y$$

$$\tilde{E}_{\mathrm{P}_0} = A[C(x) - \mathrm{i}S(x)][C(y) - \mathrm{i}S(y)]$$

A represents the spherical wave from the source, a distance D away,

$$A = \frac{\mathrm{i}\alpha}{2D}\mathrm{e}^{-\mathrm{i}kD}$$

If we treat the aperture function as a simple constant, C and S are integrals of the form

$$C(x) = \int_{x_1}^{x_2}\cos[g(x)]\mathrm{d}x \tag{26a}$$

$$S(x) = \int_{x_1}^{x_2}\sin[g(x)]\mathrm{d}x \tag{26b}$$

The integrals, $C(x)$ and $S(x)$, have been evaluated numerically and are found in tables of Fresnel integrals. To use the tabulated values for the integrals, [26] must be written in a general form

$$C(w) = \int_0^w \cos\left(\frac{\pi u^2}{2}\right)\mathrm{d}u \tag{27}$$

$$S(w) = \int_0^w \sin\left(\frac{\pi u^2}{2}\right)\mathrm{d}u \tag{28}$$

The variable u is an aperture coordinate, measured relative to the stationary point, $\mathbf{S}(x_0,y_0)$, in units of

$$\sqrt{\frac{\lambda\rho}{2}}$$

$$u = \sqrt{\frac{2}{\lambda\rho}}(x - x_0) \text{ or } u = \sqrt{\frac{2}{\lambda\rho}}(y - y_0) \tag{29}$$

The parameter w in [27] and [28] specifies the location of the aperture edge relative to the stationary point **S**. The parameter w is calculated through the use of [29] with x and y replaced by the coordinates of the aperture edge.

The Cornu Spiral

The plot of $S(w)$ versus $C(w)$, shown in Figure 8, is called the Cornu spiral in honor of M. Alfred Cornu (1841–1902) who was the first to use this plot for graphical evaluation of the Fresnel integrals.

To use the Cornu spiral, the limits w_1 and w_2 of the aperture are located along the arc of the spiral. The length of the straight line segment drawn from w_1 to w_2 gives the magnitude of the integral. For example, if there were no aperture present, then, for the x-dimension, $w_1 = -\infty$ and $w_2 = \infty$. The length of the line segment from the point $(-1/2,1/2)$ to $(1/2,1/2)$ would be the value of E_{P_0}, i.e., $\sqrt{2}$. An identical value is obtained for the y-dimension, so that

$$\tilde{E}_{\mathrm{P}_0} = 2A = \frac{\alpha\mathrm{e}^{-\mathrm{i}kD}}{D}$$

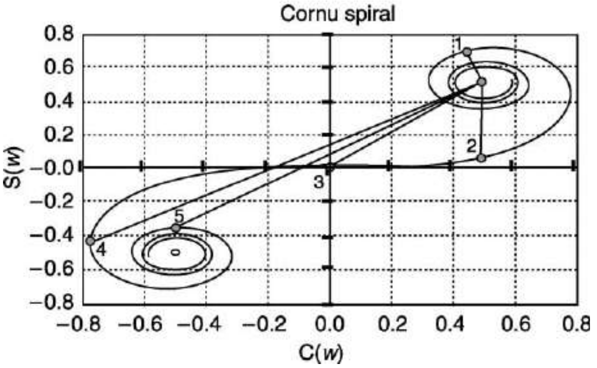

Figure 8 Cornu spiral obtained by plotting the values for *C*(*w*) versus *S*(*w*) from a table of Fresnel integrals. The numbers indicated along the arc of the spiral are values of *w*. The points 1–5 correspond to observation points shown in Figure 9 and are used to calculate the light intensity diffracted by a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

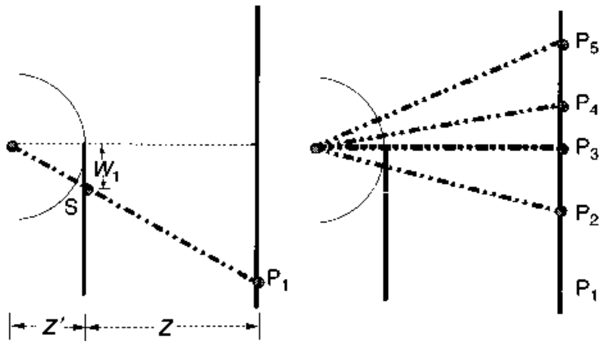

Figure 9 A geometrical construct to determine w_1 for the stationary point S associated with five different observation points. The values of *w* are then used in Figure 8 to calculate the intensity of light diffracted around a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Table 1 Fresnel integrals for straight edge

w_1	$C(w_1)$	$S(w_1)$	I_P/I_0	P_j
∞	0.5	0.5	0	
2.0	0.4882	0.3434	0.01	
1.5	0.4453	0.6975	0.021	P_1
1.0	0.7799	0.4383	0.04	
0.5	0.4923	0.0647	0.09	P_2
0	0	0	0.25	P_3
−0.5	−0.4923	−0.0647	0.65	
−1.0	−0.7799	−0.4383	1.26	P_4
−1.2	−0.7154	−0.6234	1.37	
−1.5	−0.4453	−0.6975	1.16	
−2.0	−0.4882	−0.3434	0.84	P_5
−2.5	−0.4574	−0.6192	1.08	
−∞	−0.5	−0.5	1.0	

This is the spherical wave that is expected when no aperture is present.

As an example of the application of the Fresnel integrals, assume the diffracting aperture is an infinitely long straight edge reducing the problem to one dimension:

$$f(x,y) = \begin{Bmatrix} 1 & x \geq x_1 \\ 0 & x < x_1 \end{Bmatrix}$$

The value of E_{P_0} in the y-dimension is from above $\sqrt{2}$ since $w_1 = -\infty$ and $w_2 = \infty$. In the x-dimension, the value of w at the edge is

$$w_{x_1} = \sqrt{\frac{2}{\lambda\rho}}(x_1 - x_0)$$

We will assume that the straight edge blocks the negative half-plane so that $x_1 = 0$. The straight edge is treated as an infinitely wide slit with one edge located at infinity so that $w_2 = \infty$ in the upper half-plane. When the observation point is moved, the coordinate x_0 of S changes and the origin of the aperture's coordinate system moves. New values for the w's must therefore be calculated for each observation point.

Figure 9 shows the assumed geometry. A set of observation points, P_1–P_5, on the observation screen are selected. The origin of the coordinate system in the aperture plane is relocated to a new position (given by the position of S) when a new observation point is selected. The value of w_1, the position of the edge with respect to S, must be recalculated for each observation point. The distance from the origin to the straight edge, w_1, is positive for P_1 and P_2, zero for P_3, and negative for P_4 and P_5.

Figure 8 shows the geometrical method used to calculate the intensity values at each of the observation points in Figure 9 using the Cornu spiral. The numbers labeling the straight line segments in Figure 8 are associated with the labels of the observation points in Figure 9.

To obtain an accurate calculation of Fresnel diffraction from a straight edge, a table of Fresnel integrals provides the input to the following equation

$$I_P = I_0\left\{\left[\tfrac{1}{2} - C(w_1)\right]^2 + \left[\tfrac{1}{2} - S(w_1)\right]^2\right\}$$

where $I_0 = 2A^2$. Table 1, shows the values extracted from the table of Fresnel integrals to find the relative intensity at various observation points. The result obtained by using either method for calculating the light distribution in the observation plane, due to the straight edge in Figure 9, is plotted in Figure 10. The relative intensities at the observation points

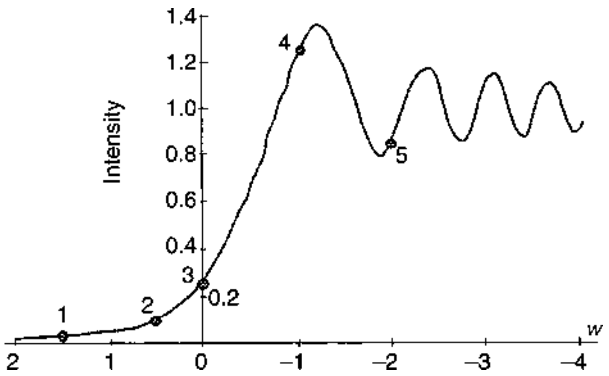

Figure 10 Light diffracted by a straight edge with its edge located at $w = 0$. The points labeled 1–5 were found using the construction in Figure 9 to obtain w. This was then used in Figure 8 to find a position on the Cornu spiral. The length of the lines shown in Figure 8 from the (1/2,1/2) point to the numbered positions led to the intensities shown. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

depicted in Figure 9 are labeled on the diffraction curve of Figure 10.

Fresnel Zones

Fresnel used a geometrical construction of zones to evaluate the Huygens–Fresnel integral. The Fresnel zone is a mathematical construct, serving the role of a Huygens source in the description of wave propagation. Assume that at time t, a spherical wavefront from a source at P_1 has a radius of R'. To determine the field at the observation point, P_0, due to this wavefront, a set of concentric spheres of radii, $Z, Z + (\lambda/2), Z + 2(\lambda/2), \ldots, Z + j(\lambda/2), \ldots$ are constructed, where Z is the distance from the wavefront to the observation point on the line connecting P_1 and P_0 (see Figure 11). These spheres divide the wavefront into a number of zones, $\zeta_1, \zeta_2, \ldots, \zeta_j \ldots$, called Fresnel zones, or half-period zones.

We treat each zone as a circular aperture illuminated from the left by a spherical wave of the form

$$\tilde{E}_j(\mathbf{R}') = \frac{A\mathrm{e}^{-\mathrm{ik}\cdot\mathrm{R}'}}{R'} = \frac{A\mathrm{e}^{-ikR'}}{R'}$$

R' is the radius of the spherical wave. The field at P_0 due to the jth zone is obtained by using [5]

$$\tilde{E}_j(\mathrm{P}_0) = \frac{A}{R'}\mathrm{e}^{-\mathrm{ik}\cdot\mathrm{R}'}\iint_{\zeta_j} C(\varphi)\frac{\mathrm{e}^{-\mathrm{ik}\cdot\mathrm{R}}}{R}\mathrm{d}s \qquad [30]$$

For the integration over the jth zone, the surface element is

$$\mathrm{d}s = R'2\sin\theta\mathrm{d}\theta\,\mathrm{d}\phi \qquad [31]$$

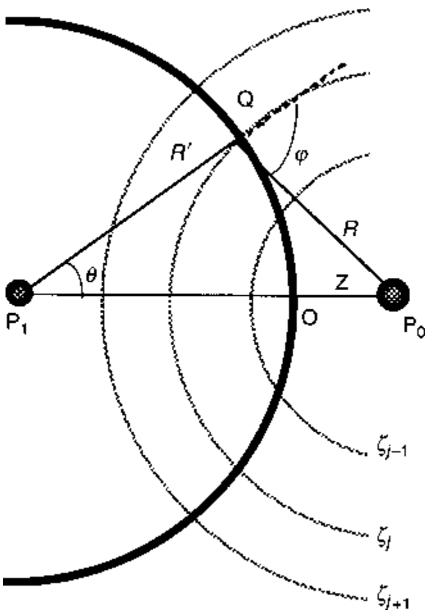

Figure 11 Construction of Fresnel zones observed from a position P_0 on a spherical wave originating from a source at point P_1. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

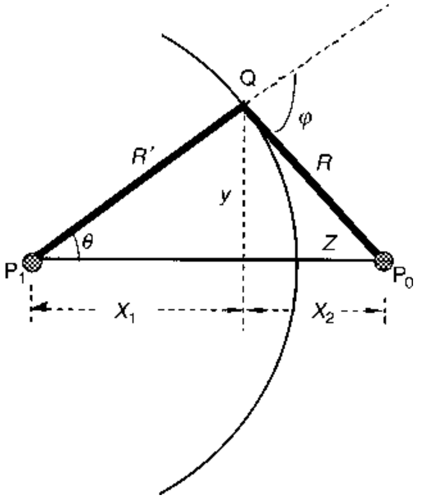

Figure 12 Geometry for finding the relationship between R and R' yielding eqn [32]. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The limits of integration extend over the range

$$Z + (j-1)\tfrac{\lambda}{2} \le R \le Z + j\tfrac{\lambda}{2}$$

The variable of integration is R; thus, a relationship between R' and R must be found. This is accomplished by using the geometrical construction shown in Figure 12.

A perpendicular is drawn from the point Q on the spherical wave to the line connecting P_1 and P_0, in Figure 12. The distance from the source to the observation point is $x_1 + x_2$ and the distance from the source to the plane of the zone is the radius of the incident spherical wave, R'. The distance from P_1 to P_0 can be written

$$x_1 + x_2 = R' + Z$$

The distance from the observation point to the zone is

$$R^2 = x_2^2 + y^2 = [(R' + Z) - x_1]^2 + (R' \sin\theta)^2$$
$$= R'2 + (R' + Z)^2 - 2R'(R' + Z)\cos\theta$$

The derivative of this expression yields

$$R\,\mathrm{d}R = R'(R' + Z)\sin\theta\,\mathrm{d}\theta \qquad [32]$$

Substituting [32] into [31] gives

$$\mathrm{d}s = \frac{R'}{R' + Z} R\mathrm{d}R\mathrm{d}\phi \qquad [33]$$

The integration over ϕ is accomplished by rotating the surface element about the P_1P_0 axis. After integrating over ϕ between the limits of 0 and 2π, we obtain

$$\tilde{E}_j(P_0) = \frac{2\pi A}{R' + Z} e^{-ik\cdot R'} \int_{Z+(j-1)\frac{\lambda}{2}}^{Z+\frac{j\lambda}{2}} e^{-ik\cdot R} C(\varphi)\mathrm{d}R \qquad [34]$$

We assume that R', $Z \gg \lambda$ so that the obliquity factor is a constant over a single zone, i.e., $C(\varphi) = C_j$. The obliquity factor is not really a constant as we earlier assumed but rather a very slowly varying parameter of j, changing by less than two parts in 10^4 when 'j' changes by 500 (for $Z = 1$ meter and $\lambda = 500$ nm). In fact it only changes by 2 parts in 10^{-7} across one zone. We are safe in applying the assumption that the obliquity factor is a constant over a zone and this allows the integral in [34] to be calculated:

$$\tilde{E}_j(P_0) = \frac{2\pi i C_j A}{k(R' + Z)} e^{-ik(R'+Z)} e^{-ikj\frac{\lambda}{2}} \left[1 - e^{-ik\frac{\lambda}{2}}\right] \qquad [35]$$

Using the identity $k\lambda = 2\pi$ and the definition for the distance between the source and observation point [24] modified for this geometry, $D = R' + Z$ [35] can be simplified to

$$\tilde{E}_j(P_0) = 2i\lambda(-1)^j \frac{C_j A}{D} e^{-ikD} \qquad [36]$$

The physical reasons for the behavior predicted by [36] are quite easy to understand. The distance from P_0 to a zone changes by only $\lambda/2$ as we move from zone to zone and the area of a zone is almost a constant, independent of the zone number; thus, the amplitudes of the Huygens' wavelets from each zone should be approximately equal. The alternation in sign, from zone to zone, is due to the phase change of the light wave from adjacent zones because the propagation paths for adjacent zones differ by $\lambda/2$.

To find the total field strength at P_0 due to N zones, the collection of Huygens' wavelets is added:

$$E(P_0) = \sum_{j=1}^{N} \tilde{E}_j(P_0) = \frac{2i\lambda A}{D} e^{-ikD} \sum_{j=1}^{N} (-1)^j C_j \qquad [37]$$

To evaluate the sum, the elements of the sum are regrouped and rewritten as

$$-\sum_{j=1}^{N} (-1)^j C_j = \frac{C_1}{2} + \left(\frac{C_1}{2} - C_2 + \frac{C_3}{2}\right) + \left(\frac{C_3}{2} - C_4 + \frac{C_5}{2}\right) + \ldots$$

Because the C's are very slowly varying functions of j, even out to 500 zones, we are justified in setting the quantities in parentheses equal to zero. With this approximation, the summation can be set equal to one of two values, depending upon whether there is an even or odd number of terms in the summation

$$-\sum_{j=1}^{N} (-1)^j C_j = \left\{ \begin{array}{ll} \frac{1}{2}(C_1 + C_N) & N \text{ odd} \\ \frac{1}{2}(C_1 - C_N) & N \text{ even} \end{array} \right\}$$

For very large N, the obliquity factor approaches zero, $C_N \to 0$, as was demonstrated in Figure 4. Thus, the theory has led us to the conclusion that the total field produced by an unobstructed wave is equal to one half the contribution from the first Fresnel zone, i.e., $E = E_1/2$. Stating this result in a slightly different way, we obtain a surprising result – the contribution from the first Fresnel zone is twice the amplitude of the unobstructed wave!

The zone construction can be used to analyze the effect of the obstruction of all or part of a zone. For example, by constructing a circular aperture with a diameter equal to the diameter of the first Fresnel zone, we have just demonstrated that it is possible to produce an intensity at the point P_0 equal to four times the intensity that would be observed if no aperture were present. To analyze the effects of a smaller aperture, we subdivide a half-period zone into a number of subzones such that there is a constant phase difference between each subzone. The individual vectors form a curve know as the vibrational curve. Figure 13 shows a vibrational curve produced by the addition of waves from nine subzones of the first Fresnel zone. The vibrational curve in Figure 13 is an arc with the appearance of a half-circle.

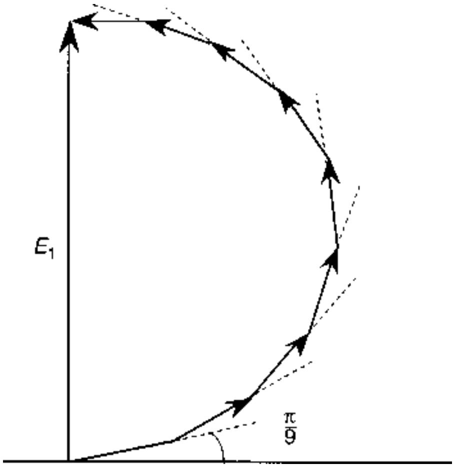

Figure 13 Vector addition of waves from nine subzones constructed in the first Fresnel zone. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If the radius of curvature of the arc were calculated, we would discover that it is a constant, except for the contribution of the obliquity factor,

$$\rho = \frac{\lambda R'}{R' + Z} C(\varphi)$$

Because the obliquity factor for a single zone is a constant to 2 parts in 10^{-7}, the radius of curvature of the vibration curve can be considered a constant over a single zone. If we let the number of subzones approach infinity, the vibrational curve becomes a semicircle whose chord is equal to the wavelet produced by zone one, i.e., E_1. If we subdivide additional zones, and add the subzone contributions, we create other half-circles whose radii decrease at the same rate as the obliquity factor. The vibrational curve for the total wave is a spiral, constructed of semicircles which converges to a point halfway between the first half-circle, see Figure 14. The length of the vector from the start to the end of the spiral is $E = E_1/2$, as we derived above. When this same construction technique is applied to a rectangular aperture, the vibrational curve generated is the Cornu spiral.

Circular Aperture

The vector addition technique, described in Figure 13, can be used to evaluate Fresnel diffraction, at a point P_0, from a circular aperture and yields the intensity distribution along the axis of symmetry of the circular aperture. The zone concept will also allow a qualitative description of the light distribution normal to this axis.

To develop a quantitative estimate of the intensity at an observation point on the axis of symmetry of the circular aperture, we construct a spiral, Figure 14, to represent the Fresnel zones of a spherical wave incident on the aperture. The point B on the spiral shown in Figure 14 corresponds to the portion of the spherical wave unobstructed by the screen. The length of the chord, AB, represents the amplitude of the light wave at the observation point P_0. As the diameter of the aperture increases, B moves along the spiral, in a counterclockwise fashion, away from A. The first maximum occurs when B reaches the point labeled A_1 in Figure 14; the aperture has then uncovered the first Fresnel zone. At this point, the amplitude is twice what it would be with no obstruction. Four times the intensity!

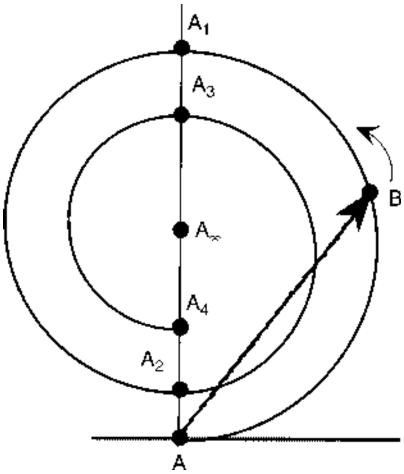

Figure 14 Vibration curve for determining the Fresnel diffraction from a circular aperture. The change of the diameter of the half-circles making up the spiral has been exaggerated for easy visualization. The actual changes are one part in 10^{-5}. The position of point B is determined by the aperture size. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If the aperture's diameter continues to increase, B reaches the point labeled A_2 in Figure 14 and the amplitude is very nearly zero; two zones are now exposed in the aperture. Further maxima occur when an odd number of zones are in the aperture and further minima when an even number of zones are exposed. Figure 15 shows an aperture containing four exposed Fresnel zones. The amplitude at the observation point would correspond to the chord drawn from A to A_4 in Figure 14.

The aperture diameter can be fixed and the observation point P_0 can move along, or perpendicular to, the axis of symmetry of the circular aperture. As P_0 is moved away from the aperture, along the symmetry axis, i.e., as Z increases, the radius of the Fresnel zones increase without limit. For small values of n, the radius of the nth zone can be approximated by

$$r_n \approx \sqrt{nZ\lambda} \quad [38]$$

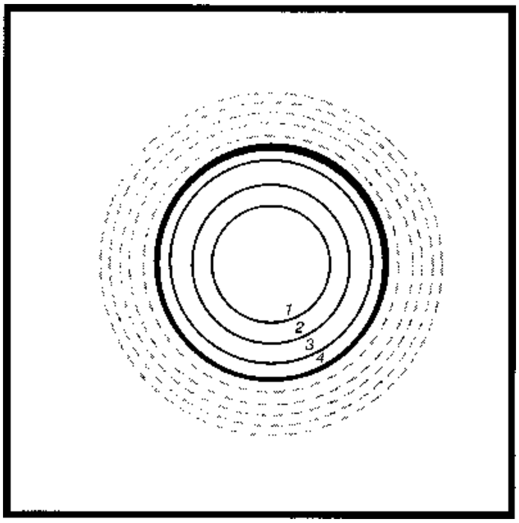

Figure 15 Aperture with four Fresnel zones exposed. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

At $Z_{\max}$, the light intensity is a maximum, given by the chord length from A to A_1 in Figure 14. If $\lambda = 500$ nm and $a = 0.5$ mm then this maximum occurs when $Z = 0.5$ m

$$Z_{\max} = \frac{a^2}{\lambda} \qquad [39]$$

If we start at $Z_{\max}$ and move toward the aperture, along the axis, as Z decreases in value, a point will be reached when the intensity on the axis becomes a minimum. The value of Z where the first minimum in intensity is observed is equal to

$$Z_{\min} = \frac{a^2}{2\lambda}$$

In Figure 14 the chord would extend from A to A_2.

As the observation point, P_0, is moved along the axis toward the aperture, and Z assumes values less than $Z_{\min}$, the point B in Figure 14 spirals inward, toward the center of the spiral and the intensity cycles through maximum and minimum values. The cycling of the intensity, as the observation point moves toward the aperture, will not continue indefinitely. At some point, the field on the axis will approach the field observed without the aperture because the distance between $Z_{\max}$ and $Z_{\min}$ shrinks to the wavelength of light making the intensity variation unobservable.

For values of Z that exceed $Z_{\max}$ of [39], the aperture radius, a, will be smaller than the radius of the first zone, and Fraunhofer diffraction will be observed because the aperture contains only one zone and the phase in the aperture is a constant.

Opaque Screen

If the screen containing a circular aperture, of radius a, is replaced by an opaque disk of radius a, the intensity distribution on the symmetry axis, behind the disk, is found to be equal to the value that would be observed with no disk present. This prediction was first derived by Poisson, to demonstrate that wave theory was incorrect; however, experimental observation supported the prediction and verified the theory.

We construct a spiral, shown in Figure 16, similar to the one for the circular aperture in Figure 14. The point B on the spiral represents the edge of the disk. The shaded portion of the spiral from A to B does not contribute because that portion of the wave is covered by the disk and the zones, associated with that portion of the wave, cannot be seen from the observation point. The amplitude at P_0 is the length of the chord from B to A_∞ shown in Figure 16. If the observation point moves toward the disk, then B moves along the spiral toward A_∞. There is always intensity on the axis for this configuration, though it slowly decreases until it reaches zero when the observation point reaches the disk; this corresponds to point B reaching point A_∞ on the spiral. Physically, zero intensity occurs when the disk blocks the entire light wave. There are no maxima or minima observed as the disk diameter, a, increases or as the observation distance changes. If the observation point is moved perpendicular to the symmetry axis, a set of concentric bright rings are observed. The origin of these bright rings can be explained using Fresnel zones, in a manner similar to the one used to explain the bright rings observed in a Fresnel diffraction pattern from a circular aperture.

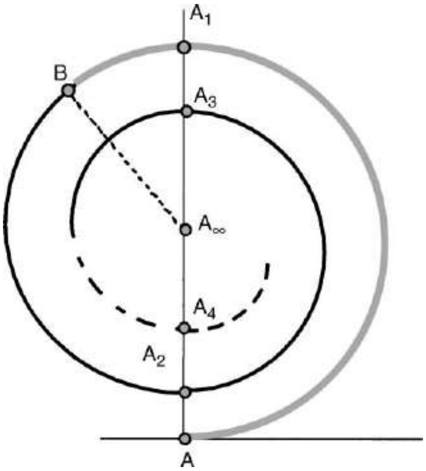

Figure 16 Vibration curve for opaque disk. The shaded region makes no contribution because it is associated with the portion of the wave obstructed by the opaque object. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Zone Plate

In the construction of Fresnel zones, each zone was assumed to produce a Huygens wavelet, out of phase with the wavelets produced by its nearest neighbors. If every other zone were blocked, then there would be no negative contributions to [37]. The intensity on-axis would be equal to the square of the sum of the amplitudes produced by the in-phase zones – exceeding the intensity of the incident wave. An optical component made by the obstruction of either the odd or the even zones could therefore be used to concentrate the energy in a light wave.

The boundaries of the opaque zones, used to block out-of-phase wavelets, are seen from [38] to increase as the square root of the integers. An array of opaque rings, constructed according to this prescription is called a zone plate, see Figure 17.

A zone plate will perform like a lens with focal length

$$f = \pm \frac{r_1^2}{\lambda} \qquad [40]$$

The zone plate, shown in Figure 17, will act as both a positive and negative lens. What we originally called the source can now be labeled the object point, O, and what we called the observation point can now be labeled the image point, I. The light passing through the zone plate is diffracted into two paths, labeled C and D in Figure 17. The light waves, labeled C, converge to a real image point I. For these waves the zone plate performs the role of a positive lens with a focal length given by the positive value of [40]. The light waves, labeled D in Figure 17, appear to originate from the virtual image point labeled I. For these waves the zone plate performs the role of a negative lens with a focal length given by the negative value of [40].

The zone plate will not have a single focus, as is the case for a refractive optical element, but rather will have multiple foci. As we move toward the zone plate, from the first focus, given by [40], the effective Fresnel zones will decrease in diameter. The zone plate will no longer obstruct out-of-phase Fresnel zones and the light intensity on the axis will decrease. However, additional maxima, of the on-axis intensity, will be observed at values of Z for which the first zone plate opening contains an odd number of zones. These positions can also be labeled as foci of the zone plate; however, the intensity at each of these foci will be less than the intensity at the primary focus.

Lord Rayleigh suggested that an improvement of the zone plate design would result if, instead of blocking every other zone, we shifted the phase of alternate zones by 180°. The resulting zone plate, called a phase-reversal zone plate, would, more efficiently, utilize the incident light. R.W. Wood was the first to make such a zone plate. Holography provides an optical method of constructing the phase-reversal zone plates in the visible region of the spectrum. Semiconductor lithography has been used to produce zone plates in the x-ray region of the spectrum.

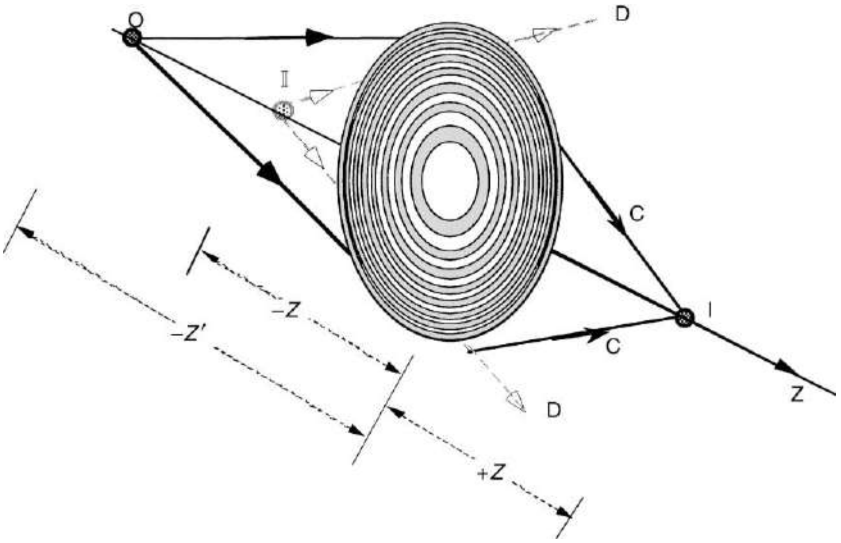

Figure 17 A zone plate acts as if it were both a positive and a negative lens. Light from the object, O, is diffracted into waves traveling in both the D and the C directions. The light traveling in the C direction produces a real image of O at I. The light traveling in the D direction appears to originate from a virtual image at I. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

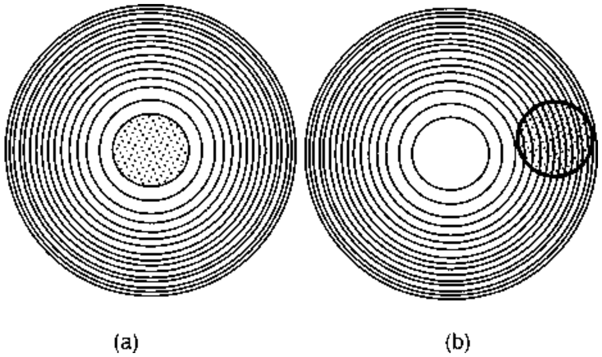

Figure 18 (a) A set of Fresnel zones has been constructed about the optical path taken by some light ray. If the optical path of the ray is varied over the cross-hatched area shown in the figure, then the optical path length does not change. This cross-hatched area is equal to the first Fresnel zone and is described as the neighborhood of the light ray. (b) The neighborhood defined in (a) is moved so that it surrounds an incorrect optical path for a light ray. We see that this region of space would contribute no wave amplitude at the observation point because of the destructive interference between the large number of partial zones contained in the neighborhood. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The resolving power of a zone plate is a function of the number of zones contained in the plate. When the number of zones exceeds 200, the zone plate's resolution approaches that of a refractive lens.

Fermat's Principle

The Fresnel zone construction provides physical insight into the interpretation of Fermat's principle which states that, if the optical path length of a light ray is varied in a neighborhood about the true path, there is no change in path length. By constructing a set of Fresnel zones about the optical path of a light ray, we can discover the physical meaning of a neighborhood.

The rules for constructing a Fresnel zone require that all rays passing through a given Fresnel zone have the same optical path length. The true path will pass through the center of the first Fresnel zone constructed on the actual wavefront. A neighborhood must be the area of the first Fresnel zone, for it is over this area that the optical path length does not change. Figure 18a shows the neighborhood, about the true optical path, as a cross-hatched region equal to the first Fresnel zone.

Light waves do travel over wrong paths but we do not observe them because the phase differences for those waves that travel over the 'wrong' paths are such that they destructively interfere. By moving the neighborhood, defined in the previous paragraph as the first Fresnel zone, to a region displaced from the true optical path, we can use the zone construction to see that this statement is correct. In Figure 18b the neighborhood is constructed about a ray that is improper according to Fermat's principle. We see that this region of space would contribute no energy at the observation point because of destructive interference between the large number of partial zones contained in the neighborhood.

This leads us to another interpretation of diffraction. If an obstruction is placed in the wavefront so that we block some of the normally interfering wavelets, we see light not predicted by Fermat's principle, i.e., we see diffraction.

Further Reading

Backer BB and Copson EJ (1969) *The Mathematical Theory of Huygens' Principle*. London: Oxford University Press.

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Rubinowicz A (1984) *Tahe Miyamoto–Wolf diffraction wave*. In: Wolf E (ed.) Progress in Optics, vol. IV.

Sayanagi I (1967) Pinhole Imagery. *Journal of the Optical Society of America*. 57: 1091–1099.

Sears FW (1949) *Optics*. Reading, MA: Addison-Wesley.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

Stegliani DJ, Mittra R and Semonin RG (1967) Resolving power of a zone plate. *Journal of the Optical Society of America* 57: 610–613.

Stone JM (1963) *Radiation and Optics*. New York: McGraw-Hill.

DIFFRACTIVE SYSTEMS

Contents

Aberration Correction with Diffractive Elements
Applications of Diffractive and Micro-Optics in Lithography
Design and Fabrication of Diffractive Optical Elements
Diffractive Laser Resonators
Diffractives in Animals
Microstructure Fibers
Omnidirectional Surfaces and Fibers
Wave Optical Modeling and Design

Aberration Correction with Diffractive Elements

N Lindlein, University of Erlangen-Nürnberg,
Erlangen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Diffractive optical elements (DOEs) can be used for the correction of aberrations in optical systems because of two fundamental properties:

1. The phase function Φ of a DOE can be chosen quite arbitrarily so that for a given wavelength λ the DOE can act like an aspheric optical element.
2. The dispersion of a DOE due to the local grating equation is very high and its sign is opposite to that of the dispersion of a refractive optical element with normal dispersion of the refractive index.

The first property allows the correction of monochromatic aberrations in an optical system similar to the case of using an aspheric surface in a refractive optical system. The second property allows the construction of hybrid achromatic optical elements that consist of a refractive lens and a diffractive lens whose optical powers have the same sign (e.g., two positive lenses for a positive achromatic element or two negative lenses for a negative achromatic element).

This article will first describe two methods for the simulation of DOEs by ray tracing because ray tracing is the basic simulation tool for aberration correction. Afterwards, the properties of DOEs for the correction of monochromatic and chromatic aberrations and some applications will be discussed. The name DOE will always mean either binary, or multiple-step or blazed surface relief diffractive optical elements which change only the phase of the incident wave and not the amplitude. All examples will be given for transmission elements but in principle reflective elements can also be described with quite similar formulas.

Ray Tracing Simulation of DOEs

First of all it seems straightforward to use a diffractive analysis to simulate a diffractive optical element. However, the global diffractive analysis of a general DOE with perhaps 1000 or more grating structures and locally changing grating period and especially the diffractive analysis of a complete optical system containing DOEs and other elements is so complex that it can be done only for very simple DOEs and systems. Therefore, the diffractive analysis of a DOE is mostly restricted to a local analysis using some periods (e.g., 10–50 periods) to calculate, e.g., the different local diffraction orders, their diffraction efficiencies and possibly local phase shifts. For the global analysis another simulation method has to be used.

The most important simulation method for the correction of aberrations in traditional optical systems is ray tracing. Therefore, ray tracing methods are also used for DOEs. These methods can calculate the path of a ray for a given diffraction order, but they cannot calculate the diffraction efficiency for this ray. For the calculation of the diffraction efficiency local diffractive analysis methods (scalar or rigorous methods depending on the local grating frequency and the angle of the incident ray) have to be used in addition. In particular, for applications with a broad wavelength spectrum (e.g., an achromatic lens) or a broad spectrum of angles of the incident rays, the diffraction efficiency has to be analyzed carefully to obtain the energy distribution in the different diffraction orders. But this is beyond the scope of this article about aberration correction and we assume that we know the local diffraction efficiency for each ray.

Sweatt Model for the Ray Tracing Simulation of DOEs

An aspheric refractive surface with height h as a function of the lateral coordinates x and y and a refractive index n below the surface and air (refractive index 1) above the surface can be approximately replaced by a DOE with the phase function (see Figure 1)

$$\Phi(x, y) = 2\pi \frac{(n-1)h(x, y)}{\lambda} \qquad [1]$$

Here, λ is the wavelength of the light for which the DOE is designed and the phase function is a continuous representation of the structure of the DOE. To encode the phase function of the DOE the wrapped phase function $\Phi \bmod 2\pi$ has to be considered. The encoding itself can be done by building a blazed, a multiple-step or a binary structure (see Figure 1), whereby the first-order diffraction efficiency decreases from the blazed to the binary structure (at maximum 100% diffraction efficiency for the blazed structure and about 40% for the binary structure). The replacement of an aspheric refracting surface by a DOE using eqn [1] is only valid in the paraxial approximation since the phase function of the DOE is encoded in a plane whereas the refracting surface of the aspheric is typically curved. But, the agreement between the aspheric surface and the DOE with a phase function calculated with the help of eqn [1] increases with increasing refractive index n since then the curvature of the aspheric is quite small and the refraction also nearly takes place in a plane. Of course, the refractive index of real materials cannot exceed a value of more than about 3.5 (for silicon used in the near infrared). But, in a numerical simulation the refractive index can be arbitrarily high, e.g., several thousand. This is the reason why the so called Sweatt model for tracing rays through DOEs uses a weakly curved aspheric surface with height h_{sweatt} and a very high refractive index n_{sweatt} to simulate a diffractive phase function Φ:

$$h_{\text{sweatt}}(x, y) = \frac{\lambda}{n_{\text{sweatt}} - 1} \frac{\Phi(x, y)}{2\pi} \qquad [2]$$

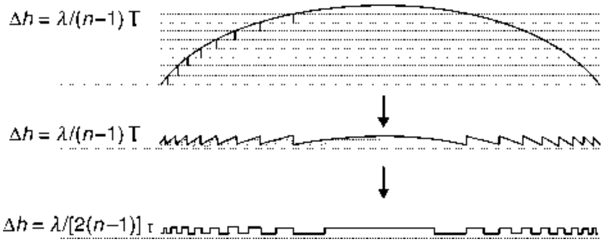
$\Delta h = \lambda/(n-1)$
$\Delta h = \lambda/(n-1)$
$\Delta h = \lambda/[2(n-1)]$

Figure 1 Transition from an aspheric refracting surface to a blazed diffractive optical element and finally to a binary diffractive optical element.

To take into account the wavelength dispersion of the DOE the refractive index n_{sweatt} in the Sweatt model has to be proportional to the wavelength λ

$$n_{\text{sweatt}}(\lambda) = n_0 \frac{\lambda}{\lambda_0} \qquad [3]$$

where n_0 is the refractive index at the wavelength λ_0.

The validity of the Sweatt model increases with increasing refractive index and becomes nearly exact for $n_{\text{sweatt}} \to \infty$ (in practice values of several thousand for the refractive index are normally sufficient and a rule of thumb is to take the numerical value of the utilized wavelength specified either in nanometers for infrared light or angstroms for visible light). The Sweatt model was introduced as a means to simulate DOEs in ray tracing programs unable to implement DOEs directly but capable of simulating aspheric refractive surfaces. However, a disadvantage of this model is the numerical error generated by the finite value of the refractive index.

Modern ray tracing programs therefore use a more appropriate model based on the local grating equation.

Local Grating Model for the Ray Tracing Simulation of DOEs

The basic equation for the simulation of DOEs by the local grating model is that the phase Φ_{out} of a diffracted ray in the plane of the DOE can be calculated from the phase Φ_{in} of the incident ray by adding the phase function Φ of the DOE multiplied by the diffraction order m:

$$\Phi_{\text{out}}(x, y, z_0) = \Phi_{\text{in}}(x, y, z_0) + m\Phi(x, y) \qquad [4]$$

Here, z_0 is the z-coordinate of the DOE plane which is assumed to be perpendicular to the z-axis (see Figure 2). The diffraction order m will be only one for an ideal blazed DOE but it can have an arbitrary integer value for binary DOEs. In practice, the optical designer will calculate the system for one diffraction order and analyze the system for other diffraction orders afterwards if necessary.

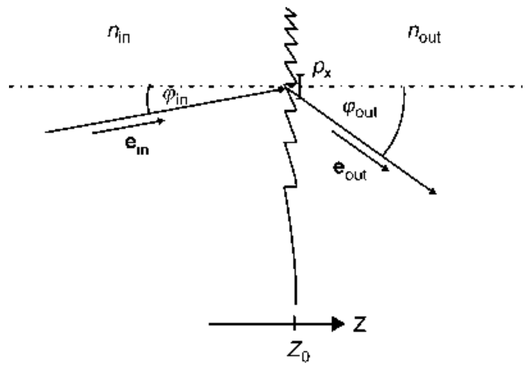

Figure 2 Schematic view in two-dimensions of the path of a ray which is diffracted at a DOE.

For the phases Φ_{in} of the incident and Φ_{out} of the diffracted ray the eikonal equation of geometrical optics is valid,

$$|\nabla\Phi_j|^2 = k_j^2 = \left(\frac{2\pi n_j}{\lambda}\right)^2 \qquad [5]$$

where the index j can have the two values $j \in \{\text{in}, \text{out}\}$. n_{in} or n_{out} are the refractive indices of the materials in which the incident or diffracted rays propagate. λ is the wavelength in vacuum. This form of the eikonal equation is obtained from the more common form $|\nabla L_j|^2 = n_j^2$, which is derived in any textbook of optics, by multiplying it with $(2\pi/\lambda)^2$ and using the definition of the optical phase $\Phi_j = 2\pi L_j/\lambda$. Here, L_j is the optical path length which is often called the eikonal. The curves of constant phase $\Phi_j(x,y,z) = \text{constant}$ are the wavefronts and the gradient $\nabla\Phi_j(x,y,z)$ points in the direction of the local ray which is perpendicular to the wavefront at the point (x,y,z) and describes its direction of propagation. Furthermore, the unit vectors $\mathbf{e}_{\text{in}}$ and $\mathbf{e}_{\text{out}}$ along the ray direction of the incident and diffracted ray, respectively, can be calculated by normalizing the gradient:

$$\mathbf{e}_j = \frac{\nabla\Phi_j}{|\nabla\Phi_j|} = \frac{\lambda}{2\pi n_j}\nabla\Phi_j\,; \qquad j \in (\text{in}, \text{out}) \qquad [6]$$

Since $\mathbf{e}_j$ is a unit vector, i.e., $e_{x,j}^2 + e_{y,j}^2 + e_{z,j}^2 = 1$, it can be seen from eqn [6] (and also from the original eqn [5]) that the z-component $e_{z,j}$ of the ray direction vector can be calculated (apart from its sign) if the two partial derivatives of the phase with respect to x and y are known. This means that it is sufficient to know the phase in a plane as stated in eqn [4].

By calculating the partial derivatives of eqn [4] with respect to x and y the equations for the ray tracing at a DOE are derived:

$$\begin{aligned}
\frac{\partial\Phi_{\text{out}}(x,y)}{\partial x} &= \frac{\partial\Phi_{\text{in}}(x,y)}{\partial x} + m\frac{\partial\Phi(x,y)}{\partial x} \\
&\Rightarrow n_{\text{out}}e_{x,\text{out}}(x,y) \\
&= n_{\text{in}}e_{x,\text{in}}(x,y) + m\frac{\lambda}{2\pi}\frac{\partial\Phi(x,y)}{\partial x} \\
\frac{\partial\Phi_{\text{out}}(x,y)}{\partial y} &= \frac{\partial\Phi_{\text{in}}(x,y)}{\partial y} + m\frac{\partial\Phi(x,y)}{\partial y} \\
&\Rightarrow n_{\text{out}}e_{y,\text{out}}(x,y) \\
&= n_{\text{in}}e_{y,\text{in}}(x,y) + m\frac{\lambda}{2\pi}\frac{\partial\Phi(x,y)}{\partial y}
\end{aligned} \qquad [7]$$

The local grating frequencies ν_x and ν_y in x and y direction are calculated from the phase function Φ by:

$$\begin{aligned}
\nu_x(x,y) &= \frac{1}{2\pi}\frac{\partial\Phi(x,y)}{\partial x} \\
\nu_y(x,y) &= \frac{1}{2\pi}\frac{\partial\Phi(x,y)}{\partial y}
\end{aligned} \qquad [8]$$

Therefore, given $\mathbf{e}_{\text{in}}$ and eqn [6], the ray direction vector $\mathbf{e}_{\text{out}}$ of the diffracted ray is

$$n_{\text{out}}\mathbf{e}_{\text{out}} = n_{\text{out}}\begin{pmatrix} e_{x,\text{out}} \\ e_{y,\text{out}} \\ e_{z,\text{out}} \end{pmatrix} = \begin{pmatrix} n_{\text{in}}e_{x,\text{in}} + m\lambda\nu_x \\ n_{\text{in}}e_{y,\text{in}} + m\lambda\nu_y \\ \sqrt{n_{\text{out}}^2 - (n_{\text{in}}e_{x,\text{in}} + m\lambda\nu_x)^2 - (n_{\text{in}}e_{y,\text{in}} + m\lambda\nu_y)^2} \end{pmatrix} \qquad [9]$$

The dependence of the lateral coordinates x and y is left out in eqn [9] to simplify notation. It can be easily checked that this equation fulfills the required condition for the unit vector: $e_{x,\text{out}}^2 + e_{y,\text{out}}^2 + e_{z,\text{out}}^2 = 1$.

As stated above, for a blazed DOE fabricated to the exact depth and illuminated by the design wavelength only the first diffraction order $m = 1$ is generated. But in practice other diffraction orders m are generated due to fabrication errors or illuminating with other wavelengths. In particular, by using binary DOEs there are inherently different diffraction orders. Therefore, it may be necessary to analyze the optical system for several values m one after the other. Each value m defines one wavefront generated by the DOE and the wavefront can be calculated by evaluating the optical path lengths along the rays at different points. In practice, the optical system has to be analyzed for those values m with sufficiently high diffraction efficiency. Then, it can be checked whether undesired diffraction orders of the DOE are blocked in the optical system by aperture stops or whether they will disturb the output of the optical system. The local diffraction efficiency of each ray for a value m can be calculated to a good approximation by using a diffractive analysis of the corresponding local grating, i.e., that of an infinitely extended grating which would have the same grating period and the same structure shape (e.g., depth) as the DOE at the specific point.

If the y component of the incident ray and the local grating frequency ν_y are both zero the ray components in the x direction can be expressed via the sine of the angle of incidence φ_{in} or the angle of the diffracted ray φ_{out} (see Figure 2). In this case the first row of eqn [9] can be written as

$$n_{\text{out}} \sin \varphi_{\text{out}} = n_{\text{in}} \sin \varphi_{\text{in}} + m \frac{\lambda}{p_x} \qquad [10]$$

where the local grating period $p_x = 1/\nu_x$ has been introduced. This is the well-known grating equation.

In this article eqns [4]–[10] assume that the DOE is on a planar substrate and that x and y are the lateral coordinates on the substrate. However, the equations can also be written in a vector formulation which is valid for DOEs on curved substrates (see Further Reading).

So, we have now a powerful simulation method to design and analyze in the following several optical systems containing DOEs to correct monochromatic or chromatic aberrations.

Correction of Monochromatic Aberrations

For practical applications it is important to have an efficient algorithm to calculate the phase function of a DOE to correct the aberrations in a given optical system. Several different cases have to be distinguished. An easy case is the design of a DOE to produce a desired output wavefront Φ_{out} from a known input Φ_{in}. In this case, eqn [4] can be used to calculate the phase function Φ of the DOE directly. Two examples are given in the following. Afterwards, some aspects of a more general numerical optimization procedure are described.

Aberration Correction for Interferometrically Recorded Holograms

Although the following example of the aberration correction of interferometrically recorded holograms by using DOEs seems to be quite special it shows exemplarily several principal steps which are necessary to design a DOE for correcting monochromatic aberrations.

In this article the name DOE is used for diffractive elements with a microstructured surface (binary, multistep or blazed). Such elements have the property that their diffraction efficiency depends only weakly on the wavelength or the angle of incidence of the illuminating light. Another interesting class of diffractive elements are holograms in volume media, which are recorded interferometrically using two coherent recording waves. The refractive index in a volume hologram is modulated periodically and the thickness has to amount to several periods of the interference pattern along the z-direction in order to have a so-called thick hologram with a high diffraction efficiency. If there are only a few periods along the z-direction the quite small modulation of the refractive index (typically about 0.02) will result in a low diffraction efficiency. Therefore, if the angle between the two recording waves is small the resulting interference pattern will have large periods and only a few periods are recorded in the photosensitive layer along the z-direction. But this means that the holographic optical element (HOE) can only have a small diffraction efficiency. From this fact, it is clear that also by using the recorded HOE the deflection angle between the incident and the diffracted wave must not be smaller than a certain value because otherwise the thickness has to be very big or the diffraction efficiency decreases. Therefore, such elements with high diffraction efficiency, which will be named in the following holographic optical elements (HOE), are always off-axis elements because for on-axis elements all rays in the neighborhood of the axis are only deflected by a small angle and then the diffraction efficiency would be low. For the correct wavelength and angle of incidence, so that the Bragg condition of holography is fulfilled, HOEs, like blazed DOEs, exhibit only one dominant diffraction order $m = m_{\text{dom}}$ (normally $m_{\text{dom}} = 1$). But, if the angle of incidence or the wavelength change, the diffraction efficiency in the order m_{dom} decreases rapidly and light passes through the HOE without being deflected, i.e., most of the light is then in the order $m = 0$. For this reason, a HOE can be used for example for a head-up display to fade in some information under the Bragg angle without hindering the normal view onto the environment.

Traditional materials for HOEs like dichromated gelatin (DCG) are sensitive only in the blue–green spectral region (e.g., for light of an argon ion laser with $\lambda = 488$ nm). However, once recorded it will be illuminated in many applications (e.g., optical interconnects) by a wavelength in the red–infrared spectral region. Unfortunately, the wavelength mismatch between recording and reconstruction illumination introduces aberrations.

If the two desired reconstruction waves at the wavelength λ_2 have the phases $\Phi_{\mathrm{r},\lambda 2}$ and $\Phi_{\mathrm{o},\lambda 2}$ in the surface plane of the HOE (the subscripts r and o refer to interferometry reference and object, respectively), the phase function Φ of the HOE which has to be recorded is

$$\Phi(x, y) = \Phi_{\mathrm{o},\lambda 2}(x, y) - \Phi_{\mathrm{r},\lambda 2}(x, y) \qquad [11]$$

If our discussion here is restricted to holographic lenses, the two reconstruction waves will be spherical (or plane) waves (see Figure 3b) with the center of curvature at the points $\mathbf{r}_{\mathrm{r},\lambda 2}$ and $\mathbf{r}_{\mathrm{o},\lambda 2}$ and eqn [11] is:

$$\begin{aligned}\Phi(x,\ y) = {} & \frac{2\pi\sigma_{\mathrm{o},\lambda 2}}{\lambda_2}\sqrt{(x - x_{\mathrm{o},\lambda 2})^2 + (y - y_{\mathrm{o},\lambda 2})^2 + z_{\mathrm{o},\lambda 2}^2} \\ & - \frac{2\pi\sigma_{\mathrm{r},\lambda 2}}{\lambda_2}\sqrt{(x - x_{\mathrm{r},\lambda 2})^2 + (y - y_{\mathrm{r},\lambda 2})^2 + z_{\mathrm{r},\lambda 2}^2}\end{aligned} \qquad [12]$$

The parameters $\sigma_{\mathrm{o},\lambda 2}$ and $\sigma_{\mathrm{r},\lambda 2}$ are either $+1$ for a divergent spherical wave or -1 for a convergent spherical wave. The surface of the HOE itself is in the plane at $z = 0$.

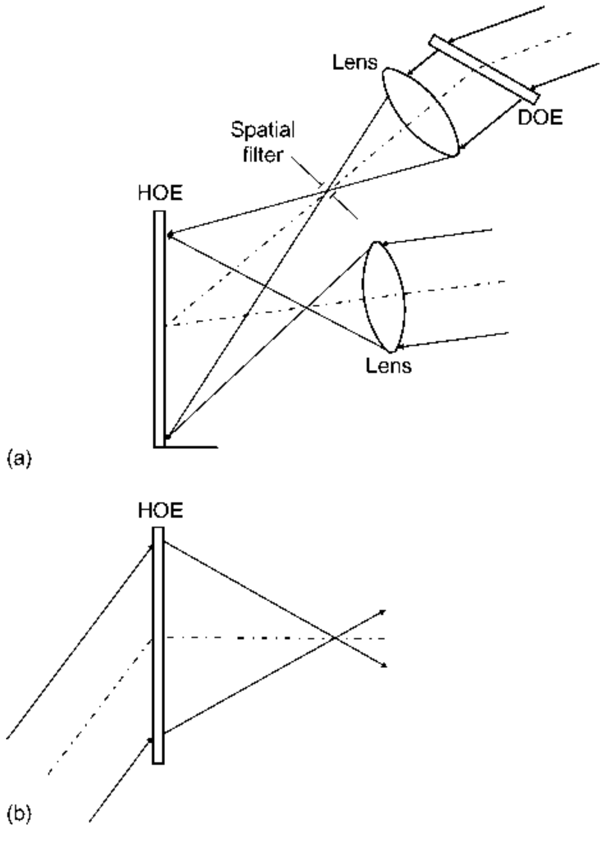

Figure 3 Schematic view of the (a) recording and (b) reconstruction setup for a holographic lens. The recording is made at wavelength λ_1 and the reconstruction at wavelength λ_2.

The task is now to find two recording waves which construct nearly the same three-dimensional interference pattern in the hologram at the recording wavelength λ_1. This means that eqn [12] has to be fulfilled exactly for the difference of the phases of the two recording waves in the plane $z = 0$ and, additionally, the Bragg condition has to be fulfilled at least approximately. So, the next step is to calculate one of the recording waves with phase $\Phi_{\mathrm{r},\lambda 1}$ so that the Bragg condition is fulfilled. This can be done approximately with a third-order theory resulting in a spherical wave or nearly exactly with an iterative numerical procedure resulting generally in an aspheric recording wave. If $\Phi_{\mathrm{r},\lambda 1}$ is known, the other recording wave with phase $\Phi_{\mathrm{o},\lambda 1}$ is calculated by

$$\Phi_{\mathrm{o},\lambda 1}(x, y) = \Phi(x, y) + \Phi_{\mathrm{r},\lambda 1}(x, y) \qquad [13]$$

At least this recording wave will be aspheric. To generate such an aspheric wave it is traced by ray tracing to a plane where the two recording waves are spatially separated. There, either a DOE or a combination of a DOE and a lens are used to generate the aspheric wave. The calculations of the phase function of the DOE are done again by using an equation like [11] whereby one wave is the propagated aspheric wave (possibly propagated also through a lens) and the other is normally a plane off-axis wave with which the DOE will be illuminated in the real recording setup (see Figure 3a). Here, the off-axis angle is necessary to allow spatial filtering of disturbing diffraction orders of the DOE which will be normally fabricated as a binary DOE.

Experiments of our own showed (see the literature) that nearly diffraction-limited holographic lenses with a numerical aperture of NA = 0.4 can be fabricated by correcting the aberrations of about 28 wavelengths peak-to-valley (for the reconstruction at 633 nm wavelength) due to the wavelength mismatch between recording at 488 nm wavelength and reconstruction at 633 nm wavelength by using a DOE in the recording setup.

Testing of Aspheric Surfaces by Using a DOE as Aberration Compensating Element

An important application for DOEs is testing aspheric surfaces and nowadays more and more companies use DOEs to test aspheric surfaces which are used for example in the objectives of lithographic wafer steppers to produce highly integrated microelectronic circuits.

In this testing setup the DOE is illuminated by a wave with phase Φ_{in} to generate a wave with phase Φ_{out} that is incident perpendicular to the aspheric surface (see Figure 4). After reflection at the asphere

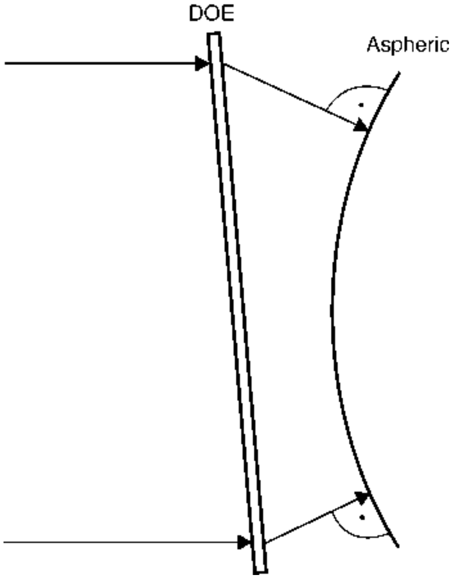

Figure 4 Scheme showing the basic elements for testing an aspheric surface.

the wave is retro-reflected and takes the same path if the asphere has no surface deviations and no misalignments. This means that the setup is completely symmetric for a reflection at the aspheric surface. Therefore, the optical path lengths have to be identical for all rays on the asphere. To calculate the phase function of the DOE the following steps are made:

- A grid of N rays starting perpendicularly on the aspheric surface and having all phase zero (due to the fact that the phase has to be identical it can be set to zero) is traced to the plane of the DOE. N depends on the shape and numerical aperture of the asphere and the grid will be in most cases two dimensional and regular or one dimensional (along one radial section of the asphere) and regular if the asphere and the system are completely rotationally symmetric. As a rule of thumb the number N of rays has to be more than twice the number of polynomial coefficients if the asphere is described by a polynomial.
- The phase of these rays is the function $-\Phi_{\text{out}}$ (the negative sign is necessary because the rays take the opposite direction) and is now known at some points (x_i, y_i) with $i \in \{1, 2, \ldots, N\}$.
- The phase Φ_{in} of the incident wave which is in practice either a plane wave or a spherical wave is calculated in the plane of the DOE. If the incident wave is generated by a lens having aberrations it is again calculated by ray tracing, otherwise it can be calculated analytically.
- Now, the phase function Φ of the DOE can in principle be calculated with the help of eqn [4].

However, the phase Φ_{out} (and possibly also Φ_{in}) is known only at some points (x_i, y_i), but it has to be known at all points of the DOE. Therefore in practice an analytical representation of the phase functions of the incident and output wave has to be calculated. This can be done by fitting a polynomial to the phase values using a least squares fit. In this case the coordinates, at which the phase values are known, can be quite arbitrary as long as the sampling is dense enough. For a rotationally symmetric system the polynomial should also be rotationally symmetric, i.e., a polynomial P_{rot} depending only on the radius coordinate r

$$r = \sqrt{x^2 + y^2} \qquad [14]$$

$$P_{\text{rot}}(r) = \sum_{j=0}^{G} a_j r^j \qquad \text{or} \qquad P_{\text{rot}}(r) = \sum_{j=0}^{G/2} a_{2j} r^{2j} \qquad [15]$$

where G is the degree of the polynomial, i.e., the highest power in r. In these two cases the number N of sampling points lying along one radial section of the rotationally symmetric optical system should be larger than $2G$ or G, respectively.

For general non-symmetric systems a Cartesian polynomial P_{cart} can be used, e.g.,

$$P_{\text{cart}}(x, y) = \sum_{j=0}^{G} \sum_{k=0}^{G-j} a_{jk} x^j y^k \qquad [16]$$

Here, the number of coefficients of the polynomial is $(G+1)(G+2)/2$ and the number N of sampling points should again be larger than twice the number of coefficients, i.e., $N > (G+1)(G+2)$.

For both types of polynomials the coefficients a_j or a_{jk} have to be calculated so that the merit function M

$$M(\mathbf{a}) = \sum_{i=1}^{N} (P(x_i, y_i) - \Phi_j(x_i, y_i))^2; \quad j \in \{\text{in, out}\} \qquad [17]$$

is minimized. The symbol P means either P_{rot} or P_{cart}. The vector $\mathbf{a}$ symbolizes the coefficients $\mathbf{a}_j$ or $\mathbf{a}_{jk}$. By differentiating M with respect to all coefficients and setting the partial derivatives to zero a system of linear equations is obtained which can be solved by standard procedures. As mentioned before, the number N of independent sampling points has to be at least double the size of the number of coefficients of the polynomial. Otherwise, the polynomial can 'swing' between the sampling points and the polynomial fits well to the desired phase function only at the sampling points but not at other points in between.

For very irregular aspherics with a fast change of the local curvature it might by necessary not to take a global polynomial fit but a spline fit, i.e., a set of adjacent local polynomials with several continuity conditions at the border area of each local polynomial. In this case the total number of coefficients used to describe the phase functions and Φ_{in} and Φ_{out} of the DOE will be increased compared to the description by a global polynomial. Then, the number N of rays for the calculation has also to be increased accordingly.

So, in this example the DOE acts as a null-compensator for the asphere and at first sight it is not a typical example for aberration correction. However, by regarding the complete system the DOE compensates the aberrations, i.e., deviations from a spherical wave, generated by the asphere which are primary and higher-order spherical aberrations if the asphere is rotationally symmetric and more general aberrations otherwise. This example shows how the phase function of a DOE can be designed straightforwardly for correcting aberrations if there is only one well-defined wavefront propagating through the optical system. This design method can also be used for other optical systems where the input and the output wavefront of the DOE are well defined.

Numerical Optimization of DOEs in Optical Systems

In a general optical system, where a DOE is not only used to compensate the aberrations for one specific incident wave but for different incident waves (optical system with a field), it is not possible to calculate the optimum phase function in such a straightforward way as given above. However, a DOE is in principle similar to an asphere in an optical system, so that the design rules for including aspherics in optical systems can be used in some cases as long as monochromatic aberrations are considered. Nevertheless, one should keep in mind that a DOE is, according to the Sweatt model, an asphere with high refractive index and small height, i.e., the deflection of the light rays occurs in a plane (or on a curved surface if the DOE is situated on a curved substrate). Therefore, the design rules for aspheric surfaces can be used only if the high refractive index of the Sweatt model is used in the design.

In the general case only a numerical optimization can be used. For this, it is very important to select an appropriate mathematical representation of the phase function of the DOE.

For rotationally symmetric optical systems also the phase function of the DOE should be rotationally symmetric. Therefore, in such a case it is useful to write the phase function Φ as a polynomial in the radius coordinate r as given in eqns [14] and [15]. Then, the $(G + 1)$ or $(G/2 + 1)$ coefficients a_j have to be optimized, whereby a_2 is proportional to the focal power of the DOE, whereas the other coefficients represent aspheric terms. Depending on the numerical aperture and on the application the degree G of the polynomial can be increased in order to achieve a better solution. Nevertheless, the degree in the optimization should be as small as necessary in order to avoid 'swinging effects' and numerical instabilities of the polynomial. To determine the smallest adequate degree an iterative method has to be used by starting with a quite small degree (e.g., 4). If the result of the optimization is not good enough the degree has to be increased until the result of the optimization is satisfying.

For non-rotationally symmetric optical systems a Cartesian polynomial like in eqn [16] can be taken and all the $(G + 2)(G + 1)/2$ coefficients a_{jk} have to be optimized. Since the number of free parameters is in this case larger than the number of free parameters in the rotationally symmetric case the convergence of the optimization may take a long time and the risk to break down in a local minimum of the multidimensional parameter space increases.

In principle, the calculation of a DOE used, e.g., for testing an aspheric surface can also be done by optimizing the coefficients of the polynomial representing the phase function of the DOE. Nevertheless, as a general rule it can be said that a direct calculation of the phase function is always more efficient than a numerical optimization, if the direct calculation is possible. In practice, it can be seen that the maximum degree of a polynomial, which is directly calculated by the method described in the last section, is always smaller than the degree which is necessary in a numerical optimization to achieve convergence of the optimization.

Of course, in such cases where several different types of aberrations (on-axis aberrations like spherical aberration, off-axis aberrations like astigmatism and coma or field aberrations like field curvature or distortion) have to be corrected in an optical system a direct calculation of the phase function is not possible so that only the numerical optimization can provide a solution.

Up to now only monochromatic aberrations have been discussed. But as stated at the beginning of this article the strong dispersion of a DOE having opposite sign to the dispersion of a refractive element offers additional applications of DOEs by correcting chromatic aberrations. The principle of doing this is described in the following.

Correction of Chromatic Aberrations

The focal power $1/f$ (f is the focal length of the lens) of a thin refractive lens consisting of two spherical surfaces with radii of curvature R_1 and R_2 and made of a material (glass) with refractive index n is

$$\frac{1}{f(\lambda)} = (n(\lambda) - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad [18]$$

where λ is the wavelength of the illuminating light.

This equation is valid for all wavelengths and so also for the commonly used wavelength $\lambda_d = 587.6$ nm (yellow d-line of helium) which is near the center of the visible spectrum. Then $1/R_1 - 1/R_2$ can be expressed as $1/((n_\mathrm{d} - 1)f_\mathrm{d})$ and the focal power depends on the wavelength as

$$\frac{1}{f(\lambda)} = \frac{n(\lambda) - 1}{n_\mathrm{d} - 1}\frac{1}{f_\mathrm{d}} \qquad [19]$$

where n_d and f_d are the refractive index and the focal length at $\lambda_\mathrm{d} = 587.6$ nm, respectively. So, by combining two thin lenses made of materials with refractive indices n_1 and n_2 and focal lengths f_1 and f_2 the focal power of the doublet assuming a zero distance between the two lenses is:

$$\begin{aligned}\frac{1}{f_\mathrm{ges}(\lambda)} &= \frac{1}{f_1(\lambda)} + \frac{1}{f_2(\lambda)}\\ &= \frac{n_1(\lambda) - 1}{n_{1,\mathrm{d}} - 1}\frac{1}{f_{1,\mathrm{d}}} + \frac{n_2(\lambda) - 1}{n_{2,\mathrm{d}} - 1}\frac{1}{f_{2,\mathrm{d}}} \qquad [20]\end{aligned}$$

For an achromatic doublet the focal power should be equal for two different wavelengths which are commonly the wavelengths $\lambda_\mathrm{F} = 486.1$ nm (blue F-line of hydrogen) and $\lambda_\mathrm{C} = 656.3$ nm (red C-line of hydrogen) which mark the limits of the well visible spectrum (of course the visible spectrum itself is broader but the sensitivity of the human eye to wavelengths below λ_F or beyond λ_C is quite bad). Therefore, the following equations result:

$$\begin{aligned}\frac{1}{f_\mathrm{ges}(\lambda_\mathrm{F})} = \frac{1}{f_\mathrm{ges}(\lambda_\mathrm{C})} &\Rightarrow \frac{n_1(\lambda_\mathrm{F}) - n_1(\lambda_\mathrm{C})}{n_{1,\mathrm{d}} - 1}\frac{1}{f_{1,\mathrm{d}}}\\ &= -\frac{n_2(\lambda_\mathrm{F}) - n_2(\lambda_\mathrm{C})}{n_{2,\mathrm{d}} - 1}\frac{1}{f_{2,\mathrm{d}}} \qquad [21]\end{aligned}$$

By introducing the Abbe number V_d

$$\begin{aligned}V_\mathrm{d} &= \frac{n_\mathrm{d} - 1}{n_\mathrm{F} - n_\mathrm{C}}\\ &= \frac{n(\lambda_\mathrm{d} = 587.6\ \mathrm{nm}) - 1}{n(\lambda_\mathrm{F} = 486.1\ \mathrm{nm}) - n(\lambda_\mathrm{C} = 656.3\ \mathrm{nm})} \qquad [22]\end{aligned}$$

equation [21] can be simplified to

$$V_{1,\mathrm{d}} f_{1,\mathrm{d}} = -V_{2,\mathrm{d}} f_{2,\mathrm{d}} \qquad [23]$$

Now, the wavelength dispersion of all conventional materials (glasses, polymers, etc.) is such that the refractive index decreases with increasing wavelength, i.e. the material has a so-called normal dispersion. This means that the Abbe number V_d is always positive and ranges from about 20 for high-dispersive glasses (e.g., SF59) to more than about 80 for low-dispersive glasses (e.g. Ultran 20). So, to fabricate a refractive achromatic doublet with a positive focal power it is necessary to take a positive lens with focal length f_1 made of a low-dispersive glass with a large Abbe number $V_{1,\mathrm{d}}$ and a negative lens with focal length f_2 made of a high-dispersive glass with a small Abbe number $V_{2,\mathrm{d}}$ (see Figure 5a). Therefore, the resulting focal power of the achromatic doublet

$$\frac{1}{f_{\mathrm{ges,d}}} = \frac{1}{f_{1,\mathrm{d}}} + \frac{1}{f_{2,\mathrm{d}}} = \frac{1}{f_{1,\mathrm{d}}}\left[1 - \frac{V_{2,\mathrm{d}}}{V_{1,\mathrm{d}}}\right] \qquad [24]$$

is always smaller than the focal power of the positive lens with focal power $l/f_{1,\mathrm{d}}$ alone.

The situation is completely different if a refractive and a diffractive lens are used. From the Sweatt model

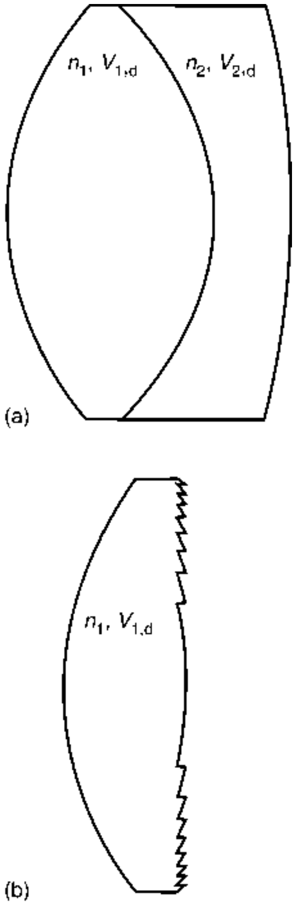
n_1, $V_{1,\mathrm{d}}$
n_2, $V_{2,\mathrm{d}}$
(a)
n_1, $V_{1,\mathrm{d}}$
(b)

Figure 5 Structure of achromats: (a) refractive achromatic doublet consisting of a positive lens with low dispersion and a negative lens with high dispersion; (b) hybrid achromatic lens consisting of a positive refractive lens and a positive diffractive lens.

of a DOE (eqn [3]) it is known that the refractive index n_{sweatt} is proportional to the wavelength λ

$$n_{\mathrm{sweatt}}(\lambda) = \frac{n_0}{\lambda_0}\lambda = \frac{n_{\mathrm{d}}}{\lambda_{\mathrm{d}}}\lambda \qquad [25]$$

Therefore, the Abbe number of a DOE is

$$\begin{aligned} V_{\mathrm{d}} &= \frac{n_{\mathrm{d}} - 1}{n_{\mathrm{F}} - n_{\mathrm{C}}} \\ &= \frac{\lambda_{\mathrm{d}}}{\lambda_{\mathrm{F}} - \lambda_{\mathrm{C}}}\left(1 - \frac{1}{n_{\mathrm{d}}}\right) \overset{n_{\mathrm{d}}\to\infty}{=} \frac{\lambda_{\mathrm{d}}}{\lambda_{\mathrm{F}} - \lambda_{\mathrm{C}}} \\ &= -3.452 \end{aligned} \qquad [26]$$

Here, we used the fact that the accuracy of the Sweatt model increases with increasing refractive index so that the Abbe number of a DOE is achieved in the limit of an infinite refractive index. So, it can be seen that the Abbe number of a DOE is a small negative constant. This means that the dispersion of a DOE is high and opposite in sign to that of a refractive element. Therefore, according to eqns [23] and [26] a positive refractive lens with a large focal power (i.e., small focal length) and a positive diffractive lens with a small focal power (i.e., long focal length) can be combined to form an achromatic doublet with a focal power larger than the focal power of one of the two lenses alone (see Figure 5b). By using eqns [19] and [25] the focal power of a diffractive lens can be calculated with the Sweatt model (n very big $\rightarrow n - 1 \approx n$):

$$\frac{1}{f_{\mathrm{DOE}}(\lambda)} = \frac{\lambda}{\lambda_{\mathrm{d}} f_{\mathrm{d}}} \qquad [27]$$

In practice, the lenses are neither thin nor is the distance between them zero. In this case the equations for the design of a hybrid achromatic doublet are more complex (see Further Reading). Nevertheless, the given equations can be used for a first design and to understand the principal behavior.

Another advantage of a hybrid achromatic doublet is that the DOE can in addition behave like an asphere. Therefore, e.g., the spherical aberration can be corrected with the aspheric terms of the DOE provided that the spherical aberration is considerably smaller than the phase term encoded in the DOE to correct the chromatic aberrations. Otherwise, the correction of the monochromatic (spherical) aberration will disturb the correction of the chromatic aberrations. A comparison of a simple hybrid achromat, consisting of a planoconvex lens with the diffractive structure made directly onto the plane surface of the refractive lens and a conventional refractive achromatic doublet shows that the optical performance of the hybrid achromat is a little bit worse than that of the refractive achromatic doublet regarding, for example, off-axis aberrations. A simple explanation of this is that a refractive achromatic doublet has three spherical surfaces and two glasses, i.e., in total five parameters, whereas a hybrid achromat has one spherical surface, one glass and the DOE with a paraxial term and an aspheric term, i.e. in total about four parameters. Of course, the merit of these different types of parameters is not identical so that this assessment is only qualitative.

In the case of hybrid achromatic elements for a broad wavelength range (e.g., for the whole visible wavelength range) the dependence of the diffraction efficiency on the wavelength may be a limiting factor because the other diffraction orders behave like stray light and reduce the maximum contrast of the image. Therefore, special laminated blazed diffractive optical elements made of laminae of two or more different materials are described in the literature which show a nearly achromatic behavior of the diffraction efficiency but the ‘normal’ dispersion of DOEs. DOEs are in the meantime used in zoom lenses (see the cited patent in Further Reading) to correct chromatic aberrations.

Summary

In this article the possibilities of correcting monochromatic and chromatic aberrations in optical systems by using diffractive optical elements have been discussed. There are several advantages and disadvantages of using DOEs.

Advantages:

- DOEs offer many degrees of freedom in the design. By using modern lithographic fabrication techniques the fabrication of a DOE with a strongly aspheric phase function is just as expensive as fabricating a simple Fresnel zone lens with a quadratic phase function. So, DOEs can be used in some cases instead of aspheric surfaces which are more expensive. In addition, the lithographic fabrication technique guarantees a quite good accuracy concerning the phase function of the DOE compared to the more difficult fabrication of well-defined aspheric surfaces.
- The strong chromatic dispersion with the opposite sign to that of a refractive element allows the correction of chromatic aberrations. DOEs also give in this case an additional degree of freedom in the design because aspheric terms of the phase function can correct in addition monochromatic aberrations of the system as long as these aberrations are considerably smaller than the quadratic term of the phase function which corrects the chromatic aberrations.

Disadvantages:

- DOEs often exhibit not only one diffraction order but several diffraction orders due to fabrication errors, a broad spectrum of application wavelengths or because some DOEs naturally have several diffraction orders like, for example, binary DOEs. This can produce disturbing light in the optical system and limit the application.
- The strong dispersion can restrict the application in monochromatic systems because it requires in some cases (e.g., for testing aspheric surfaces) not only a relative constancy, i.e., constant but the absolute value does not have to be hit exactly, of the illuminating wavelength but an absolute constancy because otherwise systematic aberrations are introduced by the DOE.

So, there are many advantages of DOEs compared to other optical elements but on the other hand a DOE is not a magic bullet in correcting aberrations in optical systems. Depending on the system a careful assessment of the advantages and disadvantages has to be performed in order to find the best solution.

List of Units and Nomenclature

Abbe number	real number	V_d
Angle	[rad]	φ
Cartesian coordinates	[m]	x, y, z
Coefficients of a polynomial	different negative powers of m	a
Components of the ray vector	real numbers	e_x, e_y, e_z
Degree of a polynomial	integer number	G
Diffraction order	integer number	m
Focal length	[m]	f
Grating frequencies in x and y direction	[m^{-1}]	ν_x, ν_y
Height of a lens surface	[m]	h
Local grating periods in x and y direction	[m]	p_x, p_y
Merit function	real valued function	M
Modulus of the wave vector	[m^{-1}]	k
Nabla differential operator $\begin{pmatrix}\partial/\partial x\\ \partial/\partial y\\ \partial/\partial z\end{pmatrix}$	[m^{-1}]	∇
Number of sampling points/rays	integer number	N
Optical phase, phase function	[rad]	Φ
Polynomial	real valued function	P
Radial coordinate	[m]	r
Refractive index	real number	n
Sign variable	either +1 or −1	σ
Symbol for partial differentiation		∂
Unit vector in ray direction	vector with three real components	$\mathbf{e}$
Wavelength	[m]	λ

See also

Geometrical Optics: Aberrations.

Further Reading

Born M and Wolf E (1997) *Principles of Optics*, 6th edn. Cambridge, UK: Cambridge University Press.

Fairchild RC and Fienup JR (1982) Computer-originated aspheric holographic optical elements. *Optical Engineering* 21: 133–140.

Falkenstörfer O, Lindlein N, Keinonen T and Schwider J (1996) Diffraction-limited holographic lenses in dichromated gelatine. *Optical Engineering* 35: 2026–2033.

Herzig HP (1997) *Micro-Optics*. London: Taylor & Francis.

Kogelnik H (1969) Coupled wave theory for thick hologram gratings. *The Bell System Technology Journal* 48: 2909–2947.

Kress B and Meyrueis P (2000) *Digital Diffractive Optics*. Chicester, UK: Wiley.

Latta JN (1971) Computer-based analysis of holography using ray tracing. *Applied Optics* 10: 2698–2710.

Lindlein N, Schwider J, Schrader M, *et al.* (1993) Wavelength mismatch correcting elements for volume holograms written on a laser pattern generator. *Journal of Modern Optics* 40: 647–661.

MacGovern AJ and Wyant JC (1971) Computer generated holograms for testing optical elements. *Applied Optics* 10: 619–624.

Meier RW (1965) Magnification and third-order aberrations in holography. *Journal of the Optical Society of America* 55: 987–992.

Nishio A (1999) Zoom lens with diffractive optical elements. United States Patent 5978153.

Offner A (1966) Ray tracing through a holographic system. *Journal of the Optical Society of America* 56: 1509–1512.

Sinzinger S and Jahns J (1999) *Microoptics*. Weinheim: Wiley-VCH.

Stone Th and George N (1988) Hybrid diffractive-refractive lenses and achromats. *Applied Optics* 27: 2960–2971.

Sweatt WC (1977) Describing holographic optical elements as lenses. *Journal of the Optical Society of America* 67: 803–808.

Upatnieks J, Vander Lugt A and Leith E (1966) Correction of lens aberrations by means of holograms. *Applied Optics* 5: 589–593.

Welford WT (1975) A vector raytracing equation for hologram lenses of arbitrary shape. *Optical Communications* 14: 322–323.

Winick K (1982) Designing efficient aberration-free holographic lenses in the presence of a construction-reconstruction wavelength shift. *Journal of the Optical Society of America* 72: 143–148.

Applications of Diffractive and Micro-Optics in Lithography

V Kettunen, Heptagon, Zurich, Switzerland

H P Herzig, University of Neuchâtel, Neuchâtel, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Most aspects of modern life are thoroughly influenced by the micro-electronics industry through microchips, such as microprocessors and memory chips. Optical lithography has been a key factor in the rapid pace of integration that drives the development of the micro-electronics industry. The photolithography process not only facilitates the fabrication of the chips, but also directly influences the limits of the obtainable chip performance in terms of speed or storage capacity, through the minimum feature size that can be created in the lithography steps. The increasing resolution demands of the IC industry, outlined in the International Technology Roadmap for Semiconductors as technology nodes, mean continuing demand for improvements in the photolithography systems.

The challenges to the improvements are encapsulated in the equation for the resolution of an optical lithography system:

$$R = k_1\lambda/NA \qquad [1]$$

which expresses the resolution R in terms of the smallest resolvable half-pitch as a function of the wavelength λ, the numerical aperture NA of the exposure system and the unit less constant k_1, which is used as a measure of lithographic difficulty, with smaller values indicating a more challenging process. Traditionally, the increase in the resolution has been obtained by decreasing the used wavelength, with the current mark for volume production systems currently standing at 193 nm. Within each wavelength, further reduction has been traditionally obtained by increasing the numerical aperture towards the mathematical limit of 1. Finally, several so-called Resolution Enhancement Techniques (RET) have been introduced to reduce k_1, in order to reach subwavelength technology nodes within each lithography wavelength generation. It is in connection with these techniques that micro-optics has found a role in modern lithography systems.

In this chapter we discuss the ways micro-optics are being utilized in lithography systems. We will start with a brief overview of conventional lithography systems, of the essential parts of such systems as well as the special resolution enhancement techniques utilized in the systems. We will then discuss the utilization of micro-optics within the context of lithography systems, along with the unique challenges in terms of design and fabrication that are associated with such applications. Finally, we will briefly discuss nonconventional lithography applications, especially those based on utilization of micro-optics. It should be noted that the motivation of this article is to describe the issues related to the use of micro-optical elements in a modern lithography system and is not intended as a review of the modern lithography systems, nor is it a review of micro-optics.

Overview of Conventional Lithography Systems

Photolithography is a process whereby a specified pattern or mask is transferred onto a wafer using photons. In the process a photosensitive material, commonly known as a resist, is initially used to record the pattern that is generated when the mask is illuminated. After exposure, the resist is developed. The resist is referred to as positive or negative, depending on whether it is removed or remains, respectively, after the development of the irradiated regions. After its development, the pattern in the resist is then transferred onto the wafer using suitable chemical processes, such as wet chemical etching or dry plasma etching.

Several different lithographical techniques exist. The oldest and simplest is contact printing, where a photomask is brought into contact with a layer of photoresist on the wafer. Although this technique can reproduce high-resolution features, the contact between the wafer and the mask can lead to problems with cross contamination, and can damage either the wafer or the mask. Contact printing is mainly used nowadays in laboratory environments for small series of photolithography steps. A slightly more complicated technique is proximity printing, in which the photomask is held in close proximity to the surface of the wafer without actual contact. This avoids the problems of contamination and damage, but due to diffraction effects, reduces the maximum resolution of the process. Nevertheless, proximity printing remains a valuable technique for lithography applications, that do not require the highest possible resolution.

Modern volume production photolithography systems are based on the use of a projection system to image a mask onto a wafer through a complex system of lenses. There exist several variations of the basic idea. In some systems the entire mask is imaged at once onto the wafer. This approach works only when the feature size on the mask is reasonably large, in the range of a few microns. Reproducing smaller features with this approach is generally not possible, as the demands on the optical quality of the projection system, in the form of allowed aberrations, become extremely difficult to meet, especially when the physical size of the optical components increases with the reducing feature size. Therefore, so-called wafer steppers are used in modern microchip fabrication. The basic idea of a wafer stepper is to image only a small region of the mask onto the wafer, and pattern the entire surface area in consecutive steps.

The centerpiece of a wafer stepper is the exposure system. Its task is to image desired patterns from the photomask to their proper positions on the surface of the wafer. In many systems this process includes a size reduction between the patterns on the photomask and on the wafer. This reduces the dimensional accuracy requirements, such as feature size and positioning for the mask fabrication by whatever is the de-magnification of the lithography system. However, it also increases the complexity of the exposure system, as significance of aberrations is increased. The requirements of microchip production mean that the image of the photomask with submicron features must be reproduced on the wafer with a dimensional accuracy of only a few tens of nanometers and aligned to a specific position within a fraction of the linewidth. Furthermore, these tolerances must be guaranteed over the entire exposure field that is typically several square centimeters. It is not surprising that the resulting complex piece of optics costs several millions dollars.

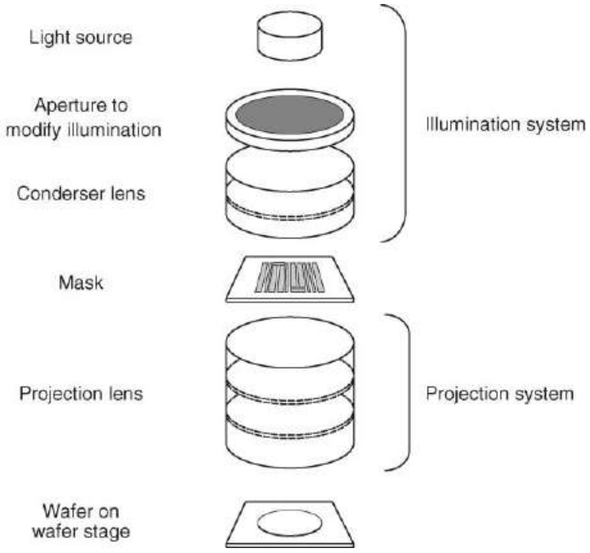

Figure 1 Schematic illustration of the exposure system.

The exposure system consists of a projection lens, illumination system, and wafer management system, as illustrated schematically in Figure 1.

A typical projection lens is a complicated lens system. As the resolution of a lens system is inversely proportional to the diameter of the lenses, the lenses in photolithographic systems can have diameters up to 250 mm, and the whole system is arranged inside a massive barrel-frame. The aim of the projection lens in the entire exposure system is to image the pattern from the photomask to the wafer. The projection lens is primarily responsible for the resolution and must therefore fulfill the highest optical requirements. To produce the smallest possible image, diffraction effects alone must essentially limit the resolution of the system. This means that the optical aberrations over the exposure field of several square centimeters must be virtually eliminated, by the wavefront deviations being less than a fraction of the wavelength.

The illumination system transfers the illumination source through the photomask to the entrance of the projection system. The main requirements for the illumination system are to collect most of the light into the system and to ensure that the irradiance of the photomask is uniform to within a few percent. In addition, resolution enhancement techniques, based on off-axis illumination, require that the illumination pattern on the photomask has a specific shape. We will discuss these requirements later in more detail. In general terms, the illumination system is less complicated than the projection system, especially with regard to the elimination of aberrations, due to the fact that precise imaging is not required in the illumination system.

Although the wafer management system is outside the scope of this article, we note in passing that it maintains the position, orientation, and movement of the wafers through the exposure system. Again, high precision is required to ensure the successful exposure of the desired pattern to the wafer.

Resolution Enhancement Techniques

In most modern lithography systems the achievable maximum resolution is increased via the use of one or more special resolution enhancement techniques (RET). These are essentially creative refinements of the lithography technology, based on theoretical optics papers published often decades before their first application in the lithography process in the late 1980s and early 1990s. There are essentially three RET approaches: Off-Axis Illumination (OAI), Optical Proximity Correction (OPC), and Phase Shifted Masks (PSM).

In optical proximity correction the basic idea is to predistort the mask pattern in order to overcome the proximity effects caused by the optics and the different processes. This is done either using a set of pre-existing rules describing the connection between a mask pattern and its realization on the wafer, using advanced models to predict the proximity effects during the lithography process, or with some combination of both. The key difficulties with OPC are related to the fabrication of the predistorted mask and to the accuracy of the used distortion rules or proximity effect models. Finally, it should be noted that OPC is strictly speaking not a true resolution enhancement technique, as it only enhances the effective resolution and not the theoretical resolution limit of the process defined by the optics.

In phase shifted masks the improvement in the resolution is obtained by altering the phase of light passing through different portions of the mask, creating regions of destructive interference in the image. This leads to higher contrast, which helps to improve the effective patterning resolution. However, as with OPC, there is no true resolution enhancement in terms of the theoretical resolution limit. There are two main difficulties with application of PSM. First, it is not straightforward to introduce required phase shifts into an arbitrary design layout, because some geometrical shapes, such as junctions are harder to realize than others. Another problem lies with the mask fabrication, which requires careful calibration of the material layers and can be very complicated and costly compared to fabrication of conventional binary masks.

For the purposes of this article, only off-axis illumination is discussed, as both OPC and PSM are essentially mask-based techniques and thus provide no opportunity for micro-optics. Off-axis illumination refers to any illumination scheme that significantly reduces or eliminates the light hitting the mask at near normal incidence, leading to the diffraction pattern of the mask being shifted on the projection lens. The principle behind the resolution enhancement, obtained by introducing such a shift, is schematically illustrated in Figure 2 for periodic pattern.

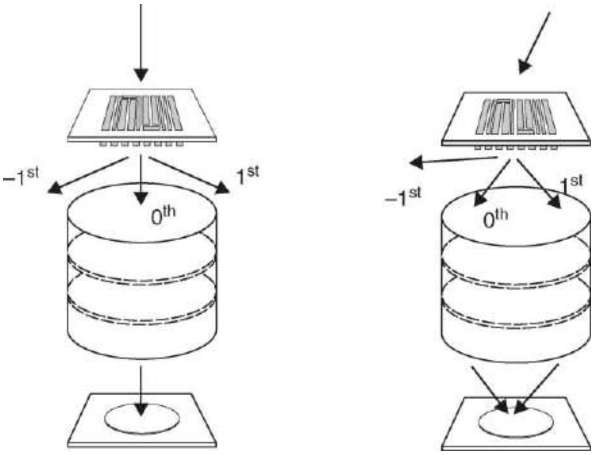

Figure 2 Schematic illustration of effect of off-axis illumination on image formation.

The mask pattern can be imaged if at least two diffraction orders generated by the mask pass through the pupil of the projection lens. Using on-axis illumination, the minimum feature size of the system is reached when the period of the mask pattern is so short that the ± 1st diffraction orders can no longer pass through the pupil. If the illumination is moved off-axis, the diffraction pattern is shifted and the second diffraction order (in this case +1st) again passes the pupil. Thus the resolution has been increased. A secondary improvement in contrast may be obtained by shaping the diffraction orders falling onto the pupil, such that the filling of the pupil is optimized, i.e., such that less unwanted light contributing to the background illumination passes the pupil. Finally, the depth of focus of the system can be optimized by ensuring that the contributing orders (in this case 0th and +1st) are symmetric with regards to the optical axis.

The conventional way to realize off-axis illumination schemes is to introduce a suitable aperture to the illumination optics between the source and the mask. In order to effectively use off-axis illumination, the shape and size of the aperture, i.e., the illumination pattern, must be optimized for the specific mask shape and pitch. This optimization leads to four basic shapes, as illustrated in Figure 3. In most modern lithography systems these shapes

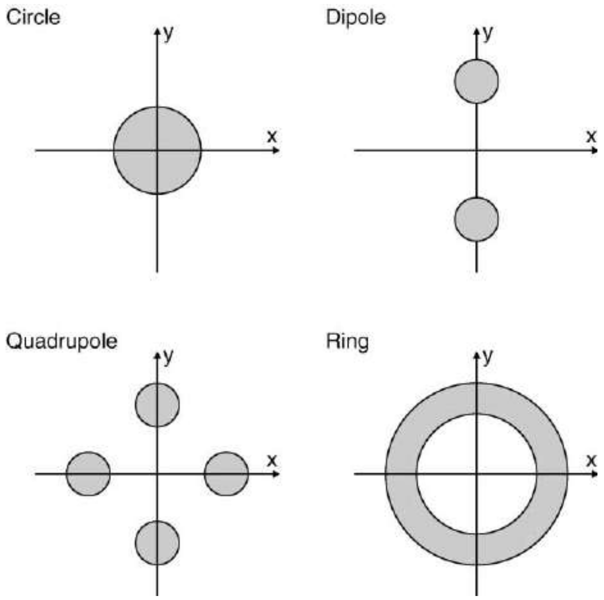

Figure 3 The four typical illumination patterns: disk, ring or annular, quadrupole, and dipole.

are available with different values of partial coherence factor, governed by the ratio of the numerical aperture of the illumination lens and projection lens, allowing control of proximity effects and edge sharpness.

Overview of Micro-Optics

There exists no generally agreed definition of micro-optics, but it typically means optical systems built to significantly smaller scales than conventional table-top systems, using components such as micro-lenses and microprisms with physical dimensions in the range of a couple of millimeters at most. Micro-optical systems also include so-called diffractive optical elements, which are usually optical components whose operation cannot be explained by Snell's law or the law of reflection. Such elements are usually macroscopically planar microstructures, consisting of features with dimensions from a couple of wavelengths to a few tens of microns, and are designed by advanced numerical algorithms based on diffraction theory.

In the past decade, micro-optics has emerged as a powerful new addition to the field of optics and optical systems, following improvements in micro-optics fabrication techniques, both via adaptation from established microelectronics fabrication processes and development of novel approaches, such as variable dose direct writing using either UV light or an electron beam. Additionally, emergence of replication techniques, such as injection molding or UV casting, allows the low-cost mass production of micro-optical elements, making micro-optics economically feasible for a wide range of consumer applications. Consequently, micro-optical elements, utilizing either refractive or diffractive surfaces, are now found in industrial applications such as material processing, in the optical telecom field, and in consumer electronics such as the digital video camera.

One of the major strengths of micro-optics, as compared to conventional optics, is their ability to realize complex optical functions into a compact form. This is especially true for micro-optics using diffractive optical elements, as diffractive elements may be designed to realize several optical functions such as focusing, filtering, or beam splitting at once allowing integration of several classical optical components into a single diffractive element.

Micro-Optics in Conventional Lithography Systems

The simplicity of contact or proximity printing means that there is little need or opportunity to incorporate additional functionality through inclusion of micro-optical elements. The only exception to this will be discussed in the final section of this article, where we will consider a nonconventional technique, which utilizes micro-optics and can be incorporated into proximity printing systems. However, in this section we will limit the discussion to projection-based lithography systems such as wafer steppers.

Let us consider the possible uses of micro-optics in the three main parts of the exposure system identified in the earlier section. It is obvious that the wafer management system, which is essentially mechanical, cannot be enhanced by inclusion of micro-optics. In principle, micro-optics could be incorporated into the projection system to reduce its weight and, to a lesser degree, complexity. However, it is important to note that using micro-optics to reduce the physical size, i.e., the diameter of the lenses in the projection system is not possible, as such reduction would increase the theoretical minimum feature size the system can resolve which is inversely proportional to the lens diameter. In practical terms, even the most advanced fabrication approaches used to realize micro-optical elements cannot yet produce elements with the precision and quality sufficient to meet the optical requirements associated with a lithographic projection lens. In general, the state of the art in micro-optics is such that it does not make sense to replace refractive lenses with diffractive ones unless the weight of the optics is a critical parameter in determining the merit of the system. This is not the case for wafer steppers.

Even though the optical requirements for the lenses in the illumination system are not as high as in the projections system, they are in practice still far too demanding for micro-optical elements. Therefore, the only area where micro-optics could be used in the illumination system is the generation of various illumination patterns needed for the off-axis illumination schemes used for resolution enhancement, as well as overall homogenization of light. An element meeting these goals is essentially a custom-designed diffuser, where both the generated angular range and the relative intensity of illumination in any given angular direction are controlled to a high degree.

Custom Diffusers for Illumination

As mentioned earlier, one of the strengths of micro-optics lies in its ability to combine different optical functions into a single element. When used in lithographic illumination systems, micro-optical elements can serve a dual purpose. The element can serve as an artificial aperture, effectively distributing the light over some predetermined angular range and at the same time minimizing light outside the range, allowing generation of various illumination patterns needed for the resolution enhancement via off-axis illumination. Unlike conventional solutions based on the use of actual apertures to generate the desired illumination patterns, distribution of light using micro-optical elements can be done with efficiencies up to 95%, with typical target efficiency being in the range of 90%, depending on the pattern complexity. At the same time, the element can be used to smooth the light distribution, improving the uniformity of the illumination. Typically the required uniformity error, i.e., allowed variations from the average illumination, is in the range of 5%, but can be as low as 1%.

In addition to the above-mentioned high-efficiency and -uniformity targets, optical quality requirements of modern lithography systems introduce unique challenges to the micro-optics used in such systems. First, micro-optical elements used in lithography systems must also exhibit low amounts of stray light, especially in formation of hot-spots. Typically, hot-spots are allowed to be only a few percent of the average illumination level. Because of the fact that the size, shape, and profile of the illumination beam can vary, partly due to variations in the source and partly because such variations improve the functionality of the lithography system, by, for example, controlling the partial coherence factor, the micro-optical elements must be virtually space-invariant. The elements must perform the same function regardless of where and how the illumination coming from the source hits the element. Furthermore, the elements themselves must have very low absorption, because the high power used could otherwise lead to destruction of the elements. This means that the materials must be carefully selected. Finally, the optical wavelengths used in advanced lithography systems are 248 nm or less. Because the feature size of a diffractive element scales down with the wavelength, the features in elements used are much smaller than in the case of elements used for applications in the visible region. Consequently the accurate fabrication of such elements becomes a formidable challenge, because most fabrication errors, such as mask misalignment, etch depth error or the way the sharp corners in diffractive elements get rounded during fabrication, increase in significance when the feature size is decreased.

There exist two distinct approaches for obtaining such custom-designed diffusers. We will now discuss the benefits and weaknesses of these two techniques in terms of efficiency, pattern uniformity, stray light, and ease of fabrication, in order to illustrate the trade-off a designer creating micro-optical elements for a lithography system has to consider.

Aperture Modulated Diffusers

One approach to acquire custom diffusers for illumination systems is to use an array of refractive or diffractive micro-lenses. More generally, the lens phase can be replaced by any phase obtained by geometrical considerations, i.e., by a geometrical map transform between the input and output planes of the optical system. The basic principle of an aperture modulated diffuser is to use the phase function to accurately control the angular divergence of the diffuser, while the shape of the desired intensity distribution is created by introducing a properly shaped aperture in which the phase function is contained. This basic principle is shown in Figure 4.

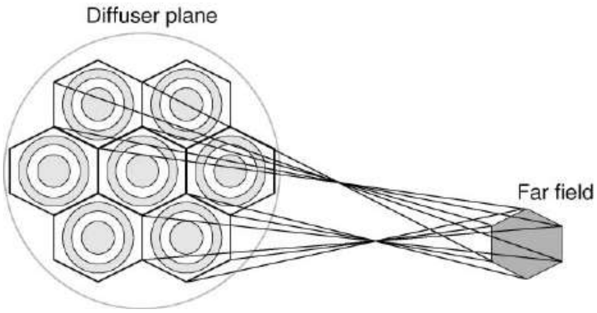

Figure 4 Basic principle of aperture modulated diffuser. The lens f-number determines the angular extend of the far field pattern, while the lens aperture shape determines the shape of the pattern.

The main benefit of using a lens phase lies in the ease of the design process, which essentially consists of selecting the numerical aperture of the lens to meet the angular size of the desired pattern, while more complicated phase profiles can be used to improve some aspects of the performance of the diffuser.

In terms of optical performance, aperture modulated diffusers typically achieve efficiency of up to 95% and low levels of stray light, easily fulfilling the target specifications outlined earlier, especially when refractive surfaces are used. However, since the lens aperture in the element plane controls the shape of the intensity distribution, the range of intensity distribution shapes is limited. This is necessary because it must be possible to tile the required aperture shape in order to realize an array of elements with high fill factor, i.e., to realize a space-invariant diffuser. Furthermore, introducing an aperture into the element plane means that intensity oscillations, due to boundary diffraction at the edge of the apertures, cannot be avoided. Consequently, the uniformity of the generated intensity distributions suffers. This is especially true in cases where the desired intensity pattern has sharp corners, where the intensity oscillations due to boundary diffraction, can easily be up to 50% of the average intensity of the desired intensity pattern. Figure 5 illustrates a typical far field intensity distribution obtainable with aperture modulated diffusers.

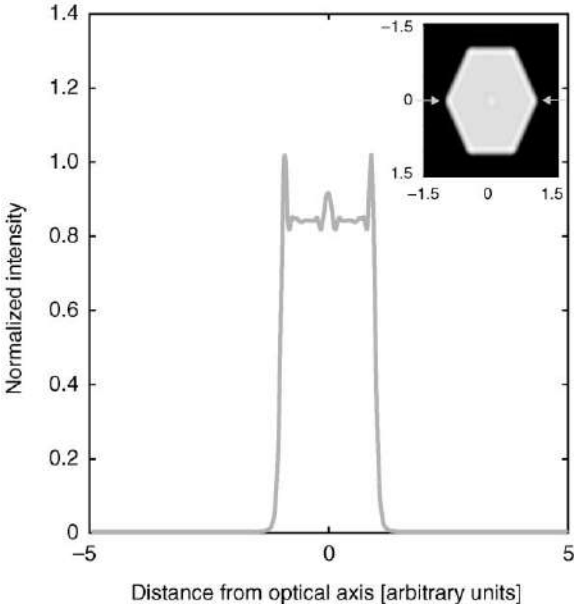

Figure 5 Typical intensity distribution obtainable with an aperture modulated diffuser. The overall distribution is smooth apart from the slow intensity oscillations at the edges resulting from diffraction at the aperture boundary.

One needs to consider several issues with regard to the fabrication of aperture modulated diffusers. For refractive aperture modulated diffusers, the main issue is the accurate realization of the required surface profile. The fabrication of large arrays of refractive elements having arbitrary surface profiles is still challenging. Resist reflow techniques, where the micro-lenses are realized by melting small islands of resist, lead to a spherical lens profile. Complicated and hard-to-control profile reshaping techniques are required if dramatic alterations from the spherical lens profiles are needed. Even though fabrication through direct writing techniques, such as laser writing offers more freedom, realization of arbitrary refractive surface remains a challenge. Therefore, use of refractive elements is mainly limited to cases where the lens profile can be close to the spherical shape. The most important of such cases are patterns with uniform circular or rectangular distributions. The former can be realized by proper selection of the numerical aperture of the lens, while the latter can be realized by combining two properly selected cylindrical lenses on opposite sides of the substrate.

For diffractive aperture modulated diffusers there are considerably less limitations in terms of what phase profiles can be attained. As long as the opening angles, i.e., the angular size of the pattern is below 10 degrees, the phase can be realized as a diffractive element. With diffractive elements, the main issue to consider is therefore the effect of typical fabrication errors such as profile depth and shape errors. First, the typical consequence of fabrication errors in the element is loss of efficiency via increase of stray light. Fortunately the errors do not generate sharp stray light peaks, with the obvious exception of the axis hot-spot. This results from the fact that both lens functions and more complicated phase functions designed by geometrical considerations, exhibit locally periodic structures, and the introduction of common fabrication errors then leads to redistribution of light away from the desired signal orders of these local structures to higher or lower orders. On the level of the whole intensity distribution, such redistribution appears as a constant rise in the background noise level, but will not lead to sharp intensity fluctuations.

Another issue related to the fabrication of aperture modulated diffusers appears when diffuser elements are fabricated using a series of photomasks, as it is typically the case with a relatively large size of the element. Although mask misalignment is essentially a random error, i.e., the direction and amount of the error are difficult to predict before hand, it easily leads to systematic and partly symmetric uniformity problems with aperture modulated diffusers.

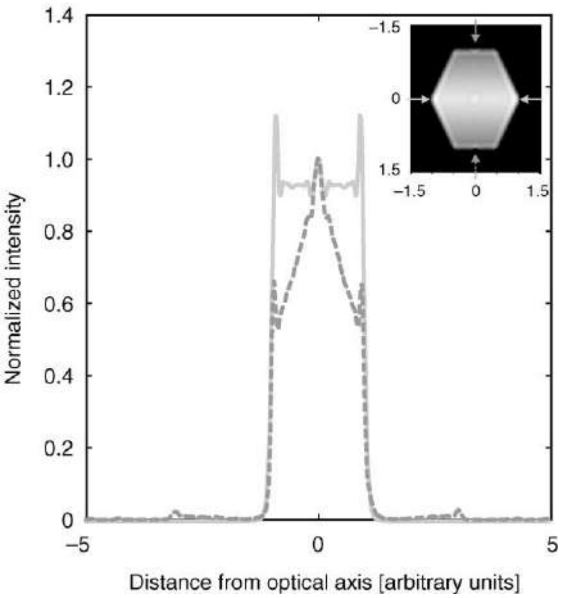

Figure 6 Effect of mask misalignment on the intensity distribution with aperture modulated diffusers. Essentially random error leads to symmetric and systematic uniformity error due to locally periodic nature of the design.

Again the reason for this lies in the locally periodic nature of the elements. A single mask misalignment along some direction results in an erroneous profile shape in the structures, which are periodic in the direction of the alignment error. Consequently, the relative intensity distribution along that direction is changed. On the other hand, structures, where the period is perpendicular to the misalignment, are realized correctly. Furthermore, the local period of the phase function is often changing from the center of the element to the edge, meaning that the relative magnitude of the mask misalignment is different in different parts of the element. These facts lead to asymmetry in the overall intensity distribution, which is especially harmful with off-axis illumination patterns, as the asymmetry results in loss of the resolution gain obtained with off-axis illumination. Furthermore, as the source of the asymmetry is essentially a random error, compensation in the design is very difficult. Figure 6 illustrates the above-mentioned phenomena.

Fan-Out Based Diffusers

An alternative technique uses specially designed fan-out elements to produce a large number of closely spaced diffraction orders arranged in a desired pattern, as illustrated in Figure 7. Such an approach offers all the flexibility of diffractive elements, allowing generation of nearly arbitrary intensity distributions, especially if the elements can be realized as so-called multilevel elements, i.e., if the surface profile can have more than two distinct levels.

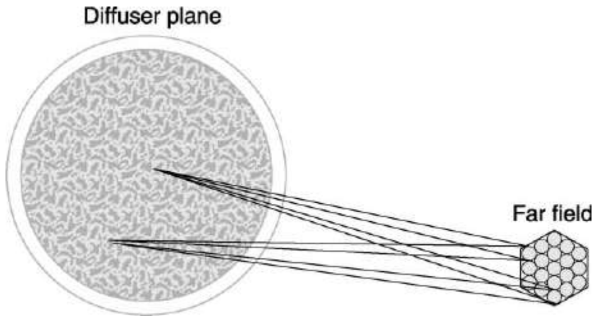

Figure 7 Basic principle of fan-out based diffusers. A grating is used to generate a high number of closely spaced diffraction orders arranged in the shape of the desired pattern.

The flexibility in the design comes at a cost, where specially developed design tools, such as the iterative Fourier transform algorithm, are required. Fortunately such tools are starting to become commercially available as emphasis and interest in wave-optical engineering, an extension of the classical optical engineering approaches, increases.

In terms of optical performance, fan-out based diffusers offer high uniformity as there are no boundary diffraction effects as with aperture modulated diffusers. Furthermore, by choosing the beam separation of the fan-out to be clearly smaller than the divergence of the incidence beam, it is possible to average out any local intensity variations from order to order, leading to a smooth overall intensity distribution. The remaining variations in the intensity distribution generated by a fan-out based diffuser can be as small as 1% of the average intensity.

As a trade-off for the improved performance in terms of uniformity compared to aperture modulated diffusers, fan-out based diffusers typically exhibit lower efficiency and, consequently, higher levels of stray light, even when no fabrication errors are present. This is a result of additional design freedoms being used to remove unwanted intensity oscillations from the far field regions, belonging to the desired intensity distribution, by introducing some noise outside these regions. Typically the efficiency of a fan-out based diffuser is in the range of 85–90%, or 5–10 percentage points lower than the efficiency of a comparable aperture modulated diffuser (should one exist). Furthermore, unlike aperture modulated diffusers, fan-out based diffusers can have stray light that forms hot-spots. If care is not taken during the design, these hot-spots may be up to 10% of the average intensity of the illumination pattern, even in the absence of fabrication errors. Figure 8 illustrates

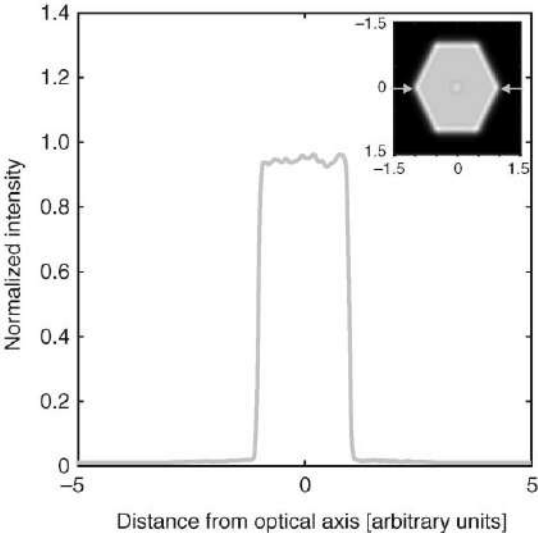

Figure 8 Typical intensity distribution obtainable with a fan-out based diffuser. Overall distribution is smooth, but stray light appears outside the main pattern.

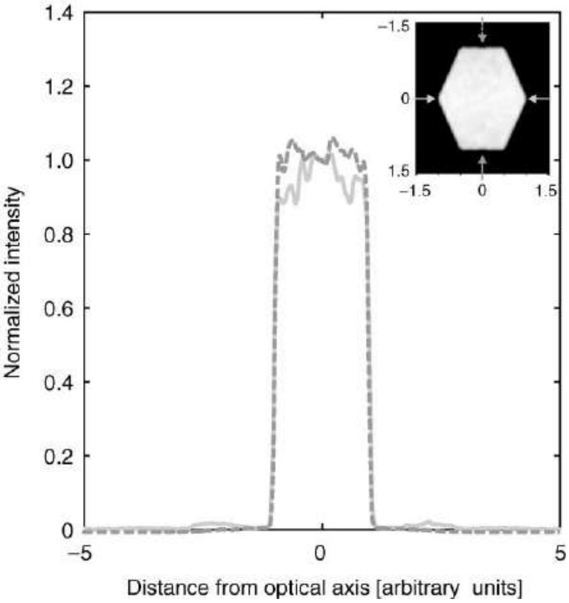

Figure 9 Effect of mask misalignment on the intensity distribution with fan-out based diffusers. The uniformity error appears randomly, allowing statistical compensation schemes.

a typical far field intensity distribution obtainable with fan-out based diffusers.

Fan-out based diffusers offer several advantages in terms of fabrication compared to mapping based diffusers. First, random errors, such as mask misalignment, do not lead to symmetric or systematic uniformity problems, but instead result in random local oscillations in the intensity pattern as illustrated in Figure 9. This allows special compensation schemes where the element is divided into segments, each of which has different fabrication tolerances but generates the same far field intensity distribution. If the fabrication tolerances of the individual designs are balanced correctly against the typical fabrication errors of the used fabrication methods, the loss of uniformity due to fabrication errors can be eliminated almost entirely. Furthermore, the flexibility of diffractive design allows for fabrication error-tolerant designs. This follows from the ability to combine several optical functions in a single optical element. The typical trade-off with designs with relaxed fabrication tolerances is that the optical performance of the element in terms of efficiency is worse than that of a conventional design if fabrication errors are not present.

Micro-Optics and Non-Conventional Lithography

In the previous section we have considered micro-optics in connection with conventional lithography approaches and shown that micro-optical elements can find a significant role in such systems. We conclude this article by considering briefly nonconventional lithography systems, focusing especially on a technique called micro-lens lithography, which is fundamentally based on the use of micro-optical elements.

The most straightforward way to improve existing lithography approaches is to further reduce the wavelength used in the exposure process. In X-ray lithography, the light source used in conventional lithography approaches is replaced by a cyclotron emitting light with wavelengths deep in the X-ray region. This leads to minimum resolution far lower than achievable by conventional techniques, but with the additional difficulty and cost of maintaining and running such a source.

The techniques that are currently considered as serious alternatives for conventional photolithography approaches are based on the use of particles other than photons in the fundamental exposure process. Techniques such as electron- or ion-beam lithography are routinely used to realize structures with feature size and positioning resolution of some tens of nanometers for use in both micro-electronics and micro-optics. The main drawback of these approaches is the cost involved in patterning large surface areas. Unlike optical lithography techniques, particle beam lithography cannot expose large surface areas at once, instead requiring painstaking scanning of the surface area with a relatively small writing beam. This leads to processing times that scale linearly with the required surface area, and consequently the throughput of such techniques is small. Therefore, particle beam lithography is not yet widely used in commercial fabrication of micro-electronics. However, it is worth noting that especially electron beam lithography is widely used in the origination of micro-optical components, as the relatively high cost of creating the master component can be offset in the final product by the inexpensive replication process. Such an approach allows the replicated end-product to meet the high optical requirements achievable only by high resolution surface control, yet remain affordable for use in, for example, consumer electronics. This has lead some in the micro-optics community to conclude that electrons are better suited for fabrication of components controlling photons, while photons are better suited for fabrication components controlling electrons.

In addition to particle beam lithography, other noteworthy techniques exist. For our purposes, the most interesting technique is the so-called micro-lens array lithography, where one or more micro-lens arrays are used to image the pattern of the photomask to the resist layer on the substrate. Although the resolution of this approach is limited to 3–5 microns, for applications such as with flat panel displays, micromechanics, or multichip module fabrication, this technique is of considerable interest.

It is well known that the geometrical aberrations scale down proportionally with the physical lens size. Thus, large areas can be imaged favorably by use of arrays of micro-lenses, where each lens images only a small section of the whole field. Such a setup is shown in Figure 10.

With this setup it is possible to expose a comparatively large area with a single exposure, thus eliminating the need for costly stepping operations and therefore increasing the throughput of the system. Analysis of the basic system shows that an array based on spherical micro-lenses of 1 mm focal length and 300 micron aperture can produce a nearly diffraction limited resolution on the order of 3–5 micrometers over a 300 micron field.

In terms of conventional lithography systems, the micro-lens array lithography has most in common with proximity printing. Both techniques achieve comparable resolutions, yet avoid damaging the mask or substrate by keeping the two separated by some finite working distance. However, micro-lens array lithography offers two distinct advantages over conventional proximity printing. First, the working distance between the imaging system and the substrate surface can be much larger, which further minimizes the risk of mask or substrate damage. Furthermore, added working distance allows lithography steps on nonflat surfaces, e.g., inside holes or grooves where conventional proximity printing would not be possible. Another advantage of the micro-lens array lithography is the larger depth of focus. With proximity printing, the resolution rapidly decreases with increasing distance. In micro-lens array lithography, the optimum resolution is obtained at the image plane, while a region of reasonable resolution extends forward and backward from this plane (Figure 11). This extension means that in the depth of field, resolution is less sensitive to changes in the working distance, which makes it better suited for practical applications.

In addition to the optical benefits of micro-lens array lithography, there also exist other practical benefits. Many micro-devices have a repetitive and

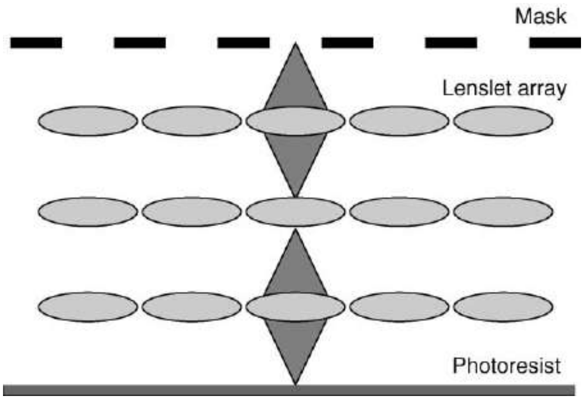

Figure 10 Basic geometry of micro-lens array lithography.

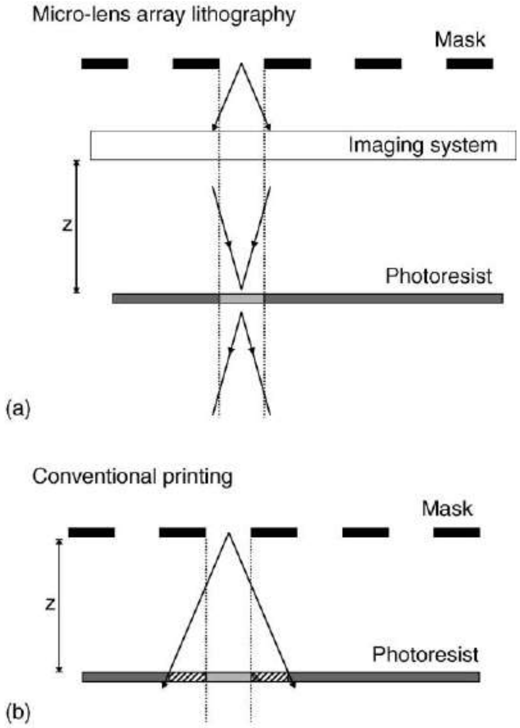

Figure 11 Depth of focus with micro-lens array lithography and conventional printing. (a) With micro-lens array lithography, the image of the mask is extended both forward and backward from the working plane at distance *z*. (b) In conventional printing, the resolution is rapidly reduced with increasing working distance *z*.

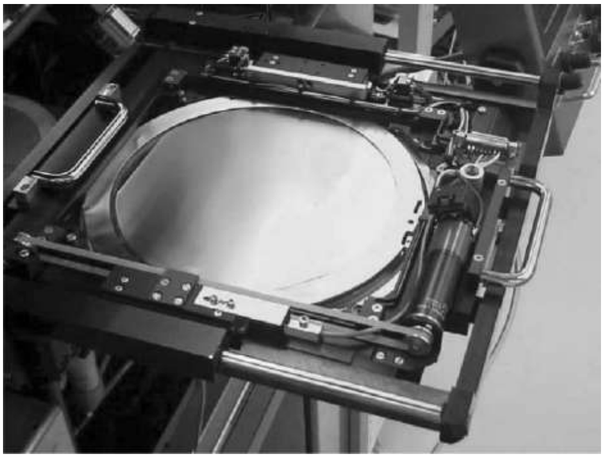

Figure 12 Ultra-flat micro-lens projection system (courtesy of R. Völkel, SUSS MicroOptics, Neuchâtel, Switzerland).

discrete nature with respect to some unit element. Such a unit element lends itself well to the imaging properties particular to lens arrays, as each lens of the array can be made to correspond to the unit cell of the particular micro-device making the imaging highly efficient. A photograph of a micro-lens projection system is shown in Figure 12.

One could also imagine that instead of an array of imaging systems, an array of pattern generators is used that generates a specific pattern, such as lines or points. In the ideal case such a system would replace the mask and the imaging system altogether. However, this approach is only possible for simple patterns and large production series.

Outlook

As long as optical lithography stays attractive, micro-optics will be used as beam-homogenizers and sophisticated diffusers to generate optimized illumination patterns. How far micro-optics will be used in the imaging system is difficult to predict. However, it can be assumed that the difference between micro-optics and macro-optics fabrication technology will disappear.

Recent research is also investigating near-field optics to further reduce the resolution limit. Micro-optics technology is well suited to realize such systems. A serious problem of near-field optics is the short distance (10 nm–100 nm) between the optical element and the interface to be patterned. In addition, near-field optics is based on scanning systems, which increases the exposure time considerable. However, there is still room for improvements or as Richard Feynman said 'there is plenty of room at the bottom'.

Further Reading

Dändliker R, Gray S, Clube F, Herzig HP and Völkel R (1995) Non-conventional techniques for optical lithography. *Microelectronic Engineering* 27: 205–211.

Herzig HP (1998) *Micro-Optics: Elements, Systems and Applications*. London: Taylor & Francis.

Herzig HP and Kipfer P (1999) Aperture modulated diffusers (AMDs). In: Asakura T (ed.) *International Trends in Optics and Photonics*. Berlin: Springer.

Naumann H and Schröder G (1992) *Bauelemente der Optik*. München, Germany: Carl Hanser Verlag.

Sheats JR and Smith BW (eds) (1998) *Microlithography: Science and Technology*. New York: Marcel Decker.

Turunen J and Wyrowski F (eds) (1997) *Diffractive Optics for Industrial And Commercial Applications*. Berlin: Wiley-VCH.

Völkel R, Herzig HP, Nussbaum P, Dändliker R and Hugle WB (1996) Microlens array imaging system for photolithography. *Optical Engineering* 35: 3323–3330.

Design and Fabrication of Diffractive Optical Elements

D W Prather, T Dillon, A Sure and X Gao,
University of Delaware, Newark, DE, USA

J N Mait, US Army Research Laboratory, Adelphi, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Diffractive optical elements (DOEs) are passive devices that redirect and focus light through the division and mutual interference of a propagating electromagnetic wave. This is in contrast to refractive elements, which redirect and focus light through variations in indices of refraction. From a historical perspective, diffraction was traditionally considered a detriment in optics and caused images to blur. However, we now know that it is possible to use diffraction to enhance the performance of optical systems. For example, DOEs can be used to correct chromatic aberrations and to implement optical functions that conventional refractive and reflective optics cannot, such as multiple beamsplitting.

Further, recent advances in computing and fabrication technologies have increased the number of applications in which diffractive elements are used, which has established diffractive optics as yet another tool in an optical designer's kit.

The first modern diffractive element was the hologram. A hologram is a photographic record of the constructive and destructive interference between two coherent beams known as the object and write beams. The object beam contains information about the object recorded. The write beam is necessary to record the hologram but contains no information itself. When the hologram is played back with a third beam, the read beam, the recorded interference pattern diffracts it in such a way that it reconstructs the object beam.

Unfortunately, the functions that a natural hologram can produce are limited to those whose wavefields can be produced in a laboratory. Thus, although it is possible to record a hologram that can focus or deflect a single beam, it is difficult to record a hologram that splits a single beam into multiple foci. However, the advent of computing technology made it possible to design such holograms analytically. The first diffractive element designed with the aid of a computer was demonstrated in 1966. Since then the ability to create elements that respond arbitrarily has had widespread application in imaging systems, telecommunications, and sensing.

The two most basic diffractive elements are the grating and the Fresnel zone plate. The diffractive behavior of a grating is predominately transverse to the optic axis. It causes light to split into discrete beams, referred to as diffraction orders, each with a different angle of propagation relative to the optic axis. The diffractive behavior of a Fresnel zone plate also divides incoming light into discrete orders, however, the orders are produced along the optic axis. This is useful in producing the diffractive equivalent of a refractive lens. Arbitrary transverse and longitudinal behavior is realized by combining characteristics of these basic two diffractive elements. In the remainder of this article, we discuss the design and fabrication of general diffractive elements.

Design

Once an optical system designer determines the functional behavior required for a particular application, for example, focusing or multiple beamsplitting, he or she needs to design a diffractive element that realizes this function. However, unlike refractive and reflective optical elements, whose material properties and surface curvatures determine their functionality, the functionality of a diffractive element is determined by a pattern of wavelength scale features. Determining this pattern is the goal of diffractive design but to do so one must first understand the physical relationship between the DOE and the wavefield it produces.

Scalar diffraction theory is one of the simpler models that relates the effect of a surface relief, or diffractive profile, $d(x,y)$, on an incident wavefield $U_{\text{in}}(x,y)$:

$$U_{\text{out}}(x,y) = \exp[\mathrm{i}\theta(x,y)]U_{\text{in}}(x,y) \qquad [1]$$

where the phase transformation $\theta(x,y)$ is related to the diffractive profile $d(x,y)$ by

$$\theta(x,y) = 2\pi(n-1)d(x,y)/\lambda \qquad [2]$$

The constant n is the refractive index of the optical substrate and λ is the wavelength of illumination in a vacuum.

The effects of the phase change are not apparent until the wavefield $U_{\text{out}}(x,y)$ propagates over some distance z. In scalar diffraction theory the relationship between $U_{\text{out}}(x,y)$ and the wavefield $U_z(x,y)$ at a plane a distance z away, is described by the Rayleigh–Sommerfeld diffraction equation:

$$U_z(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\text{out}}(\zeta,\eta) \times \frac{\exp\left\{\mathrm{i}(2\pi z/\lambda)\sqrt{1+(x-\zeta/z)^2+(y-\eta/z)^2}\right\}}{\mathrm{i}z\lambda\left[1+(x-\zeta/z)^2+(y-\eta/z)^2\right]}\mathrm{d}\zeta\,\mathrm{d}\eta \qquad [3]$$

Huygens' principle is reflected in this formulation by the convolution between the transmitted field $U_{\text{out}}(x,y)$ and a spherical wave $h(x,y)$, which represents free space propagation over the distance z:

$$U_z(x,y) = U_{\text{out}}(x,y)^{**}h(x,y) \qquad [4]$$

where

$$h(x,y) = \frac{\exp\left\{\mathrm{i}(2\pi z/\lambda)\sqrt{1+(x/z)^2+(y/z)^2}\right\}}{\mathrm{i}z\lambda\left[1+(x/z)^2+(y/z)^2\right]} \qquad [5]$$

The Rayleigh–Sommerfeld formulation is valid so long as the angles at which energy diffracts from the aperture are small. If the distance z is sufficiently large that the angles stay clustered about the optic axis, one can replace the spherical wave response of freespace with a parabolic wave:

$$h(x, y) \approx \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda} \exp[\mathrm{i}(\pi/\lambda z)(x^2 + y^2)] \qquad [6]$$

This simplification is referred to as the paraxial, or Fresnel, approximation. The diffraction integral is now given by:

$$\begin{aligned} U_z(x, y) \approx\ & \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_{\mathrm{out}}(\zeta, \eta) \\ & \times \exp\{\mathrm{i}(\pi/\lambda z)[(x - \zeta)^2 + (y - \eta)^2]\} \mathrm{d}\zeta\, \mathrm{d}\eta \\ =\ & \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda} \exp[\mathrm{i}(\pi/\lambda z)(x^2 + y^2)] \\ & \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_{\mathrm{out}}(\zeta, \eta) \exp[\mathrm{i}(\pi/\lambda z)(\zeta^2 + \eta^2)] \\ & \times \exp\{-\mathrm{i}2\pi[\zeta(x/\lambda z) + \eta(y/\lambda z)]\} \mathrm{d}\zeta\, \mathrm{d}\eta \end{aligned} \qquad [7]$$

The second representation results from an expansion of the exponent and allows the diffraction integral to be written as a Fourier transform. The quadratic phase terms can be removed if a positive lens with focal length z is placed next to the diffractive element and a second positive lens with focal length z is placed in the image plane (see Figure 1). Under these conditions, $U_{\mathrm{out}}(x, y)$ and $U_z(x, y)$ form a Fourier transform pair:

$$\begin{aligned} U_z(x, y) &= \mathrm{FT}[U_{\mathrm{out}}(x, y)]|_{u=x/\lambda z, v=y/\lambda z} \\ &= \mathrm{FT}\{\exp[\mathrm{i}\theta(x, y)]\}^{**}\mathrm{FT}[U_{\mathrm{in}}(x, y)]|_{u=x/\lambda z, v=y/\lambda z} \end{aligned} \qquad [8]$$

The objective of diffractive design is to find a surface relief $d(x, y)$ that generates a desired response $q(x, y)$ somewhere in the response $U_z(x, y)$. However, eqns [3] and [7] neglect the fabrication process, which affects the fields that can actually be realized. A diffractive optic designer can incorporate fabrication constraints into their design either directly or indirectly. In a direct approach, the designer represents the diffractive structure parametrically, for example, a multilevel phase element can be represented by a discrete set of allowed phase values and a binary phase element can be represented by the contours that separate one phase value from the other. A designer defines the diffractive element by specifying and modifying these parameters.

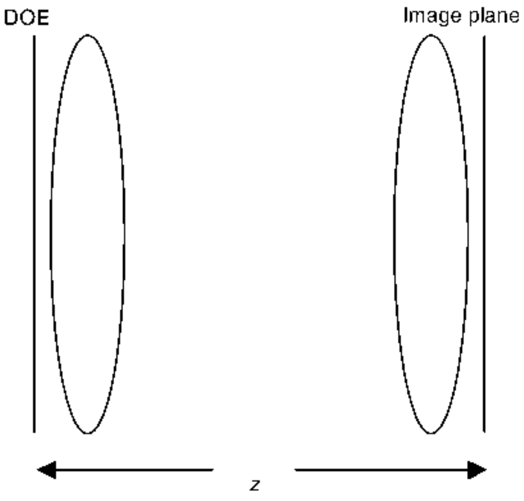

Figure 1 Optical Fourier transform module.

In an indirect approach, the designer first determines an intermediate function without consideration for fabrication and afterward applies the fabrication constraints to this function. For example, if $f(u, v)$ is the desired response one wishes to generate in the z-plane using $U_{\mathrm{in}}(x, y)$ as the input, from eqn [8], the optical function required to do is

$$Q(x, y) = F(x, y)/U_{\mathrm{in}}(x, y) \qquad [9]$$

where $F(x, y)$ is the Fourier transform of $f(u, v)$. The designer must then find a phase $\theta(x, y)$ that satisfies the fabrication constraints and insures that $\exp[\mathrm{i}\theta(x, y)] \approx Q(x, y)$.

For example, fabrication using multiple binary masks yields a phase structure quantized to 2^N levels. An effective design should, therefore, yield a quantized phase structure that minimizes reconstruction error. A naïve approach is to set the magnitude of $Q(x, y)$ in eqn [9] to unity and quantize the phase. However, this introduces considerable error. A direct approach to design searches only over the set of quantized phase structures for the one that, for example, generates the desired response with maximum energy and minimum error. An indirect design seeks first a phase structure whose response produces a minimum amount of error between it and $Q(x, y)$ and then quantizes this structure.

Performance of the element can be improved if one exploits certain freedoms available in design. For example, all systems have a finite-sized detector. It is, therefore, unnecessary to specify the performance of the diffractive element everywhere in the image plane. Its behavior within the detector window is all that is important. If only the intensity of the response is specified, then the phase of the response within the window is also a design freedom. Finally, one can also control the amount of energy contained in the signal window by scaling the amplitude of the response. This allows a designer to balance diffraction efficiency with error in the signal window.

Two general nonlinear optimization algorithms for the design of DOEs are represented in Figure 2. Indicative of the nature of data flow, Figure 2a is

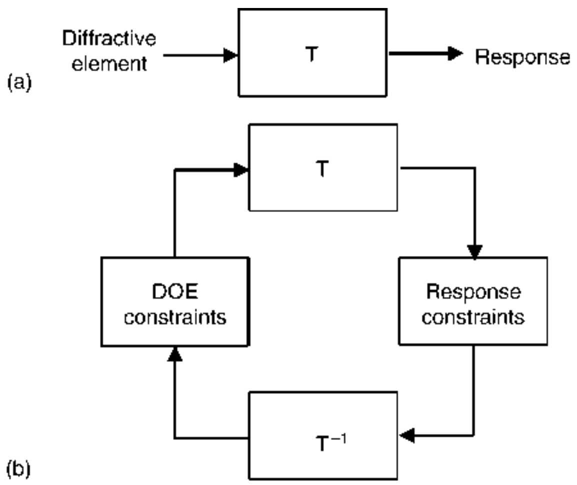

Figure 2 Algorithms used for diffractive design. (a) Unidirectional and (b) bidirectional algorithms.

referred to as a unidirectional algorithm and Figure 2b as bidirectional. In a unidirectional algorithm, changes are made only to the diffractive element. The changes are driven by how well the element performed in a previous iteration. If models exist for both forward and inverse propagation (T and T^{-1}, respectively), a designer can use a bidirectional algorithm, which allows modification of both the diffractive element and its response. This allows bidirectional algorithms to converge to a solution, in general, more rapidly than unidirectional algorithms. In scalar theory, T is a Fourier transform but more rigorous models exist. If the model of the optical system cannot be inverted, a unidirectional algorithm must be used.

Fabrication

Regardless of which design algorithm is used to determine $d(x, y)$, ultimately it needs to be realized physically. Although numerous techniques exist for doing so, by far the most common are based on microlithographic processes. The process requires exposing a layer of photoresist deposited on top of an optical substrate and etching the photoresist to transfer the pattern into the substrate. These methods are especially appealing because they leverage the existing manufacturing infrastructure of the microelectronics industry and, as a consequence, have enabled the efficient and economical mass manufacture of DOEs.

Two basic methods exist for patterning photoresist, direct writing using a modulated beam and indirect patterning using masks. Further, the mask process can be either a single or multistep process. The multistep approach uses N masks and N exposures to fabricate a multilevel diffractive element that has 2^N phase levels (see Figure 2). However, profiles that contain very small features are difficult to fabricate in this manner. As a consequence, alternative processes using single-step grayscale lithography have also been developed.

Although grayscale photolithography is more time-efficient and less prone to mask alignment errors than multistep lithography, the photoresist must respond linearly to illumination. Unfortunately, most photoresists used to fabricate microelectronic circuits are highly nonlinear. However, one can obtain a linear response by adjusting the exposure and development process. In the following, we review fabrication using both multiple binary masks and a single grayscale mask.

Multilevel Fabrication Using Binary Masks

In this fabrication method, one must first generate a sequence of N masks that are related, in a recursive manner, to the quantized phase $\theta(x, y)$:

$$m_k(x, y) = \mathrm{mod}(M_k(x, y), 2), \quad k = [1, N] \qquad [10]$$

where

$$M_{k+1}(x, y) = \frac{M_k(x, y) - m_k(x, y)}{2} \qquad [11]$$

and

$$M_1(x, y) = \theta(x, y)\frac{N}{2\pi} \qquad [12]$$

The masks can be produced using either optical or electron beam (e-beam) pattern generators from integrated circuit fabrication. The pattern generators expose a thin layer of photoresist on the surface of a chrome-covered quartz substrate. Development of the photoresist leaves the desired diffractive pattern in the unexposed areas. The diffractive pattern is transferred to the chrome layer by etching any chrome not covered by photoresist. After the chrome etch process is completed, any remaining photoresist is washed away to yield the finished lithographic mask. This process is repeated for each mask.

The completed mask set is used to fabricate the DOE using techniques similar to those used to generate the mask set. An optical substrate is coated with a thin layer of photoresist, placed in intimate contact with the first lithographic mask $m_1(x, y)$, and illuminated with an ultraviolet (UV) exposure lamp. During photoresist development, exposed resist is washed away, which leaves a binary pattern in the remaining developed photoresist.

When the pattern is transferred into the substrate material via etching, the photoresist acts as an etch stop. For the first mask, the etch depth into the substrate corresponds to a π-phase shift. The process is repeated for each subsequent mask, that is, the substrate is coated with a layer of photoresist, UV illumination exposes the photoresist through the mask, the photoresist is developed, and the substrate etched by the remaining pattern. The etch depth of each mask is half that of the previous mask. Figure 3 illustrates this process.

The substrate can be etched using either wet or dry techniques. In wet etching, one uses reactive chemicals, such as hydrofluoric acid, to etch the substrate. However, wet etching is isotropic, that is, substrate etching occurs equally in all directions. As a result, it is difficult to preserve vertical sidewalls and to achieve high fidelity in etching high-aspect ratio features. As a consequence, dry etching is more commonly used.

A dry etch uses a beam of ions to remove substrate material. The process for removal can be either chemical or kinetic. The process used is determined by the value of the plasma power that creates the source of ions relative to the bias power used to accelerate the ions toward the substrate. If the bias power is large and the plasma power small, a small stream of ions are accelerated rapidly toward the substrate. In this case, the operational mode is similar to a micromilling machine, wherein inert particles chip away at the surface by virtue of their kinetic energy. Alternatively, a small bias power and a large plasma power produce a large population of ions with no preferential directionality. This yields an isotropic process similar to wet etching.

Two methods for performing dry etching are reactive ion etching (RIE) and inductively coupled plasma (ICP) etching. However, in RIE systems the bias and plasma power are not independent due to capacitive coupling. Thus, adjusting one affects the other. ICP, on the other hand provides independent control over the bias and plasma powers. Independent control allows one to control precisely the anisotropic nature of an etch process. As mentioned above, this is achieved by varying the bias power relative to the plasma power, which controls the directionality of the acceleration.

Independent control is important because the ideal process for a given DOE depends on the substrate material and the characteristics of the diffractive profile. Independent control allows one to adjust the respective powers, as well as gas flow rates, and their partial pressures, to determine an optimal etch process. Dry etching allows one to incorporate endpoint detection, which provides precise control

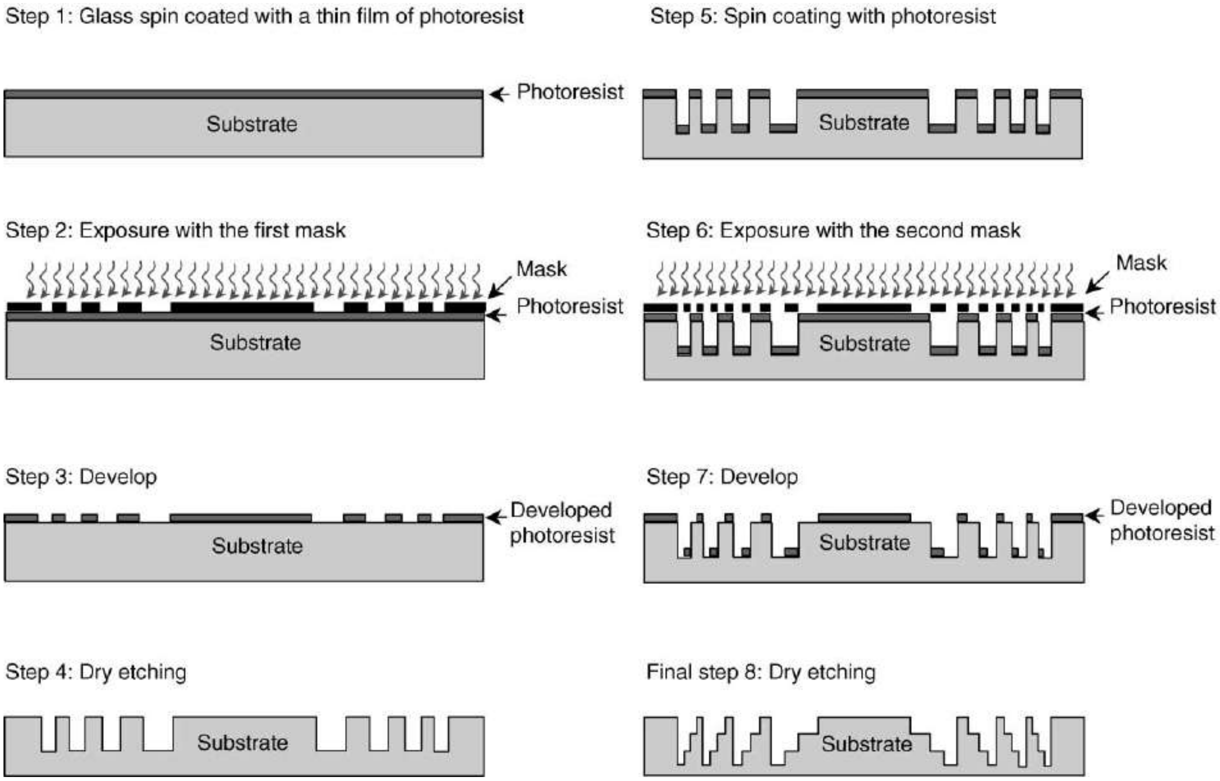
Step 1: Glass spin coated with a thin film of photoresist
Photoresist
Substrate
Step 2: Exposure with the first mask
Mask
Photoresist
Substrate
Step 3: Develop
Developed photoresist
Substrate
Step 4: Dry etching
Substrate
Step 5: Spin coating with photoresist
Substrate
Step 6: Exposure with the second mask
Mask
Photoresist
Substrate
Step 7: Develop
Developed photoresist
Substrate
Final step 8: Dry etching
Substrate

Figure 3 Fabrication of a multilevel quantized phase DOE using binary masks.

over the etch depth. As presented below, achieving the proper etch depth is important to generating responses with high diffraction efficiency.

Care must be used when determining the number of phase levels appropriate for a particular application. The number of phase levels has a direct impact on the portion of input energy one can hope to diffract into the desired response. This is referred to as the diffraction efficiency. According to scalar diffraction theory, the relationship between the maximum diffraction efficiency η and the number of masks N is:

$$\eta = \mathrm{sinc}^2\left(\frac{1}{N}\right) \qquad [13]$$

Thus, using two masks to create a 4-level phase element, as opposed to a single mask to create a binary-phase element, doubles the upper bound on the diffraction efficiency from 41% to 81%. The maximum diffraction efficiency for 8-level and 16-level phase elements is 95 and 99%, respectively.

However, this increase in diffraction efficiency incurs fabrication costs. After the first etch step, each subsequent exposure requires careful mask alignment. Alignment is accomplished using a standard tool from integrated circuit fabrication known as a mask aligner. The quality of a mask aligner is determined by its precision and the majority of commercially available aligners can register a mask and substrate to within less than a micrometer. This degree of alignment accuracy is sufficient to fabricate many useful multilevel diffractive elements.

To summarize, a multilevel surface relief profile uses a set of N masks to generate a surface profile that exhibits 2^N phase levels. The etch depth for the kth mask is $d_k = \lambda/2^k(n-1)$. In the next section, we describe a process for fabricating continuous DOE profiles in a single expose and etch cycle.

Continuous DOE Profiles Using Grayscale Lithography

Grayscale photolithography has become increasingly popular because of its flexibility and convenience in fabricating continuous diffractive profiles. Methods including directly writing partially exposed e-beam resist, resist reflow lithography, and using grayscale masks in a single step exposure. Here we discuss only the latter method because the processes are similar to those used in multiple mask fabrication. Thus, the same facilities can be used to produce both. In particular, we consider the application of high-energy-beam-sensitive (HEBS) glass to grayscale lithography and the fabrication of continuous profile DOEs.

The process for creating a diffractive element using HEBS glass is similar to fabricating a binary-phase element from a single mask. However, mask creation is markedly different. HEBS glass is a white crown glass within which a thin buried layer of silver ions is incorporated. The layer of silver ions is created via ion exchange in an aqueous acidic solution and essentially serves the same function as silver grains in photographic film. When the glass is exposed to a beam of electrons, the silver-containing crystal complexes are chemically reduced and the glass darkens. By controlling the electron dose, one controls the opacity of the glass to produce a grayscale pattern.

Although the attraction of single-step grayscale lithography is reduced costs in fabrication, the process incurs high calibration costs to insure quality pattern transfer through several steps. The parameter that has the greatest impact on the overall process is the desired size of the smallest diffractive feature. The minimum feature size affects the selection of the accelerating potential, beam current, addressing grid, and electron dose of the e-beam used to expose the HEBS glass.

The e-beam accelerating potential sets the resolution limit of the process and affects the sensitivity of the glass to the electron dose, the total charge per area incident upon the glass. For example, for a constant electron dose, increasing the accelerating potential increases the darkening response of the material. The large accelerating potential produces a large flux, which achieves the desired electron dose in a short period using energetic electrons. This, in turn, enhances the sensitivity of the glass. However, increasing the accelerating potential also increases scattering, which limits the minimum feature size achievable. Thus, low beam energies are generally used to write small features. But this, unfortunately, reduces the range of available optical densities because the material is less sensitive at low beam energies. To compensate for the reduced sensitivity, it is possible to increase the electron dose by increasing the exposure time. However, this undermines to some extent using a low energy beam to achieve small features due to broadening that results from long exposure times.

Once the accelerating potential for writing the mask is chosen, the beam current and addressing grid need to be specified. The beam current is related to the spot size and, therefore, affects directly the length of time to write the mask. High current levels produce large spots. Since write times can easily reach into hours for large structures, high current levels are desirable to reduce them.

The addressing grid represents the physical grid that the electron beam is translated over in order to complete the exposure. This grid should be chosen as finely as possible, however, too fine a grid can significantly increase the required dwell time on each grid point. In this respect, a compromise is required. High beam currents produce short dwell times, which may necessitate using a coarse address grid to insure the beam is capable of performing the exposure.

Once the process parameters are set, calibration between the HEBS transmission and the electron dose is necessary. Analyzing the optical density of written test structures by some experimental means and determining its functional dependence on the electron dose completes the calibration. This allows one to specify the electron doses required to achieve the features dictated by the device design.

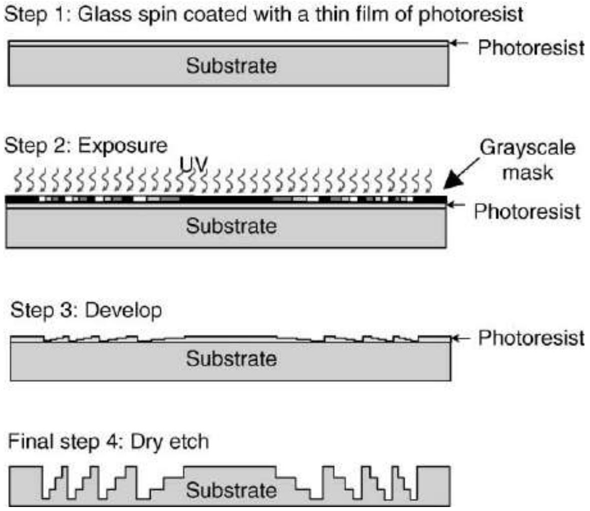

Figure 4 Fabrication of a continuous phase DOE using a grayscale mask.

If resolution requirements cannot be met through judicious selection of the accelerating potential, then some of the contrast in the mask can be sacrificed for an increase in available resolution. As we have mentioned, reducing the maximum optical density, or contrast, can alleviate scattering-induced broadening. It is possible, however, to compensate to some degree for the reduced contrast during the photolithography process, where UV exposure and develop times provide additional latitude in the pattern transfer process.

Figure 4 represents the fabrication of a continuous profile DOE using a grayscale mask. The substrate is spin-coated with a thin film of photoresist, placed in intimate contact with the grayscale mask, and UV illuminated to expose the resist film. The resist is developed and the pattern transferred into the substrate using either wet or dry etching techniques. Because the solubility of the resist is proportional to exposure, developing the photoresist produces a variable height profile.

Grayscale lithography is challenging because modern photoresists are also designed to be highly nonlinear yet linear response is required. Traditional photoresists are also designed to amplify mask contrast and produce the sharpest resist profile, thus they either dissolve away quickly and completely, or remain impervious to developer. The exposure threshold of traditional resists is, therefore, steep and considerable effort is required to

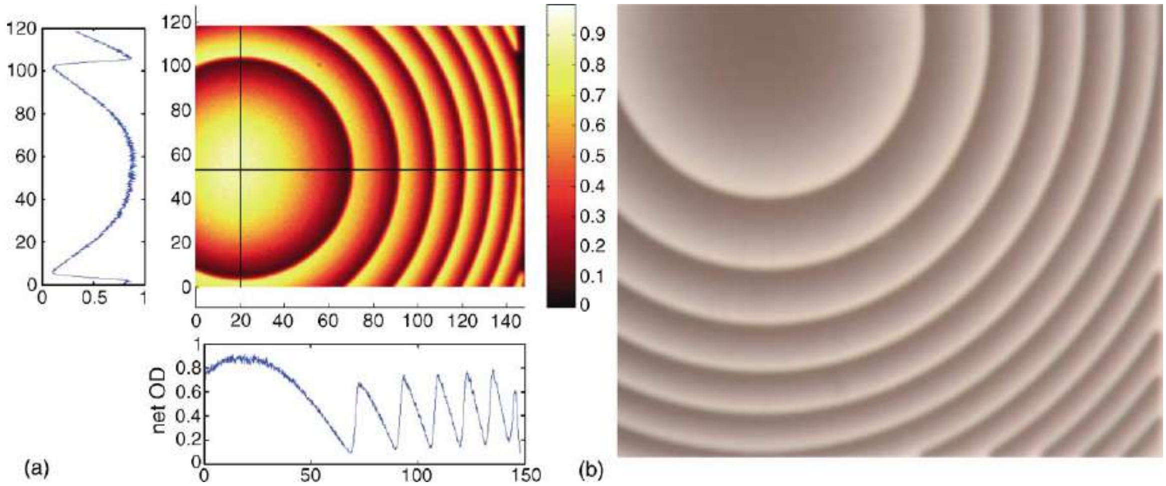

Figure 5 HEBS glass fabrication. (a) Data map of the optical density cross-section of an exposed HEBS grayscale mask and (b) its corresponding visible image.

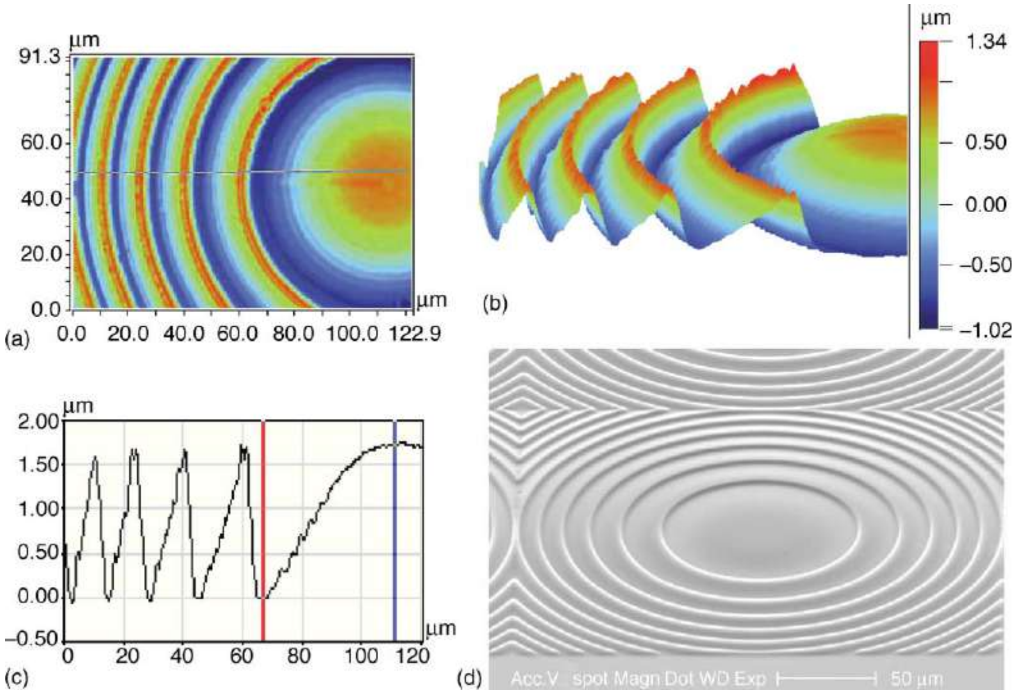

Figure 6 Characterization of a continuous phase diffractive lens. (a) Two-dimensional profile. (b) Three-dimensional profile. (c) Line scan of the two-dimensional plot. (d) Image of element from a scanning electron microscope.

insure that resist exposure is on the linear portion of the curve.

Although the resist exposure process is sensitive to all parameters, the behavior of a resist can be characterized using simulation models. The details of the photoresist models can be found elsewhere, but the simulations predict that the greatest utilization of the available resist occurs for short exposure times (seconds), and that these short exposures require long developing times (minutes). Further, the models predict that low mask contrast does not adversely affect the usable resist height. It is, therefore, not unreasonable to use a low contrast mask to achieve small features.

Through simulation it is possible to determine process parameters that maximize the linear response range of the photoresist. However, these parameters must be validated in the laboratory. The results presented in Figures 5 and 6 provide convincing evidence of the efficacy of these techniques.

Final Remarks

Advances in passive photonic devices have been made possible by leveraging fabrication processes from the electronics industry. Modifying standard photolithographic techniques and photoresist processes has produced an industrial capacity for fabricating diffractive elements. This capability will no doubt lead to future advances in telecommunications, signal processing, and imaging.

See also

Diffraction: Fresnel Diffraction. **Diffractive Systems:** Aberration Correction with Diffractive Elements. **Holography, Techniques:** Overview.

Further Reading

Dascher W, Long P, Stein R, Wu C and Lee SH (1997) Cost-effective mass fabrication of multilevel diffractive optical elements by use of a single optical exposure with a gray-scale mask on high-energy beam-sensitive glass. *Applied Optics* 36: 4675–4680.

Dillon T, Balcha A, Murakowski JA and Prather DW (2004) Continuous-tone grayscale mask fabrication using high-energy-beam-sensitive (HEBS) glass. *Journal of Microlithography, Microfabrication, and Microsystems.*

Fienup JR (1980) Iterative method applied to image reconstruction and to computer-generated holograms. *Optical Engineering* 19: 297–306.

Goodman JW (1968) *Introduction to Fourier Optics.* New York: McGraw-Hill.

Kohler M (1998) Etching in microsystem technology. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications.* Berlin: Wiley-VCH.

LeCompte M, Gao X and Prather DW (2001) Photoresist characterization and linearization procedure for the grayscale fabrication of diffractive optical elements. *Applied Optics* 40: 5921–5927.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America A* 52: 1123–1130.

Lohmann AW and Paris DP (1967) Binary Fraunhofer holograms generated by computer. *Applied Optical* 6: 1739–1748.

Mait JN (1995) Understanding diffractive optical design in the scalar domain. *Journal of the Optical Society of America A* 12: 2145–2158.

Stern MB and Rubico-Jay T (1994) Dry etching for coherent refractive microlens arrays. *Optical Engineering* 33: 3547–3551.

Swanson GJ (1989) Binary optics technology: the theory and design of multi-level diffractive optical elements, Technical Report 854, MIT.

Swanson GJ (1991) Binary optics technology: theoretical limits on the diffraction efficiency of multi-level diffractive optical elements, Technical Report 914, MIT.

Swanson GJ and Veldkamp WB, (1990) High-efficiency, multilevel diffractive optical elements. United States Patent 4,895,790.

Wang MR and Su H (1998) Laser direct-write gray-level mask and one-step etching for diffractive microlens fabrication. *Applied Optics* 37: 7568–7576.

Wu CK (1994) High Energy Beam Sensitive Glasses. U.S. Patent 5,285,517.

Wyrowski F (1991) Digital holography as part of diffractive optics. *Report on Progress in Physics* : 1481–1571.

Diffractive Laser Resonators

U D Zeitner, Fraunhofer Institute of Applied Optics and Precision Engineering, Jena, Germany

F Wyrowski, Friedrich-Schiller University, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

All lasers consist of three basic functional components: an active medium, an optical arrangement to pump the active medium, and a resonator, which we consider in detail here. The basic function of the resonator is to back-couple a portion of the emitted laser light into the active medium for further amplification. Simultaneously, in steady-state laser operation, the geometry of the resonator defines the spatial field distribution of the generated light field.

In the most simple, and most common, case the laser resonator consists of two planar or spherical mirrors. This configuration was proposed in 1958 and has been treated extensively in the literature. However, these configurations are only special cases of a more general resonator concept in which arbitrary mirror surface profiles are used. Historically, resonators consisting of surface profiled mirrors played only a small role in laser physics. This was due, not only to difficulties in their numerical design and analysis, but also to the lack of suitable fabrication methods for the mirrors.

However, since the 1990s, these difficulties have been overcome and in 1992 it was demonstrated for the first time experimentally a stable CO_2 laser resonator that has a user-defined mode shape generated by surface structured mirrors.

In the following we derive the basic design rules for stable and unstable diffractive laser resonators and discuss the advantages and disadvantages of the technology.

Modes in Laser Resonators

Before presenting the details of diffractive laser resonators we look first at some general properties of the electromagnetic field inside a laser resonator. As is typical in laser resonator theory, we use scalar and paraxial descriptions of the fields.

The complex amplitude $U(x, y, z)$ represents one scalar, monochromatic component of the electromagnetic field. To handle wave propagation inside the resonator, we use a general operator formalism in which the influence of each optical component on the wave $U(x, y, z)$ is represented by an operator. One of the most important operators is the so-called round trip operator $\mathbf{Z}$, which describes the effect of a round trip inside the resonator on the wave $U(x, y, z)$.

If $U(x, y, z)$ denotes the wave at the beginning of a resonator round trip, the resulting wave $\tilde{U}(x, y, z)$ at its end is

$$\tilde{U}(x, y, z) = \mathbf{Z}U(x, y, z) \qquad [1]$$

With reference to Figure 1, the round trip operator can be decomposed into a series of operators that represent the different elements in the resonator. For the common two-mirror resonator a round trip from mirror 1 to mirror 2, reflection at mirror 2, back propagation to mirror 1, and final reflection at

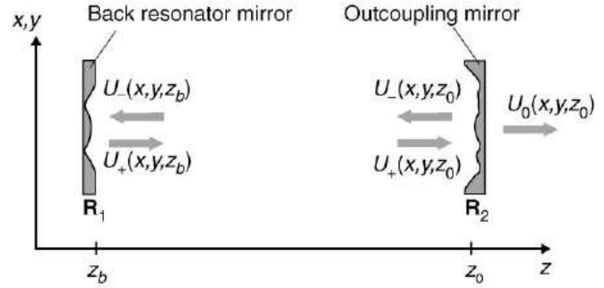

Figure 1 General diffractive laser resonator arrangement.

mirror 1, is represented by

$$\mathbf{Z} = \mathbf{R}_1 \mathbf{P}_{\Delta z} \mathbf{R}_2 \mathbf{P}_{\Delta z} \qquad [2]$$

The operators $\mathbf{R}_1$ and $\mathbf{R}_2$ represent the reflection at mirrors 1 and 2, respectively, and the propagation along the resonator length Δz is represented by the propagation operator $\mathbf{P}_{\Delta z.}$

In steady-state laser operation, the wave circulating inside the resonator reproduces itself, up to a constant factor γ, after a complete round trip:

$$\tilde{U}(x, y, z) = \gamma U(x, y, z) \qquad [3]$$

Together eqns [1] and [3] give the eigenvalue equation:

$$\gamma U(x, y, z) = \mathbf{Z} U(x, y, z) \qquad [4]$$

which, in general, has a set of solutions $U_p(x, y, z)$.

Thus, in steady-state, the waves $U_p(x, y, z)$ circulating inside the resonator are eigenfunctions of the round trip operator $\mathbf{Z}$ with the connected eigenvalue γ_p. The eigenfunctions of the resonator are called modes and the real-valued eigenvalues γ_p are connected with the round trip losses V, according to the equation:

$$V = 1 - |\gamma_p|^2 \qquad [5]$$

As we can see, the spatial properties of the modes are determined by the optical elements in the resonator. In the simple two-mirror resonator, the mirror profiles, which are related to the operators $\mathbf{R}_1$ and $\mathbf{R}_2$, can be used to specify the shape of the modes.

To derive the rules for resonator design we must look closer at the orthogonality of the resonator modes. The normalized eigenfunctions of the round trip operator $\mathbf{Z}$ do not obey the usual power-orthogonality relationship:

$$\iint U_m(x, y, z) U_n^*(x, y, z)\, \mathrm{d}x\, \mathrm{d}y = \delta_{mn} \qquad [6]$$

but instead satisfy the biorthogonality relationship:

$$\iint U_{m;+}(x, y, z) U_{n;-}(x, y, z)\, \mathrm{d}x\, \mathrm{d}y = \delta_{mn} \qquad [7]$$

The δ_{mn} is the so-called Kronecker symbol and $*$ denotes the complex conjugate. In eqn [7], $U_{m;+}(x, y, z)$ is the mode in plane z traveling in the $+z$ direction and $U_{m;-}(x, y, z)$ is the mode in the same plane z propagating in the opposite direction. The physical reason for the difference between eqns [6] and [7] are diffractive mode losses at limiting apertures. However, if a mode $U_{m;+}(x, y, z)$ travels without losses through the whole resonator, a power normalization:

$$\iint U_m(x, y, z) U_m^*(x, y, z)\, \mathrm{d}x\, \mathrm{d}y = 1 \qquad [8]$$

is valid in all planes z. Equations [7] and [8] will be used in the next section to derive the design strategy for resonators with predefined fundamental modes.

If the mirrors, either planar or spherical, are assumed to be infinite in extent, it is possible to derive analytic solutions for the modes. This yields either the Gauss–Hermite or Gauss–Laguerre modes, which are orthogonal in the sense of eqn [6]. For resonators whose mirror shape differs from either a plane or sphere, the analytical calculation of its eigenfunctions or resonator modes is difficult and in most cases even impossible.

However, in such cases the analysis can be performed numerically. One of the most common methods is the Fox-Li-algorithm which uses the physical propagation of the modes through the resonator. The algorithm begins with a random field distribution in some plane of the resonator. This field is then numerically propagated through the resonator until a round trip is completed. The effect of the resonator optical elements must be accounted for in this procedure. At the end of the resonator round trip, the starting field will have changed due to diffraction and serve as a new starting wave for a succeeding round trip calculation. This procedure is repeated until a static field distribution is obtained, which is then an eigenfunction of the resonator.

Design Principles

The condition in eqn [7] on the biorthogonality of resonator modes, together with the eqn [8] valid for lossless modes, can be used to develop principles to design resonators with an arbitrary fundamental mode shape. Consider a stable laser resonator characterized by negligible diffraction loss of the fundamental mode. Following the discussion above, its lossless modes must simultaneously satisfy eqns [7] and [8]. A comparison of both equations yields:

$$U_+(x, y, z) = U_-^*(x, y, z) \qquad [9]$$

For reasons of simplicity we neglect here, and in the following, the subscript m for the mode order if only one mode is considered. Because this equation must hold in all planes z, it is satisfied in the mirror resonator planes z_0 and z_b. Further, by definition of the operators $\mathbf{R}_1$ and $\mathbf{R}_2$:

$$U_+(x, y, z_b) = \mathbf{R}_1 U_-(x, y, z_b) \qquad [10]$$

$$U_-(x, y, z_0) = \mathbf{R}_2 U_+(x, y, z_0) \qquad [11]$$

The combination of eqns [10] and [11] with eqn [9] yields the following expressions for the mirror operators:

$$\mathbf{R}_1 = \frac{U_+(x, y, z_b)}{U_+^*(x, y, z_b)} \qquad [12]$$

$$\mathbf{R}_2 = \frac{U_-(x, y, z_0)}{U_-^*(x, y, z_0)} \qquad [13]$$

Equations [12] and [13] imply that lossless modes are phase-conjugated on reflection at the resonator mirrors. Because we have not made any assumptions on the shape of the mode, these equations represent a general property of lossless resonator modes. Furthermore, they provide a general design rule for the construction of stable laser resonators with predefined fundamental mode shape.

If one considers limiting apertures inside the resonator, an equivalent statement for lossy modes can be made by making assumptions about phase conjugation in a lossy system and showing that they lead to a contradiction. Without loss of generality, we assume an aperture on the outcoupling resonator mirror that clips the mode. We now consider the wave $U_-(x, y, z_0)$ after reflection at this mirror and assume this wave is still phase-conjugated after propagation through the resonator during its reflection at the back resonator mirror according to:

$$U_+(x, y, z_b) = U_-^*(x, y, z_b) \qquad [14]$$

If now the wave $U_+(x, y, z_b)$ is back propagated to the outcoupling mirror, we end up with the wave $U_+(x, y, z_0)$ which must, because of the phase conjugation in z_b, have the same amplitude as the wave $U_-(x, y, z_0)$ we started with. As a result, because both waves are connected to each other via the reflection operator of the outcoupling mirror by eqn [11], $\mathbf{R}_2$ cannot influence the amplitude distribution which, however, is a contradiction to the assumption of a limiting aperture. Thus, we finally can state that lossy modes do not have the property to be phase conjugated during their reflection at both mirrors. As a consequence, the design of diffractive mirrors for unstable laser resonators, in general characterized by energy outcoupling via diffractive losses, is more complex than simply calculating a phase-conjugating mirror structure for stable laser resonators.

Up to now, design methods described in the literature for unstable laser resonators with user-defined output beam shapes have all been based on an iterative calculation of the mirror reflection function. Because a detailed description of these methods is rather complex, we restrict ourselves to a sketch of the principal ideas behind them.

Without loss of generality, we consider an unstable laser resonator from which the light is coupled out by a mirror with spatially varying transmission and reflection. The second mirror is fully reflective within its aperture. The particular application defines only few design constraints which are in most cases the light wavelength, the dimensions of the resonator, and the shape of the outcoupled beam $U_0(x, y, z_0)$. As a consequence of the outcoupling principle, the shape of the mode inside the resonator is not finally defined by the outcoupled beam. In regions where the outcoupling mirror is reflective and not transmissive, the mode amplitude can be chosen freely. Also the amplitude of the mode in the plane of the back resonator mirror is not fixed but influenced by the amplitude and phase of the mode in the plane of the outcoupling mirror.

The general starting situation for an unstable resonator design is sketched in Figure 2. The design procedure must yield the complex amplitudes of four waves: the mode directly in front of the outcoupling mirror $U_-(x, y, z_0)$, the mode after reflection at the outcoupling mirror $U_+(x, y, z_0)$, and the complex amplitudes of the forward and reverse modes in the plane of the back resonator mirror $U_+(x, y, z_b)$ and $U_-(x, y, z_b)$, respectively. Since we assume no losses at

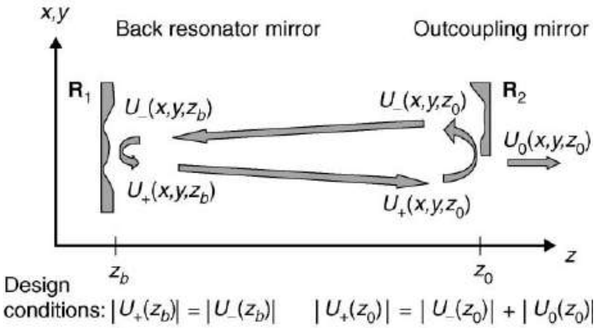

Figure 2 Starting situation for design of an unstable diffractive laser resonator.

this mirror, the amplitudes must satisfy the condition:

$$|U_+(x,y,z_b)| = |U_-(x,y,z_b)| \tag{15}$$

If we further assume no absorption at the outcoupling mirror, the amplitudes of the three waves in z_0 are related according to:

$$|U_-(x,y,z_0)| + |U_0(x,y,z_0)| = |U_+(x,y,z_0)| \tag{16}$$

To design the resonator one starts with the choice of initial intensity distributions for the wavefields $U_-(x,y,z_0)$ and $U_+(x,y,z_b)$. In order to obtain a good filling of the available mode volume, a homogeneous intensity across the mirror apertures is recommendable. To avoid strong diffraction effects at the mirror edges, the corresponding intensity distributions should be smoothed out. Together with the desired outcoupled beam amplitude $|U_0(x,y,z_0)|$ one can calculate the two remaining amplitudes for the wave fields $U_-(x,y,z_b)$ and $U_+(x,y,z_0)$ according to eqns [15] and [16].

An iterative algorithm, such as the well known Gerchberg–Saxton algorithm, can be used to find the phases of the beams $U_-(x,y,z_0)$ and $U_+(x,y,z_0)$ that insure the wavefields have almost equal amplitudes once they are propagated or back-propagated, respectively, to the plane z_b. In order to find wavefields that satisfy eqns [15] and [16], two successive iteration runs with different constraints on the complex amplitudes have to be performed. During the first iteration run no change in the initial amplitudes in plane z_0 are allowed. The iteration only look for phases for $U_-(x,y,z_0)$ and $U_+(x,y,z_0)$ which result in the amplitude $|U_+(x,y,z_b)|$ once they are propagated and back-propagated, respectively, to plane z_b. Usually the results of a Gerchberg–Saxton algorithm, with such constraints, are wavefields $U_+(x,y,z_b)$ and $U_-(x,y,z_b)$ with similar but not equal amplitude. Thus, in order to determine wavefields that satisfy eqn [15], a second iteration run with the Gerchberg–Saxton algorithm has to be added in which slight variations in wavefield amplitudes also in plane z_0 are allowed. Because the application of the constraints to ensure convergence of the iteration is rather complex, we do not go into details here but refer the reader to the Further Reading at the end of this article.

Upon completion of the iterations, the mirror reflectance operators $\mathbf{R}_1$ and $\mathbf{R}_2$ are calculated according to:

$$\mathbf{R}_1 = \frac{U_+(x,y,z_b)}{U_-(x,y,z_b)} \tag{17}$$

$$\mathbf{R}_2 = \frac{U_-(x,y,z_0)}{U_+(x,y,z_0)} \tag{18}$$

The design method described so far yields a mathematical expression for the resonator mirror reflection operator. Practical realization of the resonator requires this expression be converted into a description of a physical property of the mirror. In most practical cases this is the surface profile. Since the design method considers only phase-only functions, the physical description of the reflection operators are:

$$\begin{aligned} \mathbf{R}_1 &= \exp[-\mathrm{i}\phi_1(x,y)] \text{ and} \\ \mathbf{R}_2 &= \exp[-i\phi_2(x,y)] \end{aligned} \tag{19}$$

Using the so-called thin-element-approximation, the surface profiles $h_1(x,y)$ and $h_2(x,y)$ of the two mirrors follow directly from the phases $\phi_1(x,y)$ and $\phi_2(x,y)$:

$$\begin{aligned} h_1(x,y) &= -\frac{\lambda}{4\pi}\phi_1(x,y) \text{ and} \\ h_2(x,y) &= -\frac{\lambda}{4\pi}\phi_2(x,y) \end{aligned} \tag{20}$$

with λ being the laser wavelength. The expressions in eqn [20] contain a factor of 1/2 due to reflection. These surface profiles are commonly fabricated by microlithographic techniques. For stable laser resonators, where the mirrors perform a static phase conjugation, the mirror profile is a surface of equal phase of the fundamental mode.

Use of General Resonator Modes

Applications of diffractive laser resonators can be classified into beam shaping for a particular task external to the laser and resonator optimization via modifying the eigen-space of the round trip operator.

Internal beam shaping for external tasks is relevant to applications where it is advantageous to use a fixed beam profile other than a Gaussian, as is the case for instance in materials processing. Shaping the beam internally, avoids the energy losses usually associated with external beam shaping techniques. However, lasers that have a nonGaussian outcoupled beam profile are often less flexible in their general applicability.

However, there is considerable potential for diffractive laser resonators to enhance resonator characteristics and thereby improve the lasing operation itself. The conventional Gaussian-shaped fundamental mode in customary laser resonators varies considerably in intensity from its central maximum to its outer wings. The transfer of energy from the pumped active medium into the propagating mode is dependent upon the local transverse intensity and is optimal for homogeneous illumination. As a result, the efficiency with which energy is transferred into the fundamental mode can be increased by using diffractive resonator mirrors to match the shape of the lateral fundamental mode to that of the cross-section of the active medium.

An important resonator property, especially in high power laser systems, is the difference of round trip losses V between the fundamental mode and the higher, usually undesired, laser modes. This difference is referred to as modal discrimination. With no limiting apertures, the wavefronts of all Gauss–Hermite or Gauss–Laguerre modes of common laser resonators are phase conjugated upon reflection at the spherical mirrors. As we have shown above, there is a tight connection between the phase conjugation property of resonator modes and their property of having no round trip losses. Thus, the phase conjugation property of Gauss–Hermite and Gauss–Laguerre modes directly implies small round trip losses. In fact, in resonators without apertures they all have an eigenvalue $\gamma_p = 1$ and according to eqn [5] the round trip losses are zero. High-order modes, whose diameters are large, can be suppressed by decreasing the size of the mirror aperture, however, because the difference in mode diameter between modes of neighboring order m is small, the achievable discrimination is limited to a maximum ratio $V_{m+1}/V_m \approx 10$. The structured surface profile of the mirrors in a diffractive laser resonator whose fundamental mode is nonGaussian is tightly adapted to the phase of the design mode. The wavefront of the higher-order modes in these resonators often do not fit the mirror profile and it can be observed that their round trip losses are significantly higher. Thus, diffractive laser resonators usually show an improved modal discrimination compared to resonators with spherical mirrors. This effect can be further improved by inserting into the resonator additional phase plates with diffuser-like properties.

Shaping the surface profile of resonator mirrors or insertion of additional phase plates significantly improves the flexibility for the resonator design. As is demonstrated by the examples below, there are resonator configurations in which high modal discrimination or adaptation of the mode volume to the cross-section of the active medium are combined with a Gaussian shape of the outcoupled beam.

In the design procedure described so far, the resonators were treated as passive optical systems. The propagation of the mode through the resonator has not been influenced by the active medium. However, in real laser systems, this simplification is often not applicable. Due to the pumping process and the connected heat insertion, the refractive index distribution inside the active medium is not homogeneous, which affects the propagation of the modes. Furthermore, changes or fluctuations in the pump energy and cooling mechanisms introduce temporal changes in the optical function associated with the refractive index of the active medium. Further, although static or averaged, optical power of the active medium are included in the round trip operator and calculation of the mirror profiles, experiments with lasers that have diffractive resonator mirrors indicate that the temporal fluctuations can strongly influence the shape and round trip losses of the fundamental resonator mode. In general, diffractive laser resonators are more sensitive to deviations from design than conventional resonators.

Examples

We consider as an example the adaptation of fundamental mode shape to the cross-section of the active medium. In solid state lasers, such as Nd:YAG lasers, the active medium is often a crystal rod with a circular cross-section. Instead of using this sharp-edged cross-section shape as intensity distribution for the desired fundamental mode, a smoothed version, which is approximated by a super-Gaussian intensity distribution in the center plane of the rod, is chosen to be the fundamental mode. The super-Gaussian intensity distribution has a more desirable propagation characteristic than a sharp circular one. However, even these modes do not exhibit the advantageous property of propagation-invariant beam shape characteristic of Gaussian beams.

In our first example, we consider the design of a stable diffractive laser resonator with a super-Gaussian fundamental mode shape inside the active medium and a Gaussian beam coupled out from the resonator. The resonator configuration is based upon a beam shaping arrangement for a Gauss to super-Gauss beam transformation. Its optical setup is sketched in Figure 3. In the plane of the outcoupling

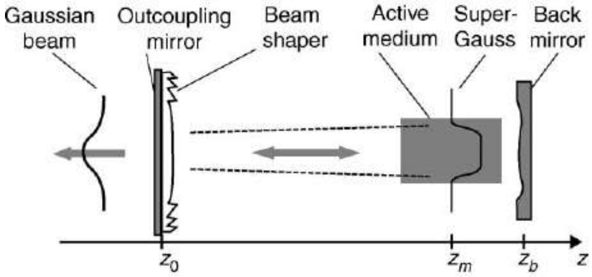

Figure 3 Setup for the Gauss-to-super-Gauss resonator.

mirror z_0, the fundamental mode has a Gaussian amplitude $U_g(x, y, z_0)$. A phase element, which is located directly in front of the outcoupling mirror, shapes the mode amplitude into a super-Gaussian one $U_{sg}(x, y, z_m)$ in the center plane z_m of the active medium. For this special beam shaping problem it is possible to calculate a phase plate surface profile that transforms the beams without energy losses. The corresponding calculation procedure is described in the Further Reading section at the end of this article.

The mirror profiles have to be realized to satisfy the previously derived phase conjugation property for the described beam. Since the Gaussian beam has its minimum waist at the outcoupling mirror, its phase is planar in that plane. Thus, to insure phase conjugation a plane mirror is all that is required in front of the beam shaping phase plate. In the plane of the back resonator mirror the beam has a modulated wavefront due to its transformation from a Gaussian to a super-Gaussian amplitude distribution. Thus, a surface modulated mirror profile is required to phase conjugate this beam during reflection. Figures 4a and b represent the phase profile of the beam shaping element and the back resonator mirror, respectively, for the following set of parameters: laser wavelength $\lambda = 1064$ nm, radius of the Gaussian and super-Gaussian beams in the plane z_0 and z_m, respectively, $w_g = 1$ mm and $w_{sg} = 1$ mm, order of the super-Gauss $n = 10$, distance $z_m - z_0 = 200$ mm, and resonator length $z_b - z_0 = 250$ mm. The amplitude distribution for the fundamental mode in this resonator simulated with the Fox–Li-algorithm is shown in Figure 4c for the plane of the outcoupling mirror and in Figure 4d for the center of the active medium.

In our second example, the flexibility of diffractive laser resonators is demonstrated by a resonator with a user-defined far-field intensity distribution. In our case the far-field intensity $|U_f(x, y, z_f)|^2$ is shaped like the Greek letter lambda, λ. Because we are concerned only with the intensity of the wavefield, the phase $\phi_f(x, y, z_f)$ of the far-field $U_f(x, y, z_f)$ can be chosen freely. As shown in the following, this phase freedom can be used to improve the interaction of the mode with the active medium, as was the case in our first design example. Again, the mode shape in the center of the active medium has a super-Gaussian intensity distribution, which insures high extraction efficiency. We used an iterative Fourier-transform algorithm to calculate a far-field phase that yields a nearly super-Gaussian intensity distribution in the outcoupling plane of the resonator. The resulting wave:

$$U_0(x, y, z_0) = a_0(x, y, z_0) \exp[\mathrm{i}\varphi_0(x, y, z_0)] \qquad [21]$$

is generated directly as the laser output. The amplitude $a_0(x, y, z_0)$ is generated as a mode inside the cavity using a surface-modulated back resonator mirror. This performs a static phase conjugation of the desired mode in the mirror plane. The phase $\varphi_0(x, y, z_0)$ is provided by an additional element on the outside of the outcoupling mirror. A sketch of the resonator configuration is shown in Figure 5.

For the experimental realization of this example the thermal index gradient of the active medium was included into the resonator design and the two diffractive elements were fabricated on a quartz substrate using electron beam lithography and ion-beam etching. The measured mode shape in the plane of the outcoupling mirror and the far-field

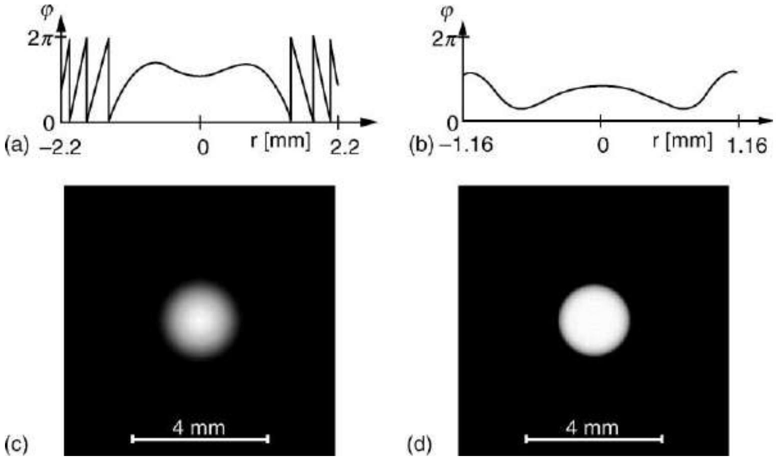

Figure 4 Gauss-to-super-Gauss resonator: (a) phase of beam shaping element; (b) phase of back resonator mirror, simulated amplitude distributions for the fundamental resonator mode; (c) in the plane of the outcoupling mirror; and (d) in the center of the active medium.

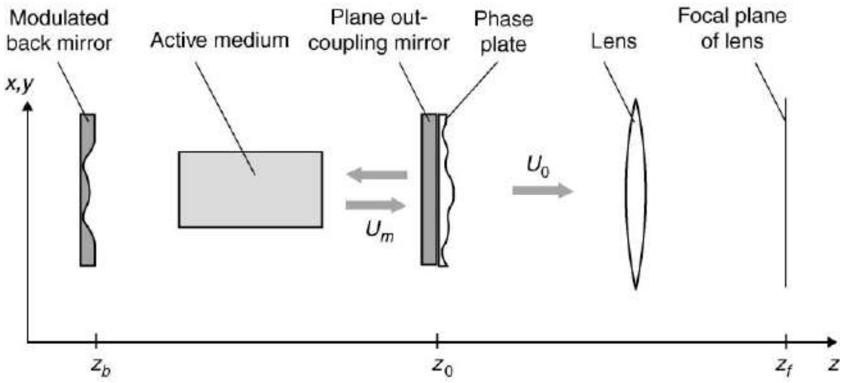

Figure 5 Configuration for resonator internal generation of a user defined far-field intensity distribution. (With permission from Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109. © 2001 IEEE.)

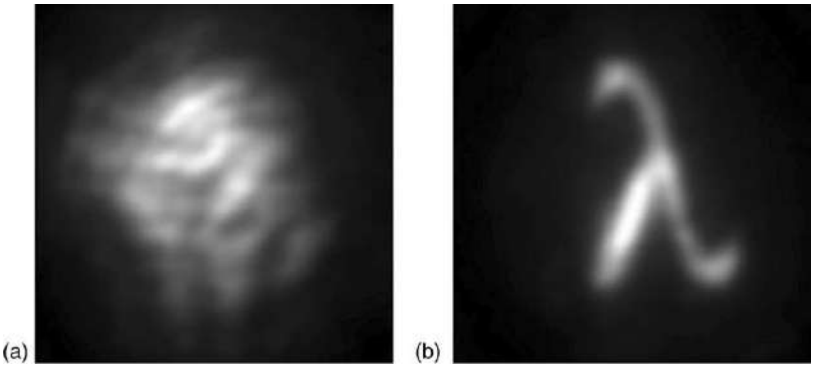

Figure 6 Measured mode intensities for the resonator of Figure 5. (a) In the plane of the outcoupling mirror; and (b) in the far-field of the resonator. (With permission from Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109. © 2000 IEEE.)

intensity distribution are shown in Figures 6a and b, respectively. The mode in the outcoupling plane of the resonator fills an area of about 2 mm diameter with nearly homogeneous intensity. The far-field shows the desired λ-shaped intensity distribution. The variations in intensity are due to the nonlinear response of the active medium. As mentioned above, the response can be improved if these effects are included in the design of the resonator mirrors.

Conclusion

Diffractive laser resonators are a generalization of common spherical mirror resonators. The use of arbitrary resonator mirror profiles and additional internal phase plates offer a huge flexibility for the generation of user defined mode shapes and optimization of resonator properties like extraction efficiency and modal discrimination. The drawback of the use of structured mirrors is an increased adjustment effort, because the diffractive mirrors have more degrees of freedom to be aligned. Furthermore, diffractive resonators tend to be more sensitive against changes of the optical properties of the active medium than spherical mirror resonators. Such changes may occur due to temporal fluctuations in pump power or cooling power.

Most promising applications for diffractive resonators are:

- lasers for materials processing which require a fixed intensity pattern for the sample treatment and a fixed laser power;
- high efficient homogeneous laser illumination, for example, for data projection or sensing applications; and
- suppression of higher order modes and maintaining single-mode operation for high brightness lasers.

Some example applications and solutions have been presented here which cannot be achieved by conventional resonator optics. For stable laser resonators, the phase-conjugation property of the modes is used for a straightforward calculation of the required diffractive mirror profiles. Unstable resonator geometries with diffractive outcoupling require more complex iterative design procedures. However, especially for high-power and high-gain lasers, the latter sometimes have more advantageous properties, such as better discrimination of undesired higher-order modes and more simple mirror cooling options.

See also

Lasers: Carbon Dioxide Lasers. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Aagedal H, Wyrowski F and Schmid M (1997) Paraxial beam splitting and shaping. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, pp. 165–188. Berlin: Akademie Verlag.

Belanger PA, Lachance RL and Pare C (1992) Super-Gaussian output from a CO_2 laser by using a graded-phase mirror resonator. *Optics Letters* 17: 739–741.

Fox A and Li T (1961) Resonant modes in a maser interferometer. *Bell Systems Technical Journal* 40: 453–488.

Gerchberg R and Saxton W (1972) A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35: 237–246.

Herzig H (1997) *Micro Optics*. London: Taylor & Francis.

Leger JR (1997) Diffractive laser resonators. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, pp. 165–188. Berlin: Akademie Verlag.

Leger JR, Chen D and Dai K (1994) High modal discrimination in a Nd:YAG laser resonator with internal phase gratings. *Optics Letters* 19: 1976–1978.

Makki S and Leger JR (2001) Mode shaping of a graded-reflectivity-mirror unstable resonator with an intracavity phase element. *IEEE Journal of Quantum Electronics* 37: 80–86.

Siegman AE (1986) *Lasers*. Mill Valley, CA: Univ. Science Books.

Zeitner UD and Wyrowski F (2001) Design of unstable laser resonators with user-defined mode shape. *IEEE Journal of Quantum Electronics* 37: 1594–1599.

Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109.

Diffractives in Animals

A R Parker, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Structural coloration involves the selective reflectance of incident light by the physical nature of a structure. Although the color effects often appear considerably brighter than those of pigments, structural colors often result from completely transparent materials. In addition, animal structures can be designed to provide the opposite effect to a light display: they can be antireflectors, causing 'all' of the incident light to penetrate a surface (like glass windows, smooth surfaces cause some degree of reflection).

The study of natural structural colors began in the seventeenth century, when Hooke and Newton correctly explained the structural colors of silverfish (Insecta) and peacock feathers (Figure 1), respectively. Nevertheless, accurate, detailed studies of the mechanisms of structural colors did not begin until the introduction of the electron microscope in 1942.

Structural colors generally may be formed by one of three mechanisms: thin-film reflectors, diffraction gratings, or structures causing scattering of light waves. This is a simplistic characterization resulting from the application of geometric optics to the reflective elements. Some structures, however, rather fall between the above categories, such as photonic crystals (Figure 2) – these become apparent when quantum optics is considered. In some cases, this has led to confusion over the reflector type. For example, the reflectors in some scarab beetles have been categorized by different authors as multilayer reflectors, three-dimensional diffraction gratings, photonic crystals, and liquid crystal displays. Perhaps all are correct! Since the above categories are academic, I place individual cases of structural colors in their most appropriate, not unequivocal, category. Also, in this article I will employ geometric optical treatments only.

The array of structural colors found in animals today results from millions of years of evolution. Structures that produce metallic colors have also been identified in extinct animals. Confirmation of this

Figure 1 Newton's subject.

fact, from ultrastructural examination of exceptionally well-preserved fossils, such as those from the Burgess Shale (Middle Cambrian, British Columbia), 515 million years old, permits the study of the role of light in ecosystems throughout geological time, and consequently its role in evolution. In some cases, such as for ostracod crustaceans, diffraction gratings have been found to increase in their optical sophistication throughout 350 million years of evolution.

Multilayer Reflectors

Light may be strongly reflected by constructive interference between reflections from the different interfaces of a stack of thin films (of actual thickness *d*) of alternately high and low refractive index (*n*). For this to occur, reflections from successive interfaces must emerge with the same phase and this is achieved when the so-called 'Bragg Condition' is fulfilled. The optical path difference between the light reflected from successive interfaces is an integral number of wavelengths and is expressed in the equation:

$$2nd \cos \Theta = (m + 1/2)\lambda \qquad [1]$$

from which it can be seen that the effect varies with angle of incidence (Θ, measured to the surface normal), wavelength (λ), and the optical thickness of the layers (nd). There is a phase change of half a wavelength in waves reflected from every low to high refractive index interface only. The optimal narrowband reflection condition is therefore achieved where the optical thickness (nd) of every layer in the stack is a quarter of a wavelength. In a multilayer consisting of a large number of layers with a small variation of index the process is more selective than one with a smaller number of layers with a large difference of index. The former therefore gives rise to more saturated colors corresponding to a narrow spectral bandwidth and these colors therefore vary more with a change of angle of incidence. Both conditions can be found in animals – different colored effects are appropriate for different functions under different conditions. For an oblique angle of incidence, the wavelength of light that interferes constructively will be shorter than that for light at normal incidence. Therefore, as the angle of the incident light changes, the observed color also changes (Figure 3).

Single layer reflectors are found in nature, where light is reflected, and interferes, from the upper and lower boundaries (Figure 4). A difference in the thickness of the layer provides a change in the color observed from unidirectional polychromatic light. The wings of some houseflies act as a single thin film and appear different colors as a result of this phenomenon. A single quarter-wavelength film of guanine in cytoplasm, for example, reflects about 8% of the incident light. However, in a multilayer reflector with ten or more high index layers, reflection efficiencies can reach 100%. Thus, animals possessing such reflectors may appear highly metallic.

The reflectance of the multilayer system increases very rapidly with increasing number of layers. If the dimensions of the system deviate from the quarter-wave condition (i.e., *nd* is not equal for all layers), then the reflector is known as 'nonideal' in a theoretical sense (may be 'ideal' for some natural situations). 'Nonideal' reflectors have a reduced proportional reflectance (not always a significant reduction) for a given number of layers and this

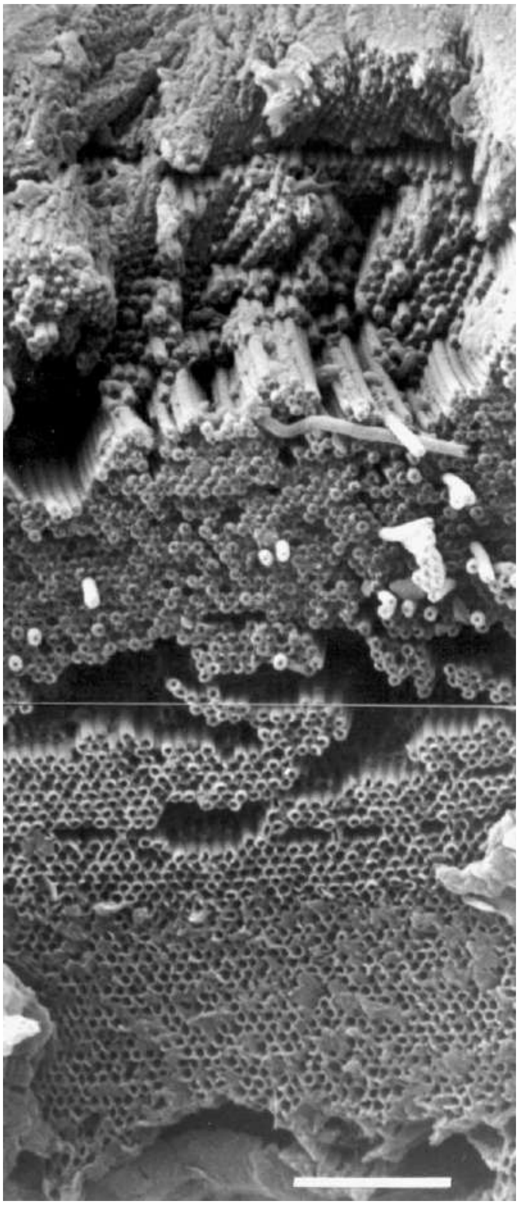

Figure 2 Scanning electron micrograph of a cross-section of the wall of a cylindrical spine of the sea mouse *Aphrodita* sp. (Polychaeta). The wall is composed of small cylinders with varying internal diameters (increasing with depth in the stack), arranged in a hexagonal array, that form a photonic crystal fiber. The opal-type photonic crystal – hexagonally packed nano-spheres (250 nm diameter) – are also found in animals (e.g. weevils). Scale bar represents 8 μm.

reflectance has a narrower bandwidth. A narrow bandwidth, less conspicuous reflection is sometimes selected for in animals, this will be discussed later in this article. Multilayer reflectors polarize light incident at Brewster's angles. This is about 54° for a quarter-wave stack of guanine and cytoplasm. At very oblique angles, all wavelengths are strongly reflected.

Multilayer reflectors are common in animals. They are usually extra-cellular, produced by periodic secretion and deposition, but sometimes occur within cells. Guanine ($n = 1.83$) is a common component in invertebrate reflectors because it is one of very few biological materials with a high refractive index and is readily available to most invertebrates as a nitrogenous metabolite. However, arthropods, including insects, crustaceans, and spiders, have largely ignored guanine in favor of pteridines. Also surprising is that the reflector material of closely related species, e.g., the molluscs *Pecten* (scallop) and *Cardium* (cockle), may differ.

Multilayers produce effects in beetle cuticle from highly metallic colors ('ideal' system) to rather dull greens ('nonideal' system in combination with scattering; Figure 5), and colors from the wings of many butterflies. Often in butterflies, layers of chitin ($n =$ about 1.56) are supported by vertical vanes of the scales. Air ($n = 1.0$) fills in the spaces and provides the alternate layers of the system. Under white light at normal incidence, the blue color of the butterfly *Arhopala micale* turns to green when the air is replaced by acetone. This is due to an increase in the refractive index of the low n layer (the actual thickness of this layer remains the same but the optical thickness increases) that effectively makes the system less 'ideal'. If the refractive indices of both layers were made equal, there would be no internal interference. A layer of melanin often underlies the reflector and intensifies the metallic colored effect by absorbing the transmitted portion of incident light. For example, in beetles, the elytra of *Anopognathus parvulus* appears metallic gold, green, or yellow in reflected light, and diffuse brown in transmitted light. Individual butterfly scales have been examined in detail to reveal a number of variations of quarter-wave stacks, sometimes in combination with other optical structures, to provide a range of colored effects.

The crustaceans *Limnadia* (Conchostraca), *Tanais tennicornis* (Tanaidacea), *Ovalipes molleri* (Decapoda), and the males of *Sapphirina* (Copepoda) all bear multilayer reflectors in their cuticles, in different forms (Figure 6). In contrast to the usual continuous thin layers, male sapphirinids have ten to fourteen layers of interconnecting hexagonal platelets within the epidermal cells of the dorsal integument. The reflector of *O. molleri* comprises layers that are corrugated and also slightly out of phase (Figure 7). At close to normal incidence this structure reflects red light, but at an angle of about 45° blue light is reflected. The corrugation, however, functions to broaden the reflectance band, at the expense of reducing the intensity of reflection.

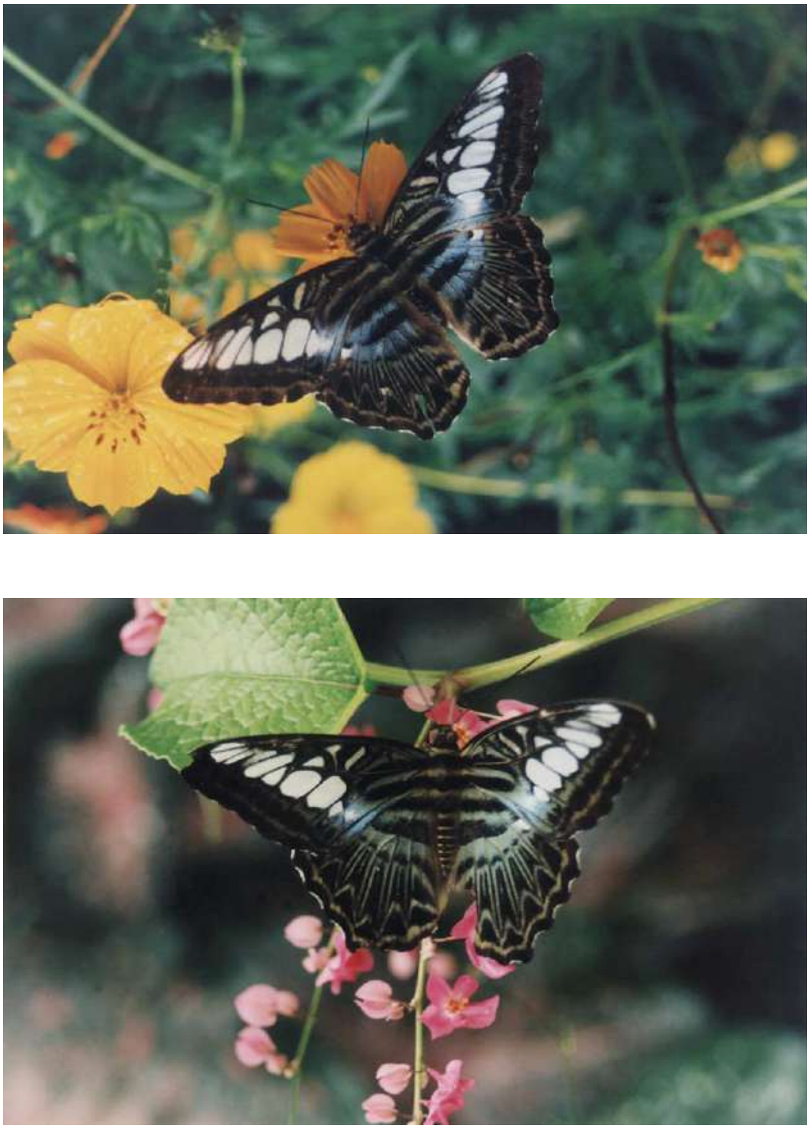

Figure 3 Butterfly (*Parthenos sylvia* (Nymphalidae) from Malaysia), showing the change in color with angle of wings. This animal may have evolved a temporal signal rather than a static one.

A broadband wavelength-independent reflectance, appearing silver or mirror-like to the human eye, can be achieved in a multilayer stack in at least three ways in invertebrates (Figure 8). These are (a) a composite of regular multilayer stacks each tuned to a specific wavelength, (b) a stack with systematically changing optical thicknesses with depth in the structure, termed a 'chirped' stack, and (c) a disordered arrangement of layer thicknesses about a mean value, termed a 'chaotic' stack (Figure 8). The nauplius eye of the copepod *Macrocyclops* (Crustacea) has regularly arranged platelets ~100 nm thick in stacks of 20–60, achieving the first condition. Silver beetles and the silver and gold chrysalis of butterflies in the genera *Euoplea* and *Amauris*, owe their reflection to the second condition. The mirror-like reflectors in the scallop *Pecten* eye comprise alternating layers of cytoplasm ($n = 1.34$) and guanine crystals ($n = 1.83$) and approximate an 'ideal' quarter-wave system in the same manner as within fish skin using the third mechanism. The ommatidia of the superposition compound eyes of *Astacus* (Crustacea) are lined with a multilayer of isoxanthopterin (a pteridine) crystals, which again fall into the third category. Multilayer reflectors can also be found in the eyes of certain spiders, butterflies, and possibly flies, where they assist vision, as discussed further below.

Squid and cuttlefish, for example, possess mirror-like reflectors in photophores (light organs) and

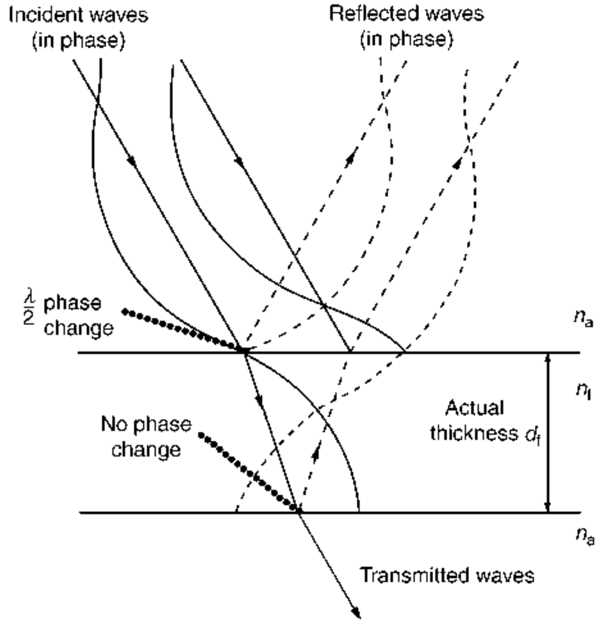

Figure 4 Schematic diagram of thin-film reflection. The direction of wave (straight line) and profile of electric (or magnetic) component are illustrated. Incident waves are indicated by a solid line, reflected waves by broken lines. Refraction occurs at each media interface. The refractive index of the film (n_f) is greater than the refractive index of the surrounding medium (n_a). Constructive interference of the reflected waves is occurring. As the angle of incidence changes, different wavelengths constructively interfere. At normal incidence constructive interference occurs where $n_f \times d_f = \lambda/4$.

iridophores. Iridophores are cells that, in this case, contain groups (iridosomes) of flexible layers of thin lamellae with cytoplasm between, forming a quarter-wave stack. The platelets of both squids and octopods develop from the rough endoplasmic reticulum and are separated by extracellular space. Euphausiid crustaceans possess photophores with very elaborate mirror-like reflectors. Up to 60 dense layers, about 70 nm thick and 75–125 nm apart, are formed from the aggregation of granules (probably a type of chitin) and surround the main photogenic mass. An intricate ring consisting of very flattened cells, forming the dense layers of a multilayer reflector (about 175 nm thick, separated by 90 nm) surrounds the lens of the photophore and reflects blue light at acute angles of incidence.

Dead invertebrates may not display their original colors. Following death, one (or both) of the layers in a multilayer reflector may become gradually reduced. For example, water may be lost from the system. This occurs in beetles of the genus *Coptocycla*; their brassy yellow color quickly changes through green, blue, and violet until the brown of melanin is finally observed. The color progression may subsequently be reversed by water uptake. This is an important consideration when examining fossils for multilayer reflectors. Multilayer reflectors have been identified in the shells of some ammonites, preserved as fossils.

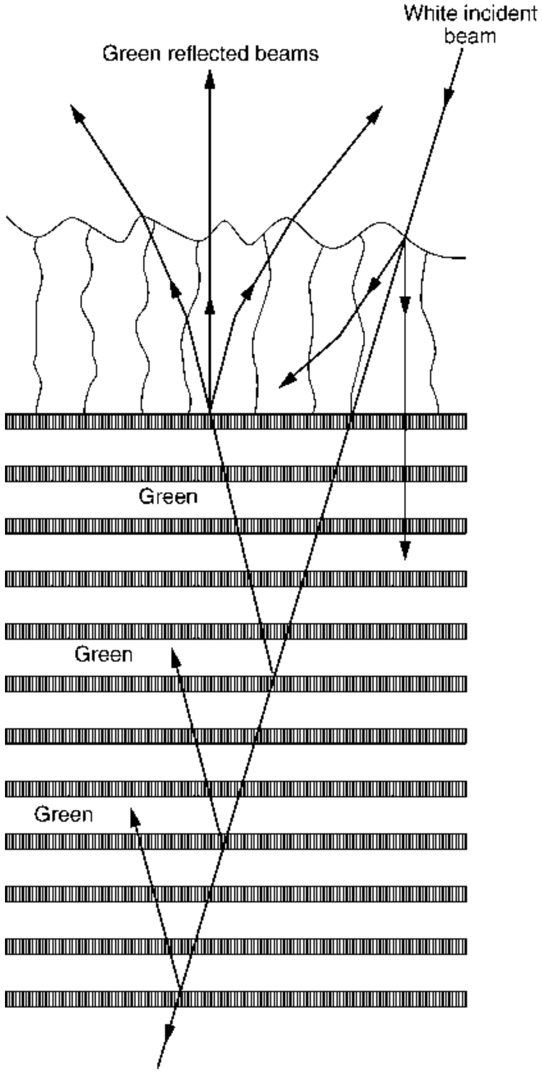

Figure 5 Generalized diagram of a multilayer reflector in the cuticle of the green beetle *Calloodes grayanus* (high refractive index material is shown shaded). Outer layer causes scattering.

Diffraction Gratings

When light interacts with a periodic surface consisting for example of a series of parallel grooves, it may be deviated from the direction of simple transmission or reflection. For this to happen, light which is scattered or diffracted from successive grooves should be out of phase by integral values of 2π. This occurs when, for a given direction of propagation, the optical path difference via successive grooves is $M\lambda$, where M is an integer known as the circle number.

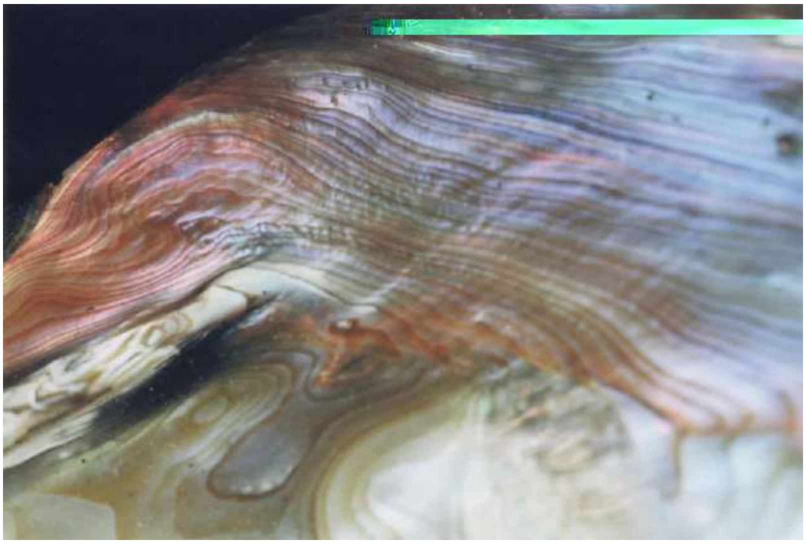

Figure 6 Bivalve shell with outer pigment removed to reveal an incidental multilayer reflector. Layers evolved to provide strength and prevent crack propagation.

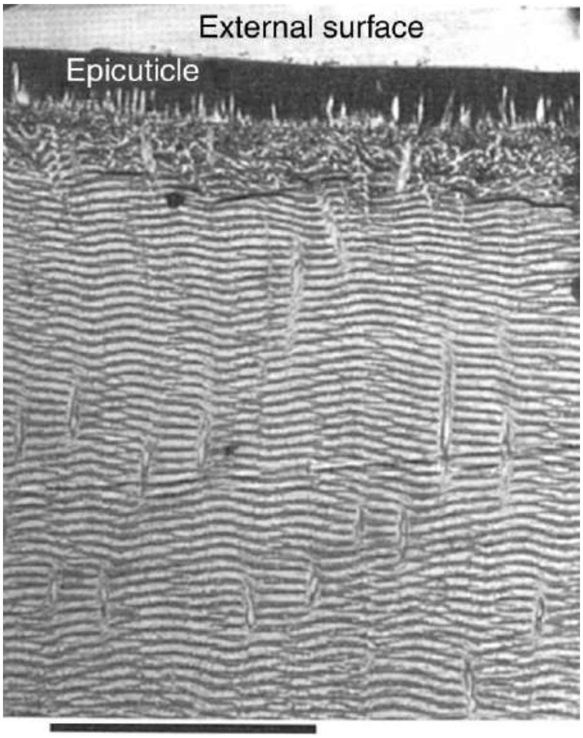

Figure 7 Transmission electron micrograph of a multilayer reflector in the cuticle of a swimming paddle of the crab *Ovalipes molleri* (Crustacea: Decapoda). Layers of reflector are slightly sinuous and out of phase; note the unusual side branches of each high refractive index (dark) layer, which provide support for these solid layers within a liquid matrix. Scale bar represents 5 μm.

This may be expressed by the grating equation:

$$2d(\sin\alpha - \sin\beta) = M\lambda \qquad [2]$$

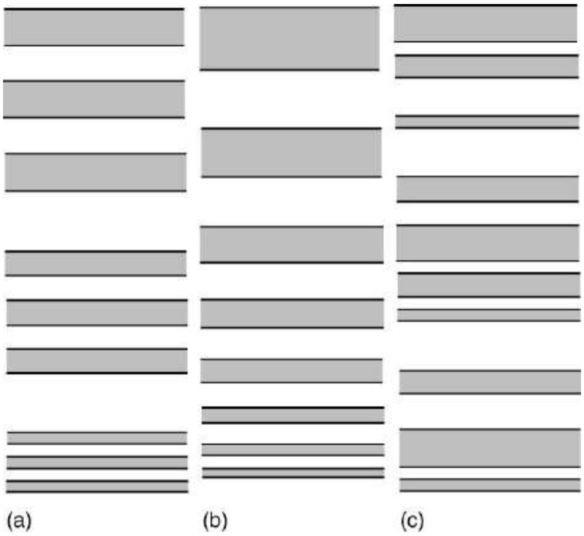

Figure 8 Three ways of achieving a broadband, wavelength-independent reflector in a multilayer stack (high refractive index material is shown shaded). (a) Three quarter-wave stacks, each tuned to a different wavelength. (b) A ‘chirped’ stack. (c) A ‘chaotic’ stack.

where α and β are angles of incidence and diffraction, and d is the period.

As with multilayers, the effect gives rise to coloration because different wavelengths are diffracted into different directions. Although the effect changes with angle of incidence, it is less critical than it is with thin films and the visual appearance is different. For a parallel beam of white light incident upon a multilayer, one wavelength will be reflected as determined by the Bragg Condition. The same beam

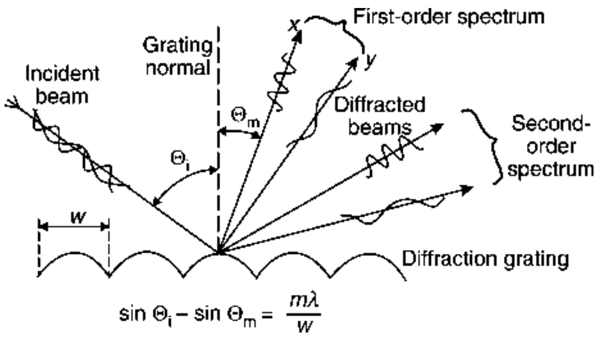

Figure 9 Reflection-type diffraction grating ($w =$ periodicity, $m =$ order of reflected beam). For white light of fixed angle of incidence (Θ_i), the color observed is dependent on the point of observation (e.g., violet light can be seen at point x, red at point y, within the first order spectrum).

incident upon a grating will be dispersed into spectra (e.g., Figure 9). The complete spectrum reflected nearest to the perpendicular (grating normal) is the first order. The first-order spectrum is reflected over a smaller angle than the second-order spectrum, and the colors are more saturated and appear brighter within the former. Diffraction gratings have polarizing properties, but this is strongly dependent on the grating profile.

Diffraction gratings were believed to be extremely rare in nature, but have recently been revealed to be common among invertebrates. They are particularly common on the seate or setules (hairs) of Crustacea, such as on certain first antennal setules of male Myodocopina ostracods or ‘seed shrimps’ (Crustacea) (Figure 10). Here, the grating is formed by the external surface of juxtaposed rings with walls circular in cross-section (Figure 11). The width of the rings, and consequently the periodicity of the grating, is about 700 nm in *Azygocypridina lowryi*. Different colors are viewed with varying angles of observation under a fixed light source. The ostracod *Euphilomedes carcharodonta*, for example, additionally houses a diffraction grating on the rostrum, a continuous flattened area of the carapace that is corrugated to form periodic ridges. The dark brown beetle *Serica sericea*, bears gratings on its elytra with 800 nm periodicity, which causes a brilliant iridescence in sunlight.

Many polychaetes possess gratings on their setae (hairs). For example, the opheliid *Lobochesis longiseta* bear gratings with periodicities in the order of 500 nm (Figure 12), appearing iridescent. The wings of the neurochaetid fly *Neurotexis primula* bear diffraction gratings only on their dorsal surfaces, and the iridescent effect remains after the insect is gold coated for electron microscopy. These gratings cause iridescence with a higher reflectance than the iridescence of the membranous wings of other insects, which reflect light by interference. Iridescence caused by interference disappears after gold coating because transmission of light through the outer surface is prevented.

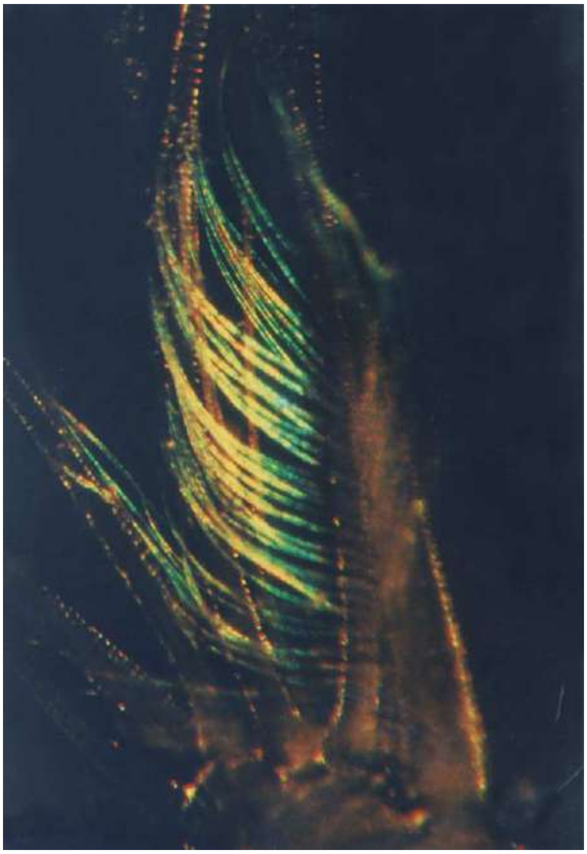

Figure 10 Diffraction gratings on the halophores (hairs) on the first antenna of the seed shrimp *Azygocypridina lowryi* (Ostracoda).

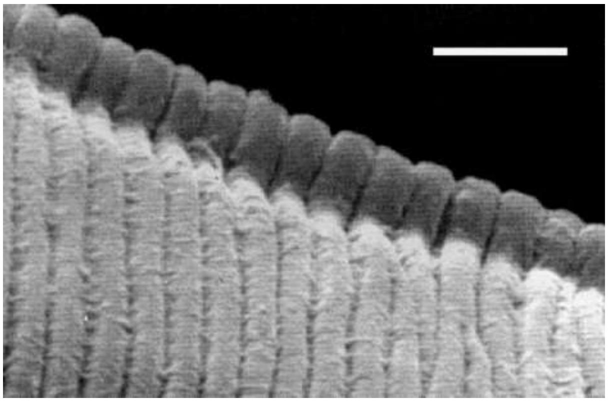

Figure 11 Scanning electron micrograph of diffraction grating on a halophore (setule) of *Azygocypridina lowryi* (Ostracoda). Scale bar represents 2 μm.

Very closely spaced, fine setules may also form the ridges of a diffraction grating. Cylindroleberidid

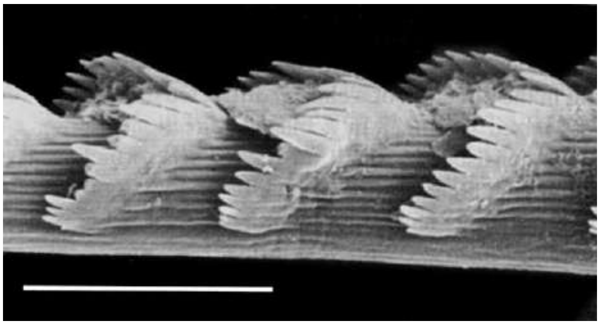

Figure 12 Scanning electron micrograph of diffraction gratings on a hair of *Lobochesis longiseta* (Polychaeta). Scale bar represents 10 μm.

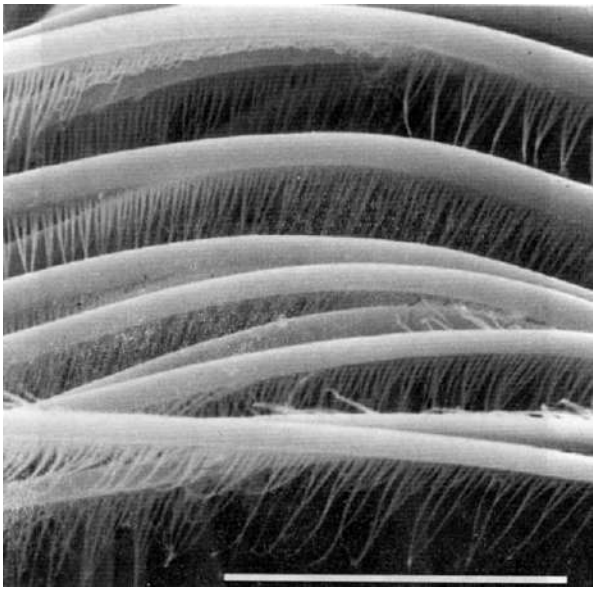

Figure 13 Scanning electron micrograph of *Tetraleberis brevis* (Ostracoda), hairs of comb of maxilla; setules (orientated vertically) form the ridges of a grating. Scale bar represents 2 μm.

ostracods (seed-shrimps) possess a comb on their maxilla bearing numerous setules on each seta (hair), collectively forming a grating with a periodicity of about 500 nm (Figure 13).

The 'helicoidal' arrangement of the microfibrils comprising the outer 5–20 μm of the cuticle (exocuticle) of certain scarabeid beetles, such as *Plusiotis resplendens*, also gives rise to metallic colors. Here, the fibrils are arranged in layers, with the fibril axis in each layer arranged at a small angle to the one above, so that after a number of layers the fibrillar axis comes to lie parallel to the first layer. Thus, going vertically down through the cuticle, two corresponding grating layers will be encountered with every 360° rotation of the fibrils – the 'pitch' of the system. Polarized light encounters an optically reinforcing plane every half turn of the helix. The system can be treated as a three-dimensional diffraction grating, with a peak reflectance at $\lambda = 2nd$, where d is the separation of analogous planes, or half the pitch of the helix. The diffracted light resembles that from a linear grating except for the polarization; the three-dimensional grating reflects light that is circularly or elliptically polarized. It should be noted that the diffracted color does not depend on the total film thickness as it does in interference, but on the layer repeat distance within the film as in the diffraction grating (analogous to 'liquid crystals').

When each groove of a grating is so formed that it independently, by means of geometrical optics, redirects the light in the direction of a chosen diffracted order, it is known as a blazed grating. In a blazed reflection grating, each groove consists of a small mirror (or prism) inclined at an appropriate angle (i.e., the grating has a 'saw-tooth' profile). Blazed gratings have been identified on the wing scales of the moth *Plusia argentifera* (Plusinae).

When the periodicity of a grating reduces much below the wavelength of light, it becomes a zero-order grating and its effect on light waves changes. This difference in optical effect occurs because when the periodicity of the grating is below the wavelength of light the freely propagating diffracted orders are suppressed and only the zero-order is reflected when the illumination is normal to the plane of the grating. To describe accurately the optical properties of a zero-order grating, rigorous electromagnetic theory is required. In contrast to gratings with freely propagating orders, zero-order structures can generate saturated colors even in diffuse illumination. Such structures occur on the setae (hairs) of the first antenna of some isopod crustaceans, such as the giant species of *Bathynomus*. Here, there are diffracted orders and the spectral content of the light within the grating is controlled by the groove profile. In an optical system that only accepts the zero-order, what is seen is white light minus that diffracted into the ±1 orders.

A zero-order grating can cause total transmission (i.e., there is no reflection). Such antireflective structures are found on the corneal surfaces of each ommatidium (visual unit) in the eye of *Zalea minor* (Diptera) (Figure 14). The periodicity of the corneal gratings of this fly is 242 nm. Another form of antireflection grating is formed on the transparent wings of the hawkmoth *Cephonodes hylas* on the corneal surface of each visual unit (ommatidium) of the eyes of moths and butterflies (Figure 15). Here, optical-impedance matching is achieved by means of a hexagonal array of tapered cylindrical protuberances, each of about 250 nm diameter, thus forming a 'tri-grating' with grooves transecting at 120°.

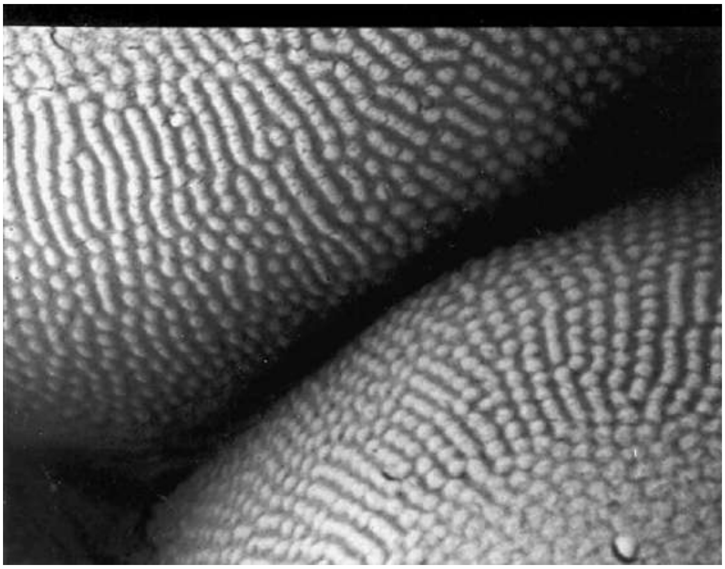

Figure 14 Scanning electron micrograph of the compound eye of *Zalea minor* (Diptera). The corneal surface at the junction of two ommatidia is shown with antireflection gratings (positioned on sloping regions, i.e., at high angles to the normal of the eye) phasing into moth-eye-type protuberances. Scale bar represents 2 μm.

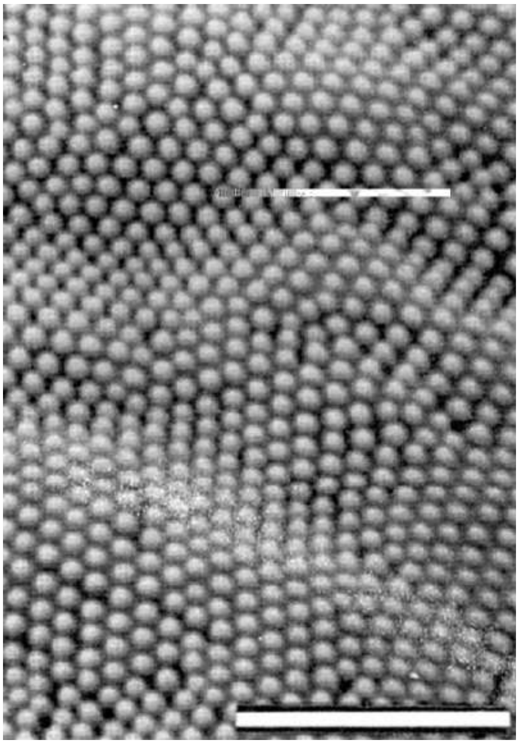

Figure 15 Tri-grating (antireflector) on the corneal surface of a butterfly eye. Scale bar represents 2 μm.

The protuberances provide a graded transition of refractive index between the air and the cornea/wing. Hence the refractive index at any depth is the average of that of air and the corneal/wing material.

The grooves of a grating may also create parallel rows in two directions, forming a bi-grating. Bi-gratings can be found in some crustaceans and flies. In the amphipod crustacean *Waldeckia australiensis*, two effectively superimposed gratings subtend angles of about 60/120° (Figure 16).

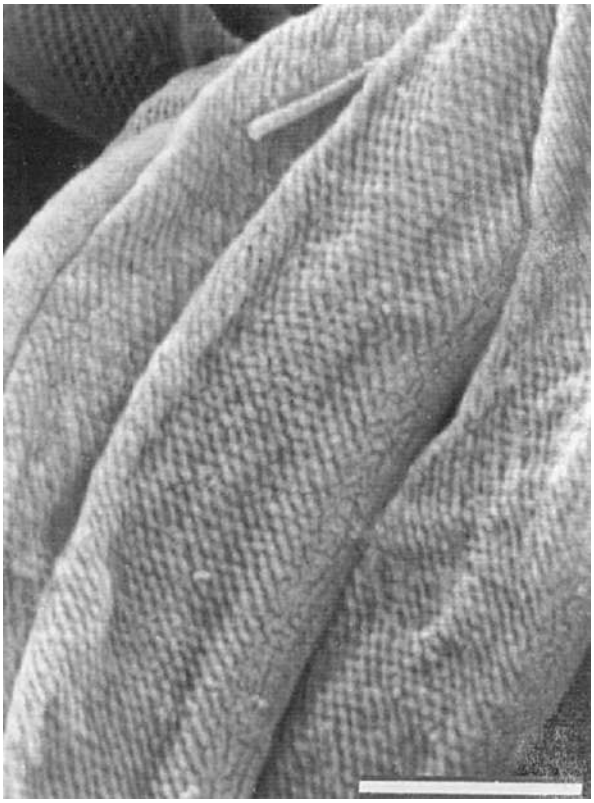

Figure 16 Bi-grating on the callynophores (hairs) of the amphipod crustacean *Waldeckia australiensis*. Scale bar represents 2 μm.

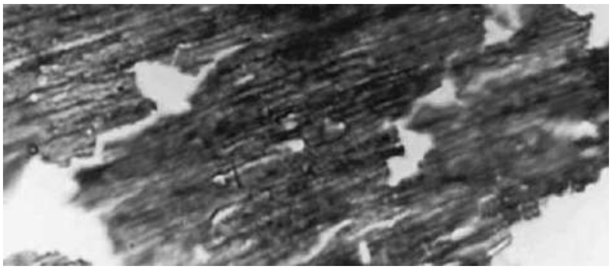

Figure 17 Light micrograph of the surface of a hair of *Canadia spinosa* (Middle Cambrian, 515 million years old, British Columbia), isolated by acid maceration of the rock matrix, showing gratings running longitudinally. Periodicity 0.9 μm.

In terms of fossil colors, diffraction gratings are responsible for the nacreous luster of brachiopod shells, such as those from the Devonian, around 360 million years old. Here, tabular aragonite platelets averaging 600 nm in thickness, each comprising a linear diffraction grating, form layers and consequently a three-dimensional diffraction grating. Antireflective, zero-order gratings have been identified on the eye of an Eocene fly, 45 million years old, preserved in Baltic amber. Linear diffraction gratings causing color have been discovered on the sclerites of *Wiwaxia corrugata* from the Burgess Shale (Middle Cambrian, 515 million years old) (Figure 17). This finding has led to the 'light switch theory', in that the evolution of the first eye caused the big bang in animal evolution.

Scattering (Incoherent Reflectors)

Simple, equal scattering of all spectral wavelengths results in the observation of a diffuse white effect. This commonly arises from the effects of a nonperiodic arrangement of colloidally dispersed matter where the different materials involved have different refractive indices (Figure 18), or from solid colorless materials in relatively concentrated, thick layers. In the colloidal system, the particles are larger than the wavelength of light and can be thought of as mirrors oriented in all directions. The reflection is polarized unless the incident light is at normal incidence on the system and, in the colloidal system, spherical or randomly arranged particles are involved.

The colloidal system involves either a gas-in-solid, gas-in-liquid, liquid-in-liquid (emulsions), or solid-in-liquid. For example, the gas-in-liquid system is partly responsible for the white body and/or tentacles of certain anemones. Light is reflected and refracted at the surfaces of the particles of matter or spaces (with dimensions >1 μm), regardless of the color of the materials involved (except opaque brown and black compounds, such as melanin). In insects, the materials involved typically have very low transparencies.

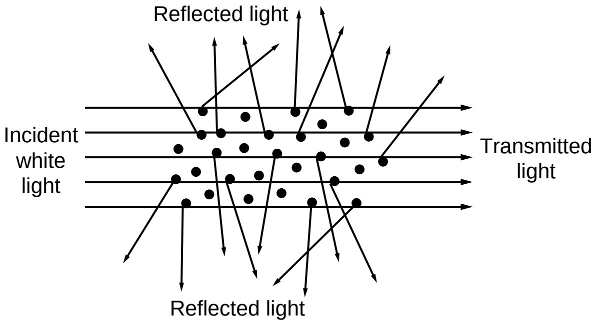

Figure 18 Scattering of white light by small particles (represented by black circles). The reflected light is randomly directed at different angles.

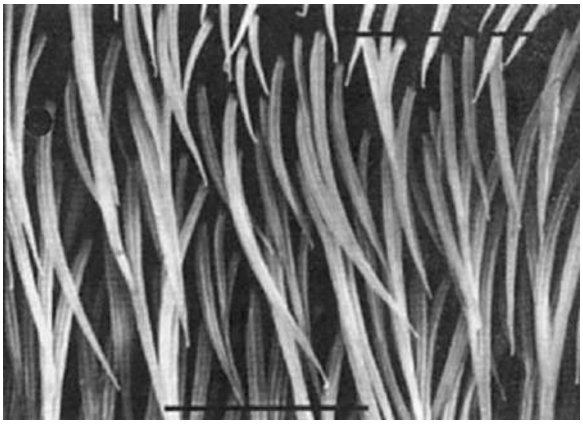

Figure 19 Scanning electron micrograph of a section of a white reflecting patch on the abdomen of the fly *Amenia* sp., showing closely packed, randomly arranged hairs that scatter incident light in all directions (Figure 11). Scale bar represents 20 μm.

An unordered (as opposed to periodic) group of closely spaced setae (hairs), such as those in patches on the fly *Amenia* sp., may form a white reflection via random scattering or reflection (Figure 19). However, if the arrangement becomes periodic to some degree, a diffraction grating may be formed, such as the grating of *Tetraleberis brevis* (Ostracoda) (Figure 13).

From some scales of butterfly wings, light is scattered uniformly and completely in all directions, due to the chaotic disposition of the surfaces. Matt or pearly whites may be observed depending on the complexity of the arrangement of the structures, which affects the relative degree of scattering. The structures may be so small that the molecular topography of the surface has an effect. The chromatic effects of the butterfly scales are greatly intensified if a dark, absorbing pigment screen lies beneath. This screen prevents reflection of white or colored light from the background that would dilute or alter the color. Additionally, if a dark pigment such as melanin is interspersed with the scattering elements, the reflection will appear a shade of gray or brown. The cells of the reflector in the photophore of a beetle ('fire-fly') are packed with sphaerocrystals of urate that cause a diffuse reflection (Figure 20).

Reflection and refraction that occur at the interfaces of strata with different refractive indices may result in the display of white light. The degree of whiteness depends upon the difference in refractive indices. This mechanism is evident in the shells of many lamellibranch molluscs. Between the outer, often pigmented layer and the mantle, is a thick middle layer of crystalline calcium carbonate. The inner surface of this (nacreous) layer is lined with multiple laminations of the same salt. In most species these laminations are sufficiently thick (>1 μm) to render the inner lining white, although in some species they become thin so as to form a multilayer reflector. Calcium carbonate similarly produces whiteness in Foraminifera and in calcareous sponges, corals, echinoderms, and crustacean cuticles. Also, in the class of white solids is silica in diatom tests and skeletons of hexactinellid sponges.

Other forms of scattering also exist and result in a blue colored effect (red when the system is viewed in transmission). Tyndall or Mie scattering occurs in a colloidal system where the particle size approximates the wavelength of light. Here, diffraction is important. Rayleigh scattering occurs in molecules in a two-photon process by which a photon is absorbed and raises the molecule to an excited electronic state from which it re-radiates a photon when it returns to the ground state. Diffraction is not involved here.

Tyndall scattered light is polarized under obliquely incident light. The intensity of the resultant blue is increased when it is viewed against a dark background, such as melanin. The relative sizes of particles determine the shade of blue. If the particles responsible for the scattering coalesce to form particles with a diameter greater than about 1 μm, then white light is observed (see above). A gradation from blue to white scattering ('small' to 'large' particles) occurs on the wings of the dragonfly *Libellula pulchella*.

Scattered blues can also be found in other dragonflies. In the aeschnids and agrionids, the epidermal cells contain minute colorless granules and a dark base. The males of libellulids and agrionids produce a waxy secretion that scatters light similarly over their dark cuticle. The green of the female *Aeschna cyanea* is the combined result of Tyndall scattering and a yellow pigment, both within the epidermal cells (degradation of the yellow pigment turns the dead dragonfly blue).

Scattered blues are also observed from the skin of the cephalopod *Octopus bimaculatus*, where a pair of ocelli are surrounded by a broad blue ring. Blue light is scattered from this region as a result of fine granules of purine material within cells positioned above melanophore cells. The color and conspicuousness of the ring are controlled by the regulation of the melanophores, by varying the distribution of melanin

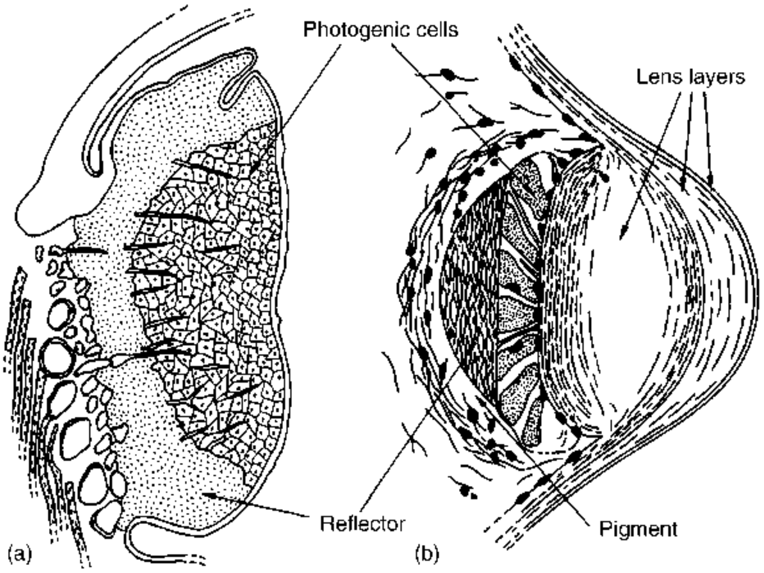

Figure 20 Sections of photophores (bioluminescent organs), external surface of animals on the right. (a) The 'fire-fly' beetle *Pyrophorus* sp., abdominal photophore of male – reflector is based on a scattering system. (b) The shrimp *Sergestes prehensilis*, showing lens layers and absorptive pigment – reflector is based on a multilayer type.

and consequently the density of the absorbing screen. The squid *Onychia caribbaea* can produce rapidly changing blue colors similarly. The bright blue patterns produced by some nudibranch molluscs result from reflecting cells containing small vesicular bodies, each composed of particles about 10 nm in diameter and therefore appropriate for Rayleigh scattering.

See also

Diffraction: Diffraction Gratings. **Scattering:** Scattering Theory.

Further Reading

Fox DL (1976) *Animal Biochromes and Structural Colours*. Berkeley, CA: University of California Press.

Fox HM and Vevers G (1960) *The Nature of Animal Colours*. London: Sidgwick and Jackson Ltd.

Gale M (1989) Diffraction, beauty and commerce. *Physics World* 2: 24–28.

Harvey EN (1952) *Bioluminescence*. New York: Academic Press Inc.

Herring PJ, Cambell AK, Whitfield M and Maddock L (1990) *Light and Life in the Sea*. Cambridge, UK: Cambridge University Press.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Land MF and Nilsson D-E (2002) *Animal Eyes*. Oxford, UK: Oxford University Press.

Pain S (1998) Seeing the light. *New Scientist* 2161: 42–46.

Parker AR (1999) Light-reflection strategies. *American Scientist* 87: 248–255.

Parker AR (2003) *In the Blink of an Eye*. London: Simon and Schuster/Cambridge (USA): Perseus Publishing.

Parker AR, Welch VL, Driver D and Martini N (2003) An opal analogue discovered in a weevil. *Nature* 426: 786–787.

Parker AR, McPhedran RC, McKenzie DR, Botten LC and Nicorovici N-AC (2001) Aphrodite's iridescence. *Nature* 409: 36–37.

Microstructure Fibers

R S Windeler, OFS Laboratories, Murray Hill, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Microstructure fibers have unique properties that cannot be obtained from traditional fibers (i.e. all glass, doped silica fibers) and can deliver functionalities superior to many of today's best transmission and specialty fibers. Their unique properties are obtained from an intricate cross-section of high and low index regions that traverse the length of the fiber. The vast majority of fibers consist of silica for the high-index region and air for the low-index region. These fibers are known by several different names including microstructure fiber, holey fiber, and photonic crystal fiber. The term microstructure fiber is used in this chapter because it is the most generic, encompassing all the fiber types.

Microstructure fiber properties vary greatly and are determined by the size and arrangement of air holes in the fiber. For example, fibers have been fabricated such that the numerical aperture of the core or cladding approaches one. They have been fabricated with unique dispersion profiles, such as a zero-group velocity dispersion in the near visible regime or an ultraflat dispersion over hundreds of nanometers wide. They have been fabricated to generate a super continuum over two octaves wide. Some guide light with an air core via bandgap guidance. In addition, the air holes in these microstructure fibers can be filled with highly tunable materials, giving one the capability of controlling the fiber properties for use in energy efficient devices.

Microstructure fibers can have doped regions like traditional fibers. These hybrid fibers combine the benefits of traditional and microstructure fibers. The doped core typically guides most of the light, but its guidance properties can be strongly influenced by the air regions. Applications for hybrid fibers include dispersion compensation, polarization maintaining, and bend insensitive fibers.

Loss is always an important factor when determining whether a microstructure fiber will compete with or replace a traditional optical fiber. The index profiles that make these fibers so special can also lead to fibers that have an inherently high loss at connections or along the length of the fiber. Loss at connections typically occurs due to mode mismatch between the two fibers, undesired hole collapse, or poor alignment. Loss along the fiber length can occur due to impurities, hole surface roughness, or poor confinement.

The dispersion profile (zero dispersion wavelength, dispersion slope, normal or anomalous) of microstructure fibers can be optimized more than a traditional fiber, because the hole size and placement can be arranged to make the cladding index strongly wavelength dependent. Dispersion causes a pulse to spread because the phase velocity is wavelength dependent. Normal dispersion is when longer wavelengths travel faster than shorter wavelengths. Anomalous dispersion is just the opposite – shorter wavelengths travel faster than longer wavelengths. Dispersion determines the amount of interaction between different wavelengths. For applications such as transmission fiber, one wants little interaction and for high nonlinear applications, one wants significant interaction.

This chapter examines the methods for fabricating microstructure fibers and reviews the different types of microstructure fibers and their properties. Fibers that guide by total internal reflection are reviewed separately from bandgap guided fibers because their properties are significantly different. Lastly, a description and example of an active device is presented in which the air regions of a microstructure fiber are filled with controllable material.

Microstructure Fiber Fabrication

At first glance, fabrication of microstructure fibers looks similar to traditional fibers in that the fibers are fabricated in a two-step process. First a preform is fabricated and then it is drawn (stretched) into a fiber. However, when the process is examined in more detail, both the preform fabrication and draw depart significantly from traditional methods. First, a novel preform fabrication method is used to incorporate air holes that run the length of the preform. Second, novel drawing procedures are used to keep the air holes open as the preform diameter is reduced by several orders of magnitude during draw.

Preform Fabrication

Microstructure preforms are cylinders of amorphous material usually less than a few centimeters in diameter that have an index profile running the length of the preform similar to the desired fiber. The most common preform consists of air and a single material such as silica, polymer, or high nonlinear glass. These air-containing preforms are relatively easy to fabricate but are difficult to draw because the structures can deform.

The basic parameters of the fiber are usually determined by the geometry of the holes in the preform. These include hole diameter (d), hole position, pitch (center-to center hole spacing, Λ), core diameter (a), and number of layers. However, changes in the size and shape of the holes in the fiber can be made purposely or accidentally during draw, which will cause deviations between the fiber and preform profiles.

Several methods are used to fabricate microstructure preforms including stacking, casting, extrusion, and drilling. Each process has its advantages and disadvantages. By far, the most common method is stacking in which tubes, rods, and core rods (rods containing doped cores fabricated by traditional optical fiber techniques) are stacked in a close packed pattern such as a triangular or hexagonal lattice (Figure 1). The bundle of tubes and rods is then bound together with a large tube, called an overclad tube. Fibers can be made with complicated or asymmetric index profiles by strategically placing tubes with the same outer diameter (for good stacking) but a different inner diameter to change the index in that location. For example, a fiber with smaller holes on opposite sides of the core produces high birefringence.

The main advantages of stacking are that no special equipment is needed for fabricating microstructure preforms and that doped cores are easily added. However, there are several disadvantages. First, the stacking method is limited to simple geometries of air holes because the tubes are stacked in a close pack arrangement. Second, unless hexagonal tubes are used, interstitial areas are created between the tubes that may not be desired in the final fiber. Third, the method is labor intensive and requires significant glass handling.

A variation of the stacking method is used for fabricating air-clad fibers. Air-clad preforms are created by placing a layer of small tubes around the perimeter of a large core rod. The assembly is then overcladed for strength (Figure 2).

Extrusion and casting of glass powder, polymer, or sol–gel slurry are also used for making single material

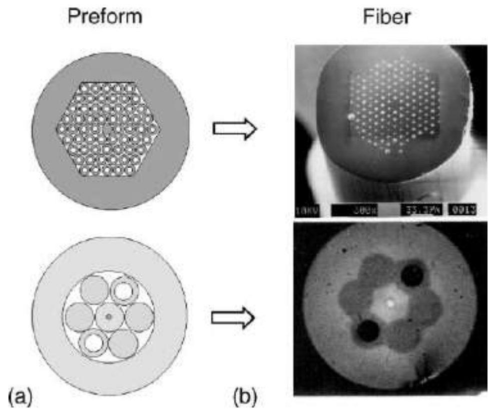

Figure 1 (a) Tubes, rods, and core rods are stacked together in a close packed arrangement and held together with an overclad tube. (b) During draw the desired air region stay open forming a microstructure fiber.

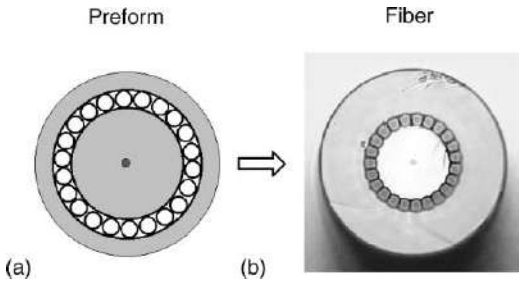

Figure 2 (a) An air-clad preform is fabricated by stacking a ring of small, thin walled tubes around a large core rod. (b) Photograph of the resulting fiber.

microstructure preforms. The main advantage of extrusion and casting is that complicated structures can be fabricated in which the position, size, and shape of the air regions are independent of one another. The disadvantage of this method is that preform fabrication is more difficult compared to the other fabricate methods. This method may become more common as complex air structures are needed to make advanced microstructure fibers.

The last method consists of drilling holes in a preform or rod. Drilling is well understood and used for other specialty fibers. The advantage of this method is that it is easy to drill holes of various sizes in any position in a preform, including doped regions. The disadvantages of drilling are that the holes cannot be drilled very deep compared to a traditional preform length; the distance between holes may be limited due to cracking; and the fiber may experience high loss due to surface roughness of the holes and impurities incorporated during drilling.

Fiber Draw

It is significantly more difficult to draw a microstructure preform with air holes than a traditional preform. This is because the air holes tend to close due to surface tension. This force can be overcome or minimized using two different techniques. The first is to draw the preform under very high tension (low temperature). Minimal hole collapse occurs when the draw tension is significantly larger than the surface tension. The problem with this method is that the break rate increases substantially at higher tension.

The second method is to apply an external pressure to the holes to counteract the surface tension. If a single pressure source is used, the holes can be made larger or smaller during draw by changing the pressure as needed. However, the process can become unstable because a large hole needs a lower pressure to maintain its size than a small hole. If holes of different sizes exist, a larger hole will grow at the expense of a smaller one regardless of the pressure used. To minimize the instability, this method is typically performed in conjunction with high-tension drawing.

To alleviate the draw problem, a second solid material can be used in the preform to obtain the desired index profile. With the air regions removed from the preform, the drawing process becomes significantly easier. In addition, when a preform consists of two solid materials, one of the materials does not have to be continuous, as in single material preforms. This allows for simpler designs such as concentric rings, which may be easier to fabricate and model (Figure 13c). The disadvantage of this approach is finding two materials that have compatible physical and chemical properties, have low loss, and have a large index of refraction difference.

Microstructure Fiber Types

Microstructure fibers are categorized by their method of guidance, properties, and function. Index guided microstructure fibers most closely resemble traditional fibers and are described first. These fibers guide light by total internal reflection and have a core index of refraction greater than the cladding. Bandgap fibers are examined next. The unique guidance of bandgap fibers allows the core index to be lower than the cladding index. Tunable microstructure fibers are described last. These microstructure fibers are filled with tunable materials so that the fiber's properties can be actively manipulated.

Index Guided Fibers

Most microstructure fibers guide light by total internal reflection. The size and spacing of air holes determine the guiding properties. For example, a solid silica core surrounded by a cladding consisting of small air holes that are spread apart creates a small core-to-cladding index difference similar to what can be achieved in traditional fibers. At the other extreme, a solid silica core surrounded by a cladding of large holes, which are closely spaced together, forms a large core-to-cladding index difference. This structure creates a large numerical aperture in the core and optically isolated regions within the fiber. The ability to create regions of significantly different indices in various parts of the fiber distinguishes the index guided microstructure fibers from traditional fiber. Different types of microstructure fiber are formed by varying the size, spacing and pattern of the air holes.

Different techniques are used to model these fibers, depending on spacing and size of the air holes relative to the wavelength of the guided light. If the hole diameter and spacing is much smaller than the wavelength of the guided light, the cladding index can be modeled as an average index weighted by the volume fraction of the two materials (typically air and silica). As the hole diameter or spacing approaches the wavelength of the guided light, complicated models are used such as finite difference method, finite element method, beam propagation method, or multipole method.

Endlessly single-mode fiber

The endlessly single-mode fiber consists of a solid core surrounded by a two dimensional array of small air holes in the cladding (Figure 3). The size and position of the air holes are arranged such that only the fundamental mode exists in the core regardless of the wavelength of the guided light.

The prevention of higher order modes can be understood by examining the V number. The V number is a dimensionless parameter that describes the guiding nature of the waveguide. In traditional fiber, when the V number is less than 2.405, the fiber will guide only the fundamental mode. The V number is given by

$$V = \frac{2\pi a}{\lambda}(n_{\text{core}}^2 - n_{\text{clad}}^2)^{1/2} \qquad [1]$$

where a is the core diameter, λ is the wavelength of the guided light, and n_{core} and n_{clad} are the index of refractions of the core and cladding, respectively. In traditional fiber, the core and cladding index (n_{core} and n_{clad}) are for the most part constant, so the V number increases as the wavelength decreases.

However, for the endlessly single mode microstructure fiber, the V number does not increase indefinitely with decreasing wavelength, but approaches a constant value. Obtaining a V number independent of wavelength is understood by examining the distance the light propagates into the cladding as a function of wavelength. As the wavelength becomes shorter, light penetrates a shorter distance into the cladding, interacting with less of the air regions. This causes the effective cladding index (n_{clad}) to increase with decreasing wavelength, such that V approaches a constant value.

Microstructure fibers are made endlessly single-moded over all wavelengths by properly designing the hole diameter-to-pitch ratio (d/Λ). When the d/Λ is low enough, the microstructure fiber guides light only in the single mode regardless of wavelength and core size (Figure 4). However, these large core diameter fibers are limited by long wavelength bend loss in the same way as large core single mode traditional fibers.

High delta core fiber

High delta core microstructure fibers (HDCMF) consist of a small (typically less than 2 microns) solid silica core surrounded by one or more layers of large air holes closely spaced together (Figure 5). This creates a very large index difference between the core and cladding. A core-to-cladding index difference of 0.4 is easily achieved in a HDCMF, which is an order of magnitude larger than can be achieved with traditional fiber.

Even though the core is very small, these fibers are almost always multimoded due to the large index difference between the core and cladding. However, because of its small core, it is difficult to launch anything other than the fundamental mode into the fiber. Once a mode is guided in the fiber it remains in that mode due to the large difference in effective indices between modes (Figure 6). When higher-order modes are purposely generated in the fiber, they also

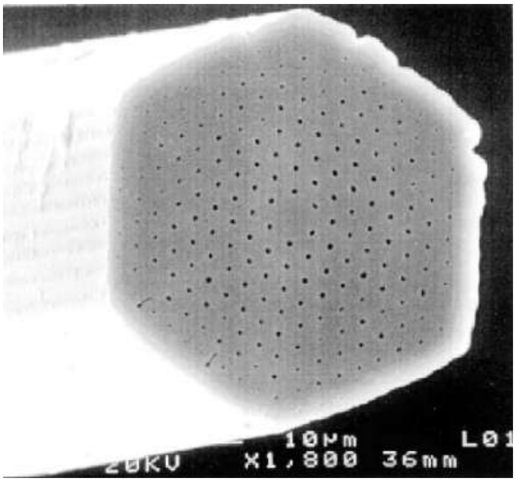

Figure 3 Photograph of an endlessly single mode fiber (Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963).

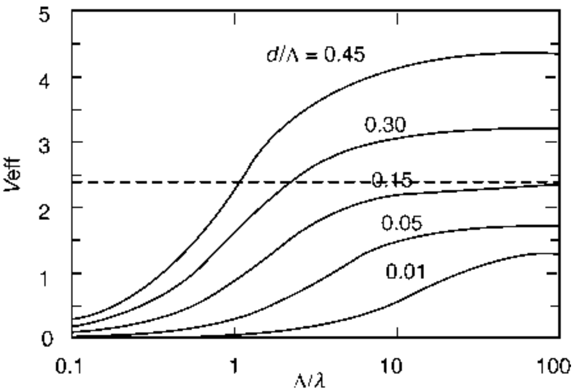

Figure 4 The effective V number as a function of Λ/λ for various d/Λ (Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963).

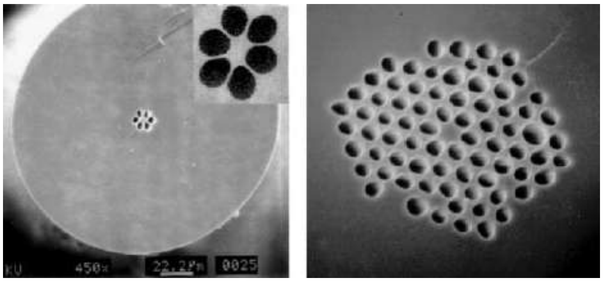

Figure 5 High delta core fibers consist of a small solid core surrounded by at least one layer of large air holes closely spaced together. Right-hand panel reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

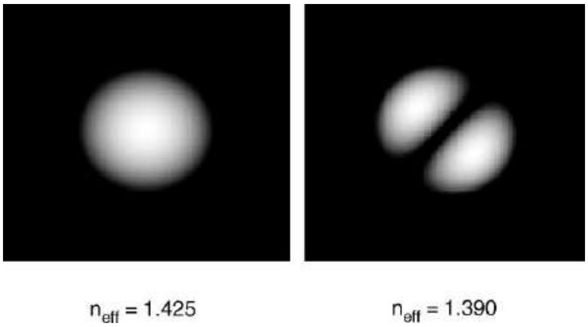

Figure 6 Individual modes do not couple well due to large differences in effective indices between modes. The figure shows the mode profile and effective index of the fundamental and the next higher mode for a fiber with a two-micron diameter core with a core-to-cladding index difference of 0.45 (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798).

guide over long lengths without coupling to other modes (Figure 7). This strong guidance makes the fiber extremely bend insensitive.

HDCMF can have a unique dispersion profile because it can guide the fundamental mode exclusively in the multimode regime. This allows greater flexibility in controlling the shape of the dispersion profile, especially the ability to move the zero dispersion to shorter wavelengths. The HDCMF is the first silica fiber with anomalous dispersion of single-mode light below 1270 nm (the wavelength at which the material dispersion of silica is zero). These fibers are often made to have a zero-group velocity dispersions around 770 nm for use with Ti-sapphire lasers.

HDCMF is ideal for performing nonlinear experiments in the near-visible regime because the fiber's small, high-index core creates high intensity light, and the fiber's low dispersion creates long interaction times. For example, a broadband continuum can be generated over two octaves using less than a meter of HDCMF fiber (Figures 8 and 9).

Tapered microstructure fiber (TMF)

Coupling light in and out of HDCMF is difficult and results in a large loss due to its small core. When free space optics are used to couple the light, frequent realignment of the beam is required to minimize loss due to drifts in the equipment. Splicing HDCMF fiber to traditional fiber also results in high loss due to a large modal mismatch. These coupling problems are solved with tapered microstructure fibers.

Tapered microstructure fibers (TMF) consist of a traditional germanium single-mode core surrounded by a small silica cladding (Figure 10). The inner

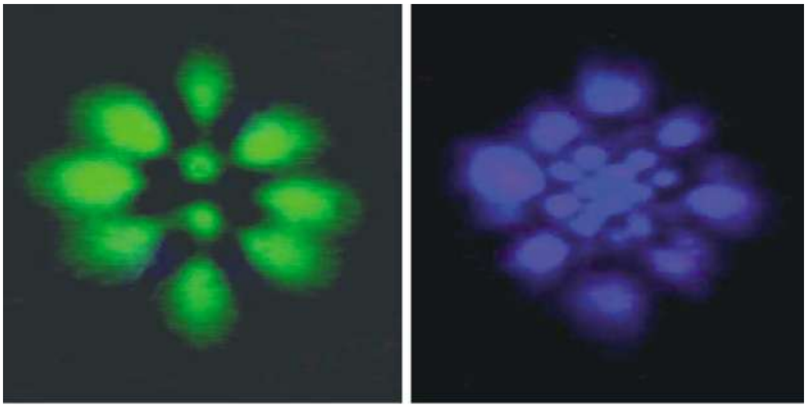

Far-field pattern, 532 nm light
5 meter fiber length

Far-field pattern, 355 nm light
2 meter fiber length

Figure 7 Far-field patterns show that higher order modes propagate long distances in the HDCMF without coupling to other modes (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798).

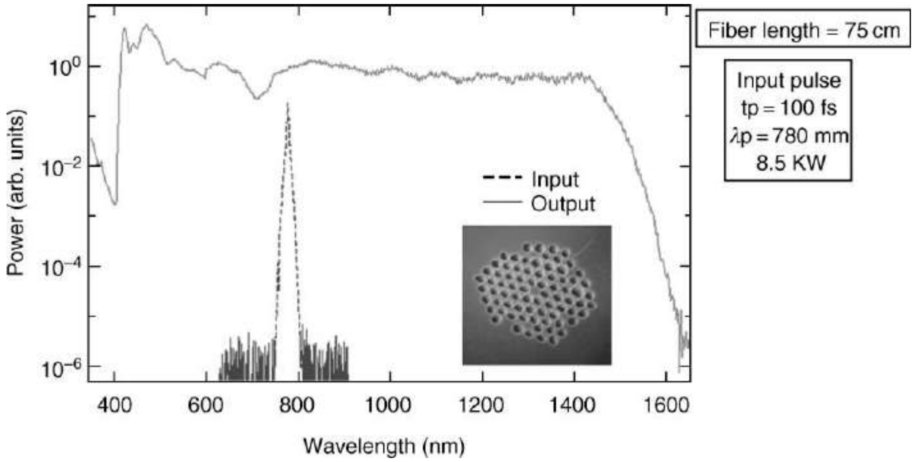

Figure 8 Plot of a supercontinuum spectrum over two octaves wide generated from a 75 cm piece of HDCMF (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25: 25–27).

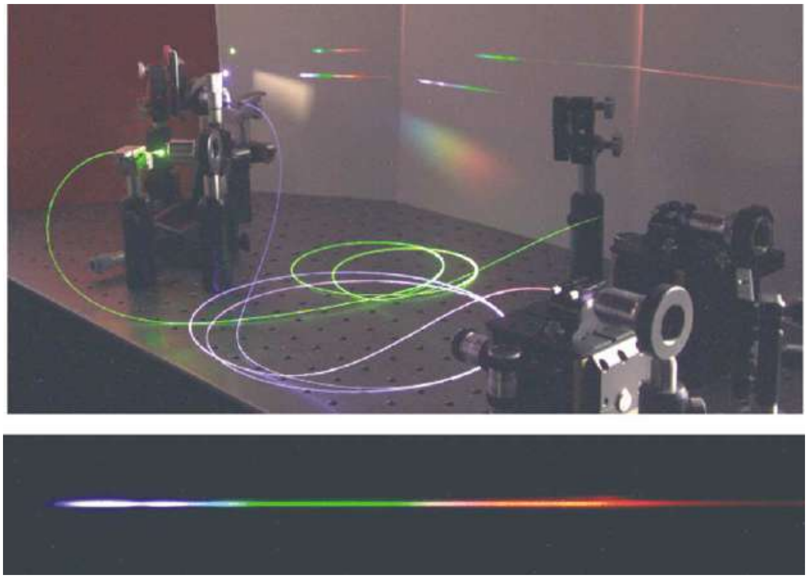

Figure 9 Photograph of fiber generating the supercontinuum. The bottom photo shows the continuum after passing through a prism.

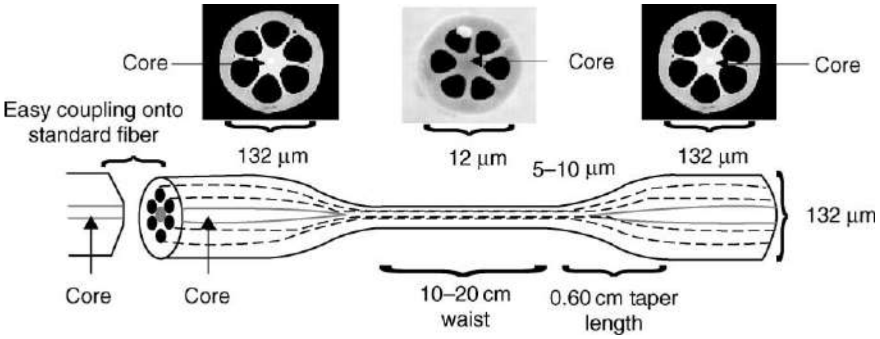
Core
Core
Core
Easy coupling onto standard fiber
132 μm
12 μm
5–10 μm
132 μm
132 μm
Core
Core
10–20 cm waist
0.60 cm taper length

Figure 10 Schematic drawing of the tapered microstructure fiber (Reprinted with permission from Chandalia JK, Eggleton BJ, Windeler RS, Kosinski SG, Liu X and Xu C (2001) Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13(1) (©2001 IEEE)).

cladding is surrounded by a layer of air holes and then a protective outer silica cladding for strength. Since the core of the TMF is very similar to a traditional single-mode fiber core, the splice loss is low. Splice losses below 0.075 dB are typical.

Properties similar to the HDCMF are obtained by adiabatically tapering the TMF while maintaining the cross-sectional profile of the fiber. In the taper region, the fundamental mode is no longer guided by the germanium core and evolves into the fundamental mode of the silica region, where it is confined by the ring of air holes. The taper waist is typically ten to twenty centimeters long and has properties identical to the HDCMF. When the fiber is adiabatically expanded back to its original size, the light is guided back into the standard diameter germanium core, making it easy to splice to traditional fiber with low loss.

The same effect can be obtained by stretching a traditional fiber from 125 microns down to 2 microns. However, there are several disadvantages compared to the tapered microstructure fiber. The taper region is much longer in the traditional fiber; the optimal diameter of the silica core is harder to maintain; the taper region must remain clean and uncoated; and the taper region is much weaker.

Air-clad fiber

Air-clad fibers consist of a doped core surrounded by a silica inner cladding. The inner cladding is surrounded by a ring of air, which in turn is surrounded by an outer silica cladding for strength (Figure 2). Thin silica webs connect the inner and outer claddings to hold the inner fiber region in place. The region consisting of air and thin silica webs creates an optical barrier, preventing most of the light from escaping the inner cladding. Such fibers are referred to as double clad fiber.

In traditional double-clad fibers, a material of lower index (typically another layer of glass or a layer of polymer) immediately surrounds the inner silica cladding. A key optical parameter describing the light guiding properties of the inner cladding is the numerical aperture (NA). The NA is defined as the sine of the largest (acceptance) angle of light that will be guided in the inner cladding. The upper NA values of traditional double-clad fibers (~0.22 and ~0.45 for glass and polymer outer claddings, respectively) are limited by the refractive indices of the available cladding materials.

For the air-clad fiber, the NA is determined by the ability of light to escape from the inner cladding through the silica webs. The NA of the air-clad fiber can be calculated by modeling the web as an infinitely long slab waveguide. As seen in Figure 11, the NA increases as the webs become thinner or the wavelength of light becomes longer. To obtain NA similar to what is achieved by coating a traditional fiber with a low index polymer (NA ~ 0.45), the web thickness must be about equal to the wavelength of the light. The NA of the air-clad fiber increases dramatically as the web thickness becomes less than half the wavelength of the guided light. Numerical apertures above 0.90 have been experimentally demonstrated.

The large NA values achievable with air-clad fibers are an important advantage for advanced double-clad

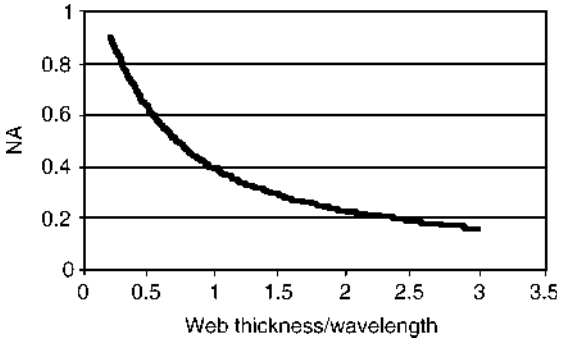

Figure 11 Predicted NA as a function of the web thickness divided by the wavelength using an infinite slab model.

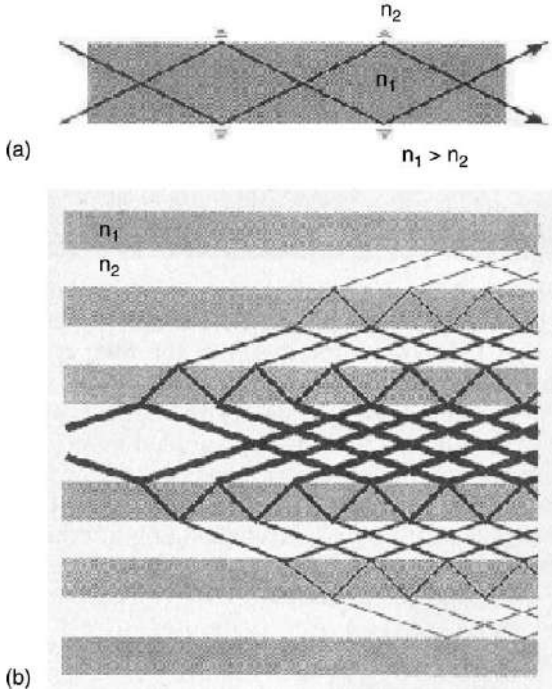

Figure 12 Schematic of the mechanism of (a) index guided light and (b) band gap guided light (Reprinted with permission from Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* 285. Copyright 1999 American Association for the Advancement of Science).

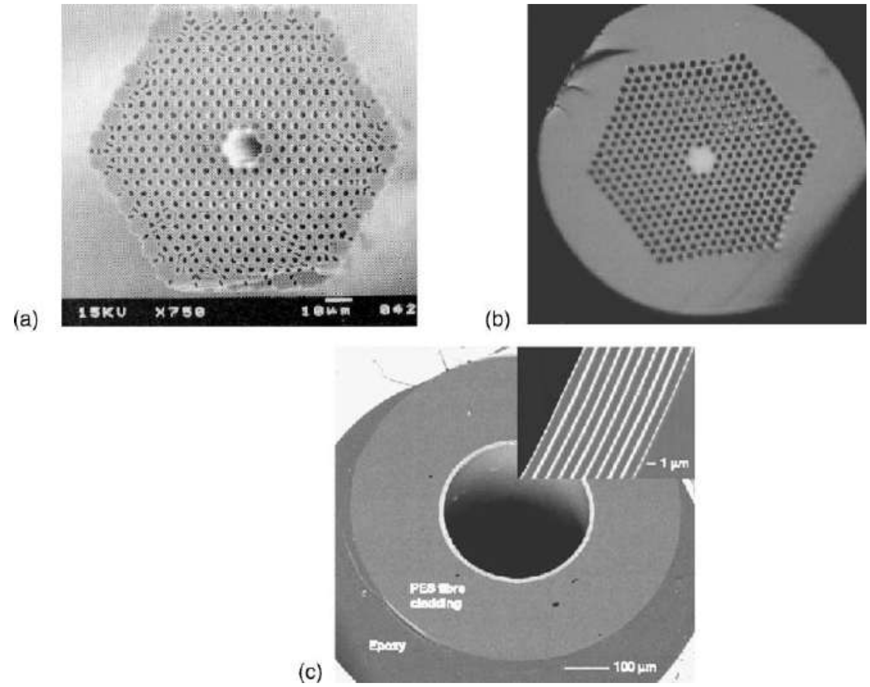

Figure 13 Photographs and drawing of three types of bandgap fibers: (a) triangular array cladding with an air core (Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* vol. 285), (b) tunable with a silica core, (c) concentric rings with an air core (Reprinted with permission from Temelkuran *et al.* *Nature* 420. Copyright 1999 American Association for the Advancement of Science).

amplifiers and lasers. A double-clad amplifier or laser (also known as cladding pumped amplifier or laser) contains a double-clad fiber with a rare earth doped core. High-power pump light, launched into and guided within the inner cladding, excites the rare earth ions as the pump light crosses through the core. The excited ions, which emit light with the same wavelength and phase as the signal, amplify the signal. The fiber's efficiency and the rare earth inversion level increase with the amount of pump light absorbed by the core. The light absorbed by the core, in turn, increases with the core-to-inner cladding area ratio and the pump light intensity in the inner cladding.

Traditional fibers have a practical limit to their core-to-cladding ratio. The maximum core diameter is limited by bend loss. And, while ultimately the minimum fiber diameter is limited by draw capabilities, in practice it is limited by the ability to couple pump light into a fiber having a limited inner-cladding NA. For a pump source of given brightness, the product of beam diameter and NA cannot be increased. Such light can be successfully coupled into a larger inner cladding with smaller NA or a smaller inner cladding with larger NA.

Air-clad fibers have other advantages over traditional double-clad fibers. The outer glass cladding in air-clad fibers is optically inactive. This allows the inner cladding diameter to be decoupled from the physical fiber diameter, eliminating draw-related constraints on minimum inner cladding size. Since only the inner cladding is optically active, its size can be optimized (made smaller) while keeping the total fiber diameter at the standard 125 microns. In addition, the pump light does not interact with the polymer coating, which can be important if the polymer properties are affected by the surroundings.

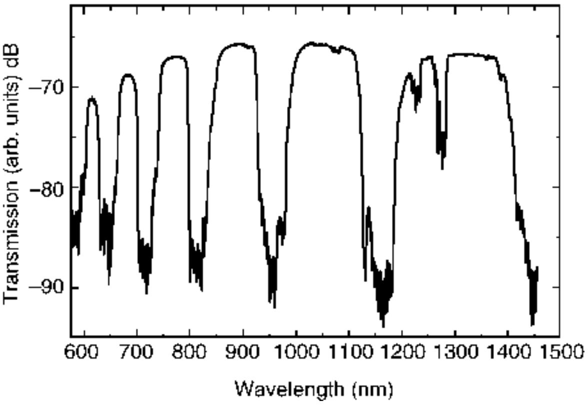

Figure 14 Spectrum of a typical bandgap fiber.

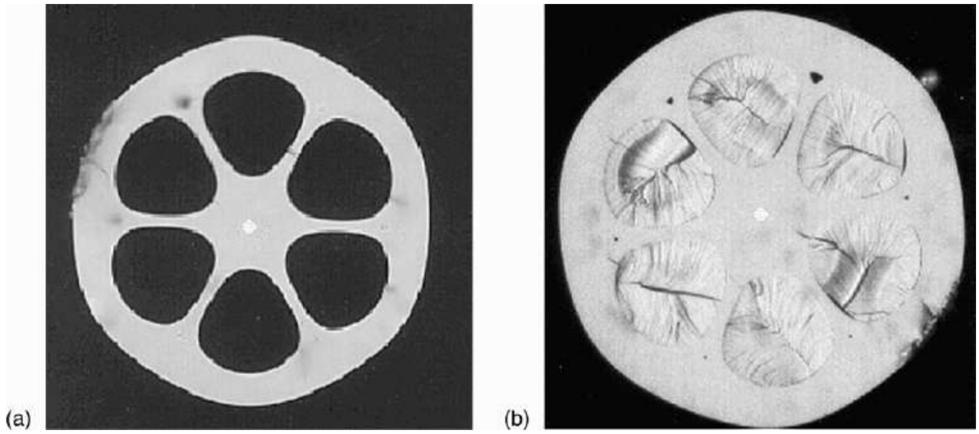

Figure 15 Holes in the microstructure fibers are filled with controllable materials to affect the fiber properties actively. Photos show (a) original fiber and (b) fiber after polymer is inserted and cured in the microstructure fiber.

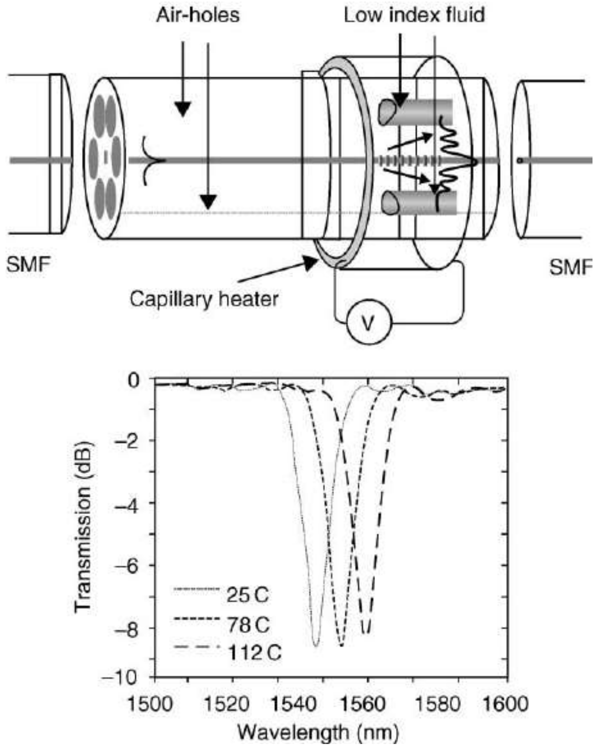

Figure 16 Schematic of a tunable device in which low index liquid is positioned over the grating. The plot shows the shift in the spectrum as a function of temperature (Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184, with permission from Elsevier).

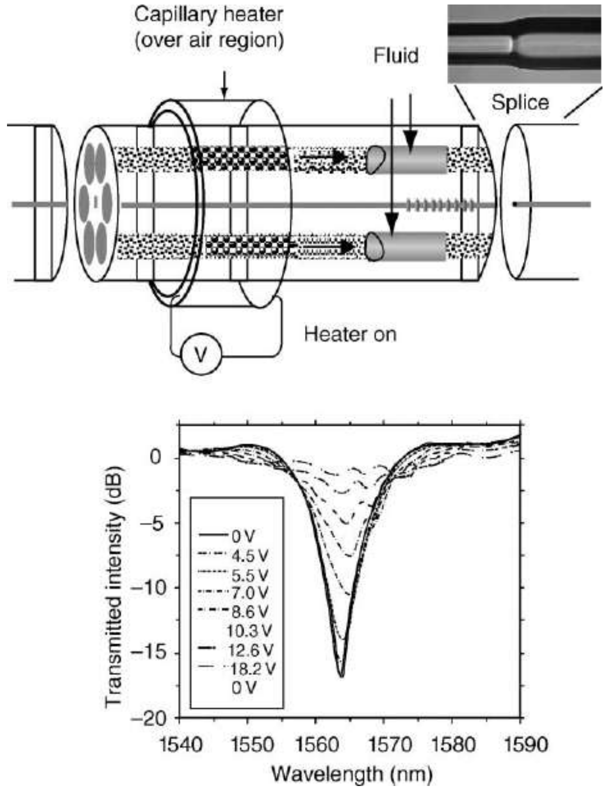

Figure 17 Schematic of a tunable device in which a high index liquid plug is moved over the grating. The plot shows the intensity of the spectrum as a function of temperature (Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber *Optics Communications* 204: 179–184, with permission from Elsevier).

Bandgap Guided Fibers

Bandgap fibers guide light in the core (also referred to as a defect) by confining the light through constructive interference due to Bragg scattering (Figure 12). Unlike traditional fibers, this mechanism allows light to propagate in a core that has an index lower than the cladding. Bandgap fibers with an air core could theoretically be very low loss, propagate high powers, have a large effective area core, and exhibit very low nonlinearities. In addition, the fibers can be used in novel devices by filling the core with special gasses or liquids for nonlinear processes.

Bandgap fibers can be generalized into three types (Figure 13). The first type of fiber has an air core surrounded by a cladding that consists of a periodic array of air holes in silica. Fibers of this type have been designed to guide a single mode and hold the greatest promise of guiding light over long distances with very low loss. The second type of bandgap fiber consists of a silica core and a silica cladding with a periodic array of holes which are filled with a high-index, tunable liquid. This design allows the bandgap properties to be tuned and is of interest for use in devices. The third fiber type consists of rings of alternating high- and low-index dielectric material with a center air core. They typically have a large, multimode core which is ideal for sending high powers. In addition, very little light is in the dielectric materials so the fiber can be designed to guide any wavelength with relatively little loss.

The spectrum of a bandgap fiber is quite different than an index guided fiber (Figure 14). Frequencies that lie within the bandgap cannot propagate in the cladding and are confined to the defect in the lattice (the core). Bandgap fibers with ideal geometry are predicted to have losses over an order of magnitude lower than can be obtained in an ideal traditional fiber. Although the loss of bandgap fibers has not been demonstrated to be less than traditional fiber at telecommunication wavelengths, bandgap fibers have been shown to have much lower loss than traditional fibers at other wavelengths.

Tunable Microstructure Fiber Devices

Unique tunable devices are made by filling the air holes of microstructure fibers with materials whose properties can be controlled actively (Figure 15). The materials can be positioned in the core or cladding to affect the corresponding modes. Most commonly, the index of refraction is changed using temperature sensitive liquids or polymers, although

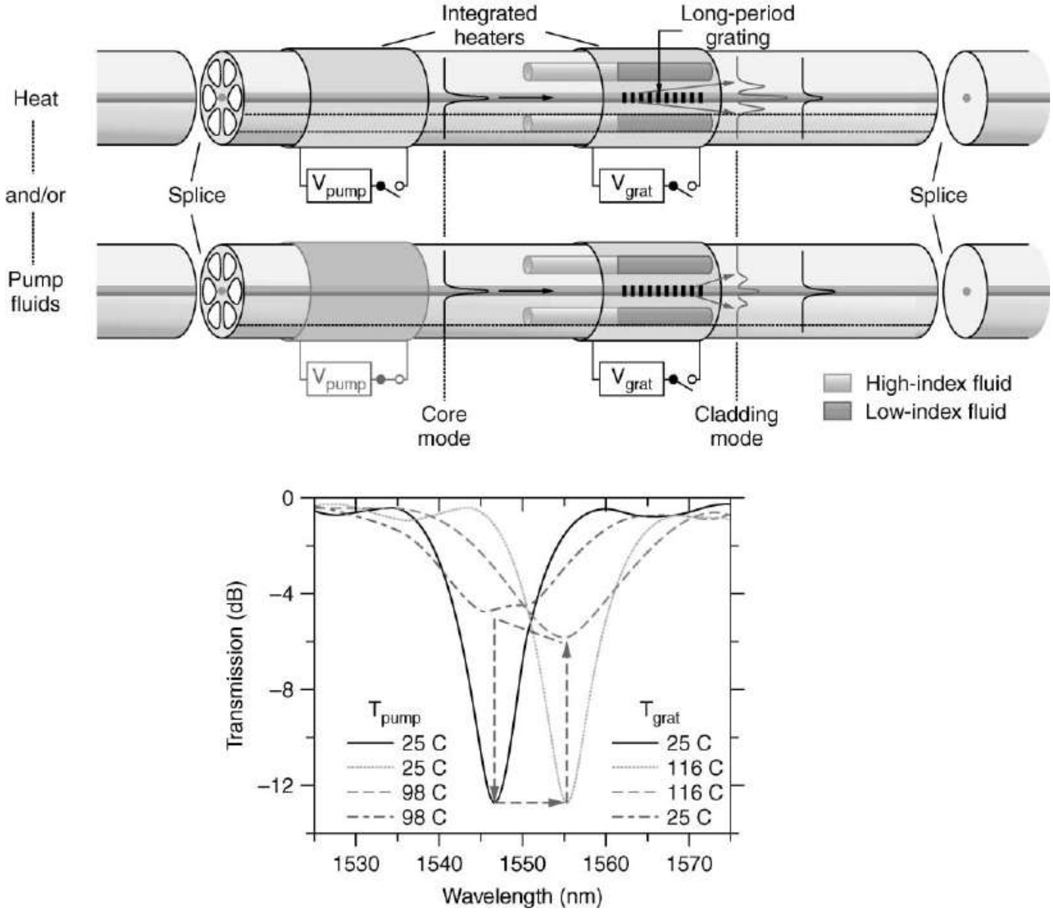

Figure 18 Schematic of a tunable device using both high and low index liquids with two heaters. The plot shows that the position and strength of the spectrum are adjusted independently (Reproduced with permission from Mach P, Dolinski M, Baldwin KW, Rogers JA, Kerbage C, Windeler RS and Eggleton BJ (2002) Tunable microfluidic optical fiber. *Applied Physics Letters* 80(23). Copyright (2002) by the American Physical Society).

materials that respond to electric or magnetic fields enable a quicker response time. These material-filled active fibers can be used to fabricate devices like tunable filters, switches, broadband attenuators, and fibers with tunable birefringence. Below, a few examples of tunable filters are presented in which the device properties depend on the type and position of the liquid inserted in the fiber holes.

A microstructure tunable filter works by controlling the spectral position, shape, or strength of a longitudinal long-period grating written in the fiber. A grating is written in the fiber before filling the holes with tunable liquid. The position of the peak absorption wavelength is controlled by placing a liquid, with an index close to silica and strongly temperature dependent, in air holes in the cladding region over the grating (Figure 16). The cladding index is tuned by varying the temperature of the liquid, using a small thin film heater located over the liquid. The position of the peak absorption wavelength moves as the cladding index is changed.

The strength or depth of the filter is controlled by inserting a moveable plug of high-index fluid in a sealed air section of a microstructure fiber (Figure 17). The strength of the grating is determined by the fraction of the grating surrounded by high index fluid. The position of the liquid plug relative to the grating can be finely adjusted with a thin film heater located over the air region adjacent to the material plug. As heat is applied to the air, the air expands and pushes the liquid plug over the grating, decreasing the coupling of the fundamental mode to higher order cladding modes. Figure 18 shows an example in which the position and strength of a filter are varied independently using two heaters and two adjacent materials.

See also

Photonic Crystals: Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22(13).

Bjarklev A, Broeng J and Bjarklev AS, *Photonic Crystal Fibers*, Boston: Kluwer Academic Publishers.

Broeng J, Mogilevstev D, Bjarklev A and Barkou SE (1999) Photonic crystal fibers: A new class of optical waveguides. *Optical Fiber Technology*, vol. 5.

Chandalia JK, Eggleton BJ, Windeler RS, Kosinski SG, Liu X and Xu C (2001) Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13(1).

Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* 285.

Eggleton BJ, Westbrook PS, White CA, Kerbage C, Windeler RS and Burdge GL (2000) Cladding-mode resonance in air-silica microstructure optical fiber. *Journal Lightwave Technology* 18(8).

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals*. New Jersey: Princeton University Press.

Johnson SG, Ibanesc M, Skorobogatiy M, Weisberg M, Engeness O, Soljacic M, Jacobs SA, Joannopoubs JD and Fink Y (2001) Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Optics Express* 9(13).

Kerbage C, Hale A, Yablon A, Windeler RS and Eggleton BJ (2001) Integrated all fiber variable attenuator based on hybrid microstructure fiber. *Applied Physics Letters* 79(19).

Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184.

Knight JC, Birks TA, Russell PJ and Atkin DM (1996) All-silica single-mode optical fiber with photonic crystal cladding. *Optics Letters* 21(19).

Knight JC, Birks TA and Russell PJ (2001) 'Holey' silica fibers. In: Markel VA and George TF (eds). *Optics of nanostructured materials*. New York: John Wiley & Sons.

Mach P, Dolinski M, Baldwin KW, Rogers JA, Kerbage C, Windeler RS and Eggleton BJ (2002) Tunable microfluidic optical fiber. *Applied Physics Letters* vol. 80.

Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25(1).

Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

Omnidirectional Surfaces and Fibers

S Hart, G Benoit and Y Fink, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Omnidirectional Reflecting Mirrors

Mirrors are probably the most prevalent of optical devices. Known to the ancients and used by them as objects of worship and beauty, mirrors are currently employed for imaging, solar energy collection, and also in laser cavities. Their intriguing optical properties have captured the imagination of scientists as well as artists and writers.

One can distinguish between two types of mirrors, the age-old metallic, and the more recent multilayer dielectric. Metallic mirrors reflect light over a broad range of frequencies incident from arbitrary angles (i.e., omnidirectional reflectance). However, at infrared and optical frequencies a few percent of the incident power is typically lost due to absorption.

Multilayer dielectric mirrors are used primarily to reflect a narrow range of frequencies incident from a particular angle or particular angular range. Unlike their metallic counterparts, dielectric reflectors can be extremely low loss. The ability to reflect light of arbitrary angle of incidence for all-dielectric structures has been associated with the existence of a complete photonic bandgap, which can exist only in a system with a dielectric function that is periodic along three orthogonal directions. In fact, a sufficient condition for the achievement of omnidirectional reflection in a periodic system with an interface, is the existence of an overlapping bandgap regime in phase space above the light cone of the ambient media.

Consider a system that is constructed of alternating dielectric layers coupled to a homogeneous medium – characterized by n_0 (such as air with $n_0 = 1$), at the interfaces. Electromagnetic waves are incident upon the multilayer film from the homogeneous medium. While such a system was analyzed extensively in the literature the possibility of omnidirectional reflectivity was not recognized until recently.

The generic system is described by the index of refraction profile in Figure 1, where h_1 and h_2 are the layer thickness, and n_1 and n_2 are the indices of refraction of the respective layers. The incident wave has a wavevector $\vec{k} = k_x\hat{e}_x + k_y\hat{e}_y$, and frequency of $\omega = c|k|$. The wavevector together with the normal to the periodic structure defines a mirror plane of symmetry which allows us to distinguish between two independent electromagnetic modes: transverse electric (TE) modes and transverse magnetic (TM) modes. For the TE mode the electric field is perpendicular to the plane, as is the magnetic field for the TM mode.

General features of the transport properties of the finite structure can be understood when the properties of the infinite structure are elucidated. In a structure with an infinite number of layers, translational symmetry along the direction perpendicular to the layers leads to Bloch wave solutions of the form:

$$u_K(x, y) = E_K(y)\, e^{iKy}\, e^{ik_x x} \qquad [1]$$

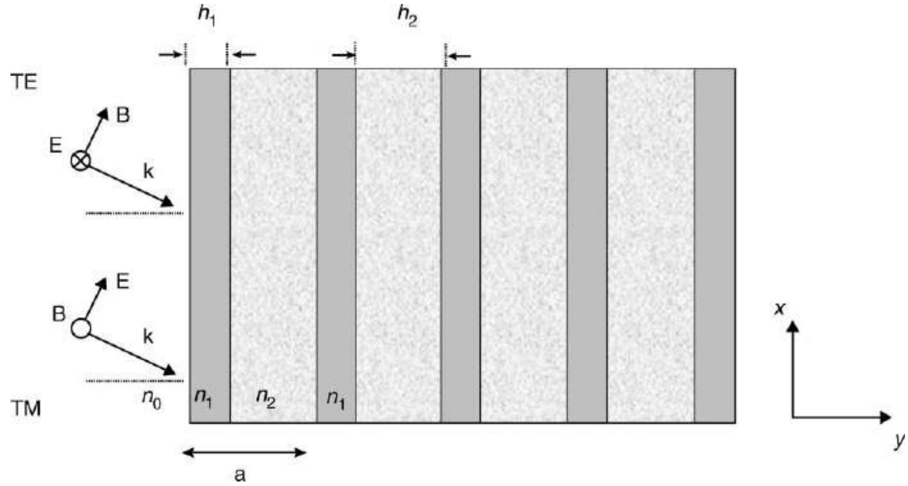

Figure 1 Schematic of the multilayer system showing the layer parameters (n_α, h_α – index of refraction and thickness of layer α), the incident wavevector $\vec{k}$ and the electromagnetic mode convention.

where $E_K(y)$ is periodic, with a period of length a, and K is the Bloch wave number. These waves represent solutions to an eigenvalue problem and are completely and uniquely defined by the specification of K, k_x and ω.

Solutions can be propagating or evanescent, corresponding to real or imaginary Bloch wave numbers, respectively. It is convenient to display the solutions of the infinite structure by projecting the $\omega(K, k_x)$ function onto the $\omega - k_x$ plane; Figure 2a,b are examples of such projected structures.

The gray background areas highlight phase space where K is strictly real, that is, regions of propagating states, while the white areas represent regions containing evanescent states. The shape of the projected bandstructures for the multilayer film can be understood intuitively. At $k_x = 0$, the bandgap for waves traveling normal to the layers is recovered. For $k_x > 0$, the bands curve upward in frequency. As $k_x \rightarrow \infty$, the modes become largely confined to the slabs with the high index of refraction and do not couple between layers (and are therefore independent of k_x).

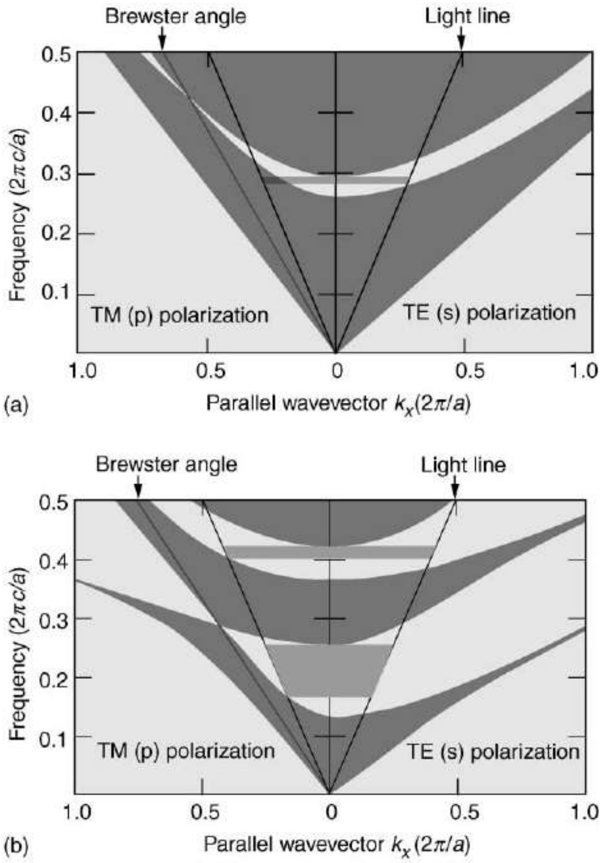

Figure 2 (a) Projected bandstructure of a multilayer film with the light line and Brewster line, exhibiting a reflectivity range of limited angular acceptance with ($n_1 = 2.2$ and $n_2 = 1.7$ and a thickness ratio of $h_2/h_1 = 2.2/1.7$). (b) Projected bandstructure of a multilayer film together with the light line and Brewster line, showing an omnidirectional reflectance range at the first and second harmonic (propagating states – dark gray, evanescent states – white, omnidirectional reflectance range – light grey). The film parameters are $n_1 = 4.6$ and $n_2 = 1.6$ with a thickness ratio of $h_2/h_1 = 1.6/0.8$. These parameters are similar to the actual polymer–tellurium film parameters measured in the experiment. Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

In a finite structure, the translational symmetry in the directions parallel to the layers is preserved, hence k_x remains a conserved quantity and can be used to label solutions even though these solutions will no longer be of the Bloch form. The relevance of the band diagram to finite structures is that it allows for the prediction of regions of phase space where waves are evanescently decaying in the multilayer structure.

Since we are primarily interested in waves originating from the homogeneous medium external to the periodic structure, we will focus only on the portion of phase space lying above the light line. Waves originating from the homogeneous medium satisfy the condition $\omega \geq ck_x/n_0$, where n_0 is the refractive index of the homogeneous medium, and therefore they must reside above the light line. States of the homogeneous medium with $k_x = 0$ are normal incident, and those lying on the $\omega = ck_x/n_0$ line with $k_y = 0$ are incident at an angle of 90°.

The states in Figure 2a, that are lying in the restricted phase space defined by the light line and that have a (ω, k_x) corresponding to the propagating solutions (gray areas) of the crystal can propagate in both the homogeneous medium and in the structure. These waves will partially or entirely transmit through the film. Those with (ω, k_x) in the evanescent regions (white areas) can propagate in the homogeneous medium but will decay in the crystal – waves corresponding to this portion of phase space will be reflected off the structure.

The multilayer system leading to Figure 2a represents a structure with a limited reflectivity cone, since for any frequency one can always find a k_x vector for which a wave at that frequency can propagate in the crystal – and hence transmit through the film. The necessary and sufficient criterion for omnidirectional reflectivity at a given frequency is that there exists no transmitting state of the structure inside the light cone – this criterion is satisfied by frequency ranges marked in light gray in Figure 2b. In fact, the system leading to Figure 2b exhibits two omnidirectional reflectivity ranges.

The omnidirectional range is defined from above by the normal incidence bandedge $\omega_h(k_y = \pi/a, k_x = 0)$ (Figure 2b) and below by the intersection of the top of the TM allowed bandedge with the light line $\omega_l(k_y = \pi/a, k_x = \omega_l/c)$ (Figure 2b).

A dimensionless parameter used to quantify the extent of the omnidirectional range is the range to mid-range ratio defined as $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$.

Figure 3 is a plot of this ratio as a function of n_2/n_1 and n_1/n_0, where ω_h and ω_l are determined by solutions of eqn [2] with quarter wave layer thickness. The contours in this figure represent various equi-omnidirectional ranges for different material index parameters and could be useful for design purposes.

At normal incidence, there is no distinction between TM and TE modes. At increasingly oblique angles, the gap of the TE mode increases, whereas the gap of the TM mode decreases. In addition, the center of the gap shifts to higher frequencies. Therefore, the criterion for the existence of omnidirectional reflectivity can be restated as the occurrence of a frequency overlap between the gap at normal incidence and the gap of the TM mode at 90°. Analytical expressions for the range to mid-range ratio can be obtained by setting:

$$\begin{aligned}\omega_h &= \frac{2c}{h_2 n_2 + h_1 n_1}\cos^{-1}\left(-\left|\frac{n_1 - n_2}{n_1 + n_2}\right|\right)\\ \omega_l &= \frac{2c}{h_2\sqrt{n_2^2 - 1} + h_1\sqrt{n_1^2 - 1}}\\ &\quad\times\cos^{-1}\left(\left|\frac{n_1^2\sqrt{n_2^2 - 1} - n_2^2\sqrt{n_1^2 - 1}}{n_1^2\sqrt{n_2^2 - 1} + n_2^2\sqrt{n_1^2 - 1}}\right|\right)\end{aligned} \qquad [2]$$

Moreover, the maximum range width is attained for thickness values that are not equal to the quarter wave stack though the increase in bandwidth gained by deviating from the quarter wave stack is typically only a few percent.

In general, the TM mode defines the lower frequency edge of the omnidirectional range; an example can be seen in Figure 2b for a particular choice of the indices of refraction. This can be proven by showing that:

$$\left.\frac{\partial\omega}{\partial k_y}\right|_{\mathrm{TM}} \geq \left.\frac{\partial\omega}{\partial k_y}\right|_{\mathrm{TE}} \qquad [3]$$

in the region that resides inside the light line. The physical reason for eqn [3] lies in the vectorial nature of the electric field. In the upper portion of the first band the electric field concentrates its energy in the high dielectric regions. Away from normal incidence, the electric field in the TM mode has a component in the direction of periodicity, this component forces a larger portion of the electric field into the low dielectric regions. The group velocity of the TM mode is therefore enhanced. In contrast, the electric field of the TE mode is always perpendicular to the direction of periodicity and can concentrate its energy primarily in the high dielectric region. While the omnidirectional reflection criteria can be used to confine light in various geometries, it is the cylindrical fiber configuration that appears to present significant application opportunities.

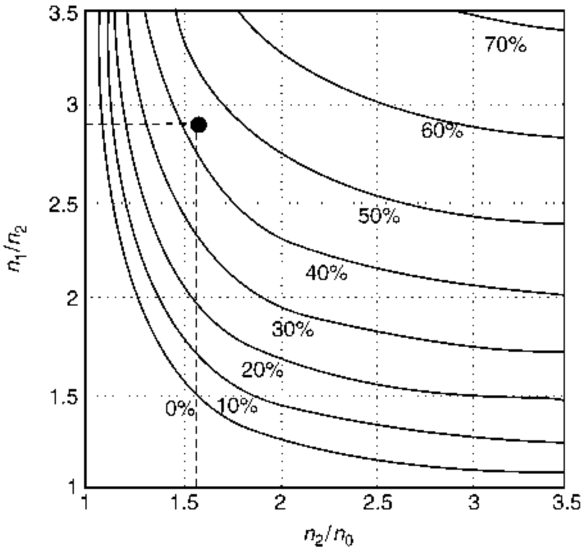

Figure 3 The range to mid-range ratio $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$, for the fundamental frequency range of omnidirectional reflection, plotted as contours. Here, the layers were set to quarter wave thickness and $n_1 > n_2$. The ratio for our materials is approximately 45%. ($n_1/n_2 = 2.875$, $n_2/n_0 = 1.6$) is located at the intersection of the dashed lines (black dot). Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

Mesostructured Fibers for External Reflection Applications

Polymer fibers are ubiquitous in applications such as textile fabrics, due to their excellent mechanical properties and the availability of low-cost, high-volume processing techniques; however, the control over their optical properties has so far remained relatively limited. Conversely, dielectric mirrors are used to precisely control and manipulate light in high performance optical applications, but the fabrication of these typically fragile mirrors has been mostly restricted to planar geometries and typically involves multiplicity of deposition sequences in a high vacuum thin film deposition system.

Fabrication Approach

The fabrication of extended lengths of omnidirectionally reflecting fibers with large bandgaps and

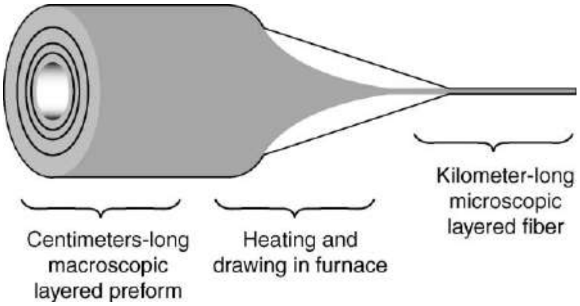
Centimeters-long macroscopic layered preform
Heating and drawing in furnace
Kilometer-long microscopic layered fiber

Figure 4 Conceptual preform based fabrication process for meso-structured fibers.

high layer counts poses considerable challenges. To illustrate the nature of this formidable task, one needs to consider the necessity of maintaining the uniformity of sub-100 nm layer thicknesses over kilometer-length scales; creating continuous layers with an aspect ratio of $\sim 10^{10}$ in a single process! To meet this and other challenges associated with the fabrication of mesostructured fibers, we have developed a preform-based fabrication approach (Figure 4). A scaled-up version of the final fiber, called a preform, is fabricated which shares the geometry and materials of the final fiber but exhibits macroscopic lateral features. The preform is heated up and drawn under tension into the fiber using a simple cylindrical furnace. The macroscopic layers are reduced in the process to microscopic dimensions while maintaining the overall geometry and symmetry of the original preform. Conservation of mass determines the final length of the fiber – typically this length is equal to the lateral reduction factor squared multiplied by the length of the preform. The nature of the process and requirements on the fiber's optical properties lead to the definition of materials selection rules. First, in order to achieve omnidirectional reflectivity, one needs to identify two solid materials exhibiting an index contrast given by the plot in Figure 3. Second, to enable the codrawing of the two dissimilar materials both will need to have viscosities that are lower than $\sim 10^8$ poise at the drawing temperature. In order to maintain high draw speeds, the majority component needs to be amorphous and the adhesion between these two materials needs to be sufficient to prevent delamination. Finally, their thermal expansion coefficients need to be close or alternatively at least one of the materials needs to be capable of relieving the stress due to CTE mismatch.

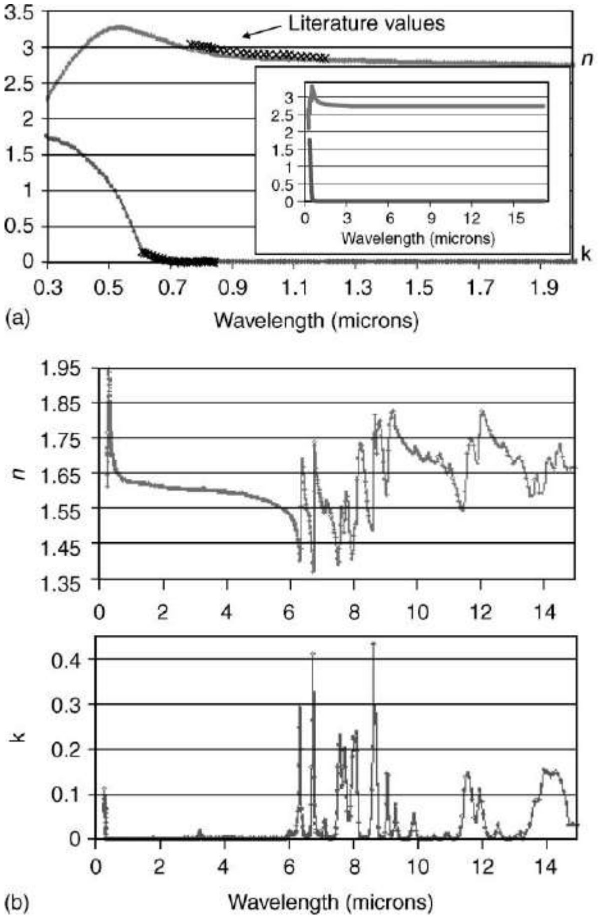

Figure 5 (a) Real part (*n* – light gray) and imaginary part (*k* – dark gray) of the refractive index of annealed As_2Se_3. The black crosses correspond to literature values. (b) Real part (*n* – upper) and imaginary part (*k* – lower) of the refractive index of PES.

Materials Selection Criteria

Pairs of materials that are compatible with the process and fiber property requirements have been identified, including: high glass transition temperature (T_g), thermoplastic polymers such as poly(ether-sulfone) (PES), poly (ether-imide) (PEI) and members of the chalcogenide glass family, such as arsenic triselenide (As_2Se_3) or arsenic trisulfide. Pairs of these materials will have substantially different refractive indices, as shown in Figure 5, which is a measurement of the real and imaginary indices of refraction of PES and As_2Se_3, obtained using a broadband spectroscopic ellipsometer (SOPRA GES5). They nevertheless exhibit similar thermo-mechanical properties within a certain thermal processing window.

Adhesion and extensional viscosity in the fluid state are difficult to measure in general, and the measurement of high-temperature surface tension is quite involved. Thus, limited data on these properties are available and it was necessary to empirically identify materials that could be used to draw out mirror-fibers. Various high-index chalcogenide (S, Se, and Te containing) glasses and low-index polymers were identified as potential candidates based on their optical properties and overlapping thermal softening regimes. Adhesion and viscosity matching were tested by thermal evaporation of a chalcogenide glass layer on top of a polymer film or rod and elongation of the coated substrate at elevated temperatures. The choice of a high-temperature polymer, PES, and a simple chalcogenide glass, As_2Se_3, resulted in excellent thermal co-deformation without film cracking or delamination. Approximate matching of extensional viscosity in this manner was also demonstrated using As_2Se_3 and PEI. The properties, processing, and applications of chalcogenide glasses have been explored extensively elsewhere. One advantage in choosing As_2Se_3 for this application is that not only is it a stable glass, but it is a stoichiometric compound that can be readily deposited in thin films through thermal evaporation or sputtering without dissociation. Additionally, As_2Se_3 is transparent to IR radiation from approximately 0.8 to 17 μm as shown in the ellipsometric data (Figure 5a), and has a refractive index of ~2.8 in the mid-IR. PES is a high-performance, dimensionally stable thermoplastic with a refractive index of ~1.6 and good transparency to EM waves in a range extending from the visible regime into the mid-IR (~5 μm), as shown in Figure 5b.

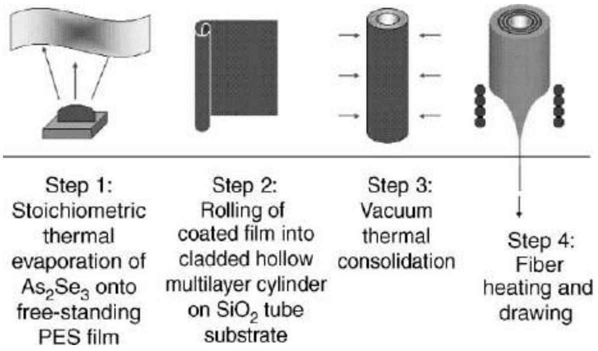
Step 1: Stoichiometric thermal evaporation of As_2Se_3 onto free-standing PES film
Step 2: Rolling of coated film into cladded hollow multilayer cylinder on SiO_2 tube substrate
Step 3: Vacuum thermal consolidation
Step 4: Fiber heating and drawing

Figure 6 Multilayer preform fabrication sequence.

Preform Construction Process and Fiber Draw

The selected materials were used to construct a multilayer preform rod, which essentially is a macroscale version of the final fiber. In order to fabricate the dielectric mirror fiber preform, an As_2Se_3 film was deposited through thermal evaporation on either side of a free-standing PES film which was then rolled on top of a PES tube substrate, forming a structure having 21 alternating layers of PES and As_2Se_3, as shown in Figure 6.

The resulting multilayer fiber preform was subsequently drawn down using an optical fiber draw tower into hundreds of meters of multilayer fiber with a precisely controlled submicron layer thickness, creating a photonic bandgap in the mid-IR. Fibers of outer diameters varying from 175–500 μm with a typical standard deviation of 10 μm from target, were drawn from the same preform to demonstrate adjustment of the reflectivity spectra through thermal deformation. The spectral position of the photonic bandgap was controlled by the optical monitoring of the outer diameter (OD) of the fiber during draw, which was later verified by reflectivity measurements on single and multiple fibers of different diameters. Scanning electron micrographs (SEMs) of the cross-section of these fibers are depicted in Figure 7.

Bandstructure for Multilayer Fibers for External Reflection Applications

In theoretically predicting the spectral response of these fibers, it is helpful to calculate the photonic

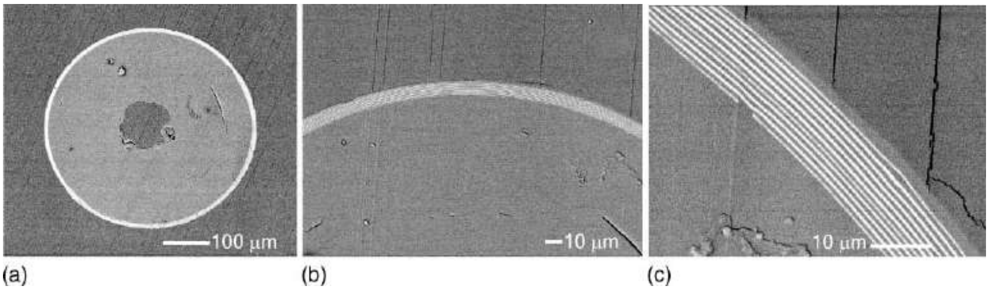

Figure 7 SEM micrographs of 400 μm OD fiber cross-section. The entire fiber is embedded in epoxy. (a) shows the entire fiber cross-section, with mirror structure surrounding the PES core; (b) demonstrates that the majority of the fiber exterior is free of significant defects and that the mirror structure adheres well to the fiber substrate; and (c) reveals the ordering and adhesion within the alternating layers of As_2Se_3 (bright layers) and PES. Stresses developed during sectioning caused some cracks in the mounting epoxy that are deflected at the fiber interface. Fibers from this batch were used in the reflectivity measurements recorded below in Figure 9a. Reproduced with permission from Hart SD, Maskaly GR, Temelkuran B, Prideaux PH, Joannopoulos JD and Fink Y (2002) External reflection from omnidirectional dielectric mirror fibers. *Science* 296: 510–513. Copyright 2002 American Association for the Advancement of Science.

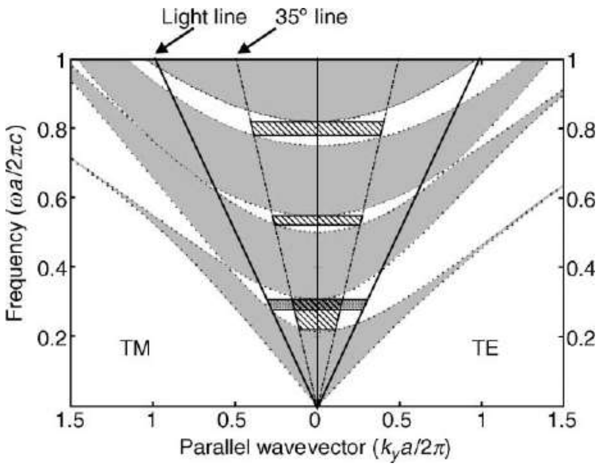

Figure 8 Photonic band diagram for a one-dimensional photonic crystal having a periodic refractive index alternating between 2.8 and 1.55. Gray regions represent propagating modes within the structure, while white regions represent evanescent modes. Hatched regions represent photonic bandgaps where high reflectivity can be expected for external EM waves over an angular range extending from normal to 35° incidence. The lower dark shaded trapezoid represents a region of external omnidirectional reflection.

bandstructure that corresponds to an infinite one-dimensional photonic crystal. This allows for the analysis of propagating and evanescent modes in the structure, corresponding to real or imaginary Bloch wave number solutions.

The electric or magnetic field vector is parallel to the mirror layer interfaces for the TE and TM polarized modes, respectively. The parallel wavevector (k_y) is the component of the incident electromagnetic (EM) wavevector that is parallel to the layer interfaces. The phase space accessible from an external ambient medium is contained between the light lines (defined by the glancing-angle condition $\omega = ck_y/n_0$), and the modes between the 35° lines correspond to those sampled experimentally. Axes are normalized to the thickness a of one mirror bilayer (a period consisting of one high- and one low-index layer). Figure 8 depicts the photonic band diagram for an infinite structure having similar periodicity and refractive indices to the mirror structures fabricated here. Three photonic bandgaps are present where high reflectivity is expected within the 0–35° angular range, and the fundamental gap contains a region of external omnidirectional reflectivity.

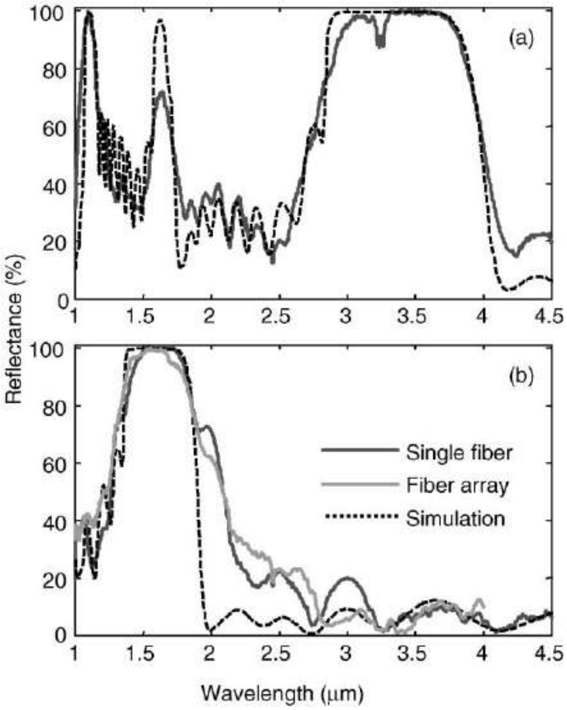

Figure 9 Measured reflectance spectra for 400 µm OD (a) and 200 µm OD (b) dielectric mirror fibers relative to gold-coated fibers of the same diameter. (a) shows a single-fiber reflectivity measurement, while (b) compares single-fiber reflectivity to that measured from a multifiber array. Simulations were performed using the transfer matrix method.

Optical Characterization of 'Mirror Fibers'

Mirror fiber reflectivity was measured from both single fibers and parallel fiber arrays using a Nicolet/SpectraTech NicPlan Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860). The microscope objective (SpectraTech 15 ×, Reflachromat) used to focus on the fibers had a numerical aperture (NA) of 0.58. This results in a detected cone where the angle of reflection with respect to the surface normal of the structure could vary from normal incidence to ~35°, which is determined by the NA of the microscope objective. As a background reference for the reflection measurements, we used gold-coated PES fibers of matching diameters. Dielectric mirror fibers, drawn to 400 µm OD, exhibited a very strong reflection band centered at 3.4 µm wavelength (Figure 9a). Measured reflectivity spectra agree well with planar-mirror transfer matrix method (TMM) simulations, where the reflectivity was averaged across the aforementioned angular range for both polarization modes. Fibers drawn down to 200 µm OD show a similar strong fundamental reflection band centered near 1.7 µm (Figure 9b). This shifting of the primary photonic bandgap clearly illustrates the precise tuning of the reflectivity spectra over wide frequency ranges through thermal deformation processing. Strong optical signatures are measurable from single fibers as small as 200 µm OD. Fiber array measurements, simultaneously sampling reflected light from multiple fibers, agree quite well with single-fiber data (Figure 9b).

These reflectivity results are strongly indicative of uniform layer thickness control, good interlayer adhesion, and low interdiffusion through multiple thermal treatments. This was confirmed by SEM inspection of fiber cross-sections (Figure 7). The layer thicknesses observed ($a = 0.90\ \mu m$ for the 400 μm fibers; $a = 0.45\ \mu m$ for the 200 μm fibers) correspond well to the measured reflectivity spectra. The fibers have a hole in the center, due to the choice of a hollow rod as the preform substrate, which experienced some nonuniform deformation during draw. The rolled-up mirror structure included a double outer layer of PES for mechanical protection, creating a noticeable absorption peak in the reflectivity spectrum at ~3.2 μm (Figure 9a).

A combination of spectral and direct imaging data demonstrates excellent agreement with the photonic band diagram. The measured gap width (range to mid-range ratio) of the fundamental gap for the 400 μm OD fiber is 27%, compared to 29% in the photonic band diagram.

Tunable 'Fabry–Perot' Fibers

The fabrication of fibers surrounded by or lined with alternating layers of materials with a large disparity in their refractive indices, presents interesting opportunities for passive and active optical devices. While a periodic multilayer structure, such as the one reported above, leads to the formation of photonic bandgaps and an associated range of high reflectivity, it is the incorporation of intentional deviations from periodicity, also called 'defects', which allows for the creation of localized electromagnetic modes in the vicinity of the defect. These structures, sometimes called optical cavities, in turn can provide the basis for a large number of interesting passive and active optical devices such as vertical cavity surface emitting lasers (VCSELs), bi-stable switches, tuneable dispersion compensators, tuneable drop filters, etc. Here we report on the fabrication of a fiber surrounded by a Fabry–Perot cavity structure and demonstrate that by application of axial mechanical stress, the spectral position of the resonant Fabry–Perot mode can be reversibly tuned.

Structure and Optical Properties of the Fabry–Perot Fibers

The fabrication technique described above allows for the accurate placement of optical cavities that can encompass the entire or partial fiber circumference. The fibers discussed above are made of As_2Se_3 and PES and have a low index Fabry–Perot cavity (Figure 10).

This structure was achieved by introducing an extra polymer layer in the middle of the periodic multilayer structure of the preform, thus generating a defect mode in the photonic bandgaps of the drawn fibers. The position of the bandgap center is linearly related to the optical thickness of the layers by the Bragg condition. Through accurate outer diameter control afforded by the laser micrometer mounted on the draw tower we have been able to place gaps (Figure 11) at wavelengths ranging from 11 microns to below 1 micron. Such cost-effective tuneable optical filters could lead to applications such as optical

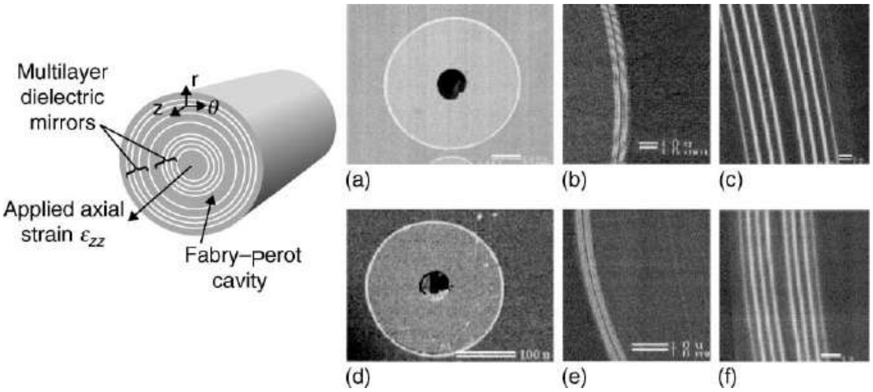
Multilayer dielectric mirrors
r
z
θ
Applied axial strain ε_{zz}
Fabry–perot cavity
(a)
(b)
(c)
(d)
(e)
(f)

Figure 10 Schematic of the structure of dielectric mirror fibers made of As_2Se_3 (light gray) and PES (dark gray) with low index Fabry–Perot cavity. The local cylindrical coordinate system is represented as well as the applied axial strain ε_{zz}. Typical radii are ~150 μm ± 100 μm. Backscattered SEM micrographs of the cross-sections of a 460 micron diameter fiber (a, b, c) and of a 240 micron diameter fiber (d, e, f) embedded in epoxy and microtomed. (a) and (d) show the entire cross-section of the fibers, (b) and (e) demonstrate long-range layer uniformity, and (c) and (f) reveal the ordering and adhesion of the Fabry–Perot cavity structure. Bar scales have been redrawn for clarity. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

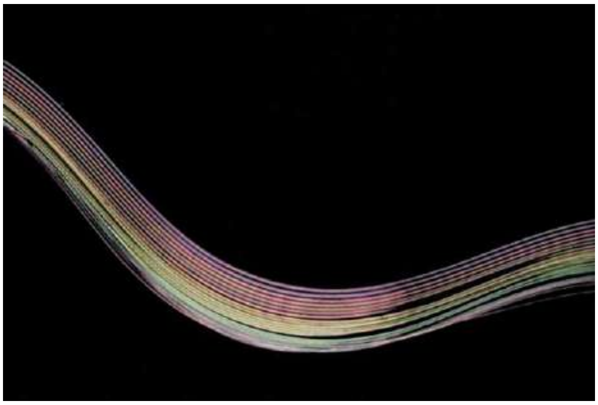

Figure 11 Array of parallel fibers with outer diameter ranging from ~420 μm (bottom) to ~100 μm (top). The colors are due to the narrow 4th order photonic bandgap in the visible.

switches for wavelength-division-multiplexing (WDM) systems and sensors.

The cross-sectional structures of a 240 micron and a 460 micron diameter fiber were observed by an SEM using a backscattered electron detector (Figure 10). The structure is composed of a hollow core polymer rod surrounded by six bilayers of As_2Se_3 and PES, separated in the middle by an extra polymer layer, which forms the Fabry–Perot cavity. An extra polymer layer protects the fiber surface. For the 240 micron (460 micron) diameter fiber, the glass layers are ~135 nm (250 nm) thick, except for the first and the last ones which are half as thick due to the fabrication technique; the polymer layers are ~270 nm (540 nm) thick and the defect layer is ~610 nm (1170 nm) thick.

Reflectivity spectra measurements were performed under a microscope using a (Nicolet SpectraTech NicPlan) Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860) with a lens numerical aperture (NA) corresponding to 30 degrees of angular spread. They exhibited a single-mode Fabry–Perot resonant mode at 1.74 and 3.2 μm for the 240 and 460 micron diameter fiber, respectively (Figure 12).

Because of the range of incident angles, the measured quality factor (Q-factor, defined as its spectral position divided by the full width half maximum (FWHM)) was equal to 31. Using the different thicknesses and the real and imaginary part of the refractive index (n & k) of As_2Se_3 and PES carefully measured with a broadband (300 nm to 16 microns) spectroscopic ellipsometer (Sopra GES-5), the reflectivity spectra of these fibers were computed with the TMM by approximating them as a one-dimensional planar stacking. By averaging the calculated spectra over the accessible incident angles and polarizations, we obtained a very good agreement between the simulation and the measurements.

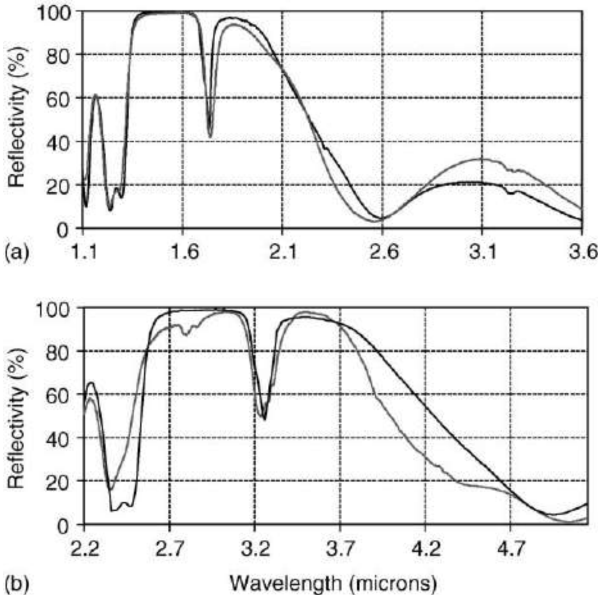

Figure 12 Computed (black lines) and measured (grey lines) reflectivity spectra for the 240 micron (a) and of the 460 micron diameter fibers (b) with Fabry–Perot resonant modes at 1.74 and 3.2 μm, respectively. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

Simulation of the Opto-Mechanical Behavior of the Fabry–Perot Fibers

We focused our analysis on the elastic regime, which ultimately limits the operational range of these fibers. The fiber's Young's modulus (E) can be approximated by modeling this multilayer structure as independent parallel springs under axial (along the z-axis of the fiber) strain assumed to be equal for all the layers, leading to a Young's modulus of 2.64 GPa (using $E_{\mathrm{PES}} = 2.4$ GPa for bulk PES and $E_{\mathrm{As_2Se_3}} = 15$ GPa reported for 1.5 μm thick films).

Neglecting the possible strain-induced refractive index variation of the materials, the normalized shift of the Fabry–Perot resonant mode can be related to the applied axial strain by calculating the radial stresses σ_{rr} (resulting from the difference between the Poisson ratios of the materials – $\nu_{\mathrm{PES}} = 0.45$ and $\nu_{\mathrm{As_2Se_3}} = 0.289$ – and the adhesion condition between the layers) and displacements u_r in each layer under axial strain. Starting from the equilibrium equations:

$$\frac{\partial \sigma_{zz}}{\partial z} = 0 \qquad [4]$$

$$\frac{\partial \sigma_{rr}}{\partial r} + \frac{\sigma_{rr} - \sigma_{\theta\theta}}{r} = 0 \Leftrightarrow \frac{\mathrm{d}^2 u_r}{\mathrm{d}^2 r} + \frac{1}{r}\frac{\mathrm{d}u_r}{\mathrm{d}r} - \frac{u_r}{r^2} = 0 \qquad [5]$$

and using Hooke's law and Lame's equations, we can derive a general expression for u_r and σ_{rr}, with two unknowns A and B per layer:

$$u_r = \frac{A}{r} + Br \qquad [6]$$

$$\sigma_{rr} = -\frac{2GA}{r^2} + 2(\eta + G)B + \lambda\varepsilon_{zz} \qquad [7]$$

where $G = E/[2(1+\nu)]$ is the shear modulus and $\eta = E\nu/[(1+\nu)(1-2\nu)]$ the Lame modulus. The continuity of the displacement and the radial stress at each interface, plus the boundary conditions (σ_{rr} vanishes at free surfaces) can be expressed as a linear system whose unique solution allows us to relate the radial strain in each layer to the applied axial strain as a linear relation $\varepsilon_{rr} = C\varepsilon_{zz}$, where C can be interpreted as an effective Poisson ratio. Finally, taking into account that the Fabry–Perot resonant mode itself is linearly shifted within the bandgap because of the different effective Poisson ratios of the glass and polymer layers, we obtain a linear relation between the normalized shift of the Fabry–Perot resonant mode and the applied axial strain:

$$\frac{\Delta\lambda}{\lambda} = -0.373\varepsilon_{zz} \qquad [8]$$

All the layers (except the outer protective polymer layer) are under tensile radial stress whose maximum (0.22 MPa under 1% axial strain) is located at the interface between the layers and the polymer core where delamination is most likely to occur.

Mechanical Tuning Experiment and Discussion

Measurements were performed on fibers ~30 cm long, which were fixed at one end with epoxy to a load cell (Transducer Techniques MDB-2.5) while the other end was attached with strong tape to a pole mounted on a stepper rotational stage (Newport PR50) (Figure 13). This end of the fiber was also screwed to the pole to further secure it in place.

The diameter of the pole was equal to 2.1 cm, leading to a normalized shift precision below 0.005%. The uncertainty on the Young's modulus, due to the precision of the load cell, was lower than 30 MPa. All the reflectivity spectra were normalized to a background taken with a flat gold mirror. The measurements were realized as far as possible from the fixed ends of the fiber where edge effects are likely to occur. These edge effects result in a reduction of the length of the fiber that deforms uniformly and consequently increase the real strain far from the edges by a factor of 1.14 and 1.15 for the 240 and the 460 micron diameter fiber, respectively (determined experimentally by measuring the position of two reference points on the fiber) compared to the strain calculated from the rotation of the pole. Moreover, to avoid measuring slight variations in the spectral position of the bandgap resulting from outer diameter variations, typically of the order of 4–5 microns over meters of fibers, the measurements were realized at a fixed reference position on the fiber.

By focusing on the fundamental bandgap of the 240 and the 460 micron diameter fiber, we demonstrated the tuning of the Fabry–Perot resonant mode under increasing axial strain (Figure 14).

A drop in the reflectivity of 13% was observed at 1.71 μm (dash line) for the 240 micron diameter fiber when increasing the applied axial strain from 0.23% (light gray) to 1.07% (dark gray). The normalized-shift versus strain curves (Figure 15) appeared to be linear for both fibers up to approximately 0.9% (dash line) with a slope equal to −0.3859 and −0.3843 for the 240 and the 460 micron diameter fiber, respectively, close to the predicted value (−0.373, eqn [8]).

The normalized shift was equal to −0.347% for 0.9% applied axial strain, which seems to be the limit of the elastic regime and corresponds to small applied

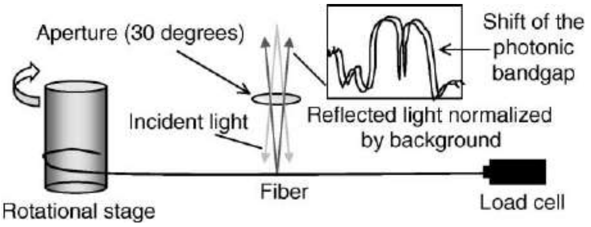
Aperture (30 degrees)
Shift of the photonic bandgap
Incident light
Reflected light normalized by background
Fiber
Rotational stage
Load cell

Figure 13 Experimental setup used for mechanical tuning demonstration. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

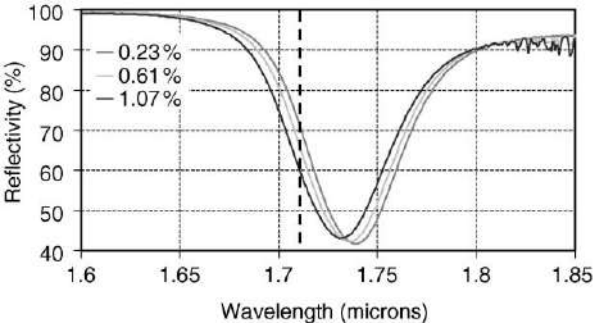

Figure 14 Reflectivity versus wavelength plot showing the shift of the Fabry–Perot resonant mode of the 240 micron diameter fiber for three increasing values of the applied axial strain. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

loads: 76 and 300 g for the 240 and 460 micron diameter fiber, respectively. Under higher strains, the normalized-shift versus strain curves were no longer linear and the measured loads would start decreasing slightly with time, which could be a sign of delamination between the layers and the polymer core, of plastic deformation and/or of relaxation in the layers. The stress–strain curves also exhibit an elastic regime up to approximately 1% strain with a corresponding Young's modulus equal to 2.59 GPa for the 240 μm OD fiber and to 2.39 GPa for the 460 μm OD fiber. These values are slightly lower than the predicted value, possibly due to relaxation effects in the polymer and the glass layers (for As_2Se_3 and PES, $T_g = 175$ and 220 °C, respectively).

Wavelength-Scalable Hollow Optical Fibers with Large Photonic Bandgaps for CO_2 Laser Transmission

Hollow optical transmission fibers offer the potential to circumvent fundamental limitations associated with conventional index guided fibers and thus have been the subject of active research in recent years. Here we report on the materials selection, design, fabrication, and characterization of extended lengths of hollow optical fiber lined with an interior omnidirectional dielectric mirror. These fibers consist of a hollow air core surrounded by multiple alternating submicron-thick layers of a high-refractive-index glass and a low-index polymer, resulting in large infrared photonic bandgaps. These gaps provide strong confinement of optical energy in the hollow fiber core and lead to light guidance in the fundamental and up to fourth-order gaps. We show that the fiber transmission windows can be scaled over a large wavelength range covering at least 0.75 to 10.6 microns. The utility of our approach is further demonstrated by the design and fabrication of tens of meters of hollow photonic bandgap fibers for 10.6 micron radiation transmission. We demonstrate transmission of carbon dioxide (CO_2) laser light with high power-density through more than 4 meters of hollow fiber and measure the losses to be less than 1.0 dB/m at 10.6 microns. This establishes suppression of fiber waveguide losses by orders of magnitude compared to the intrinsic fiber material losses.

Silica optical fibers have been extremely successful in telecommunications applications, and other types of solid-core fibers have been explored at wavelengths where silica is not transparent. However, all fibers that rely on light propagation principally through a solid material have certain fundamental limitations stemming from nonlinear effects, light absorption by electrons or phonons, material dispersion, and Rayleigh scattering that limit maximum optical transmission power and increase attenuation losses. These limitations have, in turn, motivated the study of a fundamentally different light-guiding methodology: the use of hollow waveguides having highly reflecting walls. Light propagation through air in a hollow fiber eliminates or greatly reduces the problems of nonlinearities, thermal lensing, and end-reflections, facilitating high-power laser guidance and other applications which may be impossible using conventional fibers. Hollow metallic or metallo-dielectric waveguides have been studied fairly extensively and found useful practical application, but their performance has been bounded by the notable losses occurring in metallic reflections at visible and infrared (IR) wavelengths, as well as by the limited length and mechanical flexibility of the fabricated waveguides. Hollow all-dielectric fibers, relying on specular or attenuated total reflection, have also been explored, but high transmission losses have prevented their broad application. More recently, all-dielectric fibers,

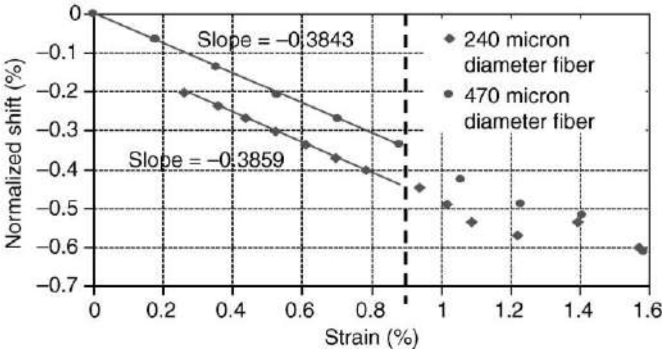

Figure 15 Normalized shift of the Fabry–Perot resonant mode (defined as $\Delta\lambda/\lambda$) versus applied axial strain for the 240 micron (diamonds) and the 460 micron diameter fiber (dots). The curve obtained for the 240 micron diameter fiber is lower because the experimental 0% strain is likely to correspond to a nonzero positive strain, necessary to keep this thinner fiber straight for the measurement (equivalent to a load less than 10 g). Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

consisting of a periodic array of air holes in silica, have been used to guide light through air using narrow photonic bandgaps. Solid-core, index-guiding versions of these silica photonic crystal fibers have also been explored for interesting and important applications, such as very large core single-mode fibers, nonlinear enhancement and broadband supercontinuum generation, polarization maintenance, and dispersion management. However, the air-guiding capabilities of such waveguides thus far remain inferior to transmission through solid silica, due to various factors such as the difficulties in fabricating long, uniform fibers which must have a high volume fraction of air and many air–hole periods, as well as by the large electromagnetic (EM) penetration depths associated with the small photonic bandgaps achievable in these air–silica structures. In our fiber, the hollow core is surrounded by a solid high-refractive-index-contrast multilayer structure leading to large photonic bandgaps and omnidirectional reflectivity. The pertinent theoretical background and recent analyses indicate that such fibers may be able to achieve ultralow losses and other unique transmission properties. The large photonic bandgaps result in very short EM penetration depths within the layer structure, significantly reducing radiation and absorption losses while increasing robustness. Omnidirectional reflectivity is expected to reduce intermode coupling losses.

To achieve high index contrast in the layered portion of the fiber, we combined a chalcogenide glass with a refractive index of ~2.8 As_2Se_3, and a high-performance polymer with a refractive index of ~1.55 PES. We recently demonstrated that these materials could be thermally co-drawn into precisely layered structures without cracking or delamination, even under large temperature excursions. The same polymer was used as a cladding material, resulting in fibers composed of ~98% polymer by volume (not including the hollow core) and thus combine high optical performance with polymeric processability and mechanical flexibility. We fabricated a variety of fibers by depositing a 5–10 micron thick As_2Se_3 layer through thermal evaporation onto a 25–50 micron thick PES film and the subsequent 'rolling' of that coated film into a hollow multilayer tube called a fiber preform. This hollow macroscopic preform was consolidated by heating under vacuum and cladded with a thick outer layer of PES; the layered preform was then placed in an optical fiber draw tower and drawn down into tens or hundreds of meters of fiber having well-controlled submicron layer thicknesses. The nominal positions of the photonic bandgaps were determined by laser monitoring of the fiber OD during the draw process. Typical standard deviations in the fiber OD were ~1% of the OD. The resulting fibers were designed to have large hollow cores, useful in high-energy transmission.

SEM analysis (Figure 16) reveals that the drawn fibers maintain proportionate layer thickness ratios and that the PES and As_2Se_3 films adhere well during rigorous thermal cycling and elongation. Within the multilayer structure shown in Figure 1, the PES layers (gray) have a thickness of 900 nm, and the As_2Se_3 layers (bright) are 270 nm thick (except for the first and last As_2Se_3 layers, which are 135 nm). Broadband fiber transmission spectra were measured with a Fourier transform infrared (FTIR) spectrometer (Nicolet Magna 860), using a parabolic mirror to couple light into the fiber and an external detector. The results of these measurements are shown in the lower panel of Figure 2 for fibers having two different layer structures. For each spectrum, light is guided at the fundamental and high-order photonic bandgaps. Also shown in the upper panel of Figure 2, is the corresponding photonic band diagram for an infinite periodic multilayer structure calculated using the experimental parameters of our fiber (layer thicknesses and indices). Good agreement is found between the positions of the measured transmission peaks and the calculated bandgaps, corroborated with the SEM-measured layer thicknesses, verifying that transmission is dominated by the photonic bandgap mechanism. In order to demonstrate 'wavelength scalability' (i.e., the control of transmission through the fiber's structural parameters) another fiber was produced, having the same cross-section but with thinner layers. We compared the transmission

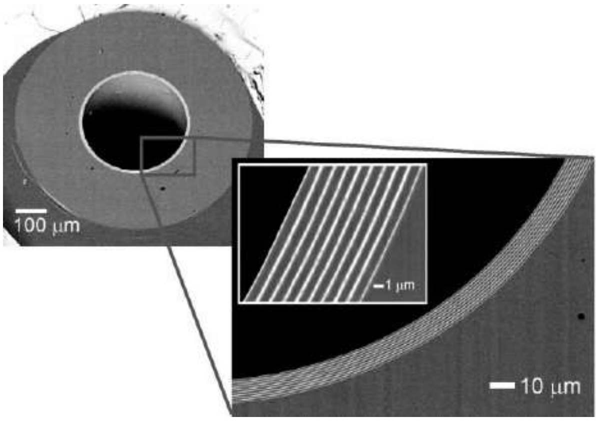

Figure 16 Cross-sectional SEM micrographs at various magnifications of hollow cylindrical multilayer fiber mounted in epoxy. The hollow core appears black, the PES layers and cladding gray, and the As_2Se_3 layers bright white. This fiber has a fundamental photonic bandgap at a wavelength of ~3.55 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

spectra for the original 3.55 micron bandgap fibers to the fiber with the scaled-down layer thicknesses.

Figure 17 shows the shifting of the transmission bands, corresponding to fundamental and high-order photonic bandgaps, from one fiber to the next. The two fibers analyzed in Figure 17 were fabricated from the same fiber preform using different draw-down ratios (fibers with a fundamental bandgap centered near 3.55 microns have an OD of 670 microns; those with a gap at 3.1 microns have an OD of 600 microns). The high-order bandgaps are periodically spaced in frequency, as expected for such a photonic crystal structure.

The wavelength scalability of our fibers was further demonstrated in the fabrication of hollow fibers designed for the transmission of 10.6 micron EM radiation. This not only shows that these structures can be made to guide light at extremely disparate wavelengths, but that specific useful bandgap wavelengths can be accurately targeted during fabrication and fiber drawing. Powerful and efficient CO_2 lasers are available that emit at 10.6 microns and are used in such applications as laser surgery and materials processing, but waveguides operating at this wavelength have remained limited in length or loss levels. Using the fabrication techniques outlined above, we produced fibers having hollow core diameters of 700–750 microns and ODs of 1300–1400 microns with a fundamental photonic bandgap spanning the 10–11 micron wavelength regime, centered near 10.6 microns. Figure 5 depicts a typical FTIR transmission spectrum for these fibers, measured using ~30 cm long straight fibers.

In order to quantify the transmission losses in these 10.6 micron bandgap hollow fibers, fiber cutback measurements were performed. This involved the comparison of transmitted intensity through ~4 meters of straight fiber with the intensity of transmission through the same section of fiber cut to shorter lengths (Figure 18 inset). This test was performed on multiple sections of fiber, and the results found to be nearly identical for the different sections tested. The measurements were performed using a 25 watt CO_2 laser (GEM-25, Coherent-DEOS) and high power detectors (Newport 818T-10). The fiber was held straight, fixed at both ends as well as at multiple points in the middle to prevent variations in the input coupling and propagation conditions during fiber cutting. The laser beam was sent through focusing lenses as well as 500 micron diameter pinhole apertures and the input end face of the fiber was coated with a metal film to prevent accidental

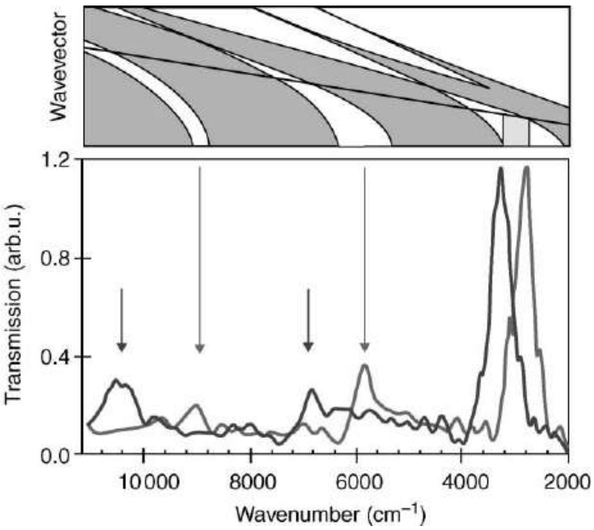

Figure 17 Upper panel: Calculated photonic bandstructure associated with the dielectric mirror lining of the hollow fiber. Modes propagating through air and reflected by the fiber walls lie in the bandgaps (white) and within the light cone defined by the glancing-angle condition (black line). The gray regions represent modes radiating through the mirror. The fundamental bandgap has the widest range-to-mid-range ratio, with a region of omnidirectional reflectivity highlighted in black. Lower panel: Comparison of transmission spectra for two different hollow fibers of ~30 cm length having similar structure but scaled photonic crystal (multilayer) period dimensions. The spectrum in gray is from a fiber with a fundamental photonic bandgap at 3.55 microns; the black spectrum is from a fiber where the corresponding bandgap is at 3.1 microns. High-order bandgaps (indicated by arrows) are periodically spaced in frequency. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

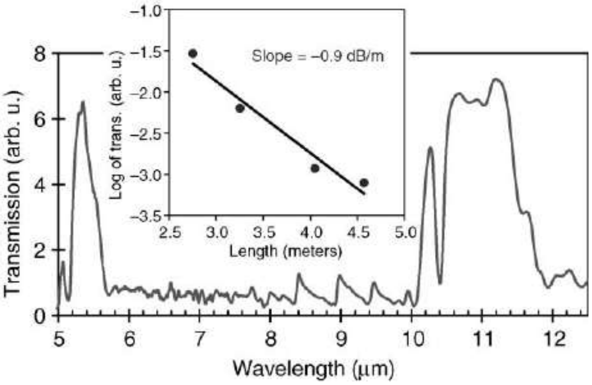

Figure 18 Typical transmission spectrum of hollow fibers designed to transmit CO_2 laser light. The fundamental photonic bandgap is centered near a wavelength of 10.6 microns and the second-order gap is at ~5 microns. Inset: Log of transmitted power (arbitrary units) versus length of fiber (meters). The slope of this graph is the loss in dB/m. The measured fiber has a hollow core diameter of 700 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

laser damage from misalignment. The transmission losses in the fundamental bandgap at 10.6 microns were measured to be 0.95 dB/m, as shown in the inset of Figure 18, with an estimated measurement uncertainty of 0.15 dB/m. These loss measurements are comparable to some of the best reported loss values for other types of waveguides operating at 10.6 microns. A bending analysis for fibers with a bandgap centered at 10.6 microns revealed bending losses below 1.5 dB for 90 degree bends with bending radii from 4–10 cm. We expect that these loss levels could be lowered even further by increasing the number of layers, through optimization of the layer thickness ratios and by creating a cylindrically symmetric multilayer fiber with no inner seam (present here because of the 'rolling' fabrication method). In addition, using a polymer with lower intrinsic losses should greatly improve the transmission characteristics.

One reasonable figure of merit for optical transmission losses through hollow all-dielectric photonic bandgap fibers is to compare the hollow fiber losses to the intrinsic losses of the materials used to make the fiber. As_2Se_3 has been explored as an IR-transmitting material, yet the losses at 10.6 microns reported in the literature are ~10 dB/m for highly purified material, and more typically are greater than 10 dB/m for commercially available materials such as those used in our fabrication. Based on FTIR transmission and spectroscopic ellipsometer measurements that we have performed on PES, the optical losses associated with propagation through solid PES should be greater than 40 000 dB/m at 10.6 microns. This demonstrates that guiding light through air in our hollow bandgap fibers leads to waveguide losses that are orders of magnitude lower than the intrinsic fiber material losses, which has been one of the primary goals of hollow photonic bandgap fiber research. These comparatively low losses are made possible by the very short penetration depths of EM waves in the high refractive index contrast photonic crystal structure, allowing these materials to be used at wavelengths that may have been thought improbable. Another long-standing motivation of infrared fiber research has been the transmission of high-power laser light. As a qualitative demonstration of the potential of these fibers for such applications, both straight and smoothly bent fibers of lengths varying from 0.3–2.5 meters were used to transmit enough CO_2 laser energy to burn holes through paper. The maximum laser power density coupled into our fibers in these trials was approximately 300 W/cm^2, more than sufficient to burn a homogeneous polymer material, including PES. No damage to the fibers was observed when the laser beam was properly coupled into the hollow fiber core. These results indicate the feasibility of using hollow multilayer photonic bandgap fibers as a low-loss wavelength-scalable transmission medium for high-power laser light.

See also

Diffractive Systems: Design and Fabrication of Diffractive Optical Elements. **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber. **Lasers:** Carbon Dioxide Lasers. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Abeles F (1950) Investigations on the propagation of sinusoidal electromagnetic waves in stratified media. Application to thin films. *Annales de Physique* 5: 706.

Benoit G and Fink Y, Spectroscopic Ellipsometry Database, http://mit-pbg.mit.edu/Pages/Ellipsometry.html

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. New York: Pergamon Press, p. 67.

Engeness TD, Ibanescu M, Johnson SG, *et al.* (2003) Dispersion tailoring and compensation by modal interactions in OmniGuide fibers. *Optics Express* 11: 1175–1198. http://www.opticsexpress.org/abstract.cfm?URI = OPEX-11-10-1175

Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682.

Fink Y, Ripin DJ, Fan S, Chen C, Joannopoulos JD and Thomas EL (1999) Guiding optical light in air using an all-dielectric structure. *Journal of Lightwave Technology* 17: 2039–2041.

Hart SD, Maskaly GR, Temelkuran B, Prideaux PH, Joannopoulos JD and Fink Y (2002) External reflection from omnidirectional dielectric mirror fibers. *Science* 296: 510–513.

Ibanescu M, Johnson SG, Soljacic M, *et al.* (2003) Analysis of mode structure in hollow dielectric waveguide fibers. *Physics Review E* 67: 046608-1–8.

Johnson SG, Ibanescu M, Skorobogatiy M, *et al.* (2001) Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Optics Express* 9: 748–779. http://www.opticsexpress.org/abstract.cfm?URI = OPEX-9-13-748

Kuriki K, Shapira O, Hart SD, *et al.* (2004) Hollow multilayer photonic bandgap fibers for NIR applications. *Optics Express* 12: 8.

Soljacic M, Ibanescu M, Johnson SG, Joannopoulos JD and Fink Y (2003) Optical bistability in axially modulated OmniGuide fibers. *Optics Letters* 28: 516–518.

Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

Yeh P, Yariv A and Hong C-H (1977) Electromagnetic propagation in periodic stratified media. 1. General theory. *Journal of the Optical Society of America* 67(4): 423.

Yeh P, Yariv A and Marom E (1978) Theory of Bragg fiber. *Journal of the Optical Society* 68: 1196–1201.

Wave Optical Modeling and Design

F Wyrowski, Friedrich-Schiller University, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Why do ray optics dominate the literature on optical engineering? The major practical reason may be that historically optical engineering has dealt with systems for image forming. These systems transform divergent spherical waves from points in an object, into spherical waves converging to focal points in the image. Thus, the basic operation of an imaging system performed by a lens is changing the radius of curvature of spherical fields. The propagation of spherical waves is modeled well by geometrical optics, so long as truncation by apertures and other high frequency aberrations are not significant. Moreover, the basic merit function in imaging evaluates the appearance of the image point at its correct location by so-called dot diagrams. Evaluation of this merit function requires only a ray bundle representation of the output field. The combination of a ray representation with geometrical optics propagation is called raytracing, which explains the dominance of raytracing in conventional optical engineering. Because the design of illumination systems is also often based on principles of imaging, raytracing remains a suitable technique in that case too. However, systems that perform general light transformations must apply wave-optical engineering techniques, which constitutes a generalization of conventional optical engineering. In this sense it includes the modeling of image-forming systems as a special case.

Modeling Principles

Modeling in wave-optical engineering addresses the propagation of electromagnetic fields through systems as illustrated in Figure 1. The introduction of operators and operands as mathematical tools simplifies a systematic discussion of the modeling concepts. Formally, an electromagnetic field may be expressed by the operand $\mathbf{f} \in \mathsf{F}$ with the set of fields F. The actual form of $\mathbf{f} \in \mathsf{F}$ depends on the physical approach chosen to model electromagnetic fields. For instance, $\mathbf{f} \in \mathsf{F}$ may stand for a ray bundle, a complex amplitude, or a coherence function. The photon source generates a field which acts as the input field $\mathbf{f}$ into the optical system. The quality of the output field $\mathbf{f}_{\text{out}}$ is evaluated with the help of a merit function $\boldsymbol{\Omega}$. This function may consist of M components which yields:

$$\boldsymbol{\Omega}(\mathbf{f}_{\text{out}}) = \{\Omega_1(\mathbf{f}_{\text{out}}), \ldots, \Omega_\mu(\mathbf{f}_{\text{out}}), \ldots, \Omega_M(\mathbf{f}_{\text{out}})\} \qquad [1]$$

with $\Omega_\mu : \mathsf{F} \to \mathsf{R}^+$ and $\mu = 1, \ldots, M$. Each component Ω_μ defines a mathematical way to obtain one positive valued quantity for a given $\mathbf{f}_{\text{out}}$. Examples are the rms-value to evaluate the quality of a wavefront, M^2-value to judge the profile of a laser beam, center of gravity of a light distribution, light transformation efficiency, and the value of an overlap integral to determine the coupling efficiency into a fiber. Any property of the output field can be evaluated by an appropriate definition of a function Ω_μ, as long as the demanded field parameters are accessible for the field model in use. However, it should be mentioned, that it is not clear *a priori*, that all quality criteria Ω_μ can be satisfied simultaneously. How the different Ω_μ are related is dependent on the application.

The propagation of the input field through the system is mathematically summarized by the operator equation:

$$\mathbf{f}_{\text{out}} = S(p_\alpha)\mathbf{f}_{\text{in}} \qquad [2]$$

with the operator $S : \mathsf{F} \to \mathsf{F}$. p_α stands for all parameters p_1, p_α, p_{A} which specify the system, for instance position of elements, refractive index, surface profiles, and layer thickness. Obviously a full electromagnetic model of S does not exist for arbitrary systems. The propagation problem eqn [2] is much too complex for a wave-optical solution. However, for a sequential approach the situation is different. If the propagation through a system is performed sequentially through the elements and the homogenous regions, at least approximate wave-optical methods are available to model the propagation S. In a sequential approach, either the transmitted or reflected field is considered per region

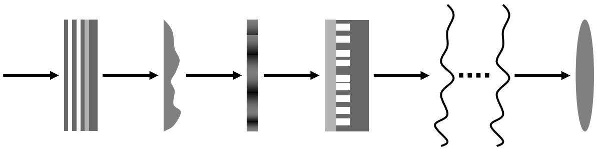

Figure 1 An optical system may include various types of elements to perform the desired optical function. Layered media, free-form surfaces, index-modulated regions, microstructured interfaces, lenses, and other structured media offer the flexibility needed to realize general light transformations. Typically structured media and homogenous regions follow each other.

of the system. Although this is an approximation, it turns out to be very useful in wave-optical engineering. With regard to Figure 2 the sequential approach is mathematically expressed by

$$\mathbf{f}_{\text{out}} = S(p_\alpha)\mathbf{f}_{\text{in}} = \mathcal{P}_J(n_J, \Delta z_J)S_{J-1}(p_{\alpha_{J-1}})\cdots S_2(p_{\alpha_2}) \times \mathcal{P}_1(n_1, \Delta z_1)\mathbf{f}_{\text{in}} \qquad [3]$$

$\mathcal{P}_j(n_j, \Delta z_j)$ expresses the propagation through a homogenous dielectric with refractive index n_j and extent Δz_j. This type of propagation is also called free propagation. Sometimes the term free propagation is restricted to propagation in vacuum; Here it is used more generally. $S_j(p_{\alpha_j})$ models the propagation through an optical component which is specified by the parameters p_{α_j}. Optical engineering needs physical models for the field operand $\mathbf{f}$, the free propagation $\mathcal{P}$, and the propagation S_j through elements. The most fundamental demand in wave-optical engineering is a wave-optical model to express the field $\mathbf{f}$ in homogenous regions. It is common to restrict to harmonic fields. The generalization to more general radiation is often possible by suitable superposition of harmonic fields. This approach assumes linear optical systems as typical in optical engineering. From Maxwell's equations it follows that harmonic fields in homogenous regions are completely specified if two independent field components are determined in one plane z_0 within the region. Therefore, the electromagnetic field operand takes the form:

$$\mathbf{f}(x, y, z_0) = [E_x(x, y, z_0), E_y(x, y, z_0)] \qquad [4]$$

with the complex amplitudes E_x and E_y of the x and y components of the electric field vector. Because the knowledge of $\mathbf{f}(x, y, z_0)$ allows the direct calculation of the missing complex amplitudes E_z, H_x, H_y, H_z via Maxwell's equations, the entire electromagnetic harmonic field in a homogenous region is specified by eqn [4].

With respect to the two components of the field operand $\mathbf{f}$ propagation operators S are 2×2-matrices of the form:

$$S = \begin{Bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{Bmatrix} \qquad [5]$$

In the case of a diagonal matrix, that is $S_{xy} = S_{yx} = 0$, E_x and E_y propagate independently and are considered decoupled channels of the system. If in addition $S_{xx} = S_{yy}$, both channels behave identically, and it is sufficient to model only one channel. This is often referred to as the scalar model. Of course, also in the scalar model the electromagnetic field still consists of two independent components and maintains its vectorial nature.

Numerical implementation of eqn [3] for the field operand of eqn [4] constitutes the backbone of modeling in wave-optical engineering.

Free Propagation

The propagation $\mathcal{P}$ of harmonic fields $\mathbf{f}$ through homogenous regions of a system has a rigorous electromagnetic solution in form of the spectrum of plane waves integral which yields:

$$\begin{aligned} \mathbf{f}(x,y,z \geq z_0) &= \mathcal{P}_{\text{SPW}}\mathbf{f}(x,y,z_0) \\ &= \begin{pmatrix} \mathcal{P}_{\text{SPW}} & 0 \\ 0 & \mathcal{P}_{\text{SPW}} \end{pmatrix} \begin{pmatrix} E_x(x,y,z_0) \\ E_y(x,y,z_0) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{F}^{-1}\left([\mathcal{F}E_x(x,y,z_0)]\mathrm{e}^{\mathrm{i}(k_x x + k_y y + k_z \Delta z)}\right) \\ \mathcal{F}^{-1}\left([\mathcal{F}E_y(x,y,z_0)]\mathrm{e}^{\mathrm{i}(k_x x + k_y y + k_z \Delta z)}\right) \end{pmatrix} \end{aligned} \qquad [6]$$

with the Fourier transform $\mathcal{F}$ and $\Delta z = z - z_0$. The term $\mathcal{F}E(x,y,z_0)$ is called spectrum of planewaves (SPW). The operator matrix is diagonal and $\mathcal{P}_{\text{SPW}} = \mathcal{P}_{xx} = \mathcal{P}_{yy}$, that means the x- and y-components propagate independent of each other according to the same formula in homogenous dielectrics.

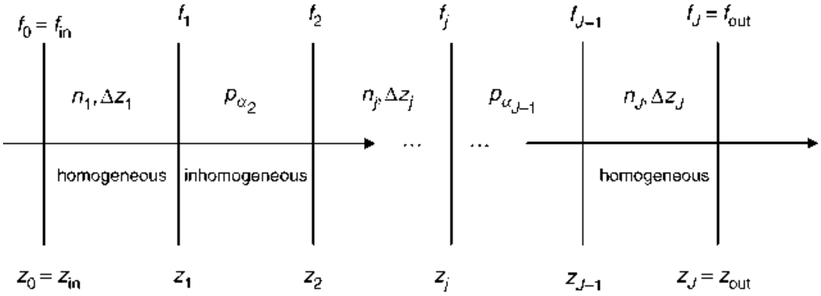

Figure 2 Formally a system may be subdivided into a sequence of homogenous dielectrics like air and inhomogeneous regions which include the elements. We indicate a region by the letter *j*.

In this sense, free propagation is a scalar problem which allows the convenient use of $U(x,y,z)$ instead of two field components for the sake of simpler mathematical expressions.

Equation [6] allows the derivation of other rigorous versions of the free propagation operator $\mathcal{P}_{\mathrm{SPW}}$, for instance Rayleigh's integral formula, which is well-known from diffraction theory. Moreover, eqn [6] leads to important approximate operators for far-field propagation and the propagation of paraxial fields. Propagation of a field into its far field is given by the operator:

$$\begin{aligned} U(r\mathbf{s}) &= \mathcal{P}_{\mathrm{far}} U(x,y,z_0) \\ &= -\mathrm{i}2\pi k s_z [\mathcal{F} U(x,y,z_0)]|_{(ks_x,ks_y)} \frac{\exp(\mathrm{i}kr)}{r} \end{aligned} \qquad [7]$$

with the wave number $k = 2\pi n/\lambda$, $r = |\mathbf{r}| = (x^2 + y^2 + (\Delta z)^2)^{1/2}$, and the direction vector $\mathbf{s} = (s_x, s_y, s_z) = \mathbf{r}/r$. Here, and in what follows, λ denotes the vacuum wavelength. $|_{(ks_x,ks_y)}$ indicates the variables to be inserted into the Fourier transformed field. Though the far-field formula is derived for $r \to \infty$, it is also of great practical value for the propagation of fields generated by nonparaxially emitting sources like laser diodes. Roughly speaking eqn [7] predicts a spherical wave in the far field which is modulated by the spectrum of planewaves of the field in z_0. $\mathcal{P}_{\mathrm{far}}$ propagates from the plane z_0 to the spherical surface with radius r and not to a plane z.

The paraxial approximation of eqn [6] follows if the spectrum of planewaves $\mathcal{F} U(x,y,z_0)$ possesses significant values only for small k_x and k_y, that is if both electric field components propagate approximately along the z-axis. The resulting propagation operator is known as Fresnel's integral formula. In practice, its formulation with the Fourier transform $\mathcal{F}$ is typically more suitable:

$$\begin{aligned} U(x,y,z \geq z_0) &= \mathcal{P}_{\mathrm{paraxial}} U(x,y,z_0) \\ &= \frac{k \exp[\mathrm{i}k\Delta z]}{\mathrm{i}2\pi\Delta z} \exp\left[\frac{\mathrm{i}k(x^2+y^2)}{2\Delta z}\right] \\ &\times \left| \left\{ \mathcal{F}\left(\exp\left[\frac{\mathrm{i}k(x^2+y^2)}{\Delta z}\right] U(x,y,z_0) \right) \right\} \right|_{(x,y)n/\lambda\Delta z} \end{aligned} \qquad [8]$$

$|_{(x,y)n/\lambda\Delta z}$ indicates the variables to be inserted after Fourier transform. The far-field approximation of eqn [8] leads to the well-known Fraunhofer diffraction formula.

Equations [6]–[8] provide fundamental operators for the propagation of fields in wave-optical engineering. All integrals are based on Fourier transforms which allows the application of the fast Fourier transform (FFT) in practice. Dependent on the propagation technique, one or two FFTs must be performed. In practice, the sampling of fields is of great concern. Naturally the amount of data should be as small as possible. According to the rigorous propagation operator eqn [6] the (x,y)-coordinate system and therefore the sampling period in z_0 and z are identical, that is, scaling does not occur. As illustrated in Figure 3, that can lead to an enormous amount of data for large propagation distances. For demonstration purpose the field amplitude depicted at the top is chosen. The field window has the size 1 mm × 4 mm. The wavelength λ is 632.8 nm, $n = 1$, and the sampling distance 10 μm. That results in 100 × 400 sampling points. Propagation of this field with the spectrum of planewave operator maintains the size of the field window. The resulting field for a propagation distance $\Delta z = 10$ mm is shown in the middle. Because of diffraction, the field widens. As long as the initial window is large enough to encompass the propagated field, the propagation method works well. At the bottom, the resulting field for $\Delta z = 200$ mm is shown. The field exceeds the size of the window and that leads to numerical errors (aliasing). This can be avoided by embedding the initial field into a window which is large enough with respect to the expected size of the propagated field. As a result, one obtains the

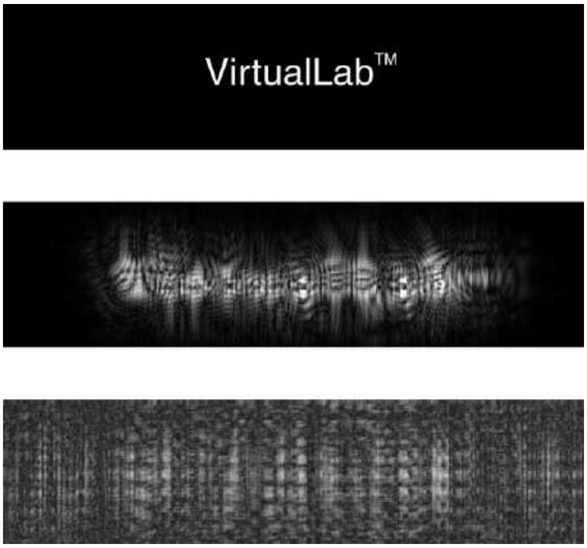

Figure 3 Illustration of numerical properties of the spectrum of plane wave propagation technique. The field amplitude depicted at the top is numerically propagated in free space and the distributions in the middle and the bottom are obtained. The simulations in this and all following figures were performed. The simulations in this and all following figures were performed with the wave-optical engineering software VirtualLab™ from LightTrans GmbH.

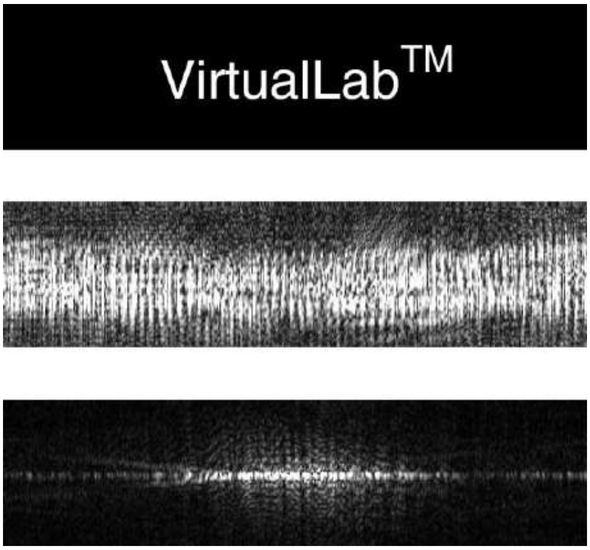

Figure 4 Starting from the same input field and the same distances as in Figure 3 the propagated fields depicted in the middle and the bottom are obtained using the paraxial propagation operator. Because of the inherent scaling the field window sizes are 632.8 μm × 632.8 μm and 12.656 mm × 12.656 mm, respectively. The number of sampling points remains unchanged. Obviously, the field in the middle is too small to encompass the propagated field and numerical errors occur. For the long distance the field size is perfectly matched to the field size obtained by propagation.

demanded accuracy at cost of increasing data memory requirements and computation time. Thus in practice the SPW propagation operator is particularly suitable for short distances. Fortunately the paraxial propagation operator eqn [6] possesses a complementary behavior. Because of $(x, y)n/\lambda\Delta z$, the (x, y)-coordinate system changes its scaling with the distance. As illustrated in Figure 4, that makes it well suited for large propagation distances but problems occur for small ones. Therefore, all paraxial propagation problems have an efficient numerical solution by a suitable combination of $\mathcal{P}_{\text{paraxial}}$ and $\mathcal{P}_{\text{SPW}}$. In the nonparaxial case the far-field operator eqn [7] often helps to realize a numerically efficient propagation for large distances. However, there are sometimes situations, in which free propagation of nonparaxial fields causes significant numerical problems.

Propagation Through Elements

The propagation of harmonic fields $\mathbf{f}$ through elements is a challenging task. For example, Figure 5 illustrates propagation through smooth or microstructured interfaces between two homogenous media. The field $\mathbf{f}_1(x, y, z_1)$ should be determined if $\mathbf{f}_0(x, y, z_0)$, the shape of the interface, and the refractive indices n_0 and n_1 are known. The reflection mode of operation can be formulated analogously. It is remarkable that this basic problem possesses a rigorous solution only for a few special interface geometries. Of particular importance are the rigorous solutions for plane and periodic interfaces. In the first case, the solution comprizes Snell's law of refraction, the law of reflection, and Fresnel's formulae. Grating theory provides the solution for periodic interfaces. It is possible to use these methods also for nonperiodic interfaces but at the cost of immense computational effort. Thus, they are only practicable for structures with detail no larger than a few wavelengths. For general interfaces, accurate approximate methods are required. One approach is to decompose a general propagation problem into a set of simpler problems that can be solved rigorously. The results are then combined to obtain an approximate solution of the original problem. Two basic concepts for subdividing the general problem are: (i) Decomposition of the incident field into laterally truncated fragments which illuminate elementary fractions of the interface. (ii) Decomposition of the response of the interaction of the incident field with the interface into approximately independent ones related to local interface features. The first technique is referred to as local elementary interface approximation (LEIA) and the second one as local independent response approximation (LIRA). Elementary interfaces of particular importance are plane, periodic, and spherical ones.

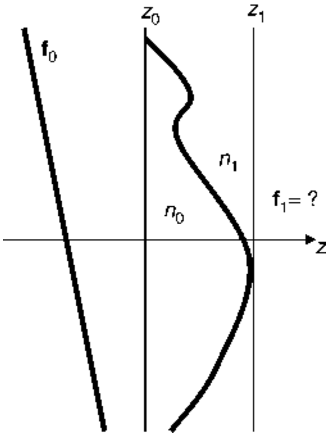

Figure 5 A planewave $\mathbf{f}_0$ propagates through an interface between two homogeneous dielectrics with refractive index n_0 and n_1. The interface is of general shape and therefore a rigorous solution of the propagation problem does not exist. The virtual planes in z_0 and z_1 encompass the interface in a way which minimizes the distance $z_1 - z_0$.

The local plane interface approximation (LPIA) is of particular importance in wave-optical engineering. Figure 6 illustrates the basic concept for the example

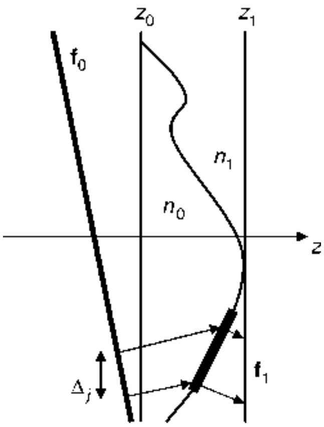

Figure 6 If a lateral fraction of the input field is propagated towards the interface, it is often possible to laterally restrict the major part of the propagated field to a local plane interface. Because of diffraction it is never possible to do that rigorously. Neglecting diffraction in the region between z_0 and z_1 provides a simple model to propagate a section of the incident plane field through the interface. Fresnel's formulae and Snell's law of refraction can then be applied directly to obtain the transmitted plane field, which is propagated to the plane z_1.

of a plane incident field. It has been proven that the approximation of geometrical optics propagation within the regions of the component, here that means between z_0 and z_1, provides results which are surprisingly accurate in numerous situations of practical interest. A full wave-optical version of LPIA is currently under research and development. In general, however, LPIA is applicable if the characteristic lateral extent of the local modulation of the interface is not close to the wavelength. It should be emphasized that LPIA is, as most other approximate propagation techniques, a 2×2 matrix, like eqn [5], with nonvanishing components. That means, for instance, that in general, an E_y-component occurs in z_1 even if $E_y = 0$ in z_0. That is, the two field channels experience crosstalk. A further approximation in geometrical optics LPIA, popular in Fourier optics holography, and diffractive optics is the thin element approximation (TEA). Geometrical optics LPIA leads to TEA if refraction and the local dependence of Fresnel losses at the interface are neglected. Obviously, that is allowed only if all angles involved, including the slope of the interface and the direction of the input field, are small. Therefore, TEA is a paraxial and geometrical optics approximation. Moreover, small local slopes of the interface lead to a restriction to thin elements, which explains the name TEA. It is also referred to as scalar approximation, because the operator matrix becomes diagonal with $S_{\mathrm{TEA}} = S_{xx} = S_{yy}$, that means the two field channels propagate independently according the same formula through the interface. In the thin element only approximation the dependence of the optical path through the element on position (x, y) is considered and mathematically takes the form:

$$\begin{aligned} U(x,y,z_1) &= S_{\mathrm{TEA}} U(x,y,z_0) \\ &= c_{\mathrm{Fresnel}} U(x,y,z_0) \exp\left[\frac{\mathrm{i}2\pi n_1(z_1 - z_0)}{\lambda}\right] \\ &\quad \times \exp\left[\frac{\mathrm{i}2\pi h(x,y)(n_0 - n_1)}{\lambda}\right] \end{aligned} \qquad [9]$$

with the vacuum wavelength λ. The profile height $h(x, y)$ is expressed as the distance from z_0 to the interface. Propagation through an interface between dielectrics modeled with TEA affects the phase of the incident field proportional to the profile height. The magnitude remains unchanged besides a constant factor c_{Fresnel}, due to Fresnel's equations. If complex valued refractive indices are taken into consideration, absorption results in an amplitude modulation completely analogous to the phase modulation described in eqn [9]. This way it is also possible to address the propagation through apertures and the combination of S_{TEA} with $\mathcal{P}$ leads to the well-known diffraction integrals. In fact, Kirchhoff's boundary conditions of field propagation through a hole in a completely opaque screen can be interpreted as TEA for complex valued refractive indices.

The simple form of the TEA operator allows a straightforward numerical implementation. The field operand and the profile are equidistantly sampled and the resulting field is sampled analogously. That makes the combination of S_{TEA} with numerical implementations of the free propagation operators $\mathcal{P}$ easy. As soon as refraction is considered in geometrical optics LPIA, the equidistant sampling of the output field is destroyed which enforces additional interpolation effort before a free propagation by fast Fourier transforms can be implemented, because fast numerical Fourier transform algorithms require an equidistant sampling.

In wave-optical engineering the combination of geometrical optics LPIA with a wave-optical field model is mandatory. In this context it is instructive to mention, that its combination with a ray-bundle representation of the incident electromagnetic field leads to raytracing through the interface, that is it reduces to the basic technique in conventional optical engineering.

Similar to LPIA it is possible to consider linear gratings as elementary local interface and the local linear grating approximation (LLGA) results. It is of special concern for all types of diffractive elements, for which the interface can locally considered periodically. An important example is given for diffractive lenses in the outer zones of high numerical aperture lenses. In the center and for low numerical aperture TEA is appropriate for propagating through diffractive lenses.

Design Principles

The design problem in wave-optical engineering can be stated as follows: specify system parameters p_α in a way that ensures for a given input field $\mathbf{f}_{\text{in}}$ an output field $\mathbf{f}_{\text{out}}$ which satisfies the quality criteria defined by the merit functions Ω_μ. In addition, one may demand a system which is most simple, particularly cheap, or various other system properties of concern in practice.

Design constitutes an optimization problem. Restrictions to the system parameters and the merit function sets design constraints as well as design freedoms.

System parameters can be restricted, for instance, by the materials available, the maximum size of a system, and types of element profiles. Within these limits the system parameters are chosen freely during optimization to achieve the optical function.

The merit functions determines which parameters of the output field are important and should be emphasized, for instance the intensity distribution. All output fields, whose parameters satisfy the constraints satisfy the design problem. In most applications, not all field parameters are constrained. For instance, in laser materials processing or illumination applications, the merit function is related to energy quantities and then the phase of the output field can be used as a free parameter distribution. Moreover, the output field is typically not constrained in the entire output plane, but only in so-called signal regions, which introduces freedom of amplitude outside the signal regions. Hence, dependent on the application, there exist field parameters that can be chosen partly or completely free during design.

Successful design strategies use the freedoms to improve the matrix function. It is often not clear a priori, if a solution does exist. Then the field constraints are satisfied as good as possible in terms of the merit function.

In imaging systems, typical system parameters include, for instance, distances, refractive indices, curvature of lenses, and parameters to specify aspherical surfaces. For high-end lens systems, the resulting number of parameters can be large, but they are still small enough to utilize parameter optimization algorithms if the designer starts with a reasonable initial guess. In wave-optical engineering, the number of system parameters can become huge due to the generalized nature of the structures, which typically prevents the use of system parameter optimization techniques. For example, the profile of diffractive elements is typically specified by 10^4–10^6 parameters. Thus a direct optimization of the system parameters, which requires the evaluation of $S(p_\alpha)$ (see eqn [3]) after any change of system parameters p_α, is not realistic.

Instead of optimizing the system parameters directly, which is referred to as design in the structural embodiment of the system, a technique wellknown from Fourier optics and diffractive optics can be generalized for wave-optical engineering. It does not start with structure related element parameters but replaces all or at least some of the operators $S_j(p_{\alpha_j})$ of eqn [3] by mathematical operators $\mathcal{T}_j(t_{\tau_j})$. t_τ denotes the parameters of $\mathcal{T}$. The propagation operator through a system expressed in this functional embodiment has the form:

$$\begin{aligned}\mathbf{f}_{\text{out}} &= S_{\text{functional}}\mathbf{f}_{\text{in}} \\ &= \mathcal{P}_J(n_J, \Delta z_J)\mathcal{T}_{J-1}(t_{\tau_{J-1}})\cdots\mathcal{T}_2(t_{\tau_2}) \\ &\quad \times \mathcal{P}_1(n_1, \Delta z_1)\mathbf{f}_{\text{in}}\end{aligned} \qquad [10]$$

if all propagations through elements are expressed in functional form. A very important example of a functional operator $\mathcal{T}$ is given by a simple multiplication of both field components with identical or different complex functions $t(x, y)$, that is:

$$\mathcal{T}_t(t_\tau)\mathbf{f}_{\text{in}} = \begin{pmatrix} t_{xx}(x, y) & 0 \\ 0 & t_{yy}(x, y) \end{pmatrix}\mathbf{f}_{\text{in}}(x, y) \qquad [11]$$

The values of both $t(x, y)$ form the parameters t_τ of $\mathcal{T}$. The functions $t(x, y)$ are also called ideal transmission functions of elements. Of special interest are phase-only transmission functions $t(x, y) = \exp[\mathrm{i}\tau(x, y)]$ (see discussion below). If $t(x, y) = t_{xx}(x, y) = t_{yy}(x, y)$, the design is simplified. Fortunately, in most actual applications this assumption is valid.

Designing in the functional embodiment requires the evaluation of $S_{\text{functional}}$ instead of S of eqn [3], which reduces the numerical complexity enormously. That improves the likelihood for developing efficient design algorithms to optimize the distances Δz_j, the refractive indices n_j, and parameters t_{τ_j}.

The design in the functional embodiment does not deliver the structure parameters of elements but only parameters t_τ of the operator $\mathcal{T}$ which gives information about how a field should be changed at the position z where $\mathcal{T}$ takes effect. Thus, in a next design step, elements must be obtained which cause the required effect. This so-called structure design step is also generally challenging because it requires the solution of an inverse propagation problem through interfaces. For a further discussion the most important $\mathcal{T}$, that is the transmission type operator of eqn [11], is considered. For the sake of simplicity $t(x,y) = t_{xx}(x,y) = t_{yy}(x,y)$ is assumed. Then, a solution which uses geometrical optics LPIA in combination with phase-only transmissions is known in the form of a recursive structure design strategy. In its paraxial approximation, that is TEA, the solution becomes very simple. For a transmission $t(x,y) = \exp[i\tau(x,y)]$, obtained from a functional design, an element with a surface profile:

$$h(x,y) = \frac{\lambda\tau(x,y)}{2\pi(n_{\text{element}} - n_{\text{free}})} \qquad [12]$$

results from eqn [9]. The refractive index of the element is denoted by n_{element} and the homogenous region behind the element has the index n_{free}, eqn [12] constitutes a direct relationship between the phase values of the transmission function and the profile $h(x,y)$. This structure design technique is accurate for all paraxial situations. However, it is also a reasonable method for nonparaxial structure design since it provides an initial profile for further optimization. Obviously it is applicable for phase-only transmission functions only. That underlines the importance of this type of functional operator for the design in the functional embodiment. Often the fabrication of profiles is done with lithographic techniques which result in surface reliefs with a few discrete height levels, that means the profile $h(x, y)$ is quantized. According to eqn [12] a quantization of $h(x, y)$ directly leads to quantized phase values $\tau(x,y)$. This quantization must be taken into account when designing $\tau(x,y)$. An example should serve to discuss further aspects of structure design according to eqn [9]. A fundamental transmission in various applications is the phase of a spherical wave and its paraxial version, the quadratic phase. An example of a spherical phase $\tau(x,y)$ is shown in Figure 7 where the 2π-modularity of the exponential function is taken into account – as normal for phase data which has been obtained by numerical calculations. Inserting this phase into eqn [9] leads to the profile of a diffractive lens. Because of the use of inverse TEA for structure design it works as demanded for moderate numerical apertures and starts to suffer from aberrations for large numerical apertures because of electromagnetic effects in the very outer zones. It is also possible to unwrap the phase before inserting it into eqn [9]. Then a smooth surface, that is a refractive spherical lens, results. It is well-known that a spherical lens works accurately in the paraxial domain and suffers from aberrations for moderate and large numerical apertures, which are more severe than for the diffractive lens. The use of inverse geometrical optics LPIA as a structure design technique is not restricted to the paraxial domain and therefore it leads to a suitable aspherical surface in order to ensure a spherical phase transmission for larger numerical apertures too. This is illustrated in Figure 8. The lens example shows another advantage of the two-step design method, which consists of the design in the functional embodiment prior to the one in the structural form of the element. In the functional

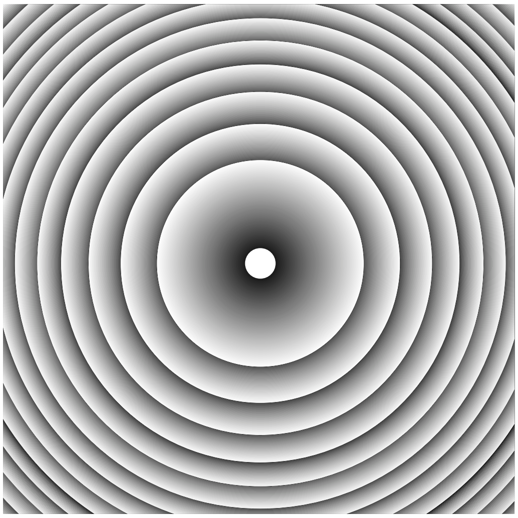

Figure 7 The phase of a spherical phase in the 2π-modulo representation. This phase can be unwrapped which leads to a smooth spherical phase which varies from 0 to various 2π. The noisy edges are due to the use of a cartesian coordinate system.

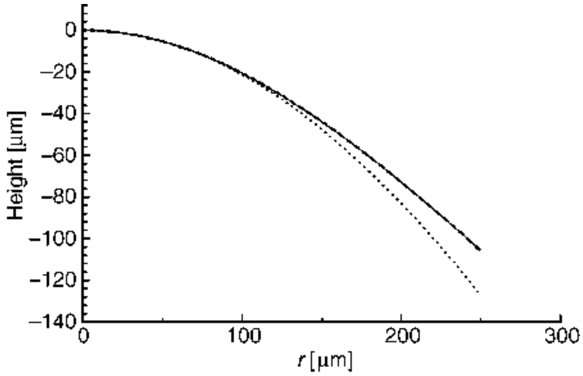

Figure 8 Application of the thin lens approximation for structure design leads to a spherical surface for a spherical phase-only transmission. The corresponding surface is depicted by the dotted line. If geometrical optics LPIA is used for structure design, an aspherical surface is obtained as required. The aspherical surface is depicted by the solid line. The height of the surfaces are shown as function of the radius *r*. Naturally the difference between TEA and geometrical optics LPIA occurs for larger *r* only.

form the transmission is obtained. It is not necessary to decide about the use of a diffractive or a refractive element in an early stage of the design. Moreover, it provides the designer with insight into selecting optical elements to achieve the function. If the phase-only transmission can be unwrapped, a refractive as well as a diffractive element can be used. An optical engineer has the freedom to decide about the suitable choice taking fabrication, size, costs, and other additional constraints into account. However, as soon as the phase-only transmission includes so-called vortices, unwrapping is not possible and diffractive optics is mandatory. An example of a vortex is shown in Figure 9. It turns out that vortices are more likely to appear in wave-optical engineering the less image-like the optical function is (see also last section). It is important to note, that functional design assumes a diagonal operator matrix, that is, it excludes crosstalk of the two field channels. Transferring the system into a structural form may violate this assumption because of vectorial effects in the nonparaxial domain. For instance, geometrical optics LPIA in general does not correspond to a diagonal operator. Therefore, it is necessary to analyze the final system for $\mathbf{f}_{\mathrm{in}}$ – even in the case of linear polarized light – using more accurate propagation techniques than S_{TEA} if it is not ensured that the paraxial approximation is accurate enough.

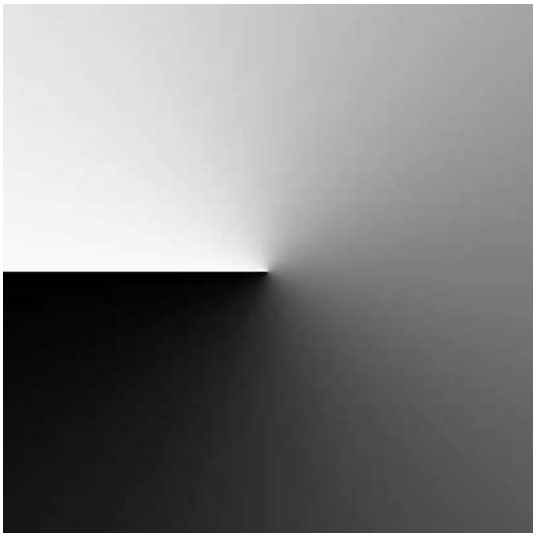

Figure 9 Illustration of a single vortex of charge one. The phase varies from 0 to 2π. Because of the vortex, this phase is not unwrappable, that means a smooth phase function cannot be obtained. The transition in the phase function requires a microstructure in the profile of the element, which generates this phase-only transmission.

The so-called thin lens, or ideal lens, in conventional optical engineering constitutes a special case of the two-step design strategy. In the first design step, distances and lens focal lengths are determined, without taking into account the actual shape of a lens. In the second step lenses, or even lens systems, are inserted according to the position and the focal length specified in the first step. Thus, wave-optical engineering extends our understanding by broadening established methods in optical engineering.

Fundamental Systems

A smart approach to design minimizes the number of elements required to achieve a desired light transformation. Thus, it is logical to try a functional design with a system of the form:

$$\begin{aligned} U_{\mathrm{out}}(x,y) &= S_{\mathrm{functional}} U_{\mathrm{in}}(x,y) \\ &= \mathcal{P}(n_{\mathrm{free}}, \Delta z_{\mathrm{out}}) \exp[\mathrm{i}\tau(x,y)] \\ &\quad \times \mathcal{P}(n_{\mathrm{free}}, \Delta z_{\mathrm{in}}) U_{\mathrm{in}}(x,y) \end{aligned} \qquad [13]$$

which is illustrated in Figure 10. One phase-only transmission, properly placed between the input and the output plane, should realize the desired transformation of the function. Scalar terminology is used, because both components are decoupled in the functional embodiment.

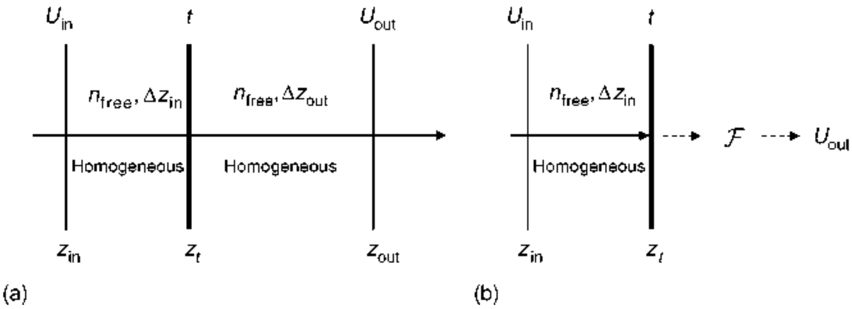

Figure 10 A basic system for light transformation can be described functionally by two propagations through homogeneous dielectrics with refractive index n_{free} and one phase-only transmission function. (b) If the second propagation is replaced by an operation in functional form, here the Fourier transform, the setup shown on the right side results.

If the output field is completely specified by the merit function, for instance: if the complex amplitude of a laser beam should be transformed into another one, the so-called amplitude matching strategy is a very powerful technique to solve the design problem stated by eqn [13]. In amplitude matching one determines the Distances Δz_{in} and Δz_{out} such that the inverse propagated U_{out} and the forward propagated U_{in} possess identical magnitudes. If such distances exist, the transformation can be realized using a phase-only transmission in the plane z_t. The phase simply follows from the ratio of the propagated fields. If no distances exist where amplitude matching occurs, the design problem has no solution and at least one more transmission function must be introduced.

As mentioned above, in numerous applications the output field is not specified completely. In particular, phase is often a free parameter. Then, an even simpler system approach than the one suggested by eqn [13], often leads to reasonable solution. Instead of $\mathcal{P}(n_{\mathrm{free}}, \Delta z_{\mathrm{out}})$, a Fourier transform is assumed, then if,

$$\begin{aligned} U_{\mathrm{out}}(k_x, k_y) &= S_{\mathrm{functional}} U_{\mathrm{in}}(x, y) \\ &= \mathcal{F}\{\exp[i\tau(x, y)]\mathcal{P}(n_{\mathrm{free}}, \Delta z_{\mathrm{in}}) U_{\mathrm{in}}(x, y)\} \qquad [14] \end{aligned}$$

The output field is expressed in k-coordinates as long as the system is discussed in functional form. For a functional design the distance Δz_{in} is of no concern, however it must be chosen together with the structure design as discussed below. Thus, in functional design, one concentrates only on the design of the transmission function. The iterative Fourier transform algorithm (IFTA), an extremely powerful algorithm, has been developed to accomplish this design. Its principle is illustrated in Figure 11. The algorithm sequentially applies constraints on the transmission and in the Fourier plane. Properly controlled, the algorithm converges to a Fourier pair for which the transmission plane constraint is satisfied completely and the Fourier domain constraint is satisfied as accurately as required by the merit function. The initial distribution is of enormous importance for a successful transmission design. Typically the initial distribution is defined in the Fourier plane as shown in Figure 11. A field $U_{\mathrm{sig}}(k_x, k_y)$ is chosen, which satisfies the merit function. It is called the signal field. Because of design freedoms such as phase, typically there exists a set of signal fields $\mathsf{F}_{\mathrm{sig}}$. If phase freedom is available, any field with the desired amplitude and an arbitrary phase is an element of $\mathsf{F}_{\mathrm{sig}}$ and can be used as initial distribution of the IFTA. Dependent on the application, the choice of the phase is of fundamental concern. In the following we assume phase freedom is available, but IFTA is also useful if this is not the case.

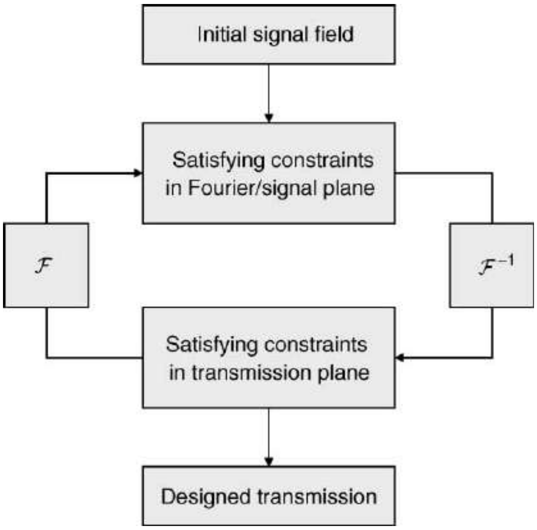
Initial signal field
Satisfying constraints in Fourier/signal plane
$\mathcal{F}$
$\mathcal{F}^{-1}$
Satisfying constraints in transmission plane
Designed transmission

Figure 11 Illustration of the basic principle of the iterative Fourier transform algorithm (IFTA). By a sequential satisfaction of constraints in the signal and the transmission plane, the design in the functional embodiment can be accomplished. If both constraints are satisfied accurate enough, the IFTA is terminated with a final satisfaction of the transmission plane constraints. The depicted IFTA version can be generalized by replacing $\mathcal{F}$ by $\mathcal{P}$ or any S.

It is helpful to distinguish between two types of signal fields. If amplitudes for only a finite number of discrete k-values, (i.e., angles or directions), need to be determined, the transmission acts as a diffractive beamsplitter. In all other cases more general light transformations are concerned.

The constraints in the transmission plane follow from the structure design. In most cases the structure design is done by TEA according to eqn [12]. Then, phase-only transmissions are required. Typically the fabrication of the corresponding element profile demands a phase quantization in addition. Both constraints have been intensively investigated for the IFTA.

As mentioned above, practically signal fields are always restricted to one or more signal regions of finite extent. For example, a detector has finite extent. Outside of these regions the signal may be unrestricted or only partially restricted. This freedom in amplitude outside the signal regions is fundamental for a successful design of phase-only transmissions. It has been shown, that the only phase-only function whose Fourier transform exists over a finite area is a linear phase. The transform of all other phase-only functions contribute energy outside of the signal regions. In other words, the use of amplitude freedom is mandatory. Energy losses due to stray light outside the signal region cannot be avoided. As a result, it is not possible to generate the desired response with 100% efficiency. For any given signal field U_{sig} an upper bound $\eta_{\text{bound}}(U_{\text{sig}})$ of the achievable efficiency $\eta(U_{\text{sig}})$ of light transformation with a phase-only transmission has been derived, that is $\eta(U_{\text{sig}}) \leq \eta_{\text{bound}}(U_{\text{sig}})$. The equal sign is only valid for the special case of a linear phase transmission, for which the signal field consists of one k-direction only. Then $\eta_{\text{bound}}(U_{\text{sig}}) = \eta(U_{\text{sig}}) = 1$ is obtained. After structure design for a linear phase-only transmission with TEA, a so-called blazed grating results. If TEA is a sufficient approximation one concludes that a blazed grating has a theoretical efficiency of 100%. For all other signals the theoretical efficiency is smaller than 100%. Moreover, the upper bound is a function of U_{sig} and thus also of its phase. Therefore, in a first step of a transmission design it is recommendable to perform a phase synthesis which maximizes the upper bound. This first design step delivers information about achieveable efficiency and a suitable initial signal field for the IFTA.

If fabrication requires phase quantization, amplitude freedom can be used to separate quantization noise from the signal regions. Unfortunately this reduces the achievable efficiency even more. If the phase to be quantized is homogenously distributed, the upper bound for a phase-only transmission quantized to Q equidistant phase levels is given by:

$$\eta_{\text{bound}}(U_{\text{sig}}, Q) = \text{sinc}^2\left[\frac{1}{Q}\right]\eta_{\text{bound}}(U_{\text{sig}}) \qquad [15]$$

For the special case of linear phase, the well-known formula $\eta(Q) = \text{sinc}^2[1/Q]$ follows. The values of $\text{sinc}^2[1/Q]$ and therefore of quantized blazed gratings in the paraxial approximation are 0.405, 0.812, 0.950, 0.987, and 0.997 for $Q = 2, 4, 8, 16$, and 32, respectively. This result does not mean that, for example, a diffractive beamsplitter with a splitting ratio $1 : M_{\text{beams}}$ and 16 height levels will provide ~99% efficiency. According to eqn [15], the theoretical values are smaller and depend on M_{beams} because $\eta_{\text{bound}}(U_{\text{sig}})$ depends via U_{sig} on M_{beams}. If fabrication processes and the inherent imperfections are taken into consideration, the values decrease even more. As a rule of thumb, it is possible to claim that efficiencies larger than 90% are always challenging to obtain. With sophisticated design and fabrication techniques, efficiencies larger than 80% are reasonable.

Up to now, we have discussed only the importance of the signal field phase to efficiency. But the phase of the initial signal field has another extremely important effect on the result of the transmission design. From mapping-type considerations, a smooth initial signal field phase can be constructed in order to realize a light transformation. An example is illustrated in Figure 12. It shows the central part of the phase of the transmission function which transforms a Gaussian input beam into a flat-top line beam. Because of the smooth signal phase also the transmission phase is smooth and therefore unwrappable. This approach is typical for a so-called beam shaping design. It requires a well-defined input beam and high accuracy in adjustment of the beam shaping optics to obtain the output field with good quality. If, instead of a smooth initial phase, a random one is applied, the situation illustrated in Figure 13 (left) is achieved. Here the phase of the transmission looks random, which is typical for beamsplitter and diffuser. Because such phase functions possess numerous vortices, they do not allow an unwrapping. Thus, diffractive elements are the logical result of structure design. An example is shown in Figure 13 (right).

Figure 14 shows how light from a laser diode passes through the diffuser element of Figure 13 (right). The radiation is transformed into an intensity pattern that depicts the 'skyline' of Jena. Although globally the intensity possesses the requested shape, a closer look shows that the intensity distribution is speckled. However, in numerous applications this is not a problem so long as the size of the speckles is well adapted to the resolution of the detector.

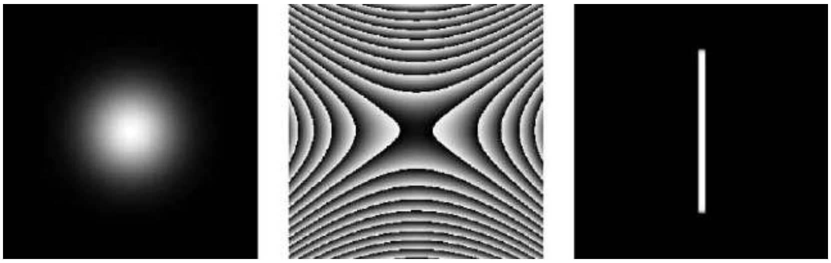

Figure 12 The input beam on the left propagates through the transmission function the central part of the phase of which is shown in the middle and in the signal plane the amplitude shown on the right results. This Gaussian-to-line transformation is a typical beam shaping design task.

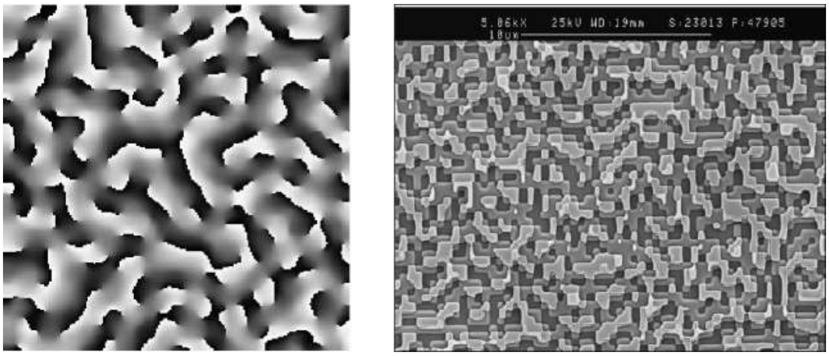

Figure 13 A typical phase distribution of a transmission (left) which results from design with IFTA for a random initial signal field phase. Numerous vortexes prevent its unwrapping. Therefore, structure design leads to diffractive elements. Then, a quantization of the phase is demanded in most cases. An example is shown on the right-handed side. It depicts a portion of a four level diffuser which results in the output field shown in **Figure 14**. The pixel size is 400 nm and the corresponding element was designed and fabricated in a cooperation of LightTrans GmbH and the Institute of Applied Physics at the University of Jena.

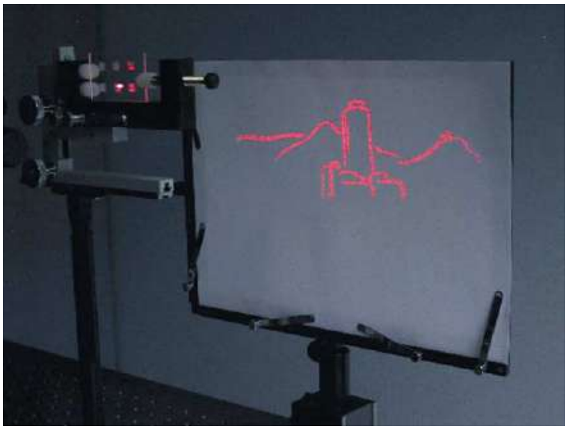

Figure 14 Optical demonstration of the use of a diffuser (see in Figure (right)) transforming a laser diode beam into an intensity displaying the ‘skyline’ of Jena. No additional optical element is needed! Stray light appears but is laterally separated from the desired signal and thus it is easy to be filtered.

This technique is called diffuser-type light transformation. The appearance of the global intensity distribution is robust against variations in the input beam and thus is well-suited also for partial coherent illumination, for example radiation from an excimer laser and LED. In that case, even the appearance of speckles is reduced or vanishes completely. The mapping-type and the diffuser-type approach to choose the initial signal phase constitute two extreme cases of a great variety to control basic aspects of light transformation using different initial phases.

After choosing a suitable initial signal field $U_{\mathrm{sig}}(k_x, k_y)$ the IFTA typically performs well and a transmission function is obtained, which leads to $|U_{\mathrm{out}}(k_x, k_y)|^2 \propto |U_{\mathrm{sig}}(k_x, k_y)|^2$ in the signal region with minimum error. In the next step the structural design must be performed. The structure of the element is calculated via eqn [12].

Besides designing structure to produce the desired transmission function one needs to realize optically the Fourier transformation. Placing the observation plane in the far field behind the element is the most straightforward way to achieve this. However, field Fourier transformation can also be realized by lenses as shown in Figure 15. From simple geometric considerations using paraxial approximation one obtains:

$$\alpha = \frac{2\pi n_{\mathrm{free}}}{\lambda f_{\mathrm{e}}} \qquad [16]$$

Each system produces a Fourier transform that is scaled specially by α, where the output field is considered in a dielectric with refractive index n_{free}. The effective focal length f_{e} is equal to the focal length of the lens if the element is placed in front of the lens (see top of Figure 15), the distance of the element to the lens is of no concern for the resulting intensity, however it is recommended to minimize the distance between element and lens for elements with deflection angles close to the paraxial limit. That avoids light losses due to truncation by the lens aperture. If the focal length itself is chosen as the distance of the element to the lens, a so-called $2f$-setup results, which generate, in addition to the correct magnitude, the correct phase of the Fourier transformation in the focal plane. If an alternative distance to the lens is used, an additional quadratic phase occurs in the focal plane. If the element is placed behind the lens (see middle row of Figure 15), f_{e} is the distance between element and focal plane. This setup provides some freedom in adjusting the output scaling. On the other hand, it requires an accurate adjustment of position to obtain a specific scaling. This is not required for the first setup. In order to obtain a specific

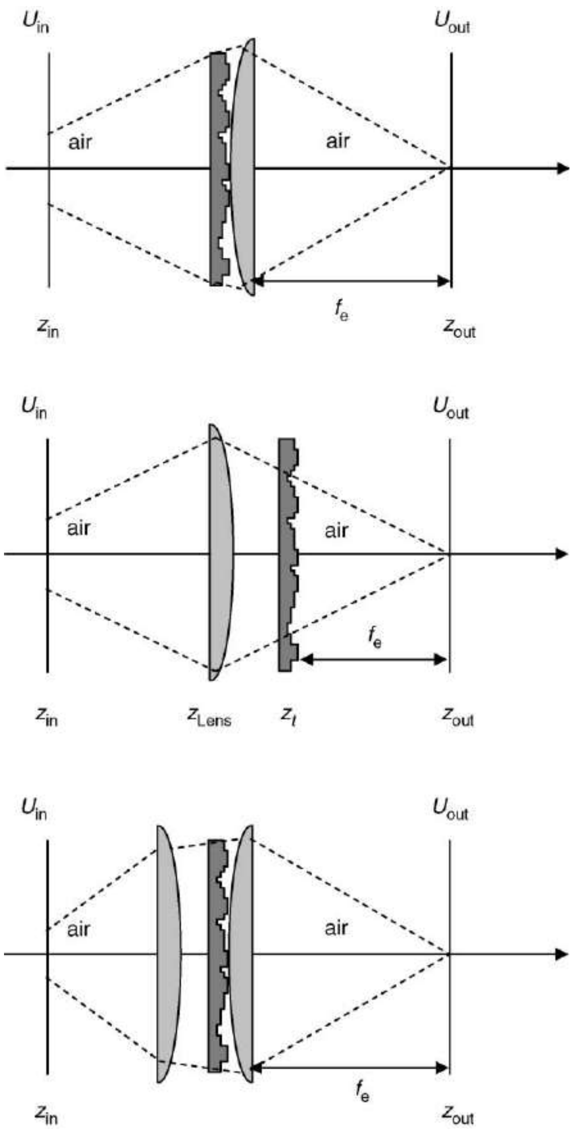

Figure 15 Three setups to realize a Fourier transformation for a beam shaping or diffuser/splitter element, which deflects the light only in a paraxial way. In all cases the output is obtained in the focal plane of the lens.

scaling of the output field $U_{\text{out}}(\alpha x, \alpha y)$, the effective focal length and the size of $U_{\text{sig}}(k_x, k_y)$ in the k-space must be chosen properly according to eqn [16]. In addition to the size of the output field, its resolution is also of fundamental importance. According to eqn [14] and the convolution theorem the output field is given by

$$U_{\text{out}}(k_x, k_y) = \mathcal{F} \exp[\mathrm{i}\tau(x, y)] * \mathcal{F}\mathcal{P}(\Delta z_{\text{in}}) U_{\text{in}}(x, y) \qquad [17]$$

where $*$ denotes the convolution. For a constant transmission function one would obtain the focus of the input beam. In functional embodiment that is $\mathcal{F}\mathcal{P}(\Delta z_{\text{in}}) U_{\text{in}}(x, y)$ only. In the structural form, the focus of the beam must be calculated by propagating the input field through the system without the transmission function. The Fourier transformed transmission function, that is $\mathcal{F} \exp[\mathrm{i}\tau(x, y)]$, is convolved with this beam focus in order to obtain the resulting output field. In beamsplitting that means a replication of the beam. In general the output field can be understood as being composed of numerous replicated beam foci which are overlapping. That makes plausible that the size of the focus of the input beam limits the resolution of the whole output field. The size of the focus is determined by the basic beam focusing lens system and not by the diffractive or refractive element. In conclusion, a highly resolved output field requires a focusing optic with high numerical aperture. For examples, it is not possible to obtain a flat-top beam profile from a Gaussian beam where focus has the same size as the flat-top should have. The steeper the sidewalls of the flat-top profile, the smaller must be the focal spot without the beam shaping element. The choice of a suitable numerical aperture of the lens system must be done before the transmission design is performed because it also determines the extent of the transmission function. The distance Δz_{in} must be chosen in such a way that ensures that the propagated input field illuminates the whole aperture of the system.

If a high numerical aperture on the output side is required to achieve a high resolution, the setup in the middle of Figure 15 should be avoided and the first setup is preferable. Otherwise the element would be illuminated by a highly nonparaxial field which violates the structure design with TEA. With the same argument, but now on the input side, an input field with high divergence should be at least partly collimated to achieve a paraxial illumination of the element. Then, the setup at the bottom of Figure 15 results.

For nonparaxial light transformation, that means the output field is large and hence the transformation requires nonparaxial deflection of the input field; the described transmission and structure design is still useful to obtain an initial design for a further optimization by wave-optical modeling techniques.

Though a Fourier approach to obtain general light transformations seems rather restricted, it covers a broad field of applications. The combination of suitable beam optics with the refractive free-form or diffractive element can be very simple (see optical setup of Figure 14) or very complex. That mainly depends on the resolution and size of the demanded output field as it is well known from imaging systems.

Each design should be finalized with a complete wave-optical analysis, typically including adjustment and fabrication error tolerancing, to achieve a full understanding of the optical properties of the system. After this final investigation it is ready for fabrication and application providing a further innovation through optics and photonics on the base of wave-optical engineering.

See also

Fourier Optics.

Further Reading

Aagedal H, Schmid M, Beth T, Teiwes S and Wyrowski F (1996) Theory of speckles in diffractive optics and its application to beam shaping. *Journal of Modern Optics* 43(7): 1409–1421.

Aagedal H, Schmid M, Egner S, Müller-Quade J, Beth T and Wyrowski F (1997) Analytical beam shaping with application to laser-diode arrays. *Journal of the Optical Society of America A* 14(7): 1549–1553.

Aagedal H, Wyrowski F and Schmid M (1997) Paraxial beamsplitting and shaping. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, chapter 6, pp. 165–188. Berlin: Akademie Verlag.

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. New York: Pergamon Press.

Bryngdahl O (1974) Optical map transformations. *Optics Communications* 10(2): 164–168.

Bühling S, Schimmel H and Wyrowski F (2003) Wave-optical engineering with virtuallab. *Proceedings SPIE*, 5182: 24–33.

Bühling S, Schimmel H and Wyrowski F (2003) Challenges in industrial applications of diffractive beam splitting. *Proceedings SPIE* 5182: 123–131.

Collier RJ, Burckhardt CB and Lin LH (1971) *Optical Holography*. London: Academic Press.

Fienup JR (1980) Iterative method applied to image reconstruction and to computer-generated holograms. *Optical Engineering* 19(3): 297–305.

Gerchberg RW and Saxton WO (1972) A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35(2): 237–246.

Goodmann JW (1968) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Goodman JW and Silvestri AM (1970) Some effects of Fourier-domain phase quantization. *IBM Journal of Reserved Development* 14: 478–484.

Krackhardt U, Mait JN and Streibl N (1992) Upper bound on the diffraction efficiency of phase-only fanout elements. *Applied Optics* 31(1): 27–37.

Lajunen H, Tervo J, Turunen J, Vallius T and Wyrowski F (2003) Simulation of light propagation by local spherical interface approximation. *Applied Optics* 34: 6804–6810.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge: Cambridge University Press.

Pfeil AV and Wyrowski F (2000) Wave-optical structure design with the local plane-interface approximation. *Journal of Modern Optics* 47(13): 2335–2350.

Pfeil AV, Wyrowski F, Drauschke A and Aagedal H (2000) Analysis of optical elements with the local plane-interface approximation. *Applied Optics* 39(19): 3304–3313.

Schimmel H, Bühling S and Wyrowski F (2003) Use of ray tracing in wave-optical engineering. *Proceedings SPIE* 5182: 6–14.

Schimmel H and Wyrowski F (2000) Amplitude matching strategy for wave-optical design of monofunctional systems. *Journal of Modern Optics* 47(13): 2295–2321.

Sergienko N, Stamnes JJ, Kettunen V, Kuittinen M, Turunen J, Vahimaa P and Friberg AT (1999) Comparison of electromagnetic and scalar methods for evaluation of diffractive lenses. *Journal of Modern Optics* 46(1): 65–82.

Smith WJ (1997) *Practical Optical System Layout*. New York: McGraw-Hill.

Smith WJ (2000) *Modern Optical Engineering*, 3rd edn. Bellingham: SPIE Press.

Turunen J (1997) Diffraction theory of microrelief gratings. In: Herzig HP (ed.) *Micro-Optics Elements, Systems and Applications*, chapter 2, pp. 31–52. London: Taylor & Francis.

Turunen J and Wyrowski F (1997) *Diffractive Optics for Industrial and Commercial Applications*. Berlin: Akademie Verlag.

Vallius T, Jefimovs K, Kettunen V, Kuittinen M, Laakkonen P and Turunen J (2001) Design of non-paraxial array illuminators by step-transition perturbation approach. *Journal of Modern Optics* 48(12): 1869–1879.

Wyrowski F (1990) Diffractive optical elements: Iterative calculation of quantized, blazed phase structures. *Journal of the Optical Society of America A* 7: 961–969.

Wyrowski F (1991) Digital phase encoded inverse filter for optical pattern recognition. *Applied Optics* 30: 4650–4657.

Wyrowski F (1991) Considerations on convolution and phase factors. *Optics Communications* 81(6): 353–358.

Wyrowski F (1991) Upper bound of the diffraction efficiency of diffractive phase elements. *Optics Letters* 16(24): 1915–1917.

Wyrowski F (1993) Design theory of diffraction elements in the paraxial domain. *Journal of the Optical Society of America A* 10: 1553–1561.

Wyrowski F (2003) Comments on wave-optical engineering. *Proceedings SPIE* 5182: 1–5.

Wyrowski F and Turunen J (2002) Wave-optical engineering. In: Guenther AH (ed.) *International Trends in Applied Optics*, chapter 21, pp. 471–496. Bellingham: SPIE Press.

DISPERSION MANAGEMENT

A E Willner, Y-W Song, J Mcgeehan and Z Pan, University of Southern California, Los Angeles, CA, USA

B Hoanca, University of Alaska Anchorage, Anchorage, AK, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical communications systems have grown explosively in terms of the capacity that can be transmitted over a single optical fiber. This trend has been fueled by two complementary techniques, those being the increase in data-rate-per-channel coupled with the increase in the total number of parallel wavelength channels. However, there are many considerations as to the total number of wavelength-division-multiplexed (WDM) channels that can be accommodated in a system, including cost, information spectral efficiency, nonlinear effects, and component wavelength selectivity.

Dispersion is one of the critical roadblocks to increasing the transmission capacity of optical fiber. The dispersive effect in an optical fiber has several ingredients, including intermodal dispersion in a multimode fiber, waveguide dispersion, material dispersion, and chromatic dispersion. In particular, chromatic dispersion is one of the critical effects in a single mode fiber (SMF), resulting in a temporal spreading of an optical bit as it propagates along the fiber. At data rates ≤ 2.5 Gbit/s, the effects of chromatic dispersion are not particularly troublesome. For data rates ≥ 10 Gbit/s, however, transmission can be tricky and the chromatic dispersion-induced degrading effects must be dealt with in some way, perhaps by compensation. Furthermore, the effects of chromatic dispersion rise quite rapidly as the bit rate increases – when the bit rate increases by a factor of four, the effects of chromatic dispersion increase by a factor of 16! This article will only deal with the management of chromatic dispersion in single mode fiber.

One of the critical limitations of optical fiber communications comes from chromatic dispersion, which results in a pulse broadening as it propagates along the fiber. This occurs as photons of different frequencies (created by the spreading effect of data modulation) travel at different speeds, due to the frequency-dependent refractive index of the fiber core. Compounding the problems cause by chromatic dispersion is the fact that as bit rates rise, chromatic dispersion effects rise quadratically with respect to the increase in the bit rate. One can eliminate these effects using fiber with zero chromatic dispersion, known as dispersion shifted fiber (DSF). However, with zero dispersion, all channels in a wavelength-division-multiplexed (WDM) system travel at the same speed, in-phase, and a number of deleterious nonlinear effects such as cross-phase modulation (XPM) and four-wave mixing (FWM) result. Thus, in WDM systems, some amount of chromatic dispersion is necessary to keep channels out-of-phase, and as such chromatic dispersion compensation is required. Any real fiber link may also suffer from 'dispersion slope' effects, in which a slightly different dispersion value is produced in each WDM channel. This means that while one may be able to compensate one channel exactly, other channels may progressively accumulate increasing amounts of dispersion, which can severely limit the ultimate length of the optical link and the wavelength range that can be used in a WDM system. This article will address the concepts of chromatic dispersion and dispersion slope management followed by some examples highlighting the need for tunability to enable robust optical WDM systems in dynamic environments. Some dispersion monitoring techniques are then discussed and examples given.

Chromatic Dispersion in Optical Fiber Communication Systems

In any medium (other than vacuum) and in any waveguide structure (other than ideal infinite free space), different electromagnetic frequencies travel at different speeds. This is the essence of chromatic dispersion. As the real fiber-optic world is rather distant from the ideal concepts of both vacuum and infinite free space, dispersion will always be a concern when one is dealing with the propagation of electromagnetic radiation through fiber. The velocity in fiber of a single monochromatic wavelength is constant. However, data modulation causes a broadening of the spectrum of even the most monochromatic laser pulse. Thus, all modulated data have a nonzero spectral width which spans several wavelengths, and the different spectral components of modulated data travel at different speeds. In particular, for digital data intensity modulated on an optical carrier, chromatic dispersion leads to pulse broadening – which in turn leads to chromatic dispersion limiting the maximum data rate that can be transmitted through optical fiber (see Figure 1).

Considering that the chromatic dispersion in optical fibers is due to the frequency-dependent nature of the propagation characteristics, for both the material (the refractive index of glass) and the waveguide structure, the speed of light of a particular wavelength λ will be expressed as follows, using a Taylor series expansion of the value of the refractive index as a function of the wavelength:

$$v(\lambda) = \frac{c_0}{n(\lambda)} = \frac{c_0}{n_0(\lambda_0) + \dfrac{\partial n}{\partial \lambda}\delta\lambda + \dfrac{\partial^2 n}{\partial \lambda^2}(\delta\lambda)^2} \qquad [1]$$

Here, c_0 is the speed of light in vacuum, λ_0 is a reference wavelength, and the terms in $\partial n/\partial\lambda$ and $\partial^2 n/\partial\lambda^2$ are associated with the chromatic dispersion and the dispersion slope (i.e., the variation of the chromatic dispersion with wavelength), respectively. Transmission fiber has positive dispersion, i.e. longer wavelengths result in longer propagation delays. The units of chromatic dispersion are picoseconds per nanometer per kilometer, meaning that shorter time pulses, wider frequency spread due to data modulation, and longer fiber lengths will each contribute linearly to temporal dispersion. Higher data rates inherently have both shorter pulses and wider frequency spreads. Therefore, as network speed increases, the impact of chromatic dispersion rises precipitously as the square of the increase in data rate. The quadratic increase with the data rate is a result of two effects, each with a linear contribution. On one hand, a doubling of the data rate makes the spectrum twice as wide, doubling the effect of dispersion. On the other hand, the same doubling of the data rate makes the data pulses only half as long (hence twice as sensitive to dispersion). The combination of a wider signal spectrum and a shorter pulse width is what leads to the overall quadratic impact. Moreover, the data modulation format used can significantly affect the sensitivity of a system to chromatic dispersion. For example, the common nonreturn-to-zero (NRZ) data format, in which the optical power stays high throughout the entire time slot of a '1' bit, is more robust to chromatic dispersion than is the return-to-zero (RZ) format, in which the optical power stays high in only part of the time slot of a '1' bit. This difference is due to the fact that RZ data have a much wider channel frequency spectrum compared to NRZ data, thus incurring more chromatic dispersion. However, in a real WDM system, the RZ format increases the maximum allowable transmission distance by virtue of its reduced duty cycle (compared to the NRZ format), making it less susceptible to fiber nonlinearities as can be seen in Figure 2.

A rule for the maximum distance over which data can be transmitted is to consider a broadening of the pulse equal to the bit period. For a bit period B, a dispersion value D and a spectral width $\Delta\lambda$, the dispersion-limited distance is given by

$$L_{\mathrm{D}} = \frac{1}{D \cdot B \cdot \Delta\lambda} = \frac{1}{D \cdot B \cdot (cB)} \propto \frac{1}{B^2} \qquad [2]$$

(see Figure 3). For example, for single mode fiber, $D = 17$ ps/nm/km, so for 10 Gbit/s data the distance is $L_{\mathrm{D}} = 52$ km. In fact, a more exact calculation

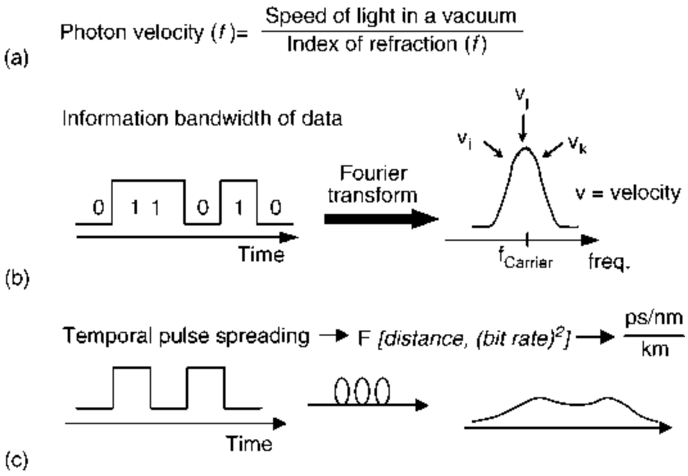
Photon velocity (f)= Speed of light in a vacuum / Index of refraction (f)
(a)
Information bandwidth of data
0
1 1
0
1
0
Time
Fourier transform
v_i
v_j
v_k
v = velocity
$f_{Carrier}$
freq.
(b)
Temporal pulse spreading → F [distance, (bit rate)²] → ps/nm / km
000
Time
(c)

Figure 1 The origin of chromatic dispersion in data transmission. (a) Chromatic dispersion is caused by the frequency-dependent refractive index in fiber. (b) The nonzero spectral width due to data modulation. (c) Dispersion leads to pulse broadening, proportional to the transmission distance and the data rate.

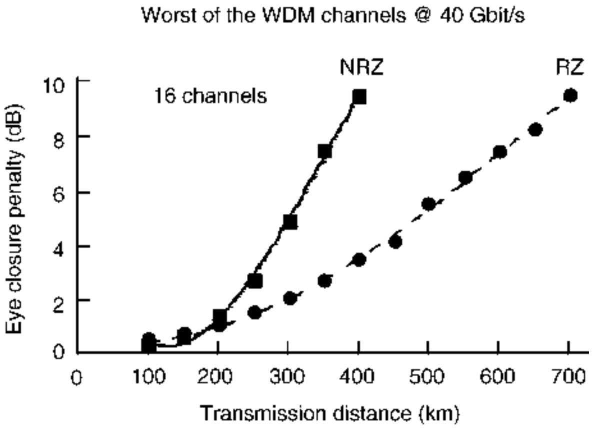

Figure 2 Performances of RZ and NRZ formats in a real fiber transmission link. (Reproduced with permission from Hayee I and Willner AE (1999) NRZ versus RZ in 10–40-Gb/s dispersion-managed WDM transmission systems. *IEEE Photon. Tech. Lett.* 11(8): 991–993.)

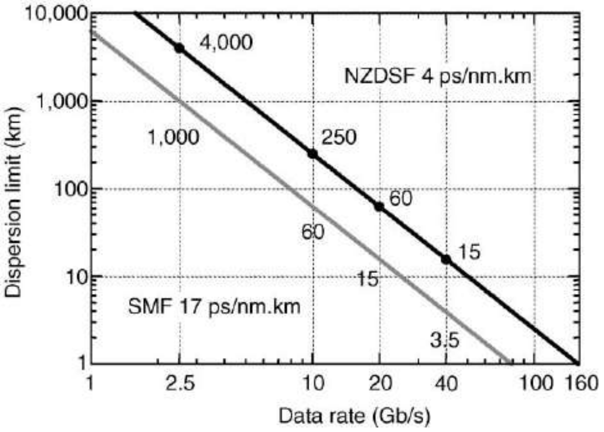

Figure 3 Transmission distance limitations due to uncompensated dispersion in SMF as a function of data rate for intensity modulated optical signals. (Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference.*)

shows that for 60 km, the dispersion induced power penalty is less than 1 dB (see Figure 4). The power penalty for uncompensated dispersion rises exponentially with transmission distance, and thus to maintain good signal quality, dispersion compensation is required.

Chromatic Dispersion Management

Optical Nonlinearities as Factors to be Considered in Dispersion Compensation

Even though it is possible to manufacture fiber with zero dispersion, it is not practical to use such fiber for WDM transmission, due to large penalties induced by fiber nonlinearities. Most nonlinear effects originate from the nonlinear refractive index of fiber, which is not only dependent on the frequency of light but also on the intensity (optical power), and is related to the optical power as:

$$\bar{n}(f, P) = n(f) + n_2 \frac{P}{A_{\mathrm{eff}}} \qquad [3]$$

where $n(f)$ is the linear part of the refractive index, P is the optical power inside the fiber, and n_2 is the nonlinear-index coefficient for silica fibers. The typical value of n_2 is 2.6×10^{-20} $\mathrm{m^2/W}$. This number takes into account the averaging of the polarization states of the light as it travels in the fiber. The intensity dependence of the refractive index gives rise to three major nonlinear effects.

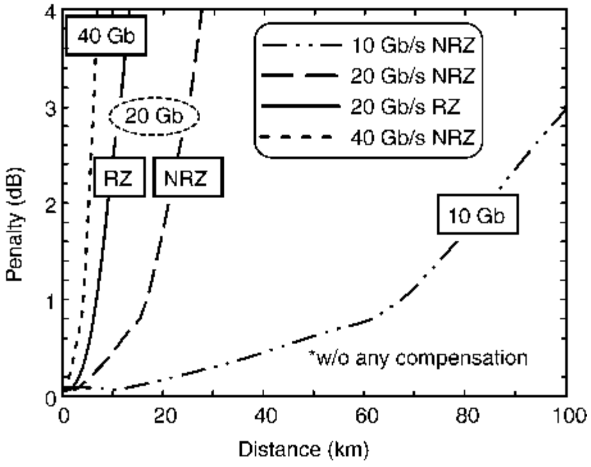

Figure 4 Power penalties due to uncompensated dispersion in single mode fiber (SMF) as a function of distance and data rate. (Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference.*)

Self-phase modulation (SPM)

A million photons 'see' a different glass than does a single photon, and a photon traveling along with many other photons will slow down. SPM occurs because of the varying intensity profile of an optical pulse on a single WDM channel. This intensity profile causes a refractive index profile and, thus, a photon speed differential. The resulting phase change for light propagating in an optical fiber is expressed as:

$$\Phi_{\mathrm{NL}} = \gamma P L_{\mathrm{eff}} \qquad [4]$$

where the quantities γ and L_{eff} are defined as:

$$\gamma = \frac{2\pi n_2}{\lambda A_{\mathrm{eff}}} \text{ and } L_{\mathrm{eff}} = \frac{1 - e^{-\alpha L}}{\alpha} \qquad [5]$$

where A_{eff} is the effective mode area of the fiber and α is the fiber attenuation loss. L_{eff} is the effective nonlinear length of the fiber that accounts for fiber loss, and γ is the nonlinear coefficient measured in rad/km/W. A typical range of values for γ is between 10–30 rad/km/W. Although the nonlinear coefficient is small, the long transmission lengths and high optical powers, that have been made possible by the use of optical amplifiers, can cause a large enough nonlinear phase change to play a significant role in state-of-the-art lightwave systems.

Cross-phase modulation (XPM)

When considering many WDM channels co-propagating in a fiber, photons from channels 2 through N can distort the index profile that is experienced by channel 1. The photons from the other channels 'chirp' the signal frequencies on channel 1, which will interact with fiber chromatic dispersion and cause temporal distortion. This effect is called cross-phase modulation. In a two-channel system, the frequency chirp in channel 1, due to power fluctuation within both channels, is given by

$$\Delta B = \frac{\mathrm{d}\Phi_{\mathrm{NL}}}{\mathrm{d}t} = \gamma L_{\mathrm{eff}}\frac{\mathrm{d}P_1}{\mathrm{d}t} + 2\gamma L_{\mathrm{eff}}\frac{\mathrm{d}P_2}{\mathrm{d}t} \qquad [6]$$

where, $\mathrm{d}P_1/\mathrm{d}t$ and $\mathrm{d}P_2/\mathrm{d}t$ are the time derivatives of the pulse powers of channels 1 and 2, respectively. The first term on the right-hand side of the above equation is due to SPM, and the second term is due to XPM. Note that the XPM-induced chirp term is double that of the SPM-induced chirp term. As such, XPM can impose a much greater limitation on WDM systems than can SPM, especially in systems with many WDM channels.

Four-wave-mixing (FWM)

The optical intensity propagating through the fiber is related to the electric field intensity squared. In a WDM system, the total electric field is the sum of the electric fields of each individual channel. When squaring the sum of different fields, products emerge that are beat terms at various sum and difference frequencies to the original signals. Figure 5 depicts that if a WDM channel exists at one of the four-wave-mixing beat-term frequencies, then the beat term will interfere coherently with this other WDM channel and potentially destroy the data.

Dispersion Maps

While zero-dispersion fiber is not a good idea, a large value of the accumulated dispersion at the end of a fiber link is also undesirable. An ideal solution is to have a 'dispersion map,' alternating sections of positive and negative dispersion as can be seen in Figure 6. This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, eliminating FWM and XPM, but the total dispersion at the end of the fiber link is zero, so that no pulse broadening is induced (Table 1). The most advanced systems require periodic dispersion compensation, as well as pre- and post-compensation (before and after the transmission fiber).

The addition of negative dispersion to a standard fiber link has been traditionally known as 'dispersion compensation,' however, the term 'dispersion management' is more appropriate. SMF has positive dispersion, but some new varieties of nonzero dispersion-shifted fiber (NZDSF) come in both positive and negative dispersion varieties. Some examples are shown in Figure 7. Reverse dispersion fiber is also now available, with a large dispersion

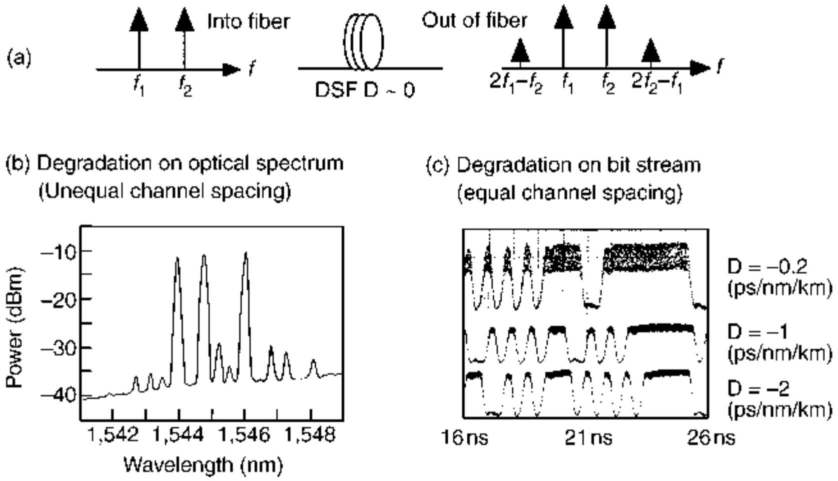

Figure 5 (a) and (b) FWM induces new spectral components via nonlinear mixing of two wavelength signals. (c) The signal degradation due to FWM products falling on a third data channel can be reduced by even small amounts of dispersion. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghieri F, Gnauck AH and Derosier RM (1995) Four-photon mixing and high-speed WDM systems. *Journal of Photon Technology* 13(5): 841–849.)

comparable to that of SMF, but with the opposite sign. When such flexibility is available in choosing both the magnitude and sign of the dispersion of the fiber in a link, dispersion-managed systems can be fully optimized to the desired dispersion map using a combination of fiber and dispersion compensation devices (see Figure 8).

Dispersion is a linear process, so first-order dispersion maps can be understood as linear systems. However, the effects of nonlinearities cannot be ignored, especially in WDM systems, with many tens of channels, where the launch power may be very high. In particular, in systems deploying dispersion compensating fiber (DCF), the large nonlinear coefficient of the DCF can dramatically affect the dispersion map.

Corrections to Linear Dispersion Maps

Chromatic dispersion is a necessity in WDM systems, to minimize the effects of fiber nonlinearities. A chromatic dispersion value as small as a few ps/nm/km is usually sufficient to make XPM and FWM negligible. To mitigate the effects of nonlinearities but maintain small amounts of chromatic dispersion, NZDSF is commercially available. Due to these nonlinear effects, chromatic dispersion must be managed, rather than eliminated.

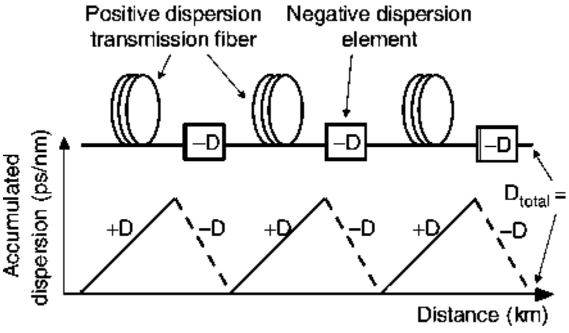
Positive dispersion transmission fiber
Negative dispersion element
–D
–D
–D
Accumulated dispersion (ps/nm)
+D
–D
+D
–D
+D
–D
D_total = 0
Distance (km)

Figure 6 Dispersion map of a basic dispersion managed system. Positive dispersion transmission fiber alternates with negative dispersion compensation elements such that the total dispersion is zero end-to-end.

If a dispersion-management system was perfectly linear, it would be irrelevant whether the dispersion along a path is small or large, as long as the overall dispersion is compensated to zero (end to end). Thus, in a linear system the performance should be similar, regardless of whether the transmission fiber is SMF, and dispersion compensation modules are deployed every 60 km, or the transmission fiber is NZDSF (with approximately a quarter of the dispersion value of SMF) and dispersion compensation modules are deployed every 240 km. In real life, optical nonlinearities are very important, and recent results seem to favor the use of large, SMF-like, dispersion values in the transmission path and correspondingly high dispersion compensation devices. A recent study of performance versus channel spacing showed that the capacity of SMF could be more than four times that of NZDSF. This is because the nonlinear coefficients are much higher in NZDSF than in SMF, and for dense WDM the channel interactions become a limiting factor. A critical conclusion is that not all dispersion compensation maps are created equal: a simple calculation of the dispersion compensation, to cancel the overall dispersion value, does not lead to optimal dispersion map designs.

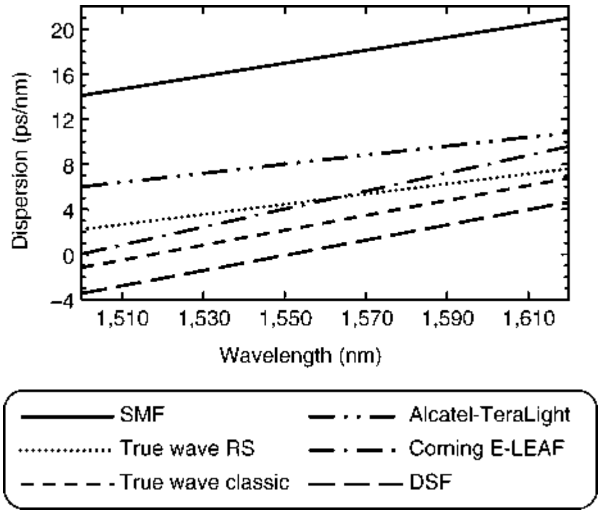

Figure 7 Chromatic dispersion characteristics of various commercially available types of transmission fiber.

Table 1 Commercially available fibers and their characteristics

	Dispersion @ 1,550 nm [ps/nm/km]	*Dispersion slope [ps/nm^2/km]*	*Attenuation [dB]*	*Mode field diameter [μm]*	*PMD [ps/km$^{0.5}$]*
SMF	18	0.08	≤0.25	11	≤0.5
Tera Light	8.0	0.058	≤0.2	9.2	≤0.04
TW-RS	4.4	0.043	≤0.25	8.4	≤0.03
LEAF	4.0	0.085	≤0.25	9.6	≤0.08
Standard DCF	−90	−0.22	≤0.5	5.2	≤0.08

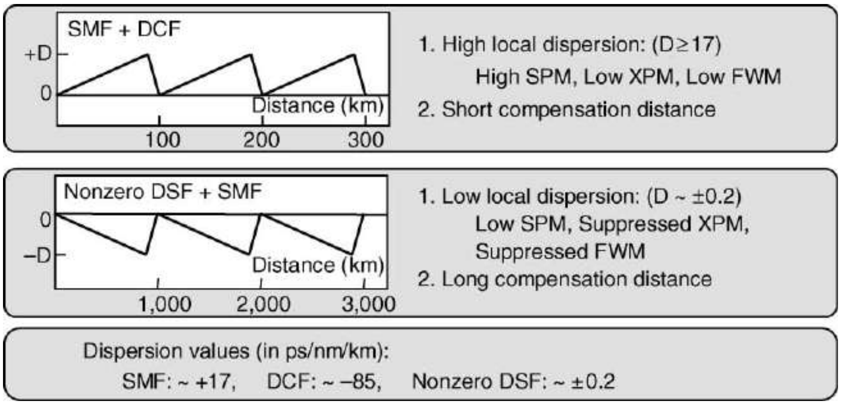
SMF + DCF
+D
0
Distance (km)
100
200
300
1. High local dispersion: (D≥17)
High SPM, Low XPM, Low FWM
2. Short compensation distance
Nonzero DSF + SMF
0
−D
Distance (km)
1,000
2,000
3,000
1. Low local dispersion: (D ~ ±0.2)
Low SPM, Suppressed XPM,
Suppressed FWM
2. Long compensation distance
Dispersion values (in ps/nm/km):
SMF: ~ +17, DCF: ~ −85, Nonzero DSF: ~ ±0.2

Figure 8 Various dispersion maps for SMF-DCF and NZDSF-SMF.

Additionally, several solutions have been shown to be either resistant to dispersion, or have been shown to rely on dispersion itself for transmission. Such solutions include chirped pulses (where prechirping emphasizes the spectrum of the pulses so that dispersion does not broaden them too much), dispersion assisted transmission (where an initial phase modulation tailored to the transmission distance leads to full-scale amplitude modulation at the receiver end due to the dispersion), and various modulation formats robust to chromatic dispersion and nonlinearities.

Dispersion Management Solutions

Fixed Dispersion Compensation

From a systems point of view, there are several requirements for a dispersion compensating module: low loss, low optical nonlinearity, broadband (or multichannel) operation, small footprint, low weight, low power consumption, and clearly low cost. It is unfortunate that the first dispersion compensation modules, based on DCF only, met two of these requirements: broadband operation and low power consumption. On the other hand, several solutions have emerged that can complement or even replace these first-generation compensators.

Dispersion compensating fiber (DCF)

One of the first dispersion compensation techniques was to deploy specially designed sections of fiber with negative chromatic dispersion. The technology for DCF emerged in the 1980s and has developed dramatically since the advent of optical amplifiers in 1990. DCF is the most widely deployed dispersion compensator, providing broadband operation and stable dispersion characteristics, and the lack of a dynamic, tunable DCF solution has not reduced its popularity.

As can be seen in Figure 9, the core of the average dispersion compensating fiber is much smaller than that of standard SMF, and beams with longer wavelengths experience relatively large changes in mode size (due to the waveguide structure) leading to greater propagation through the cladding of the fiber, where the speed of light is greater than that of the core. This leads to a large negative dispersion value. Additional cladding layers can lead to improved DCF designs that can include negative dispersion slope to counteract the positive dispersion slope of standard SMF.

In spite of its many advantages, DCF has a number of drawbacks. First, it is limited to a fixed compensation value. In addition, DCF has a weakly guiding structure and has a much smaller core cross-section, 19 μm^2, compared to the 85 μm^2 of SMF. This leads to higher nonlinearity, higher splice losses, as well as higher bending losses. Last, the length of DCF required to compensate for SMF dispersion is rather long, about one-fifth of the length of the transmission fiber for which it is compensating. Thus DCF modules induce loss, and are relatively bulky and heavy. The bulk is partly due to the mass of fiber, but also due to the resin used to hold the fiber securely in place. One other contribution to the size of the module is the higher bend loss associated with the refractive index profile of DCF; this limits the radius of the DCF loop to 6–8 inches, compared to the minimum bend radius of 2 inches for SMF.

Traditionally, DCF-based dispersion compensation modules are usually located at amplifier sites. This serves several purposes. First, amplifier sites offer relatively easy access to the fiber, without requiring any digging or unbraiding of the cable. Second, DCF has high loss (usually at least double that of standard SMF), so a gain stage is required

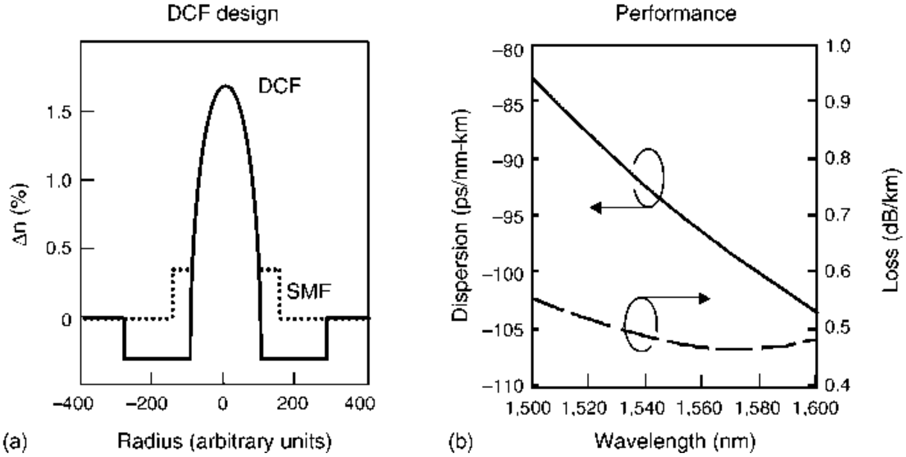

Figure 9 Typical DCF (a) refractive index profile and (b) dispersion and loss as a function of wavelength. Δ*n* is defined as refractive index variation relative to the cladding.

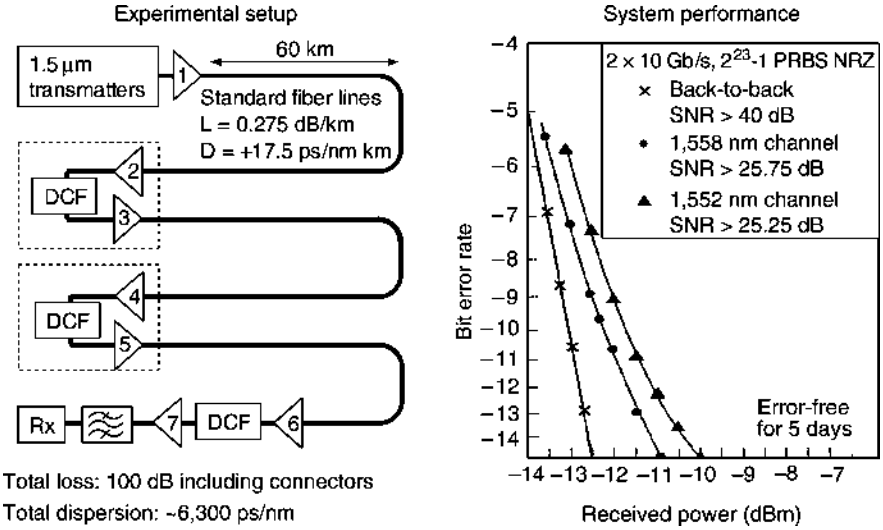

Figure 10 System demonstration of dispersion compensation using DCF. (Reproduced with permission from Park YK, Yeates PD, Delavaux J-MP, *et al.* (1995) A field demonstration of 20-Gb/s capacity transmission over 360 km of installed standard (non-DSF) fiber. *Photon. Technol. Lett.* 7(7): 816–818.)

before the DCF module to avoid excessively low signal levels. DCF has a cross-section four times smaller then SMF, hence a higher nonlinearity, which limits the maximum launch power into a DCF module. The compromise is to place the DCF in the mid-section of a two-section EDFA. This way, the first stage provides pre-DCF gain, but not to a power level that would generate excessive nonlinear effects in the DCF. The second stage amplifies the dispersion compensated signal to a power level suitable for transmission though the fiber link. This launch power level is typically much higher than could be transmitted through DCF without generating large nonlinear effects. Many newer dispersion compensation devices have better performance than DCF, in particular lower loss and lower nonlinearities. For this reason, they may not have to be deployed at the mid-section of an amplifier. Figure 10 shows the real demonstration results using the DCF.

Chirped fiber Bragg gratings

Fiber Bragg gratings have emerged as major components for dispersion compensation because of their low loss, small footprint, and low optical nonlinearity. Bragg gratings are sections of single-mode fiber in which the refractive index of the core is modulated in a periodic fashion, as a function of the spatial coordinate along the length of the fiber. When the spatial periodicity of the modulation matches what is known as a Bragg condition with respect to the wavelength of light propagating through the grating, the periodic structure acts like a mirror, reflecting the optical radiation that is traveling through the core of the fiber. An optical circulator is traditionally used to separate the reflected output beam from the input beam.

When the periodicity of the grating is varied along its length, the result is a chirped grating which can be used to compensate for chromatic dispersion. The chirp is understood as the rate of change of the spatial frequency as a function of position along the grating. In chirped gratings the Bragg matching condition for different wavelengths occurs at different positions along the grating length. Thus, the roundtrip delay of each wavelength can be tailored by designing the chirp profile appropriately. Figure 11 compares the chirped FBG with uniform FBG. In a data pulse that has been distorted by dispersion, different frequency components arrive with different amounts of relative delay. By tailoring the chirp profile such that the frequency components see a relative delay which is the inverse of the delay of the transmission fiber, the pulse can be compressed back. The dispersion of the grating is the slope of the time delay as a function of wavelength, which is related to the chirp.

The main drawback of Bragg gratings is that the amplitude profile and the phase profile as a function of wavelength have some amount of ripple. Ideally, the amplitude profile of the grating should have a flat (or rounded) top in the passband, and the phase profile should be linear (for linearly chirped gratings) or polynomial (for nonlinearly chirped gratings). The grating ripple is the deviation from the ideal profile shape. Considerable effort has been expended on reducing the ripple. While early gratings were plagued by more than 100 ps of ripple, published results have shown vast improvement to values close to ±3 ps.

Higher order mode dispersion compensation fiber

One of the challenges of designing standard DCF is that high negative dispersion is hard to achieve unless the cross-section of the fiber is small (which leads to high nonlinearity and high loss). One way to reduce both the loss and the nonlinearity is to use a higher-order mode (HOM) fiber (LP_{11} or LP_{02} near cutoff instead of the LP_{01} mode in the transmission fiber).

Such a device requires a good-quality mode converter between LP_{01} and LP_{02} to interface between the SMF and HOM fiber. HOM fiber has a dispersion per unit length greater than six times that of DCF. Thus, to compensate for a given transmission length in SMF, the length of HOM fiber required is only one sixth the length of DCF. Thus, even though losses and nonlinearity per unit length are larger for HOM fiber than for DCF, they are smaller overall, because of the shorter HOM fiber length. As an added bonus, the dispersion can be tuned slightly by changing the cutoff wavelength of LP_{02} (via temperature tuning). A soon to be released HOM fiber-based commercial dispersion compensation module is not tunable, but can fully compensate for dispersion slope.

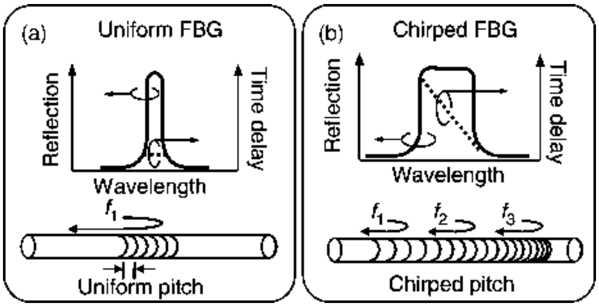

Figure 11 Uniform and chirped FBGs. (a) A grating with uniform pitch has a narrow reflection spectrum and a flat time delay as a function of wavelength. (b) A chirped FBG has a wider bandwidth, a varying time delay and a longer grating length.

Tunable Dispersion Compensation

The need for tunability

In a perfect world, all fiber links would have a known, discrete, and unchanging value of chromatic dispersion. Network operators would then deploy fixed dispersion compensators periodically along every fiber link to exactly match the fiber dispersion. Unfortunately, several vexing issues may necessitate that dispersion compensators are tunability, that they have the ability to adjust the amount of dispersion to match system requirements.

First, there is the most basic business issue of inventory management. Network operators typically do not know the exact length of a deployed fiber link nor its chromatic dispersion value. Moreover, fiber plants periodically undergo upgrades and maintenance, leaving new and nonexact lengths of fiber behind. Therefore, operators would need to keep in stock a large number of different compensator models, and even then the compensation would only be approximate. Second, we must consider the sheer difficulty of 40 Gbit/s signals. The tolerable threshold for accumulated dispersion for a 40 Gbit/s data channel is 16 times smaller than at 10 Gbit/s. If the compensation value does not exactly match the fiber to within a few percent of the required dispersion value, then the communication link will not work. Tunability is considered a key enabler for this bit rate (see Figures 12 and 13). Third, the accumulated dispersion changes slightly with temperature, which begins to be an issue for 40 Gbit/s systems and 10 Gbit/s ultra long-haul systems. In fiber, the zero-dispersion wavelength changes with temperature at a

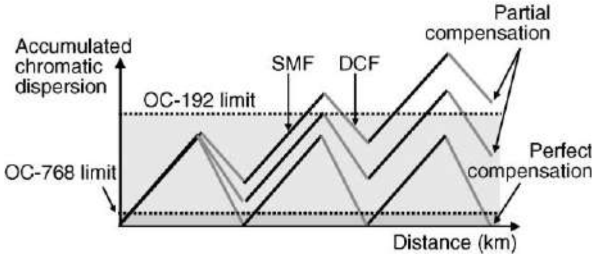
Accumulated chromatic dispersion
Partial compensation
SMF
DCF
OC-192 limit
Perfect compensation
OC-768 limit
Distance (km)

Figure 12 The need for tunability. The tolerance of OC-768 systems to chromatic dispersion is 16 times lower than that of OC-192 systems. Approximate compensation by fixed in-line dispersion compensators for a single channel may lead to rapid accumulation of unacceptable levels of residual chromatic dispersion.

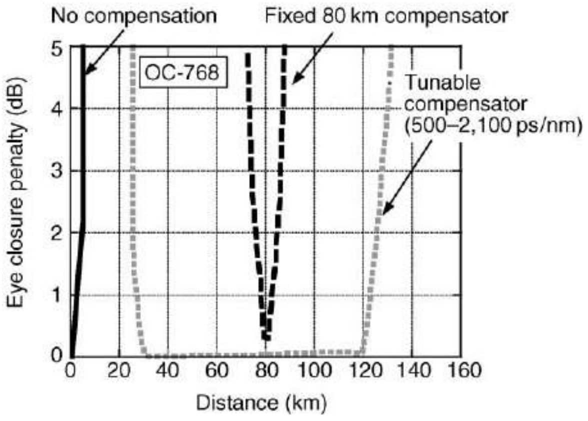

Figure 13 Tunable dispersion compensation at OC-768 (40 Gb/s) is essential for achieving a comfortable range of acceptable transmission distances (80 km for tunable, only ~4 km for fixed compensation).

typical rate of 0.03 nm/°C. It can been shown that a not-uncommon 50 °C variation along a 1,000 km 40 Gbit/s link can produce significant degradation (see Figure 14). Fourth, we are experiencing the dawn of reconfigurable optical networking. In such systems, the network path, and therefore the accumulated fiber dispersion, can change. It is important to note that even if the fiber spans are compensated span-by-span, the pervasive use of compensation at the transmitter and receiver suggests that optimization and tunability based on path will still be needed.

Other issues that increase the need for tunability include: (i) laser and (de)mux wavelength drifts for which a data channel no longer resides on the flat-top portion of a filter, thereby producing a chirp on the signal that interacts with the fiber's chromatic dispersion; (ii) changes in signal power that change both the link's nonlinearity and the optimal system dispersion map; and (iii) small differences that exist in transmitter-induced signal chirp.

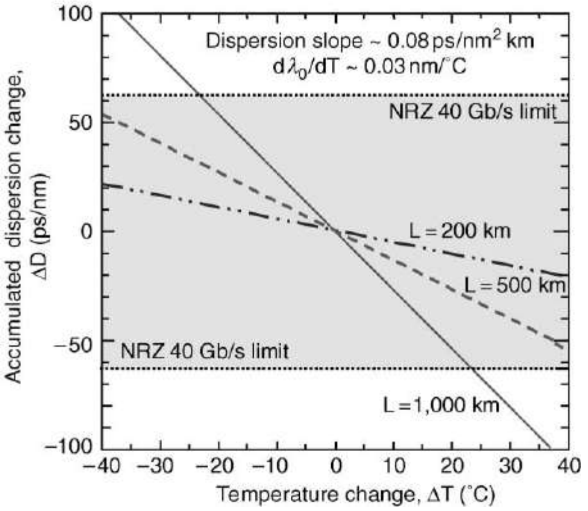

Figure 14 Accumulated dispersion changes as a function of the link length and temperature fluctuation along the fiber link.

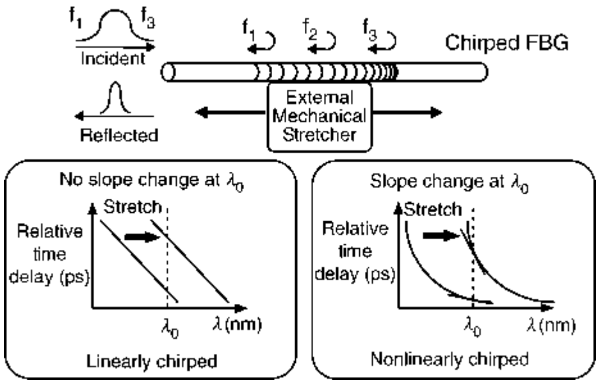
f_1
f_3
Incident
f_1
f_2
f_3
Chirped FBG
Reflected
External Mechanical Stretcher
No slope change at λ_0
Stretch
Relative time delay (ps)
λ_0
λ (nm)
Linearly chirped
Slope change at λ_0
Stretch
Relative time delay (ps)
λ_0
λ (nm)
Nonlinearly chirped

Figure 15 Tuning results for both linearly and nonlinearly chirped FBGs using uniform stretching elements. The slope of the dispersion curve at a given wavelength λ_0 is constant when the linearly chirped grating is stretched, but changes as the nonlinearly chirped grating is stretched.

Approaches to tunable dispersion compensation

A host of techniques for tunable dispersion compensation have been proposed in recent years. Some of these ideas are just interesting research ideas, but several have strong potential to become viable technologies.

Fiber gratings offer the inherent advantages of fiber compatibility, low loss, and low cost. If a FBG has a refractive-index periodicity that varies nonlinearly along the length of the fiber, it will produce a time delay that also varies nonlinearly with wavelength (see Figure 15). Herein lies the key to tunability. When a linearly chirped grating is stretched uniformly by a single mechanical element, the time delay curve is shifted towards longer wavelengths, but the slope of the ps-vs.-nm curve remains constant at all wavelengths within the passband. When a nonlinearly-chirped grating is stretched, the time delay curve is shifted toward longer wavelengths, but the slope of the ps-vs.-nm curve at a specific channel wavelength changes continuously. Ultimately, tunable dispersion compensators should accommodate multichannel operation. Several WDM channels can be accommodated by a single chirped FBG in one of two ways: fabricating a much longer (i.e., meters-length) grating, or using a sampling function when writing the grating, thereby creating many replicas of transfer function of the FBG in the wavelength domain (see Figure 16).

One free-space-based tunable dispersion compensation device is the virtually imaged phased array (VIPA), based on the dispersion of a Fabry–Perot interferometer. The design requires several lenses, a movable mirror (for tunability), and a glass plate with a thin film layer of tapered reflectivity for good mode matching. Light incident on the glass plate undergoes several reflections inside the plate. As a result, the beam is imaged at several virtual locations, with a spatial distribution that is wavelength-dependent.

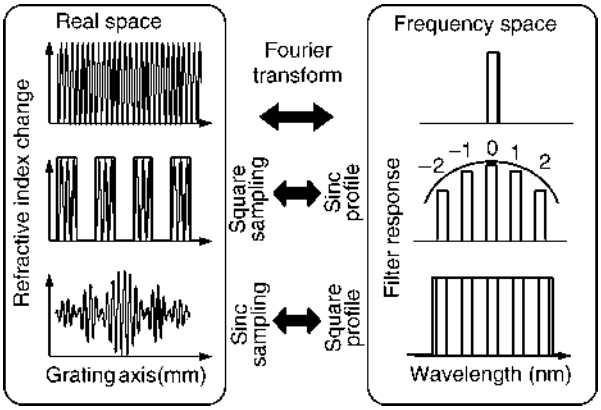

Figure 16 The concept of ‘sampled’ FBGs, where a superstructure is written on top of the grating that produces a Fourier transform in the frequency domain, leading to multiple grating passbands.

Several devices used for dispersion compensation can be integrated on a chip, using either an optical chip media (semiconductor-based laser or amplifier medium) or an electronic chip. One such technology is the micro-ring resonator, a device that, when used in a structure similar to that of an all-pass filter (see Figure 17), can be used for dispersion compensation on a chip-scale. Although these technologies are not yet ready for deployment as dispersion compensators, they have been used in other applications and have the potential to offer very high performance at low cost.

As the ultimate optical dispersion compensation devices, photonic bandgap fibers (holey fibers) are an interesting class in themselves (see Figure 18). These are fibers with a hollow structure, with holes engineered to achieve a particular functionality. Instead of being drawn from a solid preform, holey fibers are drawn from a group of capillary tubes fused together. This way, the dispersion, dispersion slope, the nonlinear coefficients could in principle all be precisely designed and controlled, up to very small or very large values, well outside the range of those of the solid fiber.

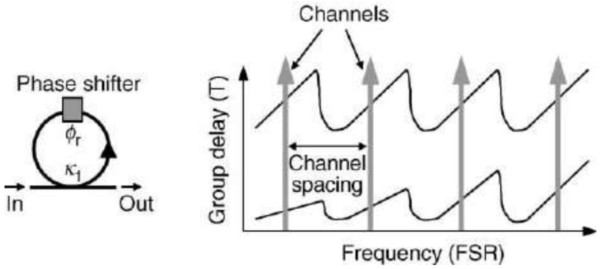

Figure 17 Architecture of an all-pass filter structure for chromatic dispersion and slope compensation. (Reproduced with permission from Madsen CK, Lenz G, Bruce AJ, *et al.* (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11(12): 1623–1625.)

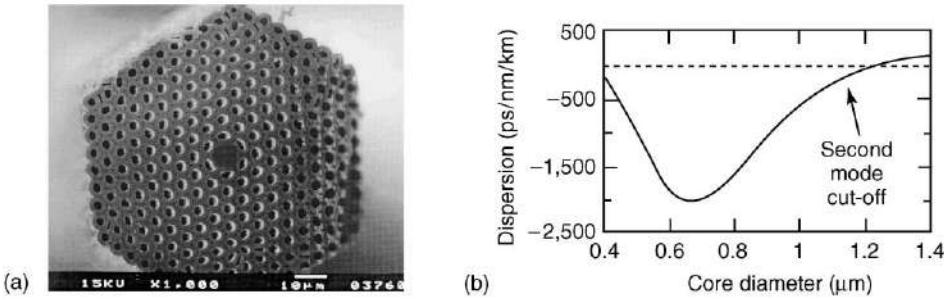

Figure 18 (a) SEM image of a photonic crystal fiber (holey fiber), and (b) net dispersion of the fiber at 1550 nm as a function of the core diameter. (Reproduced with permission from Birk TA, Mogihetsev D, Knight JC and Russell PSt.J (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11(6): 674–676.)

Dispersion Slope Mismatch

Transmission fiber, especially fiber with dispersion compensation built in, may suffer from a dispersion slope in which a slightly different dispersion value is produced for each WDM channel (see Figure 19). Even though the compensator would be able to cancel the dispersion of the fiber at the design wavelength, there will be residual dispersion left at the other wavelength channels unless the compensator can match the slope of the dispersion curve of the fiber as well. Some solutions for dispersion slope compensation are described in this section.

First, DCF, with negative dispersion slope, is a prime candidate for deployment as a dispersion slope compensator even though it cannot easily be made tunable. By designing the DCF with the same ratio of dispersion to dispersion slope as that of a real fiber link, new types of DCF can be used to compensate for both dispersion and dispersion slope, much like DCF is used for dispersion compensation today. DCF's popularity, wideband functionality, and stable dispersion characteristics make this a particularly attractive solution. Designing the DCF to match the dispersion characteristics of the transmission fiber is the critical engineering challenge in this slope compensation scheme. Second, third-order nonlinearly chirped FBG can act as a tunable dispersion slope compensator. A simple modification of the nonlinearly chirped FBG allows tuning of the compensated dispersion slope value via stretching the grating. The grating is prepared such that the time delay as a function of wavelength has a cubic profile that covers several WDM channels over a continuous bandwidth of many nanometers. Since the resulting dispersion curve is quadratic over the grating bandwidth, the dispersion slope experienced by the WDM channels can be tuned by stretching the grating using a single mechanical element (see Figure 20). Third, combining the VIPA, with either a 3D mirror or diffraction grating, can also provide tunable free-space dispersion slope compensation. Slope tuning is achieved by dynamically controlling the MEMS-based 3D mirror or the diffraction grating. Fourthly, using an FBG with many spaced thin-film heater sections can also enable tunable dispersion slope. Each heater can be individually electrically controlled, allowing the time delay profile of the grating to be dynamically tuned via changing the temperature along the length of the grating. Advanced applications of this technique can allow alterations to the entire time delay profile of the grating, providing a truly flexible dispersion and dispersion slope compensation mechanism.

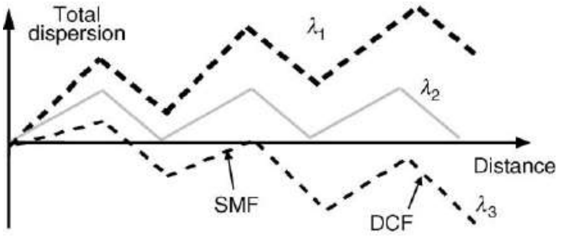

Figure 19 The chromatic dispersion slope mismatch caused by the different slopes of transmission fiber (SMF or NZDSF) and DCF.

Chromatic Dispersion Monitoring

Another important issue related to dispersion management is dispersion monitoring techniques. In a reconfigurable system, it is necessary to reconfigure

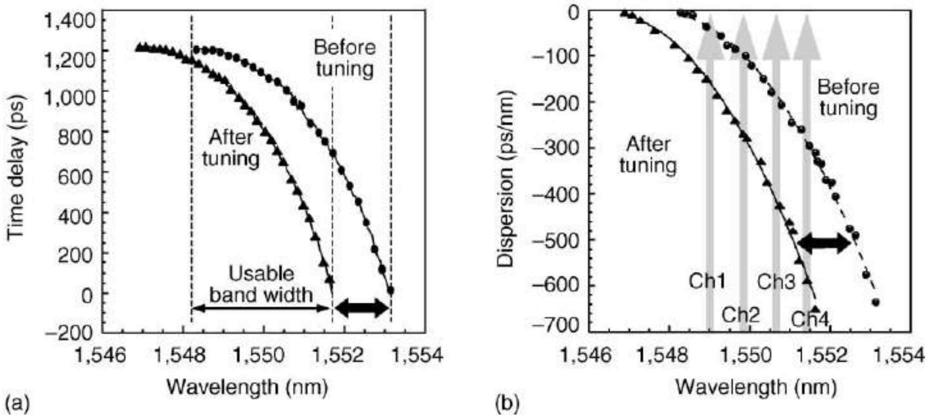

Figure 20 Tunable dispersion slope compensation using a third-order nonlinearly chirped FBG. (a) Cubic time delay curves of the grating, and (b) quadratic dispersion curves showing the change in dispersion in each channel before and after tuning. (Reproduced with permission from Song YW, Motoghian SMR, Starodubov D, *et al.* (2002) Tunable dispersion slope compensation for WDM systems using a non-channelized third-order-chirped FBG. *Optical Fiber Communication Conference* 2002. Paper ThAA4.)

any tunable chromatic dispersion compensation modules on the fly as the network changes. An in-line chromatic dispersion monitor can quickly measure the required dispersion compensation value while data are still being transmitted through the optical link. This is very different from the more traditional chromatic dispersion measurement techniques where dark fiber is used and the measurement is done off-line over many hours (or days).

Chromatic dispersion monitoring is most often done at the receiving end, where the Q-factor of the received data or some other means is employed to assess the accumulated dispersion. Existing techniques that can monitor dispersion in-line, fast, and with relatively low cost include: (i) general performance monitoring using the bit-error rate (BER) or eye opening, but this approach cannot differentiate among different degrading effects; (ii) detecting the intensity modulation induced by phase modulation at the transmitter; (iii) extracting the bit-rate frequency component (clock) from photo-detected data and monitoring its RF power (see Figure 21); (iv) inserting a subcarrier at the transmitter and subsequently monitoring the sub-carrier power degradation; (v) extracting several frequency components from the optical data using very narrow-band optical filters and detecting the optical phase; (vi) dithering the optical-carrier frequency at the transmitter and measuring the resultant phase modulation of the clock extracted at receiver with an additional phase-locked loop; and (vii) using an optical filter to select the upper and lower vestigial sideband (VSB) signals in transmitted optical data and determine the relative group delay caused by dispersion (see Figure 22).

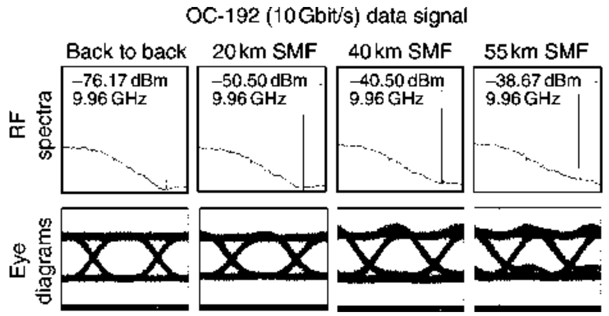

Figure 21 Clock regenerating effect due to chromatic dispersion for NRZ data. As the amount of residual dispersion increases, so does the amount of power at the clock frequency. This power can be used to monitor the amount of uncompensated chromatic dispersion. (Reproduced with permission from Pan Z, Yu Q, Xie Y, *et al.* (2001) Chromatic dispersion monitoring and automated compensation for NRZ and RZ data using clock regeneration and fading without adding signalling. *Optical Fiber Communication Conference 2001.* Paper WH5.)

Conclusion

Chromatic dispersion is a phenomenon with profound implications for optical fiber communications systems. It has negative effects, broadening data pulses, but it also helps reduce the effects of fiber nonlinearities. For this reason, managing dispersion, rather than trying to eliminate it altogether, is the key. Fixed dispersion components are suitable for point-to-point OC-192 systems. Tunable dispersion components are essential for dispersion management in reconfigurable and OC-768 systems. These dispersion compensation elements must have low loss, low nonlinearity, and must be cost effective.

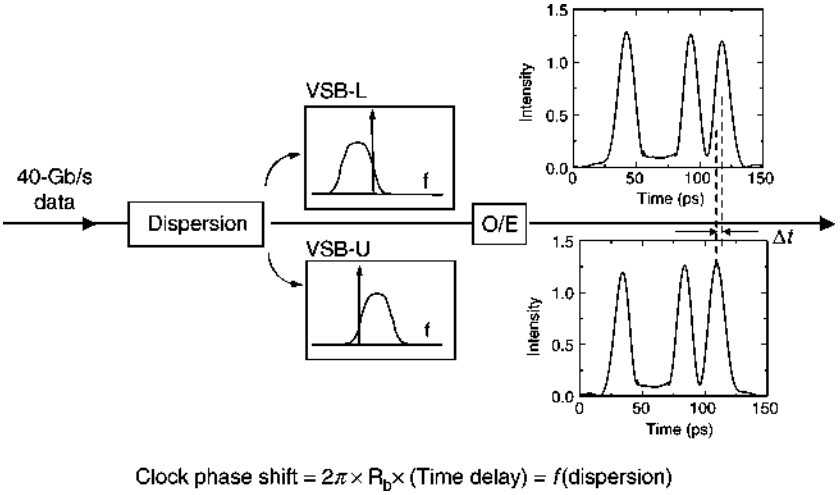

Figure 22 Chromatic dispersion monitoring using the time delay (Δt) between two VSB signals, which is a function of chromatic dispersion. (Reproduced with permission from Yu Q, Yan L-S, Pan Z and Willner AE (2002) Chromatic dispersion monitor for WDM systems using vestigial-sideband optical filtering. *Optical Fiber Communication Conference 2002.* Paper WE3.)

Although several technologies have emerged that meet some or all of the above requirements, no technology is a clear winner. The trend is towards tunable devices, or even actively self-tunable compensators, and such devices will allow system designers to cope with the shrinking system margins and with the emerging rapidly reconfigurable optical networks.

List of Units and Nomenclature

Chromatic dispersion	The time-domain pulse broadening in an optical fiber caused by the frequency dependence of the refractive index. This results in photons at different frequencies traveling at different speeds [ps/nm/km].
Chromatic dispersion slope	Different wavelengths have different amounts of dispersion values in an optical fiber [ps/nm^2/km].
Conventional single mode fiber (SMF)	Fiber that transmits only a single optical mode by virtue of a very small core diameter relative to the cladding. Provides lowest loss in the 1,550 nm region and has zero dispersion at 1,300 nm.
Cross-phase modulation (XPM)	A nonlinear Kerr effect in which a signal undergoes a nonlinear phase shift induced by a copropagating signal at a different wavelength in a WDM system.
Dispersion compensating fiber (DCF)	Optical fiber that has both a large negative dispersion and dispersion slope around 1,550 nm.
Fiber Bragg gratings (FBGs)	A small section of optical fiber in which there is a periodic change of refractive index along the core of the fiber. An FBG acts as a wavelength-selective mirror, reflecting only specific wavelengths (Bragg wavelengths) and passing others.
Four-wave mixing (FWM)	A nonlinear Kerr effect in which two or more signal wavelengths interact to generate a new wavelength.
Kerr effect	Nonlinear effect in which the refractive index of optical fiber varies as a function of the intensity of light within the fiber core.
Modulation	The process of encoding digital data onto an optical signal so it can be transmitted through an optical network.
Nonlinear effects	Describes the nonlinear responses of a dielectric to intense electromagnetic fields. One such effect is the intensity dependence of the refractive index of an optical fiber (Kerr effect).
Nonzero dispersion-shifted fiber (NZDSF)	Optical fiber that has a small amount of dispersion in the 1,550 nm region in order to reduce deleterious nonlinear effects.
Polarization mode dispersion (PMD)	Dispersion resulting from the fact that different light polarizations within the fiber core will travel at different speeds through the optical fiber [ps/km$^{0.5}$].
Self-phase modulation (SPM)	A nonlinear Kerr effect in which a signal undergoes a self-induced phase shift during propagation in optical fiber.
Wavelength division multiplexing (WDM)	Transmitting many different wavelengths down the same optical fiber at the same time in order to increase the amount of information that can be carried.

See also

Nonlinear Optics, Basics: Four-Wave Mixing. **Photonic Crystals:** Nonlinear Optics in Photonic Crystal Fibers.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*. Rochester, NY: John Wiley & Sons.

Agrawal GP (1997) *Nonlinear Fiber Optics*, 2nd edn. Academic Press.

Gowar J (1993) *Optical Communication Systems*, 2nd edn. Prentice Hall.

Kaminow IP and Koch TL (1997) *Optical Fiber Telecommunications IIIA & IIIB*. Academic Press.

Kashyap R (1999) *Fiber Bragg Gratings*. Academic Press.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Artech House.

Willner AE and Hoanca B (2002) Fixed and tunable management of fiber chromatic dispersion. In: Kaminow IP and Li T (eds) *Optical Fiber Telecommunications IV*. Academic Press.

Yariv A and Yeh P (1984) *Optical Waves in Crystals*. John Wiley & Sons.

DISPLAYS

R L Donofrio, Display Device Consultants LLC, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article is a brief review of display technology being used today. It emphasizes some of the optical aspects of displays. The details of the basic construction of the display devices are also discussed to aid in understanding. The displays covered are the CRT, VFD, FED, PDP, LED, OLED, LCD-AMLCD, transmissive and transflective LCDs.

Definition of a Display

A display can be defined as an object which transfers information to the viewer and/or an output unit that gives a visual representation of data.

Optical Characteristics of a Display

A display has (in general) the optical characteristics of luminance, reflectivity, contrast ratio, color, surface gloss, resolution, and image sharpness (sometimes called small area contrast). Measurement methods can be obtained from the EIA (Electronics Industries Alliance), JT-31 (Optical Characteristics of Display Devices), SAE (Society of Automotive Engineers and ISO Standards) and VESA Standards, to name a few.

Luminance

The luminance of a display is its light output and is given in units of candela/square meter or cd/m^2 (in the English system fL). Luminance is measured using a photometer, photometer/colorimeter, or spectroradiometer.

Illuminance

The illuminance is the light incident on the surface being measured. Its units are Lumen per m^2 or Lux (and in the English system Ft.cd). This is measured using illuminance meters.

Reflectivity

The reflectivity for CRTs (cathode ray tubes) is measured with incident light at 45 degrees (in some cases for LCDs, a smaller 30 degree angle is used). The reflected light from the surface is measured with a detector normal to (perpendicular to) the surface. Devices such as 'tube face reflectivity' (TFR) can be used to measure this parameter and it is published in the EIA-JT-31 measurement methods. Other similar methods are used for flat panel displays.

Color

Although color perception is different for different individuals, standard responses to color stimuli have been developed and we can measure the color of different objects and display images. Color can be measured with such equipment as a colorimeter or spectroradiometer. The output data are given in various color systems. Color systems, such as the 1931 CIE (Commission International de I'Eclairage) Color System (x, y), the 1960 CIE (u, v) or 1976 CIE (u', v') color coordinates are all used. All these color systems can be converted to each other by simple mathematical relations and these conversions are usually available to the user in the data output. The choice of the color space/system is dependent on application but the system in most general use for displays is the 1976 CIE (u', v'). This color system has become more widely used because equally perceived color differences in this color system are at almost equal distances in this color space. In reality, the quest for a system with equal distances equating to equal color differences can be achieved through a complicated matrix approach in which the constants used in the matrix vary from point to point in the color space.

Gloss

Gloss is a measure of surface reflectivity under Snell's Law criteria (equal incidence and reflection angles). The measurement angles of 20, 45, 60, and 85 degrees are used to measure the gloss of common surfaces. Display glass surface measurements tend to use the 60 degree gloss angle. The surface of a display can be made rougher to scatter incident light away from the viewer or it can be optically coated to reduce the reflected light to achieve improved viewing. Some times both methods are used. Reduced gloss surfaces also are involved with the reduction in surface glare. The level of increased roughness of the surface affects the user's ability to see detail.

Contrast Ratio

When people talk about contrast they generally mean 'contrast ratio'. The contrast ratio (CR) of an area on a display is a unitless parameter and is defined as the excited luminance (L_e) divided by the unexcited luminance (L_u). The excited luminance (L_e) equals the glass transmission (T) times some excited internal luminance (L_{ei}) and for simplicity, L_u equals T times an internal zero ambient unexcited luminance light level of (L_{ui}). Factors, such as internal scattered light in LCDs and scattered electrons and scattered light in CRTs, give a nonzero un-excited luminance L_u.

When dealing with the CR, the effect of ambient illumination A contributes to the numerator and the denominator. Thus, when the ambient illumination is increased, it lowers the contrast ratio. In a simple case where the ambient illumination is reflected from the surface of the display whose reflectivity is ρ, the contrast ratio is given below:

$$\mathrm{CR} = (L_e + A\rho)/(L_u + A\rho) \qquad [1]$$

For the case of the reflectivity of a phosphor screen (reflectance R) coated panel with a glass transmission (T), the reflectivity to first approximation is

$$\rho = cRT^2 \qquad [2]$$

where c is a constant.

So we see that changes in contrast can come about through changing the display's glass panel transmission. For example, if the L_e of the white field of a CRT was 100 fL with a 90% panel glass (along with $R = 0.5$) and had a L_u of 2 fL, then the zero ambient contrast would be 50. However, with an ambient illumination of 50 Ftcd the contrast would be 4.76. Now, when we use a 50% transmission panel glass at the ambient illumination of 50 Ft.cd the CR is increased to approximately 8.39. Equations [1] and [2] show how to calculate the contrast. It should be noted that decreased panel glass transmission or the use of 'dark glass' increases contrast because the reflectivity is a 'T squared' effect and the luminance L is a function 'T'. There are second-order effects involved with the reduction of reflected light and unexcited luminance which make the equations more complicated. These second-order effects will not be covered as they are beyond the basic discussion of this article.

In comparing display CR, one must compare the displays at the same ambient illumination. Additionally, you will find that a number of flat panel display advertising reports will report the CR at zero ambient illumination. As we have shown, these CR values are the highest ones for the display. The CR for emissive displays tends to be constant over a wide range of viewing angles, whereas the contrast ratio for liquid crystal displays (LCDs) is more variable over large-angle viewing. Recent new electrode position developments, such as the 'in-plane switching' technology for active matrix liquid crystal displays (AMLCDs) has gone a long way to reduce this reduced angle-of-view effect in that type of LCD display.

Figure 1 shows how a viewer in a sport utility vehicle (SUV) would see the changes in contrast with viewing angle for two types of twisted nematic LCDs (TN-LCDs). One of the TN LCDs does not

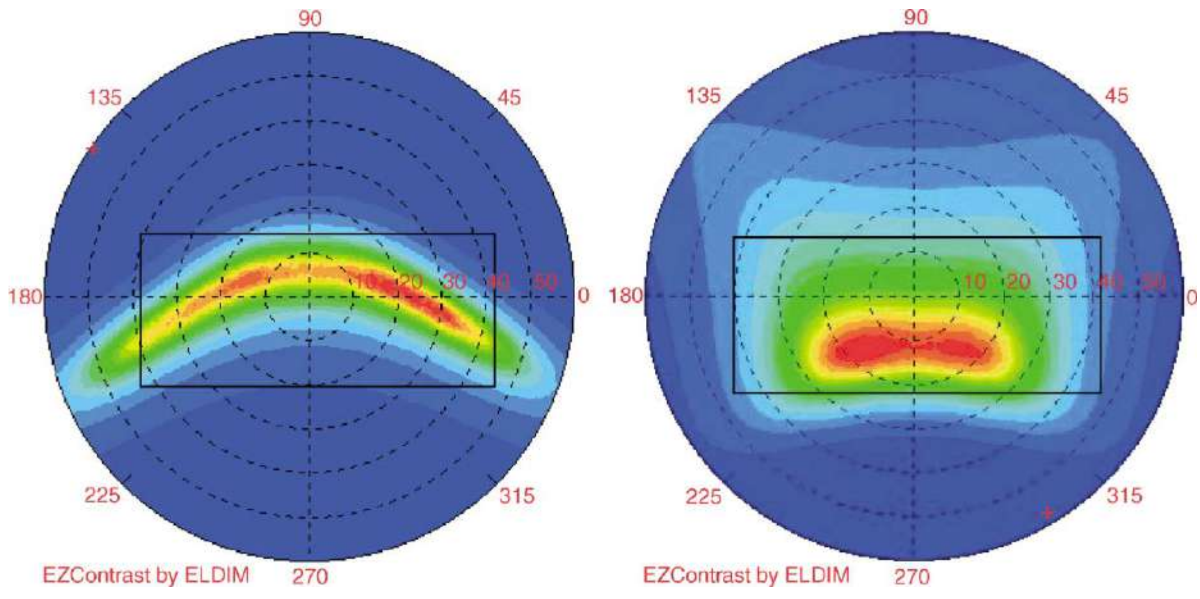

Figure 1 A typical rectangular viewing envelope for a mid-size SUV vehicle superimposed on LCD iso-contrast plots. Left is standard TN LCD. Right, is film-compensated TN LCD. Reproduced from Smith-Gillespie R, *et al.* (2001) Design requirements for automotive entertainment displays. *Proceedings of the 8th Annual Symposium on Vehicle Displays (2001)*. Permission for reprint courtesy of the Society for Information Displays.

have the film compensation and the other does. For relative contrast appraisal, the red areas show the highest contrast regions and the blue the lowest. It is obvious that different technologies can give significant changes in viewed contrast.

Phosphor Aging

The fall off of luminance as the phosphor screen is bombarded with the electron beam is called phosphor aging. This effect is due to two main factors: (i) the thermal quenching of the phosphor luminescence (which is reversible); and (ii) a nonreversible aging due to damage to the screen components (such as phosphor, binder, and aluminum coating). It has been found that the aging (when we are using an electron beam) is a function of the number of electrons bombarding the screen. Aging models help us to understand and predict the amount of aging a tube will show after a fixed amount of time. The first aging model in 1951 was on ion bombardment and showed that:

$$I = I_0/(1 + AN) \qquad [3]$$

where I_0 is the initial luminance before aging, and I is the final luminance.

In 1961 a similar (one parameter) Model $I = I_0/(1 + CN)$ was used by Pfhanl, where C is the aging constant to half luminance and N is the number of coulombs. In 1981, a two-component model for low current aging was presented, where:

$$I = I_0/(1 + N(B/Q^D)) \qquad [4]$$

Here we have two constants B and D, and Q is the cumulated charge in number of coulombs. When $D = 0$ then the equation given above goes to the Pfhanl case.

Later, in 1996 the three-parameter model was described and in 1997 the Gaussian beam aging model was described.

Ambient Illumination and Reflection

To see the information shown on a display, the information must be spatially detected by the human eye over the reflected ambient illumination (such as room and sunlight). The ambient illumination is generally from localized or diffuse sources. As we have seen previously, the effect of diffuse illumination can be reduced through the use of neutral density filters. But, when we are dealing with very high ambient light such as that entering the canopy of a fighter aircraft then, the monochromatic information from the display can be viewed using selective spectral filters. However, for the usual computer monitor, the 'specular reflection' (Snell's Law type of reflection) off the front of the display is due to a localized external light source. This reflection is reduced by methods such as anti-glare (AG) and anti-reflecting (AR) surfaces applied to the front surface of the display. A flat versus curved front surface of the display can also be a glare reduction technique. Surface roughness is used in AG surfaces to scatter a portion of the specular light away from the viewer. Surface roughness can be achieved through surface abrading, surface plastic coating, surface silicate coating, and surface acid etching to name a few methods. AR surfaces use optical interference techniques to reduce the reflected light to the viewer. In some cases, both methods are applied to a display surface. Surface gloss of the display is usually measured with a gloss meter in 60 degree gloss values but 20, 45, and 80 degree angles are also used. Resolution is measured in line pairs/mm or pixels/mm and image sharpness is associated with modulation transfer function (MTF) measurements. We should note here that resolution and image sharpness are not the same thing. High resolution is equated with the high frequency limit of the MTF measurement. Since the reduction in specular reflection through the use of surface roughness can affect the image sharpness, we see that a system of trade-offs can be developed to optimize a display for a given viewing environment. AR coatings improve glare reduction without scattering and thus give higher-resolution capability to the displays but this is usually obtained at a higher price then the AG coatings.

Types of Displays

Displays can be divided into two types; emissive and nonemissive. By emissive we mean that the display surface gives off light. In the nonemissive, the display or a small portion of the display acts as a shutter or spatial filter which also can be reflective in nature. In this nonemissive mode, the display can pass light from a source in back of or in front of the shutter (or both, as we shall see later).

Table of Emissive and Nonemissive Displays

Thus from Table 1 we see that under the title of emissive displays we have the cathode ray tube (CRT), projection CRT, vacuum fluorescent display (VFD), field emissive display (FED), electroluminescent (EL) display, light emitting diode (LED), organic light emitting diode (OLED), and plasma. These devices, which are in the realm of emissive displays, use certain types of emissive materials called

Table 1 Table of emissive and nonemissive displays

Emissive	*Nonemissive*
CRT	LCD and AMLCD
VFD	Bi-Stable LCD
FED	Digital Mirror D
EL – inorganic	LCOS (Liquid Crystal on Si)
LED	Electronic Inks
OLED	
PLASMA	

lumophors (or phosphors). The lumophors are called cathodoluminescent, photoluminescent, and electroluminescent, depending whether the excitation action was due to the interaction of the phosphor with cathode rays, photons, or low-voltage electrons. Cathodoluminescent lumophors are used in cathode ray tubes (CRTs – television tubes, monitor tubes, projection CRTs), (VFDs), and (FEDs).

EL displays can use inorganic lumophors (phosphors) or organic lumophors such as those used in OLEDs. Under photoluminescence displays, we have ‘plasma displays’ (plasma TV), IR-up conversion, and fluorescent lamps. In plasma displays, the excitation of the plasma in a localized region gives rise to UV photons which strike the phosphor screen in that region. In IR-up conversion, IR photons are absorbed in such a way that two IR photons can excite an electron with twice the energy of one IR photon. This then allows the use of IR lasers to create displays through the excitation of these up-conversion phosphors. In fluorescent lamps, the UV created by the excitation of Hg or Xe, strikes the special lamp phosphors (coated on the inside of the glass envelope), which become luminescent.

Under the heading of shutters we can consider LCDs which consist of AMLCDs, super twisted nematic liquid crystal displays (STN LCDs) and bistable LCDs. Another shutter type is the liquid crystal on silicon (LCOS) display and the movable digital mirror display (DMD), and the digital light processing (DLP) projecting engines. Another spatial display type is both normal and dynamic inks. A dynamic ink can be a ‘polarizable’ ink display in which a two shade (light/dark) ink sphere element can rotate under the presence of an electric field to show the light or dark side. Although they are very interesting, we will not cover ink displays in this article.

The shutter action is also used in some projection TV sets (or projectors) in which the LCD either transmits or reflects (such as LCOS) the light from an arc lamp. The DMD movable mirrors reflect the digital information to the lens system and the unwanted rays are reflected away from the lens to a black light capture area.

The Cathode Ray Tube (CRT)

The CRT is a display which has been in use for over 100 years (the Braun CRT tube was developed in 1898). The CRT uses a phosphor screen which is bombarded by energetic electrons and emits visible light radiation. The single electron gun CRT was used for years in early monochrome television sets and monochrome computer displays. Monochrome green, amber, and white phosphors have been used in monitor tubes. Some color selectivity can be obtained with the single electron gun using the following three methods; (i) voltage penetration phosphors, (ii) current sensitive phosphors and (iii) beam index methods. In a CRT, the penetration of the electron beam follows Hans Bethe’s energy loss equation for electron penetration in solids and thus is a function of the electron energy or CRT voltage.

Hans Bethe’s energy loss equation is:

$$\mathrm{d}E/\mathrm{d}x = -2\Pi e^4(N_e/E)\ln(\alpha E/I) \qquad [5]$$

where: E = energy of the incident electron; x = depth of material; N_e = density of atomic electrons; I = mean excitation energy; and α = a constant.

In the voltage penetration tube, a cascaded screen is used to give a color change when the electron voltage is increased. For example, if a cascade screen is made from a green phosphor layer placed on top of the red phosphor layer then the electrons strike the green phosphor first. Then with increased voltage, the electrons pass through the top phosphor and excite the bottom red phosphor. In the current sensitive color tube, the phosphor screen is composed of (for example) red and green phosphor whose luminance versus current curves are different. In these cases, we are referring to the linearity of the change in luminescence from the phosphor as a result of a change in electron beam current. The red phosphor (generally a rare earth oxide) is a linear phosphor, whereas the green is nonlinear. Changing the current thus can change the color of the display. In the beam index tube, a UV (or secondary electron emitters have also been used) emitting index layer is put along side a red, green, or blue triad (for example, between the blue and the red phosphor stripes). The scanning electron beam, through the use of a UV sensor (or secondary electron detector) mounted in the rear portion of the tube and the scanning electronics, can determine the location of the red, green, and blue phosphors. The electron beam can then excite only the individual phosphor areas of interest. With this tube, generally, the electron beam is at a low current to give a small enough cross-section to only excite one color and thus has limited application.

Another application of the monochrome is the projection CRT system. Here, three tubes for red, green, and blue are used to project an image on a viewing screen. The terminology is such that, when the CRTs are in the back of the screen and the viewer in front of the screen, it is called a ‘rear’ projection set and if the CRTs and the viewer are in front of the screen it is a ‘front’ projection set. The principle here is that, either using one CRT tube at a time or by a superposition two or three of the CRTs, the color of the displayed image will be within the color gamut defined by the primary colors.

The Color CRT

The more ubiquitous method of obtaining a color display is the device called the shadow mask CRT (and the similar Trinitron CRT). This color CRT is designed in such a manner that through the use of an aperture mask, (seen in Figure 1) the electrons from the red electron gun strike the red phosphor and likewise electrons from the green and blue electron guns strike the green and blue phosphors. The combination allows the viewer to see a color image. This color separation action is seen in Figure 2. Although Figure 2 shows a delta tri-color design, there are also in-line systems in which the mask consists of an array of slots which are somewhat rectangular in nature. In this array, many slots can be on a line, or a mask made can be made of a series of wires as used in the Trinitron design.

The disadvantage of the shadow mask technique is that the mask can capture as much as 80% of the electrons from the electron guns. This yields lower screen luminance and lower contrast due to scattered electrons. As the screen size gets larger from 19 inches to 36 inches or more, additional temperature compensation methods must be used to reduce the effects of mask heating. A number of methods to control the shift of the electron beam on the phosphor as the mask heats up have been developed. Initially, axially positioned bimetal springs were used on masks to compensate for the heating effect. Then lighter masks were used with corner lock suspension systems. Another method is using mask materials with lower coefficients of expansion such as INVAR.

Contrast in color CRTs has been improved by first applying a carbon coating between the panel glass and the phosphor screen. This requires additional process steps which form ‘windows’ in the carbon coating (matrix) through which the excited phosphor is seen. The use of pigmented phosphors also helped to reduce the reflectivity of the phosphor/matrix screen.

The phosphors used in the shadow mask color CRT are the red Y_2O_2S:Eu or, in some cases, the Y_2O_3:Eu phosphor. The green phosphor is ZnS:Cu and the blue phosphor is ZnS:Ag.

The CRT seen in Figure 3 shows the new high-definition television-type (HDTV) tube with the 16 × 9 aspect ratio as contrasted with the older 4 × 3 aspect ratio (screen width × height).

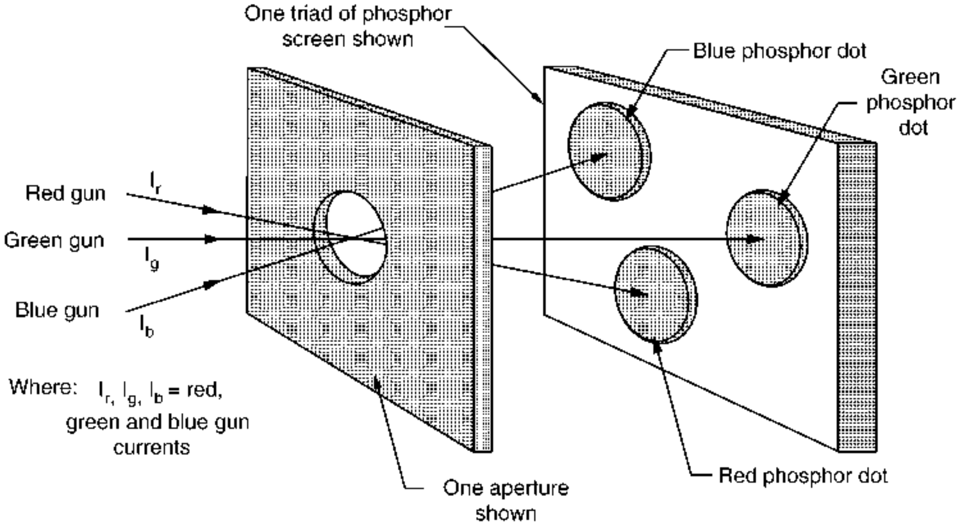

Figure 2 The action of a shadow mask.

Figure 3 A typical HDTV color CRT. Reproduced from Donofrio RL (2002) The future of direct-view CRTs. *SID Information Display Magazine* 6: 10. Permission for reprint courtesy of the Society for Information Display.

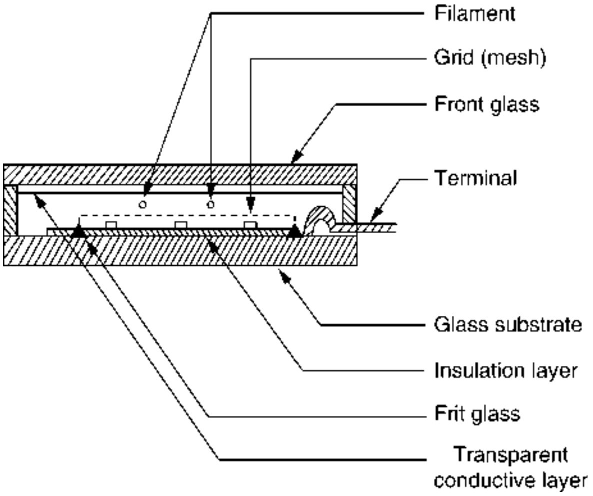

Figure 4 The cross-section of a vacuum fluorescent display assembly.

The Vacuum Fluorescent Display (VFD)

The VFD can be thought of as a low-voltage CRT. Just as in the CRT, electrons from a cathode are controlled by a grid which allows them to be accelerated and strike the phosphor screen anode. However, the electrons in a CRT can be accelerated with a voltage of 30 000 V, whereas the VFD accelerates the electrons with voltages of 12 to 40 V.

In the operation of the VFD, electrons are emitted by the filament. The filament is a very thin tungsten wire coated with a tri-carbonate oxide (barium, strontium, and calcium carbonates) which are the same constituents of the CRT oxide cathode. The filament is held between a filament support and an anchor which applies tension to the filament wire. A current is passed through the filament to achieve a temperature to ~600 °C. At that temperature, thermonic emission occurs. Between the filament and the phosphor elements is a stainless steel mesh grid. By the application of a positive or negative voltage on this grid, electrons can either be cut off or accelerated to the phosphor on the anode. The anode can be a conductor such as graphite, coated with a phosphor layer such as ZnO:Zn (zinc oxide). This is a low-voltage cathodoluminescent phosphor which emits a green color with an emission maxima of about 505 nm. Other phosphors are also used to obtain red, orange, and blue (Figure 4).

In addition to the filament, grid and anode, the VFD is a vacuum device (like the CRT) and contains a getter to assist achieving and maintaining a good vacuum and an internal transparent conductive coating to protect the display from external electrostatic fields. The VFD display is used frequently in the auto industry and aging of the phosphor is very small, achieving 100 000 hrs to half of the initial luminance. VFD units have been made with phosphor elements on either side of the filament. This gives rise to a 3D effect. Graphic and fixed segment VFDs have been made for multimedia products and 16 × 16 graphics for Kanji and other characters.

Experimentally, 2D VFD arrays have been made similar to the FEDs covered in the next section.

Field Emissive Displays (FEDs)

The FED is similar to the CRT and the VFD in that electrons from a source bombard a phosphor screen. In both the CRT and the VFD thermionic electrons are used. However, in the FED the electrons arise from field emission. Field emission devices create electrons though quantum mechanical tunneling from the emitter surface into the vacuum by the action of a large electrical field. This process is governed by the Fowler–Nordheim equation. This equation takes the form:

$$J = A(\beta V)^2/\Phi \exp(-B\Phi^{3/2}/\beta V) \qquad [6]$$

where, J is the electron current density; V is the applied voltage; Φ is the work function of the emitting tip; A & B are constants; and β is the field enhancement factor.

The emitters can be of the Molybdenum Spindt type and also the polydiamond or diamond-like carbon (DLC) coated emitters. We have the field enhancement with a sharp tip and reduced work function with the DLC. There has been renewed interest in FEDs, with the development of the 'carbon nanotube' as an electron field emitting source. The carbon nanotube is a fullerene material, which is a distinct form of carbon. This 'bucky ball'-like

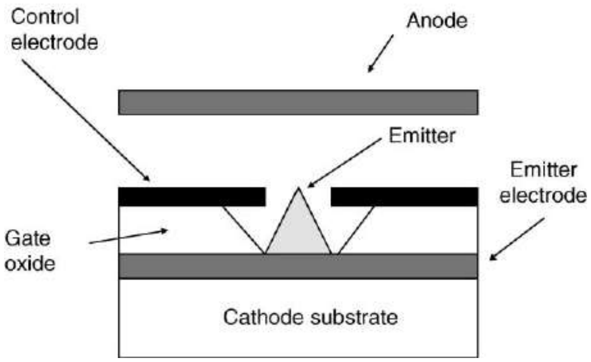

Figure 5 Field emission display cross-section. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.

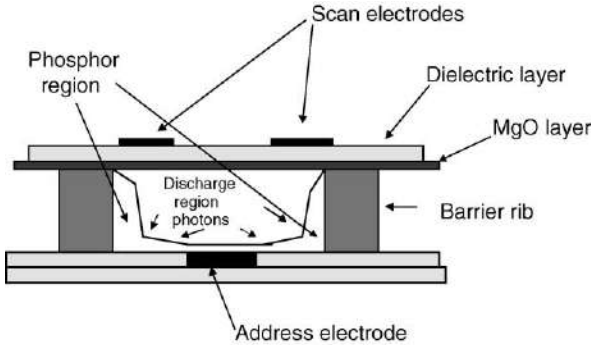

Figure 6 ACPDP partial cross-section view. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.

material is cylindrical instead of spherical in nature, has great strength, and high conductivity.

There are two types of FEDs, a low-voltage (LVFED) and a high-voltage (HVFED). The HVFED uses an aluminum film over the phosphor to protect the phosphor screen from aging. The HVFED can use the high voltage (2–10 KV) electrons to bombard the aluminized screen and it can use the more efficient, somewhat standard CRT phosphors ZnS:Cu green, ZnS:Ag Blue, and Y_2O_3:Eu red. The LVFED (12–200 V) cannot use the standard CRT phosphors but uses the low-voltage phosphor such as the ZnO:Zn green phosphor. This does not use an aluminum layer, since about 2 KV is needed to penetrate the Al layer. The commonly used CRT red phosphor Y_2O_2S:Eu was found unsuitable for HVFED use due to release of the sulfur during electron bombardment.

Figure 5 shows that for this basic FED structure, the electrons from the Spindt emitter are controlled by the voltages of the anode, control electrode and emitter electrode.

Plasma Displays

Earlier plasma displays, made using excited Ne gas with the characteristic amber glow required no phosphor and were used in the printing industry. Present plasma displays used for television and information signs are photoluminescent displays. As shown in Figure 6, the excitation of the plasma gas made of elements such as He, Ne, Ar, and Xe give rise to a glow discharge and the creation of vacuum ultraviolet photons that strike the screen and cause it to luminesce. In construction, the display is similar to the FED in that it is a matrix addressed device. It has been described in literature that the similarity goes even further with the need for anode and cathode surfaces, a means of evacuating and a perimeter seal. There are two types of plasma displays, DC and AC. The AC design appears to be the more successful. Typical PDP phosphors for the triad are (Y,Gd)BO_3:Eu for the red, Zn_2SiO_4:Mn for the green, and $BaMgAl_{10}O_{17}$:Eu (called BAM for short) for the blue. Figure 6 below shows the basic structure of an AC plasma display. One phosphor element is shown with barrier ribs and address and scan electrodes. The discharge of the plasma takes place in the region of the phosphor and excites the photoluminescent phosphor, which is shaped (in this example) along the side of the barrier ribs and over the address electrode.

Electroluminescence

Electroluminescence was first seen in the 1930s by Destriau. EL is the generation of (nonthermal) light by use of an electric field. High electric field EL units emit light through the impact of high energy electrons in luminescent centers in a semi-conducting material. The electric field can be alternating current (AC) or direct current (DC). Powder phosphor AC EL displays appear to be one of the successful inorganic EL display methods. Because of their rugged nature, alternating current electroluminescence ACEL displays have been used in trucks, tractors and the M1 Abrams Tank. Thick-film EL panels are also used in some vehicle displays as back lights for LCDs. These light panels are free from mercury and are more environmentally friendly.

Inorganic LED Displays

Light-emitting diodes are being used as backlights for LCD displays, replacing some of the fluorescent light backlighting methods. They are also used in indicator lights. Large displays have been made of LEDs. Using a matrix array row and column connectors, we can digitally select a given color LED and pattern in an LED array and generate a color image. Similar methods were carried out many years ago using incandescent lights. However, now they are bright and fast reacting. LEDs operate by semiconductor P–N junction rules and the flow of electrons and holes under forward bias operation give rise to junction light emission. Materials such as GaP:Zn,O are used to create a red LED with an emission maxima wavelength of 699 nm.

Red, green, and blue-UV/blue LEDs are now available, with an increasing demand for white LEDs. Two methods are used to produce a white LED. One method is to use a blue LED with the front of the LED coated with a green photoluminescent phosphor. If the phosphor is thin enough, the blue light penetrates the screen and excites the green to give a blue/green white. Another method is to use a blue LED and coat it with a white photoluminescent phosphor.

Organic LEDs

The basic process of light generation in OLEDs is the formation of a P–N junction. Electrons injected from the N-type material and holes from the P-type material recombine at the junction and light is generated. There are presently two competing methods to fabricate an OLED. The first developed by Kodak in 1987 is the small molecule molecular weight material approach. In OLED operation, the electron hole pair recombine to form excitons (on the average one singlet and three triplet excitons are formed for every four electron hole pairs). The light for the Kodak OLED is from the recombination of singlet excitons and the energy transferred to the fluorescent dye. The terminology is such that we have small molecule-OLEDs (sm-OLEDs) and polymer-OLED (PLED). In later developments, the sm-OLED device structure was reported to have three layers. The first is the hole transport layer (hole transport) enhancing layer copper phthalocyanine (CuPc), a layer NPB (60 nm), the light emitting aluminum tris-8-hydroxyquinolate Alq layer (75 nm), and electron transport layer. The electron transport layer is in contact with the three layer structure. The top electrode is a low work function material, such as MgAg (75–200 nm thick) called an electron injection layer and topped with a transparent indium tin oxide (ITO) layer for transparent OLEDS. All of these coatings are applied in a vacuum chamber.

In the phosphorescent material method, the light emission is from the triplet excitons through the use of phosphorescent dyes and the polymer-based devices. The common base material is a polyphenylene-vinylene PPV and other polymers such as polyfluorene have been used. The construction of the phosphorescent OLED is similar in that the ITO is also used as a starting substrate but the hole transport layer is applied by spin casting from solution. Spin casting is also used for the electron transfer material. After drying (removal of solvents) a calcium cathode material is applied. A contrast of 3000/1 has been achieved (for a military type application) using a green OLED with an AR coated circular polarizer and at an ambient light level of 500 lux, and a 6/1 contrast ratio has been achieved with an ambient illumination of 50 000 lux and an output of 370 cd/m^2. Additionally, OLEDs can be made on flexible (F) substrates and are called FOLEDs. The use of a transparent (T) cathode gives rise to a top emission OLED or a TOLED.

The first OLED display used in auto products was developed by Pioneer and used in a radio display and the Ford Motor Company has shown a White OLED display in one of their ‘Concept Cars’. This display allows them to have a reconfigurable instrument

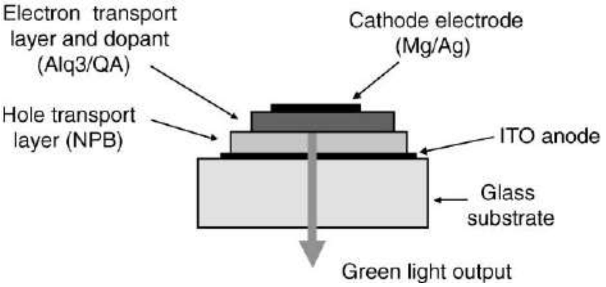
Electron transport layer and dopant (Alq3/QA)
Cathode electrode (Mg/Ag)
Hole transport layer (NPB)
ITO anode
Glass substrate
Green light output

Figure 7 OLED partial cross-section – Kodak type. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.

Figure 8 Philips poled display in an auto mirror. Reproduced with permission of Philips Mobile Display Systems.

panel (IP) and is made up of a number of 5 inch × 8 inch displays across the instrument panel (Figure 7).

The POLED as seen in Figure 8 is now being used by Philips Mobile Systems in a rear view car mirror, as seen below. It is the display from a video camera looking at the rear of a car to see any obstacles to backing up safely.

Thin Film EL (TFEL) Displays

A simple type of TFEL display is the color by white structure shown below. In this case, the color is generated using a white phosphor and red, green, and blue optical filters. To make this ACEL display function properly, insulators are used on either side of the phosphor. In Figure 9, the ITO conducting stripes (depicted) horizontally on the top and vertically on the bottom, give the means of addressing each color pixel. The white phosphor, for example, is a ZnS:Mn, SrS:Ce blend which gives a white electroluminescence.

Other phosphors such as ZnS:Mn can be used but its yellow luminescence requires the use of a green and red filter and only two metal electrodes on the bottom to give a two-color display. Better red green color separation can be achieved through the use of a $Zn_{1-x}Mg_xS$:Mn and ZnS:Mn phosphor mix. New EL blue phosphor developments have allowed larger color gamuts to be achieved.

One of the important characteristics of the alternating current thin-film electroluminescence (ACTFEL) is that it has an operating temperature range from −40 °C to 85 °C which is key for automotive use. It has a viewing angle of 160 degrees with a life of more than 50 000 hours. Active matrix (active addressed) AMEL structures allow fast

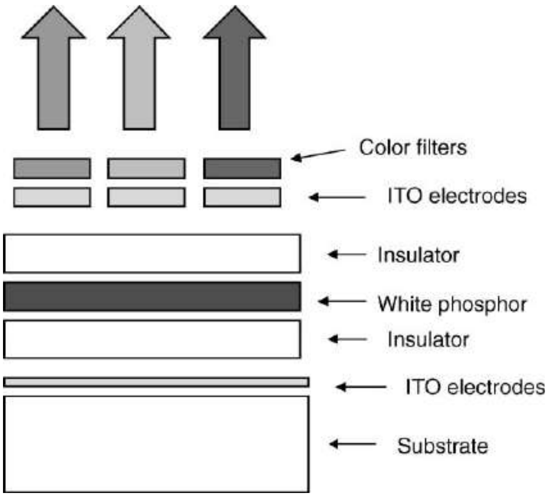

Figure 9 Color thin film electroluminescent display (TFEL) cross-section. Reproduced from Dakkala A (1998) New EL displays for vehicular applications. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98.* Permission for reprint courtesy of the Society for Information Display.

Figure 10 AMLCDs used in automobiles. Reproduced from Donofrio RL (2003) The road show. *SID Information Display Magazine* 7: 26. Permission for reprint courtesy of the Society for Information Display.

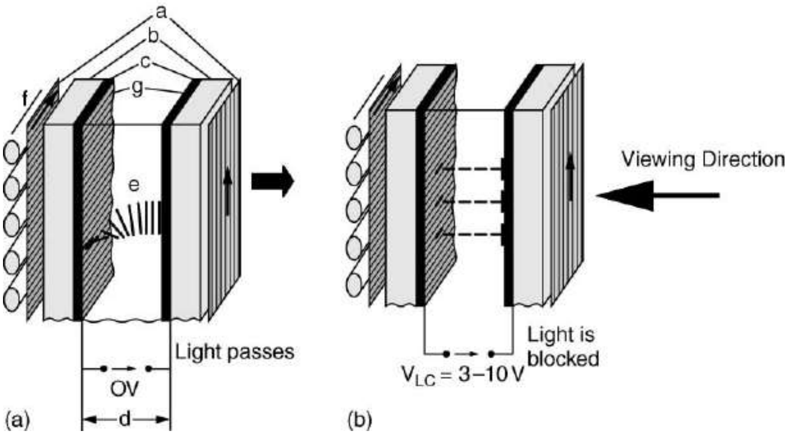

Figure 11 The operation of a simple liquid crystal display. Reproduced from Lueder E (1997) Fundamentals of passive and active-addressed liquid crystal displays. *SID Short Course S-1* (1997). Permission for reprint courtesy of the Society for Information Display.

response and moving graphic displays. These have found application in 'heads up displays' (HUD) for the military and medical industries and in consumer goods such as portable computers, pagers, cell phones, and 3D gaming.

The Active Matrix LCD

The active matrix liquid crystal display (AMLCD) device consists of a series of liquid crystal cells or windows whose characteristics can be controlled by an applied voltage to pass or restrict the passage of light through the windows via the external polarizers and the polarizing action of the LCD cells. These displays are used in laptop and tabletop computers, TVs, and for automotive displays (for driving information and entertainment). Figure 10 shows some present applications for AMLCDs in the automobile.

When the simple LCD is used without a transistor matrix, it is termed a passive display rather than an active one. Figure 11 below shows how a simple LCD cell operates.

The action shown in this figure is the twisted alignment nature of the liquid crystal material. When the polarizer on the left has its high transmission axis horizontal and the polarizer on the right of this cell has perpendicular alignment then the light from the backlight (f) is passed through the crossed polarizers. However, when voltage is applied to the cell, the liquid crystal is no longer twisted and the crossed polarizers prevent the light from going though the cell to the viewer. If the polarizers were not crossed but the polarization transmission axis was in the same direction in both polarizers, then the opposite effect would occur.

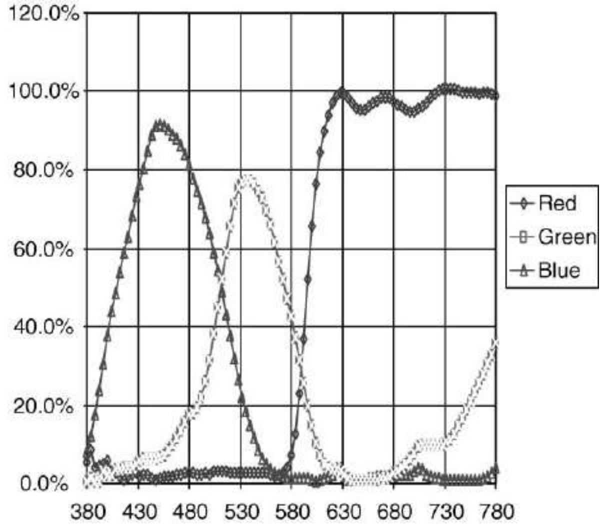

Figure 12 Typical color LCD filter transmission characteristics. Reproduced from Donofrio RL and Abileah A (1998) LCD backlighting system – phosphors and colors. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98, p. 123*. Permission for reprint courtesy of the Society for Information Display.

In order to show a color image, the LCD is designed to have three or more cells clustered per color pixel. The following (Figure 12) is the transmission profiles of each of these red, green, and blue filters.

The backlight used is a fluorescent lamp with a white phosphor (made up of red, green, and blue photoluminescent phosphors). Figure 13 shows what would happen if only the blue phosphor was used and also shows the superposition of the mercury lines that penetrate the phosphor and contribute to the final spectral energy distribution.

Reflective LCD

For applications where there is suitable external light to view the display, no internal light source is used.

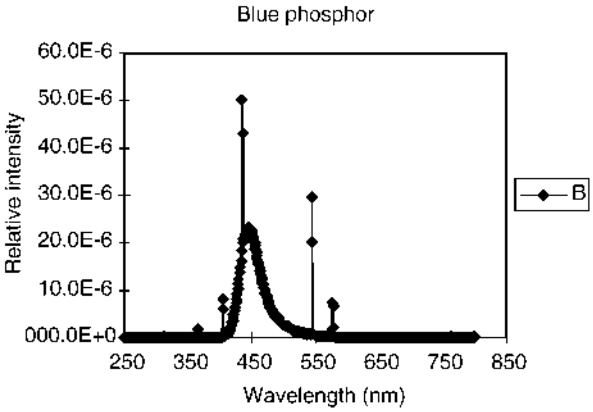

Figure 13 The relative luminance spectrum of the LCD. Backlight blue phosphor showing mercury lines. Reproduced from Donofrio RL and Abileah A (1998) LCD backlighting system – phosphors and colors. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98, p. 123*. Permission for reprint courtesy of the Society for Information Display.

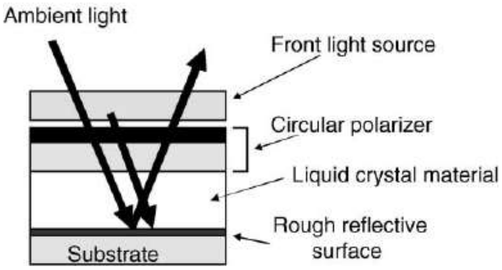

Figure 14 The action and construction of a reflective LCD. Reproduced with permission from Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35. Permission for reprint courtesy of the Society for Information Display.

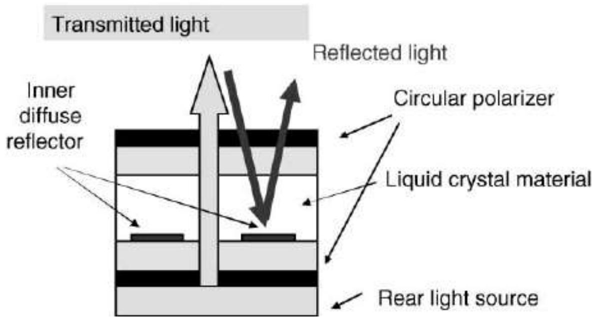

Figure 15 The action and cross-section of a transflective LCD. Reproduced with permission from Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35. Permission for reprint courtesy of the Society for Information Display.

This reflective LCD consists of a polarizer and retarder plates stacked on the front side of the cell. The cell with spacers contains the liquid crystal material and on the bottom of the cell is located a reflecting surface. It has been found that a dimpled reflecting surface makes a diffuse reflecting surface (Figure 14).

Transflective LCD

The importance of the transflective device is that it has better contrast under high ambient illumination than the standard transmission back lighted LCD and fair contrast for low ambient illumination.

The construction of this device consists of front and rear circular polarizers, back light, reflective elements, and novel filters. The back light is in the rear of the display. However, like the reflective LCD, internal reflectors are used on each pixel. The filters are designed spatially to be more transmissive for the reflective light than for the back light. Additionally the cell gap for the reflective light is smaller then the transmitted light region because the reflected light passes the cell twice whereas the transmitted light passes the cell once. This type of device seen in Figure 15, is finding uses in the auto industry and where the LCD is needed for both low and high ambient light levels.

Conclusions

In this brief review of displays we have tried to cover many different types of displays. The reader will in time find that there are many other variations on the displays discussed. The object is to give the reader an up-to-date review of the topics involved with displays and the various types of display in use today. By its nature it is a broad approach.

See also

Color and the World. **Instrumentation:** Photometry. **Light Emitting Diodes**. **Materials for Nonlinear Optics:** Liquid Crystals for NLO.

Further Reading

Donofrio RL (1981) Low Current Density Aging. *Proceedings of the Electro-Chemical Society*, Abstract #152, May 11, 1981.

Donofrio RL (2003) The road show. *Information Display* 7(3): 26–29.

Donofrio RL and Abileah A (1998) LCD backlighting. *Flat Panel and Vehicle Displays '98* : 119–124. Metropolitan Detroit Chapter of the SID.

Donofrio RL, Musa S and Wise K (eds) (1999) Vehicle Displays and Microsensors '99. *Proceedings of the 6th Strategic Forum and Technical Symposium*, Sept 22–23, 1999.

Flat Panel Display Measurements Standard, Version 1.0 VESA (Video Electronics Standards Association) May 15, 1998.

Forrest S (2002) *Fundamentals of Organic Light-Emitting Devices*. Short Course S-2 SID.

Green P and Haven D (1998) *Fundamentals of Emissive Displays*, Short Course S-3, SID.

Keller P (1991) *The Cathode-Ray Tube, Technology, History, and Applications*. New York: Palisades Press.

Leverence HW (1968) *An Introduction to Luminescence of Solids*. Dover Publications.

MacDonald W and Lowe A (1997) *Display Systems – Design and Applications*. SID Series in Display Technology, Wiley and Sons Ltd.

Sherr S (1993) *Electronic Displays*, 2nd edn. John Wiley.

Shionoya S and Yen WM (1998) *Phosphor Handbook*. CRC Press.

Stupp EH and Brennesholtz MS (1999) *Projection Displays*. SID Series in Display Technology, Wiley and Sons Ltd.

Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35.

Web site atis.org/tg2k/. American National Standard TI.523-2001 'Def of Display Device'.

Wu ST and Yang DK (2001) *Reflective Liquid Crystal Displays*. SID Series in Displays Technology, Wiley and Sons Ltd.

E

ELECTROMAGNETICALLY INDUCED TRANSPARENCY

J P Marangos, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In this article we examine how the process of electromagnetically induced transparency (EIT) can be used to increase the efficiency of non-linear optical frequency mixing schemes. We illustrate the basic ideas by concentrating upon the enhancement of resonant four-wave mixing schemes in atomic gases. To start we introduce the quantum interference phenomenon that gives rise to EIT. The calculation of EIT effects in a three-level atomic medium using a density matrix approach is presented. Next we examine the modified susceptibilities that result and that can be used to describe nonlinear mixing, and show how large enhancements in nonlinear conversion efficiency can result. Specific examples are briefly discussed along with a further discussion of the novel aspects of refractive index and pulse propagation modifications that result from EIT. The potential benefits to nonlinear optics of electromagnetically induced transparency are discussed. In an article of this nature it is not possible to credit all of the important contributions to this field, but we include a bibliography for further reading indicating some of the seminal contributions.

Electromagnetically induced transparency is the term applied to the process by which an otherwise opaque medium can be rendered transparent through laser-induced quantum mechanical interference processes. The linear susceptibility is substantially modified by EIT. Absorption is cancelled due to destructive interference between excitation pathways. The dispersion of the medium is likewise modified such that where there is no absorption the refractive index is unity and there can be a large value of normal dispersion leading to low values of group velocity. In contrast to the linear susceptibility the nonlinear susceptibility involving these same resonant laser fields will undergo constructive rather than destructive interference. This leads to a strong enhancement in nonlinear frequency mixing. For a normally transparent medium the nonlinear optical couplings will be small unless large-amplitude fields are applied. The most important consequence of EIT, in contrast to the usual case when a medium is transparent, is that the dispersion is not vanishing and the nonlinear couplings can be very large.

To understand how this comes about we must consider the system illustrated in Figure 1a where because of the resonant condition satisfied for the applied fields the atom can be simplified to just three-levels. The important parameters in this model system are that state |2> is metastable and has a dipole-allowed coupling to state |3>. The coupling between states |1> and |3> is also dipole-allowed

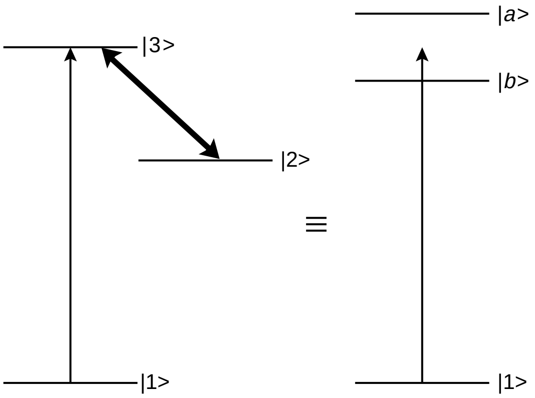

Figure 1 A three-level atomic system coupled to laser fields in the lambda configuration is shown on the left-hand side. In the limit of a strong coupling field this is equivalent to the dressed states |a> and |b> coupled to the ground state by the probe field alone as shown on the right-hand side.

and as a consequence state |3> can decay radiatively to either |1> or |2>. Let a coupling field be applied resonantly to the |2> − |3> transition, which may be cw or pulsed but should in the latter case be transform-limited. We define the Rabi coupling as $\Omega_{23} = \frac{\mu_{23}E}{\hbar}$ where E is the laser electric field amplitude and μ_{23} is the dipole moment of the transition. The Rabi frequency needs to be larger than the inhomogeneous broadening of the sample. A second field, the probe, typically of much lower intensity, is then applied in resonance with the |1> − |3> transition. If the condition $\Omega_{23} \gg \Omega_{13}$ is satisfied then it is convenient to replace the bare atomic states |2> and |3> with the dressed states (see Figure 1b):

$$|a> = \frac{1}{\sqrt{2}}[|2> + |3>] \qquad [1a]$$

$$|b> = \frac{1}{\sqrt{2}}[|2> - |3>] \qquad [1b]$$

Transitions from state |1> induced by the probe field to this pair of near-resonant dressed states are subject to exact cancellation at resonance if |2> is perfectly metastable. This is because the only contribution to the excitation amplitude comes from the matrix elements $<1|\mathbf{r}\cdot\mathbf{E}|3>$ as the |1> − |2> transition is dipole forbidden. The contributions from the equally detuned states |a> and |b> thus enter into the overall amplitude with opposite signs and equal magnitude as can be seen by inspection of eqns [1a] and [1b]. This leads to cancellation of the absorption amplitude. This type of absorption cancellation is well known and closely related to the so-called *dark states*.

Theoretical Treatment of EIT in a Three-Level Medium

It was realized by several workers in the 1970s that laser-induced interference effects could lead to a cancellation of absorption at certain frequencies. To gain a more quantitative understanding of the effects of the coupling field upon the optical properties of a dense ensemble of three-level atoms we require a treatment that computes the optical susceptibility of the medium. A treatment originally carried out by Harris *et al.* for a Λ scheme similar to that illustrated in Figure 1 was the first to derive the modified susceptibilities that will be discussed below. In that treatment the state amplitudes in the three-level atom were solved in the steady-state limit and from these the linear and nonlinear susceptibilities (see below) are then derived. In what follows we adopt a density matrix treatment as employed by various workers. This readily allows the inclusion of dephasing as well as population decay terms. The critical parameters in this system are the strengths of the fields (described in terms of the Rabi couplings), the detuning of the fields from resonance $\Delta\omega_{13} = \omega_{13} - \omega_P$ and $\Delta\omega_{23} = \omega_{23} - \omega_C$ (see Figure 2), the radiative decays from |3> to |1> and |2>, γ_c and γ_d and from |2>−|1> γ_a (although the latter is anticipated to be smaller). Extension to other configurations of the three states, such as a V or ladder scheme is implicit within this general treatment.

Figure 2 illustrates the system considered. This treatment will also address the nonlinear response in a four-wave mixing scheme completed by the two-photon resonant coupling applied between state |1> and |2>.

We write the interaction Hamiltonian as:

$$H = H_0 + V \qquad [2]$$

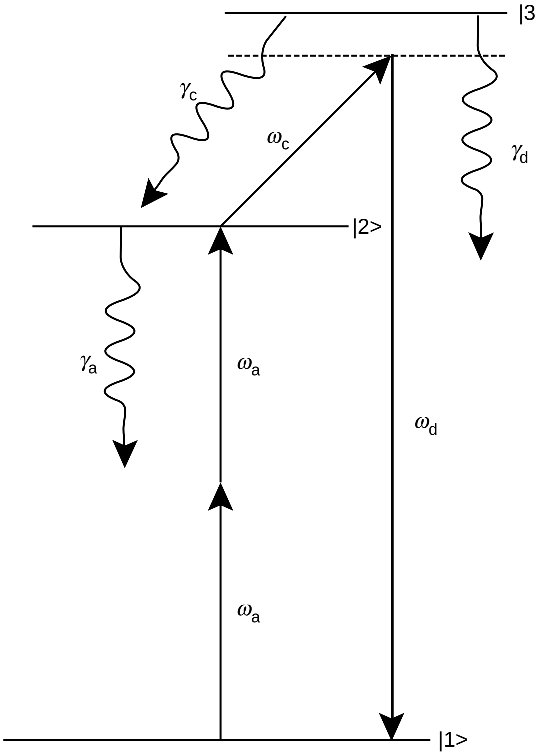

Figure 2 A four-wave mixing scheme incorporating the three-level lambda system in which electromagnetically induced transparency has been created for the generated field ω_d by the coupling field ω_c. The decay rates γ_i from the three atomic levels are also shown. For a full explanation of this system see text.

where H_0 is the unperturbed Hamiltonian of the system and is written as

$$H_0 = \hbar\omega_1 |1\rangle\langle 1| + \hbar\omega_2 |2\rangle\langle 2| + \hbar\omega_3 |3\rangle\langle 3| \qquad [3]$$

and V is the interaction Hamiltonian and can be expressed

$$\begin{aligned} V = {} & \hbar\Omega_{\mathrm{a}} \mathrm{e}^{-\mathrm{i}\omega_{\mathrm{a}}\mathrm{t}} |2\rangle\langle 1| + \hbar\Omega_{\mathrm{c}} \mathrm{e}^{-\mathrm{i}\omega_{\mathrm{c}}\mathrm{t}} |2\rangle\langle 3| \\ & + \hbar\Omega_{\mathrm{d}} \mathrm{e}^{-\mathrm{i}\omega_{\mathrm{d}}\mathrm{t}} |3\rangle\langle 1| + c.c \end{aligned} \qquad [4]$$

Note that the Rabi frequencies in the equation can be described as $\hbar\Omega_{ij} = \mu_{ij}|E(\omega_{ij})|$, where μ_{ij} is the dipole moment of the transition between states |i> and |j>, the Rabi frequency Ω_{a} is a two-photon Rabi frequency that characterizes the coupling between the laser field a and the atom for this two-photon transition. We have assumed for simplicity that $\omega_{\mathrm{a}} = \omega_{\mathrm{b}} = \omega_{12}/2$.

Assuming all the fields lie close to the resonance, the rotating wave approximation can be applied to the interaction picture and the interaction Hamiltonian V^I is given as

$$\begin{aligned} V^I = {} & \hbar\Omega_{\mathrm{a}} \mathrm{e}^{\mathrm{i}\Delta_{\mathrm{a}}\mathrm{t}} |2\rangle\langle 1| + \hbar\Omega_{\mathrm{c}} \mathrm{e}^{\mathrm{i}\Delta_{\mathrm{c}}\mathrm{t}} |2\rangle\langle 3| \\ & + \hbar\Omega_{\mathrm{d}} \mathrm{e}^{\mathrm{i}\Delta_{\mathrm{d}}\mathrm{t}} |3\rangle\langle 1| + c.c \end{aligned} \qquad [5]$$

where Δ_{a}, Δ_{c} and Δ_{d} refer the detunings of the fields and can be written as:

$$\begin{aligned} \Delta_{\mathrm{a}} &= \omega_{12} = 2\omega_{\mathrm{a}} \\ \Delta_{\mathrm{c}} &= \omega_{32} - \omega_{\mathrm{c}} \\ \Delta_{\mathrm{d}} &= \omega_{13} - \omega_{\mathrm{d}} \end{aligned} \qquad [6]$$

For the evaluation of the density matrix with this interaction V^I, the Schrödinger equation can be restated in terms of the density matrix components. This form is called the Liouville equation and can be written as:

$$\hbar \frac{\partial}{\partial t} \varrho_{ij}(t) = -\mathrm{i} \sum_k H_{ik}(t) \varrho_{ij}(t) + \mathrm{i} \sum_k \varrho_{ij}(t) H_{ik}(t) + \Gamma_{ij} \qquad [7]$$

where Γ_{ij} represents phenomenologically added decay terms (i.e. spontaneous decays, collisional broadening, etc.). This formalism leads to a set of nine differential equations for nine different density matrix elements that describe the three-level system.

To remove the optical frequency oscillations, a coordinate transform is needed and to incorporate the relevant oscillatory detuning terms into the off-diagonal elements we make the substitution:

$$\begin{aligned} \tilde{\varrho}_{12} &= \varrho_{12} \mathrm{e}^{-\mathrm{i}\Delta_{\mathrm{a}}\mathrm{t}} \\ \tilde{\varrho}_{23} &= \varrho_{23} \mathrm{e}^{-\mathrm{i}\Delta_{\mathrm{c}}\mathrm{t}} \\ \tilde{\varrho}_{31} &= \varrho_{31} \mathrm{e}^{-\mathrm{i}\Delta_{\mathrm{d}}\mathrm{t}} \end{aligned} \qquad [8]$$

This operation conveniently eliminates the time dependencies in the rate equations and the equations of motion for the density matrix are given by:

$$\begin{aligned} \frac{\partial}{\partial t}\varrho_{11} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}\tilde{\varrho}_{12} + \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}\tilde{\varrho}_{13} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}^{*}\tilde{\varrho}_{21} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}^{*}\tilde{\varrho}_{31} + \Gamma_{11} \\ \frac{\partial}{\partial t}\varrho_{22} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}^{*}\tilde{\varrho}_{23} + \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}^{*}\tilde{\varrho}_{21} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{32} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}\tilde{\varrho}_{12} + \Gamma_{22} \\ \frac{\partial}{\partial t}\varrho_{33} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}^{*}\tilde{\varrho}_{23} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}^{*}\tilde{\varrho}_{31} + \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}\tilde{\varrho}_{13} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{32} + \Gamma_{33} \\ \frac{\partial}{\partial t}\varrho_{23} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{22} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{33} - \mathrm{i}\Delta_{\mathrm{a}}\tilde{\varrho}_{23} \\ &\quad + \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}\tilde{\varrho}_{21} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}\tilde{\varrho}_{13} + \Gamma_{23} \\ \frac{\partial}{\partial t}\varrho_{21} &= \mathrm{i}\Omega_{\mathrm{a}}\tilde{\varrho}_{22} + \mathrm{i}\Omega_{\mathrm{a}}\tilde{\varrho}_{33} + \mathrm{i}\Omega_{\mathrm{d}}\tilde{\varrho}_{23} - \mathrm{i}\Delta_{\mathrm{c}}\tilde{\varrho}_{21} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{31} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}\tilde{\varrho}_{31} + \Gamma_{21} \\ \frac{\partial}{\partial t}\varrho_{31} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}\tilde{\varrho}_{22} + \mathrm{i}\Omega_{\mathrm{d}}\tilde{\varrho}_{33} + \mathrm{i}\Omega_{\mathrm{d}}\tilde{\varrho}_{23} - \mathrm{i}\Delta_{\mathrm{c}}\tilde{\varrho}_{21} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{31} - \mathrm{i}\frac{1}{2}\Omega_{a}\tilde{\varrho}_{31} + \Gamma_{31} \end{aligned} \qquad [9]$$

Using the set of equations above and the relation $\tilde{\varrho}_{ij} = \tilde{\varrho}_{ij}^{*}$ we obtain equations for $\tilde{\varrho}_{12}$, $\tilde{\varrho}_{32}$, and $\tilde{\varrho}_{13}$. For the incoherent population relaxation the decays can be written:

$$\begin{aligned} \Gamma_{11} &= \gamma_{\mathrm{a}}\varrho_{22} + \gamma_{\mathrm{d}}\varrho_{33} \\ \Gamma_{22} &= -\gamma_{\mathrm{a}}\varrho_{22} + \gamma_{\mathrm{c}}\varrho_{33} \\ \Gamma_{33} &= -(\gamma_{\mathrm{c}} + \gamma_{\mathrm{d}})\varrho_{33} \end{aligned} \qquad [10]$$

and for the coherence damping:

$$\begin{aligned} \Gamma_{21} &= -\left\{\frac{1}{2}[\gamma_{\mathrm{a}} + \gamma_{\mathrm{c}}] + \gamma_{21}^{\mathrm{col}}\right\}\varrho_{21} \\ \Gamma_{23} &= -\left\{\frac{1}{2}[\gamma_{\mathrm{a}} + \gamma_{\mathrm{c}} + \gamma_{\mathrm{d}}] + \gamma_{23}^{\mathrm{col}}\right\}\varrho_{23} \\ \Gamma_{31} &= -\left\{\frac{1}{2}[\gamma_{\mathrm{d}} + \gamma_{\mathrm{c}}] + \gamma_{31}^{\mathrm{col}}\right\}\varrho_{31} \end{aligned} \qquad [11]$$

where $\gamma_{ij}^{\mathrm{col}}$ represents collisional dephasing terms which may be present.

This system of equations can be solved by various analytical or numerical methods to give the individual density matrix elements. Analytical solutions are possible if we can assume for instance that $\Omega_c \gg \Omega_a$, Ω_d and that there are continuous-wave fields so a steady-state treatment is valid. For pulsed fields or in the case where the generated field may be significant numerical solutions are in general required.

We are specifically interested in the optical response to a probe field at ω_d (close to resonance with the $|1> - |3>$ transition) that is governed by the magnitude of the coherence ϱ_{13}. We find ϱ_{13} making the assumption that only the coupling field is strong, i.e., that $\Omega_c \gg \Omega_a, \Omega_d$ holds. From this quantity the macroscopic polarization is obtained from which the susceptibility can be computed. The macroscopic polarization P at the transition frequency ω_{13} can be related to the microscopic coherence ϱ_{13} via the expression:

$$P_{13} = N\mu_{13}\varrho_{13} \qquad [12]$$

where N is the number of equivalent atoms in the ground state within the medium, and μ_{13} is the dipole matrix element associated with the transition. In this way real and imaginary parts of the linear susceptibility χ at frequency ω can be directly related to ϱ_{13} via the macroscopic polarization since the latter can be defined as:

$$P_{13}(\omega) = \varepsilon_0\chi(\omega)E \qquad [13]$$

where E is the electric field amplitude at frequency ω_d. The linear susceptibility (real and imaginary parts) is given by the following expressions:

$$\mathrm{Re}\chi_D^{(1)}(-\omega_d,\omega_d) = \frac{|\mu_{13}|^2 N}{\varepsilon_0\hbar} \times\left[\frac{-4\Delta_{21}(|\Omega_2|^2 - 4\Delta_{21}\Delta_{32}) + 4\Delta_{31}\gamma_a^2}{4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2}\right] \qquad [14]$$

$$\mathrm{Im}\chi_D^{(1)}(-\omega_d,\omega_d) = \frac{|\mu_{13}|^2 N}{\varepsilon_0\hbar} \times\left[\frac{8\Delta_{21}^2\gamma_d + 2\gamma_a(|\Omega_2|^2 + \gamma_d\gamma_a)}{(4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2}\right] \qquad [15]$$

We now consider the additional two-photon resonant field ω_a. This leads to a four-wave mixing process that generates a field at ω_d (i.e., the probe frequency). The nonlinear susceptibility that describes the coupling of the fields in a four-wave mixing process is given by the expression:

$$\chi_D^{(3)}(-\omega_d,\omega_a,\omega_a,\omega_c) = \frac{\mu_{23}\mu_{13}N}{6\varepsilon_0\hbar^3((\Delta_{21} - j\gamma_a/2)(\Delta_{31} - j\gamma_d/2) - (|\Omega_c|^2/4))^*} \times\sum_i \mu_{i1}\mu_{i3}\left(\frac{1}{\omega_i - \omega_a} + \frac{1}{\omega_i - \omega_b}\right)$$

Nonlinear Optical Processes

The dependence of the susceptibilities upon the detuning is plotted in Figure 3. In this plot the effects of inhomogeneous (Doppler) broadening are

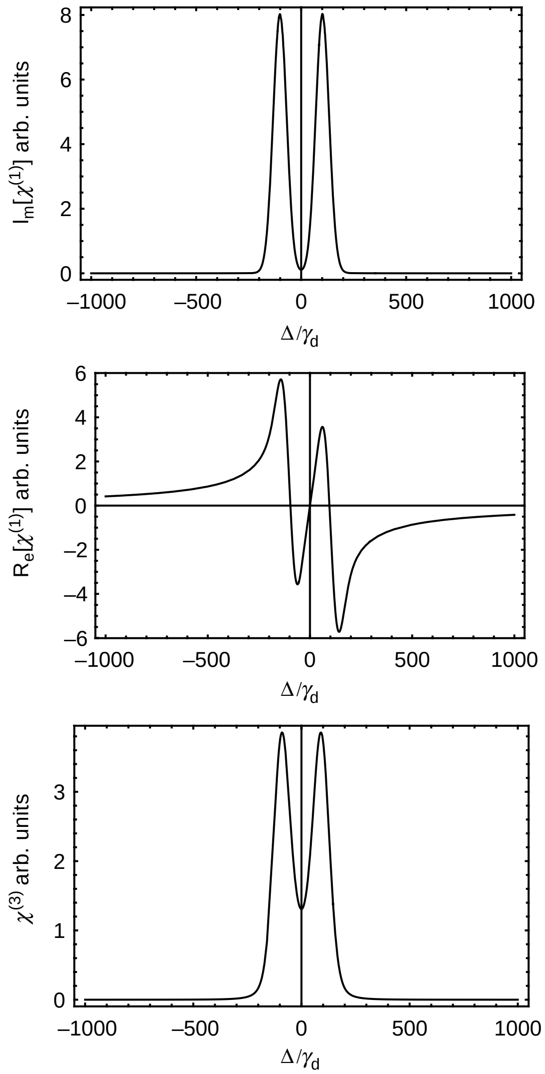

Figure 3 Electromagnetically induced transparency is shown in the case of significant Doppler broadening. The Doppler averaged values of the imaginary $\mathrm{Im}[\chi^{(1)}]$ and real $\mathrm{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. The Doppler width is taken to be $20\gamma_d$, $\Omega_c = 100\gamma_d$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

also incorporated by computing the susceptibilities over the inhomogeneous profile (see below for more details).

By inspection of eqn [14] we see that the absorptive loss at the minimum of $\mathrm{Im}[\chi^{(1)}]$ varies as γ_a/Ω_c^2. This dependence is a consequence of the interference discussed above. In the absence of interference (i.e., just simply Autler–Townes splitting) the minimum loss would vary as γ_d/Ω_c^2. Since $|3\rangle - |1\rangle$ is an allowed decay channel (in contrast to $|2\rangle - |1\rangle$) it follows that $\gamma_b \gg \gamma_a$ and so the absorption is much less as a consequence of EIT. $\mathrm{Re}[\chi^{(1)}]$ in eqn [15] and Figure 3 is also modified significantly. The resonant value of refractive index is equal to the vacuum value where the absorption is a minimum, the dispersion is normal in this region with a gradient determined by the strength of the coupling laser, a point we will return to shortly. For an unmodified system the refractive index is also unity at resonance but in that case there is high absorption and steep anomalous absorption. Reduced group velocities result from the steep normal dispersion that accompanies EIT. Inspection of the expression [16] and Figure 3 shows that $\chi^{(3)}$ is also modified by the coupling field. The nonlinear susceptibility depends upon $1/\Omega_c^2$ as is expected for a laser dressed system, however in this case there is not destructive but rather constructive interference, between the field-split components. This result is of great importance: it ensures that the absorption can be minimized at frequencies where the nonlinear absorption remains large.

As a consequence of constructive interference the nonlinear susceptibility remains resonantly enhanced whilst the linear susceptibility vanishes or becomes very small at resonance due to destructive interference. Large nonlinearity accompanying vanishing absorption (transparency) of course match conditions for efficient frequency mixing as a large atomic density can then be used. Moreover the dispersion (controlling the refractive index) also vanishes at resonance; this leads to perfect phase matching (i.e., zero wavevector mismatch between the fields) in the limit of a simple three-level system. As a result of these features a large enhancement of the conversion efficiency in this type of scheme can be achieved.

To compute the generated field strength Maxwell's equations must be solved using these expression for the susceptibility to describe the absorption, refraction, and nonlinear coupling in the medium. We will treat this within the slowly varying envelope approximation since the fields are cw or nanosecond pulses. To be specific we assume that the susceptibilities are time independent, i.e., that we are in the steady-state (cw) limit. We make the assumptions also that there is no pump field absorption and that we have plane waves. Under these assumptions the generated field amplitude A_d is given in terms of the other field amplitudes A_i by:

$$\frac{\partial}{\partial z}A_\mathrm{d} = \mathrm{i}\frac{\omega_\mathrm{d}}{4c}\chi^{(3)}A_\mathrm{a}^2A_\mathrm{c}\mathrm{e}^{-\mathrm{i}\Delta k_\mathrm{d}z} - \frac{\omega_\mathrm{d}}{2c}\mathrm{Im}[\chi^{(1)}]A_\mathrm{d} + \mathrm{i}\frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]A_\mathrm{d} \qquad [16]$$

where the wavevector mismatch is given by:

$$\Delta k_\mathrm{d} = k_\mathrm{d} + k_\mathrm{c} - 2k_\mathrm{a} \qquad [17]$$

The wavevector mismatch will be zero on resonance for the three-level atom considered in this treatment. In fact the contribution to the refraction from all the other atomic levels must be taken into account whilst computing Δk_d and it is implicit that these make a finite contribution to the wavevector mismatch.

We can solve this first-order differential equation with the boundary condition $A_\mathrm{d}(z=0)=0$ to find the generated intensity $I(\omega_\mathrm{d})$ after a length z:

$$I(\omega_\mathrm{d}) = \frac{3n\omega_\mathrm{d}^2}{8Z_0c^2}|\chi^{(3)}|^2|A_\mathrm{a}|^4|A_\mathrm{c}|^2 \times \frac{\left\{1+\mathrm{e}^{-\frac{\omega_\mathrm{d}}{c}\mathrm{Im}[\chi^{(1)}]z} - 2\mathrm{e}^{-\frac{\omega_\mathrm{d}}{c}\mathrm{Im}[\chi^{(1)}]z}\cos\left(\Delta k + \frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]\right)z\right\}}{\frac{\omega_\mathrm{d}^2}{4c^2}\mathrm{Im}[\chi^{(1)}]^2 + \left\{\Delta k + \frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]\right\}^2}$$

where Z_0 is the impedance of free space. This expression is quantitatively correct for the case of the assumptions made. More generally the qualitative predictions and general scaling remain valid in the limit where the pulse duration is significantly longer than the time required to traverse the medium. Note that both real and imaginary parts of $\chi^{(1)}$ and $\chi^{(3)}$ play an important role, as we would expect for resonant frequency mixing. The influence of that part of the medium refraction which is due to other levels is contained in the terms with Δk. In the case of a completely transparent medium this becomes a major limit to the conversion efficiency.

Propagation and Wave-Mixing in a Doppler Broadened Medium

Doppler shifts arising from the Maxwellian velocity distribution of the atoms in the medium lead to a corresponding distribution in the detunings for the various members of the atomic ensemble. The response of the medium, as characterized by the susceptibilities, must therefore include the Doppler effect by performing a weighted sum over possible detunings. The weighting is determined by the Gaussian form of the Maxwellian velocity distribution. From this operation the effective values of the susceptibilities at a given frequency are obtained, and these quantities can be used to calculate the generated field. This step is of considerable practical importance as in most up-conversion schemes it is not possible to achieve Doppler-free geometries and the use of laser cooled atoms, although in principle possible, limits the atomic density that can be employed. The interference effects persist in the dressed profiles providing the coupling laser Rabi frequency is comparable to or larger than the inhomogeneous width. This is because the Doppler profile follows a Gaussian distribution which falls off much faster in the wings of the profile than the Lorentzian profile due to the natural broadening.

In the case considered with weak probe field, excited state populations and coherences remain small. The two-photon transition need not be strongly driven (i.e., a small two-photon Rabi frequency can be used) but a strong coupling laser is required. The coupling laser must be intense enough that its Rabi frequency is comparable to or exceeds the inhomogenous widths in the system (i.e., Doppler width), and a laser intensity of above 1 MW cm^{-2} is required for a typical transition. This is trivially achieved even for unfocused pulsed lasers, but does present a serious limit to cw lasers unless a specific Doppler-free configuration is employed. The latter is not normally suitable for a frequency up-conversion scheme if a significant up-conversion factor is required, e.g., to the vacuum ultraviolet (VUV); however recent experiments report significant progress in cw frequency up-conversion using EIT and likewise a number of other possibilities, e.g., laser-cooled atoms and standing-wave fields, have been proposed.

A transform-limited single-mode laser pulse is essential for the coupling laser field since a multimode field will cause an additional dephasing effect on the coherence, resulting in a deterioration of the quality of the interference. In contrast, whilst it is advantageous for the field driving the two-photon transition to be single mode (in order to achieve optimal temporal and spectral overlap with the EIT hole induced by the dressing laser), this is not essential since this field does not need to drive the coherence responsible for interference.

When a pulsed laser field is used additional issues must be considered. The group velocity is modified for pulses propagating in the EIT large reductions, e.g., by factors down to $<c/100$, in the group velocity have been observed. Another consideration beyond that found in the simple steady-state case is that the medium can only become transparent if the pulse contains enough energy to dress all the atoms in the interaction volume. The minimum pulse energy to prepare a transparency is:

$$E_{\text{preparation}} = \frac{f_{13}}{f_{23}} NL\hbar\omega \qquad [18]$$

where f_{ij} are the oscillator strengths of the transitions and NL the product of the density and the length. Essentially the number of photons in the pulse must exceed the number of atoms in the interaction volume to ensure all atoms are in the appropriate dressed state. This puts additional constraints on the laser pulse parameters.

Up-conversion to the UV and vacuum UV has been enhanced by EIT in a number of experiments. Only pulsed fields have so far been up-converted to the VUV with EIT enhancement. The requirements on a minimum value of $\Omega_c > \Delta_{\text{Doppler}}$ constrains the conversion efficiency that can be achieved because the $1/\Omega_c^2$ factor in eqn [17] ultimately leads to diminished values of $\chi^{(3)}$. The use of gases of higher atomic weight at low temperatures is therefore highly desirable in any experiment utilizing EIT for enhancement of four-wave mixing to the VUV. Conversion efficiencies, defined in terms of the pulse energies by E_d/E_a or E_d/E_c of a few percent have been achieved using the EIT enhancement technique. It is typically most beneficial to maximize the conversion efficiency defined by the first ratio since ω_a is normally in the UV and is the lower energy of the two applied pulses.

Nonlinear Optics with a Pair of Strong Coupling Fields in Raman Resonance

An important extension of the EIT concept occurs when two strong fields are applied in Raman resonance between a pair of states in a three-level system. Considering the system illustrated in Figure 1 we can imagine that both applied fields are now strong. Under appropriate adiabatic conditions the system evolves to produce the maximum possible value for the coherence $\varrho_{12} = 0.5$. Adiabatic evolution into the maximally coherent state is achieved by adjusting either the Raman detuning or the pulse sequence (counter to-intuitive order). The pair of fields may also be in single-photon resonance with a third level, in which case the EIT-like elimination of absorption will be important. This situation is equivalent to the formation of a darkstate, since neither of the two strong fields is absorbed by the medium. For sufficiently strong fields the single-photon condition need not be satisfied and a maximum coherence will still be achieved.

An additional field applied to the medium can participate in sum- or difference-frequency mixing with the two Raman resonant fields. The importance of the large value of coherence is that it is the source polarization that drives the new fields generated in the frequency mixing process. Complete conversion can occur over a short distance that greatly alleviates the constraints usually set by phase-matching in nonlinear optics. Recently near unity conversion efficiencies to the far-UV were reported in an atomic lead system where maximum coherence had been created. In a molecular medium large coherence between vibrational or rotational levels has also been achieved using adiabatic pulse pairs. Efficient multi-order Raman sideband generation has been observed to occur. This latter observation may lead the way to synthesizing very short duration light pulses since the broadband Raman sideband spectrum has been proved to be phase-locked.

Pulse Propagation and Nonlinear Optics for Weak CW Fields

In a Doppler-free medium a new regime can be accessed. This is shown in Figure 4 where the possibility now arises of extremely narrow transparency dips since very small values of Ω_c are now sufficient to induce EIT. The widths of these features are typically subnatural and are therefore accompanied by very steep normal dispersion, which corresponds to a much reduced group velocity. The ultraslow propagation of pulses is one of the most dramatic manifestations of EIT in this regime. Nonlinear susceptibilities are now very large as there is constructive interference controlling the value and the splitting of the two states is negligible compared to their natural width. Nonlinear optics at very low light levels, i.e., at the few-photon limit, is possible in this regime.

Propagation of pulses is significantly modified in the presence of EIT. Figure 4 shows the changes to $\mathrm{Re}[\chi^{(1)}]$ that arise. Within the transparency dip there exists a region of steep normal dispersion. In the vicinity of resonance this is almost linear and it becomes reasonable to consider the leading term only that describes the group velocity. An analysis of the refractive changes has been provided by Harris who expanded the susceptibilities (both real and imaginary parts) of the dressed atom around the resonance frequency to determine various terms in $\mathrm{Re}[\chi^{(1)}]$. The first term of the series (zero order) $\mathrm{Re}[\chi^{(1)}](\omega_{13}) = 0$ corresponds to the vanishing dispersion at resonance. The next term $\partial[\mathrm{Re}\chi^{(1)}](\omega)/\partial\omega$ gives the slope of the dispersion curve; at ω_{13} this takes the value:

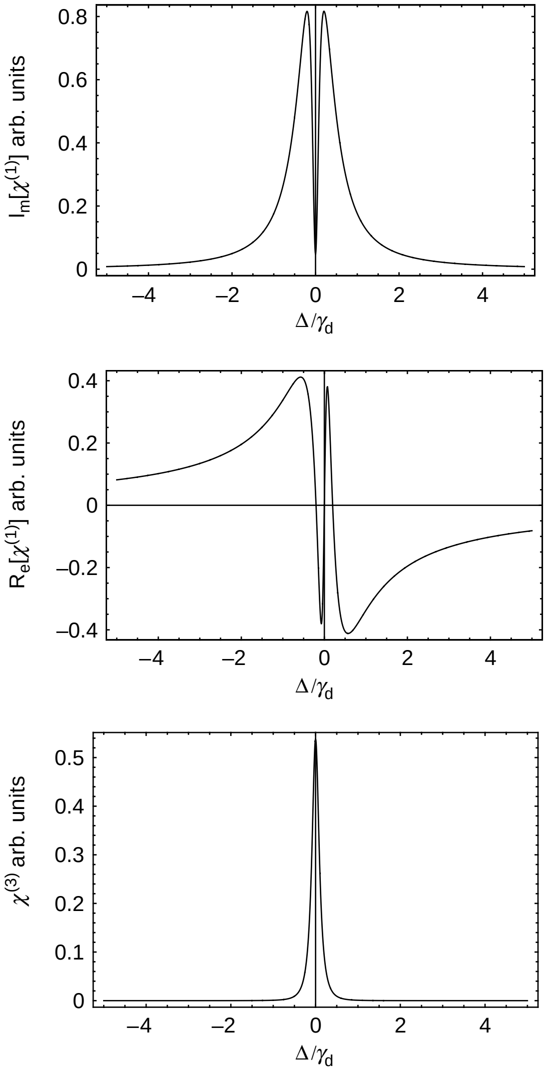

Figure 4 Electromagnetically induced transparency is shown in the case where there is no significant Doppler broadening. The values of the imaginary $\mathrm{Im}[\chi^{(1)}]$ and real $\mathrm{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. We take $\Omega_c = \gamma_d/5$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

$$\frac{\partial \mathrm{Re}\chi^{(1)}(\omega_{13})}{\partial \omega} = \frac{|\mu_{13}|^2 N 4(\Omega_c^2 - \gamma_a^2)}{2\pi\varepsilon_0(\Omega_c^2 + \gamma_a\gamma_d)} \qquad [19]$$

The slope of the dispersion profile leads to a reduced group velocity v_g:

$$\frac{1}{v_g} = \frac{1}{c} + \frac{\pi}{\lambda}\frac{\partial\chi^{(1)}}{\partial\omega} \qquad [20]$$

From the expression for $\partial\chi/\partial\omega$ we see that this slope is steepest (and so v_g mimimum) in the case where $\Omega_c \gg \Gamma_2$ and $\Omega_c^2 \gg \Gamma_2\,\Gamma_3$ but is still small compared to Γ_3 (i.e., $\Omega_C < \Gamma_3$) and hence $\partial\chi/\partial\omega \propto 1/\Omega_c^2$. In the limit of small Ω_c the following expression for v_g therefore holds:

$$v_g = \frac{\hbar c\varepsilon_0[\Omega_c]^2}{2\omega_d|\mu_{13}|^2 N} \qquad [21]$$

Extremely low group velocities, down to a few meters per second, are achieved in this limit using excitation of the hyperfine split ground states in either laser cooled atomic samples or Doppler-free configurations in finite temperature samples. Recently similar light slowing has been observed in solids. Storage of the optical pulse within the medium has also been achieved by adiabatically switching off the coupling field and thus trapping the optical excitation as an excitation within the hyperfine ground states for which the storage time can be very long (>1 ms) due to very low dephasing rates. Since the storage scenario should be valid even for single photons this process has attracted considerable attention recently as a means to enable quantum information storage.

Extremely low values of Ω_c are sufficient to induce complete transparency (albeit in a very narrow dip) and at this location the nonlinear response is resonantly enhanced. Very high efficiency nonlinear frequency mixing and phase conjugation at low light levels have been reported under these conditions. It is expected that high-efficiency nonlinear optical processes will persist to ultralow intensities (the few photon level) in an EIT medium of this type.

One example of highly enhanced nonlinear interactions is the predicted large values of the Kerr type nonlinearity (nonlinear refractive index). The origin of this can be seen by considering the steep dispersion profile in the region of the transparency dip in Figure 4. Imagine that we apply an additional field ω_f, perhaps a very weak one, at a frequency close to resonance between state |2> and a fourth level |4>. The ac Stark shift caused by this new field to the other three level, although small, will have a dramatic effect upon the value of the refractive index because of the extreme steepness of the dispersion profile. Essentially even a very small shift of the resonant wavelength causes a large change in the refractive index for a field applied close to this frequency. It is predicted that strong cross-phase modulations will be created in this process between the fields ω_f and ω_d, even in the quantum limit for these fields. This is predicted to lead to improved schemes for quantum nondemolition measurements of photons through the measurement of the phase-shifts they induce (through cross-phase modulation) on another quantum field. This type of measurement has direct application in quantum information processing.

See also

Nonlinear Optics, Applications: Phase Matching; Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Arimondo E (1996) Coherent population trapping in laser spectroscopy. *Progress in Optics* 35: 257–354.

Harris SE (1997) Electromagnetically induced transparency. *Physics Today* 50: 36–42.

Harris SE and Hau LV (1999) Non-linear optics at low light levels. *Physical Review Letters* 82: 4611–4614.

Harris SE, Field JE and Imamoglu A (1990) Nonlinear optical processes using electromagnetically induced transparency. *Physical Review Letters* 64: 1107–1110.

Hau LV, Harris SE, Dutton Z and Behroozi CH (1999) Light speed reduction to 17 metres per second in an ultra-cold atomic gas. *Nature* 397: 594–598.

Marangos JP (2001) Electromagnetically induced transparency. In: Bass M, *et al.* (ed.) *Handbook of Optics*, vol. IV, ch. 23. New York: McGraw-Hill.

Merriam AJ, Sharpe SJ, Xia H, *et al.* (1999) Efficient gas-phase generation of vacuum ultra-violet radiation. *Optics Letters* 24: 625–627.

Schmidt H and Imamoglu A (1996) Giant Kerr nonlinearities obtained by electromagnetically induced transparency. *Optics Letters* 21: 1936–1938.

Scully MO (1991) Enhancement of the index of refraction via quantum coherence. *Physical Review Letters* 67: 1855–1858.

Scully MO and Suhail Zubairy M (1997) Quantum Optics. Cambridge, UK: Cambridge University Press.

Sokolov AV, Walker DR, Yavuz DD, *et al.* (2001) Femtosecond light source for phase-controlled multi-photon ionisation. *Physical Review Letters* 87: 033402-1.

Zhang GZ, Hakuta K and Stoicheff BP (1993) Nonlinear optical generation using electromagnetically induced transparency in hydrogen. *Physical Review Letters* 71: 3099–3102.

ENVIRONMENTAL MEASUREMENTS

Contents

Doppler Lidar
Hyperspectral Remote Sensing of Land and the Atmosphere
Laser Detection of Atmospheric Gases
Optical Transmission and Scatter of the Atmosphere

Doppler Lidar

R M Hardesty, National Oceanic and Atmospheric Administration, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lidars developed for the measurement of atmospheric winds make use of the Doppler effect, in which radiation scattered or emitted from a moving object is shifted in frequency as a result of the movement of the particles. Atmospheric Doppler lidars irradiate a volume of atmosphere with a pulse of very narrowband, laser-produced radiation, then detect the change in frequency of the radiation backscattered from atmospheric aerosol particles or molecules present in the volume. Although Doppler lidars are conceptually similar to Doppler weather radars, lidar wavelengths are 3–4 orders of magnitude shorter than radar wavelengths, producing some important differences in propagation and scattering characteristics. Radiation at wavelengths in the optical region of the spectrum is efficiently scattered by aerosols and molecules, therefore Doppler lidars, unlike commonly used Doppler weather radars, do not require the presence of hydrometeors or insects to obtain useful results. Also, optical radiation can be tightly collimated, virtually eliminating effects of ground clutter and enabling lidar probing of small volumes to within a few meters of terrain or structures.

The primary disadvantage of lidar techniques is the severe attenuation of the optical radiation by cloud water droplets and fog. Doppler lidars do not typically probe beyond the edge of most atmospheric clouds – the one exception being tenuous ice clouds such as cirrus, which often are characterized by low optical extinction and high backscatter, making them excellent lidar targets. Doppler lidars also do not have the extended range, which can exceed 100 km, available from contemporary meteorological radars when scatterers are present.

In clear air, however, the characteristics and capabilities of Doppler lidar are well suited for obtaining detailed wind and turbulence observations for a wide variety of applications. Lidar beams can be easily scanned to characterize motions within very confined three-dimensional spaces such as shallow atmospheric boundary layers, narrow canyons, and turbulent structures. The relative compactness of Doppler lidars makes deployment on aircraft or other moving platforms extremely viable, and a Doppler lidar deployed on a satellite has been proposed as a way to obtain global measurements of atmospheric wind fields. By scanning a lidar beam from an orbiting satellite and analyzing the returns backscattered from clouds, aerosols, and molecules, a satellite-based instrument could provide important wind information for numerical forecast models.

Doppler Lidar Fundamentals

Backscattered Light

As laser light propagates through the atmosphere, some of the incident energy is scattered by the atmospheric molecules and constituents through which the light passes. In lidar applications, the light backscattered (scattered directly back toward the source) is collected and analyzed. The backscattering properties of an atmospheric particle depend on its refractive index, shape, and size. In the special case when the particle is much smaller than the wavelength of the incident radiation, as is the case for laser radiation scattered by atmospheric molecules, the scattering process is characterized as Rayleigh scattering. Because, in the Rayleigh scattering regime the energy backscattered by a particle increases proportionally as the inverse of the fourth power of the wavelength, Doppler lidar systems designed for molecular scatter operate at short wavelengths. Molecular Doppler lidar systems typically operate in the visible or ultraviolet spectral regions.

Aerosol particles, the other component of the atmosphere that scatters laser light, produce Mie scattering, which applies when the diameter of the scatterers is not orders of magnitude smaller than the incident wavelength. Aerosol particles present in the atmosphere include dust, soot, smoke, and pollen, as well as liquid water and ice. Although for Mie scattering the relationship between the power backscattered by an ensemble of aerosol particles and the incident wavelength is not simply characterized, most studies have shown that the energy backscattered increases roughly proportional to the first or second power of the inverse of the incident wavelength, depending on the characteristics of the scattering aerosol particles. In a polluted environment, such as in the vicinity of urban areas, an abundance of large particles often results in a roughly linear relationship between the inverse of the incident wavelength and the backscattered energy, while in more pristine environments, such as the free troposphere, the inverse wavelength/backscatter relationship can approach or exceed a square-law relationship.

The primary objective in Doppler lidar is to measure the Doppler frequency shift of the scattered radiation introduced by the movements of the scattering particles. Figure 1 shows a typical spectrum of the radiation collected at the lidar receiver for a volume of atmosphere irradiated by a monochromatic laser pulse. Molecular scattering produces the broadband distribution in Figure 1, where the broadening results from the Doppler shifts of the radiation backscattered from molecules moving at their thermal velocities. The width of the molecular velocity distribution in the atmosphere ranges from ~320 to 350 m s^{-1}, scaling as the square root of the temperature. In the center of the spectrum is a much narrower peak resulting from scattering of the light by aerosol particles. Because the thermal velocity of the much larger aerosol particles is very low, the width of the distribution of the aerosol return is determined by the range of velocities of particles moved about by small-scale turbulence within the scattering volume. This is typically on the order of a few m s^{-1}. Also shown in Figure 1 is an additional broadband distribution, due to scattered solar radiation collected at the receiver. If the laser source is not monochromatic, the backscattered signal spectrum is additionally broadened, with the resulting spectrum being the convolution of the spectrum shown in Figure 1 with the spectrum of the laser pulse. As seen in the figure, the entire spectrum is Doppler-shifted in frequency, relative to the frequency of the laser pulse. The object for a Doppler lidar system is to measure this shift, given by $\delta f = 2v_{\text{rad}}/\lambda$, where v_{rad} is the component of the mean velocity of the particles in the direction of propagation of the lidar pulse and λ is the laser wavelength.

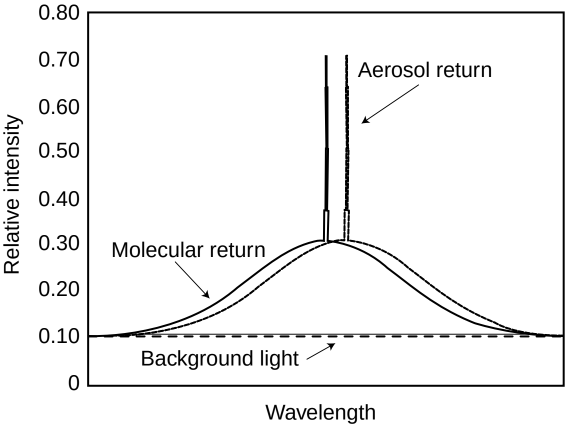

Figure 1 Distribution of wind velocities in a lidar return, showing Doppler shift resulting from wind-induced translation of the scatterers. The ratio of aerosol to molecular signals increases with increasing wavelength.

Elements of a Doppler Lidar

Doppler lidar systems can be designed primarily to measure winds from aerosol-scattered radiation, molecule-scattered radiation, or both. The type of system places specific requirements on the primary components that comprise a Doppler lidar system. A Doppler lidar consists of a laser transmitter to produce pulses of energy that irradiate the atmospheric volume of interest, a receiver which collects the backscattered photons and estimates the energy and Doppler shift of the return, and a beam pointing mechanism for directing the laser beam and receiver field of view together in various directions to probe different atmospheric volumes. Independent of whether the primary scatterers are molecules or aerosol particles, the system design criteria for a Doppler system are driven by a fundamental relationship between the error in the estimate of mean frequency shift δf_1, the bandwidth of the return f_2 (proportional to the distribution of velocities), and the number of incident backscattered photons detected N as

$$\delta f_1 \propto f_2/N^{1/2} \tag{1}$$

Thus, the precision of the velocity estimate is improved by increasing the number of detected signal photons and/or decreasing the bandwidth of the backscattered signal. It is obvious from the equation that a significantly greater number of photons are required to achieve the same precision in a Doppler measurement from a molecular backscattered signal, characterized by higher bandwidth f_2, compared to the number required to compute velocities from aerosol returns. The improved measurement precision gained from a narrow bandwidth return also indicates that the laser transmitter in a Doppler lidar system should be characterized by narrow spectral width (considerably narrower than the spectral broadening caused by the distribution of velocities), and to increase the number of photons detected, maximum transmit energy.

Narrowband, single-frequency pulses are obtained by employing a low-power, precisely frequency-stabilized master oscillator laser, whose radiation is either used to seed a stabilized, higher-energy laser cavity (injection seeding) or amplified in one or more optical amplifiers (master-oscillator, power amplifier, (MOPA)) to produce frequency-stabilized pulses.

The lidar receiver gathers the backscattered photons and extracts the wind velocity as a function of the range to the scattering volume. This requires a telescope, to gather and focus the scattered radiation, and a receiver to detect the scattered radiation and determine the Doppler shift. Frequency analysis in a Doppler lidar receiver is carried out using one of two techniques: coherent detection (also known as heterodyne detection) or direct detection (alternately labeled incoherent detection). The techniques differ fundamentally. In a heterodyne receiver, frequency analysis is carried out on a digitized time series created by mixing laser radiation with the backscattered radiation, while in direct detection lidar an interferometer optically analyzes the backscattered radiation to produce a light pattern which contains the information on the frequency content.

Coherent (Heterodyne) Doppler Lidar

Description

Coherent or heterodyne lidar is implemented by optically mixing the backscattered laser light with radiation from a stable, continuous-wave, local oscillator (LO) laser whose frequency is precisely controlled to be a known frequency offset, typically on the order of tens of MHz, from that of the laser transmitter (Figure 2). The mixing process at the face of an optical detector generates, after high pass filtering, an electrical signal with amplitude proportional to amplitude of the backscattered electromagnetic field and frequency equal to the difference between the backscattered field frequency and the LO laser field frequency. This signal is sampled, and then digitally processed to estimate the range-dependent mean frequency shift of the backscattered signal, from which the radial wind component can be derived. A property of coherent lidar is that, because of constructive and destructive interference of the returns from individual scatterers, single pulse returns are characterized by random fluctuations in the amplitude and phase of the detected time series and Fourier spectrum. Consequently, averaging (sometimes referred to as accumulation) over multiple pulses usually is necessary to produce a sufficiently precise estimate of the signal mean frequency estimate. Because the phase of the detected time series is random within each pulse, averaging must be carried out in the power spectrum or autocorrelation function domain, rather than on the detected time series.

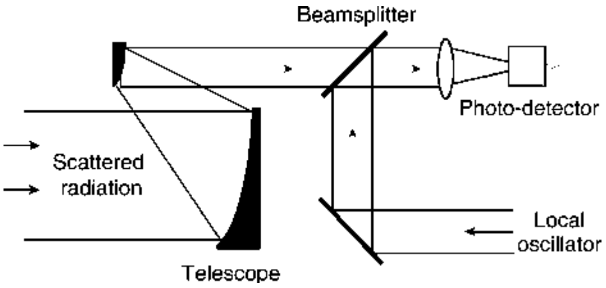

Figure 2 Schematic of coherent detection of backscattered radiation.

The optical mixing process in coherent Doppler lidar provides both benefits and design challenges. Because receiver efficiencies (equivalent fraction of incident photons entering the receiver that are converted to electrons) are quite high, wind velocities can be estimated from very weak backscattered signals. Also, signal processing is performed on a time series derived from the mixed signal, enabling the use of electrical filters to produce a narrow receiver bandwidth and effectively eliminating broadband solar background light as a major source of noise. As a result, unlike many lidars coherent Doppler lidar performance is not degraded under daylight conditions.

The primary limitation of coherent detection results from the added photon noise from the local oscillator radiation (which is usually much stronger than the backscattered radiation). The effect of local oscillator photon noise is to define a system noise level below which very weak signals cannot be practically extracted, even with substantial multiple-pulse averaging. In addition, because efficient mixing requires phase coherence of the backscattered signal field across the detector, coherent lidar performance at longer ranges is degraded by strong optical turbulence along the path of the laser pulse. These adverse effects of turbulence increase for shorter wavelength systems.

Although measurements at wavelengths near 1.06 μm have been demonstrated, coherent Doppler lidar systems used for regular atmospheric probing operate in the eye-safe, infrared portion of the spectrum at wavelengths longer than 1.5 μm. Currently, the two most common system wavelengths for coherent lidar wind systems are in atmospheric window spectral regions around 2 and 10.6 μm. Early coherent Doppler lidar measurements, beginning in the 1970s, employed CO_2 laser transmitters and local oscillators operating at wavelengths near 10.6 μm. Pulsed CO_2 laser transmitters with as much as 10 J of energy have since been demonstrated, and systems with 1 J lasers have probed the atmosphere to ranges of 30 km or more. In the late 1980s, solid state laser transmitters operating near 2 μm wavelengths were introduced into coherent lidar wind-measuring systems. The compact size and potential reliability advantages of solid-state transmitters, in which the transmitter laser is optically pumped by an array of laser diodes, provide advantages over larger CO_2 laser technology. Also, because for a given measurement accuracy the range-resolution obtainable is proportional to wavelength, 2 μm instruments have an enhanced capability to probe small-scale features. However, although development of higher energy, single-frequency coherent lidar sources operating at 2 μm region, as well as at 1.5 μm, is currently an active research area, solid state lasers with pulse energies greater than several tens of mJ have yet to be incorporated into lidar systems for atmospheric probing.

Applications of Coherent Doppler Lidar

Coherent lidars have been used to measure winds for a variety of applications, and from an assortment of platforms, such as ships and aircraft. Since these lidars operate in the infrared where aerosol scattering dominates molecular scattering, they require aerosol particles to be present at some level to obtain usable returns. Although clouds also provide excellent lidar targets, most of the more useful applications of coherent lidars involve probing the atmospheric boundary layer or lower troposphere where aerosol content is highest. Because of the capability to scan the narrow lidar beam directly adjacent to terrain, a unique application of lidar probing is the measurement of wind structure and evolution in complex terrain such as mountains and valleys. Over the past two decades, Doppler lidar studies have been used, for example, to study structure of damaging windstorms on the downwind side of mountain ranges, advection of pollution by drainage flows from valleys, and formation of mountain leeside turbulence as a potential hazard to landing aircraft. The interactions of wind flows with complex terrain can produce dangerous conditions for transportation, especially aircraft operations. Figure 3 shows a strong mountain wave measured by a Doppler lidar near Colorado Springs, during an investigation of the effects of downslope winds and turbulence on aircraft operations at the Colorado Springs airport. The mountains are on the right of the figure, where the strong wave with wavelength of about 6 km can be seen. Note the reversal in winds near the location where the wave descends to the surface. The Colorado Springs study was prompted by the crash of a passenger jet while landing at the airport during a period when the winds were down the slope of the Rocky Mountains just to the west of the airport.

A Doppler lidar was recently deployed in an operational mode for wind shear detection at the new Hong Kong International Airport. Because winds flowing over mountainous terrain just south of the

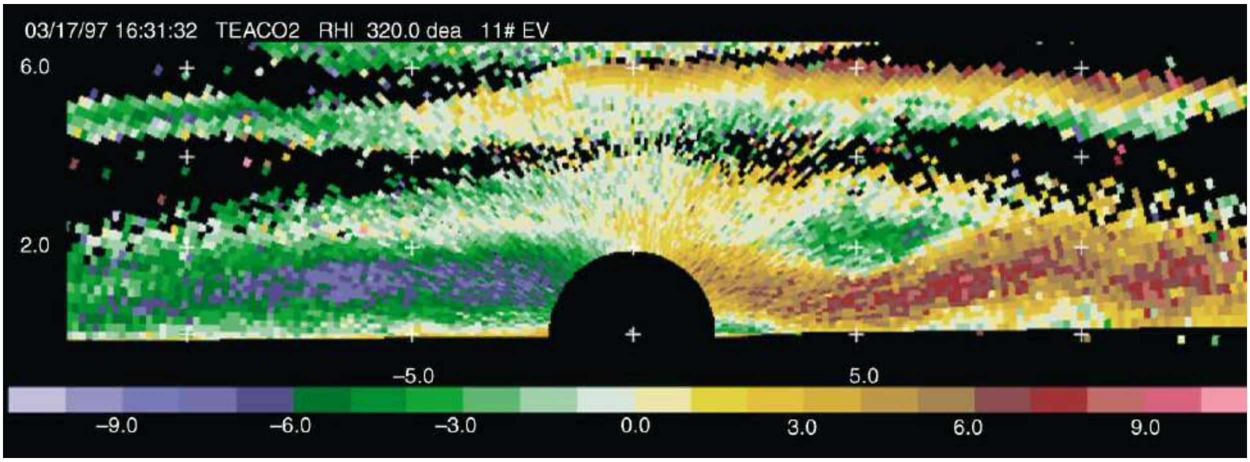

Figure 3 Vertical scan of the winds in a mountain wave measured by a 10.6 μm coherent Doppler lidar near Colorado Springs, Colorado. Courtesy L. Darby, NOAA.

airport can produce mountain waves and channeled flows, wind shear is frequently encountered during landing and takeoff operations. Figure 4 shows a gust front situated just to the west of the airport, characterized by a sharp wind change (which would produce a corresponding change in airspeed) approaching 13 m s^{-1} over a distance of less than one kilometer. Providing a capability to detect such events near the airport and warn pilots during approach or preparation for takeoff was the primary reason for deployment of the lidar. Within the first year of operation the Hong Kong lidar was credited with improving wind shear detection rates and providing more timely warnings for pilots of the potentially hazardous conditions. As of August 2003 this lidar had logged more than 10 000 hours of operation.

Because coherent Doppler lidars are well matched to applications associated with probing small-scale, turbulent phenomena, they have also been deployed at airports to detect and track wing tip vortices generated by arriving or departing aircraft on nearby parallel runways. In the future, a network of ground-based lidars could provide information on vortex location and advection speed as well as wind shear, leading to a potential decrease in congestion at major airports. Also, compact lidar systems looking directly ahead of research aircraft have shown the capability to detect wind changes associated with potentially hazardous clear-air turbulence. In the future, a Doppler lidar installed on the commercial aircraft fleet could potentially look ahead and provide a warning to passengers to fasten seat belts before severe turbulence is encountered.

The high resolution obtainable in a scanning lidar enables visualization of finely structured wind and turbulence layers. Figure 5 shows an image of turbulence associated with a nocturnal low-level jet just 50 m above the surface obtained at night over flat

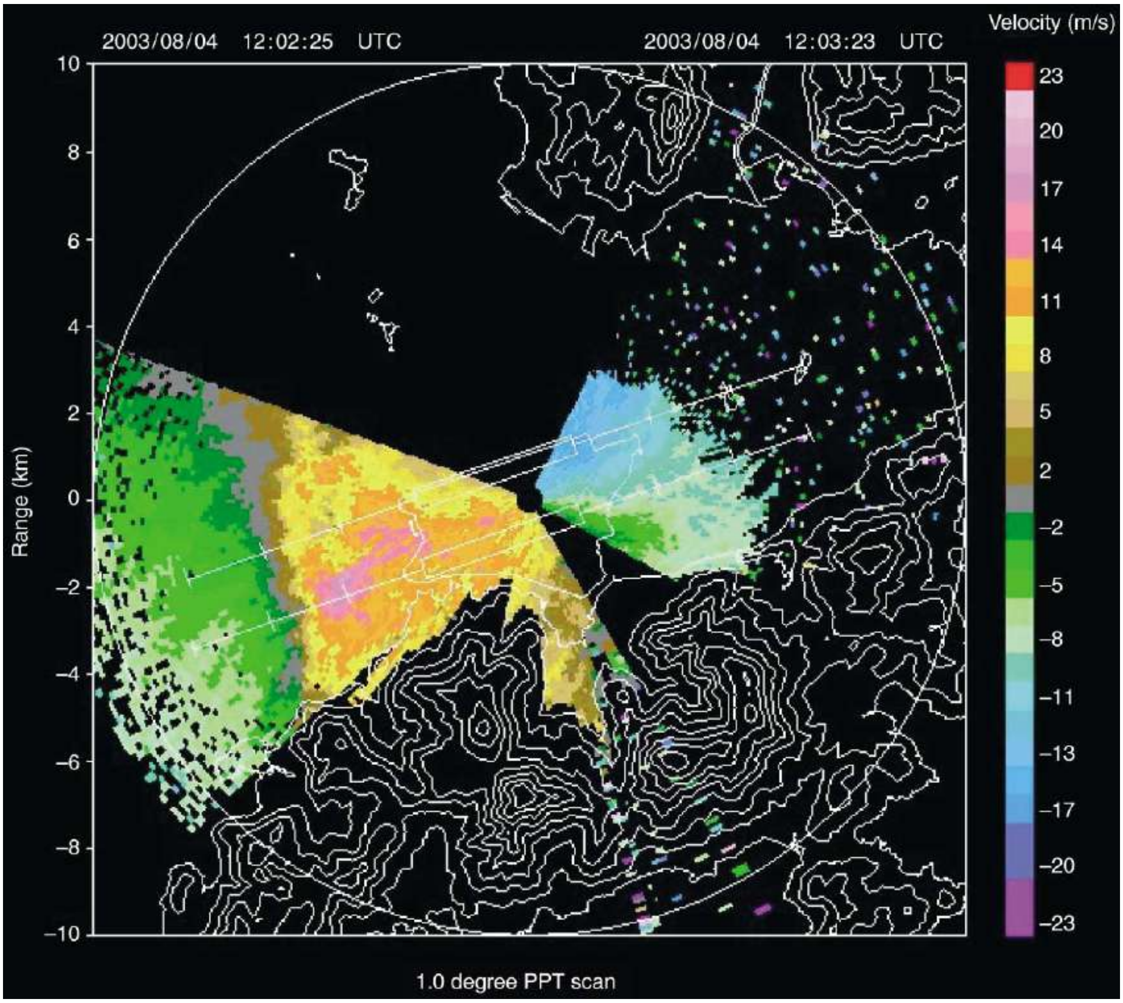

Figure 4 Horizontal depiction of the radial component of the wind field measured at Hong Kong Airport by a 2.02 μm Doppler lidar during a gust front passage, showing strong horizontal shear just west of the airport. Courtesy C. M. Shun, Hong Kong Observatory.

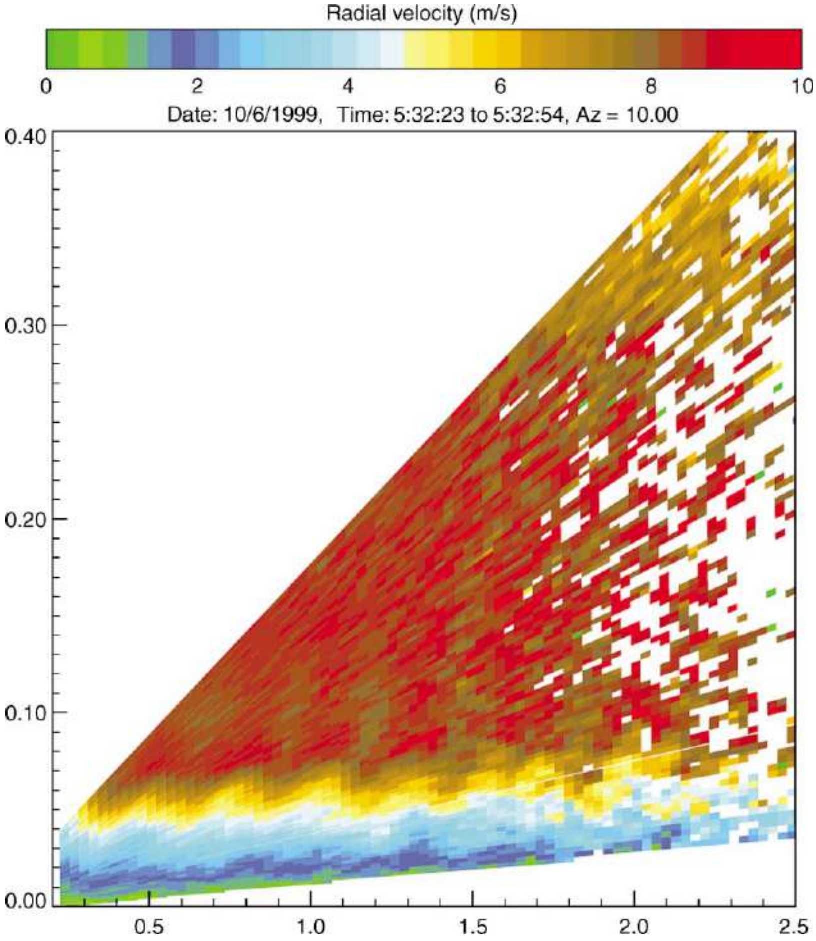

Figure 5 Turbulence structure within a nocturnal stable layer at a height of 50 m. Note the stretching of the vertical scale. Courtesy R. Banta, NOAA.

terrain in Kansas. Although a low-level jet was present on just about every evening during the experiment, similar images obtained at different times by the Doppler lidar illustrated markedly different characteristics, such as a wide variation in the observed mechanical turbulence along the interface. Such observations enable researchers to improve parameterizations of turbulence in models and better understand the conditions associated with vertical turbulent transport and mixing of atmospheric constituents such as pollutants.

By deploying Doppler lidars on moving platforms such as ships and aircraft, the spatial coverage of the measurements can be greatly increased. Although removal of platform motion and changes in orientation are not trivial, aircraft-mounted lidar can map out such features as the low-level jet and the structure of boundary layer convective plumes. Figure 6 shows horizontal and vertical motions associated with convective plumes measured along an approximately 60 km path over the southern Great Plains of the United States. From such measurements estimates of turbulence intensity, integral scale, and other boundary layer characteristics can be computed.

Direct Detection Doppler Lidar

Description

Direct detection or incoherent Doppler lidar has received significant attention in recent years as an alternative to coherent lidar for atmospheric wind measurements. In contrast to coherent lidar, in which an electrical signal is processed to estimate Doppler shift, an optical interferometer, usually a Fabry–Perot etalon, serves as the principal element in a direct detection lidar receiver for determining the frequency shift of the backscattered radiation.

One implementation of a direct-detection Doppler lidar receiver is the ‘fringe imaging’ technique. In this design, the interferometer acts as a spectrum analyzer. The backscatter radiation is directed through a

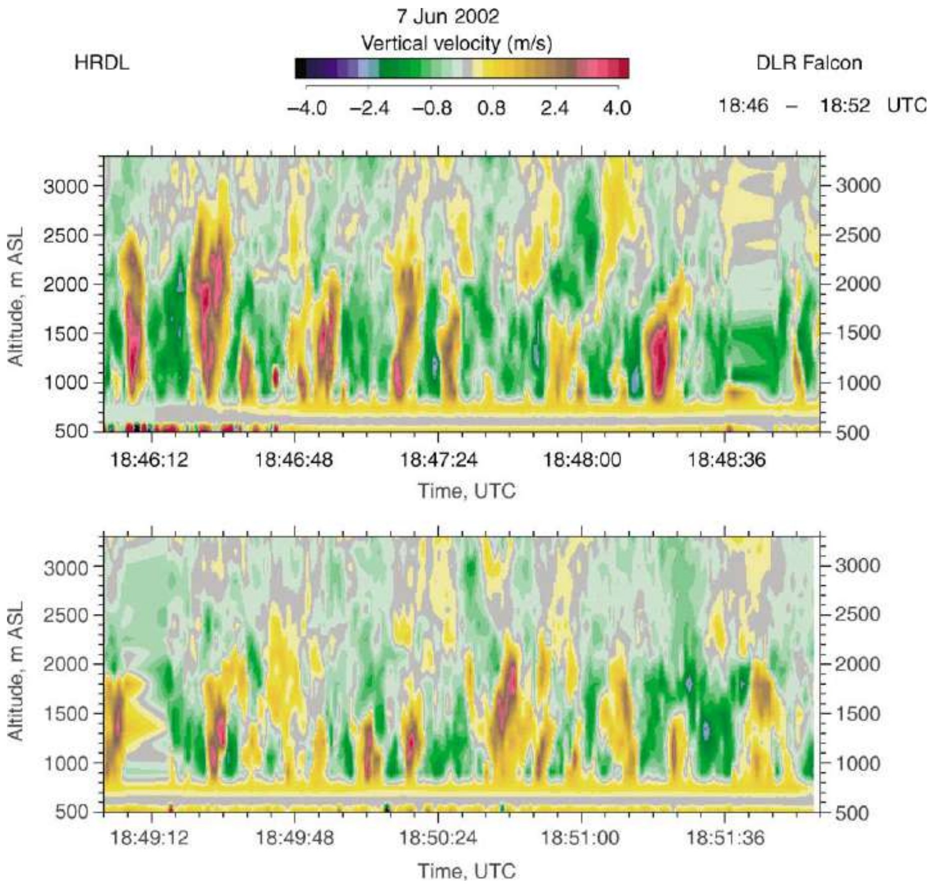

Figure 6 Vertical motion in convective plumes measured by a vertically pointing airborne coherent Doppler lidar. Total horizontal extent of the measurements is approximately 60 km, reddish colors correspond to upward motions.

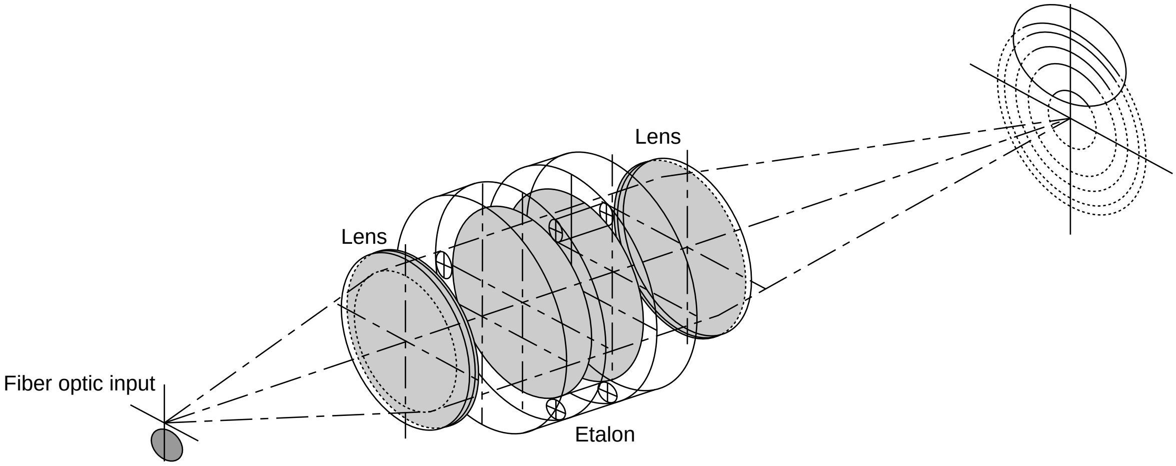

Figure 7 Schematic of Fabry–Perot etalon in a direct detection, fringe-imaging lidar receiver. Courtesy P. Hays, Michigan Aerospace Corporation.

Fabry–Perot interferometer, which produces a ring pattern in the focal plane (Figure 7). The spectral content information of the incident radiation is contained in the radial distribution of the light. Each ring corresponds to an order of the interferometer and is equivalent to a representation of the backscattered signal frequency spectrum. As the mean frequency of the backscattered radiation changes, the rings move inward or outward from the center. To extract a spectrum of the backscattered light, one or more of the rings are imaged onto a two-dimensional detector, and the resulting pattern analyzed. The rings can either be detected using a multi-element ring detector, or can be converted to a linear pattern with a circle-to-line converter optic and then imaged on to a charge-coupled device array device.

An alternate direct detection receiver configuration, generally called the double edge technique, is to employ two interferometers as bandpass filters, with the center wavelength of each filter set above and below the laser transmitter wavelength, as shown in Figure 8. The incoming radiation is split between the two interferometers, and the wavelength shift is computed by examining the ratio of the radiation transmitted by each interferometer. Both the double edge and fringe-imaging techniques have demonstrated wind measurements to heights well into the stratosphere. The major challenge associated with the double edge receiver is optimizing the instrument when both aerosol and molecular scattered radiation are present, since in general the change in transmission as a function of velocity is different for the aerosol and molecular signals. Double edge receivers optimized for both aerosol and molecular returns place the bandpass of the etalon filters at the precise wavelength where the change in transmission with change in Doppler shift is the same for both aerosol and molecular returns.

For both types of direct detection receivers described above, much of the radiation incident on the interferometer is reflected out of the system, reducing the overall efficiency of the receiver. Recently, designs that incorporate fiber optics to collect a portion of the reflected radiation and 'recycle' it back into the etalon have been demonstrated as a method to improve the transmission efficiency of the etalon in a fringe imaging receiver.

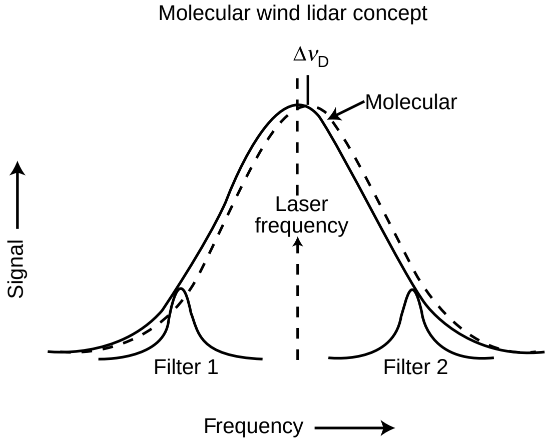

Figure 8 Spectrum of molecular lidar return showing placement of bandpass filters for a dual channel (double edge) direct detection receiver. Courtesy B. Gentry, NASA.

Doppler Wind Measurements Based on Molecular Scatter

One of the primary advantages of direct detection Doppler lidar is its capability for measurements based on scatter from atmospheric molecules. Measurement of Doppler shifts from molecular scattered radiation is challenging because of the large Doppler-broadened bandwidth of the return. Because one wants to measure a mean wind velocity with a precision of a few m s^{-1} or better, in the order of 10^5 photons are required. Some combination of multiple pulse averaging, powerful lasers, and large receiver optics is required to obtain these high photon counts from backscattered returns.

Molecular-scatter wind measurements have been demonstrated in the visible spectral region at 532 nm wavelength as well as at 355 nm in the near ultraviolet. The ultraviolet region has the dual advantages of enhanced molecular scatter and less restrictive laser eye-safety restrictions. Figure 9 shows the time series of a wind profile measured in the troposphere from both aerosols and clouds using a molecular-scatter, ground-based 355-nm wavelength Doppler fringe-imaging lidar. The figure shows measurements from receiver channels optimized for the wideband molecular signal and the narrowband aerosol return. For this measurement, backscattered photons were collected by a 0.5 m receiver aperture, averaged for 1 minute, and processed. In the absence of clouds, direct detection Doppler lidars have measured wind profiles continuously from the surface to beyond 15 km height. Figure 10 shows that the estimated wind error for the same 355 nm lidar is less than 1 m s^{-1} to about 10 km, and less than 4 m s^{-1} at 15 km height.

Heterodyne and Direct-Detection Doppler Trade-Offs

Lively debates within the lidar community have occurred over the past decade regarding the relative merits of heterodyne versus direct detection Doppler lidars. To a large extent, the instruments are complementary. Generally, heterodyne instruments are much more sensitive when significant aerosols are present. Coherent lidar processing techniques have been developed that can produce accurate wind measurement rates using only a few lidar pulses with as few as 100 detected photons, such that several wind observations per second can be obtained for nominal pulse rates. This inherent sensitivity has led to numerous applications in which a lidar beam has been scanned rapidly over a large volume to obtain time-varying, three-dimensional wind measurements.

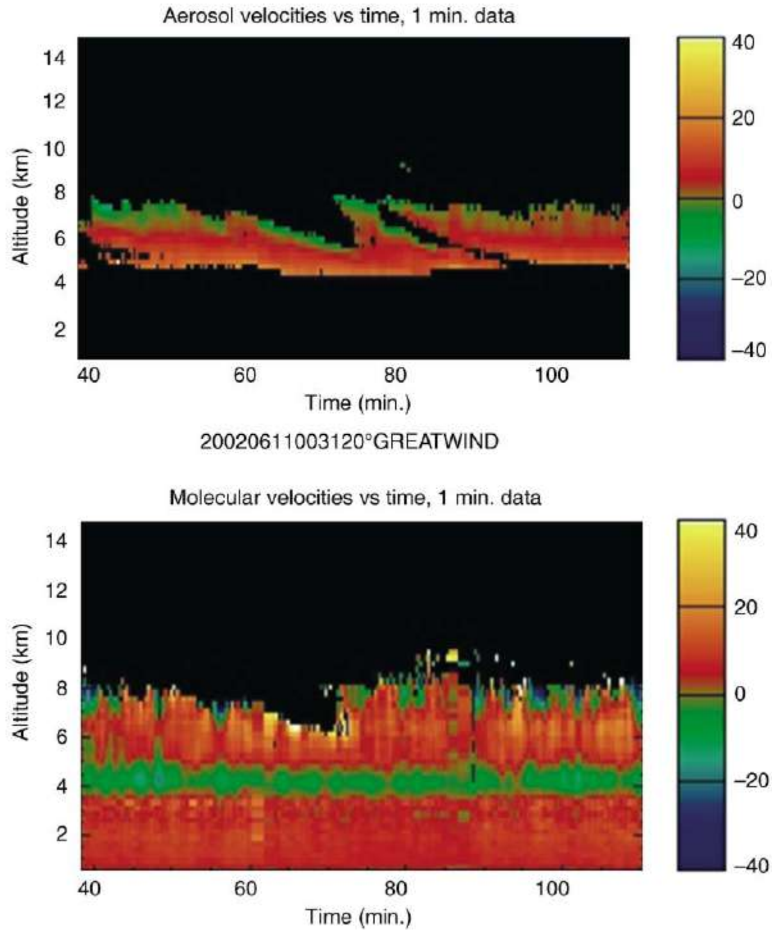

Figure 9 Time series of radial wind speed profiles measured by a 355 nm fringe-imaging direct detection lidar aerosol channel (top) and molecular channel (bottom) at Maura Loa Observatory, HI. Change in wind speed and blocking of the return by clouds is clearly seen. Courtesy C. Nardell, Michigan Aerospace Corp.

The primary advantage of direct detection instruments is their demonstrated capability to measure winds from molecular-backscattered returns in the middle and upper troposphere. In very pristine air, direct detection offers the only method for long-range wind measurements, even though significant averaging may be required. Direct-detection lidars have the additional advantage of being unaffected by atmospheric refractive turbulence. Unlike heterodyne lidars, which require a very pure laser pulse and a diffraction-limited receiver field of view matched to and precisely aligned with the transmitted beam, direct detection systems can have a wider bandwidth transmitter and a receiver field of view several times diffraction-limited. In direct detection lidar design, the field of view is usually constrained by the need to limit background light during daytime operation. A major design challenge for direct detection instruments is holding the Fabry–Perot etalon plate spacing stable over a range of temperatures and in high vibration environments.

Global Wind Measurements

A satellite-based Doppler lidar has frequently been proposed as a way to measure wind fields over most of the Earth. At present, winds are the one major meteorological variable not well-measured from orbiting platforms. Measurement of winds is especially important over regions of the Earth that are not currently well sampled, such as over Northern Hemisphere oceans, as well as over most of the tropics and Southern Hemisphere. Wind profile information is currently obtained from radiosondes and by tracking cloud and water vapor inhomogeneities using satellite imagers. Doppler lidar wind measurements would greatly augment the current data set by providing wind estimates throughout the troposphere under clear conditions, and highly height-resolved observations down to cloud tops when cloud decks are present. Observing system simulation experiments conducted in recent years indicate that satellite-based lidar global wind

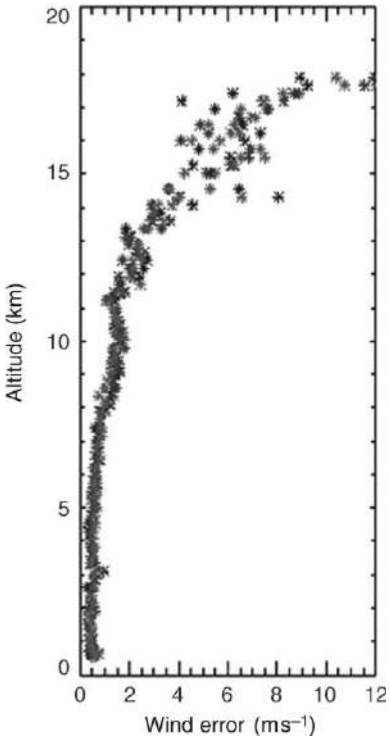

Figure 10 Estimated error in the radial velocity estimate versus altitude for a 355 nm direct detection, fringe imaging Doppler lidar system at Mauna Loa observatory on Hawaii. The lidar provides data for assimilation into local mesoscale forecast models. Courtesy C. Nardell, Michigan Aerospace Corp.

measurements could lead to a significant improvement in long-term forecast skill, provided the wind fields can be observed with sufficient accuracy and spatial resolution.

In a Doppler lidar wind mission, a satellite carrying a lidar system would orbit the Earth in a nearly polar orbit. The pulsed laser beam would be scanned conically about the nadir to obtain different components of the wind velocity. The scanning could be either continuous or ‘stop and stare’. After sufficient returns are averaged at a given pointing angle to produce an acceptable estimate, the radial component of the velocity would be computed and assimilated directly into numerical analysis and forecast models.

Doppler lidar measurement of winds from space is theoretically feasible but technologically difficult. Depending on the orbital height, the scattering volume is anywhere from 450 to ~850 km from the satellite, which challenges the sensitivity of current system types. Because weight and power consumption are critical parameters for space systems, telescope diameter and laser power cannot be easily increased to obtain the necessary sensitivity. Similarly, the ability to average returns from multiple pulses is also limited by time limitations. Because a satellite moves at about 7 km s^{-1}, in order to obtain measurements over a horizontal distance of 300 km (the resolution of the radiosonde network) only about 45 seconds are available to make enough observations from multiple look angles to obtain a useful measurement. It should also be noted that, as a result of the high orbital velocity of the satellite, precise knowledge of beam pointing is extremely critical for measurements from satellites. For a lidar operating at a nadir angle of 45 degrees, an error in the knowledge of pointing angle of just 1 mrad results in an error of about 5 ms^{-1} in the measured radial component of the wind.

Despite the challenge of employing a satellite-based Doppler lidar, efforts are continuing to develop the appropriate technology and to assess the impact of the observations. The European Space Agency is planning a Doppler wind lidar demonstration mission for the late 2000s that would incorporate a nonscanning, direct-detection instrument with both aerosol and molecular channels. Doppler lidar technology research and numerical simulations aimed at satellite-based wind sensing is ongoing at several research centers within the United States, Europe, and Japan. One option currently being studied is a ‘hybrid’ lidar system that would combine a direct detection lidar for measurement of winds in the clear free troposphere with a low-energy coherent system that would obtain its observations from clouds and the aerosol-rich boundary. Although a hybrid instrument reduces some requirements on system power and aperture size, combining the two techniques on a single platform will likely require innovative engineering and new approaches to beam combining and scanning.

See also

Environmental Measurements: Optical Transmission and Scatter of the Atmosphere. **Scattering:** Scattering Theory.

Further Reading

Baker WE, Emmitt GD, Robertson F, *et al.* (1998) Lidar-measured wind from space: a key component for weather and climate prediction. *Bulletin of the American Meteorological Society* 79: 581–599.

Banta RM, Olivier LD and Levinson DH (1993) Evolution of the Monterey Bay sea breeze layer as observed by pulsed Doppler lidar. *Journal of Atmospheric Science* 50 (24): 3959–3982.

Banta RM, Olivier LD, Gudiksen PH and Lange R (1996) Implications of small-scale flow features to modeling dispersion over complex terrain. *Journal of Applied Meteorology* 35: 330–342.

Banta RM, Darby LS, Kaufmann P, Levinson DH and Zhu C-J (1999) Wind flow patterns in the Grand Canyon as revealed by Doppler lidar. *Journal of Applied Meteorology* 38: 1069–1083.

Frehlich R (1996) Coherent Doppler lidar measurements of winds in the weak signal regime, 1997. *Applied Optics* 36: 3491–3499.

Gentry B, Chen H and Li SX (2000) Wind measurements with a 355 nm molecular Doppler lidar. *Optics Letters* 25: 1231–1233.

Grund CJ, Banta RM, George JL, *et al.* (2001) High-resolution Doppler lidar for boundary-layer and cloud research. *Journal of Atmosphere and Oceanic Technology* 18: 376–393.

Huffaker RM and Hardesty RM (1996) Remote sensing of atmospheric wind velocities using solid state and CO_2 coherent laser systems. *Proceedings of the IEEE* 84: 181–204.

Menzies RT and Hardesty RM (1989) Coherent Doppler lidar for measurements of wind fields. *Proceedings of the IEEE* 77: 449–462.

Post MJ and Cupp RE (1990) Optimizing a pulsed Doppler lidar. *Applied Optics* 29: 4145–4158.

Rees D and McDermid IS (1990) Doppler lidar atmospheric wind sensor: reevaluation of a 355 nm incoherent Doppler lidar. *Applied Optics* 29: 4133–4144.

Rothermel J, Cutten DR, Hardesty RM, *et al.* (1998) The multi-center airborne coherent lidar atmospheric wind sensor. *Bulletin of the American Meteorological Society* 79: 581–599.

Skinner WR and Hays PB (1994) Incoherent Doppler lidar for measurement of atmospheric winds. *Proceedings of SPIE* 2216: 383–394.

Souprayen C, Garnier A and Hertzog A (1999) Rayleigh–Mie Doppler wind lidar for atmospheric measurements. II: Mie scattering effect, theory, and calibration. *Applied Optics* 38: 2422–2431.

Souprayen C, Garnier A, Hertzog A, Hauchecorne A and Porteneuve J (1999) Rayleigh–Mie Doppler wind lidar for atmospheric measurements. I: Instrumental setup, validation, first climatological results. *Applied Optics* 38: 2410–2421.

Hyperspectral Remote Sensing of Land and the Atmosphere

W H Farrand, Space Science Institute, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hyperspectral remote sensing is a true marriage of imaging technology with spectroscopy. Hyperspectral remote sensing systems (also known as imaging spectrometer systems) fully sample the optical wavelength range of interest, whether it be the reflected solar range (0.35 to 2.5 μm) or the range of thermal emission from the Earth's surface (3.0 to 14 μm). A hyperspectral sensor views the Earth's surface in a series of contiguous and spectrally narrow image bands. Figure 1 presents the concept of hyperspectral remote sensing, wherein each spatial element of an image has an associated full resolution spectrum. The calibrated reflectance or emittance spectra collected by a hyperspectral system are meant to be directly comparable with that of materials measured in the laboratory. With such high spectral resolution, it thus becomes possible to do reflectance or emittance spectroscopy of the Earth's surface from an overhead perspective.

Underlying Principles

Optical remote sensing is done over wavelength intervals, or windows, in which the atmosphere is largely transparent. For the reflective solar portion of the spectrum, the atmosphere is mostly transparent from approximately 0.35 to 2.5 μm. Water, CO_2, and other gases have absorptions of varying strength in this range. This window is subdivided into the range of human vision (the visible) and the region of slightly longer wavelengths known as the near infrared; together they are the visible/near infrared or VNIR and extend from approximately 0.35 to 1.0 μm. The region from approximately 1.0 to 2.5 μm is known as the short wavelength infrared or SWIR. In the range of emitted terrestrial radiation there are two window regions. The first extends from approximately 3 to 5 μm and is known as the medium wavelength infrared or MWIR, and the second, the long wavelength infrared or LWIR, extends from approximately 8 to 14 μm.

The light that is intercepted by the entrance aperture of a sensor is known by the quantity of radiance which is measured in units of microwatts per square centimeter per nanometer per unit of solid angle. Physical quantities related to the material properties of Earth surface materials are reflectance and emittance. Both properties are the result of ratios and so are unitless. One definition of reflectance is that it is the ratio of radiance reflected from a material, divided by the radiance reflected from an identically illuminated perfectly diffuse reflector. Likewise, a definition for emittance is the ratio of radiance emitted from a material, divided by the ratio

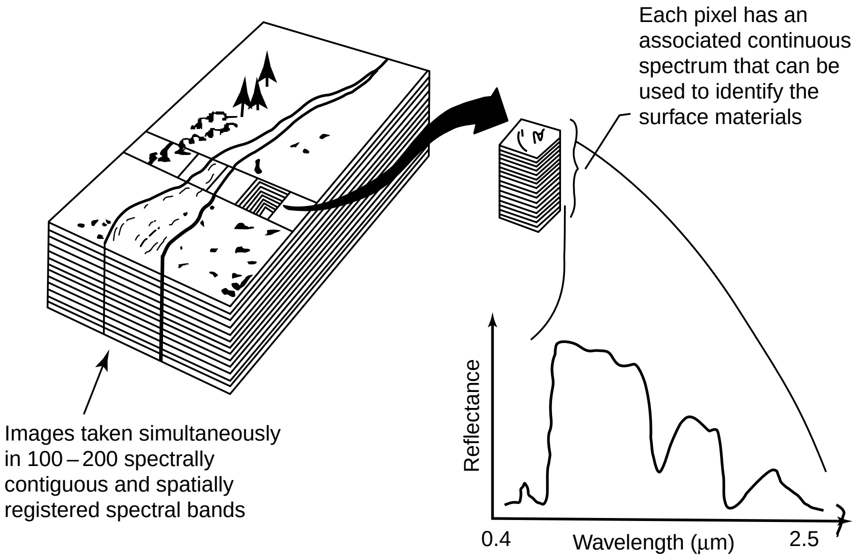
Each pixel has an associated continuous spectrum that can be used to identify the surface materials
Images taken simultaneously in 100 – 200 spectrally contiguous and spatially registered spectral bands
Reflectance
0.4
Wavelength (µm)
2.5

Figure 1 Hyperspectral remote sensing concept.

of radiance emitted from a perfectly emitting material of the same temperature.

Materials covering the Earth's surface, or gases in the atmosphere, can be identified in hyperspectral data on the basis of absorption features, or bands, in the spectra recorded by the sensor. Absorption bands result from the preferential absorption of energy in some wavelength interval. The principal types of processes by which light can be absorbed include, in order of decreasing energy of the process, electronic charge transfers, electronic crystal field effects, and molecular vibrations. The former two processes are observed in minerals and manmade materials that contain transition group metal cations, most usually, iron. Charge transfer absorptions are the result of cation-to-anion or cation-to-cation electron transfers. Crystal field absorptions are explainable by the quantum mechanical theory of atomic structure, wherein an atom's electrons are contained in orbital shells. The transition group metals have incompletely filled *d* orbital shells. When a transition metal cation, such as iron, is surrounded by anions, and potentially selects other cations, an electric field, known as the crystal field, exists. Crystal field absorptions occur when radiant energy causes that cation's orbital shell energy levels to be split by interaction with the crystal field. The reflectance spectra of different iron-bearing minerals are unique, because the iron cation has a unique positioning with respect to the anions and other cations that compose the mineral.

A material's component molecules have bonds to other molecules and as a result of interaction with radiant energy, these bonds can stretch or bend. These molecular vibrations, and overtones of those vibrations, are less energetic processes than the electronic absorption processes described above and so the resulting absorption features occur at longer wavelengths. Fundamental vibrational features occur in the MWIR and LWIR. Overtones and combination overtones of these vibrations are manifested as weaker absorption features in the SWIR. These overtone features in the SWIR include absorptions diagnostic of carbonate and certain clay minerals. Absorption features caused by vibrational overtones of C–H, O–H, and N–H stretches are also manifested in the SWIR and these absorption features are characteristic of important vegetative biochemical components such as cellulose, lignin, starch, and glucose. An example of the effect of the changes in reflected energy recorded at a sensor, due to the absorption of incident energy by green vegetation and other surface materials, is provided in Figure 2.

Laboratory spectrometers have been used for many years to analyze the diagnostic absorptions of materials caused by the above effects. More recently, spectrometers were mounted on telescopes to determine the composition of the Moon and the other solid bodies in our solar system. Technology progressed to where profiling spectrometers (instruments that measure a successive line of points on the surface) could be mounted on aircraft. The next logical step was the construction of imaging spectrometers (hyperspectral sensors) that measured spectra in two spatial dimensions. In fact, the data produced by a hyperspectral sensor are often thought of as an image cube (Figure 3) because they consist of three dimensions: two spatial and one spectral.

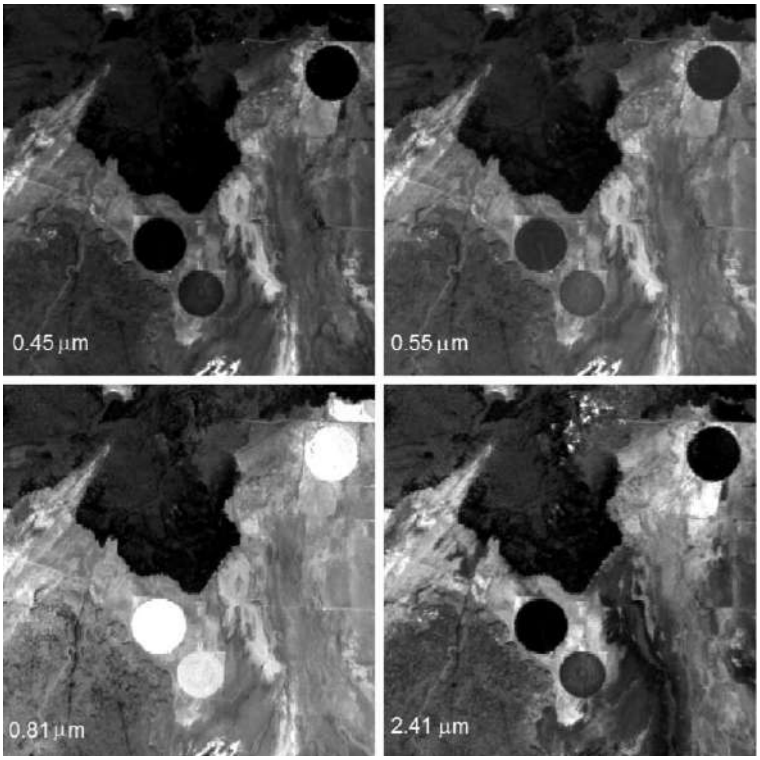

Figure 2 Portion of a scene from the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) from Utah that includes a basalt flow, circularly irrigated grasses sparsely vegetated plains, and damp alluvial sediments. AVIRIS channels centered at 0.45, 0.55, 0.81, and 2.41 μm are shown. Green vegetation is brightest in the 0.81 μm band which samples the peak of the near infrared plateau (see Figure 7) just beyond the 'red edge'. The green vegetation and damp sediments are darkest in the 2.41 μm band where water has a very low reflectance (note also Figure 8). The bright patches in the basalt flow in the 2.41 μm image are occurrences of oxidized red cinders which have a high reflectance in the SWIR.

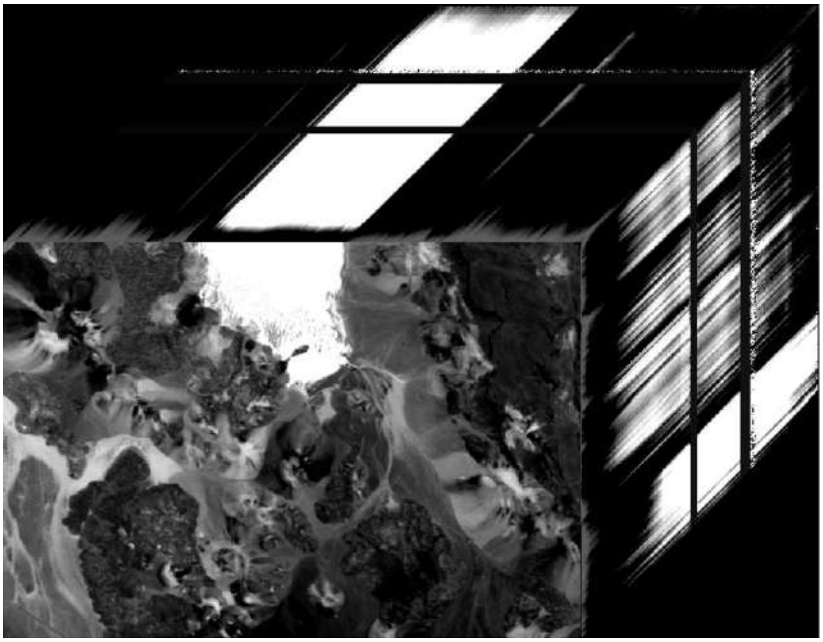

Figure 3 Visualization of a hyperspectral image cube. A 1.7 μm image band from the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the Lunar Crater Volcanic Field, Nevada is shown with each of the 224 rightmost columns and 224 topmost lines arrayed behind it. In the lower right-hand corner, the low reflectance of ferric oxide-rich cinders in the blue and green show up as dark in the stacking of columns with the high reflectance in the infrared showing up as bright.

Advances in hyperspectral sensor technology have been accompanied by advances in associated technologies, such as computing power and the evolution of data processing algorithms.

Data Processing Approaches

An early impediment to the widespread use of hyperspectral data was the sheer volume of the data sets. For instance, a single scene from NASA's Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) is 614 samples by 512 lines by 224 bands. With its short integer storage format, such a scene takes up 140.8 Mbytes. However, since the dawn of the hyperspectral era, in the 1980s, computing power has expanded immensely and computing tasks, which once seemed prohibitive, are now relatively effortless. Also, numerous processing algorithms have been developed which are especially well suited for use with hyperspectral data.

Spectral Matching

In one class of algorithms, the spectrum associated with each spatial sample is compared against one or a group of model spectra and some similarity metric is applied. The model spectrum can be measured by a laboratory or field portable spectrometer of pure materials or can be a single pixel spectrum, or average of pixel spectra, covering a known occurrence of the material(s) to be mapped. A widely used similarity metric is the Spectral Angle Mapper (SAM). SAM is the angle, θ, obtained from the dot product of the image pixel spectrum, t, and the library spectrum, r, as expressed by

$$\theta = \cos^{-1}\left(\frac{t * r}{\|t\| * \|r\|}\right) \qquad [1]$$

SAM is insensitive to brightness variations and determines similarity based solely on spectral shape. Lower θ values mean more similar spectra.

In a related approach, matching is done, not on the entire spectrum, but rather on characteristic absorption features of the materials of interest. In Figure 4, the geometry of an absorption band is plotted. To map a given mineral, with a specific absorption feature, a high resolution laboratory spectrum is resampled to the spectral resolution of the hyperspectral sensor. The spectrum is subsampled to match the specific absorption feature (i.e., only bands from the short wavelength shoulder of the absorption to the long wavelength shoulder are included). A straight line continuum (calculated, based on a line between the two-band shoulders) is divided out from both the laboratory and the sensor spectra. Contrast between the library and the sensor spectra is mitigated through the use of an additive constant. This constant is incorporated into a set of equations which are solved through the use of standard least squares. The result of such a band-fitting algorithm is two data numbers per spatial sample of the hyperspectral scene. First the band depth (or, alternatively, band area) of the feature is calculated, and second, a goodness-of-fit parameter is calculated. These two parameters are most often combined in a band-depth times goodness-of-fit image. More sophisticated versions of the band-mapping algorithm target multiple absorption features and are, essentially, expert systems.

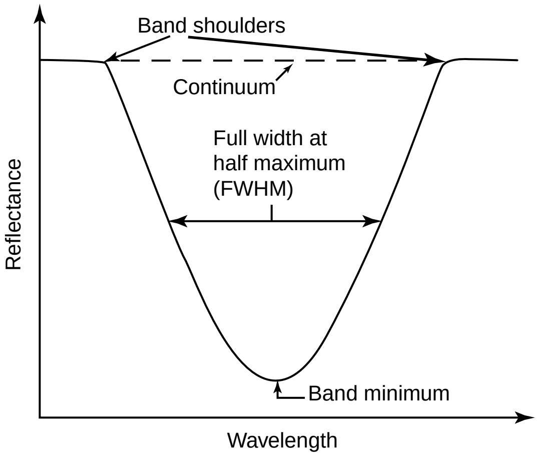

Figure 4 Parameters that describe an absorption band.

Spectral Mixture Analysis

A fundamentally different methodology for analyzing hyperspectral data sets is to model the measured spectrum of each spatial sample as a linear combination of endmember spectra. This methodology is based on the fact that the ground area corresponding to any given spatial sample likely will be covered by more than one material, with the consequence that the measured reflectance or emittance spectrum is a mixed pixel spectrum. The objective of linear SMA is to try to determine the fractional abundance of the component materials, or endmembers. The basis of linear spectral unmixing is to model the response of each spatial element as a linear combination of endmember spectra. The basis equation for linear SMA is

$$r(x, y) = \alpha \mathbf{M} + \varepsilon \qquad [2]$$

where $r(x, y)$ = the relative reflectance spectrum for the pixel at position (x, y), α = the vector of endmember abundances, $\mathbf{M}$ = the matrix of endmember spectra, and ε = the vector of residuals between the modeled and the measured reflectances. Application of SMA results in a series of fraction images for each endmember wherein the data numbers range, ideally, between 0 and 1. Fraction image pixels with a digital number (DN) of 0 are taken to be devoid of the endmember material. Fraction image pixels with a DN of 1.0 are taken to be completely covered with the endmember material.

Techniques related to SMA have been developed which also map the abundance of a target material on a per pixel basis. These include implementations highly similar to SMA such as Orthogonal Subspace Projection (OSP) wherein, as with SMA, all endmembers must be determined a priori. Other approaches such as Constrained Energy Minimization (sometimes referred to as 'Matched Filter') do not require a priori knowledge of the endmembers. Instead, only the reflectance or emittance spectrum of the target need be known and the undesired spectral background is estimated, and ultimately compensated for, using the principal eigenvectors of the sample correlation or covariance matrix of the hyperspectral scene.

Expert Systems and Artificial Neural Networks

Another means of extracting information from hyperspectral data sets is to use computer approaches that in some way mimic the human thought process. This can take the form of an expert system which applies a set of rules or tests as each pixel spectrum is successively analyzed. For example, if a spectrum has an absorption at 2.2 μm it could tentatively be classified as being caused by a clay mineral. Additionally, if the absorption is a doublet, the specific clay mineral is likely a kaolinite. In practice, expert system codes are lengthy and complex although excellent results can be demonstrated from their use, albeit at the expense of processing time. Another approach in this vein is the use of an artificial neural network (ANN) for data processing. The use of ANNs is motivated by their power in pattern recognition. ANN architectures are well suited for a parallel processing approach and thus have the potential for rapid data processing.

Hyperspectral Remote Sensing of the Land

Geology

Early efforts with airborne hyperspectral sensors focused on geologic remote sensing. This was, at least in part, a consequence of the mineralogic diversity of the Earth's surface and the fact that many minerals have absorption features which are unique and diagnostic of the mineral's identity. The expectation, which was soon borne out by results from these early sensors, was that, given exposure of the surface, mineralogic maps as detailed as the spatial resolution of the sensor could be derived from hyperspectral data. As some of the processing techniques discussed above have become more readily available, it has become possible to produce mineralogic maps from hyperspectral data far more rapidly and cost-effectively than geologic maps produced by standard means (e.g., by walking over the ground and manually noting which lithologies are underfoot).

The propensity for transition metals, most especially iron, to absorb energy through charge transfer and crystal field effects, is noted above. Fe–O and Fe^{2+}–Fe^{3+} charge transfers cause a profound absorption in the reflectance spectrum of Fe-bearing minerals shortwards of 0.4 μm. The wing of this absorption causes the low blue and green reflectance of Fe-bearing minerals. Crystal field bands cause an absorption feature in the 0.9 to 1.0 μm region. The ability to discriminate subtle differences among Fe-bearing minerals in hyperspectral data has proven extremely valuable in applications such as mineral exploration, volcanology, and in the characterization of abandoned mine lands.

In the SWIR, absorptions caused by vibrational overtones of molecular bonds within minerals. These include absorptions in the 2.2 and 2.3 μm region which are characteristic of many Al- and Mg-bearing clay minerals. Absorptions in the SWIR are narrower in width than the Fe-generated crystal field bands discussed previously. In fact, the requirement to efficiently resolve these vibrational overtone absorption bands (which have full width at half maximum (FWHM) bandwidths of 20 to 40 nm) helped to drive the selection of the nominal 10 nm bandwidth of early and most current hyperspectral sensors. Certain clay minerals can be indicators of hydrothermal activity associated with economic mineral deposits; thus, the SWIR is an important spectral region for mineral exploration. Clay minerals, by definition, have planar molecular structures that are prone to failure if subjected to shearing stresses. The ability to uniquely identify and map these minerals, using hyperspectral data, has thus been used to good effect to identify areas on volcanoes and other hydrothermally altered terrains that could be subject to landslides.

Minerals with adsorbed or molecularly bound OH^- and/or water have vibrational overtone absorptions near 1.45 and 1.9 μm, although these features are masked in remotely sensed data by prominent atmospheric water vapor bands at 1.38 and 1.88 μm. The reflectance of water and OH-bearing minerals decrease to near zero at wavelengths longwards of

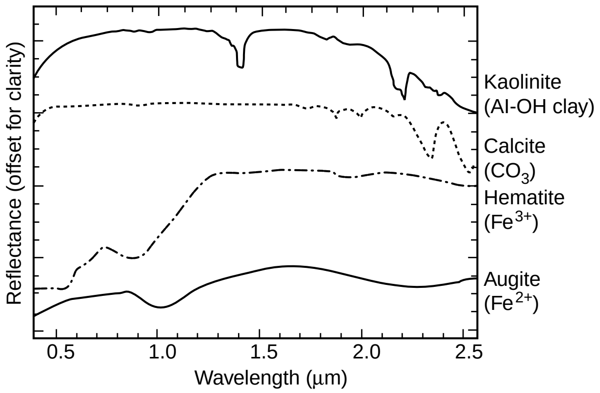

Figure 5 Reflectance spectra of representative minerals. Kaolinite represents clay minerals with its diagnostic 2.2 μm band caused by an overtone vibration of the Al–OH bond. The 2.335 μm band of calcite is caused by an overtone vibration of the carbonate molecule. The features shortwards of 1.0 μm in the hematite spectrum are caused by crystal field and charge transfer absorptions due to Fe^{3+}. The bands near 1.0 and longwards of 2.0 μm in the augite spectrum are the result of Fe^{2+} crystal field bands.

2.5 μm, due to the presence of water and OH fundamental vibrations near 3.0 μm. Minerals bearing the carbonate (CO_3^{-2}) molecule are abundant on the Earth's surface due to the ubiquitous occurrences of limestone and dolomite. Calcite and dolomite have overtone absorptions centered at 2.335 and 2.315 μm, respectively. A good test for the spectral resolution and spectral calibration of a hyperspectral sensor is its ability to successfully discriminate calcite from dolomite on the basis of the aforementioned absorption features. Figure 5 shows representative spectra of carbonate, clay, and ferric and ferrous iron-bearing minerals.

Absorptions resulting from fundamental molecular absorptions are manifested in the LWIR. These absorptions are of great geologic interest because they include absorptions resulting from vibrations of the Si–O bond and silicate minerals form the bulk of the Earth's crust. The wavelength at which the Si–O stretching feature occurs is dictated by the level of polymerization (or molecule-to-molecule bonding) of the silicate mineral. Highly polymerized framework silicate minerals, such as quartz and feldspar, have a shorter wavelength absorption than do silicate minerals, such as olivine which are composed of disconnected SiO_4 molecules. In Figure 6, laboratory emission spectra of highly and poorly polymerized silicate minerals are shown.

Vegetation and the Environment

While different minerals are, by definition, generally composed of diverse component molecules, different species of vegetation represent variations of the same general biochemical constituents (e.g., chlorophyll, proteins, lignin, cellulose, sugar, starch, etc.). In the thermal IR, vegetation has a generally flat emittance, thus it is in the reflective solar spectrum that most vegetative remote sensing is performed. The major features in the reflectance spectrum of green vegetation are shown in Figure 7. In the visible, the major absorption features are caused by the presence of chlorophyll. Chlorophyll has strong absorptions in the blue and the red, leaving a reflectance maximum (the green peak) at 0.55 μm. In the NIR, scattering in the leaf structure causes high reflectance leaving a profound absorption edge (the red edge) between

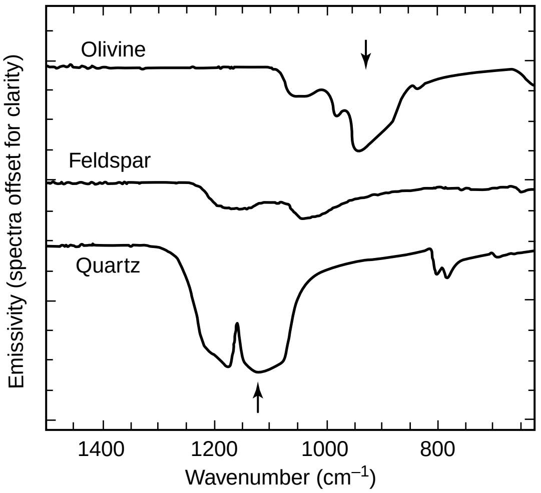

Figure 6 Emissivity spectra of SiO_4-bearing minerals showing the shift in band minimum to higher wave number with increasing polymerization.

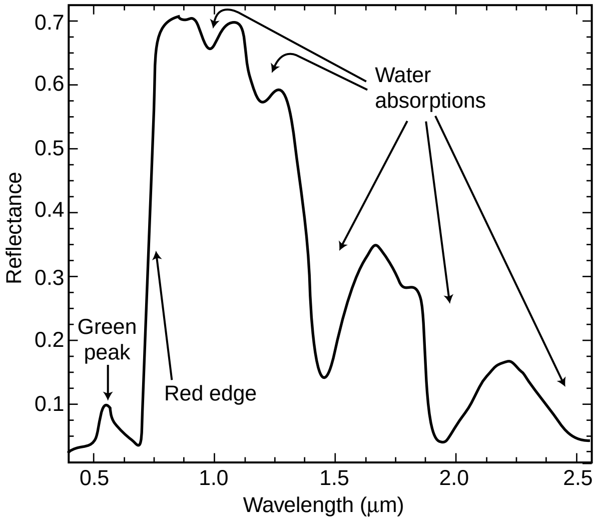

Figure 7 Major spectral features in the reflectance spectrum of green vegetation.

0.7 and 0.78 μm. Past the red edge and into the SWIR, the spectrum of green vegetation is dominated by water in the leaf, with leaf water absorptions occurring at 0.97, 1.19, 1.45, 1.93, and 2.50 μm. Studies of vegetation, using multispectral systems, have made use of a number of broadband indices. Hyperspectral systems allow for the discrimination of individual absorption features and subtle spectral shape differences in vegetation spectra which are unresolvable using broadband multispectral systems. For example, the absorptions caused by chlorophyll can be used for studies of chlorophyll content in vegetative canopies. The unique interaction of chlorophyll with leaf structures can provide a morphology to the green peak that is unique to a given species and thus is mappable using a spectral feature fitting approach.

A desired goal in the study of vegetation and ecosystem processes is to be able to monitor the chemistry of forest canopies. The foliar biochemical constituents making up forest canopies have associated absorption features in the SWIR, that result from vibrational overtones and combinations of C–O, O–H, C–H, and N–H molecular bonds. However the absorptions from these biochemical components overlap so that the chemical abundance of any one plant component cannot be directly related to any one absorption feature. An even more serious challenge is introduced by the strong influence of water in the SWIR in green vegetation spectra. Water constitutes 40 to 80% of the weight of leaves. However, the unique characteristics of high-quality hyperspectral data (e.g., excellent radiometric and spectral calibration, high signal-to-noise ratio, high spectral resolution) have been used to detect even these subtle features imprinted on the stronger water features.

The amount of water present in plant leaves is, in itself, a valuable piece of information. Hyperspectral data can be used to determine the equivalent water thickness present in vegetative canopies. Conversely, the amount of dry plant litter and/or loose wood can be estimated using spectral mixture analysis and related techniques. Taken together, all these data: vegetation species maps, determinations of leaf water content and the relative fractions of live vegetation versus litter, can be used to characterize woodlands for forest fire potential. The ability to map different species of vegetation has also proved useful in the field of agriculture, for assessing the impact of invasive weed species on farm and ranch land. Being able to assess the health of crops, in terms of leaf water content, is also important for agriculture.

Another subtle vegetative spectral feature, that provides information on plant vitality, is the aforementioned red edge. Shifts in the position of the red edge have been linked to vegetation stress. Studies have shown that vegetation stress and consequent shifts in the position of the red edge can be caused by a number of factors, including insufficient water intake or the intake of pernicious trace metals. Vegetation growing over mineral deposits can display red edge shifts and this can be used to assist in mineral exploration efforts.

Another application area, in which hyperspectral remote sensing has found great utility, is in the analysis of snow-covered areas. Over the reflective solar spectrum, the reflectance of snow varies from values near zero in the SWIR to values near one in the blue. Differences in the grain size of snow result in differences in reflectance as is illustrated in Figure 8. The spectral variability of snow makes it easily distinguishable from other Earth surface materials.

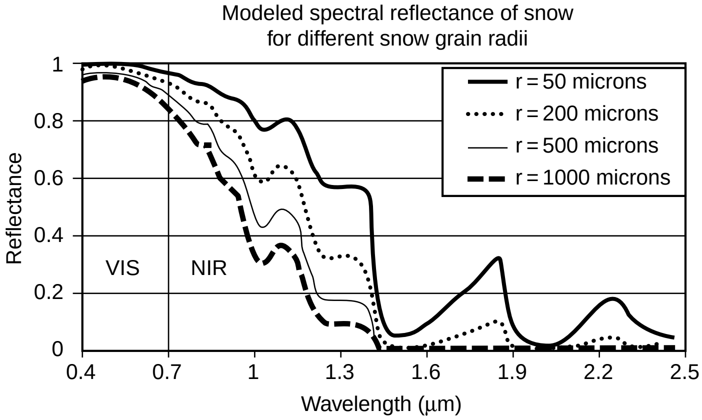

Figure 8 Differences in snow reflectance spectra resulting from differences in snow grain size. Courtesy of Dr Anne Nolin of Oregon State University.

Linear spectral unmixing, using hyperspectral data, has been shown to successfully discriminate between snow, vegetation, rock, and clouds and to accurately map snow cover fraction in mixed pixels. Accurate maps of snow cover, types of snow, and fractions of liquid water admixed with snow grains, are required for forecasting snowmelt runoff and stream discharge in watersheds dominated by snow cover.

Hyperspectral Remote Sensing of the Atmosphere

The physical quantity most often sought in land remote sensing studies is surface reflectance. However, the quantity recorded by a hyperspectral sensor is radiance at the entrance aperture of the sensor. In order to obtain reflectance, the background energy level of the Sun and/or Earth must be removed and the scattering and absorbing effects of the atmosphere must be compensated for. In the VNIR through SWIR, there are seven atmospheric constituents with significant absorption features: water vapor, carbon dioxide, ozone, nitrous oxide, carbon monoxide, methane, and oxygen. Molecular scattering (commonly called Rayleigh scattering) is strong in the blue but decreases rapidly with increasing wavelength. Above 1 μm, its effect is negligible. Scattering caused by atmospheric aerosols, or Mie scattering, also is more prominent at shorter wavelengths and decreases with increasing wavelength, but the dropoff of effects from Mie scattering is not as profound as that for Rayleigh scattering. Consequently, effects from aerosol scattering can persist into the SWIR.

There are three primary categories of methods for removing the effects of atmosphere and solar insolation from hyperspectral imagery, in order to derive surface reflectance. Methods of atmospheric corrections can be considered as being either an image-based, empirical, or model-based approach. An image-based, or 'in-scene' approach uses only data measured by the instrument. Empirical methods make use of the remotely sensed data in combination with field measurements of reflectance, to solve a simplified equation of at-sensor radiance such as eqn [3].

$$L_s = A\rho + B \qquad [3]$$

where L_s is the at-sensor radiance, ρ is the reflectance of the surface, and A and B are quantities that incorporate, respectively, all multiplicative and additive contributions to the at-sensor radiance. All the quantities expressed in eqn [3] can be considered as varying as a function of wavelength, λ. Approximations for A and B from eqn [3] constitute a set of gains and offsets derived from the empirical approach. In some empirical approaches, atmospheric path radiance is ignored and only a multiplicative correction is applied.

Model-based approaches seek to model what the at-sensor radiance should be on a pixel-by-pixel basis, including the contribution of the atmosphere. The at-sensor radiance, L_s, at any wavelength, λ, can be expressed as

$$L_s = \frac{1}{\pi}(E\rho + M_T)\tau_\theta + L_u \qquad [4]$$

where E is the irradiance at the surface of the Earth, ρ is the reflectance of the Earth's surface, M_T is the spectral radiant exitance of the surface at temperature, T, τ_θ is the transmissivity of the atmosphere at zenith angle θ, and L_u is the spectral upwelling path radiance of the atmosphere. The ability to solve eqn [4] and do atmospheric correction on a pixel-by-pixel basis is appealing, in that it negates the shortcomings of an empirical approach where a correction based on calibration targets in one part of a scene might not be appropriate for another part of the scene, due to differences in atmospheric path-length or simple atmospheric heterogeneity. Model-based approaches are also able to take advantage of the greater spectral dimensionality of hyperspectral data sets for the derivation of atmospheric properties (e.g., the amount of column water vapor, CO_2 band depths, etc.) directly from the data.

The main atmospheric component, affecting imagery in the reflective solar portion of the spectrum, is water vapor. Model-based atmospheric correction techniques determine the column water vapor for each pixel in the scene, based on the depth of the 940 and/or 1130 nm atmospheric water bands. Thus for each pixel, an appropriate amount of water can be removed. The maps of atmospheric water vapor distribution are themselves of interest to atmospheric scientists. Absorption features caused by well-mixed gases, such as the 760 nm O_2 band, can be used by model-based programs, to produce images of approximate scene topography. More advanced model-based approaches also solve for the scattering effects of aerosols on the hyperspectral data.

As noted above, atmospheric constituents, including industrial effluents, can be detected and spatially mapped by hyperspectral sensors. The fundamental vibrational absorption of gases of interest occurs in the MWIR and the LWIR. Overtones of these gaseous molecular vibrations occur in the VNIR to SWIR, but are generally too weak to be detected. The ability to detect anthropomorphically produced gases depends on a number of factors, including the temperature difference between the gas cloud and the background atmosphere, the concentration of gas within the cloud, and the size of the cloud (e.g., the path length of light through the cloud).

See also

Imaging: Infrared Imaging. **Instrumentation:** Spectrometers.

Further Reading

Clark RN (1999) Spectroscopy of rocks and minerals, and principles of spectroscopy. In: Rencz AN (ed.) *Remote Sensing for the Earth Sciences*, pp. 3–57. New York: John Wiley and Sons.

Farrand WH and Harsanyi JC (1997) Mapping the distribution of mine tailings in the Coeur d'Alene River Valley, Idaho through the use of a Constrained Energy Minimization technique. *Remote Sensing Environment* 59: 64–76.

Gao B-C and Goetz AFH (1995) Retrieval of equivalent water thickness and information related to biochemical components of vegetation canopies from AVIRIS data. *Remote Sensing Environment* 52: 155–162.

Gao B-C, Heidebrecht KH and Goetz AFH (1993) Derivation of scaled surface reflectances from AVIRIS data. *Remote Sensing Environment* 44: 165–178.

Goetz AFH (1989) Spectral remote sensing in geology. In: Asrar G (ed.) *Theory and Applications of Optical Remote Sensing*, pp. 491–526. New York: John Wiley & Sons.

Goetz AFH, Vane G, Solomon JE and Rock BN (1985) Imaging spectrometry for Earth remote sensing. *Science* 228: 1147–1153.

Hapke B (1993) *Theory of Reflectance and Emittance Spectroscopy*. Cambridge, UK: Cambridge University Press.

Kirkland LE, Herr KC and Salisbury JW (2001) Thermal infrared spectral band detection limits for unidentified surface materials. *Applied Optics* 40: 4852–4862.

Martin ME, Newman SD, Aber JD and Congalton RG (1998) Determining forest species composition using high spectral resolution remote sensing data. *Remote Sensing Environment* 65: 249–254.

Painter TH, Roberts DA, Green RO and Dozier J (1998) The effect of grain size on spectral mixture analysis of snow-covered area with AVIRIS data. *Remote Sensing Environment* 65: 320–332.

Pieters CM and Englert PAJ (eds) *Remote Geochemical Analysis: Elemental and Mineralogical Composition*. Cambridge, UK: Cambridge University Press.

Pinzon JE, Ustin SL, Castaneda CM and Smith MO (1997) Investigation of leaf biochemistry by hierarchical foreground/background analysis. *IEEE Transactions on Geoscience and Remote Sensing* 36: 1–15.

Roberts DA, Green RO and Adams JB (1997) Temporal and spatial patterns in vegetation and atmospheric properties using AVIRIS. *Remote Sensing Environment* 62: 223–240.

Wessman CA, Aber JD and Peterson DL (1989) An evaluation of imaging spectrometry for estimating forest canopy chemistry. *International Journal of Remote Sensing* 10: 1293–1316.

Laser Detection of Atmospheric Gases

E V Browell, W B Grant and S Ismail, NASA Langley Research Center, Hampton, VA, USA

Published by Elsevier Ltd.

Introduction

Lidar (light detection and ranging) systems are able to measure profiles of atmospheric aerosols, clouds, and gases by transmitting a pulsed laser beam into the atmosphere and collecting the backscattered radiation from aerosols and molecules in the atmosphere with a receiver located near the transmitter. The differential absorption lidar (DIAL) approach is the most widely used technique for measuring a variety of gases. In this approach, two closely spaced laser wavelengths are used: one which is absorbed by the gas of interest, and the other which is only weakly, or not at all, absorbed by the gas. A differential with respect to range and wavelength is calculated, to determine the average gas concentration along any segment of the lidar path, using the familiar Beer–Lambert law for an absorbing medium.

The DIAL equation can be expressed in its simple form as

$$N = \Big(1\big/\big(2(R_2 - R_1)(\sigma_{\mathrm{on}} - \sigma_{\mathrm{off}})\big)\Big) \times \ln\big(P_{\mathrm{off}}(R_2)P_{\mathrm{on}}(R_1)\big)\big/\big(P_{\mathrm{off}}(R_1)P_{\mathrm{on}}(R_2)\big) \qquad [1]$$

where N is the average gas concentration, R is the range, σ_{on} and σ_{off} are the absorption cross sections at the on- and off-line wavelengths, and $P_{\mathrm{on}}(R)$ and $P_{\mathrm{off}}(R)$ are the powers received at the on- and off-line wavelengths. However, since it is performed in the atmosphere and not a laboratory, great care must be taken to ensure that the data are analyzed properly to minimize random and systematic errors. Random errors arise from noise in the backscattered signal, the solar background signal, and the inherent detector noise, and this type of error can be reduced by signal averaging. Systematic errors arise from uncompensated instrument and atmospheric effects and must be carefully considered in system design and operation and data processing.

Another approach, the Raman lidar approach, uses inelastic scattering (scattered/emitted light has different wavelength than the illuminating light) from gases where the wavelength shift corresponds to vibrational or rotational energy levels of the molecules. Any illuminating laser wavelength can be used, but since the Raman scattering cross-section varies as λ^{-4}, where λ is the wavelength, the shorter laser wavelengths, such as in the near UV spectral region, are preferred. While even shorter wavelengths in the solar blind region below ~300 nm would permit operation in daytime, the atmospheric attenuation, due to Rayleigh scattering and UV molecular absorption, limits the measurement range. High-power lasers are used due to the low Raman scattering cross-sections. The Raman lidar measurements have to be carefully calibrated against a profile measured using another approach, such as is done for water vapor (H_2O) measurements with the launch of hygrometers on radiosondes, since the lidar system constants and atmospheric extinctions are not usually well known or modeled. In order to obtain mixing ratios of H_2O with respect to atmospheric density, the H_2O signals are ratioed to the nitrogen Raman signals. However, care must be exercised in processing these data due to the spectral dependences of atmospheric extinction.

Surface-Based Lidar Systems

Surface-based lidar systems measure the temporal evolution of atmospheric profiles of aerosols and gases. The first lidar systems used to remotely measure atmospheric gases were Raman lidar systems. The Raman lidar systems were thought to be very promising since one high-power laser could be used to measure a variety of gases. They are easier to develop and operate than DIAL systems, because the laser does not have to be tuned to a particular gas absorption feature. However, the Raman scattering cross-section is low, and the weak, inelastically scattered signal is easily contaminated by daylight background radiation, resulting in greatly reduced performance during daytime. Raman lidar systems have been developed primarily for measuring H_2O and retrieving atmospheric temperature.

Surface-based UV DIAL systems, that are part of the Network for the Detection of Stratospheric Change (NDSC) have made important contributions to the understanding of stratospheric O_3. There are DIAL systems at many locations around the Earth, with sites in Ny-Ålesund, Spitzbergen (78.9°N, 11.9°E), Observatoire Haute Provence, France (43.9°N, 5.7°E), Table Mountain, California (34.4°N, 118.2°W), Mauna Loa, Hawaii (19.5°N, 155.6°W), and Lauder, New Zealand (45.0°S, 169.7°E). First established in the 1980s, these DIAL systems are strategically located so that O_3 in different latitude bands can be monitored for signs of change. In addition, they can provide some profiles for comparison with space-based O_3 measuring instruments. The parameters of a typical ground-based UV DIAL system used in the NDSC are given in Table 1.

Airborne DIAL Systems

Airborne lidar systems expand the range of atmospheric studies beyond those possible by surface-based lidar systems, by virtue of being located in aircraft that can be flown to high altitudes and to remote locations. Thus, they permit measurements at locations inaccessible to surface-based lidar systems. In addition, they can make measurements of large atmospheric regions in times that are short compared with atmospheric motions, so that the large-scale patterns are discernible. Another advantage of airborne lidar operation is that lidar systems perform well in the nadir (down) direction since the atmospheric density and aerosol loading generally increases with decreasing altitude towards the surface, which helps to compensate for the R^{-2} decrease in the lidar

Table 1 Parameters for the Jet Propulsion Laboratory's Mauna Loa DIAL systems for stratospheric O_3 measurements

L	XeCl	Nd:YAG (3rd Harmon.)
Wavelength (nm)	308 (on)	355 (off)
Pulse energy (mJ)	300	150
Pulse repetition frequency (Hz)	200	100
R		
Area (m^2)	0.79	
Optical efficiency (%)	~40	
Wavelengths–Rayleigh (nm)	308	355
N_2 Raman (nm) (for aerosol correction)	332	387
[illegible]		
Measurement range* (km)	15–55	
Vertical resolution (km)	3 at bottom, 1 at O_3 peak, 8–10 at top	
Measurement averaging time (hours)	1.5	
Measurement accuracy	<5% at peak, 10–15% at 15 km, >40% at 45 km	

*Chopper added to block beam until it reaches 15 km in order to avoid near field signal effects.

signal with range. For the zenith direction, the advantage is that the airborne lidar system is closer to the region being measured.

Field Measurement Programs

Global O_3 Measurements

The first airborne DIAL system, which was developed by the NASA Langley Research Center (LaRC), was flown for O_3 and aerosol investigations in conjunction with the Environmental Protection Agency's Persistent Elevated Pollution Episodes (PEPE) field experiment, conducted over the east coast of the US in the summer of 1980. This initial system has evolved into the advanced UV DIAL system that will be described in the next section. Airborne O_3 DIAL systems have also been developed and used in field measurement programs by several other groups in the United States, Germany, and France.

The parameters of the current NASA LaRC airborne UV DIAL system are given in Table 2. The on-line and off-line UV wavelengths of 288.2 and 299.6 nm are used for DIAL O_3 measurements during tropospheric missions, and 301 and 310 nm are used for stratospheric missions. This system also transmits 1,064 and 600 nm beams for aerosol and cloud measurements. The time delay between the on- and off-wavelength pulses is 400 μs, which is sufficient time for the return at the first set of wavelengths to end, but short enough that the same region of the atmosphere is sampled. This system has a demonstrated absolute accuracy for O_3 measurements of better than 10% or 2 ppbv (parts per billion by volume), whichever is larger, and a measurement precision of 5% or 1 ppbv with a vertical resolution of 300 m and an averaging time of 5 minutes (about 70 km horizontal resolution at typical DC-8 ground speeds).

The NASA LaRC airborne UV DIAL systems have made significant contributions to the understanding of both tropospheric and stratospheric O_3, aerosols, and clouds. These systems have been used in 18 international and 3 national field experiments over the past 24 years, and during these field experiments, measurements were made over, or near, all of the oceans and continents of the world. A few examples of the scientific contributions made by these airborne UV DIAL systems are given in Table 3.

The NASA LaRC airborne UV DIAL system has been used extensively in NASA's Global Tropospheric Experiment (GTE) program which was started in the early 1980s, had as its primary mission the study of tropospheric chemistry in remote regions of the Earth, in part to study and gain a better understanding of atmospheric chemistry in the unperturbed atmosphere. A related goal was to document the Earth's atmosphere in a number of places during seasons when anthropogenic influences are largely absent, then return to these areas later to document the changes. The field missions have included campaigns in Africa, Alaska, the Amazon Basin,

Table 2 Parameters of the NASA LaRC airborne UV DIAL system

L[illegible]: Nd:YAG-pumped dye lasers, frequency doubled into the UV			
Pulse repetition frequency (Hz)	30		
Pulse length (ns)	8–12		
Pulse energy (mJ) at 1.06 μm	250–300		
Pulse energy (mJ) at 600 nm	50–70		
UV pulse energy (mJ)			
For troposphere at 288/300 nm	20		
For stratosphere at 301/310 nm	20		
Dimensions (l × w × h) (cm)	594 × 102 × 109		
Mass (kg)	1735		
Power requirement (kW)	30		
		Wavelength region (nm)	
R[illegible]	289–311	572–622	1064
Area (m^2)	0.086	0.086	0.864
Receiver optical efficiency (%)	30	40	30
Detector quantum efficiency (%)	26 (PMT)	8 (PMT)	40 (APD)
Field-of-view (mrad)	≤1.5	≤1.5	≤1.5
[illegible]			
Measurement range (km)	up to 10–15 (nadir and zenith)		
Vertical resolution (m)	300–1500, depending on range		
Horizontal resolution (km)	≤70		
Measurement accuracy	≤10% or 2 ppbv, whichever is greater		

Table 3 Examples of significant contributions of airborne O_3 DIAL systems to the understanding of tropospheric and stratospheric O_3 (see also **Figures 1–5**)

[illegible]:
- Air mass characterizations in a number of remote regions
- Continental pollution plume characterizations (see **Figures 1** and **2**)
- Study of biomass burn plumes and the effect of biomass burning on tropospheric O_3 production (see **Figure 3**)
- Case study of warm conveyor belt transport from the tropics to the Arctic
- Study of stratospheric intrusions and tropopause fold events (see **Figure 4**)
- Observation of the decay of a cutoff low
- Power plant plume studies

[illegible]:
- Contributions to the chemical explanation of behavior of Antarctic O_3
- Quantification of O_3 depletion in the Arctic (see **Figure 5**)
- Polar stratospheric clouds – particle characterizations
- Intercomparison with surface-based, airborne and space-based instruments
- Quantification of O_3 reduction in tropical stratosphere after the June 1991 eruption of Mount Pinatubo and characterization of the tropical stratospheric reservoir edge
- Cross-vortex boundary transport

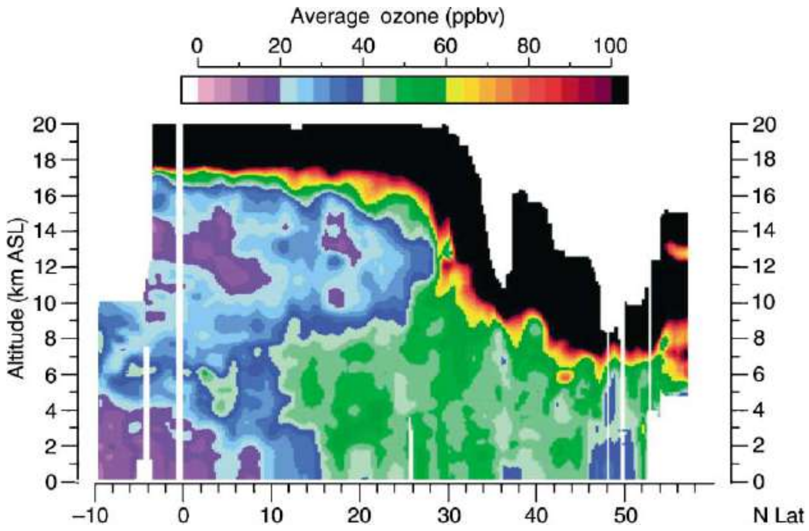

Figure 1 Latitudinal distribution of ozone over the western Pacific Ocean obtained during the second Pacific Exploratory Mission (PEM West B) in 1994.

Canada, and North America, and over the Pacific Ocean from Antarctica to Alaska, with concentrated measurements off the east coast of Asia and in the tropics. One of the bigger surprises of the two decades of field missions, was the discovery in the mid-1990s, that the tropical and South Pacific Ocean had very high tropospheric O_3 concentrations in plumes from biomass burning in Africa and South America during the austral spring. There were few indications from surface-based or space-based measurements that there were extensive biomass burn plumes in the area, primarily since the plumes were largely devoid of aerosols due to being stripped out during cloud convective lofting. Once over the ocean, horizontal transport appears to proceed relatively unimpeded unless a storm system is encountered.

The NASA LaRC airborne UV DIAL system has also been flown in all the major stratospheric O_3 campaigns starting in 1987 with the Airborne Antarctic Ozone Experiment (AAOE) to determine the cause of the Antarctic ozone hole. The UV DIAL system documented the O_3 loss across the ozone hole region. Later, when attention was turned to the Arctic and the possibility of an ozone hole there, the system was used to produce an estimate of O_3 loss during the winter season as well as to better characterize the polar stratospheric cloud (PSC) particles. The UV DIAL system was also used to study O_3 loss in the tropical stratospheric reservoir following the eruption of Mount Pinatubo in June 1991. The loss was spotted by the Microwave Limb Sounder (MLS) on the Upper Atmospheric Research Satellite (UARS), but the MLS

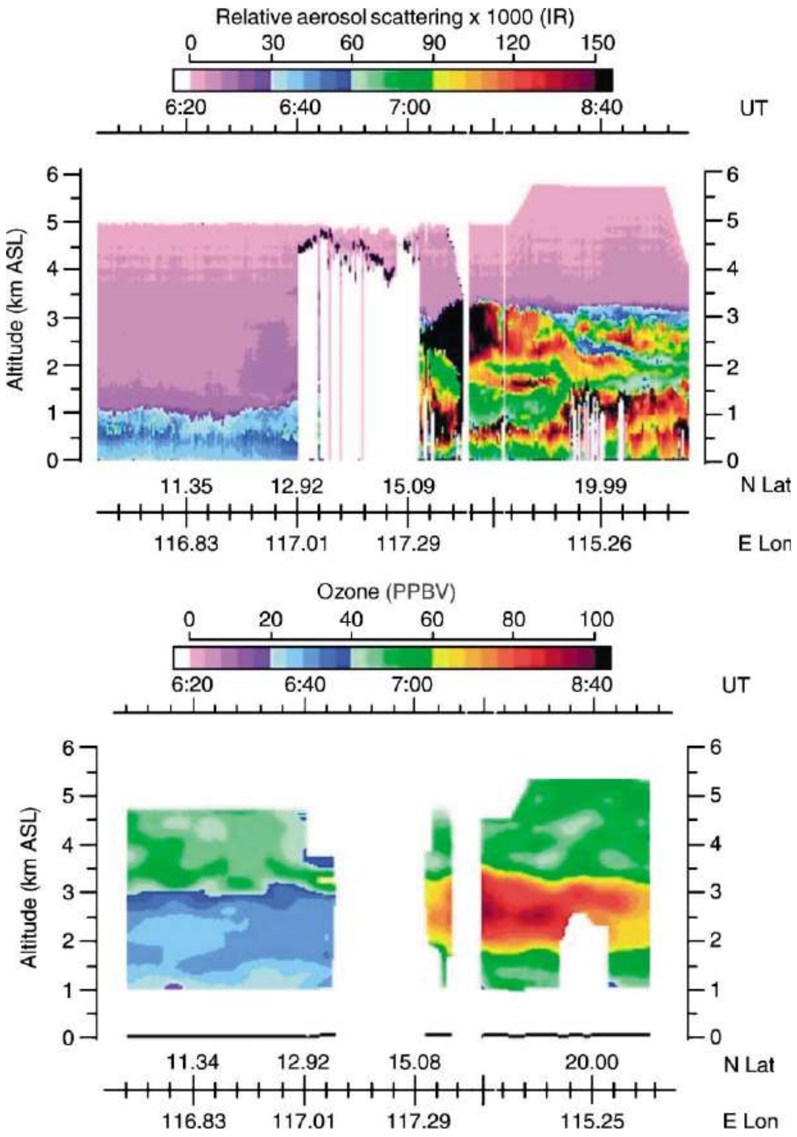

Figure 2 Pollution outflow from China over the South China Sea (right side of figure) with clean tropical air on south side of a front (left side of figure), observed during PEM West B in 1994.

was unable to study the loss in detail due to its low vertical resolution (5 km) compared to the small-scale (2–3 km) features of the O_3 loss. Other traditional space-based O_3 measuring instruments also had difficulty during this period, due to the high aerosol loading in the stratosphere following the eruption.

Space-Based O_3 DIAL System

In order to obtain nearly continuous, global distributions of O_3 in the troposphere, a space-based O_3 DIAL system is needed. A number of key issues could be addressed by a space-based O_3 DIAL system including: the global distribution of photochemical O_3 production/destruction and transport in the troposphere; location of the tropopause; and stratospheric O_3 depletion and dynamics. High-resolution airborne O_3 DIAL and other aircraft measurements show that to study tropospheric processes associated with biomass burning, transport of anthropogenic pollutants, tropospheric O_3 chemistry and dynamics, and stratosphere–troposphere exchange, a vertical profiling capability from space with a resolution of 2–3 km is needed, and this capability cannot currently be achieved using passive remote sensing satellite instruments. An example of the type of latitudinal O_3 cross-section that could be provided by a space-based O_3 DIAL system is shown in Figure 1. This figure shows many different aspects of O_3 loss and production; vertical and horizontal transport; and stratosphere–troposphere exchange that occurs from the tropics to high latitudes.

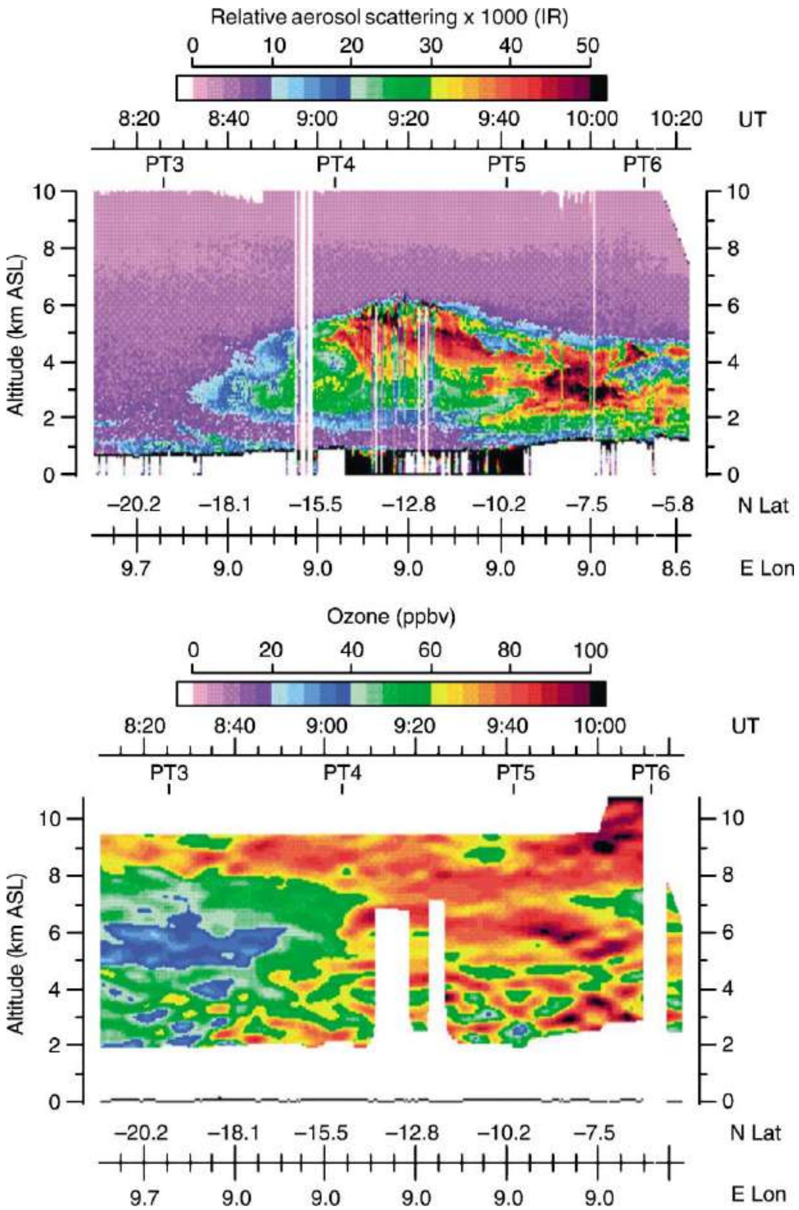

Figure 3 Biomass burning plume over the Atlantic Ocean arising from biomass burning in the central part of western Africa, observed during the TRACE-A mission in 1992.

This type of data would be available from just one pass from a space-based O_3 DIAL system. A space-based O_3 DIAL system optimized for tropospheric O_3 measurements (see system description in Table 4a,b) would also permit high-resolution O_3 measurements in the stratosphere (1 km vertical, 100 km horizontal), along with high-resolution aerosol measurements (100 m vertical, 10 km horizontal). In addition, these DIAL measurements will be useful in assisting in the interpretation of passive remote sensing measurements and in helping to improve their data processing algorithms.

Global H_2O Measurements

H_2O and O_3 are important to the formation of OH in the troposphere, and OH is at the center of most of the chemical reactions in the lower atmosphere. In addition H_2O is an excellent tracer of vertical and horizontal transport of air masses in the troposphere, and it can be used as a tracer of stratosphere–troposphere exchange. Increased aerosol sizes, due to high relative humidities, can also affect heterogeneous chemical processes and radiation budgets in the boundary layer and in cloud layers. Knowledge of H_2O is important to weather forecasting and climate

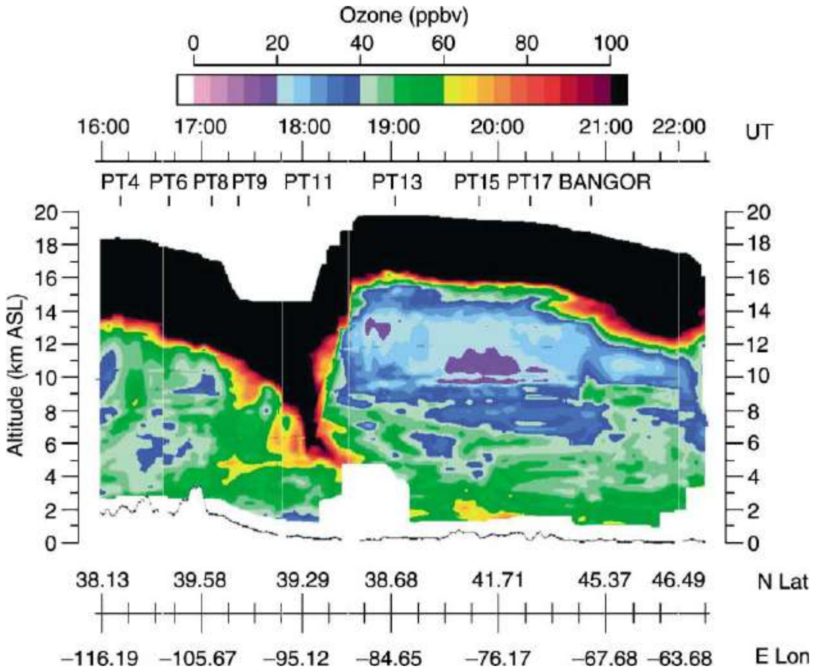

Figure 4 Ozone distribution observed on flight across US during the SASS (Subsonic Assessment) Ozone and Nitrogen Experiment (SONEX) in 1997. A stratospheric intrusion is clearly evident on left side of figure, and low ozone air from tropics transported to mid-latitudes can be seen in the upper troposphere on the right.

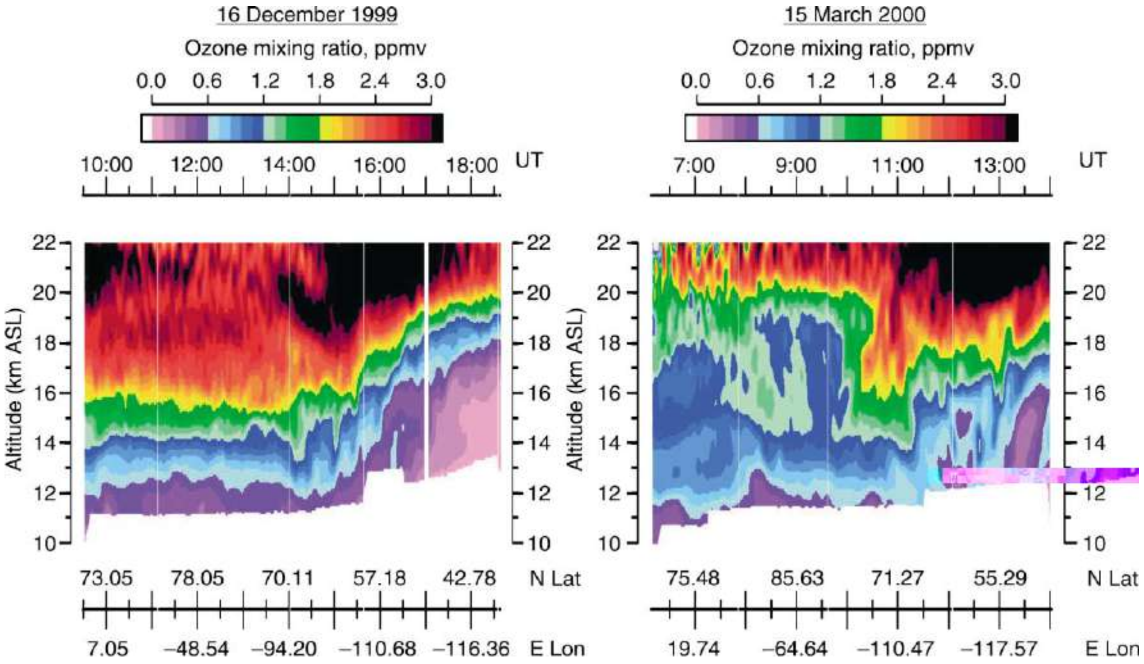

Figure 5 Ozone cross-sections in the stratosphere measured in the winter of 1999/2000 during the SOLVE mission. The change in ozone number density in the Arctic polar vortex due to chemical loss during the winter is clearly evident at latitudes north of 72°N.

predictions. Thus, H_2O distributions can be used in several different ways to better understand chemical, transport, radiation and meteorological processes in the global troposphere.

H_2O Raman Lidar Systems

The first Raman lidar measurements of H_2O were made in the late 1960s, but not much progress in using the Raman approach for H_2O was made until a system using an abandoned searchlight mirror was employed by the NASA Goddard Space Flight Center in the mid-1980s, to show that if the signal collector (telescope) was large enough, useful Raman measurements could be made to distances of several kilometers. Until recently, Raman lidar measurements of H_2O were largely limited to night-time, due to the high sunlight background interference. Measurements are now made during daytime, using very narrow telescope fields of view and very narrowband filters in the receiver. A good example of a Raman lidar system used to measure H_2O is that at the Department of Energy's Atmospheric Radiation Measurement – Cloud and Radiation Test Bed (ARM-CART) site in Oklahoma. The site is equipped with a variety of instruments aimed at studying the radiation properties of the atmosphere. The system parameters are given in Table 4a,b.

Raman lidar systems have been used for important measurements of H_2O from a number of ground-based locations. Some of the important early work dealt with the passage of cold and warm fronts. The arrival of the wedge-shaped cold front, pushing the warm air up, was one of these studies. Raman lidar have also been located at the ARM-CART site in Kansas and Oklahoma, where they could both provide a climatology of H_2O as well as provide correlative measurements of H_2O for validation of space-based instruments.

Table 4a Parameters for the US Department of Energy's surface-based Raman lidar system for measurements of H_2O

L	Nd:YAG Laser
Wavelength (nm)	355
Pulse energy (mJ)	400
Pulse repetition frequency (Hz)	30
R	
Area (m^2)	1.1
Wavelengths (nm)	
Water vapor	407
Nitrogen	387
[illegible]	
Measurement range (km)	
Night-time	near surface to 12
Daytime	near surface to 3
Range resolution (m)	39 at low altitudes, 300 >9 km
Measurement accuracy (%)	
Night-time	<10 <7 km (1 min avg) 10–30 at high altitudes (10–30 min avg)
Daytime	10 <1 km; 5–15 for 1–3 km (10 min avg)

Table 4b Parameters for the NASA Goddard Space Flight Center's Scanning Raman lidar system for measurements of H_2O

	D/N	*N*
L	Nd:YAG Laser	XeF
Wavelength (nm)	355	351
Pulse energy (mJ)	300	30–60
Pulse repetition frequency (Hz)	30	400
R		
Area (m^2)	1.8	1.8
Wavelengths (nm)		
Water vapor	407	403
Nitrogen	387	382
[illegible]		
Measurement range (km)		
Night-time	near surface to 12	
Daytime	near surface to 4	
Range resolution (m)	7.5 at low altitudes, 300 >9 km	
Measurement accuracy (%)		
Night-time	<10 <5 km (10 sec and 7.5 m range resolution) <10 <7 km (1 min avg) 10–30 at high altitudes (10–30 min avg)	
Daytime	<10 <4 km (5 min avg)	

H_2O DIAL Systems

H_2O was first measured with the DIAL approach using a temperature-tuned ruby laser lidar system in the mid-1960s. The first aircraft-based H_2O DIAL system was developed at NASA LaRC, and was flown in 1982, as an initial step towards the development of a space-based H_2O DIAL system. This system was based on Nd:YAG-pumped dye laser technology, and it was used in the first airborne H_2O DIAL atmospheric investigation, which was a study of the marine boundary layer over the Gulf Stream. This laser was later replaced with a flashlamp-pumped solid-state alexandrite laser, which had high spectral purity, i.e., little out-of-band radiation, a requirement since water vapor lines are narrow, and this system was used to make accurate H_2O profile measurements across the lower troposphere.

A third H_2O DIAL system, called LASE (Lidar Atmospheric Sensing Experiment) was developed as a prototype for a space-based H_2O DIAL system, and it was completed in 1995. This was the first fully autonomously operating DIAL system. LASE uses a Ti:sapphire laser that is pumped by a double-pulsed, frequency-doubled Nd:YAG to produce laser pulses in the 815 nm absorption band of H_2O (see Table 5). The wavelength of the Ti:sapphire laser is controlled by injection seeding with a diode laser that is frequency locked to a H_2O line using an

Table 5 Parameters of LASE H_2O DIAL system

Laser	Ti:sapphire
Wavelength (nm)	813–818
Pulse energy (mJ)	100
Pulse-pair repetition frequency	5 (on- and off-line pulses separated by 300 μs)
Linewidth (pm)	<0.25
Stability (pm)	<0.35
Spectral purity (%)	>99
Beam divergence (mrad)	<0.6
Pulse width (ns)	35
Receiver	
Area (m^2)	0.11
Receiver optical efficiency (%)	50 (night), 35 (day)
Avalanch Photodiode (APD) detector quantum efficiency (%)	80
Field-of-view (mrad)	1.0
Noise equivalence power (W $Hz^{-0.5}$)	2×10^{-14}
Excess noise factor	3
System performance	
Measurement range (altitude) (km)	15
Range resolution (m)	300–500
Measurement accuracy (%)	5

Table 6 Examples of significant contributions of airborne H_2O DIAL systems to the understanding of H_2O distributions (see also Figures 6–8)

NASA Langley H_2O DIAL systems (including LASE)
Study of marine boundary layer over Gulf Stream
Observation of H_2O transport at a land/sea edge
Study of large-scale H_2O distributions across troposphere (see Figure 6)
Correlative in situ and remote measurements
Observations of boundary layer development (see Figure 7)
Cirrus cloud measurements
Hurricane studies (see Figure 8)
Relative humidity effects on aerosol sizes
Ice supersaturation in the upper troposphere
Studies of stratospheric intrusions
H_2O distributions over remote Pacific Ocean
Other airborne H_2O DIAL systems
Boundary layer humidity fluxes
Lower-stratospheric H_2O studies

absorption cell. Each pulse pair consists of an on-line and off-line wavelength for the H_2O DIAL measurements. To cover the large dynamic range of H_2O concentrations in the troposphere (over 3 orders of magnitude), up to three line pair combinations are needed. LASE uses a novel approach of operating from more than one position on a strongly absorbing H_2O line. In this approach, the laser is electronically tuned at the line center, side of the line, and near the wing of the line to achieve the required absorption cross-section pairs (on and off). LASE has demonstrated measurements of H_2O concentrations across the entire troposphere using this 'side-line' approach. The accuracy of LASE H_2O profile measurements was determined to be better than 6% or 0.01 g/kg, whichever is larger, over the full dynamic range of H_2O concentrations in the troposphere. LASE has participated in over eight major field experiments since 1995. See Table 6 for a listing of topics studied using airborne H_2O DIAL systems (Figures 6–8).

Space-Based H_2O DIAL System

The technology for a space-based H_2O DIAL system is rapidly maturing in the areas of: high-efficiency, high-energy, high-spectral-purity, long-life lasers with tunability in the 815- and 940-nm regions; low-weight, large-area, high-throughput, high-background-rejection receivers; and high-quantum-efficiency, low-noise, photon-counting detectors. With the expected advancements in lidar technologies leading to a 1-J/pulse capability, a space-based H_2O DIAL system could be flown on a long-duration space mission this decade.

Space-based DIAL measurements can provide a global H_2O profiling capability, which when combined with passive remote sensing with limited vertical resolution, can lead to 3-dimensional measurements of global H_2O distributions. High vertical resolution H_2O ($\leq$1 km), aerosol ($\leq$100 m), and cloud top ($\leq$50 m) measurements from the lidar along the satellite ground-track, can be combined with the horizontally contiguous data from nadir passive sounders to generate more complete high-resolution H_2O, aerosol, and cloud fields for use in the various studies indicated above. In addition, the combination of active and passive measurements can provide significant synergistic benefits leading to improved temperature and relative humidity measurements. There is also strong synergism with aerosol and cloud imaging instruments and with future passive instruments that are being planned or proposed for missions addressing atmospheric chemistry, radiation, hydrology, natural hazards, and meteorology.

Tunable Laser Systems for Point Monitoring

Tunable diode laser (TDL) systems are also used to measure atmospheric gases on a global scale from aircraft and balloons. TDLs are small lasers that emit extremely narrowband radiation and can be tuned in the near IR spectral region using a combination of temperature and current. TDLs can be built into very sensitive and compact systems and located on the surface or flown on aircraft or balloons and used to measure such species as CO, HCl, CH_4, and oxides of nitrogen such as N_2O and NO_2. They derive their

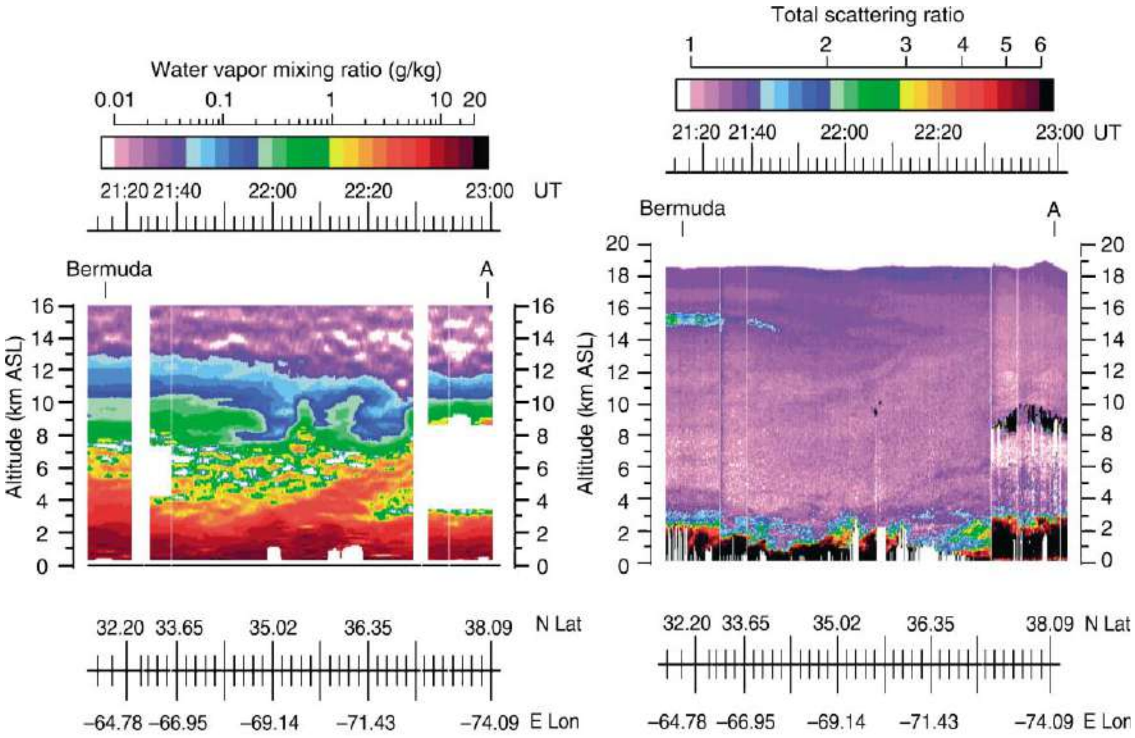

Figure 6 LASE measurements of water vapor (left) and aerosols and clouds (right) across the troposphere on an ER-2 flight from Bermuda to Wallops during the Tropospheric Aerosol Radiative Forcing Experiment (TARFOX) conducted in 1996.

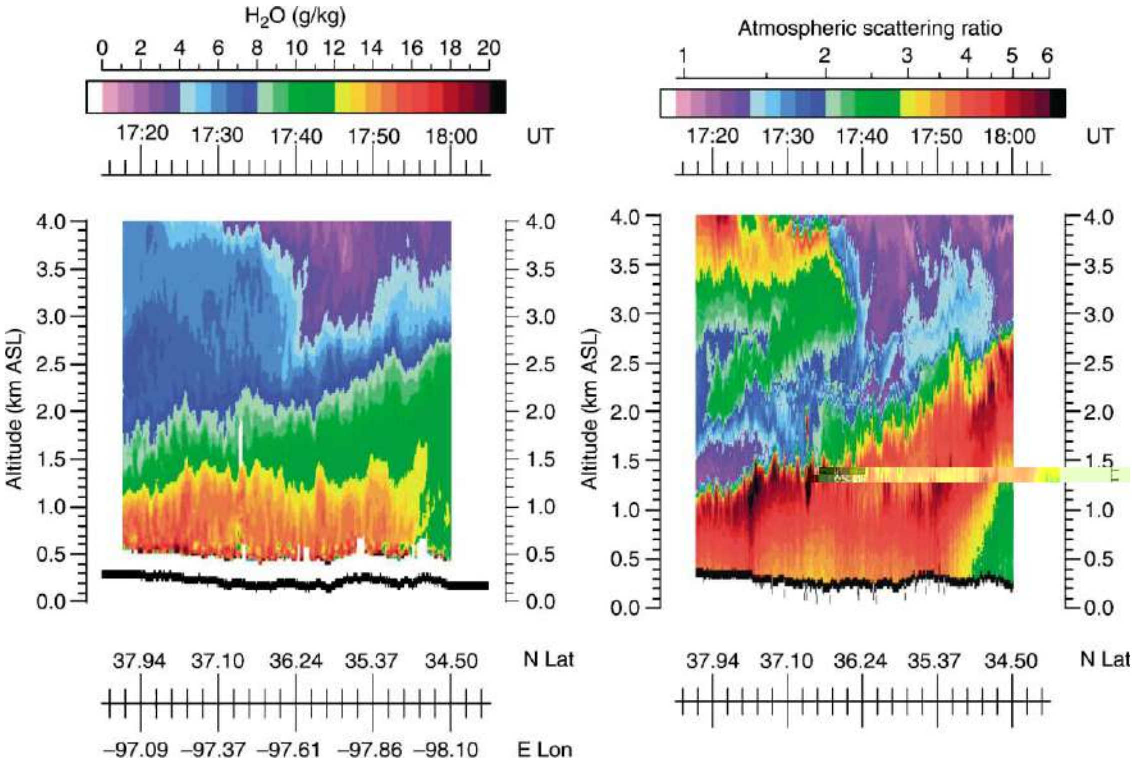

Figure 7 Water vapor and aerosol cross-section obtained on a flight across a cold front during the Southern Great Plains (SGP) field experiment conducted over Oklahoma in 1997.

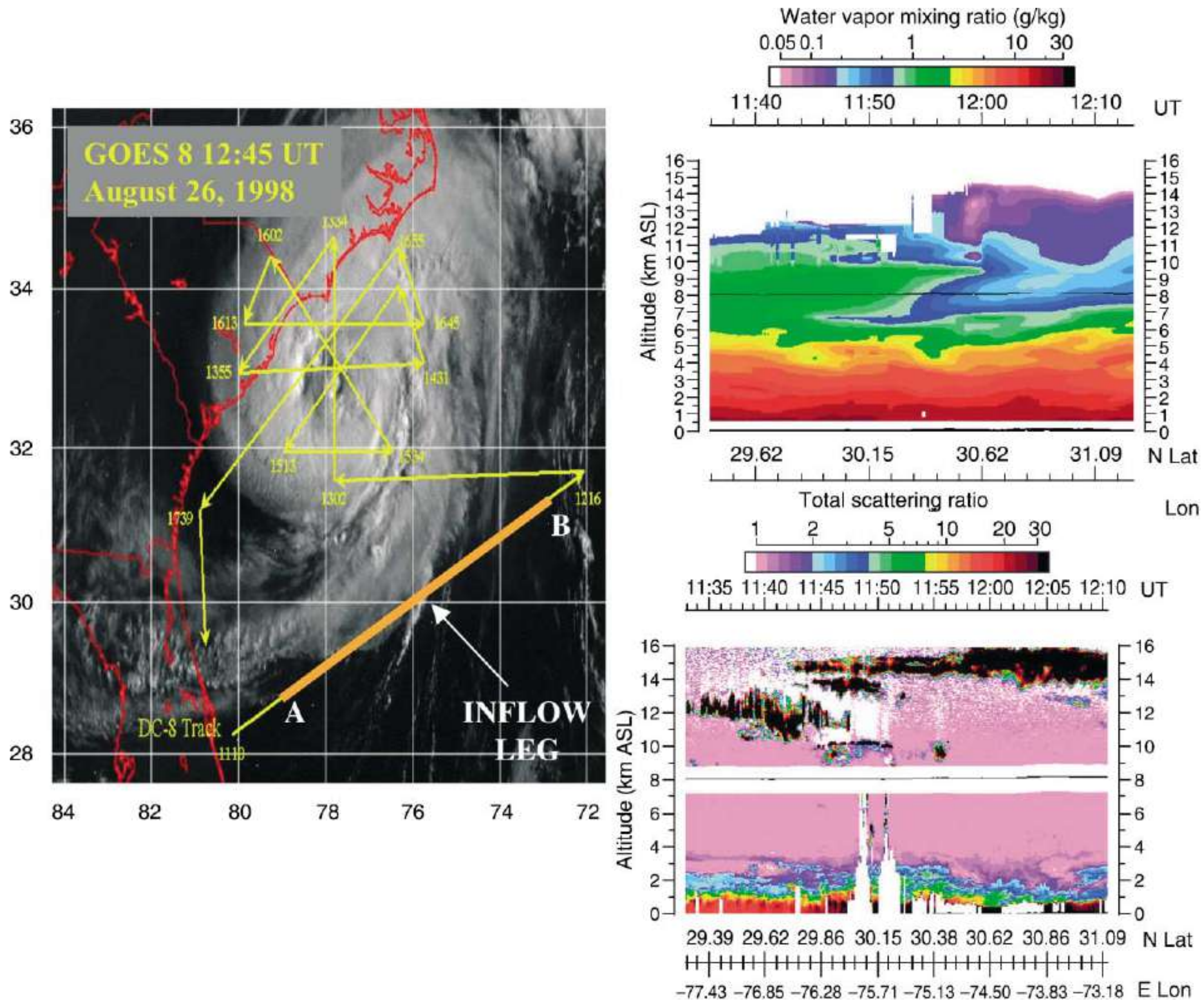

Figure 8 Measurements of water vapor, aerosols, and clouds in the inflow region of Hurricane Bonnie during 1998 Convection and Moisture Experiment (CAMEX-3). A rain band can be clearly seen at the middle of Leg-AB on the satellite and LASE cross-sections.

high sensitivity to the fact that the laser frequency is modulated at a high frequency, permitting a small spectral region to be scanned rapidly. The second- or fourth-harmonic of the scan frequency is used in the data acquisition, effectively eliminating much of the low-frequency noise due to mechanical vibrations and laser power fluctuations. In addition, multipass cells are employed, thereby generating long paths for the absorption measurements.

TDL systems have been used in a number of surface-based measurement programs. One system was mounted on a ship doing a latitudinal survey in the Atlantic Ocean and monitored NO_2, formaldehyde (HCHO), and H_2O_2. Another TDL system was located at the Mauna Loa Observatory and was used to monitor HCHO and H_2O_2 during a photochemistry experiment, finding much lower concentrations of both gases than models had predicted. The TDL system used to measure HCHO has been used in a number of additional ground-based measurement programs and has also been flown on an aircraft in a couple of tropospheric missions.

One TDL system, called DACOM (Differential Absorption CO Measurement), has been used in a large number of NASA GTE missions. It makes measurements of CO in the 4.7 μm spectral region and CH_4 in the 3.3 or 7.6 μm spectral region. This instrument is able to make measurements at a 1 Hz rate and with a precision of 0.5–2.0%, depending on the CO value, and an accuracy of ±2%. DACOM has been very useful in determining global distributions of CO due to the wide-ranging nature of the GTE missions. It has also contributed to the characterization and understanding of CO in air masses encountered during the GTE missions (Table 7).

A pair of TDL instruments, the Airborne Tunable Laser Absorption Spectrometer (ATLAS) and the Aircraft (ER-2) Laser Infrared Absorption Spectrometer (ALIAS), have been flown on several NASA missions to explore polar O_3 chemistry and atmospheric transport. The ATLAS instrument measures N_2O, and ALIAS is a 4-channel spectrometer that measures a large variety of gases, including HCl, N_2O, CH_4, NO_2, and CO and is currently configured to measure water isotopes across the tropopause.

A new class of lasers, tunable quantum-cascade (QC) lasers, are being added to the list of those available for *in situ* gas measurement systems. Using a cryogenically cooled QC laser during a series of 20 aircraft flights beginning in September 1999 and extending through March 2000, measurements were made of CH_4 and N_2O up to ~20 km in the stratosphere over North America, Scandinavia, and Russia, on the NASA ER-2. Compared with its companion lead salt diode lasers, that were also flown on these flights, the single-mode QC laser, cooled to 82 K, produced higher output power (10 mW), narrower laser linewidth (17 MHz), increased measurement precision (a factor of 3), and better spectral stability (~0.1 cm^{-1} K). The sensitivity of the QC laser channel was estimated to correspond to a minimum-detectable mixing ratio of approximately 2 ppbv of CH_4.

Laser Long-Path Measurements

Laser systems can also be used in long-path measurements of gases. The most important of such programs entailed the measurement of hydroxyl radical (OH) in the Rocky Mountains west of

Table 7 Examples of significant contributions of airborne tunable diode laser systems to the understanding of atmospheric chemistry and transport

B[illegible]
- The first [illegible] measurements of the suite NO_2, NO, O_3, and the NO_2 photolysis rate to test NO_x (NO_2 + NO) photochemistry
- The first [illegible] stratospheric measurements of NO_x over a full diurnal cycle to test N_2O_5 chemistry
- [illegible] measurements of NO_2 and HNO_3 over the 20–35 km region to assess the effect of Mt. Pinatubo aerosol on heterogeneous atmospheric chemistry
- Measurements of HNO_3 and HCl near 30 km
- Measurements of CH_4, HNO_3, and N_2O for validation of several satellite instruments
- Intrusions from midlatitude stratosphere to tropical stratospheric reservoir

A[illegible]
- *C[illegible]*
  - Measurement of CO from biomass burning from Asia, Africa, Canada, Central America, and South America
  - Detection of thin layers of CO that were transported thousands of miles in the upper troposphere
  - Observed very high levels of urban pollution in plumes off the Asian continent
  - Emission indices for many gases have been calculated with respect to CO
- *M[illegible]*
  - Determined CH_4 flux over the Arctic tundra, which led to rethinking of the significance of tundra regions as a global source of CH_4
  - Found that biomass burning is a significant source of global CH_4
  - Found strong enhancements of CH_4 associated with urban plumes

A[illegible]
- *P[illegible]*
  - Extreme denitrification observed in Antarctic winter vortex from NO_y:N_2O correlation study
  - Observed very low N_2O in Antarctic winter vortex, which helped refute the theory that the O_3 hole is caused by dynamics
  - Contributed to the study of transport out of the lower stratospheric Arctic vortex by Rossby wave breaking
  - Measurement of concentrations of gases involved in polar stratospheric O_3 destruction and production
  - Activation of chlorine in the presence of sulfate aerosols
- *M[illegible]*
  - Measurements in aircraft exhaust plumes in the lower stratosphere, especially of reactive nitrogen species and CO
  - Vertical profiles of CO in the troposphere and lower stratosphere
  - Determination of the hydrochloric acid and the chlorine budget of the lower stratosphere
  - Measurement of NO_2 for testing atmospheric photochemical models
  - Trends in HCl/Cl_y in the stratosphere <21 km, 1992–1998
  - Gas concentration measurements for comparison with a balloon-borne Fourier transform spectrometer observing the Sun
  - Near-IR TDL laser hygrometers (ER-2, WB57, DC-8) for measuring H_2O and total water in the lower stratosphere and upper troposphere
  - Remnants of Arctic winter vortex detected many months after breakup
- *[illegible]*
  - Tropical entrainment time scales inferred from stratospheric N_2O and CH_4 observations

Boulder, Colorado. OH was measured using a XeCl excimer laser operating near 308 nm and transmitting a beam to a retroreflector 10.3 km away. The measurements were quite difficult to conduct, primarily since OH abundance is very low (10^5–$10^7\,\text{cm}^{-3}$), yielding very low absorption (~0.02% for an abundance of $10^5\,\text{cm}^{-3}$) over the 20.6 km path. In addition, the excimer laser could generate OH from ambient H_2O and O_3, unless the laser energy density was kept low. To help ensure good measurements, a white light source was also employed during the measurements to monitor H_2O, O_3, and other gases. The measurements were eventually very successful and led to new values for OH abundances in the atmosphere.

Laser-Induced Fluorescence (LIF)

The laser-induced fluorescence approach has been used to measure several molecular and ionic species *in situ*. The LIF approach has been used to measure OH and HO_2 (HO_x) on the ground and on aircraft platforms. One airborne LIF system uses a diode-pumped Nd:YAG-pumped, frequency-doubled dye laser to generate the required energy near 308 nm. The laser beam is sent into a White cell where it can make 32–36 passes through the gas in the cell to increase the LIF signal strength. NO is used to convert HO_2 to OH. The detection limit in 1 minute of about 2–3 ppqv (10^{-15}) above 5 km altitude, which translates into a concentration of about 4×10^4 molec/cm^3 at 5 km and 2×10^4 molec/cm^3 at 10 km altitude. One of the interesting findings from such measurements is that HO_x concentrations are up to 5 times larger than model predictions based on NO_x concentrations, suggesting that NO_x emissions from aircraft could have a greater impact on O_3 production than originally thought.

NO is detected using the LIF technique in a two-photon approach: electrons are pumped from the ground state using 226 nm radiation and from that state to an excited state using 1.06 μm radiation. The 226 nm radiation is generated by frequency doubling a dye laser to 287 nm and then mixing that with 1.1 μm radiation derived from H_2 Raman shifting of frequency-mixed radiation from a dye laser and a Nd:YAG laser. From the excited level, 187–201 nm radiation is emitted. In order to measure NO_2, it is first converted to the photofragment NO via pumping at 353 nm from a XeF excimer laser. One of the interesting findings from airborne measurements in the South Pacific is that there appeared to be a large missing source for NO_x in the upper troposphere.

Summary

DIAL and Raman lidar systems have played important roles in studying the distribution of gases such as O_3 and H_2O on local, regional, and global scales, while TDL systems have played corresponding roles for such gases as CO, HCl, and oxides of nitrogen. It is anticipated that these approaches will continue to yield valuable information on these and other gases, as new and more capable systems are developed. Within the next decade, it is expected that a DIAL system will be placed in orbit to make truly global measurements of O_3, H_2O, and/or carbon dioxide.

See also

Imaging: Lidar. **Scattering:** Raman Scattering.

Further Reading

Browell EV (1994) Remote sensing of trace gases from satellites and aircraft. In: Calvert J (ed.) *Chemistry of the Atmosphere: The Impact on Global Change*, pp. 121–134. Cambridge, MA: Blackwell Scientific Publications.

Browell EV, Ismail S and Grant WB (1998) Differential absorption lidar (DIAL) measurements from air and space. *Applied Physics. B* 67: 399–410.

Dabas A, Loth C and Pelon J (eds) (2001) *Advances in Laser Remote Sensing*. Selected Papers presented at the 20th International Laser Radar Conference (ILRC), Vichy, France, 10–14 July 2000, pp. 357–360. Palaiseau Cedex, France: A. Edition d'Ecole Polytechnique.

Goldsmith JEM, Bisson SE, Ferrare RA, *et al.* (1994) Raman lidar profiling of atmospheric water vapor: simultaneous measurements with two collocated systems. *Bulletin American Meteorological Society* 75: 975–982.

Goldsmith JEM, Blair FH, Bisson SE and Turner DD (1998) Turn-key Raman lidar for profiling atmospheric water vapor, clouds, and aerosols. *Applied Optics* 37: 4979–4990.

Grant WB (1995) Lidar for atmospheric and hydrospheric studies. In: Duarte F (ed.) *Tunable Laser Applications*, pp. 213–305. New York: Marcel Dekker Inc.

Grant WB, Browell EV, Menzies RT, Sassen K and She C-Y (eds) (1997) *Laser Applications in Remote Sensing*. SPIE Milestones series, 690 pp, 86 papers.

Grisar R, Booettner H, Tacke M and Restelli G (1992) *Monitoring of Gaseous Pollutants by Tunable Diode Lasers*. Proceedings of the International Symposium held in Freiburg, Germany, 17–18 October 1991, 372 pp. Dordrecht, The Netherlands: Kluwer Academic Publishers.

McDermid IS, Walsh TD, Deslis A and White M (1995) Optical systems design for a stratospheric lidar system. *Applied Optics* 34: 6201–6210.

Mégie G (1988) Laser measurements of atmospheric trace constituents. In: Measures RM (ed.) *Laser Remote Chemical Analysis*, pp. 333–408. New York: J Wiley & Sons.

Podolske JR and Loewenstein M (1993) Airborne tunable diode laser spectrometer for trace gas measurements in the lower stratosphere. *Applied Optics* 32: 5324–5330.

Pougatchev NS, Sachse GW, Fuelberg HE, *et al.* (1999) Pacific Exploratory Mission-Tropics carbon monoxide measurements in historical context. *Journal of Geophysics Research* 104: 26,195–26,207.

Sachse GW, Collins Jr JE, Hill GF, *et al.* (1991) Airborne tunable diode laser sensor for high precision concentration and flux measurements of carbon dioxide and methane. *Proceedings of the SPIE International Society for Optical Engineering* 1443: 145–156.

Svanberg S (1994) Differential absorption lidar (DIAL). In: Segrist MW (ed.) *Air Monitoring Spectroscopic Techniques*, pp. 85–161. New York: J Wiley & Sons.

Webster CR, *et al.* (1994) Aircraft (ER-2) laser infrared absorption spectrometer (ALIAS) for in situ measurements of HCl, N_2O, CH_4, NO_2, and HNO_3. *Applied Optics* 33: 454–472.

Whiteman DN, Melfi SH and Ferrare RA (1992) Raman lidar system for the measurement of water vapor and aerosols in the Earth's atmosphere. *Applied Optics* 31: 3068–3082.

Zanzottera E (1990) Differential absorption lidar techniques in the determination of trace pollutants and physical parameters of the atmosphere. *Critical Reviews in Analytical Chemistry* 21: 279–319.

Optical Transmission and Scatter of the Atmosphere

S M Adler-Golden and A Berk,
Spectral Sciences, Inc., Burlington, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ability to understand and model radiative transfer (RT) processes in the atmosphere is critical for remote sensing, environmental characterization, and many other areas of scientific and practical interest. At the Earth's surface, the bulk of this radiation, which originates from the sun, is found in the ultraviolet to infrared range between around 0.3 and 4 μm. This article describes the most significant RT processes for these wavelengths, which are absorption (light attenuation along the line of sight LOS) and elastic scattering (redirection of the light). Transmittance, T, is defined as one minus the fractional extinction (absorption plus scattering). At longer wavelengths (in the mid- and long-wave infrared) the major light source is thermal emission.

A few other light sources are mentioned here. The moon is the major source of visible light at night. Forest fires can be a significant source of mid-wave infrared radiation. Manmade light sources include continuum sources such as incandescent lamps and spectrally narrow sources such as fluorescent lamps and lasers. In general, spectrally narrow sources need to have a different RT treatment than continuum sources due to the abundance of narrow spectral absorption lines in the atmosphere, as is discussed below.

The challenge of atmospheric RT modeling is essentially to solve the following equation that describes monochromatic light propagation along the LOS direction:

$$\left(\frac{1}{k}\right)\frac{\mathrm{d}I}{\mathrm{d}u} = -I + J \qquad [1]$$

where I is the LOS radiance (watts per unit area per unit wavelength per steradian), k is the extinction coefficient for the absorbing and scattering species (per unit concentration per unit length), u is the material column density (in units of concentration times length), and J is the radiance source function. I is a sum of direct (i.e., from the sun) and diffusely scattered components. The source function represents the diffuse light scattered into the LOS, and is the angular integral over all directions Ω_i of the product of the incoming radiance, $I(\Omega_i)$, and the scattering phase function, $p(\Omega_o, \Omega_i)$:

$$J(\Omega_o) = \int_{\Omega_i} p(\Omega_o, \Omega_i) I(\Omega_i)\mathrm{d}\Omega_i \qquad [2]$$

The scattering phase function describes the probability density for incoming light from direction Ω_i scattering out at angle Ω_o, and is a function of the difference (scattering) angle θ.

The direct radiance component, I', is described by eqn [1] with the source function omitted. Integrating along the LOS leads to the well-known Beer's Law equation for transmittance:

$$T = I'/I'_0 = \exp(-ku) \qquad [3]$$

where I'_0 is the direct radiance at the boundary of the LOS. The quantity $ku = \ln(1/T)$ is known as the optical depth.

Atmospheric Constituents

The atmosphere has a large number of constituents, including numerous gaseous species and suspended liquid and solid particulates. Their contributions to extinction are depicted in Figure 1. The largest category is gases, of which the most important are water vapor, carbon dioxide, and ozone. Of these, water vapor is the most variable and carbon dioxide the least, although the CO_2 concentration is gradually increasing. In atmospheric RT models, carbon dioxide is frequently taken to have a fixed and altitude-independent concentration, along with other 'uniformly mixed gases' (UMGs). The concentration profiles of the three major gas species are very different. Water vapor is located mainly in the lowest 2 km of the atmosphere. Ozone has a fairly flat profile from the ground through the stratosphere (~30 km). The UMGs decline exponentially with altitude, with a scale height of around 8 km.

Gases

Gas molecules both scatter and absorb light. Rayleigh scattering by gases scales inversely with the fourth power of the wavelength, and is responsible for the sky's blue color. For typical atmospheric conditions, the optical depth for Rayleigh extinction is approximately $0.009/\lambda^4$ per air mass (λ is in μm and air mass is defined by the vertical column from ground to space). The Rayleigh phase function has a $(1 + \cos^2 \theta)$ dependence.

Absorption by gases may consist of a smooth spectral continuum (such as the ultraviolet and visible electronic transitions of ozone) or of discrete spectral lines, which are primarily rotation lines of molecular vibrational bands. In the lower atmosphere (below around 25 km altitude), the spectral shape of these lines is determined by collisional broadening and described by the normalized Lorentz line shape formula:

$$f_{\text{Lorentz}}(\nu) = \frac{\alpha_c/\pi}{\alpha_c^2 + (\nu - \nu_0)^2} \qquad [4]$$

Here ν is the wavenumber (in cm^{-1}) [$\nu = (10\,000\ \mu\text{m/cm})/\lambda$], ν_0 is the molecular line transition frequency and α_c is the collision-broadened half-width (in cm^{-1}), which is proportional to pressure. The constant of proportionality, known as the pressure-broadening parameter, has a typical value on the order of $0.06\ \text{cm}^{-1}\ \text{atm}^{-1}$ at ambient temperature. The extinction coefficient $k(\nu)$ is the product of $f_{\text{Lorentz}}(\nu)$ and the integrated line strength S, which is commonly in units of $\text{atm}^{-1}\ \text{cm}^{-2}$.

At higher altitudes in the atmosphere the pressure is reduced sufficiently that Doppler broadening

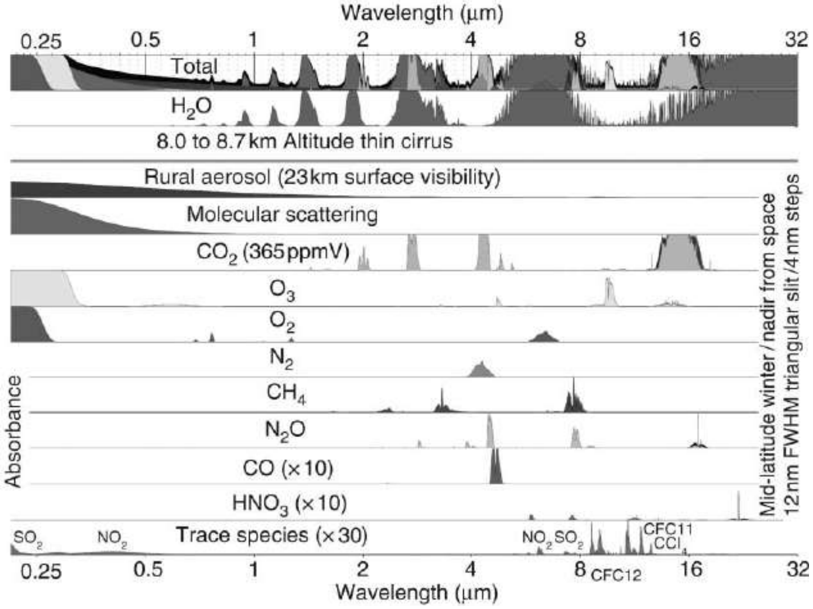

Figure 1 Spectral absorbance (1 − transmittance) for the primary sources of atmospheric extinction. The 12 nm resolution data were generated by MODTRAN for a vertical path from space with a mid-latitude winter model atmosphere.

becomes competitive with collisional broadening, and the Lorentz formula becomes inaccurate. The general lineshape for combined collisional and Doppler broadening is the Voigt lineshape, which is proportional to the real part of the complex error (probability) function, w:

$$f_{\mathrm{Voigt}}(\nu) = \frac{1}{\alpha_{\mathrm{d}}\sqrt{\pi}} \mathrm{Re}\left[w\left(\frac{\nu - \nu_0}{\alpha_{\mathrm{d}}} + \frac{\alpha_{\mathrm{c}}}{\alpha_{\mathrm{d}}} \mathrm{i} \right) \right] \qquad [5]$$

Here, α_{d} is the Doppler $1/e$ half-width.

Comprehensive spectral databases have been compiled of the transition frequencies, strengths, and pressure-broadened half-widths for atmospherically important molecules throughout the electromagnetic spectrum. Perhaps the most notable of these databases is HITRAN, which was developed by the US Air Force Research Laboratory and is currently maintained at the Harvard-Smithsonian Center for Astrophysics in Cambridge, MA.

Liquids and Solids

The larger particulates in the atmosphere (greater than a few μm in radius) typically belong to clouds. Low-altitude clouds consist of nearly spherical water droplets, while high-altitude cirrus clouds are mainly a collection of ice crystals. Other large particulates include sand dust. Their light scattering is close to the geometric limit at visible and ultraviolet wavelengths. This means that the extinction is nearly wavelength-independent, and the scattering phase function and single-scattering albedo may be reasonably modeled with ray-tracing techniques that account for the detailed size and shape distributions of the particles. However, Mie scattering theory is typically used to calculate cloud optical properties because it is exact for spherical particles of any size.

The smaller particulates in the atmosphere belong to aerosols, which are very fine liquid particles, and dusts, which are solids such as minerals and soot. These particulates are concentrated mainly in the lower 2 km or so of the atmosphere; however, they are also present at higher altitudes in smaller concentrations. Their optical properties are typically modeled using Mie theory. The wavelength dependence of the scattering is approximately inversely proportional to a low power of the wavelength, typically between 1 and 2, as befits particulates intermediate in size between molecular and geometric-limit. The scattering phase functions have a strong forward-scattering peak; values of the asymmetry parameter g (the average value of $\cos\theta$) typically range from 0.6 to 0.8 at solar wavelengths.

Solution Methods

Geometry

Atmospheric properties are primarily a function of altitude, which determines pressure, temperature and species concentration profiles. Accordingly, most RT methods define a stratified atmosphere. The most accurate treatments of transmission and scattering account for the spherical shape of the layers and refraction; however, most RT models use a plane-parallel approximation for at least some computations, such as multiple scattering.

Spectral Resolution

Optical instruments have finite, and frequently broad, wavelength responses. Nevertheless, modeling their signals requires accounting for the variation of absorption on an extremely fine wavelength scale, smaller than the widths of the molecular lines.

‘Exact’ monochromatic methods

The most accurate RT solution method involves explicitly solving the RT problem for a very large number of monochromatic wavelengths. This line-by-line method is used in a number of RT models, such as FASCODE. It allows Beer’s law to be applied to combine transmittances from multiple LOS segments, and provides an unambiguous definition of the optical parameters. It is suitable for use with spectrally structured light sources, such as lasers. The one major drawback of this method is that it is computationally intensive, and therefore may not be practical for problems where large wavelength ranges, multiple LOS views and multiple atmospheric conditions need to be treated.

To alleviate the computational burden of monochromatic calculations, some approximate methods have been developed that model RT in finite spectral intervals, as described below.

Statistical band models

The band model method represents spectral lines in a narrow interval, $\Delta\nu$, statistically using such parameters as the total line strength, the mean pressure-broadening parameter, and the effective number of lines in the interval. An example of a popular band model-based RT algorithm is MODTRAN, which is described below.

A key to the success of band models is the availability of approximate analytical formulas for the integrated absorption for an individual molecular transition of strength S, known as the single-line total equivalent width, W_{sl}:

$$W_{\mathrm{sl}} = \int_{-\infty}^{\infty} [1 - \exp(-Suf_{\nu})]\mathrm{d}\nu \qquad [6]$$

In the optically thin (small absorption) limit, W_{sl} is proportional to the molecular species column density, while in the optically thick (large absorption) limit it scales as the square root of the column density. A further assumption made by MODTRAN is that the line centers are randomly located within the interval, i.e., spectrally uncorrelated. With this assumption, the net transmittance can be expressed as the product of the transmittances for each individual line, whether the line belongs to the same molecular species or a different species.

A particular challenge in band models is treatment of the inhomogeneous path problem – that is, the variations in path properties along a LOS and their effect on the statistical line parameters, which arise primarily from differences in pressure and hence line width. The Curtis–Godson path averaging method provides a reasonable way to define 'equivalent' homogeneous path parameters for the band model. Another challenge is to define an effective extinction optical depth for each layer in order to solve the RT problem with scattering. One option in MODTRAN is to define it by computing the cumulative transmittance through successive layers in a vertical path.

Correlated-*k* model

Another well-known approximate RT algorithm for spectral intervals is the correlated-*k* method. This method starts with an 'exact' line-by-line calculation of extinction coefficients (k's) within the interval on a fine spectral grid, from which a *k*-distribution (vs. cumulative probability) is computed. A database of *k* values and probabilities summing to 1 is built from these *k*-distributions for a grid of atmospheric pressures and temperatures, and for all species contributing to the spectral interval. Inhomogeneous paths are handled by recognizing that the size order of the k's is virtually independent of pressure and typically only weakly dependent on temperature. LOS transmittances, LOS radiances, and fluxes are calculated by interpolating the database over temperature and pressure to define the k's for each LOS segment, solving the monochromatic RT equation at each fixed distribution location, and finally integrating over the distribution. The correlated-*k* method has been found to be quite accurate for atmospheric paths containing a single molecular species; however, corrections must be applied for spectral intervals containing multiple species.

Scattering Methods

When the diffuse light field is of interest, scattering methods are used to calculate the source function J of eqns [1] and [2].

Single scattering

If scattering is weak, the approximation may be made that the solar radiation scatters only once. Thus the integral over the scattering phase function, eqn [2], is straightforwardly calculated using the direct radiance component, which is given by eqn [3]. The neglect of multiple scattering (i.e., the diffuse contribution to the source function) means that the diffuse radiance is underestimated; however, single scattering is sufficiently accurate for some atmospheric problems in clear weather and at infrared wavelengths.

Multiple scattering

A number of different methods have been developed to solve the multiple scattering problem. Two-stream methods, which are the simplest and fastest, resolve the radiance into upward and downward directions. These methods generally produce reasonably accurate values of hemispherically averaged radiance, which are also referred to as horizontal fluxes or irradiances.

A much more accurate approach to the multiple scattering problem is the method of discrete ordinates. It involves expansion of the radiation field, the phase function and the surface reflectance as a series of spherical harmonics, leading to a system of linear integral equations. Evaluation of the integrals by Gaussian quadrature leads to a solvable system of linear differential equations. An important approximation called delta-M speeds up convergence of the discrete ordinates method, especially when scattering phase functions are strongly forward peaked, by representing the phase function as the sum of a forward direction δ-function and a remainder term. For most scattering problems, the solution is converged upon with a modest number ($\sim$8 to 16) of quadrature points (streams).

A very different type of multiple scattering technique, called the Monte Carlo method, is based on randomly sampling a large number of computer-simulated 'photons' as they travel through the atmosphere and are absorbed and scattered. The basic idea here is that sensor radiance can be expressed as a multiple path integral over the local source terms, and Monte Carlo methods solve integrals by sampling the integrand. The major advantage of this method is that it is flexible enough to allow for all of the complexity of a realistic atmosphere, often neglected by other methods. The major drawback is its computational burden, as a large number of photons is required for reasonable convergence; the Gaussian error in the calculation declines with the square root of the number of photons. The convergence problem is moderated to a large extent by using the physics of the problem being solved to bias the selection of photon paths toward those trajectories which contribute most; mathematically, this is equivalent to requiring that the integrand be sampled most often where its contributions are most significant.

An Example Atmospheric RT Model: MODTRAN

MODTRAN, developed collaboratively by the Air Force Research Laboratory and Spectral Sciences, Inc., is the most widely used atmospheric radiation transport model. It defines the atmosphere using stratified layering and computes transmittances, radiances, and fluxes using a moderate spectral resolution band model with IR through UV coverage. The width of the standard spectral interval, or bin, in MODTRAN is 1 cm^{-1}. At this resolution, spectral correlation among extinction sources is well characterized as random. Thus, the total transmittance from absorption and scattering of atmospheric particulates and molecular gases is computed as the product of the individual components.

Rayleigh, aerosol, and cloud extinction are all spectrally slowly varying and well represented by Beer's Law absorption and scattering coefficients on a 1 cm^{-1} grid. Calculation of molecular absorption is more complex because of the inherent spectral structure and the large number of molecular transitions contributing to individual spectral bins. As illustrated in Figure 2, MODTRAN partitions the spectral bin molecular attenuation into 3 components:

- *Line center absorption* from molecular transitions centered within the spectral bin;
- *Line tail absorption* from the tails of molecular lines centered outside of the spectral bin but within 25 cm^{-1}; and
- H_2O and CO_2 *continuum absorption* from distant (>25 cm^{-1}) lines.

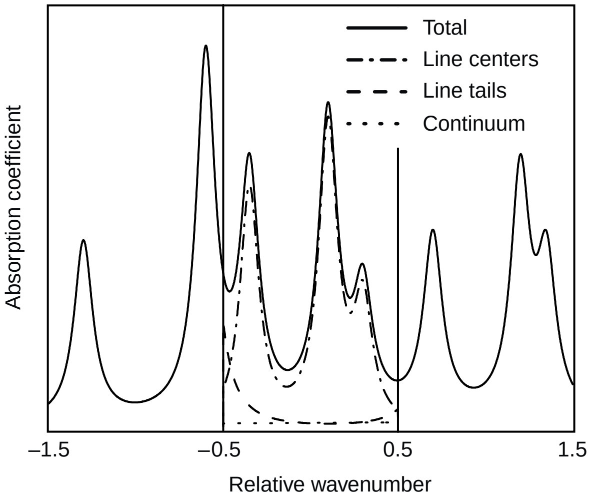

Figure 2 Components of molecular absorption. The line center, line tail and continuum contributions to the total absorption are illustrated for the central 1 cm^{-1} spectral bin.

Within the terrestrial atmosphere, only H_2O and CO_2 have sufficient concentrations and line densities to warrant inclusion of continuum contributions. These absorption features are relatively flat and accurately modeled using 5 cm^{-1} spectral resolution Beer's Law absorption coefficients.

Spectral bin contributions from neighboring line tails drop off, often rapidly, from their spectral bin edge values, but the spectral curves are typically simple, containing at most a single local minimum. For MODTRAN, these spectral contributions are pre-computed for a grid of temperature and pressure values, and fit with Padé approximants, specifically the ratio of quadratic polynomials in wavenumber. These fits are extremely accurate and enable line tail contributions to be computed on an arbitrarily fine grid. MODTRAN generally computes this absorption at a resolution equal to one-quarter the spectral bin width, i.e., 0.25 cm^{-1} for the 1.0 cm^{-1} band model.

The most basic *ansatz* of the MODTRAN band model is the stipulation that molecular line center absorption can be approximated by the absorption of n identical Voigt lines randomly distributed within the band model spectral interval, $\Delta\nu$. Early in the development of RT theory, Plass derived the expression for the transmittance from these n randomly distributed lines:

$$T = \left(1 - \frac{W'_{\mathrm{sl}}}{\Delta\nu}\right)^n \qquad [7]$$

MODTRAN's evolution has resulted in a fine tuning of the methodology used to define both the effective line number n and the in-band single-line equivalent width W'_{sl}. The effective line number is initially estimated from a relationship developed by Goody, in which lines are weighted according the to square root of their strength, but MODTRAN combines nearly degenerate transitions into single lines because these multiplets violate the random distribution assumption. The initial effective line number values are refined to insure a match with higher resolution transmittance predictions degraded to the band model resolution. The in-band equivalent width is computed for an off-centered Voigt line of strength *S*. Lorentz and Doppler half-widths are determined as strength-weighted averages. The off-center distance is fixed to insure that the weak-line Lorentz equivalent width exactly equals the random line center value.

MODTRAN scattering calculations are optimally performed using the DISORT discrete ordinates algorithm developed by Stamnes and co-workers. Methods for computing multiple scattering such as DISORT require additive optical depths, i.e., Beer's Law transmittances. Since the in-band molecular transmittances of MODTRAN do not satisfy Beer's Law, MODTRAN includes a correlated-*k* algorithm option. The basic band model *ansatz* is re-invoked to efficiently determine *k*-distributions; tables of *k*-data are pre-computed as a function of Lorentz and Doppler half-widths and effective line number, assuming spectral intervals contain *n* randomly distributed identical molecular lines. Thus, MODTRAN *k*-distributions are statistical, only dependent on band model parameter values, not on the exact distribution of absorption coefficients in each spectral interval.

Applications

Among the many applications of atmospheric transmission and scattering calculations, we briefly describe two complementary ones in the area of remote sensing, which illustrate many of the RT features discussed earlier as well as current optical technologies and problems of interest.

Earth Surface Viewing

The first example is Earth surface viewing from aircraft or spacecraft with spectral imaging sensors. These include hyperspectral sensors, such as AVIRIS, which have typically a hundred or more contiguous spectral channels, and multispectral sensors, such as Landsat, which typically have between three and a few tens of channels. These instruments are frequently used to characterize the surface terrain, materials and properties for such applications as mineral prospecting, environmental monitoring, precision agriculture, and military uses. In addition, they are sensitive to properties of the atmosphere such as aerosol optical depth and column water vapor. Indeed, in order to characterize the surface spectral reflectance, it is necessary to characterize and remove the extinction and scattering effects of the atmosphere.

Figure 3 also shows an example of data collected by the AVIRIS sensor at ~3 km altitude over thick vegetation. The apparent reflectance spectrum is the observed radiance divided by the Top-of-Atmosphere horizontal solar flux. Atmospheric absorption by water vapor, oxygen, and carbon dioxide is evident, as well as Rayleigh and aerosol scattering. Figure 3 shows the surface reflectance spectrum inferred by modeling and then removing these atmospheric effects (this process is known as atmospheric removal, compensation, or correction). It has the characteristic smooth shape expected of vegetation, with strong chlorophyll absorption in the visible and water bands at longer wavelengths. A detailed analysis of such a spectrum may yield information on the vegetation type and its area coverage and health.

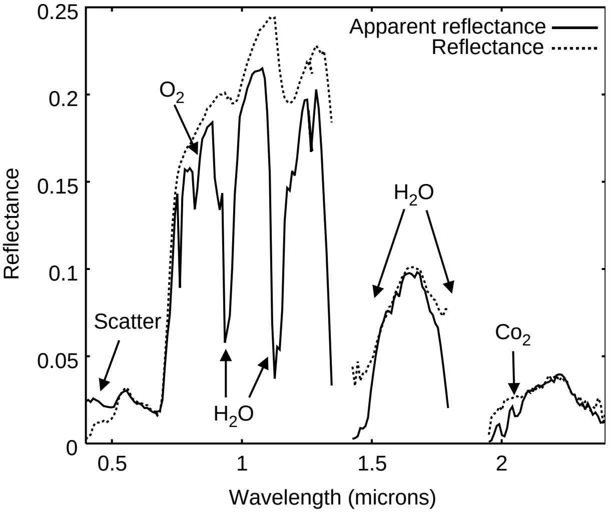

Figure 3 Vegetation spectrum viewed from 3 km altitude before (apparent reflectance) and after (reflectance) removal of atmospheric scatter and absorption.

Sun and Sky Viewing

The second remote sensing example is sun and sky viewing from the Earth's surface with a spectral radiometer, which can yield information on the aerosol content and optical properties as well as estimates of column concentrations of water vapor, ozone, and other gases. Figure 4 shows data from a Yankee Environmental Systems, Inc. multi-filter rotating shadow-band radiometer, which measures both 'direct flux' (the direct solar flux divided by the cosine of the zenith angle) and diffuse (sky) flux in narrow wavelength bands. The plot of ln(direct signal) versus the air mass ratio is called a Langley plot, and is linear for most of the bands, illustrating Beer's Law. The extinction coefficients (slopes) vary with wavelength consistent with a combination of Mie and Rayleigh scattering. The water-absorbing 940 nm band has the lowest values and a curved

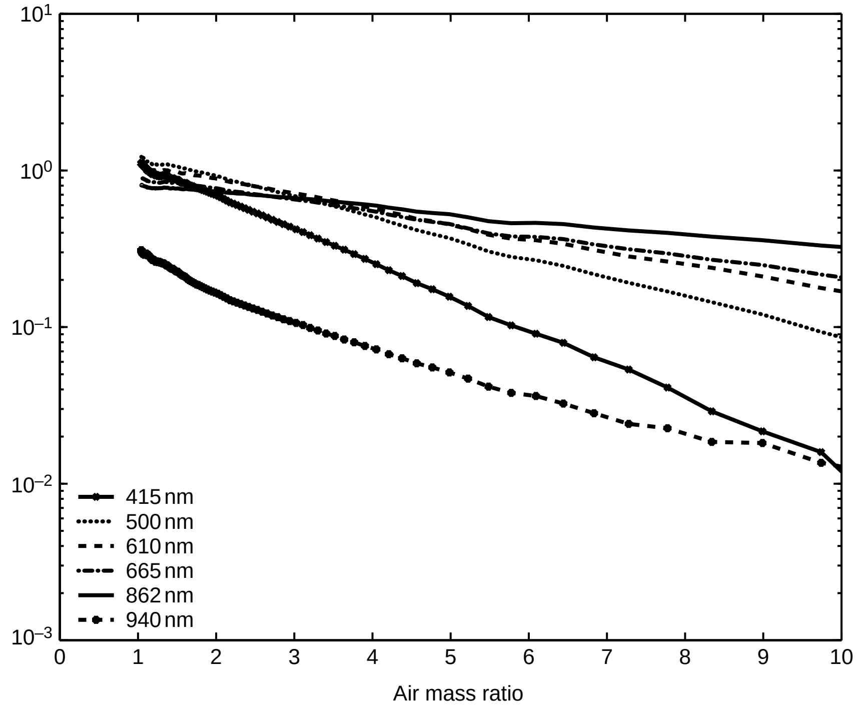

Figure 4 Direct solar flux versus air mass at the surface measured by a Yankee Environmental Systems, Inc. multi-filter rotating shadow-band radiometer at different wavelengths.

plot, in accordance with the square-root dependence of the equivalent width for optically thick lines. The diffuse fluxes, which arise from aerosol and Rayleigh scattering, have a very different dependence on air mass than the direct flux. In particular, the diffuse contributions often increase with air mass for high sun conditions.

The ratio of the diffuse and direct fluxes is related to the total single-scattering albedo, which is defined as the ratio of the total scattering coefficient to the extinction coefficient. Results from a number of measurements showing lower than expected diffuse-to-direct ratios have suggested the presence of black carbon or some other continuum absorber in the atmosphere, which would have a significant impact on radiative energy balance at the Earth's surface and in the atmospheric boundary layer.

List of Units and Nomenclature

Line shape function [cm]	f_ν
Lorentz line shape function [cm]	$f_{\mathrm{Lorentz}}(\nu)$
Voigt line shape function [cm]	$f_{\mathrm{Voigt}}(\nu)$
Scattering asymmetry parameter	g
Direct plus diffuse radiance (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I
Direct radiance (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I'
Direct radiance at the LOS boundary (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I'_0
Source function (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	J
Extinction coefficient (cm^{-1} atm^{-1})	k
Effective number of lines in bin	n
Scattering phase function (sr^{-1})	p
Line strength (cm^{-2} atm^{-1})	S
Transmittance	T
Column density (atm cm)	u
Complex error (probability) function	w
Single-line total equivalent width (cm^{-1})	W_{sl}
Single-line in-band equivalent width (cm^{-1})	W'_{sl}
Collision-broadened half-width at half maximum (cm^{-1})	α_{c}
Doppler half-width at 1/*e* (cm^{-1})	α_{d}
Spectral interval (bin) width (cm^{-1})	$\Delta\nu$
Scattering angle (radian)	θ
Wavelength (μm)	λ
Wavenumber (cm^{-1})	ν
Molecular line transition frequency	ν_0
Direction of incoming light (sr)	Ω_{i}
Direction of outgoing light (sr)	Ω_{o}

See also

Environmental Measurements: Laser Detection of Atmospheric Gases. **Instrumentation:** Spectrometers. **Scattering:** Scattering from Surfaces and Thin Films.

Further Reading

Berk A, Anderson GP, Acharya PK, Bernstein LS, Shettle EP, Adler-Golden SM, Lee J and Muratov L (2004) MODTRANS User's Manual. Airforce Research Laboratory, Hanscom AFB, MA, in press.

Chandrasekhar S (1960) *Radiative Transfer*. New York: Dover Publications, Inc., Originally published by Oxford University Press, London (1950).

Fenn RW, Clough SA, Gallery WO, Good RE, Kneizys FX, Mill JD, Rothman LS, Shettle EP and Volz FE (1985) Optical and infrared properties of the atmosphere. In: Jursa AS (ed.) *Handbook of Geophysics and the Space Environment*, chap. 18, US Air Force Geophysics Laboratory, Hanscom AFB, MA, AFGL-TR-85-0315, available from NTIS ADA 167000.

Goody RM and Yung YL (1989) *Atmospheric Radiation, Theoretical Basis*, 2nd edn. New York: Oxford University Press.

HITRAN Special Issue, J. Quant. Spectrosc. Rad. Transfer, vol. 82, Nos. 1–4, 15 November–15 December (2003).

Houghton J (2001) *The Physics of Atmospheres*, 3rd edn, Cambridge, UK: Cambridge University Press.

Killinger DK, Churnside JH and Rothman LS (1994) Atmospheric optics. In *Handbook of Optics*, 2nd edn, 2-Volume Set, chap. 44. New York: McGraw-Hill.

Liou KN (1997) *An Introduction to Atmospheric Radiation*, 2nd edn. London: Academic Press.

Smith FG (ed.) (1993) *The Infrared and Electro-Optical Systems Handbook*, vol. 2, *Atmospheric Propagation of Radiation*. Ann Arbor: Environmental Research Institute of Michigan.

Thomas GE and Stamnes K (1999) *Radiative Transfer in the Atmosphere and Ocean*. Cambridge, UK: Cambridge University Press.

Van de Hulst HC (1981) *Light Scattering by Small Particles*. New York: Dover Publications, Inc., Originally published in 1957, New York: John Wiley & Sons, Inc.

F

FIBER AND GUIDED WAVE OPTICS

Contents

Overview
Dispersion
Fabrication of Optical Fiber
Light Propagation
Measuring Fiber Characteristics
Nonlinear Effects (Basics)
Nonlinear Optics
Optical Fiber Cables
Passive Optical Components

Overview

A Mickelson, University of Colorado, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is said that some in ancient Greece knew that light could be guided inside slabs of transparent material. From this purported little-known curiosity in the ancient world, guided wave optics has grown to be the technology of the physical level of the systems which transfer most of the information about the world; the waveguide of the worldwide telecommunications network is the optical fiber. The optical fiber itself is but a passive waveguide, and guided wave optics is the technology which includes all of the passive and active components which are necessary to prepare optical signals for transmission, regenerate optical signals during transmission, route optical signals through systems and code onto optical carriers, and decode the information from optical carriers, to more conventional forms. In this article, some introduction to this rather encompassing topic will be given.

This article will be separated into four parts. In the first section, discussion will be given to fiber optics, that is, the properties of the light guided in optical waveguides which allow the light to be guided in distinctly nonrectilinear paths over terrestrial distances. The second section will then turn to the components which can be used along with the fiber optical waveguides in order to form useful systems. These components include sources and detectors as well as optical amplifiers. In the third section, we will discuss the telecommunications network which has arisen due to the availability of fiber optics and fiber optic compatible components. The closing section will discuss integrated optics, the field of endeavor which has such great promise to form the future of optical technology.

Fiber Optics

Fiber optics is a term which generally refers to a technology in which light (actually infrared, visible, or ultraviolet radiation) is transmitted through the transparent cores of small threads of composite material. These threads, or fibers as they are called, when surrounded by a cladding material and coated with polymer for environmental protection (see Figure 1) can be coiled like conventional wire and

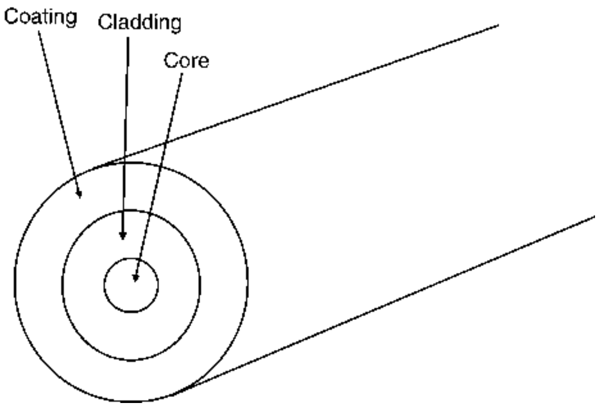

Figure 1 A depiction of the structure of an optical fiber. The innermost cylinder is the core region in which the majority of the light is confined. The next concentric region is the cladding region which is still made of a pure material but one of a lower index of refraction than the innermost core region, such that the light of the lightwave decays exponentially with extension into this region. The outermost region is a coating which protects the fused silica of the core and cladding from environmental contaminants such as water. For telecommunications fibers, the core is comprised of fused silica doped with germanium and typically has an index of refraction of 1.45, whereas the cladding differs from this index by only about 1%. The coating is often a polyimide plastic which has a higher index (perhaps 1.6) but no light guided light should see the cladding-coating interface.

when cabled can resemble (a very lightweight flexible version of) conventional transmission wire. Although most of the optical fiber in use today is fabricated by a process of gas phase chemical deposition of fused silica doped with various other trace chemicals, fiber can be made from a number of different material systems and in a number of different configurations for use with various types of sources. In what follows in this opening section, we will limit discussion to the basic properties of the light guided by the fiber and leave more technological discussion to following sections.

There are two complementary mathematical descriptions of the propagation of light in an optical waveguide. In the ray description, light incident on a fiber endface is considered to be made up of a bundle of rays. In a uniform homogeneous medium, each ray is like an arrow that exhibits rectilinear propagation from its source to its next interface with a dissimilar material. These rays satisfy Snell's laws of reflection and refraction at interfaces between materials with dissimilar optical properties that they encounter along their propagation path. That is, at an interface, a fraction of the light is reflected backwards at an angle equal to the incident angle and a portion of the light is transmitted in a direction which is more directed toward the normal to the interface when the index increases across the boundary and is directed more away from the normal when the index decreases. At the input endface of a waveguide, a portion of the energy guided by each ray is refracted due to the change in refractive index at the guide surface and then exhibits a more interesting path within the fiber. In a step index optical fiber, where the index of refraction is uniformly higher in the fiber core than in a surrounding cladding, the rays will propagate along straight paths until encountering the core cladding interface. Guided rays (see Figure 2) are totally internally reflected back into the fiber core to again be totally internally reflected at the next core cladding interface and so on. Radiating rays (see Figure 3) will be only partially reflected at the core cladding interface and will rapidly die out in propagating down the fiber when it is taken into account that typical distances between successive encounters with the boundary may be sub-millimeter and total propagation distances may be many kilometers. In graded index fibers, the refractive index within the fiber core varies continuously from a maximum somewhere within the core to a minimum at which the core ends and attaches continuously to a cladding. The ray paths within such fibers are curved and the guided rays are characterized by the fact that once in the fiber they never encounter the cladding. Radiating rays encounter the cladding and are refracted out of the fiber. This description of fiber propagation is quite simple and pleasing but does not take into account that each ray actually is carrying a clock that remembers how long it has been following its given path. When two rays come together, they can

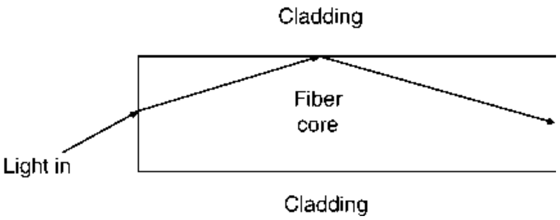

Figure 2 Schematic depiction of the geometrical optics interpretation of the coupling of light into an optical fiber. Here a ray that is incident on the center of the fiber core at an upward angle is first refracted at the fused silica–air interface into the fused silica core. The ray is then totally internally reflected at the core cladding interface such that the ray power remains in the core. The ray picture of light coupling is applicable to single rays coupling to a multimode fiber. When there are multiple rays, interference between these rays must be taken into account which is difficult to do in the ray picture. When there is only one (diffraction limited) mode in the fiber, the interference between a congruence of (spatially coherent) incident rays is necessary to describe the coupling. This description is more easily effected by a quasi-monochromatic mode picture of the propagation.

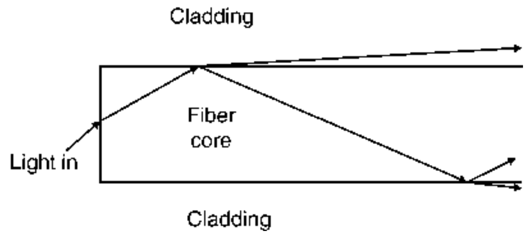

Figure 3 Schematic depiction of the existence of a cut-off angle for coupling into an optical fiber. A ray incident from air on the center of the fused silica fiber core is refracted at the air fused silica interface and then split when it comes to the core cladding interface within the fiber. Were the refracted angle of this ray zero (along the core cladding interface) we would say the ray is totally internally reflected and no power escapes. As it is, there is refracted power into the cladding and the ray will rapidly attenuate as it propagates along the fiber. Although most telecommunications fiber in this day and age is graded index (the rays will curve rather than travel along straight paths), a typical repeat period in such a fiber is a millimeter, whereas propagation distances may be 100 km. With a loss of even only 0.1% at each reflection form the cladding, the light light at cut-off will be severely attenuated in one meter, no loss, one kilometer.

either add or subtract, depending on the reading on their respective clocks. When the light is nearly monochromatic, the clocks' readings are simply a measure of the phase of the ray when it was radiated from its source and the interference between rays can be quite strong. This interference leads to conditions which only allow certain ray paths to be propagated. When the light has a randomly varying phase in both space and time (and is therefore polychromatic), the interference all but disappears. But the condition that allows coupling to a specific ray path (mode) also is obliterated by the phase randomness, and coupling to a single mode guide, for example, becomes inefficient.

When we need to preserve phase relations in order to preserve information while it propagates along a fiber path, we need to use single modes excited by nearly transform limited sources, that is sources whose time variation exhibits well-defined (nonrandom) phase variation. As a monochromatic wave carries no information, we will require a source which can be modulated. When modulation can be impressed on a carrier without loss of phase information of either information or carrier, information can be propagated at longer distances than when the frequency spectrum of the carrier is broadened by noise generated during the modulation. We refer to such a source that can be modulated without broadening as a coherent or nearly coherent source. We cannot easily adapt the ray picture to the description of propagation of such coherent radiation in an optical fiber. Instead we must resort to using the time harmonic form of Maxwell's equations, the equations which describe electromagnetic phenomena, whereas rays can be described by a simplification of the time-dependent form of these equations. In the time harmonic approach, we assume that the source is monochromatic and then solve for a set of modes of the fiber at the assumed frequency of the source. These modes have a well-defined phase progression as a function of carrier frequency and possess a given shape in the plane transverse to the direction of propagation. Information can be included in the propagation by assuming the quasi-monochromatic variation of the source, that is, one assumes that the source retains the phase relations between the various modulated frequency components even while its amplitude and phase is being varied externally. When one assumes that time harmonically (monochromatic) varying fields are propagating along the fiber axis, one obtains solutions of Maxwell's equations in the form of a summation of modes. These are guided modes, ones that propagate down the fiber axis without attenuation. There are also radiation modes that never couple into a propagating mode in the fiber. These modes are analogous to the types of rays we see in the ray description of fiber propagation. There are also another set of modes which are called evanescent modes which show up at discontinuities in the fiber or at junctions between the fiber and other components. In the modal picture of fiber propagation, each of these modes is given an independent complex coefficient (that is, a coefficient with both amplitude and phase). These coefficients are determined first by any sources in the problem and then must be recalculated at each discontinuity plane along the propagation path in the fiber. When the source(s) in the problem is not transform limited, the phases of the coefficients become smeared out and the coupling problems at discontinuities take on radically different solutions.

In some limit, these randomized solutions must appear as the ray solutions. Optical fibers are generally characterized by their numerical aperture (NA), as well as the number which characterizes its transverse dimension. The numerical aperture is essentially the sine of the maximum angle into which the guide will radiate into free space. When the guide is multimoded, then the transverse dimension is generally given as a diameter. When a guide is single-moded, the transverse dimension is generally given as a spotsize, that is, by a measure of the size of a unity magnification image of the fiber endface. Multimode guides can be excited by even poorly coherent sources so long as they radiate into the NA or capture angle at the input of the guide. Single-mode fibers require an excitation which matches the shape and size of their fundamental mode.

Active Fiber Compatible Components

Our discussion of propagation in fiber optic waveguides would be a rather sterile one if there were not a number of fiber compatible components that can generate, amplify, and detect light streams available to us in order to construct fiber optic systems. In this section, we will try to discuss some of the theory of operation of active optical components in order to help elucidate some of the general characteristics of such components and the conditions that inclusion of such components in a fiber system impose on the form that the system must take.

In a passive component such as an optical fiber, we need only consider the characteristics of the light that is guided. Operation of active components requires that the optical fields interact with a medium in such a manner that energy can be transferred from the field through an excitation of the medium to the controlling electrical stream and/or vice versa. Generally one wants the exchange to go only one way, that is from field to electrical stream or from electrical stream to field. In a detector this is not difficult to achieve, whereas in lasers and amplifiers it is quite hard to eliminate the back reaction of the field on the device. Whereas a guiding medium can be considered as a passive homogeneous continuum, the active medium has internal degrees of freedom of its own as well as a granularity associated with the distribution of the microscopic active elements. At least, we must consider the active medium as having an active index of refraction with a mind (set of differential equations anyway) of its own. If one also wants to consider noise characteristics, one needs to consider the lattice of active elements which convert the energy. The wavelength of operation of that active medium is determined by the energy spacing between the upper and lower levels of a transition which is determined by the microscopic structure of these grains, or quanta, that make up the medium. In a semiconductor, we consider these active elements to be so uniformly distributed and ideally spaced that we can consider the quanta (electron hole pairs) to be delocalized but numerous and labeled by their momentum vectors. In atomic media such as the rare earth doped optical fibers which serve as optical amplifiers, we must consider the individual atoms as the players. In the case of the semiconductor, a current, flowing in an external circuit, controls the population of the upper (electronic) and lower (hole) levels of each given momentum state within the semiconductor. When the momentum states are highly populated with electrons and holes, there is a flow of energy to the field (at the transition wavelength) and when the states are depopulated, the field will tend to be absorbed and populate the momentum states by giving up its energy to the medium. The situation is similar with the atomic medium except that the pump is generally optical (rather than electrical) and at a different wavelength than the wavelength of the field to be amplified. That is to say, one needs to use a minimum of three levels of the localized atoms in the medium in order to carry out an amplification scheme.

The composition of a semiconductor determines its bandgap, that is, the minimum energy difference between the electron and hole states of a given momentum value. A source, be it a light emitting diode (LED) or laser diode (see Figure 4) will emit light at a wavelength which corresponds to an energy slightly above the minimum gap energy (wavelength equals the velocity of light times Planck's constant divided by energy) whereas a detector can detect almost any energy above the bandedge. Only semiconductors with bandgaps which exhibit a minimum energy at a zero momentum transfer can be made to emit light strongly. Silicon does not exhibit a direct gap and although it can be used as a detector, it cannot be used as a source material. The silicon laser is and has been the 'Holy Grail' of electronics because of the ubiquity of silicon electronics. To even believe that a 'Holy Grail' exists requires a leap of faith. Weak luminescence has been

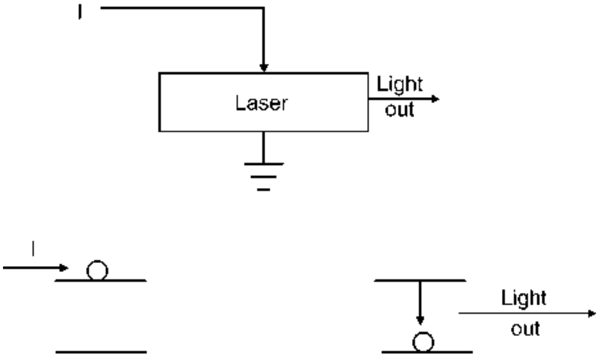

Figure 4 A schematic depiction of the workings of a semiconductor laser light source. The source is fabricated as a diode, and normal operation of this diode is in forward bias, that is, the p-side of the junction is biased positively with respect to ground which is attached to the n-side of the junction. With the p-side positively biased, current should freely flow through the junction. This is not quite true as there is a heterojunction region between the p- and n-regions in which electrons from the n-side may recombine with holes from the p-side. This recombination gives off a photon which is radiated. In light emitting diodes (LEDs), the photon simply leaves the light source as a spontaneously emitted photon. In a laser diode, the heterojunction layer serves as a waveguide and the endfaces of the laser as mirrors to provide feedback and allow laser operation in which the majority of the photons generated are generated in stimulated processes.

observed in certain silicon structures. This luminescence has been used for mid-infrared sources (where thermal effects mask strong luminescence) and in concert with rare earth dopants to make visible light emitting diodes. The silicon laser is still only a vision.

Source materials are, therefore, all made of compound semiconductors. Detectors are in essence the inverse of sources (see Figure 5) but also must detect at wavelengths above their gap energy. For primarily historical reasons, as will be further discussed below, telecommunications employs essentially three windows of wavelengths for different applications, wavelengths centered about 0.85 microns in the first window, wavelengths about 1.3 microns in the second window, and wavelengths about 1.55 microns in the third window. Materials made of layers of different compositions of AlGaAs and mounted on GaAs substrates are used in the first band, while the other bands require the increased degree of freedom allowed by the quaternary alloys of InGaAsP mounted on InP substrates. Other material mixes are possible, these above-mentioned materials are the most common. Rare earth ions exhibit many different transitions. The most useful ones have proven to be the transitions of Er in the third telecommunications window, the one that covers the spectral region to either side of 1.55 microns. Although both Nd and Pr can be made to amplify near the 1.3 micron window, amplifiers made of these materials have not proven to be especially practical.

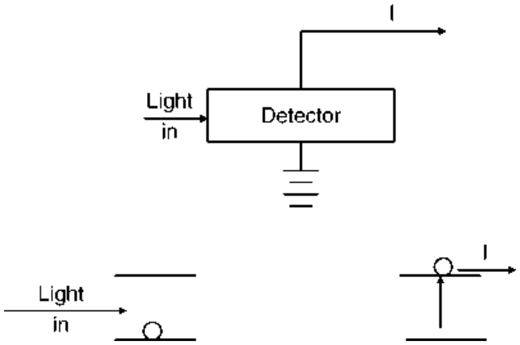

Figure 5 A schematic depiction of the workings of a semiconductor detector. As with a semiconductor source, the detector is fabricated to operate as a diode, that is a p-n junction, but in the case of the detector this diode is to be operated only in reverse bias, that is, with the n-material being biased positively with respect to the ground which is attached to the p-material's contact. In this mode of operation, no current should flow through the junction as there are no free carriers in the material, they have all been drawn out through the contacts. A photon of sufficiently high energy can change this situation by forming an electron hole pair in the material. The electron and hole are thereafter drawn to the n- and p-contacts, respectively, causing a detectable current to flow in the external circuit.

Telecommunications Technology

The physical transmission layer of the worldwide telecommunications network has come to be dominated by fiber optic technology, in particular, by wavelength division multiplexed or WDM single-mode fiber optics operating in the 1.55 micron telecommunications window. In the following, we will first discuss how this transformation to optical communications took place and then go on to discuss some of the specifics of the technology.

Already by the middle of the 1960s, it was clear that changes were going to have to take place in order that the exponential growth of the telephone system in the United States, as well as in Europe, could continue. The telephone system then, pretty much as now, consisted of a hierarchy of tree structures connected by progressively longer lines. A local office is used to connect a number of lines emanating in a tree structure to local users. The local offices are connected by trunk lines which emanate from a toll office. The lines from there on up the hierarchy are long distance ones which are termed long lines and can be regional and longer. The most pressing problem in the late 1960s was congestion in the so-called trunk lines which connect the local switching offices. The congestion was naturally most severe in urban areas. These trunk lines were generally one kilometer in length at that time. The problem was that there was no more space in the ducts that housed these lines. A solution was to use time division multiplexing or TDM to increase the traffic that could be carried by each of the lines already buried in the conduit. The problem was that the twisted pair lines employed would smear out the edges of the time varying bit streams carrying the information at the higher bitrates (aggregated rates due to the multiplexing) due to the inherent dependence of signal propagation velocity on frequency known as dispersion. After discussion and even development of a number of possible technologies, in 1975 a demonstration of a fiber optic system which employed multimode semiconductor lasers feeding multimode optical fibers all operating at 0.85 micron wavelength, proved to be the most viable model for trunk line replacement. The technology was successful and already in 1980 advances in single-mode laser and single-mode fiber technology operating at the 1.3 micron wavelength had made the inclusion of fiber into the long lines viable as well. For roughly the decade from 1985 onward, single-mode fiber systems dominated long line replacement for terrestrial as well as transoceanic links. The erbium doped fiber amplifier which operated in the 1.55 micron wavelength third telecommunication window had proven itself to be viable for extending repeater periods by around 1990. Development of the InGaAsP quaternary semiconductor system allowed for reliable lasers to be manufactured for this window as development of strained lattice laser technology in the InGaAs system allowed for efficient pump lasers for the fiber amplifiers. As the optical amplifiers could amplify across the wavelength band that could be occupied by many aggregated channels of TDM signals, the move of the long line systems to the third telecommunications window was accompanied by the adoption of wavelength division multiplexing or WDM. That is, electronics becomes more expensive as the analog bandwidth it can handle increases. Cost-effective digital rates are now limited to a few Gb/s, and analog rates to perhaps 20 GHz. Optical center frequencies are in the order of 2×10^{14} Hz and purely optical filters can be made to cleanly separate information signals spaced as closely as 100 GHz apart on optical carriers. An optical carrier can then be made to carry hundreds of channels of 10 Gb/s separated by 100 GHz and this signal can be propagated thousands of kilometers through the terrestrial fiber optics network. Such is today's Worldwide Web.

Integrated Optics as a Future of Guided Wave Optics

As electronics has been pushed to ever greater levels of integration, its power has increased and its price has dropped. A goal implicit in much of the development of planar waveguide technology has been that optics could become an integrated technology in the same sense as electronics. This has not occurred and probably will not occur. This is not to say that integrated optical processing circuits are not being developed and that they will be not employed in the future. They most definitely will.

Integrated optics was coined as a term when it was used in 1969 to name a new program at Bell Laboratories. This program was aimed at investigation of all the technologies that were possible with which to fabricate optical integrated circuits. The original effort was in no way driven by fiber optics as the Bell System only reluctantly moved to a fiber optic network solution in 1975. The original integrated optics efforts were based on such technologies as the inorganic crystal technologies that allowed for large second-order optical nonlinearities, a necessity in order that a crystal also exhibits a large electro-optic coefficient. An electro-optic coefficient allows one to change the index of refraction of a crystal by applying a low frequency or DC electromagnetic field across the crystal.

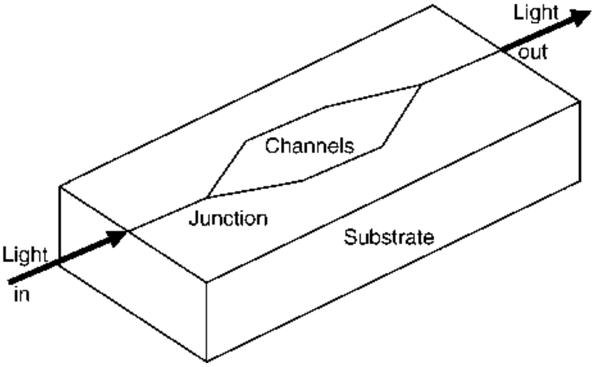

Figure 6 A schematic depiction of an integrated optical device which is often used as a high speed modulator. The device is an integrated version of a Mach–Zehnder interferometer. The basic idea behind its operation is that spatially and temporally coherent light is input into a single mode optical channel. That channel is then split into two channels by a Y-junction. Although not depicted in the figure, the two arms should not have completely equivalent propagation paths. That is to say, that if the light in one of those paths propagates a slightly longer distance (as measured in terms of phase fronts of the wave) then one cannot recombine the power from the two arms. That is, in the second Y-junction, a certain amount of the light which we are trying to combine from the two arms will be radiated out from the junction. In fact, if the light were temporally incoherent (phase fronts not well defined), exactly half of the light would be radiated form the junction. This is a statement of the brightness theorem which was thought to be a law of propagation before there were coherent sources of light. In a high speed modulator, the substrate is an electro-optic crystal and electrodes are placed over the two arms which apply the electrical signal to be impressed on the optical carrier to the channels.

Lithium niobate technology is a technology that has lasted to the present as an optical modulator technology (see Figure 6) and an optical switch technology (see Figure 7), as well as a technology for parametric wavelength conversion. Unfortunately, although low loss passive waveguide can be fabricated in lithium niobate, the crystal is not amenable to any degree of monolithic integration. Glass was also investigated as an integration technology from the early days of integrated optics, but the lack of second-order nonlinearity in glass strictly limited its applicability. Attention for a period turned to monolithic integration in the semiconductor materials which were well progressing as for use as lasers, detectors, and integrated circuits. Semiconductor crystals, however, are too pure to allow for low loss light propagation which requires the material defects to be so numerous that optical wavelengths cannot sample them. Passive waveguides in semiconductors incur huge losses that can only be mitigated by almost constant optical amplification which, unfortunately, cannot track the rapid optical field variations necessary for information transfer. Early on, attention turned to silicon

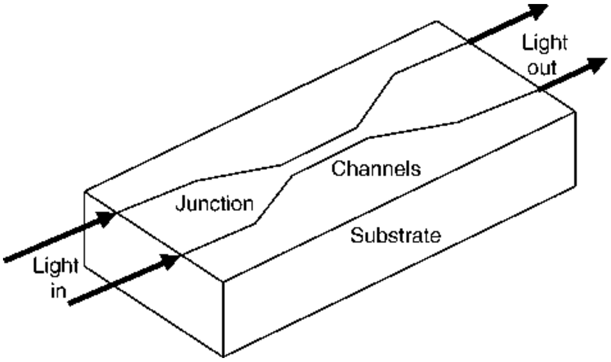

Figure 7 A schematic depiction of an integrated optical device which is often used as an optical switch. This device is interferometric but really has no free space optics counterpart as the Mach–Zhender interferometer does. In this interferometer, spatially and temporally coherent light is input into a single mode input channel. The two single optical mode input channels, which are initially so far apart from each other in terms of their mode sizes that they are effectively uncoupled, are then brought ever closer together until they are strongly coupled and their modes are no longer confined to their separate waveguides but are shared between the two channels. As the channels are symmetric, but the initial excitation, if from a single waveguide alone, is not, the pattern must evolve with propagation. Were both input channels excited, symmetry would still require a symmetry of their phases. If the phases were either equal to each other or completely out of phase, then there would be no evolution because we would have excited coupler modes. If the coupled region is the proper length, light input to one channel will emerge from the other after the channels are pulled apart. If the substrate is an electro-optic crystal and electrodes are placed over the channels, application of a voltage can affect the evolution of the interference in such a manner so as to 'decouple' the channels and switch the light back to the original channel.

optical bench technology, a technology which has presently spawned micro-electro-mechanical-system (MEMs) technology and still persists as an expensive vehicle to hybrid electro-optical integration. In silicon bench, individual components are micro-mounted with solder or silicon pop-up pieces on a silicon wafer which serves as a micro-optical bench. The technology is expensive and hybrid but pervasive in the highest end of cost classes of optical systems.

The tremendous success of the fiber optical network spawned a movement in which groups tried to achieve every possible system functionality in fibers themselves. A notable success has been the optical fiber amplifier. Significantly less successful in the fiber optic network has been the fiber Bragg grating. Attempts to carry out WDM functionalities in all fiber configuration leads to unwieldy configurations requiring large amounts of fiber and significant propagation delays, although, Bragg grating sensors have proven to be somewhat useful in sensing applications. WDM functions have been implemented in a large number of different hybrid configurations, involving various types of glasses, fused silica on silicon and simple hybrids of birefringent crystals.

The telecommunications network is not the driver that it once was. That this is so is perhaps the strongest driving force yet for integrated optics. The driver now is new applications that must be implemented in the most efficient manner. The most efficient manner is quite generally the one in which there is a maximum degree of integration. In the long run, polymer has always been the winning technology for optics of any kind, due to the flexibility of the technology and the drop in cost that accompanies mass production. Indeed, polymer integrated optics progresses. There are also a number of researchers involved in investigating strongly guiding optics, so-called photonic crystal optics. That is, in order to achieve low loss as well as low cost, the fiber and integrated optic waveguides up to the present have used small variations in optical properties of material to achieve guidance at the cost of having structures that are many wavelengths in size. In radio frequency (RF) and microwave technology, for example, guides are tiny fractions of a wavelength. This can lead to an impedance matching problem. RF and microwave impedance matching costs (versus circuit impedance matching costs) are generally quite high, precluding mass application. That radio frequency systems, such as cell phone transceivers link to the rest of the world by antenna, allows that the overall circuit dimension can be kept less than a single wavelength and impedance matching can be foregone in this so-called circuit limit. It is usually hard to keep an overall microwave system to less than a wavelength in overall extent except in the cell phone case or in a microwave oven. Optical miniaturization will require high index contrast guides, and will require optical impedance matching. There are few complete optical systems which will comprise less than an optical wavelength in overall extent. But then this so-called photonic crystal technology will likely be polymer compatible and there may be ways to find significant cost reduction. The future of integrated optics is unclear but bright.

See also

Fiber and Guided Wave Optics: Dispersion; Fabrication of Optical Fiber; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Nonlinear Optics; Optical Fiber Cables; Passive Optical Components. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1995) *Fiber-Optic Communication Systems*. New York: Wiley.

Bette, DeMarchis G and Iannone (1995) *Coherent Communications*. New York: Wiley.

Haken H (1985) *Light*. New York: Elsevier North-Holland.

Hecht E (1998) *Optics*. Reading, MA: Addison-Wesley.

Klein MV and Furtak TE (1986) *Optics*. New York: Wiley.

Marcuse D (1982) *Light Transmission Optics*. New York: Van Nostrand Reinhold.

Marcuse D (1991) *Theory of the Dielectic Optical Waveguide*. Boston, MA: Academic Press.

Mickelson AR, Lee Y-C and Basavanhally NR (1997) *Optoelectronic Packaging*. New York: Wiley.

Saleh BAA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Yariv A (1997) *Optical Electronics in Modern Communications*. New York: Van Nostrand Reinhold.

Dispersion

L Thévenaz, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Dispersion designates the property of a medium to propagate the different spectral components of a wave with a different velocity. It may originate from the natural dependence of the refractive index of a dense material to the wavelength of light, and one of the most evident and magnificent manifestations of dispersion is the appearance of a rainbow in the sky. In this case dispersion gives rise to a different refraction angle for the different spectral components of the white light, resulting in an angular dispersion of the sunlight spectrum and explicating the etymology of the term dispersion. Despite this spectacular effect, dispersion is mostly seen as an impairment in many applications, in particular for optical signal transmission. In this case, dispersion causes a rearrangement of the signal spectral components in the time domain, resulting in temporal spreading of the information and eventually in severe distortion.

There is a trend to designate all phenomena resulting in a temporal spreading of information as a type of dispersion, namely the polarization dispersion in optical fibers that should be properly named as polarization mode delay (PMD). In this article chromatic dispersion will be solely addressed, that designates the dependence of the propagation velocity on wavelength and that corresponds to the strict etymology of the term.

Effect of Dispersion on an Optical Signal

An optical signal may always be considered as a sum of monochromatic waves through a normal Fourier expansion. Each of these Fourier components propagates with a different phase velocity if the optical medium is dispersive, since the refractive index n depends on the optical frequency ν. This phenomenon is linear and causal and gives rise to a signal distortion that may be properly described by a transfer function in the frequency domain.

Let $n(\nu)$ be the frequency-dependent refractive index of the propagation medium. When the optical wave propagates in a waveguiding structure the effective wavenumber is properly described by a propagation constant β, that reads:

$$\beta(\nu) = \frac{2\pi\nu}{c_0} n(\nu) \qquad [1]$$

where c_0 is the vacuum light velocity. The propagation constant corresponds to an eigenvalue of the wave equation in the guiding structure and normally takes a different value for each solution or propagation mode. For free-space propagation this constant is simply equal to the wavenumber of the corresponding plane wave.

The complex electrical field $E(z,t)$ of a signal propagating in the z direction may be properly described by the following expression:

$$\begin{aligned} &E(z,t) = A(z,t)\,\mathrm{e}^{\mathrm{i}(2\pi\nu_0 t - \beta_0 z)} \\ &\quad \text{with} \quad \nu_0\text{: the central optical frequency} \\ &\quad \text{and} \quad \beta_0 = \beta(\nu_0) \end{aligned} \qquad [2]$$

where $A(z,t)$ represents the complex envelope of the signal, supposed to be slowly varying compared to the carrier term oscillating with the frequency ν_0. Consequently, the signal will spread over a narrow band around the central frequency ν_0 and the propagation constant β can be conveniently approximated by a limited expansion to the second order:

$$\beta(\nu) = \beta_0 + \left.\frac{\mathrm{d}\beta}{\mathrm{d}\nu}\right|_{\nu=\nu_0} (\nu - \nu_0) + \left.\frac{\mathrm{d}^2\beta}{\mathrm{d}\nu^2}\right|_{\nu=\nu_0} (\nu - \nu_0)^2 \qquad [3]$$

For a known signal $A(0,t)$ at the input of the propagation medium, the problem consists in determining the signal envelope $A(z,t)$ after propagation over a distance z. The linearity and the causality of the system make possible a description using a transfer function $H_z(\nu)$ such as:

$$\tilde{A}(z,\nu) = H_z(\nu)\,\tilde{A}(0,\nu) \qquad [4]$$

where $\tilde{A}(z,\nu)$ is the Fourier transform of $A(z,t)$.

To make the transfer function $H_z(\nu)$ explicit, let us assume that the signal corresponds to an arbitrary harmonic function:

$$A(0,t) = A_0\ \mathrm{e}^{\mathrm{i}2\pi ft} \qquad [5]$$

Since this function is arbitrary and the signal may always be expanded as a sum of harmonic functions through a Fourier expansion, there is no loss of generality. The envelope identified as a harmonic function actually corresponds to a monochromatic wave of optical frequency $\nu = \nu_0 + f$ as defined by eqn [2]. Such a monochromatic wave will experience the following phase shift through propagation:

$$\begin{aligned} E(z,t) &= A_0\ \mathrm{e}^{\mathrm{i}[2\pi(\nu_0+f)t-\beta(\nu_0+f)z]} \\ &= A_0\ \mathrm{e}^{\mathrm{i}2\pi ft}\mathrm{e}^{\mathrm{i}[2\pi\nu_0 t-\beta(\nu_0+f)z]} \end{aligned} \qquad [6]$$

On the other hand, an equivalent expression may be found using the linear system described by eqns [2] and [4]:

$$\begin{aligned} E(z,t) &= A(z,t)\ \mathrm{e}^{\mathrm{i}(2\pi\nu_0 t-\beta_0 z)} \\ &= A_z\ \mathrm{e}^{\mathrm{i}2\pi ft}\ \mathrm{e}^{\mathrm{i}(2\pi\nu_0 t-\beta_0 z)} \end{aligned} \qquad [7]$$

Since eqns [6] and [7] must represent the same quantities and using the definition in eqn [4], a simple comparison shows that the transfer function must take the following form:

$$H_z(\nu) = \mathrm{e}^{-\mathrm{i}[\beta(\nu)z-\beta_0 z]} \qquad [8]$$

The transfer function takes a more analytical form using the approximation in eqn [3]:

$$H_z(\nu) = \mathrm{e}^{-\mathrm{i}2\pi(\nu-\nu_0)\tau_\mathrm{D}}\ \mathrm{e}^{-\mathrm{i}\pi D_\nu(\nu-\nu_0)^2 z} \qquad [9]$$

In the transfer function interpretation the first term represents a delay term. It means that the signal is delayed after propagation by the quantity:

$$\tau_\mathrm{D} = \frac{1}{2\pi}\frac{\mathrm{d}\beta}{\mathrm{d}\nu_0}z = \frac{z}{V_\mathrm{g}} \qquad [10]$$

where V_g represents the signal group velocity. This term therefore brings no distortion for the signal and thus states that the signal is replicated at the distance z with a delay τ_D.

The second term is the distortion term which is similar in form to a diffusion process and normally results in time spreading of the signal. In the case of a light pulse it will gradually broaden while propagating along the fiber, like a hot spot on a plate gradually spreading as a result of heat diffusion. The effect of this distortion is proportional to the distance z and to the coefficient D_ν, named group velocity dispersion (GVD):

$$D_\nu = \frac{1}{2\pi}\frac{\mathrm{d}^2\beta}{\mathrm{d}\nu^2} = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{1}{V_\mathrm{g}}\right) \qquad [11]$$

It is important to point out that the GVD may be either positive (normal) or negative (anomalous) and the distortion term in the transfer function in eqn [9] may be exactly cancelled by propagating in a medium with D_ν of opposite sign. It means that the distortion resulting from chromatic dispersion is reversible and this is widely used in optical links through the insertion of dispersion compensators. These are elements made of specially designed fibers or fiber Bragg gratings showing an enhanced GVD coefficient, with a sign opposite to the GVD in the fiber.

From the transfer function in eqn [9] it is possible to calculate the impulse response of the dispersive medium:

$$h_z(t) = \frac{1}{\sqrt{i|D_\nu|z}}\,\mathrm{e}^{\mathrm{i}\pi\frac{(t-\tau_\mathrm{D})^2}{D_\nu z}} \qquad [12]$$

so that the distortion of the signal may be calculated by a simple convolution of the impulse response with the signal envelope in the time domain.

The effect of dispersion on the signal can be more easily interpreted by evaluating the dispersive propagation of a Gaussian pulse. In this particular case the calculation of the resulting envelope can be carried out analytically. If the signal envelope takes the following Gaussian distribution at the origin:

$$A(0,t) = A_0\ \mathrm{e}^{-\frac{t^2}{\tau_0^2}} \qquad [13]$$

with τ_0 the $1/e$ half-width of the pulse, the envelope at distance z is obtained by convoluting the initial envelope with the impulse response $h_z(t)$:

$$\begin{aligned} A(z,t) &= h_z(t)\otimes A(0,t) \\ &= A_0\sqrt{\frac{\mathrm{i}z_0}{z+\mathrm{i}z_0}}\,\mathrm{e}^{\mathrm{i}\pi\frac{(t-\tau_\mathrm{D})^2}{D_\nu(z+\mathrm{i}z_0)}} \end{aligned} \qquad [14]$$

where

$$z_0 = -\frac{\pi\tau_0^2}{D_\nu} \qquad [15]$$

represents the typical dispersion length, that is the distance necessary to make the dispersion effect noticeable.

The actual pulse spreading resulting from dispersion can be evaluated by calculating the intensity of the envelope at distance z:

$$|A(z,t)|^2 = A_0 \frac{\tau_0}{\tau(z)} e^{-\frac{2(t-\tau_D)^2}{\tau^2(z)}} \qquad [16]$$

that is still a Gaussian distribution centered about the propagation delay time τ_D, with $1/e^2$ half-width:

$$\tau(z) = \tau_0\sqrt{1+(z/z_0)^2} \qquad [17]$$

The variation of the pulse width $\tau(z)$ is presented in Figure 1 and clearly shows that the pulse spreading starts to be nonnegligible from the distance $z = z_0$. This gives a physical interpretation for the dispersion length z_0. It must be pointed out that there is a direct formal similarity between the pulse broadening of a Gaussian pulse in a dispersive medium and the spreading of a free-space Gaussian beam as a result of diffraction. Asymptotically for distances $z \gg z_0$, the pulsewidth increases linearly:

$$\tau(z) \simeq |D_\nu| \frac{z}{\pi\tau_0} \qquad [18]$$

It must be pointed out that the width increases proportionally to the dispersion D_ν, but also inversely proportionally to the initial width τ_0. This results from the larger spectral width corresponding to a narrower pulsewidth, giving rise to a stronger dispersive effect.

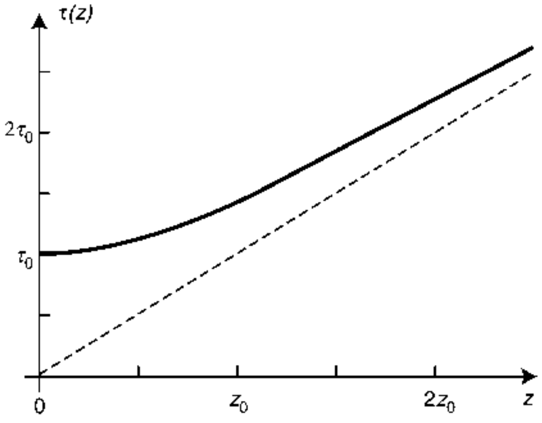

Figure 1 Variation of the $1/e^2$ half-width of a Gaussian pulse, showing the pulse spreading effect of dispersion. The dashed line shows the asymptotic linear spreading for large propagation distance.

The chromatic dispersion does not modify the spectrum of the transmitted light, as any linear effect. This can be straightforwardly demonstrated by evaluating the intensity spectrum of the signal envelope at any distance z, using the eqns [4] and [8]:

$$|\tilde{A}(z,\nu)|^2 = |H_z(\nu)\tilde{A}(0,\nu)|^2 = |e^{-i[\beta(\nu)z-\beta_0 z]}|^2|\tilde{A}(0,\nu)|^2 = |\tilde{A}(0,\nu)|^2 \qquad [19]$$

It means that the pulse characteristics in the time and frequency domains are no longer Fourier-transform limited, since after the broadening due to dispersion, the spectrum should normally spread over a narrower spectral width. This feature results from a re-arrangement of the spectral components within the pulse. This can be highlighted by evaluating the distribution of instantaneous frequency through the pulse. The instantaneous frequency ω_i is defined as the time-derivative of the wave phase factor $\phi(t)$ and is uniformly equal to the optical carrier pulsation $\omega_i = 2\pi\nu_0$ for the initial pulse, as can be deduced from the phase factor at $z = 0$ by combining eqns [2] and [13]. This constant instantaneous frequency means that all spectral components are uniformly present within the pulse at the origin.

After propagation through the dispersive medium the phase factor $\phi(t)$ can be evaluated by combining eqns [2] and [14] and evaluating the argument of the resulting expression. The instantaneous frequency ω_i is obtained after a simple time derivative of $\phi(t)$ and reads:

$$\omega_i(t) = \omega_0 + \frac{2\pi z}{D_\nu(z^2+z_0^2)}(t-\tau_D) \qquad [20]$$

For $z > 0$ the instantaneous frequency varies linearly over the pulse, giving rise to a frequency chirp. The frequency components are re-arranged in the pulse, so that the lower frequency components are in the leading edge of the pulse for a normal dispersion $D_\nu > 0$ and the higher frequencies in the trailing edge. For an anomalous dispersion $D_\nu < 0$, the arrangement is opposite, as can be seen in Figure 2. The effect of this frequency chirp can be visualized in Figure 3, showing that the effect of dispersion is equivalent to a frequency modulation over the pulse. The chirp is maximal at position z_0 and the pulsewidth takes its minimal value τ_0 when the chirp is zero. It is evident with this description that the pulse broadening may be entirely compensated through propagation in a medium of opposite

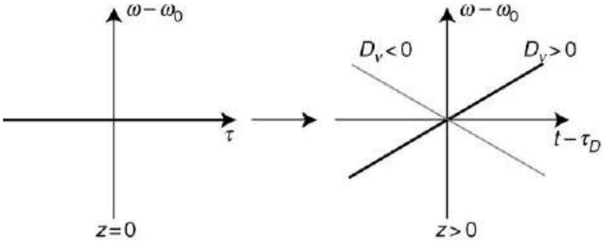

Figure 2 Distribution of the instantaneous frequency through a Gaussian pulse. At the origin the distribution is uniform (left) and the dispersion induces a frequency chirp that depends on the sign of the GVD coefficient D_ν.

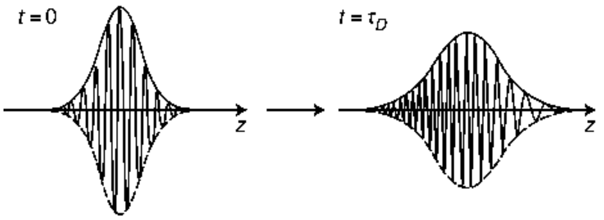

Figure 3 The dispersion results in a pulse broadening together with a frequency chirp, here for a normal GVD, that can be seen like a frequency modulation.

group velocity dispersion; this is equivalent to reversing the time direction in Figure 3. Moreover a pre-chirped pulse can be compressed to its Fourier-transform limited value after propagation in a medium with the proper dispersion sign. This feature is widely used for pulse compression after pre-chirping through propagation in a medium subject to optical Kerr effect.

The above description of the propagation of a Gaussian pulse implicitly states that the light source is perfectly coherent. In the early stages of optical communications it was not at all the case, since most sources were either light emitting diodes or multimode lasers. In this case the spectral extent of the signal was much larger than actually required for the strict need of modulation. Each spectral component may thus be considered as independently propagating the signal and the total optical wave can be identified to light merged from discrete sources emitting simultaneously the same signal at a different optical frequency. The group velocity dispersion will cause a delay $\delta\tau$ between different spectral components separated by a frequency interval $\delta\nu$ that can be simply evaluated by a first-order approximation:

$$\delta\tau = \frac{\mathrm{d}\tau_\mathrm{D}}{\mathrm{d}\nu}\delta\nu = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{z}{V_\mathrm{g}}\right)\delta\nu = D_\nu z\delta\nu \qquad [21]$$

where eqns [10] and [11] have been used. The delay $\delta\tau$ is thus proportional to the GVD coefficient D_ν, to the propagation distance z and the frequency separation $\delta\nu$. This description can be extended to a continuous frequency distribution with a spectral width σ_ν, resulting to the following temporal broadening σ_τ:

$$\sigma_\tau = |D_\nu|\sigma_\nu z \qquad [22]$$

Traditionally the spectral characteristics of a source are given in units of wavelength and the GVD coefficient is expressed in optical fibers accordingly. Following the same description as above, the temporal broadening σ_τ, for a spectral width σ_λ in wavelength units, reads:

$$\sigma_\tau = |D_\lambda|\sigma_\lambda z \qquad [23]$$

Since equal spectral widths must give equal broadening the value of the GVD in units of wavelength can be deduced from the natural definition in frequency units:

$$D_\lambda = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{V_\mathrm{g}}\right) = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{1}{V_\mathrm{g}}\right)\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = -\frac{c_0}{\lambda^2}D_\nu \qquad [24]$$

It must be pointed out that the coefficient D_λ takes a sign opposite to D_ν; in other words, a normal dispersion corresponds to a negative GVD coefficient D_λ. It is usually expressed in units of picoseconds of temporal broadening, per nanometer of spectral width and per kilometer of propagating distance, or ps/nm km. For example, a pulse showing a spectral width of 1 nm propagating through a 100 km fiber having a dispersion D_λ of $+10$ ps/nm km, will experience, according to eqn [23], a pulse broadening σ_τ of $10 \times 1 \times 100 = 1000$ ps or 1 ns.

Material Group Velocity Dispersion

Any dense material shows a variation of its index of refraction n as a function of the optical frequency ν. This natural property is called material dispersion and is the dominant contribution in weakly guiding structures such as standard optical fibers. This natural dependence results from the noninstantaneous response of the medium to the presence of the electric field of the optical wave. In other words, the polarization field $P(t)$ corresponding to the material response will vary with some delay or inertia to the change of the incident electric field $E(t)$. This delay between cause and effect generates a memory-type response of the medium that may be described using a time-dependent medium susceptibility $\chi(t)$. The relation between medium polarization at time t and incident field results from the weighted superposition of the effects of $E(t')$ at all previous times $t' < t$. This takes the form of the following convolution:

$$P(t) = \varepsilon_0 \int_{-\infty}^{+\infty} \chi(t - t')E(t')\mathrm{d}t' \qquad [25]$$

Through application of a simple Fourier transform this relation reads in the frequency domain as:

$$P(\nu) = \varepsilon_0 \chi(\nu) E(\nu) \qquad [26]$$

showing clearly that the noninstantaneous response of the medium results in a frequency-dependent refractive index using the standard relationship with the susceptibility χ:

$$n(\nu) = \sqrt{1 + \chi(\nu)} \quad \text{where} \quad \chi(\nu) = \mathrm{FT}\{\chi(t)\} \qquad [27]$$

This means that a beautiful natural phenomenon such as a rainbow, originates on a microscopic scale from the sluggishness of the medium molecules to react to the presence of light. For signal propagation, it results in a distortion of the signal and in most cases in a pulse spreading, but the microscopic causes are in essence identical. This noninstantaneous response is tightly related to the molecules' vibrations that also give rise to light absorption. For this reason it is convenient to express the propagation in an absorptive medium by adding an imaginary part to the susceptibility $\chi(\nu)$:

$$\chi(\nu) = \chi'(\nu) + \mathrm{i}\chi''(\nu) \qquad [28]$$

so that the refractive index $n(\nu)$ and the absorption coefficient $\alpha(\nu)$ reads in a weakly absorbing medium:

$$\begin{aligned} n(\nu) &= \sqrt{1 + \chi'(\nu)} \\ \alpha(\nu) &= -\frac{2\pi\chi''(\nu)}{\lambda n(\nu)} \end{aligned} \qquad [29]$$

Since the response of the medium, given by the time-dependent susceptibility $\chi(t)$ in eqn [25], is real and causal, the real and imaginary part of the susceptibility in eqn [28] are not entirely independent and are related by the famous Kramers–Kronig relations:

$$\begin{aligned} \chi'(\nu) &= \frac{2}{\pi} \int_0^{\infty} \frac{s\chi''(\nu)}{s^2 - \nu^2} \mathrm{d}s \\ \chi''(\nu) &= \frac{2}{\pi} \int_0^{\infty} \frac{\nu\chi'(\nu)}{\nu^2 - s^2} \mathrm{d}s \end{aligned} \qquad [30]$$

Absorption and dispersion act in an interdependent way on the propagating optical wave and knowing either the absorption or the dispersion spectrum is theoretically sufficient to determine the entire optical response of the medium. The interdependence between absorption and dispersion is typically illustrated in Figure 4. The natural tendency is to observe a growing refractive index for increasing frequencies in low absorption regions. In this case, the dispersion is called normal and this is the most observed situation in transparency regions of a material, which obviously offer the largest interest for optical propagation. In an absorption line the tendency is opposite, a diminishing index for increasing wavelength, and such a response is called anomalous dispersion. The dispersion considered here is the phase velocity dispersion, represented by the slope of $n(\nu)$, that must not be mistaken with the group velocity dispersion (GVD) that only matters for signal distortion. The difference between these two quantities is clarified below.

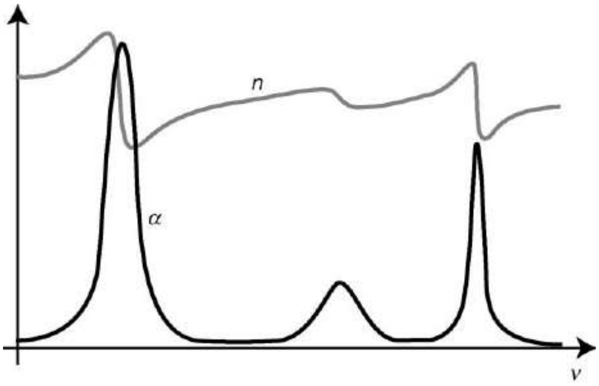

Figure 4 Typical optical response of a transparent medium, showing the spectral interdependence between the absorption coefficient α and the index of refraction n.

To demonstrate that a frequency-dependent refractive index $n(\nu)$ gives rise to a group velocity dispersion and thus a signal distortion we use eqns [1] and [10], so the group velocity V_g can be expressed:

$$V_g = \frac{c_0}{N} \quad \text{with} \quad N = n + \nu\frac{\mathrm{d}n}{\mathrm{d}\nu} = n - \lambda\frac{\mathrm{d}n}{\mathrm{d}\lambda} \qquad [31]$$

N is called group velocity index and differs from the phase refractive index n only if n shows a spectral dependence. In a region of normal dispersion ($\mathrm{d}n/\mathrm{d}\nu) > 0$, the group index is larger than the phase index and this is the situation observed in the great majority of transparent materials. From eqn [31] and using eqn [24] the GVD coefficient D_λ can be expressed as a function of the refractive index $n(\lambda)$:

$$D_\lambda = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{V_g}\right) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{N}{c_0}\right) = -\frac{\lambda}{c_0}\frac{\mathrm{d}^2 n}{\mathrm{d}\lambda^2} \qquad [32]$$

The GVD coefficient is proportional to the second derivative of the refractive index n with respect to wavelength and is therefore minimal close to a point of inflexion of $n(\lambda)$. As can be seen in Figure 4 such a point of inflexion is always present where absorption is minimal, at the largest spectral distance from two absorption lines. It means that the conditions of low-group velocity dispersion and high transparency are normally fulfilled in the same spectral region of an optical dielectric material. In this region the phase velocity dispersion is normal at any wavelength, but the group velocity is first normal for shorter wavelengths, then is zero at a definite wavelength corresponding to the point of inflexion of $n(\lambda)$, and finally becomes anomalous for longer wavelengths. The situation in which normal phase dispersion and anomalous group dispersion are observed simultaneously is in no way exceptional.

In pure silica the zero GVD wavelength is at 1273 nm, but is subject to be moderately shifted to larger wavelengths in optical fibers, as a result of the presence of doping species to raise the index in the fiber guiding core. This shift normally never exceeds 10 nm using standard dopings; larger shifts are observed resulting from waveguide dispersion and this aspect will be addressed in the next section. The zero GVD wavelength does not strictly correspond to the minimum attenuation in silica fibers, because the dominant source of loss is Rayleigh scattering in this spectral region and not molecular absorption. This scattering results from fluctuations of the medium density as observed in any amorphous materials such as vitreous silica and is therefore a collective effect of many molecules that does not impinge on the microscopic susceptibility $\chi(t)$. It has therefore no influence on the material dispersion characteristics and this explains the reason for the minimal attenuation wavelength at 1550 nm, mismatching and quite distant from the zero material GVD at 1273 nm.

Material GVD in amorphous SiO_2 can be accurately described using a three-term Sellmeier expansion of the refractive index:

$$n(\lambda) = \sqrt{1 + \sum_{j=1}^{3} \frac{C_j \lambda^2}{\lambda^2 - \lambda_j^2}} \qquad [33]$$

and performing twice the wavelength derivative. The coefficients C_j and λ_j are found in most reference handbooks and result in the GVD spectrum shown in Figure 5. Such a spectrum explains the absence of interest for propagation in the visible region through optical fibers, the dispersion being very important in this spectral region. It also explains the large development of optical fibers in the 1300 nm region as a consequence of the minimal material dispersion there. It must be pointed out that it is quite easy to set up propagation in an anomalous GVD regime in optical fibers, since this regime is observed in the lowest attenuation spectral region. Anomalous dispersion makes possible interesting propagation features when combined with a third-order nonlinearity as the optical Kerr effect, namely soliton propagation and efficient spectral broadening through modulation instability.

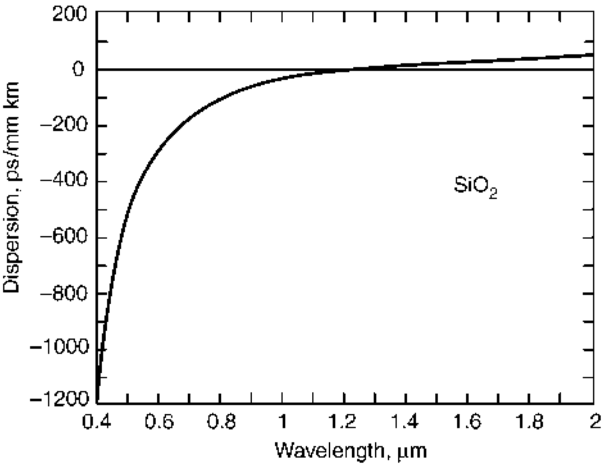

Figure 5 Material dispersion of pure silica. The visible region (0.4–0.7 µm) shows a strong normal GVD that decreases when moving into the infrared and eventually vanishes at 1273 nm. In the minimum attenuation window (1550 nm) the material GVD is anomalous.

Waveguide Group Velocity Dispersion

Solutions of the wave equation in an optical dielectric waveguide such as an optical fiber are discrete and limited. These solutions, called modes, are characterized by an unchanged field distribution along the waveguides and by a uniform propagation constant β over the wavefront. This last feature is particularly important if one recall that the field extends over regions presenting different refractive indices in a dielectric waveguide. For a given mode, the propagation constant β defines an effective refractive index n_{eff} for the propagation by similarity to eqn [1]:

$$\beta = \frac{2\pi\nu}{c_0} n_{\text{eff}} \qquad [34]$$

The value of this effective refractive index n_{eff} is always bound by the value of the core index n_1 and of the cladding index n_2, so that $n_2 < n_{\text{eff}} < n_1$. For a given mode, the propagation constant β, and so the effective refractive index n_{eff}, only depend on a quantity called normalized frequency V that essentially scales the light frequency to the waveguide optical parameters:

$$V = \frac{2\pi}{\lambda} a\sqrt{n_1^2 - n_2^2} \qquad [35]$$

where a is the core radius. Figure 6 shows the dependence of the propagation constant β of the fundamental mode LP_{01} on the normalized frequency V. The variation is nonnegligible in the single-mode region and gives rise to a chromatic dispersion, since V depends on the wavelength λ, even in the fictitious case of dispersion-free refractive indices in the core and the cladding materials. This type of chromatic dispersion is called waveguide dispersion.

To find an expression for the waveguide dispersion in function of the guiding properties, let us define another normalized parameter, the normalized phase constant b, such as:

$$\beta = \frac{2\pi}{\lambda}\sqrt{n_2^2 + b(n_1^2 - n_2^2)} \qquad [36]$$

The parameter b takes values in the interval $0 < b < 1$, is equal to 0 when $n_{\text{eff}} = n_2$ at the mode cutoff, and is equal to 1 when $n_{\text{eff}} = n_1$. This latter situation is never observed and is only asymptotic for very large normalized frequencies V. Solving the wave equation provides the dispersion relation $b(V)$ and it is important to point out that this relation between normalized quantities depends only on the shape of the refractive index profile. Step-index, triangular or multiple-clad index profiles will result in different $b(V)$ relations, independently of the actual values of the refractive indices n_1 and n_2 and of the core radius a.

From the definitions in eqns [10] and [35] and in the fictitious case of an absence of material dispersion, the propagation delay per unit length reads:

$$\frac{1}{V_g} = \frac{1}{2\pi}\frac{d\beta}{d\nu} = \frac{1}{2\pi}\frac{d\beta}{dV}\frac{dV}{d\nu} = \frac{\lambda}{2\pi}V\frac{d\beta}{dV} \qquad [37]$$

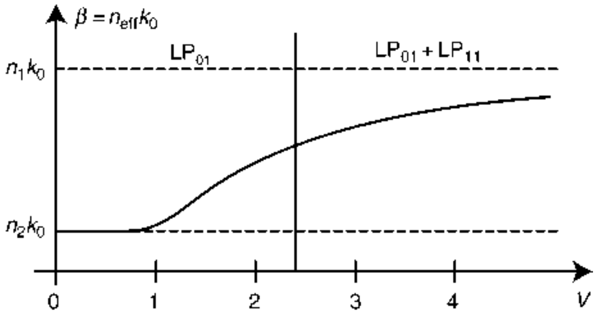

Figure 6 Propagation constant β as a function of the normalized frequency V in a step-index optical fiber. The single mode region is in the range $0 < V < 2.405$.

so that the waveguide group velocity dispersion can be expressed using the relation in eqn [24]:

$$\begin{aligned} D_\lambda^{\text{w}} &= \frac{d}{d\lambda}\left(\frac{1}{V_g}\right) = \frac{d}{d\nu}\left(\frac{1}{V_g}\right)\frac{d\nu}{d\lambda} \\ &= -\frac{\nu}{\lambda}\frac{d}{d\nu}\left(\frac{1}{V_g}\right) = -\frac{1}{2\pi c_0}V^2\frac{d^2\beta}{dV^2} \qquad [38] \end{aligned}$$

The calculation of the combined effect of material and waveguide dispersions results in very long expressions in which it is difficult to highlight the relative effect of each contribution. Nevertheless, by making the assumption of weak guidance:

$$\frac{n_1 - n_2}{n_2} \simeq \frac{N_1 - N_2}{N_2} \ll 1 \qquad [39]$$

where N_1 and N_2 are the group indices in the core and the cladding, respectively, defined in eqn [31], the complete expression can be drastically simplified to obtain for the delay per unit length:

$$\frac{1}{V_g} = \frac{1}{c_0}\left[N_2 + (N_1 - N_2)\frac{d(bV)}{dV}\right] \qquad [40]$$

and for the total dispersion:

$$\begin{aligned} D_\lambda = {} & D_2 + (D_1 - D_2)\frac{d(bV)}{dV} \\ & - (N_1 - N_2)\frac{N_2}{n_1}\frac{1}{\lambda c_0}V\frac{d^2(bV)}{dV^2} \qquad [41] \end{aligned}$$

where

$$D_1 = -\frac{\lambda}{c_0}\frac{d^2 n_1}{d\lambda^2} \quad \text{and} \quad D_2 = -\frac{\lambda}{c_0}\frac{d^2 n_2}{d\lambda^2} \qquad [42]$$

are the material GVD in the core and the cladding, respectively.

The first two terms in eqn [41] represent the contribution of material dispersion weighted by the relative importance of core and cladding materials for the propagating mode. In optical fibers, the difference between D_1 and D_2 is small, so that this contribution can be often well approximated by D_2, independently of any guiding effects.

The last term represents the waveguide dispersion and is scaled by 2 factors:

- *The core-cladding index difference* $n_1 - n_2 \simeq N_1 - N_2$. The waveguide dispersion will be significantly enhanced by increasing the index difference between core and cladding.
- *The shape factor* $V(d^2(bV)/dV^2)$. This factor uniquely depends on the shape of the refractive index profile and may substantially modify the

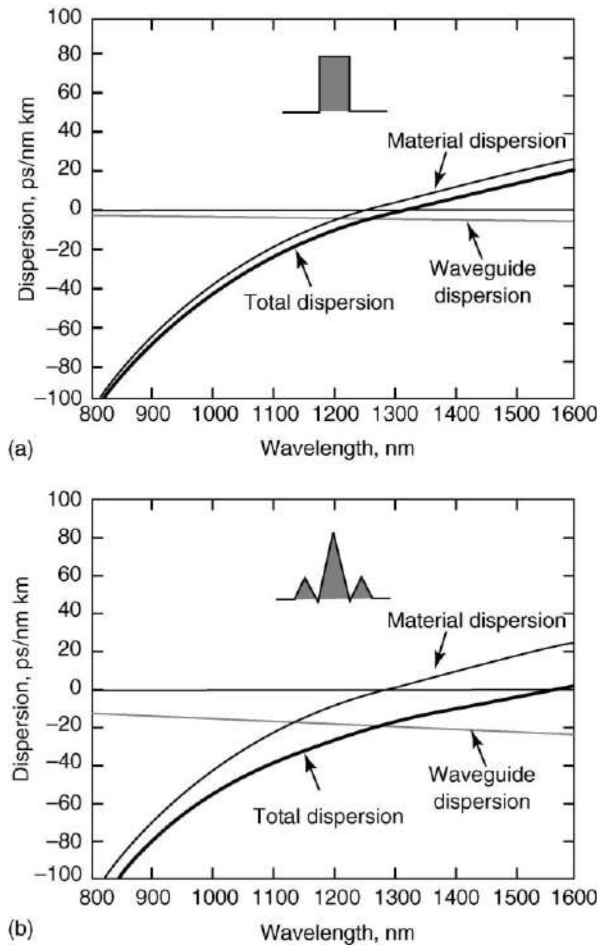

Figure 7 Material, waveguide and total group velocity dispersions for: (a) step-index fiber; (b) triangular profile fiber. The waveguide dispersion can be significantly enhanced by changing the shape of the index profile, making possible a shift of the zero GVD to the minimum attenuation window.

spectral dependence of the waveguide dispersion, making possible a great variety of dispersion characteristics.

Due to this degree of freedom brought by waveguide dispersion it is possible to shift the zero GVD wavelength to the region of minimal attenuation at 1550 nm in silica optical fibers. Figure 7a shows the total group velocity dispersion of a step-index core fiber, together with the separate contributions of material and waveguide GVD. In this case, the material GVD is clearly the dominating contribution, while the small waveguide GVD results in a shift of the zero GVD wavelength from 1273 nm to 1310 nm. A larger shift could be obtained by increasing the core-cladding index difference, but this also gives rise to an increased attenuation from the doping and no real benefit can be expected from such a modification. In Figure 7b, the core shows a triangular index profile to enhance the shape factor in eqn [41], so that the contribution of waveguide GVD is significantly increased with no impairing attenuation due to excessive doping. This makes it possible to realize the ideal situation of an optical fiber showing a zero GVD at the wavelength of minimum attenuation. These dispersion-shifted fibers (DSF) have now become successful in modern telecommunication networks.

Nevertheless, the absence of dispersion favors the efficiency of nonlinear effects and several classes of fibers are now proposed, showing a small but nonzero GVD at 1550 nm, with positive or negative sign, nonzero DSF (NZDSF). By interleaving fibers with positive and negative GVDs it is possible to propagate along the optical link in a dispersive medium and thus minimize the impact of nonlinearities, while maintaining the overall GVD of the link close to zero and canceling any pulse spreading accordingly.

List of Units and Nomenclature

Chromatic dispersion [ps nm^{-1} km^{-1}]	D_λ
Electric field [V m^{-1}]	E
Group index	N
Group velocity [m s^{-1}]	V_g
Group Velocity Dispersion GVD [s^2 m^{-1}]	D_ν
Linear attenuation coefficient [m^{-1}]	α
Medium susceptibility	χ
Normalized frequency	V
Optical frequency [s^{-1}]	ν
Propagation constant [m^{-1}]	β
Propagation delay [s]	τ_D
Polarization density field [A s m^{-2}]	P
Refractive index	n
Vacuum light velocity [m s^{-1}]	c_0
Wavelength [m]	λ

See also

Dispersion Management. **Fiber and Guided Wave Optics:** Overview. **Fourier Optics**. **Nonlinear Optics, Basics:** Kramers–Kronig Relations in Nonlinear Optics.

Further Reading

Ainslie BJ and Day CR (1986) A review of single-mode fibers with modified dispersion characteristics. *IEEE Journal of Lightwave Technology* LT-4(8): 967–979.

Bass M (ed.) (1994) *Handbook of Optics*, sponsored by the Optical Society of America, 2nd edn, vol. II, chap. 10 & 33. New York: McGraw-Hill.

Buck JA (1995) In: Goodman JW (ed.) *Fundamentals of Optical Fibers*, chap. 1, 5, 6. New York: Wiley Series in Pure and Applied Optics.

Gloge D (1971) Dispersion in weakly guiding fibers. *Applied Optics* 10: 2442–2445.
Jeunhomme LB (1989) *Single Mode Fiber Optics: Principles and Applications*. Optical Engineering Series, No. 23. New York: Marcel Dekker.
Marcuse D (1974) *Theory of Dielectric Optical Waveguides*. Series on Quantum Electronics – Principles and Applications, chap. 2. New York: Academic Press.
Marcuse D (1980) Pulse distortion in single-mode fibers. *Applied Optics* 19: 1653–1660.
Marcuse D (1981) Pulse distortion in single-mode fibers. *Applied Optics* 20: 2962–2974.
Marcuse D (1981) Pulse distortion in single-mode fibers. *Applied Optics* 20: 3573–3579.
Mazurin OV, Streltsina MV and Shvaiko-Shvaikovskaya TP (1983) *Silica Glass and Binary Silicate Glasses (Handbook of Glass Data; Part A)* Series on Physical Sciences Data. Amsterdam: Elseviervol. 15, pp. 63–75.
Murata H (1988) *Handbook of Optical Fibers and Cables*. Optical Engineering Series, No. 15, chap. 2. New York: Marcel Dekker.
Saleh BEA and Teich MC (1991) In: Goodman JW (ed.) *Fundamentals of Photonics*, chap. 5, 8 & 22. New York: Wiley Series in Pure and Applied Optics.

Fabrication of Optical Fiber

D Hewak, University of Southampton,
Southampton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The drawing of optical fibers from silica preforms has, over a short period of time, progressed from the laboratory to become a manufacturing process capable of producing millions of kilometers of telecommunications fiber a year. Modern optical fiber fabrication processes produce low-cost fiber of excellent quality, with transmission losses close to their intrinsic loss limit. Today, fiber with transmission losses of 0.2 dB per kilometer of fiber are routinely drawn through a two-stage process that has been refined since the 1970s.

Although fibers of glass have been fabricated and used for hundreds of years, it was not until 1966 that serious interest in the use of optical fibers for communication emerged. At this time, it was estimated that the optical transmission loss in bulk glass could be as low as 20 dB km^{-1} if impurities were sufficiently reduced, a level at which practical applications were possible. At this time, no adequate fabrication techniques were available to synthesize glass of high purity, and fiber-drawing methods were crude.

Over the next five years, efforts worldwide addressed the fabrication of low-loss fiber. In 1970, a fiber with a loss of 20 dB km^{-1} was achieved. The fiber consisted of a titania doped core and pure silica cladding. This result generated much excitement and a number of laboratories worldwide actively began researching optical fiber. New fabrication techniques were introduced, and by 1986, fiber loss had been reduced close to the theoretical limit.

All telecommunications fiber that is fabricated today is made of silica glass, the most suitable material for low-loss fibers. Early fiber research studied multicomponent glasses, which are perhaps more familiar optical materials; however, low-loss fiber, could not be realized, partly due to the lack of a suitable fabrication method. Today other glasses, in particular the fluorides and sulfides, continue to be developed for speciality fiber applications, but silica fiber dominates in most applications.

Silica is a glass of simple chemical structure containing only two elements, silicon and oxygen. It has a softening temperature of about 2,000°C at which it can be stretched, i.e. drawn into fiber. An optical fiber consists of a high purity silica glass core, doped with suitable oxide materials to raise its refractive index (Figure 1). This core, typically on the order of 2–10 microns in diameter, is surrounded by silica glass of lower refractive index. This cladding layer extends the diameter to typically 125 microns. Finally a protective coating covers the entire structure. It is the phenomenon of total internal reflection at the core cladding interface that confines light to the core and allows it to be guided. The basic requirements of an optical fiber are as follows:

1. The material used to form the core of the fiber must have a higher refractive index than the cladding material, to ensure the fiber is a guiding structure.

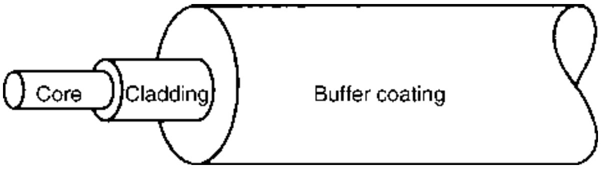

Figure 1 Structure of an optical fiber.

2. The materials used must be low loss, providing transmission with no absorption or scattering of light.
3. The materials used must have suitable thermal and mechanical properties to allow them to be drawn down in diameter into a fiber.

Silica (SiO_2) can be made into a glass relatively easily. It does not easily crystallize, which means that scattering from unwanted crystalline centers within the glass is negligible. This has been a key factor in the achievement of a low-loss fiber. Silica glass has high transparency in the visible and near-infrared wavelength regions and its refractive index can be easily modified. It is stable and inert, providing excellent chemical and mechanical durability. Moreover, the purification of the raw materials used to synthesize silica glass is quite straightforward.

In the first stage of achieving an optical fiber, silica glass is synthesized by one of three main chemical vapor processes. All use silicon tetrachloride ($SiCl_4$) as the main precursor, with various dopants to modify the properties of the glass. The precursors are reacted with oxygen to form the desired oxides. The end result is a high purity solid glass rod with the internal core and cladding structure of the desired fiber.

In the second stage, the rod, or preform as it is known, is heated to its softening temperature and stretched to diameters of the order of 125 microns. Tens to hundreds of kilometers of fiber are produced from a single preform, which is drawn continuously, with minimal diameter fluctuations. During the drawing process one or more protective coatings are applied, yielding long lengths of strong, low-loss fiber, ready for immediate application.

Preform Fabrication

The key step in preparing a low-loss optical fiber is to develop a technique for completely eliminating transition metal and OH ion contamination during the synthesis of the silica glass. The three methods most commonly used to fabricate a glass optical fiber preform are: the modified chemical vapor deposition process (MCVD); the outside vapor deposition process (OVD); and the vapor-axial deposition process (VAD).

In a typical vapor phase reaction, halide precursors undergo a high temperature oxidation or hydrolysis to form the desired oxides. The completed chemical reaction for the formation of silica glass is, for oxidation:

$$SiCl_4 + O_2 \rightarrow SiO_2 + 2Cl_2 \qquad [1]$$

For hydrolysis, which occurs when the deposition occurs in a hydrogen-containing flame, the reaction is:

$$SiCl_4 + 2H_2O \rightarrow SiO_2 + 4HCl \qquad [2]$$

These processes produce fine glass particles, spherical in shape with a size of the order of 1 nm. These glass particles, known as soot, are then deposited and subsequently sintered into a bulk transparent glass. The key to low-loss fiber is the difference in vapor pressures of the desired halides and the transition metal halides that cause significant absorption loss at the wavelengths of interest. The carrier gas picks up a pure vapor of, for example, $SiCl_4$, and any impurities are left behind.

Part of the process requires the formation of the desired core/cladding structure in the glass. In all cases, silica-based glass is produced with variations in refractive index produced by the incorporation of dopants. Typical dopants used are germania (GeO_2), titania (TiO_2), alumina (Al_2O_3), and phosphorous pentoxide (P_2O_5) for increasing the refractive index, and boron oxide (B_2O_2) and fluorine (F) for decreasing it. These dopants also allow other properties to be controlled, such as the thermal expansion of the glass and its softening temperatures. In addition, other materials, such as the rare earth elements, have also been used to fabricate active fibers that are used to produce optical fiber amplifiers and lasers.

MCVD Process

Optical fibers were first produced by the MCVD method in 1974, a breakthrough that completely solved the technical problems of low-loss fiber fabrication (Figure 2).

As shown schematically in Figure 3, the halide precursors are carried in the vapor phase by oxygen carrier gas into a pure silica substrate tube. An oxy-hydrogen burner traverses the length of the tube, which it heats externally. The tube is heated to temperatures of about 1,400°C which then oxidizes the halide vapor materials. The deposition temperature is sufficiently high to form a soot made of glassy particles which are deposited on the inside wall of the substrate tube but low enough to prevent the softened silica substrate tube from collapsing. The process usually takes place on a horizontal glass-working lathe.

During the deposition process, the repeated traversing of the burner forms multiple layers of soot. Changes to the precursors entering the tube and thus the resulting glass composition are introduced for layers which will form the cladding and then the core. The MCVD method allows germania to be doped into

Figure 2 Fabrication of an optical fiber preform by the MCVD method.

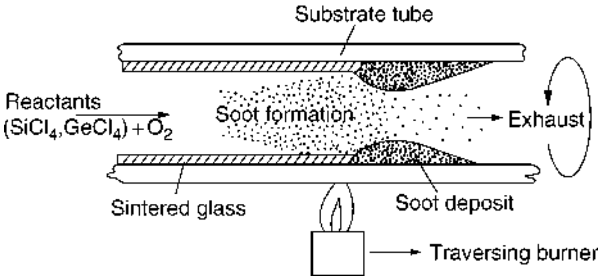

Figure 3 Schematic of preform fabrication by the MCVD method.

the silica glass and the precise control of the refractive index profile of the preform. When deposition is complete, the burner temperature is increased and the hollow, multilayered structure is collapsed to a solid rod. A characteristic of fiber formed by this process is a refractive index dip in the center of the core of the fiber. In addition, the deposition takes place in a closed system, which dramatically reduces contamination by OH^- ions and maintains low levels of other impurities. The high temperature allows high deposition rates compared to traditional chemical vapor deposition (CVD) and large performs, built up from hundreds of layers can be produced.

The MCVD method is still widely used today though it has some limitations, particularly on the preform size that can be achieved and thus the manufacturing cost. Diameter of the final preform is determined by the size of the initial silica substrate tube, which, due to the high purity required, accounts for a significant portion of the cost of the preform. A typical preform fabricated by MCVD yields about 5 km of fiber. Its success has spurred improvements to the process, in particular to address the fiber yield from a single preform.

OVD Process

The outside vapor deposition (OVD) methods, also known as outside vapor phase oxidation (OVPO), synthesize the fine glass particles within a burner flame. The precursors, oxygen, and fuel for the burner, are introduced directly into the flame. The soot is deposited onto a target rod that rotates and traverses in front of the burner. As in MCVD, the preform is built up layer by layer, though now initially by depositing the core glass and then building up the cladding layers over this. After the deposition process is complete, the preform is removed from the target rod and is collapsed and sintered into a transparent glass preform. The center hole remains, but disappears during the fiber drawing process. This technique has advantages in both size of preform which can be obtained and the fact that a high-quality silica substrate tube is no longer required. These two advantages combine to make a more economical process. From a single preform, several hundred kilometers of fiber can be produced.

VAD Process

The most recent refinement to the fabrication process was developed again to aid the mass production of high-quality fibers. In the VAD process, both core and cladding glasses are deposited simultaneously. Like OVD, the soot is synthesized and deposited by flame hydrolysis, as shown in Figure 4, the precursors are blown from a burner, oxidized, and deposited onto a silica target rod.

Burners for the VAD process consist of a series of concentrate nozzles. The first delivers an inert carrier and the main precursors $SiCl_4$, the second delivers an inert carrier glass and the dopants, the third delivers hydrogen fuel, and the fourth delivers oxygen. Gas flows are up to a liter per minute and deposition rates can be very high. With the VAD process, both core and cladding glasses can be deposited simultaneously. The main advantage is that this is a continuous process, as the soot which forms the core and cladding glasses are deposited axially onto the end of the rotating silica rod, it is slowly drawn upwards into a furnace which sinters and consolidates the soot into a transparent glass. The upward motion is such that the end at which deposition is occurring remains in a fixed position and essentially the preform is grown from this base.

The advantages of the VAD process are a preform without a central dip and, most importantly, the mass production associated with a continuous process. The glass quality produced is uniform and the resulting fibers have excellent reproducibility and low loss.

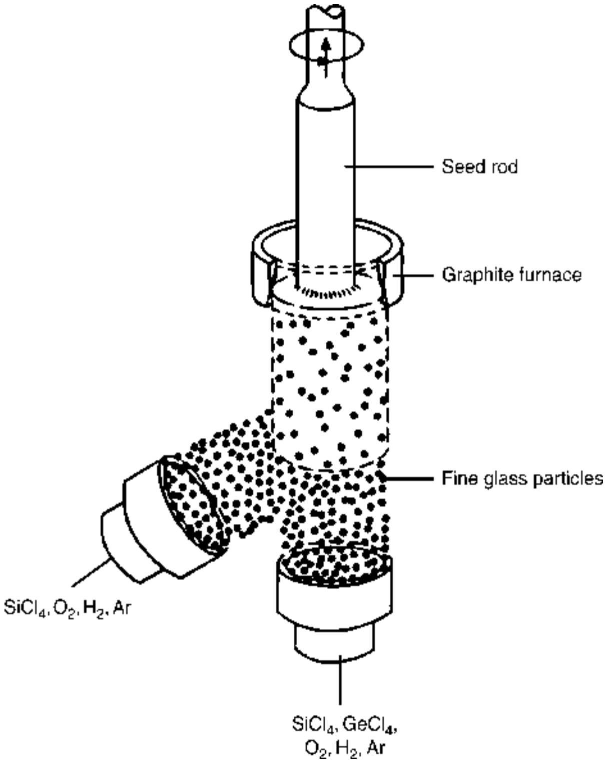

Figure 4 Schematic of preform fabrication by the VAD method.

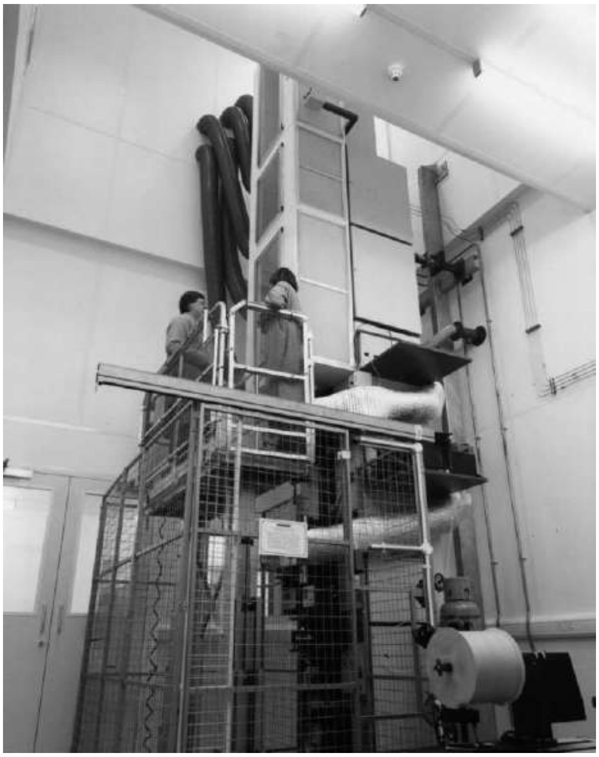

Figure 5 A commercial scale optical fiber drawing tower.

Other Methods of Preform Fabrication

There are other methods of preform fabrication, though these are not generally used for today's silica glass based fiber. Some methods, such as fabrication of silica through sol–gel chemistry could not provide the large low-loss preforms obtained by chemical vapor deposition techniques. Other methods, such as direct casting of molten glass into rods, or formation of rods and tubes by extrusion, are better suited to and find application in other glass chemistries and speciality fibers.

Fiber Drawing

Once a preform has been made, the second step is to draw the preform down in diameter on a fiber drawing tower. The basic components and configuration of the drawing tower have remained unchanged for many years, although furnace design has become more sophisticated and processes like coating have become automated. In addition, fiber drawing towers have increased in height to allow faster pulling speeds (Figure 5).

Essentially, the fiber drawing process takes place as follows. The preform is held in a chuck which is mounted on a precision feed assembly that lowers the preform into the drawing furnace at a speed which matches the volume of preform entering the furnace to the volume of fiber leaving the bottom of the furnace. Within the furnace, the fiber is drawn from the molten zone of glass down the tower to a capstan, which controls the fiber diameter, and then onto a drum. During the drawing process, immediately below the furnace, is a noncontact device that measures the diameter of the fiber. This information is fed back to the capstan which speeds up to reduce the diameter or slows down to increase the fiber diameter. In this way, a constant diameter fiber is produced. One or more coatings are also applied in-line to maintain the pristine surface quality as the fiber leaves the furnace and thus to maximize the fiber strength.

The schematic drawing in Figure 6 shows a fiber drawing tower and its key components. Draw towers are commercially available, ranging in height from 3 m to greater than 20 m, with the tower height increasing as the draw speed increases.

Typical drawing speeds are on the order of $1\ \mathrm{m\,s^{-1}}$. The increased height is needed to allow the fiber to cool sufficiently before entering the coating applicator, although forced air-cooling of the fiber is commonplace in a manufacturing environment.

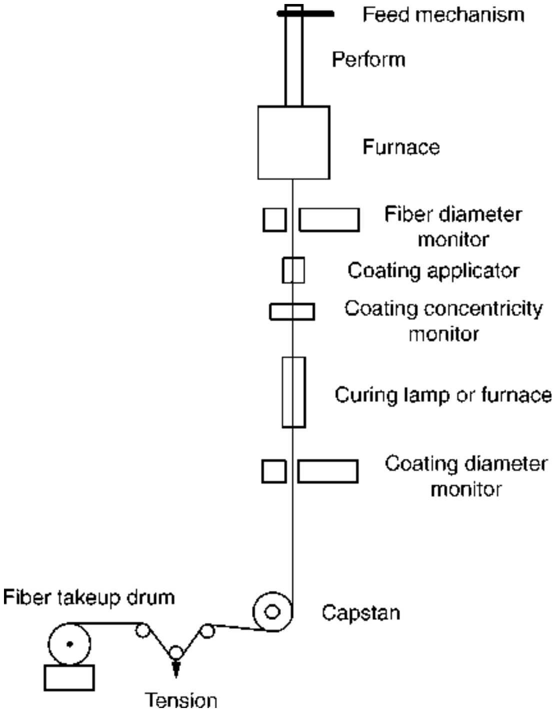

Figure 6 Schematic diagram of an optical fiber drawing tower and its components.

Furnaces

A resistance furnace is perhaps the most common and economical heating source used in a fiber drawing tower. A cylindrical furnace, with graphite or tungsten elements, provides heating to the preform by blackbody radiation. These elements must be surrounded by an inert gas, such as nitrogen or argon, to prevent oxidation. Gas flow and control is critical to prevent variations in the diameter of the fiber. In addition, the high temperature of the elements may lead to contamination of the preform surface that can then weaken the fiber.

An induction furnace provides precise clean heating through radio frequency (RF) energy that is inductively coupled to a zirconia susceptor ring. This type of furnace is cleaner than the graphite resistance type and is capable of continuous running for several months. It also has the advantage that it does not require a protective inert atmosphere and consequently the preform is drawn in a turbulence-free environment. These advantages result in high-strength fibers with good diameter control.

Diameter Measurement

A number of noncontact methods of diameter measurement can be applied to measure fiber diameter. These include laser scanning, interferometry, and light scattering techniques. To obtain precise control of the fiber diameter, the deviation between the desired diameter and the measured diameter is fed into the diameter control system. In order to cope with high drawing speeds, sampling rates as high as 1,000 times per second are used. The diameter control is strongly affected by the gas flow in the drawing furnace and is less affected by the furnace temperature variation. The furnace gas flow can be used to achieve suppression of fast diameter fluctuations. This is used in combination with drawing speed control to achieve suppression of both fast and slow diameter fluctuations. Current manufacturing processes are capable of producing several hundreds of kilometers of fiber with diameter variations of ± 1 micron.

There are two major sources of fiber diameter fluctuations: short-term fluctuations caused by temperature fluctuations in the furnace; and long-term fluctuations caused by variations in the outer diameter of the preform. Careful control of the furnace temperature, the length of the hot zone, and the flow of gas minimize the short-term fluctuations. Through optimization of these parameters, diameter errors of less than ± 0.5 microns can be realized. The long-term fluctuations in diameter are controlled by the feedback mechanism between the diameter measurement and the capstan.

Fiber Coating

A fiber coating is primarily used to preserve the strength of a newly drawn fiber and therefore must be applied immediately after the fiber leaves the furnace. The fiber coating apparatus is typically located below the diameter measurement gauge at a distance determined by the speed of fiber drawing, the tower height, and whether there is external cooling of the fiber. Coating is usually one of the limiting factors in the speed of fiber drawing. To minimize any damage or contamination of the pristine fiber leaving the furnace, this portion of the tower, from furnace to the application of the first coating, is enclosed in a clean, filtered-air chamber.

A wide variety of coating materials has been applied; however, conventional, commercial fiber generally relies on a UV curable polymer as the primary coating. Coating thickness is typically 50–100 microns. Subsequent coatings can then be applied for specific purposes. A dual coating is often used with an inner primary coating that is soft and an outer, secondary coating which is hard. This ratio of low to high elastic modulus can minimize stress on the fiber and reduce bending loss.

Alternative coatings include hermetic coatings of a low melting temperature metal, ceramics, or amorphous carbon. These can be applied in-line before the polymeric coating. Metallic coatings are applied by passing the fiber through a molten metal while ceramic or amorphous coatings utilize an in-line chemical vapor deposition reactor.

Types of Fiber

The majority of silica fiber drawn today is single mode. This structure consists of a core whose diameter is chosen such that, with a given refractive index difference between the core and cladding, only a single guide mode propagates at the wavelength of interest. With a discrete index difference between core and cladding, it is often referred to as a step index fiber. In typical telecommunications fiber, single mode operation is obtained with core diameters of 2–10 microns with a standard outer diameter of 125 microns.

Multimode fiber has core diameters considerably larger, typically 50, 62.5, 85, and 110 microns, again with a cladded diameter of 125 microns. Multimode fibers are often graded index, that is the refractive index is a maximum in the center of the fiber and smoothly decreases radially until the lower cladding index is reached. Multimode fibers find use in non-telecommunication applications, for example optical fiber sensing and medicine.

Single mode fibers, which are capable of maintaining a linear polarization input to the fiber, are known as polarization preserving fibers. The structure of these fibers provides a birefringence that removes the degeneracy of the two possible polarization modes. This birefringence is a small difference in the effective refractive index of the two polarization modes that can be guided and it is achieved in one of two ways. Common methods for the realization of this birefringence are an elliptical core in the fiber, or through stress rods, which modify the refractive index in one orientation. These fiber structures are shown in Figure 7.

While the loss minimum of silica-based fiber is near 1.55 microns, step index single-mode fiber offers zero dispersion close to 1.3 micron wavelengths and dispersion at the loss minimum is considerable. A modification of the structure of the fiber, and in particular a segmented refractive index profile in the core, can be used to shift this dispersion minimum to 1.55 microns. This fiber, illustrated in Figure 7 is known as dispersion shifted fiber. Similarly fibers, with a relatively low dispersion over a wide wavelength range, known as dispersion flattened fibers, can be obtained by the use of multiple cladding layers.

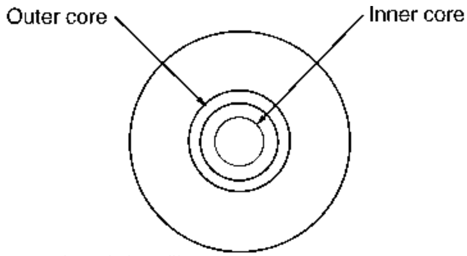

(a) Dispersion shifted fiber

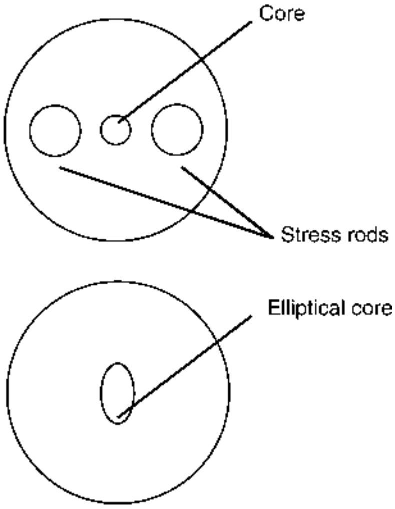

(b) Polarization preserving fiber

Figure 7 Structure of (a) dispersion shifted fiber and (b) two methods of achieving polarization preserving fiber.

List of Units and Nomenclature

Dopants	Elements or compounds added, usually in small amounts, to a glass composition to modify its properties.
Fiber drawing	The process of heating and thus softening an optical fiber preform and then drawing out a thin thread of glass.
Fiber loss	The transmission loss of light as it propagates through a fiber, usually measured in dB of loss per unit length of fiber. Loss can occur through the absorption of light in the core or scattering of light out of the core.
Glass	An amorphous solid formed by cooling from the liquid state to a rigid solid with no long range structure.
Modified chemical vapor deposition (MCVD)	A process for the fabrication of an optical preform where gases flow into the inside of a rotating tube, are heated and react

	to form particles of glass which are deposited onto the wall of the glass tube. After deposition, the glass particles are consolidated into a solid preform.
Outside vapor deposition (OVD)	A process for the fabrication of an optical preform where glass soot particles are formed in an oxy-hydrogen flame and deposited on a rotating rod. After deposition, the glass particles are consolidated into a solid preform.
Preform	A fiber blank, a bulk glass rod consisting of a core and cladding glass composite which is drawn into fiber.
Refractive index	A characteristic property of glass, which is defined by the speed of light in the material relative to the speed of light in a vacuum.
Silica	A transparent glass formed from silicon dioxide.
Vapor-axial deposition	A process similar to OVD, where the core and cladding layers are deposited simultaneously.

See also

Detection: Fiber Sensors. **Fiber and Guided Wave Optics:** Dispersion; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Optical Fiber Cables. **Fiber Gratings**. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Historical Development; Wavelength Division Multiplexing. **Optical Materials:** Optical Glasses.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*. New York: John Wiley & Sons.

Bass M and Van Stryland EV (eds) (2001) *Fiber Optics Handbook: Fiber, Devices, and Systems for Optical Communications*. New York: McGraw-Hill Professional Publishing.

Fujiura K, Kanamori T and Sudo S (1997) Fiber materials and fabrications. In: Sudo S (ed.) *Optical Fiber Amplifiers*, ch. 4, pp. 193–404. Boston, MA: Artech House.

Goff DR and Hansen KS (2002) *Fiber Optic Reference Guide: A Practical Guide to Communications Technology*, 3rd edn. Oxford, UK: Butterworth-Heinemann UK.

Hewak D (ed.) (1998) *Glass and Rare Earth Doped Glasses for Optical Fibres*. EMIS Datareview Series, INSPEC. London: The Institution of the Electrical Engineers.

Keck DB (1981) Optical fibre waveguides. In: Barnoski MK (ed.) *Fundamentals of Optical Fiber Communications*. New York: Academic Press.

Li T (ed.) (1985) *Optical Fiber Communications: Fiber Fabrication*. New York: Academic Press.

Personick SD (1985) *Fiber Optics: Technology and Applications*. New York: Plenum Press.

Schultz PC (1979) In: Bendow B and Mitra SS (eds) *Fiber Optics*. New York: Plenum Press.

Light Propagation

F G Omenetto, Los Alamos National Laboratory, Los Alamos, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Optical fibers have become a mainstay of our modern way of life. From phone calls to the internet, their presence is ubiquitous. The unmatched ability that optical fibers have to transmit light is an amazing asset that, besides being strongly part of our present, is poised to shape the future.

The foundation of the unique ability of optical fibers to transmit light hinges on the well-known physical concept of total internal reflection. This phenomenon is described by considering the behavior of light traveling from one material (A) to a different material (B). When light traverses the interface of the two materials with a specific angle, its direction of propagation is altered with respect to the normal to the surface, according to the well-known Snell law which relates the incident angle (onto the interface between A and B) to the refracted angle (away from the interface between A and B). This is written as $n_A \sin i = n_B \sin r$, where i is the angle of incidence, r the angle of refraction, and n_A and n_B are the refractive indices of the materials A and B. Under appropriate conditions, a critical value of the incidence angle (i_{CR}) exists for which the light does not propagate into B and is totally reflected back into A. This value is given by $i_{CR} = \arcsin(n_A/n_B)$. It follows that total internal reflection occurs when light is traveling from a medium with a higher index of refraction into a medium with a lower index of refraction. As an approximation, this is the basic idea behind light propagation in fibers.

Optical fibers are glass strands that include a concentric core with a cladding wrapped around it. The index of refraction of the core is slightly higher than the index of refraction of the cladding, thereby creating the necessary conditions for the confinement of light within them. Such index mismatches between the core and the cladding are determined by using different materials for each, by doping the glass matrix before pulling the fiber to create an index gradient, by introducing macroscopic defects such as air-holes in the structure of the fiber cladding (such as in the recently developed photonic crystal fibers), or by using materials other than glass, such as polymers or plastic.

The real physical picture of light guiding in fibers is more complex than the conceptual description stated above. The light traveling in the fiber is, more precisely, an electromagnetic wave with frequencies that lie in the optical range, and the optical fiber itself constitutes an electromagnetic waveguide. As such, the waveguide supports guided electromagnetic modes which can take on a multitude of shapes (i.e., of energy distribution profiles), depending on the core/cladding index of refraction profiles. These profiles are often complex and determine the physical boundary conditions that influence the field distribution within the fiber. The nature of the fiber modes has been treated extensively in the literature where an in-depth description of the physical solutions can be found.

There are many distinguishing features that identify different optical fiber types. Perhaps the most important one, in relation to the electromagnetic modes that the fiber geometry supports, is the distinction that is made between single-mode fibers and multimode fibers. As the name implies, a single-mode fiber is an optical fiber whose index of refraction profile between core and cladding is designed in such a way that there is only one electromagnetic field distribution that the fiber can support and transmit. That is to say, light is propagated very uniformly across the fiber. A very common commercially used fiber of this kind is the Corning SMF-28 fiber, which is employed in a variety of telecommunication applications. Multimode fibers have larger cores and are easier to couple light into, yet the fact that many modes are supported implies that there is less control over the behavior of light during its propagation, since it is more difficult to control each individual mode.

The condition for single-mode propagation in optical fibers is determined during the design and manufacture of the fiber by its geometrical parameters and is specific to a certain operation wavelength. A quantity V is defined as $V = (2\pi/\lambda) \times a(n_{\text{core}}^2 - n_{\text{cladding}}^2)^{1/2}$, where a is the core radius and λ is the wavelength of light. For example, a step-index fiber (i.e., a fiber where the transition from the index of refraction of the core to the index of refraction in the cladding is abrupt), supports a single electromagnetic mode if $V < 2.405$.

Light propagation in optical fibers, however, is affected by numerous other factors besides the electromagnetic modes that the fiber geometry allows.

First, the ability to efficiently transmit light in optical fiber is dependent on the optical transparency of the medium that the fiber is made of. The optical transparency of the medium is a function of the wavelength (λ) of light that needs to be transmitted through the fiber. Optical attenuation is particularly high for shorter wavelengths and the transmission losses for ultraviolet wavelengths become considerably higher than in the near infrared. The latter region, due to this favorable feature, is the preferred region of operation for optical fiber based telecommunication applications. This ultimate physical limit is determined by the process of light being scattered off the atoms that constitute the glass, a process called Rayleigh scattering. The losses due to Rayleigh scattering are inversely proportional to the fourth power of the wavelength (losses $\propto \lambda^{-4}$), and therefore become quite considerable as λ is decreased.

The ability of optical fibers to transmit light over a certain set of wavelengths, is quantified by their optical attenuation. This parameter is defined as the ratio of the output power versus the input power and is usually measured in units of decibels (dB), i.e., attenuation (dB) $= 10 \log_{10}(P_{\text{out}}/P_{\text{in}})$.

The values that are found in the literature relate optical attenuation to distance and express the attenuation per unit length (such as dB/km). Single-mode fibers used in telecommunications are usually manufactured with doped silica glasses designed to work in the near infrared, in a wavelength range between 1.3 and 1.6 μm ($1\ \mu\text{m} = 10^{-6}$ m) which provides the lowest levels of Rayleigh scattering and is located before the physical infrared absorption of silica (~1.65 μm) kicks in. Such fibers, which have been developed and refined for decades, have attenuation losses of less that 0.2 dB/km at a wavelength of 1.55 μm. Many other materials have been used to meet specific optical transmission needs. While silica remains by far the most popular material used for optical fiber manufacturing, other glasses are better suited for different portions of the spectrum: quartz glass can be particularly transmissive in the ultraviolet whereas plastic is sometimes conveniently used for short distance transmission in the visible.

The ability to collect light at the input of the fiber represents another crucial and defining parameter. Returning to total internal reflection, such an effect can only be achieved by sending light onto an interface at a specific angle. In fibers, the light needs to be coupled into a circular core and reach the core cladding interface at that specific geometrical angle. The measure of the ability to couple light in the fiber is referred to as 'coupling efficiency'. The core dimension defines the geometry of the coupling process which, in turn, determines the efficiency of the propagation of light inside the fiber. In order to be coupled efficiently and therefore guided within the fiber, the light has to be launched within a range of acceptable angles. Such a range is usually defined or derivable from the numerical aperture (NA) of the fiber which is approximately $\mathrm{NA} = (n_{\mathrm{core}}^2 - n_{\mathrm{cladding}}^2)^{1/2}$, or can be alternatively described as a function of the critical angle for total internal reflection within the fiber by $\mathrm{NA} = n_{\mathrm{core}} \sin i_{\mathrm{CR}}$. Also, light has to be focused into the core of the fiber and therefore the interplay between core size, focusability, and NA (i.e., acceptance angle) of the fiber influence the coupling efficiency into the fiber (which in turn affects transmission). Typical telecommunication fibers have a 10 μm core (a NA of ~0.2) and have very high coupling efficiencies (of the order of 70–80%). Specialized single mode fibers (such as photonic crystal fibers) can have core diameters down to a few microns, significantly lower coupling efficiencies, and high numerical apertures. Multimode fibers, in contrast, are easy to couple into given their large core size.

Other factors that contribute to losses during propagation are due to the physical path that the actual optical fiber follows. A typical optical fiber strand is rarely laid along a straight line. These types of losses are called bending losses and can be intuitively understood by thinking that when the fiber is straight, the light meets its total reflection condition at the core-cladding interface, but when the fiber bends, the interface is altered and varies the incidence conditions of light at the interface. If the bend is severe, light is no longer totally reflected and escapes from the fiber.

Another important issue that influences propagation arises when the light that travels through the fiber is sent in bursts or in pulses of light. Pulsed light is very important in fiber transmission because the light pulse is the carrier of optical information, most often representing a binary digit of data. The ability to transmit pulses reliably is at the heart of modern day telecommunication systems. The features that affect the transmission of pulses through optical fibers depend, again, largely on the structure and constituents of the fiber. One of the most important effects on pulsed light is dispersion. This phenomenon causes the broadening in time of the light pulses during their travel through the fiber. This can be easily explained by a Fourier analogy by thinking that a short pulse in time (such as the 10–20 picosecond pulse duration that is used for optical communications, $1\ \text{picosecond} = 10^{-12}$ seconds) has a large frequency content (i.e., is composed of a variety of 'colors'). Since the speed of light through bulk media, and therefore through the fiber constituent, depends on wavelength, different 'colors' will travel at different speeds, arriving at the end of the fiber slightly delayed with respect to one another. This causes a net pulse broadening and is dependent on the properties of the fiber that are depicted in the dispersion curve of the fiber. This type of dispersion is called chromatic dispersion and is dependent on the distance the pulse travels in the fiber. A measure of the pulse broadening can be calculated by multiplying the dispersion value for the fiber (measured in ps/km) at the wavelength of operation times and the distance in kilometers that the pulse travels.

There are other types of dispersion that affect pulses traveling through fibers: light with different polarization travels at a different speed, thereby causing an analogous problem to chromatic distortion. In multimode fibers, different modes travel with different properties. Dispersion issues are extremely important in the design of the next-generation telecommunication systems. As a push is made toward higher transmission capacity, pulses become shorter and these issues have to be dealt with carefully. Specifically, transmission of light in optical fibers becomes more complex because the interaction between the light and the fiber material constituents becomes nonlinear. As pulses become shorter, their energy is confined in a smaller temporal interval, making their peak power (which is equal to energy/time) increase. If the peak power becomes sufficiently high, the atoms that form the glass of which the fiber is made are excited in a nonlinear fashion adding effects that need to be addressed with caution. Energetic pulses at a certain wavelength change their shape during propagation and can become severely distorted or change wavelength (i.e., undergo a frequency conversion process), thereby destroying the information that they were meant to carry.

Several physical manifestations of optical nonlinearity affect pulsed light propagation, giving rise to new losses but also providing new avenues for efficient transmission.

Perhaps the most successful example of the positive outcome of nonlinear effects in fibers is the optical soliton. The latter is an extremely stable propagating pulse that is transmitted through the fiber undistorted, since it is a result of a peculiar balance between the linear distortion and nonlinear distortion that cancel each other by operating in a specific region of dispersion of the optical fiber (the anomalous dispersion region).

Nonlinear effects in fibers are an extremely rich area of study which, besides several drawbacks, carries immense opportunities if the nonlinearities can be controlled and exploited.

See also

Dispersion Management. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Solitons:** Optical Fiber Solitons, Physical Origin and Properties.

Further Reading

Hecht J (2001) *Understanding Fiber Optics*. New York: Prentice Hall.

Snyder AW and Love JD (1983) *Optical Waveguide Theory*. Boston: Kluwer Academic Press.

Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego: Academic Press.

Measuring Fiber Characteristics

A Girard, EXFO, Quebec, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The optical fiber is divided in two types: multimode and singlemode. Each type is used in different applications and wavelength ranges and is consequently characterized differently. Furthermore, the corresponding test methods also vary.

The optical fiber characteristics may be divided into four categories:

- The optical characteristics (transmission related);
- The dimensional characteristics;
- The mechanical characteristics; and
- The environmental characteristics.

These categories will be reviewed in the following sections, together with their corresponding test methods.

Fiber Optical Characteristics and Corresponding Tests Methods

The following sections will describe the following optical characteristics:

- attenuation;
- macrobending sensitivity;
- microbending sensitivity;
- cut-off wavelength;
- multimode fiber bandwidth;
- differential mode delay for multimode fibers;
- chromatic dispersion;
- polarization mode dispersion;
- polarization crosstalk; and
- nonlinear effects.

Attenuation

The spectral attenuation of an optical fiber follows exponential power decay from the power level at a cross-section 1 to the power level at cross-section 2, over a fiber length L, as follows:

$$P_2(\lambda) = P_1(\lambda) \cdot \mathrm{e}^{-\gamma(\lambda)L} \qquad [1]$$

$P_1(\lambda)$ is the optical power transmitted through the fiber core cross-section 1, expressed in mW; $P_2(\lambda)$ is the optical power transmitted through the fiber core cross-section 2 away from cross-section 1, expressed in mW; $\gamma(\lambda)$ is the spectral attenuation coefficient in linear units, expressed in km^{-1}; and L is the fiber length expressed in km.

Attenuation may be characterized at one or more specific wavelengths or as a function of wavelength. In the later case, attenuation is referred to spectral attenuation. Figure 1 illustrates such power decay.

Equation [1] may be expressed in relative units as follows:

$$\log_{10} P_2 = (\log_{10} P_1) \cdot - \gamma L \cdot \log_{10} e \qquad [2]$$

P is expressed in dBm units using the following definition.

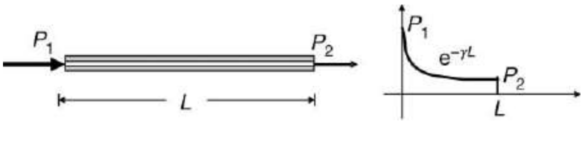

Figure 1 Power decay in an optical fiber due to attenuation.

The power in dBm is equal to 10 times the logarithm in base 10 of the power in mW; or

$$0\ \text{dBm} = 10\log_{10}(1\ \text{mW}) \qquad [3]$$

Then

$$\gamma(\lambda)[\text{km}^{-1}] = [(10\log_{10} P_1) - (10\log_{10} P_2)] \div L\log_{10} e \qquad [4]$$

$$\alpha(\lambda)[\text{dB/km}] = [P_1(\text{dBm}) - P_2(\text{dBm})]/L \qquad [5]$$

A new relative attenuation coefficient $\alpha(\lambda)$ with units of dB/km has been introduced, which is related to the linear coefficient $\gamma(\lambda)$ with units of km^{-1} as follows:

$$\alpha(\lambda) = (\log_{10} e)\cdot\gamma(\lambda) \approx 4{,}343\gamma(\lambda) \qquad [6]$$

$\alpha(\lambda)$ is the spectral attenuation coefficient of a fiber and is illustrated in Figure 2.

Test methods for attenuation

The attenuation may be measured by the following methods:

- cut-back method;
- backscattering method; and
- insertion loss method.

Cut-back method. The cut-back method is a direct application of the definition in which the power levels P_1 and P_2 are measured at two points of the fiber change of input conditions. P_2 is the power emerging from the far end of the fiber and P_1 is the power emerging from a point near the input after cutting the fiber.

The output power P_2 is recorded from the fiber under test (FUT) placed in the measurement setup. Keeping the launching conditions fixed, the FUT is cut to the cut-back length, for example 2 m from the launching point. The FUT attenuation, between the points where P_1 and P_2 have been measured, is calculated using P_1 and P_2, from the definition equations provided above.

Backscattering method. The attenuation coefficient of a singlemode fiber is characterized using bidirectional backscattering measurements. This method is also used for:

- attenuation uniformity;
- optical continuity;
- physical discontinuities;
- splice losses; and
- fiber length.

An optical time domain reflectometer (OTDR) is used for performing such characterization. Adjustment of laser pulsewidth and power is used to obtain a compromise between resolution (a shorter pulsewidth provides a better resolution but at lower power) and dynamic range/fiber length (higher power provides better dynamic range but with longer pulsewidth). An example of such an instrument is shown in Figure 3.

The measurement is applicable to various FUT configurations (e.g., cabled fiber in production or deployed in the field, fiber on a spool, etc.). Two unidirectional backscattering curves are obtained, one from each end of the fiber (Figure 4).

Each backscattering curve is recorded on a logarithmic scale, avoiding the parts at the two ends of the curves, due to parasitic reflections.

The FUT length is found from the time interval between the two ends of the backscattering loss

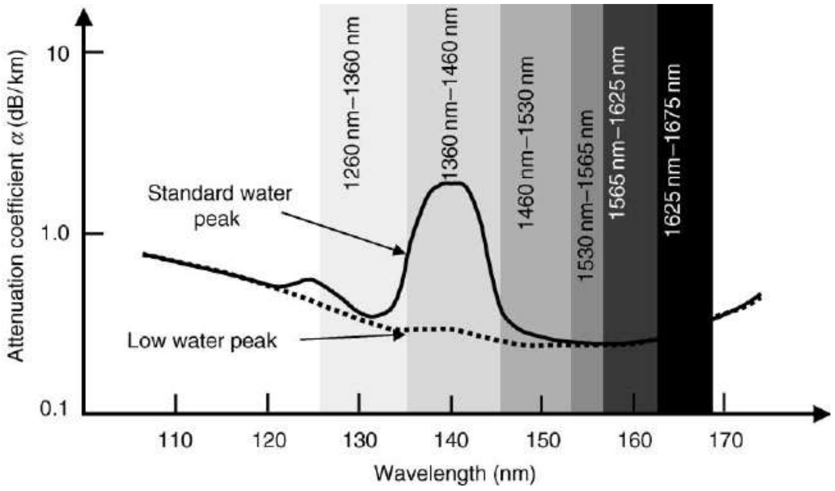

Figure 2 Typical spectral attenuation of a singlemode fiber.

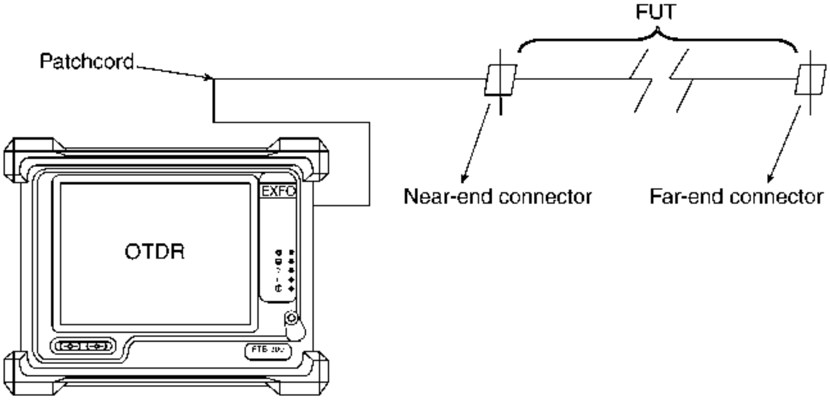
FUT
Patchcord
EXFO
OTDR
Near-end connector
Far-end connector

Figure 3 The backscattering method (OTDR).

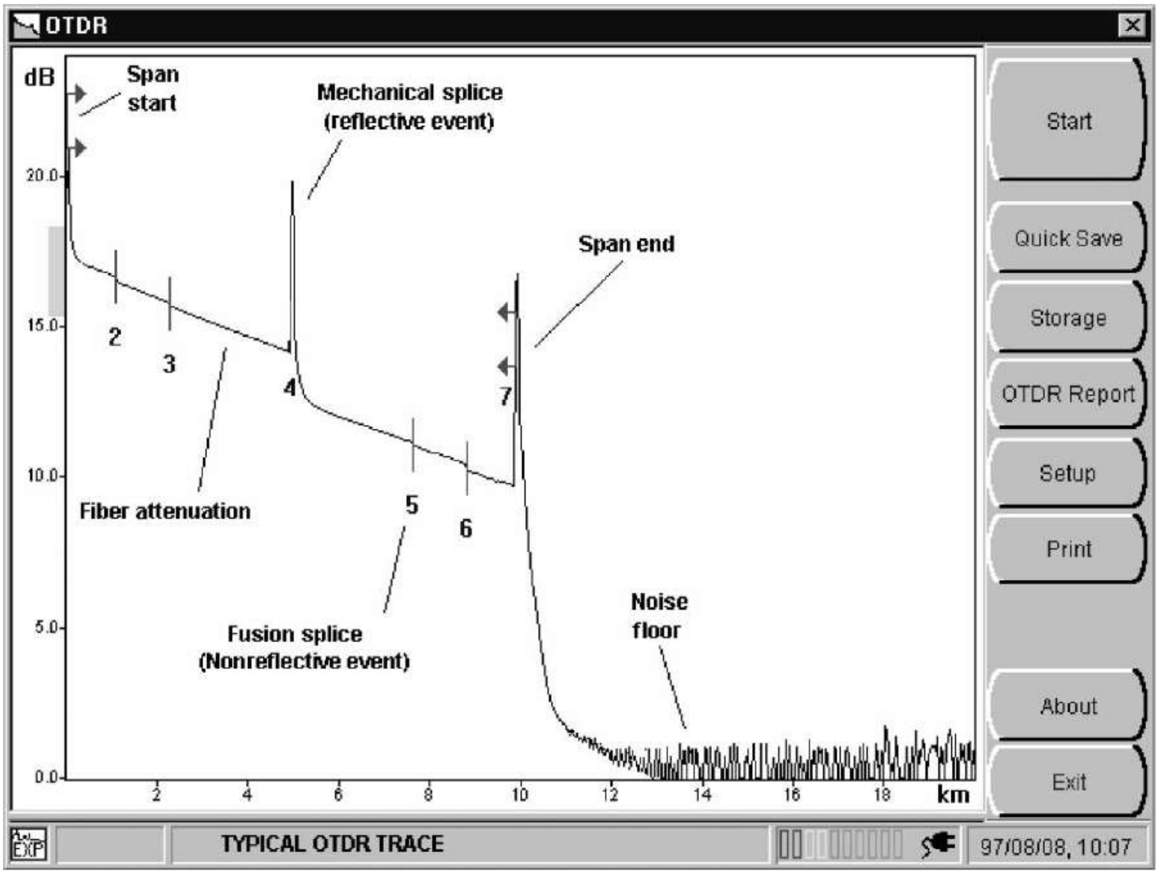
OTDR
dB
Span start
Mechanical splice (reflective event)
Span end
Fiber attenuation
Fusion splice (Nonreflective event)
Noise floor
2
3
4
5
6
7
20.0
15.0
10.0
5.0
0.0
2
4
6
8
10
12
14
16
18
km
Start
Quick Save
Storage
OTDR Report
Setup
Print
About
Exit
TYPICAL OTDR TRACE
97/08/08, 10:07

Figure 4 Unidirectional OTDR backscattering loss measurement.

curve together with the FUT group index of refraction, n_g, as:

$$L_f = c_{fs} \cdot \frac{\Delta T_f}{n_g} \qquad [7]$$

c_{fs} is the velocity of light in free space.

The bidirectional backscattering curve is obtained using two unidirectional backscattering curves and calculating the average loss between them. The end-to-end FUT attenuation coefficient is obtained from the difference between two losses divided by the difference of their corresponding distances.

Insertion loss method. The insertion loss method consists of the measurement of the power loss due to the FUT insertion between a launching and a receiving system, previously interconnected (reference condition). The powers P_1 and P_2 are evaluated in a less straightforward way than in the cut-back method. Therefore, this method is not intended for use in manufacturing.

The insertion loss technique is less accurate than the cut-back, but has the advantage of being non-destructive for the FUT. Therefore, it is particularly suitable in the field.

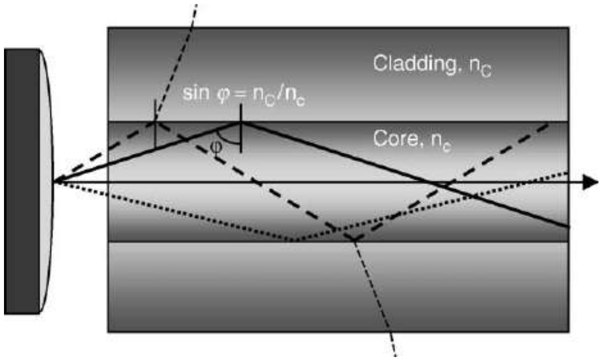
Cladding, n_C
$\sin\varphi = n_C/n_c$
Core, n_c
φ

Figure 5 Total internal reflection and macrobending effect on the light rays.

Macrobending Sensitivity

Macrobending sensitivity is the property by which there is a certain amount of light leaking (loss) into the cladding when the fiber is bent and the bending angle is such that the condition of total internal reflection is no longer met at the core-cladding interface. Figure 5 illustrates such a case.

Macrobending sensitivity is a direct function of the wavelength: the longer the wavelength and/or the smaller the bending diameter, the more loss the fiber experiences. It is recognized that 1625 nm is a wavelength that is very sensitive to macrobending (see Figure 6).

The macrobending loss is measured by the power monitoring method (OTDR, especially for field assessment, see Figure 6) or the cut-back method.

Microbending Sensitivity

Microbending is a fiber property by which the core-cladding concentricity randomly changed along

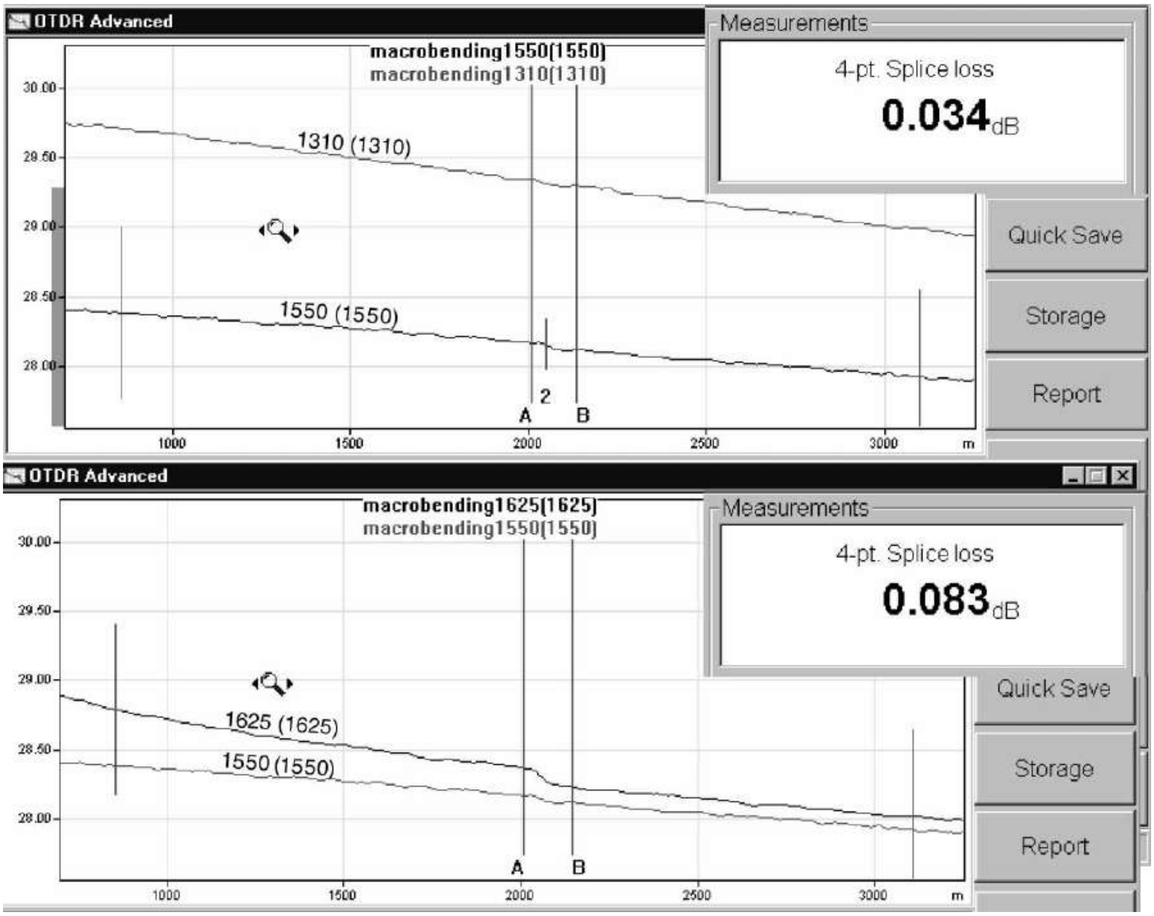
OTDR Advanced
macrobending1550(1550)
macrobending1310(1310)
30.00
29.50
29.00
28.50
28.00
1310 (1310)
1550 (1550)
2
A
B
1000
1500
2000
2500
3000
m
Measurements
Quick Save
Storage
Report
OTDR Advanced
macrobending1625(1625)
macrobending1550(1550)
30.00
29.50
29.00
28.50
28.00
1625 (1625)
1550 (1550)
A
B
1000
1500
2000
2500
3000
m
Measurements
Quick Save
Storage
Report

4-pt. Splice loss
0.034 dB

4-pt. Splice loss
0.083 dB

Figure 6 Macrobending sensitivity as a function of wavelength.

the fiber length. It causes the core to wobble inside the cladding and along the fiber length.

Four methods are available for characterizing microbending sensitivity in optical fibers:

- expandable drum for singlemode fibers and optical fiber ribbons over a wide range of applied linear pressure or loads;
- fixed diameter drum for step-index multimode, singlemode, and ribbon fibers for a fixed linear pressure;
- wire mesh and applied loads for step-index multimode and singlemode fibers over a wide range of applied linear pressure or loads; and
- 'basketweave' wrap on a fixed diameter drum for singlemode fibers.

The results from the four methods can only be compared qualitatively. The test is nonroutine for general evaluation of optical fiber.

Cut-off Wavelength

The cut-off wavelength is the shortest wavelength at which a single mode can propagate in a singlemode fiber. This parameter can be computed from the fiber refractive index profile (RIP). At wavelengths below the cut-off wavelength, several modes propagate and the fiber is no longer singlemode, but multimode.

In optical fibers, the change from multimode to singlemode behavior does not occur at a specific wavelength, but rather over a smooth transition as a function of wavelengths. Consequently, from a fiber-optic network standpoint, the actual threshold wavelength for singlemode performance is more critical. Thus an effective cut-off wavelength is described below.

The cut-off wavelength is defined as the wavelength greater than the ratio between the total power, in the higher-order modes and the fundamental mode, which has decreased to less than 0.1 dB. According to this definition, the second-order mode LP_{11} undergoes 19.3 dB more attenuation than the fundamental LP_{01} mode when the modes are equally excited.

Because the cut-off wavelength depends on the fiber length, bends, and strain, it is defined on the basis of the following three cases:

- fiber cut-off wavelength;
- cable cut-off wavelength; and
- jumper cable cut-off wavelength.

Fiber cut-off wavelength: Fiber cut-off wavelength λ_{fc} is defined for uncabled primary-coated fiber and is measured over 2 m with one loop of 140 mm radius loosely constrained, with the rest of the fiber kept essentially straight. The presence of a primary coating on the fiber usually will not affect λ_{fc}. However, the presence of a secondary coating may result in λ_{fc} being significantly shorter than that of the primary coated fiber.

Cable cut-off wavelength: Cable cut-off wavelength is measured prior to installation on a substantially straight 22 m cable length prepared by exposing 1 m of primary-coated fiber at both ends, the exposed ends each incorporating a 40 mm radius loop. Alternatively, this parameter may be measured on 22 m primary-coated uncabled fiber in the same configuration as for the λ_{fc} measurement.

Jumper cable cut-off wavelength: Jumper cable cut-off wavelength is measured over 2 m with one loop of 76 mm radius, or equivalent (e.g., split mandrel), with the rest of the jumper cable kept essentially straight.

Multimode Fiber Bandwidth for Multimode Fibers

The −3 dB bandwidth of a multimode optical fiber (or modal bandwidth) is defined as the lowest frequency where the magnitude of the baseband frequency response in optical power has decreased by 3 dB relative to the power at zero frequency. Modal bandwidth is also called intermodal dispersion as it takes into account the dispersion between the modes of propagation of the transmitted signal into the multimode fiber.

Various methods of reporting the results are available, but the results are typically expressed in terms of the −3 dB (optical power) frequency. Figure 7 illustrates modal bandwidth.

The bandwidth or pulse broadening may be normalized to a unit length, such as GHz·km, or ns/km.

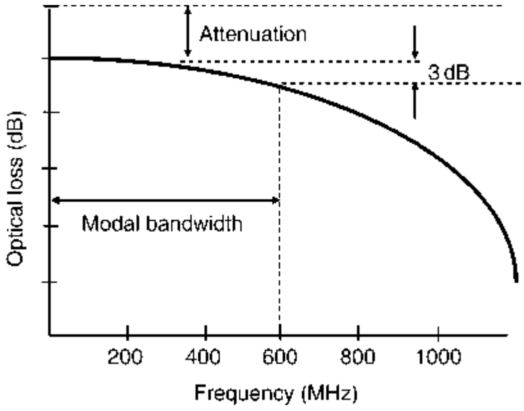

Figure 7 Determination of modal bandwidth.

Two methods are available for determining transmission capacity of multimode fibers:

- the frequency domain measurement method in which the baseband frequency response is directly measured in the frequency domain by determining the fiber response to a sinusoidally modulated light source; or
- the optical time domain measurement method (pulse distortion) in which the baseband response is measured by observing the broadening of a narrow pulse of light.

Differential Mode Delay for Multimode Fibers

Differential mode delay (DMD) characterizes the modal structure of a graded-index glass-core multimode fiber. DMD is useful for assessing the bandwidth performance of a fiber when used with short-pulse, narrow spectral-width laser sources.

The output from a singlemode probe fiber excites the multimode FUT at the test wavelength. The probe spot is scanned across the FUT endface, and the optical pulse delay is determined at specified radial offset positions between an inner and an outer limit. DMD is the difference in optical pulse delay time between the FUT fastest and slowest modes excited for all radial offset positions between and including the inner and the outer limits.

The related critical issues influencing DMD are the temporal width of the optical pulse, jitter in the timing, the finite bandwidth of the optical detector, and the mode broadening due to the source spectral width and the FUT chromatic dispersion.

The test method is commonly used in production and research facilities, but is not easily accomplished in the field. DMD may be a good predictor of the source launching conditions. DMD may be normalized to a unit length, such as ps/m.

Chromatic Dispersion

Chromatic dispersion in a singlemode fiber is a combination of material dispersion and waveguide dispersion (see Figure 8), and it contributes to pulse broadening and distortion in a digital signal.

Material dispersion is produced by the dopants used in glass and is important in all fiber types. Waveguide dispersion is produced by the wavelength dependence of the index of refraction and is critical in singlemode fibers only.

From the point of view of the transmitter, this is due to two causes:

- The presence of wavelengths in the source optical spectrum. Each wavelength has a different phase delay and group delay (different group velocities) along the fiber, because they travel under different index of refraction (or phase) as the index varies as a function of wavelengths, as shown in Figure 9.
- The other cause is the modulation of the source, which itself has two effects:
  - As bit-rates increase, the spectral width of the modulated signal increases and can become comparable to or exceed the spectral width of the source.
  - Chirp occurs when the source wavelength spectrum varies during the pulse. By convention, positive chirp at the transmitter occurs when the spectrum during the rise/fall of the pulse shifts towards shorter/longer wavelengths respectively. For a positive fiber dispersion coefficient, longer wavelengths are delayed relative to shorter wavelengths. Hence if the sign of the product of chirp and dispersion is positive, the two processes combine to produce pulse broadening. If the product is negative, pulse compression can occur over an initial fiber length until the pulse reaches a minimum width and then broadens again with increasing dispersion.

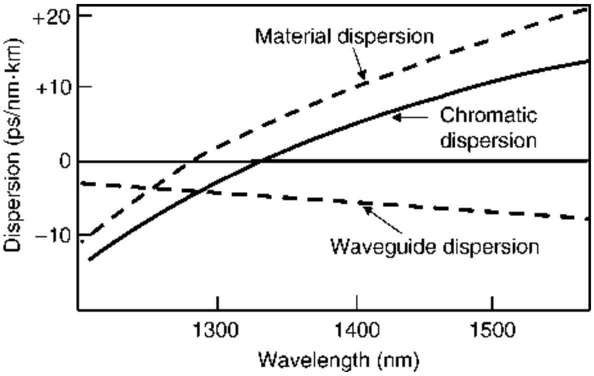

Figure 8 Contribution of the material and waveguide dispersions to the chromatic dispersion.

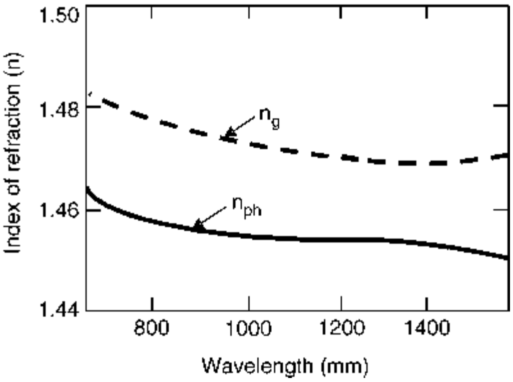

Figure 9 Difference between the phase and the group index of refraction.

The electric field propagating into the FUT may be simply described as follows:

$$E(t,z) = E_0 \sin(\omega t - \beta z) \qquad [8]$$

$\omega = 2\pi c/\lambda$ [rad/s] is the angular frequency; $\beta = kn = (\omega/c)n$ is the effective index; as β has units of m^{-1} it is often referred to as wavenumber or even sometimes propagation constant; k is the propagation constant.

The group delay τ_g is then given by:

$$d\beta/d\omega = \beta_1 = \tau_g \qquad [9]$$

An example of the group delay spectral distribution is shown in Figure 10. Assuming n is not complex; β_1 is the first-order derivative of β.

The group velocity v_g is given by

$$v_g = (d\beta/d\omega)^{-1} = -(\lambda^2/2\pi c)(d\beta/d\lambda)^{-1} \qquad [10]$$

The FUT input–output delay is given by

$$\tau_d = L/v_g \qquad [11]$$

L is the FUT length.

The group index of refraction n_g is given by

$$n_g = c/v_g = n - \lambda(dn/d\lambda) \qquad [12]$$

The dispersion parameter or dispersion coefficient D (ps/nm · km) is given by

$$D = -(\omega/\lambda)(d\tau_g/d\omega) = -(2\pi c/\lambda^2)(d^2\beta/d\omega^2)$$

$$= -(\lambda/c)(d^2 n/d\lambda^2) \qquad [13]$$

$$d^2\beta/d\omega^2 = \beta_2 \qquad [14]$$

An example of the spectral distribution of D obtained from the group delay is shown in Figure 10.

β_2 (ps²/km) is the group velocity dispersion parameter, so D may be related to β_2, as follows:

$$D = -(\omega/\lambda)\beta_2 \qquad [15]$$

When β_2 is positive then D is negative and vice-versa. The region where β_2 is positive is called normal dispersion while the negative-β_2 region is called anomalous dispersion.

At λ_0, β_1 is minimum, and $\beta_2 = 0$, then $D = 0$.

The dispersion slope S ($ps/nm^2 \cdot km$), also called the differential dispersion parameter or second-order dispersion, is given by

$$S = dD/d\lambda = (\omega/\lambda)\beta_3 + (2\omega/\lambda^2)\beta_2 \qquad [16]$$

$$\beta_3 = d\beta_2/d\omega = d^3\beta/d\omega^3$$

At λ_0, β_1 is minimum, $\beta_2 = 0$, then $D = 0$; but S is not zero and depends on β_3. An example of the spectral distribution of S and S_0 is illustrated in Figure 10.

Overall, the general expression of β is given by

$$\beta(\omega) = \beta_0 + (\omega - \omega_0)\beta_1 + (1/2)(\omega - \omega_0)^2\beta_2$$

$$+ (1/12)(\omega - \omega_0)^3\beta_3 + \cdots \qquad [17]$$

Figure 11 illustrates the difference between dispersion unshifted fiber (ITU-T Rec. G.652), dispersion shifted fiber (ITU-T Rec. G.653) and nonzero dispersion shifted fiber (ITU-T Rec. G.655).

Test methods for the determination of chromatic dispersion

All methods measure the group delay at a specific wavelength over a range and use agreed fitting functions to evaluate λ_0 and S_0.

In the phase shift method, the group delay is measured in the frequency domain, by detecting, recording, and processing the phase shift of a sinusoidal modulating signal between a reference, a secondary fiber path, and the channel signal.

Setup variances exist and some do not require the secondary reference path. For instance, by using a reference optical filter at the FUT output it is possible to completely decouple the source from the phase-meter. With such an approach, chromatic dispersion may now be measured in the field over very long links using optical amplifiers (see Figure 12).

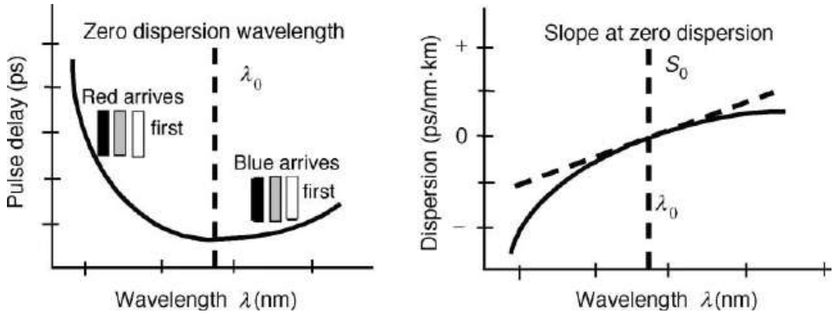

Figure 10 Relation between the pulse (group) delay and the (chromatic) dispersion.

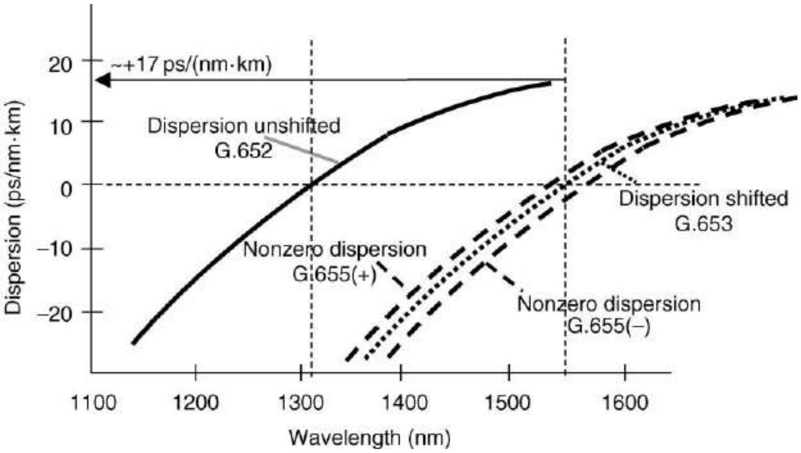

Figure 11 Chromatic dispersion for various types of fiber.

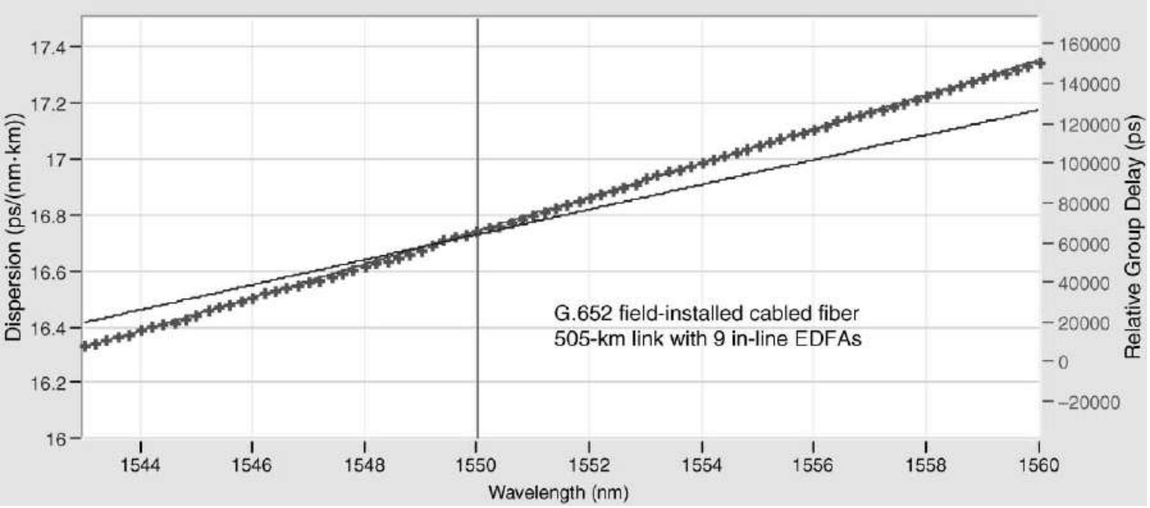

Figure 12 Test results for field-related chromatic dispersion using the phase-shift technique.

In the differential phase shift method, two detection systems are used together with two wavelength sources at the same time. In this case the chromatic dispersion may be determined directly from the two group delays. This technique usually offers faster and more reliable results but costs much more than the phase-shift technique which is usually preferred.

In the interferometric method, the group delay between the FUT and a reference path is measured by a Mach–Zehnder interferometer. The reference delay line may be an air path or a singlemode fiber standard reference material (SRM). The method can be used to determine the following characteristics:

- longitudinal chromatic dispersion homogeneity; and
- effect of overall or local influences, such as temperature changes and macrobending losses.

In the pulse delay method, the group delay is measured in the time domain, by detecting, recording, and processing the delay experienced by pulses at various wavelengths.

Polarization Mode Dispersion

Polarization mode dispersion (PMD) causes an optical pulse to spread in the time domain and may impair the performance of a telecommunications system. The effect can be related to differential phase and group velocities and corresponding arrival time of different polarization components of the pulse signal. For a sufficiently narrowband source, the effect can be related to a differential group delay (DGD), $\Delta\tau$, between a pair of orthogonally polarized principal states of polarization (PSP) at a given wavelength or optical frequency (see Figure 13a).

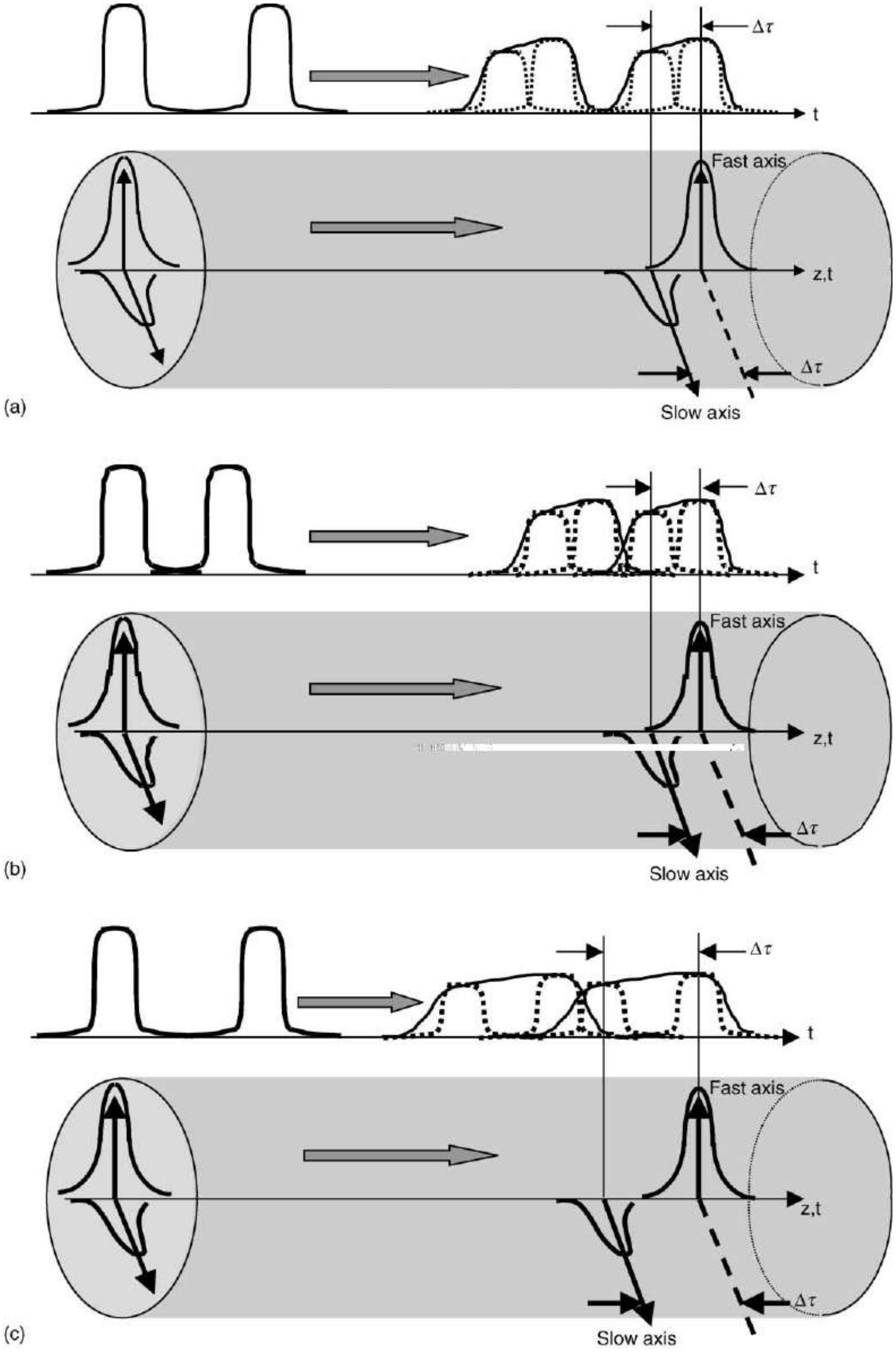
Δτ
t
Fast axis
z,t
Δτ
Slow axis
(a)
Δτ
t
Fast axis
z,t
Δτ
Slow axis
(b)
Δτ
t
Fast axis
z,t
Δτ
Slow axis
(c)

Figure 13 PMD effect on the pulse broadening and its possible pulse impairment. (a) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is too low to create an impairment. (b) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is high enough to create an impairment. (c) The DGD $\Delta\tau$ is large enough even at low bit rate to make the pulse spreading and creating impairment.

In an ideal circular symmetric fiber, the two PSPs propagate with the same velocity. However:

- a real fiber is not perfectly circular;
- the core is not perfectly concentric with the cladding;
- the core may be subjected to microbending;
- the core may present localized clusters of dopants; and
- the environmental conditions may stress the deployed cable and affect the fiber.

Each time the fiber undergoes local stresses and consequently birefringence. These asymmetry characteristics vary randomly along the fiber and in time, lead to a statistical behavior of PMD.

For a deployed cabled fiber at a given time and optical frequency, there always exist two PSPs such that the pulse spreading due to PMD vanishes, if only one PSP is excited. On the contrary, the maximum pulse spread due to PMD occurs when both PSPs are equally excited, and is related to the difference in their group delays, the DGD associated with the two PSPs. For broadband transmission, the DGD statistically varies as a function of wavelengths or frequencies and result in an output pulse that is spread out in the time domain (see Figures 13a–c). In this case, the spreading can be related to the RMS (root mean square) of DGD values $\langle \Delta\tau^2 \rangle^{1/2}$. However, if a known distribution such as the Maxwell distribution may be fit to the DGD distribution probability, then a mean (or average) value of the DGD $\langle \Delta\tau \rangle$ may be correlated to the RMS value and used as a system performance predictor in particular with a maximum value of the DGD distribution associated to a low probability of occurrence. This maximum DGD may then be used to define the quality of service that would tolerate values lower than this maximum DGD.

Test methods for polarization mode dispersion

Three methods are generically used for measuring PMD. Other methods or analyses may exist but they are generally not standardized or are limited in their applications.

- Stokes parameter evaluation (SPE)
  - Jones matrix eigenanalysis (JME)
  - Poincaré sphere analysis (PSA)
- Fixed analyzer (FA)
  - Extrema counting (EC)
  - Fourier transform (FT)
- Interferometry (INTY)
  - Traditional analysis (TINTY)
  - General analysis (GINTY).

All methods use a linearly polarized source at the FUT input and are suitable for laboratory measurements of factory lengths of fiber and cable. However, the interferometric method is the only method appropriate for measurements of cabled fiber that may be moving or vibrating such as is found in the field.

Stokes parameter evaluation. SPE determines PMD by measuring a response to a change of narrowband light (from a tuneable light source with broadband detector – JME, or a broadband source with a filtered detector such as an interferometer – PSA) across a wavelength range. The Stokes vector of the output light is measured for each wavelength. The change of these Stokes vectors with angular optical frequency (wavelength), ω and with the change in input SOP (state of polarization), yields the DGD as a function of wavelength.

For both JME and PSA analyses, three distinct and known linear SOPs (orthogonal on the Poincaré sphere) must be launched for each wavelength. Figure 14 illustrates the test setup and examples of test results.

The JME and PSA method are mathematically equivalent.

Fixed analyzer. FA determines PMD by measuring a response to a change of narrowband light across a wavelength range. For each SOP, the change in output power that is filtered through a fixed polarization analyzer, relative to the power detected without the analyzer, is measured as a function of wavelength. Figure 15 illustrates a test setup and examples of test results.

The resulting measured function can be analyzed in one of two ways:

- by counting the number of peaks and valleys (EC) of the curve and application of a formula. This analysis is considered as a frequency domain approach; and
- by taking the Fourier transform (FT) of the measured function. This FT is equivalent to the pulse spreading obtained by TINTY.

Interferometry. INTY uses a broadband light source and an interferometer. The fringe pattern containing the source auto-correlation, together with the PMD related cross-correlation of the emerging electromagnetic field, is determined by the interference pattern of the output light, i.e., the interferogram. The PMD determination for the wavelength range associated with the source spectrum is based on the envelope of the fringe

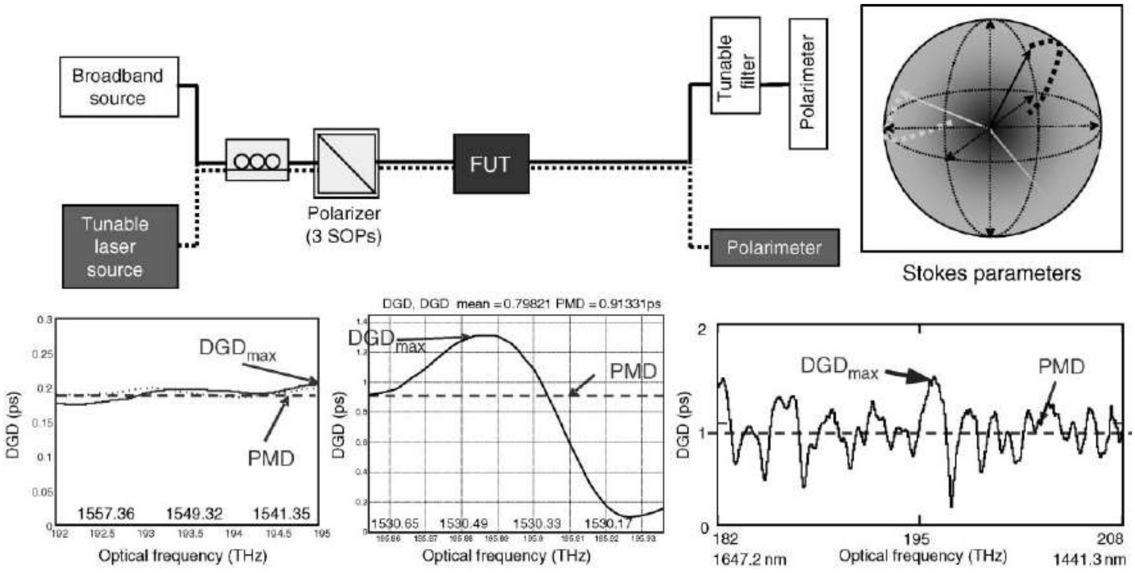
Broadband source
Tunable laser source
Polarizer (3 SOPs)
FUT
Tunable filter
Polarimeter
Polarimeter
Stokes parameters
DGD$_{max}$
PMD
DGD (ps)
Optical frequency (THz)
1557.36
1549.32
1541.35
DGD, DGD mean = 0.79821 PMD = 0.91331ps
1530.65
1530.49
1530.33
1530.17
182
1647.2 nm
195
208
1441.3 nm

Figure 14 PMD by Stokes parameter evaluation method.

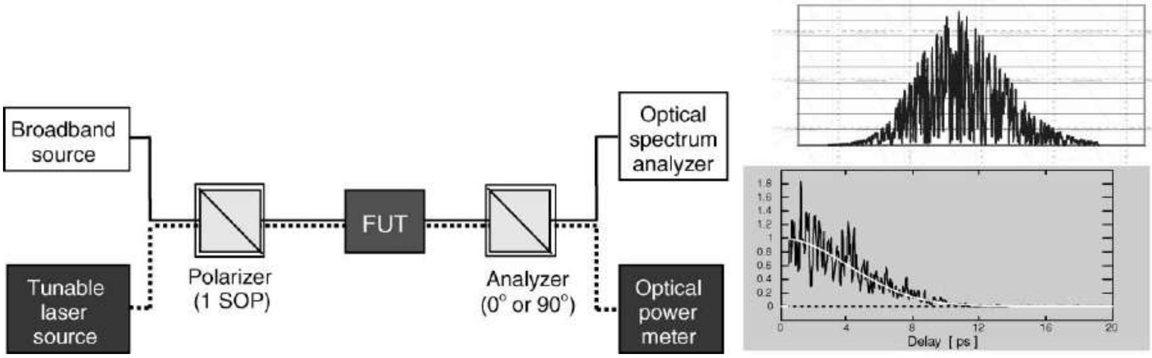
Broadband source
Tunable laser source
Polarizer (1 SOP)
FUT
Analyzer (0° or 90°)
Optical spectrum analyzer
Optical power meter
Delay [ps]

Figure 15 PMD by fixed analyzer method.

pattern interferogram. Two analyses are available to obtain the PMD:

- TINTY uses a set of specific operating conditions for its successful applications and a basic setup; and
- GINTY uses no limiting operating conditions, but in addition to the same basic setup, also using a modified setup compared to TINTY.

Figure 16 illustrates the test setup for both approaches and examples of test results.

Polarization Crosstalk

Polarization crosstalk is a characteristic of energy mixing/transfer/coupling between the two PSPs in a PMF (polarization maintaining fiber) when their isolation is imperfect. It is the measure of the strength of mode coupling or output power ratio between the PSPs within a PMF.

A PMF is an optical fiber capable of transmitting, under external perturbations, such as bending or lateral pressures, both HE_{11}^x and HE_{11}^y polarization modes whose electric field vector directions are orthogonally to each other and which have different propagation constants β_x and β_y.

Two methods are available for measuring the polarization crosstalk of PMF:

- *power ratio method*, which uses the maximum and minimum values of output power at a specified wavelength, and is applicable to fibers and connectors jointed to a PMF, and to two or more PMFs joined in series; and
- *in-line method*, which uses an analysis of the Poincaré sphere, and is applicable to single or cascaded sections of PMF, and to PMF inter-connected with optical devices.

Nonlinear Effects

When the power of the transmission signal is increased to achieve longer span lengths at high bit

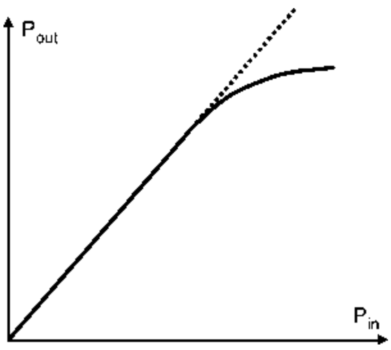

Figure 17 Power output-to-power input relationship for production of nonlinear effects.

expressed as follows:

$$n = n_0 + n_2 I \qquad [18]$$

n is the nonlinearity dependent index; n_0 is the linear part of the index; n_2 is the nonlinear index, also called the Kerr nonlinear index (2.2 to 3.4×10^{-16} cm^2/W); and I is the optical intensity inside the fiber.

The field propagation at a distance L into the fiber is described by the following equation:

$$E_{\text{out}}(L) = E_{\text{in}}(0)\exp[-a/2 + \mathrm{i}\beta + \gamma P(L,t)/2]L \qquad [19]$$

$a/2$ is the attenuation; $\mathrm{i}\beta$ is the phase of the wave; and $\gamma P(L,t)/2$ is the nonlinearity term;

$$\gamma = 2\pi n_2/\lambda A_{\text{eff}} \qquad [20]$$

γ is the nonlinearity coefficient and may be a complex number; A_{eff} is the fiber core effective area; $P(L,t)$ is the total power; and λ is the signal wavelength and t the time variable.

The nonlinear coefficient is defined as n_2/A_{eff}. This coefficient plays a critical role in the fiber and is closely related to system performance degradation due to nonlinearities when very high power is used.

Methods for measuring the nonlinear coefficient. Two methods are available for measuring the nonlinear coefficient:

- Continuous-wave dual-frequency (CWDF); and
- Pulsed single-frequency (PSF).

In the CWDF method, light from two wavelengths is injected into the fiber. At higher power, the light from the two wavelengths beat due to the nonlinearity and produce an output spectrum that is spread. The relationship of the power level to a particular spreading is used to calculate the nonlinear coefficient.

In the PSF method, the pulsed light from a single wavelength is injected into the fiber. Very short pulses (<1 ns) and their input peak power must be measured and related to the nonlinear spreading of the output spectrum.

Stimulated Brillouin scattering

In an intensity modulated system using a source with a narrow linewidth, significant optical power is transferred from the forward-propagating signal to a backward-propagating signal when the SBS power threshold is exceeded. At that point periodic regions of index of refraction produce a grating traveling at speed of sound away from the source. The grating reflects backward part of the incident light. The reflected sound waves (acoustic phonons) scatter light back to the source. Phase matching (or momentum conservation) dictates that the scattered light preferentially travels in the backward direction. The scattered light will be Doppler-shifted (downshifted or Brillouin-shifted) by approximately 11 GHz (at 1550 nm, for G.652 fiber). The scattered light has a very narrow spectrum (very coherent) and very close to the carrier signal and may be very detrimental.

Stimulated Raman scattering

SRS is an interaction between light and the fiber molecular vibrations as adjacent atoms vibrate in opposite directions (an ‘optical phonon’). Some of the energy of the main carrier (optical pump wave) is transferred to the molecules, thereby further increasing the amplitude of their vibrations. If the vibrations become large enough, a threshold is reached at which the local index of refraction changes. These local changes then scatter light in all directions similar to Rayleigh scattering. However, unlike Rayleigh scattering, the wavelength of the Raman scattered light is shifted to longer wavelengths by an amount that corresponds to the molecular vibration frequencies and the Raman signal spreads over a large spectrum.

Self-phase modulation

SPM is the effect that a powerful pulse has on its own phase, considering that in eqn [18], $I(t)$ varies in time:

- $I(t) \rightarrow n(t) = n_0 + n_2 I(t) \rightarrow$ modulates the phase $\beta(t)$ of the pulse; and
- $\mathrm{d}I/\mathrm{d}t \rightarrow \mathrm{d}n/\mathrm{d}t \rightarrow \mathrm{d}\beta/\mathrm{d}t$(chirp) → broadening in the frequency domain → broadening in the time domain.

$I(t)$ peaks at the center of the pulse (peak power) and consequently increases the index of refraction. A higher index causes the wavelengths in the center of the pulse to accumulate phase more quickly than at the wings of the pulse:

- this causes wavelength stretching (shift to longer wavelengths) at pulse leading edge (risetime); and
- this causes wavelength compression (shift to shorter wavelengths) at pulse trailing edge (falltime).

The pulse will then broaden with negative (normal) dispersion and shorten with positive (anomalous) dispersion. SPM may then be used for dispersion compensation considering that self-phase modulation imposes $C > 0$ (positive chirp). It can cancel the dispersion if properly manage as a function of the sign of the dispersion. SPM is one of the most critical nonlinear effects for the propagation of soliton or very short pulses over very long distance.

Cross-phase modulation

XPM is the effect that a powerful pulse has on the phase of an adjacent pulse from another WDM system channel traveling in phase or at slightly the same group velocity. It concerns spectral interference between two WDM channels:

- the increasing $I(t)$ at the leading edge of the interfering pulse shifts the other pulse to longer wavelength; and
- decreasing $I(t)$ at the trailing edge of the interfering pulse shifts the other pulse to shorter wavelengths.

This produces spectral broadening, which dispersion converts to temporal broadening depending on the sign of the dispersion. XPM effect is similar to SPM except it depends on the channel count.

Four-wave (four-photon) mixing

FWM is the by-product production effect from two or more WDM channels. For two channels $I(t)$ modulates the phase of each signal (ω_1 and ω_2). An intensity modulation appears at the beat frequency $\omega_1 - \omega_2$.

Two sideband frequencies are created in a similar way as harmonics generation. New wavelengths are created in a number equal to $N^2(N-1)/2$, where N = number of original wavelengths.

Fiber Dimension Characteristics and Corresponding Tests Methods

Table 1 provides a list of the various fiber dimensional characteristics and their corresponding test methods.

Table 1 Fiber dimensional characteristics

Attribute	*Measured parameter*
Fiber geometry	Core/cladding diameter Core/cladding noncircularity Core-cladding concentricity error
Numerical aperture	
Mode field diameter	
Coating geometry	
Length	

Fiber Geometry Characteristics

The fiber geometry is related to the core and cladding characteristics.

Core

The core center is the center of a circle which best fits the points at a constant level in the near-field intensity profile emitted from the central region of the fiber, using wavelengths above and/or below the cut-off wavelength.

The RIP can be measured by refracted near field (RNF) or transverse interferometry techniques and transmitted near field (TNF).

The core concentricity error is the distance between the core center and the cladding center. This definition applies very well for multimode fibers. The distance between the center of the near field profile and the center of the cladding is also used for singlemode fibers.

The mode field diameter (MFD) represents a measure of the transverse electromagnetic field intensity of the mode in a fiber cross-section and it is defined from the far-field intensity distribution.

The MF is the singlemode field distribution of the LP_{01} mode, giving rise to a spatial intensity distribution in the fiber.

The MF concentricity error is the distance between the MF center and the cladding center.

The core noncircularity is a measure of the core ellipticity. This parameter is one of the causes for creating birefringence in the fiber and consequently PMD.

Cladding

The cladding is the outermost region of constant refractive index in the fiber cross-section.

The cladding center is the center of a circle best fitting the outer limit (boundary) of the cladding.

The cladding diameter is the diameter of the circle defining the cladding center.

The cladding noncircularity is a measure of the difference between the diameters of the two circles defined by the cladding tolerance field divided by the nominal cladding diameter.

Coating

The primary coating is one or more layers of protective material applied to the cladding during or after the drawing process to protect the cladding surface (e.g., a 250 μm protective coating). The secondary coating is one or more layers of protective material applied over the primary coating in order to give additional protection or to provide a particular structure.

Measurement of the fiber geometrical attributes
The fiber geometry is measured by the following methods:

- TNF;
- RNF;
- Side-view technique/transverse interference;
- TNF image technique; and
- Mechanical diameter.

Test instrumentation may incorporate two or more methods such as the one shown in Figure 18.

Transmitted near-field technique. The cladding diameter, core concentricity error, and cladding noncircularity are determined from the near-field intensity distribution. Figure 19 provides a series of examples of test results from TNF measurements.

Refracted near-field technique. The RIP across the entire fiber (core and cladding) can be directly obtained from the RNF measurement, as shown in Figure 20.

The geometrical characteristics of the fiber can be obtained from the refractive index distribution using suitable algorithms:

- core/cladding diameter;
- core/cladding concentricity error;
- core/cladding noncircularity;
- maximum numerical aperture (NA); and
- index and relative index of refraction difference.

Figure 21 illustrates the core geometry.

Side-view technique/transverse interference. The side-view method is applied to singlemode fibers to determine the core concentricity error, cladding diameter and cladding noncircularity by measuring the intensity distribution of light that is refracted inside the fiber. The method is based on an interference microscope focused on the side view of an FUT illuminated perpendicular to the FUT axis. The fringe pattern is used to determine the RIP.

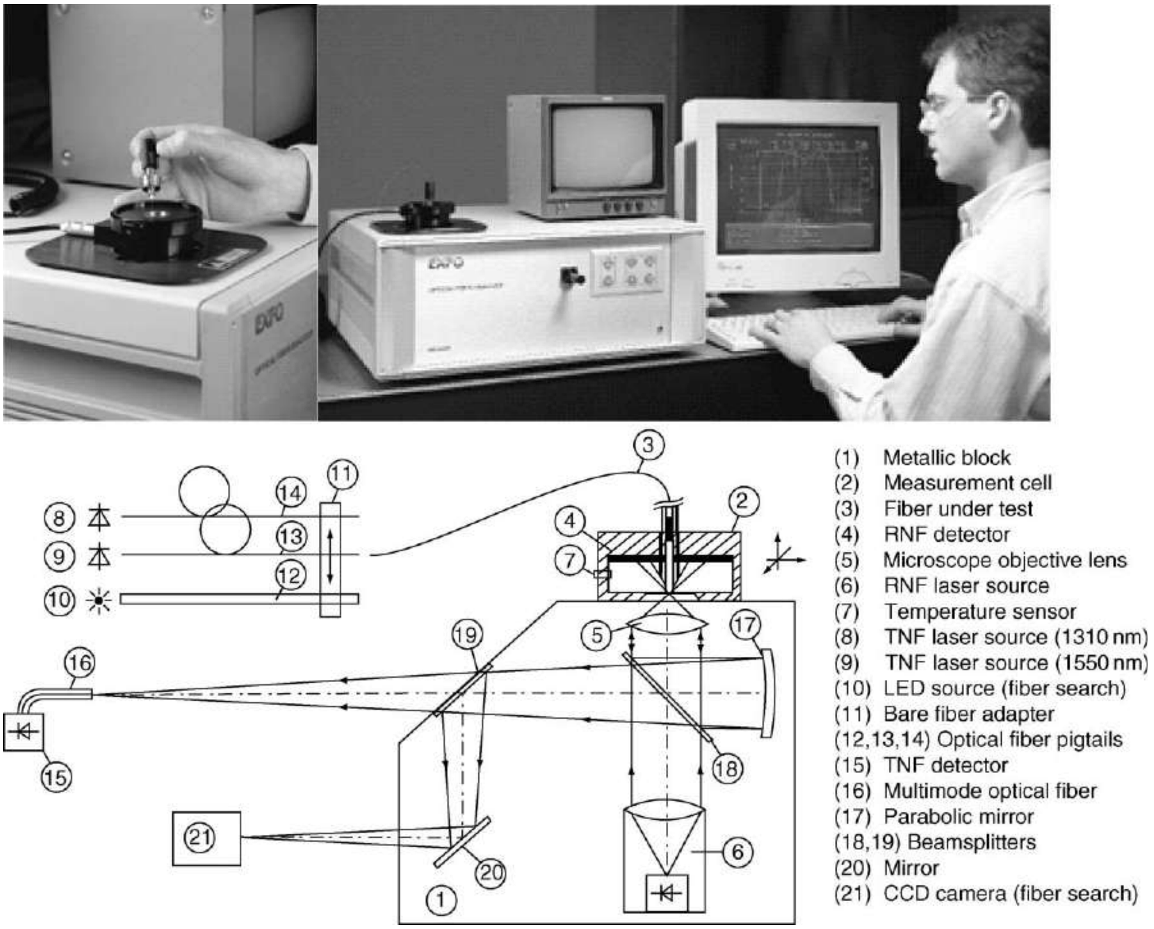
EXFO
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
(1) Metallic block
(2) Measurement cell
(3) Fiber under test
(4) RNF detector
(5) Microscope objective lens
(6) RNF laser source
(7) Temperature sensor
(8) TNF laser source (1310 nm)
(9) TNF laser source (1550 nm)
(10) LED source (fiber search)
(11) Bare fiber adapter
(12,13,14) Optical fiber pigtails
(15) TNF detector
(16) Multimode optical fiber
(17) Parabolic mirror
(18,19) Beamsplitters
(20) Mirror
(21) CCD camera (fiber search)

Figure 18 RNF/TNF combined instrumentation.

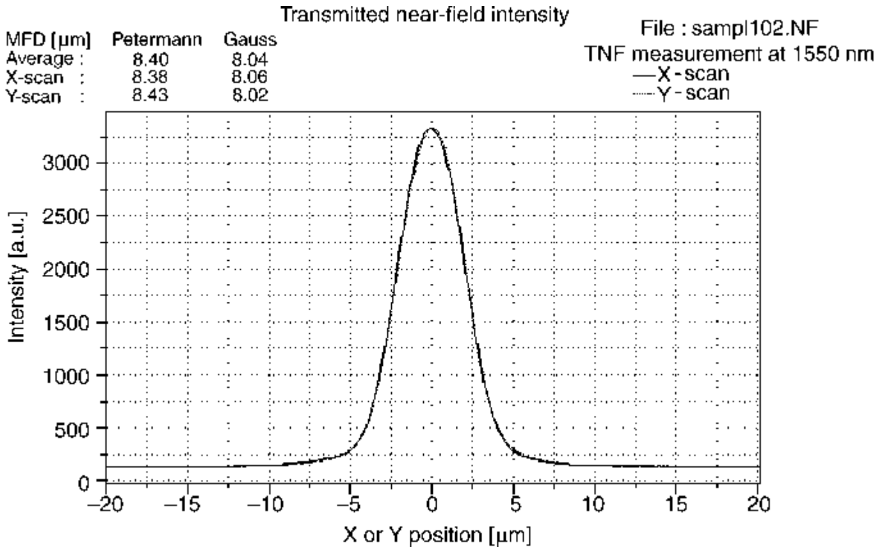

Figure 19 MFD measurement by TNF.

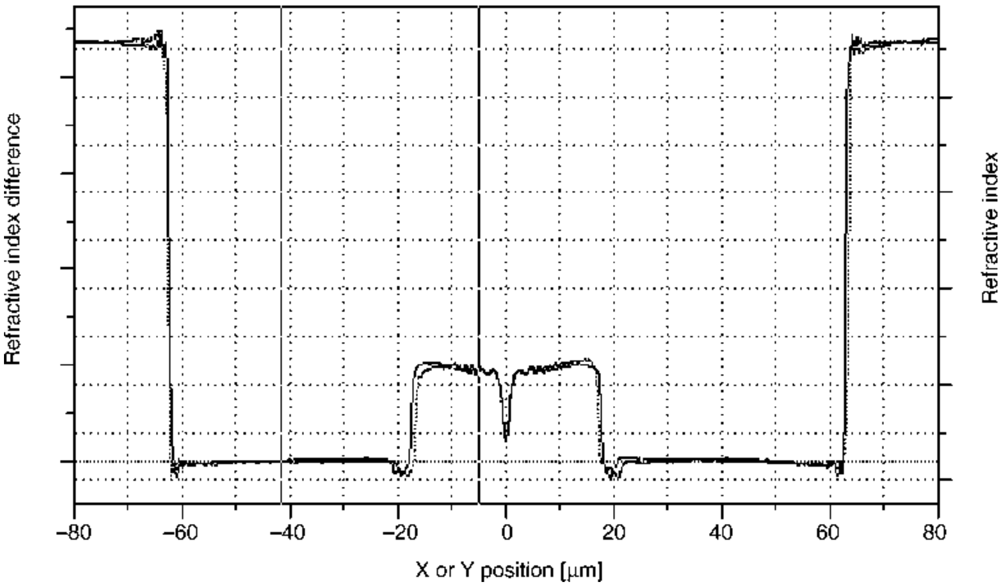

Figure 20 RIP from RNF measurement.

TNF image technique. The TNF image technique, also called near-field light distribution, is used for the measurement of the geometrical characteristics of singlemode fibers. The measurement is based on analysis of magnified images at the FUT output. Two subsets of the method are available:

- grey-scale technique which performs an x–y near-field scan using a video system; and
- Single near-field scanning technique performing a one-dimensional scan.

Mechanical diameter. This is a precision mechanical diameter measurement technique used to accurately determine the cladding diameter of silica fibers. The technique uses an electronic micrometer such as based on a double-pass Michelson interferometer. The technique is used for providing calibrated fibers to the industry as SRM.

Numerical Aperture

The NA is an important attribute for multimode fibers in order to predict their launching efficiency, joint loss at splices and micro/macrobending characteristics.

A method is available for the measurement of the angular radiant intensity distribution (far-field)

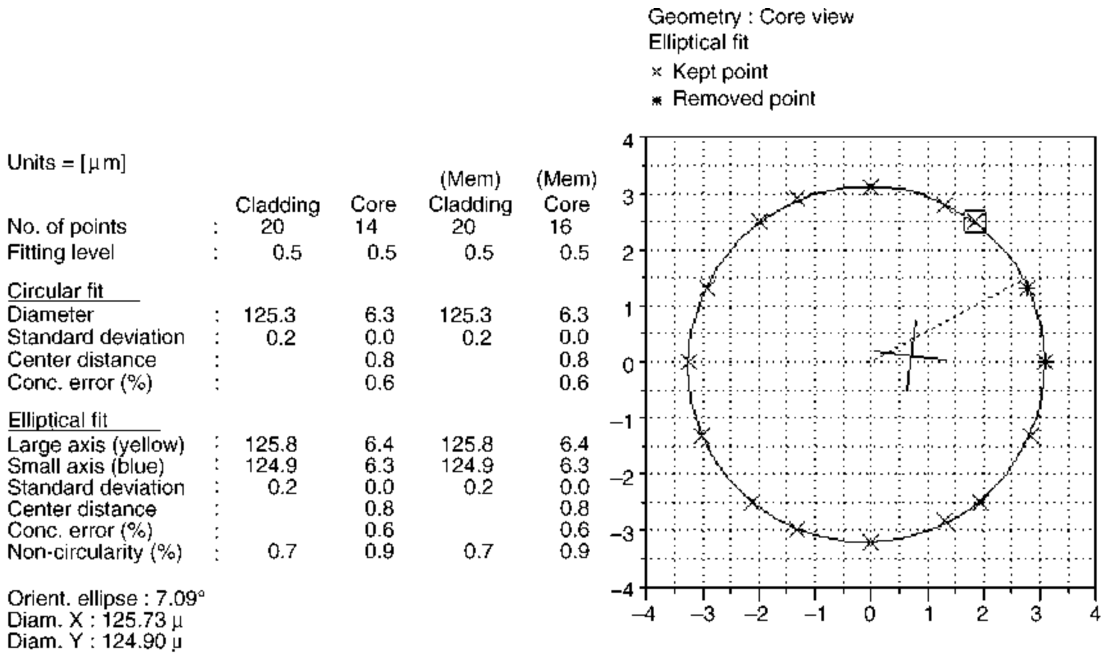

Units = [μm]

		Cladding	Core	(Mem) Cladding	(Mem) Core
No. of points	:	20	14	20	16
Fitting level	:	0.5	0.5	0.5	0.5
Circular fit					
Diameter	:	125.3	6.3	125.3	6.3
Standard deviation	:	0.2	0.0	0.2	0.0
Center distance	:		0.8		0.8
Conc. error (%)	:		0.6		0.6
Elliptical fit					
Large axis (yellow)	:	125.8	6.4	125.8	6.4
Small axis (blue)	:	124.9	6.3	124.9	6.3
Standard deviation	:	0.2	0.0	0.2	0.0
Center distance	:		0.8		0.8
Conc. error (%)	:		0.6		0.6
Non-circularity (%)	:	0.7	0.9	0.7	0.9

Orient. ellipse : 7.09°
Diam. X : 125.73 μ
Diam. Y : 124.90 μ

Figure 21 Core geometry by RNF measurement.

distribution or the RIP at the output of an FUT. NA can be determined from the analysis of the test results.

Mode Field Diameter

The definition of the MFD is given in the section describing the core center above. Four measurement methods are available:

- direct far-field scan determines MFD from the far-field intensity distribution;
- variable aperture in far field determines MFD from the complementary aperture transmission function $a(x)$, $x = d\tan\theta$ being the aperture radius and d the distance between the aperture and the FUT:

$$\mathrm{MFD} = \frac{\lambda}{\pi d}\left[\int_0^\infty a(x)\frac{x}{(x^2+d^2)^2}\,\mathrm{d}x\right]^{-1/2} \qquad [21]$$

Equation [21] is valid for small-θ approximation.

- near-field scan determines MFD from the near-field intensity distribution I_{NF}, r being the radial coordinate:

$$\mathrm{MFD} = 2\left[2\frac{\displaystyle\int_0^\infty rI_{\mathrm{NF}}(r)\,\mathrm{d}r}{\displaystyle\int_0^\infty r\left[\frac{\mathrm{d}I^{1/2}(r)}{\mathrm{d}r}\right]^2\mathrm{d}r}\right]^{1/2} \qquad [22]$$

Equation [22] is valid for small-θ approximation; and

- bidirectional backscattering uses an OTDR and bidirectional measurements to determine MFD by comparing the FUT results with a reference fiber.

Effective Area

A_{eff} is a critical nonlinearity parameter and is defined as follows:

$$A_{\mathrm{eff}} = \frac{2\pi\left[\int_0^\infty I(r)r\,\mathrm{d}r\,\right]^2}{\int_0^\infty I(r)^2 r\,\mathrm{d}r} \qquad [23]$$

$I(r)$ is the field intensity distribution of the fiber fundamental mode at radius r. The integration in the equation is carried out over the entire fiber cross-section. For a Gaussian approximation:

$$I(r) = \exp - 2\left(\frac{2r}{\mathrm{MFD}}\right)^2 \qquad [24]$$

which yields:

$$A_{\mathrm{eff}} = \frac{\pi}{4}\mathrm{MFD}^2 \qquad [25]$$

Three methods are available for the measurement of A_{eff}:

- direct far-field;
- variable aperture in far-field; and
- near-field.

Mechanical Measurement and Test Methods

Table 2 describes the various mechanical characteristics and their corresponding test methods.

Proof Stressing

The proof stress level is the value of tensile stress or strain applied to a full fiber length over a short period of time. The method for fiber proof stressing is the longitudinal tension which describes procedures for applying tensile loads to a length of fiber. The fiber stress is calculated from the applied tension. The tensile load is applied over short periods of time but not too short in order for the fiber to experience proof stress.

Residual Stress

Residual stress is the stress built up by the thermal expansion difference between core and cladding during the fiber drawing process or splicing. Methods are available for measuring residual stress based on polarization effects. A light beam produced by a rotating polarizer propagates to the x-axis direction while the beam polarization is in the $y-z$ plane and the fiber longitudinal axis in the z-axis. The light experiences different phase shift in the y- and z-axis due to the FUT birefringence. The photoelastic effect gives the relationship between this phase change and residual stresses.

Stress Corrosion Susceptibility

The stress corrosion susceptibility is related to the dependence of crack growth on applied stress. It depends on the environmental conditions and static and dynamic values may be observed.

Environmental Characteristics

Table 3 lists the fiber characteristics related to the effect of the environment.

Table 2 Fiber mechanical characteristics

Attribute
Proof stress
Residual stress
Stress corrosion susceptibility
Tensile strength
Stripability
Fiber curl

Table 3 Fiber environmental characteristics

Attribute
Hydrogen aging
Nuclear gamma irradiation
Damp heat
Dry heat
Temperature cycling
Water immersion

Hydrogen Aging for Low-Water-Peak Single-Mode Fiber

Hydrogen aging on low-water peak fibers, such as G.652.C, is based on a test performed at 1.0 atmosphere of hydrogen pressure at room temperature over a period of one month. Other proportional combinations are possible.

Nuclear Gamma Irradiation

Nuclear radiation is considered on the basis of a steady state response of optical fibers and cables exposed to gamma radiation and to determine the level of related radiation-induced attenuation produced in singlemode or multimode cabled or uncabled fibers.

The fiber attenuation generally increases when exposed to gamma radiation. This is primarily due to the trapping of radiolytic electrons and holes at defect sites in the glass (i.e., the formation of 'color centers'). Two regimes are considered:

- the low dose rate suitable for estimating the effect of environmental background radiation; and
- the high dose rate suitable for estimating the effect of adverse nuclear environments.

The effects of environmental background radiation are measured by the attenuation (cut-back method). The effects of adverse nuclear environments are tested by power monitoring before, during and after FUT exposure.

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics). **Imaging:** Interferometric Imaging. **Interferometry:** Gravity Wave Detection; Overview. **Polarization:** Introduction. **Scattering:** Scattering Phenomena in Optical Fibers.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*, 2nd edn. New York: John Wiley & Sons.

Agrawal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.
Girard A (2000) *Guide to WDM Technology and Testing*. Quebec: EXFO, pp. 194.
Hecht J (1999) *Understanding Fiber Optics*, 3rd edn. Upper Saddle River, NJ: Prentice Hall.
Masson B and Girard A (2004) *FTTx PON Guide, Testing Passive Optical Networks*. Quebec: EXFO, pp. 56.
Miller JL and Friedman E (2003) *Optical Communications Rules of Thumb*. New York: McGraw-Hill, pp. 428.
Neumann E-G (1988) *Single-Mode Fibers, Fundamentals*. Berlin: Springer-Verlag.

Nonlinear Effects (Basics)

G Millot and P Tchofo-Dinda, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many physical systems in various areas such as condensed matter or plasma physics, biological sciences, or optics, give rise to localized large-amplitude excitations having a relatively long lifetime. Such excitations lead to a host of phenomena referred to as nonlinear phenomena. Of the many disciplines of physics, the optics field is probably the one in which practical applications of nonlinear phenomena have been the most fruitful, in particular, since the discovery of the laser in 1960. This discovery has thus led to the advent of a new branch in optics, referred to as *nonlinear optics*. The applications of nonlinear phenomena in optics include the design of various kinds of laser sources, optical amplifiers, light converters, light-wave communication systems for data transmission purposes, to name a few.

In this article, we present an overview of some basic principles of nonlinear phenomena that result from the interaction of light waves with dielectric waveguides such as optical fibers. These nonlinear phenomena can be broadly divided into two main categories, namely, parametric effects and scattering phenomena. Parametric interactions arise whenever the state of the dielectric matter is left unchanged by the interaction, whereas scattering phenomena imply transitions between energy levels in the medium. More fundamentally, parametric interactions originate from the electron motion under the electric field of a light wave, whereas scattering phenomena originate from the motion of heavy ions (or molecules).

Linear and Nonlinear Signatures

The macroscopic properties of a physical system can be obtained by analyzing the response of the system under an external excitation. For example, consider at time t the response of a system, such as an amplifier, to an input signal $E_1 = A\sin(\omega t)$. In the low-amplitude limit of the output signal, the response R_1 of the system is proportional to the excitation:

$$R_1 = \alpha_1 E_1 \qquad [1]$$

where α_1 is a constant. This type of behavior corresponds to the so-called linear response. In general, a physical system executes a linear response when the superposition of two (or more) input signals E_1 and E_2 yields a response which is a superposition of the output signals, as schematically represented in Figure 1:

$$R = \alpha_1 R_1 + \alpha_2 R_2 \qquad [2]$$

Now, in almost all real physical systems, if the amplitude of an excitation E_1 becomes sufficiently large, distortions will occur in the output signals. In other words, the response of the system will no longer be proportional to the excitation, and consequently, the law of superposition of states will no longer be observed. In this case, the response of the system may take the following form:

$$R = \alpha_1 E_1 + \alpha_2 E_1^2 + \alpha_3 E_1^3 + \dots \qquad [3]$$

which involves not only a signal at the input frequency ω, but also signals at frequencies 2ω, 3ω, and so on. Thus, harmonics of the input signal are generated. This behavior, called *nonlinear response*, is at the origin of a host of important phenomena in many

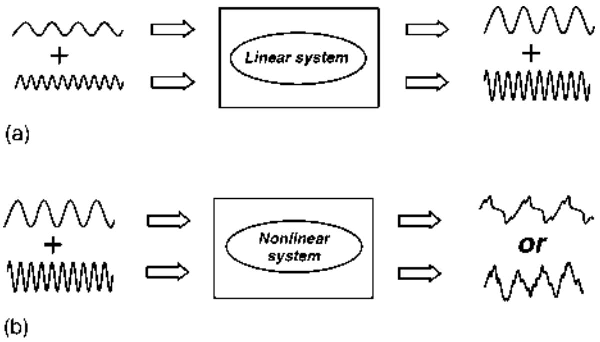

Figure 1 Schematic representation of linear and nonlinear responses of a system, to two input signals.

branches of sciences, such as in condensed matter physics, biological sciences, or in optics.

Physical Origin of Optical Nonlinearity

Optical nonlinearity originates fundamentally from the action of the electric field of a light wave on the charged particles of a dielectric waveguide. In contrast to conductors where charges can move throughout the material, dielectric media consist of *bound* charges (ions, electrons) that can execute only relatively limited displacements around their equilibrium positions. The electric field of an incoming light wave will cause the positive charges to move in the polarization direction of the electric field whereas negative charges will move in the opposite direction. In other words, the electric field will cause the charged particles to become dipoles, as schematically represented in Figure 2.

Then each dipole will vibrate under the influence of the incoming light, thus becoming a source of radiation. The global light radiated by all the dipoles represents the scattered light. When the charge displacements are proportional to the excitation (incoming light), i.e., in the low-amplitude limit of scattered radiation, the output light vibrates at the same frequency as that of the excitation. This process corresponds to Rayleigh scattering. On the other hand, if the intensity of the excitation is sufficiently large to induce displacements that are not negligible with respect to atomic distances, then the charge displacements will no longer be proportional to the excitation. In other words, the response of the medium, which is no longer proportional to the excitation, becomes nonlinear. In this case, the scattered waves will be generated not only at the excitation frequency (the Kerr effect), say ω, but also at frequencies that differ from ω (e.g., 2ω, 3ω). On the other hand, it is worth noting that when all induced dipoles vibrate coherently (that is, their relative phase does not vary randomly), their individual radiation may, under certain conditions, interfere constructively and lead to a global field of high intensity. The condition of constructive interference is the phase-matching condition.

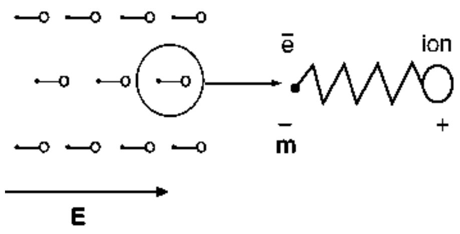

Figure 2 Electric dipoles in a dielectric medium under an external electric field.

In practice, the macroscopic response of a dielectric is given by the polarization, which corresponds to the total amount of dipole moment per unit volume of the dielectric. As the mass of an ion is much larger than that of an electron, the amplitude of the ion motion is generally negligible with respect to that of the electrons. As a consequence, the electron motion generally leads to the dominant contribution in the macroscopic properties of the medium. The behavior of an electron under an optical electric field is similar to that of a particle embedded in an anharmonic potential. A very simple model (called the Lorentz model) that provides a deep insight into the dielectric response consists of an electron of mass m and charge $-e$ connected to an ion by an elastic spring (see Figure 2). Under the electric field $E(t)$, the electron executes a displacement $x(t)$ with respect to its equilibrium position, which is governed by the following equation:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + 2\lambda\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x + (a^{(2)}x^2 + a^{(3)}x^3 + \cdots) = -\frac{e}{m}E(t) \qquad [4]$$

where $a^{(2)}$, $a^{(3)}$, and so on, are constant parameters, ω_0 is the resonance angular frequency of the electron, and λ is the damping coefficient resulting from the dipolar radiation. When the amplitude of the electric field is sufficiently large, then the restoring force on the electrons becomes a nonlinear function of x; hence the presence of terms such as $a^{(2)}x^2$, $a^{(3)}x^3$, and so on, in eqn [4]. In this situation, the macroscopic response of the dielectric is the polarization

$$P = -\sum ex(\omega, 2\omega, 3\omega, \ldots) \qquad [5]$$

where $x(\omega, 2\omega, 3\omega, \ldots)$ is the solution of eqn [4] in the frequency domain, and the summation extends over all the dipole moments per unit volume. In terms of the electric field E the polarization may be written as

$$P = \varepsilon_0(\chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots) \qquad [6]$$

where $\chi^{(1)}$, $\chi^{(2)}$, $\chi^{(3)}$, and so on, represent the susceptibility coefficients. Figure 3 (top left) illustrates schematically the polarization as a function of the electric field.

In particular, one can clearly observe that when the amplitude of the incoming electric field is sufficiently small (bottom left in Figure 3), the polarization (top right) is proportional to the electric field (bottom left), thus implying that the electric dipole radiates a wave having the same frequency as that of the incoming light (bottom right). On the other hand, Figure 4

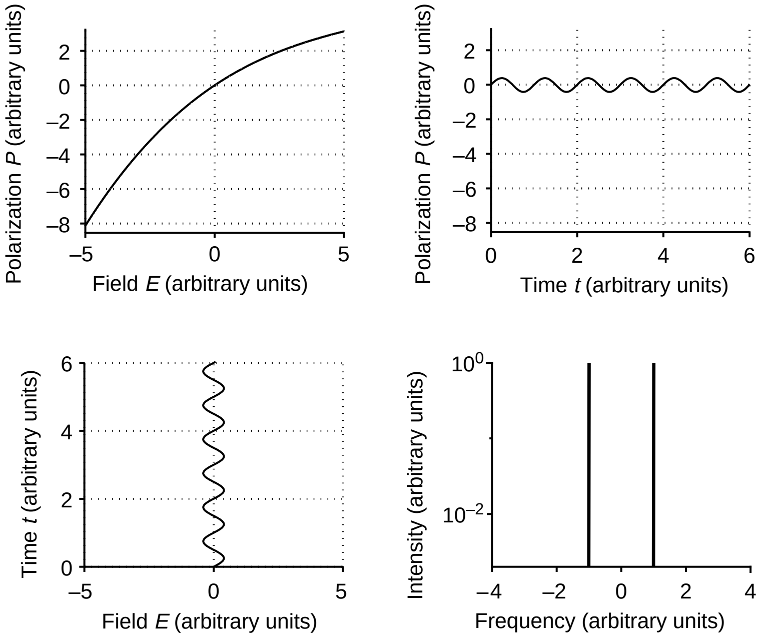

Figure 3 Polarization induced by an electric field of small amplitude. Nonlinear dependence of the polarization as a function of field amplitude (top left) and time dependence of the input electric field (bottom left). Time dependence of the induced polarization (top right) and corresponding intensity spectrum (bottom right).

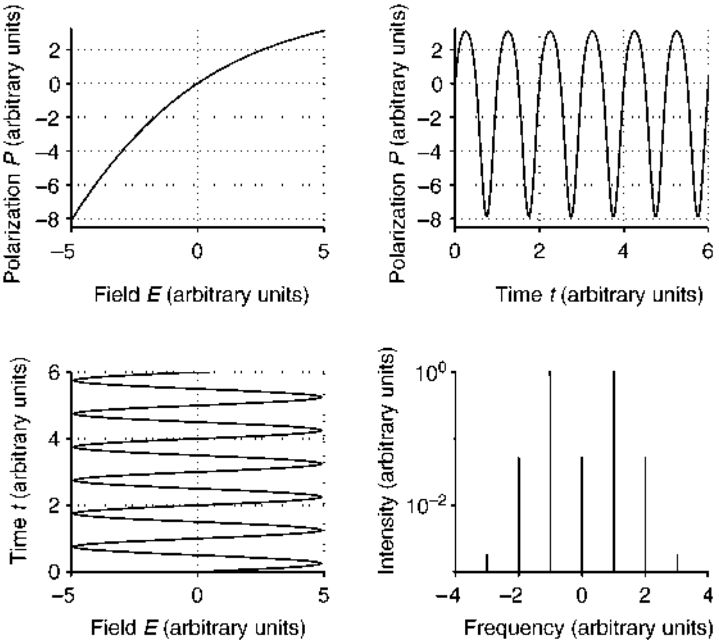

Figure 4 Polarization induced by an incoming electric field of large amplitude.

shows that for an electric field of large amplitude, the polarization is no longer proportional to the electric field, leading to radiation at harmonic frequencies (see Figure 4, bottom right).

This nonlinear behavior leads to a host of important phenomena in optical fibers, which are useful for many optical systems but detrimental for other systems.

Parametric Phenomena in Optical Fibers

In anisotropic materials, the leading nonlinear term in the polarization, i.e., the $\chi^{(2)}$ term, leads to phenomena such as the harmonic generation or optical rectification. This $\chi^{(2)}$ term vanishes in homogeneous isotropic materials such as cylindrical optical fibers, and there the leading nonlinear term becomes the $\chi^{(3)}$ term. Thus, most of outstanding nonlinear phenomena in optical fibers originate from the third-order nonlinear susceptibility $\chi^{(3)}$. Some of those phenomena are described below.

Optical Kerr Effect

The optical Kerr effect is probably the most important nonlinear effect in optical fibers. This effect induces an intensity dependence of the refractive index, which leads to a vast wealth of fascinating phenomena such as self-phase modulation (SPM), cross-phase modulation (CPM), four-wave mixing (FWM), modulational instability (MI) or optical solitons. The Kerr effect can be conveniently described in the frequency domain through a direct analysis of the polarization, which takes the following form:

$$P_{\mathrm{NL}}(\omega) = \frac{3}{4}\varepsilon_0\chi^{(3)}(\omega)|E(\omega)|^2E(\omega) \qquad [7]$$

The constant 3/4 comes from the symmetry properties of the tensor $\chi^{(3)}$. Setting $P_{\mathrm{NL}}(\omega) = \varepsilon_0\varepsilon_{\mathrm{NL}}E(\omega)$, where $\varepsilon_{\mathrm{NL}} = \frac{3}{4}\chi^{(3)}|E|^2$ is the nonlinear contribution to the dielectric constant, the total polarization takes the form

$$P(\omega) = P_{\mathrm{L}} + P_{\mathrm{NL}} = \varepsilon_0[\chi^{(1)}(\omega) + \varepsilon_{\mathrm{NL}}]E(\omega) \qquad [8]$$

As eqn [8] shows, the refractive index n, at a given frequency ω, is given by

$$n^2 = 1 + \chi^{(1)} + \varepsilon_{\mathrm{NL}} = (n_0 + \Delta n_{\mathrm{NL}})^2 \qquad [9]$$

with $n_0^2 = 1 + \chi^{(1)}$. In practice $\Delta n_{\mathrm{NL}} \ll n_0$, and then, the refractive index is given by

$$n(\omega, |E|^2) = n_0(\omega) + n_2^e|E|^2 \qquad [10]$$

where n_2^e is the nonlinear refractive index defined by $n_2^e = 3\chi^{(3)}/(8n_0)$. The linear polarization P_{L} is responsible for the frequency dependence of the refractive index, whereas the nonlinear polarization P_{NL} causes an intensity dependence of the refractive index, which is referred to as the optical Kerr effect. Knowing that the wave intensity I is given by $I = \alpha|E|^2$, with $\alpha = \frac{1}{2}\varepsilon_0 c n_0$, the refractive index can be then rewritten as

$$n(\omega, I) = n_0(\omega) + n_2 I \qquad [11]$$

with $n_2 = n_2^e/\alpha = 2n_2^e/(\varepsilon_0 c n_0)$. For fused silica fibers one has typically: $n_2 = 2.66 \times 10^{-20}\ \mathrm{m^2\ W^{-1}}$. For example, an intensity of $I = 1\ \mathrm{GW\ cm^{-2}}$ leads to $\Delta n_{\mathrm{NL}} = 2.66 \times 10^{-7}$, which is much smaller than $n_0 \approx 1.45$.

Four-Wave Mixing

The four-wave mixing (FWM) process is a third-order nonlinear effect in which four waves interact through an energy exchange process. Let us consider two intense waves, $E_1(\omega_1)$ and $E_2(\omega_2)$, with $\omega_2 > \omega_1$, called pump waves, propagating in an optical fiber. Hereafter we consider the simplest case when waves propagate with the same polarization. In this situation the total electric field is given by

$$\begin{aligned} E_{\mathrm{tot}}(\mathbf{r}, t) &= E_1 + E_2 \\ &= A_1(\omega_1)\exp[\mathrm{i}(\mathbf{k}_1\cdot\mathbf{r} - \omega_1 t)] \\ &\quad + A_2(\omega_2)\exp[\mathrm{i}(\mathbf{k}_2\cdot\mathbf{r} - \omega_2 t)] \end{aligned} \qquad [12]$$

where $\mathbf{k}_1$ and $\mathbf{k}_2$ are the wavevectors of the fields E_1 and E_2, respectively. Equation [7], which gives the nonlinear polarization induced by a single monochromatic wave, remains valid provided that the frequency spacing between the two waves is relatively small, i.e., $|\Delta\omega| = |\omega_2 - \omega_1| \ll \omega_0 = (\omega_1 + \omega_2)/2$. In this context, eqn [7] leads to

$$P_{\mathrm{NL}} \approx \frac{3}{4}\varepsilon_0\chi^{(3)}(\omega_0)|E_{\mathrm{tot}}|^2E_{\mathrm{tot}} \qquad [13]$$

Substituting eqn [12] in eqn [13] yields

$$\begin{aligned} P_{\mathrm{NL}} = 2n_0n_2\varepsilon_0\Big[&\left(|E_1|^2 + 2|E_2|^2\right)E_1 \\ &+ \left(|E_2|^2 + 2|E_1|^2\right)E_2 + E_1^2E_2^* + E_1^*E_2^2\Big] \end{aligned} \qquad [14]$$

The term $|E_1|^2E_1$ in the right-hand side of eqn [14] represents the self-induced Kerr effect on the wave ω_1. The term $|E_2|^2E_1$, which corresponds to a modification of the refractive index seen by the wave ω_1, due to the presence of the wave ω_2, represents the cross-Kerr effect. Thus, the refractive index at frequency ω_1 depends simultaneously on the intensities of the two pumps, i.e., $n = n(\omega_1, |E_1|^2, |E_2|^2)$. Similarly the third and fourth terms in the right-hand side of eqn [14] illustrate the self-induced and cross-Kerr effects on the wave ω_2, respectively. Note that the self-induced Kerr effect is responsible for a self-phase modulation, which induces a spectral broadening of a pulse propagating through the optical fiber, whereas the cross-Kerr effect leads to a cross-phase modulation that induces a spectral broadening of one wave in the presence of a second wave. The two last terms in the right-hand side of eqn [14] correspond to the generation of new waves at frequencies $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, respectively. The wave with the smallest frequency $2\omega_1 - \omega_2 = \omega_{\mathrm{s}}$ is called the Stokes wave, whereas the wave with the highest frequency $2\omega_2 - \omega_1 = \omega_{\mathrm{as}}$ is called the anti-Stokes wave. These two newly generated waves interact

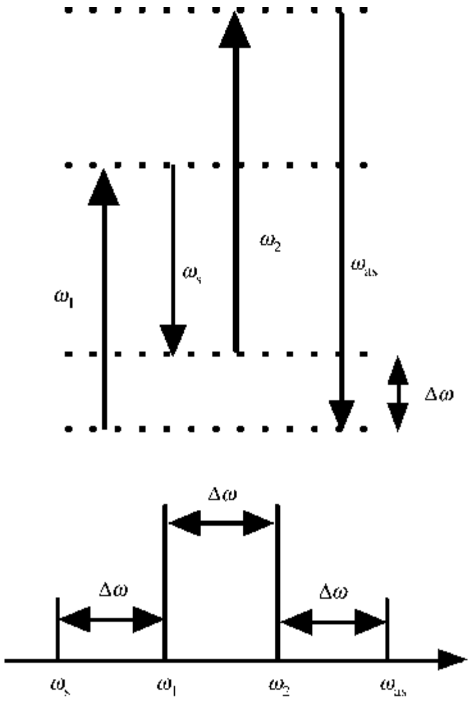

Figure 5 Schematic diagram representing a non-degenerate four-wave mixing process (top) and the corresponding frequency spectrum (bottom).

with the two pumps through an energy exchange process. This interaction is commonly referred to as a four-wave mixing process.

From a quantum mechanical point of view, four-wave mixing corresponds to a process in which two photons with frequencies ω_1 and ω_2 are annihilated with simultaneous creation of two photons at frequencies ω_s and ω_{as}, respectively. The new waves are generated at frequencies ω_s and ω_{as} such that

$$\omega_1 + \omega_2 = \omega_s + \omega_{as} \qquad [15]$$

which states that the total energy is conserved during the interaction. But the condition for this FWM process to occur is that the total momentum is conserved, that is,

$$\Delta\mathbf{k} = \mathbf{k}_s + \mathbf{k}_{as} - \mathbf{k}_1 - \mathbf{k}_2 = 0 \qquad [16]$$

or equivalently,

$$n(\omega_s)\omega_s + n(\omega_{as})\omega_{as} - n(\omega_1)\omega_1 - n(\omega_2)\omega_2 = 0 \qquad [17]$$

Equation [16] is known as the phase-matching condition. Figure 5 displays a schematic diagram of a four-wave mixing process with the corresponding frequency spectrum.

FWM with different pump frequencies $\omega_1 \neq \omega_2$ is called 'nondegenerate FWM', whereas the case $\omega_1 = \omega_2$ is referred to as 'degenerate FWM', or more simply, as three-wave mixing.

Conclusion

Optical fibers constitute a key device for many areas of optical sciences, and in particular for ultrafast optical communications. The nonlinear phenomena that arise through the nonlinear refractive index change (induced mainly by the Kerr effect) have been largely investigated these last two decades for applications such as all-optical wavelength conversion, parametric amplification, generation of new optical frequencies or ultrahigh repetition rate pulse trains. However, in many other optical systems such as optical communication systems, these nonlinear phenomena become harmful effects, and in this case control processes are being developed in order to suppress or at least reduce their impact in the system.

List of Units and Nomenclature

Nonlinear refractive index n_2 [$\mathrm{m^2\,W^{-1}}$]
Optical intensity I [$\mathrm{GW\,cm^{-2} = 10^9\,W\,cm^{-2}}$]

CPM: cross phase modulation
FWM: four wave mixing
MI: modulational instability
SPM: self-phase modulation

See also

Fiber and Guided Wave Optics: Light Propagation; Measuring Fiber Characteristics. **Lasers:** Optical Fiber lasers. **Optical Amplifiers:** Semiconductor Optical Amplifiers. **Optical Communication Systems:** Wavelength Division Multiplexing. **Scattering:** Scattering Theory. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Bloembergen N (1977) *Nonlinear Optics*. Reading, MA: Benjamin.

Boyd RW (1992) *Nonlinear Optics*. San Diego, CA: Academic Press.

Butcher PN and Cotter DN (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Dianov EM, Mamyshev PV, Prokhorov AM and Serkin VN (1989) *Nonlinear Effects in Optical Fibres*. Switzerland: Harwood Academic Publishers.

Hellwarth R (1977) Third-order optical susceptibilities of liquids and s\olids. *Progress in Quantum Electronics* 5: 1–68.

Pocholle JP, Papuchon M, Raffy J and Desurvire E (1990) Non linearities and optical amplification in single mode fibres. *Revue Technique Thomson-CSF* 22: 187–268.

Shen YR (1984) *Principles of Nonlinear Optics*. New York: Wiley.

Nonlinear Optics

K Thyagarajan and A K Ghatak, Indian Institute of Technology, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The invention of the laser provided us with a light source capable of generating extremely large optical power densities (several MW/m^2). At such large power densities, matter behaves in a nonlinear fashion and we come across new optical phenomena such as second-harmonic generation (SHG), sum and difference frequency generation, intensity dependent refractive index, mixing of various frequencies, etc. In SHG, an incident light beam at frequency ω interacts with the medium and generates a new light wave at frequency 2ω. In sum and difference frequency generation, two incident beams at frequencies ω_1 and ω_2 mix with each other producing sum $(\omega_1 + \omega_2)$ and difference $(\omega_1 - \omega_2)$ frequencies at the output. Higher-order nonlinear effects, such as self-phase modulation, four-wave mixing, etc. can also be routinely observed today. The field of nonlinear optics dealing with such nonlinear interactions is gaining importance due to numerous demonstrated applications in many diverse areas such as optical fiber communications, all-optical signal processing, realization of novel sources of optical radiation, etc.

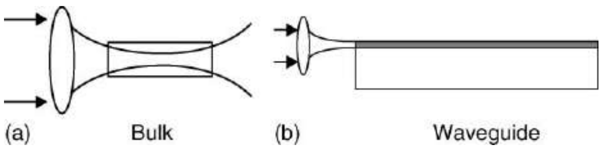

Figure 1 (a) In bulk media, tighter focusing produces larger intensities, but over shorter interaction lengths. (b) In optical waveguides diffraction effects are balanced by waveguiding and the interaction lengths are much larger.

Nonlinear optical interactions become prominent when the optical power densities are high and interaction takes place over long lengths. The usual method to increase optical intensity is to focus the light beam using a lens system. For a given optical power, the tighter the focusing, the larger will be the intensity for a given optical power; however, greater will be the divergence of the beam. Thus tighter focusing produces larger intensities, but over shorter interaction lengths (see Figure 1a). Optical waveguides, in which the light beam is confined to a small cross-sectional area, are currently being explored for realizing efficient nonlinear devices. In contrast to bulk media, in waveguides, diffraction effects are balanced by waveguiding and the beam can have small cross-sectional areas over much longer interaction lengths (see Figure 1b). A simple optical waveguide consists of a high index dielectric medium surrounded by a lower index dielectric medium so that light waves can be trapped in the high index region by the phenomenon of total internal reflection. Figure 2 shows a planar waveguide, a channel waveguide and an optical fiber. In the planar waveguide a film of refractive index n_f is deposited/diffused on a substrate of refractive index n_s and has a cover of refractive index n_c (with n_s, $n_\mathrm{c} < n_\mathrm{f}$). The waveguide has typical cross-sectional dimensions of a few micrometers. Unlike planar waveguides, in which guidance takes place only in one dimension, in channel waveguides the diffused region in a substrate is surrounded on all sides by lower index media. Optical fibers are structures with cylindrical symmetry and have a central cylindrical core of doped silica surrounded by a concentric cylindrical cladding of pure silica which has a slightly lower refractive index. Light guidance in all these waveguides takes place through the phenomenon of total internal reflection.

In contrast to bulk interactions requiring beam focusing, in the case of optical waveguides, the beam can be made to have extremely small cross-sectional

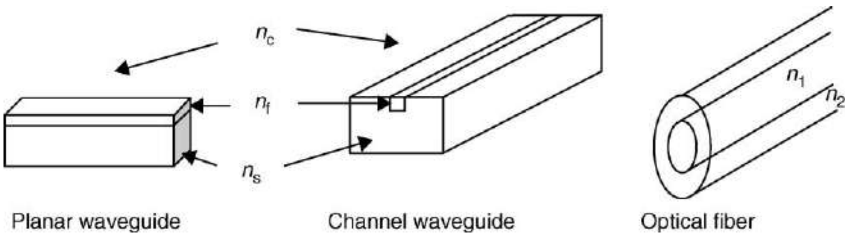

Figure 2 A planar waveguide, a channel waveguide and an optical fiber.

areas (~25 μm^2) over very long interaction lengths (~20 mm in integrated optical waveguides and tens of thousands of kilometers in the case of optical fibers). This leads to very much increased nonlinear interaction efficiencies even at moderate powers (approximately a few tens of mW).

In the following sections, we will discuss some of the important nonlinear interactions that are being studied with potential applications to various branches of science and engineering. Obviously it is impossible to cover all aspects of nonlinear optics – several books have been written in this area (see Further Reading section at the end of this article) – what we will do is to discuss the physics of some of the important nonlinear effects.

Apart from intensity and length of interaction, one of the most important requirements for many efficient nonlinear interactions is the requirement of phase matching. The concept of phase matching can be easily understood from the point of view of SHG. In SHG, the incident wave at frequency ω generates a nonlinear polarization at frequency 2ω and this nonlinear polarization is responsible for the generation of the wave at 2ω. Now, the phase velocity of the nonlinear polarization wave at 2ω is the same as the phase velocity of the electromagnetic wave at frequency ω, which is usually different from the phase velocity of the electromagnetic wave at frequency 2ω; this happens due to wavelength dispersion in the medium. When the two phase velocities are unequal, the polarization wave at 2ω (which is the source) and the electromagnetic wave at 2ω pass in and out of phase with each other as they propagate through the medium. Due to this, the energy flowing in from ω to 2ω cannot add constructively and the efficiency of second-harmonic generation is limited. If the phase velocities of the waves at ω and 2ω are matched then the polarization wave and the wave at 2ω remain in phase leading to drastically increased efficiencies. This condition is referred to as phase matching and plays a very important role in most nonlinear interactions.

Nonlinear Polarization

In a linear medium, the electric polarization P is assumed to be a linear function of the electric field E:

$$P = \varepsilon_0 \chi E \qquad [1]$$

where, for simplicity, a scalar relation has been written. The quantity χ is termed as linear dielectric susceptibility. At high optical intensities (which corresponds to high electric fields), all media behave in a nonlinear fashion. Thus eqn [1] is modified to

$$P = \varepsilon_0(\chi E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \dots) \qquad [2]$$

where $\chi^{(2)}, \chi^{(3)}, \dots$ are higher order susceptibilities giving rise to the nonlinear terms. The second term on the right-hand side is responsible for SHG, sum and difference frequency generation, parametric interactions, etc. while the third term is responsible for third-harmonic generation, intensity dependent refractive index, self-phase modulation, four-wave mixing, etc. For media possessing an inversion symmetry, $\chi^{(2)}$ is zero and there is no second-order nonlinear effect. Thus silica optical fibers, which form the heart of today's communication networks, do not possess the second-order nonlinearity.

We will first discuss second-harmonic generation and parametric amplification which arises due to the second-order nonlinear term and then go on to effects due to third-order nonlinear interaction.

Second-Harmonic Generation in Crystals

The first demonstration of SHG was made in 1961 by focusing a 3 kW ruby laser pulse ($\lambda_0 = 6943$ Å) on a quartz crystal. An incident beam from a ruby laser (red color) after passing through a crystal of KDP, gets partly converted into blue light which is the second harmonic. Ever since, SHG has been one of the most widely studied nonlinear interactions.

We consider a plane wave of frequency ω propagating along the z-direction through a medium and consider the generation of the second-harmonic frequency 2ω as the beam propagates through the medium. Now, the field at ω generates a polarization at 2ω, which acts as a source for the generation of an electromagnetic wave at 2ω. Corresponding to the frequencies ω and 2ω, the electric fields are assumed to be given by

$$E^{(\omega)} = \frac{1}{2}(E_1(z)\mathrm{e}^{\mathrm{i}(\omega t - k_1 z)} + \text{c.c.}) \qquad [3]$$

and

$$E^{(2\omega)} = \frac{1}{2}(E_2(z)\mathrm{e}^{\mathrm{i}(2\omega t - k_2 z)} + \text{c.c.}) \qquad [4]$$

respectively; c.c. stands for the complex conjugate of the preceding quantities. The quantities:

$$k_1 = \omega\sqrt{(\varepsilon_1 \mu_0)} = (\omega/c)n^{\omega} \qquad [5]$$

and

$$k_2 = 2\omega\sqrt{(\varepsilon_2 \mu_0)} = (2\omega/c)n^{2\omega} \qquad [6]$$

represent the propagation vectors at ω and 2ω, respectively; ε_1 and ε_2 represent the dielectric permittivities at ω and 2ω, and n^{ω} and $n^{2\omega}$ represent the corresponding wave refractive indices. It should be noted that the amplitudes E_1 and E_2 are assumed to be z dependent – this is because at $z = 0$ (where the beam is incident on the medium) the amplitude E_2 is zero and this would develop as the beam propagates through the medium. We will now develop an approximate theory for the generation of the second harmonic.

We start with the wave equation:

$$\nabla^2 E - \varepsilon\mu_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial^2 P_{\mathrm{NL}}}{\partial t^2} \quad [7]$$

where P_{NL} is the nonlinear polarization. An incident wave at frequency ω generates a polarization at 2ω, which acts as a source for the generation of an electromagnetic wave at 2ω.

In order to consider SHG, we write the wave equation corresponding to 2ω with the nonlinear polarization at 2ω given by

$$P_{\mathrm{NL}}^{(2\omega)} = \frac{1}{2}(\tilde{P}_{\mathrm{NL}}^{(2\omega)}\,\mathrm{e}^{2\,\mathrm{i}(\omega t - k_1 z)} + \mathrm{c.c.}) \quad [8]$$

where

$$\tilde{P}_{\mathrm{NL}}^{(2\omega)} = dE_1E_1 \quad [9]$$

and

$$d = \frac{\varepsilon_0\chi^{(2)}}{2} \quad [10]$$

represents the effective nonlinear coefficient and depends on the nonlinear material, the polarization states of the fundamental and the second harmonic and also on the propagation direction. Simple manipulations give us, for the rate of change of amplitude of the second harmonic wave:

$$\frac{\mathrm{d}E_2}{\mathrm{d}z} = -\frac{i\mu_0\, dc\omega}{n^{2\omega}} E_1^2(z)\mathrm{e}^{\mathrm{i}(\Delta k)z} \quad [11]$$

where

$$\Delta k = k_2 - 2k_1 = (2\omega/c)(n^{2\omega} - n^{\omega}) \quad [12]$$

and we have assumed:

$$\mathrm{d}^2E_2/\mathrm{d}z^2 \ll k_2(\mathrm{d}E_2/\mathrm{d}z) \quad [13]$$

In order to solve eqn [11] we neglect depletion of the fundamental field, i.e., $E_1(z)$ is almost a constant and the quantity E_1^2 on the right-hand side can be assumed to be independent of z. If we now integrate eqn [11], we obtain:

$$E_2(z) = -\frac{i\mu_0\, \mathrm{d}c\omega}{n^{2\omega}} zE_1^2\mathrm{e}^{\mathrm{i}\sigma}\frac{\sin\sigma}{\sigma} \quad [14]$$

where

$$\sigma = \frac{1}{2}(\Delta k)z = (\omega/c)(n^{2\omega} - n^{\omega})z \quad [15]$$

Now, the powers associated with the beams corresponding to ω and 2ω are given by:

$$P_1 = \frac{n^{\omega}}{2c\mu_0}S|E_1|^2, \qquad P_2 = \frac{n^{2\omega}}{2c\mu_0}S|E_2|^2 \quad [16]$$

where S represents the cross-sectional area of the beams. Substituting for $|E_2|^2$ from eqn [16], we get after some elementary simplifications:

$$\eta = \frac{P_2}{P_1} = \frac{2c^3\mu_0^3d^2\omega^2}{(n^{\omega})^2n^{2\omega}} z^2 \frac{P_1}{S}\left(\frac{\sin\sigma}{\sigma}\right)^2 \quad [17]$$

when η represents the SHG efficiency. Note that the efficiency of SHG increases if P_1/S, the intensity of the fundamental wave increases. Also η increases if the nonlinear coefficient d increases (obviously) and if the frequency increases. However, for a given power P_1, the conversion efficiency increases if the area of the beam decreases – thus a focused beam will have a greater SHG efficiency. The most important factor is $(\sin\sigma/\sigma)^2$ which is a sharply peaked function around $\sigma = 0$, attaining a maximum value of unity at $\sigma = 0$. Thus for maximum SHG efficiency:

$$\sigma = 0 \Rightarrow n^{2\omega} = n^{\omega} \quad [18]$$

i.e., the refractive index at 2ω must be equal to the refractive index at ω – this is known as the phase matching condition.

The phase matching condition can be pictorially represented by a vector diagram (see Figure 3). In Figure 3, $\mathbf{k}_1\omega$ and $\mathbf{k}_2\omega$ represent the wave vectors of the fundamental and the second harmonic respectively. To achieve reasonably effective SHG, it is very important to satisfy the phase-matching condition. Efficiencies under nonphase matched operation can be orders of magnitude lower than under phase-matched operation.

We see from eqn [17] that the smallest z for which η is maximum is:

$$z = L_{\mathrm{c}} = \frac{\pi}{\Delta k} = \frac{\pi c}{2\omega(n^{2\omega} - n^{\omega})} \quad [19]$$

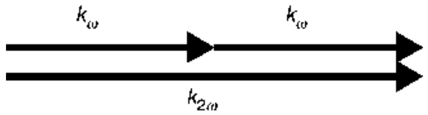

Figure 3 Vector diagram representing the phase matching condition for SHG.

where we have used eqn [15] for Δk. The length L_c is called the phase coherence length and represents the maximum crystal length up to which the second-harmonic power increases. Thus, if the length of the crystal is less than L_c, the second-harmonic power increases almost quadratically with z. For $z > L_c$, the second-harmonic power begins to reduce again.

In general, because of dispersion, it is very difficult to satisfy the phase-matching condition. However, in a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the *o*-wave for ω equals the refractive index of the *e*-wave for 2ω (see Figure 4). For media with $n_e > n_o$, the direction would correspond to that along which the refractive index of the *e*-wave for ω equals the refractive index for the *o*-wave for 2ω. This can readily be understood by considering a specific example. We consider the SHG in KDP corresponding to the incident ruby laser wavelength ($\lambda_0 = 0.6943\ \mu\text{m}$, $\omega = 2.7150 \times 10^{15}$ Hz). For this frequency:

$$\left\{\begin{matrix} n_o^{\omega} = 1.50502, & n_e^{\omega} = 1.46532 \\ n_o^{2\omega} = 1.53269, & n_e^{2\omega} = 1.48711 \end{matrix}\right\} \qquad [20]$$

The refractive index variation for the *e*-wave is given by:

$$n_e(\theta) = \left(\frac{\sin^2\theta}{n_e^2} + \frac{\cos^2\theta}{n_o^2} \right)^{-\frac{1}{2}} \qquad [21]$$

where θ is the angle that the wave propagation direction makes with the optic axis. Now, as can be seen from the above equations:

$$n_e < n_e(\theta) < n_o \qquad [22]$$

Since n_o at $0.6943\ \mu\text{m}$ lies between n_e and n_o at $0.34715\ \mu\text{m}$, there will always exist an angle θ_m along which $n_e^{2\omega}(\theta_m) = n_o^{\omega}$. We can solve for θ_m and obtain:

$$\cos\theta_m = \left[\frac{(n_o^{\omega})^2 - (n_e^{2\omega})^2}{(n_o^{2\omega})^2 - (n_e^{2\omega})^2} \right]^{1/2} \frac{n_o^{2\omega}}{n_o^{\omega}} \qquad [23]$$

For the values given by eqn [20], we find $\theta_m = 50.5°$.

Quasi Phase Matching (QPM)

As mentioned earlier, phase matching is extremely important for any efficient nonlinear interaction. Recently the technique of quasi phase matching (QPM) has become a convenient way to achieve phase matching at any desired wavelength in any material. QPM relies on the fact that the phase mismatch in the two interacting beams can be compensated by periodically readjusting the phase of interaction through periodically modulating the nonlinear characteristics of the medium at a spatial frequency equal to the wavevector mismatch of the two interacting waves. Thus in SHG, when the nonlinear polarization at 2ω and the electromagnetic wave at 2ω have an accumulated phase difference of π, then the sign of the nonlinear coefficient is reversed so that the energy flowing from the polarization to the wave can add constructively with the existing energy (see Figure 5). Thus, by properly choosing the period of the spatial modulation of the nonlinear coefficient, one could achieve phase matching. This scheme, QPM, is being very widely studied for application to nonlinear interactions in bulk and in waveguides.

In a ferroelectric material such as lithium niobate, the signs of the nonlinear coefficients are linked to the direction of the spontaneous polarization. Thus a

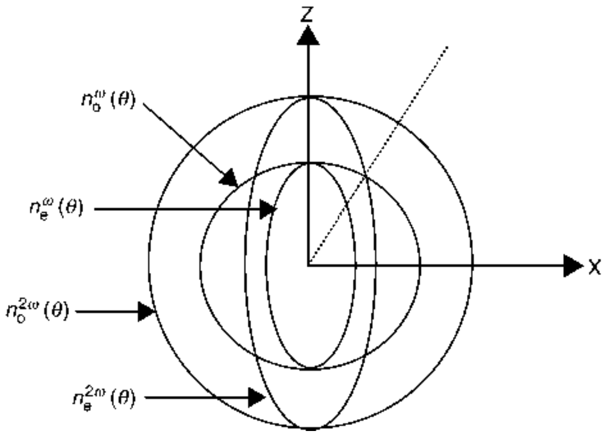

Figure 4 In a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the *o*-wave for ω equals the refractive index of the *e*-wave for 2ω.

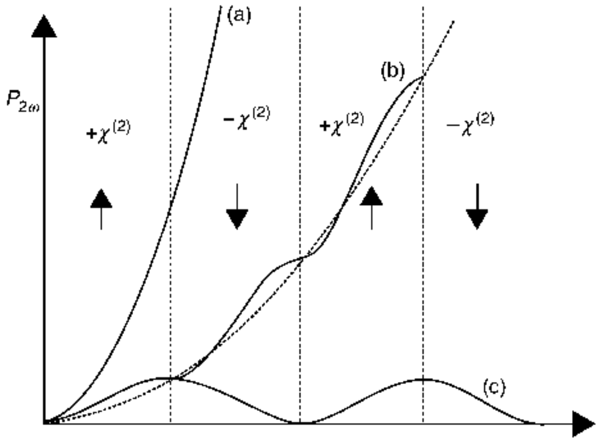

Figure 5 By reversing the sign of the nonlinear coefficient after every coherence length, the energy in the second harmonic can be made to grow. (a) Perfect phase matching; (b) Quasi phase matching; (c) Nonphase matched.

periodic reversal of the domains of the crystal can be used for achieving QPM (see Figure 6). This is the currently used technique to obtain high-efficiency SHG and other nonlinear interactions in $LiNbO_3$, $LiTaO_3$, and KTP. The most popular technique today, to achieve periodic domain reversal in $LiNbO_3$, is the technique of electric field poling. In this method, a high electric field pulse is applied to properly oriented lithium niobate crystal using lithographically defined electrode patterns to produce a permanent periodic domain reversed pattern. Such a periodically domain reversed $LiNbO_3$ crystal with the periodically reversed domains going through the entire depth of the crystal is also referred to as PPLN (periodically poled lithium niobate, pronounced 'piplin') and is now commercially available.

In order to analyze SHG in a periodically poled material, let us assume that the nonlinear coefficient d varies sinusoidally with a period Λ. In such a case we have:

$$d = d_0 \sin(Kz) \qquad [24]$$

where d_0 is the amplitude of modulation of the nonlinear coefficient and $K(= 2\pi/\Lambda)$ represents the spatial frequency of the periodic modulation. For easier understanding we are assuming the modulation to be sinusoidal; in general, the modulation will be periodic but not sinusoidal. Any periodic modulation can be written as a superposition of sinusoidal and cosinusoidal variations. Thus our discussion is valid for one of the Fourier components of the variation.

By using eqn [24], eqn [11] for the variation of the amplitude of the second harmonic becomes:

$$\frac{dE_2}{dz} = -\frac{i\mu_0 d_0 c\omega}{n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} \sin(Kz) \qquad [25]$$

which can be written as:

$$\frac{dE_2}{dz} = -\frac{\mu_0 d_0 c\omega}{2n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} [e^{iKz} - e^{-iKz}] \qquad [26]$$

Using similar arguments as earlier, it can be shown that if $\Delta k - K = 0$, then only the second term within the brackets in eqn [26] contributes to the growth of the second harmonic and similarly if $\Delta k + K = 0$, then only the first term within the brackets contributes to the growth of the second harmonic. The first condition implies that:

$$2\frac{2\pi}{\lambda_0}(n^{2\omega} - n^{\omega}) - \frac{2\pi}{\Lambda} = 0 \qquad [27]$$

where $\Lambda = 2\pi/K$ represents the spatial period of the modulation of the nonlinear coefficient, and λ_0 is the wavelength of the fundamental. Thus the modulation period Λ required for QPM SHG is:

$$\Lambda = \frac{\lambda_0}{2(n^{2\omega} - n^{\omega})} = 2\,L_c \qquad [28]$$

Figure 7 shows the vector diagram corresponding to QPM SHG. The phase mismatch between the fundamental and second harmonic is compensated by the spatial frequency vector of the periodic variation of d.

In the case of waveguides, n^{ω} and $n^{2\omega}$ would represent the effective indices of the modes at the fundamental and the second-harmonic frequency. It may be noted that the index difference between the fundamental and the second harmonic is typically 0.1 and thus the required spatial periods for a fundamental wavelength of 800 nm is ~4 μm.

The main advantage of QPM is that it can be used at any wavelength within the transparency window of the material; only the period needs to be correctly chosen for a specific fundamental wavelength. One can also choose appropriate polarization to make use of the largest nonlinear optical coefficients. Another advantage is the possibility of varying the domain reversal period (chirp) to achieve specific interaction characteristics such as increased bandwidth, etc.

In general, the spatial variation of the nonlinear grating is not sinusoidal. In this case the efficiency of interaction would be determined by the Fourier component of the spatial variation at the spatial frequency corresponding to the period given by eqn [28]. Also, since the required periods are very small, it is possible to use a higher spatial period of modulation and use one of the Fourier components for the nonlinear interaction process. Thus, in the case of periodic reversal of the nonlinear coefficient

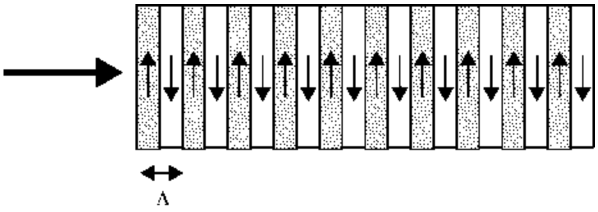

Figure 6 QPM can be achieved by a periodic reversal of the domains of a ferroelectric material. Arrows represent the direction of the spontaneous polarization of the crystal.

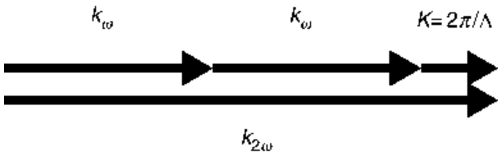

Figure 7 Vector diagram corresponding to QPM-SHG.

with a spatial period given by:

$$\Lambda_g = m\frac{\lambda_0}{2(n^{2\omega} - n^{\omega})}; \qquad m = 1, 3, 5, \ldots \qquad [29]$$

which happens to be the mth harmonic of the fundamental spatial frequency required for QPM, the corresponding nonlinear coefficient that would be responsible for SHG would be the Fourier amplitude at that spatial frequency. This can be taken into account by defining an effective nonlinear coefficient:

$$d_{\text{QPM}} = \frac{2d_0}{m\pi} \qquad [30]$$

Of course the largest effective nonlinear coefficient is achieved by using the fundamental frequency with $m = 1$. Higher spatial periods are easier to fabricate but would lead to reduced nonlinear efficiencies.

As compared to bulk devices, in the case of waveguides, the interacting waves are propagating in the form of modes having specific intensity distributions in the transverse plane. Because of this, the nonlinear interaction also depends on the overlap between the periodically inverted nonlinear medium, the fields of the fundamental and second-harmonic waves. Thus if $E_\omega(x, y)$ and $E_{2\omega}(x, y)$ represent the electric field distributions of the waveguide modes at the fundamental and second-harmonic frequency, then the efficiency of SHG depends on the following overlap integral:

$$I_{\text{ovl}} = \int\int d_0(x, y)E_\omega^2(x, y)E_{2\omega}(x, y)\,\mathrm{d}x\,\mathrm{d}y \qquad [31]$$

where $d_0(x, y)$ represents the transverse variation of the nonlinear coefficient of the waveguide. Thus optimization of SHG efficiency in the case of waveguide interactions has to take account of the overlap integral.

Since QPM relies on periodic phase matching, it is highly wavelength dependent. Thus the required period Λ is different for different fundamental frequencies. Any deviation in the frequency of the fundamental would lead to a reduction in the efficiency due to deviation from QPM condition. Thus the pump laser needs to be highly stabilized at a frequency corresponding to the fabricated period.

Apart from SHG, QPM has been used for other three-wave interaction processes such as difference frequency generation, parametric amplification, etc. Among the many materials that have been studied for SHG using QPM, the important ones are lithium niobate, lithium tantalite, and potassium titanyl phosphate (KTP). Many techniques have been developed for periodic domain reversal to achieve a periodic variation in the nonlinear coefficient. This includes electric field poling, electron bombardment, thermal diffusion treatment, etc.

Third-Order Nonlinear Effects

In the earlier sections, we discussed effects arising out of second-order nonlinearity, i.e., the term proportional to E^2 in the nonlinear polarization. This nonlinear term is found only in media not possessing an inversion symmetry. Thus amorphous materials or crystals possessing inversion symmetry do not exhibit second-order effects. The lowest-order nonlinear effects in such a medium is of an order three wherein the nonlinear polarization is proportional to E^3.

Self-phase modulation (SPM), cross-phase modulation (XPM) and four-wave mixing (FWM) represent some of the very important consequences of third-order nonlinearity. These effects have become all the more important as they play a significant and important role in wavelength division multiplexed optical fiber communication systems.

Self-Phase Modulation (SPM)

Consider the propagation of a plane light wave at frequency ω through a medium having $\chi^{(3)}$ nonlinearity. The polarization generated in the medium is given by

$$P = \varepsilon_0\chi E + \varepsilon_0\chi^{(3)}E^3 \qquad [32]$$

If we consider a single frequency wave with an electric field given by

$$E = E_0\cos(\omega t - kz) \qquad [33]$$

then

$$P = \varepsilon_0\chi E_0\cos(\omega t - kz) + \varepsilon_0\chi^{(3)}E_0^3\cos^3(\omega t - kz) \qquad [34]$$

Expanding $\cos^3\theta$ in terms of $\cos\theta$ and $\cos 3\theta$, we obtain the following expression for the polarization at frequency ω:

$$P = \varepsilon_0\left(\chi + \frac{3}{4}\chi^{(3)}E_0^2\right)E_0\cos(\omega t - kz) \qquad [35]$$

For a plane wave given by eqn [33], the intensity is

$$I = \frac{1}{2}c\varepsilon_0 n_0 E_0^2 \qquad [36]$$

where n_0 is the refractive index of the medium at low intensity. Then

$$P = \varepsilon_0\left(\chi + \frac{3}{2}\frac{\chi^{(3)}}{c\varepsilon_0 n_0}I\right)E \qquad [37]$$

The polarization P and electric field are related through the following equation:

$$P = \varepsilon_0(n^2 - 1)E \qquad [38]$$

where n is the refractive index of the medium. Comparing eqns [37] and [38], we get

$$n^2 = n_0^2 + \frac{3}{2}\frac{\chi^{(3)}}{c\varepsilon_0 n_0} I \qquad [39]$$

where

$$n_0^2 = 1 + \chi \qquad [40]$$

Since the last term in the eqn [39] is usually very small, we get

$$n = n_0 + n_2 I \qquad [41]$$

where

$$n_2 = \frac{3}{4}\frac{\chi^{(3)}}{c\varepsilon_0 n_0^2} \qquad [42]$$

is the nonlinear coefficient.

For fused silica $n_0 \approx 1.47$, $n_2 \approx 3.2 \times 10^{-20}$ m^2/W and if we consider power of 100 mW having a cross-sectional area of 100 μm^2, the resultant intensity is 10^9 W/m^2 and the corresponding change in refractive index is

$$\Delta n \approx n_2 I \approx 3.2 \times 10^{-11} \qquad [43]$$

Although this is very small, when the beam propagates over an optical fiber over long distances (a few hundred to a few thousand kilometers), the accumulated nonlinear effects can be significant.

In the case of an optical fiber, the light beam propagates as a mode having a specific transverse electric field distribution and thus the intensity is not constant across the cross-section. In such a case, it is convenient to express the nonlinear effect in terms of the power carried by the mode (rather than in terms of intensity). If the linear propagation constant of the mode is represented by β, then in the presence of nonlinearity, the effective propagation constant is given by

$$\beta_{\mathrm{NL}} = \beta + \frac{k_0 n_2}{A_{\mathrm{eff}}} P \qquad [44]$$

where $k_0 = 2\pi/\lambda_0$, P is the power carried by the mode. The quantity A_{eff} represents the effective transverse cross-sectional area of the mode and is defined by

$$A_{\mathrm{eff}} = \frac{\left[\int_0^\infty \int_0^{2\pi} \psi^2(r) r \, \mathrm{d}r \, \mathrm{d}\phi\right]^2}{\int_0^\infty \int_0^{2\pi} \psi^4(r) r \, \mathrm{d}r \, \mathrm{d}\phi} \qquad [45]$$

where $\psi(r)$ represents the transverse mode field distribution of the mode. For example, under the Gaussian approximation:

$$\psi(r) = \psi_0 \, \mathrm{e}^{-r^2/w_0^2} \qquad [46]$$

where ψ_0 is a constant and $2w_0$ represents the mode field diameter (MFD), we get:

$$A_{\mathrm{eff}} = \pi w_0^2 \qquad [47]$$

It is usual to express the nonlinear characteristic of an optical fiber by the coefficient given by

$$\gamma = \frac{k_0 n_2}{A_{\mathrm{eff}}} \qquad [48]$$

Thus, for the same input power and same wavelength, smaller values of A_{eff} lead to greater nonlinear effects in the fiber. Typically:

$$\begin{aligned} &A_{\mathrm{eff}} \sim 50\text{–}80\ \mu\mathrm{m}^2 \quad \text{and} \\ &\gamma \approx 2.4\ \mathrm{W}^{-1}\,\mathrm{km}^{-1} \ \text{to}\ 1.5\ \mathrm{W}^{-1}\,\mathrm{km}^{-1} \end{aligned} \qquad [49]$$

When a light beam propagates through an optical fiber, the power decreases because of attenuation. Thus, the corresponding nonlinear effects also reduce. Indeed, the phase change suffered by a beam in propagating from 0 to L is given by

$$\phi = \int_0^L \beta_{\mathrm{NL}} \, \mathrm{d}z = \beta L + \gamma \int_0^L P \, \mathrm{d}z \qquad [50]$$

If α represents the attenuation coefficient, then

$$P(z) = P_0 \, \mathrm{e}^{-\alpha z} \qquad [51]$$

and we get

$$\phi = \beta L + \gamma P_0 L_{\mathrm{eff}} \qquad [52]$$

where

$$L_{\mathrm{eff}} = \frac{1 - \mathrm{e}^{-\alpha L}}{\alpha} \qquad [53]$$

is referred to as the effective length. For $\alpha L \gg 1$, $L_{\mathrm{eff}} \approx 1/\alpha$ and for $\alpha L \ll 1$, $L_{\mathrm{eff}} \approx L$.

The effective length represents the length of the fiber over which most of the nonlinear effects has accumulated. For a loss coefficient of 0.20 dB/km, $L_{\mathrm{eff}} \approx 21$ km.

If we consider a fiber length much longer than L_{eff}, then to have reduced impact of SPM, we must have

$$\gamma P_0 L_{\mathrm{eff}} \ll 1 \qquad [54]$$

or

$$P_0 \ll \frac{1}{\gamma L_{\mathrm{eff}}} \approx \frac{\alpha}{\gamma} \qquad [55]$$

For $\alpha = 4.6 \times 10^{-2}$ km^{-1} (which corresponds to an attenuation of 0.2 dB/km) and $\gamma = 2.4$ W^{-1} km^{-1}, we get

$$P_0 \ll 19\ \mathrm{mW} \qquad [56]$$

Propagation of a Pulse

When an optical pulse propagates through a medium, it suffers from the following effects:

(i) attenuation;
(ii) dispersion; and
(iii) nonlinearity.

Attenuation refers to the reduction in the pulse energy due to various mechanisms, such as scattering, absorption, etc. Dispersion is caused by the fact that a light pulse consists of various frequency components and each frequency component travels at a different group velocity. Dispersion causes the temporal width of the pulse to change; in most cases it results in an increase in pulse width, however, in some cases the temporal width could also decrease. Dispersion is accompanied by chirping, the variation of the instantaneous frequency of the pulse within the pulse duration. Since both attenuation and dispersion cause a change in the temporal variation of the optical power, they closely interact with nonlinearity in deciding the pulse evolution as it propagates through the medium.

Let $E(x, y, z, t)$ represent the electric field variation of an optical pulse. It is usual to express E in the following way:

$$E(x, y, z, t) = \frac{1}{2}[A(z, t)\psi(x, y)\mathrm{e}^{\mathrm{i}(\omega_0 t - \beta_0 z)} + \text{c.c.}] \quad [57]$$

where $A(z, t)$ represents the slowly varying complex envelope of the pulse, $\psi(x, y)$ represents the transverse electric field distribution, ω_0 represents center frequency, and β_0 represents the propagation constant at ω_0.

In the presence of attenuation, second-order dispersion and third-order nonlinearity, the complex envelope $A(z, t)$ can be shown to satisfy the following equation:

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A - \beta_1\frac{\partial A}{\partial t} + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial t^2} - i\gamma|A|^2 A \quad [58]$$

Here

$$\beta_1 = \left|\frac{\mathrm{d}\beta}{\mathrm{d}\omega}\right|_{\omega=\omega_0} = \frac{1}{v_g} \quad [59]$$

represents the inverse of the group velocity of the pulse, and

$$\beta_2 = \left|\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right|_{\omega=\omega_0} = -\frac{\lambda_0^2}{2\pi c}D \quad [60]$$

where D represents the group velocity dispersion (measured in ps/km nm).

The various terms on the RHS of the eqn [58] represent the following:

I term:	attenuation
II term:	group velocity term
III term:	second-order dispersion
IV term:	nonlinear term

If we change to a moving frame defined by coordinates $T = t - \beta_1 z$, eqn [58] becomes

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} - i\gamma|A|^2 A \quad [61]$$

If we neglect the attenuation term, we obtain the following equation which is also referred to as the nonlinear Schrödinger equation:

$$\frac{\partial A}{\partial z} = i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} - i\gamma|A|^2 A \quad [62]$$

The above equation has a solution given by

$$A(z, t) = A_0 \operatorname{sech} \sigma T\, \mathrm{e}^{-\mathrm{i}gz} \quad [63]$$

with

$$A_0^2 = -\frac{\beta_2}{\gamma}\sigma^2, \qquad g = -\frac{\sigma^2}{2}\beta_2 \quad [64]$$

Equation [63] represents an envelope soliton and has the property that it propagates undispersed through the medium. The full width at half maximum (FWHM) of the pulse envelope will be given by $\tau_f = 2\tau_0$, where

$$\operatorname{sech}^2 \sigma\tau_0 = \frac{1}{2} \quad [65]$$

which gives the FWHM τ_f:

$$\tau_f = 2\tau_0 = \frac{2}{\sigma}\ln(1 + \sqrt{2}) \approx \frac{1.7627}{\sigma} \quad [66]$$

The peak power of the pulse is:

$$P_0 = |A_0|^2 = \frac{|\beta_2|}{\gamma}\sigma^2 \quad [67]$$

Replacing σ by τ_f, we obtain

$$P_0\tau_f^2 \approx \frac{\lambda_0^2}{2c\gamma}D \quad [68]$$

where we have used eqn [60]. The above equation gives the required peak for a given τ_f for the formation of a soliton pulse.

As an example, we have $\tau_f = 10$ ps, $\gamma = 2.4\ \mathrm{W}^{-1}\ \mathrm{km}^{-1}$, $\lambda_0 = 1.55\ \mu\mathrm{m}$, $D = 2$ ps/km nm and the required peak power will be $P_0 = 33$ mW.

Soliton pulses are being extensively studied for application to long distance optical fiber communication. In actual systems, the pulses have to be optically amplified at regular intervals to compensate for the loss suffered by the pulses. The amplification could be carried out using erbium doped fiber amplifiers (EDFAs) or fiber Raman amplifiers.

Spectral Broadening due to SPM

In the presence of only nonlinearity, eqn [61] becomes

$$\frac{dA}{dz} = -i\gamma|A|^2 A \qquad [69]$$

whose solution is given by

$$A(z,t) = A(z=0,t)e^{-i\gamma Pz} \qquad [70]$$

where $P = |A|^2$ is the power in the pulse. If P is a function of time, then time dependent phase term at $z = L$ becomes

$$e^{i\phi(t)} = e^{i[\omega_0 t - \gamma P(t)L]} \qquad [71]$$

We can define an instantaneous frequency as

$$\omega(t) = \frac{d\phi}{dt} = \omega_0 - \gamma L \frac{dP}{dt} \qquad [72]$$

for a Gaussian pulse:

$$P = P_0 \, e^{-2T^2/\tau_0^2} \qquad [73]$$

giving

$$\omega(t) = \omega_0 + \frac{4\gamma LTP_0 \, e^{-2T^2/\tau_0^2}}{\tau_0^2} \qquad [74]$$

Thus the instantaneous frequency within the pulse changes with time, leading to chirping of the pulse (see Figure 8). Note that since the pulsewidth has not changed, but the pulse is chirped, the frequency spectrum of the pulse has increased. Thus SPM leads to the generation of new frequencies. By Fourier transform theory, an increased spectral width implies that the pulse can now be compressed in the temporal domain by passing it through a medium with the proper sign of dispersion. This is indeed one of the standard techniques to realize ultrashort femtosecond optical pulses.

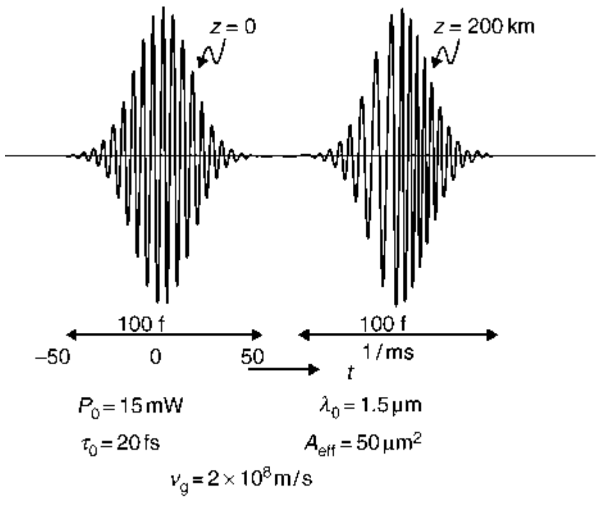

Figure 8 Due to self phase modulation, the instantaneous frequency within the pulse changes with time leading to chirping of the pulse.

Cross Phase Modulation (XPM)

Like SPM, cross-phase modulation also arises due to the intensity dependence of refractive index, leading to spectral broadening. Unlike SPM, in the case of XPM, intensity variations of a light beam at a particular frequency modulate the phase of light beam at another frequency. If the signals at both frequencies are pulses, then due to difference in group velocities of the pulses, there is a walk off between the two pulses, i.e., if they start together, they will separate as they propagate through the medium. Nonlinear interaction takes place as long as they physically overlap in the medium. Smaller the dispersion, smaller will be the difference in group velocities (assuming closely spaced wavelengths) and the longer they will overlap. This would lead to stronger XPM effects. At the same time, if two pulses pass through each other, then since one pulse will interact with both the leading and the trailing edge of the other pulse, XPM effects will be nil provided there is no attenuation. In the presence of attenuation in the medium, the pulse will still get modified due to XPM. A similar situation can occur when the two interacting pulses are passing through an optical amplifier.

To study XPM, we assume simultaneous propagation of two waves at two different frequencies through the medium. If ω_1 and ω_2 represent the two frequencies, then one obtains for the variation of the amplitude A_1 of the frequency ω_1:

$$\frac{dA_1}{dz} = -i\gamma(\tilde{P}_1 + 2\tilde{P}_2)A_1 \qquad [75]$$

where $\tilde{P}_1$ and $\tilde{P}_2$ represent the powers at frequencies ω_1 and ω_2, respectively. The first term in eqn [75] represents SPM while the second term corresponds to XPM. If the powers are assumed to attenuate at the same rate, i.e.:

$$\tilde{P}_1 = P_1 \, e^{-\alpha z}, \qquad \tilde{P}_2 = P_2 \, e^{-\alpha z} \qquad [76]$$

then the solution of eqn [75] is

$$A_1(L) = A_1(0) \, e^{-i\gamma(P_1 + 2P_2)L_{eff}} \qquad [77]$$

where, as before, L_{eff} represents the effective length of the medium. When we are studying the effect of power at ω_2 on the light beam at frequency ω_1, we will refer to the wave at frequency ω_2 as pump, and the wave at frequency ω_1 as probe or signal. From eqn [77] it is apparent that the phase of signal at frequency ω_1 is modified by the power at another frequency. This is referred to as XPM. Note also that XPM is twice as effective as SPM.

Similar to the case of SPM, we can now write for the instantaneous frequency in the presence of XPM as eqn [72]:

$$\omega(t) = \omega_0 - 2\gamma L_{\mathrm{eff}} \frac{\mathrm{d}P_2}{\mathrm{d}t} \qquad [78]$$

Hence the part of the signal that is influenced by the leading edge of the pump will be down-shifted in frequency (since in the leading edge $\mathrm{d}P_2/\mathrm{d}t > 0$) and the part overlapping with the trailing edge will be up-shifted in frequency (since $\mathrm{d}P_2/\mathrm{d}t < 0$). This leads to a frequency chirping of the signal pulse just as in the case of SPM.

If the probe and pump beams are pulses, then XPM can lead to induced frequency shifts depending on whether the probe pulse interacts only with the leading edge or trailing edge or both, as they both propagate through the medium. Let us consider a case when the group velocity of pump pulse is greater than that of the probe pulse. Thus, if both pulses enter the medium together, then since the pump pulse travels faster, the probe pulse will interact only with the trailing edge of the pump. Since in this case $\mathrm{d}P_2/\mathrm{d}t$ is negative, the probe pulse suffers a blue-induced frequency shift. Similarly if the pulses enter at different instants but completely overlap at the end of the medium, then $\mathrm{d}P_2/\mathrm{d}t > 0$ and the probe pulse would suffer a red-induced frequency shift. Indeed, if the two pulses start separately and walk through each other, then there is no induced shift due to cancellation of shifts induced by leading and trailing edges of the pump. Figure 9 shows measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber.

When pulses of light at two different wavelengths propagate through an optical fiber, due to different group velocities of the pulses, they pass through each other, resulting in what could be termed as a collision. In the linear case, the pulses will pass through without affecting each other, but when intensity levels are high, XPM induces phase shifts in both pulses. We can define a parameter termed walk off length L_{wo} which is the length of the fiber required for the interacting pulses to walk off relative to each other. The walk off length is given by

$$L_{\mathrm{wo}} = \frac{\Delta\tau}{D\Delta\lambda} \qquad [79]$$

where D represents the dispersion coefficient, and $\Delta\lambda$ represents the wavelength separation between the interacting pulses. For return to zero (RZ) pulses, $\Delta\tau$ represents the pulse duration while for nonreturn to zero (NRZ) pulses, $\Delta\tau$ represents the rise term or fall time of the pulse. Closely spaced channels will thus interact over longer fiber lengths, thus leading to greater XPM effects. Larger dispersion coefficients will reduce L_{wo} and thus the effects of XPM. Since the medium is attenuating, the power carried by the pulses decreases as they propagate and thus leading to reduced XPM effect. The characteristic length for attenuation is the effective length L_{eff}, defined by eqn [53]. If $L_{\mathrm{wo}} \ll L_{\mathrm{eff}}$, then over the length of interaction of the pulses, the intensity levels do not change appreciably and the magnitude of the XPM-induced effects will be proportional to the wavelength spacing $\Delta\lambda$. For small $\Delta\lambda$'s, $L_{\mathrm{wo}} \gg L_{\mathrm{eff}}$ and the interaction length is now determined by the fiber losses (rather than by walk off) and the XPM-induced effects become almost independent of $\Delta\lambda$. Indeed, if we consider XPM effects between a continuous wave (cw) probe beam and a sinusoidally industry modulated pump beam, then the amplitude of the XPM-induced

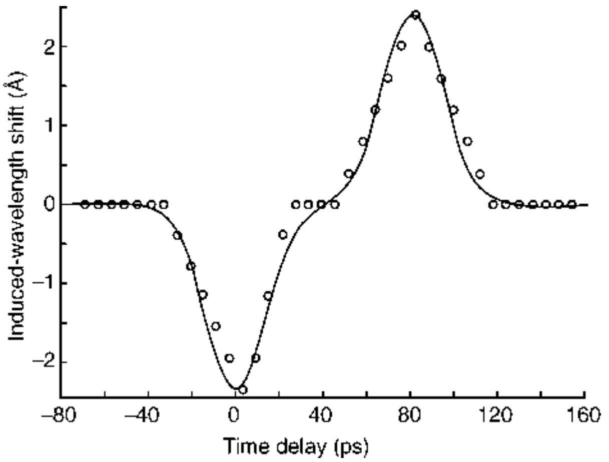

Figure 9 Measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber. (Reproduced from Baldeck PL, Alfano RR and Agrawal GP (1998) Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52: 1939–1941 with permission from the American Institute of Physics.)

phase shift ($\Delta\Phi_{\mathrm{pr}}$) in the probe beam is given by:

$$\begin{aligned} \Delta\Phi_{\mathrm{pr}} &\approx 2\gamma P_{2\mathrm{m}} L_{\mathrm{eff}} \quad \text{for} \quad L_{\mathrm{wo}} \gg L_{\mathrm{eff}} \\ \Delta\Phi_{\mathrm{pr}} &\approx 2\gamma P_{2\mathrm{m}} L_{\mathrm{wo}} \quad \text{for} \quad L_{\mathrm{wo}} \ll L_{\mathrm{eff}} \end{aligned} \qquad [80]$$

Here $P_{2\mathrm{m}}$ is the amplitude of the sinusoidal power modulation of the pump beam.

XPM-induced intensity interference can be studied by simultaneously propagating an intensity modulated pump signal and a cw probe signal at a different wavelength. The intensity modulated signal will induce phase modulation on the cw probe signal and the dispersion of the medium will convert the phase modulation to intensity modulation of the probe. Thus, the magnitude of the intensity fluctuation of the probe signal serves as an estimate of the XPM induced interference. Figure 10 shows the intensity fluctuations on a probe signal at 1550 nm, induced by a modulated pump for a channel separation of 0.6 nm. Figure 11 shows the variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. The experiment has been performed over four amplified spans of 80 km of standard single mode fiber (SMF) and nonzero dispersion shifted fiber (NZDSF). The large dispersion in SMF has been compensated using dispersion compensating chirped gratings. The probe modulation, in the case of SMF, decreases approximately linearly with 1/Δλ for all Δλs; the modulation is independent of Δλ. In contrast the NZDSF shows a constant modulation for Δλ ≤ 1 nm.

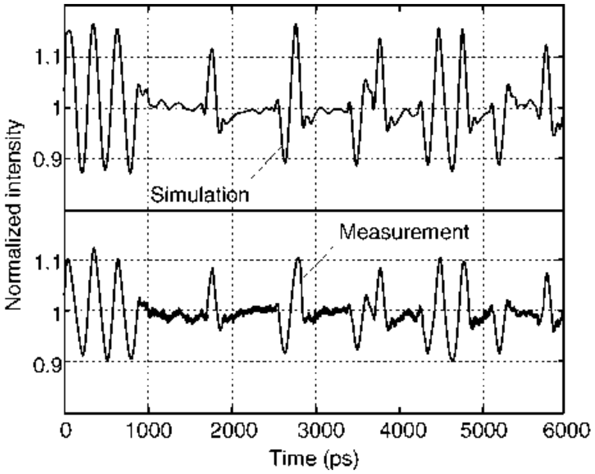

Figure 10 Intensity fluctuations induced by cross phase modulation on a probe signal at 1550 nm by a modulated pump for a channel separation of 0.6 nm. (Reproduced with permission from Rapp L (1997) Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9: 1592–1595, © 2004 IEEE.)

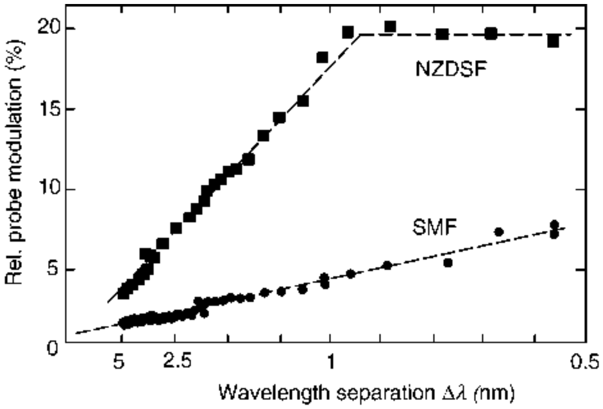

Figure 11 Variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. (Reproduced with permission from Shtaif M, Eiselt M and Garret LD (2000) Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12: 88–90, © 2004 IEEE.)

This is consistent with the earlier discussion in terms of L_{wo} and L_{eff}.

Four-Wave Mixing (FWM)

Four-wave mixing (FWM) is a nonlinear interaction that occurs in the presence of multiple wavelengths in a medium, leading to the generation of new frequencies. Thus, if light waves at three different frequencies ω_2, ω_3, ω_4 are launched simultaneously into a medium, the same nonlinear polarization that led to intensity dependent refractive index, leads to nonlinear polarization component at a frequency:

$$\omega_1 = \omega_3 + \omega_4 - \omega_2 \qquad [81]$$

This nonlinear polarization, under certain conditions, leads to the generation of electromagnetic waves at ω_1. This process is referred to as four-wave mixing due to the interaction between four different frequencies. Degenerate four-wave mixing corresponds to the case when two of the input waves have the same frequency. Thus, if $\omega_3 = \omega_4$, then inputting waves at ω_2 and ω_3 leads to the generation of waves at the frequency

$$\omega_1 = 2\omega_3 - \omega_2 \qquad [82]$$

During FWM process, there are four different frequencies present at any point in the medium. If we write the electric field of the waves as

$$E_i = \frac{1}{2}[A_i(z)\psi_i(x, y)\mathrm{e}^{\mathrm{i}(\omega_i t - \beta_i z)} + \mathrm{c.c.}], \qquad [83]$$
$$i = 1, 2, 3, 4$$

where, as before, $A_i(z)$ represents the amplitude of the wave, $\psi_i(x, y)$ the transverse field distribution, and β_i the propagation constant of the wave. The total electric field is given by

$$E = E_1 + E_2 + E_3 + E_4 \qquad [84]$$

Substituting for the total electric field in the equation for nonlinear polarization, the term with frequency ω_1 comes out as

$$P_{\mathrm{NL}}^{(\omega_1)} = \frac{1}{2}[P_{\mathrm{NL}}^{(\omega_1)}\,\mathrm{e}^{\mathrm{i}(\omega_1 t-\beta_1 z)} + \mathrm{c.c.}] \qquad [85]$$

where

$$P_{\mathrm{NL}}^{(\omega_1)} = \frac{3\varepsilon_0}{2}\chi^{(3)}A_2^* A_3 A_4 \psi_2\psi_3\psi_4\,\mathrm{e}^{-\mathrm{i}\Delta\beta z} \qquad [86]$$

and

$$\Delta\beta = \beta_3 + \beta_4 - \beta_2 - \beta_1 \qquad [87]$$

In writing eqn [86], we have only considered the FWM term, neglecting the SPM and XPM terms.

Substituting the expression for $P_{\mathrm{NL}}^{(\omega_1)}$ in the wave equation for ω_1 and making the slowly varying approximation (in a manner similar to that employed in the case of SPM and XPM), we obtain the following equation for $A_1(z)$:

$$\frac{\mathrm{d}A_1}{\mathrm{d}z} = -2i\gamma A_2^* A_3 A_4\,\mathrm{e}^{-\mathrm{i}\Delta\beta z} \qquad [88]$$

where γ is defined by eqn [48] with $k_0 = \omega/c$, ω representing the average frequency of the four interacting waves, and A_{eff} is the average effective area of the modes.

Assuming all waves to have the same attenuation coefficient α and neglecting depletion of waves at frequencies ω_2, ω_3 and ω_4, due to nonlinear conversion we obtain for the power in the frequency ω_1 as

$$P_1(L) = 4\gamma^2 P_2 P_3 P_4 L_{\mathrm{eff}}^2 \eta\,\mathrm{e}^{-\alpha L} \qquad [89]$$

where

$$\eta = \frac{\alpha^2}{\alpha^2 + \Delta\beta^2}\left[1 + \frac{4\mathrm{e}^{-\alpha L}\sin^2\dfrac{\Delta\beta L}{2}}{(1-\mathrm{e}^{-\alpha L})^2}\right] \qquad [90]$$

and L_{eff} is the effective length (see eqn [53]). Maximum four-wave mixing takes place when $\Delta\beta = 0$, since in such a case $\eta = 1$. Now:

$$\Delta\beta = \beta(\omega_3) + \beta(\omega_4) - \beta(\omega_2) - \beta(\omega_1) \qquad [91]$$

Since the frequencies are usually close to each other, we can make a Taylor series expansion about any frequency, say ω_2. In such a case, we obtain

$$\Delta\beta = (\omega_3 - \omega_2)(\omega_3 - \omega_1)\left|\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right|_{\omega=\omega_2} \qquad [92]$$

In optical fiber communication systems, the channels are usually equally spaced. Thus, we assume the frequencies to be given by

$$\omega_4 = \omega_2 + \Delta\omega, \qquad \omega_3 = \omega_2 - 2\Delta\omega \ \text{ and } \qquad \omega_1 = \omega_2 - \Delta\omega \qquad [93]$$

Using these frequencies and eqn [60], eqn [92] gives us:

$$\Delta\beta = -\frac{4\pi D\lambda^2}{c}(\Delta\nu)^2 \qquad [94]$$

where $\Delta\omega = 2\pi\Delta\nu$. Thus maximum FWM takes place when $D = 0$. This is the main problem in using wavelength division multiplexing (WDM) in dispersion shifted fibers which are characterized by zero dispersion at the operating wavelength of 1550 nm, as FWM will then lead to crosstalk among various channels. FWM efficiency can be reduced by using fiber with nonzero dispersion. This has led to the development of nonzero dispersion shifted fiber (NZDSF) which have a finite nonzero dispersion of about ±2 ps/km nm at the operating wavelength.

From eqn [94], we notice that for a given dispersion coefficient D, FWM efficiency will reduce as $\Delta\nu$ increases.

In order to get a numerical appreciation, we consider a case with $D = 0$, i.e., $\Delta\beta = 0$. For such a case $\eta = 1$. If all channels were launched with equal power P_{in} then:

$$P_1(L) = 4\gamma^2 P_{\mathrm{in}}^3 L_{\mathrm{eff}}^2\,\mathrm{e}^{-\alpha L} \qquad [95]$$

Thus the ratio of power generated at ω_1, due to FWM and that existing at the same frequency, is

$$\frac{P_{\mathrm{g}}}{P_{\mathrm{out}}} = \frac{P_1(L)}{P_{\mathrm{in}}\,\mathrm{e}^{-\alpha L}} = 4\gamma^2 P_{\mathrm{in}}^2 L_{\mathrm{eff}}^2 \qquad [96]$$

Typical values are $L_{\mathrm{eff}} = 20$ km, $\gamma = 2.4\ \mathrm{W}^{-1}\,\mathrm{km}^{-1}$. Thus:

$$\frac{P_{\mathrm{g}}}{P_{\mathrm{out}}} \approx 0.01 P_{\mathrm{in}}^2(\mathrm{mW}^2) \qquad [97]$$

Figure 12 shows the output spectrum measured at the output of a 25 km-long dispersion shifted fiber ($D = -0.2$ ps/km nm) when three 3 mW wavelengths are launched simultaneously. Notice the generation of many new frequencies by FWM. Figure 13 shows the

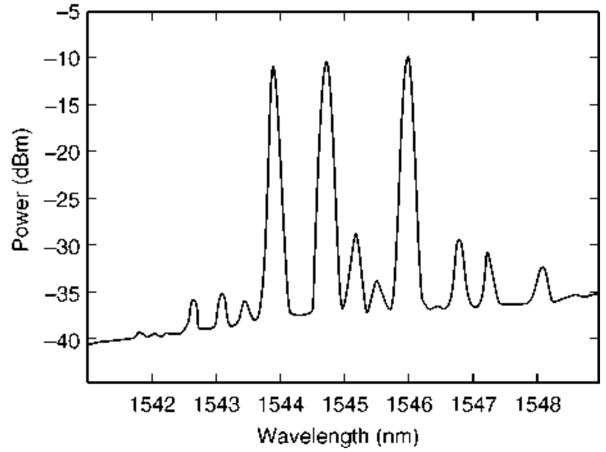

Figure 12 Generation of new frequencies because of FWM when waves at three frequencies are incident in the fiber. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.)

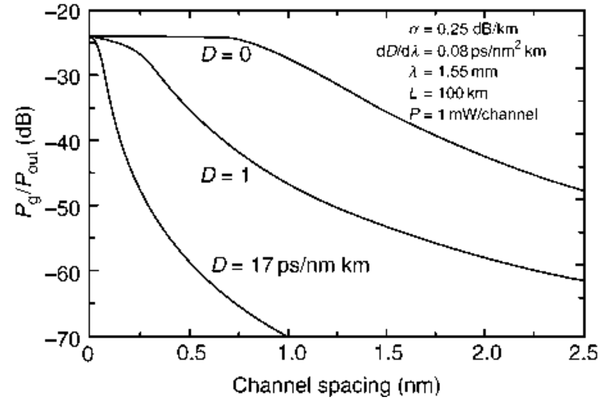

Figure 13 Ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.)

ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. It can be seen that by choosing a nonzero value of dispersion, the four-wave mixing efficiency can be reduced. Larger the dispersion coefficient, smaller can be the channel spacing for the same crosstalk.

Since dispersion leads to increased bit error rates in fiber optic communication systems, it is important to have low dispersion. On the other hand, lower dispersion leads to crosstalk due to FWM. This problem can be resolved by noting that FWM depends on the local dispersion value in the fiber, while the pulse spreading at the end of a link depends on the overall dispersion in the fiber link. If one chooses a link made up of positive and negative dispersion coefficients, then by an appropriate choice of the lengths of the positive and negative dispersion fibers, it would be possible to achieve a zero total link dispersion while at the same time maintaining a large local dispersion. This is referred to as dispersion management in fiber optic systems.

Although FWM leads to crosstalk among different wavelength channels in an optical fiber communication system, it can be used for various optical processing functions such as wavelength conversion, high-speed time division multiplexing, pulse compression, etc. For such applications, there is a concerted worldwide effort to develop highly nonlinear fibers with much smaller mode areas and higher nonlinear coefficients. Some of the very novel fibers that have been developed recently include holey fibers, photonic bandgap fibers, or photonic crystal fibers which are very interesting since they possess extremely small mode effective areas ($\sim$2.5 μm^2 at 1550 nm) and can be designed to have zero dispersion even in the visible region of the spectrum. This is expected to revolutionize nonlinear fiber optics by providing new geometries to achieve highly efficient nonlinear optical processing at lower powers.

Supercontinuum Generation

Supercontinuum (SC) generation is the phenomenon in which a nearly continuous spectrally broadened output is produced through nonlinear effects on high peak power picosecond and subpicosecond pulses. Such broadened spectra find applications in spectroscopy, wavelength characterization of optical components such as a broadband source from which many wavelength channels can be obtained by slicing the spectrum.

Supercontinuum generation in an optical fiber is a very convenient technique since it provides a very broad bandwidth (>200 nm), the intensity levels can be maintained high over long interaction lengths by choosing small mode areas, and the dispersion profile of the fiber can be appropriately designed by varying the transverse refractive index profile of the fiber.

The spectral broadening that takes place in the fiber is attributed to a combination of various third-order effects such as SPM, XPM, FWM, and Raman scattering. Since dispersion plays a significant role in the temporal evolution of the pulse, different dispersion profiles have been used in the literature to achieve broadband SC.

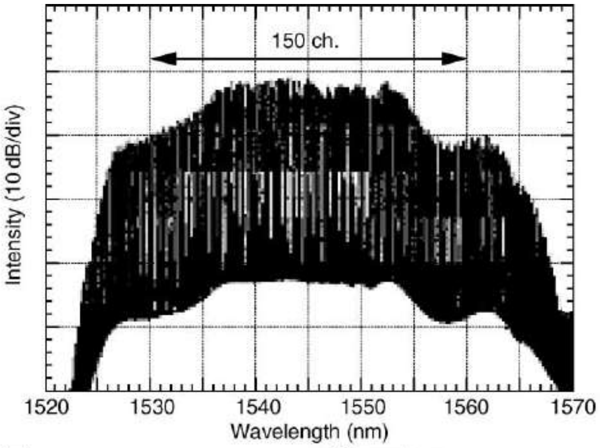

(a) SC spectrum

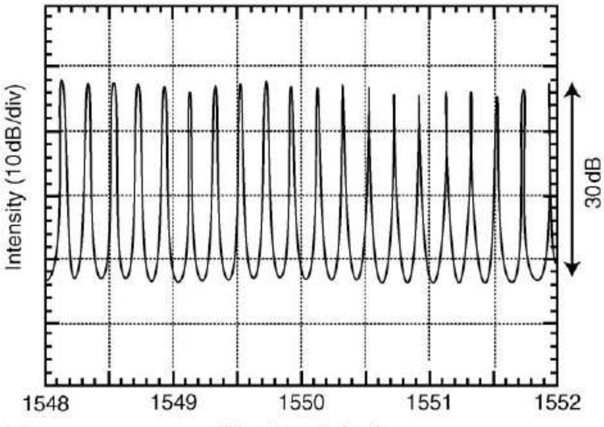

(b)

Figure 14 Super Continuum spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. (Reproduced from Yamada E, Takara H, Ohara T, Satc K, Morioka T, Jinguji K, Itoh M and Ishi M (2001) A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37: 304–306 with permission from IEEE.)

Some studies have used dispersion decreasing fibers, dispersion flattened fibers, while others have used a constant anamolous dispersion fiber followed by a normal dispersion fiber.

Figure 14 shows the SC spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. The fiber used for SC generation is a polarization maintaining dispersion flattened fiber. The output is a broad spectrum containing more than 150 spectral components at a spacing of 25 GHz with a flat top spectrum over 18 nm. The output optical powers range from −9 dBm to +3 dBm. Such sources are being investigated as attractive solutions for dense WDM (DWDM) optical fiber communication systems.

Conclusions

The advent of lasers provided us with an intense source of light and led to the birth of nonlinear optics. Nonlinear optical phenomena exhibit a rich variety of effects and coupled with optical waveguides and fibers, such effects can be observed even at moderate optical powers. With the realization of compact femtosecond lasers, highly nonlinear holey optical fibers, quasi phase matching techniques, etc, the field of nonlinear optics is expected to grow further and lead to commercial exploitation such as in compact blue lasers, soliton laser sources and multiwavelength sources and for fiber optic communication.

List of Units and Nomenclature

Attenuation	The decrease in optical power of a propagating mode usually expressed in dB/km
Birefringent medium	A medium characterized by two different modes of propagation characterized by two different polarization states of a plane light wave along any direction
Cross-phase modulation (XPM)	The phase modulation of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by another propagating light wave
Effective length (L_{eff})	The length over which most of the nonlinear effect is accumulated in a propagating light beam
Four-wave mixing (FWM)	The mixing of four propagating light waves due to intensity dependent refractive index of the medium. This mixing leads to the generation of new frequencies from a set of two or three input light beams
Mode effective area (A_{eff})	The cross sectional area of the mode which determines the nonlinear effects on the propagating mode. This is different from the core area
Nonlinear polarization (P_{NL})	When the response of the medium to an applied electric field is not proportional to the applied electric field, then this results in a nonlinear polarization
Optical waveguides	Devices which are capable of guiding optical beams by overcoming diffraction effects using

	the phenomenon of total internal reflection
Phase coherence length	The minimum crystal length upto which the second harmonic power generated by the nonlinear interaction increases
Phase matching	Condition under with the nonlinear polarization generating an electromagnetic wave propagates at the same phase velocity as the generated electromagnetic wave
Pulse dispersion	The broadening of a light pulse as it propagates in an optical fiber
Quasi-phase matching (QPM)	Periodic modulation of the nonlinear coefficient to achieve efficient nonlinear interaction
Second-harmonic generation (SHG)	The generation of a wave of frequency 2ω from an incident wave of frequency ω through nonlinear interaction in a medium
Self-phase modulation (SPM)	The modulation of the phase of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by itself
Supercontinuum generation (SC)	The phenomenon in which a nearly contiunuous spectrally broadband output is produced through nonlinear effects on high peak power picosecond and sub-picosecond pulses
Walk off length (L_{wo})	Length of the fiber required for two interacting pulses at different wavelengths to walk off relative to each other

See also

Nonlinear Optics, Applications: Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; $\chi^{(3)}$–Third-Harmonic Generation; Four-Wave Mixing. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Agrawal GP (1989) *Nonlinear Fiber Optics*. Boston, MA: Academic Press.

Armstrong JA, Bloembergen N, Ducuing J and Pershan PS (1962) Interactions between light waves in a nonlinear dielectric. *Physics Review* 127: 1918.

Baldeck PL, Alfano RR and Agrawal GP (1998) Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52: 1939–1941.

Baldi P (1992) *Generation de fluorescence parametrique guidee sur niobate et tantalite de lithium polarizes periodiquement. Etudes preliminaire d'un oscillateur parmetrique optique integree*. Doctoral Thesis of the University of Nice, France.

Chiang TK, Kagi N, Marhic ME and Kazovsky LG (1996) Cross phase modulation in fiber links with multiple optical amplifiers and dispersion compensators. *Journal of Lightwave Technology* 14: 249–259.

Franken PA, Hill AE, Peter CW and Weinreich G (1961) Generation of optical harmonics. *Physics Review Letters* 7: 118.

Ghatak AK and Thyagarajan K (1987) *Optical Electronics*. Cambridge, UK: Cambridge University Press.

Ghatak AK and Thyagarajan K (1998) *Introduction to Fiber Optics*. Cambridge, UK: Cambridge University Press.

Lim EJ, Fejer MM and Byer RL (1989) Second-harmonic generation green light in periodically poled planar lithium niobate waveguide. *Electronic Letters* 25: 174.

Lim EJ, Fejer MM, Byer RL and Kozovsky WJ (1989) *Electronic Letters* 25: 731–732.

Myers LE and Bosenberg WR (1997) Periodically poled lithium niobate and quasi phase matcher parametric oscillators. *IEEE Journal of Quantum Electronics* 33: 1663–1672.

Newell AC and Moloney JV (1992) *Nonlinear Optics*. Redwood City, CA: Addison Wesley Publishing Co.

Rapp L (1997) Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9: 1592–1595.

Shen YR (1989) *Principles of Nonlinear Optics*. New York: Wiley.

Shtaif M, Eiselt M and Garret LD (2000) Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12: 88–90.

Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of Lightwave Technology* 13: 841–849.

Wang QZ, Lim QD, Lim DH and Alfano R (1994) High resolution spectra of self phase modulation in optical fibers. *Journal of Optical Society of America* B-11: 1084.

Webjorn J, Siala S, Nan DW, Waarts RG and Lang RJ (1997) Visible laser sources based frequency doubling in nonlinear waveguides. *IEEE Journal of Quantum Electronics* 33: 1673–1686.

Yamada M, Nada N, Saitoh M and Watanabe K (1993) *Applied Physics Letters* 62: 435–436.

Yamada E, Takara H, Ohara T, Sato K, Morioka T, Jinguji K, Itoh M and Ishi M (2001) A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37: 304–306.

Yariv A (1997) *Optical Electronics in Modern Communications*, 5th ed. New York: Oxford University Press.

Optical Fiber Cables

G Galliano, Telecom Italia Lab, Torino, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In order to be put to practical use in the telecommunication network, optical fibers must be protected from environmental and mechanical stresses which can change their transmission characteristics and reliability.

The fibers, during manufacturing, are individually protected with a thin plastic layer (extruded during the drawing) to preserve directly their characteristics from damage due to the environment and handling. This coating surrounding the fiber cladding (with an external diameter of 125 μm) is known as the primary coating, which is obtained with a double layer of UV resin for a maximum external diameter of about 250 μm.

Primary coated fiber, at the end of the manufacturing process, is subjected to a dynamic traction test (proof-test) to monitor its mechanical strength. This test, carried out on all fibers manufactured, consists of the application of a well-defined strain (from 0.5% to 1%) for a fixed time (normally 1 s) and thus permits the exclusion of, from successive cabling, those fibers with poor mechanical characteristics.

At the end of the manufacturing process the optical fibers are not directly usable. It is necessary to surround the single fiber (or groups of fibers) with a structure which provides protection from mechanical and chemical hazards and allows for stresses that may be applied during cable manufacture, installation, and service.

Suitable protection is provided using fiber conditioning within the cable (secondary coating or loose cabling), while entire protection is obtained with some cable elements, such as strength members, core assembling, filling, and outer protections. Moreover, the cable structure and its design is heavily influenced by the application (e.g., number of fibers), installation (laying and splicing), and environmental conditions of service (underground, aerial, or submarine).

In this article, the main characteristics of optical cables (with both multimode or single mode fibers) are described and analyzed, together with suitable solutions.

Secondary Fiber Coating

The fiber, before cabling, can be protected with a secondary coating (a tight jacketing) to preserve it during cable stranding. However, fiber with only a primary coating may be cabled with a suitable cable protection.

In the first case the fiber can be directly stranded to form a cable core; in the second case, a single group of fibers must be inserted into a loose structure (tubes or grooves) with a proper extra-length. These two solutions are not always distinguishable; in fact, tight fibers inserted into loose tubes or grooves are often used.

The implementation of the secondary coating and the choice of proper materials requires particular attention, in order to avoid microbending effects on the fiber, which can cause degradation of the transmission characteristics. Microbending is caused by the pressure of the fiber on a microrough surface, or by the fiber buckling, due to the contraction of the structure (e.g., secondary coating) containing it. Microbending causes an increase in fiber attenuation at long wavelengths (1300 and 1550 nm) for both single and multimode fibers, but is particularly emphasized in multimode fibers.

Tight Jacket

In a tight protection, the primary coating fiber is surrounded by a second plastic jacket in contact with the fiber. This jacket mainly protects the fiber from lateral stresses that can cause the microbending.

The fiber can be individually protected with a single or double layer of plastic material or in a ribbon structure. In Figure 1 four different types of tight jacket are shown. In Figure 1a and b the fibers are singularly protected with a single or double layer up to an external diameter of 0.9 mm. The second type is normally preferred because the external hard plastic gives a good fiber protection and the inner soft plastic avoids microbending effects on the fiber.

In a ribbon structure, from 4 to 16 primary coated fibers are laid close and parallel and covered with plastic. Two types of ribbon are normally classified ‘encapsulated’ (Figure 1c) or ‘edge bonded’ (Figure 1d). In the first case, the fibers are assembled with a common secondary coating (one or two layers), in the second case, an adhesive is used to stick together the fibers with only the primary coating.

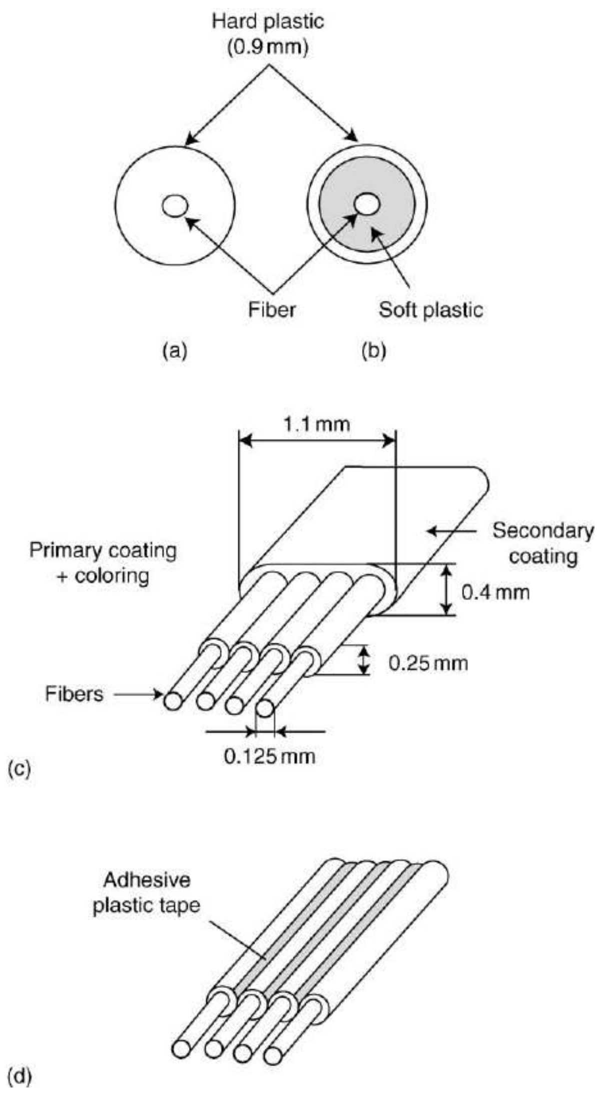

Figure 1 Different types of tight coatings.

Care should be taken in applying tight protection, in particular for a ribbon structure, to avoid an attenuation increase due to the action of the jacket on the fiber (in particular for the thermal contraction of the coating materials). The ribbon structure is designed to permit the simultaneous junction of multiple fibers; for this reason the ribbon geometry must be precision controlled.

Loose Jacket

In a loose protection the fiber can be inserted into a cable with only the primary coating. A thin layer of colored plastic may be added onto the primary coating to identify the fiber in the cable. The fibers are arranged in a tube or in a groove on a plastic core (slotted core). The diameter of the tube or the dimension of the core must be far larger than the diameter of the fiber so that the fiber is completely free inside the structure. The fibers can be inserted into a tube or into a slotted core, singularly or in groups. The space inside the tube around the fibers can be empty or filled with a suitable jelly. Some examples of loose structures are shown in Figure 2. In the tube structure, the external diameter of the jacket depends on the number of the fibers: for a single fiber up to 2 mm (Figure 2a), and for a multifiber tube (Figure 2b) up to 3 mm (for 10 fibers).

Examples of slotted core structures are shown in Figure 2c and d, respectively, for one fiber and for a group of fibers, the dimensions of grooves depending on the number of fibers.

Plastic materials with a high Young modulus are normally used for the tube that is extruded on the single or on the group of fibers. Instead, for the slotted core, plastic materials with good extrusion are used to obtain a precise profile of the grooves.

The loose structure guarantees freedom to the fibers in a defined interval of elongation and compression. Out of this interval, microbending effects can arise which increase fiber attenuation. Axial compression generally occurs when the cable is subjected to low temperatures and is generated from the thermal contraction of the cable materials that is different from that of the fibers. Axial elongation normally occurs when the cable is pulled or by the effects of high temperature.

For a correct design of a cable with a loose structure, the extra length of the fiber with respect to the tube or groove dimension and the stranding pitch must be evaluated, starting from the range of axial compression and elongation. In fact, care must be taken to prevent the fiber from being strained or forced against the tube or groove walls in the designed range.

For the tube solution, the extra length of the fibers is controlled during the tube extrusion; for the slotted core solution, the length is controlled during cabling, giving a controlled tension to the slotted core. The loose structure is often used for the cabling of fibers with tight jacketing, joining the advantages of the tight protection and loose cable. This solution is typical for ribbon cables.

Fiber Identification

To identify the fibers in a cable, a proper coloration with defined color codes must be given to the single fibers. In the tight protection, the tight buffer is directly colored during extrusion. In the loose protection, a thin colored layer is added on the primary coating of the fiber; tubes or slotted core may

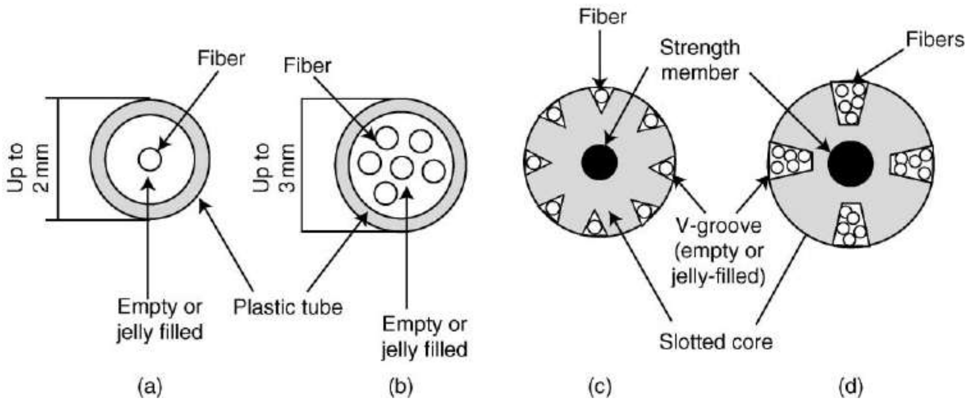

Figure 2 Different types of loose coatings.

be also colored to identify groups of fibers in high potentiality cables.

Cable Components

The fundamental components of an optical cable are: the cable core with the optical fibers, the strength member (one or more) which provide the mechanical strength of the cable, sheaths which provide the external protection, and filling materials.

Cable Core

The optical fibers with the secondary coating (tight or loose) are rejoined together in a cable core. For tight fibers or loose tubes, the cable core is obtained by stranding the fibers or the tubes around central elements that normally act as strength members too (see below).

The stranding pitch must be long enough to avoid excessive bending of the fibers and short enough to guarantee stress apportionment among the fibers in cable bends and maximum elongation of the cable without fiber strain. Normally, a helical stranding is used, but often SO stranding, that consists in an inversion of stranding direction every three of more turns, is employed. The second stranding type makes the cabling easier but may reduce its performance.

The unit core or the slotted core are made of plastic and normally incorporate a central strength member.

Strength Member

The strength member may be defined as an element that provides the mechanical strength of the cable, to withstand elongation and contraction during the installation and to ensure thermal stability. Fiber elongation, when the cable is pulled or the fiber compressed due to low temperatures, depends greatly on the characteristics of the strength member inserted in the cable.

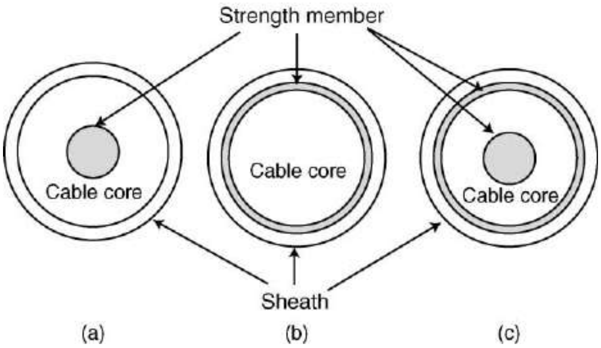

Figure 3 Different locations of the strength member in an optical cable.

As the maximum elongation values suggested for the fibers are in the range from 20 to 30% of the proof-test value, the maximum elongation allowed for a tight cable is ~0.1% instead of ~0.3% for a loose cable. To warrant these performances, a material with a high Young modulus is normally used as a strength member. Also, the strength member must be light (to avoid excessive cable weight) and flexible: these properties make it easier when laying the cable in ducts.

The strength member may be metallic or non-metallic and may be inserted in the core (Figure 3a) or in the outer part of the cable (Figure 3b), normally under the external sheath. Often two strength members are placed in an optical cable: the first one in the core and the second one under the sheath (Figure 3c).

Metallic strength members, generally one or more wires of steel, are inserted in the center or in the outer part of the cable. The steel, for its low thermal expansion coefficient and high Young modulus, behaves as a good strength member, particularly when it is arranged in the central core. Steel is a nonexpensive material, but needs protection against corrosion and electrical-induced voltages, such as by lightning or current flow in the ground.

If cables with high flexibility are requested, steel is not suitable and aramid fiber yarn or glass-fiber reinforced plastic (G-FRP) are used, particularly when the strength member is placed in the outer part of the cable. Glass fiber elements are generally used as a compressive strength member in the central cable core to avoid contraction at low temperatures, but they may be used as strength elements when the traction strength is low.

Aramid yarns are not active as resisting compression elements but are normally arranged in the outer part of the optical cable. Aramid yarns have a high Young modulus (as with steel) and low specific weight: in commercial form they consist of a large number of filaments gathered into a layer. Normally two layers stranded in opposite directions are used to avoid cable torsion during the installation.

Metal-free cables are protected by G-FRP and aramid yarns. This solution results in optical cables which are insensitive to electrical interferences.

Sheaths

The functions of the sheaths are to protect an optical cable during its life, in particular from mechanical and environmental stresses. During the installation or when the cable is in service, it may be subjected to mechanical stresses, for example, presence of humidity or water and chemical agents. The sheaths act as a barrier to avoid degradations of the fibers in the internal cable core.

The plastic sheaths (one or more, depending on the cable type and structure) are extruded over the core. This extrusion is a delicate operation in cable manufacturing because the cable core may be subjected to tension, while during the cooling the thermal compression of the plastic produces a stress on all the cable elements. Examples of cable sheaths are shown in Figure 4.

Different plastic materials, such as polyvinylchloride (PVC), polyethylene, and polyurethane are normally used. PVC has excellent mechanical properties, flexibility, and is flame retardant but is permeable to humidity. On the other hand, high-density polyethylene has a permeability lower than PVC, with good mechanical properties and low coefficient of friction, but it is more flammable than PVC and less flexible. Finally, polyurethane, for its softness and high flexibility, is often used for inner sheaths. When low toxicity is requested (e.g., in indoor cables), halogen-free materials are employed.

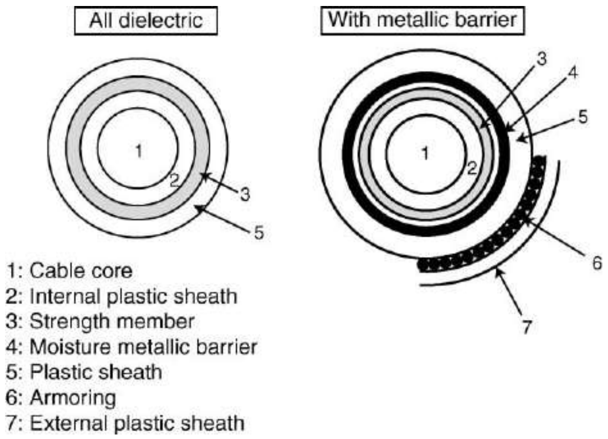

Figure 4 Examples of optical cable sheaths.

Often a metallic barrier is used under the external sheath to prevent moisture entering the cable core. The type most widely used is an aluminum ribbon (of a few tenths of a millimeter) bonded into the external polyethylene sheath, during the sheath extrusion. This structure is normally indicated as a LAP (laminated aluminum polyethylene) sheath.

When the cable is directly buried, a metallic armor (corrugated steel tape or armoring wires) is generally inserted. The metallic armor protects the cable core against radial stresses but it is also used as protection against rodents and insects.

Filling Compounds and Other Components

Fillings are used in optical cables to avoid the presence of moisture and water propagation in the cable core should a failure occur in the cable sheath. Suitable jelly compounds, with constant viscosity at a wide range of temperatures and chemical stabilities, are generally inserted in the cable core and, sometimes, in the outer protection layer. In loose structures, the jelly is in direct contact with the fibers: the compound must be compatible with them and guarantee the fiber freedom. Special compounds may be used to adsorb and permanently fix any hydrogen present in the cable core. The presence of hydrogen gas, that may be generated from the internal materials of the cable (mainly metals), is harmful to the fibers because it causes permanent attenuation increases and thus damage to them.

With the generic terminology ‘other components’, some other components used in the optical cables are included. They are often similar to those used in conventional cables. The most important are: insulated conductors, plastic fillers, and cushion layers and tapes around the fiber core.

Insulated copper conductors may be incorporated in the cable core, for example, to carry power supply or as service channels, or to detect the presence of moisture in the cable (in this case, the insulation is missing or partial). The conductors may cause problems due to the voltages induced by power lines or lighting.

Plastic fillers are often used to complete the cable core geometry. Cushion layers are used to protect the cable core against radial compression. Normally, they are based on plastic materials wound around the core. Tapes are wound around the cable core to hold the assemblies together and secondly to provide a heat barrier during the sheath extrusion. Mylar tapes are usually employed.

Optical Cable Structures, Types, and Applications

Cable Structures

The structure of an optical cable is strictly dependent on the construction method and the application type that determines the design and the grade of external protection. The main core (or inner) structures of an optical cable can be classified as: stranded structures (tight and loose); slotted core cable; or ribbon cable. In this section, a few examples of cable structures are presented. Below, separate sections are devoted to submarine cables and to special application cables, for their particular structures and features.

In the stranded tight structure, a low number of tight or loose fibers is stranded around a strength member to form a cable unit. Some of this cable unit may be joined to constitute a cable with a medium/high number of fibers (Figure 5a). Alternatively the fibers may be arranged in a multilayer structure (Figure 5b).

The main advantages of a tight stranded structure are the small dimensions of the cable core, even when a large number of fibers are involved, and also the facility of handling. Care must be taken in the design of the strength member: in fact, the fibers are immediately subjected to tension when the cable is pulled into the ducts or compressed when subjected to low temperatures.

In the stranded loose structures, two categories of cables are defined, one fiber per tube and a group of fibers (up to 20) per tube (Figure 5c). As already mentioned, in the loose structures, the fibers are free inside the tube and for this reason the cable is less critical than the tight cable. The external dimensions of loose cables are larger than tight cables, but the solution with multiform tubes allows for cables with a high number of fibers in relatively small dimensions.

Slotted core cables (Figure 2c,d) contain fibers with only the primary coating and may be divided, as in the stranded loose structure, into one fiber per groove or into a group of fibers per groove. The cable design and the choice of a suitable plastic material for the slotted core joined to a proper strength member, make it possible to obtain cables with fiber stability at a wide range of temperature. As in the loose stranded solutions, the slotted core solution is adopted to obtain cables with a high number of fibers in small dimensions.

In the ribbon solution, the fibers (from 4 to 16) are placed in a linear ribbon array. Many ribbons are stacked together to constitute the optical unit. The ribbon cable normally has a loose structure (loose tube or slotted core). Three structures are shown in Figure 6. In the first example (classical AT & T cable), 12 ribbons of 12 fibers are directly stacked and stranded in a tube (Figure 6a). In the second and third examples, ribbons are inserted into tubes or grooves (Figure 6b,c). Several tubes or slotted cores can be assembled together to give a cable with very high fiber density. The ribbon approach may be efficient when a large number of fibers must be

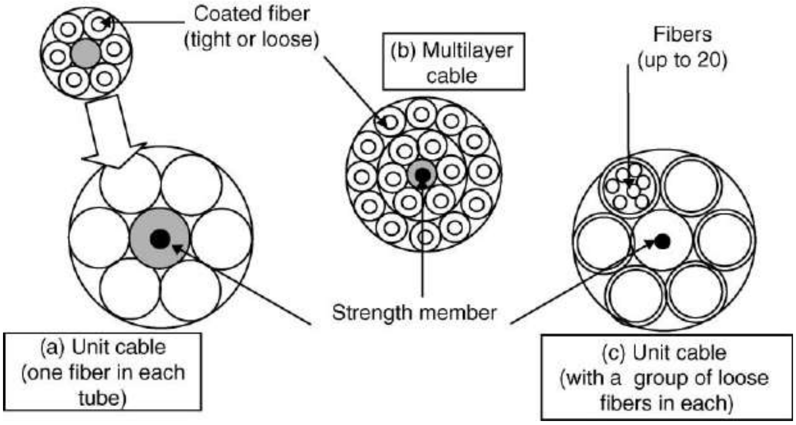

Figure 5 Structures of stranded core cables.

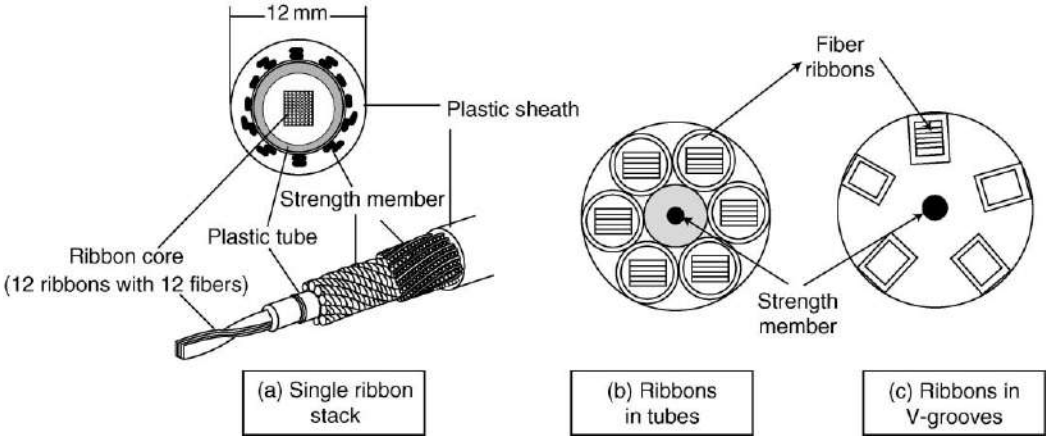

Figure 6 Structures of ribbon cables.

simultaneously handled, spliced, and terminated at connectors.

Types and Applications

The choice of an optical cable is heavily established by many factors, but mainly on the type of installation. On this basis, some different typology of cables may be identified: aerial cables, duct cables, direct buried cables, indoor cables, underwater cables, etc.

Aerial cables can be clasped onto a metal strand or can be self supporting. In the first case, the cable is not subjected to a particular tension but requires good thermal and mechanical performances. This solution is chosen when the cable must withstand strong ice and wind and when long distances between the poles are required. All the structures mentioned above may be adopted. In the second case, the cable is normally exposed to high mechanical stresses during its life and high tensile strength must be guaranted to maintain the fiber elongation at a safety level. Generally, loose structures are preferred for their greater fiber freedom. Often, in self-supporting cables, the strength member is external to the cable core. An example is shown in Figure 7.

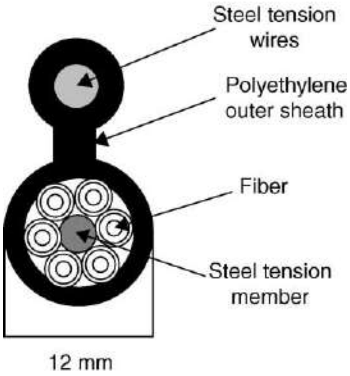

Figure 7 Self-supporting aerial cable.

Cables pulled in ducts must be resistant to pulling and torsion forces, and light and flexible to permit the installation of long sections. The cable must also protect the fibers against water and moisture that may be present in ducts or manholes. Usually the cable is filled with a jelly compound, a metallic barrier is normally used and, often, an armoring is added when protection against rodents is necessary. All the structures described above may be employed in duct cables.

The characteristics of direct buried cables are very similar to those of duct cables, but additional protections, such as metallic armoring, are required to avoid the risk of damages from digging and other earthmoving work.

Cables for indoor applications normally require a smaller number of fibers. Their main characteristics are: small dimensions, flexibility, small curvature radius, and ease in handling and jointing. For indoor applications, the sheath must not permit flame propagation and must have a low emission of toxic gases and dark smoke. Tight cables are preferably used for indoor applications due to their compactness and small dimensions.

Underwater cables are employed in the crossing of rivers, in lakes and lagoons, and for shallow water applications in general. Such cables must have stringent requirements such as resistance to moisture and water penetration and propagation, pulling resistance (during the installation, and recovery in event of failure), and high resistance to static pressure and core structure compatible with land cables (when it is part of a ground link). In addition, due to the presence of metallic structures, attention should be paid to hydrogen effects.

Submarine Cables

Submarine cable is manufactured for laying in deep-sea conditions. In designing a submarine cable, it is necessary to provide high reliability that the mechanical and transmission characteristics of the optical fibers will be stable over a long period of time. Different cables are used for submarine applications, but generally follow the basic scheme of Figure 8.

The central core (with a few fibers – up to 12 – in a slotted core or in a tight structure) is surrounded by a double layer of steel wires that act as strength members against tensile action and water pressure. Metal pipes at the inner or outer sides of the strength member, act as a water barrier and as power supply conductors (for the supply of the undersea regenerators).

The outer polyethylene insulating sheath is the ultimate protection for the ordinary deep-sea cable. For cables requiring special protection, steel wire armoring (anti-attack from fishes), and further polyethylene are added.

The cable must have stringent mechanical requirements, such as resistance to traction, torsion, crushing, impact, and to shark attack. The cable must be suitable for installation using standard cable-laying ships. Furthermore the cable materials must have low content of hydrogen and emissions. If necessary, a hydrogen absorber may be included in the cable core.

Special Cables

This category comprises cables not specifically used in ordinary telecommunication networks, but cables used for specific applications and according to specified requirements. Cables for military use are normally employed for temporary plant restoration. They must be light but crush resistant and with good mechanical characteristics. Cables for mobile applications (robots, elevators, aircrafts, submarines, etc.) generally require high flexibility and tensile strength. Special cables may be installed in particular environments, where they must be insensitive, for instance, to nuclear radiation, chemical agents, and high temperatures.

Optical cables can be installed in power lines, together with the power conductors, or into the ground wires of high-voltage overhead power lines. In this last type of cable, the central stranded steel wire, making up the ground wire, is replaced with a metal tube containing the optical unit (normally a small groups of optical fibers inserted in a tube or in a slotted core).

Blown Fiber

This technology consists of the installation of a single fiber or of small fiber bundles (or cables) into pre-installed tubes or ducts using a flux of compressed air. Initially, cables with empty tubes are installed and successively, at the time of need, the tubes can be filled by blowing fibers (singly, or in bundles) or cables. The installation time is normally fast; some compressors can blow tens of meters in a minute. The typical maximum blowing distance is about 1 km for horizontal planes, but it is reduced

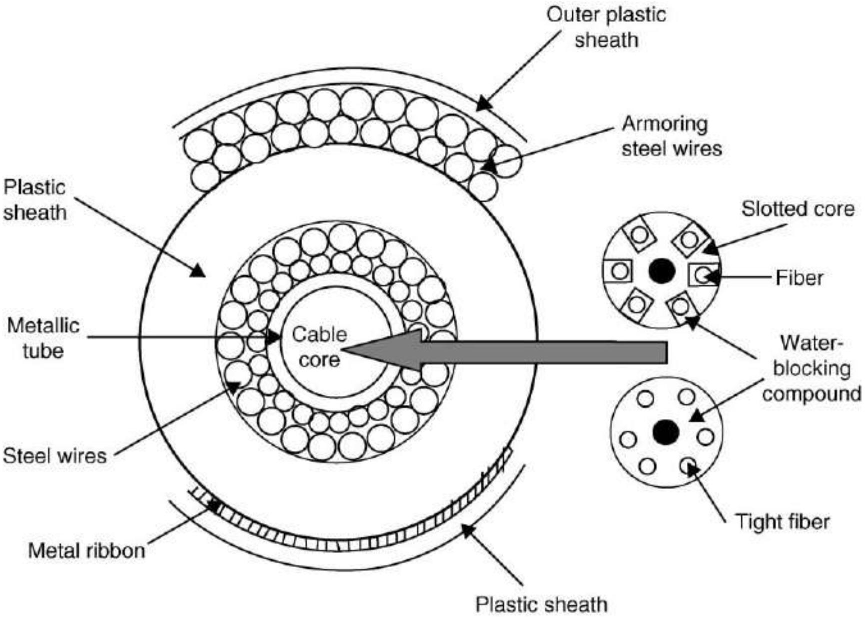

Figure 8 Basic structure of a submarine cable.

along winding roads. The advantage of this technique is to produce variable and flexible structures, increasing the number of fibers to fit the network need. Besides, substitutions or change of fibers can be quickly carried out without substitution of the cable.

See also

Fiber and Guided Wave Optics: Dispersion. **Light Emitting Diodes**.

Further Reading

Griffioen W (1989) The installation of conventional fiber-optic cables in conduits using the viscous flow of air. *Journal of Lightwave Technology* 7(2): 297–302.

Hughes H (1997) *Telecommunications Cables. Design, Manufacture and Installation.* : John Wiley and Sons.

Keiser G (1986) *Optical Fiber Communications*. McGraw-Hill.

Murata H (1988) *Handbook of Optical Fibers and Cables*. Marcel Dekker.

Technical Staff of CSELT (1990) *Fiber Optic Communications Handbook Second Edition.* : TAB Books.

Passive Optical Components

D Suino, Telecom Italia Lab, Torino, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modern telecommunications networks are based on fiber optics as signal transporters. In these networks the signals travel in the form of light confined within the fibers. As light power travels through the fibers from the source apparatus to the receiver, a number of components are employed to manage the optical signal. All other components that are not used to manage the signal are called 'passive optical components'. They serve several functions, such as to join two different fibers to distribute the signal onto more optical branches to limit the optical power to select particular wavelengths, and so on.

These components can be located anywhere in the network, depending on their function and on the network architecture. However, their main location is near the apparatus in which it is necessary to manage the signal before it is sent through the network or at the junction in the network where it is sent on to the receiver.

The main types of passive components that are found in a network can be grouped as follows:

- Joint devices:
  - Optical connectors;
  - Mechanical splices.
- Branching devices:
  - Nonwavelength-selective branching devices;
  - WDM.
- Attenuators;
- Filters;
- Isolators;
- Circulators.

Optical fibers are also divided into several groups (multimode 50/125, multimode 62.5/125, singlemode, dispersion shifted, NZD, and so on). There are also components of different typology that match up with each particular fiber. However, the general concept in terms of functionality, technology, and characteristic parameters, are the same among passive optical components of the same type (connectors, branching devices, or attenuators) but of different groups (single-mode, multimode, etc.).

Optical Joint Devices

The function of an optical joint is to perform a junction between two fibers, giving optical continuity to the line. Two different classes of joint can be considered: the connectors and the splices. The joints made by using connectors, are flexible points in the network, which means they can be opened and reconnected several times in order, for example, to reconfigure the network plan. However, a splice is a fixed joint and usually cannot be opened and remated. There are two type of splice: mechanical splices, holding joints to fibers by glue or mechanical crimp, and fusion splices fusing the two fiber heads with a suitable fusion splicer.

General Considerations

Considering that fibers can be described as pipes, in the core of which flows light, continuity between two fibers means that the two cores must be aligned in a stable and accurate way. Obviously this alignment must be carried out with minimum power loss in the junction. The main factors that affect the power loss in a joint point between two fibers are:

- $x =$ Lateral offset;
- $z =$ Longitudinal offset;
- $\theta =$ Angular misalignment; and
- $w_T/w_R =$ Mode field diameter ratio (in the single mode fiber) or numerical aperture ratio (in the multimode fiber).

These geometrical mismatching between the two fibers is represented in Figure 1.

The function that describes the joint attenuation is different for single-mode and multimode connectors, but the parameters that contribute to it are the same. The general formulation for these functions are complicated and, in particular, for the multimode connector it involves an integral on the shape distribution of the light in the fiber core. These functions become simpler in particular conditions: for example, if we consider a single-mode joint without a gap between the two fibers (that is the more usual condition in the modern connectors) the z parameter is nil, and in this case the function that describes the attenuation of the connector is

$$A(\mathrm{d}B) = -10\log\left\{\left(\frac{2w_T w_R}{w_T^2 + w_R^2}\right)^2 \times \exp\left[-2\frac{(x\lambda)^2 + (\pi n w_T w_R \sin(\vartheta))^2}{\lambda^2(w_T^2 + w_R^2)}\right]\right\} \qquad [1]$$

where n is the fiber core refractive index and λ is the wavelength of the light.

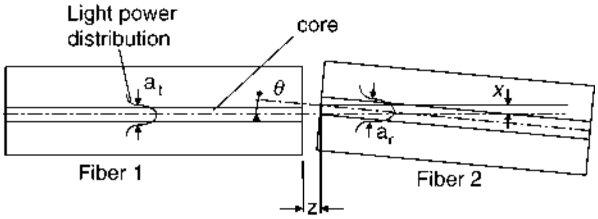

Figure 1 Main parameter that affects the joint power loss.

Figure 2 shows the trend of eqn [1], where the parameter ranges in the graphs are typical for the most common products on the market. It is evident that the main contribution to attenuation arises from the lateral offset.

For connectors of good quality, both for single-mode and multimode fibers, the attenuation values are in the range of a few tenths of dB (typically less than 0.3 dB). Another important factor for a joint, in particular for connectors used in networks for high speed or analog signals, is the return loss (RL). This is a parameter that gives a measure of the quantity of light back-reflected onto the optical discontinuity of the connector, due to the Fresnel effect. The RL is defined as the ratio in dB between the incident light power (P_{inc}) and the reflected light power (P_{ref}):

$$\mathrm{RL} = -10\log\left(\frac{P_{\mathrm{ref}}}{P_{\mathrm{inc}}}\right) \qquad [2]$$

Fresnel reflection arises when the light passes between two media with different refractive indices. In the case of connectors in which the two fibers are not in contact, the light passes from the silica of the fiber

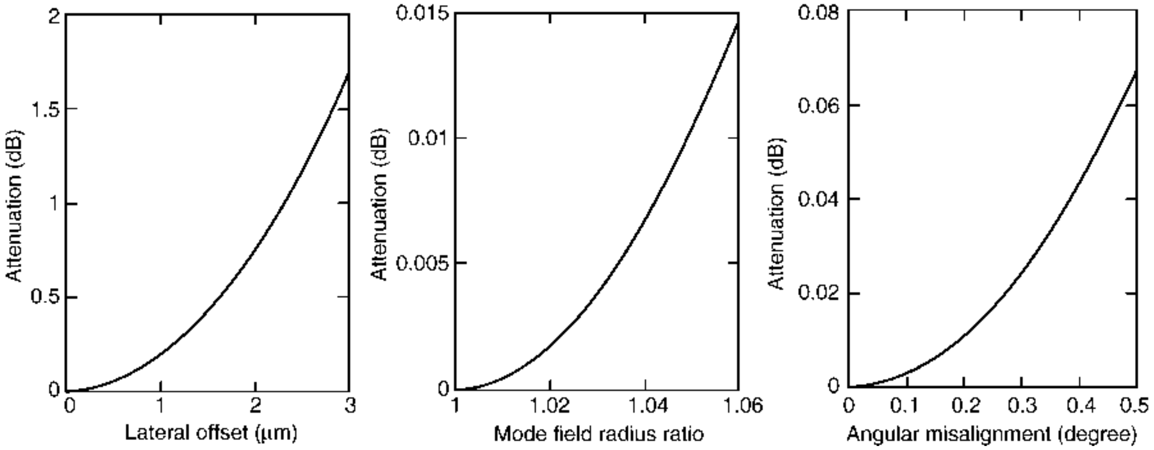

Figure 2 Single mode connector attenuation as a function of the geometrical parameters. Each curve is plotted keeping the other parameter constant.

core to the air in the gap. This causes the reflection of 4% of the incident light that, as RL, is about 14 dB.

To improve the RL performances of an optical joint, it is necessary to reduce the refractive index difference. This can be achieved either by index matching material between the two fibers that reduces the differences between the refractive index of the fiber and the gap, or by performing a physical contact between the fiber heads. The first technique is typically used in mechanical splices but, due to the problem related with the pollution contamination of the index matching, it is not a good solution for connectors that are to be opened and reconnected several times.

The physical contact solution is the most frequently adopted for these connectors. Physical contact is obtained by polishing the ferrule and the fiber end-faces in a convex shape, with the fiber core positioned in the apex (Figure 3). With this technique, the typical values of return loss for a single mode connector are in the range from 35 dB to 55 dB, depending on the polishing grade. The residual back-reflected light is generated in the thin layer of the fiber end surface, because of slight changes in the refractive index due to the polishing process.

Higher return loss values (small quantity of reflected light) are achieved with the angled physical contact (APC) technique. This is obtained by polishing the convex ferrule end face angled with respect to the fiber axis. In this way, the light is not back reflected into the fiber core, but into the cladding, and is thus eliminated (Figure 4). In APC connectors, typical return loss values are greater than 60 dB.

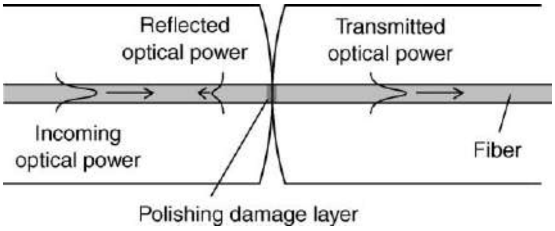

Figure 3 Back reflection effects in PC connectors.

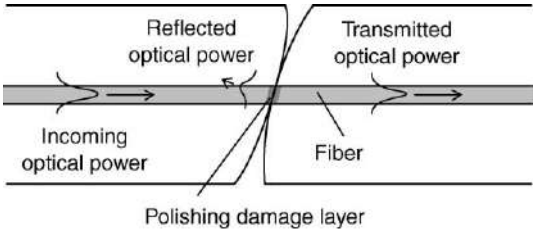

Figure 4 Back reflection effects in APC connectors.

Optical Connectors

On the market are several types of optical connectors. It is possible to divide them into three main groups:

- Connectors with direct alignment of bare fiber;
- Connectors with alignment by means of a ferrule; and
- Multifiber connectors.

Connectors with direct alignment of bare fiber

In this type of connector the bare fibers are aligned on an external structure that is usually a V-groove (Figure 5). Two configurations can be performed: plug and socket or plug–adapter–plug. In the first case, the fiber in the socket is fixed and the one in the plug aligns to it. In the plug–adapter–plug configuration, two identical plugs containing the bare fibers are locked onto an adapter in which the fibers are placed on the aligned structure. These connectors are cheaper than the connectors with a ferrule, (described below), but this is unreliable. In fact, the bare fibers are delicate and brittle and they can be affected by the external mechanical or environmental stresses.

Connectors with alignment by means ferrule

In this type of connector, the fibers are fixed in a secondary alignment structure, usually of cylindrical shape, called a ferrule. The fiber alignment is obtained by inserting the two ferrules containing the fibers into a sleeve. Usually this type of connector is in the plug–adapter–plug configuration (Figure 6). With this technique, the precision of the alignment is moved from the V-groove precision to the dimensional tolerance of the ferrule. Typical tolerance values on the parameters for a cylindrical ferrule with a diameter of 2.5 mm, are in the range of a tenth of a micrometer, and must be verified on the diameter, eccentricity between the hole (through which the fiber is fixed) and cylindricity of the external surface.

Multifiber connectors

In some types of optical cable, the fibers are organized in multiple structures, in which many fibers are glued together to form a ribbon. The ribbon can have

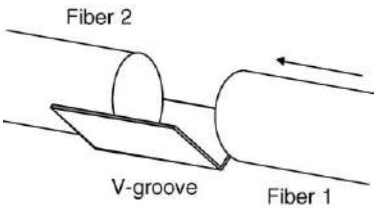

Figure 5 Scheme of the bare fibers alignment on a V-groove.

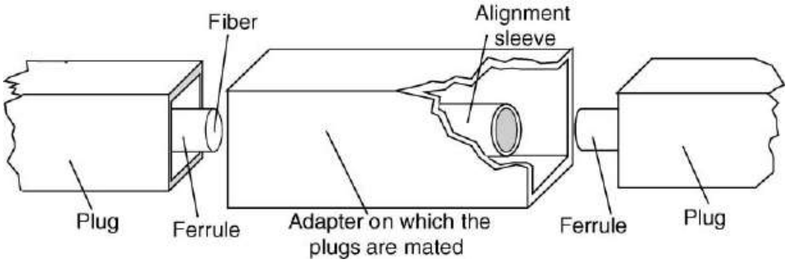

Figure 6 Scheme of a plug–adapter–plug connector with alignment by ferrules and sleeve.

2, 4, 8, 12, or 16 fibers, for example. When these cables are used, it is convenient to maintain the ribbon fiber structure in the connectors (until the point in which it is necessary to use the fibers singularly). Multifiber connectors are typically based on a secondary alignment on a plastic ferrule of rectangular shape in which the fibers maintain the same geometry that they have in the ribbon. The rectangular ferrules of the two plugs are normally aligned by means of two metallic pins placed in two hole in line with the fiber. In Figure 7, a scheme of a multifiber connector is shown.

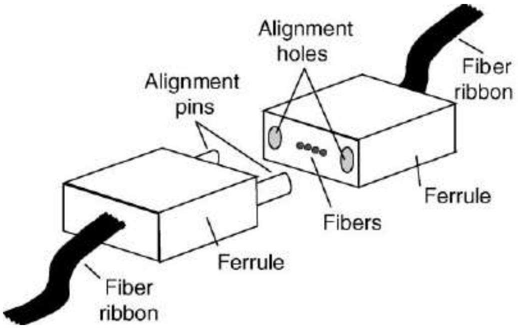

Figure 7 Multifiber connector.

Mechanical Splices

Mechanical splices are yet another way to join together two fibers. All the main considerations for the connectors are true for the mechanical splices. The difference is that the mechanical splice is a fixed joint and, normally, it cannot be opened and re-mated. Mechanical splices, both for single fiber and for multifiber structures, are the most usually manufactured. In a mechanical splice, the fibers are aligned directly in a V-groove and fixed by glue or a mechanical holder.

These components are cheaper than fusion splices but their optical, mechanical, and environmental functions are lower than the fusion splices. Figure 8, shows a mechanical splice for multifibers in which the fibers are crimped on a V-groove array. Usually, to avoid high optical reflection in the junction point, this type of joint is filled with index matching material.

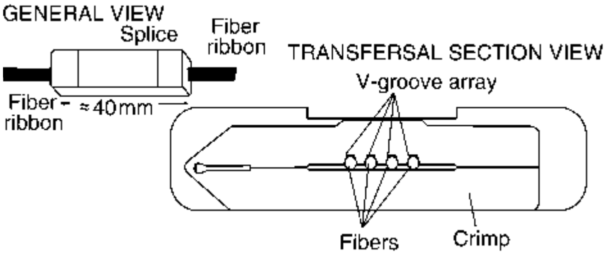

Figure 8 Example of a multifiber mechanical splice.

Branching Devices

Branching devices are passive components devoted to distributing the optical power from an input fiber to two or more fibers. These components can be classified in two broad classes, on the basis of the wavelength dependence. The nonwavelength-selective branching devices, called also couplers or splitters, share the input power to two or more fibers independently from the light wavelength. The components that perform the light distribution on more fibers on the basis of the different wavelengths of the input light are called a WDM (wavelength division (de)multiplexer).

Branching devices have three or more ports, bare fiber or connectors, for the input and/or output of optical power, and share optical power among these ports. A generic branching device can be represented as a multiport device having N input ports and M output ports, as shown in the Figure 9.

The optical characteristics of a branching device can be represented by a square transfer matrix of $n \times n$ coefficients, where n is the total number of the component ports. Each coefficient t_{ij} in the matrix is the fractional optical power transferred among designated ports, that is the ratio of the optical power P_{ij} transferred out of port j with respect to input

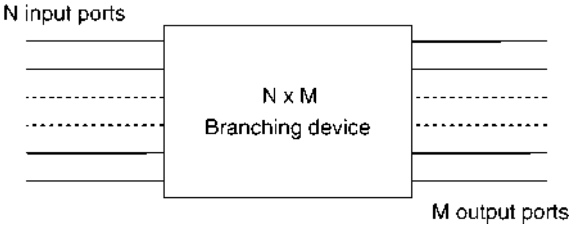

Figure 9 Scheme of a generic N × M branching device.

power P_i into port i:

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & & & \\ \vdots & & t_{ij} & \vdots \\ & & \ddots & \\ t_{n1} & & & t_{nn} \end{bmatrix} \quad [3]$$

where:

$$t_{ij} = \frac{P_{ij}}{P_i} \quad [4]$$

So, according to the relationship between two ports, the coefficients represent several parameters. For example, if the port i is an input port and j is an output port, t_{ij} is the transferred power signal through the component, while if they exist as two isolated ports the coefficients represent the crosstalk between the two ports.

The three main techniques to obtain this type of component are:

(i) All fiber components;
(ii) Micro-optics; and
(iii) Integrated optics.

Nonwavelength-Selective Branching Devices

A nonwavelength-selective branching device is a component having three or more ports, which shares an input signal among the output ports in a predetermined fashion. Usually these components are completely reversible and so they also work as a combiner of signals from more input ports onto a single output port.

There are several types of couplers for different applications:

- Y-coupler with one input port and two output ports;
- X-coupler with two input ports and two output ports; and
- Star coupler with more than two ports in input and/or in output.

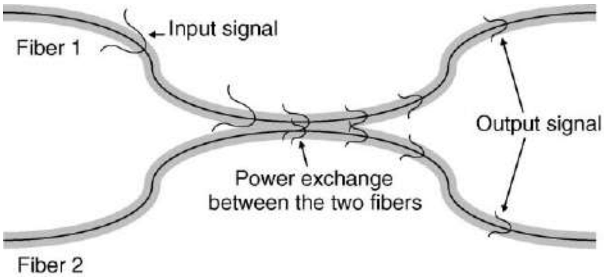

Figure 10 Scheme of a fusion 2 × 2 coupler.

Moreover the couplers can be symmetrical, and in this case they divide the input power light onto n equal output flows, or asymmetrical devices in which the quantity of power light in the output ports is shared in a predefined nonuniform way.

The techniques used to carry out this type of device are mainly the fusion technique or planar technique. The first one is based on the phenomenon of evanescent wave coupling: when the core of two optical guides (as fibers) are located in close proximity, a power exchange from a guide to another is possible. This configuration can be obtained by stretching the fibers under controlled fusion (see Figure 10).

The planar technique, that is the simpler integrated optic application, is based on the optical guides with appropriate shapes being placed directly onto a silicon wafer by means of lithographic processes (Figure 11).

The fusion technique has the advantage of being 'all in fiber' so the component is ready to be inserted into the optical networks by means of usual joint techniques. But for a high number of ports, the coupling ratio among the branches cannot be controlled in a precise way and the cost of the fusion device is proportional to the number of the branches. On the contrary, a device obtained with the planar technology must be coupled with fibers normally used and this operation may be difficult due to the different geometry of the planar guide (rectangular) and the fiber (circular). On the other hand, this technology allows a high precision in the optical characteristics of the component and the cost of the device is independent of the branch number.

WDM

A wavelength-selective branching device, usually called WDM (Wavelength Depending Multiplexer),

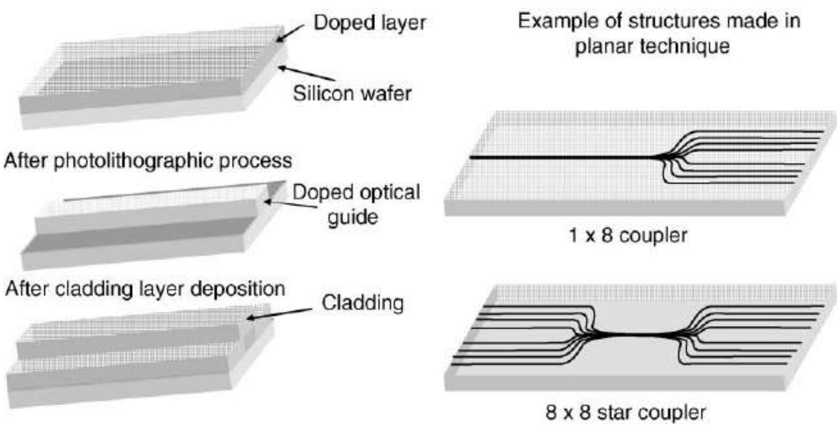
Doped layer
Silicon wafer
After photolithographic process
Doped optical guide
After cladding layer deposition
Cladding
Example of structures made in planar technique
1 x 8 coupler
8 x 8 star coupler

Figure 11 Scheme of the process to obtain a planar coupler and an example of a device.

is a component having three or more ports which shares an input signal among the output ports in a predetermined fashion, depending on the wavelength, so that at least two different wavelength ranges are nominally transferred between two different couples of ports.

WDM devices may be divided in three categories on the basis of the bandwidth of the channels (the spacing between two adjacent wavelength discriminated by the device):

(i) DWDM (Dense WDM) device operates for channel spacing equal or less than 1000 GHz (about 6–8 nm in the range of typical optical wavelength used in telecommunications systems, 1310 or 1550 nm);
(ii) CWDM (Coarse WDM) device operates for channel spacing less than 50 nm and greater than 1000 GHz; and
(iii) WWDM (Wide WDM) device operates for channel spacing equal or greater than 50 nm.

These components are used to combine, at the input side of an optical link in only one fiber, more optical signals with different wavelengths and separate them at the receiving end. This technology allows to send along a fiber, more communication channels, thus increasing the total bit rate transmitted. These types of components can be obtained using filters to select single a wavelength (as a diffractive grating that resolves the light into its monochromatic components) or using the phenomena of evanescent field coupling occurring between two adjacent fiber cores (as described for coupler devices), controlling in a precise way the coupling zone; this is, in fact, dependent on the wavelength (Figure 12).

Using wavelength filters, as they select a singular wavelength, it is necessary to perform a cascade filter structure for separating more than two wavelengths.

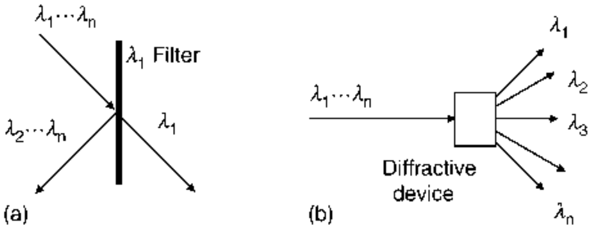
$\lambda_1 \cdots \lambda_n$
λ_1 Filter
$\lambda_2 \cdots \lambda_n$
λ_1
(a)
$\lambda_1 \cdots \lambda_n$
Diffractive device
λ_1
λ_2
λ_3
λ_n
(b)

Figure 12 (a) Scheme of the filter technique and (b) diffractive device technique.

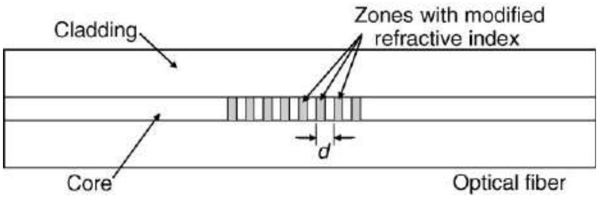
Cladding
Zones with modified refractive index
Core
d
Optical fiber

Figure 13 Scheme of FBG.

Micro-optic technologies are generally used for multimode fiber due to the critical collimation problem. For the single-mode fiber the evanescent field coupling technique is applied directly to fiber or to planar optical guides.

Another efficient technique to create WDM devices is with a fiber Bragg grating (FBG) wavelength filter directly applied to the fiber. FBGs consist in periodic modulation (see Figure 13) of the refractive index along the core of the fiber, and they act as reflection filters.

During the fabrication process (see Figure 14) the core of the optical fiber is exposed to UV radiation, and this exposure increases the refractive index of the core in defined zones, with a periodic shape obtained by the interference fringe created by crossing two coherent beams.

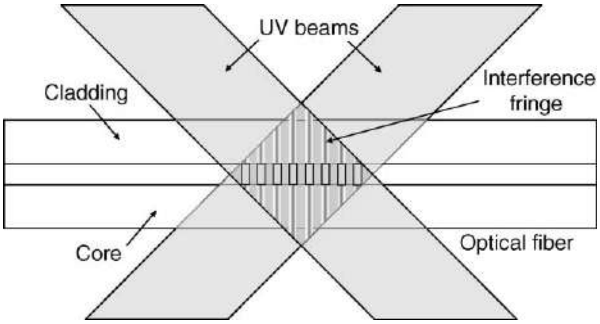

Figure 14 Scheme of FBG production.

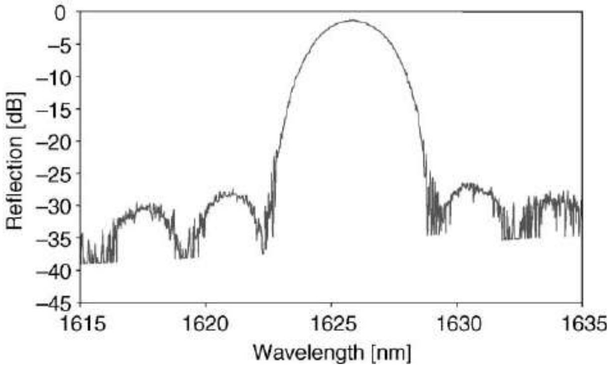

Figure 15 Transmission characteristic of FBG.

Figure 15 shows a representative spectral curve of a FBG filter centered at 1625 nm. The central wavelength, the width, and the value of reflection depend on the grating period (*d*) and on the deposited UV energy in the fiber core.

Other Passive Optical Components

In an optical network, besides the devices described above, there are a number of other passive optical components that perform particular functions on the optical signal; the main ones are:

Attenuators These are devices used to attenuate the optical power, for example, in front of a receiver when the optical line is short and the power of the received signal is still too high for the detector, or when it is necessary to equalize the signals arriving from different optical lines. It exists both in tunable and fixed versions. The first ones allow adjustment of the attenuation value, however, the second ones have a fixed and predefined value. The fixed optical attenuators can be obtained by inserting a filter or stressing a fiber in a controlled and permanent way. The tunable attenuators are based on variable optical parameters as, for example, the distance or the lateral offset in an optical connection. They exist both as connect lengths of fiber (attenuated optical jumpers) or as compact components similar to a double connector (plug stile attenuator).

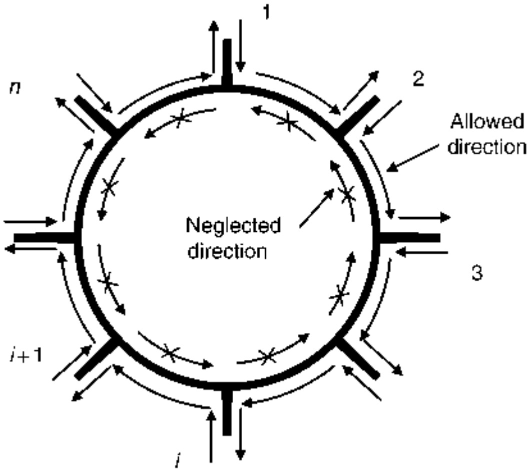

Figure 16 Principle scheme of an optical circulator.

Filters These are components with two ports employed to select or to filter some fixed wavelengths. They can be divided into the following categories:

- short-wave pass (only wavelengths lower than or equal to a specified value are passed);
- long-wave pass (only wavelengths greater than or equal to a specified value are passed);
- band-pass (only an optical window is allowed); and
- notch (only an optical window is inhibited).

It is also possible to have a combination of the above categories. The optical filtering can be obtained by inserting one or more filtering devices (as interferential films) into a fiber or creating a filter directly on the fiber by the FBG described above.

Isolators These are nonreciprocal two-port optical device in which the light flows in the forward direction, and not in the reverse one. The isolators are used to suppress backward reflections along an optical fiber transmission line, while having minimum insertion loss in the forward direction. Fiber optic isolators are commonly used to avoid back reflections into laser diodes and optical amplifiers, which can make the laser and amplifiers oscillate unstably, and cause noise in the fiber optic transmission system.

Circulators These are passive components having three or more ports numbered sequentially $(1, 2, \ldots, n)$ through which optical power is transmitted nonreciprocally from port 1 to port 2,…, from port (i) to port $(i+1), \ldots$ and from port n to port 1 (Figure 16).

See also

Diffraction: Fresnel Diffraction. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Adam MJ (1981) *An Introduction to Optical Waveguides*. New York: John Wiley & Sons.

Blonder GE, *et al.* (1989) Glass waveguides on silicon for hybrid optical packaging. *IEEE Journal of Lightwave Technology* 7: 1530.

Digonnet M and Shaw HJ (1983) Wavelength multiplexing in single-mode fiber couplers. *Applied Optics* 22: 484.

Jeunhomme LB (1983) *Single-Mode Fiber Optics – Principles and Applications*. New York: Marcel Dekker.

Kashima N (1995) *Passive Optical Components for Optical Fiber Transmission*. Norwood, MA: Artech House.

Kashima N (1993) *Optical Transmission for the Subscriber Loop*. Norwood, MA: Artech House.

Miller CM (1986) *Optical Fiber Splices and Connectors – Theory and Methods*. New York: Marcel Dekker.

SC86B IEC (1999) *IEC 60874-1: Connectors for Optical Fibres and Cables – Part 1: Generic Specification*, 4th edn. Geneva: International Electrotechnical Commission.

SC86B IEC (1999) *IEC 61073-1: Optical Fibres and Cables – Mechanical Splices and Fusion Splice Protectors – Part 1: Generic Specification*, 3rd edn. Geneva: International Electrotechnical Commission.

SC86B IEC (1999) *IEC 60869-1: Fibre Optic Attenuators – Part 1: Generic Specification*, 3rd edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 60875-1: Non-wavelength Selective Fibre Optic Branching Devices – Part 1:Generic Specification*, 4th edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 62077-1: Fibre Optic Circulators – Part 1: Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 61977-1 Ed. 1.0: Fibre Optic Filters – Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 61202-1: Fibre Optic Isolators – Part 1: Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

Technical Staff of CSELT (1990) *Fiber Optic Communications Handbook*, 2nd edn, TAB Books. New York: McGraw-Hill.

Yokohama I, *et al.* (1988) Novel mass-fabrication of fiber couplers using arrayed fiber ribbons. *Electronic Letters* 24: 1147.

FIBER GRATINGS

P S Westbrook, OFS Laboratories, Murray Hill, NJ, USA

B J Eggleton, University of Sydney, Sydney, NSW, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Fiber gratings are periodic index variations inscribed into the core of an optical fiber, and are important devices for manipulating fiber guided light. Since their first fabrication in the late 1970s, fiber grating use and manufacture has increased significantly and they are now employed commercially as optical filters, reflectors and taps in telecommunications systems, and as strain and temperature sensors in various industries. We review optical fibers and the fundamentals of fiber grating characteristics as well as fabrication. Finally we consider several of the major industrial and scientific applications of fiber gratings.

Fiber Modes

A fiber grating couples light between the modes of an optical fiber, and therefore a knowledge of these modes is essential to understanding the characteristics of fiber gratings. In what follows, we emphasize single-mode fibers used near 1550 nm because these are very important in most practical applications that use fiber gratings. A single-mode fiber can be viewed as a thin cylinder of silica with a central raised index region of small enough diameter so that only a single mode is guided through total internal reflection at the core boundary. The electric field (*E*-field) of the fundamental mode propagating in the core is described by a characteristic transverse *E*-field profile and a propagation constant *k*:

$$\begin{aligned} E &= E(r, \varphi)\mathrm{e}^{-\mathrm{i}(\omega t - kz)} \\ k &= 2\pi n_{\mathrm{eff}}/\lambda \end{aligned} \qquad [1]$$

Here r and φ are the radial and azimuthal directions, z is measured longitudinally along the fiber, ω is the radial frequency, and λ is the vacuum wavelength, and the *E*-field points in the same direction everywhere. As the second equation indicates, the fiber mode propagates in the fiber as though it has an effective refractive index n_{eff}.

The silica cladding acts to isolate the core mode from the outside, hence the core mode penetrates very little into the cladding. However, if the outer cladding surface allows total internal reflection, the cladding may also guide light. Such 'cladding modes' are also important in understanding fiber gratings, since the grating couples the core mode with the cladding modes. As shown in Figure 1, the cladding modes extend outside the core into the cladding region, and their effective index is therefore always less than that of the core mode. We also note that in the case when the outer coating of the fiber does not support total internal reflection, 'leaky' cladding modes result. In the limit that there is no reflection at the cladding boundary a continuum of 'radiation modes' results. Although these modes are not guided, they are exploited in a variety of applications and can be carefully tailored by a suitably designed fiber grating. Note that the radiation and cladding modes exist regardless of the presence of a grating; a grating simply couples power from the core mode to the cladding/radiation modes; the grating is typically only a small perturbation on the index profile of the waveguide and the corresponding mode fields.

Fiber Grating Theory

Fiber gratings are longitudinal periodic variations in the refractive index (or, more generally, the dielectric function) of the core and/or cladding of an optical fiber. The scattering from any grating may be understood simply by considering the scattering off each successive period and adding these contributions while taking the phase of the *E*-field into account. Figure 2 illustrates this analysis in the case of a fiber Bragg reflector. When the reflections from each period of the grating add constructively, a strong back-reflection results. This is known as Bragg reflection. The condition for Bragg reflection is given by:

$$\lambda_{\mathrm{Bragg}} 2 n_{\mathrm{eff}} = \Lambda_{\mathrm{grating}} \qquad [2]$$

Here λ_{Bragg} is the vacuum wavelength of the light (equivalent to its frequency, $\omega = 2\pi c/\lambda$), n_{eff} is the effective index of the mode, and $\Lambda_{\mathrm{grating}}$ is the period of the grating refractive index modulations. This Bragg reflection is evident in a narrow dip in the

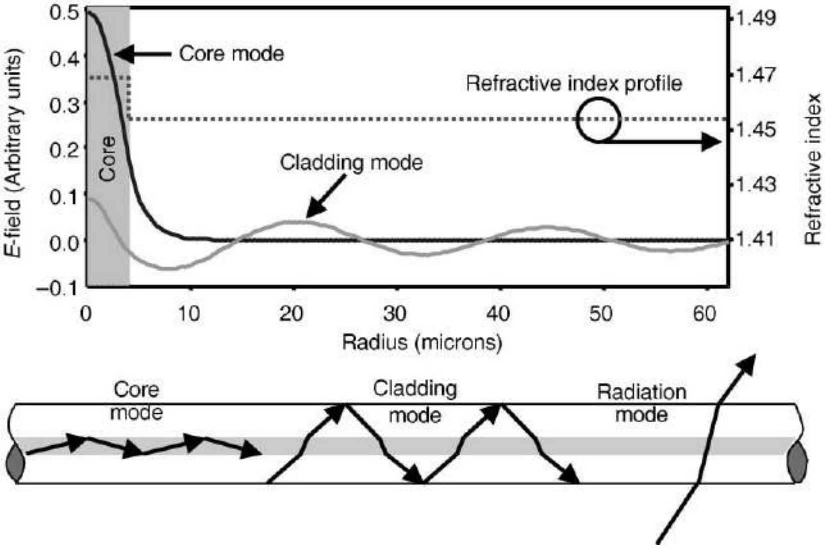

Figure 1 Fiber modes. Refractive index profile of a step index, single-mode fiber (dashed line). Radial profiles of mode *E*-fields of the fundamental (LP01) core mode, guided within the raised index region defined by the core, and one of the many higher-order cladding modes guided by the outer, air–silica interface. Also shown is a fiber indicating the ray paths for core, cladding, and radiation modes.

transmission spectrum of the grating (see Figure 3a). The ability of a fiber grating to selectively reflect the core light at one wavelength is one of its most important practical properties since it can be used as a narrowband filter or reflector. As a fiber component, the fiber grating has the advantage of compactness and provides in-line wavelength filtering.

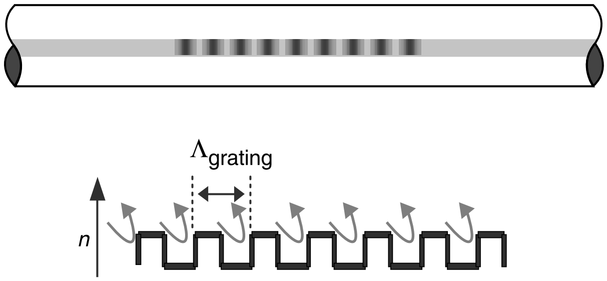

Figure 2 Bragg reflection. Each period of the grating reflects a small fraction of the incident light. When the period of the grating is adjusted so that these reflections add coherently, a strong Bragg reflection results and a narrow dip is observed in the transmission spectrum of the fiber (see **Figure 3a**). As shown in **Figures 3** and **4**, Bragg reflection can also occur into cladding modes.

The condition for Bragg reflection may be expressed more generally as a phase-matching condition. Phase matching refers to the condition when the phases of all the grating-scattered waves are the same and there is constructive interference. More specifically, this condition expresses the relationship between the spatial frequencies of the incident light, the scattered light, and the grating modulation, necessary for constructive interference in the scattered light. For a grating, the phase matching condition may be expressed in terms of the grating wavevector, $K_{\text{grating}} = 2\pi/\Lambda_{\text{grating}}$ and of the two mode propagation constants, k_{incident} and $k_{\text{scattered}}$ (defined in eqn [1])

$$k_{\text{incident}} - K_{\text{grating}} = k_{\text{scattered}} \qquad [3]$$

Using phase matching, we can understand the various transmission spectra that result from fiber

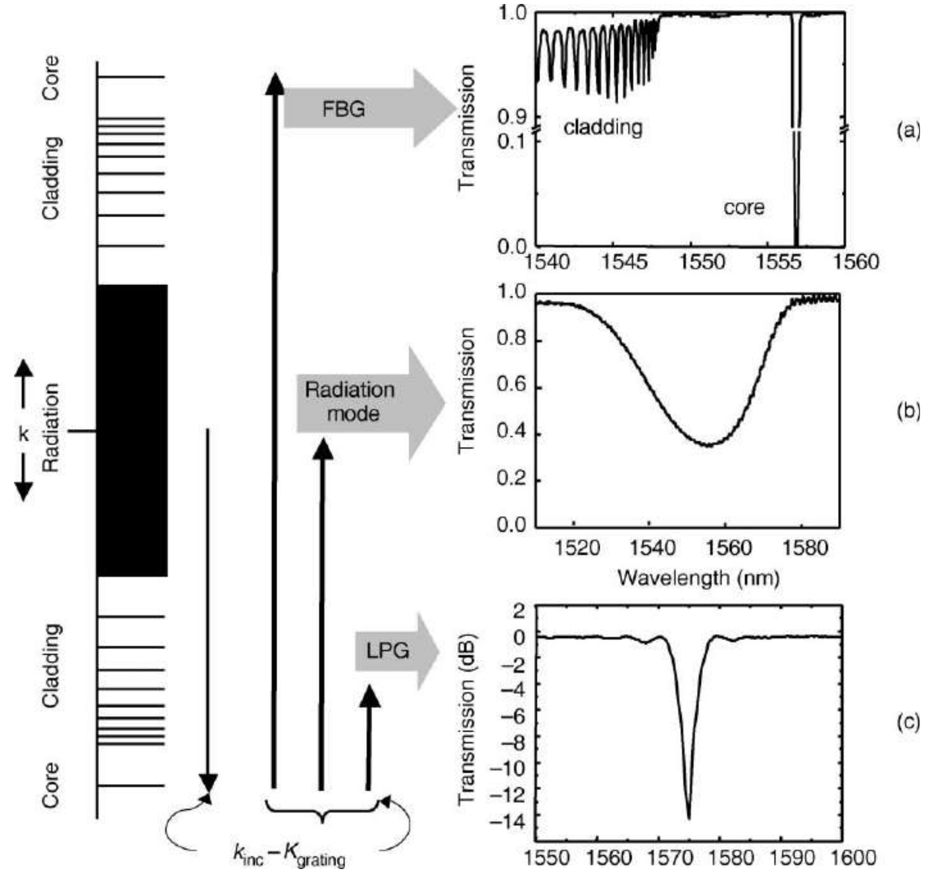

Figure 3 Phase matching. Fiber grating resonances occur when the phase matching condition is satisfied: $k_{\text{inc}} - K_{\text{grating}} = k_{\text{scat}}$. Propagation constants ($k = 2\pi n_{\text{eff}}/\lambda$) of the fiber modes are indicated as hash marks on the vertical line. The region indicates the regime of unguided radiation modes. The fiber grating is represented by its longitudinal spatial frequency (wavevector) $K_{\text{grating}} = 2\pi/\Lambda_{\text{grating}}$, where Λ_{grating} is the period of the grating. Assuming an incident forward-going core mode, the grating wavevector then phase matches to (a) counter-propagating core and cladding modes (fiber Bragg grating, FBG), (b) continuum of radiation modes, or (c) a copropagating cladding mode (long-period grating, LPG).

gratings. Figure 3 shows the transmission characteristic observed for core guided light in a single-mode fiber containing three different types of gratings: Bragg reflector, radiation mode coupler, and copropagating mode coupler. Each grating has a different type of phase matching condition. Figure 3 also depicts the phase matching condition in eqn [3] graphically. The vertical line contains a mark representing the propagation constants, k, of modes of a single mode fiber (both positive and negative for the two propagation directions) and a solid black region representing the radiation mode continuum. Bragg reflection into backward-propagating core and cladding modes occurs for short grating periods ($K_{\text{grating}} \sim 2k_{\text{core}}$). Such gratings are known as fiber Bragg gratings (FBGs) and a typical transmission spectrum with a strong Bragg reflection and several cladding mode resonances is shown in Figure 3a (a typical Bragg reflection spectrum is shown in Figure 8 below; the FBG grating period was ~500 nm). Core-cladding mode reflection results in loss resonances for wavelengths shorter than the core–core Bragg reflection. When the grating period is very long (in general, $K_{\text{grating}} < k_{\text{core}}(n_{\text{cladding}} - 1)/n_{\text{cladding}}$; typically Λ_{grating} is ~1000 times that of an FBG), phase matching to copropagating cladding modes occurs. Such gratings are known as long-period gratings (LPGs). A typical LPG spectrum is shown in Figure 3c. The LPG resonance corresponds to coupling between the core mode and a single copropagating cladding mode. For wavevectors in between these two regimes, coupling to the continuum of radiation modes is possible. An example of such a grating spectrum is shown in Figure 3b. We note that fiber cladding modes can be either well guided (resulting in the sharp grating resonances of Figure 3a) or poorly guided (resulting in a broad continuum of Figure 3b) depending on the fiber surface. The radiation mode approximation is valid in the limit that the cladding mode is very poorly guided (i.e., has high loss) within the grating region.

While phase matching yields the resonance wavelengths of a fiber grating, the strengths of these resonances must be derived from a full solution of Maxwell's equations. The most common approximation used in these calculations is known as coupled mode theory (CMT). In CMT, the longitudinal refractive index variation is assumed to be a small perturbation to the waveguide modes, which results only in a change in the propagation constant of the modes but does not change the transverse dependence of the mode E-fields (see eqn [1]). CMT then assumes that the E-field amplitude of each mode varies slowly (compared to wavelength) along the grating. A pair of coupled equations relating the two mode amplitudes as a function of position along the grating can then be derived, and these allow for computation of the grating reflection and transmission. The coupling between two modes is determined by an overlap integral κ_{12} between the product of the E-fields of the two modes (E_1 and E_2) and the periodic refractive index variation:

$$\kappa_{12} \propto \iint_{\text{Area}} E_1(r,\phi)E_2(r,\phi)\delta n(r,\phi)\,\mathrm{d}A \qquad [4]$$

where $\delta n(r,\phi)$ is derived from the refractive index modulation:

$$\delta n(r,\phi,z) = \delta n(r,\phi)\cos(K_{\text{grating}}z) \qquad [5]$$

This overlap integral is useful because it allows for a rough comparison of the strength of resonances for different modes without the need to obtain a full CMT solution. In the simple case of Bragg reflection of the core mode in a standard single-mode fiber (see Figure 1):

$$\kappa = \frac{\pi \delta n \eta}{\lambda}, \qquad [6]$$

where $\eta \approx 0.8$ is the overlap integral of the core mode intensity with the core.

To understand fiber grating spectra we first discuss a 'uniform' Bragg grating, of length L. Such a grating has uniform index modulation along its length and is shown in Figure 4a. Depending on the length and refractive index modulation of the grating, it may be classified as either weak or strong. The division between these two regimes may be understood by considering the light scattered by each successive period of the grating. The fraction of light scattered from each period can be approximated by $\delta n/2n$ using the Snell's law reflectance formula. The boundary between a strong and a weak grating occurs when the grating is either long enough or strong enough that the sum of the scattering from all periods is of order 1. This division may also be expressed in terms of κ and L: $\kappa L \sim 1$.

In the weak limit, the incident light propagates through the grating with very little scattering. In this case, both the LPG and FBG resonances have transmission spectra that are approximately related to the Fourier transform of the grating index profile. (For LPGs, this may not hold if modal dispersion is large; see the discussion below.) In the case of a uniform grating profile, this results in a $1 - \text{sinc}^2$ wavelength dependence of the transmission spectrum and a spectral width $\delta\lambda/\lambda \sim \Lambda_{\text{grating}}/L = 1/N_{\text{periods}}$ as shown in Figure 4a. An important feature of the uniform grating is the side lobes. The side lobes result from the sudden 'turn on' of the grating. In the FBG,

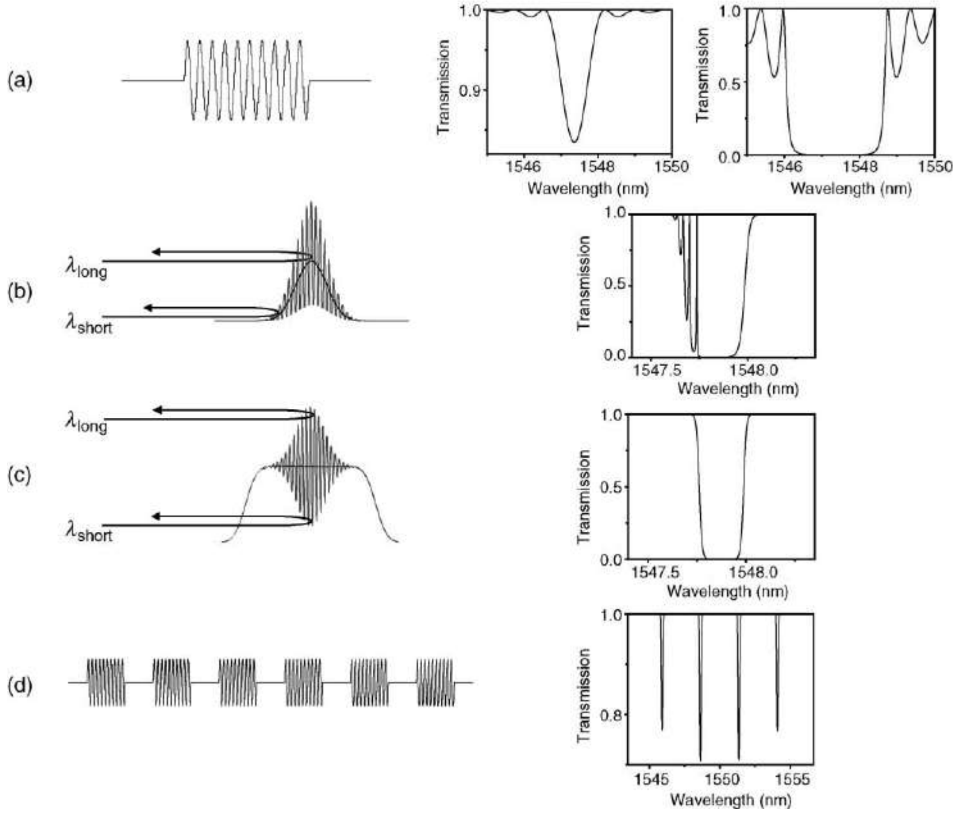

Figure 4 Fiber Bragg grating refractive index profiles and spectra. (a) A uniform profile with transmission spectra for both strong and weak index modulations. (b) An 'unapodized' Gaussian grating profile showing Fabry–Perot-like structure within the grating resonance. (c) A '100% dc apodized' Gaussian grating profile showing a symmetric transmission spectrum. (d) A superstructure grating exhibiting four resonance peaks.

they can be understood as resonances of the effective Fabry–Perot cavity defined by the sharp grating boundaries.

In the strong limit, a substantial fraction of light is scattered out of the incident mode, and the FBG and LPG transmission spectra become different. As shown in Figure 4a, the FBG spectrum widens to a width determined only by δn: $\delta\lambda/\lambda = \delta n/n$. The incident light penetrates only a short distance ($\sim \lambda/\delta n$) into the grating before being completely reflected. The strong LPG, on the other hand, becomes 'overcoupled' as the second mode is reconverted into the incident mode. The result is a higher transmission for the incident core mode (provided that the scattered mode propagates without loss).

While uniform gratings are easy to model and manufacture, most practical fiber gratings have a nonuniform profile. In a nonuniform grating profile, both the refractive index modulation amplitude, $\delta n_{\mathrm{AC}}(z)$, and the average effective index value, $\delta n_{\mathrm{DC}}(z)$, may vary along the fiber grating length:

$$\delta n(z) = \delta n_{\mathrm{DC}}(z) + \delta n_{\mathrm{AC}}(z)\cos(K_{\mathrm{grating}}z) \qquad [7]$$

By controlling the ac and dc index change, a greatly improved filter may be fabricated. Figure 4b shows the index profile and the grating transmission spectrum of a fiber Bragg grating with an 'unapodized' Gaussian profile. Note that due to the smooth increase in modulation amplitude, the sidebands of the uniform grating have been eliminated. However, because the dc index (and hence Bragg condition) is not constant within the grating, sharp resonant structure is observed within the grating resonance. This structure can be understood as Fabry–Perot resonances in the cavity formed by the nonuniform Bragg condition within the grating. Figure 4c shows a grating with a '100% dc apodized' Gaussian profile. This grating has a constant dc refractive index within the grating and therefore the resonance is smooth and symmetric. Most useful bandpass filters require a profile close to 100% apodization. Although not shown in Figure 4, a smoothly varying index profile also eliminates the sidelobes in LPG resonances.

Other important types of nonuniform fiber gratings include chirped fiber gratings and superstructure fiber gratings. In a chirped fiber grating the grating period slowly changes along the length of the grating. As discussed below, chirped fiber gratings may be used as chromatic dispersion compensators. They may also be used as broadband reflectors. LPGs may also be chirped to produce a broadband response. Superstructure gratings have oscillations in the period and/or amplitude of the refractive index modulations. The resulting grating can have several resonance peaks (see Figure 4d). The properties of nonuniform gratings may be computed using CMT and also understood intuitively using 'band-diagrams' and an effective index model of the fiber grating response.

Grating spectra showing radiation mode coupling are unlike FBGs or LPGs because there is no conversion of the radiation modes back to the incident core mode. The sharp resonances smooth out into a broad continuum. The grating transmission spectrum depends less on the refractive index profile of the grating than on the overlap of the core mode and grating with the radiation modes. Radiation mode coupling is greatly affected by tilting the grating planes with respect to the axis of the fiber. As discussed below, such tilted gratings may be designed as loss filters and optical fiber taps. Grating tilt has also been applied to decrease core mode back-reflection.

LPG resonances (and to some degree FBGs) may also exhibit strong dependence on the modal dispersion (or variation of the propagation constant, *k*, with wavelength). This wavelength dependence can dominate the wavelength dependence arising from CMT, making the resonance more or less narrow than predicted by CMT without modal dispersion. In an extreme case, very broadband LPG resonances (40 nm) may be produced between the core mode and a specially designed higher-order mode whose propagation constant varies such that it is resonant with the core mode over a large bandwidth.

Fiber Grating Fabrication

Photosensitivity

The key discovery leading to the study and practical application of fiber gratings was the phenomenon of UV-induced photosensitivity in the conventional germanosilicate optical fibers used in telecommunications systems. UV irradiation of the germanosilicate core region can produce refractive index changes in the range 0.01 to 0.0001. Such refractive index changes are large enough to produce useful filters and reflectors. The refractive index change is most efficiently produced by irradiation close to 242 nm or around 193 nm, where germanosilicate glass has strong absorptions. Numerous other dopants have been shown to be photosensitive at wavelengths ranging from 484 nm to 155 nm. These include P, Ce, Pb, Sn, N, Sn-Ge, Sb, and Ge-B. Other dopants, such as Al and F, show very little photosensitivity.

A second major discovery enabling practical manufacture of fiber gratings was the effect of 'hydrogen loading'. The fiber is placed in a pressurized hydrogen atmosphere until molecular hydrogen has diffused into the core region to a level of a few mole percent (a level similar to that of the photosensitive dopants). UV irradiation of the resulting fiber produces index changes an order of magnitude larger than in the unloaded fiber. Hydrogen loading allows gratings to be easily imprinted in conventional single-mode fibers, whose Ge content is too low to allow for strong gratings. The increase in photosensitivity is true for most photosensitive dopants. Index changes in excess of 10^{-2} have been produced with hydrogen loading. In order to avoid increased OH absorption in the telecommunications band (1520–1630 nm), deuterium (D_2) is normally used in place of hydrogen.

The origins of UV photosensitivity are not fully understood; however, at least two mechanisms are agreed to contribute to the index change. Firstly, as is evident from Figure 5, the UV absorption spectrum is modified during UV exposure. This bleaching corresponds to a change in refractive index in the infrared which may be computed using the Kramers–Kronig relations. In particular, the absorption band at

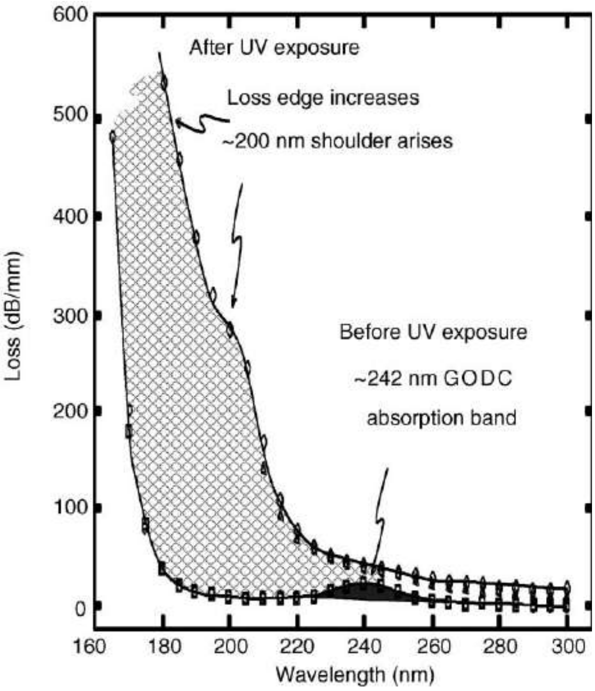

Figure 5 UV absorption spectrum of Ge-doped silica before and after irradiation with 242 nm light, showing bleaching of the Ge-oxygen deficient defect feature at 242 nm.

242 nm resulting from germanium oxygen deficient centers is bleached, indicating that these defects are modified during exposure. The infrared index change results both from this bleaching and from the change in the UV absorption edge at ~190 nm. Secondly, absorption of UV causes the silica matrix to contract. This compaction and the resulting stress distribution also change the refractive index.

Both mechanisms are enhanced by the presence of hydrogen in the silica lattice. The hydrogen reacts with oxygen, thus increasing the number of UV-absorbing germanium oxygen-deficient centers. As a result, in hydrogen loaded fibers, more of the Ge atoms participate in the index change, explaining the large increase in photosensitivity. The number of oxygen-deficient defects may also be increased during fiber manufacture by collapsing the fiber preform in a low-oxygen atmosphere. Such fibers show increased photosensitivity without the need for hydrogen loading. Co-doping a Ge-doped fiber with boron has also been shown to increase the level of photosensitivity of a fiber for a given level of index change in the core.

An important aspect of UV-induced index changes is that they require thermal annealing to ensure long-term stability. Sufficient exposure to elevated temperatures can completely erase the refractive index change in some Ge-doped gratings. Gratings can be completely erased by several hours of heating to greater than 500 °C. Thermal stability has been described by an activation energy model which assumes a distribution of energies for the defects giving rise to the refractive index change. Therefore, at a given anneal temperature a fixed fraction of the index change will quickly decay, leaving only stable defects that easily survive at lower temperatures. By annealing at several temperatures for a fixed time, a relationship between temperature and time required for a given refractive index decay may then be derived and used in accelerated aging tests of fiber gratings. For 20-year grating reliability of ~1% decrease in UV-induced refractive index change, a typical Ge-doped grating requires annealing at 140 °C for 12 hours.

Grating Inscription

The first method used in UV grating inscription is known as 'internal writing'. In this method light propagating in the fiber causes a standing wave to form within the core, and UV photosensitivity from a two-photon process results in permanent grating formation. Both LPG and FBG resonances have been fabricated using this method. The technique is inflexible, however, because the grating parameters are not easily controlled.

Another important advance that made possible the widespread use of fiber gratings is the 'transverse side-writing' technique. In this method, a UV interference pattern is projected through the side of the fiber onto the core region. The first demonstration of this technique employed a free space interferometer to produce the UV fringes. A significant improvement on this technique employs a zero-order nulled phase mask to generate the interference pattern. The phase mask acts to split the incident beam, and these two then form the interference pattern. The method is shown in Figure 6. The advantage of the phase mask

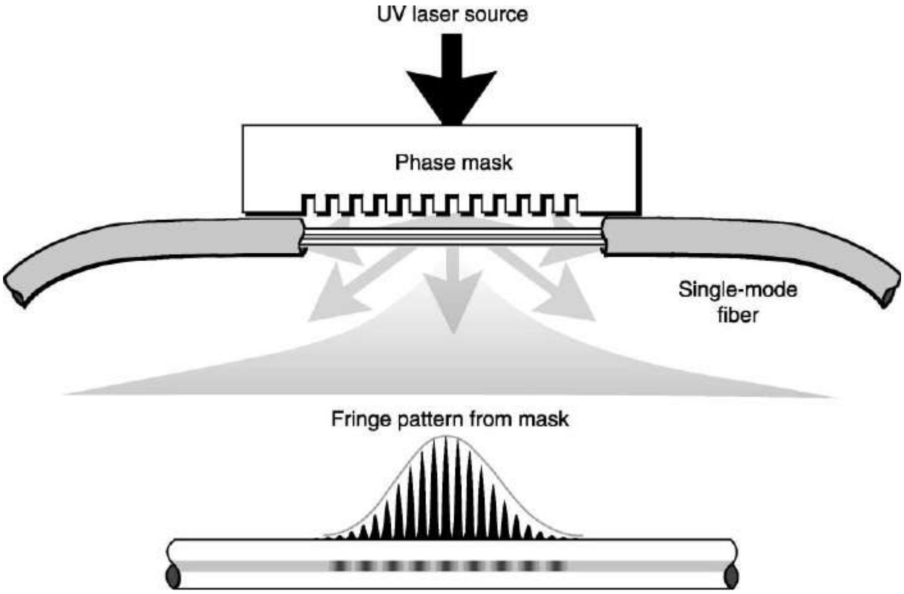

Figure 6 Phase mask fabrication of a fiber grating.

technique is that the fringes are very stable, since the UV beams propagate only a short distance. Another important technique for writing fiber gratings is the 'point-by point' writing method. In a point-by-point technique the UV interference pattern is modulated and translated along a fiber in such a way that a long grating may be formed. Although complex and sensitive, such methods can yield gratings as long as a few meters and also allow for the fabrication of complex grating profiles. Point-by-point methods have been used to fabricate broadband chirped dispersion compensating gratings and Bragg reflectors with very low dispersion.

LPGs may also be fabricated with UV irradiation. An amplitude mask is typically used. Requirements for beam coherence are greatly reduced because of the long period (typically a few hundred microns). Other techniques may also be used to fabricate LPGs. The fiber may be periodically deformed by heating or with periodic microbends. Dynamic LPGs may also be produced by propagating acoustic waves along the fiber.

Applications

Fiber gratings are key enabling technologies in a number of important applications. In this section we review these, highlighting the unique functionalities that can be provided by fiber gratings.

Fiber Grating Strain and Temperature Sensitivity

Fiber gratings are sensitive to both strain and temperature and may therefore be used as sensors. Both LPG and FBG resonances change with strain and temperature. LPGs show greater but more complicated sensitivity and typically FBGs are used in such applications. Strain sensitivity arises from the fact that gratings are extended in length. Because fibers are very thin (diameter ~125 μm), significant extension is possible. A smaller strain sensitivity arises from the stress optic effect. Typical sensitivity at 1550 nm is ~1 pm/1 $\mu\varepsilon$ in FBGs. Temperature sensitivity arises from temperature dependence of the refractive index and gives rise to a change order 0.011 nm/°C of temperature change of the FBG. FBG strain sensors have been applied to map stress distributions and can be employed in a distributed manner with many Bragg reflectors in one fiber being examined with one detector.

These sensitivities may also be used to make fiber grating filters tunable. Temperature changes are more stable but typically give 1–2 nm of change in an FBG resonance. Strain tuning is more difficult to package commercially but can give up to 4 nm in extension and more than 10 nm in compression. Passive thermal stabilization of grating resonances may also be achieved by designing packages in which strain and temperature variations cancel each other.

Fiber Grating Stabilized Diode Lasers and Fiber Lasers

One of the first important commercial applications of fiber gratings was in stabilization of fiber coupled diode lasers. Normally these lasers operate on several cavity modes making the output unstable. If feedback is provided in the form of a narrowband reflection then the laser output will be only in this narrow bandwidth. The fiber grating is implemented in the fiber pigtail of the diode laser and typically provides a few percent reflectivity over a bandwidth of ~1 nm. The application is depicted schematically in Figure 7.

More generally, fiber gratings have been used to realize fiber laser cavities and fiber distributed feedback lasers. One important example is the cascaded Raman resonator fiber laser, in which several FBGs recycle Raman pump light to produce high-power, fiber coupled light at any design wavelength.

Optical Add/Drop Multiplexers

Another important commercial FBG application is the optical add/drop multiplexer (see Figure 8). These are key components in wavelength division multiplexed (WDM) networks. WDM networks transmit large amounts of data through a single fiber by modulating several optical carrier wavelengths and sending all the wavelengths through a single fiber. The optical add/drop multiplexer allows individual wavelengths to be added or dropped from a given fiber.

Fiber Bragg gratings are useful in this application because of the very narrow resonances that are achievable (typically a few tenths of a nm at 1550 nm). For wavelengths close to the Bragg wavelength an FBG strongly reflects. The reflectivity and bandwidth are determined by the depth of the index modulation and grating length, as discussed earlier. These degrees of freedom make fiber gratings ideal filters to be used in wavelength division multiplexed lightwave networks, in which multiple channels are transmitted along long lengths of optical fiber. Figure 8 illustrates schematically the principle of

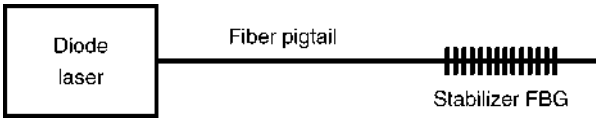

Figure 7 Fiber Bragg grating stabilized diode laser. The fiber grating reflector in the laser pigtail provides a small level of feedback to stabilize the laser output power and wavelength.

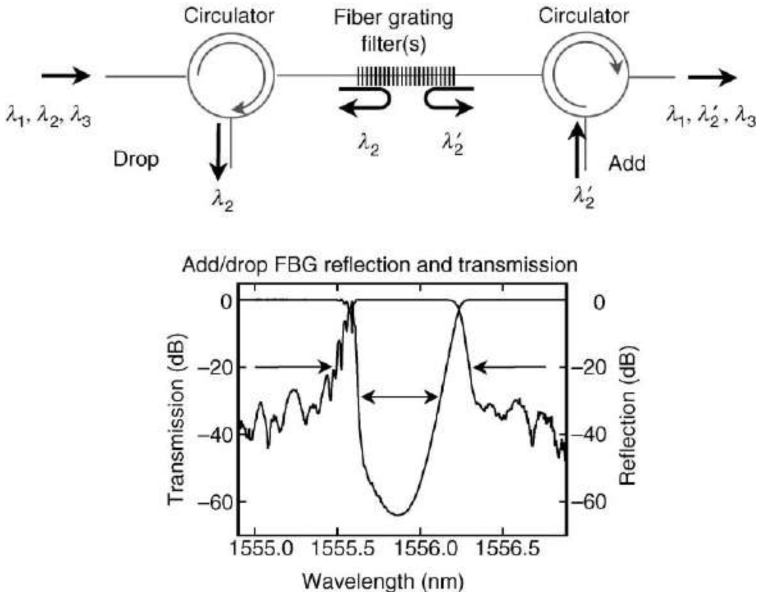

Figure 8 Top: Add/drop node employing fiber grating reflector. Bottom: characteristic reflection and transmission spectrum of an add/drop fiber grating filter.

a fiber grating based add/drop multiplexer. The grating is located between two three-port optical circulators. (A circulator is a bulk optic element that allows transmission only from port n to port $n + 1$.) Different channels are incident upon the grating from the left and only one particular channel undergoes reflection at the grating. Note that by carefully designing the grating properties the reflectivity in the adjacent channel can be made to be negligible. This one channel is reflected and re-routed by the three-port circulator. The remainder of the channels are transmitted through the grating. The grating also enables the 'add' functionality. In this case light originating from a different part of the optical network enters the link again through an optical circulator and is reflected off the fiber grating.

Gain-Flattening Loss Filters

A primary strength of fiber gratings is the diversity of filter shapes that is achievable through modification of the grating profile. Gain-flatteners are an important application requiring such filter design. Gain-flattening filters are important in WDM telecommunications systems in order to smooth the wavelength dependence of fiber amplifiers (such as Er or Raman amplifiers). Both LPGs and FBGs may be employed as gain flattening filters. The desired loss spectrum is translated into a corresponding grating profile. LPGs have the advantage of low return loss, but the LPG resonance wavelengths and strengths are very sensitive to the fiber geometry, making the computation of the grating profile less accurate. FBGs may also be used in both tilted and untilted form; however they typically have back-reflections that require careful fiber grating growth or an isolator to suppress back-reflections. Figure 9 shows an example of a gain equalizing loss filter using LPGs.

Dispersion Compensator

Chromatic dispersion can limit the transmission distances and capacities of long-haul fiber links. Chromatic dispersion arises from the variation in propagation velocity with wavelength. An initially sharp pulse is thus dispersed and begins to overlap with adjacent pulses resulting in degraded signal quality. A key technology for long-haul fiber links is therefore 'dispersion compensation'. The most common solution has been 'dispersion compensating fiber' (DCF), which is simply fiber in which the chromatic dispersion has been engineered to be the exact opposite of the dispersion in the fiber link. Limitations of DCF include nonlinearities, higher-order dispersion, and lack of tunability. Fiber gratings offer an alternative approach that overcomes the problems.

Fiber Bragg gratings can provide strong chromatic dispersion when operated in reflection with a three-port circulator (see Figure 10). In this case the grating period is made to vary along the length of the grating. This means that different wavelengths in a pulse reflect at different positions in the grating. This gives rise to wavelength-dependent group delay and therefore chromatic dispersion. The advantage of the fiber grating compared to DCF is that the device can be made to be very compact, with potentially lower

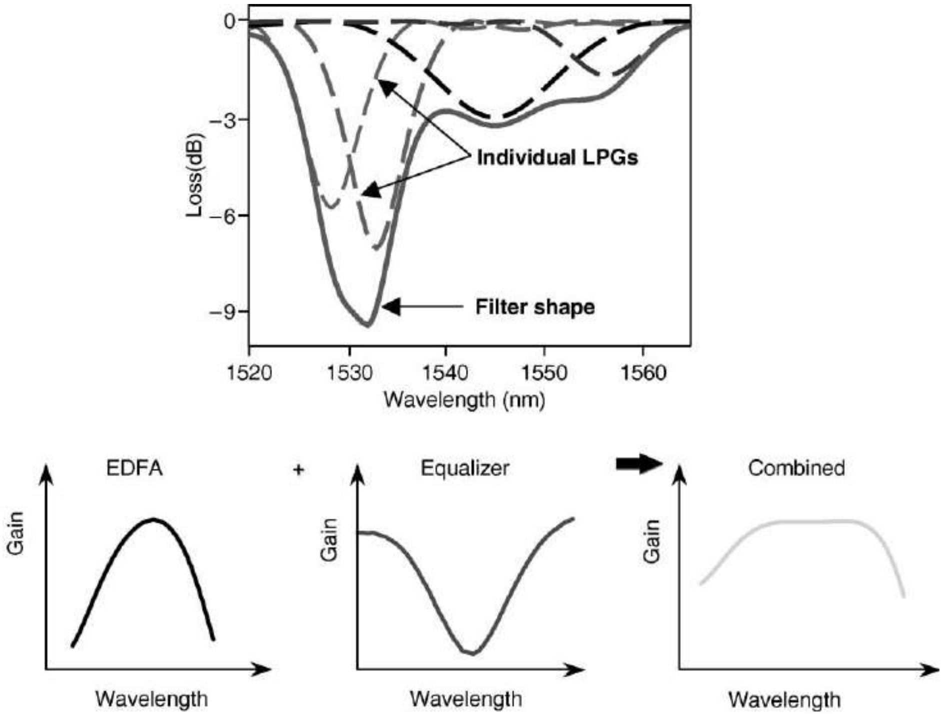

Figure 9 Long-period grating gain equalizer. Several individual gratings (dashed line) are combined to produce the overall shape (solid line).

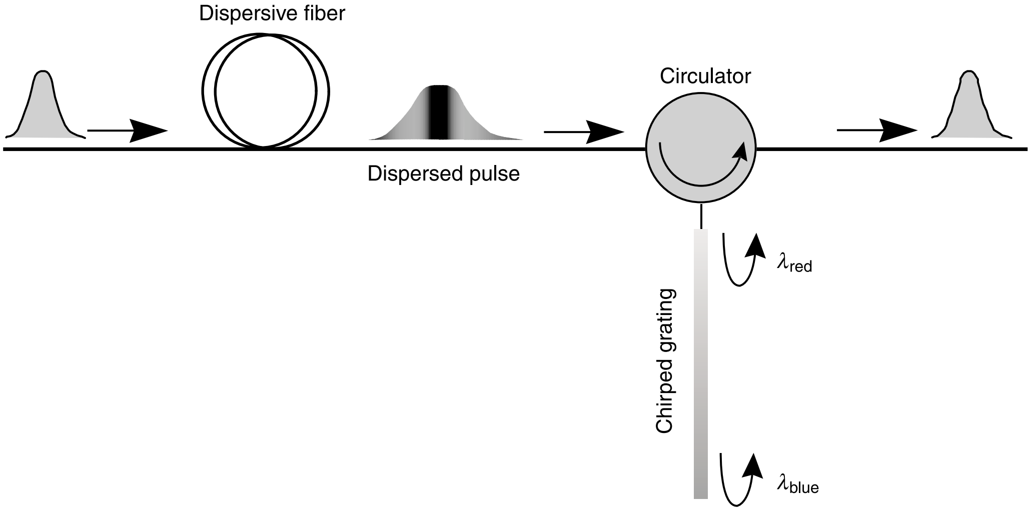

Figure 10 Chirped fiber grating dispersion compensator.

insertion loss, reduced optical nonlinearity, and controllable optical characteristics. In particular fiber gratings offer the possibility of engineering the chromatic dispersion to allow for compensation of dispersion slope and other degradations such as polarization mode dispersion.

Optical Monitors

Fiber gratings may also be employed as compact optical taps coupling core guided light to the radiation modes propagating outside the fiber (see Figure 3b). When a conventional Bragg grating is tilted with respect to the fiber axis it can scatter light out of the fiber in a highly directional manner. The angle of scatter is also wavelength dependent because of the phase matching condition discussed above. This wavelength dependence may form the basis of a compact wavelength monitor. The fiber grating is used to simultaneously couple light out of the fiber and separate the wavelengths. A simple

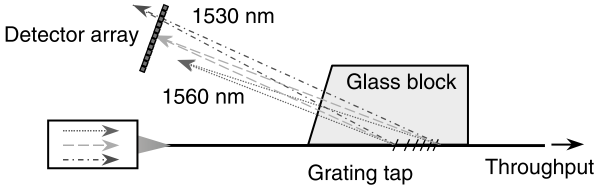

Figure 11 Fiber grating wavelength monitor. The fiber grating acts to scatter light out of fiber and simultaneously disperse the light.

detector array can then measure a spectrum of all the channels in a typical WDM system (see Figure 11).

Fiber grating taps may also scatter light in a polarization sensitive manner. This may be achieved by tilting the fiber grating by 45° with respect to the fiber axis. The resulting broadband taps may be employed in compact polarization monitoring schemes in which detectors are placed on the outside of the fiber to detect light at specific polarizations.

Advanced Topic in Fiber Gratings

Fiber Gratings in Air–Silica Microstructured Optical Fibers

Air–silica microstructure fibers (ASMFs), also known as 'holey' fibers, are fibers with air regions running along their length and allow for more complicated modal properties than conventional silica-doped fiber. It has been shown that fiber gratings can be UV written into photosensitive regions of various types of microstructure fibers. ASMF gratings open new possibilities for fiber grating devices. In addition, they are useful in studying the guidance properties of ASMFs, because they can selectively excite higher-order modes of the fiber. The phase matching condition may be used to derive the effective indices of the higher-order modes and mode profiles may be recorded by imaging a cleaved surface. Such measurements agree well with standard beam propagation computations. Figure 12 shows an example of a mode field image generated with a FBG.

One important application of ASMF gratings is a widely tunable LPG filter. Such filters employ a 'hybrid waveguide' formed by infusing a polymer into the air regions. Because the polymer index is much more sensitive to temperature than silica, the guidance properties of the cladding modes may be tuned. Thus, widely tunable LPG filters may be formed. Moreover, because the cladding modes are confined within an 'inner cladding' such filters are insensitive to external index, allowing for instance, the use of on-fiber, thin film heaters for power-efficient temperature tuning.

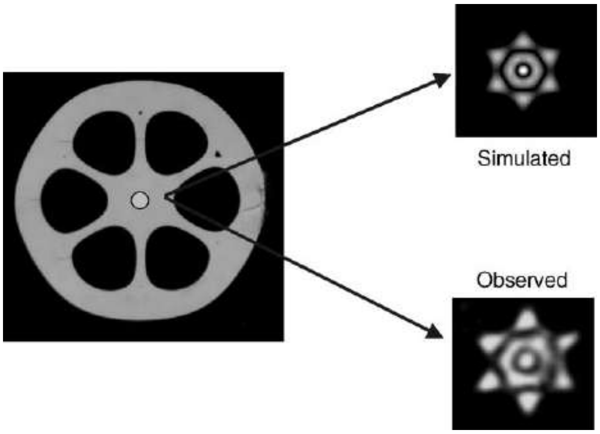

Figure 12 Cross-section of an air–silica microstructure fiber along with measured and simulated modes guided by the air regions.

Nonlinear Optics in Fiber Gratings

The linear properties of fiber gratings are well understood and as explained earlier lead to a number of interesting phenomena and applications. When the peak intensity of the optical field is very high a new class of nonlinear effects can occur. These optical nonlinearities arise from the intrinsic nonlinearity of the material system in which the grating is inscribed. In the case of optical fibers, the dominant nonlinearity is third-order nonlinearity which manifests itself in the refractive index of the material depending on the peak intensity of the optical field: $n = n_0 + n_2 I$. This nonlinearity gives rise to a number of dramatic nonlinear effects in fiber gratings, including all-optical switching, bistability, multistability, soliton generation, pulse compression, and wavelength conversion.

To understand these effects we consider an intense optical field incident upon a fiber grating with a central wavelength tuned inside the stopband of the grating where the reflectivity is high. At low intensities the pulse is reflected and only a small amount of light leaks through the grating. As the intensity of the incident optical field is increased the optical nonlinearity becomes significant and results in shifting of the local Bragg wavelength of the grating. This in turn can lead to an increase in transmission of light through the grating and bistable operation.

Theoretical studies of the continuous wave nonlinear properties of periodic media show that such media can exhibit switching as well as bistability and multistability. This prediction has been confirmed experimentally in a variety of geometries, such as in semiconductors, and in colloidal crystals. Experiments studying nonlinear effects in gratings can also be performed in optical fiber gratings. In these experiments short optical pulses are typically used in order to obtain the very high intensities. This has several key advantages. The first of these is that long, high-quality Bragg gratings can be permanently imprinted into the core of optical fibers by side illumination with interfering ultraviolet beams (see above). This allows for very long interaction lengths (compared to picosecond pulse lengths) and enables nonlinear propagation phenomena to be studied in detail. These include a number of propagation phenomena, such as the formation and propagation of Bragg solitons, modulational instability, and pulse compression. The second advantage is that although the nonlinearity is weak, it has a response time of the order of a few femtoseconds, and thus acts essentially instantaneously on the scale of all but the shortest optical pulses. Finally, the ability to engineer the fiber grating offers a number of unique geometries and potential applications of nonlinear fiber gratings.

A key idea in the interpretation of nonlinear optics in fiber Bragg gratings is the Bragg soliton. Briefly, in these solitons the enormous dispersion provided by the fiber Bragg grating is balanced by the fiber nonlinearity. This is analogous to the optical soliton that has been demonstrated in standard single-mode fiber where the dispersion is predominantly associated with the intrinsic material properties of silica. Since the dispersion of a Bragg grating can be five orders of magnitude larger than that of unprocessed fiber, Bragg solitons can be observed in gratings of only a few centimeters in length, thus rivaling the compactness of other soliton geometries, such as spatial solitons. Of course this implies that the required optical intensities to generate Bragg solitons are correspondingly higher than in uniform media.

Figure 13 shows an example of experimental results when the peak intensity of a pulse incident upon a fiber grating was approximately 10 GW/cm^2 and the incident pulse is tuned to the anomalous dispersion regime of the grating. In particular, Figure 13 shows the transmitted intensity versus time for different values of detuning. By varying the strain on the grating, we can change the detuning of the input pulse, and thus tune to different points on the dispersion curve. Note also from Figure 13 that far from the photonic bandgap edge, where the grating-induced dispersion is negligible, the pulse is unaffected and propagates through the grating at the speed of light. Closer to the edge of the photonic bandgap the pulse is substantially compressed due to the formation of a Bragg soliton. In addition to this compression, the transmitted pulse is substantially delayed in time as a result of the reduced group velocity of the pulse propagating through the grating.

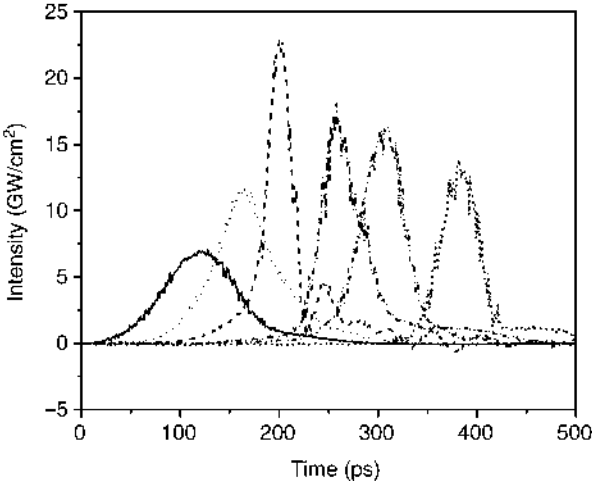

Figure 13 Pulse compression associated with nonlinear pulse propagation through a fiber Bragg grating. Compression is associated with the formation of a Bragg soliton when the incident pulse is tuned just outside of the photonic bandgap formed by the grating.

List of Units and Nomenclature

cw	Continuous wave
FBG	Fiber Bragg grating
LPG	Long-period grating

See also

Detection: Fiber Sensors. **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber; Light Propagation; Measuring Fiber Characteristics; Nonlinear Optics; Optical Fiber Cables. **Lasers:** Optical Fiber Lasers. **Optical Amplifiers:** Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Basic Concepts. **Optical Materials:** Optical Glasses. **Photon Picture of Light**.

Further Reading

Adams MJ (1981) *An Introduction to Optical Waveguides*. Chichester, UK: Wiley.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Eggleton BJ, Westbrook PS, Windeler RS, Spalter S and Strasser TA (1999) Grating resonances in air-silica microstructured optical fibers. *Optical Letters* 24: 1460–1462.

Eggleton BJ, Lenz G and Litchinitser NM (2000) Optical pulse compression schemes that use nonlinear Bragg gratings. *Fiber Integrated Optics* 19: 383–421.

Espindola RP, Windeler RS, Abramov AA, *et al.* (1999) External refractive index insensitive air-clad long period fiber grating. *Electronic Letters* 35: 327–328.

Hill KO, Fujii Y, Johnson DC and Kawasaki BS (1978) Photosensitivity in optical fiber waveguides: applications to reflection filter fabrication. *Applied Physics Letters* 32: 647–649.

IEEE Journal of Lightwave Technology (1997) Special issue on fiber gratings, photosensitivity and poling 15(8).

Kashyap R (1999) *Fiber Bragg Gratings*. New York: Academic Press.

Kogelnik H (1990) Theory of optical waveguides. In: Tamir T (ed.) *Guided-Wave Optoelectronics*. Berlin: Springer-Verlag.

Lemaire PJ, Atkins RM, Mizrahi V and Reed WA (1993) High-pressure H_2 loading as a technique for achieving ultrahigh UV photosensitivity and thermal sensitivity in GeO_2 doped optical fibers. *Electronic Letters* 29: 1191–1193.

Marcuse D (1974) *Theory of Dielectric Optical Waveguides*. New York: Academic Press.

Meltz G, Morey WW and Glenn WH (1989) Formation of Bragg gratings in optical fibres by transverse holographic method. *Optical Letters* 14: 823–825.

Othonos A and Kalli K (1999) *Fiber Bragg Gratings: Fundamentals and Applications in Telecommunications and Sensing*. Norwood, MA: Artech House.

Polladian L (1993) Graphical and WKB analysis of nonuniform Bragg gratings. *Physical Review E* 48: 4758–4767.

Westbrook PS, Eggleton BJ, Windeler RS, *et al.* (2001) Cladding mode loss in hybrid polymer-silica microstructured optical fiber gratings. *IEEE Photonic Technology Letters* 12: 495–497.

FOURIER OPTICS

S Jutamulia, University of Northern California,
Petaluma, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Coherent optical information processing is almost entirely based on the Fourier-transform property of a lens. A Fourier-transform lens is actually an ordinary lens. If the input transparency is placed in the front focal plane of the lens and illuminated with coherent collimated light (planewave), the amplitude function in the back focal plane of the lens will be the Fourier transform of the input transparency as shown in Figure 1. A coherent optical processor usually consists of two lenses, as shown in Figure 2, thus it performs two Fourier transforms successively. The first lens transforms the input function from space domain into frequency domain, and the second lens transforms back the frequency function from frequency domain to the output function in space domain.

Fourier Transform Property of Lens

Recently, the Fourier-transform property of lenses has been explained, simply using the concept of linear system and geometrical optics. In contrast, the Fourier-transform property of lenses is traditionally analyzed, based on the Fresnel–Kirchhoff diffraction theory involving quadratic phase terms (eqns [2] and [3]). In this article, we adopt the simple geometrical theory based on geometrical optics.

We start our discussion by looking at Figure 1. Consider that a coherent point source is at the center of the front focal plane of the lens, which is denoted by plane (x, y). The point source generates a collimated beam parallel to the optical axis. The amplitude function at the back focal plane of the lens, which is denoted by plane (u, v), can be described as unity. Consequently, we can express mathematically:

$$f(x, y) = \delta(x, y) \quad [1]$$

$$F(u, v) = 1 \quad [2]$$

where $f(x, y)$ is the amplitude function in the front focal plane, and $F(u, v)$ is the amplitude function in the back focal plane.

If the point source shifts a distance a along the y-axis, the amplitude function in the front focal plane is

$$f(x, y) = \delta(x, y + a) \quad [3]$$

The off-axis point source generates a tilted collimated beam. If we place a screen in the plane (u, v), we will see uniform intensity across the screen similar to that which would be seen when the source is centered on the axis. Mathematically, it means:

$$|F(u, v)|^2 = 1 \quad [4]$$

However, the phase of the amplitude function is not uniform, since the collimated beam is not parallel to

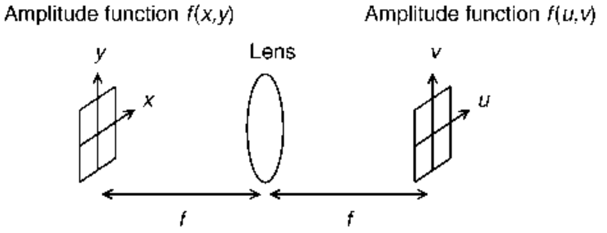
Amplitude function f(x,y)
Amplitude function f(u,v)
y
x
Lens
v
u
f
f

Figure 1 A lens Fourier-transforms amplitude function $f(x, y)$ in the front focal plane to amplitude function $F(u, v)$ in the back focal plane.

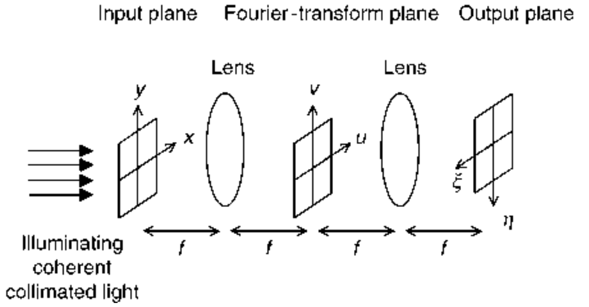
Input plane
Fourier-transform plane
Output plane
Lens
Lens
y
x
v
u
ξ
η
Illuminating coherent collimated light
f
f
f
f

Figure 2 A pair of lenses form 4-f coherent optical processor.

the optical axis. In other words, plane (u, v) is not parallel to the wavefront.

If the phase at the center of plane (u, v) is considered zero, the phase at any point at plane (u, v) can be calculated referring to Figure 3. Using simple geometry, we know that

$$d(u, v) = \frac{a}{f} v \qquad [5]$$

where $d(u, v)$ is the optical path difference. Correspondingly, the phase is

$$\phi(u, v) = \frac{2\pi}{\lambda f} av \qquad [6]$$

where λ is the wavelength of light.

In complex representation, the amplitude function is expressed as

$$F(u, v) = A(u, v) \exp[i\phi(u, v)] \qquad [7]$$

where $A(u, v)$ and $\phi(u, v)$ are magnitude and phase of the amplitude function, respectively. From eqn [4], we know that

$$A(u, v) = |F(u, v)| = 1 \qquad [8]$$

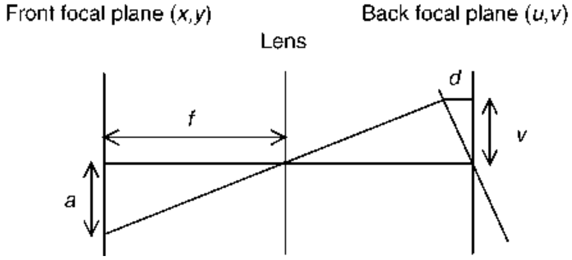
Front focal plane (x,y)
Back focal plane (u,v)
Lens
f
d
v
a

Figure 3 Calculation of phase function in the back focal plane (u, v) based on geometrical optics.

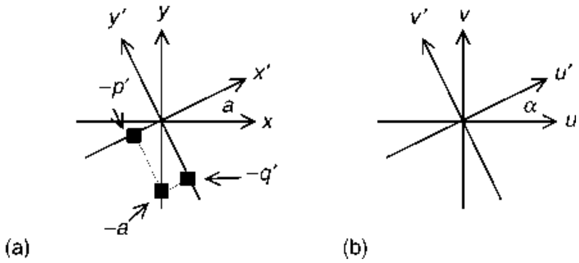
y′
y
−p′
x′
a
x
−q′
−a
(a)
v′
v
u′
α
u
(b)

Figure 4 (a) Front focal plane coordinate system (x, y) is rotated by α. (b) Similarly, back focal plane coordinate system (u, v) is rotated by α.

Substituting eqns [8] and [6] into [7] yields

$$F(u, v) = \exp\left(i \frac{2\pi}{\lambda f} av\right) \qquad [9]$$

Equation [9] is the amplitude function in the back focal plane of the lens, when the amplitude function in the front focal plane of the lens is given by eqn [3].

Consider now that coordinate (x,y) is rotated by angle α to new coordinates (x',y') as shown in Figure 4a. The amplitude function in the front focal plane now becomes

$$f(x', y') = \delta(x' + p', y' + q') \qquad [10]$$

where

$$p' = a \sin \alpha \qquad [11a]$$

$$q' = a \cos \alpha \qquad [11b]$$

Correspondingly, coordinate (u, v) is also rotated by the same angle α to new coordinates (u', v') as shown in Figure 4b. The old coordinates u and v can be described in terms of new coordinates u' and v' as follow:

$$u = u' \cos \alpha - v' \sin \alpha \qquad [12a]$$

$$v = u' \sin \alpha + v' \cos \alpha \qquad [12b]$$

Substituting eqn [12b] into eqn [6] yields

$$\phi(u', v') = \frac{2\pi}{\lambda f} a(u' \sin \alpha + v' \cos \alpha) \quad [13]$$

Further, substituting eqns [11a] and [11b] into [13] yields

$$\phi(u', v') = \frac{2\pi}{\lambda f}(p'u' + q'v') \quad [14]$$

Thus, the amplitude function in the back focal plane of lens is

$$F(u', v') = \exp\left[i \frac{2\pi}{\lambda f}(p'u' + q'v') \right] \quad [15]$$

For convenience, we now change parameters x' and y' to x and y, u' and v' to u and v, p' and q' to $-p$ and $-q$. Equations [10] and [15] become

$$f(x, y) = \delta(x - p, y - q) \quad [16]$$

$$F(u, v) = \exp\left[-i \frac{2\pi}{\lambda f}(pu + qv) \right] \quad [17]$$

We may simply consider that $f(x, y)$ is an input function and $F(u, v)$ is an output function of a system, since they are both physically the same amplitude functions. If there are two mutually coherent point sources at $(0, 0)$ and $(0, -a)$ with different amplitudes k and l in the front focal plane, the input function is

$$f(x, y) = k\delta(x, y) + l\delta(x, y + a) \quad [18]$$

These two point sources will generate two collimated beams. Since they are mutually coherent, two collimated beams will interfere, and the resultant amplitude function in the back focal plane is simply the superposition of two beams, which is

$$F(u, v) = k + l \exp\left(i \frac{2\pi}{\lambda f} av \right) \quad [19]$$

Thus, the lens is a linear system when the light is coherent. Furthermore, if input function $f(x, y)$ is a group of mutually coherent point sources with different amplitudes (i.e., an optical transparency illuminated with a planewave), it can be expressed as follows:

$$f(x, y) = \int\int_{-\infty}^{\infty} f(p, q)\delta(x - p, y - q) \mathrm{d}p \, \mathrm{d}q \quad [20]$$

Since the lens is a linear system, output function will be the same superposition as follows:

$$F(u,v) = \int\int_{-\infty}^{\infty} f(p,q) \exp\left[-i \frac{2\pi}{\lambda f}(pu + qv) \right] \mathrm{d}p \mathrm{d}q \quad [21]$$

By again changing parameters p and q to x and y, respectively, eqn [21] becomes

$$F(u,v) = \int\int_{-\infty}^{\infty} f(x,y) \exp\left[-i \frac{2\pi}{\lambda f}(xu + yv) \right] \mathrm{d}x \mathrm{d}y \quad [22]$$

which is the optical Fourier transform of the input function $f(x,y)$.

4-f Coherent Optical Processor

If the amplitude function in the front focal plane of a lens is denoted by $f(x, y)$, then the amplitude function in the back focal plane is simply $F(u, v)$ given in eqn [22], which is the Fourier transform of $f(x, y)$. Mathematically, $f(x, y)$ can also be derived from $F(u, v)$ (eqn [4]). After neglecting the normalization factor, it can be expressed as follows:

$$f(x, y) = \int\int_{-\infty}^{\infty} F(u, v) \exp\left[i \frac{2\pi}{\lambda f}(xu + yv) \right] \mathrm{d}u \, \mathrm{d}v \quad [23]$$

The normalization factor $(1/\lambda f)^2$ is caused by the coordinate scale factor λf in the Fourier-transform plane (u,v).

Equations [22] and [23] are Fourier transform and inverse Fourier transform, respectively. The operation of Fourier transform is represented by a notation $F\{\cdot\}$, and inverse Fourier transform is $F^{-1}\{\cdot\}$.

Two Fourier transform lenses can be combined into a two-lens processor, which is better known as a 4-f optical processor, as shown in Figure 2. Figure 2 illustrates the principle of a 4-f optical processor, in which $f(x, y)$, $F(u, v)$, and $f(\xi, \eta)$ are amplitude functions in input, Fourier transform, and output planes, respectively. Since $F(u, v)$ is the function of spatial frequency, Fourier-transform plane is also called a frequency plane. Since a lens performs only Fourier transform, to perform inverse Fourier transform, the output plane must be rotated by 180°. Thus, orientations of ξ and η correspond to $-x$ and $-y$.

In practice, $f(x, y)$ is represented by an optical mask, which can be a film transparency or a spatial light modulator. The optical mask with $f(x, y)$ amplitude transmittance is illuminated with collimated coherent light to produce the amplitude function $f(x, y)$. The Fourier transform $F(u, v)$ is physically formed in the frequency plane. The frequency content of the input function $f(x, y)$ can be directly altered by placing an optical mask in the frequency plane. Finally, an image with amplitude function $f(\xi, \eta)$ is formed in the output plane. Notice that all detectors, including our eyes, detect intensity that is the square of amplitude.

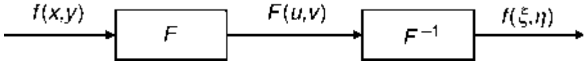

Figure 5 Block diagram of 4-f coherent optical processor.

According to eqn [22], $F(u, v)$ can be expressed in terms of $f(x, y)$ as

$$F(u, v) = F\{f(x, y)\} \qquad [24]$$

Similarly, the amplitude function in the output plane must be

$$f(x', y') = F\{F(u, v)\} \qquad [25]$$

where (x', y') is the coordinate of the output plane having the same orientation as (x, y) in the input plane. This function can be expressed in terms of inverse Fourier transform as follows:

$$f(-x', -y') = F^{-1}\{F(u, v)\} \qquad [26]$$

A rotated coordinate (ξ, η), shown in Figure 2, substitutes $(-x', -y')$, such that

$$f(\xi, \eta) = F^{-1}\{F(u, v)\} \qquad [27]$$

Thus, the successive operations can be simply illustrated by the diagram shown in Figure 5.

Spatial Filtering

The origin of optical information processing can be traced as far back as the work of Abbe of 125 years ago. His interest at that time concerned the theory of microscopy in order to establish a scientific approach to microscope design. He found that a larger aperture resulted in higher resolution. An object under the microscope, whose size was of the order of the wavelength of light, diffracted the illuminating light, i.e., the straight light was bent. If the aperture of the objective lens was not large enough to collect all of the diffracted light, the image did not correspond exactly to that object. Rather, it related to a fictitious object whose complete diffracted light pattern matched the one collected by the objective lens. These lost portions are associated with the higher spatial frequency. Their removal will result in a loss in image sharpness and resolution. The basic ideas of Abbe's theory were later clarified in a series of experiments by Porter in the early 1900s.

Since the late 1940s, the possibility of applying coherent optical techniques to image processing had been explored. In 1960, Cutrona *et al.* presented an optical system using lenses to perform Fourier transforms. Thus, every Fourier transformation in the processor can be realized by a lens. The power of this architecture is its capability of forming Fourier spectra and also being cascaded.

Consider an image formed by an optical system such as a camera; because of the error or limitation of the system, the perfect point in the object may smear. If we consider that the original object is the collection of a large number of very fine points, the formed image will be the collection of the same number of blurred spots, instead of fine points. Thus, the image formation can be expressed by convolution. The blurred spot is represented by the impulse response $h(x, y)$, which is the characteristic of the optical system.

When the input is a point (an impulse), its amplitude function is a delta function $\delta(x, y)$, and the output amplitude function is $h(x, y)$. The observed image is the intensity function $|h(x, y)|^2$. Since it indicates the spread of a point, $|h(x, y)|^2$ is called the point spread function.

The output image function $o(x, y)$ is the convolution of input function $f(x, y)$ and the impulse response $h(x, y)$ as follows:

$$o(x, y) = \iint_{-\infty}^{\infty} f(p, q)h(x - p, y - q)\mathrm{d}p\,\mathrm{d}q \qquad [28]$$

Notice that $h(x - p, y - q)$ is the blurred spot centered at (p, q). $H(u, v)$, $F(u, v)$, and $O(u, v)$ are Fourier transforms of $h(x, y)$, $f(x, y)$, and $o(x, y)$, respectively. Substitution of

$$h(x-p, y-q) = \iint_{-\infty}^{\infty} H(u,v) \times \exp\left[i\frac{2\pi}{\lambda f}[(x-p)u+(y-q)v] \right]\mathrm{d}u\,\mathrm{d}v \qquad [29]$$

then

$$F(u,v) = \iint_{-\infty}^{\infty} f(p,q)\exp\left[-i\frac{2\pi}{\lambda f}(pu+qv) \right]\mathrm{d}p\,\mathrm{d}q \qquad [30]$$

yields

$$o(x,y) = \iint_{-\infty}^{\infty} F(u,v)H(u,v)\exp\left[i\frac{2\pi}{\lambda f}(xu+yv) \right]\mathrm{d}u\,\mathrm{d}v \qquad [31]$$

Since

$$o(x,y) = \iint_{-\infty}^{\infty} O(u,v)\exp\left[i\frac{2\pi}{\lambda f}(xu+yv) \right]\mathrm{d}u\,\mathrm{d}v \qquad [32]$$

we obtain

$$O(u,v) = H(u,v)F(u,v) \qquad [33]$$

The function $H(u,v)$, which is the Fourier transform of the impulse response $h(x,y)$, is called transfer function. Actually, $H(u,v)$ changes the spectrum of input function $F(u,v)$ to become $O(u,v)$ and the spectrum of the output function becomes $H(u,v)F(u,v)$.

The first application of spatial filtering may be the restoration of a photograph. The image formed by a camera is a collection of spots. If the image is in focus, the size of the spots is very fine, and thus the image looks sharp. If the image is out of focus, the spot size is fairly large and the image is blurred. Interestingly, an out-of-focus image can be restored using spatial filtering technique.

Consider that the original image is $f(x,y)$. Because it is out of focus, a point in the image becomes a spot expressed by $h(x,y)$. If the image is in focus, $h(x,y)$ is a delta function. If the in-focus image is $f(x,y)$, its frequency function is $F(u,v)$. The defocused image $o(x,y)$ is expressed by eqn [28], and its frequency function is $F(u,v)H(u,v)$. To produce the in-focus image, the frequency function must be $F(u,v)$. Thus $F(u,v)H(u,v)$ must be multiplied by the filter function of $1/H(u,v)$ to produce $F(u,v)$ only. The filter of $1/H(u,v)$ can be made from a photographic plate to provide its magnitude, and a glass plate having two values of thickness to provide +1 and −1 signs. The glass plate is essentially a phase filter. The photographic plate with varying density function is an amplitude filter. By passing frequency function $F(u,v)H(u,v)$ through the combination of amplitude and phase filters, it will result in $F(u,v)$ only. When $F(u,v)$ is again Fourier transformed by a lens, an in-focus image $f(x,y)$ will be produced.

Complex Matched Spatial Filtering

The restoration of a photograph mentioned above, utilizes the filter function $1/H(u,v)$, which has positive and negative real values. We will see an application in which the filter function is a complex function. A complex function filter can be produced by a single hologram, instead of using a combination of a phase filter and an amplitude filter. The complex matched spatial filter can perform a correlation, as first demonstrated by Vander Lugt. The complex filter is actually a Fourier transform hologram proposed by Leith and Upatnieks.

The synthesis of a complex filter is shown in Figure 6. First, a function $g(x,y)$ is displayed in the input plane by a transparency or spatial light modulator and illuminated with a collimated coherent light beam. Its Fourier-transform $G(u,v)$ is generated in the back focal plane of the lens. The hologram is made in the Fourier-transform plane by recording the interference pattern of $G(u,v)$ and the reference beam $R(u,v)$. $R(u,v)$ is an oblique collimated beam, having magnitude of unity, which is expressed mathematically as

$$R(u,v) = \exp(iau) \qquad [34]$$

and

$$a = \frac{2\pi \sin\alpha}{\lambda} \qquad [35]$$

where α is the angle of the reference beam and λ is the wavelength of light.

The transmittance of the recorded hologram will be

$$\begin{aligned} T(u,v) &= |G(u,v) + R(u,v)|^2 \\ &= |G(u,v)|^2 + 1 + G(u,v)\exp(-iau) \\ &\quad + G^*(u,v)\exp(iau) \qquad [36] \end{aligned}$$

where * denotes the complex conjugate. The hologram, which is called a complex matched spatial filter, is then placed back in the Fourier-transform plane, as shown in Figure 7. A new input function $f(x,y)$ is displayed in the input plane. The Fourier-transform of input function, $F(u,v)$ will be generated in the Fourier-transform plane. The amplitude of light,

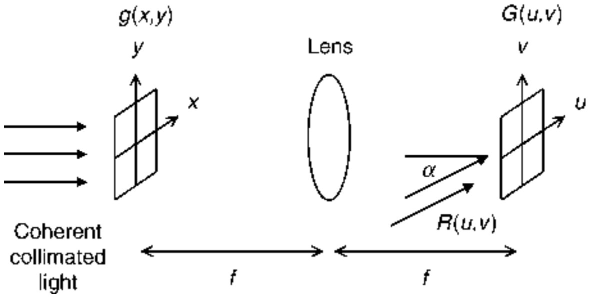

Figure 6 Synthesis of complex matched spatial filter is the recording of Fourier-transform hologram.

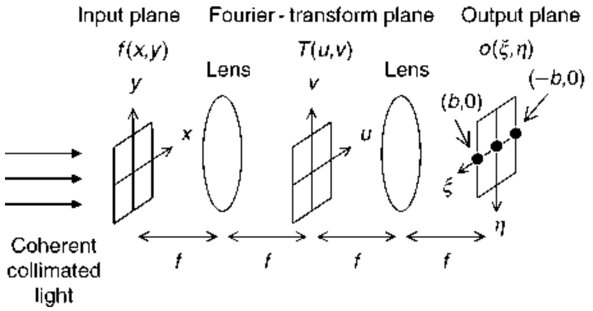

Figure 7 Optical processing using complex matched spatial filter in a 4-f processor.

immediately after the complex matched spatial filter, is

$$F(u,v)T(u,v) = (|G(u,v)|^2+1)F(u,v) + F(u,v)G(u,v)\exp(-iau) + F(u,v)G^*(u,v)\exp(iau) \qquad [37]$$

The second lens then produces the inverse Fourier transform of $F(u,v)T(u,v)$ in the output plane:

$$F^{-1}\{F(u,v)T(u,v)\} = F^{-1}\{(|G(u,v)|^2+1)F(u,v)\} + F^{-1}\{F(u,v)G(u,v)\exp(-iau)\} + F^{-1}\{F(u,v)G^*(u,v)\exp(iau)\} \qquad [38]$$

The first term will be at the center of the output plane. The second and the third terms will appear at $(b,0)$ and $(-b,0)$ in (ξ,η) plane, respectively, where

$$b = f\sin\alpha \qquad [39]$$

and f is the focal length of lens.

We are now in a position to analyze the inverse Fourier transform of $F(u,v)G(u,v)$ and $F(u,v)G^*(u,v)$. Referring to eqns [28]–[33], the inverse Fourier transform of $F(u,v)G(u,v)$ is the convolution of $f(\xi,\eta)$ and $g(\xi,\eta)$. Thus, the complex matched spatial filtering can also be applied to the restoration of photography mentioned previously without using the combination of amplitude and phase filters.

The inverse Fourier transform of $F(u,v)G^*(u,v)$ is derived as follows:

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} F(u,v)G^*(u,v) \times \exp\left[i\frac{2\pi}{\lambda f}(\xi u+\eta v)\right]\mathrm{d}u\,\mathrm{d}v \qquad [40]$$

Substitute

$$F(u,v) = \iint_{-\infty}^{\infty} f(\alpha,\beta) \times \exp\left[-i\frac{2\pi}{\lambda f}(\alpha u+\beta v)\right]\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [41]$$

and

$$G^*(u,v) = \iint_{-\infty}^{\infty} g^*(-p,-q) \times \exp\left[-i\frac{2\pi}{\lambda f}(pu+qv)\right]\mathrm{d}p\,\mathrm{d}q \qquad [42]$$

into eqn [40], one obtains

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta)\iint_{-\infty}^{\infty} g^*(-p,-q) \times \iint_{-\infty}^{\infty}\exp\left[i\frac{2\pi}{\lambda f}(\xi-\alpha-p)u + (\eta-\beta-q)v\right]\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}p\,\mathrm{d}q\,\mathrm{d}\alpha\,\mathrm{d}\beta$$

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta)\iint_{-\infty}^{\infty} g^*(-p,-q) \times \delta(\xi-\alpha-p,\eta-\beta-q)\mathrm{d}p\,\mathrm{d}q\,\mathrm{d}\alpha\,\mathrm{d}\beta$$

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta)g^*(\alpha-\xi,\beta-\eta)\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [43]$$

Equation [43] is the correlation of $f(x,y)$ and $g(x,y)$. It is apparent that the correlation output $o_3(\xi,\eta)$ will achieve the maximum when $f(x,y)$ is identical to $g(x,y)$. If $f(x,y)$ and $g(x,y)$ are different, the correlation function will be a flat function with minor random fluctuation. Optical correlators using complex matched spatial filters have been applied to automatic pattern recognition.

Joint Transform Correlator

Alternatively, the joint transform correlator can perform a correlation without recording a complex matched spatial filter in advance. In addition to the correlation, the joint transform correlator can also be used as a general optical processor.

A transparency with amplitude transmittance $f(x,y)$ is centered at $(-a,0)$, and a second transparency with amplitude transmittance $g(x,y)$, is centered at $(a,0)$. The two transparencies are illuminated by a coherent collimated beam and jointly Fourier transformed by a single lens as shown in Figure 8. In the back focal plane of the lens, the amplitude function $S(u,v)$ is

$$S(u,v) = \iint_{-\infty}^{\infty}[f(x+a,y)+g(x-a,y)] \times \exp\left[-i\frac{2\pi}{\lambda f}(xu+yv)\right]\mathrm{d}x\,\mathrm{d}y \qquad [44]$$

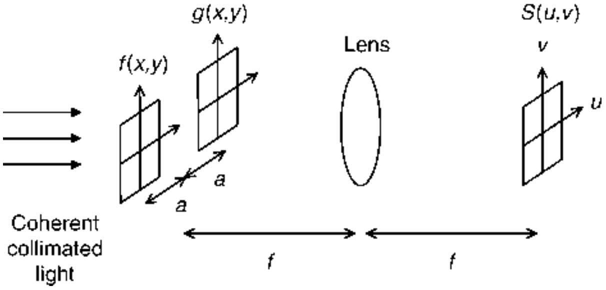

Figure 8 Joint Fourier transform of two input functions.

$S(u,v)$ is the joint Fourier spectrum of $f(x,y)$ and $g(x,y)$. It can be written as

$$S(u,v) = F(u,v)\exp\left[i\frac{2\pi}{\lambda f}au\right] + G(u,v)\exp\left[-i\frac{2\pi}{\lambda f}au\right] \qquad [45]$$

where $F(u,v)$ and $G(u,v)$ are Fourier transforms of $f(x,y)$ and $g(x,y)$, respectively.

If a square-law detector is placed in the back focal plane of the lens, the recorded pattern of joint power spectrum $|S(u,v)|^2$ is

$$\begin{aligned}|S(u,v)|^2 = {} & |F(u,v)|^2 \\ & + F(u,v)G^*(u,v)\exp\left[i\frac{4\pi}{\lambda f}au\right] \\ & + F^*(u,v)G(u,v)\exp\left[-i\frac{4\pi}{\lambda f}au\right] \\ & + |G(u,v)|^2 \end{aligned} \qquad [46]$$

where * denotes the complex conjugate. A photographic plate is a square-law detector and can be used to record $|S(u,v)|^2$. When the developed plate is placed in the input plane, as shown in Figure 9, the input transmittance is $|S(u,v)|^2$ as given in eqn [46].

The inverse Fourier transform of the joint power spectrum $|S(u,v)|^2$, in the back focal plane of the lens, is

$$o(\xi,\eta) = \iint_{-\infty}^{\infty} |S(u,v)|^2 \exp\left[i\frac{2\pi}{\lambda f}(\xi u + \eta v)\right]\mathrm{d}u\,\mathrm{d}v \qquad [47]$$

which consists of four terms, according to eqn [46]. We now analyze this equation term by term. The first term is

$$o_1(\xi,\eta) = \iint_{-\infty}^{\infty} |F(u,v)|^2 \exp\left[i\frac{2\pi}{\lambda f}(\xi u + \eta v)\right]\mathrm{d}u\,\mathrm{d}v \qquad [48]$$

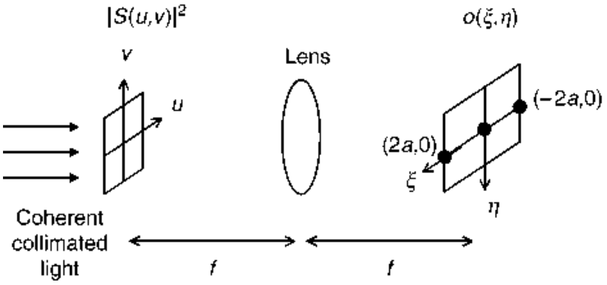

Figure 9 Fourier transform of the joint power spectrum.

which is the auto-correlation of $f(x,y)$

$$o_1(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta) f^*(\alpha-\xi,\beta-\eta)\,\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [49]$$

The second term is

$$o_2(\xi,\eta) = \iint_{-\infty}^{\infty} F(u,v)G^*(u,v)\exp\left[i\frac{4\pi}{\lambda f}au\right] \times \exp\left[i\frac{2\pi}{\lambda f}(\xi u + \eta v)\right]\mathrm{d}u\,\mathrm{d}v \qquad [50]$$

which is

$$o_2(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta) g^*[\alpha-(\xi+2a),\beta-\eta]\,\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [51]$$

Equation [51] is the correlation between $f(x,y)$ and $g(x,y)$, shifted by $-2a$ in the ξ-axis direction.

Similarly, the third term will be the correlation between $g(x,y)$ and $f(x,y)$, shifted by $2a$ in the ξ-axis direction:

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} g(\alpha,\beta) f^*[\alpha-(\xi-2a),\beta-\eta]\,\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [52]$$

And the fourth term is

$$o_4(\xi,\eta) = \iint_{-\infty}^{\infty} g(\alpha,\beta) g^*(\alpha-\xi,\beta-\eta)\,\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [53]$$

which is the auto-correlation of $g(x,y)$ and overlaps with the first term at the center of the output plane.

From the analysis given above, it is understood that the correlation of $f(x,y)$ and $g(x,y)$, and the correlation of $g(x,y)$ and $f(x,y)$, are shifted by $-2a$ and $2a$, respectively, along the ξ-axis. The correlation outputs are centered at $(-2a,0)$ and $(2a,0)$ in the back focal plane of the lens. If $f(x,y)$ and $g(x,y)$ are the same, the two correlation outputs are identical. The first and fourth terms are the autocorrelations of $f(x,y)$ and $g(x,y)$, centered at $(0,0)$. Thus the second and third terms, which are correlation signals, can be isolated. A versatile real-time joint transform correlator, using inexpensive liquid crystal televisions, is widely used in laboratory demonstration.

See also

Information Processing: Coherent Analogue Optical Processors. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Abbe E (1873) Beitrage zur Theorie des Mikroskops und der mikroskopichen Wahrnehmung. *Archiv Mikroskopie Anatomie* 9: 413–468.

Bracewell RN (1965) *The Fourier Transform and Its Applications*. New York: McGraw-Hill.

Cutrona LJ, Leith EN, Palermo CJ and Porcello LJ (1960) Optical data processing and filtering systems. *IRE Transactions on Information Theory* IT-6: 386–400.

Goodman JW (1968) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Jutamulia S (1992) Joint transform correlators and their applications. *Proceedings of the SPIE* 1812: 233–243.

Jutamulia S and Asakura T (2002) Optical Fourier-transform theory based on geometrical optics. *Optical Engineering* 41: 13–16.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America* 52: 1123–1130.

Porter AB (1906) On the diffraction theory of microscopic vision. *Philosophical Magazine* 11: 154–166.

Rau JE (1966) Detection of differences in real distributions. *Journal of the Optical Society of America* 56: 1490–1494.

Tsujiuchi J (1963) Correction of optical images by compensation of aberrations and by spatial frequency filtering. *Progress in Optics* 2: 133–180.

Vander Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* IT-10: 139–145.

Weaver CS and Goodman JW (1966) A technique for optically convolving two functions. *Applied Optics* 5: 1248–1249.

Yu FTS (1983) *Optical Information Processing*. New York: Wiley.

Yu FTS and Ludman JE (1988) Joint Fourier transform processor. *Microwave Optical Technology Letters* 1: 374–377.

Yu FTS, Jutamulia S, Lin TW and Gregory DA (1987) Adaptive real-time pattern recognition using a liquid crystal TV based joint transform correlator. *Applied Optics* 26: 1370–1372.

GEOMETRICAL OPTICS

Contents

Lenses and Mirrors
Aberrations
Prisms

Lenses and Mirrors

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lenses are carefully shaped pieces of transparent material used to form images. The term 'transparent' does not necessarily mean that the lens is transmitting visible light. The element germanium, well-known in the transistor industry, is a light-gray opaque crystal which forms images with infrared radiation, and provides the optics for night-vision instruments. The lenses used in cameras, telescopes, microscopes, and other well-known applications, are made from special glasses or – more recently – from plastics. Hundreds of different optical glass formulas are listed in the manufacturers' catalogues. Plastics have come into use in recent years because it is now possible to mold them with high precision. The way in which lenses form images has been studied for several hundred years, resulting in an enormous and very complicated literature. In this article, we show how this theory can be simplified by introducing the restriction that lenses are perfect. This limitation provides an easy introduction to the image forming process. A companion article **Geometrical Optics:** Aberrations, will deal with imperfect components, showing how their defects degrade images and how corrections are made.

The Cardinal Points and Planes of an Optical System

The location, size and orientation of an image are items of information which a designer needs to know. The nature of the image can be determined by a simple model of an optical system, based on two experimental observations. The first is well-known and is pictured in Figure 1. Parallel rays of light passing through a double convex lens (a simple magnifying glass, for example) should meet at the focal point or focus, designated as F'. For rays from the right side of the lens, another such point F exists at an equal distance from the lens. The other experiment uses the lens to magnify (Figure 2a). The amount of magnification decreases as the lens is brought closer to 'ISAAC NEWTON' (Figure 2b), and when the lens touches the paper, object and image are approximately the same size. The locations of object and image corresponding to equality in size are called the unit or principal points H and H', respectively, and the set of four points F, F', H, and H' are the cardinal points. Therefore, the planes through these points normal to the axis of the lens are the cardinal planes. Figure 3 shows an object whose image we can now determine. The coordinate orientation in this figure may appear strange at first. It is customary in optics to designate the lens axis as OZ, and the other two axes form a right-handed system. The x-axis then lies in the

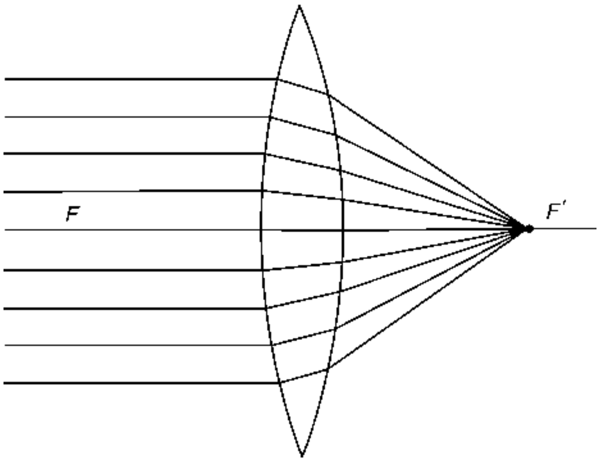

Figure 1 Focal points of a double convex lens.

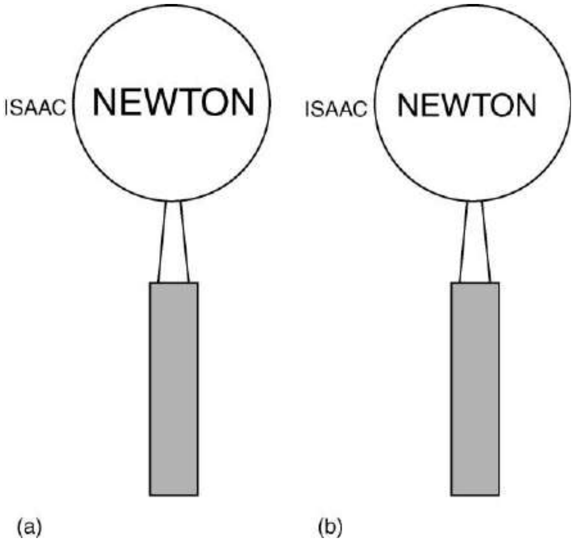

(a)

(b)

Figure 2 Magnification by a double convex lens. The magnification in (a) is reduced when the lens is brought closer to the image (b).

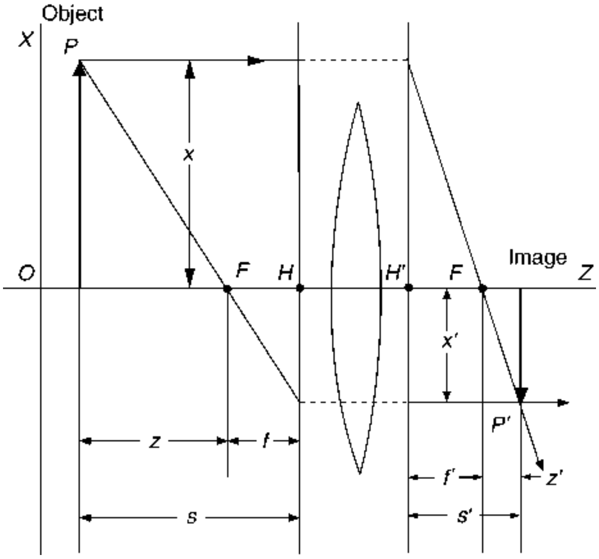

Figure 3 Ray tracing using cardinal points and planes.

plane of the figure; the y-axis, coming out of the paper and perpendicular to it, is not shown. The object, of height x, is placed to the left of the focal point F. The cardinal points occur in the order F, H, H', F', since we expect – based on the experiment corresponding to Figure 2 – that the unit points are very close to the lens. Another fact to be demonstrated later is that H and H' are actually inside the lens. Two rays are shown leaving the point P at the top of the object. The ray which is parallel to the axis strikes the unit plane through H and continues to the right, completely missing the lens. The ray then meets the unit plane through H' at a point which must be a distance x from the axis. To understand why, we recognize that if the object were relocated so as to lie on the object-space unit plane, the ray emerging at the other unit plane will be at a distance x from the axis; as required by the definition of these planes. We then assume that the ray is bent or refracted at the unit plane at H' and proceeds to, and beyond, the focal point F'. The second ray, from P through the focal point F, is easily traced since the lens is indifferent to the direction of the light. Assume temporarily that we know the location of the image point P'. Let a parallel ray leave this point and travel to the left. The procedure just given indicates that this ray will strike the unit plane at H', emerge at the other unit plane, be refracted so as to pass through F and reach the object point P. Then reversing the direction of this ray, it will start at P, emerge at the unit plane H' and travel parallel to the axis, intersecting the other ray at P'. The distance between F and H is called the object space focal length f, with f' being the corresponding image space quantity. Simple geometry, making use of the similar triangles and the distances indicated in this diagram, can be used to derive the Newton lens equation or the equivalent Gauss lens equation. Detailed derivations of these two equations will be found in either of the sources listed in the Further Reading at the end of this article. It is not usually made clear in physics texts that these equations apply only to the special case of a single, thin lens, and we shall not apply them here. The image in this figure is real, inverted, and reduced. That is, it is smaller than the object, oriented in the opposite direction, and can be projected onto a screen located at the image position. This result can be verified by using an ordinary magnifier to form the image of a distant light source on a sheet of white paper.

Paraxial Matrices

To consider how light rays behave in optical systems, we introduce the index of refraction n of a material medium, defined as the ratio of the velocity of light c in a vacuum to its velocity v in that medium, or:

$$n = \frac{c}{v} \qquad [1]$$

The velocity of light in air is approximately the same as it is in a vacuum. A light ray crossing the interface between air and a denser medium such as glass will be bent towards the normal to the surface. This is the phenomenon of refraction and the amount of bending is governed by Snell's law, which has the form:

$$n \sin\theta = n' \sin\theta' \qquad [2]$$

where n and n' are the indices of refraction of the two media and θ and θ' are the angles as measured from the normal. Figure 4 shows a ray of light leaving an object point P and striking the first surface of a lens at point P_1. It is refracted there, and proceeds to the second surface. All rays shown lie in the z, x-plane or meridional plane, which passes through the symmetry axis OZ. The amount of refraction is specified by Snell's law, eqn [2], which we shall now simplify. The sine of an angle can be approximated by the value of the angle itself, if this value is small when expressed in radians (less than 0.1 rad), and Snell's law simplifies to

$$n_1\theta_1 = n_1'\theta_1' \qquad [3]$$

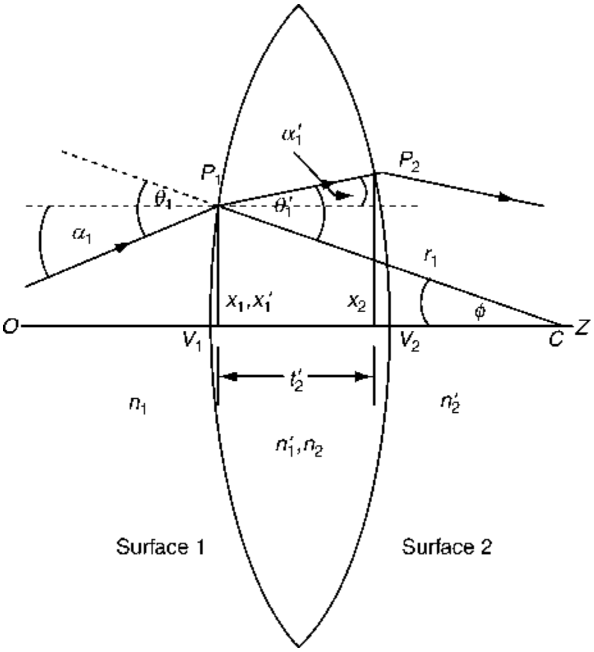

Figure 4 Ray passing through a double convex lens.

This is called the paraxial form of Snell's law; the word paraxial means 'close to the axis'. It turns out, however, that the Snell's law angles are not convenient to work with, and we eliminate them with the terms:

$$\theta_1 = \alpha_1 + \phi, \qquad \theta_1' = \alpha_1' + \phi \qquad [4]$$

where α_1 is the angle which the incident ray makes with the axis OZ, α_1' is the corresponding angle for the refracted ray, and ϕ is the angle which the radius r_1 (the line from C to P_1) of the lens surface makes at the center of curvature C. This particular angle can be specified as

$$\sin\phi = \frac{x_1}{r_1} \qquad [5]$$

but using the paraxial approximation simplifies this to

$$\phi = \frac{x_1}{r_1} \qquad [6]$$

Substituting for the angles gives

$$n_1'\alpha_1' = \frac{n_1 - n_1'}{r_1} + n_1\alpha_1 \qquad [7]$$

Notice that the distance from the point P_1 to the axis is labeled as either x_1 or x_1'. This strange notation leads to the trivial relation

$$x_1 = x_1' \qquad [8]$$

and this can be combined with eqn [7] to obtain the matrix equation:

$$\begin{pmatrix} n_1'\alpha_1' \\ x_1' \end{pmatrix} = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [9]$$

where the constant k_1 is called the refracting power of surface 1 and is defined as

$$k_1 = \frac{n_1' - n_1}{r_1} \qquad [10]$$

To review the procedure for matrix multiplication, the first element $n_1'\alpha_1'$ in the one-column product matrix is calculated by multiplying the upper left-hand corner of the 2×2 matrix with the first element of the one-column matrix to its right, obtaining $1(n_1\alpha_1)$, and to this is added the product $-k_1x_1$ of the upper right-hand element in the 2×2 matrix and the second member of the one-column matrix. This square matrix is called the refraction matrix R_1 for surface 1, defined as

$$R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \qquad [11]$$

Next, we look at what happens to the ray as it travels from surface 1 to surface 2. As it goes from P_1 to P_2, its distance from the axis becomes

$$x_2 = x'_1 + t'_1 \tan \alpha'_1 \qquad [12]$$

or using the paraxial approximation for small angles:

$$x_2 = x'_1 + t'_1 \alpha'_1 \qquad [13]$$

Another approximation we shall use is to regard t'_1 as being equal to the distance between the lens vertices V_1 and V_2. Using the identity:

$$\alpha_2 = \alpha'_1 \qquad [14]$$

leads to a second matrix equation:

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'_1/n'_1 & 1 \end{pmatrix} \begin{pmatrix} n'_1\alpha'_1 \\ x'_1 \end{pmatrix} \qquad [15]$$

where the 2×2 translation matrix T_{21} is defined as

$$T_{21} = \begin{pmatrix} 1 & 0 \\ t'_1/n'_1 & 1 \end{pmatrix} \qquad [16]$$

To get an equation which combines a refraction followed by a translation, we take the one-column matrix on the left side of eqn [9] and substitute it into eqn [15] to obtain

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = T_{21} R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [17]$$

Note that the 2×2 matrices appear in right to left order and that the multiplication of two square matrices is an extension of the rule given above. This equation gives the location and slope of the ray which strikes surface 2 after a refraction and a translation. To continue the ray trace at surface 2, we introduce a second refraction matrix R_2 by expressing k_2 in terms of the two indices at this surface and the radius. Then eqn [17] is extended to give the relation:

$$\begin{pmatrix} n'_2\alpha'_2 \\ x'_2 \end{pmatrix} = R_2 T_{21} R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [18]$$

This process can obviously be applied to any number of components. The product of the three 2×2 matrices in the above equation is known as the system matrix S_{21} and it completely specifies the effect of the lens on the incident ray passing through it. It is also written as

$$S_{21} = R_2 T_{21} R_1 = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \qquad [19]$$

where the four quantities a, b, c, and d are known as the Gaussian constants. They are used extensively in specifying the behavior of a lens or an optical system. We now state the sign and notation conventions which are necessary for the application of paraxial matrix methods:

1. Light normally travels from left to right.
2. The optical systems we deal with are symmetrical about the z-axis. The intersections of the refracting or reflecting surfaces with this axis are the vertices and are designated in the order encountered as V_1, V_2, etc.
3. Positive directions along the axes are measured from the origin in the usual Cartesian fashion, so that horizontal distances (that is, along the z-axis) are positive if measured from left to right. Angles are positive when measured up from the z-axis.
4. Quantities associated with the incident ray are unprimed; those for the refracted ray are primed.
5. A subscript denotes the associated surface.
6. If the center of curvature of a surface is to its right, the radius is positive, and vice versa.
7. There is a special rule for mirrors to be explained below.

Using the Gaussian Constants

Figure 5 shows the double convex lens of Figure 4 with the assumed locations of the cardinal points indicated. These positions, which we shall now determine accurately, are at distances designated as l_F, l'_F, l_H, and l'_H and are measured from the associated vertex. The object position can be called t, a positive quantity which is measured to the right from object to the first vertex, or it can be called t_1 if measured in the opposite direction. For the first choice, the matrix specifying the translation from object to lens will have the quantity t/n_1 in its lower left-hand corner. However, it is both logical and convenient to use the first vertex as the reference point. Hence, we replace t/n_1 in the translation matrix with the quantity $-t_1/n_1$, and remember to specify t_1 as a negative number when calculating the image position. The equation connecting object and image is obtained by starting with this matrix, multiplying it by the system matrix of eqn [19], and finally by a translation matrix

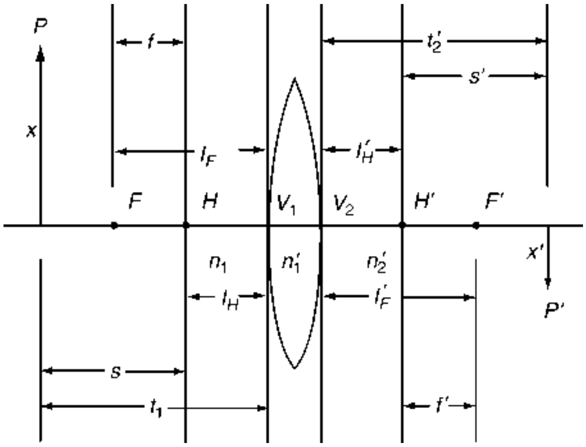

Figure 5 Definitions of cardinal plane locations.

corresponding to the translation t_2' from lens to image to obtain:

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t_2'/n_2' & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ -t_1/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \qquad [20]$$

where α or α_1 is the angle that the ray from the top of the object makes with the z-axis. This equation takes the initial value of the inclination α and of the position x of the ray leaving the object, and determines the final values, α' and x' at the image. The product of the three 2×2 matrices on the right-hand side can be consolidated into a single matrix called the object-image matrix $S_{P'P}$. Then eqn [20] can be written as

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = S_{P'P} \begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \qquad [21]$$

and this matrix has the complicated form:

$$S_{P'P} = \begin{pmatrix} b + \dfrac{at_1}{n_1} & -a \\ \dfrac{bt_2'}{n_2'} + \dfrac{at_1t_2'}{n_1n_2'} - d - \dfrac{ct_1}{n_1} & c - \dfrac{at_2'}{n_2'} \end{pmatrix} \qquad [22]$$

If we put this matrix into the previous equation, we obtain an unsatisfactory result: the value of x' will depend on the angle α_1 made by the incident ray, as can be seen by multiplying the two matrices on the right side. The ratio of x' to x is called the magnification m; that is

$$m = \frac{x'}{x} \qquad [23]$$

A perfect image can be formed only if the magnification is determined solely by the object distance and the constants of the lens. To eliminate this difficulty, the lower left-hand element in the matrix, eqn [22] is required to be equal to zero, and the magnification is then:

$$m = \frac{x'}{x} = c - \frac{at_2'}{n_2'} \qquad [24]$$

The determinant of this matrix is unity, since it is the product of three matrices whose individual determinants are unity. It follows that

$$\frac{1}{m} = b + \frac{at_1}{n_1} \qquad [25]$$

so that

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = \begin{pmatrix} 1/m & -a \\ 0 & m \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \qquad [26]$$

Using the fact that the lower left-hand element of the matrix eqn [22] is zero shows that

$$\frac{t_2'}{n_2'} = \frac{d + ct_1/n_1}{b + at_1/n_1}, \qquad \frac{t_1}{n_1} = \frac{d - bt_2'/n_2'}{-c + at_2'/n_2'} \qquad [27]$$

This equation connects the object distance t_1 with the image distance t_2', both quantities being measured from the appropriate vertex. This relation is known as the generalized Gauss lens equation. It applies to all lens systems, no matter how complicated.

We can determine the location of the unit planes by letting $m = 1$. This t_2' in eqn [27] is the location l_H' of the image space unit plane, and it becomes

$$l_H' = \frac{n_2'(c-1)}{a} \qquad [28]$$

This relation expresses the location of the unit plane on the image side as the distance from V_2 to H'. Similarly, the location of H with respect of V_1 is

$$l_H = \frac{n_1(1-b)}{a} \qquad [29]$$

To locate the focal planes, consider a set of parallel rays coming from infinity and producing a point image at F'. The terms containing t_1 in eqn [27] are much larger than d or b, and this relation becomes

$$l'_F = \frac{n'_2 c}{a} \qquad [30]$$

Letting t'_2 become infinite in the right-hand half of eqn [27], the location of F is given by the equation

$$l_F = \frac{-n_1 b}{a} \qquad [31]$$

The focal lengths f and f' shown in Figure 5 were defined earlier as the distances between their respective focal and unit planes. Hence

$$f' = l'_F - l'_H = \frac{n'_2}{a} \qquad [32]$$

and

$$f = l_F - l_H = \frac{-n_1}{a} \qquad [33]$$

If we let $n_1 = n'_2 = 1$ for air, these become

$$-f = f' = \frac{1}{a} \qquad [34]$$

The Gaussian constant a is thus the reciprocal of the focal length in air. Even when the lens is not in air, it is still true that $f' = -f$ if the lens is in a single medium. It is also true regardless of whether or not the lens is symmetric.

It is customary in optics to work with a consistent set of units, such as centimeters or millimeters, and the sign conventions used for x and y are the standard Cartesian rules. As mentioned above, curved surfaces which open to the right have a positive radius and vice versa. As an example, let a double convex lens have radii of 2.0 and 1.0, respectively, an index of refraction of 1.5, and a thickness of 0.5. These quantities are then denoted as

$$\begin{aligned} &r_1 = +2.0, \qquad r_2 = -1.0, \qquad t'_1 = t_2 = +0.5 \\ &n_1 = 1.0, \qquad n'_1 = 1.5 = n_2, \qquad n'_2 = 1.0 \end{aligned} \qquad [35]$$

By eqn [10]:

$$k_1 = \frac{n'_1 - n_1}{r_1} = \frac{1.5 - 1.0}{2.0} = 0.25 \qquad [36]$$

and

$$k_2 = \frac{1.0 - 1.5}{-1.0} = 0.50 \qquad [37]$$

The system matrix S_{21} is

$$\begin{aligned} S_{21} &= R_2 T_{21} R_1 \\ &= \begin{pmatrix} 1 & -0.50 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.5/1.5 & 1 \end{pmatrix} \begin{pmatrix} 1 & -0.25 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0.83 & -0.71 \\ 0.33 & 0.92 \end{pmatrix} \end{aligned} \qquad [38]$$

and we note that the determinant is

$$(0.83)(0.92) - (0.71)(-0.33) = 1.00 \qquad [39]$$

This property of the system matrix provides a very useful check on the accuracy of matrix multiplications. The locations of the four cardinal points can be determined from the formulas given above and their locations are shown in the scale drawing of Figure 6. Since we know how to locate the cardinal points of a lens, we are in a position to ray trace in a precise manner, using the method of Figure 3 and adding a third ray. After locating F, H, H', and F', we no longer need the lens; the cardinal planes with the proper separation are fully equivalent to the lens they replace. This is very much like what electrical engineers do when they replace a complicated circuit with a black box; this equivalent circuit behaves just like the original. Ray tracing becomes more complicated when we deal with diverging lenses (Figure 7), but the procedure gives above still applies. Parallel rays spread apart and do not come to a focus on the right-hand side of the lens. However, if the refracted

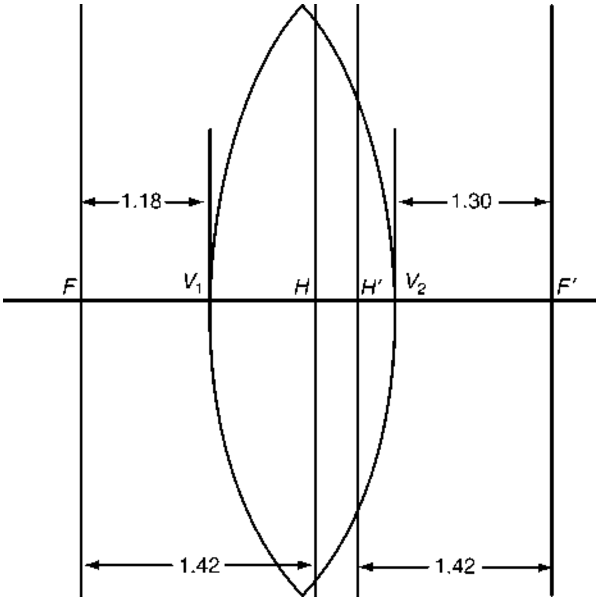

Figure 6 Positions of cardinal planes for an asymmetric lens.

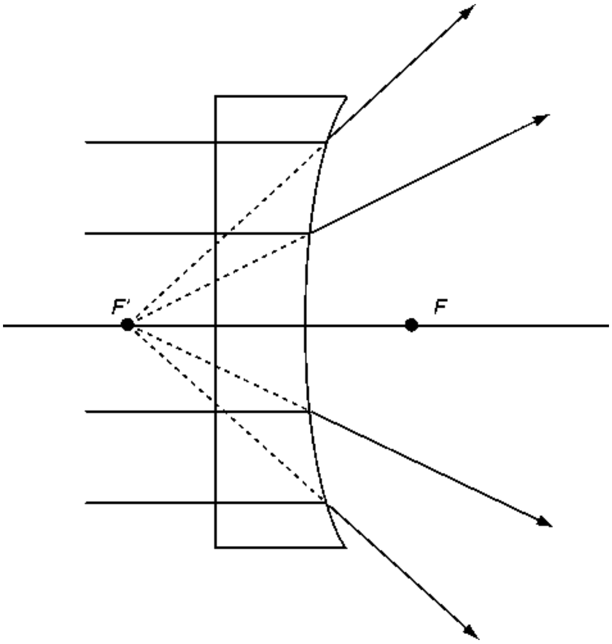

Figure 7 Ray tracing for a planar-concave lens.

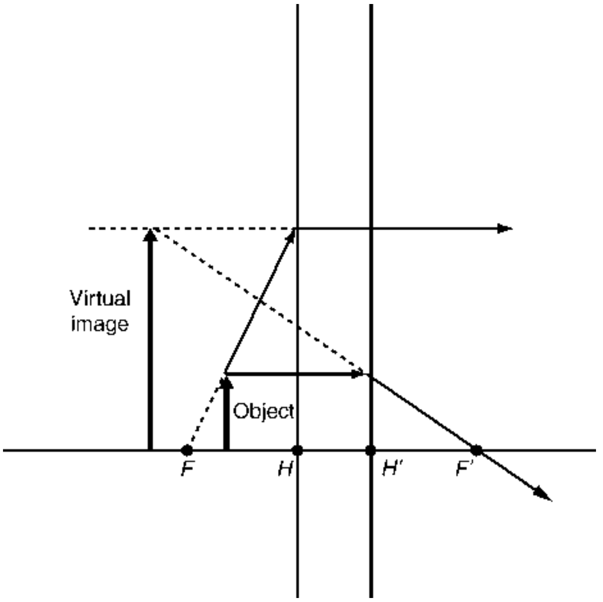

Figure 8 Formation of a virtual image by a convex lens.

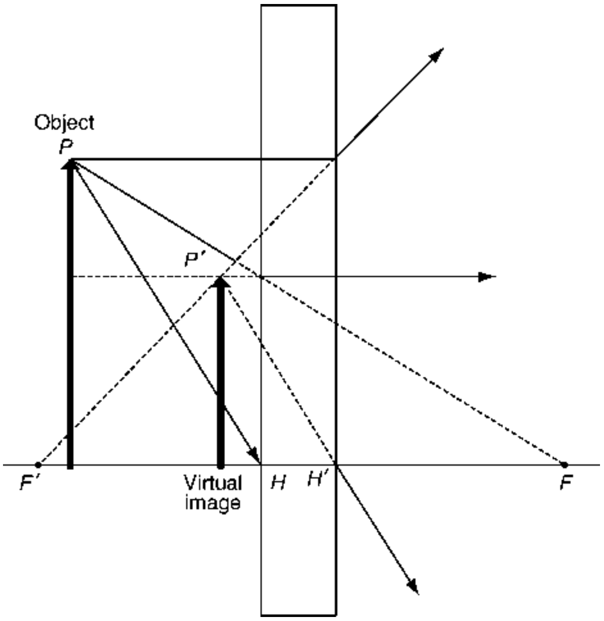

Figure 9 Formation of a virtual image by a concave lens.

rays are extended backwards, these extensions will meet as indicated. This behavior can be verified quantitatively by using the formulas developed above. It will be found that F and F' have exchanged places, but H and H' retain their original positions inside the lens. If an object is placed between the focal point F and the first vertex of a convex lens (Figure 8), then our usual procedure produces two rays which do not intersect in image space. The ray from the object which is parallel to the axis is refracted downward so that it does not intersect the other ray in image space. However, the extensions to the left of these two rays meet to form an image which is erect, magnified, and virtual. This is the normal action of a magnifying lens; the image can be seen but cannot be projected onto a screen, unlike the real, inverted image of Figure 3. This ray tracing diagram confirms what would be seen when a double convex lens is used to magnify an object lying close to the lens. For the double concave lens of Figure 9, a parallel ray leaves an object point P and is refracted upward at the unit plane H' in such a way that its extension, rather than the ray itself, passes through F'. The ray headed for F is refracted at H before it can reach the object-space focal plane and becomes parallel to the axis. The third ray, going from P to H, emerges at H' parallel to its original direction and its extension to the left passes through the image point P' already determined by the intersection of the other two rays. The resulting virtual image is upright and reduced. All three rays in this diagram behave exactly like the corresponding rays in Figure 3; the only change is the use of the extensions to locate the image. The eye receives the diverging rays from P and believes that they are coming from P'.

Nodal Points and Planes

Let a ray leave a point on the z-axis at an angle α_1, and have an angle α'_2 when it reaches image space.

Since $x = 0$ at the starting point, eqn [26] shows that

$$n'_2\alpha'_2 = n_1\alpha_1/m \qquad [40]$$

The ratio of the final angle to the initial angle in this equation is the angular magnification μ, which is then

$$\mu = n_1/mn'_2 \qquad [41]$$

The locations of object and image for unit angular magnification are called the nodal points, labeled as N and N'. The above definition shows that the linear and angular magnification are reciprocals of one another when object and image are in air or the same medium. The procedure that located the unit points will show that the nodal points have positions given by

$$\frac{l_N}{n_1} = \frac{(n'_2/n_1) - b}{a} \qquad [42]$$

and

$$\frac{l'_N}{n'_2} = \frac{c - (n_1/n'_2)}{a} \qquad [43]$$

When the two indices are identical, these relations are identical to those for the points H and H'. That is why the rays from P to H and H' to P' in Figure 3 are parallel.

Compound Lenses

A great advantage of the paraxial matrix method is the ease of dealing with systems having a large number of lenses. To start simply, consider the cemented doublet of Figure 10: a pair of lenses for which surface 2 of the first element and surface 1 of the second element match perfectly. The parameters of this doublet are given in the manner shown below.

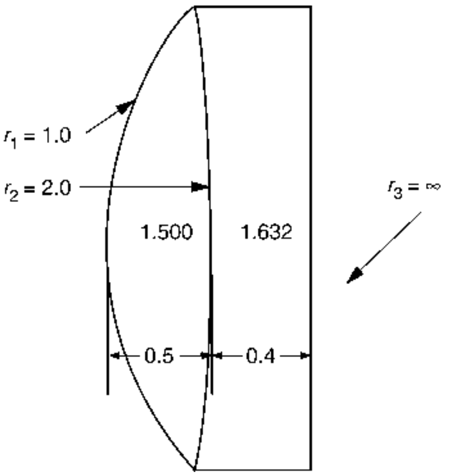

Figure 10 Specification for a cemented doublet.

It is understood that the first entry under r is r_1, the second entry is r_2, and so on. Note that this doublet has only three surfaces; if the two parts were separated by an air space, producing an air-spaced doublet, then there would be four surfaces to specify. The values of the indices n' and the spacings t' are placed between the values of r. The constants of the doublet are then

$$\begin{array}{ccc} r & n' & t' \\ 1.0 & & \\ & 1.500 & 0.5 \\ -2.0 & & \\ & 1.632 & 0.4 \\ \infty & & \end{array} \qquad [44]$$

Note that surface 3, which is flat, has a radius of infinity. Numbering the vertices in the usual manner, the system matrix is

$$S_{31} = R_3T_{32}R_2T_{21}R_1 \qquad [45]$$

where

$$k_2 = (n'_2 - n_2)/r_2 = (1.632 - 1.500)/-2.0 \qquad [46]$$

Approximations for Thin Lenses

An advantage of paraxial matrix optics is the ease of obtaining useful relations. For example, combining the several expressions for the Gaussian constant a, we obtain:

$$a = -\frac{n_1}{f} = \frac{n'_2}{f'} = k_1 + k_2 - \frac{k_1k_2t'_1}{n'_1} \qquad [47]$$

and using the definitions of k_1 and k_2 with the lens in air leads to

$$\frac{1}{f'} = (n'_1 - 1)\left[\frac{1}{r_1} - \frac{1}{r_2} + \frac{(n'_1 - 1)t'_1}{n'_1r_1r_2}\right] \qquad [48]$$

which is the lensmakers' equation. It tells how to find the focal length of a lens from a knowledge of its material and geometry. This derivation is much more direct than what you will usually find. This equation becomes simpler when we are dealing with thin lenses – those for which the third term in brackets can be neglected by assuming that the lens thickness is approximately zero. Then

$$\frac{1}{f'} = (n'_1 - 1)\left[\frac{1}{r_1} - \frac{1}{r_2}\right] \qquad [49]$$

which is the form often seen in texts. We then realize that the original form is valid for lenses of any thickness, but the thin lens form can be used if the spacing is much less than the two radii. It is also seen that when the thickness is approximately zero, then

$$b = c = 1 \qquad [50]$$

and the unit planes have locations

$$l_H = l'_H = 0 \qquad [51]$$

Thus, they are now coincident at the center of the lens. A ray from any point on the object would pass undeviated through the geometrical center of the lens, as is often shown in the elementary books. These new values of the Gaussian constants give a system matrix of the form:

$$S_{21} = \begin{pmatrix} 1 & -1/f' \\ 0 & 1 \end{pmatrix} \qquad [52]$$

This matrix contains a single constant, the focal length of the lens, so that the preliminary design of an optical system is merely a matter of specifying the focal length of the lenses involved and using the resulting matrices to determine the object-image relationship. We also note that this matrix looks like a refraction matrix; the non-zero element is in the upper right-hand corner. This implies that the refraction–translation–refraction procedure that actually occurs can be replaced, for a thin lens, by a single refraction occurring at the coinciding unit planes. Ophthalmologists take advantage of the thin lens approximation by specifying focal lengths in units called diopters. The reciprocal of the focal length in meters determines its refraction power in diopters. For example, a lens with $f' = 10$ cm will be a 10 diopter lens. If two lenses are placed in contact, the combined power is the sum of the individual powers, for multiplying matrices of the above form gives the upper right-hand element as $(-1/f'_1 - 1/f'_2)$.

The Paraxial System for Design or Analysis

The material given in this article can serve as the basis for an organized way of taking the specifications of an optical system and using them to gain a full understanding of this system. We shall demonstrate it with a practical example which has some interesting features. Chemical engineers have long known that the index of refraction of a liquid can be determined by observing the empty bore of a thick glass tube with a telescope having a calibrated eyepiece and then measuring the magnification of the bore when the liquid flows through it. Consider a liquid with an index of 1.333 and a tube with bore radius of 3 cm, outer radius of 4 cm, and glass of index equal to 1.500. The first step is to specify the parameters of the optical system. These are the radii $r_1 = -3$, $r_2 = -4$ and the thickness $t'_1 = 1$, regarding the tube as a concentric lens with an object to its left; the indices $n_1 = 4/3$ (fractions are convenient), $n'_1 = 3/2$, $n'_2 = 1$. Calculating the elements of the two refraction matrices and the translation matrix, the system matrix is

$$S_{21} = \begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 2/3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1/18 \\ 0 & 1 \end{pmatrix} \qquad [53]$$

Multiplying, the Gaussian constants are

$$a = 2/27, \quad b = 11/12, \quad c = 28/27, \quad d = -2/3 \qquad [54]$$

An important check is to verify that $bc - ad = 1$, as is the case here. The expressions given previously for the positions of the cardinal points then lead to the values

$$\begin{aligned} &l_H = 3/2, \quad l'_H = 1/2, \quad l_F = -16\ 1/2, \\ &l'_F = 14, \quad l_N = -3, \quad l'_N = -4 \end{aligned} \qquad [55]$$

These points are shown in Figure 11. Although the same scale has been used for both horizontal and vertical distances, this is usually not necessary; the vertical scale, which merely serves to verify the magnification, can be arbitrary. The entire liquid column is the object in question, but it is simpler to use a vertical radius as the object. Then the image, generated as described in all the previous examples,

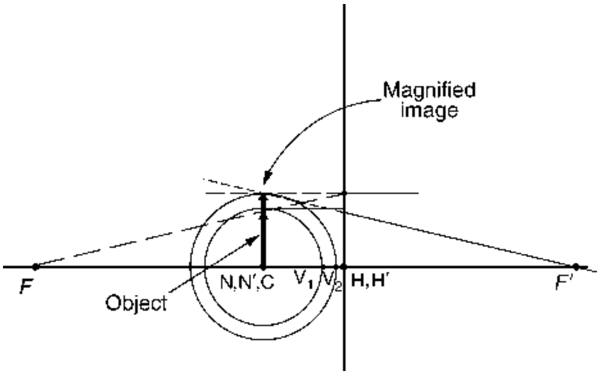

Figure 11 Ray tracing diagram for a thick-walled glass tube.

coincides with the object and it is enlarged, erect, and virtual. We confirm the image location shown by using the generalized Gaussian lens equation, giving

$$\begin{aligned} t' &= [(-2/3) + (28/27)(-3)/(4/3)]/[(11/12) \\ &\quad + (2/27)(-3)/(4/3)] \\ &= -4 \end{aligned} \qquad [56]$$

and this agrees with the sketch. The denominator of this expression was shown to be the reciprocal of the magnification $1/m$ with a value of 3/4, and m itself can be calculated from t' as $c - at' = 4/3$; this is another important check on the accuracy of the calculation; the sketch obeys this conclusion, and the purpose of this analysis is confirmed.

Reflectors

To show how reflecting surfaces are handled with matrices, consider a plane mirror located at the origin and normal to the z-axis (Figure 12). The ray which strikes it at an angle α_1 leaves at an equal angle, resulting in specular reflection. Since $k = 0$ for a flat surface, the refraction matrix reduces to the unit matrix and by eqn [9]:

$$n'_1\alpha'_1 = n_1\alpha_1 \qquad [57]$$

This contradicts Figure 12, since α_1 and α'_1 are opposite in sign.

If, however, we specify that

$$n'_1 = -n_1 \qquad [58]$$

then the difficulty is removed. Hence, reflections involve a reversal of the sign on the index of refraction, and two successive reflections restore the original sign. For a mirror, the refraction power is

$$k_1 = \frac{n'_1 - n_1}{r_1} = \frac{-1 - (1)}{r_1} = -\frac{2}{r_1} \qquad [59]$$

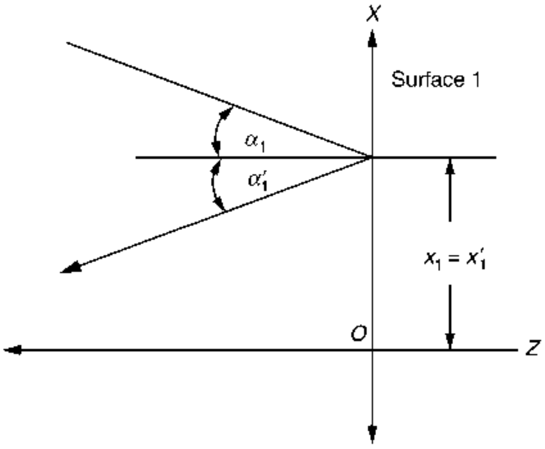

Figure 12 Specular reflection.

The system matrix for a single refracting surface thus becomes

$$S_{11} = R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2/r_1 \\ 0 & 1 \end{pmatrix} \qquad [60]$$

with

$$a = -2/r_1, \qquad b = c = 1, \qquad d = 0 \qquad [61]$$

The connection between object and image can be expressed as

$$\begin{pmatrix} n'_1\alpha'_1 \\ x'_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'/n'_1 & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -t/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [62]$$

and using the same procedure that was applied to eqns [28]–[31], the locations of the six cardinal points are

$$\begin{gathered} l_F = -n_1 b/a, \quad l'_F = n'_1 c/a \\ l_H = n_1(1-b)/a, \quad l'_H = n'_1(c-1)/a \qquad [63] \\ \frac{l_N}{n_1} = \frac{(n'_1/n_1) - b}{a}, \quad \frac{l'_N}{n'_1} = \frac{c - (n_1/n'_1)}{a} \end{gathered}$$

so that for the spherical mirror:

$$l_F = \frac{(-1)(1)}{-2/r_1} = \frac{r_1}{2} = l'_F \qquad [64]$$

$$l_H = l'_H = 0 \qquad [65]$$

and

$$l_N = \frac{-1-1}{-2/r_1} = r_1 \qquad [66]$$

but

$$\frac{l'_N}{-1} = \frac{1-(-1)}{-2/r_1}, \quad l'_N = r_1 \qquad [67]$$

Since r_1 is negative, these equations show that the unit points lie on the vertex, the foci coincide halfway between the center of curvature and the

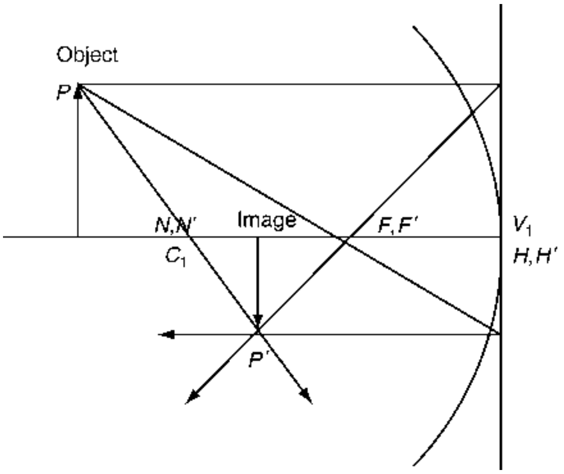

Figure 13 Ray tracing for a concave mirror.

vertex, and the nodal points are at this center. The perfect focusing is a consequence of the paraxial approximation. The true situation is presented in **Geometrical Optics:** Aberrations. Ray tracing for this mirror uses the procedure developed for lenses, which applies to optical systems of any complexity. Figure 13 shows an object to the left of the center of curvature. The ray from P parallel to the axis goes to H' and then through F', while the ray through F goes to H and then becomes parallel to the axis. Their intersection at P' produces a real, inverted image. The third set of rays represents something different: it requires a knowledge of the nodal point locations. The ray from P to N should be parallel to the ray from N' to P', by definition; in this example, they meet this requirement by being colinear.

As an example of the power of matrix optics to simplify the design or analysis of an optical system involving lenses and mirrors, consider an object that is 15 units to the left of a converging lens, with focal length $f' = 10$. A concave mirror, of radius $r_1 = 16$, is 20 units to the right of the lens. The refraction power of the mirror is $k_1 = (-1-1)/-16 = 1/8$ and using the thin lens form of the system matrix, the combined matrix is

$$\begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 20 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1/10 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -3/2 & 1/40 \\ 20 & -1 \end{pmatrix} \qquad [68]$$

This gives the constants a, b, c, and d, from which we find that the image is $-4\ 4/9$ units to the left of the mirror, its magnification is $-8/9$, and it is inverted. Compare the simplicity of this procedure with the usual way of doing this calculation, which involves first finding the image in the lens and then using that image as a virtual object.

Conclusion

It has been shown that the use of paraxial matrices provides a simple approach to an understanding of the kind of optical systems that provide perfect images. Further details and derivations of expressions given here will be found in the Further Reading below.

See also

Geometrical Optics: Aberrations.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Nussbaum A (1998) *Optical System Design*. Upper Saddle River, NJ: Prentice-Hall.

Aberrations

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The article **Geometrical Optics:** Lenses and Mirrors, describes image formation by components which are optically perfect. The images may be magnified or reduced and they may be erect or inverted, but they are otherwise faithful replicas of the object. This behavior is a consequence of ray tracing based on an approximate or paraxial form of Snell's law which we write as

$$n\theta = n'\theta' \qquad [1]$$

where n and n' are the indices of refraction on either side of an interface and θ and θ' are the angles of incidence and refraction. If, however, the following exact form of Snell's law:

$$n \sin\theta = n' \sin\theta' \qquad [2]$$

governs the ray behavior, as is the case when the rays have large inclinations with respect to the symmetry axis, then it is necessary to devise a more involved ray

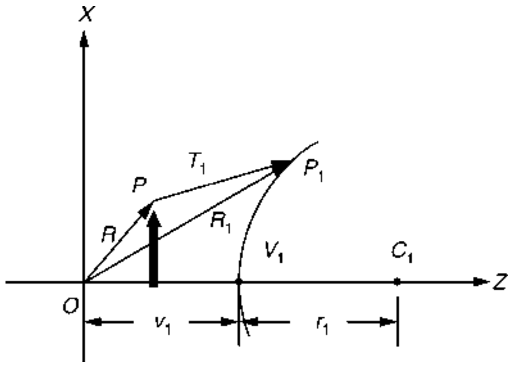

Figure 1 Ray tracing in the meridional plane.

tracing procedure. Figure 1, which shows the first surface of a lens, will be used to explain how this procedure works. A ray leaves the object point P and strikes the lens surface at the point P_1. Rays which lie completely in the plane ZOX are said to be meridional; this is the plane which passes through the symmetry axis, just as meridians on the Earth's surface are determined by planes through the geographic axis. By regarding the various distances in the figure as vectors, it is possible to start with the initial coordinates of the ray and with the direction-cosines which specify its slope and find the coordinates of the point P' where the ray strikes the lens surface. A derivation of the equation governing this process will be found in the text of A Nussbaum (see Further Reading). This derivation involves only elementary algebra and vector analysis, and will be found simpler than the usual textbooks treatment of ray tracing procedures. The equation is

$$T_1 = \frac{F}{-E + \sqrt{E^2 - c_1 F}} \qquad [3]$$

where

$$E = Bc_1/2 = c_1[(z - v_1)N + xL] - N \qquad [4]$$

and

$$F = Cc_1 = c_1\left[(z - v_1)^2 + x^2\right] - 2(z - v_1) \qquad [5]$$

Knowing the coordinates z and x of the starting point P of the ray, as well as its starting cosines N and L, plus the curvature c_1 or radius $r_1 = 1/c_1$ of the lens surface and the location v_1 of the surface's vertex, we can readily calculate the length T_1 of the ray. Then its components on the two axes give the coordinates of P'. Next, we need to find the slope of the ray after it is refracted at the lens surface, and this can be calculated from Snell's law combined with the behavior of the incident and refracted rays regarded as vectors. Again we invoke the reference cited above for a derivation of the equations giving the value of the direction-cosines after refraction; these expressions are

$$n_1' N_1' = K_1/c_1 + n_1 N_1 - K_1 z_1 \qquad [6]$$

and

$$n_1 L_1' = n_1 L_1 - K_1 x_1 \qquad [7]$$

where K_1 is known as the refracting power and has the definition

$$K_1 = c_1(n_1' \cos\theta' - n_1 \cos\theta) \qquad [8]$$

The ray can now be traced to the next surface knowing the lens thickness, and then to the image plane. These two sets of equations, when incorporated into a computer program of about two dozen lines, will trace meridional rays to any desired accuracy through an optical system containing an arbitrary number of lenses and spherical reflectors. As an example, consider a symmetric double convex lens with radii of 50 units at surface 1, −50 units at surface 2, a thickness of 15 units, and an index of 1.50. Rays parallel to the axis – one at 0.2 units above the axis, a second one at 2 units above the axis, and a third one at 20 units above the axis – are started well to the left and traced to the paraxial focal plane, which is at 47.368 units from the second surface. The ray that is 0.2 units above the axis will cross the focal plane at the axis; this is a paraxial ray, passing through the paraxial focal point F'. The ray which is 2 units above the axis crosses the focal plane slightly below the focal point, and is therefore not quite paraxial, while the ray which starts at 20 units above the axis falls well short of the focal point, intersecting the axis before it reaches the focal plane. This behavior indicates that the lens possesses a defect known as spherical aberration and these calculations imply that spherical aberration should be defined as the failure of nonparaxial rays to conform to the behavior of paraxial rays. This definition does not appear in optics texts, and the few definitions that are given appear to be incorrect. As clearly indicated in **Geometrical Optics:** Lenses and Mirrors, the focal point of a lens is strictly a paraxial concept and it is a constant, obtainable from the parameters of the lens or lens system. It is meaningless to speak of a focal point as defined by a pair of nonparaxial rays; this infinite set of ray-pairs determines a corresponding number of intersection points, all of which fall short of the true focal point F'. This situation is nicely illustrated by the next figure, which shows a large number of parallel rays, the central ones being paraxial and the others behaving differently. The rays which are very close to the axis will meet as expected at the focal point F'. The rays a little farther out – the intermediate rays – will cross the axis just a little to the left of F', and those near the edge of the lens – the *marginal* rays – will fall very short, as indicated. Rotating this diagram about the axis, we realize that spherical aberration produces a very blurry circular image at the paraxial plane. The three-dimensional envelope of the group of rays produced by this rotation is known as the caustic surface and the narrowest cross-section of this surface, slightly to the left of the paraxial focal plane, is called the circle of least confusion. If this location is chosen as the image plane, then the amount of spherical aberration can be reduced somewhat. Other methods of improving the image will be considered below. Figure 2 was obtained by adding a graphics step to the computer program described above. Figure 3 shows how to define spherical aberration quantitatively. Two kinds of aberration are specified in this figure. The place where the ray crosses the axis, lying to the left of F', is at a distance, measured from the paraxial focus, called the longitudinal spherical aberration (LSA), while the distance below the axis at the focal plane is the transverse spherical aberration (TSA). To reduce this aberration, note in Figure 2 that the amount of refraction at the two surfaces of the lens for marginal rays is not the same. Equalizing the refraction at the two surfaces will improve the situation. Suppose we alter the lens while keeping the index, thickness, and focal length constant; only the radii will change. This can be done if the new radii satisfy the paraxial equation giving the focal length in terms of the lens constants. The process of varying the shape of lens, while holding the other parameters constant, is called *bending the lens*. This can be easily accomplished by a computer program, obtaining what are known as optimal shape lenses, which are commercially available. To study their effect on spherical aberration, define the shape factor σ of a lens as

$$\sigma = \frac{r_2 + r_1}{r_2 - r_1} \qquad [9]$$

from which $\sigma = 0$ for a symmetric lens ($r_1 = -r_2$). Figure 4a shows the monitor screen for a symmetric lens, and Figure 4b illustrates the result of lens bending. This lens has had the curvature of the first surface raised and the second surface has been flattened to keep the focal length constant. Let us now consider what lens designers can do about spherical aberration. Brouwer in his book (see Further Reading) gives the specifications for a cemented doublet consisting of a double convex lens followed by a lens with two concave surfaces. The longitudinal spherical aberration of the front lens alone varies as shown in Figure 5. The way that the designer arranged for the marginal rays to meet at the paraxial focus, rather than falling short, was to recognize that a diverging lens following the front element would refract the rays away from – rather than closer to – the axis. When this second component is added, we would expect the behavior shown in Figure 6. Note that the value of LSA for the doublet has been enormously reduced, as indicated by the horizontal scale change. Spherical aberration can be reduced in more elaborate systems as part of the design process. Telescopes, in particular, make use of corrector plates which are placed in front of the large mirrors.

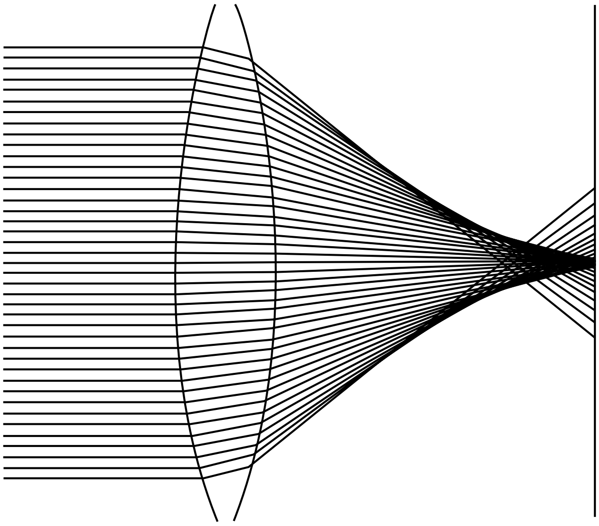

Figure 2 Spherical aberration for parallel meridional rays.

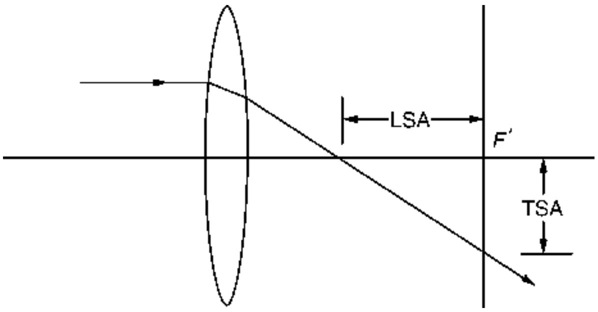

Figure 3 Definitions of transverse and longitudinal spherical aberration.

Coma

We have seen how to trace nonparaxial, meridional rays and found that spherical aberration appears because the intermediate and marginal rays do not behave like the central rays. We now wish to deal

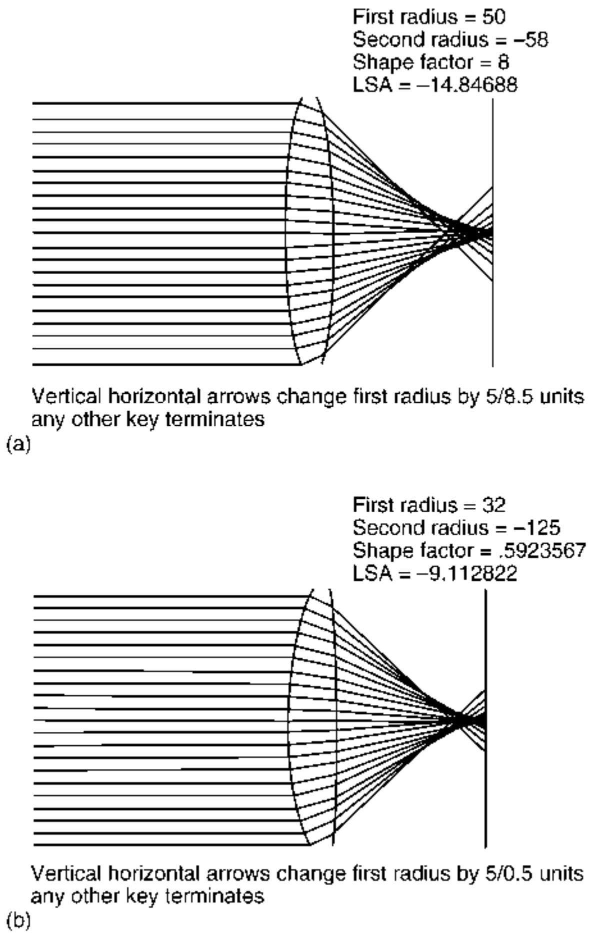
First radius = 50
Second radius = −58
Shape factor = 8
LSA = −14.84688
Vertical horizontal arrows change first radius by 5/8.5 units
any other key terminates
(a)
First radius = 32
Second radius = −125
Shape factor = .5923567
LSA = −9.112822
Vertical horizontal arrows change first radius by 5/0.5 units
any other key terminates
(b)

Figure 4 Lens bending as performed numerically.

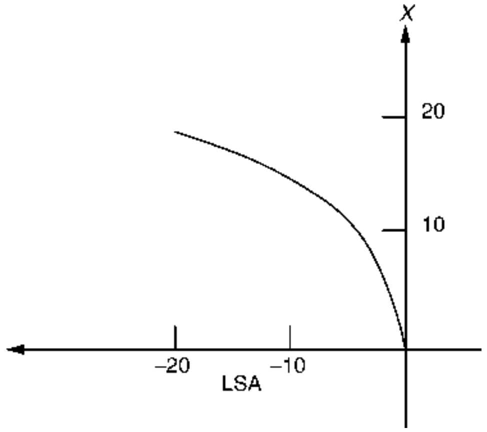

Figure 5 Spherical aberration for a single lens.

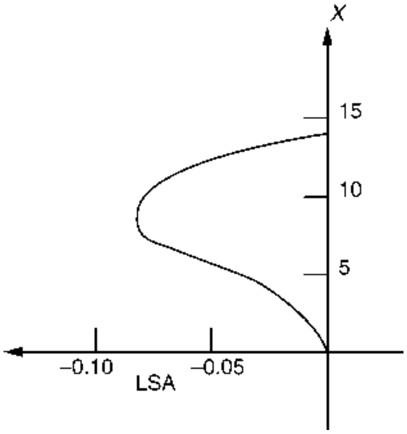

Figure 6 Reduction of spherical aberration by a doublet.

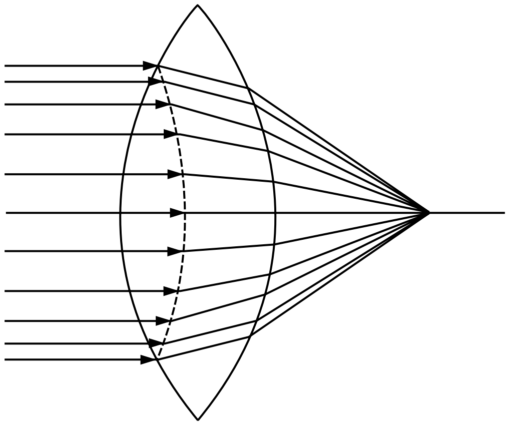

Figure 7 Configuration used to illustrate coma.

with rays that do not lie in a meridional plane; these are nonmeridional or skew rays. Calculating the behavior of such a ray would appear to be a complicated problem, but it turns out to be a simple extension of what has already been accomplished. Let a skew ray start at an arbitrary point in space with coordinates x, y, and z and let it have direction cosines L, M, and N with respect to OX, OY, and OZ. The points P and P_1 in Figure 1 are now out of the plane ZOX and when the vectors in this figure are expressed as the sum of three, rather than two, components, the equations given above for meridional rays acquire terms in y and M which have the same form as those for x and L. Now trace a set of skew rays chosen to lie on a cylinder (Figure 7) which is centered about the z-axis of a lens, so that all the rays are meridional. These rays are fairly far from the axis, so that they are also nonparaxial, but they all meet at a common image point, forming a cone whose apex is this image point. Then give this cylinder a downward displacement while holding fixed the intersection point of each ray with the dotted circle on the front of the lens. The tilting of the cylinder changes all its rays – except the top and the bottom rays – from meridional to skew. In addition, the cone on the image side will then tilt upwards; it should not change in any other way if the skew rays continue to meet at a well-defined apex. But this is not what happens in a simple lens, as will now be explained.

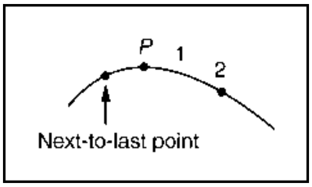

Figure 8 Starting a coma plot.

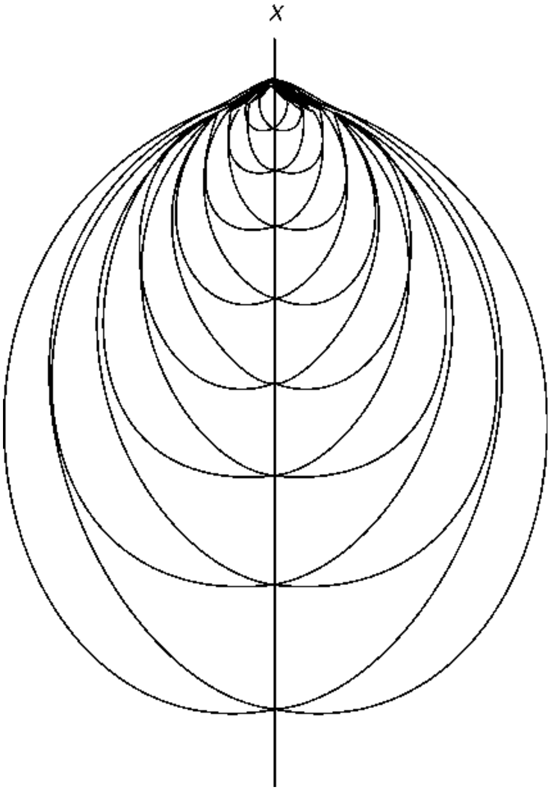

Figure 9 Coma as calculated numerically.

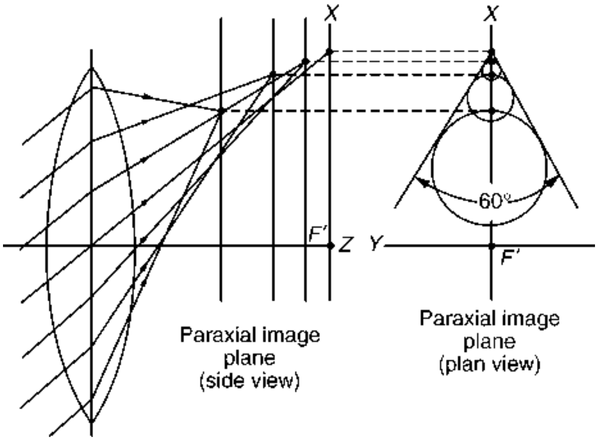
X
X
F′
Z
Y
60°
F′
Paraxial image plane (side view)
Paraxial image plane (plan view)

Figure 10 Customary way of illustrating coma.

Let the top and bottom meridional rays meet at a point *P* in image space (Figure 8) and use this point to determine the location of an image plane. The skew ray just below the uppermost ray in Figure 7 will pass through this plane fairly close to the intersection point. Let us guess (and confirm later) that it will pass through a point slightly to the right of *P* and a little below it, as we observe this plane from the lens position. This is point 1 in Figure 8. The next skew ray on the front of the lens will strike the image plane still farther below and to the right, producing point 2 in the figure. As we continue to trace these rays down the front half of the circle, we realize that a ray almost at the bottom of Figure 7 will have to be very close to the lower meridional ray and that its image point is just below but to the left of *P*; this is the point labeled 'next-to-last' in Figure 8. All of these points join to form a closed curve. The back halves of the dotted circle and of the ray cylinder are not shown in the figure, but since we are again starting at *P* in Figure 8, the pattern will repeat itself. That is, we get a second closed curve which – by symmetry – should be identical to the first curve. This is a most unusual effect; double-valued phenomena are quite rare. To confirm this prediction, we use the skew ray equations, and trace the pattern formed by a set of concentric cylinders of different sizes. The appearance of the coma pattern generated by the computer is shown in Figure 9. Each half-circle on the front side of the lens produces a closed figure (an approximation to a circle) as an image. If coma were the only aberration (when it is called ideal coma), the plots would be the perfect, coinciding circles of Figure 10. The comet-like shape explains where the name comes from. We can picture the generation of these circles in the way shown in this figure. The ray through the geometrical center of the lens produces a point in the paraxial focal plane. The top and bottom rays (the meridional rays) from the circles of increasing size produce points which define image planes that are successively closer to the lens; this is a consequence of spherical aberration. Moving the top ray parallel to itself and along the half-circle generates the image plane pattern. The lines tangential to the coma circles make an angle of 60° with each other, as can be seen by measurement on the computer output. Most optics books look at coma in a way different from that given here. Coma has been defined by WJ Smith (see Further Reading) as the variation of magnification with aperture, analogous to the definition of spherical aberration as the variation of focal length with aperture. And as was previously mentioned, focal length and magnification are purely paraxial concepts, not applicable to the description of aberrations. The definition that comes from the arguments given above is that coma is the failure of skew rays to match the behavior of meridional rays, just as spherical aberration is the failure of meridional rays to match the behavior of paraxial rays. Most texts show coma in its ideal form.

Astigmatism

We have shown that spherical aberration is the failure of meridional rays to obey the paraxial approximation and coma is the failure of skew rays to match the behavior of meridional rays. To see the connection between these two aberrations, consider the object point *P* of Figure 11. This very complicated diagram can actually be very helpful in understanding the third of the five geometrical aberrations. Let *P* be the source of a meridional fan: this is the group of rays with the top ray labeled PA and the bottom ray is PB. If the lens is completely corrected for spherical aberration, this fan will have a sharp image point P'_T lying directly below the z-axis. Now let *P* also be the source of a fan bounded by the rays PC and PD. This fan is at right-angles to the other one; we can think of these rays as having the maximum skewness possible. If the lens has no coma, this fan also will produce a sharp image point P'_s which, for a converging lens, will be farther from the lens than the tangential focal point. In other words, even though the lens has been fully corrected for both aberrations, the two corrections will not necessarily produce a common image. In the figure, the meridional fan is called a tangential fan and the skew fan is called a sagittal fan; these are the terms commonly used by lens designers. The failure of the sagittal and tangential rays to produce a single image in a lens corrected for both spherical aberration and coma is known as astigmatism. Astigmatism, coma, and spherical aberration are the point aberrations that an optical system can have. We have already noted that spherical aberration is the only one that can be associated with a point on the z-axis; the others require an off-axis object. Astigmatism is very common in the human eye; to see how it shows up, look at the sagittal image point P'_s of Figure 11. This point is a distance z_s, from the x,y-plane. A plane through this point will have a line image on it, the sagittal line, due to the tangential fan, which has already come to a focus and is now diverging. The converse effect, producing a tangential line, occurs at the other image plane. If we locate an image plane halfway between the two, then both fans contribute to the image and in the ideal case it will be a circle. As the image plane is moved forwards or backwards, these images become ellipses and eventually reduce to a sagittal or a tangential line, as shown in Figure 12. Because there are two different

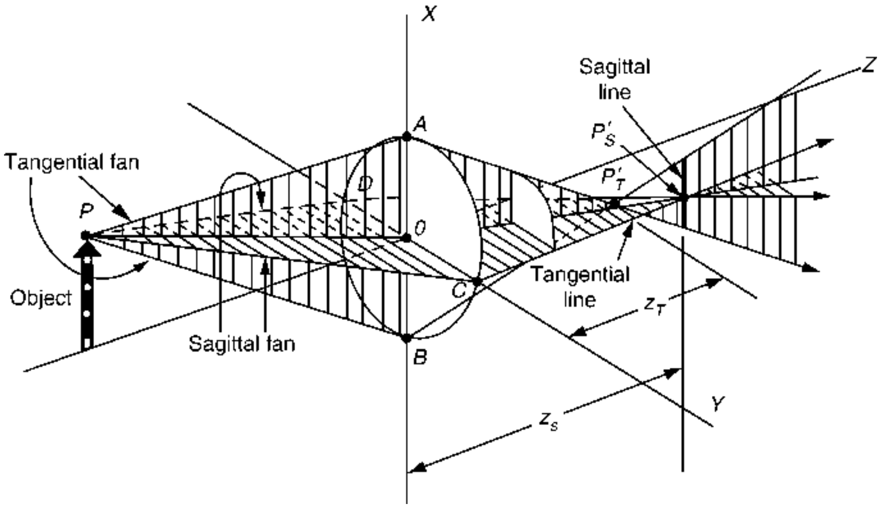

Figure 11 Tangential and sagittal fans displaying astigmatism.

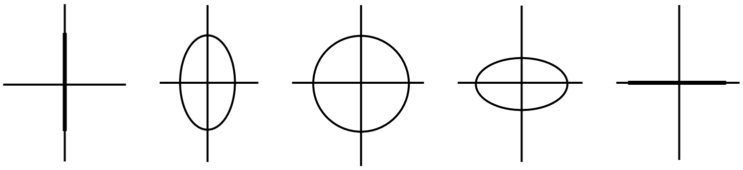

Figure 12 Astigmatic image patterns.

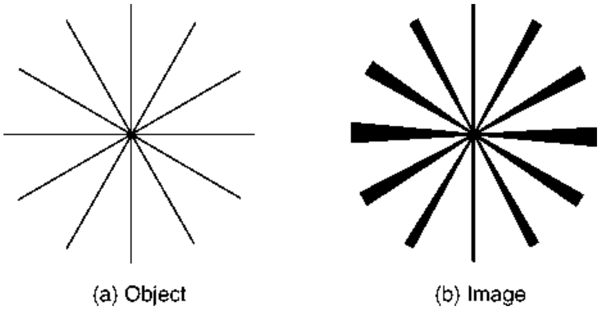

Figure 13 Image degraded by astigmatism.

image planes, an object with spokes (Figure 13) will have an image for which the vertical lines are sharp and the others get gradually poorer as they become more horizontal, or vice versa, depending on which image plane is chosen. Eye examinations detect astigmatism if the spokes appear to go from black to gray. This example of the effect of astigmatism was often based on a radial pattern of arrows, which explains the origin of the word sagittal, derived from the Latin ‘sagitta’ for arrow.

Curvature of Field and Distortion

Having covered the three point aberrations, there are two lens defects associated with extended objects. If we move the object point P in Figure 11 closer to or farther from the z-axis, we would expect the positions of the tangential and sagittal focal planes to shift, for it is only when the paraxial approximation holds that these image points are independent of x. Hence, we obtain the two curves of Figure 14, which shows what astigmatism does to the image of a two-dimensional object. If the astigmatism could be eliminated, the effect would be to make these curved image planes coincide, but we have no guarantee that the common image will be flat, or paraxial. The resulting defect is called Petzval curvature or curvature of field. For a single lens, the Petzval surface can be flattened by a stop in the proper place, and this is usually done in inexpensive cameras. Petzval curvature is associated with the z-axis. If we take the object in Figure 11 and move it along the y-axis, then all rays leaving it are skew and this introduces distortion, the aberration associated with the coordinates normal to the symmetry axis. Distortion is what causes vertical lines to bulge outward (barrel distortion) or inward (pincushion distortion) at the image plane, as shown in Figure 15. A pinhole camera will have no distortion, since there are no skew rays, and a single thin lens with a small aperture will have very little. Placing a stop near the lens to reduce astigmatism and curvature of field introduces distortion because, as

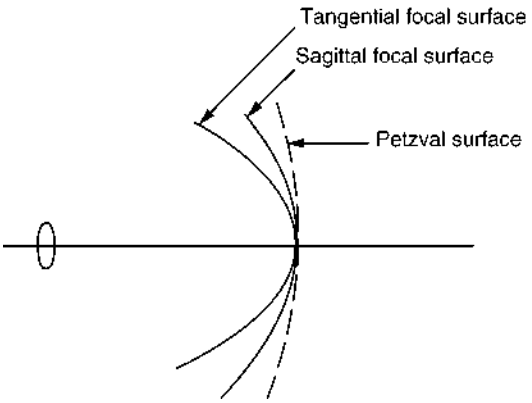

Figure 14 Petzval of field curvature.

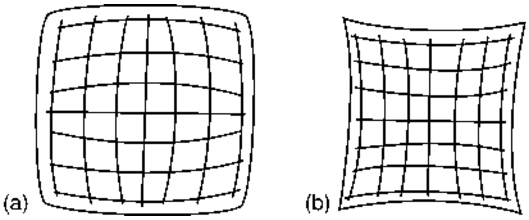

Figure 15 Barrel and pincushion distortion.

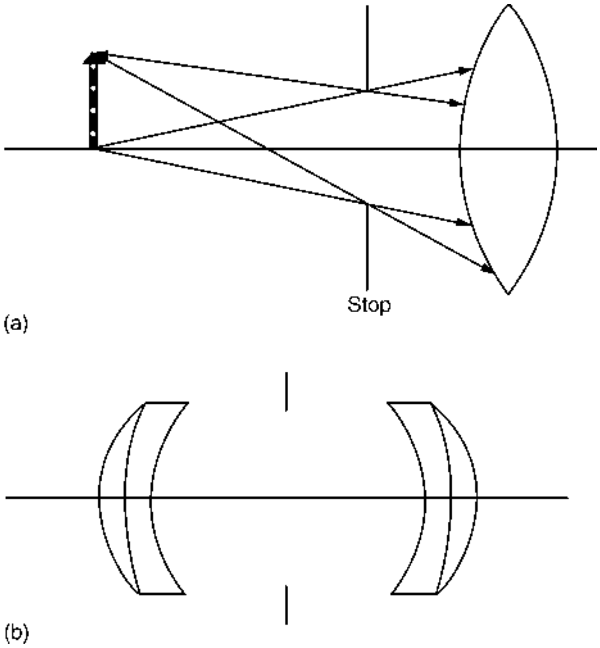

Figure 16 (a) Distortion-creating stop. (b) Symmetric-component lens that reduces distortion.

shown in Figure 16a, the rays for object points far from the axis are limited to off-center sections of the lens. The situation in this figure corresponds to barrel distortion; placing the stop on the other side of the lens produces pincushioning. Distortion can be

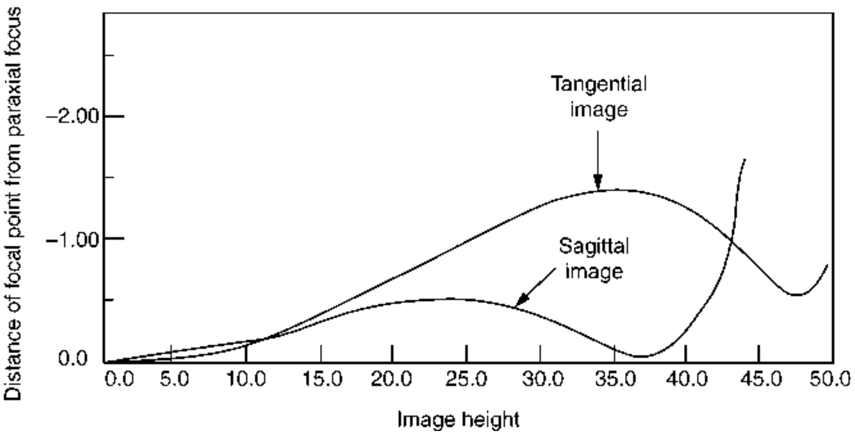

Figure 17 Reduction of astigmatism through design.

therefore reduced by a design which consists of two identical groupings (Figure 16b), with the iris diaphragm placed between them; the distortion of the front half cancels that of the back half. Figure 17 shows the output for a calculation using approximate formulas rather than exact ray tracing. Note that the designer has arranged for astigmatism to vanish at the lens margin. This example indicates the nature of astigmatism in a system for which the coma and the spherical aberration are low but not necessarily zero; it is the failure of the tangential and sagittal image planes to coincide.

Nonspherical Surfaces

So far, we have considered optical systems that use only spherical or plane reflecting and refracting surfaces. However, many astronomical telescopes use parabolic and hyperbolic surfaces. In recent years, nonspherical or aspheric glass lenses have become more common because they can be mass produced by computer-controlled machinery and plastic lenses can be molded with very strict tolerances. Ray tracing procedures given above can easily be extended to the simplest kind of aspheric surfaces: the conic sections. Iterative procedures may be applied to surfaces more complicated than the conic sections. A well-known example is the curved mirror; a spherical reflecting surface will have all the aberrations mentioned here, but a parabolic mirror will bring parallel rays to a perfect focus at the paraxial focal point, thus eliminating the spherical aberration. Headlamp reflectors in automobiles are a well-known example. A plano-convex lens with this defect corrected will be one whose first surface is flat and whose second surface is a hyperbola with its eccentricity equal to the index of refraction of the glass composing the lens. One form of astronomical telescope, which has a large primary concave mirror to collect the light and reflect it back to a smaller secondary convex mirror, is the Ritchey–Chrétien design of 1924. Both mirrors are hyperbolic in shape, completely eliminating spherical aberration. This is the design of the Hubble space telescope. The two-meter diameter primary mirror, formed by starting with a spherical mirror and converted to a hyperbola by computer-controlled polishing, was originally fabricated incorrectly and caused spherical aberration. From a knowledge of the computer calculations, it was possible to design, build, and place in orbit a large correcting optical element which removed all the aberrations from the emerging image.

Conclusion

The three-point aberrations and the two extended object aberrations have been defined and illustrated. Simple ways of reducing them have been mentioned, but the best way of dealing with aberrations is in the process of lens design. In addition, there is a completely separate defect known as chromatic aberration. All the above calculations are based on ray tracing for a single wavelength (or color) of visible light. However, the index of refraction of a transparent material varies with the wavelength of the light passing through it, and a design good for one value of the wavelength will give an out-of-focus image for any other wavelength. It has been known for many years that an optical system composed of two different kinds of glass will provide a complete correction for two colors – usually red and blue – and other colors will be partially corrected. Such lenses are said to be achromatic. For highly demanding applications, such as faithful color reproductions

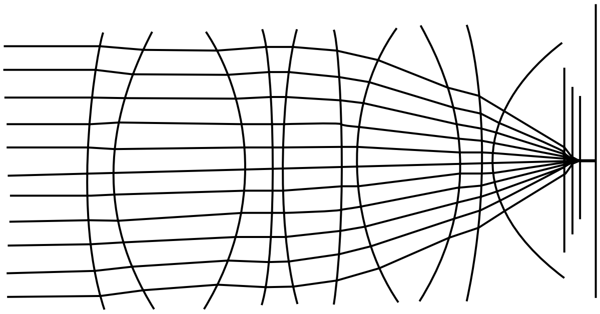

Figure 18 High-power high-resolution microscope objective.

of works of art, it is possible to design elaborate and expensive lenses which correct for yellow light as well. This kind of optical system is said to be apochromatic. With perfect imaging at the two ends and the middle of the visible spectrum, the color error elsewhere is virtually nonexistent. As an example of a design which is free of all aberrations – geometric and chromatic – Figure 18 shows the ray trace for a high-power aberration-free and apochromatic microscope objective. Note that the trace shows parallel rays brought to a focus inside the cover glass; in use, the light source would be a point on the slide and the light would converge at the eyepiece.

See also

Geometrical Optics: Lenses and Mirrors; Prisms.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Laikin M (1990) *Lens Design*. New York: Marcel Dekker.

Nussbaum A (1998) *Optical System Design*. Upper Saddle River, NJ: Prentice-Hall.

Nussbaum A and Phillips RA (1976) *Contemporary Optics for Scientists and Engineers*. Englewoods Cliffs, NJ: Prentice Hall.

Smith WJ (1990) *Modern Optical Engineering*, 2nd edn. New York: McGraw-Hill Book Co.

Prisms

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Prisms are solid structures made from a transparent material (usually glass). Figure 1 shows the simplest type of prism – one that has many uses. The vertical sides are two identical rectangles at right angles to each other and a larger third rectangle is at 45° to the other two. The top and bottom are 45°–45°–90° triangles. The most common applications of prisms are:

1. To bend light in a specified direction
2. To fold an optical system into a smaller space
3. To provide proper image orientation
4. To combine or split optical beams, using partially reflecting surfaces
5. To disperse light in optical instruments such as spectrographs.

Single Prisms as Reflectors

A prism like that of Figure 1 can act as a very effective reflector; it is more efficient than a mirror since there is no metal coating to absorb light. Figure 2 shows a light ray entering the prism at right angles to one of the smaller sides, reflected at the larger side, and emerging at right angles to the other small side. The path shown is determined by Snell's law. This law has the form:

$$n \sin \theta = n' \sin \theta' \qquad [1]$$

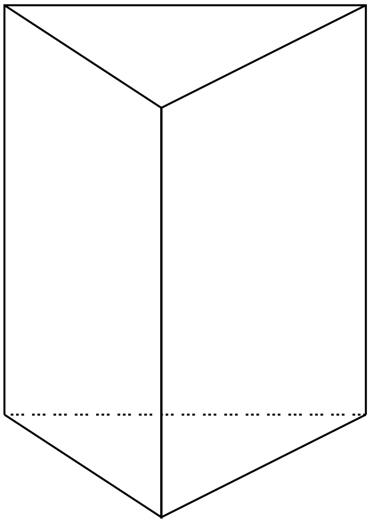

Figure 1 A 45–45–90 degree prism.

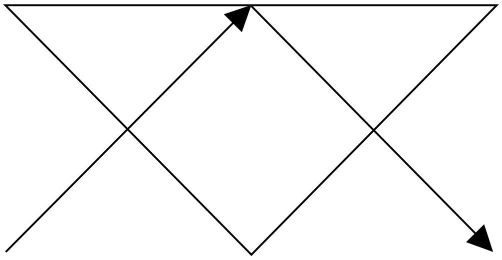

Figure 2 Total internal reflection by a prism.

This indicates that a ray of light making an angle θ to the normal at the interface between media of indices of refraction n and n', will be refracted and the angle to the normal becomes θ'. Notice that the light ray strikes the surface where the reflection takes place at an angle of 45° to the normal; this angle is too large to permit it to emerge from the prism after refraction. To understand why, we recognize that when the ray is able to leave the glass, Snell's law requires that the emerging angle be larger than the incident angle. The largest possible incident angle – known as the critical angle – will correspond to an emerging angle of 90°; that is, this ray will be parallel to the interface. For an index of 1.50, as an example, the critical angle θ_c satisfies Snell's law in the form:

$$1.5 \sin \theta_c = 1.0 \sin 90° \qquad [2]$$

from which $\theta_c = 41.8°$. The ray shown in the figure will therefore remain in the glass. Furthermore, since the indices appearing on both sides of Snell's law are now those of the glass, then the incident and refracted angles must be identical, and the prism reflects the ray in the same way that a mirror would. This phenomenon is known as total internal reflection.

This prism has an important application as an optical component of a fingerprint machine, whose action depends on modifying the total internal reflection which normally occurs. Figure 3 shows the prism and the reflecting behavior of Figure 2, with a person's finger pressed against the long side. When the ray strikes a place on the surface corresponding to a valley in the fingerprint pattern, the index at this region is that of air, and the ray is reflected in the usual manner. But a ray striking a portion of the prism which lies immediately below a raised section of the skin will make an angle with the normal which is now smaller than the critical angle, since the index is much greater than that of air, so that the ray will cross the interface and be absorbed by the finger. The reflected beam, with the fingerprint pattern sharply reproduced, is sent to a detector and the resulting signal goes to a computer, where it is processed and used to drive a printer. The fingerprint cards generated by this kind of equipment are sharper than those done by the traditional smeared-ink method. It is possible to observe the phenomenon just described by looking at your fingers while holding a glass of water – the fingerprints will show up in an enhanced manner. In addition to the better quality, the manufacturer of one version of this equipment has incorporated a feature in their computer program which can detect attempts by the person being fingerprinted to degrade the image.

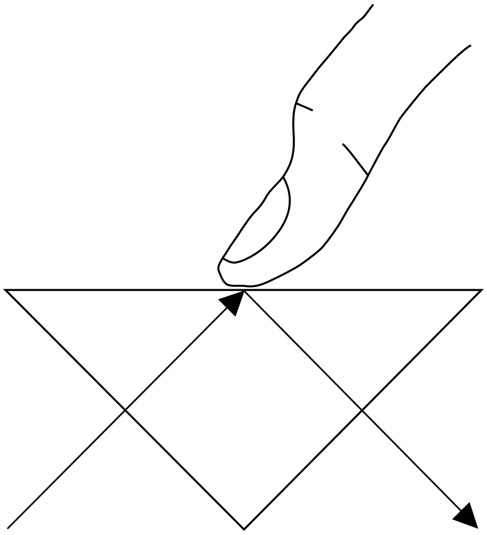

Figure 3 Use of a prism to produce a fingerprint.

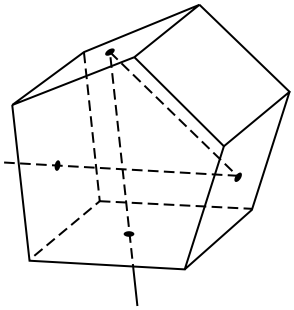

Figure 4 A pentaprism.

A more elaborate single prism – one with five sides – is the pentaprism of Figure 4, used extensively in single lens reflex cameras. This prism does not take advantage of the total internal reflection process, as described above. Below the bottom face is a mirror at a 45° angle. Light from the camera's lens is reflected upward and emerges from the left-hand face where the scene being photographed is observed by the eye.

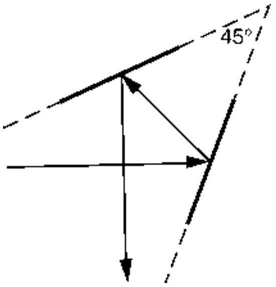

Figure 5 Action of coated faces of a pentaprism.

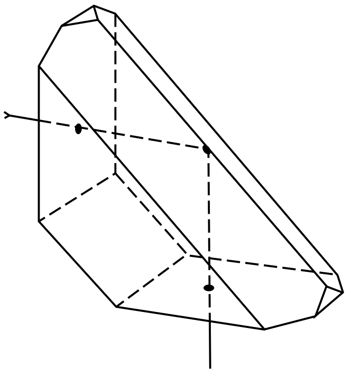

Figure 6 Amici prism.

Since there are two reflections, the reversal caused by a single mirror is eliminated. The camera's mirror is hinged and when the exposure button is pressed, it flies up and the incoming light can reach the film. The two reflecting faces, which lie at an angle of 45° to one another (Figure 5), are coated. To see why this is necessary, note that the two angles in the isosceles triangle formed by the light ray and the face extensions must equal $(180° - 45°)/2$ or 67.5°. This ray must then lie at an angle of $90° - 67.5°$ to the normal, or 22.5°, which is much less than the critical angle needed for reflection.

There are at least a dozen other forms of prism which have been designed and used. For example, the Amici prism (Figure 6) is a truncated right-angle prism with a roof section added to the hypotenuse face. It splits the image down the middle and thereby interchanges the right and left portions. These prisms must be very carefully made to avoid producing a double image. One use is in telescopes to correct the reversal of an image.

Double Prism Reflectors

Another extensive application of prisms is their use in pairs as components of telescopes and binoculars. The most widely used is a pair of Porro prisms: two right angle prisms arranged as shown in Figure 7. The double change in direction has two effects; it eliminates image reversal and it shortens the spacing between objective and eyepiece lenses, making the binoculars lighter and more compact. It is possible to calculate how light rays behave in Porro prisms with a ray tracing procedure. This method involves the use of 4×4 matrices and its development is rather complicated. A simple understanding of the way light passes through this optical component can be based on Figure 8, which shows the two prisms separated for clarity. An arbitrary ray (the heavy line) coming from the object will be reflected at a 45° angle at each of the four faces that it strikes. Using two line segments at right angles as an object, the ray from the top of the vertical arrow and the ray from the end of the horizontal arrow will behave as indicated. The image is then identical to the object but rotated by 180°. Since the optics of the binoculars – the combined effect of the objective and the eyepiece – result in an inverted image, the Porro prism corrects this problem.

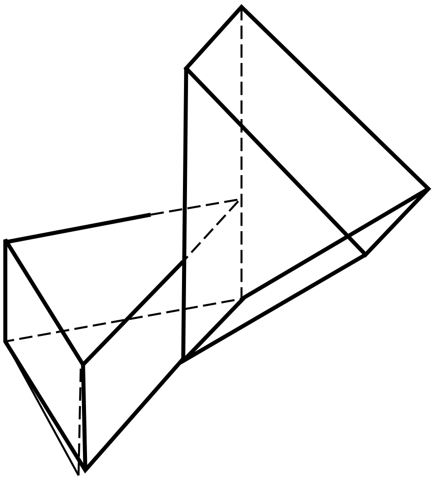

Figure 7 A pair of Porro prisms.

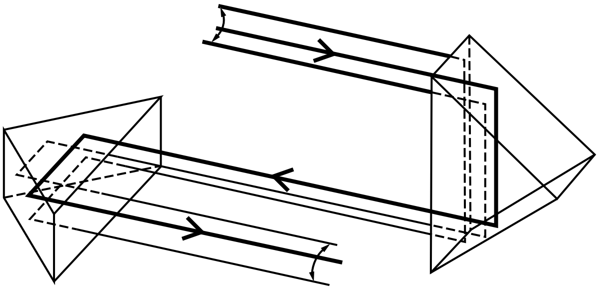

Figure 8 Action of Porro prisms.

Prisms as Instruments

Prisms are the crucial optical element in many laboratory instruments. To show how they can be used to measure the index of refraction of transparent materials – solids or liquids – consider the prism of Figure 9, with apex angle A. Light comes to the prism at an incident angle i, enters the prism at the angle of refraction r, crosses the prism and is refracted a second time when it emerges, the associated angles being designated as r' and i'. When the ray leaves the prism, it is traveling in a direction which represents an angular deviation δ from its original orientation. This deviation is the sum of the individual changes in direction at each face, or

$$\delta = (i - r) + (r' - i') \qquad [3]$$

For the triangle formed by the projected normals, which meet at the angle A:

$$r = r' + A \qquad [4]$$

Hence

$$\delta = i + i' - A \qquad [5]$$

The minimum value of the deviation is an important property of the prism; we find it by differentiation to obtain:

$$\mathrm{d}\delta = \mathrm{d}i + \mathrm{d}i' = 0 \qquad [6]$$

Differentiation of the form of Snell's law valid at each interface, namely:

$$\sin i = n \sin r, \quad \sin i' = n \sin r' \qquad [7]$$

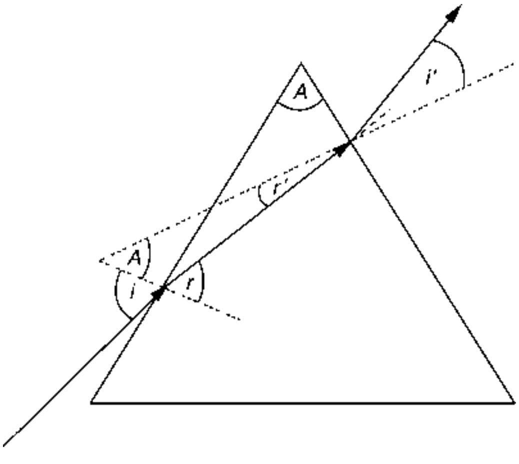

Figure 9 Calculation of prism deviation.

gives

$$\mathrm{d}i = n \cos r \, \mathrm{d}r / \cos i \qquad [8]$$

$$\begin{aligned} \mathrm{d}i' &= n \cos r' \, \mathrm{d}r' / \cos i' \\ &= n \cos r' \, \mathrm{d}r / \cos i' \qquad [9] \end{aligned}$$

where $\mathrm{d}r'$ can be replaced by $\mathrm{d}r$, as obtained from the equation connecting r, r', and A. It now follows that:

$$\begin{aligned} \mathrm{d}\delta = n\Big[&\big\{\cos r/\sqrt{(1 - n^2 \sin^2 r)}\big\} \\ &- \big\{\cos(A - r)/\sqrt{(1 - n^2 \sin^2(A - r)}\big\}\Big]\mathrm{d}r = 0 \end{aligned} \qquad [10]$$

from which

$$\begin{aligned} &\cos^2 r\{1 - n^2 \sin^2(A - r)\} \\ &\quad = \cos^2(A - r)\{1 - n^2 \sin^2 r\} \qquad [11] \end{aligned}$$

or:

$$\cos^2 r = \cos^2(A - r) \qquad [12]$$

and finally:

$$r = A/2, \quad r' = A/2, \quad i' = i \qquad [13]$$

The differentiation thus leads to the conclusion that the ray enters and leaves the prism in a symmetrical manner. The amount of deviation is now:

$$\delta = 2i - A \qquad [14]$$

or:

$$i = (\delta + A)/2 \qquad [15]$$

and:

$$n = \sin i / \sin r = \sin\{(\delta + A)/2\} / \sin(A/2) \qquad [16]$$

Using this expression to calculate n as a function of δ, it is found that δ is a minimum, so that a measurement of the minimum deviation produced by a prism gives the value of its index of refraction in a direct and accurate way. This method can also be used for liquids; a triangular glass cell with walls of reasonable thickness is made very accurately and used to measure δ. The refraction due to the walls does not affect the measurement, since the refraction occurring at the entrance wall is equal but opposite to that at the exit wall.

The description of prism behavior given so far assumes that we are dealing with monochromatic light; that is light of a single wavelength or single

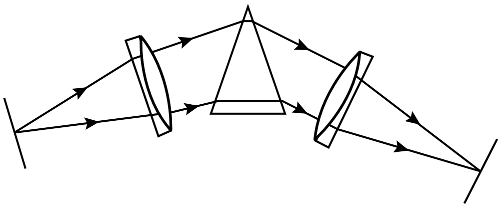

Figure 10 A prism spectrometer.

color in the visible spectrum. However, the index n of a prism varies with wavelength and white light passing through a prism will undergo dispersion – each component of a beam will emerge at a different angle and can be observed. The prisms that are part of expensive glass chandeliers display such an effect. This property of a prism is used in spectrometers – instruments which can measure the wavelength of light. Figure 10 shows a schematic diagram of this kind of instrument of high quality. Note that the light is brought into and taken out of the spectrometer with doublets. As shown in **Geometrical Optics:** Aberrations, doublets can be designed to be achromatic – that is, they focus light throughout the spectrum in approximately the same way. Various kinds of spectrometers have been put into use, with the ability to resolve closely spaced wavelengths being a crucial factor in the design.

See also

Geometrical Optics: Aberrations; Lenses and Mirrors.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Longhurst RS (1967) *Geometrical and Physical Optics*, 2nd edn. New York: John Wiley & Sons.

MIL-HDBK-141 (1962) *Military Standardization Handbook*. Washington, DC: US Government Printing Office.

H

HOLOGRAPHY, APPLICATIONS

Contents

Art Holography
High-Resolution Holographic Imaging and Subsea Holography
Holographic Recording Materials and Their Processing

Art Holography

A Pepper, Shearwater Foundation, Nottingham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Holography captures the phase and amplitude of light reflected from, or transmitted through, an object. In visual terms it captures 'true' three-dimensional space and any objects which are placed within that space. This is very different from the various optical illusions, including perspective drawing and painting, which 'trick' the eye into thinking it is seeing a dimensional image on a flat surface. Not surprisingly, artists have been intrigued by display holography since reports about it became more public in the early 1960s.

Since much of the information about the holographic process was originally published in professional scientific and optics journals, which the artistic community rarely read, or had access to, the transfer of practical information from optical scientist to visual artist was slow. As less technical reports about holography began to appear in publications, such as *Scientific American*, and art journals, such as *Leonardo*, more artists became aware of the possibilities holography offered. It is now over 50 years since its invention by Dennis Gabor, and art holography can be found in public and private collections around the world, ranging from major museums, through academic art galleries to large individual collections.

For the purpose of this article, holography can be defined as a section of a broad optical field, encompassing display holograms produced by artists, creative individuals, and groups. These works are often produced through independent or self-driven research, but can also be the result of a commissioned project. Holographic art works tend not to be produced specifically for commercial gain (although some do sell for large amounts and can be editioned in the same way prints or photographs, to increase their salability or make them more accessible to a wider audience). It is incredibly difficult to assess the number of artists working in this field today, as no specific research has been carried out, but it can be estimated as between 50–100 individuals worldwide, which includes those who have shown work in curated group and solo exhibitions. This relatively small number has made an enormous impact and includes artists who have made holography their main medium of expression, as well as those, more established in other media, who have extended their creative vocabulary by embracing the holographic process. Salvador Dali for example, had holograms made of his sculptural compositions in the early 1970s, and Bruce Nauman used the process to capture a series of self-portraits in the 1960s. Apart from these well known and established contemporary artists, much of the pioneering work has been driven by individuals who have concentrated on holography as their medium of choice.

Some artists build their own optical studio in which to experiment in a hands-on environment, some collaborate with scientific and research establishments, where the work, facilities and expertise is shared, and others design and supervise the production of their work which is produced for them by a commercial company. All of them produce a unique contribution to the visual arts.

Early Exploration

Holography offers a unique method of recording and reproducing three dimensions on a flat surface without the need for special glasses (for stereoscopic viewing), lenses (such as lenticular viewing) or linear perspective which, since the Renaissance, has been our main method of tricking the eye into believing that an image recedes into the distance and displays three dimensions.

Many of the very early works produced by artists concentrated on the obvious still life nature of the medium and explored the unusual phenomenon of objects 'frozen' in holographic space. Almost all of these early works were recorded using continuous wave helium neon lasers, either in scientific/research facilities, which artists had arranged to use, or in collaboration with established scientists and optical imaging researchers.

As early as 1969, British artist Margaret Benyon was experimenting with this new medium. Her laser transmission hologram 'Picasso', is a recording of images taken from Pablo Picasso's painting 'Demoiselles D'Avignon'. Here Margaret Benyon took a copy of the painting, cut out the characters, placed them in a three-dimensional spatial arrangement, with real objects, and then made a hologram of this 'scene'. Picasso attempted to show, through many of his paintings, multiple views of the same subject on a two-dimensional surface, one of the main thrusts of Cubism in the late twentieth century. Benyon achieved this in 'real space' by recording a hologram of the flat reproductions. By doing so, she demonstrated that holography automatically achieved what Picasso, and many other key contemporary visual artists, had been trying to do for years.

Holographic Black Holes

It has been known, from the very early days of holographic research that, when using a continuous wave laser to record a hologram, if any object in the scene moves more than half the wavelength of light, no interference pattern will be produced and therefore no recording of the object will take place. Unlike photography, where if an object moves during the exposure, it will become blurred, in holography it will simply not be recorded. Clearly, if no recording can be made, optical researchers working with holography, while noting the phenomenon, would not necessarily feel the need to make holograms which would fail. Several artists, on the other hand, not confined by research deadlines, have acknowledged this strange phenomenon and exploited this limitation of the holographic process. One of the very early examples is 'Hot Air' (Figure 1), a laser transmission hologram by Margaret Benyon and now in the Australian National Gallery Collection. Produced in 1970, it shows a three-dimensional composition of a jug, cup, and the artist's hand. The jug and cup were filled with hot water and during the recording of the holographic plate, with a continuous wave laser, the heat from the hot water disturbed the air above these objects. This tiny amount of movement resulted in the area above the jug and cup not being recorded. Similarly, the artist's hand moved slightly during the exposure and therefore did not record. Strangely, when viewing the resulting laser transmission hologram, the hand is visible. What is displayed is a holographic hand-shaped black hole, the three-dimensional recording of the space once occupied by the hand; the dimensional recording of the absence of an object.

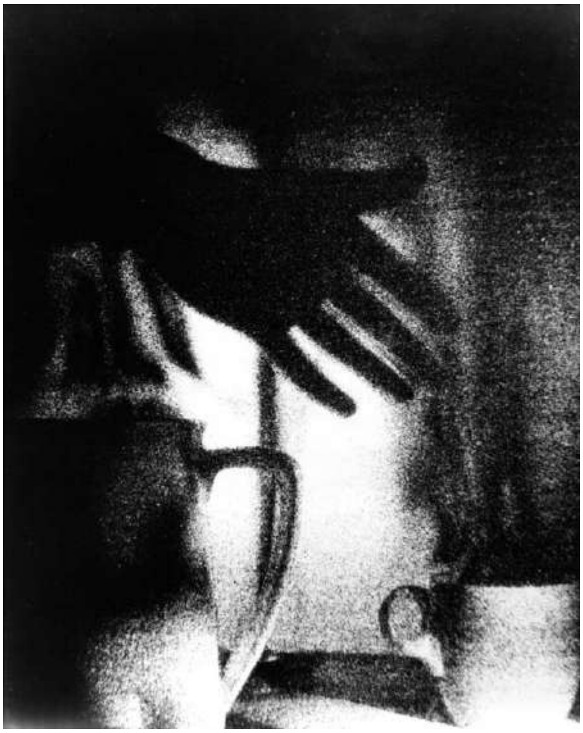

Figure 1 'Hot Air', Margaret Benyon 1970. Laser transmission hologram. Copyright © 1970 M. Benyon. Reproduced with permission.

> Artists in display holography place continual demands on the technology. Their natural tendency is to break rules. I believe that most successful holographic artists belong to this 'suck it and see' category of experimentalists. On hearing about the 'rules' of holography, they break them, just to see what happens. (Margaret Benyon, PhD Thesis, 1994).

Benyon started by breaking the rules and has continued to develop from these early pioneering laser transmission works, which helped establish a visual 'vocabulary' including many important social and political issues.

Other artists have exploited the 'limitations' of the holographic process to produce visually stunning works which question how we look at the three-dimensional world and question what is 'real'. Rick Silberman (USA) developed a shadowgram technique where the hologram is made of a laser illuminated ground glass screen. Objects are placed in front of this luminous surface and their three-dimensional silhouette is recorded. The viewer of the hologram is looking at the space where the object used to be. A pivotal work in this field is 'The Meeting' 1979 (Figure 2), a white light reflection hologram which had been recorded in such a way as to cause the three-dimensional shadow of a wine glass to protrude out of the holographic plate and be viewable several centimeters in front of it, between the observer and the holographic emulsion. The actual wine glass used in the recording has been broken and what remains (the base, stem, and small part of the bowl of the glass) is placed on a platform attached to the holographic plate. When viewing the work head on, the missing parts of the broken glass are recreated, in three dimensions, by the holographic shadow. When viewed obliquely, from the side for example, only the broken (real), glass is visible. Not only does this piece have an element of performance in the recording, breaking, mounting, and reconstruction of the wine glass, it also explores questions of visual perception and concepts of reality.

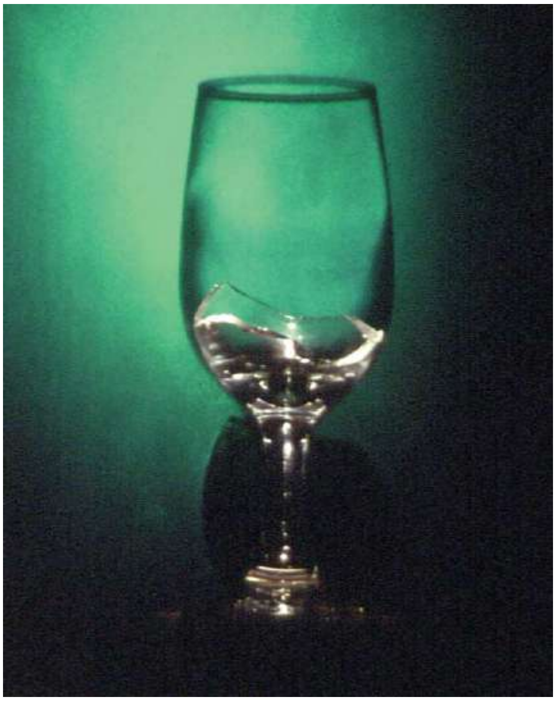

Figure 2 'The Meeting', Rick Silberman 1979. White light reflection hologram (shadowgram). Copyright © 1980 ATP. Reproduced with permission.

The recording of the shadow of an object, or objects, has now become commonplace in creative holography, with many artists employing techniques to mix different objects together, change their color and rearrange them in space. One advantage of this technique is that dependence on absolute stability of the objects used is reduced. The objects are 'encouraged' to move during the recording process and so create 'negative' images. Steve Weinstock (USA), Marie-Andrée Cossette (Canada), Michael Wenyon and Susan Gamble (USA/UK) have all contributed to this area in some of their early works.

Public Display

As the technology and optical processes of holography have developed, so greater flexibility, and public accessibility, have emerged. The single most important aspect has been the ability to display holograms in white light.

Emmett Leith and Uris Upatnieks invented laser transmission holography in 1964. They demonstrated the practicalities of Gabor's original invention, by utilizing laser light, to produce three-dimensional images. To display these recorded images, laser light needed to be shone through the holographic plate, at the original angle used by the recording reference beam (off-axis). The results were breathtaking, with full visual parallax in all directions. Although it was the laser light which made the recording and reconstruction of these unusual images possible, the laser was also a limiting factor. Each time a hologram was displayed, a laser was required. This relegated the technique to displays in scientific conferences, laboratories, or other locations with optical equipment. What was needed was a technique allowing holograms to be displayed without the need for this expensive and dangerous laser light. White light holography was the answer.

Although a laser is still required to record the original hologram, by using a copying technique, a restricted aperture or one step process, the resulting holograms can be reconstructed using a small source of white light such as a spotlight or even direct sunlight. As expensive lasers were no longer needed to display holograms, large public exhibitions could be mounted. Gallery and museum spaces could include holography in their exhibitions and individuals could buy and display holography in domestic locations. This technique also benefited the commercial display market, where holograms could now be used for advertising and promotional events.

Stephen Benton (USA) (1942–2003), inventor of white light transmission (rainbow) holography in 1968, made a major contribution to the use of the medium by artists. His technique not only allowed holograms to be viewed in white light, but they were extremely bright and offered the possibility of creating multicolor images. His ‘Crystal Beginnings’ 1977, (Figure 3), shows just how much space can be created with very simple points of light. Benton even taught master classes for artists so that he could explain how his technique functioned and explain the mathematics involved. He also worked with several artists to help them realize their individual projects.

Harriet Casdin-Silver (USA) was one of the first pioneering artists to work with Benton and produced several key works now seen as integral to the art of holography. Their early collaboration ‘Cobweb Space’, 1972, sculpted pure light and was one of the very first art holograms to be displayed using Benton’s rainbow hologram technique. A later work, ‘Equivocal Forks’, 1977, exists in different formats; as a laser transmission hologram showing the pseudoscopic display of domestic folks which protrude, inside out, from the holographic plate, and as a number of white light transmission holograms, of the same image, which were shown in an outdoor installation mounted by The Center for Advanced Visual Studies, part of the Massachusetts Institute of Technology. This demonstrated to a wide, nonscientific, audience, that holograms could be accessible, easily illuminated and weatherproof. Casdin-Silver has continued to produce art holography and has, most recently, worked with large-scale portraiture.

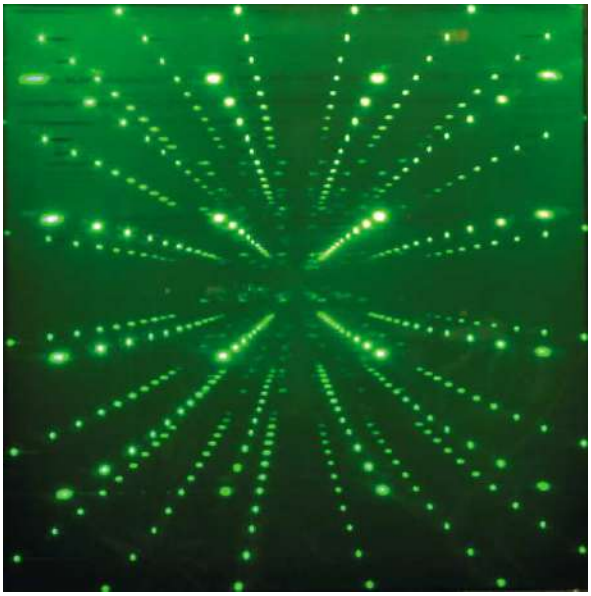

Figure 3 ‘Crystal Beginnings’ Stephen Benton 1977, White light transmission (rainbow) hologram. Jonathan Ross Collection, London, UK. One of the best known works by the inventor of the rainbow hologram. Copyright © 2004 J. Ross. Reproduced with permission.

Although Benton’s process increased the accessibility of holography, an earlier technique, invented by Yuri Denisyuk (Russia) in 1962, allowed holograms to be recorded in a single step and then displayed using white light. The political situation between the USSR and the rest of the world during the early 1960s meant that information and results from much of this early work was not openly accessible, and communication between Denisyuk and other key researchers in holography was limited. However, as soon as details of the Denisyuk technique became more widely available in the 1970s, artists began to embrace it as a practical alternative.

The work of Nicholas Phillips (UK), whose research improved chemical processing of display holograms during the late 1970s, helped artists and the rest of the holographic field to achieve much better results when making reflection holograms, a technique favored in Europe. He was also heavily involved with the rock band, ‘The Who’, and their Holoco company which mounted two very successful exhibitions of display holography at the Royal Academy of Art, London, at this time. Artists who might not have come across the holographic phenomenon became aware of its potential due to its presence in one of the UK’s more traditional art venues. Other large exhibitions in the late 1970s and early 1980s in USA, Sweden, and Japan also highlighted the immense possibilities holography offered to a nonscientific community of gallery visitors.

As interest in all aspects of holography increased, dedicated spaces opened to sell, display, and archive the output from artists, scientists, and commercial producers. One of the most influential of these was the Museum of Holography, located in the SoHo art district of Lower Manhattan, New York (Figure 4). From 1976 through to 1992, it regularly mounted group and solo exhibitions by artists, giving many of

Figure 4 Dr Dennis Gabor, inventor of holography with, Posy Jackson, Director of the Museum of Holography, during his visit to the Museum in March, 1977. Copyright © 1977 Paul D. Barefoot. Reproduced with permission.

them their first major exposure to a museum-going public. There were displays relating to the scientific, commercial, and industrial usage of the medium, as well as exceptionally popular traveling exhibitions. ‘Through the Looking Glass’, which traveled worldwide, allowed early display holography to be experienced by thousands of people who would not normally have come into contact with the medium. In the Summer of 1981, over 300,000 visitors saw ‘Through the Looking Glass’ during the 11 weeks it was on view at the Israel Museum, Jerusalem.

After the Museum of Holography closed in 1992, its entire collection was purchased by the MIT Museum, Boston, USA, and now forms part of the largest museum collection of creative and scientific holography in the world. The New York museum was not an isolated venue. During the late 1970s and early 1980s, museums of holography opened in Chicago, Paris, and Pulheim (near Cologne). Galleries were also founded in many major cities to sell art and commercial holograms.

Stability Solutions

Although being able to display holograms using inexpensive sources of white light allowed the medium to become much more accepted as a display process, there has always been the cost and complexity of the optical laboratories needed to record the holograms in the first place. Few artists were able to equip their studios with ‘off the shelf’ optical components and precision-built isolation tables most often found in scientific and research labs. The solution was to replace the commercially built isolation table with much less expensive sand tables.

Sand box holography revolutionized the field for independent artists. By building a wood or brick box filled with sand, resting on car inner tubes, artists were able to hand-make their own isolation tables. Inexpensive optics, mounted onto plastic pipes were simply thrust into the sand, allowing complex optical recording setups to be constructed. Surprisingly, many key artworks have been produced using this technique, which opened the field to anyone with basic construction skills and a laser.

The laser has remained an expensive element in the holography studio but, most recently, the prospect of replacing this with laser diodes has offered a much more economic alternative. Not only can light from these diodes provide more compact optical setups to record holograms, they can also be used to display them. Significant progress has been made in both diode recording and display, which is opening up new opportunities for the visual artist. Almost all of the early practitioners in the field were attracted to holography because of the laser transmission holograms they saw. They have a depth and dimensional quality which is extremely different from the white light viewable holograms we tend to see today. Now, with the possibility of displaying holograms illuminated by inexpensive laser diodes, the laser transmission hologram is set, once again, to be a viable medium for artists.

Object Holography

Many artists were attracted to holography because of its unusual three-dimensional properties. However, they did not have the resources to make holograms of large objects or scenes, so scaled these down and constructed small sculptural settings for recording. One of the pioneering artists in this area was Dan Schweitzer (1946–2001). Originally an actor and familiar with theatrical staging, many of his holograms are populated by tiny figures placed in specially built ‘theatrical’ settings. Holography not only recorded these specially constructed miniature ‘rooms’, but the resulting images were realistic enough to suggest to the viewer that they were looking at a scaled-down life-size event. We do this regularly when looking at photographs and accept that they are a small-scale version of a larger scene. Traditional holography does not offer this ‘scaling down’ in the same way as photography. One celebrated example is Schweitzer’s ‘Thendara’, 1978, where a sculpted, miniature figure sits and looks towards a window frame. He also produced a companion piece in which the physical model, used to make the earlier hologram, is on display as a piece of sculpture. Here the miniature window is the frame for a hologram displaying a previously recorded model of a landscape. As Schweitzer’s technical skills

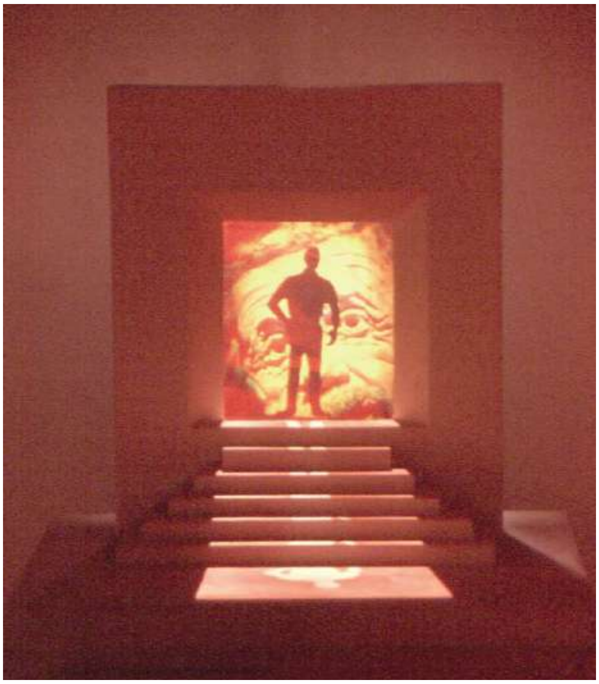

Figure 5 'Od Vent', Dan Schweitzer 1980, White light transmission hologram in sculptural mount. Copyright © 1980 ATP. Reproduced with permission.

developed, he began to produce complex multilayered works incorporating still-life models, hologram fragments, and optical animation to construct visual spaces which could not exist in the real world.

In his work 'Od Vent' (1980) (Figure 5), Schweitzer framed an exceptionally complex white light transmission hologram into a theatrical model of a doorway. The miniature stairs lead up to an opening, a portal into another world, in which the hologram is mounted. Illuminated from behind, the hologram displays multiple exposures, optical animation, and several superimpositions of objects as an observer moves from left to right in front of the piece. In this complex holographic world, the silhouette of the figure, standing in the door, casts a bright (negative) shadow down the miniature stairs. Not only are we able to look into an illusionary world through the holographic doorway, where at a certain point Einstein looks back at us, but the peculiarities of this holographic world extend into our own world as seen by the 'real' (negative) shadow cast down onto the stairs.

Another artist, who combined the use of multiple exposed objects displayed in a single space, is John Kaufman (USA). He established a reputation by recording objects (often rocks) collected from around his Californian holography studio. These found objects were captured in three dimensions and were then, through a multiple exposure and emulsion swelling technique, combined into a single still-life composition. He became expert at masking out areas of his objects so that they could appear in different, intense colors, and in many cases used the holographic process and multiple exposures to mix colors directly in the holographic space he was working with. In 'Stone Head' (1982) (Figure 6), for example, this relatively 'normal' stone is transformed by the use of intense color and the three-dimensional volume it occupies.

When considering how the hologram should be framed and displayed, several artists have utilized the mount in which the hologram sits as an integral part of a much more complex work. Frithioff Johansen (Denmark) recorded a hologram of smoke, freezing it in time and space using a pulsed laser (see below). This was then converted into a white light transmission hologram and mounted in a frame of granite. 'Hardware-Software 1' 1987–1988, allows the gallery visitor to not only view the massiveness of the granite structure, but also the delicate sculptural solidity of smoke 'frozen' in space.

Rebecca Deem (USA) recorded what appears to be an extremely simple hologram of a luminous tube, mounted onto a swivel at the end of an elegant metal support. In 'W/Hole Message' viewers can look into the small holographic plate and see the image of the tube extending away from them into the volume behind the plate. By flipping the hologram, on its swivel, to turn it over, the image in the hologram is also flipped and the tube appears in the viewer's space, between them and the plate. This three-dimensional pseudoscopic image, one of the attractive and unusual aspects of holographic display, is inside out, back to front, and upside down. The gallery visitor really does have a chance to view the whole message.

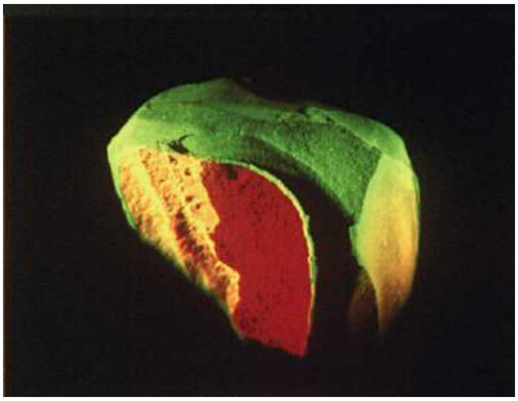

Figure 6 'Stone Head', John Kaufman 1982, White light reflection hologram. Copyright © 1982 John Kaufman. Reproduced with permission.

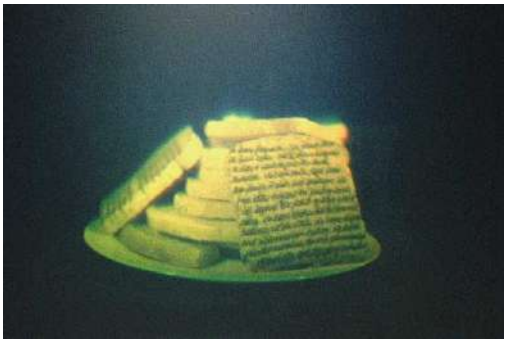

Figure 7 'Breakfast', Pearl John 2002. Laser etched toast displayed in a white light reflection hologram. Copyright © 2002 Pearl John. Reproduced with permission.

Text has become an important element in holography, as it is often incorporated in, onto, or around holographic displays. Eduardo Kac (Brazil) has produced several holograms, which display poems, words, and letters floating within the holographic space he has created.

Pearl John (UK) most recently produced a series of still-life holograms using a pulsed laser but with the added subtlety of writing onto the surface of her objects. In 'Breakfast' 2002 (Figure 7), for example, the slices of toast, making up the composition, have been laser etched with text which appears 'burnt' into the surface of the bread.

Pulsed Lasers

Continuous wave lasers placed limits on the holographic recording process and dictated that no movement in the recorded object could take place. We have seen that this 'problem' was exploited by artists at an early stage, but there was also a desire to capture living objects, in the form of the human figure. Pulsed laser holography was the solution. Because this type of laser generates a bright flash of coherent laser light, the intensity and duration of the pulse is enough to 'freeze' any physical movement during the recording process, allowing the recording of people, animals, or anything which moves and could be brought into an holography studio. Again, the cost and accessibility of this specialist equipment was prohibitive to artists, but they found ways of persuading scientists with pulsed lasers to help them. Early adopters of this technique were artists such as Ana Maria Nicholson (USA), Melissa Crenshaw and Sydney Dinsmore (Canada), Margaret Benyon (UK), Anait (USA), and Alexander (USA/UK). Recordings could now be made of live figurative compositions and individual portraits. Many of the artists, not content with producing the holographic version of a traditional photographic portrait, began to manipulate these recordings, distorting, multiple exposing and inverting these unusual three-dimensional images of people.

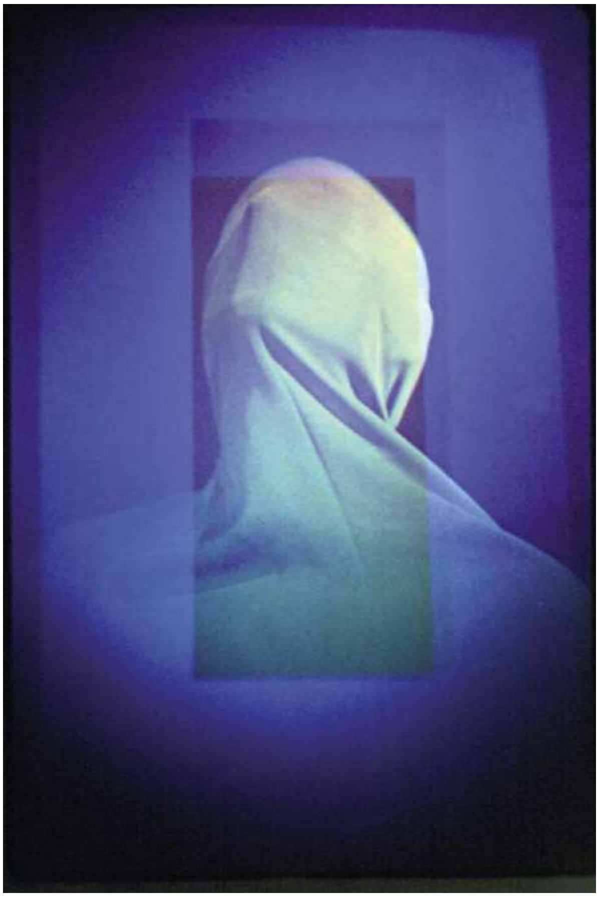

Figure 8 'Awakening', Ana Maria Nicholson, 1993. White light (pulsed) reflection hologram. Copyright © 1993 Ana Maria Nicholson. Reproduced with permission.

Holographic portraits do have an unusual quality, which painters or photographers have found difficult to deal with but art holographers have embraced. Some, like Ana Maria Nicholson (USA), have combined subtle processing techniques, to introduce color 'layers' into the portraits, combined with nonstandard views or compositions (see 'Awakening' 1993 (Figure 8)). Other artists, who emerged in the 1980s, produced complex, highly staged compositions, such as those from Patrick Boyd (UK), or they offered us a different view of the 'self portrait' like those by Martin Richardson (UK). Others, such as Shu-Min Lin (Taiwan), used the combination of multiple views of human figures recorded onto the same holographic plate. By walking past these displays, an animated effect could be experienced.

So much interest was generated in pulsed holography that several studios emerged for use by artists, including independent facilities and those within educational establishments. No longer was this the domain of the well funded scientific research facility. Several of these 'studios' still exist and are offering artists access to equipment through specialized artist-in-residence programs.

Abstraction

Although holography is an exceptional process for recording every detail of an object and the volume it occupies, many artists have wanted to work with the abstract and kinetic possibilities the process offers. Light can be recorded and displayed in three dimensions, it can be animated, inverted, compressed, sculpted, and projected. Douglas Tyler (USA), in his In the 'Dream Point Series' 1983 (Figure 9), recorded small laser transmission holograms but displayed them using normal white light, producing a smearing effect of the three-dimensional object. He then mounted these onto transparent surfaces (Plexiglas) and drew on the surface of the plastic to incorporate etched lines and shapes. Rudie Berkhout (Netherlands/USA) began his holographic exploration with stunning abstract landscapes made of luminous, multicolor, animated objects which took advantage of his mastery of the white light transmission technique. Most recently he has been recording reflection holograms which display undulating colored 'waves' of light as an observer moves past the hologram. These have been included in specially produced paintings for large-scale architectural commissions.

Sam Moree (USA) combines multiple images, exposures, diffraction gratings, and smaller holograms into a single work. The finished holograms display a complex series of image fragments and colors within the depth of the holographic space he has constructed. Ruben Nuñez (Venezuela), originally approached holography from a background in glass production and used the ability of the holographic process to 'capture' light reflecting and passing through his glass objects. The resulting white light transmission holograms display an abstract 'photonic' display of form and color. Fred Unterseher (USA), uses holography to record optical distortions and diffraction gratings which would display abstract patterns of light. Many of these early works were mandala images which displayed an intense kinetic effect as an observer walked past the surface of the holographic plate.

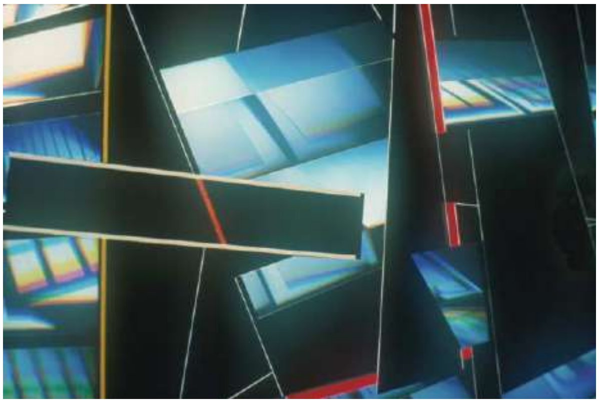

Figure 9 'Dream Passages': 32″ × 138″ (3 panels 32″ × 46″). Douglas E. Tyler 1983, Laser transmission holograms, tape and Plexiglas, illuminated with white light. Detail of center panel. Copyright © 1983 D.E. Tyler. Reproduced with permission.

Pseudo Holograms

In 1973, Lloyd Cross invented the multiplex (integral) hologram technique which used sequential two-dimensional images recorded on movie film combined with white light (rainbow) hologram techniques. The result was a holographic stereogram, which gave viewers the illusion of seeing animated images in three dimensions. The final holographic sheet was displayed either curved or as a complete cylinder, which was often motorized so that viewers could stand in front of the unit and watch the animation take place as the hologram rotated. Such a display does give the impression that an observer can walk all the way round the image, enhancing its perceived three-dimensional effect.

As the original images are recorded photographically, there are several advantages – anything which can be recorded on inexpensive black and white movie film can be converted into a holographic display. Short animated sequences can be presented while accepted film techniques such as zoom, pan, tilt, and transitions can be incorporated into the finished sequence of images. Although very popular for advertizing and promotional campaigns during the 1980s, particularly in the USA, several artists used this technique. They could make the movie film sequences themselves and then send this film to a commercial producer for a hologram to be made for them. This offered an interesting balance in the creative process. The artist could concentrate on the filmic image, using facilities and photographic techniques they were familiar with and not have to learn the holographic process or build a holographic laboratory themselves. Artists such as Anait (USA) produced several experiments in the area, one of which showed her inside the holographic display,

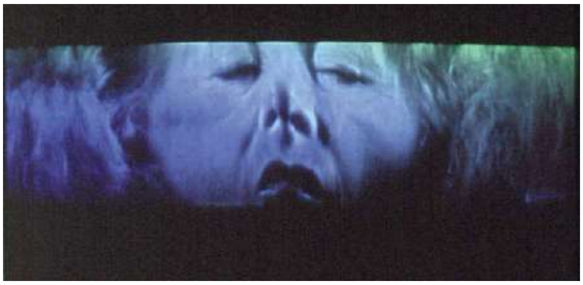

Figure 10 'Woman' Harriet Casdin-Silver 1979. Multiplex hologram (white light transmission). Copyright © 1980 ATP. Reproduced with permission.

writing, with spray cream, seemingly onto the holographic surface. Dieter Jung (Germany), used this technique to display a specially written poem which appeared to spiral around inside of the display.

Harriet Casdin-Silver used the technique to produce an unusual self-portrait. 'Woman', 1979 (Figure 10), is a multiplex hologram of the artist's head. Black and white film images were recorded of Casdin-Silver's head, through 360 degrees, and then each frame of film was converted into a thin white light transmission hologram. The artist chose not to display the resulting image in a normal cylindrical display, but to 'flatten' it out so that the front, sides, and back of her head could be seen, distorted, in a single head-on view.

As video and digital imaging processes have developed over the years they too have been combined with this pseudo holographic technique, opening up a wide avenue of possibilities. Video footage, computer generated modeling, or any process which can produce sequential 2-D images, can now be converted into an animated 3-D hologram. As digital video cameras and computer 3-D modeling software drop in price, and increase in sophistication, a whole new area of display has become available to the digital and video artist.

Multi Media

As in other areas of the visual arts, more than one medium or process are often combined. Doris Vila (USA) has produced complex environments combining large-scale holographic images which, rather than being illuminated with a single source of light, are reconstructed with multiple spotlights or via video projectors. The lighting is activated by sensors in the exhibition space, which are computer controlled and made to trigger certain sequences dependent on the input they receive from the movement of visitors.

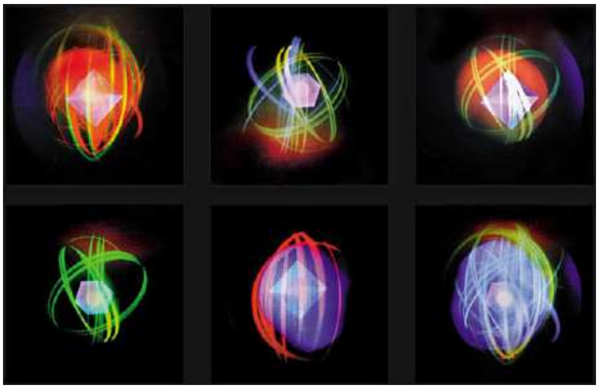

Figure 11 'Neuro Hologram', Ikuo Nakamura, 1993. White light transmission hologram with multiple, computer controlled, lighting. Copyright © 1993 Ikuo Nakamura. Reproduced with permission.

Traditional illumination of a hologram involves a single point source of light shining onto, or through, the holographic plate. This will reconstruct a single image. If multiple lights are shone onto the hologram, multiple reconstructions of the single image will result. By using many independently controlled lights on one hologram, Vila has been able to create animation and kinetic effects as the resulting multiple images overlap. Using light from a video projector also allows the artist to combine a reconstructed holographic image with two-dimensional animated footage from the video projector.

Other artists to combine holograms with structures, interactive lighting and environments, sound, and installational displays include: Georges Dyens (Canada), Marie-Christiane Mathieu (Canada), Matthew Schreiber (USA), Philippe Boissonnet (Canada) Ana MacArthur (USA) and Shunsuke Mitamura (Japan).

Ikuo Nakamura (Japan/USA) has used multiple lighting of a single computer generated hologram to produce animated effects controlled by the viewer. In his 'Neuro Hologram' 1993 (Figure 11), brain waves from the viewer, wearing a special set of sensors on their head, are measured and converted into control signals for a number of lights behind the hologram. By relaxing, or becoming agitated, the viewer can change their brain waves and actively control the appearance of the hologram, causing multiple views of the holographic image to fade in and out, giving an impression of abstract animation.

Architectural Scale

One of the great limitations of holography has been the size of the holographic plates on which the light sensitive emulsion is coated. These tend to be in the region of 20 × 23 cm, 30 × 40 cm, or meter-square formats. Generally, the larger the finished hologram, the larger the laboratory needed to produce it. This impacts a great deal on artists who do not have the facilities, or finances, to produce ‘large’ holograms. A solution has been to combine smaller holographic plates together, as a mosaic, to give the size and scale needed. In recent years, several artists have completed public commissions for large-scale interior and exterior holographic installations.

Setsuko Ishii, (Japan) has produced several installations in public spaces where individual holographic plates are mounted together to form a larger composite structure. Many of these pieces use holograms recorded on dichromate gelatin plates which produce an exceptionally bright image, ideal for public spaces, where visibility, from multiple angles, is important.

Michael Bleyenberg (Germany) produced an entire exterior holographic wall (13 × 5 meters) during 2002 (Figure 12), for the extension of a research building in Bonn. The installed graphic wall, ‘EyeFire’ (AugenFeuer), is made up of 26 smaller panels, each constructed from tiny holographic ‘pixels’ (holographic optical elements, HOE) to give an intense color effect. Working in collaboration with the Institute for Light and Building Technology, Cologne, Bleyenberg was able to design his graphic panels on a computer screen and have this information translated into the thousands of tiny holographic pixels which, when seen together, make up the complex and high intensity colour design.

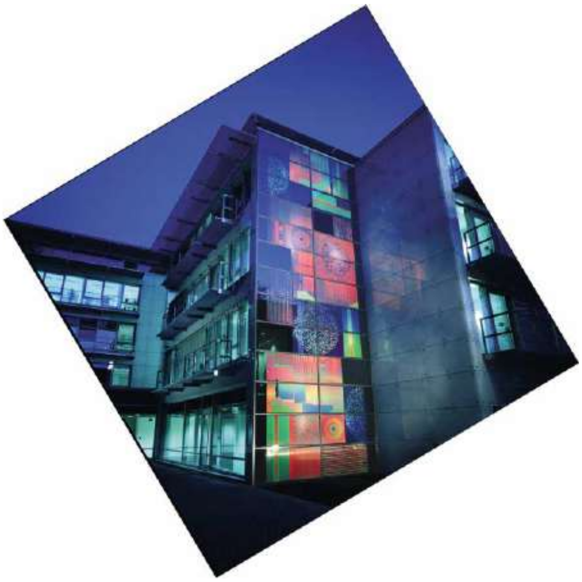

Figure 12 ‘EyeFire/AugenFeuer’, Michael Bleyenberg 2000. White light transmission holograms mounted onto building exterior. Copyright © 2000 M. Bleyenberg. Reproduced with permission.

Dieter Jung (Germany) has produced commissioned art works for large interior architectural spaces. In ‘Perpetuum Mobile’ 2001, located at the European Patent Office, The Hague, holograms are placed on the floor, which can be walked over, and an extensive holographic mobile structure hangs overhead. The light, and color, from the abstract holograms in the mobile sections, can be reflected in the abstract holographic floor and combine to produce an interior lightscape impossible to produce with more traditional media.

Holograms can be mass produced using a mechanical embossing process which transfers the optical contents of the original master hologram onto plastic sheeting for duplication in large runs. One novel use of this material has been the interior of the Canadian pavilion at Expo ‘92 (Figure 13). Melissa Crenshaw (Canada/USA) used huge quantities of embossed holographic material to cover the 3000 square foot interior walls of the specially built pavilion. Lighting was installed to illuminate this optical diffracting wall and the entire structure was covered in glass, over which water flowed. The resulting interior appeared to be a wall of liquid light which changed color and hue depending on where it was viewed from.

Sally Weber (USA) used holographic embossed material for several outdoor commissions. Structures were clad in embossed holographic sheets and made water resistant. Sunlight illuminated the installations to produce surfaces of constantly changing color. In direct (bright) sunlight, intense rainbow colors would be diffracted by the holographic surfaces. On an overcast day, pastel, more muted,

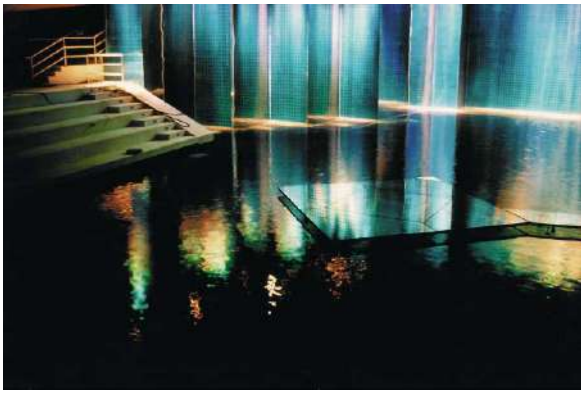

Figure 13 Interior of the Canadian Pavilion Expo ‘92, Spain. Holographic wall water feature by Melissa Crenshaw. White light transmission (embossed) hologram foil mounted into pavilion walls. Copyright © 1992 M. Crenshaw. Reproduced with permission.

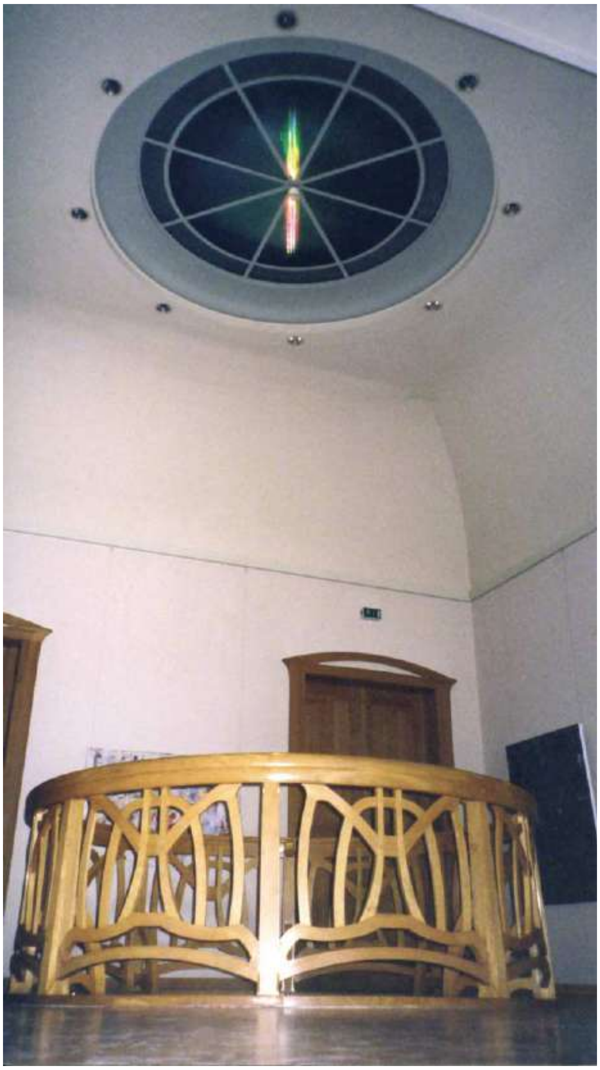

Figure 14 Signature of Source, Sally Weber 1997. Site specific window installation at the Karl Ernst Osthaus-Museum, Hagen, Germany. Copyright © 1997 Sally Weber. Reproduced with permission.

colors would appear. Although she is often known for these large-scale outdoor embossed works, one of her permanent commissions is housed in the Karl Ernst Osthaus-Museum, Hagen, Germany, where specially made holographic diffraction gratings are incorporated into an existing building. 'Signature of Source' 1997, (Figure 14), is a permanent, site-specific installation producing a line of light which enters the museum space from the skylight it is built into. This abstract line of pure color becomes a sculptural element within the space of the museum and is generated by the action of light from outside the building being manipulated and 'formed' as it passes through the holographic window.

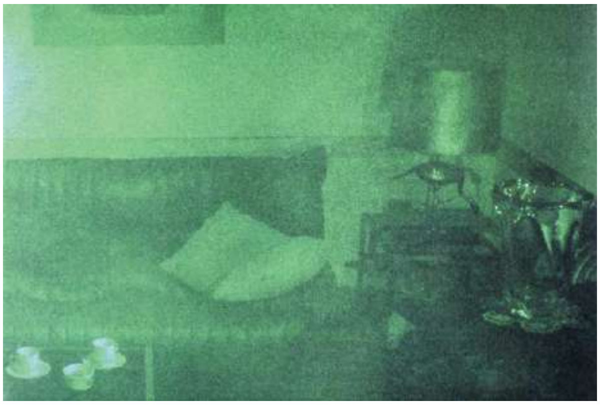

Figure 15 'There's no Place Like Home', Paula Dawson 1979–1980. Laser transmission hologram of an entire room. Copyright © 1980 Paula Dawson. Reproduced with permission.

One of the most celebrated and very early attempts to incorporate large-scale holograms into art works was the installation 'There's No Place Like Home' by Paula Dawson (Australia) in 1979–1980 (Figure 15). A laser transmission hologram of an entire room was recorded. The room was specially constructed, life size, and recorded onto a single holographic plate. The resulting hologram was then built into the window of a small 'house' constructed in the gallery. Visitors were able to approach the house, look into the window and 'see' the interior with its furniture, fixtures, and fittings. They would then enter the house and walk past the space where the room appeared to exist. When viewed from this vantage point there was nothing there except an empty space. Artists have often questioned and researched the concepts of 'inside and outside', voyeurism, reality, normality, and 'existence', but few have had the tenacity to produce an installation with such visual and conceptual impact.

Twenty-First Century Art

Art holography has developed and matured over the past 50 years to become a viable and significant element in the contemporary visual arts. Critics are still asking 'is it art?' in the same way they used to worry about whether photography was art. While they consider this problem, practitioners in the medium are continuing to find ways to push past its technical and visual boundaries (Figure 16). They continue to collaborate with the scientific and research world and use holography as a medium to comment on life and society in the early twenty-first century.

> …I hope, that one day I will be able to give you, artists, something really powerful; panoramic holograms,

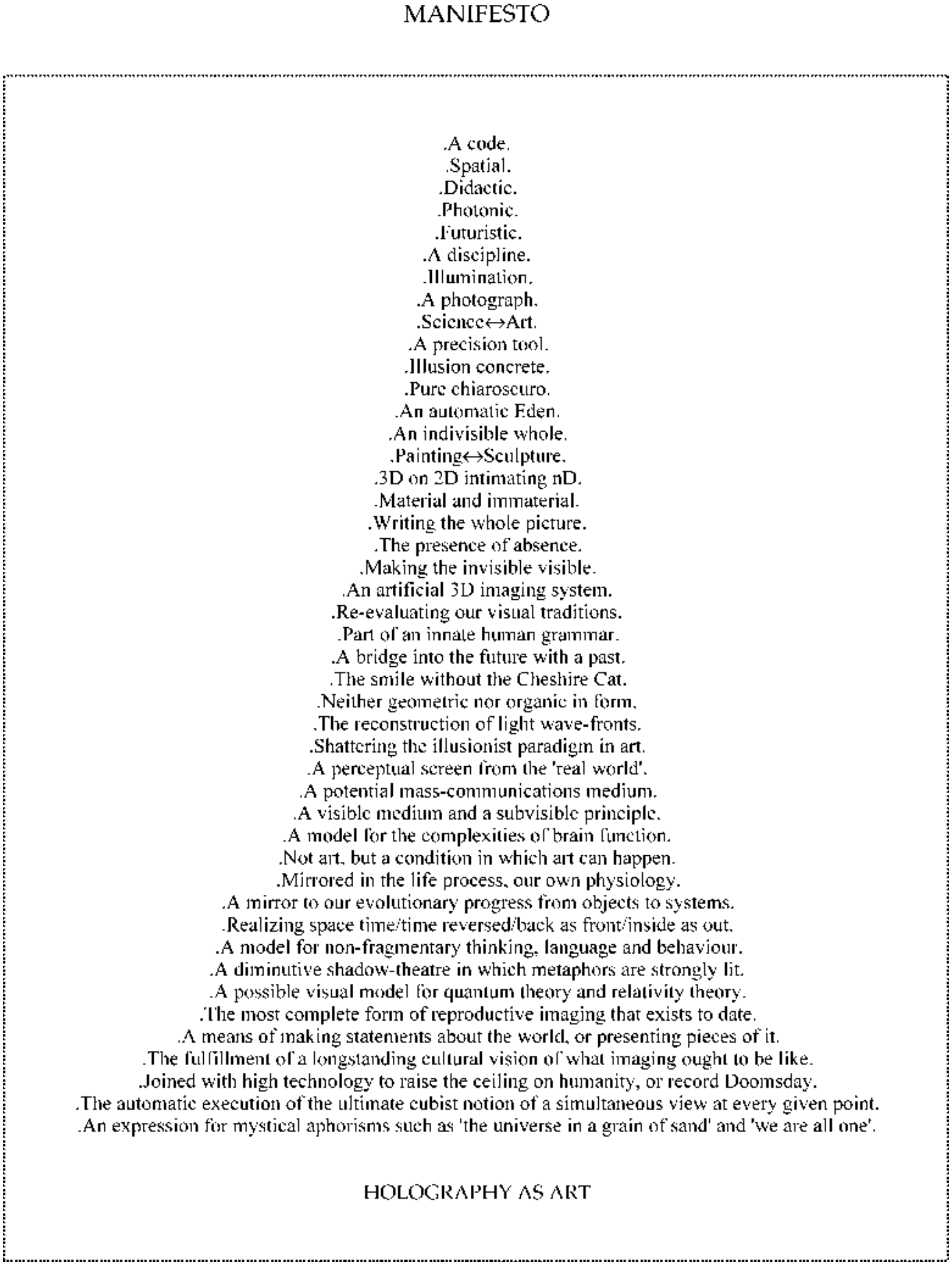

MANIFESTO

.A code.
.Spatial.
.Didactic.
.Photonic.
.Futuristic.
.A discipline.
.Illumination.
.A photograph.
.Science↔Art.
.A precision tool.
.Illusion concrete.
.Pure chiaroscuro.
.An automatic Eden.
.An indivisible whole.
.Painting↔Sculpture.
.3D on 2D intimating nD.
.Material and immaterial.
.Writing the whole picture.
.The presence of absence.
.Making the invisible visible.
.An artificial 3D imaging system.
.Re-evaluating our visual traditions.
.Part of an innate human grammar.
.A bridge into the future with a past.
.The smile without the Cheshire Cat.
.Neither geometric nor organic in form.
.The reconstruction of light wave-fronts.
.Shattering the illusionist paradigm in art.
.A perceptual screen from the 'real world'.
.A potential mass-communications medium.
.A visible medium and a subvisible principle.
.A model for the complexities of brain function.
.Not art, but a condition in which art can happen.
.Mirrored in the life process, our own physiology.
.A mirror to our evolutionary progress from objects to systems.
.Realizing space time/time reversed/back as front/inside as out.
.A model for non-fragmentary thinking, language and behaviour.
.A diminutive shadow-theatre in which metaphors are strongly lit.
.A possible visual model for quantum theory and relativity theory.
.The most complete form of reproductive imaging that exists to date.
.A means of making statements about the world, or presenting pieces of it.
.The fulfillment of a longstanding cultural vision of what imaging ought to be like.
.Joined with high technology to raise the ceiling on humanity, or record Doomsday.
.The automatic execution of the ultimate cubist notion of a simultaneous view at every given point.
.An expression for mystical aphorisms such as 'the universe in a grain of sand' and 'we are all one'.

HOLOGRAPHY AS ART

Figure 16 'Holography as Art', Margaret Benyon 1980. Manifesto. Copyright © 1980 M. Benyon. Reproduced with permission.

extending to infinity, in natural colors. Dennis Gabor, Inventor of holography in a letter to artist Margaret Benyon 28th February, 1973.

See also

Coherence: Coherence and Imaging. **Diffraction:** Diffraction Gratings. **Holography, Techniques:** Overview. **Imaging:** Multiplex Imaging. **Interferometry:** White Light Interferometry. **Introductory Article:** Early Development of the He–Ne Laser (Title TBC).

Further Reading

Benyon M (1994) *How is Holography Art?* PhD thesis, Royal College of Art, London.

Brill LM (ed.) (1989) Holography as an art medium, Leonardo. *Journal of the International Society of Arts, Sciences and Technology*, 22 (3 and 4).

Coyle R and Hayward P (1995) *Apparition, Holographic Art in Australia*. University of Sydney, NSW, Australia: Power Publications.

Casdin-Silver H (1998) *The Art of Holography Exhibition Catalogue*. Lincoln, MA, USA: DeCordova Museum and Sculpture Park.

Holographische Visionen, Bilder durch Licht zum Leben erweckt (1991) Exhibition Catalogue, Museum für Holographie und neue Visuelle Medien, Pulheim, Germany: Edition Braus.

Holosphere, Formerly published by the Museum of Holography. New York: Boston, USA: Archieved at the MIT Museum.

Jackson R and Burns J (1976) *Holografi: Det 3-Dimensionella Mediet, Exhibition Catalogue*. Museum of Holography, New York: Stockholm Cultural Center.

Jung D (ed.) (2003) *Holographic Network*. Germany: Rasch Verlag, Bramsach.

Orazem V and Schepers J (1992) *Olografia, Holografie, Holography, Exhibition Catalogue*. Rocca, Paolina, Italy: Centro Espositivo.

Lipp A and Zec P (1985) *Mehr Licht/More Light, Exhibition Catalogue*. Kabel-Verlag: Hamburger Kunsthalle.

Pepper AT (ed.) (1992) *The Creative Holography Index: The International Catalogue for Holography*. Germany: Monand Press.

Pepper AT (ed.) (1996) *Art in Holography 2*. International Symposium on Art in Holography. Nottingham, UK: University of Nottingham.

Margaret Benyon M (1989) *Phases: Exhibition Catalogue*. New York, USA: Museum of Holography.

Tyler DE (ed.) (1990) *Report, International Congress on Art in Holography*. Notre Dame, Indiana, USA: Saint Mary's College.

Tyler DE (ed.) (2003) *Leading Lights, Women in Holography, Exhibition Catalogue*. Notre Dame, Indiana, USA: Saint Mary's College.

High-Resolution Holographic Imaging and Subsea Holography

J Watson, University of Aberdeen, Aberdeen, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Holography is well-known to most of us: there can be few who are not familiar with the white-light display hologram seen in exhibitions and shops and on credit cards. The impact of holography, though, spreads far and wide into areas as diverse as vibration analysis, image processing, holographic-optical elements, and optical computers. One aspect often overlooked is its suitability for high-precision imaging and accurate dimensional measurement. It is the unique ability of holography to recreate, from a flat photographic film, a three-dimensional image that is optically indistinguishable from the original which sets it apart from other imaging techniques. In particular, holography comes into its own when recording objects in a hazardous environment or where access is difficult, such as underwater or in nuclear fuel element inspection.

We describe here the use of holography for imaging, identification, and measurement of aquatic organisms. A detailed knowledge of the distribution and dynamics of marine and freshwater organisms and particles, such as plankton, is crucial to our understanding of such systems and how they affect our environment. Traditional methods of gathering *in situ* data on aquatic particles, such as net-collection and electronic counting or photography, are not usually suited to observing precise spatial relationships and can destroy the more delicate organisms. The overriding benefit of holography is that it permits nonintrusive and nondestructive observation and analysis of organisms in their natural environment, while preserving their relative spatial distribution. Holography also offers a permanent archive. The use of a short-duration pulsed laser is essential, in order to reduce the effects of vibration or movement in any of the optical components or objects being recorded. Consequently, the object scene is effectively 'frozen' at the recording instant, thereby allowing fast moving particles to be imaged. Compared to a photograph, a hologram captures vastly more data. Sequential holograms can record changes within the volume over a defined period of time and the wide recording dynamic range afforded by a hologram allows images to be captured that would otherwise be lost in noise.

Methodology of Holographic Recording and Replay

Of the many possible recording methods of holography two in particular, 'in-line' and 'off-axis', find use for high-resolution *in situ* imaging underwater. The concepts and fundamental principles outlined here are generally applicable wherever high-precision measurement is needed from a hologram.

In an in-line reference beam hologram (ILH), a single laser beam (Figure 1) is directed through the

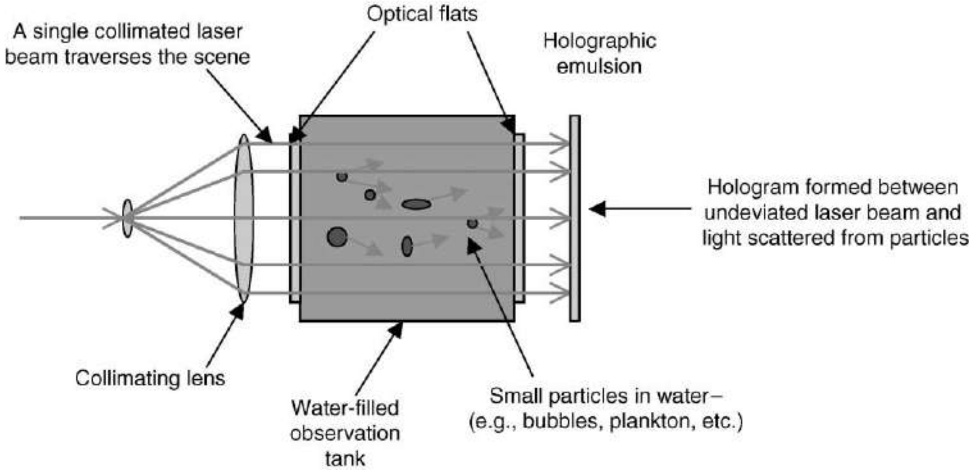
A single collimated laser beam traverses the scene
Optical flats
Holographic emulsion
Hologram formed between undeviated laser beam and light scattered from particles
Collimating lens
Water-filled observation tank
Small particles in water– (e.g., bubbles, plankton, etc.)

Figure 1 Recording an in-line hologram.

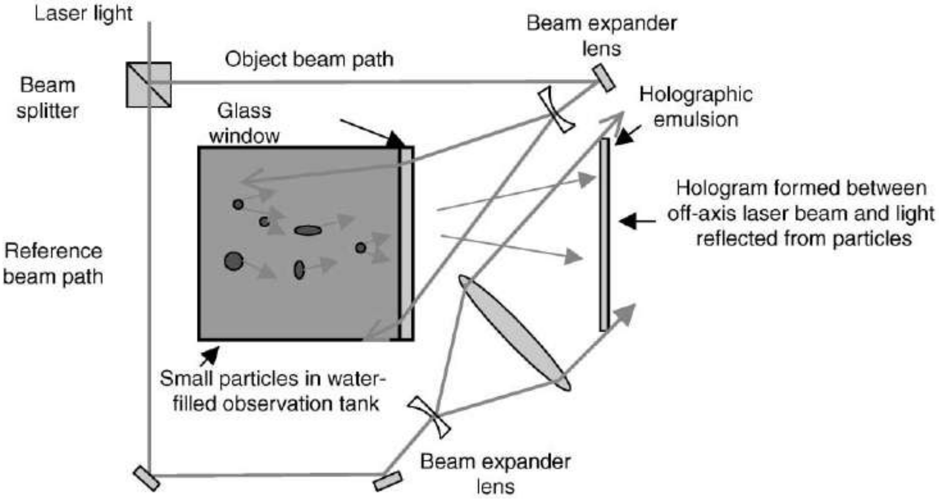
Laser light
Beam splitter
Object beam path
Beam expander lens
Holographic emulsion
Glass window
Hologram formed between off-axis laser beam and light reflected from particles
Reference beam path
Small particles in water-filled observation tank
Beam expander lens

Figure 2 Recording an off-axis hologram.

sample volume towards the holographic plate and records the optical interference between light diffracted by the object and the undeviated portion of the illuminating beam. The subject volume is illuminated in transmission and the scene needs an overall transparency of about 80% so that speckle noise does not seriously degrade image quality. Particles down to around 5 μm dimension can be readily recorded and identified. The particular benefits of ILH are its geometric simplicity, minimization of laser coherence and energy requirements, and high resolution over a large depth of field.

For recording an off-axis reference beam hologram (OAH), a two beam geometry is utilized: one beam illuminates the scene and the other directly illuminates the holographic film at an oblique incidence angle (Figure 2). Interference occurs between diffuse light reflected from the scene and the angularly separated reference beam. On replay, the real and virtual images are similarly angularly separated which makes their interrogation easier. OAH is primarily applied to opaque subjects of large volume. Although there is no real upper limit to the size of particles that can be recorded (this is determined by available energy and coherence of the laser) the practical lower limit to the resolution that can be achieved is around 50 to 100 μm. The scene can be front, side or back illuminated (or all three combined) allowing high concentrations of large particles to be recorded, provided that they reflect sufficient light to the film, and allowing subject visibility to be optimized. OAH provides a more complete record than ILH, in the sense that it gives information on the surface structure of the object, while ILH records only a silhouette.

Holograms are replayed or reconstructed in a variety of ways, depending on the recording parameters and subsequent processing of the recording medium. In scientific and engineering applications the requirements of high resolution and image fidelity demand that they are replayed with a laser, and viewed with some auxiliary optical system and (usually) a video camera. Depending on the orientation of the hologram, either the virtual image, which is apparently located behind the hologram and seen through it, or the real image, which is formed in space in front of the hologram, will be visible.

Considering first the off-axis hologram in the virtual image mode, the processed hologram is replaced in the position in which it was recorded and illuminated with an exact duplicate of the original reference wave, in terms of its wavelength, curvature, and beam angle. In this mode, which might be termed 'normal' viewing of an off-axis hologram, a virtual image (Figure 3) is observed as if viewing the scene through a window. Other than the colour, this image is almost optically indistinguishable from the original scene; it retains features such as parallax and three-dimensionality and does not suffer from perspective distortion or defocusing of the image.

For data extraction and measurement from a hologram, creation of the virtual image is not the best method of observation. Here the projected real image (Figure 4) is most useful. If we illuminate an off-axis hologram from behind, with a wave which is the exact conjugate of the original (i.e., one which possesses the same wavelength as the original but with the opposite direction and curvature), we generate an image that appears to float in the real space in front of the observer. Often a collimated reference beam is used in recording, since reconstruction is then a simple case of turning the hologram around. No lens is needed to form this image, and it is optically identical to the original wave that it is reversed left-to-right and back-to-front (pseudoscopic). Planar sections of the real image can be optically interrogated by directly projecting onto a video camera (often with the lens removed) which can be moved throughout the 3D volume. It is this 'optical sectioning' feature of a holographic recording that is so useful when nondestructive (noninvasive) evaluation must be made, at high resolution, of collections of objects which extend over a considerable depth.

A series of images taken from the real image of one hologram and captured on a video camera (Figure 5) illustrates how the viewing plane can be moved sequentially through the reconstruction to show firstly one marine organism and then another. The axial separation of the organisms is about 35 mm.

To replay an in-line hologram, it is also placed in a conjugate of the recording beam (Figure 6). The replayed hologram simultaneously forms two images, one virtual, the other real, which are located on the optic axis on opposite sides of the holographic plate. These images are co-axial, and located (for the typical case of collimated recording and replay beams) at equal distances in front of and behind the hologram plane. Although a true 3D real image is produced it has almost no parallax, and cannot easily be seen by the unaided eye. This image has inverted depth perspective and is seen in darkfield against the out-of-focus virtual image.

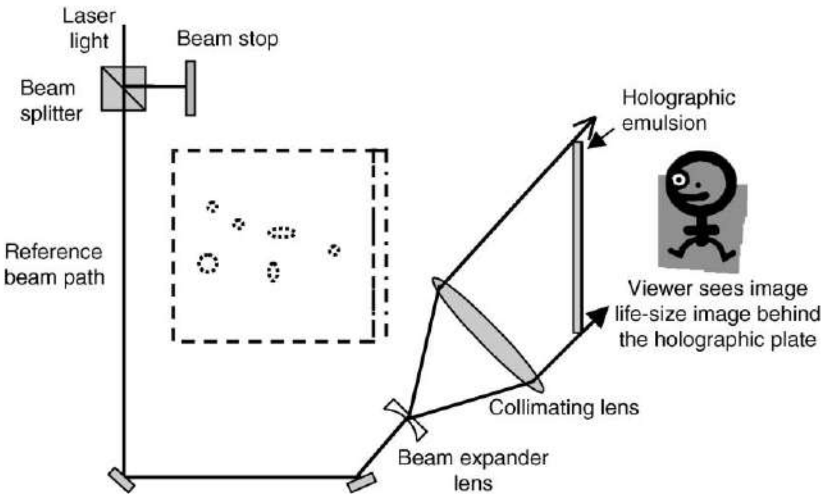

Figure 3 Replay of an off-axis hologram in virtual image mode.

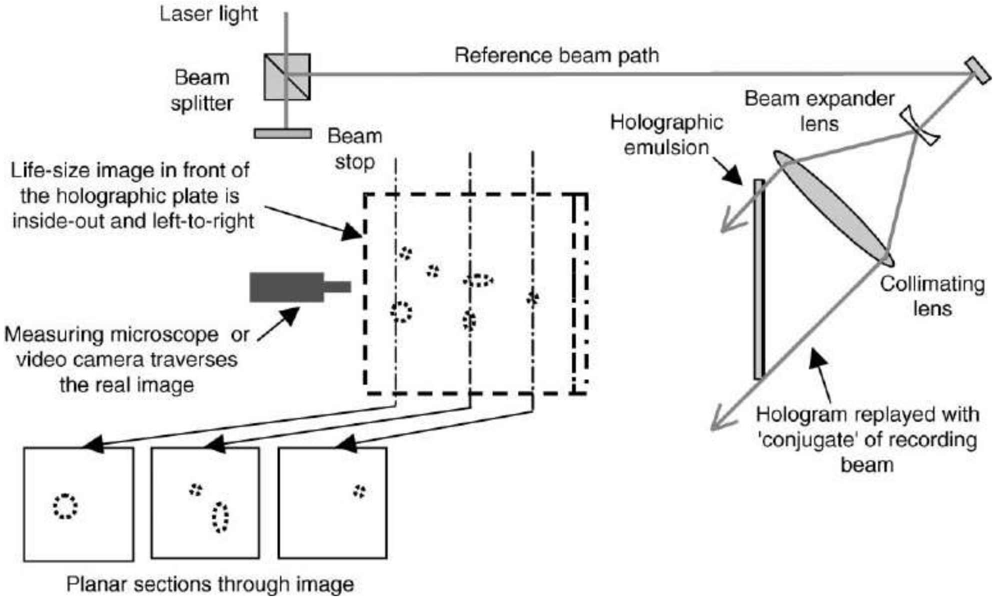
Laser light
Reference beam path
Beam splitter
Beam stop
Beam expander lens
Holographic emulsion
Life-size image in front of the holographic plate is inside-out and left-to-right
Collimating lens
Measuring microscope or video camera traverses the real image
Hologram replayed with 'conjugate' of recording beam
Planar sections through image

Figure 4 Replay of an off-axis hologram in real image mode.

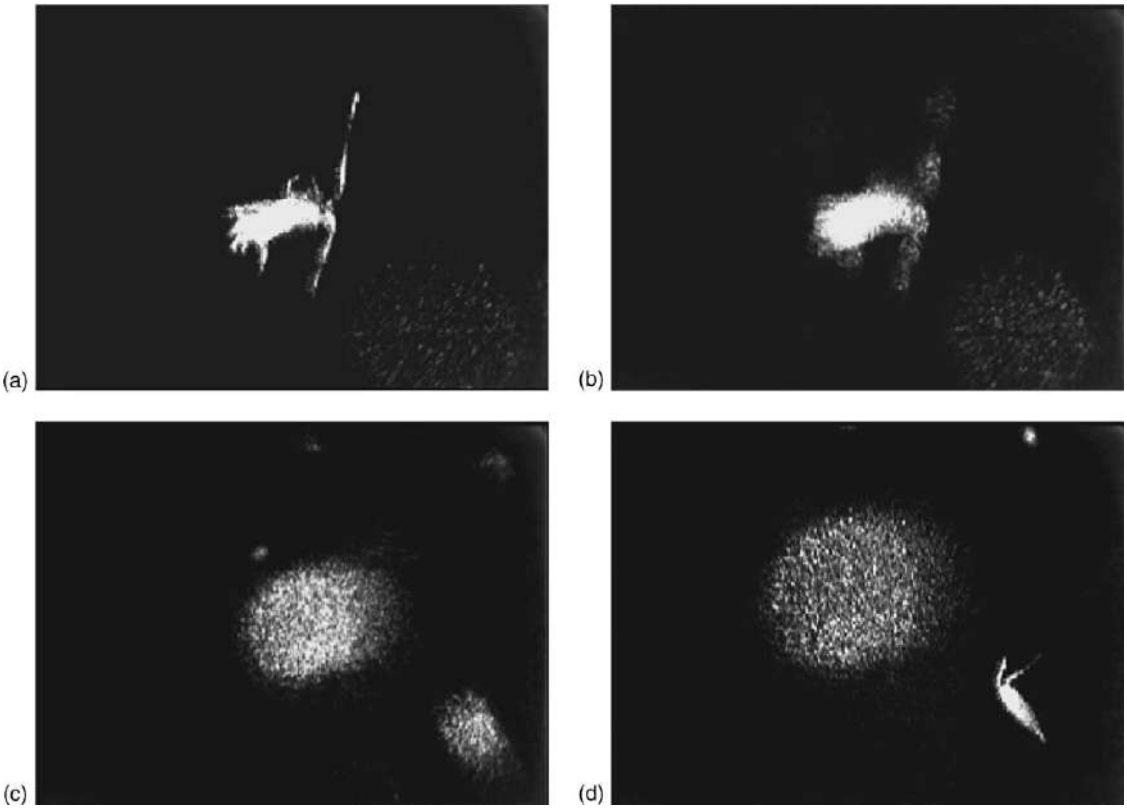

Figure 5 A series of images taken from a single hologram showing the concept of moving the viewing plane through the real image.

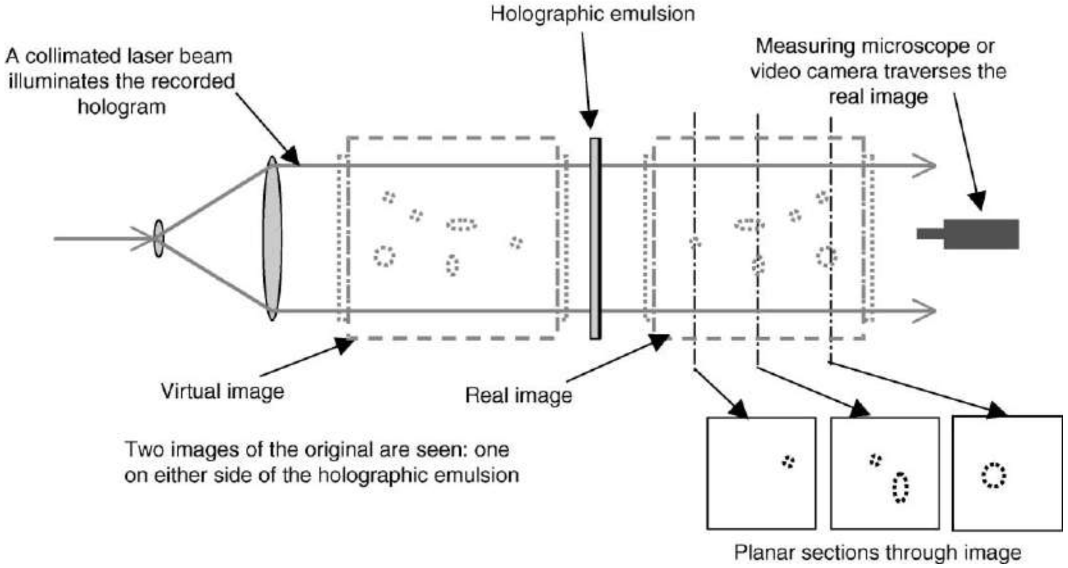
Holographic emulsion
A collimated laser beam illuminates the recorded hologram
Measuring microscope or video camera traverses the real image
Virtual image
Real image
Two images of the original are seen: one on either side of the holographic emulsion
Planar sections through image

Figure 6 Replay of an in-line hologram.

Holographic Image Quality

In general, the usefulness of holography for accurate inspection and measurement is dependent on its ability to reproduce an image of the object which is low in optical aberrations and high in image resolving power. Regardless of whether or not the hologram is recorded underwater, high resolution, high contrast, and low noise are the dominant requirements, rather than image brightness. In practice, loss of resolution and image degradation can occur at any stage in the recording and replay of a hologram. All the primary monochromatic aberrations to be found in any optical system (spherical aberration, astigmatism, coma, distortion, and field curvature), may be present in the holographic image. The most significant factors affecting this are departure from conjugacy of the reconstruction beam; positional mismatch and misalignment between reference beam and film; degradation of the reconstruction beam due to poor optical quality of the film substrate or optical components; and, distortion of the fringe pattern recorded in the holographic film as a result of chemical processing. For precise and accurate reconstruction of the real image, the reconstruction reference beam should be the exact phase conjugate of that used in recording. Under these conditions, the lateral, longitudinal, and angular magnifications of the real image with all equal unity and aberrations will be reduced to a minimum.

Diffraction-limited resolution of a holographically produced real image, in the absence of all aberrations, is usually defined from the ability to distinguish two points in the image separated by a distance r. The resolution as defined in this way is dependent upon reconstruction wavelength, λ_c, the separation between hologram and reconstructed image, s, and the effective aperture, D, of the hologram:

$$r = 1.22 s \lambda_c / D \qquad [1]$$

This relationship is well-known in imaging optics: it defines the ideal resolving power of a lens and is equally applicable to holography. In holography, though, the presence of speckle effects introduced by the coherence of the light and the finite aperture of the viewing system influence the resolving power. In practice, the minimum resolution is increased by a factor of two to three to take account of speckle. For off-axis holograms the choice of film size (effective aperture) has a considerable bearing on image resolution. Typically, for 100 mm diameter film the theoretical image resolution (neglecting speckle effects) is around 6 μm (150 lp/mm) with a 532 nm laser at a target distance of 1 m. For in-line holograms (with collimated illumination) the film size also defines the transverse dimensions of the volume recorded.

Wavelength matching between recording and reconstruction beams require that, if possible, the same type of laser be used in both replay and recording. Of course, this is not possible in field situations when the use of a pulsed laser is essential to ensure that the fringe pattern is not blurred by any movement in the system. In such applications, laboratory replay of the hologram may require the use of a tuneable laser to match wavelengths precisely, although in practice an argon laser provides a useful selection of wavelengths. A mismatch between record and replay wavelengths will lead to distortion in the image, because longitudinal and lateral magnifications will no longer have the same value.

In contrast to off-axis holograms, the interference patterns on an in-line hologram are localized about the geometric shadow of the objects. The lack of a spatially separate reference beam leads to a requirement that there be a large fraction (>80%) of the beam unobscured. This limits the size of particles that can be recorded and sets an upper limit on the recordable concentration of particles (around 40 particles cm^{-3} for 20 μm diameter particles and a recorded depth of 1 m). There is also a need to balance the size of the particles with the object-to-hologram distance. The range of object size recordable on an in-line hologram has an upper limit set by the requirement to record in the far-field of the object, so that the conditions of Fraunhofer recording are satisfied:

$$z > d^2/\lambda \tag{2}$$

where z is the object-to-film distance and d is the maximum dimension of the object to be recorded. In-line holograms are relatively easy to replay as the only requirement is to place the hologram normal to the replay beam. Following this criterion, the upper limit for good size measurement is about 2 to 3 mm particle size over a recording distance of a meter. Unlike off-axis holograms which are extremely sensitive to misalignment, the in-line geometry minimizes the effects of film rotation about its axes. A collimated recording beam is by far the easiest to duplicate upon replay. Small deviations from collimation can significantly affect conjugacy and hence resolution, introducing aberrations (mainly spherical) and a depth-dependent magnification (e.g., a cuboid replays as a trapezoidal prism).

In general, the interference field is captured on photographic emulsions; these may be laid on acetate or polyester film bases or on glass plates. If film is used, it is essential that it be held as flat as possible between thin flat glass plates or under vacuum, and should not be stretched or put under any strain during exposure. An in-line hologram does not require an extremely high-resolution recording medium. The fringe spacing is not as fine as the typical carrier fringe frequency used in off-axis holograms, and holograms can be made on conventional slow photographic emulsions. In practice, however, ultra-fine grained holographic film is almost always used since the intensity of the forward scattered light (a function of grain size) is then reduced: this is important in increasing the signal-to-noise ratio, as the replayed image is viewed against a background of the attenuated replay beam. Since the proportion of the illuminating beam that goes into the images is very low, this lack of angular separation requires that the average optical density of the hologram is high. The high optical density attenuates the undiffracted light, thus the image appears bright against a dark background, resulting in contrast reversal.

The exposed film is chemically processed to render the interference field permanent. This process is a crucial step in the holographic procedure, and especially critical for good off-axis holograms. Some of the factors, which have to be considered, include image brightness, image resolution, reconstruction wavelength, emulsion shrinkage, humidity, and noise level. It has been well established that for bright off-axis holograms on silver halide film, bleaching of the emulsion with its consequent conversion into a phase hologram is essential to maximize diffraction efficiency. Some forms of bleaching produce a nonuniform shrinkage of the emulsion which will give rise to astigmatism in the reconstructed image. Bleaching significantly increases the background noise levels of the hologram and introduces a high degree of variability into the result. In practice, the best repeatability will be obtained with amplitude processing if the low brightness can be tolerated.

Underwater Holography

When holography is extended to the inspection of underwater scenes, the hologram is recorded in water but the image is replayed, in the laboratory, in air. Because of the refractive index difference between recording and replay spaces, optical aberrations will be introduced into the reconstructed image, and will seriously impair the potential for precision measurement, unless some method of aberration reduction is incorporated. In ILH, only spherical aberration is significant since both reference and object beam angles are normal to the recording plane. However, in OAH the dominant aberrations are astigmatism and coma, which increase with the field angle of the subject in the reconstructed image. These limit resolution and introduce uncertainty in co-ordinate location, since the focal ‘point’ is now distributed over a finite region of image space. A key point here is an appreciation that the additional aberrations introduced are entirely due to the refractive index mismatch between object and image spaces and are unconnected with the holographic process itself: they are in essence the same distortions seen when looking into a fish-tank. Furthermore, the water itself may be expected to influence the quality of the images produced. An increase in the overall turbidity of the water will adversely affect both in-line and off-axis techniques and would be expected to create a background noise that will reduce image fidelity.

Techniques, such as immersing the holographic plate in water during recording or direct replay back into water, have been used to minimize underwater aberrations, but such methods are generally impractical in the field. One practical solution, unique to holography, compensates the change of effective wavelength of light as it passes through water, by allowing a deliberate mismatch of recording and replay reference beams to balance the refractive index mismatch.

For holograms recorded and replayed in air, a usual prerequisite is that the reconstruction wavelength should remain unchanged from that used in recording. By noting the dependence of wavelength on the refractive index, we can apply the more general condition that it is the ratio λ/n that must remain constant. This relationship suggests that a hologram, immersed in water, and recorded at a wavelength λ_a will produce an aberration-free image in air when replayed at a reconstruction wavelength, λ_c, which is itself equivalent to the wavelength of light when passing through water, λ_w, i.e.:

$$\lambda_c = \lambda_w = \lambda_a n_a / n_w \qquad [3]$$

where n_a and n_w are the respective refractive indices of air and water. Thus, in principle, the aberrations may be completely eliminated by the introduction of an appropriate wavelength change between recording and replay. If a green (532 nm) laser is used in recording, the ideal replay wavelength is about 400 nm (i.e., 532 nm/1.33). However, no convenient laser lines exist at exactly this wavelength. Furthermore, complete correction assumes that the entire recording system be located in water. Since this is both impractical and undesirable, holograms are usually recorded with the holographic film in air behind a planar glass window. The additional glass and air paths affect the compensation of aberrations. However, third-order aberration theory shows that if the window-to-air path length ratio is appropriately chosen for a specific replay wavelength, then aberration balancing occurs and residual aberrations are reduced to a minimum over a wide range of field angles and object locations. For an air gap of 120 mm and a BK7 window of 30 mm thickness, a blue replay wavelength of 442 nm (HeCd laser) achieves a good performance over a full field angle of about 40°.

In most underwater applications the objects (or the camera) are in motion during the exposure. The effect of this motion is to blur out the finer fringes, and thus reduce resolution and contrast. In-plane motion is the most severe, and adopting the experimentally verified criterion that the maximum allowable motion is less than one-tenth of the minimum required fringe spacing of the smallest object, then for in-line holograms the maximum object motion must be less than one-tenth of the object's dimension. For particles of 10 μm dimension, and a typical Q-switched YAG laser pulse duration of 20 ns, a high transverse velocity of up to 50 m s^{-1} can be tolerated. Off-axis holograms are more demanding in their requirements and the maximum allowed velocity is reduced to about 3 m s^{-1}. This is, however, more than adequate for most field applications of the technique.

Underwater Holographic Cameras

The first-known use of holography for recording living marine plankton was by Knox in 1966, who recorded in-line holograms (ILH), using a short coherence length ruby laser, of a variety of living marine plankton species in a tank. However, to avoid problems with refractive index and wavelength changes, the holographic plate was immersed directly in water, which is completely impractical for field use. It was not until the late 1970s that quantitative data from replayed holograms of marine particles recorded *in situ* were obtained by Carder and his team. Small volumes (10 or 20 cm^3) were recorded using ILH and a low-power HeNe laser, which limited the technique to slowly moving objects. Later, Heflinger used the off-axis geometry with both pulsed argon-ion (514 nm) and pulsed xenon (535 nm) lasers to record plankton on 35 mm film. In order to compensate for the aberrations introduced by replaying the real image in air, the real image was projected back into the water tank using a conjugate beam from a continuous-wave argon laser. The sides of their tank were flexible so that the real image could be placed close to the exit window and viewed via an external microscope. Knox later developed a ship-borne in-line holographic camera which was deployed from the *RV Ellen B Scripps*. A pulsed xenon laser was used to record water volumes of up to 10^5 cm^3 to a water depth of 100 m.

It was not until the late 1990s that significant use of underwater holography for plankton recording came to the fore again. Joseph Katz at Johns Hopkins University, Baltimore, USA developed a camera based around a ruby laser to record in-line holograms on rolls of film, in a 63 mm diameter sample volume that can be varied from 100 to 680 mm in length. The camera housing utilizes a two-tube design with one tube holding the laser and another containing film holder and power units. The recording volume is situated 890 mm above the camera to minimize disturbance to flow. The camera is designed as a 'Lagrangian drifter' to move with the water current and was successfully deployed in Chesapeake Bay, Baltimore, and Solomon's Island, Maryland. Holograms were recorded of various species of phytoplankton and zooplankton and represented a significant advance in the deployment of sub-sea holography.

From the early 1980s, a team led by Aberdeen University, Scotland, investigated the use of underwater holography for offshore applications. Primary interest was in inspection of subsea pipelines for corrosion and damage, and much of the work concentrated on an analysis of the optical aberrations introduced by recording in water and replaying in air and methods of compensating for such aberrations. However, this work was later adapted for holography of plankton and led to the consequent development of the HoloMar system. The HoloMar camera was the first to incorporate simultaneous ILH and OAH recording.

More recent developments have utilized digital or electronic recording of holograms in the in-line mode, and it is to be expected that much more use will be made of 'e-holography' in the future.

For field applications of holography, it is essential to record the holograms using a pulsed laser in order to create stable interference fringes. For subsea purposes, a laser with a wavelength in the blue-green region of the spectrum will match the oft-quoted peak transmission window of seawater. In practice, a frequency-doubled Nd-YAG laser with a wavelength of 532 nm is a good choice, although ruby lasers operating at 694 nm are also acceptable if transmission distance is not an issue. Usually OAH requires more energy and places stricter constraints on coherence properties than ILH. Beam quality is governed by quality of the optical components in the beam path and also by the output beam of the laser itself. Because off-axis holography uses diffuse front/side illumination, energies of up to several hundred millijoules are often required for large (up to a meter in diameter) poorly reflective objects, and coherence lengths of greater than a meter are needed to record large scene depths. To obtain a long coherence length, and thus enable large-volume objects to be recorded, it is necessary that the laser operate in a single longitudinal cavity mode so that the bandwidth of the laser line is kept small. To enable smooth even illumination of both subject and holographic emulsion, good spatial beam quality is needed. This usually dictates laser output beams with either a Gaussian (TEM_{00}) mode profile or a flat profile, and spatial filtering during beam expansion. Compared to off-axis holography, the demands placed upon lasers for ILH are minimal. The only essential requirement is for a good spatial mode structure. Energy requirements are moderate: as little as 10 mJ is enough since little light is lost during the passage of the beam through the recording volume. The laser coherence length can be quite short, and only depends on the

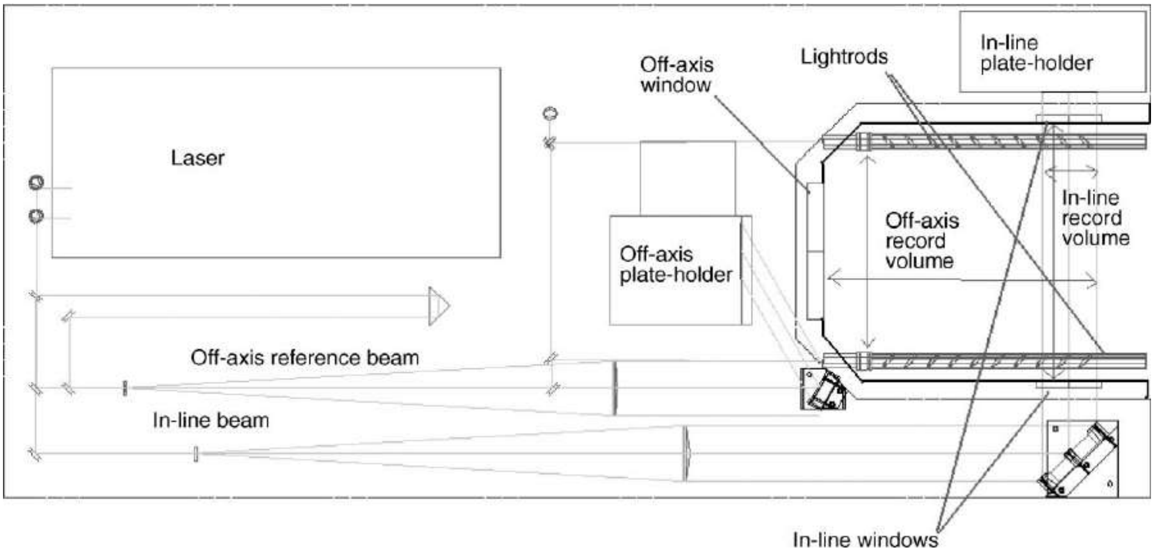

Figure 7 The HoloMar camera configuration showing the separate in-line and off-axis beam paths.

effective hologram area, thus lasers operating on more than one longitudinal mode are usually quite suitable.

The HoloMar System

The HoloMar system (shown diagrammatically in Figure 7) was developed to record simultaneous in-line and off-axis holograms of partially overlapping, but orthogonal, views of the water column. Although the recording views are orthogonal to each other, the arrangement provides an element of cross-correlation between holograms. It is designed around a Q-switched, frequency-doubled Nd-YAG laser with an output energy of 700 mJ in a single pulse of less than 10 ns duration and a coherence length in excess of 2 m. The laser output is split into two beams: one of 100 mJ and the other of 600 mJ energy. The 100 mJ beam is further split into two equal energy beams and expanded and collimated to a diameter of about 90 mm. One path forms the illuminating beam for the in-line mode and the other path forms the reference beam for the off-axis mode. The in-line path records a water column of 470 mm by 90 mm diameter (9,000 cm^3) at 400 mm from the front face of the housing, whereas the off-axis hologram records almost the entire volume (50,000 cm^3) delineated by the front window to the end of the arms that house the in-line system (see Figure 7). Three 'lightrods', encompassing hollow perspex tubes with 10 glass beam-splitter plates distributed equally along the length, positioned at 45° to the main axis, provide roughly even side illumination to the off-axis volume. The front face of the camera is shown in Figure 8; the lightrods are positioned on either side of the recording volume (two on one side, one on the other). The exit window for the in-line beam path and the off-axis window can also be seen.

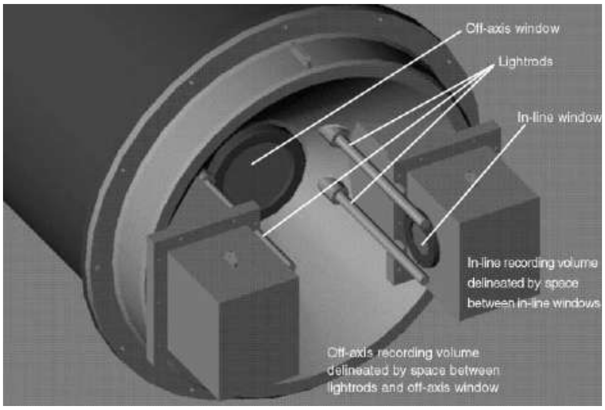
Off-axis window
Lightrods
In-line window
In-line recording volume delineated by space between in-line windows
Off-axis recording volume delineated by space between lightrods and off-axis window

Figure 8 Front-end of the HoloMar camera showing the 'lightrods' and in-line and off-axis windows.

Two motor-driven plate holders and transport mechanisms hold up to 20 plates for in-line and 24 for off-axis holograms. One pair of plates can be exposed every 10 seconds, or they can be exposed separately. The camera is remotely operated from the surface via a control console (PC). A network of micro-controllers serves as the backbone of the in-camera control system and communication between topside and camera modules is implemented using a Controller Area Network (CAN) bus via an umbilical cable. The entire camera is enclosed in a water-tight stainless steel housing of 1,000 mm diameter by 2,200 mm length. The prototype system is capable of either ship deployment or attachment to a fixed buoy and allows recording down to a water depth of 100 m.

The HoloMar camera is shown, in Figure 9, being lowered into Loch Etive, a sea loch on the west coast of Scotland, on its maiden dive. Figures 10 and 11 show holographic images of plankton recorded on this cruise. The first shows a view from a virtual

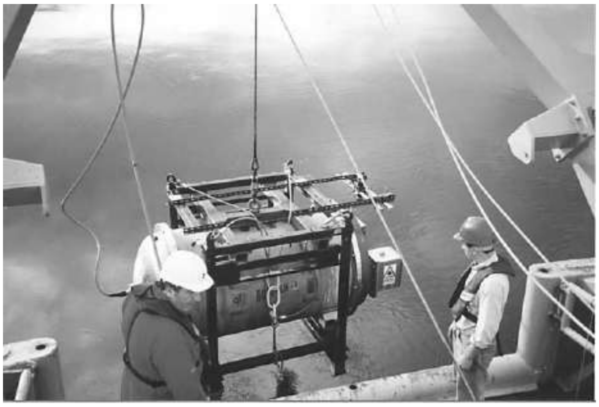

Figure 9 The HoloMar camera being lowered into Loch Etive, Scotland.

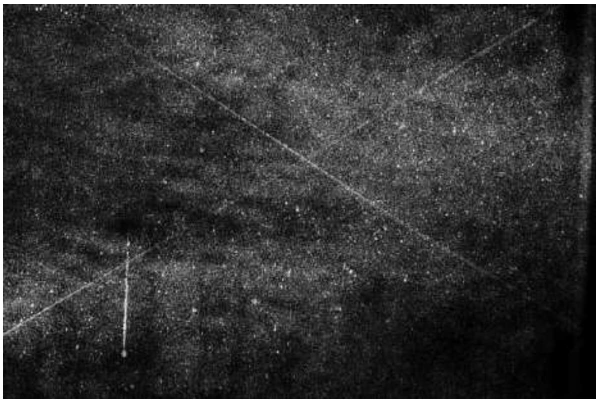

Figure 10 A photograph taken from a virtual image replay of an off-axis hologram recorded in Loch Etive.

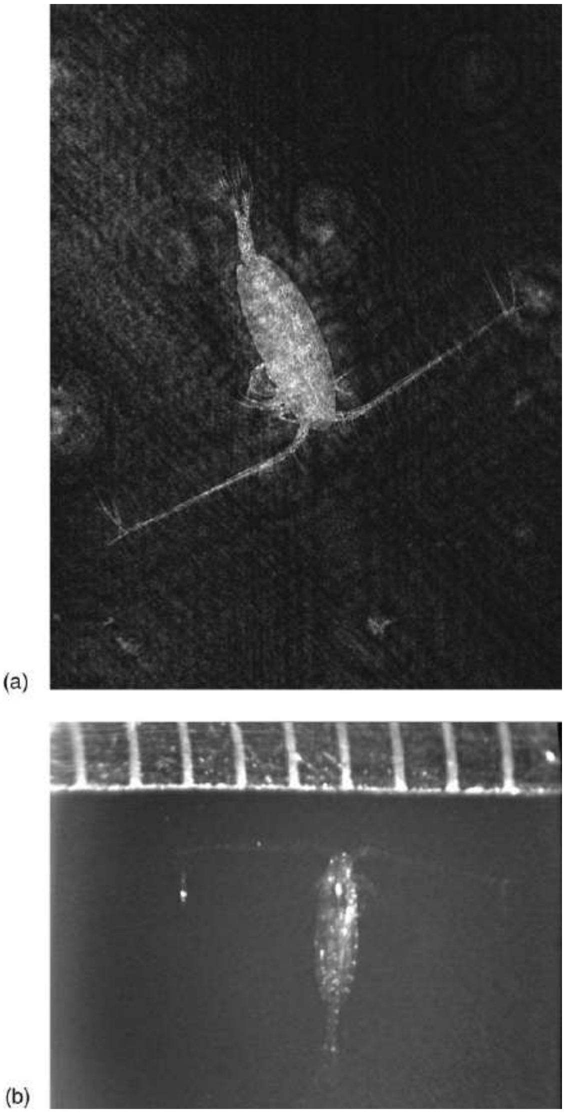

Figure 11 Images of calanoid copepods from (a) an in-line hologram recorded at 70 m depth and (b) an off-axis hologram recorded at 60 m depth, the millimeter scale is added at the replay stage. Each organism is approximately 4 mm long.

replay of an off-axis hologram at a wavelength of 514 nm: this is what a viewer would see when looking from the inside of the holocamera into the laser-illuminated water. Each bright dot signifies an object of interest; two fiducial wires can also be seen fixed diagonally across the arms at the end of the camera, some 300 mm from the off-axis window. In the lower left of the image a long 'string-like' object about 40 mm long is probably 'floc', the dead organic matter that forms the food-stuff of many zooplankton. Objects like this would be destroyed in net-collection. Figure 11 shows images taken from the real image projection of in-line and off-axis holograms. The holograms are replayed using a collimated beam at wavelengths of 514, 488, 457, or 442 nm, depending on location in the image volume. Some of the contrasting features between in-line and off-axis holograms are apparent: the in-line image shows clearly the outline of the organism with detail resolvable down to about 10 μm, whereas the off-axis hologram, although not achieving this resolution, reveals the body structure of the organism.

List of Units and Nomenclature

d	dimension of object to be recorded
D	effective aperture of hologram
n	refractive index
n_a	refractive index of air
n_w	refractive index of water
r	minimum resolvable separation between two points on the hologram
s	distance between hologram and image location
z	distance between hologram and object location
λ	wavelength
λ_a	in-air wavelength of recording beam
λ_c	reconstruction wavelength
λ_w	effective wavelength of recording beam in water

See also

Holography, Applications: Holographic Recording Materials and Their Processing. **Holography, Techniques:** Computer-Generated Holograms. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Beers JR, Knox C and Strickland JDH (1970) A permanent record of plankton samples using holography. *Limnology and Oceanography* 15: 967–970.

Carder KL (1979) Holographic microvelocimeter for use in studying ocean particle dynamics. *Optical Engineering* 18: 524–525.

Hariharan P (1984) *Optical Holography*, 2nd edn. Cambridge, UK: Cambridge University Press.

Heflinger LO, Stewart GL and Booth CR (1978) Holographic motion pictures of microscopic plankton. *Applied Optics* 17: 951–954.

Hobson PR and Watson J (2002) The principles and practice of holographic recording of plankton. *Journal of Optics A: Pure and Applied Optics* 4: S34–S49.

Katz J, Donathay PL, Zhang J, King S and Russell K (1999) Submersible holocamera for detection of particle characteristics and motions in the ocean. *Deep-Sea Research* 46: 1455–1481.

Kilpatrick JM and Watson J (1993) Underwater hologrammetry: reduction of aberrations by index compensation. *Journal of Physics D: Applied Physics* 26: 177–182.

Knox C (1966) Holographic microscopy as a technique for recording dynamic microscopic subjects. *Science* 153: 989–990.

Malkiel E, Alquaddoomi O and Katz J (1999) Measurements of plankton distribution in the ocean submersible holography. *Measurement Science and Technology* 10: 1141–1152.

Owen RB and Zozulya AA (2000) In-line digital holographic sensor for monitoring and characterizing marine particulates. *Optical Engineering* 39: 2187–2197.

Stewart GL, Beers JR and Knox C (1970) Application of holographic techniques to the study of marine plankton in the field and the laboratory. *Proceedings of SPIE* 41: 183–188.

Sun H, Dong H, Player MA, *et al.* (2002) In-line digital holography for the study of erosion processes in sediments. *Measurement Science and Technology* 13: L1–L6.

Watson J, Alexander S, Craig G, *et al.* (2001) Simultaneous in-line and off-axis subsea holographic recording of plankton and other marine particles. *Measurement Science and Technology* 12: L9–L15.

Holographic Recording Materials and Their Processing

H I Bjelkhagen, De Montfort University, Leicester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The first off-axis transmission holograms in the USA were recorded on Kodak 649F spectroscopic silver halide material (now discontinued). For the first reflection holograms, which were recorded in Russia, a very fine-grain emulsion (Lippmann emulsion) had to be prepared to resolve the fine interference pattern associated with such holograms. Since then various types of recording materials for holography have been developed and manufactured, including improved silver halide emulsions.

The following provides a brief description of the holographic recording process. For recording holograms and holographic optical elements (HOEs), dichromated gelatin, photopolymer materials, and photoresist materials are employed, in addition to silver halide materials. Existing holographic recording materials are described, as well as suitable processing methods to obtain high-quality holograms and HOEs.

Holographic Recording Materials

To produce a hologram, a micropattern, caused by interference between the object beam and the reference beam, is recorded in a light-sensitive high-resolution material. The quality of a holographic image depends on a number of factors, such as: the geometry and stability of the recording setup; the coherence of the laser light; the reference and object beam ratio; the type of hologram produced; the size of the object and its distance from the recording material; the recording material and the emulsion substrate used; the processing technique applied as well as the reconstruction conditions. The material must comply with certain requirements to be suitable for this purpose. The most important of these concerns the resolving power of the material. The recording material must be able to resolve the highest spatial frequencies of the interference pattern created by the maximal angle θ between the reference and the object beams in the recording setup (Figure 1).

If λ is the wavelength of the laser light used for the recording of a hologram and n the refractive index of the emulsion, then the closest separation d_e between the fringes in the interference pattern created by the angle θ between the reference and the object beams in the recording setup is

$$d_e = \frac{\lambda}{2n \sin(\theta/2)} \qquad [1]$$

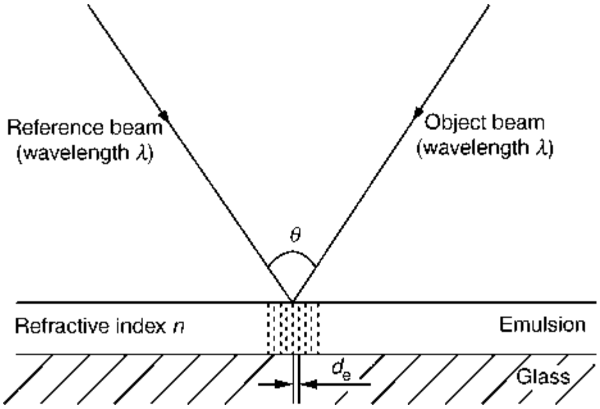
Reference beam (wavelength λ)
Object beam (wavelength λ)
θ
Refractive index n
Emulsion
d_e
Glass

Figure 1 Demand on resolution for recording a hologram.

For high-quality holograms the resolution limit of the material must be higher than the minimum value obtained according to the above formula. For a reflection hologram recorded in blue light ($\lambda = 400\,\text{nm}$) with an angle of 180° between the beams, a minimum resolving power of 7600 lines $(\text{mm})^{-1}$ is required.

In holography, the resolution of the holographic image and the resolving power of the recording material are not directly related to the way in which they are in photography.

Equation [1] gives the minimum resolving power needed to record a hologram. The resolution of the image depends on the area (diameter D) of the recording material, the recording laser wavelength (λ), and the distance (L) between the recording material and the object (Figure 2).

Theoretically, the resolution of the holographic image is the diffraction-limited resolution that can be obtained when the information is collected over an aperture equal to the size of the recording holographic plate. In principle, the larger the holographic plate the better the resolution will be. To obtain high image resolution, it is important that the recorded interference fringes will not change or be distorted during processing. To prevent that from happening, a stable support for the emulsion (like a glass plate) is needed and the processing method applied must not affect the recorded fringe pattern in the emulsion. In coherent imaging systems laser speckles are present which may also have an effect upon image resolution.

There are two main types of holograms: amplitude and phase holograms. In a pure amplitude hologram only the absorption varies with the exposure (after processing), whereas in a pure phase hologram, either the refractive index or the emulsion thickness varies with the exposure. If the hologram is thin, $d \approx 0$, phase variations are then caused by surface relief variations only. If the hologram is thick and has a negligible surface relief ($\Delta d = 0$), phase variations are caused by index variations only. In many cases, the phase modulation in a hologram is a combination of the two different extreme types (a complex hologram).

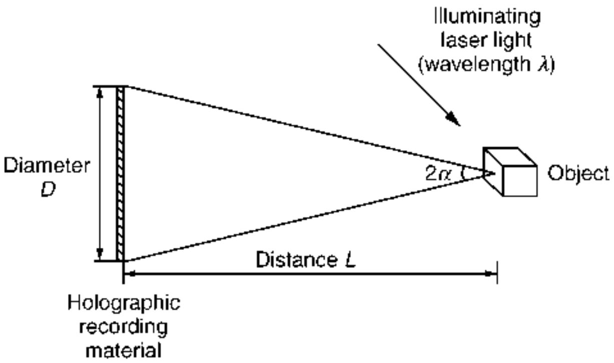

Figure 2 Resolution of a holographic image depends on the area (diameter D) of the recording material, the recording laser wavelength λ, and the distance L between the recording material and the object.

For amplitude holograms, amplitude transmission T_a against exposure or log exposure is used (Figure 3). For a phase hologram, the corresponding curve is the phase shift against the log exposure relation (Figure 4).

One of the most important hologram characteristics is its diffraction efficiency η. It is defined as the diffracted intensity E_i of the wanted diffraction order of the hologram in relation to the incident

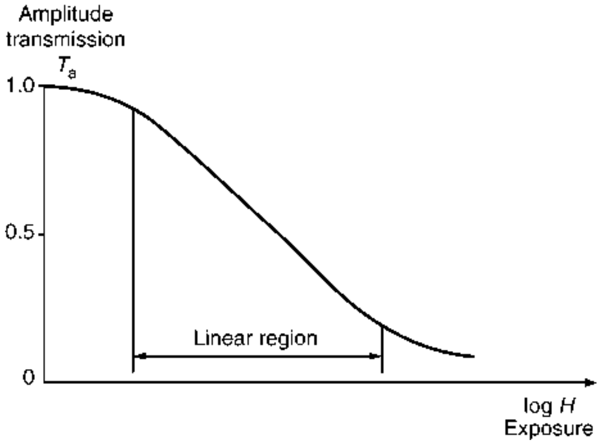

Figure 3 Amplitude transmission T_a versus the logarithm of exposure H used for the characterization of amplitude transmission holograms.

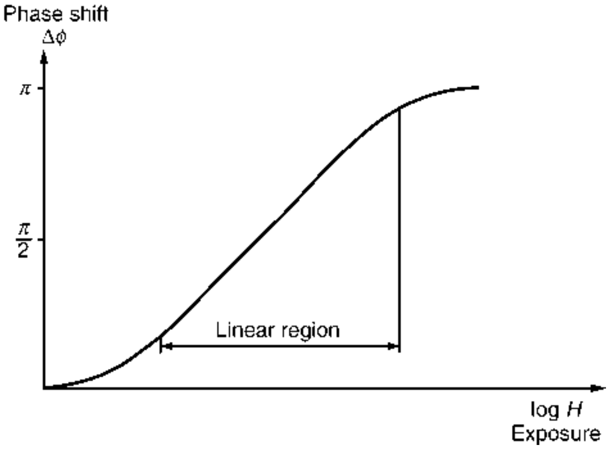

Figure 4 Phase shift $\Delta\phi$ versus the logarithm of exposure H used for the characterization of phase transmission holograms.

intensity of the reconstruction beam E_r : $\eta = E_i/E_r$. In the definition of η, the incidental light losses caused by surface reflection and base absorption should be subtracted, but are often ignored. Other important hologram characteristics are the signal-to-noise ratio (SNR) and the dynamic range of the recording material.

The resolution capability of an image reproduction process is normally described by the Modulation Transfer Function (MTF). Briefly, a test pattern containing a sinusoidal variation in illuminance combined with a continuous variation in spatial frequency along one direction is recorded. The modulation M of the pattern in the test target is

$$M = (H_{\mathrm{MAX}} - H_{\mathrm{MIN}})(\mathrm{H}_{MAX} + \mathrm{H}_{MIN})^{-1} \qquad [2]$$

where **H** is the exposure incident on the photographic material. When this pattern is recorded in the material, light scattering will take place in the emulsion, which will reduce the original contrast of the pattern. Therefore, the modulation of the pattern will be decreased, in particular at high spatial frequencies. The effective exposure modulation M' will then be

$$\mathrm{M}' = (H'_{\mathrm{MAX}} - H'_{\mathrm{MIN}})(H'_{\mathrm{MAX}} + H'_{\mathrm{MIN}})^{-1} \qquad [3]$$

where **H**′ is exposure in the emulsion.

The original modulation M is constant and accurately known; it is also independent of the spatial frequency. After the tested emulsion has been processed, the corresponding 'exposed' modulation is obtained from the density variations. The ratio between modulation M' in the emulsion and modulation M of the incident exposure is called the modulation transfer factor, also called response

$$R = M(M')^{-1} \qquad [4]$$

If the response is plotted as the function of spatial frequency, this curve will then be the modulation transfer function of the material. The MTF for both a holographic emulsion and a conventional photographic emulsion are illustrated in Figure 5.

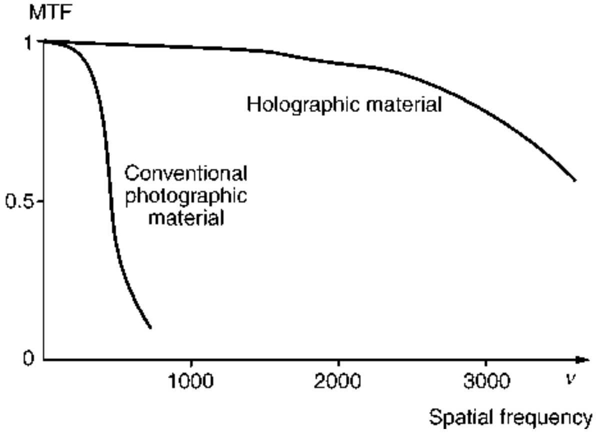

Figure 5 Graphical representation of the modulation transfer function (MTF) curves for both a conventional emulsion and a holographic emulsion are shown.

There are many ways of categorizing holograms but here we are mainly interested in definitions related to the recording material and its processing. One of the valid criteria is the thickness of the recording layer (as compared to the interference fringe spacing within the layer), i.e., the layer coated on the material substrate. Holograms can thus be classified into 'thin' or 'thick' (sometimes also called 'plane' or 'volume' holograms, respectively). To distinguish between the two types, the Q-parameter is normally used; it is defined in the following way:

$$Q = 2\pi\lambda d(n\Lambda^2)^{-1} \qquad [5]$$

where λ is the wavelength of the illuminating light, d is the thickness of the layer, n is the refractive index of the emulsion, and Λ the spacing between the recorded fringes. A hologram is considered thick if $Q \geq 10$, and thin if $Q \leq 1$. Holograms with Q-values between 1 and 10 can be treated as either thin or thick.

Recording Materials

The most common holographic materials used for recording holograms are the following: silver halide emulsions; dichromated gelatin emulsions; photopolymer materials; photoresist materials; thermoplastic materials; bacteriorhodopsin; and photorefractive crystals.

The following materials can also be considered for holographic recordings, but not many practical applications have been reported so far: chalcogenide glasses; ferroelectric-photoconductors; liquid crystals; magneto-optic films; metal and organic dye ablative films; photochromic and photodichroic materials; photothermorefractive glass; and transparent electrophotographic films.

Sensitivity of Photographic and Holographic Materials

The time of exposure for a given material will depend on the sensitivity of the material used, as well as on the intensity of the interference pattern. Some holographic materials must be processed after exposure in a specific way to obtain a hologram. The recorded intensity variations are converted during this processing step to local variations in optical density or refractive index and/or thickness of the recording layer.

Exposure H is defined as the incident intensity E times the time t of exposure of the recording material. If the intensity is constant during the whole exposure time, which is usually the case, then

$$H = Et \qquad [6]$$

Holographic materials are usually characterized using radiometric units. The radiometric equivalent of illuminance is irradiance. The unit of irradiance is $\mathrm{W(m)^{-2}}$ and the exposure will then be expressed in $\mathrm{J(m)^{-2}}$. The sensitivity of a holographic emulsion is most often expressed in $\mu\mathrm{J(cm)^{-2}}$ or $\mathrm{mJ(cm)^{-2}}$. Knowing the sensitivity of the material used and having measured the irradiance at the position of the holographic plate, the exposure time can be calculated using the above formula, i.e.,

$$\text{Exposure time} = \text{sensitivity/irradiance}$$

Holographic materials are sensitized in such a way that they are optimized for laser wavelengths commonly used in holography.

Silver Halide Materials

A silver halide recording photographic material is based on one type, or a combination of silver halide crystals embedded in a gelatin layer. The emulsion is coated on a flexible or stable substrate material. Silver-halide grain sizes vary from about 10 nanometers for the ultra-fine-grain holographic emulsions to a few micrometers for highly sensitive photographic emulsions (Table 1).

Silver salts are only sensitive to UV light, violet, and deep blue light. Therefore special sensitizers (dyes) must be added to the emulsion to make it sensitive to other parts of the spectrum. Orthochromatic emulsions are sensitive to green light. If the material has been sensitized to both green and red light, it is said to be panchromatic.

Table 1 Silver halide emulsion grain sizes

Type of emulsion	*Average grain diameter [nm]*
Ultra-fine-grain holographic emulsion	10–30
Fine-grain holographic emulsion	30–50
Fast holographic emulsion	50–100
Chlorobromide paper emulsion	200
Lithographic emulsion	200–350
Fine-grained photographic emulsion	350–700
Fast photographic emulsion	1000–2000
Fast medical X-ray emulsion	2500

In the following, silver halide materials manufactured by photographic and holographic companies are described. The main difference between the new materials and the previous emulsions from Agfa, Ilford, etc., is that for the current silver halide emulsions, the grain sizes are smaller. The present holographic silver halide emulsion manufacturing process is more or less based on the Russian ultrafine-grain emulsion technology.

Substrates for Holographic Emulsions

The material on which the emulsion is coated has a strong bearing on the final quality of the hologram. The best choice is often a glass plate as it is mechanically stable and optically inactive. High resolution imaging, hologram interferometry, holographic optical elements (HOEs) and spatial filters are a few examples where a very stable emulsion support is important. Mass produced holograms are mainly recorded on film substrates. However, the use of film has many advantages as compared to that of glass (breakage, weight, cost, size, etc.). Film substrates are used exclusively in the production of large-format holograms. Film substrates are of mainly two types: a polyester (polyethylene terephthalate); or a cellulose ester, commonly triacetate (cellulose triacetate) or acetate–butyrate. Polyester is mechanically more stable than triacetate film and it is also less sensitive to humidity. Because of the higher tensile strength, the polyester film can be made thinner than the triacetate film. On the other hand, polyester is birefringent, which can cause problems when recording holograms. In Table 2 most current commercial silver halide materials for holography are listed.

Processing of Silver Halide Emulsions

In silver halide emulsions a latent image is formed during exposure, which later is developed to a silver image. Processing (development, fixing, and bleaching) silver halide emulsions requires some expertise to obtain high-quality holograms. With the exception of DCG emulsions, other holographic materials are easier to process and more straightforward. Here it is not possible to describe all techniques employed and recipes used for silver halide processing. The reader is referred to the Further Reading section for more details.

Table 2 Commercial holographic recording materials

Material	*Thickness [μm]*	*Spectral sensitivity [nm]*	*Sensitivity [μJ (cm)$^{-2}$] at*				*Resolv. power [lp (mm)$^{-1}$]*	*Grain size [nm]*
			442	*514*	*663*	*694*		
SILVER HALIDE EMULSIONS:								
Slavich								
Red PFG-01	7	<700	–	–	80	–	>3000	35–40
Red PFG-03M	7	<700	–	–	1500	–	>5000	10–20
Green VRP-M	7	<550	–	80	–	–	>3000	35–40
Pan PFG-03C	9	400–700	1000	2000	1000	–	>5000	10–20
Colourholographic								
Red BB-700	7	<700	–	–	50	150	>2500	50–60
Red BB-640	7	<650	–	–	150	–	>4000	20–25
Green BB-520	7	<540	150	150	–	–	>4000	20–25
Blue BB-450	7	<470	150	–	–	–	>4000	20–25
Kodak								
Red 131PX	9	<650	2	–	0.5	–	>1250	70
Red 131CX	9	<650	2	–	0.5	–	>1250	70
Red 120PX	6	<750	60	–	40	40	>2500	58
Red 120CX	6	<750	60	–	40	40	>2500	50
FilmoTec-ORWO								
Red HF65	10	<650	–	–	1500	–	5000	30
Green HF53	10	<550	–	700	–	–	5000	30
Ultimate								
Ultimate 15	7	<700	–	150	150	150	>5000	15
Ultimate 08	7	<650	120	200	200	–	>7000	8
DICHROMATED GELATIN EMULSIONS								
Slavich								
Blue PFG-04	16	<515	1×10^5	2.5×10^5	–	–	10000	–
FilmoTec-ORWO								
Blue GF 40	6/20	*	*	–	–	–	–	–
* delivered: non-sensitized								
Holotec								
Custom orders only								
THERMOPLASTIC MATERIALS								
Tavex America								
Pan TCC-2	–	<800	1	1	1	1	1500	–
MD Diffusion								
Custom orders only								
PHOTORESIST MATERIALS								
Towne Technologies								
UV-Blue Shipley 1800	1.5–2.4	<450	1.5×10^5	–	–	–	1000	–
Hoya Corporation								
Custom orders only								
BACTERIORHODOPSIN MATERIALS								
MIB GmbH								
BR-WT B-type	30–100	<650	–	80×10^3	–	–	5000	–
BR-D96N M-type	30–100	<650	30×10^3	–	–	–	5000	–

The developing agent is the most important constituent in a holographic developer. In addition to the developing agent, a developer consists of the following components: a preservative (or antioxidant); a weak silver-solvent agent; an accelerator (or activator); a restrainer; and a solvent which is usually water. Employing a developer which contains a silver-solvent agent, solution-physical development or semi-physical development can be performed. Such a developer is particularly suitable for processing the ultrafine-grain emulsions.

Holographic Developers

Silver halide materials for holography are all of the fine-grain type. Such materials require the application of special processing techniques to achieve the best possible results. For the processing of amplitude holograms a high-contrast developer is required, such as Kodak D-19. For holograms which are going to be bleached (phase holograms), developers based on ascorbic acid, hydroquinone, metol, pyrocatechol, or pyrogallol are frequently used. The following developers are recommended for the new silver halide materials: D-19; CW-C2; AAC; MAA-3; PAAP; and for pulsed laser recordings, the SM-6 developer. Colloidal development, employing a highly diluted developer which contains silver solvents, is recommended for the ultrafine-grain emulsions. Colloidal silver grains formed in the emulsion during such development are very small which means that very little absorption occurs. The method produces holograms with high diffraction efficiency combined with very low light scattering noise. The Russian GP-2 developer is recommended for colloidal development. Recipes for common holographic developers are found in Table 3.

Bleach Baths

Bleaching is used to convert a developed hologram (containing a silver image) into a phase hologram, to obtain a higher diffraction efficiency than is possible for amplitude holograms. During bleaching the silver image is converted into a water-soluble or an insoluble silver compound. Holographic bleaching techniques can be divided into three categories:

1. Conventional or direct (rehalogenating) bleaching;
2. Fixation-free rehalogenating bleaching;
3. Reversal (complementary) or solvent bleaching.

In conventional bleaching, the developed silver image is, after fixing, converted into a transparent silver halide compound, i.e., after the unexposed silver halide crystals have been removed. This method is not recommended for ultrafine-grain emulsions. Instead the fixation-free rehalogenating bleaching is preferred. The hologram is bleached directly after development, without fixing, leaving the unexposed silver halide crystals in the emulsion. A rehalogenating bleach bath consists of: an oxidizing agent; an alkali halide (often KBr); and a buffer. Among the rehalogenating bleaches, the following bleach baths work well: PBQ-bleach; Ferric EDTA; and the PBU-bleaches (PBU-amidol or the PBU-metol).

In reversal bleaching, the developed silver image is not fixed. Instead it is converted into a soluble silver complex which is removed from the emulsion during bleaching, leaving only the original unexposed silver halide grains in the emulsion. A reversal bleach bath consists of an oxidizing agent and a buffer. For ultrafine-grain silver halide emulsions, the reversal method is not recommended. The very small unexposed silver halide grains are also affected (partly dissolved) during reversal bleaching, resulting in low diffraction efficiency. For materials with larger grains, a more diluted version of the dichromate Kodak R-9 bleach can be

Table 3 Developers for silver halide emulsions

Constituents	*CWC2*	*GP-2*	*MAA-3*	*PAAP*	*AAC*	*SM6*
Methylphenidone		0.2 g				
Catechol	10 g					
Ascorbic acid	5 g		10 g	18 g	16 g	18 g
Hydroquinone		5 g				
Metol			2.5 g			
Phenidone				0.5 g		6 g
Sodium sulfite	5 g	100 g				
Sodium phosphate (diabasic)				28.4 g		28.4 g
Urea	50 g					
Potassium hydroxide		5 g				
Sodium hydroxide				12 g		12 g
Sodium carbonate	30 g		55.6 g		60 g	
Ammonium thiocyanate		12 g				
Distilled water	1 l	1 l	1 l	1 l	1 l	1 l
Dilution:	no	15 ml dev. + 400 ml water	no	no	no	no

Table 4 Bleach baths for silver halide emulsions

Constituents	*Ferric-EDTA (rehal.)*	*PBQ (rehal.)*	*PBU-(amidol/metol) (rehal.)*	*Copper (rehal.)*	*Chrome (reversal)*	*PSSB (reversal)*
Amidol *or* metol*			1 g			
Cupric bromide			1 g			
Potassium persulfate			10 g			20 g
Potassium dichromate					4 g	
p-Benzoquinone		2 g				
Ferric III sulfate	30 g					30 g
Copper sulfate				35 g		
Disodium EDTA	30 g					
Boric acid		1.5 g				
Citric acid			50 g			
Potassium bromide	30 g	30 g	20 g	30 g		
Sodium hydrogen sulfate	30 g			30 g		30 g
Sulfuric acid					4 ml	
Distilled water	1 l	1 l	1 l	1 l	1 l	1 l

*To be added after all other constituents have been dissolved.

used. Recipes for common holographic bleach baths are found in Table 4.

Commercial Silver Halide Emulsions

In Table 2 most of the current silver halide holographic emulsions are listed. The main manufacturer of holographic silver halide emulsions is the Micron branch of the SLAVICH Joint Stock Company photographic company located outside Moscow (2 pl. Mendeleeva, 152140 Pereslavl-Zalessky, Russia, www.slavich.com). Both film (triacetate) and glass plates are manufactured. Many different sizes are produced, including 115 cm by 10 m film rolls. Recommended holographic developers: CW-C2, SM-6. Many rehalogenating holographic bleach baths work well and, in particular, the PBU-type bleaches are recommended.

Colourholographic Ltd is a new company based in England (Braxted Park, Great Braxted, Witham CM8 3BX, England; www.colourholographic.com). The Colourholographic materials are based on the HRT emulsion, previously manufactured in Germany by Dr Birenheide. At the present time only glass plates are manufactured in England. Recommended processing solutions for the new HRT materials are a sodium-sulfite-free ascorbic acid–metol developer followed by a rehalogenating ferric EDTA or PBU bleaches. For reflection holograms, a pyrogallol developer and the ferric-EDTA bleach can be used.

Eastman Kodak's holographic materials are produced in America (343 State Street, Rochester, New York 14650, USA; www.kodak.com/go/PCB products). These plates are usually made to order only. Materials can be ordered with (marked -01) or without (marked -02) anti-halation layer. The 120 emulsion can be used for recording holograms using a pulsed ruby laser. Kodak recommends the developer Kodak D-19 for the processing.

FilmoTec is located in Germany (Chemie Park Bitterfeld-Wolfen, Röntgenstrasse, D-06766 Wolfen, Germany; www.filmotec.de). The company manufactures ORWO holographic emulsions which are coated only on triacetate film, 104 cm wide in 10 and 30 m lengths.

In France, M. Yves Gentet has started a small-scale emulsion manufacturing company: Atelier de Création d'Art en Holographie Yves Gentet (50, rue Dubourdieu, F-33800 Bordeaux, France; www.perso.wanadoo.fr/holographie) Monochromatic and panchromatic ultrafine-grain emulsions are produced called 'Ultimate'. Plates and film up to 60 cm by 80 cm are produced. An ascorbic acid–metol developer is recommended followed by a Ferric EDTA bleach. In addition, the Russian GP-2 developer can be used. Also a special developer can be ordered from the company.

Konica in Japan is manufacturing a green-sensitive emulsion for the use in the American holographic VOXEL 3D medical imaging system and, in addition, for the Japanese market only.

Dichromated Gelatin Materials

Dichromated gelatin (DCG) is an excellent recording material for volume phase holograms and HOEs. The grainless material has its highest sensitivity in the UV region, but extends into the blue and green parts of the spectrum. There are possibilities to sensitize the emulsion in the red part of the spectrum as well. However, DCG is most often exposed with blue laser wavelengths. Depending on processing parameters, diffraction efficiency and bandwidth can be controlled. It is easy to obtain high diffraction efficiency combined with a large SNR.

During the exposure of a DCG emulsion to UV or blue light, the hexavalent chromium ion (Cr^{6+}) is photo-induced to trivalent chromium ion (Cr^{3+}) which causes cross-linking between neighboring gelatin molecules. The areas exposed to light are hardened and become less soluble than the unexposed areas. Developing consists of a water wash which removes the residual or unreacted chemical compounds. Dehydration of the swollen gelatin follows after the material has been immersed in isopropanol, which causes rapid shrinkage resulting in voids and cracks in the emulsion, thus creating a large refractive index modulation. The underlying mechanism is not completely clear because high modulation can also be caused by the binding of isopropanol molecules to chromium atoms at the cross-linked sites. The DCG material has a rather low sensitivity of about $100\,\mathrm{mJ(cm)^{-2}}$. In many cases DCG plates are prepared by fixing an unexposed silver halide emulsion and then sensitizing it in a dichromate solution. It is also, for example, possible to make a 10 μm thick emulsion on a 8″ by 10″ glass plate using 1 g ammonium dichromate mixed with 3 g photographic grade gelatin in 25 ml deionized water. Then the emulsion is spin-coated on the glass substrate or the emulsion is applied by the doctor blade coating technique. Many HOE manufacturers prefer to produce their own emulsions but there are also commercial DCG emulsions available on the market.

Processing of DCG holograms are performed in graded alcohol (propanol) baths starting with water–propanol solutions of high water content and ending with pure propanol (cold or hot) baths. Depending on the wanted HOE characteristics, one has to carefully control the propanol/water mixtures as well as the temperature of the different baths. Cold baths produce better uniformity and lower noise. Warm baths can yield high index modulation but often with increased noise. The bandwidth can be controlled by the processing temperature and the ratio of propanol/water mixtures.

Commercial DCG Materials

Slavich in Russia is one manufacturer of presensitized dichromated plates for holography. The DCG emulsion is marked PFG-04. Plates up to a size of 30 by 40 cm can be ordered.

The FilmoTec DCG emulsions are only available on 190 μm triacetate film, 104 cm wide in 10 and 30 m lengths; emulsion thickness 6 or 20 μm. Note that the film needs to be sensitized in dichromate solution before recording. This means that the company only supplies large-format gelatin-coated film.

Holotec is a German-based company which offers presensitized DCG coated on both plates and film (Jülicher Strasse 191, D-52070 Aachen, Germany; www.holotec.de). The Holotec emulsion is the high-quality DCG emulsions developed by Professor Stojanoff in Aachen. The company can supply large-format DCG glass plates (m^2 size) or film (PET-Polyethylenterephtalat), pre-sensitized and ready to use.

Photopolymer Materials

Photopolymer materials have become popular for recording phase holograms and HOEs, in particular for mass production of holograms since some photopolymer materials only require dry processing techniques.

A photopolymer recording material, such as the DuPont material, consists of three parts: a photopolymerizable monomer; an initiator system (initiates polymerization upon exposure to light); and a polymer (the binder). First, an exposure is made to the information-carrying interference pattern. This exposure polymerizes a part of the monomer. Monomer concentration gradients, formed by variation in the amount of polymerization due to the variation in exposures, give rise to diffusion of monomer molecules from the regions of high concentration to the regions of lower concentration. The material is then exposed to regular light of uniform intensity until the remaining monomer is polymerized. A difference in the refractive index within the material is obtained. The DuPont material requires only a dry processing technique (exposure to UV light and a heat treatment) to obtain a hologram. The DuPont photopolymer materials have a coated film layer thickness of about 20 μm. The photopolymer film is generally coated in a 12.5″ width on a 14″ wide Mylar® polyester base which is .002″ thick. The film is protected with a .00092″ thick Mylar® polyester cover sheet.

The recording of a hologram on DuPont polymer is rather simple. The film has to be laminated to a piece of clean glass or attached to a glass plate using an index-matching liquid. Holograms can be recorded manually, but in order to produce large quantities of holograms, a special machine is required. For hologram replication a laser line scanning technique can provide the highest production rate.

The photopolymer material needs an exposure of about 10 mJ(cm)$^{-2}$.

After the exposure is finished, the film has to be exposed to strong white or UV light. DuPont recommends about 100 mJ(cm)$^{-2}$ exposure at 350–380 nm. After that, the hologram is put in an oven at a temperature of 120 °C for two hours in order to increase the brightness of the image. The process is simple and very suitable for machine processing using, for example, a baking scroll oven.

Commercial Pholymer Materials

Currently the only manufacturer of holographic photopolymer materials is E.I. du Pont de Nemours & Co. (DuPont Holographics, POB 80352, Wilmington, DE 19880, USA; www.dupont.com/holographics). In the past it was possible to obtain photopolymer materials from DuPont, but in 2002 the company changed its policy and today it is producing materials mainly for DuPont Holographics, but approved customers may still be able to buy material from DuPont. The reason for the restrictions is that many applications of mass-produced holograms are in the field of document security, where DuPont's photopolymers are used to produce optical variable devices (OVDs).

Photoresist Materials

Photoresist materials in holography are used for making masters for display and security holograms. The recorded relief image in the photoresist plate is then used to make the nickel shim needed as the tool for embossing holograms into plastic materials.

The exposure to actinic radiation produces changes in the photoresist layer that result in a solvency differentiation as a function of exposure. Processing is done with a suitable solvent dissolving either the unexposed or exposed regions, depending on whether the resist is of the negative or the positive type. For optical recording, positive photoresist (exposed resist removed during development) is preferred to the negative type because of the higher resolving power and low scatter. On the resulting surface, relief pattern particles of nickel are deposited by electrolysis to make a mold, which can then be used as an embossing tool. The photoresist process can be used for making transmission holograms only. If an embossed hologram is mirror-backed by using, for example, an aluminum coating process, it can be used in the reflection reconstruction mode as well. Photoresists are UV and deep-blue sensitive. The material is spincoated on glass substrates to obtain a thickness of between 0.5 and 2 μm, then it is baked at about 75 °C for 15 minutes. A typical photoresist for holography (e.g., Shipley Microposit 1350) has a sensitivity of about 10 mJ(cm)$^{-2}$. The most common laser wavelengths for recording photoresist holograms are the following wavelengths: 413 nm (Krypton-ion), 442 nm (Helium–Cadmium), and 458 nm (Argon-ion). At higher wavelengths, e.g., 488 nm, the sensitivity is rather low. Development of Shipley resist is performed in Shipley 303A developer, diluted 1 part developer + 5 parts de-ionized water. Normally only a 10 second development time is needed (under agitation). After that the plate has to be rinsed in de-ionized water for 2 minutes. The plate needs to be dried quickly with a high pressure steam of compressed nitrogen or dry, filtered compressed air.

Commercial Resist Materials

The main American manufacturer of photoresist is Shipley Co. (1457 Macarthur Rd., #101, Whitehall, PA 18052, USA). There are a few companies which make plates based on Shipley resist.

Towne Technologies, Inc. (6–10 Bell Ave., Sommerville, NJ 08876, USA; www.townetech.com) produces spin-coated plates using Shipley resist (Shipley S-1800 series). Towne plates have a sublayer of iron oxide which enhances adhesion of the resist during electroplating operations. In addition, the sublayer eliminates unwanted backscatter from the glass substrate. The plates are developed in Shipley 303A developer.

Hoya Corporation (7-5 Naka-Ochiai 2-chome, Shinjuku-ku, Tokyo, Japan; www.hoya.co.jp) is a producer of mask blanks coated with 0.5 μm thick Shipley S-1800 resist (SLW plates). Different types of glass as well as quartz plates are produced. Hoya supply plates for holography, however, custom orders only.

Thermoplastic Materials

The holographic thermoplastic is a multilayer structure coated on glass or film. The substrate is first coated with a conducting layer, e.g., evaporated gold, then a photoconductor, e.g., poly-n-vinyl carbazole, PVK, that has been sensitized with, e.g., 2,4,7-trinitro-9-fluorenone (TNF), and on top of this layer a thermoplastic coating is deposited (usually a styrene-methacrylate material). The recording of a hologram starts with a uniform charging of the surface of the thermoplastic material using a corona charger. The charge is divided between the photoconductor and the thermoplastic layer. Exposure and the consequent photogeneration in the photoconductor cause charges of opposite signs to migrate to the interface with the thermoplastic layer and the substrate. This will not change the charge, but only decrease the surface potential. Before the image can be developed, the material has to be recharged with the uniform corona charger. This step adds additional charges to the exposed areas, in proportion to the reduced potential which is also proportional to the exposure. The material is then heated to a softening temperature of the thermoplastic layer which will be deformed due to the electrostatic forces acting on it. The material is then cooled and the image is fixed as a relief surface pattern in the thermoplastic layer. The processing time after the exposure necessary to obtain the image is between ten seconds and half a minute. If the material is heated to a somewhat higher temperature, the pattern disappears and the image is erased. The thermoplastic material can now be used for the recording of another hologram, repeating the recording procedure. Such a cycling technique can be repeated hundreds of times, without any serious effect on the quality of the image. That, and the fast dry processing made such materials popular for the use in holographic cameras for nondestructive testing. The sensitivity is between 10 and 100 $\mu J(cm)^{-2}$ over the whole visible electromagnetic spectrum.

Commercial Thermoplastic Materials

In the past, several companies produced thermoplastic materials and recording equipment for holography. Since there is very little demand for such materials today, there is only one company which can deliver thermoplastic equipment for holography: Tavex America Inc. (14 Garden Road, Natick, MA 01760, USA; www.tavexamerica.com). Tavex America manufactures a thermoplastic camera, TCC-2 in which 40 mm by 40 mm plates are used. For obsolete systems, thermoplastic film can still be obtained from MD Diffusion in France (93 rue d'Adelshoffen, F67300 Schiltigheim, France).

Bacteriorhodopsin

Photochromic materials, a class to which bacteriorhodopsin (BR) belongs, have not been used much in holography. Photochromics are real-time recyclable materials which need no processing for development. They can easily be erased and reused over and over again. In holography the only successful material has been the BR film which was introduced only a few years ago. BR is a biomolecule with several applications in photonics, comprising a light-driven molecular proton pump with exceptional photoelectric and photochromic properties. BR is a photochromic protein, found in the photosynthetic system of a salt-marsh bacterium, termed purple membrane (PM), of the cell membrane of *Halobacterium salinarum*. BR thin films are made from polymer suspensions of the isolated purple membrane. Some of the desirable properties of BR films are real-time writing and erasing in microsecond time-scales, durability allowing millions of write/erase cycles without degradation, very high spatial resolution of 5000 $lp(mm)^{-1}$, reasonably good light sensitivity of approximately 10 $mJ(cm)^{-2}$, the encoding of both amplitude and phase patterns, and the use of visible laser light for writing and reading. The composition of the films may also be altered by chemical means to control the optical characteristics, and the BR molecules themselves may be mutagenically changed for performance optimization. Particular properties that can be enhanced in these ways are image lifetimes and the diffraction efficiency of holograms recorded in the films. One advantage of this material is that it is reversible. Therefore it has been considered to be used in holographic storage systems and for real-time holographic interferometry applications. BR film can be used for recording polarization holograms because of BR films inherent photo-inducable optical anisotropy. The material is also possible to expose with pulsed lasers, which makes it suitable for fast holographic recording.

The BR film works in the following way. When illuminated with green light, the BR molecule undergoes an absorption resulting in a maximum shift of about 160 nm towards the red and its color changes from purple to yellow. The yellow image can be erased by illuminating the BR material with blue light. The recording–erasing cycle can be repeated about a million times. Suitable BR-films for holography are obtained by embedding purple membranes into inert matrices like polyvinylalcohol or polyacrylamide. For hologram recording and erasure two photochemical conversions B → M (B-type holograms) and M → B (M-type holograms) can be employed. After excitation by light, BR cycles through a sequence of spectrally distinguishable intermediates and returns to the initial state. The time associated with this thermal relaxation process depends on several aspects, for example, type of BR (wildtype or variant), pH value of film-matrix, film-temperature, etc. For example, the M intermediate is populated from the B state (absorption spectrum peaked at 570 nm) by a green pumping beam and information is recorded with blue light, which initiates the photochemical M → B transition. The M state has its absorption peak at 410 nm. In BR variants in which the thermal relaxation of the M state has been prolonged, photocontrolled switching between these two states can be achieved. Only when the pumping beam is incident upon the BR film can information be recorded with blue light. The pumping beam can be used as a readout beam for the recorded hologram.

Commercial BR Materials

A holographic camera based on a BR film is manufactured in Germany. The holographic system, FringeMaker-*plus*, developed at the University of Marburg is marketed by Munich Innovation Biomaterials (MIB) GmbH. (Hans-Meerwein-Str., D-35032 Marburg, Germany; www.mib-biotech.de). MIB is also a manufacturer of both wildtype and genetically modified BR film for holography and other scientific/technical applications. BR-films offered by MIB can be used in two different recording configurations: the transition from the initial B to M-state (induced with yellow light) serves as basis for optical data recording and processing (B-type recording); or the photochemically induced transition from M to B state (M-type recording). In the latter case, yellow light is used to control the population of the M-state and blue light is used to write the information. MIB offers BR-films with normal and slow thermal relaxation characteristics. Typical relaxation time ranges (RTR), after which the photochemically excited BR molecules have returned to the original state, are between 0.3 and 80 s. The BR film is sealed between two windows of high-quality optical glass. The BR film layer has a thickness of 30–100 μm, dependent on the optical density of the film. The diameter of the film device is 25 mm and the clear aperture is 19 mm.

Photorefractive Crystal Materials

Photorefractive crystals are electro-optic materials which are suitable for recording volume phase holograms in real-time. These crystals are important components in holographic storage systems and are seldom used in other holographic image applications. Holograms recorded in crystals, such as: lithium niobate ($LiNbO_3$); barium titanate ($BaTiO_3$); potassium tantalate niobate (KTN); barium sodium neobate (SBN); bismuth silicon oxide (BSO); and bismuth germanium oxide (BGO), consist of bulk space charge patterns. An interference pattern acting upon a crystal will generate a pattern of electronic charge carriers that are free to move. Carriers are moving to areas of low optical illumination where they get trapped. This effect will form patterns of net space charge, creating an electrical field pattern. As these crystals have strong electro-optic properties, the electrical field pattern will create a corresponding refractive index variation pattern, which means a phase hologram. These holograms can be immediately reconstructed with a laser beam differing from the beams creating the hologram. A hologram can also be erased and a new hologram can be recorded in the crystal. Holograms can be fixed by converting the charge patterns to the patterns of ions which are not sensitive to light. In some materials the hologram can be fixed by heating the crystal to above 100 °C based on thermally activated iconic conductivity. The resulting iconic is frozen upon cooling the crystal back to room temperature. More details about these materials and storage applications are found in the Further Reading.

See also

Holography, Techniques: Overview.

Further Reading

Bjelkhagen HI (1993) *Silver Halide Recording Materials for Holography and Their Processing*. Springer Series in Optical Sciences, vol. 66. Heidelberg: Springer-Verlag.

Bjelkhagen HI (ed.) (1996) *Selected Papers on Holographic Recording Materials*. SPIE Milestone Series, vol. MS130. Bellingham, WA, USA.

Lessard RA (ed.) (1995) *Selected Papers on Photopolymers, Physics, Chemistry, and Applications*. SPIE Milestone Series, vol. MS114. Bellingham, WA, USA.

Sincerbox GT (ed.) (1994) *Selected Papers on Holographic Storage*. SPIE Milestone Series, vol. MS95. Bellingham, WA, USA.

Smith HM (ed.) (1977) *Holographic Recording Materials*, Springer Topics in Applied Physics, vol. 20. Berlin: Springer-Verlag.

HOLOGRAPHY, TECHNIQUES

Contents

Overview
Color Holography
Computer-Generated Holograms
Digital Holography
Holographic Interferometry
Sandwich Holography and Light in Flight

Overview

C Shakher and A K Ghatak, Indian Institute of Technology, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Major Milestones

1948:	Essential concept for holographic recording by Dennis Gabor.
1960:	The first successful operation of a laser device by Theodore Maiman.
1962:	Off-axis technique of holography by Leith and Upatnieks.
1962:	Yu N Denisyuk suggested the idea of three-dimensional holograms based on thick photoemulsion layers. His holograms can be reconstructed in ordinary sunlight. These holograms are called Lippmann–Bragg holograms.
1964:	Leith and Upatnieks pointed out that a multicolor image can be produced by a hologram recorded with three suitably chosen wavelengths.
1969:	S A Benton invented 'Rainbow Holography' for display of holograms in white light. This was a vital step to make holography suitable for display applications.

Holography is the science of recording an entire optical wavefront, both amplitude and phase information, on appropriate recording material. The record is called a hologram. Unlike conventional photography, which records a three-dimensional scene in a two-dimensional format, holography records true three-dimensional information about the scene.

Holography was invented by Gabor in 1948 and his first paper introduced the essential concept for holographic recording – the reference beam. Holography is based on the interference between waves and, it provides us with a way of storing all the light information arriving at the film in such a way that it can be regenerated later.

The use of the reference beam is utilized because the physical detectors and recorders are sensitive only to light intensity. The phase is not recorded but is manifest only when two coherent waves of the same frequency are simultaneously present at the same location. In that case, the waves combine to form a single wave whose intensity depends not only on intensities of the two individual waves, but also on the phase difference between them. This is key to holography. The film record, or hologram, can be considered as a complicated diffraction grating. Holograms bear no resemblance to conventional photographs in that an image is not actually recorded. In fact, the interferometric fringes which are recorded on the recording material are not visible to an unaided eye because of extremely fine interfringe spacing (~0.5 micrometer). The fringes which are visible on the recording material are the result of dust particles in the optical system used to produce the hologram.

Gabor's original technique is now known as in-line holography. In this arrangement, the coherent light source as well as the object, which is a transparency containing small opaque details on a clear background, is located along the axis normal to the photographic plate. With such a system, an observer focusing on one image observes it superposed on the out-of-focus twin image as well as a strong coherent background. This constitutes the most serious problem of Gabor's original technique.

The first successful technique for separating the twin images was developed by Leith and Upatnieks in 1962. This is called the off-axis or side band technique of holography. We shall consider mainly the off-axis technique here.

Basic Holography Principle

Light from a laser or any light beam is characterized by spatial coherence (l_c) and temporal coherence (τ_c), which is discussed in all textbooks on optics. For typical laboratory hologram recording, the required degree of coherence of laser radiation is determined by the type of object and geometrical arrangement being used for the recording. The following condition must hold:

$$L \ll c\tau_c \qquad [1]$$

where τ_c is the coherence time, and L the maximum path length difference between the two waves chosen to record the hologram. To begin recording, two wavefronts are derived from the same laser source. One wavefront is used to illuminate the object and another is used as a reference wavefront. The wavefront derived by illuminating the object is superimposed with the reference wave and the interference pattern is recorded. These object and reference waves must:

- be coherent (derived from the same laser); and
- have a fixed relative phase at each point on the recording medium.

If the above conditions are not met, the fringes will move during the exposure and the holographic record will be smeared. Therefore, to record the hologram one should avoid air currents and vibrations. The recording medium must have sufficient resolution to record the fine details in the fringe pattern. This will be typically of the order of the wavelength. To create a three-dimensional image from the holographic process one then has to record and reconstruct the hologram. The object and reference waves are derived from the same laser.

Holograms record an interference pattern formed by interference of object and reference wavefronts, as explained above. We may mention here that for the two plane waves propagating at an angle θ between them, the spacing of the interference fringes is given by

$$\alpha = \frac{\lambda_0}{2\sin(\theta/2)} \qquad [2]$$

where λ_0 is the free space wavelength.

Processing of the hologram (developing, fixing, and washing of the recording material) yields a plate with alternating transparent and opaque parts, variation of refractive index, or variation of height corresponding to intensity variation in the fringe pattern. Such a plate can be regarded as a complicated diffraction grating. In this process, the hologram is illuminated with monochromatic, coherent light. The hologram diffracts this light into wavefronts which are essentially indistinguishable from the original waves which were diffracted from the object. These diffracted waves produce all the optical phenomena that can be produced by the original waves. They can be collected by a lens and brought into focus, thereby forming an image of the original object, even though the object has since been removed. If the reconstructed waves are intercepted by the eye of an observer, the effect is exactly as if the original waves were being observed; the observer sees what appears to be the original object in true three-dimensional form. As the observer changes his viewing position, the perspective of the image scene changes; parallactic effects are evident and the observer must refocus when the observation point is changed from a near to a distant object in the scene. Assuming that both the construction and reconstruction of the hologram are made with the same monochromatic light source, there is no visual test which can be made to distinguish between the real object and the reconstructed image of the object. It is as if the hologram were a window through which the apparent object is viewed.

Hologram of a Point Object

Consider a hologram recorded with a collimated reference wave normal to the recording plate and a point object inclined at a certain angle. If the hologram is illuminated once again with the same collimated reference wave, it reconstructs two images, one virtual true image and the other real image. However, the two images differ in one very important respect.

While the virtual image is located in the same position as the object and exhibits the same parallax properties, the real image is formed at the same distance from the hologram but in front of it. Corresponding points on the real and virtual images are located at equal distances from the plane of the hologram; the real image has the curious property that its depth is inverted. Such an image is not formed with a normal optical system; it is therefore called a pseudo image as opposed to a normal or orthoscopic image.

This depth inversion results in conflicting visual clues, which make viewing of the real image psychologically unsatisfactory. Thus, if O_1 and O_2 are two elements in the object field, and if O_1 blocks the light scattered by O_2 at a certain angle, the hologram records information only on the element O_1 at this angle and records no information about this part of O_2. An observer viewing the real image from the corresponding direction then cannot see this part of O_2, which, contrary to normal experience, is obscured by O_1, even though O_2 is in front of O_1.

Production of an Orthoscopic Real Image

An orthoscopic real image of an object can be produced by recording two holograms in succession. In the first step, a hologram is recorded of the object with a collimated reference beam. When this hologram is illuminated once again with the collimated reference beam that was used to record it, it reconstructs two images of the object at unit magnification, one of them being an orthoscopic virtual image, while the other is a pseudoscopic real image. A second hologram is then recorded of this real image with a second collimated reference beam.

When the second hologram is illuminated with a collimated beam it reconstructs a pseudoscopic virtual image located in the same position as the real image formed by the second hologram is an orthoscopic image. Since a collimated reference beam is used throughout, the final real image is the same size as the original object and free from aberrations.

In addition to these characteristic linked intensities with the three-dimensional nature of the reconstruction, the holographic recording has several other properties. Each portion of the hologram can reproduce the entire image scene. If a smaller and smaller portion of the hologram is used for reconstruction, there is loss of image intensity and reconstruction. When a hologram is reversed, such as in contact printing processes, it will still reconstruct a positive image indistinguishable from the image produced by the original hologram.

Simple Mathematical Description of Holography

Let the object and reference waves be given by

$$U_o = O(x,y)e^{i\phi(x,y)} \qquad [3]$$

and

$$U_r = R\, e^{iky\sin\theta} \qquad [4]$$

U_r describes the reference (plane wave) propagating at angle θ to the z-axis.

The intensity at the hologram plane is

$$\begin{aligned} I &= |U_r + U_o|^2 \\ &= |U_r|^2 + |U_o|^2 + U_r^* U_o + U_r U_o^* \end{aligned} \qquad [5]$$

As can be seen from the above equation, the amplitude and phase of the wave are encoded as the amplitude and phase modulation of a set of interference fringes. The material used to record the patterns of interference fringes described above is assumed to provide linear mapping of the intensity incident during the reconstruction process into amplitude transmitted by or reflected from the recorded material. Usually both light detection and wavefront modulation are performed by photographic plate/film. We assume the amplitude transmission properties of the plate/film after processing to be described by

$$T = T_o - b(E - E_o) \qquad [6]$$

where the exposure E, at the film is $E = I\tau$; here τ is the exposure time. T_0 is the transmittance of the unexposed plate.

If the hologram is reilluminated by the reference wave, the transmitted wave amplitude will be

$$U_t = U_r(T_o - b\tau(|U_o|^2 + U_r^* U_o + U_r U_o^*)) \qquad [7]$$

The first term is reference wave times a constant. The second term is the reference wave modulated by $|U_o|^2 = O(x,y)^2$; it gives small-angle scattering about the reference wave direction. The third term is proportional to U_o. It is the same as the original object wave (note this is only so if the reconstruction wave is identical to the reference wave). The fourth term is

$$-b\tau U_r^2 U_0^* = -b\tau R^2\, e^{iky2\sin\theta} O(x,y) e^{-i\phi(x,y)} \qquad [8]$$

This is essentially a wave traveling in a direction $\sin^{-1}(2\sin\theta)$ to the z-axis, with the correct object amplitude modulation but its phase reversed in sign, producing a conjugate wave.

Types of Holograms

Primarily, the holograms are classified as thin and thick, based on the thickness of the recording medium. When the thickness of the recording medium is small compared with the average spacing of the interference fringes, the hologram can be treated as a thin hologram. Such holograms are characterized by spatially varying complex amplitude transmittance:

$$\mathbf{t}(x, y) = [\mathbf{t}(x, y)]\exp[-i\phi(x, y)] \qquad [9]$$

Thin holograms can be further categorized as thin amplitude holograms or thin phase holograms. If amplitude transmittance of the hologram is such that $\phi(x, y)$ is constant while $\mathbf{t}(x, y)$ varies over the hologram, the hologram is termed an amplitude hologram. For a lossless phase hologram, $|\mathbf{t}(x, y)| = 1$, so that the complex amplitude transmittance is caused by variation in phase.

When the thickness of the hologram recording material is large compared to the fringe spacing of the interference fringes, the holograms may be considered as volume holograms. These may be treated as a three-dimensional system of layers corresponding to a periodic variation of absorption coefficient or refractive index, and the diffracted amplitude is at a maximum when Bragg's condition is satisfied. In general, the behavior of thin and thick holograms corresponds to Raman Nath and Bragg diffraction regimes. The distinctions between two regimes is made on the basis of a parameter Q, which is defined by the relation:

$$Q = \frac{2\pi\lambda_o d}{n_o \Lambda^2} \qquad [10]$$

where

Λ = spacing of fringe on hologram, measured normal to the surface;
d = thickness of recording medium;
n_o = mean refractive index of the medium; and
λ_o = wavelength of light.

Small values of Q ($Q < 1$) correspond to thin gratings, while large values of Q ($Q > 10$) correspond to volume gratings. For values of Q between 1 and 10, intermediate properties are exhibited. However, this criterion is not always adequate. The boundaries between the thin and thick holograms are given by the value of

$$P = \frac{\lambda_o^2}{\Lambda^2 n_o n_1} \qquad [11]$$

where P^{-2} is the relative power diffracted into higher orders, and $P < 1$ for thin holograms and $P > 10$ for thick holograms. For P having values between 1 and 10, the hologram may be thin or thick, depending on the other parameters involved.

Holograms can also be classified based on whether they are reconstructed in transmitted light or reflected light. For transmission holograms, during the recording stage, two interfering wavefronts make equal but opposite angles to the surfaces of recording medium and are incident on it from the same side. Reflection holograms reconstruct images in reflected light. They are recorded such that the interfering wavefronts are symmetrical with respect to the surface of the recording medium but are incident on it from opposite sides. When the angle between the interfering wavefronts is maximum (180°), the spacing between the fringe planes of the recording medium is minimum. Under such conditions, reflection holograms may have wavelength sensitivity high enough to be reconstructed, even with white light.

An important development in the field of holography was the invention of rainbow holograms by Benton in 1969. This provides a method for utilizing white light for illumination when viewing the holograms. The technique does so by minimizing the blur introduced by color dispersion in transmission holograms, at the price of giving up parallax information in one dimension.

Recording Materials

An ideal recording medium for holography should have a well matching spectral sensitivity corresponding to available laser wavelengths, linear transfer characteristics, high resolution, and low noise. It should also be either inexpensive or indefinitely recyclable. Toward achieving the above-mentioned properties, several materials have been studied, but none has been found so far that meets all the requirements. Materials investigated include: silver halide photographic emulsion; dichromated gelatin plates/films; silver halide sensitized gelatin plates/films; photo-resists; photo polymer systems; photochromics; photo thermoplastics; and ferro-electric crystals. Recently, the use of storage and processing capabilities of computers, together with CCD cameras, has been used for recording holograms.

Silver halide photographic emulsions are a commonly used recording material for holography, mainly because of relatively high sensitivity and easy availability. Manufacture, use, and processing of these emulsions have been well standardized. The need for wet processing and drying may constitute a major drawback. Dichromated gelatin can be considered an almost ideal recording material for volume phase holograms, as it has a large refractive index modulation capability, high resolution, low absorption and scattering.

Because of these features, dichromated gelatin has been extensively investigated. The most significant disadvantage is its comparatively low sensitivity. Pennington *et al.* developed an alternative technique, which combines the advantage of silver halide emulsions (high sensitivity) with that of dichromated gelatin (low absorption scattering and high stability). It involves exposing silver halide photographic emulsion and then processing it, so as to obtain a volume phase hologram consisting solely of hardened gelatin.

Thin phase holograms can be recorded on photoresists, which are light-sensitive organic films yielding relief image after exposure and development. They offer advantages of easy replication using thermoplastics but are slow in response and undergo nonlinear effects at diffraction efficiency greater than ~0.05. Shipley AZ-1350 is a widely used photoresist, with maximum sensitivity in the ultraviolet, dropping rapidly for longer wavelengths towards the blue.

Photopolymers are being keenly investigated because they offer advantages such as ease of handling, low cost, and real-time recording for the application of holography and nonlinear optics. However, they have low sensitivity and short shelf life.

Thin phase surface relief holograms can be recorded in a thin layer of thermoplastics. These materials have a reasonably high sensitivity over the whole visible spectrum, fairly high diffraction efficiency, and do not require wet processing. Their application in holographic interferometry, optical information processing, and in making compact holographic devices has been widely reported.

Photorefractive crystals such as $Bi_{12}SiO_{20}$, $LiNbO_3$, $Bi_{12}GeO_{20}$, $BaTiO_3$, $LiTaO_3$, etc., as recording materials, offer high-angular sensitivity and provide capability to read and write volume holographic data in real time. Besides the materials discussed here, several other materials have been investigated for recording holograms; these include photocromics, elastomere devices, magneto-optic materials, etc.

Application of Holography

Holography can be constructed not only with the light waves of lasers, but also with sound waves, microwaves, and other waves in the electromagnetic spectrum of radiation. Holograms made with ultraviolet light or X-rays can record images of objects/particles smaller than the wavelength of visible light, e.g., atoms or molecules. Acoustical holography uses sound waves to see through solid objects. Holography has a vast scope of practical applications, which have been classified into two major categories:

1. applications requiring three-dimensional images for visual perception; and
2. applications in which holography is used as a measuring tool.

The unique ability of holography – to record and reconstruct both electromagnetic and sound waves – makes it a valuable tool for education, science, industry, and business. Below are some of the important applications:

1. Holographic interferometry (HI) is one of the most powerful measuring tools. In HI, two states of an object, i.e., initial and deformed state are recorded on the same photographic plate. After reconstruction of the light wave corresponding to two states of an object, interferences and deformations are displayed in terms of the interference pattern. The change of distance of one tenth of a micron, or lower, can be resolved. HI provides scientists/engineers with crucial data for design of critical machine parts of power-generating equipment, in the aircraft industry, automobile industry, and nuclear installations (say, for example, in the design of containers used to transport nuclear materials, improve the design of aircraft wings and turbine blades, etc.). Presently, HI is being widely used in mechanical engineering, acoustics, and aerodynamics, for nondestructive testing, to investigate oscillation in diaphragms and flow around various objects, respectively.
2. Microwave holography can detect objects deep within spaces, by recording the radio waves they emit.
3. Another important application of holography is the design of optical elements, which possess special properties. A holographic recording of a concave mirror behaves in much the same way as the mirror itself, i.e., it can focus the light. In some cases chromatism can be introduced in the design of elements so that a location of the point, where the beams are focused, depends on the wavelength. This can be achieved by accurately choosing the recording arrangement of the focusing elements, these elements are, in fact, diffraction gratings. These can have low noise levels, freedom from astigmatism, and have other useful properties. Holographic optical elements have found applications in supermarket scanners to read barcodes, headup displays in fighter aircraft to observe critical cockpit instruments, etc.
4. A telephone credit card used in Europe and other developed countries has embossed surface holograms which carry a monetary value. When the card is inserted into the telephone, a card reader discerns the amount due and deducts the appropriate amount to cover the cost of the call.

5. Holography is having applications in analog and digital computers, offering remarkable opportunities to realize various logical operators, devices for identifying images based on matched filtering, and in computer memory units. The basic advantage of holographic memory is that a relatively large volume of information can be stored and that there are a limited number of ways to change the record. The arrival of the first prototype of optical computers, which use holograms as storage material for data, could have a dramatic impact on the holographic market. The yet-to-be-unveiled optical computers will be able to deliver trillions of bits of information faster than the current generation of computers.
6. Optical tweezers are going to become an important tool for the study of micro-organisms/bacteria, etc.
7. Holograms can be used to locate and retrieve information without knowing its location in the storage medium, only needing to know some of its content.
8. The links between computer science and holography are now well-established and are developing, with at least two aspects making computer-generated holograms extremely interesting. First, such holograms enable us to obtain visual 3-dimensional reconstructions of imagined objects. For example, one can reconstruct three dimensions of a model of an object still in the design stage. Second, computer-generated holograms can be used to reconstruct lightwaves with specified wave fronts. This means specially computed and manufactured holograms may function as optical elements that transform the incident light wave into desired wavefronts.
9. Another important applications of holography is its utilization to compensate for the distortion that occurs when viewing objects through optically heterogeneous mediums. It can be achieved based on the principle of beam reversal by the hologram.
10. Finally, let us consider the application of holography to art. The development of holography gives very effective ways of creating qualitative three-dimensional images. Thus, a new independent area of holographic creative work representational/artistic holography has appeared. The art of holographic depiction has developed along two major routes. The first is creation of the view hologram, used as holograms of natural objects that are to be displayed in exhibitions and museums; these are also known as artistic holograms. Portrait holography is also classified under this category. The progress in portrait holography is hampered partly because of imperfection of pulsed lasers and partly because of the deterioration of photographic material when exposed to pulsed electromagnetic radiation.

The principle behind the creation of elusion by using composite holograms is also very convincing for the display of objects. To synthesize the composite holograms, the photographs of various aspects of a scene are printed onto the photographic plate. The synthesis techniques for preparing the composite hologram are very complicated, while the images created by holograms are still far from perfect. However, there is no doubt that composite holography opens holography up as an artistic technique. Recently rainbow holography has been very popular to display such objects.

See also

Holography, Applications: Art Holography. **Holography, Techniques:** Color Holography; Computer-Generated Holograms; Digital Holography; Holographic Interferometry. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Benton SA (1969) On a method for reducing the information content of holograms. *Journal of the Optical Society of America* 59: 1545.

Denisyuk YuN (1962) Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Sov. Phys.-Dokl.* 7: 543.

Denisyuk YuN (1963) On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt. I. *Optical Spectroscopy* 15: 279.

Denisyuk YuN (1965) On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt II. *Optical Spectroscopy* 18: 152.

Gabor SD (1948) A new microscopic principle. *Nature* 161: 777.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceedings of the Royal Society* A197: 454.

Gabor D (1951) Microscopy by reconstructed wavefronts: II. *Proceedings of the Physical Society* B64: 449.

Hariharan P (1983) *Optical Holography: Principle, Techniques and Applications*. Cambridge, UK: Cambridge University Press.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal Optical of the Society of America* 52: 1123.

Leith EN and Upatnieks J (1963) Wavefront reconstruction with continuous-tone objects. *Journal of the Optical Society of America* 53: 1377.

Leith EN and Upatnieks J (1964) Wavefront reconstruction with diffused illumination and three-dimensional objects. *Journal of the Optical Society of America* 54: 1295.

Pennington KS and Lin LH (1965) Multicolor wavefront reconstruction. *Applied Physics Letters* 7: 56.

Stroke GW and Labeyrie AE (1966) White-light reconstruction of holographic images using the Lippmann-Bragg diffraction effect. *Physics Letters* 20: 368.

Color Holography

H I Bjelkhagen, De Montfort University, Leicester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the appearance of the first laser-recorded monochromatic holograms in the 1960s, the possibilities of recording full-color high-quality holograms have now become a reality. Over many years holographic techniques have been created to produce holograms with different colors, referred to as pseudo- or multicolor holograms. For example, in the field of display holography, it is common to make multiple-exposed reflection holograms using a single-wavelength laser and emulsion thickness manipulation to obtain an image with different colors. In this way many beautiful pseudocolor holograms have been made by artists. Applying Benton's rainbow hologram technique (for example, the one used for embossing holograms on credit cards), it has been possible to record transmission holograms as well as mass-produce embossed holograms with 'natural' colors. However, a rainbow hologram has a very limited viewing position from which a correct color image can be seen.

This article will acquaint the reader with the topic of color holography (full-color or true-color holography). This describes a holographic technique to obtain a color 3D image of an object where the color rendition is as close as possible to the color of the real object. A brief review of the development of color holography, the current status and the prospects of this new 3D imaging technique, are presented. Also included are recording techniques using three lasers providing red, green and blue (RGB) laser light as well as computer-generating techniques for producing full parallax 3D color images.

The Development of Color Holography

In theory, the first methods for recording color holograms were established in the early 1960s. Already Leith and Upatnieks (1964) proposed multicolor wavefront reconstruction in one of their first papers on holography. The early methods concerned mainly transmission holograms recorded with three different wavelengths from a laser or lasers, combined with different reference directions to avoid cross-talk. Such a color hologram was then reconstructed (displayed) by using the original laser wavelengths from the corresponding reference directions. However, the complicated and expensive reconstruction setup prevented this technique from becoming popular. More suitable for holographic color recording is reflection holography, which can be reconstructed and viewed in ordinary white light.

Most likely, the three-beam transmission technique might eventually become equally applicable, provided inexpensive multicolor semiconductor lasers become widely available. Such RGB lasers can be used to display a very realistic, deep, and sharp holographic image of, for example, a room.

Relatively few improvements in color holography were made during the 1960s and 1970s. Not until the end of the 1970s did a few examples of color holograms appear. During the 1980s, new and improved techniques were introduced. A review of various transmission and reflection techniques for color holography was published by Hariharan (1983). From 1990 until today, many high-quality color holograms have been recorded, mainly due to the introduction of new and improved panchromatic recording materials. On the market are ultra-fine-grain silver halide emulsions (manufactured by Slavich in Russia) as well as photopolymer materials (manufactured by E. I. du Pont de Nemours & Co. in the USA).

Color reflection holography presents no problems with regard to the geometry of the recording setup, but the final result is highly dependent on the recording material used and the processing techniques applied. There are some problems associated with recording color holograms in general and specifically in silver halide emulsions:

- Scattering occurring in the blue part of the spectrum typical of many holographic coarse-grain silver halide emulsions makes them unsuitable for the recording of color holograms.

- Multiple storage of interference patterns in a single emulsion reduces the diffraction efficiency of each individual recording. The diffraction efficiency of a three-color recording in a single-layer emulsion is lower than a single wavelength recording in the same emulsion.
- During wet processing, emulsion shrinkage frequently occurs, causing a wavelength shift. White-light-illuminated reflection holograms normally show an increased bandwidth upon reconstruction, thus affecting the color rendition.
- The fourth problem, related to some extent to the recording material itself, is the selection of appropriate laser wavelengths and their combination in order to obtain the best possible color rendition of the object.

In the beginning, when no suitable panchromatic emulsion existed, the sandwich technique was used to make color reflection holograms. Two plates were sandwiched together, in which, for example, two different types of recording materials were used. The most successful demonstration of the sandwich recording technique was made by Kubota (1986) in Japan. He used a dichromated gelatin (DCG) plate for the green (515 nm) and the blue (488 nm) components, and an Agfa 8E75 silver halide plate for the red (633 nm) component of the image. Kubota's sandwich color hologram of a Japanese doll recorded in 1986, demonstrated the potential of high-quality color holography for the first time. Color holograms have also been recorded in red-sensitized DCG materials.

Extensive work in the field of reflection color holography was performed by Hubel and Solymar (1991). Primarily they employed Ilford silver halide materials for the recording of color holograms applying the sandwich technique.

Very important is the research on recording materials by Usanov and Shevtsov (1990) in Russia. Their work is based on the formation of a microcavity structure in gelatin, by applying a special processing technique to silver halide emulsions. Such holograms recorded in a silver halide emulsion have a high diffraction efficiency and exhibit a typical DCG hologram quality.

Not until panchromatic ultra-fine-grain silver halide emulsions were introduced in Russia in the early 1990s, was it possible to record high-quality color holograms in a single emulsion layer as demonstrated by the present author. Although it is now possible to produce color holograms, they need to be displayed correctly. An important improvement would be to find an edge-illuminating technique to make color holograms easier to display. The display problem still remains the main obstacle preventing color holograms, as well as monochrome holograms, being more widely used.

Recording of Color Holograms

Currently, most transmission color holograms are of the rainbow type. Large-format holographic stereograms made from color-separated movie or video recordings have been produced. There are also some examples of embossed holograms in which a correct color image is reproduced. In order to generate a true color rainbow hologram, a special setup is required in which the direction of the reference beam can be changed in between the recordings of the color-separated RGB primary images. However, as already mentioned, holographic color images of the rainbow type can reconstruct a correct color image only along a horizontal line in front of the film or plate and, thus, of less interest for serious color holography applications. Reflection holography can offer full parallax, large field of view 3D color images where the colors do not change when observed from different directions.

Silver Halide Materials

To be able to record high-quality color reflection holograms it is necessary to use extremely low light-scattering recording materials. This means, for example, the use of ultra-fine-grain silver halide emulsions (grain size about 10 nm). Currently, the only producer of a commercial holographic panchromatic ultra-fine-grain silver halide material is the Micron branch of the Slavich photographic company located outside Moscow. The PFG-03c emulsion comes in standard sizes from 63 mm × 63 mm format up to 300 mm × 406 mm glass plates and is also available on film. Some characteristics of the Slavich material are presented in Table 1.

By using suitable processing chemistry for the PFG-03c emulsion it has been possible to obtain high-quality color holograms. These holograms are recorded in a single-layer emulsion, which greatly simplifies the recording process as compared with many earlier techniques. The aforesaid single-layer technique is described below.

Table 1 Characteristics of the Slavich panchromatic emulsion

Silver halide material	PFG-03C
Emulsion thickness	7 μm
Grain size	12–20 nm
Resolution	~10,000 lp(mm)$^{-1}$
Blue sensitivity	~1.0–1.5 × 10^{-3} J(cm)$^{-2}$
Green sensitivity	~1.2–1.6 × 10^{-3} J(cm)$^{-2}$
Red sensitivity	~0.8–1.2 × 10^{-3} J(cm)$^{-2}$
Color sensitivity peaked at	633 nm and 530 nm

Photopolymer Materials

The color holography photopolymer material from DuPont is another alternative recording material for color holograms. In particular, this type of material is suitable for mass production of color holograms. Although, being less sensitive than the ultra-fine-grain silver halide emulsion, it has special advantages of easy handling and dry processing (only UV-curing and baking). The DuPont color photopolymer material has a coated film layer thickness of about 20 μm. The photopolymer film is generally coated in a 12.5-inch width on a 14-inch wide Mylar® polyester base which is 0.002-inch thick. The film is protected with a 0.00092-inch thick Mylar® polyester cover sheet. However, the panchromatic polymer material is only supplied to specially selected and approved hologram producers. The reason for this is that some applications of color holograms are in the field of document security where optically variable devices (OVDs) are produced.

The recording of a color hologram on DuPont polymer is simple. The film has to be laminated to a piece of clean glass or attached to a glass plate using an index-matching liquid. To obtain the right color balance, the RGB sensitivity depends on the particular material, but typically red sensitivity is lower than green and blue sensitivities. It is difficult to obtain high red-sensitivity of photopolymer materials. Simultaneous exposure is the best recording technique for photopolymer materials. Holograms can be recorded manually, but in order to produce large quantities of holograms, a special machine is required. For hologram replication the scanning technique can provide the highest production rate. In this case, three scanning laser lines are needed, which can be adjusted in such a way that all three simultaneously can scan the film. The color photopolymer material needs an overall exposure of about 10 mJ(cm)$^{-2}$.

After the exposure is finished, the film has to be exposed to strong white or UV light. DuPont recommends about 100 mJ(cm)$^{-2}$ exposure at 350–380 nm. After that, the hologram is put in an oven at a temperature of 120 °C for two hours in order to increase the brightness of the image.

Laser Wavelengths for Color Holograms

The problem of choosing optimal primary laser wavelengths for color holography is illustrated in the 1976 CIE chromaticity diagram (Figure 1) which indicates suitable laser wavelengths for color holograms.

It may seem that the main aim of choosing the recording wavelengths for color holograms would be to cover as large an area of the chromaticity diagram as possible. However, there are many other

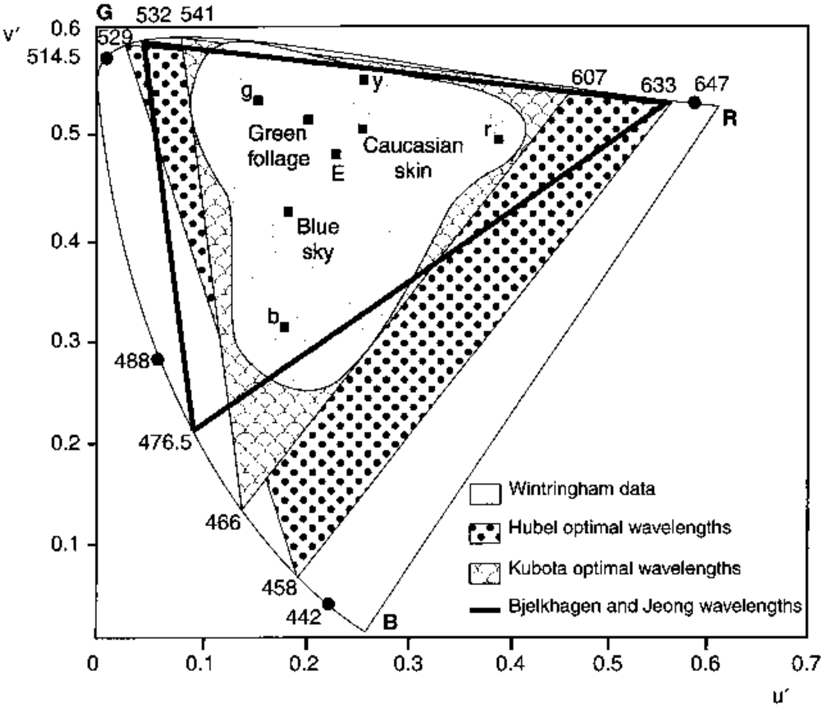

Figure 1 The 1976 CIE uniform scales chromaticity diagram shows the gamut of surface colors and positions of common laser wavelengths. Optimal color-recording laser wavelengths are also indicated.

considerations that must be taken into account when choosing the wavelengths for color holograms. One of these important considerations is the question of whether three wavelengths are really sufficient for color holography. The wavelength selection problem for color holography has been treated in several papers, for example by Peercy and Hesselink (1994).

Hubel and Solymar (1991) provided a definition of color recording in holography:

> A holographic technique is said to reproduce 'true' colors if the average vector length of a standard set of colored surfaces is less than 0.015 chromaticity coordinate units, and the gamut area obtained by these surfaces is within 40% of the reference gamut. Average vector length and gamut area should both be computed using a suitable white light standard reference illuminant.

An important factor to bear in mind when working with silver halide materials is that a slightly longer blue wavelength than the optimal one might give higher-quality holograms (better signal-to-noise ratio) because of reduced Rayleigh scattering during the recording. Another important factor to consider is the reflectivity of the object at primary spectral wavelengths. It has been shown that the reflectivity of an object at three wavelength bands, peaked at 450, 540, and 610 nm, has an important bearing on color reconstruction. These wavelengths can also be considered optimal for the recording of color holograms even though the laser lines are very narrow-band. A triangle larger than necessary can be considered in order to compensate for color desaturation (color shifting towards white) that takes place when reconstructing color reflection holograms in white light. According to Hubel's color rendering analysis, the optimal wavelengths are 464, 527, and 606 nm for the sandwich silver halide recording technique. If the calculations are performed to maximize the gamut area instead, the following set of wavelengths is obtained: 456, 532, and 624 nm. However, when approached from a different viewpoint, the optimal trio of wavelengths based on the reconstructing light source of 3400 K, a 6 μm thick emulsion with a refractive index of 1.63 and an angle of 30° between the object and the reference beam the following wavelengths were obtained: 466.0, 540.9, and 606.6 nm. Peercy and Hesselink (1994) discussed wavelength selection by investigating the sampling nature of the holographic process. During the recording of a color hologram the chosen wavelengths point-sample the surface-reflectance functions of the object. This sampling on color perception can be investigated by the tristimulus value of points in the reconstructed hologram which is mathematically equivalent to integral approximations for the tristimulus integrals. Peercy and Hesselink used both Gaussian quadrature and Riemann summation for the approximation of the tristimulus integrals. In the first case they found the wavelengths to be 437, 547, and 665 nm. In the second case the wavelengths were 475, 550, and 625 nm. According to Peercy and Hesselink, the sampling approach indicates that three monochromatic sources are almost always insufficient to preserve all of the object's spectral information accurately. They claim that four or five laser wavelengths are required.

Only further experiments will show how many wavelengths are necessary and which combination is the best for practical purposes. Another factor that may influence the choice of the recording wavelengths is the availability of cw lasers currently in use in holographic recording, for example, argon ion, krypton ion, diode-pumped solid state (DPSS) frequency-doubled Nd:YAG, helium neon, and helium cadmium lasers (Table 2).

The recent progress in DPSS laser technology has made available both red and blue DPSS lasers. These lasers are air-cooled, small, and each laser requires less than a hundred watts of electric power to operate. Usually, a set of three DPSS lasers will be the best choice of cw lasers for color holography in the future.

Setup for Recording Color Holograms

A typical reflection hologram recording setup is illustrated in Figure 2.

Table 2 Wavelengths from cw lasers

Wavelength [nm]	*Laser type*	*Single line power [mW]*
442	Helium cadmium	<100
457	DPSS blue	<500
458	Argon ion	<500
468	Krypton ion	<250
476	Krypton ion	<500
477	Argon ion	<500
488	Argon ion	<2000
497	Argon ion	<500
502	Argon ion	<400
514	Argon ion	<5000
521	Krypton ion	<100
529	Argon ion	<600
531	Krypton ion	<250
532	DPSS green	<3000
543	Green neon	<10
568	Krypton ion	<100
633	Helium neon	<80
647	Krypton ion	<2000
656	DPSS red	<1000

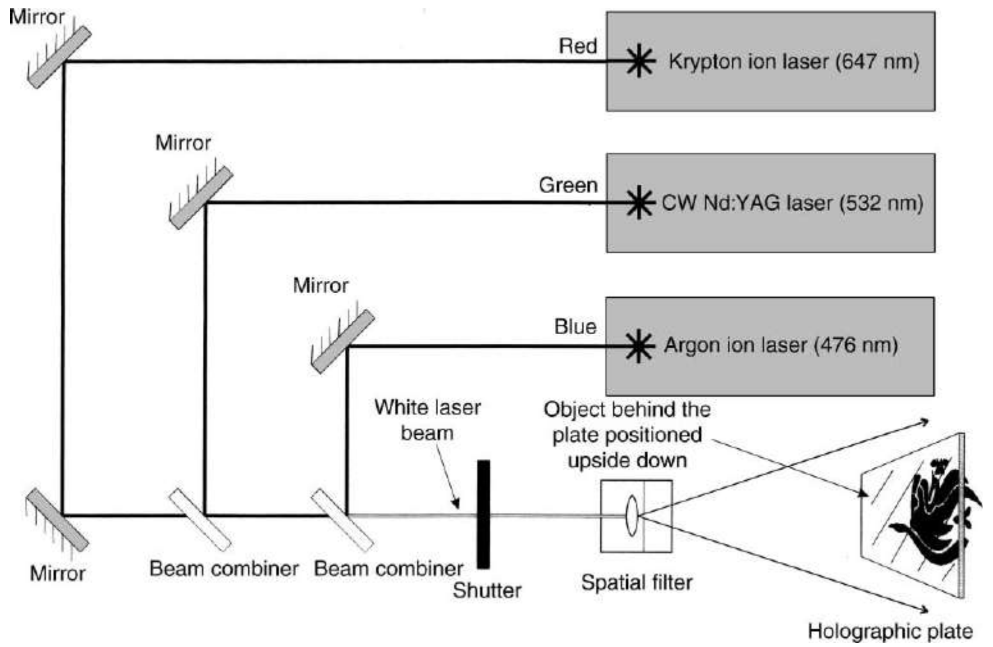

Figure 2 The setup for recording color reflection holograms.

The different laser beams necessary for the exposure of the object pass through the same beam expander and spatial filter. A single-beam Denisyuk arrangement is used, i.e., the object is illuminated through a recording holographic plate. The light reflected from the object constitutes the object beam of the hologram. The reference beam is formed by the three expanded laser beams. This 'white' laser beam illuminates both the holographic plate and the object itself through the plate. Each of the three primary laser wavelengths forms its individual interference pattern in the emulsion, all of which are recorded simultaneously during the exposure. In this way, three holographic images (a red, a green, and a blue image) are superimposed upon one another in the emulsion.

Three laser wavelengths are employed for the recording: 476 nm, provided by an argon ion laser, 532 nm, provided by a cw frequency-doubled Nd:YAG laser, and 647 nm, provided by a krypton laser. Two dichroic filters are used for the combining of the three laser beams. The 'white' laser beam goes through a spatial filter, illuminating the object through the holographic plate.

By using the dichroic filter beam combination technique, it is possible to perform simultaneous exposure recording, which makes it possible to control independently the RGB ratio and the overall exposure energy in the emulsion. The RGB ratio can be varied by individually changing the output power of the lasers, while the overall exposure energy is controlled solely by the exposure time. The overall energy density for exposure is about 3 $\mathrm{mJ(cm)}^{-2}$.

In the initial experiments performed by the author in 1993, a specially designed test object consisting of the 1931 CIE chromaticity diagram, a rainbow ribbon cable, pure yellow dots, and a cloisonné elephant, was used for the color balance adjustments and exposure tests. Later, another test target was employed – the Macbeth ColorChecker® chart, which was used for color rendering tests.

Color reflection holograms can also be produced using copying techniques so that projected real images can be obtained, however, normally associated with a restricted field of view. For many display purposes, the very large field of view obtainable in a Denisyuk color hologram is often more attractive.

Processing of Color Holograms

The dry processing of color holograms recorded on photopolymer materials has already been described. The process is simple and very suitable for machine processing using, for example, a baking scroll oven. The processing of silver halide emulsions is more difficult and critical. The Slavich emulsion is rather soft, and it is important to harden the emulsion before the development and bleaching takes place. Emulsion shrinkage and other emulsion distortions caused by the active solutions used for the processing must be avoided. In particular, when recording master color holograms intended for photopolymer replication, shrinkage control is extremely important. The processing steps are summarized in Table 3.

Table 3 Color holography processing steps

1.	Tanning in a formaldehyde solution	6 min
2.	Short rinse	5 s
3.	Development in the CWC2 developer	3 min
4.	Wash	5 min
5.	Bleaching in the PBU-amidol bleach	~5 min
6.	Wash	10 min
7.	Soaking in acetic acid bath (printout prevention)	1 min
8.	Short rinse	1 min
9.	Washing in distilled water with wetting agent added	1 min
10.	Air drying	

It is very important to employ a suitable bleach bath to convert the developed silver hologram into a phase hologram. The bleach must create an almost stain-free clear emulsion so as not to affect the color image. In addition, no emulsion shrinkage can be permitted, as it would change the colors of the image. Washing and drying must also be done so that no shrinkage occurs. Finally, to prevent any potential emulsion thickness variation by humidity variations, the emulsion needs to be protected by a glass plate sealed onto the hologram plate.

Evaluation of Color Holograms

Recorded color holograms of the two test targets are presented here. The illuminating spotlight to reconstruct the recorded color holograms was a 12-Volt 50-Watt halogen lamp. This type of spotlight is suitable for displaying color holograms. The selection of a suitable lamp for the reconstruction of color holograms is much more important than the selection of lamps for monochrome hologram display. The color balance for the recording of a color hologram must be adjusted to the type of spotlight that is going to be used for the display of the finished hologram. Figure 3 shows a typical normalized spectrum obtained from a white area of the color test target hologram.

This means that the diffraction efficiency of each color component is obtained assuming a flat spectrum of the illuminating source. One should note the high diffraction efficiency in blue, needed to compensate for the low blue light emission of the halogen spotlight. The noise level, mainly in the blue part of the spectrum, is visible and low. The three peaks are exactly at the recording wavelengths; i.e., 647, 532, and 476 nm.

In Table 4 some results of the Macbeth ColorChecker® investigation are presented. The 1931 CIE x and y coordinates are measured at both the actual target and the holographic image of the target. The measured fields are indicated in the table by color and the corresponding field number.

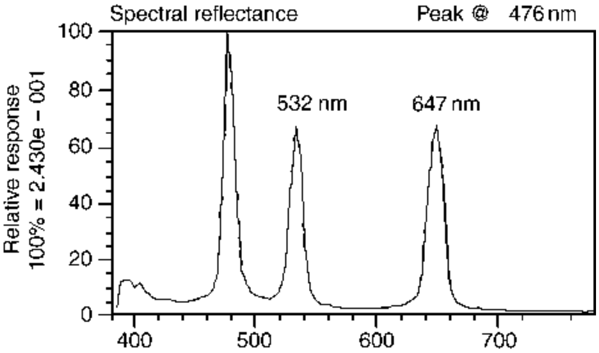

Figure 3 Normalized spectrum from a white area of a color test target hologram.

Color reproductions of two color holograms by the author are presented. A photograph of the 100 mm by 120 mm hologram of the CIE test object is shown in Figure 4. Featured in Figure 5 is a 200 mm by 250 mm hologram of a large Russian egg.

Computer-Generated Color Holograms

Today it is not possible to obtain a computer-generated hologram (CGH) with the same high information content as the laser-recorded ones of real objects. What may become possible in the future is a technique to compute and record the interference pattern that is stored in a Denisyuk color hologram which, upon illumination, can produce an image like the laser-recorded ones of real objects. Therefore, the best compromise is to use holographic stereograms which can provide high-quality computer-generated images. A holographic stereogram is created by using a series of 2D photographic images or 2D images displayed on a LCD screen from which the hologram is recorded. In order to obtain a high-quality holographic image with a large field of view, many 2D images are needed. As mentioned earlier, some color CGHs of the rainbow transmission type have been produced. However, image color changes as a function of viewer position in that type of hologram, making them less attractive than computer-generated reflection holograms. Over the last few years there has been rapid progress in color CGHs of the reflection type. Remarkable results have been achieved by Klug *et al.* (1997) at Zebra Imaging Inc. A new technique, to record very large full-parallax color reflection CGHs, has been developed. Color holograms can be produced, having both vertical and horizontal parallax, with a 100° field of view. The generation of a holographic hardcopy of either digitized images or computer graphics models is based on the following technique. The 'object'

Table 4 Chromaticity coordinates from color hologram recording tests using the Macbeth ColorChecker®

Object	*White #19*	*Blue #13*	*Green #14*	*Red #15*	*Yellow #16*	*Magenta #17*	*Cyan #18*
CIE *x y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*
Target	0.435/0.405	0.295/0.260	0.389/0.514	0.615/0.335	0.517/0.450	0.524/0.322	0.285/0.380
Image	0.354/0.419	0.335/0.362	0.337/0.449	0.476/0.357	0.416/0.437	0.448/0.338	0.295/0.366

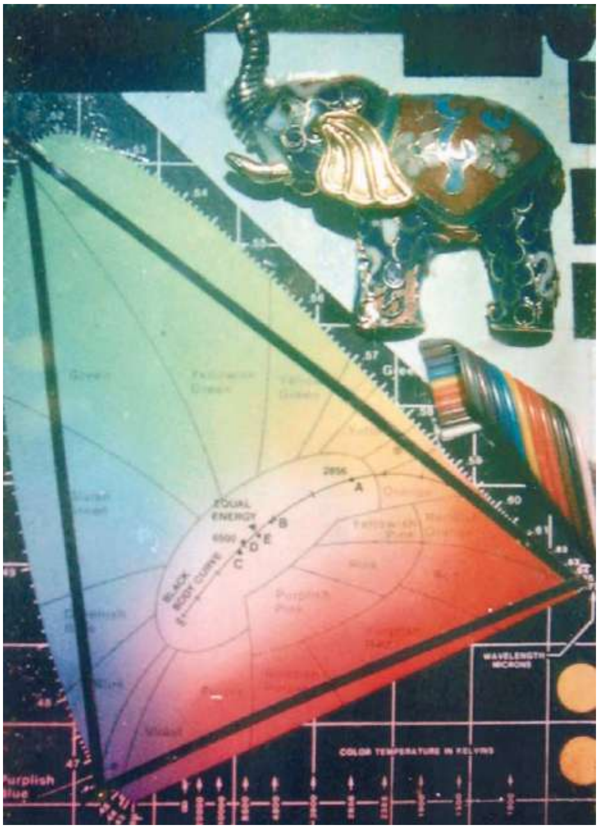

Figure 4 Hologram of the CIE test object recorded with 476, 532, and 647 nm laser wavelengths. Notice the rainbow ribbon cable, the pure yellow dots, the full range of colors, and a balanced white in the center of the CIE color test target.

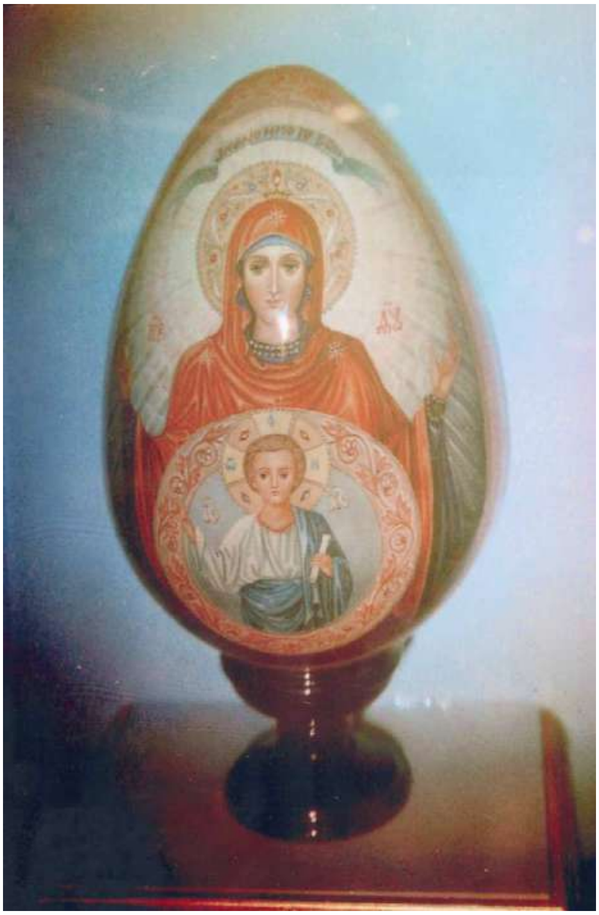

Figure 5 Color hologram of a large Russian egg, size 100 mm by 120 mm.

sub-beam is directed through a sequence of digital images on a liquid-crystal screen. Each resulting exposure, about two millimeters square, is called a 'hogel'. The full-color hogels are the holographic building blocks of a finished CGH image. In an automated step-and-repeat process, 9×10^4 hogels are formed on a flat square tile of DuPont panchromatic photopolymer film by simultaneous RGB laser exposures. The 60 cm by 60 cm tile itself is the finished hologram, or to obtain larger holograms, the 3D image is made in 60 cm by 60 cm tiled increments. So far, the largest image created was of Ford's P2000 Prodigy concept car in which ten such holograms tiled together made up the large color reflection hologram (1.2 m by 3 m) which is reproduced in Figure 6.

This technique has opened the door to real 3D computer graphics. However, generating such a large hologram is a very time-consuming process. The Ford P2000 hologram took almost two weeks (300 hours) to produce, since each individual panel requires a 24-hour recording time.

The Future of Color Holography

The manufacturing of large-format color reflection holograms is now possible. Mass production of color holograms on photopolymer materials has started in Japan. Although good color rendition can be obtained, some problems remain to be solved, such

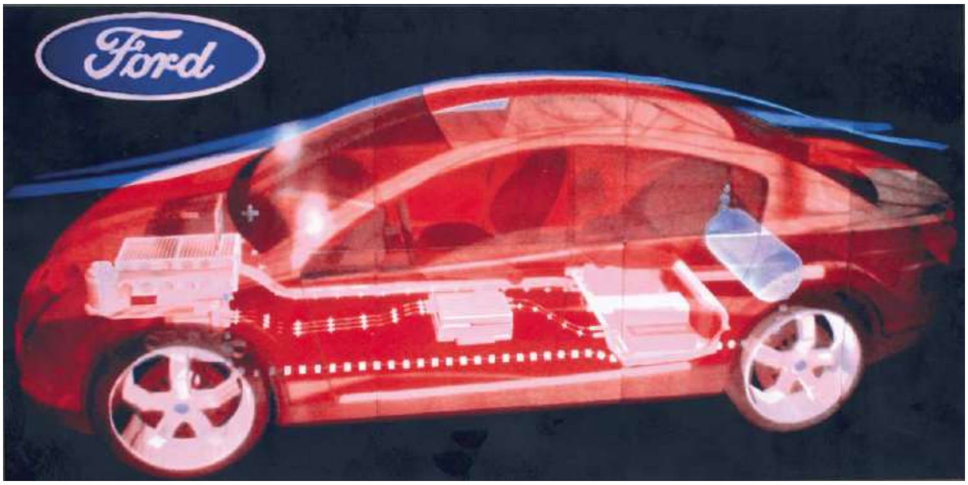

Figure 6 Computer-generated color hologram of Ford's P2000 Prodigy concept car. A ten tile full parallax Mosaic hologram produced by Zebra Imaging, Inc., size 1.2 m by 3 m.

as desaturation, image resolution, signal-to-noise ratio and dynamic range. Other limitations concerning holographic color recording include the fact that some colors we see are the result of fluorescence, which cannot be recorded in a hologram. There are some differences in the recorded colors in one of the test charts shown here. However, color rendition is a very subjective matter. Different ways of rendition may be preferable for different applications, and different people may have different color preferences.

At the present time, research on processing color reflection holograms recorded on ultra-fine-grain materials is still in progress. Work is carried out on microcavity structure holograms in order to produce low-noise reflection holograms with high diffraction efficiency. For producing large quantities of color holograms, the holographic photopolymer materials are most suitable and will be further improved in the future.

Another important field of research is the development of a three-wavelength pulsed laser. Employing such a laser, dynamic events, as well as portraits, can be recorded in a holographic form. Currently, research work is in progress at the French German Research Institute ISL, Saint Louis, France, as well as at the Geola company in Lithuania.

The virtual color image behind a color holographic plate represents the most realistic image of an object that can be obtained today. The extensive field of view adds to the illusion of beholding a real object rather than an image of it. The wavefront reconstruction process recreates accurately the three-wavelength light scattered off the object during the recording of the color hologram. This 3D imaging technique has many obvious applications, in particular, in displaying unique and expensive artifacts. There are also many potential commercial applications of this new feature of holographic imaging, provided that the display problem can be solved using some sort of integrated lighting.

Today, it is technologically possible to record and replay acoustical waves with very high fidelity. Hopefully, holographic techniques will be able to offer the same possibility in the field of optical waves, wavefront storage, and reconstruction. Further development of computer-generated holographic images will make it possible to display extremely realistic full-parallax 3D color images of nonexisting objects, which could be applied in various spheres, for example, in product prototyping, as well as in other applications in 3D visualization and three-dimensional art. Eventually it will become possible to generate true-holographic 3D color images in real-time, as computers become faster and possess greater storage capacity. However, the most important issue is the development of extremely high-resolution electronic display devices which are necessary for electronic holographic real-time imaging.

See also

Holography, Techniques: Computer-Generated Holograms.

Further Reading

Bjelkhagen HI (1993) *Silver Halide Recording Materials for Holography and Their Processing*. Springer Series in Optical Sciences, vol. 66. New York: Springer-Verlag.

Bjelkhagen HI, Jeong TH and Vukičević D (1996) Color reflection holograms recorded in a panchromatic ultrahigh-resolution single-layer silver halide emulsion. *Journal of Imaging Science & Technology* 40: 134–146.

Bjelkhagen HI, Huang Q and Jeong TH (1998) Progress in color reflection holography. In: Jeong TH (ed.) *Sixth International Symposium on Display Holography*, Proc. SPIE, vol. 3358, pp. 104–113. Bellingham, WA: The International Society for Optical Engineering.

Hariharan P (1983) Colour holography. In: Wolf E (ed.) *Progress in Optics*, vol. 20, pp. 263–324. Amsterdam: North-Holland.

Hubel PM and Solymar L (1991) Color reflection holography: theory and experiment. *Applied Optics* 30: 4190–4203.

Klug MA, Klein A, Plesniak W, Kropp A and Chen B (1997) Optics for full parallax holographic stereograms. In: Benton SA and Trout TJ (eds) *Practical Holography XI and Holographic Materials III*, Proc. SPIE, vol. 3011, pp. 78–88. Bellingham, WA: The International Society for Optical Engineering.

Kubota T (1986) Recording of high quality color holograms. *Applied Optics* 25: 4141–4145.

Leith EN and Upatnieks J (1964) Wavefront reconstruction with diffused illumination and three-dimensional objects. *Journal of the Optical Society of America* 54: 1295–1301.

Peercy MS and Hesselink L (1994) Wavelength selection for true-color holography. *Applied Optics* 33: 6811–6817.

Usanov YuE and Shevtsov MK (1990) Principles of fabricating micropore silver-halide-gelatin holograms. *Optics and Spectroscopy (USSR)* 69: 112–114.

Watanabe M, Matsuyama T, Kodama D and Hotta T (1999) Mass-produced color graphic arts holograms. In: Benton SA (ed.) *Practical Holography XIII*, Proc. SPIE, vol. 3637, pp. 204–212. Bellingham, WA: The International Society for Optical Engineering.

Computer-Generated Holograms

W J Dallas, University of Arizona, Tucson, AZ, USA

A W Lohmann, University of Erlangen-Nürnberg, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A hologram is a tool for shaping the amplitude and the phase of a light wave. Shaping the amplitude distribution is easy, one way is by a photographic plate, another way is by binary pixels as used in 'half-tone printing'. The phase distribution can be shaped by RE-fraction on a thin film of variable thickness, or by DIF-fraction on a diffraction grating that has been deliberately distorted.

Both kinds of holograms can be manufactured based on a computer design, the refractive version being called a 'kinoform', and the diffractive version being known as a computer-generated hologram (CGH). We will explain the fundamentals of computer holography with emphasis on the 'Fourier–CGH', and will also highlight the large variety of applications.

From the Classical Hologram to the Computer-Generated Hologram: CGH

Holography is a method to form images, consisting of two steps: recording and reconstruction. In the recording step, some interference fringes are recorded on a photographic plate (Figure 1a). The two interfering waves originate from the object and from a reference light source. After being developed by the usual photo-chemical treatment the photographic plate is called a hologram. In the reconstruction step, the hologram acts as a diffracting object that is illuminated by a replica of the reference light. Due to diffraction, the light behind the hologram is split into three parts (Figure 1b). One part proceeds to the observer who perceives a virtual object where the genuine object had been originally during the recording step.

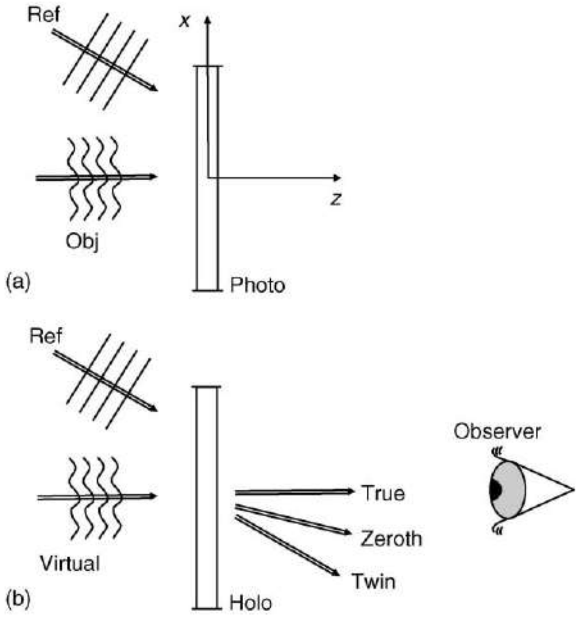

Figure 1 Holographic image formation in two steps: (a) recording the hologram, (b) reconstructing the object wave.

This was a brief description of ‘classical-holography’. In ‘computer holography’ the first step, recording, is synthetic. In other words, the propagation of the light from the object to the photographic plate is simulated digitally. The simulation includes the addition of the object wave to the reference wave and the subsequent modulus square process, which describes the transition from the two complex amplitudes to the hologram irradiance:

$$|u_0(x) + u_R(x)|^2 = I_H(x) \qquad [1]$$

The amplitude transmittance of the hologram is

$$T_H(x) = c_0 + c_1 I_H(x) = c_1 u_0 u_R^* + \cdots \qquad [2]$$

The coefficients c_0 and c_1 represent the photochemical development. In the reconstruction process (Figure 1b) the transmittance T_H is illuminated by the reference beam. Hence, we get

$$u_R(x) T_H(x) = c_1 u_0(x) |u_R(x)|^2 + \cdots \qquad [3]$$

If the reference intensity $|u_R(x)|^2$ is constant, this term represents a reconstruction of the object wave $u_0(x)$. The omitted parts in eqns [2] and [3] describe those beams that are ignored by the observer (Figure 1b). The hologram transmittance $T_H(x)$ is an analog signal:

$$0 \le T_H \le 1 \qquad [4]$$

Before the 1960s, when computer holograms were invented, the available plotters could produce only binary transmittance patterns. Hence, it was necessary to replace the analog transmission $T_H(x)$ by a binary transmission $B_H(x)$, which is possible without loss of image quality. In addition, binary computer holograms have better light efficiency and a better signal-to-noise ratio. They are also more robust when being copied or printed.

Another advantage of a CGH is that the object does not have to exist in reality, which is important for some applications. The genuine object may be difficult to manufacture or, if it is three-dimensional, difficult to illuminate. The CGH may be used to implement an algorithm for optical image processing. In that case, the term ‘object’ becomes somewhat fictitious. We will discuss more about applications, towards the end of this article.

From a Diffraction Grating to a Fourier CGH

Now we will explain, in some detail, the ‘Fourier type’ CGH. Fourier holograms are more popular than two other types: ‘image-plane’ CGHs and ‘Fresnel-type’ CGHs. The Fresnel-type will be treated in the section on ‘software’ because of some interest in 3-D holograms. Certain hardware issues will also be described.

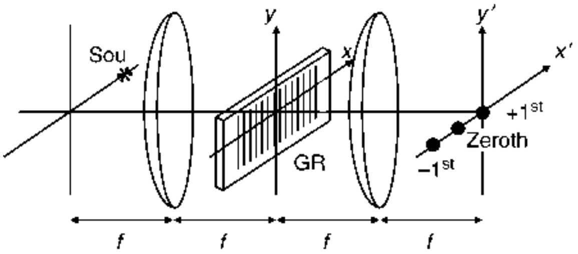

Figure 2 Grating diffraction, but with an off-axis source.

The explanation of Fourier CGHs begins with grating diffraction (Figure 2). The only uncommon feature in Figure 2 is the off-axis location of the source. It is arranged such that the plus-first diffraction order will hit the center of the output plane.

We will now convert a simple Ronchi grating (Figure 3a) into a Fourier CGH. The Ronchi grating transmittance can be expressed as the Fourier series:

$$G(x) = \sum_{(m)} C_m \exp\left(\frac{2\pi i m x}{D}\right); \quad C_0 = \frac{1}{2};$$

$$C_m = \frac{\sin(\pi m/2)}{\pi m} \qquad [5]$$

Now we shift the grating bars by an amount PD. Then we modify the width from $D/2$ to WD, as shown in Figure 3b. The Fourier coefficients are now

$$C_m = \frac{\exp(2\pi i m P)\sin(\pi m W)}{\pi m} \qquad [6]$$

The C_{+1} coefficient is responsible for the complex amplitude at the center of the readout plane:

$$C_{+1} = \frac{\exp(2\pi i P)\sin(\pi W)}{\pi} \qquad [7]$$

We are able now to control the beam magnitude, $A = (\frac{1}{\pi})\sin(\pi W)$, by the relative bar width W and the phase $\phi = 2\pi P$ by the relative bar shift P. This particular kind of phase shift is known as ‘detour phase’.

So far, the light that ends up at the center of the exit plane, leaves the grating as a plane wave with a wave vector parallel to the optical axis. Now we distort the grating on purpose (Figure 3c):

$$W \to W(x, y); \qquad P \to P(x, y) \qquad [8]$$

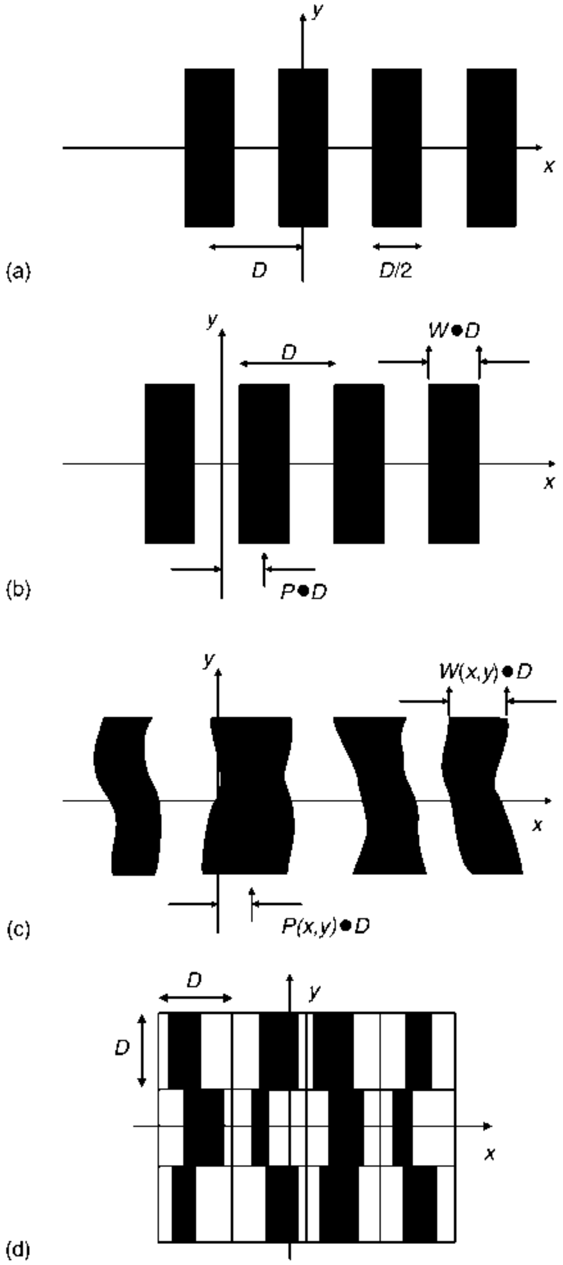

Figure 3 Modifying a grating into a hologram in 3 steps: (a) Ronchi grating, (b) Modified width and location of the grating bars, (c) Distorted grating bars, (d) as in **Figure 3c**, but now discrete distortions.

These distortions should be mild enough that the plus-first order light behind the distorted grating can be described by the complex amplitude:

$$\left(\frac{1}{\pi}\right)\sin[\pi W(x,y)]\exp[2\pi i P(x,y)] = u_{\mathrm{H}}(x,y) \qquad [9]$$

This is a generalization of eqn [7]. The phase $2\pi P(x,y)$ covers the range from $-\pi$ to $+\pi/2$ if the shift is bounded by $|P| \le \frac{1}{2}$. That is enough to cover the range of complex amplitudes within the circle of $|u_{\mathrm{H}}| \le \frac{1}{\pi}$. We will return shortly to the condition 'mild-enough distortions' in the context of Figure 4.

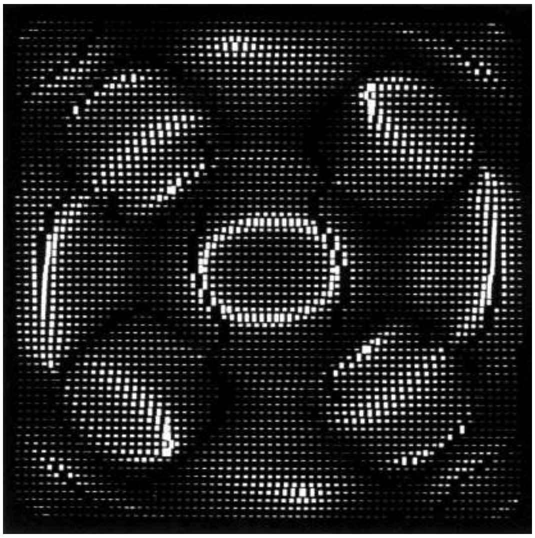

Figure 4 CGH with 64 × 64 cells.

Most plotters move their pens only in the x-direction and y-direction and printers place dots on a grid. Therefore it is convenient to replace the continuous (x,y) variations of W and P by piece-wise constant variations, as in Figure 3d. That restriction turns out to be acceptable, as demonstrated by the first computer holograms. The theoretical proof involves the sampling theorem.

Recall that the light propagation through a 2-f system (Figure 2, between grating and output plane) can be described by a Fourier transformation:

$$\iint u_{\mathrm{H}}(x,y)\exp\left[\frac{-2\pi i(xx'+yy')}{\lambda f}\right]\mathrm{d}x\,\mathrm{d}y = \tilde{u}_{\mathrm{H}}\left(\frac{x'}{\lambda f},\frac{y'}{\lambda f}\right) \qquad [10]$$

Suppose now that we wish to see a particular image: $v(x',y')$. Hence, we request that

$$v(x',y') = \tilde{u}_{\mathrm{H}}\left(\frac{x'}{\lambda f},\frac{y'}{\lambda f}\right)\mathrm{RECT}\left(\frac{x'}{\lambda f}\right) \qquad [11]$$

The RECT function indicates that only the PLUS first diffraction order is of interest. As a consequence we have to select the grating distortions $W(x,y)$ and $P(x,y)$ such that the hologram $u_{\mathrm{H}}(x,y)$ is the

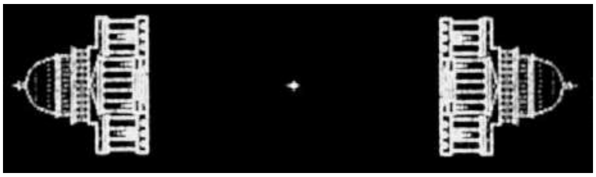

Figure 5 A reconstruction from a 1024 × 1024 CGH. The real image and the symmetrical twin image appear. The zeroth-order is blocked.

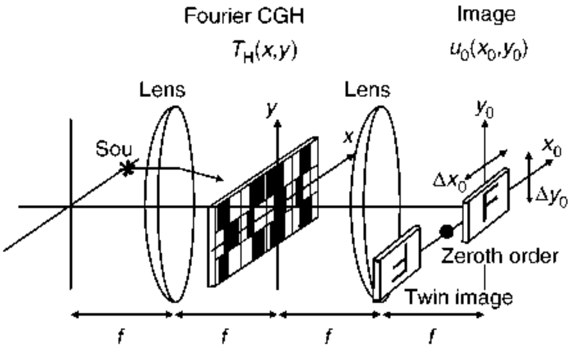

Figure 6 Reconstruction setup for a Fourier CGH.

Fourier transform of $v(x', y')$:

$$u_H(x,y) = \iint v(x',y') \exp\left[\frac{2\pi i(xx'+yy')}{\lambda f}\right] dx'\,dy' \qquad [12]$$

The condition 'mild enough distortions' has been treated in the literature. There is not enough space here to present that part of the CGH theory. But we show an actual CGH with 64 × 64 cells (Figure 4). This CGH is no longer a grating, but, it shows enough resemblance to our heuristic derivation. Hence, the CGH theory appears to be trustworthy. A CGH looks less regular if the image $v(x', y')$ contains a random phase that simulates a diffuser. The diffuser spreads light across the CGH and so levels out unacceptably large fluctuations in the amplitude.

With 512 × 512 or 1024 × 1024 cells in a CGH one can obtain an image of the quality seen in Figure 5. The reconstruction setup is shown in Figure 6, here with the letter F as the image.

About Some CGH Algorithms

The computational effort for getting the desired hologram amplitude of a Fourier hologram u_H is reasonable due to the powerful FFT algorithm. However, to use it the data must be discrete. In other words, the CGH consists of cells, centered at $x_m = mD$, $y_n = nD$ as shown in Figure 3d. The questions now are: 'What is the proper cell size D?' and 'How many cells are needed?' The answers to those questions depend on the parameters of the image. If the zero order (see Figure 6) should be at the distance Δx_0, then the quasi-grating period D should obey:

$$\frac{\lambda f}{D} \geq \Delta x_0 \qquad [13]$$

And if the size of an image pixel should be as fine as δx_0, the size Δx_H of the hologram should obey:

$$\frac{\lambda f}{\Delta x_H} \leq \delta x_0 \qquad [14]$$

This condition is plausible because the finite size Δx_H of the hologram acts like a resolution-limiting aperture. The combination of eqns [13] and [14] yields

$$\frac{\Delta x_0}{\delta x_0} \leq \frac{\Delta x_H}{D} \qquad [15]$$

This means that the number of image pixels $\Delta x_0/\delta x_0$ is bounded by the number of CGH cells $\Delta x_H/D$. The generalization to two dimensions is straightforward. Again, Fourier CGHs benefit from the fact that the light propagation from the virtual object to the hologram plane can be described by a Fourier transformation. Hence, the FFT can be used.

And now we move to the synthesis of a Fresnel hologram. A Fresnel hologram is recorded at a finite distance z from the object. Hence, we have to simulate the wave propagation in free space from the object (at $z = 0$) to the hologram at a distance z. This free space propagation can be described (in the paraxial approximation) as a convolution of object and quasi-spherical wave:

$$u_0(x) \rightarrow \int_{-\infty}^{\infty} u_0(x') \exp\left[\frac{-i\pi(x-x')^2}{\lambda z}\right] dx' = u(x,z) \qquad [16]$$

In the Fourier domain the corresponding operation is:

$$\tilde{u}_0(\mu) \rightarrow \tilde{u}_0(\mu) \exp[-i\pi\lambda z\mu^2] = \tilde{u}(\mu, z) \qquad [17]$$

The indirect execution of eqn [16], based on eqn [17], is easy: a Fourier transformation of the object converts $u_0(x)$ into $\tilde{u}_0(\mu)$, which is then multiplied by the quadratic phase factor, yielding $\tilde{u}(\mu, z)$. An inverse Fourier transformation produces $u(x, z)$. The two Fourier transformations consist typically of $4N \log N$ multiply/add operations. N is the number of pixels: image size Δx_0, divided by the pixel size δx_0.

In the case of a rectangular object $u_0(x, y)$, the number of pixels is $\Delta x_0 \Delta y_0 / \delta x_0 \delta y_0$.

Now to describe the algorithmic effort for computing the convolution, the $u_0(x)$ has N pixels, and, the lateral size of the exponential is essentially $M = z/\delta z$, where δz represents 'depth of focus'. M describes the lateral size of the diffraction cone, measured in pixel size units δx_0, and the convolution involves MN multiply/adds. A comparison of these two algorithms is dictated by the ratio

$$\frac{M}{4 \log N} \qquad [18]$$

The direct convolution is preferable if the distance z is fairly small. The integration limits of eqn [16] depend on the oscillating phase $\pi x^2/\lambda z$. Effectively, nothing is contributed to this integral, when this phase varies by more than 2π over the length, δx_0, of an object pixel. Note that the pixel size of $u(x, z)$ does not change with the distance z, because the bandwidth $\Delta\mu(z) = \Delta\mu_0$ is z-independent. This is so because the power spectrum is z-invariant:

$$|\tilde{u}(\mu, z)|^2 = |\tilde{u}_0(\mu)|^2 \qquad [19]$$

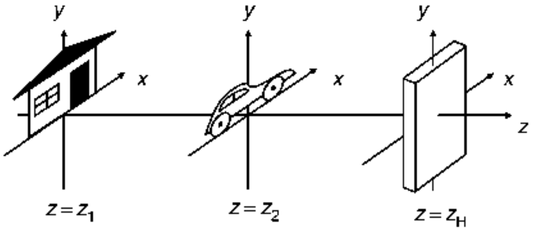

Figure 7 Synthesis of a hologram for a three-dimensional object.

It is easy to proceed from here to the CGH synthesis of a 3-D object. One starts with an object detail $u_1(x)$ at z_1, for example, the house as in Figure 7. Then, one propagates to the next object detail at z_2. This second detail may act as a multiplication (transmission) and as addition (source elements). This process is repeated as the wave moves towards the plane of the hologram. There, a reference wave has to be added, as in the Fourier case.

A very powerful algorithm for the CGH synthesis is 'IFTA' (iterative Fourier transform algorithm). Suppose, we want to design a Fourier CGH, which looks like a person and whose reconstructed image shows the signature of that person (Figure 8a,b). In other words, the two intensities

$$|u_H(x, y)|^2 \text{ and } |\tilde{u}_H(\mu, v)|^2 \qquad [20]$$

are determined. But, the phase distributions of u_H and $\tilde{u}_H$ are still at our disposal. The IFTA algorithm will yield those phases in many cases at the expense of ten, hundred or more Fourier transformations.

The IFTA algorithm is sometimes called 'Fourier Ping-Pong' because it bounces back and forth between the (x, y) and the (μ, ν) domain. The amplitudes $|u_H|$ and $|u_0|$ are enforced at every opportunity. But the associated phases are free to vary and to converge, eventually. If so, u_H and u_0 are completely known. The IFTA will not always converge, for example not if

$$|u_H(x, y)| = \delta(x, y) \text{ and } |\tilde{u}_H(\mu, v)| = \delta(\mu, v) \qquad [21]$$

But the chances are fairly good if both amplitudes are often close to one and seldom near zero.

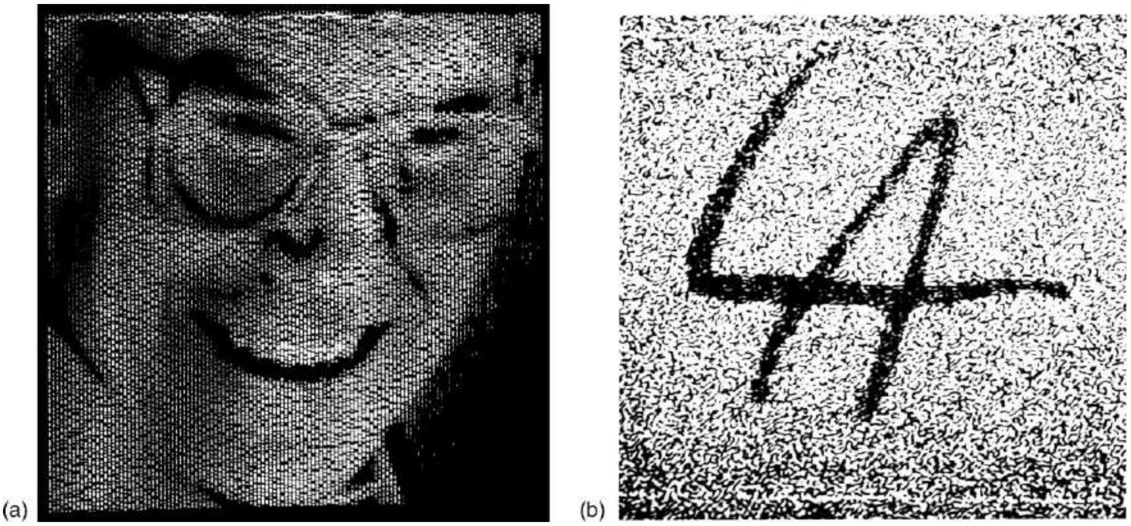

Figure 8 IFTA-CGH: (a) The Fourier CGH. (b) The optical reconstruction thereof. (Courtesy of H.O. Bartelt.)

Finally, an image plane CGH is trivial from an algorithmic point of view. The simulated propagation from object plane to image plane is usually an identity, apart from the finite bandwidth of the image-forming system. Nevertheless, image plane holograms can be quite useful, for example for interferometric testing of a nonspherical mirror of an astronomical telescope. The CGH is laid out to provide a wavefront that is suitable for an interferometric 'null test'. The CGH serves as a 'synthetic prototype'.

A fourth kind of CGH, to be named near-field CGH may emerge. It operates in the near-field and it employs evanescent waves as they occur, if object details are comparable in size with, or even smaller than, the wavelength. Conceivable topics are 'super resolution' and 'sub-lambda lithography'. Fundamentals and history of classical evanescent holography are reported in the Further Reading.

Some Hardware Aspects

We now describe some hardware aspects of the CGHs. Remembering the CGH cell structure in Figure 3d, we saw the amplitude encoded as relative width W (eqn [6]). Instead of the width, one may use the relative height $H \leq 1$ as the amplitude parameter. More efficient, but more difficult to manufacture are saw tooth-shaped hologram cells. The saw tooth prism may be approximated by a phase stair, which is possible with today's lithographic technology. The corresponding theory is called phase quantization.

Some CGH Applications

Some applications had been mentioned already before, e.g., 3-D display (see Figure 7). The data volume can be very high, hence, shortcuts such as eliminating the vertical perspective, are necessary. The largest CGHs for visual light are used for testing astronomical telescopes by means of interferometry. The IFTA algorithm has been used, among other things, for designing a CGH which acts as an optimal diffuser. Speckles and light efficiency are the critical issues in diffuser design. Another project, that also benefits from IFTA, is beam shaping. That is a broad topic with aims at, for example, homogenizing the output of laser, beam structuring for welding, cutting, and scanning.

A CGH can also be used as a spatial filter for image processing. For example, an input image may be Hilbert-transformed in the setup shown in Figure 9a. The filter (Figure 9b) is a grating, of which one sideband is shifted by one-half of the grating period. Such a geometric shift generates the π-phase shift that is needed for the Hilbert transformation. The same setup can also be used to implement Zernike's phase contrast. Now the grating is distorted such that the zero-frequency region at the optical axis is reduced in amplitude by (typically) a factor of 0.2. In addition, the phase is shifted by $\pi/2$ due to a fringe shift of a quarter period (Figure 10a). In the image plane one can see an ordinary image on-axis, and two phase-contrast images in the plus-minus first diffraction orders. One of them has positive phase contrast, the

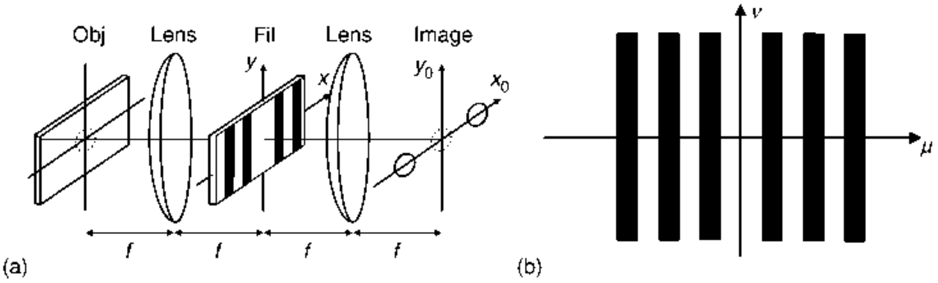

Figure 9 (a) Hilbert setup. (b) Hilbert filter.

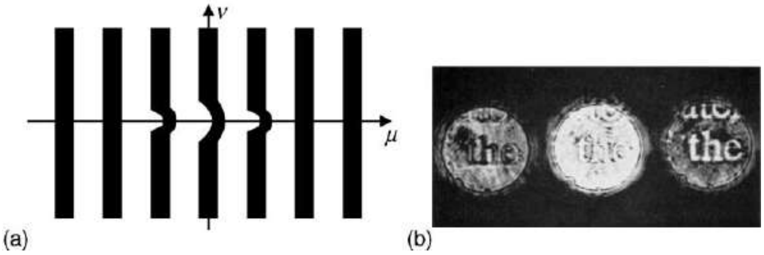

Figure 10 (a) Phase contrast filter. (b) Phase contrast output.

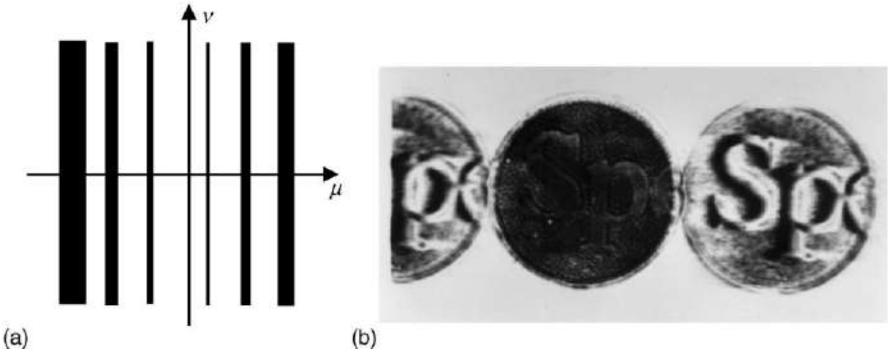

Figure 11 (a) Differentiation filter. (b) Differentiation output.

other one has negative phase contrast (Figure 10b). Another simple spatial filter (Figure 11a) produces a derivative of the input $u(x, y) : \partial u(x, y)/\partial x$ (Figure 11b). Equations [22]–[24] explain this differential filtering experiment.

$$u(x, y) \rightarrow \frac{\partial u(x, y)}{\partial x} = v(x, y) \qquad [22]$$

$$\tilde{u}(\mu, \nu) \rightarrow 2\pi i\mu \cdot \tilde{u}(\mu, \nu) = \tilde{v}(\mu, \nu) \qquad [23]$$

$$\tilde{p}(\mu, \nu) = 2\pi|\mu| \cdot \begin{Bmatrix} +i(\mu > 0) \\ -i(\mu > 0) \end{Bmatrix} \qquad [24]$$

The phase portion of the filter (Figure 11a) is the same as for the Hilbert filter, but, the filter amplitude now enhances the higher frequencies. The object used for getting the results of Figure 10b and 11b was a phase-only object: a bleached photograph.

A very famous spatial filtering experiment is ‘matched filtering’, which can be used for pattern recognition. The original matched filters were classical Fourier holograms of the target pattern. Computer hologram’s can do the same job, even with spatially incoherent objects, this work being initiated by Katyl.

Fourier holograms are also attractive for data storage, due to their robustness against local defects. However, a local error in the Fourier domain spreads out all over the image domain. The computed hologram has an advantage over classical holograms because of the freedom to incorporate any error detection and correction codes for reducing the bit error rate even further.

One of the most recent advances is in matched filtering with totally incoherent light. This step allows moving spatial filtering out of the well-guarded laboratory into hostile outdoor environments.

Finally, a word about terminology; several terms are used in place of ‘CGH’: synthetic hologram, digital hologram, holographic optical element (HOE), diffractive optical element (DOE) or digital optical element.

CGHs have also been made for other waves, such as acoustical waves, ocean waves, and microwaves. Digital antenna steering represents a case of computer holography.

See also

Diffraction: Fraunhofer Diffraction; Fresnel Diffraction. **Fourier Optics**. **Holography, Techniques:** Computer-Generated Holograms; Holographic Interferometry; Overview. **Imaging:** Information Theory in Imaging; Inverse Problems and Computational Imaging; Three-Dimensional Field Transformations. **Information Processing:** Coherent Analogue Optical Processors; Free-Space Optical Computing; Incoherent Analog Optical Processors. **Optical Communication Systems:** Free-Space Optical Communications. **Optical Processing Systems**. **Phase Control:** Wavefront Coding.

Further Reading

Andrés P (1999) *Opt. Lett.* 245: 1331.
Brown BR and Lohmann AW (1966) *Appl. Opt.* 5: 967.
Brown BR and Lohmann AW (1969) *IBM J. R&D* 13: 160.
Bryngdahl O (1973) *Progr. Opt.* 11: 168.
Bryngdahl O and Wyrowski F (1990) *Progr. Opt.* 28: 1–86.
Dallas WJ (1971) *Appl. Opt.* 10: 673.
Dallas WJ (1973) *Appl. Opt.* 12: 1179.
Dallas WJ (1980) *Topics in Appl. Phys.* 40: 291–366.
Dallas WJ (1999–2003) http://www.radiology.arizona.edu/dallas/CGH.htm. University of Arizona, Tucson, AZ.
Dresel T, Beyerlein M and Schwider J (1996) *Appl. Opt.* 35: 4615.
Fienup JR (1981) *Proc. SPIE* 373: 147.
Fienup JR (1982) *Appl. Opt.* 21: 2758.
Gerchberg RW and Saxton WO (1972) *Optik* 35: 237.
Goodman JW and Silvestri AM (1970) *IBM J. R&D* 14: 478.

Hansler RL (1968) *Appl. Opt.* 7: 1863.

Hauk D and Lohmann AW (1958) *Optik* 15: 275.

Ichioka Y and Lohmann AW (1972) *Appl. Opt.* 11: 2597.

Katyl RH (1972) *Appl. Opt.* 11: 1255.

Kozma A and Kelly DL (1965) *Appl. Opt.* 4: 387.

Lee WH (1978) *Progr. Opt.* 16: 219–232.

Lesem LB, Hirsch PM and Jordan JA (1969) *IBM J. R&D* 13: 150.

Lohmann AW and Paris DP (1967) *Appl. Opt.* 6: 1746.

Lohmann AW and Paris DP (1968) *Appl. Opt.* 7: 651.

Lohmann AW and Sinzinger S (1995) *Appl. Opt.* 34: 3172.

Lohmann AW and Werlich HW (1971) *Appl. Opt.* 10: 670.

McGovern AJ and Wyant JC (1971) *Appl. Opt.* 10: 619.

Pe'er A, Wang D, Lohmann AW and Friesem AA (1999) *Opt. Lett.* 24: 1469; (2000) 25: 776.

Sinzinger S and Jahns J (1999) *Micro Optics*. New York: Wiley.

Vander Lugt AB (1964) *IEEE, IT* 10: 139.

Wyrowski F and Bryngdahl O (1991) *Rep. Prog. Phys.* 1481–1571.

Yaroslavskii LP and Merzlyakov NS (1980) *Methods in Digital Holography*. New York: Plenum Press.

Digital Holography

W Osten, University of Stuttgart, Stuggart, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The sinusoidal nature of light, its ability to interfere, and the precise knowledge of its wavelength make it possible to measure several metrological quantities, such as distances and displacements, with high accuracy. However, the high frequency of light results in the difficulty that the primary quantity to be measured, i.e., the phase of the light wave, cannot be observed directly. All quadratic sensors (CCD, CMOS, photo plates) are only able to record intensities. Therefore, one makes use of a trick: the transformation of phase changes into recordable intensity changes as the basic principle of interferometry and holography. Holography provides the possibility to record and to reconstruct the complete information of optical wave fields. Numerous potential applications such as 3D-imaging of natural objects, fabrication of diffractive optical elements, and interferometric investigation of complex technical components become possible. Further progress has been made by recording holograms directly with electronic targets and not in photographic emulsions. In this way the independent numerical reconstruction of phase and intensity can be accomplished by the computer. One important consequence is that interferometric techniques, such as 3D-displacement analysis and 3D-contouring, can be implemented easily and with high flexibility. Furthermore, digitally recorded holograms can be transmitted via telecommunication networks and optically reconstructed at any desired place using an appropriate spatial light modulator. Consequently, the complete optical information of complex objects can be transported between different places with high velocity. This opens the possibility to compare nominal identical but physically different objects (master sample comparison) available at different locations with interferometric sensitivity. We call this remote metrology.

Direct Phase Reconstruction by Digital Holography

In digital speckle pattern interferometry (DSPI), the object is focused onto the target of an electronic sensor. Thus, an image plane hologram is formed as a result of the interference with an inline reference wave. In contrast with DSPI, a digital hologram is recorded without imaging. The target records the superposition of the reference and the object wave in the near-field region – a so-called Fresnel hologram. The basic optical setup in digital holography for recording holograms is the same as in conventional holography (Figure 1a). A laser beam is divided into two coherent beams, one illuminating the object and forming the object wave, the other entering the target directly and forming the reference wave. On this basis, very compact solutions are possible. Figure 1b shows a holographic camera where the interferometer, containing a beamsplitter and some wavefront shaping components, is mounted in front of the camera target.

For a description of the principle of digital Fresnel holography, we use a simplified version of the optical setup (Figure 2a). The object is modeled by a plane rough surface that is located in the (x, y)-plane and illuminated by laser light. The scattered wavefield forms the object wave $u(x, y)$. The target of an electronic sensor (e.g., a CCD or a CMOS) used for recording the hologram is located in the (ξ, η)-plane at the distance d from the object. Following the basic principles of holography, the hologram $h(\xi, \eta)$ originates from the interference of the object wave $u(\xi, \eta)$ and the reference wave $r(\xi, \eta)$

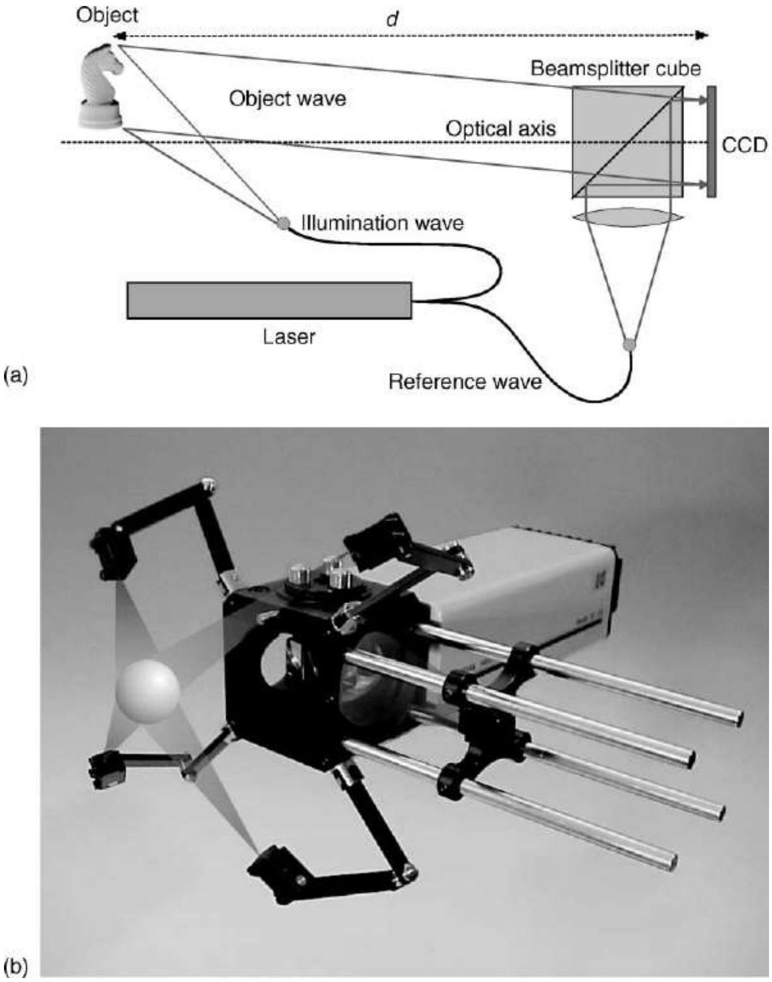

Figure 1 Setup for recording digital holograms. (a) Schematic setup for recording a digital hologram onto a CCD target. (b) Implementation of a holographic camera by mounting a compact holographic interferometer in front of a CCD-target (4 illumination directions are used). Reproduced from Seebacher S, Osten W, Baumbach Th and Jüptner W (2001) The determination of material parameters of microcomponents using digital holography. *Optics and Lasers in Engineering* 36(2): 103–126, copyright © 2001, with permission from Elsevier.

in the (ξ, η)-plane:

$$\begin{aligned} h(\xi,\eta) &= |u(\xi,\eta) + r(\xi,\eta)|^2 \\ &= rr^* + ru^* + ur^* + uu^* \end{aligned} \qquad [1]$$

The transformation of the intensity distribution into a gray value distribution, that is stored in the image memory of the computer, is considered by the characteristic function t of the sensor. This function is generally only approximately linear:

$$T = t[h(\xi,\eta)] \qquad [2]$$

Because the sensor has a limited spatial resolution, the spatial frequencies of the interference fringes in the hologram plane – the so-called micro-interferences – have to be considered. The fringe spacing g and the spatial frequency f_x, respectively, are determined by the angle β between the object and the reference wave (Figure 2b):

$$g = \frac{1}{f_x} = \frac{\lambda}{2\sin(\beta/2)} \qquad [3]$$

with λ as the wavelength. If we assume that the discrete sensor has a pixel pitch (distance between adjacent pixels) $\Delta\xi$, the sampling theorem requires at least two pixels per fringe for a correct reconstruction of the periodic function:

$$2\Delta\xi < \frac{1}{f_x} \qquad [4]$$

Consequently, we obtain for small angles β:

$$\beta < \frac{\lambda}{2\Delta\xi} \qquad [5]$$

Modern high-resolution CCD or CMOS chips have a pitch, $\Delta\xi$, of about 4 μm. In this case, a maximum angle between the reference and the

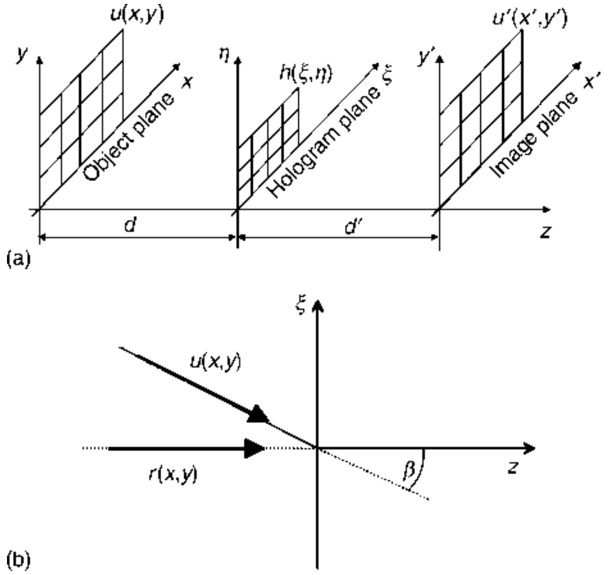

Figure 2 Geometry for recording and reconstruction of digital holograms. (a) Schematic setup for Fresnel holography. (b) Interference between the object and reference wave in the hologram plane.

object wave of only 4° is acceptable. The practical consequence of the restricted angle resolution in digital holography is a limitation of the effective object size that can be stored holographically by an electronic sensor. However, this is only a technical handicap. Larger objects can be placed at a sufficient distance from the hologram, or reduced optically by imaging with a negative lens.

The reconstruction is done by illuminating the hologram with a so-called reconstruction wave $c(\xi, \eta)$:

$$u'(x', y') = t[h(\xi, \eta)]c(\xi, \eta) \qquad [6]$$

The hologram, $t[h(\xi, \eta)]$, diffracts the wave, $c(\xi, \eta)$, in such a way that images of the object wave are reconstructed. Usually four terms are reconstructed if the wave, $u'(x', y')$, propagates in space. An assumed linear characteristic function, $t(h) = \alpha h + t_0$, delivers:

$$u' = Tc = t(h)c \qquad [7]$$

and

$$u' = \alpha \lfloor cu^2 + cr^2 + cur^* + cru^* \rfloor + ct_0 \qquad [8]$$

with two relevant image terms, $[cur^*]$ and $[cru^*]$. The appearance of the image terms depends on the concrete shape of the reconstruction wave. Usually the reference wave, $c = r$, or its conjugated version, $c = r^*$, is applied. In the case of the conjugated reference wave, a direct or real image will be reconstructed due to a converging image wave that can be imaged on a screen at the location of the original object.

However, in digital holography the reconstruction of the object wave in the image plane, $u'(x', y')$, is done by numerical simulation of the physical process, as shown in Figure 3a. The reconstruction wave with a well-known shape equal to the reference wave, $r(\xi, \eta)$, propagates through the hologram, $h(\xi, \eta)$. Following the Huygens principle, each point, $P(\xi, \eta)$, on the hologram, acts as the origin of a spherical elementary wave. The intensity of these elementary waves is modulated by the transparency, $h(\xi, \eta)$. In a given distance, $d' = d$, from the hologram, a sharp real image of the object can be reconstructed as the superposition of all elementary waves. For the reconstruction of a virtual image, $d' = -d$ is used.

Consequently, the calculation of the wavefield, $u'(x', y')$, in the image plane starts with the pointwise multiplication of the stored and transformed intensity values, $t[h(\xi, \eta)]$, with a numerical model of the reference wave, $r(\xi, \eta)$. If the hologram is assumed to be uniformly illuminated by a normally incident, monochromatic plane wave of amplitude 1 the reference wave can be modeled by $r(\xi, \eta) = 1$. After the multiplication the resulting field in the hologram plane is propagated in free space. In the distance, d', the diffracted field, $u'(x', y')$, can be found by solving the well-known Rayleigh–Sommerfeld diffraction formula, that is also known as the

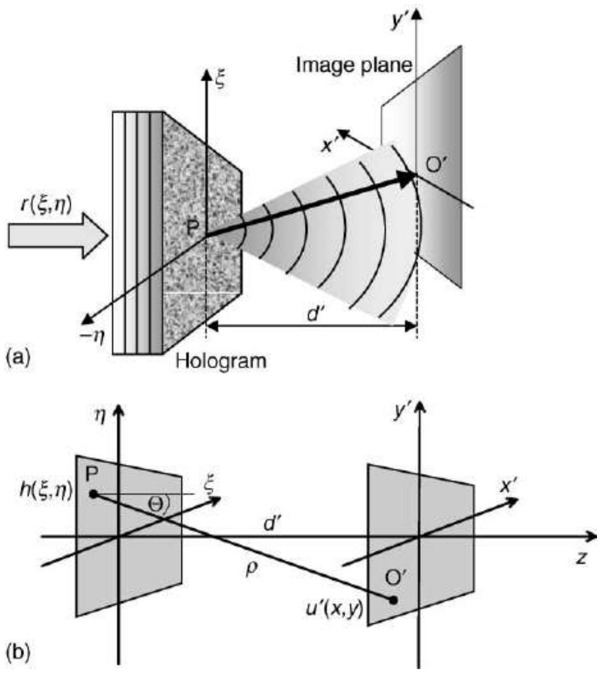

Figure 3 Reconstruction of digital holograms. (a) Principle of wave front reconstruction. (b) Light propagation by diffraction (Huygens–Fresnel principle).

Huygens–Fresnel principle:

$$u'(x',y')=\frac{1}{\mathrm{i}\lambda}\iint_{-\infty\cdots\infty} t[h(\xi,\eta)]r(\xi,\eta)\times\frac{\exp(\mathrm{i}k\rho)}{\rho}\cos\theta\,\mathrm{d}\xi\,\mathrm{d}\eta \qquad [9]$$

with

$$\rho(\xi-x',\eta-y')=\sqrt{d'^2+(\xi-x')^2+(\eta-y')^2} \qquad [10]$$

as the distance between a point $O'(x',y',z=d')$ in the image plane and a point $P(\xi,\eta,z=0)$ in the hologram plane, and

$$k=\frac{2\pi}{\lambda} \qquad [11]$$

as the wave number. The obliquity factor, $\cos\theta$, represents the cosine of the angle between the outward normal and the vector joining P to O′. This term is given as:

$$\cos\theta=\frac{d'}{\rho} \qquad [12]$$

and therefore the Huygens–Fresnel principle can be rewritten:

$$u'(x',y')=\frac{d'}{\mathrm{i}\lambda}\iint_{-\infty\cdots\infty} t[h(\xi,\eta)]r(\xi,\eta)\frac{\exp(\mathrm{i}k\rho)}{\rho^2}\mathrm{d}\xi\,\mathrm{d}\eta \qquad [13]$$

This numerical reconstruction delivers the complex amplitude of the wavefront. Consequently, the phase distribution, $\phi(x',y')$, and the intensity, $I(x',y')$, can be calculated directly from the reconstructed complex function, $u'(x',y')$:

$$\phi(x',y')=\arctan\frac{\mathrm{Im}|u'(x',y')|}{\mathrm{Re}|u'(x',y')|} \qquad [-\pi,\pi] \qquad [14]$$

$$I(x',y')=u'(x',y')u'^{*}(x',y') \qquad [15]$$

The direct approach to the phase yields several advantages for imaging and metrology applications that are discussed later.

Reconstruction Principles in Digital Holography

Different numerical reconstruction principles have been investigated: the Fresnel approximation, the convolution approach, the lens-less Fourier approach, the phase shifting approach, and the phase-retrieval approach. In this section the main techniques are briefly described.

The Fresnel Approximation

If the distance, d, between the object and hologram plane, and equivalently, $d'=d$, between the hologram and image plane, is large compared to $(\xi-x')$ and $(\eta-y')$, then the denominator of eqn [13] can be replaced by d'^2 and the parameter ρ in the numerator can be approximated by a binomial expansion for the square root (10) where only the first two terms are considered:

$$\rho\approx d'\left[1+\frac{(\xi-x')^2}{2d'^2}+\frac{(\eta-x')^2}{2d'^2}\right] \qquad [16]$$

The resulting expression for the field at (x',y') becomes

$$u'(x',y')=\frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda d'}\iint_{-\infty\cdots\infty} t[h(\xi,\eta)]r(\xi,\eta)\times\exp\left[\frac{\mathrm{i}k}{2d'}\left\{(\xi-x')^2+(\eta-y')^2\right\}\mathrm{d}\xi\,\mathrm{d}\eta\right] \qquad [17]$$

Equation [17] is a convolution integral that can be expressed as

$$u'(x',y)=\iint_{-\infty\cdots\infty} t[h(\xi,\eta)]r(\xi,\eta)H\times(\xi-x',\eta-y')\mathrm{d}\xi\,\mathrm{d}\eta \qquad [18]$$

with the convolution kernel:

$$H(x',y')=\frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda d'}\exp\left[\frac{\mathrm{i}k}{2d'}(x'^2+y'^2)\right] \qquad [19]$$

Another notation of eqn [17] is found if the term $\exp[(\mathrm{i}k/2d')(x'^2+y'^2)]$ is taken out of the integral:

$$u'(x',y')=\frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda d'}\exp\left[\mathrm{i}\frac{k}{2d'}(x'^2+y'^2)\right]\iint_{-\infty\cdots\infty}\times\left\{t[h(\xi,\eta)]r(\xi,\eta)\exp\left[\mathrm{i}\frac{k}{2d'}(\xi^2+\eta^2)\right]\right\}\times\exp\left[-\mathrm{i}\frac{2\pi}{\lambda d'}(\xi x'+\eta y')\right]\mathrm{d}\xi\,\mathrm{d}\eta \qquad [20]$$

or

$$u'(x',y')=\frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda d'}\exp\left[\mathrm{i}\frac{\mathrm{k}}{2d'}(x'^2+y'^2)\right]\times\mathrm{FT}_{\lambda d'}\left\{t[h(\xi,\eta)]r(\xi,\eta)\exp\left[\mathrm{i}\frac{k}{2d'}(\xi^2+\eta^2)\right]\right\} \qquad [21]$$

where $\mathrm{FT}_{\lambda d'}$ indicates the 2-dimensional Fourier transform that has been modified by a factor $1/(\lambda d')$. Equation [21] makes clear that the diffracted wavefront consists of the Fourier transform (FT) of the digitally stored hologram, $t[h(\xi,\eta)]$, multiplied by the reference wave, $r(\xi,\eta)$, and the chirp function, $\exp\{(i\pi/\lambda d')(\xi^2+\eta^2)\}$. This Fourier transform is scaled by the constant factor, $1/(id'\lambda)$, and a phase factor that is independent of the processed hologram.

The discrete sampling in digital processing requires the transformation of the infinite continuous integral in eqn [20] into a finite discrete sum. This results in the finite Fresnel transform:

$$u'(m,n) = \sum_{j=0}^{M-1}\sum_{l=0}^{N-1} t[h(j\Delta\xi, l\Delta\eta)]r(j\Delta\xi, l\Delta\eta) \times \exp\left[\frac{i\pi}{d'\lambda}(j^2\Delta\xi^2 + l^2\Delta\eta^2)\right] \times \exp\left\{-i2\pi\left(\frac{jm}{M}+\frac{ln}{N}\right)\right\} \quad [22]$$

where constants and pure phase factors preceding the sums have been omitted. The main parameters are the pixel number $M \times N$ and the pixel pitches $\Delta\xi$ and $\Delta\eta$ in the two directions, which are defined by the used sensor chip.

The discrete phase distribution, $\phi(m,n)$, of the wavefront and the discrete intensity distribution, $I(m,n)$, on the rectangular grid of $M \times N$ sample points can be calculated from the reconstructed complex function, $u'(m,n)$, by using eqns [14] and [15], respectively. The wrapped phase distribution is computed directly without the need of additional temporal or spatial phase modulation such as phase shifting or spatial heterodyning.

For interferometric applications with respect to displacement and shape measurement the phase difference, $\delta(m,n)$, of two reconstructed wave fields, $u'_1(m,n)$ and $u'_2(m,n)$, needs to be calculated. The algorithm can be reduced to the following equation:

$$\delta(m,n) = \phi_1(m,n) - \phi_2(m,n) \quad [23]$$

Figure 4 shows the principle of displacement measurement by digital holography. Besides the investigation of surface displacements in the region of the wavelength, digital holography can be applied advantageously for the measurement of the shape of complex objects. In Figure 5, a turbine blade was investigated with the 2-wavelength contouring method. Both, the digitally reconstructed mod 2π-phase and its demodulated version after-phase unwrapping, are presented. Phase unwrapping is necessary in digital holography, since the fringe-counting problem still remains. However, there are some efficient approaches to overcome the difficulties of this process.

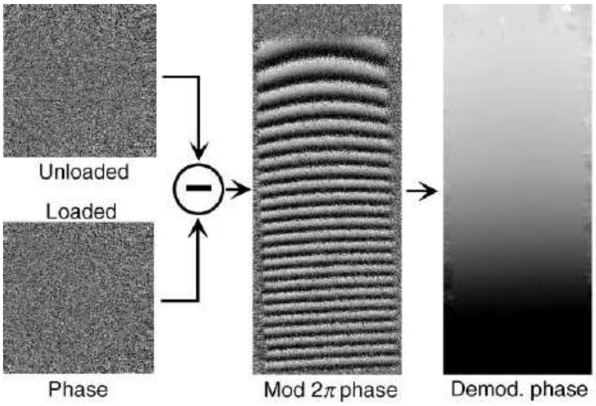

Figure 4 The principle of digital holographic interferometry demonstrated on an example of a loaded small beam. The interference phase is the result of the subtraction of the two phases which correspond to the digital holograms of both object states.

Following eqn [22] the pixel spacing in the reconstructed image is

$$\Delta x' = \frac{d'\lambda}{M\Delta\xi} \quad \text{and} \quad \Delta y' = \frac{d'\lambda}{N\Delta\eta} \quad [24]$$

Additional to the real image, a blurred virtual image and a bright dc-term, the zero-order diffracted field, is reconstructed. This term can effectively be eliminated by preprocessing the stored hologram or the different terms can be separated by using the off-axis instead of the in-line technique. However, the spatial separation between the object and the reference field requires a sufficient space-bandwidth product of the used CCD-chip, as discussed in the section on direct phase reconstruction above.

Numerical Reconstruction by the Convolution Approach

In the section above, we have already mentioned that the connection between the image term, $u'(x',y')$, and the product, $t[h(\xi,\eta)]r(\xi,\eta)$, can be described by a linear space-invariant system. The diffraction formula [17] is a superposition integral with the impulse response:

$$H(x'-\xi, y'-\eta) = \frac{\exp(ikd')}{i\lambda d'} \times \exp\left[\frac{ik}{2d'}\{(\xi-x')^2+(\eta-y')^2\}\right] \quad [25]$$

A linear space-invariant system is characterized by a transfer function G, that can be calculated as the

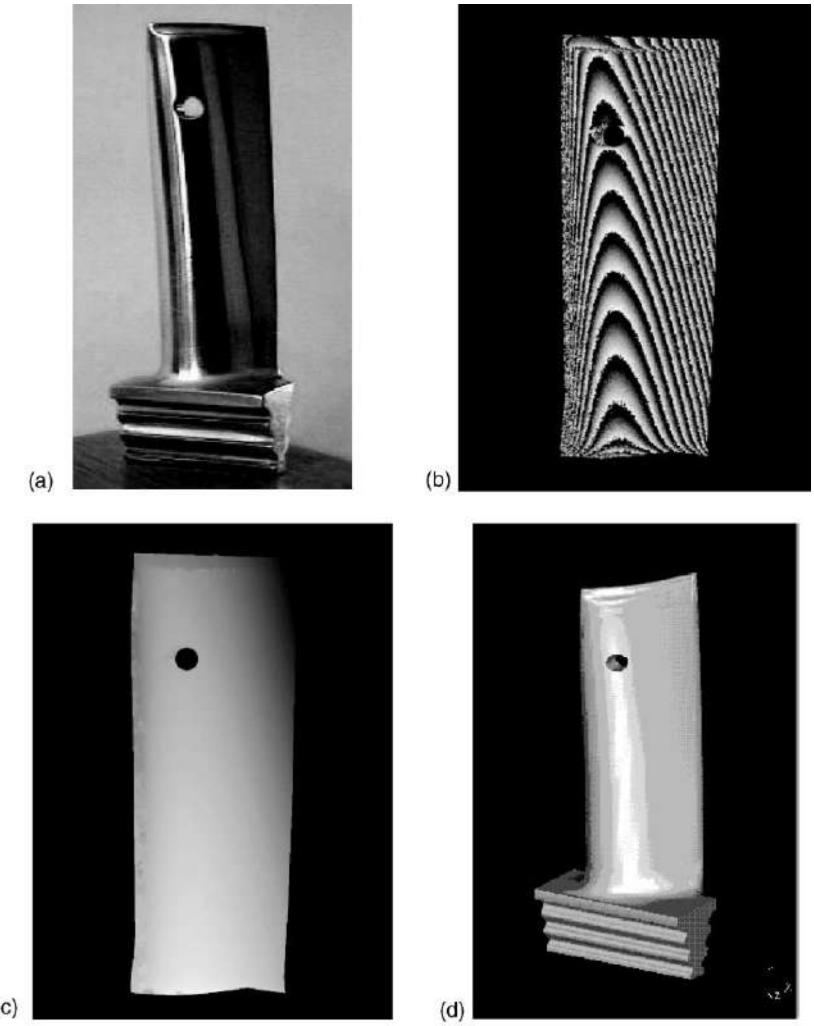

Figure 5 Phase reconstruction by digital holographic interferometry on an example of 2-wavelength contouring of a turbine blade. (a) Image of the turbine blade. (b) Mod-2π phase distribution. (c) Unwrapped phase distribution. (d) CAD reconstruction of the turbine blade.

Fourier-transform of the impulse response:

$$G(f_x, f_y) = \mathrm{FT}\{H(\xi, \eta, x', y')\} \qquad [26]$$

with the spatial frequencies f_x and f_y. Consequently, the convolution theorem can be applied, which states that the Fourier transform of the convolution, $t[h(\xi,\eta)]r(\xi,\eta)$, with H is the product of the individual Fourier transforms, $\mathrm{FT}\{t[h(\xi,\eta)]r(\xi,\eta)\}$ and $\mathrm{FT}\{H\}$. Thus $u'(x', y')$ can be calculated by the inverse Fourier transform of the product of the Fourier-transformed convolution partners:

$$u'(x', y') = \mathrm{FT}^{-1}\{\mathrm{FT}[t(h)r]\mathrm{FT}[H]\} \qquad [27]$$

$$u'(x', y') = [t(h)r] \otimes H \qquad [28]$$

with $\otimes$ as the convolution symbol. The computing effort is comparatively high: two complex multiplications and three Fourier transforms. The main difference to the Fresnel transform is the different pixel size in the reconstructed field:

$$\Delta x' = \Delta\xi \quad \text{and} \quad \Delta y' = \Delta\eta \qquad [29]$$

Numerical Reconstruction by the Lensless Fourier Approach

We have already mentioned that the limited spatial resolution of the sensor restricts the angle resolution of the digital hologram. The sampling theorem requires that the angle between the object beam and the reference beam at any point of the electronic sensor be limited in such a way that the microinterference spacing is larger than double the pixel size. In general, the angle between the reference beam and the object beam varies over the sensor's surface, and so does the maximum spatial frequency. Thus, for most holographic setups, the full spatial bandwidth of the sensor cannot be used. However, it is important to use the entire spatial bandwidth of the sensor because the lateral resolution of the reconstructed image depends on a complete evaluation of the entire information obtained from the sensor.

Even if the speed of digital signal processing is increasing rapidly, algorithms should be as simple and as fast to compute as possible. For the Fresnel approach and the convolution approach, several fast Fourier transforms and complex multiplications are necessary. Therefore, a more effective approach, such as the subsequently described algorithm, seems promising.

The lensless Fourier approach is the fastest and most suitable algorithm for small objects. The corresponding setup is shown in Figure 6. It allows us to choose the lateral resolution in a range from a few microns to hundreds of microns without any additional optics. Each point, (ξ, η), on the hologram is again considered as a source point of a spherical elementary wavefront (Huygen's principle). The intensity of these elementary waves is modulated by $t[h(\xi,\eta)]$ – the amplitude transmission of the hologram. The reconstruction algorithm for lensless Fourier holography is based on the Fresnel reconstruction. Here again, $u(x, y)$ is the object wave in the object plane, $h(\xi,\eta)$ the hologram, $r(\xi,\eta)$ the reference wave in the hologram plane, and $u'(x', y')$ the reconstructed wavefield.

For the specific setup of lensless Fourier holography, a spherical reference wave is used with an origin at the same distance from the sensor as the object itself. In the case that $d' = -d$, $x = x'$, and $y = y'$, both virtual images are reconstructed and the reconstruction algorithm then reads:

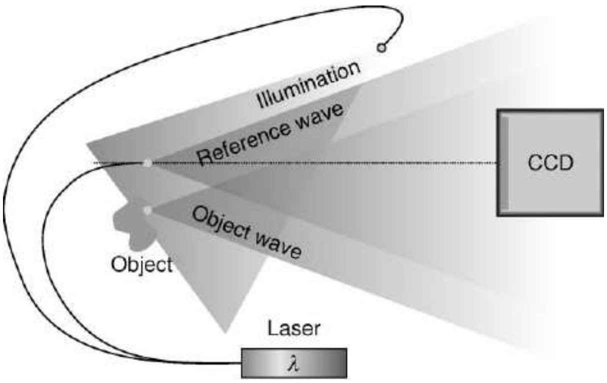

Figure 6 Scheme of a setup for digital lensless Fourier holography of diffusely reflecting objects.

$$u'(x', y') = \frac{\exp(ikd)}{i\lambda d} e^{-i\frac{k}{2d}(x^2+y^2)} \times \mathrm{FT}_{\lambda d'}\left\{t[h(\xi,\eta)]\cdot r(\xi,\eta)\cdot e^{-i\frac{k}{2d}(\xi^2+\eta^2)}\right\} \quad [30]$$

with $\mathrm{FT}_{\lambda d}$ as the 2-dimensional Fourier transformation which has been scaled by a factor $1/\lambda d$. In this recording configuration, the effect of the spherical phase factor associated with the Fresnel diffraction pattern of the object is eliminated by the use of a spherical reference wave, $r(\xi,\eta)$, with the same average curvature:

$$r(x,y) = \text{const.} \times \exp\left(i\frac{\pi}{\lambda d}(\xi^2+\eta^2)\right) \quad [31]$$

This results in a more simple reconstruction algorithm which can be described by

$$u'(x', y') = \text{const.} \times \exp\left(-i\frac{\pi}{\lambda d}(x^2+y^2)\right) \times \mathrm{FT}_{\lambda d}\{h(\xi,\eta)\} \quad [32]$$

Besides the simpler and faster reconstruction algorithm (only one Fourier transformation has to be computed), the Fourier algorithm uses the full space-bandwidth product (SBP) of the sensor chip because it adapts the curvature of the reference wave to the curvature of the object wave.

Influences of Discretization

For the estimation of the lateral resolution in digital holography, three different effects related to discretization have to be considered: averaging, sampling, and the limited sensor size. We assume a quadratic sensor with $N \times M$ pixels of the size $\Delta\xi \times \Delta\xi$. Each pixel has a light-sensitive region with a side length $\gamma\Delta\xi = \Delta\xi_{\mathrm{eff}}$. γ^2 is the so-called fill factor $(0 \leq \gamma \leq 1)$, that indicates the active area of the pixel. The sensor averages the incident light which has to be considered because of possible intensity fluctuations over this area. The continuous expression for the intensity $I(\xi,\eta) = t[h(\xi,\eta)]$, registered by the sensor, has to be integrated over the light-sensitive area. This can be expressed mathematically by the convolution of the incident intensity $I(\xi,\eta)$, with a rectangle function:

$$I^1(\xi,\eta) \propto I \otimes \mathrm{rect}_{\Delta\xi_{\mathrm{eff}},\Delta\xi_{\mathrm{eff}}} \quad [33]$$

The discrete sampling of the light field is modeled by the multiplication of the continuous assumed hologram with the 2-dimensional comb-function:

$$I^2(\xi, \eta) \propto I^1(\xi, \eta) \times \mathrm{comb}_{\Delta\xi.\Delta\xi} \qquad [34]$$

Finally, the limited sensor size requires that the comb-function has to be truncated at the borders of the sensor. This is considered by the multiplication with a 2-dimensional rect-function of the size $N \times \Delta\xi$:

$$I^3(\xi, \eta) \propto I^2 \times \mathrm{rect}_{N\Delta\xi N.\Delta\xi} \qquad [35]$$

The consequences are amplitude distortions, aliasing and speckling that have to be considered in the reconstruction procedure.

Advantages of Digital Holography

Besides the electronic processing and the direct access to the phase, some further advantages recommend digital holography for several applications such as metrology and microscopy:

- The availability of the independently reconstructed phase distributions of each individual state of the object and interferometer, respectively, offers the possibility to record a series of digital holograms with increased load amplitude. In the evaluation process, the convenient states can be compared interferometrically. Furthermore, a series of digital holograms with increasing load can be applied to unwrap the mod 2π-phase temporally. In this method, the total object deformation is subdivided into many measurement steps, in which the phase differences are smaller than 2π. By adding up those intermediate results, the total phase change can be obtained without any further unwrapping. This is an important feature, since it is essential to have an unwrapped phase to be able to calculate the real deformation data from the phase map. Figure 7 shows an example of such a measurement. The left image shows the wrapped deformation phase for a clamped coin which was loaded with heat. The right image shows the temporal unwrapped phase which has been obtained by dividing the total deformation into 85 sub-measurements. Thus, the displacement of the object can be observed almost in real time during the loading process.
- The independent recording and reconstruction of all states also gives a new degree of freedom for optical shape measurement. In the case of multiwavelength contouring, each hologram can be stored and reconstructed independently with its corresponding wavelength. This results in a drastic decrease of aberrations and makes it possible to use larger wavelength differences for the generation of shorter synthetic wavelengths.
- Because all states of the inspected object can be stored and evaluated independently, only seven digital holograms are necessary to measure the 3D-displacement field and shape of an object under test: one hologram for each illumination direction before and after the loading, respectively, and one hologram with a different wavelength (or a different source-point of illumination) which can interfere with one of the other holograms to perform two-wavelength-contouring (or two-source-point-contouring) for shape measurement. If four illumination directions are used, nine holograms are necessary.
- The direct access to all components of the wavefield makes it possible to correct wavefront aberrations effectively by computer.
- The digital hologram containing all information of the wavefront can be transmitted via the internet to any location and can be reconstructed digitally by computer or as an analogue version by a spatial light modulator. Consequently, remote access to wavefronts, that are generated at distant places, is possible.
- Finally, the possibility to miniaturize complex holographic sensor setups and to use this method for remote comparative interferometry, makes digital holography a versatile tool for the solution of numerous inspection and measurement problems.

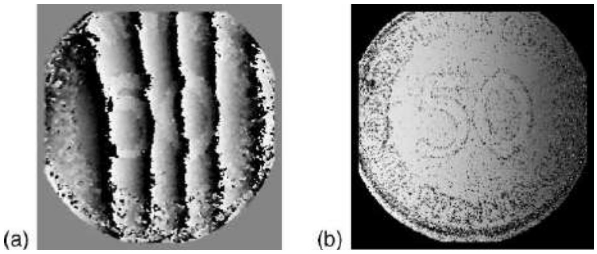

Figure 7 Temporal unwrapping by digital holography: (a) Wrapped phase of a thermal loaded coin. (b) Temporally unwrapped phase.

See also

Diffraction: Fresnel Diffraction. **Fourier Optics**. **Holography, Techniques:** Holographic Interferometry.

Further Reading

Butters JN and Leendertz JA (1971) Holographic and video techniques applied to engineering measurement. *Journal of Measurement Control* 4: 349–354.

Creath K (1993) Temporal phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*. Bristol and Philadelphia: IOP Publishing Ltd.

Demetrakopoulos T and Mittra T (1974) Digital and optical reconstruction of images from suboptical diffraction patterns. *Applied Optics* 13: 665–670.

Denisyuk YuN (1962) Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Doklady Akademii Nauk SSSR* 144: 1275–1279.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceedings of the Royal Society, A* 197: 454–487.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw Hill Comp. Inc.

Goodman JW and Lawrence RW (1967) Digital image formation from electronically detected holograms. *Applied Physics Letters* 11: 77–79.

Kolenovic E, Osten W, Klattenhoff R, Lai S, Kopylow C and Jüptner W (2003) Miniaturized digital holography sensor for distal three-dimensional endoscopy. *Applied Optics* 42(25): 5167–5172.

Kreis T (1996) *Holographic Interferometry*. Berlin: Akademie Verlag.

Kreis Th and Jüptner W (1997) Principles of digital holography. In: Jüptner W and Osten W (eds) *Proc. Fringe'97, 3rd International Workshop on Automatic Processing of Fringe Patterns*, pp. 353–363. Berlin: Akademie Verlag.

Kreis Th and Jüptner W (1997) The suppression of the dc-term in digital holography. *Optical Engineering* 36: 2357–2360.

Kujawinska M (1993) Spatial phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*. Bristol and Philadelphia: IOP Publishing Ltd.

Lee W-H (1978) Computer generated holograms: techniques and applications. In: Wolf E (ed.) *Progress in Optics*, vol. XVI, pp. 121–230. Amsterdam: North Holland.

Leith EN and Upatnieks J (1961) New techniques in wavefront reconstruction. *JOSA* 51: 1469–1473.

Osten W (2000) Holography and virtual 3D-testing. Proc. International Berlin Workshop HoloMet 2000, New Prospects of Holography and 3D-Metrology, *Strahltechnik* 14: 14–17.

Osten W (2003) Active metrology by digital holography. *Proceedings of SPIE* 4933: 96–110.

Osten W and Kujawinska M (2000) Active phase measuring metrology. In: Rastogi PK and Inaudi D (eds) *Trends in Optical Nondestructive Testing and Inspection*, pp. 45–69. Amsterdam: Elsevier Science B.V.

Osten W, Jüptner W and Seebacher S (1998) *The Qualification of Large Scale Approved Measurement Techniques for the Testing of Microcomponents*. Proc. 18th Symposium on Seebacher, Experimental Mechanics of Solids, Jachranka, pp. 43–55.

Osten W, Baumbach T, Seebacher S and Jüptner W (2001) Remote shape control by comparative digital holography. *Proc. Fringe 2001*, Bellingham: Elsevier Science, pp. 373–382.

Osten W, Seebacher S and Jüptner W (2001) The application of digital holography for the inspection of microcomponents. *Proceedings of SPIE* 4400: 1–15.

Osten W, Baumbach T and Jüptner W (2002) Comparative digital holography. *Optics Letters* 27(20): 1764–1766.

Pedrini G, Schedin S and Tiziani HJ (2001) Aberration compensation in digital holographic reconstruction of microscopic objects. *Journal of Modern Optics* 48: 1035–1041.

Powell RL and Stetson K (1965) Interferometric vibration analysis of three-dimensional objects by wavefront reconstruction. *JOSA* 55: 1593–1598.

Schnars U (1994) Direct phase determination in hologram interferometry with use of digitally recorded holograms. *Journal of the Optical Society of America A* 11: 2011–2015.

Schnars U and Jüptner W (1994) Direct recording of holograms by a CCD target and numerical reconstruction. *Applied Optics* 33(2): 179–181.

Seebacher S (2001) Application of digital holography for 3d-shape and deformation measurement of micro components. PhD thesis, University of Bremen.

Seebacher S, Osten W and Jüptner W (1997) 3-D deformation analysis of micro-components using digital holography. *Proceedings of SPIE* 3098: 382–391.

Seebacher S, Osten W, Baumbach Th and Jüptner W (2001) The determination of material parameters of microcomponents using digital holography. *Optics and Lasers in Engineering* 36(2): 103–126.

Takeda M, Taniguchi K, Hirayama T and Kogho H (1996) Single transform Fourier–Hartley fringe analysis for holographic interferometry. In: Füzessy Z, Jüptner W and Osten W (eds) *Simulation and Experiment in Laser Metrology*, pp. 67–73. Berlin: Akademie Verlag.

Wagner C, Seebacher S, Osten W and Jüptner W (1999) Digital recording and numerical reconstruction of lensless Fourier holograms in optical metrology. *Applied Optics* 38(22): 4812–4820.

Wagner C, Osten W and Seebacher S (2000) Direct shape measurement by digital wavefront reconstruction and multi-wavelength contouring. *Optical Engineering* 39(1): 79–85.

Yamaguchi I and Zhang T (1979) Phase-shifting digital holography. *Optical Letters* 22: 1268–1270.

Yaroslavsky LP and Merzlyakov NS (1980) *Methods of Digital Holography*. New York: Consultants Bureau.

Zhang Y, Pedrini G, Osten W and Tiziani H (2003) Image reconstruction for in-line holography with the Yang-Gu algorithm. *Applied Optics* 42(32): 6452–6457.

Holographic Interferometry

P Rastogi, Swiss Federal Institute of Technology,
Lausanne, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The development of holographic interferometry has redefined our perception of optical interferometry. The unique property of holographic interferometry to bring a wavefront, generated at some earlier time, stored in a hologram and released at a later time, to interfere with a comparison wavefront, has made possible the interferometric comparison of a rough surface, which is subject to stress, with its normal state. Assuming the surface deformations to be very small, the interference of the two speckled wavefronts forms a set of interference fringes that are indicative of the amount of displacement and deformation undergone by the diffuse object.

Holographic interferometry is an important and exciting area of research and development. It has established itself as a highly promising noninvasive technique in the field of optical metrology. Techniques within its folds for the measurement of static and dynamic displacements, topographic contours, and flow fields have been developed and demonstrated with success in a wide range of problems. Significant and extensive contributions to the development of holographic interferometry have been realized both from the theoretical and experimental perspectives. A wide range of procedures have been developed which make it not only possible to measure surface displacements and deformations of engineering structures to an accuracy of a fraction of a micrometer but also to detect material flaws and inhomogeneities having escaped the manufacturing process. The methodology of quantitative analysis of holographic interferograms has undergone extensive development during the last decade. This article provides a brief review of the holographic techniques with emphasis placed on bringing out its relevance in experimental mechanics and nondestructive testing. The series of books given in the Further Reading section at the end of this article treat the field in detail.

Basic Methods in Holographic Interferometry

The ability of holography to store a wavefront and release it for reconstruction at a later time offers access to the possibility to compare wavefronts, which have albeit existed at different times. There are several schemes that have been established to obtain the interferometric comparison of the wavefronts.

Double-Exposure Holographic Interferometry

In frozen form of wavefront comparison, two holograms of the object are recorded on a photographic plate. The first hologram is made with the object in its initial, unstressed state and the second hologram is recorded with the object in its final, stressed state. Upon reconstruction, the hologram releases the two stored wavefronts, one corresponding to the object in its unstressed state and the other corresponding to the object in its stressed state (Figure 1). The virtual image is overlaid by a set of

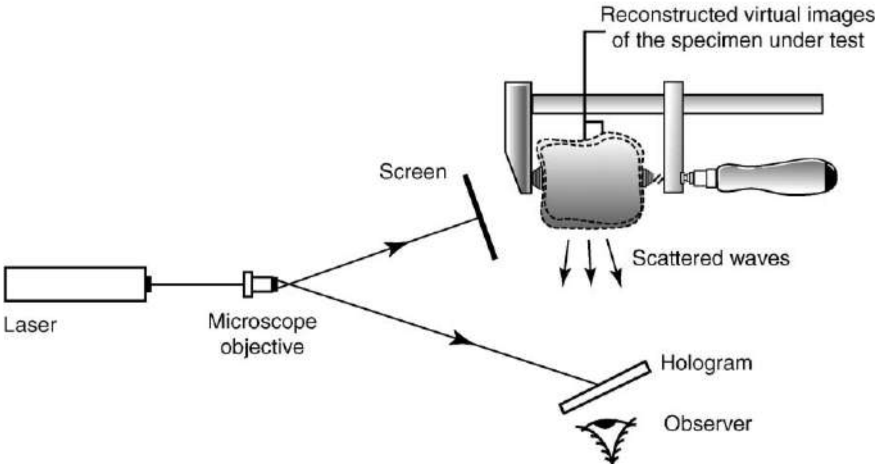

Figure 1 Layout for observing a double exposure holographic interferogram.

bright and dark interference fringes, which are due to the displacement of the object between the two exposures.

The fringe pattern is permanently stored on the holographic plate. The interference fringes denote the loci of points that have undergone an equal change of optical path between the light source and the observer. The image intensity of the holographic interferogram is

$$I(x,y) = I_0(x,y)\{1 + V(x,y)\cos(\varphi(x,y) - \varphi'(x,y))\} \qquad [1]$$

where $I_0(x,y)$ is the background intensity, $V(x,y)$ is the fringe contrast, and φ and φ' are the phases of the waves from the object in its initial and deformed states, respectively.

A bright fringe is produced whenever

$$\varphi(x,y) - \varphi'(x,y) = 2n\pi \qquad n = 0,1,2,\ldots \qquad [2]$$

where n is the fringe number. The wave scattered by a rough surface shows rapid variations of phase, which have little correlation across the wavefront. The phase difference varies randomly over the ray directions contained within the aperture of the viewing system, which would normally prevent the observation of fringes with large aperture optics near such points. However, a surface may exist at a distance from the object at which the value of phase difference is stationary over the cone of ray pairs defined by the viewing aperture. Interference fringes are localized in the region where the variation in phase difference is minimum over the range of viewing directions. This approach has been widely used to compute fringe localization for any object displacement and for any arbitrary illumination and observation geometry.

Real-Time Holographic Interferometry

A hologram is made of an arbitrarily shaped rough surface. After development, the hologram is placed back in exactly the same position. Upon reconstruction, the hologram produces the original wavefront. A person looking through the hologram sees a superposition of the original object and its reconstructed image (Figure 2). The object wave interferes with the reconstructed wave to produce a dark field due to destructive interference. If the object is now slightly deformed, interference fringes are produced which are related to the change in shape of the object. A dark fringe is produced, whenever:

$$\varphi(x,y) - \varphi'(x,y) = 2n\pi \qquad n = 0,1,2,\ldots \qquad [3]$$

The method is very useful for determining the direction of the object displacement, and for compensating on the interferogram the influence of the rigid body motion of the object when subjected to a stress field.

The optical setup for real-time holography uses an immersion tank, which is mounted on a universal stage and contains the photographic plate. The holographic plate is followed by a vidicon camera, which visualizes the interferograms directly onto the TV monitor screen connected to a video-recording tape to memorize the information. The use of thermoplastic plates as recording material has gained in importance in recent years for implementing real-time and double exposure holographic interferometry setups, as it can be processed rapidly *in situ* using electrical and thermal processing.

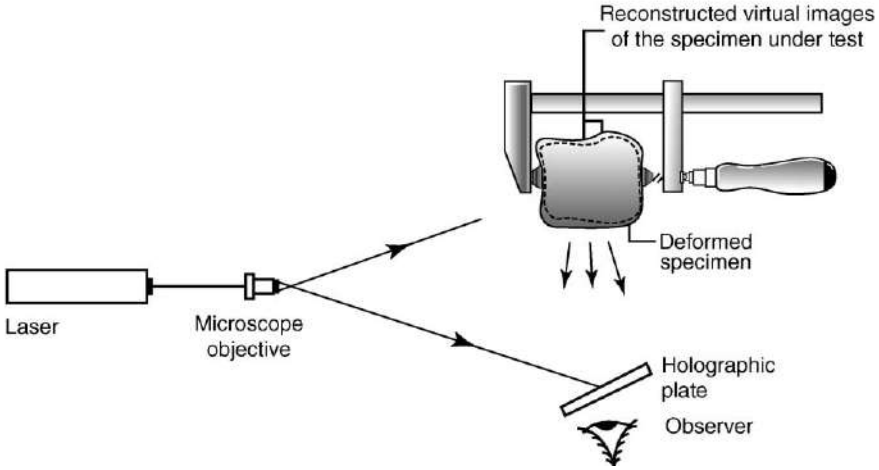

Figure 2 Layout for observing a holographic interferogram in real-time.

The plate is also cost-effective as it is erasable and reusable at least several hundred times. In thermoplastic recording the information is recorded as a thickness variation corresponding to a charge intensity pattern deposited on the thermoplastic layer. Self-developing or photorefractive crystals are other interesting alternatives to the use of film as the recording medium.

Digital Holographic Interferometry

The method consists of recording two holograms corresponding to undeformed and deformed states of an object using a CCD camera. These images are stored in a digital image processing system. The two holograms are then reconstructed separately by numerical methods from the digitaly stored holograms. The advantage of digital holography is that it allows for computing both the intensity and the phase of a holographically stored wavefront during the numerical reconstruction. The contours of constant optical phase change, due to deformation, are obtained by subtracting the reconstructed phases of the undeformed from the deformed object wavefield.

A hologram can be recorded on a CCD as long as the sampling theorem is fulfilled; that is to say, each period of the spatial variation of the hologram intensity is sampled by at least two pixels of the CCD array. The condition imposes a limit on the angle that the object beam can make with the reference beam, which must necessarily be small given that the resolution of CCDs is relatively low. The applications of the method are thus limited mainly by the pixel size in CCD array. The maximum spatial frequency $f_{\max}$ that must be resolved by the recording medium is determined by the maximum angle $\Delta\theta_{\max}$ between the reference and the object wave:

$$f_{\max} = \frac{2}{\lambda}\sin\frac{\Delta\theta_{\max}}{2} \qquad [4]$$

where λ is the wavelength of the laser source. Since the CCD cameras have resolutions of around 100 lines/mm, the maximum angle between the object and reference waves is limited to a few degrees.

Calculation of Phase Change on Object Loading

A diffusely reflecting object is deformed such that a point P on the surface moves to P′ as shown in Figure 3. In double-exposure or real-time holographic interferometry the two wavefronts related to the states of a surface before and after deformation are reconstructed simultaneously and compared. The change of optical path length due to displacement d is

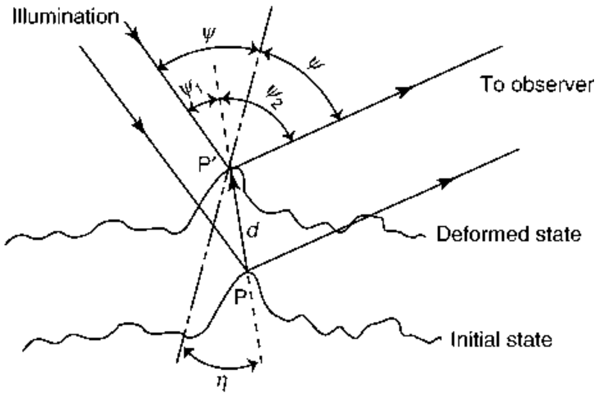

Figure 3 Calculation of the change in optical phase due to surface displacement.

$$\Delta\varphi = \frac{2\pi d}{\lambda}(\cos\psi_1 + \cos\psi_2) \qquad [5]$$

where ψ_1 and ψ_2 are the angles which the illumination and observation waves make with respect to the direction of displacement; eqn [5] can be written in the form:

$$\Delta\varphi = \frac{2\pi d}{\lambda}\cos\eta\cos\psi \qquad [6]$$

where ψ is the angle of bisection between the incident and viewing directions, and η is the angle which the bisector makes with the direction of displacement. The term $d\cos\eta$ being the resolved part of the displacement PP′ in the direction of the bisector, implies that the fringe pattern provides the measure of the displacement component along the bisector of the angle between the incident and viewing directions.

Digital Phase Measurement

The last decade has seen a rapid development of techniques for the automatic and precise reconstruction of phases from fringe patterns. These are based on the concepts of fringe tracking, Fourier transform, carrier frequency, and phase shift interferometry. Quantitative measurement of wavefront phase, for example, can be obtained in a few seconds using phase shift interferometry. With this technique one records a series of holograms by introducing artificially known steps of phase differences in the interference image in eqn [1]:

$$I_i(x,y) = I_0(x,y)\{1 + V(x,y)\cos(\Delta\varphi(x,y) + \Delta\psi_i)\}; \quad i = 0, 1, 2, \ldots \qquad [7]$$

where $\Delta\psi_i$ is the phase shift which is produced, for example, by shifting the phase of the reference wave. A minimum of three intensity patterns is required to calculate the phase at each point on the object surface. Known phase shifts are commonly introduced using a piezoelectric transducer, which serves to shorten or elongate the reference beam path by a fraction of wavelength; the phase-shifter is placed in one arm of the interferometer. Numerous phase shifting techniques have been developed and which can be incorporated to holographic interferometry to generate a phase map, $\Delta\varphi(x, y)$, corresponding to the object displacement. The calculated phase $\Delta\varphi(x, y)$ is independent of the terms $I_0(x, y)$ and $V(x, y)$, and considerably reduces the dependence of the accuracy of measurements on the fringe quality.

The introduction of phase shift interferometry not only provides accurate phase measurements but also eliminates the problem of phase sign ambiguity of the interference fringes. Dynamic phase shifting techniques allow for producing time sequences of deformation maps for studying objects subjected to time-varying loads. The concept is based on considering the normal phase variations caused by a dynamic phenomenon as equivalent to the phase shifting that is introduced in a holographic interferometer in order to perform the phase evaluation task. This method extends the possibility of applying the method to continuous deformation measurements.

Basic Interferometers for Static Displacement Measurement

Equation [6] shows that observation along a single direction yields information only about the resolved part of the surface displacement in one particular direction. The approach to measuring three-dimensional vector displacement would be to record holograms from three different directions in order to obtain three independent resolved components of displacement. The computation of the resulting system of equations would then allow for determining the complete displacement vector.

Out-of-Plane Displacement Component Measurement

The out-of-plane component of displacement is usually measured by illuminating and viewing the object surface in a near normal direction. The fringe equation corresponding to the line of sight displacement component is given by

$$w = \frac{n\lambda}{2} \qquad [8]$$

where w is the out-of-plane component of displacement. An example of a fringe pattern depicting the out-of-plane displacements in a rectangular aluminum plate, clamped along the edges and drawn out at the center by means of a bolt, subjected to three-point loading, is shown in Figure 4.

In-Plane Displacement Component Measurement

The in-plane component of displacement can be measured by illuminating the object along two directions symmetric to the surface normal and observing along a common direction. This is schematized in Figure 5. The in-plane displacements, in a direction containing the two beams, are mapped as a moiré between the interferograms due to each illumination beam. The moiré appears as a family of fringes:

$$u = \frac{n\lambda}{2\sin\theta_e} \qquad [9]$$

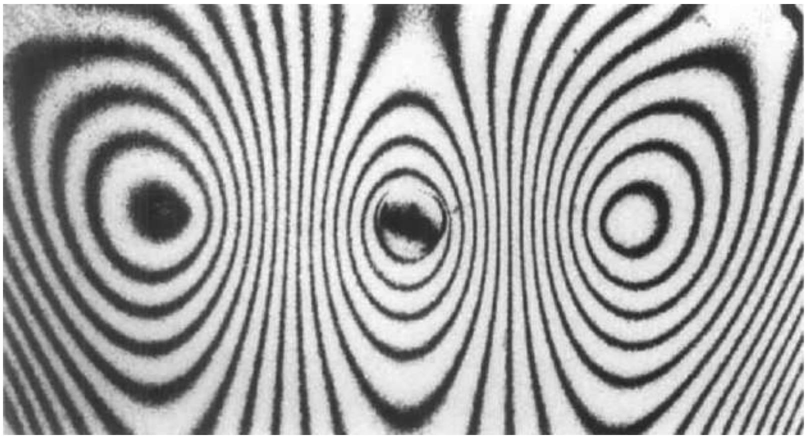

Figure 4 Example of interferogram corresponding to the whole field distribution of out-of-plane component of displacement, *w*, on an aluminum plate subjected to three-point bending.

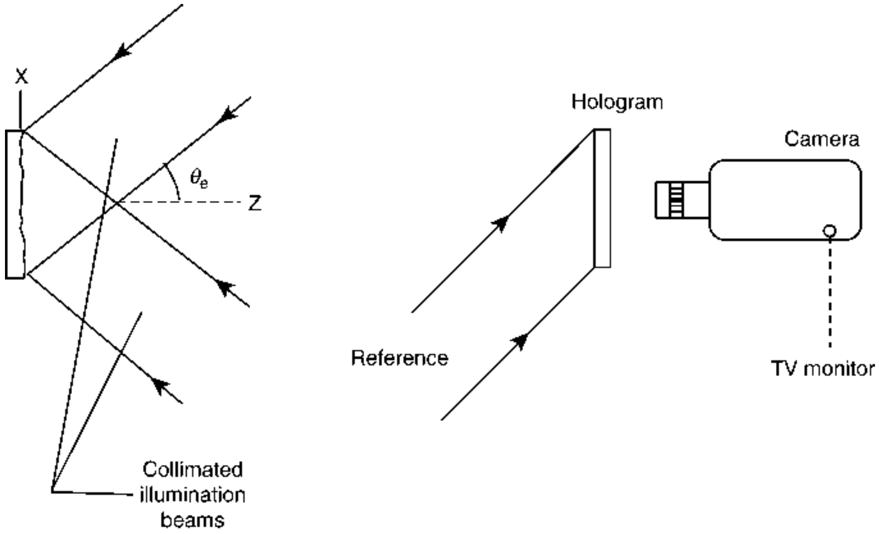

Figure 5 Schematic of the geometrical configuration for measuring the in-plane component of displacement, *u*.

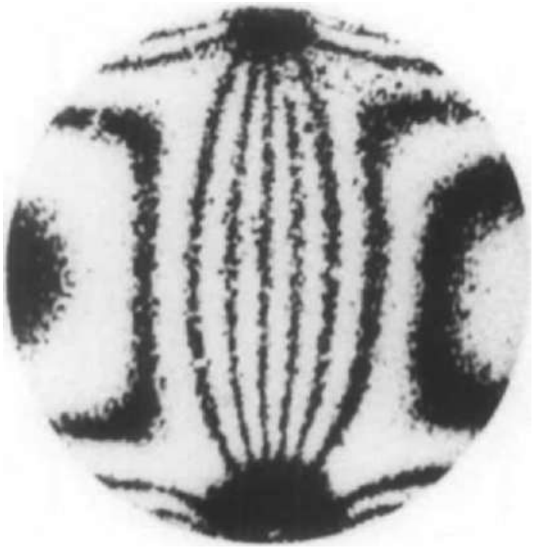

Figure 6 Fringe contours showing the in-plane displacement response of a disk under diametral compression.

where *u* is the displacement component in the *x*-direction and θ_e is the angle which the illumination beams make with the surface normal. Figure 6 shows the *u* pattern corresponding to a disk under diametral compression. The moiré pattern is obtained after optically filtering the interference image.

Comparative Holographic Interferometry

Another technique of significant interest in nondestructive testing is comparative holography that provides the contours of path variations related to the difference in displacements or shapes of two objects. The method provides information about the resolved part of the difference in displacement vector along the direction bisecting the illumination and observation rays. In the case of illumination and observation normal to the object surface, the fringe equation becomes

$$\Delta w = \frac{n\lambda}{2} \qquad [10]$$

where Δw is the difference in displacement component along *z* direction. The technique provides a tool to compare the mechanical responses of two nominally identical specimens subjected to the same loading and also the shapes of two nominally identical specimens. These features are useful in detecting anomalies in a test specimen with respect to the flaw-free master specimen. One such example is shown in Figure 7a, which displays the presence of a defect in a test specimen. In this case two plates, one flaw-free and the other with a flaw, are subjected to identical loadings. The fringe map in Figure 7b is related to the difference in displacements of test and master specimens subjected to unequal loadings. The presence of flaws is detected in both interferograms.

Basic Interferometers for Dynamic Displacement Measurement

Another important application of holographic interferometry is in the study of the resonant modes of vibration of an object.

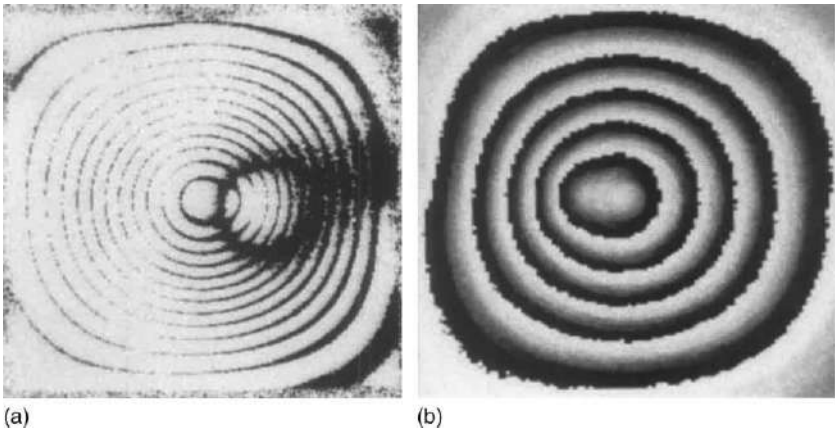

Figure 7 Comparative holography reveals the presence of defects in aluminum plates using (a) identical and (b) unequal loadings.

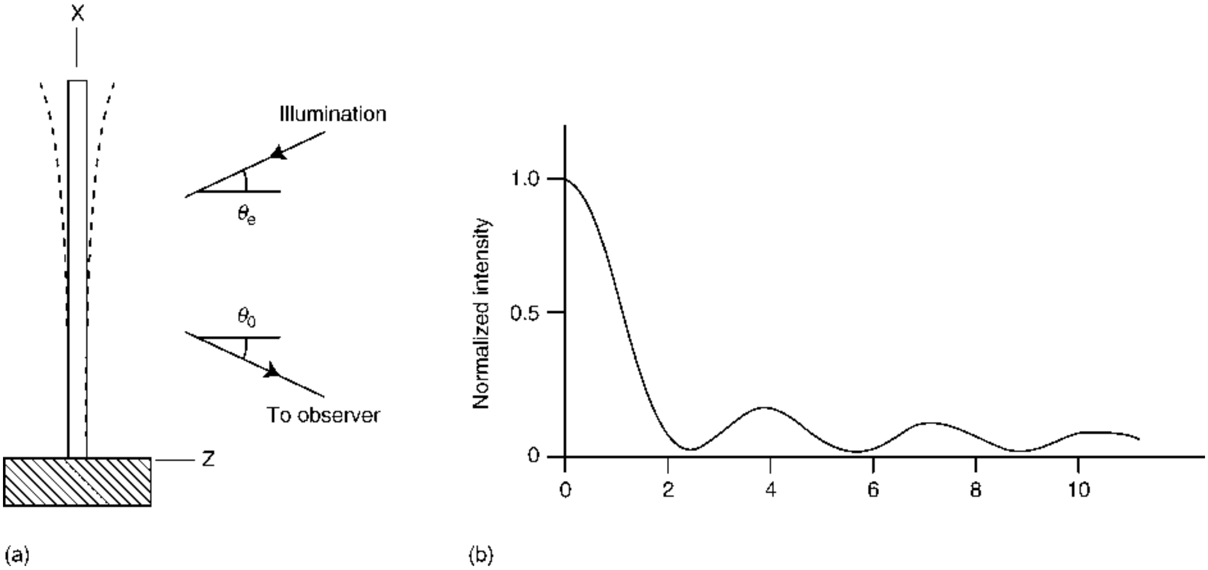

Figure 8 (a) Schematic of the geometrical configuration of a vibrating cantilever beam; (b) plot of fringe function for time-average holography.

Frozen Time-Average Holographic Interferometry

Consider that a hologram is recorded of an object, which is vibrating sinusoidally in a direction normal to its surface (Figure 8a). The exposure time is supposed to be much longer than the period of vibration. The intensity distribution of the image reconstructed by this hologram is

$$I(x,y) = I_o(x,y) J_o^2\left[\frac{2\pi}{\lambda} d(x,y)(\cos\theta_e + \cos\theta_o)\right] \quad [11]$$

where J_o is the zero-order Bessel function of the first kind. The virtual image is modulated by the $J_o^2(\xi)$ function. The dark fringes correspond to the zeros of the function $J_o^2(\xi)$. A plot of the function, shown in Figure 8b, is characterized by a comparatively brighter zero-order fringe, which corresponds to the nodes, a decreasing intensity, and an unequal spacing between the successive zeros. The interferogram in Figure 9 shows the vibrational modes of a helicopter component. The nodes that represent zero motion are clearly seen as the brightest areas in the time average reconstructed pattern.

Real-Time Time-Average Holographic Interferometry

The possibility of studying the response of vibrating objects in real-time extends the usefulness of the technique to identify the resonance states (resonant vibration mode shapes and resonant frequencies) of the object. A single exposure is made of the object in its state of rest. The plate is processed, returned to its original condition and reconstructed. The observer looking through the hologram at the sinusoidally

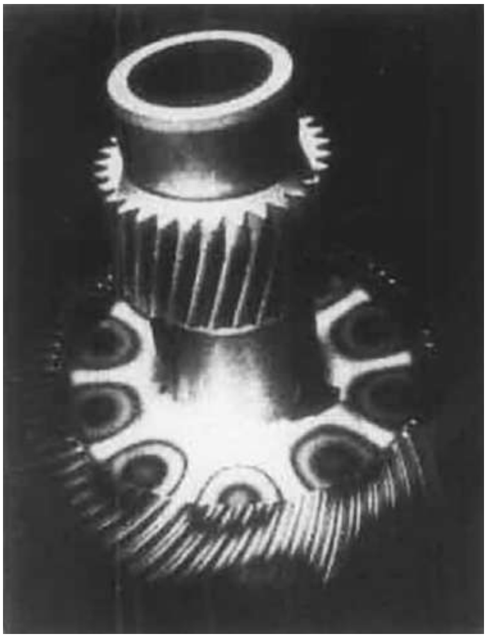

Figure 9 Fringe contours showing the mode shape of a vibrating helicopter component.

vibrating object sees the time-averaged intensity:

$$I(x,y)=I_o(x,y)\left[1-J_o\left\{\frac{2\pi}{\lambda}d(x,y)(\cos\theta_e+\cos\theta_o)\right\}\right] \qquad [12]$$

Stroboscopic Holographic Interferometry

Stroboscopic holography is another interesting variant in the study of vibrations. The hologram is recorded by exposing the photographic plate twice for short time intervals during a vibration cycle. The pulsed exposures are synchronized with the vibrating surface, which is equivalent to making the surface virtually stationary during the recording. Reconstruction of the hologram yields cosinusoidal fringes, which are characteristic of double exposure holography. In real-time stroboscopic holography, a hologram of a nonvibrating object is first recorded. If the vibrating object is illuminated stroboscopically and viewed through the hologram, the reconstructed image from the hologram interferes directly with the light scattered from the object to generate live fringes.

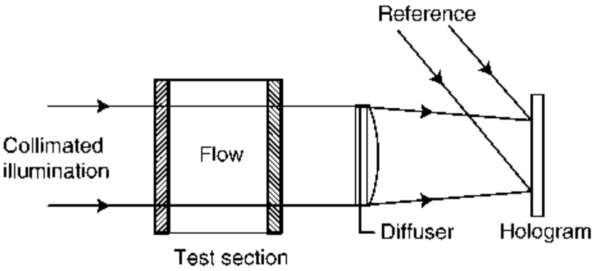

Figure 10 Schematic of a configuration for studying flow.

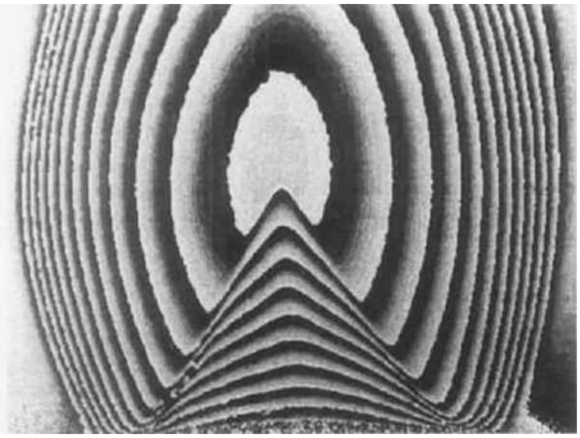

Figure 11 Fringe contours showing the phase distribution of a test flame.

Flow Measurement

Application of holographic interferometry to flow visualization and the measurement of spatial refractive index, density or temperature distributions have led to an advancement of understanding in areas such as aerodynamics, plasma diagnostics, and heat transfer. In the holographic interferometer shown in Figure 10 two consecutive exposures are made, usually the first exposure without flow and the second in the presence of a flow field. Double pulsed holography is used if the flow field is changing rapidly. The optical phase change due to flow between the exposures is

$$\Delta\varphi(x,y)=\frac{2\pi}{\lambda}\int_0^t\{n'(x,y,z)-n'_o\}\mathrm{d}z \qquad [13]$$

where $n'(x,y,z)$ is the refractive index distribution during the second exposure, n'_o is the uniform refractive index during the first exposure, and t is the length of the test section. Assuming that the properties of the flow are constant in the z direction, the expression for phase change can be expressed as

$$\Delta\tau(x,y)=KL\{\rho(x,y)-\rho_o\} \qquad [14]$$

where K is the Gladstone–Dale constant, ρ is the density of the gas, and ρ_o is the density in a no-flow situation. The interference pattern contours the change in the density field of the flow. The change in density per fringe is given by λ/Kt. An example of application related to temperature measurements in an axisymmetric premixed flame is illustrated in Figure 11.

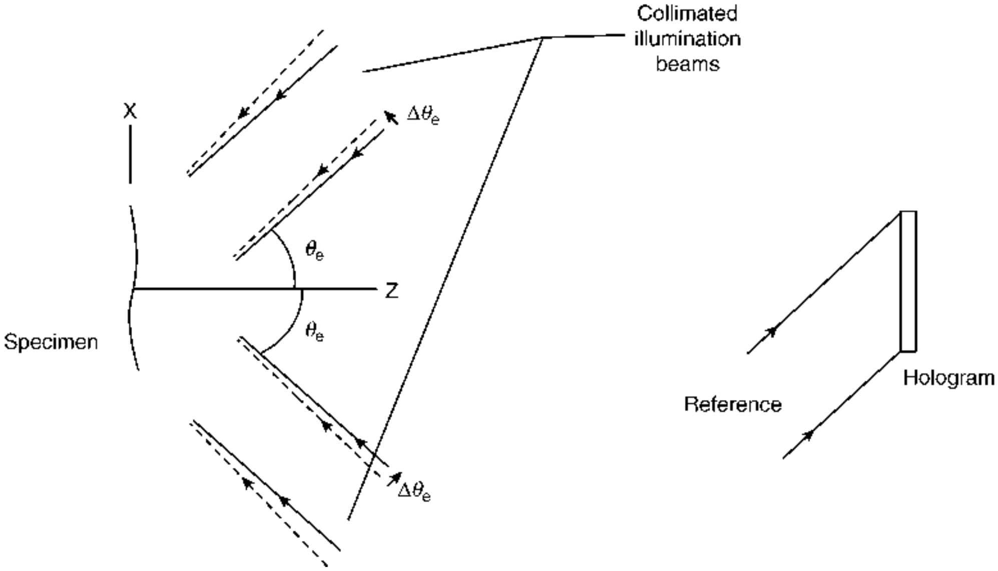

Figure 12 Schematic diagram of a two-beam method for shape measurement.

Surface Contouring

A practical way to display the shape of an object is to obtain contour maps showing the intersection of the object with a set of equidistant planes orthogonal to the line of sight.

The Dual Refractive Index Method

The method requires placing the model in a glass tank filled with a liquid (or gas) of refractive index n'_1. A hologram of the object is recorded, and the liquid contained in the tank is replaced by one of refractive index n'_2. An observer looking through the hologram sees the object surface modulated by a set of interference fringes arising from the change of optical phase in the light rays traversing the two liquids. The contour interval is given by

$$\Delta z = \frac{\lambda}{2(n'_1 - n'_2)} \qquad [15]$$

The Dual Illumination Method

In this method, the object is illuminated obliquely by collimated beams along two directions symmetric to the surface normal. The scattered waves are recorded on a holographic plate and the object and its reconstructed image are viewed through the hologram. Two sets of parallel equidistant fringes are projected onto the object surface by introducing small tilts, $\Delta\theta_e$, to the illumination beams (Figure 12). The contour sensitivity per fringe is given by

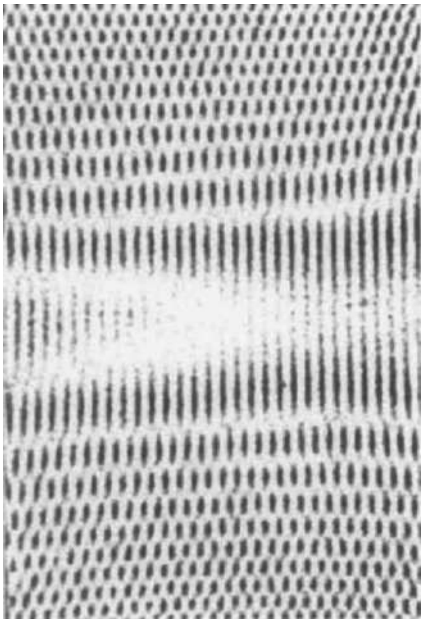

Figure 13 Example of a moiré topographic contour pattern obtained using the two-beam method.

$$\Delta z = \frac{\lambda}{2\sin\theta_e \sin\Delta\theta_e} \qquad [16]$$

The sensitivity of the method can be tuned in a wide range. An example of topographic contour pattern is shown in Figure 13. The distance between two adjacent moiré contour planes is 172 μm.

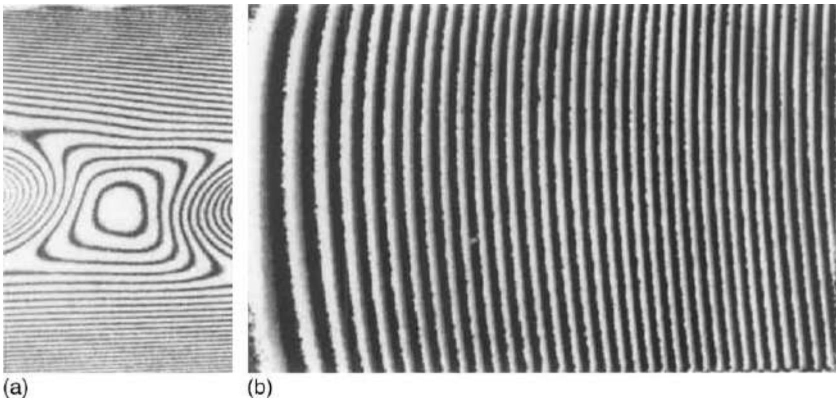

Figure 14 Examples of (a) shape measurement: the distance between two adjacent contour planes is 208 μm; and, (b) very large out-of-plane deformation measurement on a cantilever beam. The distance between two adjacent contour planes is 19 μm.

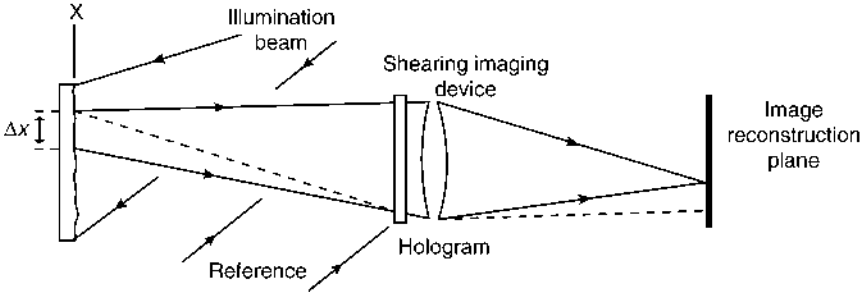
X
Illumination beam
Shearing imaging device
Δx
Image reconstruction plane
Hologram
Reference

Figure 15 Schematic diagram of the principle used in holographic shearing interferometry.

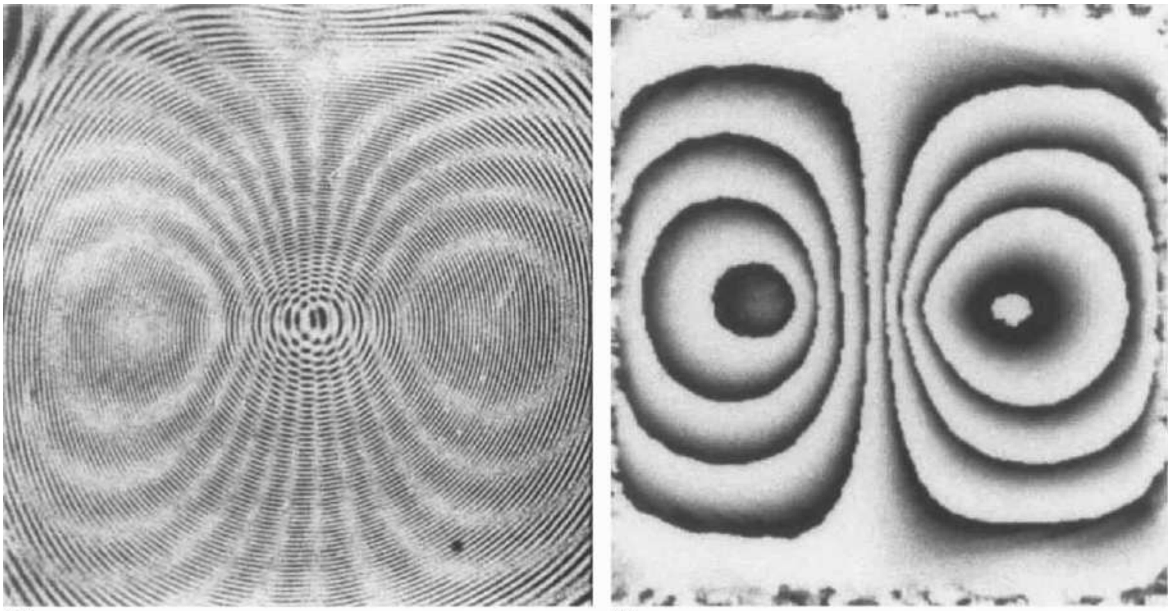

Figure 16 Examples of (a) moiré fringe contour, and (b) phase distribution corresponding to slope change produced on a centrally loaded aluminum plate clamped along its boundary.

In another method a collimated beam obliquely illuminates the object and a set of parallel equidistant fringes is projected onto the object surface by introducing a small tilt, $\Delta\theta_e$, to the illumination beam. Providing an appropriate reference beam rotation and a holographic plate translation generates contouring surfaces normal to the line of sight. A typical result using one beam illumination is shown in Figure 14a. The distance between two adjacent moiré contour planes is 208 μm.

This approach has been used to obtain measurement of large out-of-plane deformations undergone by a deformed object. The fringe pattern in Figure 14b corresponds to the out-of-plane displacement of a cantilever beam subjected to load at its free end. The distance between two adjacent contour planes is 19 μm.

Holographic Shearing Interferometry

Holographic shearing interferometry directly provides the patterns of slope change contours by laterally shearing wavefronts diffracted from the object surface. There are numerous ways to achieve shearing, such as, by introducing a Michelson type configuration, an inclined glass plate device, or a split-lens assembly in a holographic interferometer. The role of the shearing device is to enable the observation of a point on the object along two distinct neighboring directions. As a result of lateral wavefront shearing in the interferometer, a point in the image plane receives contributions from two different points on the object (Figure 15). Assuming that the illumination beam lies in the $x-z$ plane and makes an angle θ_e with the z-axis, and the observation is carried along the direction normal to the object surface, the change of optical phase due to deformation is

$$\Delta\varphi = \frac{2\pi}{\lambda}\left[\frac{\partial u}{\partial x}\sin\theta_e + \frac{\partial w}{\partial x}(1+\cos\theta_e)\right]\Delta x \qquad [17]$$

where Δx is the object shear along the x-direction. For $\theta_e = 0$, eqn [11] reduces to

$$\frac{\partial w}{\partial x} = \frac{n\lambda}{2\Delta x} \qquad [18]$$

Equation [18] displays a family of holographic moiré fringes corresponding to the contours of constant slope change. The fringe pattern shown in Figure 16a illustrates a moiré corresponding to contours of constant slope change for a centrally loaded square aluminum plate clamped along its boundary. An example of phase distribution

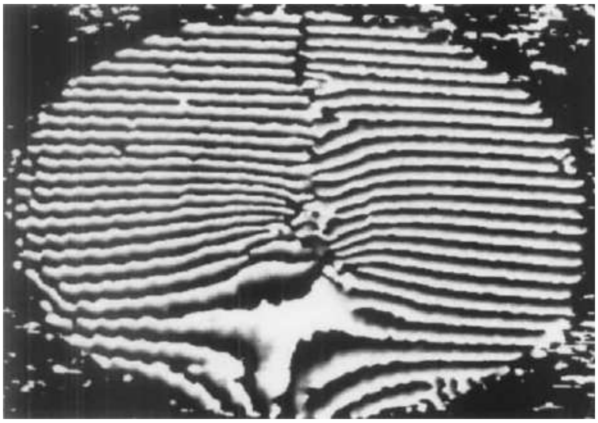

Figure 17 Example taken from the study of the fracture process zone in concrete: phase distribution corresponds to in-plane displacements on a notched concrete specimen subjected to a wedge splitting test.

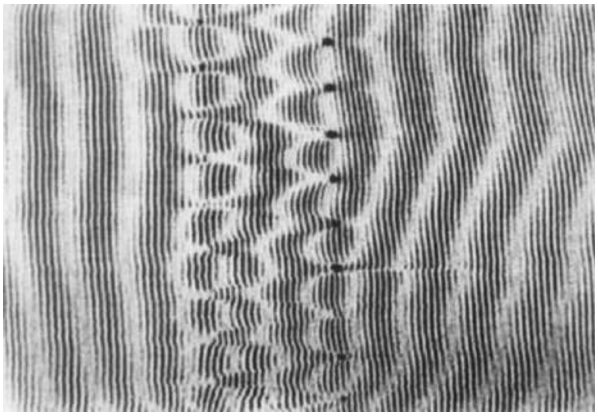

Figure 18 Example of application to nondestructive testing on a fingertip-joint wood specimen under tension using an in-plane sensitive holographic moiré setup.

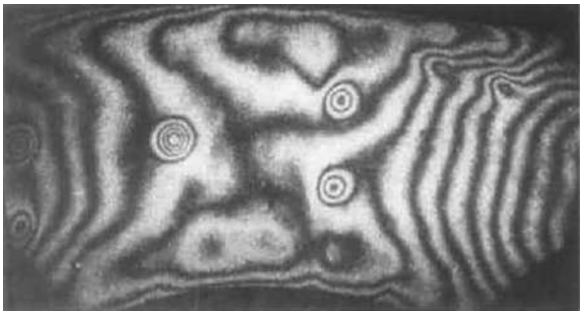

Figure 19 Application to the detection of defects in a tyre. Courtesy of H. Steinbichler, Steinbichler GmbH.

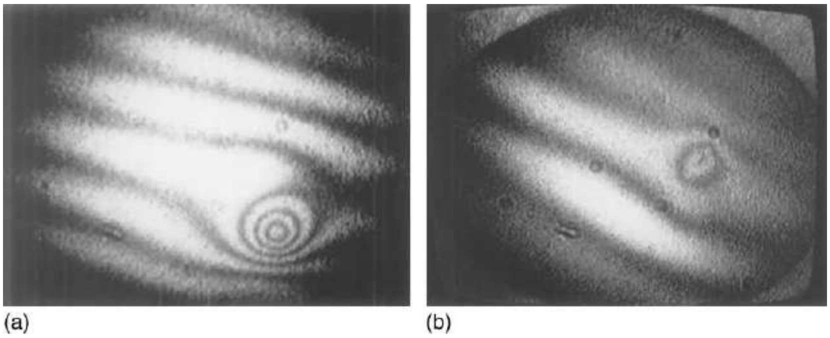

Figure 20 Fringe contours showing (a) a specific limb movement of a chick embryo during incubation, and (b) the hatching behavior. The point on the shell receiving blows is clearly visible.

corresponding to slope change along x-direction, is shown in Figure 16b.

Examples of Applications

Holographic interferometry has firmly established itself as an attractive and useful class of technique in nondestructive testing. Defects in a specimen are made apparent by the presence of anomalies in the interference fringe distribution of the deformed state of the specimen. The presence of flaws such as voids, disbonds, delaminations, and cracks weakens the strength of the specimen. The resulting strain concentration areas produced by the presence of flaws in the specimen are detected through their effect on the surface deformation map. The flaws in the specimen manifest themselves as distortion of the fringes, pockets of fringe concentrations, or localized discontinuities of the fringes in the full-field display of the deformation field of a test specimen in response to the applied load.

Figure 17 shows phase distribution corresponding to in-plane displacements of a notched fiber-reinforced concrete specimen subjected to a wedge-splitting test. The application is taken from the study of a fracture process zone in fiber-reinforced concrete. Another example of the application of holography in nondestructive testing is shown in Figure 18. The fringe pattern corresponds to in-plane displacements on a fingertip joint wood specimen under tension. Figure 19 shows an example of an application of holography in the tyre industry. The interferogram reveals, through the changes in the fringe pattern, defects in the tyre. The picture in Figure 20a is taken from the study of the embryonic behavior of embryos during incubation. The photograph shows a specific limb movement near the air space. The method also allows for a lively display of the hatching behavior of the chick. The point on the shell receiving blows is clearly visible in Figure 20b.

Conclusions

The aim of this article is to introduce readers to holographic techniques available for the noninvasive and whole-field investigation of the time-evolution and spatial distribution of the displacement and refractive index fields. Diverse techniques to obtain contourgrams of three-dimensional objects are also presented. The method offers strong potential in nondestructive testing applications, such as in the detection of flaws, in the conception of prototypes, in the evaluation of the performances of structural components, and in the verification of the reliability of finite-element computer codes.

See also

Holography, Techniques: Digital Holography; Overview. **Interferometry:** Overview; Phase Measurement Interferometry.

Further Reading

Erf RK (1974) *Holographic Non-Destructive Testing*. New York: Academic.

Ostrovsky YI, Butusov MM and Ostrovskaya GV (1980) *Interferometry by Holography*. Berlin: Springer-Verlag.

Ostrovsky YI, Schepinov VP and Yakovlev VV (1991) *Holographic Interferometry in Experimental Analysis*. Berlin: Springer-Verlag.

Rastogi PK (1994) *Holographic Interferometry – Principles and Methods*. Berlin: Springer-Verlag.

Schumann W and Dubas M (1979) *Holographic Interferometry*. Berlin: Springer-Verlag.

Schumann W, Zürcher JP and Cuche D (1985) *Holography and Deformation Analysis*. Berlin: Springer-Verlag.

Vest CM (1979) *Holographic Interferometry*. New York: Interscience.

Sandwich Holography and Light in Flight

N Abramson, Royal Institute of Technology, Stockholm, Sweden

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction to Holographic Interferometry

Holography is a photographic principle that produces 3D images that are so real that it is possible to take measurements inside this image with a precision of a fraction of a thousandth of a millimeter. Contours are formed within the holographic image that, like the relief lines of a map, indicate the displacements of every point on the object surface. The method is named holographic interferometry (*see* **Holography, Techniques:** Holographic Interferometry) and it can be used to measure deformation, vibration and also, with slightly lower resolution, dimension.

In real time holography the deformed object is directly compared to the holographic image of its undeformed state. In double exposure holography the comparison is made between two holographic recordings on the same plate, while in sandwich holography the two states are recorded on two different plates that are sandwiched together for comparison.

Sandwich holography is useful in industrial environments, because it can measure much larger deformations and also compensate for unwanted movement of the object or the setup between the two exposures. Further on in the process it will produce any sign of deformation and comparison can be made at any time without referring back to original setup. However, processing of the two plates is needed before the comparison can be made and the plates have to be positioned with high accuracy.

Interferometric Sensitivity

To calculate the interferometric sensitivity let us name (A) the point from which the laser light is emitted, (B) the point behind the hologram plate from which the observation is made, and (C) the displaced point on the object surface to be measured. Draw an ellipsoid through (C) with (A) and (B) as focal points. If (C) is displaced along the surface of the ellipsoid, the path length ACB will remain unchanged and no interference fringes are formed, while a displacement normal to this surface produces maximal interference. A diagram, the holodiagram, is based on these ellipsoids and used for the making and evaluation of holograms.

The bisector to ACB will be the normal to the ellipsoids. If the path length ACB is repeatedly increased by δL a diagram like Figure 1 is formed where the separation (s) of the ellipses, and thus the interferometric sensitivity, varies over the diagram. If δL is the wavelength of the laser light, one interference fringe is formed each time an object point has moved the distance (s) because of a deformation of the object between the two exposures. The displacement (d) can be calculated:

$$d = ns \qquad [1]$$

$$s = k0.5\delta L \qquad [2]$$

$$k = 1/\cos\alpha \qquad [3]$$

where n is number of fringes between fixed and studied point, k is constant along arcs of circles through A and B, and α is half the angle ACB.

When two slightly dissimilar patterns are placed one on top of the other, a moire pattern is formed that shows up the difference. The moire analogy to holographic interferometry is based on the fact that one fringe is formed for every ellipse that is crossed by an object point displaced between the two exposures.

If δL is the wavelength (λ) of the laser light the displacement (d) is

$$d = nk0.5\lambda \qquad [4]$$

If two-wavelength holography is used to measure the 3D shape of an undisturbed object, it will be seen intersected by ellipses where the separation (s) is

$$s = k0.5\lambda_1\lambda_2/(\lambda_1 - \lambda_2) \qquad [5]$$

If δL is the coherence length or the pulse length of the laser source, the object will be seen intersected by one single bright fringe, the thickness (s) of which is

$$s = k\delta L \qquad [6]$$

Everything outside that ellipse will be dark during reconstruction. This phenomenon can be used either to optimize the use of a limited coherence length or to explain the results of ‘light-in-flight recordings’, as described later.

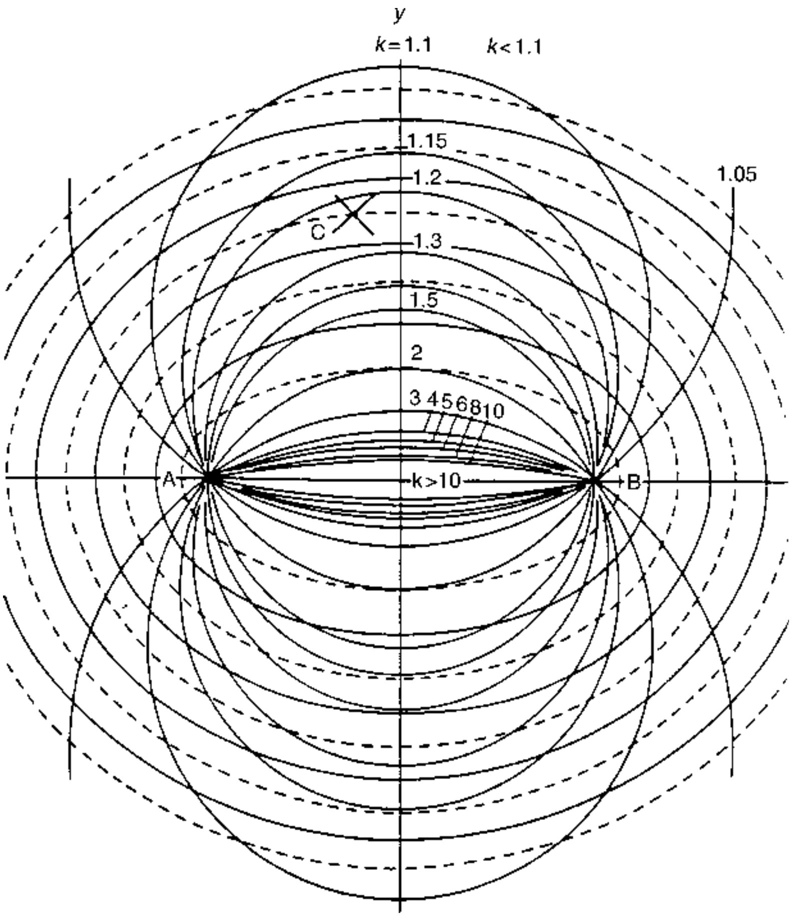

Figure 1 The holodiagram used for the evaluation of holograms.

The Principle of Sandwich Holography

In the following discussion on sandwich holography, we will only study the simplified example where the directions of illumination and observation are normal to the studied displacement and therefore α is equal to zero. The reflected light from an object point combined with the reference light produces a set of hyperbolas (Young's fringes) on the hologram plate.

When the object point is displaced between exposures on two different hologram plates, the hyperbolas will be positioned differently on the plates. There is a one-to-one relationship between the ellipsoids and the hyperbolas. Thus one point on the hologram plate will be crossed by one hyperbola if the corresponding object point is moved so that it crosses one of the ellipses of the holodiagram. Thus, the moire effect of the displaced hyperbolas at the hologram plate represents the moire effect of the ellipses in the object space. The interference pattern in the reconstructed object image can therefore be manipulated by moving one plate in relation to the other during reconstruction.

We have found that the simplest way to do this is to glue the two plates together separated by the glass thickness of one plate. To make the situation identical during exposure and reconstruction, the plates also have to be recorded in pairs. Thus, the first exposure can be made with one front plate and one back plate in contact in the hologram holder (Figure 2). The second exposure is made in exactly the same position with two new plates. After processing the front plate of the second exposure is glued in front of the back plate of the first exposure.

Then a reconstruction beam similar to the reference beam, is used to reconstruct the sandwich hologram. The interference fringes seen on the image of the object are manipulated by tilting the sandwich hologram, which results in an addition or subtraction of new fringes as in the following examples:

1. If the object is rotated by the small angle ϕ_1, the number of fringes can, during reconstruction, be

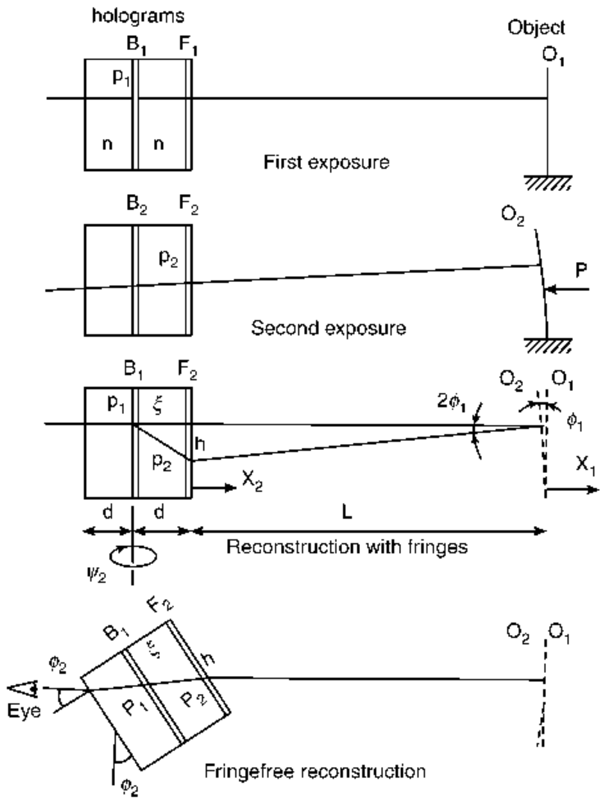

Figure 2 The exposures and reconstructions of a sandwich hologram.

eliminated by rotating the sandwich in the same direction by the much larger angle ϕ_2.

$$\phi_1 = \left[\frac{d}{2Ln} \right] \phi_2 \qquad [7]$$

2. If the sandwich during reconstruction is rotated around an axis perpendicular to that of the object rotation, new fringes are introduced at the angle γ, which is analogous to, but much larger than, the object rotation ϕ_1.

$$\phi_1 = \left[\frac{d}{2Ln} \right] \varphi_2 tg\gamma \qquad [8]$$

3. When the sandwich is rotated as described in (2) fringes are formed, the radius of curvature (r_2) of which are much larger than those of the bent object (r_1).

$$r_1 = \left[\frac{2Ln}{d} \right] \left[\frac{1}{\varphi_2} \right] r_2 \qquad [9]$$

Examples

The object consisted of three vertical steel bars (Figure 3) that were fixed by screws at their lower end to a rigid frame that surrounded all the bars. The middle bar was also supported at the upper end. Forces were applied at the middle of the bars. The right bar is deformed away from the observer while the other two are deformed towards him. After deformation of the three objects the second pair of plates were exposed. Then all four plates were processed.

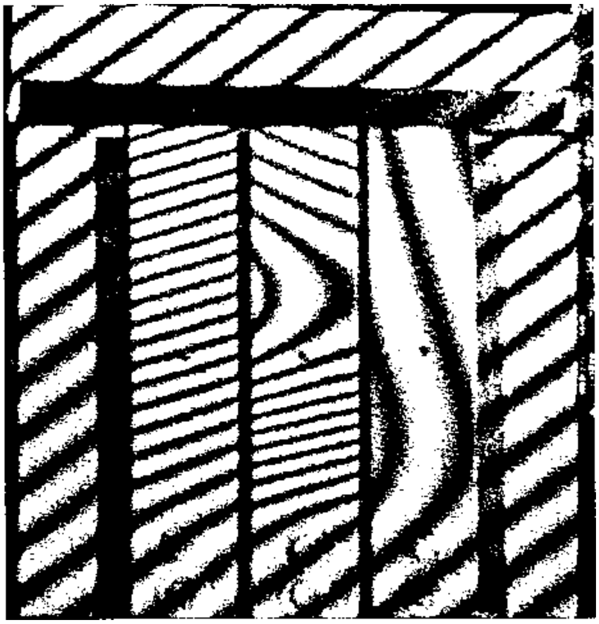

Figure 3 The movements of three steel bars in relation to a frame are to be measured by the fringes of sandwich holography. However, the unwanted motion of the total set-up makes the evaluation impossible.

The back plate of the first exposure (B_1 in Figure 2) was repositioned in the plate holder in front of the plate of the second exposure (F_2). The two plates were bonded together to form a sandwich hologram. O_1 and O_2 represent the two positions of one object, which between the two exposures was bent at angle ϕ_1 toward the hologram. (L) is the distance between the front hologram and the object (~1 m); (L) is also the distance between the point of illumination and the object. (D) is the distance between the point source of the reference beam and the hologram plate ($D = 2L$); (d) is the thickness of one hologram plate (~1.3 mm); (n) is the refractive index of the plates (~1.5 mm).

When the two plates were properly repositioned in the plate holder, fringes were seen everywhere, both on the steel bars and on the supposedly fixed frame which apparently had made an accidental movement (Figure 3). By tilting the plate holder in different directions, the fringes on the frame could be eliminated and the hologram could be evaluated as if the frame had remained fixed (Figure 4).

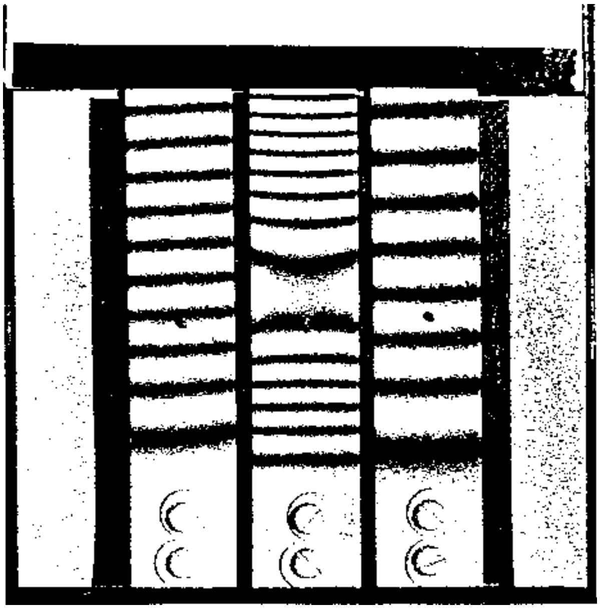

Figure 4 The unwanted fringes caused by the motion of the total set-up have been eliminated by a tilt of the sandwich hologram.

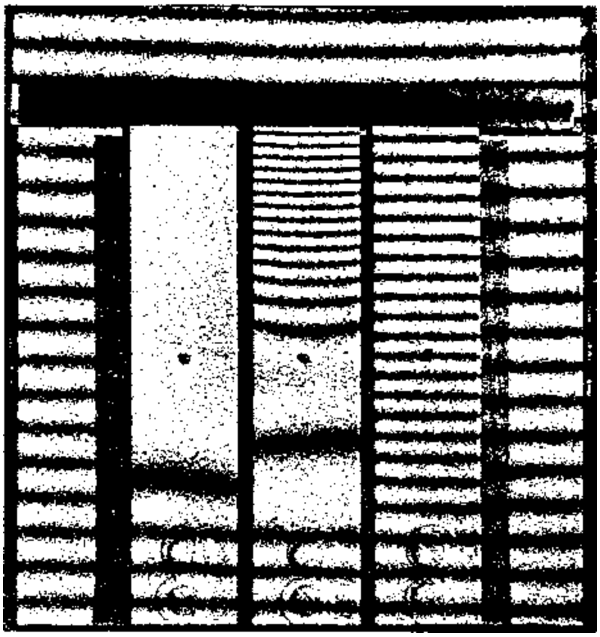

Figure 5 Another tilt of the sandwich hologram reveals that the left and the right bar have been deformed in opposite directions.

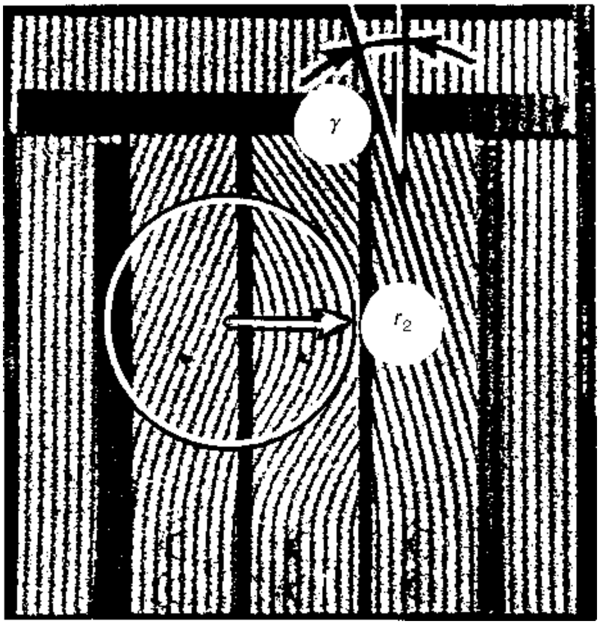

Figure 6 A rotation of the sandwich hologram around a vertical axis produces fringes that reveal direction of tilt and bending radius of all steel bars.

The plate holder was then tilted backwards at angle ϕ_2 and horizontal fringes were found on the frame while the top of the left bar became fringe-free (Figure 5). The number of fringes on the right bar increased, proving that it had been tilted in the opposite direction. If the sandwich were instead rotated around a vertical axis at angle φ_2, a set of vertical fringes were formed on the frame, while the bars became covered by fringes, the inclination and curvature of which represent a magnified view of the deformed bars as seen in profile (Figure 6).

Introduction to Light-in-flight Recording by Holography

Recording of light pulses in flight have been possible since the beginning of the 1990s, using two-photon fluorescence or ultrafast Kerr cells driven by laser pulses. Very fast optical phenomena, such as refractive-index changes in laser-produced plasmas, have been recorded by holography using short illumination pulses.

In 1968, Staselko *et al.* published a method of studying the time coherence pattern of wave trains from pulsed lasers by the use of autocorrelation in which the unknown pulse is compared to itself. The method we name 'light-in-flight recording by holography' (LIF) is based on studying both the temporal and the spatial shape of pulses by correlation, which means that we study the change in shape of a pulse before and after it has passed through the transformation we want to study. For this process, we want to start with as short a pulse (or short coherence length) as possible and then compare the transformed pulse to the original pulse.

In a hologram only those parts of an object will be recorded for which the path length from the laser to the holographic plate via the object does not differ from the path length of the reference beam by more than the pulse length (or coherence length) of the laser light used for the recording. If that length is short, during reconstruction a large object will be seen intersected by an imaginary surface in the form of an ellipsoid, one of its two foci being the point from which the spherical wavefronts of illumination are emitted (spatial filter (A) in Figure 1) and the other being the point of observation (B). A spherical wavefront appears distorted into an ellipsoid if it is divergent, and into a hyperboloid if it is convergent, while a flat wavefront appears distorted into a paraboloid. The general rule is that the apparent wavefront has the shape of a mirror that focuses light from (A) to (B).

By careful planning, it is possible to optimize the geometry of the holographic setup in such a way that the utilized portions of the ellipsoids and hyperboloids can be approximated into flat surfaces. Thus, the object is seen intersected by one flat imaginary interference surface, the depth of which can be altered during reconstruction by changing the point of observation at the hologram plate.

Light of short duration, or short coherence length, can, in this way, be used in a holographic method for finding depth contours. If, however, the 3D shape of the object is known a priori, e.g., if it is a flat surface, this method can just as well be used to study the 3D shape of the illuminating wavefront. As mentioned before, the intersections of the ellipsoids can be approximated by intersections of spherical wave fronts of illumination, if the distance between object and hologram plate is long and if the flat surface is perpendicular to the line of sight.

Examples of Wavefront Studies

As an example we will describe an experiment that consists of an opaque white diffusely reflecting flat screen, that is illuminated from the left at an oblique angle, as seen in Figure 7. A mirror is fixed to the screen so that some of the light is reflected upwards. During recording of the hologram, the reference pulse is reflected by the two mirrors (M) so that it arrives at the middle of the hologram plate at the same time as the object pulse, in the form of diffuse light, arrives from the middle of the screen. As the reference pulse passes over the hologram plate from left to right, it works like a curtain shutter (focal-plane shutter) recording early object light to the left and later light to the right. This ‘light shutter’ moves across the plate with a velocity faster than light which is possible as it is just an intersection between pulsefront and plate. The difference in time of the recordings made at the two ends of the hologram plate is the time it takes for the reference pulse to move across the plate, or about 800 picoseconds. During reconstruction with a continuous laser the early arriving object light is seen through the left part of the plate, while the later arriving light is seen through its right part. If the eye, or a TV camera scans from left to right, the pulse is seen moving continuously over the object screen, with some of it reflected upwards by the mirror (Figure 8a–d). The corresponding pulse length is 100 ps (~30 mm), and to minimize pulsefront distortions the screen is perpendicular to the line of sight.

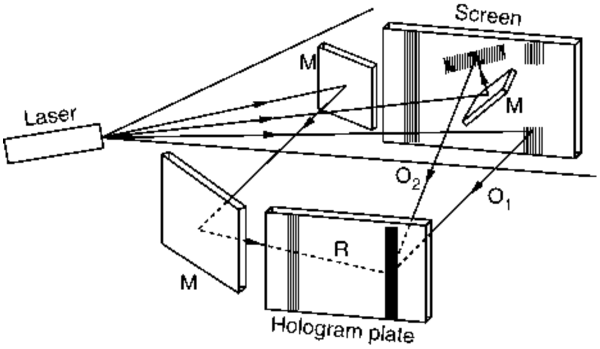

Exposure of light-in-flight

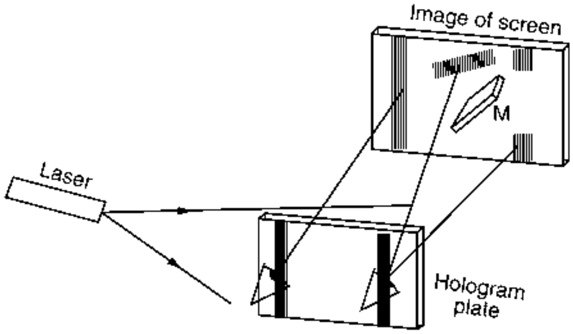

Reconstruction of light-in-flight

Figure 7 During exposure of the light-in-flight hologram, the short pulse from the laser is used to obliquely illuminate both the object and the hologram plate. Only those parts will be recorded where path length for object light and reference light are equal.

The photo in Figure 9 shows the result of another experiment. A spherical wave expands from a point source at left and moves to the right. The spherical wavefront intersects a flat observation screen consisting of a white-coated aluminum plate of 20 × 30 cm. A cylindrical lens is fixed, with its axis normal to the screen, so that it focuses the arriving light almost parallel to the screen surface. Figure 8 is a composite of several photographs taken through different parts

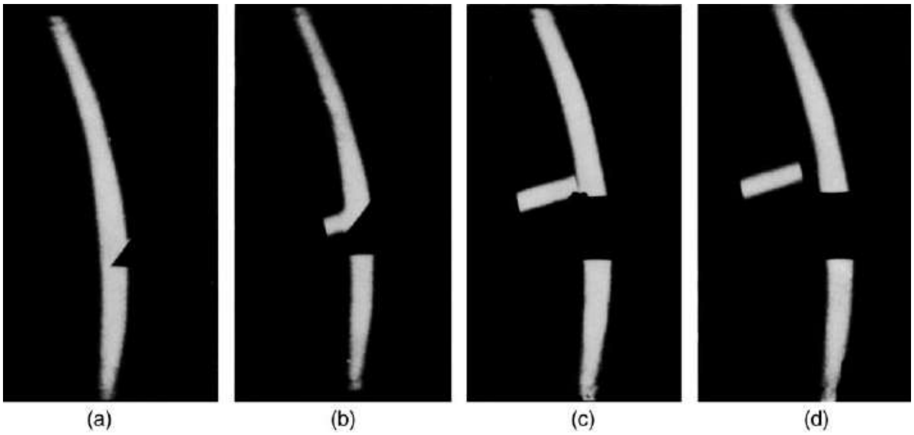

Figure 8 Resulting reconstructed images of the light reflected by a mirror at about 45°. These four photos were taken during reconstruction while the camera behind the hologram plate was moved from left to right in between each of the exposures.

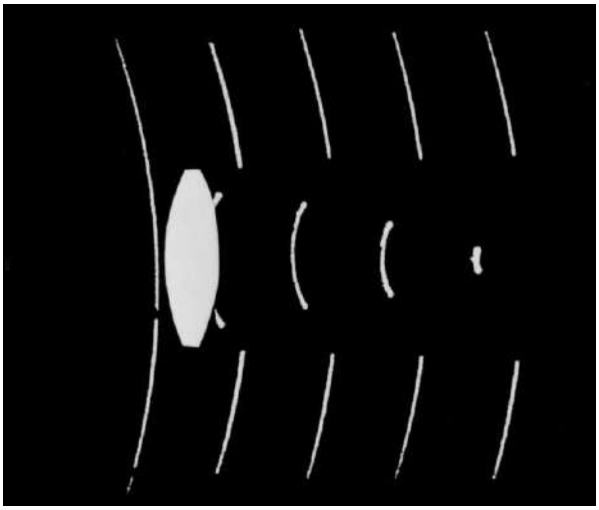

Figure 9 Light focused by a lens. A composite of several photos taken from different parts of the hologram plate. An image of the lens is included to make the picture clear.

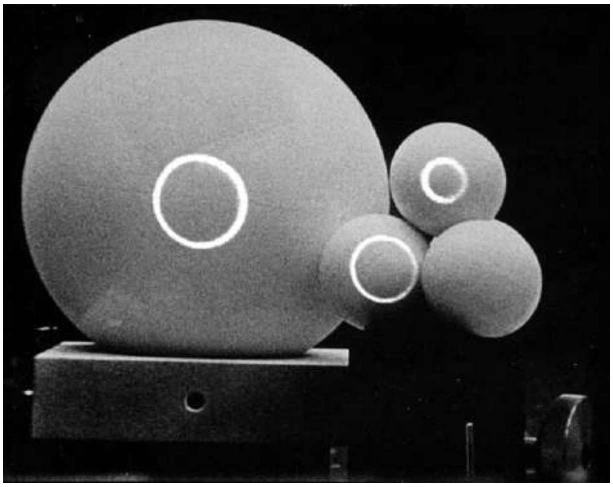

Figure 10 Spheres intersected by a thin sheet of light.

of the hologram plate. The pulse length was in this case 10 ps (~3 mm).

Measuring the Shape of 3D Objects

The wavefront (pulse front) of a short pulse forms a thin sheet of light. The intersection of this sheet by a 3D body produces on its surface a thin line representing a cross-section. It is advantageous if this cross-section represents a flat surface, which will be approximately the case if the illumination and observation beams are collimated or arrive from points at large distances.

By moving the point of observation along the hologram plate from left to right, the intersecting sheet of light can be moved in depth forward or backward. Determination of fringe order number (by finding the zero fringe) no longer represents any problem because only the zero order fringe exists. Therefore, this way of contouring has great advantages, especially when the observations and calculations are made by computer. Figure 10 shows a set of spheres intersected by a thin light sheet.

The intersecting light sheet can be manipulated in yet another way, i.e., it can be rotated. If the point of observation is moved away from the plate, the right part of an object will be seen through the right part of the plate were the recording made later than at the left part. During the time interval between left and right records, the intersecting light sheet has moved so that it intersects the object further away, resulting in a tilted cross-section of the object. This way it has been possible to rotate the cross-section up to 90 degrees, so that if, for example, a face is illuminated from the front by a flat sheet of light, the reconstructed intersection will show the face in profile.

In a limited way it is also possible to produce results similar to those of LIF by using a CCD camera instead of the holographic plate. The great advantage is its compatibility with a PC, while the main disadvantage is a lower resolution.

Other Uses

Like all methods of 'gated viewing', LIF can be used to look through scattering media, e.g., to look through human tissue. By observing only the first light the influence of the later arriving scattered light is eliminated. Thus, LIF can be used as a diagnostic method based on the idea that the scattering, absorption, or refractive index of cancer is different from that of healthy tissue. Another use of LIF is to image just a thin section of a gas or liquid filled with particles. In this way it is possible to observe the distribution and velocity in a certain localization without influence of the scattering from particles in front of or behind the studied section.

Up to now we have discussed gated viewing by LIF in the laboratory using picosecond light pulses and distances of up to a few meters. Instead we could look at astronomical objects with light pulses of days (e.g., a supernova) and distances of many light years. Even in these cases we will see particles in space illuminated by ellipsoids with (A) being the supernova and (B) a telescope on Earth. While each ellipsoid in the laboratory may represent a pathlength of some nanoseconds, each ellipsoid in space might represent light years.

However, the methods we have shown here are not limited to the study of the apparent shape of spheres of light. They can also be used to study other objects moving at ultrahigh velocities. If an exploding supernova throws out particles in all directions at a rather slow velocity, it can be seen surrounded by an expanding sphere. If, instead, this sphere expanded with the speed of light (which is not possible) it would appear to us just like an ellipsoid of the holodiagram. If, however, it expanded at a very high velocity, but lower than that of light, the sphere would appear to us in the shape of an egg with its pointed end towards the observer. The velocity of expansion could, in this case, appear superluminal (higher than the speed of light). If the velocity is (v), its approaching velocity will to a first approximation appear to be the true velocity divided by $(1 - v/c)$.

Finally, we will show a relationship between LIF and special relativity, by looking at a modification of the 'Minkowski light cones'. Figure 11 shows a diagram where (x) and (y) represent space coordinates, while (z) represents time multiplied by (c). A sphere from a short pulse of light that expands from the point A is represented by a cone with its apex at (A). A sphere from a short observation that converges to (B) is represented by a cone with its apex at (B). Only those object points, that are positioned at the elliptic intersection of the two cones, can be seen illuminated. This ellipse is identical to one of the ellipses of the holodiagram. The distance in time and space separating (A) and (B) could be either caused by the fixed distance between source of laser light and hologram plate in holographic LIF, or by a high velocity of the observer. Thus, the Lorentz contraction in special relativity could be explained as a measuring error because the high velocity of the observer traveling from (A) to (B) causes his sphere of observation to be transformed into an ellipsoid of observation. The tilt of the elliptic intersection indicates which observations that appear to be simultaneous depending on the separation of (A) and (B) along the y-axis, both in LIF and in relativity.

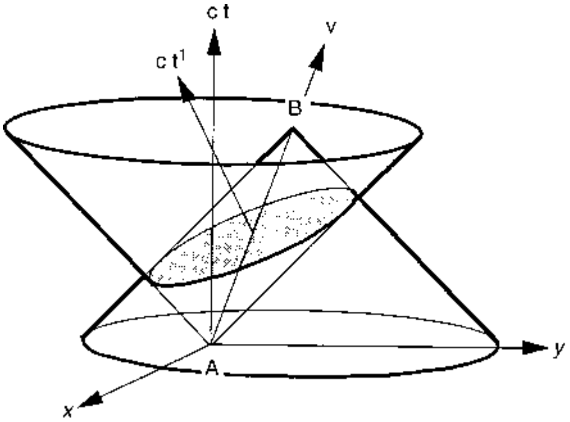

Figure 11 A modified Minkowski diagram consisting of one cone of illumination (A) and one cone of observation (B). The separation in time and space between (A) and (B) could either be static as in LIF or caused by a relativistic velocity of the observer from (A) to (B).

See also

Holography, Techniques: Holographic Interferometry.

Further Reading

Abramson N (1971) Moiré patterns and hologram interferometry. *Nature* 231: 65–67.

Abramson N (1981, 1986) *The Making and Evaluation of Holograms*. London: Academic Press.

Abramson N (1996) *Light in Flight, or, The Holodiagram – The Columbi Egg of Optics*. USA: SPIE Press, vol. PM 27.

An W and Carlsson T (1999) Digital measurement of three-dimensional shapes using light-in-flight speckle interferometry. *Optical Engineering* 38: 1366–1370.

Dugay MA and Mattick AT (1971) Ultrahigh speed photography of picosecond light pulses and echoes. *Applied Optics* 10: 2162–2170.

Garrington S (1992) Invisible jet sheds new light. *Nature* 355: 771–772.

Giordmaine J and Rentzepis P (1967) Two-photon excitation of fluorescence by picosecond light pulses. *Applied Physics Letters* 11: 216.

Hinrichs H and Hinsch KD (1997) Light-in-flight holography for flow visualization and velocimetry in three-dimensional flows. *Optics Letters* 22: 828–830.

Malin D and Allen D (1990) Echoes of the supernova. *Sky and Telescopes* 22–25, January.

Powell R and Stetson K (1965) Interferometric analysis by wavefront reconstruction. *Journal of the Optical Society of America* 55: 1593–1598.

Spears K and Serafin J (1989) Chrono-coherent imaging for medicine. *IEEE Transactions on Medical Engineering* 36: 1210–1220.

Staselko DI, Denisyuk YN and Smirnow AG (1969) Holographic recording of the time-coherence pattern of a wave train from a pulsed laser source. *Optics and Spectroscopy* 26: 413.

I

IMAGING

Contents

Information Theory in Imaging
Inverse Problems and Computational Imaging
Adaptive Optics
Hyperspectral Imaging
Imaging Through Scattering Media
Infrared Imaging
Interferometric Imaging
Lidar
Multiplex Imaging
Photon Density Wave Imaging
Three-Dimensional Field Transformations
Volume Holographic Imaging
Wavefront Sensors and Control (Imaging Through Turbulence)

Information Theory in Imaging

F O Huck and C L Fales, NASA Langley Research Center, Hampton, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Information theory is an intuitively attractive mathematical approach for combining observation with communication, or sensing with processing. So far, however, this approach has remained largely in the realm of signal processing and transmission, as first formalized by Shannon and Wiener over 50 years ago. Recent extensions of this approach to imaging have mostly addressed the interpretation of images, without explicitly accounting for the critical limiting factors that constrain image gathering and restoration. Yet, to be useful in the characterization of imaging, information theory must account for these constraints and lead to a close correlation between predicted and actual performance. The characterization that is outlined here traces the rate of information from the scene to the observer; and it pairs this rate, first, with the theoretical minimum data rate to assess the efficiency with which information can be transmitted, and, second, with the maximum-realizable fidelity to assess the quality with which images can be restored. Further details, especially for the mathematical development of figures of merit and for the design of imaging systems, can be found in the Further Reading.

Model of Imaging

Description

Modern imaging systems usually have the following three stages, as depicted in Figure 1:

1. *Image gathering*, to capture the radiance field that is either reflected or emitted by the scene and transform this field into a digital signal;

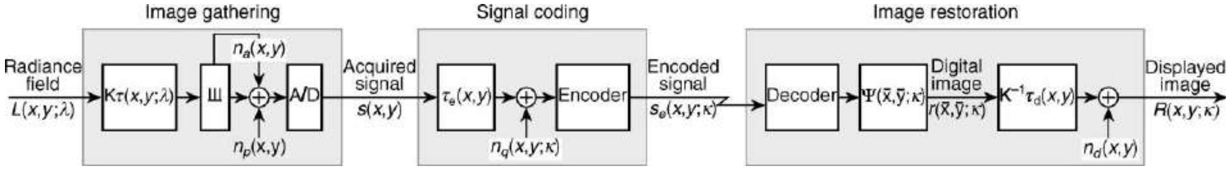

Figure 1 Model of digital imaging systems.

2. *Signal coding*, to encode the acquired signal for the efficient transmission and/or storage of data and decode the received signal for the restoration of images. The encoding may be either lossless (without loss of information) or lossy (with some loss of information), if a higher compression is required;
3. *Image restoration*, to produce a digital representation of the scene and transform this representation into a continuous image for the observer.

Human vision may be included in this model as a fourth stage, to characterize imaging systems in terms of the information rate that the eye of the observer conveys from the displayed image to the higher levels of the brain.

These stages are similar to each other in one respect: each contains one or more transfer functions, either continuous or discrete, followed by one or more sources of noise, also either continuous or discrete. The continuous transfer function of the image-gathering device (and of the human eye) is followed by sampling that transforms the captured radiance field into a discrete signal with analog magnitudes. In most devices the photodetection mechanism conveys this signal serially into an analog-to-digital (A/D) converter to produce a digital signal. However, analog signal processing in a parallel structure, akin to that in human vision, has many advantages that are increasingly emulated by neural networks.

This model of imaging differs from the classical model of communication that Shannon and Wiener addressed in two fundamental ways:

1. *The continuous-to-discrete transformation in image gathering*. Whereas communication is constrained critically only by bandwidth and noise, image gathering is constrained also by the compromise between blurring and aliasing, due to limitations in the response of optical apertures and in the sampling density of photodetection mechanisms. This additional constraint requires the rigorous treatment of insufficiently sampled signals throughout the imaging system.
2. *The sequence of image gathering followed by signal coding*. Whereas the characterization of communication addresses the perturbations that occur at the encoder and decoder or during transmission, the characterization of imaging systems must also address the perturbations in the image-gathering process prior to coding. These additional perturbations require a clear distinction between the information rate of the encoded signal and the associated theoretical minimum data rate or, more generally, between information and entropy.

The mathematical model of imaging that is outlined here accounts for the appropriate deterministic and stochastic (random) properties of natural scenes and for the critical limiting factors that constrain the performance of imaging systems. The corresponding formulations of the figures of merit are sufficiently general to account for a wide range of radiance field properties, photodetection mechanisms, and signal processing structures. However, the illustrations emphasize: (a) the salient properties of natural scenes that dominate the captured radiance field in the visible and near-infrared region under normal daylight conditions; and (b) the photodetector arrays that are used most frequently as the photodetection mechanism of image-gathering devices.

Radiance Field

The radiance field $L(x, y; \lambda)$, reflected by natural scenes, may be modeled approximately as

$$L(x, y; \lambda) = \frac{1}{\pi} I(\lambda)\rho(\lambda)\rho(x, y) \qquad [1]$$

where $I(\lambda)$ and $\rho(\lambda)$, respectively, represent the spectral irradiance and reflectance properties as a function of the wavelength λ, and $\rho(x, y)$ represents the random reflectance properties as a function of the spatial coordinates (x, y). Figure 2 illustrates a model of $\rho(x, y)$ for which the power spectral density (PSD) $\hat{\Phi}_\rho(\upsilon, \omega)$ corresponds closely to that typical of natural scenes, as given by

$$\hat{\Phi}_\rho(\upsilon, \omega) = \frac{2\pi\mu_\rho^2\sigma_\rho^2}{[1 + (2\pi\mu_\rho\zeta)^2]^{3/2}} \qquad [2]$$

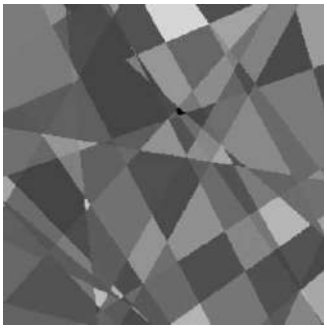

Figure 2 Model of reflectance $\rho(x, y)$.

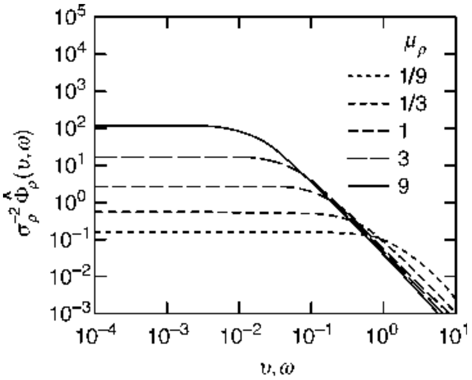

Figure 3 Normalized PSD of the reflectance $\rho(x, y)$.

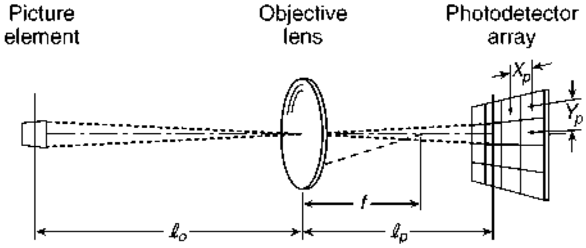

Figure 4 Optical geometry of image-gathering device.

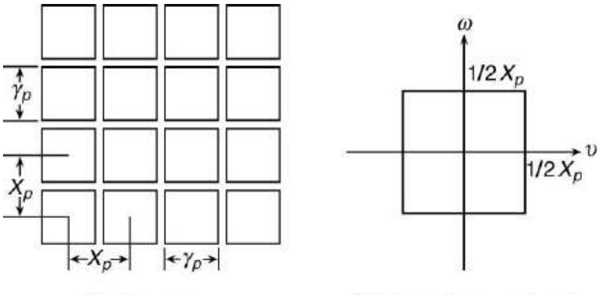

(a) Geometry (b) Sampling passband

Figure 5 Photodetector-array geometry and sampling passband.

where $\zeta^2 = \upsilon^2 + \omega^2$ and (υ, ω) are the spatial frequency coordinates of the Fourier-transform domain (see Figure 3). This model consists of random polygons, whose boundaries are distributed according to Poisson probability with a mean separation of μ_ρ, and whose magnitudes are distributed according to independent zero-mean Gaussian statistics of variance σ_ρ^2. The constraints associated with eqn [1] are satisfied by letting the mean reflectance $\bar{\rho} = 0.5$ and the standard deviation $\sigma_\rho \leq 0.2$.

Image Gathering

Figure 4 depicts the basic optical geometry of an image-gathering device, in which the objective lens forms an image of the captured radiance field across the photodetector array, and the array transforms this image into a discrete signal. The sampling lattice (X, Y), which determines the resolution of this device, is formed in angular units by the separation (X_p, Y_p) of the photodetector-array apertures and the distance ℓ_p as given for one of the dimensions by

$$X = 2\tan^{-1}\frac{X_p}{2\ell_p} \approx \frac{X_p}{\ell_p}\text{rad} = \frac{57.3X_p}{\ell_p}\text{deg} \qquad [3]$$

Normally, the sampling lattice is square, as shown in Figure 5, so that $X_p = Y_p$. Then the corresponding area of the photodetector apertures is $A_p = \gamma_p^2$, and the solid angle Ω of the instantaneous field of view (IFOV) is $\Omega \approx A_p/\ell_p^2$ ster.

The sampling lattice projected onto the scene at the distance ℓ_o has the dimensions $(X_o, Y_o) \approx (X\ell_o, Y\ell_o)$. Because these dimensions represent the finest spatial detail that normally can be restored, it is convenient to measure the mean spatial detail of the scene relative to the sampling lattice, as given by $\mu = \mu_\rho(X_oY_o)^{-1/2}$, which reduces to $\mu = \mu_\rho X_o^{-1}$ for $X = Y$. Hence, $\mu = 1$ indicates that the sampling interval is equal to the mean spatial detail.

The image-gathering process that transforms the continuous radiance field $L(x, y; \lambda)$ into the discrete signal $s(\mathbf{x}, \mathbf{y})$ may be modeled by the expression

$$s(\mathbf{x}, \mathbf{y}) = K\rho(\mathbf{x}, \mathbf{y}) * \tau(\mathbf{x}, \mathbf{y}) + n_p(\mathbf{x}, \mathbf{y}) \qquad [4]$$

where the discrete coordinates $(\mathbf{x}, \mathbf{y}) = (mX, nY)$, and the symbol $*$ denotes continuous spatial convolution. This model accounts, in the simplest form, for the reflectance-to-signal conversion gain K, the reflectance $\rho(x, y)$ of the scene, the spatial response (SR) $\tau(x, y)$ of the image-gathering device, the continuous-to-discrete transformation by the photodetector-array lattice, and the discrete noise $n_p(\mathbf{x}, \mathbf{y})$ of the photodetectors. The discrete coordinates are defined by the sampling function

$$\text{Ш} \equiv \text{Ш}(\mathbf{x}, \mathbf{y}) = XY \sum_{\mathbf{x},\mathbf{y}} \delta(x - \mathbf{x}, y - \mathbf{y}) \qquad [5]$$

where $\delta(x, y)$ is the Dirac delta function; and the gain is given by

$$K = \frac{1}{\pi} A_\ell \Omega \int I(\lambda)\rho(\lambda)r(\lambda)\mathrm{d}\lambda \qquad [6]$$

where $A_\ell = \pi D^2/4$ is the area of the objective lens with diameter D, and $r(\lambda)$ is the responsivity of the photodetectors. This gain relates the magnitude of the signal $s(\mathbf{x}, \mathbf{y})$ directly to that of the reflectance $\rho(x, y)$; and, together with the variances σ_ρ^2 and σ_p^2 of the reflectance and the photodetector noise, respectively, it also determines the rms-signal-to-rms-noise ratio (SNR) $K\sigma_\rho/\sigma_p$.

The DFT of $s(\mathbf{x}, \mathbf{y})$ is the periodic spectrum $\tilde{s}(\upsilon, \omega)$ given by

$$\tilde{s}(\upsilon, \omega) = K\hat{\rho}(\upsilon, \omega)\hat{\tau}(\upsilon, \omega) * \hat{\text{Ш}} + \tilde{n}_p(\upsilon, \omega) \qquad [7]$$

where $\hat{\rho}(\upsilon, \omega)$ is the spectrum of the reflectance, $\hat{\tau}(\upsilon, \omega)$ is the spatial frequency response (SFR) of the image-gathering device, and $\tilde{n}_p(\upsilon, \omega)$ is the spectrum of the photodetector noise. The symbol $\hat{\text{Ш}}$ is the Fourier transform of the sampling function Ш. It accounts for the sidebands generated by the sampling process and is given by

$$\hat{\text{Ш}} \equiv \hat{\text{Ш}}(\upsilon, \omega) = \sum_{\mathbf{v},\mathbf{w}} \delta(\upsilon - \mathbf{v}, \omega - \mathbf{w}) \qquad [8]$$

where $(\mathbf{v}, \mathbf{w}) = (m/X, n/Y)$. The corresponding sampling passband is

$$\hat{B} = \left\{(\upsilon, \omega);\ |\upsilon| < \frac{1}{2X},\ |\omega| < \frac{1}{2Y}\right\} \qquad [9]$$

The above expression for $\tilde{s}(\upsilon, \omega)$ can be rewritten more conveniently as

$$\tilde{s}(\upsilon, \omega) = K\hat{\rho}(\upsilon, \omega)\hat{\tau}(\upsilon, \omega) + \hat{n}(\upsilon, \omega) \qquad [10]$$

where

$$\hat{n}(\upsilon, \omega) = \hat{n}_\mathrm{a}(\upsilon, \omega) + \tilde{n}_p(\upsilon, \omega) \qquad [11]$$

is the spectrum of the total noise in the acquired signal. The first term $\hat{n}_\mathrm{a}(\upsilon, \omega)$ represents the aliased signal components caused by insufficient sampling, as given by

$$\hat{n}_\mathrm{a}(\upsilon, \omega) = K\hat{\rho}(\upsilon, \omega)\hat{\tau}(\upsilon, \omega) * \hat{\text{Ш}}_\mathrm{s} \qquad [12]$$

where

$$\hat{\text{Ш}}_\mathrm{s} \equiv \hat{\text{Ш}}_\mathrm{s}(\upsilon, \omega) = \sum_{\mathbf{v},\mathbf{w} \neq (0,0)} \delta(\upsilon - \mathbf{v}, \omega - \mathbf{w}) \qquad [13]$$

If, for spatial coordinates at either the scene or the photodetector array, the unit of (x, y) is meters, then the unit of (υ, ω) is cycles m^{-1}; and if, for angular coordinates centered at the objective lens, the unit of (x, y) is radians, then the unit of (υ, ω) is cycles rad^{-1}. Either way, it is convenient to normalize the sampling intervals to unity (i.e., $X = Y = 1$). Then the area of the sampling passband is $|\hat{B}| = 1$, the unit of (x, y) is samples, and the unit of (υ, ω) is cycles sample^{-1}.

The SFR $\hat{\tau}(\upsilon, \omega)$ of the image-gathering device is

$$\hat{\tau}(\upsilon, \omega) = \hat{\tau}_\ell(\upsilon, \omega)\hat{\tau}_p(\upsilon, \omega) \qquad [14]$$

where $\hat{\tau}_\ell(\upsilon, \omega)$ and $\hat{\tau}_p(\upsilon, \omega)$ are the SFRs of the objective lens and photodetector apertures, respectively. This SFR may be modeled approximately by the circularly symmetric Gaussian shape

$$\hat{\tau}(\upsilon, \omega) = \exp[-(\zeta/\zeta_\mathrm{c})^2] \qquad [15]$$

where, as shown in Figure 6, the optical-response index ζ_c controls the trade-off between the blurring that is caused by the decrease of the SFR inside the sampling passband and the aliasing that is caused by the extension of the SFR beyond this passband.

Signal Coding

Signal coding may be divided into two stages, as depicted in Figure 1. The first stage, which is represented by the digital operator $\tau_\mathrm{e}(\mathbf{x}, \mathbf{y})$ and the quantization noise $n_q(\mathbf{x}, \mathbf{y}; \kappa)$, accounts for manipulations of the acquired signal that seek to extract those

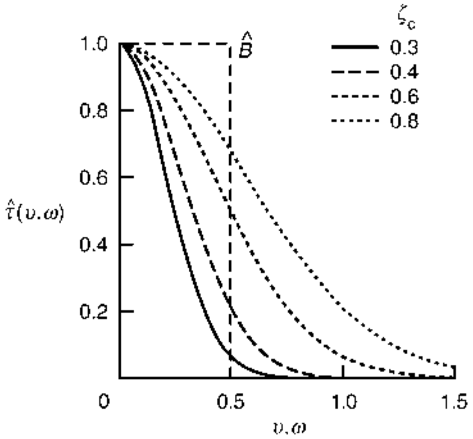

Figure 6 SFRs $\hat{\tau}(\upsilon, \omega)$ of the image-gathering device relative to the sampling passband $\hat{B}$ for unit sampling intervals (i.e., $X = Y = 1$).

features of the captured radiance field considered to be perceptually most significant, which usually entails some loss of information. The second stage, which is represented by the encoder, accounts for manipulations that seek to reduce redundancies in the acquired signal, without loss of information. The resultant encoded signal $s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa)$ may be expressed in the same form as eqn [4] for the acquired signal $s(\mathbf{x}, \mathbf{y})$ as

$$s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa) = s(\mathbf{x}, \mathbf{y}) \otimes \tau_{\mathrm{e}}(\mathbf{x}, \mathbf{y}) + n_q(\mathbf{x}, \mathbf{y}; \kappa) \qquad [16]$$

where the symbol $\otimes$ denotes discrete spatial convolution, and $n_q(\mathbf{x}, \mathbf{y}; \kappa)$ is the quantization noise. If the data transmission and/or storage are error-free, as is assumed here, then the signal that leaves the decoder remains identical to the signal that enters the encoder.

The DFT of $s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa)$ is the periodic spectrum $\tilde{s}_{\mathrm{e}}(\upsilon, \omega; \kappa)$ given in the same form as eqn [10] by

$$\tilde{s}_{\mathrm{e}}(\upsilon, \omega; \kappa) = K\hat{\rho}(\upsilon, \omega)\hat{\Gamma}_{\mathrm{e}}(\upsilon, \omega) + \hat{n}_{\mathrm{e}}(\upsilon, \omega; \kappa) \qquad [17]$$

where

$$\hat{\Gamma}_{\mathrm{e}}(\upsilon, \omega) = \hat{\tau}(\upsilon, \omega)\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega) \qquad [18]$$

is the throughput SFR from image gathering to transmission, and

$$\hat{n}_{\mathrm{e}}(\upsilon, \omega; \kappa) = \hat{n}(\upsilon, \omega)\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega) + \tilde{n}_q(\upsilon, \omega; \kappa) \qquad [19]$$

is the accumulated noise in the encoded signal. For η-bit quantization, where $\eta = \log \kappa$ and κ is the number of uniformly spaced quantization levels, $\tilde{n}_q(\upsilon, \omega; \kappa)$ may be treated (for $\eta \geq 4$) as independent Gaussian noise with the PSD

$$\tilde{\Phi}_q(\upsilon, \omega; \kappa) = \sigma_q^2 = (\sigma_{s_{\mathrm{e}}}/\kappa)^2 \qquad [20]$$

where $\sigma_{s_{\mathrm{e}}}^2$ is the variance given approximately by

$$\sigma_{s_{\mathrm{e}}}^2 \approx K^2 \int_{\hat{B}}\int \hat{\Phi}_\rho(\upsilon, \omega)|\hat{\Gamma}_{\mathrm{e}}(\upsilon, \omega)|^2 \mathrm{d}\upsilon \mathrm{d}\omega \qquad [21]$$

Image Restoration

Image restoration, as depicted in Figure 1, transforms the decoded signal $s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa)$ into the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$, as modeled by the expression

$$r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa) = \overline{|||}s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa) \otimes \Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa) \qquad [22]$$

The operation indicated by $\overline{|||}s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa)$ (called zero padding) prepares the decoded signal for image restoration with interpolation, and the operator $\Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$ produces the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$ on the interpolation lattice $\overline{|||}$ that is Z times denser than the sampling lattice $\underline{|||}$. If the interpolation in the image restoration is sufficiently dense (normally, for $Z \geq 4$), then the blurring and raster effects of the image-display process are suppressed entirely. This condition leads not only to the best realizable image quality but also to the simplest expression for the continuous displayed image, as given in the same form as eqns [10] and [17] by

$$\hat{R}(\upsilon, \omega; \kappa) = \hat{\rho}(\upsilon, \omega)\hat{\Gamma}_r(\upsilon, \omega; \kappa) + \hat{n}_r(\upsilon, \omega; \kappa) \qquad [23]$$

where

$$\hat{\Gamma}_r(\upsilon, \omega; \kappa) = \hat{\tau}(\upsilon, \omega)\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega)\hat{\Psi}(\upsilon, \omega; \kappa) \qquad [24]$$

is the throughput SFR from image gathering to display, and

$$\begin{aligned}\hat{n}_r(\upsilon, \omega; \kappa) &= K^{-1}\hat{n}_{\mathrm{e}}(\upsilon, \omega; \kappa)\hat{\Psi}(\upsilon, \omega; \kappa) \\ &= \hat{n}_{\mathrm{ar}}(\upsilon, \omega; \kappa) + \hat{n}_{pr}(\upsilon, \omega; \kappa) \\ &\quad + \hat{n}_{qr}(\upsilon, \omega; \kappa)\end{aligned} \qquad [25]$$

is the accumulated noise in the restored image. This noise may be expressed explicitly by substituting the noise components given above. The noise due to the granularity in image-display mediums is not accounted for here.

The Wiener filter, that minimizes the mean-square error due to the perturbations in image gathering and coding, is given by

$$\hat{\Psi}(\upsilon, \omega; \kappa) = \frac{\hat{\Phi}'_\rho(\upsilon, \omega)\hat{\Gamma}^*_{\mathrm{e}}(\upsilon, \omega)}{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_{\mathrm{e}}(\upsilon, \omega)|^2 + \hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega; \kappa)} \qquad [26]$$

where $\hat{\Phi}'(\upsilon, \omega) = \sigma_\rho^{-2}\hat{\Phi}_\rho(\upsilon, \omega)$ is the normalized PSD of the radiance field and

$$\begin{aligned}\hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega; \kappa) = (K\sigma_\rho)^{-2}\Big[&\hat{\Phi}_n(\upsilon, \omega)|\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega)|^2 \\ &+ \tilde{\Phi}_q(\upsilon, \omega; \kappa)\Big]\end{aligned}$$

is the normalized PSD of the accumulated noise in the encoded signal. If the photodetector noise is white so that the PSD $\tilde{\Phi}_p(\upsilon, \omega)$ is equal to its variance σ_p^2 and the PSD $\tilde{\Phi}_q(\upsilon, \omega; \kappa)$ is equal to its variance σ_q^2, then the PSD of the noise simplifies to

$$\hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega; \kappa) = \hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega) + \tilde{\Phi}'_q(\upsilon, \omega; \kappa) \qquad [27]$$

where

$$\begin{aligned}\hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega) = \Big[&\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\tau}(\upsilon, \omega)|^2 * \hat{\underline{|||}}_{\mathrm{s}} \\ &+ (K\sigma_\rho/\sigma_p)^{-2}\Big]|\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega)|^2\end{aligned} \qquad [28]$$

and

$$\hat{\Phi}'_q(\upsilon, \omega; \kappa) = (K\sigma_\rho/\sigma_{s_{\mathrm{e}}})^{-2}\kappa^{-2} \qquad [29]$$

These simplifying conditions permit the Wiener filter given above and the figures of merit given below to be expressed as a function of the SNR $K\sigma_\rho/\sigma_p$, which can be specified without accounting explicitly for either the gain K or the variances σ_ρ^2 and σ_p^2.

Figures of Merit

Information Rate $\mathcal{H}$

The information rate $\mathcal{H}$ (in bits sample^{-1}) of the encoded signal $s_\mathrm{e}(\mathbf{x}, \mathbf{y}; \kappa)$ is defined by

$$\mathcal{H} = \mathcal{E}[\tilde{s}_\mathrm{e}(\upsilon, \omega; \kappa)] - \mathcal{E}[\tilde{s}_\mathrm{e}(\upsilon, \omega; \kappa) | \hat{\rho}(\upsilon, \omega) : (\upsilon, \omega) \in \hat{B}] \qquad [30]$$

where the first term, $\mathcal{E}[\cdot]$, represents the entropy of $\tilde{s}_\mathrm{e}(\upsilon, \omega; \kappa)$ and the second term, $\mathcal{E}[\cdot|\cdot]$, represents the conditional entropy of this signal when the reflectance spectrum $\hat{\rho}(\upsilon, \omega)$ within the sampling passband $\hat{B}$ is known. For the assumptions that have been made about image gathering and coding, the conditional entropy becomes the entropy of the accumulated noise $\hat{n}_\mathrm{e}(\upsilon, \omega; \kappa)$, so that $\mathcal{H}$ simplifies to

$$\begin{aligned}\mathcal{H} &= \mathcal{E}[\tilde{s}_\mathrm{e}(\upsilon, \omega; \kappa)] - \mathcal{E}[\hat{n}(\upsilon, \omega; \kappa) : (\upsilon, \omega) \in \hat{B}] \\ &= \frac{1}{2}\int_{\hat{B}}\int \log\left[1 + \frac{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_\mathrm{e}(\upsilon, \omega)|^2}{\hat{\Phi}'_{n_\mathrm{e}}(\upsilon, \omega; \kappa)}\right] \mathrm{d}\upsilon\, \mathrm{d}\omega\end{aligned} \qquad [31]$$

The theoretical upper bound of $\mathcal{H}$ is Shannon's channel capacity $\mathcal{C}$ for a bandwidth-limited system, with white photodetector noise and an average power limitation (here the variance σ_ρ^2), as given by

$$\mathcal{C} = \frac{1}{2}\log\left[1 + (K\sigma_\rho/\sigma_p)^2\right] \qquad [32]$$

$\mathcal{H}$ can reach $\mathcal{C}$ only when

$$\hat{\Phi}_\rho(\upsilon, \omega) = \begin{cases}\sigma_\rho^2, & (\upsilon, \omega) \in \hat{B} \\ 0, & \text{elsewhere}\end{cases}$$

$$\hat{\tau}(\upsilon, \omega) = \begin{cases}1, & (\upsilon, \omega) \in \hat{B} \\ 0, & \text{elsewhere}\end{cases} \qquad [33]$$

However, neither of these two conditions can occur in practice because both the PSD $\hat{\Phi}_\rho(\upsilon, \omega)$ of natural scenes and the SFR $\hat{\tau}(\upsilon, \omega)$ of image-gathering devices decrease gradually with increasing spatial frequency, as depicted in Figures 3 and 6, respectively. Hence, $\mathcal{H}$ can never reach $\mathcal{C}$.

Theoretical Minimum Data Rate $\mathcal{E}$

The theoretical minimum data rate $\mathcal{E}$ (also in bits sample^{-1}) that is required to convey the information rate $\mathcal{H}$ is defined by

$$\mathcal{E} = \mathcal{E}[\tilde{s}_\mathrm{e}(\upsilon, \omega; \kappa)] - \mathcal{E}[\tilde{s}_\mathrm{e}(\upsilon, \omega; \kappa) | \tilde{s}(\upsilon, \omega)] \qquad [34]$$

where the second term, $\mathcal{E}[\cdot|\cdot]$ is the uncertainty of $\tilde{s}_\mathrm{e}(\upsilon, \omega; \kappa)$, when $\tilde{s}(\upsilon, \omega)$ is known. For the assumptions that have been made about the coding process, the conditional entropy becomes the entropy $\mathcal{E}[\tilde{n}_q(\upsilon, \omega; \kappa)]$ of the quantization noise, so that $\mathcal{E}$ simplifies to

$$\begin{aligned}\mathcal{E} &= \mathcal{E}[\tilde{s}_\mathrm{e}(\upsilon, \omega; \kappa)] - \mathcal{E}[\tilde{n}_q(\upsilon, \omega; \kappa)] \\ &= \frac{1}{2}\int_{\hat{B}}\int \log\left[1 + \frac{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_\mathrm{e}(\upsilon, \omega)|^2 + \hat{\Phi}'_{n_\mathrm{e}}(\upsilon, \omega)}{\tilde{\Phi}'_q(\upsilon, \omega; \kappa)}\right] \mathrm{d}\upsilon\, \mathrm{d}\omega\end{aligned} \qquad [35]$$

This expression for $\mathcal{E}$ represents the entropy of completely decorrelated data.

Information Efficiency $\mathcal{H}/\mathcal{E}$

The information efficiency of completely decorrelated data is defined by the ratio $\mathcal{H}/\mathcal{E}$. This ratio reaches its upper bound $\mathcal{H}/\mathcal{E} = 1$ when $\hat{\Phi}'_{n_\mathrm{e}}(\upsilon, \omega) \ll \hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_\mathrm{e}(\upsilon, \omega)|^2$ and $\hat{\Phi}'_{n_\mathrm{e}}(\upsilon, \omega) \ll \tilde{\Phi}'_q(\upsilon, \omega; \kappa)$, because then both eqn [31] for $\mathcal{H}$ and eqn [35] for $\mathcal{E}$ reduce to the same expression

$$\mathcal{H} = \mathcal{E} = \frac{1}{2}\int_{\hat{B}}\int \log\left[1 + \frac{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_\mathrm{e}(\upsilon, \omega)|^2}{\tilde{\Phi}'_q(\upsilon, \omega; \kappa)}\right] \mathrm{d}\upsilon\, \mathrm{d}\omega \qquad [36]$$

This upper bound can be approached with a small loss in $\mathcal{H}$ only when the aliasing and photodetector noise are small. Hence, the electro-optical design that optimizes $\mathcal{H}$ also optimizes $\mathcal{H}/\mathcal{E}$. However, a compromise always remains between $\mathcal{H}$ and $\mathcal{H}/\mathcal{E}$ in the selection of the number of quantization levels: $\mathcal{H}$ favors fine quantization, whereas $\mathcal{H}/\mathcal{E}$ favors coarse quantization.

Maximum-Realizable Fidelity $\mathcal{F}$

The maximum-realizable fidelity $\mathcal{F}$ of the images restored with the Wiener filter $\hat{\Psi}(\upsilon, \omega; \kappa)$ can be expressed as a function of the throughput SFR $\hat{\Gamma}_r(\upsilon, \omega; \kappa)$ given by eqn [24] as

$$\mathcal{F} = \int\int \hat{\Phi}'_\rho(\upsilon, \omega)\hat{\Gamma}_r(\upsilon, \omega; \kappa)\, \mathrm{d}\upsilon\, \mathrm{d}\omega \qquad [37]$$

or, equivalently, as a function of the integrand $\hat{\mathcal{H}}(\upsilon, \omega; \kappa)$ of the information rate $\mathcal{H}$ given by eqn [31] as

$$\mathcal{F} = \int\int \hat{\Phi}'_{\rho}(\upsilon, \omega)\left[1 - 2^{-\hat{\mathcal{H}}(\upsilon,\omega;\kappa)}\right] \mathrm{d}\upsilon \mathrm{d}\omega \qquad [38]$$

This relationship between fidelity and information rate can be extended to other goals. For example, the enhancement of spatial features (e.g., edges and boundaries), for robotic vision, can be accommodated by letting $\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega)$ be the desired feature-enhancement filter.

Performance and Design Trade-offs

Electro-optical Design

Figure 7 presents the information rate $\mathcal{H}$ (for lossless encoding) as a function of the optical-response index ζ_{c} and the mean spatial detail μ. The information rate $\mathcal{H}$ versus ζ_{c} is given for several SNRs $K\sigma_{\rho}/\sigma_{p}$ and $\mu = 1$, and $\mathcal{H}$ versus μ is given for the three informationally optimized designs specified in Table 1.

The curves show that $\mathcal{H}$ can reach its highest possible value only when both of the following conditions are met:

1. The relationship between the SFR $\hat{\tau}(\upsilon, \omega)$ and the sampling passband $\hat{B}$ is chosen to optimize $\mathcal{H}$ for the available SNR $K\sigma_{\rho}/\sigma_{p}$. This design is said to be informationally optimized.
2. The relationship between the sampling passband $\hat{B}$ and the PSD $\hat{\Phi}_{\rho}(\upsilon, \omega)$ is chosen to optimize $\mathcal{H}$. For the PSDs $\hat{\Phi}_{\rho}(\upsilon, \omega)$ typical of natural scenes, the best match occurs when the sampling interval is near the mean spatial detail of the scene (i.e., when $\mu \approx 1$).

These conditions are consistent with those for which the information rate $\mathcal{H}$ reaches Shannon's channel capacity $\mathcal{C}$. However, blurring and aliasing constrain $\mathcal{H}$ to values that are smaller than $\mathcal{C}$ by a factor of nearly two (i.e., $\mathcal{H} \lesssim \mathcal{C}/2$). Hence, these conditions replace the role of white and band-limited signals in classical communication theory, to establish an upper bound on the information rate in imaging systems.

The above two conditions appeal intuitively when the problem of image restoration is considered:

1. The result that the SFR $\hat{\tau}(\upsilon, \omega)$ that optimizes $\mathcal{H}$ depends on the available SNR is consistent with the observation that, in one extreme, when the SNR is low, substantial blurring should be avoided because the photodetector noise would constrain the enhancement of fine spatial detail even if aliasing were negligible. In the other extreme, when the SNR is high, substantial aliasing should be avoided so that this enhancement is relieved from any constraints except those that the sampling passband inevitably imposes.

Table 1 Informationally optimized designs

Design	$K\sigma_{\rho}/\sigma_{p}$	ζ_{c}	$\mathcal{H}^{*}$
1	256	0.3	4.4
2	64	0.4	3.3
3	16	0.5	2.2

$^{*}\mu = 1$.

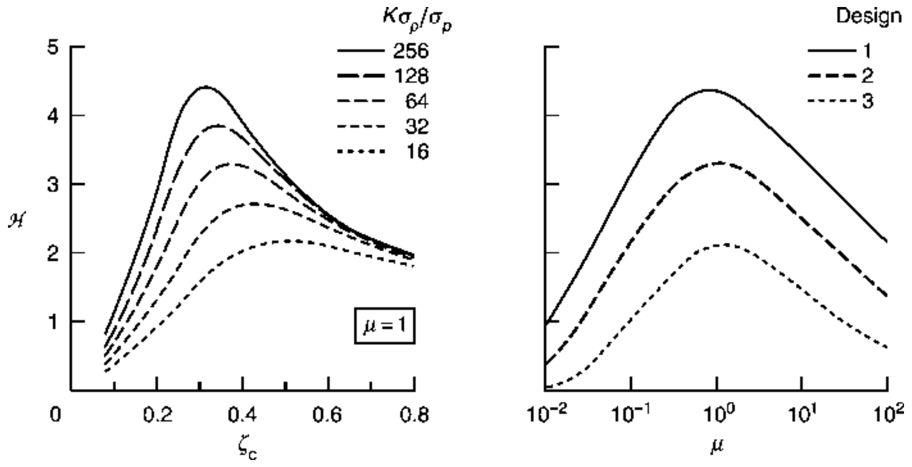

Figure 7 Information rate $\mathcal{H}$ versus the optical-response index ζ_{c} and the mean spatial detail μ (relative to the sampling interval). The designs are specified in Table 1, and the SFRs $\hat{\tau}(\upsilon, \omega)$ for ζ_{c} are shown in Figure 6.

2. The result that $\mathcal{H}$ reaches its highest value when the sampling interval is near the mean spatial detail of the scene is consistent with the observation that, ordinarily, it would not be possible to restore spatial detail that is finer than the sampling interval, whereas it would be possible to restore much coarser detail from fewer samples.

Information Rate and Data Rate

Figure 8 presents an information-entropy $\mathcal{H}(\mathcal{E})$ plot that characterizes the relationship between the information rate $\mathcal{H}$ of the encoded signal and the associated theoretical minimum data rate, or entropy, $\mathcal{E}$ for η-bit quantization. The encoder SFR $\tilde{\tau}_e(\upsilon, \omega)$ is assumed to be unity. The plot illustrates the trade-off between $\mathcal{H}$ and $\mathcal{E}$ in the selection of the number of quantization levels for encoding the acquired signal. It shows, in particular, that the design that realizes the highest information rate $\mathcal{H}$ also enables the encoder to realize the highest information efficiency $\mathcal{H}/\mathcal{E}$. This result appeals intuitively, because the perturbations due to aliasing and photodetector noise that interfere with the image restoration can also be expected to do so with the signal decorrelation.

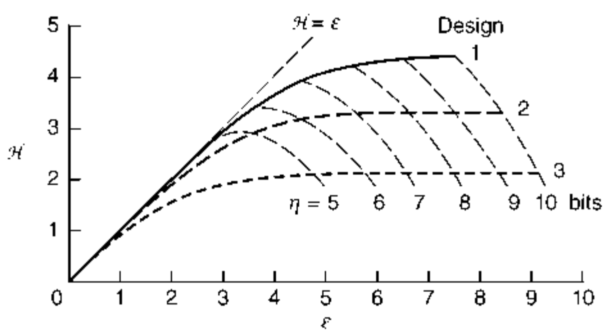

Figure 8 The information-entropy $\mathcal{H}(\mathcal{E})$ plot of the information rate $\mathcal{H}$ versus the associated theoretical minimum data rate $\mathcal{E}$ for η-bit quantization. The designs are specified in Table 1.

Throughput SFR

Figure 9 illustrates the SFRs $\hat{\tau}(\upsilon, \omega)$ and $\hat{\Psi}(\upsilon, \omega; \kappa)$ of the image-gathering device and Wiener filter, respectively, and of their product, the throughput SFR $\hat{\Gamma}_r(\upsilon, \omega; \kappa)$, for the three designs specified in Table 1. As the accumulated noise becomes negligible, $\hat{\Gamma}_r(\upsilon, \omega; \kappa)$ approaches the ideal SFR of the classical communication channel, which is unity inside the sampling passband and zero outside.

Information Rate, Fidelity and Robustness

The PSD $\hat{\Phi}_\rho(\upsilon, \omega)$ of a scene is seldom known when images are restored. Moreover, even if the PSD were known to be the one given by eqn [2], then the mean

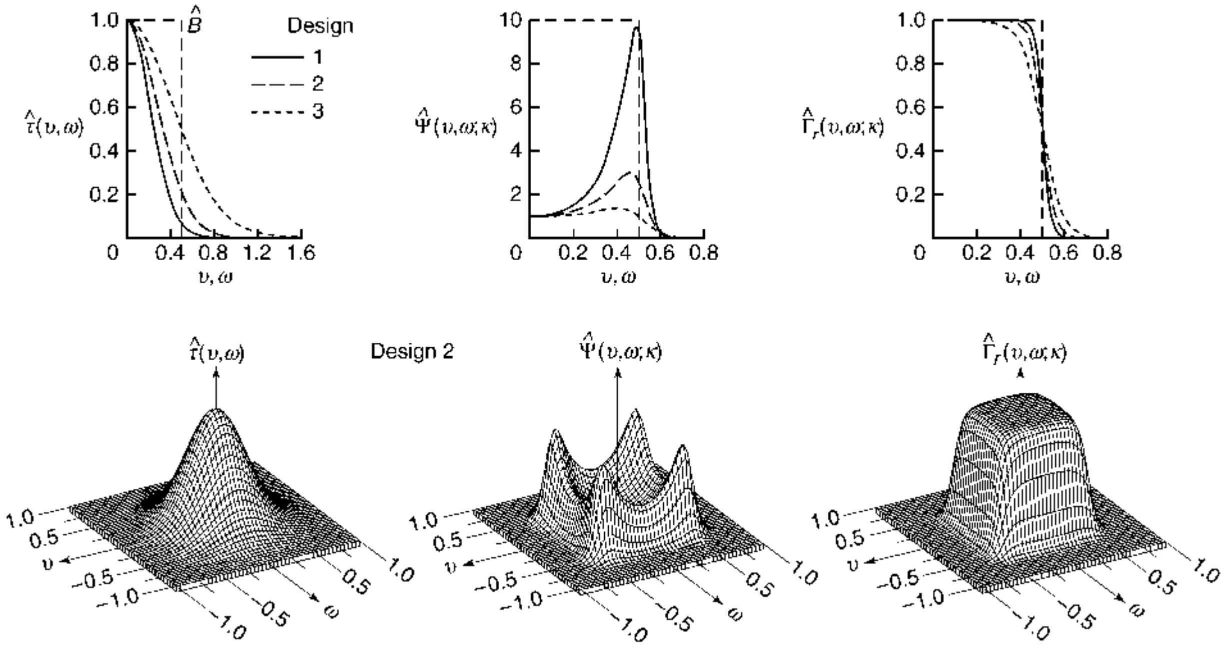

(a) Image gathering (b) Wiener filter (c) Throughput

Figure 9 SFRs of image gathering, restoration, and throughput. The designs are specified in Table 1.

spatial detail μ would usually remain uncertain because it depends, not only on the properties of the scene, but also on the angular resolution and viewing distance of the image-gathering device.

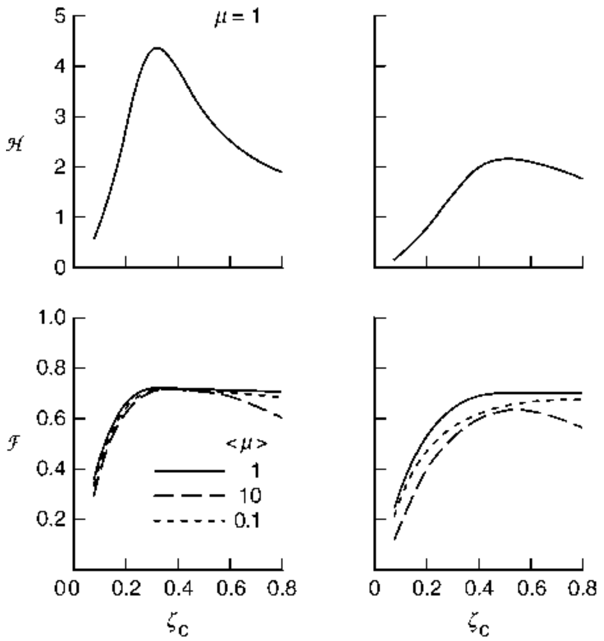

(a) $K\sigma_\rho/\sigma_p = 256$ (b) $K\sigma_\rho/\sigma_p = 16$

Figure 10 Information rate $\mathcal{H}$ and maximum-realizable fidelity $\mathcal{F}$ versus the optical-response index ζ_c for two SNRs $K\sigma_\rho/\sigma_p$. $\mathcal{F}$ is given for the matched ($\langle\mu\rangle = 1$) and two mismatched Wiener restorations.

In practice, therefore, image restorations have to rely on estimates of the statistical properties of scenes. The tolerance of the quality of the restored image to errors in these estimates is referred to as the robustness of the image restoration.

Figure 10 presents the information rate $\mathcal{H}$ and the corresponding maximum-realizable fidelity $\mathcal{F}$ for matched and mismatched Wiener restorations. These restorations represent the digital image $R(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$ or, equivalently, the continuous image $R(x, y; \kappa)$ displayed on a noise-free medium. The matched restorations use the correct value of μ, whereas the mismatched restorations use wrong estimates of μ. These curves reveal that the informationally optimized designs produce the highest fidelity and robustness, and that both improve with increasing $\mathcal{H}$. Moreover, these curves reveal that, ordinarily, one cannot go far wrong in Wiener restorations by assuming that the mean spatial detail is equal to the sampling interval (i.e., $\mu = 1$). This observation appeals intuitively because spatial detail that is much smaller cannot be resolved, and detail that is larger is not degraded substantially by blurring and aliasing.

Information Rate and Visual Quality

Figures 11 and 12 present Wiener restorations and their three components, as given by eqn [23], where $\rho(x, y) * \Gamma_r(x, y; \kappa)$ accounts for blurring, $n_{ar}(x, y; \kappa)$ accounts for aliasing, and $n_{pr}(x, y; \kappa)$ accounts for

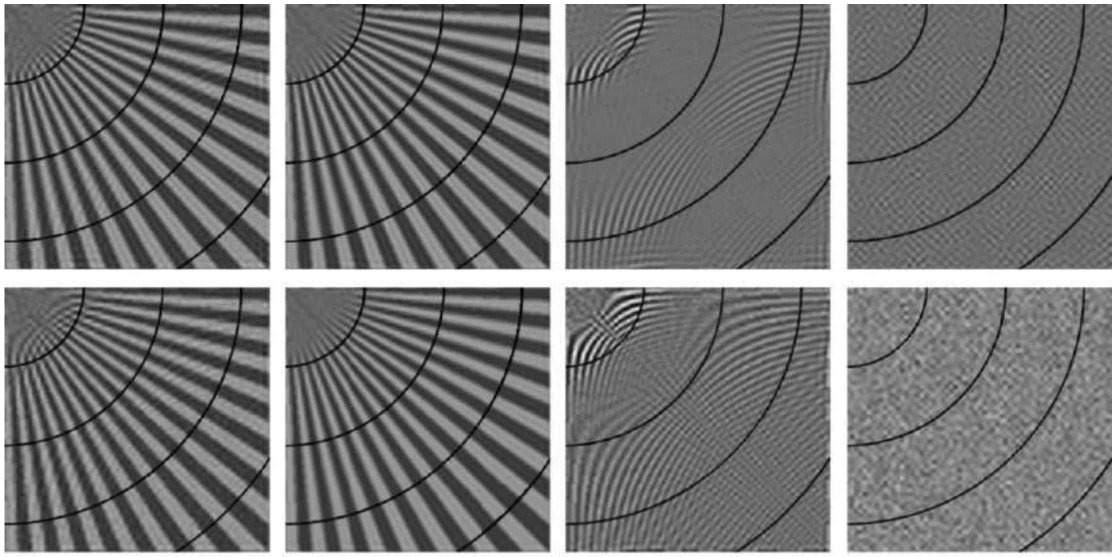

(a) $R(x,y;\kappa)$ (b) $\rho(x,y) * \Gamma_r(x,y;\kappa)$ (c) $4n_{ar}(x,y;\kappa)$ (d) $4n_{pr}(x,y;\kappa)$

Figure 11 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

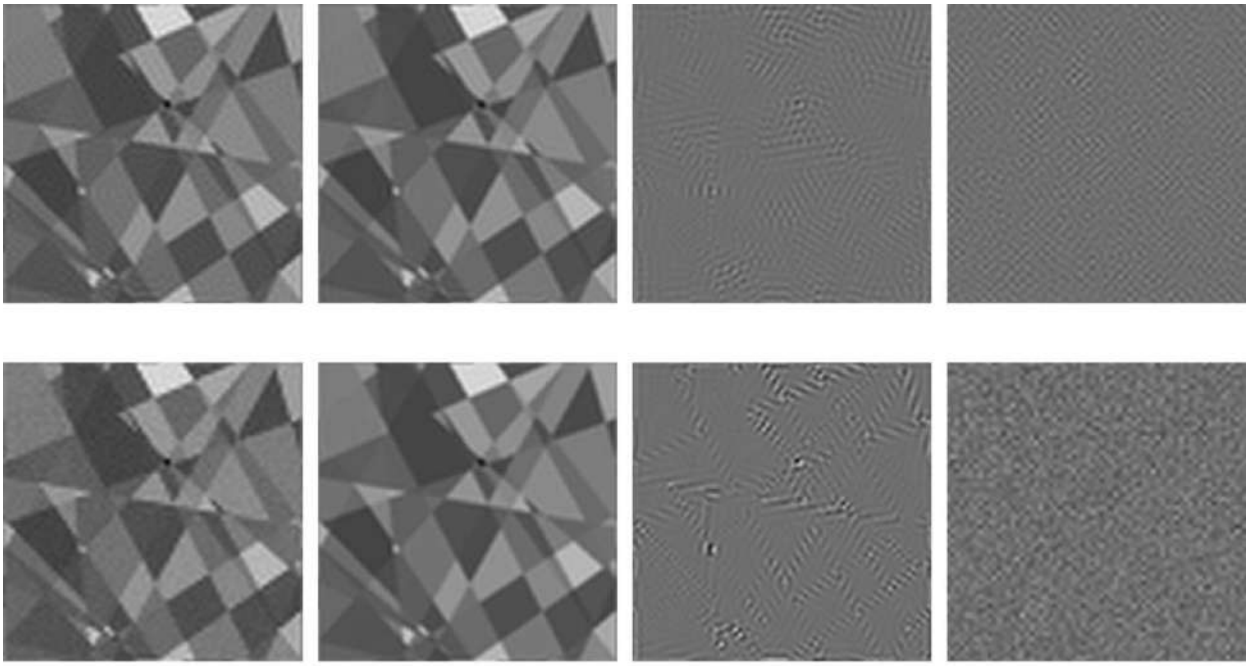

(a) $R(x,y;\kappa)$ (b) $\rho(x,y) * \Gamma_r(x,y;\kappa)$ (c) $4n_{ar}(x,y;\kappa)$ (d) $4n_{pr}(x,y;\kappa)$

Figure 12 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

photodetector noise. The contrast of the illustrations of the two noise components has been increased by a factor of four to improve their visibility. The extraneous structure that aliasing produces in the images of the resolution wedges is commonly referred to as Moiré pattern. The corresponding distortion in the images of the random polygons emerges more subtly as jagged (or staircase) edges.

The images are formed from a set of 64×64 samples with a $Z = 4$ interpolation and have a 4×4 cm format. If these images were displayed in a smaller format, then the jagged edges could not be resolved by the observer and would appear to be the result of blurring or photodetector noise instead of aliasing. Aliasing has, therefore, often been overlooked as a significant source of image degradation.

The Wiener restoration produces images in which fine spatial detail near the limit of resolution normally has a high contrast approaching that in the scene. However, these images also tend to contain visually annoying defects, due to the accumulated noise and the ringing near sharp edges (Gibbs phenomenon). Hence, it is often desirable to combine this restoration with an enhancement filter that gives the user some control over the trade-off among fidelity, sharpness and clarity.

Figure 13 presents images for which the Wiener restorations given in Figures 11 and 12 are enhanced for visual quality by suppressing some of the defects due to aliasing and ringing. This enhancement improves clarity at the cost of a small loss in sharpness. The images are produced again from 64×64 pixels, but they are displayed in three different formats. The smallest format, with a density of 64 pixels/cm, corresponds to a display of 256×256 pixels in a 4×4 cm area, which is typical of the images found in the prevalent digital image-processing literature. As can be observed, this small format leads to images that are indistinguishable from each other because they hide many of the distortions that larger formats clearly reveal.

In general, if the images are displayed in a format that is large enough to reveal the finest spatial detail near the limit of resolution of the imaging system, then distortions due to the perturbations in this system become also visible. When no effort is spared to reduce these distortions without a perceptual loss of resolution and sharpness, then it becomes strongly evident that the best visual quality with which images can be restored improves with increasing information rate, even after the fidelity has essentially reached its upper bound. This improvement continues until it is gradually ended by the unavoidable compromise among sharpness, aliasing, and ringing as well as by the granularity of the image display.

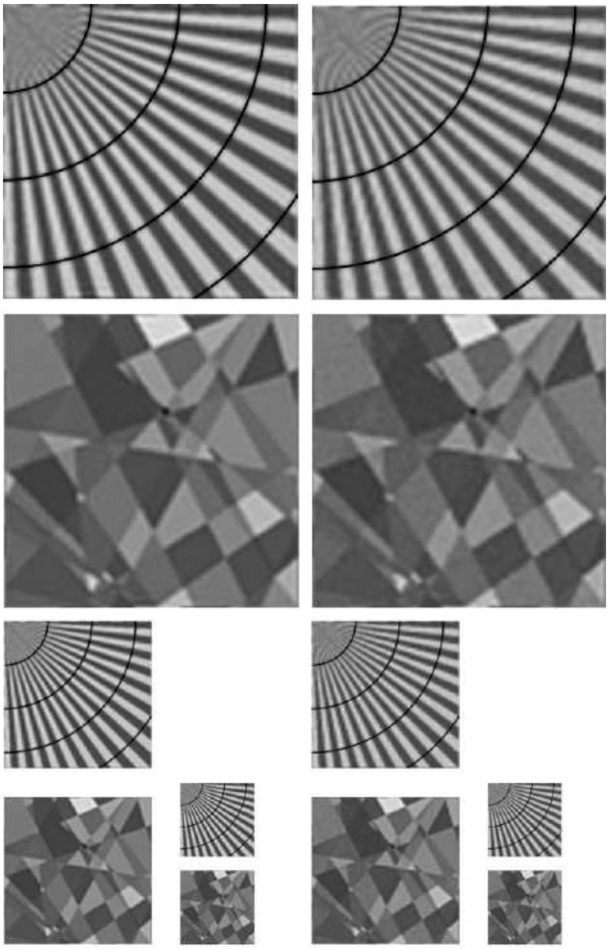

(a) Design 1 (b) Design 3

Figure 13 Images restored with the Wiener filter and enhanced for visual quality. The large, medium, and small formats contain 16, 32 and 64 pixels/cm, respectively.

Concluding Remarks

The performance of image-gathering devices and the human eye is constrained by the same critical limiting factors. When these factors are accounted for properly, then it emerges that the design of the image-gathering device that is optimized for the highest realizable information rate corresponds closely, under appropriate conditions, to the design of the human eye that evolution has optimized for viewing the real world. This convergence of information theory and evolution toward the same design clearly supports the extension of information theory to the characterization of imaging systems.

Numerous other information-theoretic characterizations of imaging are now emerging. However, none of these characterizations accounts for the critical limiting factors that constrain image gathering and restoration, as outlined here. Hence, it is not possible for them to address the efficiency and accuracy with which images can be conveyed. Moreover, none of these characterizations combines observation with communication, or sensing with processing, in a general formalization. Instead, each of them still deals with a particular measurement or imaging problem. Hence, the role of information theory in imaging is far from mature. Many challenges remain: to develop general formalizations, to account for physical phenomena, and to correlate predictions with measurements.

See also

Information Processing: Optical Digital Image Processing. **Modulators:** Electro-optics.

Further Reading

Fales CL, Huck FO, Alter-Gartenberg R and Rahman Z (1996) Image gathering and digital restoration. *Philosophical Transactions of the Royal Society of London* A354: 2249–2287.

Huck FO and Fales CL (2001) Characterization of image systems. *Encyclopedia of Imaging Science and Technology*. New York: John Wiley & Sons.

Huck FO, Fales CL and Rahman Z (1996) An information theory of visual communication. *Philosophical Transactions of the Royal Society of London* A354: 2193–2248.

Huck FO, Fales CL and Rahman Z (1997) *Visual Communication: An Information Theory Approach*. Boston, CA: Kluwer Academic Publishers.

Lesurf JCG (1995) *Information and Measurement*. Bristol, UK: Institute of Physics Pub.

O'Sullivan JA, Blahut RE and Snyder DL (1998) Information-theoretic image formation. *IEEE Transactions Information Theory* 44: 2094–2123.

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423 and 28: 623–656.

Shannon CE and Weaver W (1964) *The Mathematical Theory of Communication*. Urbana, IL: U. Illinois Press.

Wiener N (1949) *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: John Wiley & Sons.

Inverse Problems and Computational Imaging

M Bertero and P Boccacci, University of Genova, Genova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A digital image is an array (2D image) or a cube (3D image) of quantized numbers, providing a representation of a physical object or a map of some distributed property of the object (density of matter, energy, etc.). It can be obtained by digitizing the original image (for instance, a photograph) or can be directly formed by the outputs of the imaging system (digital camera, microscope, telescope, etc.). When visualized on a suitable image display device, a digital image must provide a representation of the physical object which can be easily interpreted and analyzed. In many cases, however, this condition is not satisfied. We give two typical examples.

The first is that of an image degraded by blurring which can be introduced by defects of the optical systems (such as aberrations) or by image motion, image defocusing, etc. The representation of the object does not have the quality which should be desirable for its interpretation.

The second case is even more important and arises when the output of the imaging system is not directly related to the physical quantity to be imaged. For instance, in X-ray tomography, the output of a detector is the attenuation of an X-ray pencil crossing the body while the quantity to be imaged is the density of the body.

In the situations described above, the desired image can be obtained by solving a number of mathematical problems. If we denote f as the object, namely the target of the imaging system, and g as the image of f provided by the system, then the problems to be solved can be summarized as follows:

- develop a physical model of the system to obtain a mathematical relationship between the object f and the image g; the computation of g from a given f is usually called the solution of the direct problem;
- solve the problem of obtaining the physical quantity f from given outputs g of the system; this second step is usually called the solution of the inverse problem.

The digital image of f, obtained by solving an inverse problem, is a computed one, generated by the computer where the algorithm for the solution of the inverse problem has been implemented.

It may be convenient to formulate both the direct and the inverse problem in terms of functions rather than arrays, cubes, etc., because such a formulation allows the use of powerful tools of functional analysis for their investigation. Next, the results obtained in the continuous case can be used for understanding the features of the corresponding discrete problem. In most important applications, for instance, the relationship between f and g is linear so that, in the discrete version of the direct problem, one obtains g by applying a suitable matrix A to f while, in the discrete version of the inverse problem, one obtains f from g by solving a linear algebraic system. However, the solution of the inverse problem is not so simple and the difficulties can be understood by looking at the corresponding continuous problem.

The basic reason relies on the fact that inverse problems are ill-posed in the sense of Hadamard: the solution may not exist; even if it exists it may not be unique; and, even if it exists and is unique, it may not depend continuously on the data. The last statement means that a small perturbation of the data, such as that caused by noise or any kind of experimental error, can completely modify the solution.

Inverse problems are ill-posed because imaging systems do not transmit complete information about the physical object. Therefore, the problem is to extract the useful information contained in the data or, in mathematical terms, to look for approximate solutions which are stable against noise. This is the purpose of a mathematical theory introduced in its general form by the Russian mathematician, Tikhonov and known as regularization theory.

We illustrate the general framework outlined above by means of a specific example – image deconvolution, also known as image deblurring, image restoration, etc. The object f is the image which should be recorded in the absence of degradation, while g is the image corrupted by aberrations or other causes of blurring. Both f and g are functions of 2D (or 3D) space variables $\mathbf{x}$ which are the coordinates of a point in the image domain. If the imaging system is isoplanatic, then it is described by a space-invariant point spread function (PSF) $K(\mathbf{x})$, which is the image of a point source located in the center of the image domain. The PSF provides the response of the system to any point source wherever it is located. On the other hand, its Fourier transform, the transfer function $\hat{K}(\omega)$ (the hat denotes Fourier transform and ω, the spatial frequencies associated to the space variables $\mathbf{x}$), tells us how a signal of a fixed frequency is propagated through the linear system, so that the blurring can also be viewed as a sort of frequency filtering.

Under these assumptions the image g is the convolution product of object f and PSF K:

$$g(\mathbf{x}) = \int K(\mathbf{x} - \mathbf{x}')f(\mathbf{x}')\mathrm{d}\mathbf{x}' \qquad [1]$$

and the solution of the direct problem is just the computation of a convolution product.

The image g, as given by eqn [1], is the so-called noise-free image. The actually recorded image g_r is affected by a noise term and, in general, it can be written in the following form:

$$g_r(\mathbf{x}) = g(\mathbf{x}) + n(\mathbf{x}) \qquad [2]$$

where $n(\mathbf{x})$ is a function describing noise contamination. It is a random function and, therefore, it is unknown even if one understands its statistical properties. We point out that eqn [2] does not imply that the noise is additive or signal independent. For instance, in microscopy and astronomy, the contamination of the image is due both to photon noise, which satisfies Poisson statistics, and read-out noise, which is basically Gaussian and white. Therefore the noise term in eqn [2] must only be intended as the difference between the noisy and the noise-free image.

In a first attempt at approaching the inverse problem, it seems quite natural to ignore the noise term $n(\mathbf{x})$ and to solve eqn [1] for $f(\mathbf{x})$ with $g(\mathbf{x})$ replaced by $g_r(\mathbf{x})$. Then, by taking the Fourier transform of both sides of this equation and using the well-known convolution theorem, one finds the following relationship:

$$\hat{g}_r(\omega) = \hat{K}(\omega)\hat{f}(\omega) \qquad [3]$$

which relates the Fourier transform of f and g to the transfer function of the imaging system.

The solution of this equation looks elementary. However, a first problem is due to the fact that an optical system is generally band-limited; its band Ω is the bounded domain of spatial frequencies where the transfer function is different from zero. Since the noise is not band-limited or, at least, has a band much broader than that of the optical system, outside Ω the right-hand side of eqn [3] is zero while the left-hand side is not; in other words, the solution of the deconvolution problem, as formulated above, does not exist because no object f satisfies eqn [3] everywhere. In this particular case, one can circumvent the problem by applying a suitable filter to $g_r(\mathbf{x})$, in order to suppress the out-of-band noise.

The second problem is the nonuniqueness of the solution of the problem with the filtered data. This is due to the so-called invisible objects, namely objects whose Fourier transform is zero on Ω, so that the corresponding images are exactly zero. If we find a solution of the deconvolution problem, by adding an arbitrary invisible object to this solution, we find another solution of the same problem. A well-defined solution can be obtained by requiring its Fourier transform to be zero outside Ω. In other words, we set to zero what is not transmitted by the imaging system.

The standard way for approaching the previous questions is to look for solutions in the least-squares sense, namely for objects which minimize the functional:

$$\varepsilon^2(f) = \|K * f - g_r\|^2 \qquad [4]$$

where $*$ denotes the convolution product, as defined in eqn [1], and the right-hand side is the square of the L^2-norm (which can also be interpreted as energy) of the discrepancy between the recorded image g_r and the computed image $K * f$. If one looks for a least-squares solution with minimal energy then one automatically re-obtains the solution discussed above, namely an object whose Fourier transform in Ω is given by

$$\hat{f}_{\mathrm{inv}}(\omega) = \frac{\hat{g}_r(\omega)}{\hat{K}(\omega)} \qquad [5]$$

and is zero outside Ω. Such a solution is that provided by the so-called inverse filter.

However, there is an additional difficulty in that the transfer function usually tends to zero at the boundary of the band while the Fourier transform of the noise, hence that of g_r, does not, or tends to zero in a different way. It follows that $\hat{f}_{\mathrm{inv}}(\omega)$ is divergent or very large at the boundary of the band. We conclude that its inverse Fourier transform does not exist or, if it exists, is affected by large artifacts due to noise propagation. The latter is the typical situation in the case of digital images. In Figure 1 we give the result of a numerical simulation showing the typical effect produced by the inverse filter. The object is a beautiful picture of a galaxy recorded by the Hubble Space Telescope (HST), while the PSF is a computed one describing the blurring of HST before installation of the optics correction. The blurred image is obtained by convolving the object with the PSF and by adding white noise. The restored image produced by the inverse filter is completely corrupted by noise.

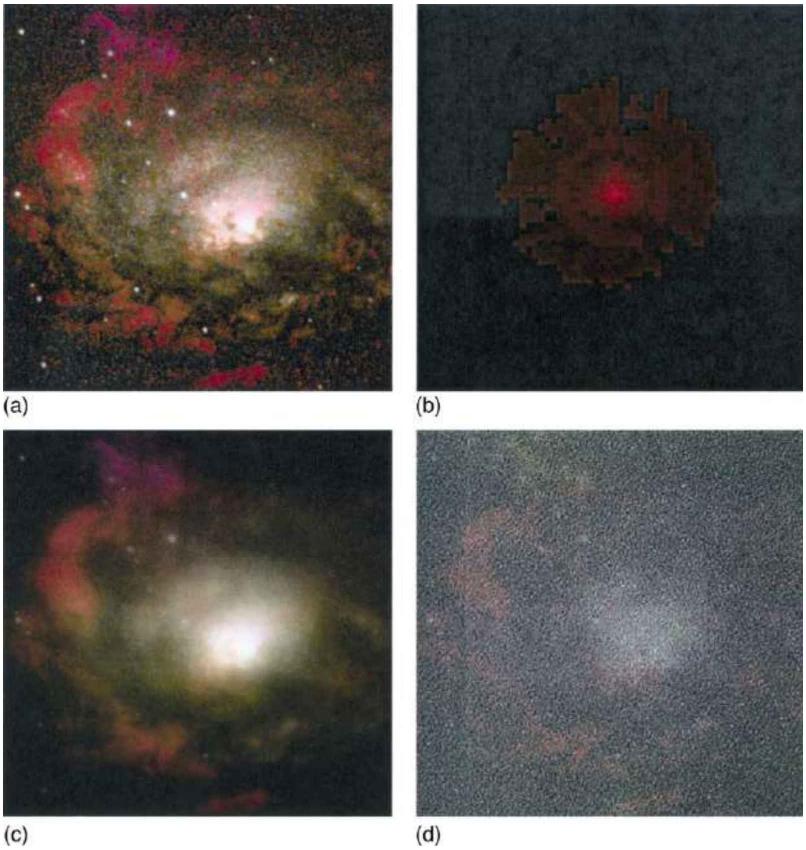

Figure 1 Illustrating the effect of the inverse filter: (a) Image of the Circinus galaxy taken on 10 April 1999, with the Wide Field Planetary Camera 2 of the Hubble Space Telescope; (b) the PSF used for blurring the RGB components of the image (here shown in red); (c) the blurred image obtained by convolving the components of the image in (a) with the corresponding PSFs and adding noise; (d) the restoration obtained by applying the inverse filter to the three components of (c). Part (a) courtesy of NASA/Space Telescope Science Institute.

The inverse filter provides a solution which fits the data in the best possible way and therefore also fits the noise in the frequency domains where the signal is weakly transmitted by the imaging system. This can be achieved at the cost of an energy (L^2-norm) of the solution which is too large. Therefore, such a conclusion suggests that one must search for a reasonable compromise between data fitting and energy of the solution. This is the basic idea of regularization theory, at least in its most simple form. The mathematical formulation is obtained by looking for objects which minimize the functional:

$$\Phi_\mu(f) = \|K * f - g_r\|^2 + \mu\|f\|^2 \qquad [6]$$

where μ, the so-called regularization parameter, is a parameter controlling the trade-off between data fitting and energy of the solution: when μ is small, the data fitting is good and the energy is large while, when μ is large, the data fitting is poor and the energy is small.

For a given μ, the function f_μ, which minimizes the functional of eqn [6] is called the regularized solution of the problem. It is easy to show that its Fourier transform is given by

$$\hat{f}_\mu(\omega) = \frac{\hat{K}^*(\omega)}{|\hat{K}(\omega)|^2 + \mu}\hat{g}_r(\omega) \qquad [7]$$

The regularized solutions form a one-parameter family of functions. When μ is small these functions are strongly contaminated by noise whose effect is gradually reduced when μ increases. For too large values of μ, one re-obtains blurred versions of the original object.

Such behavior is illustrated by the sequence of regularized solutions given in Figure 2 and corresponding to the example of Figure 1. The regularization parameter increases starting from left to right and from top to bottom. Its values are $\mu = 0$ (inverse filter), 0.001, 0.01, 0.05, 0.1, and 0.5 (the noise is of the order of a few percent and the PSF is normalized in such a way that the sum of all its value is 1). The sequence makes evident the well-established theoretical result of the existence of an optimal value of the regularization parameter. Such an optimal value can

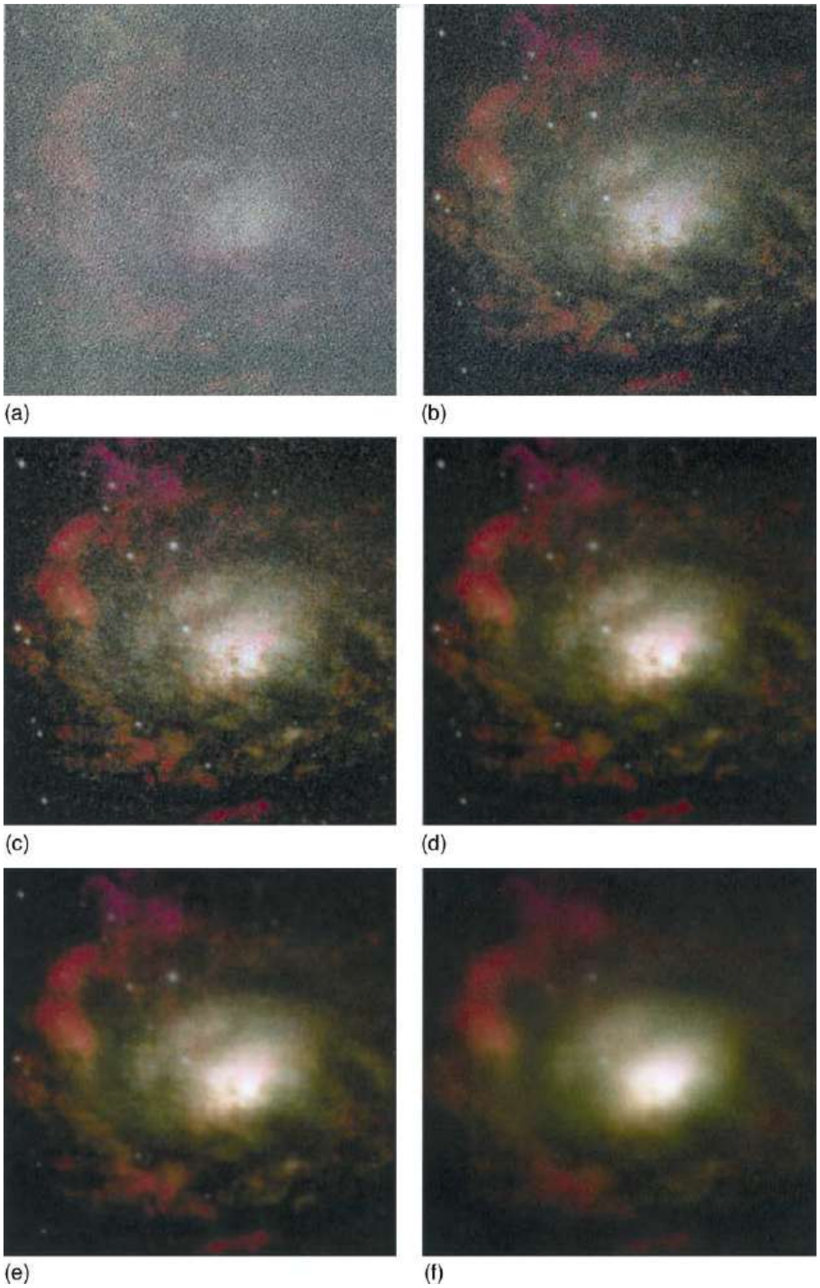

Figure 2 Illustrating the effect of the regularization parameter. The blurred image is that shown in Figure 1c. The sequence is obtained by increasing the value of the regularization parameter: (a) $\mu = 0$ (inverse filter); (b) $\mu = 0.001$; (c) $\mu = 0.01$; (d) $\mu = 0.05$; (e) $\mu = 0.1$; (f) $\mu = 0.5$. It is evident that the best restoration is that shown in (c).

be determined in the case of numerical simulations as that represented in Figure 2. However, its estimation in the case of real images is more difficult. Several methods for the selection of μ have been proposed. We only mention the so-called discrepancy principle. It consists of determining the value of μ, such that the value of the discrepancy functional at f_μ, $\varepsilon^2(f_\mu)$ coincides with an estimate of the energy of the noise. In other words, the data are fitted with an accuracy comparable with their uncertainty.

As shown by eqn [7], the regularized solution has the same band Ω of the imaging system, because $\hat{K}^*(\omega)$ is zero outside Ω. As is known, a band-limited function can be represented in terms of its samples taken at a rate which is roughly proportional to the size of the band. This is the content of the famous Shannon sampling theorem (more precisely its 2D extension). On the other hand, the sampling distance can be taken as a measure of the resolution provided by the restored image so that one concludes that the restoration method discussed above does not produce an improvement of the resolution of the imaging system. It certainly produces an improvement of the quality of the image which makes possible a visual verification of that resolution.

In many circumstances, the term super-resolution is used for describing methods which allow an improvement of resolution beyond the limit provided by the sampling theorem. Therefore, a super-resolving method must produce a restored image with a band broader than that of the imaging system. Therefore, super-resolution is related to the problem of out-of-band extrapolation.

Such an approach was already proposed in the 1960s, when it was observed that the Fourier transform of an object with a finite spatial extent (all objects have this property) is analytic; then the theorem of unique analytic continuation implies that the Fourier transform of the object can be determined everywhere from its values on the band of the imaging system.

Unfortunately this problem is ill-posed, so that out-of-band extrapolation is not feasible in practice. Only recently it has been recognized that a considerable amount of super-resolution is possible when the spatial extent of the object is not much greater than the resolution distance of the instrument. For example, super-resolving methods could be used for detecting unresolved binary star in astronomical images.

The important point is that a super-resolving method must implement explicitly the finite extent of the object: the domain of the object must be estimated and the method must search for a solution which is zero outside this domain. This introduces an important question in object restoration and, more generally, in the theory of inverse problems, namely the design of regularized solutions satisfying additional conditions (constraints). This question is also important from the theoretical point of view: the problem is ill-posed because of insufficient information on the object as transmitted by the imaging system; the use of additional constraints reduces the class of the solutions which are compatible with the data and therefore can improve the restoration. This is the use of a priori information in the solution of inverse problems.

A constraint which has been widely investigated and used is the positivity of the solution, i.e., all the values of the restored image must be positive or zero. The physical meaning of the constraint is obvious. Its beneficial effect is to reduce ringing artifacts which affect the regularized solutions previously discussed, in cases where the object contains bright spots over a black background or sharp intensity variations from zero to some positive value. The ringing is essentially related to the well-known Gibbs effect in the truncation of the Fourier series.

The requirement of positivity is the requirement of a particular lower bound on the values of the solution. Therefore, one can consider more general constraints in requiring a given lower and/or upper bound on the values of the solution, eventually combined with a constraint on the domain to produce, for instance, a super-resolved positive solution. The most general form of these constraints requires that the solution belongs to a given closed and convex set $\mathcal{C}$ in some functional space. A general method producing regularized solutions in a convex set is an iterative method which is essentially a gradient method for the minimization of the least-squares functional, with a projection on the set $\mathcal{C}$ at each iteration. In the case of object deconvolution, the least-squares functional is that given in eqn [4] and the method has the following form: if f_k is the result of the k-th iteration, then f_{k+1} is given by

$$f_{k+1} = P_{\mathcal{C}}[f_k + \tau K^{\mathrm{T}} * (g_r - K * f_k)] \qquad [8]$$

where $P_{\mathcal{C}}$ is the convex projection onto the set $\mathcal{C}$, τ is a relaxation parameter and $K^{\mathrm{T}}(\mathbf{x}) = K(-x)$. In general, the algorithm, known as the projected Landweber method, is initialized with $f_0 = 0$. An important property is that it has a regularization effect, in the sense that iterations must not be pushed to convergence but stopped after a suitable number of iterations to avoid strong noise contamination. In Figure 3 we compare, in a specific example, the result obtained by means of the linear regularization method and that obtained by means of the projected iterative method with a lower and upper bound on the solution. The reduction of ringing is evident.

Among the methods producing positive solutions we must mention the most popular one in astronomy, the so-called Richardson–Lucy method, which is an iterative method for Maximum-Likelihood estimation. If the PSF has been normalized in such a way that the sum of all its values is one, it takes the following form:

$$f_{k+1} = f_k\left(K^{\mathrm{T}} * \frac{g_r}{K * f_k}\right) \qquad [9]$$

The method is normally initialized with a uniform image and it must not be pushed to convergence to avoid noise amplification. Another algorithm suggested by statistical arguments and producing positive solution is the so-called Maximum Entropy Method (MEM).

The image restoration methods described above have wide applications both in microscopy and

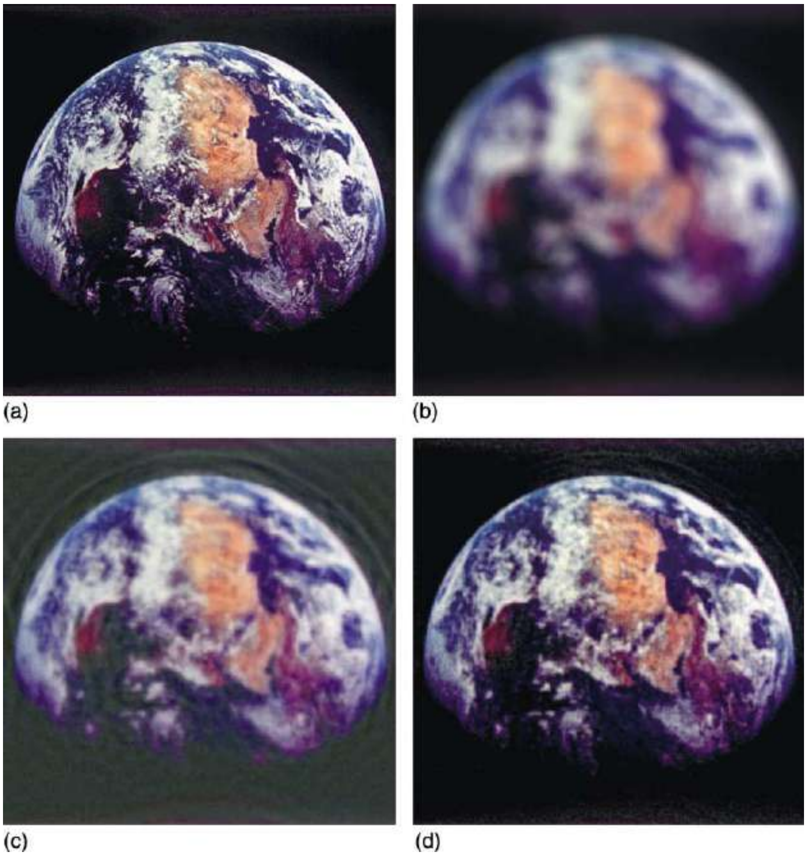

Figure 3 Illustrating the effect of constraints on image restoration: (a) Image of the Earth taken during the Apollo 11 mission; (b) blurred image of (a) obtained by assuming out-of-focus blur and adding noise; (c) optimal restoration provided by the regularization method; the ringing around the boundary of the Earth is evident; (d) optimal restoration obtained by means of the projected Landweber method with a lower and an upper bound on the solution at each iteration. Part (a) courtesy of NASA.

astronomy. For instance, the minimization of the functional eqn [6], with the additional constraint of positivity, has been used for deconvolving images in microscopy to reduce the effect of the missing cone. On the other hand, the Richardson–Lucy method was widely used for deconvolving the images of HST before installation of corrective optics.

Nowadays, image deconvolution is becoming more important for the ground-based telescopes equipped with adaptive optics. Indeed, even if adaptive optics is able to provide a considerable compensation of the atmospheric blur, a further improvement can be obtained by deconvolving the detected images. The PSF is provided by the image of a suitable guide star.

Finally, it is worth mentioning the Large Binocular Telescope, in construction on the Mount Graham in Arizona, because this instrument requires image restoration methods for a full exploitation of its imaging properties. The telescope (Figure 4) consists of two 8 m mirrors on a common mount: both mirrors are equipped with adaptive optics and are combined interferometrically to reach the resolution

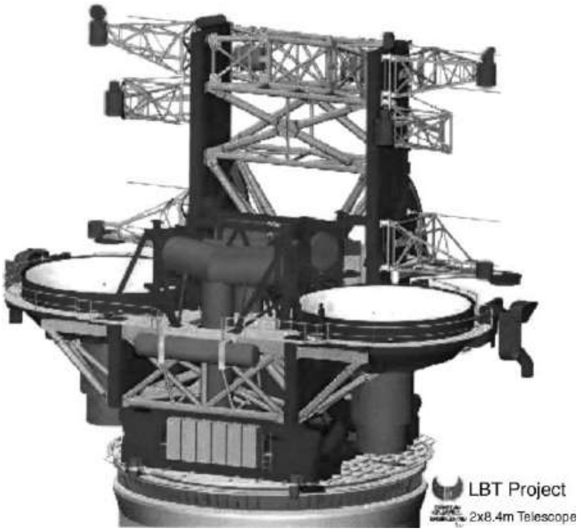

Figure 4 Picture of the Large Binocular Telescope (LBT) in construction on the Mount Graham in Arizona. It consists of two mirrors on the same mount. The images of the two mirrors are combined interferometrically to produce a high-resolution image in the direction of the line joining the centers of the mirrors (baseline).

of a 22.8 m mirror in the direction of the baseline. The telescope can be rotated to record images of the same astronomical target with different orientations of the baseline, i.e., the line joining the centers of the two mirrors. Finally the images must be processed by means of multiple images deconvolution methods to obtain a unique high-resolution image, equivalent to that produced by a 22.8 m mirror.

The applications of inverse problems with major social impact are in the area of medical imaging. The spectacular success of X-ray computerized tomography (CT), invented by GH Hounsfield at the beginning of the 1970s, has stimulated the development of other imaging modalities, based essentially on the same principle, such as positron emission tomography (PET) and single photon emission computerized tomography (SPECT). Computed images are also provided by magnetic resonance imaging (MRI). In such a case, however, the computational problem is rather simple since it consists essentially in Fourier transform inversion.

The basic principle of X-ray CT is described as a finely collimated source S (see Figure 5a) emitting a pencil of X-rays which propagates through the body along a straight line L up to a well collimated receiver R. If we know both the intensity I_0 of the source and the intensity I detected by the receiver, the logarithm of the ratio I_0/I is the integral of the linear attenuation function (roughly proportional to the density function of the body) along the line L. If the source and the receiver are moved along two parallel lines, having direction θ and defining the plane Π (in practice a planar slice of the body under consideration), one obtains the integrals of the linear attenuation function $f(\mathbf{x})$ along all parallel lines orthogonal to θ. The scanning variables in the plane Π are defined in Figure 5b: φ is the angle formed by the unit vector θ with the x-axis and s is the signed distance between the origin and the integration line L. The integrals along all parallel lines orthogonal to θ define a function of s which is called the projection of $f(\mathbf{x})$ in the direction θ:

$$P_\theta(s) = \int_{-\infty}^{+\infty} f(s\cos\varphi - u\sin\varphi, s\sin\varphi + u\cos\varphi)\mathrm{d}u \qquad [10]$$

The set of the projections of $f(\mathbf{x})$ for all possible θ is called the Radon transform of $f(\mathbf{x})$, in honor of the mathematician Johann Radon, who first investigated the problem of recovering a function of two variables from its line integrals. Today, such a problem is known as Radon transform inversion or object restoration from projections. Sampled values of the Radon transform, i.e., sampled values of the projections for a number of possible directions between 0 and π, are the outputs of the acquisition system of the medical equipment known as CT-scanners (Figure 6). Their data-processing system contains the implementation of an algorithm for Radon transform inversion and the final result is a set of computed images providing maps of slices of the human body.

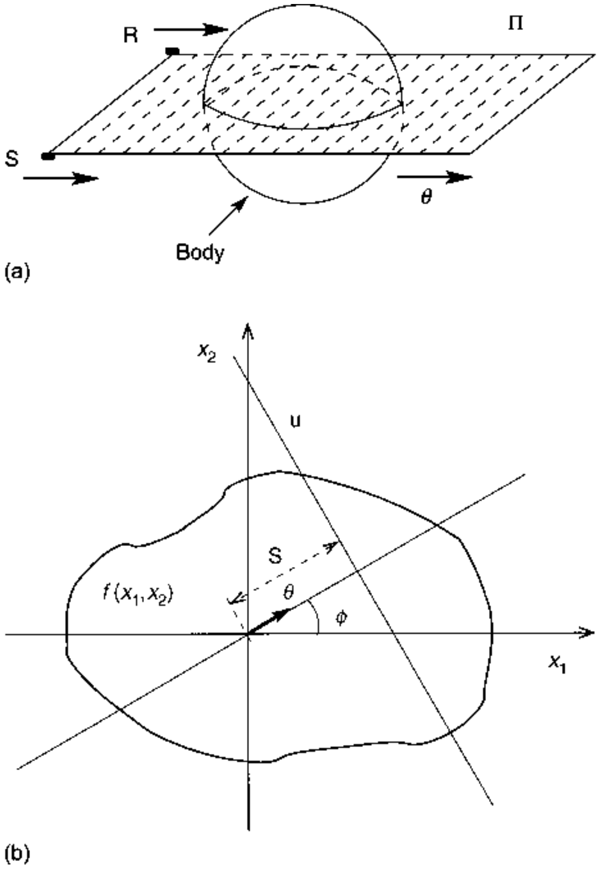

Figure 5 (a) Scheme of the scanning procedure in X-ray tomography: the source(S)−receiver(R) pair is moved along a direction θ orthogonal to the line S−R, defining the plane Π to be imaged; when the scanning has been completed the system is rotated by a certain angle, the scanning procedure is repeated, and so on. (Reproduced with permission from Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging.* Bristol, UK: IOP Publishing.) (b) Definition of the variables used in the representation of the Radon transform; φ is the angle formed by the S−R line with the x_1-axis; s is the signed distance between the origin and the integration line (orthogonal to θ); and u is the coordinate on the integration line.

The plot of the Radon transform of f in the (s, φ)-plane, obtained by representing its values as gray levels, is called the sinogram of f. In Figure 7, we give the image of the sinogram of a brain slice. The horizontal rows provide representations of the projections of the phantom. In the image, the angle φ defining the direction θ takes values between 0 and 2π

while the variable s takes values between $-a$ and a, if a is the radius of a disc containing the phantom. As is evident, the sinogram is symmetric with respect to the point $(0, \pi)$. Each point in the rectangle corresponds to a straight line crossing the phantom and the straight lines through a fixed point of the phantom describe a sinusoidal curve in the (s, φ)-plane. This property has given rise to the name sinogram.

The sinogram is, in a sense, the image of the body slice under investigation, as provided directly by a CT-scanner. However, an inspection of this image does not easily provide information about the body. Therefore, it must be processed to obtain a more significant image. Radon transform inversion is an example of inverse and ill-posed problem and its solution must be treated with extra care.

The crucial point in Radon transform inversion is the Fourier slice theorem, whose content is the following: the Fourier transform of the projection in the direction θ of the function f, is the Fourier transform of f on the straight line passing through the origin and having direction θ:

$$\hat{P}_\theta(\omega) = \hat{f}(\omega\theta) \qquad [11]$$

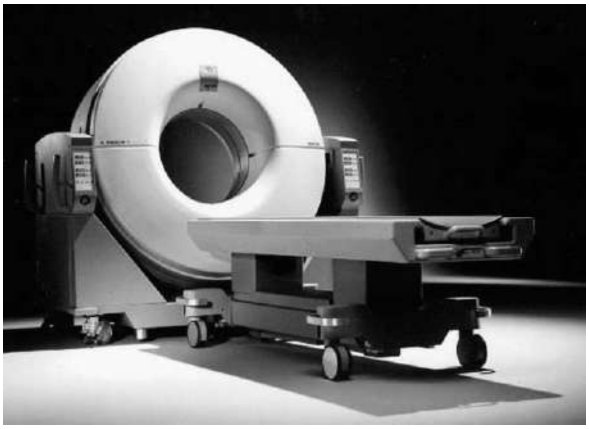

Figure 6 Picture of a scanner for X-ray tomography. The annular part contains the acquisition and scanning systems.

This theorem clarifies the information content of CT data: each projection provides the Fourier transform of the function f along a well-defined straight line in the plane of the spatial frequencies. It is also the starting point for deriving the basic algorithm of data inversion in tomography, the filtered back-projection (FBP). It is obtained from Fourier transform inversion in polar coordinates and is a two-step algorithm: the first step is a filtering of the projections by means of a ramp-filter; the second is the back-projection of the filtered sinogram. More precisely the two steps are as follows:

- *Filtering*: for each θ the projection $P_\theta(s)$ is replaced by a filtered projection given by

  $$Q_\theta(s) = \int_{-\infty}^{+\infty} |\omega| \hat{P}_\theta(\omega) e^{i\omega s}\, d\omega \qquad [12]$$

  where the multiplication by the ramp filter $|\omega|$ derives from the Jacobian of polar coordinates;
- *Back-projection*: this operation is the dual of the projection operation: indeed the projection assigns to a straight line L with coordinates (s, φ), hence to a point of the domain of definition of the Radon transform, the integral of f along L; on the other hand, the back-projection assigns the value of the Radon transform at (s, φ) to all points of the straight line L with the same coordinates (a pictorial representation of projection and back-projection is given in Figure 8); as a consequence the back-projection operator $R^\#$ assigns to each

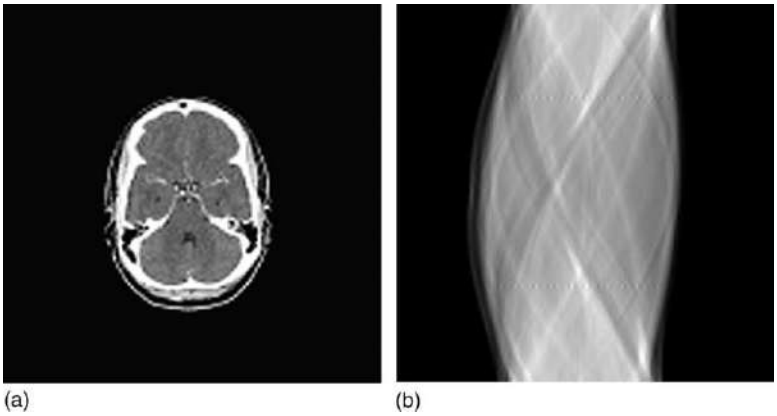

Figure 7 (a) Picture of a brain slice; (b) the corresponding sinogram. As explained in the text, the sinogram is a gray-level representation of the Radon transform in the s, φ plane. Each horizontal line in (b) corresponds to a projection of (a).

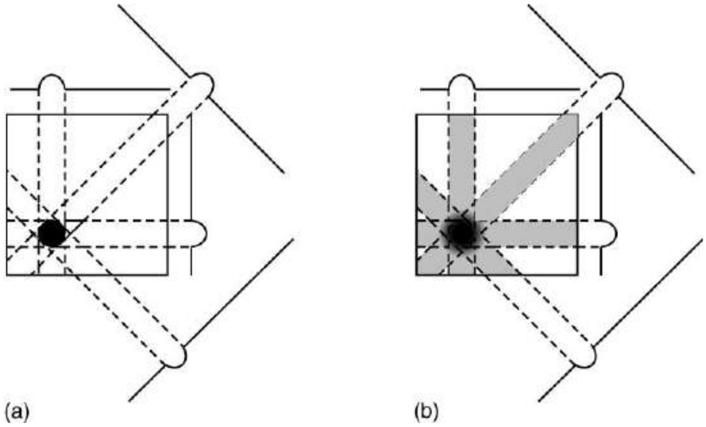

Figure 8 (a) The effect of the projection operation is shown by drawing the projections of a circular phantom corresponding to three directions. (b) shows the result obtained by applying the back-projection operation to the three projections given in (a). It is evident that the picture provided by this operation has a maximum on the domain of the circular phantom.

point **x** the sum (integral) of all values of the Radon transform corresponding to straight lines passing through **x**:

$$R^{\#}P_\theta(\mathbf{x}) = \int_0^{2\pi} P_\theta(x\cos\varphi + y\sin\varphi)\mathrm{d}\varphi \qquad [13]$$

The back-projection operator, when applied directly to the projections as in eqn [13], provides a blurred image of f; in order to get a satisfactory restoration it must be applied to the filtered projections $Q_\theta(s)$. This point is illustrated in Figure 9.

We point out that the ramp filter introduced in eqn [12] amplifies the high frequency components of the projections hence the noise corrupting these components. This remark clarifies that Radon transform inversion is an ill-posed problem and requires some kind of regularization. This is obtained by introducing an additional low-pass filter which reduces the effect of the high frequencies. The most frequently used combination of ramp and low-pass filter is the so-called Shepp–Logan filter.

X-ray tomography is also called transmission computerized tomography (TCT) because the image is obtained by detecting the X-rays transmitted by the body. It provides information about anatomical details of human organs because the map of the linear attenuation function is essentially the map of the density of the tissues.

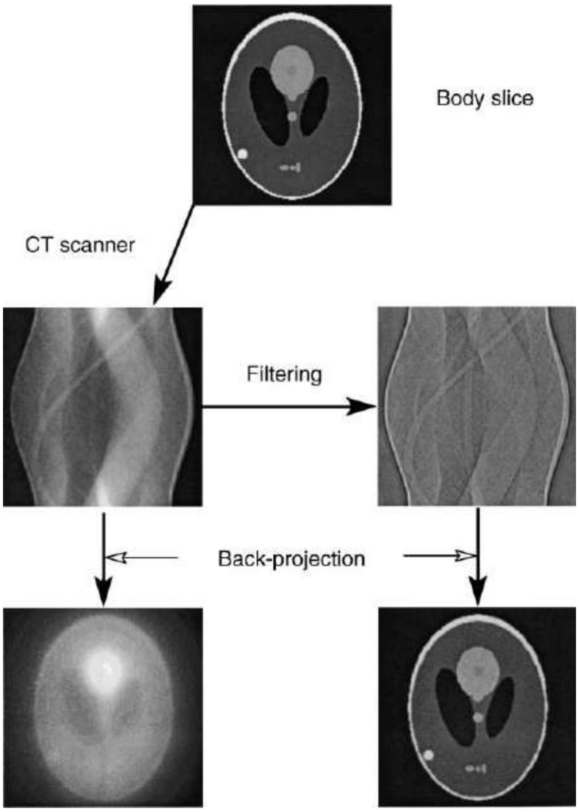

Figure 9 Illustrating the effect of the filtered back-projection. Shown here is: a slice of the so-called Shepp–Logan phantom; the corresponding sinogram (indicated by the arrow CT scanner); the filtered sinogram (indicated by the arrow filtering; the restorations obtained by applying the back-projection operation both to the original sinogram and to the filtered sinogram.

A different type of information is obtained by the so-called emission computerized tomography (ECT) which is based on the administration of radionuclide-labeled agents known as radio-pharmaceuticals. Their distribution in the body of the patient depends on factors such as blood flow, metabolic processes, etc. Then a map of this distribution is obtained by detecting the γ-rays produced by the decay of the radio-nuclides. Therefore, ECT yields functional information, in the sense that the images produced by ECT show the function of the tissues of the organs.

Two different modalities of ECT are usually considered:

- single photon emission computerized tomography (SPECT), which makes use of radio-isotopes where a single γ-ray is emitted per nuclear disintegration;
- positron emission tomography (PET), which makes use of β^+-emitters, where the final result of a nuclear disintegration is a pair of γ-rays, propagating in opposite directions, produced by the annihilation of the emitted positron in the tissues.

In both cases it is necessary to detect the γ-rays coming from well-defined regions of the body and, to this purpose, different methods are used in each case.

In SPECT, the discrimination of the γ-rays is obtained by means of a collimator, consisting of holes in a large lead slab covering the crystal detector face. In PET the collimation is obtained by means of pairs of detectors in coincidence. This technique, which is also called electronic collimation, is more accurate than the physical collimation used in SPECT and allows the design of systems with a great efficiency.

The basic reconstruction method in TCT, namely FBP, is often used in the reconstruction of SPECT and PET data. However, several corrections are necessary in practice. The principal ones are due to the collimator blur and to the scattering of photons in the body. If one develops a more accurate model of data acquisition by taking into account these effects, then the restoration problem implies the inversion of a very large matrix which is sparse and ill-conditioned. To this purpose iterative regularization methods are used such that the basic operation required at each step is matrix-vector multiplication. Other imaging modalities, which have been proposed for medical applications, are electrical impedance and microwave tomography. The underlying inverse problems are basically nonlinear so that the computational cost of the methods for their solution is, in general, too high for clinical applications. However the research in these areas is very active and, therefore, the situation could be improved in the near future.

See also

Tomography: Tomography and Optical Imaging.

Further Reading

Bertero M (1989) Linear inverse and ill-posed problems. In: Hawkes PW (ed.) *Advances in Electronics and Electron Physics*. vol. 75, pp. 1–120. New York: Academic Press.

Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging*. Bristol, UK: IOP Publishing.

Bertero M and De Mol C (1996) Super-resolution by data inversion. In: Wolf E (ed.) *Progress in Optics*. vol. XXXVI, pp. 129–178. Amsterdam, Elsevier.

Engl HW, Hanke M and Neubauer A (1996) *Regularization of Inverse Problems*. Dordrecht, Germany: Kluwer.

Groetsch CW (1993) *Inverse Problems in Mathematical Sciences*. Braunsschweig, Germany: Vieweg.

Herman GT (ed.) (1979) *Image Reconstruction from Projections*. Berlin: Springer.

Herman GT (1980) *Image Reconstruction from Projections. The Fundamentals of Computerized Tomography*. New York: Academic Press.

Huang TS (ed.) (1975) *Picture Processing and Digital Filtering*. Berlin: Springer.

Kak AC and Slaney M (1988) *Principles of Computerized Tomographic Imaging*. New York: IEEE Press.

Krestel E (ed.) (1990) *Imaging Systems for Medical Diagnostics*. Berlin: Siemens Aktiengesellschaft.

Natterer F (1986) *The Mathematics of Computerized Tomography*. New York: Wiley.

Tikhonov AN and Arsenin VY (1977) *Solutions of Ill-Posed Problems*. Washington, DC: Winston/Wiley.

Adaptive Optics

C Pernechele, Astronomical Observatory of Padova, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The image quality of a perfect optical system is ideally limited by its diffraction figure. In the real world there are many effects which degrade that physical limit, resulting in poorer image quality at the focal plane. These effects may be static (e.g., time-independent, as, for example, error in the alignment of the optical train) or dynamic (e.g., variable with time, such as optical misalignment introduced by thermo-mechanical distortion). Active and adaptive optics, the topics of this article, are technologies able to correct for the optical aberrations induced by dynamical effects.

There are a variety of causes liable for dynamical image degradation, among these time-variable optical misalignments, pointing and tracking errors (for motorized instruments), and atmospheric perturbations. In practice, both atmospheric turbulence and misaligned instrument introduce random spatial and temporal perturbations on the optical beam, distorting the spherical (ideal) wavefront incoming on to the focal plane. This leads to loss of both the diffraction limited condition and image resolution. A typical situation for astronomical application is shown in Figure 1. A target (a star in this case) emits light whose wavefront, being far from the observing system, can be considered plane. When this wavefront passes through the atmosphere it is distorted with a temporal frequency of the order of hundreds of Hertz. This dynamically distorted wavefront fills the aperture of the optical system and will once more be distorted by the dynamic misalignment of the optical system itself. The final wavefront will be deformed with respect to the ideal (spherical) wavefront.

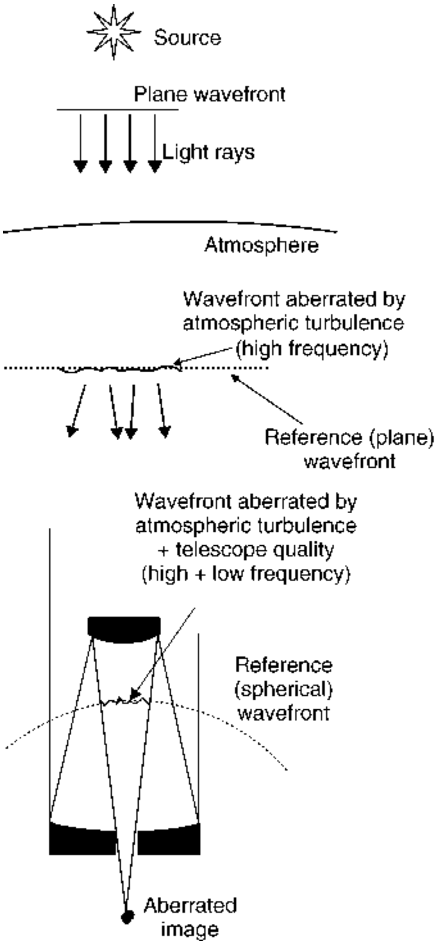
Source
Plane wavefront
Light rays
Atmosphere
Wavefront aberrated by atmospheric turbulence (high frequency)
Reference (plane) wavefront
Wavefront aberrated by atmospheric turbulence + telescope quality (high + low frequency)
Reference (spherical) wavefront
Aberrated image

Figure 1 Dynamical wavefront aberration.

Dynamic optical misalignment may depend on gravitational and thermal deformation, while pointing error may appear as very slow (drift), slow (wander), or fast (jitter) tilt of the image on the focal plane. Atmospheric perturbations are produced by air turbulence and may be natural (as in the astronomical or satellite tracking cases) or induced by the experiment itself (thermal blooming in high power laser applications). All these effects appear on very different time-scales, as shown in Figure 2. As noticeable, atmospheric turbulence introduces aberrations of all orders, at a frequency above about 1 Hz. Pointing and tracking errors introduce an image tilt with different time-scale. Gravitational effects induce mechanical deformation of the optics, mainly in astronomical applications because of the massive components (the telescope primary mirror in particular). Thermal deformation produces optical misalignment among optical components mainly resulting in defocus. Figure 2 is purely illustrative, because some effects may produce wavefront aberrations at different frequencies than indicated there. For example, drift may be faster than indicated in the figure, and thermal blooming (a thermal effect) introduces aberrations at a higher frequency, and so on.

The correction of the wavefront errors is usually made by means of devices able to change the phase wave somewhere in the light path. Though the concept for correcting aberrated wavefront at slow or high frequency is the same, for example, the use of one or more deformable elements in the optical path, the technologies developed for each area are quite

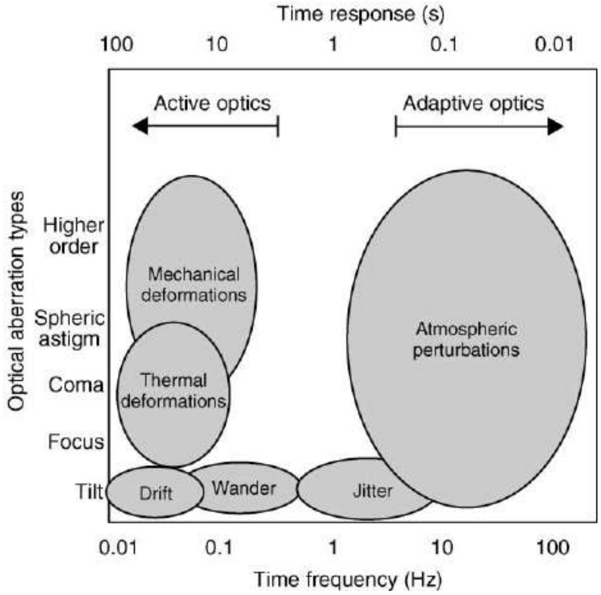

Figure 2 Wavefront aberration time-scale and associated aberrations.

different: variations at slow frequency (approximately below 0.1 Hz in Figure 2) are indicated as quasi-static errors and the techniques developed to overcome them are properly called active optics, while adaptive optics indicates the techniques to overcome the unwanted effects at high frequency (above about 1 Hz in frequency). The frequency value dividing the active from adaptive optics is not well defined and depends on the field of application and its scientific engineering communities. The important point is that a correct use of a combination of the two techniques may ultimately lead to bringing the imaging system close to its diffraction limit.

Slow Corrections: Active Optics

The analysis of the wavefront is required before its correction, and one or more deformable/movable mirrors and their control system, to operate the corrections. A typical system for low frequency correction is shown in Figure 3. A telescope contains three mirrors (M1, M2, and M3) and an instrument to detect its scientific target, usually placed on its optical axis. The dynamic deformations present in this kind of instrument are mainly third-order spherical (thermal deformation of M1), third-order astigmatism (a saddle-horse deformation of M1 due to gravity when the telescope is pointing away from the zenith), and third-order coma (due to gravitational decentering between M1 and M2). Since a telescope has a tracking system some pointing error may be present, causing motion of the image in the focal plane. The observation of an off-axis target (typically a star for astronomical applications) is made to measure the deformations of the incoming wavefront by means of a wavefront sensor (WS) system. This step is completed using some of the wavefront sensing technique described elsewhere in this encyclopedia. The WS must measure the information on the wavefront distortion with adequate spatial resolution to compensate for aberrations across the full aperture of the system. Once the wavefront distortion has been measured by the WS, the needed corrections are sent to the control system which apply the appropriate mirror deformation/movements. In the example of Figure 3 only the primary mirror (entrance pupil) is deformable and it is used for correcting spherical and astigmatism aberrations. M2 is used to compensate the decentering coma by means of lateral displacement. M3 may be tilted to correct for low-frequency pointing errors. In astronomical applications active optics is essential for telescopes with primary-mirror diameters larger than approximately four meters which cannot be manufactured rigidly at reasonable cost. Active optics requires a thin deformable or a segmented primary mirror maintained on the correct shape at every elevation. Using this method, low-frequency correction has been achieved to date in telescopes with primary mirrors of the order of ten meters.

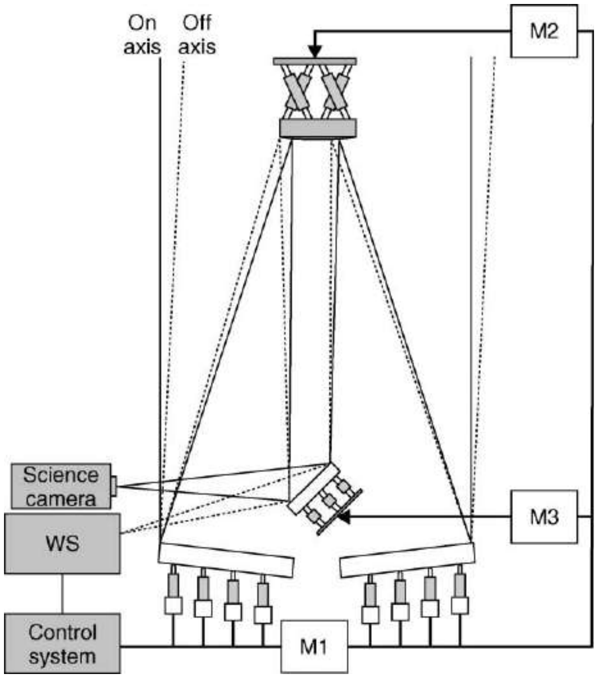

Figure 3 A typical system for correcting low-frequency dynamical aberrations (active optics).

Fast Corrections: Adaptive Optics

It is clear that corrections at high frequency are impossible using the approach discussed above, because of the large dimensions and weight of the mirror and the low response time required. For an adaptive correction it is necessary to image a smaller system pupil in the optical train and make the active correction of the wavefront. A typical configuration is shown in Figure 4. The light from the telescope focal plane (corrected for low frequency errors by means of active optics) is collimated and a pupil is located in this beam. Behind the active mirror a beamsplitter is used to send part of the light into the WS and measure wavefront distortions. The analysis results are sent to the control system which performs the active mirror deformation for the wavefront compensation. If the WS and the control system are sufficiently fast (typically faster than the time-variations of the wavefront distortion), this system is able to produce a plane wavefront which can be focused at the diffraction limit of the system. For this reason

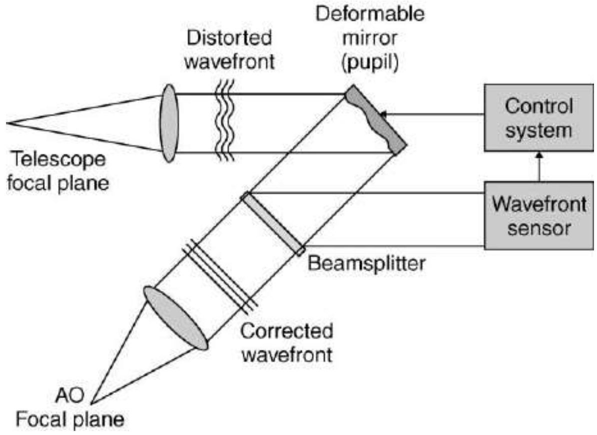
Deformable mirror (pupil)
Distorted wavefront
Control system
Telescope focal plane
Wavefront sensor
Beamsplitter
Corrected wavefront
AO Focal plane

Figure 4 A typical setup for correcting high-frequency dynamical aberrations (adaptive optics).

adaptive optics are often referred to as real-time system.

The seminal paper on the adaptive optics concept was written by H.W. Babcock of the Mount Wilson and Palomar observatories in 1953. Since then a great improvement has been made, especially in military and astronomical fields. In recent years, however, many applications of adaptive optics have developed in fields different from traditional ones, such as precision manufacturing by means of high-power lasers, telecommunications, aero-optics, medical physics, and ophthalmology.

We describe adaptive optics with particular reference to astronomical applications. We provide a brief description of other applications at the end of this article.

Imaging Through the Atmosphere

The optical effects of the atmospheric turbulence are due to local temperature variations that induce fluctuations in the refractive index of air. The refractive index changes depend on the air density as well as the range of the temperature variations. Air density is highest at sea level and decreases exponentially with altitude. Optical effects of turbulence decrease with the altitude of the site, the reason for locating astronomical telescopes on mountains. The effects of turbulence on a point source image are depicted in Figure 5. Hereafter we consider an ideal telescope, for example, one without need for active correction at low frequencies. We consider the source 1 (continous line). The radiation emitted by the source is a spherical wavefront that can be treated as a plane, because of the very long distance to the Earth's atmosphere.

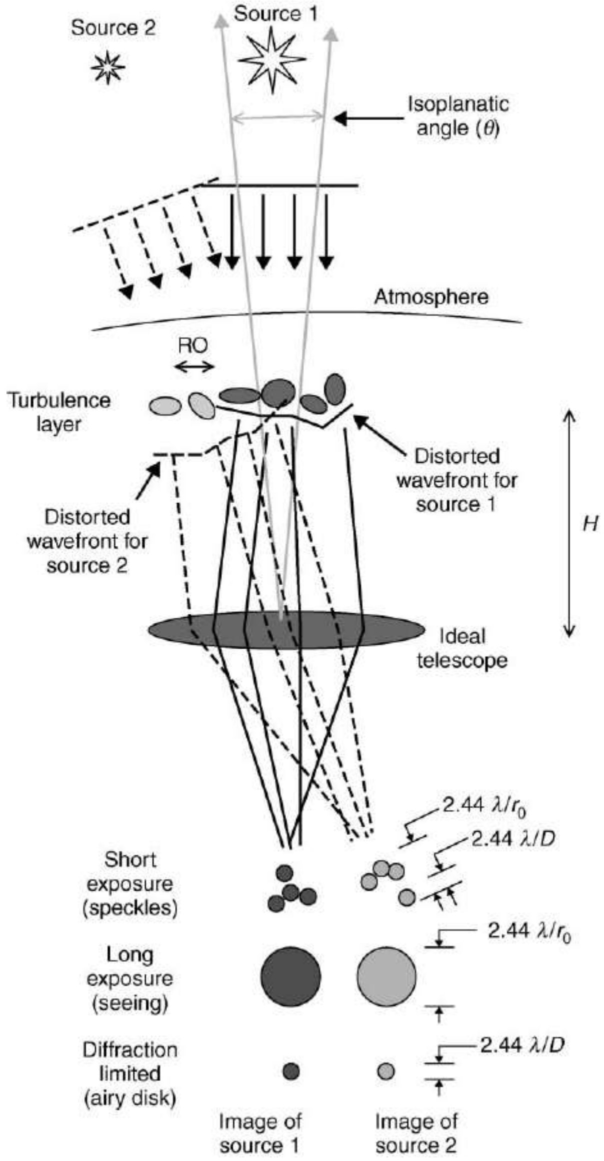
Source 2
Source 1
Isoplanatic angle (θ)
Atmosphere
RO
Turbulence layer
Distorted wavefront for source 1
H
Distorted wavefront for source 2
Ideal telescope
$2.44\ \lambda/r_0$
$2.44\ \lambda/D$
Short exposure (speckles)
$2.44\ \lambda/r_0$
Long exposure (seeing)
$2.44\ \lambda/D$
Diffraction limited (airy disk)
Image of source 1
Image of source 2

Figure 5 Imaging through the atmosphere.

In absence of wavefront distortion, the angular diameter of the image (Airy disk), is approximately 2.44 λ/D, with λ the wavelength of the observation and D the telescope aperture (the mirror diameter). In this case the image is said to be diffraction limited, and it is the best possible performance for the observation. This is the performance of a space-based ideal telescope.

In the presence of the atmosphere the incoming wavefront is distorted by the turbulence and when it reaches the telescope aperture it is no longer a plane (see in Figure 5). The so-called Fried's parameter r_0 (also called atmospheric coherence length) is the most common parameter used to quantify the distortion effects of atmospheric turbulence. It is defined as the distance over which the wavefront phase is highly correlated (in this way it provides a measure of the turbulence scale size). For an undistorted plane wave, r_0 is infinity, while for atmospheric distortion it typically varies between 5 cm (bad sites) and 20 cm (excellent sites). We depict Fried's parameter schematically in Figure 5, as bubbles of dimension r_0 in the turbulent layer (referred to as Fried's cells). r_0 is a statistical parameter, which depends on time and wavelength ($r_0 = r_0(t, \lambda^{6/5})$). Its variability with time may be considerable, sometimes changing by a factor of 2 within a few seconds. A small telescope, with a diameter less than r_0, will always operate at this diffraction limit. When the r_0 is smaller than the telescope aperture, the angular size of the image is different if we consider long- or short-time exposures. If the exposure time is very short (less than hundredths of a second) it is possible to image a structure in the focal plane (see short exposure image in Figure 5). This structure, which varies with time, is due to high-order wavefront distortions that occur within the aperture. The short exposure image is broken into a large number of speckles of diameter 2.44 λ/D, which are randomly distributed over a circular region of approximately 2.44 λ/r_0 in angular diameter. A lower-order wavefront distortion is correlated with the size of the largest disturbances produced by atmospheric turbulence (the so-called outer scale) which varies considerably (from 1 m to over 1 km). Such disturbances produce a time-dependent tilt of the overall wavefront that causes fluctuations in the angle of arrival of the light from a star and randomly displace the image. For D of the order of the Fried parameter r_0, image motion dominates, and a simple tip-tilt mirror (not just deformable) provides useful correction. For D greater than r_0, speckles dominate and a deformable mirror is necessary to account for the wavefront distortion. For long exposures the dimension of the image is determined by the ratio λ/r_0 rather than λ/D. The angular resolution is diminished due to the atmospheric turbulence (see Figure 5). For a 4 m telescope, atmospheric distortion degrades the spatial resolution by more than one order of magnitude compared to the diffraction limit, and the intensity at the center of the star image is lowered by a factor of 100 or more. Even at the best astronomical sites, ground-based telescopes observing at visible wavelengths cannot, because of atmospheric turbulence alone, achieve an angular resolution better than telescopes of about 20 cm diameter.

Most of the distortions imposed by the atmosphere are in the wave-phase and an initially plane wavefront traveling 20 km through the turbulent atmosphere accumulates, across the diameter of a large telescope, phase errors of a few micrometers.

Turbulence is distributed along the propagation path through the atmosphere, and then the wavefront errors become uncorrelated as the field angle increases. This effect is known as angular isoplanatism and is one of the major limitations to the astronomical adaptive optics. The structure becomes completely uncorrelated out of the isoplanatic angle, approximately equal to r_0/H, where H is the altitude of the turbulence layer. This decorrelation is sketched in Figure 5, where the short exposure images for objects 1 and 2 have different speckles structure.

In Earth's atmosphere significant turbulence occurs at altitudes up to 20 km, often with large peaks in the vicinity of the tropopause, at around 10 km there exist, in practice, more than one parameter H in Figure 5, one for each turbulence layer. This three-dimensional distribution of the turbulence considerably restricts the angle over which adaptive compensation is effective and leads to the design of multiconjugate adaptive optics, as we next explain.

Adaptive System Design

As stated above, an active system is composed of at least one wavefront sensor and one deformable mirror. Typically, the wavefront is corrected by making independent tip, tilt, and piston movements of the deformable mirror surface, as shown in Figure 6. As a rule of thumb, a deformable mirror used for atmospheric correction should have one actuator for each Fried's cell, such that the total number of actuators should be equal to the number of Fried's cells filling the telescope aperture, i.e., D^2/r_0^2. For instance, a near-perfect correction for an observation in visible light (0.6 μm) with an 8 m telescope placed in a good astronomical site ($r_0 = 20$ cm), would require a deformable mirror with approximately 6400 actuators in its adaptive optics system. Recalling that r_0 is strongly dependent on the wavelength (i.e., proportional to $\lambda^{6/5}$), similar performance in the infrared at 2 μm requires only 250 actuators. This is one reason why adaptive correction is easier in the infrared. A large number

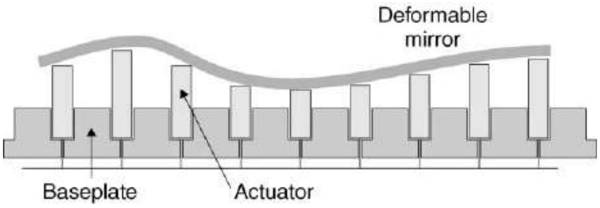

Figure 6 A sketch of a deformable mirror.

of Fried's cells requires a similarly large number of subapertures in the wavefront sensor. Commonly used WSs are based on modifications of the classic Hartmann screen test or of curvature type. In the former case the adaptive device is made with individual piezoelectric actuators (see Figure 6), while in the latter the correction is achieved with a bimorph adaptive mirror, made of two bonded piezoelectric plates. With both methods, wavefront sensing is done on a reference off-axis object (Figure 3), or even on the observed object itself (Figure 4) if it is bright enough. In the former case the angular distance of the reference object must fall inside the isoplanatic angle and must be sufficiently luminous. In typical astronomical sites the isoplanatic angle at a wavelength of 0.6 μm is ~5 arcsec, while it can be ~20 arcsec at 2 μm. This is an additional reason why the adaptive optics system is most efficient in the infrared range. Correction in the visible requires a reference star approximately 25 times brighter than in the infrared. Most current astronomical systems are designed to provide diffraction-limited images in the near-infrared (1 to 2 μm) with the capability for partial correction in the visible. However, some military systems for satellite observations do provide full correction in the visible on at least 1 m class telescopes.

The control system is generally a specialized computer that calculates, from the WS measurements, the commands to send the actuators of the deformable mirror. The calculation must be performed quickly (within 0.5 to 1 ms), otherwise the state of the atmosphere may have changed making the wavefront correction inaccurate. The required computing power can exceed several hundred million operations for each command set sent to a 250 actuator deformable mirror. As in active optics systems, zonal or modal control methods are used. In zonal control, each zone or segment of the mirror is controlled independently by wavefront signals that are measured for the subaperture of that zone. In modal control, the wavefront is expressed as the best-fit linear combination of polynomial modes of the wavefront distorted by the atmospheric perturbations.

Future Development

As explained above, a convenient method for measuring the wavefront distortion is to look for an off-axis reference star (also known as natural guide star, NGS) within the isoplanatic angle. However, it is generally impossible to find a bright reference star close to an arbitrary astronomical target. Conditions are far better in the infrared than in the visible because atmospheric turbulence (and in particular its high spatial frequencies) is, for a given AO correction, less pronounced at longer wavelengths. The spatial and temporal sampling of the disturbed wavefront can be reduced, which in turn can permit the use of fainter reference stars. Coupled with the larger isoplanatic angle in the infrared, this gives a much better outlook for AO correction than in the visible.

Nevertheless, even for observations at 2.2 μm, the sky coverage achievable by this technique (equal to the probability of finding a suitable reference star in the isoplanatic patch around the chosen target) is of the order of 0.5 to 1%. It is therefore unsurprising that most scientific applications of AO have to date been limited to the study of objects which provide their own reference object. The most promising way to overcome the isoplanatic angle limitation is by the use of artificial reference stars, also referred to as laser guide stars (LGS). These are patches of light created by the back-scattering of pulsed laser light of sodium atoms in the high mesosphere or by molecules and particles located in the low stratosphere. The laser beam is focused at an altitude of about 90 km in the first case (sodium resonance) and 10 to 20 km in the second case (Rayleigh diffusion). Such an artificial reference star can be created as close to the astronomical target as desired, and a wavefront sensor measuring the scattered laser light is used to correct the wavefront aberrations on the target object.

Military laboratories have reported the successful operation of adaptive optics devices at visible wavelengths using laser guide star on both a 60 cm telescope and a 1.5 meter telescope, achieving an image resolution of 0.15 arcsec. Nevertheless, there are still physical limitations of a LGS. The first problem, focus anisoplanatism, or cone-effect, was realized long ago. Since the artificial star is created at a relatively low altitude, back-scattered light collected by the telescope forms a conical beam, which does not cross the same turbulence-layer areas as light from the astronomical target. This leads to a phase estimation error, which in principle can be solved by the simultaneous use of several laser guide stars surrounding the targeted object. A more significant problem is image motion or tilt determination. Since the paths of the outgoing and returning light rays are the same, the centroid of the artificial light appears to be stationary in the sky, while the apparent position of an astronomical source suffers lateral motions (also known as tip/tilt). The simplest solution is to supplement the AO system and LGS with a tip/tilt corrector set on a (generally) faint close NGS. Performance is then limited by poor photon statistics in the tip/tilt correction.

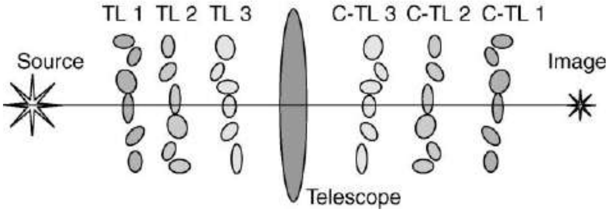

Figure 7 Conjugated planes of turbulent layers.

Adaptive optics with a multicolor laser probe is another concept that has been studied to solve the tilt determination problem of laser beacon based AO. Only applicable to sodium resonant scattering at 90 km, it excites different states of the sodium atoms and makes use of the slight variation in the refraction index of air with wavelength. Its main drawback is the limited returned flux, due to the saturation of mesospheric sodium layer.

The quality of a site for adaptive optics is determined not only by the median seeing, but also by the height and thickness of the turbulent layers (often only a few thin layers dominate). Some adaptive-optics systems can be adapted to correct for turbulence originating at a particular height by using several reference stars. This technique is known as multiconjugate adaptive optics (MCAO) and is still in his infancy. This concept deals with a 3-D view of the atmosphere, and for this reason it is also called turbulence tomography. Several wavefront sensors are needed for tomography. Turbulence tomography is also useful to correct for the cone effect when using LGS. When several deformable mirrors are placed on the appropriate conjugate planes, as shown in Figure 7, anisoplanatism reduces and the size of the corrected field of view increases. The situation is roughly depicted in Figure 7, where each turbulent layer (indicated as TL) has a proper conjugated plane (C-TL).

Another challenging optical solution is one of utilizing large secondary adaptive mirrors. This application seems especially interesting at thermal wavelengths, where any additional mirror raises the significant thermal background detected by instruments.

Other Adaptive Optics Applications

In recent years the adaptive optics technique has been developed in many applications beside astronomy and the military, for example, as in high-power laser welding or cutting, in ophthalmology and in aero-optics.

In high-power laser applications, thermal blooming takes place when laser light is absorbed by the gases and particles that make up the atmosphere. Since energy is conserved, the energy lost by the laser (its intensity diminishes) is gained by the atmosphere that heats up. This inhomogeneity at the boundary of the laser beam generates wavefront distortions by turbulence, as shown in Figure 8. This means that laser light that passes through the air at a later time interacts with a different atmosphere than the light that passed by earlier. The net effect is that, in the case of a Gaussian laser beam, light following a wavefront which passed at an earlier time, 'sees' a concave lens that deflects the successive rays out of the beam (away from the axis) causing the beam to appear to grow in radius or 'bloom'. Adaptive optics techniques for real-time compensation of blooming in laser beams were developed in the late 1960s. Given the high coherence and monochromaticity of a laser, the technique is more straightforward than for more common white-light applications of the astronomical applications.

Density fluctuations in a compressible shear layer are sufficiently high to cause a time varying index of refraction of the layer. This situation is common in a turbulent flowfield, and the study of the wavefront aberrations induced by this flowfield, when it is of relatively short propagation length in the near-field of the optical system, is referred to as aero-optics. Turbulent flowfield induces various types of degradations in the performances of an optical system. For example the jitter and defocusing of laser-based weapons cause a reduction in the amount of energy delivered to the target. Tracking systems experience bore-site errors and imaging systems experience blurring. Recent work suggests that adaptive optics compensation may successfully overcome these problems.

In ophthalmology the adaptive technique finds application in studying the retina of the eye. In this case the eye itself produces wavefront aberration, and the retina is seen as blurred. Techniques in ophthalmic laser surgery are being developed to use adaptive optics to correct for detected distortions.

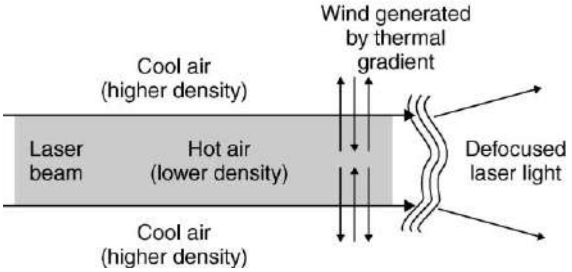
Wind generated by thermal gradient
Cool air (higher density)
Laser beam
Hot air (lower density)
Defocused laser light
Cool air (higher density)

Figure 8 Thermal blooming effect in high-power laser applications.

In the field of telecommunications, studies are in progress for using adaptive techniques in data transmission between ground-based stations and satellites.

See also

Environmental Measurements: Doppler Lidar; Optical Transmission and Scatter of the Atmosphere. **Instrumentation:** Astronomical Instrumentation; Telescopes. **Phase Control:** Wavefront Coding; Wavefront Sensors and Control (Imaging Through Turbulence).

Further Reading

Beckers JM (1993) Adaptive optics for astronomy: principles, performance and application. *ARAA* 31: 13–62.

Fried DL (1965) Statistics of a geometric representation of wavefront distortion. *JOSA* 55: 1427–1435.

Hardy JW (1998) *Adaptive Optics for Astronomical Telescopes*. New York: Oxford University Press.

Jumper EJ and Fitzgerald EJ (2001) Recent advances in aero-optics. *Progress in Aerospace Sciences* 37: 299–339.

Liang J, Williams DR and Miller DT (1997) Supernormal vision and high-resolution retinal imaging through adaptive optics. *JOSA A* 14(11): 2884–2892.

Lukin VP (1995) *Atmospheric Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.

Ragazzoni R, Marchetti E and Valente G (2000) Adaptive-optics corrections available for the whole sky. *Nature* 403(6765): 54–56.

Roddier F (1999) *Adaptive Optics in Astronomy*. Cambridge: Cambridge University Press.

Roggeman M and Welsh B (1996) *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.

Tyson RK (2000) *Introduction to Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.

Welford W (1974) *Aberrations of the Symmetrical Optical System*. New York: Academic Press.

Wilson RW and Jenkins CR (1996) Adaptive optics for astronomy: theoretical performance and limitations. *MNRAS* 268: 39.

Hyperspectral Imaging

M L Huebschman, R A Schultz and H R Garner,
University of Texas, Dallas, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hyperspectral imaging techniques on a microscopic level have paralleled the hyperspectral remote astronomical sensing which has developed over the years. It is especially apparent in remote sensing of our own planet's land masses and atmosphere, such as in the LANDSAT satellite program. In biology, biochemistry, and medicine, hyperspectral imaging and multispectral imaging are often interchanged, although there is a distinct difference in the two techniques.

A multispectral imager produces images of each of a number of discrete but relatively wide spectral bands. Multispectral imagers employ several wide bandpass filters to collect the data for the images. A hyperspectral imager produces images over a contiguous range of spectral bands and, thus, offers the potential for spectroscopic analysis of data. Hyperspectral imagers may use gratings or interferometers to produce the contiguous range of spectral bands to record.

Fluorescent dyes (also referred to as fluorophores or fluorochromes) and recently quantum dots, as markers, are a very powerful tool for analyzing organic tissues for researchers in cytology (study of cell function), cytogenetics (study of the human chromosomes), histology (study of structure and function of tissues), microarray scanning (analysis of concentrations of multiple varied probes), and biological material science. Without the correlation of the spectral and spatial data in biological images, there would be little useful information for the researcher. Therefore, hyperspectral imaging in the biological sciences must combine imaging, spectroscopy, and fluorescent probe techniques. Such imaging requires high spatial resolution to present a true physical picture, and the spectroscopy must also have high spectral resolution in order to determine the quantities of biochemical probes present. The correlation must provide the color-position relation for identification of meaningful information.

This description will begin with an initial review of multispectral imaging to lead into a discussion of hyperspectral imaging. The discussion will examine aspects of hyperspectral imaging based on an interferometric method and on a grating dispersion method similar to the methods in multispectral imaging.

Image Parameters

The analysis of fluorophores involves the excitation of the dye with a shorter wavelength and thus higher energy than the emission spectrum. The use of the appropriate excitation wavelength will yield a maximum emission intensity at a characteristic wavelength. The quality of a fluorescent image generally combines contrast and brightness. Contrast depends on the ratio of emission intensity to that of background light. Consequently, the use of filters to block the excitation wavelength and prevent it from overwhelming the field is critical in providing high contrast for the detection of fluorophore emission. However, the use of narrow filters to restrict the background can severely limit the brightness of the fluorescent image. This is due to the barrier filters allowing only a portion of the actual emission spectrum to pass. Brightness is an intrinsic property of each individual fluorophore, but is also dependent on the ability to illuminate with the peak excitation wavelength for that fluorophore.

The balance between contrast and brightness may be further complicated by applications designed to facilitate the evaluation of multiple fluorophores. The selection of appropriate excitation and emission filters within a single set to allow for the simultaneous viewing of multiple fluorophores can be very inhibiting. Although quad-band beamsplitters are currently available, commercial advances to date have yielded only triple-pass combinations for simultaneous detection, such as one of which permits the simultaneous detection of the widely used fluorophore combination fluorescein-5-isothiocyanate (FITC), Texas Red and 4′,6-diamidino-2-phenylindole (DAPI) (e.g., Chroma Technology Corp.).

Although the capture of digital images has propelled this field, efforts to collect multicolored images as digital files also suffer from technical limitations. Color CCD cameras often utilize filters or liquid crystals to segregate the red, green, and blue wavelengths onto the imaging chip. The employment of these cameras is restricted to resolving colors as discriminated by the filters used. For scientific applications, then, the greatest sensitivity and selectivity has been achieved through black and white CCD cameras with optimized filters for a specific application. While this approach is widely used, it suffers from the inherent limits of the spectra of data that can be analyzed since it is defined by the characteristics of the optical filters employed. Thus, the system is restricted to the use of fluorophores that can be effectively resolved from each other by filters, and the filters selected limit the data collected to a portion of the total emission spectrum for any one fluorophore.

Multispectral Imaging

Multispectral imaging systems in biomedical research are composed of a microscope, filters and/or interferometer for spectral selection, a charged coupled device (CCD) camera to record the data, and computer control for acquisition, analysis, and display (Figure 1a). The introduction of CCD cameras has offered a means to document results and capture images as digital computer files. This has facilitated the analysis of results with more and more complex probe sets and has allowed the use of more sophisticated computer algorithms for data analysis. The other major component of the multispectral imaging system is a light source to excite the sample's fluorescent dyes whose emissions are selected by filters or an interferometer for recording on the CCD camera.

An example of the use of multispectral imaging is in the identification of the 24 human chromosomes. A scheme was presented for the analysis of human chromosomes by differentially labeling each chromosome with a unique combination of 5 fluorophore-labeled nucleotides (building blocks of DNA and RNA, consisting of a nitrogenous base, a five-carbon sugar, and a phosphate group). These complex probes are simultaneously hybridized to metaphase chromosomes and analyzed to determine which of the fluorophores are present. Combinations of one, two, or three of the five fluorophores are enough to distinguish a particular chromosome from the others.

An interferometric multispectral system used to perform this type of analysis is the Spectral Karyotyping (chromosome typing) or SKY System (Figure 1a). This technology does not represent hyperspectral imaging, but is presented for comparison. Illumination light passes through a triple band-pass dichroic filter which allows for the simultaneous excitation of all the fluorochromes in, for instance, Speicher's scheme. Modern multi-band pass filters add little to image distortion while effectively blocking unwanted autofluorescence or excitation illumination. The fluorescence emissions from the sample are split into two separate beams by the interferometer. The beams are then recombined with an optical path difference (OPD) between them such that the two beams are no longer in register and interfere with each other. By adjusting the OPD, a particular wavelength band can be selected for, which results in constructive interference and thus can be collected by the CCD camera pixels. After adjusting the OPD to record

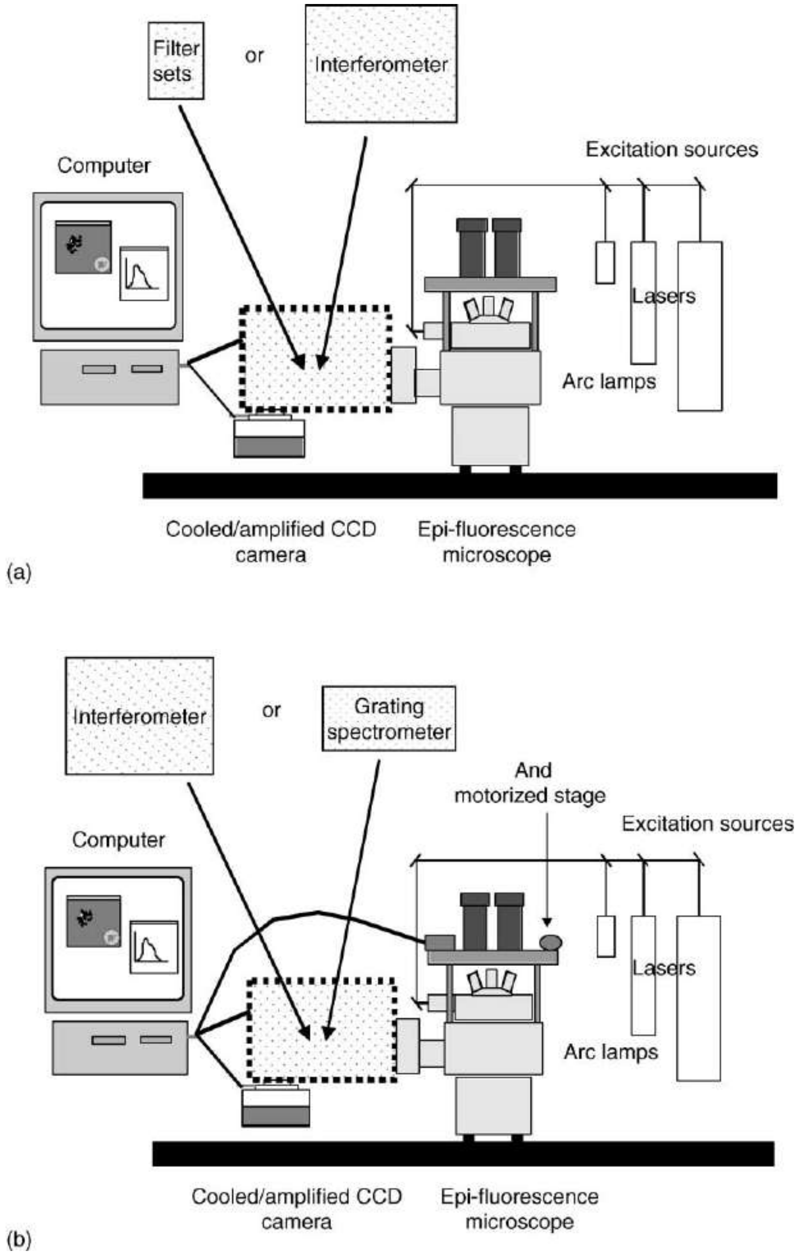
Filter sets
or
Interferometer
Computer
Excitation sources
Lasers
Arc lamps
Cooled/amplified CCD camera
Epi-fluorescence microscope
(a)
Interferometer
or
Grating spectrometer
And motorized stage
Computer
Excitation sources
Lasers
Arc lamps
Cooled/amplified CCD camera
Epi-fluorescence microscope
(b)

Figure 1 The instruments for multispectral and hyperspectral imaging are similar. (a) Multispectral imagers use interferometers or bandpass filters to collect noncontiguous emission wavelengths. (b) Hyperspectral imagers employ interferometers or gratings to collect a contiguous wavelength band. Computers, microscopes, CCD cameras, and excitation light sources are common to all the techniques. The systems which use filter sets or interferometers use motorized filter placement systems or automated interferometers. The gratings systems employ a motorized stage for the sample movement. Adapted with permission from Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R and Garner HR (2001) Hyperspectral imaging: a novel approach for microscopic analysis. *Cytometry* 43: 239–247.

the 5 dye wavelengths, each pixel's 5 records are then compared to the known probe-labeling scheme to identify the corresponding chromosome material. The image of the metaphase chromosome spread can be karyotyped using specific pseudo-colors for maximal visual discrimination of the chromosomes.

Another nonhyperspectral method is the Multi-Fluorescence *In Situ* Hybridization or M-FISH System. It uses five single-band-pass excitation/emission filter sets where each filter set is specific for one of the five fluorochromes used in the combinatorial labeling process (Figure 1a). Again, five separate exposures are taken (one exposure for each filter set) and combined to create a multicolor image. Most M-FISH systems now have motorized filter set selectors which minimize any image registration shifts that may occur between exposures. For each pixel, the presence/absence of signal from each of the five fluorochromes is assessed and then matched to the probe-coding scheme to determine which chromosome material is present. As with the SKY System, an image of the metaphase spread is converted into a karyotype and specific pseudo-colors are used to maximize visual discrimination of the chromosomes.

Hyperspectral Imaging

As has been stated, the main difference between multi- and hyper-spectral imaging is the collection of a broad contiguous band of wavelengths. The apparatus and methods for hyperspectral imaging are similar to multispectral imaging (Figure 1). Instead of collecting sections of the emission spectrum using filters or interferometers, hyperspectral imaging collects a continuous emission spectrum by dispersing the wavelength of light using gratings and measuring the wavelength intensity, or by dispersing Fourier components of light using interferometers and measuring interference intensity (Figure 1b). Multispectral images are limited to measuring total intensity within the selected wavelength band. Determining the abundance of more than one dye per pixel is less accurate when only a small segment of each dyes' spectrum is passed by the emission filter. This has not been considered a limitation in the past because most research depended on just identifying the presence and not the abundance of several spectrally broad and overlapping dyes which also occupied relatively large spatial areas. With the advent of smaller and smaller featured microarrays, the desire to use more dyes in identification schemes, and the use of quantum-dot markers which can be used to indicate both genetic makeup and quantity, the limitations of multispectral imaging has led to a move to hyperspectral imaging.

In contrast to the fluorescence multispectral imaging approaches, a hyperspectral image is fully three-dimensional. Every XY spatial pixel of the image has an arbitrarily large number of wavelengths allied with it. An intensity cube of data is collected. Two of the axes of the cube are spatial and one axis is wavelength (see Figure 2). At each of the 3D points, there is an associated intensity of emitted energy. Therefore, it is a four-dimensional information space. The individual intensity signature of differing emitting bodies can then be separated to give a complete view. Spectral decomposition is a mathematical/algorithmical process by which individual emission spectra of the dyes can be uniquely separated from a composite (full spectral) image by knowing the signature of each dye emission spectra component measured separately. Spectral decomposition may be performed using standard algorithms with codes which are readily available.

Interferometric Hyperspectral Imaging

The system architecture of the interferometric hyperspectral imager resembles that of the SKY System (Figure 1). But, instead of just sequencing the interferometer to select a set of specific wavelengths to record as in the SKY System, the interferometer in the hyperspectral imager is sequenced through

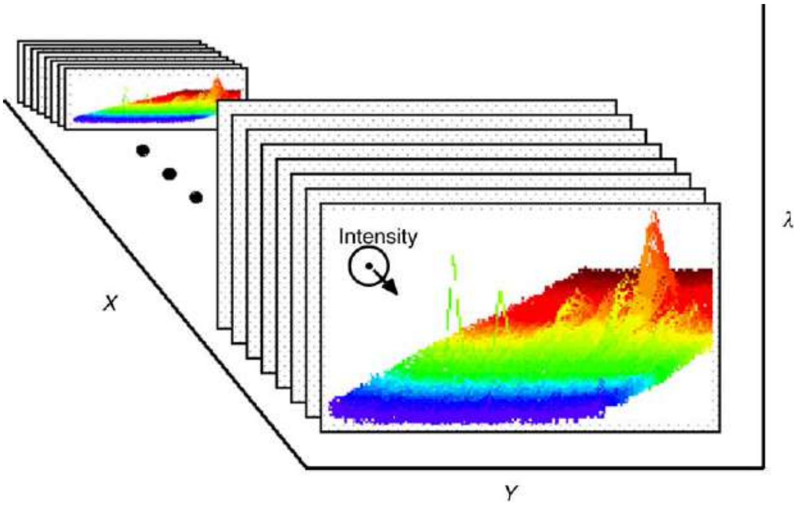

Figure 2 Intensity data cube. In this data cube, each CCD exposure captures a measurement in intensity (number of counts) by wavelength for a small spatial area in *Y* and ΔX. Many displacements in *X* with subsequent CCD acquisitions yield a 'cube' of data in *X*, *Y*, and λ. The intensity counts may be thought of as another dimension, $I(x,y,\lambda)$. Another form of the data cube is with each CCD acquisition comprising intensity counts for a spatial *X* and *Y* area for a small $\Delta\lambda$ and multiple displacements in λ for each acquisition.

discrete optical path differences to record the intensity at each pixel for each OPD. Fourier spectroscopy techniques are used to transform the OPD intensities at each pixel into the wavelength intensities at each pixel. Each pixel may be thought of as a Fourier spectrometer which is not dependent on any other pixel. Then, spectral decomposition algorithms can be run to determine which dyes are present (see the next section).

Like the multifilter methods, the interferometer approach provides a real two-dimensional image that can be observed, focused, and optimized prior to measurement. But like the filtering methods, the excitation source must be filtered for sufficient contrast to view the fluorescent emissions. Interferometers have high optical throughput, relative to other methods, and can have high and variable spectral resolution. The spectral range can cover from the ultraviolet to the infrared. The high and variable spectral resolution allows the user to trade off sensitivity, spectral resolution, and acquisition time in a way that is not possible with any other technique. And interferometers do not cause polarization of light as a filter may.

However, interferometric instruments are susceptible to image registration and lateral coherence difficulties. They are limited by the extreme precision required of the optical components which results in the continuous need for monitoring and adjustment, since aberrations are not tolerated in interferometric instruments. The throughput is reduced by loss due to the beamsplitter design and from the mirrors. Additionally, the Fourier transform instrument deconvolutes based on a two-step process; the correlation of the light passed by an interferometer followed by an inverse Fourier transform in software.

To mitigate the extreme precision and sensitivity of the most interferometric systems, Sagnac interferometers have been employed because of their common path intrinsic stability. The common path compensates any misalignment, vibration, or temperature affect. The OPD is provided by the Sagnac affect. When the Sagnac interferometer is rotated, the path traveled in one direction will be different from that of the light on the common path in the other direction. Although, this interferometer method is trading one difficulty for another, it is somewhat easier to manage the rotation of a physical system than to maintain the exacting precision of mirror alignments in a standard interferometer.

Spatial Resolution

The spatial resolution of a Sagnac Hyperspectral Imager is set by the size of the pixels of the CCD camera and the microscope system magnification, since the interferometric hyperspectral viewer sees the entire field of view at once. Let the x- and y- dimensions of the CCD pixel be δx and δy, respectively, and the magnification be M, then the spatial resolutions in the x and y directions are $2\delta x/M$ and $2\delta y/M$, respectively. For CCD pixel sizes of 10 microns square, binning together two pixels in the x- and two pixels in the y- direction, and an overall magnification of 60, the x and y resolutions are both 0.66 microns.

Field of View

For N pixel rows and C pixel columns, the field of view will be $N\delta y/M$ and $C\delta x/M$ for the length of the area along the y- and x- axis, respectively. For a CCD array of 1536 columns by 1024 rows, and the parameters above, the field of view is 171 microns by 256 microns, x and y, respectively.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution will depend on the ability to vary the OPD and the Fourier transform. While the interferometer is used basically as a filter in the multispectral imager to select a wavelength, in the hyperspectral imager it is used to vary the OPD for all the wavelengths whose energy is integrated at the CCD camera pixels.

The electric field, E, for emissions at one wavelength, λ, or wave number, $k = 1/\lambda$, radial frequency, ω, and amplitude, A, can be written as

$$E(k) = A(k)\, \mathrm{e}^{-\mathrm{i}(2\pi kx - \omega t)} \qquad [1]$$

and the intensity which is measured is the magnitude squared of the electric field:

$$I = E^{*}E \qquad [2]$$

Emissions from fluorescent sources are usually noncoherent and composed of a range of wavelengths, even with only one dye present. The intensity measured at a pixel is the superposition of many wave numbers. After traveling a certain OPD, L, the intensity integrated over all wave numbers may be written as

$$I_{\mathrm{pxl}}(L) = \frac{1}{2}\left[\int I(k)\mathrm{d}k + \int I(k)\cos(2\pi kL)\mathrm{d}k\right] \qquad [3]$$

or

$$I_{\mathrm{pxl}}(L) = C_0 + C_1 \cdot \mathrm{Re}\{\mathrm{FT}[I(k)]\} \qquad [4]$$

where FT stands for Fourier transform.

The inverse transform is

$$\int I_{\mathrm{pxl}}(L)\,\mathrm{e}^{-\mathrm{i}2\pi k'L}\,\mathrm{d}L = C_0 \int \mathrm{e}^{-\mathrm{i}2\pi k'L}\,\mathrm{d}L + C_1 \int\int I(k)\mathrm{e}^{-\mathrm{i}2\pi kL'}\,\mathrm{d}k \cdot \mathrm{e}^{\mathrm{i}2\pi k'L}\,\mathrm{d}L \qquad [5]$$

$$\int I_{\mathrm{pxl}}(L)\mathrm{e}^{-\mathrm{i}2\pi k'L}\,\mathrm{d}L = C_3 + C_4 I(k)$$

The integration constant, C_3, can be ignored for large wave numbers, as in the case of the visible range, since C_3 will be proportional to the inverse of the wave number. The intensity as a function of wave number is then

$$I(k) = C_5 \int I_{\mathrm{pxl}}(L)\mathrm{e}^{-\mathrm{i}2\pi kL}\,\mathrm{d}L \qquad [6]$$

For absolute quantitative measurements, the constant C_5 can be determined by calibration with a known emission. For discrete measurements made by changing the optical path length, then the working formula is

$$I_{\mathrm{pxl}}(k) = \sum_{L_i} I_{\mathrm{pxl}}(L_i)\mathrm{e}^{-i2\pi kL_i}\,\Delta L \qquad [7]$$

This is the discrete Fourier transform which must be preformed for each pixel to determine the intensity as a function of wavelength. Per the Nyquist sampling tenets:

$$\Delta L = \frac{1}{(N/2)\Delta k} \quad \text{or} \quad \Delta k = \frac{2}{N\Delta L} \qquad [8]$$

where Δk is the discrete wave number interval and N the number of different OPDs. Sagnac interferometers can make rotations which give at least a change of $\Delta L = \lambda/4$, so substituting in eqn [8]:

$$\Delta k = k_1 - k_0 = \frac{1}{\lambda_1} - \frac{1}{\lambda_0} = \frac{\Delta\lambda}{\lambda^2 + \lambda\Delta\lambda} = \frac{8}{N\lambda} \qquad [9]$$

$$\frac{\Delta\lambda}{\lambda^2} \approx \frac{8}{N\lambda} \quad \text{or} \quad \frac{\Delta\lambda}{\lambda} \approx \frac{8}{N}$$

A 5% resolution in wavelength (or 20 nm at 400 nm) requires $N = 160$ and, of course, a 2.5% resolution requires $N = 360$. Therefore, about 400 OPD changes and acquisitions of the intensity fringes will give calculated resolutions of less than 10 nm in wavelength.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Software techniques for displaying data will be described in the next section.

Grating-Dispersion Hyperspectral Imaging

A hyperspectral imaging microscope collects the complete visible emission spectrum from a microscope slide by using a grating spectrometer to provide the wavelength dispersion. This microscope correlates spatial and spectral information with minimal use of optical filters. A hyperspectral microscope system is composed of a standard epi-fluorescence microscope (Olympus IX70) equipped with objectives lens (including 1.25 × and 10 × Plan Fluorite, 40 × Plan Apochromat Dry, and 60 ×, and 100 × Plan Apochromat Oil), as well as a set of standard filter cubes (Chroma Technology) to allow for traditional visualization of dyes through the eyepiece. (Prototype typical components are listed in parentheses.) Multiple excitation sources can be utilized including standard 100 W mercury and xenon arc lamps, tungsten brightfield illumination, a 532 nm solid state laser (Brimrose), a helium neon laser (Uniphase), an argon ion laser (Uniphase), and a pulsed-doubled nitrogen dye laser (Laser Science, Inc.) (Figure 1b).

One side port of microscope is optically coupled to a spectrograph (Acton Research SpectraPro® 556i). A narrow entrance slit into the spectrograph allows for only one line (ΔX) of the image to be captured at a time. A 50-micron slit permits 80% of maximum throughput for the spectrometer. The spectrograph separates the polychromatic light illuminated from this single line into its light components. Normal operation uses a grating with 50 grooves/mm and a 600 nm-blaze wavelength, which allows wavelengths of approximately 400–780 nm to be measured. Different spectrograph gratings can be used for special applications from near UV to near IR. The center wavelength and grating can be shifted under computer control. The output of the spectrograph is then captured with an air/liquid-cooled CCD camera (Photometrics-Quantix) and an optional image intensifier (Video Scope International VS4-1845). The camera has a resolution of 1536 × 1024 pixels with a depth of 12 bits per pixel. The 1536 pixels are aligned in the spatial direction to allow the greatest spatial resolution. During normal operation the wavelength resolution is 'binned' to 128 pixels, sufficient to resolve ~8 different spectra in a single excitation scan. Higher resolution can be employed for special applications.

A line of excitation light can be provided by a line generator when laser excitation is used or by a combination of slits and cylindrical lenses when using traditional lamps or burners (mercury or xenon). This line of light illuminates the sample in the same position that maps the resulting emission light through the imaging spectrograph and onto the camera. Some advantages of illumination with a line of light include decreased sample bleaching, reduced background due to scattering, and efficient use of available excitation power. A traditional bandpass, a low-pass dichroic filter, or a linear variable filter (Raynard Corp.) may be employed to select or restrict the excitation wavelength during specific applications.

Linear Variable Filter (LVF)

The linear variable filter (LVF) (Raynard Corporation) can be used to replace the standard excitation filter cube to provide a capability for continuous excitation wavelength band selection. The LVF is $1'' \times 2.25''$ and is coated such that it continuously filters light from ~400 nm to ~800 nm along its length. Custom filters that span other wavelengths are available. The coating and therefore wavelength is uniform along a short distance, making it an ideal method for selecting an excitation wavelength band to paint across the slide, excite the slide's contents, and map the contents' emission into the slit of the imaging spectrograph. The bandwidth passed is about 21 nm wide. The LVF introduces ~50% attenuation of the intensity.

Spatial Resolution

The spatial resolution of a grating spectrograph Hyperspectral Imager is set by the size of the pixels of the CCD camera in the y- direction and the microscope system magnification. However, differing from the interferometric system, the spatial resolution in the x- direction depends on the spectrometer slit width and the microscope system magnification. As before, let the x- and y- dimensions of the CCD pixel be δx and δy, respectively, and the magnification be M. Let w_s be the slit width of the spectrometer, and note than the slit width is always going to be larger than a few δx. Then the spatial resolution in the y-direction is again $2\delta y/M$, but the spatial resolution in the x- direction is now $2w_s/M$. Again, for CCD pixel sizes of 10 microns square, binning together two pixels in the y-direction, and an overall magnification of 60, the y- resolution is 0.66 microns, as it was for the interferometer method. For a slit width of 50 microns, the x- resolution is 1.67 microns, independent of the number of pixels binned in the x-direction. The control software moves the stage the right number of steps to minimize the overlap or skipped areas. The slit width can be reduced, but at a loss of throughput or increased exposure time. The control software can compensate for different spatial resolution to display the image as 1:1.

Field of View

For N pixel rows (y- dimension) and C pixel columns (x- dimension), the field of view will be $N\delta y/M$ for the length of the area along the y-axis. If the spectrometer slit length is less than the height of the CCD, then the length is just h_s/M, where h_s is the slit length. We orient our CCD array so that the 1536-pixel dimension is along the y-axis. The y-view is thus 256 microns as in the interferometric example. To achieve the same field of view as the Sagnac method imager, we would need to take 103 stage movements and acquisitions. This highlights the difference in the two hyperspectral methods. The grating method collects the complete spectrum on each acquisition with no spectrometer adjustment but needs to step down the slide to produce the full spatial image. The interferometer method collects the complete spatial image on each acquisition with no position adjustment but needs to step the interferometer through a range of optical path differences to acquire multispectral content.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution depends on spectrometer instrument function which includes the slit width, the dispersion of the grating, the CCD pixel size, and number of pixels binned together. The dispersion for normal operation was determined to be ~40 nm/mm or 0.4 nm/pixel in the 10 micron CCD pixel example. For example, binning of 8 pixels yields a spectral resolution of 3.2 nm at all wavelengths. To achieve this spectral resolution, the interferometer method would require 1000 OPD increments and acquisitions.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Although the spatial resolutions differ considerably in the two methods, as long as the dye peaks are separated by the spectral resolution spacing, the spectral decomposition algorithms will find all dyes within the spatial resolution area. We have demonstrated that the analysis algorithms can resolve 5 dyes with peaks separated by more than 6 nm in various combinations and known quantity ratios against the dye standards. The program successfully identifies the dyes present with a 91% accuracy in the quantities.

Software

Software (*Hyper*Scope, a 'C' language program) controls all the hardware components for acquisition.

It includes features for analyzing and displaying the resultant $X-Y$ images and spectral information for the grating spectrometer Hyperspectral Imager. During a scan, individual pictures are taken in the $Y-\lambda$ plane, while the stage is incremented in the X-direction. A collection of such $Y-\lambda$ scan files, which can total from 50–1000, is merged to build an image cube (Figure 2). Another advantage of the single-line acquisition mode is that time-dependent features can be monitored continuously. The software was designed to permit repeated collection of a single line at variable time points. The utility of this feature has been demonstrated by capturing images of fluorescent objects moving through UV-transparent capillaries.

The software was designed to display the resulting $X-Y$ plane extracted in various ways from the image cube. A graph of the emission spectrum for any pixel within that image may be viewed by clicking on any pixel of interest in the $X-Y$ image (Figure 3c). A standard linear curve-fitting algorithm is used to determine the contribution of individual dyes to the measured emission spectrum. The software also features a windowing technique (integration of the emission spectra over a fixed window) which emulates standard filtering. An overlay feature within the software allows the curve fitting results to be displayed in false color and compared by viewing the composite for multiple fluorophores in a single image. Additionally, it permits the contribution of a fluorophore to be emphasized for visualization purposes by either displaying the contribution of a single fluorescent signature in the absence of the others that were present in the image or by scaling individual display intensity values up or down. Overlapping pseudo-colors assigned to represent each fluorophore (either by curve fitting or by windowing) may be displayed as added, averaged, maximum or minimum values. For each new fluorophore–excitation source combination, the characteristic emission spectra is acquired and added to a library of available spectra that can be used as components in the spectral decomposition. This permits the user to introduce new fluorophores at any time. Moreover, background spectra (defined as a region of interest by the user) can be acquired and used as one element of the decomposition. This background component can be turned off in the false color view (as any of the spectral components may be) resulting in enhanced noise reduction and visual contrast improvement.

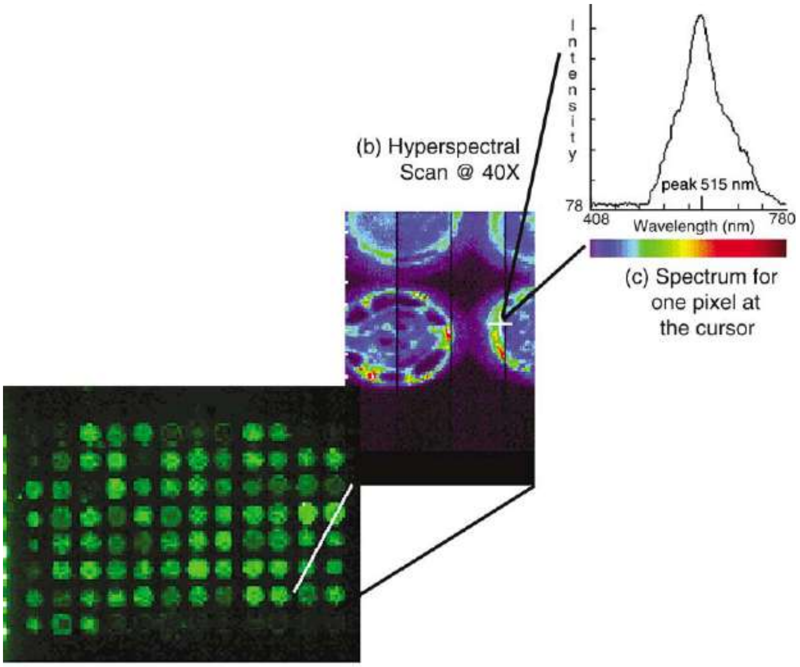

(b) Hyperspectral Scan @ 40X

(c) Spectrum for one pixel at the cursor

(a) Genepix4000 Scan

Figure 3 The benefit of spectral imaging is demonstrated in a spatial high-resolution scan of an expression microarray. (a) A portion of the scan of an expression microarray on a microscope slide using the Genepix4000 scanner. It records the intensity emitted by the Cy3 dye when excited with a 532 nm laser. (b) A portion of the scan of the same microarray with the Hyperspectral Imaging Microscope using a 40X objective and Hg excitation lamp. False color is displayed to indicate the intensity variations. (c) A *Hyper*Scope computer program display of the spectrum recorded when the cursor is placed over one pixel in a scan.

Description of Acquisition

In order to acquire a full image, a slide is placed on the microstage mover (Ludl Electronics) that allows for precise movement of the sample across the image acquisition plane. The user can position the stage via a joystick or via software to move to predetermined coordinates. The image is then scanned, with the number of scans being selected at the time of the image acquisition. Furthermore, the sample can be scanned at one magnification or resolution and then rescanned at another. Each $Y-\lambda$ image is saved as a single $Y-\lambda$ file for each increment along the sample. Thus, each image file corresponds to the excitation line of light and the emission collection region that maps through the imaging spectrograph to the CCD camera. Saving individual files permits unlimited $Y-\lambda$ image acquisition at the same Y coordinate, valuable for time-dependent studies. This feature also provides some robustness by enabling a partial $X-Y$ image to be displayed should acquisition of one or more Y scans fail for any reason. After acquiring each image, the stage is moved automatically, one or more steps in the X direction, to acquire emissions at the next line and generate the next $Y-\lambda$ file. Therefore, only the fluorescent probes in a line at a single X-location are examined at each acquisition time. Step sizes can be adjusted by the software to give: (i) the best resolution; (ii) the best 1:1 image; (iii) the minimum overlap; or (iv) the fast acquisition time.

After a variable number of rows of associated emission have been captured (typically 50–1000 scans), an image cube is loaded into memory comprising all the $Y-\lambda$ data files. A graph of the spectrum from any pixel (or bin) in the $X-Y$ plane may be displayed in a separate window by clicking on the cube (Figure 3c). A typical acquisition scan of 250 $Y-\lambda$ files, which are used to construct a $250 \times 768 \times 128$ ($\times 12$ bit) image cube, consists of approximately 100 Mb of information.

MicroArray Scanning and Customized Excitation

Another use of this technology is to analyze spotted microarrays. The Hyperspectral Imager has been compared to commercial two-color scanners with comparable sensitivity at approximately 0.5 fluors per square micron (Figure 3). Increased sensitivity (500–10 000 gain) can be provided by using a microchannel plate amplifier (Model VS4-1845, Videoscope International, Ltd.) installed between the CCD camera and the imaging spectrograph.

Since the hyperspectral imaging microscope collects images that reflect only the $Y-\lambda$ plane, it contains the potential to use a single line of light as a source for image illumination. Therefore, lights systems being developed, which are capable of generating a line of light composed of variable or complex wavelengths, could be employed to precisely control excitation. The incorporation of such an excitation source in hyperspectral imaging microscopy could be useful in many clinical and basic research applications.

See also

Environmental Measurements: Hyperspectral Remote Sensing of Land and the Atmosphere. **Imaging:** Interferometric Imaging.

Further Reading

Anguiano A (2000) Fluorescence in situ hybridization (FISH): Overview and medical applications. *Journal of Clinical Ligand Assay* 23: 33–42.

Berland K, Jacobson K and French T (1998) Electronic cameras for low-light microscopy. *Methods in Cell Biology* 56: 19–44.

Bernheim A, Vagner-Capodano AM and Couturier J (2000) From cytogenetics to cytogenomics of tumors. *S-Medical Science* 16: 528–539.

Bezrookove V, Hansson K, van der Burg M, *et al.* (2000) Individuals with abnormal phenotype and normal G-banding karyotype: improvement and limitations in the diagnosis by the use of 24-colour FISH. *Human Genetics* 106: 392–398.

Boland MV and Murphy RF (1999) Automated analysis of patterns in fluorescence-microscope images. *Trends in Cell Biology* 9: 201–202.

Cardullo RA and Alm EJ (1998) Introduction to image processing. *Methods in Cell Biology* 56: 91–115.

Garini Y, Katzir N, Cabib D, Buckwald RA, Soenksen D and Malik Z (1996) Spectral bio-imaging. In: Wang XF and Herman B (eds) *Fluorescence Imaging Spectroscopy and Microscopy*, pp. 87–122. New York: John Wiley and Sons.

Garini Y, Macville MV, du Manoir S, *et al.* (1996) Spectral Karyotyping. *Bioimaging* 4: 65–72.

Gothot A, Grosdent J and Palulus J (1996) A strategy for multiple immunophenotyping by image cytometry: Model studies using latex microbeads labeled with seven Streptaviden-bound fluorochromes. *Cytometry* 24: 214–225.

Gribble SM, Roberts I, Grace C, Andrews KM, Green AR and Nacheva EP (2000) Cytogenetics of the chronic myeloid leukemia-derived cell line K562: Karyotype clarification by multicolor fluorescence *in situ* hybridization, comparative genomic hybridization, and locus-specific fluorescence in situ hybridization. *Cancer Genet Cytogenetics* 118: 1–8.

Han H, Gao X, Su J and Nie S (2001) Quantum-dot-tagged microbeads for multiplexed optical coding of bimolecules. *Nat Biotech.* 19: 631–635.

Huebschman ML, Schultz RA and Garner HR (2002) Characteristics and capabilities of the hyperspectral imaging microscope. *Eng in Med and Bio*. 21: 104–117.

Inoue T and Gliksman N (1998) Techniques for optimizing microscopy and analysis through digital image processing. *Methods in Cell Biology* 56: 63–90.

Jain AK (1997) *Fundamental of Digital Image Processing*, pp. 80–128. New Jersey: Prentice-Hall.

Jalal SM and Law ME (1999) Utility of multicolor fluorescent in situ hybridization in clinical cytogenetics. *Genet Med*. 1: 181–186.

Johannes C, Chudoba I and Obe G (1999) Analysis of X-ray-induced aberrations in human chromosome 5 using high-resolution multicolour banding FISH (mBAND). *Chromosome Research* 7: 625–633.

Kruse F (1993) The Spectral Image Processing System (SIPS) – Interactive visualization and analysis of imaging spectrometer data. *Remote Sensing Environment* 44: 145–163.

Lu YJ, Morris JS, Edwards PAW and Shipley J (2000) Evaluation of 24-color multifluor-fluorescence *in-situ* hybridization (M-FISH) karyotyping by comparison with reverse chromosome painting of the human breast cancer cell line T-47. *Chromosome Research* 8: 127–132.

Malik Z, Cabib D, Buckwald RA, Talmi A, Garini Y and Lipson SG (1996) Fourier transform multipixel spectroscopy for quantitative cytology. *Journal of Microscopy* 182: 133–140.

Press WH, Flannery BP, Teukolsky SA and Vetterling WT (1986) *Numerical Recipes, the Art of Scientific Computing*, pp. 546. New York: Cambridge University Press.

Raap AK (1998) Advances in fluorescence in situ hybridization. *Mutat. Res. Fundam. Mol. Mech. Mutagen* 400: 287–298.

Sakamoto M, Pinkel D, Mascio L, *et al.* (1995) Semiautomated DNA probe mapping using digital imaging microscope: II. system performance. *Cytometry* 19: 60–69.

Schrock E, duManoir S, Veldman T, *et al.* (1996) Multicolor spectral karyotyping of human chromosomes. *Science* 273: 494–497.

Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R and Garner HR (2001) Hyperspecteral imaging: a novel approach for microscopic analysis. *Cytometry* 43: 239–247.

Schwartz S (1999) Molecular cytogenetics: show me the colors. *Genet Med*. 1: 178–180.

Speicher MR, Ballard SG and Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* 12: 368–375.

Wang YL (1998) Digital deconvolution of fluorescence images for biologists. *Methods in Cell Biology* 56: 305–315.

Zhao L, Hayes K and Glassman A (2000) Enhanced detection of chromosomal abnormalities with the use of RxFISH multicolor banding technique. *Cancer Genet Cytogenetics* 118: 108–111.

Imaging Through Scattering Media

A C Boccara, Ecole Supérieure de Physique et Chimie Industrielles and Centre National de la Recherche Scientifique, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We are often faced with the problem of seeing through scattering media such as smoke, fog, clothes, blood, tissues, ceramics, walls, etc. Tremendous effort has been made to introduce techniques which will allow us to overcome the intrinsic difficulty of selecting ‘good’ photons which are able to provide valuable information on structures which appear fuzzy or, for most of the time, completely hidden.

Although a number of concepts and techniques that we will discuss in this article could be applied to a large number of scattering samples, we will restrict ourselves to biological tissues.

Studying light propagation through biological tissues is becoming more and more popular among physicists, biologists, engineers, and doctors. Indeed this research domain appears to be growing rapidly as seen by the number of published articles, conferences, and international meetings. Optical techniques are attractive and are starting to compete with other well-established techniques (magnetic resonance imaging, X-ray imaging, ultrasonic imaging, positron emission tomography, etc.), because they are nonionizing, nondestructive, they can reach high resolution and are usually much cheaper. Optical techniques reveal an optical contrast which is valuable in providing morphologic as well as functional information.

Most biological tissues are weakly absorbing in the deep red and near infrared region of the spectrum; the main difficulty in performing imaging arises from the high level of scattering. Figure 1 shows the picture of a small flash light bulb placed in the mouth: it appears fuzzy and red.

In Figure 2 we consider various photon trajectories between a short pulsed source S and a fast detector D.

Unscattered photons are called 'ballistic' photons, their number decreases exponentially, and they are completely damped after one or two millimeters.

Then there are scattered photons which propagate close to the ballistic trajectory reaching the detector; these are named 'snake-like' photons and can be used for imaging purposes.

Most of the energy which reaches D is related to the scattered photons which may exhibit very long trajectories (typically ~10 times the distance S–D).

The description of various experimental methods used to overcome the difficulty introduced by scattering will be associated with the various photon families that we have mentioned: ballistic, snake-like, and scattered photons. We will first give a few orders of magnitude of the parameters which will dictate the way light interacts with the tissues. Let us emphasize that these parameters also carry the information related to the contrast mechanism. The ultimate goal is to realize an 'optical biopsy', noninvasive but as close as possible to the regular biopsy followed by histopathology.

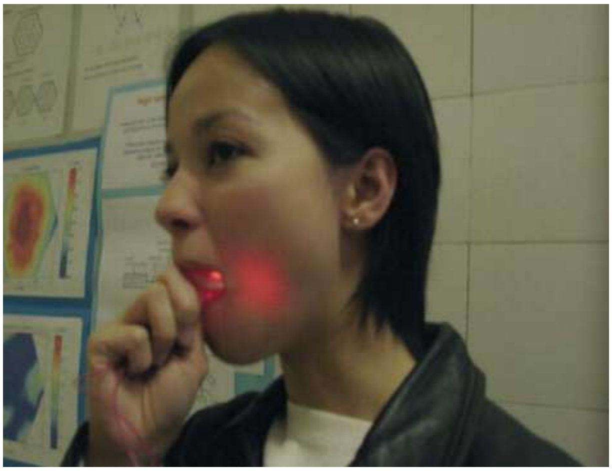

Figure 1 Biological tissues are highly scattering but the absorption is low in the red (as seen in this digital visible photography) and the near infrared.

Optical Parameters

Scattering Cross-Section

As shown in Figure 3, biological tissues are constituted of structures which exhibit various characteristic scales: from a few nanometers in membranes to about ten micrometers for an entire cell.

As we know from scattering theories these various scales correspond to various scattering properties. For small particles ('small' means much smaller than the optical wavelength λ) we are in the Rayleigh regime which is based on the evaluation of a time-varying dipole induced by a uniform time-varying (at the light frequency scale) electric field. This dipole radiates in the far field and its scattering cross-section, which is the ratio of the scattered power to the incoming irradiance for a sphere of radius r, is given by

$$\sigma = \frac{\langle P \rangle}{I} = \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \frac{8\pi}{3} r^6 k^4 \qquad [1]$$

n being the refractive index of the sphere divided by the index of the surrounding medium and $k = 2\pi/\lambda$. The scattering diagram for such particles is isotropic if we use unpolarized light. This cross-section is much smaller than the 'geometrical' section S of the particle; for instance, if $n = 1.5$, $a = 10$ nm, $\lambda = 0.6$ μm, $\sigma = 1510^{-8}$ μm^2, which is much smaller than the actual section (3×10^{-4} μm^2).

For larger particles the sphere (or the ellipse) interacting with a plane polarized wave is rather complicated because we are not restricted to the dipolar approximation (uniform field), but induced multipolar moments and their radiating fields have to be taken into account. This theory was proposed by Mie and can be found in optics textbooks. As an example, Figure 4 shows the scattering cross-section of a micron-size particle as a function of the

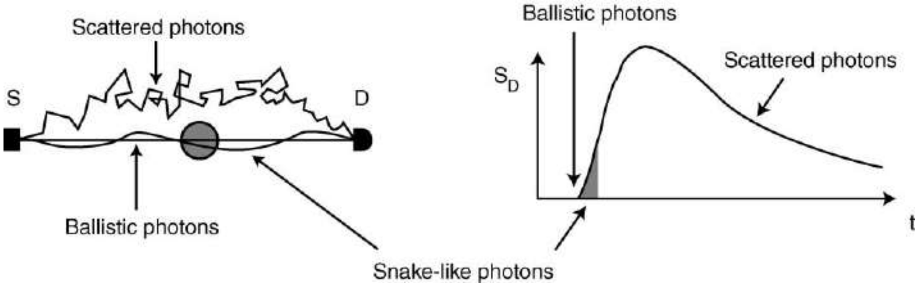

Figure 2 Between an ultrafast light source S and a detector D can be distinguished various classes of photons which can be used for imaging.

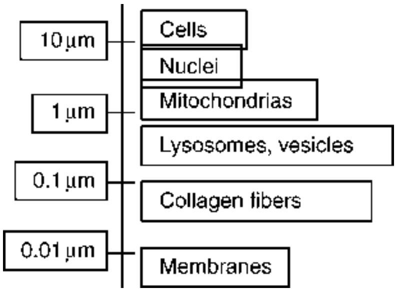

Figure 3 Various sizes of scatterers may be found in cell structures.

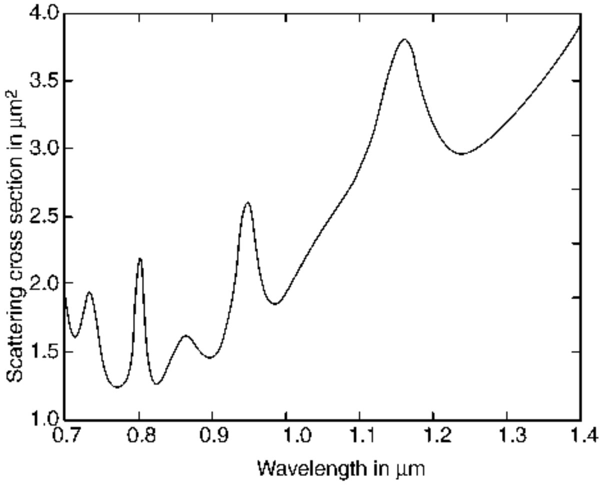

Figure 4 In the Mie scattering regime resonances are observed when varying the wavelength (here a 1 μm sphere, refractive index: 2).

wavelength: a number of 'resonances' can be observed.

The scattering diagram is no longer isotopic, it exhibits lobes, and forward scattering is much higher than back-scattering. Scattering is characterized by a scattering coefficient, μ_s, which is the inverse mean free path l_s of photons between scattering events:

$$l_s = 1/\mu_s = \sigma\rho \qquad [2]$$

where σ is the scattering cross-section and ρ the number of scatterers by unit volume.

The scattering coefficients μ_s of tissues are found within the range 100–1000 cm^{-1}. After a path length L, the number of ballistic photons will be damped by a factor $\exp(-L/l_s)$, so tissue imaging cannot rely only on ballistic photons as they survive only to a path length larger than about one mm. One needs to consider in more detail the scattering properties of the scatterers.

The dominant sizes of the scatterers in tissue (cells) demand consideration of the scattering in the Mie regime. Such a scattering event does not result in isotropic scattering angles. Instead, the scattering in tissue is biased in the forward direction.

Anisotropic scattering is quantified in a coefficient, g, which is defined as the mean cosine of the scattering angle θ, $p(\theta)$ being the probability of a particular scattering angle:

$$g \equiv \langle \cos(\theta) \rangle \equiv \frac{\int_0^{\pi} p(\theta)\cos(\theta)\sin(\theta)\mathrm{d}\theta}{\int_0^{\pi} p(\theta)\sin(\theta)\mathrm{d}\theta} \qquad [3]$$

For isotropic scattering, $g = 0$. For complete forward scattering, $g = 1$. In tissues, g varies from 0.7 to 0.95.

Photon migration, especially in the so-called diffusion regime, is often based on isotropic scattering such as for phonons (heat) or charge carriers scattering. For diffusion-like models, it has been shown that one may use an isotropic scattering model with a corrected scattering coefficient, μ_s', and obtain equivalent results using:

$$\mu_s' = \mu_s(1 - g) \qquad [4]$$

One can consider $l_s' = 1/\mu_s'$ as the length over which the incoming photon loses the memory of its initial direction, μ_s' is usually called the transport-corrected scattering coefficient.

Absorption Cross-Section

The absorption coefficient, μ_a (cm^{-1}) represents the inverse mean path length of photons before absorption. $1/\mu_a$ is also the distance in a medium at which intensity is attenuated by a factor of $1/e$ (Beer's Lambert law). Absorption in tissue is strongly wavelength dependent and is due to chromophores and water. Among the chromophores in tissue, the dominant component is the hemoglobin in blood. In the visible range, the blood absorption is relatively high, whereas it is much weaker in the near infrared. Water absorption is low in the visible and NIR region and increases above 1 μm, with tissues turning completely opaque above 2 μm. Thus, for greatest penetration of light in tissue, wavelengths in the 700–1300 nm spectrum are used. This region of the spectrum is called 'the therapeutic window'. Different spectra of chromophores allow one to separate the contribution of varying functional species in tissue such as quantification of oxy- and deoxy-hemoglobin to study tissue oxygenation.

Finally, when light is propagating in scattering media of thickness L, the damping of the incoming energy involves a combination of its absorption and scattering properties, and is given by

$$\exp(-L/a) \qquad \text{with} \qquad a^2 = 1/3\mu_a(\mu_s' + \mu_a) \qquad [5]$$

This formula reflects the fact that the photon's pathlength is increased by scattering, leading to an increase of the damping, which is related to absorption.

These few examples provide a 'static' view of light behavior in scattering media. Studying in a more detailed way, the spatial and temporal distribution of light, is more difficult.

Photon Migration in Scattering Media

Photon migration models have been inspired by other fields, such as astrophysics, soft matter physics, and neutronics, where physical media are random in both space and time. To avoid difficulties linked to the microscopic complexity such investigations rely on a statistical basis.

The most widely used theory is the time-dependent diffusion approximation to the transport equation:

$$\vec{\nabla}\cdot(D\vec{\nabla}\Phi(\vec{r},t)) - \mu_a\Phi(\vec{r},t) = \frac{1}{c}\frac{\partial\Phi(\vec{r},t)}{\partial t} - S(\vec{r},t) \qquad [6]$$

where $\vec{r}$ and t are spatial and temporal coordinates, c is the speed of light in tissue, and D is the diffusion coefficient related to the absorption and scattering coefficients by the formula:

$$D = \frac{1}{3[\mu_a + \mu_s']} \qquad [7]$$

The quantity $\Phi(\vec{r},t)$ is called the fluence, defined as the power divided by the unit area. This equation does not reflect the scattering angular dependence; the use of μ_s' instead of μ_s ensures that isotropic scattering has been reached. Indeed, for the use of the diffusion theory which implies anisotropic scattering, the diffusion coefficient is expressed in terms of the transport-corrected scattering coefficient; $S(\vec{r},t)$ is the source term.

The diffusion equation has been solved analytically for different types of measurements such as reflection and transmission modes, assuming that the optical properties remain invariant throughout the tissue. To incorporate the finite boundaries, the method of images has been used.

The diffusion approximation equation in the frequency-domain is the Fourier transform of the time-domain equation. This leads to the equation:

$$\vec{\nabla}\cdot(D\vec{\nabla}\Phi(\vec{r},\omega)) - \left[\mu_a + \frac{i\omega}{c}\right]\Phi(\vec{r},\omega) + S(\vec{r},\omega) = 0 \qquad [8]$$

Here the time variable is replaced by the frequency ω which is the modulation angular frequency of the source. In this model, the fluence can be seen as a complex number describing the amplitude and phase of the so-called 'photon density wave' in addition to a DC component:

$$\begin{aligned}\Phi(\vec{r},\omega) &= \Phi_{AC}(\vec{r},\omega) + \Phi_{DC}(\vec{r},0)\\ &= I_{AC}\exp(i\theta) + \Phi_{DC}(\vec{r},0) \qquad [9]\end{aligned}$$

where θ is the phase shift of the diffusing wave whose wavelength is:

$$2\pi\sqrt{\frac{2c}{3\mu_s'\omega}} \qquad [10]$$

Although analytical techniques provide a rapid understanding of most phenomena involved in static or dynamic situations, real biological samples are much more complex than homogeneous isotropic media. In these cases, direct and inverse algorithms map the spatially varying optical properties.

Monte Carlo simulations are very popular nowadays; a number of useful programs are available on the websites of various groups. Results may include time or polarization dependence of the photons along their paths.

The other method used in tissue optics is the random walk theory (RWT) on a lattice. RWT models the diffusion-like motion of photons in turbid media in a probabilistic manner. Using RWT, an expression may be derived for the probability of a photon arriving at any point and time given a specific starting point and time.

Imaging in the Ballistic Regime: Taking Advantage of Various Coherence Properties

Shallow Tissue Imaging Through Selection of Ballistic Photons

When imaging thin (less than 2 mm) tissue samples, it is possible to use photons which have not been scattered, called the ballistic photons which can be used to form high-resolution images in the same manner as if no scattering had taken place.

Their number, however, decreases exponentially with propagation distance and this ballistic signal is usually hidden by multiply scattered photons that obscure the image. For relatively shallow tissue depths it is possible to use various spatial filtering techniques to block the multiply scattered photons and there is much current research that aims to extend optical imaging and biopsy to depths of a few mm.

Confocal microscopy rejects much of the scattered light through spatial filtering and has been shown to form 3-D images at tissue depths of up to 200–300 μm, both *in vitro* and *in vivo*, which corresponds to about 50 dB of damping for the ballistic photons.

Two-photon fluorescence microscopy is particularly interesting in this regime since it ensures that all the detected fluorescence photons originate from the desired image plane. This technique also ensures a better penetration (using IR excitation) and a better resolution (nonlinear response) than single photon confocal microscopes.

Imaging to much greater tissue depths does not appear to be practical using optical microscopy and spatial filtering alone and more sophisticated techniques must also be employed to discriminate in favor of the ballistic photons. These techniques exploit either the fact that the ballistic signal photons retain spatial coherence of incident light or the fact that the ballistic photons keep the optical phase memory that multiply diffused photons have lost.

One of the most successful ballistic light-imaging techniques based on the latter filtering approach is optical coherence tomography (OCT) which uses coherence gating using low temporal coherence light and an interferometric detection. Coherent detection is used to discriminate against the multiply scattered photons that scatter back into the trajectories of ballistic photons. The ballistic light signal is measured by interference with a reference beam and the resulting pattern is detected with high sensitivity using homodyne or heterodyne detection. By using low coherence length sources and matching the time-of-flight of the reference beam signal to that of the desired ballistic light, OCT can acquire depth-resolved images in scattering media such as biological tissues. For powers in the mW range for tissue irradiation in the infrared, the unscattered ballistic (or more precisely single back scattered) component of the light will be limited by the shot noise detection limit after propagating through approximately 25 scattering mean free paths (MFP) of a scattering medium (or more than 100 dB). This corresponds to a ‘typical’ tissue thickness of about 1 mm tissue depth for the reflection geometry which is potentially useful for clinical applications such as screening for skin cancer, or for retinal examination.

OCT has proved clinically useful as a means of acquiring depth-resolved images in the eye and is being developed for application in strongly scattering biological tissue. Although the image acquisition requires pixel-by-pixel scanning, the use of new high-power superluminescent diodes in the 100 mW range or of ultrafast lasers to provide high average power, low coherence length radiation (sub-two cycle lasers) has resulted in an *in vivo* OCT system providing real-time imaging.

There are a number of other approaches to coherent imaging, including whole-field 2-D acquisition techniques which remove the need to scan transversely pixel by pixel.

Whole-field imaging using a temporally and spatially incoherent source and a well-balanced interferometric microscope intrinsically provides higher image acquisition rates and can take advantage of inexpensive spatially incoherent, broadband sources such as high-power LEDs, and white light thermal sources.

As an example, Figure 5 shows Linnik-type white light interferometer with optical path modulation and Figure 6 an image of an *ex-vivo* rat retina where one can recognize the main feature expected from histology.

Holography, which works in the Fourier space rather than in the object or image space, is also an interferometric technique able to discriminate between ballistic and scattered photons. Since the development of electronic holography, which takes advantage of available arrays of detectors with multimillion pixels, this approach is rapidly growing.

Real-time 3-D imaging systems based on photorefractive holography in multi-quantum-wells devices, which can potentially acquire depth-resolved images at thousands of frames per second, have been demonstrated.

These ballistic light techniques can extend the depth of tissue imaging to the mm range when

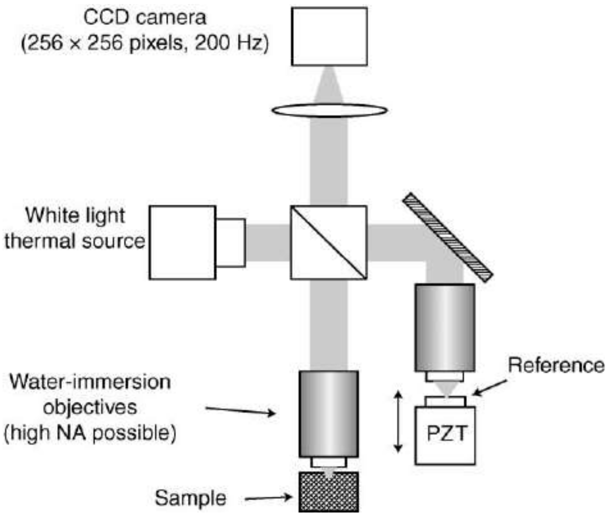

Figure 5 The full-field OCT setup based on a Linnik interferometer. The PZT actuator is used to modulate the path difference between the two arms.

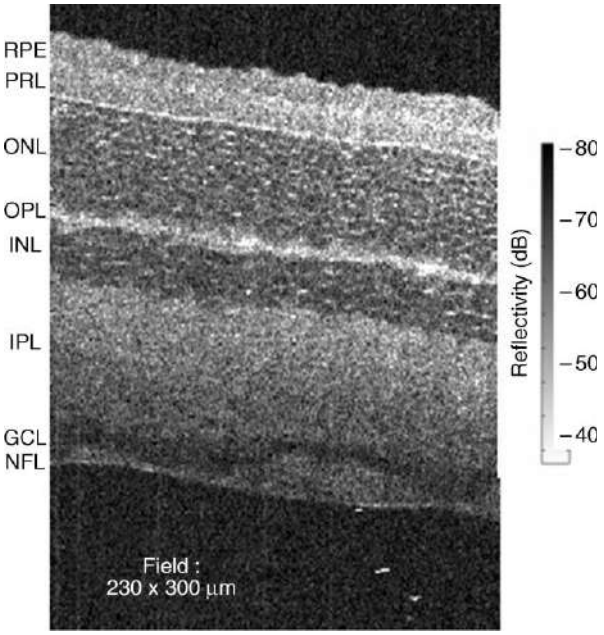

Figure 6 OCT rat retinal image: retinal pigmented epithelium. PRL: photoreceptor layer. ONL: outer nuclear layer. OPL: outer plexiform layer. INL: inner nuclear layer. IPL: inner plexiform layer. GCL: ganglion cell layer. NFL: nerve fiber layer.

imaging in reflection. When it is necessary to image through cms, rather than mms of tissue, however, there will be no detectable ballistic light signal and one must extract useful information from the scattered photons.

Imaging in the Snake-Like Regime: Taking Advantage of Time (or Frequency) Gating

The first approach is to use time gating to select the earliest arriving scattered light, which will provide the most useful image information.

For moderate tissue depths one can exploit the fact that biological tissue is highly forward scattering, with most photons being only slightly deviated from their original direction upon each scattering event. After a few cm of tissue, there can be a significant number of photons that have followed reasonably well-defined ‘snake-like’ paths about the original direction through the tissue, arriving at the detector after the ballistic light but before the fully diffuse photons whose trajectories are effectively randomized. This early arriving light can be selectively gated using ultrafast lasers and a variety of detectors such as streak cameras, time-gated optical image intensifiers, and time-gated fast photon counting systems, which provide picoseconds time gates.

An alternative, sometimes cheaper, approach to temporal gating is the use of (high) frequency signal discrimination (which is the Fourier transform of the pulse experiment). The periodic solution of the diffusion equation with a periodic source term gives rise to the so-called ‘photon density waves’. Such diffusive waves, like thermal waves, are used in the near-field regime (their amplitudes are damped by $\exp(-2\pi)$ after propagating over a single wavelength path) and they can reveal subsurface structures through their refractive and diffusive behavior. Coherence gating and polarization gating can be used when the early arriving light still retains some degree of coherence of the incident light. Indeed polarization discrimination has been used to select those photons which undergo few scattering events and therefore preserve a fraction of their initial polarization state, from the photons which experience multiple scattering resulting in a complete loss of memory of their initial polarization state.

As an example, Figure 7 shows the image of sub-mm light sources (LEDs) taken through 1 cm of polystyrene solution ($\mu_s' \sim \mu_s \sim 10\ \text{cm}^{-1}$) by selecting photons which kept the memory of their initial circular polarization (here a few less than 1%). One can see the sources quite clearly.

For many important biomedical applications, such as mammography or imaging brain function, it is necessary to penetrate through 5–10 cm of tissue, after which all the detected signal is diffuse. Detecting earlier arriving photons can provide more information but, if the time window is too narrow, the detected signal becomes too weak to use with acceptable data acquisition times.

This problem is, however, being addressed with some success using statistical models of photon transport, with different degrees of approximation ranging from full Monte Carlo simulation of photon propagation to the diffusion equation. The approach is to address the inverse problem, i.e. to measure the scattered light signal as comprehensively as is practical and to calculate what distribution of absorption and scattering properties would have produced the measured signal. This provides a means to quantify the optical properties of biological tissue, averaged over a volume corresponding to a particular distribution of photon paths, and to form relatively low-resolution tomographic images.

Multiplying the number of sources and detectors and correlating the various intensity distributions over the whole structure under examination gives, as expected, better precision in localizing the size and position of the local structure to be identified.

Despite the multiple input and multiple output approaches, real-life situations with the previously

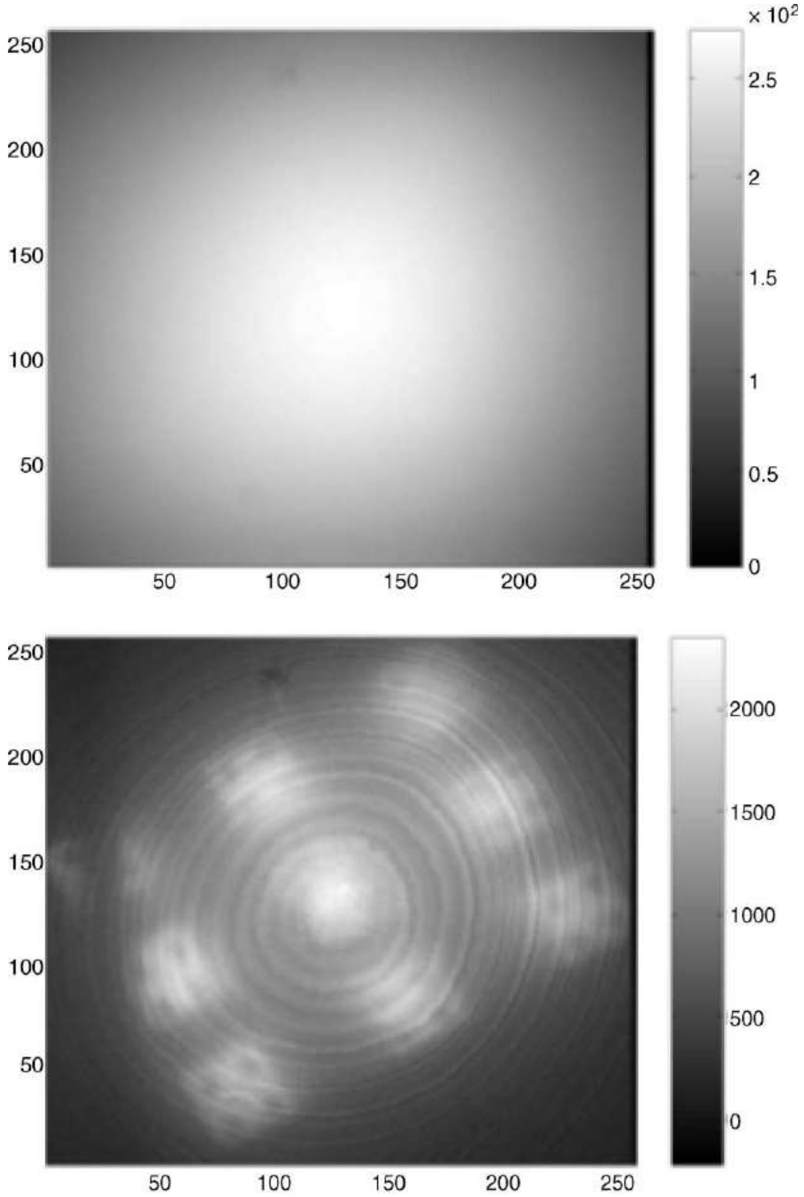

Figure 7 Polarization imaging: here the degree of circular polarization is measured in the upper image. The lower image shows the field without polarization filtering.

mentioned problem of the inhomogeneity of the optical properties of human tissues means that the results are still poor in term of resolution and subcentimeter structures can hardly be observed in breast imaging, for instance.

One important exception to this limited success in finding precisely the position, the size, and the optical properties of hidden structures is the case of 'activation' studies. This corresponds to a 'differential' situation in which an unknown structure is locally changing with time. Here the goal is not to reach a full description of the unknown structure but only to characterize the local changes induced by the activation.

The main field of application of this approach is certainly in brain activation studies. A matrix of light emitters (usually at two wavelengths which differentiate oxy- and deoxy-hemoglobin, for instance 780 and 840 nm) is coupled to a matrix of light detectors. The signals are usually balanced and any local change which breaks the symmetry of the scattered photon paths induces a differential signal, easy to detect, because there is no background signal before activation takes place.

Moreover it has been demonstrated that for weak perturbations (weak variations of the optical parameters), the inverse problem can be rigorously solved. This geometrical selection provides a cheap and simple setup with rather spectacular results. Sometimes a modulation is added and the phase shift of the photon density wave is enhanced by the differential approach.

Imaging in the Multiple Scattering Regime: Taking Advantage of all the Photons

Light Only: Transillumination and Fluorescence Imaging

The optical systems used for CW measurements are rather simple and require only the alignment of a number of light sources and detectors. Compared to time-gated techniques due to multiple scattering a much stronger path length distribution is present in the signal and the result is a loss of localization and resolution. In the so-called transillumination geometry, collimated detection is used to isolate multiply scattered photons close to the normal emergence angle.

Direct imaging with the required resolution using CW techniques in very thick tissues, such as breast tissue, has not been established even with sophisticated multi-sources/multi-detectors combinations (typically a few tens of sources and detectors at three different wavelengths).

Despite rapid progress in the modeling of light behavior in scattering media it has not yet been demonstrated that CW imaging can provide images with suitable resolution in thick tissue and time-dependent measurement techniques discussed earlier are dominantly used.

For fluorescence imaging, one has to account for the strong attenuation of light as it passes through tissue. Fluorescent dyes (porphyrin, indo-cyanine, and more recently quantum dots) have been developed that are excited and re-emit in the ‘biological window’ at near-infrared (NIR) wavelengths, where scattering and absorption coefficients are lower, which is favorable for deep fluorescence imaging. Fluorescent intensity in deep tissues depends on the location, size, concentration, and intrinsic characteristics (e.g., lifetime, quantum efficiency) of the fluorophores, and on the scattering and absorption coefficients of the tissue at both the excitation and emission wavelengths. In order to extract intrinsic characteristics of fluorophores within tissue, it is necessary to describe the statistics of photon path lengths which depend on all these different parameters. The amount of light emitted at the site of the fluorophore is directly proportional to the total amount of light impinging on the fluorophore.

Using fluorescence contrast has the unique feature of carrying two spectroscopic contrasts related to both excitation and fluorescent photons.

The next two techniques take advantage of another way to handle the optical information by coupling optics and acoustics.

Coupling Light and Sound: Acousto-Optic and Photo-Acoustic

Although the two approaches that we will now describe are different in their basic principles as well as in the experimental setups, their names are similar – opto-acoustics (sometimes called photo-acoustics) and acousto-optics.

Opto- (or Photo-)Acoustics

In this technique a pulsed (or modulated) electromagnetic beam irradiates the structure under examination. Despite their tortuous paths (about 10 to 100 cm) photons propagate through the volume in a few nanoseconds. Local absorbing centers then absorb the electromagnetic energy, they experience a fast thermal expansion, and become sources of acoustic waves.

The optical problem of tomographic reconstruction is then transformed into a simpler problem of acoustic reconstruction of source distribution because the speed of the sound is much slower than the speed of the light. More precisely we can compute the signal generated by a short light pulse impinging on a scattering medium which contains an absorber (thickness e) as can be seen on Figure 8, when $e\mu_a \ll 1$.

The amount of power per unit surface which is deposited in the absorbing slice is $E\mu_a$ (E is the fluence in W/m^2). When the light pulse duration τ is short enough, despite the propagation in the scattering medium ($\tau \ll e/v$ where v is the speed of the sound),

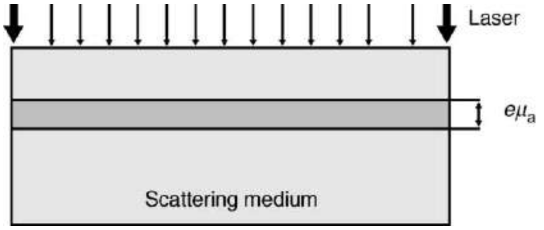

Figure 8 Photo acoustic signal generation: 1-D model.

we find a variation of the mass by volume unit:

$$\Delta\rho \sim \rho\beta\Delta T \sim \rho\frac{\beta E\mu_a e}{C_v e} \qquad [11]$$

The corresponding pressure is:

$$\Delta p = \frac{v^2}{\gamma}\Delta\rho \qquad \text{or} \qquad \Delta p \sim \rho\frac{v^2\beta E}{C_p}\mu_a \qquad [12]$$

Although this approach is new, compared to the purely optical transillumination approach, it has led to a large number of studies in which time gating brings new information to the accurate determination of the US emitter depth. It has been applied to breast tumor examination to complement standard imaging techniques such as echography and X-ray mammography. Recently this technique has demonstrated submillimeter resolution for *in vivo* blood distribution in rat brain imaging.

Acousto-Optic

Here a DC single frequency laser is used to illuminate the sample. For a typical biological sample a few cm thick the trajectories are distributed between the sample thickness value and more than ten times this value. Due to the good temporal coherence of the source all the scattered photons are likely to interfere when they emerge from the sample volume. These interferences are seen as a speckle field which can be observed at a suitable distance from the sample surface.

As seen in Figure 9, adding an ultrasonic field will mainly induce a small (smaller than the optical wavelength) displacement of the scatterers (it also induces local variation of the refractive index) which will cause a speckle modulation at the ultrasonic frequency.

If the zone irradiated by the ultrasonic field is optically absorbing, less photons will emerge from this zone than from a nonabsorbing one and thus less modulation will be seen when scanning across the sample volume. This modulation can be detected by selecting a single speckle grain and averaging over a number of uncorrelated speckle distributions to obtain a suitable average value (for instance with latex particles). This is not always possible, in particular because the overall geometry is rather stable in a semisolid tissue. Using a single detector which receives many speckle grains gives less contrast. A multiple detector and a parallel processing of many speckle grains improve greatly the speed and the signal-to-noise ratio: a typical image of absorbing spheres, embedded in a scattering phantom, is shown in Figure 10.

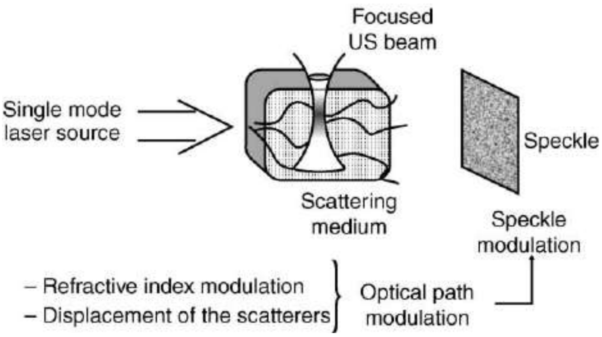

Figure 9 Acousto-optic imaging principle: photons are tagged by the acoustic field.

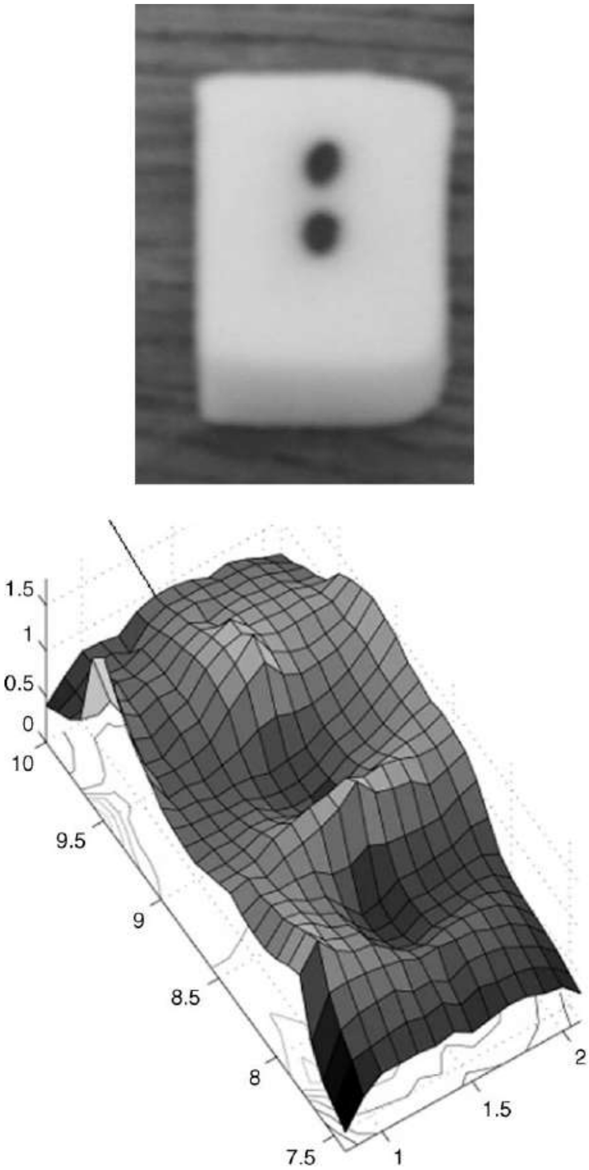

Figure 10 Acousto-optic imaging (lower image) of two absorbing spheres embedded in a scattering medium (upper image shows a section of the gel sample).

Conclusion

The complexity associated with multiple scattering in turbid media, such as tissues of the human body, prevents an easy use of light as a routine tool for in-depth examination, whereas superficial examinations have always been sources of diagnostics.

Optical techniques will obviously progress to better characterization of tissues' optical properties, improvement of laser properties and detection techniques, as well as new mathematical approaches of the inverse problem.

We believe that this field of research is still open to new ideas, and new experimental schemes leading to new breakthroughs in the delicate problem of using light in highly scattering media.

Acknowledgements

I would like to warmly thank all the research staff (Arnaud Dubois, Benoit Forget, François Ramaz, Vincent Loriette) and PhD students (Michael Atlan, Kate Grieve, Gael Moneron, Juliette Selb) at ESPCI whose results have been used to illustrate this article. I also acknowledge discussions with my colleagues Pr Paul French (Imperial College London) and Dr Amir Gandbjbakhche (NIH Bethesda).

See also

Coherence: Coherence and Imaging. **Holography, Techniques:** Digital Holography. **Imaging:** Interferometric Imaging. **Interferometry:** White Light Interferometry. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy. **Tomography:** Optical Coherence Tomography; Tomography and Optical Imaging.

Further Reading

Boccara AC and Fink M (eds) (2001) *Optical and Acoustical Imaging of Biological Media*. Compte Rendu de l'Académie des Sciences, Physics and Astrophysics série IV tome 2, no. 8. Paris: Elsevier.

Born M and Wolf E (1997) *Principles of Optics*. New York: Oxford University Press.

Goodman JW (2000) *Statistical Optics*. New York: Wiley-Interscience.

Hecht E and Zajac A (1979) *Optics*. Reading, MA: Addison-Wesley.

Sebbah P (ed.) (2001) *Waves and Imaging through Complex Media*. Dordrecht: Kluwer Academic Publishers.

Van de Hulst HC (1981) *Light Scattering by Small Particles*. New York: Dover.

Van Tiggelen B and Skipetrov S (eds) (2003) *Wave Scattering in Complex Media: From Theory to Applications*. Dordrecht: Kluwer Academic Publishers.

Infrared Imaging

K Krapels and R G Driggers, US Army Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, USA

Published by Elsevier Ltd.

Introduction

All objects above 0 Kelvin emit electromagnetic radiation associated with the thermal activity on the surface of the object. For terrestrial temperatures (around 300 Kelvin), objects emit a good portion of the electromagnetic flux in the infrared part of the electromagnetic spectrum. The visible band spans wavelengths from 0.4 to 0.7 micrometers (infrared engineers typically use micrometer/μm for wavelength rather than the nanometer or wavenumber more commonly used in other fields). Infrared imaging devices convert energy in the infrared portion of the electromagnetic spectrum into images that can be created by the human eye. The unaided human eye cannot image infrared energy because the lens of the eye is opaque to infrared radiation.

The infrared spectrum begins at the red end of the visible spectrum where the eye can no longer sense energy. It spans from 0.7 μm to 100 μm. The infrared spectrum is, by common convention, broken into five different bands (this may vary according to the application). The bands are typically defined as: near infrared (NIR) from 0.7 to 1.0 μm; short wavelength infrared (SWIR) from 1.0 to 3.0 μm; mid-wavelength infrared (MWIR) from 3.0 to 5.0 μm; long wavelength infrared (LWIR) from 8.0 to 14.0 μm; and far infrared (FIR) from 14.0 to 100 μm. These bands are depicted graphically in Figure 1, which shows the atmospheric transmission for a 1-kilometer horizontal ground path for a 'standard' day in the United States. These types of transmission graphs can be tailored for any conditions, using sophisticated atmospheric models such as MODTRAN (from www.ontar.com). Note that

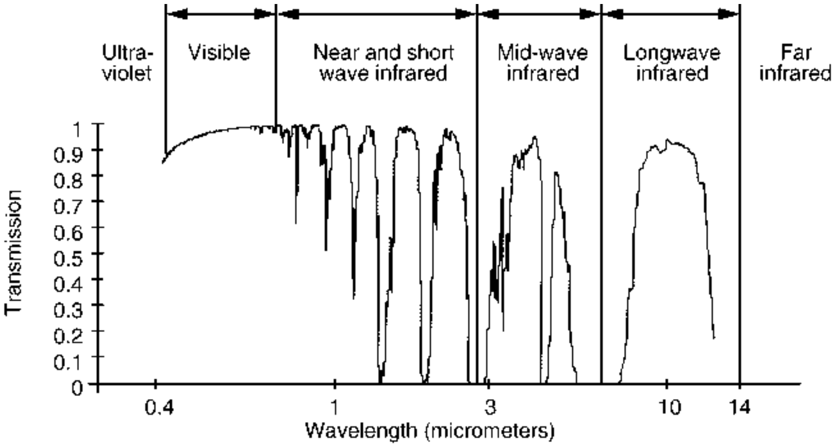

Figure 1 Visible and infrared spectra.

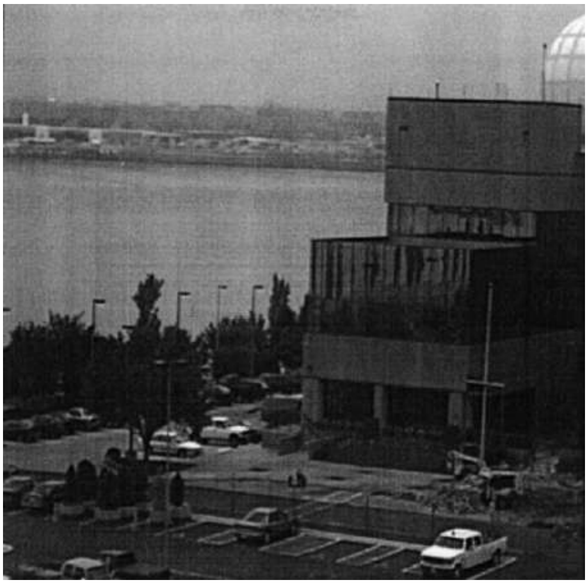

Figure 2 Visible image–reflected energy.

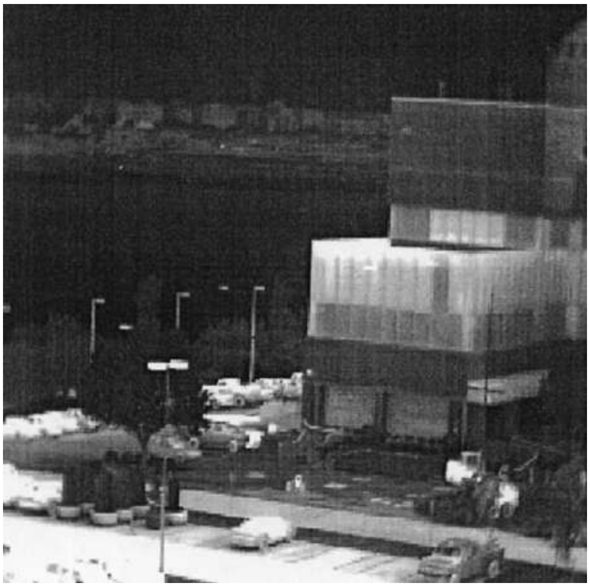

Figure 3 Infrared image–self-emitted energy.

there are many atmospheric 'windows', such that an imager designed with such a band selection can see through the atmosphere.

The primary difference between a visible spectrum camera and an infrared imager is the physical phenomonology of the radiation from the scene being imaged. The energy used by a visible camera is predominantly reflected solar, or some other illuminating energy, in the visible spectrum. The energy imaged by infrared imagers (Forward Looking Infrared, commonly known as FLIRs) in the MWIR and LWIR bands, is primarily self-emitted radiation. From Figure 1, the MWIR band has an atmospheric window in the 3 to 5 μm region and the LWIR band has an atmospheric window in the 8 to 12 μm region. The atmosphere is opaque in the 5 to 8 μm region, so it would be pointless to construct a camera that responds to this waveband.

Figures 2 and 3 are images that show the difference in the source of the radiation sensed by the two types of cameras. The visible image (Figure 2) is light that was provided by the Sun, propagated through Earth's atmosphere, reflected off the objects in the scene, traversed through a second atmospheric path to the sensor, and then imaged with a lens and a visible band detector. A key here is that the objects in the scene are represented by their reflectivity characteristics. The image characteristics can also change by any change in atmospheric path or source characteristic change. The atmospheric path characteristics from the Sun to the objects change frequently, because the Sun angle changes throughout the day and also weather and cloud conditions change. The visible imager characterization model is therefore a multipath problem that is extremely difficult.

The LWIR image given in Figure 3 is obtained primarily by the emission of radiation by objects in the scene. The amount of electromagnetic flux depends on the temperature and emissivity of the objects. A higher temperature and a higher emissivity correspond to a higher flux. The image shown is 'white-hot' or the whiter a point in the image the higher the flux leaving the object. It is interesting to note that trees have a natural self-cooling process since a high temperature can damage foliage. Objects that have absorbed a large amount of solar energy are hot and are emitting large amounts of infrared radiation. This is sometimes called 'solar loading'.

The characteristics of the infrared radiation emitted by an object are described by Planck's black-body law in terms of spectral radiant emittance

$$M_\lambda = \varepsilon(\lambda)\frac{c_1}{\lambda^5(e^{c_2/\lambda T}-1)}\ \mathrm{W/cm^2\ \mu m} \qquad [1]$$

where c_1 and c_2 are constants of 3.7418×10^4 W μm^4/cm^2 and 1.4388×10^4 μm K. The wavelength, λ, is provided in micrometers and $\varepsilon(\lambda)$ is the emissivity of the surface. A black-body source is defined as an object with an emissivity of 1.0, so it is a perfect emitter. Source emissions of black-bodies at typical terrestrial temperatures are shown in Figure 4. Often, in modeling, the terrestrial background temperature is assumed to be 300 K. The source emittance curves are shown for other temperatures for comparison purposes. One curve corresponds to an object colder than the background and two curves correspond to temperatures hotter than the background.

Planck's equation describes the spectral shape of the source as a function of wavelength. It is readily apparent that the peak shifts to the left (shorter wavelengths) as the body temperature increases. If the temperature of a black-body is increased to that of the Sun (5900 Kelvin), the peak of the spectral shape would decrease to 0.55 μm or green light. This peak wavelength is described by

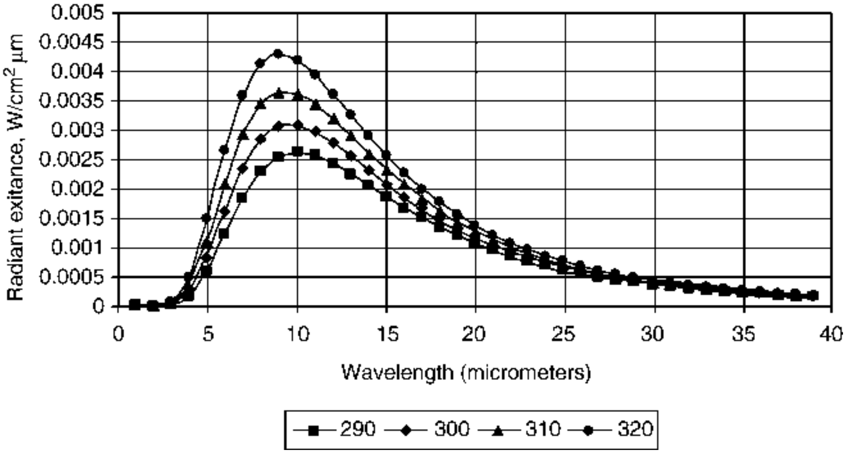

Figure 4 Planck's blackbody radiation curves for four temperatures from 290 K to 320 K.

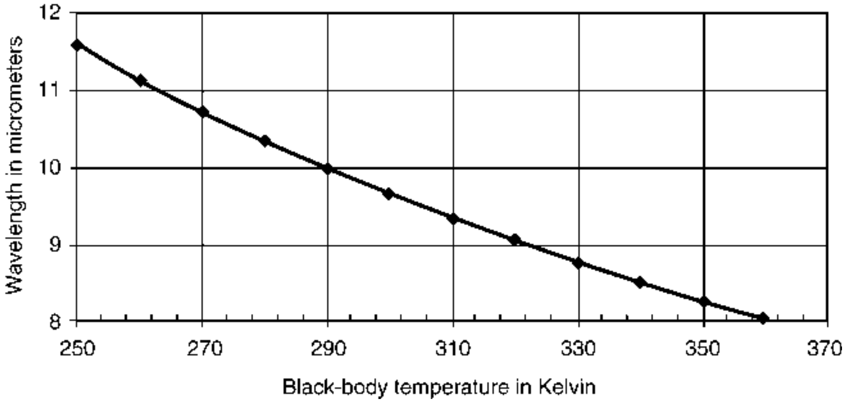

Figure 5 Location of peak of black-body radiation, Wien's Law.

Wien's displacement law

$$\lambda_{\max} = 2898/T, \ \mu\text{m} \qquad [2]$$

Figure 5 shows the peak wavelength as a function of temperature through the longwave region. It is important to note that the difference between the black-body curves is the 'signal' in the infrared bands. For an infrared sensor, if the background is at 300 K and the target is at 302 K, the signal is the difference in flux between these curves. Signals in the infrared ride on very large amounts of background flux. However, in the visible band this is not the case. For example, consider the case of a white target on a black background. The black background is generating no signal, whereas the white target is generating a maximum signal (given that the sensor gain has been adjusted). Dynamic range may be fully utilized in a visible sensor. In the case of an IR sensor, a portion of the dynamic range is used by the large background flux radiated by everything in the scene. This flux never has a small value, hence sensitivity and dynamic range requirements are much more difficult to satisfy in IR sensors than in visible sensors. Table 1 summarizes this characteristic numerically for the 300 K background and targets shown in Figure 4.

A typical infrared imaging system scenario is depicted in Figure 6. The scene consists of two major components, the target and the background. In an IR scene, the majority of the energy is remitted from the constituents of the scene. This emitted energy is transmitted through the atmosphere to the sensor. As it propagates through the atmosphere it is degraded by absorption and scattering. Obscuration by intervening objects and additional energy emitted by the path also affect the target energy. All of these contributors, which are not the target of interest, essentially reduce one's ability to discriminate the target. So, at the entrance aperture for the sensor, the target signature is reduced from the values observed at the target. Then, the signal is degraded by the optics and scanner (as applicable) of the sensor. The energy is then sampled by the detector array and converted to electrical signals. Various electronics amplify and condition this signal before it is presented to either a display for human interpretation or an algorithm like an automatic target recognizer for machine interpretation. A linear systems approach to modeling allows the components' transfer functions to be treated separately as contributors to the overall system performance. This approach allows for straightforward modifications to a performance model for changes in the sensor or environment when performing trade-off analyses.

Table 1 Signal/dynamic range limitation in IR bands

Blackbody temperature	*Exitance in 8–12 μm band (w/cm^{-2})*	*Contrast*
290 ($-10\Delta T$ K target)	1.25E − 02	−8.48%
300 (background)	1.48E − 02	
310 ($+10\Delta T$ target)	1.74E − 02	7.97%
320 ($+20\Delta T$ target)	2.02E − 02	15.39%

History of Infrared Imaging

Night vision systems began to be developed extensively during World War II. Satisfying the requirement to image at night was approached with two different methodologies. The first method was image intensification (I^2). This method involves amplification of any small light that was available and displaying it directly to the eye. Typically I^2 devices are sensitive to visible light and a small portion of the short-wave IR (SWIR, 0.7–3.0 μm) band. They are often classified (along with visible band imagers) as electro-optical (EO) imagers. With an I^2 imager, there must be some source of illumination for them to function well (as little as starlight is sufficient

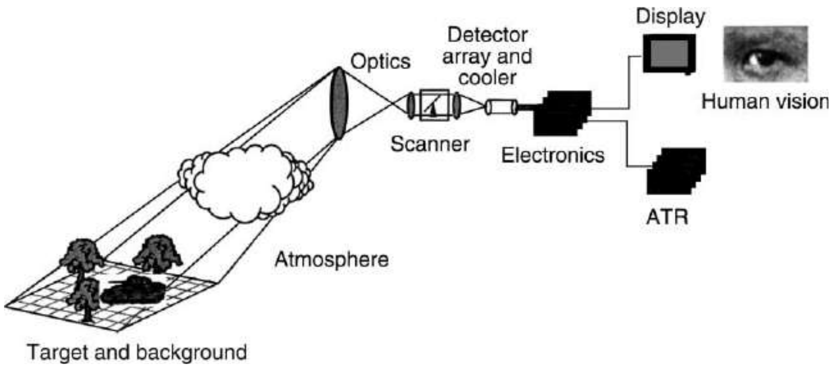

Figure 6 Typical infrared imaging scenario.

Table 2 Applications of infrared imaging

Category	*User*	*Application*
Military	Intelligence analyst	Intelligence surveillance and reconnaissance
	Pilot, driver	Navigation/pilotage
	Gunner, weapons officer	Targeting/fire control
	Air defense	Search and track
Commercial	*Civil government*	
	Police	Surveillance, fugitive tracking
	Fire	Rescue, hot-spot location
	Environmental	
	EPA	Emission tracking
	Interior department	Resource management
	NIMA\USGS	Mapping
	NOAA\MWS	Weather forecasting
	Industrial	
	Manufacturers	Machine vision
	Maintenance	Non-destructive testing
	Medical	
	Doctors	Diagnostic imaging

for approx 20/40 acuity with current systems). The second type of night vision device is the infrared or thermal imager (also known as FLIRs). FLIRs are sensitive to the radiation in either the mid-wave IR (MWIR, 3–5 μm) or long-wave IR (LWIR, 8–14 μm) bands.

The first infrared cameras used photographic film that was sensitive to infrared wavelengths. Intelligence and reconnaissance imaging from aircraft drove the infrared system development as a night augmentation to existing visible cameras. These imagers used either continuous moving film or stationary film, like an ordinary camera. A slit aperture was used in the continuous moving film cameras and aircraft motion provided the scan of the scene along the film to build a continuous image. The introduction of television, as electronic imaging in the visible spectrum, led to the development of electronic imagers in the infrared band. Electronic infrared imagers were not restricted to looking down and using aircraft motion for scan, hence these devices came to be known as FLIRs (Forward Looking InfraRed). Today, an electronic infrared imager may be called a FLIR or a thermal imager, with the trend being to use FLIR to describe military applications.

Early FLIRs were often accompanied by illumination devices. They were not very sensitive compared to modern FLIRs so active illumination was necessary. Infrared searchlights or illuminators were used to get higher signal-to-noise ratios. These types of FLIRs are called 'active' IR imagers, while FLIRs that do not use illuminators are 'passive' imagers. Generally, active imagers are used in civil applications and are declining in military use. Broadly, applications for thermal imagers fall into two categories; military and commercial. Table 2 lists some of the purposes and users for modern thermal imagers. The design and performance criteria vary widely for some of the applications.

Types of Infrared Imagers

Infrared imagers are classified by different characteristics: scan type, detector material/cooling requirements, and detector physics. The scan type refers to the method used to construct the electronic image. The camera may use a single detector which is raster scanned over the input scene to build an image. Alternatively, a parallel scan uses a linear array of detectors scanned across a scene to build an image. The latest advances in materials have led to staring arrays of detectors. In a staring system, a detector is present in two dimensions to represent each image pixel and no mechanical motion of the focal plane is necessary to construct the image. There are some hybrid FLIR types that combine the different imaging techniques. Usually this combination leads to an improvement in signal-to-noise ratio or to reduce undersampling.

The second classification, by detector material/cooling requirements, usually describes FLIRs built using differing materials. For example, a typical LWIR FLIR material is mercury cadmium telluride (HgCdTe or MCT). In order to achieve high sensitivity, these devices are cooled to decrease dark current. Usually the cooling is based on liquid nitrogen or a cryogenic cooler, and the detectors operate at 77 K. Another common detector material is indium antimonide (InSb), which is used for MWIR FLIRs and is also cooled. A newer class of infrared cameras is not cooled, being referred to as 'uncooled' FLIRs. These uncooled FLIRs are microbolometers (resistive elements) or pyrometers (capacitive elements) and have less sensitivity than cooled imagers. Typically, the cooled imagers are comprised of photon detectors while the uncooled imagers are based on thermal detectors, the uncooled FLIRs being used in lower-performance applications.

Infrared Imager Performance

There are three general categories of infrared sensor performance characterizations. The first is sensitivity and the second is resolution. When end-to-end, or human-in-the-loop (HITL), performance is required, the third type of performance characterization describes the visual acuity of an observer through a sensor. The former two are both related to the hardware and software that comprises the system, while the latter includes both the sensor and the observer. Sensitivity is determined through radiometric analysis of the scene/environment and the quantum electronic properties of the detectors. Resolution is determined by analysis of the physical optical properties, the detector array geometry, and other degrading components of the system in much the same manner as complex electronic circuit/signals analysis.

Sensitivity describes how the sensor performs with respect to input signal level. It relates noise characteristics, responsivity of the detector, light gathering of the optics, and the dynamic range/quantization of the sensor. Radiometry describes how much light leaves the object and background and is collected by the detector. Optical design and detector characteristics are of considerable importance in sensor sensitivity analysis. In infrared systems, noise equivalent temperature difference (NETD) is often a first-order description of the system sensitivity. The 3D noise model describes more detailed representations of sensitivity parameters (see Further Reading).

The second type of measure is resolution. Resolution is the ability of the sensor to image small targets and to resolve fine detail in large targets. Modulation transfer function (MTF) is the most widely used resolution descriptor in infrared systems. Alternatively, it may be specified by a number of descriptive metrics such as the optical Rayleigh criterion or the instantaneous field-of-view of the detector. Where these metrics are component-level descriptions, the system MTF is an all-encompassing function that describes the system resolution.

Sensitivity and resolution can be competing system characteristics and they are the most important issues in initial studies for a design. For example, given a fixed sensor aperture diameter, an increase in focal length can provide an increase in resolution, but may decrease sensitivity. Typically, visible band systems have plenty of sensitivity and are resolution-limited, whereas infrared imagers have been more sensitivity-limited. With staring infrared sensors, the sensitivity has seen significant improvements.

Often metrics, such as NETD and MTF, are considered to be separable. However, in an actual sensor, sensitivity and resolution performance are not independent. As a result, minimum resolvable temperature difference (MRT or MRTD) has become the primary performance metric for infrared systems. MRT is a quantitative performance measure in terms of both sensitivity and resolution and a simple MRT curve is shown in Figure 7. The performance is bounded by the sensor's limits and the observer's limits. The temperature difference (or thermal contrast) required to image smaller details in a scene increases with detail size. The inclusion of observer performance yields a single sensor performance characterization, MRT. It describes sensitivity as a function of resolution, and includes the human visual system.

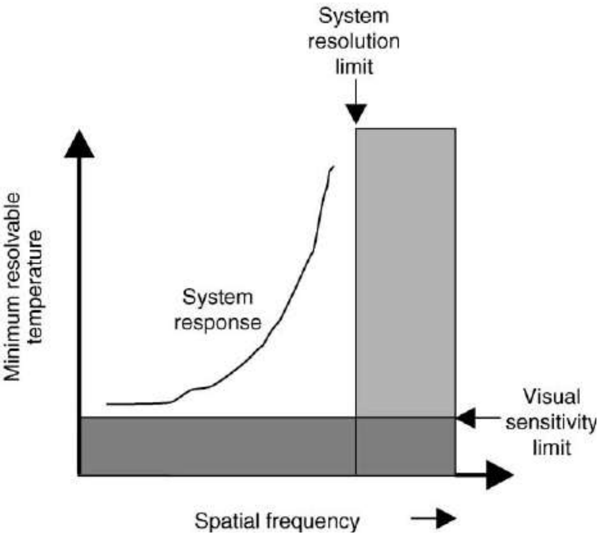

Figure 7 IR imager performance.

Many different sensor characteristics may be considered to increase the fidelity of a sensor model. A list of typical infrared imaging system parameters is given in Figure 8. When considering performance in the system sense, the groupings are often blocks in the system diagram, each with separate transfer functions. This approach works well unless the linear shift invariance assumption is not valid.

General Characteristics of Infrared Imagers

There are a number of imaging characteristics that make infrared systems a little different than conventional visible imaging systems. These include source flux levels, detector charge storage and sensitivity, detector size, diffraction blur, sampling, and uniformity characteristics.

First, we consider the source flux levels. The daytime and night-time flux levels (in photons per second per square centimeter) on Earth in the visible (0.3 to 0.7 μm) is 1.5×10^{17} and around 1×10^{10}, respectively. In the mid-wave (3 to 5 μm), the daytime and night-time flux levels are 4×10^{15} and 2×10^{15} μm where the flux is a combination of

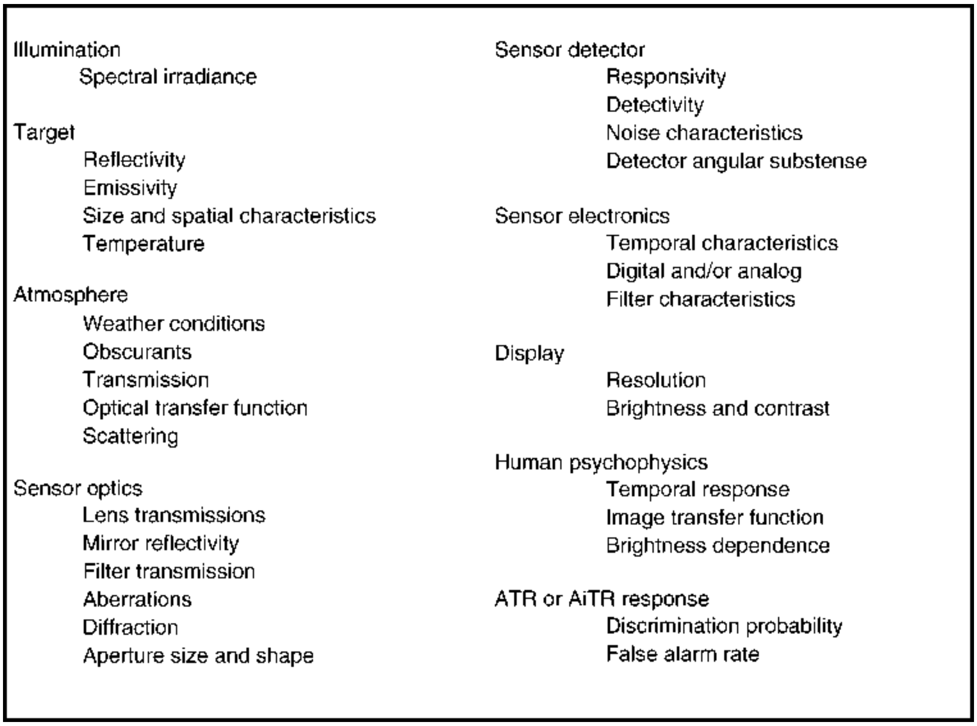

- Illumination
  - Spectral irradiance
- Target
  - Reflectivity
  - Emissivity
  - Size and spatial characteristics
  - Temperature
- Atmosphere
  - Weather conditions
  - Obscurants
  - Transmission
  - Optical transfer function
  - Scattering
- Sensor optics
  - Lens transmissions
  - Mirror reflectivity
  - Filter transmission
  - Aberrations
  - Diffraction
  - Aperture size and shape
- Sensor detector
  - Responsivity
  - Detectivity
  - Noise characteristics
  - Detector angular substense
- Sensor electronics
  - Temporal characteristics
  - Digital and/or analog
  - Filter characteristics
- Display
  - Resolution
  - Brightness and contrast
- Human psychophysics
  - Temporal response
  - Image transfer function
  - Brightness dependence
- ATR or AiTR response
  - Discrimination probability
  - False alarm rate

Figure 8 Factors in infrared sensor modeling.

emitted and solar reflected flux. In the longwave, the flux is primarily emitted where both day and night yield a 2×10^{17} μm frequency, the AMOP drops rapidly to zero. The effect is to extend the prediction of the MRT beyond the half sample rate and introduce a parameter called minimum temperature difference perceived (MTDP) to replace MRT in the range performance estimate and also in the lab to characterize an undersampled system.

The concept of MTDP follows from observation of standard 4-bar images as the pattern frequency passes the imager half sample rate. The image changes from four bars to three, two, and finally one (can be perceived). The phase of the target image on the detector array may need to be adjusted to observe this progression. The standard MRT measurement requires that all four bars be resolvable by the observer during the measurement process. For use in the lab, the MTDP is defined as the minimum temperature difference at which two, three, or four bars can be resolved in the image of the standard 4-bar test pattern by an observer, with the test pattern positioned at optimum phase. The optimum phase is the phase at which two, three, or four bars are 'best perceived'. TRM3 uses the standard definition of MRT as for adequately sampled imagers. The TRM3 Approach Model is implemented if the imager is considered undersampled as defined when the prefilter MTF at the half-sample frequency is larger than 10%. The MTDP equation is given by

$$\mathrm{MTDP}(f) = \frac{\frac{\pi}{2}\mathrm{SNR}_{\mathrm{th}}\Psi(f)}{\mathrm{AMOP}(f)} \qquad [3]$$

where $\mathrm{SNR}_{\mathrm{th}}$ is a threshold signal-to-noise ratio and Ψ is the total system filtered noise. MTDP is used in the same manner as MRTD in range calculations.

Another approach is the triangle orientation discrimination (TOD) threshold. In the TOD, the test pattern is a (nonperiodic) equilateral triangle in four possible orientations (apex up, down, left, or right), and the measurement procedure is a robust 4AFC (four alternative forced-choice) psychophysical procedure. In this procedure, the observer has to indicate which triangle orientation he sees, even if he is not sure. Variation of triangle contrast and/or size leads to a variation in the percentage correct between 25% (complete guess) and 100%, and by interpolation the exact 75% correct threshold can be obtained. A complete TOD curve (comparable to an MRTD curve) is obtained by plotting the contrast thresholds as a function of the reciprocal of the triangle angular size (Figure 9).

The TOD method has a large number of theoretical and practical advantages: it is suitable for under-sampled and well-sampled electro-optical and optical

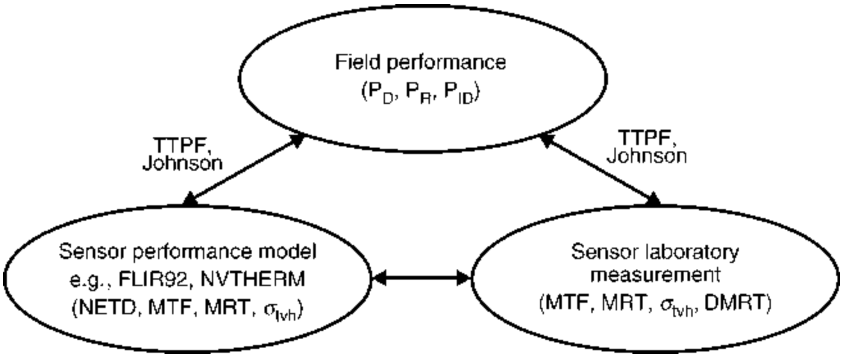

Figure 9 Modeling, testing and performance triangle.

imaging systems in both the thermal and visual domains, it has a close relationship to real target acquisition, and the observer task is easy. The results are free from observer bias and allow statistical significance tests. The method may be implemented in current MRTD test equipment with little effort, and the TOD curve can be used easily in a target acquisition (TA) model such as ACQUIRE. Two validation studies with real targets show that the TOD curve predicts TA performance for under-sampled and well-sampled imaging systems very well. Currently, a theoretical sensor model to predict the TOD (comparable to NVTherm or TRM3) is under development. The lab measurement and field performance appear sound, but the model is not yet available.

Testing Infrared Imagers: NETD, MTF, and MRTD (Sensitivity, Resolution and Acuity)

In general, imager sensitivity is a measure of the smallest signal that is detectable by a sensor. Sensitivity is determined using the principles of radiometry and the characteristics of the detector. For infrared imaging systems, noise equivalent temperature difference (NETD) is a measure of sensitivity. The system intensity transfer function (SITF) can be used to estimate the NETD, which is the system noise rms voltage over the noise differential output. NETD is the smallest measurable signal produced by a large target (extended source).

Equation [4] describes NETD as a function of noise voltage and the system intensity transfer function. The measured values are determined from a line of video stripped from the image of a test target, as depicted in Figure 10. A square test target is placed before a black-body source. The delta T is the difference between the black-body temperature and the mask. This target is then placed at the focal point of an off-axis parabolic mirror. The mirror serves the purpose of a long optical path length to the target, yet relieves the tester from concerns over atmospheric losses to the temperature difference. The image of the target is shown in Figure 10b. The SITF slope for the scan line in Figure 10c is the $\Delta S/\Delta T$ where ΔS is the signal measured for a given ΔT. The N_{rms} is the background signal on the same line.

$$\mathrm{NETD} = \frac{N_{\mathrm{rms}}[\mathrm{volts}]}{\mathrm{SITF\_Slope}[\mathrm{volts/K}]} \qquad [4]$$

Resolution is a general term that describes the size of resolvable features in an imager's output. With infrared imaging systems, the resolution is described by the system modulation transfer function (MTF). Consider Figure 11 for the determination of MTF, from either point spread function or edge spread function (psf or esf). In Figure 11a, a test target is placed at the focal point of a collimator. The target may be a point, edge, or line as shown in Figure 11b. The IR system images this target. The output of a detector scan, or a line of detectors in the staring array case, is taken for analysis, in Figure 11c. For the edge spread function, the spatial derivative is taken to get the psf. The Fourier transform is then calculated to give the system MTF as shown in Figure 11d. The point spread function is really the impulse response of the system, so a smaller psf is desirable, where a wide MTF is desirable. Such a psf/MTF combination gives a higher resolution.

The MRTD of an infrared system is defined as the differential temperature of a four bar target that makes the target just resolvable by a particular observer. It is a measure of observer visual acuity when a typical observer is using the infrared imager. It results in a descriptor, which is a function not just of a single value. It is a plot of sensitivity as a function of spatial frequency.

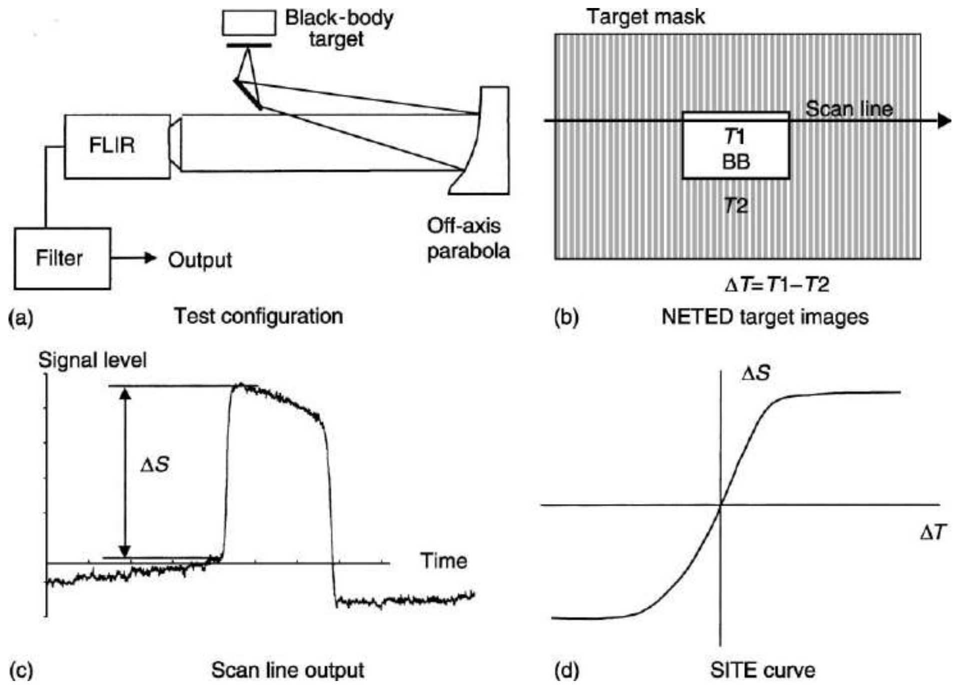
Black-body target
FLIR
Filter
Output
Off-axis parabola
(a)
Test configuration
Target mask
Scan line
$T1$
BB
$T2$
$\Delta T = T1 - T2$
(b)
NETED target images
Signal level
ΔS
Time
(c)
Scan line output
ΔS
ΔT
(d)
SITE curve

Figure 10 NETD measurement.

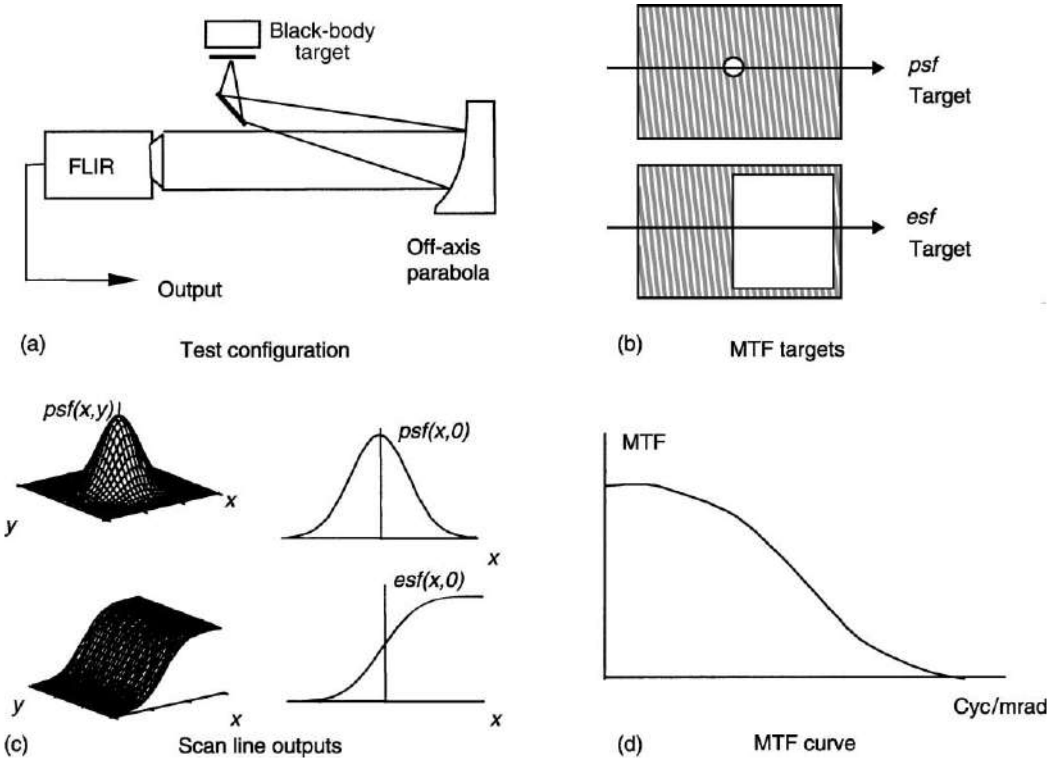
Black-body target
FLIR
Output
Off-axis parabola
(a)
Test configuration
psf
Target
esf
Target
(b)
MTF targets
psf(x,y)
x
y
psf(x,0)
x
esf(x,0)
y
x
x
(c)
Scan line outputs
MTF
Cyc/mrad
(d)
MTF curve

Figure 11 MTF measurement.

This parameter can be extended to field performance using Johnson's criteria.

As a laboratory measurement, it uses a 7:1 aspect ratio bar target. The procedure is shown in Figure 12. A bar target mask, with a black-body illuminator, is placed at the focal plane of a collimator (Figure 12a and 12b). The temperature is increased from a small value until the bars are just resolvable to a trained

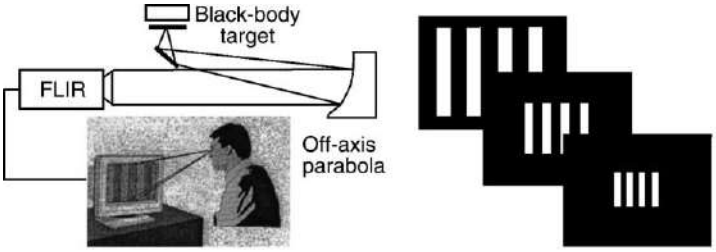

(a) Test configuration (b) MRT targets

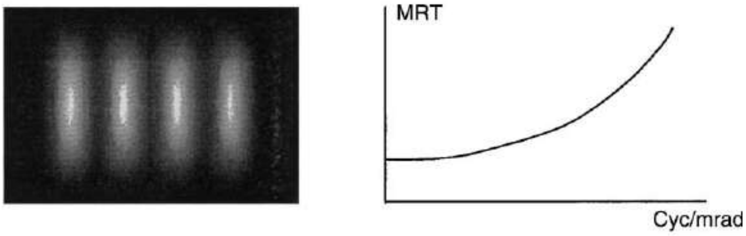

(c) MRT target image (d) MRT curve

Figure 12 MRTD measurement.

observer (Figure 12c). Then the temperature difference is decreased until the bars are no longer visible. These are averaged for a particular target. The same procedure is performed with a negative contrast bar target. The average of the positive and negative values is the MRT for a particular spatial frequency. These data points are plotted and a curve may be fitted to interpolate/extrapolate performance at other than the discrete spatial frequencies of the targets, as in Figure 12d.

Summary

We have provided a general description of infrared imaging systems in terms of characteristics, modeling, field performance, and performance measurement. The characteristics of infrared imagers continue to change. Significant changes have occurred in the past five years, to include the development of higher performance uncooled imagers and ultra-narrow field of view photon systems. Large format detector arrays are commercially available in the mid-wave and are becoming more available in the longwave. Still in the research phase are dual-band focal plane level. At first sight, it appears that the longwave flux characteristics are as good as a daytime visible system; however, there are two other factors limiting performance. First, the energy bandgaps of infrared photons are much smaller than those photons in the visible, so the detectors suffer from higher dark current. The detectors are usually cooled to reduce this effect. Second, the reflected light in the visible is modulated with target and background reflectivities that typically range from 7 to 20%. In the infrared, all terrestrial objects are emitting with an emission temperature of around 300 Kelvin. Typically, a two-degree equivalent black-body difference in photon flux is considered a high contrast target to background flux. The flux difference between two black-bodies of 302 K compared to 300 K can be calculated in a manner similar to that shown in Figure 4. This is the difference in signal that provides an image, so note the difference in signal compared to the ambient background flux. In the longwave, the signal is three percent of the mean flux and in the mid-wave, it is six percent of the mean flux. This means that there is a large flux pedestal associated imaging in the infrared. Unfortunately, there are two problems accompanying this large pedestal. First, photon noise is the dominant noise and is determined by the mean of the pedestal and is compared to the small signal differences. Second, the charge well storage in infrared detectors is limited to around 10^7 charge carriers. A longwave system in a hot desert background would generate 10^{10} charge carriers in a 33 millisecond integration time. Smaller F-numbers, spectral bandwidths, and integration times are used so that the charge well does not saturate. This results in SNR of 10 to 30 times below the ideal. It has been suggested that some on-chip compression may be a solution to the well pedestal problem. In many scanning FLIR systems, the pedestal is eliminated by AC-coupling the detector signals. Infrared focal plane array (IRFPA) readout circuits have been previously designed and fabricated to perform charge skimming and charge partitioning.

Another major difference in infrared systems, from that of visible systems, is the size of the detector and diffraction blur. For infrared photon detectors, typical sizes range from 25 to 50 μm, where visible systems can be fabricated with 3 μm detectors. Also, the diffraction blur for the longwave is more than ten times larger than the visible blur and mid-wave blur is eight times larger than visible blur. Therefore, the image blur, due to diffraction and detector size, is much larger in an infrared system than a visible system. Also, it is common for infrared staring arrays to be sampling limited where the sample spacing is larger than the diffraction blur and the detector size. Dither and microscanning are frequently used to enhance performance.

Finally, both infrared photon and thermal detectors in staring arrays have responsivities that vary dramatically and it is current practice to correct for the resulting nonuniformity (nonuniformity correction or NUC). The nonuniformity can cause fixed pattern noise in the image that can limit the performance of the system even more than temporal noise (especially in static imaging or stabilized target acquisition imaging).

Modeling Infrared Imagers

Otto Schade Sr. developed the earliest imaging system models and performance measures. His work describing television in the 1950s and 1960s, pioneered the way for the characterizations of imaging sensors used by engineers today. His performance measure for still pictures, moving pictures and television systems was based on an observer resolving a three bar periodic target.

John Johnson developed the technique of relating the acquisition of military targets in the field to laboratory measurements of resolvable bar targets. He divided the discrimination tasks into four categories: detection, orientation, recognition, and identification. This technique was termed an 'equivalent bar pattern approach'. It related performance on a simple test target to performance with complex object targets. Johnson viewed scale models and bar targets in the laboratory against a bland background. The smallest discernible barchart target yielded the maximum resolvable bar pattern frequency. These results, tabulated in Table 3 as cycles across the minimum dimension, were the basis for the discrimination methodology in widespread use today. Note that detection took only one cycle, but as the tasks got more specific or difficult, the requirement went as high as 8 cycles across the target minimum dimension.

Recognizing that typical vehicle aspect ratios were somewhat limited, Johnson and Lawson conducted further experiments. Paul Moser conducted some of his own, reanalyzed Johnson and Lawson's, and developed the concept that this task was related to the average or critical dimension instead of the minimum dimension. This led to the conversion of line pairs to resolution elements, introducing the second dimension. Additionally, he performed a similar experiment on marine targets, which is summarized in Table 4. Lloyd and Sendall developed the MRT metric, which combines both resolution and sensitivity characteristics. This measurement of MRT is described in detail in a following article, but there is a theoretical model that is used to evaluate new sensor designs and concepts. It relates the minimum temperature difference between bar pairs at which they are just resolvable as a function of spatial frequency. MRT describes the infrared imager sensitivity as a function of resolution. More specifically, MRT is a measure of thermal contrast sensitivity as a function of spatial frequency.

Ratches *et al.* developed the NVL (US Army Night Vision and Electronic Sensors Directorate formerly known as the Night Vision Laboratory) Static Performance Model, which predicted the end-to-end performance. It started with the target signature and carried through to the observer. The MRT, a laboratory measurement or a modeled value, was related to the probabilities of discrimination through the target transfer probability function, cumulative percentages related to Johnson's criteria.

Table 3 Johnson's criteria

Target	*Resolution per minimum dimension*			
Broadside view	*Detection*	*Orientation*	*Recognition*	*Identification*
Truck	0.9	1.25	4.5	8
M-48 Tank	0.75	1.2	3.5	7
Stalin Tank	0.75	1.2	3.5	6
Centurion Tank	0.75	1.2	3.5	6
Half-track	1	1.5	4	5
Jeep	1.2	1.5	4.5	5.5
Command car	1.2	1.5	4.3	5.5
Soldier (standing)	1.5	4.8	3.8	8
105 howitzer	1	1.5	4.8	6
Average	1 ± 0.25	1.4 ± 0.35	4.0 ± 0.8	6.4 ± 1.5

Table 4 Moser's data

Discrimination task	Line pairs/critical dimension
Detect ship	Aperiodic treatment
Classify as combatant	4
Recognize type	10

D'Agostino and Webb created the 3D noise model, both spatial and temporal considerations, for analyzing system noise. This technique divided the noise components into manageable and understandable components. It also simplified the integration of these effects into models.

The introduction of focal plane array (FPA) imagers has created a requirement to improve the model performance to account for the difference between scanning arrays and staring arrays. This change has made sampling an important area for improvement. A new model called NVTHERM is a result of the NVESD model improvement effort, which addresses undersampled imaging system performance.

Sensor characterization is seen in three ways: theoretical models, field performance, and laboratory measurements (Figure 9). Theoretical models are developed that describe sensor sensitivity, resolution, and human performance for the purpose of evaluating new conceptual sensors. Acquisition models, and other models, are developed to relate the theoretical models to field performance. This link allows theoretical models to be converted to field performance quantities (e.g., probabilities of detection, recognition, and identification). Field performance is measured in the field so that the theoretical models can be refined and become more accurate with advanced sensor developments. Since field performance activities are so expensive, methods for the direct measurement of sensor performance are developed for the laboratory. Field performance testing of every infrared sensor built, including buy-off, acceptance, and life-cycle testing, is ridiculous and out of the question. Laboratory measurements of sensor performance are developed such that, given these measurements, the field performance of a system can be predicted. Sensor characterization programs require accurate sensor models, field performance estimates with acquisition models, and repeatable laboratory measurements.

There are a few alternatives to the US NVTherm model. One candidate approach to undersampled imager modeling is Germany's TRM3, or the MDTP approach. The impact of undersampling in the image of a 4-bar target can readily be seen by simply observing the change in the image as a function of spatial frequency and phase. The spatial frequency is defined as line pairs per angular extent of the target, and the phase specifies the relative location of the target image to the detector array raster. These effects are obviously not independent of each other, but for each target there can be found an optimal phase where the observer can see the maximum amplitude modulation in the image of the target. MRT measurements in the past have utilized this variation with phase by allowing the observer to optimize the displayed image through target or system line-of-sight changes during the measurement process. For undersampled imagers, TRM3 addresses the problem of the MRT calculation's inability to predict beyond the half sample rate of the sensor, by replacing the MTF in the denominator of the MRT equation with an appropriately scaled term called the average modulation at optimum phase (AMOP). AMOP is the average signal difference in the image of the 4-bar standard pattern, with the test pattern positioned at optimum phase. AMOP oscillates between the presample MTF and the bar modulation. Beyond a frequency of 0.8 times the sample rate or 1.6 times the half sample arrays and quantum well detector systems. These systems will find their place in applications to include both military and commercial sectors.

Further Reading

Bijl P and Valeton MJ (1998) Triangle orientation discrimination: the alternative to minimum resolvable temperature difference and minimum resolvable contrast. *Optical Engineering* 37(7): 1976–1983.

D'Agostino J and Webb C (1991) 3-D analysis framework and measurement methodology for imaging system noise. *Proceedings of SPIE* 1488: 110–121.

Driggers RG, Cox P and Edwards T (1998) *Introduction to Infrared and Electro-Optical Systems*. Boston, MA: Artech House, p. 8.

Holst GC (1995) *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 347.

Holst GC (1995) *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 432.

Johnson J (1958) Analysis of image forming systems. *Proceedings of IIS* 249–273.

Johnson J and Lawson W (1974) Performance modeling methods and problems. *Proceedings of IRIS*.

Ratches JA (1973) *NVL Static Performance Model for Thermal Viewing Systems*. USA Electronics Command Report ECOM 7043, AD-A011212.

Schade OH (1948) Electro-optical characteristics of television systems. *RCA Review* IX(1–4).

Sendall R and Lloyd JM (1970) Improved specifications for infrared imaging systems. *Proceedings of IRIS* 14(2): 109–129.

Vollmerhausen RH and Driggers RG (1999) NVTHERM: next generation night vision thermal model. *Proceedings of IRIS Passive Sensors* 1.

Wittenstein W (1998) Thermal range model TRM3. *Proceedings of SPIE* 3436.

Interferometric Imaging

D L Marks, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Interferometric imaging is image formation by measuring the interference between electromagnetic signals from incoherently emitting or illuminated sources. To learn about interferometric image formation using coherent light, *see* **Holography, Techniques:** Overview. To understand what optical coherence is, *see* **Coherence:** Overview. To learn what kinds of sources produce incoherent light, *see* **Incoherent Sources:** Lamps. To learn about interferometric instruments, *see* **Interferometry:** Overview. To learn about interferometric instruments utilizing incoherent light, *see* **Interferometry:** White Light Interferometry. To better understand the effect of the coherence of light sources on image formation, *see* **Coherence:** Coherence and Imaging.

Interferometric imaging attempts to improve the resolution of images by interferometrically combining the signals from several apertures. By doing so, a resolution can be achieved that is superior to that achievable through standard image formation by any of the individual apertures. This is especially useful in astronomy, where resolving more distant and smaller stellar objects is always desirable. It is also finding increasing use in microscopy. It has several advantages over conventional noninterferometric imaging methods. The method can combine the light gathered from several imaging instruments to form an image superior to that which can be formed by any one of the instruments. It can also obviate the need to produce a single very large aperture to achieve the equivalent resolution, which in many cases is of impractical size. In addition, because it can produce a phase-resolved measurement, the data can be processed more flexibly to form a computed image estimate. There are also considerable disadvantages to interferometric imaging compared to conventional imaging methods. The amount of signal gathered is far less than could be gathered with an equivalent single aperture. It is difficult to mechanically or feedback stabilize the time delay between two widely separated apertures to obtain an accurate phase measurement. Often only a small bandwidth of the source light can be utilized, further reducing the available signal. The interference component of the combined signal can be very small. Typically an interferometric image must be computed from the data, rather than being directly measured on photographic film or an electronic focal plane array. Even with these disadvantages, interferometric imaging is frequently more feasible than building a single large aperture to achieve the highest available resolution images.

In contrast to interferometric imaging, a standard image forming instrument such as a telescope has a resolution limited by the telescope mirror aperture in the absence of aberrations and atmospheric turbulence. As telescope mirrors become very large, they become bulky, extremely heavy, and difficult to manipulate, as well as deform under their own weight and temperature gradients. However, adaptive optics is being more frequently utilized to dynamically correct for instrument aberrations as well as atmospheric turbulence. The images captured by telescopes are typically intensity-only images, which are amenable to image processing and deconvolution, but nevertheless lack phase information. Interferometric imaging is an alternative, where the resolution of a large aperture can be attained by measuring the interference between individual points of the electromagnetic field within an area equivalent to the large aperture extent. The interferometric combining of light from sub-apertures to achieve the resolution of a larger aperture is called aperture synthesis.

Unlike image formation by a lens, interferometric images are never physically formed. Conventional images are formed when diverging spherical waves emanating from sources are focused by a lens to converging spherical waves on a photographic film or an electronic sensor. In this case, the image information is directly contained in the exposure at each position on the sensor. Interferometric measurements contain the image information as statistical correlations between the electromagnetic fields at pairs of spatial points. The statistical properties of electromagnetic waves are modeled by optical coherence theory. Interferometric imaging works by measuring the statistical correlations between various pairs of points in the electromagnetic field, and then uses optical coherence theory to infer what the image of the source of the radiation is.

The light that emanates from most natural and many artificial sources is incoherent. Incoherent radiation is optical random noise. It is analogous to the sound of static heard on a radio receiving no signal. Because it is random, the fluctuations of the electromagnetic field produced by an incoherent source are completely unrelated to the fluctuations of any other source. The origin of the randomness is usually from microscopic processes such as the thermal motions of electrons or the spontaneous emission of an excited atom. Because these microscopic processes tend to act independently of one another, the light produced by different sources tends to be mutually incoherent. Because incoherent sources do not radiate a deterministic field, the average amount of power they radiate is usually specified rather than the field itself, which is random.

However, as the fields propagate away from their sources, the fields can become partially coherent. To see this, consider the light waves propagating away from a single point-like incoherent source. The field produced at the source point consists of random fluctuations in time, but these fluctuations travel outward in a spherical wave at the speed of light. The fluctuations of the field of any two points an equal distance from the source will be the same, or perfectly correlated, because they arrive from the source at the same time. This is illustrated in Figure 1a, where a random wave is emanated from a point incoherent source. Because of the complete correlation, we can say that these two points are fully spatially coherent.

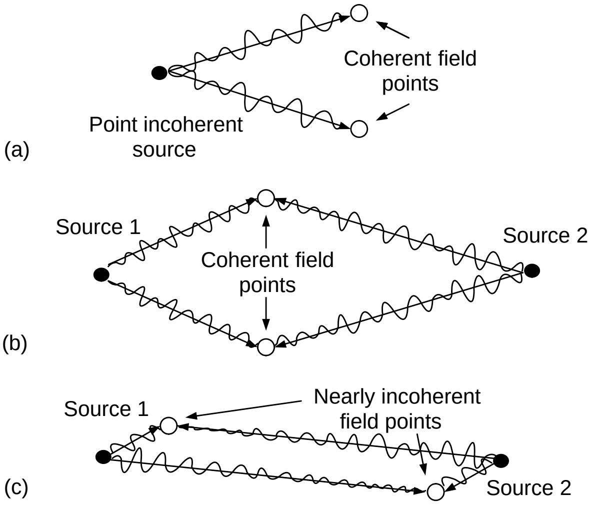

Figure 1 The distance between incoherent sources and two field points determines the coherence between the field points. (a) Because there is only one source, the field everywhere is determined by the radiation of that source. Therefore, the same fluctuations of the wave arrive at two points an equal distance from the source at the same time, and produce identical fields at both points, and so are completely coherent. (b) Two field points are now the same distance from two incoherent sources. Because both sources contribute identical fields to both points (because the fluctuations of the wave from both sources arrive at both points at the same time), the superposition of the fields from both sources at both field points are identical, and so the fields are coherent. (c) Two field points are now near two sources. Because each field point receives radiation mostly from its nearby source and very little from the other source, the fields at each field point are almost completely mutually incoherent.

So while the source itself produces random noise, the fluctuations of the field at points distant from the source become correlated because they originate from the same source.

Partially coherent waves occur when the measured field contains a superposition of the fields from more than one incoherent source. If there are now two point sources instead of one, each field point in space will receive radiation from both sources. The amount of the contribution from each source to each point in space depends on the distance between the field point and the source. In general, because both sources can be different distances from the two field points, the fluctuations at the field points will not be the same. For example, two field points that are the same distances from both sources will be perfectly correlated, or coherent, because they receive the same combination of fields from both sources. This situation is illustrated in Figure 1b. However, if one considers two points that are near both respective sources, the points will usually receive light from the nearby source. Because the sources are incoherent, and each point is mostly receiving radiation from its closer source, these points will be nearly uncorrelated, as shown in Figure 1c.

A spatial distribution of incoherent sources radiating various amounts of power will create a pattern of correlations in the field remote from the source. By using interferometry, the correlation between pairs of field points remote from the sources can be measured. The interference between two points in the electromagnetic field is achieved by relaying the two fields (e.g., with mirrors) to a common location that has an equal travel time for the fields from both points, where the fields are superimposed. The power of the superimposed beams is then detected, by a photocathode, for example. If the two signals being combined are incoherent, then no interference will take place, and the total power measured on the photodetector is simply the sum of the power of the two signals. However, if the two signals are partially coherent, there will be a deviation in the measured power from the sum of the power of the two signals. The magnitude of this deviation relative to the power of the constituent signals indicates the degree of partial coherence of the signal.

An interferometer, such as depicted in Figure 2, gathers the light from two areas of the electromagnetic field remote from an object and combines them together at the same point to form an interferogram. In this example, two telescopes of a known separation gather light from a star. The primary mirror of each telescope is of insufficient size to resolve the star as anything but a point. However, a telescope is used in practice to collect enough light to produce a

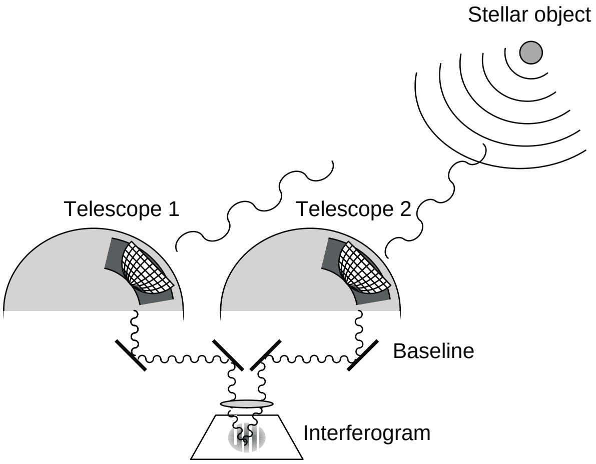

Figure 2 Two telescopes receive radiation from a stellar object. The light is collimated and directed down the baseline to a point halfway between the telescopes, where the two images are interfered together onto a focal plane.

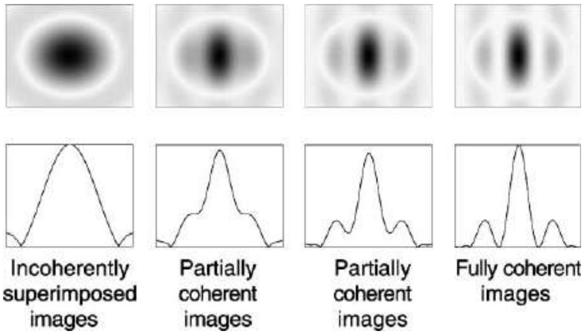

Figure 3 The image of two interfering wavefronts from a point object with various states of coherence.

measurable interference signal. Each telescope primary mirror forms a point image of the star. At the image plane of each telescope a mirror directs the point image to a measurement plane equidistant between the two telescopes. This line that joins the two telescopes is called the baseline, and it is the baseline length rather than the individual telescope mirror size that determines the minimally resolvable size. The image from each telescope is superimposed by mirrors (or perhaps beamsplitters) and imaged onto a focal plane, to get an image similar to that of Figure 3. If the light arriving from one telescope is blocked, then the image formed looks like an Airy ring pattern formed by an image, a point-like object, which is the incoherent image of Figure 3. However, when both signals are allowed to interfere, the image will have vertical stripes superimposed on it. If these stripes are very prominent, so that the image is fully darkened inside the stripes, then an image similar to the rightmost image of Figure 3 is obtained, complete intereference is occurring and the two field points at which the telescopes are situated are coherent or completely correlated. If the stripes disappear, then no interference is occurring and the two field points at which the telescopes reside are incoherent. These examples are shown in Figure 3, with images of two superimposed points with varying degrees of coherence on the top row, and a line plot of the measured field through the center of the images in the bottom row. The degree to which the contrast of the stripes is enhanced by coherence is referred to as the 'modulation' and indicates the degree or magnitude of partial coherence.

In addition to having a degree of coherence, the correlation between two electromagnetic field points also has a relative phase shift. This phase shift is given by the relative position of the interference maxima, as shown in Figure 4. As the relative phase between the

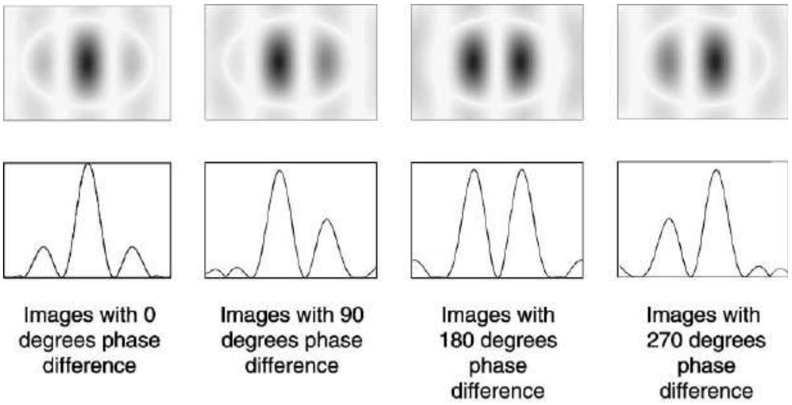

Figure 4 The image of two interfering wavefronts from a point object with various phase differences between the two wavefronts.

fields collected by the telescopes changes (which is a function of the path length difference between the light arriving at the two telescopes), the peaks of the interferogram between the two telescope images shift position. By observing the depth of the formed interference fringes and the position of the fringes, both the amplitude and the phase of the correlation between the two field points can be simultaneously measured. The amplitude of the modulation and the phase of the interference are used to compute the complex degree of coherence that is used in the computation of interferometric images.

In practice, the phase of the complex degree of coherence is much more difficult to measure than its magnitude. This is because while the amplitude tends to change very slowly as the positions of the two correlated points change, the phase is sensitive to changes in relative path length on the order of an optical wavelength. Even though a long baseline produces better resolution, the baseline length must be known to a small fraction of a wavelength to obtain a meaningful phase measurement. Minute vibrations in the positions of the mirrors on the baseline can easily cause fluctuations in the path length greater than a wavelength. This is observed as wild oscillations in the position of the fringes on the image plane. In addition, atmospheric turbulence can also produce large random phase shifts. Achieving a stable phase measurement is one of the most challenging engineering aspects of interferometric imaging.

To model realistic sources, we must determine the coherence produced by sources of a finite size. Finite size incoherent radiators can be regarded as an arrangement of an infinite number of point-like incoherent sources radiating various amounts of power. When an interferogram is made of a field which has two superimposed incoherent sources, the power of the interferograms that each would make alone is added together, rather than the fields of the interferograms. This is because when the fields of the two sources interfere, the phases of both sources are varying randomly over time. The two fields will randomly vary between constructively and destructively interfering with each other. Over a long time, the constructive and destructive interference averages out to no interference. We can then calculate what the interferogram of many incoherent sources combined is by computing the interferogram of a point source and adding the power of many such interferograms together.

To figure out what the power of an interferogram of an arbitrary object is, first consider what the interferogram of a single point source is. Consider a situation depicted in Figure 5, but with a single point illuminator at the source plane positioned at (x,y,z) with radiant power I. We will have a receiving plane at a distance z from the source plane with two telescopes located symmetrically about the origin at $(\Delta x/2,\Delta y/2,0)$ and $(-\Delta x/2,-\Delta y/2,0)$. The fields received at the telescopes from the point will be interfered together at the center of the baseline, which is at the origin. We will assume that the point source produces a narrow band of radiation centered at wavelength λ. The point source produces a spherical wave that is received at both telescopes. The interference power is proportional to the two intensities measured at each telescope with the relative phase at the two telescopes causing constructive or destructive interference:

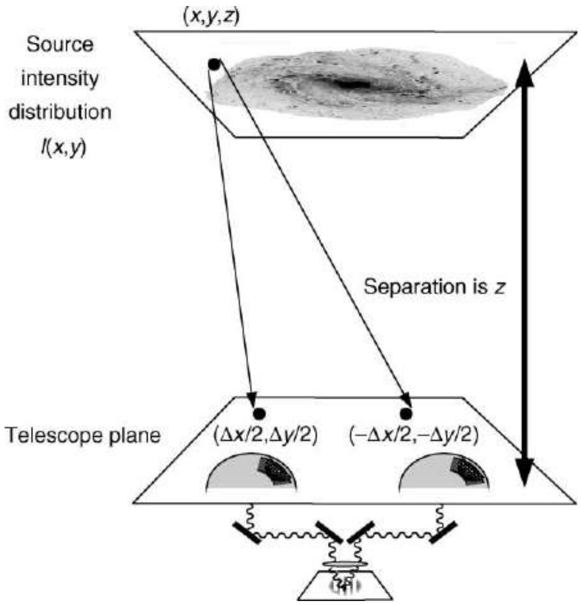

Figure 5 Diagram of source and telescope plane used for derivation of Van Cittert–Zernike theorem. A source at a distance z from the telescope plane has a power density $I(x,y)$. The telescope plane has two movable apertures symmetric about the origin of the telescope plane separated by $(\Delta x,\Delta y)$. The light from both apertures is directed to the center of the baseline at the origin of the telescope plane, where it is interfered on the sensor.

$$P \propto I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\phi \qquad [1]$$

where I_1 and I_2 are the intensities of the field at each telescope, and ϕ is the phase difference between the two fields at each telescope. We will assume that the point (x,y,z) is very far away from the telescopes, and so the intensities at the two telescopes are equal and proportional to the point intensity I, so that $P \propto I(1+\cos\phi)$. The phase ϕ is given by the difference in distance from the the source point at (x,y,z) to the receivers at $(\Delta x/2,\Delta y/2,0)$ and $(-\Delta x/2,-\Delta y/2,0)$. The phase is given by the difference in the Euclidean distances:

$$\phi = \frac{2\pi}{\lambda}\left(\sqrt{z^2 + (x + \Delta x/2)^2 + (y + \Delta y/2)^2} - \sqrt{z^2 + (x - \Delta x/2)^2 + (y - \Delta y/2)^2}\right) \approx \frac{2\pi(x\Delta x + y\Delta y)}{z\lambda} \qquad [2]$$

where the second term gives the difference in the distance to a first-order approximation in x and y. The interference intensity relation is then $P \propto I(1 + \cos(2\pi(x\Delta x + y\Delta y)/z\lambda))$.

If the source is now of a finite site, we can regard it as radiating an intensity (power per unit area) $I(x, y)$. To compute the total intensity of the interferogram, we integrate over all the sources in the source plane:

$$P(\Delta x,\Delta y) \propto \int I(x,y)(1+\cos(2\pi(x\Delta x+y\Delta y)/z\lambda))\mathrm{d}x\mathrm{d}y = \int I(x,y)\mathrm{d}x\mathrm{d}y + \int I(x,y)\cos(2\pi(x\Delta x+y\Delta y)/z\lambda)\mathrm{d}x\mathrm{d}y \qquad [3]$$

This result can be interpreted as follows. The intensity of the interference when the two telescopes are separated by a distance (Δx,Δy) is proportional to the real part of the two-dimensional Fourier transform of the intensity distribution $I(x,y)$ evaluated at spatial frequencies ($\Delta x/z\lambda$,$\Delta y/z\lambda$), plus a constant intensity. This result, related to the Van Cittert–Zernike theorem, is a commonly used result in interferometric imaging. It is covered in more detail in **Coherence:** Overview.

By moving the telescopes to various horizontal and vertical separations and measuring the interference intensity, the real part of the Fourier transform of the intensity is inferred. In practice the imaginary part can also be found by varying the baseline optical path length difference by a quarter wavelength. With these samples of the Fourier transform, the inverse Fourier transform computes the image intensity distribution $I(x,y)$. This is how simple computational image formation occurs with interferometric data.

One application of the Van Cittert–Zernike theorem and interferometric imaging is the measurement of the diameter of stars. When observing coherence, the correlations between field points tend to decrease as the separation between the points grows. By observing this decrease of the coherence, the angular size of a star can be measured. This method was pioneered by Albert Michelson in 1890 with the invention of the Michelson Stellar Interferometer. It is needed because the aperture size required to measure the angular diameter of most nearby stars in the Milky Way is larger than the size of most telescope apertures. Stars usually have a circular profile that produces a circularly symmetric pattern of correlations. The coherence between field points as the distance between them is increased does not decrease monotonically, but disappears at certain separations, and has many revivals. The first separation distance at which interference disappears (and the degree of coherence is zero) for a circular profile object is $\mathrm{d}_0 = 0.61\lambda/\alpha$, where λ is the wavelength of the light measured from the star, α is the angular diameter (in radians), and d_0 is the first separation from zero at which the interference disappears. In practice, each wavelength produces a null in interference at a different separation, so all but a small bandwidth of the star's light is filtered out before measurement to achieve a suitably strong interference null. For example, for the star Sirius, which has an angular diameter of 0.0068 arc seconds viewed from the Earth, measuring at a wavelength of $\lambda = 500$ nm (blue-green in the visible spectrum), the first separation distance at which no interference will be observed is at 18.5 m. This is larger than the largest available telescope aperture, and therefore an image cannot be directly formed that can resolve the size of Sirius.

Modern examples of stellar interferometry include the European Southern Observatory in Paranal, Chile, the CHARA Array at Mt Wilson, California, USA, and the NASA Jet Propulsion Laboratory Keck Interferometer in Hawaii, USA. The ESO main telescope system is the Very Large Telescope Interferometer (VLTI) consisting of four 8.2 m diameter telescopes as well as several smaller ones that can be moved independently, and will achieve milliarcsecond resolution through the interferometric combination of the telescopes. These telescopes are connected by underground tunnels through which the collected light from each telescope is combined. To measure the phase of the interference, extreme positioning precision is required in the relative delay between the light collected by the telescopes along the baseline, of 50 nm over 120 m, or 1 part in 2400 million, equivalent to 1.6 cm over the circumference of the Earth. Achieving this requires combinations of electronic and mechanical feedback systems with laser stabilized interferometric position measurement. The CHARA array has a very large 330-meter baseline, with 1-meter telescope apertures. The Keck Interferometer consists of two 10 m diameter telescopes with a baseline separation of 85 m. One of the main goals of achieving such high resolution is to be able to observe planets and planet formation around distant stars, to identify possible signs of life in other solar systems.

A similar principle is used for interferometric image formation in radio astronomy. The chief differences are the methods used to collect, detect, and correlate the electromagnetic radiation. Radio astronomical observatories, such as the Very Large Array in New Mexico, USA, and MERLIN at the Jodrell Bank Observatory in Cheshire, UK, use collection dishes and radio antennas to directly measure the electric field, which is not feasible at optical frequencies. These fields are converted to electrical voltages, which can then be directly correlated using electronic mixers, rather than detecting the correlation indirectly through interference. Because the wavelength is much larger, the stability requirements are greatly relaxed. Radio and optical frequency interferometric images are formed in essentially the same way, using the Van Cittert–Zernike theorem.

Interferometric imaging methods have also become of interest for microscopy applications. Aperture synthesis is not needed in microscopy, because the objects of interest are very small and therefore single apertures can be used to achieve diffraction-limited resolution. However, interferometric detection of partially coherent light affords greater flexibility because an image may be computed from the data rather than being directly detected. Typically this is achieved by creating an incoherent hologram of the object, typically through shearing interferometry. For example, by using a Rotational Shearing Interferometer, one can measure an incoherent hologram that can be used to produce a representation of an object with a greatly increased depth-of-field. In addition, direct detection of the partial coherence of sources can allow the methods of computed tomography to be used to produce three-dimensional images of volumes of incoherently emitting objects. Because of the versatility that interferometric detection can provide, interferometric imaging will surely be increasingly used outside of astronomy.

See also

Coherence: Coherence and Imaging; Overview. **Holography, Techniques:** Overview. **Incoherent Sources:** Lamps. **Interferometry:** Overview; White Light Interferometry.

Further Reading

Baldwin JE and Haniff CA (2002) The application of interferometry to optical astronomical imaging. *Philosophical Transactions of the Royal Society of London A* 360: 969–986.

Beckers JM (1991) Interferometric imaging with the very large telescope. *Journal of Optics* 22: 73–83.

Born M and Wolf E (1997) *Principles of Optics*. New York: Cambridge University Press.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw-Hill Inc.

Goodman JW (2000) *Statistical Optics*. New York: Wiley-Interscience.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. New York: Cambridge University Press.

Marks DL, Stack RA, Brady DJ, Munson DC Jr and Brady RB (1999) Visible cone-beam tomography with a lensless interferometric camera. *Science* 284: 2164–2166.

Monnier JD (2003) Optical interferometry in astronomy. *Reports on Progress in Physics* 66: 789–857.

Potuluri P, Fetterman MR and Brady DJ (2001) High depth-of-field microscopic imaging using an interferometric camera. *Optics Express* 8: 624–630.

Thompson AR, Moran JM and Swenson GW (2001) *Interferometry and Synthesis in Radio Astronomy*. New York: Wiley-Interscience.

Lidar

M L Simpson and D P Hutchinson, Oak Ridge National Laboratory, Oak Ridge, TN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Light detection and ranging (LIDAR) refers to an optical system that measures any or all of a variety of target parameters including range, velocity, and chemical constituents. Usually a LIDAR system consists of a transmitting laser source, transmitter optics, receiver optics, and a detector. Several configurations of LIDAR systems are possible, based on the type of laser source and detection scheme selected. Basically, sources can be continuous wave (CW), pulsed, or chirped. Sources that are CW are most often used to measure velocity or chemical constituents. Pulsed and chirped sources provide the additional capability of measuring range. LIDAR receivers are either incoherent (direct detection – amplitude only) or coherent (heterodyne detection – amplitude and frequency). Generally, for velocity detection, a coherent receiver with a local oscillator (LO) laser is required. A coherent receiver is also needed if a chirped laser source is used, for instance, in frequency modulation continuous wave (FM-CW) LIDAR.

For the special case of imaging LIDAR, the detector has multiple elements. In an imaging LIDAR system, the detector, or focal plane array (FPA), and the receiver optics are designed to image independent volumes in object space onto individual elements in the FPA. Each detector element, therefore, contains information for a specific volume in space and can be thought of as an independent LIDAR system. The advantage of an imaging LIDAR system is that multiple measurements can be made on extended target(s) at the same time. Thus, characterization of rapidly changing events (like wake vortices), speckle reduction by spatial averaging, synthetic aperture radar, multiple target tracking, and target identification are all possible without the sophisticated algorithms and hardware platforms required in single channel scanning systems. In addition because the multiple pixels allow parallel image capture, a true imaging LIDAR can acquire data much more rapidly than a scanning system.

In this chapter, for clarity, the emphasis will be on imaging LIDAR systems applied to hard body targets for the measurement of range and velocity. The basic configurations presented here, though, are extensible to distributed targets and measurements of wind profiles and atmospheric constituents. Also, since general textbooks are available that cover LIDAR systems and incoherent imaging systems, the concentration will be on architectures, optics, detectors, and design considerations that are specifically applicable to coherent imaging LIDAR.

Frequency Modulated Continuous Wave (FM-CW) LIDAR

One LIDAR technique, which is particularly suited for imaging due to the simplicity of the processing, is FM-CW LIDAR. FM-CW LIDAR simultaneously measures range and velocity (Doppler shift) by using a frequency chirped waveform. Referring to Figure 1, assume that a continuous wave (CW) laser frequency is linearly modulated (chirped) by a triangular wave about an offset frequency, such that the modulated signal is given as:

$$f(t) = f_{\mathrm{o}} + \alpha\tau \qquad [1]$$

where f_{o} is the subcarrier offset frequency and α is the linear modulation rate in frequency/second, and τ is a time variable. The transmitted beam is focused by the optical system onto the target surface, reflected (or absorbed or scattered) and arrives at the receiver after τ seconds, where

$$\tau = 2\frac{R}{c} \qquad [2]$$

and R is the distance to the target surface, and c is the speed of light. The returning signal is optically mixed with the LO on the detector and produces a sum/difference beat frequency (Δf). Figure 1 shows the incident wave with its linear or chirped frequency characteristics and the delayed version returned from the target. Only the difference frequency can be optically detected since only it falls within the detector's bandwidth. Assuming the target is not moving, the result is

$$\Delta f = f(t+\tau) - f(t) = f_{\mathrm{o}} + \alpha(t+\tau) - (f_{\mathrm{o}} + \alpha t) = \alpha\tau \qquad [3]$$

where $\alpha = B/T$, the chirp rate of the modulator in units of Hz/s, B is the modulation bandwidth, and T is the period of the chirp. Substituting into eqn [3] for

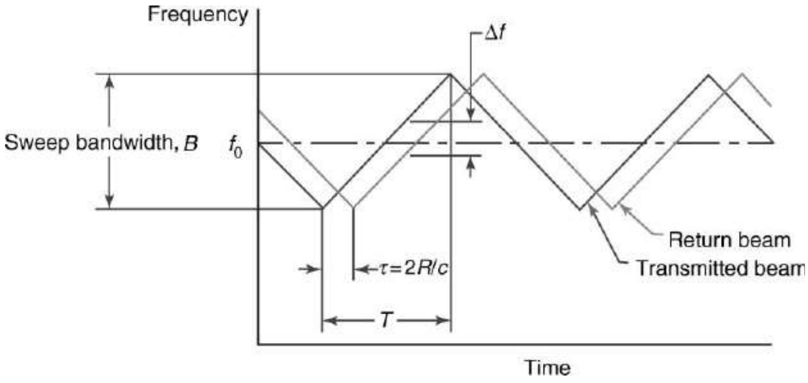

Figure 1 Triangular-wave modulation Envelope on the transmitted beam (black) and return beam (gray).

τ and α we have

$$\Delta f = \frac{B}{T}\left(2\frac{R}{c}\right) \qquad [4]$$

Finally, one can solve for the distance to the target, R, as a function of the beat frequency and chirp rate:

$$R = \frac{cT}{2B}\Delta f \qquad [5]$$

The difference frequency, Δf, is the same for the up and down ramp of the modulation signal. Figure 2 demonstrates the return signal for the case of a target moving toward the LIDAR system. The difference frequency now contains a component proportional to the target induced Doppler shift given by

$$f_{\mathrm{d}} = 2\frac{v}{\lambda} \qquad [6]$$

where v is the line-of-sight target velocity and λ is the laser wavelength. From Figure 2, the Doppler shift causes the difference frequency on the up ramp to be less than the range induced frequency shift given by eqn [5] and the net difference frequency is greater on the down slope. In this case, the difference frequency is given by

$$\Delta f = \frac{B}{T}\left(2\frac{R}{c}\right) + f_{\mathrm{d}} \qquad [7]$$

The range and Doppler contributions to the difference frequency may be separated by measuring the frequency difference over both the up and down chirp of the waveform. The Doppler frequency $f_{\mathrm{d}} = \Delta f(\mathrm{up}) - \Delta f(\mathrm{down})$ and the range is given by the average of $\Delta f(\mathrm{up})$ and $\Delta f(\mathrm{down})$ over a complete cycle of the modulating triangular wave. Therefore, substituting in eqns [5] and [6], the range may be determined from the equation

$$R = \frac{cT}{2B}\left\{\Delta f(\mathrm{up}) - \Delta f(\mathrm{down})\right\} \qquad [8]$$

and the line-of-sight (LOS) velocity is

$$v = \frac{\lambda}{2}\left\{\frac{\Delta f(\mathrm{up}) + \Delta f(\mathrm{down})}{2}\right\} \qquad [9]$$

The range resolution of a FM-CW LIDAR system is the range for which the difference frequency equals 1 cycle in the sweep time interval of T or $\Delta f = 1/T$. Then, from eqn [5], the range resolution, ΔR, is given by

$$\Delta R = \frac{cT}{2B}\frac{1}{T} = \frac{c}{2B} \qquad [10]$$

Therefore, the range resolution is only a function of the sweep bandwidth and not the laser wavelength. However, the Doppler shift is dependent on the laser frequency. The accuracy with which the difference frequency may be determined is a function of the number of cycles over which the measurement is made. The number of cycles is maximized for the case of $\tau = T/2$ or T should be set for twice the time delay for the maximum range of interest.

Numerical Example

A range resolution of 30 cm requires a chirp bandwidth, B, of 500 MHz. For a total range of interest of 10 km, the chirp time, T, should be 133.33 μs or the chirp rate should be 3750 Hz (1 up and 1 down ramp per cycle). The difference frequency for a 10 km range target will be 250 MHz, which demonstrates the need for wideband detectors to capture high-resolution range images (Table 1).

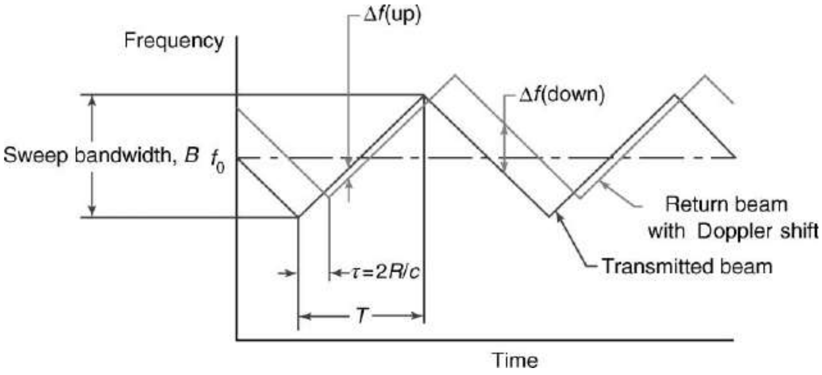

Figure 2 Lidar return with Doppler shift proportional to target velocity.

Table 1 Summary of FM-CW LIDAR characteristics

Parameter	*Value*
Range resolution (m)	0.3
Chirp rate (Hz)	3750
Chirp bandwidth (Hz)	500×10^6
Maximum range (km)	10

Optical Configuration for Imaging LIDAR

Imaging LIDAR systems can be constructed for virtually any wavelength range with appropriate sources, optical components, and detectors. The design of an optical receiver system capable of recording coherent images requires additional considerations over those necessary for a video camera or single-channel receiver. First, the collection optics must have resolution capabilities consistent with both the detector grid dimensions and the coherence properties of the source. Next, the local oscillator (LO) power must be applied to each array element. For a typical infrared (HgCdTe) detector, one to ten milliwatts of LO power might be required to reach the shot noise limit. For large arrays, this may result in unacceptable heat loads, even for momentary application. Thus, detector arrays with substantial spacing between detector elements benefit from the application of individual LO 'beamlets' to shield the insensitive portions of the array substrate. These beamlets can be generated by illuminating a mask containing an array of holes with the LO Gaussian laser beam. This scheme is illustrated in Figure 3.

In the configuration shown in the figure, the signal beam (from the target) and the LO beamlets are combined at the beamsplitter prior to being focused onto the detector plane by the objective lens. The detector plane coincides with the far-field image plane of the objective lens, and any additional optical elements utilized to enhance the camera's resolution are not shown. An expanded LO beam illuminates an aperture array located at the far-field object plane of the ancillary lens. For instance, if the objective lens and the ancillary lens are identical, the LO mask is imaged onto the detector plane with unity magnification. The detector surface is assumed to be flat with uniform video and heterodyne quantum efficiency.

Several aspects of this design are discussed below. We first discuss how the optical parameters may be adjusted to give nearly-Gaussian LO profiles at each detector position. The receiver antenna array is then shown to consist of nonoverlapping sub-beams spaced with the detector array geometry. Signal considerations for both incoherent and coherent sources are then discussed, and it is found that optimal detector size and array-spacing is dependent on the coherence properties of the intended object field.

Aperture Array

As an alternative to holographic LO beam array generation, the LO mask design presented here consists of an array of apertures with size and spacing equal to the detector dimensions. The LO is expanded to coherently illuminate the entire mask with approximately uniform intensity. The mask is located at the front focal plane of the ancillary lens and its image is located at the back focal plane (detector plane) of the objective lens. The coherent image of the mask formed at the detector plane by the ancillary-objective lens combination can be described by a two-stage Fraunhofer diffraction process.

Finite dimensions of the ancillary pupil results in spatial filtering which can affect the image profile of the LO mask. If desired, the ancillary pupil may be chosen to provide nearly-Gaussian LO beamlet profiles with beam waists at each detector location.

It is useful to first consider the coherent image of a single aperture uniformly illuminated with plane LO wavefronts. In one dimension, this wave train diffracts from the aperture to form a far-field diffraction pattern at the Fraunhofer plane of the ancillary lens given by

$$E_{\mathrm{F}}(\beta) = E_{\mathrm{F}} \frac{\sin(\beta)}{\beta} \qquad [11]$$

where $\beta = \frac{1}{2}kb\sin(\theta)$, $k = 2\pi/\lambda$, λ is the wavelength, b is the aperture width, and θ is measured from the optical axis. For small θ, $\beta = \pi b\nu/(\lambda f)$ where ν measures a point on the Fraunhofer plane and f is the focal length of the ancillary lens. The finite diameter D (determined by an aperture stop in the Fraunhofer plane) of the ancillary lens spatially filters the diffraction pattern and this limits the range of β values in eqn [11]. We define an aperture-limited pupil as one which limits β to the range $-\pi < \beta < \pi$; this will be the case when $f/D = b/2\lambda$.

Consider now a linear aperture array consisting of N individual apertures of width b and separation $a > b$. Assume that each aperture is illuminated by plane, uniform, mutually coherent LO wavefronts of equal intensity. In this case, the LO profile at the Fraunhofer plane is characterized by the diffraction pattern for a coherently illuminated linear diffraction grating:

$$E_{\mathrm{F}}(\theta) = E_0 \operatorname{sinc}(\beta) \frac{\sin(N\alpha)}{N\sin(\alpha)} \qquad [12]$$

where $\alpha = \frac{1}{2}ka\sin\theta$ and where β, θ, and k are defined above.

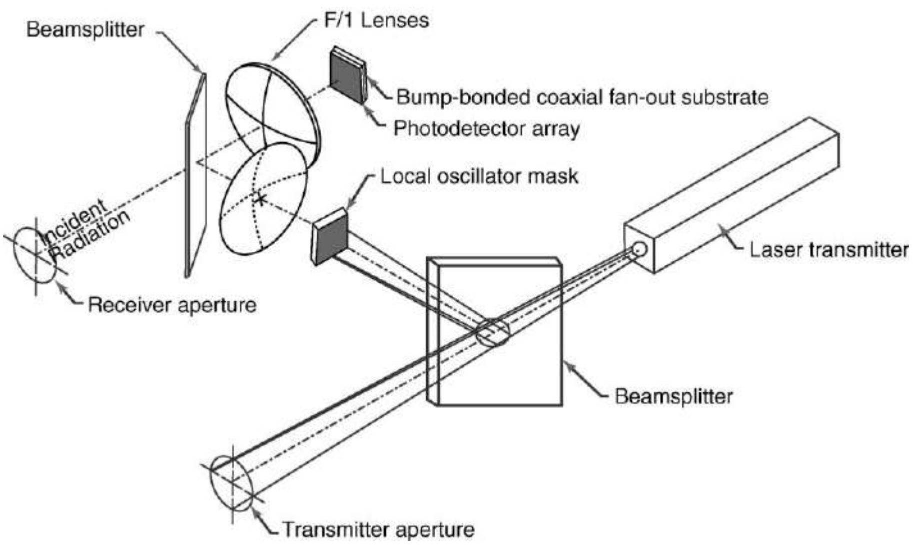

Figure 3 The individual beamlets are generated by the local oscillator or aperture mask. Reproduced with permission from Simpson ML, Bennett CA, Emery MS, *et al.* (1997) Coherent imaging with two-dimensional focal-plane arrays: design and applications. *Applied Optics* 36: 6913–6920.

Propagation of the LO from the ancillary lens Fraunhofer plane to the detector plane can be described by applying the Fraunhofer diffraction integral to eqn [12] with integration limits appropriate to the ancillary pupil dimensions. The larger pupil gives increased illumination at the detector edges while the aperture-limited pupil gives a nearly Gaussian LO profile with 1/*e* points close to the detector edges.

Receiver Antenna Array

The requirement of wavefront matching between the LO field and the signal field leads to the notion of an 'antenna beam' or equivalently, 'backward propagating LO' (BPLO) to define the field of view of an individual heterodyne receiver element. This beam is formed by allowing a reverse projection of the local oscillator field to diffract from the detector surface, through all collection optics in the signal beam path to a point on the object field of the imaging system. A receiver array will produce an array of such BPLO beams which must maintain the detector array geometry in the object field if the LO and signal beams are to mix efficiently.

As above, spatial filtering by the receiver pupil (assumed to be located at the Fraunhofer plane of the objective lens) will affect the BPLO profile. In the Fraunhofer plane of the objective lens, the uniformly illuminated single-detector BPLO profile is given by eqn [11], and we define an aperture-limited detector as one which limits the range of β to $\pm\pi$. Similarly, eqn [12] characterizes the BPLO profile resulting from a detector array uniformly illuminated with an extended uniform LO wavefront. The uniform illumination BPLO profiles in the object field differ from the detector plane profiles only in horizontal scale which now is consistent with the magnified detector dimensions.

For our purposes, we may define the far field of the objective lens according to the condition for maximum Gaussian beam collimation:

$$\omega_{\mathrm{d}}\omega_0 = \frac{f\lambda}{\pi} \qquad [13]$$

where ω_0 is an effective BPLO Gaussian beam waist at the Fraunhofer plane and ω_{d} is an effective BPLO beam waist at the detector. Under this circumstance, the BPLO Rayleigh range is $Z_{\mathrm{D}} = \pi\omega_0^2/\lambda = f(\omega_0/\omega_{\mathrm{d}})$. Identifying $\omega_0/\omega_{\mathrm{d}}$ as the magnification (notice that for the BPLO, the magnification is greater than unity) gives $d_{\mathrm{o}} = (d_{\mathrm{i}}/f)Z_{\mathrm{D}} \sim Z_{\mathrm{D}}$ where d_{o} is the object distance and d_{i} is the image distance. Thus the far-field of the objective lens is at least one BPLO Rayleigh range away.

The object field BPLO profiles for an aperture-limited detector array consists of nearly Gaussian sub-beams with 1/*e* points located approximately at the magnified detector edge locations. Since the far-field object distance exceeds Z_{D}, the divergence of each BPLO sub-beam is constant and remains distinct and nonoverlapping for all $d_{\mathrm{o}} > Z_{\mathrm{D}}$.

Imaging Applications Involving Incoherent Sources

For an object plane in the far field of aperture-limited collection optics, the nearly Gaussian-profile BPLO propagates from the receiver aperture according to the condition for maximum Gaussian beam collimation. The waist is at the Fraunhofer plane where the BPLO wavefront is nearly plane, while at the object field the BPLO wavefronts are spherical with radius equal to

$$R(d_o) = d_o\left[1 + \left(\frac{Z_D}{d_o}\right)^2\right] \qquad [14]$$

where we have set $d_o - f \approx d_o$.

It will generally be the case that an imaging system will be used to analyze extended objects whose size exceeds the resolution limit. Following convention, we define a spatially coherent source as one which is illuminated with spatially coherent light with appreciable structure and roughness of depth within the coherence length of the illuminating light, and where the target motion does not appreciably perturb the speckle field during the measurement process. The speckle effects, which so seriously degrade the information content in images formed of spatially coherent objects, are not present in images formed of incoherent objects such as thermal emitters. Speckle may be reduced by time-averaging speckle fluctuations resulting from target motion, atmospheric effects, and independent transmitter pulses, and in the limit such objects become incoherent pseudothermal sources. Similar remarks apply to spatial averages over detector subarrays.

Independent radiators on the object field produce disturbances which can diffract to illuminate the entire receiver aperture as well as a finite coherence area in the image field A_{ci} given by

$$A_{ci} \gg \frac{\lambda^2}{\Omega_{ci}} \qquad [15]$$

where Ω_{ci} is the solid angle subtended by the receiver aperture at the image field. Alternatively, we may interpret A_{ci} as the area of an aperture in the object plane which, if illuminated with wavefronts characterized by eqn [14] (reciprocal illumination), would just fill the receiver aperture with a central diffraction order. Thus the combination of independent disturbances within the detector's field of view on the object plane, combine to form a spatially coherent disturbance over the receiver aperture and within A_{ci}.

According to the Van Cittert–Zernike theorem, the complex degree of spatial coherence μ_s is a unity amplitude wavetrain that propagates exactly as a signal beam resulting from the reciprocal coherent illumination of an object-field aperture of area A_{ci}. Thus, μ_s and the BPLO are Helmholtz-reciprocal pairs whose overlap determines the mixing efficiency with which heterodyne signal is generated from radiation collected from an extended incoherent source. Following convention, we refer to the overlap of μ_s and the BPLO as the effective receiver aperture.

Detectors larger than the aperture-limit, whose area exceeds A_{ci}, exhibit reduced receiver apertures and a corresponding decrease in heterodyne mixing efficiency. For example, varying the receiver pupil, while maintaining constant LO illumination, results in a heterodyne signal proportional to the received power for detector diameters much less than the aperture-limit, while detector diameters much greater than the aperture-limit produce heterodyne signal only in proportion to the square-root of the received power. Thus, for constant LO illumination and incoherent sources, increasing the entendue past the aperture limit will always increase the heterodyne signal, but less than for direct detection.

Optimal Detector Size

As noted above, an aperture-limited system is one where detectors of diameter d are matched with an objective lens with $f/D = d/2\lambda$. Commercially available detectors for use at 10 μm are commonly 100 μm wide suggesting the use of a $f/5$ lens, however aspheric lenses as fast as $f/1$ are also commercially available. Apart from aberrations, it would be advantageous to utilize a 5×5 subarray of 20 μm detectors with an $f/1$ lens in place of a single 100 μm detector for applications involving coherent sources. Applications involving incoherent thermal sources would generally favor a wide detector/fast lens combination since thermal sources provide weak heterodyne signals.

Noise Analysis and Performance

Radar Range Equation

The laser power requirements for the LIDAR system depend on the signal-to-noise ratio of the FPA, the i.f. amplifier performance, atmospheric transmission, and the optics system design. The following calculations are used to determine the performance of the LIDAR system for a single pixel of the FPA. The radar range equation is

$$P_r = P_t\left(\frac{G_t}{4\pi R^2}\right)\left(\frac{\sigma_{pt}}{4\pi R^2}\right)\left(\frac{\pi D^2}{4}\right)\eta_{atm}\eta_{sys} \qquad [16]$$

where P_r = received power; P_t = transmitted power; G_t = antenna gain; σ_{pt} = radar cross-section; R = range; D = transmitting and receiving aperture diameter; η_{atm} = atmospheric transmission; and η_{sys} = optical system efficiency.

The beam divergence of the transceiver is

$$\theta_t = \left[\left(\frac{D}{R}\right)^2 + \left(\frac{\lambda}{D}\right)^2\right]^{0.5} \qquad [17]$$

where λ = wavelength and the antenna gain is

$$G_t = \frac{4\pi}{\theta_t^2} \qquad [18]$$

The radar cross-section for a lambertian diffuse point target is $\sigma_{pt} = 4\rho_t \mathrm{dA}$. The target reflectivity is ρ_t and dA is the target area, which is effectively the size of a pixel of the FPA projected onto the target. For purposes of the following calculations, a surface reflectivity of 0.1 is assumed. The pixel size at the target (assuming Gaussian beam optics) is given by

$$\mathrm{dA} = \pi\left(\frac{D}{3}\right)^2\left[1 + \left[\frac{\lambda R}{\pi}\left(\frac{3}{D}\right)^2\right]^2\right] \qquad [19]$$

where D = transmitter and receiver mirror diameter and it is assumed that the target is larger than dA.

Finally, the signal-to-noise (SNR) ratio for a coherent receiver is given by

$$\mathrm{SNR} = \frac{\eta P_r}{h\nu B_{if}} \qquad [20]$$

where η = detector quantum efficiency; P_r = received power from eqn [16]; h = Planck's constant; ν = frequency of laser; and B_{if} = post-detection bandwidth of signal processor.

Numerical example

Assuming an atmospheric transmission efficiency, η_{atm}, of 98%, an optical system efficiency, η_{sys}, of 30%, and a detector efficiency, η, of 10%, the SNR was calculated as a function of the transmitting/receiving telescope (assumed to be a mirror) diameter, D, with target range as a parameter and plotted in Figure 4. This calculation assumed that the pixel field-of-view (FOV) at the target was illuminated with 1 watt of power. Even at a range of 10 km, the 32 × 32 element imaging LIDAR receiver with a SNR of 100 for each pixel will require only 75 watts (approximately 74-milliwatts per pixel) of total transmitted power for a mirror diameter of 30 cm. From Figure 4, it appears that the telescope mirror size could be reduced to a diameter of 20 cm and still maintain an excellent SNR. However, the SNR is not the only consideration on the transmitting telescope diameter. The cross-range target resolution is determined by the projected pixel size at the target. This target pixel size is plotted versus the telescope mirror diameter in Figure 5. Diffraction from the transmitting and receiving optic sets the limiting resolution that can be achieved with the LIDAR system and is only a function of the diameter and wavelength. A transmit/receive telescope aperture of 30 cm seems to represent a good trade-off between cross-range resolution and SNR.

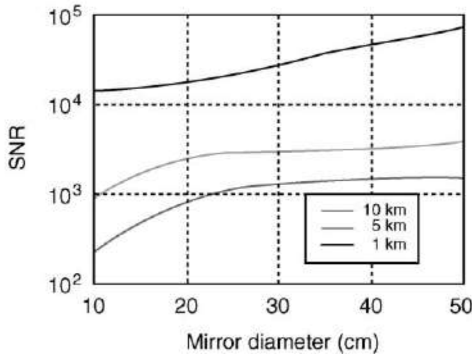

Figure 4 The SNR saturates for a mirror diameter greater than 30 cm at long range.

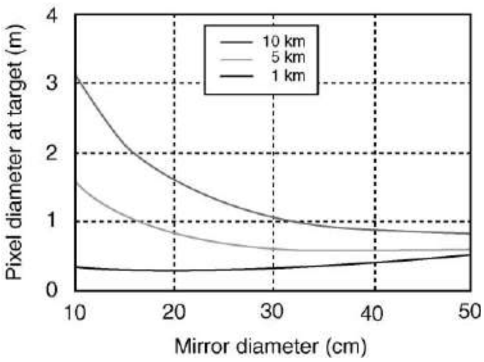

Figure 5 The pixel size determines the target cross-range image resolution.

Applications and Future Directions

The remote sensing community has recognized the advantages of coherent receivers over conventional direct IR detection with respect to increased measurement sensitivity and multidimensional information capability. These advantages, however, come at a price of added system complexity, power, size, weight, and cost. In addition, choices of components, particularly in the long wavelength infrared (LWIR) region of the spectrum, have been limited to liquid nitrogen-cooled nonimaging detectors and table-top-sized gas lasers. Thus, implementations of imaging LIDAR systems have been primarily single detector, single source scanning systems which require high pulse repetition frequency lasers, complex tracking systems, and platforms, and augmentation of automated target recognition input to compensate for multiple, simultaneous measurements on the same target. New technology developments in quantum well infrared photodetector (QWIP) detectors and arrays, and in HgCdZnTe technology are now paving the way toward thermo-electrically cooled LWIR imaging array detectors. Also waveguide, folded cavity, laser resonator designs are providing compact, lightweight packages for CO_2 lasers. Coherent receivers, that use the new focal plane arrays and compact sources, provide the capability of making parallel LIDAR measurements without the added hardware and software overhead of scanning systems. In the following section, several examples are provided that demonstrate the versatility of imaging LIDAR systems that use detector arrays for measurement applications such as chemical plumes, velocity gradients, and speckle reduction.

Wideband heterodyne receivers are useful in measuring chemical species whose absorption peaks are not well aligned with the laser source line. If the species absorption peak corresponds directly with the laser line, an absorption measurement can be made using direct detection (amplitude only). For off-line (off-resonance) measurements, heterodyning is needed, where the bandwidth of the heterodyne receiver dictates how far away from the laser line an absorption peak can be measured. For example, the CO_2 laser line spacing in the 9 to 12 micron region of the spectrum averages around 50 GHz. With the new QWIP detectors theoretically having heterodyne bandwidths of 30 GHz or more, full high-resolution spectral coverage for the long-wave IR is within grasp. For chemical detection with heterodyne receivers, there are two modes of operation. The first is termed passive heterodyne radiometry where the source is a black-body radiator and the absorption is measured for chemical species located between the blackbody and the receiver. The second mode of operation is thermoluminescence where a laser transmitter is used to excite a chemical species in the atmosphere and the resulting emission (luminescence) is measured with the heterodyne receiver. The advantage of imaging in both cases is the ability to capture chemical plumes in the presence of wind and other anomalies.

To illustrate the ability of heterodyne imaging receivers to spectroscopically resolve chemical plumes, Figure 6 shows a series of 10 × 10 pixel images of a small bottle of concentrated ammonium hydroxide against a 212 °C blackbody recorded in both passive heterodyne radiometry and direct detection modes. Figure 6a shows a heterodyne image of the bottle with the top in place where, in Figure 6b the top is removed and the absorbing plume is clearly imaged. The heterodyne images in Figures 6a and 6b were recorded with the LO adjusted to emit the 9R(30) line which is known to

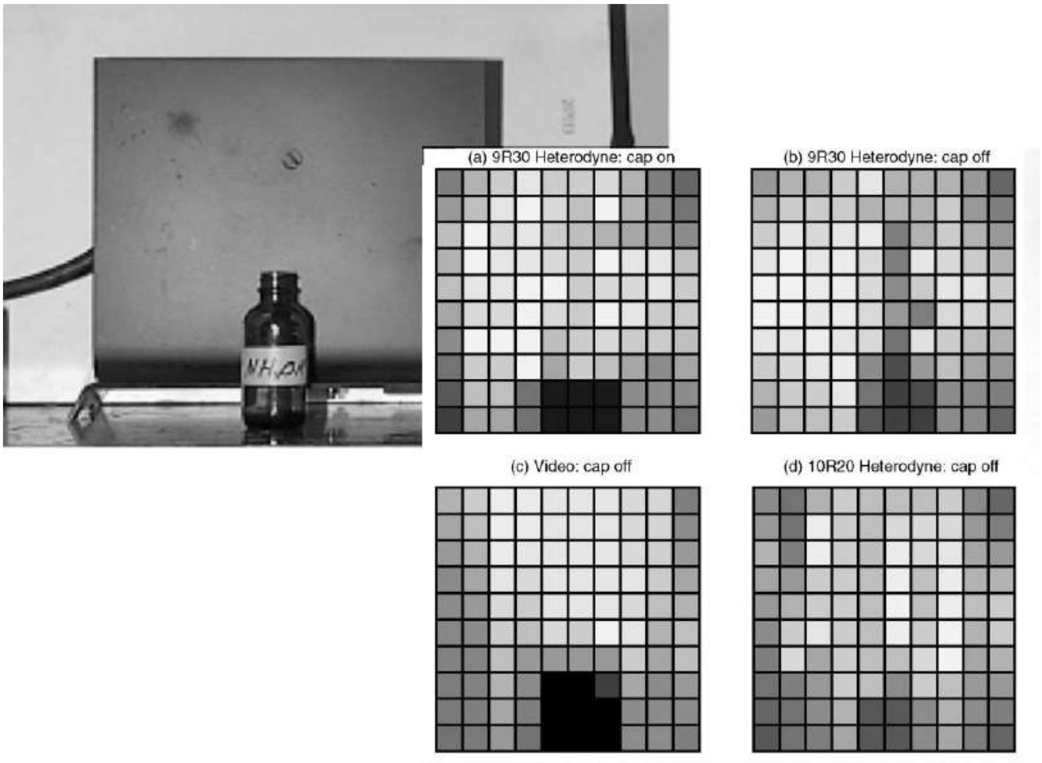
NH4OH
(a) 9R30 Heterodyne: cap on
(b) 9R30 Heterodyne: cap off
(c) Video: cap off
(d) 10R20 Heterodyne: cap off

Figure 6 Ammonia plume characterization.

be absorbed strongly by NH_3. In Figure 6c, the same image was recorded by direct detection and fails to detect the ammonia plume. Similarly, in Figure 6d, the LO was adjusted to emit a nearby line known not to be absorbed by NH_3 and the plume is again not detected. Both the objective and ancillary lens were $f/1$ 38 mm, and the object field was about three meters away. The top of the bottle was approximately 2 cm in diameter and is located at the bottom of the field of view. Measurements utilizing hot ammonium hydroxide imaged against a cold background resulted in similar images (albeit at substantially reduced signal to noise ratios) where the plume was bright and the background was dark.

Another use of imaging LIDAR is the measurement of velocity profiles. Applications include wake vortex measurements for aircraft, wind shear detection, and vibration signature measurements for identification. In another experiment, active mode images were recorded by illuminating the target with a portion of the LO laser beam which had been shifted in frequency by 40 MHz using an acousto-optic modulator. The resulting Doppler image is recorded using an 8 × 8 HgCdTe array with discrete electronics for each pixel element in the array. Figure 7 shows the active image of a vertical squirrel cage fan where each pixel is rendered to represent the peak Doppler shift measured with the detector at the corresponding location in the image plane. The moving target scatters incoherently and thus speckle effects do not degrade the image. The measured Doppler shifts are consistent with the known rotational velocity of the target.

Speckle effects associated with coherent mixing of laser light reflected from an extended target (constructive and destructive interference) are a major source of noise in LIDAR measurements. These speckle effects can be reduced by signal averaging. For single detector LIDAR, averaging over many transmitter pulses can reduce speckle but this technique also reduces temporal resolution. The alternative considered here is to sacrifice spatial resolution by averaging over an array of detectors, all of which record the signal returned from a single transmitter pulse. Detectors separated by a coherence area diameter will record signals that approach statistical independence, and in this case speckle effects can be reduced by an amount approaching the square-root of the number of detectors in the subarray.

The following laboratory experiment demonstrates speckle reduction by spatial averaging with a fixed focal plane array. The receiver used in this experiment consists of a 3 × 3 array of HgCdTe detectors. The 50-micron diameter detectors are arranged in a square pattern, with 100 micron center-to-center spacing. Custom electronics multiplex the output of each detector for subsequent processing. As with the velocity profile experiment, LIDAR illumination is produced by frequency shifting a portion of the LO beam by 40 MHz with an acousto-optic modulator prior to illuminating the object field. Polarization of the illumination beam in the experiment is parallel to the LO prior to scattering from the target. The measured receiver bandwidth, limited mostly by the dewar design, is about 100 MHz, which permits the acquisition of the 40 MHz heterodyne signal. The collection optic consists of a 38 mm $f/2$ asphere which produces a 2 mm diameter pixel image at a 3 m object distance.

The target for this experiment is a 30 cm diameter disk coated with an aluminum powder. The particle size of this powder ranges from 1 to 100 microns as determined by electron microscopy. The disk is attached to a stepper motor that is incremented between frames to provide statistically independent speckle fields. Figure 8 shows the distribution of

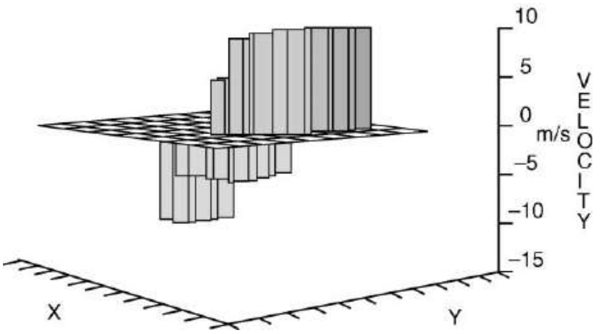

Figure 7 Doppler image of a rotating squirrel cage fan.

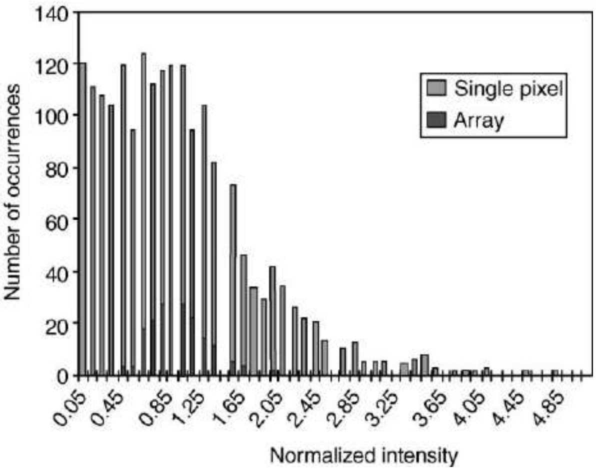

Figure 8 Speckle reduction using spatial averaging over a coherent array.

measurements taken with a single pixel superimposed on the distribution taken by averaging over the focal plane array. A reduction in speckle noise by a factor of the square root of the number of averaged pixels is obtained as predicted.

In this chapter, an overview of imaging LIDAR systems has been presented. Imaging LIDAR system design, performance, and application were discussed. With new technology developments in wide-bandwidth QWIP detector arrays, thermo-electrically cooled detectors, and compact sources, substantial potential exists for imaging LIDAR systems in the next generation in remote sensing. Due to the parallel, multifunction measurement capability of these new receivers a single sensor system can provide three-dimensional information, including range, shape, velocity, and chemical composition. In addition to the sensor fusion advantages, snapshot imaging allows the capture of high-speed, rapidly evolving events that elude present scanning systems.

See also

Environmental Measurements: Doppler lidar. **Imaging:** Infrared Imaging.

Further Reading

Born M and Wolf E (1964) *Principles of Optics*. London: Pergamon.

Forrestor AT (1956) On coherence properties of light waves. *American Journal of Physics* 24: 192–196.

Goodman JW (1975) Statistical properties of laser speckle patterns. In: Dainty JC (ed.) *Laser Speckle and Related Phenomena*, Chap. 2, pp. 9–75. New York: Springer-Verlag.

Haner AB and Isenor NR (1970) Intensity correlations from pseudothermal light sources. *American Journal of Physics* 38: 748–750.

Holzhauser E and Massig JH (1978) An analysis of optical mixing in plasma scattering experiments. *Plasma Physics* 20: 867–877.

Jelalian AV (1992) *Laser Radar Systems*, pp. 3–5. Norwood, WA: Artech House.

Kogelnik H and Li T (1966) Laser beams and resonators. *Applied Optics* 5: 1550–1566.

Kusters JF, Rye BJ and Walker AC (1989) Spatial weighting in laboratory light scattering experiments. *Applied Optics* 28: 657–664.

Martienssen W and Spiller E (1964) Coherence and fluctuations in light beams. *American Journal of Physics* 32: 919–926.

Measures RM (1984) *Laser Remote Sensing Fundamentals and Applications*. New York: John Wiley.

Rye BJ (1979) Antenna parameters for incoherent backscatter heterodyne lidar. *Applied Optics* 18: 1390–1398.

Shannon RR (1997) *The Art and Science of Optical Design*. New York: Cambridge University Press.

Siegman AE (1966) The antenna properties of optical heterodyne receivers. *Applied Optics* 5: 1588–1594.

Simpson ML, Bennett CA, Emery MS, *et al.* (1997) Coherent imaging with two-dimensional focal-plane arrays: design and applications. *Applied Optics* 36: 6913–6920.

Veldkamp WB (1983) Holographic local oscillator beam mutiplexing for array heterodyne detection. *Applied Optics* 22: 891–900.

Veldkamp WB and Van Allen EJ (1983) Binary holographic LO beam multiplexer for IR imaging detector arrays. *Applied Optics* 22: 1497–1506.

Zernike F (1938) The concept of degree of coherence and its application to optical problems. *Physica* V8: 785–795.

Multiplex Imaging

A Lacourt, Université de Franche-Comté, Besancon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ability to project an image onto a screen in a darkened room has been known since antiquity and Leonardo da Vinci (1515) precisely described principles of the Camera oscura. When one exposes a paper sheet directly to sunlight, it is uniformly lighted. However, if one pierces a little hole through the paper sheet then an image of the sun can be displayed on a screen beyond the hole (Figure 1). Similarly, the shape of the light spot we can see in a dark room, from the rays passing through a small aperture of a closed shutter, is not that of the aperture. Generally, the spot has a circular shape, as an image of the sun. Actually, it is neither a shadow of the hole nor an image of the sun but a combination of both. Light incident on an aperture contains the entire information about the light distribution of the object, but this information has to be 'deblurred', or decoded. The aperture acts as an elementary spatial 2D-filter that selects and preserves the direction and intensity of light rays coming from each point of the scene. The projection on the screen of each pencil of light passing through the hole is a small light spot, the geometry of which is the same as the aperture, and

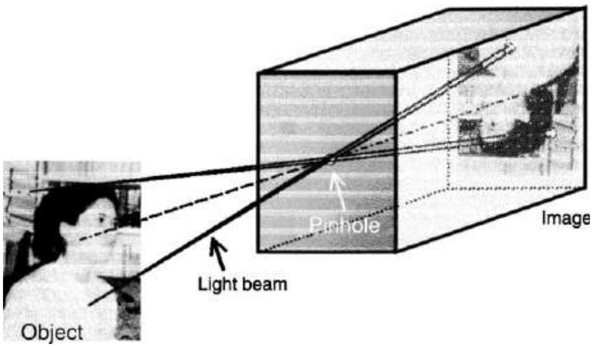

Figure 1 Principle of darkroom.

the intensity is proportional to that of the emitting point. The smaller the hole, the more detail the image exhibits, but with less luminosity. Thus, we must adopt a compromise between the resolution and the illumination of the image.

Now, if the mask has several holes, the amount of light reaching the screen, being proportional to the whole aperture area, is extensively increased. On the other hand, as each hole displays an image on the screen, the overlapping of these multiple images gives a new but blurred image. However, the information exists and we should be able to retrieve it if we know the characteristics of the complex aperture. This is the basic concept of multiplex imaging. The crucial question is: what is a good aperture design and a good way to reconstruct images having high resolution and with high illumination levels? The suitable masks with this aim are named synthetic apertures.

Porta (1589) and Kepler (1611) found the perfect solution, by setting up a lens in front of a large hole. Indeed, one can regard a lens as a wide aperture made of continuously juxtaposed microprisms (Figure 2). The lens obtains such refractive properties that the image displayed by each of these microprisms is moved toward a center. Moreover, as the angle of the microprisms varies continuously, it focuses on a single point, all the rays coming from a given object point. Thus, the blurring is suppressed and the illumination is increased. The way was opened toward the invention of optical instruments and photography and the 'pinhole' has fallen into oblivion for a long time.

Two new problems awake interest in synthetic apertures, around 1960. First, the domain of imaging has extended out of visible light: infrared and ultraviolet light, γ-rays, X-rays, ultrasound waves, and others. Materials suitable for making lenses or mirrors, operating with such waves, are expensive or often do not exist. For example, most of the transparent materials have a refractive index close to 1 for γ-rays and X-rays. However, some are opaque enough to enable the use of masks.

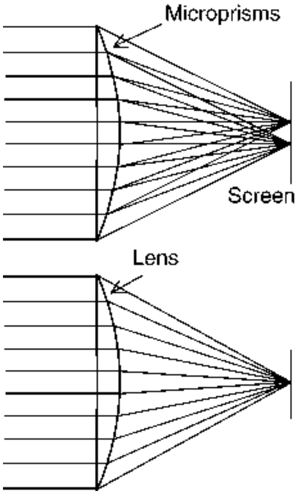

Figure 2 A lens can be modeled as a juxtaposition of microprisms with continously variable angles.

The second challenge appeared with progress in astronomy. In order to explore the universe more, further from Earth and closes to the Big Bang, astrophysicists needed telescopes providing very high luminosity and high resolution, since distant stars, nebulae, and other galaxies are extremely weak and small luminous objects. This exploration can be achieved by increasing the diameter of the primary mirror of telescopes. Indeed, due to diffraction of light by the pupil of the primary mirror, the theoretical limit of resolution of a telescope (i.e., the smaller distance between two points that one can discriminate), is inversely proportional to the diameter of the mirror. In addition, the illumination is proportional to its area. However, the working of large mirrors is a technical prowess that is very expensive and has high risks. Defect of a fraction of micrometer when polishing the surface can have dramatic consequences, e.g., Hubble. Also, such mirrors, being heavy, can bend out of shape under their own weight. Nevertheless, outstandingly large telescopes have been implemented, based on Adaptive optics: the deformations of the mirror are measured in real time by an interference method and compensated by activating a mosaic of jacks (New Technology Telescope – NTT – European program, 1989). Another way consists in replacing the solid-state mirror by a mosaic of 36 mirrors accurately directed (Keck telescope, USA). However, in either case, the mirror diameter does not exceed 8 to 10 meters. Synthetic apertures could be a solution to that difficulty, as earlier demonstrated by Fizeau (1868) and Michelson (1890).

In the first part of this article, any fundamentals are briefly recalled about properties of light and imaging.

The second part describes recent experimental methods and results obtained in synthetic aperture imaging. The last part reports the important programs in progress today relating to stellar interferometry.

Background

Electromagnetic Waves Propagation

Light power, which is involved in image formation, is carried by the electric field of an electromagnetic wave (Figure 3). The electric field $\vec{E}$ has properties of a vector orthogonal to the direction of propagation $\vec{N}$ of the light ray. It is characterized by its amplitude $\vec{E}_0$, the angular time frequency of vibration ω, and the wavevector $\vec{k}$, which specifies the propagating direction of the ray. Let $P(\vec{d})$ be a current point on a plane wavefront, defined by the position vector $\vec{d}$ from the origin of the reference frame. The electric field propagation of a monochromatic wave is expressed as:

$$\vec{E}(\vec{d},t)=\vec{E}_0\exp[\mathrm{i}(\vec{k}\cdot\vec{d}-\omega t)] \qquad [1]$$

The light power carried by the electromagnetic field is proportional to $I=\|\vec{E}_0\|^2=\vec{E}\cdot\vec{E}^*$ where $\vec{E}^*$ is the complex conjugate of $\vec{E}$. The time frequency $\nu=\omega/2\pi$, or the wavelength $\lambda=c/\nu$, where c is the electromagnetic wave celerity, characterize the hue of the light. The wavevector $\vec{k}=\vec{N}\omega/c$, where $\|\vec{N}\|=1$ is an angular space frequency vector, as $\vec{n}=\vec{N}/\lambda=\vec{N}\nu/c$ is a space frequency vector. We have $\vec{k}=2\pi\vec{n}$. As λ is a constant for a given monochromatic light, $\vec{n}$ defines the direction of propagation of a ray.

Spatially Coherent Light

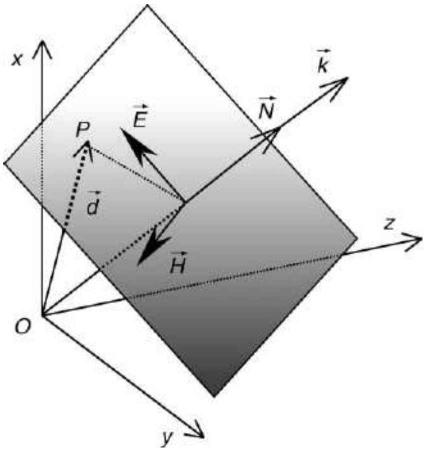

Figure 3 Electromagnetic wave propagation.

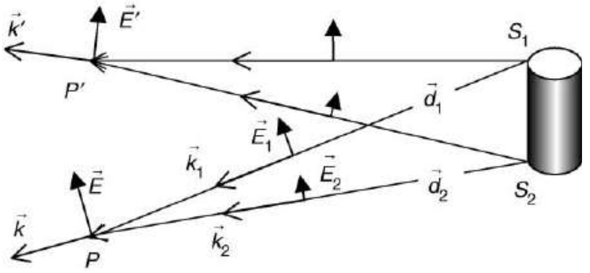

Figure 4 Spatially coherent light.

Consider a very small source stated as the point source S_1, emitting uniform lightwaves in any direction (Figure 4). The wave reaching a farfield point, P, from S_1, propagates along a direction defined by the wavevector $\vec{k}_1$. The vectors $\vec{k}_1$ and $\vec{d}_1=\overrightarrow{S_1P}$ are collinear. Another point source, S_2 is characterized from P by the wavevector $\vec{k}_2$, such that $\vec{k}_2$ and $\vec{d}_2=\overrightarrow{S_2P}$ are collinear. When the distances d_1 and d_2 are close to each other, the amplitudes $\vec{E}_{01}$ and $\vec{E}_{02}$ of the waves at P are proportional to those of the corresponding emitting sources, S_1 and S_2. According to the principle of superposition, the resulting field at the point P, $\vec{E}=\vec{E}_1+\vec{E}_2$, and the instant wavevector $\vec{k}$, are unique vectors.

If S_1 and S_2 are two synchronous sources, the amplitude $\vec{E}_0$ of $\vec{E}$ and the wavevector $\vec{k}$ are constant versus time at any point P. This is still the case when an extended object is made of an arrangement of synchronous points. That situation defines a spatially coherent object. Thus, one cannot recover multiple data about the object's distribution, from the field framework at a single point P, since the relationship between the two is multi-unequivocal. Nevertheless, due to diffraction, one can retrieve them by knowing the complex amplitude distribution of light on a set of wavevectors available over an extensive area of a screen. Let $\vec{E}_0(\vec{r})$ be the amplitude of the electric field at the object point $M(\vec{r})$ and let $\vec{E}_0(\vec{n})$ be the one diffracted by the object to farfield toward the direction corresponding to the space frequency $\vec{n}$. Fourier transforms give the connection between $\vec{E}_0(\vec{r})$, and $\vec{E}_0(\vec{n})$:

$$\left\{\begin{array}{l}\vec{E}_0(\vec{n})=\displaystyle\int\vec{E}_0(\vec{r})\exp[\mathrm{i}2\pi\vec{n}\cdot\vec{r}]\mathrm{d}\vec{r},\\[2ex] \vec{E}_0(\vec{r})=\displaystyle\int\vec{E}_0(\vec{n})\exp[-\mathrm{i}2\pi\vec{n}\cdot\vec{r}]\mathrm{d}\vec{n}.\end{array}\right\} \qquad [2]$$

This is the situation encountered when an object is illuminated by a laser or by a point source such as a single star.

Spatially Incoherent Light

When the point sources S_1 and S_2 are independent – spatially incoherent – both wavevectors $\vec{k}_1$ and $\vec{k}_2$

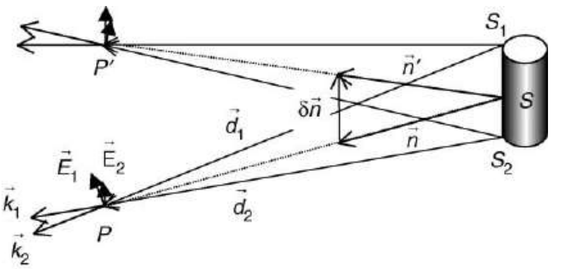

Figure 5 Spatially incoherent light.

coexist at the point P, each characterizing the angular position of S_1 and S_2 (Figure 5). If S_1 and S_2 are part of an extended intensity distribution $I_0(\vec{r})$, fans of wavevectors and fields are mixed at P, each coming from an independent point source S. Together, they carry the whole data on the object distribution. Those data are partially uncorrelated at any two points P and P' on a screen, related to the different optical paths from any point S to P and P' and to the finite time of coherence of light. The correlation between data available to farfield in two directions $\vec{n}$ and $\vec{n}'$ from the object, is given by the mutual intensity function, $\Gamma(\delta\vec{n}) = \int \vec{E}(\vec{n})\cdot\vec{E}^*(\vec{n}+\delta\vec{n})\mathrm{d}\vec{n}$, where $\delta\vec{n} = \vec{n}' - \vec{n}$. Fourier transforms give the relation between $I_0(\vec{r})$ and $\Gamma(\delta\vec{n})$:

$$\left\{ \begin{array}{l} \Gamma(\delta\vec{n}) = \displaystyle\int I(\vec{r}\,)\exp[\mathrm{i}2\pi\delta\vec{n}\cdot\vec{r}]\mathrm{d}\vec{r}, \\ I(\vec{r}) = \displaystyle\int \Gamma(\delta\vec{n})\exp[-\mathrm{i}2\pi\delta\vec{n}\cdot\vec{r}\,]\mathrm{d}\delta\vec{n} \end{array} \right\} \qquad [3]$$

Therefore, by drawing a map of $\Gamma(\delta\vec{n})$, one can recover the object distribution $I_0(\vec{r}\,)$ by a Fourier transformation. Most of time, one uses the spatial degree of coherence function, $\gamma(\delta\vec{n}) = \Gamma(\delta\vec{n})/\Gamma(\vec{0})$. When the object illuminates a screen, at a large distance d, the degree of coherence of light between two points P and P' on the screen is $\gamma(\delta\vec{\rho}/\lambda d)$, with $\delta\vec{\rho} = \overrightarrow{PP'}$. The 2D extension of $\gamma(\delta\vec{\rho}/\lambda d)$ defines the area of coherence of light on the screen. The width of the area of coherence, $\Delta\rho$, and that of the object distribution, Δr, are inversely proportional:

$$\Delta\rho\cdot\Delta r \sim \lambda d \qquad [4]$$

If a mask perforated by two pinholes, separated by a distance $\delta\vec{\rho}$, replaces the screen, one can observe Young fringes of interference beyond the mask, the visibility of them being precisely $\gamma(\delta\vec{\rho}/\lambda d)$. Therefore, one can built a map of $\gamma(\delta\vec{\rho}/\lambda d)$ versus $\delta\vec{\rho}$ from a set of 2D measurements of visibilities of fringes, then reconstruct the object distribution by a Fourier transformation.

Notice that the entire data of the object distribution is available from a single coherence area. That is, a pinhole covering exactly one area of coherence would select the entire available data from the object and would display to farfield the most accurate image possible. Generally, such a pinhole is so small that practically no light passes through it. However, we can regard each area of coherence as an independent channel of information. Then an imaging device based on a juxtaposition of such reduntant channels can be considered as a multiplex imaging device.

Self-luminous objects such as extended or double stars, and artificial thermic sources are spatially incoherent sources.

Multiplex Imaging

Methods of multiplex – or lensless – imaging are performed either with coherent of incoherent light. The very different properties of the two types of light leads to completely different methods, even if all are related to holography. Holography with coherent light is an interference method for local recording of both the direction and modulus of the wavevectors and the amplitude of the field contained in many juxtaposed grains of speckle, scattered by the object. Each grain is an independent channel of imaging.

Synthetic Apertures with Spatially Incoherent Objects

As interferences are impossible by this method with incoherent light, we must improve other methods of imaging. Synthetic aperture is a concept consisting of using a redundant multichannel system to increase the luminosity of images without loss of resolution. The basic principle is that of a multiple holes darkroom. The processing is in two steps: recording and reading the image. The recording step consists in illuminating a photosensitive receptor with the object through a mask (Figure 6). Generally, the receptor is a

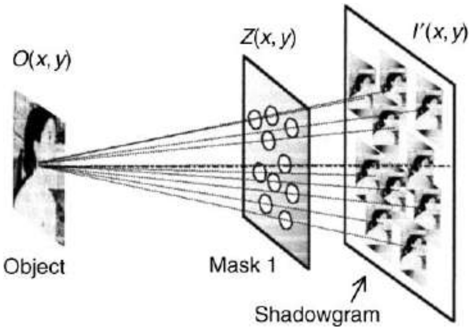

Figure 6 Principle of synthetic aperture: recording.

photographic plate. Obviously, the light distribution on the plate is blurred. Indeed, any elementary transparencies on the mask project on the receptor inverted images of the object located the shadow of the mask. Such a distribution is expressed by a convolution. Let $O(x, y)$ be the conical projection of the object on the receptor and $Z(x, y)$ that of the mask. The shadows distribution on the receptor is (the variable being normalized):

$$
\begin{aligned}
I'(x', y') &= Z(x, y) * O(x, y) \\
&= \int Z(x, y) O(x' - x, y' - y) \mathrm{d}x \, \mathrm{d}y \qquad [5]
\end{aligned}
$$

The reading step is a decorrelation processing from the recorded intensity (Figure 7). It consists in achieving a new correlation of I' with an arbitrary function T:

$$
\begin{aligned}
I''(x, y) &= T(x, y) * [Z(x, y) * O(x, y)] \\
&= [T(x, y) * Z(x, y)] * O(x, y) \qquad [6]
\end{aligned}
$$

$R[x, y] = [T(x, y) * Z(x, y)]$ is the system point spread function (SPSF) of the processing. It expresses the image distribution of a single-point. The perfect image is obtained when $R(x, y)$ is a 2D *Dirac* – or *delta*– distribution, $\delta(x, y)$:

$$
\begin{aligned}
R(x, y) = \delta(x, y), \quad \text{as} \quad &\int \delta(x, y) O(x' - x, y' - y) \\
&\times \mathrm{d}x \, \mathrm{d}y = O(x', y') \qquad [7]
\end{aligned}
$$

Fresnel zone plates

First Mertz and Young used Fresnel zone plate (FZP) as a synthetic aperture for X-ray imaging purposes. FZPs are made of alternate transparent and opaque concentric crowns of equal area. The radius of the nth circle is $r_1\sqrt{n}$, r_1 being the smallest. When illuminated by a cylindrical light beam, the FZP is a multiple foci diffractive lens (Figure 8).

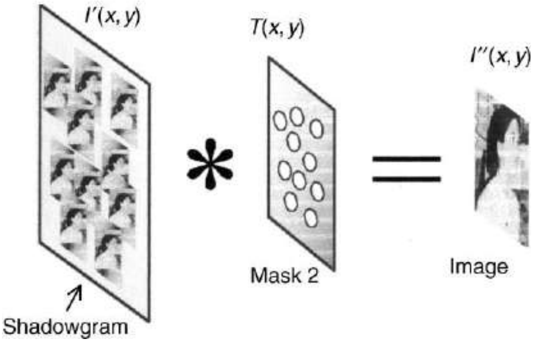

Figure 7 Synthetic aperture: reading.

The reading step is either analogical or numerical. Through the analogical method, we consider the recorded plate as a pseudohologram (Figure 9). When the 'shadowgram' is lighted with a parallel beam of laser light, the FZPs recorded around each point concentrates the light on its foci. Then an image of the object is reconstructed on the focal planes; a method that is simple and elegant. Nevertheless, the processing is not linear since we record an intensity distribution at the recording step, while the reading step by diffraction involves amplitude. Moreover, the S/N ratio decreases as the amount of data increases, recorded on the photographic plate because of its finite dynamic.

Mertz and Young had the idea of using a second FZP to perform the deconvolution processing by optical methods. In this case, the SPSF, $R(x, y)$, is the autocorrelation function of the FZP transparency (Figure 10). It presents a narrow central peak, emerging from a pyramidal ground (Figure 11). The optical setup is the same as the recording step.

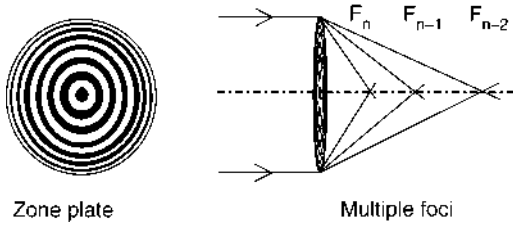

Figure 8 Properties of FZP.

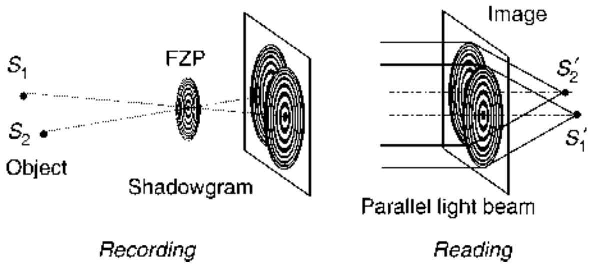

Figure 9 Recording and analogical reading with a FZP.

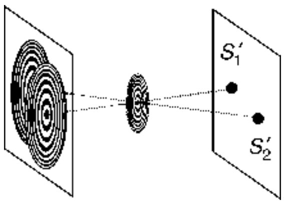

Figure 10 Optical decorrelation.

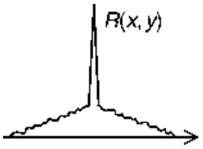

Figure 11 System point spread function with a Fresnel zone plate.

The FZP scale factor only must be adjusted to compensate for the magnification ratio generated by the conical projection at the recording step.

In either digital or analogical deconvolution, the image reconstructed is accurate, with good sharpness, if the number of zones is sufficient. Ceglio *et al.* obtained relevant images of 8 μm targets with FZP having 100 zones. Nevertheless, the multiple foci of FZP induces decreasing of the S/N ratio which is an important limitation. Besides, we pointed out above that the larger the object, the more the decrease in the S/N ratio.

Random and uniformly redundant arrays

Dicke and Ables substituted a random array to the FZP. The mask is made of randomly distributed holes and the expected advantage is doubled. First, one can reduce the transparent area of the mask by less than 50% in order to increase the S/N ratio when the object is wide. Second, the autocorrelation function of such an array is theoretically a Dirac distribution when its width is infinite. Thus, one can hope to narrow the system point response. However, when the width is finite, the pyramidal ground of the point response mentioned above is present and the advantage is not clearly demonstrated. In order to correct this effect, Chris Brown used a 'mismatched' function, for the postprocessing deconvolution. This function is calculated, first by achieving the complement of the recording mask function, then subtracting from it. Thus, the 1 transparencies result in +1 while the 0 transparencies are changed by −1. Brown shows images digitally reconstructed of X-ray sources, 30 sec. of arc wide. Fenimore and Cannon conceived a uniformly redundant array. The center random motif of the mask is half replicated on each side with only the center area being used for the recording step. The deconvolution is calculated by taking into account the whole area in a mismatched way. The result is an exact delta function. The authors show images of 200 μm-wide microspheres with a similar resolution but a signal 7100 times stronger than with a comparable pinhole camera.

Theta Rotating Interferometer

The theta shearing interferometer enables us to record complete data from the Fourier transform of a spatially incoherent object (Figure 12). The source can be any natural one, but monochromatic. This is easily practicable using a good filter or monochromator.

The device is based on a Michelson interferometer, the plane mirrors being replaced by two roof prisms. One of the roof prisms is fixed so the other can rotate around the axis of the interferometer. As a result, we have two images of the object. Let θ be the angle between the two edges of the prisms. The angle between the images is 2θ. Since the original object is noncoherent, and two homologous points being mutually coherent, the light they emit interferes, resulting in fringes.

The intensity, space frequency, and direction of the fringes characterize the intensity and the position of the point source. One shows that the farfield visibility distribution of fringes $V(\delta\vec{n})$ and the object distribution $O(\vec{r})$ are related by Fourier transforms:

$$\left\{\begin{array}{l} V(\delta\vec{n}) = \int O(\vec{r})\exp\left[i2\pi\frac{\delta\vec{n}\cdot\vec{r}}{2\sin\theta}\right]\frac{d\vec{r}}{2\sin\theta}, \\ O(\vec{r}) = \int V(\delta\vec{n})\exp\left[-i2\pi\frac{\delta\vec{n}\cdot\vec{r}}{2\sin\theta}\right]d\delta\vec{n} \end{array}\right\} \qquad [8]$$

Then, one can see that the visibility of the interference pattern is an image of the degree of coherence $\gamma(\delta\vec{n})$, with a scale factor $1/2\sin\theta$.

Such a pattern recorded on a photographic plate is a Fourier hologram of the noncoherent object. An image can be reconstructed by diffraction at infinity or near a focus (see Figure 13, examples performed in 1972 by the author). The technique was successfully applied to the determination of the modulation transfer function of optical systems. C Roddier and F Roddier applied the method to astronomy.

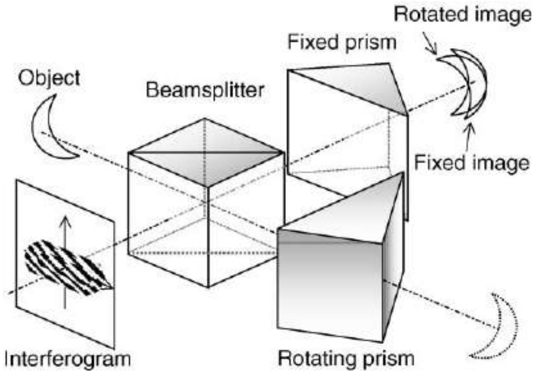

Figure 12 Theta shearing interferometer.

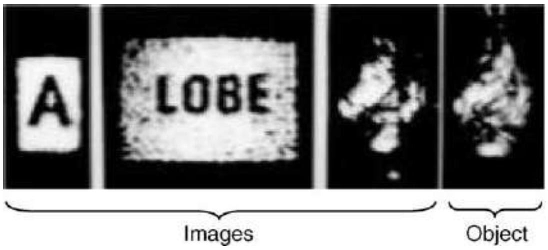

Figure 13 Incoherent light holograms.

For instance, they examined Alpha Orionis and discovered the evolution of its dust envelope and the possible existence of a companion.

High Resolution Imaging in Astronomy

In astronomy, the limit of resolution depends chiefly on two factors. The first is a technical one. Adaptive optics and very large telescopes (VLT) brought spectacular progress to clarity and resolution of telescopes. The 8.2 m VLT of Kueyen set up in 1989, shows images easily resolved of HIC 69495 Centauri, separated by 0.12 sec. of arc. Nevertheless, working of large mirrors for telescopes reaches its limits. Therefore, their clarity remains too weak. On the other hand, atmospheric turbulences make the refractive index of air unstable and the position of the images in the focal plane of telescopes unsettled. Thus, the point response is a granular spot named ‘speckle’. Short exposure times relieve that inconvenience, but they prohibit the observation of weak stars. In 1970, Labeyrie proposed a method of speckle interferometry consisting of a statistical analysis of a set of speckles. The result is a display of a diffraction-limited image. He resolved many tens of single and double stars. Only Hubble, above the atmosphere, is able to do it by direct imaging. However, the primary mirror of Hubble being only 2.40 m wide, has clarity much lower. Stellar interferometry is a very promising alternative solution.

Principles of Stellar Interferometry

Stellar interferometry is based on eqn [3]. The light emitted by double or wide stars is noncoherent, so that one can access an image of them trough 2D measurements of the degree of coherence of light. We showed above that one could achieve this by using interference devices such as Young holes. Eqn [3] shows that if we measure the visibility of Young fringes at increasing distances $\delta\rho$ between the pinholes, the distance $\Delta\rho$ from which the visibility is zero gives the angular width of the star, $\Delta\alpha$, under the formula:

$$\Delta\rho{\cdot}\Delta\alpha \sim \lambda \qquad [9]$$

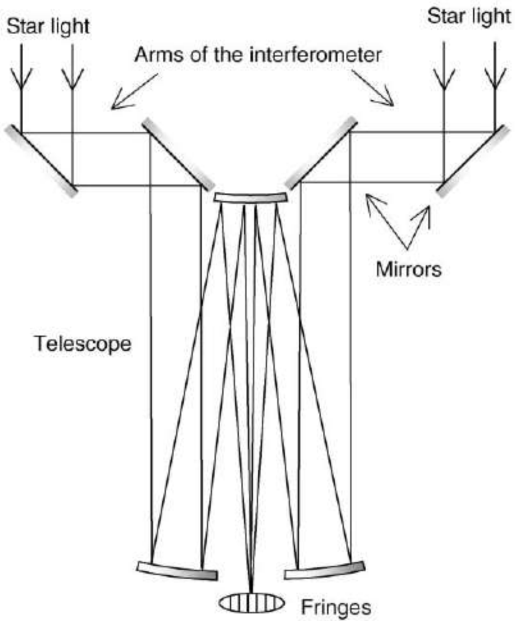

Figure 14 Stellar interferometer of Michelson.

where $\Delta\alpha = \Delta r/d$. In the case of a double star, the visibility varies periodically versus $\delta\rho$ when the pinholes are aligned along the axis of two stars. Let $\Delta\alpha$ be the angular distance between the stars: the smallest distance $\Delta\rho$ for which the visibility is zero is given by:

$$\Delta\rho{\cdot}\Delta\alpha = \lambda/2 \qquad [10]$$

The larger the uttermost distance between the holes, the more the limit of resolution of the processing is potentially sharp.

In order to exploit this idea, Stephan placed and rotated a mask with two holes in front of the objective of a refractive telescope. He established that the diameter of the stars he observed was much less than 0.16 sec. of arc. Then, Michelson built up, on top of a 30 cm refractive telescope, a stellar interferometer consisting of two arms catching the light at 7 m of distance from each other (Figure 14). He measured correctly the diameter of the satellites of Jupiter. Hale carried the distance of the base up to 15 m in 1920.

The Two Telescopes Interferometer of Labeyrie

In 1975, Labeyrie improved the idea by setting up two separate 25 cm diameter Cassegrain-coudé telescopes on a 12 m baseline, oriented north–south (Figure 15). The setting and mechanical stability of the device are extremely sensitive. The tolerances do not exceed any μm. A mobile table near the common focal plane of the two telescopes compensates for the optical path differences due to stars tracking. In spite

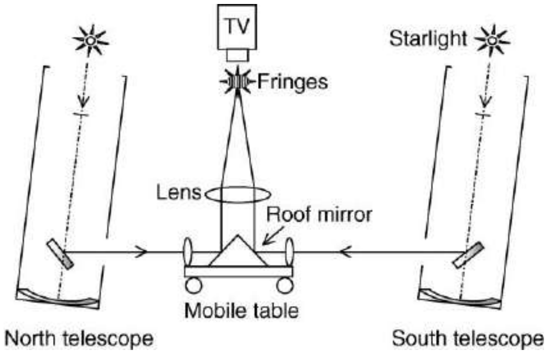

Figure 15 The two telescopes interferometer of Labeyrie.

of the acute difficulties, he succeeded in observing fringes on Vega. He obtained a resolution of 0.01 sec. of arc versus 0.5 sec. with a single 25 cm telescope. Later, one telescope was rendered mobile bringing the base width variable between 4 and 67 m. The resolution reaches up to 0.00015 sec. of arc. Note that the rotation of Earth also enables east–west analysis with resolution of 0.005 sec., and today, the 25 cm telescopes are replaced by 1.5 m ones.

The Very Large Telescope Interferometer (VLTI)

The two telescopes interferometer of Labeyrie is a first step toward a very ambitious project: the telescope at synthetic aperture and very large base. Proofs of reliability of such a principle was already given by the very large array (VLA) experimented in radio astronomy. Labeyrie with Lena, and the team of the ESO (Ecole Supérieure d'Optique, Paris), imagined a very large telescope interferometer (VLTI) made of 6 telescopes of 2 m, in a 2D structure on a base of 100 m. The European VLTI was built at Cerro Paranal. It is made of four 8.2 m telescopes corrected in real time by adaptive optics (NTT technology) and three 1.8 m telescopes. The fourth 8.2 m telescope was put in service in 2001. Some days, each of the 8.2 m telescopes used separately gives a resolution close to that of Hubble. However, the clarity is much better because of the larger primary mirror diameter (8.2 m against 2.4 m). In May 2003, the team of J-G Cuby observed a galaxy the farthest away ever seen, only 900 millions years old since the Big Bang. Obviously, a drastic gain of resolution (by a factor 10) is expected from the interference configuration.

See also

Coherent Lightwave Systems. **Diffraction:** Fresnel Diffraction. **Imaging:** Adaptive Optics; Interferometric imaging. **Instrumentation:** Telescopes. **Interferometry:** Overview.

Further Reading

Labeyrie A (1976) High resolution techniques in optical astronomy. In: Wolf E (ed.) Progress in Optics, vol. XIV, pp. 49–85. New York: North-Holland.

Mertz L (1965) *Transformations in Optics*. Amsterdam, Oxford: John Wiley & Sons, Inc.

Photon Density Wave Imaging

V Toronov, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Photon density wave imaging (PDWI) employs diffused light, usually in the near-infrared band, for imaging inhomogeneous structures in a highly scattering medium. The other names for PDWI are diffuse optical tomography and photon migration imaging. The term photon density wave (PDW) refers to the wave of light density spreading in the highly scattering medium from the source whose power changes in time. Therefore, rigorously speaking, photon density wave imaging uses intensity-modulated light sources, such as pulsed lasers, or sources whose power is harmonically modulated. However, many applications employ steady-state light sources. Most applications of PDWI are in the field of imaging of biological tissues, including medicine and cognitive neuroscience.

As in the other imaging methods, such as x-ray, computed tomography (CT) or ultrasound imaging, the problem of PDWI consists in the reconstruction of the internal structure of the medium from the measurements made outside the medium (on its surface). In the case of PDWI the internal structure of the medium is represented by the spatial distribution of its optical properties. The reconstruction procedure is essentially an iterative mathematical algorithm, which includes the modeling of the light propagation inside the medium assuming a certain distribution of the optical properties, and the correction of this distribution to match the model prediction with the actual measurement. The modeling part of the imaging problem is usually referred to as the forward problem, and the correction procedure is known as an inverse problem.

Forward Problem: Photon Density Waves

Unlike optically transparent media, in a highly scattering medium the coherence of light from the coherent source very rapidly decays with distance. This happens due to the interference of many randomly and multiply scattered light waves at every point of the medium that is distanced from the source much farther than the photon mean free path l, which is the mean distance at which a light wave undergoes a single scattering event. For biological tissues, where light scattering occurs on cell membranes and other heterogeneities of the index of refraction, l is typically about 1 mm. Although high scattering eliminates the coherence of the electromagnetic waves at the short time intervals corresponding to oscillation at optical frequencies ($\sim 10^{-15}$ s), in the case of an intensity-modulated light source, the macroscopic wave of the density of light energy, called 'photon density wave', forms in the highly scattering medium. The properties of PDW depend on the type of the source modulation and on the local optical properties of the medium, which are the scattering coefficient μ_s [cm^{-1}] and the absorption coefficient μ_a [cm^{-1}].

Because of the complexity and incoherence of the macroscopic electromagnetic field in highly scattering media, i.e. media with $\mu_a \ll \mu_s$, the fundamental Maxwell equations of electrodynamics do not provide a convenient approach to the solution of the problem of light propagation in such media at distances much larger than l. Therefore, statistical approaches have been developed, such as the photon transport model and its diffusion approximation. In the photon transport model the light is characterized by the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-2} steradian^{-1}] (where the steradian is the unit solid angle), which is the power of radiation near the point $\mathbf{r}$ propagating within a small solid angle around the direction $\hat{\Omega}$ at time t. The photon transport model accounts for the possible nonisotropy of the radiation flux. For the given spatio-temporal distribution of the optical properties, μ_a and μ_s, and for the given spatial and angular distribution of light source power $S(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-3} steradian^{-1}] the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ can be found from the linear radiation transfer equation, also called the photon transport equation, or simply the transport equation (TE):

$$\frac{1}{v}\frac{\partial L(\mathbf{r}, \hat{\Omega}, t)}{\partial t} + \nabla \cdot L(\mathbf{r}, \hat{\Omega}, t)\hat{\Omega} + (\mu_s + \mu_a)L(\mathbf{r}, \hat{\Omega}, t) = \mu_s \int L(\mathbf{r}, \hat{\Omega}', t) f(\hat{\Omega}, \hat{\Omega}')\mathrm{d}\hat{\Omega}' + S(\mathbf{r}, \hat{\Omega}, t) \qquad [1]$$

where $f(\hat{\Omega}, \hat{\Omega}')$ is the phase function, which represents the portion of light energy scattered from the direction $\hat{\Omega}$ to the direction $\hat{\Omega}'$, and v is the speed of light in the medium. TE is equivalent to the Boltzmann transport equation describing the transport of particles undergoing scattering.

Equation [1] is difficult to solve even numerically. Therefore, a hierarchy of approximations to the PTM was developed. This hierarchy, known as the P_N approximation, is based on the expansion of the radiance in a series of spherical harmonics $Y_{nm}(\hat{\Omega})$ truncated at $n = N$. In the P_1 approximation the coefficients of the series are proportional to the radiation fluence

$$\Phi(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t)\mathrm{d}\hat{\Omega}, \qquad [2]$$

and the flux

$$J(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t)\hat{\Omega}\mathrm{d}\hat{\Omega}, \qquad [3]$$

both having units of [W cm^{-2}]. By substituting the P_1 series for $L(\mathbf{r}, \hat{\Omega}, t)$ and $S(\mathbf{r}, \hat{\Omega}, t)$ into eqn [1] and assuming isotropic light sources and sufficiently slow change of source intensity, one can obtain the equation that includes only $\Phi(\mathbf{r}, t)$ as the unknown. Usually, this equation

$$v\mu_a U(\mathbf{r}, t) + \frac{\partial U(\mathbf{r}, t)}{\partial t} - D\Delta U(\mathbf{r}, t) = Q(\mathbf{r}, t) \qquad [4]$$

is written in terms of the value $U(\mathbf{r}, t) = \frac{1}{v}\Phi(\mathbf{r}, t)$ [J cm^{-3}], which has units of energy density, and is called the photon density. In eqn [4] $D = v/(3\mu_a + 3\mu_s')$ [cm^2/s] is the light diffusion coefficient, and $\mu_s' = \mu_s(1 - g_1)$ [cm^{-1}] is the reduced scattering coefficient; g_1 is the average cosine of the scattering angle. The source term $Q(\mathbf{r}, t)$ [J cm^{-3} s^{-1}] in the right-hand side of the equation [4] describes the spatial distribution of light sources (in the isotropic approximation, i.e., neglecting the dipole and higher momenta of $S(\mathbf{r}, \hat{\Omega}, t)$) and the temporal modulation of source intensity.

Equation [4] is mathematically equivalent to the equation describing the density of the medium consisting of the particles undergoing linear diffusion. Therefore, it is called the diffusion equation (DE), and the P_1 approximation is also known as the diffusion approximation. The diffusion approximation is valid under the following three conditions: (a) the light sources are isotropic or the distance between the source and the detector is much larger than l; (b) characteristic source modulation frequencies ν are low enough that $v\mu_s'/(2\pi\nu) \gg 1$; and (c) $\mu_a \ll \mu_s'$. In the biomedical applications of PDWI the condition (a) is fulfilled when the source–detector distance is larger than 1 cm. The condition (b) is valid when the characteristic modulation frequencies of the light source are less than 1 GHz. The condition (c) is fulfilled in most biological tissues, for which the scattering is highly isotropic, i.e. $g_1 \ll 1$. Thus, the forward problem of PDWI, i.e., the modeling of light propagation, can be solved using the diffusion approximation in a large variety of situations.

PDWs described by equation [4] are the scalar damped traveling waves, characterized by the coherent changes of light energy (photon density) in different locations on time intervals from microseconds to nanoseconds. Usually pulsed or harmonically modulated light sources are used to generate PDWs. The properties of the PDWs from a harmonic source are fundamental, because the PDW from the source of any type of modulation can be considered as a superposition of harmonic waves from sources having different frequencies using the Fourier transform.

In order to get an idea of the basic properties of PDW, one can consider the analytical solution of [4] for the case of a harmonically modulated point-like source placed in an infinite homogeneous scattering medium. This solution is

$$U(r,t) = U_{dc}(r) + U_{ac}(r)\exp[-i\omega t + i\phi(r)] \qquad [5]$$

where U_{ac}, U_{dc}, and ϕ are the ac, dc and phase, respectively, of the wave measured at the distance r from the source, and $\omega = 2\pi\nu$. U_{ac}, U_{dc}, and ϕ depend on the medium optical properties, source intensity Q_0, modulation amplitude A_0, and initial phase ψ_0 as

$$U_{ac}(r) = \frac{Q_0 A_0}{4\pi D} \times \frac{\exp\left[-r\left(\frac{v\mu_a}{2D}\right)^{1/2}((1+x^2)^{1/2}+1)^{1/2}\right]}{r} \qquad [6a]$$

$$U_{dc}(r) = \frac{Q_0}{4\pi D}\;\frac{\exp\left[-r\left(\frac{v\mu_a}{D}\right)^{1/2}\right]}{r} \qquad [6b]$$

$$\phi(r) = r\left(\frac{v\mu_a}{2D}\right)^{1/2}\sqrt{(1+x^2)^{1/2}-1} \qquad [6c]$$

where $x = \omega/(v\mu_a)$. Equations [5]–[6] describe a spherical wave whose wavelength at $\mu_a = 0.1\,\text{cm}^{-1}$ and $\mu_s' = 8\,\text{cm}^{-1}$ (typical for biological tissues) and at modulation frequency $\nu = 100$ MHz, is close to 35 cm, so that the phase of the wave changes approximately by 10° per cm. In the same time, the wave decays exponentially at a distance of about 3 cm, so that PDWs are strongly damped. A similar analytical solution in the form of a spherical wave can be obtained for the case of a semi-infinite homogeneous medium and point source placed on the surface of the medium (see Figure 1). These analytical solutions are employed to measure μ_a and μ_s' of the homogeneous medium. The method is based on fitting measured ac, dc, and phase of the PDW as functions of source–detector distance by the analytical solution.

Inhomogeneities in the optical properties of a turbid medium cause distortion of the PDW from sphericity (see Figure 1). PDWs in homogeneous media of limited size or with strong surface curvature also usually have complex structure. For a limited number of configurations one can obtain solutions to the DE using analytical methods of mathematical physics, such as the method of Green's functions. However, in practice the complexity of the inhomogeneities and surfaces usually requires numerical solution of the DE. This can be achieved using numerical methods for the solution of partial differential equations, such as the finite element method and the finite difference method. The finite difference method can also be used to solve the TE. This may be necessary in situations when the DE is not valid, for example to accurately account for the influence of the cerebrospinal fluid (which has relatively low scattering) on light transport in the human skull. However, the solution of the TE requires formidable computational resources, which increase proportionally to the volume of the medium. Therefore, hybrid DE-TE methods were developed, in

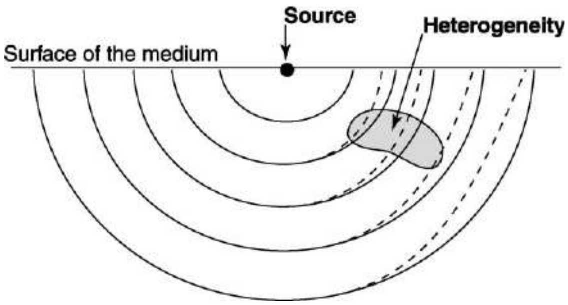

Figure 1 Equiphase surfaces of the PDW induced by a point source in a semi-infinite homogeneous medium (solid lines) and in a medium including heterogeneity (dashed lines).

which DE is used to describe light propagation in most of the medium except small low-scattering areas described by the TE.

One should note that although the term 'photon' suggests quantum properties of light, no consistent application of quantum theory to problems related to PDWI has been developed so far. Also, it should be noted that the notion of a photon as a particle of light having certain energy, momentum, and localization at the same time is not completely consistent with quantum electrodynamics. However, since the theoretical models of light transport in a scattering medium (such as TE and DE) are mathematically equivalent to the models describing systems of randomly moving particles, the notion of a photon as a 'particle of light' may be accepted as adequate and is widely used in the limits of PDWI.

Inverse Problem and Imaging Algorithm

Mathematically the forward problem can be expressed in the form of a general nonlinear operator F:

$$G = F[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}] \qquad [7]$$

where G is the set of measured values (ac, dc, and phase), $\mathbf{r}$ is the coordinate vector of a medium element (voxel), and $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}$ denotes the set of tissue optical properties corresponding to whole collection of voxels. The problem of finding the spatial distribution of medium optical properties can be formalized as

$$\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} = F^{-1}[G] \qquad [8]$$

Neither the transport model nor the diffusion model allow inversion in a closed analytical form for a general boundary configuration and spatial distribution of optical properties. The basic reason for the complexity of the inverse problem is in the complexity of the light paths in the highly scattering medium. This situation is opposite to that in CT, where the x-ray radiation propagates through the medium along straight lines almost without scattering, which significantly simplifies the forward problem and allows its analytical inversion. From the mathematical point of view, the complexity of the inverse problem of PDWI is due to its nonlinearity and ill-posedness. Only approximate methods of inversion have been developed so far.

The simplest method is called backprojection by analogy with the method of the same name used in CT. The method attempts to reconstruct the actual spatial distribution of the optical properties μ_a and/or μ_s' from the 'measured' bulk values μ_a^m and $\mu_s'^m$ obtained at a variety of locations assuming homogeneity of the medium. The measured values are considered as a result of the volumetric 'projection'

$$\{\mu_a^m(\mathbf{r}), \mu_s'^m(\mathbf{r})\} = \int \{\mu_a(\mathbf{r}'), \mu_s'(\mathbf{r}')\} h(\mathbf{r} - \mathbf{r}')\mathrm{d}^3\mathbf{r}' \qquad [9]$$

of the actual distribution, where $\mathbf{r}$ and $\mathbf{r}'$ are the locations of the measurement and reconstruction, respectively. A weight function $h(\mathbf{r})$ accounts for the contribution of the particular region to a measured value and should be derived from the forward model. Different versions of the backprojection method use different weight functions $h(\mathbf{r})$. In the backprojection method the values of μ_a and/or μ_s' are obtained by inverting the convolution [9]. In practice the convolution integral is usually estimated by a sum over a limited number of voxels for a limited number of measurements. Therefore, the inversion of the convolution reduces to the solution of the set of algebraic equations corresponding to [9]. The advantage of the method is high speed, which allows real-time imaging of the dynamic processes, for example, functional hemodynamic changes in the brain. However, the backprojection method provides a phenomenological rather than a self-consistent inversion of the forward problem, and its quantitative accuracy is low.

A more consistent method is based on the perturbation approach to the solution of the inverse problem [8]. If we have an estimate $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ that is close to the ideal solution, then $G_e = F[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e]$ (see eqn [7]) is close to S. Then the difference $G - G_e$ can be written as a Taylor series:

$$G - G_e = J[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e]\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} + \ldots \qquad [10]$$

where $J[\ldots]$ is the Jacobian, and $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$. Neglecting terms after the first one in [10], the problem of finding the correction $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\}$ to the estimate $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ reduces to the calculation of the Jacobian and to the solution of a linear algebraic equation.

This approach provides a basis for the PDWI algorithm, which is the following. First one makes an initial guess at the optical properties of the medium. Then one calculates the ac, dc, and phase of the PDW using a model of light propagation through the tissue and the geometry of the source–detector configuration. The calculated values are compared with the measured ones. If the error is above a set tolerance level, one calculates corrections to the current values of μ_a and μ_s' by solving eqn [10]. The algorithm is completed when the required error level is achieved. Thus, the PDWI algorithm iteratively applies both the forward and inverse problems. Since at each of many iterations one should solve the forward problem, which in the simplest case is a partial differential equation, the algorithm typically requires considerable computational resources.

One should note that the success of the method is largely dependent on the closeness of the initial guess to the actual values of the optical properties. In many studies the ‘first guess’ of the tissue optical properties was obtained by measuring the bulk absorption and reduced scattering coefficients. However, because of the ill-posedness of the inverse problem, homogeneous initial conditions did not always lead to the correct reconstruction. A more productive approach to the problem of the initial conditions is to assign approximate initial values of the optical properties known for example from *in vitro* measurements to different structures revealed by nonoptical imaging techniques, such as magnetic resonance imaging (MRI) or CT.

Applications

PDWI was proposed and developed as a noninvasive biomedical imaging technique. Although the spatial resolution of PDWI cannot compete with that of CT or MRI, PDWI has the advantages of low cost, high temporal resolution, and high biochemical specificity. The latter is based on the optical spectroscopic approach. Biological tissues are relatively transparent to light in the near-infrared band between 700 and 1000 nm (a near-infrared window). The major biological chromophores in this window are oxy- and deoxyhemoglobin, and water. One can separate contributions of these chromophores into light absorption by employing light sources at different wavelengths. Multi-wavelength PDWI can also be used to measure the redox state of the cytochrome oxidase, which is the terminal electron acceptor of the mitochondrial electron transport chain and responsible for 90% of cellular oxygen consumption.

One of the hot topics in cognitive neuroscience is the hemodynamic mechanism of the oxygen supply to neurons and removal of products of cerebral metabolism. High temporal resolution and biochemical specificity make PDWI a unique noninvasive tool for studies of functional cerebral hemodynamics, although light can penetrate no deeper than the very surface of the adult brain. Figure 2 displays a typical arrangement of light sources and detectors used in PDWI of functional cerebral hemodynamics, and the images of the activated area in the brain.

Because of its low cost, noninvasiveness, and compact size of the optical sensors, PDWI is also considered to be a promising tool for mammography and structural imaging of anomalies in the brain. PDWI is an especially promising tool for neonatology, because the infant tissues have low absorption.

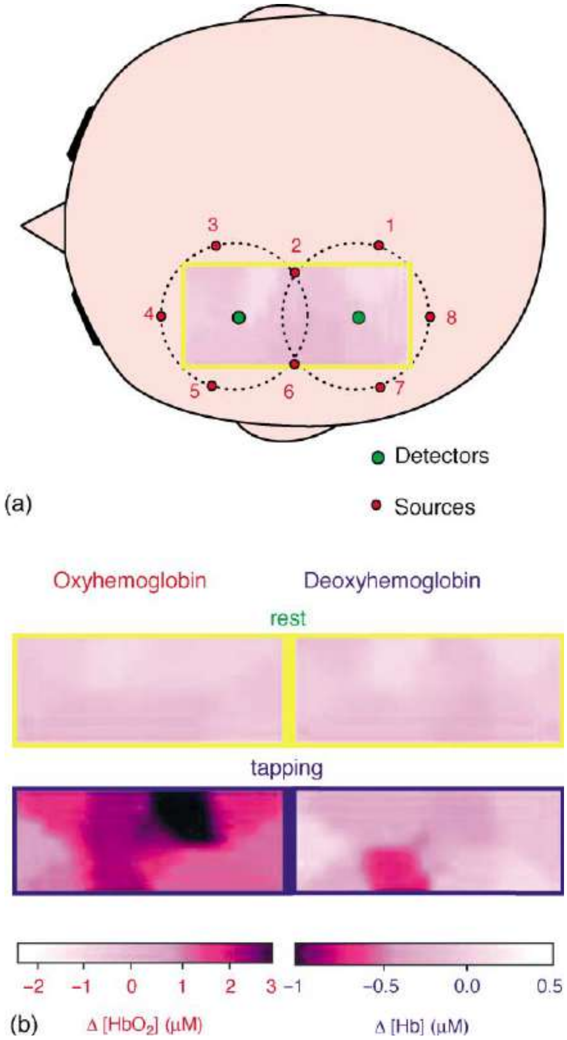

Figure 2 (a) Geometrical arrangement of the 16 sources and two detectors on the head of a human subject. Each one of the eight small red circles (numbered 1–8) represents one pair of source fibers at 758 and 830 nm. Two larger green circles represent detector fiber bundles. The rectangle inside the head outline is the 4 × 9 cm^2 imaged area. (b) Maps of oxyhemoglobin ([HbO_2], left panels) and deoxyhemoglobin ([Hb] right panels) concentration changes at rest (top) and at the time of maximum response during the third tapping period (bottom). The measurement protocol involved hand-tapping with the right hand (contralateral to the imaged brain area). Adapted with permission from Franceschini MA, Toronov V, Filiaci M, Gratton E, Fantini S (2000) On-line optical imaging of the human brain with 160-ms temporal resolution. *Optics Express* **6**(3), 49–57 (Figures 1 and 4).

PDWI instruments employ lasers, light-emitting diodes and gas-discharge tubes as sources of light. Among the detectors there are photomultiplier tubes, avalanche photodiodes, and CCD cameras. Both sources and detectors can be coupled with the tissue by means of optical fibers. A number of commercially produced near-infrared instruments for PDWI are available on the market. According to the type of source modulation, there are three groups of instruments: frequency domain (harmonically modulated sources), time domain (pulsed sources), and steady state. Steady-state instruments are simpler and cheaper than frequency-domain and time-domain ones, but their capabilities are much lower. Time-domain instruments can theoretically provide a maximum of information about the tissue structure, but technically it is difficult to build a time-domain instrument with high temporal resolution and high signal-to-noise ratio. Frequency-domain instruments provide temporal resolution up to several milliseconds and have high quantitative accuracy in the determination of optical properties.

Nomenclature

Absorption coefficient, μ_a [cm^{-1}]
Alternating part of PDW, alternating current (ac)
Average cosine of the scattering angle, g_1
Circular modulation frequency, ω [s^{-1}]
Computed tomography (CT)
Difference between the estimate and actual optical properties, $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ [cm^{-1}]
Diffusion coefficient, $D = v/(3\mu_a + 3\mu_s')$ [cm^2/s]
Diffusion equation (DE)
Distribution of light source power $S(\mathbf{r}, \hat{\Omega}, t)$ [$\mathrm{W\ cm}^{-3}$ $\mathrm{steradian}^{-1}$]
Divergence operator, ∇ [cm^{-1}]
Estimate of optical properties, $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ [cm^{-1}]
General nonlinear operator, F
Jacobian, J
Laplace operator, Δ [cm^{-2}]
Magnetic resonance imaging (MRI)
Measured bulk optical properties (absorption and reduced scattering coefficients) $\{\mu_a^m \mu_s'^m\}$ [cm^{-1}]
Modulation frequency, ν [s^{-1}]
Non-alternating part of PDW, direct current (dc)
Optical properties of the medium (absorption and reduced scattering coefficients) as functions of spatial coordinate, $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}$ [cm^{-1}]
Phase as a function of source–detector distance $\phi(r)$
Phase function, $f(\hat{\Omega}, \hat{\Omega}')$
Photon density, $U(\mathbf{r}, t)$ [$\mathrm{J\ cm}^{-3}$]
Photon density wave imaging (PDWI)
Photon density wave (PDW)
Photon mean free path, l [cm]
Radiance, $L(\mathbf{r}, \hat{\Omega}, t)$ [$\mathrm{W\ cm}^{-2}$ $\mathrm{steradian}^{-1}$]
Radiation fluence, $\Phi(\mathbf{r}, t)$ [$\mathrm{W\ cm}^{-2}$]
Radiation flux, $J(\mathbf{r}, t)$ [$\mathrm{W\ cm}^{-2}$]
Reduced scattering coefficient, μ_s' [cm^{-1}]
Scattering coefficient, μ_s [cm^{-1}]
Set of measured values (ac, dc, and phase), G
Set of predicted values (ac, dc, and phase), G_e
Source term $Q(r, t)$ [$\mathrm{J\ cm}^{-3}$ s^{-1}]
Source–detector distance, r [cm]
Spatial coordinate, $\mathbf{r}$ [cm]
Speed of light in medium, v [cm/s]
Spherical harmonics $Y_{nm}(\hat{\Omega})$
Transport equation (TE)
Weight function, $h(\mathbf{r})$
$x = \omega/(v\mu_a)$

See also

Coherence: Coherence and Imaging. **Fiber and Guided Wave Optics:** Light Propagation. **Imaging:** Imaging Through Scattering Media. **Optical Materials:** Heterogeneous Materials. **Scattering:** Scattering Theory.

Further Reading

Arridge SR (1999) Optical tomography in medical imaging. *Inverse Problems* 15: R41–R93.

Boas DA (1996) *Diffuse Photon Probes of Structural and Dynamical Properties of Turbid Media: Theory and Biomedical Applications*. PhD Thesis, University of Pennsylvania.

National Research Council (US), Institute of Medicine (US) (1996) *Mathematics and Physics of Emerging Biomedical Imaging*. Washington, DC: National Academy Press.

O'Leary MA (1996) *Imaging with Diffuse Photon Density Waves*. PhD Thesis, University of Pennsylvania.

Special Issue (1997) Near-infrared spectroscopy and imaging of living systems. *Philosophical Transactions of the Royal Society B* 352(1354).

Tuchin V (2002) *Handbook of Optical Biomedical Diagnostics*. Bellingham, WA: SPIE Press.

Tuchin VV, Luo Q and Ulyanov SS (2001) *Imaging of Tissue Structure and Function*. Proceedings of SPIE 4427. Bellingham, WA: SPIE Press.

Villringer A and Chance B (1997) Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neuroscience* 20: 435–442.

Villringer A and Dirnagl U (1993) *Optical Imaging of Brain Function and Metabolism*. New York: Plenum Press.

Villringer A and Dirnagl U (1997) *Optical Imaging of Brain Function and Metabolism 2: Physiological Basis and Comparison to Other Functional Neuroimaging Methods*. New York: Plenum Press.

Three-Dimensional Field Transformations

R Piestun, University of Colorado, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Conventional optical imaging is understood as an isomorphic mapping from an object space onto an image space. The object space is composed of scattering or emitting objects. The scattering objects can be opaque, transparent, or partially absorbing, while emitting objects can be as diverse as stars in astronomical imaging or fluorescent molecules in microscopic imaging. The image space is an optical field representation of the object.

In most classical applications of optical imaging the goal is to optimize the mapping between one plane in the object space and one plane in the image space. Modern optical systems, though, are aimed at acquiring three-dimensional information regarding the object. The image is thus a three-dimensional field representation of the object.

While many imaging systems are aimed at obtaining visually appealing images, many modern imaging systems register the field information in some type of sensitive device such as film or CCD cameras for later processing. This processing is typically performed electronically by a computer and the final image is ultimately a mathematical digital representation of the object. This type of imaging is called computational imaging. An example of computational imaging is optical tomography, which involves the registration of a set of images that do not necessarily resemble the object, and off-line digital processing to obtain a three-dimensional representation of the object.

In this article we will focus on imaging in the conventional sense but will also focus on the three-dimensional aspects of the image.

Geometrical Optics Transformations

In many optical problems it is appropriate to consider the wavelength of the electromagnetic wave as infinitely small. This approximation is the origin of the field of geometrical optics in which the propagation of light is formulated in terms of geometrical laws. Accordingly, the energy is transported along light rays and their trajectories are determined by certain differential or integral equations that can be directly derived from Maxwell's equations. One of these formulations uses Fermat's principle that states the principle of the shortest optical path. The optical path between two points P_1 and P_2 is defined as:

$$OP = \int_{P_1}^{P_2} n(P)\mathrm{d}s \qquad [1]$$

where n is the refractive index of the medium. Fermat's principle states that the optical path of a real ray is always a local minimum.

In ideal optical imaging systems, every point P_o of the so-called object space will generate a sharp image point P_i, that is a point where infinite rays emerging from P_o pass through P_i. Perfect imaging occurs when every curve defined by object points is geometrically similar to its image. Perfect imaging can occur between three-dimensional spaces or between surfaces.

An interesting result of geometrical optics is Maxwell's theorem. This theorem states that if an optical instrument produces a sharp image of a three-dimensional domain, the optical length of any curve in the object is equal to the optical length of its image. More explicitly, if the object space has index n_o and the image space index n_i, the theorem states that:

$$\int_{C_\mathrm{o}} n_\mathrm{o}\,\mathrm{d}s_\mathrm{o} = \int_{C_\mathrm{i}} n_\mathrm{i}\,\mathrm{d}s_\mathrm{i} \qquad [2]$$

where C_i is the image of the object curve C_o.

An important consequence of Maxwell's theorem is a theorem by Carathéodory which states that any imaging transformation has to be a projective transformation, an inversion, or a combination of them. Another consequence of this theorem is that if the two media are homogeneous, perfect imaging can only occur if the magnification is equal to the ratio of the refractive indices in the object and image space. If the media have the same index, the imaging is trivial with unit magnification. The simplest example of a perfect imaging instrument is a planar mirror. Therefore, if we want to achieve nontrivial imaging between homogeneous spaces with equal index, we must give up either the perfect imaging (similarity) or the sharpness of the image.

The geometrical optics approximation is very useful in the design of optical imaging systems. However, it fails to provide a complete description of wave optical effects, notably diffraction. Therefore, the following sections provide a description of field transformations based on wave optics.

Systems Approach to 3D Imaging

In this article we deal with linear imaging systems. Hence the relation between three-dimensional (3D) input (object) signals and 3D output (image) signals can be completely specified by the 3D impulse response or point spread function (PSF) of the system $h(x, y, z; x'y', z')$. This function specifies the response of the system at image space coordinates (x, y, z) to a δ-function input (point source) at object coordinates (x', y', z'). The optical system object and image signals are related by the superposition integral:

$$\mathrm{i}(x,y,z) = \int\int\int_{\Omega} \mathrm{o}(x',y',z')h(x,y,z;x'y',z')\mathrm{d}x'\,\mathrm{d}y'\,\mathrm{d}z' \quad [3]$$

where Ω is the object domain. It is clear that for a generic linear imaging system the image signal is completely specified by the response to input impulses located at all possible points of the image domain.

If the system is shift invariant, the PSF due to a point source located anywhere in the object domain will be a linearly shifted version of the response due to a reference point source. This approximation is in general not valid and the system's PSF is a function of the position of the point source in the object domain. Shift invariance is traditionally divided into radial and axial shift invariance. Radial shift variance is the change in the image of a point source as a function of its radial distance from the optical axis. Radial shift variance is caused by lens aberrations. Axial shift variance is the change in the image of a point source as a function of its axial position, z', in the object space. Axial shift variance is inherent to any rotationally symmetric lens system because they can only have one pair of conjugate surfaces, i.e., object–image corresponding surfaces.

In most practical cases the image is detected at a single plane, $z = z_1$ transverse to the optical axis Oz. In those situations, we need to specify the impulse response on this plane:

$$h(x, y, z_1; x'y', z') = h_{z'}(x, y; x'y') \quad [4]$$

The condition for lateral shift invariance is then expressed as $h_{z'}(x, y; x'y') = \tilde{h}_{z'}(x - x', y - y')$. In this case, the superposition integral is a two-dimensional (2D) convolution and it is possible to define the transfer function of the system that is the 2D Fourier transform of the impulse response $\tilde{h}_{z'}(x - x', y - y')$.

Point Spread Function for Coherent, Incoherent, and Partially Coherent Illumination

The input–output relations of an optical imaging system relate different magnitudes according to the type of illumination. If the illumination is coherent, the system is linear in the complex amplitude, while if the illumination is fully incoherent the system is linear in intensity. The 3D impulse response of the incoherent (h_{Inc}) and coherent (h_{Coh}) systems are related as follows:

$$h_{\mathrm{Inc}}(x, y, z; x'y', z') = \left|h_{\mathrm{Coh}}(x, y, z; x'y', z')\right|^2 \quad [5]$$

Notwithstanding, many optical systems cannot be considered fully coherent or fully incoherent. In these situations the system is said partially coherent. If the optical path differences within the optical system are much shorter than the coherence length of the illumination, the light effectively has complete temporal coherence and the magnitude of interest is the mutual intensity. The mutual intensity is defined as the statistical correlation of the complex wavefunction $U(\mathbf{r}, t)$ between different points in space:

$$J(\mathbf{r}_1, \mathbf{r}_2) = \langle U^*(\mathbf{r}_1, t)U(\mathbf{r}_2, t)\rangle \quad [6]$$

where $\mathbf{r}_i = (x_i, y_i, z_i), i = 1, 2$. The illumination is then called quasi-monochromatic. For the case of partially coherent quasi-monochromatic illumination, the system is linear in the mutual intensity. The partially coherent impulse response (h_{PC}) is also related to the coherent impulse response as follows:

$$h_{\mathrm{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) = h^*_{\mathrm{Coh}}(\mathbf{r}_1; \mathbf{r}'_1)h_{\mathrm{Coh}}(\mathbf{r}_2; \mathbf{r}'_2) \quad [7]$$

The 3D superposition integral that relate object and image mutual intensities is now:

$$J_{\mathrm{image}}(\mathbf{r}_1, \mathbf{r}_2) = \int\int\int_{\Omega} h_{\mathrm{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) \times J_{\mathrm{object}}(\mathbf{r}'_1, \mathbf{r}'_2)\mathrm{d}\mathbf{r}'_1\,\mathrm{d}\mathbf{r}'_2 \quad [8]$$

Therefore, we conclude that knowledge of the coherent PSF is enough to characterize the optical system for illumination of various degrees of coherence.

Spatial Frequency Content of the 3D Point Spread Function

As previously stated, the coherent impulse response of the optical system completely defines its imaging characteristics. Therefore, it is justified to wonder what the limitations are in the specification of this response. In this section we analyze the frequency domain of coherent and incoherent impulse responses.

The impulse response of the optical system will satisfy the corresponding wave equations regardless of the particulars of the optical system. In particular, assuming a homogeneous isotropic medium in the imaging domain, the coherent response $h(\mathbf{r})$ to an impulse at $\mathbf{r}'$ will satisfy the homogeneous Helmholtz equation in the image domain:

$$\nabla^2 h(\mathbf{r}) + k^2 h(\mathbf{r}) = 0 \qquad [9]$$

where $h(\mathbf{r})$ is the complex amplitude and $k = 2\pi/\lambda$, with λ the wavelength in the medium. If the field at a given origin of coordinates $z = 0$, $f(x, y)$, is known, $h(\mathbf{r})$ can be calculated using the angular spectrum representation as follows:

$$h(x,y,z) = \iint_{-\infty}^{\infty} F(f_x, f_y) \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\, \mathrm{d}f_y \qquad [10]$$

where $f_z = \sqrt{\lambda^{-2} - (f_x^2 + f_y^2)}$ and $F(f_x, f_y) = \mathrm{FT}\{f(x,y)\}$; FT represents a 2D Fourier transformation and f_x, f_y are the spatial frequencies along the x and y axes respectively.

Equation [10] can be written as a 3D integral as follows:

$$h(x,y,z) = \iint_{-\infty}^{\infty} F(f_x, f_y)\cdot\delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\, \mathrm{d}f_y\, \mathrm{d}f_z \qquad [11]$$

where δ represents the dirac impulse. Next we represent the 3D impulse response $h(x,y,z)$ as a 3D Fourier decomposition:

$$h(x,y,z) = \iint_{-\infty}^{\infty} H(f_x, f_y, f_z) \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\, \mathrm{d}f_y\, \mathrm{d}f_z \qquad [12]$$

where $H(f_x, f_y, f_z) = \mathrm{FT}^{3\mathrm{D}}\{h(x,y,z)\}$. Comparing eqns [11] and [12] we can infer that:

$$H(f_x,f_y,f_z) = F(f_x,f_y)\cdot\delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \qquad [13]$$

This relation implies that the domain of the Fourier transform of the impulse response is confined to a spherical shell of radius $1/\lambda$. Alternatively, $H(f_x,f_y,f_z)$ is zero everywhere, except from the surface of the sphere.

The previous derivation did not consider any limitation in the range of spatial frequencies allowed through the system. We now assume a finite aperture that will introduce an effective radial cutoff frequency $f_c < 1/\lambda$ and will further limit the spatial spectral domain of $h(x,y,z)$. This is depicted in Figure 1. Note that the transverse and longitudinal spatial frequencies are related, hence the longitudinal spatial frequencies are located within a bandpass of width $\Delta f_z = 1 - \sqrt{\lambda^{-2} - f_c^2}$.

Let us now consider the frequency content of possible incoherent PSFs. From eqn [5] we derive the 3D Fourier transform of the incoherent impulse response:

$$\begin{aligned} H_{\mathrm{Inc}}(f_x, f_y, f_z) &= \mathrm{FT}\left\{h_{\mathrm{Inc}}(x,y,z;x'y',z')\right\} \\ &= \mathrm{FT}\left\{|h_{\mathrm{Coh}}(x,y,z;x'y',z')|^2\right\} \\ &= H \otimes H(f_x, f_y, f_z) \end{aligned} \qquad [14]$$

where $\otimes$ represents a 3D correlation. The domain of H_{Inc} now becomes a volume of revolution as shown in Figure 2. Moreover, because h_{Inc} is real, H_{Inc} is conjugate symmetric, i.e., $H_{\mathrm{Inc}} = H_{\mathrm{Inc}}^*$. Now, the transverse cutoff frequency is $f_{\mathrm{c}}^{\mathrm{Inc}} = 2f_{\mathrm{c}}$ while the longitudinal cutoff frequency is $f_{\mathrm{cl}}^{\mathrm{Inc}} = 2\Delta f_z$.

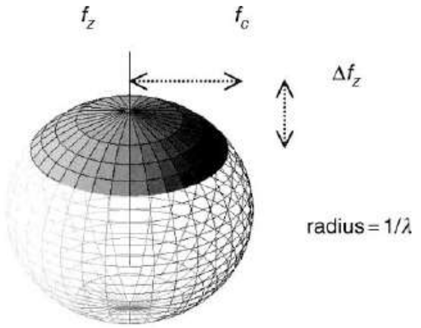

Figure 1 Three-dimensional spatial-frequency domain of the coherent point spread function.

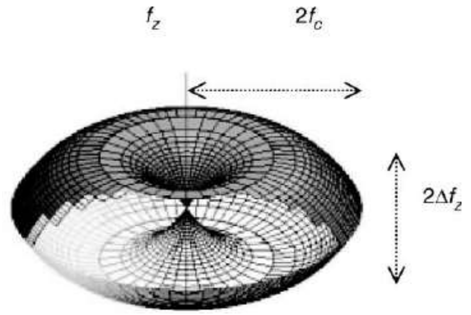

Figure 2 Three-dimensional spatial-frequency domain of the incoherent point spread function.

Diffraction Calculation of the 3D Point Spread Function

The 3D-image field created by a point source in the object space can be predicted by proper modeling of the optical system. In the absence of aberrations, the distortion of the PSF is originated in the diffraction created by the limiting aperture of the optical system. Diffraction is thus the ultimate limitation to the resolution of the optical system. Under the design conditions of the system, i.e., assuming all aberrations are corrected, the image of a point source is the diffraction pattern produced when a spherical wave converging to the (geometrical) image point is diffracted by the limiting aperture of the system. This is called a diffraction limited image.

There are many formulations to describe the 3D diffraction pattern obtained with a converging wave through an aperture. Here we review the Kirchhoff formulation. Let us consider the aperture **A** and the field at a point P_{d} originated from a converging spherical wave focused at point P_{F} (see Figure 3). We define the vectors **r** joining a generic point on the aperture P_0 and P_{F}, and **s** joining P_0 and P_{d}. The Kirchhoff formulation states that the complex amplitude $U(P_{\mathrm{d}}/P_{\mathrm{F}})$ is the integral over the aperture of the complex amplitude at each point propagated to P_{d} by a spherical wave $\exp(iks)/s$ modified by the directional factor $\cos(\hat{\mathbf{n}},\mathbf{r}) - \cos(\hat{\mathbf{n}},\mathbf{s})$. $\hat{\mathbf{n}}$ is the unit vector normal to the aperture:

$$U(P_{\mathrm{d}}/P_{\mathrm{F}}) = \frac{-iC}{2\lambda}\iint_{\mathbf{A}} \frac{\exp[-ik(r-s)]}{rs} \times [\cos(\hat{\mathbf{n}},\mathbf{r}) - \cos(\hat{\mathbf{n}},\mathbf{s})]\mathrm{d}S \qquad [15]$$

where $r = \|\mathbf{r}\|, s = \|\mathbf{s}\|$, and C is the amplitude of the spherical wave.

The directional factor varies slowly and in many situations can be approximated by a constant equal to 2 for small image-field angles. The amplitude over the aperture is also slowly varying and can be approximated by a constant, i.e., $C/r \approx C_0$. Different formulations of the Kirchhoff integral differ in the estimates of the factor $1/s$ and the phase $k(r-s)$. For example, one such formulation leads to the following expression for the 3D diffraction pattern for a focal point on the axis of an optical system with circular aperture:

$$U(P_{\mathrm{d}}/P_{\mathrm{F}}) = \frac{-ika^2C_0}{z_d}\exp\left[ik\left(z_{\mathrm{d}} - z_{\mathrm{F}} + \frac{r_{\mathrm{d}}^2}{2z_{\mathrm{d}}}\right)\right] \times \int_0^1 J_0\left(\frac{ka\rho r_{\mathrm{d}}}{z_{\mathrm{d}}}\right)\exp\left[\frac{-ika^2\rho^2(z_{\mathrm{d}} - z_{\mathrm{F}})}{2z_{\mathrm{d}}z_{\mathrm{F}}}\right]\rho\mathrm{d}\rho \qquad [16]$$

where a is the radius of the aperture, r_{d} is the radial coordinate of P_{d}, and ρ is the radial coordinate of P_0.

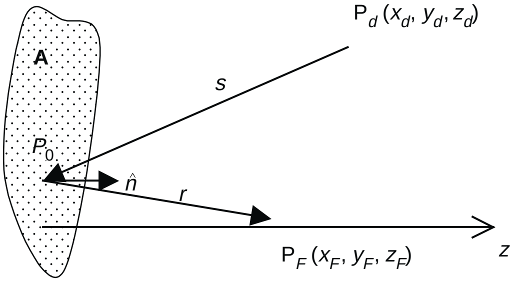

Figure 3 Diffraction of a spherical wave by a planar aperture.

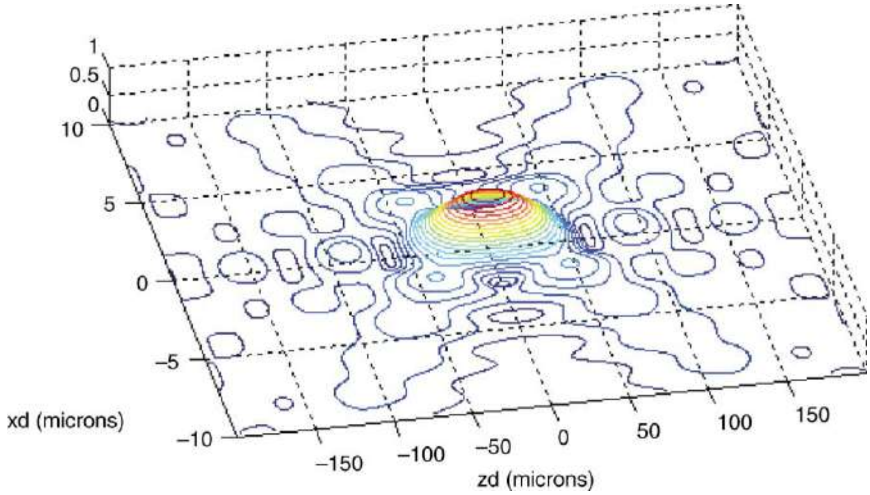

Figure 4 Lines of equal intensity in a meridional plane near the focus of a converging spherical wave diffracted at a circular aperture ($\lambda = 0.5\ \mu$, $a = 0.5$ cm, $z_{\mathrm{F}} = 4$ cm).

To illustrate the use of eqn [16], we calculated the intensity around the focus of a spherical wave diffracted through a circular aperture of radius $a = 0.5$ cm, focusing at distance $z_{\mathrm{F}} = 4$ cm from the screen. We assumed the wavelength to be $\lambda = 0.5\ \mu$. Figure 4 depicts the magnitude of the complex amplitude in a meridional plane near the focus of the converging spherical wave.

Models that take into account off-axis focal points have also been developed. For the most accurate calculations, it is necessary to consider the full electromagnetic nature of the optical field.

Conclusions

Classical imaging theory can be applied to the 3D imaging transformation from object to image. The systems approach to the imaging problem implies that knowledge of the 3D coherent PSF suffices to predict the 3D response of the system under illumination of varying degrees of coherence. The PSF response can be predicted using well-known formulations of diffraction.

See also

Coherent Lightwave Systems. **Fourier Optics**.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. New York: Cambridge University Press.

Gibson SF and Lanni F (1989) Diffraction by a circular aperture as a model for three-dimensional optical microscopy. *Journal of the Optical Society of America A* 6: 1357–1367.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Li Y and Wolf E (1984) Three-dimensional intensity distributions near the focus in systems of different Fresnel numbers. *Journal of the Optical Society of America A* 1: 801–808.

Piestun R and Shamir J (2002) Synthesis of 3D light fields and applications. *Proceedings of the IEEE* 90: 222–244.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Stamnes JJ (1986) *Waves in Focal Regions*. Bristol: Hilger.

Volume Holographic Imaging

G Barbastathis, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005 Published by Elsevier Ltd.

Volume holographic imaging (VHI) refers to a class of imaging techniques which use a volume hologram in at least one location within the optical path. The volume hologram acts as a 'smart lens,' which processes the optical field to extract spatial information in three dimensions (lateral as well as longitudinal) and spectral information.

Figure 1 is a generic VHI system. The object is either illuminated by a light source (e.g., sunlight or a pump laser) as shown in the figure, or it may be self-luminous. Light scattered or emitted by the object is first transformed by an objective lens and then illuminates the volume hologram. The role of the objective is to form an intermediate image which serves as input to the volume hologram. The volume hologram itself is modeled as a three-dimensional (3D) modulation $\Delta\varepsilon(\mathbf{r})$ of the dielectric index within a finite region of space. The light entering the hologram is diffracted by $\Delta\varepsilon(\mathbf{r})$ with efficiency η, defined as

$$\eta = \frac{\text{Power diffracted by the volume hologram}}{\text{Power incident to the volume hologram}} \qquad [1]$$

We assume that diffraction occurs in the Bragg regime. The diffracted field is Fourier-transformed by the collector lens, and the result is sampled and measured by an intensity detector array (such as a CCD or CMOS camera).

Intuitively, we expect that a fraction of the illumination incident upon the volume hologram is Bragg-matched and is diffracted towards the Fourier-transforming lens. The remainder of the incident illumination is Bragg-mismatched, and as result is transmitted through the volume hologram un-diffracted. Therefore, the volume hologram acts as a filter which admits the Bragg-matched portion of the object and rejects the rest. When appropriately designed, this 'Bragg imaging filter' can exhibit very rich behavior, spanning the three spatial dimensions and the spectral dimension of the object.

To keep the discussion simple, we consider the specific case of a transmission geometry volume

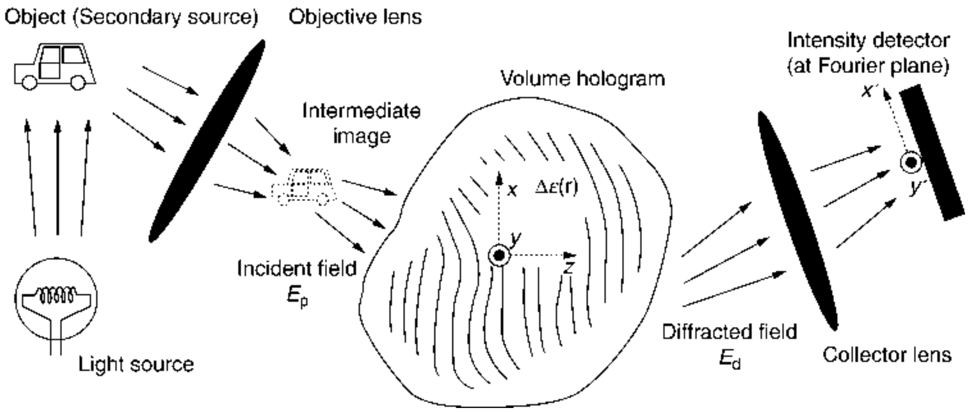

Figure 1 Volume holographic imaging (VHI) system.

hologram, described in Figure 2. The volume hologram is created by interfering a spherical wave and a planewave, as shown in Figure 2a. The spherical wave originates at the coordinate origin. The planewave is off-axis and its wavevector lies on the *xz* plane. As in most common holographic systems, the two beams are assumed to be at the same wavelength λ, and mutually coherent. The volume hologram results from exposure of a photosensitive material to the interference of these two beams.

First, assume that the object is a simple point source. The intermediate image is also approximately a point source, that we refer to as 'probe,' located somewhere in the vicinity of the reference point source. Assuming the wavelength of the probe source is the same as that of the reference and signal beams, volume diffraction theory shows that:

(i) If the probe point source is displaced in the y direction relative to the reference point source, the image formed by the volume hologram is also displaced by a proportional amount. Most common imaging systems would be expected to operate this way.

(ii) If the probe point source is displaced in the x direction relative to the reference point source, the image disappears (i.e., the detector plane remains dark).

(iii) If the probe point source is displaced in the z direction relative to the reference point source, a defocused and faint image is formed on the detector plane. 'Faint' here means that the fraction of energy of the defocused probe transmitted to the detector plane is much smaller than the fraction that would have been transmitted if the probe had been at the origin.

Now consider an extended, monochromatic, spatially incoherent object and intermediate image, as in Figure 2f. According to the above description, the volume hologram acts as a 'Bragg slit' in this case: because of Bragg selectivity, the volume hologram transmits light originating from the vicinity of the y axis, and rejects light originating anywhere else. For the same reason, the volume hologram affords depth selectivity (like the pinhole of a confocal microscope). The width of the slit is determined by the recording geometry, and the thickness of the volume hologram. For example, in the transmission recording geometry of Figure 2a, where the reference beam originates a distance z_0 away from the hologram, the planewave propagates at angle θ with respect to the optical axis z (assuming $\theta \ll 1$ radian), and the hologram thickness is L (assuming $L \ll z_0$), the width of the slit is found to be

$$\Delta x \approx \frac{\lambda z_0}{L\theta} \qquad [2]$$

The imaging function becomes richer if the object is polychromatic. In addition to its Bragg slit function, the volume hologram exhibits then dispersive behavior, like all diffractive elements. In this particular case, dispersion causes the hologram to image simultaneously multiple Bragg slits, each at a different color and parallel to the original slit at wavelength λ, but displaced along the z axis. Light from all these slits finds itself in focus at the detector plane, thus forming a 'rainbow image' of an entire slice through the object, as shown in Figure 2g.

To further exploit the capabilities of volume holograms, we recall that in general it is possible to 'multiplex' (superimpose) several volume gratings

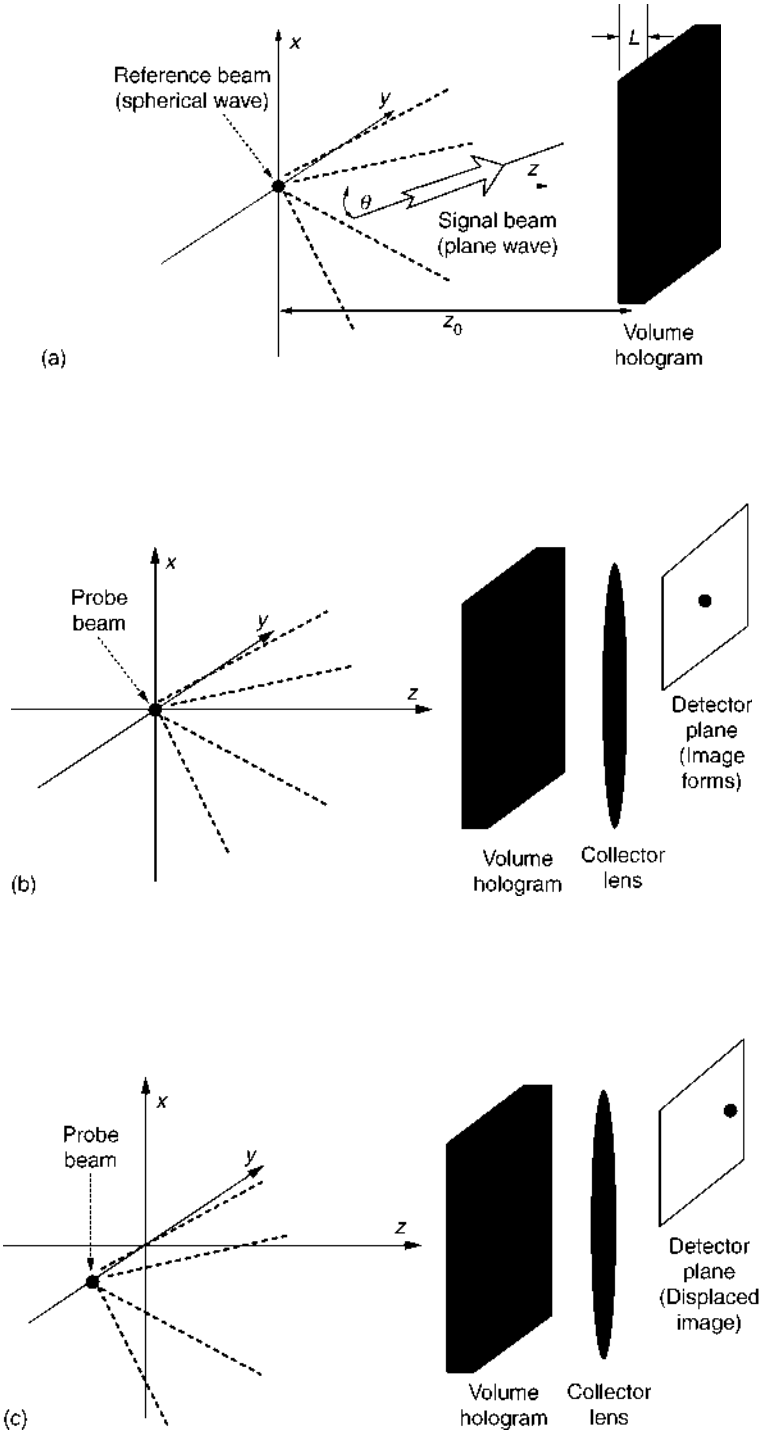
(a)
x
Reference beam
(spherical wave)
y
z
θ
Signal beam
(plane wave)
L
z_0
Volume
hologram
(b)
x
Probe
beam
y
z
Volume
hologram
Collector
lens
Detector
plane
(Image
forms)
(c)
x
Probe
beam
y
z
Volume
hologram
Collector
lens
Detector
plane
(Displaced
image)

Figure 2 (a) Recoding of a transmission-geometry volume hologram with a spherical wave and an off-axis plane wave with its wave-vector on the *xz* plane. (b) Imaging of a probe point source that replicates the location and wavelength of the reference point source using the volume hologram recorded in part (a). (c) Imaging of a probe point source at the same wavelength but displaced in the *y* direction relative to the reference point source. (d) Imaging of a probe point source at the same wavelength but displaced in the *x* direction relative to the reference point source. (e) Imaging of a probe point source at the same wavelength but displaced in the *z* direction relative to the reference point source. (f) 'Bragg slitting:' Imaging of an extended monochromatic, spatially incoherent object using a volume hologram recorded as in (a). (g) Joining Bragg slits from different colors to form rainbow slices: Imaging of an extended polychromatic object using a volume hologram recorded as in (a). (h) Multiplex imaging of several slices using a volume hologram formed by multiple exposures.

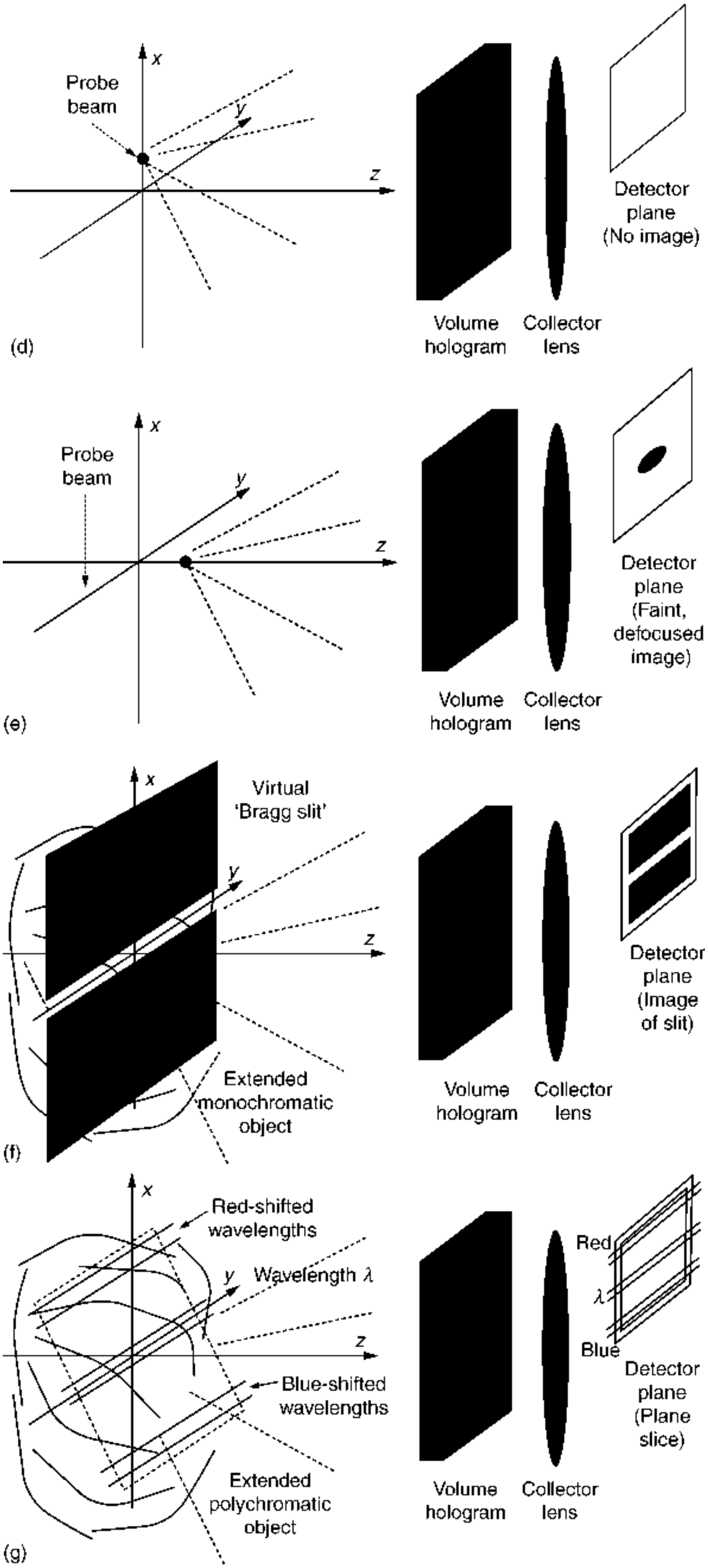
x
Probe beam
y
z
Detector plane (No image)
Volume hologram
Collector lens
(d)
x
Probe beam
y
z
Detector plane (Faint, defocused image)
Volume hologram
Collector lens
(e)
x
Virtual 'Bragg slit'
y
z
Detector plane (Image of slit)
Extended monochromatic object
Volume hologram
Collector lens
(f)
x
Red-shifted wavelengths
y
Wavelength λ
z
Blue-shifted wavelengths
Extended polychromatic object
Red
λ
Blue
Detector plane (Plane slice)
Volume hologram
Collector lens
(g)

Figure 2 *Continued.*

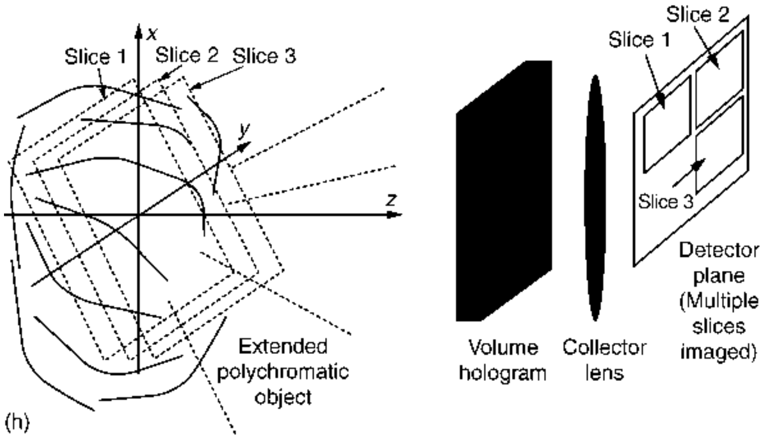

Figure 2 *Continued.*

within the same volume by successive exposures. In the imaging context, suppose that we multiplex several gratings similar to the grating described in Figure 2a but with spherical reference waves originating at different locations and plane signal waves at different orientations. When the multiplexed volume hologram is illuminated by an extended polychromatic source, each grating forms a separate image of a rainbow slice, as described earlier. By spacing appropriately the angles of propagation of the plane signal waves, we can ensure that the rainbow images are formed on nonoverlapping areas on the detector plane, as shown in Figure 2h. This device is now performing true four-dimensional (4D) imaging: it is separating the spatial and spectral components of the object illumination so that they can be measured independently by the detector array. Assuming the photon count is sufficiently high and the number of detector pixels is sufficient, this 'spatio-spectral slicing' operation can be performed in real time, without need for mechanical scanning.

The complete theory of VHI in the spectral and spatial domains is given in the Further Reading section below. Experimental demonstrations of VHI have been performed in the context of a confocal microscope where the volume hologram performs the function of a 'Bragg pinhole' as well as a real-time 4D imager. The limited diffraction efficiency η of volume holograms poses a major concern for VHI systems. Holograms, however, are known to act as matched filters. It has been shown that the matched filtering nature of volume holograms as imaging elements is superior to other filtering elements (e.g., pinholes or slits) in an information-theoretic sense. Efforts are currently underway to strengthen this property by design of volume holographic imaging elements which are more elaborate than described here.

See also

Holography, Techniques: Overview. **Imaging:** Three-Dimensional Field Transformations.

Further Reading

Barbastathis G and Brady DJ (1999) Multidimensional tomographic imaging using volume holography. *Proceedings of the IEEE* 87(12): 2098–2120.

Barbastathis G and Sinha A (2001) Information content of volume holographic images. *Trends in Biotechnology* 19(10): 383–392.

Barbastathis G, Balberg M and Brady DJ (1999) Confocal microscopy with a volume holographic filter. *Optics Letters* 24(12): 811–813.

Barbastathis G, Sinha A and Neifeld MA (2001) Information-theoretic treatment of axial imaging. In: *OSA Topical Meeting on Integrated Computational Imaging Systems*, pp. 18–20. Albuquerque, NM.

Liu W, Psaltis D and Barbastathis G (2004) Real time spectral imaging in three spatial dimensions. *Optics Letters* in press.

van-der Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* IT-10(2): 139–145.

Wavefront Sensors and Control (Imaging Through Turbulence)

C L Matson, Air Force Research Laboratory,
Kirtland, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In many situations, distant objects are imaged using optical or near-infrared imaging systems. Examples include terrestrial surveillance from space, tactical surveillance from airborne imaging systems, and ground-based astronomical imaging. Ideally, the resolution in these images is limited only by the wavelength of light λ used to create the images divided by the imaging system's effective diameter D. The ratio λ/D is called the diffraction limit of the imaging system. Unfortunately, when the imaging system is located in the Earth's atmosphere, it is often atmospheric turbulence fluctuations that limit the level of detail in the long-exposure images, not the diffraction limit of the telescope. For example, the Keck telescopes, located on top of Mauna Kea in Hawaii, are 10 m diameter telescopes that have a diffraction-limited angular resolution of 0.2 μrad for an imaging wavelength of 2 μm. However, atmospheric turbulence limits the achievable resolution in long-exposure images to the order of 1–5 μrad at this wavelength.

The important discovery by Anton Labeyrie in 1970 that short-exposure images of objects retain meaningful diffraction-limited information about the object, where 'short' means that the exposure time is less than the atmospheric turbulence coherence time, has revolutionized imaging through turbulence. Atmospheric coherence times depend upon a number of factors, including wind speeds, wind altitudes, and observation wavelength. At astronomical observatories for visible observation wavelengths, a rule of thumb is that atmospheric turbulence stays essentially constant for exposure times of one to ten milliseconds under average clear-sky weather conditions. A variety of techniques that exploit this discovery have been developed. These techniques can be grouped into three different categories: techniques that remove blurring from the detected intensity images (post-detection compensation), techniques that modify the optical field prior to detecting the light (pre-detection compensation), and hybrid techniques that use both post-detection and pre-detection compensation. In this article, a variety of these techniques are presented that are representative of the myriad of methods that have been developed during the past decades. To simplify the discussion, it is assumed that the blurring due to atmospheric turbulence is the same at every point in the image (isoplanaticity), that the images are formed using incoherent light as emitted by the sun and stars, and that a single filled aperture is used to create the images. Interferometric techniques that use multiple apertures to estimate the mutual coherence function of the optical field and calculate an image from it, are discussed in a separate article (*see* **Imaging:** Interferometric Imaging).

The emphasis in this article is on identifying and removing atmospheric blurring in images. As part of this process, regularization techniques must be included in the image reconstruction algorithms in order to obtain the best possible resolution for an acceptable level of noise. Although regularization techniques are not discussed in this article, information can be found in the Further Reading section at the end of this article (*see* **Imaging:** Inverse Problems and Computational Imaging).

Imaging and Atmospheric Turbulence

In Figure 1, a conceptual diagram illustrates the impact that atmospheric turbulence has upon resolution in classical imaging systems. A distant star is being imaged by a telescope/detector system. In the absence of atmospheric turbulence, the field emitted by the star has a planar wavefront over the extent of the lens since the star is a long distance from the imaging system and thus unresolved by the telescope. When imaged by the telescope, the diffraction pattern of the telescope is obtained, resulting in a diffraction-limited image (Figure 2). However, in the presence of atmospheric turbulence near the telescope, the planar wavefront is randomly distorted by the turbulence, resulting in a corrugated wavefront with a spatial correlation scale given by the Fried parameter r_o. For visible wavelengths, at good astronomical observing sites, r_o ranges from 5 cm to 20 cm. When this corrugated wavefront is converted to an image by the telescope, a distorted image is obtained (Figure 3). Because atmospheric turbulence is time-varying, the distortions in the image also change over time. If the imaging shutter on the telescope remains open for many occurrences of atmospheric turbulence, the resulting image will be an average of all the individually distorted images, resulting in what is known as the seeing disk (Figure 4). It is the seeing disk that typically limits spatial resolution in long-exposure images, not the diffraction pattern of the

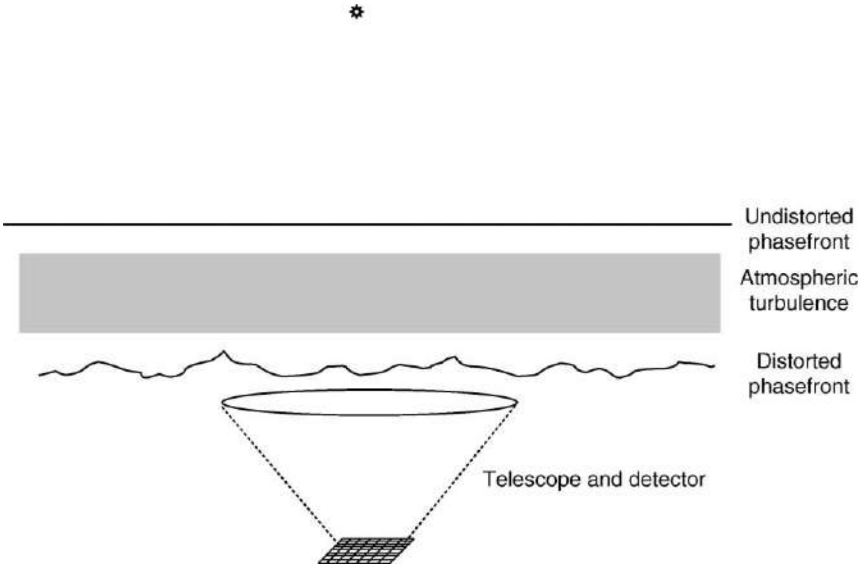

Figure 1 Conceptual diagram of imaging through atmospheric turbulence.

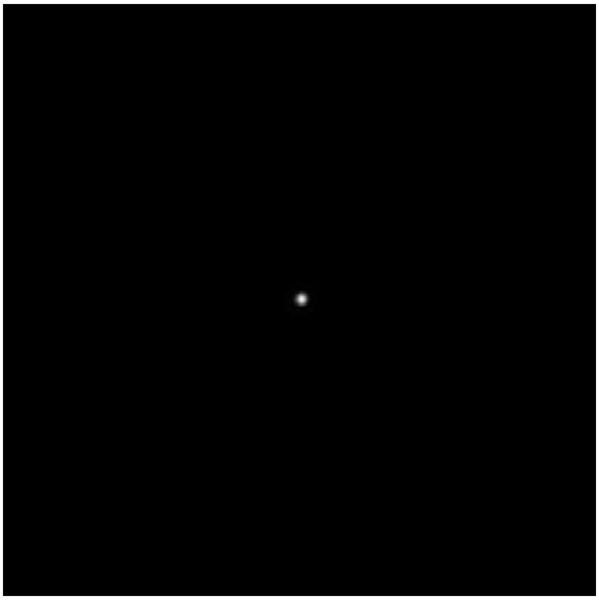

Figure 2 Computer-simulated diffraction-limited image of a point source.

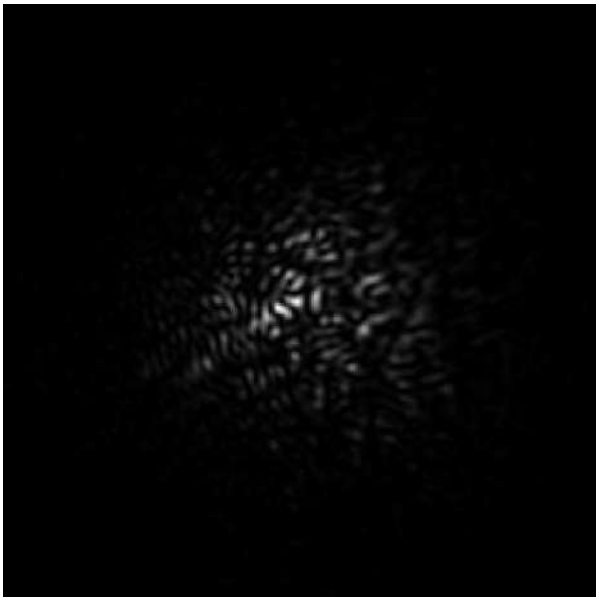

Figure 3 Computer-simulated short-exposure image of a point source through atmospheric turbulence.

telescope. As a result, the effective spatial resolution in long-exposure images in the presence of atmospheric turbulence is given by λ/r_o, not λ/D.

Atmospheric turbulence distorts a field's wavefront by delaying it in a way that is spatially and temporally random. The origin of these delays is the random variation in the index of refraction of the atmosphere as a result of turbulent air motion brought about by temperature variations. A Fourier optics imaging model is useful in characterizing the effects of atmospheric turbulence on spatial resolution in images. Let $i(\vec{x})$ be an image of an object $o(\vec{x})$, where $\vec{x}$ is a two-dimensional spatial vector, and let $I(\vec{f})$ and $O(\vec{f})$ be their spatial Fourier transforms, where $\vec{f}$ is a two-dimensional spatial frequency vector. In the absence of atmospheric turbulence, the diffraction-limited Fourier transform $I(\vec{f})$ is given by

$$I_d(\vec{f}) = O(\vec{f})H_d(\vec{f}) \quad [1]$$

where the subscript 'd' denotes the diffraction-limited case and where $H_d(\vec{f})$ is the diffraction-limited

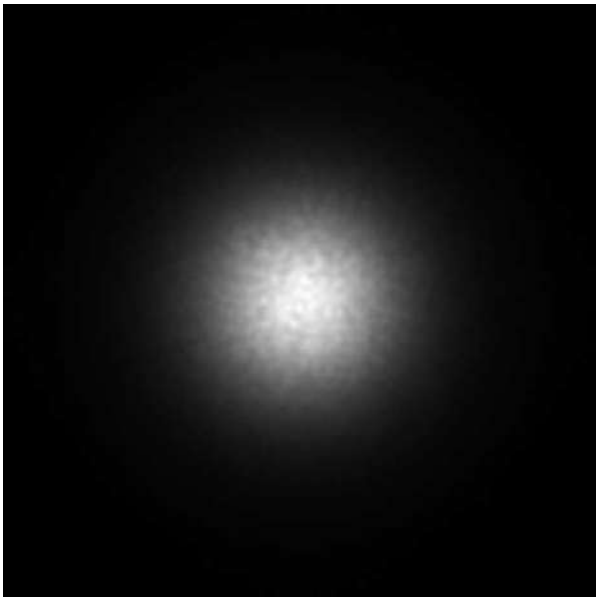

Figure 4 Computer-simulated long-exposure image of a point source through atmospheric turbulence.

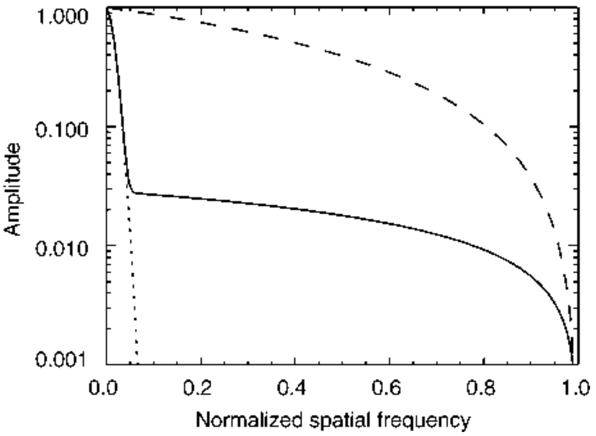

Figure 5 Plots of the Fourier amplitudes of the diffraction-limited transfer function (dashed line), Fourier amplitudes of the long-exposure transfer function (dotted line), and the square root of the average short exposure energy spectrum. The spatial frequency axis is normalized to one at the diffraction limit of the telescope.

telescope transfer function. In the presence of atmospheric turbulence, the Fourier transform of a long exposure image $i_{\mathrm{le}}(\vec{x})$ is given by

$$I_{\mathrm{le}}(\vec{f}) = O(\vec{f})H_{\mathrm{d}}(\vec{f})H_{\mathrm{le}}(\vec{f}) = I_{\mathrm{d}}(\vec{f})H_{\mathrm{le}}(\vec{f}) \qquad [2]$$

where the subscript 'le' denotes long exposure. If the turbulence satisfies Kolmogorov statistics, as is typically assumed, then:

$$H_{\mathrm{le}}(\vec{f}) = \exp\left[-3.44\left(\frac{\lambda d|\vec{f}|}{r_{\mathrm{o}}}\right)^{5/3}\right] \qquad [3]$$

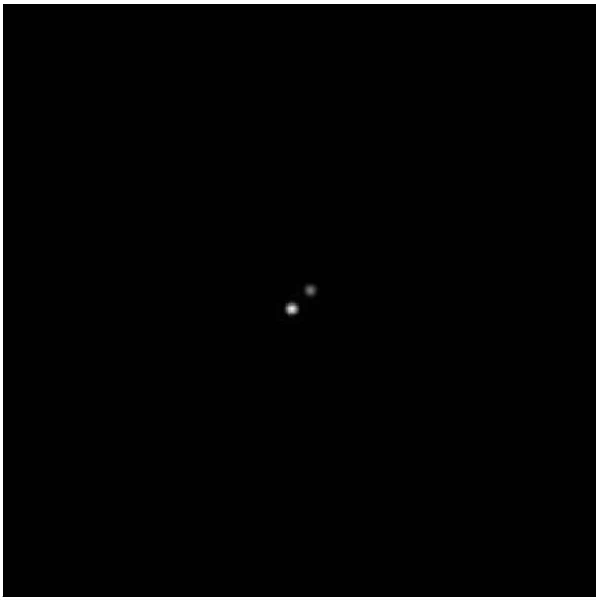

Figure 6 Computer-simulated diffraction-limited image of a binary star.

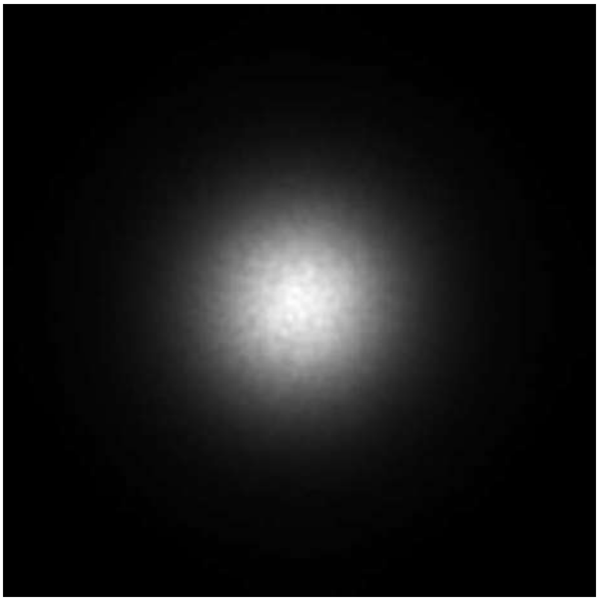

Figure 7 Computer-simulated long-exposure image of a binary star.

where d is the distance between the imaging system's exit pupil and detector plane. As can be seen from eqn [2], the effect of atmospheric turbulence on a long-exposure image is a multiplication of the diffraction-limited image's Fourier transform by a transfer function due to atmospheric turbulence. In Figure 5, an amplitude plot of a diffraction-limited transfer function is displayed along with an amplitude plot of a long-exposure transfer function for a 2.3 m diameter telescope, an imaging wavelength of 0.5 μm, and an r_o value of 10 cm. Notice that the long exposure transfer function significantly reduces the amount of high spatial frequency information in an image. Diffraction-limited and long-exposure turbulence-limited images of a binary star are shown in Figures 6 and 7, illustrating the significant loss of spatial resolution brought on by atmospheric turbulence.

Post-Detection Compensation

Speckle Imaging

Labeyrie's 1970 discovery was motivated by his analysis of short-exposure star images. In these images (see Figure 3), a number of speckles are present, where the individual speckle sizes are roughly the size of the diffraction-limited spot of the telescope. From this, he realized that high spatial frequency information is retained in short-exposure images, albeit distorted. He then showed that the energy spectra $E_I(\vec{f})$ of the individual short-exposure images can be averaged to obtain an estimate of the average energy spectrum that retains high signal-to-noise ratio (SNR) information at high spatial frequencies. The square root of the average energy spectrum (the amplitude spectrum) of a star, calculated using Labeyrie's method (also known as speckle interferometry), is shown in Figure 5, along with the diffraction-limited and long-exposure amplitude spectra. Notice that meaningful high spatial frequency information is retained in the average amplitude spectrum out to the diffraction limit of the telescope. Because the atmospherically corrupted amplitude spectrum estimate is significantly lower than the diffraction-limited amplitude, a number of short exposure images must be used to build up the SNRs at the higher spatial frequencies.

To reconstruct an image, both the Fourier amplitudes and Fourier phases of the image must be known. In fact, it is well known that the Fourier phase of an image carries most of the information in the image. Because only the Fourier amplitudes are available using Labeyrie's method, other techniques were subsequently developed to calculate the Fourier phase spectrum from a sequence of short-exposure images. Techniques that combine amplitude spectrum estimation using Labeyrie's method along with phase spectrum estimation are called speckle imaging techniques. The two most widely used phase spectrum estimation methods are the Knox–Thompson (or cross spectrum) and bispectrum algorithms. Both of these algorithms calculate Fourier-domain quantities that retain high spatial frequency information about the Fourier phase spectrum when averaged over multiple short-exposure images. The cross spectrum and bispectrum techniques calculate the average cross spectrum $C(\vec{f}, \Delta\vec{f})$ and bispectrum $B(\vec{f}_1, \vec{f}_2)$, respectively, that are given by

$$C(\vec{f}, \Delta\vec{f}) = I(\vec{f})I^*(\vec{f} + \Delta\vec{f}) \qquad [4]$$

$$B(\vec{f}_1, \vec{f}_2) = I(\vec{f}_1)I(\vec{f}_2)I^*(\vec{f}_1 + \vec{f}_2) \qquad [5]$$

where the superscript $*$ denotes complex conjugate. Although the cross spectrum is defined for all values of $\vec{f}$ and $\Delta\vec{f}$, only for values of $\Delta\vec{f}$ or $\vec{f}$ less than r_o/λ are the cross spectrum elements obtained with sufficiently high SNRs to be useful in the reconstruction process. The equivalent constraint on the bispectrum elements is that either $\vec{f}_1, \vec{f}_2$, or $\vec{f}_1 + \vec{f}_2$ must be less than r_o/λ. Once the average cross spectrum or bispectrum is calculated from a series of short-exposure images, the phase spectrum $\phi_I(\vec{f})$ of the underlying image can be calculated from their phases using one of a number of algorithms based upon the following equations:

$$\phi_I(\vec{f} + \Delta\vec{f}) = \phi_I(\vec{f}) - \phi_C(\vec{f} + \Delta\vec{f}) \qquad [6]$$

for the cross spectrum and:

$$\phi_I(\vec{f}_1 + \vec{f}_2) = \phi_I(\vec{f}_1) + \phi_I(\vec{f}_2) - \phi_B(\vec{f}_1 + \vec{f}_2) \qquad [7]$$

for the bispectrum. Because the phases are only known for units of 2π, phasor versions of eqns [6] and [7] must be used. For the cross spectrum technique, each short-exposure image must be centroided prior to calculating its cross spectrum and adding it to the total cross spectrum. The bispectrum technique is insensitive to tilt, so no centroiding is necessary; however, the bispectrum technique requires setting two image phase spectrum values to arbitrary values to initialize the phase reconstruction process. Typically, the image phase spectrum values at $\vec{f} = (0, 1)$ and $\vec{f} = (1, 0)$ are set equal to zero, which centroids the reconstructed image but does not affect its morphology.

To obtain an estimated Fourier transform $I_e(\vec{f})$ of the image from the energy and phase spectrum estimates, the square root of the average energy spectrum is combined with the average phase spectrum calculated from either the bispectrum or cross spectrum. The result is given by

$$I_e(\vec{f}) = [E_I(\vec{f})]^{1/2} \exp[j\phi_I(\vec{f})] \qquad [8]$$

where $j = \sqrt{-1}$. The result in eqn [8] includes atmospheric attenuation of the Fourier amplitudes, so it must be divided by the amplitude spectrum of the atmosphere to obtain the desired object spectrum estimate. The atmospheric amplitude spectrum estimate is obtained by using Labeyrie's technique on a sequence of short-exposure images of a nearby unresolved star. No phase spectrum correction is necessary because it has been shown theoretically and verified experimentally that atmospheric turbulence has a negligible effect upon $\phi_{\mathrm{I}}(\vec{f})$.

Blind Deconvolution

Speckle imaging techniques are effective as well as easy and fast to implement. However, it is necessary to obtain a separate estimate of the atmospheric energy spectrum using an unresolved star in order to reconstruct a high-resolution image. Usually, the star is not co-located with the object and the star images are not collected simultaneously with the images of the object. As a result, the atmospheric energy spectrum estimate may be inaccurate. At the very least, collecting the additional star data expends valuable observation time. In addition, there is often meaningful prior knowledge about the object and the data, such as the fact that measured intensities are positive (positivity), that many astronomical objects are imaged against a black background (support), and that the measured data are band-limited, among others. Speckle imaging algorithms do not exploit this additional information. As a result, a number of techniques have been developed to jointly estimate the atmospheric blurring effects as well as an estimate of the object being imaged using only the short-exposure images of the object and available prior knowledge. These techniques are called blind deconvolution techniques. Two of the most common techniques are known as multi-frame blind deconvolution (MFBD) and projection-based blind deconvolution (PBBD).

The MFBD algorithm is based upon maximizing a figure of merit to determine the best object and atmospheric short-exposure point spread function (PSF) estimates given the data and available prior knowledge. This figure of merit is created using a maximum likelihood problem formulation. Conceptually, this technique produces estimates of the object and the atmospheric turbulence short-exposure PSFs that most likely generated the measured short-exposure images. Mathematically, the MFBD algorithm obtains these estimates by maximizing the conditional probability density function (pdf) $\mathrm{p}[i_{\mathrm{m}}(\vec{x}) | o(\vec{x}); h_{\mathrm{m}}(\vec{x}),\ m = 1, \ldots, M]$, where $i_{\mathrm{m}}(\vec{x})$ is the mth of M short exposure images and is given by

$$i_{\mathrm{m}}(\vec{x}) = h_{\mathrm{m}}(\vec{x}) * o(\vec{x}) + n_{\mathrm{m}}(\vec{x}) \qquad [9]$$

where $h_{\mathrm{m}}(\vec{x})$ and $n_{\mathrm{m}}(\vec{x})$ are the atmospheric/system blur and detector noise functions for the mth image, respectively, and the asterisk denotes convolution. The noise is assumed to be zero mean with variance $\sigma^2(\vec{x})$. The conditional pdf $\mathrm{p}[i_{\mathrm{m}}(\vec{x}) | o(\vec{x}); h_{\mathrm{m}}(\vec{x}),\ m = 1, \ldots, M]$ is the pdf of the noise with a mean equal to $h_{\mathrm{m}}(\vec{x}) * o(\vec{x})$ and, as a function of $o(\vec{x})$ and $h_{\mathrm{m}}(\vec{x})$, it is known as a likelihood function. For spatially independent Gaussian noise, as is the case for camera read noise, the likelihood function is given by

$$\mathrm{L_G}[o(\vec{x}); h_{\mathrm{m}}(\vec{x}),\ m = 1, \ldots, M] = \prod_m \prod_{\vec{x}} \frac{1}{\sqrt{2\pi\sigma^2(\vec{x})}} \times \exp\left\{ -\left[i(\vec{x}) - (o * h_{\mathrm{m}})(\vec{x})\right]^2 / 2\sigma^2(\vec{x}) \right\} \qquad [10]$$

For Poisson noise, the likelihood function is given by

$$\mathrm{L_P}[o(\vec{x}); h_{\mathrm{m}}(\vec{x}), \mathrm{m} = 1, \ldots, M] = \prod_m \prod_{\vec{x}} \frac{\left[(o * h_{\mathrm{m}})(\vec{x})\right]^{i(\vec{x})} \exp\left\{ -\left[(o * h_{\mathrm{m}})(\vec{x})\right] \right\}}{i(\vec{x})!} \qquad [11]$$

In practice, logarithms of eqns [10] and [11] are taken prior to searching for the estimates of $o(\vec{x})$ and $h_{\mathrm{m}}(\vec{x})$ that maximize the likelihood function.

In general, there are an infinite number of estimates of $o(\vec{x})$ and $h_{\mathrm{m}}(\vec{x})$ that reproduce the original dataset when convolved together. However, when additional prior knowledge is included to constrain the solution space, such as positivity, support, and the knowledge that $h_{\mathrm{m}}(\vec{x})$ is band-limited, along with regularization, the MFBD algorithm almost always converges to the correct estimates of $o(\vec{x})$ and $h_{\mathrm{m}}(\vec{x})$.

The PBBD technique also uses prior knowledge constraints to jointly estimate $o(\vec{x})$ and $h_{\mathrm{m}}(\vec{x})$. However, instead of creating and maximizing a likelihood function, the PBBD technique uses the method of convex projections to find estimates of $o(\vec{x})$ and $h_{\mathrm{m}}(\vec{x})$ that are consistent with the measured short-exposure images and all the prior knowledge constraints. The data consistency constraint and the prior knowledge constraints are mathematically formulated as convex sets and the PBBD technique repeatedly takes the current estimates of $o(\vec{x})$ and $h_{\mathrm{m}}(\vec{x})$ and projects them onto the convex constraint sets. As long as all the sets are truly convex, the algorithm is guaranteed to converge to a solution that is in the intersection of all the sets as long as the algorithm doesn't stagnate at erroneous solutions called traps. These traps are the projection equivalent of local minima in cost function minimization algorithms.

A key component of this algorithm is the ability to formulate the measured short-exposure images and prior knowledge constraints into convex sets. As an example, consider the convex set $C_{\text{pos}}[u(\vec{x})]$ that corresponds to the positivity constraint. This set is given by

$$C_{\text{pos}}[u(\vec{x})] \triangleq \{u(\vec{x}) \text{ such that } u(\vec{x}) \geq 0\} \qquad [12]$$

It is straightforward to show that $C_{\text{pos}}[u(\vec{x})]$ is a convex set. Other constraints can be converted into convex sets in a similar manner. Although all constraints can be formulated into sets, most but not all of these sets are convex.

The projection process is straightforward to implement. For example, the projections of the current estimates of $o(\vec{x})$ and $h_{\text{m}}(\vec{x})$ onto $C_{\text{pos}}[u(\vec{x})]$ are carried out by zeroing out the negative values in these estimates. Similarly, the projections of $o(\vec{x})$ and $h_{\text{m}}(\vec{x})$ onto the convex sets that correspond to their support constraints are carried out by zeroing out all their values outside of their supports. In practice, some versions of PBBD algorithms have been formulated to use minimization routines as part of the projection process in order to minimize the possibility of stagnation.

Deconvolution from Wavefront Sensing

A key distinguishing feature of post-detection compensation algorithms is the type of additional data required to estimate and remove atmospheric blurring in the measured short-exposure images. Blind deconvolution techniques need no additional data. Speckle imaging techniques estimate average atmospheric blurring quantities and need an estimate of the average atmospheric energy spectrum. Because average quantities are calculated when using speckle imaging techniques, only an estimate of the average atmospheric energy spectrum is needed, not the specific energy spectra that corrupted the measured short-exposure images. As a result, the additional data need not be collected simultaneously with the short-exposure images. The technique described in this section, deconvolution from wavefront sensing (DWFS), seeks to deconvolve the atmospheric blurring in each short-exposure image by estimating the specific atmospheric field phase perturbations that blur each short-exposure image. Thus, the additional data must be collected simultaneously with the short-exposure images. The atmospheric field phase perturbations are calculated from data obtained with a wavefront sensor using a different region of the optical spectrum than used for imaging so that no light is taken from the imaging sensor. Most commonly, the wavefront sensor data consists of the first derivatives of the phasefront (phase differences) as obtained with Shack–Hartmann or shearing interferometer wavefront sensors, but curvature sensing wavefront sensors are also used that produce data that are proportional to the second derivatives of the phasefront.

The original algorithm to implement the DWFS technique produced an estimate $\hat{O}(\vec{f})$ of the Fourier transform $O(\vec{f})$ of the object being imaged using the following estimator:

$$\hat{O}(\vec{f}) = \frac{\sum\limits_{m} I_{\text{m}}(\vec{f})\tilde{H}^{*}_{\text{m}}(\vec{f})}{\sum\limits_{m} |\tilde{H}_{\text{m}}(\vec{f})|^{2}} \qquad [13]$$

where $\tilde{H}_{\text{m}}(\vec{f})$ is the overall system transfer function (including the telescope and atmospheric effects) for the mth short-exposure image estimated from the wavefront sensor data. To create $\tilde{H}_{\text{m}}(\vec{f})$, the atmospheric field phase perturbations corrupting the mth image must be calculated from the wavefront sensor measurements. This topic is discussed in more detail in **Imaging**: Adaptive Optics. Then $\tilde{H}_{\text{m}}(\vec{f})$ is obtained using standard Fourier optics techniques. Although eqn [13] is easily implemented given an existing system that includes a wavefront sensor, it suffers from the fact that it is a biased estimator. An unbiased version of eqn [13] is given by

$$\hat{O}(\vec{f}) = \frac{\sum\limits_{m} I_{\text{m}}(\vec{f})\tilde{H}^{*}_{\text{m}}(\vec{f})}{\sum\limits_{n} H_{\text{n,ref}}(\vec{f})\tilde{H}^{*}_{\text{n,ref}}(\vec{f})} \qquad [14]$$

where $\tilde{H}_{\text{n,ref}}(\vec{f})$ and $H_{\text{n,ref}}(\vec{f})$ are estimates of the overall system transfer function obtained from data collected on a reference star using wavefront sensor data and image-plane data, respectively. Thus, to obtain an unbiased estimator, two more datasets must be collected. The number of frames in the reference star dataset need not be the same as for the imaging dataset. As for speckle imaging, the atmospheric statistics must be the same for the reference star and imaging datasets.

Another technique that uses at least one additional data channel along with the measured image is called phase diversity. In the additional data channels, images are also recorded, but with known field phase aberrations intentionally introduced to add known diversity to the unknown field phase corrupting the measurements. The most commonly introduced field phases aberration in an additional channel is defocus due to its ease of implementation. In the case of defocus, the additional data channel along with the imaging channel can be viewed as samples of the focal volume of the imaging system. Both the noise-free object and the unknown field phase aberrations can be recovered from the measured data using optimization techniques to enforce data consistency along with other prior knowledge constraints. These types of systems can use higher-order aberrations as well. The phase diversity concept has been generalized to an approach known as computational imaging, where an imaging system is designed by taking into account its intended use, the expected image degradation mechanisms, and the fact that the image will be the product of both an optical system and a postprocessing algorithm.

Pre-Detection Compensation

The primary pre-detection-only compensation technique in use today employs dynamical electro-optical systems to sense and remove the phases in the impinging wavefront due to atmospheric turbulence. This technique, known as adaptive optics, provides much higher-quality data than post-detection compensation techniques, when there is enough light from the object being imaged or a guidestar (artificial or natural) to drive the adaptive optics system at a bandwidth high enough to keep up with the temporal changes in atmospheric turbulence. It consists of a wavefront sensor that measures the portion of the phasefront in the pupil of the telescope that is due to atmospheric turbulence, a wavefront reconstructor that generates the deformable mirror drive signals from the wavefront sensor data, and a deformable mirror that applies the conjugate of the atmospheric phase distortions to a reimaged pupil. Some of the light reflected off the deformable mirror is then sent to the imaging sensor. Adaptive optics systems are discussed in much greater detail in **Imaging**: Adaptive Optics. Here, it suffices to say that no adaptive optics system produces imagery that is diffraction limited in quality, although, under the right conditions, the images can be near diffraction limited. However, atmospheric seeing conditions fluctuate significantly over a night as well as seasonally. Therefore, virtually all adaptive optics images benefit from applying post-compensation techniques, as is discussed next.

Hybrid Techniques

Adaptive Optics and Post-Processing

Often, an adaptive optics system does not fully correct for atmospheric turbulence. This can occur when atmospheric conditions become worse than the system is designed to handle. Also, an adaptive optics system may be intentionally designed to only partially correct for atmospheric turbulence in order to lower the cost of the system. As a result, there still may be significant atmospheric blurring in systems that use adaptive optics. For this reason, additional post-processing of adaptive optics imagery can result in significant improvement in image quality. All of the techniques discussed in the section on post-detection compensation can be applied to adaptive optics imagery. Applying these techniques should never degrade image quality, and will improve the quality if the adaptive optics system doesn't fully correct for atmospheric turbulence. Generally speaking, the amount of improvement decreases as the performance of the adaptive optics system increases. Because the adaptive optics system's performance parameters can vary as atmospheric conditions vary, the blind deconvolution algorithms are especially applicable to processing adaptive optics imagery since they estimate the overall system transfer function from the imaging data. Because speckle imaging techniques require a separate estimate of the overall system transfer function, it can be difficult and time-consuming to obtain all of the additional data necessary to apply these techniques to adaptive optics data. Since these post-detection compensation techniques require short-exposure images, the adaptive optics system needs to have a camera capable of collecting short-exposure images. Finally, even if an adaptive optics system fully corrects for the atmosphere, deconvolution techniques that remove the diffraction-limited transfer function amplitude attenuation can improve image quality.

Partially Redundant Pupil Masking

Although adaptive optics systems can function very well, currently they are expensive and require significant maintenance. The primary benefit of adaptive optics systems is that they remove atmospheric turbulence phase distortions in the pupil of the telescope, thus permitting light from all areas of the pupil to interfere in phase when creating an image. An interferometric view of imaging reveals that multiple regions of the light in the pupil contribute to a single spatial frequency in the image. When the phasefront in the pupil is distorted by the atmosphere, these multiple regions tend to not interfere in phase, reducing the amplitude of the image's Fourier component at that spatial frequency. This amplitude reduction effect occurs for all spatial frequencies greater than the limit imposed by the atmosphere, r_0/λ. Therefore, techniques have been developed to modify the wavefront by masking part of the pupil prior to measuring the field with an imaging sensor. This class of techniques has various names including optical aperture synthesis, pupil masking, and aperture masking. Here this class of techniques will be referred to as partially redundant pupil masking (PRPM).

PRPM techniques function by masking the pupil in such a way as to minimize the amount of out-of-phase interference of light on the imaging detector. Early efforts used pupil masks that were opaque plates with r_0-sized holes in them. Because the atmospheric coherence length is r_0, all the light from an r_0-sized hole is essentially in phase. By choosing the holes in the mask so that only two holes contribute to a single spatial frequency in the image, out-of-phase interference problems are removed. This type of mask is called a nonredundant pupil mask. Of course, a significant amount of light is lost due to this masking, so only bright objects benefit from this particular implementation. Often, the pupil mask needs to be rotated in order to obtain enough spatial frequencies in the image to be able to reconstruct a high-quality estimate of the object. To increase the light levels in the image, masks have been made to allow some out-of-phase interference for a given spatial frequency, thus trading off the amplitude of the spatial frequencies allowed in the image with the amount of light in the image. Masks with this property are called partially redundant pupil masks.

One version of PRPM is called the variable geometry pupil (VGP) technique. It is envisioned to work in conjunction with an adaptive optics system that does not fully correct the wavefront for atmospheric phase distortions. It employs a spatial light modulator or a micro-electromechanical mirror between the deformable mirror and the imaging sensor. Either of these two devices is driven by signals from the wavefront sensor in the adaptive optics system in such a way that the portions of the pupil where the phasefront has a sufficiently large second derivative is blocked, while the remainder of the phasefront is transmitted.

All PRPM techniques benefit from post-compensation processing. For the case of nonredundant pupil masks, post-processing techniques are required to convert the discrete-spatial-frequency information into an image with reasonable quality. These techniques are essentially the same techniques as used in interferometric imaging, not those described in this article. When partially redundant pupil masks are used (including VGP masks), the image may include enough spatial frequency content that the methods described in the post-detection compensation section can be applied.

See also

Imaging: Adaptive Optics; Interferometric Imaging; Inverse Problems and Computational Imaging.

Further Reading

Ayers GR, Northcott MJ and Dainty JC (1988) Knox–Thompson and triple-correlation imaging through atmospheric turbulence. *Journal of the Optical Society of America A* 5: 963–985.

Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging*. Bristol, UK: Institute of Physics Publishing.

Craig IJD and Brown JC (1986) *Inverse Problems in Astronomy*. Bristol, UK: Adam Hilger Ltd.

Dainty JC (1984) Stellar speckle interferometry. In: Dainty JC (ed.) *Topics in Applied Physics, vol. 9: Laser Speckle and Related Phenomena*, 2nd edn, pp. 255–320. Berlin: Springer-Verlag.

Goodman JW (1985) *Statistical Optics*. New York: John Wiley and Sons.

Haniff CA and Buscher D (1992) Diffraction-limited imaging with partially redundant masks I: Infrared imaging of bright objects. *Journal of the Optical Society of America A* 9: 203–218.

Katsaggelos AK and Lay KT (1991) Maximum likelihood identification and restoration of images using the expectation-maximization algorithm. In: Katsaggelos AK (ed.) *Digital Image Restoration*, pp. 143–176. Berlin: Springer-Verlag.

Nakajima T and Haniff CA (1993) Partial adaptive compensation and passive interferometry with large ground-based telescopes. *Publications of the Astronomical Society of the Pacific* 105: 509–520.

Roggemann MC and Welsh B (1996) *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.

Schulz TJ (1993) Multiframe blind deconvolution of astronomical images. *Journal of the Optical Society of America A* 10: 1064–1073.

Tyson RK (1991) *Principles of Adaptive Optics*. San Diego, CA: Academic Press.

Youla DC and Webb H (1982) Image restoration by the method of convex projections: Part 1 – Theory. *IEEE Transactions on Medical Imaging* M1-1: 81–94.

INCOHERENT SOURCES

Contents

Lamps
Synchrotrons

Lamps

V Roberts, Roberts Research and Consulting, Inc., Burnt Hills, NY, USA

© 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

Introduction

Electric lamps are by far the most common source of light used on Earth, other than our sun. These sources are composed of a very large number of individual radiating elements, all acting independently. This independence causes the light generated by electric lamps to have a very low degree of spatial and temporal coherence. When compared to lasers, lamps are usually considered to be incoherent sources.

Electric lamps are widely used for general lighting, such as residential, commercial and industrial space, and task lighting. A much less frequent use of incoherent sources, but perhaps of more interest to the reader, is as wavelength, intensity, and total light output standard used to calibrate various coherent light sources. We will discuss the general characteristics of incandescent, gas discharge, and solid state lamps, with sources in the Further Reading section at the end of this article, which are used as calibration standards for coherent sources.

Definitions

Before discussing individual lamp technologies, we need to define some of the key terms that are used for describing the performance of electric lamps. Most of these definitions are based on or abstracted from the IESNA Lighting Handbook:

- *Efficacy*: The total light output of a lamp, measured in lumens, divided by the power input in watts. Efficacy has the units of lumens per watt, or $\mathrm{l\,m\,W^{-1}}$, and can be greater than 100. The efficacy of common light sources varies from $10\ \mathrm{l\,m\,W^{-1}}$ to over $200\ \mathrm{l\,m\,W^{-1}}$.
- *Correlated Color Temperature* (*of a light source*) (*CCT*): the absolute temperature of a blackbody whose chromaticity most nearly resembles that of the light source. CCT is measured in Kelvin (K). The CCT of typical light sources varies from 2000 K to 6000 K, with the most popular sources being in the range from 2700 K to perhaps 4500 K.
- *Color Rendition Index* (*of a light source*) (*CRI*): a measure of the degree of color shift objects undergo when illuminated by the light source as compared with those same objects when illuminated by a reference source of comparable color temperature. The highest possible CRI is 100. The lowest theoretical CRI is less than zero, but CRI less than 50 is generally considered to have no value in determining the color quality of a light source.
- *Photopic Luminous Efficiency Function* (V_λ): a measure of the sensitivity of the human eye in lumens per watt of radiation received, as a function of wavelength, for luminance levels higher than $\sim 3\ \mathrm{cd\,m^{-2}}$. V_λ is based on relatively few observations, and individuals may differ significantly from the standard V_λ function. V_λ provides the basis for measuring the efficacy of all light sources.
- *Visible Light*: generally accepted to mean radiation between 380 nm and 780 nm.

Incandescent Lamps

Incandescent lamps are quasi-blackbody thermal radiators. A true blackbody radiator would generate light according to Planck's Law, as shown in eqn [1] and Figure 1, for various temperatures. Most incandescent lamps use tungsten filaments, and the emissivity of tungsten, the ratio of its radiative output to that of a blackbody at the same temperature, is about 0.44 in the visible region of the spectrum and about 0.33 in the infrared. The radiative output of a tungsten incandescent lamp is, therefore, less than half that of a blackbody of the same temperature and the intensity versus wavelength curve does not exactly follow Plank's law, being slightly favorable to visible

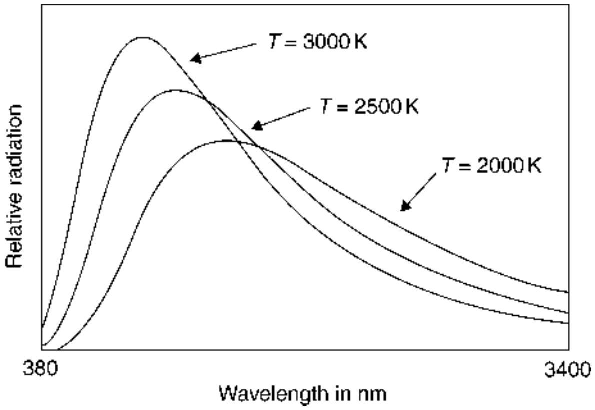

Figure 1 Radiation from blackbody source at various temperatures. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

radiation:

$$E(\lambda) = \frac{2\pi hc^2 10^{-9}}{\lambda^5(\exp^{hc/k\lambda T} - 1)} \qquad [1]$$

where:

E = power per square meter of surface area per nm of wavelength
λ = wavelength in meters
h = Plank's constant, 6.63×10^{-34} Joule seconds
c = speed of light, 3×10^8 meters $\sec^{-1}$
κ = Boltzman's constant, 1.38×10^{-23} Joule K^{-1}
T = temperature of the source in K

Tungsten lamps have excellent color quality, reflected in a perfect CRI of 100, are inexpensive, available in a great variety of sizes, shapes, and operating voltages, have a small source size that permits excellent light control with relatively small reflectors and lenses, run equally well on AC or DC power, can be easily dimmed, can have very long life, are insensitive to operating position, are relatively insensitive to ambient temperature, contain no hazardous materials, and light virtually instantly, though not as fast as LEDs. However, incandescent lamps have one major disadvantage; they are very inefficient. Most of the energy consumed by incandescent lamps is radiated in the infrared, while only 5% to 10% is radiated in the visible portion of the spectrum. The reason for this is shown in Figure 2, which shows the visible region of the spectrum superimposed on the blackbody radiation curves.

There a number of techniques that have been used or are being investigated to increase the efficacy of incandescent lamps. First, as can be seen in Figure 2, increasing the filament temperature shifts the wavelength for peak emission toward lower wavelengths, and will result in a larger fraction of the radiated energy being in the visible region. Unfortunately, increasing filament temperature also increases the rate of tungsten evaporation and therefore shortens the life of tungsten lamps, since evaporation of tungsten is the dominant incandescent lamp failure mode. In a similar manner, the life of incandescent lamps can be increased to arbitrarily long times by decreasing filament temperature. This technique, which is the means used in far too many 'long life' incandescent lamps, unfortunately leads to a decrease in the already low efficacy of incandescent lamps. The approximate dependences of life, efficacy, and color temperature on operating voltage, for common incandescent lamps operating between 90% to 110% of rated voltage, are shown in eqns [2], [3], and [4]:

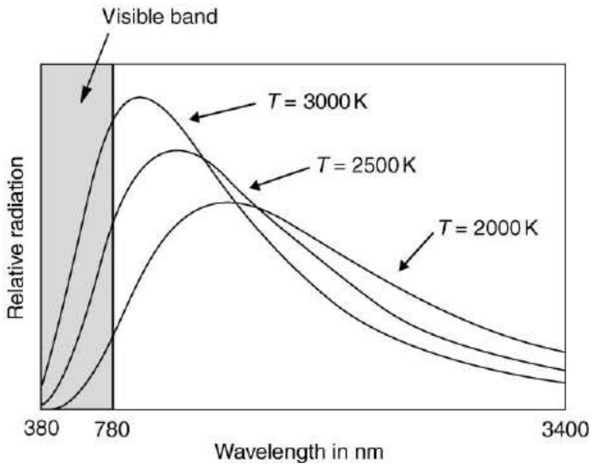

Figure 2 Radiation from blackbody source at various temperatures with visible band. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

$$\frac{\mathrm{LIFE}_1}{\mathrm{LIFE}_2} = \left(\frac{\mathrm{VOLTS}_2}{\mathrm{VOLTS}_1}\right)^{13} \qquad [2]$$

$$\frac{\mathrm{EFFICACY}_1}{\mathrm{EFFICACY}_2} = \left(\frac{\mathrm{VOLTS}_1}{\mathrm{VOLTS}_2}\right)^{1.9} \qquad [3]$$

$$\frac{\mathrm{COLOR\_TEMPERATURE}_1}{\mathrm{COLOR\_TEMPERATURE}_2} = \left(\frac{\mathrm{VOLTS}_1}{\mathrm{VOLTS}_2}\right)^{0.42} \qquad [4]$$

Combining eqns [2] and [3] yields eqn [5]:

$$\frac{\mathrm{LIFE}_1}{\mathrm{LIFE}_2} = \left(\frac{\mathrm{EFFICACY}_2}{\mathrm{EFFACACY}_1}\right)^{6.84} \qquad [5]$$

Under the assumption that the color temperature of an incandescent lamp is approximately the same as the operating temperature of the filament, we can combine eqns [2] and [4] to obtain eqn [6], and combine eqns [3] and [4] to obtain eqn [7]:

$$\frac{\mathrm{LIFE}_1}{\mathrm{LIFE}_2} = \left(\frac{\mathrm{TEMPERATURE}_2}{\mathrm{TEMPERATURE}_1}\right)^{30.95} \qquad [6]$$

$$\frac{\mathrm{EFFACACY}_1}{\mathrm{EFFICACY}_2} = \left(\frac{\mathrm{TEMPERATURE}_1}{\mathrm{TEMPERATURE}_2}\right)^{4.52} \qquad [7]$$

One way to allow operation with higher filament temperatures, while not decreasing filament life, is to fill the lamp with a high molecular weight and/or high pressure gas, as either will retard tungsten evaporation. Lamps using krypton or xenon in place of the normal argon fill allow for a moderate increase in filament temperature. Better yet is a fill of halogen gas, such as HBr, CH_3Br, or CH_2Br^2 plus Kr or Xe at total pressures that range from 2 to 15 atmospheres. The high total pressure inhibits tungsten evaporation from the filament, and the halogen fill gas reacts with tungsten that deposits on the wall of the lamp to remove this tungsten and maintain light output over the life of the lamp. The tungsten removed from the wall of the lamp is redeposited on the tungsten filament, increasing lamp life under some circumstances. Due to the high gas pressures involved, tungsten halogen incandescent lamps are constructed in small 'filament tubes' to minimize both the amount of gas needed and the total explosive energy. The filament tubes are manufactured from quartz or special high temperature glass. For general purpose applications, the quartz or high temperature glass filament tube is enclosed in a larger glass envelope to protect the user should the filament tube explode and to protect the filament tube from environmental contaminants, such as salts from human fingers.

Since 90% to 95% of the energy consumed by an incandescent lamp is radiated as infrared energy, it is easy to see that the potential exists to dramatically increase the efficacy of incandescent lamps by reflecting some of this infrared energy back onto the filament, thus allowing the filament to be maintained at its normal operating temperature with far less electrical power. The reflector must have very low absorption in the visible region of the spectrum, high reflectivity in the IR and a reasonably sharp transition between the two, because the greatest IR emission is just below the visible band. The reflector and filament must also be precisely positioned so that the reflected IR radiation hits the filament, and the reflector must also be low enough in cost to be commercially viable. Successful IR reflectors have been made from multilayer dielectric films. These are similar to the coatings used to make high reflectivity mirrors for He–Ne lasers, but they have much broader bandwidth and lower reflectivity. The reflector-filament alignment problem is most easily solved using the geometry of a tungsten halogen filament tube as the carrier for the IR reflector. The typical efficacy gain is about 50%: a 60-watt IR-halogen lamp produces about as much light as a 90-watt halogen lamp.

Other methods being investigated to increase the efficacy of incandescent lamps involve modification of the filament to decrease its emissivity in the IR and/or increase its emissivity in the visible, and hence turn it into more of a 'selective radiator' than natural tungsten. Three general methods are under investigation: one involves creating a dense pattern of submicron size cavities on the surface of the tungsten to inhibit IR radiation; a second involves coating the tungsten with a material that is itself a selective radiator that favors the visible portion of the spectrum, while a third involves replacing the tungsten filament with a material that is a natural selective radiator. No products exist that employ these techniques, but active R&D projects are underway that may yield a more efficient incandescent lamp in the future.

Incandescent lamps can be used as standards for total flux and color temperature if carefully constructed and operated. The National Institutes for Standards and Technology (NIST) in the US sells 1000-watt tungsten halogen lamps calibrated to an accuracy of 0.6% in luminous intensity and 8 K in color temperature at a color temperature of 2856 K. Similar service are provided by standards organizations in other countries.

Discharge Lamps

The most efficient lamps produced today are based on electric discharges in gasses. These include 'low pressure' lamps such as fluorescent lamps and low pressure sodium lamps; and 'high pressure' or high intensity discharge (HID) lamps such as metal halide, high pressure sodium, high pressure mercury, and sulfur lamps. These lamps are all more efficient than incandescent lamps because they are 'selective radiators' instead of blackbody sources and have been designed to generate as much radiant energy as possible in the visible band while minimizing energy produced outside the visible band. The best white light discharge lamps have conversion efficiencies from electric power to visible light of 25% to 30%, making them almost five times as efficient as the typical incandescent lamp.

Fluorescent Lamps

Fluorescent lamps, as shown in Figure 3, are generally constructed in a cylindrical soda lime glass envelope with a length to diameter ratio significant greater than one. Most contain thermionic electrodes at each end to generate an electric field that establishes and sustains the discharge. The lamps are filled with about 2 Torr of Argon and/or another rare gas, and a few mg of liquid mercury, while the inside surface of the tube is coated with a phosphor that converts UV radiation generated by excited mercury atoms into visible light.

As with any other low pressure gas discharge, various discharge regions are aligned along the axis of the tube:

- Aston Dark Space
- Cathode Glow
- Cathode Dark Space
- Negative Glow
- Faraday Dark Space
- Positive Column
- Anode Glow
- Anode Dark Space

The axial lengths of all these regions, except for the positive column, are fixed by the properties of the cathode and anode, the fill gas type and pressure, and the discharge current. The length of the positive column is variable and fills the space between the Faraday Dark Space and the Anode Glow. The cathode and anode regions of the discharge have a total axial length of about twice the diameter of the discharge tube. Therefore, if the length-to-diameter ratio of the discharge tube is substantially greater than one, the majority of the tube is filled by the Positive Column, as shown in Figure 3. The positive column is the part of the discharge that creates the UV photons that excite the phosphor and cause it to generate visible light. The remainder of this discussion will therefore focus on the positive column.

The basic operation of a fluorescent lamp positive is shown in Figure 4. Free electrons are accelerated by the electric field established between the two electrodes. These electrons collide with both the mercury and rare gas atoms, but selectively excite and ionize mercury atoms, since the excitation and ionization energies of mercury are significantly lower than that of any rare gas used in the lamps. Ionization of mercury creates more free electrons to sustain the discharge. The gas density in the lamp is low enough to require that most of the recombination of mercury ions and free electrons takes place on the wall of the tube, since this recombination requires a three-body collision. The excited mercury atoms generate both UV and visible photons, and the UV photons generate visible light once they reach the phosphor coated on the inner surface of the lamp.

Positive column

Figure 3 Linear fluorescent lamp. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

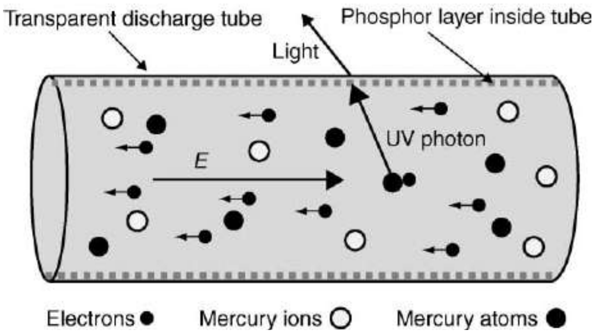

Figure 4 Operation of fluorescent lamp positive column. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

The high efficiency of fluorescent lamps is directly related to the fact that 60% to 65% of the electric power consumed by the positive column of the fluorescent lamp discharge is converted into 245 nm UV photons as a result of the resonance transition from the 6^3P_1 state of mercury at 4.86 eV to the 6^1S_0 ground state. The resonance transition from the 6^1P_1 state of mercury generates 185 nm UV photons that carry about 10% of the energy. About 3% of the positive column energy ends up in mercury visible lines at 405 nm, 436 nm, 546 nm, 577 nm, and 579 nm.

The 245 nm and 185 nm photons generated by the excited mercury atoms travel only a fraction of a mm before they are absorbed by a ground-state mercury atom. However, due to the relative absence of other de-excitation mechanisms, a large fraction of these absorbed photons are re-emitted by the absorbing atoms, and most of the UV photons generated in the core of the discharge eventually reach the phosphor coated interior wall of the discharge tube where they are converted into visible light (and waste heat).

The phosphors used in modern fluorescent lamps are a blend of three rare-earth phosphors:

- Blue: $BaMg_2Al_{16}O_{27}:Eu^{2+}$
- Green: $CeMgAl_{11}O_{19}(Ce^{3+}):Tb^{3+}$
- Red: $Y_2O_3:Eu^{3+}$

Rare earth phosphors provide a combination of high efficiency, high color rendition index, and high resistance to degradation that far exceeds the performance of the halophosphate phosphors used in older generations of fluorescent lamps.

However, due to the Stokes Shift, there is a substantial energy loss converting the UV photons produced by excited mercury atoms into visible light, even using a 'perfect' phosphor that has quantum efficiency of 1.0. The average energy of a visible photon is 2.43 eV, using the 380 nm to 780 nm limits for visible radiation discussed at the start of this article. When one 245 nm photon that has 4.885 eV of energy is converted to one visible photon with an average energy of 2.43 eV, a bit over 50% of the photon's energy is lost. The situation is worse for the 185 nm photons, which each have 6.707 eV of energy. They lose 64% of their energy when they are converted to visible photons, so it is good that the fluorescent lamp discharge generates six to seven 245 nm photons for each 185 nm photon. The Stokes shift is responsible for the largest single energy loss mechanism in fluorescent lamps. Work is underway to develop phosphors with a quantum efficiency greater than 1.0, so less energy will be lost in the UV-to-visible light conversion process, but no practical phosphors have been developed to date.

The mercury pressure for optimum UV production is about 6×10^{-3} Torr, which is the vapor pressure of mercury over a liquid pool at 40 °C. At higher mercury pressure there is too much absorption by ground state mercury atoms, and at lower pressures there is an insufficient number of mercury atoms for excitation by the electrons. The diameter of most linear lamps is chosen so that the coolest spot on the interior of the lamp wall will be at 40 °C when the lamp is operated at its design power in an ambient temperature of 25 °C. Usually this 'cold spot' will be near the middle of the lamp, but in some higher power linear lamps the wall temperature is well over 40 °C. In these lamps a cold spot is often created at the end of the lamp by use of a longer electrode mount and a heat shield mounted between the electrode and the end of the lamp. Obviously, the performance of fluorescent lamps designed in this manner is highly sensitive to ambient temperature. In some smaller, high-power density lamps, most commonly compact fluorescent lamps (CFLs), it is not possible to create a region with a temperature as low as 40 °C. In these lamps the mercury pressure is controlled by amalgamating the mercury with Pb, Bi, or Sn, creating a compound that both reduces the mercury vapor pressure at any given temperature and also reduces the variability of mercury vapor pressure with temperature. These mercury amalgams have the secondary advantage of making the fluorescent lamp less sensitive to ambient temperature.

Mercury, which is the key ingredient in fluorescent lamps, is considered to be a hazardous material by regulatory agencies in many countries. At least as early as the mid-1980s researchers began to study ways to reduce or eliminate mercury from fluorescent lamps.

Development of a mercury-free fluorescent lamp is very difficult and the only mercury-free fluorescent lamp that exists, as of mid-2003 has an efficacy of about $10\,\mathrm{l\,mW}^{-1}$, one-tenth the efficacy of the best mercury-based fluorescent lamp (Osram Sylvania LINEX® Linear Excimer Lamp System, Model LINEX A3-10W40). There are three key problems that have to be solved to develop an efficient mercury-free fluorescent lamp system:

(i) Identify a nontoxic material that generates UV photons and has its optimum gas density at a convenient and safe temperature. The most promising candidate is xenon, used as either an atomic radiator or an excimer.

(ii) Develop a phosphor that produces high-quality white light with high efficiency when excited by UV photons from the chosen mercury-replacement material. This has proven a major challenge. The resonance line of atomic Xe has a wavelength of 147 nm, while the output from a xenon excimer discharge is at 172 nm. The Stokes shift loss converting each of these to visible light is 71% for atomic xenon and 66% for xenon excimers. To reduce the Stokes shift loss, a quantum-splitting phosphor, one that generates more than one visible photon for each UV photon, is needed. Most of the effort to develop mercury-free fluorescent lamps has been concentrated in this area. No viable phosphor has been developed, but work continues.

(iii) Develop a 'reservoir' for xenon or the other working gas. The mass of 'active' gas needed for a fluorescent lamp is small; a 4-foot, 1.5-inch diameter lamp needs 0.1 mg of mercury in the gas phase to operate properly. However, during life, some of this gas is consumed through physical attachment or chemical reaction and must be replaced. The drop of liquid mercury or small mass of mercury amalgam used in conventional fluorescent lamps provides a reservoir that maintains the mercury gas pressure in the presence of these consumption mechanisms.

In contrast to the failure to develop a mercury-free fluorescent lamp, there has been great success in reducing the amount of mercury used in each lamp. A typical 4-foot, 1.5-inch diameter fluorescent lamp produced in the 1970s would have 50 mg to 100 mg of mercury. This high amount of mercury was mostly the result of the type of manufacturing equipment used, but mercury consumption required that each lamp had 15 mg to 20 mg of mercury in order to survive for its rated 20 000-hour life. The development of coatings and other materials that reduce mercury consumption, when combined with new manufacturing technology, has allowed the amount of mercury used in modern 4-foot fluorescent lamps to be reduced to 3 mg to 5 mg, and further reductions are possible with improved technology.

As stated above, the length-to-diameter ratio of most fluorescent lamps is significantly greater than one. The reason for this is that there is a 10 to 15 volt drop at the electrodes, and the power lost in these regions does not produce a significant amount of light. A large length-to-diameter ratio increases the voltage and power of the positive column, where the bulk of the UV photons are generated, relative to the power lost in the electrode regions. When the energy crunch hit in the mid-1970s, there was strong interest in developing a fluorescent lamp that could replace the inefficient incandescent lamp. However, the need for high length-to-diameter ratios meant that it would not be practical to make a fluorescent lamp shaped like an incandescent lamp. This need created the technology to manufacture CFLs which are constructed of long tubes bent or coiled into shapes that allow them to fit into spaces occupied by incandescent lamps.

Fluorescent lamps and other discharge lamps have a negative incremental voltage vs. current (V/I) characteristic, which means that their operating voltage decreases as the operating current increases. Due to this, such discharges need to be operated from a current limited power source, commonly known as a ‘ballast’ in the lighting industry. Ballasts were originally simple series inductors and lamps operated at the power line or mains frequency, though many line frequency ballasts are now considerable more complex than an inductor. In the 1940s, it was discovered that the efficacy of fluorescent lamps could be increased by operation at frequencies of a few kHz. It was not until the 1970s, however, until the cost of power electronics came down to the point where it was possible to make affordable high frequency electronic ballasts. Electronic ballasts operating in the 20 kHz to 100 kHz range are now used with almost all CFLs and in most commercial applications using linear fluorescent lamps.

The desire for fluorescent lamps shaped like incandescent lamps also sparked the development of electrodeless fluorescent lamps. Since the need for high length-to-diameter ratios is driven by the losses at the lamp electrodes, researchers realized they could make fluorescent lamps in short, wide shapes if they eliminated the electrodes. As shown in Figure 5, an electrodeless lamp is just like a transformer, except it has a single-turn discharge secondary instead of a multiturn copper or aluminum secondary. An open core version of an electrodeless fluorescent lamp with integral ballast is shown in Figure 6. The electric field is created by a time-varying magnetic field. Because of the relative high voltage-per-turn imposed by the discharge secondary, these lamps operate only at high frequency. There are three electrodeless fluorescent lamps on the market. The closed core design

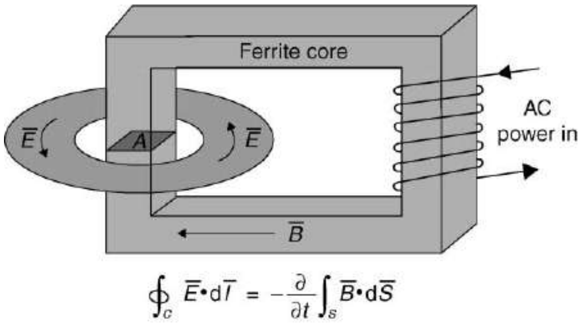

$$\oint_c \overline{E}\cdot d\overline{l} = -\frac{\partial}{\partial t}\int_s \overline{B}\cdot d\overline{S}$$

Figure 5 Basic electrodeless fluorescent lamp with closed ferrite core. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

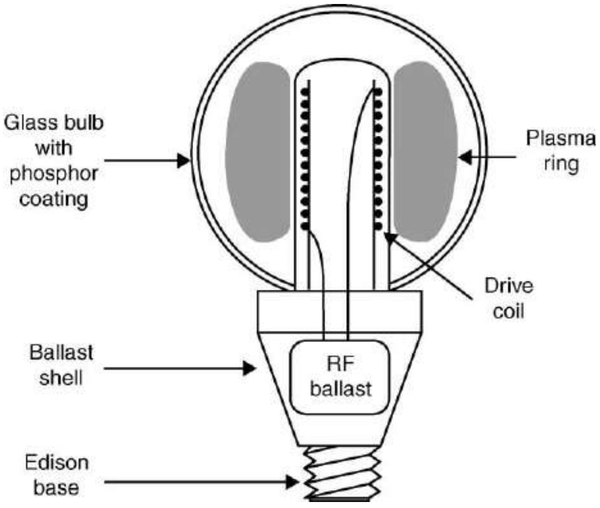

Figure 6 Open core electrodeless lamp with integral ballast. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

shown in Figure 7 operates near 250 kHz with an external ballast, while the open core design shown in Figure 8 operates near 2.5 MHz. Figure 9 shows an open core design with an integral ballast that also operates near 2.5 MHz. Some people have suggested that electrodeless fluorescent lamps are fundamentally different from normal fluorescent lamps and the high frequency field drives the phosphor to emit light. This is incorrect. Other than the manner in which the electric field is created, and the new shapes this allows, electrodeless fluorescent lamps operate in the same manner as normal fluorescent lamps.

Due to the strong dependence of fluorescent performance on mercury temperature, these lamps are not useful as flux standards. However, the five mercury visible lines listed above, plus the 365 nm line of mercury, make the ordinary fluorescent lamp an ideal source for wavelength calibration of spectrometers. Germicidal lamps, which do not use phosphor but are constructed in quartz envelopes and use the same type of mercury-rare gas discharge as fluorescent lamps, are a convenient source of 254 nm radiation that can be used to calibrate UV spectrometers. Care should be taken when using these lamps, as the 254 nm radiation can damage human eyes and skin.

High Intensity Discharge Lamps

If the mercury pressure in a discharge lamp is raised from 6×10^{-3} Torr to an atmosphere or

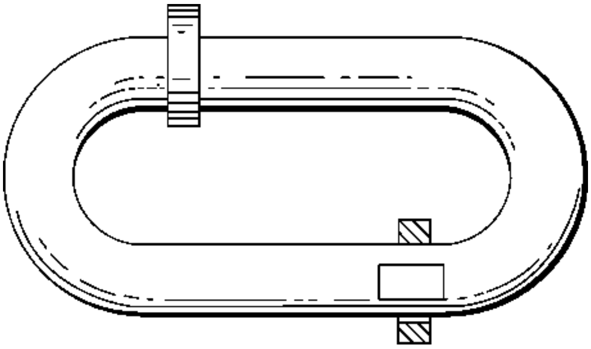

Figure 7 Closed core electrodeless fluorescent lamp, Osram Sylvania Icetron/Endura Lamp, from US Patent 5,834,905.

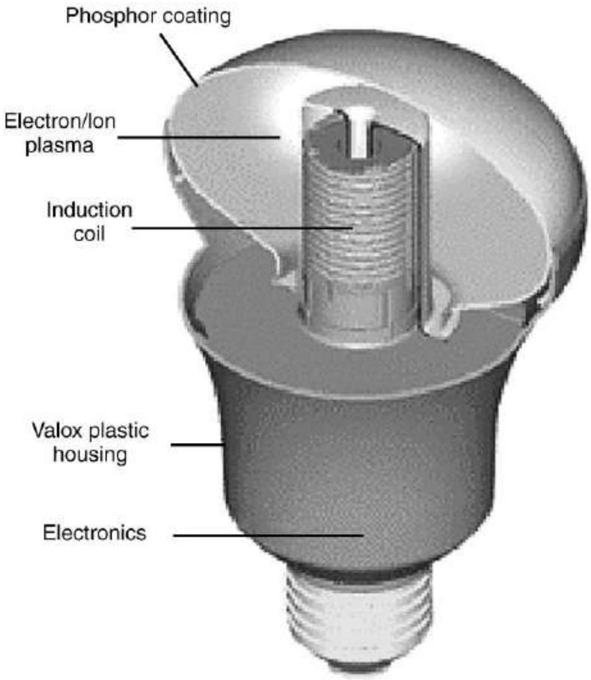

Figure 9 GE Genura® open core electrodeless fluorescent lamp with integral ballast. Courtesy of General Electric Company.

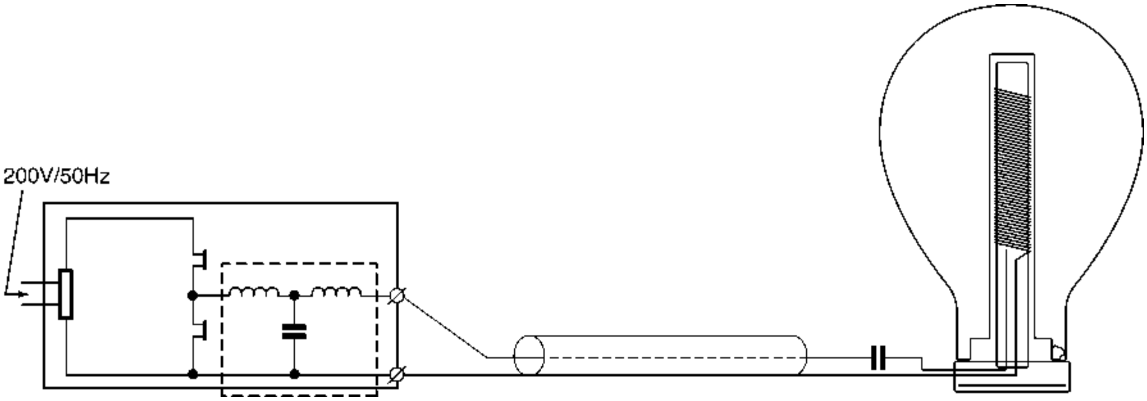

Figure 8 Open core electrodeless fluorescent lamp with separate ballast, Philips QL lamp, from US Patent 6,373,198 B1.

greater, the 245 nm and 185 nm resonance lines are trapped via absorption by ground-state mercury atoms and the radiation output from the discharge shifts to the five optically thin visible lines listed in the discussion of fluorescent lamps, plus the 365 nm near-UV line and a number of lines near 300 nm. The result is a high pressure mercury lamp that operates with an efficacy of about $55\,\mathrm{l\,m\,W^{-1}}$ and has a CRI of about 50. The high-mercury vapor pressure requires a wall temperature of at least 360 °C, which in turn requires that the arc tube be constructed from quartz and operated at high power density. A 400-watt high-pressure mercury arc tube has a diameter of 22 mm and length of about 100 mm. In contrast, a 32-watt fluorescent lamp has a diameter of 25.4 mm and a length of about 1200 mm. This accounts for the classification of these high pressure lamps as high intensity discharge (HID) lamps. Due to the high temperature of the arc tube, most HID lamps are constructed with the arc tube mounted inside a hard glass outer jacket that protects both the user and the arc tube. The high-pressure mercury lamp is a direct visible radiator, but the arc tube does generate a small amount of radiation in the 254 nm to 365 nm region, that radiation is trapped by the outer glass jacket. Some high-pressure mercury lamps have a phosphor coating on the inside of the outer jacket to convert this UV radiation into visible radiation, mostly in regions of the spectrum where the high-pressure mercury arc is deficient in radiation, thus improving the color of the lamp.

Most high-pressure mercury lamps operate at a mercury pressure of about four atmospheres. Because their spectrum is dominated by mercury line emission, the output of these lamps is deficient in the red portion of the spectrum and the lamps have poor color rendition. In 1992, Philips received a US patent for a new type of ultra-high pressure mercury lamp that they call UHP for 'Ultra High Performance'. This lamp operates at a mercury pressure of 200 atmospheres or more. The output spectrum is characterized by a significant amount of continuum radiation from quasi-molecular states of mercury and this continuum includes energy in the red portion of the spectrum that is missing from conventional high pressure mercury lamps. UHP lamps also include a small amount of halogen, such as bromine, which creates a tungsten-halogen cycle similar to that discussed earlier in relation to tungsten-halogen incandescent lamps. During lamp operation, tungsten is sputtered and evaporated from the electrodes. In the absence of the halogen fill, this tungsten rapidly deposits on the walls, leading to very short lamp life. The halogen combines with most of the tungsten before it has a chance to reach the wall of the arc tube. The halogen will also combine with any tungsten that has deposited on the wall of the arc tube. The tungsten-halogen compound formed is gaseous at the operating temperature of the lamp and will diffuse back toward the core of the discharge where the high temperature of the arc dissociates the tungsten-halogen compound. The tungsten is deposited back on the electrodes, thereby greatly extending the life of the lamp. Due to their very high pressures, UHP lamps are constructed in small discharge tubes and have a very short discharge length, typically 1 to 2 mm, making these lamps a virtual point source. Due to their excellent spectral characteristics and small source size, UHP lamps have found application in various projection applications, including LCD-based electronic projectors.

The performance of high-pressure mercury lamps can be substantially improved by adding metals such as sodium and scandium to the high-pressure mercury arc. The sodium emits near the peak of the eye sensitivity curve, providing high efficacy, while the scandium has many transitions in the visible, which fill out the spectrum and provide high CRI. If metallic elements are added directly to the high-pressure mercury arc they would deposit on or react with the wall of the arc tube. To prevent this they are added as iodides, such as NaI or $\mathrm{ScI_3}$. This creates a regenerative cycle in which the NaI and $\mathrm{ScI_3}$ are heated and dissociated in the 5000 K core of the arc, allowing the free metals to radiate. As the metals diffuse toward the cooler sections of the arc near the walls, they recombine with the iodine to reform the iodides. These metal halide lamps have produce light with an efficacy of about $90\,\mathrm{l\,m\,W^{-1}}$ to $100\,\mathrm{l\,m\,W^{-1}}$ and a CRI of about 65 to 75.

The high-pressure sodium (HPS) lamp is an HID lamp that uses hot sodium vapor to generate a golden white light. The lamp uses a mixture of sodium, mercury, and rare gas such as xenon. The core temperature of the sodium-mercury-xenon discharge is typically about 4000 K, while the temperature at the edge of the discharge and the walls of the ceramic arc tube is typically about 1500 K. At low pressure, emission from sodium is dominated by the resonance lines at 589.0 and 589.6 nm, so the light is virtually monochromatic. However, in the HPS lamp, the sodium pressure is typically 75 Torr, which is high enough to almost completely trap emission from the resonance lines, forcing radiant energy to escape in the wings of these lines, thereby increasing the color quality of the light. Since sodium will attack quartz, this mixture is contained in a translucent alumina (aluminum oxide) ceramic arc tube. As with other HID lamps, the alumina arc tube is contained in a hard glass outer bulb. HPS lamps are the most efficient generators of 'white light' available. A typical 400 watt HPS lamp produces light with an efficacy of $125\,\mathrm{l\,m\,W^{-1}}$, this high efficacy being due to the fact that the emission matches the peak of the eye sensitivity curve. The light has a CRI of only 22; however, the HPS lamp is widely used for roadway and other outdoor lighting applications.

Solid State Sources

The newest incoherent sources to be used for general purpose lighting are solid state devices: light emitting diodes (LEDs), and organic light emitting diodes (OLEDs). LEDs are very similar to laser diodes but without the two mirrors used to form the laser cavity in laser diods. Photons are generated as a result of electron-hole recombination as electrons move across the energy gap of a biased p-n junction. By appropriate selection of materials, photons can be created at visible wavelengths. Low brightness infrared and red LEDs have been available for quite some time. It was not until the mid-1980s that high brightness red LEDs were developed. These were followed by increasingly more challenging yellow, green, and then blue high brightness LEDs, developed in 1994 by Shuji Nakamura, of Nichia Chemical Industries. Nakamura added phosphor to his high brightness blue LEDs in 1996 and developed the first 'white' LED, thus triggering the current level of excitement concerning the use of LEDs to replace incandescent and discharge lamps. Monochromatic high brightness LEDs have already replaced color-filtered incandescent lamps in many applications, such as traffic lights, vehicle signal lights, and signage. In these applications, LEDs are the clear winner in life and efficacy. However, when compared to unfiltered sources, the best white LEDs are about as efficient as tungsten halogen incandescent lamps and far less efficient than fluorescent and metal halide lamps. They also cost far more per lumen than even the most expensive discharge lamp. LEDs have high internal quantum efficiency, so the main technical challenges are to fabricate devices to minimize internal light trapping in order to improve device efficiency and reduce the cost per lumen produced.

Organic light emitting diodes (OLEDs) generate light by the passage of electric current through special organic molecules, using a process known as electrophosphorescence. The electrophosphorescent material is usually coated on a transparent and perhaps flexible substrate, and overcoated with a transparent conductor. OLEDs currently have lower efficacy and shorter life than LEDs, but they also have lower manufacturing costs. If LEDs are the 'new' point sources that replace incandescent lamps and perhaps CFLs, then OLEDs are the 'new' diffuse sources that some say will replace liner fluorescent lamps.

See also

Light Emitting Diodes.

Further Reading

Cobine JD (1958) *Gaseous Conductors: Theory and Engineering Applications*. New York: Dover Publications.

de Groot JJ and van Vliet JAJM (1986) *The High Pressure Sodium Lamp*. Philips Technical Library. Deventer–Antwerprn: MacMillian Education.

Elenbass W (1951) *The High Pressure Mercury Vapor Discharge*. Amsterdam: North-Holland Publishing Company.

Fisher HE (1996) *High Pressure Mercury Discharge Lamp*, U.S. Patent, 5,497,049.

Fisher HE and Hörster H (1998) *High-Pressure Mercury-Vapor Discharge Lamp*, U.S. Patent 5,109,181.

Godyak VA (2002) Bright Idea: Radio-frequency light sources. *IEEE Industry Applications Magazine* 8: 42–49.

Hirsh MN and Oskam HJ (1978) *Gaseous Electronics, Volume 1, Electrical Discharges*. New York: Academic Press.

Mitchel AGC and Zemansky MW (1971) *Resonance Radiation and Excited Atoms*, 2nd edn. London: Cambridge University Press.

Nichia's Shuji Nakamura: Dream of the Blue Laser Diode, ScienceWatch, January/February 2000. Available at: http://www.sciencewatch.com/jan-feb2000/sw_jan-feb2000_page3.htm

Pekarski P, Hechtfischer U, Heusler G, Koerber A, Moonch H, Retsch JP, Weichmann U and Kitz A (2003) *UHP Lamps for Projection Systems*, XXVI International Conference on Phenomena in Ionized Gasses, Greifswald, Germany. Available at: http://www.icpig.uni-greifswald.de/proceedings/data/Pekarski_1

Raizer YP (1991) *Gas Discharge Physics*. Berlin: Springer-Verlag.

Rea MS (2000) *IESNA Lighting Handbook*, 9th edn. New York: Illuminating Engineering Society of North America.

Roberts VD (2003) *Radio-Frequency Lighting*. McGraw-Hill Yearbook of Science & Technology, pp. 349–352. New York: The McGraw-Hill Companies, Inc.

Schewe PF and Stein B (1999) *Physics News Update*, Number 437 (Story #2), The American Institute of Physics Bulletin of Physics News. Available at: http://www.aip.org/enews/physnews/1999/split/pnu437-2.htm

Waymouth JF (1971) *Electric Discharge Lamps*. Cambridge, Massachusetts: MIT Press.

Wharmby DO (1993) Electrodeless Lamps for Lighting: A Review. *IEE Proceedings-A* 140(6): 465–473.

Synchrotrons

R Clarke, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Since their discovery by Röntgen in 1895, X-rays have been used in a wide variety of applications ranging from medical and industrial radiography to crystallographic studies of the structure of biomolecules. The short wavelength ($\lambda \sim 0.1$ nm) and high penetrating power of hard X-rays makes them an ideal probe of the internal structure and composition of condensed matter. Indeed, Watson and Crick's elucidation of the double helix structure of DNA in the early 1950s, based on the X-ray diffraction patterns of Rosalind Franklin, has revolutionized biology and provided an atomistic view of the stuff of life.

The quest for ever higher spatial resolution, and the desire to obtain high-quality X-ray data on shorter time-scales, have both been prime motivations for the development of high intensity hard X-ray sources since the mid-1970s. This evolutionary trend has led to the design and construction of large synchrotron radiation facilities, serving many thousands of users and enabling forefront experimental studies across a broad range of disciplines. One such example of a national facility, dedicated to high-brightness X-ray synchrotron radiation, is the Advanced Photon Source (APS), located at Argonne National Laboratory in the United States (Figure 1). Facilities of a similar scope and purpose have also been constructed in Europe and Japan, and a number of new high brightness sources are currently under construction in several countries around the world (see Table 1).

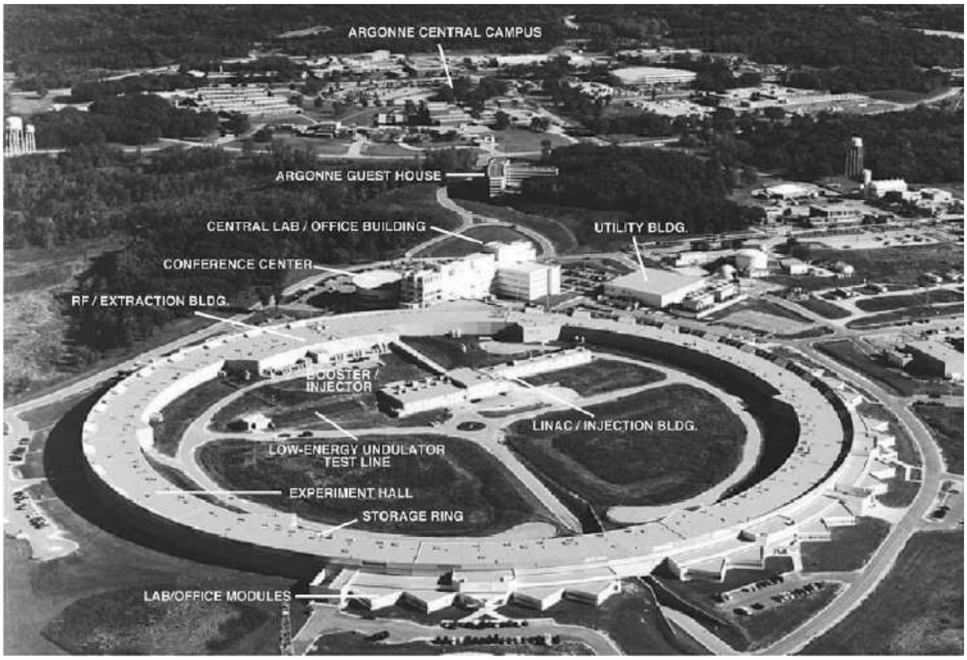

Figure 1 Aerial view of the Advanced Photon Source, Argonne National Laboratory. Courtesy of Advanced Photon Source.

Table 1 Third-generation synchrotron radiation sources

Location	*Facility*	*Electron energy (GeV)*
Grenoble, France	European Synchrotron Radiation Facility (ESRF)	6
Argonne, Illinois, USA	Advanced Photon Source (APS)	7
Nishi Harima, Japan	8-GeV Super Photon Ring (SPring-8)	8
Berkeley, CA, USA	Advanced Light Source (ALS)	1.5–1.9
Canberra, Australia	Boomerang[a]	3
Didcot, UK	DIAMOND[a]	3
Orsay, France	SOLEIL (LURE)	2.75
Karlsruhe, Germany	Angströmquelle-Karlsruhe (ANKA)	2.5
Hamburg, Germany	PETRA III[b]	6
Indore, India	INDUS II[a]	2.5
Villigen, Switzerland	Swiss Light Source (SLS)	2.4
Saskatoon, Canada	Canadian Light Source[a]	2.9
Stanford, CA, USA	SPEARS	3
Upton, New York, USA	NSLS II[b]	2.5–2.8
Trieste, Italy	Elettra	2.0–2.4

[a] Under construction as of March 2004.
[b] Planning stage.

In this article we will draw on examples from APS to illustrate the characteristics and performance of the current generation of synchrotron sources in the hard X-ray region.

From the earliest days of particle physics studies in the 1930s and 40s, it was realized that copious amounts of X-rays could be produced by accelerating charged particles in a curved trajectory, as in a cyclotron or a synchrotron. For particle physics such radiation is highly detrimental because it limits the ultimate energy of particle accelerators and forces their circumference to become inordinately large in order to reduce radiation losses.

Originally considered to be a major source of concern in particle accelerator design, synchrotron radiation soon began to emerge as a boon to X-ray science. This came about in the early 1970s after the capability was developed to produce and store significant circulating currents of high-energy charged particles (electrons or positrons) for colliding beam experiments. At that time X-ray scientists were tolerated as ‘parasitic’ experimenters at such storage rings, taking X-ray radiation from the bending magnets which maintained the stored current in an overall circular trajectory. The PETRA (Positron–Electron Tandem Ring Accelerator) facility at HASYLAB/DESY in Hamburg, Germany, and the Cornell Electron Storage Ring (CESR)/Cornell High-Energy Synchrotron Source (CHESS) at Cornell University, Ithaca, New York, are examples of accelerator facilities that were primarily intended for particle physics experiments and then became partly dedicated to X-ray production. Such shared synchrotron facilities are referred to as ‘first-generation sources’.

The first generation sources demonstrated several attractive features of synchrotron radiation, including a high degree of collimation in the vertical direction with a divergence half-angle given approximately by the relativistic factor $\gamma^{-1} = m_0c^2/E$, where m_0c^2 is the rest-mass energy (0.51 MeV) of the electron (or positron) and E is the energy of the stored particle beam. For a bending magnet source with $E = 2-3$ GeV the beam would have a vertical collimation of ~ 0.3 mrad (17 mdeg). This would produce an X-ray beam with a vertical extent of ~ 3 mm at a distance of 10 m from the bending magnet source. The horizontal fan of bend magnet radiation is broad and is usually limited by a mask or a slit to a few mrad. The most useful characteristic of bend magnet radiation is its large bandwidth whereby the X-ray spectrum is continuous and increasing in intensity (see Figure 2) until some critical energy above which the intensity begins to fall off. Bending magnet radiation is therefore useful for X-ray measurements requiring a ‘white’ beam, for example, Laue X-ray diffraction, energy dispersive diffraction, and Extended X-ray Absorption Fine Structure (EXAFS) analysis.

Beginning in the 1980s, several second-generation storage rings were constructed in various locations around the world. The design of these facilities for the first time was optimized for the production of bright beams of X-ray and XUV (soft X-ray to ultraviolet) and their operation was dedicated exclusively to synchrotron radiation science. The increase in brightness provided by these new storage ring sources, compared to standard laboratory X-ray sources (e.g., rotating anode generators), was dramatic (5 or 6 orders of magnitude) and this opened up many new experiments that were not previously feasible. However, it was soon realized that the capabilities of such sources could be extended to

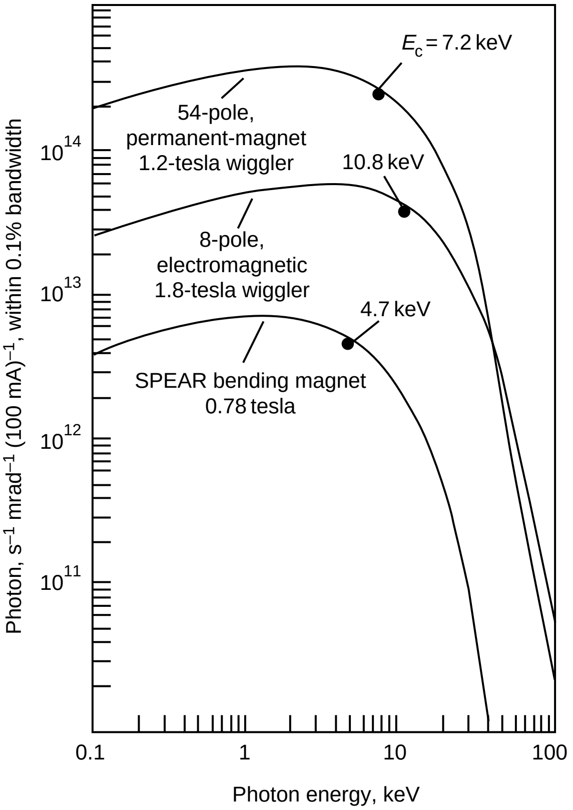

Figure 2 Comparison of bending magnet and wiggler sources. With permission from Stanford Synchrotron Radiation Laboratory, Stanford Linear Accelerator Center, Menlo Park, CA.

even higher levels of intensity and to significantly higher X-ray energies.

One of the limitations of bending magnet sources relates to the X-ray critical energy alluded to above:

$$E_c = 3\hbar c\gamma^3/4\pi\rho \qquad [1]$$

where ρ is the radius of curvature imposed on the particle beam trajectory by the bending magnet. For a given particle acceleration energy, E, eqn [1] leads to a critical X-ray energy given by:

$$E_c\,(\text{keV}) = 2.2\,E^3\,(\text{GeV})/\rho\,(\text{m}) \qquad [2]$$

For a typical second-generation machine with $E = 2.5$ GeV and $\rho = 7$ m (limited by the magnetic field generated by the dipole electromagnet), the critical energy is only about 5 keV. There is useable flux well above this energy, but X-ray photons become scarce in the energy range above 10 keV for a bending magnet source with these characteristics. Since E is effectively fixed by the design parameters of the storage ring, the only possibility to achieve higher photon energies is to install additional magnetic devices which will introduce perturbations of the trajectory with effectively smaller radius of curvature in the denominator of eqn [2]. Such a device, known as a 'wiggler', shifts the effective critical X-ray energy to higher values in proportion to the strength of the magnetic field B (see Figure 2). Using wigglers based on superconducting magnets ($B \sim 5$ Tesla), E_c can be raised to >20 keV, a value which would require accelerator energies of more than 6 GeV using conventional bending magnets alone. In addition to shifting the output spectrum to higher energies, it is possible to enhance the intensity at a given energy by producing multiple wiggles with no net displacement of the particle trajectory. Multipole wigglers are able to produce significant increases in flux in the hard X-ray range. The 27-pole hybrid wiggler installed on beamline X25 at the National Synchrotron Light Source, a second generation facility at Brookhaven National Laboratory, Upton, New York, is one of the most successful designs of its kind, providing $\sim 6 \times 10^{14}$ photons/sec per 1% bandwidth at 8 keV. When focused with a toroidal X-ray mirror, this beam is about an order of magnitude brighter than the standard bending magnet sources at NSLS. For reasons that will become clear later in this article, the radiation from multiple wiggles in such a device adds incoherently.

Insertion Devices

The idea of using wigglers to enhance the characteristics of synchrotron radiation sources has led to the design and construction of storage rings with many straight sections to accommodate wigglers and other types of 'insertion devices'. Such facilities are referred to as third-generation sources. Coming on-line in the mid-1990s, several such sources are now in mature operation and are supporting a large international community of X-ray scientists.

One of the most important types of insertion device to emerge from the development of third-generation facilities is the undulator. Conceptually, the idea behind an undulator is similar to that of the wiggler, i.e., it provides additional tightly curved transverse displacements in the particle beam. There are two important differences: in an undulator the displacements are smaller and they are more numerous, being produced by a linear periodic array of many short alternating dipoles (Figure 3). Under these conditions the synchrotron radiation emitted from each 'undulation' of the particle beam trajectory adds coherently and therefore the peak intensity from such a device can be enhanced by a factor of N^2 where N is the number of dipole magnet pairs, or periods.

A simple theoretical description of the basic operation of an undulator can be discussed in terms of radiation from a charged particle moving at

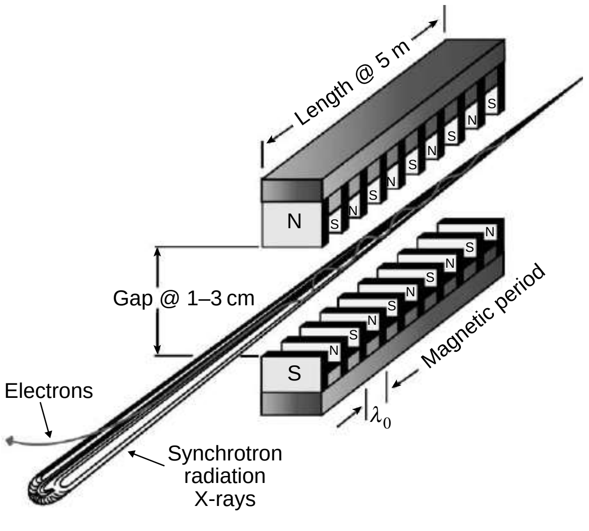

Figure 3 Schematic of undulator. Courtesy of Advanced Photon Source.

relativistic speeds through an alternating magnetic field. The magnetic field in the vertical direction (y) is taken to be periodic:

$$B_y = B_0 \sin[(2\pi/\lambda_{\mathrm{u}})z] \qquad [3]$$

where z is in the direction of average particle motion.

The transverse particle trajectory is described by:

$$x = (\lambda_{\mathrm{u}}\phi_{\max}/2\pi)\sin[(2\pi/\lambda_{\mathrm{u}})z] \qquad [4]$$

where λ_{u} is the magnet period, and $\phi_{\max}$ is the maximum particle deflection angle:

$$\phi_{\max} = (eB_0\lambda_{\mathrm{u}}c)/2\pi m_0c^2\gamma = K/\gamma \qquad [5]$$

where

$$K = 93.4B_0\ (\text{Tesla})\lambda_{\mathrm{u}}\ (\text{meters}) \qquad [6]$$

The field strength parameter K serves as a useful quantity for discussing the characteristics of various insertion device designs. For $K \leq 1$, which defines the undulator regime, the transverse angular deflections are less than or on the order of the angular spread of the emitted radiation and in this case the radiation from each of the deflections adds coherently. This is distinguished from the case of a multipole wiggler where $K \gg 1$ leading to incoherent superposition.

Most of the interesting features of undulator radiation are revealed by considering the trajectory of the particle in a Lorentz frame moving at speed v^* with respect to the laboratory frame. In this frame the particle will execute oscillatory motion with the period of oscillations T' given by the Lorentz transformation:

$$T' = \gamma^*(T - \beta^*\lambda_{\mathrm{u}}/c) = \lambda_{\mathrm{u}}/(v^*\gamma^*) \qquad [7]$$

with a corresponding oscillation frequency $\omega' = 2\pi/T' = (2\pi c/\lambda_{\mathrm{u}})\gamma^*$.

For $K \ll 1$, the motion in the co-moving frame is purely transverse and nonrelativistic. However, for $K \sim 1$, the oscillatory motion in this frame can have a longitudinal component and in this case the transverse motion becomes nonharmonic with frequency components at odd multiples of the fundamental frequency ω'. This nonharmonic motion is a consequence of the fact that in the laboratory frame the particle is moving with constant speed $\approx c$ but its path length is slightly longer than λ_{u}, owing to the transverse motion. This constrains the motion in the co-moving frame to be nonharmonic.

Undulator Spectrum

It now remains to transform ω' into the laboratory frame so that we can obtain the spectral distribution of the output of the undulator. Invoking the Doppler transformation:

$$\omega = \gamma^*\omega'(1 + \beta^*\cos\theta') \qquad [8a]$$

$$\tan\theta = (\sin\theta')/[\gamma^*(\cos\theta' + \beta^*)] \qquad [8b]$$

where θ is the angle at which the radiation is observed:

$$\omega_{\mathrm{n}}(\theta) = (4\pi cn\gamma^2)/(\lambda_{\mathrm{u}}[1 + \tfrac{1}{2}K^2 + (\gamma\theta)^2]) \qquad [9a]$$

or in terms of wavelength:

$$\lambda_{\mathrm{n}}(\theta) = (\lambda_{\mathrm{u}}/2n\gamma^2)[1 + \tfrac{1}{2}K^2 + (\gamma\theta)^2] \qquad [9b]$$

Equation [9] is the central expression describing the spectral output of an undulator. Notice how the macroscopic period of the undulator λ_{u} is transformed into a microscopic wavelength $\lambda_{\mathrm{n}}(\theta)$ by the relativistic factor γ^2. For third-generation hard X-ray sources with particle energies in the 6–8 GeV range, $\gamma \sim 14\,000$, taking cm scales down to the angstrom level.

Equation [9] describes a spectrum of monochromatic peaks, for on-axis radiation ($\theta = 0$), with wavelengths that increase with increasing K. Since $K \sim B_0\lambda_{\mathrm{u}}$ (eqn [6]) any particular harmonic can be tuned either by changing B_0 or λ_{u}. In most undulator designs, λ_{u} is fixed by the construction of the permanent magnet array, so the tuning is conveniently accomplished by varying B_0. In practice this is done by changing the vertical distance between the poles

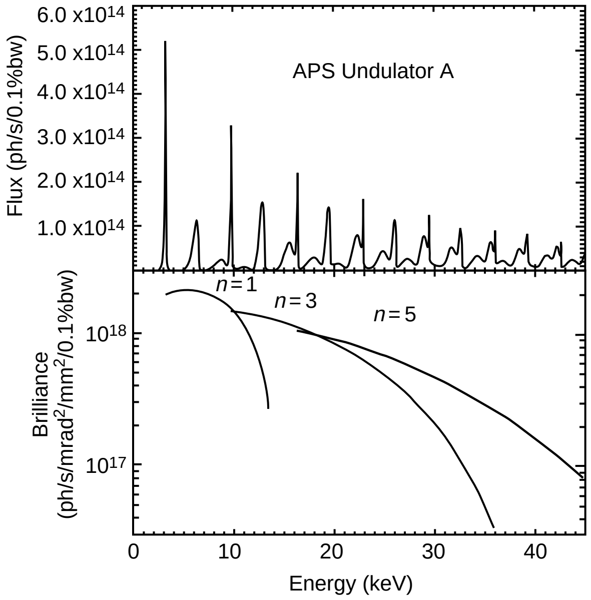

Figure 4 Undulator spectrum (upper panel) and tuning range for first, third and fifth harmonics (lower panel) of Advanced Photon Source Undulator A.

of the magnets (the 'undulator gap') similar to opening the jaws of a vise. Typical tuning ranges for the $n = 1$, 3, and 5 harmonics are illustrated in Figure 4.

Since the number of undulator oscillations of the particle beam is finite, the nth harmonic is not strictly monochromatic and has a spectral broadening of order $1/nN$, where N is the number of oscillator periods. In practice with N typically ~100, a bandwidth of ~1% is expected for the first harmonic. Additional broadening of a similar magnitude might also be expected from the θ-dependence in eqn [9], depending on the acceptance angle of the optics that is used in a particular experiment. This quasi-monochromatic radiation is sometimes referred to as 'pink beam'.

The storage ring energy of the APS was carefully chosen so that the tuning ranges of the different harmonics would overlap in energy, thus providing a continuous tuning of the X-ray energy over a wide range (Figure 4). Note also that the discrete spectral characteristics of the undulator can be deliberately smeared out to mimic the broadband (white beam) characteristics of a bend magnet (with consequent reduction in brightness). This is done simply by tilting the jaws of the undulator with respect to each other so that B_0 is made to vary along the length of the undulator. The ability to taper the undulator in this way is a standard feature of most designs.

The radiation from an undulator is sharply peaked in the forward direction in both the vertical and horizontal directions. The half-angle of the emission cone is given by:

$$\theta_{1/2} = (1/\gamma)[(1 + \tfrac{1}{2}K^2)/2Nn]^{1/2} = [\lambda_n/L]^{1/2} \qquad [10]$$

where L is length of the undulator. For a typical third-generation source undulator, such as the Type A undulator at the Advanced Photon Source, $L = 2.4$ m, giving $\theta_{1/2} \approx 6$ μrad for the first harmonic tuned to 12 keV. This corresponds to an X-ray beam size of ~180 μm at a distance 30 m from the source.

The very tight collimation of undulator radiation (eqn [10]) highlights the importance of distinguishing between the brightness and the 'brilliance' of the beam. The latter term takes into account the natural collimation of the beam, as reflected in its units of photons/sec/mm^2/mrad^2/0.1% bandwidth. Undulator radiation is well matched to the acceptance angle of ideal crystal monochromators and is ~10 000 times more brilliant than the radiation from a bend magnet source. The high brilliance of undulator X-ray beams is extremely advantageous in X-ray imaging applications, X-ray microprobe measurements, and in experiments where the sample size is very small (e.g., high-pressure diamond anvil cell measurements).

In practice the emission cone of the source might be broadened a little when the undulator characteristics are convolved with those of the storage ring. The largest effect is in the horizontal direction where the photon beam divergence is dominated by the electron beam divergence. This is a consequence of the particular conditions under which the storage ring is operated in order to stabilize the particle beam orbit.

Equation [10] demonstrates that preserving the exquisite performance capabilities of an undulator insertion device places severe demands on the stability and precision of the storage ring orbit. An important measure of this precision is characterized by the 'emittance' of the particle beam defined in terms of the area of phase space (position and angular divergence) occupied by the stored beam. In general, storage rings are designed to achieve very low emittances and in current third-generation sources they are close to being diffraction limited. Table 2 summarizes the characteristics of the APS storage ring operated in 'high-emittance' and 'low-emittance' modes.

The large power densities resulting from the extreme brightness of undulator sources (up to 160 kW/mrad^2 for APS Undulator A) necessitates careful design of the X-ray optical components that are used to condition the beam. For example, if a double-bounce Silicon crystal pair is used as the first optical component to monochromatize the undulator

Table 2 Advanced photon source storage ring parameters

Parameter	*Low emittance*	*High emittance*
Storage ring energy, E	7.0 GeV	7.0 GeV
Storage ring current, I	100 mA	100 mA
Beam energy spread, $\delta E/E$	0.096%	0.096%
Horizontal emittance, ε_x	3.5×10^{-9} m-rad	7.7×10^{-9} m-rad
Vertical emittance, ε_y	3.5×10^{-11} m-rad	7.7×10^{-11} m-rad
Coupling constant	1%	1%
Horizontal beta function, β_x	14.4 m	16.1 m
Vertical beta function, β_y	4.0 m	4.3 m
Dispersion function, η_x	0.124 m	0.0 m
Horizontal beam size, σ_x	254 μm	351 μm
Vertical beam size, σ_y	12 μm	18 μm
Horizontal beam divergence, $\sigma_{x'}$	15.6 μrad	21.8 μrad
Vertical beam divergence, $\sigma_{y'}$	3.0 μrad	4.2 μrad

beam, the first crystal is normally cooled with liquid nitrogen to dissipate the heat. Alternately, high thermal conductivity diamond crystals are sometimes used as monochromators.

Helical Undulators

The use of undulators to produce extremely bright, tunable X-ray beams has enabled many new applications of X-ray science, as well as offering significant improvements to existing techniques in terms of quality of data and the time required to accumulate a high degree of statistical accuracy. One of the areas that has benefited significantly from novel undulator design is research on magnetic materials and other spin-dependent phenomena. It is often advantageous to probe such systems with circularly polarized X-rays, for example to study magnetic X-ray circular dichroism (MXCD). Early MXCD measurements utilized the left and right elliptical polarization of bending magnet radiation above and below the plane of the ring. Such measurements are difficult because the intensity falls off dramatically out of the plane. The problem to produce high-brightness beams of circularly polarized X-rays is effectively solved by designing an undulator which produces a transverse periodic helical magnetic field of constant magnitude. Several helical undulators are now in operation at third-generation sources. The soft X-ray region (500–1500 eV) is particularly interesting because it covers the L-edges of the 3d transition elements.

Coherence

While this article describes what would be nominally considered as 'incoherent' sources in the parlance of conventional optics, it should be pointed out that undulators are in fact excellent sources of coherent X-rays. This is by virtue of the high degree of collimation inherent in the emission of radiation from the undulations of a highly relativistic charged particle (eqn [10]). The spatial (transverse) coherence angle is given by:

$$\lambda_n/2d \qquad [11]$$

where d is the source size. This is the angle within which the beam is spatially coherent.

For APS undulator A, the source size is ~15 μm (vertical) × 240 μm (horizontal), giving a coherence angle in the vertical direction of 6.5 μrad and 0.4 μrad in the horizontal direction. The X-ray beam divergence is approximately 4 μrad (vertical) by 15 μrad (horizontal). This means that the beam has complete spatial coherence in the vertical direction and is coherent over about 3% of the horizontal fan. At a distance of 30 m from the source, an on-axis pinhole of ~12 micron diameter will be completely illuminated by coherent X-rays. Transversely coherent X-ray beams are finding novel applications in X-ray photon correlation spectroscopy, coherent X-ray scattering, and phase reconstruction (lensless) imaging.

Timing Structure

X-ray storage rings produce pulsed X-rays. In the Advanced Photon Source, one or more long trains of electron bunches (15 ns long) are accelerated by the linac to 325 MeV. The electron bunches are injected into a booster that ramps the single bunch from 325 MeV to 7 GeV in about 225 ms. At the end of the booster cycle, the bunch is extracted and injected into the storage ring. The bunch can be injected into any of 1296 'buckets' that are spaced by 2.842 ns, forming any arbitrary pattern after several injection cycles.

The APS storage ring rf frequency of 351.927 MHz determines the bucket spacing of 2.842 ns. The circumference of 1104 m gives a revolution frequency of 271.5 kHz which allows for 1296 buckets in 3683 ns. The storage ring is filled with a selected bunch pattern depending on the requirements for a particular operations cycle. The standard 'singlets' bunch pattern at APS is 102 mA of average current in a train of 23 bunches (4.43 mA per bunch) each spaced by 153 ns (54 buckets). There is a gap of 306 ns (108 buckets) in this pattern, useful as a timing guide. The singlets pattern has a lower lifetime than patterns with less charge per bunch, due to natural internal scattering processes. A 'top-up' mode has been devised to overcome the problem of overall intensity fall-off due to the exponential decay of the stored current. In top-up mode, new bunches are injected into the storage ring every few minutes to maintain the stored current at 102 mA.

A second fill pattern, known as the 'hybrid' pattern, is again 102 mA but distributed in one bunch of 5 mA, plus 8 groups of 7 consecutive bunches (1.73 mA/bunch), spaced by 68 ns (24 buckets). The 8 groups are diametrically opposite the 5 mA bunch, allowing for a 1.5 μs gap on both sides of the 5 mA bunch. This hybrid mode permits a single bunch to be isolated using a mechanical chopper or a fast framing camera. Since the RMS width of the single bunch is ~36 ps, relatively fast time-resolved experiments can be performed stroboscopically using this pulsed X-ray beam. In another implementation of time-resolved experiments, the singlet pattern is used, isolating each bunch with a fast avalanche photodiode detector, or an X-ray streak camera. With the streak camera it is possible to achieve time resolutions of a few picoseconds. This approach has been used to study impulsive phonon excitations and ultrafast strain wave propagation in solids.

The Future

This article highlights current (third-generation) storage ring sources based on undulator-type insertion devices. These sources will provide high-brightness tunable X-rays in support of a wide range of experiments for many years to come. The brightness of these sources will increase steadily into the future as continual improvements are made, such as further reducing the emittance of the particle beam, improved undulator design, increasing the average storage ring current, and making changes to the storage ring lattice. These incremental improvements will be important to maintain the vitality and usefulness of national user facilities based on third-generation sources. New storage ring sources with designs and specifications optimized for low emittance and new types of undulators (e.g., in-vacuum undulators) will come online and will undoubtedly lead to new technical advances, supporting new and exciting science. It seems likely that linear accelerator (linac) based undulator sources will provide a promising route towards higher brilliance in the future. The superconducting Energy Recovery Linac (ERL) proposed by CHESS is an interesting concept which uses regeneration of the electron beam to overcome the high energy cost of single pass particle acceleration.

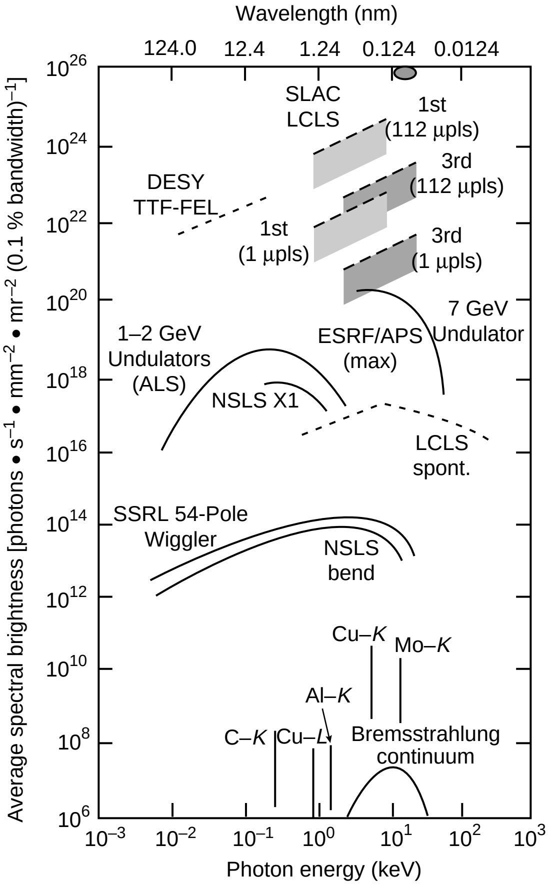

Figure 5 Comparison of spectral brilliance of X-ray sources. Courtesy of Linear Coherent Light Source, Stanford Linear Acceleration Center.

A new generation of X-ray sources (the 'fourth generation') is envisioned to take on the task of providing vastly brighter beams (>10 orders more peak brilliance than at present, see Figure 5) with much shorter pulses (~100 fs). These facilities will also be based on linear accelerators. Two such *Free Electron Laser* sources are currently being designed and should be operational towards 2010: the Linac Coherent Light Source (LCLS), to be built at the Stanford Linear Accelerator Center (SLAC), and the X-ray Free Electron Laser (XFEL) Source, to be constructed by the TESLA collaboration at HASYLAB, Hamburg.

See also

Coherence: Overview. **Physical Applications of Lasers:** Free Electron Lasers in Physics.

Further Reading

Brown G and Lavender W (1991) Synchrotron radiation spectra. In: Brown GS and Moncton DE (eds) *Handbook on Synchrotron Radiation*, vol. 3. Amsterdam: North Holland.

Duke PJ (1993) Synchrotron radiation sources. In: Michette AG and Buckley CJ (eds) *X-Ray Science and Technology*. Bristol: Institute of Physics Publishing Ltd.

Duke PJ (2000) *Synchrotron Radiation: Production and Properties*. Oxford, UK: OUP.

Mills DM (ed.) (2002) *Third Generation Hard X-ray Synchrotron Radiation Sources*. Berlin: Wiley, VCH.

INFORMATION PROCESSING

Contents

All-Optical Multiplexing/Demultiplexing
Coherent Analog Optical Processors
Free-Space Optical Computing
Incoherent Analog Optical Processors
Optical Bit-Serial Computing
Optical Digital Image Processing
Optical Neural Networks

All-Optical Multiplexing/Demultiplexing

Z Ghassemlooy, Northumbria University, Newcastle upon Tyne, UK

G Swift, Sheffield Hallam University, Sheffield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ever-increasing aggregate demand of electrically based time division multiplexing systems should have coped with the steady growth rate of voice traffic. However, since 1990, the explosive growth of the Internet and other bandwidth-demanding multimedia applications, has meant that long-haul telecommunication traffic has been increasingly dominated by data, not voice traffic. Such systems suffer from a bandwidth bottleneck due to speed limitations of the electronics. This limits the maximum data rate to considerably less than the THz bandwidth offered by an optical fiber. Optical technology is proposed as the only viable option and is expected to play an ever increasing role in future ultrahigh-speed links/networks. There are a number of multiplexing techniques, such as space division multiplexing (SDM), wavelength division multiplexing (WDM), and optical time division multiplexing (OTDM), that are currently being applied to effectively utilize the bandwidth of optical fiber as a means to overcome the bandwidth bottleneck imposed by electrical time division multiplexing (TDM). In SDM a separate optical fiber is allocated to each channel, but this is the least preferred option for increasing channel numbers. In WDM, a number of different data channels are allocated to discrete optical wavelengths for transmission over a single fiber. Dense WDM technology has been improving at a steady rate in recent years, with the latest systems capable of operating at a data rate of >1 T bps, using a large number of wavelengths over a single fiber link. However, there are a number of problems associated with the WDM systems such as:

- Performance of WDM is highly dependent on the nonlinearities associated with fiber, i.e.:
  - *Stimulated Raman scattering*: degrades the signal-to-noise (SNR) as the number of channels increases;
  - *Four-wave mixing*: limits the channel spacing;
  - *Cross-phase modulation*: limits the number of channels.
- Relatively static optical paths, thus offering no fast switching with high performance within the network;

- Switching is normally carried out by separating each wavelength of each fiber onto different physical outputs. Space switches are then used to spatially switch the separated wavelengths, an extremely inefficient way of utilizing network resources;
- The need for amplifiers with high gain and flat spectra.

In order to overcome these problems, OTDM was introduced that offers the following:

- Flexible high bandwidth on demand (>1 Tbit/s compared to the bit rates of 2.5–40 Gbit/s per wavelength channel in WDM systems);
- The total bandwidth offered by a single channel network is equal to DWDM;
- In a network environment, OTDM provides potential improvements in:
  - Network user access time, delay and throughput, depending on the user rates and statistics.
  - Less complex end node equipment (single-channel versus multichannels).
- Self-routing and self-clocking characteristics;
- Can operate at second- and third-transmission windows:
  - 1500 nm (like WDM) due to Erbium doped fiber amplifier (EDFA);
  - 1300 nm wavelengths.
- Offers both broadcast and switched based networks.

Principle of OTDM

Figure 1 show the generic block diagram of a point-to-point OTDM transmission link, where N optical data channels, each of capacity M Gbps, are multiplexed to give an aggregate rate of $N \times M$ Gbps. The fundamental components are a pulsed light source, an optical modulator, a multiplexer, channel, add/drop unit, and a demultiplexer. The light source needs to have good stabilities and be capable of generating ultrashort pulses (<1 ps). Direct modulation of the laser source is possible but the preferred method is based on external modulation where the optical signal is gated by the electronic data. The combination of these techniques allows the time division multiplexed data to be encoded inside a subnanosecond time slot, which is subsequently interleaved into a frame format. Add/drop units provide added versatility (see Figure 2) allowing the 'adding' and 'dropping' of selected OTDM channels to interchange data at chosen points on the link. At the receiving end, the OTDM pulse stream is demultiplexed down to the individual channels at the initial M Gbps data rate. Data retrieval is then within the realm of electronic devices and the distinction between electronic and optical methods is no longer relevant. Demultiplexing requires high-speed all optical switching and can be achieved using a number of methods, which will be discussed in more detail below.

The optical multiplexing (or interleaving) can be carried out at the bit level (known as bit interleaving) or at the packet level (known as packet interleaving), where blocks of bits are interleaved sequentially. This is in accord with the popular conception of packet switched networks. The high data rates required and consequently narrow time slots necessitate the need for strict tolerances at processing nodes, e.g., switches. As such, it is important that the duration

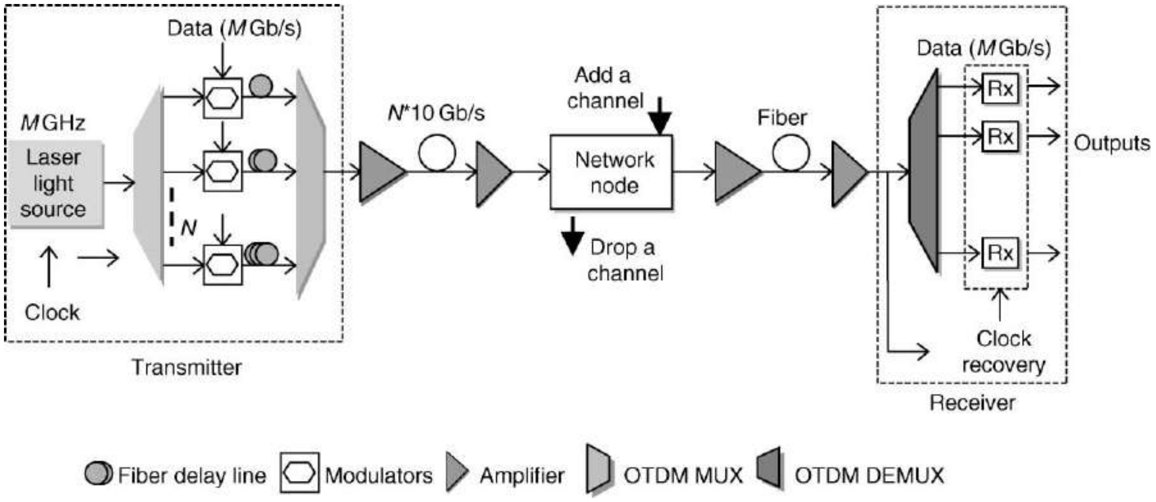

Figure 1 Block diagram of a typical OTDM transmission system.

of the optical pulses is chosen to be significantly shorter than the bit period of the highest multiplexed line rate, in order to reduce the crosstalk between channels.

Bit Interleaved OTDM

A simple conceptual description of a bit interleaved multiplexer is shown in Figure 3. It uses a number of different length optical fiber delay lines (FDL) to interleave the channels. The propagation delay of each FDL is chosen to position the optical channel in its corresponding time slot in relation to the aggregate OTDM signal. Prior to this, each optical pulse train is modulated by the data stream. The output of the modulators and an undelayed pulse train, labeled the framing signal, are combined, using a star coupler or combiner, to produce the high bit rate OTDM signal (see Figure 3b). As shown in Figure 3b, the framing pulse has a higher intensity for clock recovery purpose. At the demultiplexer, the incoming OTDM pulse train is split into two paths (see Figure 4a). The lower path is used to recover the framing pulse by means of thresholding, this is then delayed by an amount corresponding to the position of the ith (wanted) channel (see Figure 4b). The delayed framing pulse and the OTDM pulse stream are then passed through an AND gate to recover the ith channel. The AND operation can be carried out all optically using, for example, a nonlinear loop mirror, a terahertz optical asymmetric demultiplexer, or a soliton-trapping gate.

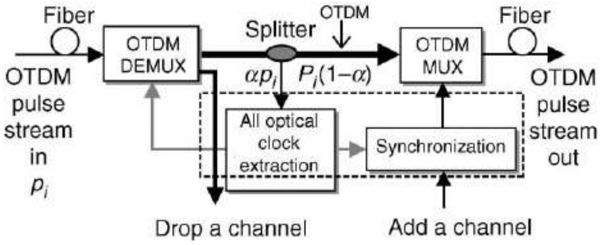
Fiber
OTDM
Splitter
Fiber
OTDM DEMUX
OTDM MUX
OTDM pulse stream in
P_i
αp_i
$P_i(1-\alpha)$
All optical clock extraction
Synchronization
OTDM pulse stream out
Drop a channel
Add a channel

Figure 2 Add/drop unit in an OTDM network node.

Packet Interleaved OTDM

Figure 5a shows the block diagram of a system used to demonstrate packet interleaving. Note that the time interval between successive pulses now needs to be much less than the bit interval T. This is achieved by passing the low bit rate output packet of the modulator into a compressor, which is based on a feed forward delay line structure. The feed forward delay lines are based on a cascade of passive M–Z interferometers with unequal arm lengths. This configuration generates the delay times required, i.e., $T-\tau, 2(T-\tau)\ldots(2^{n-1})(T-\tau)$, etc., where $n = \log_2 k$ is the number of stages, T and τ are the

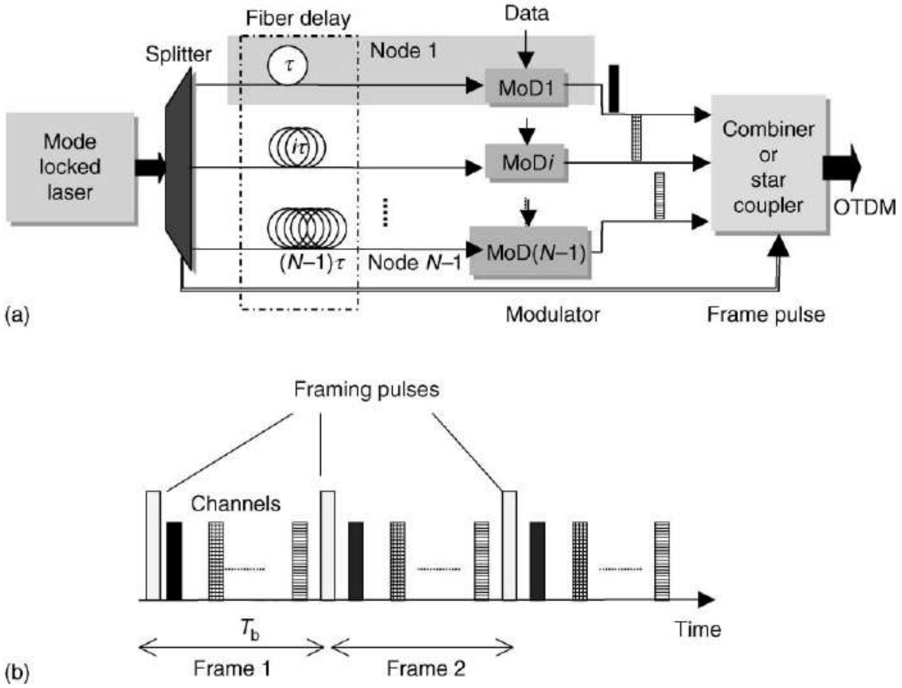
Fiber delay
Data
Splitter
Node 1
τ
MoD1
Mode locked laser
$i\tau$
MoDi
Combiner or star coupler
OTDM
$(N-1)\tau$
Node $N-1$
MoD$(N-1)$
(a)
Modulator
Frame pulse
Framing pulses
Channels
T_b
Time
Frame 1
Frame 2
(b)

Figure 3 Bit interleaved OTDM; (a) block diagram and (b) timing waveforms.

incoming bit duration and the outgoing bit duration, respectively, and k is the packet length in bits. Pulse i's location at the output is given by $(2^n-1)(T-t)+(i-1)\tau$. The signal at the input of the compressor is:

$$I_{\text{in}}(t)=\sum_{i=0}^{k-1}\delta(t-iT)A_i \qquad [1]$$

Where $\delta(\cdot)$ is the optical carrier pulse shape, and A_i is the ith data bit in the packet.

As shown in **Figure 5b**, each bit is split, delayed, and combined to produce the four bit packets O_i:

$$O_i(t)=\frac{1}{2^{n+1}}\sum_{i=0}^{k-1}I_i[t-i(T-\tau)] \qquad [2]$$

Note the factor of $1/(2^{n-1})$ is due to the signal splitting at the 2 × 2 3-dB coupler at the input of each stage.

The combined signal is shown in **Figure 5b**, and is defined as:

$$I_{\text{out}}=\sum_{i=0}^{k-1}O_i=\frac{1}{2^{n+1}}\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}I_i[t-(i+j)T+j\tau]A_i \qquad [3]$$

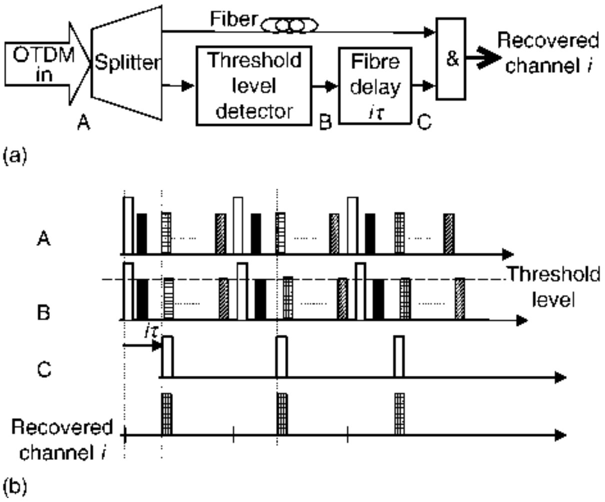
OTDM in
Splitter
A
Fiber
Threshold level detector
B
Fibre delay iτ
C
&
Recovered channel i
(a)
A
B
Threshold level
iτ
C
Recovered channel i
(b)

Figure 4 Bit interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

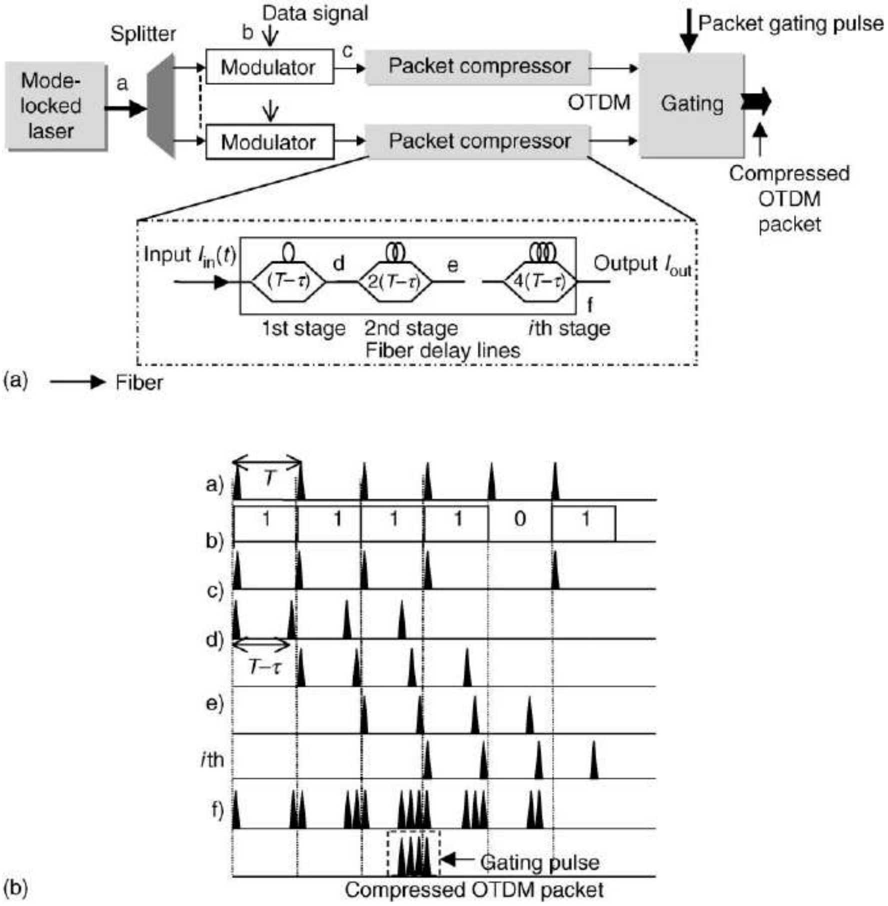
Mode-locked laser
a
Splitter
Data signal
b
c
Modulator
Modulator
Packet compressor
Packet compressor
OTDM
Packet gating pulse
Gating
Compressed OTDM packet
Input I_in(t)
(T−τ)
d
2(T−τ)
e
4(T−τ)
Output I_out
f
1st stage
2nd stage
ith stage
Fiber delay lines
(a)
Fiber
a)
T
b)
1
1
1
1
0
1
c)
d)
T−τ
e)
ith
f)
Gating pulse
Compressed OTDM packet
(b)

Figure 5 Packet interleaving multiplexer: (a) block diagram and (b) typical waveforms.

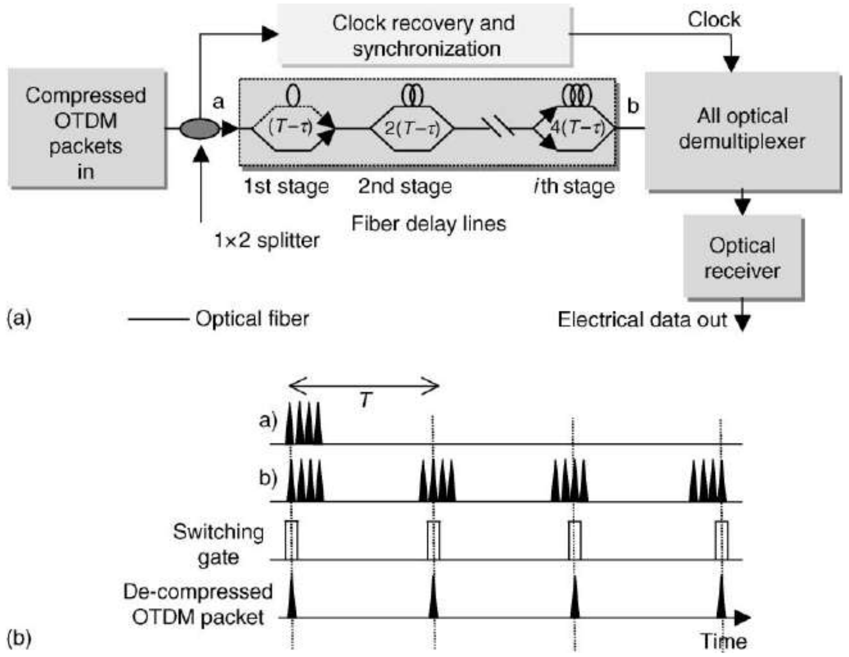

Figure 6 Packet interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

For a packet delay of $(i+j)T$ packets are being built up. For the case $(i+j)=3$, four bits are compressed into a packet (see Figure 5b). An optical gating device is used to select the only complete usable compressed copy of the incoming packets from the unwanted copies.

At the receiving end, demultiplexing is carried out by decompressing the OTDM packet stream using the same delay line structure (see Figure 6). Multiple copies of the compressed packet are generated, and each is delayed by $(T-\tau)$. An optical demultiplexer placed at the output of the delay lines generates a sequence of switching window at the rate of $1/T$. By positioning the switching pulses at the appropriate position, a series of adjacent packet (channel) bits from each of the copied compressed packets is extracted (see Figure 6b). The output of the demultiplexer is the decompressed packet signal, which can now be processed at a much slower rate using electronic circuitry.

Components of an OTDM System

Optical Sources

In an ultrahigh speed OTDM system, it is essential that the optical sources are capable of generating transform-limited subnanosecond pulses having low duty cycles, tuneability, and a controllable repetition rate for synchronization. A number of suitable light sources are: gain-switch distribute feedback laser (DFB), active mode locked lasers (MLL) (capable of generating repetitive optical pulses), and harmonic mode-locking Erbium-doped fiber (EDF) lasers. Alternative techniques include super-continuum pulse generation in a dispersion shifted fiber with EDF pumping, adiabatic soliton compression using dispersion-flattened dispersion decreasing fiber, and pedestal reduction of compressed pulses using a dispersion-imbalanced nonlinear optical loop mirror. As shown in Figure 1, the laser light source is power split to form the pulse source for each channel. These are subsequently modulated with an electrical data signal (external modulation). External modulation is preferred in an OTDM system as it can achieve narrow carrier linewidth, thus reducing the timing jitter of the transmitted pulse. For ultrahigh bit rate OTDM systems, the optical pulses emerging from the external modulators may also need compressing. One option is to frequency-chirp the pulses and pass them through an anomalous dispersive medium. Using this approach, it is essential that the frequency chirp be linear throughout the duration of the pulse, in addition the midpoint of the linear-frequency chirp should coincide with the center of the pulse. When a frequency-chirped pulse passes through a dispersive medium, different parts of the pulse travel at different speeds, due to a temporal variation of frequency. If the trailing edge travels faster than the leading edge, the result would be pulse compression. An optical fiber cable or a semiconductor laser amplifier (SLA) can be used to create the frequency chirp.

Multiplexers

Multiplexing of the pulses generated by the optical sources can be implemented either passively or actively. The former method is commonly implemented using a mono-mode optical fiber. This method has the advantage of being simple and cost-effective. The latter method uses devices more complex in nature, for example, electro-optic sampling switches, semiconductor optical amplifiers, and integrated optics.

An integrated active multiplexer can be made by integrating semiconductor laser amplifiers (SOA) into a hybrid planar light-wave circuit (PLC). If the optical path length is kept short then temperature control is made easier. An alternative to this approach is to use lithium niobate (LiNb) technology. Periodically poled LiNb (PPLN) based OTDM multiplexers offer compact size and low noise (due to the absence of amplifier spontaneous emission noise and pattern effect, which is a feature of SOA based devices). Figure 7 is a schematic of a PLC-based OTDM multiplexer composed of two PLCs (R and L), composed of 1×8 and 8×1 couplers, eight 2×1 couplers, and eight different path length PPLN waveguides. The input clock pulse is split into eight by the 1×8 coupler in PLC-L, and are combined with the modulated optical pulse trains using the 2×1 couplers. The outputs of the 2×1 couplers are then passed through a PPLN waveguide to generate a return-to-zero (RZ) optically modulated signal. These are then combined, using the 8×1 coupler in PLC-R. When the path-length difference between the waveguides is set to one time slot (equal to 1/8 of the base data rate), then the output of the 8×1 coupler is the required high-bit rate OTDM pulse.

Demultiplexers

In contrast to multiplexing, demultiplexing must be performed as an active function. It can be implemented electro-optically or optically. The former method needs to complete demultiplexing of all channels in order to extract a single channel (see Figure 8). The demultiplexer in Figure 8 uses two LiNb Mach–Zehender (M–Z) modulators in tandem. The first and second modulators are driven with a sinusoidal signal of amplitudes $2V_\pi$ and V_π, respectively, to down-covert the N-bit rate to $N/2$ and $N/4$, respectively. Channels can be selected by changing either the DC-bias V + to the M–Zs, or the electrical phase delay. At ultrahigh speed implementation of an electro-optics demultiplexer becomes increasingly difficult due to the higher drive voltage requirement by the M–Z. An alternative is to use all optical demultiplexing based on the nonlinear effect in a fiber and optical active devices offering switching resolution in the order of picoseconds. There are a number of methods available to implement all optical demultiplexing. The most popular methods that use fast phase modulation of an optical signal are based on M–Z and Sagnac interferometers. Four-wave mixing is another popular method.

Mach–Zehnder (M–Z) Interferometers

The key to interferometric switching is the selection of an appropriate material for the phase modulation. Semiconductor materials, often in the form of an SLA, are suitable devices for this purpose.

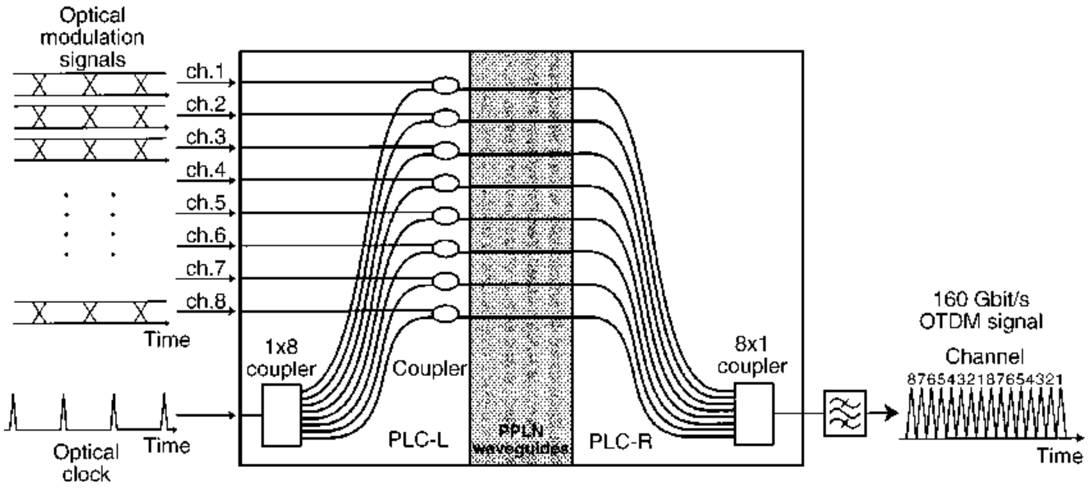

Figure 7 Configuration of a PLC-OTDM-MUX. Reproduced with permission from Ohara T, Takara H and Shake I, *et al.* (2003). 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15(2): 302–304. © 2003 IEEE.

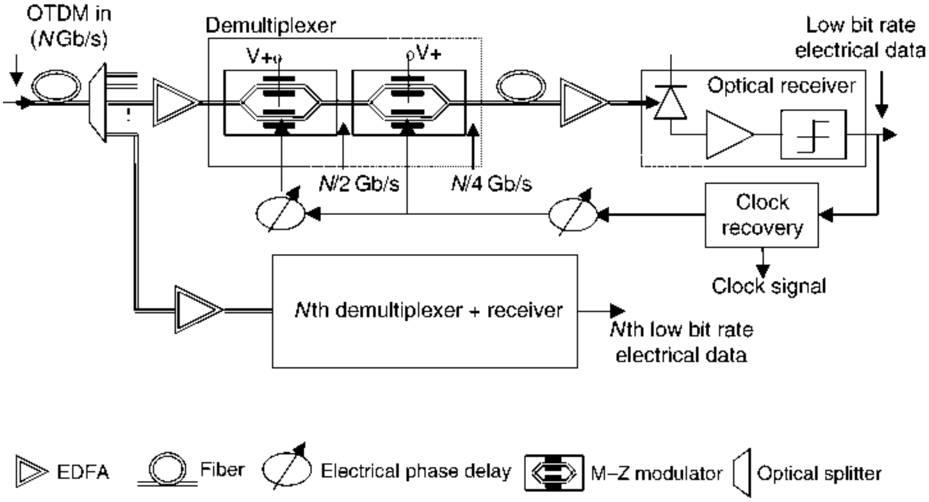

Figure 8 Electro-optics demultiplexer and receiver system block diagram.

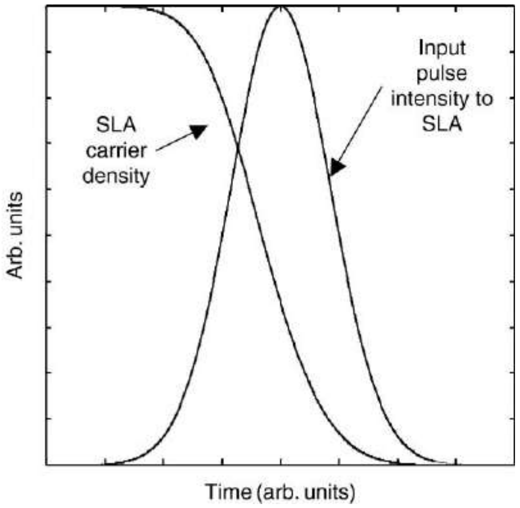

Figure 9 SLA carrier density response to input pulse.

The refractive index of an SLA is a function of the semiconductor carrier density, which can be modulated optically resulting in fast modulation. If a high-intensity pulse is input to an SLA the carrier density changes nonlinearly, as shown in Figure 9. Phase modulation is affected via the material chirp index (dN/dn), which represents the gradient of the refractive index carrier density curve for the material. The phase modulation is quite strong and, in contrast to the intensity-based nonlinearity in an optical fiber, (see Kerr effect below), is a consequence of a resonant interaction between the optical signal and the material. The nonlinearity is relatively long-lived and would be expected to limit the switching speed when using semiconductors. Semiconductors, when in excitation from an optical field, tend to experience the effect quite quickly (ps) with a slow release (hundreds of picoseconds) time. Advantage can be taken of this property by allowing the slow recovery to occur during the time between channels of an OTDM signal, as this time may be of the order of hundreds of pico-seconds or more. A Mach–Zehnder configuration using two SLAs, one in each arm of the interferometer placed asymmetrically, is shown in Figure 10. An optical control pulse entering the device via port 3 initiates the nonlinearity. The transmission equation relating the input signal (port 1) to the output port is given by

$$\frac{I_{\mathrm{out}}(t)}{I_{\mathrm{in}}(t)} = 0.25(G_{\mathrm{SLA1}}(t) + G_{\mathrm{SLA2}}(t) \pm 2\sqrt{G_{\mathrm{SLA1}}(t)G_{\mathrm{SLA2}}(t)}\cos\Delta\phi(t)) \qquad [4]$$

where $G_{\mathrm{SLA1}}(t)$ and $G_{\mathrm{SLA2}}(t)$ refer to the gain profiles of the respective SLAs, and $\Delta\phi(t)$ is the time-dependent phase difference between them. Assuming that the gain and phase profiles of an excited amplifier are given by $G(t)$ and $\phi(t)$, respectively, then the signals passing through the upper and lower arms experience optical properties of the material given by:

$$G(t),\ G(t - T_{\mathrm{d}}),\ \phi(t) \text{ and } \phi(t - T_{\mathrm{d}}) \qquad [5]$$

where T_{d} is given by $2L_{\mathrm{d}}N_{\mathrm{SLA}}/c$, L_{d} is the distance between SLA1 and SLA2, N_{SLA} is the SLA index, and c is the speed of light in a vacuum. As the width of the $\Delta\phi(t)$ profile is dependent on the distance between the SLAs, placing them in close proximity allows high-resolution switching. Less emphasis has been placed on the SLA gain as this is considered to be less effective when phase differences of π are reached. It only remains for the gain and phase

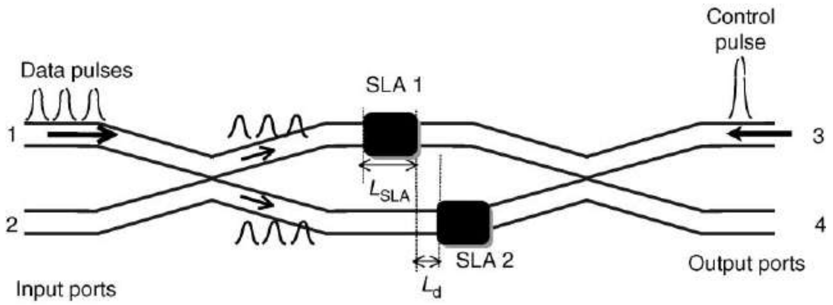

Figure 10 Asymmetric TWSLA Mach–Zehnder devices.

modulation to recover, and as indicated previously, this is allowable over a time-scale commensurate with tributary rates.

Sagnac Interferometers

There are two main types; the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD).

NOLM

In this method of switching the inherent nonlinearity of an optical fiber known as the Kerr effect is used. The phase velocity of any light beam passing through the fiber will be affected by its own intensity and the intensity of any other beams present. When the intrinsic third-order nonlinearity of silica fibers is considered via the intensity-dependent nonlinear refraction component, then the signal phase shift is

$$\Delta\phi_{\mathrm{signal}} = \frac{2\pi}{\lambda_{\mathrm{ds}}} n_2 L I_{\mathrm{ds}} + 2\frac{2\pi}{\lambda_{\mathrm{ds}}} n_2 L I_{\mathrm{c}} \qquad [6]$$

where n_2 is the Kerr coefficient, I_{ds} is the intensity of the data signal to be switched, I_{c} is the intensity of the control signal used to switch the data signal, λ_{ds} is the data signal wavelength, and L is the fiber length. The optical loop mirror consists of a long length of mono-mode fiber formed into a fiber coupler at its free ends (see Figure 11). The input to the loop comprises the high-frequency data stream plus a control pulse at the frame rate. The data split at the coupler and propagate around the loop in contra directions (clockwise E_{CW} and counter-clockwise E_{CCW}) recombining back at the coupler. In the absence of a control pulse, the pulse exits via port 1. If a particular pulse in the loop (in this example E_{CW}) is straddled by the control pulse (see Figure 12), then that pulse experiences cross-phase modulation, according to the second term on the right-hand side of eqn [6], and undergoes a phase change relative to E_{CW}. The difference in the phase between E_{CW} and E_{CCW} causes the pulse to exit via port 2. The phase shift profile experienced by the co-propagating pulse is

$$\Delta\phi(t) = \frac{4\pi}{\lambda_{\mathrm{ds}}} n_2 \int_0^L I_{\mathrm{c}}(t)\,\mathrm{d}x \qquad [7]$$

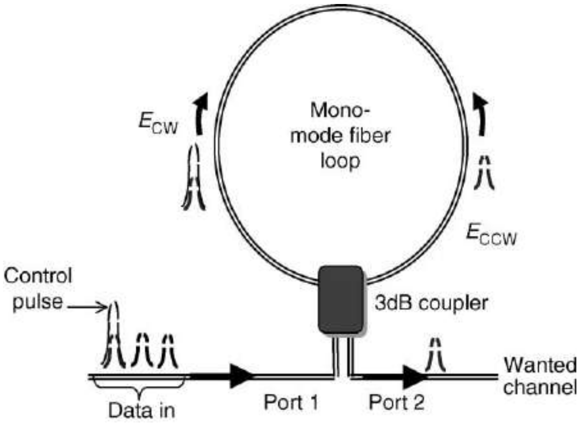

Figure 11 Nonlinear optical loop mirror.

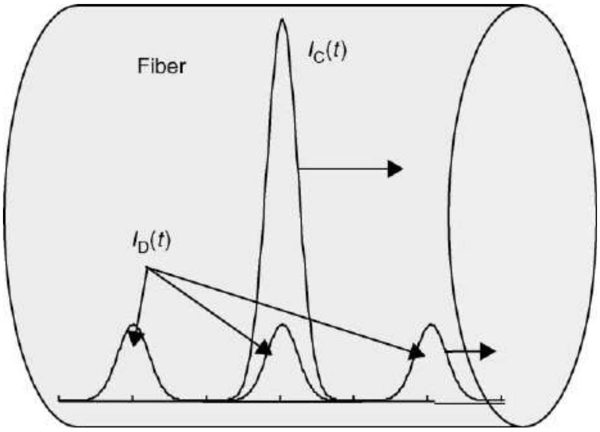

Figure 12 Control and data pulses propagation in the fiber.

Assuming unity gain around the loop, the transmittance of the NOLM is

$$
\begin{aligned}
T_x &= 1 - \cos^2\left(\frac{\Delta\phi}{2}\right) = \Delta\phi(t) \\
&= 1 - \cos^2\left(\frac{2\pi}{\lambda_{\text{ds}}} n_2 \int_0^L I_C(t)\,dx\right) \qquad [8]
\end{aligned}
$$

As it stands, the switching resolution is determined by the width of the control pulse; however, it is possible to allow the signal and control to 'walk off' each other, allowing the window width to be increased by an amount determined by the 'walk off'. The phase shift is now

$$\Delta\phi(t) = 2\frac{2\pi}{\lambda_{\text{ds}}} n_2 \int_0^L I_C(t - T_w x)dx \qquad [9]$$

where the parameter T_w, is the walk off time per unit length between control and signal pulses. The increased window width is accompanied by a lower peak transmission (see Figure 13).

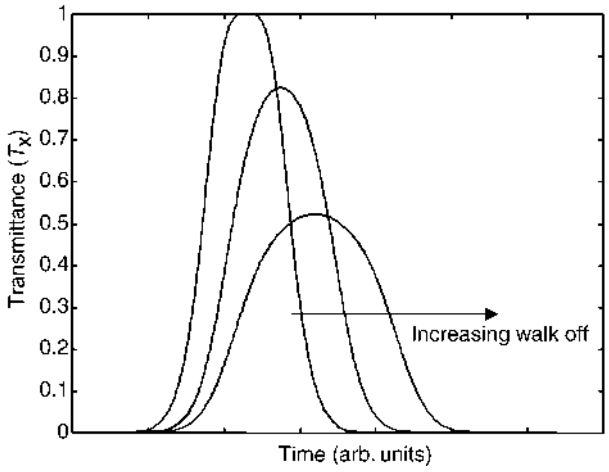

Figure 13 Transmittance profiles with the walk-off time as a parameter.

TOAD

The TOAD uses a loop mirror architecture that incorporates an SLA and only needs a short length of fiber loop (see Figure 14a). The SLA carrier density is modulated by a high-intensity control pulse, as in the M–Z-based demultiplexer. The operation of the TOAD is similar to the NOLM, where the loop transmittance is determined by the phase difference between CW and CCW traveling pulses. Strictly speaking, the gain of the SLA must also be taken into account; however, without any loss of generality the effect is adequately described by considering only the phase property. Figure 14a shows the timing diagram associated with a TOAD demultiplexer. The control pulse, shown in Figure 14b(1), is incident at the SLA at a time t_1; the phase profiles for the CCW and CCW data pulses are shown in Figures 14b(2) and (3), respectively. The resulting transmission window is shown in Figure 14b(4). The window width is given by $t_{\text{asy}} = 2\Delta x/c_f$, where Δx is the SLA offset from the loop center and c_f is the speed of light in the fiber. As in NOLM, if the phase difference is of sufficient magnitude, then data can be switched to port 2. The switch definition is, in principle, determined by how close the SLA is placed to the loop center when the asymmetry is relatively large. However, for small asymmetries the switching window is asymmetric, which is due to the CW and CCW gain profiles being different (see Figure 15). Assuming the phase modulation dominates the transmission then the normalized transmission for a small asymmetry loop is as depicted in Figure 16. The gain response (not shown) would have a similar shape and temporal position as the phase response.

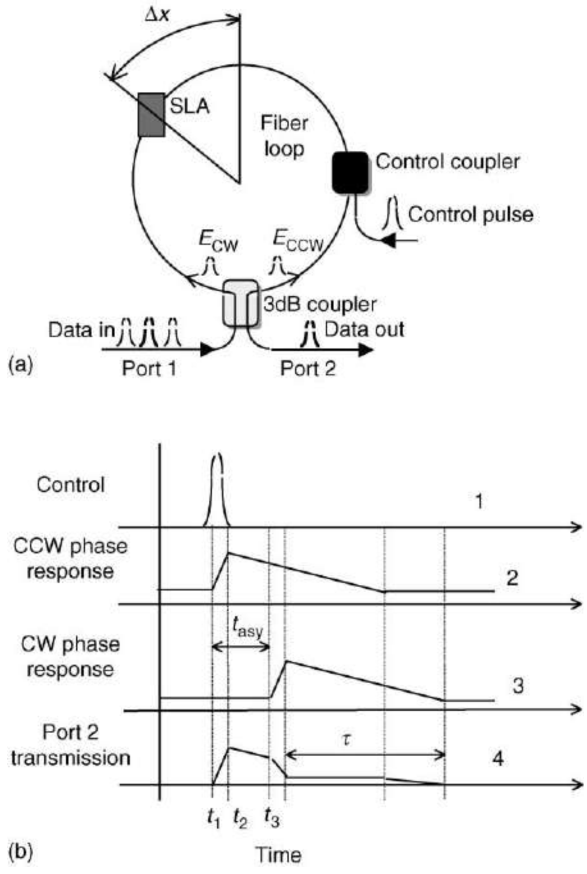

Figure 14 TOAD: (a) architecture and (b) timing diagrams.

Four-Wave Mixing (FWM)

FWM demultiplexing uses a concept whereby two optical signals of different wavelengths are mixed together in a nonlinear medium to produce harmonic components. The nonlinearity in this case arises from the third-order nonlinear susceptibility $\chi^{(3)}$ of an optical fiber, such that the polarization P induced on a pair of electric fields propagating through the fiber is

$$P = \varepsilon_0 \chi^{(3)} (E_{\mathrm{ds}} + E_{\mathrm{c}})^3 \qquad [10]$$

where ε_0 is the vacuum permittivity of the medium, E_{ds} and E_{c} are the electric fields of data and control signals, respectively. The mixing of the two signals takes place in a long length of fiber in which the control signal propagates alongside the channel to be demultiplexed (see Figure 17). The control signal is of sufficient intensity to cause the fiber to operate in the nonlinear regime. The nonlinear relationship causes a number of frequencies to be generated, with the ones of interest having a frequency given by

$$f_{\mathrm{FWM}} = 2f_{\mathrm{c}} - f_{\mathrm{ds}} \qquad [11]$$

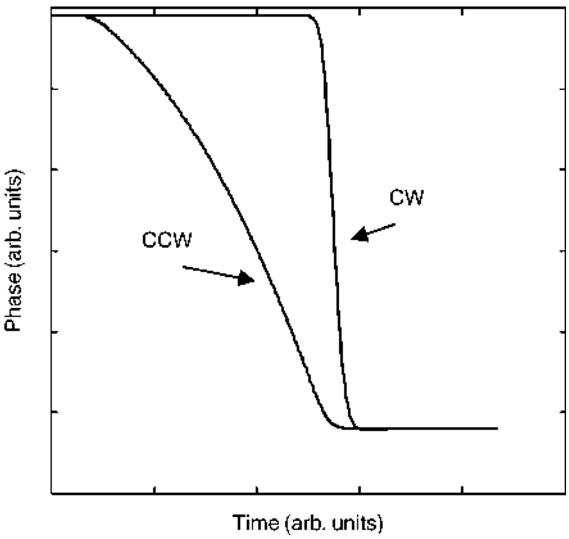

Figure 15 Phase responses for CW and CCW components of TOAD.

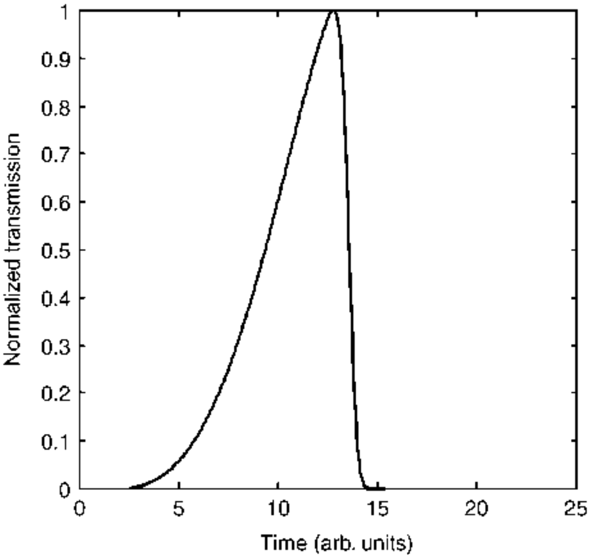

Figure 16 TOAD transmission window profile for small asymmetry loop.

An optical filter is then used to demultiplex the required channel from the composite signal. The nonlinearity is detuned from the resonant frequency of the fiber glass and as such tends to be weak, requiring long lengths of fiber to give a measurable effect. The power in the demultiplexed signal, for given data and control signal wavelengths and fiber material, depends on the input power and the fiber length according to

$$P_{\mathrm{FWM}} = kP_{\mathrm{ds}}P_{\mathrm{c}}L_{\mathrm{E}}^2 \qquad [12]$$

where k is a constant, P_{ds} is the signal launch power, P_{c} is the control signal launch power, and L_{E} is the fiber effective length. FWM is essentially an inefficient method as power is wasted in the unused frequency components of the four-wave signal. More in line with integrated structures, the nonlinear properties of a semiconductor laser amplifier can be used. Here a relatively high-power optical signal is input to the SLA (see Figure 18). This enables saturation and operation in the nonlinear regime

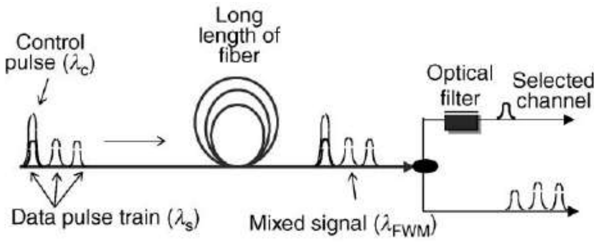
Control pulse (λ_c)
Long length of fiber
Optical filter
Selected channel
Data pulse train (λ_s)
Mixed signal (λ_{FWM})

Figure 17 Block diagram of FWM demultiplexer.

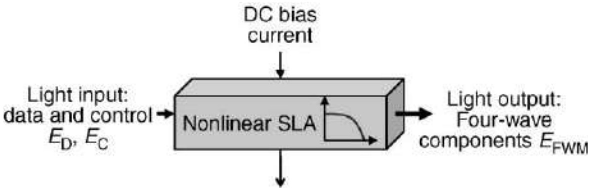
DC bias current
Light input: data and control E_D, E_C
Nonlinear SLA
Light output: Four-wave components E_{FWM}

Figure 18 Four-wave mixing in SOA.

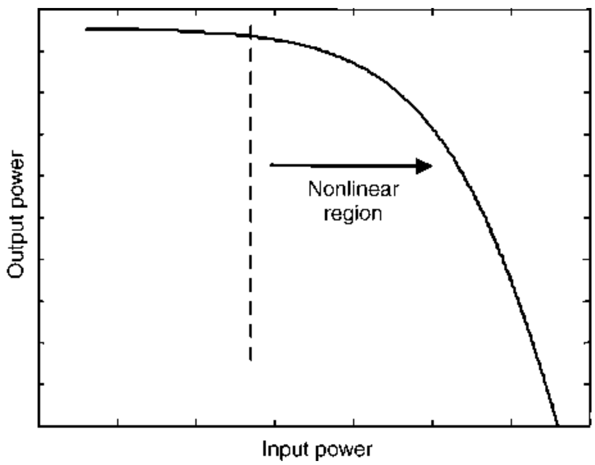

Figure 19 SLA output–input power curve.

(see Figure 19). Operating the SLA in saturation allows the nonlinear effects to produce the FWM components as in the fiber method.

Clock and Data Synchronization in OTDM

In common with electronically based TDM systems, clock recovery is fundamental to the recovery of data in ultrahigh-speed OTDM systems. Two main methods are proposed: (i) clock signal transmitted with the OTDM signal (i.e., multiplexed); and (ii) extraction of clock signal from the incoming OTDM pulse stream. In a packet-based system, synchronization between the clock and the data packet is normally achieved by sending a synch pulse with each packet. However, techniques based on optical-phased locked loops are proving popular and remove the need for a separate clocking signal.

Clock Multiplexing

(i) *Space division multiplexing*: This is conceptually the simplest to implement, where the clock signal is carried on a separate fiber from the data. However, it is susceptible to any differential delay between the different paths taken by clock and data due to temperature variation. It is difficult to justify in systems where the installed fiber base is at a premium.

(ii) *Wavelength division multiplexing*: Here different wavelengths are allocated to the clock, and payload. It is only really practical for predetermined path lengths between nodes in single-hop networks such as point–point links or broadcast-and-select star networks. It also suffers from random delays between the clock and the payload, which is problematic in an asynchronous packet switched based network, where the optical path length a packet may take is nondeterministic.

(iii) *Orthogonal polarization*: This is suitable for small links, where separate polarizations are used for the clock and data. However, in large networks it is quite difficult to maintain the polarization throughout the transmission link due to polarization mode dispersion and other nonlinear effects.

(iv) *Intensity division multiplexing*: This uses higher-intensity optical clock pulses to differentiate it from the data pulses as discussed above. However, in long-distance transmission links, it is difficult to maintain both the clock intensity and its position, due to the fiber nonlinearity.

(v) *Time division multiplexing*: In this scheme a single clock pulse, which has the same wavelength, polarization, and amplitude as the payload pulses, is separated in time, usually ahead of the data pulses.

Synchronization – Optical Phased Locked Loops (PLL)

The PLL is a common technique used for clock recovery in electronic TDM systems. However, the speed of conventional electronic PLLs tends to be limited by the response of the phase comparators used. There are a number of approaches based on opto-electronic PLL. Opto-electronic PLLs based on four-wave mixing in a traveling wave laser amplifier are complex and can suffer from frequency modulation in the recovered clock. However, others based on balanced photo-detectors, result in low timing jitter and good phase stability. In contrast, a number of all optical methods clock recovery scheme exist, one technique based on the TOAD (see above) is depicted in Figure 20. The high-speed data stream enters the TOAD at a rate $n \times R$ where n is the number of channels per OTDM frame, and R is the frame rate. A pulse generator, such as a mode locked fiber laser (MLFL) clocked by a local oscillator (LO), is used as the TOAD control input (MLFL-C). The OTDM data is switched by the TOAD at a frequency of say, $R + \Delta f$ Hz. Thus, the switching window samples data pulses at a rate higher or lower than $n \times R$ Hz and uses this signal for cross correlation in the PLL unit. The output of the phase comparator is used to regulate a voltage controlled oscillator (VCO)

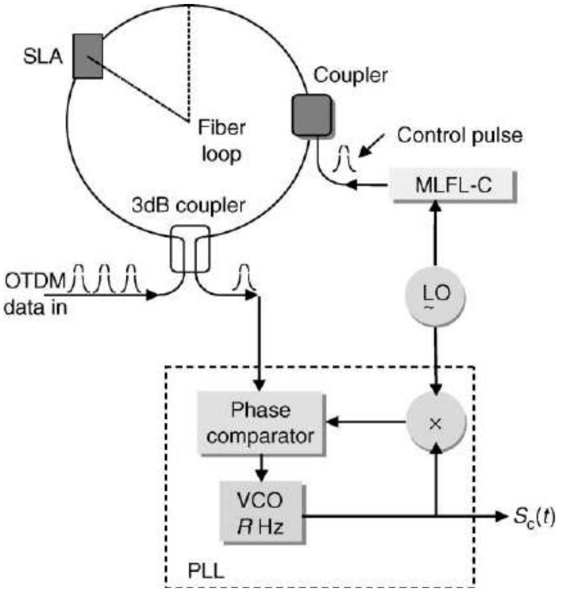
SLA
Coupler
Fiber loop
Control pulse
MLFL-C
3dB coupler
OTDM data in
LO
~
Phase comparator
×
VCO
R Hz
$S_c(t)$
PLL

Figure 20 Ultrafast clock recovery using TOAD.

running at R Hz, which in turn feeds the control signal. The PLL circuit locks into the clock frequency, generating the clock signal $S_c(t)$.

OTDM Bit-Error Rate (BER) Performance

Figure 21 shows a typical block diagram of a high-speed optical receiver for an OTDM system, composed of a NOLM or a TOAD demultiplexer, an optical pre-amplifier, an optical bandpass filter, and a conventional optical receiver using a PIN photodiode. Due to the crosstalk introduced in the demultiplexing process, the demultiplexed optical signal contains not only the target channel but may also contain nontarget channels with reduced amplitude (see the inset in Figure 21). The intensity of the demultiplexed optical signal is boosted by the optical pre-amplifier (EDFA). Amplified spontaneous emission (ASE) from the optical preamplifier adds a wide spectrum of optical fields onto the demultiplexed optical signal. Although an optical bandpass filter (BPF) can reduce the ASE, it still remains one of the major noise sources. The function of the optical filter is to reduce the excess ASE to within the range of the signal spectrum. The PIN photodiode converts the received optical power into an equivalent electrical current, which is then converted to a voltage signal by an electrical amplifier. Finally, the output voltage is sampled periodically by a decision circuit for estimating the correct state (mark/space) of each bit.

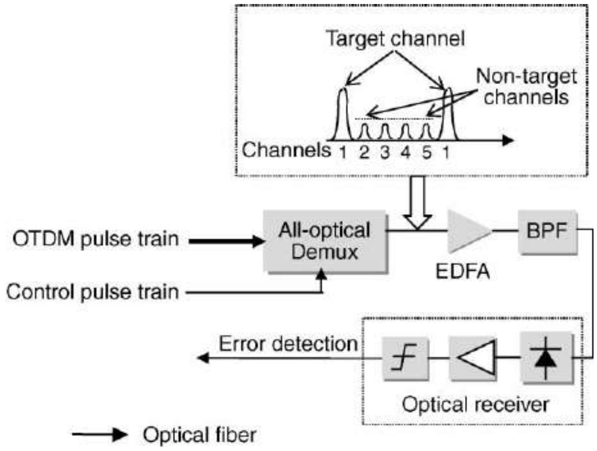
Target channel
Non-target channels
Channels 1 2 3 4 5 1
OTDM pulse train
Control pulse train
All-optical Demux
EDFA
BPF
Error detection
Optical receiver
Optical fiber

Figure 21 Block diagram of OTDM receiver.

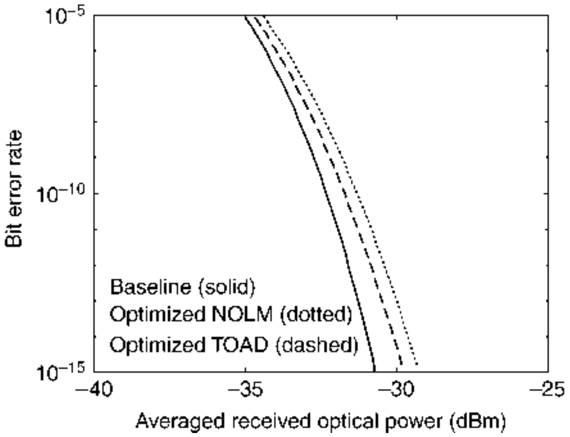

Figure 22 BER versus the average received optical power for 100 Gb/s (10 channels) OTDM system.

The BER performance of an OTDM system deteriorates because of the noise and crosstalk introduced by the demultiplexer and is:

$$\mathrm{BER} = \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-Q^2/2\right)}{Q} \qquad [13]$$

where Q is defined as:

$$Q = \frac{\bar{I}_\mathrm{m} - \bar{I}_\mathrm{s}}{\sqrt{\sigma^2_{\mathrm{RIN,m}} + \sigma^2_{\mathrm{RIN,s}} + \sigma^2_{\mathrm{amp,m}} + \sigma^2_{\mathrm{amp,s}} + \sigma^2_{\mathrm{rec,m}} + \sigma^2_{\mathrm{rec,s}}}} \qquad [14]$$

and $\bar{I}_\mathrm{m}$ and $\bar{I}_\mathrm{s}$ are the average photocurrents for a mark and a space, respectively. $\sigma_{x,i}$ are the variances of the relative intensity noise (RIN), further optical pre-amplifier and receiver for a mark and a space.

For 100 Gb/s (10 channels) OTDM system, the BER against average received optical power for optimized NOLM and TOAD demultiplexers, is shown in Figure 22.

List of Units and Nomenclature

A_i	ith data bit in a packet
c	Speed of light in a vacuum
c_f	Speed of light within the fiber
E_c	Electric fields of control signals
E_ds	Electric fields of data signals
f_FWM	Frequency due to four-wave mixing
$G_\mathrm{SLA}(t)$	Gain profiles of SLA
I_c	Intensity of the control signal
I_D	Intensity of the data signal
$\bar{I}_\mathrm{m}$	Average photo current for a mark
$\bar{I}_\mathrm{s}$	Average photo current for a space
k	Packet length in bits
L	Fiber length
L_d	Distance between two SLAs
L_E	Fiber effective length
n	Number of stages
n_2	Kerr coefficient
N_SLA	SLA index
P	Polarization
P_c	Control signal launch power
P_ds	Data signal launch power
t_asy	The window width is given by $t_\mathrm{asy} = 2\Delta x/c_f$
T	Incoming bit interval
T_w	Walk off time per unit length between control and data pulses
T_x	Transmittance of the NOLM
$\delta(\cdot)$	Optical carrier pulse shape
$\Delta\phi(t)$	Time-dependent phase difference
Δx	SLA offset from the fiber loop center
ε_0	Vacuum permittivity of the medium
λ_s	Data signal wavelength
$\sigma_\mathrm{amp,m}$	Variances for optical pre-amplifier for a mark
$\sigma_\mathrm{amp,s}$	Variances for optical pre-amplifier for a space
$\sigma_\mathrm{rec,m}$	Variances for receiver for a mark
$\sigma_\mathrm{rec,s}$	Variances for receiver for a space
$\sigma_\mathrm{RIN,m}$	Variances of RIN for a mark
$\sigma_\mathrm{RIN,s}$	Variances of RIN for a space
τ	Outgoing bit interval
$\phi(t)$	Phase profile of SLA
$\chi^{(3)}$	Third-order non-linear susceptibility of an optical fiber

See also

Interferometry: Overview. **Nonlinear Optics, Basics**: Four-Wave Mixing. **Optical Communication Systems**: Optical Time Division Multiplexing; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (2002) *Fiber-Optic Communication Systems*, 3rd edn. New York: J Wiley and Sons Inc.

Chang K (ed.) (2003) *Handbook of Optical Components and Engineering*, 2nd edn. New York: Wiley and Sons Inc.

De Marchis G and Sabella R (1999) *Optical Networks Design and Modelling*. Dordrecht: Kluwer Academic.

Hamilton SA, Robinson BS, Murphy TE, Savage SJ and Ippen EP (2002) 100 Gbit/s optical time-division multiplexed networks. *Journal of Lightwave Technology* 20(12): 2086–2100.

Jhon YM, Ki HJ and Kim SH (2003) Clock recovery from 40 gbps optical signal with optical phase-locked loop based on a terahertz optical asymmetric demultiplexer. *Optics Communications* 220: 315–319.

Kamatani O and Kawanishi S (1996) Ultrahigh-speed clock recovery with phase locked loop based on four wave mixing in a traveling-wave laser diode maplifier. *Journal of Lightwave Technology* 14: 1757–1767.

Nakamura S, Ueno Y and Tajima K (2001) Femtosecond switching with semiconductor-optical-amplifier-based symmetric-Mach–Zehnder-type all-optical switch. *Applied Physics Letters* 78(25): 3929–3931.

Ohara T, Takara H, Shake I, *et al.* (2003) 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15(2): 302–304.

Ramaswami R and Sivarajan KN (2002) *Optical Network; a Practical Perspective*. Morgan Kaufman.

Reman E (2001) Trends and evolution of optical networks and technology. *Alcatel Telecommun. Rev. 3rd Quarter* 173–176.

Sabella R and Lugli P (1999) *High Speed Optical Communications*. Dordrecht: Kluwer Academic.

Seo SW, Bergman K and Prucnal PR (1996) Transparent optical networks with time division multiplexing. *Journal of Lightwave Technology* 14(5): 1039–1051.

Sokolloff JO, Prucnal PR, Glesk I and Kane M (1993) A terahertz optical asymmetric demultiplexer (TOAD). *IEEE Photonics Technology Letters* 5(7): 787–790.

Stavdas A (ed.) (2001) *New Trends in Optical Network Design and Modelling*. Dordrecht: Kluwer Academic.

Ueno Y, Takahashi M, Nakamura S, Suzuki K, Shimizu T, Furukawa A, Tamanuki T, Mori K, Ae S, Sasaki T and Tajima K (2003) Control scheme for optimizing the interferometer phase bias in the symmetric-Mach–Zehnder all-optical switch (invited paper). Joint special issue on recent progress in optoelectronics and communications. *IEICE Trans. Electron.* E86-C(5): 731–740.

Coherent Analog Optical Processors

H H Arsenault, S Roy and D Lefebvre, University of Laval, Quebec, PQ, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Linear filtering is a useful tool commonly used in image processing. The correlation operation can be defined as:

$$f(x,y)*h(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)h(x-x',y-y')\mathrm{d}x'\mathrm{d}y' \qquad [1]$$

where $*$ denotes the correlation. The operation of correlation and the Fourier transform

$$G(\mu,\nu)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\mathrm{e}^{-j2\pi(\mu x+\nu y)}\mathrm{d}x\mathrm{d}y \qquad [2]$$

are intimately related through the correlation theorem of Fourier theory, which states that the correlation between two functions $f(x,y)$ and $h(x,y)$ is equal to the inverse Fourier transform of the product of the Fourier transforms:

$$f(x,y)*h(x,y)=FT^{-1}[F(\mu,\nu)H^{*}(\mu,\nu)] \qquad [3]$$

where $F(\mu,\nu)$ and $H(\mu,\nu)$ are the Fourier transforms of $f(x,y)$ and of $h(x,y)$ respectively, and where H^{*} designates the complex conjugate of H. The ability to carry out and display Fourier transforms in two dimensions, by means of optical systems, is at the root of optical information processing.

One of the main advantages of correlation is that the operations do not require segmentation of the regions of interest. This means that the locations of objects do not have to be segmented from the scene and that the detection is shift-invariant, that is the correlation peaks appear at locations in the scene corresponding to locations of the targets. These advantages are clearer for detection in uncontrolled conditions, for example for detecting vehicles in natural scenes. But they are also useful in security applications, such as identity verification or for detecting defects in assembly lines.

The Optical Fourier Transform

The Fourier transform of eqn [2] is widely used in the field of information processing. The inverse Fourier transform is simply defined as:

$$g(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(\mu,\nu)\mathrm{e}^{i2\pi(\mu x+\nu y)}\mathrm{d}\mu\mathrm{d}\nu \qquad [4]$$

We will now discuss how to perform Fourier transforms by means of four different optical setups.

FT Without a Lens

If we illuminate a transparency $U(x_0,y_0)$ with a coherent plane wave, the diffraction pattern we observe at a point d from the input plane is given by the following relation, known as the Fresnel approximation for diffraction (Figure 1):

$$U(x_1,y_1)=\frac{\exp\left(i\frac{2\pi d}{\lambda}\right)}{j\lambda d}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0,y_0) \times\exp\left\{i\frac{\pi}{\lambda d}\left[(x_0-x_1)^2+(y_0-y_1)^2\right]\right\}\mathrm{d}x_0\mathrm{d}y_0 \qquad [5]$$

where λ is the wavelength of the light and d is the distance between the diffracting screen and the observed pattern.

In order to use the Fresnel approximation we need only to satisfy the geometrical optics paraxial conditions. If we add an additional approximation, we can further simplify this equation to yield the Fourier transform of $U(x_0,y_0)$. The extra condition is known as the Fraunhofer or far-field approximation, which is only satisfied when the diffracted image is observed far away from the diffracting screen or when the extent of $U(x_0,y_0)$ is very small. Mathematically this condition is expressed by:

$$\frac{\pi}{\lambda d}(x_0^2+y_0^2)\ll 1 \qquad [6]$$

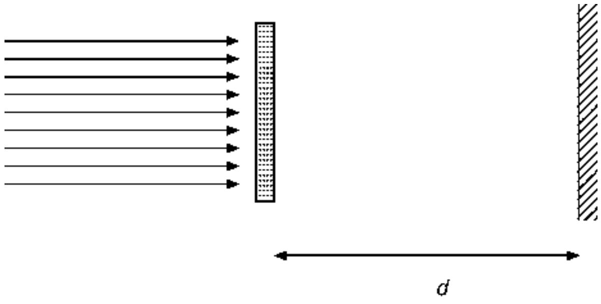

Figure 1 Fresnel diffraction without a lens.

After expanding the squares in the exponentials, there will be three terms: one can go outside the integral and one will be approximately equal to unity so the diffracted pattern becomes

$$U(x_1, y_1) = \frac{\exp\left(i\frac{2\pi d}{\lambda}\right)}{i\lambda d} \times \exp\left[i\frac{\pi}{\lambda d}(x_1^2 + y_1^2)\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda d}(x_0 x_1 + y_0 y_1)\right] \mathrm{d}x_0 \mathrm{d}y_0 \qquad [7]$$

This is the Fourier transform of $U(x_0, y_0)$, scaled by a factor λd, with an extra quadratic phase factor

$$\exp\left[\frac{i\pi}{\lambda d}(x_1^2 + y_1^2)\right]$$

The factor $\exp(i2\pi d/\lambda)/i$ is usually of no importance since it is a phase factor distributed across the output plane. In some applications we only want to know the power spectrum (squared modulus) of the Fourier transform. In such cases the quadratic phase factor disappears because the squared modulus of any complex exponential is equal to 1. Performing a Fourier transform without a lens is possible within the Fraunhofer approximation. But if it is required to obtain the Fourier transform of an image which extends about 1 cm from the optical axis using a coherent He–Ne laser ($\lambda = 632.8$ nm), the condition to be satisfied is $d \gg 496$ m: the required observation distance d is at least a few kilometers. This kind of optical setup is usually impractical, but can fortunately be alleviated by means of lenses.

Object on the Lens

Instead of having only the transparency in the input plane, we place a lens immediately after it as shown in Figure 2. For mathematical simplicity the lens and input plane are superimposed (more realistic cases are considered below). The lens acts as a phase factor

$$\exp\left[-i\frac{\pi}{\lambda f}(x_0^2 + y_0^2)\right]$$

where f is the focal length of the lens. Now in the input plane the object $U(x_0, y_0)$ is multiplied by the phase factor of the lens:

$$U'(x_0, y_0) = U(x_0, y_0)\exp\left[-i\frac{\pi}{\lambda f}(x_0^2 + y_0^2)\right] \qquad [8]$$

Inserting this expression into the integral formula for Fresnel diffraction yields the diffraction pattern observed in the focal plane of the lens:

$$U(x_1, y_1) = \frac{\exp\left(i\frac{2\pi f}{\lambda}\right)}{i\lambda f} \times \exp\left[i\frac{\pi}{\lambda f}(x_1^2 + y_1^2)\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda f}(x_0 x_1 + y_0 y_1)\right] \mathrm{d}x_0 \mathrm{d}y_0 \qquad [9]$$

The lens exactly compensates for the other phase factor that was in the integral. This expression for the Fraunhofer approximation is almost the same as before, but the focal length f of the lens replaces the distance d. So by placing a lens immediately after the input plane we can have a much more compact setup since lenses are available with a wide range of focal lengths, usually from a few centimeters to a few meters. Another way to look at this is to consider that the effect of a lens is to bring the pattern at infinity to the focal plane of the lens – except for the phase factor.

Object Before the Lens

Another setup used to perform the Fourier transform is shown in Figure 3, where the lens is placed somewhere between the input and the output planes. Using the Fresnel formula for diffraction it can be

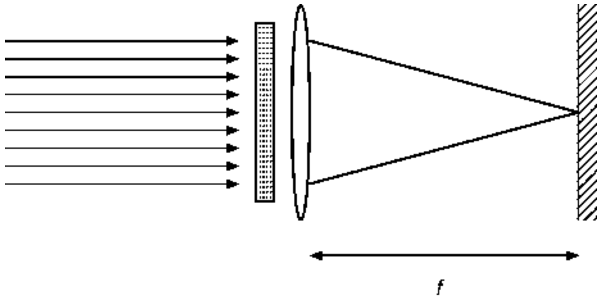

Figure 2 Diffraction with the object on the lens.

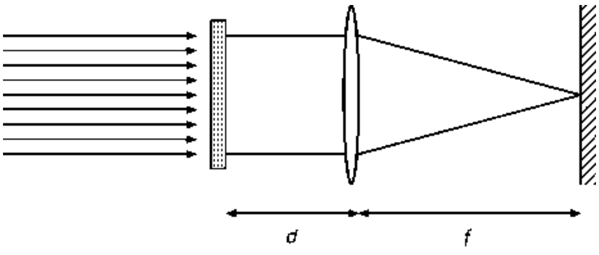

Figure 3 Diffraction with the object before the lens.

shown that the diffraction pattern in the focal plane is:

$$U(x_1, y_1) = \frac{1}{i\lambda f} \times \exp\left[i\frac{\pi}{\lambda f}\left(1 - \frac{d}{f}\right)(x_1^2 + y_1^2)\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda f}(x_0x_1 + y_0y_1)\right]\mathrm{d}x_0\mathrm{d}y_0 \qquad [10]$$

where d is the distance between the input plane and the lens. If we set d equal to the focal length f of the lens, the quadratic phase factor in front in the integral vanishes and we have an exact Fourier transform. This setup with the object and filter in the front and back focal planes of the lens respectively is the one most frequently seen in scientific articles and is known as the *f-f* Fourier transform system.

Object After the Lens

Still another configuration is the one shown in Figure 4, where the lens is placed at a distance d before the input plane. For this configuration the Fresnel diffraction yields the following pattern in the focal plane of the lens:

$$U(x_1, y_1) = \frac{\exp\left[i\frac{\pi}{\lambda d}(x_1^2 + y_1^2)\right]}{i\lambda d}\frac{f}{d}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda d}(x_0x_1 + y_0y_1)\right]\mathrm{d}x_0\mathrm{d}y_0 \qquad [11]$$

Once again the Fourier transform of $U(x_0, y_0)$ is multiplied by a quadratic phase factor. It is interesting to note that, in this case, the Fourier transform is scaled by a factor λd. This can be used to control the scale of the Fourier transform by changing the distance between the lens and the object.

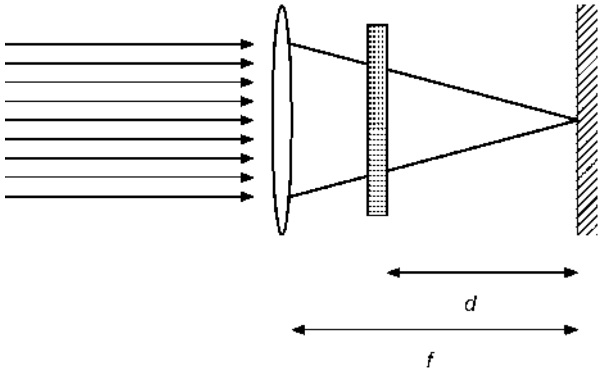

Figure 4 Diffraction with the object after the lens.

Note that although two of the three methods yield Fourier transforms that are accompanied by a quadratic phase factor, the latter is not necessarily an impediment to correlation, contrary to popular belief. Indeed all three configurations have been used for correlation pattern recognition, and each configuration has its advantages.

Optical Correlation

The last section showed how to obtain Fourier transforms by means of an optical setup. The proposed optical setups cannot perform the inverse Fourier transform. Because the Fourier transform and its inverse are very similar, the Fourier transform is performed instead, yielding the mirror image of the correlation:

$$(f * h)(-x, -y) = \mathrm{FT}[F(\mu, \nu)H^*(\mu, \nu)] \qquad [12]$$

The correlation can be considered as the result of three operations: a Fourier transform; a product; and an additional Fourier transform. The Fourier transform of a 2D function can be obtained by means of coherent optics. The equivalent of multiplication in optics is transmissivity, so multiplication is accomplished by sending light successively through two superimposed transparencies.

There are two main categories of optical correlators, serial correlators and joint transform correlators.

Serial Correlators

A serial correlator functions in two stages. The first stage performs the Fourier transform of the input image $f(x, y)$, which is multiplied in the Fourier plane with a transparency $H(\mu, \nu)$. The filter $H(\mu, \nu)$ can be a hologram, a spatial light modulator (SLM) or a simple transparency. A second Fourier transform displays in the output plane the correlation of the input image with the impulse response $h(x, y)$, which is the Fourier transform of the filter function $H(\mu, \nu)$.

Serial correlators directly implement the two Fourier transforms and the multiplication operation in a pipelined manner. The input plane and the correlation plane are conjugate planes of the system, that is in the absence of a filter, the output plane displays an inverted image of the input $f(x, y)$. Between those conjugate planes, at the conjugate plane of the source with respect to the first lens, is the Fourier plane where the spatial frequency filter is placed. When the input is illuminated with parallel light, the frequency plane is at the back focal plane of the first lens. A SLM can be placed in the input plane to feed the images to be correlated; a second SLM can be placed in the Fourier plane to display the filter. The filter is often calculated by a digital computer, and usually takes the form of a hologram if phase information must be used in the filter.

The 4-*f* correlator

The 4-*f* correlator is the most intuitive optical correlator system. We showed above that an *f*-*f* system can be used to optically obtain a Fourier transform (Figure 5). If $f(x, y)$ is in the front focal plane, its Fourier transform $F(\mu, \nu)$ appears at the back focal plane of the first lens. If a filter function $H^*(\mu, \nu)$ is inserted at this point, by means of a SLM or a hologram, then a second *f*-*f* system can perform a second Fourier transform to yield the correlation plane at the back focal plane of the second lens.

Convergent beam correlator

This kind of correlator consists of two conjugate systems. This time, the Fourier transform is obtained by placing the input object in a convergent beam after the first lens. The second lens is placed so that the output plane is conjugate to the input plane, and as before, the filter is placed at the point of convergence of the light after the first lens. It is important to note that this is not the focal plane of the lens, but the conjugate plane of the source of light. In order to have an imaging system between the input and output planes, the distances must satisfy the conjugate relation:

$$\frac{1}{s_0} + \frac{1}{s_i} = \frac{1}{f_2} \qquad [13]$$

where s_i and s_0 are defined in Figure 6. As previously discussed for convergent light, the Fourier transform will have an additional parabolic phase factor. This must be taken into account when constructing the filter *H*.

This configuration has some advantages over the 4-*f* correlator. By moving the object between the first lens and the filter, it is possible to control the scale of the Fourier transform. And because the first lens is imaging only a single point on the optical axis, only spherical aberration needs to be compensated for. In 4-*f* systems, the first lens must be corrected for all aberrations for the whole input plane, and for a given extent of the input object, the first lens must be larger than that for the convergent system in order to avoid space variance effects. Unfortunately, specially designed expensive lenses are required to correct for those aberrations, so well-corrected Fourier transform lenses for 4-*f* systems are expensive, whereas a simple doublet is sufficient to correct for on-axis spherical aberration, which is all that is required for the convergent light system. In addition, the convergent light system does not require an expensive collimator that is required in the 4-*f* system to obtain the parallel illuminating beam. Aberration correction requirements for the second lens are less stringent

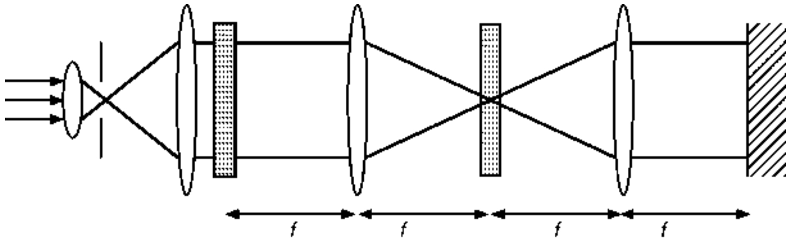

Figure 5 4-*f* correlator setup.

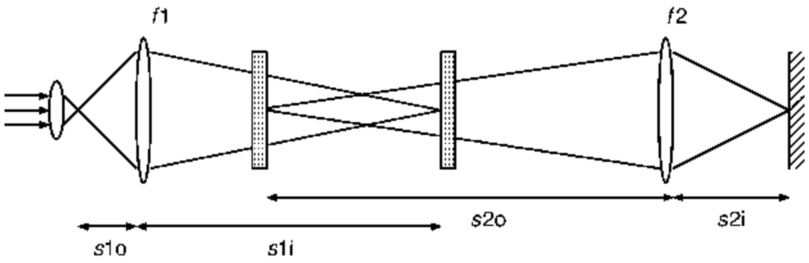

Figure 6 Convergent beam correlator setup.

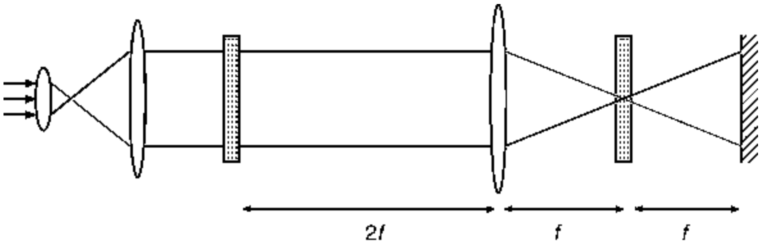

Figure 7 Single lens correlator setup.

than for the first, because in most cases, a slight deformation of the output plane can be acceptable, whereas comparable distortions in the Fourier plane can cause misregistrations between the Fourier transform and the filter that can seriously degrade the correlation.

Divergent beam correlators

Instead of placing the input object in the convergent beam, as in the previous case, another configuration that has been used places the input object in the divergent beam before the first lens. This configuration is rarely seen today, because it exacerbates the space variance and aberration problems of the 4-*f* system, and the control it allows over the scale of the Fourier transform can be achieved with fewer penalties by means of the convergent light system.

Single lens correlator

Consider an imaging system using only one lens illuminated with a parallel beam of coherent light (Figure 7). The input and output planes are located at distances $2f$ on either side of the lens – the shortest conjugate distances possible. If a filter $H(\mu,\nu)$ is placed at the focal plane of the lens on the object's Fourier transform, the correlation will appear on the output plane. This system requires satisfying impractical conditions: aberration corrections for both the Fourier transform and for the imaging operation, so it is rarely seen.

Filters

Many types of filters can be used for different applications. Table 1 lists some commonly used filters for pattern recognition. The classical matched filter yields the optimum signal to noise ratio in the presence of additive noise. The phase-only filter (POF) is frequently used because of its high discrimination capability. The inverse filter is sometimes used because it theoretically yields very sharp correlation peaks, but is very sensitive to noise, and when the spectrum $F(\mu,\nu)$ has some zeros, the realization of $1/F(\mu,\nu)$ can be impossible. A close relative of the inverse filter is the Wiener filter, which avoids the problem of the inverse filter by introducing a quantity $S(\mu,\nu)$ in the denominator of the filter function. The function $S(\mu,\nu)$ can sometimes be a constant. $H(\mu,\nu)$ is usually chosen to minimize the effect of noise or to optimize some parameter such as the peak to sidelobe ratio, or the discrimination ability of the system.

Table 1 Commonly used filters for image processing

Name	*Filter*
Matched filter	$H_{\mathrm{CMF}}(\mu,\nu) = F(\mu,\nu)$
Phase-only filter	$H_{\mathrm{POF}}(\mu,\nu) = \dfrac{F(\mu,\nu)}{\lvert F(\mu,\nu)\rvert}$
Inverse filter	$H_{\mathrm{INV}}(\mu,\nu) = \dfrac{1}{F(\mu,\nu)}$
Wiener filter	$H_{\mathrm{W}}(\mu,\nu) = \dfrac{F^*(\mu,\nu)}{\lvert F(\mu,\nu)\rvert^2 + S(\mu,\nu)}$

Joint Transform Correlators

In serial transform correlators, the filter is generally a complex function. Spatial light modulators are usually conceived with only one degree of freedom for each pixel, which can be optimized for amplitude modulation or for phase modulation. In order to achieve a fully complex filter, two modulators are required, or some spatial bandwidth of a modulator is sacrificed to encode a computer-generated hologram.

The joint-transform correlator (JTC) does not require a fully complex filter. The filter is stored in the spatial domain so there is usually no need for the computation of a filter in the spectral domain. In the JTC, the correlation of a scene with a reference image is contained in the power spectrum of an input plane consisting of the scene placed side by side with the reference. The correlation terms are revealed when the Fourier transform of the power spectrum is computed. The joint transform correlation can be summarized in three steps:

1. The scene and the reference are located side by side in the input plane.
2. The Fourier transform of the input plane is computed or carried out optically by one of the methods previously described. The transform intensity is measured and constitutes the joint power spectrum. Some optional enhancements, based on spatial filtering and/or nonlinear processing on the power spectrum, can also be performed.
3. The Fourier transform of this joint power spectrum is carried out, yielding the correlation plane containing the correlation of the reference with the scene.

A mathematical analysis is required to understand the location of the correlation peaks. The side-by-side input scene $s(x, y)$ and reference image $r(x, y)$ are

$$i(x, y) = r(x - x_\mathrm{r}, y - y_\mathrm{r}) + s(x - x_\mathrm{s}, y - y_\mathrm{s}) \qquad [14]$$

The joint power spectrum $E(\mu, \nu)$ is equal to the modulus squared of the Fourier transform of the input plane:

$$\begin{aligned} I(\mu, \nu) &= R(\mu, \nu)\mathrm{e}^{-i2\pi(\mu x_\mathrm{r} + \nu y_\mathrm{r})} + S(\mu, \nu)\mathrm{e}^{-i2\pi(\mu x_\mathrm{s} + \nu y_\mathrm{s})} \\ E(\mu, \nu) &= \|I(\mu, \nu)\|^2 = \\ &\quad \|R(\mu, \nu)\|^2 + \|S(\mu, \nu)\|^2 \\ &\quad + R(\mu, \nu)S^*(\mu, \nu)\mathrm{e}^{-i2\pi[\mu(x_\mathrm{r} - x_\mathrm{s}) + \nu(y_\mathrm{r} - y_\mathrm{s})]} \\ &\quad + R^*(\mu, \nu)S(\mu, \nu)\mathrm{e}^{-i2\pi[\mu(x_\mathrm{s} - x_\mathrm{r}) + \nu(y_\mathrm{s} - y_\mathrm{r})]} \end{aligned} \qquad [15]$$

The joint power spectrum contains the Fourier transforms of the autocorrelation of the input scene, the autocorrelation of the reference, the correlation of the scene with the reference and its complex conjugate. The correlation peaks are obtained from the Fourier transform of the joint power spectrum:

$$\begin{aligned} C(x', y') = {} & R_\mathrm{ss}(x', y') + R_\mathrm{rr}(x', y') \\ & + R_\mathrm{sr}(x' - (x_\mathrm{r} - x_\mathrm{s}),\ y' - (y_\mathrm{r} - y_\mathrm{s})) \\ & + R_\mathrm{sr}(x' + (x_\mathrm{r} - x_\mathrm{s}),\ y' + (y_\mathrm{r} - y_\mathrm{s})) \end{aligned} \qquad [16]$$

where

$$R_\mathrm{ab}(x', y') = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(\xi, \zeta)b^*(\xi - x', \zeta - y')\mathrm{d}\xi\mathrm{d}\zeta \qquad [17]$$

Multiple targets and multiple references

The input scene of the JTC may contain multiple targets. Each target in the scene will correlate with the other targets. The additional correlation peaks can clutter the correlation plane, so some care must be taken when positioning the scene with respect to the reference. Let $s(x, y)$ contain multiple targets $s_i(x, y)$:

$$s(x, y) = \sum_{i=1}^{N} s_i(x - x_i, y - y_i) \qquad [18]$$

The joint power spectrum becomes:

$$\begin{aligned} E(\mu, \nu) = {} & \|R(\mu, \nu)\|^2 + \sum_{i=1}^{N} \|S_i(\mu, \nu)\|^2 + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} \\ & \times S_i(\mu, \nu)S_k^*(\mu, \nu)\mathrm{e}^{-i2\pi[\mu(x_i - x_k) + \nu(y_i - y_k)]} + \text{C.C.} \\ & + \sum_{i=1}^{N} R(\mu, \nu)S_i^*(\mu, \nu)\mathrm{e}^{-i2\pi[\mu(x_\mathrm{r} - x_i) + \nu(y_\mathrm{r} - y_i)]} + \text{C.C.} \end{aligned} \qquad [19]$$

where C.C. denotes the complex conjugate of the preceding term. The correlation plane will now be:

$$\begin{aligned} C(x', y') = {} & R_\mathrm{rr}(x', y') + \sum_{i=1}^{N} R_{\mathrm{s}_i\mathrm{s}_i}(x', y') \\ & + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} R_{\mathrm{s}_i\mathrm{s}_k}(x' - (x_i - x_k),\ y' - (y_i - y_k)) \\ & + \sum_{i=1}^{N} R_{\mathrm{rs}_i}(x' - (x_\mathrm{r} - x_i),\ y' - (y_{\mathrm{r}i} - y_i)) \\ & + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} R_{\mathrm{s}_k\mathrm{s}_i}(x' + (x_i - x_k),\ y' + (y_i - y_k)) \\ & + \sum_{i=1}^{N} R_{\mathrm{s}_i\mathrm{r}}(x' + (x_\mathrm{r} - x_i),\ y' + (y_{\mathrm{r}i} - y_i)) \end{aligned} \qquad [20]$$

The second lines of eqns [9] and [20] are the cross-correlation terms between the targets. In order to avoid confusion between reference–target correlations and target–target correlations, special care must be taken in positioning the reference with respect to the scene. All the target–target correlations will be inside a centered box that has twice the length and twice the width of the input scene, so the reference should be placed at least at twice the length or width of the scene.

Multiple targets in the input scene can be viewed as multiple references, and the correlation of multiple references with multiple targets are all computed simultaneously. This property has led to applications in multiclass pattern recognition problems and in optical neural networks, where each neuron of a layer is coded as a reference object. This is one of the advantages of the JTC compared to serial correlators, since serial correlators are limited to one filter at a time.

Example

Here is an example of joint transform correlation. The input plane is shown in Figure 8a. The plane is composed of the input scene in the top third of the input plane and the reference is placed in the center. This input scene contains two targets. The joint power spectrum is shown in Figure 8b with its pixels mapped to a logarithmic intensity scale. The correlation plane is shown in Figure 8c and d with the central peak truncated. The central peak contains the autocorrelation terms R_{ss} and R_{rr}. On each side of the autocorrelation peak are the cross-correlations between the different targets of the input scene. The correlation between the reference image and the input scene is located in the top part of the correlation plane. The correlation plane is symmetrical with respect to the origin, so the correlation peaks are reproduced on the bottom half of the output plane.

Joint transform correlator architectures

Different architectures have been used for JTCs. Three setups are presented here. This is not an exhaustive list. The simplest optical setup shown in Figure 9 is an optoelectronic implementation of a power spectrum machine. A convergent beam

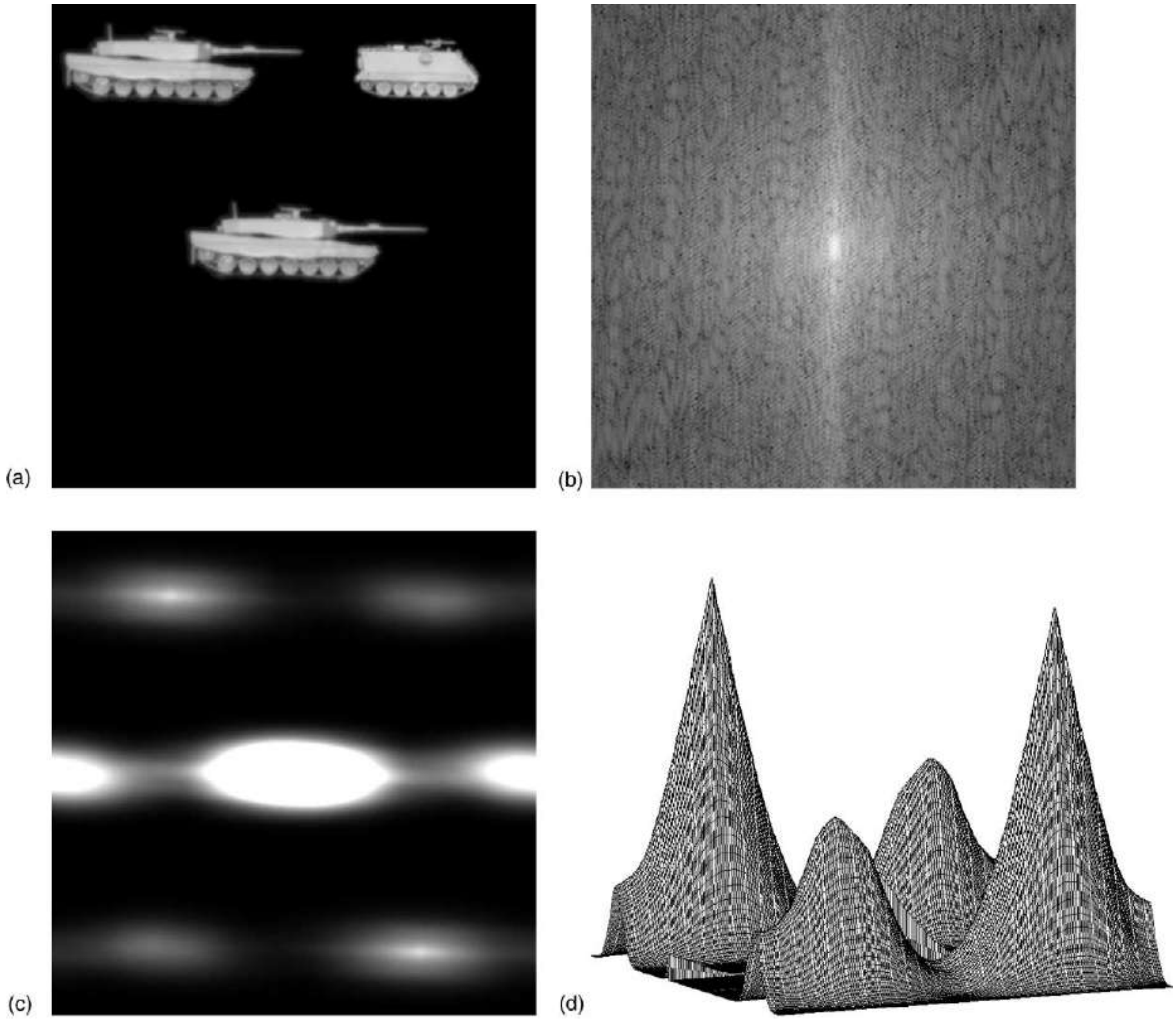

Figure 8 Joint transform correlation example. (a) Input scene; (b) joint power spectrum; (c) correlation plane (top view); (d) correlation plane (mesh view).

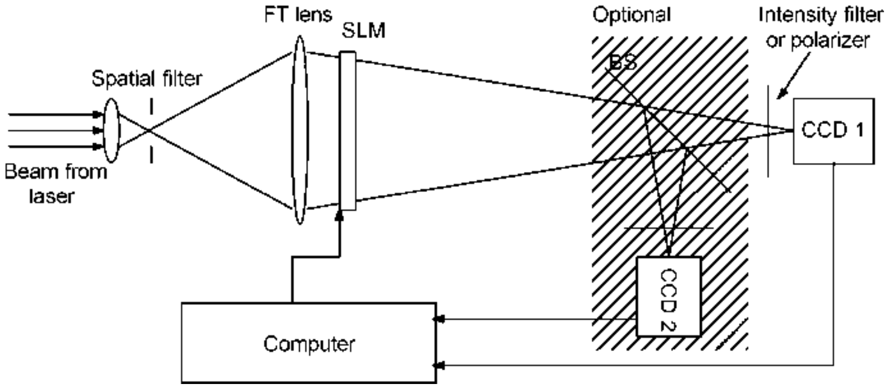

Figure 9 Joint transform correlator implementation as an optical power spectrum machine.

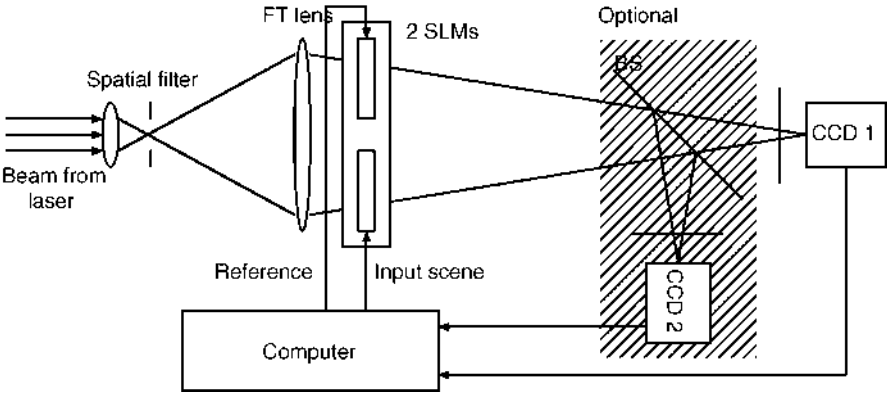

Figure 10 Optical JTC using two SLMs in the input plane.

optical Fourier transform system is used to produce the Fourier transform of an image displayed on a single spatial light modulator (SLM). A CCD camera is placed in the Fourier plane to measure the joint power spectrum. The joint transform correlation is a two-step process: the input plane containing the scene and the reference is first displayed on the SLM, and the camera simultaneously captures the joint power spectrum. In the second step, the joint power spectrum is displayed on the SLM, so the correlation plane is now on the camera. An optional second camera with intensity filters can be added to this setup. In this case, one camera serves to digitize the joint power spectrum during step 1 and the second camera captures the correlation plane at step 2 and transmits it to a display device. This optional configuration is useful to keep both the correlation plane and the joint power spectrum within the dynamic range of their specific cameras.

The setup shown in Figure 9 has two main disadvantages: the correlation is performed in two steps and requires some processing to construct the input plane, and only half of the spatial light modulator can be used to display the input scene. There are alternative setups to alleviate those drawbacks, but they require the use of two or three SLMs. Two modulators can be used in the input plane to simulate a full modulator for the scene, as shown in Figure 10. The original setup can also be doubled with one optical Fourier transform system for the joint power spectrum and one for the correlation plane (Figure 11). With three modulators, it is possible to perform the joint transform correlation in a single step with a full modulator for the input scene.

Comparison to serial transform correlators

The joint transform correlator has some advantages over the serial transform correlator. Since the phase of the Fourier transform is destroyed in the JTC, there are much fewer constraints on the design. So it is advantageous to use an optical convergent beam setup to obtain the Fourier transform. As mentioned before, there is no need for a fully complex spatial light modulator in the joint transform correlator.

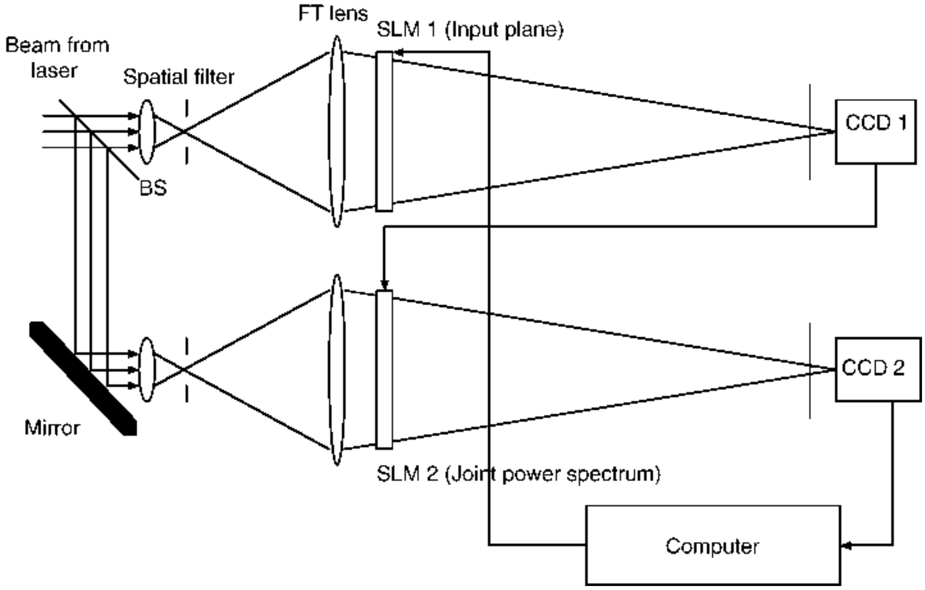

Figure 11 JTC implementation using two power spectrum machines.

This allows the use of commercially available spatial light modulators primarily designed for projectors or for displays, where phase modulation is not required. Since there is a wider market for such devices, they are cheaper.

For pattern recognition, the joint transform correlator has some properties that a serial transform correlator cannot reproduce. Some pattern recognition techniques require a linear combination of multiple amplitude correlations of the scene with different references. This is impossible with serial correlators because only the intensity of the correlation plane is detected. In a JTC, the joint power spectrum is linked to the correlation plane by the Fourier transform. So it is possible to obtain linear combinations of correlation planes by performing the corresponding operations on the joint power spectra. Some recent pattern recognition techniques such as morphological correlation and the sliced orthogonal nonlinear generalized (SONG) correlation, have been optically implemented using a JTC, but the required operations could not be carried out by means of serial optical correlators.

Whereas optical serial transform correlators always carry out linear filtering operations, optical joint transform correlators can be made to carry out nonlinear filtering. This is achieved by imposing a nonlinear characteristic curve on the joint power spectrum, instead of the usual squaring operation inherent to the detection process. The nonlinearity can be intrinsic to the components (saturation, dynamic range) or obtained by introducing electronics between the CCD camera and the second SLM, or by using an optically addressed spatial light modulator. Nonlinear processing can enhance the weights of the high frequencies in the correlation, resulting in sharper correlation peaks and improved discrimination. To illustrate this effect, the joint power spectrum of the example of Figure 8 has been modified with a logarithmic transformation:

$$E(\mu, \nu) = \ln(1 + \beta\|I(\mu, \nu)\|^2) \qquad [21]$$

with $\beta = 0.25$. The correlation plane for the transformed joint power spectrum is shown in Figure 12.

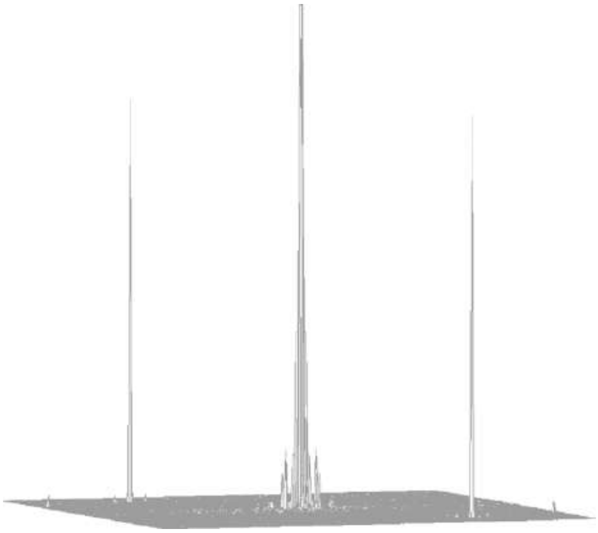

Figure 12 Correlation plane with a logarithmic transformation of the joint power spectrum.

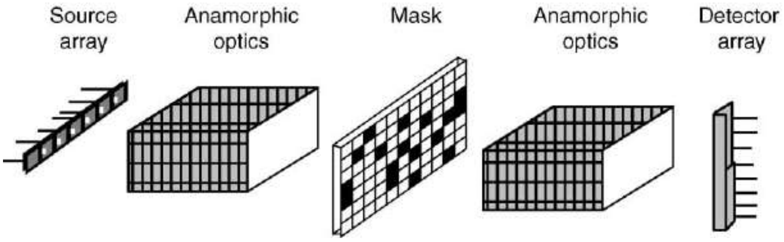

Figure 13 Optical vector–matrix multiplier.

It is clear that the discrimination has improved over that of the previous case.

Optical Vector–Matrix Multiplier

An operation that is frequently required in information processing is the product of a matrix and a vector. This can be accomplished by means on anamorphic optics as shown in Figure 13.

Here the input vector is a set of vertical fans of light from the source array, that are spread out, each over one row of the matrix mask by means of anamorphic optics. The inverse operation is carried out in the second half, so that the vertical array of spots in the output contains the vector–matrix product between the intensities of the source array and the transmissivities of the mask array.

A vector–matrix product can also be carried out by means of a conventional optical correlator as shown in Figure 14.

The input vector **a** is placed in a parallel beam at the focal plane of a conventional 2D correlator. A hologram in the Fourier plane contains a vector array **H**. The vector–matrix product **aH** appears along the axis in the correlation plane.

The above figure shows an input vector $f(x,b)\delta(y-b)$ containing the vector components $f(x1,x2,\ldots,xN)$ arrayed along the horizontal axis x, and shifted a distance a from the horizontal axis. If the Fourier transform of the mask in the filter plane is $h(x,y)$, then the output plane has the distribution:

$$g(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mu,v)\delta(\mu-b)h(x+\mu,y+v)\mathrm{d}\mu\mathrm{d}v \quad [22]$$

$$g(x,y)=\int_{-\infty}^{\infty} f(\mu,a)h(x+\mu,b-y)\mathrm{d}\mu \quad [23]$$

$$g(x,b)=\int_{-\infty}^{\infty} f(\mu,b)h(x+\mu,b)\mathrm{d}\mu \quad [24]$$

$$g(0,b)=\int_{-\infty}^{\infty} f(\mu,b)h(\mu,b)\mathrm{d}\mu \quad [25]$$

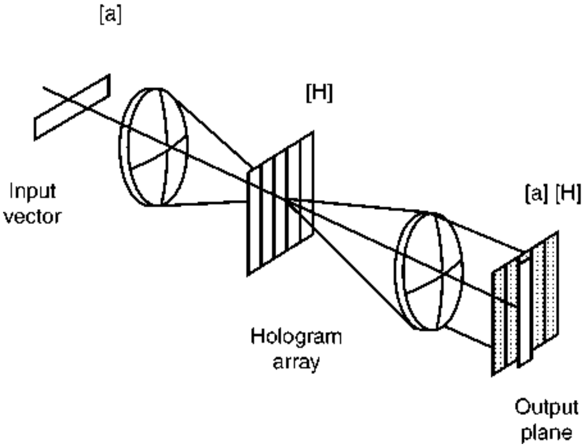

Figure 14 A vector–matrix product using a conventional 2D optical correlator.

So the output value along the y-axis at a distance a from the horizontal axis is equal to the scalar product of the vector $f(x,b)$ with the vector $h(x,b)$. Now if input vectors are stacked up one over the other in the input plane, each one will yield along the output y-axis a value that is the scalar product of the corresponding vector with the vector **h**. The set of values along the y-axis is therefore the vector–matrix product of the matrix constituted by the set of vectors in the input plane, with the vector **h** that is the impulse response of the filter. The cost for this operation is some considerable loss of light, since only the light along the y-axis is used in the output.

Applications of the optical vector–matrix product are optical crossbar, numerical matrix operations such as the solving of linear equations, optical neural networks, spectrum analysis and discrete Fourier transforms.

Conclusion

It has been estimated that serial coherent optical correlators should be able to attain data rates of up to 10^{12} operations per second, using the fastest spatial light modulators available. Joint transform correlators are somewhat slower, due to the input plane being divided between the object and the reference, to geometrical constraints in placement of targets, and due to the need to use electronics to carry out some of the operations. Until now, speeds have been limited by the data rates at the input, but as the speeds of optically addressed SLMs increase, the problem could be shifted to the output, where the lack of optical high spatial bandwidth threshold devices could cause a bottleneck.

List of Units and Nomenclature

$\mathbf{A}$	Vector or matrix
CCD	Charge coupled device
$F(\mu, \nu)$	Fourier transform of $f(x, y)l$
FT	Fourier transform
FT^{-1}	Inverse Fourier transform
H^*	Complex conjugate of H
JTC	Joint-transform correlator
POF	Phase-only filter
R_{ab}	Correlation between functions a and b
(x, y)	Spatial coordinates
λ	Wavelength of light
(μ, ν)	Spatial frequencies
$*$	Correlation
SLM	Spatial light modulator
SONG	Sliced orthogonal nonlinear generalized correlation

See also

Coherence: Coherence and Imaging. **Diffraction:** Fraunhofer Diffraction; Fresnel Diffraction. **Fourier Optics**. **Geometrical Optics:** Aberrations. **Holography, Techniques:** Computer-Generated Holograms. **Imaging:** Information Theory in Imaging. **Information Processing:** Incoherent Analog Optical Processors; Optical Neural Networks.

Further Reading

Arsenault HH, Szoplik T and Macukow B (eds) (1989) *Optical Processing and Computing*. New York: Academic Press.

Bartelt H (1987) Unconventional correlators. In: Horner JL (ed.) *Optical Signal Processing*, pp. 97–128. New York: Academic Press.

Garcia-Martinez P, Arsenault HH and Ferreira C (2001) Time sequential nonlinear correlations for optical pattern recognition. Optoelectronic Information processing: optics for information systems. *Proceedings of SPIE* CR81-12: 186–207.

Goodman JW (1988) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Javidi B (1994) Nonlinear joint transform correlators. In: Javidi B and Réfrégier P (eds) *Workshop on Optical Pattern Recognition*. Bellingham, WA: SPIE – The International Society for Optical Engineering, PM-12.

Javidi B and Horner JL (eds) (1994) *Real-Time Optical Information Processing*. New York: Academic Press.

Javidi B and Johnson KM (eds) (1996) *Optoelectronic Devices and Systems for Processing*. Bellingham, WA: SPIE – The International Society for Optical Engineering, CR65.

Lu XL, Yu FTS and Gregory D (1990) Comparison of Vander Lugt and joint transform correlators. *Applied Physics B* 51: 153–164.

McAulay AD (1991) *Optical Computer Architectures*. New York: Wiley & Sons.

Preston K (1972) *Coherent Optical Computers*. New York: McGraw-Hill.

Vander Lugt AB (1992) *Optical Signal Processing*. New York: John Wiley & Sons.

Free-Space Optical Computing

A A S Awwal, Lawrence Livermore National Laboratory, Livermore, CA, USA

M Arif, Infineon Technologies, Longmont, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Free space optical computing encompasses a vast field, from digital and discrete to analog and hybrid optical computing. This article gives an overview of various algorithms, such as shadow-casting, symbolic substitution to nonlinear optics-based processors. Some of the core devices, such as modulators and nonlinear logic elements, are discussed. The optical computing concepts, originated with nonlinear optics and analog optical computing, are more commonly known as optical signal processing. Now they have culminated into digital optical computing, opening the door to highly specialized parallel computing architecture.

Parallel computation is a direct approach to enhance the speed of digital computations which are otherwise inhibited by the sequential processing bottleneck. Physical limitations – such as limited interconnections, interactions amongst electrons, with the consequent inability to share a data bus simultaneously, and constant propagation delay of electronic circuits – place an upper bound on the speed of the fastest operation possible with an electronic computer, without a modification to the computing architecture. Optics provides an architectural advantage over electronics, due to the fact that the most natural way to use optics lends itself to a parallel implementation. Optical signals propagate in parallel, cross each other without interference (whereas charged electrons interact heavily with each other), have the lowest possible propagation delay of all signals, and can provide million channel free-space interconnections with simple lenses (whereas an unrealistic number of connections are necessary for electrons). Thus, optics is an excellent candidate for implementing parallel computing machines using the 3D space connectivity. While electronic architecture is naturally 2D, optical free space architecture is 3D, exploiting the third dimension of space. The communication bottleneck forces electronic computers to update the computation state space sequentially, whereas the use of optics will empower one to change the entire computation space in parallel.

Historical Perspective

Historically, digital optical computing was proposed as a natural extension of nonlinear optics. Just as transistors, which are nonlinear electronic devices, led to the realization of the computing logic using transistors, it was envisioned that nonlinear optical effects would give rise to optical transistors which would lead to development of optical logic gates. One of the limitations of nonlinear optical device-based computation is the high laser power requirement to demonstrate any nonlinear optical effect.

One of the effects used to demonstrate nonlinearity is to change the refractive index with the incident intensity. In other words, most materials will behave linearly with low optical energy. The refractive index expressed as $n = n_0 + n_2 I$ is nominally linear because the nonlinear co-efficient n_2 is usually very small. In the presence of high power, the $n_2 I$ is significant enough to cause an optical path change by altering the refractive index. When such a medium is placed inside a Fabry–Perot resonator, it will allow one to control the transmission of the filter by changing the incident intensity. Thus, at high power, it shows a nonlinear behavior or bistability. The term optical bistability was coined, which meant the presence of two stable states of output for the same intensity value.

Another development took place at the University of Michigan by the development of a precision optical processor; this was an optical processor which could process radar signals using optics, this is the example of an analog optical computing processor.

A third area which is in between the digital and the analog are the discrete level devices which most often take the form of matrix-based processors. Free space optical matrix vector multipliers are examples of such processors.

The latest field of optical computing research is in the algorithms and architecture area. The motivation here is to develop new algorithms which take advantage of optics. The thing that is lacking is the presence of appropriate device technology that can implement the algorithm in a given architecture. In terms of practical realization, analog optical computing is the most successful one. Adaptive optic processors have found success both in astronomy as well as vision sciences, and more recently in optical imaging through the atmosphere.

Development of free space optical computing has two main driving forces: the technology (device) base and the algorithm base. In this classification, since the field is in its infancy, both device-based classes and architectural (algorithm) classes are listed side-by-side as separate entities, ignoring the natural overlap that may be present in some instances. For example, nonlinear optical devices could be implemented into a variety of architecture that is listed below. Holographic processor is another technology-based class while some other architectures, such as, symbolic substitution may involve use of holograms, but it does not exclude it being a class by itself. The following list gives a classification of optical free space computing:

1. Digital Optical Computing:
    (a) Pattern coded computing or cellular logic:
        (i) Optical shadow-casting processor;
        (ii) Symbolic substitution;
        (iii) Theta logic based computation.
    (b) Parallel arithmetic circuits:
    (c) Hybrid optical computing:
    (d) Nonlinear optical device based:
    (e) Threshold logic.
2. Discrete optical computing:
    (a) Optical fuzzy logic:
    (b) Optical matrix processors:
        (i) Vector matrix multiplier;
        (ii) Differential equation solver;

    - (iii) Linear algebra processor;
    - (iv) DFT processor;
    - (v) Programmable logic array;
    - (vi) Associative memory processor.
  - (c) Optical neural network.
3. Analog optical computing:
   - (a) Adaptive optics processor;
   - (b) Partially coherent processing (white light processing):
   - (c) Fourier transform based processor:
   - (d) Nonlinear contact screen based:
   - (e) Holographic processors:
   - (f) Linear optical processors (image differentiation).

We start by discussing some of the device based classes and their underlying concepts: (a) nonlinear optics; (b) SLMs; and (c) hologram based.

Core Device Technologies Used for Free-Space Optical Computing

One of the most difficult challenges of optical computing (OC) is inventing devices that can realize aforementioned key advantages of optics. There exist, however, several key technology areas, where a breakthrough in any one of these will pave the way to real-world optical computing. We discuss here some of the major device technologies that relate to OC.

Spatial light modulator (SLM) converts information from different forms to the optical domain. As a result, SLM is used as an input device to an optical computing system. High contrast ratio, very high speed, and parallel addressing scheme are desirable properties of an SLM as an input device. For example, the SLM could be used to input a 2-dimensional signal or an image into a Fourier transform-based processor. Several applications of SLM are possible:

1. *Optical encoding*: Pattern coded logic can perform logic operations if inputs can be encoded using certain pattern;
2. *Encoding complex filter*: It could encode phase, amplitude or both; and
3. *As phase modulator*: for adaptive optics.

Two types of SLMs are liquid-crystal based and micro-electromechanical (MEM) based. The former uses liquid crystals to change the phase of transmitted or reflected light leading to modulation of phase. Although the most successful commercial application of MEMs is its use in consumer electronics, such as projection TVs or video projectors, in free-space optical computing it can be used as:

- (a) an interconnection device;
- (b) an SLM for matrix; or
- (c) a phase modulating device in AO close loop control of optical phase aberrations or phase only filter.

The heart of the MEMs is tiny electrically controllable mirrors which can steer the incoming beam in a different direction or bend to create a relative optical path difference for the reflected beam.

Algorithms, where an SLM plays a major role, are pattern coded computing, optical matrix processors, adaptive optic processors, partially coherent processing, Fourier transform based processors, and linear optical processors.

Nonlinear Optical Element

A nonlinear optical element is an architecture or device structure which allows for utilization of the nonlinearity of the optical material that produces detectable changes of optical intensity. Figure 1 shows the basic block diagram for an optical computing basic element. To achieve an optical logic gate, one needs a scheme where light can be controlled by light, as current is controlled by current in a transistor. An optical device showing optical bistability can be used for this purpose. Optical bistability is defined as the situation where two different output intensities are possible for a given input intensity.

For example, as shown in Figure 2 below, at the intensity level of C, there are two stable states, one HIGH and one LOW. Just as the transistor has two stable states ON and OFF, the optical bistable device exhibits two states ON and OFF. The fact that such nonlinear behavior could implement logic behavior

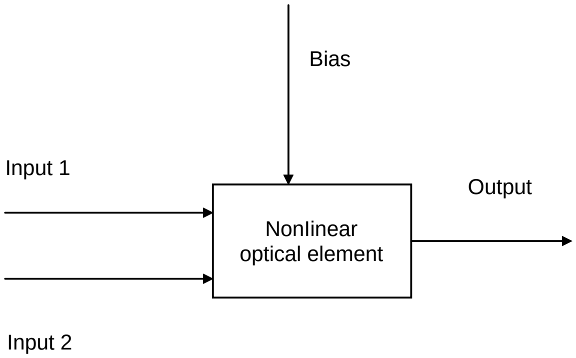

Figure 1 Block diagram of an optical logic gate using nonlinear optical element.

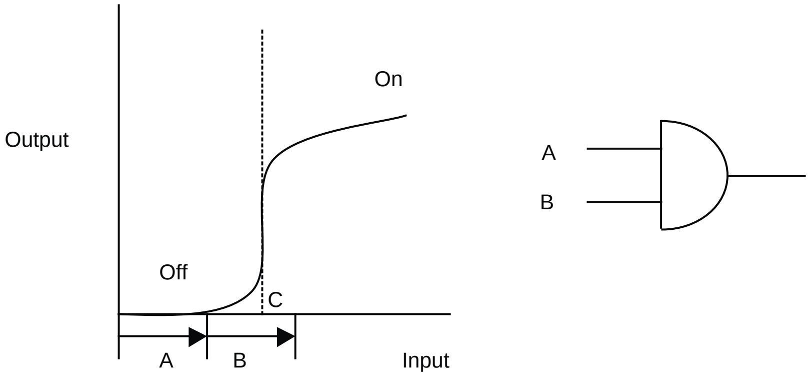

Figure 2 Bistable behavior, a typical AND behavior with inputs A and B.

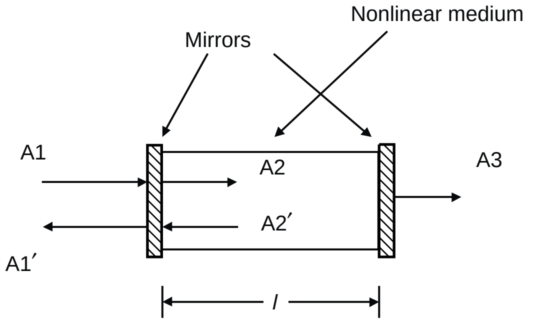

Figure 3 Bistable optical device, nonlinear medium in a Fabry–Perot cavity.

can be explained as follows. Assume that a two input device possesses a nonlinear characteristic as shown in Figure 2. Assume that the input intensity value of 1 is indicated by an intensity level equal to A. When both A and B are equal to 1, then the combined intensity switches the device to HIGH output state. However, if only one input is present, the output will be LOW. Thus, this bistable device acts as a logical AND gate.

To achieve optical bistability, two features are required, namely, nonlinearity and feedback. One way to achieve this is by placing a nonlinear medium in a Fabry–Perot cavity. Figure 3 shows a schematic of these devices. In Figure 3, A_1 is the incident field, A_1' is the reflected field, A_2 and A_2' are the forward and backward fields inside the cavity, respectively, and l is the cavity length.

The coupled wave equation for this system is given by

$$A_2' = \rho A_2 e^{2ikl-\alpha l} \qquad [1a]$$

$$A_2 = \tau A_1 + \rho A_2' \qquad [1b]$$

where, ρ is the amplitude reflectivity and τ is the amplitude transmissivity. The wave vector $k = n\omega/c$ and α is the absorption coefficient. In deriving the above equations, it was assumed that k and α are spatially invariant and real. Solving for A_2, from eqns [1a] and [1b] yields

$$A_2 = \frac{\tau A_1}{1 - \rho^2 e^{2ikl-\alpha l}} \qquad [2]$$

Equation [2] is the so-called Airy's equation which describes the properties of the Fabry–Perot cavity. From eqn [2] it can be seen that if k or α is a sufficiently nonlinear function, this device will show bistability. In general, both k and α can show nonlinearity. Hence optical bistability can be either absorptive or dispersive.

Absorptive Bistability

If the absorption coefficient α depends nonlinearly on the intensity, assuming k is constant, after some manipulation, eqn [2] can be written as

$$I_2 = \frac{TI_1}{[1 - R(1 - \alpha l)]^2} \qquad [3]$$

where, $I_i \propto |A_i|^2$, $T = |\tau|^2$ and $R = |\rho|^2$. In deriving eqn [3], it was assumed that the mirror separation is such that the system is in resonance and $\alpha l \ll 1$. If α depends on the intensity, it can be assumed that it follows the relationship of two level saturable absorbers, for simplicity, which is

$$\alpha = \frac{\alpha_0}{1 + I/I_s} \qquad [4]$$

where, α_0 denotes the unsaturated absorption coefficient, I is the local value of intensity, and I_s is the saturation intensity. Using eqn [4] in eqn [3] yields

$$I_1 = TI_2\left(1 + \frac{C_0}{1 + 2I_2/I_s}\right)^2 \qquad [5]$$

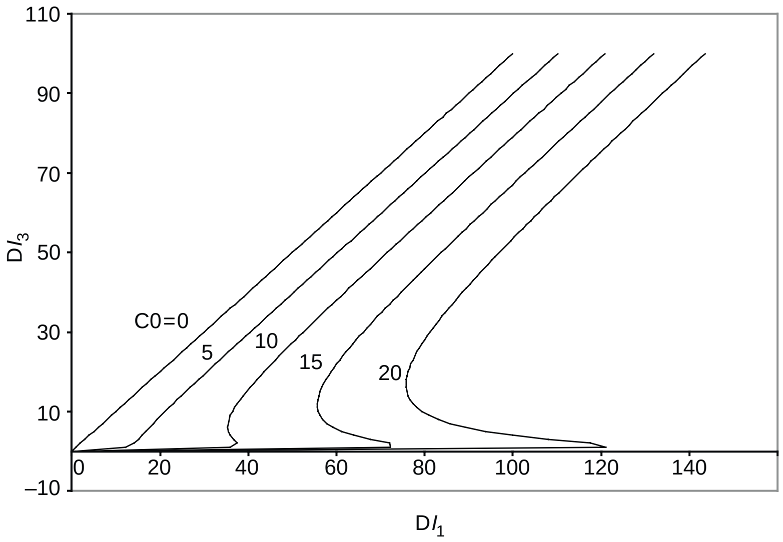

Figure 4 The input–output relationship of absorptive bistability predicted by eqns [5] and [6].

where, $C_0 = R\alpha_0 l/(1 - R)$. Also the output intensity I_3 is related to I_2 by

$$I_3 = TI_2 \qquad [6]$$

Using eqns [5] and [6], it can be shown that there can be multiple solutions to the above input–output relationship which predicts the optical bistability. Figure 4 shows a plot of DI_1 versus DI_3, where $D = 2/TI_s$, as a function of various C_0. As can be seen from Figure 4, for a high enough value of C_0, the system possesses multiple solutions.

Dispersive Bistability

The intensity dependence of the refractive index can be written as

$$n = n_0 + n_2 I \qquad [7]$$

If the refractive index n varies nonlinearly with intensity, assuming $\alpha = 0$ and using eqn [7], eqn [2] can be written as

$$A_2 = \frac{\tau A_1}{1 - \rho^2 e^{2ikl}} = \frac{\tau A_1}{1 - Re^{i\delta}} \qquad [8]$$

where, $\delta = \delta_0 + \delta_2$. δ_0 and δ_2 are given by

$$\begin{aligned} \delta_0 &= \varphi + 2n_0 \frac{\omega c}{l} \\ \delta_2 &= 2n_2 I \frac{\omega c}{l} \end{aligned} \qquad [9]$$

where, φ is the phase associated with ρ and the fact that $k = n\omega/c$ is used.

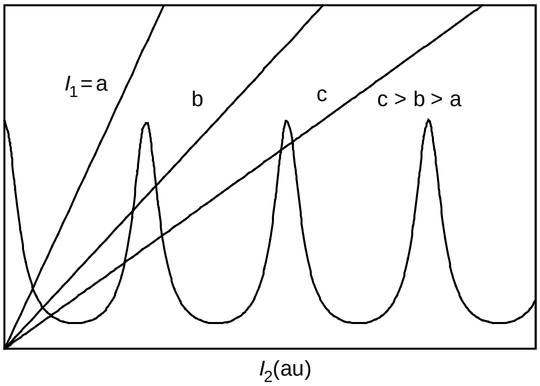

Figure 5 Plots of both sides of eqn [10] as a function of I_2 and for increasing values of input intensity I_1. The oscillatory curve represents the right-hand side of eqn [10] and the straight lines represent the left-hand side with increasing values.

After some manipulation, eqn [8] can be rewritten in terms of intensity as

$$\frac{I_2}{I_1} = \frac{1/T}{1 + (4R/T^2)\sin^2(\delta/2)} \qquad [10]$$

and

$$\delta = \delta_0 + \left(4n_2 \frac{\omega}{c} l\right) I_2 \qquad [11]$$

Equation [10] can be solved graphically to show the introduction of bistable behavior as a function of input intensity. Figure 5 shows the plots of both sides of eqn [10] as a function of I_2 and for increasing values of input intensity I_1. The oscillatory curve represents the right-hand side of eqn [10] and the straight lines represent the left-hand side with increasing values of I_1. It can be readily seen from Figure 5 that at sufficiently high values of I_1, multiple solutions are possible for eqn [10]. This again will give rise to the bistable behavior.

Nonlinear Interference Filters

The optical bistability from the nonlinear interference filters (or etalons) can be used to achieve optical logic gates. The transmission property of these etalons can be modified by another optical beam, known as a bias beam. A typical logic gate and the transmission property are shown in the following figure. If the bias beam is such that, $P_{\text{switch}} - (P_a + P_{\text{bias}}) < P_b$, then the transmitted beam shown in Figure 6b readily follows an AND gate. The truth table for such configurations is shown in Figure 6d.

Four-Wave Mixing

Four Wave Mixing (FWM) can be used to implement various computing functions. Using the photorefractive effect and FWM in a third-order nonlinear optical material, optical phase conjugate beams can be obtained. A schematic of FWM and phase conjugation is shown in Figure 7. The forward pump beam interferes with the probe beam to produce a grating in the photorefractive material. The backward pump beam is then diffracted to produce a beam which retraces the path of the probe beam but has an opposite phase. It can be seen from Figure 7, that the optical phase conjugation can be readily used as an AND logical operator. By using the four-wave mixing, an optical counterpart of transistors has been introduced. FWM have also been used in AND-based optical symbolic substitution (OSS) operations. FWM is also used to produce matched filter-based optical correlators.

Another device showing the optical bistability is the self electro-optic effect device (SEED). SEEDs were used to implement optical logic circuits. The SEEDs are produced by putting multiple quantum wells (MQW) in the intrinsic region of a p-i-n structure. The MQWs are created by placing alternating thin layers of high (barriers) and low (wells) bandgap materials. An electric field is applied to the SEEDs externally. The absorption coefficient is a function of applied bias voltage. When a light beam is incident on these devices, it creates a photocurrent, consequently lowering the bias voltage across the MQW. This, in turn, increases the absorption and when the input optical power is high enough, the system sees a peak absorption and suddenly switches to a lower state and thus shows the bistable behavior. A schematic of SEED is shown in Figure 8.

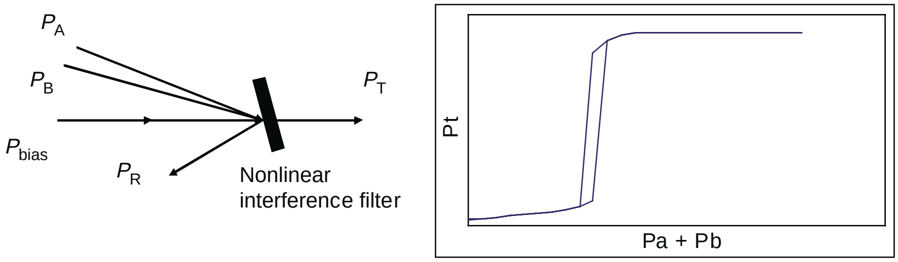

(a) (b)

Truth table for an AND function from the above setup from the transmitted beam

P_A= input 1	P_B = input 2	P_T = output
0	0	0
0	1	0
1	0	0
1	1	1

(d)

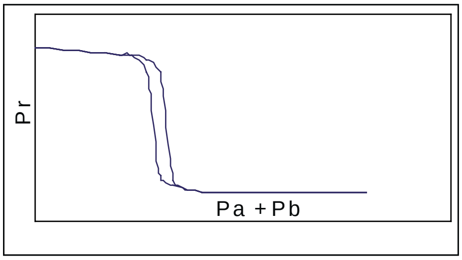

(c)

Figure 6 A typical optical logic gate configuration using nonlinear interference filter: (a) schematic; (b) the transmitted power output; (c) reflected power output (both as functions of inputs); and (d) the truth table for AND operation with beams A and B as inputs.

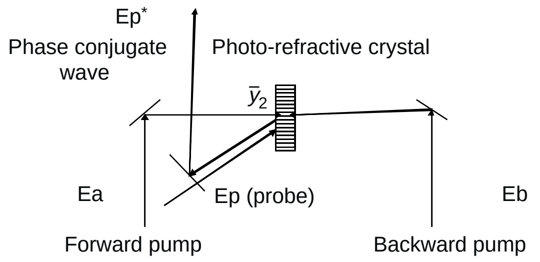

Figure 7 Four-wave mixing setup.

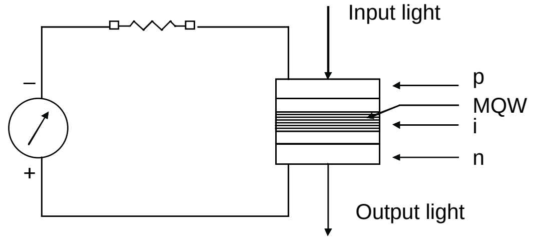

Figure 8 Schematic of self electro-optic effect devices.

Replacing the resistive load by a photodiode, a diode-based SEED (D-SEED) can be constructed, which is analogous to the way diode transistor logic was developed from resistor transistor logic in electronics. Electrically connecting two MQW p-i-n structure symmetric SEEDs (S-SEED) are constructed, which are less sensitive to optical power fluctuations and effectively provide isolation between input and output.

Pattern Coded Optical Computing: Symbolic Substitution or Cellular Logic

This is a keyword which is equivalent to simple digital logic operation. In digital logic, a truth table can describe the relationship between input and output variable. In symbolic substitution logic, this relationship is expressed as a 2D pattern rule. In any specific rule (equivalent to a row in the truth table) of SS logic, a specific 2D pattern is identified and replaced by another pattern. For example, in the case of binary logic, we can express the addition of two binary bits with the following truth table:

Rule number	*Input*		*Output*	
	A	*B*	*C*	*S*
1	0	0	0	0
2	0	1	0	1
3	1	0	0	1
4	1	1	1	0

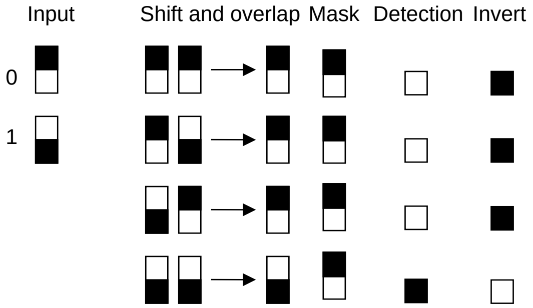

Figure 9 Symbolic substitution example, the two inputs are coded using dual rails. The second column shows the four combinations of inputs (00, 01, 10, and 11). They are shifted and overlapped creating the fourth column. A mask detects the output as a dark pixel appearing through the transparent opening. The dark is inverted to create a detection output.

Here A and B are the two digits to be added, C and S are the sum and the carry bit, respectively. To convert this operation to an SS operation, we can say that a two bit binary pattern is identified and replaced by a sum bit and a carry bit. If we now express the above in terms of symbolic substitution, we will say let 0 and 1 be represented by the following 2D symbol. Then the rule for the carry bit generation is shown in Figure 9.

Note that each of these rules has to be implemented individually. However, one can implement these by only generating the 1 output or the zero output, whichever leads to the minimum rule. Here implementing the last rule is convenient, one can make a copy of the whole pattern and then overlap it with its left shift version. This will result in a dark pixel at the origin. This is inverted and then produces a detection output. Natural extension of SS is in signed digit computing, where carry free addition can be achieved, allowing parallel implementation of numerical algorithm using free space optics.

Optical Shadow Casting (OSC)

All digital logic is nonlinear. The nonlinearity of the digital logic is converted into nonlinearity of coding. The OSC system originally proposed by Tanida and Ichioka uses two-dimensional spatially encoded patterns as data input and light emitting diodes (LEDs) as light sources. In the original shadow-casting system, the inputs were represented by vertical and horizontal stripes of opaque and transparent bars. With this system, programmability was attained by changing the LED pattern and it was possible to realize 16 logical operations between two

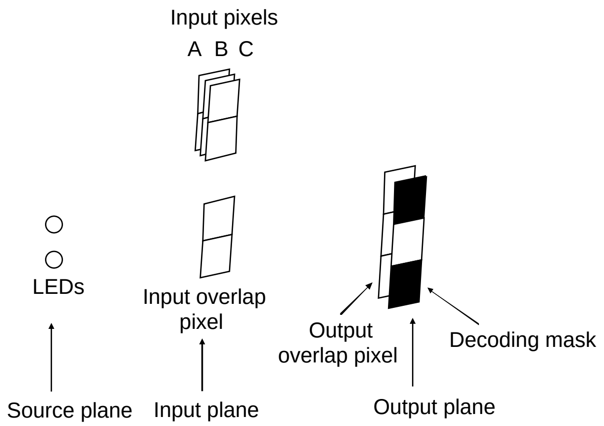

Figure 10 An optical shadow-casting system.

variables as well as a half addition as an example of arithmetic operation. Later, polarized codes were introduced into shadow-casting systems. This improved technique, called the polarization-encoded optical shadow casting (POSC), enabled complex combinational logic units like binary and trinary full adders, as well as image processing applications, to be realized.

The lensless OSC system, as shown in Figure 10, consists of spatially encoded 2D binary pixel patterns as the inputs. The input patterns are kept in close contact at the input plane and the resulting input overlapped pixel pattern is illuminated by a set of LEDs from the source plane. The light originating from each of the LEDs located at the source plane produces a shadow of the input overlap pixel pattern at the output plane. The overlap of these relatively displaced shadows results in an output overlap pixel pattern at the output plane. A decoding mask placed at the output plane is then used to spatially filter and detect the logical output. To prevent crosstalk, specific spacing among the different processing elements must be maintained.

Discrete Processors: Optical Matrix Processor

Most of the discrete optical processing schemes revolve around optical matrix–vector multiplication. Many numerically useful operations such as: (a) linear algebraic operations; (b) signal processing algorithms; (c) Boolean logic; (d) digital multiplication by analog convolution; and (e) neural networks, can be formulated as matrix–vector multiplication. An example of the digital multiplication by analog convolution (DMAC) can be understood by the following numerical example:

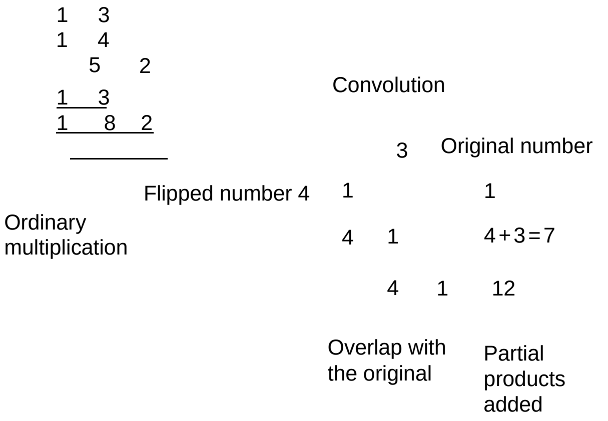

$$1\ 7\ 12 = 1\times100 + 7\times10 + 12\times1 = 182$$

The two numbers are convolved as shown above. This is done by reversing one of the numbers (4 1), while keeping the other number in the same order (1 3), then sliding the reversed number to the right, and collecting the element by element products. The second partial products are 4 × 1 and 1 × 3 which are then added to give the convolution results. Surprisingly, the weighted number in mixed radix (since numbers that have digits higher than 9, are no longer decimal) has the same value as the product shown in the upper left. It is possible to have fast parallel convolution optically, resulting in fast parallel multiplier. However, an efficient electronic post processor is necessary to convert the mixed radix numbers to the radix of the to-be-multiplied numbers.

A basic optical matrix–vector multiplier is shown in Figure 11. Here the vector, represented by an expanding column of light, is projected onto the 2D matrix, possibly represented by a glass plate or electronically addressable spatial light modulator. As the light representing the vector passes through the transmitting medium, each element of the vector (**I**) gets multiplied by a column of matrix represented by the transmittance of the matrix (T). The product (**I**T)

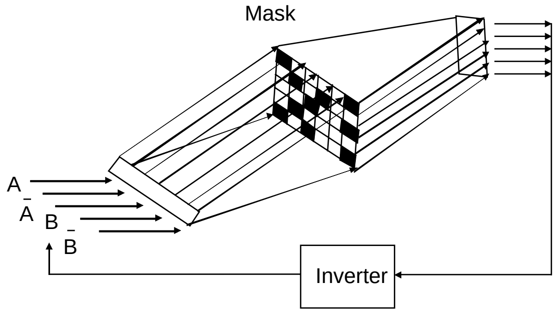

Figure 11 A matrix–vector multiplier.

is integrated horizontally using a cylindrical lens. This leads to an addition operation of the partial products, resulting in the generation of the output product vector. Mathematically:

$$y = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} a_{11}b_1 + a_{12}b_2 + a_{13}b_3 + a_{14}b_4 \\ a_{21}b_1 + a_{22}b_2 + a_{23}b_3 + a_{24}b_4 \\ a_{31}b_1 + a_{32}b_2 + a_{33}b_3 + a_{34}b_4 \\ a_{41}b_1 + a_{42}b_2 + a_{43}b_3 + a_{44}b_4 \end{bmatrix} \qquad [12]$$

The above can be achieved if the vector b is expanded as

$$\begin{bmatrix} b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \end{bmatrix}$$

and a point-by-point multiplication is performed between the expanded vector and the original matrix. Then the addition is performed by the cylindrical lenses.

Another simple application of vector matrix multiplier is in digital logic, known as programmable logic array (PLA). PLA is an electronic device that consists of a set of AND gates followed by a set of OR gates, which can be organized in AND–OR format suitable for realizing any arbitrary general purpose logic operation. AND logic is performed when light representing a logic level is passed through a transparency representing the other variable; OR is achieved by converging light to a common detector.

$$\text{Assume a function } F = AC + \bar{A}BD + A\bar{B}C \qquad [13]$$

where $\bar{A}$ represents the logical inverse of A. Using DeMorgans' theorem in digital logic, this can be expressed as

$$\begin{aligned} F &= \overline{(\bar{A} + \bar{C})} + \overline{(A + \bar{B} + \bar{C})} + \overline{(\bar{A} + B + \bar{C})} \\ &= \bar{y}_1 + \bar{y}_2 + \bar{y}_2 \end{aligned} \qquad [14]$$

Now the y_is can be generated by the vector matrix product, as shown below:

$$\text{where} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} A \\ \bar{A} \\ B \\ \bar{B} \\ C \\ \bar{C} \\ D \\ \bar{D} \end{pmatrix} \qquad [15]$$

Note here that the AND has been converted to OR, which needs to be inverted and then ORed to generate the function F. A setup shown in Figure 11 can be used to accomplish this. In step 1, the y_is are generated, and then in step 2 they are inverted and summed by a second pass through the same system with a different mask. Other applications of vector–matrix multipliers are in neural networks, where the interconnection weights are represented by the matrix and the inputs are the vector.

Analog Optical Computing

Correlation is a way to detect the presence of signal in additive noise. Optics offer a fast way of performing correlations. The practical motivation of this area comes from the fact that a lens performs a Fourier transform of a two-dimensional signal at its focal plane. The most famous analog optical computing is known as optical matched filter. The theoretical basis of an optical matched filter is in the Fourier theorem of matched filtering which states that the cross correlation between two signals can be formed by multiplying the Fourier transform of the signal and the complex conjugate of the second signal and followed by a subsequent Fourier transform operation. Mathematically:

$$\text{Correlation}(f, g) = \mathrm{F}^{-1}[\mathrm{F}\{f\}^{*}\text{conjugate}(\mathrm{F}\{g\})] \qquad [16]$$

where F represents a Fourier transform operation and F^{-1} represents an inverse transform. Since the lens can easily perform the Fourier transform, and multiplication can be performed by light passing through a medium, the only problem that has to be

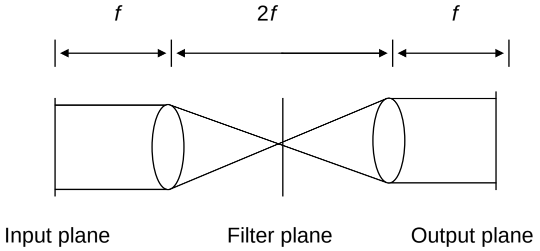

Figure 12 An optical 4-*f* set up. Lenses perform Fourier transform at their focal plane.

solved is to represent the signal and the filter in the optical domain. Since an SLM can represent both the real signal and the complex filter, thus the whole operation can be performed in a 4-*f* setup as shown in Figure 12. The input is displayed in the input plane by an SLM, the first lens performs a Fourier transform. This filter (conjugate F{*g*}) is encoded using another SLM and placed at the filter plane. When light from the F{*f*} passes through the filter, it performs the required multiplication. The second lens performs a second Fourier transform and produces the correlation output at the output plane. Variations of the basic setup exist as a joint transform correlator (JTC) or binary phase-only filter (BPOF). While the JTC achieves the product of the complex domain by adding the two signals and squaring, the BPOF is achieved by simplifying the complex domain filter by numerically approximating it with the binarized phase. Other applications of linear optical processing can be performed in the basic 4-*f* optical setup where filters such as low-pass or high-pass filters, can be set up in the correlation plane.

The hologram base computing treats holograms as a legitimate way of signal transformation from one form to another, just as a lens transforms from one domain to another. Holograms are complex representations that can act on a signal. A lens can be approximated by a hologram, in reality functions of many optical elements, such as lens, grating, prism, and other transformation such as beamsplitting, it can be combined into single computing element known as computer-generated hologram (CGH). The smartness of this type of system depends on how many operations have been replaced by a single CGH.

It is envisioned that optical computing is better suited for special purpose computing rather than general purpose computing. However, the future of optical computing lies in the marriage of appropriate device technology with the right architecture at the right applications. The development of quantum computing may hold new promises for optical computing.

List of Units and Nomenclature

DMAC	Digital multiplication by analog convolution
JTC	Joint transform correlator
OC	Optical computing
PLA	Programmable logic array
QC	Quantum computing
SLM	Spatial light modulator

See also

Nonlinear Optics, Basics: Four-Wave Mixing; Nomenclature and Units. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Arsenault H, Szoplik T and Macukow B (eds) (1989) *Optical Processing and Computing*. San Diego, CA: Academic Press.

Boyd RW (1992) *Nonlinear Optics*. San Diego, CA: Academic Press.

Gaskill J (1978) *Linear Systems Fourier Transform and Optics*. New York: Wiley.

Gibbs HM (ed.) (1982) *Optical Bistabiligy II*. New York: Plenum Press.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw Hill.

Ichioka Y and Tanida J (1984) Optical parallel logic gates using a shadow-casting system for optical digital computing. *Proceedings of the IEEE* 72: 787.

Karim MA and Awwal AAS (1992) *Optical Computing: An Introduction*. New York: Wiley.

Koren I (1993) *Computer Arithmetic Algorithms*. New Jersey: Prentice Hall.

McAulay A (1992) *Optical Computing Architecture*. New York: Wiley.

Reynolds GO, DeVelis JB, Parrent GB and Thompson BJ (eds) (1989) *Physical Optics Notebook: Tutorials in Fourier Optics*. Bellingham, CA: SPIE Optical Engineering Press.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley & Sons.

VanderLugt A (1992) *Optical Signal Processing*. New York: Wiley.

Yu FTS and Jutamulia S (1992) *Optical Signal Processing, Computing and Neural Networks*. New York: Wiley.

Incoherent Analog Optical Processors

S Jutamulia, University of Northern California, Petaluma, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

After low-cost lasers became widely available, it seemed that incoherent analog optical processing was no longer important. In general, incoherent optical processing has no apparent significant advantages, as compared with its coherent optical counterpart or digital computer processing. However, although nobody now attempts to build an incoherent optical processor, such as an incoherent correlator, most instruments are still operating using incoherent or natural light. These incoherent instruments relate to imaging, including cameras, microscopes, telescopes, projection displays, and lithographic equipment. From this point of view, it is still very important to study the principles of incoherent analog optical processing, in order to further progress or improve the incoherent image.

Incoherent Image Formation

The image formation in a coherent optical system can be explained using the linear system. A perfect point $\delta(x, y)$ in the input plane is imaged in the output plane as the impulse response of the system $h(x, y)$. When the input is only a single point, the output amplitude is simply $h(x, y)$ and the output intensity is $|h(x, y)|^2$. Accordingly, the function of $|h(x, y)|^2$ is called point spread function (PSF).

If we consider that the input object is the collection of a large number of very fine points, under coherent light illumination, the output image will be the collection of the same number of blurred spots. The shape of each blurred spot is $h(x, y)$. If the amplitude function of input object is $f(x, y)$, the amplitude function of output image is

$$o(x, y) = \iint_{-\infty}^{\infty} f(p, q) h(x - p, y - q) \mathrm{d}p \, \mathrm{d}q \qquad [1]$$

which is a convolution of $f(x, y)$ and $h(x, y)$. The intensity function of output image is

$$|o(x, y)|^2 = \left| \iint_{-\infty}^{\infty} f(p, q) h(x - p, y - q) \mathrm{d}p \, \mathrm{d}q \right|^2 \qquad [2]$$

According to the convolution theorem, in the Fourier transform domain or frequency domain, we have

$$O(u, v) = F(u, v) H(u, v) \qquad [3]$$

where $O(u, v)$, $F(u, v)$, and $H(u, v)$ are the Fourier transforms of $o(x, y)$, $f(x, y)$, and $h(x, y)$, respectively. $H(u, v)$ is the transfer function of the coherent imaging system. To distinguish it from the incoherent imaging system, it is also called coherent transfer function (CTF).

The same optical imaging system is now illuminated with incoherent light instead of coherent light. The amplitude impulse response is still the same $h(x, y)$, and the point spread function is also the same $|h(x, y)|^2$. Similarly, the input object is the collection of perfect points, and the output image is the collection of blurred images of each point. However, since the illuminating light is incoherent, light from any point in the input object is not coherent with light from any other points. Thus, there is no interference with light from different points. The output image is simply the addition of intensity patterns generated by each point in the input object. The amplitude function of the output image is not computable. However, the intensity function of the output image is

$$|o(x, y)|^2 = \iint_{-\infty}^{\infty} |f(p, q)|^2 |h(x - p, y - q)|^2 \mathrm{d}p \, \mathrm{d}q \qquad [4]$$

In the frequency domain, we now have

$$O_I(u, v) = F_I(u, v) H_I(u, v) \qquad [5]$$

where $O_I(u,v)$, $F_I(u,v)$, $H_I(u,v)$ are Fourier transforms of $|o(x,y)|^2$, $|f(x,y)|^2$, and $|h(x,y)|^2$, respectively. The intensity impulse response $|h(x,y)|^2$ is the PSF. Note that the index I of O_I, F_I, and H_I denotes intensity. The function $H_I(u,v)$ is now the transfer function of the incoherent imaging system, which is called optical transfer function (OTF).

Referring to Fourier analysis, if $H(u, v)$ is the Fourier transform of $h(x, y)$, the Fourier transform of $|h(x, y)|^2$ is the autocorrelation of $H(u, v)$ as follows:

$$H_I(u, v) = \iint_{-\infty}^{\infty} H(p, q) H^*(p - u, q - v) \mathrm{d}p \, \mathrm{d}q \qquad [6]$$

The OTF (the function $H_I(u, v)$) is the autocorrelation of CTF (the function $H(u, v)$), which is a complex function in general. It can be written as

$$\mathrm{OTF} = |\mathrm{OTF}| \exp(i\phi) \qquad [7]$$

We can further write

$$\mathrm{MTF} = |\mathrm{OTF}| \qquad [8]$$

where MTF stands for modulation transfer function, and

$$\mathrm{PTF} = \phi \qquad [9]$$

where PTF stands for phase transfer function.

Measurement of MTF

The MTF of an incoherent imaging system can be measured empirically using an input that is a cosine grating having intensity transmittance function:

$$|f(x)|^2 = 1 + \cos(2\pi a x) \qquad [10]$$

where a is the frequency of grating. The output or the image intensity function will be

$$|o(x)|^2 = 1 + m\cos(2\pi a x + \varphi) \qquad [11]$$

where m, the contrast, can be measured by

$$m = \frac{|o(x)|^2_{\max} - |o(x)|^2_{\min}}{|o(x)|^2_{\max} + |o(x)|^2_{\min}} \qquad [12]$$

In the frequency domain, the Fourier transform of the input function is

$$F_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a) + \tfrac{1}{2}\delta(u+a) \qquad [13]$$

The Fourier transform of the output function will be

$$O_I(u) = F_I(u)\mathrm{OTF}(u) \qquad [14]$$

Substitution of eqn [13] yields

$$O_I(u) = \delta(u)\mathrm{OTF}(0) + \tfrac{1}{2}\delta(u-a)\mathrm{OTF}(a) + \tfrac{1}{2}\delta(u+a)\mathrm{OTF}(-a) \qquad [15]$$

Notice that

$$\mathrm{OTF}(0) = 1 \qquad [16]$$

due to normalization, and

$$\mathrm{OTF}(a) = \mathrm{OTF}(-a) \qquad [17]$$

Because OTF is an autocorrelation function, it must be a symmetric function. Substitution of eqns [16] and [17] into eqn [15] yields

$$O_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a)\mathrm{OTF}(a) + \tfrac{1}{2}\delta(u+a)\mathrm{OTF}(a) \qquad [18]$$

Remembering that OTF is a complex function given in eqn [7], we can write

$$\mathrm{OTF}(a) = \mathrm{MTF}(a)\exp(i\phi(a)) \qquad [19]$$

Substitution of eqn [19] into eqn [18] yields

$$O_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a)\mathrm{MTF}(a)\exp(i\phi(a)) + \tfrac{1}{2}\delta(u+a)\mathrm{MTF}(a)\exp(i\phi(a)) \qquad [20]$$

The inverse Fourier transform of eqn [20] is

$$|o(x)|^2 = 1 + \mathrm{MTF}(a)\cos(2\pi a x + \phi(a)) \qquad [21]$$

By comparing eqn [21] and eqn [11], we find

$$\mathrm{MTF}(a) = m \qquad [22]$$

and the PTF

$$\phi(a) = \varphi \qquad [23]$$

Therefore, the MTF at frequency $u=a$ of the incoherent imaging system can be measured using eqn [12]. To get the complete MTF(u), the measurement is repeated using cosine gratings with different frequencies. The PTF can also be measured at the same time.

Incoherent Spatial Filtering

In the 4f coherent optical processor, the first lens performs the Fourier transform, and the second lens performs the inverse Fourier transform, as shown in Figure 1. The Fourier transform domain or frequency domain is materialized in the frequency plane that is the back focal plane of the first lens.

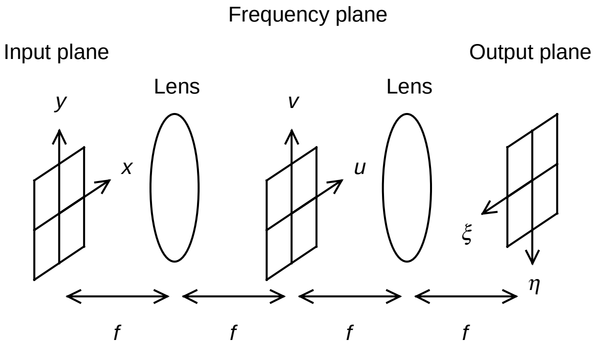

Figure 1 4f coherent optical processor consists of two Fourier transform lenses.

Thus, the Fourier transform of the object can be visualized in the frequency plane. For example, if the object is a grating, we will see its frequency components in the frequency plane. However, if the coherent illumination (e.g., from a laser) is replaced with incoherent illumination (e.g., from a lightbulb), the pattern of Fourier transform in the frequency plane disappears. Does it mean that we can no longer perform spatial filtering? No, we still can perform spatial filtering, although the pattern of Fourier transform is no longer visualized.

Consider that we have a spatial filter in the frequency plane whose physical form can be expressed by the function $H(u,v)$. For illustration, this spatial filter is simply a small one-dimensional window $H(u)$, as shown in Figure 2a. If the 4f optical processor is illuminated with coherent light, the CTF of the system is $H(u)$ itself as shown in Figure 2b. However, if the 4f optical processor is illuminated with incoherent light, we will use OTF instead of CTF in the analysis of linear system. The OTF is the autocorrelation of CTF, which is graphically depicted in Figure 2c. In this example, since the OTF is real and positive, the MTF is identical to the OTF.

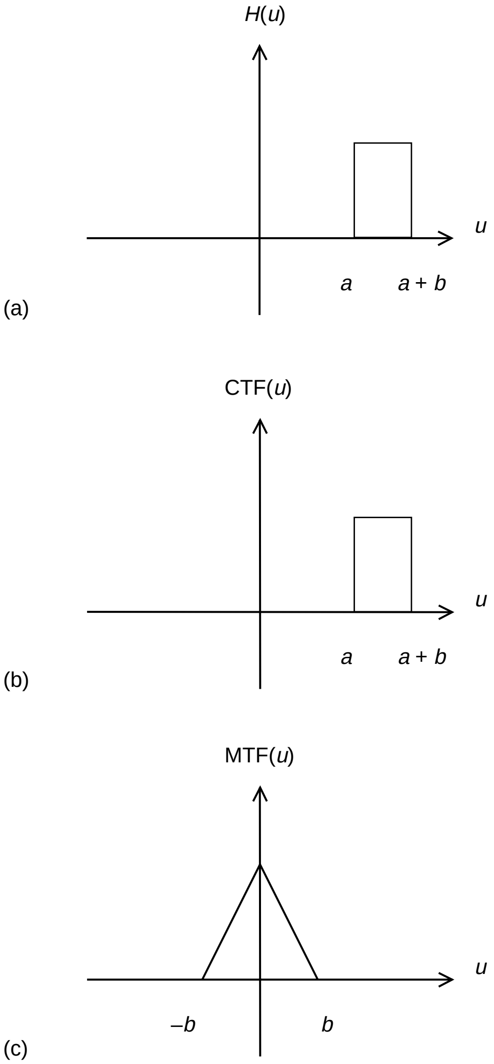

Figure 2 (a) One-dimensional window in frequency plane functions as a spatial filter. (b) CTF and (c) OTF or MTF produced by spatial filter (a).

For simplicity, we analyze and show a one-dimensional filter only in a two-dimensional graph. It certainly can be extended to the two-dimensional filter. And we need a three-dimensional graph to show the two-dimensional autocorrelation.

It is important to note that, although the CTF is a high pass filter that passes frequencies in the band of $u = a$ to $u = a + b$, the MTF is a low pass filter that passes frequencies from $u = -b$ to $u = b$, independent of a. This indicates that an incoherent processor cannot enhance high frequency. It is always a low pass processor. It is interesting to note that the OTF is independent of the location of the filter $u = a$, but it is determined by the width of the filter $\Delta u = b$ only.

Accordingly, an aperture made up of randomly distributed pinholes behaves as a low pass filter in an incoherent imaging system. The cut-off frequency of such a filter can be derived from the diameter of the pinhole, which is equivalent to Figure 2b. The procedure for using this filter is simple. If a camera is used to take a picture, the random pinhole filter can be simply attached to the camera lens. The picture taken will not consist of high-frequency components, because the filter acts like a low pass filter. The low pass filtering can also be done by reducing the aperture of camera. However, by doing so, the light entering the camera is also reduced. The use of a random pinhole filter removes high frequencies without reducing the light intensity. When the random pinhole filter is placed in the frequency plane of a 4f coherent optical processor, it can be used to reduce the speckle noise. Note that the speckle noise is the main drawback in coherent optical processing.

Incoherent Complex Matched Spatial Filtering

Lohmann and Werlich pointed out that the Vander Lugt correlator can also be operated with incoherent light. In the coherent Vander Lugt correlator the correlation term is:

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta) g*(\alpha-\xi,\beta-\eta)\,\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [24]$$

Its intensity distribution is

$$|o_3(\xi,\eta)|^2 = \left| \int\int_{-\infty}^{\infty} f(\alpha,\beta) g*(\alpha-\xi,\beta-\eta) \mathrm{d}\alpha \mathrm{d}\beta \right|^2 \tag{25}$$

However, when the input object is illuminated with incoherent light instead of coherent light, the intensity distribution of the correlation term becomes

$$|o_3(\xi,\eta)|^2 = \int\int_{-\infty}^{\infty} |f(\alpha,\beta)|^2 |g(\alpha-\xi,\beta-\eta)|^2 \mathrm{d}\alpha \mathrm{d}\beta \tag{26}$$

Equation [26] shows the correlation of $|f(x,y)|^2$ and $|g(x,y)|^2$. Therefore, when $f(x,y)$ is identical to $g(x,y)$, the correlation peak always appears regardless of whether the illumination is coherent or incoherent. However, the detailed structures of the coherent and incoherent correlation outputs are different. Note that the joint transform correlator cannot be operated using incoherent light, since no joint Fourier spectra will be formed.

Depixelization of Projection Display

Recently, a technique to remove the pixel structure from an image projected by liquid crystal projection display has been reported. A liquid crystal projection display must have a pixel structure to form an image. Can we remove this pixel structure from the projected image to obtain a movie-like image? It is well known that a single hole in the frequency plane can perform a coherent spatial filtering, such that the pixel structure will disappear from the output image.

Because of the pixel structure, we get multiple Fourier spectra in the frequency plane. By passing only one spectrum order through the hole in the frequency plane, the pixel structure can be removed as shown in Figure 3.

It is interesting to note that selecting any one of the spectra will produce the image at the same location.

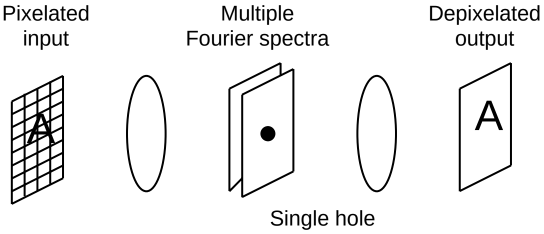

Figure 3 Removal of pixel structure using a single hole in frequency plane. Most of the energy is wasted and the output intensity is dim.

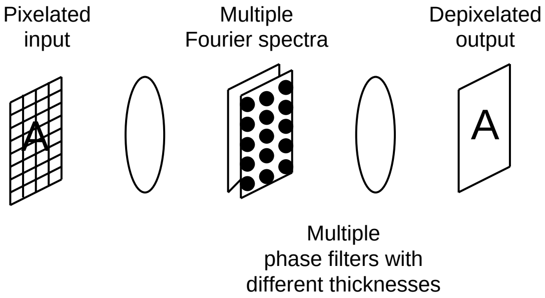

Figure 4 Removal of pixel structure using multiple phase filters with different thickness to cover each Fourier spectrum order. No energy is wasted and the output intensity is very bright.

However, if two spectra are passed, on the top of the produced image, Young interference fringes are also produced. If all spectra are passed, the produced interference fringes are in fact the pixel structure itself. By passing only one spectrum order and blocking most spectrum orders, we lose most of the energy. The projected image will be very dim, although the pixel structure is removed.

To overcome this problem, we may cover each spectrum order with a transparent material having different thickness as shown in Figure 4. Thus, every spectrum order is passed and delayed by a phase filter with a different thickness. If the delay produced by the phase filter is larger than the coherent length of light, every spectrum order is no longer coherent to each other. In other words, the resultant image is the sum of the intensities of all images. As a result, the pixel structure does not appear, and no intensity is lost. For a white light source, since the frequency bandwidth is large, the coherent length is typically on the order of several tens of μm.

This technique can significantly improve the quality of the liquid crystal projection display. Figure 5 shows the experimental result of a depixelated projection image. Figure 5a shows the projected image of an input with pixel structure when no phase filter is applied in the frequency plane. Figure 5b shows the projected image of the same input with pixel structure when phase filters are applied in the frequency plane. The pixel structure is successfully removed as shown in Figure 5b.

Computed Tomography

The computed tomography or CT using X-ray is usually considered beyond optical information processing. However, it is interesting to review the principle of the X-ray CT, since the X-ray source is incoherent, and the CT image reconstruction utilizes Fourier transformation that is commonly used in

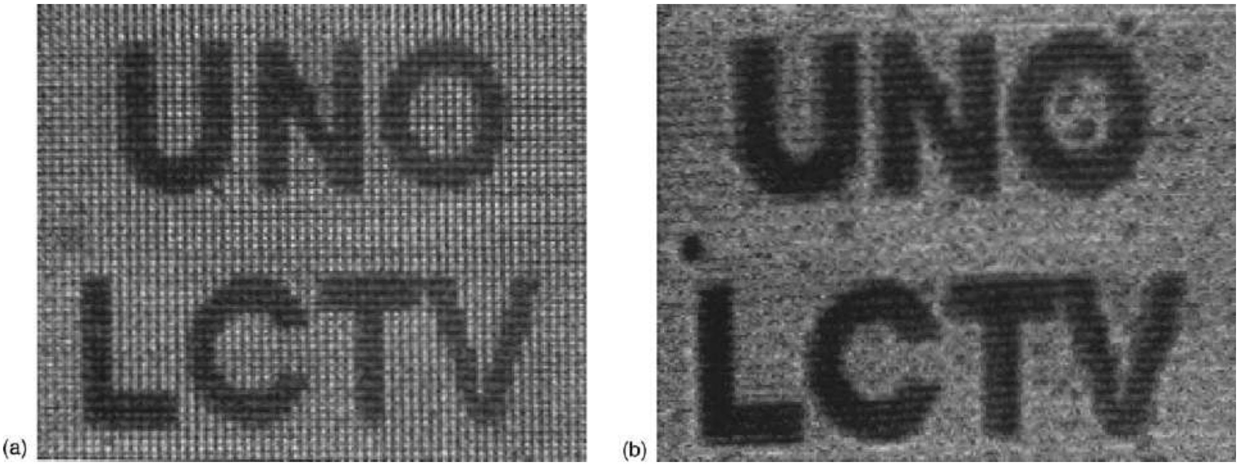

Figure 5 (a) Projected image consists of pixel structure when no phase filter is applied in the frequency plane. (b) Depixelated image is produced when phase filters are applied in the frequency plane.

optical information processing. The X-ray CT is used for taking the cross-section picture of the human body. Note that a conventional X-ray picture only provides a projection instead of a cross-section picture.

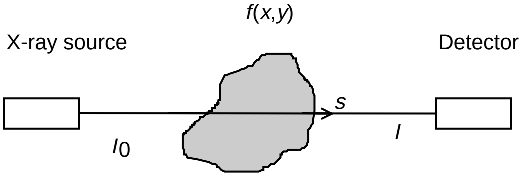

Figure 6 X-ray passes through object having density function $f(x, y)$.

Consider the case when an X-ray penetrates through an object characterized by its attenuation coefficient or density function $f(x, y)$ as shown in Figure 6. The detected X-ray intensity is

$$I = I_0 \exp - \left[\int f(x, y) \mathrm{d}s \right] \qquad [27]$$

where I_0 is the original intensity. Equation [27] can be written as

$$\int f(x, y) \mathrm{d}s = -\ln \frac{I}{I_0} \qquad [28]$$

Apparently, we cannot directly measure $f(x, y)$, although we can obtain $\int f(x, y) \mathrm{d}s$ from the measurement of I and I_0. However, we want to get the cross-section picture which is the density function $f(x, y)$. The basic idea is to compute $f(x, y)$ from its Fourier transform $F(u, v)$, and $F(u, v)$ is derived from eqn [28]. $F(u, v)$ can be expressed as

$$F(u, v) = \int \int_{-\infty}^{\infty} f(x, y) \exp[-\mathrm{i}2\pi(ux + vy)] \mathrm{d}x\, \mathrm{d}y \qquad [29]$$

For $v = 0$, eqn (29) becomes

$$F(u, 0) = \int \int_{-\infty}^{\infty} f(x, y) \exp[-\mathrm{i}2\pi ux] \mathrm{d}x\, \mathrm{d}y \qquad [30]$$

which can be written

$$F(u, 0) = \int_{-\infty}^{\infty} p(x) \exp[-\mathrm{i}2\pi ux] \mathrm{d}x \qquad [31]$$

where

$$p(x) = \int_{-\infty}^{\infty} f(x, y) \mathrm{d}y \qquad [32]$$

The projection function $p(x)$ given by eqn [32] can be obtained using an X-ray parallel beam as shown in Figure 7. Then $F(u, 0)$ can be computed using computer based on eqn [31]. $F(u, 0)$ is the Fourier transform along the u-axis. To get other data, we rotate the coordinate (u, v) to (u', v') by a small angle α. Correspondingly, the coordinate (x, y) is also rotated to (x', y') by the same angle α while the object is not rotated as shown in Figure 8.

From the measurement shown in Figure 8, we can get $p(x')$, which, in turn, provides $F(u', v' = 0)$. By increasing the rotation angle α, we will get other data of $F(u'', v'' = 0)$. The completion of rotation of 180 degrees will give us the data in the frequency domain (u, v) as shown in Figure 9.

After the data $F(u, v)$ is collected, as shown in Figure 9, $F(u, v)$ can be inverse Fourier transformed to

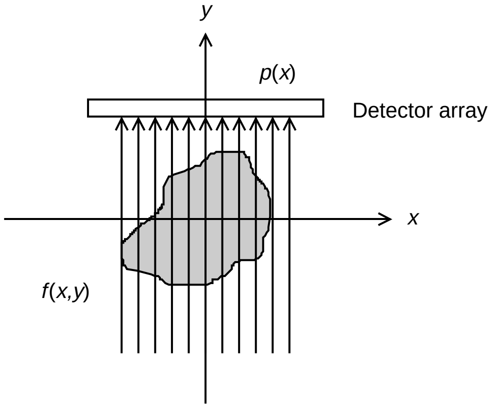

Figure 7 X-ray parallel beam passes through the object to produce projection function $p(x)$.

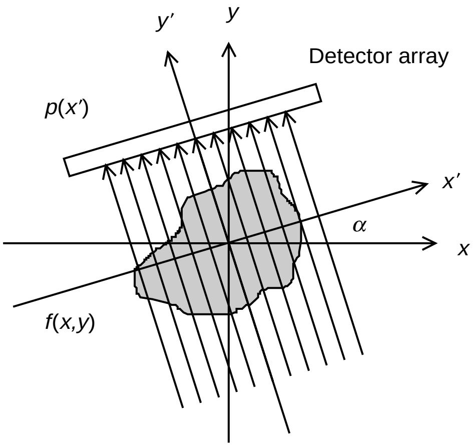

Figure 8 Rotated X-ray parallel beam produces the rotated projection function $p(x')$.

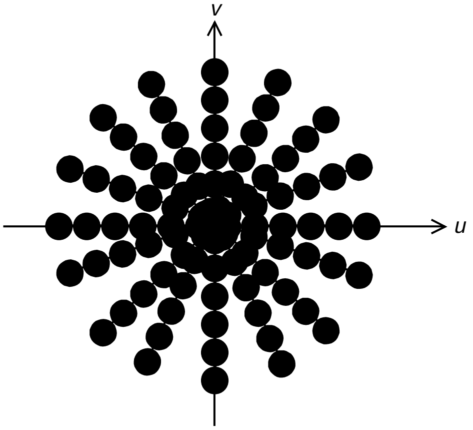

Figure 9 $F(u,v)$ computed from a set of rotated projection functions.

produce $f(x,y)$. There are two approaches in the inverse Fourier transform. The first approach is to interpolate the computed value $F(u,v)$ such that the value of $F(u,v)$, on a regular grid structure, can be defined before an inverse Fourier transform can be taken.

The second approach is to apply polar coordinates, instead of Cartesian coordinates, to the inverse Fourier transform, such that no interpolation in frequency domain (u,v) is required. The inverse Fourier transform in Cartesian coordinates is as follows:

$$f(x,y)=\int\int_{-\infty}^{\infty} F(u,v)\exp[\mathrm{i}2\pi(ux+vy)]\mathrm{d}u\,\mathrm{d}v \qquad [33]$$

It can be written in polar coordinates as:

$$f(x,y)=\int_0^{2\pi}\int_0^{\infty} F(\rho\cos\phi,\rho\sin\phi) \times\exp[\mathrm{i}2\pi\rho(x\cos\phi+y\sin\phi)]|\rho|\mathrm{d}\rho\,\mathrm{d}\phi \qquad [34]$$

where

$$u=\rho\cos\phi \qquad [35]$$

$$v=\rho\sin\phi \qquad [36]$$

Therefore $f(x,y)$ can be obtained from the collected $F(u,v)$, as shown in Figure 9, without further interpolation using eqn [34]. Note that eqns [31], [33], and [34] are computed using a computer program.

See also

Spectroscopy: Fourier Transform Spectroscopy.

Further Reading

Bracewell RN (1965) *The Fourier Transform and its Applications*. New York: McGraw-Hill.

Hounsfield GN (1973) Computerized transverse axial scanning (tomography): Part I. *British Journal of Radiology* 46: 1016–1022.

Iizuka K (1985) *Engineering Optics*. New York: Springer-Verlag.

Jutamulia S, Toyoda S and Ichihashi Y (1995) Removal of pixel structure in liquid crystal projection display. *Proceedings of SPIE* 2407: 168–176.

Klein MV (1970) *Optics*. New York: Wiley.

Lohmann AW and Werlich HW (1971) Incoherent matched filtering with Fourier holograms. *Applied Optics* 10: 670–672.

Reynolds GO, DeVelis JB, Parrent GB Jr and Thompson BJ (1989) *The New Physical Optics Notebook: Tutorials in Fourier Optics*. Bellingham, WA: SPIE.

Vander Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Trans. Inf. Theo.* IT-10: 139–145.

Yang X, Jutamulia S and Li N (1996) Liquid-crystal projection image depixelization by spatial phase scrambling. *Applied Optics* 35: 4577–4580.

Yu FTS and Wang EY (1973) Speckle reduction in holography by means of random spatial sampling. *Applied Optics* 12: 1656–1659.

Optical Bit-Serial Computing

A D McAulay, Lehigh University, Bethlehem, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Why Optical Bit-Serial Computing?

Optical bit-serial computing will lower cost and improve dataflow rates in optical telecommunications and associated computing. Moreover, the electromagnetic immunity of optics will enhance security for computing as well as for communications. Also, absence of sparking with optics enables safe operation, even in hazardous environments.

Currently, optical bit-serial communication involves passing terabits of data per second (10^{12} bps) – thousands of encyclopedias per second – along an optical fiber by switching light of multiple colors on and off, representing ‘1’ and ‘0’ bits respectively. Replacing electronic computing functions, constrained to tens of gigabits per second (10^9 bps), with integrated optics, ones of much higher speed, enables faster flow rates – allowing continuation of the annual doubling of capacity-distance that has occurred since the 1970s. Cost is reduced by eliminating the demultiplexing and multiplexing for optic-electronic-optic (OEO) conversions. Two illustrative areas of optical bit-serial computing for replacing electronic computing functions are described: computing at up to 40 Gbps with semiconductor optical amplifiers (SOAs) and computing at higher bit rates with integrated optic components. Several books address the relevant technical background: free-space optical information processing, optical computer architectures, optical fiber communications, and integrated optics.

Computing with Semiconductor Optical Amplifiers

Semiconductor optical amplifiers, SOAs, allow nonlinear operations (switching, logic) at low power (mW) because of amplification – typically up to 1000. An SOA is a laser diode with antireflection coatings in place of reflecting facets. Light pulses at bit rates in excess of 40 Gbps are amplified by absorbing power from an electronic pump. SOAs are used in the following, at bit rates up to 40 Gbps, for nonlinear operations in cross-gain, cross-phase, or cross-polarization.

SOA-XGM for Logic

Figure 1 shows an SOA in cross-gain modulation (XGM) performing a NOR function. Either A or B inputs at a ‘1’ level, driving the SOA through couplers into saturation, lowering its gain. A continuous wave at C, having a different frequency λ_3, from that of A, λ_1 and B, λ_2, experiences the decrease in gain, resulting in an output ‘0’ level at frequency λ_3. The filter in Figure 1 can be avoided by counter-propagating A and B through the SOA from right to left. XOR gates use interference between two incoming streams of bits (locked to identical frequency, phase, polarization) to correlate headers on internet protocol (IP) address headers – similar to matching zip codes in postal mail routing machines. Bit-serial-adders are constructed from combinations of XOR and NOR gates; subsequently the adders are used to construct ripple-carry-adders for word addition. In addition, NOR gates are used to construct flip-flops; hence registers which form short-term memory.

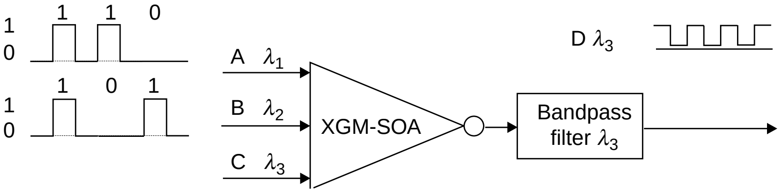

Figure 1 Optical XGM-SOA NOR gate.

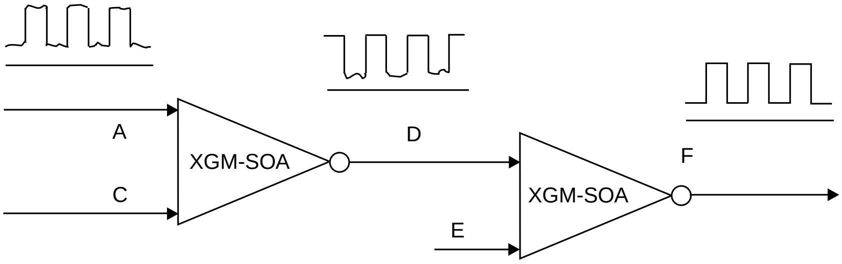

Figure 2 Optical XGM-SOA 2R signal regeneration.

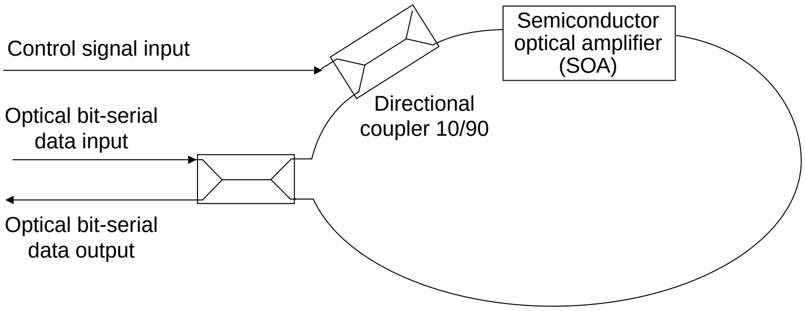
Control signal input
Optical bit-serial data input
Optical bit-serial data output
Directional coupler 10/90
Semiconductor optical amplifier (SOA)

Figure 3 XPM-SOA in Sagnac interferometer for optical manipulation of bits at high data rates.

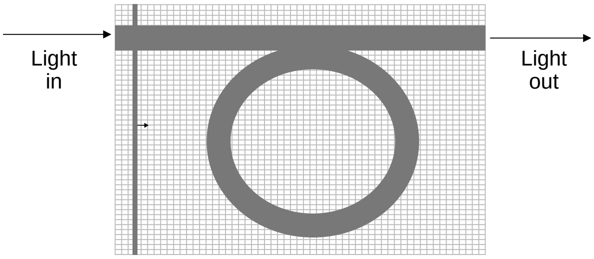

Figure 4 Integrated optic Optiwave layout for microring resonator.

SOA-XGM for Frequency Conversion

Ignoring input B, the bit information in Figure 1 has been transferred from carrier frequency A onto the carrier frequency C, with inversion. This enables switching of signals in wavelength division multiplexing (WDM), for example, color reuse is simplified and a color red is routed to the location where blue was previously routed; this is equivalent to time slot interchangers for switching in time division multiplexing (TDM).

SOA-XGM for 2-R Signal Regeneration

Figure 2 show how two SOAs may be used in sequence to restore signal levels distorted by propagation or computation; reamplification and reshaping (2-R). The ‘1’ level at the output of the first SOA is clipped by gain saturation, removing noise, this becoming the ‘0’ level at the output of the second SOA. The second SOA clips the ‘1’ level at the output.

SOA-XPM with Interferometer

In cross-phase modulation (XPM), the signal at A, Figure 1, changes the phase of carrier for signal C passing through the SOA (signal levels may be less than for XGM). Inclusion of an SOA in an interferometer, Mach–Zehnder or Sagnac, converts phase to intensity for frequency conversion, switching or high-speed bit manipulation. Figure 3 shows a control pulse entering a coupler at the left and providing a control input to the SOA at one frequency. The bits circulating round the loop at a different frequency can be switched on and off by the control pulse, to synchronize or manipulate high bit rate signals or for frequency conversion.

Computing with Integrated Optics

Combining several discrete optical components into a single optical chip is more beneficial than in electronic very large-scale integration (VLSI) because optical connections require greater precision; moreover bit rates are higher because photons, unlike electrons, have no mass or charge (only spin).

Integrated Optic Microring Resonator

Figure 4 shows a microring resonator in an integrated optic chip for optical filtering, light entering at the left couples into the loop. For loop lengths that are a multiple of wavelength, light cycling the loop is in phase and resonance occurs. At resonant frequencies,

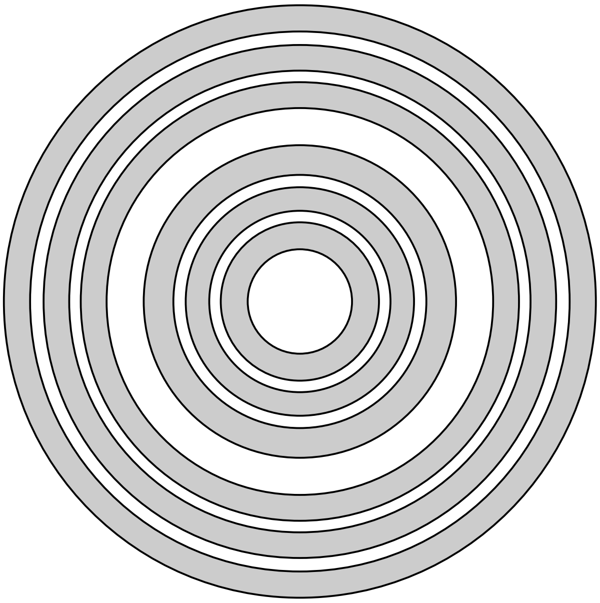

Figure 5 Integrated optic Optiwave layout for Bragg microring resonator.

light exiting the loop is out of phase with that in the linear guide, causing cancellation of the input. As frequency varies away from resonance, light passes and the device acts as a notch filter to block light at the resonant frequencies. With nonlinear Kerr material, low-level light intensity may be used to change a resonant frequency to make a switch.

Bragg Integrated Optic Microring Resonator

Performance is improved by using a Bragg microring resonator, Figure 5, which confines light in the central light colored ring by dielectric mirror reflection from Bragg gratings on either side. The refractive index is less in the ring than in the surrounding regions, in contrast to a conventional waveguide (used in Figure 4) that guides by total internal reflection. The higher reflectivity, smaller ring radius and lower refractive index – these help to reduce power loss, volume, and travel time around the ring. This increases filter selectivity, drives up peak power for enhanced nonlinearity for switching, decreases stored energy for fast response, and increases frequency range (free spectral range, *FSR*). In this case, an equivalent photonic crystal may be more expensive because it requires higher resolution lithography and greater refractive index contrast.

Photonic Crystal Mach–Zehnder Switch

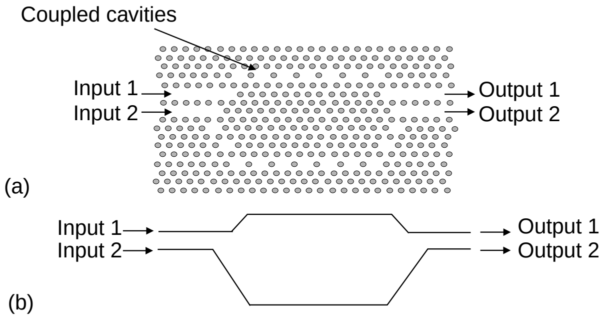

Figure 6 Proposed photonic crystal Mach–Zehnder interferometer.

Recently developed photonic crystal technology allows a reduction in integrated optic chip size from centimeters to submillimeters. The 2D or 3D periodic structure of dielectric variations emulates crystals and provides dielectric mirror reflections in multiple dimensions for lossless waveguide bends, high *Q* resonators, and slowing of wave propagation to enhance effective nonlinearity of materials. Figure 6 shows a proposed 2D photonic crystal Mach–Zehnder interferometer switch in which light is split into two paths by a coupler and propagation velocity is slowed for enhancing nonlinearity by using coupled-resonator optical waveguides (CROWs), in which alternate posts of different dielectric constant remain in the waveguides to form coupled cavities.

Conclusion

Optical bit-serial computing can reduce cost and increase flow-rate for expanding telecommunication network capacity, while providing enhanced security from electromagnetic sabotage and tampering. Advances in integrated optics (microring resonators), progress in optoelectronic components (SOAs), improved materials and fabrication techniques – these technologies enable cost-effective optical bit-serial computing solutions for future telecommunication evolution.

See also

All-Optical Signal Regeneration. **Information Processing:** All-Optical Multiplexing/Demultiplexing.

Further Reading

Agrawal G (2002) *Fiber-Optic Communication Systems*. New York: Wiley.

Connelly MJ (2002) *Semiconductor Optical Amplifiers*. Boston, MA: Kluwer Academic Publishers.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw-Hill.

Heuring VP, Jordan HF and Pratt JP (1992) A bit serial architecture for optical computing. *Applied Optics* 31(17): 3213–3224.
Hunsperger RG (2002) *Integrated Optics*, 5th edn. New York: Springer.
Joannopoulios JD, Meade RD and Winn JN (1995) *Photonic Crystals*. Princeton, NJ: Princeton University Press.
McAulay AD (1991) *Optical Computer Architectures*. New York: Wiley.
McAulay AD (2004) Novel all-optical flip-flop using semiconductor optical amplifiers in innovating frequency-shifting inverse-threshold pairs. *Optical Engineering* 43(5): 1115–1120.
Scheuer J and Yariv A (2003) Two-dimensional optical ring resonators based on radial Bragg resonance. *Optics Letters* 28(17): 1528–1530.
Xu L, Glesk I, Baby V and Prucnal PR (2004) All-optical wavelength conversion using SOA at nearly symmetric position in a fiber-based Sagnac interferometric loop. *IEEE Photonics Technical Letters* 16(2): 539–541.
Yariv A (1997) *Optical Electronics in Modern Communications*, 5th edn. New York: Oxford University Press.

Optical Digital Image Processing

B L Shoop, United States Military Academy, West Point, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Optical approaches to image processing provide a natural method by which to obtain the inherent advantages associated with optics including massive parallelism, high-speed processing, and the inherent compatibility with image formats. Early applications of optics employed analog processing techniques with the most common being the optical Fourier transform, matrix-vector processors, and correlators. Later, advances in digital signal processing (DSP), digital storage, and digital communication systems demonstrated the potential for higher resolution, improved flexibility and functionality, and increased noise immunity over analog techniques and quickly became the preferred method for accurate signal processing. Consequently, optical processing architectures, that incorporated digital processing techniques, were explored. Advances in very large-scale integration (VLSI) electronic circuitry and optoelectronic devices later enabled smart pixel technology, which integrates the programmability and processing of electronic circuitry with the two-dimensional (2D) nature of an optical architecture and made possible a particular realization of optical digital image processing. This chapter describes an application of optical digital image processing, based on smart pixel technology, called digital image halftoning.

Natural images are, by definition, continuous in intensity and color. Halftoning is the process by which a continuous-tone, gray-scale image is converted to one containing only binary output pixels. Halftoning can be thought of as an image compression technique whereby a high-resolution image is transformed to a low-resolution image containing only black-and-white pixels. The transformation from a continuous-tone, gray-scale image to one containing only binary-valued pixels is also similar, in principle, to low-resolution analog-to-digital (A/D) conversion.

An early manual approach to printing continuous-tone images, using only the presence or absence of ink, is the Mezzotint. Ludwig von Siegen (1609–1680), an amateur Dutch printmaker, first invented the mezzotint, or ‘half tone process’, in 1640. The process later came into prominence in England during the early eighteenth century. Mezzotints are produced on copper plates where the surface of the plate is roughed with a tool called the Mezzotint Rocker, shaped like a wide chisel with a curved and serrated edge. By rocking the toothed edge backwards and forwards over the plate, a rough burr is produced which holds the ink. The dark regions of an image were roughed in a random fashion, while the areas to be lightened were scraped and burnished. This process was found to be especially useful for the reproduction of paintings, due to its ability to capture the subtlest gradations of tone from rich, velvety blacks to glowing highlights. The mezzotint is, therefore, an early artistic predecessor to the modern day halftone.

Optical halftoning has been in use by commercial printers for over 100 years. Commercial halftone screens are based on a discovery made by William Henry Fox Talbot (1800–1877), in 1852. He demonstrated the feasibility of optical halftoning by photographing an image through a loosely woven fabric or ‘screen’. In the 1890s, this process came into practical use when the halftone screen, consisting of two ruled glass plates cemented together, became commercially available. Commercial halftone screens produce an effect of variably sized dots on the photographic plate that gives the illusion of a continuous tone image.

Digital image halftoning, sometimes referred to as spatial dithering, is the process of converting an electronic continuous-tone, gray-scale image to one containing only binary-valued pixels, for the purpose of displaying, printing, or storing the image. The underlying concept is to provide the viewer of the image with the illusion of viewing a continuous-tone image when, in fact, only black and white pixel values are used in the rendering. This process is particularly important in applications such as laser printing, bilevel displays, xerography, and more recently, facsimile.

There are a number of different methods by which this digital image halftoning can be accomplished, which are generally classified as either point or neighborhood processes. A point process is one which computes the output pixel value based strictly on some characteristic of the corresponding input pixel. In contrast, a neighborhood process computes a single output pixel based on a number of pixels in a neighborhood or region of the input image. Ordered dither, which is considered a point process, produces an output by comparing a single continuous-tone input value against a deterministic periodic array of threshold values. If the value of the input pixel under consideration is greater than the corresponding threshold value, the corresponding output pixel is rendered white. If the intensity is less that the threshold, it is rendered black. Dispersed-dot ordered dither is a subset of ordered dither in which the halftone dots are of a fixed size while clustered-dot ordered dither uses variable-sized dots and simulates the variable-sized dots of a printer's halftone screen in the rendering. Among the advantages of point process halftoning, in general and ordered dither halftoning specifically, are simplicity and speed of implementation. The primary disadvantage is that ordered dither produces patterns in the halftoned image, which are visually undesirable.

In contrast, halftoning using the error diffusion algorithm, first introduced by Floyd and Steinberg in 1975, employs neighborhood operations and is currently the most popular neighborhood process. In this approach to halftoning, the output pixel value is determined, not solely by the value of the input pixel and some deterministic pattern, but instead by a weighted average of values from pixels in a neighborhood surrounding the specific pixel being computed. Here, the error of the quantization process is computed and spatially distributed or diffused within a local neighborhood in order to influence future pixel quantization decisions within that neighborhood and thereby improve the overall quality of the halftoned image. In classical unidirectional error diffusion, the image is processed sequentially, proceeding from the upper-left to the lower-right. Starting in the corner of the image, the first pixel is thresholded and the quantizer error is calculated. The error is then diffused to neighboring pixels that have not yet been processed according to the scaling and interconnect defined by the error diffusion filter. The remainder of the pixels in the image are subsequently processed with each quantizer input, now being the algebraic sum of the corresponding original input pixel intensity and the weighted error from previously processed pixels. While error diffusion produces halftone images of superior quality, the unidirectional processing of the algorithm continues to introduce visual artifacts that are directly attributable to the algorithm itself.

In an effort to overcome the implementation constraints of serial processing and improve overall halftone image quality, a number of different parallel architectures have been investigated, including several based on neural network algorithms. While neural network approaches can provide distinct advantages in improved halftone image quality, they also present challenges to hardware implementations, including speed of convergence and the physical interconnect requirements. These challenges, in addition to the natural compatibility with imaging media, provide the motivation for developing optical image processing architectures for these types of applications.

The Error Diffusion Algorithm

Figure 1 shows a block diagram of the error diffusion architecture used in digital image halftoning. The input image is represented as $x_{m,n}$, $w_{m,n}$ is the impulse response of a 2D causal, unity gain error diffusion filter, and the output quantized image is described by $y_{m,n} \in \{-1, 1\}$. The quantizer $q[u_{m,n}]$ provides the one-bit thresholding functionality necessary to convert each analog input pixel to a low-resolution digital output pixel. The quantizer error $\varepsilon_{m,n}$ is computed as the difference between the output and input to the quantizer and distributed to adjacent pixels according to the weighting and interconnection specified by the error diffusion filter. The unity gain constraint ensures

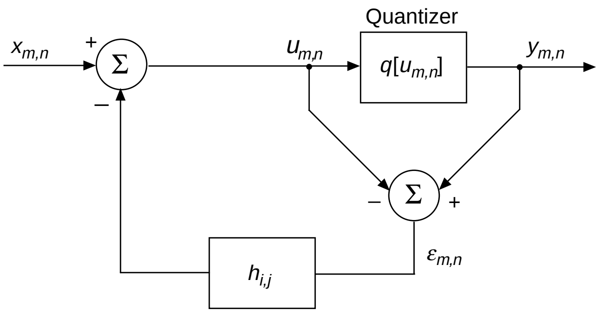

Figure 1 Block diagram of recursive error diffusion architecture.

that no amplification or attenuation of the error signal occurs during the error diffusion process and preserves the average intensity of the image.

In this architecture, the error associated with the quantizer decision at spatial coordinates (m, n) is diffused within a local 2D region to influence adjacent quantization decisions. All of the state variables in this architecture are scalar values since this architecture represents a recursive algorithm in which each pixel in the image is processed sequentially. In the general case, where $w_{m,n}$ is assumed to be a 2D finite impulse response (FIR) filter with coefficients that span a region of support defined by $R_{m,n}$, the quantizer input state $u_{m,n}$ can be written as

$$u_{m,n} = x_{m,n} - \sum_{i,j \in R_{m,n}} w_{i,j} \varepsilon_{m-i,n-j} \qquad [1]$$

Here, $w_{i,j} = 0$, for all $i = j$, since the error signal is diffused spatially within the region of support $R_{m,n}$ and does not influence the pixel under consideration. In this unidirectional error diffusion, the 2D error diffusion filter is therefore noncausal. Figure 2 shows the impulse response of several popular error diffusion filters that have been used in rectangular grid architectures. Figure 2a shows the original error diffusion filter weights developed by Floyd and Steinberg. They argued that the four weights were the smallest number of weights that would produce good halftone images. Figure 2b is a filter that contains a larger region of support developed by Jarvis, Judice, and Ninke in 1976. Finally, the filter in Figure 2c, developed by Stucki, contains the same region of support with coefficients that are a multiple of 2 for digital compatibility and computational efficiency. Here, '•' represents the pixel being processed and the weights describe the local region over which the error is diffused. The normalization factors ensure that the filter coefficients sum to one and therefore meet the unity gain criterion.

To understand the effect of error diffusion and its impact on the quantizer error, it is instructive to describe the effect of error diffusion mathematically. Consider the relationship between the quantizer error $\varepsilon_{m,n} \equiv y_{m,n} - u_{m,n}$ and the overall quantization error $\hat{\varepsilon}_{m,n} \equiv y_{m,n} - x_{m,n}$ in the frequency domain. If we assume that the error is uncorrelated with the input and has statistical properties consistent with a white process, z-transform techniques can be applied to show that the feedback architecture provides the following noise-shaping characteristics:

$$H_{\text{ns}}(z_1, z_2) \equiv \frac{\hat{E}(z_1, z_2)}{E(z_1, z_2)} = 1 - W(z_1, z_2) \qquad [2]$$

Here, $\hat{E}(z_1, z_2)$, $E(z_1, z_2)$, and $W(z_1, z_2)$ represent the z-transforms of $\hat{\varepsilon}_{m,n}$, $\varepsilon_{m,n}$, and $w_{m,n}$, respectively. Equation [2] shows that the noise-shaping characteristic of the error diffusion architecture is directly related to the spectral characteristics of $W(z_1, z_2)$. Appropriate selection of $W(z_1, z_2)$ can spectrally shape the quantizer noise in such a way that minimizes the effect of the low-resolution quantization process on the overall halftoning process. An error diffusion filter with low-pass spectral characteristics produces an overall noise-shaping characteristic $1 - W(z_1, z_2)$, which has the desired high-pass frequency characteristics. This noise-shaping function suppresses the quantization noise within those frequencies occupied by the image and spectrally shapes the noise to high frequencies, which are less objectionable to the human visual system. Circular symmetry is another important characteristic in the error diffusion filter since the human visual system is particularly sensitive to directional artifacts in the image.

In the following qualitative description of digital image halftoning, a 348 × 348 natural image of the Cadet Chapel at West Point was selected as the test image. This image provides a variety of important image characteristics, with regions of uniform grayscale, edges, and areas with fine detail. This image was scanned from a high-resolution black and white photograph at 150 dots per inch (dpi) and then printed using a 300 dpi laser printer.

Figure 3 shows the test image of the Cadet Chapel rendered using dispersed-dot ordered dither and results in 256 gray-levels at 300 dpi. Figure 4 shows the halftone image of the Cadet Chapel using the error diffusion algorithm and the Floyd–Steinberg filter coefficients shown in Figure 2a. The unidirectionality of the processing and the causality of the diffusion filter result in undesirable visual artifacts in the halftone image. These include directional hysteresis, which is manifested as 'snakes' running from upper-left to lower-right, and transient behavior near boundaries, which appears as 'shadows' below and to the right of sharp-intensity changes. The directional

$$\left(\tfrac{1}{16}\right)\times \begin{array}{ccc} & \bullet & 7 \\ 3 & 5 & 1 \end{array}$$

(a)

$$\left(\tfrac{1}{48}\right)\times \begin{array}{ccccc} & & \bullet & 7 & 5 \\ 3 & 5 & 7 & 5 & 3 \\ 1 & 3 & 5 & 3 & 1 \end{array}$$

(b)

$$\left(\tfrac{1}{42}\right)\times \begin{array}{ccccc} & & \bullet & 8 & 4 \\ 2 & 4 & 8 & 4 & 2 \\ 1 & 2 & 4 & 2 & 1 \end{array}$$

(c)

Figure 2 Three common error diffusion filters used in rectangular grid architectures. (The '•' represents the origin.)

hysteresis is particularly objectionable in uniform gray intensity areas such as the cloud structure in the upper-left of Figure 4. Similar artifacts are also present in images halftoned using the error diffusion algorithm and other causal diffusion kernels. The logical conclusion to draw from this limited, qualitative analysis is that if we could diffuse the error symmetrically and simultaneously process the entire image in parallel, we could reduce some of these visual artifacts and thereby improve overall halftone image quality.

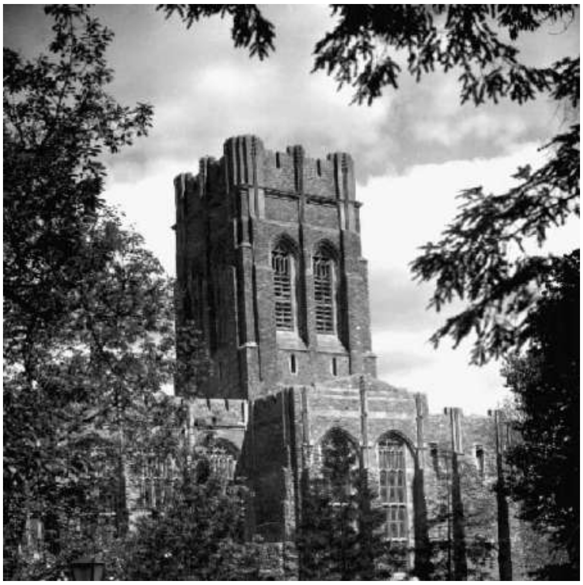

Figure 3 Original gray-scale image of the Cadet Chapel rendered using dispersed-dot ordered dither.

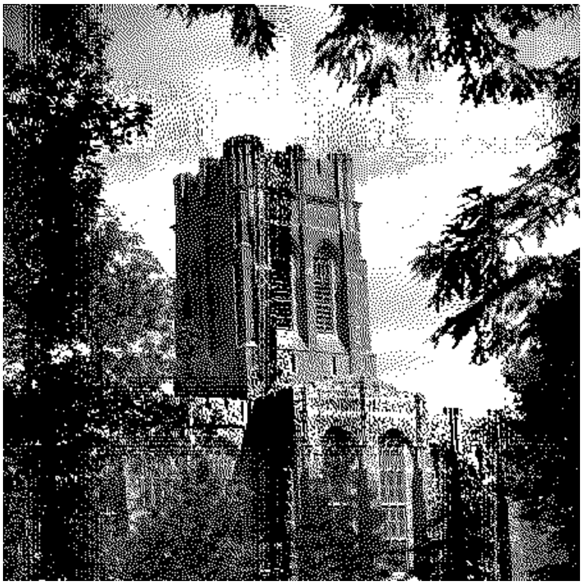

Figure 4 Halftone image of the Cadet Chapel produced using the error diffusion algorithm and the Floyd–Steinberg weights.

The Error Diffusion Algorithm and Neural Networks

The popularity of the neural network-based approach to signal and image processing lies in the ability to minimize a particular performance metric associated with a highly nonlinear system of equations. Specifically, the problem of creating a halftone image can be cast in terms of a nonlinear quadratic optimization problem where the performance measure to be minimized is the difference between the original and the halftone images.

The Hopfield-Type Neural Network

Tank and Hopfield first proposed a mathematical description of the functionality of the neural processing network for signal processing applications. Figure 5 shows an electronic implementation of a four-neuron architecture. Here, a single neuron is comprised of both a standard and inverting amplifier and the synapses or neural interconnections are represented by the physical connections between the input and output of the amplifiers. If the input to amplifier i is connected to the output of amplifier j by a resistor with resistance $R_{i,j}$, the amplitude of the connection $T_{i,j}$ is the conductance $1/R_{i,j}$.

The dynamic behavior of an N-neuron Hopfield-type neural network can be described by the following system of N nonlinear differential equations:

$$c\frac{\mathrm{d}u_i(t)}{\mathrm{d}t} = -u_i(t) + \sum_j T_{i,j}\Im[u_j(t)] + x_i \qquad [3]$$

where $i = 1, 2, \ldots N$, $\Im[\cdot]$ is a monotonically increasing sigmoid function, x_i is an input vector containing N

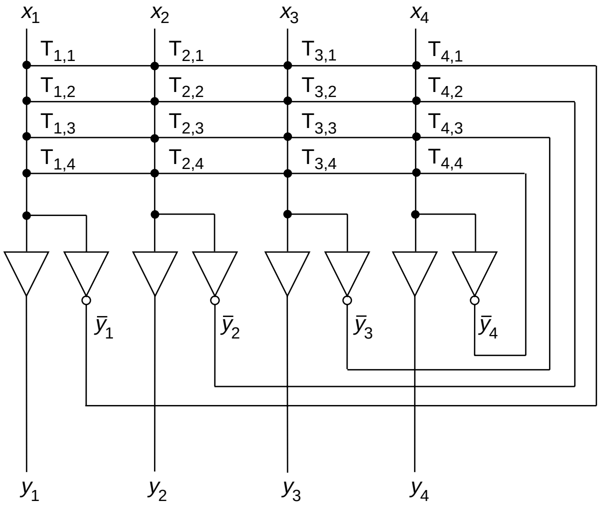

Figure 5 Electronic realization of a four-neuron Hopfield-type neural network.

elements, and c is a scaling factor. In equilibrium, eqn [3] implies that

$$u_i = x_i + \sum_j T_{i,j}\Im[u_j] \qquad [4]$$

Hopfield showed that when the matrix of interconnection weights $\mathbf{T}$ is symmetric with zero diagonal elements and the high-gain limit of the sigmoid $\Im[\cdot]$ is used, the stable states of the N functions $y_i(t) = \Im[u_i]$ are the local minima of the energy function:

$$E = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{T}\mathbf{y} - \mathbf{x}^{\mathrm{T}}\mathbf{y} \qquad [5]$$

where $\mathbf{y} \in \{-1, 1\}$ is the N-vector of quantized states. As a result, if the neural network can be shown to be stable, as the network converges the energy function is minimized. In most neural network applications, this energy function is designed to be proportional to a performance metric of interest and therefore as the network converges, the energy function and consequently the performance metric is also minimized.

The Error Diffusion Neural Network

An understanding of the operation and characteristics of the Hopfield-type neural network can be applied to the development and understanding of a 2D extension of the error diffusion algorithm. Figure 6 shows an electronic implementation of a four-neuron error diffusion-type neural network, where the individual neurons are represented as amplifiers and the synapses by the physical connections between the input and output of the amplifiers.

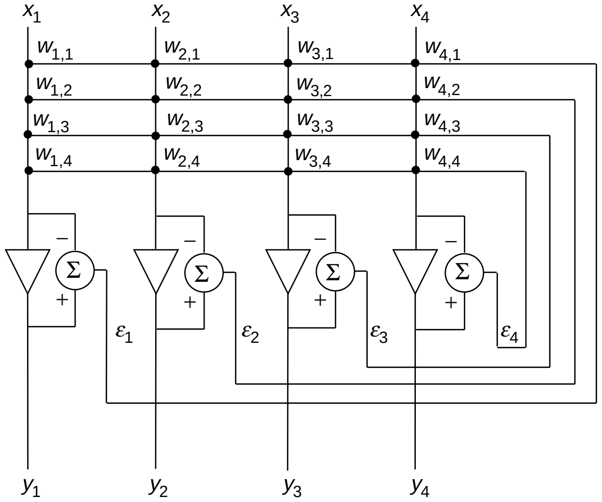

Figure 6 Four-neuron electronic implementation of the error diffusion neural network.

In equilibrium, the error diffusion neural network can be shown to satisfy

$$\mathbf{u} = \mathbf{W}(\mathbf{y} - \mathbf{u}) + \mathbf{x} \qquad [6]$$

Here, $\mathbf{u}$ is the state vector of neuron inputs, $\mathbf{W}$ is a matrix containing the interconnect weights, $\mathbf{x}$ is the input state vector, and $\mathbf{y}$ is the output state vector. For an $N \times N$ image, $\mathbf{W}$ is an $N^2 \times N^2$ sparse, circulant matrix derived from the original error diffusion weights $w_{m,n}$. If we define the coordinate system such that the central element of the error diffusion kernel is $(i, j) = (0, 0)$, then the matrix $\mathbf{W}$ is defined as $W(i, j) = -w[(j - i) \text{ div } N, (j - i) \bmod N]$ where:

$$x \text{ div } y = \left\lfloor \frac{x}{y} \right\rfloor \text{ if } xy \geq 0 \text{ or } \left\lceil \frac{x}{y} \right\rceil \text{ if } xy < 0 \qquad [7]$$

An equivalence to the Hopfield network can therefore be described by

$$\mathbf{u} = \mathbf{A}(\mathbf{W}\mathbf{y} + \mathbf{x}) \qquad [8]$$

where $\mathbf{A} = (\mathbf{I} + \mathbf{W})^{-1}$. Effectively, the error diffusion network includes a pre-filtering of the input image $\mathbf{x}$ by the matrix $\mathbf{A}$ while still filtering the output image $\mathbf{y}$ but now with a new matrix, $\mathbf{AW}$. Recognizing that $\mathbf{AW} = \mathbf{I} - \mathbf{A}$ and adding the arbitrary constant $k = \mathbf{y}^{\mathrm{T}}\mathbf{y} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, we can write the energy function of the error diffusion neural network as

$$\hat{E}(x, y) = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \qquad [9]$$

This energy function is a quadratic function, which can be factored into

$$\hat{E}(x, y) = [\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}]^{\mathrm{T}}[\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}] \qquad [10]$$

where $\mathbf{A} = \mathbf{B}^{\mathrm{T}}\mathbf{B}$. From eqn [10] we find that as the error diffusion neural network converges and the energy function is minimized, so too is the error between the output and input images.

If the neurons update independently, the convergence of the error diffusion network is guaranteed if

$$\forall k : [\mathbf{AW}]_{k,k} \geq 0 \qquad [11]$$

We find in practice, that even in a synchronous implementation, the halftoned images converge to a solution, which results in significantly improved halftone image quality over other similar halftoning algorithms.

The Error Diffusion Filter

The purpose of the error diffusion filter is to spectrally shape the quantization noise in such a way that the quantizer error is distributed to higher spatial frequencies, which are less objectionable to the human visual system. In this application, a feedback filter for the error diffusion neural network was designed using conventional 2D filter design techniques resulting in the 7 × 7 impulse response shown in Figure 7. Figure 8 shows the same image of the Cadet Chapel halftoned using the error diffusion neural network algorithm and the 2D error diffusion filter shown in Figure 7. There is clear improvement in halftone image quality over the image produced using sequential error diffusion in Figure 4. Notice particularly the uniform distribution of pixels in the cloud formation in the upper-left of Figure 8. Also noteworthy is the improvement around the fine detail portions of the tree branches and next to the vertical edges of the chapel.

0.0003	0.0019	0.0051	0.0068	0.0051	0.0019	0.0003
0.0019	0.0103	0.0248	0.0328	0.0248	0.0103	0.0019
0.0051	0.0248	0.0583	0.0766	0.0583	0.0248	0.0051
0.0068	0.0328	0.0766	•	0.0766	0.0328	0.0068
0.0051	0.0248	0.0583	0.0766	0.0583	0.0248	0.0051
0.0019	0.0103	0.0248	0.0328	0.0248	0.0103	0.0019
0.0003	0.0019	0.0051	0.0068	0.0051	0.0019	0.0003

Figure 7 Impulse response of a 7 × 7 symmetric error diffusion filter. ('•' represents the origin.)

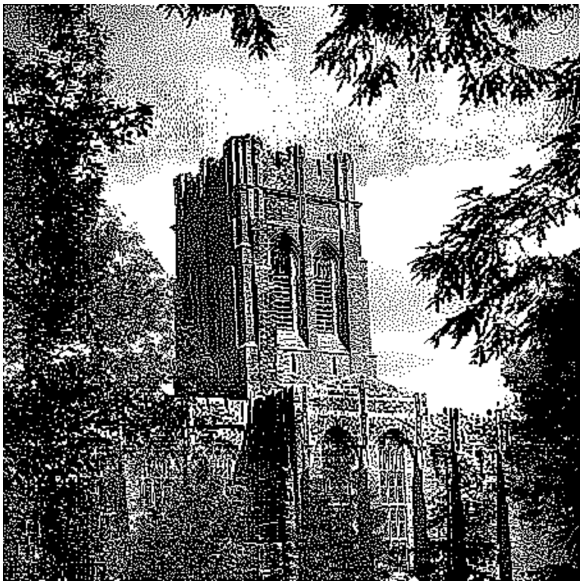

Figure 8 Halftone image of the Cadet Chapel using the error diffusion neural network and the 7 × 7 symmetric error diffusion filter shown in Figure 7.

Smart Pixel Technology and Optical Image Processing

The concept underlying smart pixel technology is to integrate electronic processing and individual optical devices on a common chip, to take advantage of the complexity of electronic processing circuits and the speed of optical devices. Arrays of these smart pixels then allow the advantage of parallelism that optics provides. There are a number of different approaches to smart pixel technology that differ primarily in the kind of opto-electronic devices, the type of electronic circuitry, and the method of integration of the two. Common opto-electronic devices found in smart pixels include: light-emitting diodes (LEDs); laser diodes (LDs); vertical cavity surface emitting lasers (VCSELs); multiple quantum well (MQW) modulators; liquid crystal devices (LCDs); and photodetectors (PDs). The most common types of electronic circuitry are silicon-based semiconductors such as complementary metal oxide semiconductors (CMOS) and compound semiconductors such as gallium arsinide (GaAs)-based circuitry. There are a number of different approaches to smart pixels, which generally differ in the way in which the electronic and optical devices are integrated. Monolithic integration, direct epitaxy, and hybrid integration are the three most common approaches in use today.

Smart pixel technology provides a natural methodology by which to implement optical image processing architectures. The opto-electronic devices provide a natural optical interface while the electronic circuitry provides the ability to perform either analog or digital computation. It is important to understand that any functionality that can be implemented in electronic circuitry can be integrated into a smart pixel architecture. The only limitations arise from physical space constraints imposed by the integration of the opto-electronic devices with the electronic circuitry. While smart pixels can be fabricated with digital or analog circuitry, the smart pixel architecture described subsequently uses mixed-signal circuitry.

A Smart Pixel Implementation of the Error Diffusion Neural Network

A smart pixel hardware implementation of the error diffusion neural network provides the potential to simultaneously achieve the computational complexity of electronic circuitry and the parallelism and high-speed switching of optics. The specific smart pixel architecture described here integrates MQW modulators called self electro-optic effect devices (SEEDs) with silicon CMOS

VLSI circuitry, using a hybrid integration approach called flip-chip bonding. This type of smart pixel is commonly referred to as CMOS-SEED smart pixel technology.

To provide an example of the application of smart pixel technology to digital image halftoning, a 5 × 5 CMOS-SEED smart pixel array was designed and fabricated. The CMOS circuitry was produced using a 0.5 μm silicon process and the SEED modulators were subsequently flip-chip bonded to silicon circuitry using a hybrid integration technique. The central neuron of this new smart pixel array consists of approximately 160 transistors while the complete 5 × 5 array accounts for over 3,600 transistors. A total of 50 optical input/output channels are provided in this implementation.

Figure 9 shows the circuitry associated with a single neuron of the smart pixel error diffusion neural network. All state variables in each circuit of this architecture are represented as currents. Beginning in the upper-left of the circuit and proceeding clockwise, the input optical signal incident on the SEED is continuous in intensity and represents the individual input analog pixel intensity. The input SEED at each neuron converts the optical signal to a photocurrent and subsequently, current mirrors are used to buffer and amplify the signal. The width-to-length ratio of the metal oxide semiconductor field effect transistors (MOSFETs) used in the CMOS-SEED circuitry provide current gain to amplify the photocurrent. The first circuit produces two output signals: $+I_u$ which represents the state variable $u_{m,n}$ as the input to the quantizer and $-I_e$ which represents the state variable $-\varepsilon_{m,n}$ as the input to the feedback differencing node. The function of the quantizer is to provide a smooth, continuous threshold functionality for the neuron that produces the output signal I_{out}, corresponding to the output state variable $y_{m,n}$. This second electronic circuit is called a modified wide-range transconductance amplifier and produces a hyperbolic tangent sigmoidal function when operated in the sub-threshold regime. The third circuit takes as its input I_{out}, produces a replica of the original signal, and drives

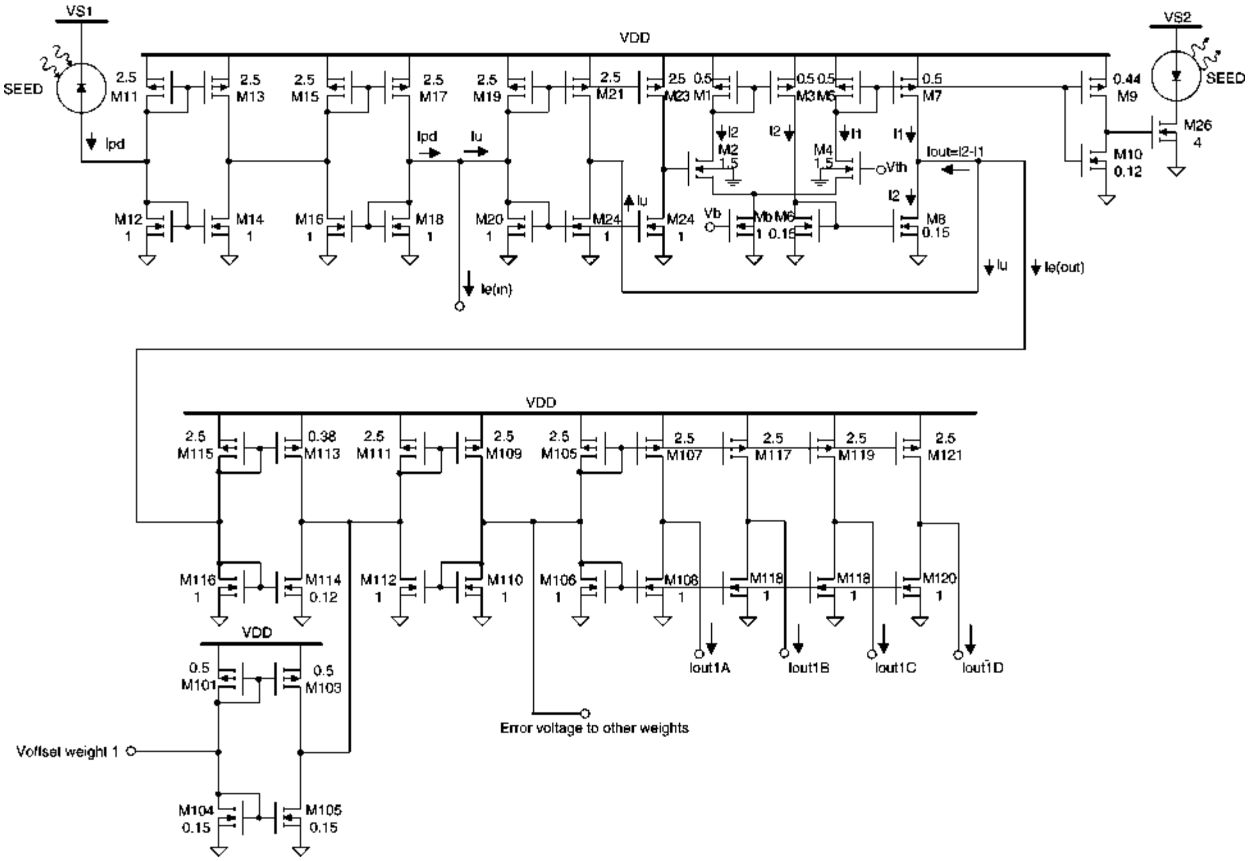
VS1
SEED
Ipd
VDD
M11
M12
M13
M14
M15
M16
M17
M18
M19
M20
M21
M23
M24
Ipd
Iu
Ie(in)
I2
I1
M1
M2
M3
M4
M5
M6
M7
M8
Vth
Vb
Iout=I2-I1
Iu
Ie(out)
VS2
SEED
M9
M10
M26
VDD
M101
M103
M104
M105
M106
M107
M108
M109
M110
M111
M112
M113
M114
M115
M116
M117
M118
M119
M120
M121
Voffset weight 1
Error voltage to other weights
Iout1A
Iout1B
Iout1C
Iout1D

Figure 9 Circuit diagram of a single neuron and a single error weight of the 5 × 5 error diffusion neural network based on a CMOS-SEED-type smart pixel architecture. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.6, p. 222. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

the output optical SEED. In this case, the output optical signal is a binary quantity represented as the presence or absence of light. When the SEED is forward-biased, light is generated through electroluminescence. The last circuit at the bottom of the schematic implements a portion of the error weighting and distribution function of the error diffusion filter. The individual weights are implemented by scaling the width-to-length ratio of the MOSFETs to achieve the desired weighting coefficients. The neuron-to-neuron interconnections are accomplished using the four metalization layers of the 0.5 μm silicon CMOS process. In this design, the error diffusion filter was limited to a 5 × 5 region of support because of the physical constraints of the circuitry necessary to implement the complete error diffusion neural network in silicon circuitry. The impulse response of the 5 × 5 filter used in this particular smart pixel architecture is shown in Figure 10. The error weighting circuitry at the bottom of Figure 9 represents only the largest weight (0.1124) with interconnects to its four local neighbors ($I_{\text{outA}} - I_{\text{outD}}$).

0.0023	0.0185	0.0758	0.0185	0.0023
0.0185	0.0254	0.1124	0.0254	0.0185
0.0758	0.1124	●	0.1124	0.0758
0.0185	0.0254	0.1124	0.0254	0.0185
0.0023	0.0185	0.0758	0.0185	0.0023

Figure 10 Error diffusion filter coefficients used in the 5 × 5 CMOS-SEED error diffusion architecture. ('•' represents the origin.)

Figure 11 shows a photomicrograph of a 5 × 5 error diffusion neural network, which was fabricated using CMOS-SEED smart pixels. Figure 12 shows a photomicrograph of a single neuron of the 5 × 5 CMOS-SEED neural network. The rectangular features are the MQW modulators while the silicon circuits are visible between and beneath the modulators. The MQW modulators are approximately 70 μm × 30 μm and have optical windows, which are 18 μm × 18 μm.

Simulations of this network predict individual neuron switching speeds of less than 1 μs, which corresponds to network convergence speeds capable of providing real-time digital image halftoning. Individual component functionality and dynamic operation of the full 5 × 5 neural array were both experimentally characterized. Figure 13 shows CCD images of the operational 5 × 5 CMOS-SEED smart pixel array. The white spots show SEED MQW modulators emitting light. Figure 13a shows the fully functioning array while Figure 13b shows the 5 × 5 CMOS-SEED neural array under 50% gray-scale input. Here 50% of the SEED modulators are shown to be in the *on*-state demonstrating correct operation of the network and the halftoning architecture.

Both the simulations and the experimental results demonstrate that this approach to a smart pixel implementation of the error diffusion neural network provides sufficient accuracy for the digital halftoning application. The individual neuron switching speeds also demonstrate the capability for this smart pixel hardware implementation to provide real-time halftoning of video images.

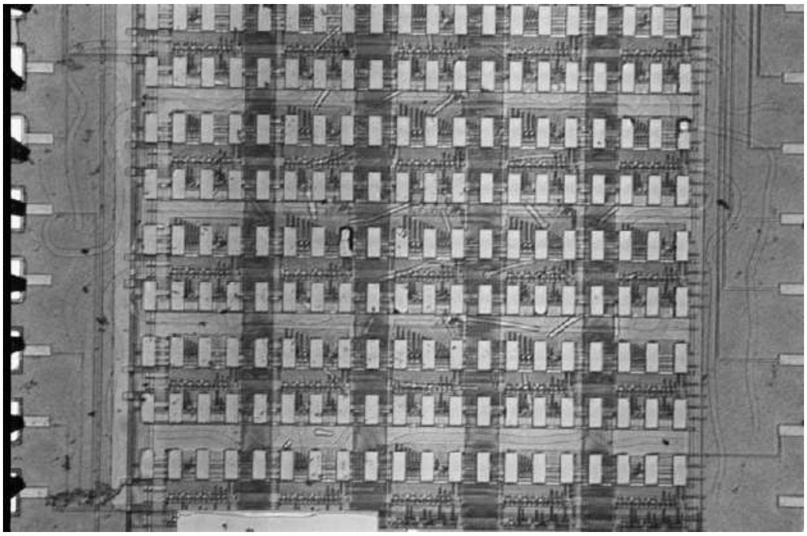

Figure 11 Photomicrograph of a smart pixel implementation of a 5 × 5 CMOS-SEED error diffusion neural network for digital image halftoning. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.5, p. 221. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

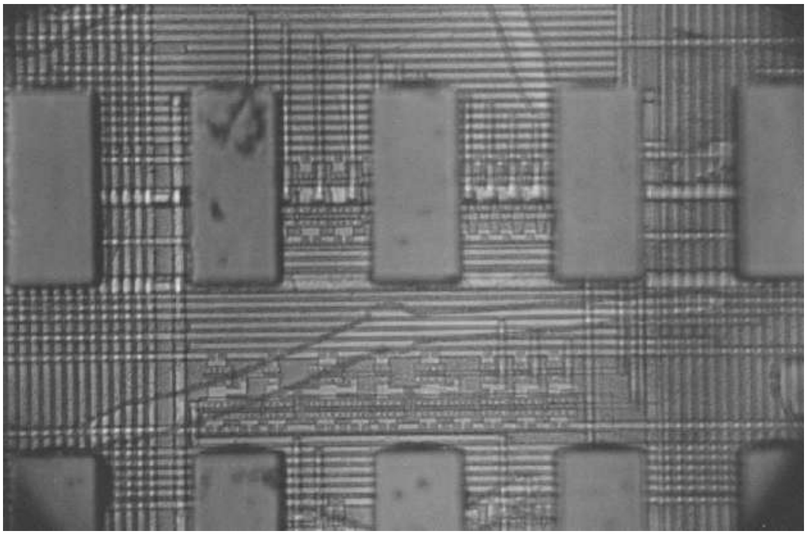

Figure 12 Photomicrograph of a single neuron of the 5 × 5 error diffusion neural network. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.4, p. 220. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

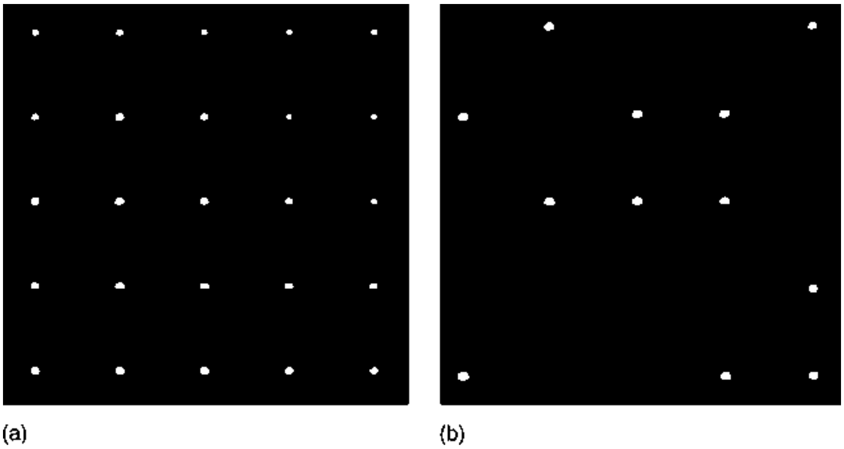

Figure 13 CCD images of the 5 × 5 CMOS-SEED array. (a) Fully-operational array, and (b) under 50% gray scale illumination. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Figs. 8.19 and 8.20, p. 236. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

Image Processing Extensions

Other image processing functionality is also possible by extending the fundamental concepts of 2D, symmetric error diffusion to other promising image processing applications. One important application includes color halftoning while other extensions include edge enhancement and feature extraction. In the basic error diffusion neural network, the error diffusion filter was specifically designed to produce visually pleasing halftoned images. Care was taken to ensure that the frequency response of the filter was circularly symmetric and that the cutoff frequency was chosen in such a way as to preserve image content. An analysis of the error diffusion neural network shows that the frequency response of this filter directly shapes the frequency spectrum of the output halftone image and therefore directly controls halftone image content. Other filter designs with different spectral responses can provide other image processing features such as edge enhancement, which could lead to feature extraction and automatic target recognition applications.

See also

Detection: Smart Pixel Arrays. **Information Processing:** Optical Neural Networks. **Optical Processing Systems**.

Further Reading

Dudgeon DE and Mersereau RM (1984) *Multidimensional Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Lau DL and Arce GR (2001) *Modern Digital Halftoning*. New York: Marcel Dekker.

Lim JS (1990) *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ: Prentice Hall.

Mead CA (1989) *Analog VLSI and Neural Systems*. New York: Addison-Wesley.

Poon T-C and Banerjee PP (2001) *Contemporary Optical Image Processing with Matlab*. New York: Elsevier.

Russ JC (1998) *The Image Processing Handbook*. Boca Raton, FL: CRC Press.

Shoop BL (2001) *Photonic Analog-to-Digital Conversion*. Springer Series in Applied Optics, vol. 81. Berlin: Springer-Verlag.

Shoop BL, Sayles AH and Litynski DM (2002) New devices for optoelectronics: smart pixels. In: De Cusatis (ed.) *Fiber Optic Data Communication: Technological Trends and Advances*, pp. 352–421. San Diego, CA: Academic Press.

Ulichney R (1987) *Digital Halftoning*. Cambridge, MA: The MIT Press.

Wax C (1990) *The Mezzotint: History and Technique*. New York: Harry N. Abrahms, Inc.

Optical Neural Networks

H J Caulfield, Fisk University, Nashville, TN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

An artificial neural network (ANN) is a nonlinear signal processing system based on the neural processes observed in animals. Usually they have multiple inputs and often multiple outputs also. Conventionally, each input sends its signal to many neurons, and each neuron receives signals from many inputs. The neuron forms an intermediate sum of weighted inputs and transforms that sum, according to some nonlinearity, to form an output signal. Often the output signal from one neuron is used as the input signal for another in feed forward, feedback, or mixed mode complex neural networks.

ANNs are of value in helping us understand how biological neural networks perform and in a variety of technical applications, most of which involve pattern association. An auto-associative neural network, that associates patterns with themselves, can have useful noise cleanup or pattern completion properties. A hetero-associative neural network is often used in pattern recognition to associate input patterns with preselected class labels.

Neural Networks – Natural and Artificial

The most primitive one-celled organisms have no nerve cells, but they do exhibit various chemical means for interacting with their environments, within their own cells, and among individuals of the same or different types. As organisms increased in size, a common chemical environment became impractical, and nerves arose as a means for long-distance communication within larger organisms. One thing seen in evolution is that widespread useful adaptations are often modified but seldom discarded. Thus, it should not be surprising to find that chemical interactions play a major role in the most complex thing we know of – the human brain. Chemicals play a huge role in making us who we are, setting our moods, and so forth. Realizing this, in the late twentieth century, psychiatrists started to make serious progress in treating mental diseases through the use of chemistry.

There are a variety of processes that we use as humans, but the primary means is neural. Only some animals and no plants have neurons, but soon after neurons evolved, they began to concentrate in special regions of the body that developed into primitive brains. Not only were there many neurons, but also they were densely and complexly interconnected to form neural networks.

Neural networks did not replace chemical interactions, they used and supplemented them. Chemical communication across the synaptic cleft dominates communication between networks, but the neural networks also control chemical production and distribution within the brain, and those chemicals change the behavior of the neurons that cause their secretion. The functioning of real brains is incredibly complex, possibly irreducibly complex in terms of current mathematics and logic.

These things are noted to avoid a confusion that is distressingly common. Technologists produce simplified ANNs and then try to interpret the behavior of brains in those terms. Quite often, ANNs yield useful insight into brain behavior but seldom offer a detailed account of them. Reality is far more complex than any ANN. Current mathematics provides no means to describe continuous sets of events, each of which both causes and is caused by the others.

ANNs themselves have grown much more complicated as well. Starting out as simple, feed-forward, single-layer systems, they rapidly became bidirectional and even multilayered as well. Higher-order neural networks processed the input signals before inserting them into the neural network. Pulse coupled neural networks (PCNNs) arose that more closely modeled the pulsations in real brains. Learning was broken into learning with instruction, following example, and self-organizing, the latter leading to self organized maps and adaptive resonance theory, and many related subjects. These are but crude pointers to the wealth and variety of concepts now part of the ANN field.

The history of ANNs is both complicated and somewhat sordid. This is not the occasion to retrace the unpleasant things various members of this community have done to others in their quests for stature in the field. One result of that history is a dramatically up-and-down pattern of interest in and support for ANNs. The first boom followed the development of a powerful single-layer neural network called the Perceptron by its inventor Rosenblatt. Perceptrons were biologically motivated, fast, and reasonably effective; so interest in the new field of ANNs was high. That interest was destroyed by a book on the field arguing that Perceptrons were only linear discriminants and most interesting problems are not linearly discriminable. Ironically, the same argument has been made about optical Fourier correlators, and it is true there as well. But optical processing folk largely ignored the problem and kept working on it. Sociologists of science might enjoy asking why reactions differed so much in the two fields. Funding and interest collapsed and the field went into hibernation.

It was widely recognized that multilayer ANNs would comprise nonlinear discriminants, but training them presented a real problem. The simple rewards and punishment of weights depending on performance was easy when each output signal came from a known set of weighted signals that could be rewarded or punished jointly, according to how the output differed from the sought-after output. With a multilayer perceptron, credit assignment of inner layers is no longer simple. It was not until Werbos invented 'backward error propagation' that a solution to this problem was available and the field started to boom again. Also key to the second boom were enormously popular articles by Hopfield on rather simple auto- and hetero-associative neural networks. Basically, those networks are almost never used, but they were of great value in restarting the field. In terms of research and applications, we are still in that second boom, that began during the 1980s.

> An ANN is a signal processor comprised of a multiplicity of nodes which receive one or more input signals and generate one or more output signals in a nonlinear fashion wherein those nodes are connected with one or two way signaling channels.

Note also that biological inspiration is no longer a part of the definition. Most ANNs used for association are only vaguely related to their biological counterparts.

Optical Processors

Optics is well-suited for massive parallel interconnections, so it seems logical to explore optical ANNs. The first question to address is "Why bother to use optical processors at all?" as electronic digital computers seem to have everything:

- Between the writing of this article (on an electronic computer) and its publication, the speed of electronic computers will almost double and they will become even cheaper and faster.
- They can have arbitrary dynamic range and accuracy.
- Massively improved algorithms also improve their speed. For instance, the wonderful fast Fourier transform (FFT) that opticists have used to simulate optics for decades has been replaced by an even faster fastest Fourier transform in the West (FFTW). And for wavelets, the fast wavelet transform (FWT) is even faster.

Optical processors, on the other hand, have severely limited analog accuracy and few analog algotectures or archirithms. They are specialized systems (equivalent in electronics, not to computers but to application specific integrated circuits (ASICs)). They are usually larger and clumsier than electronic systems and always cost more.

There are several reasons why optics may still have a role in some computations. First, many things on which we wish to perform computations are inherently optical. If we can do the computations optically before detection, we can avoid some time and energy and sensitivity penalties inflicted, if we first detect the signals then process them electronically, and then (sometimes) convert them back to optics. Examples include spectroscopy and optical communication.

Second, pure optics is pure quantum mechanics. Input of data and setup of the apparatus is experiment preparation. Readout of data is classical measurement. Nature does the rest free of charge. There is a unitary transformation from input to output that requires no human intervention and is, in fact, destroyed by human intervention. Those intermediate (virtual) computations require no energy and no time, beyond the time taken for light to propagate through the system. Speed and power consumption are limited only by the input and output. This applies for everything from Fourier optics to digital optics to logic gates. Thus, optics can have speed and power consumption advantages over electronics.

Third, if quantum computers involving entanglement ever emerge, they may well be optical. Such computers would be used for nondeterministic solutions of hard problems. That is, they would explore all possible paths at once and see which one does best. This is something we can already do electronically or optically for maze running (using classical, not quantum entanglement) means.

Fourth, most optical processors are analog, which is sometimes superior to digital. In those cases, we would still have to choose between analog optics and analog electronics, of course.

So, optical processing (like all other processing) is a niche market. But some of the niches may be quite large and important. It is not a universal solution, but neither is digital electronics.

The Major Branches of Optical Processing and Their Corresponding Neural Networks

Every branch of optical processing leads to its own special purpose optical neural networks, so those branches become a good place to start an article such as this one.

Optical Linear Algebra

Many neural network paradigms use linear algebra followed by a point-by-point nonlinearity in each of multiple stages. The linear operation is a matrix–vector multiplication, a matrix–matrix multiplication, or even a matrix–tensor multiplication. In any case, the key operations are multiplication and addition – operations very amenable to performance with analog optics. With effort, it is possible to combine simple electronics with analog optics operating on non-negative signals (the amount of light emitted from a source, the fraction of light passing through a modulator, etc.) to expand to real or even complex numbers. With even more effort, we can use multiple analog non-negative signals to encrypt a digital signal, allowing digital optical processors. Almost all optical neural networks using linear algebra use only real numbers. In most of these, the real number is encoded as a positive number and a negative number. Data are read in parallel using spatial light modulators (SLMs), source arrays, acousto-optic modulators, etc. They are operated upon by other SLMs, acousto-optic systems, and so forth. Finally, they are detected by individual detectors or detector arrays. The required nonlinearities are then done electronically. It is a theory that holds for all linear algebraic processes up to computational complexity $O(N^4)$ that they can be performed with temporal complexity $O(N^0)$ if the extra dimensions of complexity are absorbed in space, up to $O(N^2)$ through 2D arraying and in fanin/fanout, the ability of multiple beams of light to be operated upon in parallel by the same physical modulator – this has the same complexity as the 2D array of modulators, namely $O(N^2)$. Thus, we can invert a matrix, an order $O(N^4)$ for an $N \times N$ matrix, $O(N^0)$ in time, that is, independently of the size of N, as long as that matrix is accommodated by the apparatus. This offers a speed electronics cannot match.

Most of the optical neural networks of the mid-1980s had optical vector–matrix multipliers at their heart. Most were a few layers of feed forward systems, but by the 1990s feedback had been incorporated as well. These are typically N inputs connected to N outputs through an $N \times N$ matrix, with N varying from 100 to 1000. With diffuser interconnection and a pair of SLMs, it is possible to connect an $N \times N$ input with an $N \times N$ output using N^4 arbitrary weights, but that requires integrating over a time up to N^2 intervals. Most of the N^4-dimensional matrices can be approximated well by fewer than N^4 terms using singular value decomposition (SVD) and (as this version of the interconnection uses outer products), that reduces the number of integration times considerably in most cases. We have shown, for instance, that recognizing M target images can be accomplished well with only M terms in the SVD.

Coherent Optical Fourier Transformation

It took opticists several years, during the early 1960s, to realize that if Fourier mathematics was useful in describing optics, optics might be useful in performing Fourier mathematics. The first applications of optical Fourier transforms were in image filtering and particularly in pattern recognition. Such processors are attractive not only for their speed but also for their ability to locate the recognized pattern in 1D, 2D, or even 3D. It was soon obvious that a Fourier optical correlator was a special case of a single layer perceptron. Of course, it inherits the weaknesses of a single layer perceptron too. Soon, workers had added feedback to create a winner-takes-all type of auto-associative optical neural network, a system that associated images with themselves. This proved useful in restoring partial or noisy images. It was not until 2003 that we learned how to make powerful hetero-associative networks with Fourier optical systems. This creates a nonlinear discrimination system that maintains the target location and generalizes well from a few samples of multiple classes to classify new data quite accurately.

Imaging

Imaging is the central act of optics and might seem to be of little value in optical processing. But multiple imaging (with lenslets, self-focusing fibers, holograms, and the like) has been the basis of numerous optical neural networks. Multiple imaging is a nonrandom way of distributing information from one plane to the next. Diffusers have also been used to produce fully interconnected optical neural networks. If each broad image point contains the whole image (something easily arranged with a diffuse hologram), then we have an auto-associative memory. Different such auto-associative networks can be superimposed in a single hologram. Then feedback after nonlinear image filtering can be used to make a general auto-associative memory.

Even deliberately poor imaging has been used. A blurred or out-of-focus image can accomplish the lateral signal spreading needed to facilitate a PCNN. These networks are substantially more like real brain networks than those discussed so far. It starts with a 2D array of integrate-and-fire units. The firing rate of each depends only in the incident power (how long integration to the thresholding firing takes). But if the fired signal is spread to its neighbors, the firing of a neighbor can hasten any unit's firing. This causes firings to synchronize and results in 'autowaves' of firing patterns moving across the 2D plane. Thus an incident 2D image is converted to a 3D signal – the two spatial dimensions and time. This results, for all but some very simple cases, in a chaotic signal occupying a strange attractor. It is interesting but not very revealing in itself. But suppose we 'integrate out' either the two spatial dimensions or the time dimension. If we integrate out the spatial dimensions, we are left with a time signal. Its strange attractors are remarkably useful in pattern recognition, because they are more syntactic than statistical (in pattern recognition terms). They describe the shape as similar, independently of the size, rotation, perspective, or even shading of the pattern! If we integrate out the time dimension, we restore the original image – almost. In fact, we can use this aspect of PCNNs for optimum blind segmentation of images. The autowaves can be made to 'leave tracks' so they can be reversed, which allows them to solve maze problems nondeterministically, because they can take all the paths. When the first wave exits, we simply retrace it to find the shortest path through the maze – a prototypical nondeterministic polynomial operation.

Holograms

Holograms are the most general, complex, and flexible means of light manipulation so far invented. Naturally, they have many applications in optical neural networks. One such application is the interconnection of each of an $N \times N$ array of inputs to each of an $N \times N$ array of outputs with its own unique connection strength (weight in the language of neural networks). The practical limit seems to be around $N = 1000$, yielding 10^{12} weights, roughly the same number as in the human brain! The brain, however, is not fully interconnected, so this produces an entirely different type of neural network. The resulting network would have fixed weights, but it could perform a variety of useful operations and be both forward and backward connected.

Special on-axis holograms, called diffractive optical elements, can be mass manufactured very effectively and inexpensively to perform tasks such as multiple imaging.

Photorefractive Neural Networks

A very complex thing happens when a pattern of fringes, such as those that form an interference pattern for hologram recording, strike materials such as lithium niobate, strontium barium niobate, gallium arsenide, and many others that fall into this category of 'photorefractive' materials. Of course, 'photo' refers to light and 'refractive' refers to the speed of light in a material – the speed of light in a vacuum divided by the index of refraction. These materials are photoconductive, so the light releases charge carriers in what would otherwise be an electrical insulator. The periodic charge pattern sets up a periodic space charge within the material. But the material is also electro-optic, that is, its index of refraction changes with applied electric field, thus generating a periodic index of refraction grating in the material. Finally, that index of refraction pattern (a phase hologram) diffracts the incoming light. Usually, but not always, a steady-state situation is reached after time that depends on the material and the photon density, although chaos can arise under some conditions. This makes for a dynamic hologram, one that can change or adapt over time as the input pattern changes. Once a satisfactory photorefractive hologram is formed, it can be stabilized (fixed) if desired. Most optical neural network uses envision a dynamic situation with changing interconnections.

What has not been discussed above is that the index of refraction varies with the light's electric field vector cubed, so photorefractives provide a convenient way to accomplish such nonlinear operations as winner-takes-all.

Sometimes, we can record holograms with photorefractives that cannot be recorded by conventional holograms in conventional materials. To explain this, a hologram is a transducer between two optical wavefronts – usually called the reference and object beams. If the reference beam is incident on the hologram, it is (fully or partially) converted into the object beam. If the precisely reversed (phase conjugated) reference beam is incident on it, the precisely reversed (phase conjugated) object beam is derived from it. Likewise, the object wavefront (or its phase conjugate) produces the reference wavefront (or its phase conjugate). The hologram is made by recording the interference pattern between the two wavefronts: the reference and object beams. The ability of two wavefronts to form a temporally stable interference pattern at some point in space is called the mutual coherence of those wavefronts. Usually, both wavefronts are derived from the same laser beam to make achieving high mutual coherence possible. And, of course, both beams must be present to form an interference pattern. In some cases of photorefractive holograms, all of those conditions can be violated. The two wavefronts can derive from two sources that are different in wavelength, polarization, etc., and need not be simultaneously present. This phenomenon, called mutual phase conjugation, converts each wavefront into the phase conjugate of the other, but for effects due to any differences in wavelength and polarization. This allows new freedoms that have been exploited in optical neural networks.

Photorefractive materials also tend to be piezoelectric. It has been found that up to about 20 totally independent reflective holograms can be stored with different applied electric field in lithium niobate. Changing electric fields so changes the fringe spacings and indices of refraction, that only one of those holograms has any appreciable diffraction efficiency at any one electric field value. The result is an electrically selectable interconnection pattern. The same hologram can contain the information needed for many neural networks or multiple layers of fewer of them. There is no need, in general, to have more than an input layer, an output layer, and two intermediate or hidden layers.

Others have used photorefractive material to implement learning in neural networks. In biological neural networks the synaptic strengths (the weights between neurons) continue to adapt. With photorefractive holograms, we can achieve this in an optical ANN.

As previously discussed, a key development in the history of ANNs was a way to train multilayer neural networks and thus accomplish nonlinear discrimination. The first, and still most popular, way to do this is called backward error propagation. Several groups have been implemented in photorefractive optical neural networks.

Conclusions

Essentially every major type of neural network has been implemented optically. Every major tool of optics has found its way into these optical neural networks. Yet, it remains the case that almost every neural network used for practical applications is electronic. The mere existence proof of a use for optical neural networks may be what the field needs to move forward from the demonstration stage to the application stage. The tools exist and there are niches where optics seems more appropriate than electronics. Hopefully, the next review of the field will include a number of commercial successes. Nothing spurs science better than the funds that follow successful applications.

List of Units and Nomenclature

Adaptive resonance theory (ART)	Another self-organized neural net categorizer that works by mutual adjustment between bottom-up and top-down neural connections.
Computational complexity	The measure of how the number of required calculations scales with the problem size, usually denoted by N. The scaling is represented by the Big O form of (N), where $f(n)$ shows how it scales asymptotically (large N). Problems that scale as $O(N^P)$ are said to be of polynomial complexity. Unfortunately, some of the most important problems have exponential complexity.

Fanin/Fanout	Beams of light (unlike beams of electrons or holes) do not interfere with one another, so many of them can be modulated by the same physical elements quite independently, so long as they can be identified. This is usually done by sending the different beams in at different angles, giving rise to the words fanin and fanout, as suggested by the drawing below: 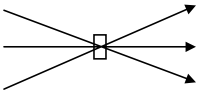
Self organized maps (Kohonen networks)	A means of self-organization of inputs into (normally 2D) maps wherein every concept falls into some category and the categories themselves develop out of application of the Kohonen's algorithm and user-selected parameters applied to the given data.
Singular value decomposition (SVD)	Any matrix can be broken into (decomposed) a weighted sum of simpler, rank-one matrices. The simplest such decomposition (in the sense of fewest number of terms) is the SVD. The weights are the singular values and are indicators of the importance of that term in the decomposition.
Spatial light modulator (SLM)	A 1D or 2D array of elements under electronic or optical control that change some property of the light incident on them. Usually, it is the index of refraction in some direction that is changed. That change can be used to modulate the amplitude, the polarization, or the phase of the light.

See also

Fourier Optics. **Holography, Techniques:** Overview. **Nonlinear Optics, Basics:** Photorefraction. **Optical Communication Systems:** Local Area Networks. **Optical Processing Systems**.

Further Reading

Abu-Mostafa YS and Psaltis D (1987) Optical neural computers. *Scientific American* 256(3): 88–95.

Anderson DZ and Erie MC (1987) Resonator memories and optical novelty filters. *Optical Engineering* 26: 434–444.

Brady D, Psaltis D and Wagner K (1988) Adaptive optical networks using photorefractive crystals. *Applied Optics* 27: 1752–1759.

Caulfield HJ, Kinser J and Rogers SJ (1987) Optical neural networks. *Proc. IEEE* 77: 1573–1583.

Chavel P, Lalanne P and Taboury J (1989) Optical inner-product implementation of neural network models. *Applied Optics* 28: 377–385.

Collins SA, Ahalt SC, Krishnamurthy AK and Stewart DF (1993) Optical implementation and applications of closest-vector selection in neural networks. *Applied Optics* 32: 1297–1303.

Farhat NH, Psaltis D, Prata A and Paek E (1985) Optical implementation of the Hopfield model. *Applied Optics* 24: 1469–1475.

Galstyan T, Frauel Y, Pauliat G, Villing A and Roosen G (1997) Topological map from a photorefractive self-organizing neural network. *Optics Communications* 135: 179–188.

Goodman JW, Dias AR and Woody JM (1978) Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Optics Letters* 2: 1–3.

Johnson JL (1994) Pulse-coupled neural nets: Translation, rotation, scale, distortion, and intensity signal invariance for images. *Applied Optics* 33: 6239–6253.

Jutamulia S (1994) *Selected Papers on Optical Neural Networks*. Bellingham: SPIE Press.

Kinser JM, Caulfield HJ and Shamir J (1988) Design for a massive all-optical bidirectional associative memory: the big BAM. *Applied Optics* 27: 3442–3444.

Neifeld MA and Psaltis D (1993) Optical implementations of radial basis classifiers. *Applied Optics* 32: 1370–1379.

Owechko Y, Dunning GJ, Marom E and Soffer BH (1987) Holographic associative memory with nonlinearities in the correlation domain. *Applied Optics* 26: 1900–1910.

Paek EG and Psaltis D (1987) Optical associative memory using Fourier-transform holograms. *Optical Engineering* 26: 428–433.

Psaltis D and Farhat NH (1985) Optical information processing based on associative-memory model of neural nets with thresholding and feedback. *Optics Letters* 10: 98–100.

Soffer BH, Dunning GJ, Owechko Y and Marom E (1986) Associative holographic memory with feedback using phase-conjugate mirrors. *Optics Letters* 11(2): 118–120.

Yu FTS (1993) Optical neural networks – architecture, design and models. In: Wolf E (ed.) *Progress in Optics*, II, pp. 63–144. Amsterdam: North-Holland.

INSTRUMENTATION

Contents

Astronomical Instrumentation
Ellipsometry
Photometry
Scatterometry
Spectrometers
Telescopes

Astronomical Instrumentation

J Allington-Smith, University of Durham, Durham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article describes the requirements placed on astronomical instruments and the methods used to achieve them, making reference to generic techniques described elsewhere in this encyclopedia. It is limited in scope to electromagnetic radiation with wavelengths from the atmospheric cutoff in the ultraviolet (0.3 μm) to the point in the near-infrared where the technology changes radically (5 μm), taking in the transition from uncooled to cooled instruments at ~1.7 μm. Thus, it excludes the extreme UV and X-ray regimes, where standard optical techniques become ineffective (e.g., requiring grazing incidence optics) and all mid-infrared, sub-mm, THz, and radio-frequency instruments. Terrestrial telescopes are the main focus but space-borne telescopes are subject to similar design considerations, except that the diffraction limit of the optics may be achieved at short wavelengths. Adaptive optics (AO), of increasing importance in astronomy and almost essential for the next generation of extremely large telescopes (ELTs), is covered elsewhere in this encyclopedia.

Astronomical Data

The purpose of astronomical instruments is to measure the following properties of photons arriving on or near to the Earth and brought to a focus by an astronomical telescope:

(i) direction of arrival;
(ii) energy;
(iii) polarization state; and
(iv) time of arrival.

In practice, these properties cannot be measured simultaneously, requiring the following generic classes of instrument:

- *Imagers* to measure (i) and, if equipped with ultra-low-noise detectors, (iv);
- *Spectrographs* to measure (ii), although some can be configured to do a good job on (i) as well. *Integral field spectrographs* can measure (i) and (ii) simultaneously; and
- *Polarimeters* for (iii) plus (i) and *spectropolarimeters for* (iii) plus (ii).

In principle, energy-resolving detectors, such as arrays of superconducting tunnel-junction devices (STJs) can replace (i), (ii), and (iv).

These must interface to the telescope which provides a focal surface with an image scale which is approximately constant over the field. Currently, almost all telescopes are based on the Cassegrain or Ritchy-Chretien reflecting configurations, comprising a large curved primary mirror (M1) which defines the collecting aperture (diameter D_T) and a curved secondary mirror (M2). This feeds the Cassegrain focus via a hole in M1 in line with the central obstruction caused by M2, or the Nasmyth focus if a tertiary (usually plane) mirror is added to direct light from M2 along the elevation axis to a location in which the gravity vector changes only in rotation about the optical axis. For the Cassegrain focus, the full range of change in the gravity vector is experienced as the telescope tracks objects during long exposures or slews between targets. It is also possible to dispense with M2 and mount the

Table 1 Optical interfaces to representative types of telescope and achievable angular resolution

Aperture diameter D_T (m)	*Primary focus*			*Final focus*			*Diffraction limit $2.44\lambda/D_T$(mas)*	
	F/#	*Image scale (μm/″)*	*FOV (′)*	*F/#*	*Image scale (μm/″)*	*FOV (′)*	*$\lambda = 0.5\ \mu$m*	*$\lambda = 5.0\ \mu$m*
2.4^{i}	2.3			24	279	25	101	1007
4.2^{ii}	2.8	60	40	11	221	15	60	600
8.0^{iii}	1.8	(70)	(45)	16	621	10	32	315
30^{iv}	1–1.5	~180	~20	15–19	~2500	~20	8.4	84
100^{v}	~1.4			~6.5	~3000	~10	2.5	25
			For enclosed telescopes				500–1500	
All	Obtainable angular resolution for terrestrial telescopes		Free-atmosphere seeing at good site		Best		250	160
					Median		420	270
			With adaptive optics (approx)				50–200	*Diff. limit*^{a}

^{a}The diffraction limit listed in the column above for $\lambda = 5.0\ \mu$m applies.

Exemplars: (i) Hubble Space Telescope; (ii) Herschel Telescope, La Palma; (iii) Gemini telescopes, Hawaii/Chile; (iv) CELT/GSMT (USA); (v) OWL (Europe).

instrument directly at the focus of M1 (prime focus), although this imposes severe restrictions on the mass and size of the instrument. An alternative mounting for instruments is via optical fibers. The input, with any fore-optics required to couple the telescope beam into the fiber, is mounted at the appropriate focus while the output, possibly several tens of meters distant, is mounted where the instrument is not subjected to variations in the gravity vector.

The useful field of view (FOV) of the telescope is determined by how rapidly the image quality degrades with radius in the field; restrictions on the size of the Cassegrain hole and the space set aside for the star trackers required for accurate guiding and auxiliary optics involved in target acquisition, AO, etc.

Typical telescope parameters are shown in Table 1, of which the last two entries are speculative examples of ELTs. The table also lists the diffraction limit of the telescopes and the limits imposed by seeing, with and without the effects of the telescope structure and enclosure, and approximate limits for performance with AO which approaches the diffraction limit in the infrared.

Instrument Requirements

Efficiency

Almost all astronomical observations are of very faint objects or require very high signal-to-noise (SNR) to extract the desired subtle information. The former, for example, might reveal the initial phase of galaxy or star formation, while the latter could reveal the mix of elements in the big bang or uncover changes in fundamental constants on cosmological timescales ($\sim 10^{10}$ years). The term 'efficiency' has several components:

(i) Throughput of the optical system, including the efficiency of the optical components (lenses or mirrors), dispersers, filters, and also detectors and also includes losses due to transmission/reflection, diffraction, scattering, and vignetting by internal obstructions;
(ii) Noise from all sources: the detector, including time-independent (read-noise) and time-dependent (dark-current) components; thermal emission from the instrument, telescope, enclosure and sky; straylight (see above), and electronic interference;
(iii) Efficient use of the instrument and telescope to maximize the fraction of the time spent accumulating useful photons instead of, say, calibrating; reconfiguring the instrument; or acquiring targets;
(iv) Parallelizing the observations so that the same instrument records data from many objects at the same time. This is discussed further below.

Angular Resolution

The required angular resolution is summarized in Table 2 for key features in galaxies. The resolutions currently achievable with AO systems (with severe restrictions on sky coverage pending routine availability of laser beacons), or without, are shown in bold. Finer resolution is possible using interferometric techniques, but is limited to a handful of bright stars.

Table 2 Typical requirements for angular resolution and single-object field of view

Target	*Distance*[a]	*Active nucleus*		*Star cluster (10 pc)*	*Visible extent*
		Central engine (10 AU)	*Dynamical core (0.5 pc)*		
Milky Way	8 kpc	1 mas	10″	4′	
Local group galaxy (M31)	800 kpc	0.01 mas	100 mas	3″	100″
Distant galaxy	>1 Gpc	10^{-5} mas	0.1 mas	2 mas	0.5–2″
Cluster of galaxies	>1 Gpc				3′
Cosmological structure	>1 Gpc				~1°

[a] 1 pc = 3.1×10^{16} m.

Field of View and Multiplexing

The FOV of an individual observation must be sufficient to cover the angular extent of the object with the minimum number of separate pointings of the telescope. As well as minimizing the total observation time, this reduces photometric errors caused by changes in the ambient conditions (e.g., flexure, thermal effects, and changes in the sky background). (Photometry is the measurement of the flux density received from an astronomical object, usually done with the aid of an imager employing filters with standardized broad passbands. Spectrophotometry is the same but in the much narrower passbands obtained by dispersing the light.) The last column of Table 2 indicates the required size of FOV. It is also highly desirable to observe many objects simultaneously since most studies require large statistically homogenous samples. The precise requirement depends on the surface number density of potential targets and the sampling strategy adopted to create the sample. Cosmological studies of the large-scale distribution of galaxies and clusters emphasize the size of the field ($\geq 0.5°$), since the targets may be sparsely sampled, while studies of clustered objects (galaxies or stars) aim to maximize the surface density of targets (≤ 1000 arcmin^{-2}) within a modest FOV ($\sim 5'$).

Spectral Resolution

The required energy resolution is specified in terms of the spectral resolving power, $R \equiv \lambda/\delta\lambda$, where $\delta\lambda$ is the resolution in wavelength defined usually in terms of the Rayleigh criterion. One of the main data products required in astrophysics is the velocity along the line of sight (radial velocity) obtained through measurement of the wavelength of a known spectral feature (emission or absorption line). This may be determined to an accuracy of:

$$\delta v = \frac{Qc}{R} \qquad [1]$$

where Q is the fraction of the width of the spectral resolution element, $\delta\lambda$, to which the centroid can be determined. This is limited by either the SNR of the observation of the line via $Q \approx 1/\text{SNR}$, or by the stability of the spectrograph via:

$$Q = \frac{\Delta x}{\chi}\left(\frac{d\chi}{dx}\right) \qquad [2]$$

where χ is the angular width of the slit measured on the sky and $\delta\chi/dx$ is the image scale at the detector. Δx is the amount of uncorrectable flexure in units of the detector pixel size. This is the part of the motion of the centroid of the line which is not susceptible to modeling with the aid of frequent wavelength calibration exposures. Typically, $\Delta x \approx 0.2$ pixels is achievable (roughly 3 μm at the detector or 10 μm at the slit for an 8 m telescope) implying $Q \approx 1/30$ for a typical $(d\chi/dx) = 0.1\text{–}0.2''$/pixel and $\chi = 0.5\text{–}1''$. Table 3 gives typical requirement for R and the value of Q required to achieve it, given the spectral resolution obtainable for different types of current spectrograph. Extrasolar planet studies require very high stability.

Some spectral observations require very accurate measurements of the shape of the spectral features to infer the kinematics of ensembles of objects which are individually not resolvable (e.g., stars in distant galaxies, gas in blackhole accretion disks). Other applications require accurate measurements of the relative flux of various spectral features within the same spectrum (e.g., to reveal abundances of elements or different types of star).

Optical Principles

Imaging

The simplest type of imager consists of a detector, comprising a 2D array of light-sensitive pixels (Table 4) placed at a telescope focal surface without any intervening optics except for the filters required for photometry. Although maximizing throughput, this arrangement removes flexibility in changing the image scale since the physical size of pixels (typically 15–25 μm) is fixed. Thus the Gemini telescopes

Table 3 Typical requirements for spectral resolution and stability

Target	*δv or required data product*	*Required λ/δλ*	*Spectral regime*	*Available R*	*Stability required = 1/Q*	*Simultaneous wavelength range*
Cosmological parameters	1000 km/s (line centroid)	300	*Low dispersion*	300–3000	1	≳1 octave
Galaxies						
Intra-cluster	100 km/s (line centroid)	3000			1–10	100 nm
Element abundance	Ratio of line fluxes	1000–3000				1 octave
Stellar orbits	10 km/s (line shapes)	3×10^4	*Medium*	$1–3 \times 10^4$	(1)	10 nm
Individual stars	1 km/s (line centroid)	3×10^5			10–30	10 nm
Extrasolar planets	<1 m/s (line centroid)	$>3 \times 10^8$	*High*	$\sim 10^5$	>3000	<1 nm

Table 4 Detectors in current use for astronomy

Wavelength (μm)	*Material/ type*	*Pixel size (μm)*	*Format*		*Read noise (e-/pix)*	*Dark current (e-/pix/hr)*	*QE range (%)*
			Single	*Mosaic*			
		Typical	*State-of-the-art*				
0.3–1.0	CCD	10–20	4k × 4k	12k × 8k	2	Negligible	10–90[a]
	CMOS	5–10	4k × 4k		5–30		
1–2.5 +	HgCdTe	15–20					
0.6–5	InSb	20–30	2k × 2k	4k × 4k	10	0.01–0.1	70–90

[a] Quantum efficiency varies strongly over quoted wavelength range.

($D_T = 8$ m) provides an uncorrected image scale of 25–40 mas/pixel at its sole F/16 focus. Although well-suited to high-resolution observations with AO, this is inappropriate for most observations of faint galaxies where 100–200 mas/pixel is required. This arrangement also provides limited facilities for defining the wavelength range of the observations since there is no intermediate focal surface at which to place narrowband filters, whose performance is sensitive to field angle and are, therefore, best placed at an image conjugate. There is also no image of the telescope pupil at which to place a cold stop to reject the external and instrumental thermal radiation which is the dominant noise source at $\lambda > 1.5$ μm.

The alternative is to use a focal-reducer consisting of a collimator and camera. The auxiliary components (filters, cold-stops, Fabry–Perot etalons, etc.) then have an additional image and pupil conjugate available to them and the image scale may be chosen by varying the ratio of the camera and collimator focal lengths. The extra complexity of optics inevitably reduces throughput but, through the use of modern optical materials, coatings such as Sol–Gel and the avoidance of internally obstructed reflective optics, losses may be restricted to 10–20% except at wavelength shorter than ~0.35 μm. An imaging focal reducer is a special case of the generic spectrograph described in the next section.

Spectroscopy

Basic principles

Astronomical spectrographs generally employ plane diffraction gratings. The interference between adjacent ray paths as they propagate through the medium is illustrated in Figure 1. From this, we obtain the grating equation:

$$m\rho\lambda = n_1 \sin\alpha + n_2 \sin\beta \equiv G \qquad [3]$$

where α and β are the angles of incidence and diffraction respectively, $\rho = 1/a$ is the ruling density, λ is the wavelength, and m is the spectral order. The refractive indices in which the incident and diffracted rays propagate are n_1 and n_2 respectively.

The layout of a generic spectrograph employing a plane diffraction grating in a focal reducer arrangement is shown in Figure 2. The spectrograph re-images the slit onto the detector via a collimator and camera. The disperser is placed in the collimated beam close to a conjugate of the telescope pupil.

A field lens placed near the focus is often incorporated in the collimator for this purpose.

The geometrical factor, G, defined in the equation is constrained by the basic angle of the spectrograph:

$$\psi = \alpha - \beta \qquad [4]$$

which is the (normally fixed) angle between the optical axes of the collimator and camera. $\psi = n\pi$, corresponds to the Littrow configuration (for integer n).

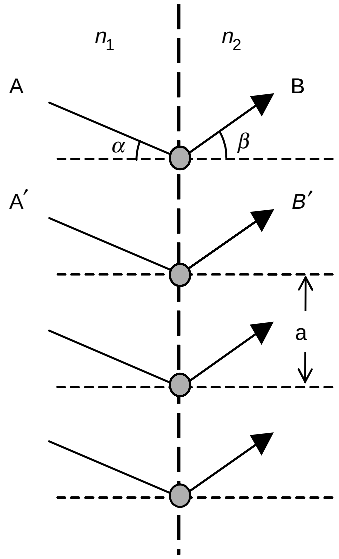

Figure 1 Derivation of the grating equation by consideration of the optical path difference between neighboring ray paths AB and A′B′.

Considering the case of a diffraction grating in vacuo, differentiation of the grating equation with respect to the diffracted angle yields the angular dispersion:

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\beta} = \frac{\cos\beta}{m\rho} \qquad [5]$$

from which the linear dispersion is obtained by considering the projection of dβ on the detector, dx:

$$\frac{\mathrm{d}\lambda}{\mathrm{d}x} = \frac{\mathrm{d}\lambda}{\mathrm{d}\beta}\frac{\mathrm{d}\beta}{\mathrm{d}x} = \frac{\cos\beta}{m\rho f_2} \qquad [6]$$

where f_2 is the focal length of the camera (see Figure 2).

In the diffraction-limited case, where the slit is arbitrarily narrow, the resolving power is given simply by the total number of rulings multiplied by the spectral order:

$$R^* = m\rho W \qquad [7]$$

where w is the length of the disperser as defined below. But in astronomy, it is usually determined by the width of the image of the slit, s, projected on the detector, s'. By conservation of Etendue:

$$s' = s\frac{F_2}{F_1} \qquad [8]$$

where F_1 and F_2 are the collimator and camera focal ratios, respectively. The spectral resolution of the spectrograph is simply the width of this expressed in

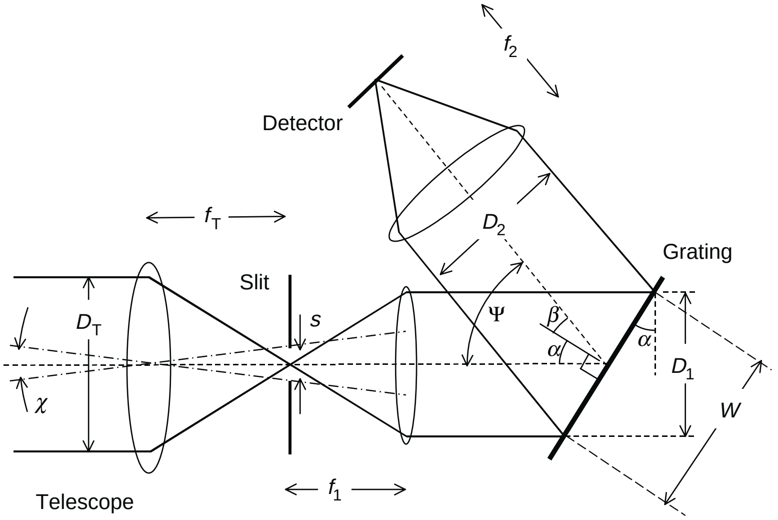

Figure 2 Generic spectrograph employing a plane diffraction grating. Note that according to the sign convention implicit in Figure 1, the diffraction angle, β, is negative in this configuration.

wavelength units:

$$\delta\lambda = \left(\frac{d\lambda}{dx}\right)s' = \frac{\cos\beta}{m\rho f_2} s \frac{F_2}{F_1} = \frac{sD_1 \cos\beta}{m\rho D_2 f_1} \qquad [9]$$

The length of the intersection between the collimated beam and the plane of the grating (not necessarily the actual physical length of the grating) is

$$W = \frac{D_2}{\cos\beta} = \frac{D_1}{\cos\alpha} \qquad [10]$$

Using this to eliminate cos β, the spectral resolution becomes

$$\delta\lambda = \frac{s}{m\rho F_1 W} \qquad [11]$$

and the resolving power is

$$R \equiv \frac{\lambda}{\delta\lambda} = \frac{m\rho\lambda F_1 W}{s} = G\frac{F_1 W}{s} \qquad [12]$$

where G is the quantity defined in eqn [3]. Note that R is now independent of the details of the camera. This expression is useful for the laboratory since it is expressed in terms of the parameters of the experimental apparatus; but, for astronomy, it is more useful to express the resolving power in terms of the angular slit width, χ, and the telescope aperture diameter, D_T, via:

$$s = \chi f_T = \chi F_T D_T = \chi F_1 D_T \qquad [13]$$

since the collimator and telescope focal ratios are the same if the spectrograph is directly beam-fed from the telescope. Note that even if the telescope focal surface is re-imaged onto the slit, the expression for the resolving power still holds due to the conservation of Etendue in the re-imaging optics. Thus the resolving power is

$$R = \frac{m\rho\lambda W}{\chi D_T} = \frac{GW}{\chi D_T} \qquad [14]$$

Note that the resolving power obtained with a nonzero slit width is always less than the theoretical maximum, $R \le R_*$, for wavelengths:

$$\lambda < \lambda_* = \chi D_T \qquad [15]$$

Thus, long-wavelength applications may approach the theoretical limit; in which case they are said to be diffraction-limited. If so, the resolving power is independent of the slit width, simplifying the interface to different telescopes. For a nondiffraction-limited spectrograph, the resolving power obtained depends on the aperture of the telescope to which it is fitted.

Optical configuration

The non-Littrow configuration shown in Figure 2 introduces anamorphism into the recorded spectra such that a monochromatic image of the slit mask is compressed in the dispersion direction by a factor:

$$A = \frac{D_2}{D_1} = \frac{\cos\beta}{\cos\alpha} \qquad [16]$$

which varies, depending on the choice of angles required to place the chosen wavelength on the detector using eqns [3] and [4].

Practical configurations that maximize R, do so by increasing $W = D_1/\cos\alpha$. This implies maximizing cos β, since $\psi = \alpha - \beta$, which results in $A > 1$. This is known as the grating-to-camera configuration since the grating normal points more towards the camera than to the collimator. The grating-to-collimator configuration may also be used at the penalty of lower R but does permit a smaller camera. Thus the camera aperture must be oversized in the dispersion direction since $D_2 > D_1$. The Littrow configuration is rarely used with reflective designs when large FOV is required because of internal vignetting and consequent light loss. However, it can be used effectively with transmissive dispersers.

If the disperser is replaced by a plane mirror (or if the disperser is removed in a transmissive system), the system acts as a versatile imager, with $A = 1$. This also facilitates target acquisition which requires that the telescope attitude be adjusted until the targets are aligned with apertures cut in an opaque mask placed at the input of the spectrograph. These apertures (normally slits) are usually oversized in the direction perpendicular to the dispersion to allow the spectrum of the sky to be recorded directly adjacent to that of the target to permit accurate subtraction of the background.

The slit width is chosen by trading off the desire to maximize the amount of light entering the instrument from the target (but not so much as to admit excessive background light) and the need to maximize spectral resolving power, R, by reducing the slit width, χ, according to eqn [14].

Since $|G| \le 2$, the maximum attainable resolving power, according to eqn [14], is determined chiefly by the telescope aperture and angular slitwidth. The only important parameter of the spectrograph is the illuminated grating length, W. Therefore, maintaining the same resolving power as the telescope aperture increases, requires W to increase in direct proportion. Because of limits placed on the geometry by the need to avoid obstructions, etc., this means that both the collimated beam diameter and the size of the disperser must increase. This is one of the major problems in devising instruments for ELTs. Current instruments on 8 m telescopes feature collimated beam diameters of $D_1 \approx 150$ mm. For ELTs, the implied sizes are 0.5–2 m, requiring huge optics, accurately co-aligned mosaics of dispersing elements, and enormous support structures to provide the required stability. A simple calculation indicates that the ruling stylus for a classically ruled diffraction grating would have to cover a thousand kilometers without significant wear unless the disperser was made from smaller pieces.

Choice of dispersing element

Astronomical spectrographs mostly use surface-relief diffraction (SL) gratings. These are ruled with groove densities $50 < \rho < 3000\ \mathrm{mm}^{-1}$.

Blazing diffraction gratings. The intensity of the multibeam interference pattern from N rulings of finite width is given by

$$I = \left(\frac{\sin^2 N\phi}{\sin^2 \phi}\right)\left(\frac{\sin^2 \theta}{\theta^2}\right) \qquad [17]$$

where 2ϕ is the phase difference between the center of adjacent rulings and θ is the phase difference between the center and edge of a single ruling. To be useful for astronomy, the peak in this pattern, determined by the second term, must coincide with a useful order, such as $m = 1$, rather than zero order. This can be done by introducing a phase shift into the second term, which represents the diffraction of a single ruling, such that:

$$\theta = \frac{\pi \cos \gamma}{\rho\lambda}(\sin i - \sin r) \qquad [18]$$

where the groove angle, γ, is the angle by which the facets are tilted with respect to the plane of the grating. With the aid of Figure 3, it can be seen that $i = \alpha - \gamma$ and $r = \gamma - \beta$ so that the maximum efficiency, which occurs when $\theta = 0$, corresponds to $i = r$, which is equivalent to simple reflection from the facets, and $\alpha + \beta = 2\gamma$.

Thus, through this process of blazing, the grating eqn [3], at the blaze wavelength, λ_B, becomes:

$$\rho m \lambda_\mathrm{B} = 2 \sin \gamma \cos \frac{\psi}{2} \qquad [19]$$

where ψ, is given by eqn [4] using the identity: $\sin x + \sin y \equiv 2 \sin\left(\frac{x+y}{2}\right) \cos\left(\frac{x+y}{2}\right)$. Thus, the resolving power at blaze is obtained from eqns [10], [14], and [19] as:

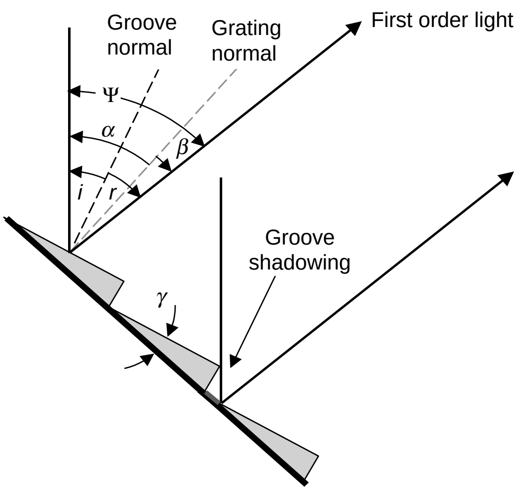

Figure 3 Blazing of ruled diffraction grating.

$$R_\mathrm{B} = \frac{2D_1}{\chi D_\mathrm{T}} \frac{\sin\gamma \cos(\psi/2)}{\cos(\gamma + \psi/2)} \qquad [20]$$

Use of grisms. An alternative configuration uses transmission gratings. The advantage here is that the collimator and camera can share the same optical axis in a straight-through configuration with unit anamorphism, $A = 1$, which does not require oversizing of the camera. The same phase-shift can be applied as before by making the facets into prisms with the required groove angle. However, this would only allow zero-order to propagate undeviated into the camera so an additional prism is required with vertex angle $\phi = \gamma$ to allow first-order light to pass undeviated into the camera. This composite of blazed transmission grating and prism is called a grism (Figure 4).

The grating equation (eqn [3]) is modified for a grism in the blaze condition to $m\rho\lambda_B = (n-1)\sin\phi$ since $\alpha = -\beta = \phi$ and $n_1 \equiv n, n_2 = 1$. Noting that $\psi = 0$, the resolving power at blaze is then:

$$R_\mathrm{B} = \frac{D_1}{\chi D_\mathrm{T}}(n-1)\tan \phi \qquad [21]$$

where n is the refractive index of the prism and grating material. Due to problems such as groove shadowing (illustrated for reflection gratings in Figure 3), these lack efficiency when $\phi \gtrsim 30°$ restricting their use to low-resolution spectroscopy.

Use of volume phase holographic gratings. A newer alternative to surface-relief (SR) dispersers are volume phase holographic (VPH) gratings in which the interference condition is provided by a variation of refractive index of a material such as dichromated gelatine, n_g, which depends harmonically on position inside the grating material as:

$$n_g(x,z) = \bar{n}_g + \Delta n_g \cos[2\pi\rho_g(x \sin\gamma + z\cos\gamma)] \qquad [22]$$

where ρ_g is the density of the lines of constant n_g and $\bar{n}_g$ and Δn_g are the mean refractive index and its modulation amplitude, respectively (Figure 5). These are described by the same grating (eqn [3]), as before leading to an identical expression for the resolving power at blaze, where $m\rho\lambda_B = 2\sin\alpha$, to that of a blazed plane diffraction grating in the Littrow configuration (eqn [20]):

$$R_B = \frac{D_1}{\chi D_T} 2\tan\alpha \qquad [23]$$

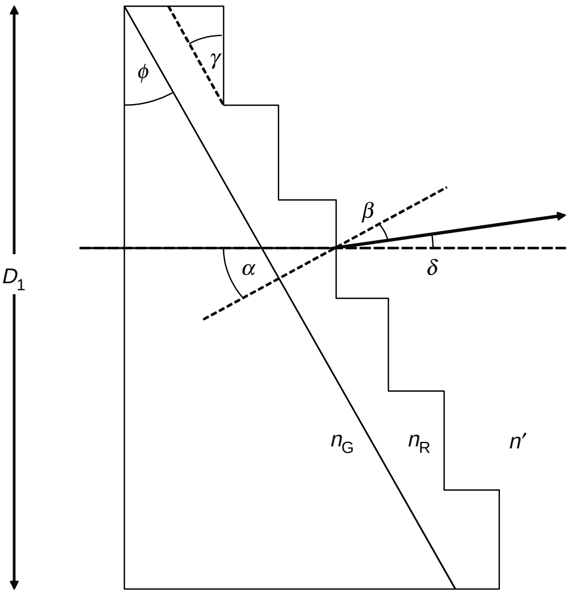

Figure 4 Typical configuration of a grism. The prism vertex and groove angles are chosen for maximum efficiency on the assumption that refractive indices, $n_G = n_R = n$ and $n' = 1$. The blaze condition occurs when $\delta = 0$. In this configuration, $\phi = \gamma = \alpha$.

Unlike SR gratings, the blaze condition of VPH gratings may be varied by changing the basic angle of the spectrograph, $\psi = 2(\gamma + \alpha)$, although this is often mechanically inconvenient. They may also be sandwiched between prisms to provide the necessary incident and diffraction angles while retaining the advantage of a straight-through configuration, to form an analogue of a grism, sometimes termed a vrism. Unlike grisms, these are not constrained by groove shadowing and so can be efficient at relatively high dispersion, subject to the constraint imposed by the size of the device.

Use of echelle gratings. Another option to increase the resolving power is to use a coarse grating in high order via the use of an echelle format. The resolving power in the near-Littrow configuration in which it is usually operated, is given by eqn [20] with $\psi = 0$.

$$R_B = \frac{D_1}{\chi D_T} 2\tan\gamma \qquad [24]$$

where γ is the groove angle.

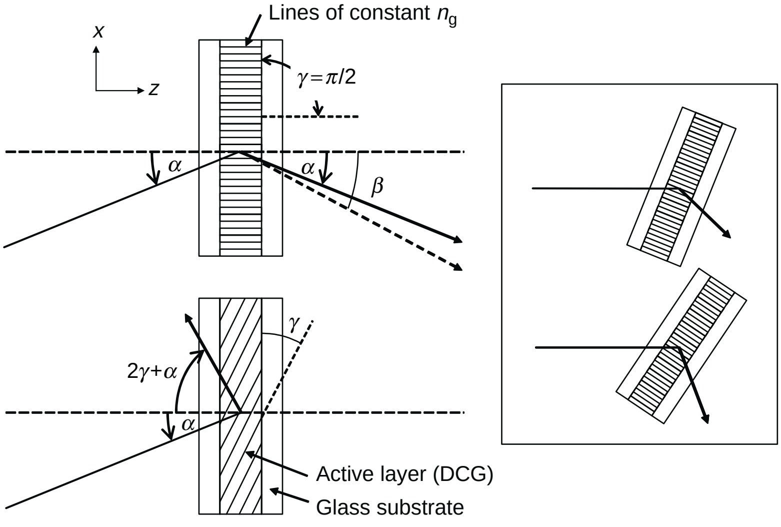

Figure 5 Various configurations of volume phase holographic gratings. The blaze condition is when $\beta = \alpha$. The inset shows how the blaze condition may be altered by changing the basic angle of the spectrograph.

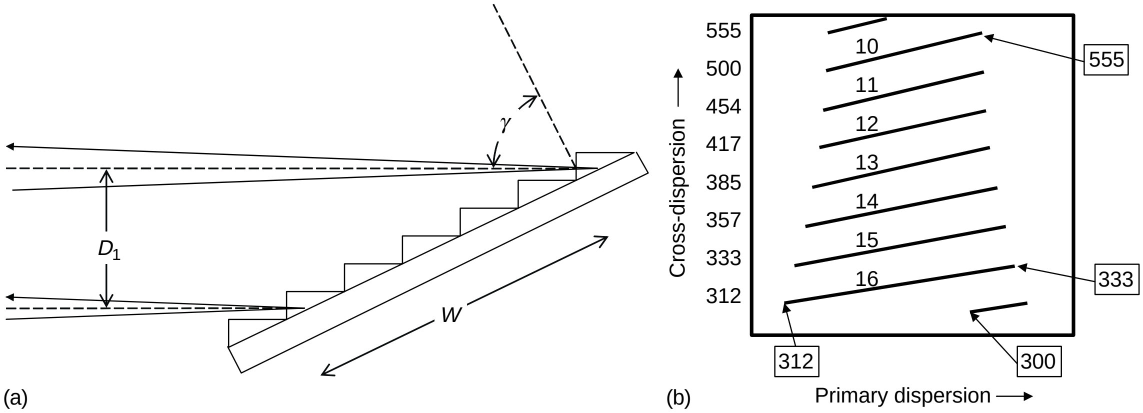

Figure 6 (a) Basic layout of an Echelle grating used in a near-Littrow configuration. (b) Illustrative cross-dispersed spectrogram showing a simplified layout on the detector. The numbers 10–16 label the different spectral orders. The numbers labeling the vertical axis are the wavelength (nm) at the lowest end of each complete order. Other wavelengths are labeled for clarity. For simplicity the orders are shown evenly spaced in cross-dispersion.

In the configuration shown in Figure 6, R is maximized by increasing W. This also means that $G = m\rho\lambda$ is large. In order to avoid excessively high ruling densities, the grating may be used in high orders. However, this presents the problem of order overlap since the wavelength in order m occurs in the same place as wavelength $\lambda_n = (m/n)\lambda_m$ in order n. Limiting the observation to a single order via the use of order-sorting filters is one solution while another is to use cross-dispersion via a chain of prisms or a grating with dispersion perpendicular to that of the echelle to separate the orders out. This option severely curtails the use of such spectral formats in multi-object spectroscopy.

Increasing resolving power using immersed gratings. All the methods of using diffraction gratings discussed so far are subject to various geometric constraints which ultimately limit the maximum obtainable resolving power by placing limits on W (e.g., in eqn [14]). These may be partly overcome by optically coupling the dispersers to immersing prisms. This can lead to a doubling of the resolving power in some cases.

Use of prisms. The final method of dispersion is to use a prism. This has the advantage that, since it does not work by multibeam interference, the light is not split into separate orders which removes the problem of order-overlap and improves efficiency. However, it does not produce the high spectral resolution required for some applications and the dependency of dispersion on λ is often markedly nonlinear.

From a consideration of Fermat's principle, it can be shown that the resolving power of a prism is:

$$R = \frac{\lambda t}{\chi D_{\mathrm{T}}}\left(\frac{\mathrm{d}n}{\mathrm{d}\lambda}\right) \qquad [25]$$

where t is the baselength of the prism. Using modern materials, $R \lesssim 300$ may be obtained. The problem of the nonlinearity in dispersion can be alleviated by the use of composite materials with opposite signs of dispersion.

Multiobject spectroscopy (MOS)

As discussed above, there is great value in increasing the multiplex gain. This can be done in two ways:

(i) *Increasing the number of apertures in the slit mask (Multislits).* This requires that each slit is carefully cut in the mask because, although the telescope attitude (translation and rotation) may be adjusted to suit one aperture, errors in the relative positions of slits cannot be compensated since the image scale is fixed. Each aperture produces a spectrum on the detector and the mask designer must ensure that the targets and their matching slits are chosen to avoid overlaps between spectra and orders. This has the effect of limiting the surface density of targets which can be addressed in one observation. Passband filters can be used to limit the wavelength range and reduce the overlap problem. This option requires superior optics able to accommodate both a wide FOV and large disperser.

(ii) *Using optical fibers (Multifibers).* A number of fibers may be deployed at the telescope focus to direct light from the various targets to a remote spectrograph whose input focus consists of one or more continuous pseudoslits. These are made up of the fiber outputs arranged in a line. The continuous nature of the slit means that spectral overlap is avoided without restricting the surface density of addressable targets; although there is a limit imposed by the distance of closest approach of the fibers. The method of deploying the fiber inputs may be a plugplate consisting of pre-cut holes into which encapsulated fibers are manually plugged, or a pick-and-place robot which serially positions the fiber inputs at the correct location on a magnetized field plate. The latter is highly versatile but mechanically complex with a significant configuration time that may erode the actual on-target integration time.

The two systems have contrasting capabilities (Figure 7). The multislit approach provides generally better SNR since the light feeds into the spectrograph directly, but is compromised in terms of addressable surface density of targets, by the need to avoid spectral overlap and in FOV by the difficulty of the wide-field optics. A current example of this type of instrument, GMOS, is shown in Figure 8 and described in Table 5. The multifiber approach is limited in SNR by modal noise in the fibers and attendant calibration uncertainties and the lack of contiguous estimates of the sky background. However, it is easier to adapt to large fields since the spectrograph field can be much smaller than the field over which the fiber inputs are distributed. In summary, multislit systems are best for narrow-but-deep surveys while the multifiber systems excels at wide-but-shallow surveys. Fiber-fed instruments may further prove their worth in ELTs where technical problems may require that these bulky instruments are mounted off the telescope (see below for further discussion).

Of paramount importance in MOS is the quality of the background subtraction as discussed above. Traditionally, this requires slits which sample the sky background directly adjacent to the object. An alternative is nod-and-shuffle (va-et-vient) in which nearby blank regions are observed alternately with the main field using the same slit mask by moving the telescope (the nod). In the case of CCDs, the photo-generated charge from the interleaved exposures is stored temporarily on the detector by moving it to an unilluminated portion of the device (the shuffle). After many repeats on a timescale less than that of variations in the sky background (a few minutes), the accumulated charge is read out, incurring the read-noise penalty only once. Although requiring an effective doubling of the exposure time and an

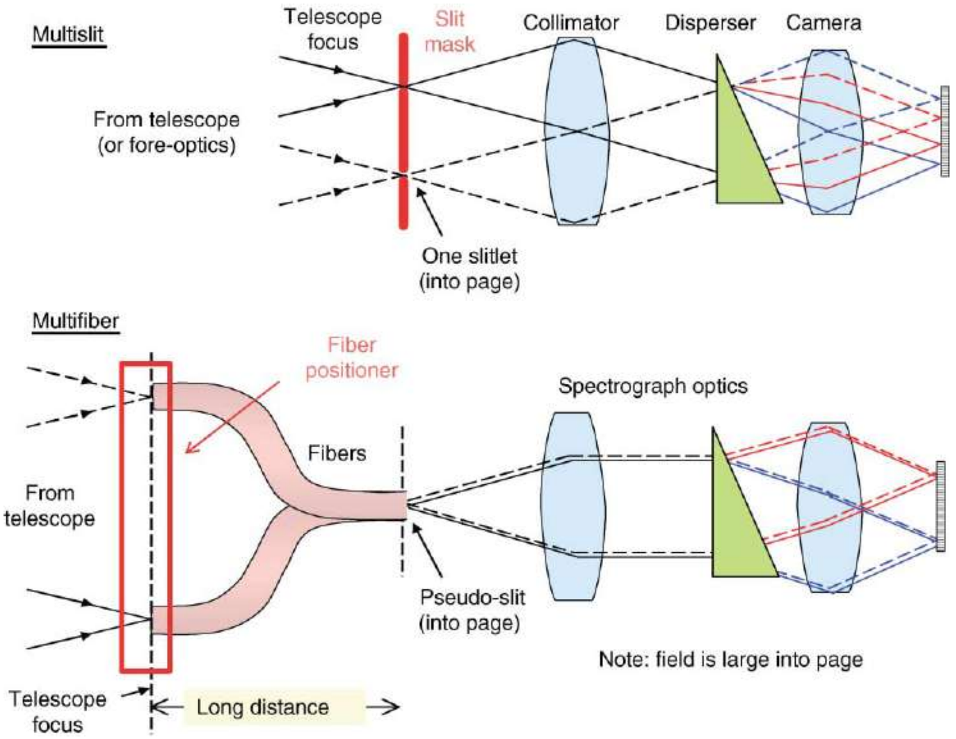

Figure 7 Illustration of the difference between the multislit and multifiber approaches to multi-object spectroscopy.

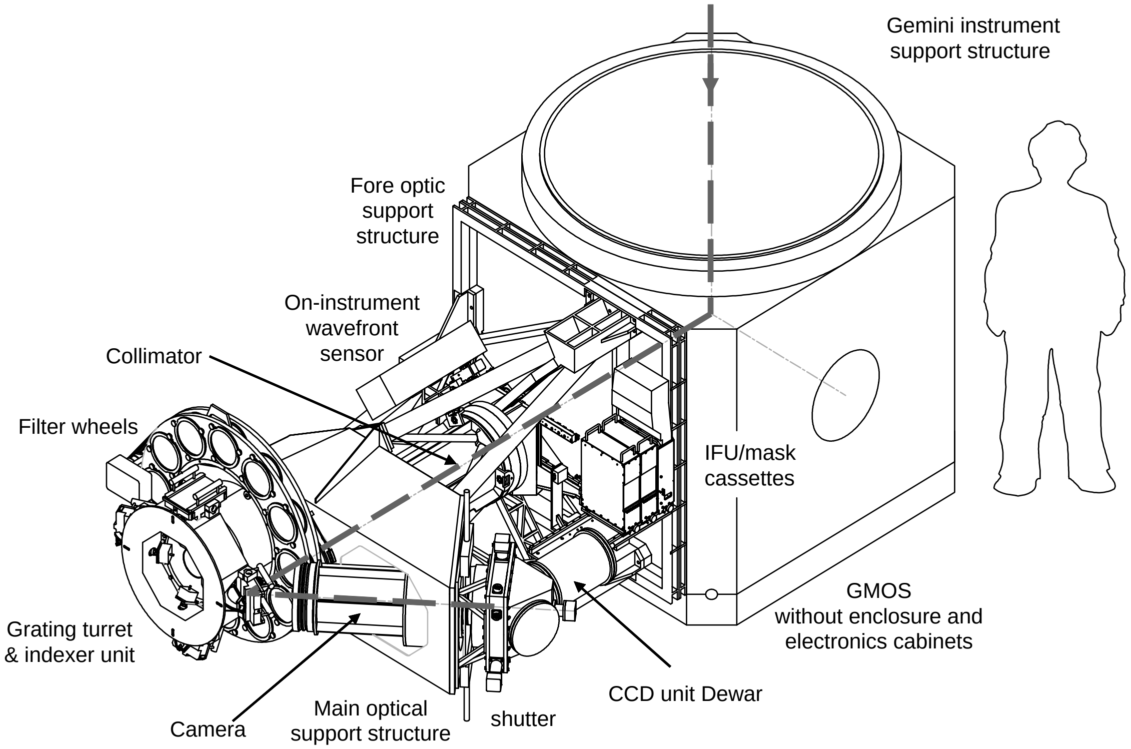
Gemini instrument support structure
Fore optic support structure
On-instrument wavefront sensor
Collimator
Filter wheels
IFU/mask cassettes
GMOS without enclosure and electronics cabinets
Grating turret & indexer unit
CCD unit Dewar
Main optical support structure
shutter
Camera

Figure 8 The Gemini Multiobject Spectrograph (GMOS). One of two built for the two Gemini 8-m telescopes by a UK–Canadian consortium. It includes an integral field capability provided by a fiber-lenslet module built by the University of Durham. The optical configuration is the same as shown in Figure 2. It is shown mounted on an instrument support unit which includes a mirror to direct light from the telescope into different instruments. The light path is shown by the red dashed line. The slit masks are cut by a Nd-YAG laser in 3-ply carbon fiber sheets. See Table 5 for the specification.

Table 5 Main characteristics of the Gemini multiobject spectrographs

Image scale	72 mas/pixel	FOV	5.5′ × 5.5′
Detector	CCD with 13.5 μm pixels Format: 3 × (4608 × 2048)	Slit/mask configuration	$\leq$few × 100 slits of width $\geq$ 0.2″
Wavelength range	Total: 0.4–1.0 μm Simultaneous: $\leq$ 1 octave	Integral field unit	1500 × 0.2″ samples in 50 sq.″
Spectral resolving power	$R \leq 5000$ with 0.5″ slit	Dispersion options	3 gratings + mirror for imaging

increase in the size of the detector, this technique allows the length of the slits to be reduced since no contiguous sky sample is required, thereby greatly increasing the attainable multiplex gain if the number density of potential targets is sufficiently high.

Integral field spectroscopy (IFS)

IFS provides a spectrum of each spatial sample simultaneously within a contiguous FOV. Other approaches provide the same combination of imaging and spectroscopy but require a series of nonsimultaneous observations. Examples include imaging through a number of narrow passband filters with different effective wavelength and spectroscopy with a single slit which is stepped across the object. Other such techniques are Fabry–Perot, Fourier-transform and Hadamard spectroscopy. Since the temporal variation of the sky background is a major limitation in astronomy of faint objects, IFS is the preferred technique for terrestrial observations of faint objects, but nonsimultaneous techniques are preferable in certain niche areas, and relatively more important in space where the sky background is reduced.

The main techniques (Figure 9) are as follows:

(i) *Lenslets*. The field is subdivided by placing an array of lenslets at the telescope focus (or its conjugate following re-imaging fore-optics) to form a corresponding array of micropupils. These are re-imaged on the detector by a focal

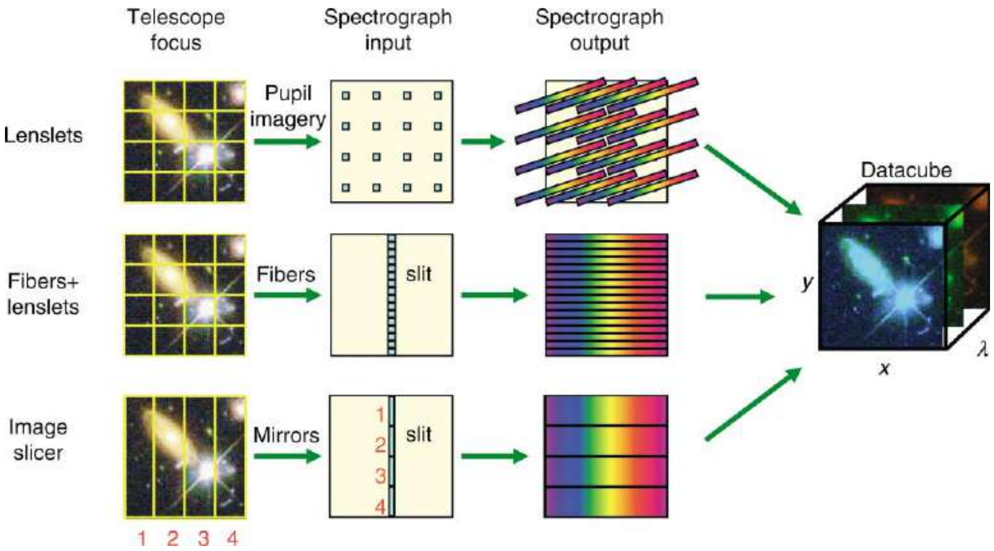

Figure 9 Main techniques of integral field spectroscopy.

reducer using a conventional disperser. The spectra are, therefore, arranged in the same pattern as that of the lenslet array. Spectral overlap is reduced by angling the dispersion direction away from the symmetry axes of the lenslet array and by the fact that the pupil images are smaller than the aperture of the corresponding lenslet.

(ii) *Fibers + lenslets*. The field is subdivided as in (i) but the pupil images are relayed to a remote spectrograph using optical fibers which reformat the field into linear pseudoslits. This avoids the problem of overlap, allowing a greater length of spectrum than in (i), but is more complex since the arrays of fibers and lenslets must be precisely co-aligned and are subject to modal noise in the fibers.

(iii) *Image slicer.* The field is subdivided in only one dimension by a stack of thin slicing mirrors placed at a conjugate of the telescope focus (Figure 10). Each mirror is angled so as to direct light to its own pupil mirror which re-images the field to form part of a linear pseudoslit. Thus the slices into which the image is divided are rearranged end-to-end to form a continuous slit at the spectrograph input. Unlike the other techniques, this retains spatial information along the length of each slice. This is dispersed by conventional means via a focal-reducer. An additional optic is often required at the image of each slice at the slit to reimage the micropupil images produced by the slicing mirrors onto a common pupil inside the spectrograph. In principle this is the most efficient of the three methods since the one-dimensional slicing produces fewer diffraction losses in the spectrograph than the two-dimensional division used by the others, and minimizes the fraction of the detector surface which must be left unilluminated in order to avoid cross-talk between noncontiguous parts of the field. However the micromirrors are difficult to make since they require diamond-turning or grinding in metal or glass with a very fine surface finish (typically with RMS ~1 nm for the optical and ~10 nm in the infrared). This sort of system is, however, well-matched to cold temperatures, since the optical surfaces and mounts may be fabricated from the same material (e.g., Al) or from materials with similar thermal properties.

The data are processed into a datacube whose dimensions are given by the two spatial coordinates, x, y, plus wavelength. The datacube may be sliced up in ways analogous to tomography to understand the physical process operating in the object.

Spectropolarimetry and Polarimetry

Spectropolarimetry and polarimetry are analogous to spectroscopy and imaging, where the polarization state of the light is measured instead of the total intensity. Like IFS, spectropolarimetry is a photon-starved (i.e., limited in SNR by photon noise) area which benefits from the large aperture of current and projected telescopes.

The Stokes parameters of interest are I, Q, and U. V is generally very small for astronomical objects.

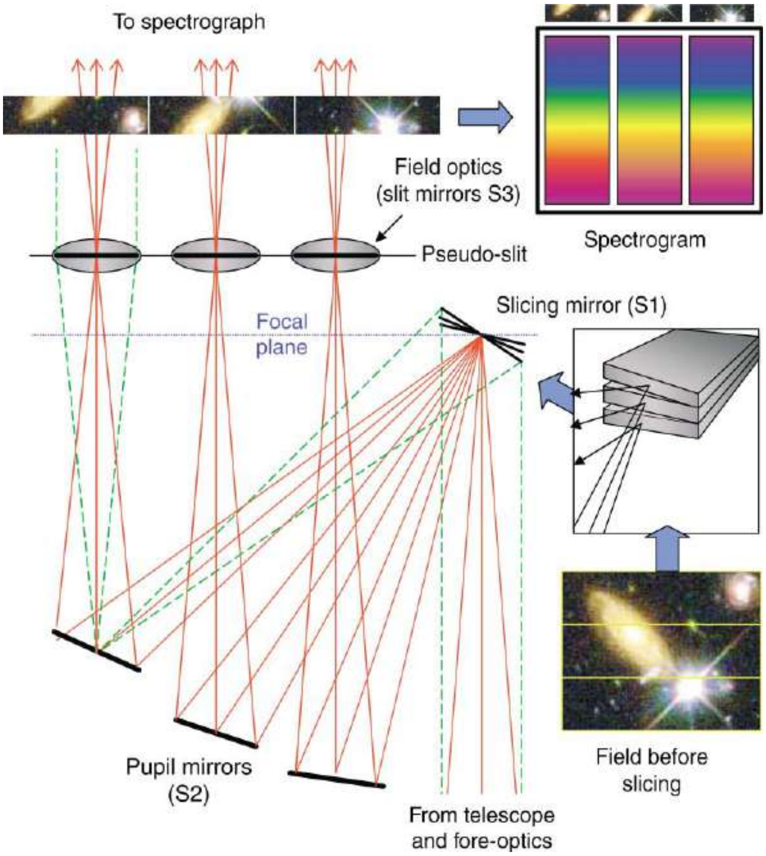

Figure 10 Principle of the advanced image slicer. Only three slices are shown here. Real devices have many more, e.g., the IFU for the *Gemini Near-Infrared Spectrograph* has 21 to provide a 5″ × 3″ field with 0.15″ × 0.15″ sampling. Reproduced from Content R (1997) A new design for integral field spectroscopy with 8 m telescopes. *Proceedings SPIE* 2871: 1295–1305.

From these, the degree and azimuthal angle of linear polarization are obtained as:

$$\pi_L = \left(\frac{Q^2 + U^2}{I^2}\right) \qquad \psi = \frac{1}{2}\arctan\left(\frac{U}{Q}\right) \qquad [26]$$

The Stokes parameters may be estimated using a modified spectrograph with a rotatable achromatic half-wave retarder, characterized by the angle of rotation about the optical axis, θ, placed before the instrument and a polarizing beamsplitter which separates the incident beam into orthogonal polarization states (Figure 11). The two states are recorded simultaneously on different regions of the detector. The separation is achieved usually through the angular divergence produced by a Wollaston prism placed before the disperser or through the linear offset produced by a calcite block placed before the slit. The intensity as a function of the waveplate rotation angle, $S(\theta)$, is recorded for each wavelength and position in the spectrum or image. The Stokes parameters for each sample are then given by

$$\begin{aligned}\frac{Q}{I} &= \frac{S(0) - S(\pi/2)}{S(0) + S(\pi/2)} \\ &= \left\langle \frac{A_1 - \omega_{AB}B_1}{A_1 + \omega_{AB}B_1}, \frac{A_2 - \omega_{AB}B_2}{A_2 + \omega_{AB}B_2} \right\rangle \\ \frac{U}{I} &= \frac{S(\pi/4) - S(3\pi/4)}{S(\pi/4) + S(3\pi/4)} \\ &= \left\langle \frac{C_1 - \omega_{CD}D_1}{C_1 + \omega_{CD}D_1}, \frac{C_2 - \omega_{CD}D_2}{C_2 + \omega_{CD}D_2} \right\rangle \\ I &= S(0) + S(\pi/2) \\ &= \left\langle \frac{A_1 + \omega_{AB}B_1}{\eta_{A1}}, \frac{A_2 + \omega_{AB}B_2}{\eta_{A2}} \right\rangle \text{ etc.}\end{aligned} \qquad [27]$$

where A and B are the signals (functions of position and/or λ) recorded with the waveplate at $\theta = 0$ and $\theta = \pi/2$ and the subscripts 1 and 2 refer to the two polarization states produced by the beamsplitter.

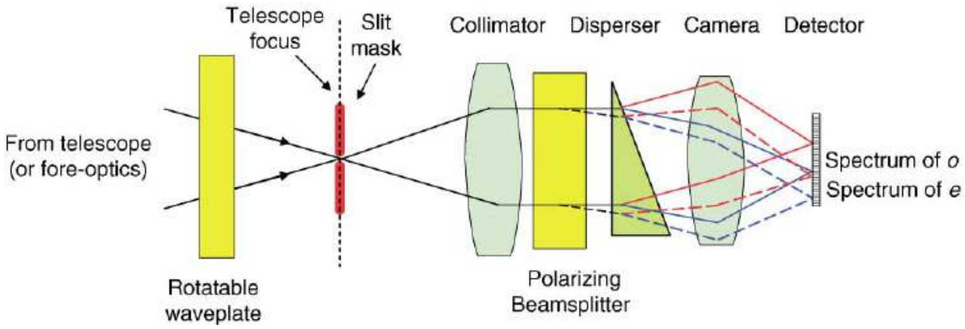

Figure 11 Schematic layout of spectropolarimeter using a polarizing beamsplitter in the collimated beam (e.g., a Wollaston prism producing *e*- and *o*-states) to produce separate spectra for orthogonal polarization states on the detector. Alternative configurations in which the two polarizations have a linear offset (e.g., a calcite block) may be used in which case the beamsplitter is placed before the slit.

Likewise, C and D are a pair of observations taken with $\theta = \pi/4$ and $\theta = 3\pi/4$. The angled brackets indicate a suitable method of averaging the quantities inside. The factors ω_{AB} and ω_{CD} are estimated as

$$\omega_{AB} = \sqrt{\frac{A_1 A_2}{B_1 B_2}} \qquad \omega_{CD} = \sqrt{\frac{C_1 C_2}{D_1 D_2}} \qquad [28]$$

and η_{A1} is a single calibration factor for polarization state 1 of observation A, etc. One problem is instrumental polarization caused by oblique reflections from telescope mirrors upstream of the polarizing waveplate, such as the fold mirror external to GMOS (see Figure 8). This must be carefully accounted for by taking observations of stars with accurately known polarization. The dispersing element is also likely to be very strongly dependent on the polarization state of the incident light with widely different efficiency as a function of wavelength. Although this effect is cancelled out using the procedure described above, it introduces a heavy toll in terms of SNR reduction.

Technology Issues

Use of Optical Fibers in Astronomy

As discussed above, optical fibers are commonly used in astronomy for coupling spectrographs to wide, sparsely sampled FOVs in the case of MOS, or small contiguously sampled FOVs in the case of IFS. For both applications, the most important characteristics are:

(i) *Throughput*. Near-perfect transmission is required for wavelength intervals of ~1 octave within the overall range of 0.3–5 μm. For MOS, a prime focus fiber feed to an off-telescope spectrograph implies fiber runs of several tens of meters for 8 m telescopes, increasing proportionally for ELTs. For IFS, runs of only ~1 m are required for self-contained par-focal IFUs. For $\lambda < 2.5$ μm, fused silica is a suitable material, although it is not currently possible to go below ~0.35 μm, depending on fiber length. Newer techniques and materials may improve the situation. For $\lambda > 2.5$ μm, alternative materials are required, such as ZrF4 and, at longer wavelengths, chalcogenide glasses, but these are much more expensive and difficult to use than fused silica.

(ii) *Conservation of Etendue*. Fibers suffer from focal ratio degradation, a form of modal diffusion, which results in a speeding up of the output beam with respect to the input. Viewed as an increase in entropy, it results in a loss of information. In practice, this implies either reduced throughput, as the fiber output is vignetted by fixed-size spectrograph optics, or a loss in spectral resolution if the spectrograph is oversized to account for the faster beam from the fibers. This effect may be severe at the slow focal ratios pr oduced by many telescopes ($F > 10$) but is relatively small for fast beams ($F < 5$). Thus, its effect may be mitigated by speeding up the input beam by attaching lenslets directly to the fibers, either individually (for MOS) or in a close-packed array (for IFS). If the spectrograph operates in both fiber- and beam-fed modes (e.g., Figure 8), it is also necessary to use lenslets at the slit to convert the beam back to the original speed.

(iii) *Efficient coupling to the telescope*. This requires that the fibers are relatively thick to match the physical size of the sampling at the telescope focus which is typically $\chi_s = 0.1''$ for IFS and $\chi_s = 2''$ for MOS. By conservation of Etendue, the physical size of the fiber aperture is

$$d_f = \chi_s D_T F_f \qquad [29]$$

where F_f is the focal ratio of the light entering the fiber and D_T is the diameter of the telescope aperture. Using the above values with $F_f = 4$, as required to maintain Etendue, implies $15 < d_f < 300$ μm for an 8 m or $60 < d_f < 1200$ μm for a 30 m telescope. Except at the lower end, this implies the use of multimode (step-index) fibers. However, the need in IFS to oversize the fibers to account for manufacturing errors and current requirements of $\chi_s \geq 0.2''$ to produce sufficient SNR in a single spatial sample gives a practical limit of $d_f > 50$ μm which is in the multimode regime.

Recently, the use of photonic crystal fibers in astronomy has been investigated. These, together with other improvements in material technology may improve the performance of optical fibers, which will be of special relevance to ELTs where, for example, there will be a need to couple very fast beams into fibers.

Cooling for the Near-Infrared

As shown in Figure 12, it is necessary to cool instruments for use in the infrared. This is to prevent thermal emission from the instrument becoming the dominant noise source. The minimum requirement is to place a cold stop at a conjugate of the telescope pupil. However, much of the rest of the instrument requires cooling because of the nonzero emissivity of the optics. The telescope also requires careful design. For example, the Gemini telescopes are optimized for the NIR by undersizing M2 so that the only stray light entering the final focal surface is from the cold night sky.

Irrespective of the thermal background and wavelength, all detectors (see Table 4), require cooling to reduce internally generated noise to acceptable levels. For the CCDs used in the optical, the cryogen is liquid nitrogen, but for the infrared, cooling with liquid helium is required.

Structure

As discussed above, the necessity for instruments to scale in size with the telescope aperture cannot be achieved by simply devising a rigid structure, since the required materials are generally not affordable (instruments on 8 m telescopes cost roughly €/$ 5–15 M, including labor). Furthermore, the stability also depends critically on the mounting of the individual optical elements, where the choice of material is constrained by the need for compliance with the optical materials. The solution adopted for GMOS was to design the collimator and camera mounting so that variations in the gravity vector induce a translation in each component orthogonal to the optical axis without inducing tilts or defocus. Therefore, the only effect of instability is a translation of the image of the slit on the detector. Since the slit area is rigidly attached to the telescope interface, the image of a celestial object does not move with respect

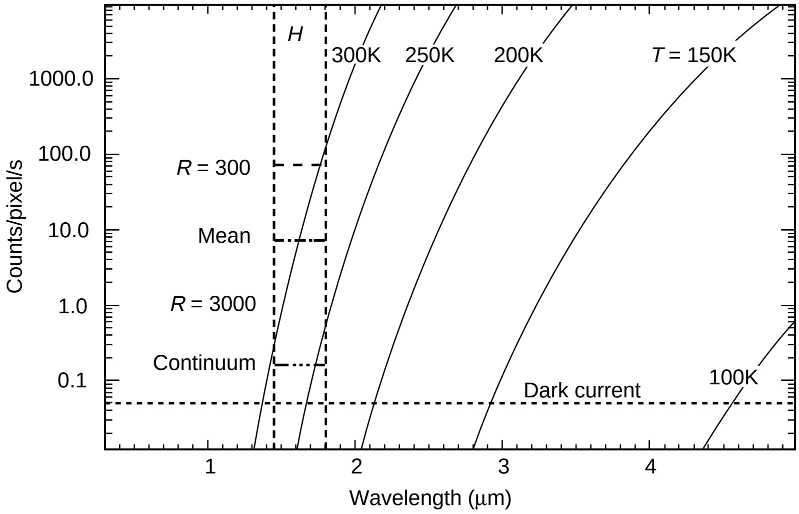

Figure 12 Illustration of the need for cooling in NIR instruments. This is modeled for the case of a spectrograph on an 8 m telescope with 0.2″ × 0.2″ sampling with both throughput and total emissivity of 50%. The curves indicate the cumulative thermal background for wavelengths shortward of that shown on the scale. For comparison, a typical dark current is shown for NIR detectors (see Table 4). Also shown is the background from the night sky for the case of the *H*-band, for two values of spectral resolution, *R*. The mean sky background level is shown for both values of *R*. For $R = 3000$, the continuum level found between the strong, narrow OH-lines which make up most of the signal is also shown. To observe in the *H*-band, cooling to −40 °C is required for high spectral resolution, but is unnecessary at low resolution. At longer wavelengths, fully cryogenic cooling with liquid nitrogen is generally required.

to the slit providing that the telescope correctly tracks the target. The motion of the image of the slit on the detector is then corrected by moving the detector orthogonal to the optical axis to compensate. Taking account of nonrepeatable motion in the stucture and the detector mounting, it proved possible to attain the desired stability in open-loop by measuring the flexure experienced with the control system turned off, making a look-up table to predict the required detector position for each attitude setting (elevation and attitude) and applying the correction by counting pulses sent to the stepper motors which control the detector translation.

Some instruments require much greater stability than GMOS does. The flexure-control outlined above may be augmented by operation in closed-loop so that nonrepeatable movements may be accounted for. However, this ideally requires a reference light source which illuminates the optics in the same way as the light from the astronomical target and which is recorded by the science detector without adversely impacting the observation. Alternatively, partial feedback may be supplied by repeated metrology of key optical components such as fold-mirrors.

One strategy to measure radial velocities to very small uncertainities (a few meters per second) is to introduce a material (e.g., iodine) into the optical path, which is present throughout the observations. This absorbs light from the observed object's continuum at one very well-defined wavelength. Thus, instrument instability can be removed by measuring the centroid of the desired spectral line in the object relative to the fixed reference produced by the absorption cell. However, flexure must still be carefully controlled to avoid blurring the line during the course of a single exposure.

Mounting of instruments via a fiber-feed remotely from the telescope, at a location where they are not subjected to variations in the gravity vector, is a solution for applications where great stability is required. But, even here, care must be taken with the fore-optics and pickoff system attached to the telescope, and to account for modal noise induced by changes in the fiber configuration as the telescope tracks.

Finally, the instrument structure has other functions. For cooled instruments, a cryostat is required inside which most of the optical components are mounted. For uncooled instruments, careful control of temperature is needed to avoid instrument motion due to thermal expansion. This requires the use of an enclosure which not only blocks out extraneous light, but provides a thermal buffer against changes in the ambient temperature during observations and reduces thermal gradients by forced circulation of air.

See also

Diffraction: Diffraction Gratings. **Fiber and Guided Wave Optics**: Fabrication of Optical Fiber. **Imaging**: Adaptive Optics. **Instrumentation**: Telescopes; Spectrometers. **Spectroscopy**: Fourier Transform Spectroscopy; Hadamard Spectroscopy and Imaging.

Further Reading

Allington-Smith JR, Murray G, Content C, *et al.* (2002) Integral field spectroscopy with the Gemini Multiobject Spectrograph. *Publications of the Astronomical Society of the Pacific* 114: 892.

Bjarklev A, Broeng J and Sanchez Bjarklev AS (2003) *Photonic Crystal Fibers*. Kluwer.

Bowen IS (1938) The Image-Slicer: a device for reducing loss of light at slit of stellar spectrograph. *Astrophysical Journal* 88: 113.

Carrasco E and Parry I (1994) A method for determining the focal ratio degradation of optical fibres for astronomy. *Monthly Notices of the Royal Astronomical Society* 271: 1.

Courtes G (1982) *Instrumentation for Astronomy with Large Optical Telescope, IAU Colloq, 67*. In: Humphries CM (ed.) Astrophysics and Space Science Library, vol. 92, pp. 123. Dordrecht: Reidel.

Glazebrook K and Bland-Hawthorn J (2001) Microslit nod-shuffle spectrocopy: a technique for achieving very high densities of spectra. *Publications of the Astronomical Society of the Pacific* 113: 197.

Lee D, Haynes R, Ren D and Allington-Smith JR (2001) Characterisation of microlens arrays for astronomical spectroscopy. *PASP* 113: 1406.

McLean I (1997) *Electronic Imaging in Astronomy.* Chichester, UK: John Wiley and Sons Ltd.

Palmer C (2000) *Diffraction Grating Handbook*. Richardson Grating Laboratory, 4th edn. Rochester, NY.

Ridgway SH and Brault JW (1984) Astronomical Fourier transform spectroscopy revisited. *Ann. Rev. Astron. Astrophys.* 22.

Saha SK (2002) Modern optical astronomy: technology and impact of interferometry. *Review of Modern Physics* 74: 551.

Schroeder D (2000) *Astronomical Optics*, 2nd edn. San Diego, CA: Academic Press.

van Breugel W and Bland-Hawthorn J (eds) *Imaging the Universe in Three Dimensions*, ASP Conference Series, vol. 195.

Weitzel L, Krabbe A, Kroker H, *et al.* (1996) 3D: The next generation near-infrared imaging spectrometer. *Astron. & Astrophys. Suppl.* 119: 531.

Ellipsometry

J N Hilfiker, J A Woollam & Co., Inc., Lincoln, NE, USA

J A Woollam, University of Nebraska, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ellipsometry measures a change in polarization as light reflects from or transmits through a material structure. The polarization-change is represented as an amplitude ratio, Ψ, and a phase difference, Δ. The measured response is dependent on optical properties and thickness of each material. Thus, ellipsometry is primarily used to determine film thickness and optical constants. However, it is also applied to the characterization of composition, crystallinity, roughness, doping concentration, and other material properties associated with a change in optical response.

Interest in ellipsometry has grown steadily since the 1960s as it provided the sensitivity necessary to measure nanometer-scale layers used in microelectronics. Today, the range of applications has spread to basic research in physical sciences, semiconductor, data storage, flat panel display, communication, biosensor, and optical coating industries. This widespread use is due to increased dependence on thin films in many areas and the flexibility of ellipsometry to measure most material types: dielectrics, semiconductors, metals, superconductors, organics, biological coatings, and composites of materials.

This article provides a fundamental description of ellipsometry measurements along with the typical data analysis procedures. The primary applications of ellipsometry are also surveyed.

Light and Materials

Ellipsometry measurements involve the interaction between light and material.

Light

Light can be described as an electromagnetic wave traveling through space. For ellipsometry, it is adequate to discuss the electric field behavior in space and time, also known as polarization. The electric field of a wave is always orthogonal to the propagation direction. Therefore, a wave traveling along the z-direction can be described by its x- and y-components. If the light has completely random orientation and phase, it is considered unpolarized. For ellipsometry, we are interested in the case where the electric field follows a specific path and traces out a distinct shape at any point. This is known as polarized light. When two orthogonal light waves are in-phase, the resulting light will be linearly polarized (Figure 1a). The relative amplitudes determine the resulting orientation. If the orthogonal waves are 90° out-of-phase and equal in amplitude, the resultant light is circularly polarized (Figure 1b). The most general polarization is 'elliptical', which combines orthogonal waves of arbitrary amplitude and phase (Figure 1c). This is how ellipsometry gets its name.

Materials

Two values are used to describe optical properties, which determine how light interacts with a material.

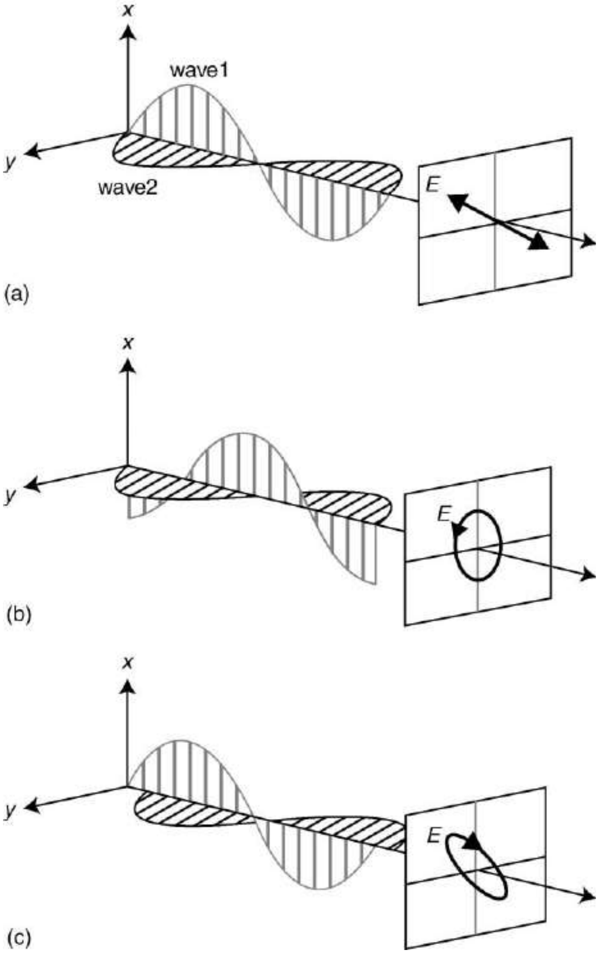

Figure 1 Orthogonal waves combine to demonstrate polarization: (a) linear; (b) circular; and (c) elliptical.

They are generally represented as a complex number. The complex refractive index ($\tilde{n}$) consists of the index (n) and extinction coefficient (k):

$$\tilde{n} = n + ik \qquad [1]$$

Alternatively, the optical properties can be represented as the complex dielectric function:

$$\tilde{\varepsilon} = \varepsilon_1 + i\varepsilon_2 \qquad [2]$$

with the following relation between conventions:

$$\tilde{\varepsilon} = \tilde{n}^2 \qquad [3]$$

The index describes the phase velocity for light as it travels through a material compared to the speed of light in vacuum, c:

$$v = \frac{c}{n} \qquad [4]$$

Light slows as it enters a material with higher index. Because frequency remains constant, the wavelength will shorten. The extinction coefficient describes the loss of wave energy to the material. It is related to the absorption coefficient, α, as:

$$\alpha = \frac{4\pi k}{\lambda} \qquad [5]$$

Light loses intensity in an absorbing material, according to Beer's Law:

$$I(z) = I(0)\,e^{-i\alpha z} \qquad [6]$$

Thus, the extinction coefficient relates how quickly light vanishes in a material. These concepts are demonstrated in Figure 2, where a light wave travels from air into two different materials of varying properties.

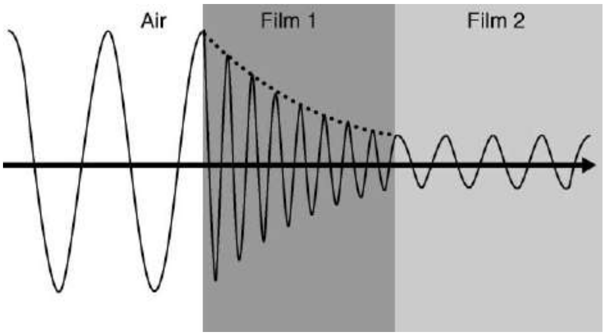

Figure 2 Wave travels from air into absorbing Film 1 and then transparent Film 2. The phase velocity and wavelength change in each material depending on index of refraction (Film 1: $n = 4$, Film 2: $n = 2$).

The optical constants for TiO_2 are shown in Figure 3 from the ultraviolet (UV) to the infrared (IR). The optical constants are wavelength dependent with absorption ($k > 0$) occurring in both UV and IR, due to different mechanisms that take energy from the light wave. IR absorption is commonly caused by molecular vibration, phonon vibration, or free-carriers. UV absorption is generally due to electronic transitions, where light provides energy to excite an electron to an elevated state. A closer look at the optical constants in Figure 3 shows that real and imaginary optical constants are not independent. Their shapes are mathematically coupled together through Kramers–Krönig consistency. Further details are covered later in this article.

Interaction Between Light and Materials

Maxwell's equations must remain satisfied when light interacts with a material, which leads to boundary conditions at the interface. Incident light will reflect and refract at the interface, as shown in Figure 4. The angle between the incident ray and sample normal

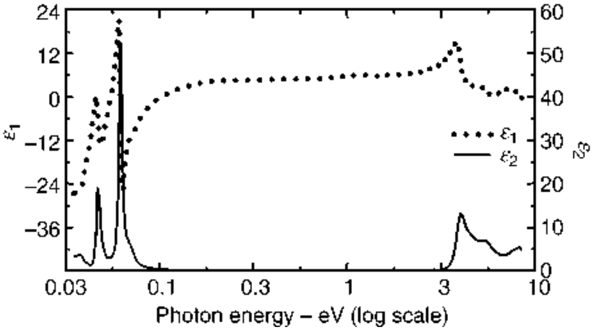

Figure 3 Complex dielectric function for TiO_2 film covering wavelengths from the infrared (small eV) to the ultraviolet (high eV).

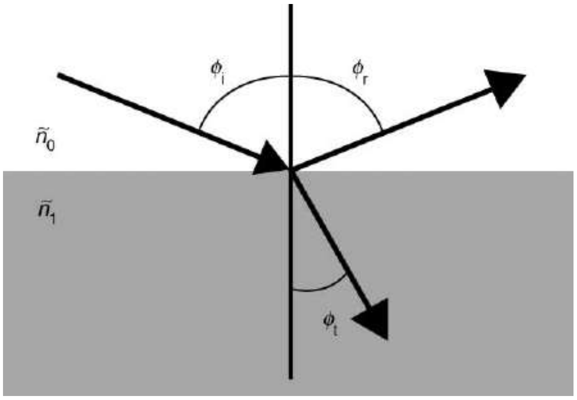

Figure 4 Light reflects and refracts according to Snell's law.

(ϕ_i) will be equal to the reflected angle, (ϕ_r). Light entering the material is refracted to an angle (ϕ_t) given by:

$$n_0 \sin(\phi_i) = n_1 \sin(\phi_t) \qquad [7]$$

The same occurs at every interface where a portion reflects and the remainder transmits at the refracted angle. This is illustrated in Figure 5. The boundary conditions provide different solutions for electric fields parallel and perpendicular to the sample surface. Therefore, light can be separated into orthogonal components with relation to the plane of incidence. Electric fields parallel and perpendicular to the plane of incidence are considered p- and s-polarized, respectively. These two components are independent for isotropic materials and can be calculated separately. Fresnel described the amount of light reflected and transmitted at an interface between materials:

$$r_s = \left(\frac{E_{0r}}{E_{0i}}\right)_s = \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t} \qquad [8a]$$

$$r_p = \left(\frac{E_{0r}}{E_{0i}}\right)_p = \frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_i \cos\theta_t + n_t \cos\theta_i} \qquad [8b]$$

$$t_s = \left(\frac{E_{0t}}{E_{0i}}\right)_s = \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_t \cos\theta_t} \qquad [8c]$$

$$t_p = \left(\frac{E_{0t}}{E_{0i}}\right)_p = \frac{2n_i \cos\theta_i}{n_i \cos\theta_t + n_t \cos\theta_i} \qquad [8d]$$

Thin film and multilayer structures involve multiple interfaces, with Fresnel reflection and transmission coefficients applicable at each. It is important to track the relative phase of each light component to correctly determine the overall reflected or transmitted beam. For this purpose, we define the film phase thickness as:

$$\beta = 2\pi\left(\frac{t_1}{\lambda}\right) n_1 \cos\theta_1 \qquad [9]$$

The superposition of multiple light waves introduces interference that is dependent on the relative phase of each light wave. Figure 5 illustrates the combination of light waves in the reflected beam and their corresponding Fresnel calculations.

Ellipsometry Measurements

For ellipsometry, primary interest is measurement of how p- and s-components change upon reflection or transmission relative to each other. In this manner, the reference beam is part of the experiment. A known polarization is reflected or transmitted from the sample and the output polarization is measured. The change in polarization is the ellipsometry measurement, commonly written as:

$$\rho = \tan(\Psi)\, e^{i\Delta} \qquad [10]$$

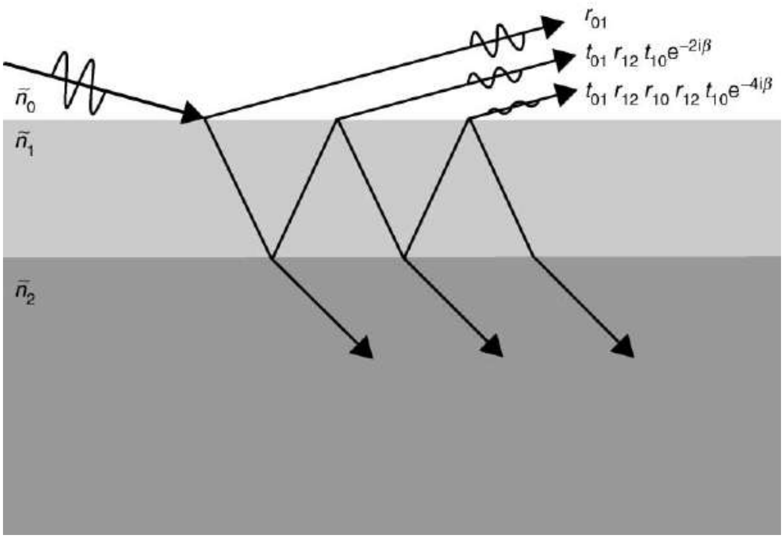
r_{01}
$t_{01} r_{12} t_{10} e^{-2i\beta}$
$t_{01} r_{12} r_{10} r_{12} t_{10} e^{-4i\beta}$
$\tilde{n}_0$
$\tilde{n}_1$
$\tilde{n}_2$

Figure 5 Light reflects and refracts at each interface, which leads to multiple beams in a thin film. Interference between beams depends on relative phase and amplitude of the electric fields. Fresnel reflection and transmission coefficients can be used to calculate the response from each contributing beam.

An example ellipsometry measurement is shown in Figure 6. The incident light is linear with both p- and s-components. The reflected light has undergone amplitude and phase changes for both p- and s-polarized light, and ellipsometry measures their changes.

The primary methods of measuring ellipsometry data all consist of the following components: light source, polarization generator, sample, polarization analyzer, and detector. The polarization generator and analyzer are constructed of optical components that manipulate the polarization: polarizers, compensators, and phase modulators. Common ellipsometer configurations include rotating analyzer (RAE), rotating polarizer (RPE), rotating compensator (RCE), and phase modulation (PME).

The RAE configuration is shown in Figure 7. A light source produces unpolarized light, which is sent through a polarizer. The polarizer passes light of a preferred electric field orientation. The polarizer axis is oriented between the p- and s-planes, such that both arrive at the sample surface. The linearly polarized light reflects from the sample surface, becoming elliptically polarized and travels through a continuously rotating polarizer (referred to as the analyzer). The amount of light allowed to pass will depend on the polarizer orientation relative to the electric field 'ellipse' coming from the sample. The detector converts light from photons to electronic signal to determine the reflected polarization. This information is used with the known input polarization to determine the polarization change caused by the sample reflection. This is the ellipsometry measurement of Ψ and Δ.

Single Wavelength Ellipsometry (SWE)

SWE uses a single frequency of light to probe the sample. This results in a pair of data, Ψ and Δ, used to determine up to two material properties. The optical design can be simple, low-cost, and accurate. Lasers are an ideal light source with well-known wavelength and significant intensity. The optical elements can be optimized to the single wavelength. However, there is relatively low information content (two measurement values). SWE is excellent for quick measurement of

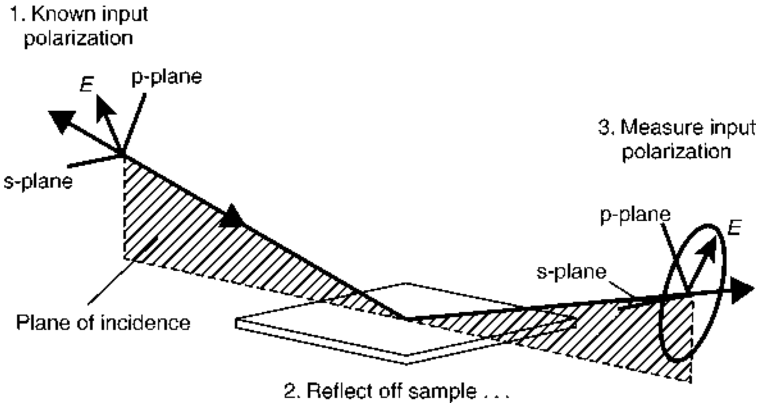
1. Known input polarization
E
p-plane
s-plane
Plane of incidence
3. Measure input polarization
p-plane
E
s-plane
2. Reflect off sample . . .

Figure 6 Typical ellipsometry configuration, where linearly polarized light is reflected from the sample surface and the polarization change is measured to determine the sample response.

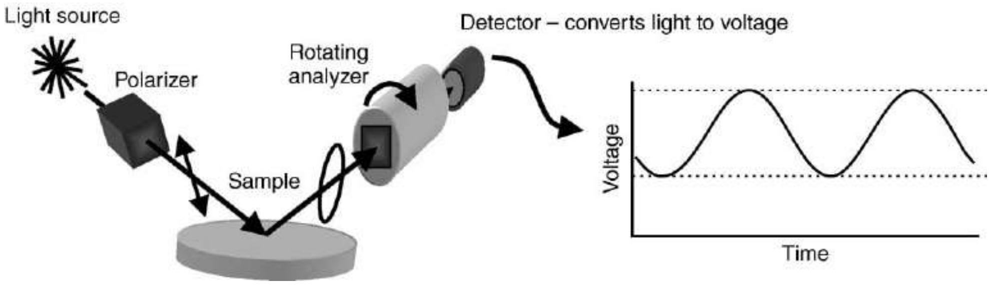

Figure 7 Rotating analyzer ellipsometer configuration uses a polarizer to define the incoming polarization and a rotating polarizer after the sample to analyze the outgoing light. The detector converts light to a voltage with the time-dependence leading to measurement of the reflected polarization.

nominally known films, like SiO_2 on Si. Care must be taken when interpreting unknown films, as multiple solutions occur for different film thickness. The data from a transparent film will cycle through the same values as the thickness increases. From Figure 5, it can be seen that this is related to interference between the multiple light beams. The refracted light may travel through different thicknesses to emerge in-phase with the first reflection, but is the returning light delayed by one wavelength or multiple wavelengths? Mathematically, the thickness cycle can be determined as:

$$D_\phi = \frac{\lambda}{2\sqrt{\tilde{n}_1^2 - \tilde{n}_0^2 \sin^2\phi}} \qquad [11]$$

This is demonstrated in Figure 8, where the complete thickness cycle is shown for SiO_2 on Si at 75° angle of incidence and 500 nm wavelength. A bare silicon substrate would produce data at the position of the star. As the film thickness increases, the data will move around this circle in the direction of the arrow. From eqn [11], the thickness cycle is calculated as 226 nm. Thus the cycle is completed and returns to the star when the film thickness reaches 226 nm. Any data point along the cycle represents a series of thicknesses which depend on how many times the cycle has been completely encircled. For example, the square represents a data point for 50 nm thickness. However, the same data point will represent 276 and 502 nm thicknesses. A second data point at a new wavelength or angle would help determine the correct thickness.

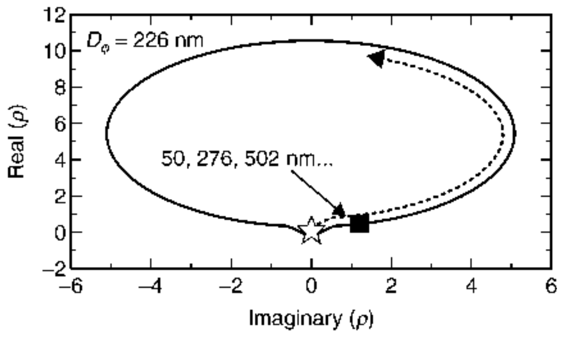

Figure 8 Data from a single wavelength will cycle through the same points as the film thickness increases. The star represents the starting point, with no SiO_2 film on Si. As the thickness increases, the data follow the cycle in the direction of the arrow. After the cycle is complete (thickness = D_ϕ), the data repeat the cycle. Thus, any point along the curve represents multiple possible thicknesses – example shown by a square at 50 nm thickness, which is also equal to 276 and 502 nm.

Spectroscopic Ellipsometry (SE)

Spectroscopic measurements solve the 'period' problem discussed for SWE. Data at surrounding wavelengths insure the correct thickness is determined, even as there remain multiple solutions at one wavelength in the spectrum. This is demonstrated in Figure 9, showing the spectroscopic data for the first three thickness results. The SE data oscillations clearly distinguish each thickness. Thus, a single thickness solution remains to match data at multiple wavelengths.

SE provides additional information content for each new wavelength. While the film thickness will remain constant regardless of wavelength, the optical constants will change across the spectrum. The optical constant shape contains information regarding the material microstructure, composition, and more. Different wavelength regions will provide the best information for each different material property. For this reason, SE systems have been developed to cover very wide spectral regions from the UV to the IR.

Spectroscopic measurements require an additional component, a method of wavelength selection. The most common methods include the use of monochromators, Fourier-transform spectrometers, and gratings or prisms with detection on a diode array. Monochromators are typically slow as they sequentially scan through wavelengths. Spectrometers and diode arrays allow simultaneous measurement at multiple wavelengths. This has become popular with the desire for high-speed SE measurements.

Variable Angle Ellipsometry

Ellipsometry measurements are typically performed at oblique angles, where the largest changes in polarization occur. The Brewster angle, defined as:

$$\tan(\phi_B) = \frac{n_1}{n_0} \qquad [12]$$

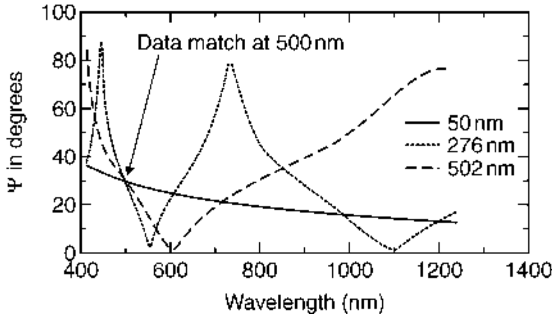

Figure 9 Three spectroscopic ellipsometry measurements that match at 500 nm wavelength, but are easily distinguishable as a function of wavelength.

gives the angle where reflection of p-polarized light goes through a minimum. Ellipsometry measurements are usually near this angle, which can vary from 55° for low-index dielectrics to 75° or 80° for semiconductors and metals.

It is also common to measure at multiple angles of incidence in ellipsometry. This allows additional data to be collected for the same material structure under different conditions. The most important change introduced by varying the angle is the change in path length through the film as the light refracts at a different angle. Multiple angles do not always add new information about a material structure, but the extra data builds confidence in the final answer.

In Situ

Within the last decade, it has become common to employ optical diagnostics during processing. *In situ* ellipsometry allows the optical response to be monitored in real-time. This has also led to feedback control for many processes. While *in situ* is generally restricted to a single angle of incidence, there is a distinct advantage to collecting data at different 'times' during processing to get unique 'glimpses' of the sample structure as it changes.

In situ SE is commonly applied to semiconductors. Conventional SE systems measure from the UV to NIR, which matches the spectral region where semiconductors absorb due to electronic transitions. The shape and position of the absorption can be very sensitive to temperature, composition, crystallinity, and surface quality, which allows SE to closely monitor these properties. *In situ* SE has also been used to monitor numerous material types, including metals and dielectrics. A promising *in situ* application is for multilayer optical coatings, where thickness and index can be determined in real-time to allow correction of future layers, to optimize optical performance.

Data Analysis

Ellipsometry measures changes in polarization. However, it is used to determine the material properties of interest, such as film thickness and optical constants. In the case of a bulk material, the equations derived for a single reflection can be directly inverted to provide the 'pseudo' optical constants from the ellipsometry measurement, ρ:

$$\langle \tilde{\varepsilon} \rangle = \sin^2(\phi)\left[1 + \tan^2(\phi)\left(\frac{1-\rho}{1+\rho}\right)\right] \qquad [13]$$

This equation assumes there are no surface layers of any type. However, there is typically a surface oxide or roughness for any bulk material and the direct inversion would incorporate these into the bulk optical constants. The more common procedure used to deduce material properties from ellipsometry measurements follows the flowchart in Figure 10. Regression analysis is required because an exact equation cannot be written. Often the answer is overdetermined with hundreds of experimental data points for a few unknowns. Regression analysis allows all of the measured data to be included when determining the solution.

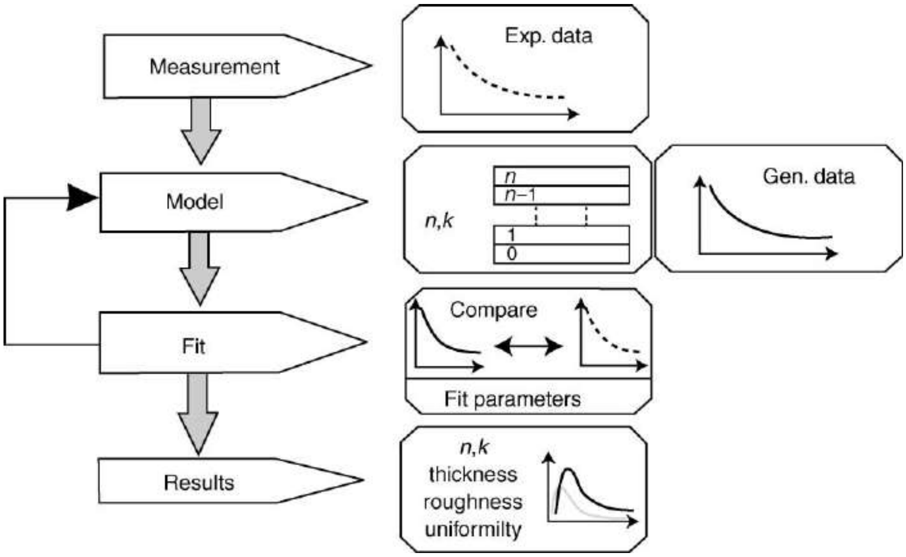

Figure 10 Flowchart for ellipsometry data analysis.

Data analysis proceeds as follows. After the sample is measured, a model is constructed to describe the sample. The model is used to calculate the predicted response from Fresnel's equations. Thus, each material must be described with a thickness and optical constants. If these values are not known, an estimate is given for the purpose of the preliminary calculation. The calculated values are compared to experimental data. Any unknown material properties can then be varied to improve the match between experiment and calculation. The number of unknown properties should not exceed the information content contained in the experimental data. For example, a single-wavelength ellipsometer produces two data points (Ψ,Δ) which allows a maximum of two material properties to be determined. Finding the best match between the model and experiment is typically done through regression. An estimator, like the mean squared error (MSE), is used to quantify the difference between curves. The unknown parameters are allowed to vary until the minimum MSE is reached.

Care must be taken to ensure the best answer is found, corresponding to the lowest MSE. For example, Figure 11a shows the MSE curve versus film thickness for a transparent film on silicon. There are multiple 'local' minima, but the lowest MSE value occurs at a thickness of 749 nm. This corresponds to the correct film thickness. It is possible that the regression algorithm will mistakenly fall into a 'local' minima, depending on the starting thickness and the MSE structural conditions. Comparing the results by eye for the lowest MSE and a local minima easily distinguish the true minima (see Figures 11b and c).

Ellipsometry Characterization

The two most common material properties studied by ellipsometry are film thickness and optical constants. In addition, ellipsometry can characterize material properties that affect the optical constants. Crystallinity, composition, resistivity, temperature, and molecular orientation can all affect the optical constants and in turn be measured by ellipsometry. This section details many of the primary applications important to ellipsometry.

Film Thickness

The film thickness is determined by interference between light reflecting from the surface and light traveling through the film. Depending on the relative phase of the rejoining light to the surface reflection, there can be constructive or destructive interference.

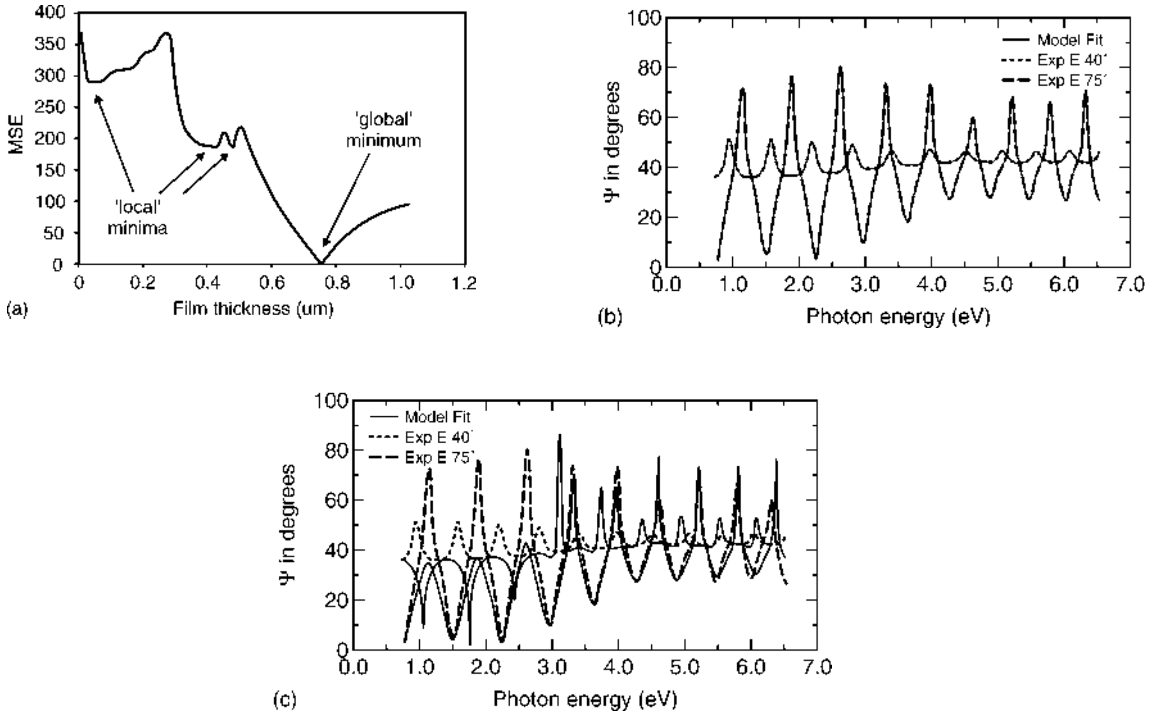

Figure 11 (a) MSE curve versus thickness shows the 'global' minimum where the best match between model and experiment occurs, and 'local' minima that may be found by the regression algorithm, but do not give the final result. (b) The experimental data and corresponding curves generated for the model at the 'global' minimum. (c) Similar curves at the 'local' minimum near 0.45 microns thickness is easily distinguishable as an incorrect result.

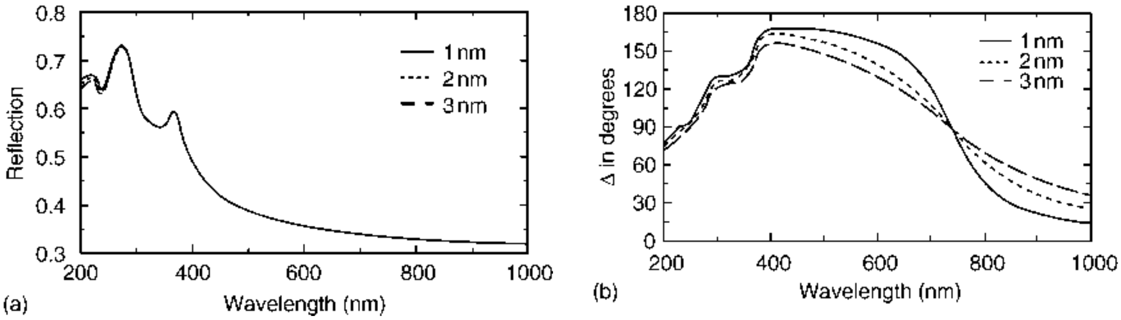

Figure 12 (a) Reflected intensity and (b) ellipsometric delta for three thin oxides on silicon show the high sensitivity of Δ to nanometer-scale films not available from the intensity measurement.

The interference involves both amplitude and phase information. The phase information from Δ is very sensitive to films down to submonolayer thickness. Figure 12 compares reflected intensity and ellipsometry for the same series of thin SiO_2 layers on Si. There are large variations in Δ, while the reflectance for each film is nearly the same.

Ellipsometry is typically used for films with thickness ranging from sub-nanometers to a few microns. As films become greater than several tens of microns thick, it becomes increasingly difficult to resolve the interference oscillations, except with longer infrared wavelengths, and other characterization techniques become preferred.

Thickness measurements also require a portion of the light to travel through the entire film and return to the surface. If the material is absorbing, thickness measurements by optical instruments will be limited to thin, semi-opaque layers. This limitation can be circumvented by measuring in a spectral region where there is lower absorption. For example, an organic film may strongly absorb UV and IR light, but remain transparent at mid-visible wavelengths. For metals, which strongly absorb at all wavelengths, the maximum layer for thickness determination is typically ~100 nm.

Optical Constants

Thickness measurements are not independent of the optical constants. The film thickness affects the path length of light traveling through the film, but the index determines the phase velocity and refracted angle. Thus, both contribute to the delay between surface reflection and light traveling through the film. Both n and k must be known or determined along with the thickness to get the correct results from an optical measurement.

The optical constants for a material will vary for different wavelengths and must be described at all wavelengths probed with the ellipsometer. A table of optical constants can be used to predict the response at each wavelength. However, it is less convenient to adjust unknown optical constants on a wavelength-by-wavelength basis. It is more advantageous to use all wavelengths simultaneously. A dispersion relationship often solves this problem, by describing the optical constant shape versus wavelength. The adjustable parameters of the dispersion relationship allow the overall optical constant shape to match the experimental results. This greatly reduces the number of unknown 'free' parameters compared to fitting individual n, k values at every wavelength.

For transparent materials, the index is often described using the Cauchy or Sellmeier relationship. The Cauchy relationship is typically given as:

$$n(\lambda) = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \qquad [14]$$

where the three terms are adjusted to match the refractive index for the material. The Sellmeier relationship enforces Kramers–Krönig (KK) consistency, which ensures the optical dispersion retains a realistic shape. The Cauchy is not constrained by KK consistency and can produce unphysical dispersion. The Sellmeier relationship can be written as:

$$\varepsilon_1 = \frac{A\lambda^2\lambda_0^2}{(\lambda^2 - \lambda_0^2)} \qquad [15]$$

Absorbing materials will often have a transparent wavelength region that can be modeled with the Cauchy or Sellmeier relationships. However, the absorbing region must account for both real and imaginary optical constants. Many dispersion relationships use oscillator theory to describe the absorption for materials, which include the Lorentz, Harmonic, and Gaussian oscillators. They all share similar attributes, where the absorption features are described with an amplitude, broadening, and center energy (related to frequency of light).

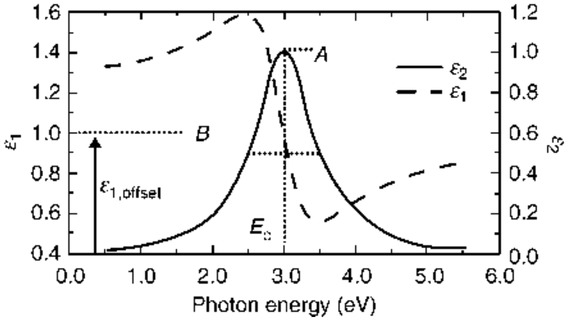

Figure 13 Lorentz oscillator illustrating the primary oscillator parameters: amplitude (*A*), broadening (*B*), and center energy (E_c) to describe the imaginary dielectric function shape and an offset ($\varepsilon_{1,\text{offset}}$) to help match the real component after KK transformation has defined its shape.

Kramers–Krönig consistency is used to calculate the shape of the real component after the imaginary behavior is described by the oscillator. An offset to the real component is added to account for extra absorption outside the measured spectral region. The Lorentz oscillator can be written as:

$$\tilde{\varepsilon} = \varepsilon_{1,\text{offset}} + \frac{AE_c}{E_c^2 - E^2 - \mathrm{i}BE} \qquad [16]$$

where the parameters for amplitude (*A*), broadening (*B*), center energy (E_c), and offset ($\varepsilon_{1,\text{offset}}$) are also shown in Figure 13 for a typical Lorentz oscillator. The energy, *E*, is related to the frequency of a wave, ν:

$$E = h\nu \cong \frac{1240}{\lambda_{\text{nm}}} \qquad [17]$$

where *h* is Planck's constant and the wavelength, λ, is given in nanometers. More advanced dispersion models, like the Tauc–Lorentz and Cody–Lorentz, will include terms to describe the bandgap energy.

Mixing Materials

When two or more materials are mixed on a microscopic level, an effective medium approximation (EMA) may be used to determine the resulting optical constants. In the case of the Bruggemann EMA, which is commonly used, the mixture optical constants ($\tilde{\varepsilon}_{\text{eff}}$) relate to those of the individual materials, as:

$$f_a \frac{\varepsilon_a - \varepsilon_{\text{eff}}}{\varepsilon_a + 2\varepsilon_{\text{eff}}} + f_b \frac{\varepsilon_b - \varepsilon_{\text{eff}}}{\varepsilon_b + 2\varepsilon_{\text{eff}}} = 0 \qquad [18]$$

This can be interpreted as small particles of material A suspended in host material B. The length-scale for mixed particles must satisfy certain electromagnetic equations: typically smaller than one tenth the wavelength of light. In practice, EMA theory is useful for studying very thin surface roughness or interfacial intermixing of materials. These cases are both generally approximated by mixing the two surrounding materials in equal portions. In the case of surface roughness, the second material is void ($n = 1$). EMA theory has also been extended for application to porous materials where the directional dependence of the inclusions is handled mathematically.

Crystallinity

Semiconductors such as Si, are widely used materials, but their properties depend strongly on crystallinity. The UV absorption features in crystalline silicon are broadened and shifted as the material becomes more amorphous. This change in optical properties related to the degree of crystallinity has been used to advantage with ellipsometry measurements to monitor semiconductors and other films. Polysilicon films are used in both the display and semiconductor industry. The degree of crystallinity varies for different process conditions and can be monitored optically to ensure consistent material properties.

Composition

The composition of many alloys will affect the optical constants. The strongest changes often occur in the absorbing region, with shifts in position and amplitude of absorption features. For example, the electronic transitions in $\mathrm{Hg}_{1-x}\mathrm{Cd}_x\mathrm{Te}$ move to higher energy with Cd concentration increase (Figure 14). This material is used for IR detectors that require precise control of composition. Spectroscopic ellipsometry performs this task in real-time with instant feedback to correct the composition during processing. Other applications include AlGaN and InGaN for optoelectronics.

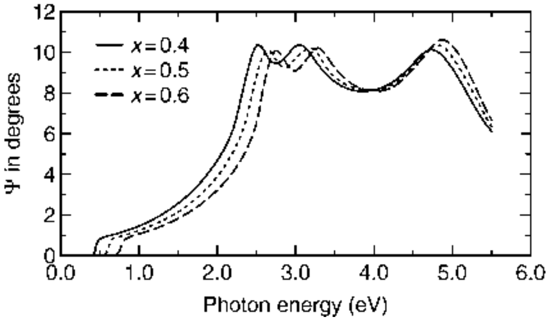

Figure 14 Optical properties for $\mathrm{Hg}_{1-x}\mathrm{Cd}_x\mathrm{Te}$ vary with changes in composition.

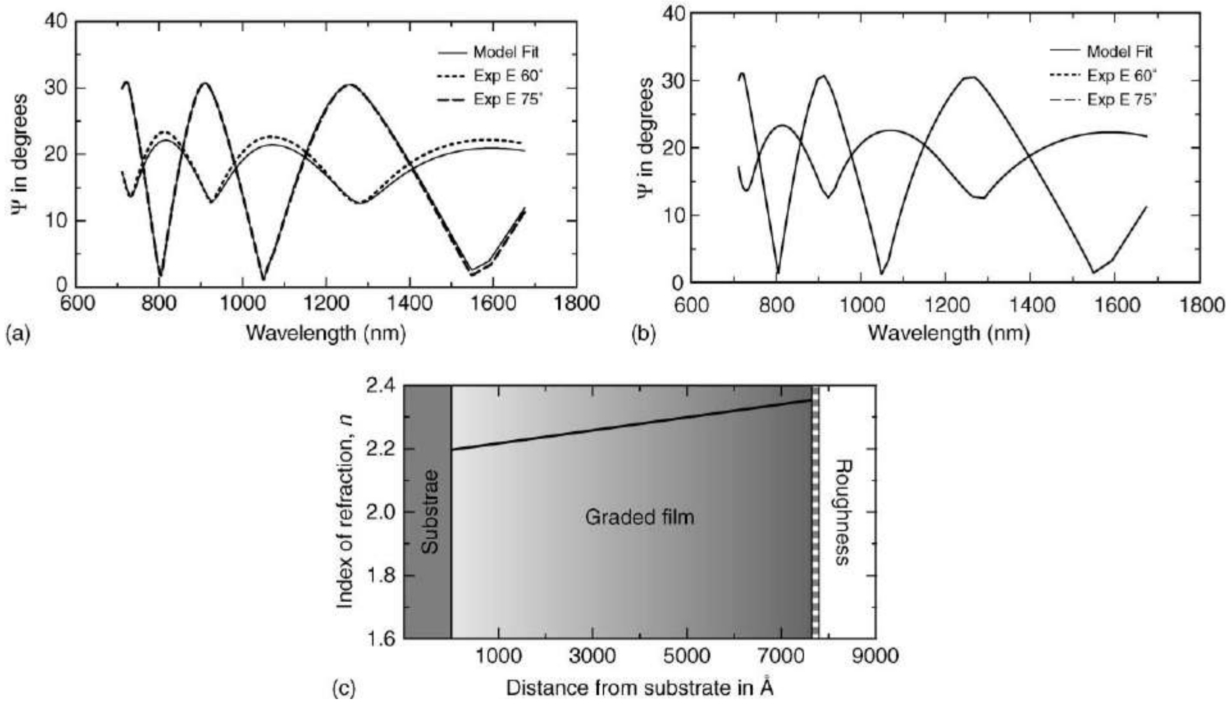

Figure 15 Experimental ellipsometry data and corresponding fit when the model is described as (a) a homogeneous single-layer and (b) a graded film with index variation through the film. (c) The best model to match subtleties in experimental data allows the index to increase toward the surface of the film, with a thin roughness layer on the surface.

Doping Concentration

Dopants in a material will introduce absorption in the infrared due to free-carriers. Spectroscopic ellipsometry measurements at long wavelengths can characterize this optical absorption, thus characterizing the doping concentration. This is common for highly doped semiconductors and transparent conductors such as indium tin oxide (ITO).

Optical Variation (Grading)

Many thin film properties change vertically throughout the film (along the direction perpendicular to the surface). This is most often induced by processing conditions, whether intentional or unintentional. Figure 15a shows the fit to experimental spectroscopic ellipsometry data taken from a single-layer film when it is modeled as a homogeneous layer. To improve the agreement between experimental and model-generated curves, the index was allowed to vary in a linear manner throughout the film. The best fit is shown in Figure 15b, where the model includes the index variation and a thin surface roughness layer. The sample structure is demonstrated in Figure 15c.

Optical Anisotropy

Many materials are optically anisotropic; i.e., their optical properties vary in different film directions. Typical ellipsometry measurements assume no cross-coupling between p- and s-polarizations. This cannot be assumed with anisotropic materials, which has led to ‘generalized’ ellipsometry measurements. Generalized ellipsometry measures additional information regarding the p- to s- and s- to p- conversion upon reflection. This allows characterization of anisotropic materials, which contain directionally dependent optical constants.

Conclusions

Ellipsometry is a common optical technique for measuring thin films and bulk materials. It relies on the polarization changes due to reflection or transmission from a material structure to deduce material properties, like film thickness and optical constants. This technique continues to develop as the requirement for thin film characterization increases.

See also

Geometrical Optics: Lenses and Mirrors; Prisms. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Measurement of Optical Properties of Solids.

Polarization: Introduction; Matrix Analysis. **Semiconductor Physics:** Band Structure and Optical Properties. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Aspnes DE (1985) The accurate determination of optical properties by ellipsometry. In: Palik ED (ed.) *Handbook of Optical Constants of Solids*, pp. 89–112. Orlando FL: Academic Press.

Azzam RMA and Bashara NM (1987) *Ellipsometry and Polarized Light*. Amsterdam, The Netherlands: Elsevier Science B.V.

Boccara AC, Pickering C and Rivory J (eds) (1993) *Spectroscopic Ellipsometry*. Amsterdam: Elsevier Publishing.

Collins RW, Aspnes DE and Irene EA (eds) (1998) *Proceedings from the Second International Conference on Spectroscopic Ellipsometry*. Thin Solid Films, vols. 313–314.

Gottesfeld S, Kim YT and Redondo A (1995) Recent applications of ellipsometry and spectroellipsometry in electrochemical systems. In: Rubinstein I (ed.) *Physical Electrochemistry: Principles, Methods, and Applications*. New York: Marcel Dekker.

Herman IP (1996) *Optical Diagnostics for Thin Film Processing*, pp. 425–479. San Diego, CA: Academic Press.

Johs B, Woollam JA, Herzinger CM, *et al.* (1999) Overview of variable angle spectroscopic ellipsometry (VASE), Part II: Advanced applications. *Optical Metrology*, vol. CR72, pp. 29–58. Bellingham, Washington: SPIE, Bellingham.

Johs B, Hale J, Ianno NJ, *et al.* (2001) Recent developments in spectroscopic ellipsometry for *in situ* applications. In: Duparré A and Singh B (eds) *Optical Metrology Roadmap for the Semiconductor, Optical, and Data Storage Industries II*, vol. 4449, pp. 41–57. Bellingham, Washington: SPIE.

Roseler A (1990) *Infrared Spectroscopic Ellipsometry*. Berlin: Akademie-Verlag.

Rossow U and Richter W (1996) Spectroscopic ellipsometry. In: Bauer G and Richter W (eds) *Optical Characterization of Epitaxial Semiconductor Layers*, pp. 68–128. Berlin: Springer-Verlag.

Tompkins HG (1993) *A User's Guide to Ellipsometry*. San Diego, CA: Academic Press.

Tompkins HG and Irene EA (eds) (in press) *Handbook of Ellipsometry*. New York: William Andrew Publishing.

Tompkins HG and McGahan WA (1999) *Spectroscopic Ellipsometry and Reflectometry*. New York: John Wiley & Sons, Inc.

Woollam JA (2000) Ellipsometry, variable angle spectroscopic. In: Webster JG (ed.) *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 109–116. New York: John Wiley & Sons.

Woollam JA and Snyder PG (1992) Variable angle spectroscopic ellipsometry. In: Brundle CR, Evans CA and Wilson S (eds) *Encyclopedia of Materials Characterization: Surfaces, Interfaces, Thin Films*, pp. 401–411. Boston, MA: Butterworth-Heinemann.

Woollam JA, Johs B, Herzinger CM, *et al.* (1999) Overview of variable angle spectroscopic ellipsometry (VASE), Part I: Basic theory and typical applications. *Optical Metrology*, vol. CR72, pp. 3–28. Washington: SPIE, Bellingham.

Photometry

J Schanda, University of Veszprém, Veszprém, Hungary

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fundamentals of Photometry

Photometry is defined by the International Commission on Illumination (internationally known as CIE from the abbreviation of its French name: Commission Internationale de l'Eclairage) as 'measurement of quantities referring to radiation as evaluated according to a given spectral efficiency function, e.g., $V(\lambda)$ or $V'(\lambda)$.' A note to the above definition states that in many languages it is used in a broader sense, covering the science of optical radiation measurement. We will restrict our treatise to the above fundamental meaning of photometry.

Under spectral efficiency function we understand the spectral luminous efficiency function of the human visual system. The internationally agreed symbols are $V(\lambda)$, for photopic vision (daylight conditions) and $V'(\lambda)$ for scotopic vision (nighttime conditions). CIE standardized the $V(\lambda)$ function in 1924 and the $V'(\lambda)$ function in 1951; their spectral distribution is shown in Figure 1. The $V(\lambda)$ function was determined using mainly the so-called flicker photometric technique, where the light of a reference stimulus and of a test stimulus of varying wavelengths are alternatively shown to the human observer at a frequency at which the observer is already unable to perceive the difference in color (due to the different wavelength of the two radiations), but still perceives a

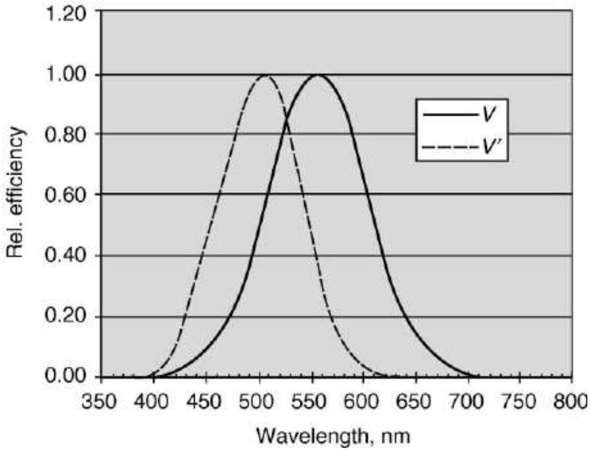

Figure 1 Spectral efficiency functions of the human eye under photopic $V(\lambda)$ and scotopic $V'(\lambda)$ conditions (see CIE The basis of physical photometry. CIE 18.2:1983).

flicker sensation if the luminances of the two stimuli are different. Adjusting the radiance of the test stimulus, we can reach the situation of minimum flicker sensation. In this case we state that we set the two stimuli to equal luminance. Luminance is the quantity that the human observer perceives as brightness (in the case of near white stimuli, see below); radiance is the physical counterpart measured in W m^{-2} sr^{-1}. For the $V'(\lambda)$ function one can just project the two stimuli side by side, as under scotopic adaptation we cannot distinguish colors, only brightness differences are perceived, thus one can set equal scotopic luminance for the two stimuli by adjusting the brightness of the test stimulus until it agrees with that of the reference stimulus.

Illuminating engineering was mainly interested in describing the effect of near white light, and therefore a psychophysical correlation was selected that described the perceived brightness sensation reasonably well. For near white stimuli, additivity holds, i.e., if A, B, C, and D are four stimuli, and the A stimulus matches the B stimulus and the C stimulus matches the D stimulus, then the superposition of the A and C stimuli matches the superposition of the B and D stimuli. The ‘match’ word is used here to describe stimuli that produce the same perception. It could be shown that for these stimuli, if the radiance is weighted with the $V(\lambda)$ function, the constructed luminances will be equal, i.e:

if $\quad L_\mathrm{A} = L_\mathrm{B}$ and $L_\mathrm{C} = L_\mathrm{D}$

then $\quad L_\mathrm{A} + L_\mathrm{C} = L_\mathrm{B} + L_\mathrm{D}$, or $L_\mathrm{A} + L_\mathrm{D} = L_\mathrm{B} + L_\mathrm{C}$

where

$$L_\mathrm{x} = K_\mathrm{m} \int_{380\ \mathrm{nm}}^{780\ \mathrm{nm}} S_{\lambda \mathrm{x}} V(\lambda) \mathrm{d}\lambda \qquad [1]$$

and x refers to A, or B, or C, or D, and $S_{\lambda\mathrm{x}}$ is the spectral radiance distribution that produces the stimulus

$$S_\lambda = \frac{\mathrm{d}S(\lambda)}{\mathrm{d}\lambda}$$

where $S(\lambda)$ is the spectral radiance. K_m is the maximum value of the luminous efficacy of radiation, its value is 683 lm W^{-1} (see the discussion of luminous flux in the sub-section on photometric quantities). To be precise the integration should go from 0 nm to infinity, but it is usual to define the lower and upper wavelength limits of the visible spectrum as 380 nm and 780 nm. $V(\lambda)$ is defined between 360 nm and 830 nm, and $V'(\lambda)$ between 380 nm and 780 nm, see Figure 1.

We have to stress that the concept of luminance – and the entire system of the present-day photometry – is not to quantify brightness perception; it is only a reasonable approximation for near white stimuli. For colored lights a brightness–luminance discrepancy exists (called the Helmhotz–Kohlrausch effect: saturated colors look brighter than predicted by luminance). Luminance is, however, a good description for the visibility of fine details, thus it is a good concept for lighting calculations.

Photometry has a unique situation in the SI system of units: the candela (cd) is a base unit of the SI system, the only one that is connected to psychophysical phenomena. By 1979, in the definition of the 16th General Conference of Weights and Measures, the candela was traced back to radiation quantities; nevertheless it was kept as the base unit of photometry:

> The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of 1/683 W sr^{-1}. (540×10^{12} Hz corresponds approximately to 555 nm.)

To be able to calculate a photometric quantity of radiation of other wavelengths, one of the psychophysical functions $V(\lambda)$ or $V'(\lambda)$ has to be used and eqn [1] applied. We will see later that besides $V(\lambda)$ and $V'(\lambda)$ in modern applications, a number of further spectral luminous efficiency functions might be used.

Photometric Quantities

As discussed in the previous section, from the point of view of vision, the most important quantity is luminance. Looking, however, at the definition of the base unit, it is obvious that from the physical point of view the definition of a quantity corresponding to radiant power, measured in watts, can help to bridge the gap between photometry and radiometry: luminous flux, measured in lumens (lm), with the symbol Φ, is defined by eqn [1], with $S_\lambda(\lambda)$ inserted in W m^{-1}.

Based on this more practical quantity of luminous flux and its unit, lm, the different quantities used in photometry can be built up as follows:

$$\textit{Luminous flux}: \quad \Phi = K_m \int_0^\infty \frac{\mathrm{d}\Phi_e(\lambda)}{\mathrm{d}\lambda} V(\lambda)\mathrm{d}\lambda \qquad [2]$$

where $\Phi_e(\lambda)$ is the radiant flux measured in W,

$$\Phi_{e,\lambda} = \frac{\mathrm{d}\Phi_e(\lambda)}{\mathrm{d}\lambda}$$

is the spectral distribution (or spectral concentration) of the radiant flux (W m^{-1}). $K_m = 683$ lm W^{-1}, is the maximum value of the luminous efficacy of radiation for $\lambda_m \approx 555$ nm.

Similar equations can be written for scotopic luminous flux, with $V(\lambda)$ exchanged with $V'(\lambda)$, where $K'_m = 1700$ lm W^{-1} for $\lambda_m \approx 507$ nm. All further quantities can be defined both for photopic and scotopic conditions. Here we write them only as photopic quantities.

$$\textit{Luminous intensity}: \quad I = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega} \qquad [3]$$

where $\mathrm{d}\Phi$ is the luminous flux traveling in an elementary solid angle $\mathrm{d}\Omega$, assuming a point source (see Figure 2). The unit of luminous intensity is the candela (cd = lm sr^{-1}).

$$\textit{Luminance}: \quad L = \frac{\partial^2\Phi}{\partial A \cos\Theta\, \partial\Omega} \qquad [4]$$

where $\mathrm{d}\Phi$ is the luminous flux traveling in an elementary solid angle $\mathrm{d}\Omega$, $\mathrm{d}A$ is the elementary surface area emitting the radiation, and Θ is the angle between the normal of $\mathrm{d}A$ and the direction of luminance measurement (see Figure 3). The unit of luminance is cd m^{-2} (in some older literature called nit).

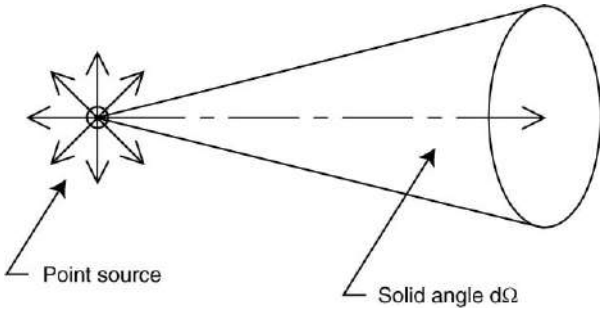

Figure 2 Concept of point source, solid angle, and luminous intensity.

$$\textit{Illuminance}: \quad E = \frac{\mathrm{d}\Phi}{\mathrm{d}A} \qquad [5]$$

where $\mathrm{d}\Phi$ is the luminous flux incident on the $\mathrm{d}A$ element of a surface (see Figure 4). The unit of illuminance is the lux (lx = lm m^{-2}).

A remark on the use of SI units and related quantities

No prefixes can be added to the SI units, thus irrespective whether one measures luminance based

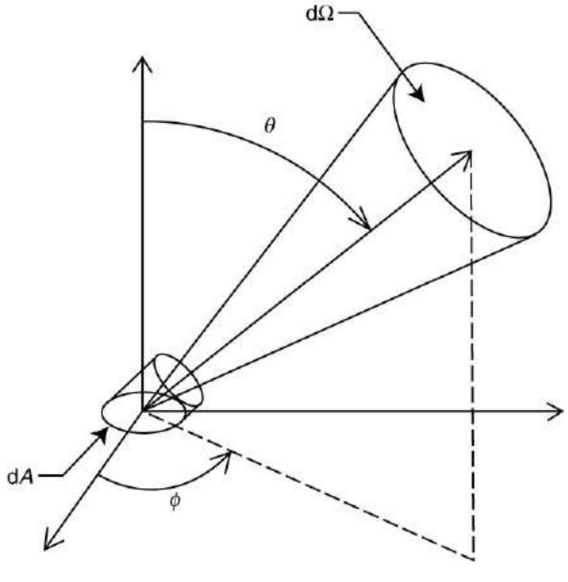

Figure 3 Geometry for the definition of luminance.

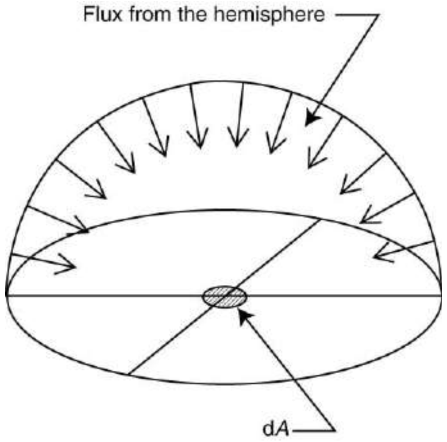

Figure 4 Illuminance is the total luminous flux per unit area incident at a point coming from a hemispherical solid angle.

on the photopic $V(\lambda)$ function, or the scotopic $V'(\lambda)$ function, the so-determined photopic luminance or scotopic luminance is measured in cd m^{-2} (no photopic or scotopic lumen, candela, etc., exist!). This often creates confusion, because only for a 555 nm monochromatic radiation will a 1 cd m^{-2} photopic or scotopic luminance produce equal visual sensation, for every other wavelength the two are not commensurable. One can only relate photopic measurement results to photopic ones, and scotopic measurement results to scotopic ones.

Concepts of Advanced Photometry

The quantities and units, as described in the previous section, are the quantities and units internationally agreed by the Meter Convention and the International Standards Organization (ISO). Modern photometric applications need, however, some further quantities and weighting functions. Thus CIE, the international organization for the development of standards in the field of optical radiation measurements, has defined a series of further weighting functions and quantities. The most important ones are the following:

$V_M(\lambda)$ function

In the blue part of the spectrum (below 460 nm) the values of the $V(\lambda)$ function turned out to be too low. For decades this was only of concern for the vision research community, but with the introduction of blue LEDs and other short wavelength emitting sources (e.g., blue channel of displays) this shortcoming of the $V(\lambda)$ function became of practical importance. The 'CIE 1988 2° spectral luminous efficiency function for photopic vision' corrects this anomaly (see Figure 5).

$V_{10}(\lambda)$ function

The official $V(\lambda)$ function is valid only for foveal vision (i.e., for targets that are smaller than 4° of visual angle). The foveal area of the retina is covered with a yellow pigmented layer (macula lutea) that absorbs in the blue part of the spectrum, at larger visual angles this screening is not effective anymore. The $V_{10}(\lambda)$ function was determined for a visual angle of 10° (see Figure 5). Its international recommendation is still under consideration at the time of writing this article and if accepted, it will be recommended for targets seen at approximately 10° off-axis, e.g., for measuring the photometric properties of traffic signs and signals, where the driver has to observe the information in the periphery of his or her eye.

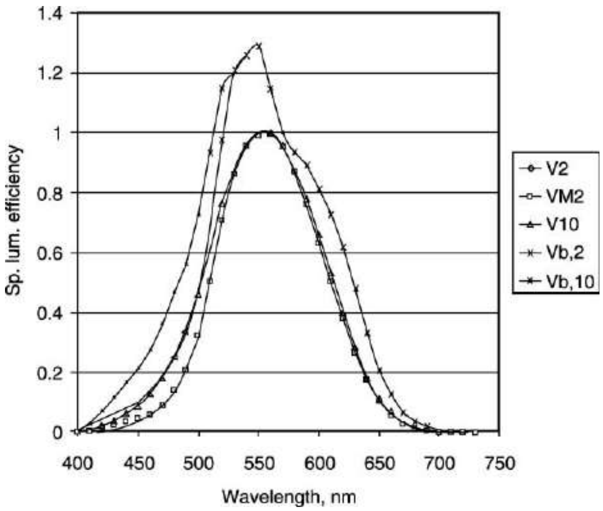

Figure 5 Spectral luminous efficiency (SPL) functions defined or under consideration for international adoption: V2: standard $V(\lambda)$ function; VM2: CIE 1988 $V_M(\lambda)$ function, this is equivalent to the brightness SPL for a point source; V10: 10° visual field SPL, Vb,2: 2° visual field brightness SPL, $V_{b,2}(\lambda)$; Vb,10: 10° visual field brightness SPL, $V_{b,10}(\lambda)$.

Brightness matching functions

As already mentioned in the Introduction, luminance is not a good correlate of brightness, which is a human perception. To find a better correlation of brightness the problem to be addressed is that brightness is a nonadditive phenomenon, i.e., in eqn [1] one cannot add (integrate) the monochromatic radiations to get a brightness correlation for a nonmonochromatic radiation. Brightness evaluating spectral luminous efficiency functions can be used only to compare monochromatic radiations. The CIE has compiled such functions for point sources, 2° and 10° visual field sizes. Figure 5 also shows these functions.

Due to the fact that the brightness perception is nonadditive in respect of the stimuli that produce it, no brightness photometry can be built that uses equations, as shown in eqn [1]. For brightness description we have to rely on the concept of equivalent luminance, a term the definition of which has recently been updated.

Equivalent luminance

Of a field of given size and shape, for a radiation of arbitrary relative spectral distribution L_{eq} : Luminance of a comparison field in which monochromatic radiation of frequency 540×10^{12} Hz has the same brightness as the field considered under the specified photometric conditions of measurement; the comparison field must have a specified size and shape which may be different from that of the field considered.

To build an instrument that measures this quantity is a real challenge for the future.

Advanced Use of Photometry

Based on above newly defined quantities, several attempts are under way to extend the usefulness of photometry in designing the human visual environment. The eye is an optical system and as in every such system, the depth of focus and the different aberrations of the system will decrease with decreasing pupil size. Pupil size will decrease with increasing illumination, and in the case of constant luminance with higher content of short wavelength radiation. Thus, there exists a school of researchers who advocate that increased blue content in the light has beneficial effects on vision, and one should extend the classical photopic-based photometry with a scotopic-based one to properly describe the visual effect of lighting.

Other investigations are concerned about the visibility at low light levels, the range used in street lighting (from about a few thousands of a candela per square meter to about a few candelas per square meter, according to one definitions: 10^{-3}cd m^{-2}–3 cd m^{-2}). In this mesopic range both rods and cones are contributing to vision, and this changes with lighting level and direction of view (for foveal vision, i.e., looking straight ahead, photopic photometry seems to hold even at low light levels). For peripheral visual angles brightness perception and the perception of an object (a signal, sign or obstacle in a nighttime driving situation) seem to have different spectral responsivity. In driving situations the necessary reaction time of the driver is an important parameter, thus experiments are going on to define a photometric system based on reaction time investigations.

In indoor situations apart from the necessary level of illumination, the observed glare is a contributor whether an environment will be accepted as pleasing or annoying. Illuminating engineering distinguishes between two types of glare: disability glare reduces visibility, discomfort glare is just an annoying experience without influencing the short-term task performance. An interesting question is the spectral sensitivity to discomfort glare, as it can influence not only indoor but also outdoor activity. Preliminary experiments seem to show that luminance sensitivity and discomfort glare sensitivity have different spectral distribution; glare sensitivity seems to peak at shorter wavelengths.

The above might be related to a further question, lying already at the boundaries of photometry, but that has to be considered in photometric design and measurement: the human daily and yearly rhythm (circadian and seasonal rhythm) of human activity coupled to hormone levels. They are influenced by light, as, for example, the hormone melatonin production is influenced by light exposure. Physiological investigations showed that melatonin production suppression has a maximum around 460 nm and might be coupled to a radiation sensitive ganglion cell in the retina. Whether discomfort sensation is mediated via the same neural pathway or via a visual one has not yet been decided. But photometry has to take these also into consideration and in the future, measurement methods and instruments to determine them, will have to be developed.

Advances in Photometric Measurements

Primary Standards

The main concern in photometry is that the uncertainty of photometric measurements is still much higher than that in other branches of physics. This is partly due to the higher uncertainty in radiometry and partly to the increase in uncertainty within the chain of uncertainty propagation from the National Laboratory to the workshop floor measurement.

In National Standards Laboratories, very sophisticated systems are used to determine the power of the incoming radiation and then elaborated spectroradiometric techniques are used to evaluate the radiation in the form of light, i.e., perform photometric measurements (e.g., NIST CIRCUS equipment, where multiple laser sources are used as power sources, and highly sophisticated methods to produce a homogeneous nonpolarized radiation field for the calibration of secondary photometric detectors).

There is, however, also a method to supply end users with absolute detectors for the visible part of the spectrum. Modern high-end Si photovoltaic cells have internal quantum efficiencies in the visible part of the spectrum of over 99.9%. Reflection losses at the silicon surface are minimized by using three or more detectors arranged in a trap configuration, where the light reflected from one detector is fed to the second one, from there to the third one, and eventually to some further ones. In a three detector configuration, as shown in Figure 6, the light from the third detector is reflected back to the second and from there to the first one. As every detector reflects only a small amount of radiation, by the fifth reflection practically

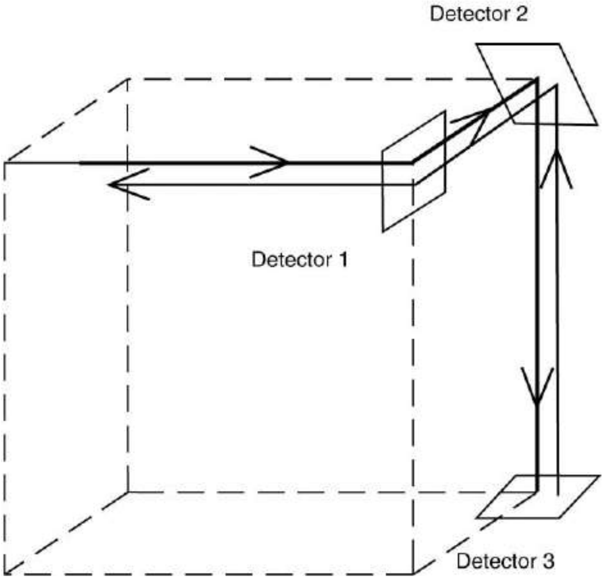

Figure 6 Schematic layout of a three Si-cell trap detector: light comes in from the left, is first partly absorbed, partly reflected on Detector 1, then on Detector 2, then on Detector 3, from where it is reflected back to Detector 2 and 1.

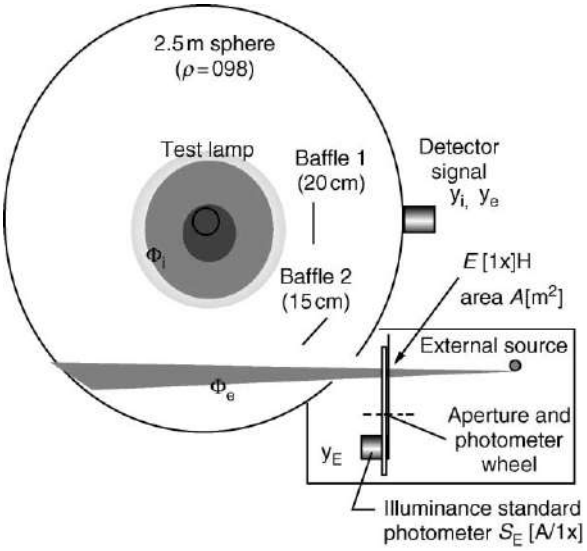

Figure 7 Arrangement of the absolute integrating sphere system developed at NIST for the detector-based total luminous flux calibration. By permission of IESNA from Ohno Y and Bergman R (2003) Detector-referenced integrating sphere photometry for industry. *J. IES Summer* 21–26.

all the radiation is absorbed and contributes to the electric signal. Such trap detectors have an almost 100% quantum efficiency in the visible part of the spectrum, and can be used as photometric detectors if a well designed color correcting filter is applied in front of the detector.

Secondary Type Measurements

In practical photometry the three most important quantities to be measured are the total luminous flux of different lamps, the illuminance in a plane and the luminance.

Light source measurement

Total luminous flux. The two methods to measure the total luminous flux is to use a goniophotometer or a photometer (Ulbicht) sphere. In goniophotometry, recent years have not brought major breakthroughs, the automation of the systems got better, but the principles are unchanged.

The integrating sphere photometer (a sphere with inner white diffuse coating, where the lamp is in the middle of the sphere) used to be a simple piece of equipment to compare total luminous flux lamps against flux standards. In recent years a new technique has been developed at NIST–USA. This enables the absolute measurement of luminous flux from illuminance measurement, the fundamentals of this new arrangement being shown in Figure 7: the test lamp is as usual in the middle of the sphere, but now light from an external source is introduced into the sphere. An illuminance meter measures the flux entering from this source. The sphere detector compares the two signals (y_i and y_e). Knowing the absolute characteristics of the sphere (a difficult measurement) one can determine the total luminous flux of the test lamp using the absolute illuminance value. As illuminance is easily determined from luminous intensity (from eqns [3] and [5] one gets with $d\Omega = dA/r^2$ that $E = I/r^2$, where r is the distance between the source and the illuminated surface, supposed to be perpendicular to the direction to the source), this technique permits us to derive the total luminous flux scale from illuminance or luminous intensity measurement using an integrating sphere.

Luminous intensity of LEDs. An other major break-through achieved during the past years was the unified measurement of LED intensity. Light-emitting diodes became, in recent years, important light sources for large scale signaling and signing, and it is foreseen that they will become important contributors in every field of light production (from car headlamps to general illumination). The most fundamental parameter of the light of an LED is its luminous intensity. The spatial power distribution of LEDs is usually collimated, but the LEDs often squint, as seen in Figure 8. In the past, some manufacturers measured the luminous intensity in the direction of maximum emission, others used the direction of the optical axis for this quantity. The highly collimated character of the radiation made measurements in far field rather difficult. Therefore, CIE recommended a new term and measuring geometry: average LED intensity can be measured under two measuring conditions, as shown in Figure 9. The detector has to be set in the direction of the LED mechanical axis (discussions are still going on as to what the reference direction should be with modern surface mounted LEDs, as with those the mechanical axis is ill-defined, the normal to the base-plane could be a better reference direction). The detector has to have an exactly 1.00 cm^2 circular aperture, and the distance between this aperture and the tip of the LED is for condition A, $d = 0.316$ m, and for Condition B, $d = 0.100$ m (these two distances with the 1.00 cm^2 detector area provide 0.001 sr (steradians) and 0.01 sr opening angles). Recent international round-robins have shown that, based on the new recommendations, agreement between different laboratories decreased from the 10 to 20% level to 1 to 2%. The remaining difference is mainly due to the fact that the LEDs emit in narrow wavelength bands, and the transfer of the calibration value for the 100 mm^2 detector from a white (CIE Standard Illuminant A color temperature) incandescent lamp to the narrow band LED emission is still uncertain, mainly due to stray light effects in the spectral responsivity and emission measurement.

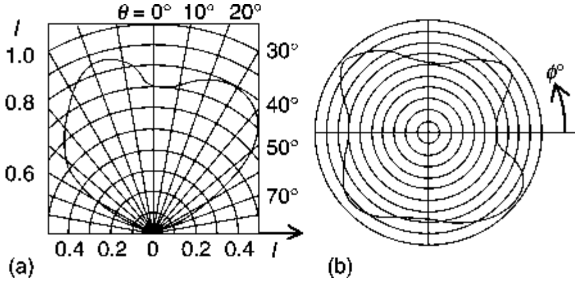

Figure 8 Spatial light distribution of an LED, Figure 8a shows the distribution of a 'squinting' LED in a plane including the optical axis, Figure 8b shows light distribution in a plane perpendicular to the optical axis. By permission of the Commission Internationale de l'Eclairage, from the Publication "Measurement of LEDs" CIE 127-1997; CIE Publications are obtainable from the CIE Central Bureau: Kegelgasse 27, A-1033 Wien, Austria.

Luminance distribution measurement

The human observer sees luminance (and color) differences. Thus for every illuminating engineering design task the luminance distribution in the environment is of utmost importance. Traditionally this was measured using a spot-luminance meter, aiming the device into a few critical directions. The recent development of charge coupled device (CCD) two-dimensionally sensitive arrays (and other, e.g., MOS-FET, CMOS, Charge injection device (CID), charge imaging matrix (CIM) systems: as for the time being the CCD technology provides best performance, we will refer to two-dimensional electronic image capture devices as to CCD cameras) opened the possibility of using an image-capturing camera for luminance distribution measurements. Such measurements are badly needed in display calibration, near-field photometry (photometry in planes nearer than in which the inverse square law holds), testing of car headlamp light distribution, glare, and homogeneity studies indoors and outdoors, etc.

Solid-state cameras have the big advantage over older type vacuum-tube image capturing devices, that the geometric position and alignment of the single pixels is well-defined and stays constant. Nowadays, CCD image detector chips are mass produced, and one can get a variety of such devices and cameras starting with some small resolution

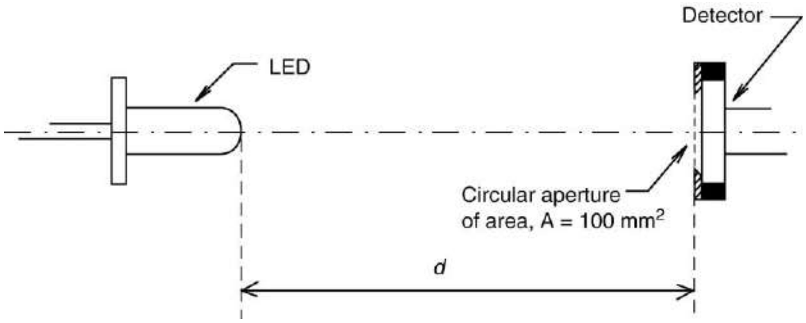

Figure 9 Schematic diagram of CIE Standard Conditions for the measurement of Average LED Intensity. Distance $d = 0.316$ m for Condition A and $d = 0.100$ m for Condition B. By permission of the Commission Internationale de l'Eclairage, from the Publication "Measurement of LEDs" CIE 127-1997; CIE Publications are obtainable from the CIE Central Bureau: Kegelgasse 27, A-1033 Wien, Austria.

(few thousand pixels) devices up to cameras with tens of mega pixel resolution. Detectors are now available with internal intensification enabling measurements down to a few photons per second intensity levels.

Main problems with these detectors are:

- spectral and absolute nonuniformities of the single pixels (see Figure 10), where a spatial homogeneity map of a CCD two-dimensional array detector is shown; the irregular 3% sensitivity change on the surface of the detector is negligible for imaging purposes, but has to be corrected in cases of photometric measurements. Even if the receptor chip would have an absolutely homogeneous sensitivity, there would be a drop in response from the middle of the imaging area to the boarders: Light reaching the edges of the detector reach the detector at an oblique angle and this produces a decrease of sensitivity with α^4, where α is the angle of incidence, measured from the middle of the lens to the given pixel of the detector and the surface normal of the detector.
- aliasing effects if the pixel resolution is not large enough to show straight lines as such when they are not in line with a pixel row or column.
- nonlinearity and cross-talk among the adjacent pixels. Figure 11 shows the so-called 'inverse gamma' characteristic of a CCD camera. The digital electronic output of the camera shows a $Y = E^{-\gamma}$ type function, where Y is the output DAC (digital-analog converter) values, E is the irradiance of the pixel, and γ is the exponent (this description comes from the film industry, where the film density depends exponentially on the irradiation; display devices have usually a nonlinear input (DAC value) – output (luminance) characteristic, and the camera inverse gamma value corrects for this output device characteristic. This is, however, not required if the camera is used for photometric measurements; this built-in nonlinearity has to be corrected in the evaluating software if the camera is intended for photometric measurements.
- to be able to perform photometric measurements the camera has to have a spectral responsivity corresponding to the CIE $V(\lambda)$-function. Many cameras have built-in filters to do this – eventually also red and blue filters to be able to capture color – but the color correction of cameras, where the correction is made by small adjacent filter chips, is usually very poor. Figure 12 shows the spectral

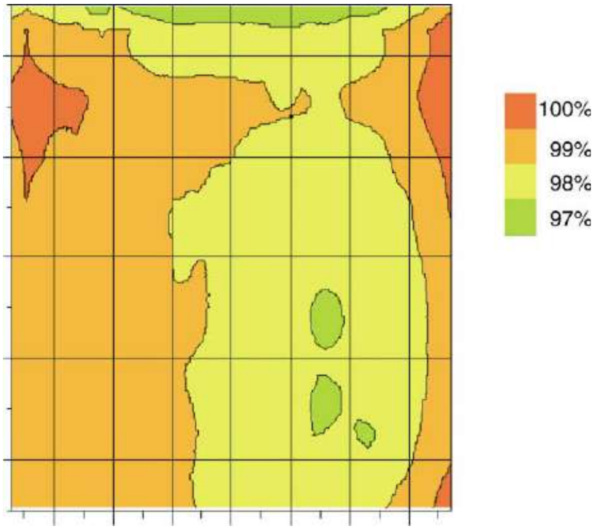

Figure 10 Spatial homogeneity of a two-dimensional CCD array.

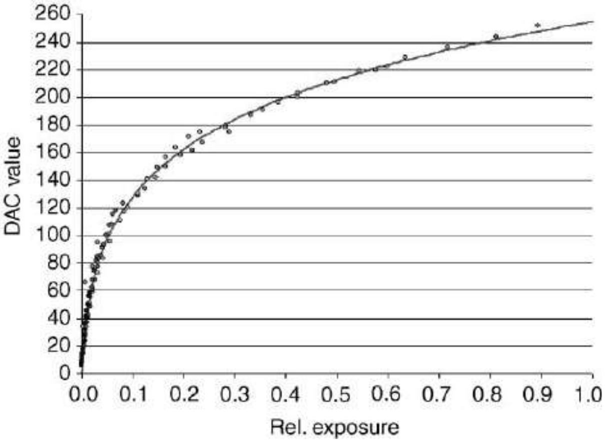

Figure 11 'Inverse gamma' characteristic of a commercial digital photographic camera, measurement points are shown at different speed settings, curve is a model function representative of the camera response.

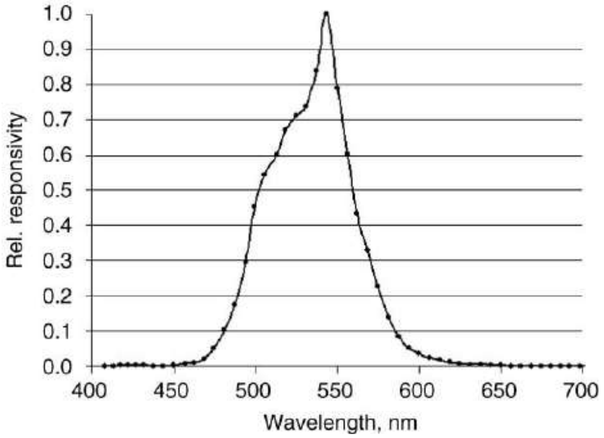

Figure 12 Spectral sensitivity of a commercial digital photographic camera.

sensitivity curve of a digital photographic camera; the output signal is produced by an internal matrix transformation of the signals produced by adjacent pixels equipped with different color filters. Figure 13 shows the spectral sensitivity of a CCD camera specially designed for photometric measurements. Naturally, meaningful photometric measurements can be made only with such a camera. Very often, however, an approximate luminance distribution is enough, and then a picture captured by a digital photographic camera, plus one luminance measurement of a representative object for absolute calibration, suffices.

Above nonspectral systematic errors can be corrected by appropriate soft-ware, so a CCD camera photometer, as shown schematically in Figure 14, is well suited to measure display characteristics, indoor and outdoor luminance distributions. The challenge for illuminating engineering is at present how the many millions of luminance values can be evaluated to get to meaningful light measurement data.

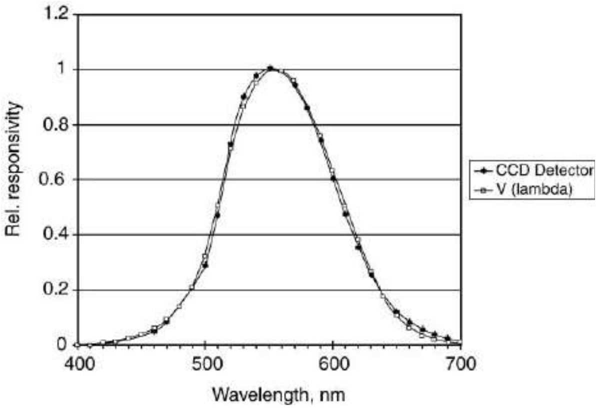

Figure 13 Spectral sensitivity of a CCD camera specially designed for photometric measurements. Kindly supplied by INPHORA Inc.

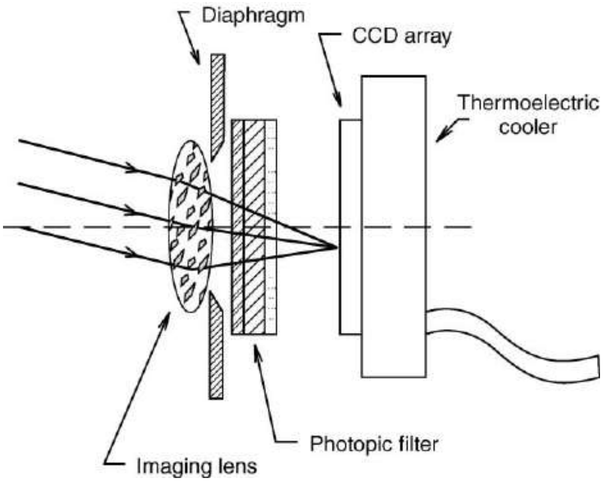

Figure 14 Cross-section of a photometric CCD camera.

The real challenge will come if visual science provides better hints how the human visual system evaluates the illuminance distribution on the retina, and instrument manufacturers will be able to capture signals corresponding to those produced by the receptor cells and provide the necessary algorithms our brain uses to get to brightness, lightness, color, luminance contrast, and glare type of output information.

Concluding Remarks

Advances in optical instrumentation, both in the field of light sources and detectors – coupled with the possibilities modern digital computation (eventually in the future increasing the use of neural networks) – provide already many new measurement technical solutions and further ones are certainly underway.

The use of LEDs needs the rethinking of many classical illuminating engineering concepts, from visibility and glare evaluation, evenness of illumination to color rendering. All of them are coupled with problems in basic photometry. Thus, there is a need to re-evaluate concepts used in design techniques. The new area-sensitive detectors provide methods of determining classical photometric quantities of entire visual fields in one shot, but already foreshadow the development of new quantities that correlate better with visual perceptions.

List of Units and Nomenclature

Terms with an $*$ refer to definitions published by the CIE in the International Lighting Vocabulary, CIE 17.4:1986, where further terms and definitions related to light and lighting are to be found.

Candela*: SI unit of luminous intensity: The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of 1/683 watt per steradian. (16th General Conference of Weights and Measures, 1979).

$$1 \text{ cd} = 1 \text{ lm sr}^{-1}$$

Equivalent luminance*: Luminance of a comparison field in which monochromatic radiation of frequency 540×10^{12} Hz has the same brightness as the field considered under the specified photometric conditions of measurement; the comparison field must have a specified size and shape which may be different from that of the field considered.

$$\text{unit} : \text{cd m}^{-2}$$

Notes:

1. Radiation at a frequency of 540×10^{12} Hz has a wavelength in standard air of 555.016 nm.
2. A comparison field may also be used in which the radiation has any relative spectral distribution, if the equivalent luminance of this field is known under the same conditions of measurement.

Far field photometry: Photometry where the inverse square law is valid.

Flicker photometer*: Visual photometer in which the observer sees either an undivided field illuminated successively, or two adjacent fields illuminated alternately, by two sources to be compared, the frequency of alteration being conveniently chosen so that it is above the fusion frequency for colours but below the fusion frequency for brightnesses.

Fovea*: Central part of the retina, thin and depressed, which contains almost exclusively cones and forming the site of most distinct vision.

Note: The fovea subtends an angle of about 0.026 rad (1.5 degree) in the visual field.

Goniophotometer*: Photometer for measuring the directional light distribution characteristics of sources, luminaires, media or surfaces.

Illuminance*: Quotient of the luminous flux $d\Phi_v$ incident on an element of the surface containing the point, by the area dA of that element.

Equivalent definition. Integral, taken over the hemisphere visible from the given point, of the expression $L_v \cos\theta \, d\Omega$, where L_v is the luminance at the given point in the various directions of the incident elementary beams of solid angle $d\Omega$, and θ is the angle between any of these beams and the normal to the surface at the given point.

$$E_v = \frac{d\Phi_v}{dA} = \int_{2\pi sr} L_v \cos\theta \, d\Omega$$

$$\text{unit} : \text{lx} = \text{lm m}^{-2}$$

Inverse square law: The illumination at a point on a surface varies directly with the luminous intensity of the source, and inversely as the square of the distance between the source and the point if the source is seen as a point source.

Lumen*: SI unit of luminous flux: Luminous flux emitted in unit solid angle (steradian) by a uniform point source having a luminous intensity of 1 candela. (9th General Conference of Weights and Measures, 1948).

Equivalent definition. Luminous flux of a beam of monochromatic radiation whose frequency is 540×10^{12} hertz and whose radiant flux is 1/683 watt.

Luminance*: Quantity defined by the formula

$$L = \frac{\partial^2\Phi}{\partial A \cos\theta \, \partial\Omega}$$

where $\partial\Phi$ is the luminous flux transmitted by an elementary beam passing through the given point and propagating in the solid angle $d\Omega$ containing the given direction; dA is the area of a section of that beam containing the given point, θ is the angle between the normal to that section and the direction of the beam.

$$\text{unit} : \text{cd m}^{-2}$$

Luminous intensity*: Quotient of the luminous flux $d\Phi_v$ leaving the source and propagated in the element of solid angle $d\Omega$ containing the given direction, by the element of solid angle:

$$I_v = \frac{d\Phi_v}{d\Omega}$$

$$\text{unit} : \text{cd} = \text{lm sr}^{-1}$$

Lux*: SI unit of illuminance: Illuminance produced on a surface of area 1 square meter by a luminous flux of 1 lumen uniformly distributed over that surface.

$$1 \text{ lx} = 1 \text{ lm m}^{-2}$$

Note: Non-metric unit: **lumen per square foot** (lm ft^{-2}) or **footcandle** (fc) (USA) = 10.764 lx.

Near-filed photometry: photometry made in the vicinity of an extended source, so that the inverse square law is not valid.

Photometer (or Ulbicht) sphere*: A hollow sphere, whitened inside. Owing to the internal reflexions in the sphere, the illumination on any part of the sphere's inside surface is proportional to the luminous flux entering the sphere, or produced inside the sphere by a lamp. The illuminance of the internal sphere wall is measured via a small window.

Pixel: The individual picture elements of an image or elements in a display that can be addressed individually.

Radiance*: Quantity defined by the formula

$$L = \frac{\partial^2\Phi}{\partial A \cos\theta \, \partial\Omega}$$

where $\partial\Phi$ is the radiant flux transmitted by an elementary beam passing through the given point and propagating in the solid angle $d\Omega$ containing the given direction; dA is the area of a section of that beam containing the given point, θ is the angle between the normal to that section and the direction of the beam.

$$\text{unit} : \text{W m}^{-2} \text{ srad}$$

Retina*: Membrane situated inside the back of the eye that is sensitive to light stimuli; it contains photoreceptors, the cones and the rods, and nerve cells that interconnect and transmit to the optic nerve the signals resulting from stimulation of the photoreceptors.

Spectral*: An adjective that, when applied to a quantity X pertaining to electromagnetic radiation, indicates:

- either that X is a function of the wavelength λ, symbol: $X(\lambda)$,
- or that the quantity referred to is the spectral concentration of X, symbol: $X_\lambda \equiv \mathrm{d}X/\mathrm{d}\lambda$.

X_λ is also a function of λ and in order to stress this may be written $X_\lambda(\lambda)$ without any change of meaning.

Spectral luminous efficiency function: for photopic vision, $V(\lambda)$; for scotopic vision $V'(\lambda)$*: Ratio of the radiant flux at wavelength λ_m to that at wavelength λ such that both radiations produce equally intense luminous sensations under specified photometric conditions and λ_m is chosen so that the maximum value of this ratio is equal to 1.

Unless otherwise indicated, the values used for the spectral luminous efficiency in photopic vision are the values agreed internationally in 1924 by the CIE (*Compte Rendu* 6^e session, p.67), completed by interpolation and extrapolation (*Publications CIE* No.18 (1970), p.43 and No. 15 (1971), p.93), and recommended by the International Committee of Weights and Measures (CIPM) in 1972. For *scotopic vision*, the CIE in 1951 adopted, for young observers, the values published in *Compte Rendu* 12^e session, Vol. 3, p. 37, and ratified by the CIPM in 1976. These values define respectively the $V(\lambda)$ or $V'(\lambda)$ functions.

Total luminous flux: luminous flux of a source emitted into 4π steradians.

Trap detector: Detector array prepared from detectors of high internal quantum efficiency, where the reflected radiation of the first detector is directed to the second one, and so on so that practically all the radiation is absorbed by one of the detectors (the radiation is "trapped" in the detector system).

See also

Displays. **Incoherent sources:** Lamps.

Further Reading

Commission Internationale de l'Eclairage: *The basis of physical photometry* (1983) CIE 18.2.

Commission Internationale de l'Eclairage: Proceedings of the CIE Symposium 2002 on *Temporal and Spatial Aspects of Light and Colour Perception and Measurement*, August 2002: Publ.CIE x025:2003.

Cromer CL, Eppledauer G, Hardis JE, Larason TC and Parr AC (1993) National Institute of Standards and Technology detector-based photometric scale. *Applied Optics* 32: 2936–2948.

Decusatis C (ed.) (1997) *Handbook of Applied Photometry*. Woodbury, New York: American Institute of Physics.

Rea MS (ed.) (2000) Illuminating Engineering Society of North America. *The IESNA Lighting Handbook: Reference and Application*, 9th edn.

Kaiser PK and Boynton RM (1966) *Human Color Vision*, 2nd edn. Washington, DC: Optical Society of America.

McCluney WR (1994) *Introduction to radiometry and photometry*. Boston, London: Artech House.

Ohno Y and Bergman R (2003) Detector-referenced integrating sphere photometry for industry. *J IES Summer* 21–26.

Walsh JWT (1953) *Photometry*. London: Constable.

Scatterometry

J C Stover, The Scatter Works, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Scattered light is a limiting source of optical noise in many advanced optical systems, but it can also be a sensitive indicator of optical component quality. Consider the simple case of a telescope successfully used to image a dim star against a dark background; however, if light from a bright source (such as the moon located well out of the field of view) enters the telescope, it will scatter from the interior walls and the imaging optics themselves. Some of this light eventually reaches the detector and creates a dim background haze that washes out the image of the distant star. A good telescope design accounts for these effects by limiting potential scatter propagation paths and by requiring that critical elements in the optical system meet scatter specifications. This means doing a careful system analysis and a means to quantify the scattering properties of the telescope components. This article discusses modern techniques for quantifying, measuring, and analyzing scattered light, and reviews their development.

Like many scientific advances moving scatterometry from an art to a reliable metrology was done in a series of small hops (not always in the same direction), rather than a single leap. It started in 1961 when a paper by Hal Bennett and Jim Porteous reported measurements made by gathering most of the light scattered from front surface mirrors and normalizing this signal by the much larger specular reflection. They defined this ratio as the total integrated scatter (TIS), and using a scalar diffraction theory result drawn from the radar literature, related it to the surface root mean square (rms) roughness. By the mid-1970s, several angle-resolved scatterometers had been built as research tools in university, government, and industry labs. Unfortunately, instrument operation and data manipulation were generally poor, and meaningful comparison measurements were virtually impossible due to instrument differences, sample contamination, and confusion over what parameters should be compared. Analysis of scatter data, to characterize sample surface roughness, was the subject of many publications. A derivation of what is commonly called 'BRDF' (bidirectional reflectance distribution function) was published by Nicodemus and co-workers at the National Bureau of Standards (now the National Institute of Science and Technology or NIST) in 1970, but did not gain common acceptance as a way to quantify scatter measurements until the late 1980s when the advent of small computers, combined with inspection requirements for defense-related optics, dramatically stimulated the development of scatter metrology. Commercial laboratory instrumentation became available that could measure and analyze as many as 50 to 100 samples a day, and the number (and sophistication) of measurement facilities increased dramatically. The first ASTM Standards were published (TIS in 1987 and BRDF in 1991), but it was still several years before most publications correctly used these quantifying terms. Government defense funding decreased dramatically in the early 1990s, following the end of the Cold War, but the economic advantages of using scatter metrology and analysis for space applications and in the rapidly advancing semiconductor industry, continued state-of-the-art advancements.

The following sections detail how scatter is quantified when related to area (roughness) and local (pit/particle) generating sources. Instrumentation and the use of scattering models are also briefly reviewed.

Quantifying Scattered Light

Scatter signals can be easily quantified as scattered light power per unit solid angle (in watts per steradian); however, in order to make the results more meaningful, these signals are usually normalized, in some fashion, by the light incident on the scatter source. The three ways commonly employed to do this are defined below.

If the scattering feature in question is uniformly distributed across the illuminated spot on the sample (such as surface roughness), then it makes sense to normalize the collected scattered power in watts/steradian by the incident power. This simple ratio, which has units of inverse steradians, was commonly referred to as 'the scattering function.' Although this term is occasionally still found in the literature, it has been generally replaced by the closely related BRDF, which is defined by the differential ratio of the sample radiance normalized by its irradiance. After some simplifying assumptions are made, this reduces to the original scattering function with a cosine of the polar scattering angle in the denominator. The BRDF, defined in this manner, has become the standard way to report angle-resolved scatter from features that uniformly fill the illuminated spot. The cosine term results from the fact that NIST used radiometric terms to define BRDF:

$$\mathrm{BRDF} = \frac{\mathrm{Ps}/\Omega}{\mathrm{Pi}\cos\theta_s} \qquad [1]$$

The scatter function is often referred to as the 'cosine corrected BRDF' and is simply equal to the BRDF multiplied by the cosine of the polar scattering angle. Figure 1 gives the geometry for the situation, and defines the polar and azimuthal angles (θ_s and ϕ_s), as well as the solid collection angle (Ω). Other common abbreviations are BSDF, for the more generic bidirectional scatter distribution function, and BTDF for quantifying transmissive scatter.

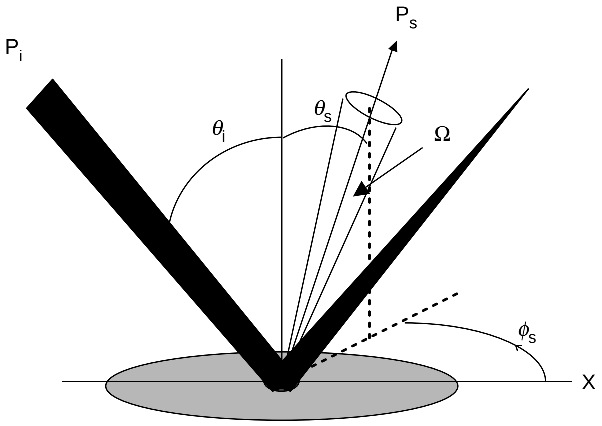

Figure 1 Scatter analysis uses standard spherical coordinates to define terms.

Integration of the scatter signal over much of the scattering hemisphere allows calculation of TIS, as the ratio of the scatter signal to the reflected specular power. This integration is usually carried out experimentally in such a way that both the incident beam and reflected specular beam are excluded. In the most common TIS situation, the beam is incident at a small angle near surface normal, and the integration is done from small values of θ_s to almost 90 degrees. If the fraction of light scattered from the specular reflection is small and if the scatter is caused by surface roughness, then it can be related to the rms surface roughness of the reflecting surface. As a ratio of powers, the TIS is a dimensionless quantity. The normalization is by Pr (instead of Pi) because reductions in scatter caused by low reflectance do not influence the roughness calculation. The pertinent relationships are given below, where σ is the rms roughness and λ is the light wavelength:

$$\mathrm{TIS} = \mathrm{Ps/Pr} \cong (4\pi\sigma/\lambda)^2 \qquad [2]$$

Of course, all scatter measurements are integrations over a detector collection aperture, but the TIS designation is reserved for situations where the aim is to gather as much scattered light as possible, while 'angle resolved' designs are created to gain information from the distribution of the scattered light. Notice that TIS values become very large when measured from a diffuse surface, where the specular reflection is very small. Although TIS can be measured for any surface, the diffuse reflectance (equal to Ps/Pi) would often be more appropriate for diffuse surfaces. The various restrictions associated with relating TIS to rms roughness are detailed below.

Scatter from discrete features, such as particles and pits, which do not completely fill the illuminated spot, must be treated differently. This is because changes in spot size, with no corresponding change in total incident power, will change the incident intensity (watts/unit area) at the feature and thus also change the scatter signal (and BRDF) without any corresponding changes in the scattering feature. Clearly this is unacceptable if the object is to characterize the defect with scatter measurements. The solution is to define another quantification term, known as the differential scattering cross-section (DSC), where the normalization is the incident intensity at the feature (the units for DSC are area/steradian). Because this was not done in terms of radiometric units at the time it was defined, the cosine of the polar scattering angle is not in the definition. The same geometrical definitions, found in Figure 1, also apply for the DSC:

$$\mathrm{DSC} = \frac{\mathrm{Ps}/\Omega}{Ii} \qquad [3]$$

If the DSC is integrated over the solid angle associated with a collection aperture then the value has units of area. Because relatively small area focused laser beams are often used as a source, area is most commonly given in micrometers squared.

These three scatter parameters, the BRDF, the TIS, and the DSC, are obviously functions of system variables such as geometry, scatter direction (both in and out of the incident plane), incident wavelength and polarization, as well as feature characteristics. It is the dependence of the scatter signal on these system parameters that makes the scatter models useful for optimizing instrument designs. It is their dependence on feature characteristics that makes scatter measurement a useful metrology tool.

A key point needs to be stressed. When applied appropriately, TIS, BRDF, and DSC are absolute terms, not relative terms. The DSC of a 100 nm PSL in a given direction for a given source is a fixed value, which can be repeatedly measured and even accurately calculated from models. The same is true for TIS and BRDF values associated with surface roughness of known statistics. Scatter measuring instruments, such as particle scanners or lab scatterometers, can be calibrated in terms of these quantities. As has already been pointed out, the user of a scanner will almost always be more interested in characterizing defects than in the resulting scatter values, but the underlying instrument calibration can always be expressed in terms of these three quantities. This is true even though designers and users may find it convenient to use other metrics (such as polysterene latex (PSL) spheres) as a way to relate calibration.

Angle Resolved Scatterometers

The diagram in Figure 2 shows the most common scatterometer configuration. The source is fixed and the sample is rotated to the desired incident angle. The receiver is then rotated about the sample during scatter measurement. Most commonly, scatterometers operate just in the plane of incidence; however, instruments capable of measuring at virtually any location in either the reflective or transmissive hemispheres have been built. Although dozens of instruments have been built following this general design, other configurations are in use. For example, the source and receiver may be fixed and the sample rotated so that the scatter pattern moves past the receiver. This is easier mechanically than moving the receiver at the end of an arm, but complicates analysis because the incident angle and the observation angle change simultaneously. Another combination is to fix the source and sample together, at constant incident angle, and rotate this unit (about the point of illumination on the sample) so that the scatter pattern moves past a fixed receiver. This has the advantage that a long receiver/sample distance can be used without motorizing a long (heavy) receiver arm. It has the disadvantage that heavy (or multiple) sources are difficult to deal with. Other configurations, with everything fixed, have been designed that employ several receivers to merely sample the BSDF and display a curve fit of the resulting data. This is an economical solution if the BSDF is relatively uniform without isolated diffraction peaks. The goniometer section of a real instrument, similar to that of Figure 2, is shown in Figure 3.

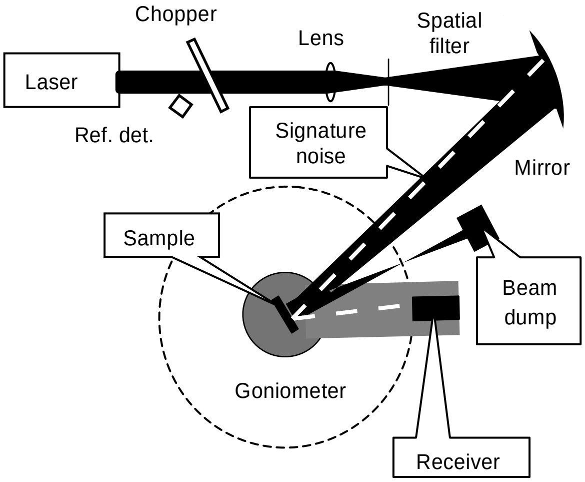

Figure 2 Basic elements of an incident plane scatterometer are shown.

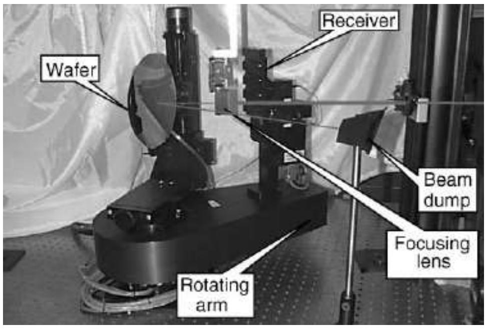

Figure 3 The author's scatterometer, which is similar to the diagram of Figure 2 is shown. In this case a final focusing lens is introduced to produce a very small illuminated spot on the silicon wafer sample. The white background was introduced to make the instrument easier to view.

Computer control of the measurement is essential to maximize versatility and minimize measurement time. The software required to control the measurement plus the display and analysis of the data can be expected to be a significant portion of total instrument development cost. The following reviews typical design features (and issues) associated with the source, sample mount and receiver components.

The source in Figure 2 is formed by a laser beam that is chopped, spatially filtered, expanded, and finally brought to a focus on the receiver path. The beam is chopped to reduce both optical and electronic noise. This is usually accomplished through the use of lock-in detection in the electronics package which suppresses all signals except those at the chopping frequency. Low noise, programmable gain electronics are essential to reducing system noise. The reference detector is used to allow the computer to ratio out laser power fluctuations and, in some cases, to provide the necessary timing signal to the lock-in electronics. Polarizers, wave plates, and neutral density filters are also commonly placed prior to the spatial filter. The spatial filter removes source scatter from the laser beam and presents a point source which is imaged by the final focusing element, in this case a mirror, to the detector zero position. Focusing the beam at this location allows near specular scatter to be more easily measured. Lasers are convenient sources, but are not necessary. Broadband sources are often required to meet a particular application or to simulate the environment where a sample will be used. Monochromators and filters can be used to provide scatterometer sources of arbitrary wavelength. The noise floors with these tunable incoherent sources increases as the spectral bandpass is narrowed, but they have the advantage that the scatter pattern does not contain laser speckle.

The sample mount can be very simple or very complex. In principal, six degrees of mechanical freedom are required to fully adjust the sample. The order in which these stages are mounted affects the ease of use (and cost) of the sample holder. In practice, it often proves convenient to either eliminate, or occasionally duplicate, some of these degrees of freedom. In addition, some of these axes may be motorized to allow the sample area to be raster-scanned to automate sample alignment or to measure reference samples. As a general rule, the scatter pattern is insensitive to small changes in incident angle but very sensitive to small angular deviations from specular. Instrumentation should be configured to allow location of the specular reflection (or transmission) very accurately. Receiver designs vary, but changeable entrance apertures, bandpass filters, lenses, and field stops are generally positioned in front of the detector.

A serious measurement problem is getting light scattered by the instrument, called instrument signature, confused with light scattered by the sample. An example of instrument signature is shown by the dotted line in Figure 2, which represents scatter from the final mirror. The signature is often measured in the straight through (transmission) direction, multiplied by the measured specular reflectance and then compared to the measured sample BRDF. Another issue is the fact that the measured BRDF is really the convolution of the receiver aperture with the actual (incremental) BRDF. When the scatter signal varies slowly across the aperture the measurement is virtually identical to the true BRDF. Near the specular reflection, or at diffraction peaks, the differences between the measurement (or convolution) and the actual (incremental) BRDF can be huge. Measurements made using invisible sources and measurements of curved samples present additional problems. These problems and the issues of calibration and accuracy are covered in the Further Reading section at the end of this article.

TIS Instruments

The two common methods of making TIS measurements are shown in Figures 4 and 5. The first one is based on a hemispherical mirror (or Coblentz sphere) to gather scattered light from the sample and image it onto the scatter detector. The specular beam enters and leaves the hemisphere through a small circular hole. The diameter of that hole defines the near specular limit of the instrument. The reflected beam (not the incident beam) should be centered in the hole

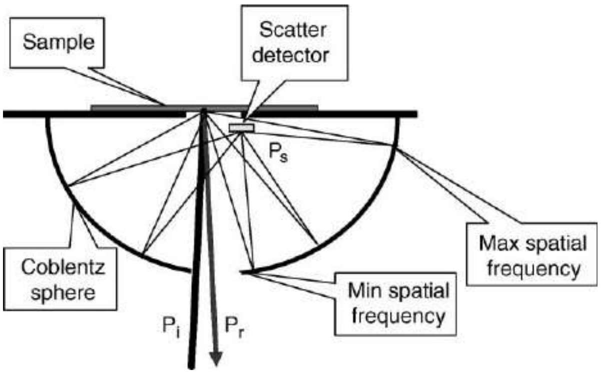

Figure 4 A diagram showing the Coblentz sphere approach to TIS measurements used in the early development of scatter instrumentation.

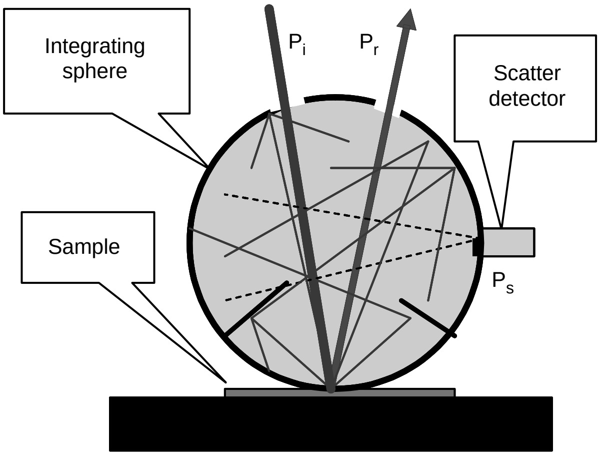

Figure 5 More modern TIS instruments make use of an integrating sphere approach which is easier to align and does not suffer from problems associated with measuring high-angle scatter from the sample.

because the BRDF will be symmetrical about it. Alignment of the hemispherical mirror is critical, and not trivial, in this approach. The second approach involves the use of an integrating sphere. A section of the sphere is viewed by a recessed detector. If the detector field of view (FOV) is limited to a section of the sphere that is not directly illuminated by scatter from the sample, then the signal will be proportional to total scatter from the sample. Again, the reflected beam should be centered on the exit hole. The Coblentz sphere method presents more signal to the detector; however, some of this signal is incident on the detector at very high angles. Thus, this approach tends to discriminate against high-angle scatter (which is not a problem for many samples). The integrating sphere is easier to align, but has a lower signal to noise ratio (less signal on the detector) and is more difficult to build in the IR where uniform diffuse surfaces are harder to obtain. A common mistake with TIS measurements is to assume that for near normal incidence, the orientation between source polarization and sample orientation is not an issue. TIS measurements made with a linearly polarized source on a grating at different orientations will quickly demonstrate this dependence.

TIS measurements can be made very near the specular reflection by utilizing a diffusely reflecting plate with a small hole in it. A converging beam is reflected off the sample and through the hole. Scatter is diffusely reflected from the diffuse plate to a receiver designed to uniformly view the plate. The reflected power is measured by moving the plate so the specular beam misses the hole and then taking that measurement. The ratio of the two scatter measurements gives the TIS. Measurements starting closer than 0.1 degrees from specular can be made in this manner and it is an excellent way to check incoming optics or freshly coated optics for low scatter.

Analyzing Scatter from Surface Roughness

The preceding sections have concentrated on obtaining and quantifying accurate scatter data, but that leaves the question of what to do with it once you have it. In rare situations you may be given a scatter (BRDF) specification – such as, the BRDF from the mirror must be less than 10^{-4} sr^{-1} 10 degrees from specular when measured at a wavelength of 633 nm incident at 5 degrees with an S polarized source. Unfortunately this is very uncommon. If the issue is limiting scatter as a noise source, you will probably have to generate your own specification based on specific system requirements. More difficult, and often of more economic value, is the situation where scatter measurements are being used as a metrology to learn something about the sample characteristics – like roughness or defect size and/or type.

The relationship between the measured BRDF and reflector roughness statistics was a subject of intense interest from the mid-1970s through the early 1990s. Dozens of papers, and even some books, have been written on the subject, and it can only be outlined here. The relatively easy case of scatter from roughness on a clean, optically smooth, front surface reflector was first published in 1975; however it was several years later before confirming experiments were completed. The deceptively simple relationship, based on vector perturbation theory, is shown below.

$$\mathrm{BRDF} = (16\pi^2/\lambda^4)\cos\theta_\mathrm{i}\cos\theta_\mathrm{s}\,QS(f_x, f_y) \qquad [4]$$

Q is the polarization factor and is determined by the material constants of the reflector, as well as the system geometry. In many cases, it is numerically about equal to the specular reflectance and this approximation is often justified. Exact expressions are available in the literature. $S(f_x, f_y)$ is the surface power spectral density function (or PSD). It may be thought of as roughness power (surface height variations squared) per unit spatial frequency (undulations per unit distance instead of per unit time). Integration of the PSD over spatial frequency space results in the mean square roughness over that band of frequencies. Taking the square root gives the root mean square roughness (or rms). Frequencies in both the x and y directions on the surface are involved and they are defined by the well-known grating equations as:

$$f_x = (\sin\theta_\mathrm{s}\cos\phi_\mathrm{s} - \sin\theta_\mathrm{i})/\lambda \qquad \text{and}$$

$$f_y = (\sin\theta_\mathrm{s}\sin\phi_\mathrm{s})/\lambda \qquad [5]$$

Thus eqn [4] becomes a model for surface roughness that allows BRDF measurement to be used to find and/or verify surface roughness specifications. Of course, there are exceptions. If the 'clean, optically smooth, front surface reflector' limitations are violated, then the surface will have more than just roughness as a source of scatter and the PSD found from eqn [4] will be too large. Obtaining the same PSD from BRDF measurements made at different wavelengths, or different polarizations, is an indication that the surface is scattering 'topographically' and the PSD can be found using this technique. Because some scatter measurements can be made very rapidly, there are industry situations where scatter metrology offers a very fast means of monitoring surface quality.

A little study of eqns [4] and [5] makes it clear that ranges of spatial frequencies in the PSD correspond directly to angular ranges of the BRDF. Scatter from a single spatial frequency corresponds to a single scatter direction.

Following the above discussion, it becomes clear why the pioneering integrated (TIS) scatter measurements could be used to produce surface rms values as indicated by eqn [2]. Unfortunately, the result of eqn [4] was not available in 1961. Instead eqn [2] was derived for the special case of a smooth surface with Gaussian statistics. Nobody was thinking about spatial bandwidths and angular limits. When striking differences were found between roughness measurements made by TIS and profilometer, the Gaussian assumption became the 'whipping boy,' and TIS scatter measurements took an underserved hit. In fact, integration of eqn [4] results in the TIS result given in eqn [2] under the generally true small angle assumption that most of the scatter is close to the specular reflection. The differences make sense when the concept of spatial frequency bandwidths (or appropriate angular limits) is introduced. It becomes clear that different wavelengths, incident angles and scatter collection angles will also generate different rms values (for the same surface) and why so much confusion resulted following the definition of TIS in terms of rms roughness. There is no such thing as a unique rms roughness for a surface, anymore than there is a single spatial bandwidth for the PSD, or a single set of angles over which to integrate the BRDF. TIS measurements and rms measurements should always be given with enough information to determine bandwidth limits.

Measuring and Analyzing Scatter from Isolated Surface Features

Understanding scatter from discrete surface features has led to big changes in the entertainment business (CDs, DVDs, digitized music, and films, etc.) as well as providing an important source of metrology for the semiconductor industry as they develop smaller faster chips for a variety of modern uses. Thus, just about everybody in the modern world utilizes our understanding of scatter from discrete surface features.

On the metrology side, the roughness signals described in the last section are a serious source of background noise that limits the size of the smallest defects that can be found. Discrete surface features come in a dazzling array of types, sizes, materials, and shapes – and they all scatter differently. A 100 nm silicon particle scatters a lot differently than a 100 nm silicon oxide particle (even if they have the same shape), and a 100 nm diameter surface pit will have a different scatter pattern. Models describing scatter from a variety of discrete surface features have been developed. Although many of these are kept confidential for competitive reasons, NIST offers some models publicly through a web site.

Confirming a model requires knowing exactly what is scattering the light. In order to accomplish this, depositions of PSLs of known size are made on the surface. Scatter is measured from one or more spheres. A second measurement of background scatter is then subtracted and the net BRDF is converted to DSC units using the known (measured) illuminated spot size. Measurements of this type have been used to confirm discrete feature scatter models. The model is then used to calculate scatter from different diameters and materials. Combined with the model for surface roughness to evaluate system noise, this capability allows signal to noise evaluation of different defect scanners designs.

Practical Industrial Instrumentation

Surface defects and particles, smaller than 100 nm are now routinely found on silicon wafers using scatter instruments known as particle scanners. Thousands of these instruments (with price tags approaching a million dollars each depending on type and use) are in daily use. These instruments are truly amazing; they inspect 200 mm wafers at a rate of one every thirty seconds reporting feature location, approximate size, and in some cases even type (pit or particle) by analyzing scatter signals that last only about 100 nanoseconds. Similar systems are now starting to be used in the computer disk industry and flat panel display inspection of surface features will follow.

Scanners are calibrated with PSLs of different sizes. Because it is impossible to identify all defect types and because diameter has little meaning for irregularly shaped objects, feature 'size' is reported in 'PSL equivalent diameters.' This leads to confusing situations for multiple detector systems, which (without some software help) would report different sizes for real defects that scatter much differently than PSLs. These difficulties have caused industry confusion similar to the 'Gaussian statistics/bandwidth limited' issues encountered before roughness scatter was understood. The publication of international standards relating to scanner calibration has reduced the level of confusion.

Scatter Related Standards

Early scatter related standards for BRDF, TIS, and PSD calculations were written for ASTM (the American Society for Testing Materials), but as the industrial need for these documents moved to the semiconductor industry, there was pressure to move the documents to SEMI (Semiconductor Equipment and Materials Inc.), which is an international organization. By 2004, the process of rewriting the ASTM documents in SEMI format was well underway in the Silicon Wafer Committee. Topics covered include: surface defect specification (M35), defect capture rate (M50), scanner specifications (M52), scanner calibration (M53), particle deposition testing (M58), BRDF measurement (ME1392), and PSD calculation (MF1811). This body of literature is probably the only place where all aspects of these related problems are brought together in one place.

Conclusion

The bottom line is that scatter measurement and analysis has moved from an art to a working metrology. Dozens of labs around the world can now take the same sample and get about the same measured BRDF from it. Thousands of industrial surface scanners employing scattered light signals are in use every day on every continent. In virtually every house in the modern world there is at least one entertainment device that depends on scatter signals. In short – scatter works.

See also

Scattering: Scattering Theory.

Further Reading

Bennett JM and Mattsson L (1995) *Introduction to Surface Roughness and Scattering*, 2nd edn. Washington, DC: Optical Society of America.

Bennett HE and Porteus JO (1961) Relation between surface roughness and specular reflectance at normal incidence. *Journal of the Optical Society of America* 51: 123.

Cady FM, Bjork DR, Rifkin J and Stover JC (1989) BRDF error analysis. *Proceedings of SPIE* 1165: 154–164.

Church EL and Zavada JM (1975) Residual surface roughness of diamond-turned optics. *Applied Optics* 14: 1788.

Doicu A, Eremin Y and Wriedt T (2000) *Acoustic & Electromagnetic Scattering Analysis Using Discrete Sources*. San Diego, CA: Academic Press.

Gu ZH, Dummer RS, Maradudin AA and McGurn AR (1989) Experimental study of the opposition effect in the scattering of light from a randomly rough metal surface. *Applied Optics* 28(3): 537.

Nicodemus FE, Richmond JC, Hsia JJ, Ginsberg IW and Limperis T (1977) *Geometric Considerations and Nomenclature for Reflectance*. NBS Monograph 160, US Dept. of Commerce.

Scheer CA, Stover JC and Ivakhnenko VI (1998) Comparison of models and measurements of scatter from surface-bound particles. *SPIE Proceedings* 3275.

Schiff TF, Stover JC and Bjork DR (1992) Mueller matrix measurements of scattered light. *Proceedings of SPIE* 1753: 269–277.

Stover JC (1975) Roughness characterization of smooth machined surfaces by light scattering. *Applied Optics* 14(8): 1796.

Stover JC (1995) *Optical Scattering: Measurement and Analysis*, 2nd edn. Bellingham, WA: SPIE.

Stover JC (2001) Calibration of particle detection systems. In: Alain Diebold (ed.) *Handbook of Silicon Semiconductor Metrology*, chap. 21. New York: Marcel Dekker.

Wolfe WL and Bartell FO (1982) Description and limitations of an automated scatterometer. *Proceedings of SPIE* 362-30.

Spectrometers

K A More, SenSyTech, Inc. Imaging Group, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Sp7ectrometers were developed after the discovery that glass prisms disperse light. Later, it was discovered that diffraction from multiple, equally spaced, wires or fibers also dispersed light. About 100 years ago, Huygens proposed his wave theory of light and Fraunhofer developed diffraction theory, which allowed scientific development of diffraction gratings.

These discoveries then led to the development of spectrometers. Light theory was sufficiently developed such that spectrometer designs, developed over 100 years ago, are still being used today. These theories and designs are briefly described along with comments on how current technology has improved upon these designs. This is followed by some examples of imaging spectrometers which have wide spectral coverage from 450 nm to 14 µm and produce images of more than 256 spectral bands.

The basic elements of a spectroscopic instrument are shown in Figure 1. The source, or more usually an image of the source, fills an entrance slit and the radiation is collimated by either a lens or mirror. The radiation is then dispersed, by either a prism or a grating, so that the direction of propagation of the radiation depends upon its wavelength. It is then brought to a focus by a second lens or mirror and the spectrum consists of a series of monochromatic images of the entrance slit. The focused radiation is detected, either by an image detector such as a photographic plate, or by a flux detector such as a photomultiplier, in which case the area over which the flux is detected is limited by an exit slit. In some cases the radiation is not detected at this stage, but passes through the exit slit to be used in some other optical system. As the exit slit behaves as a monochromatic source, the instrument can be regarded as a wavelength filter and is then referred to as a monochromator.

Prisms

The wavelength dependence of the index of refraction is used in prism spectrometers. Such an optical

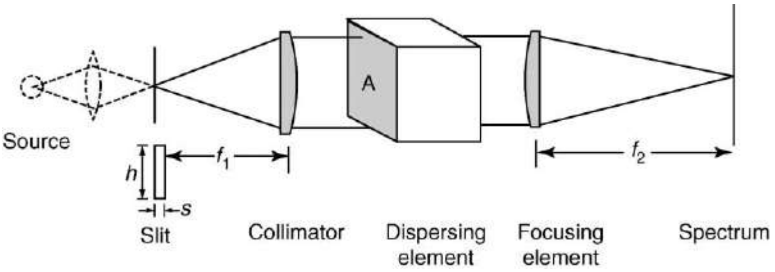

Figure 1 The basic elements of a spectroscopic instrument. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

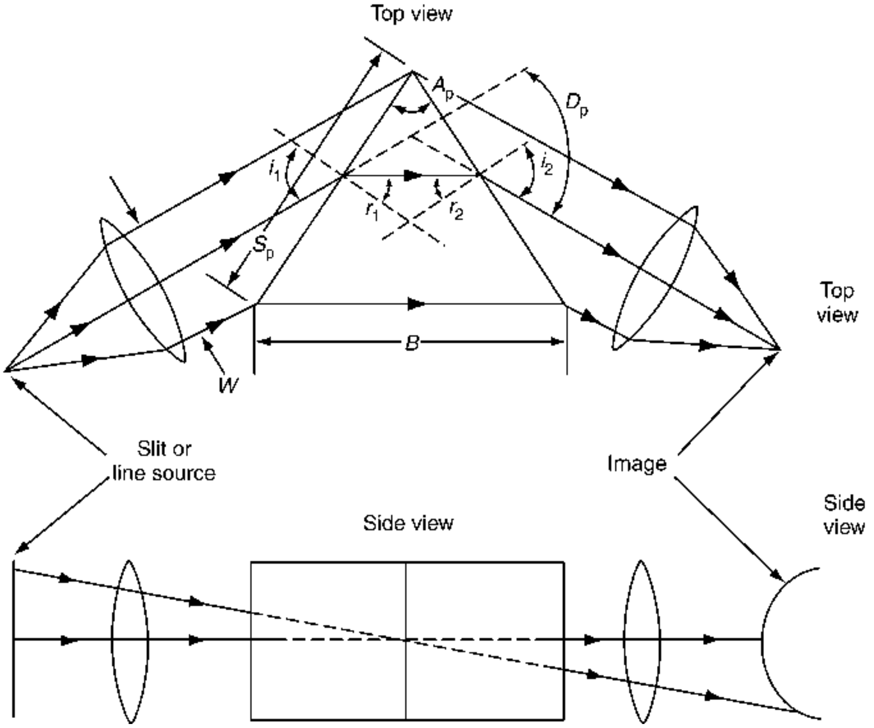

Figure 2 Elementary prism spectrometer schematic. *W* is the width of the entrance beam; S_p is the length of the prism face; and *B* is the prism base length. Reproduced with permission from *The Infrared Handbook* (1985). Ann Arbor, MI: Infrared Information Analysis Center.

element disperses parallel rays or collimated radiation into different angles from the prism according to wavelength. Distortion of the image of the entrance slit is minimized by the use of planewave illumination. Even with planewave illumination, the image of the slit is curved because not all of the rays from the entrance slit can traverse the prism in its principal plane. The prism is shown in the position of minimum angular deviation of the incoming rays in Figure 2. At minimum angular deviation, maximum power can pass through the prism. For a prism adjusted to the position of minimum deviation:

$$r_1 = r_2 = A_p/2 \qquad [1]$$

and

$$i_1 = i_2 = (D_p + A_p)/2 \qquad [2]$$

where

D_p = angle of deviation
A_p = angle of prism
r_1 and r_2 = internal angles of refraction
i_1 and i_2 = angles of entry and exit.

The angle of deviation, D_p, varies with wavelength. The resulting angular dispersion is defined as $\mathrm{d}D_p/\mathrm{d}\lambda$, while the linear dispersion is $\mathrm{d}x/\mathrm{d}\lambda = F\,\mathrm{d}D_p/\mathrm{d}\lambda$, where F is the focal length of the camera or imaging lens and x is the distance across the image plane.

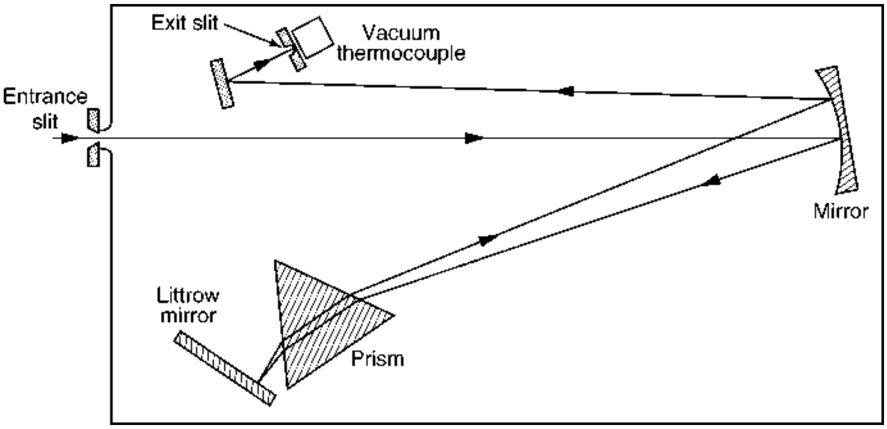

Figure 3 Infrared spectrograph of the Littrow-type mount with a rock salt prism. Reproduced with permission from *The Infrared Handbook* (1985) Ann Arbor, MI: Infrared Information Analysis Center.

It can be shown that:

$$\mathrm{d}D_\mathrm{p}/\mathrm{d}\lambda = [B/W][\mathrm{d}n/\mathrm{d}\lambda] = [\mathrm{d}D_\mathrm{p}/\mathrm{d}n][\mathrm{d}n/\mathrm{d}\lambda] \quad [3]$$

where

B = base length of the prism,
W = width of the illumination beam
n = index of refraction

while

$$\mathrm{d}x/\mathrm{d}\lambda = F[B/W][\mathrm{d}n/\mathrm{d}\lambda] \quad [4]$$

One may define the resolving power, RP, of an instrument as the smallest resolvable wavelength difference, according to the Rayleigh criterion, divided into the average wavelength in that spectral region. Thus:

$$\mathrm{RP} = \lambda/\Delta\lambda = [\lambda/\mathrm{d}D_\mathrm{p}][\mathrm{d}D_\mathrm{p}/\mathrm{d}\lambda]$$

$$= [\lambda/\mathrm{d}D_\mathrm{p}][B/W][\mathrm{d}n/\mathrm{d}\lambda] \quad [5]$$

The limiting resolution is set by diffraction due to the finite beamwidth, or effective aperture of the prism, which is rectangular. Thus:

$$\mathrm{RP} = [\lambda/(\lambda/W)][B/W][\mathrm{d}n/\mathrm{d}\lambda] \quad [6]$$

so that:

$$RP = B[\mathrm{d}n/\mathrm{d}\lambda] \quad [7]$$

If the entire prism face is not illuminated, then only the illuminated base length must be used for B.

Littrow showed that aberrations would be minimized by making the angle of incidence equal to the angle of refraction β (also known as the Littrow Configuration). Littrow used a plane mirror behind the prism for a double pass through the prism, as shown in Figure 3.

Gratings

Rowland is credited with the development of a grating mount that reduced aberrations in the spectrogram. He found that a grating ruled on a concave surface of radius R and locating the entrance and exit slits on the same radius would give the least aberrations, as shown in Figure 4. $r = R\cos\alpha$ is the distance to the entrance slit *(r)* and $r_1 = R\cos\beta$ is the distance to the focal point for the exit slit (r_1). Rowland showed that:

$$\cos\alpha/R - \cos^2\alpha/r + \cos\beta/R - \cos^2\beta/r_1 = 0 \quad [8]$$

One solution to this equation is for α and β to each be zero and

$$r = R\cos\alpha$$

and

$$r_1 = R\cos\beta$$

This condition is met if r and r_1 lie on the Rowland circle.

There are various ways in which the Rowland circle condition may be satisfied and some of them are shown in Figure 5. The simplest mounting of all is that due to Paschen and Runge (Figure 5a) in which the entrance slit is positioned on the Rowland circle and a photographic plate (or plates) is constrained to fit the Rowland circle. Alternatively, for photoelectric detection, a series of exit slits is arranged around the Rowland circle each having its own detector. In the latter case, the whole spectrum is not recorded, only a series of predetermined wavelengths, but when used in this ‘polychromator’ form, it is very rugged and convenient for applications such as the routine analysis of samples of metals and alloys. In this case, slits and detectors are set up to measure the light in various spectral lines, each characteristic of a particular component or trace element.

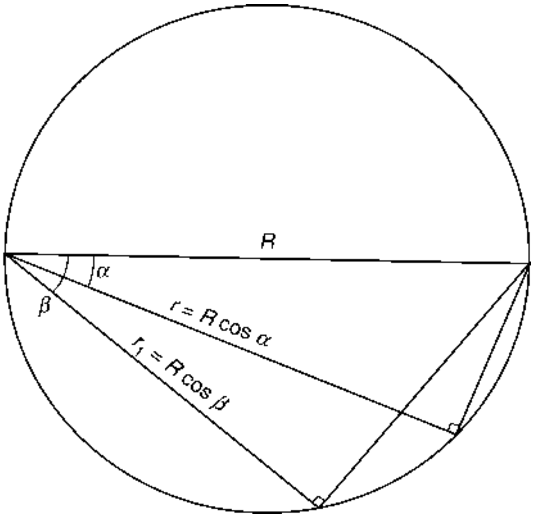

Figure 4 The construction of the Rowland Circle. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

The oldest concave grating mount is that designed by Rowland himself and which bears his name (Figure 5b). In this case, the grating and photographic plates are fixed at opposite ends of the diameter of the Rowland circle by a moveable rigid beam. The entrance slit remains fixed above the intersection of two rails at right angles to each other and along which the grating plate holder (or the exit slit) is free to move. In this way, this entrance slit, grating, and plate holder are constrained always to lie on the Rowland circle and it has the advantage that the dispersion is linear, which is useful in the accurate determination of wavelengths. Unfortunately, this mounting is rather sensitive to small errors in the position of the entrance slit and in the orthogonality of the rails, and is now very rarely used.

A variation on this mounting was devised by Abney, who again mounted the grating and the plate holder on a rigid bar at opposite ends of the diameter of the Rowland circle (Figure 5c). The entrance slit is mounted on a bar of length equal to the radius of the Rowland circle. In this way, the slit always lays on the Rowland circle, but it had to rotate about its axis in order that the jaws should remain perpendicular to the line from the grating to

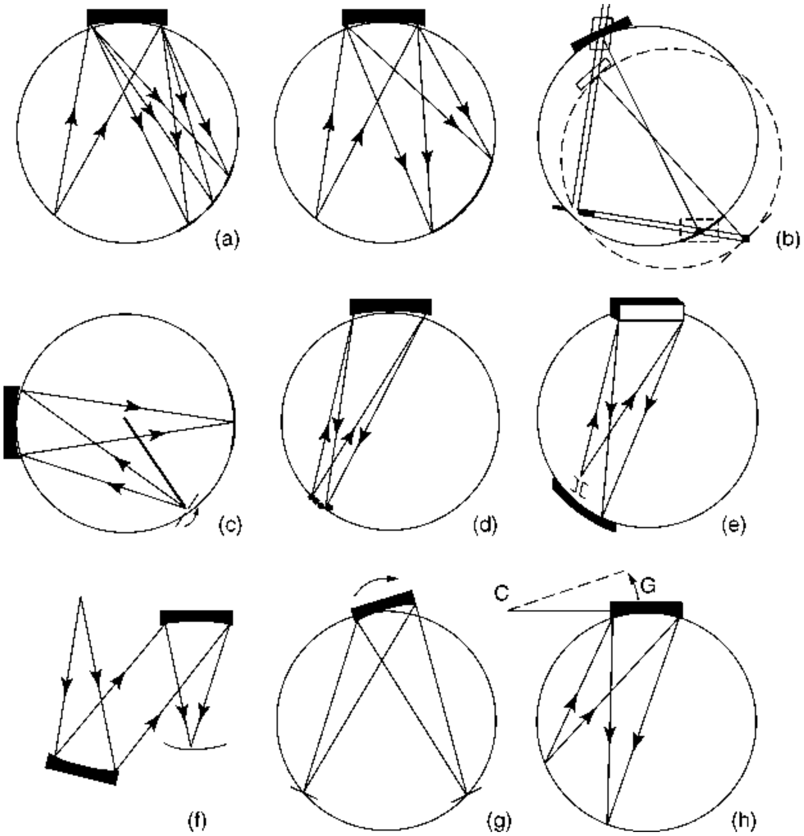

Figure 5 Mountings of the concave grating: (a) Paschen–Runge; (b) Rowland; (c) Abney; (d) and (e) Eagle; (f) Wadsworth; (g) Seya–Namioka; (h) Johnson–Onaka. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

the slit. It also had the disadvantage that the source must move with the exit slit, which could be inconvenient.

In the Eagle mounting (Figures 5d and 5e), the angles of incidence and diffraction are made equal, or very nearly so, as with Littrow mounting for plane gratings. Optically, this system has the advantage that the astigmatism is generally less than that of the Paschen–Runge or Rowland mounting, but on the other hand, the dispersion is nonlinear. From a mechanical point of view, it has the disadvantage that it is necessary with great precision both to rotate the grating and to move it nearer to the slits in order to scan the spectrum. However, it does have the practical advantage that it is much more compact than other mountings, and this is of particular importance when we bear in mind the need to enclose the instrument in a vacuum tank. Ideally, the entrance slit and exit slit or photographic plate should be superimposed if we are to set $\alpha = \beta$. In practice, of course, the two are displaced either sideways, in the plane of incidence as shown in Figure 5, or out of the plane of incidence, in which case the entrance slit is positioned below the meridonal plane and the plate holder just above it, as shown in Figure 5. The out-of-plane configuration is generally used for spectrographs and the in-plane system for monochromators. The penalty incurred in going out of plane is that coma is introduced in the image, and slit curvature becomes more important. This limits the length of the ruling that can effectively be used.

The second well-known solution to the Rowland equation is the Wadsworth mounting (Figure 5f), in which the incident light is collimated, so r is set at infinity and the focal equation reduces to $r_1 = R\cos^2\beta/(\cos\alpha + \cos\beta)$. One feature of this mounting is that the astigmatism is zero when the image is formed at the center of the grating blank, i.e., when $\beta = 0$. It is a particularly useful mounting for applications in which the incident light is naturally collimated (for example, in rocket or satellite astronomy, spectroheliography and in work using synchrotron radiation). However, if the light is not naturally collimated, the Wadsworth mount requires a collimating mirror, so one has to pay the penalty of the extra losses of light at this mirror. The distance from the grating to the image is about half that for a Rowland circle mounting which makes the instrument more compact and, since the grating subtends approximately four times the solid angle, there is a corresponding increase in the brightness of the spectral image.

Not all concave grating mountings are solutions to the Rowland equation. In some cases, other advantages may compensate for a certain defect of focus. A particularly important example of this is the Seya–Namioka mounting (Figure 5g), in which the entrance slit and exit slit are kept fixed and the spectrum is scanned by a simple rotation of the grating (Figure 5g). In order to achieve the optimum conditions for this mounting, we set $\alpha = \varphi + \theta$ and $\beta = \theta - \varphi$, where 2φ is the angle subtended at the grating by the entrance and exit slit and θ is the angle through which the grating is turned. The amount of defocus is given by:

$$F(\theta, \varphi, r, r_1) = [\cos^2(\theta + \varphi)/r] + [\cos(\theta + \varphi)/R] + [\cos^2(\theta - \varphi)/r_1] + [\cos(\theta - \varphi)/R] \quad [9]$$

and the optimum conditions are those for which $F(\theta,\varphi,r,r_1)$ remains as small as possible as θ is varied over the required range. Seya set F and three derivatives of F with respect to θ equal to zero for $\theta = 0$ and obtained the result:

$$\varphi = \sin^{-1}(1/\sqrt{3}) = 35°15' \quad [10]$$

and

$$r = r_1 = R\cos\varphi \quad [11]$$

which corresponds to the Rowland circle either in zero order or for zero wavelength. In practice, it is usual to modify the angle slightly so that the best focus is achieved in the center of the range of interest rather than at zero wavelength.

The great advantage of the Seya–Namioka mounting is its simplicity. An instrument need consist only of a fixed entrance and exit slit, and a simple rotation of the grating is all that is required to scan the spectrum. It is, in fact, simpler than instruments using plane gratings. Despite the fact that at the ends of the useful wavelength range the resolution is limited by the defect of focus and the astigmatism is particularly bad, the Seya–Namioka mounting is very widely used, particularly for medium-resolution rather than high-resolution work.

A similar simplicity is a feature of the Johnson–Onaka mounting (Figure 5h). Here again, the entrance and exit slits remain fixed, but the grating is rotated about an axis which is displaced from its center; in this way, it is possible to reduce the change of focus that occurs in the Seya–Namioka mounting. The system is set up so that at the center of the desired wavelength range, the slits and grating lie on the Rowland circle, as shown in Figure 5h. The optimum radius of rotation, i.e., the distance GC, was found by

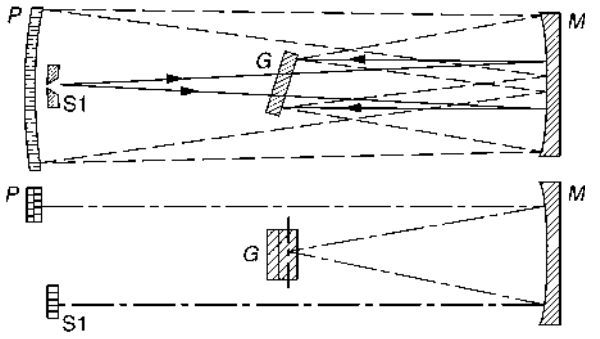

Figure 6 Ebert mounting of the plane grating designed by Fastie. 'Sl' is the entrance slit; *G* is the grating; *M* is the concave mirror; and *P* is the photographic plate. The horizontal section is at the top and the vertical section is at the bottom. Reproduced with permission from *The Infrared Handbook* (1985) Ann Arbor, MI: Infrared Information Analysis Center.

Onaka to be:

$$\mathrm{GC}_{\mathrm{opt}} = [R \sin\tfrac{1}{2}(\alpha+\beta)]/[1 - \tfrac{1}{2}\tan(\alpha+\beta) \times (\tan\beta - \tan\alpha)] \qquad [12]$$

Another non-Rowland spectrometer is the Ebert–Fastie mount, which mounts the slit, flat grating, and a detector array in the arrangement shown in Figure 6. Ebert first developed the design using two separate concave mirrors, one for collimating the incident beam and the second to focus the diffracted spectrum. Fastie used a single but larger concave mirror, which simplified the mounting structure and produced a rugged, compact spectrometer that has been used in rocket flights and space satellite observatories for astronomical and upper atmospheric applications. The Czerny–Turner mount is similar to the Ebert mount, except the flat grating is located in the same plane that contains the entrance and exit slits.

Advanced Spectrometers

While the above spectrometer designs are still used, major advances in implementation are now available. Ray tracing allows the designer to quantify the aberrations and determine solutions to remove them. Aspheric optic elements can now be fabricated to correct aberrations. Gratings can be ruled on aspherical surfaces to not only disperse the light beam, but also be an element in the optical design. Holography has been developed to etch gratings for use over a wide spectral range and to not only disperse the light, but also work as an optical element. Because holographic gratings are chemically etched, there are no machine burrs to scatter light and the hologram is free of periodic differences in groove widths that create ghosts in ruled gratings. Linear and array photodetectors have replaced film. An advantage of film was its ability to fit curved focal planes. Photodetectors are etched into flat wafers of the photodiode material. To adapt flat arrays to a curved focal plane, fiber optic face plate couplers have been ground on one side to match the focal plane curvature and the flat back side is either optically coupled to the detector or closely coupled for proximity focusing. The photodetector arrays are available in multiple materials to cover the spectral range, from soft X-rays to the thermal infrared. Often cooling is required to obtain low noise. They are now available with sensitivities never reached with film and the ease of coupling the output data to a computer for real-time analysis has taken much of the labor out of analyzing the spectrograph. Another advance is precision motor drives with precision motion sensors, some using laser interferometry, to provide feedback on automated movement of gratings, slits, and detectors as a spectrometer cycles through its wavelength range.

Imaging Spectrometers

Spectrometers that image the spectral characteristics in each pixel of the image to form a data cube, as shown in Figure 7, are named imaging spectrometers, and if the spectra have high resolution and blocks of consecutive neighbors, then the data are called hyperspectral. Imaging spectrometers have several applications that range from medical to remote sensing of land use and environment. This article will cover examples for remote sensing imaging spectrometers which require spectral coverage in all of the atmospheric windows from the UV to thermal IR.

Imaging spectrometers require a combination of spectrometers, light collecting optics, and scan mechanisms to scan the instantaneous field of view of the spectrometer over a scene. Remote sensing of the Earth applications requires an aerial platform. Either a helicopter, an aircraft, or orbital space satellites are used. The platform motion is used as part of the scanning process so that the optics image a single point on the ground so that a scanner (called a line scanner) scans a long line that is cross tracked to the platform motion. Or, the optics image a slit that is parallel to the platform track that covers many scan lines and a scanner moves the slit cross track to the platform motion. This scanner is called a whiskbroom scanner. Or, the optics image a large slit so no scan mechanism is needed other than the platform

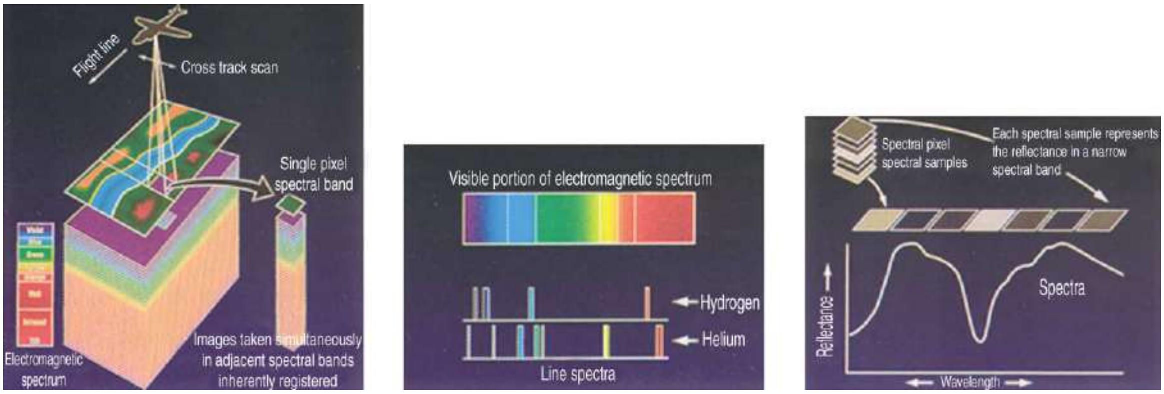
Flight line
Cross track scan
Single pixel spectral band
Electromagnetic spectrum
Images taken simultaneously in adjacent spectral bands inherently registered
Visible portion of electromagnetic spectrum
Hydrogen
Helium
Line spectra
Spectral pixel spectral samples
Each spectral sample represents the reflectance in a narrow spectral band
Spectra
Reflectance
Wavelength

Figure 7 Hyperspectral Data Cube. Hyperspectral imagers divide the spectrum into many discrete narrow channels. This fine quantization ofn spectral information on a pixel by pixel basis enables researchers to discriminate the individual constituents in an area much more effectively. For example, the broad spectral bands of a multispectral sensor allow the user only to coarsely discriminate between areas of deciduous and coniferous forest, plowed fields, etc., whereas a hyperspectral imager provides characteristic signatures which can be correlated with specific spectral templates to help determine the individual constituents and possibly even reveal details of the natural processes which are affecting them.

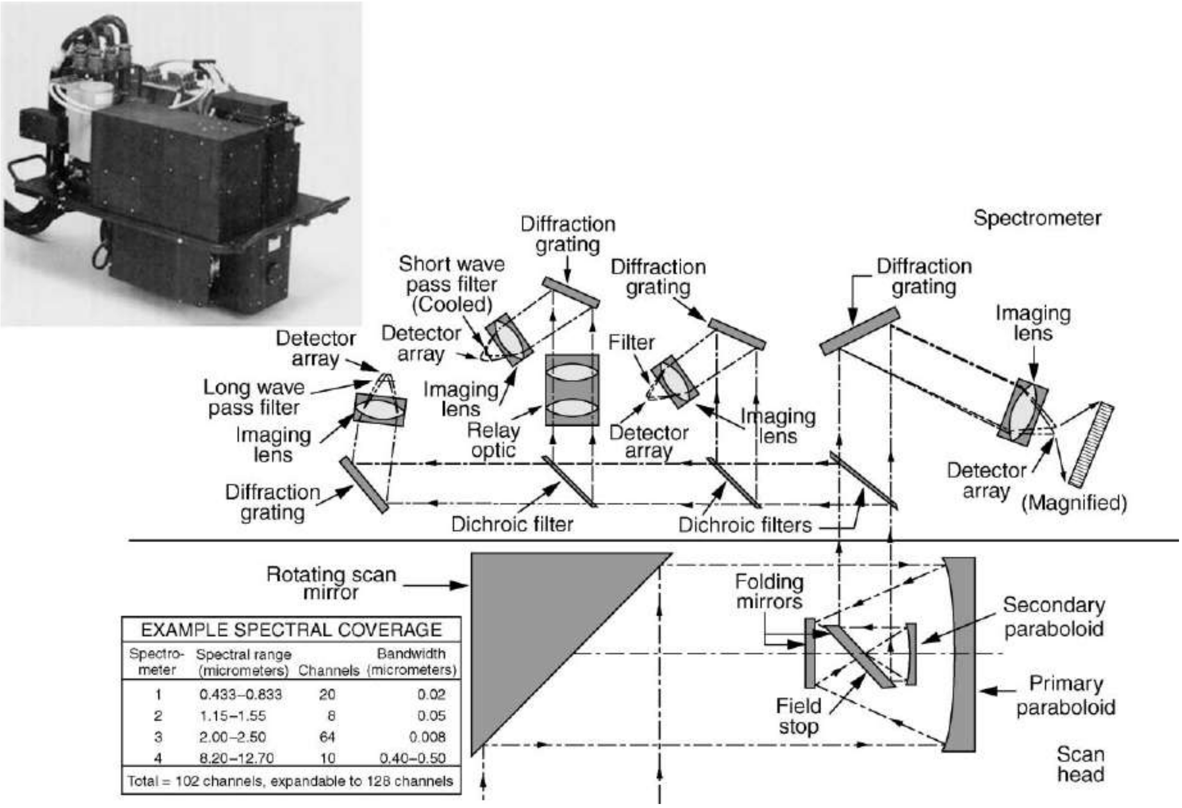
Spectrometer
Diffraction grating
Short wave pass filter (Cooled)
Diffraction grating
Diffraction grating
Imaging lens
Detector array
Detector array
Filter
Long wave pass filter
Imaging lens
Imaging lens
Imaging lens
Relay optic
Detector array
Detector array
(Magnified)
Diffraction grating
Dichroic filter
Dichroic filters
Rotating scan mirror
Folding mirrors
Secondary paraboloid
Field stop
Primary paraboloid
Scan head

EXAMPLE SPECTRAL COVERAGE

Spectro-meter	Spectral range (micrometers)	Channels	Bandwidth (micrometers)
1	0.433–0.833	20	0.02
2	1.15–1.55	8	0.05
3	2.00–2.50	64	0.008
4	8.20–12.70	10	0.40–0.50

Total = 102 channels, expandable to 128 channels

Figure 8 Multispectral infrared and visible imaging spectrometer optical schematic.

motion to form an image. This scanner is called a pushbroom scanner.

One important requirement is that all spectral measurements of a pixel be coregistered. Most airborne imaging spectrometers use a common aperture for a line scanner, or a slit for whisk and pushbroom scanners, so that platform instability from pitch roll and yaw and from inability to fly in a straight line do not compromise the coregistration of spectral data on each pixel. The image data may require geometric correction, but the spectral data are not compromised.

There is one other type of imaging spectrometer that uses a linear variable filter (also known as a wedge filter) over a 2D array of photodetectors. Each image contains a different spectral band over each row of pixels so that each frame images a scene with a different spectral band over each row of pixels. The array is oriented so the rows of spectral bands are perpendicular to the flight track. After the platform motion moves over one ground pixel, the whole array is read out and the frame is shifted one row of pixels so the second frame adds a second spectral band to each row imaged in the first frame. This frame stepping is carefully timed to the platform velocity and is repeated until each row of pixels is imaged in all spectral bands. This type of imaging spectrometer has been flown in aircraft and satellites. In aircraft, the platform motion corrupts the coregistration of the spectrum in each pixel. Extensive ground processing is required to geometrically correct each frame to improve spectral coregistration. This is a difficult task and this type of imaging spectrometer has lost favor for airborne use. Stabilized satellites have been a better platform and the wedge type of imaging spectrometer has been used successfully in space.

As examples of airborne and ground-based imaging spectrometers, a line scanner imaging

Table 1 MIVIS physical properties

	Height		*Width*		*Depth*[a]		*Weight*	
	in	*cm*	*in*	*cm*	*in*	*cm*	*lbs*	*kg*
Scan head	26.5	67.0	20.6	52.0	28.1	71.0	220	100
Electronics	40.0	102.0	19.0	48.3	24.0	61.0		
Total system weight (approx.)							460	209

[a] Not including connectors or cable bends.

Table 2 MIVIS spectral coverage (μm)

OPTICAL PORT 1			*OPTICAL PORT 3*						*OPTICAL PORT 4*		
CH #	*Band Edges*		*CH #*	*Band Edges*		*CH #*	*Band Edges*		*CH #*	*Band Edges*	
	Lower	*Upper*		*Lower*	*Upper*		*Lower*	*Upper*		*Lower*	*Upper*
1	0.43	0.45	29	2.000	2.008	61	2.250	2.258	93	8.20	8.60
2	0.45	0.47	30	2.008	2.016	62	2.258	2.266	94	8.60	9.00
3	0.47	0.49	31	2.016	2.023	63	2.266	2.273	95	9.00	9.40
4	0.49	0.51	32	2.023	2.031	64	2.273	2.281	96	9.40	9.80
5	0.51	0.53	33	2.031	2.039	65	2.281	2.289	97	9.80	10.20
6	0.53	0.55	34	2.039	2.047	66	2.289	2.297	98	10.20	10.70
7	0.55	0.57	35	2.047	2.055	67	2.297	2.305	99	10.70	11.20
8	0.57	0.59	36	2.055	2.063	68	2.305	2.313	100	11.20	11.70
9	0.59	0.61	37	2.063	2.070	69	2.313	2.320	101	11.70	12.20
10	0.61	0.63	38	2.070	2.078	70	2.320	2.328	102	12.20	12.70
11	0.63	0.65	39	2.078	2.086	71	2.328	2.336			
12	0.65	0.67	40	2.086	2.094	72	2.336	2.344			
13	0.67	0.69	41	2.094	2.102	73	2.344	2.352			
14	0.69	0.71	42	2.102	2.109	74	2.352	2.359			
15	0.71	0.73	43	2.109	2.117	75	2.359	2.367			
16	0.73	0.75	44	2.117	2.125	76	2.367	2.375			
17	0.75	0.77	45	2.125	2.133	77	2.375	2.383			
18	0.77	0.79	46	2.133	2.141	78	2.383	2.391			
19	0.79	0.81	47	2.141	2.148	79	2.391	2.398			
20	0.81	0.83	48	2.148	2.156	80	2.398	2.406			
OPTICAL PORT 2			49	2.156	2.164	81	2.406	2.414			
21	1.15	1.20	50	2.164	2.172	82	2.414	2.422			
22	1.20	1.25	51	2.172	2.180	83	2.422	2.430			
23	1.25	1.30	52	2.180	2.188	84	2.430	2.438			
24	1.30	1.35	53	2.188	2.195	85	2.438	2.445			
25	1.35	1.40	54	2.195	2.203	86	2.445	2.453			
26	1.40	1.45	55	2.203	2.211	87	2.453	2.461			
27	1.45	1.50	56	2.211	2.219	88	2.461	2.469			
28	1.50	1.55	57	2.219	2.227	89	2.469	2.477			
			58	2.227	2.234	90	2.477	2.484			
			59	2.234	2.242	91	2.484	2.492			
			60	2.242	2.250	92	2.492	2.500			

spectrometer and a ground-based scanned push broom scanner are described below.

MIVIS (multispectral infrared and visible imaging spectrometers) hyperspectral scanner developed by SenSyTech Imaging Group (formerly Daedalus Enterprises) for the CNR (Consiglio National Researche) of Italy is a line scanner imaging spectrometer. The optical schematic is shown in Figure 8. A rotating 45° mirror scans a line of pixels on the ground. Each pixel is collimated by a parabolic mirror in a Gregorian telescope mount. The collimated beam is reflected by a Pfund assembly to an optical bench that houses four spectrometers. An aperture pinhole in the Pfund assembly defines a common instantaneous field of view for each pixel. The collimated beam is then split off with either thin metallic coated or dielectric coated dichroics to four spectrometers. The thin metallic mirrors reflect long wavelengths and transmit short wavelengths. The multiple dielectric layers cause interference such that light is reflected at short wavelengths and transmitted at long wavelengths. After splitting off four wide bands (visible, near infrared, mid-wave infrared, and thermal infrared) each band is dispersed in its own spectrometers.

The wavelength coverage for each spectrometer is based on the wavelength sensitivity of different photodetector arrays. The visible spectrometer uses a silicon photodiode array; the near infrared spectrometer, an InGaAs array, the mid-infrared a InSb array and the thermal infrared, a MCT (mercury doped cadmium teluride) photo-conductor array. Note the beam expander in spectrometer 3 for the

Table 3 LAFS technical specifications

Spatial resolution	1.0 mrad IFOV vertical and horizontal (square pixels). Optical lens for 0.5 mrad IFOV
Spatial coverage	15° TFOV horizontal, nominal. A second, 7.5° TFOV, is an option
Focus range	5′ to infinity
Spectral resolution	5.0 nm per spectral channel, sampled at 2.5 nm interval
Spectral range	400–1100 nm
Dynamic range	8 bit (1 part in 256)
Illumination	Solar illumination ranging from 10:00 a.m. to 2:00 p.m. under overcast conditions to full noon sunshine
Acquisition time	1.5 sec for standard illuminations. Option for longer times for low light level conditions
Viewfinder	Near real time video view of the scene
Calibrations	Flat fielding to compensate for CCD variations in responsivity. Special calibration
Data displays	Single band imagery, selectable Waterfall Chart (one spatial line by spectral, as acquired by the CCD). Single pixel spectral display
Data format	Convertible to format compatible with image processors
Data storage	Replaceable hard disc, 14 data cubes/disk

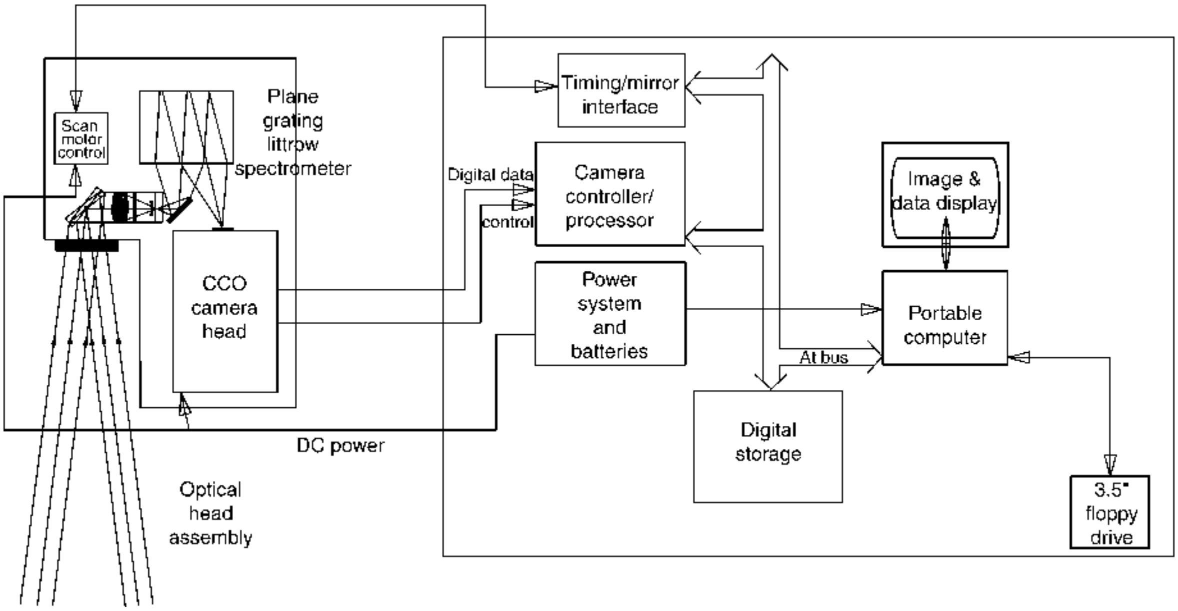

Figure 9 LAFS system block diagram.

mid-infrared (IR). Since wavelength resolution increases with the size of the illuminated area of a diffraction grating a beam expander was needed to meet the specified mid-IR resolution.

The physical characteristics of MIVIS are given in Table 1 and the spectral coverage in Table 2.

An example of a push broom imaging spectrometer is the large area fast spectrometer (LAFS), which is a field portable, tripod mounted, imaging spectrometer. LAFS uses a push broom spectrometer with a galvanometer driven mirror in front of the entrance slit. The mirror is stepped to acquire a 2D image × 256 spectral bands data cube.

LAFS was developed for the US Marine Corps under the direction of the US Navy Coastal Systems Station of the Dahlgren Division. The technical specifications are shown in Table 3. Various concepts for the imaging spectrometer were studied and the only concept that could meet the above specification was a Littrow configuration grating spectrometer that imaged one line in the scene and dispersed the spectra perpendicular to the line image onto a CCD array. A galvanometer mirror stepped the line image over the scene to generate a spectral image of the scene. A block diagram of LAFS is shown in Figure 9 and a photograph of the prototype in Figure 10. Figure 11 is a photograph of the optical head. Data from a single frame are stored in computer (RAM) memory then transferred to a replaceable hard disk for bulk storage. Data are collected in a series of 256 CCD frames and each CCD frame has a line of 256 pixels along one axis and 256 spectral samples along the second axis. The 256 CCD frames constitute a data cube as shown in Figure 12. This does not allow field viewing of an image in a single spectral band. Earth View™ software is used in the field portable computer

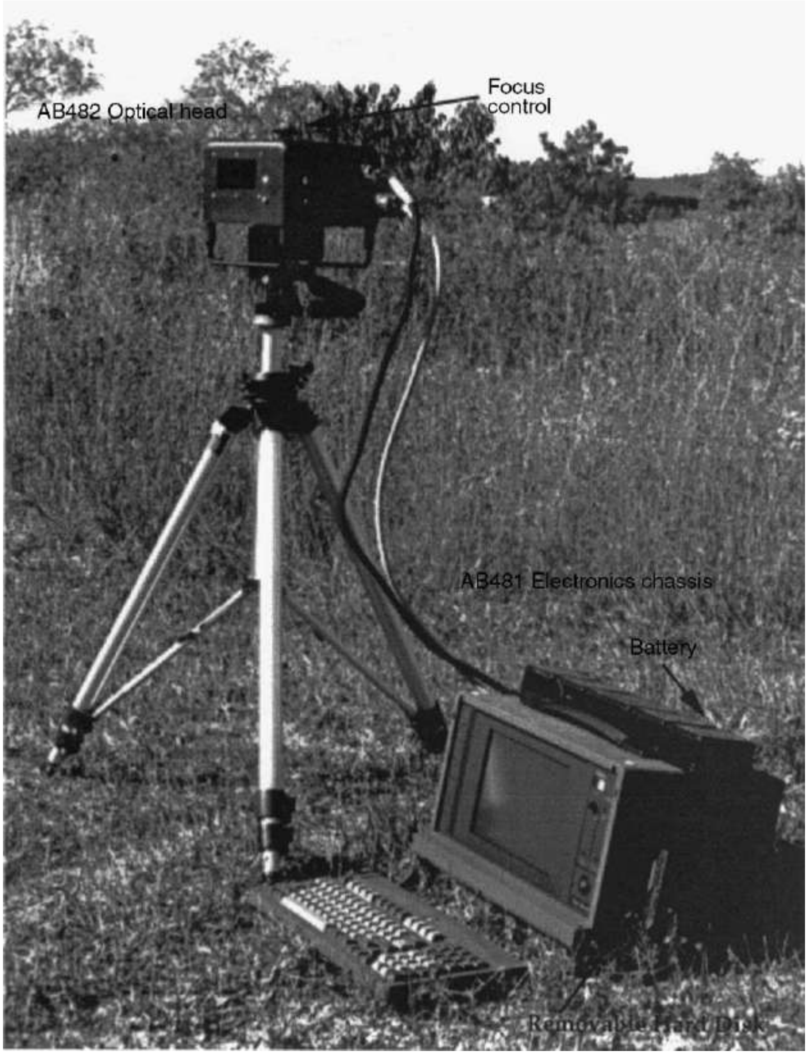

Figure 10 LAFS.

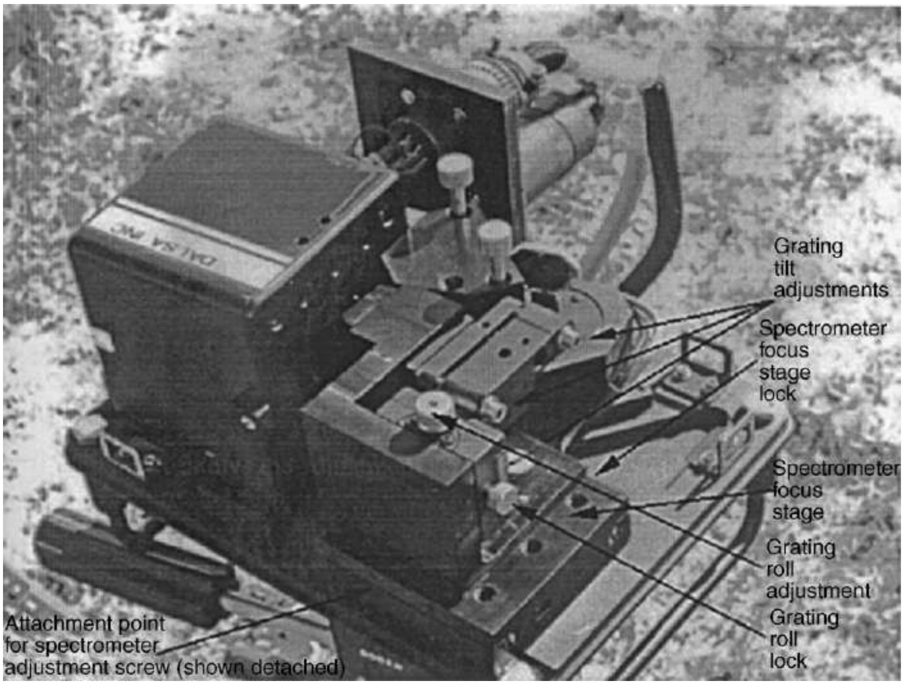
Grating tilt adjustments
Spectrometer focus stage lock
Spectrometer focus stage
Grating roll adjustment
Grating roll lock
Attachment point for spectrometer adjustment screw (shown detached)

Figure 11 LAFS optical head.

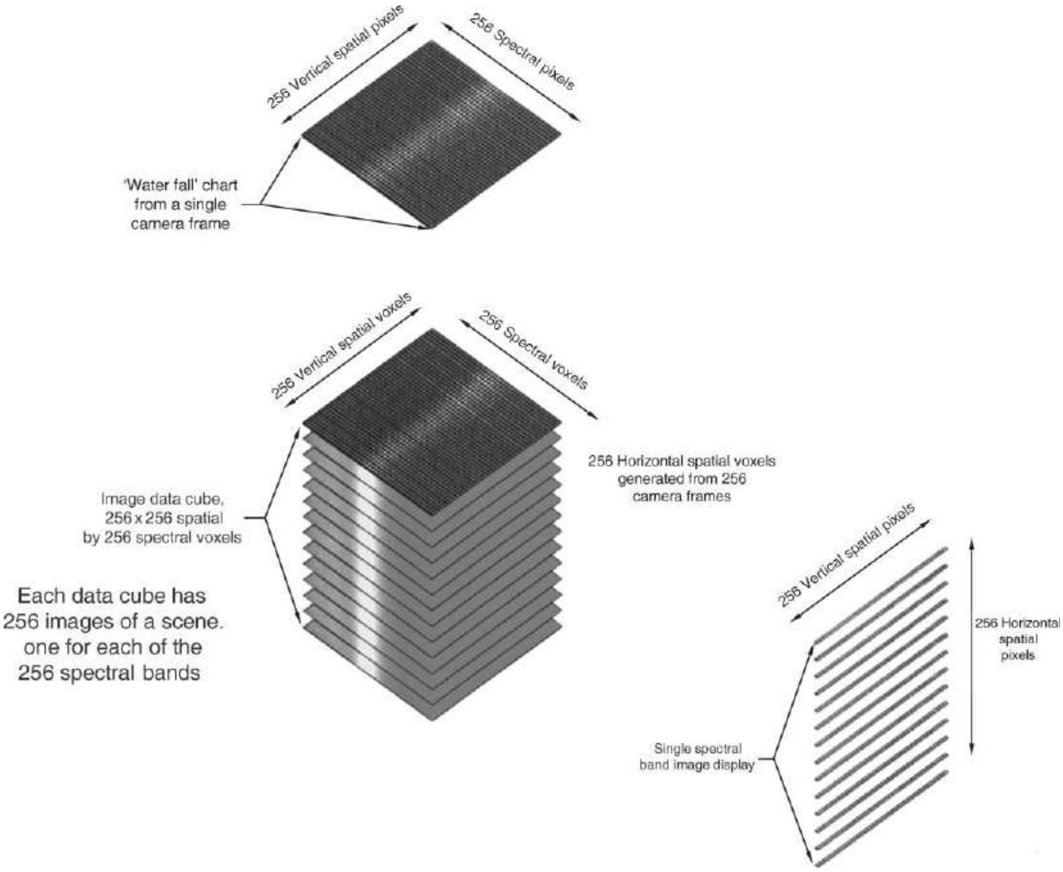
256 Vertical spatial pixels
256 Spectral pixels
'Water fall' chart from a single camera frame
256 Vertical spatial voxels
256 Spectral voxels
256 Horizontal spatial voxels generated from 256 camera frames
Image data cube, 256 x 256 spatial by 256 spectral voxels
Each data cube has 256 images of a scene. one for each of the 256 spectral bands
256 Vertical spatial pixels
256 Horizontal spatial pixels
Single spectral band image display

Figure 12 Arrangement of image memory and construction of image displays.

Table 4 LAFS physical properties

Dimensions and weight	*Size (inches)*	*weight (lbs)*
Optical head	7 × 9 × 8	11
Electronics	12 × 15.5 × 9	30
Battery	4.5 × 16.5 × 2	7
		48 + cables

Power: Less than 150 W at 12 VDC; Battery time – 45 min/charge.

to reorder the data cube into 256 spatial images, one for each spectral band as shown in Figure 12.

The prototype is packaged in two chassis, a compact optical head and a portable electronic chassis. The size, weight, and power are shown in Table 4.

LAFS was designed to also work as an airborne imaging spectrometer. The framing mirror can be locked so it views a scene directly below the aircraft (push broom mode) or the mirror can be programmed to sweep a 256 lines cross track to the aircraft flight path (whisk broom mode). In the whisk broom mode, the scan mirror rotation arc can be increased for a wider field of view than in the push broom mode. LAFS illustrates the decrease in size, weight, and power of a push broom imaging spectrometer that results from a 256 × 256 pixel array rather than single pixel line arrays of a line scanner. The array increases integration time of a line by a factor of 256, which allows longer detector integration dwell time on each pixel and thus smaller light collecting optics.

List of Units and Nomenclature

Data cube	Multiple images of a scene in many spectral bands. The spectra of each pixel can be obtained by plotting the spectral value on the same pixel in each spectral band image.
Pixel	An image point (or small area) defined by the instantaneous-field-of-view of either the optics, entrance aperture or slit, and the detector area, or by some combination of these components. A pixel is the smallest element in a scene that is resolved by the imaging system.
Instantaneous-field-of-view (IFOV)	The IFOV is used in describing scanners to define the size of the smallest field-of-view (usually in milliradians or microradians) that be resolved by a scanner system.
Field-of-view	The total size in angular dimensions of a scanner or imager.
Hyperspectral	A data cube with many (>48) spectral bands.
Line scanners	An optical/mechanical system that scans one pixel at a time along a line of a scene.
Whisk broom	An optical/mechanical system that scans two or more lines at a time.
Push broom	An optical system that images a complete line at a time.

See also

Diffraction: Diffraction Gratings; Fraunhofer Diffraction. **Fiber Gratings**. **Geometrical Optics:** Prisms. **Imaging:** Hyperspectral Imaging; Interferometric Imaging. **Interferometry:** Overview. **Modulators:** Acousto-Optics. **Optical Materials:** Color Filters and Absorption Glasses. **Semiconductor Materials:** Dilute Magnetic Semiconductors; Group IV Semiconductors, Si/SiGe; Large Gap II–VI Semiconductors; Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures. **Spectroscopy:** Fourier Transform Spectroscopy; Raman Spectroscopy.

Further Reading

Baselow R, Silvergate P, Rappaport W, *et al.* (1992) *HYDICE (Hyperspectral Digital Imagery Collection Experiment Instrument Design*. International Symposium on Spectral Sensing Research.

Bianchi, R, Marino, CM and Pignatti, S *Airborne Hyperspectral Remote Sensing in Italy.* Pomezia (Roma), Italy: CNR, Progetto LARA, 00040.

Borough HC, Batterfield OD, Chase RP, Turner RW and Honey FR (1992) *A Modular Wide Field of View Airborne Imaging Spectrometer Preliminary Design Concept.* International Symposium on Spectral Sensing Research.

Carmer DC, Horvath R and Arnold CB (1992) *M7 Multispectral Sensor Testbed Enhancements*. International Symposium on Spectral Sensing Research.

Hackwell JA, Warren DW, Bongiovi RP, *et al.* (1996) LWIR/MWIR Imaging Hyperspectral Sensor for Airborne and Ground-Based Remote Sensing. In: Descour M and Mooney J (eds) *Imaging Spectrometry II*. Proc. SPIE 2819, pp. 102–107.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

James JF and Sternberg RS (1969) *The Design of Optical Spectrometers*. London: Chapman and Hall.

Kulgein NG, Richard SP, Rudolf WP and Winter EM (1992) *Airborne Chemical Vapor Detection Experiment*. International Symposium on Spectral Sensing Research.

Lucey PG, Williams T, Horton K, Hinck K and Buchney C (1994) *SMIFTS: A Cryogenically Cooled, Spatially Modulated, Imaging Fourier Transform Spectrometer for Remote Sensing Applications*. International Symposium on Spectral Sensing Research.

Mourales P and Thomas D (1998) Compact Low-Distortion Imaging Spectrometer for Remote Sensing. *SPIE Conference on Imaging Spectrometry IV* 3438: 31–37.

Pagano TS (1992) *Design of the Moderate Resolution Imaging Spectrometer – NADIR (MODIS-N).* American Society of Photogrametry & Remote Sensing, pp. 374–385.

Stanich CG and Osterwisch FG (1994) *Advanced Operational Hyperspectral Scanners MIVIS and AHS.* First International Airborne Remote Sensing Conference and Exhibition, Strasbourg, France: Environmental Research Institute of Michigan.

Sun X and Baker JJ (2001) *A Hight Definition Hyperspectral Imaging System.* Proceedings, Fifth International Airborne Remote Sensing Conference and Exhibition, Veridian.

Torr MR, Baselow RW and Torr DG (1982) Imaging Spectrascopy of the Thermosphere from the Space Shuttle. *Applied Optics* 21: 4130.

Torr MR, Baselow RW and Mounti J (1983) An imaging spectrometric observatory for spacelab. *Astrophys. Space Science*. 92: 237.

Vane G, Green RO, Chrien TG, Enmark HT, Hansen EG and Porter WM (1993) The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote Sensing of the Environment* 44: 127–143.

Warren CP, Anderson RD, Velasco A, Lin PP, Lucas JR and Speer BA (2001) *High Performance VIS and SWIR Hyperspectral Imagers for Airborne Applications.* Fifth International Conference and Exhibition, Veridian.

Wolfe WL (1997) *Introduction to Imaging Spectrometers.* SPIE Optical Engineering Press.

Wolfe WL and Zisfis GJ (eds) (1985) *The Infrared Handbook*, revised edition. Washington, DC: Office of Naval Research, Department of the Navy.

Zywicki RW, More KA, Holloway J and Witherspoon N (1994) *A Field Portable Hyperspectral Imager: The Large Area Fast Spectrometer.* Proceedings of the International Symposium on Spectral Sensing Research.

Telescopes

M M Roth, Astrophysikalisches Institut Potsdam, Potsdam, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In a broad sense, telescopes are optical instruments which provide an observer with an improved view of a distant object, where improved may be defined in terms of magnification, angular resolution, and light collecting power. Historically, the invention of the telescope was a breakthrough with an enormous impact on fundamental sciences (physics, astronomy), and, as a direct consequence, on philosophy; but also on other technological, economical, political, and military developments. It seems that the first refractive telescopes were built in the Netherlands near the end of the sixteenth century by Jan Lipperhey, Jakob Metius, and Zacharias Janssen. Based on reports of those first instruments, Galileo Galilei built his first telescope, which was later named after him, and made his famous astronomical observations of sunspots, the phases of Venus, Jupiter's moons, the rings of Saturn, and his discovery of the nature of the Milky Way as an assembly of very many stars. The so-called astronomical telescope was invented by Johannes Kepler, who in 1611 also published in *Dioptrice* the theory of this telescope. The disturbing effects of spherical and chromatic aberrations were soon realized by astronomers, opticians, and other users, but it was not before Moor Hall (1792) and John Dollond (1758) that at least the latter flaw was corrected for by introducing the achromat. Later, the achromat was significantly improved by Peter Dollond, Jesse Ramsden, and Josef Fraunhofer. The catadioptric telescope was probably introduced by Lenhard Digges (1571) or Nicolas Zucchius (1608). In 1671, Sir Isaac Newton was the first to use a reflector for astronomical observations. Wilhelm Herschel improved this technique and began in 1766 to build much larger telescopes with mirror diameters up to 1.22 m. When, in the nineteenth century, the conventional metal mirror was replaced by glass, it was Léon Foucauld (1819–1868) who applied a silver coating in order to improve the reflectivity of its surface.

Modern refractors can be built with superb apochromatic corrections, and astronomical reflectors have advanced to mirror diameters of up to 8.4 m (monolithic mirror), and 10 m (segmented mirror).

While the classical definition of a telescope involves an observer, i.e., the eye, in modern times often electronic detectors have replaced human vision. In the following, we will, therefore, use a more general definition of a telescope as an optical instrument, whose purpose is to image a distant object, either in a real focal plane, or in an afocal projection for the observation by eye.

Basic Imaging Theory

In the most simple case, a telescope may be constructed from a single lens or a single mirror, creating a real focal plane. Let us, therefore, introduce some basic principles of imaging using optical elements with spherical surfaces.

Refraction at a Spherical Surface

Figure 1 shows an example of refraction at a spherical surface. A ray emerging from object O hits the surface between two media of refractive index n and n' at point P. After refraction, the ray continues at a deflected angle ε' towards the normal and intersects the optical axis at point O′. In the ideal case, all rays emerging from O at different angles σ are collected in O′, thus forming a real image of object point O in O′. In the paraxial approximation, we have $\sin\sigma = \tan\sigma = \sigma$, and $p \approx 0$. In triangle (O, O′, P): $\sigma' - \sigma = i - i'$. Using Snell's law $ni = n'i'$, we obtain:

$$\sigma' - \sigma = i'\frac{n' - n}{n} \qquad [1]$$

Since $\phi = \sigma' + i'$, and in the paraxial approximation:

$$\sigma' n' - \sigma n = \phi(n' - n) \qquad [2]$$

$$\frac{h}{s'}n' - \frac{h}{s}n = \frac{h}{r}(n' - n) \qquad [3]$$

we obtain finally:

$$\frac{n'}{s'} - \frac{n}{s} = \frac{n' - n}{r} \qquad [4]$$

For an object at infinity, where $s = \infty$, s' becomes the focal distance f':

$$f' = r\frac{n'}{n' - n} \quad \text{and} \quad \tilde{f} = -r\frac{n}{n' - n} \qquad [5]$$

Lenses

Combining two or more refractive surfaces, as in the preceding section, allows us to describe a single lens or more complex optical systems with several lenses in series. For a single lens in air, we just give the lens equation (without deriving it in detail):

$$s_2'(F') = \frac{1}{n - 1}\,\frac{r_2[nr_1 - (n - 1)d]}{(n - 1)d + n(r_2 - r_1)} \qquad [6]$$

where d is the lens thickness, measured between the vertices of its surfaces, r_1, r_2 are the radii of the surfaces, n the index of refraction of the lens material, and $s_2'(F')$ the distance of the object image from the vertex of the last surface, facing the focal plane.

The back focal distance is given by:

$$f' = \frac{1}{n - 1}\,\frac{nr_1r_2}{(n - 1)d + n(r_2 - r_1)} \qquad [7]$$

An important special case is where the lens thickness is small compared to the radii of the lens surfaces, i.e., when $d \ll |r_2 - r_1|$. In this case we can neglect the term of $(n - 1)d$ and write:

$$f' = \frac{1}{n - 1}\,\frac{r_1r_2}{(r_2 - r_1)} \qquad [8]$$

Reflection at a Spherical Surface

Let us now consider the other simple case of imaging object O into O′ by reflection from the concave spherical surface, as depicted in Figure 2. In the paraxial case, we have:

$$\sigma = y/s, \qquad i = \phi - \sigma, \qquad \sigma' = y/s', \qquad i' = \phi - \sigma', \qquad \phi = y/r \qquad [9]$$

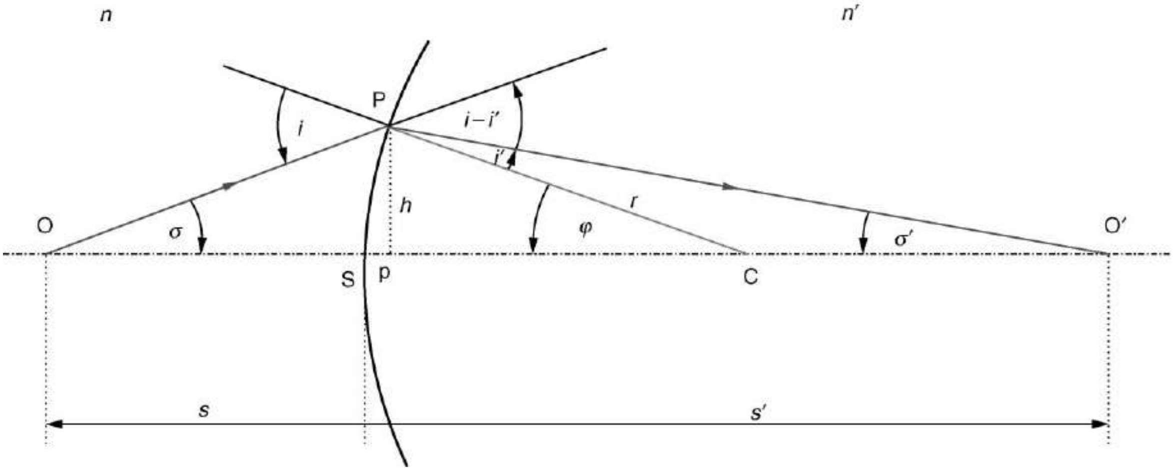

Figure 1 Refraction at a spherical surface between media with refractive indices n and n'.

With $i = -i'$, we obtain:

$$\frac{1}{s'} - \frac{1}{s} = -\frac{2}{r} \qquad [10]$$

Again, with $s = \infty$, the focal length becomes:

$$f' = -\frac{r}{2} \qquad [11]$$

In what follows, we will now assume that the distance of the object is always large (∞), which, in general, is the situation when a telescope is employed. The image is formed at a surface defined by O′, ideally a focal plane (paraxial case). We will consider deviations from the ideal case below.

Simple Telescopes

As we have observed above, a telescope is in principle nothing but an optical system, which images an object at infinity. In the most simple case, a lens with positive power or a concave mirror will satisfy this condition. The principle of a refractor and a reflecting telescope is shown in Figures 3 and 4. In both cases, parallel light coming from the distant object is entering through the entrance pupil, experiences refraction (reflection) at the lens (mirror), respectively, and is converging to form a real image in the focal plane.

Two point sources, for example two stars in the sky, separated by an angle δ, will be imaged in the focal plane as two spots with separation Δy. When f is the focal length of the telescope, we therefore have:

$$\tan \delta = \frac{\Delta y}{f} \qquad [12]$$

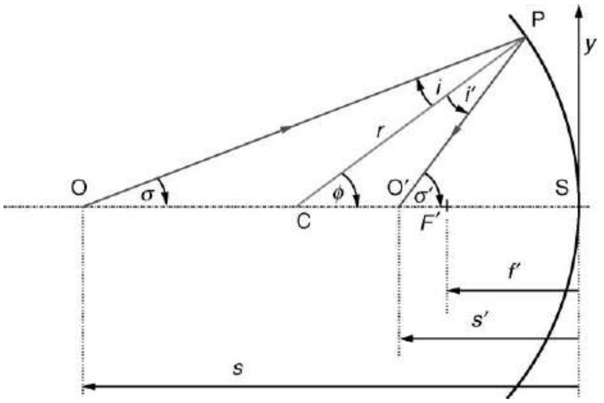

Figure 2 Reflection at a spherical surface with radius r.

The so-called plate scale m of the telescope in units of arcsec/mm is given by:

$$m = 3600 \operatorname{atan} \frac{1}{f} \quad (f \text{ in mm}) \qquad [13]$$

In the case of the refractor, we have also shown an eyepiece, illustrating that historically the instrument was invented to enhance the human vision. The eye is located at the position of the exit pupil, thus receiving parallel light, i.e., observing an object at infinity, however (i) with a flux which is increased by a factor A, given by the ratio of the area of the telescope aperture with diameter D to the area of the eye's pupil d with $A = (D/d)^2$, and (ii) with a magnification M of the apparent angle between two separate objects at infinity, which is given by the ratio of the focal lengths of the objective and the eyepiece:

$$\Gamma = \frac{f_{\mathrm{obj}}}{f_{\mathrm{ocl}}} \qquad [14]$$

The magnification can also be expressed as the ratio of the diameters of the exit pupil and the entrance pupil:

$$\Gamma = \frac{D_{\mathrm{expup}}}{D_{\mathrm{enpup}}} \qquad [15]$$

The first astronomical telescopes with mirrors were equipped with an eyepiece for visual observation. The example in Figure 4 shows a configuration which is more common nowadays for professional astronomical observations, using a direct imaging detector in the focal plane (photographic plate, electronic camera).

Aberrations

Reflector Surfaces

We shall now study the behavior of a real telescope beyond the ideal in the case of a reflector. Figure 5 shows the situation of a mirror, imaging an on-axis

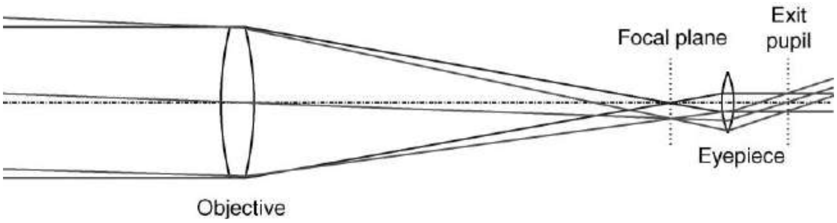

Figure 3 Refractor, astronomical telescope for visual observations with eyepiece (Kepler telescope).

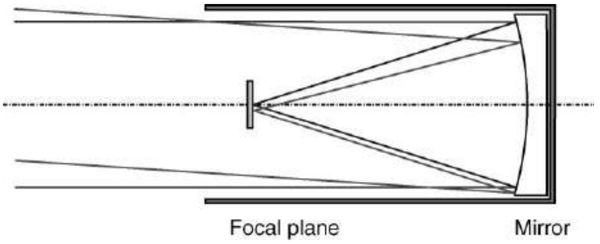

Figure 4 Reflector, astronomical telescope with prime focus imager.

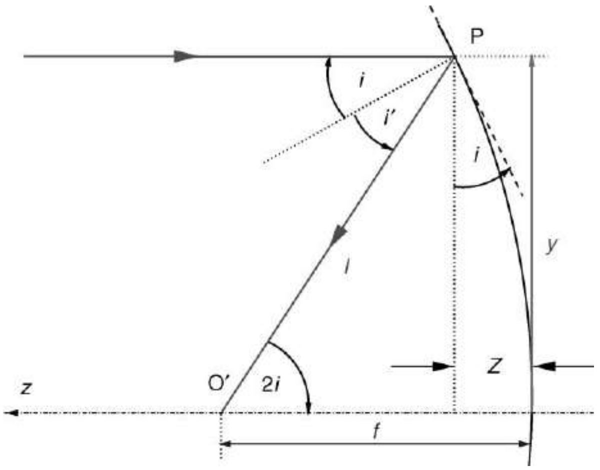

Figure 5 Imaging an object at infinity with reflector.

object at infinity into focus at point O′. With Fermat's principle, the optical paths of all rays forming the image must have identical length l. For all j, we must satisfy:

$$Z_j + f = l \qquad [16]$$

From the geometry in Figure 5, we find also:

$$y_i^2 + (f - Z_j)^2 = l^2 \qquad [17]$$

Combining eqns [16] and [17], we obtain the equation for a surface which satisfies the condition for perfect on-axis imaging:

$$y^2 = -4fz \qquad [18]$$

Obviously only a paraboloid is capable of fulfilling the condition for all rays parallel to the optical axis at whatever separation y. Also, we see that spherical mirrors are less than ideal imagers: rays at increasing distance y from the axis will suffer from increasing path length differences. The effect is called spherical aberration. We will investigate aberrations more quantitatively further below.

Elementary axisymmetrical surfaces can be expressed as follows, with parameter C being the conic constant:

$$y^2 - 2Rz + (1 + C)z^2 = 0 \qquad [19]$$

and the following characteristics:

$C > 0$	ellipsoidal, prolate
$C = 0$	spherical
$-1 < C < 0$	ellipsoidal, oblate
$C = -1$	paraboloidal
$C < -1$	hyperboloidal

Let us now leave the paraxial approximation and determine the focal length of rays which are parallel to, but further away at a distance y from the optical axis. Figure 5 shows the relevant geometrical relations. The intersection of the reflected ray with the optical axis at distance f from the vertex can be found from:

$$f = z + (f - z) \quad \text{where} \quad \frac{y}{f - z} = \tan(2i) \qquad [20]$$

We use the slope of the tangent in point P:

$$\frac{\mathrm{d}z}{\mathrm{d}y} = \tan i \qquad [21]$$

differentiate eqn [19] to get:

$$\frac{\mathrm{d}z}{\mathrm{d}y} = \frac{y}{R - (1 + C)z'} \qquad [22]$$

use the trigonometrical relation:

$$\tan 2\phi = \frac{2 \tan \phi}{1 - \tan^2 \phi} \qquad [23]$$

and substitute for tan(2i) in eqn [20] to obtain:

$$f = \frac{R}{2} + \frac{(1 - C)z}{2} - \frac{y^2}{2(R - (1 + C)z)} \qquad [24]$$

In order to express eqn [24] solely in terms of y, we rewrite eqn [19]:

$$z = \frac{R}{1 + C}\left(1 - \sqrt{1 - \frac{y^2}{R^2}(1 + C)}\right) \qquad [25]$$

Expanding z in a series and neglecting any orders higher than 6:

$$z = \frac{y^2}{2R} + (1 + C)\frac{y^4}{8R^3} + (1 - C)^3\frac{y^6}{16R^5} \qquad [26]$$

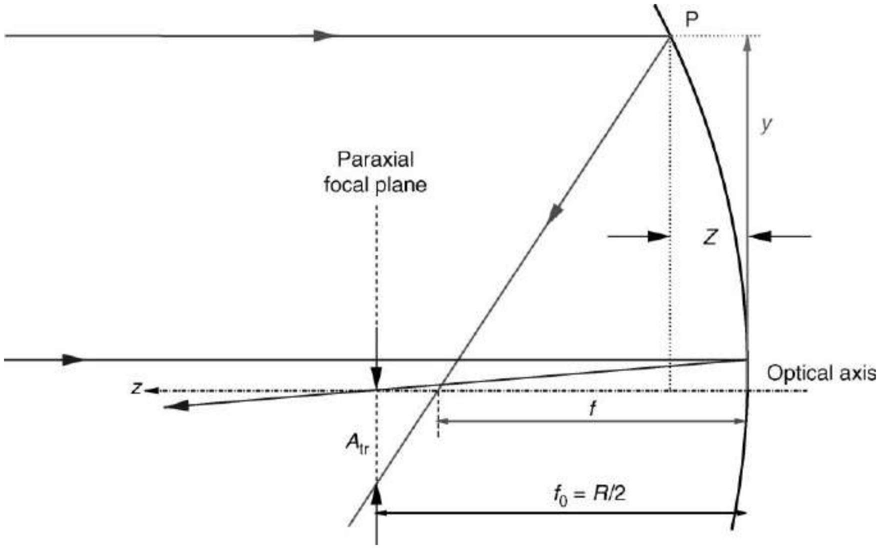

Figure 6 Transverse spherical aberration.

we obtain finally:

$$f = \frac{R}{2} - \frac{1+C}{4R}y^2 - \frac{(1+C)(3+C)}{16R^3}y^4 \qquad [27]$$

.

Transverse Spherical Aberration

Obviously, for any $C \neq -1$, rays of different distance y to the optical axis are focused at different focal lengths f (spherical aberration). The net effect for a distant point source is that the image is blurred, rather than forming a sharp spot. The lateral deviation of nonparaxial rays intersecting the nominal focal plane at $f_0 = R/2$ is described by the transverse spherical aberration A_{tr} (Figure 6):

$$A_{\mathrm{tr,sph}} = \frac{y}{f-z}(f_0 - f) \qquad [28]$$

Using eqns [19] and [27], one obtains:

$$A_{\mathrm{tr,sph}} = -\frac{1+C}{2R^2}y^3 + \frac{3(1+C)(3+C)}{8R^4}y^5 + \cdots \qquad [29]$$

Angular Spherical Aberration

It is sometimes more convenient to consider the angular spherical aberration, since a telescope is naturally measuring angles between objects at infinity. To this end, we compare an arbitrary reflector surface according to eqn [19] with a reference paraboloid, which is known to be free of on-axis spherical aberration (Figure 7). For any given ray, we observe how the reflected ray under study is tilted towards the reference ray, which would be reflected from the ideal surface of the paraboloid. While the former is measured under an exit angle of $2i$ towards the optical axis (see Figure 5), we designate the latter $2i_{\mathrm{p}}$. The angular spherical aberration is defined as the difference of these two angles:

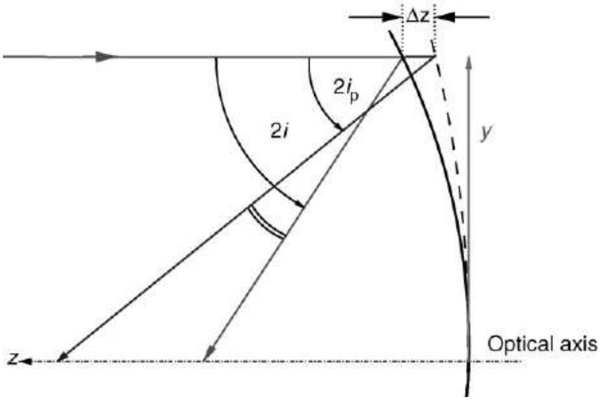

Figure 7 Angular aberration determined with reference to a paraboloid.

$$A_{\mathrm{ang,sph}} = 2i - 2i_{\mathrm{p}} = 2(i - i_{\mathrm{p}}) \qquad [30]$$

Again, we can use the slope of the tangent to determine the change of angles $i \rightarrow i_{\mathrm{p}}$:

$$A_{\mathrm{ang,sph}} = 2\left(\frac{\mathrm{d}z}{\mathrm{d}y} - \frac{\mathrm{d}z_{\mathrm{p}}}{\mathrm{d}y}\right) = 2\frac{\mathrm{d}}{\mathrm{d}y}\Delta z \qquad [31]$$

It is, therefore, sufficient to consider the difference in z between the surface under study and the reference paraboloid. From eqn [26], we obtain as an approximation to third order:

$$A_{\mathrm{ang,sph}} = -(1 + C)\frac{y^3}{R^3} \qquad [32]$$

Aberrations in the Field

Since the use of a telescope exclusively on-axis would be somewhat limited, let us finally investigate the behavior under oblique illumination, i.e., when an object at infinity is observed under an angle Θ towards the optical axis of the system. This will allow us to assess aberrations in the field. For simplicity, we are considering a paraboloid, remembering that on-axis the system is free of spherical aberration. Again, we are employing an imaginary reference paraboloid which is pointing on-axis towards the object at angle Θ (Figure 8). Its coordinate system (O', z', y') is displaced and tilted with regard to the system under study (O, z, y). One can show that the offset Δz, which is used to derive the angular aberration in the same way as above, is given by:

$$2\Delta z = a_1 \frac{y^3\Theta}{R^2} + a_2 \frac{y^2\Theta^2}{R} + a_3 y\Theta^3 \qquad [33]$$

The angular aberration in the field is then:

$$A_{\mathrm{ang,f}} = 3a_1 \frac{y^2\Theta}{R^2} + 2a_2 \frac{y\Theta^2}{R} + a_3\Theta^3 \qquad [34]$$

The coefficients are called:

a_1: coma
a_2: astigmatism
a_3: distortion

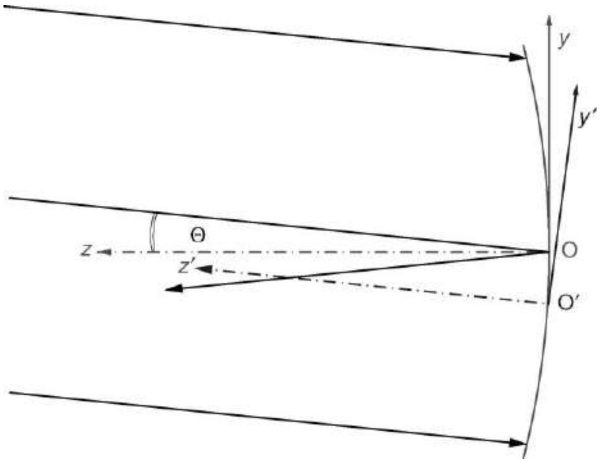

Figure 8 Aberration in the field, oblique rays.

Chromatic Aberration

So far we have only considered optical aberrations of reflecting telescopes. Spherical aberration in the case of a lens is treated in analogous ways and will not be discussed again. However, the wavelength dependence of the index of refraction in optical media (dispersion) is an important factor and gives rise to chromatic aberrations which are only encountered in refractive optical systems. We can distinguish two principal effects in the paraxial approximation: (i) longitudinal chromatic aberration, resulting in different axial image positions as a function of wavelength; and (ii) lateral chromatic aberration, which can be understood as an aberration of the principal ray as a function of wavelength, leading to a varying magnification as a function of wavelength (plate scale in the case of a telescope).

As an example, let us determine the wavelength dependence of the focal length of a single thin lens. Opticians are conventionally using discrete wavelengths of spectral line lamps of different chemical elements, e.g., the Mercury e line at 546.1 nm, Cadmium F' 480.0 nm, or Cadmium C' 643.8 nm, covering more or less the visual wavelength range of the human eye. Differentiation of eqn [8] yields for the wavelength of e:

$$\frac{f'}{\mathrm{d}n} = -\frac{1}{(n_e - 1)^2}\frac{r_1 r_2}{r_2 - r_1} \qquad [35]$$

Substituting

$$\frac{r_1 r_2}{r_2 - r_1} = f'_e(n_e - 1) \qquad [36]$$

we get:

$$\mathrm{d}f' = -\frac{\mathrm{d}n}{n_e - 1} f'_e \qquad [37]$$

Replacing the differential by the differences $\Delta f' = f'_{F'} - f'_{C'}$ and $\Delta n = n_{F'} - n_{C'}$, then:

$$f'_{F'} - f'_{C'} = -\frac{n_{F'} - n_{C'}}{n_e - 1} f'_e = -\frac{f'_e}{\nu_e} \qquad [38]$$

ν_e is called Abbe's Number:

$$\nu_e = \frac{n_e - 1}{n_{F'} - n_{C'}} \qquad [39]$$

It is a tabulated characteristic value of dispersion for any kind of glass and allows one to quickly estimate the difference in focal length at the extreme wavelengths F' and C' when the nominal focal length of a lens is known as e.

Angular Resolution

In the previous sections, we have exclusively employed geometrical optics for the description of telescope properties. With geometrical optics and in the absence of aberrations, there is no limit to the angular resolution of a telescope, i.e., there is no limiting separation angle $\delta_{\rm lim}$, below which two distant point sources can no longer be resolved. However, real optical systems experience diffraction, leading to the finite width of the image of a point source (Airy disk), whose normalized intensity distribution in the focal plane as a function of radius r' from the centroid for the case of a circular aperture of diameter D can be conveniently described by a Bessell function (λ: wavelength, f': focal length of the objective):

$$\frac{I(r)}{I_0} = \left[\frac{2J_1(\nu)}{\nu}\right]^2 \quad \text{where} \quad \nu = \frac{\pi D r'}{\lambda f'} \qquad [40]$$

The intensity distribution is plotted in Figure 9. The first four minima of the diffraction ring pattern are listed in Table 1. Two point sources are defined to be at the resolution limit of the telescope when the maximum of one diffraction pattern happens to coincide with just the first diffraction minimum of the other one (Figure 9, right). As a rule of thumb, the angular resolution is of a telescope with diameter D is given by:

$$\delta_{\rm min} = 1.22\frac{\lambda}{D} \rightarrow \frac{138}{D} \qquad [41]$$

the latter in units of arcsec for a wavelength of 550 nm and D in mm.

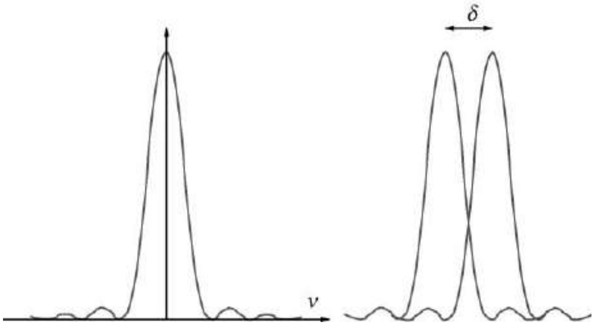

Figure 9 Left: radial intensity distribution of diffraction pattern (Airy disk). Right: two overlapping point source images with same intensity at limiting angular separation δ.

Table 1 Diffraction minima of airy disk

n	ν_n	$r'D/2f'\lambda$
1	3.83	0.61
2	7.02	1.12
3	10.2	1.62
4	13.3	2.12

Types of Telescopes

There are very many different types of telescopes which were developed for different purposes. In order to obtain an impression, we will only briefly introduce a nonexhaustive selection of refractors and reflectors, which by no means is representative of the whole. For a more complete overview, the reader is referred to the literature.

Refractor Objectives

Refractors are first of all characterized by their objectives. Figure 10 shows a sequence of lenses, indicating the progress of lens design developments from left to right. All lenses are groups 2 or 3, designed to compensate for chromatic aberration through optimized combinations of crown and flint glasses. The first lens is a classical C achromat, e.g., using BK7 and SF2 glass, which is essentially the design which was introduced by Hall and Dollond. It is still a useful objective for small telescopes, where a residual chromatic aberration is tolerable. The second lens is the Fraunhofer E type, which is characterized by an air gap between the two lenses (BK7, F2), which helps to optimize the image quality of the system, reducing spherical aberration and coma. As a disadvantage, the two additional glass–air interfaces lead to ghost images (reflected light) and Fresnell losses. The objective is critical with regard to the alignment of tip and tilt. The third example is an improved AS type achromat (KzF2, BK7), which is similar to the E type in

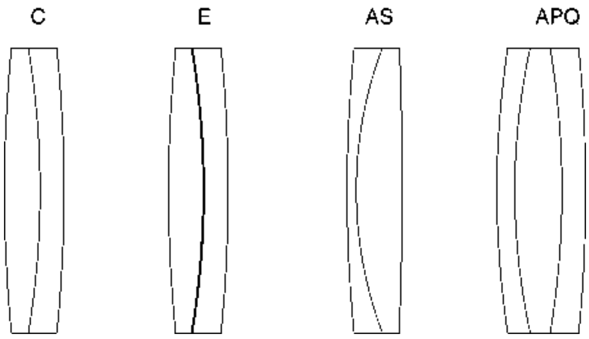

Figure 10 Different types of refractor objectives, from left to right: classical C achromat, Fraunhofer E achromat with air gap, AS achromat, APQ apochromat.

that it also possesses an air gap, however with an improved chromatic aberration correction in comparison to the Fraunhofer objective. The lenses show a more pronounced curvature and are more difficult to manufacture. The beneficial use of KzF2 glass is an example of the successful research for the development of new optical media, which was pursued near the end of the nineteenth century in a collaboration between Otto Schott and Ernst Abbé. The last example is a triplet, employing ZK2, CaF_2, and ZK2. This APQ type of objective is an apochromat, whose chromatic aberration vanishes at 3 wavelengths. It is peculiar for its use of a crystal (calcium fluoride, CaF_2) as the optical medium of the central lens. The manufacture of the CaF_2 lens is critical and expensive, but it provides for an excellent correction of chromatic and spherical aberration of this high quality objective. The triplet is also difficult to manufacture because of the coefficient of thermal expansion, which is roughly different by a factor of two between CaF_2 and the glass. In order to avoid disrupture of the cemented lens group, some manufacturers have used oil immersion to connect the triplet. Since CaF_2 not only has excellent dispersion properties to correct for chromatic aberration, but also exhibits an excellent transmission beyond the visual wavelength regime, this type of lens is particularly useful for applications in the blue and UV.

Finally, Figure 11 shows the complete layout of an astrograph, which is a telescope for use with photographic plates. The introduction of this type of instrument meant an important enhancement of astrophysical research by making available a significantly more efficient method, compared to the traditional visual observation of individual stars, one by one. Using photographic plates, it became possible to measure position and brightness of hundreds and thousands of stars, when previously one could investigate tens. Typical astrographs were built with apertures between 60 and 400 mm, and a field-of-view as large as 15 degrees. Despite the growing importance of space-based observatories for astrometry, some astrographs are still in use today.

Basic Oculars

While the general characteristics of a refractor are dominated by the objective, the eyepiece nevertheless plays an important role for the final appearance to the observer, with an effect on magnification, image quality, chromatism, field of view, and other properties of the entire system. From a large variety of different ocular types, let us briefly consider a few simple examples.

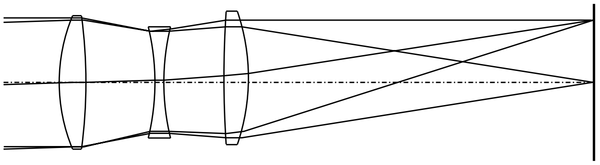

Figure 11 Astrograph.

The first eyepiece in Figure 12 is the one of the Kepler telescope, which was already introduced in Figure 3. It has a quite simple layout with a single lens of positive power. The lens is located after the telescope focal plane at a distance which is equal to the ocular focal length f_oc. The telescope and the ocular focal planes are coinciding, and as a result, the outcoming beam is collimated, forming the system exit pupil after a distance f_oc. This is where ideally the observer's eye is located. The focal plane is creating a real image, which is a convenient location for a reticle. The Kepler telescope is, therefore, a good tool for measurement and alignment purposes. We note that the image is inverted, which is relatively unimportant for a measuring telescope and astronomy, but very annoying for terrestrial use.

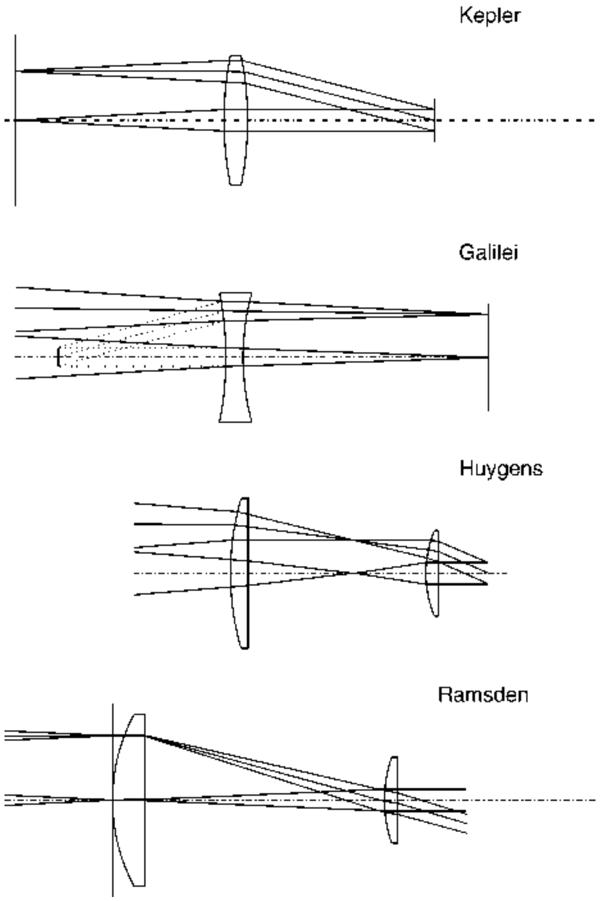

Figure 12 Elementary eyepieces after Kepler, Galilei, Huygens, and Ramdsen.

This latter property has been avoided in the next example, which is the ocular of the Galilean telescope, which was built by Galileo and used for his famous observations of planets and moons in the solar system. It uses a single lens, but contrary to the Kepler type with negative power. The lens is located in the telescope beam in front of the focal plane of the objective, which means that the exit pupil is virtual and not accessible to the observer (indicated by dotted lines in Figure 12). The overall length of the telescope is somewhat shorter than Kepler's, however at the expense of the lack of a real focal plane, so no reticle can be applied.

The third example is the 2-lens-ocular of Christian Huygens, who was able to prove with his design that it is possible to correct paraxially the lateral chromatic aberration. The eyepiece is still being built today.

Another elementary ocular is shown in Figure 12, which is the eyepiece after Ramsden. This system is also corrected for lateral chromatic aberration. The real image is formed on the front surface of the second lens, providing a location for a reticle or a scale for measuring purposes. The exit pupil is interior to the ocular. In addition to these selected basic types, there are many more advanced multilens oculars in use today, with excellent correction and large field of view (up to 60°).

Reflectors

Let us now describe several basic types of reflectors (see Figure 13). The first example is the Newton reflector, which uses a flat folding mirror to create an accessible exterior focus. It is conceptually simple and still in use today for amateur astronomy. When the main mirror is a paraboloid, the spherical aberration is corrected on the axis of the system.

The second example represents an important improvement over the simple 1-mirror design of Newton's telescope, in that the combination of a parabolic first mirror (main mirror, M1) with a hyperbolic secondary mirror (M2) eliminates spherical aberration, but not astigmatism and coma. The Cassegrain telescope has a curved focal surface which is located behind the primary mirror. Due to the onset of coma and astigmatism, the telescope has a modest useful field-of-view. Besides the correction properties, the system has significant advantages in terms of mechanical size and weight, playing an important role for the construction of large telescopes. The basic type of this design has practically dominated the construction of the classical large astronomical telescopes with M1 diameters of up to 5 m, and, more recently, the modern 8–10 m class telescopes. On the practical side, the external focus of the Cassegrain on the bottom of the mirror support cell is a convenient location for mounting focal plane instruments (e.g., CCD cameras, spectrographs, and others).

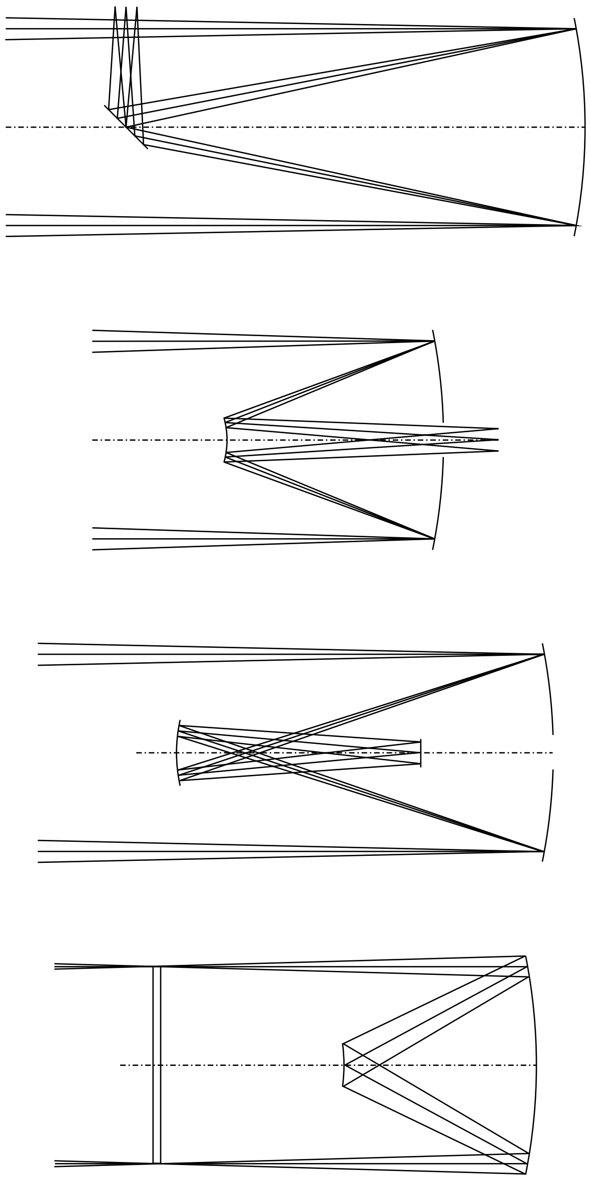

Figure 13 Reflectors, from top to bottom: Newton, Cassegrain, Gregory, Schmidt.

The more common variant of the Cassegrain type, which was actually used for the design of these telescopes, is the layout of Ritchey (1864–1945) and Cretien (1879–1956), which combines two hyperbolic mirrors for M1 and M2 (RCC), but otherwise has the same appearance as the Cassegrain. Besides spherical aberration, it also corrects for coma, giving a useful field-of-view of up to 1°. Like in the case of the Cassegrain telescope, the RCC exhibits some curvature of the focal plane.

Our third example is the Gregory reflector, which also is a 2-mirror system, however with a concave elliptical secondary mirror. The correction of this telescope is similar, but somewhat inferior to the one of the Cassegrain. The use of a concave M2 is advantageous for manufacture and testing. Nevertheless, the excessive length of the system has precluded the use of this layout for large astronomical telescopes, except for stationary beams, for example, in solar observatories.

The last example is the design put forward by Bernhard Schmidt (1879–1935) who wanted to create a telescope with an extremely wide field-of-view. The first telescope of his new type achieved a field-of-view of 16° – something which had not been feasible with previous techniques. The basic idea is to use a fast spherical mirror, and to correct for spherical aberration by means of an aspherical corrector plate in front of the mirror, which also forms the entrance pupil of the telescope. The focal surface is strongly curved, making it necessary to employ a detector with the same radius of curvature (bent photographic plates), or to compensate the curvature by means of a field lens immediately in front of the focus. The presence of the corrector plate and, if used, the field flattener, gives rise to significant chromatism. Nevertheless, the Schmidt telescope was a very successful design and dominated wide-angle survey work in astronomy for many decades.

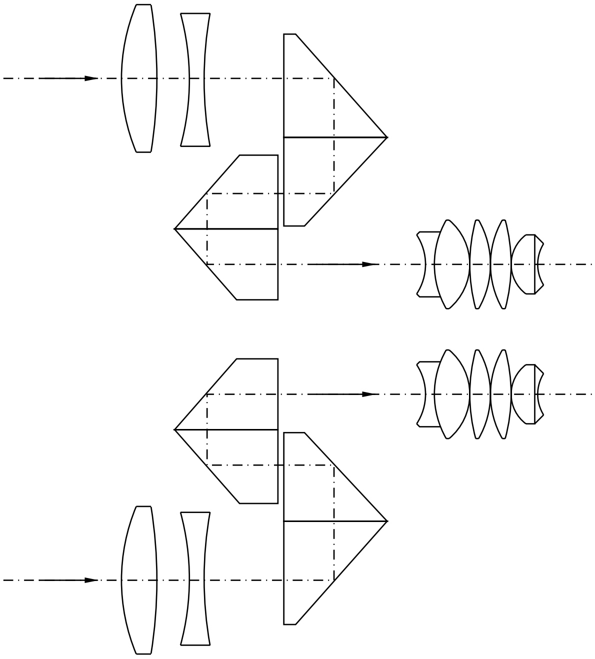

Figure 14 Principle of prism binoculars with folded beams.

Applications

General Purpose Telescopes

According to the original meaning of the Greek 'tele–skopein', the common conception of the telescope is the one of an optical instrument, which provides an enlarged view of distant objects. In addition, a large telescope aperture, compared to the eye, improves night vision and observations with poorly illuminated scenes. There are numerous applications in nautical, security, military, hunting, and other professional areas. For all of these applications, an intuitive vision is important, essentially ruling out the classical astronomical telescope (Kepler telescope, see Figure 3), which produces an inverted image.

Binoculars with Porro prisms are among the most popular instruments with upright vision, presenting a compact outline due to the folded beam geometry (Figure 14), but also single tube telescopes with re-imaging oculars are sometimes used.

Metrology

As we have realized in the introduction, telescopes are instruments which perform a transformation from angle of incidence in the aperture to distances from the origin in the focal plane. If the focal length of the objective is taken sufficiently long, and a high-quality eyepiece is used, very small angles down to the arcsec regime can be visually observed. Telescopes are, therefore, ideal tools to measure angles to very high precision.

In Figure 15 we see a pair of telescopes, facing each other, where one instrument is equipped with a light source, and the other one is used to observe the first one. The first telescope is also called a collimator. It possesses a lamp, a condensor system, a focal plane mask (usually a pinhole, a crosshair, or some other useful pattern imprinted on a transparent plate), and the telescope objective. The focal plane mask is projected to infinity. It can thus be observed with the

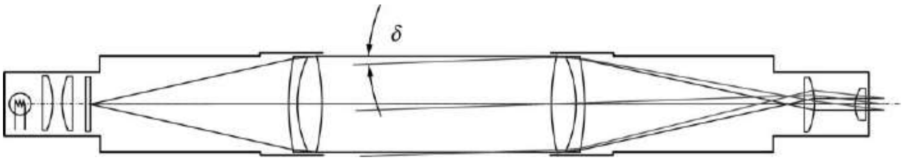

Figure 15 Collimator and telescope.

measuring telescope, such that an image of the collimator mask appears in the telescope focal plane. Also the telescope is equipped with a mask in its focal plane, usually with a precision scale or another pattern for convenient alignment with the collimator mask. If now the collimator is tilted by a small angle δ against the optical axis of the telescope, the angle of incidence in the measuring telescope changes by this same amount, giving rise to a shift of the image by $\Delta x = f \tan \delta$. Note that the measurement is completely insensitive to parallel shifts between the telescope and the collimator, since it is parallel light which is transmitted between the two instruments.

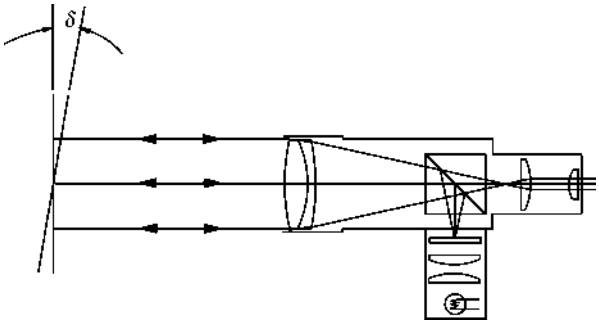

Figure 16 Autocollimator.

Another special device, which combines the former two instruments into one, is the Autocollimator. The optical principle is shown in Figure 16: in principle, the autocollimator is identical to a measuring telescope, except that a beamsplitter is inserted between the focal plane and the objective. The beamsplitter is used to inject light from the same kind of light source device as in the normal collimator. The axis of the folded light source beam and of the illuminated collimator mask are aligned to match the focal plane mask of the telescope section: the optical axis of the emerging beam is identical to the optical axis of the telescope. Under this condition, the reflected light from a test object will be exactly centered on the telescope focal plane mask, when it is perfectly aligned to normal incidence. Normally, a high-quality mirror is attached to the device under study, but sometimes also the test object itself possesses a plane surface of optical quality, which can be used for the measurement.

Figure 17 shows schematically an example for such a measurement, where a mirror is mounted on a linear stage carriage for the purpose of measuring the flexure of the stage to very high precision. With a typical length of ≈ 1 m, this task is not a trivial problem for a mechanical measurement.

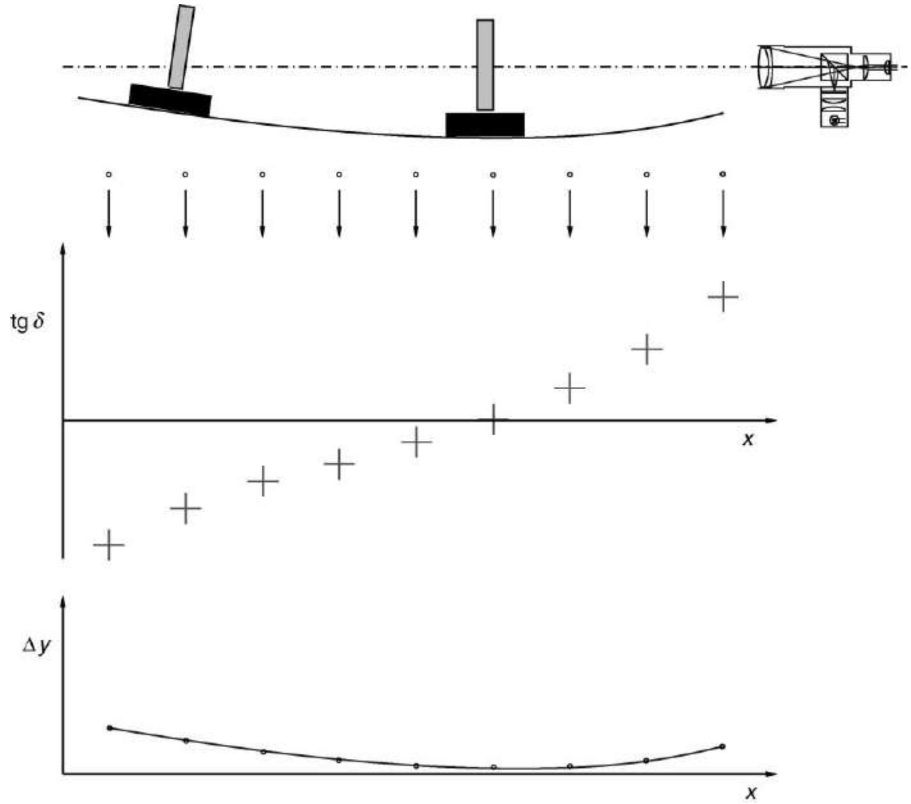

Figure 17 Using an autocollimator to measuring the flexure of a linear stage.

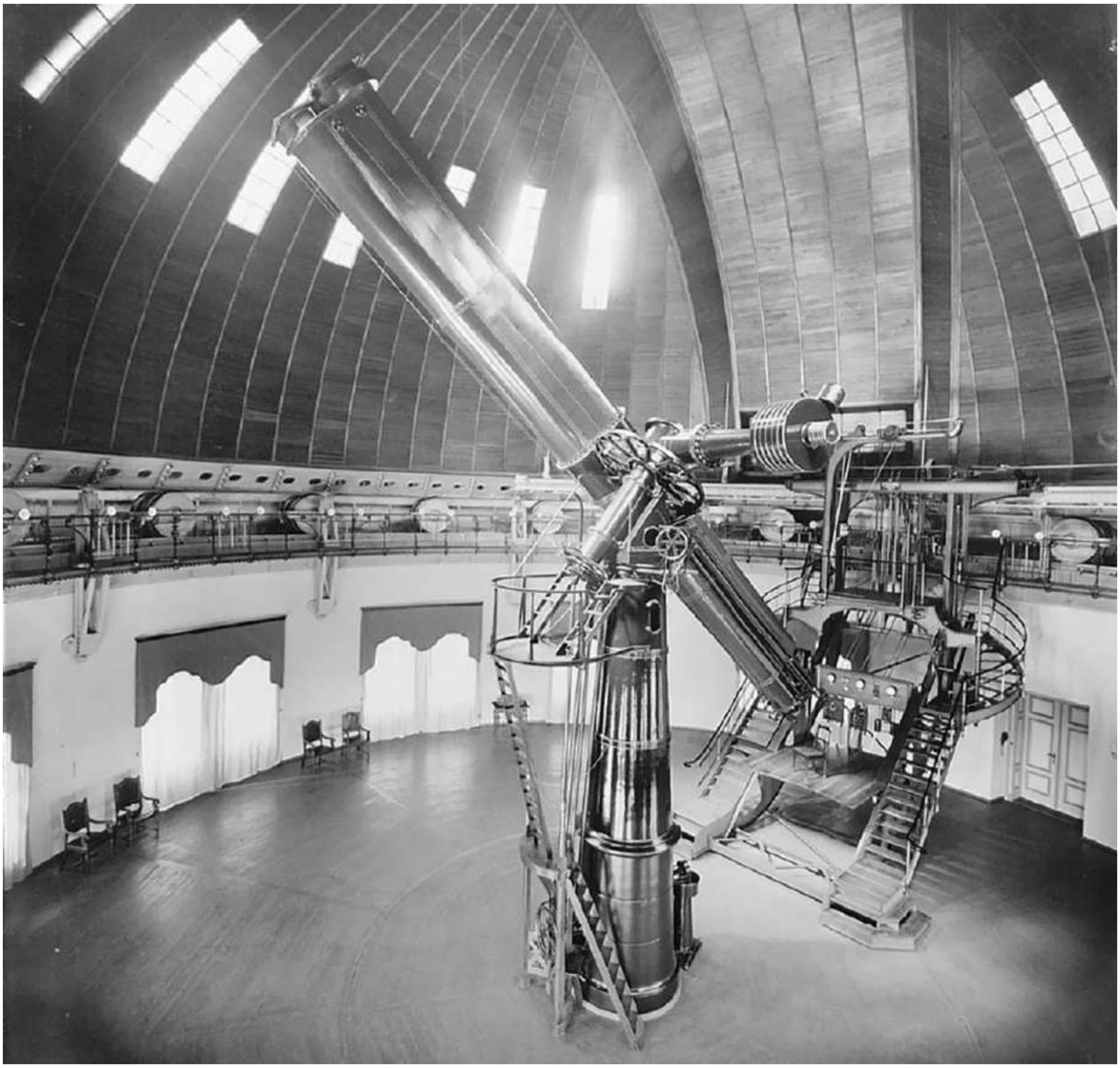

Figure 18 The 80 cm + 60 cm ‘Great Refractor’ at the Astrophysical Observatory Potsdam. The correction of the 80 cm objective of this telescope led Johannes Hartmann to the development of the ‘Hartmann Test’, which was the first quantitative method to measure aberrations. This method is still in use today, and has been further developed to become the ‘Shack-Hartmann’ technique of wavefront sensing. Courtesy of Astrophysikalisches Institut Potsdam.

By measurement with the autocollimator the tilt of the mirror consecutively for a sufficiently large number of points x_i along the rail, we can now plot the slope of the rail at each point x_i as tan d:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \tan \delta \qquad [42]$$

Integrating the series of slope values reproduces the shape of the rail $y = f(x)$.

Other applications include multiple autocollimation, wedge measurements, and right angle measurements.

Multiple autocollimation is achieved by inserting a fixed plane-parallel and semi-transparent mirror between the autocollimator and the mirror under study. Light coming from the test mirror will be partially transmitted through the additional mirror towards the autocollimator, and partially reflected back to the test mirror, and so forth. After k reflections, a deflection of $k\delta$ is measured, when δ is the tilt of the test mirror.

The wedge angle α of a plane-parallel plate with refractive index n_w is easily measured by observing the front and back surface reflections from the plate, keeping in mind that the beam bouncing back from the rear will also experience refraction in the plate. If δ is the measured angle difference between the two front and rear beams, then $\alpha = \delta/2n_\mathrm{w}$.

Two surfaces of an object can be tested for orientation at right angles by attaching a mirror to surface 1 and observe the mirror with the fixed autocollimator. In a second step, an auxiliary pentaprism is inserted into the beam, deflecting the light towards surface 2, where the mirror is attached next. If a precision pentaprism is used, any deviation from right angles between the two surfaces is seen as an offset between step 1 and 2.

These are only a few examples of the very many applications for alignment and angle measurements in optics and precision mechanics.

Without going into much detail, we are just mentioning that also the theodolite and sextant make use of the same principle of angular measurements, however without the need of a collimator or autocollimation.

Astronomical Telescopes

Despite its importance as a measurement and testing device for optical technology and manufacture as described in the preceeding paragraph, it is the original invention and subsequent improvement of the telescope for *Astronomy* which is probably the most fascinating application of any optical system known to the general public. The historical development of the astronomical telescope spans from the 17th century Galileian refractor with an aperture of just a few centimeters in diameter, over the largest refractors built for the Lick and Yerkes observatories with objectives of up to ≈1 m in diameter (see also Figures 18–20), the 4 m-class reflectors of the second half of the 20th century like e.g. the Hale Observatory 5m, the Kitt Peak and Cerro Tololo 4 m, the La Silla 3.6 m, or the La Palma 4.2 m telescopes (to name only a few), to the current state-of-the-art of 8–10 m class telescopes, prominent examples being the two Keck Observatory 10 m telescopes on Mauna Kea (Hawaii), or the ESO Very Large Observatory (Paranal, Chile), consisting of four identical 8.2 m telescopes (Figures 21–24).

This development has always been driven by the discovery of fainter and fainter celestial objects, whose quantitative study is the subject of modern astrophysics, involving many specialized disciplines such as atom/quantum/nuclear physics, chemistry, general relativity, electrodynamics, hydrodynamics, magneto-hydrodynamics, and so forth. In fact, cosmic objects are often referred to as 'laboratories'

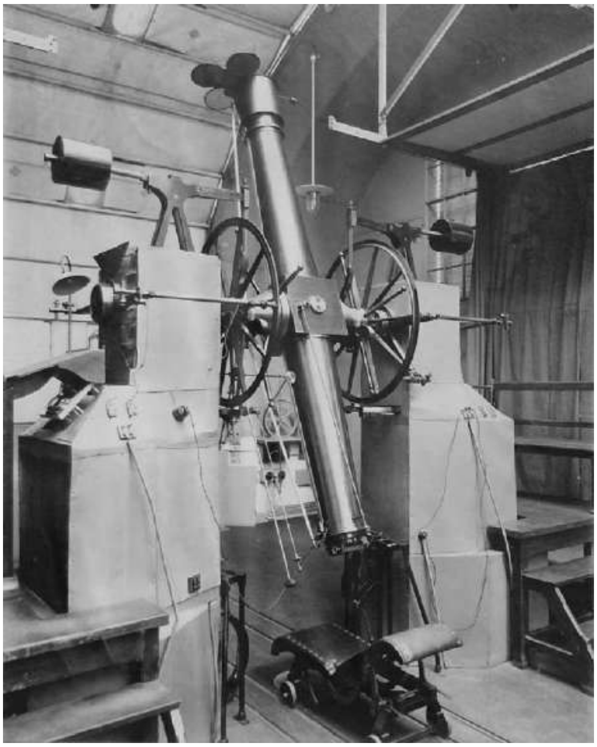

Figure 19 Pistor & Martin Meridian Circle at the former Berlin Observatory in Potsdam Babelsberg. The instrument was one of the finest art telescopes with ultra-high precision for astrometry, built in 1868. Courtesy of Astrophysikalisches Institut Potsdam.

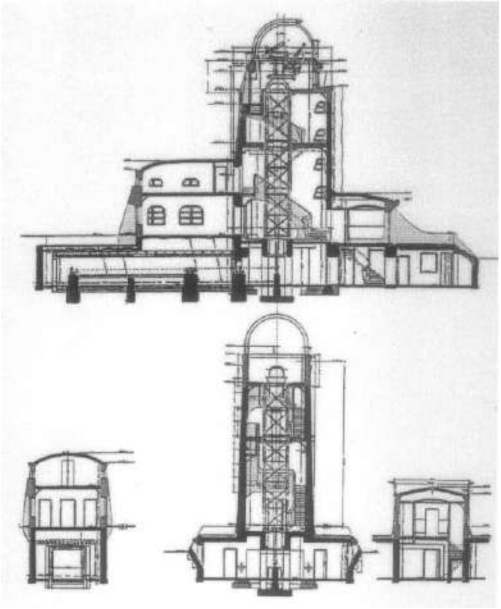

Figure 20 Architect's sketch of historical 'Einsteinturm' solar telescope, Astrophysical Observatory Potsdam. The telescope objective is mounted in the tower structure, facing upward a siderostat which is mounted inside the dome. The light is focused in the basement after deflection from a folding mirror on the slit of a bench-mounted high resolution spectrograph. The telescope represented state-of-the-art technology at the beginning of the twentieth century. Courtesy of Astrophysikalisches Institut Potsdam.

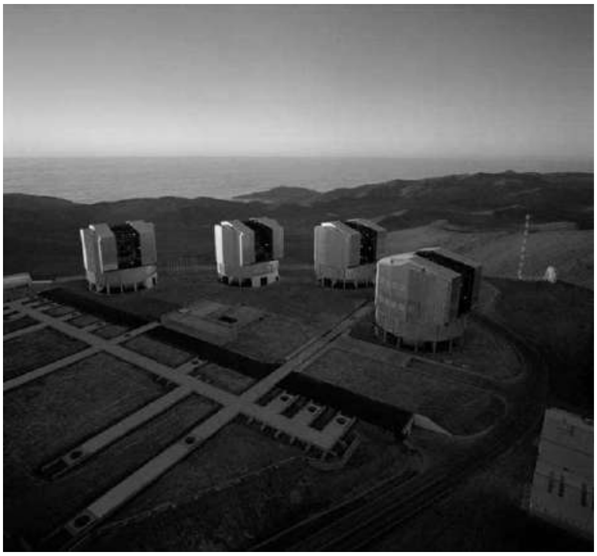

Figure 21 The Very Large Telescope Observatory (VLT) on Mt Paranal, Chile, with VLT Interferometer (VLTI). The four large enclosures are containing the active optics controlled VLT unit telescopes, which are operated either independently of each other, or in parallel to form the VLTI. The structures on the ground are part of the beam combination optics/delay lines. Courtesy of European Southern Observatory.

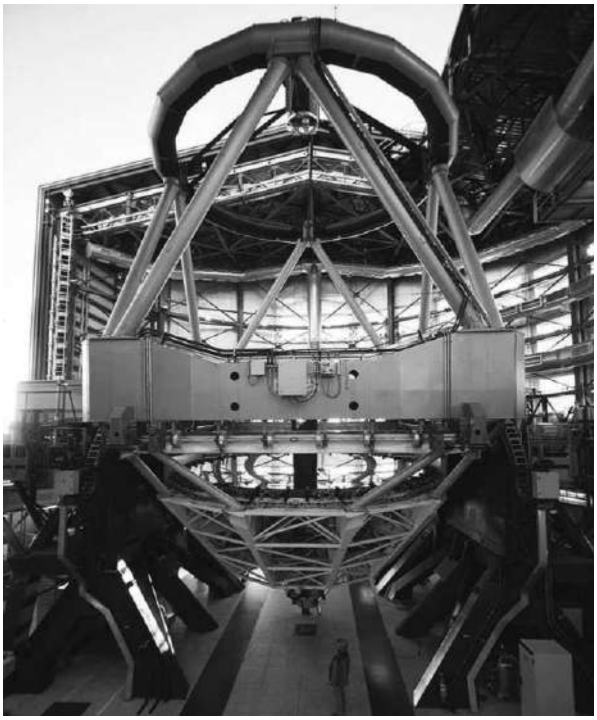

Figure 22 Total view of Kueyen, one of four unit telescopes of the VLT observatory. Courtesy of European Southern Observatory.

with extreme conditions of density, temperature, magnetic field strength, etc., which would be impossible to achieve in any laboratory on Earth. Except for particles (neutrinos, cosmic rays), gravitational waves, and the physical investigation of meteorites and other solar system material collected by space probes, telescopes (for the entire wavelength range of the electromagnetic spectrum) are the *only* tool to accomplish quantitative measurements of these cosmic laboratories.

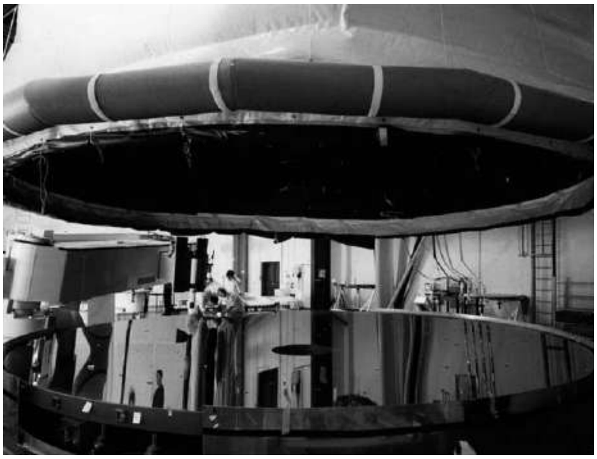

Figure 23 8.2 m thin primary mirror for the VLT. The VLT design is based on thin meniscus mirrors, made out of ZERODUR, which have low thermal inertia and which are always kept at ambient temperature, thus virtually removing the effect of thermal turbulence (mirror seeing). Seeing is the most limiting factor of ground-based optical astronomy. The optimized design of the VLT telescopes has introduced a significant leap forward to a largely improved image quality from the ground, and, consequently, improved sensitivity. The thin mirrors are constantly actively controlled in their mounting cells, using numerous piston actuators on the back surface of the mirror. Another key technology under development is ‘adaptive optics’, a method to compensate atmospheric wavefront deformations in real-time, thus obtaining diffraction limited images. Courtesy of European Southern Observatory.

In an attempt to unravel the history of the birth and evolution of the universe, to discover and to measure the first generation of stars and galaxies, and to find evidence for traces of life outside of planet Earth, plans are currently underway to develop yet another new generation of giant optical telescopes with apertures of 30–100 m in diameter (ELT: extremely large telescope). The technological challenges in terms of precision and stability are outstanding.

Along with the impressive growth of light collecting area, we must stress the importance of angular resolution, which has also experienced a significant evolution over the history of astronomical telescopes: from ≈5 arcsec of the Galilei telescope, which already provided an order of magnitude improvement over the naked eye, over roughly 1 arcsec of conventional telescopes in the 20th century, a few hundredths of an arcsec for the orbiting Hubble Space

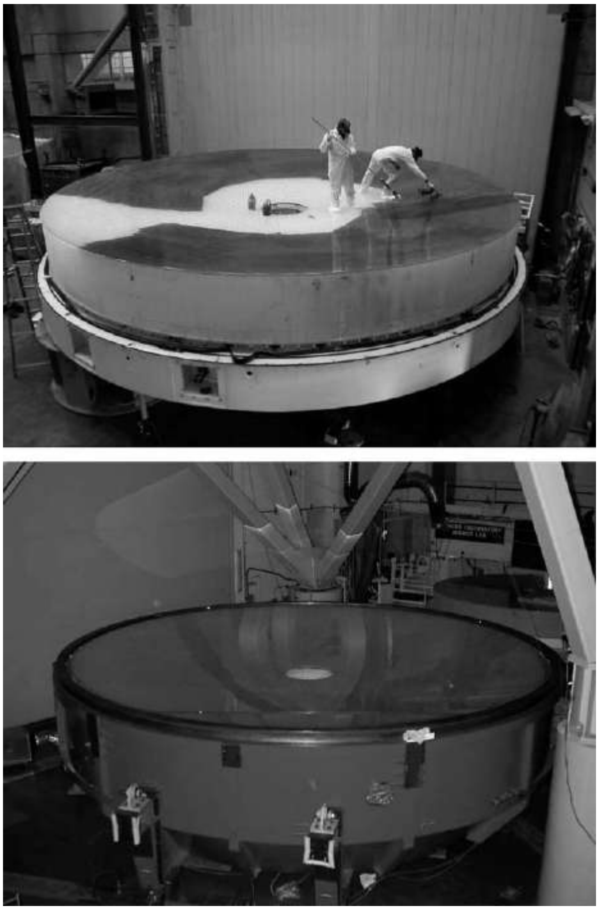

Figure 24 Top: spin-cast 8.4 m honeycomb mirror #1 for the Large Binocular Telescope, the largest monolithic mirror of optical quality in the world. Shown in the process of applying a protective plastic coating to the end-polished surface. Bottom: LBT mirror #1, finally mounted in mirror cell. Courtesy of Large Binocular Telescope Observatory.

Telescope (Figure 25), down to milli-arcsec resolution of the Keck and VLT Interferometers. The development of the aforementioned ELTs is thought to be only meaningful if operated in a diffraction-limited mode, using the emerging technique of *adaptive optics*, which is a method to correct for wavefront distortions caused by atmospheric turbulence (image blurr) in real-time.

The combination of large optical/infrared telescopes with powerful detectors and focal plane instruments (*see* **Instrumentation:** Astronomical Instrumentation) has been a prerequisite to make astrophysics one of the most fascinating disciplines of fundamental sciences.

See also

Geometrical Optics: Aberrations. **Imaging:** Adaptive Optics. **Instrumentation:** Astronomical Instrumentation.

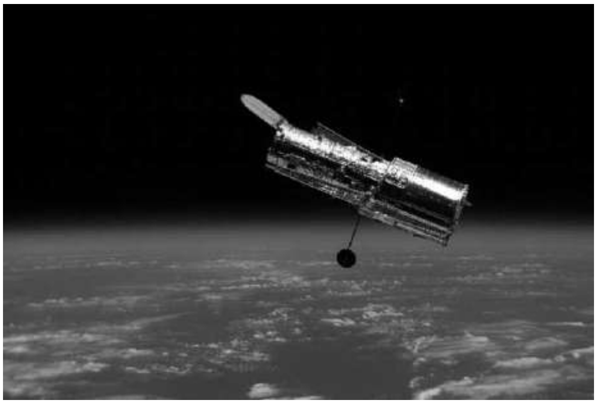

Figure 25 Hubble Space Telescope (HST) in orbit. This 2.4 m telescope is probably the most productive telescope which was ever built. Due to the absence of the atmosphere, the telescope delivers diffraction-limited, extremely sharp images, which have revealed unprecedented details of stars, star clusters, nebulae and other objects in the Milky Way and in other galaxies. Among the most spectacular results, HST has provided the deepest look into space ever, revealing light from faint galaxies that was emitted when the universe was only at 10% of its present age. Courtesy of Space Telescope Science Institute, Baltimore.

Further Reading

Bely PY (ed.) (2003) *The Design and Construction of Large Optical Telescopes*. New York: Springer.

Born M and Wolf E (1999) *Principles of Optics*, 7th ed. Cambridge: Cambridge University Press.

Haferkorn H (1994) *Optik*, 3rd ed. Leipzig: Barth Verlagsgesellschaft.

Korsch D (1991) *Reflective Optics*. San Diego: Academic Press.

Laux U (1999) *Astrooptik*, 2nd ed. Heidelberg: Verlag Sterne und Weltraum.

McCray WP (2004) *Giant Telescopes: Astronomical Ambition and the Promise of Technology*. Cambridge: Harvard University Press.

Osterbrock DE (1993) *Pauper and Prince: Richey, Hale, and Big American Telescopes*. Tucson: University of Arizona Press.

Riekher R (1990) *Fernrohre und ihre Meister*, 2nd ed. Berlin: Verlag Technik GmbH.

Rutten HGJ and van Venrooij MAM (1988) *Telescope Optics*. Richmond, VA: Willman-Bell.

Schroeder DJ (1987) *Astronomical Optics*. San Diego: Academic Press.

Welford WT (1974) *Aberrations of the Symmetrical Optical System*. London: Academic Press.

Wilson RN (1996) *Reflecting Telescope Optics I: Basic Design Theory and its Historical Development*. Berlin: Springer.

Wilson RN (1999) *Reflecting Telescope Optics II: Manufacture, Testing, Alignment, Modern Techniques*. Berlin: Springer.

INTERFEROMETRY

Contents

Overview
Gravity Wave Detection
Phase-Measurement Interferometry
White Light Interferometry

Overview

J C Wyant, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interferometers are a powerful tool used in numerous industrial, research, and development applications, including measuring the quality of a large variety of manufactured items such as optical components and systems, hard disk drives and magnetic recording heads, lasers and optics used in CDs and DVD drives, cameras, laser printers, machined parts, components for fiber optics systems, and so forth. Interferometers can also be used to measure distance, spectra, and rotations, etc. The applications are almost endless.

There are many varieties of interferometers. Some interferometers form interference fringes using two beams and some use multiple beams. Some interferometers are common path; some use lateral shear; some use radial shear; and some use phase-shifting techniques. Examples of the various types are given below, but first we must give the basic equation for two-beam interference.

Two-Beam Interference

When two quasi-monochromatic waves interfere, the irradiance of the interference pattern is given by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos[\phi] \qquad [1]$$

where the I's represent the irradiance of the individual beams and ϕ is the phase difference between the two interfering beams. The maximums of the interference occur when $\phi = 2m\pi$, where m is an integer. In most applications involving interferometers, ϕ is the quantity of interest since it is related to the quantity being measured. Figure 1 shows typical two-beam interference fringes.

There are many two-beam interferometers and we will now look at a few of them.

Fresnel Mirrors

Fresnel Mirrors are a very simple two beam interferometer. A beam of light reflecting off two mirrors, set at a slight angle to one another, will produce an interference pattern. The fringes are straight and equally spaced. The fringes become more finely spaced as the angle between the two mirrors increases, d increases, and D becomes smaller (Figure 2).

Plane Parallel Plate

When a monochromatic source illuminates a plane parallel plate, as shown in Figure 3, the interference

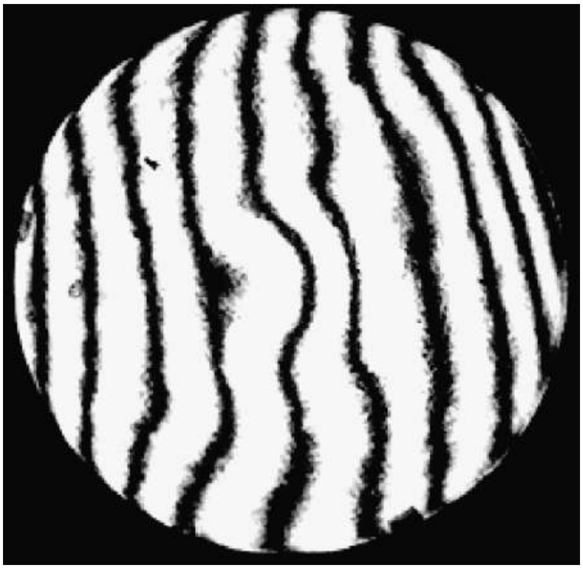

Figure 1 Typical two-beam interference fringes.

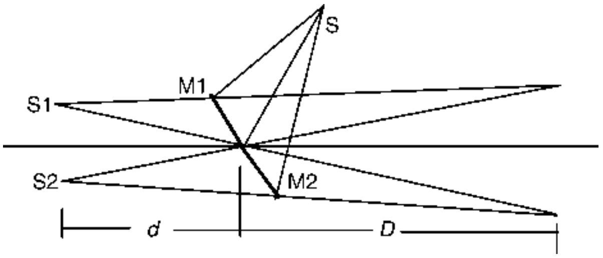

Figure 2 Fresnel mirrors.

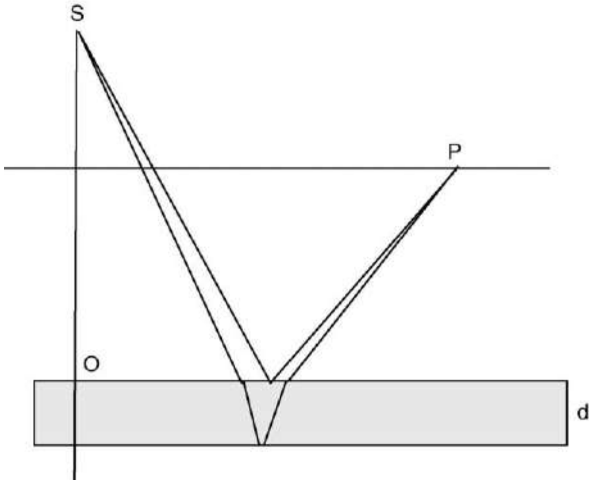

Figure 3 Plane parallel plate interferometer.

fringes are circles centered on the normal to plate. The fringes are called Haidinger fringes, or fringes of equal inclination since, for a given plate thickness, they depend on the angle of incidence. If the maximum or minimum occurs at the center, the radii of the fringes are proportional to the square root of integers. If the plate has a slight variation in thickness and it is illuminated with a collimated beam then the interference fringes are called Fizeau fringes of equal thickness and they give the thickness variations in the plate.

Michelson Interferometer

A Michelson interferometer is shown in Figure 4. Light from the extended source is split into two beams by the beamsplitter. The path difference can be viewed as the difference between the mirror M1 and the image of mirror M2, M2′. With the M1 and M2′ parallel, the fringes are circular and localized at infinity. With M1 and M2′ at a slight angle, the interference fringes are straight lines parallel to the equivalent intersection of the mirrors and localized approximately at the intersection. When the mirrors are only a few wavelengths apart, white light fringes appear, and can be used to determine their coincidence.

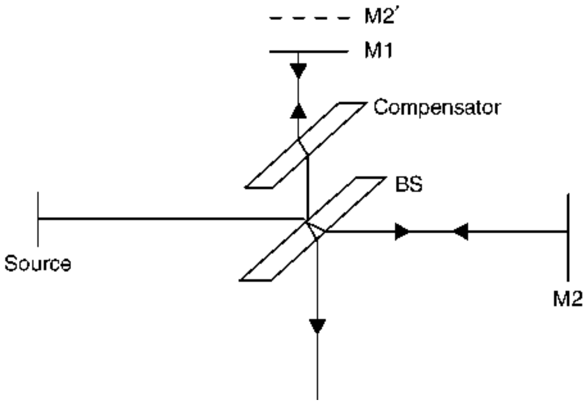

Figure 4 Michelson interferometer.

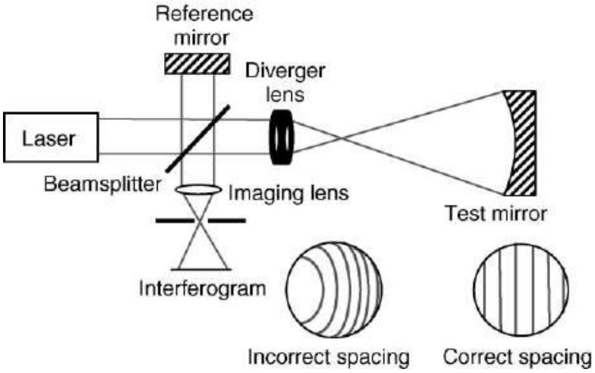

Figure 5 Twyman–Green interferometer for testing a concave spherical mirror.

Twyman–Green Interferometer

If a point source is used in a Michelson interferometer it is generally called a Twyman–Green interferometer. Twyman–Green interferometers are often used to test optical components such as flat mirrors, curved mirrors, windows, lenses, and prisms. Figure 5 shows a Twyman–Green interferometer for the testing of a concave spherical mirror. The interferometer is aligned such that the focus of the diverger lens is at the center of curvature of the spherical mirror. If the mirror under test is perfect, straight equi-spaced fringes are obtained. Figure 6 shows two interference fringes and the relationship between surface height error for the mirror being tested and fringe deviation. λ is the wavelength of the light.

Fizeau Interferometer

The Fizeau interferometer is a simple device used for testing optical surfaces, especially flats and spheres

(Figure 7). The fringes are fringes of equal thickness. It is sometimes useful to tilt the reference surface a little to get several nearly straight interference fringes. The relationship between surface height error and fringe deviation from straightness is the same as for the Twyman–Green interferometer.

Mach–Zehnder Interferometer

A Mach–Zehnder interferometer is sometimes used to look at samples in transmission. Figure 8 shows a typical Mach–Zehnder interferometer for looking at a sample in transmission.

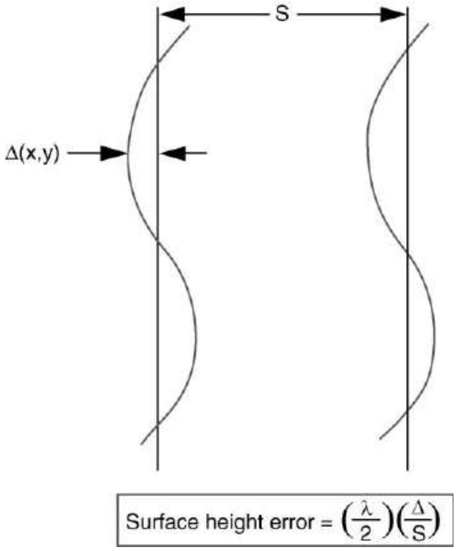
S
Δ(x,y)
Surface height error = $\left(\frac{\lambda}{2}\right)\left(\frac{\Delta}{S}\right)$

Figure 6 Relationship between surface height error and fringe deviation.

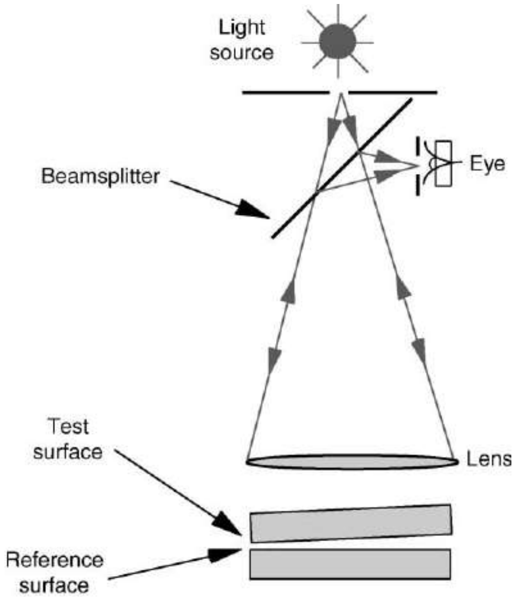
Light
source
Beamsplitter
Eye
Test
surface
Lens
Reference
surface

Figure 7 Fizeau interferometer.

Murty Lateral Shearing Interferometer

Lateral shearing interferometers compare a wavefront with a shifted version of itself. While there are many different lateral shear interferometers, one that works very well with a nearly collimated laser beam is a Murty shearing interferometer shown in Figure 9. The two laterally sheared beams are produced by reflecting a coherent laser beam off a plane parallel plate. The Murty interferometer can be used to measure the aberrations in the lens or it can be used to determine if the beam leaving the lens is collimated. If the beam incident upon the plane parallel plate is collimated a single fringe results, while if the beam is not perfectly collimated straight equi-spaced fringes result where the number of fringes gives the departure from collimation.

Radial Shear Interferometer

Radial shear interferometers compare a wavefront with a magnified or demagnified version of itself. While there are many different radial shear

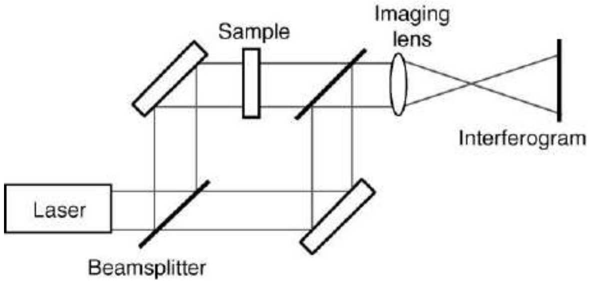
Sample
Imaging
lens
Interferogram
Laser
Beamsplitter

Figure 8 Mach–Zehnder interferometer.

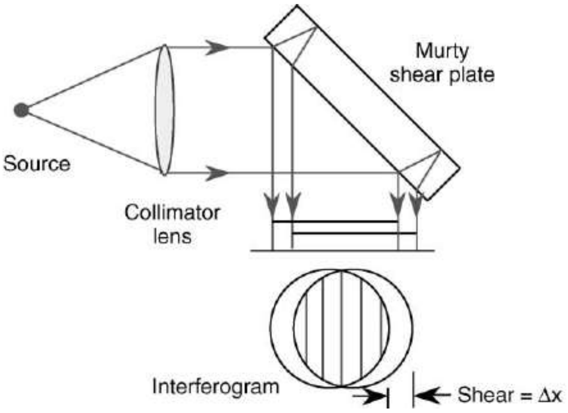
Murty
shear plate
Source
Collimator
lens
Interferogram
Shear = Δx

Figure 9 Murty lateral shear interferometer.

interferometers, one that works very well with a nearly collimated beam is shown in Figure 10. It is essentially a Mach–Zehnder interferometer where, in one arm, the beam is magnified and in the other arm the beam is demagnified. Interference fringes result in the region of overlap. The sensitivity of the interferometer depends upon the amount of radial shear. If the two interfering beams are approximately the same size there is little sensitivity, while if the two beams greatly differ in size there is large sensitivity. Sometimes radial shear interferometers are used test the quality of optical components.

Scatterplate Interferometer

A scatterplate interferometer, invented by Jim Burch in 1953, is one of the cleverest interferometers. Almost any light source can be used and no high-quality optics are required to do precision interferometry. The critical element in the interferometer is the scatterplate which looks like a piece of ground glass, but the important item is that the scattering points are arranged so the plate has inversion symmetry as shown in Figure 11. The light that is scattered the first time through the plate, and unscattered the second time through the plate, interferes with the light unscattered the first time and scattered the second time. The resulting interference fringes show errors in the mirror under test. This type of interferometer is insensitive to vibration because it is what is called a common path interferometer, in that the test beam (scattered–unscattered) and the reference beam (unscattered–scattered) travel along almost the same paths in the interferometer.

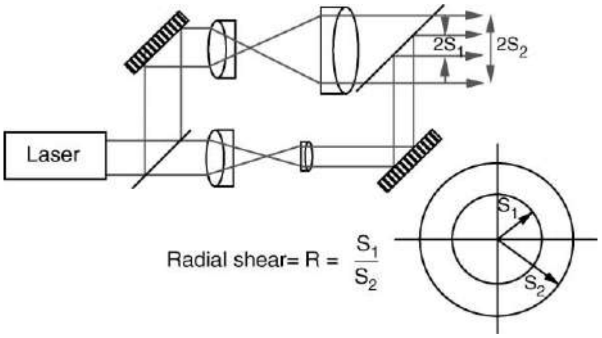

Figure 10 Radial shear interferometer.

Smartt Point Diffraction Interferometer

Another common path interferometer used for the testing of optics is the Smartt point diffraction interferometer (PDI) shown in Figure 12. In the PDI the beam of light being measured is focused onto a partially transmitting plate containing a pinhole. The pinhole removes the aberration from the light passing through it (reference beam), while the light passing through the plate (test beam) does not have the aberration removed. The interference of these two beams gives the aberration of the lens under test.

Sagnac Interferometer

The Sagnac interferometer has beams traveling in opposite directions, as shown in Figure 13. The interferometer is highly stable and easy to align. If the interferometer is rotated with an angular velocity there will be a delay between the transit times of the clockwise and the counterclockwise beams. For this reason, the Sagnac interferometer is used in laser gyros.

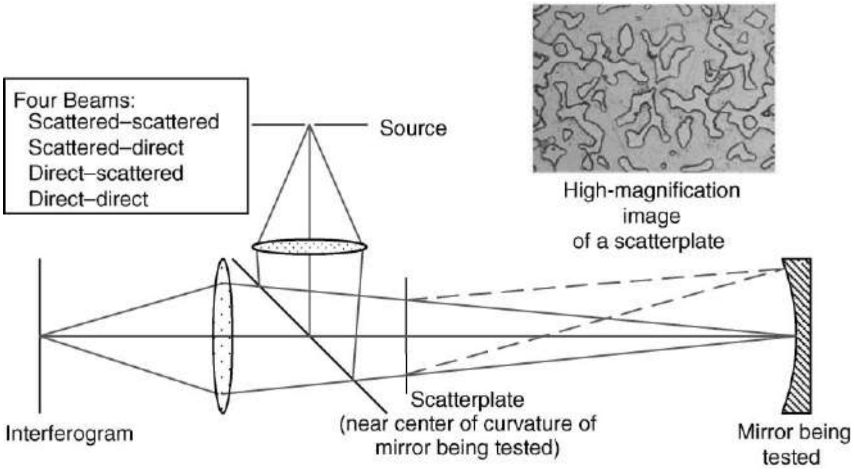

Figure 11 Scatterplate interferometer.

Fiber Interferometers

Fiber interferometers were first used for rotation sensing by replacing the ring cavity in the Sagnac interferometer with a multiloop made of a single-mode fiber. Fiber-interferometer rotation sensors are attractive as rotation sensors because they are small and low-cost.

Since the optical path in a fiber is affected by its temperature and it also changes when the fiber is stretched, fiber interferometers can be used as sensors for mechanical strains, temperature, and pressure. They can also be used for the measurement of magnetic fields by bonding the fiber to a magneto-restrictive element. Electric fields can be measured by bonding the fiber to a piezoelectric film.

Multiple Beam Interferometers

In general, multiple beam interferometers provide sharper interference fringes than a two beam interferometer. If light is reflected off a plane parallel plate there are multiple reflections. If the surfaces of the plane parallel plate have a low reflectivity, the higher-order reflections have a low intensity and the multiple reflections can be ignored and the resulting interference fringes have a sinusoidal intensity profile, but if the surfaces of the plane parallel plate have a high reflectivity, the fringes become sharper. Figure 14 shows the profile of the fringes for both the reflected and the transmitted light for different values of surface reflectivity. Multiple beam interference fringes are extremely useful for measuring the spectral content of a source. One such multiple beam interferometer, that is often used for measuring spectral distributions of a source, is a Fabry–Perot interferometer that consists of two plane-parallel plates separated by an air space, as shown in Figure 15.

Phase-Shifting Interferometry

The equation for the intensity resulting from two-beam interference, contains three unknowns, the two individual intensities and the phase difference between the two interfering beams. If three or more measurements are made of the intensity of the two-beam interference as the phase difference between the two beams is varied in a known manner, all three unknowns can be determined. This technique, called phase-shifting interferometry, is an extremely powerful tool for measuring phase distributions. Generally the phase difference is

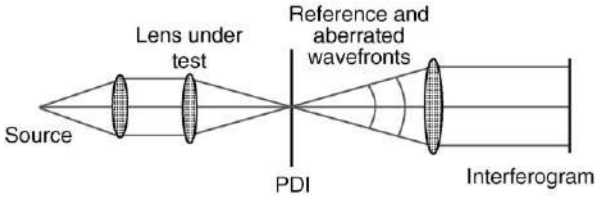

Figure 12 Smartt point diffraction interferometer.

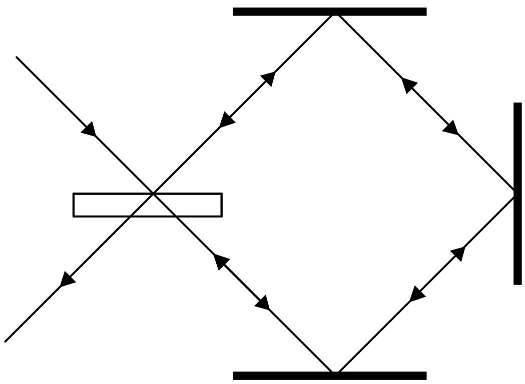

Figure 13 Sagnac interferometer.

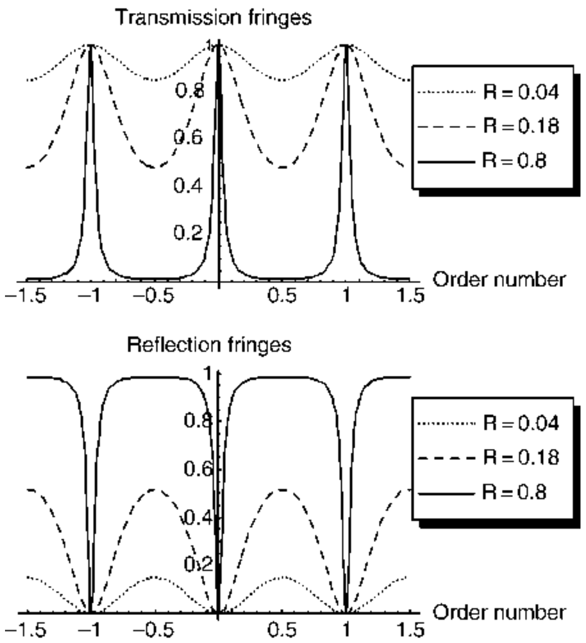

Figure 14 Multiple beam interference fringes for the transmitted and reflected light for a plane parallel plate.

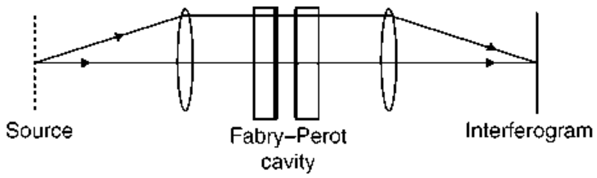

Figure 15 Fabry–Perot interferometer.

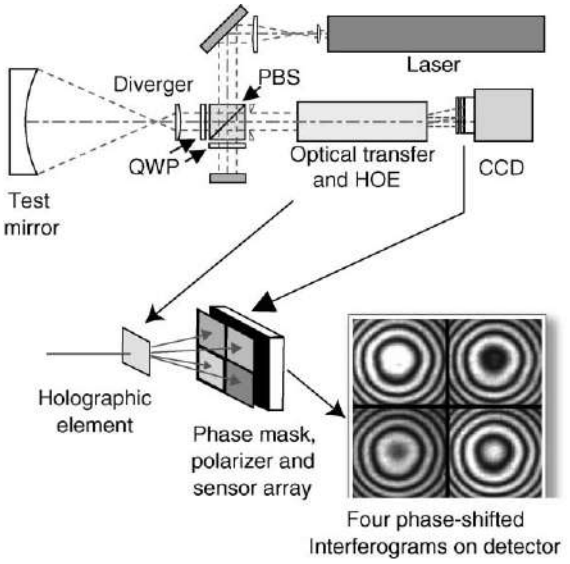
Laser
PBS
Diverger
Test mirror
QWP
Optical transfer and HOE
CCD
Holographic element
Phase mask, polarizer and sensor array
Four phase-shifted Interferograms on detector

Figure 16 Vibration-insensitive phase-shifting interferometer.

changed by 90 degrees between consecutive measurements of the interference intensity in which case the three intensities can be written as

$$\begin{aligned} I_a(x,y) &= I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y) - 90^\circ) \\ I_b(x,y) &= I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y)) \\ I_c(x,y) &= I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y) + 90^\circ) \end{aligned}$$

and the phase distribution is given by

$$\phi(x,y) = \mathrm{ArcTan}\left(\frac{I_a(x,y) - I_c(x,y)}{-I_a(x,y) + 2I_b(x,y) - I_c(x,y)}\right)$$

Phase-shifting interferometry has greatly enhanced the use of interferometry in metrology since it provides a fast low noise way of getting interferometric data into a computer.

Vibration-Insensitive Phase-Shifting Interferometer

While phase-shifting interferometry has greatly enhanced the use of interferometry in metrology, there are many applications where it cannot be used because of the environment, especially vibration. One recent phase-shifting interferometer that works well in the presence of vibration is shown in Figure 16. In this interferometer, four phase-shifted frames of data are captured simultaneously. The test and reference beams have orthogonal polarization. After the test and reference beams are combined they are passed through a holographic element to produce four identical beams. The four beams are then transmitted through wave retardation plates to cause 0, 90, 180, and 270 phase difference between the test and reference beams. After passing through a polarizer the four phase-shifted interferograms fall on the detector array. Not only can the effects of vibration be eliminated, but by making short exposures to freeze the vibration, the vibrational modes can be measured. Movies can be made showing the vibration. Likewise, flow fields can be measured.

Combining modern electronics, computers, and software with old interferometric techniques, provides for very powerful measurement capabilities.

List of Units and Nomenclature

I	irradiance
R	intensity reflectance
λ	wavelength
ϕ	phase

See also

Coherence: Speckle and Coherence. **Detection:** Fiber Sensors. **Holography, Techniques:** Computer–Generated Holograms; Digital Holography; Holographic Interferometry. **Imaging:** Adaptive Optics; Interferometric Imaging; Wavefront Sensors and Control (Imaging Through Turbulence). **Interferometry:** Phase Measurement Interferometry; White Light Interferometry; Gravity Wave Detection. **Microscopy:** Interference Microscopy. **Tomography:** Optical Coherence Tomography.

Further Reading

Candler C (1951) *Modern Interferometers*. London: Hilger and Watts.
Francon M (1966) *Optical Interferometry*. New York: Academic Press.
Hariharan P (1991) Selected papers on interferometry. *SPIE*, vol. 28.
Hariharan P (1992) *Basics of Interferometry*. San Diego: Academic Press.
Hariharan P (2003) *Optical Interferometry*. San Diego: Academic Press.
Hariharan P and Malacara D (1995) Selected papers on interference, interferometry, and interferometric metrology. *SPIE*, vol. 110.
Malacara D (1990) Selected papers on optical shop metrology. *SPIE*, vol. 18.
Malacara D (1992) *Optical Shop Testing*. New York: Wiley.
Malacara D (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Decker.
Robinson D and Reid G (1993) *Interferogram Analysis*. Brisol, UK: Institute of Physics.
Steel WH (1985) *Interferometry*. Cambridge: Cambridge University Press.
Tolansky S (1973) *An Introduction to Interferometry*. New York: Longman.

Gravity Wave Detection

N Christensen, Carleton College, Northfield, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the later part of the nineteenth century Albert Michelson performed extraordinary experiments that shook the foundations of the physics world. Michelson's precise determination of the speed of light was an accomplishment that takes great skill to reproduce today. Edward Morley teamed up with Michelson to measure the velocity of the Earth with respect to the aether. The interferometer that they constructed was exquisite, and through amazing experimental techniques the existence of the aether was disproved. The results of Michelson and Morley led to a revolution in physics, and provided evidence that helped Albert Einstein to develop the general theory of relativity. Now the Michelson interferometer may soon provide dramatic confirmation of Einstein's theory through the direct detection of gravitational radiation.

An accelerating electric charge produces electromagnetic radiation – light. It should come as no surprise that an accelerating mass produces gravitational light, namely gravitational radiation. In 1888 Heinrich Hertz had the luxury to produce and detect electromagnetic radiation in his laboratory. There will be no such luck with gravitational radiation because gravity is an extremely weak force.

Albert Einstein postulated the existence of gravitational radiation in 1916, and in 1989 Joe Taylor and Joel Weisberg indirectly confirmed its existence through observations of the orbital decay of the binary pulsar 1913 + 16. Direct detection of gravity waves will be difficult. For humans to have any chance of detecting gravity waves there must be extremely massive objects accelerating up to relativistic velocities. The only practical sources are astrophysical: supernovae, pulsars, neutron star or black hole binary systems, black hole formation or even the Big Bang. The observation of these types of events would be extremely significant for contributing to knowledge in astrophysics and cosmology. Gravity waves from the Big Bang would provide information about the Universe at its earliest moments. Observations of supernovae will yield a gravitational snapshot of these extreme cataclysmic events. Pulsars are neutron stars that can spin on their axes at frequencies in the hundreds of hertz, and the signals from these objects will help to decipher their characteristics. Gravity waves from the final stages of coalescing binary neutron stars could help to accurately determine the size of these objects and the equation of state of nuclear matter. The observation of black hole formation from these binary systems would be the *coup de grace* for the debate on the existence of black holes, and the ultimate triumph for general relativity.

Electromagnetic radiation has an electric field transverse to the direction of propagation, and a charged particle interacting with the radiation will experience a force. Similarly, gravitational radiation will produce a transverse force on massive objects, a tidal force. Explained via general relativity it is more accurate to say that gravitational radiation will deform the fabric of space-time. Just like electromagnetic radiation there are two polarizations for gravity waves. Let us imagine a linearly polarized gravity wave propagating in the z-direction, $h(z,t) = h_{0+}e^{i(kz-\omega t)}$. The fabric of space is stretched due to the strain created by the gravity wave. Consider a length L_0 of space along the x-axis. In the presence of the gravity wave the length oscillates like

$$L(t) = L_0 + \frac{h_{0+}L}{2}\cos\omega t$$

hence there is a change in its length of

$$\Delta L_x = \frac{h_{0+}L}{2}\cos\omega t$$

A similar length L_0 of the y-axis oscillates, like

$$L(t) = L_0 - \frac{h_{0+}L}{2}\cos\omega t$$

or

$$\Delta L_y = \frac{-h_{0+}L}{2}\cos\omega t$$

One axis stretches while the perpendicular one contracts, and then vice versa, as the wave propagates through. The other polarization (h_{0x}) produces a strain on axes 45° from (x,y). Imagine some astrophysical event produces a gravity wave that has amplitude h_{0+} on Earth; in order to detect a small distance displacement ΔL one should have a detector that spans a large length L. A supernova within our own Galaxy might possibly produce a gravity wave of size $h \sim 10^{-18}$ with characteristic frequencies around

1 kHz, but the occurrence of an event such as this will be extremely rare. More events will come from novae in other galaxies, but due to the great distances and the fact that the magnitude of the gravity wave falls of as $1/r$, such events will be substantially diminished in magnitude.

A Michelson interferometer, with arms aligned along the x- and y- axes, can measure small phase differences between the light in the two arms. Therefore, this type of interferometer can turn the length variations of the arms produced by a gravity wave into changes in the interference pattern of the light exiting the system. This was the basis of the idea from which modern laser interferometric gravitational radiation detectors have evolved. Imagine a gravity wave of amplitude h is incident on an interferometer. The change in the arm length will be $h \sim \Delta L/L_0$, so in order to optimize the sensitivity it is advantageous to make the interferometer arm length L_0 as large as possible. Detectors coming on-line are attempting to measure distance displacements that are of order $\Delta L \sim 10^{-18}$ m or smaller, less that the size of an atomic nucleus! If the interferometers can detect such distance displacements, and directly detect gravitational radiation, it will be one of the most spectacular accomplishments in experimental physics. The first implementation of a laser interferometer to detect a gravity wave was by Forward who used earphones to listen to the motion of the interference signal. The engineering of signal extraction for modern interferometers is obviously far more complex.

Numerous collaborations are building and operating advanced interferometers in order to detect gravitational radiation. In the United States there is the Laser Interferometer Gravity Wave Observatory (LIGO), which consists of two 4-km interferometers located in Livingston, Louisiana and Hanford, Washington. In addition, there is an additional 2-km interferometer within the vacuum system at Hanford. An Italian–French collaboration (VIRGO) has a 3-km interferometer near Pisa, Italy. GEO, a German–British collaboration, is a 600-m detector near Hanover, Germany. TAMA is the Japanese 300-m interferometer in Tokyo. The Australians (AIGO) are constructing a 500-m interferometer in Western Australia. All of the detectors will be attempting to detect gravity waves with frequencies from about 50 Hz to 1 kHz.

As will be described below, there are a number of terrestrial noise sources that will inhibit the performance of the interferometric detectors. The sensitivity of detection increases linearly with interferometer arm length, which implies that there could be advantages to constructing a gravity wave detector in space. This is the goal of the Laser Interferometer Space Antenna (LISA) collaboration. The plan is to deploy three satellites in a heliocentric orbit with a separation of about 5×10^6 km. The launch for LISA is planned for around 2015, and the detector will be sensitive to gravity waves within the frequency range of 10^{-3} to 10^{-1} Hz. Due to the extremely long baseline, LISA is not strictly an interferometer, as most light will be lost as the laser beams expand while traveling such a great distance. Instead, the phase of the received light will be detected and used to lock the phase of the light that is re-emitted by another laser.

Joe Weber pioneered the field of gravitational radiation detection in the 1960s. He used a 1400-kg aluminum cylinder as an antenna. Gravitational waves would hopefully excite the lowest-order normal mode of the bar. Since the gravity wave is a strain on space-time, experimentalists win by making their detectors longer in length. This provides a long-term advantage for laser interferometers, which can scale in length relatively easily. However, as of 2002, interferometers and bars (now cooled to 4 K) have comparable sensitivities.

Interferometer Styles

The Michelson interferometer is the tool to be used to detect a gravitational wave. Figure 1 shows a basic system. The beamsplitter and the end mirrors would be suspended by wires, and effectively free to move in the plane of the interferometer. The arms have lengths L_1 and L_2 that are roughly equal on a kilometer scale. With a laser of power P and wavelength λ incident on the beamsplitter, the light

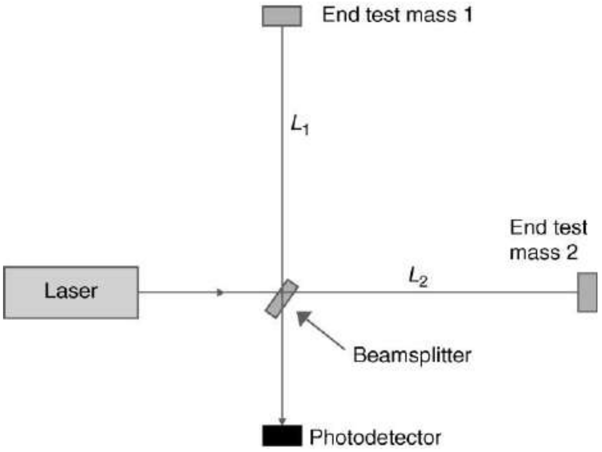

Figure 1 A basic Michelson interferometer. The photodetector receives light exiting the dark port of the interferometer and hence the signal.

exiting the dark port of the interferometer is

$$P_{\mathrm{out}} = P \sin^2\left[\frac{2\pi}{\lambda}(L_1 - L_2)\right]$$

The interferometer operates with the condition that in the absence of excitation the light exiting the dark port is zero. This would be the case for a simple and basic interferometer. However, interferometers like those in LIGO will be more sophisticated, and will use a heterodyne detection strategy. If E_0 is the amplitude of the electric field from the laser, and assuming the use of a 50–50 beamsplitter, the electric field (neglecting unimportant common phase shifts) for the light incident on the photodetector would be

$$\begin{aligned} E_{\mathrm{out}} &= \frac{E_0}{2}\left(\mathrm{e}^{\mathrm{i}\delta\phi_1} - \mathrm{e}^{\mathrm{i}\delta\phi_2}\right) \approx \mathrm{i}\frac{E_0}{2}(\phi_1 - \phi_2) \\ &= \mathrm{i}E_0\frac{2\pi}{L}(L_1 - L_2) \end{aligned}$$

The laser light will be phase modulated at a frequency in the MHz regime. As such, the deconvolved current from the photodiode that detects light at the interferometer output will be proportional to the phase acquired by the light, namely $I \propto \frac{2\pi}{\lambda}(L_1 - L_2)$.

A gravity wave of optimal polarization normally incident upon the interferometer plane will cause one arm to decrease in length while the other increases. The Michelson interferometer acts as a gravity wave transducer; the change in arm lengths results in more light exiting the interferometer dark port. The mirrors in the interferometer are suspended via wires so that they are free to move under the influence of the gravity wave.

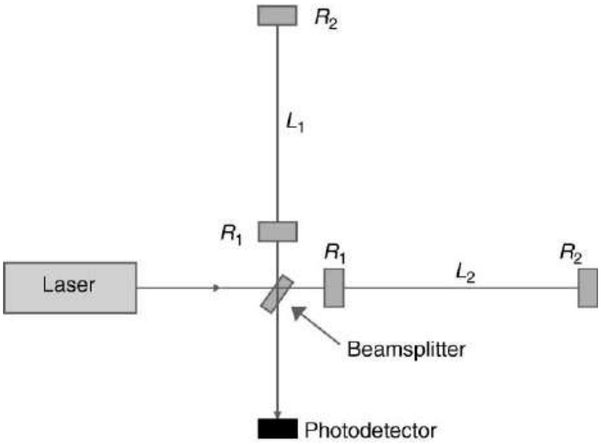

Figure 2 A Michelson interferometer with Fabry–Perot cavities in each arm. The front cavity mirrors have reflectivity R_1 while the end mirrors have $R_2 \sim 1$. By using Fabry–Perot cavities LIGO will increase the effective arm length by a factor of 75.

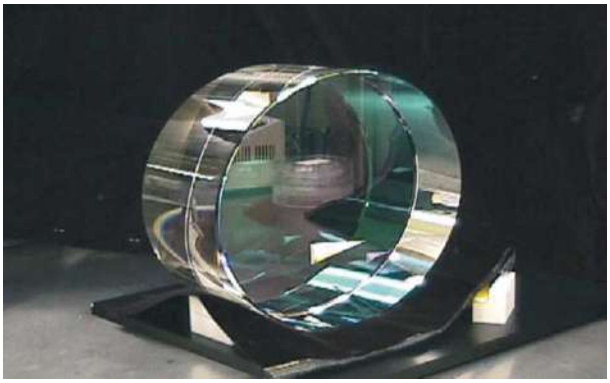

Figure 3 A picture of a mirror and test mass for LIGO. The fused silica component is 10.7 kg in mass and 25 cm in diameter. Photograph courtesy of Caltech/LIGO.

An interferometer's sensitivity increases with arm length, but geographical and financial constraints will limit the size of the arms. If there could be some way to bounce the light back and forth to increase the effective arm length it would increase the detector performance. Fabry–Perot cavities on resonance light have a storage time of $2L/[c(1 - \sqrt{R_1 R_2})]$. Figure 2 shows the system of a Michelson interferometer with Fabry–Perot cavities. This gravity wave interferometer design was proposed and tested in the late 1970s by Ron Drever. The far mirror R_2 has a very high reflectivity ($R_2 \sim 1$) in order to ultimately direct the light back towards the beamsplitter. The front mirror reflectivity R_1 is such that LIGO's effective arm length increases to $L \sim 300$ km. The optical properties of the mirrors of the Fabry–Perot cavities must be exquisite in order to achieve success. LIGO's mirrors were tested, and the root mean squared surface uniformity is less than 1 nm, scattered light is less than 50 parts per million (ppm), absorption is less than 2 ppm, and the radii of curvature for the mirrors are matched to less than 3%. A LIGO test mass (and therefore a Fabry–Perot mirror) can be seen in Figure 3.

In 1888 Michelson and Morley, with their interferometer, had a sensitivity that allowed the measurement of 0.02 of a fringe, or about 0.126 radian. Prototype interferometers constructed by the LIGO science team have already demonstrated a phase noise spectral density of $\phi(f) = 10^{-10}$ radian$/\sqrt{\mathrm{Hz}}$ for frequencies above 500 Hz. Assuming a 1 kHz signal with 1 kHz bandwidth this implies a phase sensitivity of $\Delta\phi = 3.2 \times 10^{-9}$ radian. This is about the phase sensitivity that LIGO hopes to accomplish in the 4-km Fabry–Perot system. There has been quite an evolution in interferometry since Michelson's time.

The noise sources that inhibit the interferometer performance are discussed below. However, let us

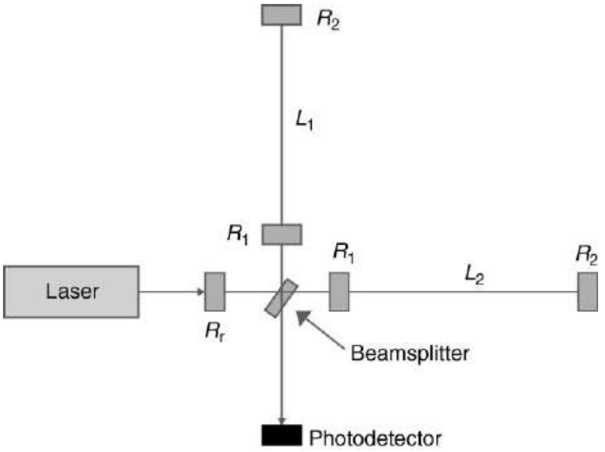

Figure 4 A power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light would exit the interferometer through the light port and head back to the laser. Installation of the recycling mirror with reflectivity R_r sends the light back into the system. A Fabry–Perot cavity is formed between the recycling mirror and the first mirror (R_1) of the arms. For LIGO this strategy will increase the power circulating in the interferometer by a factor of 50.

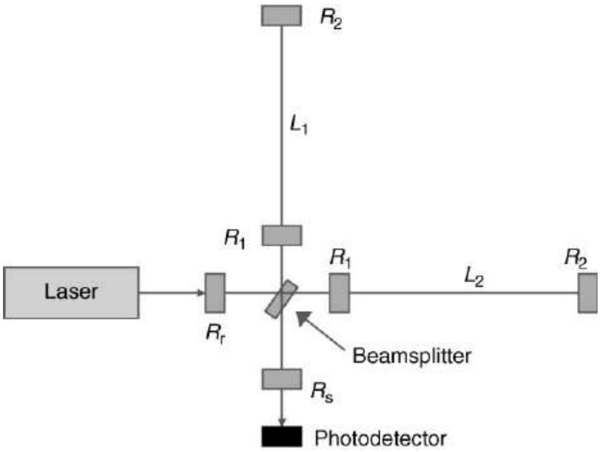

Figure 5 A signal recycled and power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light containing the gravity wave signal would exit the interferometer through the dark port and head to the photodetector. Installation of the signal recycling mirror with reflectivity R_s sends the light back into the system. The phase of the light acquired from the gravity wave will build up at a particular frequency determined by the reflectivity R_s.

consider one's ability to measure the relative phase between the light in the two arms. The Heisenberg uncertainty relation for light with phase ϕ and photon number N is $\Delta\phi\Delta N \sim 1$. For a measurement lasting time τ using laser power P and frequency f, the photon number is $N = P\lambda\tau/hc$, and with Poisson statistics describing the light $\Delta N = \sqrt{N} = \sqrt{P\lambda\tau/hc}$. Therefore

$$\Delta\phi\Delta N = \frac{2\pi}{\lambda}\Delta L\sqrt{P\lambda\tau/hc} = 1$$

implies that

$$\Delta L = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P\tau}}$$

With more light power the interferometer can measure smaller distance displacements and achieve better sensitivity. LIGO will use about 10 W of laser power, and will eventually work towards 100 W. However, there is a nice trick one can use to produce more light circulating in the interferometer, namely power recycling. Figure 4 displays the power recycling interferometer design. The interferometer operates such that virtually none of the light exits the interferometer dark port, and the bulk of the light returns towards the laser. An additional mirror, R_r in Figure 4, recycles the light. For LIGO, recycling will increase the effective light power by another factor of 50. The higher circulating light power therefore improves the sensitivity.

There is one additional modification to the interferometer system that can further improve sensitivity, but only at a particular frequency. A further Fabry–Perot system can be made by installing what is called a signal recycling mirror; this would be mirror R_s in Figure 5. Imagine light in arm 1 on the interferometer that acquires phase as the arm expands due to the gravity wave. The traveling gravity wave's oscillation will subsequently cause arm 1 to contract while arm 2 expands. If the light that was in arm 1 could be sent to arm 2 while it is expanding, then the beam would acquire additional phase. This process could be repeated over and over. Mirror R_s serves this purpose, with its reflectivity defining the storage time for light in each interferometer arm. The storage time defined by the cavity formed by the signal recycling mirror, R_s, and the mirror at the front of the interferometer arm cavity, R_1, determines the resonance frequency. Signal recycling will give a substantial boost to interferometer sensitivity at a particular frequency, and will eventually be implemented in all the main ground-based interferometric detectors.

The LIGO interferometers are infinitely more complex than the relatively simple systems displayed in the figures of this paper. Figure 6 presents an aerial view of the LIGO site at Hanford, Washington. The magnitude of the 4-km system is apparent.

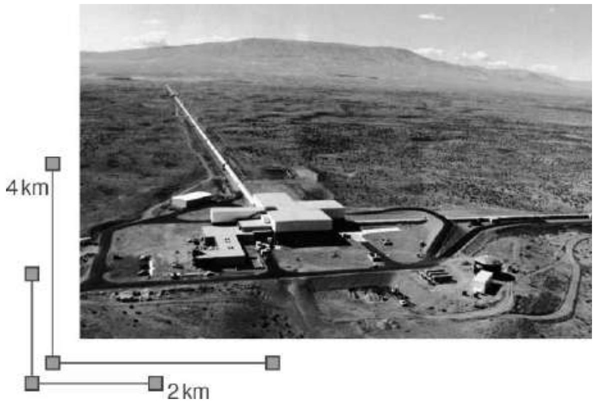

Figure 6 Arial view of the LIGO Hanford, Washington site. The vacuum enclosure at Hanford contains both 2-km and 4-km interferometers. Photograph courtesy of Caltech/ LIGO.

Noise Sources and Interferometer Sensitivity

If the interferometers are to detect distance displacements of order $\Delta L \sim 10^{-18}$ m then they must be isolated from a host of deleterious noise sources. Seismic disturbances should not shake the interferometers. Thermal excitation of components will affect the sensitivity of the detector and should be minimized. The entire interferometer must be in an adequate vacuum in order to avoid fluctuations in gas density that would cause changes in the index of refraction and hence modification of the optical path length. The laser intensity and frequency noise must be minimized. The counting statistics of photons influences accuracy. If ever there was a detector that must avoid Murphy's law this is it; little things going wrong cannot be permitted if such small distance displacements are to be detected. The expected noise sensitivity for the initial LIGO interferometers is displayed in Figure 7.

In the best of all worlds the interferometer sensitivity will be limited by the counting statistics of the photons. A proper functioning laser will have its photon number described by Poisson statistics, or shot noise; if the mean number of photons arriving per unit time is N then the uncertainty is $\Delta N = \sqrt{N}$, which as noted above implies an interferometer displacement sensitivity of

$$\Delta L = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P\tau}}$$

or a noise spectral density of

$$\Delta L(f) = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P}} \quad \text{(in units of m}/\sqrt{\text{Hz}})$$

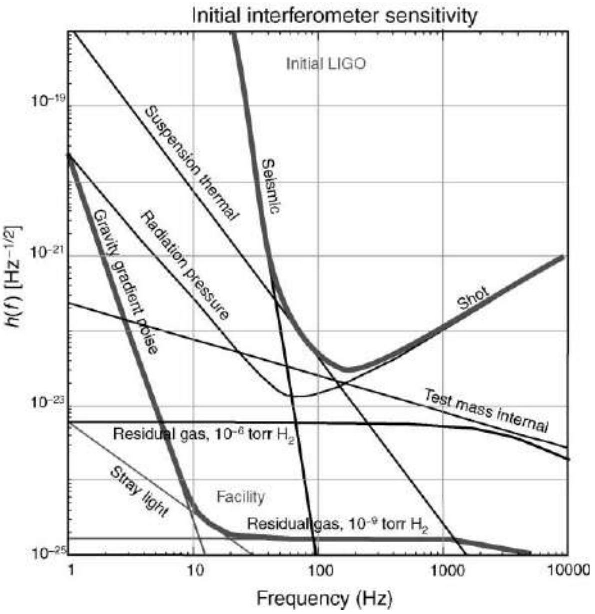

Figure 7 The target spectral density of the noise for the initial LIGO system. LIGO will be dominated by seismic noise at low frequencies (10–100 Hz), thermal noise (from the suspension system and internal modes within the mirrors) in the intermediate regime (100–300 Hz), and photon shot noise thereafter. Other sources of noise are also noted, specifically gravity gradients (gravitational excitation of masses from the seismic motion of the ground), radiation pressure of photons on the mirrors, stray (scattered) light, and index of refraction fluctuations from residual gas in the vacuum.

Note also that the sensitivity increases as the light power increases. The reason for this derives from the statistics of repeated measurements. The relative lengths of the interferometer arm lengths could be measured, once, by a photon. However, the relative positions are measured repeatedly with every photon from the laser, and the variance of the mean decreases as $\sqrt{N}$, where N is the number of measurements (or photons) involved. The uncertainty in the difference of the interferometer arm lengths is therefore inversely proportional to photon number, and hence the laser's power. In terms of strain sensitivity this would imply

$$h(f) = \frac{1}{2\pi L}\sqrt{\frac{hc\lambda}{P}} \quad \text{(in units of } 1/\sqrt{\text{Hz}})$$

This assumes the light just travels down the arm and back once. With Fabry–Perot cavities the light is stored, and the typical photon takes many trips back and forth before exiting the system. In order to maximize light power the end mirrors ($R_2 \sim 1$) and the strain sensitivity is improved to

$$h(f) = \frac{1}{2\pi L'}\sqrt{\frac{hc\lambda}{P}}$$

where $L' = 4L/(1 - R_1)$.

As the frequency of gravity waves increases the detection sensitivity will decrease. If the gravity wave causes the interferometer arm length to increase, then decrease, while the photons are still in the arm cavity, then the phase acquired from the gravity wave will be washed away. This is the reason why interferometer sensitivity decreases as frequency increases, and explains the high-frequency behavior seen in Figure 7. Taking this into account, the strain sensitivity is

$$h(f) = \frac{1}{2\pi L'}\sqrt{\frac{hc\lambda}{P}}\left(1 + \left(\frac{2\pi L' f}{c}\right)^2\right)^{1/2}$$

where $L' = 4L/(1 - R_1)$ and f is the frequency of the gravity wave.

If the gravitational wave is to change the interferometer arm length then the mirrors that define the arm must be free to move. In systems like LIGO wires suspend the mirrors; each mirror is like a pendulum. While allowing the mirrors to move under the influence of the gravity wave is a necessary condition, the pendulum itself is the first component of an elaborate vibration isolation system. Seismic noise will be troublesome for the detector at low frequencies. The spectral density of the seismic noise is about $(10^{-7}/f^2)\mathrm{m}/\sqrt{\mathrm{Hz}}$ for frequencies above 1 Hz. A simple pendulum, by itself, acts as a motion filtering device. Above its resonance frequency the pendulum filters motion with a transfer function like $T(f) \propto (f_0/f)^2$. Detectors such as LIGO will have a pendulum with resonant frequencies of about $f_0 \approx 1$ Hz, thus providing an isolation of 10^4 when looking for signals at $f = 100$ Hz. The various gravity wave detector collaborations have different vibration isolation designs. The mirrors in these interferometers will be suspended in elaborate vibration isolation systems, which may include multiple pendulum and isolation stacks. Seismic noise will be the limiting factor for interferometers seeking to detect gravity waves in the tens of hertz range, as can be seen in the sensitivity curve presented in Figure 7.

Due to the extremely small distance displacements that these systems are trying to detect it should come as no surprise that thermal noise is a problem. This noise enters through a number of components in the system. The two most serious thermal noise sources are the wires suspending the mirrors in the pendulum, and the mirrors themselves. Consider the wires; there are a number of modes that can oscillate (i.e. violin modes). At temperature T each mode will have energy of $k_B T$, but distributed over a band of frequencies determined by the quality factor (or Q) of the material. Low-loss (or high-Q) materials work best; for the violin modes of the wires there will be much noise at particular frequencies (in the hundreds of hertz). The mirror is a cylindrical object, which will have normal modes of oscillation that can be thermally excited. The first generation of LIGO will have these masses composed of fused silica, which is typical for optical components. The Qs for the internal modes are greater than 2×10^6. A switch may eventually be made to sapphire mirrors, which have better thermal properties. The limitation to the interferometers' sensitivity due to the thermal noise internal to the mirrors can be seen in Figure 7, and will be a worrying noise source for the first-generation LIGO in the frequency band around 100–400 Hz.

The frequency noise of the laser can couple into the system to produce length displacement sensitivity noise in the interferometer. With arm lengths of 4 km, it will be impossible to hold the length of the two arms absolutely equal. The slightly differing arm spans will mean that the light sent back from each of the two Fabry–Perot cavities will have slightly differing phases. As a consequence, great effort is made to stabilize the frequency of the light entering the interferometer. The LIGO laser can be seen in Figure 8. The primary laser is a Lightwave Model 126 nonplanar ring oscillator. High power is generated from this stabilized laser through the use of optical amplifiers. The beam is sent through four optical amplifiers, and then retro-reflected back

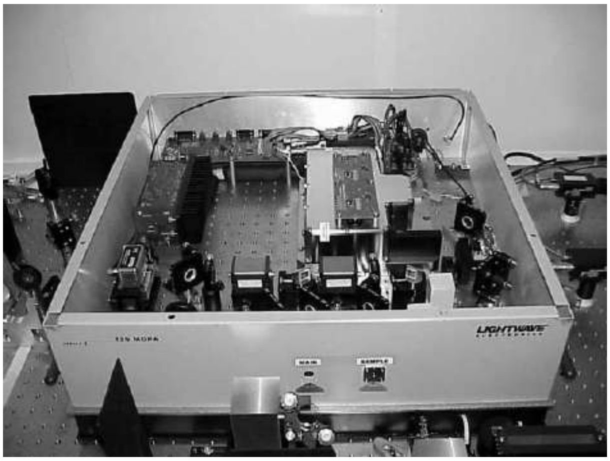

Figure 8 A picture of the Nd:YAG laser and amplifier system that produces 10 W of light for LIGO. Photograph courtesy of Caltech/LIGO.

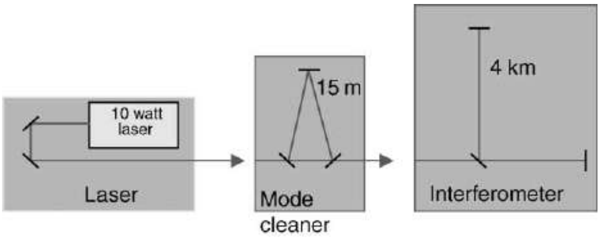

Figure 9 The optical system for the LIGO system. Light moves from the frequency-stabilized laser to a mode cleaner, and then to the interferometer. Presently 8 W of TEM_{00} laser power is delivered to the interferometer.

through the amplifiers again. For LIGO, the laser is locked and held to a specific frequency by use of signals from a reference cavity, a mode cleaner cavity, and the interferometer. For low-frequency stabilization the temperature of the ring oscillator is adjusted. At intermediate frequencies adjustment is made by signals to a piezo-electric transducer within the ring oscillator cavity. At high frequencies the noise is reduced with the use of an electro-optic crystal. The LIGO lasers currently have a frequency noise of $2\times10^{-2}\,\text{Hz}/\sqrt{\text{Hz}}$ at frequencies above 1 kHz.

It will prove important to worry about the stability of the laser power for the interferometric detectors. The hope is to be shot noise limited at frequencies above a few hundred hertz. The Nd:YAG power amplifiers used are pumped with an array of laser diodes, so the light power is controlled through feedback to the laser diodes. The LIGO requirements for the fluctuations on the power P are $\Delta P/P < 10^{-8}/\sqrt{\text{Hz}}$. The spatial quality of the light is ensured through the use of a mode-cleaning cavity. LIGO uses a triangular array of mirrors separated by 15 m. LIGO's input optical system can be seen in Figure 9. The current LIGO optical system yields 8 W of 1.06 μm light in the TEM_{00} mode.

Conclusion

The attempt to measure gravitational radiation with laser interferometers could possibly be the most difficult optical experiment of our time. Over a hundred years ago Michelson succeeded in carrying off experiments of amazing difficulty as he measured the speed of light and disproved the existence of the aether. Gravity wave detection is an experiment worthy of Michelson, and there are hundreds of physicists striving to make it a reality.

Great success has already been achieved. The TAMA 300-m interferometer is operational and has achieved a sensitivity $h(f) = 5\times10^{-21}/\sqrt{\text{Hz}}$ in the 700 Hz to 1.5 kHz frequency band. The LIGO interferometers are now operating, and scientists are presently de-bugging the system in order to achieve the target sensitivity. The first scientific data taking for LIGO commenced in 2002, and is continuing at the time of writing.

This will be a new telescope to peer into the heavens. With every new means of looking at the sky there has come unexpected discoveries. Physicists do know that there will be signals that they can predict. Binary systems of compact objects (neutron stars or black holes) will produce chirp signals that may be extracted by matched filtering techniques. A supernova will produce a burst that will hopefully rise above the noise. Pulsars, or neutron stars spinning about their axes at rates sometimes exceeding hundreds of revolutions per second, will produce continuous sinusoidal signals that can be seen by integrating for sufficient lengths of time. Gravity waves produced by the Big Bang will produce a background stochastic noise that can possibly be extracted by correlating the outputs from two or more detectors. These are exciting physics results that will come through tremendous experimental effort.

See also

Imaging: Interferometric Imaging. **Interferometry:** Overview.

Further Reading

Barish BC (1997) Gravitational wave detection. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 155–161. Tokyo: Universal Academic.

Blair DG, Taniwaki M, Zhao CN, *et al.* (1997) Progress in the development of technology for advanced laser interferometric gravitational wave detectors. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 75–94. Tokyo: Universal Academic.

Brillet A (1997) Virgo status report. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 163–173. Tokyo: Universal Academic.

Danzmann K (1997) LISA–A gravitational wave observatory in heliocentric orbit. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 111–116. Tokyo: Universal Academic.

Drever RWP, Ford GM, Hough J, *et al.* (1980) A gravity-wave detector using optical cavity sensing. In: Schmutzer E (ed.) *Proceedings of the Ninth International Conference on General Relativity and Gravitation*, pp. 265–267. Berlin: Springer.

Forward RL (1978) Wideband laser-interferometer gravitational-radiation experiment. *Physical Review* D17: 379.

Hough J and the GEO 600 Team (1997) GEO 600 Current status and some aspects of the design. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 175–182. Tokyo: Universal Academic.

Robertson NA (2000) Laser interferometric gravitational wave detectors. *Classical Quantum Gravity* 17: R19–R40.

Saulson PR (1994) Fundamentals of Interferometric Gravitational Wave Detectors. Singapore: World Scientific.

Taylor JH and Weisberg JM (1989) Further experimental tests of relativistic gravity using the binary pulsor PSR 1913 + 16. *Astrophysical Journal* 345: 434–450.

Thorne KS (1987) Gravitational radiation. In: Hawking S and Isreal W (eds) *300 Years of Gravitation*, pp. 330–458. Cambridge, UK: Cambridge University Press.

Tsubono K and the TAMA collaboration (1997) TAMA Project, Gravitational wave detection. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 183–191. Tokyo: Universal Academic.

Weiss R (1972) Electromagnetically coupled broadband gravitational antenna. *MIT Quarterly Progress Report* no. 105.

Phase-Measurement Interferometry

K Creath, Optineering and University of Arizona, Tucson, AZ, USA

J Schmit, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase-measurement interferometry is a way to measure information encoded in interference patterns generated by an interferometer. Fringe patterns (fringes) created by interfering beams of light are analyzed to extract quantitative information about an object or phenomenon. These fringes are localized somewhere in space and require a certain degree of spatial and temporal coherence of the source to be visible. Before these techniques were developed in the late 1970s and early 1980s, fringe analysis was done either by estimating fringe deviation and irregularity by eye or by manually digitizing the centers of interference fringes using a graphics tablet. Digital cameras and desktop computers have made it possible to easily obtain quantitative information from fringe patterns. The techniques described in this article are independent of the type of interferometer used.

Interferometric techniques using fringe analysis can measure features as small as a micron wide or as large as a few meters. Measurable height ranges vary from a few nanometers up to 10s of mm, depending upon the interferometric technique employed. Measurement repeatability is very consistent, and typically repeatability of 1/100 rms of a fringe is easily obtainable while 1/1000 rms is possible. Actual measurement precision depends on what is being measured and how the technique is implemented, while accuracy depends upon comparison to a sanctified standard.

There are many different types of applications for fringe analysis. For example, optical surface quality can be determined using a Twyman–Green or Fizeau interferometer. In addition, wavefront quality measurements of a source or an optical system can be made in transmission, and the index of refraction and homogeneity of optical materials can be mapped out. Many nonoptical surfaces can also be measured. Typically, surface topography information at some specific spatial frequency scale is extracted. These measurements are limited by the resolution of the optical system and the field of view of the imaging system. Lateral and vertical dimensions can also be measured. Applications of nonoptical surfaces include disk and wafer flatness, roughness measurement, distance and range sensing. Phase measurement techniques used in holographic interferometry, TV holography, speckle interferometry, moiré, grating interferometry, and fringe projection are used for nondestructive testing to measure surface structure as well as displacements due to stress and vibration. This article outlines the basics of phase measurement interferometry (PMI) techniques as well as the types of algorithms used. The bibliography lists a number of references for further reading.

Background

Basic Parts of a Phase-Measuring Interferometer

A phase-measuring interferometer consists of a light source, an illumination system (providing uniform illumination across the test surface), a beamsplitter (usually a cube or pellicle so both beams have the same optical path), a reference surface (needs to be good because these techniques measure the difference between the reference and test surface), a sample fixture, an imaging system (images a plane in space where the surface is located onto the camera), a camera (usually a monochrome CCD), an image digitizer/frame grabber, a computer system, software (to control the measurement process and calculate the surface map), and often a spatial or temporal phase shifter to generate multiple interferograms.

Steps of the Measurement Process

To generate a phase map, a sample is placed on a sample fixture and aligned, illumination levels are adjusted, the sample image is focused onto the camera, the fringes are adjusted for maximum contrast, the phase shifter or fringe spacing is adjusted or calibrated as necessary, a number of images is obtained and stored with the appropriate phase differences, the optical path difference (OPD) is calculated as the modulo 2π phase and then is unwrapped at each pixel to determine the phase map.

To make consistent measurements, some interferometers need to be on vibration isolation systems and away from heavy airflows or possible acoustical coupling. It helps to cover any air paths that are longer than a few millimeters. Consideration needs to be made for consistency of temperature and humidity.

The human operating the interferometer is also a factor in the measurement. Does this person always follow the same procedure? Are the measurements sampled consistently? Are the test surfaces clean? Is the sample aligned the same and correctly? Many different factors can affect a measurement. To obtain repeatable measurements it is important to have a consistent procedure and regularly verify measurement consistency.

Common Interferometer Types

One of the most common interferometers used in optical testing is the Twyman–Green interferometer (Figure 1). Typically, a computer controls a mirror pushed by a piezo-electric transducer (PZT). The test surface is imaged onto the camera and the computer has a frame grabber that takes frames of fringe data.

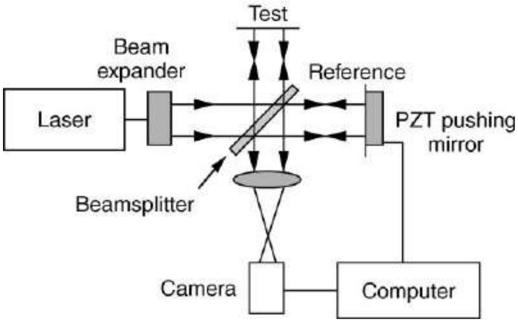

Figure 1 Twyman–Green interferometer.

The most common commercially available interferometer for optical testing is the Fizeau interferometer (Figure 2). This versatile instrument is very insensitive to vibrations due to the large common path for both interfering wavefronts. The reference surface (facing the test object) is moved by a PZT to provide the phase shift.

Interference microscopes (*see* **Microscopy:** Interference Microscopy. **Interferometry:** White Light Interferometry) are used for looking at surface roughness and small structures (see Figure 3). These instruments can employ Michelson, Mirau, and Linnik interference objectives with a laser or white light source, or Fizeau-type interference objectives that typically use a laser source because of the unequal paths in the arms of the interferometer. The phase shift is accomplished by moving the sample, the reference surface, or parts of the objective relative to the sample. Figure 3 shows a schematic of a Mirau-type interferometric microscope for phase measurement.

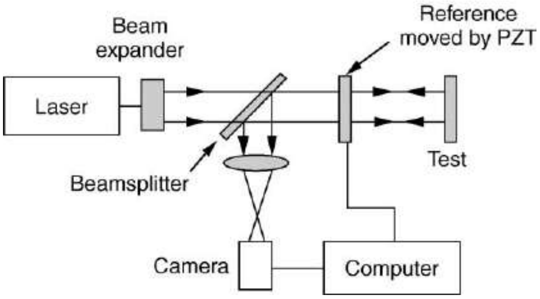

Figure 2 Fizeau interferometer.

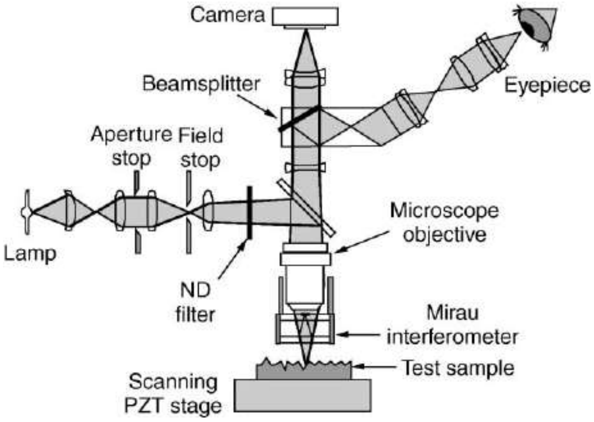

Figure 3 Mirau interference microscope.

Determination of Phase

The Interference Equation

Interference fringes from a coherent source (e.g., a laser) are theoretically sinusoidal (Figure 4a), while fringes from an incoherent source (e.g., white light) are localized in a wavepacket at a point in space (Figure 4b) where the optical paths of the arms of the interferometer are equal and marked on Figure 4 as scanner positions $z = 0$ (*see* **Interferometry:** White Light Interferometry). For generality, this analysis considers the determination of phase within the wave packet for fringes localized in space.

Interference fringes at any point in the wavepacket can be written in the following form:

$$I(x, y) = I_0\{1 + \gamma(x, y, z)\cos[\phi(x, y)]\} \qquad [1]$$

where I_0 is the dc irradiance, γ is the fringe visibility (or contrast), $2I_0\gamma$ is the modulation (irradiance amplitude or the ac part of the signal), and ϕ is the phase of the wavefront as shown in Figure 5. For simplicity, this drawing assumes that the interference fringe amplitude is constant.

Fringe visibility can be determined by calculating

$$\gamma = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \qquad [2]$$

where $I_{\max}$ is the maximum value of the irradiance for all phase values and $I_{\min}$ is the minimum value. The fringe visibility has a real value between 0 and 1 and varies with the position along the wavepacket. The fringe visibility, as defined here, is the real part of the complex degree of coherence.

Types of Phase Measurement Techniques

Phase can be determined from either a number of interference fringe patterns or from a single interferogram with appropriate fringe spacing. Temporal techniques require an applied phase shift between the test and reference beams as a function of time while multiple frames of interference fringe data are obtained. Spatial techniques can obtain data from a single interferogram that requires a carrier pattern of almost straight fringes to either compare phases of adjacent pixels or to separate orders while performing operations in the Fourier domain. Spatial techniques may also simultaneously record multiple interferograms with appropriate relative phase shift differences separated spatially in space. Multiple frame techniques require more data than single frame techniques. Temporal techniques require that the interference fringes be stable over the time period it takes to acquire the number of images. While single-frame spatial techniques require less data and can be done with a single image, they generally have a reduced amount of resolution and less precision than temporal techniques.

There are literally hundreds of algorithms and techniques for extracting phase data from interference fringe data. The references listed in the Further Reading offer more details of these techniques.

Temporal Phase Measurement

Temporal techniques use data taken as the relative phase between the test and reference beams is modulated (shifted). The phase (or OPD) is calculated at each measured point in the interferogram. As the phase shifter is moved, the phase at a single point in the interferogram changes. The effect looks like the fringes are moving across the interferogram, and because of these techniques are sometimes called fringe scanning or fringe shifting techniques. However, the fringes are not really moving; rather the irradiance at a single detector point is changing (hopefully sinusoidally) in time (see Figure 6). A 180° or π phase causes a bright fringe to become a dark fringe.

Phase Modulation Techniques

There are many ways to introduce a phase modulation (or shift). These include moving a mirror or the

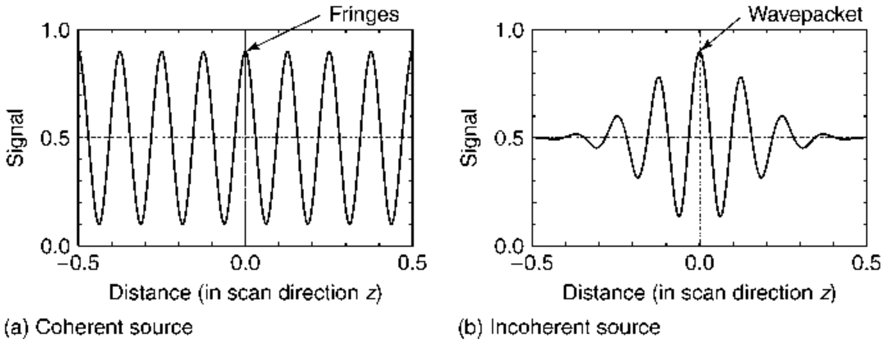

(a) Coherent source

(b) Incoherent source

Figure 4 Interference fringes for coherent and incoherent sources as observed at point $x, y, z = 0$ corresponds to equal optical path lengths of the reference and object beams.

sample, tilting a glass plate, moving a diffraction grating, rotating a polarizer, analyzer, or half-wave plate, using a two-frequency (Zeeman) laser source, modulating the source wavelength, and switching an acousto-optic Bragg cell, or magneto-optic/electro-optic cell. While any of these techniques can be used for coherent sources, special considerations need to be made for temporally incoherent sources (*see* **Interferometry:** White Light Interferometry). Figure 7 shows how moving a mirror introduces a relative phase shift between object and reference beams in a Twyman–Green interferometer.

Extracting Phase Information

Including the phase shift, the interference equation is written as

$$I(x,y) = I_0(x,y)\{1 + \gamma_0(x,y)\cos[\phi(x,y) + \alpha(t)]\} \quad [3]$$

where $I(x,y)$ is the irradiance at a single detector point, $I_0(x,y)$ is the average (dc) irradiance, $\gamma_0(x,y)$ is the fringe visibility before detection, $\phi(x,y)$ is the phase of the wavefront being measured, and $\alpha(t)$ is the phase shift as a function of time.

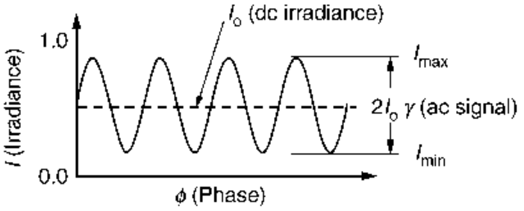

Figure 5 Variable definitions for interference fringes.

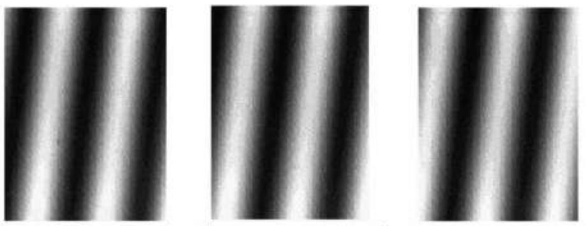

No phase shift (0°) $\pi/2$ phase shift (90°) π phase shift (180°)

Figure 6 Interference fringe patterns corresponding to different relative phase shifts between test and reference beams.

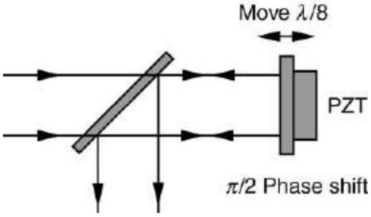

Figure 7 Moving a mirror to shift relative phase by $\lambda/4$.

Since the detector has to integrate for some finite time, the detected irradiance at a single point becomes an integral over the integration time Δ (Figure 8) where the average phase shift for the ith frame of data is α_i:

$$I_i(x,y) = \frac{1}{\Delta}\int_{\alpha_i-(\Delta/2)}^{\alpha_i+(\Delta/2)} I_0(x,y)\{1 + \gamma_0(x,y) \times \cos[\phi(x,y) + \alpha(t)]\}\,\mathrm{d}\alpha(t) \quad [4]$$

After integrating over $\alpha(t)$, the irradiance of the detected signal becomes

$$I_i(x,y) = I_0(x,y)\left\{1 + \gamma_0(x,y)\operatorname{sinc}\left(\frac{\Delta}{2}\right) \times \cos[\phi(x,y) + \alpha_i]\right\} \quad [5]$$

where $\operatorname{sinc}(\Delta/2) = \sin(\Delta/2)/(\Delta/2)$, which reduces the detected visibility.

Ramping Versus Stepping

There are two different ways of shifting the phase; either the phase shift can be changed in a constant and linear fashion (ramping) (see Figure 8) or it can be stepped in increments. Ramping provides a continuous smooth motion without any jerking motion. This option may be preferred if the motion does not wash out the interference fringes. However, ramping requires good synchronization of the camera/digitizer and the modulator to get the correct OPD (or phase) changes between data frames. Ramping allows faster data taking but requires electronics that are more sophisticated. In addition, it takes a finite time for a mass to move linearly. When ramping, the first frame or two of data usually needs to be discarded because the shift is not correct until the movement is linear.

The major difference between ramping and stepping the phase shift is a reduction in the modulation of the interference fringes after detection (the sinc($\Delta/2$) term in eqn [5]). When the phase shift is stepped ($\Delta = 0$), the sinc term has a value of one. When the phase shift is ramped ($\Delta = \alpha$) for a phase

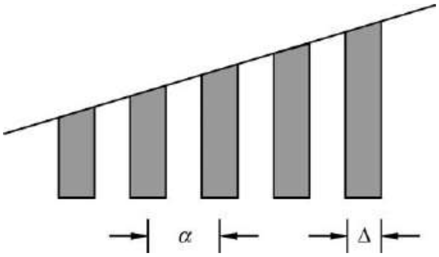

Figure 8 Change in OPD integrating over a Δ phase shift with sample separation α.

shift of $\alpha = 90° = \pi/2$, this term has a value of 0.9. Therefore, ramping slightly reduces the detected fringe visibility.

Signal Modulation

Phase measurement techniques assume that the irradiance of the interference fringes at the camera covers as much of the detector's dynamic range as possible and that the phase of the interference fringe pattern is modulating at each individual pixel as the phase between beams is modulated. If the irradiance at a single detector point does not modulate as the relative phase between beams is shifted, the height of the surface cannot be calculated. Besides the obvious reductions in irradiance modulation due to the detector sampling and pixel size or a bad detector element, scattered light within the interferometer, and defects or dirt on the test object, can also reduce signal modulation. Phase measurement techniques are designed to take into account the modulation of the signal at each pixel. If the signal does not modulate enough at a given pixel, then the data at that pixel are considered unusable, flagged as 'bad' and are often left blank. Phase values for these points may be interpolated from surrounding pixels if there are sufficient data.

Three-Frame Technique

The simplest method to determine phase uses three frames of data. With three unknowns, three sets of recorded fringe data are needed to reconstruct a wavefront providing a phase map. Using phase shifts of $\alpha_i = \pi/4$, $3\pi/4$, and $5\pi/4$, the three fringe measurements at a single point in the interferogram may be expressed as

$$I_1 = I_0\left[1 + \gamma\cos\left(\phi + \frac{\pi}{2}\right)\right] = I_0\left[1 + \frac{\sqrt{2}}{2}\gamma(\cos\phi - \sin\phi)\right] \qquad [6]$$

$$I_2 = I_0\left[1 + \gamma\cos\left(\phi + \frac{3\pi}{2}\right)\right] = I_0\left[1 + \frac{\sqrt{2}}{2}\gamma(-\cos\phi - \sin\phi)\right] \qquad [7]$$

$$I_3 = I_0\left[1 + \gamma\cos\left(\phi + \frac{5\pi}{2}\right)\right] = I_0\left[1 + \frac{\sqrt{2}}{2}\gamma(-\cos\phi + \sin\phi)\right] \qquad [8]$$

Note that (x, y) dependencies are still implied. The choice of the specific phase shift values is to make the math simpler. The phase at each detector point is

$$\phi = \tan^{-1}\left(\frac{I_3 - I_2}{I_1 - I_2}\right) \qquad [9]$$

In most fringe analysis techniques, we are basically trying to solve for terms such that we end up with the tangent function of the phase. The numerator and denominator shown above are respectively proportional to the sine and cosine of the phase. Note that the dc irradiance and fringe visibility appear in both the numerator and denominator. This means that variations in fringe visibility and average irradiance from pixel to pixel do not affect the results. As long as the fringe visibility and average irradiance at a single pixel is constant from frame to frame, the results will be good. If the different phase shifts are multiplexed onto multiple cameras, the results will be dependent upon the gain of corresponding pixels.

Bad data points with low signal modulation are determined by calculating the fringe visibility at each data point using:

$$\gamma = \frac{\sqrt{(I_3 - I_2)^2 + (I_1 - I_2)^2}}{\sqrt{2}I_0} \qquad [10]$$

It is simpler to calculate the ac signal modulation $(2I_0\gamma)$ and then set the threshold on the modulation to a typical value of about 5–10% of the dynamic range. If the modulation is less than this value at any given data point, the data point is flagged as bad. Bad points are usually caused by noisy pixels and can be due to scratches, pits, dust and scattered light.

Although three frames of data are enough to determine the phase, this and other 3-frame algorithms are very sensitive to systematic errors due to nonsinusoidal fringes or nonlinear detection, phase shifter miscalibration, vibrations and noise. In general, the larger the number of frames of data used to determine the phase, the smaller the systematic errors.

Phase Unwrapping

The removal of phase ambiguities is generally called phase unwrapping, and is sometimes known as integrating the phase. The phase ambiguities owing to the modulo 2π arctangent calculation can simply be removed by comparing the phase difference between adjacent pixels. When the phase difference between adjacent pixels is greater than π, a multiple of 2π is added or subtracted to make the difference less than π. For the reliable removal of

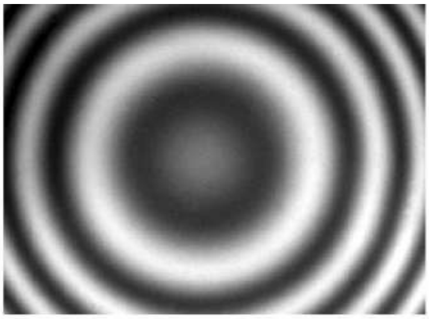

(a) Fringes for concave surface

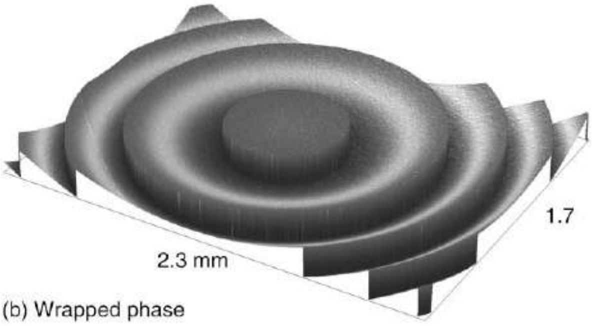

(b) Wrapped phase

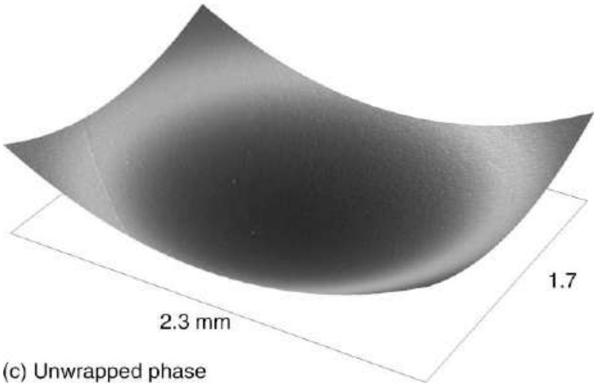

(c) Unwrapped phase

Figure 9 Fringes, wrapped, and unwrapped phase maps.

discontinuities, the phase must not change by more than π ($\lambda/2$ in optical path (OPD)) between adjacent pixels. Figure 9 shows an example of wrapped and unwrapped phase values.

Given a trouble-free wrapped phase map, it is enough to search row by row (or column by column) for phase differences of more than π between neighboring pixels. However, fringe patterns usually are not perfect and are affected by systematic errors. Some of the most frequently occurring error sources are noise, discontinuities in the phase map, violation of the sampling theorem and invalid data points (e.g., due to holes in the object or low modulation regions). Some phase maps are not easily unwrapped. In these cases different techniques like wave packet peak sensing in white light interferometry are used.

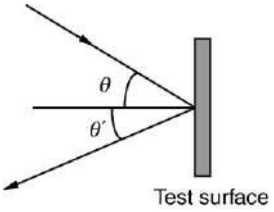

Figure 10 Angles of illumination and viewing.

From Wavefront to Surface

Once the phase of the wavefront is known, surface shape can be determined from the phase map. Surface height H of the test surface relative to the reference surface at a location (x, y) is given by

$$H(x,y) = \frac{\phi(x,y)\lambda}{2\pi(\cos\theta + \cos\theta')} \qquad [11]$$

where λ is the wavelength of illumination, and θ and θ' – the angles of illumination and viewing with respect to the surface normal – are shown in Figure 10. For interferometers (e.g., Twyman–Green or Fizeau) where the illumination and viewing angles are normal to the surface ($\theta = \theta' = 0$), the surface height is simply:

$$H(x,y) = \frac{\lambda}{4\pi}\phi(x,y) \qquad [12]$$

Since the wavefront measured represents the relative difference between the interfering reference and test wavefronts, this phase map only directly corresponds to the surface under test when the reference wavefront is perfectly flat. In practice, the shape of the reference surface needs to be accounted for by measuring it using a known test surface and subtracting this reference measurement from subsequent measurements of the test surface.

Phase Change on Reflection

Phase shifting interferometry measures the phase of reflected light to determine the shape of objects. The reflected wavefront will represent the object surface (within a scaling factor) if the object is made of a single material. If the object is comprised of multiple materials that exhibit different phase changes on reflection, the measured wavefront needs to be corrected for these phase differences (*see* **Interferometry:** White Light Interferometry).

Overview of Phase Measurement Algorithms and Techniques

There are literally hundreds of published algorithms and techniques. The optimal algorithm depends on the application. Most users prefer fast algorithms using a minimal amount of data that are as accurate and repeatable as possible, immune to noise, adaptable, and easy to implement. In practice, there are trade-offs that must be considered when choosing a specific algorithm or technique. This section provides an overview of the types of algorithms to aid the reader in sifting through published algorithms.

Synchronous Detection

One of the first techniques for temporal phase measurement utilized methods of communication theory to perform synchronous detection. To synchronously detect the phase of a noisy sinusoidal signal, the signal is first correlated (or multiplied) with sinusoidal and cosinusoidal reference signals (signals in quadrature) of the same frequency and then averaged over many periods of oscillation. This method of synchronous detection as applied to phase measurement can be extracted from the least squares estimation result when the phase shifts are chosen such that N measurements are equally spaced over one modulation period. With phase shifts α_i such that:

$$\alpha_i = \frac{2\pi}{N} i, \qquad \text{with } i = 1, \ldots, N \tag{13}$$

the phase can be calculated from

$$\phi(x, y) = \tan^{-1}\left[\frac{\sum I_i(x, y)\sin\alpha_i}{\sum I_i(x, y)\cos\alpha_i}\right] \tag{14}$$

Note that N can be any number of frames (or samples). The more frames of data, the smaller the systematic errors. This technique does not take large amounts of memory for a large number of frames, because only the running sums of the fringes multiplied by the sine and cosine of the phase shift need to be remembered. The 4-frame algorithm from Table 1 is an example of a direct adaptation of synchronous detection where simple values of 1, -1 or 0 for every $\pi/2$ phase shift can be assigned to the sine and cosine functions.

Algorithm Design

In the last ten years, a lot of work has been done to generalize the derivation of fringe analysis algorithms. This work has enabled the design of algorithms for specific applications, which are insensitive to specific systematic error sources.

Most algorithms use polynomials for the numerator and denominator. Given fringe data:

$$I_i = I_0[1 + \gamma\cos(\phi + \alpha_i)] \tag{15}$$

the phase is calculated using

$$\phi = \tan^{-1}\left[\frac{\sum\limits_i n_i I_i}{\sum\limits_i d_i I_i}\right] \tag{16}$$

Table 1 Sampling function weights for a few selected algorithms

N frames	*Phase shift*	*Coefficients*
3	$\frac{\pi}{2}$	$\frac{1, -1, 0}{0, 1, -1}$
4	$\frac{\pi}{2}$	$\frac{0, -1, 0, 1}{1, 0, -1, 0}$
5	$\frac{\pi}{2}$	$\frac{0, -2, 0, 2, 0}{1, 0, -2, 0, 1}$
7	$\frac{\pi}{3}$	$\frac{\sqrt{3}[0, 1, 1, 0, -1, -1, 0]}{-1, -1, 1, 2, 1, -1, -1}$
8	$\frac{\pi}{2}$	$\frac{1, 5, -11, -15, 15, 11, -5, -1}{1, -5, -11, 15, 15, -11, -5, 1}$
12	$\frac{\pi}{3}$	$\frac{\sqrt{3}[0, -3, -3, 3, 9, 6, -6, -9, -3, 3, 3, 0]}{2, 1, -7, -11, -1, 16, 16, -1, -11, -7, 1, 2}$

The numerator and denominator of the arctangent argument are both polynomials. The numerator is a sum proportional to the sine (imaginary part) and the denominator is a sum proportional to the cosine (real part),

$$\text{num} = 2kI_0\gamma\sin\alpha_i = \sum_i n_i I_i \tag{17}$$

$$\text{den} = 2kI_0\gamma\cos\alpha_i = \sum_i d_i I_i \tag{18}$$

where the constant k depends on the values of coefficients. From this the fringe visibility is given by

$$\gamma = \frac{\sqrt{(\text{num})^2 + (\text{den})^2}}{2kI_0} \tag{19}$$

The coefficient vectors for the numerator and denominator are window functions. For an algorithm such as the 4-frame technique, the weights of all samples are equal $[1, 1, 1, 1]$. This makes the coefficients all equal. For other algorithms, such as the 5-frame technique, the weights are larger on the middle frames than on the outer frames. The weights for the 5-frame technique are $[1, 2, 2, 2, 1]$. A property of the coefficient vectors is that the sum of the coefficients for each, the numerator and denominator, should be zero. Examples of coefficient vectors for a few selected algorithms are given in Table 1.

Heterodyne Interferometry

Historically, heterodyne techniques were developed and used before temporal phase-shifting techniques. These techniques generally determine the phase electronically at a single point by counting fringes and fractions of fringes. Areas are analyzed by scanning a detector.

Phase shifts are usually obtained using two slightly different frequencies in the reference and test beams. The beat frequency produced by the interference between the reference and test beams is compared to a reference sinusoidal signal, which may be produced either optically or electronically. The time delay (or distance traveled) between the crossing of the zero phase points of the test and reference sinusoidal signals is a measure of the phase. Every time the test signal passes through another zero in the same direction as the test surface is moved, another fringe can be counted. This is how fringe orders are counted. If the beam is interrupted as the detector is scanned across the interferogram, the fringe count is corrupted and the measurement needs to be started again. Frequency multiplication (harmonics) can also be used to determine fractions of fringes. Today, heterodyne techniques are used mainly in distance measuring interferometers. The precision and accuracy of distance measuring interferometers is at least on the order of 1 part in 10^6.

Fourier-Transform Technique

The Fourier-transform technique is a way to extract phase from a single interferogram. It is used a lot in nondestructive testing and stellar interferometry where it is difficult to get more than a single interferogram. The basic technique is shown schematically in Figure 11. The recorded interferogram distribution is Fourier transformed, and one order (usually the + or − first order) is either isolated and shifted to zero frequency or filtered out using a rectangular window. After an inverse Fourier transform, the result is the phase.

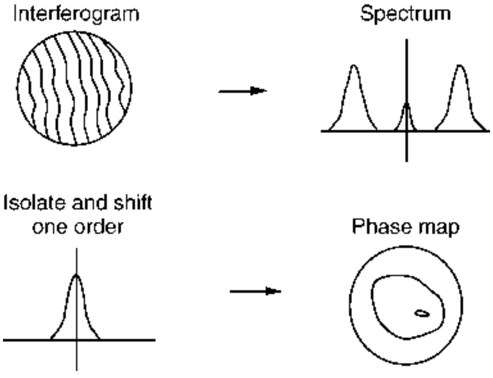

Figure 11 Fourier transform technique.

To illustrate this technique mathematically, the interference equation is rewritten as

$$I(x,y) = I_0(x,y) + c(x,y)\exp(\mathrm{i}2\pi f_0 x) + c^*(x,y)\exp(-\mathrm{i}2\pi f_0 x) \qquad [20]$$

where $c(x,y) = I_0(x,y)\gamma(x,y)\exp[\mathrm{i}\phi(x,y)]$ and the $*$ indicates a complex conjugate. The term $c(x,y)$ contains the phase information we wish to extract. After performing a one-dimensional Fourier transform:

$$I(\xi,y) = I_0(\xi,y) + c(\xi - f_0, y) + c^*(\xi - f_0, y) \qquad [21]$$

where ξ is the spatial frequency in the x direction, and italics indicate Fourier transforms. The next step is to filter out and isolate the second term, and then inverse Fourier transform to yield $c(x,y)$. The wavefront modulo 2π is then given by

$$\phi(x,y) = \tan^{-1}\left\{\frac{\mathrm{Im}[c(x,y)]}{\mathrm{Re}[c(x,y)]}\right\} \qquad [22]$$

where Re and Im refer to the real and imaginary part of the function.

This technique has limitations. If the fringes are nonsinusoidal, there will not be a simple distribution in the frequency space; there will be many orders. Another problem is overlapping orders. There needs to be a carrier frequency present that ensures that the orders are separated in frequency space. This carrier frequency is produced by adding tilt fringes until the orders are separated. This means that the aberration (fringe deviation) has to be less than the fringe spacing. Another problem is aliasing. If the interferogram is not sampled sufficiently, there will be aliasing and it will not be possible to separate the orders in frequency space. Finally, large variations in average fringe irradiance and fringe visibility across the interferogram can also cause problems.

Spatial Carrier-Frequency Technique

This is essentially the equivalent of the Fourier transform technique but is performed in the spatial domain. It is also used when there is only one interferogram available and its major applications include nondestructive testing and measurement of large optics.

These techniques relate closely to the temporal phase-measurement methods; however, instead of using a number of interferograms they can obtain all the information from a single interferogram. As an example, let's assume that the fringes are vertical and parallel to the columns on the detector array.

The carrier frequency (i.e., the number of tilt fringes) is adjusted so that there is an α phase change from the center of one pixel to the next. As long as there is not much aberration (deviation), the phase change from pixel to pixel across the detector array will be approximately constant.

When the fringes are set up this way, the phase can be calculated using adjacent pixels (see Figure 12). If one fringe takes up 4 pixels, the phase shift α between pixels will be 90°. An algorithm such as the three-frame, four-frame, or five-algorithm can be used with adjacent pixels as the input. Therefore, 3, 4, or 5 pixels in a row will yield a single-phase point. Then, the analysis window is shifted sideways one pixel and phase is calculated at the next point. This technique assumes that the dc irradiance and fringe visibility do not change over the few pixels used to calculate each phase value.

Spatial Multichannel Phase-Shift Techniques

These techniques detect all phase maps simultaneously and multiplex the phase shift using static optical elements. This can be done by using either separate cameras as illustrated below or by using different detector areas to record each of the interferograms used to calculate the phase. The phase is usually calculated using the same techniques that are used for the temporal phase techniques.

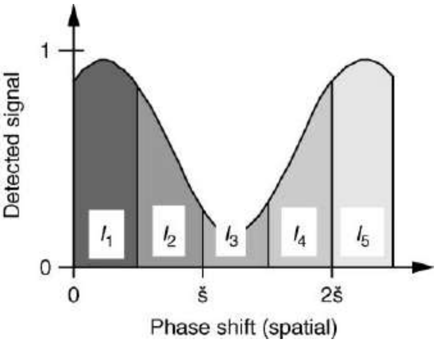

Figure 12 Spatial phase shifting. Different relative phase shifts are on adjacent detector elements.

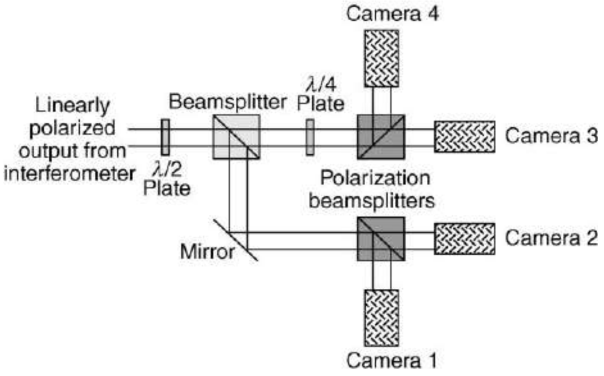

Figure 13 Spatial phase shifting using four cameras.

As an example, a four-channel interferometer can be made using the setup shown in Figure 13 to record four interferograms with 90° phase shifts between them. Camera 1 will yield fringes shifted 180° with respect to camera 2, and cameras 3 and 4 will have phase shifts of 90° and 270°. The optical system may also utilize a holographic optical element to split the beam to multiplex the four phase shifts on four quadrants of a single camera.

Signal Demodulation Techniques

The task of determining phase can be broadened by looking toward the field of signal processing. For communication via radio, radar, and optical fibers electrical engineers have developed a number of ways of compressing and encoding a signal as well as decompressing a signal and decoding it. An interference fringe pattern looks a lot like an am radio signal. Thus, it can be demodulated in similar ways. In recent years many new algorithms have been developed by drawing on techniques from communication theory and applying them to interferogram processing. Many of these use different types of transforms such as Hilbert transforms for straight fringes or circular transforms for closed fringes.

Extended Range Phase Measurement Techniques

A major limitation of phase measurement techniques is that they cannot determine surface discontinuities larger than $\lambda/4$ ($\lambda/2$ in optical path (OPD)) between adjacent pixels. One obvious solution is to use longer wavelength sources in the infrared where optically rough surfaces look smooth and their shape can be measured. An alternative is two-wavelength interferometry where two measurements at different wavelengths are taken and the measurable height limitation is now determined by the equivalent wavelength:

$$\lambda_{\mathrm{eq}} = \frac{\lambda_1 \lambda_2}{|\lambda_1 - \lambda_2|} \qquad [23]$$

Another method that allows for measurement of smooth surfaces with large height discontinuities that is limited only by the working distance of the objective combines a white light and a phase measurement interferometric technique in one long scan. The position of the wave packet resolves 2π ambiguities that result from the arctangent function (see Phase unwrapping section above). The phase of the fringe close to the wave packet maximum is determined without ambiguity. Sometimes it is not the step height discontinuity that is a problem but rather the high slope of the measured surface. If we know that the surface is continuous, then the unwrapping procedure can take advantage of this a priori information to look for a continuous.

Techniques for Deformation Measurement

Some of the techniques for deformation measurement have already been described earlier (see sections on Single frame techniques and Spatial multichannel phase-shift techniques). However, fringes in any interferometer can be analyzed not only in the x, y planes but also in the x, z, or y, z planes for which the carrier frequency of the fringes is introduced. Analysis of fringes in any plane has the same restrictions.

If the initial static object shape is measured using conventional phase measurement techniques, not only can the motion of each object point be determined but also the deformation of the whole object in time. If the motion of the object is periodic, then the object motion can be 'frozen' by using stroboscopic illumination of the same frequency as the object's motion. Once the motion is frozen, any temporal technique can be used to measure object shape at some phase of its motion. Changing the time offset between stroboscopic illumination and the periodic signal driving the motion, the object can be 'frozen' and thus measured at different phases of its periodic motion.

Systematic Errors

Noise Sensitivity

Measurement noise mostly arises from random fluctuations in the detector readout and electronics. This noise reduces precision and repeatability. Averaging multiple measurements can reduce effects of these random fluctuations.

Phase Shifter Errors

Phase shifter errors can be due to both miscalibration of the system and nonlinearities in the phase shift. It is possible to purchase very linear phase shifters. It is also possible to correct nonlinearities by determining the voltage signal making the phase shifter provide a linear phase shift.

Linear phase shifter errors (miscalibration) of the phase shift have an error signature that is at twice the frequency of the fringes. If there are two fringes across the field of view, the error signature will have four across the field of view. Figure 14 shows the difference between a calibrated and an uncalibrated phase shifter as well as the difference in error for two different phase measurement algorithms. Some algorithms are obviously more sensitive than others to this type of error.

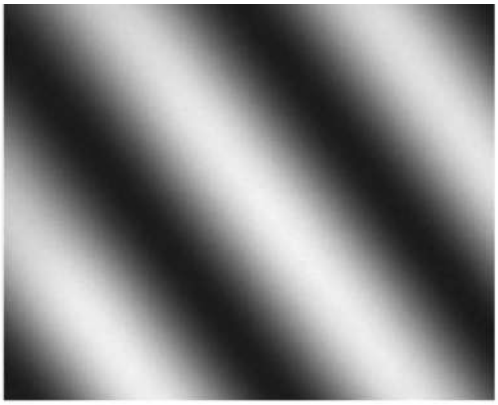

(a) Fringes on flat

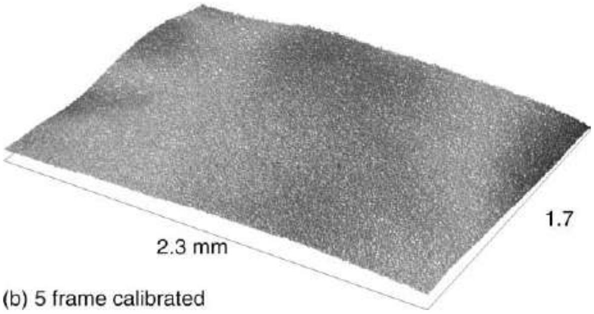

(b) 5 frame calibrated

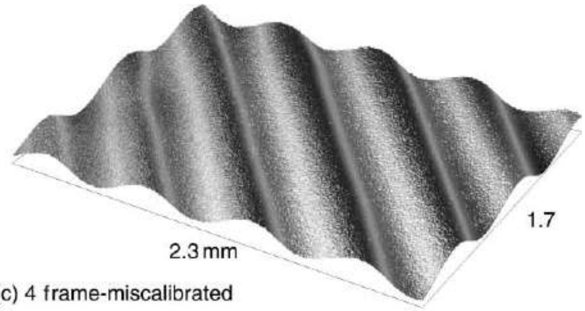

(c) 4 frame-miscalibrated

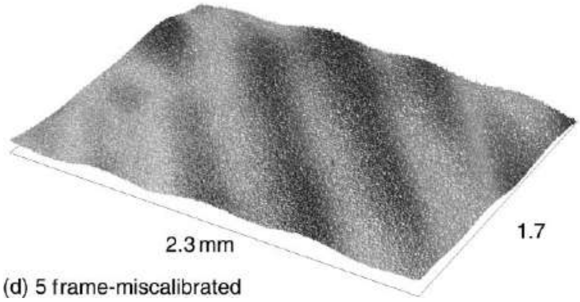

(d) 5 frame-miscalibrated

Figure 14 Comparison of phase maps for calibrated and uncalibrated phase shifts using two different phase measurement algorithms.

Other Types of Systematic Errors to Consider

Other types of errors to consider are detector nonlinearities, quantization errors due to analog-to-digital converters, and dissimilar materials (see Phase change upon reflection). For temporal phase measurement techniques errors due to vibration and

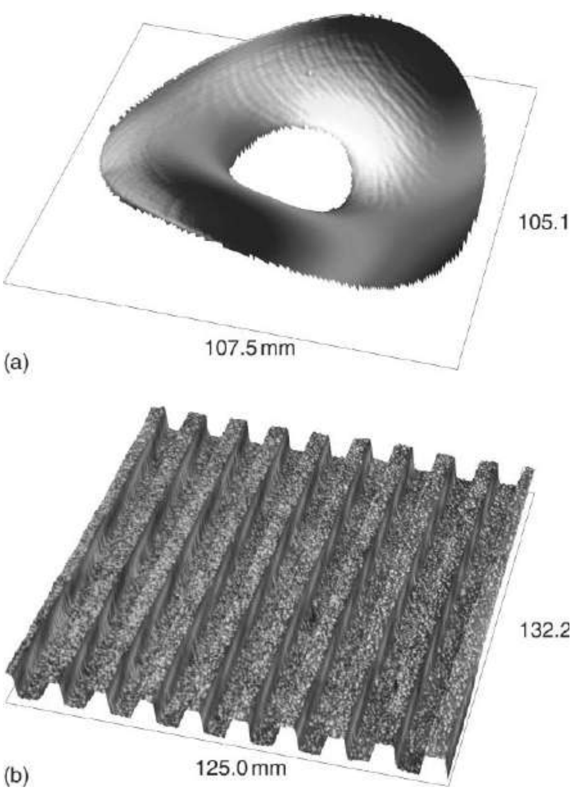

Figure 15 Measurements of (a) hard disc substrate with a vertical range is 2.2 micrometers and (b) binary grating roughness standard.

air turbulence need to be considered as well. Spatial phase measurement techniques are sensitive to miscalibrated tilt (wrong carrier frequency), unequally spaced fringes, and sampling and windowing in Fourier transform techniques.

Choosing an Algorithm or Technique

Each algorithm and type of measurement technique is sensitive to different types of systematic errors. Choosing a proper algorithm or technique for a particular type of measurement depends on the specific conditions of the test itself and reducing the systematic errors for a particular type of measurement. This is the reason that so many algorithms exist. The references in the bibliography will help the reader determine what type of algorithm will work best for a specific application.

Examples of Applications

Phase shifting interferometry can be used for measurements such as hard disk flatness, quality of optical elements, lens curvature, dimensions and quality of air-bearing surfaces of magnetic read/write heads, cantilevers, and semiconductor elements. Figure 15 shows results for measurements of a hard disk substrate and a roughness grating.

Conclusions

Phase measurement interferometry techniques have increased measurement range and precision enabling the production of more complex and more precise components. As work continues on development of interferometric techniques, phase measurement techniques will continue to become more robust and less sensitive to systematic errors. Anticipated advances will enable measurements of objects that were unimaginable 20 or 30 years ago.

See also

Interferometry: White Light Interferometry. **Microscopy:** Interference Microscopy.

Further Reading

Bruning JH (1978) Fringe scanning interferometers. In: Malacara D (ed.) *Optical Shop Testing*, pp. 409–438. New York: Wiley.

Creath K (1988) Phase-measurement interferometry techniques. In: Wolf E (ed.) *Progress in Optics*, vol. 26, pp. 349–393. Amsterdam: Elsevier.

Creath K (1993) Temporal phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*, pp. 94–140. Bristol: IOP Publishing.

Creath K and Schmit J (1996) N-point spatial phase-measurement techniques for non-destructive testing. *Optics and Lasers in Engineering* 24: 365–379.

Ghiglia DC and Pritt MD (1998) *Two-Dimensional Phase Unwrapping*. New York: Wiley.

Greivenkamp JE and Bruning JH (1992) Phase shifting interferometry. In: Malacara D (ed.) *Optical Shop Testing*, 2nd edn, pp. 501–598. New York: Wiley.

Hariharan P (1992) *Basics of Interferometry*. New York: Academic Press.

Kerr D (ed.) (2004) FASIG special issue on fringe analysis. *Optics and Lasers in Engineering* 41: 597–686.

Kujawinska M (1993) Spatial phase measurement methods. In: Robinson DW and Reid GT (eds). *Interferogram Analysis*, pp. 141–193. Bristol, UK: IOP Publishing.

Larkin KG (1996) Neat nonlinear algorithm for envelope detection in white light interferometry. *Journal of Optical Society of America A* 13: 832–842.

Malacara D (1992) *Optical Shop Testing*, 2nd edn. New York: Wiley.

Malacara D, Servin M and Malacara Z (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Dekker.

Robinson DW and Reid GT (1993) *Interferogram Analysis: Digital Processing Techniques for Fringe Pattern Measurement*. Bristol, UK: IOP Publishing.

Schmit J and Creath K (1996) Window function influence on phase error in phase-shifting algorithms. *Applied Optics* 35: 5642–5649.

Surrel Y (2000) Fringe analysis. *Photo-Mechanics: Topics in Applied Physics* 77: 55–102.

White Light Interferometry

J Schmit, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In 1960 the development of the laser injected new possibility into an old discipline, interferometry (*see* **Interferometry**: Overview). The excellent spatial and temporal coherence (ideally a single point and a single wavelength source) of the laser allowed for the formation of nonlocalized interference fringes. With lasers, researchers could easily generate good contrast fringes practically anywhere where the beams overlapped. These excellent-quality fringes from laser interferometers enabled high-resolution distance and displacement measurements on the order of meters and made possible noncontact surface probing with nanometer and subnanometer resolution. This kind of precision, which was supported by the advent of computers and the advancement of detectors, was previously unimaginable.

However, interference phenomena were used in metrology long before lasers and can be observed without any complicated device or source. On a rainy day we can observe the colorful interference patterns that have formed on a thin layer of gasoline in a puddle. The relationship between colors and the small distance between reflecting surfaces, observed as far back as 1665 by Hook, is still used to approximate layer thicknesses. The colors in interference patterns have been used to determine the spectral components of beams. By the end of the nineteenth century, Michelson and Benoit were using interference phenomena to determine distances, long before the invention of the computer and the laser.

While the application of lasers in interferometry has advanced the science, white light interferometry (WLI), which uses a spatially and temporally incoherent source and creates fringes localized over a few microns in space, has also benefited from technological advancements. The development of solid state detectors, fast computers, electronic signal processing, and precise scanning stages has allowed for incredibly fast analysis of white light interferograms. In fact, WLI is used in many applications from film thickness and surface measurement, through spectroscopy to astronomy.

This chapter focuses on the formation of white light fringes. We first examine the existence of white light fringes in the everyday world. A description of different interferometric setups and their uses follows. The bulk of this article details the formation of white light fringes and examines their different applications with an emphasis on the analysis of object structure. Suggestions for further reading, both within this encyclopedia and from outside sources, can be found at the end of this article. A basic theory of interferometry and coherence is to be found in corresponding articles of this encyclopedia.

White Light Interference in Everyday Life

The effects of interference display spectacular colors when observed in white light, such as the sun. We often see these effects in oil or gasoline spills, in soap bubbles, or between two smooth pieces of glass in contact (separated by a very thin layer of air film) – which are the simplest everyday white light interferometers. The colors we see in these displays are interference colors, and their origin is interference rather than dispersion as in the colors of a rainbow or light reflected from a CD. The picture in Figure 1 is of a gasoline spill under cloudy sky illumination; we see beautiful interference colors that are hard to find in a rainbow: iron-gray, magenta, grayish blue, whitish green, and brownish yellow. Interference colors will differ not only with the thickness of the layer but also with the layer's absorption and dispersion, the relative indices of refraction of the film, the surrounding media, and the illumination. Interference colors can be observed for layers from a fraction of a micron to a few microns thick.

Interferometers for White Light Observation

In a white light interferometer, either the colors or the intensity distribution of the fringes is typically analyzed to retrieve the necessary encoded information, such as the film thickness, birefringence, index of refraction, dispersion, spectral properties, and surface structure. White light interference can be observed only in interferometric designs where the optical paths in both arms of the interferometer are (nearly) equal and the system is (nearly) compensated for dispersion.

Interference in Thin Film

Beams of light of different wavelengths incident on a transparent thin film (such as the ones in the puddle) are partially reflected from the top air/film interface

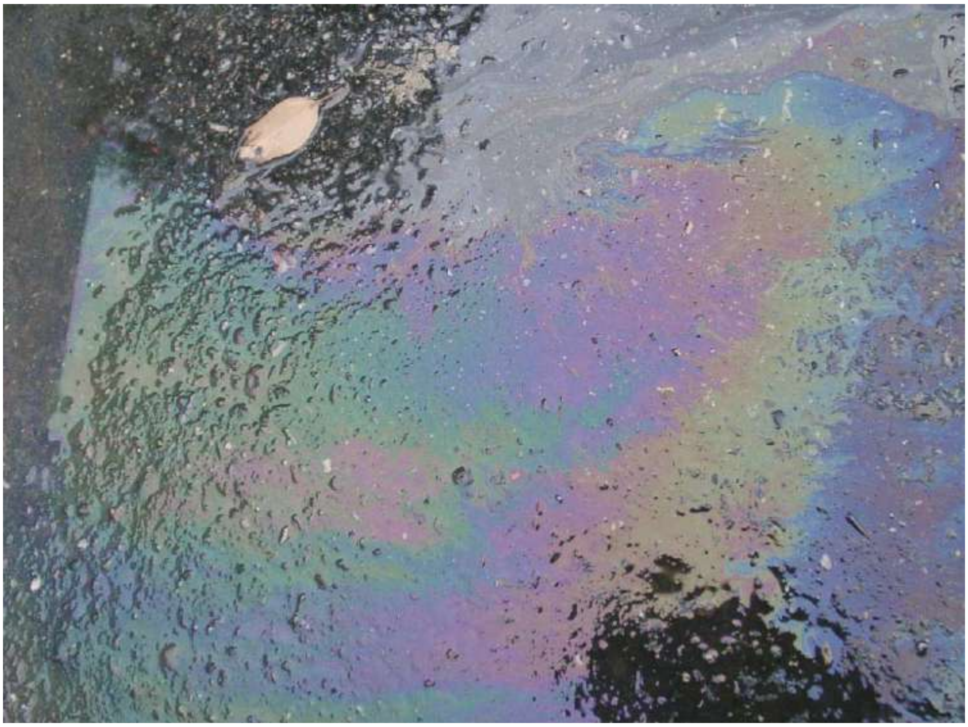

Figure 1 Interference colors in a gasoline spill in a puddle. This picture would closely resemble a picture from an interference microscope, if not for the leaf in the top left corner.

and partially from the bottom film/substrate interface and then beams of the same wavelength interfere with each other. The optical path difference (OPD) traveled by the interfering beams is related to the film thickness, the index of refraction for the given wavelength and the angle of incidence.

Fringes of equal thickness will be observed if the film thickness (distance between interfering wavefronts) varies. These fringes represent points of equal distance between the two surfaces of the film and are formed close to the film (see Figure 2). This situation is typical in any interferometer since rarely are wavefronts perfectly plane and parallel.

For any individual wavelength, λ_i, where the OPD is equal to $m\lambda_i$, and where m is an integer, a bright fringe of the color of the wavelength, λ_i, will be observed due to constructive interference. For white light illumination, the color at a given point will be dominated by the color of the wavelength for which the interference will be constructive. This color can be used to estimate the optical thickness of the film.

An additional factor needs to be taken into account when color is used for estimating optical thickness, namely the phase change on reflection. Different colors will be observed if the ratio of indices of refraction for both film interfaces is <1 (or >1) as opposed to when the ratio of indices is <1 for one interface and >1 for the other.

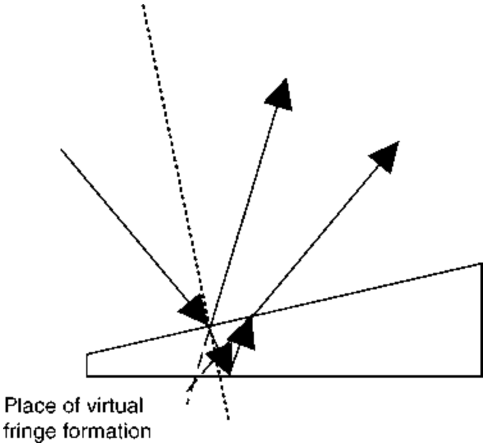

Figure 2 Formation of fringes of equal thickness for film with wedge.

Polarization Microscope

Interference colors that correspond to equal thickness fringes can represent the birefringence of an object; these colors are often observed using a polarization microscope (*see* **Microscropy:** Interference Microscopy). The polarization microscope is a conventional microscope with a set of polarizers with crossed polarization axes, one placed above and the other placed below the tested birefringent object.

Interference colors have been tabulated for many years and for many different purposes. Newton devised his color scale to describe interference fringes for two pieces of glass in contact with each other. Michel-Lévy and Lacroix in 1889, created a color scale to help recognize different rock forming minerals. For more information about colors in white light interference see sections on *White Light Interference* and *Spectral Interference* below.

Michelson Interferometer

A Michelson interferometric setup, shown in Figure 3, is often used to analyze white light interference, but the intensity of the fringes is observed as one mirror is scanned rather than the color of fringes that is usually observed.

Two plane mirrors (equivalent to the top and bottom surfaces of a thin film) return the beams to the beamsplitter, which recombines parts of the returned beams and directs them towards the detector where the interference is observed. Beam S1 travels through the parallel plate three times while beam S2 passes through the plate only once causing the system to be not well compensated (balanced) for dispersion. Thus, for observation of white light interference, an identical plate is placed in the path of beam S2. As one of the plane mirrors moves along the optical axis to change the OPD, the detector collects the irradiance, the intensity. Fringes will be observed if the system is well-compensated for dispersion and for optical path lengths. Any spectral changes or changes in optical path lengths in the interferometer affect the shape or position of the fringes, and the interferometer measures these changes.

Common-path interferometers, like the Mach–Zender or Jamin interferometers, are naturally compensated interferometers and can be used for measurement of dispersion in gases. Other interferometers, such as the Twyman–Green or Fizeau interferometers, make use of their unequal interferometer arms and their nonlocalized fringes from laser sources for testing different optical elements and systems in reflection and transition (*see* **Interferometry:** Phase Measurement Interferometry).

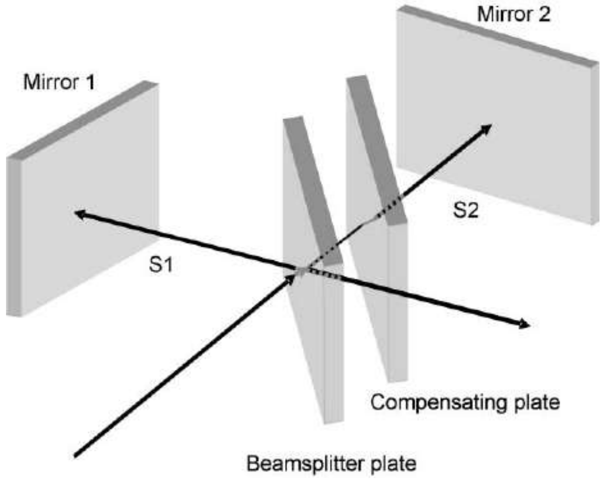

Figure 3 Michelson interferometer.

White Light Interference

A white light source consists of a wide spectrum of wavelengths in visible spectrum, from about 380 up to 750 (violet to red) nanometers. However, the principles of WLI described in this article basically apply when any low coherence source is used. Low coherence refers to a source of not only low temporal but also low spatial coherence; WLI can also be referred to as low coherence interferometry. We will be concerned mainly with temporal effect of the source, the source spectrum.

Since different wavelengths from the source spectrum are mutually incoherent, we will first look at interference between two waves of a selected monochromatic illumination with wave number $k = 2\pi/\lambda$, where λ is the wavelength. The intensity of the interference fringes at point x, y (these coordinates are omitted in all equations), as one of the mirrors is scanned along the optical axis, z can be described as

$$I(k,z) = I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma(z)|\cos(kz) \qquad [1]$$

or can be written in the form:

$$I(k,z) = I_0\left[1 + \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}|\gamma(z)|\cos(kz)\right] \qquad [2]$$

where I_1 and I_2 are intensities of each of the beams and $I_0 = I_1 + I_2$, $|\gamma(z)|$ is the modulus of the complex mutual coherence function (*see* **Coherence:** Overview) assumed here to equal 1 (perfect coherence for each wavelength). The optical path difference z equals $z_1 - z_2$ where z_1 and z_2 are the optical path lengths that the interfering waves have traveled. The difference in the traveled paths, z, corresponds to the position of the scanning mirror.

White light interference (polychromatic interference) is the overlaying of all the monochromatic fringes created for each wavelength of the source spectrum (Figure 4a). A detector observes the sum of all the fringe intensities.

In mathematical form, this interference can be described as the integral of all the fringes $I(k,z)$ for all

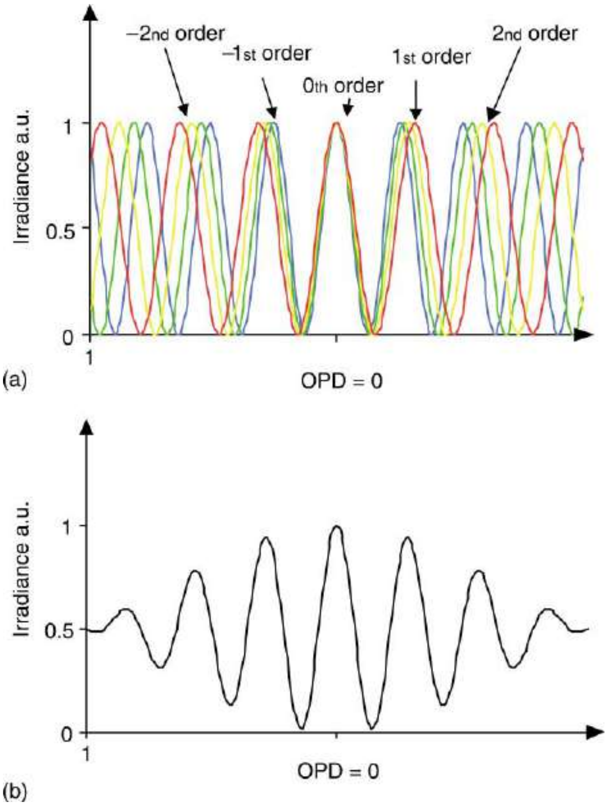

Figure 4 Formation of white light fringes: (a) fringes for individual wavelengths; (b) fringes for white light.

wavenumbers k:

$$I(z) = \int_{k_1}^{k_2} S(k)D(k)I(k,z)\,\mathrm{d}k \qquad [3]$$

where $S(k)$ is the spectral distribution of the light source, with $S(k) = 0$ outside of k_1 through k_2 range of wave numbers, and $D(k)$ is a spectral response of the detector. We assume that the detector's response $D(k)$ equals 1 over the whole spectrum. Because the spacing of the fringes for each wavelength emitted by the white light source is different, the maxima of fringes will align around only one point for a well-compensated interferometer, as shown in Figure 4a. This alignment around a single point, occurs because in a well-compensated interferometer there is one point for which the OPD is zero for all wavelengths. Away from this point, the observed sum of the intensities quickly falls off, as shown in Figure 4b. The maximum fringe, the fringe that marks the zero OPD, is called the zero-order fringe, and each next fringe of smaller amplitude on either side is called +1 and −1, +2 and −2 order fringe, and so on.

Because white light fringes are localized and can only be found within microns or tens of microns of the zero OPD, white light interferometers are excellent distance and 3D position sensors. However, this same characteristic makes them more difficult to align than interferometers with nonlocalized fringes.

Envelope of Fringes Due to Source Spectrum

The resultant intensity of interference from a broad spectrum source (see eqn [3]) can be described in general form as

$$I(z) = I_0[1 + V(z)\cos(k_0 z)] \qquad [4]$$

where $I_0 = I_1 + I_2$ is the background intensity, $V(z)$ is the fringe visibility function or coherence envelope, and $k_0 = 2\pi/\lambda_0$ is the central wave number for fringes under the envelope. The $V(z)$ is proportional to the modulus of the Fourier transform of the source spectrum $S(k)$. Generally, if the light source has a Gaussian spectrum $S(k)$, then the envelope of the fringes can be described also as a Gaussian function $V(z)$. If the spectrum of the source is rectangular, then the envelope of the fringes will be a sinc function. The wider that the spectrum of the source is the narrower the width of the envelope will be, as shown in Figure 5. The width of the fringe envelope determines the coherence length of the source (*see* **Coherence:** Overview); for a white light source this width is in the order of 1–2 microns.

Different white light sources, such as tungsten-halogen, incandescent, or arc lamps, have different spectra and thus create different coherence envelopes, as shown in Figure 5. The spectra of semiconductor light sources, such as LEDs and SLDs, are similar in shape to a Gaussian function.

The fact that the coherence envelope is related to the spectrum of the source by its Fourier transform is commonly used in Fourier transform spectroscopy, where the Fourier transform of the detected fringes is calculated to find the spectral components of the beams.

Position of Fringes Under Envelope Due to Reflection of Dielectric

Thus far we have assumed that in WLI the interferometer is compensated for all wavelengths, and thus, for this position the maximum of the fringes aligns with the maximum of the envelope, namely where there is a zero phase shift $\varphi_0 = 0$ between fringe and envelope maxima.

If there is an odd number of reflections from dielectric surfaces in one arm of the interferometer and an even number in the other, the fringes will be shifted by $\varphi_0 = 180°$ under the coherence envelope and the minimum of the fringe will align with the maximum of the coherence envelope. Thus eqn [4]

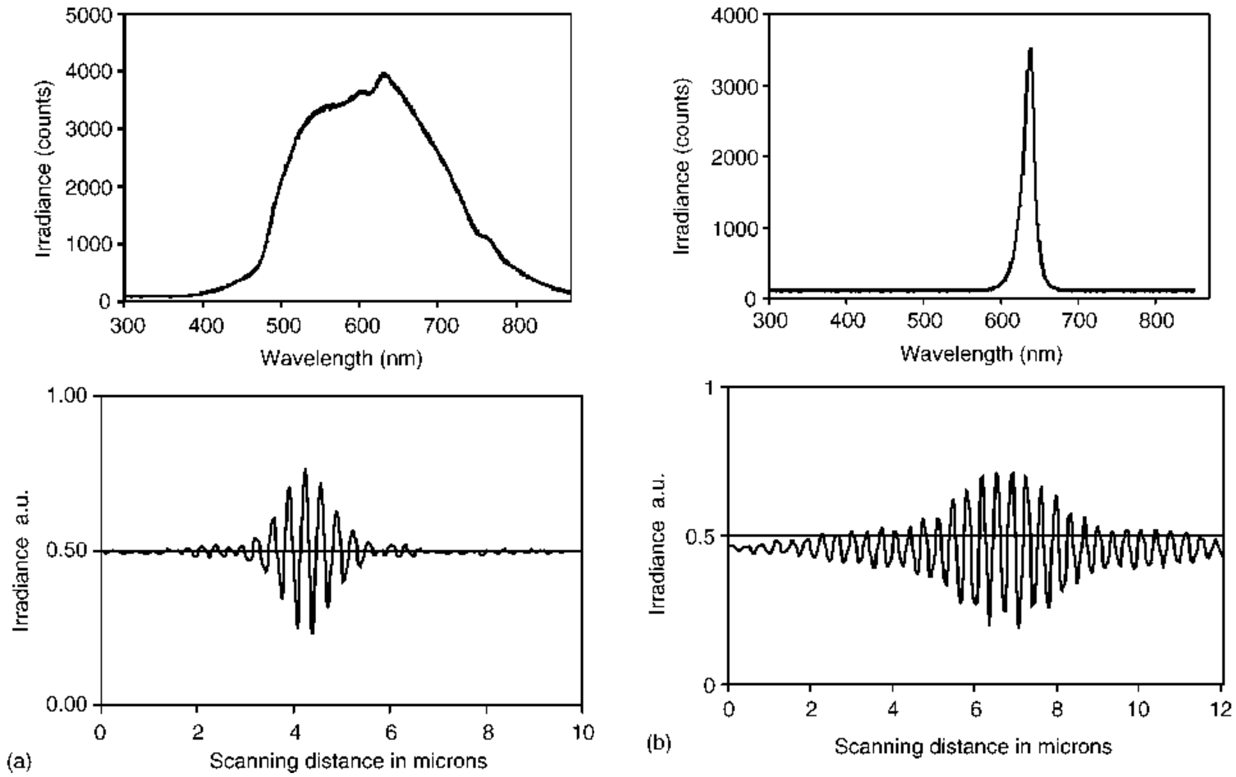

Figure 5 Spectrum and interferogram for (a) tungsten-halogen lamp and for (b) red LED sources.

can be expressed in a more general form:

$$I(z) = I_0[1 + V(z)\cos(k_0 z + \varphi_0)] \qquad [5]$$

Figure 6 shows color fringes for such a case; the dark fringe marks the zero OPD, and this fringe is surrounded by the greenish blue colors of shorter wavelengths. In contrast, the bright fringe at the maximum envelope position or marking the zero OPD would be surrounded by the reddish colors of longer wavelengths.

In a real system the fringes may be shifted with respect to the envelope by any amount of φ_0, and this shift may be due to any number of factors, such as the phase shift on reflection from nondielectric surfaces and dispersion, which we consider next.

Changes in Envelope and Fringes Due to Reflection of Nondielectric Materials

The relative position of the envelope maximum and fringe position will be different if the beam is reflected from different nondielectric materials. This difference exists because the phase change on reflection from a test surface, like metals or heavily doped semiconductors, varies with each wavelength. This variance in fringe and peak coherence position may be predicted and corrected for in surface height calculations. The linear dependence of the phase change on reflection versus wave number shifts the location of the coherence envelope peak position of fringes by different amounts. The constant phase change on reflection and higher-order terms only shift the fringes underneath the coherence envelope. As long as the change on reflection has a small dependence on the second- and higher-order of the wave number, the shape of the coherence envelope is preserved.

Changes in Envelope and Fringes Due to Dispersion

Dispersion in an interferometer that is not balanced, perhaps because a dispersive element was placed in one arm or the compensating plate has an incorrect thickness, will influence fringe formation. The phase delay between interference patterns for individual wavelengths is proportional to the product of the geometrical path and the index of refraction equal to $d \times n(k)$.

The intensity may be described as:

$$I(z) = \int_{k_1}^{k_2} \langle 1 + V(z)\cos\{kz - kd[n(k)]\}\rangle \, \mathrm{d}k \qquad [6]$$

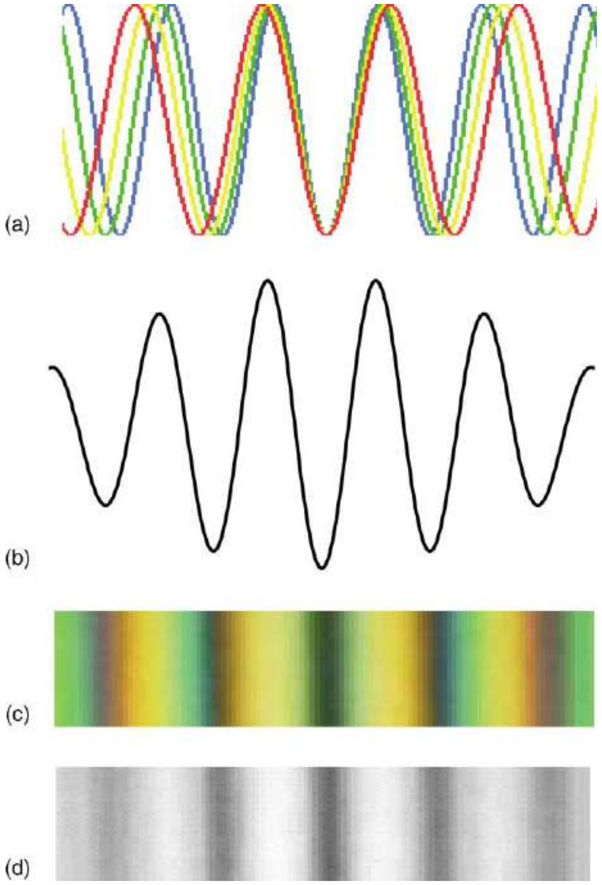

Figure 6 Formation of white light fringes with destructive interference for OPD = 0.

The dependence of the refractive index on the wave number k can be described as a linear expansion:

$$n(k) = n(k_0) + \frac{\mathrm{d}n}{\mathrm{d}k}(k - k_0) \qquad [7]$$

The linear dispersion shifts the envelope by the group index of refraction times the thickness of the dispersive element; this dispersion also shifts the fringes under the envelope slightly. In other words, the linear phase delay for a range of wavelengths causes a group delay of the whole envelope (wave packet). Higher-order dispersion, absorption of the elements, or effects due to thin films can cause the envelope to widen or even become asymmetrical, the position of the fringes under the envelope to shift, the fringes to lose contrast, and the period of the fringes to change or vary with the z position. Dispersion effects will be stronger for sources that have a wider spectrum; however, the observed changes will be different for different shapes of spectra.

The phase of the fringes under the envelope and the position of the envelope itself are parameters often used in astronomy. Highly accurate white light fringe estimation, using the optical path-length delay between the two arms of the interferometer, is a cornerstone of stellar interferometry.

Fourier Analysis of White Light Interferograms

Wavelength-dependent changes in a white light interferogram can be more easily analyzed in the spectrum rather than the spatial domain. The Fourier transform of a white light interferogram yields two symmetrical side lobes at the mean wavelength of the interferogram; analysis of just one of these side lobes is sufficient. The spectral amplitude of the side lobe contains information about the spectral components of the interfering beams, while the spectral phase in regions with appreciable spectral amplitude supplies information about any unbalanced dispersion in the interferometer, as shown in Figure 7. Fourier transform analysis is extensively used in Fourier transform spectrometry (*see* **Spectroscopy:** Fourier Transform Spectroscopy).

For a dispersion-balanced interferometer, the interferogram is symmetrical around the zero OPD position. For a symmetrical interferogram, the spectral phase will be zero if the zero OPD position of the sampled interferogram is in the middle of sampling range; otherwise a linear factor appears in the spectral phase. This linear term, while useful to surface profiling, because it determines the position of the coherence signal with respect to scanner sampling, is unwanted in Fourier spectroscopy and needs to be corrected for. Dispersion and possible thin-film effects will commonly introduce a nonlinearity in the spectral phase.

Controlled Phase Shift of Fringes Under the Envelope – Geometric Phase Shift

Many techniques in interferometry depend on shifting the phase of the interfering wavefronts. Mechanical shifters used in white light interferometry will introduce the same shift measured in nanometers for all wavelengths; however, when measured in degrees or radians, the shift will vary for different wavelengths. Geometric phase shifters (achromatic phase shifters operating on the principle of geometric phase) introduce for all wavelengths the same shift when measured in degrees and are based on polarization elements like a rotating wave plate in circularly polarized light or a rotating polarizer in a circularly polarized beam. Fringes underneath the coherence envelope shift, as shown in Figure 8, while the coherence envelope stays in place.

The advantage to these techniques is that only the phase of fringes changes, not the fringe contrast

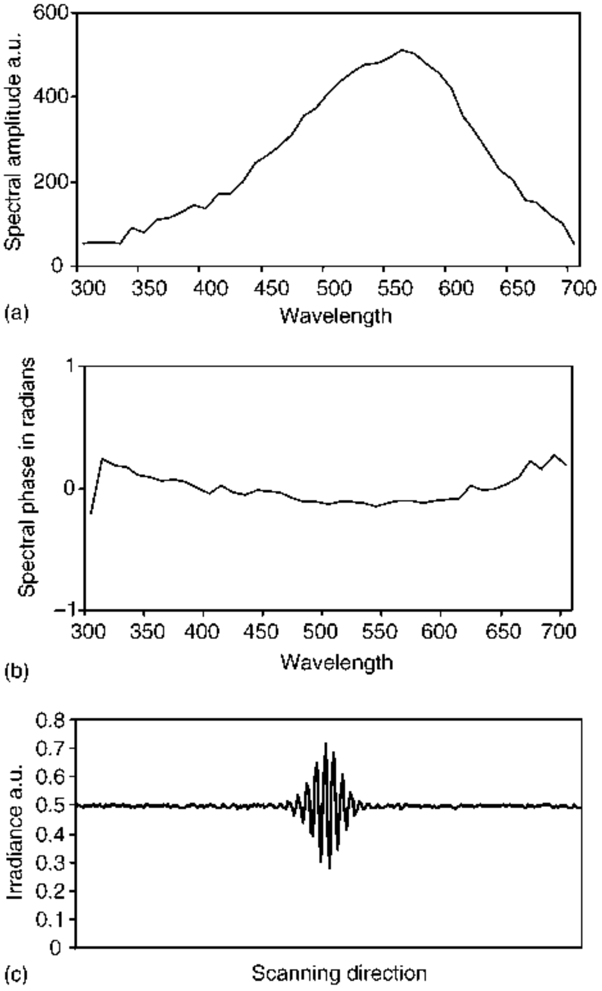

Figure 7 Spectral amplitude (a) and spectral phase (b) for input from unbalanced interferometer (c).

(*see* **Interferometry:** Phase Measurement Interferometry). A geometrical phase shifter can be very useful in polarization microscopes (Figure 9), white light shearing interferometers, or any system where the phase of the white light fringes needs to be measured.

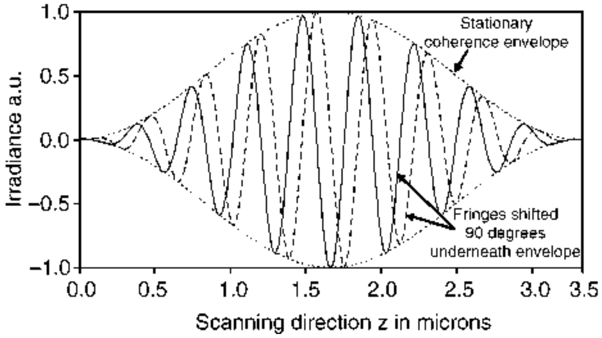

Figure 8 Fringes for two different geometric phase shifts.

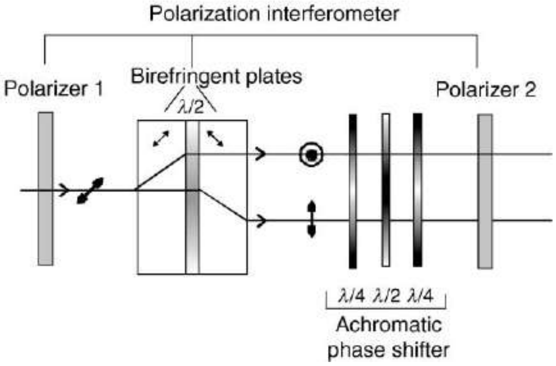

Figure 9 Polarization interferometer with geometric phase shift.

Spectral Interference

White light fringes, because they are made up of fringes of many wavelengths, are observed only over very small path differences. If we filter only a narrow band from the white light spectrum, fringes would be visible over a much larger scan. For different wavelengths we would observe different color and frequency fringes; this is simply the reverse process of the description in the white light interference section.

If we place a spectrometer in the observation plane of the white light interferometer (Figure 10), we will observe fringes with continuously changing wavelengths in dispersed light. These fringes are called fringes of chromatic order or channeled spectrum fringes and find their application in film thickness measurement or absolute distance measurement in the range up to 1 mm. Channeled spectra were used for analysis of dispersion, thin film, and spectroscopic measurements. The number of observed fringes for a given wavelength is directly proportional to the measured optical path difference. The optical path difference can be determined if the difference in fringe numbers is determined for two well-known wavelengths (this is equivalent to two-wavelength interferometry). The optical path difference can be also quickly estimated from the frequency of fringes for a given wavelength; the larger the optical path difference, the denser the fringes.

Surface Topography and Object Structure Measurement

Although WLI has many applications, this section focuses on white light interference as applied to surface topography measurement. Interference microscopes that use white light illumination are often based on the Michelson setup shown in Figure 11.

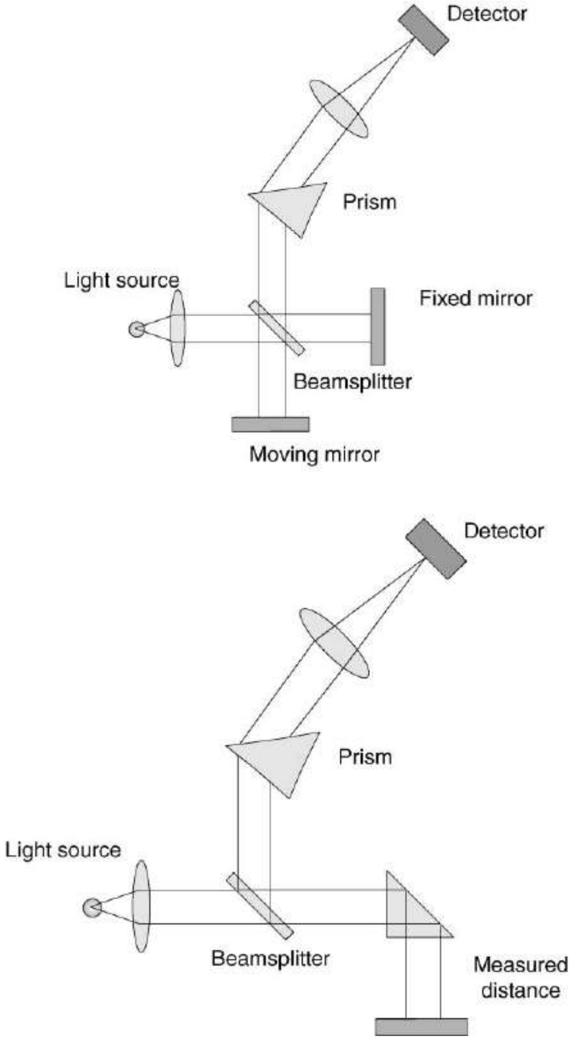

Figure 10 Interferometers with prism spectrometer to observe fringes of equal chromatic order for OPD measurement.

A beamsplitter cube is placed underneath the bright field objective, and one mirror is placed to the side at the focus plane of the objective while the other mirror is replaced by the measured surface with respect to which the whole interference objective is scanned. The narrowly localized fringes at the best focus position for each point of the surface corresponding to a pixel on the CCD camera is observed during the scan. Because of the design of the Michelson objective, in order to accommodate higher magnification objectives constraints, other interference setups like Mirau and Linnik were developed (*see* **Microscopy:** Interference Microscopy). However, these designs are still based on the Michelson interferometer, being also compensated interferometers with equal lengths and amounts of glass in each arm.

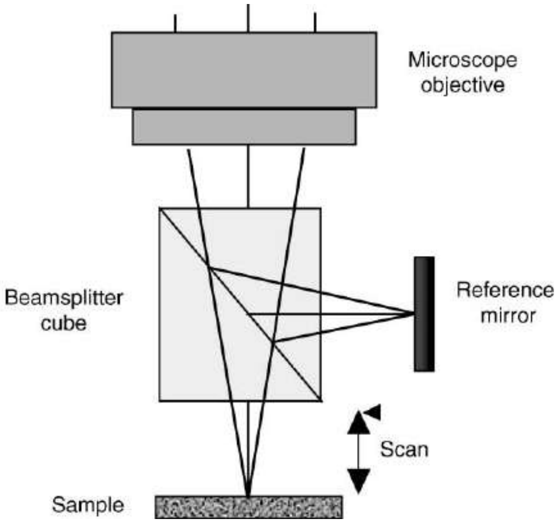

Figure 11 Michelson interferometric objective.

These interference microscopes typically have two complementary modes of operation; one mode uses monochromatic illumination and the other employs a white light source. Monochromatic illumination provides excellent vertical resolution but is limited in its range. It cannot correctly assign fringe order to discontinuous surfaces larger than 160 nanometers. Figure 12a shows monochromatic fringes for the surface of the profile shown in Figure 13. We see that with the monochromatic illumination the height of the grooves remains unknown because it is impossible to assign order numbers to these fringes.

To resolve this ambiguity, white light illumination is employed because it allows for easy identification of the zero order fringe. Figure 12b shows white light fringes created for the same surface. We see the zero-order fringe for the zero OPD as well as the next orders due to the decreasing contrast of the fringes. The position of the zero-order fringe can be followed and approximate heights can be determined visually. Thus, white light illumination permits the measurement of a broader range of surfaces that are rough or have large height variations up to a few millimeters.

The white light interferometer is a great focus detector (Figure 14) at each point of the field of view. The principle behind WLI is finding these individual focus positions using localized fringes observed during the surface scan. WLI interferometers provide high lateral resolution and large vertical range in the order of hundreds of nanometers or micrometers (Figure 15). White light interferometers are commonly used to measure magnetic heads, MEMS devices, binary optic, and machined surfaces.

Signal Processing of White Light Interferograms

For WLI topographic measurements, each pixel registers an interferogram whose position varies

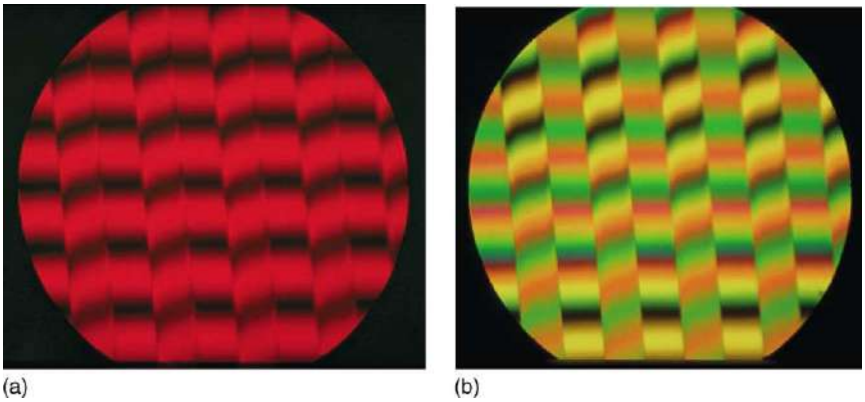

Figure 12 Fringes in quasimonochromatic and white light for object similar to one as presented in Figure 13.

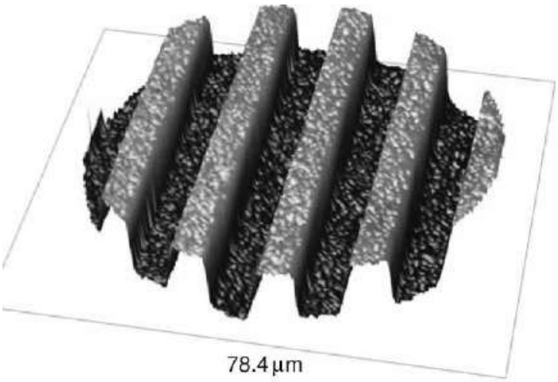

Figure 13 3D object profile – binary grating.

with the surface height. Figure 16 shows two interferograms for two positions in the field of view of a measured step height surface; the first shows the top surface of the step height and the second shows the bottom surface of the step height.

Algorithms have been developed to analyze the signal in different ways. Some focus on finding the fringe envelope position and others examine the position of the fringe underneath the envelope. These algorithms work under the assumption that white light fringes can be described by the same function over the whole field of view of the interferometer. Algorithms are applied to a signal so as to find the position of the coherence envelope for each pixel. This process is called coherence gating, and all algorithms perform coherence gating.

Some algorithms look for the position of the envelope's maximum in a process whereby first the coherence envelope of the fringes is found using signal filtering. Then, the curve is fit to a few points around the envelope's maximum, and finally the position of the envelope's maximum is found. Other algorithms calculate the center of mass of the coherence signal, as described in eqn [8]:

$$H(x,y)=\frac{\sum_{n=1}^{N-1}\left(I_i(x,y)-I_{i+1}(x,y)\right)^2 z_i}{\sum_{n=1}^{N-1}\left(I_i(x,y)-I_{i+1}(x,y)\right)^2} \qquad [8]$$

This method is very fast and computationally efficient. Center of mass calculations are equivalent to calculations of the maximum of the envelope position, but only for a symmetrical signal. For an asymmetrical signal a piston is introduced for each point; however, this piston does not affect the whole measurement. Still another method calculates the spectral phase slope in the Fourier domain to determine the position of the coherence envelope in a process that is equivalent to finding the center of mass. Finally, a different group of algorithms tracks the position of the bright or dark fringe close to the envelope's maximum.

Scanner Speed – Sampling Rate

A practical sampling rate of the white light interference signal is around four pixels per fringe. Because in most algorithms it is not the position of the fringes that we want to determine but rather the position of the envelope of the fringes, sampling the envelope of the fringes with four samples is usually sufficient. However, this sufficiency comes at the expense of lower repeatability (higher noise). The advantage to this sampling rate is short measurement time; higher sampling rates increase measurement time. Measurement speed can be increased by up to 25 times when the envelope of the fringes is widened by filtering light of a bandwidth about 20–40 nm from the white light source.

Scanner Nonlinearity

Scanner motion is assumed to be linear, but nonlinearity in this motion impairs measurement

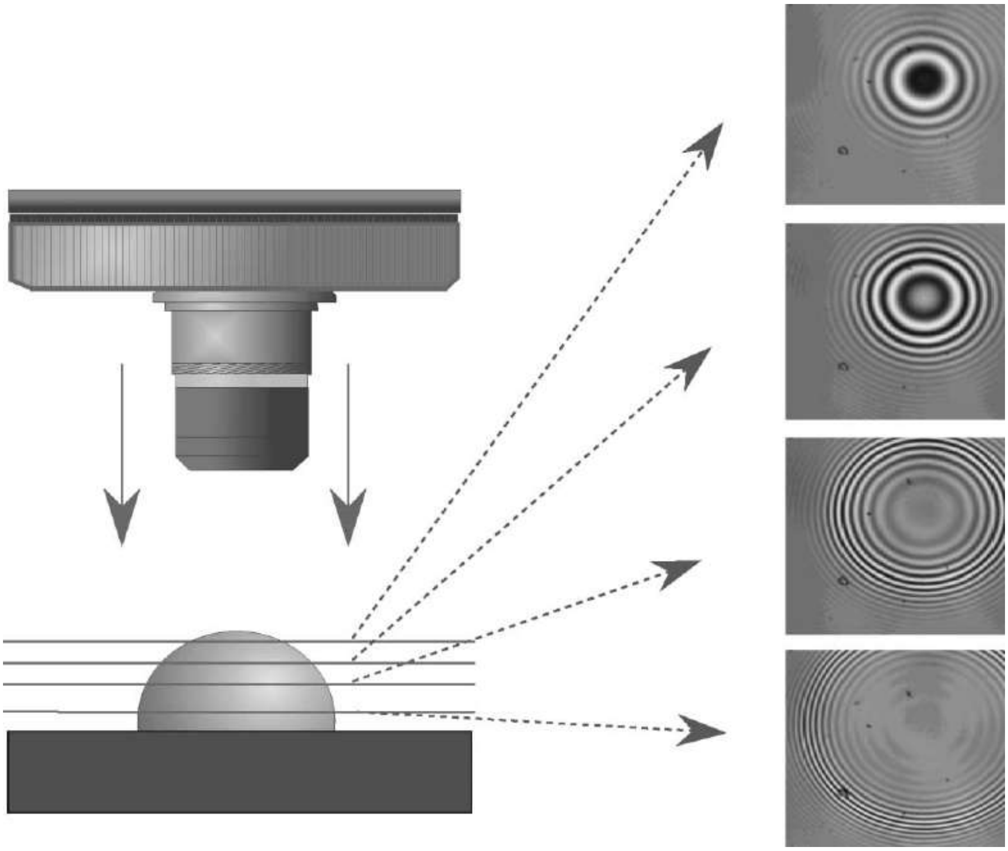

Figure 14 Operation of white light interferometer for surface topography measurement.

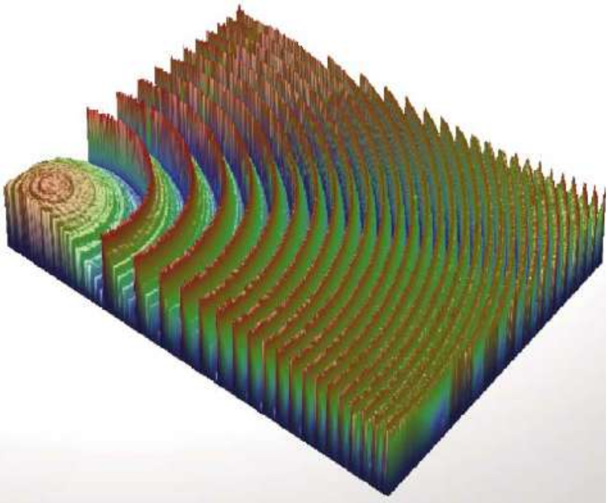

Figure 15 WLI measurement of Fresnel lens in a high-resolution mode. Courtesy of Glimmerglass.

accuracy. To account for this, simple algorithms can calculate the average steps along each interferogram in scanning direction (Figure 17) and measured steps can be used in coherence detection algorithms.

In order to have continuous information about the scan steps, the fringes should be visible somewhere in the field of view at each moment of the scan. This task may require introducing a large tilt for samples with large discontinuities such as stepped surfaces. For large discontinuities where introducing sample tilt may not be sufficient, the envelope of the fringes can be extended by increasing the spatial or temporal coherence of the light source. An alternate solution involves measuring the scanner motion with a distance-measuring interferometer or triggering the camera to collect intensity data at every equal scan step, which can be determined by using the zero-crossing technique, as is commonly done in Fourier transform spectroscopy. Equal scan steps are more suitable for techniques based on transform techniques that assume equal sampling. Other peak detection algorithms, such as center of mass calculations, make use of measured but not necessarily equal steps directly in the algorithm.

Observing the phase along the interferogram can also provide information that is important in optical fiber sensing. These observed changes, which assume that sampling rates of the interferogram are known,

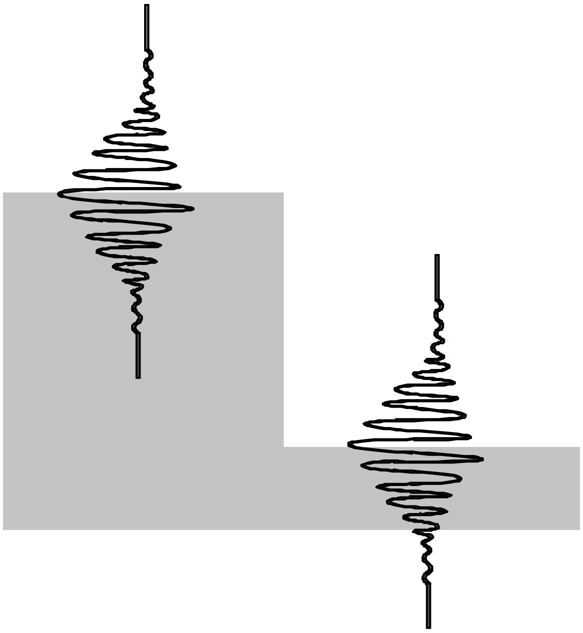

Figure 16 White light fringes as seen by two pixels during scan through focus for object in form of step.

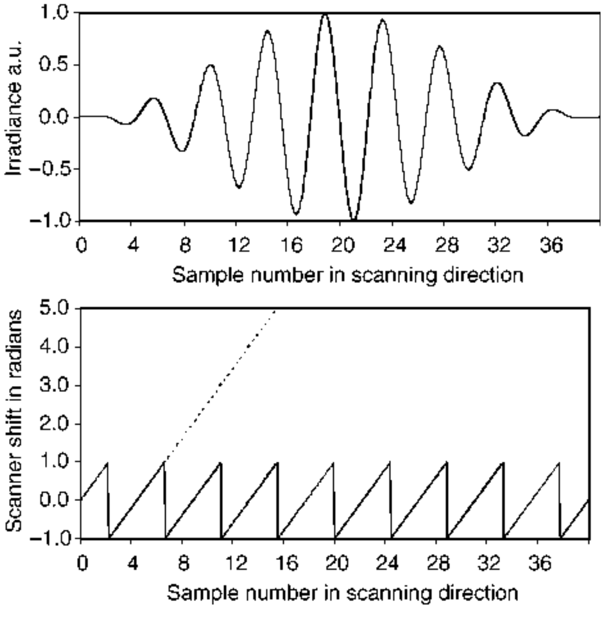

Figure 17 Phase calculated along the white light interferogram. This calculated phase can be used to determine scanner motion.

enable correction for changes in wave number value k_0 which can be due to such things as changes in the working voltage of bulb, an introduced higher-order dispersion, or a large tilt of the sample.

Increased Resolution White Light Interferometry

Interferometric methods that employ a monochromatic light source to detect the phase of the fringes (*see* **Interferometry:** Phase Measurement Interferometry) can achieve about 10 times better (0.3 nanometers versus 3 nanometers) vertical resolution than the WLI methods described so far. Combining coherence position detection from WLI to determine fringe order with the phase detection of phase shifting techniques allows for the measurement of samples with height discontinuities larger than 160 nanometers with the resolution and accuracy of phase shifting interferometry (PSI). This combination is particularly well-suited for determining the shape of smooth surfaces with large height differences such as binary diffractive optics or micro-electromechanical systems (MEMS) (see Figure 15).

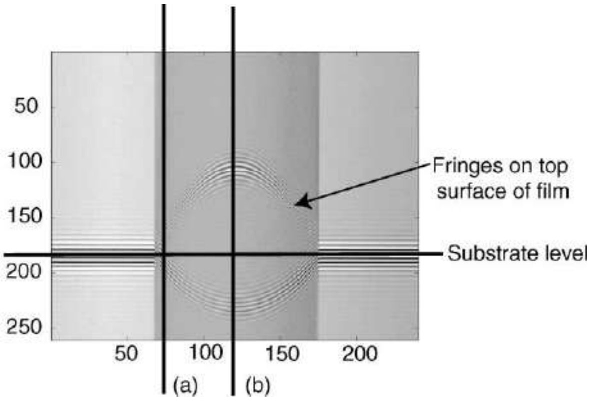

Figure 18 Fringes for (a) thin and (b) thick film.

Using this combined method we obtain both a lower-resolution map of the envelope position and a higher-resolution map of the phase (position) of the zero order fringe. These maps may differ slightly due to effects similar to those discussed in sections *Changes in Envelope and Fringes Due to Reflection of Nondielectric* and *Changes in Envelope and Fringes Due to Dispersion* above. In interference microscopes, shifts in the envelope and fringe position may be introduced by field-dependent and chromatic aberrations of the system and the system's effective numerical aperture. These shifts can vary for different points on the tested surface, but for simplicity are assumed to be constant over the field. Some correction for these effects can be applied.

Film Thickness Measurement

White light interferograms that are obtained in an optical profiler can be used to measure transparent film thicknesses because the character of the interferogram changes with the thickness of the film (Figure 18). Two different techniques, a thin or thick film technique, are used depending of the range of the film thickness. A thick film technique is used if two

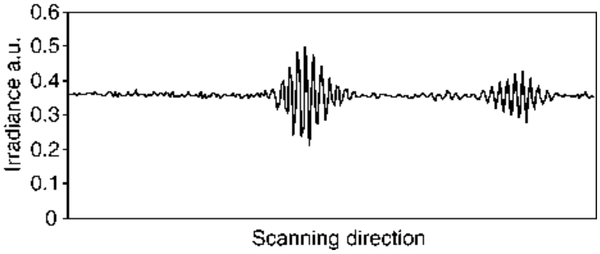

Figure 19 Thick film (see Figure 18b) irradiance.

sets of best contrast white light fringes from each interface are clearly separated, meaning that no interference occurs between the wavefronts reflected from the top and bottom surfaces of the film. A thin film technique is employed when the two sets of fringes overlap.

Thick film measurement

A simple technique for finding the relative position of the peaks of the fringe envelopes can be implemented to find the thickness of a film. Figure 19 shows two clearly separated sets of fringes formed for the air/film and film/substrate interfaces. Whether these two sets of fringes are separated or not depends mainly on the geometrical thickness of the film and its group index of refraction, which is determined by the dispersion of the film. The typical range of measurable film thicknesses runs from 3 to 150 microns, depending on the dispersion of the film. This measurement allows for the detection of flows on the surface and interface of the film. Similar principles based on finding the position of the coherence envelopes are used for distance sensing, thickness measurement of plates, and the cornea of the eye, in low coherence reflectometry and structure measurements in biological samples in optical coherence tomography.

Thin film measurement

For thicknesses of five microns down to tens of nanometers, the white light interferogram is created from the interference between the beams reflected from the reference mirror and the two beams reflected from the thin film layer. Once the interferogram is registered while the objective is scanned vertically with a constant velocity, the spectral phase is calculated by applying a Fourier transform to the measured signal at each pixel as described in the section *Fourier Analysis of White Light Interferograms* above. The phase slope is subtracted and the dispersion of the system needs to be known. The spectral phase for the thin film interference has the form of a polynomial; thus, the polynomial for the chosen film model (n and k) is fitted, and regression analysis is used to find the best fit and, therefore, the film thickness (Figure 20).

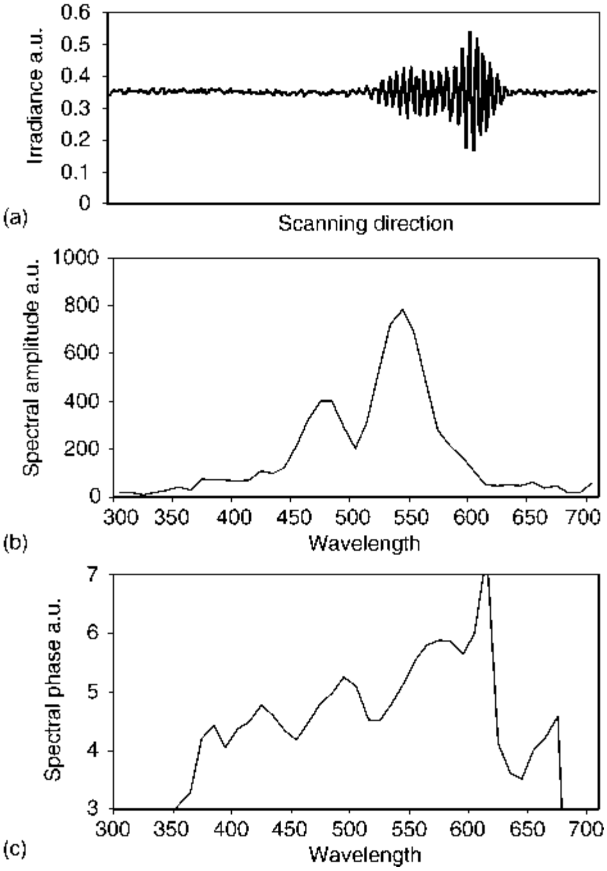

Figure 20 Thin film (see Figure 18a): (a) irradiance, (b) spectral amplitude and (c) spectral phase.

Spatial Coherence Effects in the Interference Microscope

So far we have been discussing temporal coherence effects, but everywhere that the source has a wavelength bandwidth, spatial coherence may play an important role. In an interference microscope for surface topography measurement, the size of the aperture of the condenser is on the order of the aperture of the objective that illuminates the surface with a wide range of angles. For large angles, which are determined by the numerical aperture (NA) of the objective, where NA = 0.5–0.95, the combination of the influence of the spatial and temporal coherence is clearly visible. The additional spatial effects include a reduction of the fringe envelope width and a decrease in fringe spacing corresponding to k_0. Calibration of fringe spacing is typically done on the system and accounts for spatial effects as well as for any uncertainty of k_0 due to the source spectra, the working temperature of the source, the spectral response of the detector and the influence of other factors like the intensity distribution in the illuminating aperture. The spatial influences can be reduced by stopping down the condenser, which causes an increase in the contrast of the fringes.

Rough Surfaces

Rough surfaces are difficult to measure using interferometric techniques, but under certain coherence conditions white light interference can do the job. For rough surfaces, if the microstructure of the object is not resolved by the imaging system, speckles, rather than fringes, are observed. Each speckle has a random phase which is approximately constant in the whole speckle area. If the rough surface is scanned through focus, each individual speckle exhibits the intensity modulation that is typical for WLI. These speckles enable the measurement, but they also introduce noise proportional to the roughness of the measured surface.

Despite the noise that speckles introduce into the WLI measurement, WLI has an advantage because it rejects the light that has undergone scattering outside of a small sample volume, thus allowing precise noninvasive measurement of object structure, even in dense media.

Applications

WLI is used in many disciplines and instruments such as:

- Fourier transform spectroscopy – source and material properties;
- Michelson stellar interferometer – angular size of star, binary stars, delay measurement in optical paths of interferometer;
- Shearing interferometry – structure measurement;
- DIC Nomarski interferometry – structure measurement;
- Speckle interferometry – structure measurement;
- Holography – structure measurement;
- Optical sensors – temperature, pressure, distance;
- Optical coherence tomography – structure measurement.

See also

Coherence: Coherence and Imaging; Overview. **Holography, Techniques:** Holographic Interferometry. **Interferometry:** Phase Measurement Interferometry. **Microscopy:** Interference Microscopy. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Born M and Wolf E (1997) *Principles of Optics*. New York: Cambridge University Press.

Creath K (1997) *Sampling requirements for white light interferometry*. Fringe'97 Proceedings of the 3rd International Workshop on Automatic Processing of Fringe Patterns, pp. 52–59.

Danielson BL and Boisbert CY (1991) Absolute optical ranging using low coherence interferometry. *Applied Optics* 30: 2975–2979.

Davis SP, Abrams MC and Brault JW (2001) *Fourier Transform Spectrometry*. London: Academic Press.

de Groot P and Deck L (1995) Surface profiling by analysis of white light interferograms in the spatial frequency domain. *Journal of Modern Optics* 42: 389–401.

Dresdel T, Hausler G and Venzke H (1992) Three dimensional sensing of rough surfaces by coherence radar. *Applied Optics* 31(7): 919–925.

Harasaki A, Schmit J and Wyant JC (2001) Offset envelope position due to phase change on reflection. *Applied Optics* 40: 2102–2106.

Hariharan P (1992) *Basics of Interferometry*. New York: Academic Press.

Hariharan P, Larkin KG and Roy M (1994) The geometric phase: interferometric observation with white light. *Journal of Modern Optics* 41: 663–667.

Larkin KG (1996) Neat nonlinear algorithm for envelope detection in white light interferometry. *Journal of the Optical Society of America A* 13: 832–842.

Michelson AA and Benoit JR (1895) Détermination expérimentale de la valeur du métre en longueurs d'ondes lumineuses. *Travaux et Mémoires des Bureau International des Poids et Mesures* 11: 1.

Olszak AG and Schmit JG (2003) High stability white light interferometry with reference signal for real time correction of scanning errors. *Optical Engineering* 42(1): 54–59.

Park M-C and Kim S-W (2000) Direct quadratic polynomial fitting for fringe peak detection of white light scanning interferograms. *Optical Engineering* 39: 952–959.

Pavlicek P and Soubusta J (2004) Measurement of the influence of dispersion on white light interferometry. *Applied Optics* 43: 766–770.

Schmit JG and Olszak A (2002) High precision shape measurement by white light interferometry with real time scanner error correction. *Applied Optics* 41(28): 5943–5950.

Sheppard CJR and Larkin KG (1995) Effects of numerical aperture on interference fringe spacing. *Applied Optics* 34: 4731–4734.

Sheppard CJR and Roy M (2003) Low-coherence interference microscopy. In: Torok P and Kao F-J (eds) *Optical Imaging and Microscopy*, pp. 257–273. Berlin, Heidelberg, New York: Springer-Verlag.

Steel W (1987) *Interferometry*. New York: Cambridge University Press.

Turyshew SG (2003) Analytical modeling of the white light fringe. *Applied Optics* 42(1): 71–90.

L

LASERS

Contents

Carbon Dioxide Laser
Dye Lasers
Edge Emitters
Excimer Lasers
Free Electron Lasers
Metal Vapor Lasers
Noble Gas Ion Lasers
Optical Fiber Lasers
Organic Semiconductors and Polymers
Planar Waveguide Lasers
Semiconductor Lasers
Up-Conversion Lasers

Carbon Dioxide Laser

C R Chatwin, University of Sussex, Brighton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article gives a brief history of the development of the laser and goes on to describe the characteristics of the carbon dioxide laser and the molecular dynamics that permit it to operate at comparatively high power and efficiency. It is these commercially attractive features and its low cost that has led to its adoption as one of most popular industrial power beams. This outline also describes the main types of carbon dioxide laser and briefly discusses their characteristics and uses.

Brief History

In 1917 Albert Einstein developed the concept of stimulated emission which is the phenomenon used in lasers. In 1954 the MASER (Microwave Amplification by Stimulated Emission of Radiation) was the first device to use stimulated emission. In that year Townes and Schawlow suggested that stimulated emission could be used in the infrared and optical portions of the electromagnetic spectrum. The device was originally termed the optical maser, this term being dropped in favor of LASER, standing for: Light Amplification by Stimulated Emission of Radiation. Working against the wishes of his manager at Hughes Research Laboratories, the electrical engineer Ted Maiman created the first laser on the 15 May 1960. Maiman's flash lamp pumped ruby laser produced pulsed red electromagnetic radiation at a wavelength of 694.3 nm. During the most active period of laser systems discovery Bell Labs made a very significant contribution. In 1960, Ali Javan, William Bennet and Donald Herriot produced the first Helium Neon laser, which was the first continuous wave (CW) laser operating at 1.15 μm. In 1961, Boyle and Nelson developed a continuously operating Ruby laser and in 1962, Kumar Patel, Faust, McFarlane and Bennet discovered five noble gas lasers and lasers using oxygen mixtures. In 1964, C.K.N. Patel created the high-power carbon dioxide laser operating at 10.6 μm. In 1964, J.F. Geusic and R.G. Smith produced the first Nd:Yag laser using neodymium doped yttrium aluminum garnet crystals and operating at 1.06 μm.

Characteristics

Due to its operation between low lying vibrational energy states of the CO_2 molecule, the CO_2 laser has a high quantum efficiency, ~40%, which makes it extremely attractive as a high-power industrial materials processing laser (1 to 20 kW), where energy and running costs are a major consideration. Due to the requirement for cooling to retain the population inversion, the efficiency of electrical pumping and optical losses – commercial systems have an overall efficiency of approximately 10%. Whilst this may seem low, for lasers this is still a high efficiency. The CO_2 laser is widely used in other fields, for example, surgical applications, remote sensing, and measurement. It emits infrared radiation with a wavelength that can range from 9 μm up to 11 μm. The laser transition may occur on one of two transitions: $(00^01) \rightarrow (10^00)$, $\lambda = 10.6\ \mu m$; $(00^01) \rightarrow (02^00)$, $\lambda = 9.6\ \mu m$, see Figure 1. The 10.6 μm transition has the maximum probability of oscillation and gives the strongest output; hence, this is the usual wavelength of operation, although for specialist applications the laser can be forced to operate on the 9.6 μm line.

Figure 1 illustrates an energy level diagram with four vibrational energy groupings that include all the significantly populated energy levels. The internal relaxation rates within these groups are considered to be infinitely fast when compared with the rate of energy transfer between these groups. In reality the internal relaxation rates are at least an order of magnitude greater than the rates between groups.

Excitation of the upper laser level is usually provided by an electrical glow discharge. However, gas dynamic lasers have been built where expanding a hot gas through a supersonic nozzle creates the population inversion; this creates a nonequilibrium region in the downstream gas stream with a large population inversion, which produces a very high-power output beam (135 kW – Avco Everett Research Lab). For some time the gas dynamic laser was seriously considered for use in the space-based Strategic Defence Initiative (SDI-USA). The gas mixture used in a CO_2 laser is usually a mixture of carbon dioxide, nitrogen, and helium. The proportions of these gases varies from one laser system to another, however, a typical mixture is 10%-CO_2;

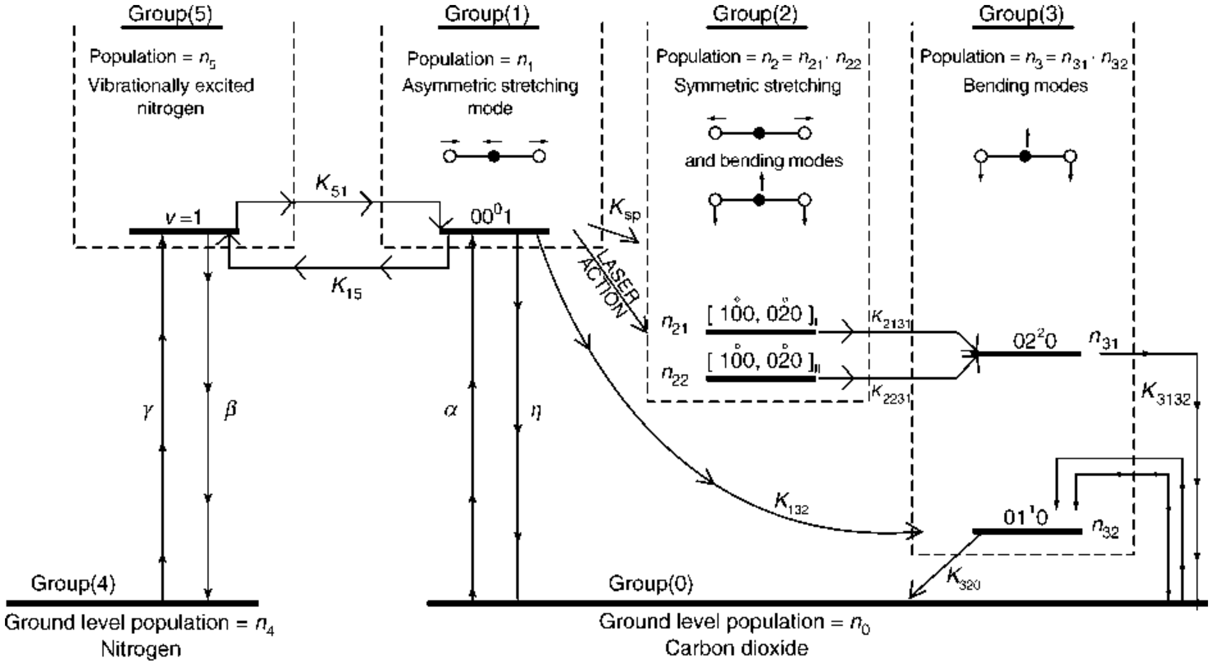
Group(5)
Population = n_5
Vibrationally excited nitrogen
Group(1)
Population = n_1
Asymmetric stretching mode
Group(2)
Population = $n_2 = n_{21} \cdot n_{22}$
Symmetric stretching
and bending modes
Group(3)
Population = $n_3 = n_{31} \cdot n_{32}$
Bending modes
K_{51}
$v = 1$
00^01
K_{sp}
K_{15}
LASER ACTION
n_{21}
$[100, 020]_I$
n_{22}
$[100, 020]_{II}$
K_{2131}
K_{2231}
02^20
n_{31}
K_{3132}
γ
β
α
η
K_{132}
01^10
n_{32}
K_{320}
Group(4)
Ground level population = n_4
Nitrogen
Group(0)
Ground level population = n_0
Carbon dioxide

Figure 1 Six level model used for the theoretical description of CO_2 laser action.

10%-N_2; 80%-He. Helium plays a vital role in the operation of the CO_2 laser in that it maintains the population inversion by depopulating the lower laser level by nonradiative collision processes. Helium is also important for stabilization of the gas discharge; furthermore it greatly improves the thermal conductivity of the gas mixture, which assists in the removal of waste heat via heat exchangers.

Small quantities of other gases are often added to commercial systems in order to optimize particular performance characteristics or stabilize the gas discharge; for brevity we only concern ourselves here with this simple gas mixture.

Molecular Dynamics

Direct Excitation and De-excitation

It is usual for excitation to be provided by an electrical glow discharge. The direct excitation of carbon dioxide (CO_2) and nitrogen (N_2) ground state molecules proceeds via inelastic collisions with fast electrons. The rates of kinetic energy transfer are α and γ, respectively, and are given by eqns [1] and [2]:

$$\alpha = \frac{F_{\mathrm{CO}_2} \times \mathrm{IP}}{E_{00^01} \times n_0} (\mathrm{sec}^{-1}) \qquad [1]$$

$$\gamma = \frac{F_{\mathrm{N}_2} \times \mathrm{IP}}{E_{v=1} \times n_4} (\mathrm{sec}^{-1}) \qquad [2]$$

where:

F_{CO_2} = Fraction of the input power (IP) coupled into the excitation of the energy level E_{00^01}, n_0 is the CO_2 ground level population density;
F_{N_2} = Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$; and n_4 is the N_2 ground level population density.

The reverse process of the above occurs when molecules lose energy to the electrons and the electrons gain an equal amount of kinetic energy; the direct de-excitation rates are given by η and β, eqns [3] and [4], respectively:

$$\eta = \alpha \times \exp\left(\frac{E_{00^01}}{E_\mathrm{e}}\right)(\mathrm{sec}^{-1}) \qquad [3]$$

$$\beta = \gamma \times \exp\left(\frac{E_{v=1}}{E_\mathrm{e}}\right)(\mathrm{sec}^{-1}) \qquad [4]$$

where E_e is the average electron energy in the discharge.

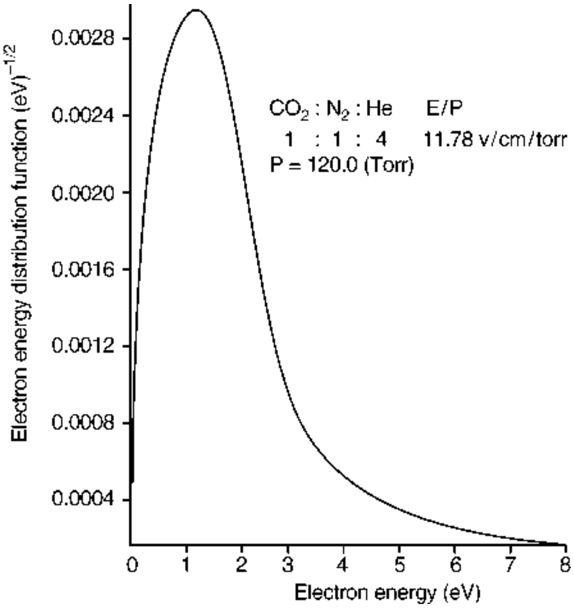

Figure 2 Electron energy distribution function.

E_e, F_{CO_2}, and F_{N_2} are obtained by solution of the Boltzmann transport equation (BTE); the average electron energy can be optimized to maximize the efficiency (F_{CO_2}, F_{N_2}) with which electrical energy is utilized to create a population inversion. Hence, the discharge conditions required to maximize efficiency can be predicted from the transport equation. Figure 2 shows one solution of the BTE for the electron energy distribution function.

Resonant Energy Transfer

Resonant energy transfer between the CO_2 (00^01) and N_2 ($v = 2$) energy levels (denoted 1 and 5 in Figure 1) proceeds via excited molecules colliding with ground state molecules. A large percentage of the excitation of the upper laser level takes place via collisions between excited N_2 molecules and ground state CO_2 molecules. The generally accepted rate of this energy transfer is given by eqns [5] and [6]:

$$K_{51} = 19\,000\, P_{\mathrm{CO}_2}\ (\mathrm{sec}^{-1}) \qquad [5]$$

$$K_{15} = 19\,000\, P_{\mathrm{N}_2}\ (\mathrm{sec}^{-1}) \qquad [6]$$

where P_{CO_2} and P_{N_2} are the respective gas partial pressures in Torr.

Hence, CO_2 molecules are excited into the upper laser level by both electron impact and impact with excited N_2 molecules. The contribution from N_2 molecules can be greater than 40% depending on the discharge conditions.

Collision–Induced Vibrational Relaxation of the Upper and Lower Laser Levels

The important vibrational relaxation processes are illustrated by Figure 1 and can be evaluated from eqns [7–10]; where the subscripts refer to the rate between energy levels 1 and 32; 21 and 31; 22 and 31; 32 and 0, respectively:

$$K_{132} = 367\, P_{CO_2} + 110\, P_{N_2} + 67\, P_{He}\ (\mathrm{sec}^{-1}) \qquad [7]$$

$$K_{2131} = 6 \times 10^5\, P_{CO_2}\ (\mathrm{sec}^{-1}) \qquad [8]$$

$$K_{2231} = 5.15 \times 10^5\, P_{CO_2}\ (\mathrm{sec}^{-1}) \qquad [9]$$

$$K_{320} = 200\, P_{CO_2} + 215\, P_{N_2} + 3270\, P_{He}\ (\mathrm{sec}^{-1}) \qquad [10]$$

K_{132}, K_{2131}, and K_{2231} are vibration/vibration transfer rates and K_{320} is a vibration/translation transfer rate. Note the important effect of helium on eqn [10]; helium plays a major role in depopulating the lower laser level, thus enhancing the population inversion. P_{He} is the partial pressure of helium in Torr.

Radiative Relaxation

Spontaneous radiative decay is not a major relaxation process in the CO_2 laser but it is responsible for starting laser action via spontaneous emission. The Einstein '*A*' coefficient for the laser transition is given by eqn [11]:

$$A = (\tau_{sp})^{-1} = 0.213\ (\mathrm{sec}^{-1}) \qquad [11]$$

Gain

The gain (g) is evaluated from the product of the absorption coefficient (σ) and the population inversion, eqn [12]:

$$g = \sigma\left(n_{00^01} - \frac{g_1}{g_2} n_{10^00}\right)\ \mathrm{cm}^{-1} \qquad [12]$$

For most commercial laser systems the absorption coefficient is that for high-pressure collision-broadening ($P > 5.2$ Torr) where the intensity distribution function describing the line shape is Lorentzian. The following expression describes the absorption coefficient, eqn [13]:

$$\sigma = \frac{692.5}{T n_{CO_2}\left(1 + 1.603\frac{n_{N_2}}{n_{CO_2}} + 1.4846\frac{n_{He}}{n_{CO_2}}\right)}(\mathrm{cm}^2) \qquad [13]$$

where T is the absolute temperature and n refers to the population density of the gas designated by the subscript.

This expression takes account of the different constituent molecular velocity distributions and different collision cross-sections for $CO_2 \rightarrow CO_2$, $N_2 \rightarrow CO_2$ and $He \rightarrow CO_2$ type collisions. Equation [13] also takes account of the significant line broadening effect of helium.

Neglecting the unit change in rotational quantum number, the energy level degeneracies g_1 and g_2 may be dropped. n_{10^00} is partitioned such that $n_{10^00} = 0.1452 n_2$ and eqn [12] can be re-cast as eqn [14]:

$$g = \sigma(n_1 - 0.1452 n_2)\ \mathrm{cm}^{-1} \qquad [14]$$

where n_1 and n_2 are the population densities of energy groups '1' and '2' respectively.

Stimulated Emission

Consider a laser oscillator with two plane mirrors, one placed at either end of the active gain medium, with one mirror partially transmitting (see Figure 4). Laser action is initiated by spontaneous emission that happens to produce radiation whose direction is normal to the end mirrors and falls within the resonant modes of the optical resonator. The rate of change of photon population density (I_p) within the laser cavity can be written as eqn [15]:

$$\frac{dI_P}{dt} = I_p c g - \frac{I_p}{T_0} \qquad [15]$$

where the first term on the right-hand side accounts for the effect of stimulated emission and the second term represents the number of photons that decay out of the laser cavity, T_0 is the photon decay time, given by eqn [16], and is defined as the average time a photon remains inside the laser cavity before being lost either through the laser output window or due to dispersion; if dispersion is ignored, I_p/T_0, is the laser output:

$$T_0 = \frac{2L}{c \log_e\left(\frac{1}{R_B R_F}\right)} \qquad [16]$$

where L is the distance between the back and the front mirrors, which have reflectivities of R_B and R_F, respectively. The dominant laser emission occurs on a rotational–vibrational P branch transition $P(22)$, that is $(J = 21) \rightarrow (J = 22)$ line of the $(00^01) \rightarrow (10^00)$, $\lambda = 10.6\ \mu m$ transition, where J is the rotational quantum number. The rotational level relaxation rate is so rapid that equilibrium is maintained between rotational levels so that they feed all their energy through the $P(22)$ transition. This model simply assumes constant intensity, basing laser performance on the performance of an average unit volume. By introducing the stimulated emission term into the molecular rate equations, which describe the rate of transfer of molecules between the various energy levels illustrated in Figure 1, a set of molecular rate eqns [17–21] can be written that permit simulation of the performance of a carbon dioxide laser:

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = \alpha n_0 - \eta n_1 + K_{51}n_5 - K_{15}n_1 - K_{\mathrm{sp}}n_1 - K_{132}\left(n_1 - \left[\frac{n_1}{n_{32}}\right]_{\mathrm{e}} n_{32}\right) - I_{\mathrm{p}}cg \qquad [17]$$

$$\frac{\mathrm{d}n_2}{\mathrm{d}t} = K_{\mathrm{sp}}n_1 + I_{\mathrm{p}}cg - K_{2131}\left(n_{21} - \left[\frac{n_{21}}{n_{31}}\right]_{\mathrm{e}} n_{31}\right) - K_{2231}\left(n_{22} - \left[\frac{n_{22}}{n_{31}}\right]_{\mathrm{e}} n_{31}\right) \qquad [18]$$

$$\frac{\mathrm{d}n_3}{\mathrm{d}t} = 2K_{2131}\left(n_{21} - \left[\frac{n_{21}}{n_{31}}\right]_{\mathrm{e}} n_{31}\right) - 2K_{2231}\left(n_{22} - \left[\frac{n_{22}}{n_{31}}\right]_{\mathrm{e}} n_{31}\right) + K_{132}\left(n_1 - \left[\frac{n_1}{n_{32}}\right]_{\mathrm{e}} n_{32}\right) - K_{320}\left(n_{32} - \left[\frac{n_{32}}{n_0}\right]_{\mathrm{e}} n_0\right) \qquad [19]$$

$$\frac{\mathrm{d}n_5}{\mathrm{d}t} = \gamma n_4 - \beta n_5 - K_{51}n_5 + K_{15}n_1 \qquad [20]$$

$$\frac{\mathrm{d}I_{\mathrm{p}}}{\mathrm{d}t} = I_{\mathrm{p}}cg - \frac{I_{\mathrm{p}}}{T_0} \qquad [21]$$

The terms in square brackets ensure that the system maintains thermodynamic equilibrium; subscript 'e' refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium. The set of five simultaneous differential equations can be solved using a Runge–Kutta method. They can provide valuable performance prediction data that is helpful in optimizing laser design, especially when operated in the pulsed mode. Figure 3a illustrates some simulation results for the transverse flow laser shown in Figure 9. The results illustrate the effect of altering the gas mixture and how this can be used to control the gain switched spike that would result in unwelcome work piece plasma generation if allowed to become too large. Figure 3b shows an experimental laser output pulse from the high pulse repetition frequency (prf–5kHz) laser illustrated in Figure 9. This illustrates that even a quite basic physical model can give a good prediction of laser output performance.

Optical Resonator

Figure 4 shows a simple schematic of an optical resonator. This simple optical system consists of two mirrors (the full reflector is often a water cooled gold coated copper mirror), which are aligned to be orthogonal to the optical axis that runs centrally along the length of the active gain medium in which there is a population inversion. The output coupler is a partial reflector (usually dielectrically coated zinc selenide – ZnSe–that may be edge cooled) so that some of the electromagnetic radiation can escape as an output beam. The ZnSe output coupler has a natural reflectivity of about 17% at each air–solid interface. For high power lasers (2 kW) 17% is sufficient for laser operation; however, depending on the required laser performance the inside surface is often given a reflective coating. The reflectivity of the inside face depends on the balance between the gain (eqn [14]), the output power and the power stability requirements. The outside face of the partial reflector must be anti-reflection (AR) coated.

Spontaneous emission occurs within the active gain medium and radiates randomly in all directions; a fraction of this spontaneous emission will be in the same direction as the optical axis, perpendicular to the end mirrors, and will also fall into a resonant mode of the optical resonator. Spontaneous emission photons interact with CO_2 molecules in the excited upper laser level, excited state (00^01), which stimulates these molecules to give up a quanta of vibrational energy as photons via the radiative transition $(00^01) \rightarrow (10^00)$, $\lambda = 10.6\,\mu\mathrm{m}$. The radiation given up will have exactly the same phase and direction as the stimulating radiation and thus will be coherent with it. The reverse process of absorption also occurs, but so long as there is a population inversion there will be a net positive output. This process is called light amplification by stimulated emission of radiation (LASER). The mirrors continue to redirect the photons parallel to the optical axis and so long as the population inversion is not depleted, more and more photons are stimulated by stimulated emission which dominates the process and also dominates the spontaneous emission, which is important to initiate laser action.

Light emitted by lasers contains several optical frequencies, which are a function of the different modes of the optical resonator; these are simply the standing wave patterns that can exist within the resonator structure. There are two types of

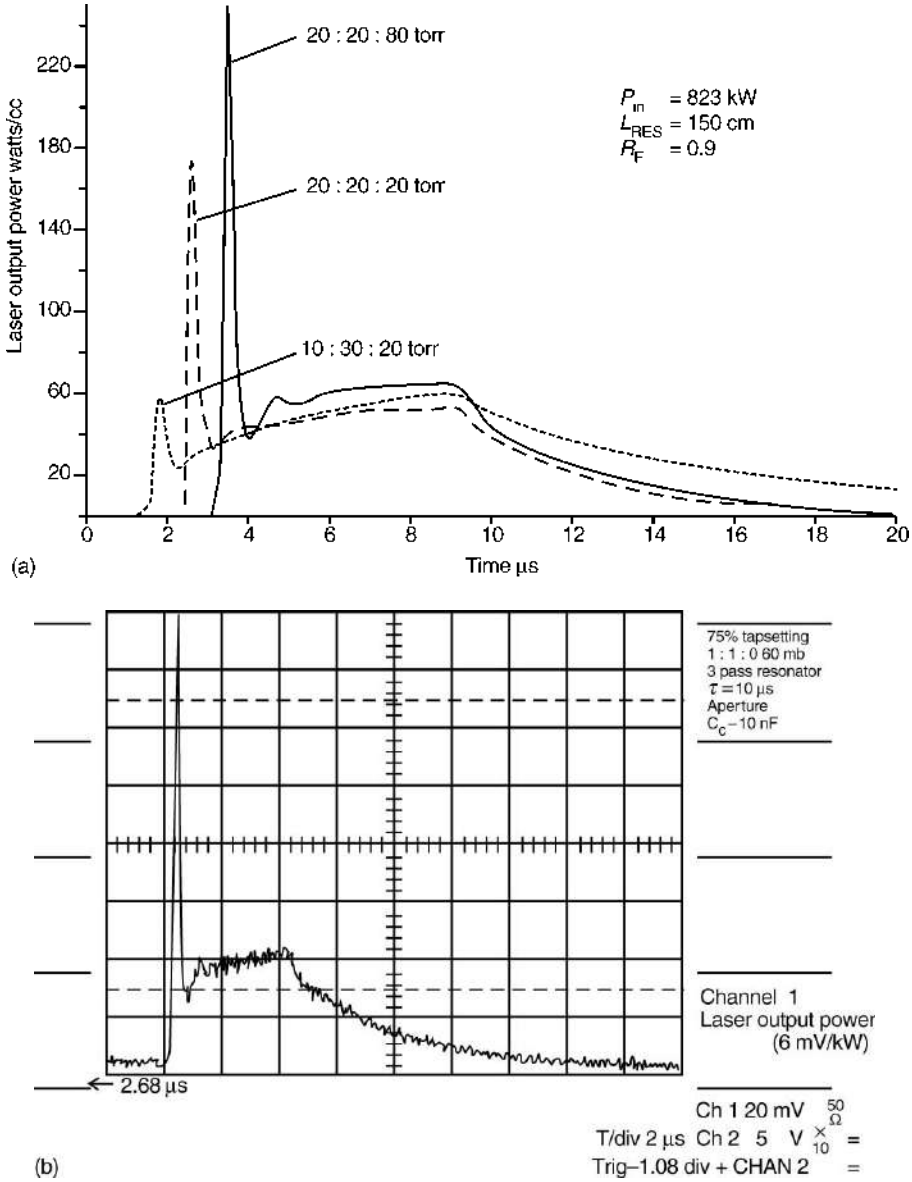

Figure 3 (a) Predicted output pulses for transverse flow CO_2 laser for different gas mixtures, (b) Experimental output pulse from transverse flow CO_2 laser.

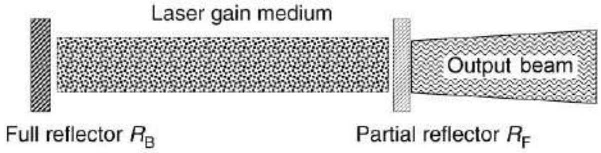

Figure 4 Optical resonator.

resonator modes: longitudinal and transverse. Longitudinal modes differ from one another in their frequency of oscillation whereas transverse modes differ from one another in their oscillation frequency and field distribution in a plane orthogonal to the direction of propagation. Typically CO_2 lasers have a large number of longitudinal modes; in CO_2 laser applications these are of less interest than the transverse modes, which determine the transverse beam intensity and the nature of the beam when focused. In cylindrical coordinates the transverse modes are labelled TEM_{pl}, where subscript '*p*' is the number of radial nodes and '*l*' is the number of angular nodes. The lowest order mode is the TEM_{00}, which has a Gaussian-like intensity profile with its maximum on the beam axis. A light beam emitted from an optical resonator with a Gaussian profile is said to be operating in the 'fundamental mode' or the TEM_{00} mode. The decrease in irradiance (I) with distance '*r*' from the axis (I_0) of the Gaussian beam is described by eqn [22]:

$$I(r) = I_0 \exp(2r^2/w^2) \qquad [22]$$

where w is the radial distance, where the power density is decreased to $1/e^2$ of its axial value. Ideally a commercial laser should be capable of operation in the fundamental mode as, with few exceptions, this results in the best performance in applications. Laser cutting benefits from operation in the fundamental mode; however, welding or heat treatment applications may benefit from operation with higher-order modes. Output beams are usually controlled to be linearly or circularly polarized, depending upon the requirements of the application. For materials processing applications the laser beam is usually focused via a water cooled ZnSe lens or, for very high power lasers, a parabolic gold coated mirror. Welding applications will generally use a long focal length lens and cutting applications will use a short focal length, which generates a higher irradiance at the work piece than that necessary for welding. The beam delivery optics are usually incorporated into a nozzle assembly that can deliver cooling water and assist gases for cutting and anti-oxidizing shroud gases for welding or surface engineering applications.

Laser Configuration

CO_2 lasers are available in many different configurations and tend to be classified on the basis of their physical form and the gas flow arrangement, both of which greatly affect the output power available and beam quality. The main categories are: sealed off lasers, waveguide lasers, slow axial flow, fast axial flow, diffusion cooled, transverse flow, transversely excited atmospheric lasers, and gas dynamic lasers.

Sealed-Off and Waveguide Lasers

Depopulation of the lower laser level is via collision with the walls of the discharge tube, so the attainable output power scales with the length of the discharge column and not its diameter. Output powers are in the range 5 W to 250 W. Devices may be constructed from concentric glass tubes with the inner tube providing the discharge cavity and the outer tube acting to contain water-cooling of the inner discharge tube. The inner tube walls act as a heat sink for the discharge thermal energy (see Figure 5). The DC electrical discharge is provided between a cathode and anode situated at either end of the discharge tube. A catalyst must be provided to ensure regeneration of CO_2 from CO. This may be accomplished by adding about 1% of H_2O to the gas mixture, or alternatively, recombination can be achieved via a hot (300 °C) Ni cathode, which acts as a catalyst. RF-excited all metal sealed-off tube systems can deliver lifetimes greater than 45 000 hours.

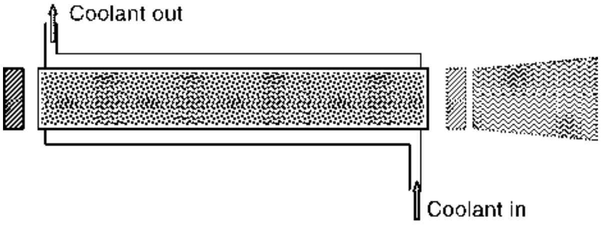

Figure 5 Schematic of sealed-off CO_2 laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

Diffusion-cooled slab laser technology will also deliver reliable sealed operation for 20 000 hrs. Excitation of the laser medium occurs via RF excitation between two water-cooled electrodes. The water-cooled electrodes dissipate (diffusion cooled) the heat generated in the gas discharge. An unstable optical resonator provides the output coupling for such a device (see Figure 6). Output powers are in the range 5 W to 300 W and can be pulsed from 0 to 100 kHz. These lasers are widely used for marking, rapid prototyping, and cutting of nonmetals (paper, glass, plastics, ceramics) and metals.

Waveguide CO_2 lasers use small bore tubes (2–4 mm) made of BeO or SiO_2 where the laser radiation is guided by the tube walls. Due to the small tube diameter, a gas total pressure of 100 to 200 Torr is necessary, hence the gain per unit length is high. This type of laser will deliver 30 W of output power from a relatively short (50 cm long) compact sealed-off design; such a system is useful for microsurgery and scientific applications. Excitation can be provided from a longitudinal DC discharge or from an RF source that is transverse to the optical axis; RF excitation avoids the requirement for an anode and cathode and results in a much lower electrode voltage.

Slow Axial Flow Lasers

In slow flow lasers the gas mixture flows slowly through the laser cavity. This is done to remove the products of dissociation that will reduce laser efficiency or prevent it from operating at all, and the main contaminant is CO. The dissociated gases (mainly CO and O_2) can be recombined using a catalyst pack and then reused via continuous recirculation. Heat is removed via diffusion through the walls of the tube containing the active gain medium. The tube is frequently made of Pyrex glass with a concentric outer tube to facilitate water-cooling of the

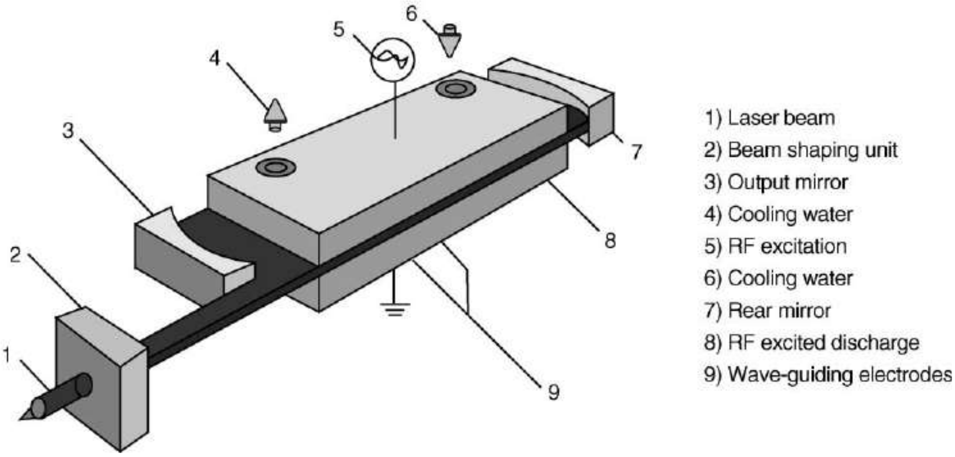

Figure 6 Schematic for sealed-off slab laser and diffusion cooled laser with RF excitation (courtesy of Rofin).

laser cavity (see Figure 7). Slow flow lasers operate in the power range 100 W to 1500 W, and tend to use a longitudinal DC electrical discharge which can be made to run continuously or pulsed if a thyratron switch is build into the power supply; alternatively, electrical power can be supplied via transverse RF excitation. The power scales with length, hence high power slow flow lasers have long cavities and require multiple cavity folds in order to reduce their physical size.

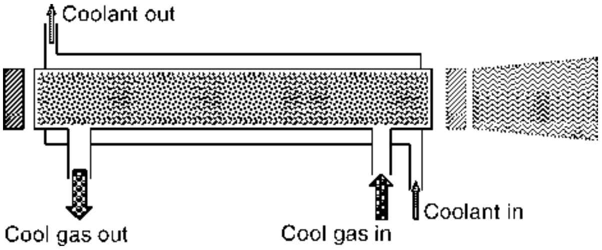

Figure 7 Slow flow CO$_2$ laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

Fast Axial Flow Lasers

The fast axial flow laser, Figure 8, can provide output powers from 1 kW to 20 kW; it is this configuration that dominates the use of CO$_2$ lasers for industrial applications. Industrial lasers are usually in the power range 2–4 kW. The output power from these devices scales with mass flow, hence the gas mixture is recycled through the laser discharge region at sonic or supersonic velocities. Historically this was achieved using Rootes blowers to compress the gas upstream of the laser cavity. Rootes compressors are inherently inefficient and the more advanced laser systems utilize turbine compressors, which deliver greater efficiency and better laser stability. Rootes compressors can be a major source of vibration. With this arrangement heat exchangers are required to remove heat after the laser discharge region and also after the compressor stage, as the compression process heats up the laser gases. Catalyst packs are used to regenerate gases but some gas replacement is often required. These laser systems have short cavities and use folded stable resonator designs to achieve higher output powers with extremely high-quality beams that are particularly suitable for cutting applications. They also give excellent results when used for welding and surface treatments.

Fast axial flow lasers can be excited by a longitudinal DC discharge or a transverse RF discharge. Both types of electrical excitation are common. For materials processing applications it is often important to be able to run a laser in continuous wave (CW) mode or as a high pulse repetition rate (prf) pulsed laser and to be able to switch between CW and pulsed in real time; for instance, laser cutting of accurate internal corners is difficult using CW operation but very easy using the pulsed mode of operation. Both methods of discharge excitation can provide this facility.

Diffusion Cooled Laser

The diffusion-cooled slab laser is RF excited and gives an extremely compact design capable of delivering 4.5 kW pulsed from 8 Hz to 5 kHz prf or CW with good beam quality (see Figure 6). The optical resonator is formed by the front and rear mirrors and two parallel water cooled RF-electrodes. Diffusion cooling is provided by the RF-electrodes, removing the requirement for conventional gas recirculation via Rootes blowers or turbines.

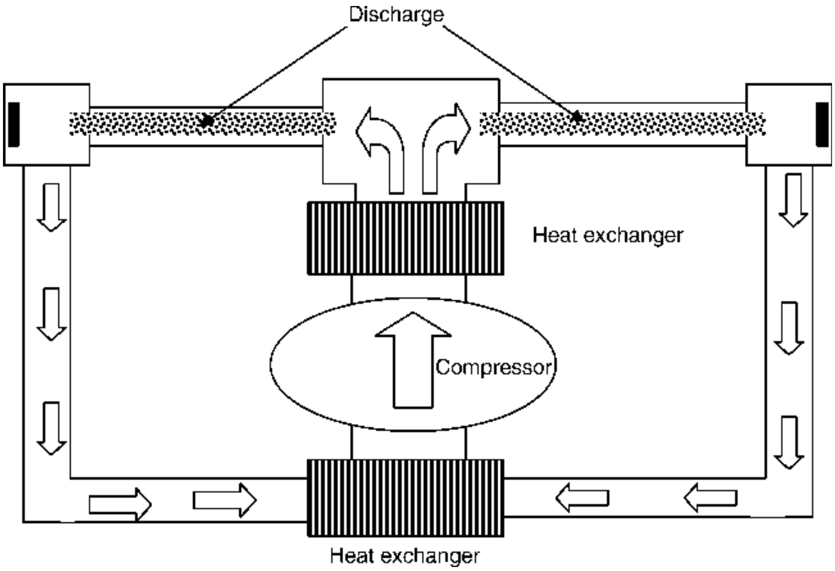

Figure 8 Fast axial flow carbon dioxide laser.

This design of laser results in a device with an extremely small footprint that has low maintenance and running costs. Applications include: cutting, welding, and surface engineering.

Fast Transverse Flow Laser

In the fast transverse flow laser (Figure 9a) the gas flow, electrical discharge, and the output beam are at right angles to each other (Figure 9b). The transverse discharge can be high voltage DC, RF, or pulsed up to 8 kHz (Figure 9c). Very high output power per unit discharge length can be obtained with an optimal total pressure (*P*) of ~100 Torr; systems are available delivering 10 kW of output power, CW or pulsed (see Figures 3a and b). The increase in total pressure requires a corresponding increase in the gas discharge electric field, *E*, as the ratio *E*/*P* must remain constant, since this determines the temperature of the discharge electrons, which have an optimum mean value (optimum energy distribution, Figure 2) for efficient excitation of the population inversion. With this high value of electric field, a longitudinal-discharge arrangement is impractical (500 kV for a 1 m discharge length); hence, the discharge is applied perpendicular to the optical axis. Fast transverse flow gas lasers provided the first multikilowatt outputs but tend to be expensive to maintain and operate. In order to obtain a reasonable beam quality, the output coupling is often obtained using a multipass unstable resonator. As the population inversion is available over a wide rectangular cross-section, this is a disadvantage of this arrangement and beam quality is not as good as that obtainable from fast axial flow designs. For this reason this type of laser is suitable for a wide range of welding and surface treatment applications.

Transversely Excited Atmospheric (TEA) Pressure

If the gas total pressure is increased above ~100 Torr it is difficult to sustain a stable glow discharge, because above this pressure instabilities degenerate into arcs within the discharge volume. This problem can be overcome by pulse excitation; using submicrosecond pulse duration, instabilities do not have sufficient time to develop; hence, the gas pressure can be increased above atmospheric pressure and the laser can be operated in a pulsed mode. In a mode locked format optical pulses shorter than 1 ns can be produced. This is called a TEA laser and with a transverse gas flow is capable of producing short high power pulses up to a few kHz repetition frequency. In order to prevent arc formation, TEA lasers usually employ ultraviolet or e-beam preionization of the gas discharge just prior to the main current pulse being applied via a thyratron switch. Output coupling is usually via an unstable resonator. TEA lasers are used for marking, remote sensing, range-finding, and scientific applications.

Conclusions

It is 40 years since Patel operated the first high power CO_2 laser. This led to the first generation of lasers which were quickly exploited for industrial laser materials processing, medical applications, defense,

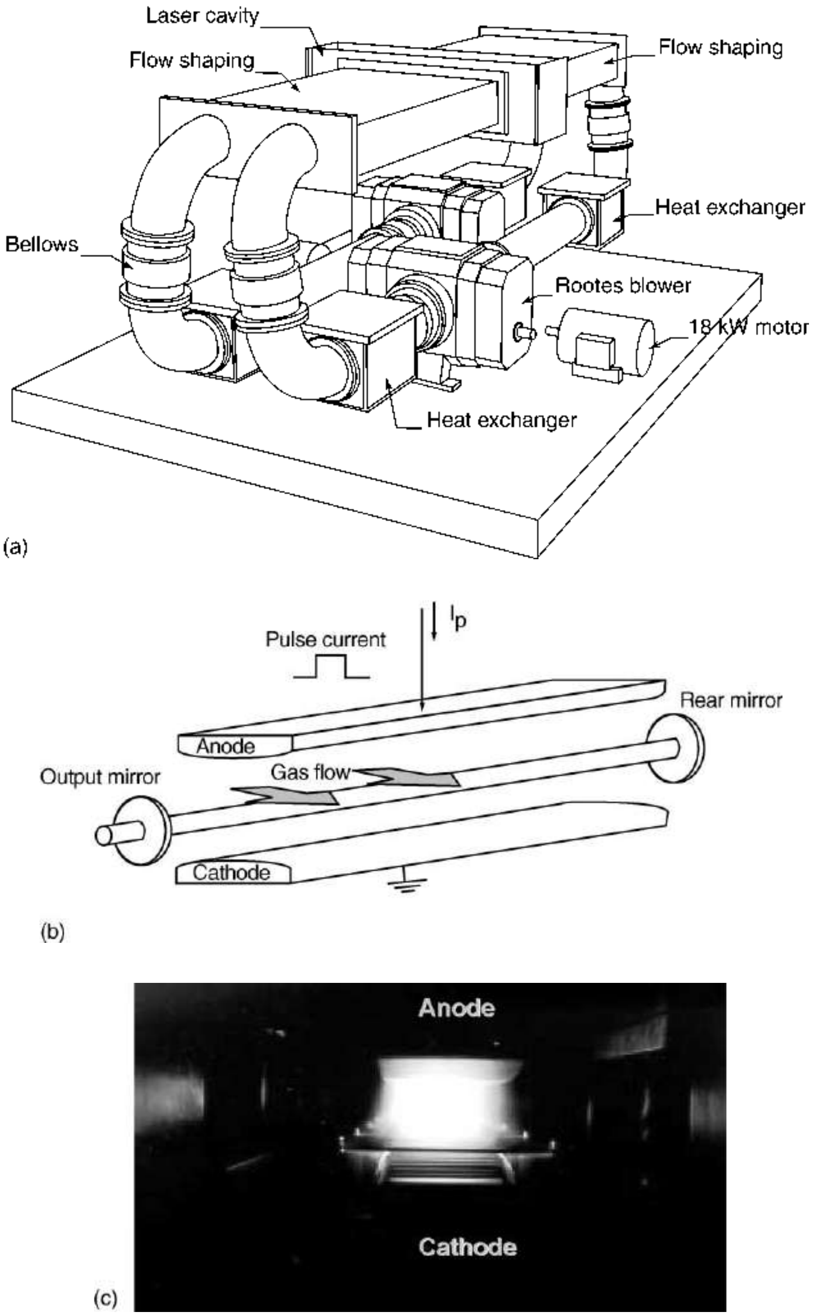
Laser cavity
Flow shaping
Flow shaping
Heat exchanger
Bellows
Rootes blower
18 kW motor
Heat exchanger
(a)
I_p
Pulse current
Rear mirror
Anode
Output mirror
Gas flow
Cathode
(b)
Anode
Cathode
(c)

Figure 9 (a) Transverse flow carbon dioxide laser gas recirculator, (b) Transverse flow carbon dioxide electrodes, (c) Transverse flow carbon dioxide gas discharge as seen from the output window.

and scientific research applications; however, the first generation of lasers were quite unreliable and temperamental. After many design iterations, the CO_2 laser has now matured into a reliable, stable laser source available in many different geometries and power ranges. The low cost of ownership of the latest generation of CO_2 laser makes them a very attractive commercial proposition for many industrial and scientific applications. Commercial lasers incorporate many novel design features that are beyond the scope of this article and are often peculiar to the particular laser manufacturer. This includes gas additives and catalysts that may be required to stabilize the gas discharge of a particular laser design; it is this optimization of the laser design that has produced such reliable and controllable low-cost performance from the CO_2 laser.

List of Units and Nomenclature

α	The direct excitation of carbon dioxide (CO_2) ground state molecules (sec^{-1})
β	The direct de-excitation of nitrogen (N_2) (sec^{-1})
γ	The direct excitation of nitrogen (N_2) ground state molecules (sec^{-1})
η	The direct de-excitation of carbon dioxide (CO_2) (sec^{-1})
λ	wavelength (m)
σ	absorption coefficient (cm^2)
τ	Electrical current pulse length (μs)
τ_{sp}	Spontaneous emission life time of the upper laser level (sec)
$A = (\tau_{sp})^{-1} = 0.213(sec^{-1})$	The Einstein '*A*' coefficient for the laser transition
c	velocity of light ($cm\ sec^{-1}$)
C_c	Coupling capacitance (nF)
CO_2	Carbon dioxide
e	subscript 'e' refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium
E	Electric Field ($V\ cm^{-1}$)
F_{CO_2}	Fraction of the input power (IP) coupled into the excitation of the energy level E_{00^01}
F_{N_2}	Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$
g	gain (cm^{-1})
g_1 and g_2	energy level degeneracy's of levels 1 and 2
He	Helium
I	beam irradiance ($W\ cm^{-2}$)
I_0	beam irradiance at the center of a Gaussian laser beam ($W\ cm^{-2}$)
I_p	Photon population density ($photons\ cm^{-3}$)
$\mathrm{I_p}$	Input current (A)
IP	Electrical input power ($W\ cm^{-3}$)
J	the rotational quantum number
K_{51}, K_{15}	Resonant energy transfer between the CO_2 (00^01) and $N_2(v = 2)$ energy levels proceeds via excited molecules colliding with ground state molecules (sec^{-1})
$K_{132}, K_{2131}, K_{2231}$	are vibration/vibration transfer rates (sec^{-1}) between energy levels 1 and 32; 21 and 31; 22 and 31, respectively (see Figure 1)
K_{320}	is a vibration/translation transfer rate (sec^{-1}) between energy levels 32 and 0 (see Figure 1)
K_{sp}, A	Spontaneous emission rate (sec^{-1})
L	The distance between the back and the front mirrors, which have reflectivity's of R_B and R_F (cm)
n	molecular population ($molecules\ cm^{-3}$)
N_2	Nitrogen
P	Pressure (Torr)
P_{CO_2}, P_{He} and P_{N_2}	The respective gas partial pressures (Torr)
P_{in}	Electrical input power (kW)
r	radius of laser beam (cm)
R_B	Back mirror reflectivity
R_F	Front mirror reflectivity
t	time (sec)
T	Temperature (deg K)
T_0	the photon decay time (sec)
w	the radial distance where the power density is decreased to $1/e^2$ of its axial value

See also

Fiber and Guided Wave Optics: Overview. **Lasers:** Noble Gas Ion Lasers.

Further Reading

Anderson JD (1976) *Gasdynamic Lasers: An Introduction*. New York: Academic Press.

Chatwin CR, McDonald DW and Scott BF (1991) Design of a High P.R.F. Carbon Dioxide Laser for Processing High Damage Threshold Materials. Selected Papers on Laser Design, SPIE Milestone Series, pp. 425–433. Washington: SPIE Optical Engineering Press.

Cool AC (1969) Power and gain characteristics of high speed flow lasers. *Journal of Applied Physics* 40(9): 3563.

Crafer RC, Gibson AF, Kent MJ and Kimmit MF (1969) Time-dependent processes on CO_2 laser amplifiers. *British Journal of Applied Physics* 2(2): 183.

Gerry ET and Leonard AD (1966) Measurement of 10.6-μ CO_2 laser transition probability and optical broadening cross sections. *Applied Physics Letters* 8(9): 227.

Gondhalekar A, Heckenberg NR and Holzhauer E (1975) The mechanism of single-frequency operation of the hybrid CO_2 laser. *IEEE Journal of Quantum Electronics* QE-11(3): 103.

Gordiets BF, Sobolev NN and Shelepin LA (1968) Kinetics of physical processes in CO_2 lasers. *Soviet Physics JETP* 26(5): 1039.

Herzberg G (1945) *Molecular Spectra & Molecular Structure*. Infra-red and Raman Spectra of Polyatomic Molecules, vol. 2. New York: Van Nostrand.

Hoffman AL and Vlases GC (1972) A simplified model for predicting gain, saturation and pulse length for gas dynamic lasers. *IEEE Journal of Quantum Electronics* 8(2): 46.
Johnson DC (1971) Excitation of an atmospheric pressure CO_2–N_2–He laser by capacitor discharges. *IEEE Journal of Quantum Electronics* Q.E.-7(5): 185.
Koechner W (1988) *Solid State Laser Engineering*. Berlin: Springer Verlag.
Kogelnik H and Li T (1966) Laser beams and resonators. *Applied Optics* 5: 1550–1567.
Levine AK and De Maria AJ (1971). *Lasers*, vol. 3. New York: Marcel Dekkar, Chapter 3.
Moeller G and Ridgen JD (1965) *Applied Physics Letters* 7: 274.
Moore CB, Wood RE, Bei-Lok Hu and Yardley JT (1967) Vibrational energy transfer in CO_2 lasers. *Journal of Chemical Physics* 11: 4222.
Patel CKN (1964) *Physics Review Letters* 12: 588.
Siegman A (1986) *Lasers*. Mill Valley, California: University Science.
Smith K and Thomson RM (1978) *Computer Modeling of Gas Lasers*. New York and London: Plenum Press.
Sobolev NN and Sokovikov VV (1967) CO_2 lasers. *Soviet Physics USPEKHI* 10(2): 153.
Svelto O (1998) *Principles of Lasers*, 4th edn. New York: Plenum.
Tychinskii VP (1967) Powerful lasers. *Soviet Physics USPEKHI* 10(2): 131.
Vlases GC and Money WM (1972) Numerical modelling of pulsed electric CO_2 lasers. *Journal of Applied Physics* 43(4): 1840.
Wagner WG, Haus HA and Gustafson KT (1968) High rate optical amplification. *IEEE Journal of Quantum Electronics* Q.E.-4: 287.
Witteman W (1987) *The CO_2 Laser.* Springer Series in Optical Sciences, vol. 53.

Dye Lasers

F J Duarte, Eastman Kodak Company, New York, NY, USA

A Costela, Consejo Superior de Investigaciones Cientificas, Madrid, Spain

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Background

Dye lasers are the original tunable lasers. Discovered in the mid-1960s these tunable sources of coherent radiation span the electromagnetic spectrum from the near-ultraviolet to the near-infrared (Figure 1). Dye lasers spearheaded and sustained the revolution in atomic and molecular spectroscopy and have found use in many and diverse fields from medical to military applications. In addition to their extraordinary spectral versatility, dye lasers have been shown to oscillate from the femtosecond pulse domain to the continuous wave (cw) regime. For microsecond pulse emission, energies of up to hundreds of joules per pulse have been demonstrated. Further, operation at high pulsed repetition frequencies (prfs), in the multi-kHz regime, has provided average powers at kW levels. This unrivaled operational versatility is summarized in Table 1.

Dye lasers are excited by coherent optical energy from an excitation, or pump, laser or by optical energy from specially designed lamps called flash-lamps. Recent advances in semiconductor laser

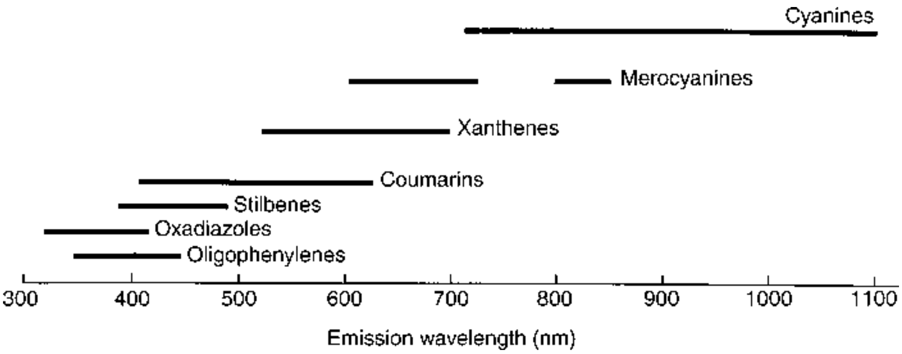

Figure 1 Approximate wavelength span from the various classes of laser dye molecules. Reproduced with permission from Duarte FJ (1995) *Tunable Laser Handbook*. New York: Academic Press.

Table 1 Emission characteristics of liquid dye lasers

Dye laser class	*Spectral coverage*[a]	*Energy per pulse*[b]	*Prf*[b]	*Power*[b]
Laser-pumped pulsed dye lasers	350–1100 nm	800 J[c]	13 kHz[d]	2.5 kW[d]
Flashlamp-pumped dye lasers	320–900 nm	400 J[e]	850 Hz[f]	1.2 kW[f]
CW dye lasers	370–1000 nm			43 W[g]

[a] Approximate range.
[b] Refers to maximum values for that particular emission parameter.
[c] Achieved with an excimer-laser pumped coumarin dye laser by Tang and colleagues, in 1987.
[d] Achieved with a multistage copper-vapor-laser pumped dye laser using rhodamine dye by Bass and colleagues, in 1992.
[e] Reported by Baltakov and colleagues, in 1974. Uses rhodamine 6G dye.
[f] Reported by Morton and Dragoo, in 1981. Uses coumarin 504 dye.
[g] Achieved with an Ar^+ laser pumped folded cavity dye laser using rhodamine 6G dye by Baving and colleagues, in 1982.

technology have made it possible to construct very compact all-solid-state excitation sources that, coupled with new solid-state dye laser materials, should bring the opportunity to build compact tunable laser systems for the visible spectrum. Further, direct diode-laser pumping of solid-state dye lasers should prove even more advantageous to enable the development of fairly inexpensive tunable narrow-linewidth solid-state dye laser systems for spectroscopy and other applications requiring low powers. Work on electrically excited organic gain media might also provide new avenues for further progress.

The literature of dye lasers is very rich and many review articles have been written describing and discussing traditional dye lasers utilizing liquid gain media. In particular, the books *Dye Lasers*, *Dye Laser Principles*, *High Power Dye Lasers*, and *Tunable Lasers Handbook* provide excellent sources of authoritative and detailed description of the physics and technology involved. In this article we offer only a survey of the operational capabilities of the dye lasers using liquid gain media in order to examine with more attention the field of solid-state dye lasers.

Brief History of Dye Lasers

1965: Quantum theory of dyes is discussed in the context of the maser (R. P. Feynman).
1966: Dye lasers are discovered (P. P. Sorokin and J. R. Lankard; F. P. Schäfer and colleagues).
1967: The flashlamp-pumped dye laser is discovered (P. P. Sorokin and J. R. Lankard; W. Schmidt and F. P. Schäfer).
1967–1968: Solid-state dye lasers are discovered (B. H. Soffer and B. B. McFarland; O. G. Peterson and B. B. Snavely).
1968: Mode-locking, using saturable absorbers, is demonstrated in dye lasers (W. Schmidt and F. P. Schäfer).
1970: The continuous-wave (cw) dye laser is discovered (O. G. Peterson, S. A. Tuccio, and B. B. Snavely).
1971: The distributed feedback dye laser is discovered (H. Kogelnik and C. V. Shank).
1971–1975: Prismatic beam expansion in dye lasers is introduced (S. A. Myers; E. D. Stokes and colleagues; D. C. Hanna and colleagues).
1972: Passive mode-locking is demonstrated in cw dye lasers (E. P. Ippen, C. V. Shank, and A. Dienes).
1972: The first pulsed narrow-linewidth tunable dye laser is introduced (T. W. Hänsch).
1973: Frequency stabilization of cw dye lasers is demonstrated (R. L. Barger, M. S. Sorem, and J. L. Hall).
1976: Colliding-pulse-mode locking is introduced (I. S. Ruddock and D. J. Bradley).
1977–1978: Grazing-incidence grating cavities are introduced (I. Shoshan and colleagues; M. G. Littman and H. J. Metcalf; S. Saikan).
1978–1980: Multiple-prism grating cavities are introduced (T. Kasuya and colleagues; G. Klauminzer; F. J. Duarte and J. A. Piper).
1981: Prism pre-expanded grazing-incidence grating oscillators are introduced (F. J. Duarte and J. A. Piper).
1982: Generalized multiple-prism dispersion theory is introduced (F. J. Duarte and J. A. Piper).
1983: Prismatic negative dispersion for pulse compression is introduced (W. Dietel, J. J. Fontaine, and J-C. Diels).
1987: Laser pulses as short as six femtoseconds are demonstrated (R. L. Fork, C. H. Brito Cruz, P. C. Becker, and C. V. Shank).
1994: First narrow-linewidth solid-state dye laser oscillator (F. J. Duarte).
1999–2000: Distributed feedback solid-state dye lasers are introduced (Wadsworth and colleagues; Zhu and colleagues).

Molecular Energy Levels

Dye molecules have large molecular weights and contain extended systems of conjugated double bonds. These molecules can be dissolved in an adequate organic solvent (such as ethanol, methanol, ethanol/water, and methanol/water) or incorporated into a solid matrix (organic, inorganic, or hybrid). These molecular gain media have a strong absorption generally in the visible and ultraviolet regions, and exhibit large fluorescence bandwidths covering the entire visible spectrum. The general energy level diagram of an organic dye is shown in Figure 2. It consists of electronic singlet and triplet states with each electronic state containing a multitude of overlapping vibrational–rotational levels giving rise to broad continuous energy bands. Absorption of visible or ultraviolet pump light excites the molecules from the ground state S_0 into some rotational–vibrational level belonging to an upper excited singlet state, from where the molecules decay nonradiatively to the lowest vibrational level of the first excited singlet state S_1 on a picosecond time-scale. From S_1 the molecules can decay radiatively, with a radiative lifetime on the nanosecond time-scale, to a higher-lying vibrational–rotational level of S_0. From this level they rapidly thermalize into the lowest vibrational–rotational levels of S_0. Alternatively, from S_1, the molecules can experience nonradiative relaxation either to the triplet state T_1 by an intersystem crossing process or to the ground state by an internal conversion process. If the intensity of the pumping radiation is high enough a population inversion between S_1 and S_0 may be attained and stimulated emission occurs. Internal conversion and intersystem crossing compete with the fluorescence decay mode of the molecule and therefore reduce the efficiency of the laser emission. The rate for internal conversion to the electronic ground state is usually negligibly small so that the most important loss process is intersystem crossing into T_1 that populates the lower metastable triplet state. Thus, absorption on the triplet–triplet allowed transitions could cause considerable losses if these absorption bands overlap the lasing band, inhibiting or even halting the lasing process. This triplet loss can be reduced by adding small quantities of appropriate chemicals that favor nonradiative transitions that shorten the effective lifetime of the T_1 level. For pulsed excitation with nanosecond pulses, the triplet–triplet absorption can be neglected because for a typical dye the intersystem crossing rate is not fast enough to build up an appreciable triplet population in the nanosecond time domain.

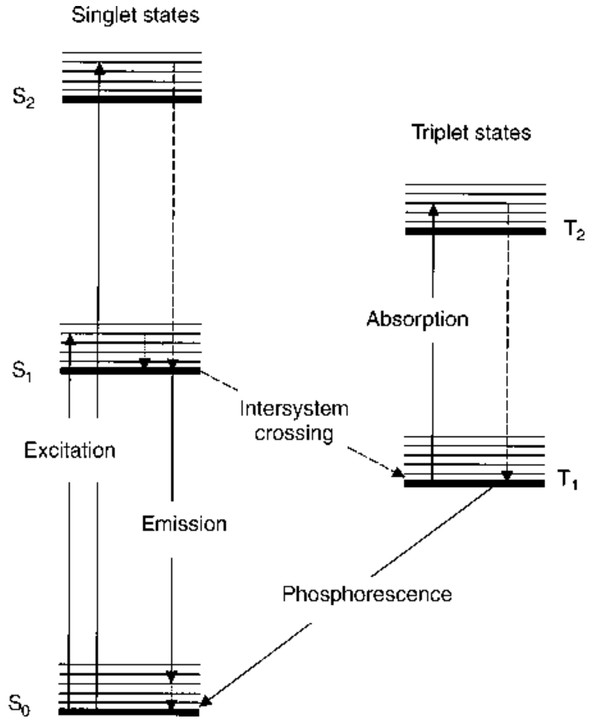

Figure 2 Schematic energy level diagram for a dye molecule. Full lines: radiative transitions; dashed lines: nonradiative transitions; dotted lines: vibrational relaxation.

Dye molecules are large (a typical molecule incorporates 50 or more atoms) and are grouped into families with similar chemical structures. A survey of the major classes of laser dyes is given later. Solid-state laser dye gain media are also considered later.

Liquid Dye Lasers

Laser-Pumped Pulsed Dye Lasers

Laser-pumped dye lasers use a shorter wavelength, or higher frequency, pulsed laser as the excitation or pump source. Typical pump lasers for dye lasers are gas lasers such as the excimer, nitrogen, or copper lasers. One of the most widely used solid-state laser pumps is the frequency doubled Nd:YAG laser which emits at 532 nm.

In a laser-pumped pulsed dye laser the active medium, or dye solution, is contained in an optical cell often made of quartz or fused silica, which provides an active region typically some 10 mm in length and a few mm in width. The active medium is then excited either longitudinally, or transversely, via a focusing lens using the pump laser. In the case of transverse excitation the pump laser is focused to a beam ~10 mm in width and ~0.1 mm in height. Longitudinal pumping requires focusing of the excitation beam to a diameter in the 0.1–0.15 mm range. For lasers operated at low prfs (a few pulses per second), the dye solution might be static. However, for high-prf operation (a few thousand pulses per second) the dye solution must be flowed at speeds of up to a few meters per second in order to dissipate the heat. A simple broadband optically pumped dye laser can be constructed using just the pump laser, the active medium, and two mirrors to form a resonator. In order to achieve tunable, narrow-linewidth, emission, a more sophisticated resonator must be employed. This is called a dispersive tunable oscillator and is depicted in Figure 3. In a dispersive tunable oscillator the exit side of the cavity is comprised of a partial reflector, or an output coupler, and the other end of the resonator is composed of a multiple-prism grating assembly. It is the dispersive characteristics of this multiple-prism grating assembly and the dimensions of the emission beam produced at the gain medium that determine the tunability and the narrowness, or spectral purity, of the laser emission.

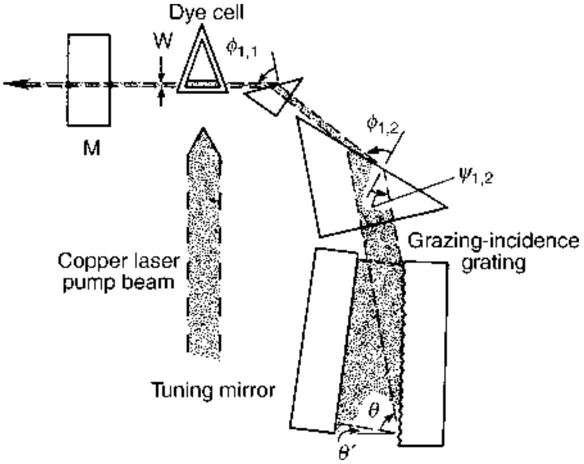

Figure 3 Copper-vapor-laser pumped hybrid multiple-prism near grazing incidence (HMPGI) grating dye laser oscillator. Adapted with permission from Duarte FJ and Piper JA (1984) Narrow linewidth high prf copper laser-pumped dye-laser oscillators. *Applied Optics* 23: 1391–1394.

In order to selectively excite a single vibrational–rotational level of a molecule such as iodine (I_2), at room temperature, one needs a laser linewidth of $\Delta\nu \approx 1.5$ GHz (or $\Delta\lambda \approx 0.0017$ nm at $\lambda = 590$ nm). The hybrid multiple-prism near-grazing-incidence (HMPGI) grating dye laser oscillator illustrated in Figure 4 yields laser linewidths in the 400 MHz $\leq \Delta\nu \leq$ 650 MHz range at 4–5% conversion efficiencies whilst excited by a copper-vapor laser operating at a prf of 10 kHz. Pulse lengths are ~10 ns at full-width half-maximum (FWHM). The narrow-linewidth emission from these oscillators is said to be single-longitudinal-mode lasing because only one electromagnetic mode is allowed to oscillate.

The emission from oscillators of this class can be amplified many times by propagating the tunable narrow-linewidth laser beam through single-pass amplifier dye cells under the excitation of pump lasers. Such amplified laser emission can reach enormous average powers. Indeed, a copper-vapor-laser pumped dye laser system at the Lawrence Livermore National Laboratory (USA), designed for the laser isotope separation program, was reported to

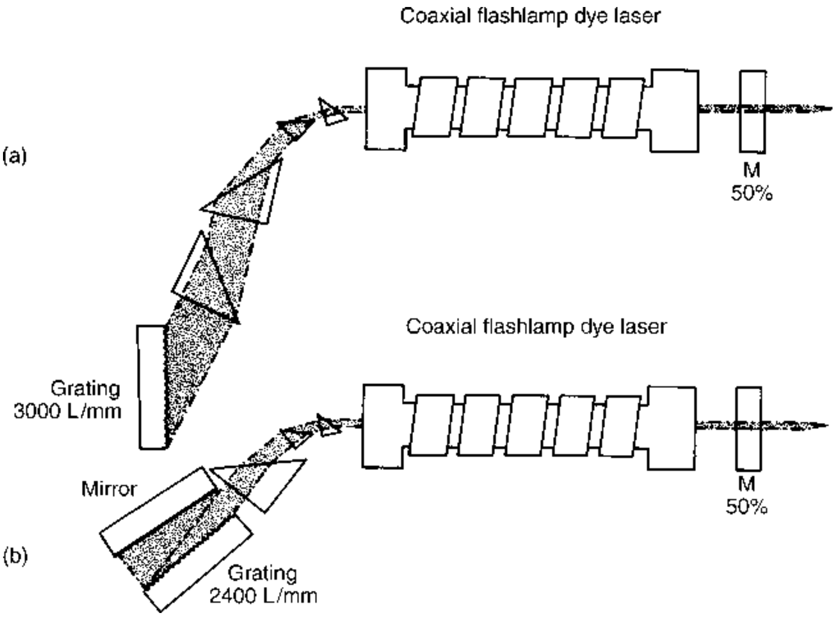

Figure 4 Flashlamp-pumped multiple-prism grating oscillators. From Duarte FJ, Davenport WE, Ehrlich JJ and Taylor TS (1991) Ruggedized narrow-linewidth dispersive dye laser oscillator. *Optics Communications* 84: 310–316. Reproduced with permission from Elsevier.

yield average powers in excess of 2.5 kW, at a prf of 13.2 kHz, at a better than 50% conversion efficiency as reported by Bass and colleagues in 1992. Besides high conversion efficiencies, excitation with copper-vapor lasers, at $\lambda = 510.554$ nm, has the advantage of inducing little photodegradation in the active medium thus allowing very long dye lifetimes.

Flashlamp-Pumped Dye Lasers

Flashlamps utilized in dye laser excitation emit at black-body temperatures in the 20 000 K range, thus yielding intense ultraviolet radiation centered around 200 nm. One further requirement for flashlamps, and their excitation circuits, is to deliver light pulses with a fast rise time. For some flashlamps this rise time can be less than a few nanoseconds.

Flashlamp-pumped dye lasers differ from laser-pumped pulsed dye lasers mainly in the pulse energies and pulse lengths attainable. This means that flashlamp-pumped dye lasers, using relatively large volumes of dye, can yield very large energy pulses. Excitation geometries use either coaxial lamps, with the dye flowing in a quartz cylinder at the center of the lamp, or two or more linear lamps arranged symmetrically around the quartz tube containing the dye solution. Using a relatively weak dye solution of rhodamine-6G (2.2×10^{-5} M), a coaxial lamp, and an active region defined by a quartz tube 6 cm in diameter and a length of 60 cm, Baltakov and colleagues, in 1974, reported energies of 400 J in pulses 25 μs long at FWHM.

Flashlamp-pumped tunable narrow-linewidth dye laser oscillators described by Duarte and colleagues, in1991, employ a cylindrical active region 6 mm in diameter and 17 cm in length. The dye solution is made of rhodamine 590 at a concentration of 1×10^{-5} M. This active region is excited by a coaxial flashlamp. Using multiple-prism grating architectures (see Figure 4) these authors achieve a diffraction limited TEM_{00} laser beam and laser linewidths of $\Delta\nu \approx 300$ MHz at pulsed energies in the 2–3 mJ range. The laser pulse duration is reported to be $\Delta t \approx 100$ ns. The laser emission from this class of multiple-prism grating oscillator is reported to be extremely stable. The tunable narrow-linewidth emission from these dispersive oscillators is either used directly in spectroscopic, or other scientific applications, or is utilized to inject large flashlamp-pumped dye laser amplifiers to obtain multi-joule pulse energies with the laser linewidth characteristics of the oscillator.

Continuous Wave Dye Lasers

CW dye lasers use dye flowing at linear speeds of up to 10 meters per second which are necessary to remove the excess heat and to quench the triplet states. In the original cavity reported by Peterson and colleagues, in 1970, a beam from an Ar^+ laser was focused on to an active region which is contained within the resonator. The resonator comprised dichroic mirrors that transmit the blue-green radiation of the pump laser and reflect the red emission from the dye molecules. Using a pump power of about 1 W, in a TEM_{00} laser beam, these authors reported a dye laser output of 30 mW. Subsequent designs replaced the dye cell with a dye jet, an introduced external mirror, and integrated dispersive elements in the cavity. Dispersive elements such as prisms and gratings are used to tune the wavelength output of the laser. Frequency-selective elements, such as etalons and other types of interferometers, are used to induce frequency narrowing of the tunable emission. Two typical cw dye laser cavity designs are described by Hollberg, in 1990, and are reproduced here in Figure 5. The first design is a linear three-mirror folded cavity. The second one is an eight-shaped ring dye laser cavity comprised of mirrors M_1, M_2, M_3, and M_4. Linear cavities exhibit the effect of *spatial hole burning* which allows the cavity to lase in more than one longitudinal mode. This problem can be overcome in ring cavities (Figure 5b) where the laser emission is in the form of a traveling wave.

Two aspects of cw dye lasers are worth emphasizing. One is the availability of relatively high powers in single longitudinal mode emission and the other is the

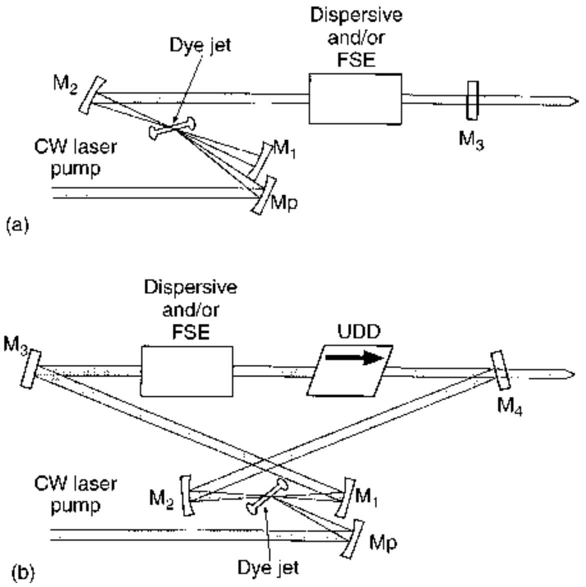

Figure 5 CW laser cavities: (a) linear cavity and (b) ring cavity. Adapted from Hollberg LW (1990) CW dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 185–238. New York: Academic Press. Reproduced with permission from Elsevier.

demonstration of very stable laser oscillation. First, Johnston and colleagues, in 1982, reported 5.6 W of stabilized laser output in a single longitudinal mode at 593 nm at a conversion efficiency of 23%. In this work eleven dyes were used to span the spectrum continuously from ~400 nm to ~900 nm. In the area of laser stabilization and ultra narrow-linewidth oscillation it is worth mentioning the work of Hough and colleagues, in 1984, that achieved laser linewidths of less than 750 Hz employing an external reference cavity.

Ultrashort-Pulse Dye Lasers

Ultrashort-pulse, or femtosecond, dye lasers use the same type of technology as cw dye lasers configured to incorporate a saturable absorber region. One such configuration is the ring cavity depicted in Figure 6. In this cavity the gain region is established between mirrors M_1 and M_2 whilst the saturable absorber is deployed in a counter-propagating arrangement. This arrangement is necessary to establish a collision between two counter-propagating pulses at the saturable absorber thus yielding what is known as colliding-pulse mode locking (CPM) as reported by Ruddock and Bradley, in 1976. This has the effect of creating a transient grating, due to interference, at the absorber thus shortening the pulse. Intracavity prisms were incorporated in order to introduce negative dispersion, by Dietel and colleagues in 1983, and thus subtract dispersion from the cavity and ultimately provide the compensation needed to produce femtosecond pulses.

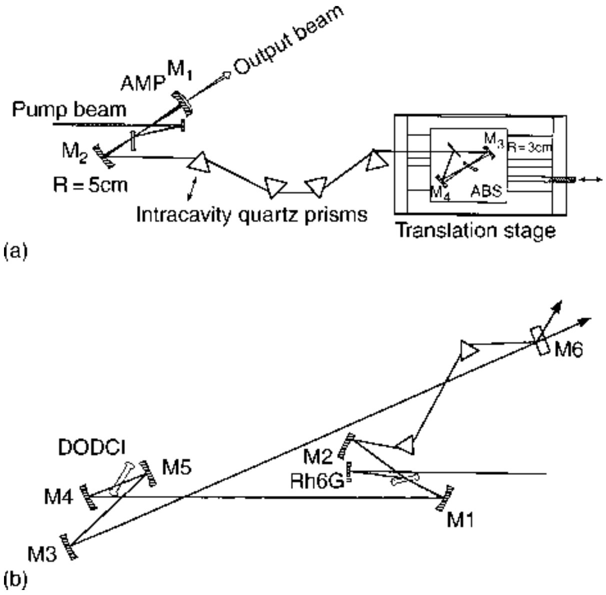

Figure 6 Femtosecond dye laser cavities: (a) linear femtosecond cavity and (b) ring femtosecond cavity. Adapted from Diels J-C (1990) Femtosecond dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 41–132. New York: Academic Press. Reproduced with permission from Elsevier.

The shortest pulse obtained from a dye laser, 6 fs, has been reported by Fork and colleagues, in 1987, using extra-cavity compression. In that experiment, a dye laser incorporating CPM and prismatic compensation was used to generate pulses that were amplified by a copper-vapor laser at a prf of 8 kHz. The amplified pulses, of a duration of 50 fs, were then propagated through two grating pairs and a four-prism sequence for further compression.

Solid-State Dye Laser Oscillators

In this section the principles of linewidth narrowing in dispersive resonators are outlined. Albeit the discussion focuses on multiple-prism grating solid-state dye laser oscillators, in particular, the physics is applicable to pulsed high-power dispersive tunable lasers in general.

Multiple-Prism Dispersion Grating Theory

The spectral linewidth in a dispersive optical system is given by

$$\Delta\lambda \approx \Delta\theta(\nabla_\lambda \theta)^{-1} \qquad [1]$$

where $\Delta\theta$ is the light beam divergence, $\nabla_\lambda = \partial/\partial\lambda$, and $\nabla_\lambda\theta$ is the overall dispersion of the optics. This identity can de derived either from the principles of geometrical optics or from the principles of generalized interferometry as described by Duarte, in 1992.

The cumulative single-pass generalized multiple-prism dispersion at the *m*th prism, of a multiple-prism array as illustrated in Figure 7, as given by Duarte and Piper, in 1982,

$$\begin{aligned}\nabla_\lambda\phi_{2,m} = {} & H_{2,m}\nabla_\lambda n_m + (k_{1,m}k_{2,m})^{-1} \\ & \times\left(H_{1,m}\nabla_\lambda n_m \pm \nabla_\lambda\phi_{2,(m-1)}\right) \qquad [2]\end{aligned}$$

In this equation

$$k_{1,m} = \cos\psi_{1,m}/\cos\phi_{1,m} \qquad [3a]$$

$$k_{2,m} = \cos\phi_{2,m}/\cos\psi_{2,m} \qquad [3b]$$

$$H_{1,m} = \tan\phi_{1,m}/n_m \qquad [4a]$$

$$H_{2,m} = \tan\phi_{2,m}/n_m \qquad [4b]$$

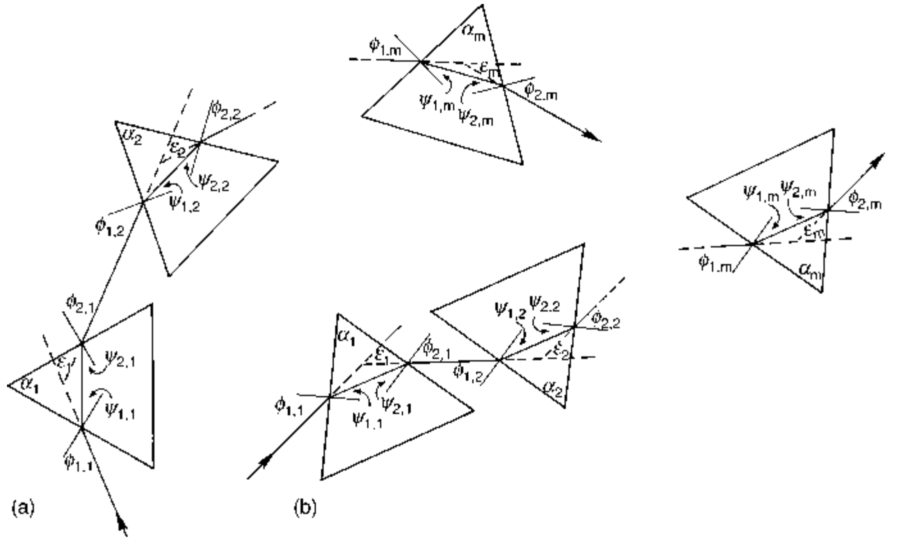

Figure 7 Generalized multiple-prism arrays in (a) additive and (b) compensating configurations. From Duarte FJ (1990) Narrow-linewidth pulsed dye laser oscillators. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 133–183. New York: Academic Press. Reproduced with permission from Elsevier.

Here, $k_{1,m}$ and $k_{2,m}$ represent the physical beam expansion experienced by the incident and the exit beams, respectively.

Equation [2] indicates that $\nabla_\lambda \phi_{2,m}$, the cumulative dispersion at the mth prism, is a function of the geometry of the mth prism, the position of the light beam relative to this prism, the refractive index of the prism, and the cumulative dispersion up to the previous prism $\nabla_\lambda \phi_{2,(m-1)}$.

For an array of r identical isosceles, or equilateral, prisms arranged symmetrically, in an additive configuration, so that the angles of incidence and emergence are the same, the cumulative dispersion reduces to

$$\nabla_\lambda \phi_{2,r} = r \nabla_\lambda \phi_{2,1} \qquad [5]$$

Under these circumstances the dispersions add in a simple and straightforward manner. For configurations incorporating right-angle prisms, the dispersions need to be handled mathematically in a more subtle form.

The generalized double-pass, or return-pass, dispersion for multiple-prism beam expanders was introduced by Duarte, in 1985:

$$\nabla_\lambda \Phi_P = 2M_1M_2 \sum_{m=1}^{r} (\pm 1) H_{1,m} \left(\prod_{j=m}^{r} k_{1,j} \prod_{j=m}^{r} k_{2,j} \right)^{-1} \nabla_\lambda n_m + 2 \sum_{m=1}^{r} (\pm 1) H_{2,m} \left(\prod_{j=1}^{m} k_{1,j} \prod_{j=1}^{m} k_{2,j} \right) \nabla_\lambda n_m \qquad [6]$$

Here, M_1 and M_2 are the beam magnification factors given by

$$M_1 = \prod_{m=1}^{r} k_{1,m} \qquad [7a]$$

$$M_2 = \prod_{m=1}^{r} k_{2,m} \qquad [7b]$$

For a multiple prism expander designed for an orthogonal beam exit, and Brewster's angle of incidence, eqn [6] reduces to the succinct expression given by Duarte, in 1990:

$$\nabla_\lambda \Phi_P = 2 \sum_{m=1}^{r} (\pm 1)(n_m)^{m-1} \nabla_\lambda n_m \qquad [8]$$

Equation [6] can be used to either quantify the overall dispersion of a given multiple-prism beam expander or to design a prismatic expander yielding zero dispersion, that is, $\nabla_\lambda \Phi_P = 0$, at a given wavelength.

Physics and Architecture of Solid-State Dye-Laser Oscillators

The first high-performance narrow-linewidth tunable laser was introduced by Hänsch in 1972. This laser yielded a linewidth of $\Delta\nu \approx 2.5$ GHz (or $\Delta\lambda \approx 0.003$ nm at $\lambda \approx 600$ nm) in the absence of an intracavity etalon. Hänsch demonstrated that the laser linewidth from a tunable laser was narrowed significantly when the beam incident on the tuning grating was expanded using an astronomical telescope. The linewidth equation including the intracavity beam magnification factor can be written as

$$\Delta\lambda \approx \Delta\theta (M \nabla_\lambda \Theta_G)^{-1} \qquad [9]$$

From this equation, it can be deduced that a narrow $\Delta\lambda$ is achieved by reducing $\Delta\theta$ and increasing the overall intracavity dispersion ($M\nabla_\lambda\Theta_G$). The intracavity dispersion is optimized by expanding the size of the intracavity beam incident on the diffractive surface of the tuning grating until it is totally illuminated.

Hänsch used a two-dimensional astronomical telescope to expand the intracavity beam incident on the diffraction grating. A simpler beam expansion method consists in the use of a single-prism beam expander as disclosed by several authors (Myers, 1971; Stokes and colleagues, 1972; Hanna and colleagues, 1975). An extension and improvement of this approach was the introduction of multiple-prism beam expanders as reported by Kasuya and colleagues, in 1978, Klauminzer, in 1978, and Duarte and Piper, in 1980. The main advantages of multiple-prism beam expanders, over two-dimensional telescopes, are simplicity, compactness, and the fact that the beam expansion is reduced from two dimensions to one dimension. Physically, as explained previously, prismatic beam expanders also introduce a dispersion component that is absent in the case of the astronomical telescope. Advantages of multiple-prism beam expanders over single-prism beam expansion are higher transmission efficiency, lower amplified spontaneous emission levels, and the flexibility to either augment or reduce the prismatic dispersion.

In general, for a pulsed multiple-prism grating oscillator, Duarte and Piper, in 1984, showed that the return-pass dispersive linewidth is given by

$$\Delta\lambda = \Delta\theta_R (MR\nabla_\lambda\Theta_G + R\nabla_\lambda\Phi_P)^{-1} \qquad [10]$$

where R is the number of return-cavity passes. The grating dispersion in this equation, $\nabla_\lambda\Theta_G$, can be either from a grating in Littrow or near grazing-incidence configuration. The multiple-return-pass equation for the beam divergence was given by Duarte, in 2001:

$$\Delta\theta_R = (\lambda/\pi w)\left(1 + (L_R/B_R)^2 + (A_R L_R/B_R)^2\right)^{1/2} \qquad [11]$$

Here, $L_R = (\pi w^2/\lambda)$ is the Rayleigh length of the cavity, w is the beam waist at the gain region, while A_R and B_R are the corresponding multiple-return-pass elements derived from propagation matrices.

At present, very compact and optimized multiple-prism grating tunable laser oscillators are found in two basic cavity architectures. These are the multiple-prism Littrow (MPL) grating laser oscillator, reported by Duarte in 1999 and illustrated in Figure 8, and the hybrid multiple-prism near grazing-incidence (HMPGI) grating laser oscillator, introduced by Duarte in 1997 and depicted in Figure 9. In early MPL grating oscillators the individual prisms integrating the multiple-prism expander were deployed in an additive configuration thus adding the cumulative dispersion to that of the grating and thus contributing to the overall dispersion of the cavity. In subsequent architectures the prisms were deployed in compensating configurations so as to yield zero dispersion and thus allow the tuning characteristics of the cavity to be

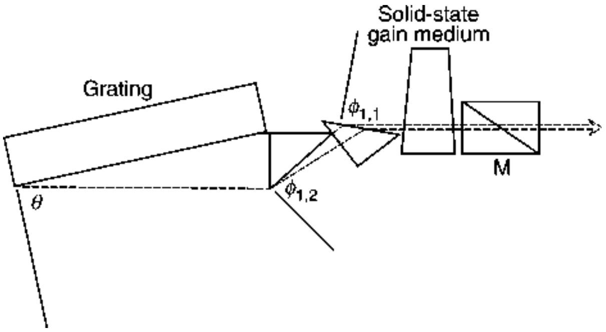

Figure 8 MPL grating solid-state dye laser oscillator: optimized architecture. The physical dimensions of the optical components of this cavity are shown to scale. The length, of the solid-state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1999) Multiple-prism grating solid-state dye laser oscillator: optimized architecture. *Applied Optics* 38: 6347–6349.

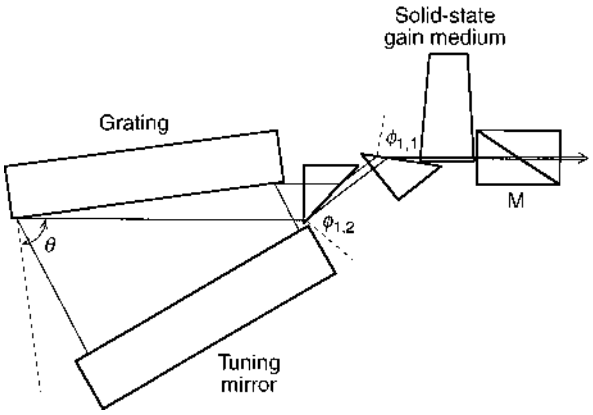

Figure 9 HMPGI grating solid-state dye laser. Schematics to scale. The length of the solid state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1997) Multiple-prism near-grazing-incidence grating solid-state dye laser oscillator. *Optics and Laser Technology* 29: 513–516.

determined by the grating exclusively. In this approach the principal role of the multiple-prism array is to expand the beam incident on the grating thus augmenting significantly the overall dispersion of the cavity as described in eqns [9] or [10]. In this regard, it should be mentioned that beam magnification factors of up to 100, and beyond, have been reported in the literature. Using solid-state laser dye gain media, these MPL and HMPGI grating laser oscillators deliver tunable single-longitudinal-mode emission at laser linewidths 350 MHz $\leq \Delta\nu \leq$ 375 MHz and pulse lengths in the 3–7 ns (FWHM) range. Long-pulse operation of this class of multiple-prism grating oscillators has been reported by Duarte and colleagues, in 1998. In these experiments laser linewidths of $\Delta\nu \approx 650$ MHz were achieved at pulse lengths in excess of 100 ns (FWHM) under flashlamp-pumped dye laser excitation.

The dispersive cavity architectures described here have been used with a variety of laser gain media in the gas, the liquid, and the solid state. Applications to tunable semiconductor lasers have also been reported by Zorabedian, in 1992, Duarte, in 1993, and Fox and colleagues, in 1997.

Concepts important to MPL and HMPGI grating tunable laser oscillators include the emission of a single-transverse-mode (TEM_{00}) laser beam in a compact cavity, the use of multiple-prism arrays, the expansion of the intracavity beam incident on the grating, the control of the intracavity dispersion, and the quantification of the overall dispersion of the multiple-prism grating assembly via generalized dispersion equations. Sufficiently high intracavity dispersion leads to the achievement of return-pass dispersive linewidths close to the free-spectral range of the cavity. Under these circumstances single-longitudinal-mode lasing is readily achieved as a result of multipass effects.

A Note on the Cavity Linewidth Equation

So far we have described how the cavity linewidth equation

$$\Delta\lambda \approx \Delta\theta(\nabla_\lambda \theta)^{-1}$$

and the dispersion equations can be applied to achieve highly coherent, or very narrow-linewidth, emission. It should be noted that the same physics can be applied to achieve ultrashort, or femtosecond, laser pulses. It turns out that intracavity prisms can be configured to yield negative dispersion as described by Duarte and Piper, in 1982, Dietel and colleagues, in 1983, and Fork and colleagues in 1984. This negative dispersion can reduce significantly the overall dispersion of the cavity. Under those circumstances, eqn [1] predicts broadband emission which according to the uncertainty principle, in the form of,

$$\Delta\nu\Delta t \approx 1 \qquad [12]$$

can lead to very short pulse emission since $\Delta\nu$ and $\Delta\lambda$ are related by the identities

$$\Delta\lambda \approx \lambda^2/\Delta x \qquad [13]$$

and

$$\Delta\nu \approx c/\Delta x \qquad [14]$$

Distributed-Feedback Solid-State Dye Lasers

Recently, narrow-linewidth laser emission from solid-state dye lasers has also been obtained using a distributed-feedback (DFB) configuration by Wadsworth and colleagues, in 1999, and Zhu and colleagues, in 2000. As is well known, in a DFB laser no external cavity is required, the feedback being provided by Bragg reflection from a permanently or dynamically written grating structure within the gain medium as reported by Kogelnik and Shank, in 1971. By using a DFB laser configuration where the interference of two pump beams induces the required periodic modulation in the gain medium, laser emission with linewidth in the 0.01–0.06 nm range have been reported. Specifically, Wadsworth and colleagues reported a laser linewidth of 12 GHz (0.016 nm at 616 nm) using Perylene Red doped PMMA at a conversion efficiency of 20%.

It should also be mentioned that the DFB laser is also a dye laser development that has found wide and extensive applicability in semiconductor lasers employed in the telecommunications industry.

Laser Dye Survey

Cyanines

These are red and near-infrared dyes which present long conjugated methine chains (!CH = CH!) and are useful in the spectral range longer than 800 nm, where no other dyes compete with them. An important laser dye belonging to this class is 4-dicyanomethylene-2-methyl-6-(*p*-dimethylamino-styryl)-4H-pyran (DCM, Figure 10a), with laser emission in the 600–700 nm range depending on the pump and solvent, liquid or solid, used.

Xanthenes

These dyes have a xanthene ring as the chromophore and are classified into rhodamines, incorporating

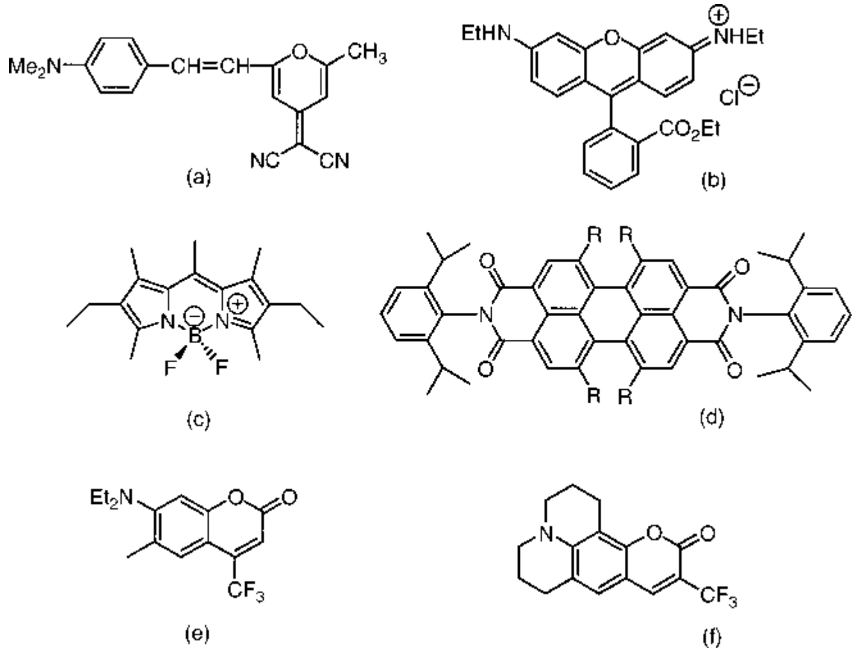

Figure 10 Molecular structures of some common laser dyes: (a) DCM; (b) Rh6G; (c) PM567; (d) Perylene Orange (R = H) and Perylene Red (R = C_4H_6O); (e) Coumarin 307 (also known as Coumarin 503); (f) Coumarin 153 (also known as Coumarin 540A).

amino radicals substituents, and fluoresceins, with hydroxyl (OH) radical substituents. They are generally very efficient and chemically stable and their emission covers the wavelength region from 500 to 700 nm.

Rhodamine dyes are the most important group of all laser materials, with rhodamine 6G (Rh6G, Figure 10b), also called rhodamine 590 chloride, being probably the best known of all laser dyes. Rh6G exhibits an absorption peak, in ethanol, at 530 nm and a fluorescence peak at 556 nm, with laser emission typically in the 550–620 nm region. It has been demonstrated to lase efficiently both in liquid and solid solutions.

Pyrromethenes

Pyrromethene.BF_2 complexes are a new class of laser dyes synthesized and characterized more recently as reported by Shah and colleagues, in 1990, Pavlopoulos and colleagues, in 1990, and Boyer and colleagues, in 1993. These laser dyes exhibit reduced triplet–triplet absorption over their fluorescence and lasing spectral region while retaining a high quantum fluorescence yield. Depending on the substituents on the chromophore, these dyes present laser emission over the spectral region from the green/yellow to the red, competing with the rhodamine dyes and have been demonstrated to lase with good performance when incorporated into solid hosts. Unfortunately, they are relatively unstable because of the aromatic amine groups in their structure, which render them vulnerable to photochemical reactions with oxygen as indicated by Rahn and colleagues, in 1997. The molecular structure of a representative and well-known dye of this class, 1,3,5,7,8-pentamethyl-2,6-diethylpyrromethene-difluoroborate complex (pyrromethene 567, PM567) is shown in Figure 10c. Recently, analogs of dye PM567 substituted at position 8 with an acetoxypolymethylene linear chain or a polymerizable methacryloyloxy polymethylene chain have been developed. These new dipyrromethene.BF_2 complexes have demonstrated improved efficiency and better photostability than the parent compound both in liquid and solid state.

Conjugated Hydrocarbons

This class of organic compounds includes perylenes, stilbenes, and *p*-terphenyl. Perylene and perylimide dyes are large, nonionic, nonpolar molecules (Figure 10d) characterized by their extreme photostability and negligible singlet–triplet transfer, as discussed by Seybold and Wagenblast and colleagues, in 1989, which have high quantum efficiency because of the absence of nonradiative relaxation. The perylene dyes exhibit limited solubility in conventional solvents but dissolve well in acetone, ethyl acetate, and methyl methacrylate, with emission wavelengths in the orange and red spectral regions.

Stilbene dyes are derivatives of uncyclic unsaturated hydrocarbons such as ethylene and butadiene, with emission wavelengths in the 400–500 nm range. Most of them end in phenyl radicals. Although they are chemically stable, their laser performance is inferior to that of coumarins. *p*-Terphenyl is a valuable and efficient *p*-oligophenylene dye with emission in the ultraviolet.

Coumarins

Coumarin derivatives are a popular family of laser dyes with emission in the blue-green region of the spectrum. Their structure is based on the coumarin ring (Figure 10e,f) with different substituents that strongly affect its chemical characteristics and allow covering an emission spectral range between 420 and 580 nm. Some members of this class rank among the most efficient laser dyes known, but their chemical photobleaching is rapid compared with xanthene and pyrromethene dyes. An additional class of blue-green dyes is the tetramethyl derivatives of coumarin dyes introduced by Chen and colleagues, in 1988. The emission from these dyes spans the spectrum in the 453–588 nm region. These coumarin analogs exhibit improved efficiency, higher solubility, and better lifetime characteristics than the parent compounds.

Azaquinolone Derivatives

The quinolone and azaquinolone derivatives have a structure similar to that of coumarins but with their laser emission range extended toward the blue by 20–30 nm. The quinolone dyes are used when laser emission in the 400–430 nm region is required. Azaquinolone derivatives, such as 7-dimethylamino-1-methyl-4-metoxy-8-azaquinolone-2 (LD 390), exhibit laser action below 400 nm.

Oxadiazole Derivatives

These are compounds which incorporate an axodiazole ring with aryl radical substituents and which lase in the 330–450 nm spectral region. Dye 2-(4-biphenyl)-5-(4-*t*-butylphenyl)-1,3,4-oxadiazole (PBD), with laser emission in the 355–390 nm region, belongs to this family.

Solid-State Laser Dye Matrices

Although some first attempts to incorporate dye molecules into solid matrices were already made in the early days of development of dye lasers, it was not until the late 1980s that host materials with the required properties of optical quality and high damage threshold to laser radiation began to be developed. In subsequent years, the synthesis of new high-performance dyes and the implementation of new ways of incorporating the organic molecules into the solid matrix resulted in significant advances towards the development of practical tunable solid-state dye lasers. A recent detailed review of the work done in this field has been given by Costela *et al.*

Inorganic glasses, transparent polymers, and inorganic–organic hybrid matrices have been successfully used as host matrices for laser dyes. Inorganic glasses offer good thermal and optical properties but present the difficulty that the high melting temperature employed in the traditional process of glass making would destroy the organic dye molecules. It was not until the development of the low-temperature sol-gel technique for the synthesis of glasses that this limitation was overcome. The sol-gel process, based on inorganic polymerization reactions performed at about room temperature and starting with metallo-organic compounds such as alkoxides or salts, as reported by Brinker and Scherrer, in 1989, provides a safe route for the preparation of rigid, transparent, inorganic matrix materials incorporating laser dyes. Disadvantages of these materials are the possible presence of impurities embedded in the matrix or the occurrence of interactions between the dispersed dye molecules and the inorganic structure (hydrogen bonds, van der Waals forces) which, in certain cases, can have a deleterious effect on the lasing characteristics of the material.

The use of matrices based on polymeric materials to incorporate organic dyes offers some technical and economical advantages. Transparent polymers exhibit high optical homogeneity, which is extremely important for narrow-linewidth oscillators, as explained by Duarte, in 1994, good chemical compatibility with organic dyes, and allow control over structure and chemical composition making possible the modification, in a controlled way, of relevant properties of these materials such as polarity, free volume, molecular weight, or viscoelasticity. Furthermore, the polymeric materials are amenable to inexpensive fabrication techniques, which facilitate miniaturization and design of integrated optical systems. The dye molecules can be either dissolved in the polymer or linked covalently to the polymeric chains.

In the early investigations on the use of solid polymer matrices for laser dyes, the main problem to be solved was that of the low resistance to laser radiation exhibited by the then existing materials. In the late 1980s and early 1990s new modified polymeric organic materials began to be developed with a laser-radiation-damage threshold comparable to that of inorganic glasses and crystals as reported by Gromov and colleagues, in 1985, and Dyumaev and colleagues, in 1992. This, together with the above-mentioned advantages, made the use of polymers in solid-state dye lasers both attractive and competitive.

An approach that intends to bring about materials in which the advantages of both inorganic glasses and polymers are preserved but the difficulties are avoided is that of using inorganic–organic hybrid matrices. These are silicate-based materials, with an inorganic Si–O–Si backbone, prepared from organosilane precursors by sol-gel processing in combination with organic cross-linking of polymerizable monomers as reported by Novak, in 1993, Sanchez and Ribot, in 1994, and Schubert, in 1995. In one procedure, laser dyes are mixed with organic monomers, which are then incorporated into the porous structure of a sol-gel inorganic matrix by immersing the bulk in the solution containing monomer and catalyst or photoinitiator as indicated by Reisfeld and Jorgensen, in 1991, and Bosch and colleagues, in 1996. Alternatively, hybrids can be obtained from organically modified silicon alkoxides, producing the so-called ORMOCERS (organically modified ceramics) or ORMOSILS (organically modified silanes) as reported by Reisfeld and Jorgensen, in 1991. An inconvenient aspect of these materials is the appearance of optical inhomogeneities in the medium due to the difference in the refractive index between the organic and inorganic phases as well as to the difference in density between monomer and polymer which causes stresses and the formation of optical defects. As a result, spatial inhomogeneities appear in the laser beam, thereby decreasing its quality as discussed by Duarte, in 1994.

Organic Hosts

The first polymeric materials with improved resistance to damage by laser radiation were obtained by Russian workers, who incorporated rhodamine dyes into modified poly(methyl methacrylate) (MPMMA), obtained by doping PMMA with low-molecular weight additives. Some years later, in 1995, Maslyukov and colleagues demonstrated lasing efficiencies in the range 40–60% with matrices of MPMMA doped with rhodamine dyes pumped longitudinally at 532 nm, with a useful lifetime (measured as a 50% efficiency drop) of 15 000 pulses at a prf of 3.33 Hz.

In 1995, Costela and colleagues used an approach based on adjusting the viscoelastic properties of the material by modifying the internal plasticization of the polymeric medium by copolymerization with appropriate monomers. Using dye Rh6G dissolved in a copolymer of 2-hydroxyethyl methacrylate (HEMA) and methyl methacrylate (MMA) and transversal pumping at 337 nm, they demonstrated laser action with an efficiency of 21% and useful lifetime of 4500 pulses (20 GJ/mol in terms of total input energy per mole of dye molecule when the output energy is down to 50% of its initial value). The useful lifetime increased to 12 000 pulses when the Rh6G chromophore was linked covalently to the polymeric chains as reported by Costela and colleagues, in 1996.

Comparative studies on the laser performance of Rh6G incorporated either in copolymers of HEMA and MMA or in MPMMA were carried out by Giffin and colleagues, in 1999, demonstrating higher efficiency of the MPMMA materials but superior normalized photostability (up to 240 GJ/mol) of the copolymer formulation.

Laser conversion efficiencies in the range 60–70% for longitudinal pumping at 532 nm were reported by Ahmad and colleagues, in 1999, with PM567 dissolved in PMMA modified with 1,4-diazobicyclo[2,2,2] octane (DABCO) or perylene additives. When using DABCO as an additive, a useful lifetime of up to 550 000 pulses, corresponding to a normalized photostability of 270 GJ/mol, was demonstrated at a 2 Hz prf.

Much lower efficiencies and photostabilities have been obtained with dyes emitting in the blue-green spectral region. Costela and colleagues, in 1996, performed some detailed studies on dyes coumarin 540A and coumarin 503 incorporated into methacrylate homopolymers and copolymers, and demonstrated efficiencies of at most 19% with useful lifetimes of up to 1200 pulses, at a 2 Hz prf, for transversal pumping at 337 nm. In 2000, Somasundaram and Ramalingam obtained useful lifetimes of 5240 and 1120 pulses, respectively, for coumarin 1 and coumarin 490 dyes incorporated into PMMA rods modified with ethyl alcohol, under transversal pumping at 337 nm and a prf of 1 Hz.

A detailed study of photo-physical parameters of R6G-doped HEMA-MMA gain media was performed by Holzer and colleagues, in 2000. Among the parameters determined by these authors are quantum yields, lifetimes, absorption cross-sections, and emission cross-sections. A similar study on the photophysical parameters of PM567 and two PM567 analogs incorporated into solid matrices of different acrylic copolymers and in corresponding mimetic liquid solutions was performed by Bergmann and colleagues, in 2001.

Preliminary experiments with rhodamine 6G-doped MPMMA at high prfs were conducted using a frequency-doubled Nd:YAG laser at 5 kHz by Duarte and colleagues, in 1996. In those experiments, involving longitudinal excitation, the pump laser was allowed to fuse a region at the incidence window of the gain medium so that a lens was formed. This lens and the exit window of the gain medium comprised a short unstable resonator that yielded broadband emission. In 2001, Costela and colleagues demonstrated lasing in rhodamine 6G- and pyrromethene-doped polymer matrices under copper-vapor-laser excitation at prfs in the 1.0–6.2 kHz range. In these experiments, the dye-doped solid-state matrix was rotated at 1200 rpm. Average powers of 290 mW were obtained at a prf of 1 kHz and a conversion efficiency of 37%. For a short period of time average powers of up to 1 W were recorded at a prf of 6.2 kHz. In subsequent experiments by Abedin and colleagues the rpf was increased to 10 kHz by using as pump laser a diode-pumped, *Q*-switched, frequency doubled, solid state Nd:YLF laser. Initial average powers of 560 mW and 430 mW were obtained for R6G-doped and PM567-doped polymer matrices, respectively. Lasing efficiency was 16% for R6G and the useful lifetime was 6.6 min (or about 4.0 million shots).

Inorganic and Hybrid Hosts

In 1990, Knobbe and colleagues and McKiernan and colleagues, from the University of California at Los Angeles, incorporated different organic dyes, via the sol-gel techniques, into silicate (SiO_2), aluminosilicate (Al_2O_3–SiO_2), and ORMOSIL host matrices, and investigated the lasing performance of the resulting materials under transversal pumping. Lasing efficiencies of 25% and useful lifetimes of 2700 pulses at 1 Hz repetition rate were obtained from Rh6G incorporated into aluminosilicate gel and ORMOSIL matrices, respectively. When the dye in the ORMOSIL matrices was coumarin 153, the useful lifetime was still 2250 pulses. Further studies by Altman and colleagues, in 1991, demonstrated useful lifetimes of 11 000 pulses at a 30 Hz prf when rhodamine 6G perchlorate was incorporated into ORMOSIL matrices. The useful lifetime increased to 14 500 pulses when the dye used was rhodamine B.

When incorporated into xerogel matrices, pyrromethene dyes lase with efficiencies as high as the highest obtained in organic hosts but with much higher photostability: a useful lifetime of 500 000 pulses at 20 Hz repetition rate was obtained by Faloss and colleagues, in 1997, with PM597 using a pump energy of 1 mJ. The efficiency of PM597 dropped to 40% but its useful lifetime increased to over 1 000 000 pulses when oxygen-free samples were prepared, in agreement with previous results that documented the strong dependence of laser parameters of pyrromethene dyes on the presence of oxygen. A similar effect was observed with perylene dyes: deoxygenated xerogel samples of Perylene Red exhibited a useful lifetime of 250 000 pulses at a prf of 5 Hz and 1 mJ pump energy, to be compared with a useful lifetime of only 10 000 pulses obtained with samples prepared in normal conditions. Perylene Orange incorporated into polycom glass and pumped longitudinally at 532 nm gave an efficiency of 72% and a useful lifetime of 40 000 pulses at a prf of 1 Hz as reported by Rhan and King, in 1998.

A direct comparison study of laser performance of Rh6G, PM567, Perylene Red, and Perylene Orange in organic, inorganic, and hybrid hosts was carried out by Rahn and King in 1995. They found that the nonpolar perylene dyes had better performance in partially organic hosts, whereas the ionic rhodamine and pyrromethene dyes performed best in the inorganic sol-gel glass host. The most promising combinations of dye and host for efficiency and photostability were found to be Perylene Orange in polycom glass and Rh6G in sol-gel glass.

An all-solid-state optical configuration, where the pump was a laser diode array side-pumped, Q-switched, frequency-doubled, Nd:YAG slab laser, was demonstrated by Hu and colleagues, in 1998, with dye DCM incorporated into ORMOSIL matrices. A lasing efficiency of 18% at the wavelength of 621 nm was obtained, and 27 000 pulses were needed for the output energy to decrease by 10% of its initial value at 30 Hz repetition rate. After 50 000 pulses the output energy decreased by 90% but it could be recovered after waiting for a few minutes without pumping.

Violet and ultraviolet laser dyes have also been incorporated into sol-gel silica matrices and their lasing properties evaluated under transversal pumping by a number of authors. Although reasonable efficiencies have been obtained with some of these laser dyes, the useful lifetimes are well below 1000 pulses.

More recently, Costela and colleagues have demonstrated improved conversion efficiencies in dye-doped inorganic–organic matrices using pyrromethene 567 dye. The solid matrix in this case is poly-trimethyl-silyl-methacrylate cross-linked with ethylene glycol and copolymerized with methyl methacrylate. Also, Duarte and James have reported on dye-doped polymer-nanoparticle laser media where the polymer is PMMA and the nanoparticles are made of silica. These authors report TEM_{00} laser beam emission and improved dn/dT coefficients of the gain media that results in reduced $\Delta\theta$.

Dye Laser Applications

Dye lasers generated a renaissance in diverse applied fields such as isotope separation, medicine, photochemistry, remote sensing, and spectroscopy. Dye lasers have also been used in many experiments that have advanced the frontiers of fundamental physics.

Central to most applications of dye lasers is their capability of providing coherent narrow-linewidth radiation that can be tuned continuously from the near ultraviolet, throughout the visible, to the near infrared. These unique coherent and spectral properties have made dye lasers particularly useful in the field of high-resolution atomic and molecular spectroscopy. Dye lasers have been successfully and extensively used in absorption and fluorescence spectroscopy, Raman spectroscopy, selective excitations, and nonlinear spectroscopic techniques. Well documented is their use in photochemistry, that is, the chemistry of optically excited atoms and molecules, and in biology, studying biochemical reaction kinetics of biological molecules. By using nonlinear optical techniques, such as harmonic and sum frequency and difference frequency generation, the properties of dye lasers can be extended to the ultraviolet.

Medical applications of dye lasers include cancer photodynamic therapy, treatment of vascular lesions, and lithotripsy as described by Goldman in 1990.

The ability of dye lasers to provide tunable sub-GHz linewidths in the orange-red portion of the spectrum, at kW average powers, made them particularly suited to atomic vapor laser isotope separation of uranium as reported by Bass and colleagues, in 1992, Singh and colleagues, in 1994, and Nemoto and colleagues, in 1995. In this particular case the isotopic shift between the two uranium isotopes is a few GHz. Using dye lasers yielding $\Delta\nu \leq 1$ GHz the ^{235}U can be selectively excited in a multistep process by tuning the dye laser(s) to specific wavelengths, in the orange-red spectral region, compatible with a transition sequence leading to photoionization. Also, high-prf narrow-linewidth dye lasers, as described by Duarte and Piper, in 1984, are ideally suited for guide star applications in astronomy. This particular application requires a high-average power diffraction-limited laser beam at $\lambda \approx 589$ nm to propagate for some 95 km, above the surface of the Earth, and illuminate a layer of sodium atoms. Ground detection of the fluorescence from the sodium atoms allows measurements on the turbulence of the atmosphere. These measurements are then used to control the adaptive optics of the telescope thus compensating for the atmospheric distortions.

A range of industrial applications is also amenable to the wavelength agility and high-average power output characteristics of dye lasers. Furthermore, tunability coupled with narrow-linewidth emission, at large pulsed energies, make dye lasers useful in military applications such as directed energy and damage of optical sensors.

The Future of Dye Lasers

Predicting the future is rather risky. In the early 1990s, articles and numerous advertisements in commercial laser magazines predicted the demise and even the oblivion of the dye laser in a few years. It did not turn out this way. The high power and wavelength agility available from dye lasers have ensured their continued use as scientific and research tools in physics, chemistry, and medicine. In addition, the advent of narrow-linewidth solid-state dye lasers has sustained a healthy level of research and development in the field. Today, some 20 laboratories around the world are engaged in this activity.

The future of solid-state dye lasers depends largely on synthetic and manufacturing improvements. There is a need for strict control of the conditions of fabrication and a careful purification of all the compounds involved. In the case of polymers, in particular, stringent control of thermal conditions during the polymerization step is obligatory to ensure that adequate optical uniformity of the polymer matrix is achieved, and that intrinsic anisotropy developed during polymerization is minimized. From an engineering perspective what are needed are dye-doped solid-state media with improved photostability characteristics and better thermal properties. By better thermal properties is meant better dn/dT factors. To a certain degree progress in this area has already been reported. As in the case of liquid dye lasers, the lifetime of the material can be improved by using pump lasers at compatible wavelengths. In this regard, direct diode laser pumping of dye-doped solid-state matrices should lead to very compact, long-lifetime, tunable lasers emitting throughout the visible spectrum. An example of such a device would be a green laser-diode-pumped rhodamine-6G doped MPMMA tunable laser configured in a multiple-prism grating architecture.

As far as liquid lasers are concerned, there is a need for efficient water-soluble dyes. Successful developments in this area were published in the literature with coumarin-analog dyes by Chen and colleagues, in 1988, and some encouraging results with rhodamine dyes have been reported by Ray and colleagues, in 2002. Efficient water-soluble dyes could lead to significant improvements and simplifications in high-power tunable lasers.

See also

Coherent Control: Experimental. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Quantum Optics:** Entanglement and Quantum Information. **Relativistic Nonlinear Optics**.

Further Reading

Costela A, García-Moreno I and Sastre R (2001) In: Nalwa HS (ed.) *Handbook of Advanced Electronic and Photonic Materials and Devices*, vol. 7, pp. 161–208. San Diego, CA: Academic Press.

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena*. New York: Academic Press.

Demtröder W (1995) *Laser Spectroscopy*, 2nd edn. Berlin: Springer-Verlag.

Duarte FJ (ed.) (1991) *High Power Dye Lasers*. Berlin: Springer-Verlag.

Duarte FJ (ed.) (1995) *Tunable Lasers Handbook*. New York: Academic Press.

Duarte FJ (ed.) (1995) *Tunable Laser Applications*. New York: Marcel Dekker.

Duarte FJ (2003) *Tunable Laser Optics*. New York: Elsevier Academic.

Duarte FJ and Hillman LW (eds) (1990) *Dye Laser Principles*. New York: Academic Press.

Maeda M (1984) *Laser Dyes*. New York: Academic Press.

Schäfer FP (ed.) (1990) *Dye Lasers*, 3rd edn. Berlin: Springer-Verlag.

Radziemski LJ, Solarz RW and Paisner JA (eds) (1987) *Laser Spectroscopy and its Applications*. New York: Marcel Dekker.

Weber MJ (2001) *Handbook of Lasers*. New York: CRC.

Edge Emitters

J J Coleman, University of Illinois, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The semiconductor edge emitting diode laser is a critical component in a wide variety of applications, including fiber optics telecommunications, optical data storage, and optical remote sensing. In this section, we describe the basic structures of edge emitting diode lasers and the physical mechanisms for converting electrical current into light. Laser waveguides and cavity resonators are also outlined. The power, efficiency, gain, loss, and threshold characteristics of laser diodes are presented along with the effects of temperature and modulation. Quantum well lasers are outlined and a description of grating coupled lasers is provided.

Like all forms of laser, the edge emitting diode laser is an oscillator and has three principal components; a mechanism for converting energy into light, a medium that has positive optical gain, and a mechanism for obtaining optical feedback. The basic edge emitting laser diode is shown schematically in Figure 1. Current flow is vertical through a pn junction and light is emitted from the ends. The dimensions of the laser in Figure 1 are distorted, in proportion to typical real laser diodes, to reveal more detail. In practical laser diodes, the width of the stripe contact is actually much smaller than shown and the thickness is smaller still. Typical dimensions are (stripe width × thickness × length) $2 \times 100 \times 1000\ \mu\text{m}$. Light emission is generated only within a few micrometers of the surface, which means that 99.98% of the volume of this small device is inactive.

Energy conversion in diode lasers is provided by current flowing through a forward-biased pn junction. At the electrical junction, there is a high density of injected electrons and holes which can recombine in a direct energy gap semiconductor material to give an emitted photon. Charge neutrality requires equal numbers of injected electrons and holes. Figure 2 shows a simplified energy versus momentum (E–k) diagrams for (a) direct, and (b) indirect semiconductors. In a direct energy gap material, such as GaAs, the energy minima have the same momentum vector, so recombination of an electron in the conduction band and a hole in the valence band takes place directly. In an indirect material, such as silicon, momentum conservation requires the participation of a third particle – a phonon. This third-order process is far less efficient than direct recombination and, thus far, too inefficient to support laser action.

Optical gain in direct energy gap materials is obtained when population inversion is reached.

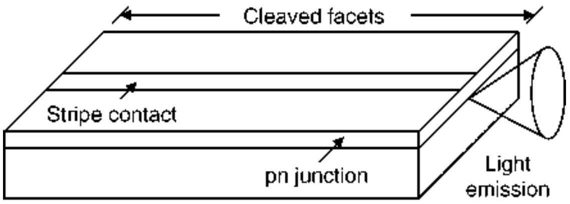

Figure 1 Schematic drawing of an edge emitting laser diode.

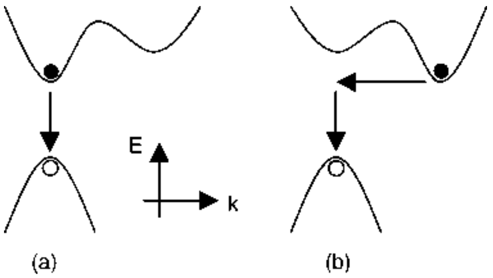

Figure 2 Energy versus momentum for (a) direct, and (b) indirect semiconductors.

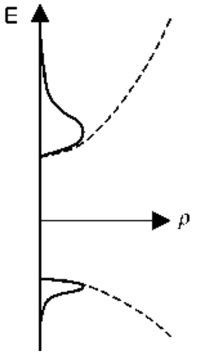

Figure 3 Density of states versus energy for conduction and valence bands.

Population inversion is the condition where the probability for stimulated emission exceeds that of absorption. The density of excited electrons n, in the conduction band is given by

$$n = \int_{E_c}^{\infty} \rho_n(E) f_n(E) \mathrm{d}E \ \mathrm{cm}^{-3} \qquad [1]$$

where $\rho_n(E)$ is the conduction band density of states function and $f_n(E)$ is the Fermi occupancy function. A similar equation can be written for holes in the valence band. Figure 3 shows the density of states ρ versus energy as a dashed line for conduction and valence bands. The solid lines in Figure 3 are the $\rho(E)f(E)$ products, and the areas under the solid curves are the carrier densities n and p.

The Fermi functions are occupation probabilities based on the quasi-Fermi levels E_{Fn} and E_{Fp}, which are, in turn, based on the injected carrier densities, δn and δp. Population inversion implies that the difference between the quasi-Fermi levels must exceed the emission energy which, for semiconductors, must be greater than the bandgap energy:

$$E_{Fn} - E_{Fp} \geq \hbar\omega \sim E_g \qquad [2]$$

Charge neutrality requires that $\delta n = \delta p$. Transparency is the point where these two conditions are just met and any additional injected current results in optical gain. The transparency current density J_o, is given by

$$J_o = \frac{q n_o L_z}{\tau_{spon}} \qquad [3]$$

where n_o is the transparency carrier density ($\delta n = \delta p = n_o$ at transparency), L_z is the thickness of the optically active layer, and τ_{spon} is the spontaneous carrier lifetime in the material (typically 5 nsec).

A resonant cavity for optical feedback is obtained simply by taking advantage of the natural crystal formation in single crystal semiconductors. Most III–V compound semiconductors form naturally into a zincblende lattice structure. With the appropriate choice of crystal planes and surfaces, this lattice structure can be cleaved such that nearly perfect plane–parallel facets can be formed at arbitrary distances from each other. Since the refractive index of these materials is much larger (~3.5) than that of air, there is internal optical reflection for normal incidence of approximately 30%. This type of optical cavity is called a Fabry–Perot resonator.

Effective lasers of all forms make use of optical waveguides to optimize the overlap of the optical field with material gain and cavity resonance. The optical waveguide in an edge emitting laser is best described by considering it in terms of a transverse waveguide component, defined by the epitaxial growth of multiple thin heterostructure layers, and a lateral waveguide component, defined by conventional semiconductor device processing methods. One of the simplest, and yet most common, heterostructure designs is shown in Figure 4. This five-layer separate confinement heterostructure (SCH) offers excellent carrier confinement in the active layer as well as a suitable transverse waveguide.

Clever choice of different materials for each layer results in the energy band structure, index of refraction

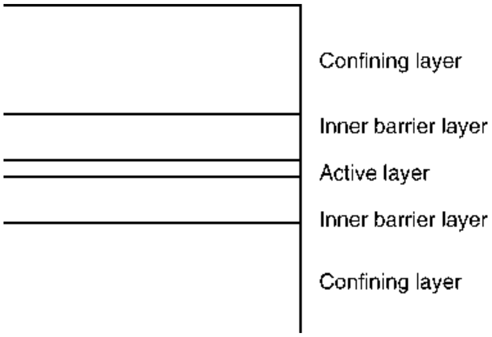

Figure 4 Schematic cross-section of a typical five-layer separate confinement heterostructure (SCH) laser.

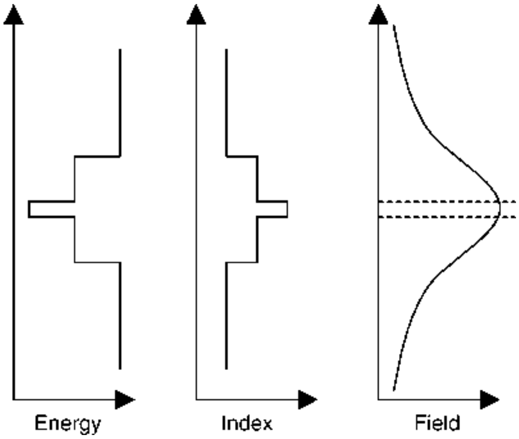

Figure 5 Energy band structure, index of refraction profile, and optical field mode profile of the SCH laser of Figure 4.

profile, and optical field mode profile shown in Figure 5. The outer confining layers are the thickest (~1 μm) and have the largest energy bandgap and lowest refractive index. The inner barrier layers are thinner (~0.1 μm), have intermediate bandgap energy and index of refraction, and serve both an optical role in defining the optical waveguide and an electronic role in confining carriers to the active layer. The active layer is the narrowest bandgap, highest index material, and is the only layer in the structure designed to provide optical gain. It may be a single layer or, as in the case of a multiple quantum well laser, a combination of layers. The bandgap energy, E, of this active layer plays a key role in determining the emission wavelength, λ, of the laser:

$$E = \frac{hc}{\lambda} \qquad [4]$$

where h is Planck's constant and c is the speed of light.

The bandgap energy of the SCH structure is shown in Figure 5. The lowest energy active layer collects injected electrons and holes and the carrier confinement provided by the inner barrier layers allows the high density of carriers necessary for population inversion and gain. The layers in the structure also have the refractive index profile shown in Figure 5 which forms a five-layer slab dielectric waveguide. With appropriate values for thicknesses and indices of refraction, this structure can be a very strong fundamental mode waveguide with the field profile shown. A key parameter of edge emitting diode lasers is Γ, the optical confinement factor. This parameter is the areal overlap of the optical field with the active layer, shown as dashed lines in the figure and can be as little as a few percent, even in very high performance lasers.

Lateral waveguiding in a diode laser can be accomplished in a variety of ways. A common example of an index guided laser is the ridge waveguide laser shown in Figure 6. After the appropriate transverse waveguide heterostructure sample is grown, conventional semiconductor processing methods are used to form parallel etched stripes from the surface near, but not through, the active layer. An oxide mask is patterned such that electrical contact is formed only to the center (core) region between the etched stripes, the core stripe being only a few microns wide. The lateral waveguide that results arises from the average index of refraction (effective index) in the etched regions outside the core being smaller than that of the core. Figure 7 shows the asymmetric near field emission profile of the ridge waveguide laser and cross-sectional profiles of the laser emission and effective index of refraction.

The optical spectrum of an edge emitting diode laser below and above threshold is the superposition of the material gain of the active layer and the resonances provided by the Fabry–Perot resonator. The spectral variation of material gain with injected carrier density (drive current) is shown in Figure 8. As the drive is increased, the intensity and energy of peak gain both increase.

The Fabry–Perot resonator has resonances every half wavelength, given by

$$L = \frac{m}{2}\frac{\lambda}{n} \qquad [5]$$

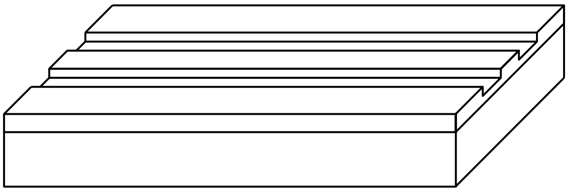

Figure 6 Schematic diagram of a real index guided ridge waveguide diode laser.

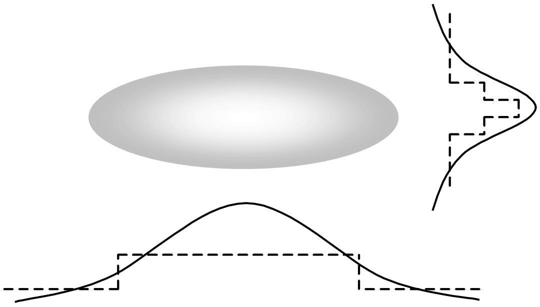

Figure 7 Near field emission pattern, cross-sectional emission profiles (solid lines), and refractive index profiles (dashed lines) from an index guided diode laser.

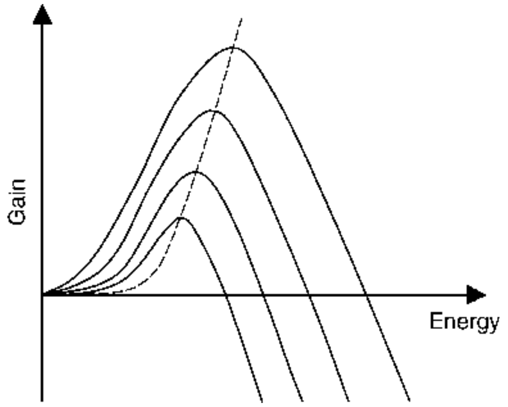

Figure 8 Gain versus emission energy for four increasing values of carrier (current) density. Peak gain is shown as a dashed line.

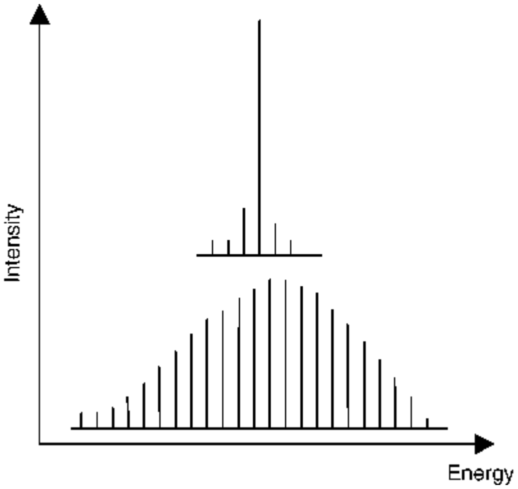

Figure 9 Emission spectra from a Fabry–Perot edge emitting laser structure at, and above, laser threshold.

where L is the cavity length, λ/n is the wavelength in the medium (n is the refractive index), and m is an integer. For normal length edge emitter cavities, these cavity modes are spaced only 1 or 2 Å apart. The result is optical spectra, near laser threshold and above, that look like the spectra of Figure 9. Many cavity modes resonate up to threshold, while above threshold, the superlinear increase in emission intensity with drive current tends to greatly favor one or two of the Fabry–Perot longitudinal cavity modes.

Laser threshold can be determined from a relatively simple analysis of the round trip gain and losses in the Fabry–Perot resonator. If the initial intensity is I_o, after one round trip the intensity will be given by

$$I = I_o R_1 R_2 e^{(g-\alpha_i)2L} \qquad [6]$$

where R_1 and R_2 are the reflectivities of the two facets, g is the optical gain, α_i are losses associated with optical absorption (typically 3–15 cm^{-1}), and L is the cavity length. At laser threshold, $g = g_{th}$ and $I = I_o$, i.e., the round trip gain exactly equals the losses. The result is

$$g_{th} = \alpha_i + \frac{1}{2L} \ln \frac{1}{R_1 R_2} \qquad [7]$$

where the second term represents the mirror losses α_m. Of course, mirror losses are desirable in the sense that they represent useable output power. Actually, the equation above does not take into account the incomplete overlap of the optical mode with the gain in the active layer. The overlap is defined by the confinement factor Γ, described above, and the equation must be modified to become

$$\Gamma g_{th} = \alpha_i + \frac{1}{2L} \ln \frac{1}{R_1 R_2} \qquad [8]$$

For quantum well lasers, the material peak gain, shown in Figure 8, as a function of drive current is given approximately by

$$g_t = \beta J_o \ln \frac{J}{J_o} \qquad [9]$$

where J_o is the transparency current density. The threshold current density J_{th}, which is the current density where the gain exactly equals all losses, is given by

$$J_{th} = \frac{J_o}{\eta_i} \exp \frac{\alpha + \frac{1}{2L} \ln \frac{1}{R_1 R_2}}{\Gamma \beta J_o} \qquad [10]$$

Note that an additional parameter η_i has appeared in this equation. Not all of the carriers injected by the drive current participate in optical processes; some are lost to other nonradiative processes. This is accounted for by including the internal quantum efficiency term η_i, which is typically at least 90%. The actual drive current, of course, must include the geometry of the device and is given by

$$I_{th} = wLJ_{th} \qquad [11]$$

where L is the cavity length, and w is the effective width of the active volume. The effective width is basically the stripe width of the core but also includes current spreading and carrier diffusion under the stripe.

The power–current characteristic (P–I) for a typical edge emitting laser diode is shown in Figure 10. As the drive current is increased

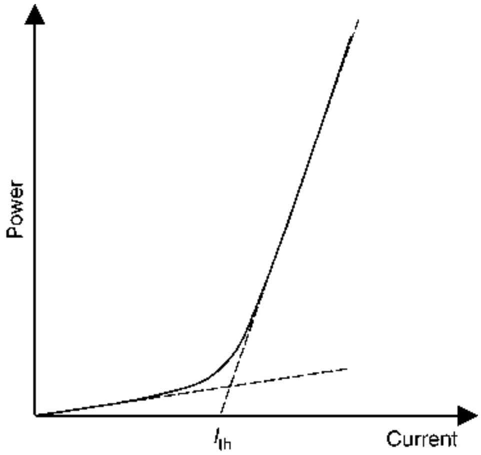

Figure 10 Power versus current ($P-I$) curve for a typical edge emitting laser diode.

from zero, the injected carrier densities increase and spontaneous recombination is observed. At some point the gain equals the internal absorption losses, and the material is transparent. Then, as the drive current is further increased, additional gain eventually equals the total losses, including mirror losses, and laser threshold is reached. Above threshold, nearly all additional injected carriers contribute to stimulated emission (laser output). The power generated internally is given by

$$P = wL(J - J_{\mathrm{th}})\eta_{\mathrm{i}}\frac{h\nu}{q} \qquad [12]$$

where wL is the area and $h\nu$ is the photon energy. Only a fraction of the power generated internally is extracted from the device

$$P_{\mathrm{o}} = wL(J - J_{\mathrm{th}})\eta_{\mathrm{i}}\frac{h\nu}{q}\frac{\alpha_{\mathrm{m}}}{\alpha_{\mathrm{i}} + \alpha_{\mathrm{m}}} \qquad [13]$$

Clearly a useful design goal is to minimize the internal optical loss α_{i}. The slope of the $P-I$ curve of Figure 10 above threshold is described by an external differential quantum efficiency η_{d}, which is related to the internal quantum efficiency by

$$\eta_{\mathrm{d}} = \eta_{\mathrm{i}}\left[\frac{\alpha_{\mathrm{m}}}{\alpha_{\mathrm{i}} + \alpha_{\mathrm{m}}}\right] \qquad [14]$$

Other common parameters used to characterize the efficiency of laser diodes include the power conversion efficiency η_{p}

$$\eta_{\mathrm{p}} = \frac{P_{\mathrm{o}}}{V_{\mathrm{F}}I} \qquad [15]$$

and the wall plug efficiency η_{w}

$$\eta_{\mathrm{w}} = \frac{P_{\mathrm{o}}}{I} \qquad [16]$$

The power conversion efficiency η_{p}, recognizes that the forward bias voltage V_{F} is larger than the photon energy by an amount associated with the series resistance of the laser diode. The wall plug efficiency η_{w}, has the unusual units of WA^{-1} and is simply a shorthand term that relates the common input unit of current to the common output unit of power. In high performance low power applications, the ideal device has minimum threshold current since the current used to reach threshold contributes little to light emission. In high power applications, the threshold current becomes insignificant and the critical parameter is external quantum efficiency.

Temperature effects may be important for edge emitting lasers, depending on a particular application. Usually, concerns related to temperature arise under conditions of high temperature operation ($\mathrm{T} > 50\,^{\circ}\mathrm{C}$), where laser thresholds rise and efficiencies fall. It became common early in the development of laser diodes to define a characteristic temperature T_{o} for laser threshold, which is given by

$$I_{\mathrm{th}} = I_{\mathrm{tho}} \exp\left(\frac{T}{T_{\mathrm{o}}}\right) \qquad [17]$$

where a high T_{o} is desirable. Unfortunately, this simple expression does not describe the temperature dependence very well over a wide range of temperatures, so the appropriate temperature range must also be specified. For applications where the emission wavelength of the laser is important, the change in wavelength with temperature $\mathrm{d}\lambda/\mathrm{d}T$ may also be specified. For conventional Fabry–Perot cavity lasers, $\mathrm{d}\lambda/\mathrm{d}T$ is typically 3–5 $\mathrm{\AA C}^{-1}$.

In order for a diode laser to carry information, some form of modulation is required. One method for obtaining this is direct modulation of the drive current in a semiconductor laser. The transient and temporal behavior of lasers is governed by rate equations for carriers and photons which are necessarily coupled equations. The rate equation for carriers is given by

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{J}{qL_{\mathrm{z}}} - \frac{n}{\tau_{\mathrm{sp}}} - (c/n)\beta(n - n_{\mathrm{o}})\varphi(E) \qquad [18]$$

where J/qL_{z} is the supply, n/τ_{sp} is the spontaneous emission rate, and the third term is the stimulated emission rate $(c/n)\beta(n - n_{\mathrm{o}})\varphi(E)$ where β is the gain coefficient and $\varphi(E)$ is the photon density.

The rate equation for photons is given by

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t} = (c/n)\beta(n - n_{\mathrm{o}})\varphi + \frac{\theta n}{\tau_{\mathrm{sp}}} - \frac{\varphi}{\tau_{\mathrm{p}}} \qquad [19]$$

where θ is the fraction of the spontaneous emission that couples into the mode (a small number), and τ_{p} is the photon lifetime (~1 psec) in the cavity. These rate equations can be solved for specific diode laser materials and structures to yield a modulation frequency response. A typical small signal frequency response, at two output power levels, is shown in Figure 11. The response is typically flat to frequencies above 1 GHz, rises to a peak value at some characteristic resonant frequency, and quickly rolls off. The characteristic small signal resonant frequency ω_{r}, is given by

$$\omega_{\mathrm{r}}^2 = \frac{(c/n)\beta\bar{\varphi}}{\tau_{\mathrm{p}}} \qquad [20]$$

where $\bar{\varphi}$ is the average photon density. Direct modulation of edge emitting laser diodes is limited by practicality to less than 10 GHz. This is in part because of the limits imposed by the resonant frequency and in part by chirp. Chirp is the frequency modulation that arises indirectly from direct current modulation. Modulation of the current modulates the carrier densities which, in turn, results in modulation of the quasi-Fermi levels. The separation of the quasi-Fermi levels is the emission energy (wavelength). Most higher-speed lasers make use of external modulation schemes.

The discussion thus far has addressed aspects of semiconductor diode edge emitting lasers that are common to all types of diode lasers irrespective of the choice of materials or the details of the active layer structure. The materials of choice for efficient laser devices include a wide variety of III–V binary compounds and ternary or quaternary alloys. The particular choice is based on such issues as emission wavelength, a suitable heterostructure lattice match, and availability of high-quality substrates. For example, common low-power lasers for CD players ($\lambda \sim 820$ nm) are likely to be made on a GaAs substrate using a GaAs–$\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ heterostructure. Lasers for fiberoptic telecommunications systems ($\lambda \sim 1.3$–1.5 μm) are likely to be made on an InP substrate using an InP–$\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}_y\mathrm{P}_{1-y}$ heterostructure. Red laser diodes ($\lambda \sim 650$ nm) are likely to be made on a GaAs substrate using a GaAs–$\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}_y\mathrm{P}_{1-y}$ heterostructure. Blue and ultraviolet lasers, a relatively new technology compared to the others, make use of $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{N}$–GaN–$\mathrm{In}_y\mathrm{Ga}_{1-y}\mathrm{N}$ heterostructures formed on sapphire or silicon carbide substrates.

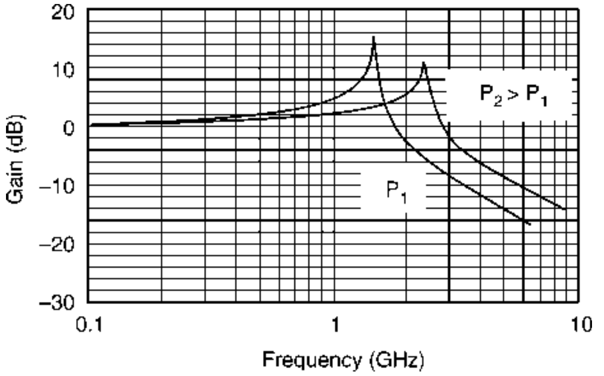

Figure 11 Direct current modulation frequency response for an edge emitting diode laser at two output power levels.

An important part of many diode laser structures is a strained layer. If a layer is thin enough, the strain arising from a modest amount of lattice mismatch can be accommodated elastically. In thicker layers, lattice mismatch results in an unacceptable number of dislocations that affect quantum efficiency, optical absorption losses, and, ultimately, long-term failure rates. Strained layer GaAs–$\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ lasers are a critical component in rare-earth doped fiber amplifiers.

The structure of the active layer in edge emitting laser diodes was originally a simple double heterostructure configuration with a single active layer of 500–1000 Å in thickness. In the 1970s however, advances in the art of growing semiconductor heterostructure materials led to growth of high-quality quantum well active layers. These structures, having thicknesses that are comparable to the electron wavelength in a semiconductor (<200 Å), revolutionized semiconductor lasers, resulting in much lower threshold current densities, higher efficiencies, and a broader range of available emission wavelengths. The energy band diagram for a quantum well laser active region is shown in Figure 12.

This structure is a physical realization of the particle-in-a-box problem in elementary quantum mechanics. The quantum well yields quantum states in the conduction band at discrete energy levels, with odd and even electron wavefunctions, and breaks the degeneracy in the valence band, resulting in separate energy states for light holes and heavy holes. The primary transition for recombination is from the $n = 1$ electron state to the $h = 1$ heavy hole state and takes place at a higher energy than the bulk bandgap energy. In addition, the density of states for quantum wells becomes a step-like function and optical gain is enhanced. The advantages in practical edge

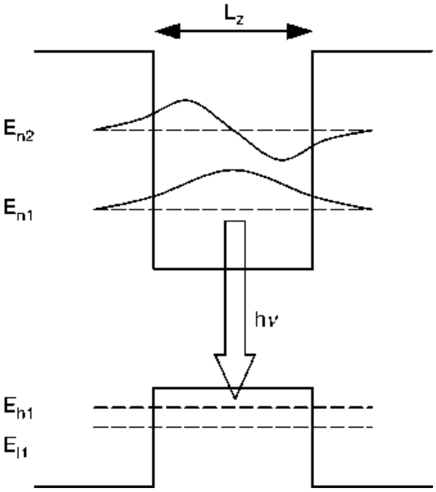

Figure 12 Energy band diagram for a quantum well heterostructure.

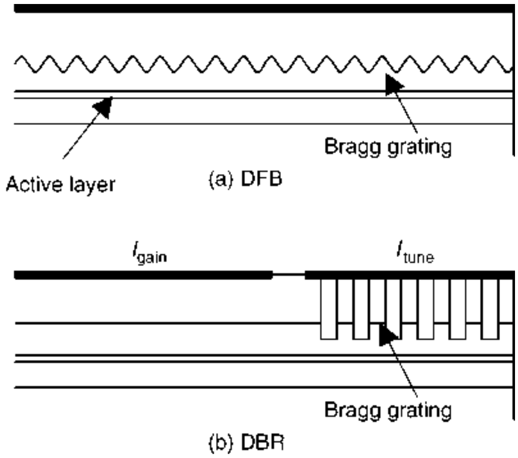

Figure 13 Schematic diagram of distributed feedback (DFB) and distributed Bragg reflector (DBR) edge emitting laser diodes.

emitting lasers that are the result of one-dimensional quantization are such that virtually all present commercial laser diodes are quantum well lasers. These kinds of advantages are driving development of laser diodes with additional degrees of quantization, including two dimensions (quantum wires) and three dimensions (quantum dots).

The Fabry–Perot cavity resonator described here is remarkably efficient and relatively easy to fabricate, which makes it the cavity of choice for many applications. There are many applications, however, where the requirements for linewidth of the laser emission or the temperature sensitivity of the emission wavelength make the Fabry–Perot resonator less desirable. An important laser technology that addresses both of these concerns involves the use of a wavelength selective grating as an integral part of the laser cavity. A notable example of a grating coupled laser is distributed feedback laser shown schematically in Figure 13a. This is similar to the SCH cross-section of Figure 4 with the addition of a Bragg grating above the active layer.

The period Λ, of the grating is chosen to fulfill the Bragg condition for *m*th-order coupling between forward- and backward-propagating waves, which is

$$\Lambda = \frac{m\lambda}{2n} \qquad [21]$$

where λ/n is the wavelength in the medium. The index step between the materials on either side of the grating is small and the amount of reflection is also small. Since the grating extends throughout the length of the cavity, however, the overall effective reflectivity can be large. Another variant is the distributed Bragg reflector (DBR) laser resonator, shown in Figure 13b. This DBR laser has a higher index contrast, deeply etched surface grating located at one or both ends of the otherwise conventional SCH laser heterostructure. One of the advantages of this structure is the opportunity to add a contact for tuning the Bragg grating by current injection.

Both the DBR and DFB laser structures are particularly well suited for telecommunications or other applications where a very narrow single laser emission line is required. In addition, the emission wavelength temperature dependence for this type of laser is much smaller, typically 0.5 $\text{Å}\,\text{C}^{-1}$.

See also

Semiconductor Physics: Quantum Wells and GaAs-based Structures.

Further Reading

Agrawal GP (ed.) (1995) *Semiconductor Lasers Past, Present, and Future*. Woodbury, NY: American Institute of Physics.

Agrawal GP and Dutta NK (1993) *Semiconductor Lasers*. New York: Van Nostrand Reinhold.

Bhattacharya P (1994) *Semiconductor Optoelectronic Devices*. Upper Saddle River, NJ: Prentice-Hall.

Botez D and Scifres DR (eds) (1994) *Diode Laser Arrays*. Cambridge, UK: Cambridge University Press.

Chuang SL (1995) *Physics of Optoelectronic Devices*. New York: John Wiley & Sons.

Coleman JJ (1992) *Selected Papers on Semiconductor Diode Lasers*. Bellingham, WA: SPIE Optical Engineering Press.

Einspruch NG and Frensley WR (eds) (1994) *Heterostructures and Quantum Devices*. San Diego, CA: Academic Press.

Iga K (1994) *Fundamentals of Laser Optics*. New York: Plenum Press.

Kapon E (1999) *Semiconductor Lasers I and II*. San Diego, CA: Academic Press.

Nakamura S and Fasol G (1997) *The Blue Laser Diode*. Berlin: Springer.

Verdeyen JT (1989) *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

Zory PS Jr (ed.) (1993) *Quantum Well Lasers*. San Diego, CA: Academic Press.

Excimer Lasers

J J Ewing, Ewing Technology Associates, Inc., Bellevue, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We present a summary of the fundamental operating principles of ultraviolet, excimer lasers. After a brief discussion of the economics and application motivation, the underlying physics and technology of these devices is described. Key issues and limitations in the scaling of these lasers are presented.

Background: Why Excimer Lasers?

Excimer lasers have become the most widely used source of moderate power pulsed ultraviolet sources in laser applications. The range of manufacturing applications is also large. Production of computer chips, using excimer lasers as the illumination source for lithography, has by far the largest commercial manufacturing impact. In the medical arena, excimer lasers are used extensively in what is becoming one of the world's most common surgical procedures, the reshaping of the lens to correct vision problems. Taken together, the annual production rate for these lasers is a relatively modest number compared to the production of semiconductor lasers. Current markets are in the range of $0.4 billion per year (Figure 1). More importantly, the unique wavelength, power, and pulse energy properties of the excimer laser enable a systems and medical procedure market, based on the excimer laser, to exceed $3 billion per year. The current average sales price for these lasers is in the range of $250 000 per unit, driven primarily by the production costs and reliability requirements of the lasers for lithography for chip manufacture. Relative to semiconductor diode lasers, excimer lasers have always been, and will continue to be, more expensive per unit device by a factor of order 10^4. UV power and pulse energy, however, are considerably higher.

The fundamental properties of these lasers enable the photochemical and photophysical processes used in manufacturing and medicine. Short pulses of UV light can photo-ablate materials, being of use in medicine and micromachining. Short wavelengths enable photochemical processes. The ability to provide a narrow linewidth using appropriate resonators enables precise optical processes, such as lithography. We trace out some of the core underlying properties of these lasers that lead to the core utility. We also discuss the technological problems and

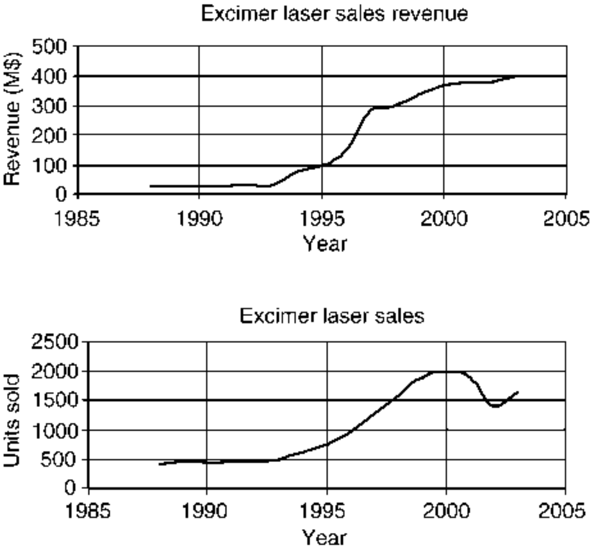

Figure 1 The sales of excimer lasers have always been measured in small numbers relative to other, less-expensive lasers. For many years since their commercial release in 1976 the sales were to R&D workers in chemistry, physics and ultimately biology and medicine. The market for these specialty but most useful R&D lasers was set to fit within a typical researcher's annual capital budget, i.e., less than $100K. As applications and procedures were developed and certified or approved, sales and average unit sales price increased considerably. Reliance on cyclical markets like semiconductor fabrication led to significant variations in production rates and annual revenues as can be seen.

solutions that have evolved to make these lasers so useful.

The data in Figure 1 provide a current answer to the question of ‘why excimer lasers?’ At the dawn of the excimer era, the answer to this question was quite different. In the early 1970s, there were no powerful short wavelength lasers, although IR lasers were being scaled up to significant power levels. However, visible or UV lasers could in principle provide better propagation and focus to a very small spot size at large distances. Solid state lasers of the time offered very limited pulse repetition frequency and low efficiency. Diode pumping of solid state lasers, and diode arrays to excite such lasers, was a far removed development effort. As such, short-wavelength lasers were researched over a wide range of potential media and lasing wavelengths. The excimer concept was one of those candidates.

We provide, in Figure 2, a ‘positioning’ map showing the range of laser parameters that can be obtained with commercial or developmental excimer lasers. The map has coordinates of pulse energy and pulse rate, with diagonal lines expressing average power. We show where both development goals and current markets lay in a map. Current applications are shown by the boxes in the lower right-hand corner. Excimer lasers have been built with clear apertures in the range of a fraction of 0.1 to 10^4 cm^2, with single pulse energies covering a range of 10^7. Lasers with large apertures require electron beam excitation. Discharge lasers correspond to the more modest pulse energy used for current applications. In general, for energies lower than ~5 J, the excitation method of choice is a self-sustained discharge, clearly providing growth potential for future uses, if required.

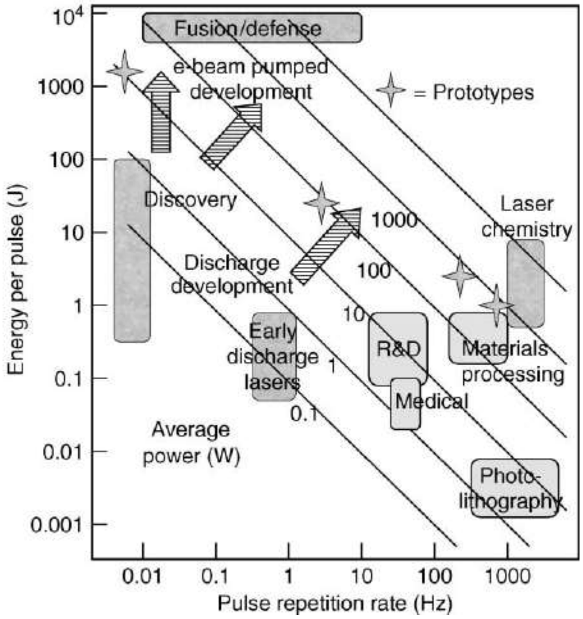

Figure 2 The typical energy and pulse rate for certain applications and technologies using excimer lasers. For energy under a few J, a discharge laser is used. The earliest excimer discharge lasers were derivatives of CO_2 TEA (transversely excited atmospheric pressure) and produced pulse energy in the range of 100 mJ per pulse. Pulse rates of order 30 Hz were all that the early pulsed power technology could provide. Early research focused on generating high energy. Current markets are at modest UV pulse energy and high pulse rates for lithography and low pulse rates for medical applications.

The applications envisaged in the early R&D days were in much higher energy uses, such as in laser weapons, laser fusion, and laser isotope separation. The perceived need for lasers for isotope separation, laser-induced chemistry, and blue-green laser-based communications from satellites, drove research in high pulse rate technology and reliability extension for discharge lasers. This research resulted in prototypes, with typical performance ranges as noted on the positioning map that well exceed current market requirements. However, the technology and underlying media physics developed in these early years made possible advances needed to serve the ultimate real market applications.

The very important application of UV lithography requires high pulse rates and high reliability. These two features differentiate the excimer laser used for lithography from the one used in the research laboratory. High pulse rates, over 2000 Hz in current practice, and the cost of stopping a computer chip production line for servicing, drive the reliability requirements toward 10^{10} shots per service interval. In contrast, the typical laboratory experiments are more often in the range of 100 Hz and, unlike a lithography production line, do not run 24 hours a day, 7 days a week. The other large market currently is in laser correction of myopia and other imperfections in the human eye. For this use the lasers are more akin to the commercial R&D style laser in terms of pulse rate and single pulse energy. Not that many shots are needed to ablate tissue to make the needed correction, and shot life is a straightforward requirement. However, the integration of these lasers into a certified and accepted medical procedure and an overall medical instrument drove the growth of this market.

The excimer concept, and the core technology used to excite these lasers, is applicable over a broad range of UV wavelengths, with utility at specific wavelength ranges from 351 nm in the near UV to 157 nm in the vacuum UV (VUV). There were many potential excimer candidates in the early R&D phase (Table 1). Some of these

Table 1 Key excimer wavelengths

Excimer emitter	*λ (nm)*	*Comment*
Ar_2	126	Very short emission wavelength, deep in the Vacuum UV; inefficient as laser due to absorption, see Xe_2.
Kr_2	146	Broad band emitter and early excimer laser with low laser efficiency. Very efficient converter of electric power into Vacuum UV light.
F_2	157	Molecule itself is not an excimer, but uses lower level dissociation; candidate for next generation lithography.
Xe_2	172	The first demonstrated excimer laser. Very efficient formation and emission but excited state absorption limits laser efficiency. Highly efficient as a lamp.
ArF	193	Workhorse for corneal surgery and lithography.
KrCl	222	Too weak compared to KrF and not as short a wavelength as ArF.
KrF	248	Best intrinsic laser efficiency; numerous early applications but found major market in lithography. Significant use in other materials processing.
XeI	254	Never a laser, but high formation efficiency and excellent for a lamp. Historically the first rare gas halide excimer whose emission was studied at high pressure.
XeBr	282	First rare gas halide to be shown as a laser; inefficient laser due to excited state absorption, but excellent fluorescent emitter; a choice for lamps.
Br_2	292	Another halogen molecule with excimer like transitions that has lased but had no practical use.
XeCl	308	Optimum medium for laser discharge excitation.
Hg_2	335	Hg vapor is very efficiently excited in discharges and forms excimers at high pressure. Subject of much early research, but as in the case of the rare gas excimers such as Xe_2 suffer from excited state absorption. Despite the high formation efficiency, this excimer was never made to lase.
I_2	342	Iodine UV molecular emission and lasing was first of the pure halogen excimer-like lasers demonstrated. This species served as kinetic prototype for the F_2 'honorary' excimer that has important use as a practical VUV source.
XeF	351, 353	This excimer laser transition terminates on a weakly bound lower level that dissociates rapidly, especially when heated. The focus for early defense-related laser development; as this wavelength propagates best of this class in the atmosphere.
XeF	480	This very broadband emission of XeF terminates on a different lower level than the UV band. Not as easily excited in a practical system, so has not had significant usage.
HgBr	502	A very efficient green laser that has many of the same kinetic features of the rare gas halide excimer lasers. But the need for moderately high temperature for the needed vapor pressure made this laser less than attractive for UV operation.
XeO	540	This is an excimer-like molecular transition on the side of the auroral lines of the O atom. The optical cross-section is quite low and as a result the laser never matured in practice.

candidates can be highly efficient, >50% relative to deposited energy, at converting electrical power into UV or visible emission, and have found a parallel utility as sources for lamps, though with differing power deposition rates and configurations. The key point in the table is that these lasers are powerful and useful sources at very short wavelengths. For lithography, the wavelength of interest has consistently shifted to shorter and shorter wavelengths as the printed feature size decreased. Though numerous candidates were pursued, as shown in Table 1, the most useful lasers are those using rare gas halide molecules. These lasers are augmented by dye laser and Raman shifting to provide a wealth of useful wavelengths in the visible and UV.

All excimer lasers utilize molecular dissociation to remove the lower laser level population. This lower level dissociation is typically from an unbound or very weakly bound ground-state molecule. However, halogen molecules also provide laser action on excimer-like transitions, that terminate on a molecular excited state that dissociates quickly. The vacuum UV transition in F_2 is the most practical method for making very short wavelength laser light at significant power. Historically, the rare gas dimer molecules, such as Xe_2, were the first to show excimer laser action, although self absorption, low gain, and poor optics in the VUV limited their utility. Other excimer-like species, such as metal halides, metal rare gas continua on the edge of metal atom resonance lines, and rare gas oxides were also studied. The key differentiators of rare gas halides from other excimer-like candidates are strong binding in the excited state, lack of self absorption, and relatively high optical cross-section for the laser transition.

Excimer Laser Fundamentals

Excimer lasers utilize lower-level dissociation to create and sustain a population inversion. For example, in the first excimer laser demonstrated, Xe_2, the lasing species is one which is comprised of two atoms that do not form a stable molecule in the ground state, as xenon dimers in the ground state are not a stable molecule. In the lowest energy state, that with neither of the atoms electronically excited, the interaction between the atoms in a collision is primarily one of repulsion, save for a very weak 'van der Waals' attraction at larger internuclear separations. This is shown in the potential energy diagram of Figure 3. If one of the atoms is excited, for example by an electric discharge, there can be a chemical binding in the electronically excited state of the molecule relative to the energy of one atom with an electron excited and one atom in the ground state. We use the shorthand notation * to denote an electronically excited atomic molecular species, e.g., Xe^*. The excited dimer, excimer for short, will recombine rapidly at high pressure from atoms into the Xe_2^* electronically excited molecule. Radiative lifetimes of the order 10 nsec to 1 µsec are typical before photon emission returns the molecule to the lower level. Collisional quenching often reduces the lifetime below the radiative rate. For Xe_2^* excimers, the emission is in the vacuum ultraviolet (VUV).

There are a number of variations on this theme, as noted in Table 1. Species, other than rare gas pairs, can exhibit broadband emission. The excited state binding energy can vary significantly, changing the fraction of excited states that will form molecules. The shape of the lower potential energy curve can vary as well. From a semantic point of view heteronuclear, diatomic molecules, such as XeO, are not excimers, as they are not made of two of the same atoms. However, the more formal name 'exciplex' did not stick in the laser world; excimer laser being preferred. Large binding energy is more efficient for capturing excited atoms into excited excimer-type molecules in the limited time they have before emission. Early measurements of the efficiency of converting electrical energy deposited into fluorescent radiation showed that up to 50% of the energy input could result in VUV light emission. Lasers were not this efficient due to absorption.

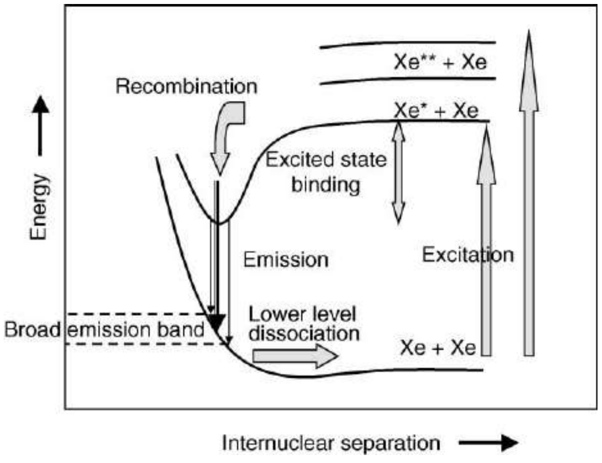

Figure 3 A schematic of the potential energy curves for an excimer species such as Xe_2. The excited molecule is bound relative to an excited atom and can radiate to a lower level that rapidly dissociates on psec timescales since the ground state is not bound, save for weak binding due to van der Waals forces. The emission band is intrinsically broad.

The diagram shown in Figure 3 is simplified because it does not show all of the excited states that exist for the excited molecule. There can be closely lying excited states that share the population and radiate at a lower rate, or perhaps absorb to higher levels, also not shown in Figure 3. Such excited state absorption may terminate on states derived from higher atomic excited states, noted as Xe^{**} in the figure, or yield photo-ionization, producing the diatomic ion-molecule plus an electron, such as:

$$Xe_2^* + h\nu \rightarrow Xe_2^+ + e \qquad [1]$$

Such excited state absorption limited the efficiency for the VUV rare gas excimers, even though they have very high formation and fluorescence efficiency.

Rare gas halide excimer lasers evolved from 'chemi-luminescence' chemical kinetics studies which were looking at the reactive quenching of rare gas metastable excited states in flowing after glow experiments. Broadband emission in reactions with halogen molecules was observed in these experiments, via reactions such as:

$$Xe^* + Cl_2 \rightarrow XeCl^* + Cl \qquad [2]$$

At low pressure the emissions from molecules, such as $XeCl^*$, are broad in the emissions bandwidth. Shortly after these initial observations, the laser community began examination of these species in high-pressure, electron-beam excited mixtures. The results differed remarkably from the low-pressure experiments and those for the earlier excimers such as Xe_2^*. The spectrum was shifted well away from the VUV emission. Moreover, the emission was much sharper at high pressure, though still a continuum with a bandwidth of the order of 4 nm. A basis for understanding is sketched in the potential energy curves of Figure 4. In this schematic we identify the rare gas atoms by Rg, excited rare gas atoms by

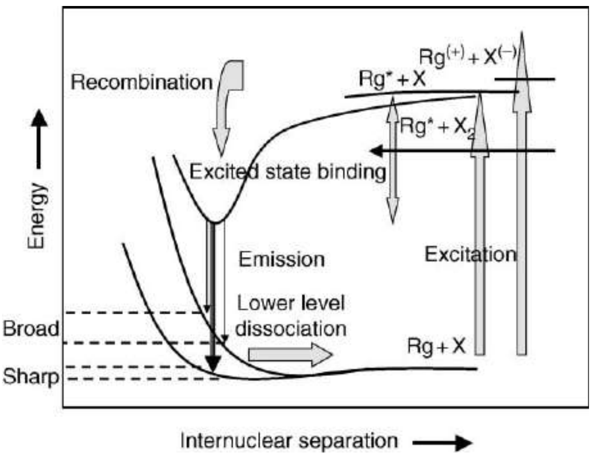

Figure 4 The potential energy curves for rare gas halides are sketched here. Both relatively sharp continuum emission is observed along with broad bands corresponding to different lower levels. The excited ion pair states consist of 3 distinct levels, two that are very close in energy and one that is shifted up from the upper laser level (the B state) by the spin orbit splitting of the rare gas ion.

Rg^*, and halogen atoms by X. Ions play a very important role in the binding and reaction chemistry, leading to excited laser molecules.

The large shift in wavelength from the VUV of rare gas dimer excimers, ~308 nm in $XeCl^*$ versus 172 nm in Xe_2^*, is due to the fact that the binding in the excited state of the rare gas halide is considerably stronger. The sharp and intense continuum at high pressure ($> \sim 100$ torr total pressure) is due to the fact that the 'sharp' transition terminates on a 'flat' portion of the lower-level potential energy curve. Indeed, in XeCl and XeF, the sharp band laser transition terminates on a slightly bound portion of the lower-level potential energy curve. Both the sharp emissions and broad bands are observed, due to the fact that some transitions from the excited levels terminate on a second, more repulsive potential energy curve.

An understanding of the spectroscopy and kinetic processes in rare gas plus halogen mixtures is based on the recognition that the excited state of a rare gas halide molecule is very similar to the ground state of the virtually isoelectronic alkali halide. The binding energy and mechanism of the excited state is very close to that of the ground state of the most similar alkali halide. Reactions of excited state rare gases are very similar to those of alkali atoms. For example, Xe^* and Cs are remarkably similar, from a chemistry and molecular physics point of view, as they have the same outer shell electron configuration and similar ionization potentials. The ionization potential of Xe^* is similar to that of cesium (Cs). Cs, and all the other alkali atoms, react rapidly with halogen molecules. Xenon metastable excited atoms react rapidly with halogen molecules. The alkali atoms all form ionically bonded ground states with halogens. The rare gas halides are effectively strongly bonded ion pairs, $Xe^{(+)}X^{(-)}$, that are very strongly bonded relative to the excited states that yield them. The difference between, for example, $XeCl^*$ and CsCl, is that $XeCl^*$ radiates in a few nanoseconds while CsCl is a stable ground-state molecule. The binding energy for these ionic bonded molecules is much greater than the more covalent type of bonding that is found in the rare gas excimer excited states. Thus XeI^*, which is only one electron different (in the I) than Xe_2^*, emits at 254 nm instead of 172 nm.

The first observed and most obvious connection to alkali atoms is in formation chemistry. Mostly, rare gas excited states react with halogen molecules rapidly. The rare gas halide laser analog reaction sequence is shown below, where some form of electric discharge excitation kicks the rare gas atom up to an excited state, so that it can react like an alkali atom, eqn [3]:

$$e_{\text{fast}} + \text{Kr} \rightarrow \text{Kr}^* + e_{\text{slow}} \qquad [3]$$

$$\text{Kr}^* + \text{F}_2 \rightarrow \text{KrF}^*(v, J)_{\text{high}} + \text{F} \qquad [4]$$

There are subtle and important cases where the neutral reaction, such as that shown in eqn [2] or [4], is not relevant as it is too slow compared to other processes. Note that the initial reaction does not yield the specific upper levels for the laser but states that are much higher up in the potential well of the excited rare gas halide excimer molecule. Further collisions with a buffer gas are needed to relax the states to the vibrational levels that are the upper levels for the laser transition. Such relaxation at high pressure can be fast, which leads the high-pressure spectrum to be much sharper than those first observed in flowing after-glow experiments.

The excited states of the rare gas halides are ion pairs bound together for their brief radiative lifetime. This opens up a totally different formation channel, indeed the one that often dominates the mechanism, ion–ion recombination. Long before excimer lasers were studied, it was well known that halogen molecules would react with 'cold' electrons (those with an effective temperature less than a few eV) in a process called attachment. The sequence leading to the upper laser level is shown below for the Kr/F_2 system, but the process is generically identical in other rare gas halide mixtures.

KrF^* formation kinetics via ion channel:

$$\text{Ionization:} \qquad e + Kr \rightarrow Kr^+ + 2e \qquad [5]$$

$$\text{Attachment:} \qquad e + F_2 \rightarrow F^{(-)} + F \qquad [6]$$

Ion–ion recombination:

$$Kr^+ + F^{(-)} + M \rightarrow KrF^*(v, J)_{\text{high}} + M \qquad [7]$$

Relaxation:

$$KrF^*(v, J)_{\text{high}} + M \rightarrow KrF^*(v, J)_{\text{low}} + M \qquad [8]$$

Finally, there is one other formation mechanism, displacement, that is unique to the rare gas halides. In this process, a heavier rare gas will displace a lighter rare gas ion in an excited state to form a lower-energy excited state:

$$ArF^* + Kr \rightarrow KrF^*(v, J)_{\text{high}} + Ar \qquad [9]$$

In general, the most important excimer lasers tend to have all of these channels contributing to excited state formation for optimized mixtures.

These kinetic formation sequences express some important points regarding rare gas halide lasers. The source of the halogen atoms for the upper laser level disappears via both attachment reactions with electrons and by the reaction with excited states. The halogen atoms eventually recombine to make molecules, but at a rate too slow for sufficient continuous wave (cw) laser gain. For excimer-based lamps, however, laser gain is not a criterion and very efficient excimer lamps can be made. Another point to note is that forming the inversion requires transient species (and in some cases halogen fuels) that absorb the laser light. Net gain and extraction efficiency are a trade-off between pumping rate, an increase in which often forms more absorbers and absorption itself. Table 2 provides a listing of some of the key absorbers. Finally, the rare gas halide excited states can be rapidly quenched in a variety of processes. More often than not, the halogen molecule and the electrons in the discharge react with the excited states of the rare gas halide, removing the contributors to gain. We show, in Table 3, a sampling of some of the types of formation and quenching reactions with their kinetic rate constant parameters. Note that one category of reaction, three-body quenching, is unique to excimer lasers. Note also that the three-body recombination of ions has a rate coefficient that is effectively pressure dependent. At pressures above ~3 atm, the diffusion of ions toward each other is slowed and the effective recombination rate constant decreases.

Though the rare gas halide excited states can be made with near-unity quantum yield from the primary excitation, the intrinsic laser efficiency is not this high. The quantum efficiency is only ~25% (recall rare gas halides all emit at much longer wavelengths than rare gas excimers); there is inefficient extraction due to losses in the gas and windows. In practice there are also losses in coupling the power into the laser, and wasted excitation during the finite start-up time. The effect of losses is shown in Table 4 and Figure 5 where the pathways are charted and losses identified. In KrF, it takes ~20 eV to make a rare gas ion in an electron beam excited mixture,

Table 2 Absorbers in rare gas halide lasers

Species	*XeF*	*XeCl*	*KrF*	*ArF*
Laser λ (nm)	351, 353	308	248	193
F_2 absorption σ (cm^2)	8×10^{-21}	NA	1.5×10^{-20}	
Cl_2 absorption σ (cm^2)	NA	1.7×10^{-19}	NA	NA
HCl absorption σ (cm^2)	NA	Nil	NA	NA
$F^{(-)}$ absorption σ (cm^2)	2×10^{-18}	NA	5×10^{-18}	5×10^{-18}
$Cl^{(-)}$ absorption σ (cm^2)	NA	2×10^{-17}	NA	NA
$Rg_2^{(+)}$ diatomic ion absorption σ (cm^2)^a	Ne: $\sim 10^{-18}$ Ar: $\sim 3 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-17}$	Ne: $\sim 10^{-17}$ Ar: $\sim 4 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-18}$	Ne: $\sim 2.5 \times 10^{-17}$ Ar: $\sim 2 \times 10^{-18}$ Kr $< 10^{-18}$	Ne: $\sim 10^{-18}$ Ar: NIL
Other transient absorbers	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}-10^{-17}$	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}-10^{-17}$	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}-10^{-17}$	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}-10^{-17}$
Other	Lower laser level absorbs unless heated	Weak lower level absorption $\sigma \sim 4 \times 10^{-16}$	Windows and dust can be an issue	Windows and dust can be an issue

[a] Note that the triatomic rare gas halides of the form Rg_2X will exhibit similar absorption as the diatomic rare gas ions as the triatomic rare gas halides of this form are essentially ion pairs of an $Rg_2^{(+)}$ ion and the halide ion.

Table 3 Typical formation and quenching reactions and rate coefficients

Formation	
Rare gas plus halogen molecule	$Kr^* + F_2 \rightarrow KrF^*(v, J) + F$ $k \sim 7 \times 10^{-10}$ cm^3/sec; Branching ratio to $KrF^* \sim 1$ $Ar^* + F_2 \rightarrow ArF^*(v, J) + F$ $k \sim 6 \times 10^{-10}$ cm^3/sec; Branching ratio to $ArF^* \sim 60\%$ $Xe^* + F_2 \rightarrow XeF^*(v, J) + F$ $k \sim 7 \times 10^{-10}$ cm^3/sec; Branching ratio to $XeF^* \sim 1$ $Xe^* + HCl(v = 0) \rightarrow Xe + H + Cl$ $k \sim 6 \times 10^{-10}$ cm^3/sec; Branching ratio to $XeCl^* \sim 0$ $Xe^* + HCl(v = 1) \rightarrow XeCl^* + H$ $k \sim 2 \times 10^{-10}$ cm^3/sec; Branching ratio to $XeCl^* \sim 1$
Ion–ion recombination	$Ar^{(+)} + F^{(-)} + Ar \rightarrow ArF^*(v, J)$ $k \sim 1 \times 10^{-25}$ cm^6/sec at 1 atm and below $k \sim 7.5 \times 10^{-26}$ cm^6/sec at 2 atm; Note effective 3-body rate constant decreases further as pressure goes over ~ 2 atm $Kr^{(+)} + F^{(-)} + Kr \rightarrow KrF^*(v, J)$ $k \sim 7 \times 10^{-26}$ cm^6/sec at 1 atm and below; rolls over at higher pressure
Displacement	$ArF^*(v, J) + Kr \rightarrow KrF^*(v, J) + Ar$ $k \sim 2 \times 10^{-10}$ cm^3/sec; Branching ratio to $KrF^* \sim 1$
Quenching	
Halogen molecule	$XeCl^* + HCl \rightarrow Xe + Cl + HCl$ $k \sim 5.6 \times 10^{-10}$ cm^3/sec; $KrF^* + F_2 \rightarrow Kr + F + F$ $k \sim 6 \times 10^{-10}$ cm^3/sec; All halogen molecule quenching have very high reaction rate constants
3-body inert gas	$ArF^* + Ar + Ar \rightarrow Ar_2F^* + Ar$ $k \sim 4 \times 10^{-31}$ cm^6/sec $KrF^* + Kr + Ar \rightarrow Kr_2F^* + Ar$ $k \sim 6 \times 10^{-31}$ cm^6/sec $XeCl^* + Xe + Ne \rightarrow Xe_2Cl^* + Ne$ $k \sim 1 \times 10^{-33}$ cm^6/sec; 4×10^{-31} cm^6/sec with Xe as 3rd body These reactions yield a triatomic excimer at lower energy that can not be recycled back to the desired laser states
Electrons	$XeCl^* + e \rightarrow Xe + Cl + e$ $k \sim 2 \times 10^{-7}$ cm^3/sec In 2-body quenching by electrons, the charge of the electron interacts with the dipole of the rare gas halide ion pair at long range; above value is typical of others
2-body inert gas	$KrF^* + Ar \rightarrow Ar + Kr + F$ $k \sim 2 \times 10^{-12}$ cm^3/sec $XeCl^* + Ne \rightarrow Ne + Xe + Cl$ $k \sim 3 \times 10^{-13}$ cm^3/sec 2-body quenching by rare gases is typically slow and less important than the reactions noted above

somewhat less in an electric discharge. The photon release is ~5 eV; the quantum efficiency is only 25%. Early laser efficiency measurements (using electron beam excitation) showed laser efficiency in the best cases slightly in excess of 10%, relative to the deposited energy. The discrepancy relative to the quantum efficiency is due to losses in the medium. For the sample estimate in Table 4 and Figure 5 we show approximate density of key species in the absorption chain, estimated on a steady-state approximation of the losses in a KrF laser for an excitation rate of order 1 MW/cc, typical of fast discharge lasers. The net small signal gain is of the order of 5%/cm. The key absorbers are the parent halogen molecule, F_2, the fluoride ion $F^{(-)}$, and the rare gas excited state atoms. A detailed pulsed kinetics code will provide slightly different results due to transient kinetic effects and the intimate coupling of laser extraction to quenching and absorber dynamics. The ratio of gain to loss is usually in the range of 7 to 20, depending on the actual mixture used and the rare gas halide wavelength. The corresponding extraction efficiency is then limited to the range of ~50%. The small signal gain, g_o, is related to the power deposition

Table 4 Typical ground and excited state densities shown for KrF at $P_{in} \sim 1$ MW/cm^3

Species	*Amount particles/cm^3*	*Absorption σ (cm)^2*	*Loss %/cm*
F_2	$\sim 1.2 \times 10^{17}$	1.5×10^{-20}	0.18
Kr	$\sim 3 \times 10^{18}$	–	–
Buffer	4×10^{19}	–	–
$F^{(-)}$	$\sim 10^{14}$	5×10^{-18}	0.05
Rg^*, Rg^{**}	1×10^{15}	$\langle 10^{-18} \rangle$ depends on ratio of Rg^{**}/Rg^*	0.1
KrF (before dissociating)	$\sim 10^{12}$	2×10^{-16}	0.02
Rg_2F^*	8×10^{14}, Note this species density decreases as flux grows	1×10^{-18}	0.08
KrF^* (B)	2.5×10^{14}	2×10^{-16}	5 (gain)

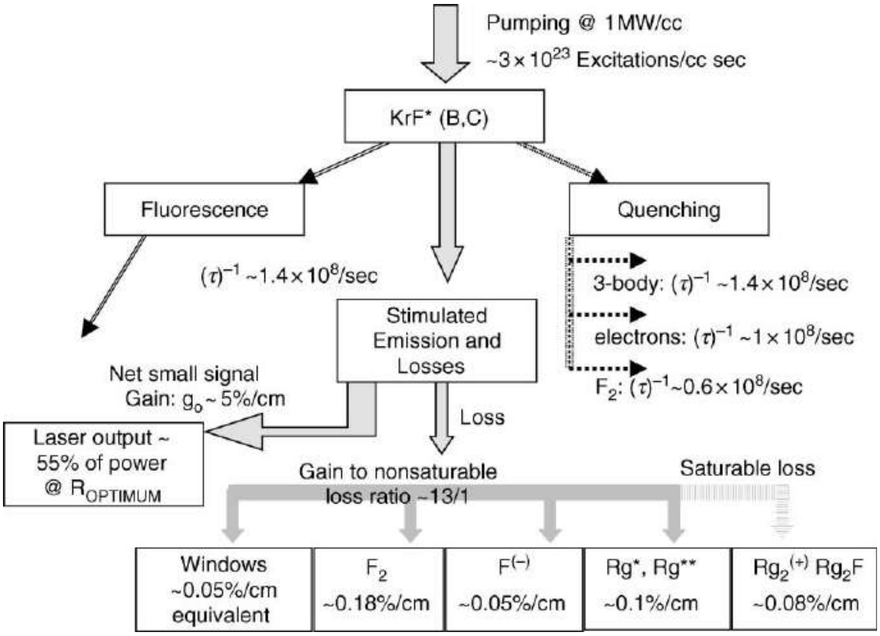
Pumping @ 1MW/cc
~3×10²³ Excitations/cc sec
KrF* (B,C)
Fluorescence
Quenching
(τ)⁻¹ ~1.4×10⁸/sec
Stimulated Emission and Losses
3-body: (τ)⁻¹ ~1.4×10⁸/sec
electrons: (τ)⁻¹ ~1×10⁸/sec
F₂: (τ)⁻¹~0.6×10⁸/sec
Net small signal Gain: g₀~5%/cm
Loss
Laser output ~ 55% of power @ R_OPTIMUM
Gain to nonsaturable loss ratio ~13/1
Saturable loss
Windows ~0.05%/cm equivalent
F₂ ~0.18%/cm
F⁽⁻⁾ ~0.05%/cm
Rg*, Rg** ~0.1%/cm
Rg₂⁽⁺⁾ Rg₂F ~0.08%/cm

Figure 5 The major decay paths and absorbers that impact the extraction of excited states of rare gas halide excimers. Formation is via the ion–ion recombination, excited atom reaction with halogens and displacement reactions, not shown. We note for a specific mixture of KrF the approximate values of the absorption by the transient and other species along with the decay rates for the major quenching processes. Extraction efficiency greater than 50% is quite rare.

rate per unit volume, $P_{\text{deposited}}$, by eqn [10].

$$P_{\text{deposited}} = g_0 E^* (\tau_{\text{upper}} \eta_{\text{pumping}} \eta_{\text{branching}} f \sigma)^{-1} \qquad [10]$$

E^* is the energy per excitation, ~20 eV or 3.2×10^{-18} J/excitation, f is the fraction of excited molecules in the upper laser level, ~40% due to KrF^* in higher vibrational states and the slowly radiating 'C' state. The laser cross-section is σ, and the η terms are efficiencies for getting from the primary excitation down to the upper laser level. The upper state lifetime, τ_{upper}, is the sum of the inverses of radiative and quenching lifetimes. When the laser cavity flux is sufficiently high, stimulated emission decreases the flow of power into the quenching processes.

The example shown here gives a ratio of net gain to the nonsaturating component of loss of 13/1. Of the excitations that become KrF^*, only ~55% or so will be extracted in the best of cases. For other less ideal rare gas halide lasers, the gain to loss ratio, the extraction efficiency, and the intrinsic efficiency are all lower. Practical discharge lasers have lower practical wall plug efficiency due to a variety of factors, discussed below.

Discharge Technology

Practical discharge lasers use fast-pulse discharge excitation. In this scheme, an outgrowth of the CO_2 TEA laser, the gas is subjected to a rapid high-voltage pulse from a low-inductance pulse forming line or network, PFL or PFN. The rapid pulse serves to break down the gas, rendering it conductive with an electron density of $> \sim 10^{14}$ electrons/cm^3. As the gas becomes conductive, the voltage drops and power is coupled into the medium for a short duration, typically less than 50 nsec. The discharge laser also requires some form of 'pre-ionization', usually provided by sparks or a corona-style discharge on the side of the discharge electrodes. The overall discharge process is intrinsically unstable, and pump pulse duration needs to be limited to avoid the discharge becoming an arc.

Moreover, problems with impedance matching of the PFN or PFL with the conductive gas lead to both inefficiency and extra energy that can go into post-discharge arcs, which both cause component failure and produce metal dust which can give rise to window losses. A typical discharge laser has an aperture of a few cm^2 and a gain length of order 80 cm and produces outputs in the range of 100 mJ to 1 J. The optical pulse duration tends to be ~20 nsec, though the electrical pulse is longer due to the finite time needed for the gain to build up and the stimulated emission process to turn spontaneously emitted photons into photons in the laser cavity.

The pre-ionizer provides an initial electron density on the order of 10^8 electrons/cm^3. With spark pre-ionization or corona pre-ionization, it is difficult to make a uniform discharge over an aperture greater than a few cm^2. For larger apertures, one uses X-rays for pre-ionization. X-ray pre-ionized discharge excimer lasers have been scaled to energy well over 10 J per pulse and can yield pulse durations over 100 nsec. The complexity and expense of the X-ray approach, along with the lack of market for lasers in this size range, resulted in the X-ray approach remaining in the laboratory.

Aside from the kinetic issues of the finite time needed for excitation to channel into the excimer upper levels and the time needed for the laser to 'start up' from fluorescent photons, these lasers have reduced efficiency due to poor matching of the pump source to the discharge. Ideally, one would like to have a discharge such that when it is conductive the voltage needed to sustain the discharge is less than half of that needed to break the gas down. Regrettably for the rare gas halides (and in marked distinction to the CO_2 TEA laser) the voltage that the discharge sustains in a quasi-stable mode is ~20% of the breakdown voltage, perhaps even less, depending on the specific gases and electrode shapes.

Novel circuits, which separate the breakdown phase from the fully conductive phase, can readily double the efficiency of an excimer discharge laser, though this approach is not necessary in many applications. The extra energy that is not dissipated in the useful laser discharge often results in arcing or 'hot' discharge areas after the main laser pulse, leading to electrode sputtering, dust that finds its way to windows, and the need to service the laser after ~10^8 pulses. When one considers the efficiency of the power conditioning system, along with the finite kinetics of the laser start-up process in short pulses, and the mismatch of the laser impedance to the PFN/PFL impedance, lasers that may have 10% intrinsic efficiency in the medium in e-beam excitation may only have 2% to 3% wall plug efficiency in a practical discharge laser.

A key aspect of excimer technology is the need for fast electrical circuits to drive the power into the discharge. This puts a significant strain on the switch that is used to start the process. While spark gaps, or multichannel spark gaps, can handle the required current and current rate of rise, they do not present a practical alternative for high-pulse rate operation. Thyratrons and other switching systems, such as solid state switches, are more desirable than a spark gap, but do not offer the desired needed current rate of rise. To provide the pumping rate of order 1 MW/cm^3 needed for enough gain to turn the laser on before the discharge becomes unstable, the current needs to rise to a value of the order of 20 kA in a time of ~20 nsec; the current rate of rise is ~10^{12} A/sec. This rate of rise will limit the lifetime of the switch that drives the power into the laser gas. As a response to this need, magnetic switching and magnetic compression circuits were developed so that the lifetime of the switch can be radically increased. In the simplest cases, one stage of magnetic compression is used to make the current rate of rise to be within the range of a conventional thyratron switch as used in radar systems. Improved thyratrons and a simple magnetic assist also work. For truly long operating lifetimes, multiple stages of magnetic switching and compression are used, and with a transformer one can use a solid state diode as the start switch for the discharge circuit. A schematic of an excimer pulsed power circuit, that uses a magnetic assist with a thyratron and one stage of magnetic compression, is shown in Figure 6. The magnetic switching approach can also be used to provide two separate circuits to gain high efficiency by having one circuit to break down the gas (the spiker) and a second, low-impedance 'sustainer' circuit, to provide the main power flow. In this approach, the laser itself can act as part of the circuit by holding off the sustaining voltage for a μ-sec or so during the charge cycle and before the spiker breaks down the gas. A schematic of such a circuit is shown in Figure 7. Using this type of circuit, the efficiency of a XeCl laser can be over 4% relative to the wall plug. A wide variety of technological twists on this approach have been researched.

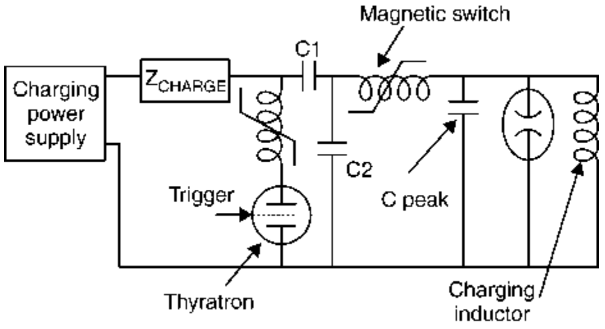

Figure 6 One of the variants on magnetic switching circuits that have been used to enhance the lifetime of the primary start switch. A magnetic switch in the form of a simple 1 turn inductor with a saturable core magnetic material such as a ferrite allows the thyratron to turn on before major current flows. Other magnetic switches may be used in addition to peak up the voltage rate of rise on the laser head or to provide significant compression. For example one may have the major current flow in the thyratron taking place on the 500 nsec time scale while the voltage pulse on the discharge is in the 50 nsec duration. A small current leaks through and is accommodated by the charging inductor.

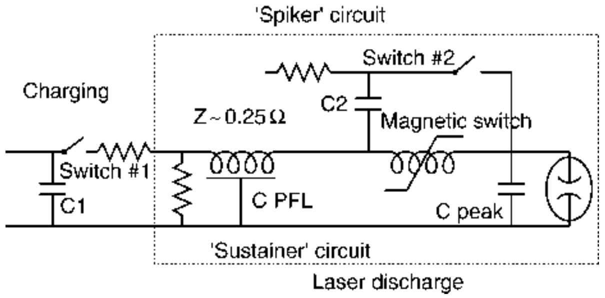

Figure 7 For the ultimate in efficiency the circuit can be arranged to provide a leading edge spike that breaks the gas down using a high impedance 'spiker' and a separate circuit that is matched to the discharge impedance when it is fully conductive. High coupling efficiency reduces waste energy that can erode circuit and laser head components via late pulse arcs. The figure shows a magnetic switch as the isolation but the scheme can be implemented with a second laser head as the isolation switch, a rail gap (not appropriate for long life) or a diode at low voltage.

Excimer lasers present some clear differences in resonator design compared to other lasers. The apertures are relatively large, so one does not build a TEM_{00} style cavity for good beam quality and expect to extract any major portion of the energy available. The typical output is divergent and multimode. When excimer lasers are excited by a short pulse discharge, ~30 nsec, the resulting gain duration is also short, ~20 nsec, limiting the number of passes a photon in a resonator can have in the gain. The gain duration of ~20 nsec only provides ~3.5 round trips of a photon in the cavity during the laser lifetime, resulting in the typical high-order multimode beam. Even with an unstable resonator, the gain duration is not long enough to collapse the beam into the lowest-order diffraction-limited output. If line narrowing is needed, and reasonable efficiency is required, an oscillator amplifier configuration is often used with the oscillator having an appropriate tuning resonator in the cavity. If both narrow spectral linewidth and excellent beam quality are needed one uses a seeded oscillator approach where the narrowband oscillator is used to seed an unstable resonator on the main amplifier. By injecting the seed into the unstable resonator cavity, a few nsec before the gain is turned on, the output of the seeded resonator can be locked in wavelength, bandwidth, and provide good beam quality. Long pulse excimer lasers, such as e-beam excited lasers or long pulse X-ray pre-ionized discharge devices can use conventional unstable resonator technology with good results.

Excimer lasers are gas lasers that run at both high pressure and high instantaneous power deposition rates. During a discharge pulse, much of the halogen bearing 'fuel' is burned out by the attachment and reactive processes. The large energy deposition per pulse means that pressure waves are generated. All of these combine with the characteristic device dimensions to require gas flow to recharge the laser mixture between pulses. For low pulse rates, less than ~200 Hz, the flow system need not be very sophisticated. One simply needs to flush the gas from the discharge region with a flow having a velocity of the order of 5 m/sec for a typical discharge device. The flow is transverse to the optical and discharge directions. At high pulse rates, the laser designer needs to consider flow in much more detail to minimize the power required to push the gas through the laser head. Circulating pressure waves need to be damped out so that the density in the discharge region is controlled at the time of the next pulse. Devices called acoustic dampers are placed in the sides of the flow loop to remove pressure pulses. A final subtlety occurs in providing gas flow over the windows. The discharge, post pulse arcs, and reactions of the halogen source with the metal walls and impurities creates dust. When the dust coats the windows, the losses in the cavity increase, lowering efficiency. By providing a simple gas flow over the window, the dust problem can be ameliorated. For truly long life, high-reliability lasers, one needs to take special care in the selection of materials for electrodes and insulators and avoid contamination, by assembling in clean environments.

See also

Lasers: Carbon Dioxide Lasers. **Laser-Induced Damage of Optical Materials**. **Nonlinear Optics,**

Applications: Pulse Compression via Nonlinear Optics; Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Ballanti S, Di Lazzaro P, Flora F, *et al.* (1998) Ianus the 3-electrode laser. *Applied Physics B* 66: 401–406.

Brau CA (1978) Rare gas halogen lasers. In: Rhodes CK (ed.) *Excimer Lasers, Topics of Applied Physics*, vol. 30. Berlin: Springer Verlag.

Brau CA and Ewing JJ (1975) Emission spectra of XeBr, XeCl, XeF, and KrF. *Journal of Chemical Physics* 63: 4640–4647.

Ewing JJ (2000) Excimer laser technology development. *IEEE Journal of Selected Topics in Quantum Electronics* 6(6): 1061–1071.

Levatter J and Lin SC (1980) Necessary conditions for the homogeneous formation of pulsed avalanche discharges at high pressure. *Journal of Applied Physics* 51: 210–222.

Long WH, Plummer MJ and Stappaerts E (1983) Efficient discharge pumping of an XeCl laser using a high voltage prepulse. *Applied Physics Letters* 43: 735–737.

Rhodes CK (ed.) (1979) *Excimer Lasers, Topics of Applied Physics*, vol. 30. Berlin: Springer-Verlag.

Rokni M, Mangano J, Jacob J and Hsia J (1978) Rare gas fluoride lasers. *IEEE Journal of Quantum Electronics* QE-14: 464–481.

Smilansksi I, Byron S and Burkes T (1982) Electrical excitation of an XeCl laser using magnetic pulse compression. *Applied Physics Letters* 40: 547–548.

Taylor RS and Leopold KE (1994) Magnetic-spiker excitation of gas discharge lasers. *Applied Physics B, Lasers and Optics* 59: 479–509.

Free Electron Lasers

A Gover, Tel-Aviv University, Tel-Aviv, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The Free Electron Laser (FEL) is an exceptional kind of laser. Its active medium is not matter, but charged particles (electrons) accelerated to high energies, passing in vacuum through a periodic undulating magnetic field. This distinction is the main reason for the exceptional properties of FEL: operating at a wide range of wavelengths – from mm-wave to X-rays with tunability, high power, and high efficiency.

In this article we explain the physical principles of FEL operation, the underlying theory and technology of the device and various operating schemes, which have been developed to enhance performance of this device.

The term 'Free Electron Laser' was coined by John Madey in 1971, pointing out that the radiative transitions of the electrons in this device are between free space (more correctly – unbound) electron quantum states, which are therefore states of continuous energy. This is in contrast to conventional atomic and molecular lasers, in which the electron performs radiative transition between bound (and therefore of distinct energy) quantum states. Based on these theoretical observations, Madey and his colleagues in Stanford University demonstrated FEL operation first as an amplifier (at $\lambda = 10.6\ \mu\text{m}$) in 1976, and subsequently as an oscillator (at $\lambda = 3.4\ \mu\text{m}$) in 1980.

From the historical point of view, it turned out that Madey's invention was essentially an extension of a former invention in the field of microwave-tubes technology – the Ubitron. The Ubitron, a mm-wave electron tube amplifier based on a magnetic undulator, was invented and developed by Philips and Enderbry who operated it at high power levels in 1960. The early Ubitron development activity was not noticed by the FEL developers because of the disciplinary gap, and largely because its research was classified at the time. Renewed interest in high-power mm-wave radiation emission started in the 1970s, triggered by the development of pulsed-line generators of 'Intense Relativistic Beams' (IRB). This activity, led primarily by plasma physicists in the defense establishment laboratories of Russia (mostly IAP in Gorky – Nizhny Novgorod) and the US (mostly N.R.L. – DC) led to development of high-gain high-power mm-wave sources independently of the development of the optical FEL. The connection between these devices and between them to conventional microwave tubes (as Traveling Wave Tubes – TWT) and other electron beam radiation schemes, like Cenenkov and Smith-Purcell radiation that may also be considered FELs, was revealed in the mid-1970s, starting with the theoretical works of P. Spangle, A. Gover and A. Yariv who identified that all these devices satisfy the same dispersion equation as the TWT derived by John Pierce in the 1940s. Thus, the optical FEL could be conceived as a kind of immense electron tube, operating with a high-energy electron beam in the low gain regime of the Pierce TWT dispersion equation.

The extension of the low-gain FEL theory to the general ‘electron-tube’ theory is important because it led to development of new radiation schemes and new operating regimes of the optical FEL. This was exploited by physicists in the discipline of accelerator physics and synchrotron radiation, who identified, starting with the theoretical works of C. Pellegrini and R. Bonifacio in the early 1980s, that high-current, high-quality electron beams, attainable with further development of accelerators technology, could make it possible to operate FELs in the high-gain regime, even at short wavelengths (vacuum ultra-violet – VUV and soft X-ray) and that the high-gain FEL theory can be extended to include amplification of the incoherent synchrotron spontaneous emission (shot noise) emitted by the electrons in the undulator. These led to the important development of the ‘self (synchrotron) amplified spontaneous emission (SASE) FEL’, which promised to be an extremely high brightness radiation source, overcoming the fundamental obstacles of X-ray lasers development: lack of mirrors (for oscillators) and lack of high brightness radiation sources (for amplifiers).

A big boost to the development of FEL technology was given during the period of the American ‘strategic defense initiative – SDI’ (Star-Wars) program in the mid-1980s. The FEL was considered one of the main candidates for use in a ground-based or space-based ‘directed energy weapon – DEW’, that can deliver megawatts of optical power to hit attacking missiles. The program led to heavy involvement of major American defense establishment laboratories (Lawrence–Livermore National Lab, Los-Alamos National Lab) and contracting companies (TRW, Boeing). Some of the outstanding results of this effort were demonstration of the high-gain operation of an FEL amplifier in the mm-wavelength regime, utilizing an Induction Linac (Livermore, 1985), and demonstration of enhanced radiative energy extraction efficiency in FEL oscillator, using a ‘tapered wiggler’ in an RF-Linac driven FEL oscillator (Los-Alamos, 1983). The program has not been successful in demonstrating the potential of FELs to operate at the high average power levels needed for DEW applications. But after the cold-war period, a small part of the program continues to support research and development of medical FEL application.

Principles of FEL Operation

Figure 1 displays schematically an FEL oscillator. It is composed of three main parts: an electron accelerator, a magnetic wiggler (or undulator), and an optical resonator.

Without the mirrors, the system is simply a synchrotron undulator radiation source. The electrons in the injected beam oscillate transversely to their propagation direction z, because of the transverse magnetic Lorenz force:

$$\mathbf{F}_\perp = -ev_z\hat{\mathbf{e}}_z \times \mathbf{B}_\perp \tag{1}$$

In a planar (linear) wiggler, the magnetic field on axis is approximately sinusoidal:

$$\mathbf{B}_\perp = B_\mathrm{w}\hat{e}_y \cos k_\mathrm{w} z \tag{2}$$

In a helical wiggler:

$$\mathbf{B}_\perp = B_\mathrm{w}(\hat{\mathbf{e}}_y \cos k_\mathrm{w} z + \hat{\mathbf{e}}_x \sin k_\mathrm{w} z) \tag{3}$$

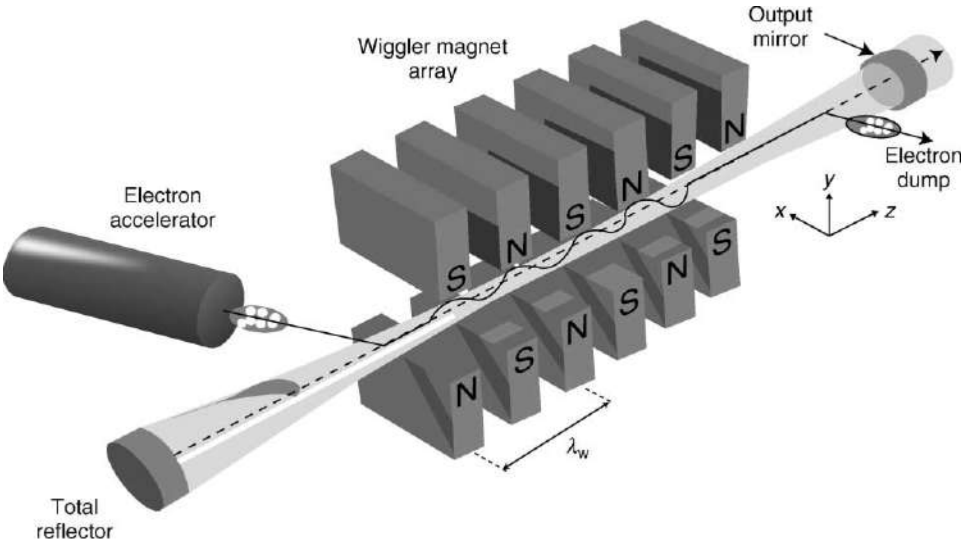

Figure 1 Components of a FEL-oscillator. (Reproduced from Benson SV (2003) Free electron lasers push into new frontiers. *Optics and Photonics News* 14: 20–25. Illustration by Jaynie Martz.)

In either case, if we assume constant (for the planar wiggler – only on the average) axial velocity, then $z = v_z t$. The frequency of the transverse force and the mechanical oscillation of the electrons, as viewed transversely in the laboratory frame of reference, is:

$$\omega_{0s} = k_w v_z = 2\pi \frac{v_z}{\lambda_w} \qquad [4]$$

where $\lambda_w = 2\pi/k_w$ is the wiggler period.

The oscillating charge emits an electromagentic radiation wavepacket. In a reference frame moving with the electrons, the angular radiation pattern looks exactly like dipole radiation, monochromatic in all directions (except for the frequency-line-broadening due to the finite oscillation time, i.e., the wiggler transit time). In the laboratory reference-frame the radiation pattern concentrates in the propagation ($+z$) direction, and the Doppler up-shifted radiation frequency depends on the observation angle Θ relative to the z-axis:

$$\omega_0 = \frac{\omega_{0s}}{1 - \beta_z \cos\Theta} \qquad [5]$$

On axis ($\Theta = 0$), the radiation frequency is:

$$\omega_0 = \frac{ck_w\beta_z}{1 - \beta_z} = (1 + \beta_z)\beta_z\gamma_z^2 ck_w \cong 2\gamma_z^2 ck_w \qquad [6]$$

where $\beta_z \equiv v_z/c$, $\gamma_z \equiv (1 - \beta_z^2)^{-1/2}$ are the axial (average) velocity and the axial Lorenz factor, respectively, and the last part of the equation is valid only in the (common) highly relativistic limit $\gamma_z \gg 1$.

Using the relations $\beta_z^2 + \beta_\perp^2 = \beta^2$, $\beta_\perp = a_w/\gamma$, one can express γ_z:

$$\gamma_z = \frac{\gamma}{1 + a_w^2/2} \qquad [7]$$

(this is for a linear wiggler, in the case of a helical wiggler the denoninator is $1 + a_w^2$):

$$\begin{aligned}\gamma &\equiv (1 - \beta^2)^{-1/2} = 1 + \frac{\varepsilon_k}{mc^2} \\ &= 1 + \varepsilon_k\,[\text{MeV}]/0.511 \qquad [8]\end{aligned}$$

and a_w – (also termed K) 'the wiggler parameter' is the normalized transverse momentum:

$$a_w = \frac{eB_w}{k_w mc} = 0.093 B_w\,[\text{KGauss}]\lambda_w\,[\text{cm}] \qquad [9]$$

Typical values of B_w in FEL wigglers (undulators) are of the order of Kgauss', and λ_w of the order of CMs, and consequently $a_w < 1$. Considering that electron beam accelerator energies are in the range of MeV to GeV, one can appreciate from eqns [6]–[8], that a significant relativistic Doppler shift factor $2\gamma_z^2$, in the range of tens to millions, is possible. It, therefore, provides incoherent synchrotron undulator radiation in the frequency range of microwave to hard X-rays.

Synchrotron undulator radiation was studied in 1951 and since then has been a common source of VUV radiation in synchrotron facilities. From the point of view of laser physics theory, this radiation can be viewed as 'spontaneous synchrotron radiation emission' in analogy to spontaneous radiation emission by electrons excited to higher bound-electron quantum levels in atoms or molecules. Alternatively, it can be regarded as the classical shot noise radiation, associated with the current fluctuations of the randomly injected discrete charges comprising the electron beam. Evidently this radiation is incoherent, and the fields it produces average in time to zero, because the wavepackets emitted by the randomly injected electrons interfere at the observation point with random phase. However, their energies sum up and can produce substantial power.

Based on fundamental quantum-electrodynamical principles or Einstein's relations, one would expect that any spontaneous emission scheme can be stimulated. This principle lies behind the concept of the FEL, which is nothing but stimulated undulator synchrotron radiation. By stimulating the electron beam to emit radiation, it is possible, as with any laser, to generate a coherent radiation wave and extract more power from the gain medium, which in this case is an electron beam, that carries an immense amount of power. There are two kinds of laser schemes which utilize stimulation of synchrotron undulator radiation:

(i) *A laser amplifier.* In this case the mirrors in the schematic configuration of Figure 1 are not present, and an external radiation wave at frequencies within the emission range of the undulator is injected at the wiggler entrance. This requires, of course, an appropriate radiation source to be amplified and availability of sufficiently high gain in the FEL amplifier.

(ii) *A laser oscillator.* In this case an open cavity (as shown in Figure 1) or another (waveguide) cavity is included in the FEL configuration. As in any laser, the FEL oscillator starts building up its radiation from the spontaneous (synchrotron undulator) radiation which gets trapped in the resonator and amplified by stimulated emission along the wiggler. If the threshold condition is satisfied (having single path gain higher than the round trip losses), the oscillator arrives to saturation and steady state coherent operation after a short transient period of oscillation build-up.

Because the FEL can operate as a high-gain amplifier (with a long enough wiggler and an electron beam of high current and high quality), also a third mode of operation exists: self amplified spontaneous emission (SASE). In this case, the resonator mirrors in Figure 1 are not present and the undulator radiation generated spontaneously in the first sections of the long undulator is amplified along the wiggler and emitted at the wiggler exit at high power and high spatial coherence.

The Quantum-Theory Picture

A free electron, propagating in unlimited free space, can never emit a single photon. This can be proven by examining the conservation of energy and momentum conditions:

$$\varepsilon_{k_\mathrm{i}} - \varepsilon_{k_\mathrm{f}} = \hbar\omega \qquad [10]$$

$$\mathbf{k}_\mathrm{i} - \mathbf{k}_\mathrm{f} = \mathbf{q} \qquad [11]$$

that must be satisfied, when an electron in an initial free-space energy and momentum state $(\varepsilon_{k_\mathrm{i}}, \hbar k_\mathrm{i})$ makes a transition to a final state $(\varepsilon_{k_\mathrm{f}}, \hbar k_\mathrm{f})$, emitting a single photon of energy and momentum $(\hbar\omega, q)$. In free space:

$$\varepsilon_k = c\sqrt{(\hbar \mathbf{k})^2 + (mc^2)^2} \qquad [12]$$

$$\mathbf{q} = \frac{\omega}{\mathrm{c}}\hat{\mathbf{e}}_q \qquad [13]$$

and eqns [10]–[13] have only one solution, $\omega = 0$, $\mathbf{q} = 0$. This observation is illustrated graphically in the energy–momentum diagram of Figure 2a in the framework of a one-dimensional model. It appears that if both eqns [10] and [11] can be satisfied, then the phase velocity of the emitted radiation wave $v_\mathrm{ph} = \omega/q$ (the slope of the chord) will equal the electron wavepacket group velocity $v_\mathrm{g} = v_\mathrm{z}$ at some intermediate point $k^* = p^*/\hbar$:

$$v_\mathrm{ph} = \frac{\omega}{q} = \frac{\varepsilon_{k\mathrm{i}} - \varepsilon_{k\mathrm{f}}}{\hbar(k_\mathrm{i} - k_\mathrm{f})} = \left.\frac{\partial\varepsilon}{\partial p}\right|_{\mathrm{p}^*} = v_\mathrm{g} \qquad [14]$$

For a radiation wave in free space (eqn [13]), this results in $c = v_\mathrm{g}$, which contradicts special relativity.

The reason for the failure to conserve both energy and momentum in the transition is that the photon momentum $\hbar q$ it too small to absorb the large momentum shift of the electron, as it recoils while releasing radiative energy $\hbar\omega$. This observation leads to ideas on how to make a radiative transition possible:

(i) *Limit the interaction length*. If the interaction length is L, the momentum conservation

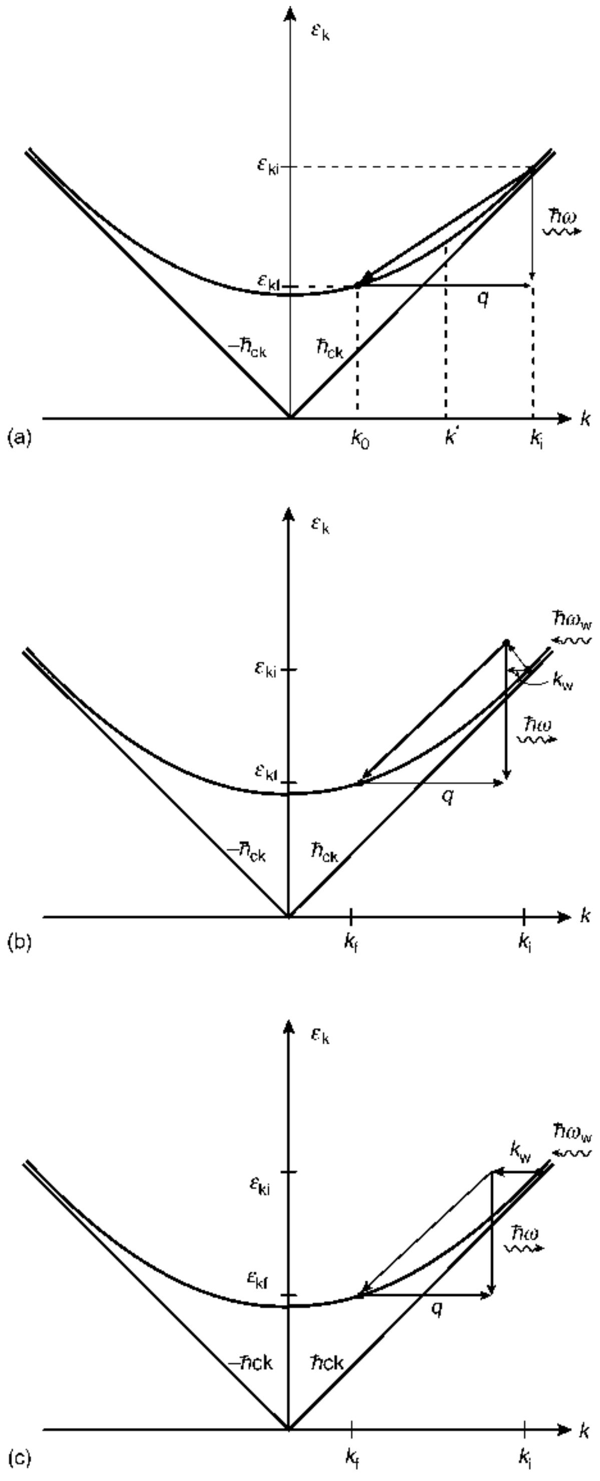
ε_k
ε_ki
$\hbar\omega$
ε_kf
q
$-\hbar_\mathrm{ck}$
$\hbar_\mathrm{ck}$
(a)
k_0
k'
k_i
k
ε_k
$\hbar\omega_\mathrm{w}$
ε_ki
k_w
$\hbar\omega$
ε_kf
q
$-\hbar_\mathrm{ck}$
$\hbar_\mathrm{ck}$
(b)
k_f
k_i
k
ε_k
$\hbar\omega_\mathrm{w}$
k_w
ε_ki
$\hbar\omega$
ε_kf
q
$-\hbar\mathrm{ck}$
$\hbar\mathrm{ck}$
(c)
k_f
k_i
k

Figure 2 Conservation of energy and momentum in forward photon emission of a free electron: (a) The slope of the tangent to the curve at intermediate point k^*, $\partial\varepsilon_k(k^*)/\partial k$ may be equal to the slope of the chord $\hbar\omega/q$ which is impossible in free space. (b) electron radiative transition made possible with an electromagnetic pump (Compton Scattering). (c) The wiggler wavenumber – k_w conserves the momentum in electron radiative transition of FEL.

condition in eqn [11] must be satisfied only within an uncertainty range $\pm\pi/L$. This makes it possible to obtain radiative emission in free electron radiation effects like 'Transition Radiation' and in microwave tubes like the Klystron.

(ii) *Propagate the radiation wave* in a 'slow wave' structure, where the phase velocity of the radiation wave is smaller than the speed of light, and satisfaction of eqn [14] is possible. For example, in the Cerenkov effect, charged particles pass through a medium (gas) with index of refraction $n > 1$. Instead of eqn [13], $\mathbf{q} = n(\omega)(\omega/c)\hat{\mathbf{e}}_q$, and consequently $q_z = n(\omega)(\omega/c)\cos\Theta_q$, where we assume radiative emission at an angle Θ_q relative to the electron propagation axis z. Substitution in eqn [14] results in the Cerenkov radiation condition $v_g n(\omega)\cos\Theta_q = 1$.

Another example for radiation emission in a slow-wave structure is the Traveling Wave Tube (TWT). In this device, a periodic waveguide of periodicity λ_w permits (via the Floquet theorem) propagation of slow partial waves (space harmonics) with increased wavenumber $q_z + mk_w$ ($m = 1, 2, \ldots$), and again eqn [14] can be satisfied.

(iii) *Rely on a 'two-photon' radiative transition.* This can be 'real photon' Compton scattering of an intense radiation beam (electromagnetic pump) off an electron beam, or 'virtual photon' scattering of a static potential, as is the case in bremsstrahlung radiation and in synchrotron–undulator radiation. The latter radiation scheme may be considered as a 'magnetic brehmsstrahlung' effect or as 'zero frequency pump' Compton scattering, in which the wiggler contributes only 'crystal momentum' $\hbar k_w$, to help satisfy the momentum conservation condition in eqn [11]. The Compton scattering scheme is described schematically for the one-dimensional (back scattering) case in Figure 3, and its conservation of energy and momentum diagram is depicted in Figure 2b (a 'real photon' (ω_w, k_w) free-space pump wave is assumed with $k_w = \omega_w/c$). The analogous diagram of a static wiggler ($\omega_w = 0$, $k_w = 2\pi/\lambda_w$) is shown in Figure 2c. It is worth noting that the effect of the incident scattered wave or the wiggler is not necessarily a small perturbation. It may modify substantially the electron energy-dispersion diagram of the free electron and a more complete 'Brillouin diagram' should be used in Figure 2c. In this sense, the wiggler may be viewed as the analogue of a one-dimensional crystal, and its period λ_w analogous to the crystal lattice constant. The momentum conservation during a radiation transition, with the aid of the wiggler 'crystal momentum' $\hbar k_w$ is quite analogous to the occurrence of vertical radiative transitions in direct bandgap semiconductors, and thus the FEL has, curiously enough, some analogy to microscopic semiconductor lasers.

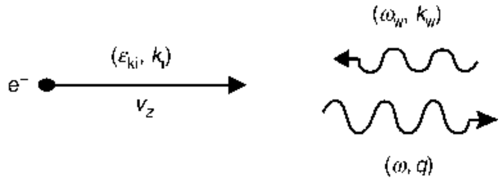

Figure 3 The scheme of backward scattering of an electromagnetic wave off an electron beam (Doppler shifted Compton scattering).

All the e-beam radiation schemes already mentioned can be turned into stimulated emission devices, and thus may be termed 'free electron lasers' in the wide sense. The theory of all of these devices is closely related, but most of the technological development was carried out on undulator radiation (Magnetic brehmsstrahlung) FELs, and the term FEL is usually reserved for this kind (though some developments of Cerenkov and Smith–Purcell FELs are still carried out).

When considering a stimulated emission device, namely enhanced generation of radiation in the presence of an external input radiation wave, one should be aware, that in addition to the emission process described by eqns [10] and [11] and made possible by one of the radiation schemes described above, there is also a stimulated absorption process. Also, this electronic transition process is governed by the conservation of energy and momentum conditions, and is described by eqns [10] and [11] with k_i and k_f exchanged.

Focusing on undulator-radiation FEL and assuming momentum conservation in the axial (z) dimension by means of the wiggler wavenumber k_w, the emission and absorption quantum transition levels and radiation frequencies are found from the solution of equations:

$$\varepsilon_{k_{zi}} - \varepsilon_{k_{zf}} = \hbar\omega_e \qquad [15a]$$

$$k_{zi} - k_{zf} = q_{ze} + k_w \qquad [15b]$$

$$\varepsilon_{k_{za}} - \varepsilon_{k_{zi}} = \hbar\omega_a \qquad [16a]$$

$$k_{za} - k_{zi} = q_{za} + k_w \qquad [16b]$$

For fixed k_w, fixed transverse momentum and given e-beam energy $\varepsilon_{k_{zi}}$ and radiation emission angle Θ_q ($q_z = (\omega/c)\cos\Theta_q$), eqns [15] and [16] have separately

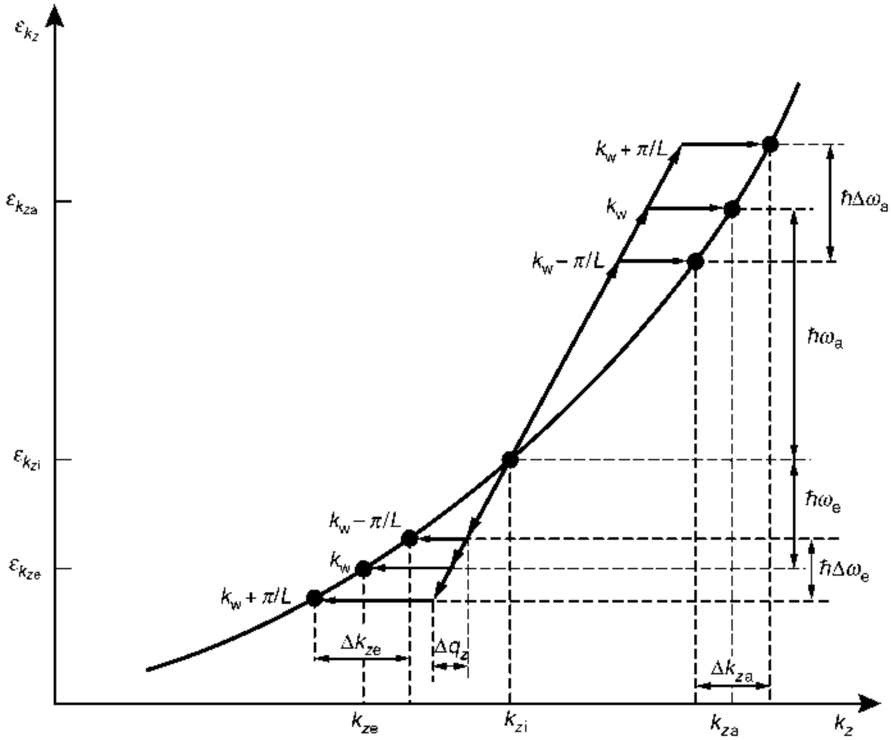

Figure 4 The figure illustrates that the origin of difference between the emission and absorption frequencies is the curvature of the energy dispersion line, and the origin of the homogeneous line broadening is momentum conservation uncertainty $\pm\pi/L$ in a finite interaction length. (Reproduced with permission from Friedman A, Gover A, Ruschin S, Kurizki G and Yariv A (1988) Spontaneous and stimulated emission from quasi-free electrons. *Reviews of Modern Physics* 60: 471–535. Copyright (1988) by the American Physical Society.)

distinct solutions, defining the electron upper and lower quantum levels for radiative emission and absorption respectively. The graphical solutions of these two set of equations are shown in Figure 4, which depicts also the 'homogeneous' frequency-line broadening $\hbar\Delta\omega_e$, $\hbar\Delta\omega_a$ of the emission and absorption lines due to the uncertainty in the momentum conservation $\pm\pi/L$ in a finite interaction length. In the quantum limit of a cold (monoenergetic) e-beam and a long interaction length L, the absorption line center ω_a is larger than the emission line center ω_e, and the linewidths $\Delta\omega_e \cong \Delta\omega_a = \Delta\omega_L$ are narrower than the emission and absorption lines spacing $\omega_a - \omega_e$, as shown in Figure 5a. The FEL then behaves as a 3-level quantum system, with electrons occupying only the central level, and the upper level is spaced apart from it more than the lower level (Figure 4).

In the classical limit $\hbar \to 0$, one can Taylor-expand ε_{kz} around k_{zi}. Using:

$$v_z = \frac{1}{\hbar}\frac{\partial \varepsilon_{kz}}{\partial k_z}, \qquad \frac{1}{\gamma_z^2 \gamma m} = \frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{kz}}{\partial k_z^2}$$

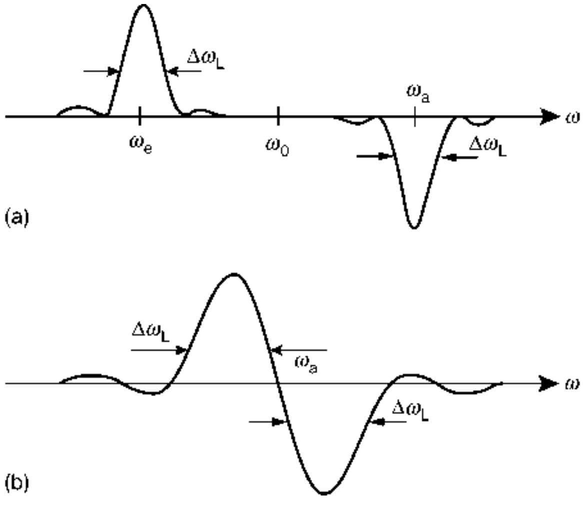

Figure 5 Net gain emission/absorption frequency lines of FEL: (a) in the quantum limit: $\omega_a - \omega_e \gg \Delta\omega_L$, (b) in the classical limit: $\omega_a - \omega_e \ll \Delta\omega_L$.

one obtains:

$$\omega_e \cong \omega_a \cong \omega_0 = v_z(q_{z0} + k_w) \qquad [17]$$

which for $q_{z0} = (\omega/c)\cos\Theta_q$ reproduces the classical synchronism condition in eqn [5]. The homogeneous broadening linewidth is found to be:

$$\frac{\Delta\omega_L}{\omega_0} = \frac{1}{N_\mathrm{w}} \qquad [18]$$

where $N_\mathrm{w} = L/\lambda_\mathrm{w}$ is the number of wiggler periods.

The classical limit condition requires that the difference between the emission and absorption line centers will be smaller than their width. This is expressed in terms of the 'recoil parameter ε' (for $\Theta_q = 0$):

$$\varepsilon = \frac{\omega_\mathrm{a} - \omega_\mathrm{e}}{\Delta\omega_L} = \frac{1+\beta}{\beta_z^2}\frac{\hbar\omega_0}{\gamma mc^2}N_\mathrm{w} \ll 1 \qquad [19]$$

This condition is satisfied in all practical cases of realizable FELs. When this happens, the homogeneous line broadening dominates over the quantum-recoil effect, and the emission and absorption lines are nearly degenerate (Figure 5b). The total quantum-electrodynamic photonic emission rate expression:

$$\frac{\mathrm{d}\nu_q}{\mathrm{d}t} = \Gamma_\mathrm{sp}[(\nu_q + 1)F(\omega - \omega_\mathrm{e}) - \nu_q F(\omega - \omega_\mathrm{a})] \qquad [20]$$

reduces then into:

$$\frac{\mathrm{d}\nu_q}{\mathrm{d}t} = \nu_q\Gamma_\mathrm{sp}\varepsilon\Delta\omega_L\frac{\mathrm{d}}{\mathrm{d}\omega}F(\omega - \omega_\mathrm{e}) + \Gamma_\mathrm{sp}F(\omega - \omega_0) \qquad [21]$$

Here ν_q is the number of photons in radiation mode q, Γ_sp – the spontaneous emission rate, and $F(\omega - \omega_0)$ is the emission (absorption) lineshape function. Figure 5b depicts the transition of the net radiative emission/absorption rate into a gain curve which is proportional to the derivative of the spontaneous emission lineshape function (first term in eqn [21]). Equation [21] presents a fundamental relation between the spontaneous and stimulated emission of FELs, which was observed first by John Madey (Madey's theorem). It can be viewed as an extension of Einstein's relations to a classical radiation source.

The Classical Picture

The spontaneous emission process of FEL (synchrotron undulator radiation) is nothing but dipole radiation of the undulating electrons, which in the laboratory frame of reference is Doppler shifted to high frequency. The understanding of the stimulated emission process requires a different approach.

Consider a single electron, following a sinusoidal trajectory under the effect of a planar undulator magnetic field in eqn [2] (Figure 1):

$$v_x = v_\mathrm{w}\cos k_\mathrm{w}z_\mathrm{e}(t) \qquad [22]$$

$$x = x_\mathrm{w}\sin k_\mathrm{w}z_\mathrm{e}(t) \qquad [23]$$

where $v_\mathrm{w} = ca_\mathrm{w}/\gamma$ and $x_\mathrm{w} = v_\mathrm{w}/(v_z k_\mathrm{w})$. An electromagnetic wave $E_x(z,t) = E_0\cos(\omega t - k_z z)$ propagates collinearly with the electron. Figure 6 displays the electron and wave 'snap-shot' positions as they propagate along one wiggler period λ_w.

If the electron, moving at average axial velocity v_z, enters the interaction region $z = 0$ at $t = 0$, its axial position is $z_\mathrm{e}(t) = v_z t$, and the electric force it experiences is $-eE_x(z_\mathrm{e}(t), t) = -eE_0\cos(\omega - k_z v_z)t$. Clearly, this force is (at least initially at $t = 0$) opposite to the transverse velocity of the electron $v_x = v_\mathrm{w}\cos(k_\mathrm{w}v_z)t$ (imply in deceleration) and the power exchange rate $-ev_\mathrm{e}\cdot E = -ev_x E_x$ corresponds to transfer of energy into the radiation field on account of the electron kinetic energy. Because the phase velocity of the radiation mode is larger than the electron velocity, $v_\mathrm{ph} = \omega/k_z > v_z$, the electron phase $\varphi_\mathrm{e} = (\omega - k_z v_z)t$ grows, and the power exchange rate $-ev_x E_x$ changes. However, if one synchronizes the electron velocity, so that while the electron traverses one wiggler period ($t = \lambda_\mathrm{w}/v_z$), the electron phase advances by 2π: $(\omega - k_z v_z)\cdot\lambda_\mathrm{w}/v_z = 2\pi$, then the power exchange rate from the electron to the wave remains non-negative through the entire interaction length, because then the electron transverse velocity

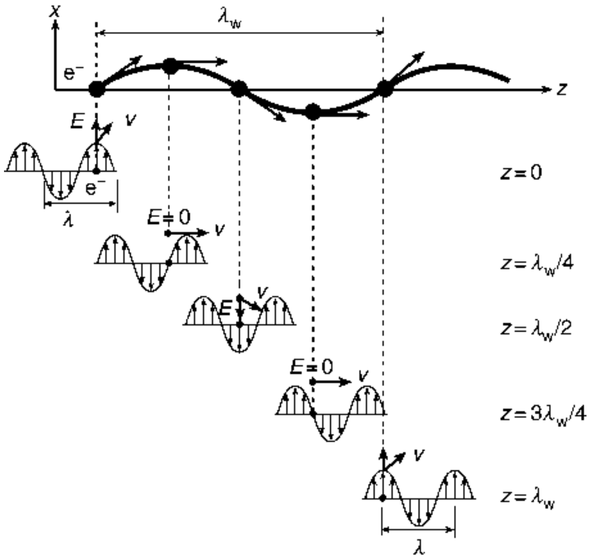

Figure 6 'Snapshots' of an electromagnetic wave period slipping relative to an undulating electron along one wiggler period λ_w. The energy transfer to the wave $-e\mathbf{v}\cdot\mathbf{E}$ remains non-negative all along.

v_x and the wave electric field E_x reverse sign, at each period exactly at the same points $(\lambda_w/4, 3\lambda_w/4)$. This situation is depicted in Figure 6, which shows the slippage of the wave crests relative to the electron at five points along one wiggler period. The figure describes the synchronism condition, in which the radiation wave slips one optical period ahead of the electron, while the electron goes through one wiggle motion. In all positions along this period, $v \cdot E \geq 0$ (in a helical wiggler and a circularly polarized wave this product is constant and positive $v \cdot E > 0$ along the entire period). Substituting $\lambda_w = 2\pi/k_w$, this phase synchronism condition may be written as:

$$\frac{\omega}{v_z} = k_z + k_w \qquad [24]$$

which is the same as eqns [17] and [5].

Figure 6 shows that a single electron (or a bunch of electrons of duration smaller than an optical period) would amplify a co-propagating radiation wave, along the entire wiggler, if it satisfies the synchronism condition in eqn [24] and enters the interaction region $(z = 0)$ at the right (decelerating) phase relative to the radiation field. If the electron enters at the opposite phase it accelerates (on account of the radiation field energy which is then attenuated by 'stimulated absorption'). Thus, when an electron beam is injected into a wiggler at the synchronism condition with electrons entering at random times, no net amplification or absorption of the wave is expected on the averages. Hence, some more elaboration is required, in order to understand how stimulated emission gain is possible then.

Before proceeding, it is useful to define the 'pondermotive force' wave. This force originates from the nonlinearity of the Lorenz force equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma m \mathbf{v}) = -e(\mathbf{E} \times \mathbf{v} \times \mathbf{B}) \qquad [25]$$

At zero order (in terms of the radiation fields), the only field force on the right-hand side of eqn [25] is due to the strong wiggler field (eqns [2] and [3]), which results in the transverse wiggling velocity (eqn [22] for a linear wiggler). When solving next eqn [25] to first order in terms of the radiation fields:

$$\begin{aligned} \mathbf{E}_s(r, \mathbf{t}) &= \mathrm{Re}[\tilde{\mathbf{E}}_s \, \mathrm{e}^{\mathrm{i}(k_z z - \omega t)}] \\ \mathbf{B}_s(r, \mathbf{t}) &= \mathrm{Re}[\tilde{\mathbf{B}}_s \, \mathrm{e}^{\mathrm{i}(k_z z - \omega t)}] \end{aligned} \qquad [26]$$

the cross product $\mathbf{v} \times \mathbf{B}$ between the transverse components of the velocity and the magnetic field generates a longitudinal force component:

$$\mathrm{F}_{\mathrm{pm}}(z, t) = \mathrm{Re}[\hat{\mathrm{e}}_z \tilde{\mathrm{F}}_{\mathrm{pm}} \mathrm{e}^{\mathrm{i}(k_z + k_w)z - \mathrm{i}\omega t}] \qquad [27]$$

that varies with the beat wavenumber $k_s + k_w$ at slow phase velocity $(v_{\mathrm{ph}} = \omega/(k_z + k_w) < c)$. This slow force-wave is called the pomdermotive (PM) wave. Assuming the signal radiation wave in eqn [26] is polarization-matched to the wiggler (linearly polarized or circularly polarized for a linear or helical wiggler respectively), the PM force amplitude is given by:

$$|\tilde{F}_{\mathrm{pm}}| = e|\tilde{E}_{s\perp}|a_w/\gamma\beta_z \qquad [28]$$

With large enough k_w, it is always possible to slow down the phase velocity of the pondermotive wave until it is synchronized with the electron velocity:

$$v_{\mathrm{ph}} = \frac{\omega}{k_z + k_w} = v_z \qquad [29]$$

and can apply along the interaction length a decelerating axial force, that will cause properly phased electrons to transfer energy to the wave on account of their longitudinal kinetic energy.

This observation is of great importance. It reveals that even though the main components of the wiggler and radiation fields are transverse, the interaction is basically longitudinal. This puts the FEL on an equal footing with the slow-wave structure devices as the TWT and the Cerenkov–Smith–Purcell FELs, in which the longitudinal interaction takes place with the longitudinal electric field component of a slow TM radiation mode. The synchronism condition in eqn [29] between the pondermotive wave and the electron, which is identical with the phase-matching condition in eqn [24], is also similar to the synchronism condition between an electron and a slow electromagnetic wave (eqn [14]).

Using the pondermotive wave concept, we can now explain the achievement of gain in the FEL with a random electron beam. Figure 7 illustrates the interaction between the pondermotive wave and electrons, distributed at the entrance $(z = 0)$ randomly within the wave period. Figure 7a shows 'snap-shots' of the electrons in one period of the pondermotive wave $\lambda_{\mathrm{pm}} = 2\pi/(k_z + k_w)$ at different points along the wiggler, when it is assumed that the electron beam is perfectly synchronous with the pondermotive wave $v_{\mathrm{ph}} = v_0$. As explained before, some electrons are slowed down, acquiring negative velocity increment Δv. However, for each such electron, there is another one, entering the wiggler at an accelerating phase of the wave, acquiring the same positive velocity increment Δv. There is then no net change in the energy of the e-beam or the wave, however, there is clearly an effect of 'velocity-bunching' (modulation), which turns along the wiggler into 'density-bunching' at the same

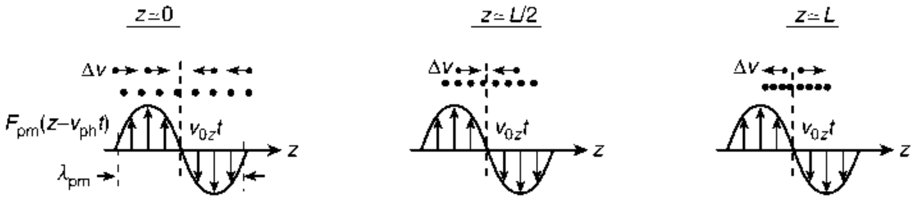

(a) Pure bunching : $v_{ph} = v_{z0}$, $\Delta\varphi(z) = 0$

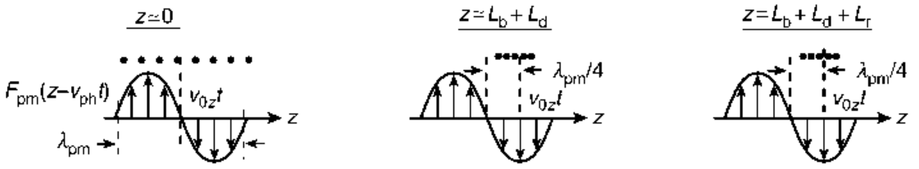

(b) Optical Klystron : $v_{ph} = v_{z0}$, $\Delta\varphi[L_b + L_d] = \frac{\pi}{2}$

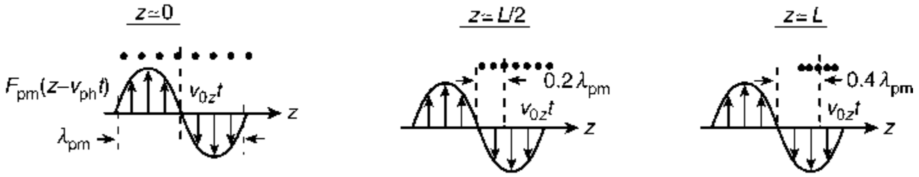

(c) FEL : $v_{z0} \geq v_{ph}$, $\Delta\varphi_c(L) = 2.6$

Figure 7 'Snapshots' of a pondermotive wave period interacting L with an initially uniformly distributed electron beam taking place respectively along the interaction length $0 < z < L_b$. (a) Exact synchronism in a uniform wiggler (bunching). (b) Energy bunching, density bunching and radiation are in the energy buncher, dispersive magnet $L_b < z < L_b + L_d$ and radiating wiggler $L_b + L_d < z < L_b + L_d + L_r$ sections of an Optical–Klystron. (c) Slippage from bunching phase to radiating phase at optimum detuning off synchronism in a uniform wiggler FEL.

frequency ω and wavenumber $k_z + k_w$ as the modulating pondermotive wave. The degree of density-bunching depends on the amplitude of the wave and the interaction length L. In the nonlinear limit the counter propagating (in the beam reference frame) velocity modulated electrons may over-bunch namely cross over, and debunch again.

Bunching is the principle of classical stimulated emission in electron beam radiation devices. If the e-beam had been prebunched in the first place, we would have injected it at a decelerating phase relative to the wave and obtained net radiation gain right away (super radiant emission). This is indeed the principle behind the 'Optical-Klystron' (OK) demonstrated in Figure 7b. The structure of the OK is described ahead in Figure 19. The electron beam is velocity (energy) modulated by an input electromagnetic radiation wave in the first 'bunching-wiggler section' of length L_b. It then passes through a drift-free 'energy-dispersive magnet section' (chicane) of length L_d, in which the velocity modulation turns into density bunching. The bunched electron beam is then injected back into a second 'radiating-wiggler section', where it co-propagates with the same electromagnetic wave but with a phase advance of $\pi/2 - m2\pi$, $m = 1, 2, \ldots$ (spatial lag of $\lambda_{pm}/4 - m\lambda_{pm}$ in real space) which places the entire bunch at a decelerating phase relative to the PM-wave and so amplifies the radiation wave.

The principle of stimulated-emission gain in FEL, illustrated in Figure 7c, is quite similar. Here the wiggler is uniform along the entire length L, and the displacement of the electron bunches into a decelerating phase position relative to the PM-wave is obtained by injecting the electron beam at a velocity v_{z0}, slightly higher than the wave v_{ph} (velocity detuning). The detailed calculation shows that detuning corresponding to a phase shift of $\Delta\Psi(L) = [(\omega/v_{z0}) - (k_z + k_w)]L = -2.6$ (corresponding to spatial bunch advance of $0.4\lambda_{pm}$ along the wiggler length), provides sufficient synchronism with the PM-wave in the first half of the wiggler to obtain bunching, and sufficient deceleration-phasing of the created bunches in the second part of the wiggler to obtain maximum gain.

Principles of FEL Theory

The 3D radiation field in the interaction region can be expanded in general in terms of a complete set of free-space or waveguide modes $\{\varepsilon_q(x,y)\mathcal{H}_q(x,y)\}$:

$$E(r,t) = \mathrm{Re}\Big[\sum_q c_q(z)\varepsilon_q(x,y)\mathrm{e}^{\mathrm{i}(k_{zq}z-\omega t)}\Big] \qquad [30]$$

The mode amplitudes $C_q(z)$ may grow along the wiggler interaction length $0 < z < L$, according to the mode excitation equation:

$$\frac{\mathrm{d}}{\mathrm{d}z}C_q(z) = -\frac{1}{4\mathcal{P}}\mathrm{e}^{-\mathrm{i}k_{zq}}\iint \mathrm{J}(x,y,z)\cdot\varepsilon_q^*(x,y)\mathrm{d}x\,\mathrm{d}y \qquad [31]$$

where $\mathcal{P} = -\frac{1}{2}\mathrm{Re}\iint \varepsilon_q \times \mathcal{H}_q\cdot\hat{e}_z\,\mathrm{d}x\,\mathrm{d}y$ is the mode normalization power, and $\tilde{\mathrm{J}}$ is the bunching current component at frequency ω, that is phase matched to the radiation waves, and needs to be calculated consistently from the electron force equations.

The FEL Small Signal Regime

We first present the basic formulation of FEL gain in the linear (small signal) regime, namely the amplified radiation field is assumed to be proportional to the input signal radiation field, and the beam energy loss is negligible. This is done in the framework of a one-dimensional (single transverse radiation mode) model.

The electron beam charge density, current density, and velocity modulation are solved in the framework of a one-dimensional plasma equations model (kinetic or fluid equations). The longitudinal PM-force in eqn [27] modulates the electron beam velocity via the longitudinal part of the force eqn [25]. This brings about charge modulation $\rho(z,t) = \mathrm{Re}[\tilde{\rho}(k_z + k_\mathrm{w},\omega)\mathrm{e}^{\mathrm{i}(k_z+k_\mathrm{w})z-\mathrm{i}\omega t}]$ and consequently, also longitudinal space-charge field $\tilde{E}_\mathrm{sc}(k_{zq} + k_\mathrm{w},\omega)$ and longitudinal current density modulation $\tilde{J}_z(k_z + k_\mathrm{w},\omega)$, related through the Poison and continuity equations:

$$\mathrm{i}(k_z + k_\mathrm{w})\tilde{E}_\mathrm{sc}(k_z + k_\mathrm{w},\omega) = \tilde{\rho}(k_z + k_\mathrm{w},\omega)/\varepsilon \qquad [32]$$

$$(k_z + k_\mathrm{w})\tilde{J}_z(k_z + k_\mathrm{w},\omega) = \omega\tilde{\rho}(k_z + k_\mathrm{w},\omega) \qquad [33]$$

Solving the force eqn [25] for a general longitudinal force $F_z(z,t) = \mathrm{Re}[\tilde{F}_z(k_z,\omega)\mathrm{e}^{\mathrm{i}(k_z z-\omega t)}]$ results in a linear longitudinal current response relation:

$$\tilde{J}_z(k_z,\omega) = -\mathrm{i}\omega\chi_\mathrm{p}(k_z,\omega)\tilde{F}_z(k_z,\omega)/(-e) \qquad [34]$$

where $\chi_\mathrm{p}(k_z,\omega)$ is the longitudinal susceptibility of the electron beam 'plasma'.The beam charge density in the FEL may be quite high, and consequently the space charge field $\tilde{E}_\mathrm{sc}$, arising from the Poison eqn [32], may not be negligible. One should take into consideration then that the total longitudinal force $\tilde{F}_z$ is composed of both the PM-force of eqn [27] and an arising longitudinal space-charge electric force $-e\tilde{E}_\mathrm{sc}$. Thus, one should substitute in eqn [34]:

$$\tilde{F}_z(k_z + k_\mathrm{w},\omega) = -e[\tilde{E}_\mathrm{pm}(k_z + k_\mathrm{w},\omega) + \tilde{E}_\mathrm{sc}(k_z + k_\mathrm{w},\omega)] \qquad [35]$$

and solve it self-consistently with eqns [32] and [33] to obtain the 'external-force' response relation:

$$\tilde{J}_z(k_z + k_\mathrm{w},\omega) = \frac{-\mathrm{i}\omega\chi_\mathrm{p}(k_z + k_\mathrm{w},\omega)}{1 + \chi_\mathrm{p}(k_z + k_\mathrm{w},\omega)/\varepsilon}\tilde{E}_\mathrm{pm}(k_z + k_\mathrm{w},\omega) \qquad [36]$$

where we defined the PM 'field': $\tilde{E}_\mathrm{pm} = \tilde{F}_\mathrm{pm}/(-e)$.

In the framework of a single-mode interaction model, we keep in the summation of eqn [30] only one mode q (usually the fundamental mode, and in free space – a Gaussian mode). The transverse current density components in eqn [31] $\tilde{J}_\perp = \frac{1}{2}\tilde{\rho}\,\tilde{v}_w$ are found using eqns [22], [33], and [36]. Finally, substituting $C_q(z) = \tilde{C}_q\,\mathrm{e}^{\mathrm{i}\delta kz}$ (where $\delta k \equiv k_z - k_{qz}$ and $k_z \cong k_{qz}$ is the wavenumber of the radiation wave modified by the interaction with the electrons) results in the general FEL dispersion relation:

$$(k_z - k_{zq})[1 + \chi_\mathrm{p}(k_z + k_\mathrm{w},\omega)/\varepsilon] = \kappa\chi_\mathrm{p}(k_z + k_\mathrm{w},\omega)/\varepsilon_0 \qquad [37]$$

Equation [37] is a general expression, valid for a wide variety of FELs, including Cerenkov–Smith–Purcell and TWT. They differ only in the expression for κ. For the conventional (magnetic wiggler) FEL:

$$\kappa = \frac{1}{4}\frac{A_\mathrm{e}}{A_\mathrm{em}}\frac{a_\mathrm{w}^2}{\gamma^2\beta_z^2}\frac{\omega}{c}A_\mathrm{JJ}^2 \qquad [38]$$

where A_e is the cross-section area of the electron beam, and $A_\mathrm{em} \equiv \mathcal{P}_q/[\frac{1}{2}\sqrt{\varepsilon_0/\mu_0}|\varepsilon_{q\perp}(0,0)|^2]$ is the effective area of the interacting radiation-mode q, and it is assumed that the electron beam, passing on axis $(x_\mathrm{e}, y_\mathrm{e}) = (0,0)$, is narrow relative to the transverse mode variation $A_\mathrm{e}/A_\mathrm{em} \ll 1$. The 'Bessel-functions coefficient' A_JJ is defined for a linear wiggler only, and is given by:

$$A_\mathrm{JJ} = J_0\left[\frac{a_\mathrm{w}^2}{2(a_\mathrm{w}^2 + 2)}\right] - J_1\left[\frac{a_\mathrm{w}^2}{2(a_\mathrm{w}^2 + 2)}\right] \qquad [39]$$

In a helical wiggler, $A_\mathrm{JJ} \equiv 1$. Usually $a_\mathrm{w} \ll 1$, and therefore $A_\mathrm{JJ} \cong 1$.

The Pierce Dispersion Equation

The longitudinal plasma response susceptibility function $\chi_\mathrm{p}(k_z,\omega)$ has been calculated, in any plasma formulation, including fluid model, kinetic model, or even quantum-mechanical theory. If the electron beam axial velocity spread is small enough (cold beam), then the fluid plasma equations can be used. The small signal longitudinal force equation derived from eqn [25], together with eqn [33] and the small signal current modulation expression:

$$\tilde{J}_z \cong \rho_0 \tilde{v}_z + v_z \tilde{\rho} \qquad [40]$$

result in:

$$\chi_{\mathrm{p}}(k_z, \omega) = -\frac{\omega_{\mathrm{p}}^{\prime 2}}{(\omega - k_z v_z)^2}\varepsilon \qquad [41]$$

where $\omega'_{\mathrm{p}} = (\mathrm{e}^2 n_0/\gamma\gamma_z^2 \varepsilon m)^{1/2}$ is the longitudinal plasma frequency, n_0 is the beam electrons density ($\rho_0 = -en_0$), v_z is the average axial velocity of the beam.

In this 'cold-beam' limit, the FEL dispersion eqn [37] reduces into the well-known 'cubic dispersion equation' derived first by John Pierce in the late 1940s for the TWT:

$$\delta k(\delta k - \theta - \theta_{\mathrm{pr}})(\delta k - \theta + \theta_{\mathrm{pr}}) = Q \qquad [42]$$

where $\delta k = k_z - k_{zq}$, θ is the detuning parameter (off the synchronism condition of eqn [24]):

$$\theta \equiv \frac{\omega}{v_z} - k_{zq} - k_{\mathrm{w}} \qquad [43]$$

$$\theta_{\mathrm{p}} = \frac{\omega_{\mathrm{p}}}{v_z} \qquad [44]$$

$$Q = \kappa \theta_{\mathrm{p}}^2 \qquad [45]$$

Here $\theta_{\mathrm{pr}} = r_{\mathrm{p}}\theta_{\mathrm{p}}$, where $r_{\mathrm{p}} < 1$ is the plasma reduction factor. It results from the reduction of the longitudinal space-charge field $\tilde{E}_{\mathrm{sp}}$ in a beam of finite radius r_{b} due to the fringe field effect ($r_{\mathrm{p}} \to 1$ when the beam is wide relative to the longitudinal modulation wavelength: $r_{\mathrm{b}} \gg \lambda_{\mathrm{pm}} = 2\pi/(k_{zq} + k_{\mathrm{w}})$).

The cubic equation has, of course, three solutions δk_i ($i = 1, 2, 3$), and the general solution for the radiation field amplitude and power is thus:

$$C_q(z) = \sum_{j=1}^{3} A_j\, \mathrm{e}^{\mathrm{i}\delta k_j z} \qquad [46]$$

$$P(z) = |C_q(z)|^2 \mathcal{P}_q \qquad [47]$$

The coefficients A_i can be determined from three initial conditions of the radiation and e-beam parameters $C_q(0)$, $\tilde{v}(0)$, $\tilde{i}(0)$, and can be given as a linear combination of them (here $\tilde{i} = A_{\mathrm{e}}\tilde{J}_z$ is the longitudinal modulation current):

$$A_j = A_j^{\mathrm{E}}(\omega)C_q^{\omega}(0) + A_j^{v}(\omega)\tilde{v}^{\omega}(0) + A_j^{i}(\omega)\tilde{i}^{\omega}(0) \qquad [48]$$

Alternatively stated, the exit amplitude of the electromagnetic mode can in general be expressed in terms of the initial conditions:

$$C_q(L) = H^{\mathrm{E}}(\omega)C_q(\omega, 0) + H^{\mathrm{v}}(\omega)\tilde{v}_q(\omega, 0) + H^{\mathrm{i}}(\omega)\tilde{i}(\omega, 0) \qquad [49]$$

where

$$H^{(\mathrm{E,v,i})}(\omega) = \sum_{j=1}^{3} A_j^{(\mathrm{E,v,i})}(\omega)\mathrm{e}^{\mathrm{i}\delta k_j L} \qquad [50]$$

In the conventional FEL, electrons are injected in randomly, and there is no velocity prebunching ($\tilde{v}(\omega, 0) = 0$) or current prebunching ($\tilde{i}(\omega, 0) = 0$) (or equivalently $\tilde{n}(\omega, 0) = 0$). Consequently, $C_q(z)$ is proportional to $C_q(0)$ and one can define and calculate the FEL small-signal single-path gain parameter:

$$G(\omega) \equiv \frac{P(L)}{P(0)} = \frac{|C_q(\omega, L)|^2}{|C_q(\omega, 0)|^2} = |H^{\mathrm{E}}(\omega)|^2 \qquad [51]$$

The FEL Gain Regimes

At different physically meaningful operating regimes, some parameters in eqn [42] can be neglected relative to others, and simple analytic expressions can be found for δk_{i}, A_{i}, and consequently $G(\omega)$. It is convenient to normalize the FEL parameters to the wiggler length: $\bar{\theta} = \theta L$, $\bar{\theta}_{\mathrm{pr}} = \theta_{\mathrm{pr}} L$, $\bar{Q} = QL^3$. An additional figure of merit parameter is the 'thermal' spread parameter:

$$\bar{\theta}_{\mathrm{th}} = \frac{v_{z\mathrm{th}}}{v_z}\frac{\omega}{v_z}L \qquad [52]$$

where $v_{z\mathrm{th}}$ is the axial velocity spread of the e-beam (in a Gaussian velocity distribution model: $f(v_z) = \exp[(v_z - v_{z0})/v_{z\mathrm{th}}]/\sqrt{\pi}v_{z\mathrm{th}})$. The axial velocity spread can result out of beam energy spread or angular spread (finite 'emittance'). It should be small enough, so that the general dispersion relation of eqn [37] reduces to eqn [42] (the practical 'cold beam' regime).

Assuming now a conventional FEL ($\tilde{v}_z(0) = 0$, $\tilde{i}(0) = 0$), the single path gain in eqn [51] can be calculated. We present next this gain expression in the different regimes. The maximum values of the gain expression in the different regimes are listed in Table 1.

Low gain

This is the regime where the differential gain in a single path satisfies $G - 1 = [P(L) - P(0)]/P(0) \ll 1$. It is not useful for FEL amplifiers but most FEL oscillators operate in this regime.

Table 1 The gain regimes maximum gain expressions

	Gain regime	*Parameters domain*	*Max. gain expression*
I	Tenuous beam low-gain	$\bar{Q}, \bar{\theta}_{\mathrm{pr}}, \bar{\theta}_{\mathrm{th}} < \pi$	$\frac{P(L)}{P(0)} = 1 + 0.27\bar{Q}$
II	Collective low-gain	$\bar{\theta}_{\mathrm{pr}} > \frac{\bar{Q}}{2}, \bar{\theta}_{\mathrm{th}}, \pi$	$\frac{P(L)}{P(0)} = 1 + \bar{Q}/2\bar{\theta}_{\mathrm{pr}}$
III	Collective high-gain	$\bar{Q}/2 > \bar{\theta}_{\mathrm{pr}} > \bar{Q}^{1/3}, \bar{\theta}_{\mathrm{th}}, \bar{Q} > \pi$	$\frac{P(L)}{P(0)} = \frac{1}{4}\exp(\sqrt{2\bar{Q}/\bar{\theta}_{\mathrm{pr}}})$
IV	Strong coupling high-gain	$\bar{Q}^{1/3} > \bar{\theta}_{\mathrm{pr}}, \bar{\theta}_{\mathrm{th}}, \bar{Q} > \pi$	$\frac{P(L)}{P(0)} = \frac{1}{9}\exp(\sqrt{3}\bar{Q}^{1/3})$
V	Warm beam	$\bar{\theta}_{\mathrm{th}} > \bar{\theta}_{\mathrm{pr}}, \bar{Q}^{1/3}, \pi$	$\frac{P(L)}{P(0)} = \exp(3\bar{Q}/\bar{\theta}_{\mathrm{th}}^2)$

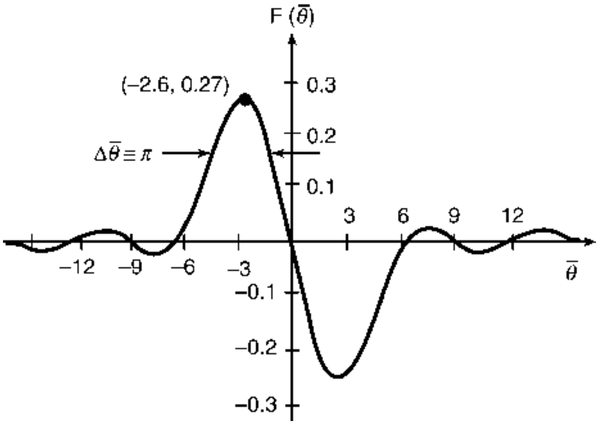

Figure 8 The low-gain cold-beam small-signal gain curve of FEL as a function of the detuning parameter $\bar{\theta}(\omega)$.

The three solutions of eqn [42] – namely the terms of eqn [46] – are reminiscent of the three eigenwaves of the uncoupled system ($\kappa = Q = 0$) : the radiation mode and the two plasma (space-charge) waves of the e-beam (the slow and fast waves, corresponding respectively to the forward and backward propagating plasma-waves in the beam rest reference-frame). In the low-gain regime, all three terms in eqn [46] are significant. Calculating them to first order in κ, results in analytical gain expressions in the collective ($\bar{\theta}_{\mathrm{pr}} \gg \pi$) and tenuous-beam ($\bar{\theta}_{\mathrm{pr}} \ll \pi$) regimes (note that $\bar{\theta}_{\mathrm{pr}}/2\pi = f'_{\mathrm{pr}} L/v_z$ is the number of plasma oscillations within the wiggler transit time L/v_z).

In most practical situations the beam current density is small enough, and its energy high enough, to limit operation to the tenuous-beam regime. The gain curve function is then:

$$G(\omega) - 1 = \bar{Q}F(\bar{\theta}(\omega)) = \bar{Q}\frac{\mathrm{d}}{\mathrm{d}\bar{\theta}}\operatorname{sinc}^2(\bar{\theta}/2) \qquad [53]$$

$$\bar{\theta}(\omega) \equiv \theta(\omega)L = 2\pi\frac{\omega - \omega_0}{\Delta\omega_L} \qquad [54]$$

where $\operatorname{sinc}(u) \equiv (\sin u)/u$, and in free space (no waveguide) propagation ($k_{zq} = \omega/c$), the FWHM frequency bandwidth of the $\operatorname{sinc}^2(\bar{\theta}/2)$ function is:

$$\frac{\Delta\omega_L}{\omega_0} = \frac{1}{N_{\mathrm{w}}} \qquad [55]$$

The small signal gain curve is shown in Figure 8. There is no gain at synchronism – $\omega = \omega_0$. Maximum gain – $G - 1 = 0.27\bar{Q}$, is attained at a frequency slightly smaller than ω_0, corresponding to $\bar{\theta} = -2.6$. The small gain curve bandwidth is $\Delta\omega_{\mathrm{SG}} \cong \Delta\omega_L/2$, namely:

$$\frac{\Delta\omega_{\mathrm{SG}}}{\omega_0} = \frac{1}{2N_{\mathrm{w}}} \qquad [56]$$

High gain

This is the regime where the FEL gain in a single path satisfies $G = P(L)/P(0) \gg 1$. It is useful, of course, when the FEL is used as an amplifier.

Since the coefficients of the cubic eqn [42] are all real, the solutions δk_i ($i = 1, 2, 3$) must be either all real, or composed of one real solution δk_3 and two complex solutions, which are complex conjugate of each other: $\delta k_1 = \delta k_2^*$. In the first case, all terms in eqn [46] are purely oscillatory, there is no exponential growth, and the FEL operates in the low gain regime. In the second case, assuming $\mathrm{Im}(\delta k_1) < 0$, $\mathrm{Im}(\delta k_2) > 0$, the first term grows exponentially, and if L is long enough it will dominate over the other decaying ($j = 2$) and oscillatory ($j = 3$) terms, and result in an exponential gain expression:

$$G(\omega) = \left(\frac{A_1}{A_1 + A_2 + A_3} \right)^2 \mathrm{e}^{2(\mathrm{Im}\ \delta k_1)L} \qquad [57]$$

If we focus on the tenuous-beam strong coupling (high–gain) regime $\theta_{\mathrm{pr}} \ll |\delta k|$, then the cubic eqn [42] gets the simple form:

$$\delta k[(\delta k)^2 - \theta^2] = \Gamma^3 \qquad [58]$$

where

$$\Gamma = Q^{1/3} = \left(\pi \frac{a_{\mathrm{w}}^2}{\gamma^3 \gamma_z^2 \beta_z^5} \frac{\omega/c}{A_{\mathrm{em}}} \frac{I_{\mathrm{b}}}{I_{\mathrm{A}}} A_{\mathrm{JJ}}^2 \right)^{1/3} \qquad [59]$$

and $I_{\mathrm{A}} = 4\pi\varepsilon_0 m_{\mathrm{e}} c^3/e \cong 17$ kA is the Alfven current. The solution of eqn [58] near synchronism ($\theta \cong 0$) is:

$$\delta k_1 = \frac{1 - \sqrt{3}\mathrm{i}}{2}\Gamma, \qquad \delta k_2 = \frac{1 + \sqrt{3}\mathrm{i}}{2}\Gamma, \qquad [60]$$

$$\delta k_3 = -\Gamma$$

resulting in:

$$H^{\mathrm{E}}(\omega) = \frac{C_q(z)}{C_q(0)}$$

$$= \frac{1}{3}\left[\mathrm{e}^{\frac{\sqrt{3}+\mathrm{i}}{2}\Gamma z} + \mathrm{e}^{\frac{-\sqrt{3}+\mathrm{i}}{2}\Gamma z} + \mathrm{e}^{-\mathrm{i}\Gamma z} \right] \qquad [61]$$

and for $\Gamma L \gg 1$:

$$G \cong \frac{1}{9}\mathrm{e}^{\sqrt{3}\Gamma L} \qquad [62]$$

The FEL gain is then exponential and can be very high. The gain exponential coefficient is characterized then by its third-order root scaling with the current, $\alpha I_{\mathrm{b}}^{1/3}$. The high-gain frequency detuning curve (found by solving eqn [58] to second order in θ) is:

$$G \cong \frac{1}{9}\mathrm{e}^{\sqrt{3}\Gamma L} \exp(-\theta^2/\Gamma L 3^{3/2})$$

$$\equiv \frac{1}{9}\mathrm{e}^{\sqrt{3}\Gamma L} \exp\left[-\frac{(\omega - \omega_0)^2}{\Delta\omega_{\mathrm{HG}}^2} \right] \qquad [63]$$

where $\Delta\omega_{\mathrm{HG}}$ is the 1/e half-width of the gain curve:

$$\frac{\Delta\omega_{\mathrm{HG}}}{\omega_0} = \frac{3^{3/4}}{2\pi} \frac{\lambda_{\mathrm{w}}}{(L/\Gamma)^{1/2}} \qquad [64]$$

Super-Radiance, Spontaneous-Emission and Self Amplified Spontaneous Emission (SASE)

Intense coherent radiation power can be generated in a wiggler or any other radiation scheme without any input radiation signal ($C_q(\omega, 0) = 0$) if the electron beam velocity or current (density) are prebunched. Namely, the injected e-beam has a frequency component $\tilde{v}(\omega)$ or $\tilde{i}(\omega)$ in the frequency range where the radiation device emits. In the case of pure density bunching ($\tilde{v}(\omega) = 0$), the coherent power emitted is found from eqns [46, 47, 49]:

$$(P_q)_{\mathrm{SR}} = P_q |H^i(\omega)|^2 |\tilde{i}(\omega, 0)|^2 \qquad [65]$$

A 'prebunched-beam FEL' emits coherent radiation based on the process of Super-radiant Emission (in the sense of Dike). Because all electrons emit in phase radiation wavepackets into the radiation mode, the resultant field amplitude is proportional in this case to the beam current I_{b} and the radiation. By contrast, spontaneous emission from a random electron beam (no bunching) is the result of incoherent superposition of the wavepackets emitted by the electrons and its power is expected to be proportional to the current I_{b}.

When the current to radiation field transfer function $H^i(\omega)$ is known, eqn [65] can be used to calculate the superradiant power, and in the high-gain regime also the amplified-superradiant power. The latter is the amplification of the superradiant radiation in the downstream sections of a long wiggler. Such unsaturated gain is possible only when the beam is partly bunched $\tilde{i}(\omega)(I_{\mathrm{b}})$ (because the FEL gain process requires enhanced bunching).

The expressions for the current to field transfer function, in the superradiant gain and the high-gain amplified superradiance limits respectively, are:

$$|H^i(\omega)| = \frac{(P_{\mathrm{pb}}/P_q)^{1/2}}{I_{\mathrm{b}}} \mathrm{sinc}(\theta L/2) \qquad [66]$$

$$\left|H^i(\omega)\right| = \frac{(P_{\mathrm{pb}}/P_q)^{1/2}}{3\Gamma L I_{\mathrm{b}}} \mathrm{e}^{(\sqrt{3}/2)\Gamma L}\ \mathrm{e}^{-(\omega-\omega_0)^2/2(\Delta\omega)_{\mathrm{HG}}^2} \qquad [67]$$

where

$$P_{\mathrm{pb}} = \frac{I_{\mathrm{b}}^2 Z_q}{32} \left(\frac{a_{\mathrm{w}}}{\gamma\beta_z} \right)^2 \frac{L^2}{A_{\mathrm{em}}} \qquad [68]$$

Z_q is the radiation mode impedance (in free-space $Z_q = \sqrt{\mu_0/\varepsilon_0}$). From these expressions one can calculate the power and spectral power of both coherent (superradiant) and partially coherent (spontaneous emission) radiation of FEL in the negligible gain and high gain regimes. The corresponding super-radiant power is in the negligible superradiance gain limit:

$$P_{\mathrm{SR}} = P_{\mathrm{pb}} \left| \frac{\tilde{i}(\omega)}{I_{\mathrm{b}}} \right|^2 \mathrm{sinc}^2(\theta L/2) \qquad [69]$$

(proportional, as expected, to the modulation current squared) and in the high-gain amplified superradiance limit (assuming initial partial bunching $|i(\omega)/I\mathrm{b}| \ll 1$):

$$P_{\mathrm{SR}} = P_{\mathrm{pb}}\left|\frac{\tilde{i}(\omega)}{I_{\mathrm{b}}}\right|^2 \frac{1}{9(\Gamma L)^2}\,\mathrm{e}^{\sqrt{3}\Gamma L}\,\mathrm{e}^{-(\omega-\omega_0)^2/(\Delta\omega)^2_{\mathrm{HG}}} \qquad [70]$$

The discussion is now extended to incoherent (or partially coherent) spontaneous emission. Due to its particulate nature, every electron beam has random frequency components in the entire spectrum (shot noise). Consequently, incoherent radiation power is always emitted from an electron-beam passing through a wiggler, and its spectral-power can be calculated through the relation:

$$\frac{\mathrm{d}P_q}{\mathrm{d}\omega} = \frac{2}{\pi}\mathcal{P}_q|H^i(\omega)|^2\frac{\langle|\tilde{i}(\omega)|^2\rangle}{T} \qquad [71]$$

Here $\tilde{i}(\omega)$ is the Fourier transform of the current of randomly injected electrons $i(t) = -e\sum_{j=1}^{N_T}\delta(t - t_{oj})$, where N_T is the average number of electrons in a time period T, namely, the average (DC) current is $I_{\mathrm{b}} = -eN_T/T$. For a randomly distributed beam, the shot noise current is simply $\langle|i(\omega)|^2\rangle/T = eI_{\mathrm{b}}$, and therefore the spontaneous emission power of the FEL, which is nothing but the 'synchrotron-undulator radiation', is given by (see eqn [66]):

$$\frac{\mathrm{d}P_q}{\mathrm{d}\omega} = \frac{1}{16\pi}eI_{\mathrm{b}}Z_q\frac{L^2}{A_{\mathrm{em}}}\left(\frac{a_{\mathrm{w}}}{\gamma\beta_z}\right)^2\mathrm{sinc}^2(\theta L/2) \qquad [72]$$

If the wiggler is long enough, the spontaneous emission emitted in the first part of the wiggler can be amplified by the rest of it (SASE). In the high-gain limit (see eqn [67]), the amplified spontaneous emission power within the gain bandwidth of eqn [64] is given by:

$$P_q = \frac{2}{\pi}\mathcal{P}_q eI_{\mathrm{b}}\int_0^\infty |H^i(\omega)|^2\mathrm{d}\omega = \frac{1}{9}P_{\mathrm{sh}}\,\mathrm{e}^{\sqrt{3}\Gamma L} \qquad [73]$$

where P_{sh} is an 'effective shot-noise input power':

$$P_{\mathrm{sh}} = \frac{2}{\sqrt{\pi}}\frac{eP_{\mathrm{pb}}}{I_0(\Gamma L)^2}(\Delta\omega)_{\mathrm{HG}} \qquad [74]$$

Saturation Regime

The FEL interaction of an electron with an harmonic electromagnetic (EM) wave is essentially described by the longitudinal component of the force in eqn [25], driven by the pondermotive force of eqns [27] and [28]:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma_i m v_{zi}) = |\tilde{F}_{\mathrm{pm}}|\cos[\omega t - (k_z + k_{\mathrm{w}})z_i] \qquad [75]$$

$$\mathrm{d}z_i/\mathrm{d}t = v_{zi} \qquad [76]$$

As long as the interaction is weak enough (small signal regime), the change in the electron velocity is negligible – $v_{zi} \cong v_{z0}$, and the phase of the force-wave, experienced by the electron, is linear in time $\Psi_i(t) = [\omega - (k_z + k_{\mathrm{w}})v_{z0}](t - t_{0i}) + \omega t_{0i}$. Near synchronism condition $\theta \cong 0$ (eqn [24]), eqn [75] results in bunching of the beam, because different acceleration/deceleration forces are applied on each electron, depending on their initial phase $\Psi_i(0) = \omega t_{0i}(-\pi < \Psi_i(0) < \pi)$ within each optical period $2\pi/\omega$ (see Figure 7). Taylor expansion of v_{zi} around v_{z0} in eqns [75] and [76], and use of conservation of energy between the e-beam and the radiation field, would lead again to the small signal gain expression eqn [53] in the low gain regime.

When the interaction is strong enough (the nonlinear or saturation regime), the electron velocities change enough to invalidate the assumption of linear time dependence of Ψ_i and the nonlinear set of eqns [75] and [76] needs to be solved exactly.

It is convenient to invert the dependence on time $z_i(t) = \int_{t_{0i}}^t v_{zi}(t')\mathrm{d}t'$, and turn the coordinate z to the independent variable $t_i(z) = \int_0^z \mathrm{d}z'/v_{zi}(z') + t_{0i}$. This, and direct differentiation of $\gamma_i(v_{zi})$, reduces eqns [75] and [76] into the well-known pendulum equation:

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}z} = K_{\mathrm{s}}^2\sin\Psi_i \qquad [77]$$

$$\frac{\mathrm{d}\Psi_i}{\mathrm{d}z} = \theta_i \qquad [78]$$

where

$$\Psi_i = \int_0^z (\omega/v_{zi}(z') - k_z - k_{\mathrm{w}})\,\mathrm{d}z' \qquad [79]$$

$$\theta_i = \frac{\omega}{v_{zi}} - k_z - k_{\mathrm{w}} \qquad [80]$$

are respectively the pondermotive potential phase and the detuning value of electron i at position z.

$$K_{\mathrm{s}} = \frac{k\sqrt{a_{\mathrm{w}}a_{\mathrm{s}}A_{\mathrm{JJ}}}}{\gamma_0\gamma_{z0}\beta_{z0}^2} \qquad [81]$$

is the synchrotron oscillation wavenumber, where a_{w} is given in eqn [9], $a_{\mathrm{s}} = e|\tilde{E}|\omega/mc$, and $\gamma_0 = \gamma(0)$, $\gamma_{z0} = \gamma_z(0)$, and $\beta_{z0} = \beta_z(0)$ are the initial parameters of the assumed cold beam.

The pendulum eqns [77] and [78] can be integrated once, resulting in:

$$\tfrac{1}{2}\theta_i^2(z) - K_{\mathrm{s}}^2\cos\Psi_i(z) = C_i \qquad [82]$$

and the integration constant is determined for each electron by its detuning and phase relative to the pondermotive wave at the entrance point ($z = 0$): $C_i = \frac{1}{2}\theta_i^2(0) - K_s^2 \cos \Psi_i(0)$.

The $\theta(z)$, $\Psi(z)$ phase-space trajectories of eqn [82] are shown in Figure 9 for various values of C_i (corresponding to the initial conditions $\theta_i(0)$, $\Psi_i(0)$). The trajectories corresponding to $|C_i| > K_s^2$ are open; namely electrons on these trajectories, while oscillating, can slip-off out of the pondermotive–potential wave period to adjacent periods, ahead or backward, depending on the value of their detuning parameter $\bar{\theta}$. The trajectories corresponding to $|C_i| < K_s^2$ are closed, namely the electrons occupying these trajectories are 'trapped', and their phase displacement is bound to a range $|\Psi_i(z) - n\pi| < \Psi_{\mathrm{im}} \equiv \arccos(|C_i|/K_s^2) < \pi$ within one pondermotive-wave period. The trajectory $C_i = K_s^2$ defines the 'separatrix':

$$\theta_i(z) = \pm 2K_s \cos(\Psi_i/2) \qquad [83]$$

which is sometimes referred to as the 'trap' or 'bucket'. Every electron within the separatrix stays trapped, and the ones out of it are free (untrapped). The height of the separatrix (maximum detuning swing) is $\Delta\theta_{\mathrm{trap}} = 4K_s$. The oscillation frequency of the trapped electrons can be estimated for deeply trapped electrons ($\Psi_m \ll 2\pi$). In this case the physical pendulum eqns [77] and [78] reduce to the mathematical pendulum equation with an oscillation frequency K_s, in the z coordinate. This longitudinal oscillation, called 'synchrotron oscillation', takes place as a function of time at the 'synchrotron frequency' $\Omega_s = K_s v_z$.

Differentiation of $\theta_i(v_{zi})$ and $v_{zi}(\gamma_i)$ permits to describe the phase-space dynamics in terms of the more physical parameters $\delta v_{zi} = v_{zi} - v_{\mathrm{ph}}$ and $\delta\gamma_i = \gamma_i - \gamma_{\mathrm{ph}}$, where:

$$v_{\mathrm{ph}} = \frac{\omega}{k_z + k_w} \qquad [84]$$

is the phase velocity of the pondermotive wave and $\gamma_{\mathrm{ph}} \equiv (1 - \beta_{\mathrm{ph}}^2)^{-1/2}$:

$$-\theta_i = \frac{\omega}{c^2 \beta_{z0}^2} \delta v_{zi} = \frac{k}{\beta_{z0}^3 \gamma_{z0}^2 \gamma_0} \delta\gamma_i \qquad [85]$$

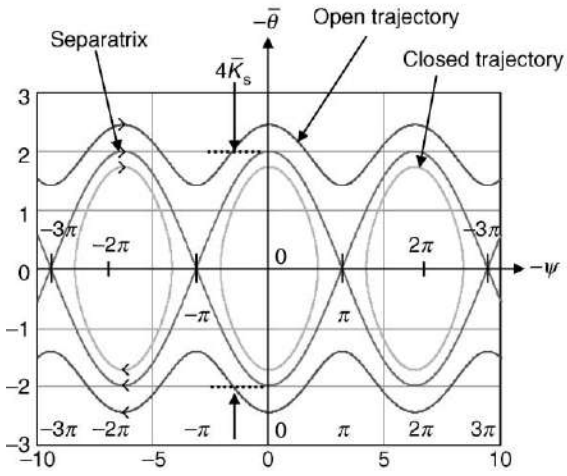

Figure 9 The (θ–Ψ) phase-space trajectories of the pendulum equation.

Figure 10 displays a typical dynamics of electron beam phase-space (γ, Ψ) evolution for the case of a cold beam of energy $\gamma(0)$ entering the interaction region at $z = 0$ with uniform phase distribution (random arrival times t_{0i}). The FEL is assumed to operate in the low-gain regime (typical situation in an FEL oscillator), and, therefore, the trap height (corresponding to $\Delta\theta_{\mathrm{trap}} = 4k_s$):

$$\Delta\gamma_{\mathrm{trap}} = 8\beta_z^2 \gamma_z^2 \gamma K_s / k \qquad [86]$$

remains constant along the interaction length. Figure 10a displays the e-beam phase-space evolution in the small signal regime. The uniform phase distribution evolves along the wiggler into a bunched distribution (compare to Figure 7c), and its average kinetic energy goes down $(\Delta E_k) = [\langle\gamma_i(L)\rangle - \gamma(0)]mc^2 < 0$, contributing this energy to the field of the interacting radiation mode, $\Delta P_q = (\Delta E_k) I_0/e$. In this case (corresponding in an FEL oscillator to the early stages of oscillation build-up), the electrons remain free (untrapped) along the entire length L.

Figure 10b displays the e-beam phase-space evolution in the large signal (saturation) regime (in the case of an oscillator – at the steady-state saturation stage). Part of the electrons are found inside the trap, immediately upon entering the interaction region ($z = 0$), and they lose energy of less than (but near) $mc^2 \Delta\gamma_{\mathrm{trap}}$ as they pass through the interaction region ($z = L$). A portion of the electrons remain outside the traps upon entrance. They follow open trajectories and lose less energy or may even become accelerated due to their interaction with the wave.

It can be appreciated from this discussion that a good design strategy in attempting to extract maximum power from the electron beam in the FEL interaction, is to set the parameters determining the synchrotron oscillation frequency K_s in eqn [81] so that only half a synchrotron oscillation period will be performed along the interaction length:

$$K_s L = \pi \qquad [87]$$

This is controlled in an amplifier by keeping the input radiation power $P_q(0)$ (and consequently a_s) small enough, so that K_s will not exceed the value set

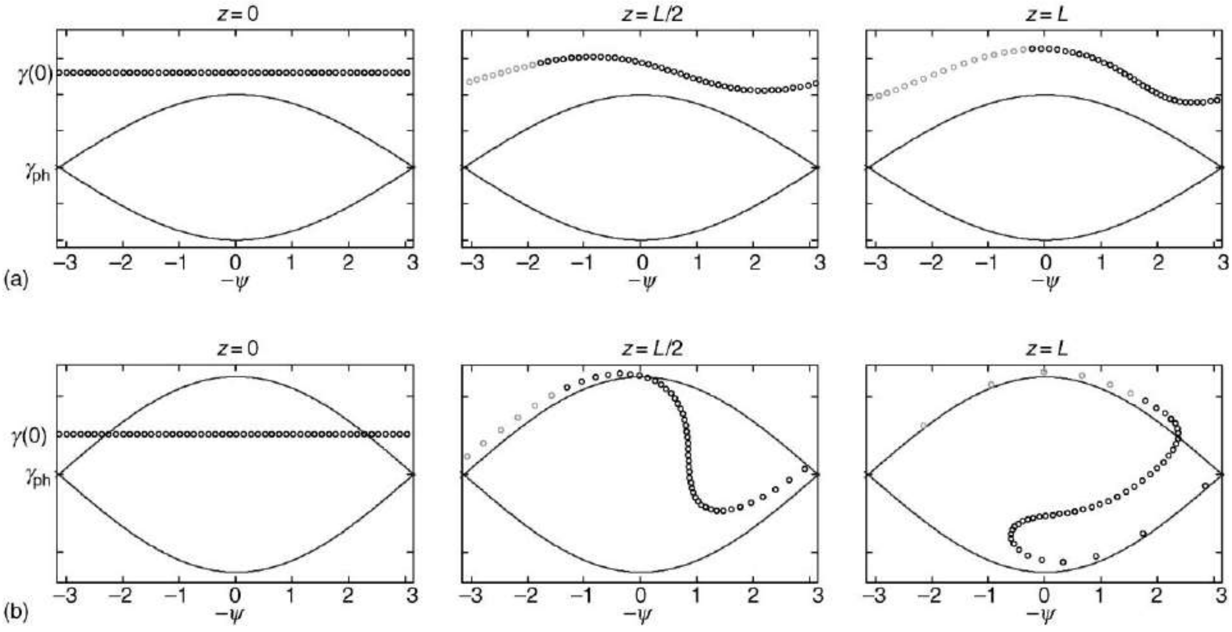

Figure 10 'Snapshots' of the ($\gamma-\Psi$) phase-space distribution of an initially uniformly distributed cold beam relative to the PM-wave trap at three points along the wiggler (a) Moderate bunching in the small-signal low gain regime. (b) Dynamics of electron beam trapping and synchrotron oscillation at steady state saturation stage of a FEL oscillator ($K_sL = \pi$).

by eqn [87]. In an oscillator, this is controlled by increasing the output mirror transmission sufficiently, so that the single path incremental small signal gain G-1 will not be much larger than the round trip loss, and the FEL will not get into deep saturation. When the FEL is over-saturated ($K_sL > \pi$), the trapped electrons begin to gain energy as they continue to rotate in their phase-space trajectories beyond the lowest energy point of the trap, and the radiative energy extraction efficiency drops down.

A practical estimate for the FEL saturation power emission and radiation extraction efficiency can be derived from the following consideration: the electron beam departs from most of its energy during the interaction with the wave, if a significant fraction of the electrons are within the trap and have positive velocity δv_{zi} relative to the wave velocity v_{ph} at $z = 0$, and if at the end of the interaction length ($z = L$), they complete half a pendulum swing and reverse their velocity relative to the wave $\delta v_{zi}(L) \cong -\delta v_{zi}(0)$. Correspondingly, in the energy phase-space diagram (Figure 10b) the electrons perform half a synchrotron oscillation swing and $\delta\gamma_i(L) = \gamma_i(L) - \gamma_{ph} = -\delta\gamma_i(0)$. In order to include in this discussion also the FEL amplifier (in the high gain regime), we note that in this case the phase velocity of the wave v_{ph} in eqn [84], and correspondingly γ_{ph}, are modified by the interaction contribution to the radiation wavenumber – $k_z = k_{z0} + \mathrm{Re}(\delta k)$, and also the electron detuning parameter (relative to the pondermotive wave) θ_i in eqn [80] differs from the beam detuning parameter θ in eqn [43]: $\theta_i = \theta - \mathrm{Re}(\delta k)$. Based on these considerations and eqn [85], the maximum energy extraction from the beam in the saturation process is:

$$\Delta\gamma = 2\delta\gamma_i(0) = 2\beta_{z0}^3\gamma_{z0}^2\gamma_0\frac{\mathrm{Re}\,\delta k - \theta}{k} \qquad [88]$$

where θ is the initial detuning parameter in eqn [43].

In an FEL oscillator, operating in general in the low-gain regime, $|\mathrm{Re}\,\delta k| \ll |\theta|$, oscillation will start usually at the resonator mode frequency, corresponding to the detuning parameter $\theta(\omega) = -2.6/L$, for which the small signal gain is maximal (see Figure 8). Then the maximum radiation extraction efficiency can be estimated directly from eqn [88]. It is, in the highly relativistic limit ($\beta_{z0} \cong 1$):

$$\eta_{ext} = \frac{\Delta\gamma}{\gamma_0} \cong \frac{1}{2N_w} \qquad [89]$$

In an FEL amplifier, in the high-gain regime $\mathrm{Re}\,\delta k = \Gamma/2 \gg |\theta|$, and consequently in the same limit:

$$\eta_{ext} \cong \frac{\Gamma\lambda_w}{4\pi} \qquad [90]$$

It may be interpreted that the effective wiggler length for saturation is $L_{eff} = 2\pi/\Gamma$.

Equation [90], derived here for a coherent wave, is considered valid also for estimating the saturation efficiency also in SASE-FEL. In this context, it is also called 'the efficiency parameter' $-\rho$.

FEL Radiation Schemes and Technologies

Contrary to conventional atomic and molecular lasers, the FEL operating frequency is not determined by natural discrete quantum energy levels of the lasing matter, but by the synchronism condition of eqn [24] that can be predetermined by the choice of wiggler period, $\lambda_w = 2\pi/k_w$, the resonator dispersion characteristics $k_{zq}(\omega)$, and the beam axial velocity v_z.

Because the FEL design parameters can be chosen at will, its operating frequency can fit any requirement, and furthermore, it can be tuned over a wide range (primarily by varying v_z). This feature of FEL led to FEL development efforts in regimes where it is hard to attain high-power tunable conventional lasers or vacuum-tube radiation sources – namely in the sub-mm (far infrared or THz) regimes, and in the VUV down to soft X-ray wavelengths.

In practice, in an attempt to develop short wavelength FELs, the choice of wiggler period λ_w is limited by an inevitable transverse decay of the magnetic field away from the wiggler magnets surface (a decay range of $\approx k_w^{-1}$) dictated by the Maxwell equations. To avoid interception of electron beam current on the walls or on the wiggler surfaces, typical wiggler periods are made longer than $\lambda_w > 1$ cm. FELs (or FEMs – free electron masers) operating in the long wavelengths regime (mm and sub-mm wavelengths) must be based on waveguide resonators to avoid excessive diffraction of the radiation beam along the interaction length (the wiggler). This determines the dispersion relation $k_{zq}(\omega) = (\omega^2 - \omega_{coq}^2)^{1/2}/c$ where ω_{coq} is the waveguide cutoff frequency of the radiation mode q. The use of this dispersion relation in eqn [24] results in an equation for the FEL synchronism frequency ω_0. Usually the fundamental mode in an overmoded waveguide is used (the waveguide is overmoded because it has to be wide enough to avoid interception of electron beam current). In this case ($\omega_0 \gg \omega_{co}$) and certainly in the case of an open resonator (common in FELs operating in the optical regime) $k_{zq} = \omega/c$, and the synchronism condition in eqn [24] simplified to the well-known FEL radiation wavelength expression in eqn [6]:

$$\lambda = (1 + \beta_z)\beta_z \gamma_z^2 \lambda_w \cong 2\gamma_z^2 \lambda_w \qquad [91]$$

where γ_z, a_w are defined in eqns [7]–[9].

To attain strong interaction, it is desirable to keep the wiggler parameter a_w large (eqn [38]), however, if $a_w > 1$, this will cause reduction in the operating wavelength according to eqns [7] and [91]. For this reason, and also in order to avoid harmonic frequencies emission (in case of a linear wiggler), $a_w < 1$ in common FEL design. Consequently, considering the practical limitations on λ_w, the operating wavelength eqn [91] is determined primarily by the beam relativistic Lorentz factor γ (eqn [8]).

The conclusion is that for a short wavelength FEL, one should use an electron beam accelerated to high kinetic energy E_k. Also, tuning of the FEL operating-wavelength can be done by changing the beam energy. Small-range frequency tuning can be done also by changing the spacing between the magnet poles of a linear wiggler. This varies the magnetic field experienced by the e-beam, and effects the radiation wavelength through change of a_w (see eqns [7] and [91]).

Figure 11 displays the operating wavelengths of FEL projects all over the world versus their e-beam energy. FELs were operated or planned to operate over a wide range of frequencies, from the microwave to X-ray – eight orders of magnitude. The data points fall on the theoretical FEL radiation curve eqns [7], [8], and [91].

FEL Accelerator Technologies

The kind of accelerator used is the most important factor in determining the FEL characteristics. Evidently, the higher the acceleration energy, the shorter is the FEL radiation wavelength. However, not only the acceleration beam energy determines the shortest operating wavelength of the FEL, but also the e-beam quality. If the accelerated beam has large energy spread, energy instability, or large emittance (the product of the beamwidth with its angular spread), then it may have large axial velocity spread v_{zth}. At high frequencies, this may push the detuning spread parameter $\bar{\theta}_{th}$ (eqn [52]) to the warm beam regime (see Table 1), in which the FEL gain is diminished, and FELs are usually not operated.

Other parameters of the accelerator determine different characteristics of the FEL. High current in the electron beam enables higher gain and higher power operation. The e-beam pulse shape (or CW) characteristics, affect, of course, the emitted radiation waveform, and may also affect the FEL gain and saturation characteristics. The following are the main accelerator technologies used for FEL construction. Their wavelength operating-regimes (eqn [91]) (determined primarily by their beam acceleration energies), are displayed in Figure 12.

Modulators and pulse-line accelerators

These are usually single pulse accelerators, based on high voltage power supplies and fast discharge stored

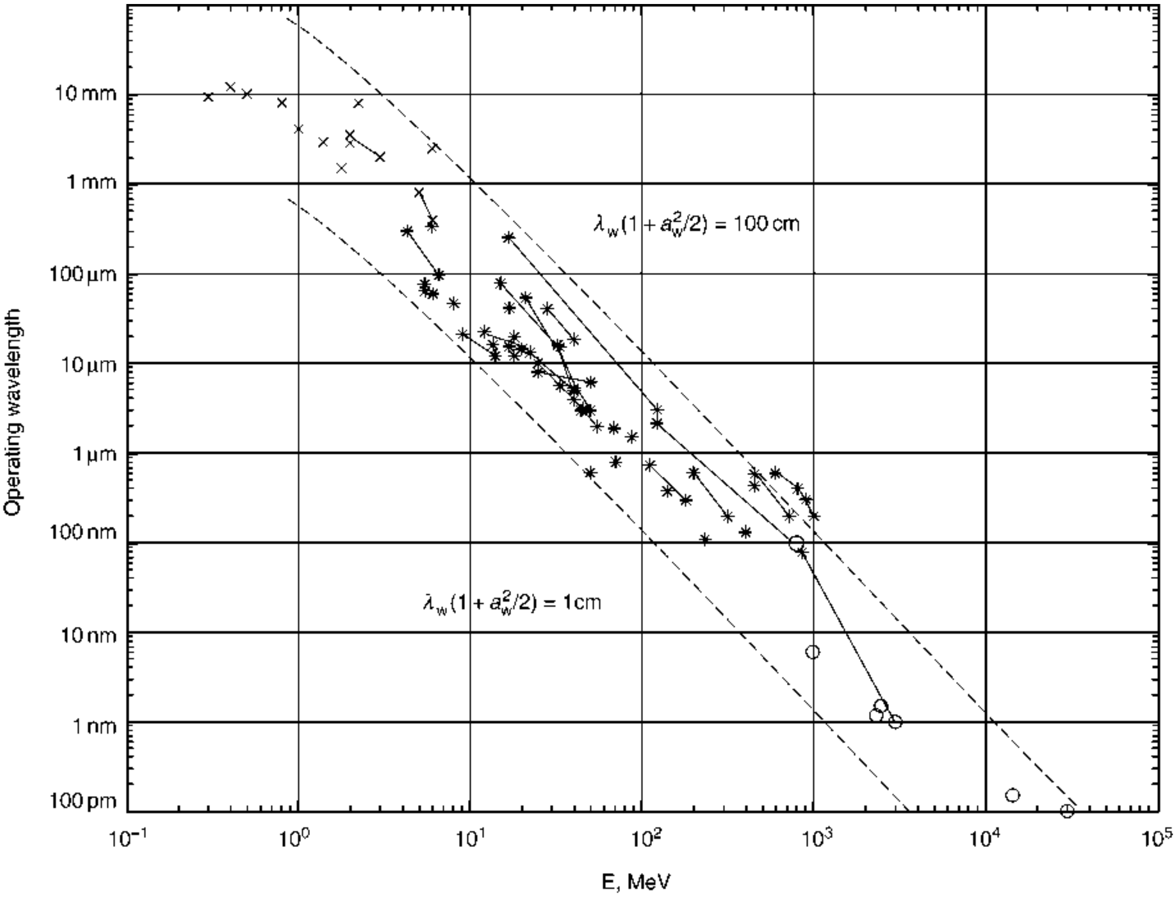

Figure 11 Operating wavelengths of FELs around the world vs. their accelerator beam energy. The data points correspond in ascending order of accelerator energy to the following experimental facilities: NRL (USA), IAP (Russia), KAERI (Korea), IAP (Russia), JINR/IAP (Russia), INP/IAP (Russia), TAU (Israel), FOM (Netherlands), KEK/JAERI (Japan/Korea), CESTA (France), ENEA (Italy), KAERI-FEL (Korea), LEENA (Japan), ENEA (Italy), FIR FEL (USA), mm Fel (USA), UCSB (USA), ILE/ILT (Japan), MIRFEL (USA), UCLA-Kurchatov (USA/Russia), FIREFLY (GB), JAERI-FEL (Japan), FELIX (Netherlands), RAFEL (USA), ISIR (Japan), UCLA-Kurchatov-LANL (USA/RU), ELSA (France), CLIO (France), SCAFEL (GB), FEL (Germany), BFEL (China), KHI-FEL (Japan), FELI4 (Japan), iFEL1 (Japan), HGHG (USA), FELI (USA), MARKIII (USA), ATF (USA), iFEL2 (Japan), VISA (USA), LEBRA (Japan), OK-4 (USA), UVFEL (USA), iFEL3 (Japan), TTF1 (Germany), NIJI-IV (Japan), APSFEL (USA), FELICITAI (Germany), FERMI (Italy), UVSOR (Japan), Super-ACO (France), TTF2 (Germany), ELETTRA (Italy), Soft X-ray (Germany), SPARX (Italy), LCLS (USA), TESLA (Germany). X, long wavelengths; $*$, short wavelengths; ○, planned short wavelengths SASE-FELs. Data based in part on H. P. Freund, V. L. Granatstein, Nucl. Inst. and Methods In Phys. Res. A249, 33 (1999), W. Colson, Proc. of the 24th Int. FEL conference, Argone, Ill. (ed. K. J. Kim, S. V. Milton, E. Gluskin). The data points fall close to the theoretical FEL radiation condition expression (91) drawn for two practical limits of wiggler parameters.

electric energy systems (e.g., Marx Generator), which produce short pulse (tens of nSec) Intense Relativistic Beam (IRB) of energy in the range of hundreds of keV to few MeV and high instantaneous current (order of kAmp), using explosive cathode (plasma field emission) electron guns. FELs (FEMs), based on such accelerators, operated mostly in the microwave and mm-wave regimes. Because of their poor beam quality and single pulse characteristic, these FELs were, in most cases, operated only as Self Amplified Spontaneous Emission (SASE) sources, producing intense radiation beams of low coherence at instantaneous power levels in the range of 1–100 MW. Because of the high e-beam current and low energy, these FEMs operated mostly in the collective high-gain regime (see Table 1).

Some of the early pioneering work on FEMs was done in the 1970s and 1980s in the US (NRL, Columbia Univ., MIT), Russia (IAP), and France (Echole Politechnique), based on this kind of accelerators.

Induction linacs

These too are single pulse (or low repetition rate) accelerators, based on induction of electromotive potential over an acceleration gap by means of an electric-transformer circuit. They can be cascaded to high energy, and produce short pulse (tens to hundreds

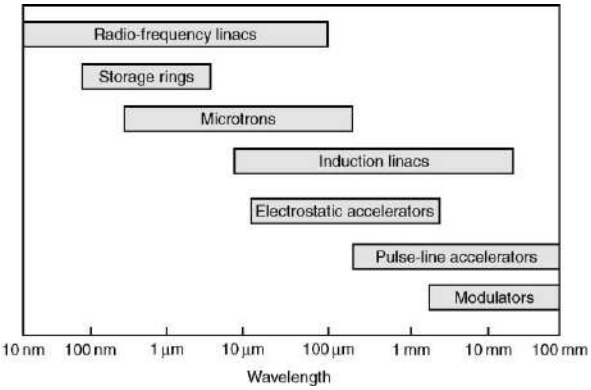

Figure 12 Approximate wavelength ranges accessible with FELs based on current accelerator and wiggler technologies.

of nSec), high current (up to 10 kA) electron beams, with relatively high energy (MeV to tens of MeV). The interest in FELs, based on this kind of accelerator technology, stemmed in the 1980s primarily from the SDI program, for the propose of development of a DEW FEL. The main development of this technology took place on a 50 MeV accelerator – ATA (for operating at 10 μm wavelength) and a 3.5 MeV accelerator – ETA (for operating at 8 mm wavelength). The latter experiment, operating in the high-gain regime, demonstrated record high power (1 GW) and energy extraction efficiency (35%).

Electrostatic accelerators

These accelerators are DC machines, in which an electron beam, generated by a thermionic electron-gun (typically 1–10 Amp) is accelerated electrostatically. The charging of the high voltage terminal can be done by mechanical charge transport (Van de Graaff) or electrodynamically (Crockford–Walton accelerator, Dynamitron). The first kind can be built at energies up to 25 MeV, and the charging current is less than mAmp. The second kind have terminal voltage less than 5 MeV, and the charging current can be hundreds of mAmps.

Because of their DC characteristics, FELs based on these kinds of accelerators can operate at arbitrary pulse shape structure and in principle – continuously (CW). However, because of the low charging current, the high electron beam current (1–10 Amp), required for FEL lasing must be transported without any interception along the entire way from the electron gun, through the acceleration tubes and the FEL wiggler, and then decelerated down to the voltage depressed beam-collector (multistage collector), closing the electric circuit back to the e-gun (current recirculation). The collector is situated at the e-gun potential, biased by moderate voltage high current power supplies, which deliver the current and power needed for circulating the e-beam and compensates for its kinetic energy loss in favor of the radiation field in the FEL cavity. This beam current recirculation is, therefore, also an ‘Energy retrieval’ scheme, and can make the overall energy transfer efficiency of the electrostatic-accelerator FEL very high.

In practice, high-beam transport efficiency in excess of 99.9% is needed for CW lasing, and has not been demonstrated yet. To avoid HV-terminal voltage drop during lasing, electrostatic-accelerator FELs are usually operated in a single pulse mode. Few FELs of this kind have been constructed. The first and main facility is the UCSB FEL shown in Figure 13. It operates in the wavelength range of 30 μm to 2.5 mm (with three switchable wigglers) in the framework of a dedicated radiation user facility. This FEL operates in the negatively charged terminal mode, in which the e-gun and collector are placed in the negatively charged HV-terminal inside the pressurized insulating gas tank, and the wigglers are situated externally at ground potential. An alternative operating mode of positively charged terminal internal cavity electrostatic-accelerator FEM was demonstrated in the Israeli Tandem–Accelerator FEM and the Dutch F.O.M. Fusion-FEM projects. This configuration enables operating with long pulse, high coherence, and very high average power. Linewidth of $\Delta\omega/\omega \cong 10^{-5}$ was demonstrated in the Israeli FEM and high power (730 kW over few microseconds) was demonstrated in the Dutch FEM, both at mm-wavelengths. The goal of the latter development project (which was not completed) was quasicontinuous operation at 1 MW average power for application in fusion plasma heating.

Radio-frequency (RF) accelerators

RF-accelerators are by far the most popular electron-beam sources for FELs. In RF accelerators, short electron beam bunches (bunch duration 1–10 pSec) are accelerated by the axial field of intense RF radiation (frequency about 1 GHz), which is applied in the acceleration cavities on the injected short e-beam bunches, entering with the accelerating-phase of the RF field. In microtrons, the electron bunches perform circular motion, and get incremental acceleration energy every time they re-enter the acceleration cavity. In RF-LINACs (linear accelerator), the electron bunches are accelerated in a sequence of RF cavities or a slow-wave structure, which keep an accelerating-phase synchronization of the traversing electron bunches with the RF field along a long linear acceleration length. The bunching of the electrons, prior to the acceleration step, is traditionally performed by bunching RF-cavities and a dispersive magnet (chicane) pulse compression system.

The FEL small signal gain, must be large enough to build-up the radiation field in the resonator from noise to saturation well within the macropulse duration.

RF-Linacs are essential facilities in synchrotron radiation centers, used to inject electron beam current into the synchrotron storage ring accelerator from time to time. Because of this reason, many FELs based on RF-LINACs were developed in synchrotron centers, and provide additional coherent radiation sources to the synchrotron radiation center users.

Figure 15 displays FELIX – a RF-LINAC FEL which is located in one of the most active FEL radiation user-centers in FOM – Holland.

Storage rings

Storage rings are circular accelerators in which a number of electron (or positron) beam bunches (typically of 50–500 pS pulse duration and hundreds of ampere peak current) are circulated continuously by means of a lattice of bending magnets and quadrupole lenses. Typical energies of storage ring accelerators are in the hundreds of MeV to GeVs range. As the electrons pass through the bending magnets, they lose a small amount of their energy due to emission of synchrotron radiation. This energy is replenished by a small RF acceleration cavity placed in one section of the ring. The electron beam bunch dimensions, energy spread, and emittance parameters are set in steady state by a balance between the electrons oscillations within the ring lattice and radiation damping due to the random synchrotron emission process. This produces high-quality (small emittance and energy spread) continuous train of electron beam bunches, that can be used to drive a FEL oscillator placed as an insertion device in one of the straight sections of the ring between two bending magnets.

Demonstrations of FEL oscillators, operating in a storage ring, were first reported by the French (LURE-Orsay) in 1987 (at visible wavelengths) and the Russians (VEPP-Novosibirsk) in 1988 (in the ultraviolet). The short wavelength operation of storage-ring FELs is facilitated by the high energy, low emittance and low energy spread parameters of the beam.

Since storage ring accelerators are at the heart of all synchrotron radiation centers, one could expect that FEL would be abundant in such facilities as insertion devices. There is, however, a problem of interference of the FEL operating as an insertion device in the normal operation of the ring itself. The energy spread increase, induced in the electron beam during the interaction in a saturated FEL oscillator, cannot be controlled by the synchrotron radiation damping process, if the FEL operating power is too high.

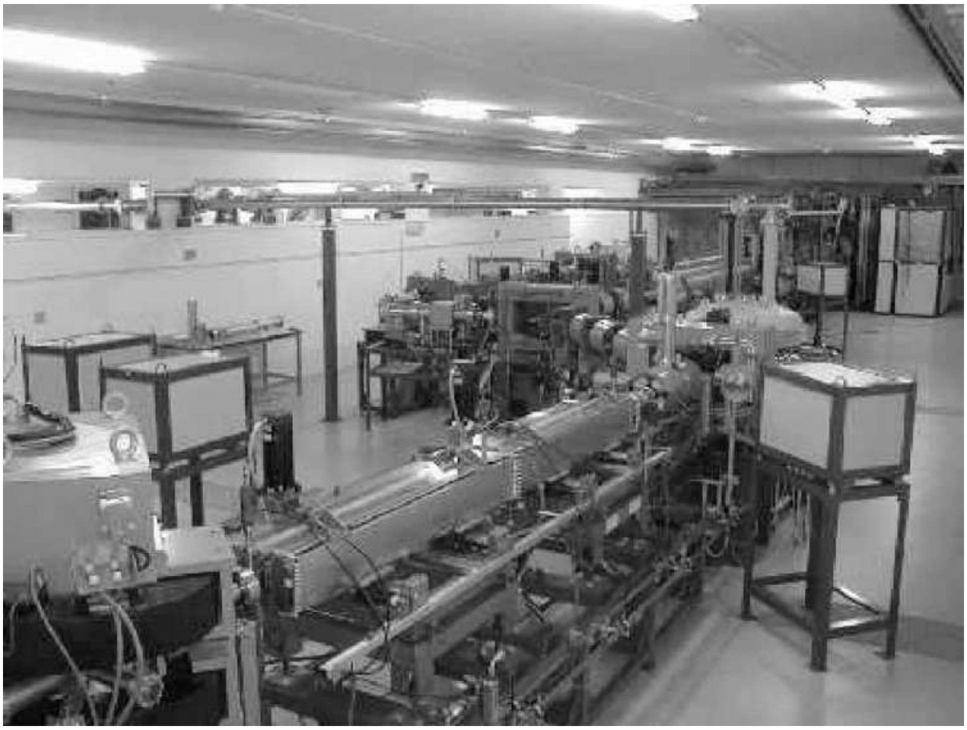

Figure 15 The FELIX RF-Linac FEL operating as a radiation users center in F.O.M. Netherlands. (Courtesy of L. van der Meer, F.O.M.)

This limits the FEL power to be kept as a fraction of the synchrotron radiation power dissipation all around the ring (the 'Renieri Limit'). The effect of the FEL on the e-beam quality, reduces the lifetime of the electrons in the storage ring, and so distrupts the normal operation of the ring in a synchrotron radiation user facility.

To avoid the interference problems, it is most desirable to operate FELs in a dedicated storage ring. This also provides the option to leave long enough straight sections in which long enough wigglers provide sufficient gain for FEL oscillation. Figure 16 displays the Duke storage ring FEL, which is used as a unique radiation user facility, providing intense coherent short wavelength radiation for applications in medicine, biology, material studies, etc.

Superconducting (SC) RF-LINACS

When the RF cavities of the accelerator are superconducting, there are very low RF power losses on the cavity walls, and it is possible to maintain continuous acceleration field in the RF accelerator with a moderate-power continuous RF source, which delivers all of its power to the electron beam kinetic energy. Combining the SC-RF-LINAC technology with an FEL oscillator, pioneered primarily by Stanford University and Thomas Jefferson Lab (TJL) in the US and JAERI Lab in Japan, gave rise to an important scheme of operating such a system in a current recirculating energy retrieval mode. This scheme revolutionized the development of FELs in the direction of high-power, high-efficiency operation, which is highly desirable, primarily for industrial applications (material processing, photochemical production, etc.).

In the recirculating SC-RF-LINAC FEL scheme the wasted beam emerging out of the wiggler after losing a fraction of only few percents (see eqn [89]) out of its kinetic energy, is not dumped into a beam-dump, as in normal cavity RF accelerators, but is re-injected, after circulation, into the SC-RF accelerator. The timing of the wasted electron bunches re-injection is such that they experience a deceleration phase along the entire length of the accelerating cavities. Usually, they are re-injected at the same cell with a fresh new electron bunch injected at an acceleration phase, and thus the accelerated fresh bunch receives its acceleration kinetic energy directly from the wasted beam bunch, that is at the same time decelerated. The decelerated wasted beam bunches are then dumped in the electron beam dump at much lower energy than without recirculation, at energies that are limited primarily just by the energy spread induced in the beam in the FEL laser-saturation process. This scheme, not only increases many folds the over-all energy transformation efficiency from wall-plug to radiation, but would solve significant heat dissipation and radioactivity activation problems in a high-power FEL design.

Figure 17 displays the TJL Infrared SC-RF-LINAC FEL oscillator, that demonstrated for the first time

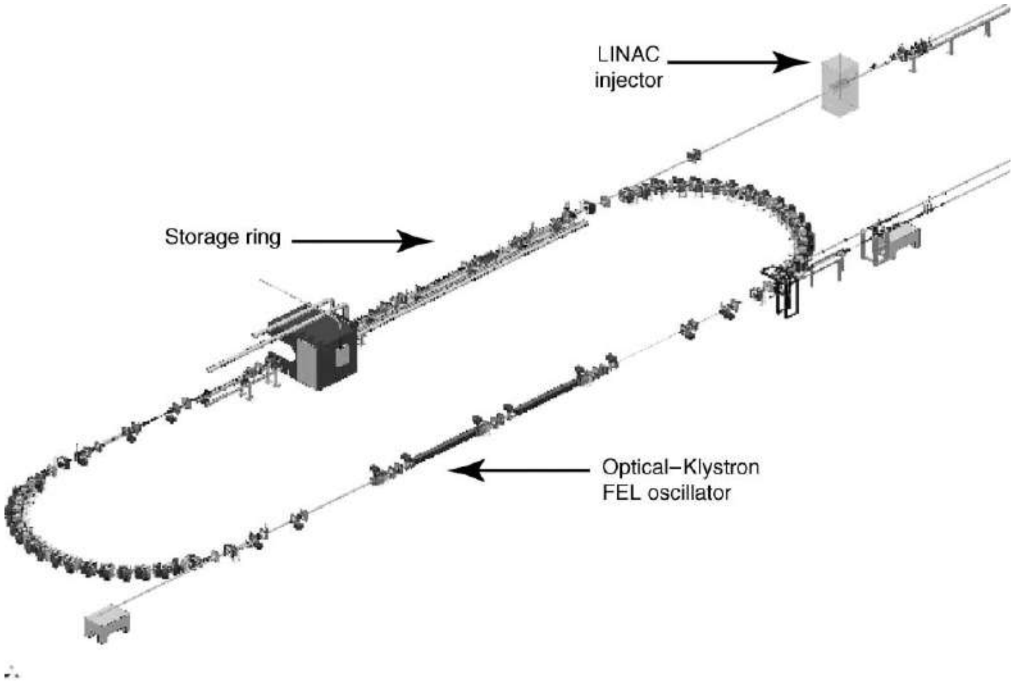

Figure 16 The Duke – University Storage Ring FEL operating as a radiation-users center in N. Carolina, USA. (Mendening: Matthew Busch, courtesy of Glenn Edwards, Duke FEL Lab.)

record high average power levels – nearly 10 kWatt at optical frequencies (1–14 μm). The facility is in upgrade development stages towards eventual operation at 100 kWatt in the IR and 1 kWatt in the UV. It operates in the framework of a laser material processing consortium and demonstrates important material processing applications, such as high-rate micromachining of hard materials (ceramics) with picoSecond laser pulses.

The e-beam current recirculation scheme of SC-RF-LINAC FEL has a significant advantage over the e-beam recirculation in a storage ring. As in electrostatic accelerators, the electrons entering the wiggler are 'fresh' cold-beam electrons from the injector, and not a wasted beam corrupted by the laser saturation process in a previous circulation through the FEL. This also makes it possible to sustain high average circulating current despite the disruptive effect of the FEL on the e-beam. This technological development has given rise to a new concept for a radiation-user facility light-source 4GLS (fourth-generation light source), which is presently in a pilot project development stage at Daresbury Lab in the UK (see Figure 18). In such a scheme, IR and UV FEL oscillators and XUV SASE-FEL can be operated together with synchrotron magnet dipole and wiggler insertion devices without disruptive interference. Such a scheme, if further developed, can give rise to new radiation-user, light-source facilities, that can provide a wider range of radiation parameters than synchrotron centers of previous generation.

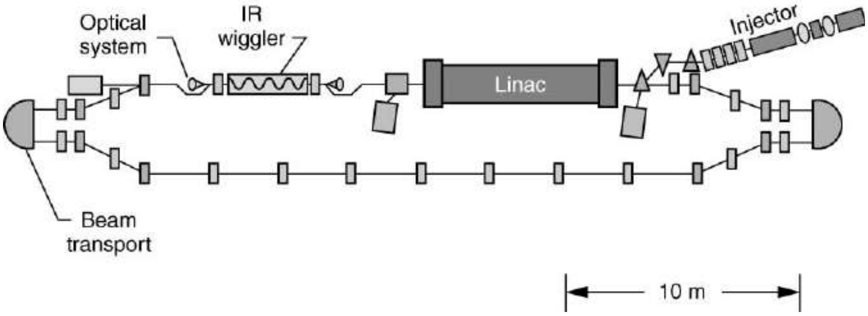
Optical system
IR wiggler
Injector
Linac
Beam transport
10 m

Figure 17 The Thomas Jefferson Lab. recirculating beam-current superconducting Linac FEL operating as a material processing FEL-user center in Virginia USA (Courtesy of S. Benson, Thomas Jefferson Laboratory).

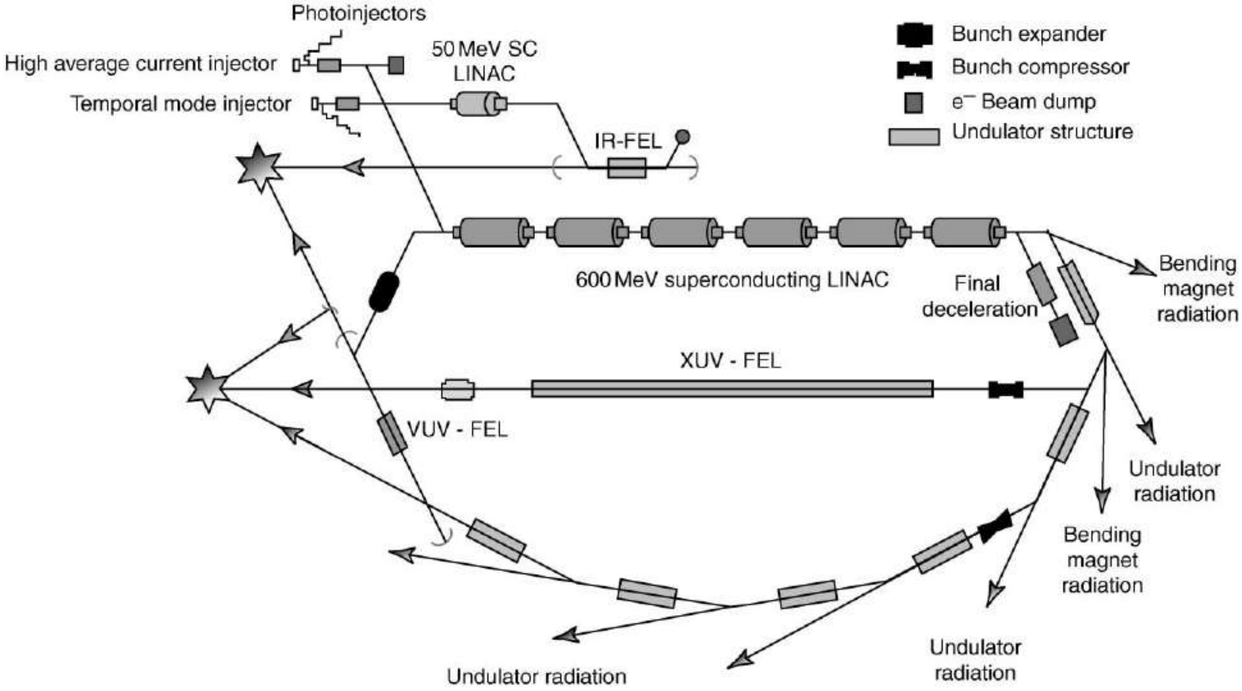
Photoinjectors
High average current injector
Temporal mode injector
50 MeV SC LINAC
IR-FEL
Bunch expander
Bunch compressor
e⁻ Beam dump
Undulator structure
600 MeV superconducting LINAC
Final deceleration
Bending magnet radiation
XUV - FEL
VUV - FEL
Undulator radiation
Bending magnet radiation
Undulator radiation
Undulator radiation

Figure 18 The Daresbury Fourth Generation Light-Source concept (4GLS). The circulating beam-current superconducting Linac includes SASE-FEL, bending magnets and wigglers as insertion devices. (Courtesy of M. Poole, Damesbury Laboratory)

Magnetic Wiggler Schemes

The optical klystron

The stimulated emission process in FEL (see Figure 7c) is based on velocity (energy) bunching of the e-beam in the first part of the wiggler, which turns into density bunching along the central part of the wiggler, and then the density-bunched electron beam performs 'negative work' on the radiation wave and emits radiative energy in the last part of the wiggler. In the OK, these steps are carried out in three separate parts of the wiggler: the energy bunching wiggler section, the dispersive magnet density buncher, and the radiating wiggler section (see Figure 7b).

A schematic of the OK is shown in Figure 19. The chicane magnetic structure in the dispersive section brings all electrons emerging from the bunching wiggler back onto the axis of the radiating wiggler, but provides variable delay $\Delta t_{\mathrm{d}i} = \int_{L_{\mathrm{b}}}^{L_{\mathrm{b}}+L_{\mathrm{d}}} (v_{zi}^{-1} - v_{\mathrm{ph}}^{-1})\,\mathrm{d}z = [\mathrm{d}(\Delta t_{\mathrm{d}})/\mathrm{d}\gamma]\delta\gamma_i$ relative to the pondermotive wave phase to different electrons, which acquired different energy modulation increments $\delta\gamma_i = \gamma_i - \gamma_{\mathrm{ph}}$ in the final section. The radiation condition is satisfied whenever the bunch-center phase satisfies $\Delta\varphi_{\mathrm{d}} = \omega\Delta t_{\mathrm{d}} = \pi/2 - 2m\pi$ (see Figure 7b). However, because the energy dispersion coefficient $\mathrm{d}(\Delta t_{\mathrm{d}})/\mathrm{d}\gamma$, is much larger in the chicane than in a wiggler of the same length, the density bunching amplitude, and consequently the OK gain, are much larger than in a uniform wiggler FEL of the same length.

The OK was invented by Vinokurov and Skrinsky in 1977 and first demonstrated in 1987 at visible wavelengths in the ACO storage ring of LURE in Orsay, France, and subsequently in 1988 at UV wavelengths, in the VEPP storage ring in Novosibirsk, Russia. The OK is an optimal FEL configuration, if used as an insertion device in a storage ring, because it can provide sufficient gain to exceed the high lasing threshold at the short operating wavelengths of a high-energy storage-ring FEL, and still conform with the rather short straight sections available for insertion devices in conventional synchrotron storage rings. It should be noted that the OK is equivalent to a long wiggler FEL of length L_{eff} of equal gain and therefore its axial velocity spread acceptance is small (this is determined from the cold beam limit $\bar{\theta}_{\mathrm{th}} \ll \pi$ with L_{eff} used in eqn [52]). This too is consistent with storage ring accelerators, which are characterized by small energy spread and emittance of the electron beam.

Radiation emission at harmonic frequencies

In a linear wiggler (eqn [2]), the axial velocity:

$$\beta_z = [\beta^2 - (a_{\mathrm{w}}/\gamma)^2 \cos^2 k_{\mathrm{w}} z]^{1/2} \qquad [92]$$

is not constant. It varies with spatial periodicity $\lambda_{\mathrm{w}}/2$, and in addition to its average value $\bar{\beta}_z = [\beta^2 - a_{\mathrm{w}}^2/2\gamma^2]^{1/2}$, contains Fourier components of spatial frequencies $2mk_{\mathrm{w}}$ $(m = 1, 2, \ldots)$. When $a_{\mathrm{w}} \gg 1$, the axial oscillation deforms the sinusoidal trajectory of the electrons in the wiggler (eqns [22] and [23]), and in a frame of reference moving at the average velocity $\bar{\beta}_z$, the electron trajectories in the wiggling $(x-z)$ plane forms an figure 8 shape, rather than a pure transverse linear motion. In the laboratory frame this leads to synchrotron undulator emission in the forward direction at all odd harmonic frequencies of ω_0, corresponding to substitution of $k_{\mathrm{w}} \to (2m+1)k_{\mathrm{w}}$ $(m = 1, 2, 3, \ldots)$ in eqn [6]:

$$\omega_{2m+1} = (2m+1)\omega_0 \cong 2\gamma_z^2 c(2m+1)k_{\mathrm{w}} \qquad [93]$$

All the stimulated emission gain expressions, presented earlier for the fundamental harmonic, are valid with appropriate substitution of

$$\theta_{2m+1} = \frac{\omega}{v_z} - k_z - (2m+1)k_{\mathrm{w}} \qquad [94]$$

instead of eqn [43], and substitution of the harmonic-weight Bessel-function coefficient of

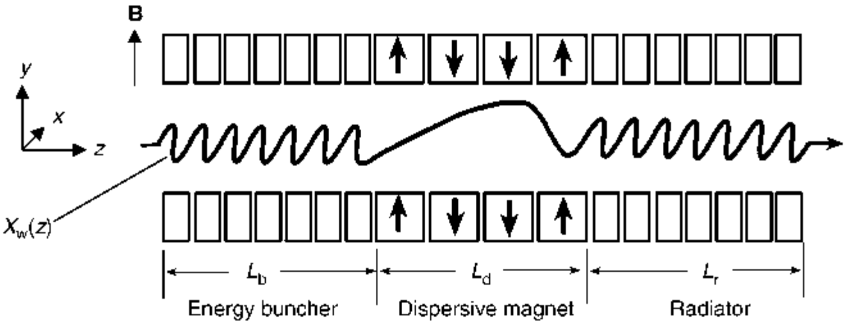

Figure 19 Schematics of the Optical–Klystron, including an energy bunching wiggler, a dispersive magnet bunching section and a radiating wiggler.

synchronism with the beam. Slowing down the PM wave can be done by the gradual increase of the wiggler wavenumber $k_w(z)$ (or decrease of its period $\lambda_w(z)$), so that eqns [29] or [91] keep being satisfied for a given frequency, even if v_z (or γ_z) goes down.

A more correct description of the nonlinear interaction dynamics of the electron beam in a saturated tapered-wiggler FEL is depicted in Figure 20: the electron trap synchronism energy $\gamma_{ph}(z)$ tapers down (by design) along the wiggler, while the trapped electrons are forced to slow down with it, releasing their excess energy by enhanced radiation. An upper limit estimate for the extraction efficiency of such a tapered wiggler FEL would be:

$$\eta_{ext} = \frac{\gamma_{ph}(0) - \gamma_{ph}(L)}{\gamma_{ph}(0)} \qquad [99]$$

and the corresponding radiative power generation would be: $\Delta P = \eta_{ext} I_b E_k / e$. In practice, the phase-space area of the tapered wiggler separatrix is reduced due to the tapering, and only a fraction of the electron beam can be trapped, which reduces correspondingly the practical enhancement in radiative extraction efficiency and power.

An alternative wiggler tapering scheme consists of tapering the wiggler field $B_w(z)$ (or wiggler parameter amplitude $a_w(z)$). If these are tapered down, the axial velocity and axial energy (eqn [7]) can still keep constant (and in synchronism with the PM wave) even if the beam energy γ goes down. Thus, in this scheme, the excess radiative energy extracted from the beam comes out of its transverse (wiggling) energy.

Efficiency and power enhancement of FEL by wiggler tapering have been demonstrated experimentally both in FEL amplifiers (first by Livermore, 1985) and oscillators (first by Los-Alamos, 1983). This elegant way to extract more power from the beam still has some limitations. It can operate efficiently only at a specified high radiation power level for which the tapering was designed. In an oscillator, a long enough untapered section must be left to permit sufficient small signal gain in the early stages of the laser oscillation build-up process.

FEL Oscillators

Most FEL devices are oscillators. As in any laser, in order to turn the FEL amplification process into an oscillation process, one provides a feedback mechanism by means of an optical resonator. In steady state saturation, $GR_{rt} = 1$, where R_{rt} is the round trip reflectivity factor of the resonator and $G = P(L)/P(0)$ is the saturated single-path gain coefficient of the FEL. To attain oscillation, the small signal (unsaturated) gain, usually given by the small gain expression in eqn [53], must satisfy the lasing threshold condition $G > 1/R_{rt}$, as in any laser.

When steady state oscillation is attained, the oscillator output power is:

$$P_{out} = \frac{T}{1 - R_{rt}} \Delta P_{ext} \qquad [100]$$

where $\Delta P_{ext} = \eta_{ext} I_0 (\gamma_0 - 1) mc^2 / e$ and η_{ext} is the extraction efficiency, usually given by eqn [89] (low-gain limit).

Usually, FEL oscillators operate in the low-gain regime, in which case $1 - R_{rt} = L + T \ll 1$ (where L is the resonator internal loss factor). Consequently, then $P_{out} \cong \Delta P_{ext} T/(L + T)$, which would give a maximum value, depending on the saturation level of the oscillator. In the general case, one must solve the nonlinear force equations together with the resonator feedback relations of the oscillating radiation mode, in order to maximize the output power (eqn [100]) or efficiency by choice of optimal T for given L.

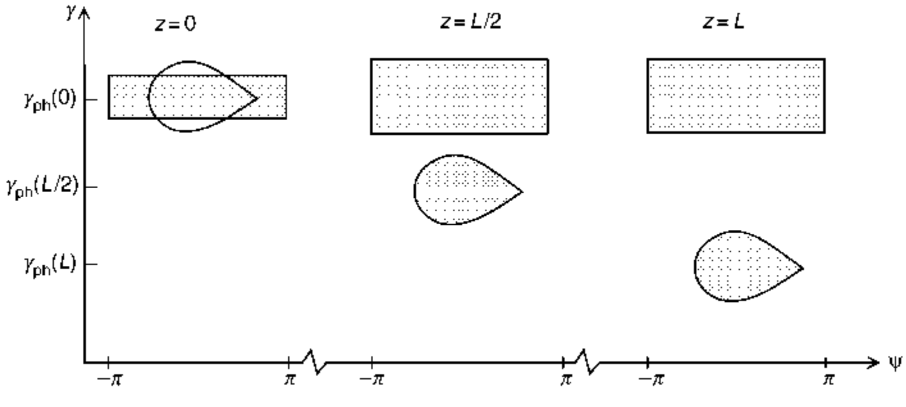

Figure 20 'Snapshots' of the trap at three locations along a tapered wiggler FEL.

In an FEL oscillator operating with periodic electron bunches (as in RF-acclerator based FEL), the solution for the FEL gain and saturation dynamics requires extension of the single frequency solution of the electron and electromagnetic field equations to the time domain. In principle, the situation is similar to that of a mode-locked laser, and the steady state laser pulse train waveform constitutes a superposition of the resonator longitudinal modes that produces a self-similar pulse shape with the highest gain (best overlap with the e-beam bunch along the interaction length). Because the e-beam velocity v_{z0} is always smaller (in an open resonator) than the group velocity of the circulating radiation wavepacket, the radiation wavepacket slips ahead of the electron bunch one optical period λ in each wiggling period (Slippage effect). This reduces the overlap between the radiation pulse and the e-beam bunch along the wiggler (see Figure 14) and consequently decreases the gain. Fine adjustment of the resonator mirrors (as shown in Figure 14) is needed to attain maximal power and optimal radiation pulse shape. The pulse-slippage gain reduction effect is negligible only if the bunch length is much longer than the slippage length $N_w\lambda$, which can be expressed as:

$$\tau_p \gg 2\pi/\Delta\omega_L \qquad [101]$$

where $\Delta\omega_L$ is the synchrotron undulator radiation frequency bandwidth (eqn [55]). This condition is usually not satisfied in RF-accelerator FELs operating in the IR or lower frequencies, and the

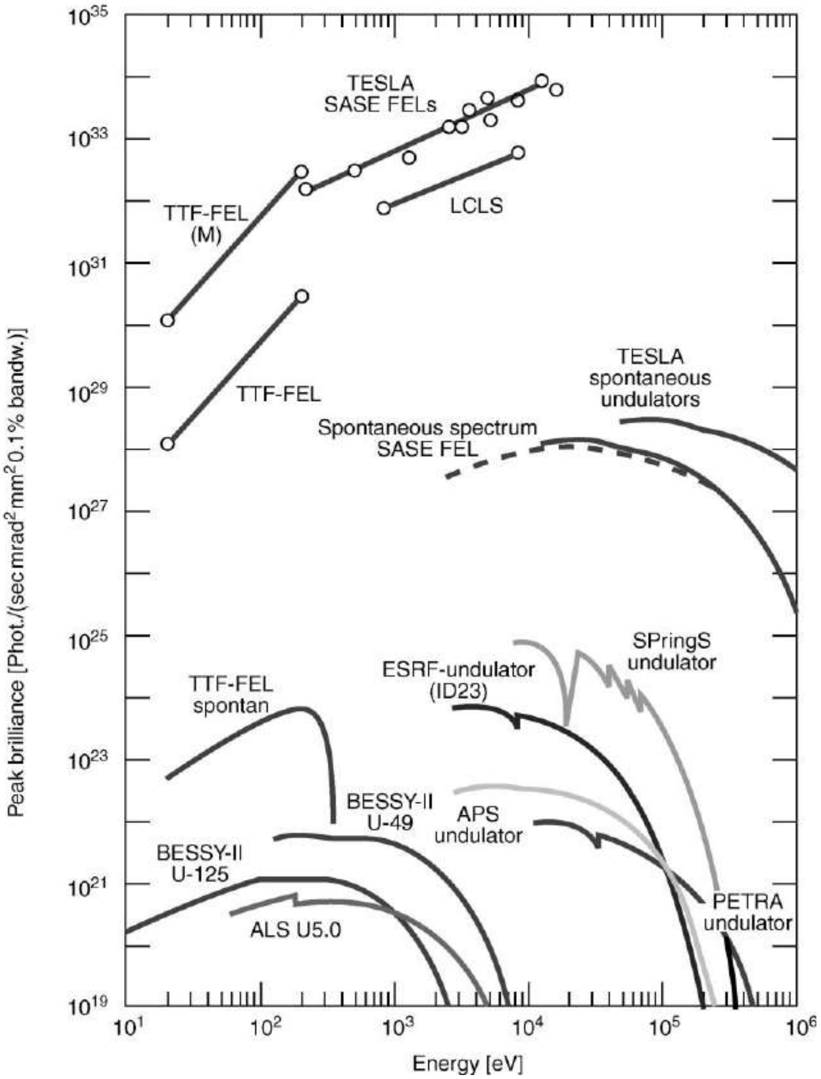

Figure 21 Anticipated peak brightness of SASE FELs (TTF-DESY, LCLS-SLAC) in comparison to the undulators in present third generation Synchrotron Radiation sources. Figure courtesy of DESY, Hamburg, Germany.

slippage effect gain reduction must be then taken into account.

An FEL operating in the cold-beam regime constitutes an 'homogeneous broadening' gain medium in the sense of conventional laser theory. Consequently, the longitudinal mode competition process that would develop in a CW FEL oscillator, leads to single-mode operation and high spectral purity (temporal coherence) of the laser radiation. The minimal (intrinsic) laser linewidth would be determined by an expression analogous to the Schawlow–Towns limit of atomic laser:

$$(\Delta f)_{\mathrm{int}} = \frac{(\Delta f_{1/2})^2}{I_b/e} \qquad [102]$$

where $\Delta f_{1/2}$ is the spectral width of the cold resonator mode. Expression [102] predicts extremely narrow linewidth. In practice, CW operation of FEL was not yet attained, but Fourier transform limited linewidths in the range of $\Delta f/f_0 \cong 10^{-6}$ were measured in long-pulse electrostatic accelerator FELs. In an FEL oscillator, based on a train of e-beam bunches (e.g., an R.F. accelerator beam), the linewidth is very wide and is equal to the entire gain bandwidth (eqn [56]) in the slippage dominated limit, and to the Fourier transform limit $\Delta\omega \cong 2\pi/\tau_p$ in the opposite negligible-slippage limit (eqn [101]). Despite this slippage, it was observed in RF-LINAC FEL that the radiation pulses emitted by the FEL oscillator are phase corrected with each other, and therefore their total temporal coherence length may be as long as the

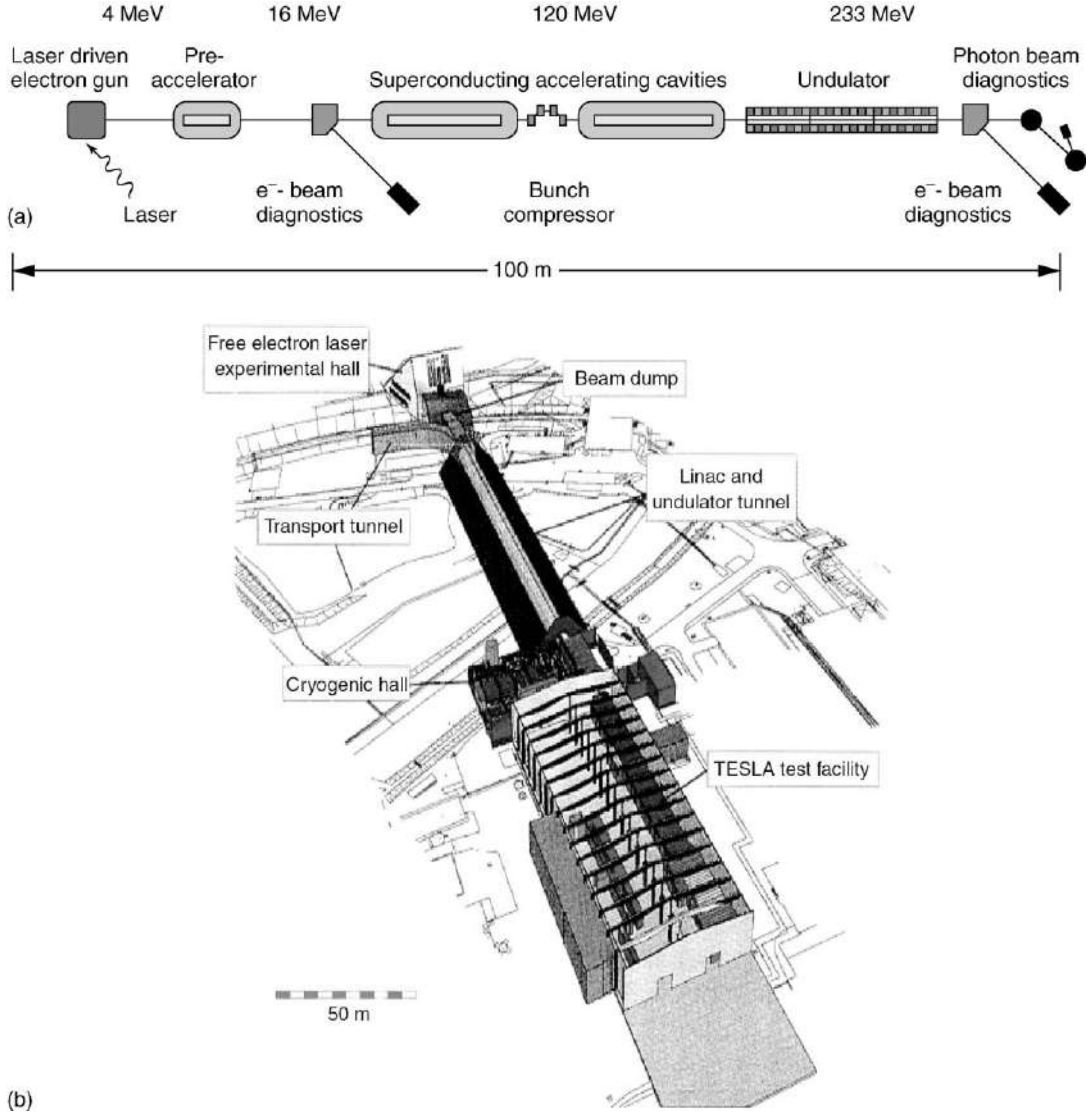
4 MeV
16 MeV
120 MeV
233 MeV
Laser driven electron gun
Pre-accelerator
Superconducting accelerating cavities
Undulator
Photon beam diagnostics
(a)
Laser
e⁻- beam diagnostics
Bunch compressor
e⁻- beam diagnostics
100 m
Free electron laser experimental hall
Beam dump
Linac and undulator tunnel
Transport tunnel
Cryogenic hall
TESLA test facility
50 m
(b)

Figure 22 Phase 1 of the SASE FE L (TTF VUV-FEL1): (a) Accelerator layout scheme; (b) General view of the TESLA test facility. Figure courtesy of DESY, Hamburg, Germany.

Metal Vapor Lasers

D W Coutts, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Metal vapor lasers form a class of laser in which the active medium is a neutral or ionized metal vapor usually excited by an electric discharge. These lasers fall into two main subclasses, namely cyclic pulsed metal vapor lasers and continuous-wave metal ion lasers. Both types will be considered in this article, including basic design and construction, power supplies, operating characteristics (including principal wavelengths), and brief reference to their applications.

Self-Terminating Resonance-Metastable Pulsed Metal Vapor Lasers

The active medium in a self-terminating pulsed metal vapor laser consists of metal atoms or ions in the vapor phase usually as a minority species in an inert buffer gas such as neon or helium. Laser action occurs between a resonance upper laser level and a metastable lower laser level (Figure 1). During a fast pulsed electric discharge (typically with a pulse duration of order 100 ns) the upper laser level is preferentially excited by electron impact excitation because it is strongly optically connected to the ground state (resonance transition) and hence has a large excitation cross-section. For a sufficiently large metal atom (or ion) density, the resonance radiation becomes optically trapped, thus greatly extending the lifetime of the upper laser level such that decay from the upper laser level is channelled through the emission of laser radiation to the metastable lower laser level. Lasing terminates when the electron temperature falls to a point such that preferential pumping to the upper laser level is no longer sustained, and the build-up of population in the metastable lower laser level destroys the population inversion. Therefore after each excitation pulse, the resulting excited species in the plasma (in particular the metastable lower laser levels which are quenched by collisions with cold electrons) must be allowed sufficient time to relax and the plasma must be allowed to partially recombine before applying the next excitation pulse. The relaxation times for self-terminating metal vapor lasers correspond to operating pulse repetition frequencies from 2 kHz to 200 kHz.

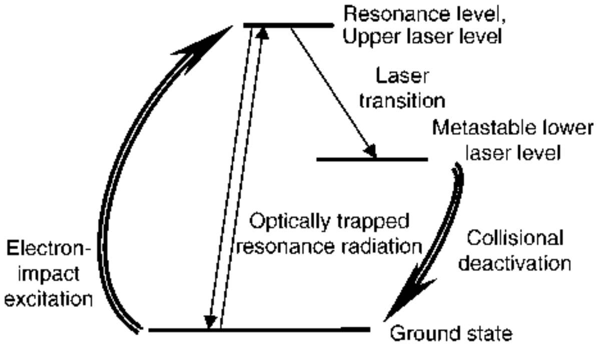
Resonance level,
Upper laser level
Laser transition
Metastable lower laser level
Optically trapped resonance radiation
Collisional deactivation
Electron-impact excitation
Ground state

Figure 1 Resonance-metastable energy levels for self-terminating metal vapor lasers.

Many metal vapors can be made to lase in the resonance-metastable scheme and are listed together with their principal wavelengths and output powers in Table 1. The most important self-terminating pulsed metal vapor laser is the copper vapor laser and its variants which will be discussed in detail in the following sections. Of the other pulsed metal vapor lasers listed in Table 1, only the gold vapor laser (principal wavelengths 627.8 nm and 312.3 nm), and the barium vapor laser which operates in the infrared (principal wavelengths 1.5 μm and 2.55 μm) have had any commercial success. All the self-terminating pulsed metal vapor lasers have essentially the same basic design and operating characteristics as exemplified by the copper vapor laser.

Copper Vapor Lasers

Copper vapor lasers (CVLs) are by far the most widespread of all the pulsed metal vapor lasers. Figure 2 shows the energy level scheme for copper. Lasing occurs simultaneously from the $^2P_{3/2}$ level to the $^2D_{5/2}$ level (510.55 nm) and from the $^2P_{1/2}$ level to the $^2D_{3/2}$ level (578.2 nm). Commercial devices are available with combined outputs of over 100 W at 510.55 nm and 578.2 nm (typically with a green-to-yellow power ratio of 2:1).

A typical copper vapor laser tube is shown in Figure 3. High-purity copper pieces are placed at intervals along an alumina ceramic tube which typically has dimensions of 1–4 cm diameter and 1–2 m long. The alumina tube is surrounded by a solid fibrous alumina thermal insulator, and a glass or quartz vacuum envelope. Cylindrical electrodes made of copper or tantalum are located at each end of the plasma tube to provide a longitudinal discharge

Table 1 Principal self-terminating resonance-metastable metal vapor lasers

Metal	*Principal wavelengths (nm)*	*Powers*		*Total efficiencies*	*Pulse repetition frequency (kHz)*	*Technological development*
		Typical (W)	*Maximum (W)*			
Cu	510.55	2–70	2500 total	1%	4–40	Highly developed and commercially available
	578.2	1–50				
Au	627.8	1–8	20	0.13%	2–40	Commercially available
	312	0.1–0.2	1.2		1–8	
Ba	1500	2–10	12	0.5%	5–15	Have been produced commercially, but largely experimental
	2550	1	1.5			
	1130	0.5	1.0			
Pb	722.9		4.4	0.15%	10–30	Experimental
Mn	534.1 (>50%)		12 total	0.32%	~10	Experimental
	1290					

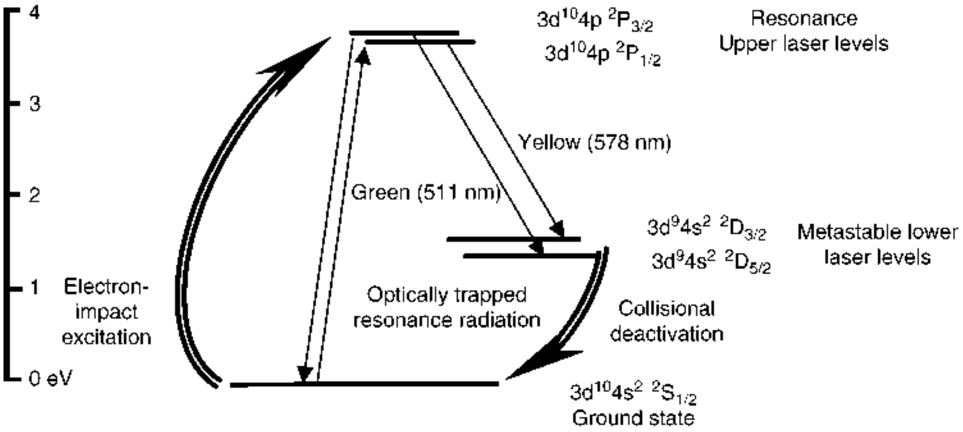
4
3
2
1
0 eV
$3d^{10}4p\ ^2P_{3/2}$
$3d^{10}4p\ ^2P_{1/2}$
Resonance
Upper laser levels
Yellow (578 nm)
Green (511 nm)
$3d^94s^2\ ^2D_{3/2}$
$3d^94s^2\ ^2D_{5/2}$
Metastable lower
laser levels
Electron-
impact
excitation
Optically trapped
resonance radiation
Collisional
deactivation
$3d^{10}4s^2\ ^2S_{1/2}$
Ground state

Figure 2 Partial energy level scheme for the copper vapor laser.

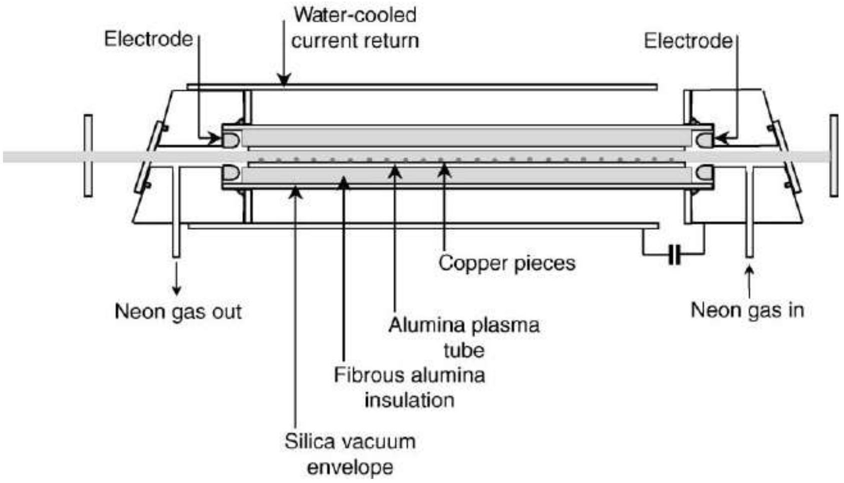
Electrode
Water-cooled
current return
Electrode
Copper pieces
Neon gas out
Neon gas in
Alumina plasma
tube
Fibrous alumina
insulation
Silica vacuum
envelope

Figure 3 Copper vapor laser tube construction.

arrangement. The cylindrical electrodes and silica laser end windows are supported by water-cooled end pieces. The laser windows are usually tilted by a few degrees to prevent back reflections into the active medium. The laser head is contained within a water cooled metal tube to provide a coaxial current return for minimum laser head inductance. Typically a slow flow (~5 mbar l min^{-1}) of neon at a pressure of 20–80 mbar is used as the buffer gas with an approximately 1% H_2 additive to improve the afterglow plasma relaxation. The buffer gas provides a medium to operate the discharge when the laser is cold and slows diffusion of copper vapor out of the ends of the hot plasma tube. Typical copper fill times are of order 200–2000 hours (for 20–200 g copper load). Sealed-off units with lifetimes of order 1000 hours have been in production in Russia for many years.

During operation waste heat from the repetitively pulsed discharge heats the alumina tube up to approximately 1500 °C at which point the vapor pressure of copper is of approximately 0.5 mbar which corresponds to the approximate density required for maximum laser output power. Typical warm-up times are therefore relatively long at around one hour to full power.

One way to circumvent the requirement for high temperatures required to produce sufficient copper density by evaporation of elemental copper (and hence also reduce warm-up times) is to use a copper salt with a low boiling point located in one or more side-arms of the laser tube (Figure 4). Usually copper halides are used as the salt, with the copper bromide laser being the most successful. For the CuBr laser, a temperature of just 600 °C is sufficient to produce the required Cu density by dissociation of CuBr vapor in the discharge. With the inclusion of 1–2% H_2 in the neon buffer gas, HBr is also formed in the CuBr laser, which has the additional benefit of improving recombination in the afterglow via dissociative attachment of free electrons: $HBr + e^- \rightarrow H + Br^-$, followed by ion neutralization: $Br^- + Cu^+ \rightarrow Br + Cu^*$. As a result of the lower operating temperature and kinetic advantages of HBr, CuBr lasers are typically twice as efficient (2–3%) as their elemental counterparts. Sealed-off CuBr systems with powers of order 10–20 W are commercially produced.

An alternative technique for reducing the operating temperature of elemental CVLs is to flow a buffer gas mixture consisting of ~5% HBr in neon at approximately 50 mbar l min^{-1} and allow this to react with solid copper metal placed within the plasma tube at about 600 °C to produce CuBr vapor *in situ*. The so-called Cu HyBrID (hydrogen bromide in discharge) laser has the same advantages as the CuBr laser (e.g., up to 3% efficiency) but at the cost of requiring flowing highly toxic HBr in the buffer gas, a requirement which has so far prevented commercialization of Cu HyBrID technology.

The kinetic advantages of the hydrogen halide in the CuBr laser discharge can also be applied to a conventional elemental CVL through the addition of small partial pressure of HCl to the buffer gas in addition to the 1–2% H_2 additive. HCl is preferred to HBr as it is less likely to dissociate (the dissociation energy of HCl at 0.043 eV is less than HBr at 0.722 eV). Such kinetic enhancement leads to a doubling in average output power, a dramatic increase in beam quality through improved gain characteristics, and shifts the optimum pulse repetition frequency for kinetically enhanced CVLs (KE-CVLs) from 4–10 kHz up to 30–40 kHz.

Preferential pumping of the upper laser levels requires an electron temperature in excess of the 2 eV range, hence high voltage (10–30 kV), high current (hundreds of A), short (75–150 ns) excitation pulses are required for efficient operation of copper vapor lasers. To generate such excitation pulses CVLs are typically operated with a power supply incorporating a high-voltage thyratron switch. In the most basic configuration, the charge-transfer circuit shown

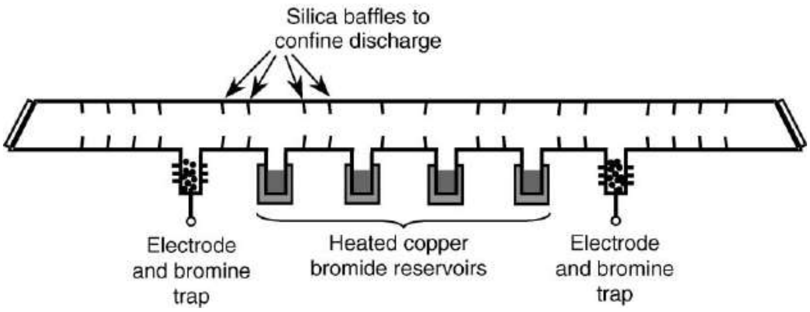
Silica baffles to confine discharge
Electrode and bromine trap
Heated copper bromide reservoirs
Electrode and bromine trap

Figure 4 Copper bromide laser tube construction.

in Figure 5a, a dc high-voltage power supply resonantly charges a storage capacitor (C_S, typically a few nF) through a charging inductor L_C, a high-voltage diode and bypass inductor L_B, up to twice the supply voltage V_S in a time of order 100 μs. When the thyratron is triggered, the storage capacitor discharges through the thyratron and the laser head on a time-scale of 100 ns. Note that during the fast discharge phase, the bypass inductor in parallel with the laser head can be considered to be an open circuit. A peaking capacitor C_P ($\sim 0.5\ C_S$) is provided to increase the rate of rise of the voltage pulse across the laser head. Given the relatively high cost of thyratrons, more advanced circuits are now often used to extend the service lifetime of the thyratron to several thousand hours. In the more advanced circuit (Figure 5b) an LC inversion scheme is used in combination with magnetic pulse compression techniques and operates as follows. Storage capacitors C_S are resonantly charged in parallel to twice the dc supply voltage V_S as before. When the thyratron is switched, the charge on C_{S1} inverts through the thyratron and transfer inductor L_T. This LC inversion drags the voltage on the top of C_{S2} down to $-4V_S$. When the voltage across the first saturable inductor L_{S1} reaches a maximum ($-4V_S$) L_{S1} saturates, and allows current to flow from the storage capacitors (now charged in series) to the transfer capacitor ($C_T = 0.5C_{S2}$) thereby transferring the charge from C_{S1} and C_{S2} to C_T in a time much less than the initial LC inversion time. At the moment when the charge on transfer capacitor C_T reaches a maximum (also $-4V_S$), L_{S2} saturates, and the transfer capacitor is discharged through the laser tube, again with a peaking capacitor to increase the voltage rise time. By using magnetic pulse compression the thyratron switched voltage can be reduced by 4 and the peak current similarly reduced (at the expense of increased current pulse duration) thereby greatly extending the thyratron lifetime. Note that in both circuits a 'magnetic assist' L_A saturable inductor is provided in series with the thyratron to delay the current pulse through the thyratron until after the thyratron has reached high conductivity thereby reducing power deposition in the thyratron.

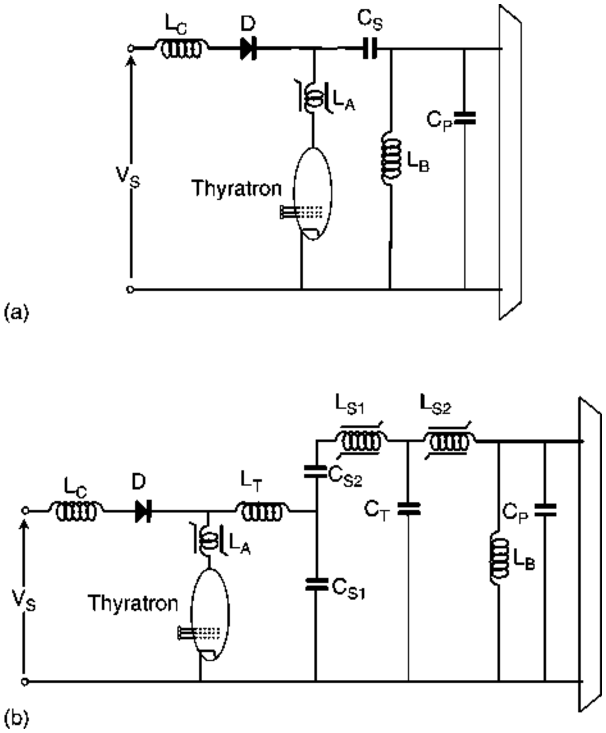

Figure 5 Copper vapor laser excitation circuits. (a) Charge transfer circuit; (b) LC inversion circuit with magnetic pulse compression.

Copper vapor lasers produce high average powers (2–100 W available commercially, with laboratory devices producing average powers of over 750 W) and have wall-plug efficiencies of approximately 1%. Copper vapor lasers also make excellent amplifiers due to their high gains, and the amplifiers can be chained together to produce average powers of several kW. Typical pulse repetition frequencies range from 4 to 20 kHz, with a maximum reported pulse repetition frequency of 250 kHz. An approximate scaling law states that for an elemental device with tube diameter D (mm) and length L (m) the average output power in watts will be of order $D \times L$. For example, a typical 25 W copper vapor laser will have a 25 mm diameter by 1 m long laser tube, and operate at 10 kHz corresponding to 2.5 mJ pulse energy and 50 kW peak power (50 ns pulse duration). Copper vapor lasers have very high single-pass gains (greater than 1000 for a 1 m long tube), large gain volumes and short gain durations (20–80 ns, sufficient for the intracavity laser light to make only a few round trips within the optical resonator). Maximum output power is therefore usually obtained in a highly 'multimode' (spatially incoherent) beam by using either a fully stable or a plane–plane resonator with a low reflectivity output coupler (usually Fresnel reflection from an uncoated optic is sufficient). To obtain higher beam quality a high magnification unstable resonator is required (Figure 6). Fortunately, copper vapor lasers have sufficient gain to operate efficiently with unstable resonators with magnifications ($M = R_1/R_2$) up to 100 and beyond. Resonators with such high magnifications impose very tight geometric constraints on propagation of radiation on repeated round-trips within the resonator, such that after two round-trips the divergence is typically diffraction-limited. Approximately half the stable-resonator

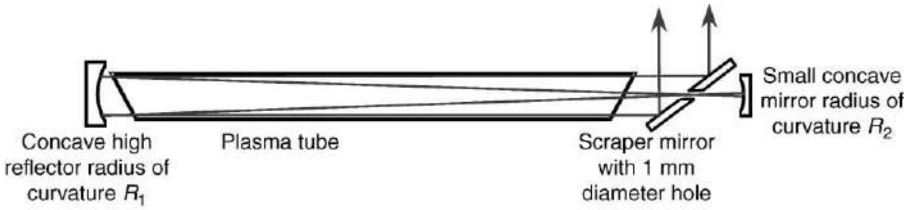
Concave high reflector radius of curvature R_1
Plasma tube
Scraper mirror with 1 mm diameter hole
Small concave mirror radius of curvature R_2

Figure 6 Unstable resonator configuration often used to obtain high beam quality from copper vapor lasers.

output power can therefore be obtained with near diffraction-limited beam quality by using an unstable resonator. Often, a small-scale oscillator is used in conjunction with a single power amplifier to produce high output power with diffraction-limited beam quality. Hyperfine splitting combined with Doppler broadening lead to an inhomogeneous linewidth for the laser transitions of order 8–10 GHz corresponding to a coherence length of order 3 cm.

The high beam quality and moderate peak power of CVLs allows efficient nonlinear frequency conversion to the UV by second harmonic generation (510.55 nm → 255.3 nm, 578.2 nm → 289.1 nm) and sum frequency generation (510.55 nm + 578.2 nm → 271.3 nm) using β-barium borate β-BaB_2O_4 (BBO) as the nonlinear medium. Typically average powers in excess of 1 W can be obtained at any of the three wavelengths from a nominally 20 W CVL. Powers up to 15 W have been obtained at 255 nm from high-power CVL master-oscillator power-amplifier systems using cesium lithium borate; $CsLi\ B_6O_{10}$ (CLBO) as the nonlinear crystal.

Key applications of CVLs include pumping of dye lasers (principally for laser isotope separation) and pumping Ti:sapphire lasers. Medically, the CVL yellow output is particularly useful for treatment of skin lesions such as port wine stain birth marks. CVLs are also excellent sources of short-pulse stroboscopic illumination for high-speed imaging of fast objects and fluid flows. The high beam quality, visible wavelength and high pulse repetition rate make CVLs very suited to precision laser micromachining of metals, ceramics and other hard materials. More recently, the second harmonic at 255 nm has proved to be an excellent source for writing Bragg gratings in optical fibers.

Afterglow Recombination Metal Vapor Lasers

Recombination of an ionized plasma in the afterglow of a discharge pulse provides a mechanism for achieving a population inversion and hence laser output. The two main afterglow recombination metal vapor lasers are the strontium ion vapor laser (430.5 nm and 416.2 nm) and calcium ion vapor laser (373.7 nm and 370.6 nm) whose output in the violet and UV spectral regions extends the spectral coverage of pulsed metal vapor lasers to shorter wavelengths.

A population inversion is produced by recombination pumping where doubly ionized Sr (or Ca) recombines to form singly ionized Sr (or Ca) in an excited state: $Sr^{++} + e^- + e^- \rightarrow Sr^{+*} + e^-$. Note that recombination rates for a doubly ionized species are much faster than for singly ionized species hence recombination lasers are usually metal ion lasers. Recombination (pumping) rates are also greatest in a cool dense plasma, hence helium is usually used as the buffer gas as it is a light atom hence promotes rapid collisional cooling of the electrons. Helium also has a much higher ionization potential than the alkaline-earth metals (including Sr and Ca) which ensures preferential ionization of the metal species (up to 90% may be doubly ionized).

Recombination lasers can operate where the energy level structure of an ion species can be considered to consist of two (or more) groups of closely spaced levels. In the afterglow of a pulsed discharge electron collisional mixing within each group of levels will maintain each group in thermodynamic equilibrium yielding a Boltzmann distribution of population within each group. If the difference in energy between the two groups of levels (5 eV >> kT) is sufficiently large then thermal equilibrium between the groups cannot be maintained via collisional processes. Given that recombination yields excited singly ionized species (usually with a flux from higher to lower ion levels), it is possible to achieve a population inversion between the lowest level of the upper group and the higher levels within the lower group. This is the mechanism for inversion in the Sr^+ (and analogous Ca^+) ion laser as indicated in the partial energy level scheme for Sr^+ shown in Figure 7.

Strontium and calcium ion recombination lasers have similar construction to copper vapor lasers described above. The lower operating temperatures (500–800 °C) mean that minimal or no thermal insulation is required for self-heated devices.

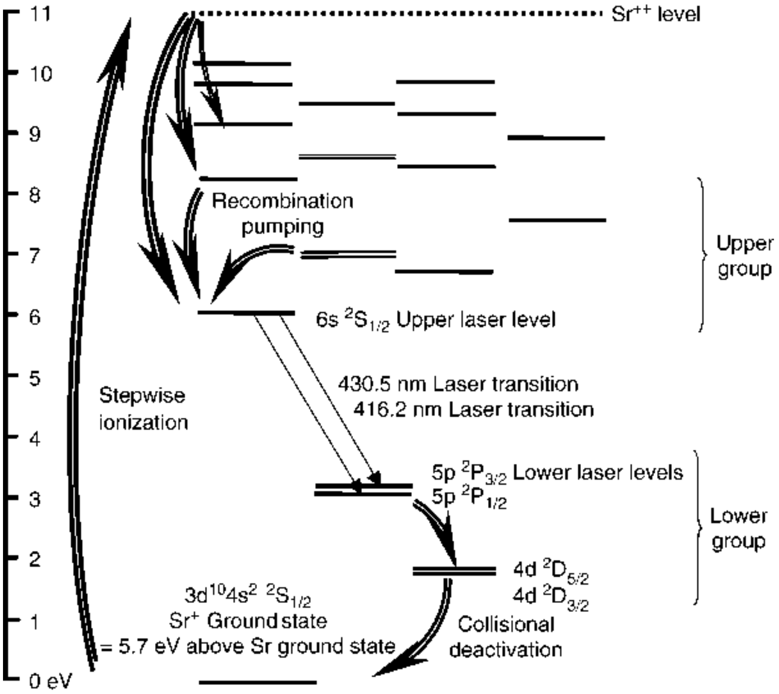
11
10
9
8
7
6
5
4
3
2
1
0 eV
Sr⁺⁺ level
Recombination pumping
6s ²S₁/₂ Upper laser level
Upper group
Stepwise ionization
430.5 nm Laser transition
416.2 nm Laser transition
5p ²P₃/₂ Lower laser levels
5p ²P₁/₂
Lower group
4d ²D₅/₂
4d ²D₃/₂
3d¹⁰4s² ²S₁/₂
Sr⁺ Ground state
= 5.7 eV above Sr ground state
Collisional deactivation

Figure 7 Partial energy level scheme for the strontium ion laser.

For good tube lifetime BeO plasma tubes are required due to the reactivity of the metal vapors. Usually helium at a pressure of up to one atmosphere is used as the buffer gas. Typical output powers at 5 kHz pulse repetition frequency are of order 1 W (~0.1% wall plug efficiency) at 430.5 nm from the He Sr^+ laser and 0.7 W at 373.7 nm from the He Ca^+ laser. For the high specific input power densities (10–15 W cm^{-3}) required for efficient lasing, overheating of the laser gas limits aperture scaling beyond 10–15 mm diameter. Slab laser geometries have been used successfully to aperture-scale strontium ion lasers. Scaling of the laser tube length beyond 0.5 m is not practical as achieving high enough excitation voltages for efficient lasing becomes problematic. Gain in strontium and calcium ion lasers is lower than in resonance-metastable metal vapor lasers such as the CVL, hence optimum output coupler reflectivity is approximately 70%. The pulse duration is also longer at around 200–500 ns resulting in moderate beam quality with plane–plane resonators.

For both strontium and calcium ion lasers the two principal transitions share upper laser levels and hence exhibit gain competition such that without wavelength-selective cavities usually only the longer wavelength of the pair is produced. With wavelength-selective cavities up to 60% (Sr^+) and 30% (Ca^+) of the normal power can be obtained at the shorter wavelength.

Many potential applications exist for strontium and calcium ion lasers, given their ultraviolet (UV)/violet wavelengths. Of particular importance is fluorescence spectroscopy in biology and forensics, treatment of neonatal jaundice, stereolithography, micromachining and exposing photoresists for integrated circuit manufacture. Despite these many potential applications, technical difficulties in power scaling mean that both strontium ion and calcium ion lasers have only limited commercial availability.

Continuous-Wave Metal Ion Lasers

In a discharge excited noble gas, there can be a large concentration of noble gas atoms in excited metastable states and noble gas ions in their ground states. These species can transfer their energy to a minority metal species (M) via two key processes: either charge transfer (Duffendack reactions) with the noble gas ions (N^+):

$$\mathrm{M} + \mathrm{N}^+ \rightarrow \mathrm{M}^{+*} + \mathrm{N} + \Delta \mathrm{E}$$

(where M^{+*} is an excited metal ion state)

or Penning ionization with a noble gas atom in an excited metastable state (N^*):

$$\mathrm{M} + \mathrm{N}^* \rightarrow \mathrm{M}^{+*} + \mathrm{N} + e^-$$

In a continuous discharge in a mixture of a noble gas and a metal vapor, steady generation of excited metal ions via energy transfer processes can lead to a steady state population inversion on one or more pairs of levels in the metal ion and hence produce cw lasing.

Several hundred metal ion laser transitions have been observed to lase in a host of different metals. The most important such laser is the helium cadmium laser, which has by far the largest market volume by number of unit sales of all the metal vapor lasers. In a helium cadmium laser, Cd vapor is present at a concentration of about 1–2% in a helium buffer gas which is excited by a dc discharge. Excitation to the upper laser levels of Cd (Figure 8) is primarily via Penning ionization collisions with He^+ 2^3S_1 metastable ions produced in the discharge. Excitation via electron-impact excitation from the ion ground state may also play an important role in establishing a population inversion in Cd^+ lasers. Population inversion can be sustained continuously because the $^2P_{3/2}$ and $^2P_{1/2}$ lower laser levels decay via strong resonance transitions to the $^2S_{1/2}$ Cd^+ ground state, unlike in the self-terminating metal vapor lasers. Two principal wavelengths can be produced, namely 441.6 nm (blue) and 325.0 nm (UV) with cw powers up to 200 mW and 50 mW available respectively from commercial devices.

Typical laser tube construction (Figure 9) consists of a 1–3 mm diameter discharge channel typically 0.5 m long. A pin anode is used at one end of the laser tube together with a large-area cold cylindrical cathode located in a side-arm at the other end of the tube. Cadmium is transported to the main discharge tube from a heated Cd reservoir in a side-arm at 250–300 °C. With this source of atoms at the anode end of the laser a cataphoresis process (in which the positively charged Cd ions are propelled towards the cathode end of the tube by the longitudinal electric field) transports Cd into the discharge channel. Thermal insulation of the discharge tube ensures that it is kept hotter than the Cd reservoir to prevent condensation of Cd from blocking the tube bore. A large-diameter Cd condensation region is provided at the cathode end of the discharge channel. A large-volume side arm is also provided to act as a gas ballast for maintaining correct He pressure. As He is lost through sputtering and diffusion through the Pyrex glass tube walls, it is replenished from a high-pressure He reservoir by heating a permeable glass wall separating the reservoir from the ballast chamber which allows He to diffuse into the ballast chamber. Typical commercial sealed-off Cd lasers have operating lifetimes of several thousand hours. Overall laser construction is not that much more

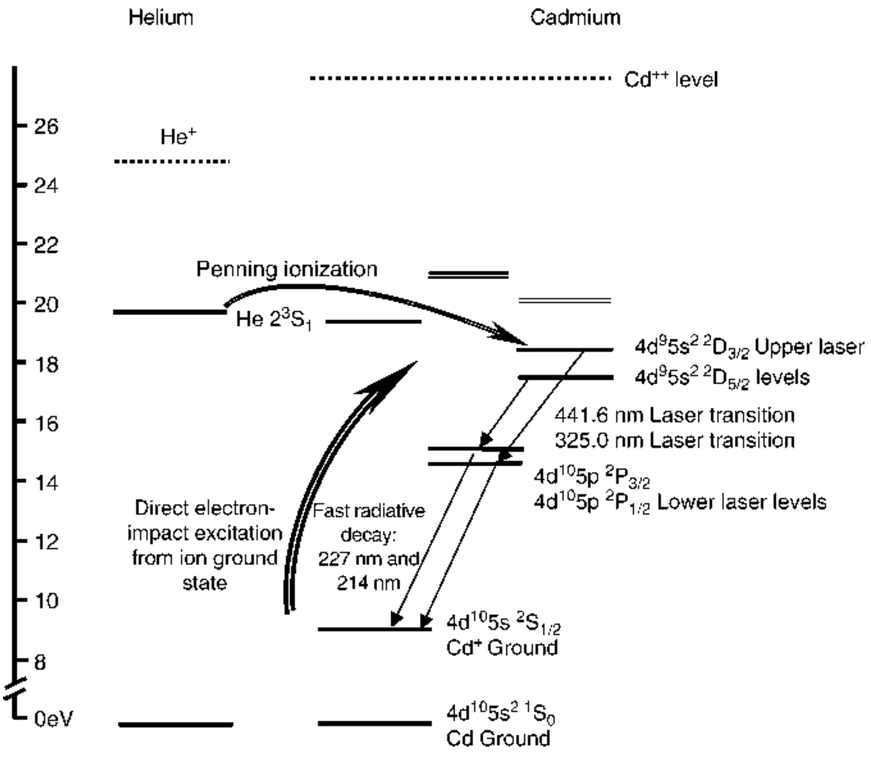
Helium
Cadmium
Cd^{++} level
26
24
22
20
18
16
14
12
10
8
0eV
He^+
Penning ionization
He 2^3S_1
$4d^95s^2\ ^2D_{3/2}$ Upper laser
$4d^95s^2\ ^2D_{5/2}$ levels
441.6 nm Laser transition
325.0 nm Laser transition
$4d^{10}5p\ ^2P_{3/2}$
$4d^{10}5p\ ^2P_{1/2}$ Lower laser levels
Direct electron-impact excitation from ion ground state
Fast radiative decay: 227 nm and 214 nm
$4d^{10}5s\ ^2S_{1/2}$
Cd^+ Ground
$4d^{10}5s^2\ ^1S_0$
Cd Ground

Figure 8 Partial energy level diagram for helium and cadmium giving HeCd laser transitions.

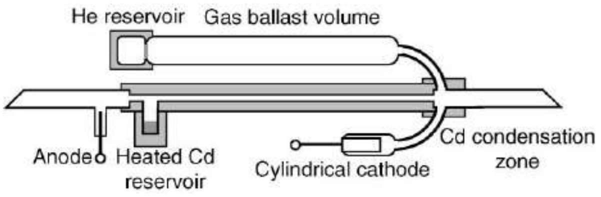

Figure 9 Helium cadmium laser tube construction.

complex than a HeNe laser, hence unit costs are considerably lower than low-power argon ion lasers which also provide output in the blue. Usually Brewster angle windows are provided together with a high Q stable resonator (97–99% reflectivity output coupler) to provide polarized output.

With their blue and UV wavelengths and relatively low cost (compared to low-power argon ion lasers), HeCd lasers have found wide application in science, medicine and industry. Of particular relevance is their application for exposing photoresists where the blue wavelength provides a good match to the peak photosensitivity of photoresist materials. A further key application is in stereolithography where the UV wavelength is used to cure an epoxy resin. By scanning the UV beam in a raster pattern across the surface of a liquid epoxy, a solid three-dimensional object may be built up in successive layers.

List of Units and Nomenclature

Boltzmann constant	[eV K^{-1}]	k
Excited metal ion		M^{+*}
Excited noble gas atom		N^*
Energy difference	[eV]	ΔE
Electron		e^-
Gas flow	[mbar l min^{-1}]	
Mirror curvatures	[m]	R_1, R_2
Metal atom		M
Noble gas atom		N
Noble gas ion		N^+
Pulse duration	[ns]	
Pulse repetition frequency	[kHz]	
Quality factor		Q
Temperature	[K]	T
Tube diameter	[mm]	D
Tube length	[m]	L
Wavelength	[nm], [μm]	
Unstable resonator magnification		M

See also

Nonlinear Sources: Harmonic Generation in Gases.

Further Reading

Little CE (1999) *Metal Vapor Lasers*. Chichester, UK: John Wiley.

Ivanov IG, Latush EL and Sem MF (1996) *Metal Vapor Ion Lasers, Kinetic Processes and Gas Discharges*. Chichester, UK: John Wiley.

Little CE and Sabotinov NV (eds) (1996) *Pulsed Metal Vapor Lasers*. Dordrecht, The Netherlands: Kluwer.

Lyabin NA, Chursin AD, Ugol'nikov SA, Koroleva ME and Kazaryan MA (2001) Development, production, and application of sealed-off copper and gold vapor lasers. *Quantum Electronics* 31: 191–202.

Petrash GG (ed.) (1989) *Metal Vapor and Metal Halide Vapor Lasers*. Commack, NY: Nova Science Publishers.

Silfvast WT (1996) *Laser Fundamentals*. New York: Cambridge University Press.

Noble Gas Ion Lasers

W B Bridges, California Institute of Technology, Pasadena, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

History

The argon ion laser was discovered in early 1964, and is still commercially available in 2004, with about $70 million in annual sales 40 years after this discovery. The discovery was made independently and nearly simultaneously by four different groups; for three of the four, it was an accidental result of studying the excitation mechanisms in the mercury ion laser (historically, the first ion laser), which had been announced only months before. For more on the early years of ion laser research and development, see the articles listed in the Further Reading section at the end of this article.

The discovery was made with pulsed gas discharges, producing several wavelengths in the blue and green portions of the spectrum. Within months, continuous operation was demonstrated, as well as oscillation on many visible wavelengths in ionized krypton and xenon. Within a year, over 100 wavelengths were observed to oscillate in the ions of neon, argon, krypton, and xenon, spanning the spectrum from ultraviolet to infrared; oscillation was also obtained in the ions of other gases, for example, oxygen, nitrogen, and chlorine. A most complete listing of all wavelengths observed as gaseous ion lasers is given in the Laser Handbook cited in the Further Reading section. Despite the variety of materials and wavelengths demonstrated, however, it is the argon and krypton ion lasers that have received the most development and utilization.

Continuous ion lasers utilize high current density gas discharges, typically 50 A or more, and 2–5 mm in diameter. Gas pressures of 0.2 to 0.5 torr result in longitudinal electric fields of a few V/cm of discharge, so that the power dissipated in the discharge is typically 100 to 200 W/cm. Such high-power dissipation required major technology advances before long-lived practical lasers became available. Efficiencies have never been high, ranging from 0.01% to 0.2%. A typical modern ion laser may produce 10 W output at 20 kW input power from 440 V three-phase power lines and require 6–8 gallons/minute of cooling water. Smaller, air-cooled ion lasers, requiring 1 kW of input power from 110 V single-phase mains can produce 10–50 mW output power, albeit at even lower efficiency.

Theory of Operation

The strong blue and green lines of the argon ion laser originate from transitions between the 4p upper levels and 4s lower levels in singly ionized argon, as shown in Figure 1. The 4s levels decay radiatively to the ion ground state. The strongest of these laser lines are listed in Table 1. The notation used for the energy levels is that of the L–S coupling model. The ion ground state electron configuration is $3s^23p^5(^2P^o_{3/2})$. The inner ten electrons have the configuration $1s^22s^22p^6$, but this is usually omitted for brevity. The excited states shown in Figure 1 result from coupling a 4p or 4s electron to a $3s^23p^4(^3P)$ core, with the resulting quantum numbers S (the net spin of the electrons), L (the net angular momentum of the electrons), and J (the angular momentum resulting from coupling S to L) are represented by $^{2S+1}L_J$, where $L = 0, 1, 2, \ldots$ is denoted S, P, D, F, … The superscript 'o' denotes an odd level, while even levels omit a superscript. Note that some weaker transitions involving levels originating from the $3s^23p^4(^1D)$ core configuration also oscillate. Note also that the quantum mechanical selection rules for the L–S coupling model are not rigorously obeyed, although the stronger laser lines satisfy most or all of these rules. The selection rule $|\Delta J| = 1$ or 0, but not $J = 0 \rightarrow J = 0$ is always obeyed. All transitions shown in Figure 1 and Table 1 belong to the second spectrum of argon, denoted Ar II. Lines originating from transitions in the neutral atom make up the first spectrum, Ar I; lines originating from transitions in doubly ionized argon are denoted Ar III, and so forth for even more highly ionized states.

Much work has been done to determine the mechanisms by which the inverted population is formed in the argon ion laser. Reviews of this extensive research

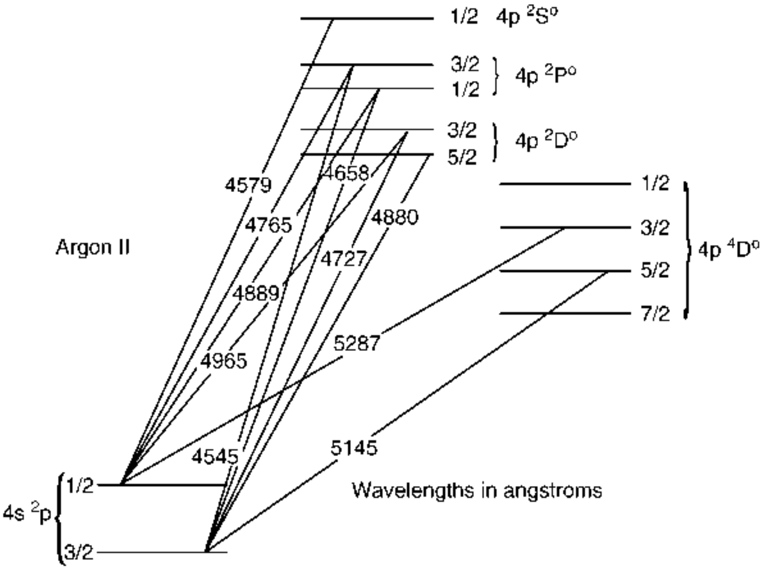

Figure 1 4p and 4s doublet levels in singly ionized argon, showing the strongest blue and green laser transitions.

Table 1 Ar II laser blue-green wavelengths

Wavelength (nanometers)	*Transition[a] (upper level) → (lower level)*	*Relative strength[b] (Watt)*
454.505	$4p\ ^2P^o_{3/2} \rightarrow 4s\ ^2P_{3/2}$	0.8
457.935	$4p\ ^2S^o_{1/2} \rightarrow 4s\ ^2P_{1/2}$	1.5
460.956	$(^1D)4p\ ^2F^o_{7/2} \rightarrow (^1D)4s\ ^2D_{5/2}$	–
465.789	$4p\ ^2P^o_{1/2} \rightarrow 4s\ ^2P_{3/2}$	0.8
472.686	$4p\ ^2D^o_{3/2} \rightarrow 4s\ ^2P_{3/2}$	1.3
476.486	$4p\ ^2P^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	3.0
487.986	$4p\ ^2D^o_{5/2} \rightarrow 4s\ ^2P_{3/2}$	8.0
488.903	$4p\ ^2P^o_{1/2} \rightarrow 4s\ ^2P_{1/2}$	[c]
496.507	$4p\ ^2D^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	3.0
501.716	$(^1D)4p\ ^2D^o_{5/2} \rightarrow 3d\ ^2D_{3/2}$	1.8
514.179	$(^1D)4p\ ^2F^o_{7/2} \rightarrow 3d\ ^2D_{5/2}$	[c]
514.532	$4p\ ^4D^o_{5/2} \rightarrow 4s\ ^2P_{3/2}$	10
528.690	$4p\ ^4D^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	1.8

[a] All levels are denoted in $L-S$ coupling with the (^3P) core unless otherwise indicated. Odd parity is denoted by the superscript 'o'.

[b] Relative strengths are given as power output from a commercial Spectra-Physics model 2080-25S argon ion laser.

[c] These lines may oscillate simultaneously with the nearby strong line, but are not resolved easily in the output beam.

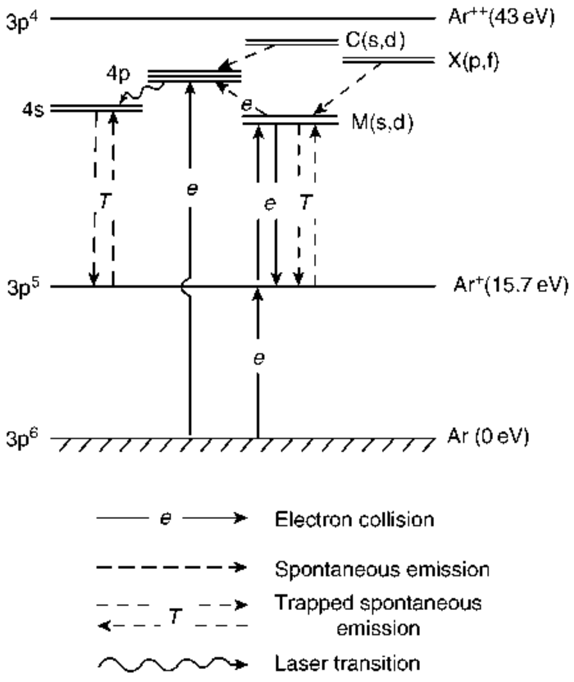

Figure 2 Schematic representation of energy levels in neutral and singly ionized argon, indicating alternative pathways for excitation and de-excitation of the argon ion laser levels.

are found in the Further Reading section. While a completely quantitative picture of argon ion laser operation is lacking to this day, the essential processes are known. Briefly, the 4p upper levels are populated by three pathways, as illustrated in Figure 2:

(i) by electron collision with the $3p^6$ neutral ground state atoms. This 'sudden perturbation process' requires at least 37 eV electrons, and also singles out the $4p\ ^2P^o_{3/2}$ upper level, which implies that only the 476 and 455 nm lines would oscillate. This is the behavior seen in pulsed discharges at very low pressure and very high axial electric field. This pathway probably contributes little to the 4p population under ordinary continuous wave (cw) operating conditions, however.

(ii) by electron collision from the lowest-lying s and d states in the ion, denoted M(s,d) in Figure 1. This only requires 3–4 eV electrons. These states have parity-allowed transitions to the ion ground state, but are made effectively metastable by radiation trapping (that is, there is a high probability that an emitted photon is re-absorbed by another ground state ion before it escapes the discharge region), or by requiring $|\Delta J|$ to be 2 to make the transition (forbidden by quantum selection rules). Thus, these levels are both created and destroyed primarily by electron collision, causing the population of the M(s,d) states to follow the population of the singly ionized ground state, which, in turn, is approximately proportional to the discharge current. Since a second electron collision is required to get from the M(s,d) states to the 4p states, a quadratic variation with current for the laser output power would be expected, and that is what is observed over some reasonable range of currents between threshold and saturation. Note that radiative decay from higher-lying opposite-parity p and f states, denoted X(p,f), can also contribute to the population of M(s,d), but the linear variation in population of the M(s,d) with discharge current is assured by electron collision creation and destruction.

(iii) by radiative decay from higher-lying opposite-parity s and d states, denoted C(s,d). These states are populated by electron collision with the $3p^5$ ion ground states, and thus have populations that also vary quadratically with discharge current. The contribution of this cascade process has been measured to be 20% to 50% of the 4p upper laser level population.

Note that it is not possible to distinguish between processes (ii) and (iii) by the variation of output power with discharge current; both give the observed quadratic dependence.

The radiative lifetimes of the 4s ^2P levels are sufficiently short to depopulate the lower laser levels by radiative decay. However, radiation trapping greatly lengthens this decay time, and a bottleneck can occur. In pulsed ion lasers, this is exhibited by the laser pulse terminating before the excitation current pulse ends, which would seem to preclude continuous operation. However, the intense discharge used in continuous operation heats the ions to the order of 2300 K, thus greatly Doppler broadening the absorption linewidth and reducing the magnitude of the absorption. Additionally, the ions are attracted to the discharge tube walls, so that the absorption spectrum is further broadened by the Doppler shift due to their wall-directed velocities. The plasma wall sheath gives about 20 V drop in potential from the discharge axis to the tube wall, so most ions hit the wall with 20 eV of energy, or about ten times their thermal velocity. A typical cw argon ion laser operates at ten times the gas pressure that is optimum for a pulsed laser, and thus the radiation trapping of the $4s \rightarrow 3p^5$ transitions is so severe that it may take several milliseconds after discharge initiation for the laser oscillation to begin.

As the discharge current is increased, the intensities of the blue and green lines of Ar II eventually saturate, and then decrease with further current. At these high currents, there is a buildup in the population of doubly ionized atoms, and some lines of Ar III can be made to oscillate with the appropriate ultraviolet mirrors. Table 2 lists the strongest of these lines, those that are available in the largest commercial lasers. Again, there is no quantitative model for the performance in terms of the discharge parameters, but the upper levels are assumed to be populated by processes analogous to those of the Ar II laser. At still higher currents, lines in Ar IV can be made to oscillate as well.

Much less research has been done on neon, krypton, and xenon ion lasers, but it is a good assumption that the population and depopulation processes are the same in these lasers. Table 3 lists both the Kr II and Kr III lines that are available from the largest commercial ion lasers. Oscillation on lines in still-higher ionization states in both krypton and xenon have been observed.

Operating Characteristics

A typical variation of output power with discharge current for an argon ion laser is shown in Figure 3. This particular laser had a 4 mm diameter discharge in a water-cooled silica tube, 71 cm in length, with a 1 kG axial magnetic field. The parameter is the argon pressure in the tube before the discharge was struck. Note that no one curve is exactly quadratic, but that the envelope of the curves at different filling pressures is approximately quadratic. At such high discharge current densities (50 A in the 4 mm tube is approximately 400 A/cm^2) there is substantial pumping of gas out of the small-bore discharge region. Indeed, a return path for this pumped gas must be provided from anode to cathode ends of the discharge to keep the discharge from self-extinguishing. The axial electric field in this discharge was 3–5 V/cm, so the input power was of the order of 10 to 20 kW, yielding an efficiency of less than 0.1%, an unfortunate characteristic of all ion lasers.

Table 2 Ultraviolet argon ion laser wavelengths

Wavelength (nanometers)	*Spectrum*	*Transition*[a] *(upper level) → (lower level)*	*Relative strength*[b] *(Watt)*
275.392	III	$(^2D^o)4p\ ^1D_2 \rightarrow (^2D^o)4s\ ^1D^o_2$	0.3
275.6	?	?	0.02
300.264	III	$(^2P^o)4p\ ^1P_1 \rightarrow (^2P^o)3d\ ^1D^o_2$	0.5
302.405	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)4s\ ^3P^o_2$	0.5
305.484	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)4s\ ^3P^o_1$	0.2
333.613	III	$(^2D^o)4p\ ^3F_4 \rightarrow (^2D^o)4s\ ^3D^o_3$	0.4
334.472	III	$(^2D^o)4p\ ^3F_3 \rightarrow (^2D^o)4s\ ^3D^o_2$	0.8
335.849	III	$(^2D^o)4p\ ^3F_2 \rightarrow (^2D^o)4s\ ^3D^o_1$	0.8
350.358	III	$(^2D^o)4p\ ^3D_2 \rightarrow (^2D^o)4s\ ^3D^o_2$	0.05
350.933	III	$(^4S^o)4p\ ^3P_0 \rightarrow (^4S^o)4s\ ^3S^o_1$	0.05
351.112	III	$(^4S^o)4p\ ^3P_2 \rightarrow (^4S^o)4s\ ^3S^o_1$	2.0
351.418	III	$(^4S^o)4p\ ^3P_1 \rightarrow (^4S^o)4s\ ^3S^o_1$	0.7
363.789	III	$(^2D^o)4p\ ^1F_3 \rightarrow (^2D^o)4s\ ^1D^o_2$	2.5
379.532	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)3d\ ^3P^o_2$	0.4
385.829	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)3d\ ^3P^o_1$	0.15
390.784	III	$(^2P^o)4p\ ^3D_1 \rightarrow (^2P^o)3d\ ^3P^o_0$	0.02
408.904	IV?	?	0.04
414.671	III	$(^2D^o)4p\ ^3P_2 \rightarrow (^2P^o)4s\ ^3P^o_2$	0.02
418.298	III	$(^2D^o)4p\ ^1P_1 \rightarrow (^2D^o)4s\ ^1D^o_2$	0.08

[a] All levels are denoted in $L-S$ coupling with the core shown in ( ). Odd parity is denoted by the superscript ‘o’.

[b] Relative strengths are given as power output from a commercial Spectra-Physics model 2085-25S argon ion laser.

Table 3 Krypton ion laser wavelengths

Wavelength (nanometers)	*Spectrum*	*Transition*[a] *(upper level) → (lower level)*	*Relative strength*[b] *(Watt)*
337.496	III	$(^2P^o)5p\ ^3D_3 \rightarrow (^2P^o)5s\ ^3P^o_2$	–
350.742	III	$(^4S^o)5p\ ^3P_2 \rightarrow (^4S^o)5s\ ^3S^o_1$	1.5
356.432	III	$(^4S^o)5p\ ^3P_1 \rightarrow (^4S^o)5s\ ^3S^o_1$	0.5
406.737	III	$(^2D^o)5p\ ^1F_3 \rightarrow (^2D^o)5s\ ^1D^o_2$	0.9
413.133	III	$(^4S^o)5p\ ^5P_2 \rightarrow (^4S^o)5s\ ^3S^o_1$	1.8
415.444	III	$(^2D^o)5p\ ^3F_3 \rightarrow (^2D^o)5s\ ^1D^o_2$	0.3
422.658	III	$(^2D^o)5p\ ^3F_2 \rightarrow (^2D^o)4d\ ^3D^o_1$	–
468.041	II	$(^3P)5p\ ^2S^o_{1/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.5
476.243	II	$(^3P)5p\ ^2D^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.4
482.518	II	$(^3P)5p\ ^4S^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.4
520.832	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)5s\ ^4P_{3/2}$	–
530.865	II	$(^3P)5p\ ^4P^o_{5/2} \rightarrow (^3P)5s\ ^4P_{3/2}$	1.5
568.188	II	$(^3P)5p\ ^4D^o_{5/2} \rightarrow (^3P)5s\ ^2P_{3/2}$	0.6
631.024	III	$(^2D^o)5p\ ^3P_2 \rightarrow (^2P^o)4d\ ^3D_1$	0.2
647.088	II	$(^3P)5p\ ^4P^o_{5/2} \rightarrow (^3P)5s\ ^2P_{3/2}$	3.0
676.442	II	$(^3P)5p\ ^4P^o_{1/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.9
752.546	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	1.2
793.141	II	$(^1D)5p\ ^4F^o_{7/2} \rightarrow (^3P)4d\ ^2F_{5/2}$	0.3
799.322	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)4d\ ^4D_{1/2}$	

[a] All levels are denoted in $L-S$ coupling with the core shown in ( ). Odd parity is denoted by the superscript 'o'.

[b] Relative strengths are given as power output from a commercial Spectra-Physics model 2080RS ion laser.

Technology

With such high input powers required in a small volume to produce several watts output, ion laser performance has improved from 1964 to the present only as new discharge technologies were introduced. The earliest laboratory argon ion lasers used thin-walled (≈ 1 mm wall thickness) fused silica discharge tubes, cooled by flowing water over the outside wall of the tube. The maximum input power per unit length of discharge was limited by thermal stresses in the silica walls caused by the temperature differential from inside to outside. Typically, ring-shaped cracks would cause the tube to fail catastrophically. Attempts to make metal–ceramic structures with alumina (Al_2O_3) discharge tubes to contain the plasma were made early on (1965) but were not successful. Such tubes invariably failed from fracture by thermal shock as the discharge was turned on. Later, successful metal–ceramic tubes were made with beryllia (BeO), which has a much higher thermal conductivity than silica or alumina and is much more resistant to thermal shock. Today, all of the lower power (less than 100 mW output) are made with BeO discharge tubes. Some ion lasers in the 0.5 to 1 W range are also made with water-cooled BeO discharge tubes.

A typical low-power air cooled argon ion laser is shown in Figure 4. The large metal can (2) on the right end of the tube contains an impregnated-oxide hot cathode, heated directly by current through ceramic feed-through insulators (3). The small (≈ 1 mm diameter) discharge bore runs down the center of the BeO ceramic rod (1), and several smaller diameter gas return path holes run off-axis parallel to the discharge bore to provide the needed gas equalization between cathode can and anode region. A copper honeycomb cooler is brazed to the cathode can, two more to the outer wall of the BeO cylinder, and one to the anode (4) at the left end of the tube. The laser mirrors (5) are glass-fritted to the ends of the tube, forming a good vacuum seal. Note that in this small laser, the active discharge bore length is less than half of the overall length.

The very early argon ion lasers were made with simple smooth dielectric tubes, allowing the continuous variation in voltage along the length of the tube required by the discharge longitudinal electric field. However, this variation can be step-wise at the discharge walls, and still be more or less smooth along the axis. Thus, the idea arose of using metal tube segments, insulated one from another, to form the discharge tube walls. The first version of this idea (1965) used short (≈ 1 cm) metal cylinders supported by metal disks and stacked inside a large diameter silica envelope, with each metal cylinder electrically isolated from the others. The tubes and disks were made of molybdenum, and were allowed to heat to incandescence, thus radiating several kilowatts of heat through the silica vacuum envelope to a water-cooled collector outside. While this eliminated the problems of thermal shock and poor thermal conductivity inherent in dielectric wall discharges, it made another

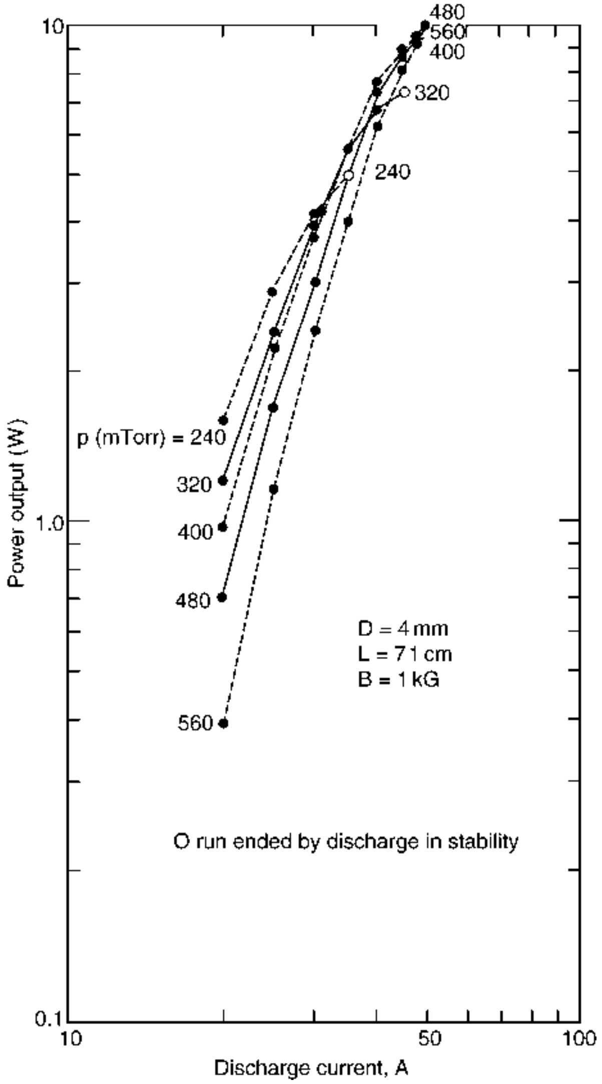

Figure 3 Laser output power (summed over all the blue and green laser lines) versus discharge current for a laser discharge 4 mm in diameter and 71 cm long, with a 1 kilogauss longitudinal magnetic field. The parameter shown is the argon fill pressure in mTorr prior to striking the discharge. The envelope of the curves exhibits the quadratic variation of output power with discharge current.

problem painfully evident. The intense ion bombardment of the metal tube walls sputtered the wall material, eventually eroding the shape of the metal cylinders and depositing metal films on the insulating wall material, thus shorting one segment to another.

Many different combinations of materials and configurations were investigated in the 1960s and 1970s to find a structure that would offer good laser performance and long life. It was found that simple thin metal disks with a central hole would effectively confine the discharge to a small diameter (on the order of the hole size) if a longitudinal d-c magnetic field of the order of 1 kiloGauss were used. The spacing between disks can be as large as 2–4 discharge diameters. Of the metals, tungsten has the lowest sputtering yield for argon ions in the 20 eV range (which is approximately the energy they gain in falling to the wall across the discharge sheath potential difference). An even lower sputtering yield is exhibited by carbon, and graphite cylinders contained within a larger diameter silica or alumina tube were popular for a while for ion laser discharges. Unfortunately, graphite has a tendency to flake or powder, so such laser discharge tubes became contaminated with 'dust' which could eventually find its way to the optical windows of the tube. Beryllia and silica also sputter under argon ion bombardment, but with still lower yields than metals or carbon; however, their smaller thermal conductivities limit them to lower-power applications.

The material/configuration combination that has evolved for higher-power argon ion lasers today is to use a stack of thin tungsten disks with 2–3 mm diameter holes for the discharge. These disks, typically 1 cm in diameter, are brazed coaxially to a larger copper annulus (actually, a drawn cup with a 1 cm hole on its axis). A stack of these copper/tungsten structures is, in turn, brazed to the inside wall of a large diameter alumina vacuum envelope. The tungsten disks are exposed to and confine the discharge, while the copper cups conduct heat radially outward to the alumina tube wall, which, in turn, is cooled by fluid flow over its exterior. Thus, the discharge is in contact only with a low-sputtering material (tungsten) while the heat is removed by a high thermal conductivity material (copper). Details differ among manufacturers, but this 'cool disk' technology seems to have won out in the end. A photo of a half-sectioned disk/cup stacked assembly from a Coherent Innova™ ion laser is shown in Figure 5. The coiled impregnated-tungsten cathode is also shown.

Sputtering of the discharge tube walls is not uniform along the length of the gas discharge. The small diameter region where the laser gain occurs is always joined to larger diameter regions containing cathode and anode electrodes (as shown in Figure 5, for example). A plasma double sheath (that is, a localized increase in potential) forms across the discharge in the transition region between large and small diameter regions (the discharge 'throats,' which may be abrupt or tapered). This sheath is required to satisfy the boundary conditions between the plasmas of different temperatures in the different regions. Such a double sheath imparts additional energy to the ions as they cross the sheath, perhaps

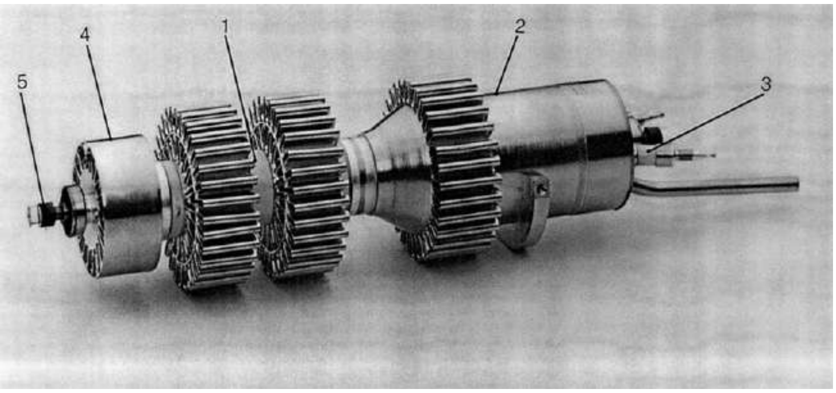

Figure 4 A typical commercial low-power, air-cooled argon ion laser. The cathode can (2) is at the right and the beryllia bore (1) and anode (4) is at the left. The cathode current is supplied through ceramic vacuum feed-throughs (3). Mirrors (5) are glass fritted onto the ends of vacuum envelope. (Photo courtesy of JDS Uniphase Corp.)

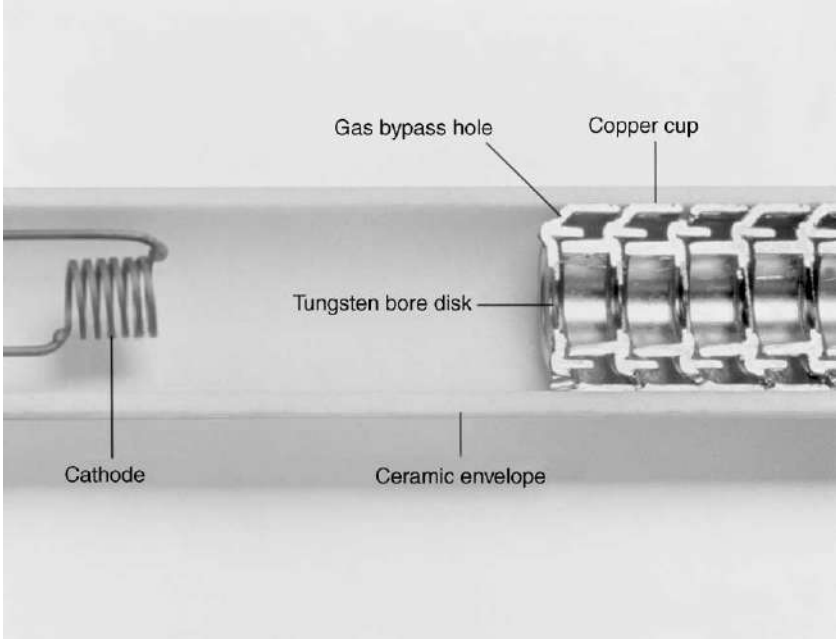

Figure 5 Discharge bore structure of a typical commercial high-power argon ion laser, the Coherent Innova™. Copper cups are brazed to the inner wall of a ceramic envelope, which is cooled by liquid flow over its outer surface. A thin tungsten disk with a small hole defining the discharge path is brazed over a larger hole in the bottom of the copper cup. Additional small holes in the copper cup near the ceramic envelope provide a gas return path from cathode to anode. Also shown is the hot oxide-impregnated tungsten cathode. (Photo courtesy of Coherent, Inc.)

an additional 20 eV. When these ions eventually hit the discharge walls near the location of the sheath, they have 40 eV of energy rather than the 20 eV from the normal wall sheath elsewhere in the small diameter discharge. Sputtering yield (number of sputtered wall atoms per incident ion) is exponentially dependent on ion energy in this low ion energy region, so the damage done to the wall in the vicinity of the 'throat' where the double sheath forms may be more than ten times that elsewhere in the small diameter bore region. This was the downfall of high-power operation of dielectric discharge bores, even BeO; while sputtering was acceptable elsewhere in the discharge, the amount of material removed in a small region near the discharge throat would cause catastrophic bore failure at that point. This localized increase in sputtering in the discharge throat is common to all ion lasers, including modern cooled tungsten disk tubes. Eventually, disks near the throat are eroded to larger diameters, and more material is deposited on the walls nearby. Attempts to minimize localized sputtering by tapering the throat walls or the confining magnetic field have proven unsuccessful; a localized double sheath always forms somewhere in the throat. Because of the asymmetry caused by ion flow, the double sheath is larger in amplitude in the cathode throat than the anode throat, so that the localized wall damage is larger at the cathode end of the discharge than at the anode end.

Another undesirable feature of sputtering is that the sputtered material 'buries' some argon atoms when it is deposited on a wall. This is the basis of the well-known Vac-Ion® vacuum pump. Thus, the operating pressure in a sealed laser discharge tube will drop during the course of operation. In low-power ion lasers, this problem is usually solved by making the gas reservoir volume large enough to satisfy the desired operating life (for example, the large cathode can in Figure 4). In high-power ion lasers, the gas loss would result in unacceptable operating life even with a large reservoir at the fill pressure. Thus, most high-power ion lasers have a gas pressure measurement system and dual-valve arrangement connected to a small high-pressure reservoir to 'burp' gas into the active discharge periodically to keep the pressure within the operating range. Unfortunately, if a well-used ion laser is left inoperative for months, some of the 'buried' argon tends to leak back into the tube, with the gas pressure becoming higher than optimum. The pressure will gradually decrease to its optimum value with further operation of the discharge (in perhaps tens of hours). In the extreme case, the gas pressure can rise far enough so that the gas discharge will not strike, even at the maximum power supply voltage. Such a situation requires an external vacuum pump to remove the excess gas, usually a factory repair.

In addition to simple dc discharges, various other techniques have been used to excite ion lasers, primarily in a search for higher power, improved efficiency and longer operating life. Articles in the Further Reading section give references to these attempts. Radio-frequency excitation at 41 MHz was used in an inductively coupled discharge, with the laser bore and its gas return path forming a rectangular single turn of an air-core transformer. A commercial product using this technique was sold for a few years in the late 1960s. Since this was an 'electrode-less' discharge, ion laser lines in reactive gases such as chlorine could be made to oscillate, as well as the noble gases, without 'poisoning' the hot cathode used in conventional dc discharge lasers. A similar electrode-less discharge was demonstrated as a quasi-cw laser by using iron transformer cores and exciting the discharge with a 2.5 kHz square wave. Various microwave excitation configurations at 2.45 GHz and 9 GHz also resulted in ion laser oscillation. Techniques common to plasma fusion research were also studied. Argon ion laser oscillation was produced in Z-pinch and Θ-pinch discharges and also by high-energy (10–45 keV) electron beams. However, none of these latter techniques resulted in a commercial product, and all had efficiencies worse than the simple dc discharge lasers.

It is interesting that the highest-power output demonstrations were made in less than ten years after the discovery. Before 1970, 100 W output on the argon blue-green lines was demonstrated with dc discharges two meters in length. In 1970, a group in the Soviet Union reported 500 W blue-green output from a two-meter discharge with 250 kW of dc power input, an efficiency of 0.2%. Today, the highest output power argon ion laser offered for sale is 50 W output.

Manufacturers

More than 40 companies have manufactured ion lasers for sale over the past four decades. This field has now (2004), narrowed to the following:

Coherent, Inc.	http://www.coherentinc.com
INVERSion Ltd.	http://inversion.iae.nsk.su
JDS Uniphase	http://www.jdsu.com
Laser Physics, Inc.	http://www.laserphysics.com
Laser Technologies GmbH	http://www.lg-lasertechnologies.com
LASOS Lasertechnik GmbH	http://www.LASOS.com
Lexel Laser, Inc.	http://www.lexellaser.com
Melles Griot	http://lasers.mellesgriot.com
Spectra-Physics, Inc.	http://www.spectraphysics.com

Further Reading

Bridges WB (1979) Atomic and ionic gas lasers. In: Marton L and Tang CL (eds) *Methods of Experimental Physics, Vol 15: Quantum Electronics, Part A*, pp. 31–166. New York: Academic Press.

Bridges WB (1982) Ionized gas lasers. In: Weber MJ (ed.) *Handbook of Laser Science and Technology, Vol. II, Gas Lasers*, pp. 171–269. Boca Raton, FL: CRC Press.

Bridges WB (2000) Ion lasers – the early years. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 885–898.

Davis CC and King TA (1975) Gaseous ion lasers. In: Goodwin DW (ed.) *Advances in Quantum Electronics*, vol. 3, pp. 169–469. New York: Academic Press.

Dunn MH and Ross JN (1976) The argon ion laser. In: Sanders JH and Stenholm S (eds) *Progress in Quantum Electronics*, vol. 4, pp. 233–269. New York: Pergamon.

Weber MJ (2000) *Handbook of Laser Wavelengths*. Boca Raton, FL: CRC Press.

Optical Fiber Lasers

G E Town, Macquarie University, NSW, Australia

N N Akhmediev, Australian National University, ACT, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Optical fiber lasers were first demonstrated in the 1960s, and since then have developed to become versatile optical sources with many desirable properties. Aided by developments in associated technologies, such as fiber design and fabrication methods, semiconductor pump diode technology, and fiber-coupled and in-fiber components such as Bragg grating filters, optical fiber lasers now compete with other laser technologies in many applications, from telecommunications to materials processing.

An optical fiber laser is fundamentally an optical oscillator, which converts input pump power to coherent optical output power at one or more well-defined wavelengths. Optical oscillators require two basic elements: optical gain, and optical feedback. For sustained oscillation to occur the round-trip gain in the laser cavity must be unity, and the round-trip phase a multiple of 2π. In optical fiber lasers the optical gain is provided by an optical fiber amplifier using one or a combination of fundamental physical processes, e.g., stimulated emission, stimulated scattering, or nonlinear parametric processes. Early fiber amplifiers used stimulated Raman scattering to produce optical gain; however, rare-earth-doped fiber amplifiers, in which gain is provided by stimulated emission, are now more common. The reader is referred to **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems, for more details on doped fiber amplifiers, and to **Scattering:** Stimulated Scattering, **Nonlinear Optics, Applications:** Raman Lasers, **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed), and **Scattering:** Raman Scattering, for more details on stimulated scattering and parametric gain processes. Optical feedback may be provided in two fundamental ways, e.g., by using a closed ring of fiber, or from reflections from nonuniformities and discontinuities in the waveguide such as in Bragg grating filters or at fiber ends. The reader is referred to **Fiber Gratings** for more details on fiber Bragg gratings.

The main advantages of optical fiber lasers are derived from the confinement of the pump and signal in a small optical waveguide. In contrast with bulk lasers, the pump intensity in the fiber waveguide is largely independent of the laser length, resulting in large amplifier gain and low laser threshold, even for gain media with small absorption and emission cross-sections. The large gain enables lossy elements such as optical fiber-coupled devices and bulk optical elements to be incorporated into the laser cavity, providing additional control over the optical signal being generated. To obtain a large gain optical fiber amplifiers must usually be relatively long (i.e., from several centimeters to many meters), and so the linear and nonlinear properties of the fiber waveguide can have a significant influence on the optical signal being generated, resulting in some interesting and useful phenomena, e.g., as in soliton fiber lasers. The reader is referred to **Fiber and Guided Wave Optics:** Optical Fiber Cables, for fiber-based components, to **Fiber and Guided Wave Optics**: Dispersion, Light Propagation, and Nonlinear Effects (Basics), for a review of the linear and nonlinear properties of optical fibers, and to **Solitons:** Soliton Communication Systems, and Temporal Solitons, for the theory and applications of temporal solitons.

Whilst optical fibers may be fabricated in a range of materials, including polymer and crystalline materials, most fibers and fiber lasers are currently made of silica containing rare-earth ion dopants – see **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber, for details. Lasers in silica optical fiber have much in common with other glass-host lasers, including a wide range of potential pump and lasing wavelengths, and broadband absorption and gain due to homogeneous and inhomogeneous broadening of the lasing energy levels in the amorphous glass host. For example, erbium-doped amplifiers in alumino-silicate glass fibers at room temperature have lasing transitions with homogeneous and

Table 1 Commonly used pump and signal transitions in rare-earth doped fiber lasers

Rare earth dopant	*Pump wavelengths [nm]*	*Signal wavelengths [nm]*	*Excited state lifetime [ms]*	*Energy levels*
Praseodymium	480, 585	885, 1080	0.1	4
Neodymium	590, 800	920 (900–950)	0.5	3
		1060 (1055–1140)		4
Samarium	488	651	1.5	4
Holmium	455, 650, 1150	2040	0.6	3
Erbium	800, 975, 1480	1550 (1530–1610)	10	3
Thulium	790, 1210 (1060–1320)	1480 (1460–1520)		
		1850 (1650–2050)	0.3	3
Ytterbium	920 (840–980)	975	0.8	3
		1040 (1010–1160)		4

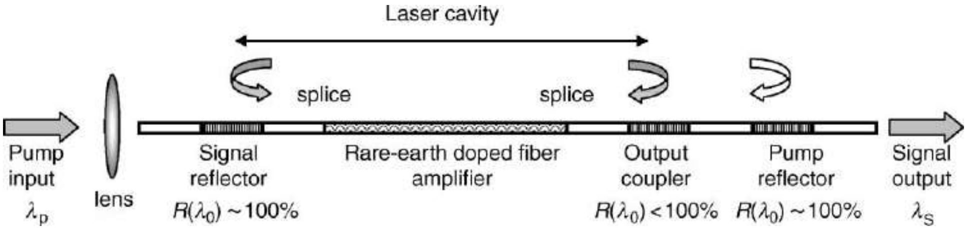
Laser cavity
splice
splice
Pump input λ_p
lens
Signal reflector $R(\lambda_0) \sim 100\%$
Rare-earth doped fiber amplifier
Output coupler $R(\lambda_0) < 100\%$
Pump reflector $R(\lambda_0) \sim 100\%$
Signal output λ_S

Figure 1 Example of a simple all-fiber optical oscillator. The rare-earth doped fiber provides optical gain, and the Bragg gratings provide narrowband optical feedback.

inhomogeneous linewidths of several nanometers, and can be optimized to provide around 30 dB gain over an 80 nm bandwidth centered around 1570 nm. The large gain-bandwidth is useful for achieving either wide tunability of the lasing wavelength and/or ultrashort pulse generation. Table 1 summarizes the most commonly used pump and signal transitions for a range of rare-earth dopants in silicate optical fibers.

Other desirable properties of silica fiber lasers include high efficiency, excellent thermal dissipation, substantial energy storage, high power capability, and compatibility with a wide range of optical fiber devices and systems. The characteristics of optical fiber lasers can be optimized for some very different applications, ranging from narrow linewidth and low-noise lasers to broadband pulsed lasers with high energy and/or high peak power. In the following sections the fundamentals of laser theory as applied to fiber lasers will be reviewed, highlighting the reasons underlying such versatility.

Fiber Laser Fundamentals

An example of a simple all-fiber laser is shown schematically in Figure 1. The laser contains the basic elements of optical gain (e.g., a rare-earth doped fiber amplifier) and optical feedback (e.g., a pair of Bragg grating filters). In the laser shown, the two Bragg gratings provide optical feedback only over a narrow range of wavelengths, which further restricts the range of possible lasing wavelengths. One grating must have a reflectivity less than unity to allow a proportion of the lasing signal to be coupled out of the laser cavity. At threshold, the gain in the amplifier exactly compensates for the loss in the output coupler.

The optical pump, typically from a semiconductor laser diode, is absorbed as is propagates along the fiber amplifier, e.g., by rare-earth ion dopants, which are raised to an excited state. The absorbed energy is stored and eventually emitted at a longer wavelength by either spontaneous or stimulated emission. Spontaneous emission produces unwanted noise and is associated with a finite excited state lifetime, whilst stimulated emission produces optical gain provided the gain medium is ‘inverted’, i.e., with more active ions in the upper lasing energy level than the lower lasing energy level. In practice some energy is also lost to nonradiative (thermal) emissions, e.g., in the transfer of energy between the pump and upper lasing energy levels. In an ideal gain medium the nonradiative transfer rate between the pump and upper lasing energy levels is much larger than the spontaneous or stimulated emission rates, so the inversion may usually be assumed to be independent of nonradiative transfer rates. Energy level diagrams for three- and four-level atoms are shown in Figure 2.

The local optical gain coefficient (i.e. power gain per unit length, in nepers[2]) in a fiber amplifier is proportional to the local inversion, i.e., $g(z) \sim \sigma_S \Delta N(z)$, where σ_S is the emission cross-section at the

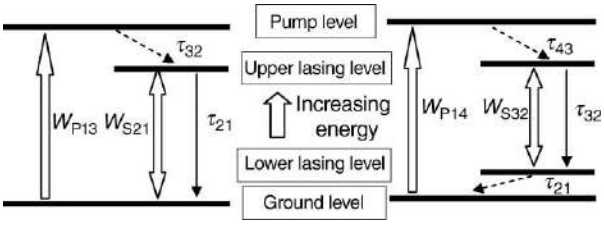

Figure 2 Energy level diagrams for three- and four-level laser transitions. *W* represents the probability of an optical transition per unit time associated with stimulated absorption or emission, and $1/\tau$ is the rate of spontaneous emission. Dotted lines represent nonradiative transitions (e.g., between the pump and upper lasing level).

signal wavelength, and $\Delta N(z)$ the difference in population density between the upper and lower lasing energy levels ($\Delta N > 0$ if the gain medium is inverted). The local inversion can be calculated by solving the rate equations including all stimulated and spontaneous transitions between the laser energy levels. The end-to-end amplifier power gain, $G = P_S(L_A)/P_S(0)$, in length L_A of amplifying fiber is then

$$G = \exp\left[\int_0^{L_A} g(z)\,dz\right] \qquad [1]$$

which in decibels is

$$G = 4.34\int_0^{L_A} g(z)\,dz$$

A large gain coefficient is desirable for amplification with minimal added noise (i.e., low noise figure), and a large gain per unit pump power is desirable for low threshold lasing. For a given gain medium the only factor usually accessible to control the inversion and hence gain is the pump rate, i.e., the rate at which pump energy is absorbed by the gain medium, which for an unbleached amplifier is proportional to pump intensity. For example, in an ideal four-level unsaturated laser medium with uniform pump intensity, the small signal gain is approximately

$$\int_0^{L_A} g(z)\,dz \approx \sigma_S \frac{P_{\mathrm{Pabs}}\tau}{h\nu_P A} \qquad [2]$$

in which τ is the lifetime of the upper lasing level, P_{Pabs} is the pump power absorbed in the fiber amplifier with cross-sectional area A, h is Planck's constant, and ν_P is the pump frequency.

For a three-level laser under the same conditions, the small signal gain is

$$\int_0^{L_A} g(z)\,dz \approx -\sigma_S N L_A + \sigma_S \frac{2P_{\mathrm{Pabs}}\tau}{h\nu_P A} \qquad [3]$$

in which the first term corresponds to ground state absorption of the signal, in which N is the dopant ion density, and the factor of 2 in the second term is due to the fact that in an ideal three-level system every absorbed pump photon increases the inversion, ΔN, by two. Equation [3] can be approximated by multiplying [2] by the factor $(P_{\mathrm{Pabs}} - P_{\mathrm{Psat}})/(P_{\mathrm{Pabs}} + P_{\mathrm{Psat}})$, where $P_{\mathrm{Psat}} = h\nu_P A/(\sigma_P \tau)$ is the pump power which must be absorbed to reduce the pump absorption coefficient to half its unpumped value (or the signal absorption coefficient to zero), and σ_P is the pump absorption cross-section. Unlike four-level lasers in which gain is available as soon as pump power is applied, three-level lasers require half the active ions to be excited before the inversion becomes positive and gain is produced.

The main advantages of optical fiber amplifiers are now clear; in an optical fiber waveguide the light is confined in a very small cross-sectional area, typically less than ten microns diameter. Consequently even a small coupled pump power can have a large intensity and produce a large gain coefficient, or low laser threshold. Even if the pump absorbed per unit length in the fiber amplifier is small (e.g., due to low dopant concentration or low absorption coefficient), this can often be compensated by using a long length of fiber with small intrinsic loss. Furthermore, rare-earth dopants have relatively long metastable state lifetimes, τ, which further assists in producing a large inversion throughout the fiber, and hence large overall gain or low lasing threshold, even at low pump powers. For example, a typical four-level neodymium doped fiber amplifier providing gain at $\lambda_S = 1.06$ μm, pumped at $\lambda_P = 0.8$ μm, with core diameter 7 μm, $\tau = 0.5$ ms, $\sigma_S = 1.4 \times 10^{-20}$ cm^2, and $\sigma_P = 23 \times 10^{-21}$ cm^2, could theoretically provide a small signal gain of 0.3 dB for every milliwatt of absorbed pump power.

Whilst the above analysis highlights the main advantages of fiber amplifiers, it is approximate in a number of respects. For example, eqns [2] and [3] assume both pump and signal have a uniform intensity distribution across the gain medium, and completely overlap. In single mode optical fibers the intensity distribution across the fiber core is not uniform but approximately Gaussian, with a width and effective area that depends on the numerical aperture of the waveguide and the wavelength. Consequently the overlap of the pump and signal beams with the doped fiber core and with each other is less than 100%, and the pump and signal intensities are higher in the center of the fiber than at the core-cladding boundary. The nonuniform distribution of pump and signal beams slightly modifies both the gain efficiency and gain saturation behavior (i.e., the dependence of gain on signal power) in fiber amplifiers. More accurate analysis would also take into account factors such as the rare-earth dopant distribution (which may not be uniform), the variation of pump and signal intensities along the fiber, spectral variation in the absorption and emission cross-sections, spatial and spectral hole-burning in the gain medium, degeneracies in the energy levels, excited state absorption and other loss mechanisms, temperature, and spontaneous emission noise.

Continuous Wave Fiber Lasers

The ideal continuous wave (cw) laser converts input pump power with low coherence to a highly coherent optical output signal which is constant in amplitude and wavelength, with the spectral purity of the output (i.e., linewidth) limited only by cavity losses. Practical continuous wave lasers may be characterized by four parameters: the pump power required for the onset of oscillation (i.e., threshold); the conversion efficiency of pump to signal power above threshold; the peak wavelength of the optical output; and the spectral width of the optical output. Continuous wave fiber lasers often have a low threshold, high efficiency, and narrow linewidth relative to other types of laser. Output powers approaching 100 watts have been achieved using specialized techniques.

The threshold power is determined by the total cavity loss and gain efficiency (i.e., gain per unit of absorbed pump power) of the fiber amplifier. For example, for the laser shown in Figure 1 the internal losses are minimal, and the pump power required to reach threshold may be calculated by setting the product of the round-trip small-signal gain, and the output mirror reflectivity, G^2R, equal to unity. For the laser configuration shown in Figure 1, and using eqns [1] and [2] for a four-level laser, the pump power which must be absorbed in the amplifier for the laser to reach threshold is

$$P_{\mathrm{th}} = \alpha_{\mathrm{cav}} L_{\mathrm{C}} \frac{h\nu_{\mathrm{P}} A}{\sigma_{\mathrm{S}} \tau} \qquad [4]$$

in which L_{C} is the cavity length, and $\alpha_{\mathrm{cav}} = \alpha_{\mathrm{int}} + \frac{1}{2} L_{\mathrm{C}} \ln(1/R)$ is the total cavity loss per unit length, comprising both internal losses, α_{int}, and outcoupling losses through a mirror with reflectivity R. For example, if the internal losses were negligible and the transmittance, $T = 1 - R$, of the output reflector in the laser shown in Figure 1 was 50% (i.e. 3 dB), and the other fiber parameters were the same as above (i.e. $G = 0.3$ dB/mW), then the laser threshold would be 5 mW. In three-level fiber lasers ground state absorption of the signal is usually the dominant loss mechanism which must be overcome before a net gain is produced, and hence the threshold of three-level lasers is given approximately by eqn [4] with $\alpha_{\mathrm{int}} = \sigma_{\mathrm{S}} N L_{\mathrm{A}} 2 L_{\mathrm{C}}$, and increases with amplifier length. In either case the pump power required at the input of the laser to reach threshold, $P_{\mathrm{P}}(0)$, may be approximated using the relation $P_{\mathrm{Pabs}} = P_{\mathrm{P}}(0) - P_{\mathrm{P}}(L_{\mathrm{A}}) \approx P_{\mathrm{P}}(0)[1 - \exp(-\sigma_{\mathrm{P}} N L_{\mathrm{A}})]$.

For either three- or four-level lasers pumped above threshold, the inversion of the gain medium remains at its threshold level as the internal round-trip gain must remain unity; however a coherent optical output signal at the peak emission wavelength grows from noise until its amplitude is limited by saturation of the amplifier gain. Provided the output coupling loss is not too large (e.g., $T < 60\%$) such that the total signal intensity is constant along the whole length of the fiber amplifier, it can be shown that changes in pump power above threshold cause a proportional change in the signal power coupled out of the Fabry–Perot cavity in one direction, P^{+}_{Sout}, i.e.

$$P^{+}_{\mathrm{Sout}} = \frac{(1 - R)}{\delta_0} \frac{\nu_{\mathrm{S}}}{\nu_{\mathrm{P}}} (P_{\mathrm{Pabs}} - P_{\mathrm{th}}) \qquad [5]$$

in which $\delta_0 = 2\alpha_{\mathrm{int}} L + \ln 1/R$ represents the total round-trip loss at threshold. The relationship between absorbed pump power and output signal power is shown schematically in Figure 3. The constant of proportionality is known as the slope or conversion efficiency, defined as $\eta_{\mathrm{S}} = P_{\mathrm{Sout}}/(P_{\mathrm{Pabs}} - P_{\mathrm{th}})$. In both three- and four-level fiber lasers with small intrinsic losses $\alpha_{\mathrm{int}} \approx 0$, and small output coupling (i.e., $R \approx 1$), the slope efficiency can approach the intrinsic quantum efficiency, $\nu_{\mathrm{S}}/\nu_{\mathrm{P}}$.

Sustained oscillation can only occur if the optical path length around the cavity is an integral number of wavelengths. Wavelengths satisfying this condition are called ‘modes’ of the cavity, determined by

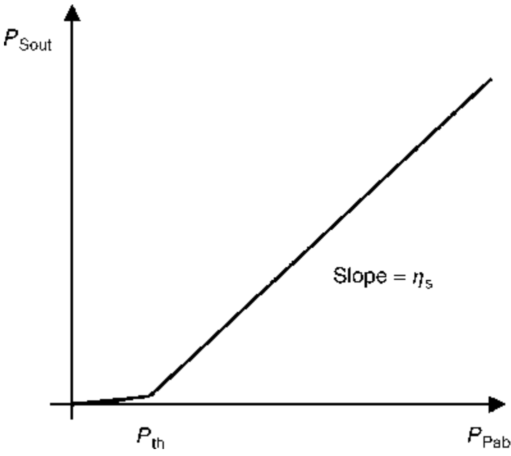

Figure 3 Representative plot of laser output power versus absorbed pump power.

$2nL_C = m\lambda_0$, where n is the effective refractive index seen by light propagating in the cavity, L_C is the cavity length, m is an integer, and λ_0 is the free-space wavelength. In Figure 1, the cavity length comprises the amplifying fiber, and two Bragg grating filters which may be regarded as having an effective length (i.e., less than the actual grating length) which depends on the grating characteristics. In fiber lasers the cavity is typically between several centimeters and many meters in length, hence there are usually very many potential lasing modes within the gain-bandwidth of the amplifying medium. For example, in the Fabry–Perot cavity of Figure 1, if the cavity length was 5 meters, the mode spacing would be $\nu_{m+1} - \nu_m = c/(2nL_C) = 20$ MHz, i.e. orders of magnitude less than the bandwidth of the optical amplifier.

Ideally the first mode to reach threshold would determine the laser output wavelength and the spectral width, or linewidth, of the laser output would be $\Delta\nu_L = h\nu/(2\pi\tau_C^2 P_{\text{Sout}})$, in which τ_C is the lifetime of photons in the cavity and determined by cavity losses, including outcoupling, and given by $\tau_C = t_R/\varepsilon$, where t_R is the cavity round trip time and ε the fractional energy loss in the cavity per round trip. In practice, the linewidth is usually significantly larger than the latter theoretical limit due to transient lasing of different modes within the bandwidth of the optical amplifier and environmental perturbations to the fiber cavity. Fiber amplifiers typically have a large homogeneously broadened gain-bandwidth, and being long and flexible are susceptible to acoustic and thermal perturbations, hence care must be taken to physically stabilize the laser cavity to minimize both mode-hopping and the output linewidth.

Whilst the CW laser shown in Figure 1 has the advantage of simple construction, it would not be ideal for narrow linewidth CW generation. The reason is that counterpropagating waves in the Fabry–Perot cavity form a standing wave, i.e., in which the local signal intensity varies due to interference between forward and backward propagating waves. The standing wave can cause spatial hole-burning in the gain medium (i.e., a spatial variation in gain, linked to the spatial variation in intensity) which in turn reduces the average gain for the lasing mode and promotes mode-hopping and hence spectral broadening of the output.

A preferable arrangement for many fiber lasers is a traveling wave laser, which can be realized using a ring configuration, as shown in Figure 4. In this configuration the optical isolator ensures unidirectional lasing and avoidance of spatial hole-burning. Additional components in the cavity can include a wavelength selective coupler (WSC), i.e., to couple the pump wavelength into the laser cavity whilst not coupling the signal wavelength out, and a narrowband filter (e.g. a small Fabry–Perot resonator) to further stabilize and/or narrow the lasing linewidth. Rare-earth doped fiber lasers with similar cavity configurations have been realized with linewidths of approximately 10 kHz, limited by environmental (e.g., acoustic) perturbations.

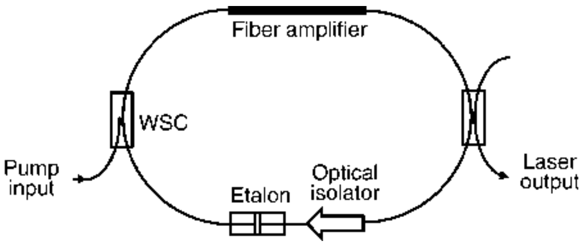

Figure 4 Schematic of a traveling wave laser, useful for low noise continuous wave oscillation. The wavelength selective coupler (WSC) is required to couple pump power into the laser cavity without coupling signal out.

Pulsed Fiber Lasers

There are two main types of fiber laser useful for generating short high-power pulses: *Q*-switched lasers, and mode-locked lasers. *Q*-switched fiber lasers are useful for generating large energy pulses (e.g., microjoules) with very high peak power (e.g., kilowatts) with relatively long pulse duration (e.g., tens of nanoseconds), whilst mode-locked fiber lasers are typically capable of generating ultrashort pulses (e.g., sub-picosecond) with moderate energy (e.g., nanojoules) and moderate-to-high peak power (e.g., tens of watts). Pulsed lasers typically produce less average output power than CW lasers (e.g., up to about 10 W). Both types of pulsed laser typically contain an element for loss modulation within the cavity, however the pulse generation mechanisms are very different.

Q-switched lasers operate by rapid switching of the cavity loss. Whilst the intracavity loss is high and the laser below threshold, energy is transferred from the pump to the lasing medium. The long excited state lifetime and small emission cross-sections of rare-earth dopants assists greatly in the latter process, so that a significant amount of energy can be stored in long fiber amplifiers before amplification of spontaneous emission noise begins to deplete the inversion. After the gain medium is fully inverted the cavity loss is suddenly reduced, and the laser is taken well above threshold, resulting in a rapid build-up of noise and the formation of a pulse which ideally extracts all the stored energy within a few round-trips in the laser cavity. The situation is shown schematically in Figure 5.

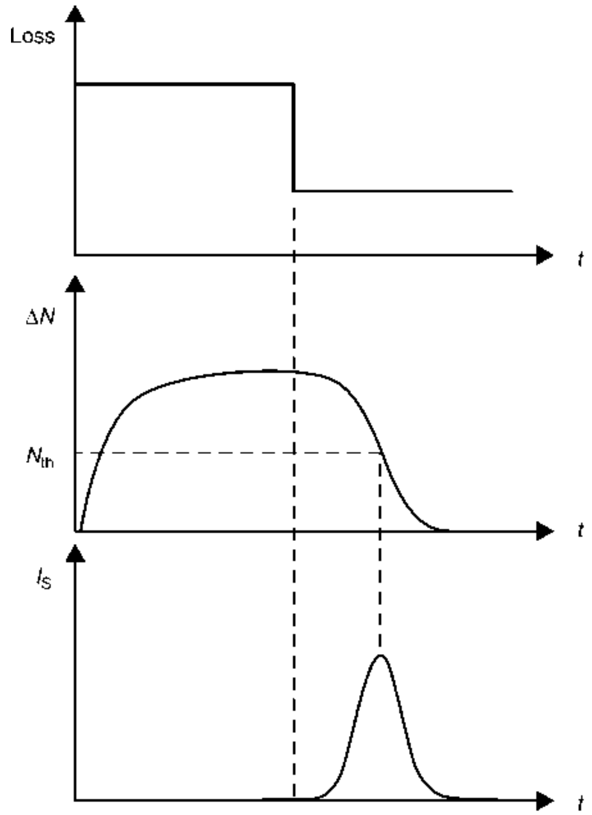

Figure 5 Pulse generation in a *Q*-switched laser. Initially the intracavity loss is high, the laser remains below threshold, and the inversion of the gain medium increases. When the intracavity loss is suddenly reduced, the laser goes well above threshold and spontaneous emission noise is rapidly amplified into a large pulse which extracts much of the energy previously stored in the gain medium.

An upper limit on the pulse energy obtainable from a Q-switched laser may be determined by assuming all active ions in the laser cavity are excited before the pulse is generated (i.e., at the time the cavity Q is switched high), that no ions remain excited immediately after the pulse is generated, and that the pulse generated is much shorter than the time required to pump all the ions into their excited state. The energy stored in the gain medium, and hence the maximum conceivable pulse energy, would then be $E = h\nu_S NV$, where N is the density of active ions and V is the cavity volume. For example, a typical erbium-doped fiber laser of length 10 m and dopant density $N = 10^{19}$ ions/cm in a single-mode fiber with core diameter 7 μm could ideally store up to 500 μJ of energy.

Two things limit the maximum inversion and energy storage achievable in practice: spontaneous emission noise (due to the finite excited-state lifetime) which is amplified and depletes the inversion, and unwanted feedback (e.g., due to the finite extinction ratio of the Q-switching element) which results in cw lasing which limits the inversion. Following from eqns [1] and [2], the unidirectional gain per unit of stored energy in a fiber laser may be expressed as $G = 4.34/(A_{\text{core}}F_{\text{sat}}) = 4.34/E_{\text{sat}}$ decibels per joule, where $F_{\text{sat}} = h\nu_S/\sigma_S$ is the saturation fluence, and $E_{\text{sat}} = A_{\text{sat}}F_{\text{sat}}$ is the saturation energy. For example, if the maximum gain achievable before the onset of cw lasing was 30 dB (i.e., even when the Q-switch introduces a large loss), then the maximum energy storable in the gain medium would be $E = 6.9E_{\text{sat}}$. For the same erbium-doped fiber parameters as used in the previous paragraph, the maximum storable energy would be $E \approx 70$ μJ.

Only a proportion of the energy stored in the gain medium may usually be converted into a Q-switched output pulse. The actual Q-switched pulse energy, duration, and peak power can be calculated by solving the rate equations for the photon flux and the population difference between the lasing energy levels in the gain medium. The equations must usually be solved numerically, however assuming the Q changes much more rapidly than the pulse build-up time, which is in turn much shorter than the time taken for a significant change in inversion to either pumping or spontaneous emission, the pulse parameters may be approximated analytically and are found to be determined by only two parameters, i.e., the population inversion (or stored energy) just before Q-switching, and the lifetime of photons within the cavity, τ_C, defined previously.

Under the conditions described above, standard analysis gives the Q-switched pulse duration as

$$T_P \approx \frac{\tau_C(r-1)}{(r-1-\ln r)} \qquad [6]$$

where r is the ratio of the inversion just before Q-switching (low Q) to the threshold inversion for cw lasing after Q-switching (Q high). In fiber lasers the gain can be very high, hence it is common for $r \gg 1$ and almost all available energy stored in the gain medium is extracted as a Q-switched laser pulse. The Q-switched pulse duration is then determined primarily by the cavity lifetime; shorter pulses can be obtained only by shortening the laser cavity and/or increasing the cavity loss. For example, for a 10 m long ring cavity in which the sum of cavity and outcoupling losses is 10 dB and $r \gg 1$, the pulse duration would be $T_P\tau_C = 22$ ns.

The exact proportion of energy stored in the gain medium that is converted to a Q-switched pulse depends on the number of energy levels in the lasing medium, the degeneracy of those levels, and the rate of nonradiative energy transfer between those levels. For example, for an ideal three-level gain medium (e.g., erbium dopant) with no internal cavity losses the maximum available Q-switched pulse energy is half the stored energy, or $E_{\mathrm{pulse}} = E/2$. It can be shown that the peak of the output pulse power occurs as the inversion decreases through threshold, and may be approximated by dividing the pulse energy by the pulse duration. Even for moderate pulse energies the peak power is potentially very high. For the erbium-doped fiber laser parameters used previously, the peak pulse power would be approximately 1600 W, i.e., sufficient to cause significant spectral broadening of the output wavelength due to stimulated Raman scattering and other nonlinear effects within the optical fiber.

Mode-locked lasers generate optical pulses by forcing many cavity modes to be locked in phase. As per Fourier theory, the pulse repetition rate is related to the spacing of the lasing modes, and the more modes that add coherently, the greater the pulse bandwidth and the shorter the transform-limited pulse duration. For example, the relationship between mode spacing and pulse repetition rate, and between the spectral and temporal envelopes of a train of Gaussian pulses is shown in Figure 6.

In mode-locked lasers a nonlinear or time-varying element must be present which synchronously modulates the optical signal circulating in the cavity in amplitude or phase at a frequency which is a harmonic of the cavity mode spacing. There are two main methods of mode locking:

(i) active mode locking (e.g., by an electro-optic modulator), in which amplitude or phase modulation is driven by an external source; and
(ii) passive mode locking (e.g., by a fast or slow saturable absorber), in which an element with nonlinear gain or loss responds to the optical signal in the laser cavity itself.

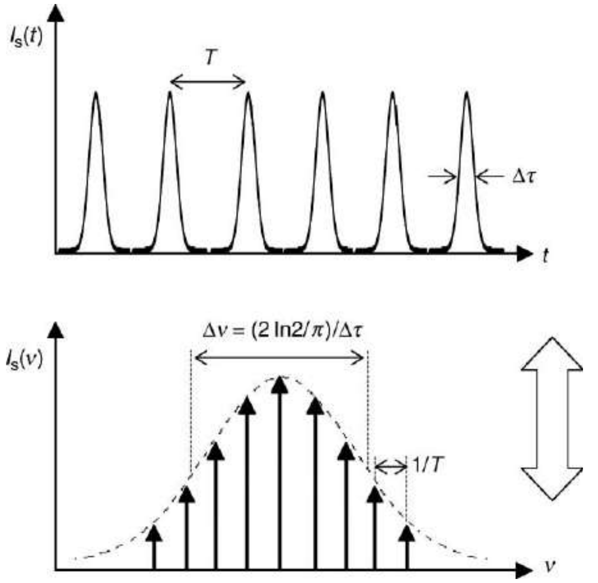

Figure 6 Schematic showing the Fourier-transform relationship between an infinite train of Gaussian pulses and the lasing modes in a mode-locked laser. The temporal and spectral widths are measured full width at half the maximum intensity.

In both cases the modulation provided by the mode-locking element distributes energy between neighboring modes, effectively injection-locking them in phase. A variation of mode locking techniques often present in fiber lasers is soliton mode-locking, in which the active or passive modulation is weak, and a balance between nonlinearity and dispersion is the dominant process underlying pulse formation. Hybrid mode-locking techniques can also be used to combine the best aspects of different pulse-generation methods.

Whilst the latter frequency-domain view of mode locking is helpful, nonlinear effects are almost always significant in mode-locked fiber lasers, and hence the linear concept of a mode is not appropriate for detailed analysis. Hence it is usually preferable to describe pulse generation in mode-locked fiber lasers in the time domain; for stable operation the pulse at any point in the cavity must self-replicate after each round-trip, with the net effects of temporal and spectral loss, gain, dispersion and nonlinearity all canceling.

An optical pulse propagating in a fiber laser is usually described by a slowly varying complex envelope, $\Psi(z,t)$, where $|\Psi(z,t)|^2$ is the instantaneous power in a frame of reference moving at the pulse group velocity, and $\arg[\Psi(z,t)]$ is the instantaneous phase relative to the optical carrier. The various elements in the laser cavity, such as the passive fiber, optical fiber amplifier, filters, and mode-locking devices, each modify the complex pulse envelope as it propagates around the laser cavity.

For example, in the absence of other effects, group velocity dispersion (GVD) causes different wavelengths to travel at different velocities along the fiber, resulting in 'chirping' and temporal broadening of optical pulses by an amount proportional to the distance propagated. The effect can be described mathematically as the accumulation of a phase shift which is quadratic in frequency, i.e., $\Psi(z,\omega) = \Psi(0,\omega)\exp[\mathrm{i}\beta_2\omega^2 z/2]$. Taking the derivative with respect to propagation distance, and using $\partial/\partial t = -\mathrm{i}\omega$, the incremental change in the complex envelope with propagation distance is

$$\frac{\partial \Psi}{\partial z} = -\mathrm{i}\frac{\beta_2}{2}\frac{\partial^2 \Psi}{\partial t^2} \quad [7]$$

If the GVD parameter, β_2, is positive then long wavelengths propagate faster than shorter wavelengths, and the dispersion is called 'normal'. If the reverse is true, the dispersion is 'anomalous'. In standard single-mode step index optical fibers the dispersion passes through zero at wavelengths around 1300 nm, and is anomalous for longer wavelengths. It is often necessary to control the net dispersion in a fiber laser, and this can be done by incorporating specially designed fibers with wavelength-shifted or otherwise modified dispersion characteristics, or by using chirped Bragg grating reflectors.

Whilst silica is not commonly regarded as a nonlinear medium (i.e., in which the optical properties are a function of the intensity of light), in ultrashort pulsed fiber lasers the peak powers are usually such that nonlinear effects cannot be ignored. In silica the main nonlinearity is associated with the Kerr effect, and is readily observed as an intensity-dependent refractive index, i.e. $n = n_0 + n_2 I$, where the nonlinear index, n_2, is typically about 3×10^{-20} m^2/W in silica fibers. Other nonlinear effects related to the Kerr nonlinearity which can sometimes be important are cross-phase-modulation and four-wave mixing. In the absence of other effects, the nonlinear index results in self-phase modulation (SPM), or a phase shift in the optical carrier which is proportional to the instantaneous intensity of the signal. Mathematically, $\Psi(z,t) = \Psi(0,t)\exp[\mathrm{i}\gamma|\Psi(0,t)|^2 z]$, in which $\gamma = n_2\omega_0/(cA)$ is the nonlinear coefficient of the fiber, ω_0 is the radian frequency of the optical carrier, and A is the effective core area for the guided mode. The incremental change per unit length in the complex pulse envelope induced by self-phase modulation is then

$$\frac{\partial \Psi}{\partial z} = \mathrm{i}\gamma|\Psi|^2\Psi \quad [8]$$

For example, self-phase-modulation is used to realize a nonlinear switching element in the passively mode-locked laser cavity shown schematically in Figure 7. For obvious reasons the laser is called a 'figure-8' laser, though it is actually a ring laser into which a device known as a nonlinear amplifying loop-mirror (NALM) has been incorporated. The NALM acts as a fast saturable absorber, with transmittance dependent on the instantaneous optical intensity. The NALM operates as follows: light propagating through the optical isolator is divided equally into co- and counter-propagating waves in the loop-mirror on the right-hand side. The light propagating in a clockwise direction is amplified immediately, and propagates around the loop with high intensity, whilst the counter-clockwise propagating signal has a low intensity for most of its time in the loop. The difference in intensity means that the co- and counter-propagating waves experience different nonlinear phase shifts. When they are recombined at the coupler, high-intensity light is transmitted back into the ring, whilst low-intensity light is reflected back to the isolator, and lost. In this way the NALM provides both gain and a nonlinear transmittance which depends on the instantaneous power of the input, described by

$$T \approx G\left\{1 - \frac{1}{2}\left[1 + \cos\left(\frac{1}{2}\gamma L P_{\mathrm{in}} G\right)\right]\right\} \quad [9]$$

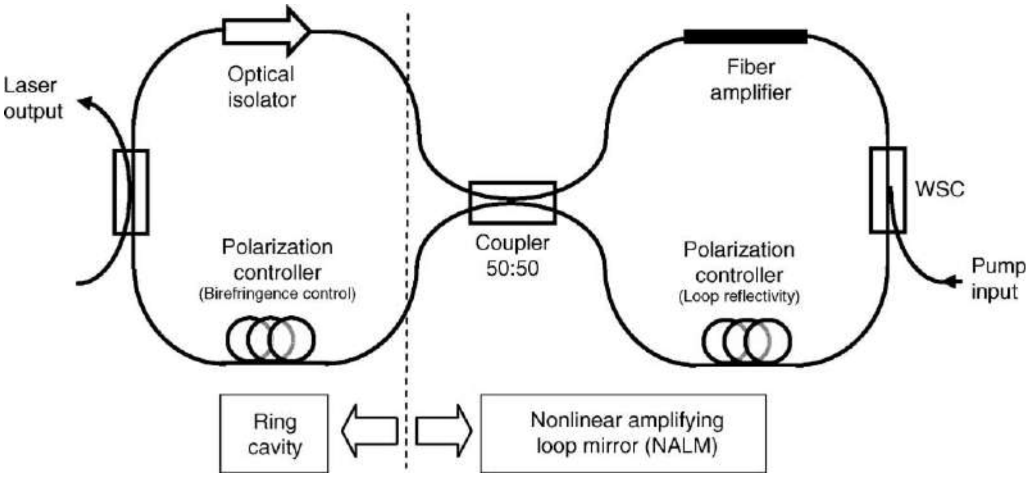

Figure 7 Schematic of a 'figure-8' fiber laser. The laser is passively mode-locked by the nonlinear amplifying loop mirror on the right, which acts as a fast saturable absorber, transmitting and amplifying high-power optical signals whilst reflecting low optical power signals.

in which G is the linear gain of the amplifier, and L the loop length, P_{in} is the input power, and T the ratio of transmitted power to input power. The nonlinear transmittance of the NALM promotes pulse formation and shortening in the laser cavity, and hence passively mode-locks the laser. The polarization controllers are to compensate for unwanted birefringence in the optical fiber, and to control the transmittance of the loop-mirror at low input powers. The fiber dispersion and nonlinearity, together with other factors, further influence the pulses generated in the laser, as described in the following paragraphs.

The combined effects of dispersion and self-phase-modulation can give rise to some interesting and useful physical phenomena. In lossless optical fibers, or fiber lasers in which loss is balanced by gain, pulse propagation is described by the nonlinear Schrödinger equation (NLSE), i.e., a nonlinear wave equation including the effects of both group velocity dispersion and intensity-dependent refractive index. The NLSE may be derived by combining eqns [8] and [9], i.e.

$$\mathrm{i}\frac{\partial \Psi}{\partial z} - \frac{\beta_2}{2}\frac{\partial^2 \Psi}{\partial t^2} + \gamma|\Psi|^2\Psi = 0 \qquad [10]$$

It can be shown that if the dispersion is anomalous ($\beta_2 < 0$) then for a particular shape of pulse envelope the distributed effects of dispersion and nonlinearity can exactly cancel, so $\partial\Psi/\partial z = 0$, and the pulse propagates without changing shape. Such pulses are called fundamental solitons, the solution for which is $\Psi(z,t) = \Psi_0 \operatorname{sech}(t/T_0)\exp[\mathrm{i}z/(2L_{\mathrm{D}})]$, in which the peak amplitude is $\Psi_0 = 1/\sqrt{\gamma L_{\mathrm{D}}}$, T_0 is the pulse width parameter, and $L_{\mathrm{D}} = T_0^2/\beta_2$ is known as the dispersion length. For example, in a standard silica fiber at $\lambda = 1.55\ \mu\mathrm{m}$, $\beta_2 \approx -20\ \mathrm{ps}^2/\mathrm{km}$ and $\gamma \approx 10\ \mathrm{W}^{-1}\mathrm{km}^{-1}$, hence a soliton with $T_0 = 1$ ps would have peak power 2 W and energy $E_0 = 2|\beta_2|/(\gamma T_0) = 4$ pJ. The energy of a soliton is quantized, in that pulses with initial energy between 0.5 and 1.5 times E_0 can evolve into a fundamental soliton during propagation. Consequently solitons behave somewhat like particles, and are robust even in the presence of significant perturbations. The pulses which form spontaneously in the figure-8 fiber laser have many of the properties of solitons.

Mode-locked fiber lasers, such as that shown in Figure 7, can often be described by a generalized nonlinear Schrödinger equation, in which extra terms are included in eqn [10] to account for the average effects of an excess linear gain or loss, nonlinear gain and loss (e.g., due to a saturable absorber), spectral filtering, and temporal modulation, etc. Assuming the effects of the added elements can be regarded as distributed over the length of the cavity rather than lumped, such that the envelope of any pulse generated in the laser does not change shape significantly as it propagates around the cavity, then the master differential equation describing propagation of an optical signal within a fiber laser which is passively mode-locked by a fast saturable absorber can be written

$$\mathrm{i}\frac{\partial U}{\partial \zeta} + \left(\frac{D}{2} - \mathrm{i}\beta\right)\frac{\partial^2 U}{\partial \tau^2} + (1 - \mathrm{i}\varepsilon)|U|^2 U$$
$$= \mathrm{i}\delta U - (\nu - \mathrm{i}\mu)|U|^4 U \qquad [11]$$

Equation [11] is known as the Ginzburg–Landau equation, which describes propagation of the normalized pulse envelope, $U(\zeta,\tau)$, as a function of delayed time, τ, and normalized distance, ζ, and under the combined influence of group velocity dispersion ($D = \pm 1$ for anomalous or normal group velocity dispersion, respectively), dispersion due to band-limiting by amplification in a homogeneously broadened gain medium ($\beta > 0$), nonlinear refractive index, nonlinear gain or loss (ε), linear gain (δ), and quintic terms, μ and ν, which if less than zero relate to saturation of the nonlinear gain and refractive index, respectively. Note that if the right-hand side of the equation is set to zero, and $\varepsilon = \beta = 0$, then the equation reduces to the normalized form of the nonlinear Schrödinger equation. Stable pulses generated within the laser are solutions of the above equation with $\partial U/\partial\zeta = 0$, and hence the effect of all terms must balance. Consequently it is possible for stable pulses to be generated even with normal dispersion in the cavity, though the pulses are then generally strongly chirped, and have an envelope which is approximately Gaussian shaped rather than the hyperbolic secant shape expected with anomalous dispersion. Stable soliton-like pulses can be found by solving the Ginzburg–Landau equation for specific ranges of parameters. There are many numerical solutions, and some analytical solutions, but in general the dynamics of pulse formation and propagation in passively mode-locked fiber lasers can be quite complex, and for some sets of parameters remarkable and unexpected numerical solutions can be found. One such example which has been observed experimentally is the 'exploding' soliton, which unlike the usual soliton is periodically unstable, as shown in the plot of pulse evolution in Figure 8.

Other Fiber Lasers

The lasers discussed in the previous sections are indicative of the main types of fiber laser in terms of their output characteristics; however it should be

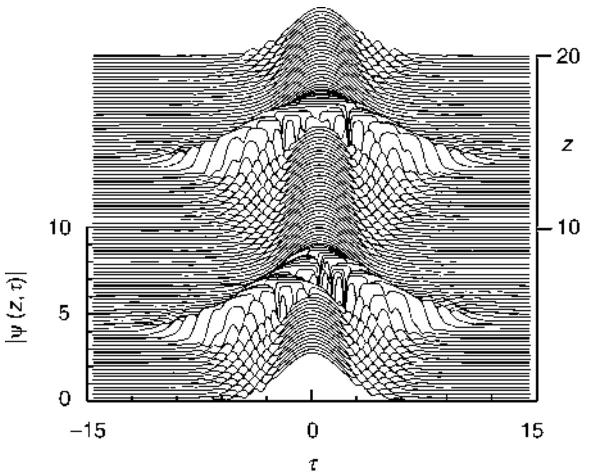

Figure 8 Two periods of the evolution of an exploding soliton found by numerical solution of the quintic Ginzburg–Landau equation with normalized parameters $\varepsilon = 1.0$, $\delta = 0.1$, $\beta = 0.125$, $\mu = 0.1$, and $\nu = 0.6$. The process never repeats itself exactly in successive 'periods'; however the pulse always returns to the same shape. Such solutions cannot be found in analytic form; nevertheless they are as common as stationary pulse solutions and exist for a wide range of parameters.

noted that there is a wide range of other fiber laser configurations with similar behaviors incorporating a variety of other fiber and/or bulk elements. For example, specially designed fibers in which the pump and/or signal has a large cross-sectional area can be used to realize fiber lasers with a high output power, comparable to that in many bulk laser systems. Linear and/or nonlinear processing of the output of fiber lasers can be used to increase their peak output power and/or pulse energy, or to change their output wavelength (i.e., by harmonic generation) from that achievable in the laser cavity itself. In mode-locked fiber lasers it is now common to include passive semiconductor devices which provide a slow saturable absorber action to initiate and stabilize soliton mode-locking.

It should also be noted that there are a number of fiber lasers which generate optical outputs using different or modified physical processes compared to those described in previous sections. For example, it is possible to realize up-conversion fiber lasers in which the output wavelength is shorter than the pump wavelength, multiwavelength fiber lasers (pulsed and cw) which produce outputs at several wavelengths simultaneously, and lasers with limited output coherence such as chaotic fiber lasers and broadband superfluorescent sources, etc. Lastly, whilst the rare-earth doped fiber lasers dealt with in previous sections have some very desirable properties, it should be recognized that optical fiber lasers based on other optical gain mechanisms, such as stimulated scattering and parametric effects, can also be realized (e.g., as in fiber Raman lasers, and fiber Brillouin lasers). In the latter cases the lasing wavelength is limited only by available pump wavelengths. The reader is referred to the references in the Further Reading for information on the wide range of fiber lasers that are possible in practice.

List of Units and Nomenclature

Absorption cross-section	σ_a	$[\mathrm{m}^{-2}]$
Amplifier length	L_A	[m]
Area	A	$[\mathrm{m}^2]$
Cavity length	L_C	[m]
Complex amplitude	Ψ	$[\mathrm{W}^{0.5}]$
Dispersion length	L_D	[m]
Distance	z	[m]
Emission cross-section	σ_e	$[\mathrm{m}^{-2}]$
Frequency	v	[Hz]
Gain coefficient	g	$[\mathrm{m}^{-1}]$
Group velocity dispersion	β_2	$[\mathrm{s}^2\mathrm{m}^{-1}]$
Inversion ratio	r	
Lifetime	τ	[s]
Loss coefficient	α	$[\mathrm{m}^{-1}]$
Nonlinear coefficient	γ	$[\mathrm{m}^{-1}\mathrm{W}^{-1}]$
Nonlinear refractive index	n_2	$[\mathrm{m}^2\mathrm{W}^{-1}]$
Number density	N	$[\mathrm{m}^{-3}]$
Power gain	G	
Pump power	P_P	[W]
Pump wavelength	λ_P	[m]
Reflectivity	R	
Saturation energy	E_{sat}	[J]
Saturation fluence	F_{sat}	$[\mathrm{J.m}^{-2}]$
Signal power	P_S	[W]
Signal wavelength	λ_S	[m]
Threshold power	P_{th}	[W]

See also

Fiber and Guided Wave Optics: Dispersion; Fabrication of Optical Fiber; Light Propagation; Nonlinear Effects (Basics); Optical Fiber Cables. **Fiber Gratings**. **Lasers:** Up Conversion Lasers. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed). **Solitons:** Temporal Solitons. **Scattering:** Raman Scattering; Stimulated Scattering. **Ultra-Fast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Agrawal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Akhmediev NN and Ankiewicz A (1997) *Solitons*. London: Chapman and Hall.
Becker PC, Olsson NA and Simpson JR (1999) *Erbium-Doped Fiber Amplifiers: Fundamentals and Technology.* San Diego, CA: Academic Press.
Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.
Digonnet MJF (ed.) (2001) *Rare-Earth-Doped Fiber Lasers and Amplifiers,* 2nd edn. New York: Marcel Dekker.
Duling IN (ed.) (1995) *Compact Sources of Ultrashort Pulses*. Cambridge, UK: Cambridge University Press.
France PW (ed.) (1991) *Optical Fibre Lasers and Amplifiers*. Glasgow, UK: Blackie.
Jeunhomme LB (1990) *Single-Mode Fiber Optics: Principles and Applications*, 2nd edn. New York: Dekker.
Kashyap R (1999) *Fiber Bragg Gratings*. San Diego, CA: Academic Press.
Siegman AE (1986) *Lasers*. Oxford, UK: Oxford University Press.
Weber MJ (ed.) (1995) *CRC Handbook of Laser Science and Technology.* Boca Raton, FL: CRC Press.

Organic Semiconductors and Polymers

G A Turnbull, University of St Andrews, Fife, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Organic semiconductors, and conjugated polymers in particular, are emerging as an interesting new class of visible laser media. The polymers are electroluminescent, broadband visible emitters and exhibit large optical gain coefficients. As plastics, they are amenable to simple processing, and can be shaped into complex structures. This combination of photonic, electronic, and processing properties make these materials extremely well suited to forming novel microlasers, with future prospects as inexpensive, plastic diode lasers operating throughout the visible spectrum.

Conjugated Polymers

Plastics are widely used in low-cost optical components such as lenses, diffraction gratings, optical fibers, and adhesives. Their common use derives from a combination of qualities: good optical transmission, a wide range of mechanical properties and simple fabrication via molding: extrusion or solution processing. Functional optical polymers, including liquid crystals and photorefractives, have also been widely applied to polarization control and for nonlinear- and electro-optics. Conjugated polymers are a remarkable class of functional plastics that exhibit some unusual properties. Firstly, they are plastics that can conduct electricity; and in fact are electrically semiconducting. Secondly, they are plastics that can be stimulated either optically or electrically to emit visible light.

These two unusual properties derive from a particular chemical structure specific to this family of plastics. In all conjugated polymers, the backbone of the molecule is a chain of carbon atoms joined by alternating single and double covalent bonds. The simplest example is illustrated in Figure 1a, which shows the chemical structure of the polymer poly(acetylene). Structural integrity of the polymer is provided by a string of single σ-bonds between adjacent carbon atoms. The additional bonding electrons in the double bonds are located in π-orbitals oriented perpendicular to the polymer chain. Adjacent π-orbitals overlap, and their electrons can be widely delocalized along the polymer backbone. The delocalized electrons can move relatively freely along the chain, giving the electronic properties of the material. The π-orbitals can be transformed between bonding and anti-bonding states by the absorption or emission of light, leading to strong optical transitions. While poly(acetylene) is a very inefficient light emitter, other conjugated polymers, notably those that include phenyl rings in the backbone (Figure 1b–d), can be highly emissive. Photoluminescence quantum

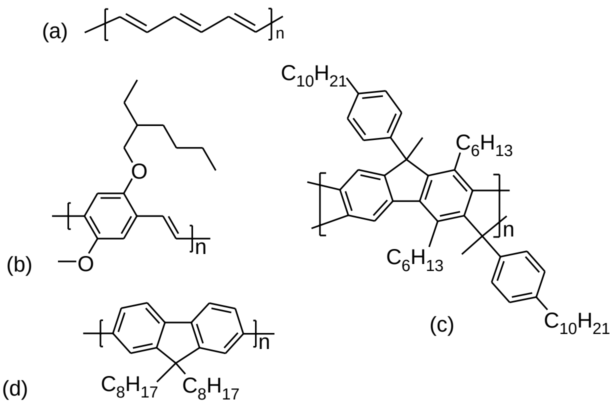

Figure 1 Chemical structures of several conjugated polymers. (a) poly[acetylene]; (b) poly[2-methoxy-5-(2′-ethylhexyloxy)-1,4-phenylene vinylene], (MEH-PPV); (c) ladder-type poly[1,4-phenylene], (MeLPPP); (d) poly[9,9-dioctylfluorene], (PFO).

efficiencies, approaching unity in solution and as high as 60% in the semiconducting solid state, have been realized.

Organic Semiconductor Lasers

While organic dye lasers have been widely used since the mid-1960s, the history of organic semiconductor lasers is much more recent. Other than a few early observations of lasing in molecular crystals in the 1970s by Karl and others, the field of organic semiconductor lasers began in earnest following the related discovery in 1990 by Burroughes *et al.* of electroluminescence in a conjugated polymer. In 1992, Moses demonstrated lasing for the first time in a solution of the conjugated polymer MEH-PPV (Figure 1b). However, it was another four years – once material synthesis had been sufficiently refined – before stimulated emission and then true laser action were observed in polymers in the semiconducting solid state. In the following year the first demonstrations were made of dye-doped small molecular semiconductor lasers. Since that time, there has been a great growth in interest in the field with an average of ~30 research papers per year. Further details of the development of the field of organic semiconductor lasers can be found in the Further Reading.

So what are the particular properties of conjugated polymers that have stimulated interest in developing organic semiconductor lasers? Firstly these materials are broadband visible emitters. Through appropriate chemical design, their emission can be tuned throughout the visible spectrum, from 400 nm to 700 nm (Figure 2). The polymers exhibit large optical gain coefficients, because the optical transitions are dipole-allowed. Compared with small molecular dyes, polymers exhibit a relatively low degree of self-absorption of the emitted light, and much reduced quenching of the luminescence with increased concentration. This means that it is possible to work with much higher chromophore densities than conventional laser dyes, which is advantageous for achieving very low threshold lasing in extremely compact resonators. Furthermore, because these materials are semiconducting, they have the potential for direct electrical excitation. This could lead to semiconductor diode lasers operating throughout the visible spectrum and, in particular, in the blue and green spectral bands where currently there are few commercially viable inorganic semiconductor sources. Finally, these materials are plastics and so are amenable to simple processing, and can be shaped and structured into complex feedback resonators.

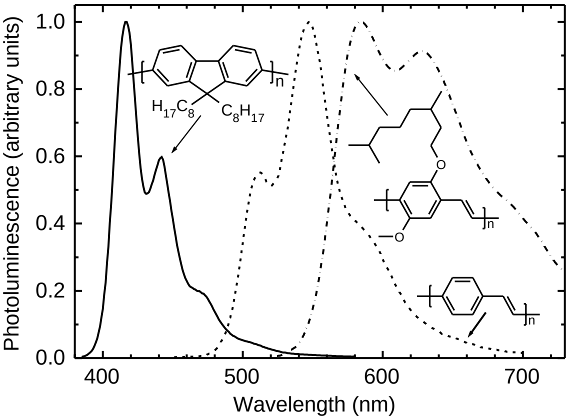

Figure 2 Photoluminescence spectra of three commonly used conjugated polymers; chemical structures inset.

Organic Semiconductor Gain Materials

A great advantage of organic laser media is that a very wide range of materials can be readily synthesized, and through minor changes in the molecular structure one can relatively simply tune the emission properties of the material. Many different organic semiconductors have been applied as lasers, though they fall into three major categories. These are conjugated polymers, dye-doped small molecular semiconductors, and molecular single crystals. Conjugated polymers make up the most diverse materials set, most notably including derivatives of poly(paraphenylene-vinylene), poly(paraphenylene), and poly(fluorene) (Figure 1b–d). The second category comprises small molecular semiconductors that are co-evaporated under vacuum with a dopant laser dye. In this blend of materials, the organic semiconductor functions simply as an energy transfer host for the emitting dye molecules. In the third category are organic single crystals, including materials such as polyacenes and thiophene oligomers. These materials have the advantage of having the highest electrical mobilities of any organic semiconductors, which is important for electrical excitation. However, they can be demanding to grow into high-quality single crystals, and so lack the key processing advantages of the other families of organic semiconductors.

Figure 3a shows a typical exciton energy level structure of a semiconducting polymer. Similar to a conventional laser dye, the energy levels comprise a manifold of singlet states and a manifold of triplet states, each with vibrational sublevels. The laser transition occurs from the bottom of the first excited singlet state S_1 to the vibronic overtones of the ground

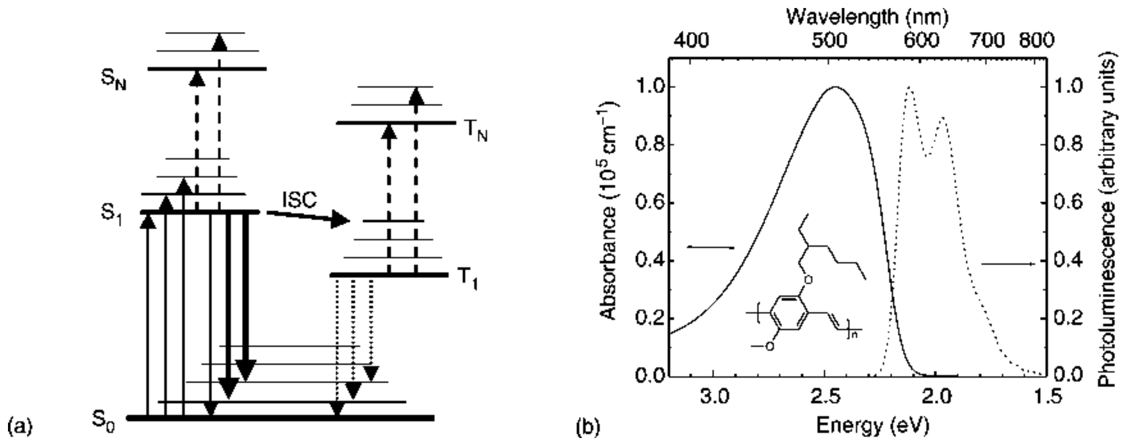

Figure 3 (a) Generic energy level structure and optical transitions of a conjugated polymer. S_N refers to the Nth singlet state; T_N refers to the Nth triplet state; ISC: intersystem crossing. (b) Absorption and photoluminescence spectra of the polymer MEH-PPV.

state S_0, so forming a classic 4-level laser system. The optical transition is dipole-allowed and so has a high oscillator strength. As a result, conjugated polymers typically exhibit simulated emission cross-sections in the range 10^{-16} cm^2 to 10^{-15} cm^2, while the population inversion lifetime is as short as a few hundred picoseconds.

There are, however, several other transitions that can compete with the stimulated emission process. There can be excited-state absorption from the S_1 singlet state into higher-lying singlet states. Additionally intersystem crossing from the S_1 state to the lowest excited triplet state T_1 can subsequently lead to excited triplet absorption. Other processes in the polymer can also impede the stimulated emission. Firstly, charged excitations can exhibit absorption bands that overlap with the emission. Secondly, at high excitation densities, exciton–exciton annihilation can significantly deplete the excited states available for stimulated emission. Ultimately, singlet–singlet annihilation is the limiting factor for the maximum excitation density possible, and while polymer films can have a repeat-unit density of $\sim 10^{21}$ cm^{-3}, the excited state density is typically limited to around 10^{18} cm^{-3}. In many materials this is not restrictive on achieving optical gain because stimulated emission can occur at densities of 10^{17} cm^{-3} and below.

An important factor that distinguishes the photophysics of semiconducting polymers from conventional laser dyes is an internal process of exciton migration between the absorption and emission of a photon. Excitations are confined to short straight sections of the polymer between kinks or twists in the chain. These straight sections, or 'sites', are typically a few repeat units in length, though there is a statistical spread of site lengths in any given material. When formed, excitons tend to hop to the longest, lowest-energy, sites before emitting a photon. The exciton migration process has two effects. First, it means the emission takes place from a subset of the chain segments, resulting in a lower degree of inhomogeneous broadening of the emission spectrum (vibronic peaks are better resolved than in absorption (see Figure 3b)). Second, the peaks of the absorption and emission spectra tend to be widely separated in wavelength, reducing the level of self-absorption at the lasing wavelength.

As mentioned above, blending materials is a particularly attractive approach towards further reducing self-absorption in organic semiconductor lasers. In such systems, a narrow bandgap dopant chromophore is blended with a concentration of 1–2% in a wider bandgap semiconducting host. Emission from the host is strongly quenched via nonradiative Förster energy transfer to the guest molecule. The blended material exhibits an absorption profile that is dominated by the host absorption, while the emission profile is almost completely that of the guest dopant. By strongly separating the emission peak from the absorption edge of the host, it is possible to achieve a substantially reduced (by perhaps an order of magnitude or more) residual absorption coefficient at the lasing wavelength. In doing so, the combined material acts much more like an ideal 4-level laser medium and can thereby have substantially lower pump thresholds for stimulated emission.

Measuring Gain

The dynamics of the excited state population in semiconducting polymers have been widely studied using femtosecond transient absorption experiments.

This is a pump–probe technique in which the transmission of the broadband probe pulse is compared in the presence and absence of a pump pulse that excites molecules into the first excited state. Such measurements have the time resolution of the pump pulses used and so can typically be ~100 fs. By systematically delaying the arrival of the broadband probe pulse, it is possible to map out the evolution of the gain dynamics across the emission spectrum of the material with very high time resolution.

Figure 4 shows a series of snapshots of the absorption induced by a pumping pulse in a poly(thiophene) derivative. This figure shows the changing absorption of the sample between 200 fs and 400 ps after excitation. Within the absorption band of the material, one finds that the level of absorption is slightly reduced because the pump pulse has partly depleted the ground state. In the emission band there is also a reduction in the absorption that mimics in shape the photoluminescence spectrum of the polymer. This negative absorption – or optical gain – shows direct evidence for stimulated emission through much of the emission band. In the low energy tail of the emission, however, there is an increase in absorption that arises due to the excited-state absorption process mentioned earlier. Clearly, for a good laser material, the spectra of excited-state and charge-induced absorptions should overlap as little as possible with the photoluminescence.

Many other investigations of gain in semiconducting polymer films have used a phenomenon known as spectral line narrowing to explore indirectly the amplifying properties of the medium. The experimental procedure for such studies is as follows. A thin film of the semiconducting material, of typically 100 nm thickness, is excited by a pulsed pump laser focused to a stripe of dimensions ~100 μm wide by a few millimeters in length. A significant fraction of the light emitted by the material is trapped in the film by total internal reflection at the polymer–air and polymer–substrate interfaces, and is thereby waveguided along the length of the excitation stripe. This waveguided spontaneous emission can be amplified by stimulated emission before being emitted out the edge of the film.

One finds that, above a particular threshold of pump energy, there is a superlinear increase in the output energy with pump intensity. Associated with this is a dramatic narrowing of the broad emission spectrum to a linewidth of typically less than 10 nm, as illustrated in Figure 5. These two features are signatures of gain narrowing via the process of amplified spontaneous emission. The line-narrowed feature usually appears at the first vibronic overtone of the optical transition, where the gain exceeds the losses in the material by the greatest amount. This particular wavelength experiences the greatest amplification and, above a particular pumping density, undergoes a runaway effect in which most of the light is stimulated to emit at that wavelength at the expense of the rest of the spectrum.

In the case of amplified spontaneous emission or mirrorless lasing as it is also known, the spontaneous emission guided along the length of the stripe acts as

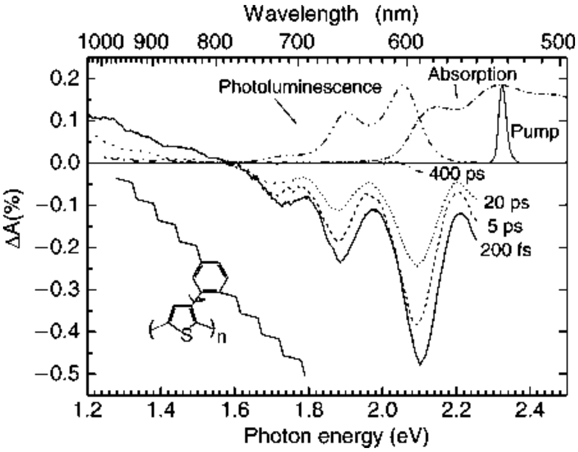

Figure 4 Transient absorption spectra of poly[3-(2,5-dioctylphenyl)-thiophene]; chemical structure inset. (Reproduced from Ruseckas A, Theander M, Valkunas L, Andersson MR, Ingänas O and Sundström V (1998) Energy transfer in a conjugated polymer with reduced inter-chain coupling. *Journal of Luminescence* 76 & 77: 474–477. Copyright (1998), with permission from Elsevier.)

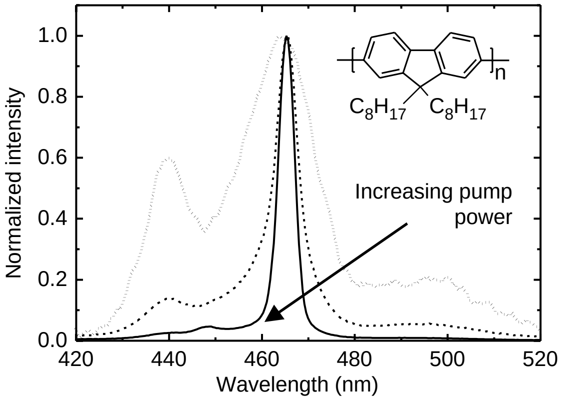

Figure 5 The change in spectral shape with increasing excitation density (from 10 to 133 μJ cm^{-2}) of the edge emission from a PFO film – the normalized spectra show a significant spectral line narrowing at high excitation densities. (Adapted from Heliotis G, Bradley DDC, Turnbull GA and Samuel IDW (2002) Light amplification and gain in polyfluorene waveguides. *Applied Physics Letters* 81: 415–417. Copyright (2002), with permission from American Institute of Physics.)

the 'probe pulse'. In these experiments it is therefore necessary to extract information indirectly about the net gain in the material. By integrating the spontaneous and stimulated emission along the length of the stripe, one finds that the wavelength-dependent output intensity $I(\lambda)$ of the spectral line narrowed light is given by the relationship

$$I(\lambda) \propto \frac{1}{g(\lambda)}[\exp[g(\lambda)l] - 1] \qquad [1]$$

where I_p is pumping intensity, l is the length of the stripe, and $g(\lambda)$ is the net gain coefficient. Therefore, by monitoring the intensity of the line-narrowed emission as a function of the length of the pump stripe, one can extract the value of the net gain for the material. Similarly, by progressively moving the stripe further and further away from the edge of the film, it is possible to evaluate the waveguide losses of light propagating through an unpumped region of the waveguide. Waveguide losses in conjugated polymers typically lie in the range of 3–50 cm^{-1}. Net gains can be as high as 60 cm^{-1} or 260 dB cm^{-1} at modest pumping densities of 4 kW cm^{-2}. These substantial gains, in very small propagation lengths, therefore highlight the attraction of waveguide-based resonators for polymer lasers. In the next section we will discuss the most commonly used feedback structures that have been used for polymer lasers.

Polymer Laser Resonators

The first demonstration of cavity lasing in a conjugated polymer was made by Moses in 1992. The resonator consisted simply of two flat dielectric mirrors on either side of a cuvette containing the polymer in solution. Lasing occurred at the peak wavelength of the gain profile. Subsequent developments in solid-state polymer lasers have involved more sophisticated resonators that exploit the capacity of these materials for very simple processing. For example, solution-processing methods such as spin-casting or dip-coating, make it possible to fabricate very high-quality optical waveguides. Wavelength-scale microstructures can also be readily formed, allowing the straightforward fabrication of laser microresonators.

Many different resonators have been explored for solid-state polymer lasers. Some of the key types are shown in Figure 6, including different geometries of both microcavities and distributed feedback structures. The first resonator used, by Tessler *et al.*, was a planar microcavity (Figure 6a), in which the polymer film was sandwiched between two mirrors. The structure supported standing-wave resonances perpendicular to the plane of the film, at three wavelengths in the emission band of the polymer. Above laser threshold, light was preferentially stimulated into only one of these modes. Microcavities of other geometries have also been studied, including annular and spherical microresonators (Figure 6b,c), in which the optical field is confined in waveguide modes, whispering gallery modes, or a superposition of the two. Such structures provide a longer interaction length in the gain medium for stimulated emission; though tend to lase on many longitudinal modes and have no directional output beam.

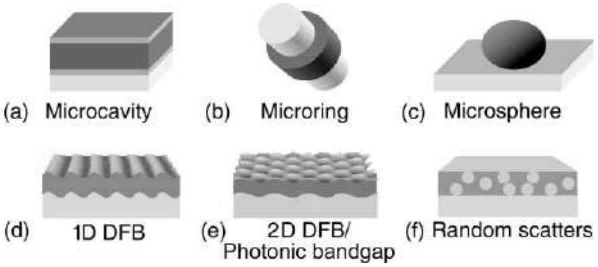

Figure 6 Resonators used for conjugated polymer lasers. (a) Planar microcavity; (b) annular microcavity; (c) spherical microcavity; (d) distributed feedback resonator; (e) 2D DFB/ photonic crystal resonator; (f) random laser.

The final key group of resonators are distributed feedback (DFB) structures, which researchers have progressively favored to attain the low pump threshold densities that will be required to realize electrically pumped lasing. Figure 6d shows a typical DFB laser structure. The device is based on a thin polymer film that is typically 100 nm in thickness deposited on top of a silica substrate. The silica–polymer–air structure forms an asymmetric slab waveguide that supports only the lowest order transverse electric mode TE$_0$. To provide the feedback, the substrate is modulated with a wavelength-scale periodic grating structure. These corrugations give rise to Bragg scattering of the waveguide mode. Through careful selection of the grating period, one may arrange that optical wavelengths close to the peak of the polymer gain will be Bragg scattered from the propagating TE$_0$ mode to the counter-propagating TE$_0$ mode. This is achieved when the Bragg equation

$$2n_{\mathrm{eff}}\Lambda = m\lambda_{\mathrm{B}} \qquad [2]$$

is satisfied, where Λ is the grating period, n_{eff} is the effective refractive index of the waveguide, m is an integer, and λ_{B} is the Bragg wavelength. In so doing, one may arrange that two counter-propagating waveguide modes can be coupled together, thereby providing the distributed resonant feedback for the laser process. Commonly in semiconducting polymer lasers, a second-order (i.e., $m = 2$) Bragg grating has been favored to provide the feedback for the laser, so that the first-order scattering provides a surface-emitted output coupling, perpendicular to the plane of the guide. This surface-emitting geometry is particularly attractive in polymer lasers because it is relatively difficult to achieve a high optical quality edge to the waveguide. DFB lasers have the advantage over simple microcavity resonators in that they can combine long interaction lengths, limited by the scattering length of the Bragg grating, with excellent wavelength selection, which is controlled by the grating period.

In addition to conventional 1D DFB gratings, a number of more exotic 2D and 3D DFB, and photonic crystal resonators have been explored in semiconducting polymers. Even randomly located scattering sites in a polymer film have been used to provide a weak closed-loop cavity feedback. Examples of 2D gratings are given in Figure 7, which shows atomic force micrographs of an eggbox structure, formed from two orthogonal gratings, and a circular DFB grating. Such 2D DFB structures can significantly improve the operation of surface-emitting polymer lasers compared with 1D gratings. Improved feedback yields lower oscillation thresholds, combined with higher slope efficiencies and significantly improved beam quality. Indeed, such surface emitting DFB lasers can produce nearly diffraction limited laser beams.

Typical operating characteristics of an egg-box DFB polymer laser are shown in Figure 8. This laser, which is based on the material MEH-PPV, is transversely pumped with a pulsed pump laser focused onto a small region of the waveguide; up to 90% of the incident pump light is absorbed in the 100 nm thick film. The laser has a threshold pump energy of 4 nJ and an external slope efficiency of >7%, and operates on a single frequency close to the Bragg wavelength of the periodic waveguide. Lasing does not occur exactly at the Bragg wavelength, because at this wavelength the propagation of light is suppressed by the periodic structure. This leads to a photonic stopband within which the emission is reduced. At the edges of this stopband the microstructure can support standing wave electric field patterns of the same spatial period. At these wavelengths there can be a strong interaction between the electric field and excited states in the gain medium, which leads to low-threshold band edge lasing as observed in the figure.

Although DFB resonators have advantages over microcavity devices, the complicated lithographic and etching steps involved in their fabrication can detract from the key material advantage of simple processing. To address this issue, several different techniques for simply replicating the complicated microstructures have been explored. These include UV embossing of flexible plastic substrates, and hot embossing and micromolding of the active

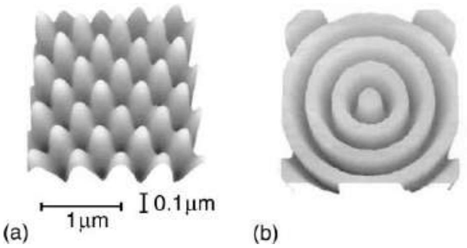

Figure 7 Atomic force microscope images of (a) 'egg-box' DFB grating, and (b) circular DFB grating; grating period is 400 nm in each structure.

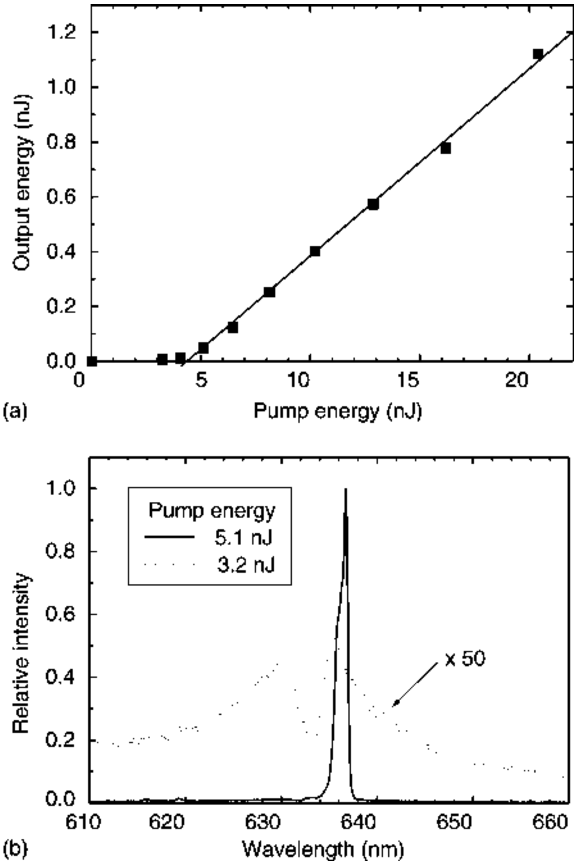

Figure 8 (a) Energy characteristic of an egg-box DFB laser; (b) normalized emission spectra above and below lasing threshold, showing lasing in the edge of the photonic stopband. (Reprinted from Turnbull GA, Andrew P, Barnes WL and Samuel IDW (2002) Operating characteristics of a semiconducting polymer laser pumped by a microchip laser. *Applied Physics Letters* 82: 313–315. Copyright (2003). With permission from American Institute of Physics.)

polymer itself. Such techniques can very readily reproduce structures with feature sizes as small as 100 nm and could allow future mass production of photonic microstructures in these materials.

Towards Plastic Diode Lasers

A typical experimental configuration for characterising optically pumped polymer lasers is shown in Figure 9. The laser is mounted in a vacuum chamber in order to isolate the polymer from oxygen and water. A pulsed visible or ultraviolet pump laser is attenuated and then focused onto the waveguide to transversely pump the polymer laser. The emission is collected using a CCD spectrograph for spectral measurements and energy meter and beam profilers for other characteristics. Typical pulsed pump lasers, which have been used, include nitrogen laser pumped dye lasers, and spectral harmonics of regeneratively amplified Ti^{3+}-sapphire lasers or flashlamp pumped Nd^{3+}:YAG lasers. Recent advances have shown the thresholds to be low enough to be successfully pumped with a modest Nd^{3+} pulsed microchip laser.

So while optically pumped polymer lasers are now proving to be compact, functional light sources, a major motivation for research in this field remains the prospect of electrically driven plastic diode lasers. Such lasers could be very low-cost, flexible light sources that would access any wavelength throughout the visible spectrum. In particular, there are exciting prospects of producing blue and green diode lasers in spectral regions currently difficult to access with inorganic semiconductors. The lowest threshold densities reported for optically pumped organic semiconductor lasers are of the order of 100 W cm^{-2}. To achieve similar excitation densities electrically, one requires pump current densities of ~200 A cm^{-2}. This is substantially more than the few mA cm^{-2} commonly used in polymer LEDs for display applications. Presently, such high current densities are infeasible with continuous excitation, but can be exceeded with short current pulses of a few tens of nanoseconds duration.

Electroluminescence in conjugated polymers was first achieved in 1990 by Burroughes and co-workers. Since that time there has been remarkable progress made in the science and technology of organic LEDs. Commercial products are now on the market, and the feasibility of both flexible and full-color light-emitting displays, in which the individual pixels have been ink-jet printed, have been demonstrated in the laboratory. Despite such rapid progress in the field of organic LEDs, there have been no successful demonstrations to date of electrically pumped lasing in any organic semiconductor. In this section we highlight some of the challenges that must be faced if electrically pumped lasing is to be achieved. To understand the problems associated with electrical driving of organic semiconductor lasers, it is useful first to review the generic structure of organic light-emitting diodes.

A generic organic LED structure is shown in Figure 10. The device consists of one or more organic layers sandwiched between two electrical contacts. Usually a transparent anode is used, commonly made from ITO, which allows light to be emitted from the front face of the LED. A low work-function metal, such as calcium or aluminum, is used as the cathode. The organic part of the device may consist of several layers. These include one or more charge transport layers, plus a light-emitting layer in which the electrons and holes combine to form excitons before emitting light. Typically, the combined thickness of these organic layers is as little as 100 nm. Such thin devices are necessary to keep operating voltages low, to overcome potential barriers for injection, and because charge mobilities in organic semiconductors are very low (10^{-5} to 10^{-2} cm^2 V^{-1} s^{-1}).

The inclusion of electrical contacts into organic semiconductor laser structures can be problematic since they need to be situated very close to the emitting region. Metallic contacts can quench the luminescence and introduce losses to the waveguide mode, thereby increasing the oscillation thresholds. Furthermore, the ITO anode has a high refractive index and so can reduce the spatial overlap of the

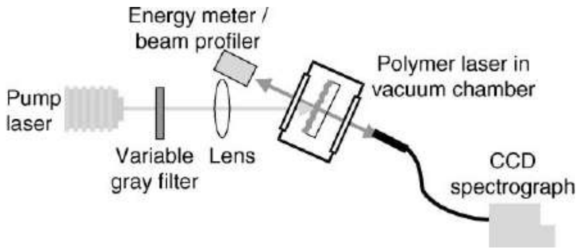

Figure 9 Schematic of a typical experimental configuration for characterizing polymer lasers.

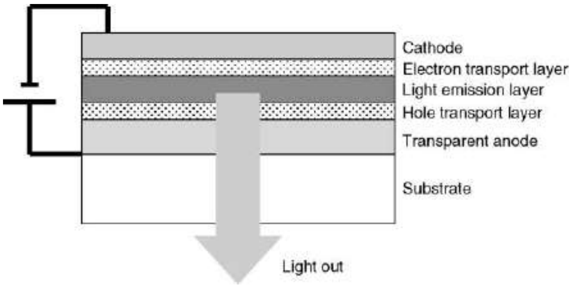

Figure 10 Schematic structure of a generic organic light-emitting diode.

guided mode with the active medium. Such issues indicate that the design of organic LEDs needs to be somewhat modified for achieving electrically pumped lasing. In particular, there needs to be a trade-off between the optimum electronic and photonic designs. To address these issues, various device structures have been proposed that distance the optical mode from metallic contacts, while reducing the thickness of the ITO layer to optimize the mode confinement in the active material.

A more demanding obstacle to electrically pumped lasing is the additional excited state absorption associated with the electrical pumping scheme. As many as three-quarters of the generated excitons in an organic LED are formed in the triplet state which is nonemissive and can lead to excited-state absorption, albeit principally at lower energies than the lasing wavelength. Additionally, injected charges that have not yet formed excitons, exhibit strong absorption, which overlaps significantly with the peak of the material gain. Indeed, this charge-induced absorption has so far prevented the demonstration of injection lasing in conjugated polymers, even in optimized device structures mentioned above. A possible alternative route for efficient pumping of polymer lasers could be through indirect electrical pumping. For example, an inorganic diode laser or LED, or even an organic LED, could be used as an electrically efficient optical pump for the semiconducting polymer laser. Recent advances in GaN-based LEDs and lasers provide interesting prospects for such alternative systems.

Research on semiconducting polymer lasers continues, both to reduce the excitation densities needed for threshold and to investigate new material systems that could overcome the obstacle of charge-induced absorption. By building on recent developments in resonator geometries, material blends, and compact pumping schemes, semiconducting polymer laser systems, whether optically or electrically pumped, should soon become practical, visible light sources for a range of applications.

List of Units and Nomenclature

Charge mobility	[cm ^{2} V^{-1} s^{-1}]
Current density	[A cm^{-2}]
Energy density	[μJ cm^{-2}]
Excited state densities	[cm^{-3}]
Gain coefficient α (exp(αz))	[cm^{-1}]
Intensity/power density	[kW cm^{-2}]
Optical gain	[dB cm^{-1}]
Stimulated emission cross-section	[cm^2]
Waveguide propagation loss coefficient γ (exp($-\gamma z$))	[cm^{-1}]
Wavelength	[nm]

See also

Diffractive Systems: Diffractive Laser Resonators. **Lasers:** Dye Lasers; Planar Waveguide Lasers. **Optical Amplifiers:** Basic Concepts. **Optical Materials:** Plastics. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Burroughes JH, Bradley DDC, Brown AR, *et al.* (1990) Light-emitting-diodes based on conjugated polymers. *Nature* 347: 539–541.

Friend RH, Gymer RW, Holmes AB, *et al.* (1999) Electroluminescence in conjugated polymers. *Nature* 397: 121–128.

Karl N (1972) Laser emission from an organic molecular crystal. *Physica Status Solidi (a)* 13: 651–655.

Kranzelbinder G and Leising G (2000) Organic solid-state lasers. *Reports on Progress in Physics* 63: 729–762.

McGehee MD and Heeger AJ (2000) Semiconducting (conjugated) polymers as materials for solid-state lasers. *Advanced Materials* 12: 1655–1668.

Miyata S and Nalwa HS (eds) (1997) *Organic Electroluminescent Materials and Devices*. Amsterdam: OPA.

Moses D (1992) High quantum efficiency luminescence from a conducting polymer in solution – a novel polymer laser-dye. *Applied Physics Letters* 60: 3215–3216.

Pope M and Swenberg CE (1999) *Electronic Processes in Organic Crystals and Polymers*. New York: Oxford University Press.

Samuel IDW (2000) Polymer electronics. *Philosophical Transactions of the Royal Society of London A* 358: 193–210.

Scherf U, Riechel S, Lemmer U and Mahrt RF (2001) Conjugated polymers: lasing and stimulated emission. *Current Opinion Solid State and Materials Science* 5: 143–154.

Tessler N (1999) Lasers based on semiconducting organic materials. *Advanced Materials* 11: 363–370.

Tessler N, Denton GJ and Friend RH (1996) Lasing from conjugated-polymer microcavities. *Nature* 382: 695–697.

Planar Waveguide Lasers

S Bhandarkar, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lasers have evolved to become true marvels of technology. From their invention in the early 1960s, laser light has changed the way society operates, its application stretching to virtually every aspect of human activity. The versatility of lasers is derived from the fact that there are several different types, each with a unique set of advantages. Solid-state lasers, made from specific semiconducting materials, are commonly used in the optical networks that underlie the greatest machine on Earth – the Internet. High power gas lasers, such as CO_2 lasers, are used in industrial machining, whereas rare-earth doped and dye lasers have found applications ranging from eye surgery to tattoo removal.

Planar waveguide lasers are a relatively new class of lasers made possible by advances in the field of optical waveguides. Over the past decade, with the aim to miniaturize optical components, planar waveguide technology has been developed to a high degree of sophistication. Likewise, Er and other rare-earth doping was studied extensively owing to its importance in all-optical amplification. These elements turned out to be the building blocks for the development of fiber lasers and subsequently planar waveguide lasers.

This chapter will discuss the basic concepts underlying this category of lasers, the variations therein, and the main characteristics. It is important to understand the basic features of waveguides and lasers, which is included as a prelude. We will then discuss planar waveguide lasers with emphasis on the essential elements: the laser materials, the various designs, their performance, and the resultant advantages as well as the applications.

Waveguides

A waveguide is a general conduit for efficient propagation of an electromagnetic field. It is usually defined by making a region of higher refractive index within a dielectric medium. This behaves as an attractive potential well that confines light in the form of bound modes, much like a potential well that confines electrons to a bound state. The propagating field is referred to as the lightwave and is governed by Maxwell's equation. For a weakly guiding approximation, as is the case when the refractive index contrast (Δn) is small, Maxwell's equation can be reduced to a scalar wave equation for the transverse electric field. For a given wavelength (λ) and polarization, the scalar wave equation in a nonmagnetic medium is

$$\nabla^2 E(x) + \frac{4\pi^2}{\lambda^2} n(x)^2 E(x) = 0 \qquad [1]$$

where E is the electric field. For a waveguide aligned in the z direction, the field is the sum of the two propagating transverse modes, each with the form $E(x) = \Phi(x, y)e^{ik_z z}$ where $\Phi(x, y)$ is the field distribution and k_z is the propagation constant. The wave equation for the transverse mode reduces to

$$\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Phi(x, y) + n(x, y)^2\Phi(x, y)$$
$$= (\bar{n})^2\Phi(x, y) \qquad [2]$$

Here, $\bar{n}$ is referred to as the effective refractive index and is given by $\bar{n} = k_z\lambda/2\pi$. The effective index dictates the occurrence of bound modes – these states occur when $\bar{n}$ is between n_{core} and n_{cladding}. For smaller values of $\bar{n}$, there is a continuum of unbound and radiation modes. These are leaky modes that are dispersed quickly over a short distance. Only bound modes are guided in the core of the waveguide structure.

Planar Waveguides

Waveguides formed on a flat substrate are called planar waveguides. These are typically made by stepwise deposition of films of dielectric materials (typically glass). The waveguide core is defined by one of several methods, the most common of which uses lithography and etching. In this case, a film of an appropriately higher index material is deposited on an 'under-cladding' film and then selectively etched to define a core. This is generally covered with a suitable 'upper-cladding' layer so that the index contrast is controlled in all directions. Other methods of forming the core include patterned ion exchange or ion implantation or increasing the index of the medium using a femtosecond laser, etc. In all these cases, the end result is the formation of an embedded higher-index region whose locus defines the optical path for the lightwave. A circuit made from these waveguides is known as planar lightwave circuits (PLC).

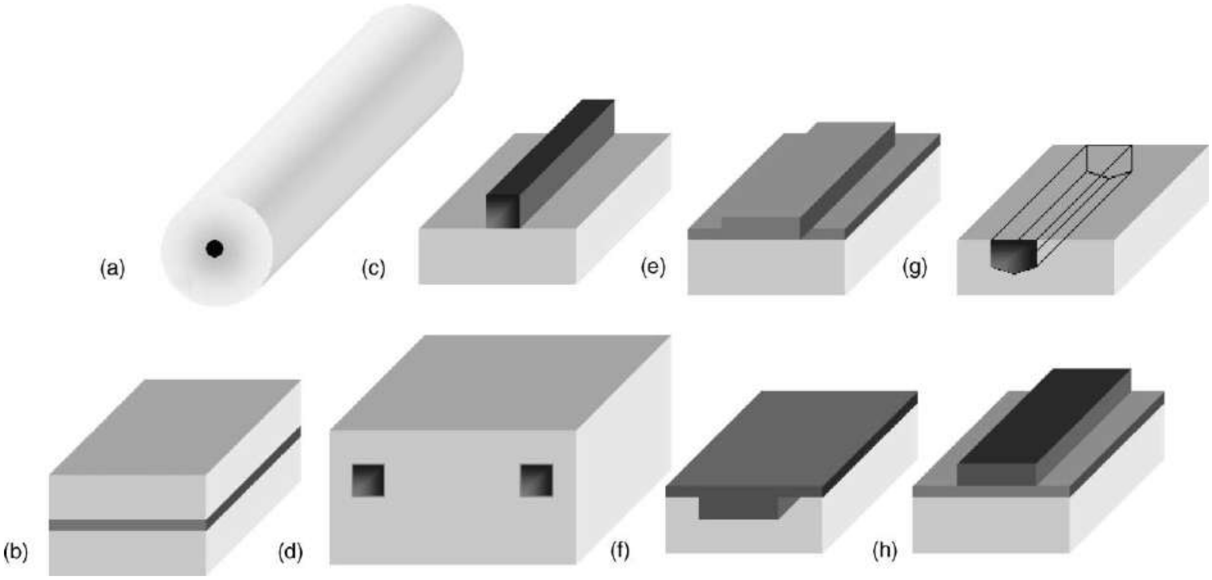

Figure 1 Range of different waveguide designs, with the core represented by the darker regions: (a) Optical fiber; (b) slab waveguide; (c) channel waveguide; (d) embedded channel waveguides with a layer of the overcladding; (e) rib waveguide; (f) inverted rib; (g) ion diffused; and (h)strip loaded, comprising of a strip of a different higher index material (that serves to confine light) on a core layer.

Waveguide cores may be designed in several different configurations, as shown in Figure 1. These designs allow for different fabrication processes as well as the control over the modal properties of the lightwave. For a given index contrast, the number of modes traveling in the waveguide is primarily determined by the dimensions of the waveguide. The aspect ratio of the waveguide affects the propagation of the transverse electric (TE) and the transverse magnetic (TM) modes, due to different propagation constant k_z for each mode.

Lasers

The generation of laser light requires four basic conditions to be satisfied:

- a source capable of light emission;
- an energy supply source – typically electrical, optical, or chemical;
- population inversion to the excited state in the lasing medium; and
- absence of parasitic effects that would absorb the emitted light.

Laser light is often a result of either electronic or vibrational excitation. This is because these excitations involve transfer between energy states that coincide with the radiation of UV, visible, or infrared light. Population inversion is an important requirement in lasers since the emission is stimulated. This involves having a greater fraction of the atoms or molecules in the excited state so that a stimulating photon can cause a radiative decay to the lower energy level. The stimulated light from most lasing media is weak, which requires unacceptably long lengths to generate sufficient power. Instead, as shown in Figure 2, lasers are made with two parallel mirrors with an axial light-generating source in between. This configuration is known as a resonant cavity, a resonator, or an oscillator. The back mirror is usually fully reflective but a partial front mirror allows a portion of the light in the cavity to pass through. The two mirrors cause the light to bounce to and fro, forcing it into making multiple passes through the cavity. Light emanating from this structure is thus amplified and is also coherent (that is the lightwaves are in phase

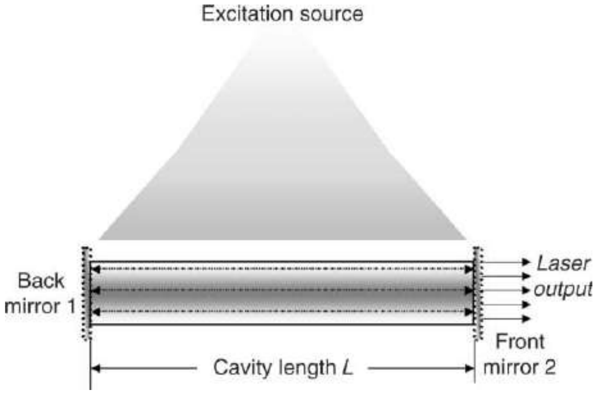

Figure 2 Conceptual representation of a simple Fabry–Perot laser.

with each other spatially and temporally). This has come to represent the signature output of a laser.

Planar Waveguide Lasers (PWL)

By way of an example, let us consider a solid state, rare-earth doped laser, such as an Nd-doped YAG laser. All rare-earth ions are able to fluoresce, some (e.g., Nd) more efficiently than others, and hence can be light emitters. These ions have to be dissolved (doped) into a solid host, say, a glass or a single crystal such as YAG (Yttrium Aluminum Garnet). The excitation of these ions is done optically with light of a suitably higher energy (or lower wavelength). When this 'pump' light is made incident on a rod of the rare earth doped material, it is absorbed by the rare earth ions and used to reach an excited electronic state. Amongst all the possible higher energy states, only certain ones are sustainable because the electron resides in that state for an acceptably long period of time (called the lifetime). Hence, population inversion can be more readily achieved in these longer lifetime states and stimulated processes such as amplification and lasing are readily possible. The laser light generated in the doped rod is emitted across its entire cross-section.

Planar waveguide lasers are designed to miniaturize the kinds of stand-alone solid state lasers describe above. In principle, a PWL is a laser where the light emitting resonant cavity is a planar waveguide. This topic of our interest is a relatively recent development that renders some unique advantages. These can be broadly classified into two categories: one relating to making efficient high-power lasers and the other, to the integration of lasers with other optical components.

An example of the first type is a Nd-YAG laser configured into a planar geometry so that the laser light is now contained within the waveguide. One design that has evolved to achieve this well is shown in Figure 3. The light is guided by the waveguiding structure setup by the sapphire planes. Since both the excitation (or pump) and the lasing modes are well confined, high modal overlap is achieved. These PWLs can be edge-pumped or face-pumped and high powers (measured in the order of Watts) can be achieved.

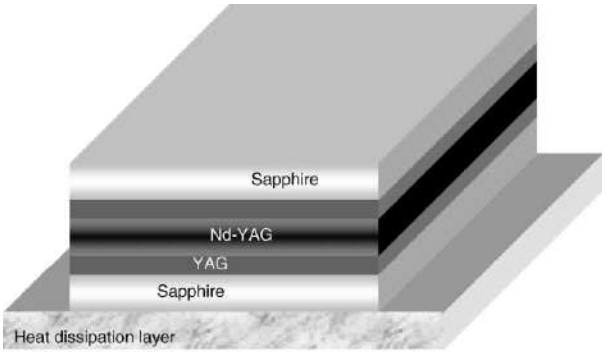

Figure 3 End view of a planar waveguide form of a Nd-YAG laser made using sapphire (single crystal Al_2O_3) for cladding.

In the above design, the cavity is defined by two parallel reflectors bracketing the light-generating medium and is known as a Fabry–Perot etalon. The distance between the two mirrors sets up a wavelength filtering mechanism whereby only resonant wavelengths are sustained. These are related to the intra-mirror distance by the simple equation shown below:

$$L_c = \frac{a\lambda}{2n} \qquad [3]$$

where L_c is the length of the cavity, a is an integer, n is the refractive index of the medium, and λ is the wavelength in vacuum. As the cavity length increases, the number of possible resonant wavelengths increases and the spacing between wavelengths is reduced. These are called the longitudinal modes (Figure 4a). Likewise, lateral modes (Figure 4b) can be generated in the cavity as well, tending to make the operation of the laser multimoded, as shown in Figure 5. This is characteristic of a Fabry–Perot laser, which can be overcome by embellishing the basic etalon design.

As stated above, another big advantage here is that the platform for fabricating PWLs can be chosen to be the same as to make PLCs. Waveguide comprizing the laser can be continued further and configured into other desirable components, such as splitters, couplers, etc. Thus the laser does not have to be separately coupled to the waveguide. This is a major advantage as this connecting process is time-consuming and incurs loss of light. In the field of photonics, the ability to integrate active components with passive ones onto the same chip is considered to be the first step toward photonic chips, analogous to the integrated circuit chip that has revolutionalized our world. This is the driver behind continued research and development of active elements such as PWLs that can be coupled with other actives, such as thin film electro-optic modulators or acousto-optic filters, etc.

Waveguide Laser Materials

The waveguiding structure is comprised of materials that are either dielectrics or semiconductors, both inorganic and organic. For their homogeneity and broad transmittance, high silica glasses are ideal

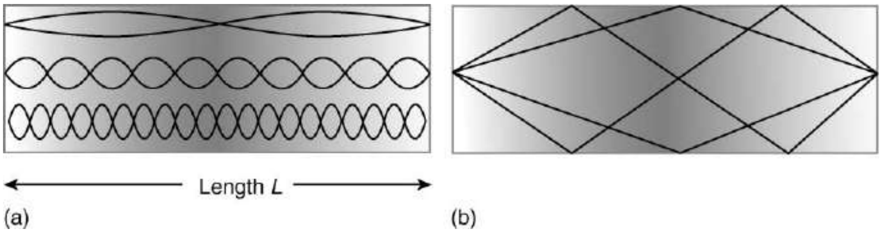

Figure 4 Generation of longitudinal (a) and lateral modes (b) in a Fabry–Perot laser.

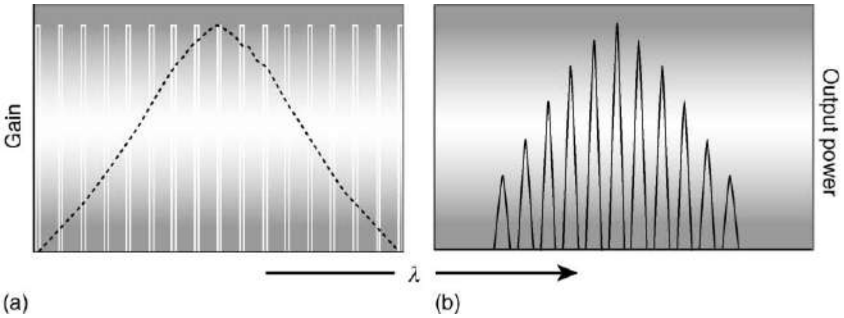

Figure 5 The output of a Fabry–Perot laser: (b) is governed by the number of modes within the gain spectrum of the lasing medium shown by the dotted lines (a). The spacing between the modes can be altered changing the length of the cavity. The smallest mode in output spectrum (b) needs to have enough gain to overcome the lasing threshold.

materials for waveguiding. The chief advantages of high silica waveguides are easy index and mode matching to standard optical fibers, low propagation loss, and good durability. When crystalline materials are used, they have to be single-crystalline to avoid grain boundary scattering (for instance, Ti-diffused waveguide in single-crystal $LiNbO_3$). If semiconductors are used, the bandgap of the material needs to be sufficiently high so as to avoid absorption of the photons. Silicon waveguide (bandgap = 1.1 eV) on a silica cladding (SOI) is commonly used system in PLCs used for telecommunication applications.

The PWLs described above are fabricated in rare-earth doped glass waveguides. Waveguides have to be much shorter than fiber, so the challenge of incorporating the rare-earth ions at high concentration in the host material has to be overcome. Planar waveguide lasers are a compressed version fiber lasers that have the unique advantage that they can be directly integrated with other components on the same chip using very large-scale integration (VLSI)-style processing. The concentration of Er (or the lasing ion in general) in a PWL needs to be many times that in fiber amplifiers (~100 ppm). The need for high gain in a short length requires high pump powers and high lasing ion concentration, both of which lead to nonideal behavior. Since rare earth ions tend to have multiple excited electronic states, there can be absorption at the excited state level as well. When a higher energy state is separated from the upper emission level by the energy in one laser photon, the ions can undergo excited state absorption or ESA (Figure 6). Rare-earth ions in close proximity tend to undergo cooperative upconversion processes, where two excited ions transfer energy between each other so that neither ion may fluoresce at the desired wavelength. These processes can negatively affect the

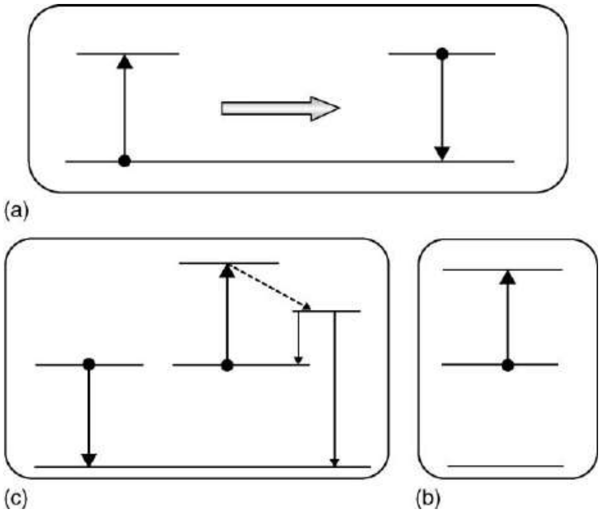

Figure 6 The energy level diagram for a representative two level system. The inset (a) shows the excitation and emission process. Excited state absorption is shown in (b) and cooperative upconversion is shown in (c).

useful gain of the system. Ensuring high spatial dispersion of the Er atoms is critical to minimize this effect. Special glass compositions, such as phosphate glasses, are used here to ensure that the Er is not phase separated or clustered. These glasses also provide a high induced cross-section for gain. Glasses with low phonon energy are also desirable to minimize phonon-mediated nonradiative decays in the rare earths. Light emitting semiconductors and polymers containing fluorescent dyes have also been used to make PWLs.

Now that we have gained an overall perspective of PWLs, we can discuss different manifestations. The following discussion uses a Er doped system as a model since there is a substantial body of work on this system.

Distributed Bragg Reflector (DBR) PWL

In planar waveguide systems, it is difficult to construct efficient mirrors or reflective facets perpendicular to the waveguide. However, mirrors can be replaced by Bragg gratings, which be fabricated relatively easily. In-line Bragg gratings result in wavelength specific reflection as per eqn [4]:

$$\Lambda = \frac{m\lambda}{2\bar{n}} \qquad [4]$$

where Λ is the grating pitch, m is the order of the grating, and λ is the central Bragg wavelength. Gratings have another important benefit in that they ensure that unwanted frequencies outside the limited reflection spectrum are dispersed before they reach the lasing threshold.

A DBR laser is similar in principle to a Fabry–Perot design but uses a Bragg grating on either end of planar Er-doped waveguide. This is shown schematically in Figure 7. Gratings can be made in a number of different ways. One popular way is to exploit the photosensitivity of glasses using 193 nm or 244 nm UV light to periodically alter the refractive index of a glass waveguide. Photosensitivity occurs by a combination of molecular changes (such as the formation of oxygen deficiency sites or color centers) as well as changes in the material strain in the region exposed to the UV light (Figure 7c). Alternatively, gratings can also be formed using etching to make periodic grooves in the waveguide. These are called corrugated or surface relief gratings (Figure 7a and 6b).

The back Grating 1 has a high reflectance of almost 100%, whereas Grating 2 is a weaker grating. The reflectance strength of Grating 2 is determined by taking into account the gain coefficient of the material and the output power required from the laser. The reflection intensity can be altered according to the following equation:

$$R = \tanh^2\left[\frac{\pi L_G \Delta n_{\mathrm{eff}}\, \eta(V)}{\lambda_B}\right] \qquad [5]$$

where L_G is the grating length, Δn_{eff} is the index modulation of the grating, λ_B is the Bragg wavelength, $\eta(V)$ is the confinement factor, a function of the waveguide parameter V that represents the fraction of the integrated mode intensity contained in the core. Increasing the reflectance strength also widens the linewidth of reflected light. The overlap of the spectral width between the back (Grating 1) and the front grating (Grating 2) becomes the effective output window within the gain spectrum of the lasing material (Figure 8).

Not unexpectedly, a DBR structure generates longitudinal and lateral modes. These modes are, however, limited to the overlapping section of the reflection spectrum of the two gratings, as opposed to the entire gain spectrum as in Fabry–Perot lasers (Figure 9). Yet, this can lead to problems such as spectral hole burning and mode hopping, whereby the dominant mode hops from one wavelength to another. The separation between the modes needs to

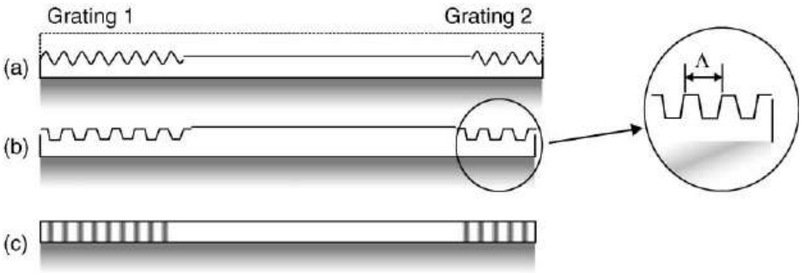

Figure 7 DBR planar waveguide laser. The PWLs shown in (a) and (b) have sinusoidal and trapezoidal surface-relief gratings respectively etched in their cores. An in-line grating of the kind seen in (c) could be fabricated using the photosensitivity of the glass. The top cladding is shown only in (a).

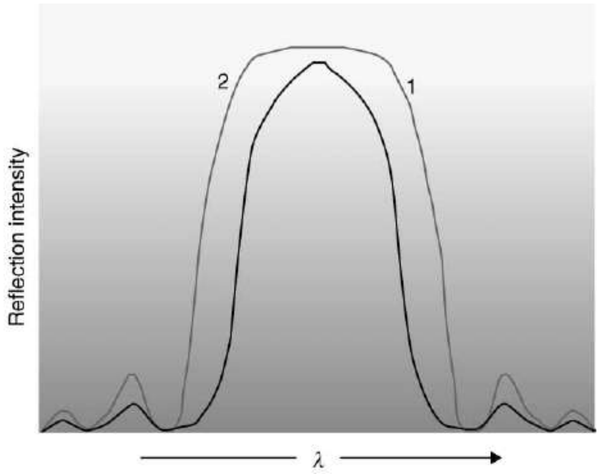

Figure 8 The reflection spectrum of the back (1) and the front (2) gratings. As the intensity of the reflected light is made to increase, so does the linewidth. Hence, the narrower spectrum of partially reflecting grating 2 tends to dictate the operating wavelength range of a DBR.

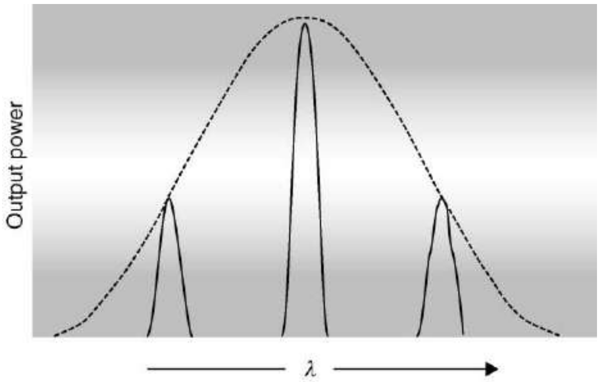

Figure 9 A sample output from a DBR laser, shown in this case with 3 modes. The dotted line shows the effective overlap between the two gratings. The existence of the modes depends on several factors, in particular the pump intensity.

be made sufficiently high compared to the spectral overlap so that single-mode operation can be achieved. But in practice the length of the cavity (L_c in eqn [3]) is determined by factors such as power and space and may not be independently altered to achieve this. However, the reflection spectrum of the back and the front gratings could be slightly shifted in opposite directions to realize a smaller overlap within which only one or a few modes may exist (Figure 10). The occurrence of a single mode also depends on the gain profile of the lasing material and the optical pump power. In Er-doped PWLs, modes other than the primary one can be suppressed by using a pump power below a threshold limit and stable operation can be thus achieved.

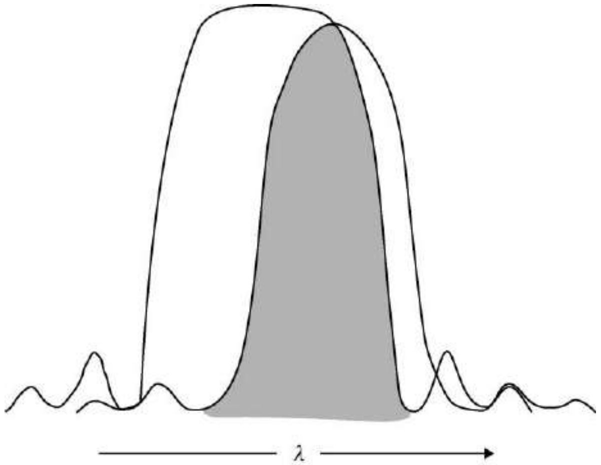

Figure 10 Shifting the front and back gratings is one means of controlling the width of the overlap region, shown in gray. This way a single mode operation can be achieved even for a longer cavity with many modes.

Distributed Feedback (DFB) PWL

The modal problems listed above can be addressed by having a single, distributed grating that is placed along the entire length of the lasing waveguide itself (Figure 11). Alternatively, the gratings may be also placed in the upper or the lower claddings to minimize loss of optical power due to scattering because of the longer grating. In this case, it is the evanescent portion of the light that interacts with the grating. In all cases, the periodicity of the grating leads to a condition for a single lasing wavelength based on constructive interference. The strongest mode occurs when the period Λ of the Bragg grating satisfies the primary Bragg condition (for first order grating or $m = 1$) in eqn [4]. The device will also function if the grating pitch is equal to small integer multiples of ($\lambda/2\bar{n}$).

The grating needs to be made strong enough to generate sufficient feedback (reflections), which also widens the linewidth of the spectrum. To ensure a single narrow linewidth output, a quarter wave phase shift is introduced in the grating, as seen in Figure 11. This phase shift creates a transmission fringe in the spectrum (Figure 12) that significantly narrows the linewidth of the output signal. This design is considered to be standard for the DFB lasers of today. Figure 13 shows a typical output of a DFB laser.

Any of the geometries seen in Figure 1 can be used to make the waveguide for the laser. Due to processing constraints, rib (c) and ion diffused (d) configurations are commonly used. Waveguide losses, characteristically due to scattering from sidewall roughness, have to be low. Otherwise, this adds a

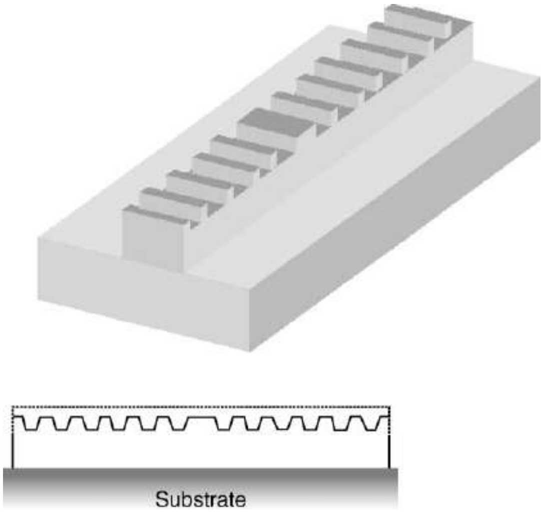

Figure 11 A DFB planar waveguide laser.

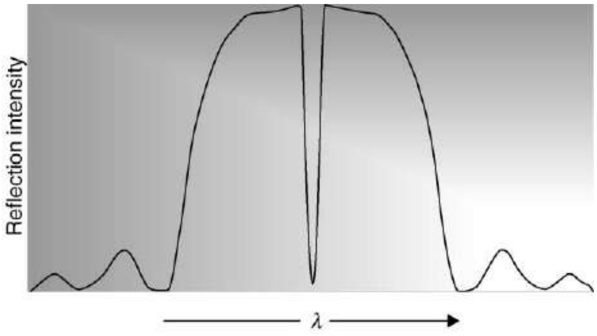

Figure 12 The reflection spectrum of a grating with a central λ/4 shift in the pitch, showing the transmission fringe.

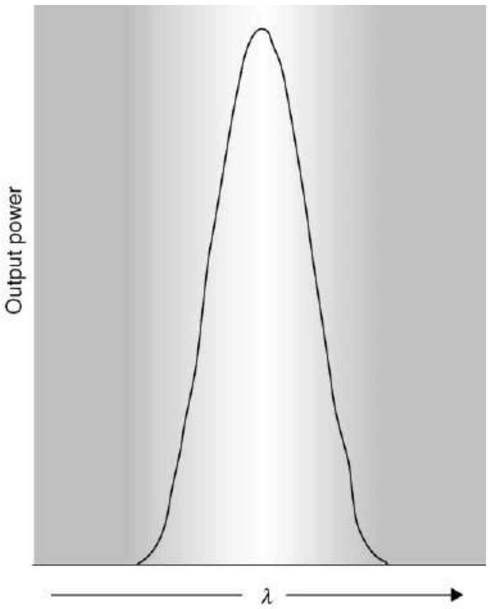

Figure 13 The typical output of a DFB PWL. The linewidth can be made to be as narrow as a few hundred kHz.

significant overhead onto the pump power and lowers the output of the PWL. Birefringence caused by thermal expansion coefficient differences and excessive sidewall roughness also need to be minimized as they can substantially alter the loss of the TE mode compared to TM.

Though their spectral properties are ideal as sources for long haul transmission, DFB lasers tend to be less powerful than the corresponding DBR designs. But the DBR PWLs are very sensitive to reflections that can cause a broadening of the laser linewidth or can act as a source of noise. In both cases, single mode lasers with very narrow linewidths in the kHz range with low radiance and noise, have been demonstrated. But neither is known to have been commercialized yet.

PWL Arrays

One of the unique advantages of the PWLs is that they are often processed using VLSI techniques. One important implication of this is the ability to make a PWL array where multiple lasers are formed simultaneously on the same chip, each with a unique output (Figure 14). Such arrays are likely to be very useful in future dense wave division multiplexing (DWDM) optical network systems where as many as 128 channels (and more) are planned. Since packaging of individual lasers can be up to 75% of the total cost, making chips with multiple lasers is very attractive. The lasers may be DBRs or DFBs whose output wavelength is controlled by changing the grating parameters. To vary the individual wavelengths either the effective index and/or the pitch of the grating can be changed. Thus, any output spectrum can be designed and executed merely by making the appropriate lithography mask.

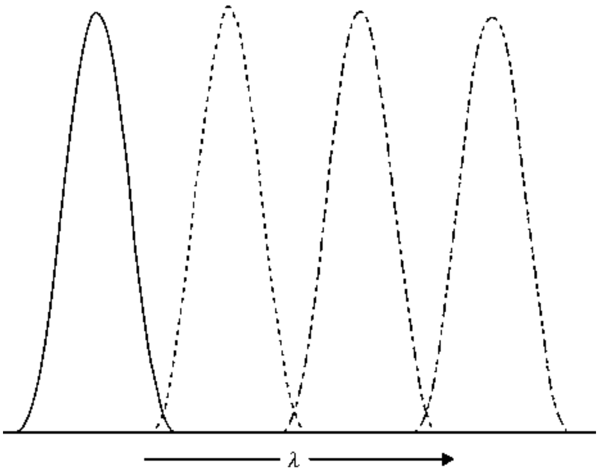

Figure 14 Idealized output from a 4-PWL array.

The gratings are often written holographically, where the wafer is exposed to an interference pattern caused by two coherent light beams. The wafer underneath can be sensitized with a photosensitive glass layer or a film a photoresist. It is not efficient to do multiple holographic exposures to individually change each of the gratings in the array. One way of doing this with a single holographic exposure is by constructing the waveguides at different relative angles. Then the component of the index modulation, resolved along the axis of the waveguide, would vary for each waveguide depending on the relative waveguide angle (Figure 15); this would lead to differences in $\bar{n}$. Alternatively, the waveguide array can be made of the same height but of different widths. This is easily done since waveguide height is usually deposited uniformly but the width is a function of the mask layout. Here the waveguide and the grating are orthogonal to each other and the variation in wavelengths is achieved by the fact that the effective propagation constant for each of the waveguides is different.

Arrays of PWLs have been made on a silicon substrate using deposition of an Er containing glass that is subsequently patterned into waveguide, and ultimately laser, arrays. Another approach has been to start with a phosphate glass as the host and the substrate. A phosphate glass is a superior host for rare-earth ions such as Er since the sensitization efficiency is nearly unity and much high doping levels are possible before concentration quenching effects set in. Waveguides are created by ion-exchange by immersion in a hot bath containing suitable ions (Na for Li, for instance). When starting with a uniformly doped glass, one issue in the assimilation of the laser array with other components is the fact that Er-doped region does not start and end with the waveguide laser section. Recently, the Er has been deposited selectively in the laser waveguide in phosphate glass using methods such as ion implantation, bringing the concept of integration closer to reality.

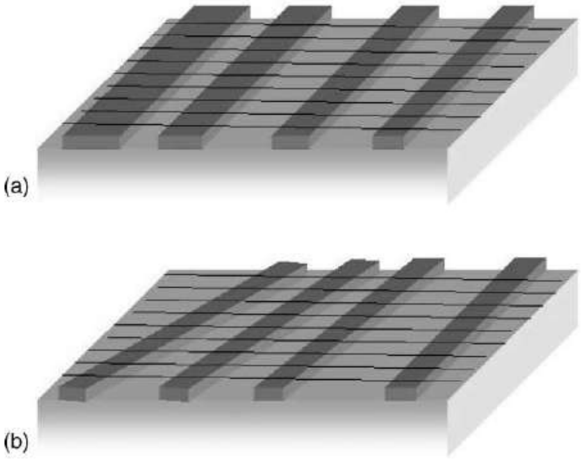

Figure 15 Two different means of making planar waveguide laser array with a single holographic exposure to write the gratings. The darker rectangular blocks represent the core channels or ribs. In (a) the width and hence the effective index of the waveguide is varied whereas the relative angle between the waveguides is altered in (b). The lines represent the direction of the gratings, which is orthogonal only to the left most waveguide in (b).

Pumping

One of the key issues here is the optical pumping of the rare-earth PWLs to generate acceptably high output power. Individual diode lasers or an array of diode lasers all on a single chip (called a diode bar) have been utilized as pumps. Both single and multimode diode lasers are available and have been used. There are pros and cons for either choice: whereas single mode lasers are less powerful (in the range of few 100 mW), it is difficult to convert the multimoded output of their counterparts to useful power. The choice is determined by the waveguide design, which needs to be optimized for high overlap of the lasing mode with the pump. Carefully tapered waveguides can be used distribute the pump power from a diode bar to the various lasers in an array without significant loss.

PWL Stability

Stable operation of a PWL is one of the big challenges. In DBR PWLs, it is seen that there is a threshold for pump power only below which single moded operation can be achieved. Additionally, any PWL die needs to be made immune to ambient temperature changes. These changes can affect many of the critical output parameters of the PWL: wavelength, linewidth, power and in the case of an array, the wavelength separation. This occurs primarily due to the thermal shift in the grating pitch as well as the thermo-optic effect. It is overcome by supplanting the laser die onto a heat spreader chip made from conducting material such as Cu or BeO, followed by a thermoelectric cooler. Ironically, the laser's output wavelength and linewidth are fine-tuned and 'fixed' using temperature to overcome any drifts due to processing errors or other reasons.

Reliability is another important issue that needs to be addressed, especially for telecommunication applications where these types of components have to pass high quality standards such as the Telecordia qualification in the USA.

Other PWL Systems

Many of the PWLs to date have been developed using Er doping, which emits light in the range of interest for telecommunications (the C-band and even the L-band). Here, Yb can be added as a co-dopant to improve power conversion efficiency. Other fluorescing ions have also been used: Nd, Tm, Ho, and Cr. Several research teams have successfully demonstrated Nd-YAG PWLs, in particular. Ridge waveguide lasers have been made using semiconductor materials InGaAs/InP heterojunctions in both Fabry–Perot and DFB configurations. These systems are electrically pumped. DFB PWLs have also been made using dye doped polymers.

It must be mentioned that there are other ways to use waveguides to make lasers. The primary example is a configuration where the waveguide resonator is external to the active region. Designs, such as ring resonator lasers and an external cavity lasers, fall in this category.

Future research and development of PWLs would undoubtedly utilize photonic crystal structures. These structures have a periodic fluctuation in the index of refraction in two or three dimensions. Owing to the resultant Bragg reflections, specific allowed and forbidden states are created which can be used to control the confinement of light. Hitherto, these structures have been used to form mirrors or hexagonal ring resonators. These more sophisticated laser designs are likely to enable new functions such as tunability of the output spectrum.

Summary

Planar waveguide lasers are a special class of laser where light is confined to a waveguide. They have distinctive advantages that include high optical gains, low laser thresholds, narrow linewidths in the kHz range, and optimal thermal management. Significantly, they afford a platform that can be readily expanded to include an array of lasers, each with an unique output, and can ultimately be combined with multiple optical components on the same chip. Such compact integrated devices are bound to take photonics to a new high level of capability.

List of Units and Nomenclature

a	integer denoting number of laser modes
E	electric field
k_z	propagation constant for travel in z direction
L_c	length of laser cavity
L_G	length of grating
m	order of the grating, an integer
n	refractive index of the medium
$\bar{n}$	modal effective refractive index
R	reflectance
V	normalized frequency or the V parameter of the waveguide
Δn_{eff}	modulation of the effective index of the grating
η	confinement factor or fraction of the light confined in the region of interest
λ	wavelength of the lightwave
λ_B	resonant wavelength of a Bragg grating or Bragg wavelength
Λ	grating pitch
Φ	electric field distribution

See also

Lasers: Carbon Dioxide Lasers; Dye Lasers. **Optical Amplifiers:** Erbium Doped Fiber Amplifiers for Lightwave Systems.

Further Reading

Carroll JE, Whiteaway JEA, Plumb RGS and Plumb D (1998) *Distributed Feedback Semiconductor Lasers*. London: IEE Publishing.

Ghafouri-Shiraz H (2003) *Principles of Semiconductor Laser Diodes and Amplifiers: Analysis and Transmission Line Laser*. London: Imperial College.

Hecht J (2001) *Understanding Lasers: An Entry Level Guide*. New York: Wiley Interscience.

Lee JR, Baker HJ, Friel JG, Hilton GJ and Hall DR (2002) High-average-power Nd:YAG planar waveguide laser, face-pumped by 10 laser diode bars. *Optical Letters* 27: 524–526.

Milonni PW and Eberly JH (1988) *Lasers*. New York: John Wiley and Sons.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley & Sons.

Shepherd DP, Hettrick SJ, Li C, *et al.* (2001) High-power planar dielectric waveguide lasers. *Journal of Physics D: Applied Physics* 34: 2420–2432.

Siegman AE (1986) *Lasers*. Harendon, Virginia: University Science Books.

Svelto O (ed.) (1998) *The Principles of Lasers*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Veasey DL, Funk DS and Sanford A (1999) Arrays of distributed-Bragg-reflector waveguide lasers at 1536 nm in Yb/Er codoped phosphate glass. *Applied Physics Letters* 74(6): 789–791.

Young M (2001) *Optics and Lasers: Including Fibers and Optical Waveguides*. NewYork: Springer-Verlag New York, Inc.

Semiconductor Lasers

S W Koch, Philipps-University, Marburg, Germany

M R Hofmann, Ruhr-Universität Bochum, Bochum, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article discusses the principle of operation and the basic material engineering aspects for the design of semiconductor lasers. The physics of the gain medium is analyzed in detail and the role of many-body effects due to the Coulomb interaction of the carriers for the optical properties of the active semiconductor material is demonstrated. The two main types of laser resonators for diode lasers, Fabry–Perot resonators and vertical cavity structures, are described. Finally, the dynamics of semiconductor lasers, including the principles of short pulse generation and nonequilibrium gain dynamics, is addressed.

Basic Principles

The optical emission from semiconductor lasers arises from the radiative recombination of charge carrier pairs, i.e., electrons and holes in the active area of the device. The conduction-band electron fills the valence-band hole by simultaneously transferring the energy difference to the light field. This is called radiative recombination. Hence, the frequency of the laser light is mainly determined by the bandgap of the semiconductor laser material. In order to achieve lasing, one needs carrier inversion, i.e., a sufficient number of electrons in the conduction band so that the probability of light absorption is lower than the probability of emission.

It has been shown that a large variety of semiconductor materials is suited for semiconductor lasers. The most stringent requirement for the material is that it has a direct bandgap. In other words, the maximum of the valence band and the minimum of the conduction band have to be at the same position in *k*-space (in the center of the Brillouin zone). The elemental semiconductors silicon and germanium do not fulfill this condition and therefore cannot be used for semiconductor lasers.

In order to achieve carrier inversion, it is necessary to pump the semiconductor laser, i.e., to excite a sufficient number of electrons from the valence band into the conduction band. Even though this pumping in semiconductors, as in other solid state laser materials, can be done optically, the large technological impact of semiconductor lasers is at least in part due to the possibility of electrical pumping with a few tens of milli-amperes at low voltages.

The original version of such a semiconductor laser device is shown in Figure 1, The laser structure consists of a GaAs-based p-n junction which is biased in the forward direction. Electrons are injected from the n region and holes from the p region, respectively. Due to the diode characteristics of the p-n junction, semiconductor lasers are often also called laser diodes. In the device shown in Figure 1, the laser mirrors are formed by the cleaved facets of the semiconductor crystal. Because of the high refractive index of GaAs, the semiconductor–air interface provides a reflectivity of about 32% which is sufficient for laser emission.

However, the early type of laser, the so-called homojunction device, was not very efficient because of the low carrier density in the active area and because of the weak overlap between the inverted region and the optical mode. Considerable improvement towards room-temperature continuous wave (cw) operation was achieved by the introduction of the so-called double heterostructure. Its principle is shown in Figure 2.

The basic idea behind the double heterostructure laser is that the active material (e.g., GaAs) is embedded in another semiconductor with a slightly larger bandgap (e.g., (AlGa)As). This causes a potential well in which electrons and holes are confined in order to achieve high carrier densities in the active area. Moreover, the smaller refractive index of the surrounding high-bandgap material helps to guide the optical field within the region of highest inversion. Additional ridge-shaped lateral structuring is frequently used to form a complete optical waveguide.

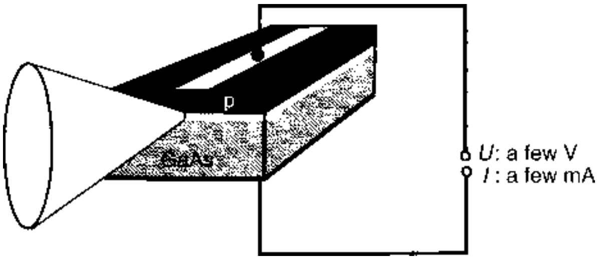

Figure 1 Principle of operation of a semiconductor laser.

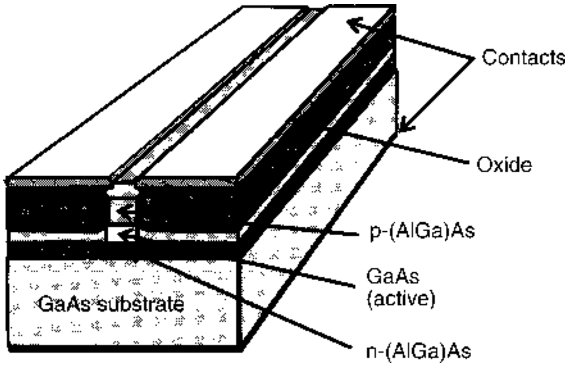

Figure 2 Double heterostructure laser diode.

Nowadays, even more advanced laser structures can be fabricated, as modern crystal growth techniques such as molecular beam epitaxy (MBE) and metal organic chemical vapor deposition (MOCVD) have become available. These techniques allow the grower not only to determine the composition of the semiconductors with remarkable precision, but also to define the shape virtually on an atomic scale. In particular, it is now possible to grow microstructures so small that their electronic and optical properties deviate substantially from those of the corresponding bulk materials. So-called quantum-well (QW) structures, in particular, turned out to be superior to bulk materials in many respects and are currently the active medium in most commercial semiconductor lasers. The gain medium in QW laser structures consists of one or more films with a thickness of a few atomic monolayers. These are embedded in a barrier material with a larger bandgap that serves to confine the carriers to the wells. In QWs the quantum confinement restricts the free carrier motion to the two dimensions within the plane of the films, since the well thickness is comparable to or less than the thermal wavelength of the carriers. Hence, one speaks of quasi-two-dimensional structures where no free motion perpendicular to the films is possible. The corresponding energy states of the carriers are quantized, i.e., only discrete values are allowed for that part of the carrier kinetic energy which is related to the confined direction.

A major consequence of this restriction of the carrier motion to two dimensions is a change in the (single particle) density of states. This density of states is altered in QWs in comparison to bulk material. The density of states for bulk material is zero at the bandgap energy and rises with the square root of the energy for higher energies. In contrast, the density of states for a two-dimensional carrier system in a QW is a step-like function with a density of states different from zero at the QW-well bandgap (the bandgap energy of the active material plus the quantization energies of electrons and holes in the well). This increase of the density of states at the bandgap promises important advantages of QW structures over bulk material. Most importantly, the number of states that have to be occupied before inversion is obtained is considerably reduced in QW structures. This reduces the threshold current and increases the modulation speed of the laser structures. Accordingly, most of the modern laser structures are QW structures.

These arguments have encouraged many researchers to develop structures with further increased carrier confinement in two dimensions (quantum wires) and three dimensions (quantum dots). The single-particle density of states is even more favorable in quantum wires and quantum dots than in QWs: in both cases the single-particle density of states peaks at the effective bandgap energy. This argument seems to promise even better performance of quantum wire and quantum dot lasers as compared to QW devices. However, the single-particle density of states description completely neglects the Coulomb interaction of the carriers which strongly influences the optical and electronic properties of semiconductor structures. The absorption is modified by excitonic effects that lead to additional resonances below the bandgap and to changes of the continuum absorption. These changes are often quantified by the so-called Sommerfeld factor, which depends on the dimensionality of the system, i.e., it differs for bulk, QW, quantum wire and quantum dot structures. For example, the Sommerfeld factor completely suppresses the singularity of the density of states at the bandgap of quantum wire structures. Accordingly, with inclusion of Coulomb effects, the benefit of the quantum confinement is much less than expected from the single-particle density of states.

The use of quantum wire and quantum dot structures as a semiconductor laser gain medium is still in its infancy even though they can be grown with various degrees of precision. However, quantum dot structures, especially, hold great promise for the future, since they can be grown in large numbers using techniques similar to those already used for QWs. The possibility of layer growth under strained conditions, such as QWs between barrier material with a slightly different atomic lattice constant, also results in a greater flexibility to design the laser gain medium with the desired emission wavelength.

In principle, the emission wavelength of a semiconductor laser can be roughly designed by the choice of the material. Different semiconductors have different bandgaps between the valence band and conduction band and the bandgap energy determines the spectral range of the emission. One can tune the bandgap and thus the spectral range of the emission by fabricating semiconductor alloys as, e.g., (GaIn)As or (AlGa)As. Presently, a large variety of different materials is commercially available covering, for example, the red ((AlGaIn)P), the near infrared ((AlGa)As, (GaIn)As), the telecom ((GaIn)(AsP)) and even the blue–ultraviolet range ((GaIn)N).

Great care has to be taken that the crystal used for a semiconductor laser is of almost perfect quality. Dislocations in the crystal, for example, have to be avoided as well as any kinds of defects since they act as nonradiative recombination centers. Nonradiative recombination competes with the radiative recombination that is essential for the laser process. Nonradiative recombination effects include Auger recombination and recombination via defects. Auger recombination is the process where an electron and a hole recombine and completely transfer the energy difference to another carrier (electron or hole) which is excited to a higher state. This is an intrinsic effect that depends mainly on the band structure and the relevant values of the transition probabilities (transition matrix elements). The probability for Auger recombination is temperature- and carrier-density-dependent and can be minimized, e.g., by properly cooling the device and by keeping the carrier densities as low as possible.

Nonradiative recombination via defects is an extrinsic effect that depends on the material quality and can thus be minimized by optimized growth. However, trying to avoid dislocations limits the choice of materials for laser structures severely. The growth processes require ‘host’ crystals, so-called substrates, on which the real structure can be grown. GaAs, InP, SiC and sapphire wafers are available in sufficient quality as standard substrates for optoelectronic applications. The epitaxially grown material adapts to the crystal structure of the substrate which implies that the lattice constant of the substrate and structure material should be very similar. If the lattice constants do not agree, i.e., if there is a structural mismatch, a certain amount of strain develops which increases with increasing lattice mismatch. Too much strain leads to the formation of dislocations which are detrimental to laser applications. Control of the dislocation density is currently one of the critical issues for the wide-gap III–V materials, such as GaN, because no substrate is available that matches the GaN lattice constant. However, a small amount of strain is often even desirable as long as it does not cause dislocations. Strain influences the band structure of the material in a predictable way and can therefore be used to tailor the laser emission properties.

Thus laser materials for a certain desired wavelength range have to fulfill two important requirements: first, the bandgap of the material has to fit the desired photon energy; and second, and more crucial, an appropriate substrate with the right lattice constant has to be available. This second condition excludes many material compositions from laser applications. In many cases, both conditions can only be simultaneously met with quaternary semiconductor alloys (for example (GaIn)(AsP) on InP substrates).

Physics of the Gain Medium

From a material physics point of view, the important structural aspects of a certain laser material are summarized in the band structure of the gain medium, i.e., in the quantum mechanically allowed energy states of the material electrons. The calculation of the band structure of a particular gain medium is therefore a crucial aspect of semiconductor laser modelling. Additionally, however, one has to understand the relevant properties of the excited electron–hole plasma in the active region of the laser.

In its ground state, i.e., without excitation and at very low temperatures, a perfect semiconductor is an insulator where no electrons are in the conduction-band states. The presence of a population inversion, i.e., occupied conduction band and empty valence band states in a certain range of frequencies, is, however, a requirement for obtaining optical gain and light amplification from a semiconductor. Sufficiently strong pumping leads to the generation of an electron–hole plasma, where the term ‘electron’ specifically refers to conduction-band electrons and ‘holes’ to the missing electrons in the originally full valence bands. Holes are also quasiparticles, just as the electrons. Electrons in a solid are quasiparticles with an effective mass that is considerably different from the free-electron mass in vacuum and that is determined by the band structure, i.e., the interaction with all the ions of the solid. The properties of the missing electrons, such as a positive charge for the missing negative electron charge, a spin opposite to that of the missing electron, etc., are attributed to the holes. As a consequence of their electrical charges, the excited electrons and holes interact via the Coulomb potential which is repulsive for equal charges (electron–electron, hole–hole) and attractive for opposite charges (electron–hole). Furthermore, electrons and holes obey Fermi–Dirac statistics and the Pauli exclusion principle that forbids two Fermions from occupying the same quantum state. The consequences of the Pauli principle are often referred to as ‘band-filling effects’ in the physics of semiconductor lasers.

The theory therefore has to account not only for the band-structure but also for this band filling, i.e., the detailed population of the electron and hole states. Taking into account only the band structure and band filling leads to a so-called free-carrier theory. But the electron–hole plasma in a semiconductor gain medium is an interacting many-body system because of the carrier interactions and the exclusion principle. In a many-body system the properties of any one carrier are influenced by all the other carriers in their respective quantum states. Generally, advanced many-body techniques are needed to systematically analyze such an interacting electron–hole plasma and to compute the resulting optical properties, such as gain, absorption or refractive index, all of which are changing in a characteristic way with the changing electron–hole density.

It is often helpful to investigate the characteristic time-scales and the most direct consequences of the various interaction effects even though the consequences of the different many-body effects are in general not additive. The fastest interaction, under typical laser conditions, is the carrier–carrier Coulomb scattering that acts on scales of femto- to picoseconds to establish Fermi–Dirac distributions of the carriers within their bands. These distributions are often referred to as 'quasi-equilibrium distributions' since they describe electrons and holes in the conduction and valence bands which may be characterized by an 'electronic temperature' or 'plasma temperature.' The plasma temperature is a measure for the mean kinetic energy of the respective carrier ensemble and relaxes towards the lattice temperature as a consequence of electron–phonon (= quantized lattice vibration) coupling, i.e., the carriers can emit or absorb phonons. The coupling to longitudinal optical phonons in polar materials is especially strong with scattering times in the picosecond range. The corresponding times for the electron–acoustic phonon scattering are in the nanosecond range. Generally, the carrier–phonon interaction provides an exchange of kinetic energy of the carriers with the crystal lattice in addition to the relaxation of the initial nonequilibrium distribution. Deviations between plasma and lattice temperature may occur, especially when lasers are subject to strong pumping or rapid dynamic changes.

The Coulomb interaction not only causes rapid carrier scattering but also directly influences the laser gain spectrum. Gain, i.e., negative absorption, occurs in the spectral region of population inversion, in other words, where the net probability for a test photon to be absorbed is lower than the probability to be amplified by stimulated electron–hole recombination. Energetically, the gain region is bounded from below by the effective (or renormalized) semiconductor bandgap and from above by the electron–hole chemical potential. Here, the chemical potential is the energy at which neither absorption nor amplification occurs, i.e., where the semiconductor medium is effectively transparent.

The effective bandgap energy for elevated carrier densities is significantly below the fundamental absorption edge of an unexcited semiconductor. This 'bandgap renormalization' is a consequence of the fact that the Pauli exclusion principle reduces the repulsive Coulomb interaction between carriers of equal charge. Furthermore, the Coulomb screening, i.e., the density-dependent weakening of the effective Coulomb interaction potential, contributes to the reduction of the bandgap.

Another many-body effect originates in the Coulomb attraction between an electron and a hole, and leads to an excitonic or Coulomb enhancement of the interband transition probability. All these effects contribute to the detailed characteristics of the optical spectra of a semicondcutor gain medium. Figure 3 shows an example of calculated spectra in comparison to experimental data.

The example demonstrates that the microscopic theory correctly accounts for the changes in the gain spectrum with changing carrier density by means of a consistent treatment of band structure, band-filling, and many-body Coulomb effects.

The Resonator

In addition to a gain medium, a laser needs a positive feedback process for the amplified stimulated emission. Hence, the inverted medium has to be

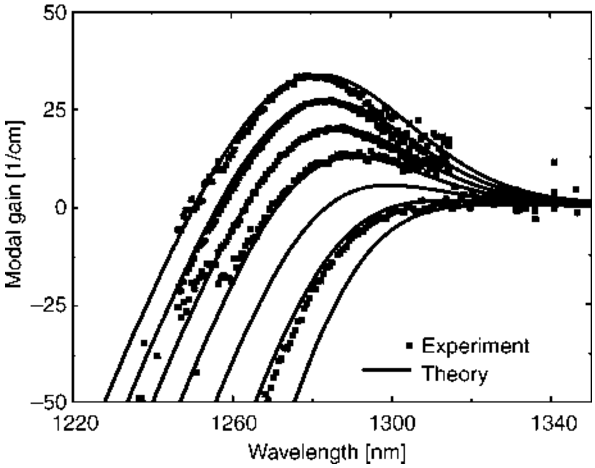

Figure 3 Comparison of experimental and theoretical gain spectra of a (GaIn)(Nas)/GaAs laser.

embedded into a resonator. One of the conceptually simplest versions of such a resonator is a so-called Fabry–Perot resonator which consists of two parallel plane mirrors. A double heterostructure Fabry–Perot laser is shown in Figure 2. Such Fabry–Perot resonators are very easy to fabricate for semiconductor lasers: the semiconductor crystal is cleaved perpendicular to the optical waveguide. The cleaved edges provide a reflectivity of about 32% due to the high refractive index of the semiconductor. This is sufficient to provide enough feedback for laser operation.

However, despite its conceptual simplicity, the cleaved-edge Fabry–Perot concept also introduces a number of problems. First, the light output occurs on two sides of the laser, whereas for most applications all the output power should be emitted out of one facet only. For practical applications, this problem can be solved by depositing coatings onto the cleaved facets of the structure. For example, deposition of a 10% reflectivity coating on one facet and of a 90% reflectivity coating on the other facet concentrates the laser emission almost completely to the facet of lower reflectivity. Antireflection coatings with reflectivities below 10^{-4} are used when laser diodes are coupled to external cavities, e.g., in tunable lasers.

A second problem with Fabry–Perot lasers is that the emission wavelength is often not well defined. A Fabry–Perot resonator of length L (optical length nL) has transmission maxima – so-called modes – at all multiples of $c/2nL$ where c is the speed of light and n the refractive index. That means that for typical devices with a length of 500 μm, hundreds of modes are present even within the limited gain bandwidth of a semiconductor. These modes often compete with each other resulting in multimode emission which may become particularly complex and uncontrollable when the laser is modulated for high-speed operation.

This problem can be overcome in so-called distributed feedback (DFB) lasers. In these devices, the refractive index is periodically modulated along the waveguide. In a simplified picture, this periodic pattern causes multiple reflections at different locations within the waveguide. For certain wavelengths, these multiple reflections interfere constructively to provide enough 'distributed' feedback for laser operation. With the DFB concept, single-mode emission can be realized even for high-speed modulation of the lasers. Alternatively, the region of distributed feedback can be separated from the region of amplification in so-called distributed Bragg reflector (DBR) lasers. These are devices with multiple sections where one section is responsible for the optical gain and another section (the DBR section) is responsible for the optical feedback.

So far, we have discussed so-called edge emitting lasers where the light emission is in the plane of the active area of the device. The lengths of these lasers are typically hundreds of microns and thus correspond to a few hundreds of optical wavelengths. This causes one of the major disadvantages of edge emitters – the multimode emission. Another disadvantage is the poor elliptic beam profile due to diffraction at the small rectangular output aperture. Finally, the fabrication procedure of edge emitters is complex since it requires multiple steps of cleaving before it is even possible to test the laser.

A novel alternative semiconductor laser concept was realized in the early 1990s: the vertical-cavity surface-emitting laser (VCSEL). A typical VCSEL structure is shown in Figure 4. In VCSELs, the light emission is parallel to the growth direction, i.e. perpendicular to the epitaxial layers of the structure. The cavity length of a VCSEL is in the order of a few (1–5) multiples of half the wavelength of the emitted light in the material. These small dimensions imply that the interaction length between the active medium and the light field in the resonator is very short. Therefore, the mirrors of a VCSEL have to be of very good quality; reflectivities of more than 99% are required in most configurations.

Very high reflectivities can be achieved with dielectric multilayer structures, so-called Bragg reflectors. Bragg reflectors consist of a series of pairs of layers of different refractive indices, where each layer has the thickness of a quarter of the emission wavelength. Most preferably, Bragg reflectors are epitaxially grown together with the active area of the device in only one growth process.

Small VCSEL resonators with Bragg reflectors at both ends of the cavity are called microcavities. They are designed in such a way that only one longitudinal mode is present in the region of the semiconductor gain spectrum. Accordingly, VCSELs are well suited to provide single-mode characteristics. Moreover, a round aperture for light emission can be realized with a diameter of a few micrometers so that

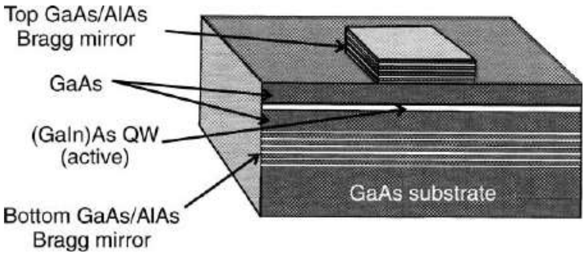

Figure 4 Vertical-cavity surface-emitting laser (VCSEL).

diffraction is weak and the beam profile is almost perfect.

VCSELs can be tested on wafer without further processing and can be arranged in two-dimensional arrays. A major disadvantage of VCSELs is their temperature dependence. The bandgap of the active material and the cavity mode depend on temperature in such a way that the cavity mode shifts away from the region of maximum gain when the temperature is varied. This severely limits the temperature range of operation for VCSELs.

Semiconductor Laser Dynamics

The dynamical behavior of semiconductor lasers is rather complex because of the coupled electron–hole and light dynamics in the active material. Like other laser materials, semiconductor lasers exhibit a typical threshold behavior when the pump intensity is increased. For weak pumping, the emission is low and spectrally broad; the device operates below threshold and shows (amplified) spontaneous emission. At threshold, the optical gain compensates the losses in the resonator (i.e., internal losses and mirror losses) and lasing action starts. Ideally the output power increases linearly with the injection current above threshold. The derivative of this curve is proportional to the differential quantum efficiency. The differential quantum efficiency can reach values close to unity and even the overall efficiency of the diode laser, i.e., the conversion rate from electrical power into optical power, can be as high as 60% which is far higher than for any other laser material.

Ideally, the laser emission is single mode above threshold. However, as discussed above, edge-emitting diode lasers tend to emit on multiple longitudinal modes and the competition of these modes can lead to complex dynamical behavior which is significantly enhanced as soon as the laser receives small amounts of optical feedback. Semiconductor lasers are extremely sensitive to feedback and tend to show dynamically unstable, often even chaotic, behavior under certain feedback conditions.

Like other lasers, semiconductor lasers respond with relaxation oscillations when they are suddenly switched on or disturbed in stationary operation. Relaxation oscillations are the damped periodic response of the total carrier density and the optical output intensity as a dynamically coupled system. Both are important for many aspects of semiconductor laser dynamics. The relaxation oscillation frequency determines the maximum modulation frequency of diode lasers. This frequency is strongly governed by the differential gain in the active material, i.e., by the variation of the optical gain with carrier density.

An important application of relaxation oscillations is short pulse generation by gain switching of laser diodes. The idea of gain switching is to inject a short and intense current pulse or an optical pump pulse into the laser structure. This pulse initiates relaxation oscillations but it is so short that already the second oscillation maximum is not amplified. As a consequence, only one intense pulse is emitted. The shortest pulse widths achieved with this scheme so far are in the few picoseconds range. For short pulse generation, a further complexity of the diode laser comes into play. The emitted pulses are observed to be strongly chirped, i.e., the center frequency changes during the pulse. This complex self-phase-modulation is a result of the strong coupling of the real and the imaginary parts of the susceptibility in the semiconductor. If the gain (which is basically given by the imaginary part of the susceptibility) changes during a pulse emission, the refractive index (basically the real part of the susceptibilty) also changes considerably. This leads to a self-phase-modulation if a strong optical pulse is emitted. This strong phase–amplitude coupling is also responsible for the high feedback sensitivity of diode lasers, for complex spatio-temporal dynamics (e.g., self-focusing, filamentation), and for a considerable broadening of the emission linewidth. The phase–amplitude coupling is often quantified with the so-called linewidth enhancement or α factor. This parameter, however, is a function of carrier density, photon wavelength, and temperature.

Diode lasers are also suited for the generation or amplification of extremely short pulses with durations below 1 picosecond. The concept used for such extremely short pulse generation is called mode-locking. The idea of mode-locking is to synchronize the longitudinal modes of the laser so that their superposition is a sequence of short pulses. The most successful concept to achieve mode-locking of diode lasers is to introduce a saturable absorber into the laser cavity and to additionally modulate the injection current synchronously to the cavity roundtrip frequency. A saturable absorber exhibits nonlinear absorption characteristics: the absorption is high for weak intensities and weak for high intensities so that high peak power pulses are supported. Such absorbers can be integrated as a separate section into multiple-section diode lasers which have already been successfully used for subpicosecond pulse generation. Furthermore, electro-absorption modulators, passive waveguide sections, and tunable DBR segments for wavelength tuning can also be integrated into such multiple-section devices.

However, when subpicosecond pulses are generated or amplified in semiconductor laser amplifiers the dynamics of the pulse generation or pulse amplification is strongly governed by complex nonequilibrium carrier dynamics in the semiconductor. The most prominent effects are spectral hole burning and carrier heating. Spectral hole burning occurs if an intense laser field removes carriers from a certain area in *k*-space faster than carrier–carrier scattering can compensate for. This leads to nonthermal carrier distributions and a gain suppression in certain energy regimes. As mentioned above, a nonthermal carrier distribution relaxes quickly (100 femtoseconds) into a thermal (quasi-equilibrium) distribution if the intense laser field is switched off, but this carrier distribution may have a plasma temperature much higher than the lattice temperature. This effect, which is referred to as carrier heating, also leads to a transient reduction of the gain on a few picoseconds timescale until the carrier distributions have cooled down to the lattice temperature by interactions with phonons. These ultrafast effects are of course particularly important for amplification and generation of short pulses but they generally influence the whole dynamical behavior of diode lasers including, e.g., the modulation behavior.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Quantum Optics:** Entanglement and Quantum Information.

Further Reading

Chow WW and Koch SW (1999) *Semiconductor-Laser Fundamentals: Physics of the Gain Materials*. Berlin: Springer-Verlag.

Chuang SL (1995) *Physics of Optoelectronic Devices*. New York: Wiley.

Delfyett PJ, Dienes A, Heritage JP, Hong MY and Chang YH (1994) Femtosecond hybrid mode-locked semiconductor laser and amplifier dynamics. *Applied Physics* B 58: 183–195.

Haug H and Koch SW (2004) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, 4th edn. Singapore: World Scientific Publisher.

Hofmann M (1999) Gain and emission dynamics of semiconductor lasers. *Recent Research Developments, in Applied Physics* 2: 269–290.

Koch SW, Jahnke F and Chow WW (1995) Physics of semiconductor microcavity lasers, review article. *Semiconductor Science and Technology* 10: 739–751.

Osinski M and Buus J (1987) Linewidth broadening factor in semiconductor lasers – an overview. *IEEE Journal of Quantum Electronics* QE-23: 9–29.

Up-Conversion Lasers

A Brenier, University of Lyon, Villeurbanne, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Under photoluminescence excitation, a luminescent center located inside a transparent material usually emits at a longer wavelength than the excitation. This is because of the existence of de-excitation losses inside the material. Thus, efficient down-conversion lasers can be built, for example, the commercial Nd^{3+}:$Y_3Al_5O_{12}$ laser which works at 1064 nm, under usual pumping near 800 nm. More surprisingly, it has been demonstrated that fluorescence can also be emitted at shorter wavelength than the excitation. This is possible if the losses are reduced and overcome by so-called 'up conversion mechanisms'. Generally speaking, a laser emitting at a shorter wavelength than the excitation is classified as an up-conversion laser.

These lasers are able to generate coherent light in the red, green, blue, and uv spectral ranges. A point of practical importance is that some of them can be pumped in the near-infrared range (around 800 nm or 980 nm) by commercially available laser diodes. They can be based on all-solid-state technology and share its advantages: compactness of devices, low maintenance, and efficiency. They are needed for a variety of applications: color displays, high density optical data storage, laser printing, biotechnology, submarine communications, gas-laser replacement, etc.

Because the emitted photons have higher energy than the excitation ones, more than one pump photon is needed to generate a single laser photon, so the physical processes involved in the up-conversion laser are essentially nonlinear. Two kinds of laser devices can be considered. In the first one, the excitation of the luminescent center up-migrates through the electronic levels and reaches the initial laser level. This step is due to some luminescence mechanisms. Then, in a second step, lasing at a short wavelength occurs through a transition towards the final laser level. In the second kind of device, the pump excitation activates a down-conversion lasing, but the laser wave is maintained inside the cavity, cannot escape, and is up-converted in short wavelengths by the second-order nonlinear susceptibility of the crystal host. This requires bifunctional nonlinear optical and laser crystals. The device is thus a self-frequency doubling laser or a self-sum frequency mixing laser. If the nonlinearity is not an intrinsic property of the laser crystal but is due to a second nonlinear optical crystal located inside or outside the cavity, the device operates intracavity or extracavity up conversion.

Up-Conversion from Luminescence Mechanisms

The numerous $^{2S+1}L_J$ energy levels ($4f^n$ configuration) of the trivalent rare earth ions, inserted in crystals or glasses, offer many possibilities for up conversion, particularly from long-lived intermediate levels that can be populated with infrared pumping. Crystal field induced transitions between them are comprised of sharp lines, because the $4f^n$ electrons are protected from external electric fields by the $5s^2p^6$ electrons. Their fluorescence spectra can exhibit cross-sections high enough (10^{-19} cm^2) to lead to laser operation in the visible range. For up-conversion purposes, the most popular ions are Er^{3+}, Pr^{3+}, Tm^{3+}, Ho^{3+}, and Nd^{3+}. Up conversion in Er^{3+} was used for the first time in 1959, to detect infrared radiation.

Energy Transfers

Two ions in close proximity interact electrostatically (Coulomb interaction) and can exchange energy. The ion which gives energy is called the sensitizer S or donor, and the energy receiving ion is the activator A or acceptor. Two examples discussed below are shown in Figure 1. The most typical case of such nonradiative energy transfer in trivalent rare earth ions in insulating materials is due to electrical dipole–dipole interaction. For the latter, the transition probability $W(\mathrm{s}^{-1})$ takes the form:

$$W = \frac{3\hbar^4 c^4}{4\pi n^4}\frac{1}{R^6}\frac{Q_A}{\tau_{\mathrm{rad}}}\int \frac{f_A(E) f_S(E)}{E^4}\,\mathrm{d}E \qquad [1]$$

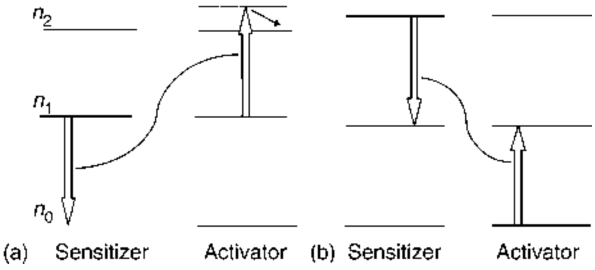

Figure 1 (a) Up-conversion cross-relaxation energy transfer. (b) Down-conversion cross-relaxation energy transfer. The example in (a) is phonon assisted, the example in (b) is resonant.

where R is the S–A distance, E is the photon energy of the transition operated by each ion, n the refractive index, and Q_A is the integrated absorption cross-section of the A-transition:

$$Q_A = \int \sigma_A(E)\mathrm{d}E, \qquad f_A = \frac{\sigma_A}{Q_A} \qquad [2]$$

In eqn [1], τ_{rad} is the radiative emission probability of S for the involved transition calculated from the emission probability A_S:

$$\frac{1}{\tau_{\mathrm{rad}}} = \int A_S(E)\mathrm{d}E, \qquad f_S = \tau_{\mathrm{rad}} A_S \qquad [3]$$

The integral in eqn [1] is called the overlap integral and shows that the transfer is efficient if the luminescence spectrum of S is resonant with the absorption spectrum of A (corresponding to the S and A transitions involved in the nonradiative transfer). The transitions are often not in perfect resonance but remain efficient enough to be used in practice. Then the energy mismatch is compensated by phonons, and the energy transfer is said to be phonon-assisted.

The energy transfer visualized in Figure 1a is an up-conversion cross-relaxation. It is often met in the rare earth ions cited above when they are sufficiently concentrated in the material. The S and A ions can be of the same or of different chemical species. Successive energy transfers can promote the activator from its ground state towards higher energy levels, up to the one from which lasing can occur. This is the so-called 'Addition de Photons par Transferts d'Energie', sometimes appointed 'APTE'. Yb^{3+} is the most used ion as sensitizer and leads to the following transfers in examples of Er^{3+} or Ho^{3+} co-doping:

$$\begin{aligned} &{}^2F_{5/2}(\mathrm{Yb}) + {}^4I_{15/2}(\mathrm{Er}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^4I_{11/2}(\mathrm{Er}) \\ &{}^2F_{5/2}(\mathrm{Yb}) + {}^4I_{11/2}(\mathrm{Er}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^4F_{7/2}(\mathrm{Er}) \end{aligned} \qquad [4]$$

$$\begin{aligned} &{}^2F_{5/2}(\mathrm{Yb}) + {}^5I_8(\mathrm{Ho}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^5I_6(\mathrm{Ho}) \\ &{}^2F_{5/2}(\mathrm{Yb}) + {}^5I_6(\mathrm{Ho}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^5S_2(\mathrm{Ho}) \end{aligned} \qquad [5]$$

If several sensitizers give their energy simultaneously to a single activator, promoting it directly from its ground state towards a high energy level without intermediate levels, the sensitization is said to be 'cooperative'.

The energy transfer shown in Figure 1b is a down-conversion cross-relaxation. It is the inverse of the previous one. At first sight it is, of course, undesirable in an up-conversion laser. However, we notice that one ion in the upper level will lead to two ions in the intermediate level. This fact can be exploited in special conditions (see below) and be beneficial for up conversion.

The dynamical study of the energy transfers, starting from the microscopic point of view contained in eqn [1], is very complicated because of the random locations of the ions or because of other processes, such as energy migration among sensitizers and back transfers. A popular method leading to useful predictions in practice is the rate equation analysis dealing with average parameters. The time evolution of the population densities n_i of the various levels i are described by first-order coupled equations, solved with adequate initial conditions and including the pumping rate. As an example, the rate equations for levels 1 and 2 in Figure 1a take the form:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} - 2kn_1^2 + \frac{\beta_{21}}{\tau_2}n_2 + W \qquad [6]$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + kn_1^2 \qquad [7]$$

$$n_0 + n_1 + n_2 = 1 \qquad [8]$$

where $\tau_{1,2}$ are the lifetimes of levels 1 and 2, k is the up-conversion transfer rate, β_{21} is the branching ratio for the $2 \rightarrow 1$ transition, and W is the pump rate (not shown in Figure 1).

Sequential Two-Photon Absorption

Pumping the ground state of a luminescent ion (Figure 2a) with a wave at angular frequency ω, leads to population with density n_1 of an excited state in a first step. Because the rare earth ions have numerous energy levels, it is possible for the pump to also be resonant with the transition from level 1 towards a higher energy level 2, and is further absorbed. Such a process leading to n_2 up-conversion population is a sequential two-photon absorption. If there is no $1 \rightarrow 2$ resonant transition for the ω frequency, it is possible to use a second pump wave at a different angular frequency ω' resonant with the $1 \rightarrow 2$ transition (Figure 2b). Of course, in practice, it is advantageous when a single pump beam is required for up conversion.

Let us show the relevant terms of the rate equation analysis describing the process in Figure 2a:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2}n_2 \qquad [9]$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 \qquad [10]$$

$$n_0 + n_1 + n_2 = 1 \qquad [11]$$

where I is the photon density of the pump (photons $s^{-1}\,cm^{-2}$) and σ_0 and σ_1 are the absorption cross-sections (cm^2) of the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions, respectively. The n_2 population density shows a nonlinear I dependence, most often close to I^2.

Photon Avalanche

In rare earth doped materials, an up-conversion emission, intense enough for visible lasing, can result from a more complex mechanism. In this case, absorption of the pump is not resonant with a transition from the ground state (weak ground-state absorption), but on the contrary, the photon pump energy matches the gap between two excited states (high excited state absorption). These levels are labeled 1 and 2 in Figure 3. Level 1 generally has a rather long lifetime in order to accumulate

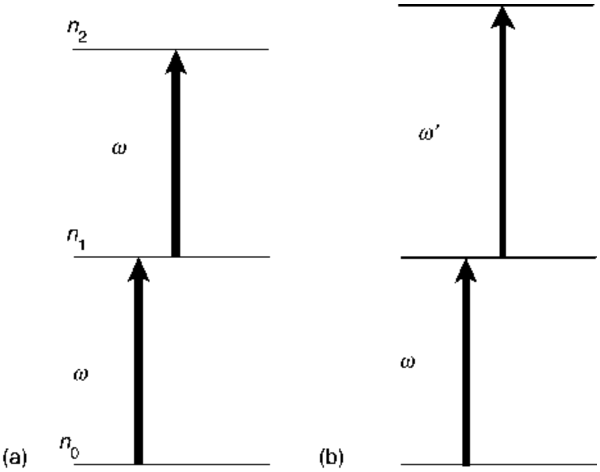

Figure 2 Two-photon absorption up-conversion.

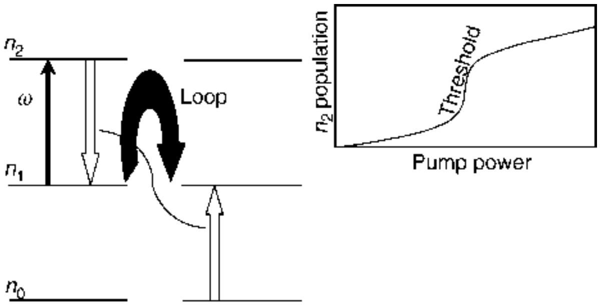

Figure 3 Photon avalanche up-conversion.

population and level 2 is the initial up-conversion laser level. When the pump light is turned on, the populations of levels 1 and 2 first grow slowly, and then can accelerate exponentially. This is generally referred to as 'photon avalanche'. The term 'looping mechanism' also applies, because we can see in Figure 3 that the excited state absorption 1 → 2 is followed by a cross-relaxation energy transfer (quantum efficiency up to two) that returns the system back to level 1. So after one loop, the n_1 population density has been amplified, and is further increased after successive loops. This pumping scheme was discovered in 1979, with $LaCl_3{:}Pr^{3+}$ crystal used as an infrared photon counter.

The rate equation analysis applied to the three-level system in Figure 3 leads to the following time evolutions:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2} n_2 + 2kn_0 n_2 \qquad [12]$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 - kn_0 n_2 \qquad [13]$$

$$n_0 + n_1 + n_2 = 1 \qquad [14]$$

where the parameters have the same meaning as in eqns [6]–[11].

The steady state limit of the populations at infinite time, upon continuous excitation starting at $t = 0$, can be obtained from canceling the time derivative in the left-hand side of the equation. Let us point out the interesting result. Two different dynamical regimes and analytical expressions for $n_2(\infty)$ are obtained. In the usual case of low pumping rate in the ground state and if the parameters of the systems satisfy: $k < \frac{1-\beta_{21}}{\tau_2}$, then the first regime is observed, whatever the pumping rate $I\sigma_1$ in the excited state.

If the parameters satisfy:

$$k > \frac{1-\beta_{21}}{\tau_2} \qquad [15]$$

the first regime is still observed at low pumping rate $I\sigma_1$ in the excited state, more precisely at pumping rate low enough such that

$$I\sigma_1 < \frac{\left(k + \dfrac{1}{\tau_2}\right)\dfrac{1}{\tau_1}}{k - \dfrac{(1-b_{21})}{\tau_2}}$$

But interestingly, the second regime, the so-called photon avalanche, is obtained at high pumping rate $I\sigma_1$ in the excited state, that is to say if:

$$I\sigma_1 > \frac{\left(k + \dfrac{1}{\tau_2}\right)\dfrac{1}{\tau_1}}{k - \dfrac{(1-b_{21})}{\tau_2}} \qquad [16]$$

So the right-hand side of eqn [16] appears to be a pump threshold separating two dynamical regimes when the condition in eqn [15] is satisfied, as the insert in Figure 3 shows.

Another way to distinguish between the two regimes is given by linearization of the system of eqns [12]–[14] ($n_0 \cong 1$). Then the description will only be valid close to $t = 0$. We know from standard mathematical methods that the solutions of eqns [12]–[14] are then a combination of $\exp(\alpha t)$ terms. If the threshold condition in eqn [16] is satisfied, a parameter α in the combination becomes positive, so the population of level 2 increases exponentially (photon avalanche) at early times and the dynamics of the system is unstable.

The physical meaning of eqns [15] and [16] is clarified by introducing the yield η_{CR} of the cross-relaxation process:

$$\eta_{CR} = \frac{k}{k + 1/\tau_2} \qquad [17]$$

and the yield η of the 2 → 1 de-excitation:

$$\eta = \frac{b_{21}/\tau_2}{k + 1/\tau_2} \qquad [18]$$

Then, the threshold condition in eqn [16] can be written as:

$$\left(\frac{I\sigma_1}{I\sigma_1 + 1/\tau_1}\right)(2\eta_{CR} + \eta) > 1 \qquad [19]$$

Starting with one ion in level 1, the first parenthesis in eqn [19] is the yield with which it will be promoted to level 2, and the second parenthesis is the yield of the backward path 2 → 1. So, the left-hand side of eqn [19] is the new number of ions after a looping path 1 → 2 → 1, which has to be higher than 1 for the avalanche to occur. If we consider a lot of successive loops, the origin of the exponential behavior is clear. The role of the first ion in level 1 (and the weak pumping rate in the ground state $I\sigma_0$) is limited to initialize the process. Note that the value of the first parenthesis in eqn [19] is less than 1, so the second parenthesis has to be higher than 1. This condition is equivalent to eqn [15].

Materials: The Energy Gap Law

As we can see from the above mechanisms, up conversion needs intermediate energy levels and is not efficient if these cannot accumulate populations due to a short lifetime. The spontaneous rate of de-excitation of an electronic level is the sum of a radiative part W^{rad} and of a nonradiative part W^{nrad}. The radiative part can be obtained through the absorption cross-section corresponding to the transition or from the Judd–Ofelt theory which also exploits the absorption spectrum. The nonradiative rate is then obtained by subtracting W^{rad} from the measured fluorescence decay rate of the level. Based on the measurements in many hosts for the different trivalent rare earth ions, at different temperatures, the nonradiative decay rate of a J-level towards the next lower J′-level was found to have the form:

$$W_{\text{JJ}'}^{\text{nrad}} = B\exp(-\beta\Delta E_{\text{JJ}'})\left(1-\exp\left(-\frac{\omega_{\max}}{kT}\right)\right)^{-p} \qquad [20]$$

where $\Delta E_{\text{JJ}'}$ is the energy gap between the two levels, $p = \Delta E_{\text{JJ}'}/\omega_{\max}$, $\omega_{\max}$ being the energy of an effective optical phonon of the host. The values of B, β and $\omega_{\max}$ are found by fitting with experimental data. An example of such a fit is represented in Figure 4.

Equation [20] is the well-known and familiar ‘energy gap law’. We have represented it for several oxide and nonoxide hosts at $T = 300$ K in Figure 5. It is clear that chloride, bromide, fluoride will be favorable hosts for up conversion due to their low phonon energy (LaBr$_3$: $\omega_{\max} = 175\ \text{cm}^{-1}$, LiYF$_4$: $\omega_{\max} = 400\ \text{cm}^{-1}$, Y$_3Al_5O_{12}$: $\omega_{\max} = 700\ \text{cm}^{-1}$, silicate glass: $\omega_{\max} = 1100\ \text{cm}^{-1}$). Let us mention also a fluorozirconate glass, ZBLAN, widely used in up-conversion fiber lasers.

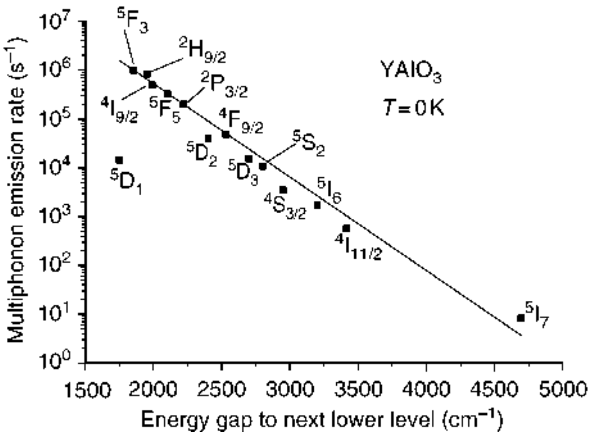

Figure 4 Energy-gap dependence of nonradiative decay rates in YAlO$_3$.

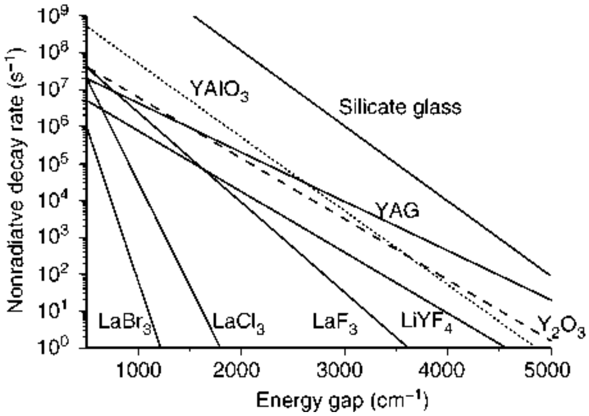

Figure 5 Energy-gap law for different hosts at 300 K.

Up-Conversion from Second-Order Optical Nonlinearity

The inter-atomic electric field acting on the electrons inside a medium has a magnitude of $\sim 10^8$ V/cm and derives from a potential that is anharmonic. The electric field **E** of an electromagnetic wave propagating through the medium drives the electrons beyond the quadratic region of the potential in the case of high **E**. So, the electron response and the associated polarization **P** take the form of a function of **E**:

$$\mathbf{P}(\mathbf{E}) = \mathscr{E}_0(\chi^{(1)} : \mathbf{E} + \chi^{(2)} : \mathbf{E}^2 + \chi^{(3)} : \mathbf{E}^3 + \cdots) \qquad [21]$$

The different terms in eqn [21] are responsible for many optical phenomena because in Maxwell electromagnetic theory, the polarization is the source of the waves. We have selected in Table 1 the terms relevant for linear effects and for the purpose of frequency up conversions: second-harmonic generation (SHG) and sum-frequency mixing, due to second-order terms. It should be noticed that the latter have a nonzero value only in noncentrosymmetric materials.

Figure 6 visualizes the up conversion from SHG: two photons of the fundamental field at angular frequency ω_1 are annihilated while one photon at twice the angular frequency is created. A similar picture could be drawn for sum-frequency up conversion: SHG is the degenerate case where $\omega_1 = \omega_2$.

The second-order nonlinear processes have an efficiency given by an effective coefficient d_{eff} given hereafter:

$$d_{\text{eff}} = \sum_{ijk} e_i(\omega_3) d_{ijk} e_j(\omega_1) e_k(\omega_2) \qquad [22]$$

where $e_i(\omega_l)$ is the ith component of the unit vector of the electric field of the wave l at angular frequency ω_l and $d_{ijk} = (1/2)\chi_{ijk}$.

Table 1 Relevant terms for frequency up-conversion

Optical process	*Term*	*New wave*
Linear		
Refractive index, absorption, stimulated emission	$\chi^{(1)}(-\omega;\omega)$	
Second order nonlinear		
Second harmonic generation	$\chi^{(2)}(-\omega_3;\omega_1,\omega_1)$	$\omega_3 = 2\omega_1$
Sum-frequency mixing	$\chi^{(2)}(-\omega_3;\omega_1,\pm\omega_2)$	$\omega_3 = \omega_1 + \omega_2$

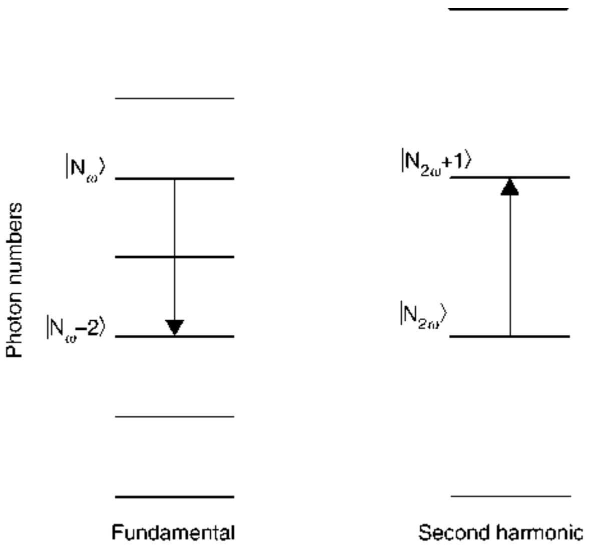

Figure 6 Up-conversion from second harmonic generation.

Moreover, the wave propagation equations in the slow varying envelope approximation, given hereafter for the E_3 electric field of the sum-frequency mixing wave:

$$\frac{dE_3}{dz} = \frac{i\omega_3 d_{\text{eff}}}{n_3 c} E_1 E_2 \quad \exp(i(k_1 + k_2 - k_3)z) \qquad [23]$$

(with $k_j = \omega_j n_j / c$)

impose that the frequency conversion is efficient in practice, only if the following phase matching condition is satisfied:

$$\omega_1 n_1 \uparrow(\theta,\varphi) + \omega_2 n_2 \uparrow(\theta,\varphi) = \omega_3 n_3 \downarrow(\theta,\varphi) \qquad [24]$$

where $n_i \uparrow$ and $n_i \downarrow$ are respectively the upper and lower refractive index in the direction of propagation (θ,φ) (n_i in eqn [23] are the same with a simplified notation). Equation [24] is restricted here for simplicity to collinear type I (waves 1 and 2 and the same polarization) and is the expression of photon momentum conservation. It is usually achieved from the birefringence of the crystals. When this is not possible, another technique can be used, namely quasi-phase matching, which involves reversing periodically the sign of the nonlinear optical coefficient.

In bifunctional crystals, the laser effect and the $\chi^{(2)}$ interaction occur simultaneously inside the same crystal. In the case of the self-frequency doubling laser, the angular frequency $\omega_1 = \omega_2$ in eqn [24] is that of the fundamental laser wave. In the case of the self-sum frequency mixing laser, ω_1 corresponds to the fundamental laser wave and ω_2 to the pump wave. As we can see from eqns [22]–[24], the polarization of the laser emission is crucial in order for the device to work. So, the laser stimulated emission cross-section σ for propagation in the phase matching direction (θ,ϕ) in the adequate polarization has to be evaluated. This can be performed with the formula:

$$\sigma(\theta,\varphi) = \frac{\sigma_X \sigma_Y \sigma_Z}{\sqrt{\sigma_Y^2 \sigma_Z^2 e_X^2 + \sigma_X^2 \sigma_Z^2 e_Y^2 + \sigma_X^2 \sigma_Y^2 e_Z^2}} \qquad [25]$$

where $\sigma_{X,Y,Z}$ are the emission cross-sections for X, Y, Z-polarizations and $e_{X,Y,Z}$ are the components of the unit vector or the electric field of the laser wave.

To summarize, the conditions necessary for a crystal to be a bi-functional material are: (i) it must be noncentrosymmetric; (ii) it must accept fluorescent doping (usually with Nd^{3+} or Yb^{3+}); and (iii) it must be phase matchable for its laser emission. In practice, only a few crystals satisfy these requirements and have been tried with some success. The main ones are $LiNbO_3$, $LaBGeO_5$, $Ba_2NaNb_5O_{15}$, β'-$Gd_2(MoO_4)_3$, $YAl_3(BO_3)_4$, $GdAl_3(BO_3)_4$, $CaY_4(BO_3)_3O$, and $CaGd_4(BO_3)_3O$.

Up-Conversion Lasers Based on Luminescence Mechanisms in Materials

Energy Transfers

Up-conversion lasers from energy transfers have the advantage of using a single pump laser. This latter populates firstly a long-lived intermediate level. A typical example, based on $LiYF_4:Er^{3+}$ crystal, is provided by Figure 7.

The pump at 802 nm from a GaAlAs diode laser, or at 969 nm, populates efficiently at the 8 ms lifetime $^4I_{11/2}$ level. The green laser, at 551 nm, corresponds to the $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition and it can be shown that energy transfer is an essential part of the mechanism by stopping abruptly the pump laser: green lasing continues to occur for several hundred microseconds because the $^4S_{3/2}$ level is still fed (an excited-state absorption would stop immediately). The involved energy transfer is:

$$^4I_{11/2} + {}^4I_{11/2} \rightarrow {}^4F_{7/2} + {}^4I_{15/2} \qquad [26]$$

and leads to a 20 mW laser threshold and 2.3 mW green output power at 50 K temperature upon 95 mW of incident power.

By using mirrors with high transmission in the green range but having high reflectivity at blue wavelengths, blue laser emission at 470 nm can be sustained, corresponding to the $^2P_{3/2} \rightarrow {}^4I_{11/2}$ transition. It originates from a mechanism involving three up-conversion energy transfers:

$$\begin{gathered} ^4I_{11/2} + {}^4I_{11/2} \rightarrow {}^4F_{7/2} + {}^4I_{15/2} \\ ^4I_{11/2} + {}^4I_{13/2} \rightarrow {}^4F_{9/2} + {}^4I_{15/2} \\ ^4S_{3/2} + {}^4F_{9/2} \rightarrow {}^4I_{15/2} + {}^2K_{13/2} \end{gathered} \qquad [27]$$

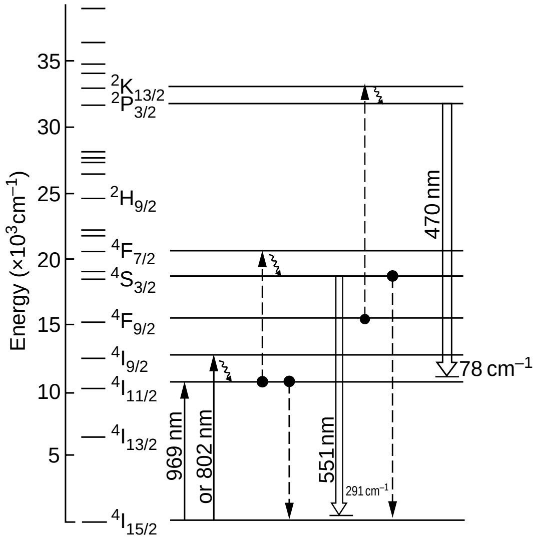

Figure 7 Energy level diagram of LiYF$_4$:Er^{3+}. The dashed lines indicate the energy transfers responsible for the green and blue up-conversion lasing. Reproduced with permission of American Institute of Physics from Macfarlane RM, Tong F, Silversmith AJ and Lenth W (1988) Violet CW neodymium upconversion lasers. *Applied Physics Letters* 52(16): 1300.

The illustration in Figure 7 shows up-conversion energy transfers between ions of the same chemical species, but co-doping the laser host with two ions of different chemical species provides the opportunity of separating their roles: one species is devoted to pump absorption (this is the sensitizer) and the other one is devoted to lasing (this is the activator). Due to the development of efficient semiconductor InGaAs diodes emitting near 980 nm, the most widely used sensitizer is Yb^{3+}. Table 2 provides a list of several up-conversion lasers based on energy transfer mechanisms, operating at room temperature. We can see that among the most efficient lasers, are those based on rare-earth ions doped glass fibers. The reason is that in fibers, the pump and laser waves remain confined inside a core (typically 5 μm diameter) and have high energy densities over a long length (tenths of cms), which is favorable for all up-conversion mechanisms.

Sequential Two-Photon Absorption

An example of an up-conversion laser pumped by a two-photon absorption mechanism is represented in Figure 8. The material is LaF$_3$:Nd^{3+} and violet lasing at 380 nm corresponds to the $^4D_{3/2} \rightarrow {}^4I_{11/2}$ transition.

Transition from the ground state up to the $^4F_{5/2}$ level is firstly obtained from an infrared beam at 790 nm in order to feed the $^4F_{3/2}$ intermediate level after a fast nonradiative de-excitation. The $^4F_{3/2}$ level has a rather long lifetime: 700 μs, so it accumulates

Table 2 Main room temperature up-conversion lasers operating from energy transfers

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
BaY$_{1.4}$Yb$_{0.59}$Ho$_{0.01}$F$_8$	670	1540 + 1054	
BaYYb$_{0.998}$Tm$_{0.002}$F$_8$	649	1054	1%
BaYYb$_{0.99}$Tm$_{0.01}$F$_8$	799–649–510–455	960	
BaYYb$_{0.99}$Tm$_{0.01}$F$_8$	348	960	
LiYF$_4$:Er(1%):Yb(3%)	551	966	37 mW
LiY$_{0.89}$Yb$_{0.1}$Tm$_{0.01}$F$_4$	810–792	969	80 mW
LiY$_{0.89}$Yb$_{0.1}$Tm$_{0.01}$F$_4$	650	969	5 mW
LiKYF$_5$:Er (1%)	550	808	150 mW
Fiber:Yb:Pr	635	849	20 mW
Fiber:Yb:Pr	635	1016	6.2 mW
Fiber:Yb:Pr	521	833	0.7 mW
Fiber:Yb:Pr	635	860	4 mW

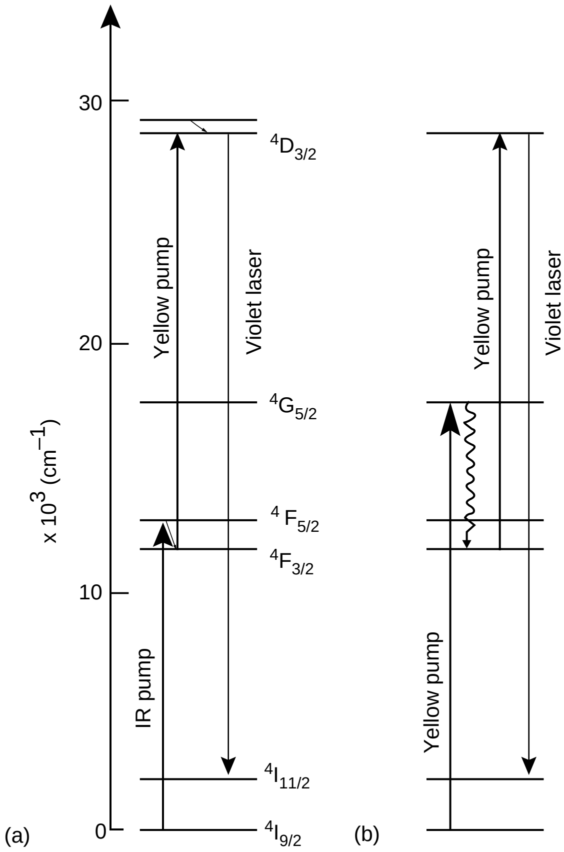

Figure 8 Simplified energy level diagram of LaF$_3$:Nd^{3+} and up-conversion lasing from two-photon absorption. (a) Two different pump wavelengths, (b) doubly resonant single pump wavelength. Reproduced with permission of American Institute of Physics from Lenth W, Silversmith AJ and Macfarlane RM (1988) Green infrared pumped erbium up conversion lasers. In: TAM AC, Gole JL and Stwalley WC (eds) *Advances in Laser Science – III*. AIP Conference Proceedings n°172: 8–12.

enough population that a second yellow pumping at 591 nm (Figure 8a) can efficiently populate the $^4D_{3/2}$ initial laser level. With 110 mW of infrared pump and 300 mW of yellow pump, 12 mW of violet output were obtained at 20 K temperature.

A simpler device is obtained if a single-pump beam is used. This is the case in Figure 8b, where a yellow pump at 578 nm provides both ground state and excited state absorption. This doubly resonant scheme is less efficient in LaF$_3$:Nd^{3+} than the previous one but in Figure 9, we show another example: LiYF$_4$:Er^{3+}, working at room temperature. The lasing $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition at 551 nm delivers 20 mW upon 1600 mW pumping at 974 nm.

The simplified Er^{3+} level scheme of Figure 9 contains four main levels with population densities: $n_0 = [^4I_{15/2}]$, $n_1 = [^4I_{13/2}]$, $n_2 = [^4I_{11/2}]$, $n_3 = [^4S_{3/2}]$. The rate equations of populations of this system can

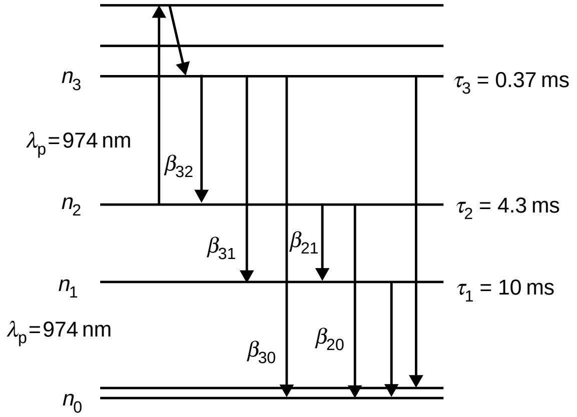

Figure 9 Simplified energy level diagram of LiYF$_4$:Er^{3+} and up-conversion lasing from two-photon absorption (doubly resonant). Reproduced with permission of Institute of Physics Publishing from Huber G (1999) Visible cw solid-state lasers. *Advances in Lasers and Applications*, Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing, 19.

Table 3 Main room temperature up-conversion lasers operating from two-photon absorption

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
Y$_3$Al$_5$O$_{12}$:Er 1%	561	647 + 810	
LiYF$_4$:Er 1%	551	647 + 810	0.95 mJ
LiYF$_4$:Er 1%	551	810	40 mW
LiYF$_4$:Er 1%	551	974	45 mW
LiYF$_4$:Er	551	974	20 mW
KYF$_4$:Er 1%	562	647 + 810	0.95 mJ
LiLuF$_4$:Er	552	974	70 mW
LiLuF$_4$:Er	552	970	213 mW
LiYF$_4$:Tm 1%	453–450	781 + 649	0.2 mJ
Fiber:Tm	480	1120	57 mW
Fiber:Tm	480	1100–1180	45 mW
Fiber:Tm	455	645 + 1064	3 mW
Fiber:Tm	803–816	1064	1.2 W
Fiber:Tm	482	1123	120 mW
Fiber:Yb:Tm	650	1120	
Fiber:Ho	540–553	643	38 mW
Fiber:Ho	544–549	645	20 mW
Fiber:Er	548	800	15 mW
Fiber:Er	546	801	23 mW
Fiber:Er	544	971	12 mW
Fiber:Er	543	800	
Fiber:Pr	635	1010 + 835	180 mW
Fiber:Pr	605	1010 + 835	30 mW
Fiber:Pr	635	1020 + 840	54 mW
Fiber:Pr	520	1020 + 840	20 mW
Fiber:Pr	491	1020 + 840	7 mW
Fiber:Nd	381	590	74 μW
Fiber:Nd	412	590	500 μW

be written and solved because all the implied parameters (lifetimes, branching ratios, ground, and excited states absorption cross-sections) have been measured in this material. Predictive models of the up-conversion green laser are successful, matching experimental measurements, and confirm that the doubly resonant mechanism is realistic at low doping concentration.

Table 3 summarizes the performances of the main up-conversion lasers based on sequential photon absorption and working at room temperature.

Photon Avalanche

Photon avalanche up-conversion was observed mainly in Nd^{3+}, Tm^{3+}, Er^{3+}, and Pr^{3+} doped materials. As an illustration of this mechanism, let us consider the case of the continuous wave Pr^{3+}–Yb^{3+}-doped fluorozirconate fiber laser (3000 ppm Pr, 20 000 ppm Yb). With two Ti:sapphire pump lasers operating at 852 and 826 nm (Figure 10), an output power at 635 nm as high as 1.02 W was obtained with an incident pump power of 5.51 W (19% slope efficiency).

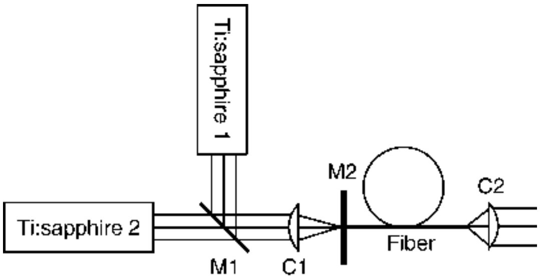

Figure 10 Experimental set-up for Yb–Pr-doped fiber up-conversion laser. M1: dichroic mirror, C1, C2: lenses, M2 dichroic input mirror. Reproduced with permission of Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

This remarkable result can be explained quantitatively by modeling the photon avalanche with a five-level diagram and restricted in Figure 11 to a single laser pump at 850 nm. The laser emission corresponds to the $^3P_0 \rightarrow {}^3F_2$ transition. The five relevant levels have the population densities: $n_0 = [^3H_4(Pr)]$, $n_1 = [^1G_4(Pr)]$, $n_2 = [^3P_0(Pr)]$, $n_3 = [^2F_{7/2}(Yb)]$, $n_4 = [^2F_{5/2}(Yb)]$.

The two components of the mechanism are:

(i) the excited state absorption of the pump by the spin allowed transition: $^1G_4 \rightarrow {}^1I_6$

(ii) the process leading to the doubling of the $^1G_4$ population, in this instance through two energy transfers:

$$\begin{aligned} &^1I_6(Pr) + {}^2F_{7/2}(Yb) \rightarrow {}^1G_4(Pr) + {}^2F_{5/2}(Yb) \\ &^2F_{5/2}(Yb) + {}^3H_4(Pr) \rightarrow {}^2F_{7/2}(Yb) + {}^1G_4(Pr) \end{aligned} \tag{28}$$

The rate equation analysis predicts that the avalanche threshold occurs at about 1 W pump and the long fluorescence rise time is somewhat shortened at higher pump power, as observed experimentally.

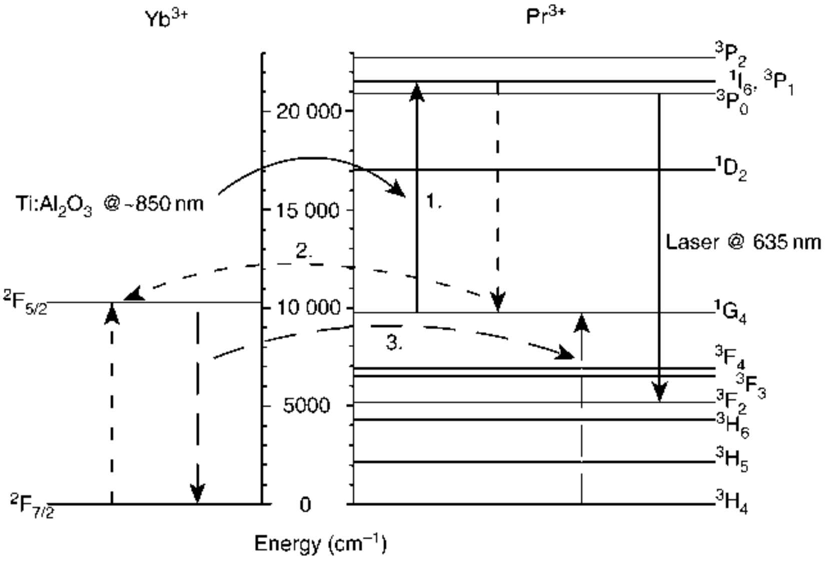

Figure 11 Energy level diagram of fluorozirconate: Yb^{3+}:Pr^{3+} fiber. The dashed lines indicate the energy transfers. Reproduced with permission of the Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

Table 4 Main room temperature up-conversion lasers operating from photon avalanche

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
BaY_2F_8:Yb:Pr	607.5	822	55 mW
BaY_2F_8:Yb:Pr	638.7	841	26 mW
$LiY_{0.89}Yb_{0.1}Pr_{0.01}F_4$	720	830	1%
$LiY_{0.89}Yb_{0.1}Pr_{0.01}F_4$	639.5	830	
Fiber:Yb:Pr	635–637	780–880	300 mW
Fiber:Yb:Pr	605–622	780–880	44 mW
Fiber:Yb:Pr	517–540	780–820	20 mW
Fiber:Yb:Pr	491–493	780–880	4 mW
Fiber:Yb:Pr	635	850	1.02 W
Fiber:Yb:Pr	635	850	2.06 W
Fiber:Yb:Pr	520	850	0.3 W

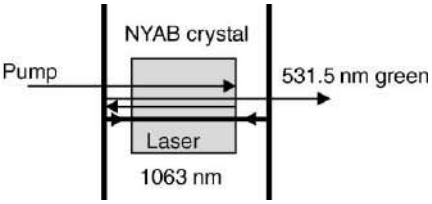

Figure 12 Scheme of a self-frequency doubling laser. Note that in reality the three beams are superimposed.

Table 4 summarizes the performances of the main up-conversion lasers based on photon avalanche mechanisms.

Up-Conversion Lasers Based on Bi-Functional Crystals

The laser stimulated emission in a bi-functional crystal was first obtained in 1969 with Tm^{3+} in $LiNbO_3$ but the most used ion is, of course, Nd^{3+} working in the two channels:

$$^4F_{3/2} \rightarrow {}^4I_{11/2} \qquad [29]$$

$$^4F_{3/2} \rightarrow {}^4I_{13/2} \qquad [30]$$

and more recently Yb^{3+} working in the channel:

$$^2F_{5/2} \rightarrow {}^2F_{7/2} \qquad [31]$$

Channels in eqns [29] and [31] operate around 1060 nm wavelength and the channel in eqn [30] near 1338 nm.

The Self-Frequency Doubling Laser

A typical laser scheme is shown in Figure 12.

Table 5 Main results of Nd^{3+} and Yb^{3+} green self-frequency doubling lasers near 530 nm (eqns [29] and [31])

Crystal	*Input power (mW)*	*Output power (mW)*	*Pumping*
$LiNbO_3$:MgO:Nd	215	1	Dye laser
$YAl_3(BO_3)_4$:Nd	870	10	Laser diode
$YAl_3(BO_3)_4$:Nd	280	3	Laser diode
$YAl_3(BO_3)_4$:Nd	400	69	Laser diode
$YAl_3(BO_3)_4$:Nd	1380	51	Laser diode
$YAl_3(BO_3)_4$:Nd	369	35	Laser diode
$LiNbO_3$:MgO:Nd	850	18	Dye laser
$LiNbO_3$:Sc_2O_3:Nd	65	0.14	Ti:sapphire laser
$LiNbO_3$:MgO:Nd	100	0.2	Laser diode
$YAl_3(BO_3)_4$:Nd	1600	225	Laser diode
$YAl_3(BO_3)_4$:Nd	2200	450	Ti:sapphire laser
$(Y,Lu)Al_3(BO_3)_4$:Nd	880	24	Laser diode
$LiNbO_3$:ZnO:Nd	430	0.65	Ti:sapphire laser
$Ba_2NaNb_5O_{15}$:Nd	270	46	Ti:sapphire laser
$CaY_4(BO_3)_3O$:Nd	900	62	Laser diode
$CaGd_4(BO_3)_3O$:Nd	1600	192	Ti:sapphire laser
$CaGd_4(BO_3)_3O$:Nd	1250	115	Laser diode
$CaGd_4(BO_3)_3O$:Nd	1560	225	Ti:sapphire laser
$YAl_3(BO_3)_4$:Yb	11 000	1100	Laser diode
$GdAl_3(BO_3)_4$:Nd	2.8 mJ/pulse	0.12 mJ/pulse	Pulsed dye laser

The laser beam cannot escape from the cavity and dichroic mirrors are used. The input mirror has high transmission at the pump wavelength and is highly reflective at laser- and second-harmonic wavelengths. The output mirror is highly reflective at laser wavelength and has high transmission for second harmonic. Channels in eqns [29] and [31] generate green light near 530 nm and the channel in eqn [30] generates red light near 669 nm.

The most exploited bi-functional crystal is $YAl_3(BO_3)_4:Nd^{3+}$, well-known as NYAB. It is a negative uniaxial trigonal crystal (extraordinary index lower than the ordinary one) and its laser emission is easily observed in ordinary polarization with a high cross-section: $2 \times 10^{-19}\,cm^2$ at 1063 nm. Type I phase matching occurs at a polar angle $\theta = 30.7°$ and 26.8° for channels in eqns [29] and [30] respectively. The effective nonlinear optical coefficient d_{eff} is close to $1.4\,pm\,V^{-1}$ at azimuthal angle $\phi = 0°$.

The Yb^{3+} ion has some advantages over the Nd^{3+} ion due to its very simple energy level scheme: there is no excited state absorption at the laser wavelength, no up-conversion losses, no concentration quenching, and no absorption in the green. The small Stokes shift between pump and laser emission reduces the thermal loading of the material during laser operation. A self-doubling laser based on $YAl_3(BO_3)_4:Yb^{3+}$ of crystal has produced 1.1 W green power upon 11 W diode pumping.

The drawback of $YAl_3(BO_3)_4$ is that it is rather difficult to grow because it is not congruent. This difficulty was overcome by the discovery at Ecole Nationale Supérieure de Chimie de Paris of the rare-earth calcium oxoborate $CaGd_4(BO_3)_3O$ which can be grown to a large size by the Czokhralski method. It is a monoclinic biaxial crystal and doped with Nd^{3+}, it was firstly exploited (channel in eqn [29]) in the *XY* principal plane at $\theta = 90°$, $\phi = 46°$ with ($d_{eff} = 0.5\,pm\,V^{-1}$). It was soon recognized that the optimum phase matching direction ($d_{eff} = 1.68\,pm \times V^{-1}$) occurred out of the principal planes, in the direction $\theta = 66.8°$, $\phi = 132.6°$, and high green power can be obtained: 225 mW under 1.56 W pump.

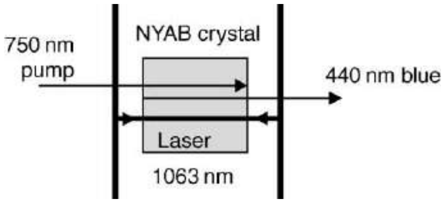

Figure 13 Scheme of a self-sum frequency mixing laser. Note that in reality the three beams are superimposed.

Table 5 summarizes the main self-frequency doubling results obtained with the channel in eqn [29].

The Self-Sum Frequency Mixing Laser

The example in Figure 13 gives the main features of such a laser. The input mirror has high transmission at the pump wavelength and both mirrors are highly reflective at laser wavelength. The output mirror has high transmission at sum frequency mixing wavelength.

The laser wave at angular frequency ω_1 in Table 1 corresponds to the channel in eqns [29] or [30] in the case of Nd^{3+}. The wave at angular frequency ω_2 in Table 1 has two different roles: first it excites the Nd^{3+} laser center and secondly its nonabsorbed part is up converted by the second-order nonlinear process. ω_2 (corresponding to the wavelength λ_2) is then chosen to match the main Nd^{3+} absorption lines:

$$\begin{aligned} &{}^4I_{9/2} \rightarrow {}^4G_{5/2} - {}^2G_{7/2}\ (\lambda_2 = 590\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{7/2} - {}^4S_{3/2}\ (\lambda_2 = 750\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{5/2} - {}^2H_{9/2}\ (\lambda = 800\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{3/2}\ (\lambda_2 = 880\ \text{nm}) \end{aligned} \qquad [32]$$

The most efficient self-frequency mixing lasers based on channels in eqns [29]–[31] are gathered in Table 6.

See also

Materials Characterization Techniques: $\chi^{(2)}$. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation.

Table 6 Main results of Nd^{3+} self-sum frequency mixing lasers based on eqn [29]

Crystal	*Pump wavelength (nm)*	*Generated wavelength (nm)*	*Output power*
$YAl_3(BO_3)_4$:Nd	740–760	436–443	0.16 mJ/pulse
$GdAl_3(BO_3)_4$:Nd	740–760	436–443	0.403 mJ/pulse
$CaGd_4(BO_3)_3O$:Nd	811	465	1 mW
$YAl_3(BO_3)_4$:Nd	585–600	377–383	0.25 mJ/pulse
$GdAl_3(BO_3)_4$:Nd	587–597	378–382	0.105 mJ/pulse
$YAl_3(BO_3)_4$:Nd	488, 515	330, 380	0.2 mW

Further Reading

Brenier A (2000) The self-doubling and summing lasers: overview and modeling. *Journal of Luminescence* 91: 121–132.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Huber G (1999) Visible cw solid-state lasers. *Advances in Lasers and Applications*. Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing.

Joubert MF (1999) Photon avalanche up-conversion in rare earth laser materials. *Optical Materials* 11: 181–203.

Kaminskii AA (1996) *Crystalline Lasers: Physical Processes and Operating Schemes*. CRC Press, Inc.

Lenth W and Macfarlane RM (1992) Upconversion lasers. *Optics & Photonics News* 3: 8–15.

Powell RC (1998) *Physics of Solid-State Laser Materials*. New York: Springer-Verlag.

Risk WP, Gosnell T and Nurmikko AV (2003) *Compact Blue-Green Lasers*. Cambridge, UK: Cambridge University Press.

Weber MJ (2001) *Handbook of Lasers*. Boca Raton, FL: CRC Press.

Yariv A (1995) *Quantum Electronics,* Third Edition. John Wiley & Sons.

LASER-INDUCED DAMAGE OF OPTICAL MATERIALS

A J Glass and A H Guenther, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Shortly after the invention of the laser, in the early 1960s, it became apparent that laser performance was limited by the ability of optical materials to withstand high energy and power densities. Broadly speaking, laser damage arises from two causes: (i) thermal effects, which are associated with high average power; and (ii) dielectric breakdown, which is associated with high values of the local electric field. Early researchers quickly recognized that most of the observed damage effects arose from extrinsic causes, such as impurities and inclusions in the bulk material, or particles or defects on exposed optical surfaces. At that time, optical materials had not been developed for high-power use, and lacked the quality needed for laser applications. In fact, in many situations, the laser provided the inspection light source used to assess the properties of the materials, such as absorption, defects, scattering centers, surface topography, and dynamic behavior. Thus, the field of laser damage studies was broadened to include many aspects of material preparation and characterization. At high-power densities, nonlinear effects become important, often limiting performance without creating any irreversible change in the material's properties. The most general definition of laser-induced damage includes all effects arising in optical materials and components that limit or degrade the performance of laser systems.

The pursuit of improved performance at high values of power and energy densities was motivated by several factors and applications. In general, improved tolerance to high flux and fluence leads to smaller, lighter, and more efficient systems. These are particularly important for military and energy applications. Since the 1970s, the laser fusion program alone has produced an extensive body of research on laser-induced damage in high-power nanosecond-scale systems, which has enabled orders of magnitude improvement in the damage resistance and extraction efficiency of glass lasers. More recently, with the development of fiber laser systems producing and transporting a kilowatt of average power in a fiber of less than 100 μm diameter, laser damage phenomena on an entirely different spatial scale have become important.

Since the early 1970s, the literature of laser damage research has been collected in a series of annual publications of the Boulder Damage Symposium, initiated by the authors of this article, which was first held in 1969. Recently the papers presented at the Damage Symposium in the years 1969 to 1998 were collected and published in CD form by the SPIE. A CD containing the proceedings of the next five years (1999–2003) will be released by SPIE this year. Also in the late 1960s, in the Soviet Union, Academician Aleksei Bonch-Bruevich launched a comprehensive research project on the nonresonant interaction of laser radiation with matter, opening up a new branch of photonics research, known in the Soviet Union as power optics. In 1969, he founded the All-Union Conference on Interaction of Optical Radiation with Matter (AURICOM), which was held every two to three years until 1990. After 1990, the conference was renamed Nonresonant Laser-Matter Interaction (NLMI), and chaired by Professor Mikhail Libenson. NLMI was held in St Petersburg in 1996, 2000, and

2003. The proceedings of the 1996 AURICOM and subsequent NLMI conferences have been published by the SPIE. We are saddened to note that Professor Libenson died in February 2004.

The Boulder Damage Symposium grew out of a one-day Symposium convened in Boulder in 1969 under the auspices of the ASTM Subcommittee on Lasers, to establish standards for laser materials. In their summary of the first Symposium, the conference organizers remarked:

> The entire question of the nature of the standards remains to be addressed. Should they take the form of energy density at which catastrophic damage is likely to occur, or should they be specified in terms of a mean number of shots at a given level of energy density before certain degradation of performance is measured? The latter would seem to express the kind of information that the buyer of laser materials would find most useful. Once a standard is agreed upon, a meaningful test configuration must be established. It is likely that this will be a well-controlled oscillator amplifier chain with the test specimen an active element of the system, but this is not certain. Another point to be determined is that of the accuracy to which its specifications should be written. This depends in part on the kind of quality control the producers of laser glass believe is feasible.

The questions raised in this first Symposium remain at the core of laser damage studies today. How do we define damage, in terms of the physical effect on the material or in terms of the degradation of performance of the system? How do we measure damage thresholds or values? How do we specify the properties and preparation of the sample, including the relevant experimental and material variables? How can we ensure that damage values are reproducible, unless we make the measurements in a very well controlled and characterized system? With what precision should damage levels be stated, in order to be useful? Extensive research was required in order to answer these questions, and to elucidate the fundamental nature of damage mechanisms. Many of the results of this research are documented in the 35 year record of the Boulder Damage Symposium and the corresponding Russian-based conferences.

From the beginning of this area of research, investigators devoted a significant effort to establishing the 'first causes' of damage phenomena. However, they quickly realized that in most cases, damage occurred at surfaces and material interfaces, or was mediated by impurities or inclusions in bulk materials. As material fabrication processes improved, and higher-quality materials became available, observed damage thresholds approached the intrinsic limits of the pure material. This was particularly true of silicate glass, metal reflectors, and in some cases, polymers. In fact, silicate glass has become one of the purest noncrystalline materials available in bulk quantities. However, for practical purposes, most laser systems are limited by damage at surfaces, coatings, or impurities. Impurities aggregate at grain boundaries in crystalline materials, and coalesce into inclusions in glasses, resulting in absorption, scattering, and field enhancement.

The coherent nature of laser light makes any scattering phenomena potentially damaging, because the scattered light interferes constructively with the laser beam, leading to the formation of intensity hot spots. Every interface in an optical system is potentially a source of constructive interference between the reflected and incident waves. Multilayer dielectric coatings are particularly vulnerable to damage, due to the presence of residual stress as well as field enhancement from Fresnel reflection. This has led to attempts to mitigate the problem by use of specific designs, such as avoiding AR coatings by using Brewster-angle surfaces, or use of nonquarter wave layers to move the peak field away from the vicinity of boundaries where impurities tend to collect. Other damage-sensitive design considerations can be found in the literature. Multiple reflections are the basis of photonic crystal optics, and applications of these devices to high-power density systems will undoubtedly be limited by field-enhanced damage phenomena.

In CW operation, of course, thermal effects dominate. They may be irreversible, creating permanent damage sites in the medium, or reversible, such as color centers that can be bleached out. As laser pulse durations vary from nanosecond to picosecond and femtosecond durations, the nature of the dominant damage mechanisms changes as well. As a rule of thumb, the scaling of damage thresholds as $\tau^{1/2}$, where τ is the pulse duration, was established empirically, and was found to be valid over a wide range of pulse durations. This scaling implies the presence of diffusion effects, either of heat or of electrons, in the vicinity of a damage center. As might be expected, this scaling does not apply for very short pulses (femtoseconds or less), where the extremely high field values can lead to instantaneous breakdown of the material.

Whether the damage is due to field enhancement or is impurity-mediated, it is difficult to establish a reproducible value for a damage threshold, unless the sample is macroscopic in extent, and the sampling laser beam is reproducibly of high spatial and temporal quality, preferably truly single-mode. Damage measurements on microscopic areas are unlikely to be representative of the entire sample. Damage will occur first at a point where the field is strongest or the material is weakest, and a sufficient area must be tested to ensure that it includes such sites. Also, unless the beam quality of the laser used for testing is well controlled and well characterized, it will not be possible to reproduce the test results. It is equally important to characterize the samples, including their method of preparation, cleaning, and test environment. For damage research, as opposed to simple threshold testing, postmortem examination is also required, to elucidate the mechanism of the observed damage effect. However, we now have several techniques, such as photothermal deflection, that can provide useful pre-catastrophic indicators of the onset of laser damage, and thus are appropriate for quality control.

Professor Roger Wood had published a series of reviews and textbooks on laser damage in optical materials, which provide an overview of the field, with appropriate emphasis on measurement techniques, characterization of the test setup and sample, and implications for system design. He has also edited a collection of seminal publications in the field, published by the SPIE in 1990.

It is gratifying to those working in this field to observe how much of the research carried out to elucidate and, if possible, avoid the deleterious effects of laser damage has been applied to the constructive purpose of laser-assisted material processing and manufacturing. The introduction of low-cost, reliable femtosecond lasers has stimulated activity in this field, since it enables the experimenter to control to exquisite precision the power and energy deposited on a material. With femtosecond pulses, one can ablate material from the surface of an object while minimizing any collateral damage to the interior material. Surface layers of dirt can be removed from priceless art works such as oil paintings, without concern for damage to the object itself. Objects with arbitrary shape can be built up from metal particles, using laser additive manufacturing techniques. Thin films and photonic crystal structures can be created using laser-assisted deposition. Laser irradiation provides a power tool for surface cleaning, conditioning, and annealing. Laser interaction with materials and laser-assisted manufacturing techniques comprise a vital area of research in the materials science community.

In both the US and Russia, study of the interaction of intense coherent light with optical materials began with a disarmingly simple objective; to understand the fundamental mechanisms of interaction. In the review and summary of the 1969 ASTM Symposium, the editors wrote:

> In view of the number of problems remaining to be resolved, it is suggested that another Symposium on laser damage be held in 1970. Hopefully, at that time a better understanding of the nature of damage in laser glass will have been obtained. Higher threshold values will have been reproducibly achieved, and some agreement can be arrived at regarding useful and realistic standards.

Thirty-five years later, damage levels in optical materials have been significantly improved, and realistic and useful standards have been developed under the auspices of the International Organization for Standardization (ISO), and are now accepted globally. However, as the operating regimes of lasers have moved to femtosecond temporal scale and micrometer dimensions, new phenomena have arisen that require further research. An entirely new technology of laser-enabled manufacturing and materials conditioning has developed, based on the same physics as laser damage. Combining the disciplines of materials science and applied optics, the interaction of intense light with materials remains a scientifically challenging and economically important field of research.

See also

Optical Coatings: Laser Damage in Thin Film Coatings.

Further Reading

Bonch-Bruevich AM, Konov VI and Libenson MN (eds) (1990) *Optical Radiation Interaction with Matter*. SPIE Proceedings, vol. 1440.

Glass AJ and Guenther AH (eds) (1969) *Damage in Laser Glass Symposium: Review and Summary*. ASTM Special Technical Publication 469.

ISO Online Catalogue, Section 31.260, http://www.iso.ch/iso/en/CatalogueListPage.CatalogueList

Laser-Induced Damage in Optical Materials, Collected Papers 1969–1998, (CD-ROM), SPIE, Bellingham, Washington (1999).

Laser-Induced Damage in Optical Materials, Collected Papers 1999–2003, (CD-ROM), SPIE, Bellingham, Washington (2004) (To be published).

Libenson MN (ed.) (1997) *Non-Resonant Laser-Matter Interaction (NLMI-9)*. SPIE Proceedings, vol. 3093.

Libenson MN (ed.) (2001) *Non-Resonant Laser-Matter Interaction (NLMI-10)*. SPIE Proceedings, vol. 4423.

Libenson MN (ed.) (2004) *Non-Resonant Laser-Matter Interaction (NLMI-11)*. SPIE Proceedings, vol. 5506, (To be published).

Wood RM (ed.) (1986) *Laser Damage in Optical Materials*. Bristol, UK: Adam Hilger.

Wood RM (ed.) (1990) *Selected Papers on Laser Damage in Optical Material*. SPIE Milestone Series, vol. MS24. SPIE Optical Engineering Press.

Wood RM (ed.) (2003) *Laser-Induced Damage in Optical Materials*. Bristol, UK: Institute of Physics.

Wood RM (ed.) (2003) *The Power- and Energy-Handling Capability of Optical Materials, Components, and Systems*. SPIE Press Tutorial Texts in Optical Engineering, vol. TT60.

LIGHT EMITTING DIODES

J Schanda, University of Veszprém, Veszprém, Hungary

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Short Historic Overview

During the second half of the nineteenth century, electric lighting became possible. In 1879, Edison demonstrated his incandescent lamp, and although some electric arc lamps competed with the incandescent lamp bulb for a few years, the first 50 years of electrical lighting was dominated by the incandescent lamp. Gas discharge lamps (both low- and high-pressure lamps) became more widely distributed only by the end of the third decade of the twentieth century, and widespread distribution of the fluorescent lamp, that still dominates our offices and many other lighting applications today, was generally accepted only after the Second World War.

The early incandescent lamps had an efficacy of a few lm/W, but by 1920 this could be increased by up to 10–15 lm/W. Gas discharge lamps started with approximately 30 lm/W, but by 1960 their efficacy increased to about 70 lm/W. Fluorescent lamps still provide only 70 to 100 lm/W; high-pressure lamp efficacy could be increased above this value. The different lamp types can be used, however, in different applications, depending on the size and the unit power of the lamp type. Nonincandescent light generation in solid-state materials dates back to the 1920s: a faint glow of a SiC crystal was observed when current passed through it at a point contact. In the 1930s, electroluminescence of ZnS powder layers was detected (Destriaux effect). By the mid-1960s, it became clear that by using the then available technology, good luminous efficiency could be expected only in III–V compounds (GaAs, GaP, and their ternary and quaternary compounds using Al, In, and N in the compounds). Pure GaAs has a bandgap in the infrared (IR), thus one can build good IR-emitters using this material, but no visible light emitter. However, GaAs proved to be an excellent coherent light-emitting material and enabled the construction of IR-emitting semiconductor lasers.

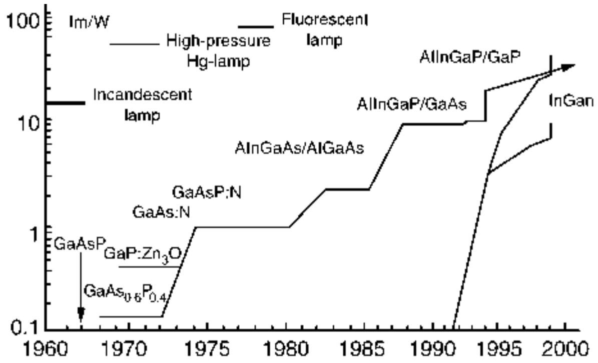

Figure 1 Increase of luminous efficacy of LEDs during the past 40 years. Courtesy of Lumileds Lighting.

First visible light-emitting diodes were fabricated using a GaAsP alloy. Figure 1 shows how the efficacy of III–V compound LEDs can be increased by introducing new compositions, and later, new technologies and better light extraction techniques. In the figure we also show the efficacy values of average incandescent, high-pressure Hg- and fluorescent lamps. The first LEDs emitted in the red part of the spectrum, and only by 1975 was the green LED invented. By 1985, practically any color between red and green became possible, and the efficacy of these lamps reached a level where for monochromatic light applications, their efficacy could compete favorably with filtered incandescent light. LEDs emit in a relatively narrow wavelength band (20 nm–30 nm bandwidth), and if one has to filter from the continuous spectrum of an incandescent lamp such a narrowband, the loss becomes considerable. LEDs needed only a few volts to function and it became general practice to build LEDs with approximately a 20 mA current load. These LEDs became very popular for interior signaling (on instruments, household appliances, and in similar applications). The goal was, however, to be able to produce white light, but the blue-emitting LED was still missing. Only after 1990 could the problem of producing p–n junctions from GaN, that had a bandgap large enough to emit blue light, be solved. Different ternary and quaternary compounds, as shown in Figure 1, enabled efficacies comparable to incandescent lamp efficacy, and the race for producing white light began.

Physical Fundamentals

Figure 2 shows a forward-biased p–n junction. As injected free electrons and holes are driven towards the junction, they can radiatively recombine, emitting a photon. The probability of photon emission versus nonradiative recombination will depend on the band

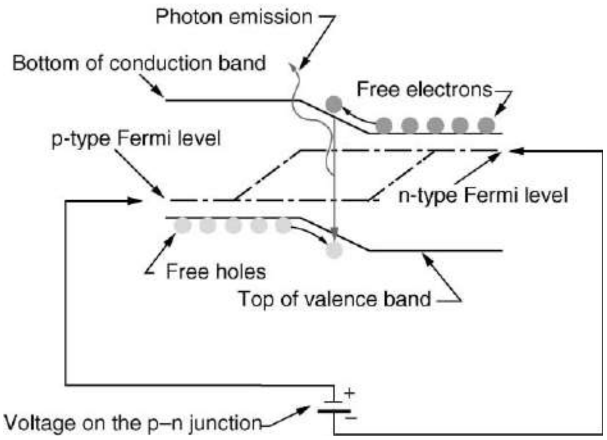

Figure 2 Forward biased p–n junction, light generation via recombination of free charge carriers.

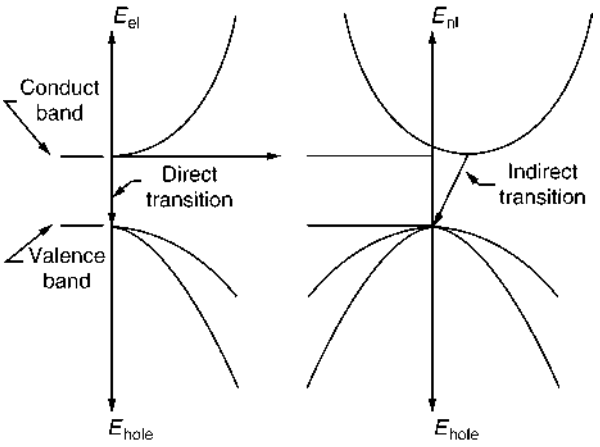

Figure 3 Band to band recombination in direct and indirect band semiconductor.

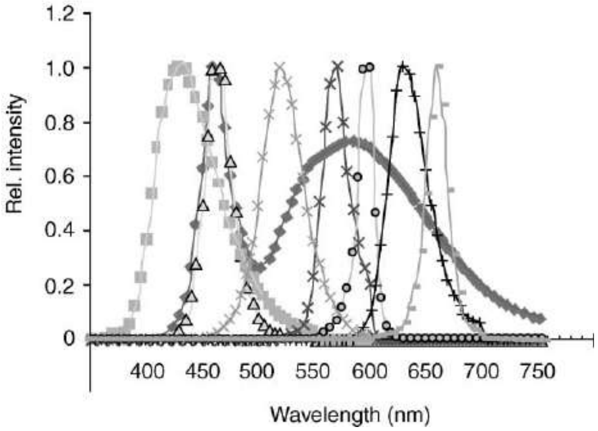

Figure 4 Relative spectral power distribution of some LED structures.

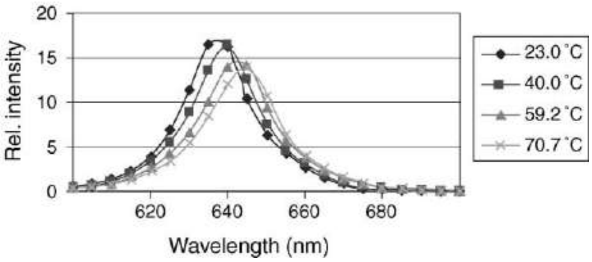

Figure 5 Spectral power distribution of a red LED at different temperatures.

structure of the semiconductor. Figure 3 shows the two possibilities of a direct and an indirect bandgap semiconductor. GaAs is, for example, a direct bandgap material. In this material free electrons at the bottom of the conduction band can recombine with free holes at the top of the valence band, as the momentum of the two ‘particles’ is the same and a photon emission takes care of the principle of energy conservation. GaP is an indirect bandgap material where the momentum of the free electrons at the bottom of the conduction band differs from the momentum of the free holes at the top of the valence band, and thus the recombination needs not only the fulfilment of the conservation of energy, but also of momentum. Thus, besides photon emission, phonon emission or absorption has to take place. Thus, the probability of the transition will decrease and competing nonradiative processes will become important. Using N doping produces very shallow donor states and from these an efficient recombination can take place. This enabled the production of early relatively efficient green-emitting LEDs.

GaN is ideal for producing LEDs, as it is a direct gap material and has a high bandgap. The technological difficulties are, however, enormous. Modern high-efficiency LEDs use (Al)InGaN-based materials for shorter wavelength LEDs, and AlInGaP-based materials for longer wavelength LEDs. Figure 4 shows a number of LED spectra of different wavelengths including a white LED. Its spectrum consists of a blue maximum produced by the LED emission and a yellow broader band produced by the luminescent material excited by the blue emission of the LED.

LEDs are semiconductor materials, and as with every semiconductor the LED characteristics are also temperature dependent. Not only the efficiency of the current carrier–photon emission process depends on the temperature, but the bandgap of the base material also depends on the temperature, and thus both the efficacy and the relative spectral distribution of the emission change with temperature. Figure 5 shows the spectral power distribution of a red LED at different temperatures.

Modern High Brightness LEDs

Besides the high quantum efficiency of the semiconductor junction, two further requirements to achieve high luminous efficacy are the efficient light extraction from the junction and keeping the temperature of the semiconductor chip low.

All semiconductors for producing LEDs have a high refractive index, thus light reflection at the semiconductor–air interface is relatively high, and for obliquely impinging rays the critical angle for total internal reflection is soon reached. Different manufacturers shaped the semiconductor chip in different forms to secure maximum light extraction. Figure 6 shows the cross-section of an LED, where the side planes have been cut obliquely to secure good light extraction.

The other problem that has to be solved is good heat conduction from the chip to a heat sink. At the time of writing, highest single chip power loads were in the order of 5 W. The structure of such a construction is seen in Figure 7. The GaN-based LED is grown on a visible light transparent sapphire (Al_2O_3) substrate. In modern LEDs the structure itself is quite complicated; in the vicinity of the p–n junction some nanometer thick layers are deposited forming quantum wells (QW(s)). The p-type original upper layer is metalized and soldered to the submount serving as heat-sink. The n-type layer gets a second soldered heat conducting current leading to the submount. As seen, the original structure grown on the sapphire substrate is inverted, and now the light can be extracted through the sapphire layer. The metalization on the p-type layer serves at the same time as a mirror. High brightness LEDs are now available in package sizes of over 100 lumen.

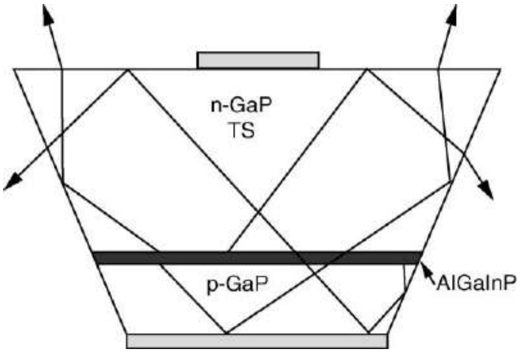

Figure 6 Truncated pyramid structure to secure good light extraction from the LED chip. Courtesy of Lumileds Lighting.

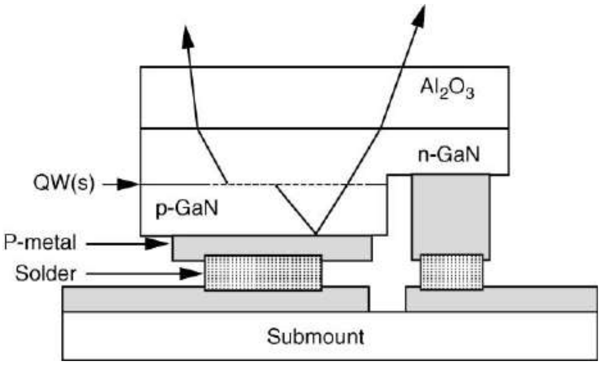

Figure 7 Inverted, flip-chip structure for good heat-sinking of the LED. Courtesy of Lumileds Lighting.

White LED Structures

The advantages of LEDs (low drive voltage, extreme long lifetime of 50000 to 100000 hours, robust shock-resistant construction) can open up a new era of lighting engineering, if white light emission can be obtained. The LED structure in itself is a narrowband emitter, thus – with inorganic LEDs – it is not possible to produce white light within one chip. Organic LEDs (OLEDs) promise the production of broadband and white light emission, but these are still in an experimental phase. At present they are used in smaller, portable displays, such as those used on portable phones, cam-coders, etc. To get further information, go to the Further Reading section at the end of this article.

At present three techniques to generate white light get thorough investigation: first, mixing the light of a red, a green, and a blue LED. This should theoretically give the highest efficiency; the problem is, however, that the three LEDs have different temperature dependence and aging characteristics, and thus it is difficult to keep the chromaticity of the mixed color constant (the three colored LEDs can be used in LED-based displays to produce full-color reproduction). At present the following technique is widely used. A blue LED excites a yellow phosphor, so that the mixture of the blue and yellow lights produce a white appearance. The color rendering of such white LEDs is not too good, and therefore alternative techniques using a green and a red phosphor are receiving increasing attention. The third method uses the same principle, but with an LED emitting in the near ultraviolet. This has the advantage that it provides a greater selection of phosphors, but as the energy difference between the UV photon and the photon emitted by the phosphor is larger than in the case of a blue-emitting LED, the theoretical efficiency is lower.

The phosphor layer is deposited on the chip surface, and for blue-emitting diodes special care has to be taken that the LED light and the phosphor light are properly mixed so that the emitted light shows no directional color shift. Early white LEDs suffered much from the fact that the spectral distribution of the light changed considerably with viewing angle.

The problems with phosphor-coated white LEDs are how a suitable correlated color temperature and color rendering can be obtained. Early white LEDs had very bluish chromaticity, their correlated color temperature was very high, above 7000 K. Preferred white light has different chromaticity in different cultural regions, in the Far East 6000 K to 6500 K is preferred, in Europe and North America for general lighting a correlated color temperature of 4000 K is well accepted and for social and home lighting even lower correlated color temperatures, 2800 K to 3200 K are preferred. Recently, white LEDs based on blue chips and special phosphor blends have been produced that meet these requirements.

A further problem is the color rendering of the white LEDs. The Commission Internationale d'Eclairage (CIE), the International Commission on Illumination, provides standards, recommendations, and technical reports in the field of light and lighting, including vision, color, and image processing. Photometry and colorimetry are based on the standards and recommendations of the CIE.

The CIE established a method of calculating a color-rendering index in 1965 and updated it in 1974. Despite several attempts, the lighting community has been unable to agree on a revision of this outdated document since the 1970s. Thus, color-rendering indices are still calculated based on this method, although several investigations have shown that the calculated indices do not correspond with the visual observation. CIE, therefore, established a technical committee (TC) dealing only with the question of LED color rendering. It is hoped that this TC will be able to come up with some recommendations soon and, therefore, that LED manufacturers can tailor the emission spectra of their LEDs for optimal efficacy and color rendering. Until this happens, however, the published color rendering indices, especially the General Color Rendering Index (R_a) have to be handled carefully. The color appearance of objects under LED illumination might deviate considerably from what one expects based on the R_a index.

Application of LEDs

LEDs are the solution for practically every signaling application. Figure 8 shows an LED traffic light. Here it is also shown that, despite the fact that LEDs have extremely long life expectancy, the failure of the electronic circuitry can lead to unpleasant dark spots and rows in the display.

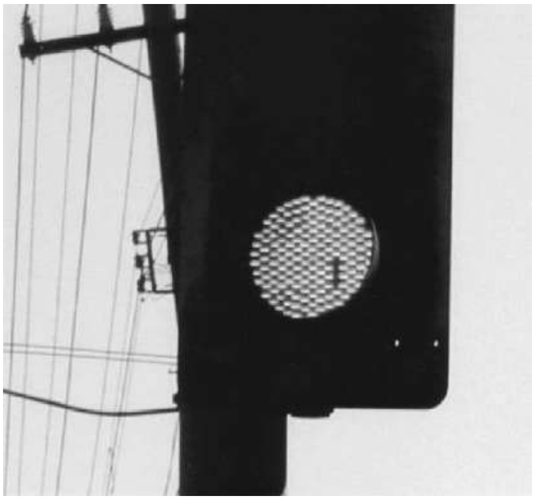

Figure 8 Green LED traffic light. Due to an electronic failure some LEDs stay dark.

LEDs are increasingly used in large active posters and displays. In such applications the highly saturated color of LED light is an advantage: LED displays can be very bright and emit saturated colors.

LEDs have an advantage in every situation where colored light has to be produced. Thus, they become the preferred automotive stoplights, dashboard lights, and are widely used in other signaling applications on the roads, at airfields, and in other traffic situations; much decorative lighting is now using LEDs.

Experiments are underway to find the best applications for white LEDs. Their drawback at the moment is still that their single unit power is below ten watts, as we often need ten to a hundred times larger units. For such applications, some LED clusters have been built, but these are still not competitive with other light sources. At places where the ruggedness of the LEDs is of an advantage, white LED lights are used – for example, in aircraft reading lights. One area where we will probably soon find higher-intensity white LEDs is car head-lamps. Both the efficacy and the luminous flux output of the single LED units is increasing continuously. Applications, unbelievable only a few years ago, can now be realized efficiently with LEDs. Thus it is very difficult to foresee in which applications – beyond signaling – LEDs will first supersede other light sources, but it is certain that within the next decade we will see many new applications of LED sources.

List of Units and Nomenclature

Chromatic adaptation	Adaptation by stimuli in which the dominant effect is that of different relative spectral distributions.
Chromaticity	Partial description of the color stimulus, describing only chromatic aspects and not the 'intensity' (the luminance) of the stimulus. In light sources, chromaticity is an important descriptor, as the level of illumination can be changed by changing the distance between the source and the illuminated object. The luminance of the object will depend on the illumination and the reflection (transmission) characteristics of the object.
Color	In color science one distinguishes between color (perception), characteristic of visual perception that can be described by attributes of hue, brightness (or lightness) and colorfulness (or saturation or chroma), and color stimulus, a specification of the optical stimulus in terms of operationally defined values, such as three tristimulus values, or luminance, and chromaticity (or dominant wavelength and saturation).
Color rendering	Measure of the degree to which the psychophysical color of an object illuminated by the test illuminant conforms to that of the same object illuminated by the reference illuminant, suitable allowance having been made for the state of chromatic adaptation.
Correlated color temperature	The temperature of the Planckian radiator whose perceived color most closely resembles that of a given stimulus at the same brightness and under specified viewing conditions.
Efficacy (luminous efficacy of a source)	Quotient of the luminous flux emitted by the power consumed by the source.
Luminous efficiency	Ratio of radiant flux weighted according to $V(\lambda)$, the spectral luminous efficiency of the human eye, to the corresponding radiant flux.
Spectral luminous efficiency (of the human observer)	Ratio of the radiant flux at wavelength λ_{m} to that at wavelength λ, such that both radiations produce equally intense luminous sensations under specified photometric conditions and λ_{m} is chosen so that the maximum value of this ratio is equal to 1.

See also

Incoherent Sources: Lamps. **Lasers:** Semiconductor Lasers.

Further Reading

Fukuda M (1991) *Reliability and Degradation of Semiconductor Lasers and LEDs*. Boston, MA: Artech House.

Rudden MN and Wilson J (1993) *Elements of Solid State Physics*. Chichester, UK: Wiley.

Sze SM (1985) *Semiconductor Devices: Physics and Technology*. New York: Wiley.

Wenckebach WT (1999) *Essentials of Semiconductor Physics*. Chichester, UK: Wiley.

Žukauskas A, Shur MS and Gaska R (2002) *Introduction to Solid-State Lighting*. New York: Wiley.

M

MAGNETO-OPTICS

Contents

Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping
Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering

Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping

H Pascher, Universität Bayreuth, Bayreuth, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A complete understanding of the band structure of a semiconductor close to the Fermi energy requires information on the fundamental energy gap, the effective masses and the effective g-factors of the valence and conduction band. Magneto-optical methods in principle allow one to determine all these quantities; they are particularly well suited to measure the spin splittings of the energy bands.

Faraday rotation is the rotation of the plane of polarization of linearly polarized light propagating through a sample parallel to an external magnetic field. The magneto-optical Kerr effect is the analog for reflection. Though in general it yields much smaller angles of rotation than the Faraday effect, it has to be applied in spectral ranges where the investigated sample is not transparent.

If a semiconductor is excited by circularly polarized light with photon energies not too high above the energy gap a spin polarization in the conduction band can be obtained (optical pumping). A small transverse magnetic field reduces the degree of this spin polarization (the Hanle effect). Owing to the spin polarization of the photo-excited conduction electrons the photoluminescence radiation is partly circularly polarized. Thus mounting the sample in a microwave resonator and applying a longitudinal magnetic field, electron spin resonance transitions can be detected optically by monitoring the degree of polarization of the photoluminescence (ODCESR: optically detected conduction electron spin resonance). Via a 'nuclear Hanle effect' mediated by the hyperfine interaction, nuclear magnetic resonance transitions can also be detected optically (ODNMR). ODCESR and ODNMR are summarized under ODMR (optically detected magnetic resonance).

In modern optics nonlinear optical effects have gained particular importance. Since electron spin resonances in semiconductors are Raman-active transitions, coherent Raman methods such as CARS (coherent anti-Stokes Raman scattering) or CSRS (coherent Stokes Raman scattering) are of great importance for the determination of spin splittings of energy bands and the investigation of magnetic structures in semiconductors.

In the present contribution the physical basis for these methods will be discussed and typical examples for their application will be presented.

For Faraday rotation and CARS the systems from which these examples will be taken are magnetic or dilute magnetic semiconductors. In these materials paramagnetic ions like Mn^{++} or Eu^{++} are incorporated. There are two types of exchange interactions: (i) between the magnetic ions which in most cases causes an antiferromagnetic coupling; and (ii) between the ions and free carriers which causes a strong modification of the spin splittings of the energy bands with respect to nonmagnetic materials. Thus the Faraday effect or CARS are particularly well-suited methods for the investigation of such systems.

As examples of optical pumping and ODMR, results obtained with III–V semiconductors will be presented, which show the most important applications of these experimental methods.

Faraday Rotation and Magneto-optical Kerr Effect

Basic Principles

As already pointed out in the Introduction, the Faraday effect is the rotation of the plane of polarization of linearly polarized light traversing a sample parallel to a magnetic field. The angle of rotation Θ_{Far} is proportional to the length l of the sample and to the strength B of the magnetic field:

$$\Theta_{\mathrm{Far}} = VlB \qquad [1]$$

where V is called Verdet's constant.

For a phenomenological understanding of the effect one can consider plane polarized light as a superposition of right- and left-hand circularly polarized components. If the indices of refraction are different for both components, then at the end of the sample, when the two components are recombined, a phase shift occurs which manifests itself as a rotation of the polarization plane. The angle resulting from this consideration is

$$\Theta_{\mathrm{Far}} = \frac{l\omega}{2c}(n_- - n_+) \qquad [2]$$

where ω is the frequency, c the velocity of light, and n_- and n_+ are the refractive indices for left- and right-hand polarized light, respectively. If one neglects ellipticity of the light behind the sample only the real parts of the complex refractive indices $\tilde{n}_\pm$ are important,

$$\tilde{n}_\pm = n_\pm - \mathrm{i}\kappa_\pm \qquad [3]$$

The imaginary parts, however, are needed to understand the magneto-optical Kerr effect which occurs in the light reflected from the sample surface. According to Fresnel's formula for vertical incidence the field amplitude of the reflected beam is:

$$\tilde{E}^{\mathrm{r}}_\pm = \frac{1 - \tilde{n}_\pm}{1 + \tilde{n}_\pm}\tilde{E}^{\mathrm{i}}_\pm = \tilde{r}_\pm \tilde{E}^{\mathrm{i}}_\pm \qquad [4]$$

where the superscript 'r' refers to the reflected beam and 'i' to the incident beam. The complex reflection coefficient can be split into an amplitude and a phase factor:

$$\tilde{r}_\pm = r_\pm \exp(\mathrm{i}\psi_\pm) \qquad [5]$$

By comparing eqns [4] and [5] the phase shifts during reflection can be expressed in terms of the refractive index:

$$\psi_\pm = \arctan\frac{2\kappa_\pm}{1 - n_\pm^2 - \kappa_\pm^2} \qquad [6]$$

Finally the Kerr angle of rotation is obtained from the phase shifts in an analogous way to the Faraday rotation:

$$\Theta_{\mathrm{Kerr}} = \tfrac{1}{2}(\psi_- - \psi_+) \qquad [7]$$

It is seen that both effects can be understood in terms of the refractive index. In a simple oscillator model the refractive index has the form given in eqn [8]:

$$\tilde{n}_\pm^2 = \tilde{\varepsilon}_\pm(\omega, B) = \varepsilon_\infty + \sum_{i,j}\frac{F^\pm_{i,j}}{\hbar\omega^\pm_{i,j} - \hbar\omega - \frac{1}{2}\mathrm{i}\Gamma^\pm_{i,j}} \qquad [8]$$

with $\hbar\omega^\pm_{i,j}$ representing the transition energies between the energy levels i and j in σ^+ and σ^- polarization, respectively, and $F^\pm_{i,j}$ and $\Gamma^\pm_{i,j}$ are the respective oscillator strength and linewidth parameter.

From eqn [8] it is seen that when tuning the frequency, ω, resonant enhancements of the rotation angles are expected whenever the frequency corresponds to an allowed electronic transition, e.g., an exciton in a wide-gap semiconductor. This is the reason why the Faraday and magneto-optical Kerr effect can yield information on interband transition energies. In the case of such a resonance, the refractive index will be dominated by the particular term in the sum, corresponding to that transition. By a Taylor expansion one can then show that the difference of the refractive indices for σ^+ and σ^- polarization and thus the angles of rotation (see eqns [2] and [7]) are proportional to the difference of the transition energies for right- and left-hand circularly polarized light, and this difference depends on the spin splittings of the valence and conduction band. So we understand why the most important information obtained from Faraday or magneto-optical Kerr effect studies is on the spin splittings of the energy bands.

A quantum mechanical description of the relation between the material properties and the magneto-optical effects beyond the simple oscillator model given here is published by Boswarva *et al.* and the reader is referred to this paper for further studies. A matrix formalism to describe the effects for a multilayer system has been developed by Nies and Kessler.

Examples of Experimental Results

The straightforward method to measure Faraday rotation is to put the sample between polarizers and sweep the magnetic field. Then the transmitted intensity is proportional to a $\cos^2$ function. Subsequent maxima demonstrate a rotation by 180°. There have been numerous successful investigations done in this way, but the possibilities of detecting small angles of rotation are very limited in this method. Therefore differential methods have been developed either using a polarizing beam splitter, or a photo-elastic modulator which periodically switches the state of polarization. Rotation angles of the order of 10^{-4} rad can be detected in this way.

Figure 1 shows the Faraday rotation of an epitaxial film of $Pb_{0.933}Eu_{0.067}Te$ (thickness 11μm) for two temperatures. The plot demonstrates behavior typical for a diluted magnetic semiconductor in its paramagnetic phase. The spin splittings of the valence and conduction band and accordingly the Faraday rotation are proportional to the magnetization of the sample which behaves like a Brillouin function. (The lines without noise are fitted Brillouin functions.)

Examples of the application of the magneto-optical Kerr effect to the investigation of semiconductors, particularly multilayer systems, are published, e.g., by Pohlt *et al.* and Krenn *et al.*

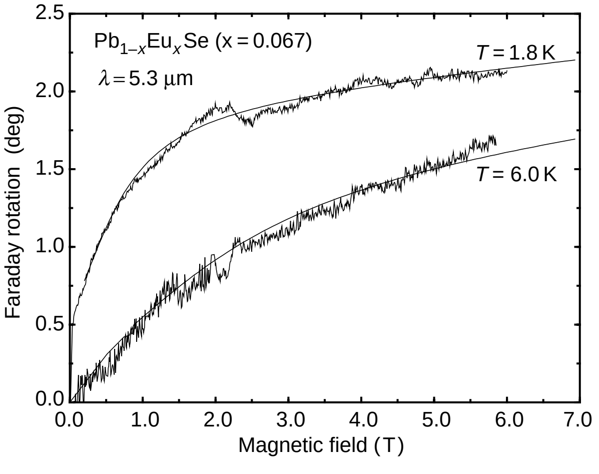

Figure 1 Faraday rotation as a function of magnetic field of a PbEuSe epitaxial film for two temperatures. Lines with noise: experimental recordings; lines without noise: fitted Brillouin functions.

Coherent Raman Scattering

Basic Principles

A more direct method for the determination of spin splittings of energy bands is provided by Raman scattering. There is a huge number of investigations of spin transitions in semiconductors using spontaneous Raman scattering (for a review of spontaneous Raman scattering in dilute magnetic semiconductors see the article by Ramdas and Rodriguez). However, there are several advantages in the spectroscopic application of the nonlinear Raman techniques compared to spontaneous Raman scattering:

(i) The scattered radiation is emitted as a collimated beam which is more easily detectable than the light spontaneously scattered into the entire solid angle. This advantage is particularly important for the investigation of narrow-gap semiconductors which have to be investigated by infrared light, a spectral range where the detectors are less sensitive than in the visible.
(ii) Fluorescence can be strongly suppressed by spatial filtering.
(iii) The spectral resolution is determined by lasers instead of a classical spectrometer.
(iv) Raman scattering cross-sections can be calculated from the intensity of a collimated beam without taking care of the aperture of the detecting system.

The coherent Raman techniques are special cases of optical four-wave mixing where two laser beams with frequencies ω_L and ω_S ($\omega_L > \omega_S$) are superimposed in the sample. Due to nonlinear interactions radiation with new frequencies is generated, among them:

$$\omega_{CARS} = 2\omega_L - \omega_S \qquad [9]$$

If one identifies ω_L with the frequency of the pump laser and ω_S with that of the scattered Stokes radiation of a spontaneous Raman experiment, then ω_{CARS} is the frequency of the anti-Stokes radiation. For the intensity of the radiation generated at ω_{CARS} one finds:

$$I_{CARS} = \frac{9}{16c_0^4\varepsilon_0^2}\frac{\omega_{CARS}^2}{n_{CARS}n_S n_L^2}\left|\chi^{(3)}\right|^2 l^2 I_S I_L^2 \times\left[\frac{\sin(\overrightarrow{\Delta k}\,\vec{l}/2)}{\overrightarrow{\Delta k}\,\vec{l}/2}\right]^2 \qquad [10]$$

where c_0 is the velocity of light, l is the length of the sample, I denotes the intensity, and n the refractive index at the respective frequency. $\chi^{(3)}_{kjkj}$ is the third-order nonlinear susceptibility of the scattering medium and

$$\overrightarrow{\Delta k} = 2\vec{k}_{\mathrm{L}} - \vec{k}_{\mathrm{S}} - \vec{k}_{\mathrm{CARS}} \qquad [11]$$

is the phase mismatch. $\overrightarrow{\Delta k} = 0$ can be achieved by noncollinear alignment of the laser beams. However, for very thin epitaxial layers with $l \ll \pi/\Delta k$ the phase factor in eqn [10] can be approximated by 1.

The information on the investigated material is contained in the nonlinear susceptibility. It can be calculated by time-dependent perturbation theory taking into account a ground state $|a\rangle$ and intermediate states $|b\rangle$, $|c\rangle$ and $|d\rangle$. It turns out that $\chi^{(3)}$ is proportional to a sum over all states of the system as intermediate states; in the sum 24 terms like the one given in eqn [12] have to be added:

$$\frac{\mu_{ijkl}}{(\omega_{ba} - \omega_{\mathrm{L}} - \mathrm{i}\Gamma_{ba})(\omega_{ca} - \omega_{\mathrm{L}} + \omega_{\mathrm{S}} - \mathrm{i}\Gamma_{ca})(\omega_{da} - \omega_{\mathrm{CARS}} - \mathrm{i}\Gamma_{da})} \qquad [12]$$

μ_{ijkl} is a product of four dipole matrix elements between states $|a\rangle{:}\,|b\rangle$; $|b\rangle{:}\,|c\rangle$; $|c\rangle{:}\,|d\rangle$ and $|d\rangle{:}\,|a\rangle$ with field polarizations as given by the indices. $\omega_{\alpha\beta} = (E_\alpha - E_\beta)/\hbar$ and $\Gamma_{\alpha\beta}$ is the linewidth of the respective transition. Resonances of the susceptibility are found by inspection of the denominators. $\chi^{(3)}$ exhibits Raman-like resonances whenever $\hbar(\omega_{\mathrm{L}} - \omega_{\mathrm{S}})$ is equal to the excitation energy of a Raman allowed transition (see the second factor in the denominator of eqn [12]). If in this case the laser beams are polarized according to the Raman selection rules then μ_{ijkl} is not zero. For spin transitions in a semiconductor in a magnetic field (free electron spin flip as well as transitions within the magnetic ions) this is the case if in the Voigt configuration ($\vec{k}_{\mathrm{L,S}} \perp \vec{B}$) the beams with frequencies ω_{L} and ω_{S} are polarized perpendicular to one another. In the case of a spin transition $\omega_{ca} = g^*\mu_{\mathrm{B}}B$ (g^* being the effective Landé-factor and μ_{B} the Bohr magneton). Thus holding the frequencies ω_{L} and ω_{S} fixed and tuning the magnetic field a resonance in I_{CARS} will be observed from the position of which the effective g-factor can be obtained. Using tunable dye lasers, of course, the laser frequencies can be tuned as well.

Close to the resonance the real part of the susceptibility becomes small and the imaginary part is not negligible. Furthermore, one-photon resonances may occur if the laser frequency ω_{L} corresponds to the energy of a real electronic transition, e.g., an allowed interband transition. If one-photon and Raman-like resonances occur simultaneously one talks about resonance CARS.

As discussed in the literature $\chi^{(3)}$ can be split into a resonant part χ_{r}, summing up all the resonant terms, and a nonresonant part χ_{nr}. In a semiconductor there are contributions to χ_{nr} due to free carriers and a background due to bound valence electrons. The background contribution in general is real. As can be seen from eqn [10],

$$I_{\mathrm{CARS}} \propto \left|\chi^{(3)}_{\mathrm{CARS}}\right|^2 = |\chi_{\mathrm{nr}}|^2 + |\chi'|^2 + |\chi''|^2 + 2\cdot\chi_{\mathrm{nr}}\chi' \qquad [13]$$

χ' and χ'' being the real and imaginary parts of the resonant contribution. Depending on the relative strength of resonant and nonresonant contributions to $\chi^{(3)}$ the last term in eqn [13] may cause complicated lineshapes which make it difficult to find the correct transition energy in spectroscopic applications of CARS. If more than one resonance is not completely resolved the lineshape is even more complicated and a careful analysis has to be performed to get the correct resonance positions and linewidths of the transitions.

Examples of Experimental Results

Figure 2 shows a scheme of a CARS setup suitable to measure spin transitions in a semiconductor. The scheme is drawn for backscattering geometry; forward detection is accomplished by directing the incoming beams onto the sample from the opposite side to that plotted.

Two dye lasers (frequencies ω_{L} and ω_{S}) are pumped by a frequency-doubled Nd:YAG laser. A Glan–Thomson prism combines the two beams. The monochromator is set to $2\omega_{\mathrm{L}} - \omega_{\mathrm{S}}$. The iris and the polarizer PZ3 reduce stray light. The experiment is performed either at fixed magnetic field and tuning one of the dye lasers (with the monochromator swept accordingly) or by keeping the laser frequencies fixed and sweeping the magnetic field. The latter method is also applicable in the mid-infrared using step-tunable CO or CO_2 lasers.

As an example, in Figure 3 a spectrum obtained from a CdMnTe alloy crystal is reproduced. There are two peaks observed due to the paramagnetic resonance of the Mn 3d electrons and due to the spin resonance of photo-excited free electrons in the conduction band. The inset shows the paramagnetic resonance on an extended scale. The linewidth of 0.5 cm^{-1} is not limited by the experimental resolution, which in the present case is about a factor of 10 better.

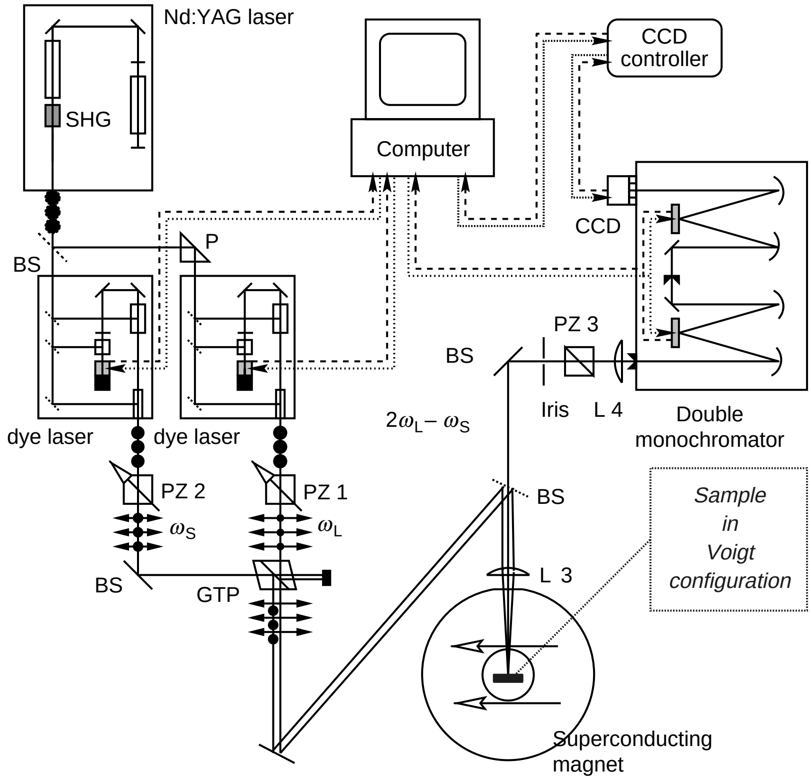
Nd:YAG laser
SHG
Computer
CCD controller
CCD
BS
P
PZ 3
BS
Iris
L 4
Double monochromator
$2\omega_L - \omega_S$
dye laser
dye laser
PZ 2
PZ 1
ω_S
ω_L
BS
BS
GTP
L 3
Sample in Voigt configuration
Superconducting magnet

Figure 2 Scheme of the experimental arrangement for CARS. Bold full lines: light beams; dashed and dotted lines: data connections; BS: beam splitter; PZ1,2: Pockels cells for rotation of linear polarization; PZ3: polarizer; GTP: Glan–Thomson prism for superposition of beams polarized orthogonally to each other; L: lenses.

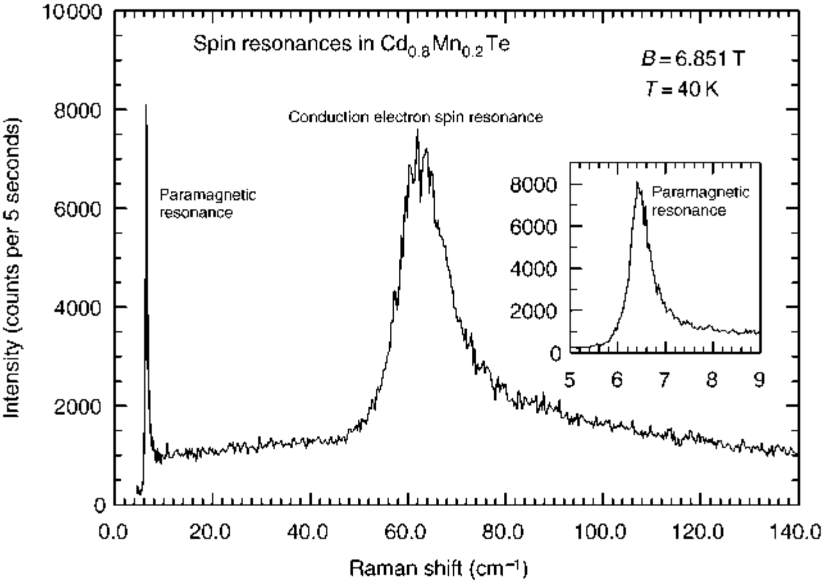

Figure 3 Coherent Raman intensity of a CdMnTe bulk crystal as a function of frequency difference $\omega_L - \omega_S$. Inset: Region of paramagnetic resonance in an extended scale. Adapted from Rupprecht R, Müller B, Pascher H, *et al.* (1998) Coherent Raman scattering from magnetic excitations in diluted magnetic semiconductors: Bulk crystals of CdMnTe. *Physical Review B* 58: 16123. Copyright (1998) by the American Physical Society.

Optical Pumping

The basic principles of optical pumping have been developed by Kastler. A material (gas or solid) is excited by absorption of circularly polarized light. Lampel was the first to demonstrate the possibility of optical spin orientation of free electrons in a semiconductor. The method has the most important applications in zincblende materials. Thus it is explained here with the example of GaAs.

Principles of Optical Spin Orientation

Most of the III–V and II–VI semiconductors have direct gaps at the center of the Brillouin zone. The conduction band (cb) is doubly degenerate with spin. The valence band consists of the heavy hole band (hh; $m_j = \pm 3/2$), the light hole band (lh; $m_j = \pm 1/2$) and the split-off band (so; $m_j = \pm 1/2$), each doubly degenerate. In bulk material hh and lh are degenerate at $k = 0$. The transition probabilities for interband transitions between hh and cb are three times larger than for transitions between lh and cb. Due to the selection rules, e.g., for the absorption of σ^+ circularly polarized light, transitions $-3/2 \rightarrow -1/2$ and $-1/2 \rightarrow +1/2$ are allowed, the first having three times higher probability. The electronic spin polarization P is defined as:

$$P = \frac{n_+ - n_-}{n_+ + n_-} \qquad [14]$$

where n_+ and n_- are the densities of electrons with the two spin orientations with respect to the direction of propagation of light. These considerations yield $P_{\mathrm{i}} = -0.5$ as the initial spin polarization under excitation with σ^+ circularly polarized light with a photon energy between the fundamental gap and the energy difference from the spin–orbit split valence band to the conduction band. Due to relaxation processes the actual electronic steady state polarization P is smaller than P_{i}:

$$P = P_{\mathrm{i}} \frac{T_1}{T_1 + \tau} \qquad [15]$$

with T_1 being the spin relaxation time, which in zero magnetic field is equal to the transverse relaxation time T_2, and τ is is the lifetime of conduction electrons.

The spin polarization of the electrons can be detected by analyzing the degree of circular polarization of the luminescence light which is defined as:

$$\Pi = \frac{I_+ - I_-}{I_+ + I_-} \qquad [16]$$

where I_+ and I_- are the intensities of σ^+ and σ^- polarized light, respectively. Since for emission due to free-electron/free-hole or free-exciton recombination the same transition probabilities as for absorption are obtained, one finds:

$$\Pi = 0.5\,P = 0.25 \frac{T_1}{T_1 + \tau} \qquad [17]$$

In the case of quantum wells, the lh-hh degeneracy is lifted due to different confinement energies. Then the maximum polarization of cb → hh as well as for cb → lh luminescence light is 0.5.

Of course these considerations are only true for intrinsic or p-type materials where the number of photo-excited electrons is high compared to the equilibrium number of electrons.

Hyperfine Interaction of Spin-Polarized Electrons and Atomic Nuclei

In GaAs three isotopes exist with spin 3/2, namely ^{69}Ga, ^{71}Ga and ^{75}As. Owing to the small energy separation of the nuclear spin levels in moderate magnetic fields, in thermal equilibrium no significant nuclear spin polarization exists even at a temperature of 2 K. However, the hyperfine interaction between electrons and nuclei allows so-called flip-flop processes, that is a simultaneous reversal of electronic and nuclear spins. This causes a considerable dynamic spin polarization of the nuclei if the electronic spin system is polarized due to optical pumping. Under most experimental conditions and when nuclear spin relaxation mechanisms other than via the hyperfine interaction are negligible, the mean nuclear spin is approximately:

$$\langle I \rangle = \frac{I(I+1)}{S(S+1)} \frac{P}{2} \qquad [18]$$

where I and S are the operators for nuclear and electron spin. These dynamically polarized nuclei act on the electrons and shift their spin resonance magnetic field by an effective nuclear field (the Overhauser shift). This shift is proportional to $\langle I \rangle$ and to the probability of finding the electrons at the nuclear sites, and is inversely proportional to g^*. In GaAs due to the small value of g^* this field can reach several tesla.

Effect of a Constant Transverse Magnetic Field – The Hanle Effect

In a transverse (with respect to the direction of propagation of light) magnetic field the spin-oriented electrons precess around the field with an angular frequency $\Omega = g^* \mu_B B/\hbar$. If the spin lifetime T_S $(1/T_S = 1/T_2 + 1/\tau)$ is long enough, this precession diminishes the average spin polarization and consequently the degree of circular polarization of the photoluminescence according to:

$$\Pi(B) = \Pi(0)\frac{1}{1+\Omega^2 T_S^2} \qquad [19]$$

This is a Lorentzian with a width at half height of $\Delta B = \hbar/(g^* \mu_B T_S)$.

It is seen that by measuring $\Pi(B)$ the spin lifetime can be determined if g^* is known. Together with eqn [17] the spin and interband lifetime can be determined in a steady-state experiment. If, on the other hand, T_S is known from a time-resolved experiment, g^* can be found from a Hanle effect measurement.

Optically Detected Conduction Electron Spin Resonance (ODCESR)

Optical spin orientation of conduction electrons is the presupposition for ODCESR. Suppose now, a longitudinal magnetic field $\mathbf{B}_0$ is applied on the sample ($\mathbf{B}_0$ parallel to $\mathbf{k}$, the direction of propagation of light, which is parallel to the direction of the optically oriented spins). If the field is not too high the considerations of optical spin orientation remain unchanged. In higher fields the spin polarization may be modified due to the energy splitting of the spin levels. In this case the spin polarization is enhanced or diminished, depending on the direction of the field parallel or antiparallel to $\mathbf{k}$. If now an additional transverse microwave field B_1 of frequency ω_0, which fulfills the resonance condition $\hbar\omega_0 = g^* \mu_B B_0$, is applied, transitions between the two spin levels will occur which resonantly decrease the spin polarization. This decrease manifests itself in a decrease of the degree of circular polarization of the photoluminescence.

The method has several advantages compared to conventional ESR. Microwave photons with frequency of about 10^{10} Hz change the state of polarization of optical photons with frequency of about 10^{14} Hz. Thus the sensitivity is that of the optical detection which is much higher than in conventional ESR. Even electrons in a single quantum well or heterostructure can be detected by the optical method. There is also the possibility to distinguish between electrons in different quantized levels in heterostructures due to the different spectral position of their photoluminescence. Another advantage is that the electrons are photo-excited; the sample does not need to be n-doped. High-purity samples can be investigated which exhibit narrower linewidths.

Optically Detected Nuclear Magnetic Resonance (ODNMR)

With the presuppositions provided in the preceeding sections the basic principle of ODNMR is easily understood. As in ODCESR we assume that an external magnetic field $\mathbf{B}_0$ is applied parallel to $\mathbf{k}$. Perpendicular to it a radio-frequency field $\mathbf{B}_2$ with frequency ω is applied. The NMR resonance frequency is $\omega_0 = \gamma_n \cdot B_0$ with γ_n being the gyromagnetic ratio of the nucleus under consideration. In the frame rotating with $\boldsymbol{\omega}$ around the direction of $\mathbf{B}_0$ an effective field $\mathbf{B}_{\text{eff}} = \mathbf{B}_0 - \boldsymbol{\omega}/\gamma_n + \mathbf{B}_2$ acts on the nuclei. The nuclear magnetization is oriented parallel to $\mathbf{B}_{\text{eff}}$. It is seen that in resonance $\mathbf{B}_{\text{eff}}$ is parallel to $\mathbf{B}_2$, which means perpendicular to $\mathbf{k}$. Via the hyperfine interaction this nuclear field acts on the electrons, diminishing their spin polarization via the Hanle effect. So the nuclear magnetic resonance can be detected as a reduction of the degree of circular polarization of the photoluminescence.

The simplified model given above does not take into account several effects which influence the lineshape, position and intensity of the observed resonance lines.

Examples of Experimental Results for Optical Pumping and Optically Detected Magnetic Resonances

These experiments can be performed at liquid helium temperature in a standard photoluminescence setup. In addition one has to provide the possibility of applying transverse (Hanle effect) or longitudinal (ODCESR, ODNMR) magnetic fields. For ODCESR the sample is mounted in a microwave resonator with optical access; for ODNMR a radio-frequency coil is wound around the sample. The state of circular polarization of the exciting laser light has to be controlled. The measured quantity is the sign and degree of circular polarization of the luminescence light at a certain wavelength. The sensitivity can be enhanced by periodically σ^+/σ^--modulating either the exciting light or the detection. Which one of these possibilities is used depends on whether or not a high nuclear spin polarization is wanted.

As pointed out above the dynamically polarized nuclei, particularly in GaAs, strongly affect the

ODCESR resonant field. To get reliable values for the effective g-factor, the nuclear spin polarization must be completely suppressed. This can be achieved by σ^+/σ^--modulation of the exciting light with a modulation period much shorter than the nuclear spin relaxation time. But due to imperfections, e.g., if the intensities of σ^+ and σ^- light or their time intervals are not equal, residual optical fields may remain. To get precise values for g^* in addition the NMR has to be saturated by a suitable radio-frequency field. On the other hand, the strong Overhauser shift of the CESR can be useful for the determination of nuclear spin relaxation times. To perform such experiments the nuclei are first polarized by optical pumping with fixed circular polarization and afterwards, for the detection, σ^+/σ^--modulation is switched on.

For the optical detection of NMR a high nuclear spin polarization is favorable. Thus for those experiments one uses fixed polarization of the exciting beam and modulates the detected light.

Figure 4 shows examples of results obtained with a p-channel AlGaAs/GaAs single heterostructure. In this sample due to the doping profile a triangular potential well at the interface in the valence band of the GaAs layer is built, in which quantized hole levels exist. The luminescence band shown in Figure 4a between 821 nm and 825 nm is due to recombination of electrons with holes in these quantized states. For the experiments shown in the other panels of the figure the spectrometer is set at a position within this band. The other lines are due to excitons and impurities and are not interesting for what follows.

In Figure 4b the results of a Hanle experiment are reproduced. The line without noise is the fitted Lorentzian. The fit yielded $\tau = 8.1$ ns and $T_1 = 4.7$ ns.

In Figure 4c a typical ODNMR spectrum of the same sample is plotted. One observes resonances due to ^{75}As at 1.36 MHz, ^{69}Ga at 1.9 MHz and ^{71}Ga at 2.4 MHz. In the inset the ^{75}As resonance is plotted in an extended scale. A quadrupole splitting into three components is clearly resolved. This result demonstrates that the tetrahedral symmetry at the nuclear site is broken either due to local defects or due to the (rather small) lattice mismatch close to the AlGaAs/GaAs interface.

Finally Figure 4d shows the ODCESR spectrum. For the experiment the exciting beam was σ^+/σ^--modulated and in addition during the sweep the NMR of all three nuclei was permanently saturated to make sure that no Overhauser shift

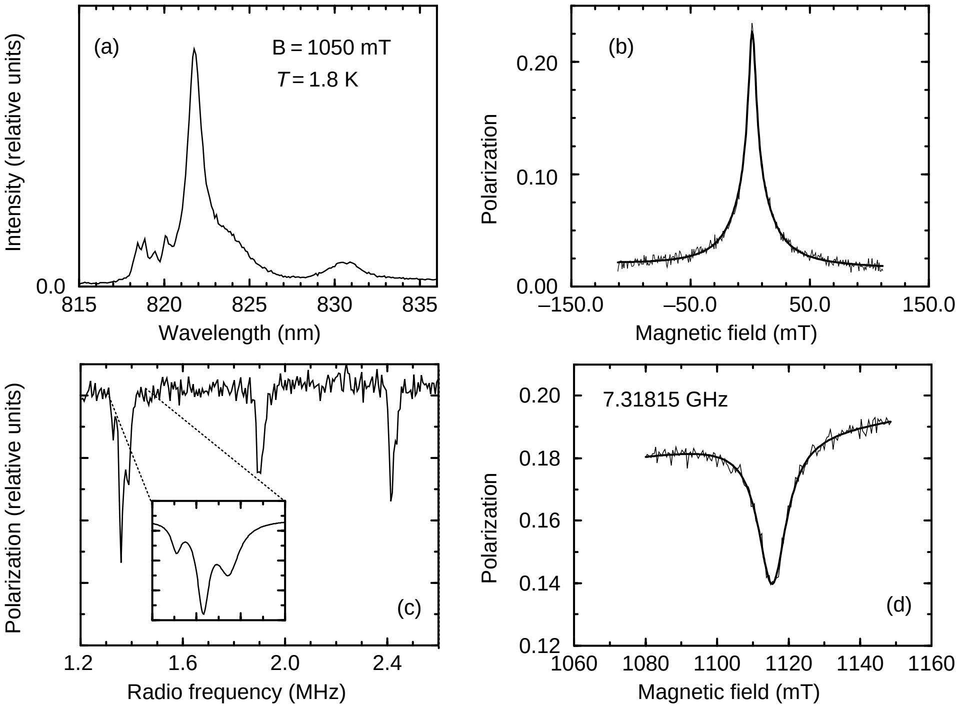

Figure 4 (a) Photoluminescence spectrum of a p-channel AlGaAs/GaAs heterostructure; (b) Hanle curve (luminescence polarization as a function of transverse magnetic field); (c) ODNMR recording; (d) ODCESR recording. Adapted from Schreiner M, Pascher H, Denninger G, *et al.* (1997) Nuclear spin relaxation in AlGaAs/GaAs heterostructures observed via optically detected magnetic resonance (ODMR) experiments. *Solid State Communications* 102: 715. Copyright (1997), with permission from Elsevier.

affects the result. The g-factor calculated from this plot is $|g^*| = 0.468$, which is larger than that found for bulk GaAs. This value was also used for fitting the Hanle curve.

List of Units and Nomenclature

Angle	[rad]
Angular frequency ω	[rad s^{-1}]
Bohr magneton μ_B	9.2742×10^{-24} J T^{-1}
Degree of circular polarization of light Π	
Electric field E	[V m^{-1}]
Gyromagnetic ratio γ	[T^{-1} s^{-1}]
Imaginary unit i	
Intensity I	[W m^{-2}]
Magnetic displacement B	[T]
Permittivity of free space ε_0	8.8542×10^{12} V s A^{-1} m^{-1}
Planck's constant $\hbar$	1.0546×10^{-34} J s
Refractive index n	
Speed of light in vacuum c_o	2.9979×10^8 m s^{-1}
Spin polarization P	
Susceptibility χ	

CARS	coherent anti-Stokes Raman scattering
CSRS	coherent Stokes Raman scattering
ODCESR	optically detected conduction electron spin resonance
ODMR	optically detected magnetic resonance
ODNMR	optically detected nuclear magnetic resonance

Acknowledgments

The author thanks W. Herbst, B. Müller, R. Rupprecht and M. Schreiner for making unpublished results available.

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Physics:** Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Boswarva IM, Howard RE and Lidiard AB (1964) Faraday effect in semiconductors. *Proceedings of the Royal Society* A269: 125–141.

Dyakonov MI and Perel VI (1984) Theory of optical spin orientation of electrons and nuclei in semiconductors. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 11–71. Amsterdam: North-Holland.

Eesley GL (1981) *Coherent Raman Spectroscopy.* Oxford: Pergamon Press.

Furdyna JK (1988) Diluted magnetic semiconductors. *Journal of Applied Physics* 64: R29–R64.

Geist F and Pascher H (1996) Resonance CARS in narrow gap semiconductors. *Journal of Raman Spectroscopy* 27: 289–295.

Geist F, Herbst W, Mejia-Garcia C, *et al.* (1997) Magneto-optical investigations of Eu-based diluted magnetic lead chalcogenides. *Physical Review B* 56: 13042.

Hermann C and Weisbuch C (1984) Optical detection of conduction electron spin resonance in semiconductors and its application to $\mathbf{k}\cdot\mathbf{p}$ perturbation theory. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 463–508. Amsterdam: North-Holland.

Krenn H, Herbst W, Pascher H, *et al.* (1999) Interband Faraday and Kerr rotation and magnetization of $Pb_{1-x}Eu_xTe$ in the concentration range $0 \times \sim 1$. *Physical Review B* 60: 8117.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy.* San Diego: Academic Press.

Nies R and Kessler FR (1989) Coherent and incoherent magnetooptical reflection and transmission effects of multilayer structures. *Phys. Stat. Sol. (a)* 111: 639–654.

Oh E, Bartholomew DU, Ramdas AK, Furdyna JK and Debeska U (1988) Interband Faraday effect in CdMnSe. *Physical Review B* 38(18): 13183–13190.

Paget D and Berkovits VL (1984) Optical investigation of hyperfine coupling between electronic and nuclear spins. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 404ff. Amsterdam: North-Holland.

Pascher H, Geist F, Springholz G, *et al.* (1997) Coherent anti-Stokes Raman scattering in diluted magnetic IV–VI epilayers and superlattices. In: Shen SC, Tang DY, Zheng GZ and Bauer G (eds) *Narrow Gap Semiconductors*, pp. 269–278. Singapore: World Scientific.

Pohlt M, Herbst W, Pascher H, Faschinger W and Bauer G (1998) Magneto-optical Kerr effect in short-period CdTe/MnTe superlattices. *Physical Review B* 57(16): 9988–9994.

Ramdas AK and Rodriguez S (1988) Raman scattering in diluted magnetic semiconductors. In: Furdyna JK and Kossut J (eds) *Diluted Magnetic Semiconductors, Semiconductors and Semimetals*, vol. 25, pp. 345–412. London: Academic Press.

Schreiner M, Pascher H, Denninger G, *et al.* (1997) Nuclear spin relaxation in AlGaAs/GaAs heterostructures observed via optically detected resonance (ODMR) experiments. *Solid State Communications* 102: 715–720.

Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The addition of a magnetic field to an optical study of a semiconductor generally improves the specificity of the experiment, but in particular affects primarily the electrons rather than the lattice so that the magneto-optical properties are determined to a large degree by the electrons alone. The major interest has been in the determination of the electronic energy band structure of semiconductors, both in their bulk and low-dimensional (two-, one-, and zero-dimensional) quantum forms. The magneto-optical experiments fall mainly into two categories according to whether the effects are associated with intraband (within a single conduction or valence band) or interband (valence to conduction band) transitions; there are also hybrid (i.e., higher-order) effects such as spin-flip Raman scattering or two-photon absorption where both inter- and intraband transitions are involved in a two-stage process. Interband transitions occur typically in the visible to mid-infrared (MIR) region, and intraband transitions in the MIR to far-infrared (FIR) region.

Theoretical Background

The most common band structure involved in semiconductor magneto-optics is the so-called quasi-Ge model, which gives an accurate description of group IV, and zincblende (III–V and II–VI) semiconductors, so it is convenient to describe the theory in terms of this model. In this model the small terms associated with inversion asymmetry are neglected to a first approximation. At the Brillouin zone center (i.e., at $k = 0$) the conduction band has s-like symmetry and the triply degenerate valence band p-like symmetry. In the presence of strong spin–orbit coupling, part of this degeneracy is raised resulting in the doubly degenerate (at $k = 0$) heavy and light mass valence bands and a so-called spin–orbit split-off band. The reader should go to the general reviews quoted under 'Further Reading' for the equivalent coverage of other common systems such as the wurzite II–VI and the lead salt systems.

The dynamics of a free electron subject to an external magnetic field have been given by a variety of authors, but the most universally used treatment is that of Landau. For the motion of an electron in a semiconductor we treat first the case of a simple parabolic energy band. For the case of the magnetic field, B, along the z-direction, k_z remains a good quantum number, but k_x and k_y are subsumed into the new magnetic orbital (Landau) quantum number, n. The original three-dimensional energy band, $E = \hbar k^2/2m^*$, is coelesced into a series of one-dimensional Landau sub-bands in the k_z-direction (where the motion along k_z remains unchanged):

$$E_n^{\pm}(k_z) = (n + 1/2)\,\hbar\omega_c + \frac{\hbar k_z^2}{2m^*} \qquad [1]$$

The cyclotron frequency is $\omega_c = eB/m^*$ and μ_B is the Bohr magneton. The resulting electron energy eigenvalues at $k_z = 0$ in an external magnetic field (i.e., Landau sub-band extrema) are referred to as Landau levels, LLs. In practice each state is doubly degenerate with spin which results in an additional term $\pm g^* \mu_\mathrm{B} B$ on the right-hand side, where g^* is the effective g-factor. (Note that in the presence of strong spin–orbit coupling and small m^*, there is a corresponding g^* which becomes negative and substantially different from the free-electron g-factor; the largest value for a binary elemental compound is that for InSb, where $g^* \approx -50$.) The situation for a simple parabolic conduction band near $k = 0$ is shown in Figure 1a,b. Intraband optical absorption associated with direct transitions between successive Landau sub-bands is referred to as cyclotron resonance (CR) absorption. There are corresponding Landau sub-bands in the valence band, and direct transitions between valence and conduction sub-bands give rise to the so-called interband magneto-optical (IMO) effects.

Expressions for the density of states (DOS) in three, two, and one dimensions in the absence of a magnetic field are given elsewhere ('semiconductor basics' – Pidgeon). In the presence of a magnetic field the three-dimensional energy band is reduced to one-dimensional sub-bands so that the energy dependence of the DOS has the one-dimensional form with the energy given by eqn [1]. Performing this substitution and multiplying by a term for the degeneracy of the Landau levels (LLs) gives the following expression for the total DOS per unit volume:

$$g(E) = \frac{(2m^*)^{1/2}}{4\pi^2\hbar}\,\frac{eB}{\hbar}\sum_n \left[E - (n + 1/2)\,\hbar\omega_c\right]^{-1/2} \qquad [2]$$

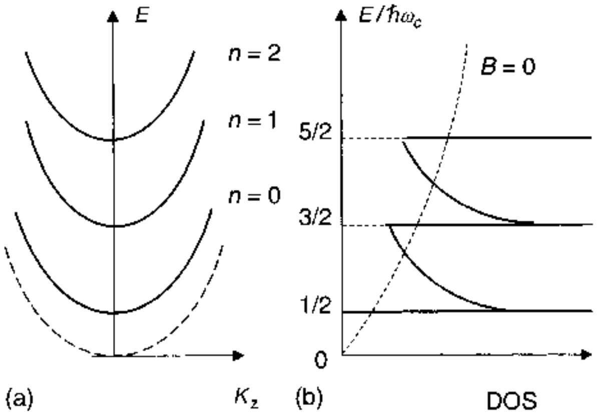

Figure 1 (a) Landau sub-bands according to eqn [1] with the magnetic field applied in the z-direction. The dashed line shows the parabolic conduction band for k along k_z and $B = 0$. (b) Schematic outline of the density of states (DOS) for free electrons in a magnetic field. It comprises the sum of identical contributions from each Landau sub-band. The density of states with $B = 0$ is shown dashed.

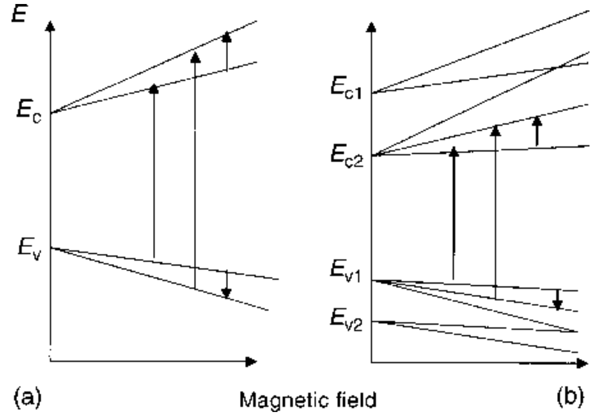

Figure 2 Fan plots of the conduction and valence band Landau levels versus magnetic field for simple parabolic bands: (a) a three-dimensional semiconductor; (b) quasi-two-dimensional systems. Some CR and interband magneto-optical transitions are shown by vertical arrows.

Thus, the DOS for Landau sub-bands is quite different from the zero-field DOS, and is a series of square-root singularities with divergences at the energies $(n + 1/2)\hbar\omega_c$. The two cases are illustrated schematically in Figure 1b. Since the optical absorption strength for both interband and intraband transitions is proportional to the joint density of states for the initial and final states, this result shows the important advantage of the use of a magnetic field in semiconductor spectroscopy because it means that the absorption lines for CR or IMO transitions are strongly peaked at the sub-band extrema (i.e., at $k_z = 0$). This allows accurate measurement of LL energy differences for transitions observed in a single experiment, and hence, by comparison to theoretical models, leads to the best determination of fundamental energy-band parameters. The basic LL selection rule can be shown straightforwardly from time-dependent perturbation theory to be $\Delta n = 0$, $\Delta m_s = 0$ for IMO transitions and $\Delta n = 1$, $\Delta m_s = 0$ for CR absorption, where m_s is the spin quantum number (i.e., $\pm\frac{1}{2}$, for spin up and spin down). The cyclotron resonance absorption occurs only for one sense of circularly polarized light, but the IMO effect can occur for all polarizations. The choice then of magnetic field direction, crystal orientation, and polarization of the optical radiation allows extremely detailed determination of the energy band parameters for particular crystal symmetries.

The data are often summarized in a so-called fan diagram of the energy levels, or transition energies, versus magnetic field. Such a plot is shown for hypothetical conduction and valence band LLs in Figure 2a, together with interband and CR transitions. The states are labeled by the Landau quantum number, with selection rules $\Delta n = 0$ and $\Delta n = 1$ for interband and CR absorption transitions, respectively (or $\Delta n = -1$ for CR emission). If the energies of the transitions themselves are plotted versus the magnetic field, then the intercept of the fan plots at $B = 0$ gives the zero-field energy gap. Note that in the presence of excitonic effects (see below) the intercept will give the energy gap plus the exciton binding energy.

For a strictly two-dimensional system the DOS per unit area is independent of energy: $g(E) = m^*/\pi\hbar^2$. In the case of quasi two-dimensional confinement in the z-direction the motion in the z-direction is quantized, and the total energy becomes (*see* **Semiconductor Physics:** Quantum Wells and GaAs-Based Structures):

$$E = E_i + \frac{\hbar^2}{2m^*}(k_x^2 + k_y^2) \qquad [3]$$

Equation [3] describes a set of confinement sub-bands (parabolas induced by geometric, as opposed to magnetic, confinement), the bottoms of which are at the energies E_i, where i is the sub-band index. The energies, E_i, depend on the shape of the confining potential. The DOS per unit area is in this case a series of step functions of height $m^*/\pi\hbar^2$. From the point of view of magneto-optics, the main effect of the quantum confinement is to suppress the k_z motion, so that in the presence of an external magnetic field along the z-direction, the motion is completely quantized. The energy eigenvalues are given by:

$$E_{n,i} = E_i + (n + 1/2)\,\hbar\omega_c \qquad [4]$$

The total DOS becomes:

$$g(E) = \frac{m^* \omega_c}{\pi \hbar} \delta[E - (n + 1/2)\, \hbar \omega_c] \qquad [5]$$

a series of delta-functions located at the energy positions of the LLs. The DOS is obviously more singular in two dimensions than in three dimensions, and leads to substantially sharper magneto-optical and magneto-transport effects. For confined structures the selection rules for interband and intersub-band transitions are identical to the bulk case for the electromagnetic (em) radiation polarized in the xy plane provided that the external magnetic field is along the confinement (z-) direction. The only difference is that the conservation of momentum in the z-direction ($\Delta k_z = 0$) is replaced by a selection rule on the sub-band index, $\Delta i = 0$. A fan plot for this configuration is shown schematically in Figure 2b. For each confinement sub-band transition there is a set of LL transitions. When the energy positions are plotted against B, this leads to a set of Landau 'fans', each of which extrapolates to the sum of a valence plus conduction confinement sub-band energy at $B = 0$. Allowed transitions are shown for the lowest two confinement sub-bands in Figure 2b. Again this picture may be complicated in the presence of excitonic effects at low temperature.

Turning again to the bulk semiconductor picture, eqns [1]–[5] apply for a single parabolic energy band and are valid for the conduction band in the quasi-Ge model. In actual semiconductors further complications arise. Most importantly the triply degenerate p-like valence band (heavy hole, light hole, and split-off band) must be treated by so-called degenerate perturbation theory. In the quasi-Ge model the conduction and three valence bands are treated together as one coupled band system using the so-called $\mathbf{k} \cdot \mathbf{p}$ method (described elsewhere by Zawdzki). With spin this results in an 8×8 matrix equation which has to be diagonalized for the eigenvalues and eigenvectors of the system. The spin–orbit coupling results in so-called complex, mixed-spin valence bands and anomalously spaced low quantum number LLs (the so-called quantum effects). The overall energy band scheme for InSb, near the fundamental direct gap in the center of the Brillouin zone for conduction and (heavy and light mass) valence bands is shown schematically in Figure 3, for no magnetic field and for an applied magnetic field. Each LL is doubly degenerate with spin, the a-set representing the so-called spin-up and the b-set the spin-down states.

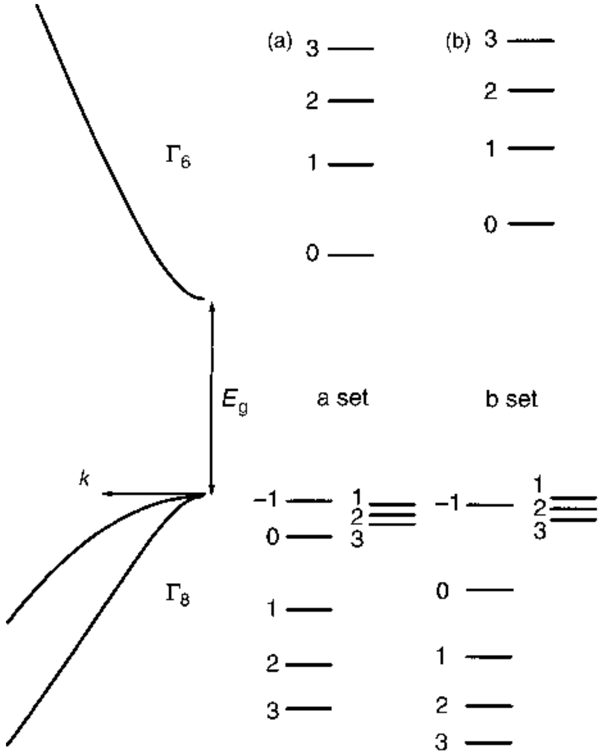

Figure 3 (a) Schematic diagram of the conduction and valence bands for InSb (symmetry labels Γ_6 and Γ_8 respectively) at $B = 0$. (b) Computed Landau levels for the conduction and valence bands of InSb at 4 K ($E_g = 0.23$ eV) with $B = 5$ T along the [100] direction, grouped according to the so-called a- and b-spin sets (see text).

Experimental

The first serious use of magneto-optics to study effective mass and energy band structure of semiconductors was in the microwave cyclotron resonance (i.e., intraband) investigations of Ge and Si in the early 1950s. Classically it can be shown that the ratio of the radiation power absorbed, $\boldsymbol{P}$, to power absorbed in zero magnetic field, $\boldsymbol{P}_0$, is given by:

$$\frac{\boldsymbol{P}}{\boldsymbol{P}_0} = \frac{1 + (\omega^2 + \omega_c^2)\tau^2}{[1 + (\omega_c^2 - \omega^2)\tau^2]^2 + 4\omega^2\tau^2} \qquad [6]$$

where τ is the collision time of the carriers. This expression for P/P_0 describes cyclotron resonance absorption and is plotted as a function of frequency in Figure 4 for different values of the parameter $\omega_c\tau$. Clearly the peak absorption occurs at $\omega = \omega_c$, namely at cyclotron resonance. Also evident is the very important requirement for observation of a clear resonance that $\omega_c\tau > 1$; this condition corresponds physically to a travel time between collisions of the particle of at least $1/2\pi$ of a revolution so that the particle can interact for a sufficiently long time

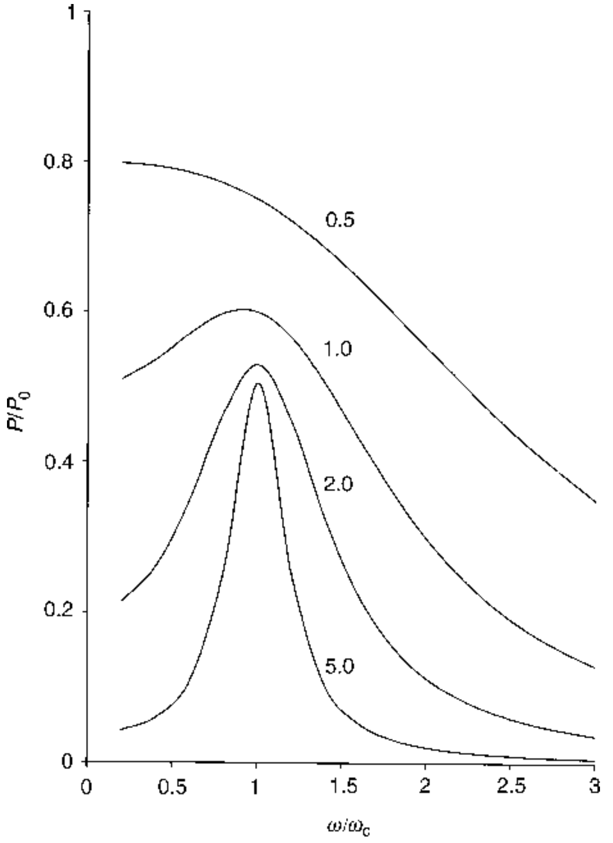

Figure 4 Ratio of the radiation power absorbed to the power absorbed at zero magnetic field, P/P_0, plotted as a function of ω/ω_c according to eqn [6], for different values of $\omega_c\tau$ from 0.5 to 5.0.

with the electromagnetic wave to absorb energy. For pure materials and low temperatures (where the condition $\omega_c\tau > 1$ is achievable) this is the best technique for obtaining the details of the electron and hole band parameters of a semiconductor near $k = 0$. The most thorough and successful studies of this type have been achieved in pure Si, Ge, and InSb where (with the aid of optical excitation to populate the levels shown in Figure 3) all possible electron, light and heavy hole cyclotron resonances have been resolved. In this way the valence band quantum effects, described above, have been observed in detail. The technique is equally applicable in low-dimensional systems where CR of both electrons and holes has been resolved in a large number of n- and p-type semiconductor heterostructures and quantum wells.

In interband – either absorption or photoluminescence (PL) – spectroscopy at low temperatures, the photo-excited electrons and holes interact and form excitons (*see* **Semiconductor Physics:** Excitons). In many cases the excitonic state may be described in the so-called envelope function approximation as a hydrogenic system, in which the electron and hole are bound by their mutual Coulomb attraction. The energy spectrum of such a hydrogenic exciton is given by

$$E_n = \frac{\mu e^4}{2\hbar^2\varepsilon^2 n^2} \qquad [7]$$

where n is the principal quantum number ($n = 1, 2, 3, \ldots$, but note that the symbol n is also used for the Landau quantum number above), μ is the electron–hole reduced effective mass, and ε is the background dielectric constant. In low-dimensional (i.e., confined) semiconductor systems of the type described above the degeneracy of the valence bands at $k = 0$ is removed, and two types of excitons are observed; these are referred to as 'heavy' and 'light' hole excitons in the literature, where heavy and light refer to the effective masses along the confinement direction. These excitons can move freely (in the remaining two unconfined directions) or be bound to an impurity or defect. In these systems exciton absorption may be observed involving not only the ground-state excitons, but also excitons associated with higher confinement sub-bands. This is shown for the so-called asymmetric quantum well system $\mathrm{Al}_y\mathrm{Ga}_{1-y}\mathrm{As}/\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}/\mathrm{GaAs}$ in Figure 5a. The data are taken both by the direct PL method (where the exciting beam has an energy much larger than the bandgap and the electrons and holes de-excite instantaneously to the bottom of the bands before recombining as exiton recombination radiation), and so-called PL excitation spectroscopy (where the exciting beam is a tunable laser, tuned into resonance with the successive exciton resonances).

The interband Landau transitions can be observed either in PL or as excitations in a magneto-absorption or magnetoreflection experiment. Excitons are affected differently by the application of an external magnetic field depending whether the ratio $\gamma = \hbar\omega_c/2E_B$ (where E_B is the exciton binding energy with $n = 1$ above) is less than or greater than one. In the so-called magnetic limit where $\gamma \gg 1$, the excitonic ground state moves approximately as $\frac{1}{2}\hbar\omega_c$, i.e., linearly with B. In this regime the magnetic field energy is the dominant term in the exciton Hamiltonian. The various states of the exciton (ground and excited states) become associated with the interband LL transition, and exhibit an energy dependence that is nearly identical to the free carrier LL transitions. The energies of the various excitonic transitions are shifted down with respect to the corresponding LL transitions by an amount equal to the binding energy of the excitonic state to that LL. Typical PL and PLE (photoluminescence excitation)

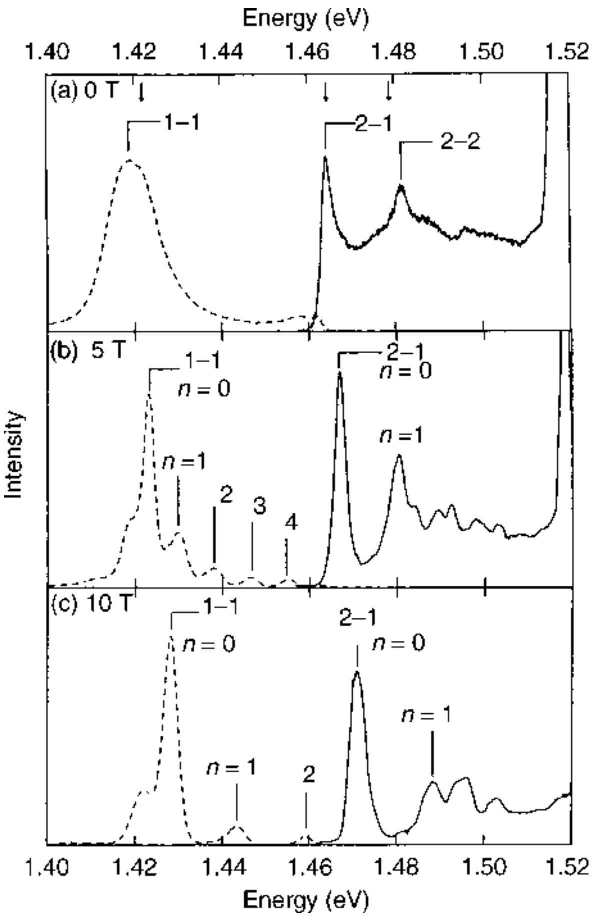

Figure 5 Interband transition data obtained in zero magnetic field (a) and fields of 5 T (b) and 10 T (c) from asymmetric quantum wells. The dashed lines indicate data obtained from PL and the solid lines data from PLE. The arrows in (a) indicate the calculated interband transition energies 1 → 1, 2 → 1, and 2 → 2, the latter being clearly resolved in PLE (the asymmetric wells mean that transitions with $\Delta i \neq 0$ are allowed). In (b) and (c) the development of the spectra into LL transitions is clearly seen, although beyond 1.48 eV the analysis is obviously extremely complicated by the overlap of LL transitions associated with all three of the above sub-bands.

spectra for interband magneto-optical transitions are shown in Figure 5b,c for a fixed applied magnetic field in the asymmetric quantum well system $\mathrm{Al}_y\mathrm{Ga}_{1-y}\mathrm{As}/\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}/\mathrm{GaAs}$.

Spin Flip Raman Scattering

In addition to the interband transitions discussed above, Raman scattering of intersub-band transitions between conduction or valence band LLs can also be studied. It provides a versatile means of probing band edge parameters of electrons and holes in semiconductors since the excitation (normally a laser pump) is not confined to the resonance frequencies under investigation, i.e., any pumping energy below the band edge will do. Electronic Raman scattering is then allowed between LLs via virtual interband transitions. In the presence of spin–orbit coupling strong spin-flip processes via electric dipole transitions become allowed giving rise to the so-called spin-flip Raman (SFR) cross-section. This is a single-particle process in which electrons or holes interacting with the radiation change their spin state. This process can be made to go stimulated particularly easily when the pump radiation is resonant with the bandgap of the semiconductor, forming the spin-flip Raman laser. The most important practical case has probably been that of InSb, resonantly pumped near 5 μm with the CO laser. In this case the small effective mass of a narrow-gap semiconductor in the presence of substantial spin–orbit coupling leads to both an extremely large anomalous effective g-factor ($g^* \sim -50$) – giving a large tuning range, $\Delta E = \hbar\omega_{\mathrm{SF}} = g^*\beta B$ – and a strong cross-section which is larger than that of the free electron by a factor $(m_0/m^*)^2$. Efficient cw operation is possible with threshold powers less than 50 mW.

SFR scattering has enabled the observation of a great variety of excitations in nonmagnetic and dilute magnetic semiconductors, including: free electrons and holes, electrons bound to donors and holes bound to acceptors, scattering from bound magnetic polarons, and scattering due to SF transitions within the Zeeman multiplets of paramagnetic ions. In addition to pure SF transitions, combined resonances with combinations such as $\omega_{\mathrm{CR}} \pm \omega_{\mathrm{SF}}$ have also been detected. The process has also been used in a similar manner to electron spin resonance to study coherent processes, including Raman photon echo spectroscopy in semiconductors.

See also

Magneto-Optics: Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Häfele HG (1991) Spin-flip raman scattering. In: Landwehr G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 207–275.

Petrou A and McCombe BD (1991) Magnetospectroscopy of confined semiconductor systems. In: Landwehra G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 679–775.

Pidgeon CR (1980) Free carrier optical properties of semiconductors. In: Balkanski M (ed.) *Optical Properties of Solids. Handbook of Semiconductors*, pp. 223–328. Amsterdam, The Netherlands: North-Holland.

Zawadzki W (1991) Intraband and interband magneto-optical transitions in semiconductors. In: Landwehr G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 483–512.

MATERIALS CHARACTERIZATION TECHNIQUES

Contents

$\chi^{(2)}$
$\chi^{(3)}$

$\chi^{(2)}$

R C Eckardt, Cleveland Crystals, Inc.,
Highland Heights, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Techniques used to measure second-order susceptibility $\chi^{(2)}$ or nonlinear optical coefficients are described in this article, and some measured values are quoted. These are the parameters that describe the property of a material that is responsible for second-harmonic generation and other three-wave nonlinear optical interactions. These other interactions include sum- and difference-frequency generation in which two waves combine to generate a third wave of a frequency that is the sum or difference of the first two waves. Also included are optical parametric generation and optical parametric oscillation, in which a single pump wave generates two resulting waves with frequencies that sum to the frequency of the pump wave. The measurement techniques involve measuring the efficiency of nonlinear frequency conversion by these processes.

The measurements require a great deal of care and precision, and it is advisable to perform a comparison with previous measurements and obtain correlation with accepted standards whenever possible. A great deal of effort continues to be expended by researchers to establish correlations between measurements and provide additional values that are accepted as standards. A number of techniques have been used for these measurements sometimes yielding results that disagree. There are, however, values derived from early measurements that have proven accurate. In these measurements it is necessary to accurately control and characterize the polarization, power, spectral distribution, and spatial and temporal intensity distribution of the incident radiation and accurately measure the frequency-converted light, which can be many orders of magnitude smaller.

The development and availability of single-longitudinal-mode and nearly diffraction-limited lasers has greatly reduced the ambiguity and simplified increased accuracy of measurements of second-order nonlinear optical coefficients. The improved optical quality of nonlinear-optical crystals has also been an important component of the increased accuracy of measurement. Measurement techniques have also improved, providing absolute measurements of optical intensity over wide dynamic ranges at different wavelengths. Until the early 1990s, there was a range of measured values for nonlinear optical coefficients of common nonlinear materials, and numerous definitions were in use. Caution is required when consulting early publications on this topic, and some of this ambiguity is carried over to present time. Different definitions vary by factors of 2 and the permittivity of free space, $\varepsilon_0 \cong 8.854 \times 10^{-12}$ C/N × m^2. It is important that the definition of a nonlinear optical coefficient is clear before a value is used. This can be done through a statement of the nonlinear electric polarization or the coupled equations for nonlinear frequency conversion.

The standard techniques that have been used to measure nonlinear optical coefficients are covered in detail in publications listed in the Further Reading at the end of this article. Brief descriptions of some of these techniques, such as Maker-fringe and wedge, optical parametric fluorescence, and phase-matched second-harmonic generation, will be presented in this article. Each of these techniques has a unique set of advantages and disadvantages. There was some lively controversy over the technique that provided the most accurate measurements and those measurements that should be dismissed. Measurements based on phase-matched second-harmonic generation had the advantage of being most similar to practical application of nonlinear optical frequency conversion. However, the large divergence and multiple-longitudinal-mode operation of early lasers made this method unreliable. Even relative comparisons of nonlinear optical coefficients between different materials were inaccurate because highly structured laser pulses had different conversion characteristics due to differences in bandwidth acceptance, group velocity walkoff, and birefringent walkoff in different materials.

Nonlinear optical coefficients of improved accuracy were required for analysis of optical harmonic conversion in the large, well-characterized laser systems used in inertial-confinement fusion research. The development of these systems provided a standardization on the value of $d_{36} = 0.39\ \mathrm{pm\,V^{-1}}$ ($1\ \mathrm{pm} = 10^{-12}\ \mathrm{m}$) for second-harmonic generation of neodymium laser radiation in the material potassium dihydrogen phosphate (KH_2PO_4 or KDP). The development of injection-seeded Q-switched lasers provided single-longitudinal-mode laser output that allowed the phase-matched second-harmonic techniques to be extended to provide more accurate absolute measurements and relative comparisons on KDP and a number of other materials. Correlation was achieved on a larger number of relative and absolute measurements of nonlinear optical coefficients to provide some systematic agreement. It is not well understood why the earlier measurements referenced to optical parametric fluorescence produced values that were higher by about 60%. Measurement techniques have been further refined, and more exacting measurements are leading agreement, using different methods of measurement on an increasing number of nonlinear optical materials.

Definitions

The definition of the nonlinear optical coefficient can be specified in a number of ways. Examples given here are how the coefficient relates to the nonlinear electric polarization in a material, the coupled equations for nonlinear frequency conversion, and expressions describing second-harmonic generation with a monochromatic plane-wave fundamental.

The second-order optical nonlinearity is reported as both the second-order susceptibility $\chi^{(2)}$ and the second-order nonlinear optical coefficient d, which usually differ in definition by a factor of two. These are third-rank tensors that relate products of components of the electric field vector E_j and E_k to the vector components of the nonlinear polarization $P_{\mathrm{i}}^{\mathrm{NL}}$. The second-order susceptibility typically is used to express instantaneous values:

$$P_{\mathrm{i}}^{\mathrm{NL}}(t) = \varepsilon_0 \sum_{j,k} \chi_{ijk}^{(2)} E_j(t) E_k(t) \qquad [1]$$

whereas the nonlinear optical coefficient more often is used to relate the vector amplitudes of frequency components of these quantities. The electric field at the fundamental angular frequency ω can be expressed as

$$E_j(\omega, \mathbf{r}, t) = \frac{1}{2}\{E_j(\omega) \exp \mathrm{i}(\mathbf{k}_\omega \cdot \mathbf{r} - \omega t) + c.c.\} \qquad [2]$$

and the nonlinear polarization at the harmonic frequency 2ω can be expressed as

$$P_{\mathrm{i}}^{\mathrm{NL}}(2\omega,\mathbf{r},t) = \frac{1}{2}\left\{P_{\mathrm{i}}^{\mathrm{NL}}(2\omega)\exp \mathrm{i}(2\mathbf{k}_\omega\cdot\mathbf{r} - 2\omega t) + c.c.\right\} \qquad [3]$$

The intensity is given by $I=(nc\varepsilon_0/2)|E|^2$. The relationship between the complex amplitudes expressed in terms on the nonlinear optical coefficient tensor is

$$P_{\mathrm{i}}^{\mathrm{NL}}(2\omega) = \varepsilon_0 \sum_{j,k} d_{ijk}(-2\omega;\omega,\omega) E_j(\omega) E_k(\omega) \qquad [4]$$

When waves of three different frequencies are involved, the expression becomes

$$P_{\mathrm{i}}^{\mathrm{NL}}(\omega_3) = 2\varepsilon_0 \sum_{j,k} d_{ijk}(-\omega_3;\omega_2,\omega_1) E_j(\omega_2) E_k(\omega_1) \qquad [5]$$

Expressing the coefficient as $d_{ijk}(-\omega_3;\omega_1,\omega_2)$ allows for dispersion and indicates that fields at frequencies ω_1 and ω_2 interact to produce a nonlinear polarization at frequency ω_3, the sum of the first two frequencies. The relationship between the second-order electrical susceptibility and the second-order nonlinear optical coefficient is

$$\chi_{ijk}^{(2)} = 2 d_{ijk} \qquad [6]$$

The dispersion of nonlinear optical coefficients is approximated by the use of Miller's delta:

$$\delta_{ijk} = \frac{d_{ijk}(-\omega_3;\omega_2,\omega_1)}{(n_{\mathrm{i}}^2(\omega_3)-1)(n_j^2(\omega_2)-1)(n_k^2(\omega_1)-1)} \qquad [7]$$

The indices of refraction in the direction of the electric field components $E_m(\omega_n)$ are $n_m(\omega_n)$ for $m=i,j,k$ and $n=1,2,3$. There is some theoretical and experimental justification for treating the components of Miller's delta as constants. Earlier definitions of Miller's delta included multiplication by ε_0, the permittivity of free space.

The nonlinear optical coefficient is usually reduced to a scalar quantity, the effective nonlinear coefficient or d_{eff}, that relates only the specific vector components of the fields involved in the interaction. The effective nonlinear optical coefficient will be discussed in more detail later. The coupled equations for second-harmonic generation using d_{eff} are

$$\frac{\mathrm{d}E_{2\omega}(z)}{\mathrm{d}z} = \mathrm{i}\frac{\omega d_{\text{eff}}}{n_{2\omega}c}E_\omega(z)E_\omega(z)\exp(-\mathrm{i}\,\Delta kz) \quad [8]$$

$$\frac{\mathrm{d}E_\omega(z)}{\mathrm{d}z} = \mathrm{i}\frac{\omega d_{\text{eff}}}{n_\omega c}E_\omega(z)E_\omega^*(z)\exp(\mathrm{i}\,\Delta kz) \quad [9]$$

Here ω is the angular frequency of the fundamental wave; z is distance in the direction of propagation, $n_{2\omega}$ and n_ω are the refractive indices of the harmonic and fundamental respectively; and c is the speed of light in free space. The complex electric field amplitudes of the harmonic and fundamental, $E_{2\omega}(z)$ and $E_\omega(z)$, respectively, are assumed to be slowly varying functions of position compared with the oscillation of the carrier waves $\exp\mathrm{i}(\mathbf{k}_{2\omega}\cdot\mathbf{r}-2\omega t)$ and $\exp\mathrm{i}(\mathbf{k}_\omega\cdot\mathbf{r}-\omega t)$, with $\mathbf{k}_{2\omega}$ and $\mathbf{k}_\omega$ the wavevectors of harmonic and fundamental; the asterisk indicates complex conjugate. The quantity Δk is the wavevector mismatch and is given by $\Delta k = k_{2\omega} - 2k_\omega$. These are simplified expressions that apply to nearly monochromatic plane waves. Two solutions to the coupled equations are

$$I_{2\omega}(z) = \frac{2\omega^2 d_{\text{eff}}^2 l^2 I_\omega^2(0)}{n_{2\omega}n_\omega^2 c^3\varepsilon_0}\frac{\sin^2(\Delta kz/2)}{(\Delta kz/2)^2} \quad [10]$$

for insignificant depletion of the fundamental amplitude and

$$I_{2\omega}(z) = I_\omega(0)\tanh^2\sqrt{\frac{2\omega^2 d_{\text{eff}}^2 z^2 I_\omega(0)}{n_{2\omega}n_\omega^2 c^3\varepsilon_0}} \quad [11]$$

for the case of perfect phase matching $\Delta k = 0$ and allowing depletion of the fundamental. The quantity $I_{2\omega}(z)$ is the harmonic intensity after propagating a distance z in the nonlinear material, and $I_\omega(0)$ is the fundamental intensity at the input surface of the nonlinear material where $I_{2\omega}(0) = 0$.

The above relationships for nonlinear polarization, the coupled equations for second-harmonic generation, or harmonic conversion of a monochromatic plane wave, each determine the definition of the second-order nonlinear optical coefficient.

Symmetry

Symmetry has an important role in the measurement of nonlinear optical coefficients. Indeed the second-order nonlinear optical coefficient must be identically zero if the material has inversion symmetry. This restricts materials to noncentrosymmetric crystals or isotropic materials in which inversion symmetry is removed by such techniques as application of an electric field or polymeric materials in which alignment of molecules is achieved by some poling process. The discussion here is limited to noncentrosymmetric crystals.

The nonlinear optical coefficient is a tensor of rank 3. Representation of the tensor in a rotated coordinate system is described by

$$d'_{ijk} = \sum_{l,m,n} R_{il}R_{jm}R_{kn}d_{lmn} \quad [12]$$

where R_{il} is the matrix that describes rotation of the initial coordinate system xyz to the new coordinate system $x'y'z'$. The representation of a vector r_l in the original coordinate system becomes $r'_{\mathrm{i}} = \sum_l R_{il}r_l$ in the new coordinate system. A required symmetry property of the nonlinear optical tensor, that exchange of the second two indices of a component does not change the value of the component $d_{ijk} = d_{ikj}$, allows the third-rank tensor to be written as a 3 by 6 matrix:

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{pmatrix} \quad [13]$$

with the components of the matrix related to the tensor components by

$$\begin{aligned} &d_{i1} = d_{i11} \quad d_{i4} = d_{i23} = d_{i32} \\ &d_{i2} = d_{i22} \quad d_{i5} = d_{i13} = d_{i31}, \; i = 1,2,3 \\ &d_{i3} = d_{i33} \quad d_{i6} = d_{i12} = d_{i21} \end{aligned} \quad [14]$$

A further property, tensor components with any permutation of the indices are equal $d_{ijk} = d_{kij}$, called Kleinman symmetry or Kleinman's conjecture, is often applied if there is no absorption at any of the wavelengths involved. Kleinman symmetry assures that the amount of energy depleted from the fundamental is equal to the amount of energy gained by the second harmonic. This reduces the number of independent components of the nonlinear optical matrix to 10: d_{11}; $d_{12} = d_{26}$; $d_{13} = d_{35}$; $d_{14} = d_{25} = d_{36}$; $d_{15} = d_{31}$; $d_{16} = d_{21}$; d_{22}; $d_{23} = d_{34}$; $d_{24} = d_{32}$; and d_{33}. Sometimes components equal, according to Kleinman symmetry, are measured individually, either as a check of the measurement or to check the validity of the conjecture.

Crystal symmetry places additional constraints on the components of the nonlinear optical matrix. These constraints determine the components that must be zero and determine additional relationships between the nonzero components. Matrices for selected crystal point groups are given in Table 1. Effective nonlinear optical coefficients can be found by a coordinate frame transformation in which the rotated axes align with the direction of propagation and the eigen polarizations. Effective nonlinear coefficients for a few crystal point groups are given in Table 2. More complete listings of both the reduced matrices and effective nonlinear optical coefficients are given elsewhere.

Table 1 Nonlinear coefficient matrices for selected crystal point groups

Orthorhombic $mm2$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{24} & 0 & 0 \\ d_{31} & d_{32} & d_{33} & 0 & 0 & 0 \end{pmatrix}$$

Hexagonal 6

$$\begin{pmatrix} 0 & 0 & 0 & d_{14} & d_{15} & 0 \\ 0 & 0 & 0 & d_{15} & -d_{14} & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$$

Tetragonal $\bar{4}2m$

$$\begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{25} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{36} \end{pmatrix}$$

Hexagonal $6mm$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$$

Trigonal $3m$ $m \perp X$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & -d_{22} \\ -d_{22} & d_{22} & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$$

Cubic $\bar{4}3m$

$$\begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{14} \end{pmatrix}$$

Measurements

Usually the symmetry of a nonlinear crystal is determined by techniques such as X-ray diffraction or surface morphology. Crystal symmetry determines the nonzero components of the second-order nonlinear optical tensor. Nonphase-matched second-harmonic-generation techniques, such as Maker-fringe and wedge measurements, allow selection of polarizations and propagation directions relative to the crystal lattice that isolates individual components of the tensor. Phase-matched measurement techniques usually require propagation in a direction for which an effective nonlinear optical coefficient applies and can involve more than one tensor component. Parametric fluorescence and sum- and difference-frequency-generation measurements are phase-matched processes that involve three wavelengths. All of the measurement techniques have been used with pulsed and continuous-wave laser output. In all of these methods, except parametric fluorescence, it is necessary to accurately characterize the incident radiation. Measurements relative to a standard material avoid some of the concerns of characterization of the fundamental radiation. Relative measurements are important in that they allow correlation between different sets of measurements. Measurement sets, that include both relative and absolute measurements on a number of materials, are useful in developing correlations that achieve more accurate compilations of values.

It is desirable to have an incident beam with a Gaussian-like transverse intensity distribution and close to diffraction limited propagation. Gaussian beams have advantages that they are most easily characterized and their propagation is described with relatively simple analytical methods. Accurate temporal characterization of the incident light is also essential. Here it is desirable to have stable single-mode laser operation to produce narrow bandwidth light that is well within phase-matching acceptance bandwidths and that has temporal structure slow enough to be easily characterized and slow enough

Table 2 Effective nonlinear optical coefficient for a few symmetries

$mm2$, a point group with biaxial birefringence		
$\phi = 0$, two waves $E \parallel Y$, one wave $E \parallel XZ$	$d_{\text{eff}} = d_{32} \sin\theta$	
$\phi = 90°$, two waves $E \parallel XY$, one wave $E \parallel Z$	$d_{\text{eff}} = d_{31} \cos^2\phi + d_{32} \sin^2\theta$	
$\phi = 90°$, two waves $E \parallel X$, one wave $E \parallel YZ$	$d_{\text{eff}} = d_{31} \sin\theta$	
Some crystal symmetry groups with uniaxial birefringence		
point group	two e rays and one o ray	two o rays and one e ray
$\bar{4}2m$	$d_{\text{eff}} = d_{14} \sin^2 2\theta \cos 2\phi$	$d_{\text{eff}} = -d_{14} \sin\theta \sin 2\phi$
$3m$	$d_{\text{eff}} = d_{22} \cos^2\theta \cos 3\phi$	$d_{\text{eff}} = d_{15} \sin\theta - d_{22} \cos\theta \sin 3\phi$
6 and $6mm$	$d_{\text{eff}} = 0$	$d_{\text{eff}} = d_{15} \sin\theta$

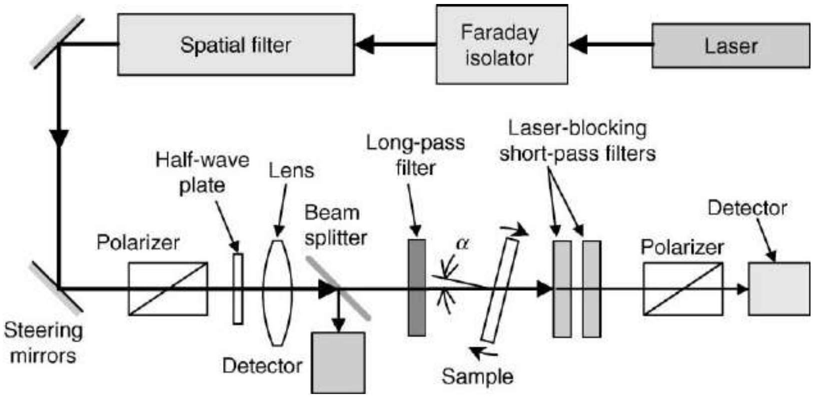

Figure 1 Schematic illustration of an experimental setup to record Maker-fringe measurements.

that group-velocity mismatch will be of no concern. Generating such laser output usually requires special techniques, such as injection-seeded pulse operation or stabilized single-frequency cw operation. Often spatial filters are required to improve the beam quality on the laser output. Even though these topics are only mentioned here, the techniques are important and demand much effort.

The analysis of measurements can also be complicated. Consider the Maker-fringe measurement technique (Figure 1). In this method a focused fundamental beam is incident on a thin slab of oriented nonlinear optical material for nonphase-matched second-harmonic generation. The thickness of the slab is large enough that the function $\sin(\Delta kz/2)$ has gone through many cycles. As the crystal is rotated both the path length in the material l and the wave vector mismatch Δk change, and $\sin^2(\Delta k \cdot l/2)$ oscillates between maxima and minima (Figure 2). The coherence length $L_c = \pi/\Delta k$ is normally small enough that conversion remains in the region of small depletion of the fundamental. Knowledge of the material dispersion and birefringence, required to calculate Δk, or L_c, can be obtained from the fringe spacing. Ideally, the incident fundamental beam will be focused to a spot size w_0, small enough to provide measurable harmonic conversion but large enough that the near-field approximation, in which diffraction is not significant, can be used for analysis of harmonic generation. In this case integration over the incident fundamental intensity distribution $I_\omega(r, z=0) = I_{\omega,0} \exp(-2r^2/w_0^2)$ yields an expression for the harmonic power:

$$W_{2\omega}(\alpha) = \frac{8\omega^2 d_{\text{eff}}^2(\alpha) W_\omega^2 L_c^2(\alpha)}{\pi^3 w_0^2 n^3 c^3 \varepsilon_0} \sin^2\left(\frac{\pi l(\alpha)}{2L_c(\alpha)}\right) \times \text{correction factors} \qquad [15]$$

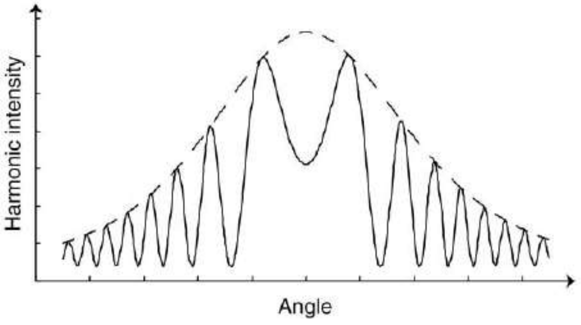

Figure 2 The envelope and spacing of the fringes allow determination of the second-order nonlinear optical coefficient and the coherence length.

where α is the angle of incidence of the fundamental radiation, $W_{2\omega}(\alpha)$ is the second harmonic power, and W_ω is the incident fundamental power. This equation applies to cw radiation. It is necessary to integrate over the temporal distributions for pulsed fundamental light. The effective nonlinear coefficient $d_{\text{eff}}(\alpha)$, the path length in the crystal $l(\alpha)$, and the coherence length $L_c(\alpha)$ change with the angle of incidence α. The correction factors include Fresnel losses at the crystal surfaces, harmonic generation at the surfaces, multiple reflections between the surfaces, absorption at either fundamental or harmonic wavelengths, and walkoff due to non-normal incidence and birefringence.

The wedge technique (Figures 3 and 4) also uses nonphase-matched second-harmonic generation. A wedged sample is translated through a focused fundamental beam, and the above equation applies with the angle α constant. The crystal length l varies with transverse position. It is important that the wedge angle and the fundamental beam spot size are chosen so that the entire beam fits well within a maximum of conversion and the beam is large enough to avoid divergence due to diffraction. Both the Maker-fringe

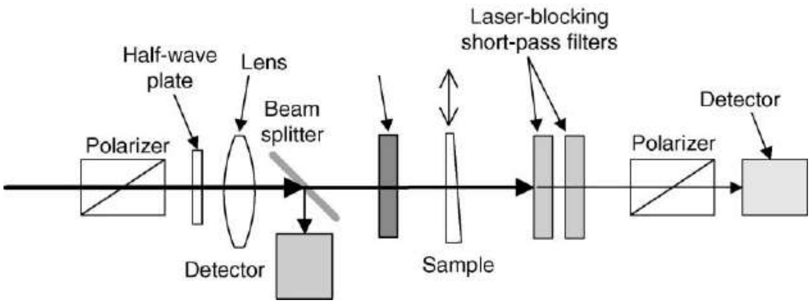

Figure 3 The setup for wedge measurements is similar to that used for Maker-fringe measurements but with a wedged sample translated in the fundamental beam to change path length.

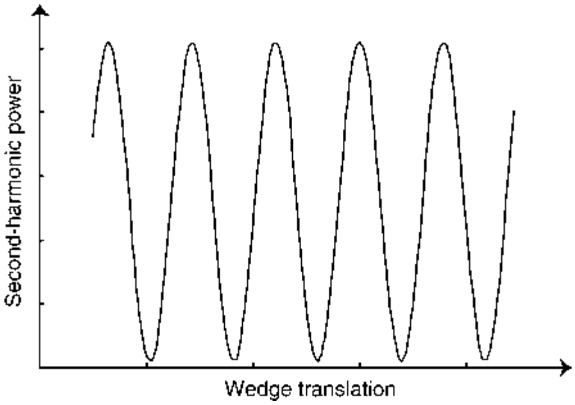

Figure 4 The harmonic intensity cycles between a minimum and maximum when the path is changed by one coherence length.

and wedge measurements have the advantage of being able to isolate individual components of the nonlinear optical tensor. Harmonic power can be low with cw beams requiring chopped, lock-in detection methods or even photon-counting techniques. Even with pulsed fundamental radiation, peak harmonic power may only be at the level of milliwatts. It is necessary to isolate the harmonic radiation from the fundamental radiation that could have peak power in the range of 10^5 W for a Q-switch laser.

Phase-matched second harmonic generation (Figure 5) offers a measurement technique that is similar to practical applications. Harmonic signals are large, thus simplifying detection. However, it is helpful to keep conversion levels below a few percent to allow the use of small depletion approximations in data analysis. Good optical quality of the nonlinear optical material and good quality and accurate characterization of the fundamental radiation are important in this measurement technique. The analysis is simplified for low levels of conversion, spectral distribution well within the phase-matching bandwidth, no absorption or other losses, and no significant diffraction or walkoff. If the fundamental intensity is a pulse with distribution of $I_\omega(r,t) = I_{\omega,0} \exp(-t^2/\tau^2)\exp(-2r^2/w_0^2)$, integration yields second-harmonic energy:

$$U_{2\omega} = \int_{-\infty}^{\infty} \mathrm{d}t \int_0^{\infty} 2\pi r\, \mathrm{d}r I_{2\omega}(r, z = l)$$

$$\approx \frac{2\sqrt{2}\omega^2 d_{\mathrm{eff}}^2 l^2 U_\omega^2}{n^3 c^3 \varepsilon_0 \pi^{3/2} \tau w_0^2} \frac{\sin^2(\Delta kl/2)}{(\Delta kl/2)^2} \qquad [16]$$

where $U_{2\omega}$ is the second harmonic energy and U_ω is the fundamental energy, r is the radial coordinate, t is time, and l is the length of the material. Harmonic conversion can be measured as Δk and varied, either by crystal rotation or temperature change, to confirm that optical quality of the nonlinear material and incident fundamental radiation are a level of tuning that is accurately predicted by theory (Figure 6). Even this simplified equation shows the importance of accurate characterization of the fundamental radiation and second-harmonic conversion in obtaining values of the nonlinear optical coefficient. Conversion efficiency is calculated more accurately with detailed analytical techniques or numerical methods, provided accurate and complete characterization of the fundamental radiation is available.

Parametric fluorescence, in principle, offers an experimental simplification over second-harmonic generation techniques. Parametric fluorescence (Figure 7) is a phase-matched technique in which a pump beam of angular frequency ω_{p} produces a signal beam and an idler beam of angular frequencies $\omega_{\mathrm{s}} + \omega_{\mathrm{i}} = \omega_{\mathrm{p}}$ by the process of spontaneous parametric generation. It is only necessary to measure the ratio of the generated signal power W_{s} and the incident pump power W_{p}. The signal power is given by

$$W_{\mathrm{s}} = \frac{2(h/2\pi)\omega_{\mathrm{s}}^4 \omega_{\mathrm{i}} d_{\mathrm{eff}}^2 n_{\mathrm{s}} l W_{\mathrm{p}} \Omega}{(2\pi)^2 \varepsilon_0 c^5 n_{\mathrm{i}} n_{\mathrm{p}} |\partial \Delta k/\partial \omega_{\mathrm{s}}|_{\omega p}} \qquad [17]$$

where h is Planck's constant, Ω is the solid angle of detection (inside the crystal), and n_{s}, n_{i}, and n_{p} are

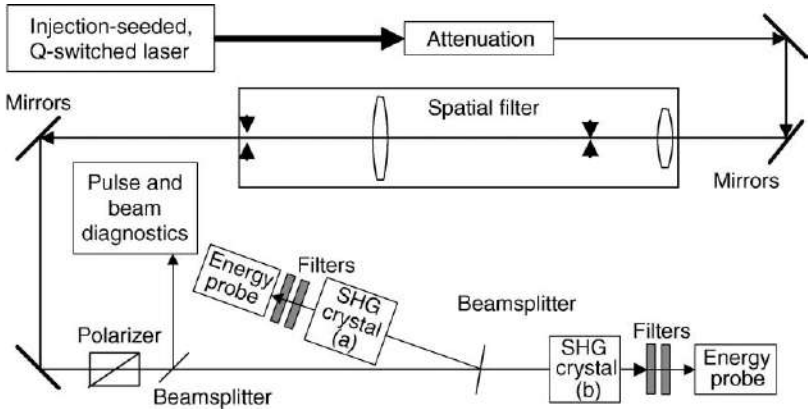

Figure 5 A setup for a relative measurement of second-order nonlinear optical coefficient based on phase-matched second-harmonic generation is shown.

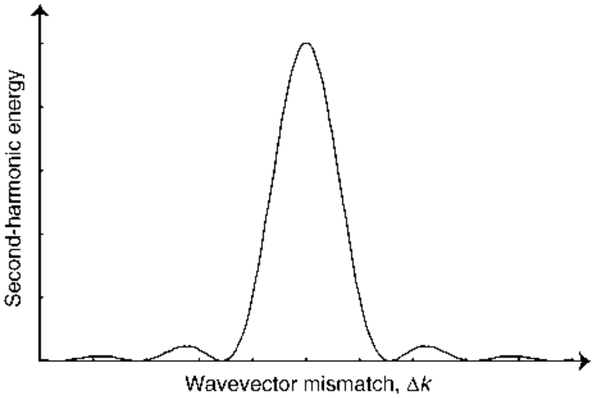

Figure 6 Observation of a theoretically predicted phase-matching tuning curve confirms required optical quality.

the indices of refraction of the signal, idler, and pump respectively. The absolute value of the partial derivative of $\Delta k = k_p - k_s - k_i$, with respect to signal frequency with the pump frequency held constant, is

$$|\partial \Delta k/\partial \omega_s|_{\omega p} = \frac{1}{c}\left|n_s - n_i + \omega_s \frac{dn_s}{d\omega_s} - \omega_i \frac{dn_i}{d\omega_i}\right| \quad [18]$$

It is still not well understood why the carefully performed earlier measurements by this method produced nonlinear optical coefficient values larger than those obtained by other techniques.

Tabulations of measurements of nonlinear optical coefficients are included in handbooks. The Internet posting ‘SNLO’ listed in the Further Reading is a good source of up-to-date values. Table 3 lists some nonlinear optical coefficients obtained from several sources. Signs of these coefficients are not always well determined, and they are reproduced where given and may only indicate relative polarity. Point group symmetry is given with some comments about axis assignments for the *mm*2 point group. The fundamental wavelength is given for second-harmonic measurements and twice the shortest wavelength is given for three-wave processes. Many of the values obtained by Shoji *et al.* are based on carefully performed wedge measurements with single frequency cw lasers and taking detailed account of multiple reflections in the analysis. Their measurements for congruent $LiNbO_3$ reported agreement between Maker-fringe and wedge measurements, and phase-matched difference-frequency generation measurements, and finally agreement with parametric fluorescence. The measurements of Alford and Smith combined optical-parametric amplification and second-harmonic generation at a number of different wavelengths and correlation with values previously reported in the literature to examine the validity of using a constant Miller’s delta, and reported best-fit values for the second-order nonlinear coefficient. Constant Miller’s delta provided reasonably accurate wavelength scaling, with good agreement for most materials in recent measurements.

The measurements of the nonlinear optical coefficients of $KNbO_3$, listed in Table 3 are of particular interest. These measurements were performed by the separated-beam wedge technique (Figure 8). With this technique, a large wedge angle is used in a relatively long crystal and with a larger fundamental beam than in the conventional wedge technique. The beam is large enough that many fringes are present in the second-harmonic at the exit surface. The fringe structure causes focused harmonic radiation to diffract into two clearly resolved spots. Integration of an earlier equation, with constant fundamental field amplitude $E_\omega(0)$, provides a simplified expression of the second-harmonic field $E_{2\omega}(z,t)$ in terms of a free wave with propagation constant

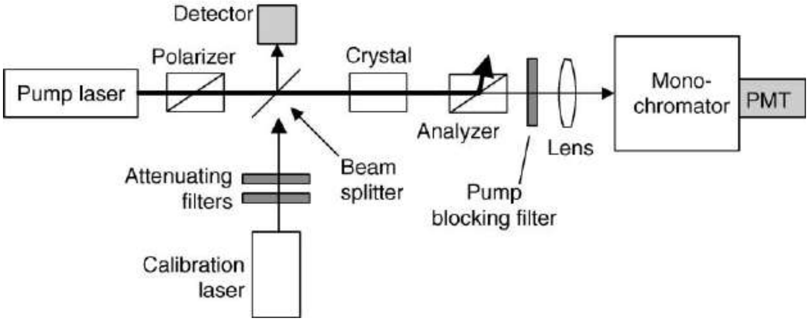

Figure 7 Spontaneous optical parametric fluorescence requires measurement of the ratio of the power of the signal fluorescent to the power of the pump radiation. The calibration laser is used to measure the signal transmission of the filtering system used to block the pump radiation from reaching the photomultiplier tube (PMT).

Table 3 Selected nonlinear optical coefficients

Material	*Point group*	*Nonlinear optical coefficient (pm V^{-1})*		*Wavelength*	*Source*
$AgGaS_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	11.2	10.6 μm	[a]
		$d_{14} = d_{24} = d_{36}$	17.5	1064 nm	[a]
$AgGaSe_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	33	10.6 μm	[a]
BaB_2O_4 (BBO)	$3m$	d_{22}	2.3	1064 nm	[a–f]
		$d_{16} = d_{21} = -d_{22}$	−2.3	1064 nm	[a]
		$d_{15} = d_{31} = d_{24} = d_{32}$	0.04	1064 nm	[d]
		d_{33}	0.04	1064 nm	[d]
CdS	$6mm$	d_{33}	32	10.6 μm	[a]
		$d_{15} = d_{31} = d_{24} = d_{32}$	−16	10.6 μm	[a]
		d_{33}	19.1	1064 nm	[e]
		$d_{15} = d_{31}$	10.1	1064 nm	[e]
		$d_{24} = d_{32}$	10.7	1064 nm	[e]
CdSe	$6mm$	d_{33}	36	10.6 μm	[a]
		$d_{15} = d_{31} d_{24} = d_{32}$	−18	10.6 μm	[a]
CdTe	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	109	1064 nm	[e]
		$d_{14} = d_{24} = d_{36}$	168	10.6 μm	[c]
$CsLiB_6O_{10}$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	0.74	1064 nm	[f]
GaAs	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	83	10.6 μm	[a]
			170	1064 nm	[e]
GaP	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	37	10.6 μm	[a]
			70.6	1064 nm	[e]
KD_2PO_4 (KD*P)	$\bar{4}2m$	d_{36}	0.37	1064 nm	[a,b]
KH_2PO_4 (KDP)	$\bar{4}2m$	d_{36}	0.39	1064 nm	[a,b,e,g,h]
$KNbO_3$	$mm2$	d_{33}	21.9	1064 nm	[h]
$Z = Z_L = x = l$		d_{15}	13	1064 nm	[h]
$X = Y_L = z = s$		d_{31}	12.4	1064 nm	[h]
$\mathrm{Y} = \mathrm{X_L} = -\mathrm{y} = \mathrm{m}$		d_{24}	9.2	1064 nm	[h]
		d_{32}	8.9	1064 nm	[h]
$KTiOAO_4$ (KTA)	$mm2$	$d_{15} = d_{31}$	4.2	1064 nm	[e]
$Z = Z_L = z = m$		d_{15}	2.91	2128 nm	[g]
$X = Y_L = -y = s$		$d_{24} = d_{32}$	2.8	1064 nm	[e]
$Y = X_L = x = l$		d_{33}	16.2	1064 nm	[e]
$KTiOPO_4$ (KTP)	$mm2$	d_{33}	14.6	1064 nm	[e]
$Z = Z_L = z = m$		d_{15}	3.7	1064 nm	[e]
$X = Y_L = -y = s$		d_{31}	3.7	1064 nm	[e]
$Y = X_L = x = l$		d_{24}	1.9	1064 nm	[e]
		d_{32}	2.2	1064 nm	[e]
LiB_3O_5 (LBO)	$mm2$	d_{33}	±0.04	1064 nm	[a,b]
$Z = Z_L = y = s$		$d_{15} = d_{31}$	±0.85	1064 nm	[a,b]
$X = Y_L = z = m$		d_{15}	1.04	1064 nm	[g]
$Y = X_L = x = l$		$d_{24} = d_{32}$	∓0.67	1064 nm	[a,b]
$LiIO_3$	6	d_{33}	4.5	1064 nm	[a,b]
		$d_{15} = d_{31}$	4.4	1064 nm	[a,b]
congruent	$3m$	d_{33}	−27.2	1064 nm	[a,b]
$LiNbO_3$		$d_{31} = d_{15} = d_{32} = d_{24}$	−4.35	1064 nm	[a,b]
		$d_{22} = -d_{21} = -d_{16}$	2.1	1064 nm	[a,b]

Table 3 (*Continued*)

Material	*Point group*	*Nonlinear optical coefficient (pm V^{-1})*		*Wavelength*	*Source*
Agreement of measurements of d_{31} by Maker fringe and wedge SHG (4.6 pm V^{-1}) and parametric fluorescence and parametric amplification (4.3 pm V^{-1})					
				1064 nm	[e]
5%MgO:LiNbO$_3$	3*m*	d_{33}	25	1064 nm	[e]
		d_{31}	4.4	1064 nm	[e]
LiTaO$_3$	3*m*	d_{33}	13.8	1064 nm	[e]
		d_{31}	0.85	1064 nm	[e]
NH$_4$H$_2$PO$_4$ (ADP)	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	0.47	1064 nm	[a,e]
αSiO$_2$	32	d_{11}	0.3	1064 nm	
ZnGeP$_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	69	10.6 μm	[a,b]

[a] Roberts DA (1992) Simplified characterization of uniaxial and biaxial nonlinear optical crystals: a plea for standardization of nomenclature and conventions. *IEEE Journal of Quantum Electronics* 28: 2057–2074.

[b] Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.

[c] Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn, chs 4 and 5. New York: Marcel Dekker.

[d] Shoji I, Nakamura H, Ohdaira K, *et al.* (1999) Absolute measurement of second-order nonlinear-optical coefficients of β-BaB$_2$O$_4$ for visible to ultraviolet second-harmonic wavelengths. *Journal of the Optical Society of America B* 16: 620–624.

[e] Shoji I (1997) Absolute scale of second-order nonlinear-optical coefficients. *Journal of the Optical Society of America B* 14: 2268–2294.

[f] Shoji I, Nakamura H, Ito R, *et al.* (2001) Absolute measurement of second-order nonlinear-optical coefficients of CsLiB$_6$O$_{10}$ for visible-to-ultraviolet second-harmonic wavelengths. *Journal of the Optical Society of America B* 18: 302–307.

[g] Alford WJ and Smith AV (2001) Wavelength variation of the second-order nonlinear coefficients of KNbO$_3$, KTiOPO$_4$, KTiOAsO$_4$, LiNbO$_3$, LiIO$_3$, β-BaB$_2$O$_4$, KH$_2$PO$_4$, and LiB$_3$O$_5$ crystals, a test of Miller wavelength scaling. *Journal of the Optical Society of America B* 18: 524–533.

[h] Pack MV, Armstrong DJ and Smith AV (2003) Measurement of the $\chi^{(2)}$ tensor of the potassium niobate crystal. *Journal of the Optical Society of America B* 20: 2109–2116.

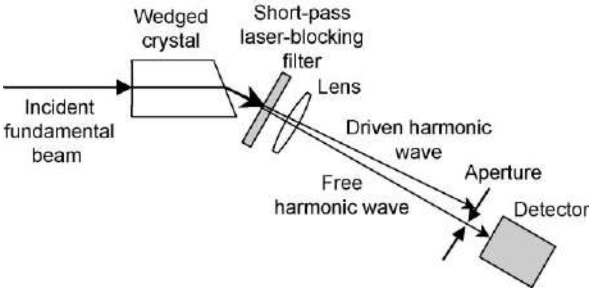

Figure 8 In the separated-beam wedge measurement technique, a fundamental beam of large transverse size generates harmonic radiation that is separated into a free harmonic wave and a driven harmonic wave after exiting the crystal.

$k_{2\omega}$ and a driven wave with propagation constant $2k_\omega$:

$$E_{2\omega}(z,t) = \frac{\omega d_{\mathrm{eff}} E_\omega^2(0)}{n_{2\omega} c} \frac{1}{\Delta k} \{\exp i(k_{2\omega} z - 2\omega t) + c.c.\} - \frac{\omega d_{\mathrm{eff}} E_\omega^2(0)}{n_{2\omega} c} \frac{1}{\Delta k} \{\exp i(2k_\omega z - 2\omega t) + c.c.\} \qquad [19]$$

The two beams with different propagation constants are refracted at different angles at the wedged exit surface. The detailed analysis must include losses due to absorption and reflections at the surfaces and also nonlinear effects at the surfaces. Precision again is required in characterizing the fundamental beam and incident and generated intensities or energies as well as the geometry of the measurement. The large wedge eliminates problems of multiple reflections, and it is no longer necessary to carefully choose the beam spot size to fit within one fringe maximum, yet be large enough to allow use of the near field approximation. The accuracy of this measurement is demonstrated when the coefficients are expressed as Miller-delta values: $\delta_{24} = 0.153\,\mathrm{pm\,V^{-1}}$, $\delta_{32} = 0.150\,\mathrm{pm\,V^{-1}}$, $\delta_{15} = 0.195\,\mathrm{pm\,V^{-1}}$, and $\delta_{31} = 0.192\,\mathrm{pm\,V^{-1}}$. These values satisfy Kleinman symmetry, $\delta_{24} = \delta_{32}$ and $\delta_{15} = \delta_{31}$, well within the reproducibility of the measurement, which was about ±3%.

The nonlinear optical coefficients of BBO d_{22}(1064 nm) = 2.3 pm V^{-1} and KDP d_{36}(1064 nm) = 0.39 pm V^{-1} are gaining agreement from a number of independent measurements. The apparent disparity in the values for GaAs d_{36}(10.6 μm) = 83 pm V^{-1} and d_{36}(1064 nm) = 170 pm V^{-1}, is resolved when the coefficients are converted to Miller's delta values: $\delta_{36} = 0.089$ pm V^{-1} and $\delta_{36} = 0.086$ pm V^{-1}, respectively. The same thing can be done for GaP, for which d_{36}(10.6 μm) = 37 pm V^{-1} and d_{36}(1064 nm) = 70.6 pm V^{-1} convert to $\delta_{36} = 0.085$ pm V^{-1} and $\delta_{36} = 0.077$ pm V^{-1}, respectively. However, the older and newer values for other materials such as CdTe and CdS do not achieve this agreement. Other materials such as CLBO and LBO need additional investigation to resolve differences. The measurement of nonlinear optical coefficients is a continuing process.

Other techniques for measurement involving different nonlinear-optical processes are possible. Parametric amplification was mentioned above. Generally more complicated processes are less well suited for measurement of fundamental properties. Integrated-optical devices such as waveguides are even more complex. Often nonlinearity is measured in fractional conversion per unit of pump power to characterize overall device performance. There are survey techniques for evaluating materials such as grinding the material to a powder, illuminating with intense laser radiation, and looking for harmonic conversion. A surprising amount of information, such as a rough estimate of nonlinearity and the possibility of birefringent phase matching, can be gleaned by sorting the power and measuring the scattered harmonic radiation. Second-harmonic reflections from surfaces and thin film can be used in some instances. Electric fields can be use to orient molecules in polymers and solute molecules in solution to measure the hyperpolarizability of individual molecules.

Reporting Frames

Standardized axis specifications avoid confusion, but nonstandard axes can prove convenient in some applications provided care is used. The crystal $KNbO_3$ has *mm*2 point group; it could also be *m*2*m* or 2*mm* depending on the choice of axes to describe the crystal properties. The axis assignment for the orthorhombic class *mm*2, as it is used in Table 3, follows the IEEE/ANSI Std. 176-1987: *Z* is the polar axis, *X* is parallel to the shorter of the remaining unit cell dimensions, and *Y* is chosen to complete a right-handed coordinate frame. Polarity of the axes can be assigned by pyroelectric properties and piezoelectric properties of the crystal. The common assignments of the axes in the literature are given by X_L, Y_L, Z_L in the table. Reference to the biaxial indices of refraction are given by the lower case *x*, *y*, *z* with $n_x < n_y < n_z$. Finally the axes are identified by the unit cell dimensions $s < m < l$. Either a plus or minus rotation about the Y axis will bring the crystal into an *xyz* frame. However, a +90 degree rotation will give coefficients that are opposite in sign from a −90 degree rotation. The signs of the nonlinear optical coefficients are often not well determined because sign is not obtained from conversion efficiency in a single crystal. However, coefficient signs can be compared in the same crystal or between two crystals to determine if they are the same or opposite.

The crystals KTP or KTA are used here as an example. A 90° rotation about the *Z*-axis gives a new coordinate system $X'Y'Z'$ aligned with the *xyz* frame determined by the indices of refraction. In the new frame, the component of the nonlinear optical tensor, in terms of the standard frame, are $d_{zxx} = d_{32}$ and $d_{zyy} = d_{31}$. All that is needed here is to exchange d_{32} and d_{31} to change the standard frame to the one commonly used in the literature, X_L, Y_L, Z_L. When the largest index of refraction is not aligned with the polar axis, as is the case for LBO and $KNbO_3$, the specification of tensor components can become confusing if a standard frame is not used.

Often crystallographers and crystal manufacturers can identify axes by surface morphology or by poling processes of ferroelectric crystals. With other crystal symmetries a change in the reporting frame that does not change linear optical properties may change the sign of some of the nonlinear optical tensor components while leaving others unchanged. For a crystal of *3m* point group, a 180° rotation about *Z* will change the sign of d_{22}, d_{21}, and d_{16} and leave the other coefficients unchanged. The *xyz* frame with $n_x < n_y < n_z$ is convenient for calculations of optical propagation, and the standard *XYZ* frame is unambiguous for reporting of tensor properties.

Conclusion

There is still a range of values for nonlinear optical coefficients of many materials. Standardization is evolving on more commonly used materials. Greater precision in measurement and extension of measurement techniques are providing better accuracy. However, caution is still required when accessing values. It is necessary to scrutinize the sources of values to obtain the highest accuracy. Accurate measurements need to be extended to difficult spectral ranges of the ultraviolet and infrared. The ambiguity of the sign of tensor components needs to be resolved. In recent measurements, carefully executed wedge measurements are producing more reliable results. The separated beam wedge technique appears to be particularly promising. Phase-matched second-harmonic generation and other practical applications of nonlinear optical materials will continue to serve as techniques to verify accuracy. Current analytical and numerical techniques provide a method for precise modeling of nonlinear optical processes and the need for more accurate material characterization.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Basics:** $\chi^{(2)}$-Harmonic Generation.

Further Reading

Alford WJ and Smith AV (2001) Wavelength variation of the second-order nonlinear coefficients of $KNbO_3$, $KTiOPO_4$, $KTiOAsO_4$, $LiNbO_3$, $LiIO_3$, β-BaB_2O_4, KH_2PO_4, and LiB_3O_5 crystals, a test of Miller wavelength scaling. *Journal of the Optical Society of America B* 18: 524–533.

Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.

Gehr RJ (1998) Separated-beam non-phase-matched second-harmonic method of characterizing nonlinear optical crystals. *Journal of the Optical Society of America B* 15: 2109–2116.

Kurtz SK (1975) Measurement of nonlinear optical susceptibilities. In: Rabin H and Tang CL (eds) *Quantum Electronics: A Treatise, vol. 1, part A*, pp. 209–281. New York: Academic Press.

Pack MV (2003) Measurement of the $\chi^{(2)}$ tensor of the potassium niobate crystal. *Journal of the Optical Society of America B* 20: 2109–2116.

Roberts DA (1992) Simplified characterization of uniaxial and biaxial nonlinear optical crystals: a plea for standardization of nomenclature and conventions. *IEEE Journal of Quantum Electronics* 28: 2057–2074.

Shoji I (1997) Absolute scale of second-order nonlinear-optical coefficients. *Journal of the Optical Society of America B* 14: 2268–2294.

SNLO nonlinear optics code available from Smith AV Sandia National Laboratories, Albuquerque, NM 87185-1423. (an Internet publication) http://www.sandia.gov/imrl/XWEB1128/xxtal.htm.

Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn, chs 4 and 5. New York: Marcel Dekker.

$\chi^{(3)}$

P P Banerjee, University of Dayton, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Over the past 30 years, several methods have been used to measure and characterize optical nonlinearities. With increasing interest in nonlinear optics beyond the traditional harmonic generation, third-order nonlinear processes have been shown to be responsible for self-focusing and defocusing of optical beams, phase conjugation, optical bistability, etc. Third-order nonlinear processes in Kerr materials arise from the cubic dependence of the polarization $P = P^{L} + P^{NL}$, $P^{NL} = \chi^{(3)} \vdots E \cdot E \cdot E$ in a material on the optical field E, through the nonlinear susceptibility tensor $\chi^{(3)}$. The cubic nonlinearity in turn results in a nonlinear refractive index coefficient n_2 defined through:

$$n = n_0 + n_2 |E_p^{\omega_0}|^2 \qquad [1]$$

where $E_p^{\omega_0}$ is the phasor part of $E : E = Re[E_p^{\omega_0} \exp \times j\omega_0 t]$. The relation between $\chi^{(3)}$ and n_2 can be expressed as:

$$n_2 = (3/8n_0)\chi^{(3)}_{1111}(-\omega_0; \omega_0, \omega_0, -\omega_0) \qquad [2]$$

In some materials, an effective $\chi^{(3)}$ or an effective n_2 can be defined as well. For instance, in a second-order nonlinear material, an effective $\chi^{(3)}$ is established through cascading of $\chi^{(2)}$s, and in the presence of phase mismatch between the fundamental and the second harmonic. In other materials, such as liquid crystals, an effective n_2 can be defined due to the nonlinear dependence of the reorientation of the director axis on the optical field. In photorefractive materials, the induced nonlinearity stems from the nonlinear dependence of the induced space charge distribution in the material on the optical field. Some of the techniques used to measure the third-order nonlinearity in materials include degenerate four-wave mixing (DFWM) or phase conjugation, interferometry, ellipse rotation, and beam self-refraction and/or distortion and bending.

For instance, in DFWM, two contradirectional propagating pumps and the probe beam having the same frequency ω_0 interact to give rise to the conjugate, whose amplitude depends on the magnitude of the cubic nonlinearity coefficient. The conservation of energy and momentum describes a parametric interaction of the form $\omega_0 + \omega_0 - \omega_0 \to \omega_0$ (in frequency space) and, $\mathbf{k}_{\text{probe}} + \mathbf{k}_{\text{pump1}} - \mathbf{k}_{\text{pump2}} \to \mathbf{k}_{\text{conj}}$ (in $\mathbf{k}$-space). The conjugate beam is clearly distinguishable by its spatial separation from the other interacting waves. However, alignment of the three waves is critical to obtain a good phase conjugate; furthermore, the sign of the nonlinearity cannot be determined.

In the interferometric method, a laser beam is split into two parts, one serving as the reference, and the other being incident on the nonlinear sample. As in the case of interferometers, if the two beams are then recombined and introduced through a lens, an interference pattern forms at the focal plane of the lens. While the information about the magnitude of the nonlinearity can be found from the fringes and their change when the incident power is varied, information of the sign of the nonlinearity is still not readily inferable.

In this article, we will concentrate on self-refraction, particularly in light of the fact that it results in a sensitive and at the same time simple method called the *z*-scan©, requiring the propagation of a single beam through the nonlinear sample. We comment that self-diffraction does also cause self-bending of an asymmetric beam due to the asymmetric phase change induced through the cubic nonlinearity; however, extracting the information of the sign and magnitude of the nonlinearity can be effectively done only if the sample is thin, implying that there is negligible linear diffraction of the beam as it passes through the sample. In the rest of the article, we will therefore concentrate on the *z*-scan method, since it is an elegant and simple single (symmetrical) beam technique that yields information on the magnitude and sign of the cubic nonlinearity, as well as its real and imaginary parts. We will describe in detail how the *z*-scan can measure the real part of the cubic nonlinearity, for both thin and thick samples, as well as showing its application in determining the reorientational nonlinearity of liquid crystals and the nonlinearity of photorefractive materials, where photovoltaic effect and diffusion of charge carriers contribute to the effective nonlinearity.

Since the mid-1990s *z*-scan has evolved as a standard method for determining effective nonlinear refractive index coefficients of nonlinear materials. The name *z*-scan is derived from the fact that the nonlinear test sample is scanned along the direction of propagation (z) of a beam about the back focal plane of an external lens illuminated by an incident Gaussian beam from a laser. The far-field diffraction pattern yields information about the nonlinearity of the test material. The *z*-scan technique has been used to measure and characterize the nonlinear refractive index coefficient and nonlinear absorption coefficient in materials such as CS_2, ZnSe, Cs atoms, multiphoton absorption in GaN, organic dyes, thermal nonlinearities in liquids, liquid crystals, photorefractive materials, etc., to name a few. *z*-scan measurements have been made using non-Gaussian incident beams, and *z*-scans have been performed using both CW and pulsed laser sources.

The first extensive series of work on the *z*-scan technique, including analysis and application to the measurement of the cubic nonlinearity, was performed by the group of Van Stryland at the Center of Electro-Optics and Lasers (CREOL) at the University of Central Florida. Other investigators later performed similar measurements and characterization during the study of nonlinear optical properties of a solution of chlorophyll in acetone, and Chinese tea.

Theory of the Thin Sample *z*-Scan

We will first analyze propagation of a (circularly symmetric) Gaussian beam which is initially focused by an external lens and is then incident on a 'thin' slice of a cubically nonlinear Kerr medium. This case is important because, as explained below, it provides a basis for the measurement of the nonlinear refractive index coefficient n_2.

Figure 1 shows a simple ray diagram to demonstrate what happens if the cubically nonlinear material is thin. The propagating Gaussian beam induces its own lens, which is positive for a material having a positive n_2, for reasons that will become clear to the reader later on in this section. Note that if the position of the induced lens is at the back focal plane of the external lens, the former does not affect the propagation of the beam behind the nonlinear material. The situation is thus the same as the propagation of the externally focused Gaussian beam in a linear diffraction-limited environment. However, if the separation between the external lens and the sample is less than the focal length f of the external lens, the beam in the far field is wider than the linear diffraction-limited case. Conversely, if the lens-sample separation is larger than f, the beam width is smaller. This is the basis of the *z*-scan: for a material with a positive n_2, the scanning of a thin nonlinear sample through the back focal plane of the external lens, starting from a position to the left of the focus, yields far-field profiles which are initially larger and eventually smaller than the width in the linear diffraction-limited case. The opposite is true for materials with a negative n_2. For very large displacements of the sample from the back focal plane of the

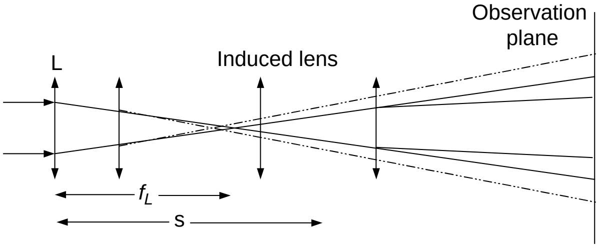

Figure 1 Geometrical optics picture of *z*-scan setup for nonlinear refractive index n_2. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

external lens, the beam power is small enough that nonlinear lensing is negligible, and hence the beam widths return to the linear diffraction-limited case. The variation of the on-axis intensity or transmittance as a function of scan distance is complementary to the variation of the width, due to conservation of power, to be explained further below. A sketch of the on-axis transmittance during a z-scan for a positive n_2 material, is shown in Figure 2.

The simpler ray picture described above can be augmented by using a q parameter approach to Gaussian beam propagation. For the sake of simplicity, we prefer to introduce the q parameter approach here, instead of the more rigorous Gaussian decomposition of the induced phase as computed by Van Stryland's group. The q of a Gaussian beam in air is defined through the relation $1/q = 1/R - \mathrm{i}(2/k_0 w^2)$, where R and w denote the radius of curvature and the beamwidth, respectively, and k_0 is the propagation constant in air. Assume a Gaussian beam with a waist w_0 (implying initially planewave fronts) incident on the external lens. The Gaussian can be described by $q = q_0 = \mathrm{i}z_{R0}$, where $z_{R0} = k_0 w_0^2/2$ is called the Rayleigh length of the incident Gaussian. The laws of q transformation of a Gaussian beam are:

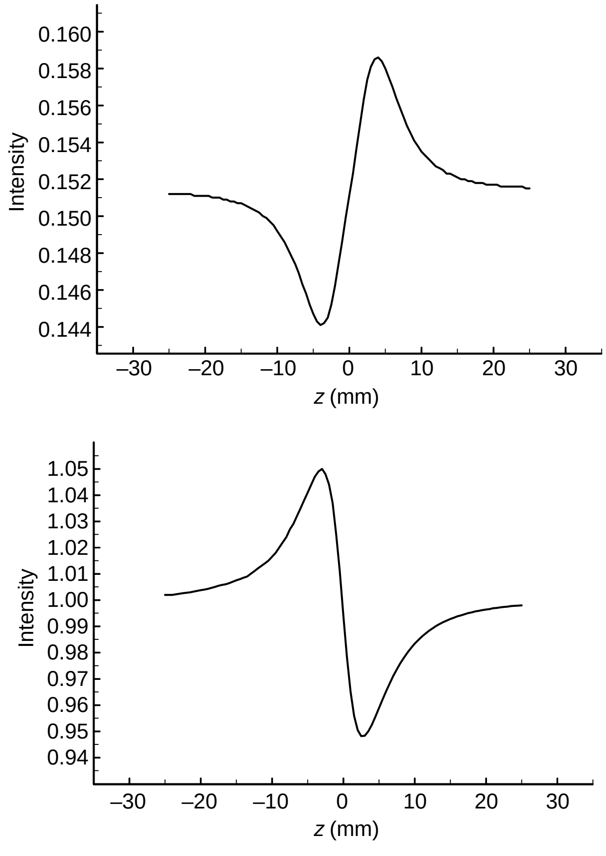

Figure 2 Typical z-scan plots for positive (top) and negative (bottom) cubic nonlinearities. The vertical axis is in arbitrary units.

(i) for translation through a distance d, the q changes to $q + d$; and
(ii) for a Gaussian with q-parameter q incident on a lens, the q immediately behind the lens, is given by $(1/q - 1/f)^{-1}$.

Using these laws, the $q = q_{\mathrm{obs}}$ of the Gaussian at a distance d from the nonlinearly induced lens of the focal length f_{ind}, which is located at a distance $s = f + \Delta s$ behind the external lens, can be found as:

$$q_{\mathrm{obs}} = \frac{f_{\mathrm{ind}}[q_0 f + (f - q_0)(f + \Delta)]}{f_{\mathrm{ind}}(f - q_0) - [q_0 f + (f - q_0)s]} + d \qquad [3]$$

In deriving eqn [3], we have used the laws of q transformation on the initial Gaussian beam of q-parameter q_0 passing through the external lens, being translated a distance $f + \Delta s$, passing through the induced lens, and finally being translated a distance d to the observation plane.

To determine $f_{\mathrm{ind}}(\Delta s)$, assume a Gaussian of the form:

$$|E_{\mathrm{p}}^{\omega_0}(x, y, z)| = E(z) \exp(-(x^2 + y^2)/w^2(z)) \qquad [4]$$

Since

$$n = n_0 + n_2 |E_{\mathrm{p}}^{\omega_0}|^2 \approx n_0 - 2n_2 E^2(z)(x^2 + y^2)/w^2(z) \qquad [5]$$

the nonlinearly induced phase change is

$$\Delta\phi = k_0 n \Delta z \approx k_0[n_0 - 2n_2 E^2(z)(x^2 + y^2)/w^2(z)]\Delta z \qquad [6]$$

so that

$$f_{\mathrm{ind}}(z) = n_0 w^2(z)/4n_2 a^2(z)\Delta z \qquad [7]$$

Note that $f_{\mathrm{ind}}(z)$ is inversely proportional to $n_2 E^2(z)$ and Δz. The longer the sample length, the more the nonlinear induced phase, implying a stronger lensing (smaller f_{ind}). Also, the stronger the nonlinearity parameter n_2 and/or on-axis intensity ($E^2(z)$), the stronger the focusing since the change in n is larger.

We must point out that the focal length f_{ind} of the induced lens is also a function of Δs, since it depends on the beam intensity at the location of the nonlinear sample. In fact, the focal length f_{ind} of the nonlinearly induced lens can be expressed in terms of the beam power P, the beam width w at the position Δs, f (the focal length of the external lens), the linear refractive index n_0, the nonlinear refractive index coefficient n_2 defined through eqn [1] where $|E_{\mathrm{p}}^{\omega_0}| = E$ is the optical field in the nonlinear material of length L, assumed 'thin' for now. The term 'thin' means that the sample length is smaller than the depth of focus of the externally focused Gaussian beam. In this case, we can assume that the beam width is a constant through the sample, and the induced focal length is then only directly dependent on the total thickness L of the sample. The focal length of a nonlinearly induced 'thin' Kerr lens can be written as

$$f_{\mathrm{ind}}(\Delta s) = \frac{n_0 w^2(\Delta s)}{4n_2|E|^2 L} \approx \frac{\pi f^4 n_0}{k_0^4 w_0^4 \eta_0 P L n_2} \qquad [8]$$

for small Δs, where η_0 is the (linear) characteristic impedance of the nonlinear medium. Note that in eqn [8], the beam width w is given by

$$w^2(\Delta s) = w_{\mathrm{f}}^2\left[1 + \frac{4(\Delta s)^2}{k_0 w_{\mathrm{f}}^4}\right] \text{ and } w_{\mathrm{f}} \cong \frac{2f}{k_0 w_0} \qquad [9]$$

In eqn [9], w_{f} denotes the waist size of the Gaussian beam at the back focal plane of the external lens. In deriving eqn [8], one has to use the power conservation relation $|E|^2 w^2 = (4/\pi)\eta P$ to re-express the dependence of the induced focal length on the field amplitude in terms of the corresponding width at the same point.

Our final interest is to find the imaginary part of $1/q_{\mathrm{obs}}(= -2/k_0 w_{\mathrm{obs}}^2)$, in order to ascertain the width w_{obs} of the Gaussian beam on the observation plane. Simple but tedious algebra leads us to the relation:

$$\frac{k_0 w_{\mathrm{obs}}^2}{2} = \frac{[(f f_{\mathrm{ind}} d - f^2 d + f^2 f_{\mathrm{ind}}) + f(f_{\mathrm{ind}} - d)\Delta s]^2 + z_{R0}^2[(f_{\mathrm{ind}} - d)\Delta s + f_{\mathrm{ind}} d]^2}{z_{R0} f^2 f_{\mathrm{ind}}^2} \qquad [10]$$

It can be shown simply that when $\Delta s = 0$, and for $|f_{\mathrm{ind}}| \gg f$, $w_{\mathrm{obs}} = w_0 d/f$ for large d. In other words, in the linear case, i.e., when $f_{\mathrm{ind}} = \infty$, as well as in the nonlinear case when $\Delta s = 0$, the far-field width of the Gaussian beam is what is predicted by geometrical optics. Furthermore, from eqn [10], the rate of change of the width with Δs around $\Delta s = 0$, (assuming $z_{R0} \gg f$, and for large d) can be found as

$$\frac{\mathrm{d}w_{\mathrm{obs}}}{\mathrm{d}\Delta s} = \frac{-2d^2 z_{R0}}{k_0 w_{\mathrm{obs}} f^2 f_{\mathrm{ind}}} = -\frac{w_0 d}{f f_{\mathrm{ind}}} \qquad [11]$$

provided we invoke the relation for w_{obs} derived above. In eqn [11], f_{ind} refers to the value of the induced focal length (eqn [8]) with $\Delta s = 0$. Equation [11] may be written as

$$\frac{\mathrm{d}w_{\mathrm{obs}}}{\mathrm{d}\Delta s} = -\frac{k_0^4 w_0^5 P L n_2}{2\pi n_0 f^5} \qquad [12]$$

This means that in a material with a positive (negative) n_2, implying a positive (negative) induced lens, the rate of change of the observed width with the position of the nonlinear sample about the focal plane of the external lens will be negative (positive), in agreement with the results from geometrical optics. Furthermore, the rate of change is proportional to the angle of convergence w_0/f of the externally focused Gaussian, and to the position d of the observation plane.

In z-scan, it is often important to relate the on-axis transmittance in the far field to the scan distance s for measurement purposes. Since for power conservation, the intensity is inversely proportional to the width of the beam in the far field, a positive n_2 would imply a positive slope as a function of scan distance with increasing s, and a negative slope for negative n_2. The value of the slope of the on-axis transmittance could give the information of the n_2 of the material. Note that the z-scan graph in this case is an odd function of the scan distance, measured about the back focal plane of the external lens.

In practice, a detector with a finite aperture area is placed on-axis, and the detected power monitored as a function of scan distance. The transmitted power through the aperture is obtained by

$$\begin{aligned} P_{\mathrm{T}} &= \frac{1}{2}\int_0^{2\pi}\int_0^{r_{\mathrm{a}}} \frac{E_{\mathrm{a}} E_{\mathrm{a}}^*}{\eta_0}\,\mathrm{d}A \\ &= \frac{\pi w_0}{2}\left(\frac{E_0^2}{2\eta_0}\right)\left(1 - \exp\left(-\frac{2r_{\mathrm{a}}^2}{w_{\mathrm{obs}}^2}\right)\right) \\ &= P\left(1 - \exp\left(-\frac{2r_{\mathrm{a}}^2}{w_{\mathrm{obs}}^2}\right)\right) \end{aligned} \qquad [13]$$

where r_{a} is the radius of aperture. The rate of change of P_{T} with Δs can be found as

$$\frac{\mathrm{d}P_{\mathrm{T}}}{\mathrm{d}\Delta s} = \frac{-4r_{\mathrm{a}}^2 P}{w_{\mathrm{obs}}^3}\exp\left(-\frac{2r_{\mathrm{a}}^2}{w_{\mathrm{obs}}^2}\right)\frac{\mathrm{d}w_{\mathrm{obs}}}{\mathrm{d}\Delta s} \qquad [14]$$

Substituting eqn [12] into eqn [14] and using the approximation $w_{\mathrm{obs}} = w_0 d/f$, eqn [14] can be rewritten as

$$\frac{\mathrm{d}P_{\mathrm{T}}}{\mathrm{d}\Delta s} \cong \frac{4r_{\mathrm{a}}^2 f^2 P}{f_{\mathrm{ind}} d^2 w_0^2}\exp\left(-\frac{2r_{\mathrm{a}}^2 f^2}{d^2 w_0^2}\right) \qquad [15]$$

from which

$$\frac{\mathrm{d}P_{\mathrm{T}}}{\mathrm{d}\Delta s}(\Delta s \approx 0) \cong \frac{4k_0^4 w_0^2 \eta_0 r_{\mathrm{a}}^2 P^2 L}{\pi f^2 d^2 n_0}\exp\left(-\frac{2r_{\mathrm{a}}^2 f^2}{d^2 w_0^2}\right) n_2 \qquad [16]$$

When r_a approaches 0;

$$\frac{dP_T}{d\Delta s}(\Delta s \approx 0) = 4\left(\frac{k_0^2 w_0 r_a P}{fd}\right)^2 \frac{\eta_0 L}{\pi n_0} n_2 \qquad [17]$$

Therefore, n_2 can be estimated as

$$n_2 = \left(\frac{\pi n_0}{\eta_0 L}\right)\left(\frac{d}{2k_0^2 w_0 P}\right)^2 \left(\frac{f}{r_a}\right)^2 \frac{dP_T}{d\Delta s} \qquad [18]$$

Thus, if one knows the slope of the z-scan graph around $\Delta s \approx 0$, the value of n_2 can be found. Another way to state this result is to say that the n_2 is related to the peak-to-valley change in the on-axis transmittance ΔT_{p-v}, which is proportional to $dP_T/d\Delta s$.

In passing, we comment that the z-scan method can be also adapted to determine the nonlinear absorption coefficient α_2 of a material, defined through the relation $\alpha = \alpha_0 + \alpha_2 I$, where α is the total absorption coefficient and I denotes the intensity. In order to achieve this, the aperture is assumed to be completely open, in order to collect all the light behind the nonlinear absorptive sample. For $\alpha_2 > 0$, maximum absorption is experienced when the nonlinear sample, assumed thin, is at the back focal plane of the lens since the optical intensity is a maximum due to the external focusing. The z-scan graph is an even function of scan distance, measured about the back focal plane of the external lens.

Thick Sample *z*-Scan

Contrary to the thin sample case, a thick sample is defined as one whose thickness may be comparable to the depth of focus of the externally focused Gaussian beam. In this case, the beamwidth changes during propagation through the sample, so that the induced focal length also changes with propagation. This makes the problem substantially more involved. Indeed, to track the far-field profile in this case, one needs to describe the propagation of the Gaussian beam inside the nonlinear sample using the q parameter formalism. Since both the width and the radius of curvature of the beam at the exit face of the sample are required to predict the far-field behavior, one can write a simple differential equation describing the variation of the q through the nonlinear sample. This equation can be composed from the laws of q-transformation described earlier, and the effects of propagation and nonlinear lensing can be added assuming each is small over an infinitesimal distance of propagation, much like the numerical method called the split step beam propagation method (BPM), often used to analyze such problems. We will not present details of the computation, rather show calculation results for a typical thick sample case, and point out differences with the thin sample predictions for z-scan.

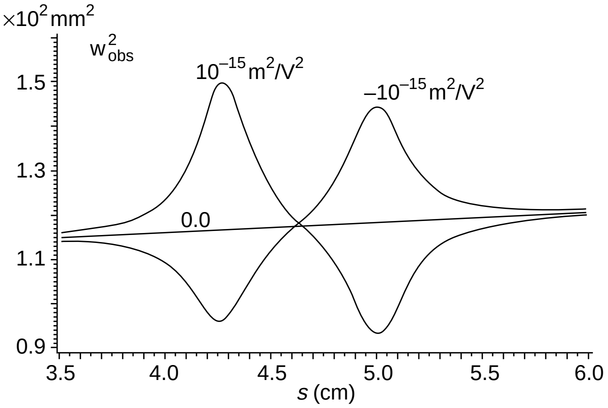

Figure 3 Plot of width of the beam in the far field as a function of the scan distance s. The light wavelength is 640 nm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

Figure 3 shows the beamwidth as a function of scan distance for the case where the observation plane is at a distance $d = 1$ m from the exit face of a 1 cm thick. Assume that a hypothetical sample with $n_0 = 1$, $n_2 = \pm 10^{-15}$ m^2/V^2 is scanned about the back focal plane of an external lens with focal length $f = 5$ cm. For comparison, the case of a linear sample is also superposed on the same graph. Note that with the variation of s, the far-field width initially increases, and then decreases before starting to increase again, for a material with a positive n_2. Note that around the point where the variations intersect the graph for the linear case, the slope of the curves are negative and positive for materials with positive and negative n_2, respectively. The graphs for the variation of on-axis transmittance are opposite to the variation of the widths, as mentioned earlier. This is also in agreement with the results predicted from geometrical optics and the q formalism for a thin sample, described earlier.

How do thick and thin sample z-scans differ? Important differences are: (i) the position of the points where the graphs for positive and negative, n_2s, intersect the graph for the linear case; and (ii) the slope of the graphs, predicted from thin and thick sample calculations. For instance, for the thin sample approximation, the value of the slope around $s = f$ is about -4 for a material with a $n_2 = 10^{-15}$ m^2/V^2. Recall from the previous section, the z-scan plot for a thin sample always intersects the graph for the linear case at the back focus of the external lens, and therefore, it is easy to determine the slope of the z-scan at this point. From the graph in the thick sample case, it is perhaps difficult to ascertain at exactly which point one must determine the slope. A figure of merit would be the minimum value, which occurs for a value of s slightly smaller than $f = 5$ cm. A conservative estimate yields a slope of approximately -1 for this particular example. Note, however, that in the thick lens analysis, we have taken a paraxial approximation to the Gaussian at every stage of its propagation through the nonlinear sample. Strictly speaking, this is not true, and to alleviate this problem, Van Stryland's group included a correction factor with the width of the Gaussian beam.

We would like to mention that values of $n_2 \approx -10^{-15}$ m^2/V^2 can indeed be measured for thermal nonlinearities in liquids. In fact, this is a typical value that we measured for our solution of chlorophyll in ethanol during our first experiments on z-scan. Values of effective $n_2 \approx 10^{-14}$ m^2/V^2 can be measured for liquid crystals with orientational nonlinearities, and will be described below.

Example of *z*-Scan: Effective Cubic Nonlinearity of Liquid Crystals

The induced refractive index in a nematic liquid crystal, due to reorientational nonlinearity, is a function of position in the crystal and depends on the intensity of the incident optical plane wave, its angle with respect to the direction of propagation, and on the applied voltage. Because of the fact that the nonlinearity depends on the direction of propagation of the planewave and the director axis of the crystal, modeling Gaussian beam propagation is rather complicated. Moreover, the director axis has a certain distribution in the material due to the applied voltage, and hence, the electrostatic field in the material, as well as the optical field. This poses a level of complexity a degree higher than that discussed above for the simple thick sample case. Since the induced refractive index is not a constant, and more importantly, a function of the spatial frequency of the optical beam (which determines the angle between a planewave component of the Gaussian beam spectrum and the director axis distribution under external field), one is compelled to analyze optical propagation using split-step BPM. A typical set of simulated z-scan plots is shown in Figure 4, assuming one transverse dimension for simplicity, and the same parameters as in the experiment described below. Actual experimental results from a z-scan of a biased liquid crystal sample is shown in Figure 5 for transmitted power versus sample position in a z-scan experiment for a liquid crystal sample for applied voltages 0.8, 1.0, 1.2, 1.8, and 2.2 V. In this case, the focal length of the external lens is approximately 50 mm and aperture diameter on the observation plane is 1 mm. The scan range is 0.15 inches or approximately 4 mm. The scan step, ΔZ, is varied from about 0.002 inches to about 0.01 inch, depending on the position of the sample. This is done so as to get the maximum information around the back focal plane of the external lens where the measurement values are to be used to find the slope of the z-scan graph, and hence the effective nonlinear refractive index coefficient.

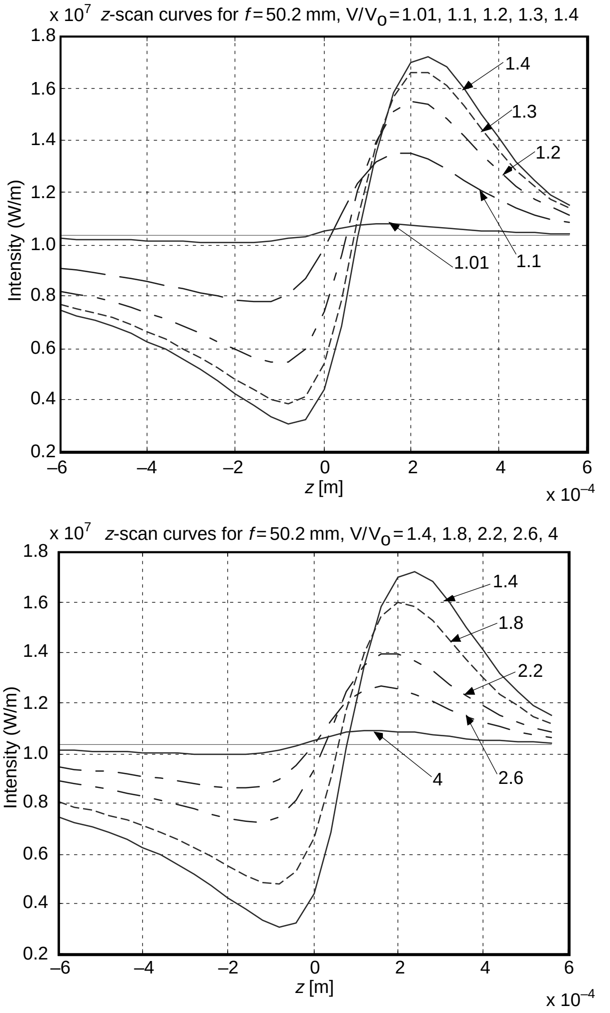

Figure 4 z-scan simulation results for $f = 50.2$ mm and $V/V_o = 1.01$, 1.1, 1.2, 1.3, 1.4 with $E_{op} = 2 \times 10^5$ V/m. z-scan simulation results for $f = 50.2$ mm and $V/V_o = 1.4$, 1.8, 2.2, 2.6, 4 with $E_{op} = 2 \times 10^5$ V/m. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

The plots of the effective $n_2 = n_{2\text{eff}}$ (simulated and measured) are shown in Figure 6a and b, respectively. As seen from both plots, the $n_{2\text{eff}}$ increases and

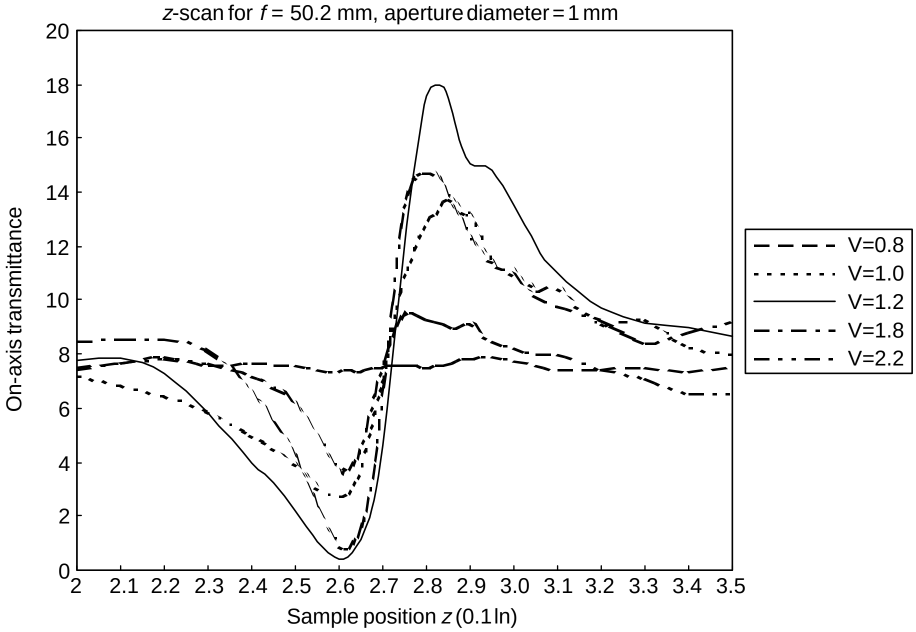

Figure 5 Measured *z*-scan graph for $f = 50.2$ mm and aperture diameter = 1 mm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

reaches a maximum at $V/V_o = 1.4$, before starting to decrease. Simulated and experimental $n_{2\mathrm{eff}}$ have similar trends. Furthermore, there is reasonable agreement between the calculated and measured values of $n_{2\mathrm{eff}}$.

Discussion: Effective n_2 of Other Materials

The *z*-scan is a powerful tool in the measurement and characterization of effective cubic nonlinearities of different materials. The method is simple and easy to set up, and can be readily automated. We provide two other examples of *z*-scans below, as examples. For instance, photorefractive materials are commonly used for dynamic holography, in order to record and reconstruct volume holograms. Photorefractive materials are doped light-sensitive electro-optic materials in which optical radiation creates mobile charge carriers which redistribute themselves to create an induced space-charge field. This electrostatic field in turn modulates the refractive index which phase modulates the optical beam. A figure of merit for the effectiveness of such a material, e.g., lithium niobate, for holographic storage, is the donor to acceptor ratio in such materials. It turns out that this ratio, along with the photovoltaic coefficient, can be related to an effective n_2 in lithium niobate. This is also responsible in producing an elliptic beam in the far field by starting out from a circularly symmetric Gaussian beam focused onto the sample. The ellipticity stems from the tensor nature of the electro-optic coefficients. Typical simulated *z*-scan plots for lithium niobate are shown in Figure 7, and carries the signature variation of the scan in the variation of the beam ellipticity with scan distance. From this plot, we are able to determine the effective n_2 and hence, the donor to acceptor concentration ratio.

In other photorefractive materials, such as barium titanate, in which diffusion of charge carriers plays a dominant effect, the induced change in the refractive index is proportional to the gradient of the intensity, rather than the intensity itself, as was the case in all previous examples. In this case, as seen in Figure 8, it can be shown that the *z*-scan plot for on-axis transmittance with a small finite aperture on the observation plane is an even function of scan distance. This, however, should not be confused with the open aperture *z*-scan that is also even, but which yields information of the nonlinear absorption coefficient, as described earlier. Experiments with a new material, manganese aluminum oxide, which is thought to have both photovoltaic and diffusive contribution to space charge field distribution, yields a *z*-scan that has both odd and even components. Through careful analysis of the *z*-scan results, one is able to determine the relative contributions of each of these effects. Care should be taken in the interpretation of the results in this case, since in an actual experiment, the aperture size cannot be infinitely small, and thus, an optimum aperture size must be determined that allows for sufficient detected power,

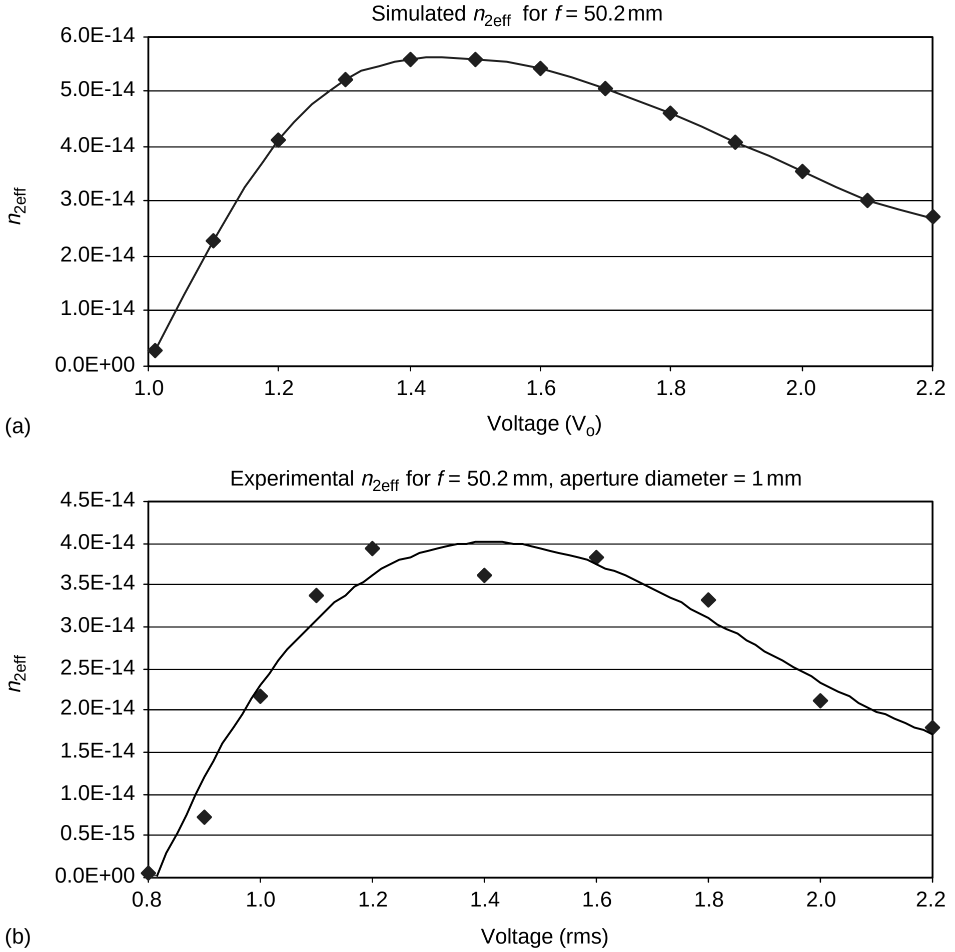

Figure 6 (a) Calculated $n_{2\mathrm{eff}}$ for $f = 50.2$ mm with $E_{\mathrm{op}} = 2 \times 10^5$ V/m; (b) Experimental $n_{2\mathrm{eff}}$ for $f = 50.2$ mm with aperture diameter = 1 mm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

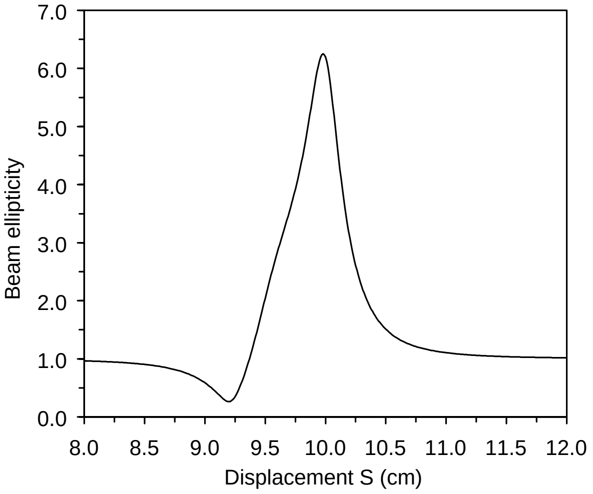

Figure 7 Typical simulated *z*-scan plot showing variation of beam ellipticity as a function of scan length. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

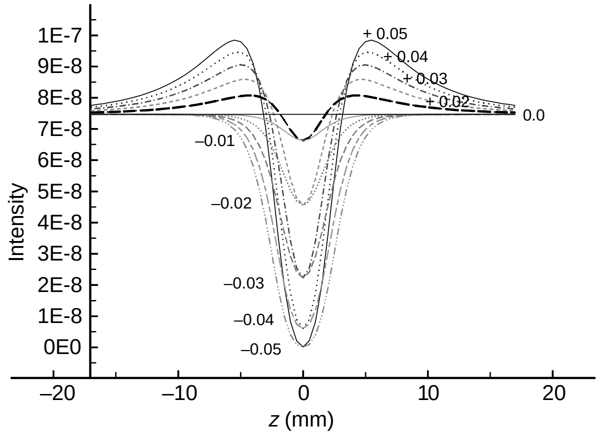

Figure 8 Typical *z*-scan plots of a diffusion dominated photorefractive material, with $\Delta n = C\nabla I$. Different values of *C* are written on the graph.

but at the same time allows one to neglect any effects of nonlinear absorption.

In summary, it is hoped that this article illuminates readers on the history of the *z*-scan, its application to thin and thick samples, its effectiveness in measurement and characterization of effective third-order optical nonlinearities for a wide class of nonlinear materials, such as $\chi^{(3)}$ materials, materials with cascaded second-order nonlinearities, liquid crystals, photorefractives, etc., and its potential for automation, while at the same time pointing out areas where care should be taken in the interpretation of the measured data. With appropriate modification in the theory, it can also be applied to characterization of linear lenses, possibly voltage controllable lenses and lenslet arrays, suitable for dynamic imaging in biomedical applications.

See also

Materials for Nonlinear Optics: Liquid Crystals for NLO. **Nonlinear Optics, Basics:** Four-Wave Mixing; $\chi^{(3)}$–Third-Harmonic Generation.

Further Reading

Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications*. New York: Marcel Dekker.

Banerjee PP, Danilieko A, Hudson T and McMillen D (1998) P-scan analysis of induced inhomogeneous optical linearities. *Journal of the Optical Society of America* 15: 2446.

Banerjee PP, Misra RM and Maghraoui M (1991) Theoretical and experimental studies of propagation of beams through a finite sample of a cubically nonlinear material. *Journal of the Optical Society of America* 8: 1072.

Boyd R (2002) *Nonlinear Optics*, 2nd edn. New York: Academic Press.

Butcher P and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Chapple PB, Staromlynska J and McDuff RG (1994) *Z*-scan studies in the thin- and the thick-sample limits. *Journal of the Optical Society of America B* 11: 975.

Noginov M, Helzer SW, Loutts GB, Banerjee PP, Morrisey M and Kim Y (2003) Study of photorefraction response and diffraction efficiency in Mn:YalO$_3$ crystals. *Journal of the Optical Society of America B* 20: 1233.

Sheik-Bahae M, Said AA and Van Stryland EW (1989) High-sensitivity, single beam n_2 measurements. *Optics Letters* 14: 955.

Sheik-Bahae M, Said AA, Wei TH, Hagan DJ and Van Stryland EW (1990) Sensitive measurement of optical nonlinearities using a single beam. *IEEE Journal of Quantum Electronics* 26: 760.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Yeh P (1993) *Introduction to Photorefractive Nonlinear Optics*. New York: Wiley.

MATERIALS FOR NONLINEAR OPTICS

Contents

Liquid Crystals for NLO
Organic Nonlinear Materials

Liquid Crystals for NLO

I C Khoo, Pennsylvania State University, University Park, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Liquid crystals are complex organic molecular systems that exhibit various mesophases characterized by the degree of order. In this review, we shall focus mainly on thermotropic liquid crystals which exhibit various mesophases in the temperature range between the crystalline and the isotropic liquid states. In particular, we shall discuss the extraordinarily large optical nonlinearities of nematic liquid crystals, as well as the electronic nonlinearities of isotropic-phase liquid crystals. Recent studies have shown that these materials are promising candidates for advanced photonic applications.

Nematic liquid crystals (NLC) possess directional ordering, i.e., when placed in thin cells with surface

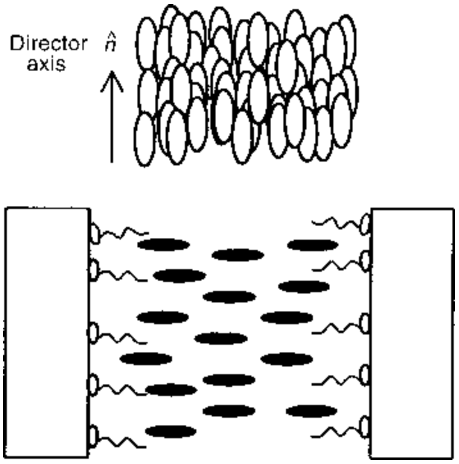

Figure 1 Nematic liquid crystals (NLC) which exhibit directional ordering characterized by the director axis. The lower figure shows how the NLC molecules are aligned in a typical cell with surfactant coated windows.

treated boundaries, the NLC molecules will self-assemble into a birefringent crystal characterized by a so-called director axis $\hat{n}$, cf. Figure 1. For light polarized parallel to $\hat{n}$, the refractive index is n_e, while for light polarized perpendicular to $\hat{n}$, the refractive index is n_0. Typically n_e is ~1.7 and n_0 is ~1.5. This large birefringence $\Delta n = n_e - n_0$ of NLC holds throughout the near UV to the far infrared spectral regime, cf. Figure 2. Perhaps the most important characteristics of NLCs is the easy susceptibility of the birefringent director (crystalline) axis orientation to externally applied fields, which has led to their ever increasing widespread use in various opto-electronics information and image display and processing devices. In analogy to dc and ac field effects, the director can easily be reoriented by an optical field, and thus cause a large optically induced index of refraction change. By doping the NLC with photosensitive or photo-charge producing agents, this process of refractive index change can be enhanced considerably. The refractive indices of NLC are also very strongly temperature dependent.

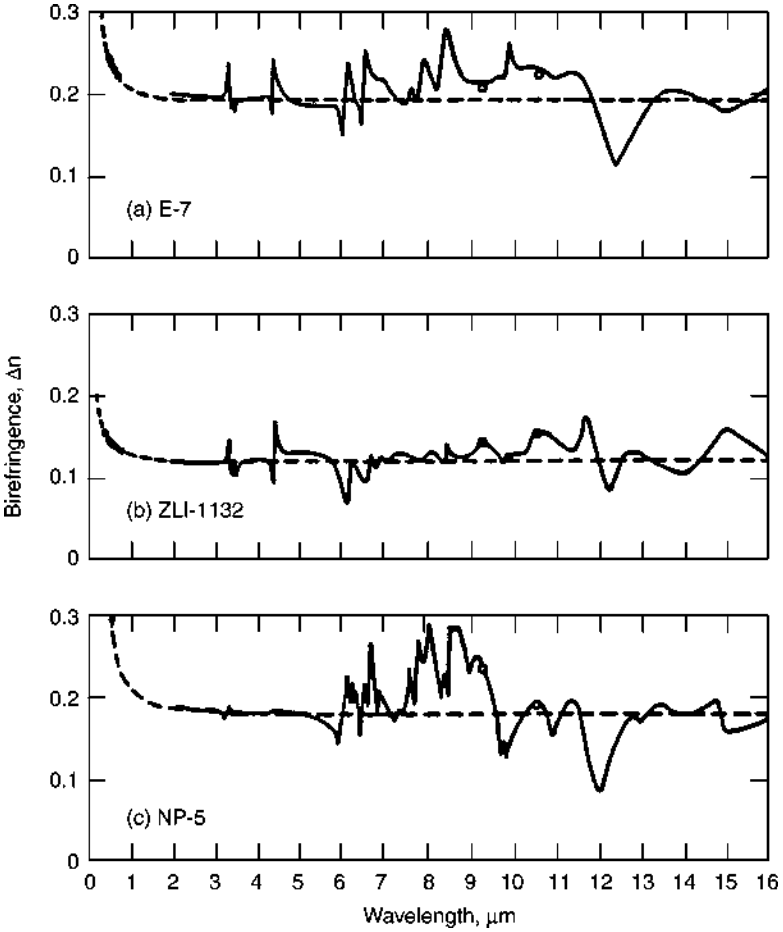

Figure 2 Broadband birefringence of nematic liquid crystals from the near ultraviolet to far infrared.

Accordingly, a large refractive index change can also be generated with laser absorption induced density and order parameter changes. In the following sections, we will broadly classify these nonlinear optical processes originating from the collective response of NLC to external fields as *nonelectronic nonlinearities*.

On the other hand, individual NLC molecules also respond to an optical field through the electronic dipolar interaction. A typical electronic energy level structure of liquid crystal molecules is shown in Figure 3. The dipolar interaction gives rise to the usual nonlinear polarization characterized by the second, third and higher-order nonlinear susceptibilities $\chi^{(2)}, \chi^{(3)}, \ldots$ etc. The complex imaginary parts of these linear and nonlinear susceptibilities account for the photonic absorption processes. Most liquid crystalline molecules possess small linear absorption in the visible region (400–700 nm), and begin to absorb near the UV region (<350 nm), cf. Figure 4. This absorption spectrum translates into a (usually) large nonlinear two-photon absorption spectrum in the visible. In some organic molecules, the intermediate absorption cross-section σ_i or the excited state absorption cross-section σ_e is much larger than the ground state linear absorption σ_g, giving rise to reverse saturable absorption (RSA) and excited state

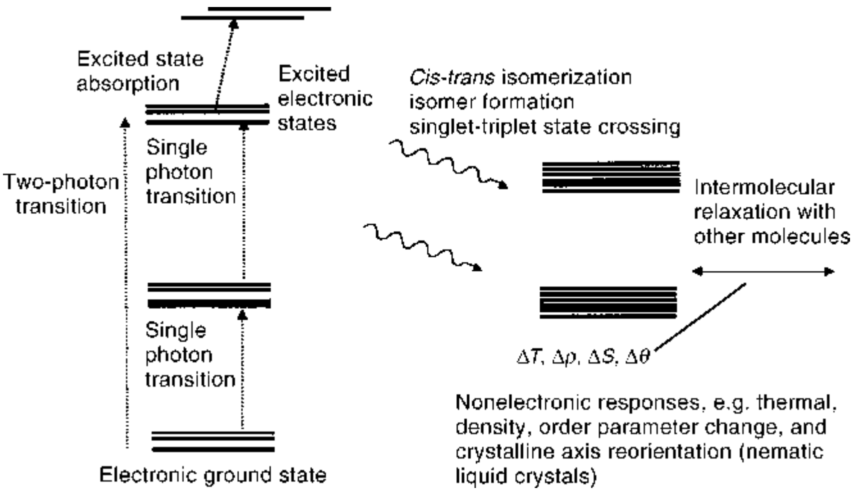
Excited state absorption
Excited electronic states
Two-photon transition
Single photon transition
Single photon transition
Electronic ground state
Cis-trans isomerization
isomer formation
singlet-triplet state crossing
Intermolecular relaxation with other molecules
$\Delta T, \Delta \rho, \Delta S, \Delta \theta$
Nonelectronic responses, e.g. thermal, density, order parameter change, and crystalline axis reorientation (nematic liquid crystals)

Figure 3 Schematic depiction of the typical energy level structure of liquid crystal molecules and various linear and nonlinear optical processes.

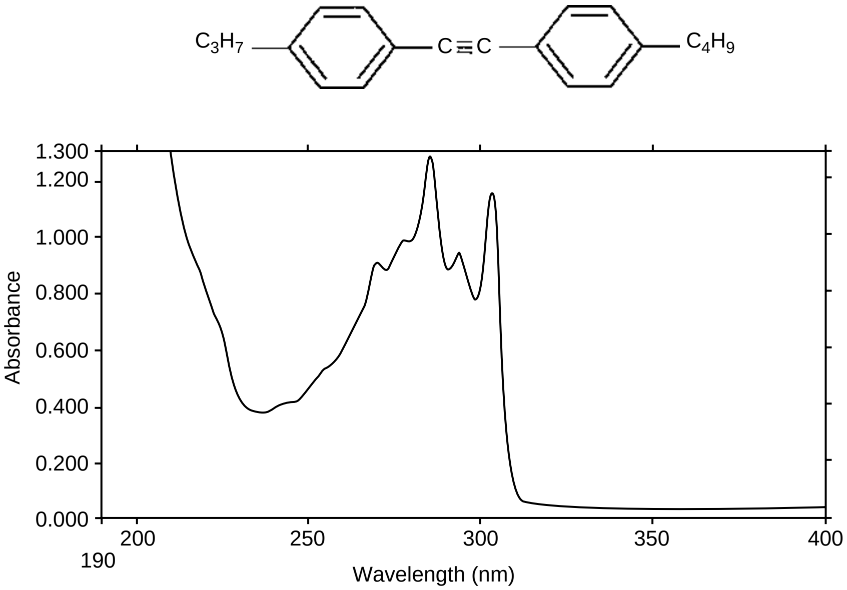

Figure 4 Molecular structure and absorption spectrum of typical isotropic liquid crystal molecules.

absorption (ESA), respectively. As we will see presently, these so-called *electronic nonlinearities* are quite useful for optical limiting applications.

Following photo-absorption, the liquid crystal molecules could undergo intersystem crossing, *cis-trans* or other conformational changes, and form isomers or charge transfer complexes. These processes will contribute to the nonelectronic nonlinearities such as reorientational, order parameter changes and thermal effects.

Orientational Nonlinearities

Purely Optically Induced Orientation

Figure 5 depicts schematically various processes that could occur in an aligned nematic film upon irradiation by an optical field. The reorientation of the director axis $\hat{n}$ is governed by the system's tendency to minimize the overall free energy, which consists of several bulk and surface interaction terms, and the interaction energy with the applied field. In purely dielectric nonabsorbing nematic liquid crystals, director axis reorientation is caused by the torque Γ_{opt} between the anisotropic nematic and the optical electric field,

$$\Gamma_{\text{opt}} = \left(\frac{\Delta\varepsilon}{8\pi}\right)(\hat{n}\cdot\mathbf{E}^*)(\hat{n} \times \mathbf{E}) \qquad [1]$$

which competes with the elastic restoring torque Γ_{el} of the surrounding NLC molecules:

$$\Gamma_{\text{el}} = \frac{(K_1 \sin^2\theta + K_3 \cos^2\theta)\mathrm{d}^2\theta}{\mathrm{d}z^2} + \left[(K_1 - K_3)\sin\theta\cos\theta\right]\left(\frac{\mathrm{d}\theta}{\mathrm{d}z}\right)^2 \qquad [2]$$

where the K's are the elastic constants, and $\Delta\varepsilon$ is the (optical) dielectric anisotropy.

As a result of the director axis reorientation, an optical wave propagating through the medium will 'see' a refractive index change. A frequently employed method for measuring the refractive index change in nematic liquid crystal films is the grating diffraction setup, as shown in Figure 6a. Two polarized coherent writing beams are overlapped on a nematic film, which is probed by a polarized reading beam. If the writing beams are copolarized, they impart an optical intensity grating onto the nematic film $I = I_0(1 + \cos qx)$, where q is the grating wavevector, cf. Figure 6b. On the other hand, if the writing beams are orthogonally polarized, the resulting intensity is uniform over the illuminated area, and instead a polarization grating is created, cf. Figure 6c. The polarization state of the light in the overlapped region varies from circular, through elliptical to linear in accordance to the phase difference between the two coherent writing beams. Using a linearly polarized probe beam, this setup allows one to distinguish various intensity and reorientation dependent effects, and assess their contribution to the induced refractive index change Δn.

The magnitude of the induced reorientation angle, and therefore the extraordinary index change, depends on whether the interaction is transient or steady state. In the steady state case, to the first order of approximation, Δn is proportional to the optical intensity I, i.e., $\Delta n = n_2^{\text{SS}} I$. Under some experimental conditions, one can also show that the transient nonlinear index coefficient n_2 (τ_{p}) is given by n_2 $(\tau_{\text{p}}) \sim n_2^{\text{SS}}\tau_{\text{p}}/\tau_{\text{r}}$, where $\tau_{\text{r}} = \gamma d^2/K\pi^2$ is the orientational relaxation time constant. For $d = 100$ μm, $\gamma = 0.1$ poise and $K = 10^{-6}$ dynes, we have $\tau_{\text{r}} \sim 1$ s. Typical observed values of $n_2^{\text{SS}}(\theta)$ are

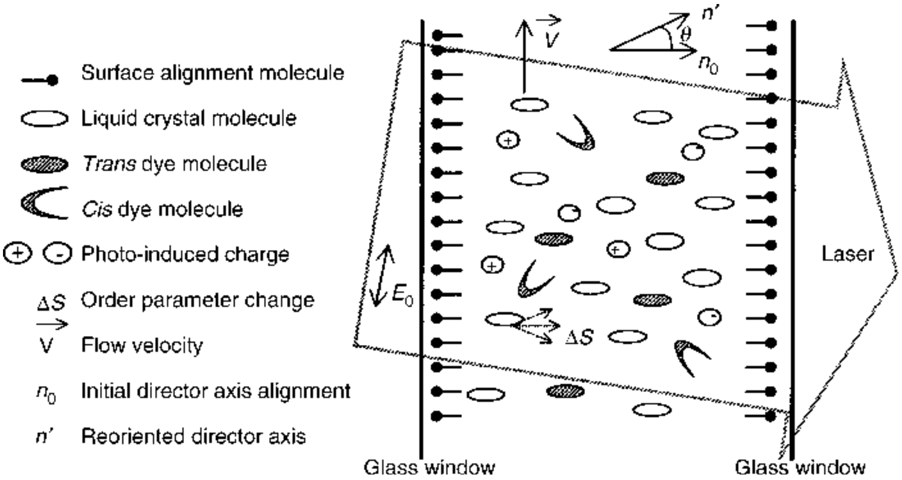

Figure 5 Schematic depiction of various laser-induced optical processes in a nematic liquid crystal film.

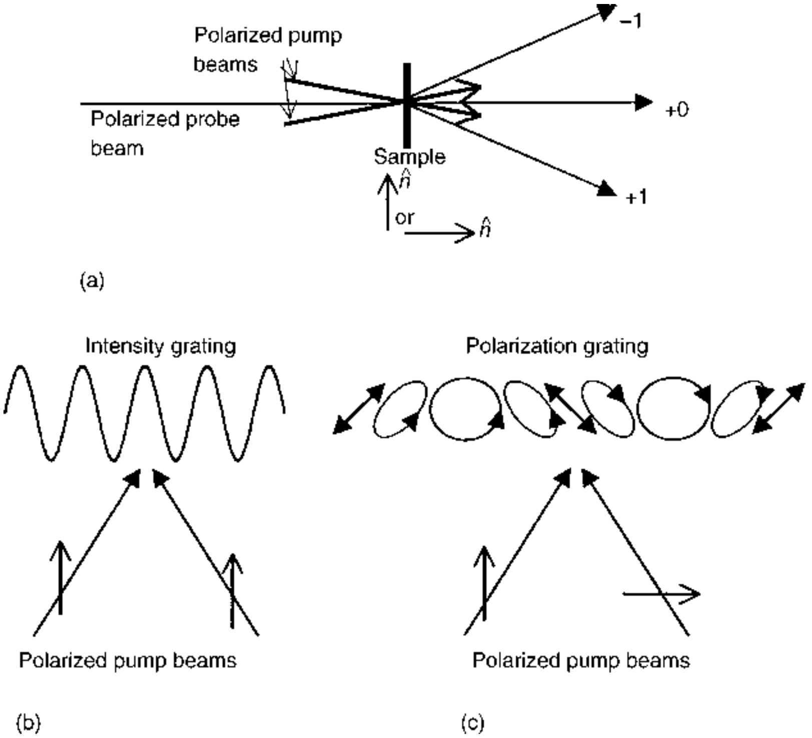

Figure 6 (a) Schematic depiction of grating formation and probe diffraction from the index grating induced in a nematic film. (b) Optical intensity grating formed by interfering two coherent copolarized writing beams. (c) Optical polarization grating formed by interfering two coherent crossed-polarized writing beams.

in the range of 10^{-5}–10^{-4} cm^2 W^{-1}. In the transient case, if $\tau_p = 1$ ns, then n_2 $(\tau_p) \sim 10^{-12}$ cm W^{-1} [$\chi^{(3)} \sim 2 \times 10^{-11}$ esu].

Excited Dye-Dopant Induced Intermolecular Torque

Studies of nematic liquid crystals doped with anthraquinone or dichroic dyes have shown that the excited dye molecules could exert an intermolecular torque Γ_{mole} that could be much larger in magnitude than Γ_{op}. By considering the interaction energies between ground and excited dyes with the surrounding LC molecules, one can show that Γ_{mole} may be written as:

$$\Gamma_{\text{mole}} = \zeta \Gamma_{\text{op}} \qquad [3]$$

where the proportionality constant ζ can be much larger than unity. The nonlinear index coefficients n_2 obtained in these films are on the order of 10^{-3} cm^2 W^{-1}. The response times for the director axis reorientation and relaxation are similar to the optically induced reorientation process discussed in the preceding section.

Photorefractive Effect

Another mechanism, namely, orientational photorefractivity, could also result in a very large optical nonlinearity. This effect occurs in nematic liquid crystals doped with photo-charge producing agents such as fullerene C_{60} or rhodamine 6G dye, in conjunction with a small applied dc field. For a 25 μm thick film with a dc bias voltage of ~1 V, the observed nonlinear coefficient n_2 is in the range of 10^{-3}–10^{-2} cm^2 W^{-1}. Figure 7 illustrates the orientational photorefractive effect. An incident optical field excites the doped liquid crystals to form charge transfer complexes (CTCs). The CTCs subsequently dissociate and create dc space charge fields through ionic diffusion, migration and other electrodynamical processes. The space charge fields, in combination with the applied dc field, create a dc field induced torque of the form:

$$\Gamma_{\text{dc}} = \left(\frac{\Delta\varepsilon_{\text{dc}}}{8\pi}\right)(\hat{n}\cdot\mathbf{E}^*_{\text{int}})(\hat{n} \times \mathbf{E}_{\text{int}}) \qquad [4]$$

Here the total internal dc field $\mathbf{E}_{\text{int}}$ contains the applied dc field $\mathbf{E}_{\text{dc}}$ and the generated dc space charge fields $\mathbf{E}_{\text{sc}}$. This causes director axis reorientation and refractive index change.

The space charge fields consist of one photorefractive and two anisotropy components. For an incident optical intensity grating function of the form $I_{\text{op}} = I_0(1 + m \cos q\xi)$, the photorefractive space charge

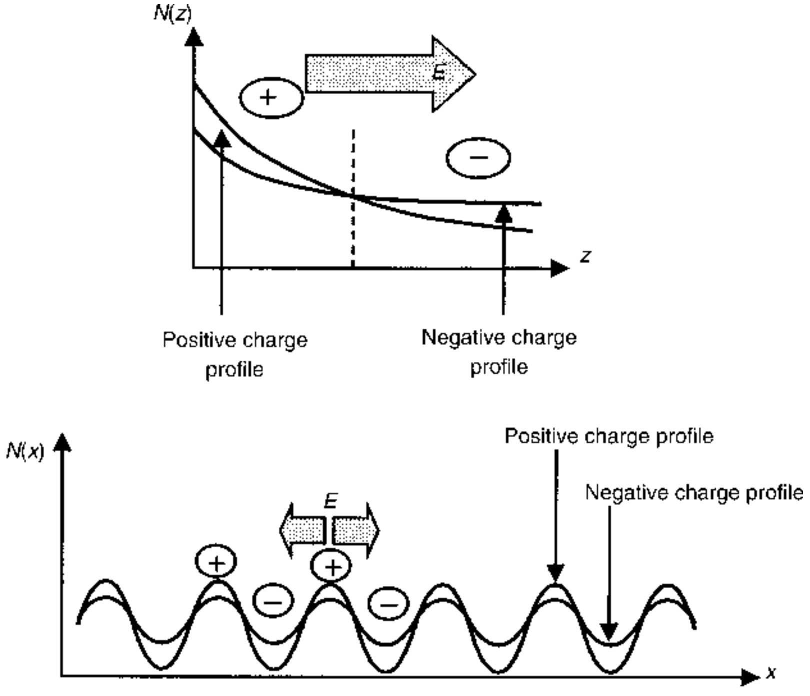

Figure 7 Schematic depiction of photo-induced space charge field formation in a nematic liquid crystal illuminated by a monotonic (upper curve) or a sinusoidal (lower curve) laser intensity function.

field is given by:

$$E_{\mathrm{ph}} = q\nu\, mk_{\mathrm{B}}T\left(\frac{\sigma - \sigma_{\mathrm{d}}}{2e\sigma}\right)\cos\left(\frac{q\xi - \pi}{2}\right) \qquad [5a]$$

where k_{B} is the Boltzmann constant, σ is the illuminated conductivity, σ_{d} is the dark state conductivity, $\nu = (D_+ - D_-)/(D_+ + D_-)$, where D_+ and D_- are the diffusion constants for the positively and negatively charged ions, respectively; m is the optical intensity modulation factor, $q = 2\pi/\Lambda$ is the magnitude of the grating wavevector, and ξ is the coordinate along q.

Two other sources of space charge fields are also at work, and they arise from the conductivity and dielectric anisotropies of the nematic under the action of the dc bias field E_{dc}:

$$E_{\Delta\sigma} = E_{\mathrm{dc}}\left(\frac{(\sigma_{//} - \sigma_{\perp})\sin\theta\cos\theta}{\sigma_{//}\sin 2\theta + \sigma_{\perp}\cos 2\theta}\right) \qquad [5b]$$

$$E_{\Delta\varepsilon} = E_{\mathrm{dc}}\left(\frac{(\varepsilon_{//} - \varepsilon_{\perp})\sin\theta\cos\theta}{\varepsilon_{//}\sin 2\theta + \varepsilon_{\perp}\cos 2\theta}\right) \qquad [5c]$$

where $(\sigma_{//} - \sigma_{\perp})$ is the low-frequency **conductivity anisotropy** and $(\varepsilon_{//} - \varepsilon_{\perp})$ the **dielectric anisotropy** of the liquid crystal, and θ is the reorientation angle of the director axis.

In addition to these space charge field induced effects, the **flows** of the charged ions and complexes under the applied dc field and the generated space charge fields also contribute significantly to reorienting the director axis. For an interaction geometry similar to Figure 5, the flow creates a shear torque Γ_{shear} of magnitude proportional to the velocity gradient

$$\Gamma_{\mathrm{shear}} = \frac{\mathrm{d}v(\zeta)}{\mathrm{d}\zeta}(\alpha_2\sin^2\theta - \alpha_3\cos^2\theta) \qquad [6]$$

where the α's are the viscosity coefficients, and $\mathbf{v}(\zeta)$ is the flow velocity. The combined action of the above space charge fields and the flow effects on director axis reorientation have been quantitatively formulated and shown to describe experimental results very well, including, in particular, the observation that maximal diffraction is obtained for a grating constant $\lambda_{\mathrm{g}} \sim 2d$. Furthermore, the flow-reorientation and adsorption of the excited dopant molecules on the surfaces are believed to be responsible for quasi-permanent and permanent director axis reorientation in rhodamine 6G- and C_{60}-doped nematic films, respectively.

Supra-Nonlinear Methyl-Red Doped NLC – Observed Phenomena

Recent studies in methyl-red doped nematic liquid crystals (MRNLCs) have ushered in the era of supra-nonlinearity. Unlike photorefractive nematic liquid crystals, no external bias field is needed. The observed n_2 values could be as large as $10\,\mathrm{cm}^2\,\mathrm{W}^{-1}$. In MRNLC, a large photovoltage drop across the sample is also observed upon irradiation. The polarity of the photovoltage is reversed if the light impinges from the opposite cell window. The extremely nonlinear optical response of MRNLC is attributed to nematic director axis reorientation by the optically induced space charge fields, similar to the effect discussed in the previous section. The unusually large response is possibly due to several factors, including the higher photo-charge producing abilities of methyl red, and the large difference in the positive and negative photo-ion diffusion constants. Another cause of director axis reorientation is the intermolecular torque exerted by the photo-excited dye molecules on the NLC. This is particularly evident in experiments using an input polarization grating on planar aligned samples. Because the laser excited dye molecules are adsorbed and strongly 'anchored' on the cell walls, they cause permanent realignment of the nematic liquid crystal. This process usually occurs upon prolonged illumination of the liquid crystal cells with a low-power laser, and can be sped up with higher laser power.

Trans–Cis Isomerization and Order Parameter Change

Some nematic liquid crystals are azo-compounds, i.e., their molecular structures are analogous to the typical one depicted in Figure 8. An azo-compound when irradiated at its absorption wavelength will change from the *trans* to the *cis* form, cf. Figure 8. *Trans–cis* isomerization of these azo-nematic liquid crystals or azo-dopants in NLC's could also give rise to a refractive index change through modification of the NLC's order parameter and/or director axis reorientation. The *trans* species are oblong in shape, and align parallel to the director axis. On the other hand, the 'bent' structure of the *cis* species is incongruous to the nematic order, and thus causes an order parameter change, and therefore a refractive index modification. Such refractive index changes owing to order parameter modification are similar to thermal effects discussed in the next section. The response [on-] time for these *trans–cis* isomerization induced order-parameter changes can be quite fast (as fast as nanoseconds in some cases). The relaxation time (typically in the millisecond regime) is decided by the director axis relaxation rate as discussed earlier.

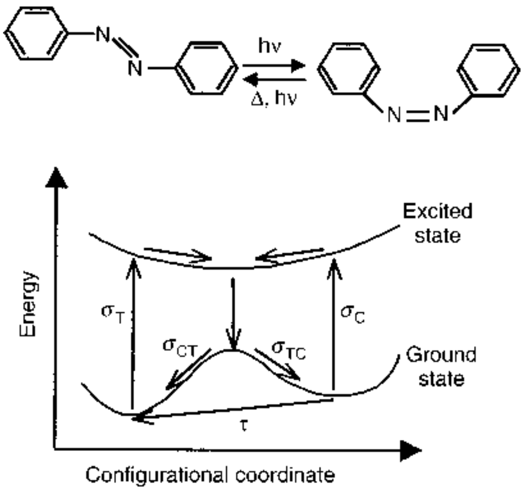

Figure 8 *Trans–cis* isomerization of azo-molecules upon optical illumination. The upper diagram depicts molecular structural changes; the lower diagram shows the excitation energy versus the molecular coordinate associated with these photo-excited processes.

The other index-changing mechanism is the tendency of the *trans* molecule to align with its transition moment perpendicular to the optical polarization in order to minimize the interaction (absorption) energies. This in turn causes the NLC director axis to realign through the intermolecular torque effect. Both mechanisms are characterized by a nonlinear index coefficient n_2 on the order of $10^{-3}\,\mathrm{cm}^2\,\mathrm{W}^{-1}$.

Thermal and Density Effect

Laser induced temperature and density changes are described by time-dependent coupled hydrodynamics equations. In grating diffraction experiments, cf. Figure 6, the problem can be simplified to a one-dimensional one, i.e., along the grating wavevector direction, if the grating spacing is smaller than the nematic film thickness and the beam size. In that case, the thermal and Brillouin (sound wave) decay time constants are given by, respectively:

$$\tau_T = \frac{\rho_0 C_v}{\lambda_T q^2} \qquad [7a]$$

$$\tau_B = \frac{2\rho_0}{\eta q^2} \qquad [7b]$$

where $q = 2\pi/\lambda_q$ is the grating wavevector, ρ_0 the unperturbed density, C_v the specific heat, λ_T the thermal conductivity, and η the viscosity. The values of these parameters depend on the director axis alignment relative to the optical fields, thermal diffusion and sound propagation directions. For typical values of $\eta = 7 \times 10^{-2}\,\mathrm{kg\,m^{-1}\,s^{-1}}$, $\rho_0 = 10^3\,\mathrm{kg\,m^{-3}}$, $D = \lambda_T/\rho_0$, $C_v \sim 1.6 \times 10^{-7}\,\mathrm{m^2\,s^{-1}}$ and a grating spacing $\lambda_q = 20\,\mu\mathrm{m}$, we have $\tau_T \sim 50\,\mu\mathrm{s}$, whereas $\tau_B \sim 200\,\mathrm{ns}$. Such a vast difference between the relaxation times allows one to distinguish the contribution from thermal and density origins.

Table 1 Nonelectronic nonlinearities of nematic liquid crystals

Nonlinear processes	*Typical magnitude*	*Time-scale*	*Liquid crystal*
Orientational nonlinearities			
Purely optically induced n_2	10^{-4} cm^2 W^{-1}	ms	pure 5CB
Excited dopant assisted n_2	10^{-3} cm^2 W^{-1}	ms	dyed 5CB
Photorefractive	10^{-3} cm^2 W^{-1}	ms	5CB
Photorefractive + Methyl-red doped 5CB	10 cm^2 W^{-1}	ms	
Trans–Cis and order parameter modification	10^{-4} cm^2 W^{-1}	ms	dyed 5CB
Thermal nonlinearities			
n_2	10^{-6} cm^2 W^{-1}	ms	E7
n_2	10^{-11} cm^2 W^{-1}	ns	E7
Density effect			
electrostrictive n_2	10^{-11} cm^2 W^{-1}	ns	E7

In the steady state, i.e., when the laser–nematic interaction time is much longer than τ_T and τ_B, the nonlinear index coefficient is given by:

$$n_2^{SS}(\Delta T) = \left(\frac{dn}{dT}\right)\frac{\alpha\tau_r}{\rho_0 C_V} \quad [8]$$

where dn/dT is the appropriate thermal index coefficient. The absorption constant α of liquid crystals varies widely in the visible–infrared spectrum. Near the phase transition temperature, $dn_{//,\perp}/dT$ also changes dramatically. The magnitude of $n_2^{SS}(\Delta T)$ could therefore vary by several orders of magnitude. For $\alpha = 100\ \text{cm}^{-1}$, $dn_{//,\perp}/dT \sim 10^{-3}\ \text{K}^{-1}$, and $\lambda_q = 20\ \mu\text{m}$, we have $n_2^{SS}(\Delta T) \sim 10^{-6}\ \text{cm}^2\ \text{W}^{-1}$. This is the typical value observed in wave mixing experiments with infrared [CO_2] or visible lasers.

In the transient regime ($\tau_p \ll \tau_r, \tau_B$), the nonlinear index coefficient is proportional to the laser pulse duration:

$$n_2(\tau_p) = n_2^{SS}(\Delta T)\tau_p/\tau_r \quad [9]$$

With $\tau_r \sim 0.5 \times 10^{-4}$ s and $\tau_p = 1$ ns, $n_2(\tau_p) \sim 5 \times 10^{-11}\ \text{cm}^2\ \text{W}^{-1}$. Such nonlinearities have been observed in transient dynamic grating diffraction studies involving nanosecond laser pulses. In these studies, it is found that the contribution from the electrostrictive effect (movement of the nematic due to the optical intensity gradient) to the refractive index change can be as large as the absorptive thermal component.

In Table 1, we give a concise summary of these nonelectronic optical nonlinearities observed in nematic liquid crystals.

Electronic Nonlinearities

Second and Third Order Nonlinear Susceptibilities

Electronic interaction involves the perturbation of the electronic wavefunctions by the impinging optical fields. The induced dipole moment of a molecule is given by:

$$\mathbf{d} = \boldsymbol{\alpha}:\mathbf{E} + \boldsymbol{\beta}:\mathbf{EE} + \boldsymbol{\gamma}:\mathbf{EEE} + \cdots \quad [10]$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are the linear, second-order and third-order nonlinear polarizability tensors, $\mathbf{E}$ is the optical electric field and : denotes tensorial operation. The macroscopic polarization $\mathbf{P}$ (dipole moment per unit volume) is

$$\mathbf{P} = \varepsilon_0\boldsymbol{\chi}^{(1)}:\mathbf{E} + \boldsymbol{\chi}^{(2)}:\mathbf{EE} + \boldsymbol{\chi}^{(3)}:\mathbf{EEE} + \cdots \quad [11]$$

where $\boldsymbol{\chi}^{(1)}$, $\boldsymbol{\chi}^{(2)}$, and $\boldsymbol{\chi}^{(3)}$ are the linear, second and third-order nonlinear susceptibilities, respectively. The macroscopic parameters $\boldsymbol{\chi}$ are related to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by the local field correction factors.

A well-known process associated with the second-order nonlinear polarization is second harmonic generation, in which two incident photons of frequency ω are converted into a photon of frequency 2ω. Symmetry rules require that a material should be non-centrosymmetric in order to exhibit second-order nonlinear polarization. In centrosymmetric nematic liquid crystals, an applied dc field or discontinuity at a surface or interface, or director orientational curvature, could provide the required symmetry-breaking mechanism.

In ferroelectric or smectic C* liquid crystals, the molecules possess permanent polarization, i.e., the centrosymmetry is broken, and it is possible to observe second harmonic generation associated with the second-order nonlinear polarization. Typically, the observed nonlinear coefficients β_{2HG} are on the order of $\sim 10^{-14}\ \text{mV}^{-1}$. In contrast to inorganic crystals of much larger dimensions, these thin liquid crystal films are not useful for practical devices. Nonetheless, because of the sensitivity of second harmonic generation to surface and interface conditions, second harmonic generation has been proven to be an effective spectroscopic tool.

In dc field induced second harmonic generation, the nonlinear polarization involved is actually the

Table 2 Electronic optical nonlinearities of nematic liquid crystals

Nonlinear optical phenomena and nonlinearities	*Magnitude*	*Time-scale (pulse duration of laser used)*
Second harmonic generation (dc field induced),		
$\chi_{//}^{(3)}$	2.7×10^{-12} esu	ns
$\chi_{\perp}^{(3)}$	0.55×10^{-12} esu	ns
Second harmonic generation β_{2HG}	10^{-12} mV^{-1}	ps (ferroelectric)
Two-photon absorption, β	5 cm GW^{-1}	80 ps
Effective nonlinear absorption, β_{eff}	200 cm GW^{-1}	10 ns
Degenerate four-wave mixing		
electronic $\chi^{(3)}(\omega;\ \omega, -\omega, \omega)$	10^{-12} esu	ns

third-order nonlinear polarization given by $P(2\omega) = \boldsymbol{\chi}^{(3)}E(0)E(\omega)E(\omega)$. Typically measured values of $\boldsymbol{\chi}^{(3)}$ are $\sim 10^{-12}$ esu, expressed in the commonly used cgs units. The conversion factor between cgs and SI unit is given by

$$\boldsymbol{\chi}^{(3)}[\text{esu}] = 3^4 \times 10^{17} \boldsymbol{\chi}^{(3)}[\text{SI unit}] \qquad [12]$$

There is relatively little work done on measuring or studying the third-order nonlinear electronic polarizabilities of nematic liquid crystals. Most of the studies are centered on the molecular correlation effects associated with the nematic–isotropic phase transition. The typical magnitude of electronic $\boldsymbol{\chi}^{(3)}$ is 10^{-12} esu.

Nonlinear Absorption and Limiting

Two-photon absorption occurs where the molecule is excited to a real electronic excited state by simultaneous absorption of two photons from the incident light. Since the absorption is proportional to the square of the optical intensity, the incident light will experience an increasing absorption effect, and its transmission at high intensity will be 'quenched' to an almost constant level. Two-photon absorption is usually characterized by the two-photon absorption coefficient β, which is related to the imaginary part of $\boldsymbol{\chi}^{(3)}$. For liquid crystals, they are on the order of 5×10^{-12}–10^{-11} mW^{-1} (equivalent to 0.5–5 cm GW^{-1} in the more commonly used units).

Liquid crystalline materials in their (less scattering) isotropic phase have been extensively studied in the last few years for application in optical limiting devices as a result of their large nonlinear absorption constant. In the nanosecond time regime, the RSA and ESA effects, cf. Figure 3, act to give rise to a very large effective nonlinear absorption coefficient β on the order of several hundreds of cm GW^{-1}. In combination with special device structure such as a nonlinear fiber array, it has become possible to clamp the transmitted nanosecond laser pulse energy to below the retina damage level (below 0.5 μJ).

In Table 2, we give a concise summary of these electronic optical nonlinearities of nematic liquid crystals.

Concluding Remarks

We have presented a concise review of the nonlinear optics of liquid crystals in their nematic and isotropic phases. Nematic liquid crystals in their various pure and doped forms possess many interesting nonlinear optical responses, with nonlinearities ranging over 16 orders of magnitude. Some of the extraordinarily large nonlinearities enable the performance of several all-optical switching and limiting, image modulation and sensing processes at unprecedented low threshold power. In particular, self-defocusing and optical limiting effects with nanowatt-power lasers, and incoherent–coherent image conversion at μW cm^{-2} optical intensity have been demonstrated. Liquid crystal films have been used in selective attenuation of glares and blinding laser light. Stimulated scattering that used to be the domain of high-power lasers can now be performed in thin nematic films with mW-power cw laser. Readers can consult the Further Reading section for some of these advanced photonic devices and applications based on nonlinear optics of liquid crystals. Current active studies by several research groups worldwide are likely to result in even more interesting nonlinear optical effects and useful applications in the near future.

List of Units and Nomenclature

Absorption coefficient	α [cm^{-1}]
Density	ρ_0 [kg m^{-3}]
Eleastic constant	K [dynes]
Intensity (optical)	I [W cm^{-2}]
Nonlinear susceptibility	$\chi^{(3)}$ [esu]
Second harmonic generation coefficient	β_{2HG} [mV^{-1}]
Specific heat	C_V [m^2 s^{-1}]
The nonlinear index coefficients	n_2 [cm^2 W^{-1}]
Time constants	τ [s]

Two-photon absorption constant	β [cm GW^{-1}]
Viscosity	η [kg m^{-1} s^{-1}]
Viscosity coefficient	γ [poise]

See also

Nonlinear Optics, Basics: Cascading; Kramers–Krönig Relations in Nonlinear Optics; Nomenclature and Units; Photorefraction.

Further Reading

Khoo IC (1995) *Liquid Crystals: Physical Properties and Nonlinear Optical Phenomena*. New York: Wiley Interscience.

Khoo IC (1996) Orientational photorefractive effects in nematic liquid crystal films. *IEEE Journal of Quantum Electronics* 32: 525–534.

Khoo IC and Liang Y (2000) Stimulated orientational and thermal scatterings and self-starting optical phase conjugation with nematic liquid crystals. *Physical Review* E62: 6722–6733.

Khoo IC and Wu ST (1993) *Optics and Nonlinear Optics of Liquid Crystals*. Singapore: World Scientific.

Khoo IC, Chen PH, Wood MV and Shih MY (1999) Molecular photonics of a highly nonlinear organic fiber core liquid for picosecond–nanosecond optical limiting effect. *Chemical Physics* 245: 517–531.

Khoo IC, Wood MV, Shih MY and Chen PH (1999) Extremely nonlinear photosensitive liquid crystals for image sensing and sensor protection. *Optics Express* 4(11): 431–442.

Khoo IC, Shih MY, Wood MV, *et al.* (1999) Dye-doped photorefractive liquid crystals for dynamic and storage holographic grating formation and spatial light modulation. *IEEE Proceedings* 87: 1897–1911.

Marrucci L and Papara D (1997) Photoinduced molecular reorientation of absorbing liquid crystals. *Physical Review* E56: 1765–1772.

Rau H (1990) Photoisomerization of Azobenzenes. In: Rabeck FJ (ed.) *Photochemistry and Photophysics*, vol. 2, chap. 4, pp. 119–141. Boca Raton, FL: CRC Press.

Shen YR (1990) *Principles of Nonlinear Optics*. New York: Wiley Interscience.

Organic Nonlinear Materials

F Kajzar and I Rau[*], Commissariat à l'Energie
Atomique, Gif sur Yvette, France

© 2005, Elsevier Ltd. All Rights Reserved.

Nonlinear Medium Polarization

Macroscopic

Photons are neutral, noninteracting bosons. Interaction is possible only through a material medium. Indeed, photons can be absorbed by electrons leading to excited states. A reverse transition may go through radiative channels, with emission of photons, usually with a different energy than the excitation energy. Interaction of photons with matter may also lead to other types of excitation, such as vibronic, rotational, libronic, etc., or just to a change in the medium polarization. This change may be expanded in the power series of the electric forcing field strength with *i*th component given by:

$$\begin{aligned} P_I(\omega_\sigma) = {} & P_{0I}(\omega_\sigma) + \varepsilon_0[K_1\chi^{(1)}_{IJ}(-\omega_\sigma;\omega_\sigma)E_J^{\omega_\sigma} \\ & + K_2\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2)E_J^{\omega_1}E_K^{\omega_2} \\ & + K_3\chi^{(3)}_{IJKL}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)E_J^{\omega_1}E_K^{\omega_2}E_{\mathrm{L}}^{\omega_3} + \cdots] \end{aligned} \qquad [1]$$

where the expansion coefficients $\chi^{(n)}$ are $(n+1)$ rank three-dimensional tensors, describing linear $(n=1)$ and nonlinear $(n>1)$ responses of the medium, $P_{0i}(\omega_\sigma)$ is its permanent polarization (at frequency ω_σ) and the coefficients K_i $(i=1,2,3,\ldots)$ take account of the conventions for the Fourier transform of electric field and polarization vectors used, as well as the degeneracy of nonlinear optical processes. For historical reasons two conventions in quantitative description of NLO processes are used:

$$\mathrm{E}(\mathrm{r},t) = \tfrac{1}{2}[\mathrm{E}(\mathrm{r})\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{c.c.}] \qquad \text{(Convention I)} \qquad [2]$$

and

$$\mathrm{E}(\mathrm{r},t) = \mathrm{E}(\mathrm{r})\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{c.c.} \qquad \text{(Convention II)} \qquad [3]$$

and similarly for the polarization field. As researchers often use different conventions, it is important to bear this in mind when comparing NLO properties of materials coming from different sources. The tabulated data and the formulas used here are within convention I, with the permutation factors included explicitly in coefficients K_i (eqn [1]). In centrosymmetric media and in dipolar approximation, the static polarization P_0, as well as all even order NLO susceptibilities, are equal to zero, as it follows from time reversal symmetry.

---

[*] On leave from POLITECHNICA University of Bucharest, Romania.

Relations Between Macroscopic and Microscopic Polarizations

Microscopic Polarization

Similar expansions to eqn [1] also obey the conventions on a molecular level. Similarly, as for bulk polarization, the ith component of the molecule dipole moment can also be expanded into the series of the local field strength giving:

$$\begin{aligned}\mu_i(\omega_\sigma) &= \mu_{0i}(\omega_\sigma) + K_1\alpha_{ij}(-\omega_\sigma;\omega_\sigma)E_j^{\omega_\sigma}\\ &\quad + K_2\beta_{ijk}(-\omega_\sigma;\omega_1,\omega_2)E_j^{\omega_1}E_k^{\omega_2}\\ &\quad + K_3\gamma_{ijkl}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)E_j^{\omega_1}E_k^{\omega_2}E_l^{\omega_3} + \cdots\\ &= \mu_{0i}(\omega_\sigma) + \mu_{\text{ind}}(\omega_\sigma,\omega_1,\omega_2,\ldots,\omega_n,\alpha,\beta\ldots,\zeta,\\ &\quad E^{\omega_1},E^{\omega_2},\ldots,E^{\omega_n}) \qquad [4]\end{aligned}$$

where α is the linear molecular polarizability, β is the first order and γ is the second molecular hyperpolarizability. μ_{in} in eqn [4] denotes the induced dipole moment. The expansions [1] and [4] may be also considered as the Tamm–Dancoff perturbation development as functions of an increasing number of interacting photons. In this case, the second term on the right-hand sides of eqns [1] and [4] describes the interaction of two photons, the third term that of three photons, etc. The electric field in eqn [4] is the effective field experienced by the molecule and is given by

$$E = fE \qquad [5]$$

where f is the local field factor, depending on the form of molecule. In eqns [1] and [4] Einstein's notation is used.

For a single crystal and for noninteracting molecules there exists a straightforward relationship between the macroscopic and microscopic quantities given by:

$$\begin{aligned}\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2) &= \sum_n f_n^{\omega_\sigma}f_n^{\omega_1}f_n^{\omega_2}N_n\\ &\quad\times\sum_{ijk} a_{Ii}a_{Jj}a_{Kk}\beta^n_{ijk}(-\omega_\sigma;\omega_1,\omega_2) \qquad [6]\end{aligned}$$

for the second-order NLO susceptibility and

$$\begin{aligned}\chi^{(3)}_{IJKL}(-\omega_\sigma;\omega_1,\omega_2,\omega_3) &= \sum_n f_n^{\omega_\sigma}f_n^{\omega_1}f_n^{\omega_2}f_n^{\omega_3}N_n\\ &\quad\times\sum_{ijkl} a_{Ii}a_{Jj}a_{Kk}a_{Ll}\gamma^n_{ijkl}(-\omega_\sigma;\omega_1,\omega_2,\omega_3) \qquad [7]\end{aligned}$$

for the third-order NLO susceptibility, respectively. The coefficients a_{Ij} in eqns [6] and [7] are Wigner rotation matrix elements, N_n denoting the number density of species (n). The subscripts ($ijkl$) refer to the molecule, and ($IJKL$) to the laboratory reference frames, respectively.

For disordered (or partly ordered) materials, the macroscopic NLO susceptibility depends on the orientation average of nonzero molecular hyperpolarizability components:

$$\begin{aligned}\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2) &= \sum_n f_n^{\omega_\sigma}f_n^{\omega_1}f_n^{\omega_2}N_n\\ &\quad\times\sum_{ijk} <\beta^n_{ijk}(-\omega_\sigma;\omega_1,\omega_2)>_{IJK} \qquad [8]\end{aligned}$$

and similarly for $\chi^{(3)}$ susceptibility. The average in eqn [8] is taken over all NLO molecule orientations and all β tensor components. Fortunately, the number of the nonzero β tensor components (21 for β tensor) is limited by the molecular symmetry and in the case of charge transfer (CT) molecules, only the component in the charge transfer direction β_{xxx} may be retained. The others are usually significantly smaller and can be neglected with respect to this one.

Second-Order Nonlinear Optical Materials

The research on nonlinear optical materials is boosted by the potentiality of their applications in photonics, and particularly in optical signal processing, image recognition, frequency conversion, and light amplification. New fields of applications open in biology, medicine (photodynamic therapy), material processing (optical engineering), etc. Depending on the targeted application, the materials showing enhanced second-, third-, and/or higher-order nonlinear optical properties are developed and optimized. The optimization is performed on both molecular and macroscopic levels. Indeed, the bulk nonlinear optical response depends not only on the individual hyperpolarizabilities but also on the way the molecules are assembled into the bulk material. As already mentioned, only the noncentrosymmetric materials exhibit the second-order nonlinear optical response.

Molecules

The most promising molecules for the application in second-order NLO devices are intramolecular CT compounds, composed from an electron donating (D), an electron accepting (A) part, linked by a π conjugated backbone (Figure 1). The CT molecules exhibit a large first-order susceptibility β tensor component β_{xxx}, enhanced in the charge transfer direction x. The value of β_{xxx} depends on the strength of electron donating and electron accepting groups (amount of the charge transfer)

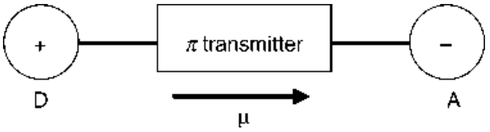

Figure 1 Schematic representation of a charge transfer molecule.

Table 1 Chemical structures of commonly used electron donating and electron accepting groups

Donors	*Acceptors*
$-N(CH_3)_2$	$-NO$
$-NH_2$	$-NO_2$
$-OCH_3$	$-CHO$
$-OH$	$-CN$
$-CH_3$	$-SO_2CH_3$
$-SCH_3$	$-SO_2CF_3$
$-N(CH_3)_2$	$=C(CN)_2$

and on the conjugation length of the transmitter. Table 1 shows the chemical structures of typical electron donating and electron accepting groups. Chemical structures of selected CT molecules are given in Table 2, together with the values of the β_{xxx} tensor component, as determined by the electric field induced second-harmonic generation technique (EFISH).

The dependence of β_{xxx} on the conjugation length L of the conjugated transmitter is given by the scaling law:

$$\beta_{xxx} \propto L^{\alpha_s} \qquad [9]$$

with $\alpha_s > 1$. Usually the conjugated backbone is composed of a number of alternate single/double (or single/triple) bonds and in this case, the scaling law is tested as a function of the number of double (or triple) bonds N:

$$\beta_{xxx} \propto N^{\alpha_\beta} \qquad [10]$$

For a number of CT oligomers with the same electron donating and electron accepting groups, differing only in the length of the conjugated transmitter, the exponent $\alpha_\beta \sim 2.5$. The β_{xxx} value saturates for a larger number of double bonds ($N > 4$, in the case of phenyl oligomers used as transmitter).

A number of studies are in progress on octupolar molecules which exhibit large β hyperpolarizabilities. However, their processing into efficient, noncentrosymmetric bulk materials is still at its initial stage and up to now no practical application has been demonstrated.

Bulk Materials

In order to obtain noncentrosymmetric materials, different approaches are used such as:

(i) single crystals growth;
(ii) epitaxy or heteroepitaxy;
(iii) Langmuir–Blodgett films;
(iv) self-assemblies;
(v) intermolecular charge transfer structures;
(vi) poled polymers; and
(vii) chiral structures.

The growth of good-quality optical single crystals is costly and difficult, especially as the charge transfer molecules usually exhibit a large ground state dipole moment (Table 2). The strong dipole–dipole interaction usually leads to their antiparallel alignment, and as a consequence, a centrosymmetric bulk structure. Only a few molecules crystallize in a noncentrosymmetric structure. One of the approaches used consists of design and synthesis of molecules with a negligible ground state dipole moment such as POM (Table 3). Therefore, a lot of effort has been put into forcing molecules to build into noncentrosymmetric bulk materials (ii)–(vi). Owing to the fact that most of applications are targeted in waveguiding configuration, a significant effort has been made to grow noncentrosymmetric thin films (ii)–(vi). While the chiral molecules form spontaneously noncentrosymmetric films, in other techniques, this is obtained more easily by application of an external electric field (poled polymers), by use of an adequate substrate (epitaxy, pseudoepitaxy, self assembly), intramolecular charge transfer in ordered layered structures (v), and interaction between molecules and molecules–substrate (Langmuir–Blodgett films). In the last case, three types of well-ordered structures, with controllable thickness are obtained: X, Y, Z (Figure 2). Only X and Z type structures give noncentrosymmetric thin films exhibiting second-order NLO properties.

Poled polymers

Up to now, the best results in getting efficient materials for second-order NLO application are obtained with poled polymers. The materials consist of a polymer matrix, exhibiting excellent optical properties, functionalized with active NLO chromophores (CT molecules). Four kinds of structures, shown schematically in Figure 3, are realized:

(i) guest–host systems;
(ii) side chain polymers;
(iii) main chain polymers; and
(iv) thermally or photo-crosslinked polymers.

Table 2 Molecular first hyperpolarizabilities and dipole moments of selected CT molecules

Molecule	*Chemical structure*	λ_{max} *[nm]*	*Solvent*	*Ground state dipole moment, μ [D]*	*Product $\mu\beta$ [10^{-77} $Cm^5 V^{-1}$]*	*β [10^{-40} $m^4 V^{-1}$]*	*λ_{meas} [nm]*	*β_0 [10^{-40} $m^4 V^{-1}$]*
p-NA (*para*-nitroaniline)	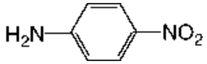	358	1,4 dioxane	5.5	1.30 0.74	71.0 40.2	1064 1907	35
MNA (2-methyl-4-nitroaniline)	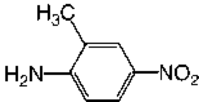	361	1,4 dioxane	7.4	1.72 0.98	69.9 39.8	1064 1907	33
NPP (*N*-(4-nitrophenyl)-(L)-prolinol)	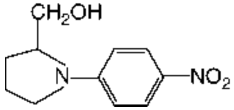	397	acetone	7.3	4.28	176	1064	67
POM (3-methyl-4-nitropyridine-1-oxide)	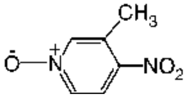	320	from crystal data	0	0	36	1064	21
PNP (1-(*N*-prolinol)-5-nitropyridine)	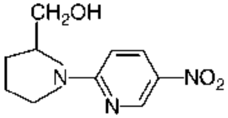	376	1,4 dioxane	7.2	2.28 1.41	95 59	1064 1907	38
MMONS (3-Methyl-4-methoxy-4′ nitrostilbene)	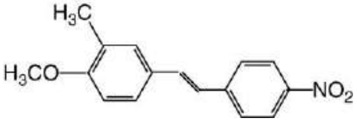	366	1,4 dioxane	5.2	2.14	123.9	1907	90
DR1 (4-[*N*-ethyl-*N*-(2-hydroxyethyl)]amino-4′-nitroazobenzene (dispersed red 1))	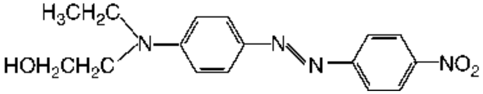	509 455	DMSO 1,4 dioxane			525 234.1	1356 1907	197 150
DADAB (4-(*N*,*N*-diethylamino)-4′-tricyanovinylazo benzene)	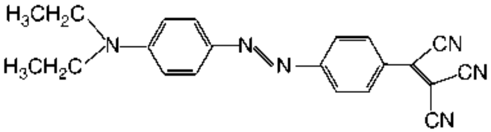	582	DMSO			1640	1580	648

1D = (1/3) × 10^{-39} Cm.

Table 3 Bulk second-order nonlinear optical properties of selected organic materials

Material	*Chemical structure*	*Form*	*Point group*	*λ_{meas} [nm]*	*Refractive index*	*SHG susceptibility $\chi_{ijk}^{(2)}(-2\omega;\omega,\omega)$ [pm V^{-1}]*	*Electrooptical coefficient, r, [pm V^{-1}]*
MNA (2-methyl-4-nitroaniline)	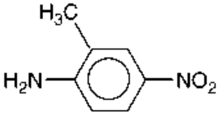	Single crystal	*m*	514.5 1064	2.0	$\chi_{111}^{(2)} = 500$ $\chi_{122}^{(2)} = 76$	70 ± 14
POM (3-methyl-4-nitropyridine-1-oxide)	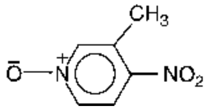	Single crystal	222	514.5 600 10 600 1064	1.712 1.919 1.638	$\chi_{123}^{(2)} = 19.2$ $\chi_{123}^{(2)} = \chi_{251}^{(2)} = \chi_{312}^{(2)}$	4.0 5.1 3.7
COANP (2 cyclooctylamino-5 nitropyridine)	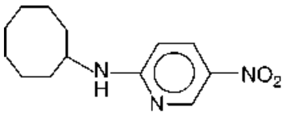	Single crystal	*mm*2	514.5 600 10 600 1064	1.672 1.781 1.647	$\chi_{322}^{(2)} = 108$ $\chi_{333}^{(2)} = 48$	9.3 ± 1.5 15.4 ± 3.2
PNP (2-(N-prolinol)-5-nitropyridine)	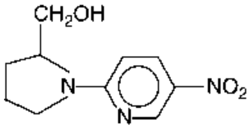	Single crystal	2	514.5 600 10 600 1064	1.990 1.788 1.467	$\chi_{211}^{(2)} = 170$ $\chi_{223}^{(2)} = 55$	8.0 ± 1.4 13.0 ± 2.2
MNBA (4′-nitrobenzylidene-3-acetamino-4-methoxyaniline)	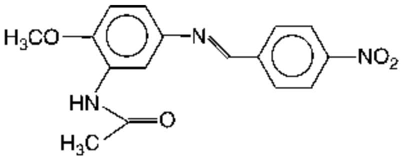	Single crystal	*m*	514.5 600 10 600 1064	2.024 1.585 1.652	$\chi_{111}^{(2)} = 437$	–
MMONS (3-Methyl-4-methoxy-4′-nitrostilbene)	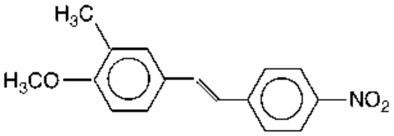	Single crystal	*mm*2	514.5 600 10 600 1064	1.569 1.693 2.129	$\chi_{333}^{(2)} = 368^{a}$ $\chi_{223}^{(2)} = 142^{a}$	19.3 ± 4 39.9 ± 7

PMMA + DR 1 (Poly(methyl methacrilate) + 4-[N-ethyl-N-(2-hydroxyethyl)] amino-4′-nitroazobenzene (Dispersed red 1))	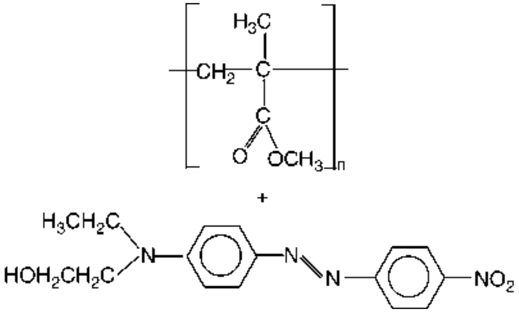	PP/GH	1580	$\chi^{(2)}_{333} = 5$	
			633		$r_{33} = 2.5$
PMMA + DCV (Poly(methyl metacrilate) + 4-dicyanovinyl-4′- [N-ethyl-(2 hydroxyethyl)] azobenzene)	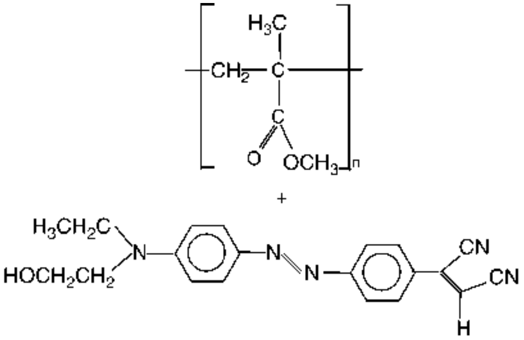	PP/SC	1580	$\chi^{(2)}_{333} = 42$	
			799		$r_{33} = 18$
PMMA + 3RDCVXY (Poly(methyl metacrilate) + 4-dicyanovinyl-4′ (diethylamino) diazobenzene)	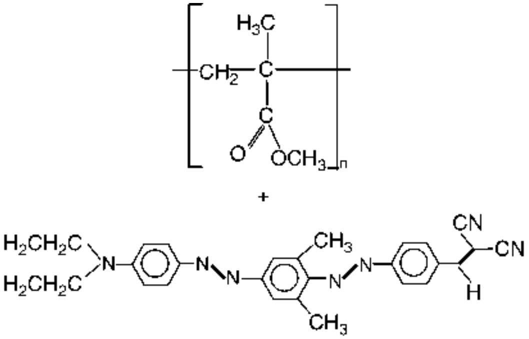	PP/SC	1064	$\chi^{(2)}_{333} = 840$	
			633		$r_{33} = 40$

Table 3 (*Continued*)

Material	*Chemical structure*	*Form*	*Point group*	λ_{meas} *[nm]*	*Refractive index*	*SHG susceptibility* $\chi^{(2)}_{ijk}(-2\omega;\omega,\omega)$ *[pm* V^{-1}*]*	*Electrooptical coefficient, r, [pm* V^{-1}*]*
NNDN + NAN (N,N-diglycidyl-4-nitroaniline + N-(2-aminophenyl)-4-nitroaniline)	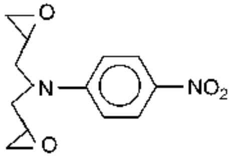 +	PP/CL		1064		$\chi^{(2)}_{311} = 28$	
	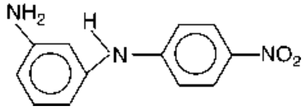			530.9			$r_{13} = 6.5$
DCANP	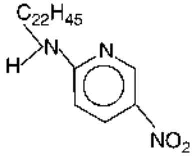	LB film		632.8	1.598 (TE$_0$-mode)		
				1064		$\chi^{(2)}_{333} = 19.5 \pm 2.5$	
FA06	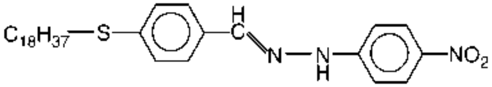	LB film		1064		$\chi^{(2)}_{322} = 20$	

PP, poled polymer; GH, guest host system; SC, side chain polymer; CL, cross linked polymer; LB, Langmuir–Blodgett film.

[a] Calibrated with KTP single crystal ($\chi^{(2)}_{333} = 27.4$ pm/V).

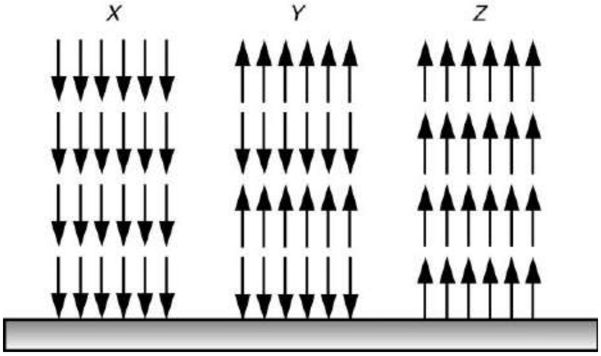

Figure 2 Schematic representation of *X, Y, Z* type Langmuir–Blodgett films. The arrows symbolize the direction of dipole moments, which usually are tilted with respect to the normal to film surface.

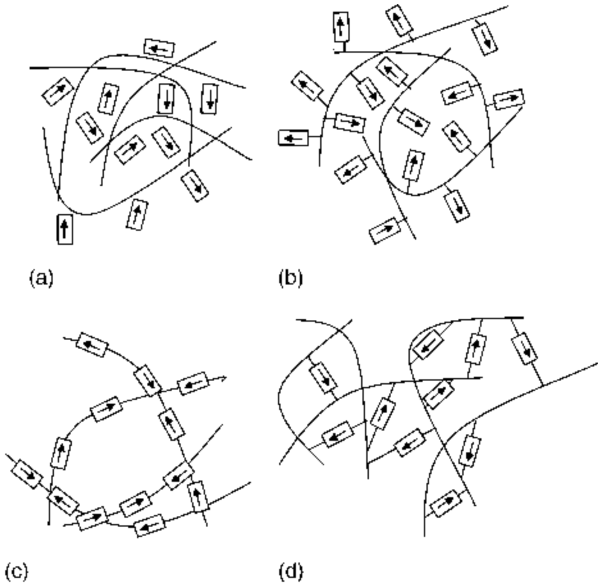

Figure 3 Schematic representation of functionalized polymers for second-order NLO effects: guest–host systems (a), side chain polymers (b), main chain polymers (c), photo and thermally cross-linking polymers (d). Arrows represent the NLO chromophores.

The guest–host systems are just solid solutions of NLO chromophores in an inert matrix. The concentration of NLO chromophores is limited because of already-mentioned strong dipole–dipole interaction leading to antiparallel alignment, resulting in an increase in propagation losses and decrease of $\chi^{(2)}$ susceptibility. The polar order is obtained by applying an external field. In order to increase the rotational mobility of chromophores, the material is heated to the glass transition temperature of polymer. The interaction of dipole moments with the external field forces them to orient in its direction. The orientation is frozen by cooling the material (usually a thin film) to room temperature. However, the stability induced in this way results in a polar order that is poor. Better stability and higher concentration of chromophores is obtained by attaching chemically the NLO active molecules to the polymer backbone (side-chain polymers). However, it requires a significant effort in chemical synthesis. The stability of induced order depends strongly on the difference between the glass transition and the operation temperatures. The larger the difference, the less relaxation is observed. Therefore, a lot of effort is going into developing functionalized high glass transition temperature polymers such as polyimides, polyurethanes, polyvinyl carbazoles, etc.

The main-chain polymers (iii) attract a relatively weak interest. This is due, not only to the difficulties encountered with chemical synthesis, but also with the efficient chromophore orientation. More promising is the last class of functionalized polymers (iv). Here the active chromophores are used to link covalently the polymer chains through a thermal or photo-crosslinking process. However, the poling process is very delicate as the glass transition temperature is increasing with the poling time, so the time process has to be carefully monitored. Also the cross-linking may lead to shrinkage and crackling of thin films and as a consequence, to the increase of propagation losses.

A special class of functionalized materials, similar to poled polymers, represent sol–gels. Here the active chromophores either fill the pores of sol–gel or are chemically bonded to the matrix. The poling process is very similar to that applied to the thermally and photo cross-linking polymers and also has to be carefully monitored. The main advantage of these materials consists of excellent optical properties, and the disadvantage on a relatively low chromophore content, limiting the bulk $\chi^{(2)}$ susceptibility value. The representative values of second-order nonlinear optical susceptibility $\chi^{(2)}_{ijk}$ for different materials for second-order NLO optics, are listed in Table 3.

Poling efficiency

The DC poling of functionalized polymers leads to birefringent noncentrosymmetric materials with two nonzero $\chi^{(2)}$ tensor components: the diagonal:

$$\chi^{(2)}_{ZZZ}(-\omega_0;\omega_1,\omega_2) = NF\beta_{xxx}(-\omega_\sigma;\omega_1,\omega_2) \times\langle\cos^3\Theta\rangle \qquad [11]$$

and the off-diagonal:

$$\chi^{(2)}_{XXZ}(-\omega_\sigma;\omega_1,\omega_2) = \frac{1}{2}NF\beta_{xxx}(-\omega_\sigma;\omega_1,\omega_2) \times\langle\sin^2\Theta\cos\Theta\rangle \qquad [12]$$

where we assumed for simplicity, that only one nonlinear species, as is usually the case. Z, in eqns [11] and [12], is the direction of the applied external field and Θ is the angle between the electric field and the molecule dipole moment. The efficiency of poling is measured by the ratio:

$$a = \frac{\chi^{(2)}_{ZZZ}(-\omega_\sigma; \omega_1, \omega_2)}{\chi^{(2)}_{XXZ}(-\omega_\sigma; \omega_1, \omega_2)} \qquad [13]$$

It varies between 1 and ∞ (perfect alignment). For isotropic poled polymers $a = 3 \div 6$ and for nematic liquid crystal side-chain polymers, values of $a \propto 18$ were obtained.

Usually, the second-order NLO susceptibility is measured by the second-harmonic generation technique ($\omega_\sigma = 2\omega$; $\omega_1 = \omega_2 = \omega$), giving the electronic part of $\chi^{(2)}$ susceptibility, with ultrafast response time as the generated harmonic wave has to follow the fundamental one. This susceptibility is important, not only for frequency conversion through the second-harmonic generation, but also for sum and difference frequency generation as well as for optical parametric oscillation (OPO) and amplification (OPA). Another important effect described by $\chi^{(2)}(-\omega; \omega, 0)$ susceptibility is the linear electro-optic effect (Pockels effect). It finds application in electro-optic modulation for signal transmission. Traditionally, this effect is described by the electro-optic coefficient r_{ijk}, which is a third-rank three-dimensional tensor. The variation of the medium refractive index under the applied electric field is given by the following formula:

$$\Delta n_i = -\frac{2}{n_i^3} r_{ij} E_j \qquad [14]$$

The electro-optic coefficient is directly linked to the $\chi^{(2)}(-\omega; \omega, 0)$ susceptibility:

$$r_{ijk} = -\frac{2\chi^{(2)}_{ijk}(-\omega; \omega, 0)}{(n_i n_j)^2} \qquad [15]$$

The measured values of electro-optic coefficient for some selected organic single crystals and thin films are also given in Table 3.

Third-Order Nonlinear Optical Materials

There is no symmetry restriction on molecules and materials for third- (or higher-odd-) order nonlinear optical effects. These materials are used in significantly larger classes of applications than the materials for second-order NLO effects, particularly in logic systems, all optical switching, frequency conversion (e.g., frequency tripling in third-harmonic generation), light amplification, optical bistability, all optical switching, etc. In the case of organic materials, it was shown that the conjugated quasi 1D π electron systems, such as semiconducting polymers (Figure 4) exhibit an enhanced third-order NLO response.

Molecules

Similarly, as in the case of second-order NLO materials, the NLO response depends strongly on the conjugation length. In the case of conjugated quasi 1D molecules, such as polymers with a chemical structure as shown in Figure 4, the γ tensor component in the conjugation direction (largest hyperpolarizability) is strongly enhanced and described by a scaling law:

$$\gamma_{xxxx} \propto L^{\alpha_t} \qquad [16]$$

where L is the conjugation length and the exponent $\alpha_t > 1$. The scaling law, eqn [16], can also be presented as dependent on γ on a number of π conjugated double or triple bonds N:

$$\gamma_{xxxx} \propto N^{\alpha_\gamma} \qquad [17]$$

It was extensively studied for a series of conjugated oligomers polyenes, oligothiophenes, etc., and the exponential coefficients are listed in Table 4, together with the calculated values. For a larger value of double (or triple) bonds, a saturation in $\chi^{(3)}$ susceptibility is observed.

The molecular second hyperpolarizability measurements are derived from solution measurements. Some typical values, obtained by the third-harmonic generation technique (THG) are listed in Table 5. Although other NLO property determination techniques were used, the THG method gives the fast, electronic part of molecular hyperpolarizability.

Bulk Materials

Importance of Order

Similarly, as in the case of materials for second-order NLO effects for practical applications the value of $\chi^{(3)}$ susceptibility is important, which on the other hand depends on the molecular second hyperpolarizability γ. Although the noncentrosymmetry is not required in the case of third-order NLO materials, the values of the bulk $\chi^{(3)}$ susceptibility depend on the order, as it follows from eqn [7]. This is particularly clear in the case of conjugated molecules, with enhanced γ_{xxxx} component. In the case of parallel polarizations

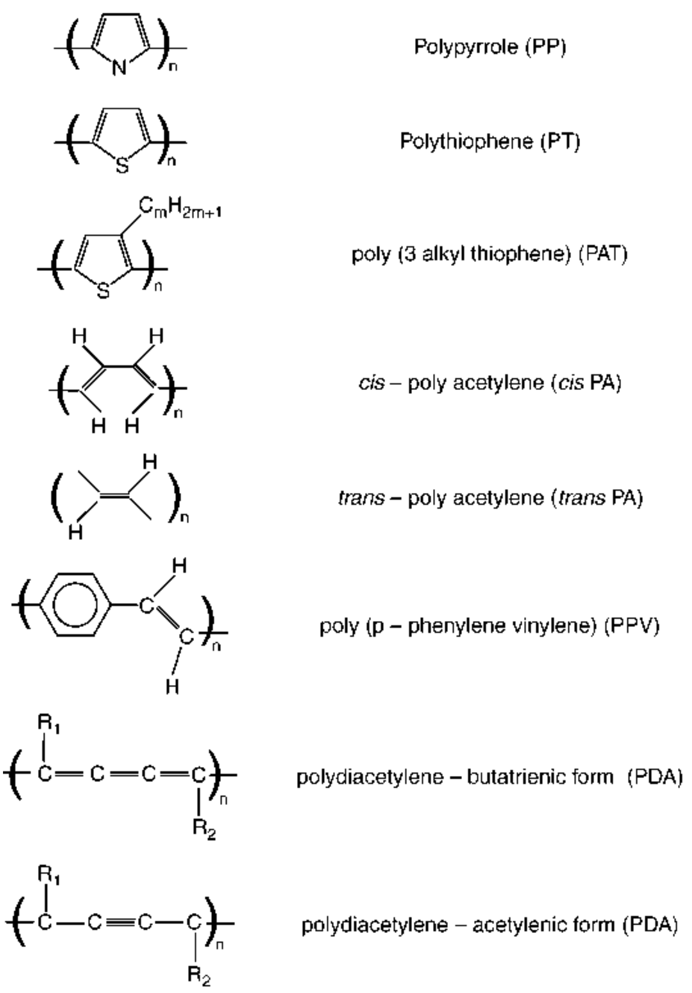
Polypyrrole (PP)
Polythiophene (PT)
poly (3 alkyl thiophene) (PAT)
cis – poly acetylene (*cis* PA)
trans – poly acetylene (*trans* PA)
poly (p – phenylene vinylene) (PPV)
polydiacetylene – butatrienic form (PDA)
polydiacetylene – acetylenic form (PDA)
N
S
C_mH_{2m+1}
H
C
R1
R2
n

Figure 4 Chemical structures of selected conjugated quasi 1D organic polymers.

Table 4 Calculated (calc) and measured (meas) values of the scaling low exponent $\gamma \propto N^{\alpha_\gamma}$ for different conjugated oligomers

Oligomers	$\alpha_{\gamma calc}$	*Remark*	$\alpha_{\gamma meas}$	*Remark*
Polyene oligomers	5	free electron model	2.82	CT polyenes $(\omega = 0)$
		INDO $(\omega = 0)$	3.5	$\lambda_\omega = 1.907$ μm
	3	SCF/STO-3G	2.3	Symmetric A-A
	3.27	SCF/3-21G		THG 1.907 μm
	3.14	$N = 2$	3.3	Asymmetric D-A
	3.67	$N = 6$, decrease		THG 1.907 μm
	3.98	For $N > 6$ DVB-PPP	4.4	Symmetric D-D
				EFISH 1.34 μm
				no saturation up to $N = 14$
	3.8	neutral solitons		
	3.85	charged polarons		
	6.32	*trans* form		
	5.4	*cis* form		
	4.7			
Thiophene oligomers	4.5	CNDO $N < 6$	4.05	DFWM in solutions
		Saturation $N > 6$	4.54	EFISH in solid solutions $N = 3, 5, 6$
			3.4	THG 1.907 μm

of $\omega_\sigma, \omega_1, \omega_2, \omega_3$ fields, the bulk susceptibility is given by:

$$\chi^{(3)}(-\omega_\sigma; \omega_1, \omega_2, \omega_3) = NF\gamma_{xxxx}(-\omega_\sigma; \omega_1, \omega_2, \omega_3) \times \langle \cos^4 \Phi \rangle \qquad [18]$$

where F is the total field factor and Φ is the angle between the polarization direction of $\omega_\sigma, \omega_1, \omega_2, \omega_3$ fields and a chosen direction in the laboratory reference frame. The value of $\langle \cos^4 \Phi \rangle$ varies between 0.2 for an isotropic distribution of polymer chains and 1 for perfectly aligned, respectively. It gives a factor of 5 difference in the macroscopic $\chi^{(3)}$ susceptibility. Some examples of third-order NLO susceptibilities for representative conjugated polymers are given in Table 6. These were obtained by third-harmonic generation technique which measures the fast, electronic contribution to $\chi^{(3)}$.

Table 5 Molecular off resonance second hyperpolarizability γ for selected organic molecules

Molecule	*$\gamma(-3\omega; \omega, \omega, \omega)$ 10^{-50} m^5/V^2*	*λ_{meas} nm*
Carbon tetrachloride	5.1	1064
	4.7	1907
Chloroform	3.8	1064
	3.2	1907
Acetone	2.5	1064
	2.6	1907
Methanol	1.16	1064
	1.21	1907
Ethanol	1.8	1064
	2.1	1907
Propanol	2.4	1064
	2.5	1907
Water	0.46	1064
Nitrobenzene	9.2	1907
Chlorobenzene	5.8	1907

Calibrated with silica value: $\chi^{(3)}(-3\omega; \omega, \omega, \omega) = 3.9 \times 10^{-22}$ m^2/V^2 at 1907 nm.

Two-Photon Absorbing Materials

Two-photon absorbing (TPA) materials are a special class of third-order NLO materials exhibiting a complex Kerr susceptibility $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$. In this case, the nonlinear refractive index, derived from eqn [1] is complex too.

$$n = n_0 + n_2 I = n_0 + (n_2^r + \kappa_2)I \qquad [19]$$

where I is the light intensity, n_2^r is the real part and κ_2 is the imaginary part (nonlinear extinction coefficient) of the nonlinear index of refraction, respectively. The nonlinear index of refraction is directly related to the Kerr susceptibility $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$:

$$n_2 = \frac{3\chi^{(3)}(-\omega; \omega, -\omega, \omega)}{4\varepsilon_0 c n_0^2} \qquad [20]$$

where c is the light velocity. If $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$ is complex (close to a resonance) then n_2 is complex too and the light beam propagating in such a medium in direction z is attenuated through the two-photon absorption process:

$$I(z) = I_0 \, \mathrm{e}^{-(\alpha_{\mathrm{L}} + \alpha_{\mathrm{TPA}} I)z} \qquad [21]$$

Table 6 Resonant and nonresonant values of cubic susceptibility $\chi^{(3)}$ for selected organic materials

Material	*Form*	*λ_{meas} nm*	*$\chi^{(3)}(-3\omega; \omega, \omega, \omega)$ 10^{-22} m^2/V^2*	*Remark*
PDA-PTS	Single crystal	1890	119 (70) × 10^3	‖ chain, 3ph
			<1400	⊥ chain
		2620	22.4 (14) × 10^3	‖ chain
			<280	⊥ chain
PDA-TCDU	Single crystal	1890	9.8 (7) × 10^3	‖ chain
			<70	⊥ chain
		2620	51.8 (20) × 10^2	‖ chain
			<40	⊥ chain
DEANS	Single crystal	1900	1.4 × 10^4	‖ (mol. axis)
			127	⊥ (mol. axis)
PDA-AFA	LB film (blue)	1350	7.8 (1.4) × 10^3	2ph
PDA-AFA	LB film (blue)	1907	11.6 (1.2) × 10^3	3ph
PA	Isotropic oriented	1907	78.4 (22) × 10^4	3ph
			2.4 × 10^6	3ph
PDA-C_4UC_4	Oriented film	1907	2.9 × 10^4	‖ chain, 3ph
			3 × 10^2	⊥ chain
C_{60}	Isotropic film	1064	11.5 (1.1) × 10^3	3ph
C_{70}	Isotropic film	1420	13 (1.4) × 10^3	3ph
PPV	Isotropic film	1450	1.96 × 10^4	3ph

Numbers in brackets are standard deviations. Data calibrated with silica: $\chi^{(3)}(-3\omega;\omega,\omega,\omega) = 3.9 \times 10^{-22}$ m^2/V^2; 2ph, two photon resonance enhancement; 3ph, three photon resonance enhancement; for chemical structures see Figure 4.

Table 7 Two photon absorption cross-sections σ_{TPA}, maximum TPA wavelength (λ_{max}), measurement wavelength (λ_{meas}), pulse duration and measurements technique for a series of organic molecules

Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
BDBAS (Bis(di-*n*-butylamino)stilbene)	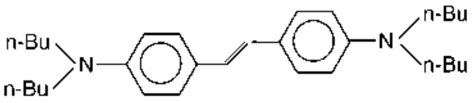	Toluene	600	310	4 ps	600	NLT
				9300	5 ns	600	
				1000	5 ns	700	
			600	210	5 ns	600	TPF
				6	5 ns	698	
				110	100 fs	620	
		Acetone		17 700	5 ns	600	NLT
	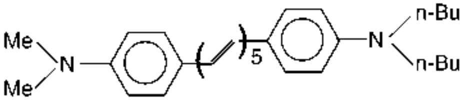	Toluene	755	10 000	5 ns	700	NLT
				17 000	5 ns	755	
				290	4 ps	700	
			730	300	5 ns	698	TPF
				790	5 ns	730	
BDPAS (Bis(diphenylamino)stilbene)	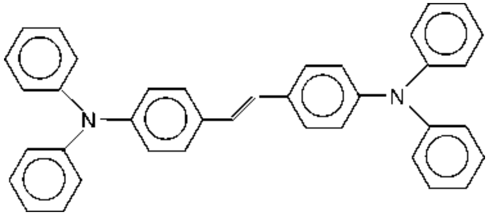	Toluene		13 000	5 ns	700	NLT
				110	4 ps	700	
Trans-stilbene	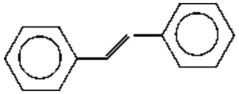	Toluene	514	12	5 ns	514	TPF
	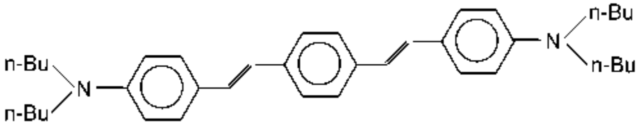	Toluene	730	995	5 ns	730	TPF
			~725	635	100 fs	725	
	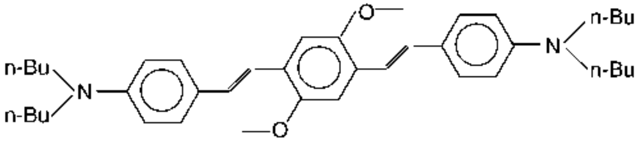	Toluene	730	900	5 ns	730	TPF
			~725	680	100 fs	725	

Table 7 (*Continued*)

Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
	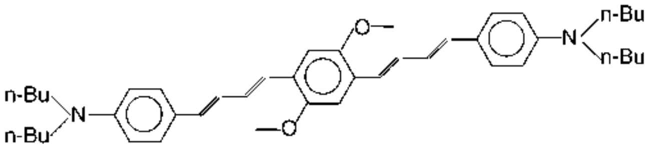	Toluene	775	1250	5 ns	775	TPF
			~750	1270	100 fs	750	
	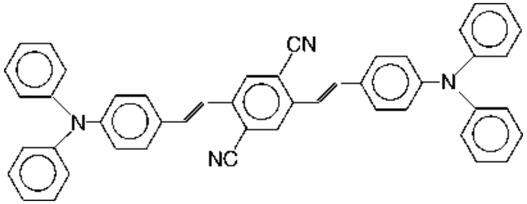	Toluene	835	1940	5 ns	835	TPF
			810	3670	100 fs	810	
	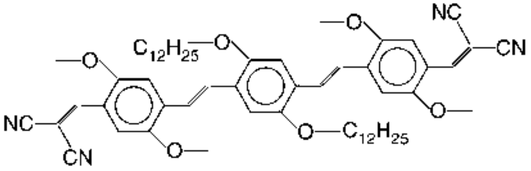	Toluene	825	480	5 ns	825	TPF
			940	620		940	
			815	650	100 fs	815	
			910	470		910	
	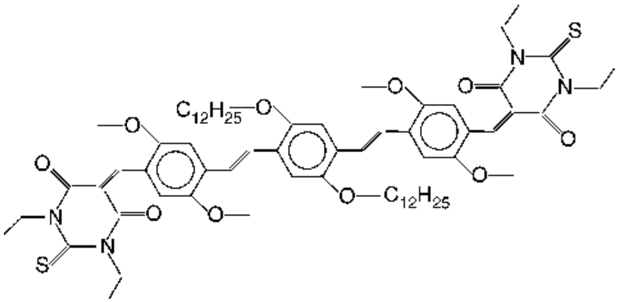	Toluene	970	1750	5 ns	970	TPF
	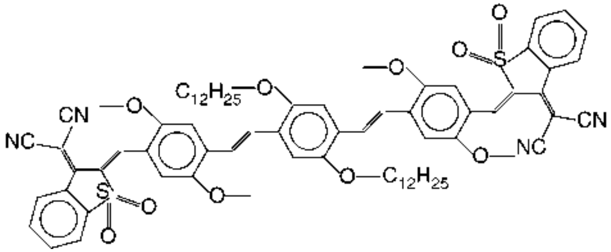	Toluene	975	4400	5 ns	975	TPF
			945	3700	100 fs	945	

	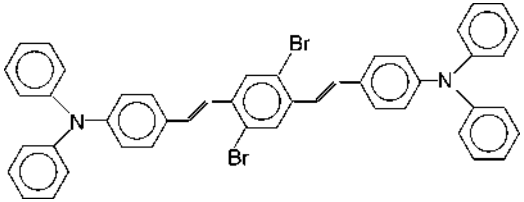	Toluene	~800	450	5 ns	800	TPF
CZ-DTT-CZ (2,6 -Bis [2-(9-ethylcarbazol-3-yl) vinyl] thiopheno [3′,2′-2,3] thiopheno [4,5-b] thiophene)	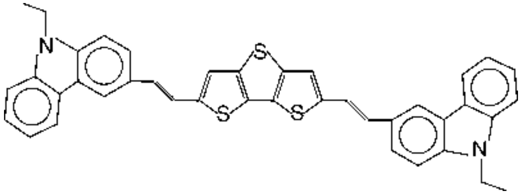	1,1,2,2-Tetra chloro-ethane		105 000	8 ns	810	NLT
CZ-DTT-OX (2-[2-(4-(5-(4-*tert*-Buthyl phenyl-1,3,4-oxadiazol-2-yl) phenyl) vinyl)-6-[2-(9-ethyl carbazol-3-yl) vinyl] thiopheno [3′,2′-2,3] thiopheno [4,5-b] thiophene)	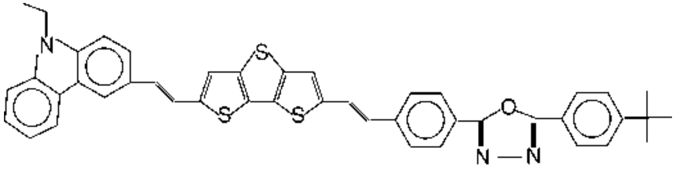	1,1,2,2-Tetra chloro-ethane		33 500	8 ns	810	NLT
TP-DTT-TP ([4-(2-(6-(2-(4-Diphenylamino) phenyl)vinyl) thiopheno [2′,3′-4,5] thiopheno [3,2-b] thiophen 2-yl-vinyl) phenyl] diphenylamine)	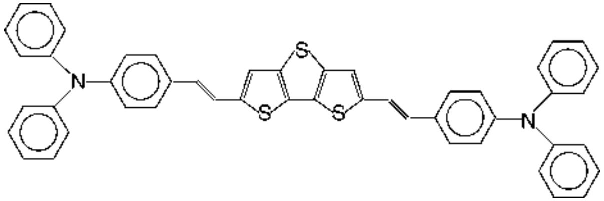	1,1,2,2-Tetra chloro-ethane		199 000	8 ns	810	NLT
TP-DTT-OX ([4-(2-(6-(2-(4-(5-(4-*tert*-Buthylphenyl-1,3,4-oxadiazol-2-yl)) phenyl) vinyl) thiopheno [2′,3′-4,5] thiopheno [3,2-b] thiophen 2-yl)-vinyl) phenyl] diphenylamine)	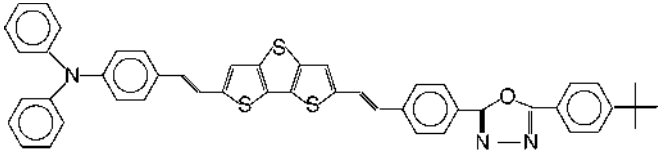	1,1,2,2-Tetra chloro-ethane		119 000	8 ns	810	NLT
Biphenyl	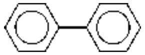			80			Theoretical calculus
Terphenyl	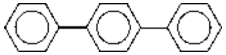			330			Theoretical calculus

Table 7 (*Continued*)

Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
Sexiphenyl	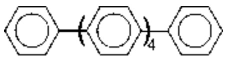			2090			Theoretical calculus
AF 50	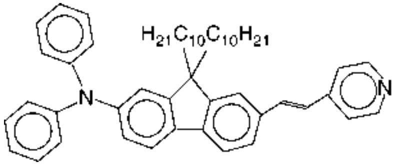	Tetrahydro-furan Benzene		11 560 19 300	8 ns	800	NLT
AF 80	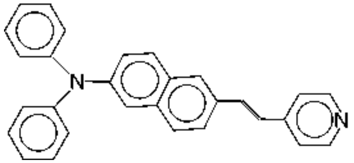	Tetrahydro-furan		6840	8 ns	800	NLT
ASPT	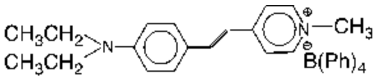	Poly hydroxyethyl metacrylate		12 000	10 ns	1064	NLT
DEANST	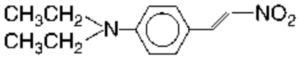	Tetrahydro-furan		200	46 ps	1064	KE
Rhodamine B	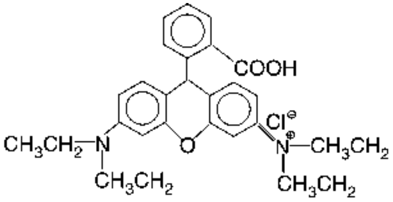	Methanol		50	100 fs	730	TPES
DR1 (4-[*N*-ethyl-*N*-(2-hydroxyethyl)] amino-4′-nitroazobenzene (dispersed red 1))	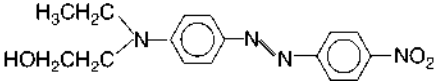	Tetrahydro-furan		102	42 ps	1064	KE

KE, optical Kerr effect; NLT, nonlinear transmission; TPF, two photon fluorescence; TPS, two photon excitation spectroscopy.
1 GM = 10^{-50} cm^4 s/photon-molecule.

Table 8 Two photon absorption coefficients for different molecules and extrapolated bulk values

Molecule	α_{TPA} *cm/GW*	λ_{max} *nm*	λ_{TPA} *nm*	*Concentration g/l*	*Solvent*	*Bulk value cm/GW*
DR 1	0.8	490	835	70	THF	11.4
DBANS	1.1	450	808	70	THF	15.7
DCNP	0.06	460	920	30	THF	2.3
DEANST	0.64	440	920	105	THF	6.1
DCM	0.1	480	950	11	DMSO	9.1
PDA-PTS	80	620	1064 not max of λ_{TPA}	Bulk	Single crystal	80
PDA-PTS	800	620	925	Bulk	Single crystal	800
Poly(3-octyl thiophene)	0.27	435 (142)	720 (94)	25	THF	9.6
Poly(3-octyl thiophene) regioregular	0.85	550 (170)	730 (70)	11.2	THF	76

λ_{max} corresponds to the maximum linear and λ_{TPA} to the maximum nonlinear absorption wavelengths, respectively, DCNP: 3-(1,1-dicyanoethenyl)-1-phenyl-4,5-dihydro-1H-pyrazole, for other molecules see Figures 4 and 5 and preceding tables.

where α_L is the linear and α_{TPA} is the two-photon absorption coefficient, respectively. The last coefficient can be explicitly expressed by the imaginary part of the nonlinear index of refraction:

$$\alpha_{TPA} = \frac{2\pi\kappa_2}{\lambda_\omega} \qquad [22]$$

where λ_ω is the wavelength of the propagating beam. On the molecular level one also defines the two-photon absorption cross-section given by

$$o_{TPA} = \frac{\hbar\omega}{N}\alpha_{TPA} \qquad [23]$$

where N is the number density of molecules. The values of TPA absorption cross-sections and two-photon absorption coefficients for selected organic molecules and solids are listed in Tables 7 and 8, respectively.

Photorefractive Materials

The photorefractive effect consists of the generation of a refractive index grating in an NLO material by its inhomogeneous illumination. This is a necessary condition but not a sufficient one. The sufficient condition consists of the possibility of creating the space charge field, which combined with the applied external field, leads to the refractive index modulation through the Pockels effect. In that case, the refractive index grating is phase mismatch with the intensity grating, leading to the transfer of photons from one beam to another, as occurs in two-beam coupling experiments, and thus applies to light amplification. Although the photorefractive effect was first observed in ferroelectric single crystals, the researchers at IBM demonstrated, at the beginning of 1990s, the photorefractive effect in organic composite materials. However, the main contribution in these materials comes from the rotational mobility of nonlinear optical (NLO) chromophores under the applied external electric field. Thus, to exhibit a photorefractive effect, the organic material has to possess the following properties:

(i) charge generation and separation capacity;
(ii) charge mobility;
(iii) linear electro-optic (Pockels) effect; and
(iv) rotational mobility.

Composite Materials

The different properties satisfying the above requirement to exhibit a photorefractive effect, are obtained by mixing together four types of molecules:

(i) photoconducting polymer;
(ii) highly second order nonlinear CT molecules;
(iii) photoensitizer for charge generation under illumination; and
(iv) plasticizer, lowering the glass transition temperature, thus increasing the rotational mobility.

Some examples of different types of molecules used are shown in Figure 5. Although these structures exhibit large exponential gain coefficients Γ and high diffraction efficiencies η, their main disadvantage consists of slow, seconds-duration response times

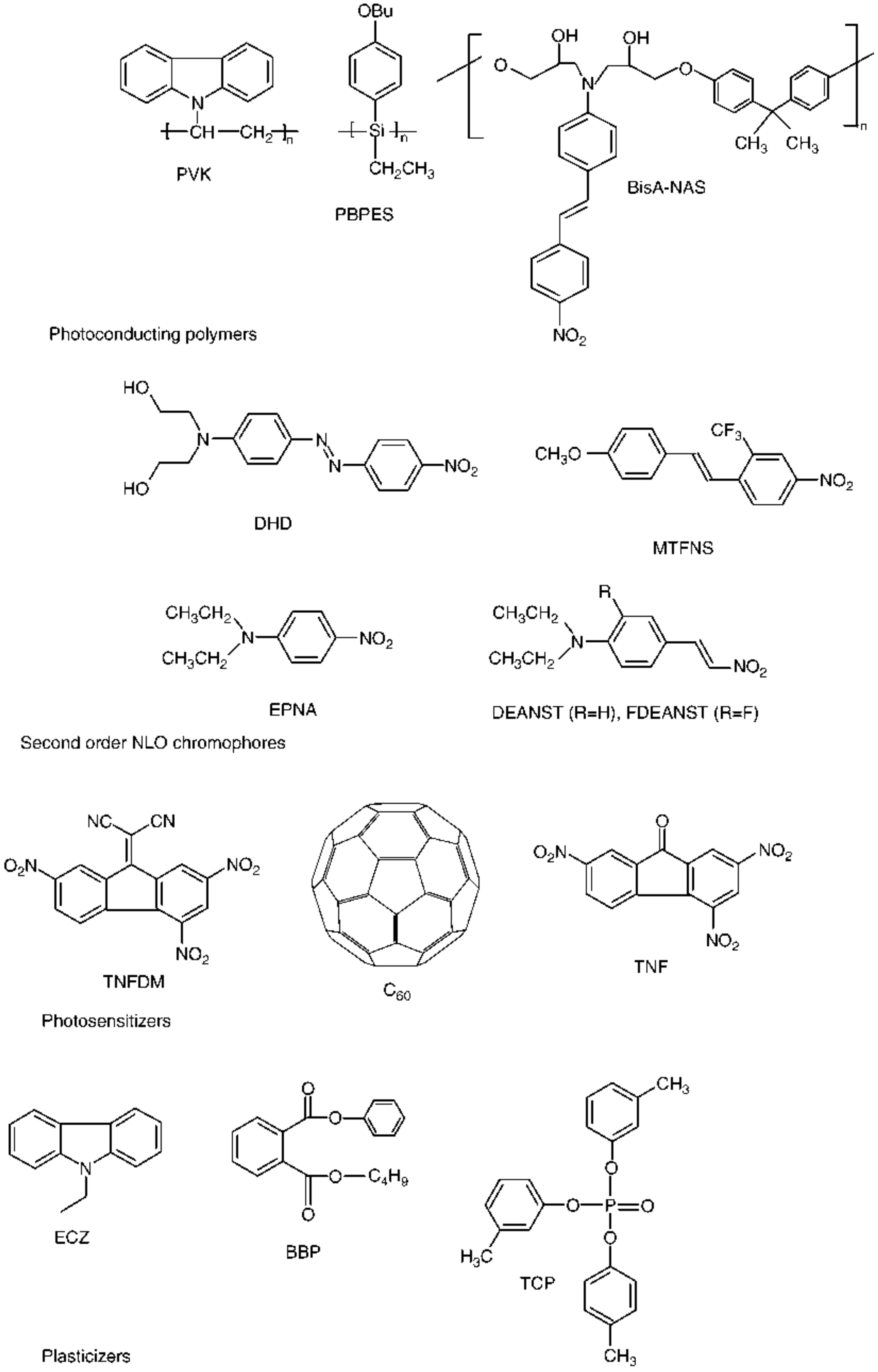
PVK
PBPES
BisA-NAS
Photoconducting polymers
DHD
MTFNS
EPNA
DEANST (R=H), FDEANST (R=F)
Second order NLO chromophores
TNFDM
C60
TNF
Photosensitizers
ECZ
BBP
TCP
Plasticizers

Figure 5 Chemical structures of photoconducting polymers, CT chromophores, photosensitizers and plasticizers commonly used for the fabrication of photorefractive polymers.

Table 9 Photorefractive properties of selected single crystals at zero external field, composite photorefractive polymers and liquid crystals

Structure	*Form*	*Thickness* μm	*Composition w%*	*Laser intensity W/cm*^{2}	*Exponential gain* Γ *cm*$^{-1}$	*Electric field V/*μm	*Response time s*	*Diffraction efficiency %*	λ_{meas} *nm*
COANP: TCNQ	Single crystal			0.08			100–1000 depending on grating period Λ		514.5
MNBA	Single crystal			0.1	2		10–100		514.5
DMNPAA/PVK/TNF/ECZ	Composite material	105	50: 33: 16: 1		220 (p-polarized) -40 (s-polarized)	90	1	86	514.5
Doped: 5 CB/C_{60}	Composite material	25	95: 5		2890	0.07	0.3		514.5
Doped: 8 OCB/5 CB /PER/NI	Composite material	37			640	0.04	0.04		514.5
SCLCP/E7/C_{60}	Composite material	50	49.975: 49.975: 0.05		600	0.7	3.5		632.8
PVA/E44/C_{60}	PDLC	25				40	On: 8 s Off: >3 days	52	
PMMA/E49/ETZ/TNF	PDLC	53	45: 33: 21: 1		41	22		8	675
POT/LC	HPCPLC	10			965	0.6	0.011	26	514.5
POT/DR1/LC	HPCPLC	10			2600	0.6	0.010	31	514.5
PVK/TNF/LC	HPCPLC	10			3700	0.9	0.010	45	514.5

LC, liquid crystals; PDLC, polymer dispersed liquid crystals; HPCPLC, hybrid photoconducting polymer–liquid crystal; COANP-2-cyclooctalamino-5 nitropyridine, TCNQ, 7,7,8,8-tetracyanoquinodimethane; MNBA4′, nitrobenzylidene-3 acetamino-4-methoxyaniline; 5-CB-4′, (*n*-octyloxy)-4-cyanobiphenyl; 8OCB-4′, (*n*-octhyloxy)-4-cyanobiphenyl; PER, perylene; NI, *N*,*N*-di(*n*-octyl)1,4,5,8-naphtalenediimide; SCLCP, side chain liquid crystalline polymer (PMMA-4-cyanophenyl benzoate); E7, E44, E49, commercial LC mixtures (Merck, Germany); other molecules in **Figure 5** and **Tables 2** and **3**.

and require large external operation electric fields (~100 V/μm).

Photorefractive Liquid Crystals

In order to improve the performances of organic photorefractive materials the use of nematic liquid crystals was proposed as active media. These materials exhibit a large rotational mobility, requiring a relatively small electric field with response time of milliseconds. The reorientation of domains requires 2 ÷ 3 orders of magnitude electric fields as compared with the composite materials (reorientation of individual molecules). Several types of structures of liquid crystal based photorefractive structures are realized, such as:

- composite materials in which the liquid crystals mesogens are NLO chromophores;
- polymer dispersed liquid crystals (PDLCs);
- hybrid photoconducting polymers – liquid crystals (HPCPLCs), in which the rotational mobility is separated from the charge mobility. Moreover the use of photoconducting polymer allows the modulation of the electric field on the liquid crystal layer, enhancing the effect.

The performances of different, representative, organic photorefractive structures are listed in Table 9.

See also

Nonlinear Optics, Basics: Photorefraction; $\chi^{(2)}$–Harmonic Generation. **Polarization:** Introduction.

Further Reading

Agranovich VM and La Rocca GC (eds) (2002) Organic Nanostructures: Science and Applications. *Proceed. International School of Physics "Enrico Fermi", Course CXLIX*. Amsterdam: IOS Press.

Bosshard Ch, Sutter K, Pretre Ph, Hulliger J, *et al.* (1995) In: Kajzar F and Garito AF (eds) *Organic Nonlinear Optical Materials*, *Advances in Nonlinear Optics*, vol. 1. Amsterdam: Gordon and Breach Publishers.

Bredas JL and Silbey R (eds) (1991) *Conjugated Polymers*. Dordrecht: Kluwer Academic Publishers.

Charra F, Agranovich V and Kajzar F (2003) *Organic Nanophotonics*. NATO Science Series 2. Dordrecht: Kluwer Academic Publishers.

Chemla DS and Zyss J (eds) (1987) *Nonlinear Optical Properties of Organic Molecules and Polymers*. Orlando, FL: Academic Press.

Graja A, Bulka B, Kajzar F (eds) (2002) *Molecular Low Dimensional-and Nano-Structured Materials for Advanced Applications*. NATO Science Series 2. Dordrecht: Kluwer Academic Publishers.

Hornac LA (ed.) (1992) *Polymers for Lightwave and Integrated Optics. Technology and Applications*. New York: Marcel Dekker, Publishers.

Kajzar F and Agranovich VM (eds) (2000) *Multiphoton and Light Driven Multielectron Processes in Organics: New Phenomena, Materials and Applications*. Dordrecht: Kluwer Academic Publishers.

Kajzar F and Reinisch R (eds) (1998) *Beam Shaping and Control with Nonlinear Optics*. NATO ASI Series, Series B: Physics, vol. 369. New York: Plenum Press.

Kajzar F and Swalen J (eds) (1996) *Thin Films for Waveguiding Nonlinear Optics*. Amsterdam: Gordon and Breach Scientific Publishers.

Kajzar F, Agranovich VM and Lee CY-C (1996) *Photoactive Organic Materials: Science and Application*. Dordrecht: Kluwer Academic Publishers.

Kuzyk M and Dirk C (eds) (1998) *Characterization Techniques and Tabulations for Organic Nonlinear Optical Materials*. New York: Marcel Dekker.

Lee KS and Wegner G (eds) (2003) *Polymers for Photonics Applications II: Nonlinear Optical, Photorefractive and Two-Photon Absorption Polymers*. Advances in Polymer Sciences 161. Berlin: Springer-Verlag.

Lessard R and Franke H (eds) (1997) *Materials Characterization and Optical Probe Techniques*. Critical Reviews of Optical Science and Technology, vol. CR69. Bellingham: SPIE.

Messier J, Kajzar F, Prasad P and Ulrich D (1989) *Nonlinear Optical Effects in Organic Polymers*. Dordrecht: Kluwer Academic Publishers.

Messier J, Kajzar F and Prasad P (1991) *Organic Molecules for Nonlinear Optics and Photonics*. Dordrecht: Kluwer Academic Publishers.

Miyata S and Sasabe H (eds) (1997) *Poled polymers and their Application to SHG and EO Devices*. Amsterdam: Gordon and Breach Scientific Publishers.

Prasad PN and Williams DJ (1991) *Introduction to Nonlinear Optical Effects in Molecules and Polymers*. New York: Wiley.

Richardson T (ed.) (1999) *Functional Polymers and Polymeric Materials*. Chichester, UK: John Wiley & Sons.

MICROSCOPY

Contents

Overview
Confocal Microscopy
Imaging Multiple Photon Fluorescence Microscopy
Interference Microscopy
Nonlinear Microscopy
Phase Contrast Microscopy

Overview

C J R Sheppard, National University of Singapore, Singapore

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The basic construction and imaging properties of the optical microscope, one of the most important optical instruments, are described. But the restrictive meaning of a microscope as a device that the user looks down to see a magnified image is far from the present terminology that includes scanning systems (where the image is stored in a computer), nonvisible radiation, such as electrons or X-rays, and a variety of different contrast mechanisms.

According to the *Oxford English Dictionary* a microscope is 'an instrument magnifying objects by means of lenses so as to reveal details invisible to the naked eye'. Thus, basically a microscope forms a magnified image of an object. The first microscopes used visible light and formed an image showing variations in the intensity of light scattered by the object. This intensity is in general related to the optical properties of the object in a complicated way.

Nowadays there are many different types of microscope. Our understanding of a microscope must therefore be generalized in the following ways.

- The microscope may not be a conventional imaging system using lenses or mirrors. For example, it may be a scanning imaging system. A confocal system is a combination of a conventional and a scanning system. Another important category of microscopes is that of probe microscopes, including near-field optical microscopes. Some probe microscopes, for example the atomic force microscope (AFM) do not even seem at first glance to rely on the direct use of radiation.
- The microscope may not use visible light, but other electromagnetic radiation, or even other forms of radiation.
- The image can be formed using a variety of different contrast mechanisms. Some of these, such as phase contrast, are designed to image particular optical properties of the sample. Others image the generation of a form of radiation when the object is stimulated by another form of radiation. In particular, in principle, virtually any form of spectroscopy can be the basis of building up an image by measuring the spatial variations in the signal.

Different Forms of Microscopes

The Conventional Microscope

In a simple microscope, an objective lens forms a real, magnified, and inverted image of an object. In a compound microscope, an eyepiece is added, forming a virtual image that is viewed by eye to give real image on the retina. The eye can be replaced by CCD camera to record moving or still images. In early microscopes, the objective lens and eyepiece are mounted in the two ends of a brass tube. The objective screws into the tube using an RMS (Royal Microscopical Society) thread (0.8$''$ Whitworth 36 threads per inch), until its mounting face is flush with the end of the tube. The length of the tube is the mechanical tube length. The distance from the mounting face of the objective to the plane of the real image is the optical tube length, which varied according to different manufacturers in the range 160–210 mm. Now most objectives have infinity tube length, so that they collimate the light from the object. The collimated light is brought to a focus by an additional tube lens. An objective of correct tube length should always be used as otherwise spherical aberration is introduced. As the objective, tube lens, and eyepiece are designed together as a system, great care should be exercised when mixing components from different manufacturers.

Abbe showed that in order to form a perfect lateral image the objective must satisfy a different condition called the Herschel condition. Otherwise the image of an off-axis point will suffer from coma. An aberration-free system satisfying the sine condition is called an aplanatic system. A microscope produces a three-dimensional (3D) image of a 3D object. A point of the object closer to the lens appears further from the lens in the image. If the lateral magnification is M, the longitudinal magnification is approximately equal to M^2. However, it can be shown that it is impossible to devise an instrument that can produce a perfect 3D image of a 3D object, because to have a perfect transverse image the system must satisfy Abbe's sine condition, but to have a perfect axial image the system must satisfy the Herschel condition. These two conditions are not mutually compatible. Actually, we find that the longitudinal magnification is not exactly equal to M^2, and is not even constant in space. Different depths of the object are imaged differently: the microscope objective only produces a good image for one plane of focus, and other planes will exhibit spherical aberration. However, all of these problems can be overcome by bringing the part of the object under observation into focus, without refocusing the eye or other detector. A 3D image can thus be generated by scanning the object stage in the axial direction, and recording a stack of images from different depths in a computer. For modern objectives of infinite tube length, focusing can be alternatively achieved by piezoelectric scanning of the objective lens.

Resolution

For an object consisting of a single bright point object in a dark background, the image produced by a perfect microscope, according to paraxial diffraction theory, is the so-called Airy disk, consisting of a bright central spot, surrounded by a series of rings. The intensity in a meridional cross-section through the Airy disk is the same for any aperture of objective, or wavelength of light, but its width varies. The intensity of the first bright ring is 1.75% of that at the peak. The radius of the first dark ring of the Airy disk is

$$r_0 = \frac{0.61\lambda}{n \sin \alpha} \qquad [1]$$

where λ is the wavelength, n is the refractive index of the immersion medium, and α is the angle subtended at the edge of the objective aperture. The radius depends on $n \sin \alpha$, which is called the numerical aperture and should be as large as possible for good resolution. Thus, high-magnification lenses often use an immersion fluid, usually oil with a refractive index of 1.518 (ISO 8036/1).

The optical coordinate v is defined as

$$v = \frac{2\pi}{\lambda} rn \sin \alpha \qquad [2]$$

so that the dark ring of the Airy disk occurs at a value of v of $2\pi \times 0.61 = 3.83$. Resolution is sometimes expressed in terms of the full width at half maximum (FWHM) of the image of a point object, which is 3.232 in optical coordinates. A full nonparaxial theory does not give a very different figure.

Resolution of a microscope is often specified by two-point resolution, which describes whether the image of two points can be distinguished from that of a single point. According to the Rayleigh criterion of two-point resolution, two points are just resolved if the second point is placed on the first dark ring of the first. The separation is then $0.61\lambda/(n \sin \alpha)$. A cross-section through the image in an incoherent optical system of two bright point objects of equal strength and different separations is shown in Figure 1. The intensity in the image between the points decreases as the separation is increased. The ratio of the intensity midway between the points to that at the points for the Rayleigh criterion to be satisfied is then found to be 0.735.

The concept of resolution should be distinguished from sensitivity or precision. An object much smaller than the resolution limit can still be detected, perhaps weakly, in a microscope: this detection of weak contrast depends on the sensitivity of the microscope rather than its resolution. The size of an object can also be measured with a precision much greater than the resolution of the microscope.

Illumination

If the object is not self-luminous, in order to be imaged in a microscope it must be illuminated. A semitransparent object, such as a biological slice, is illuminated in transmission. For observation of bulk objects, or surfaces, we use illumination in the reflection geometry, called epi-illumination. Either a tungsten halogen lamp or an arc source is usually used. In critical illumination, the source is focused on to the object by a condenser lens. The disadvantage of this approach is that variations in emission of the source are imaged directly into the image.

Cheaper microscopes avoid this problem by using a diffuser. Better microscopes use instead the Köhler illumination system, where the source is placed in the front focal plane of the condenser lens. Figure 2 shows a Köhler illumination system in which the source is imaged into the front focal plane of the condenser lens by another, projector, lens. The system incorporates a separate field stop and aperture stop. In practice, the size of the field stop should be reduced to illuminate as small a region of the object as necessary in order to minimize stray light. An important part of setting up a microscope to operate properly consists in centering the aperture stop of the condenser. This is often done using a so-called Bertrand lens that allows imaging of the aperture stop.

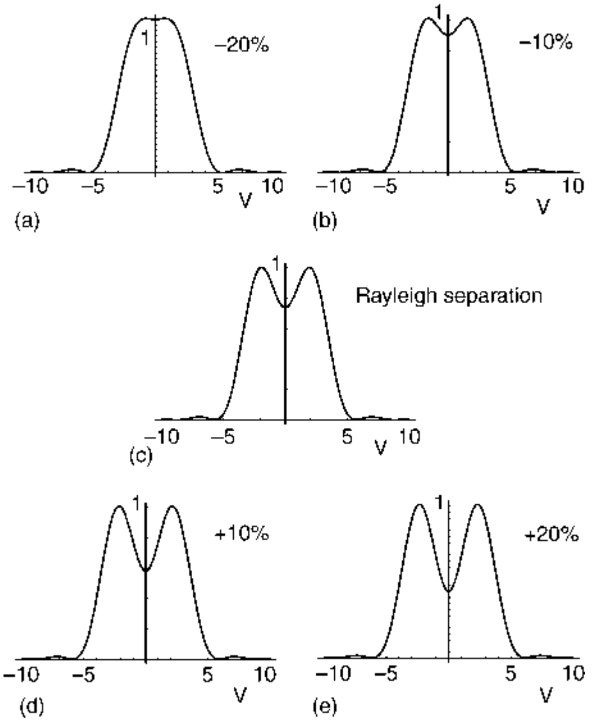

Figure 1 A cross-section through the image of two equal point objects in an incoherent microscope. The separation of the points is that for the Rayleigh criterion to be satisfied, and for changes of ±10, 20% of the Rayleigh separation.

Image formation

The Rayleigh criterion was originally specified for an incoherent optical system. It is thus applicable in fluorescence microscopy, as fluorescence is an incoherent process as two fluorescent point objects fluoresce independently. But for the image of a trans-illuminated object, the resolution depends on the coherence of the illumination. This is controlled by the numerical aperture of the condenser lens as compared with that of the objective. The ratio of these numerical apertures is called the coherence parameter, S, so that

$$S = \frac{n_c \sin \alpha_c}{n \sin \alpha} \qquad [3]$$

For $S = 0$ the illumination is purely coherent, while for $S \to \infty$ imaging becomes purely incoherent, thus corresponding to the Rayleigh criterion, so S should really be called an incoherence parameter. An important in-between case is when $S = 1$, when the apertures of objective and condenser are equal. This is termed matched, full, or complete illumination. According to the generalized Rayleigh criterion, the points are just resolved when the ratio of the intensity midway between the points to that at the points is 0.735, which is called the generalized Rayleigh criterion. We require that the distance between the points is as small as possible when this occurs. The generalized Rayleigh separation is the same for incoherent illumination and for $S = 1$, and corresponds to a distance of $2v_0 = 3.83$ in optical coordinates. Note, however, that for other separations of the points, the ratio is different for these two cases: it is not true, but often erroneously stated, that $S = 1$ corresponds to incoherent imaging. Resolution improves as the aperture of the condenser is increased ($2v_0 = 5.15$ in optical coordinates for coherent illumination) reaching a maximum when $S = 1.46$, when resolution is 9% better than for incoherent imaging ($2v_0 = 3.58$). This maximum in resolution is achieved when the numerical aperture of the condenser is larger than that of the objective, which is not

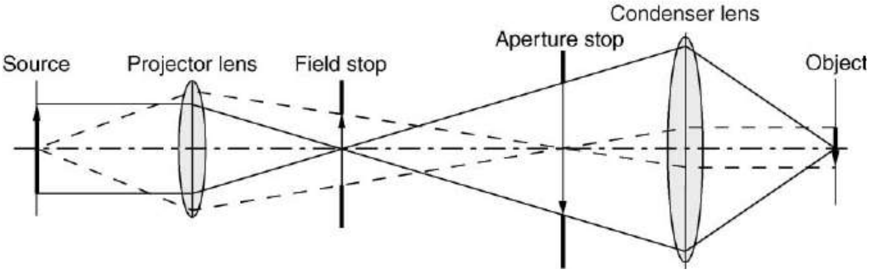

Figure 2 Geometry of the Köhler illumination system. In practice an arrangement with a smaller number of lenses is usually used.

physically achievable if the objective has the highest possible numerical aperture. Similarly, it is impossible to achieve incoherent illumination with very large objective numerical apertures.

The effect of coherence on the resolution of an image depends on the form of the object, and on the particular resolution criterion employed. In practice, contrast also depends on the aperture of the condenser. Usually, opening the condenser improves resolution, but reduces contrast, so there will be an optimum condenser aperture size.

Abbe theory of microscope imaging

Abbe argued that for coherent illumination of a grating, the grating spectra can be observed in the back focal plane of the objective. The tube lens then forms an image from the grating components. Thus, this is an early description of Fourier optics. The strength of the grating components can be altered in the back focal plane to give various optical effects, such as phase contrast. Abbe's theory was not properly appreciated at the time because it was known that in practice, resolution could be improved by using a larger condenser aperture, which does not give coherent illumination. There was much controversy about the merits of Köhler versus critical illumination, and different sizes of condenser aperture.

We now know that Köhler and critical illumination are in principle equivalent. The aberrations of the condenser lens are not important, so that the source and condenser together behave simply as a partially coherent effective source. Imaging can still be described by Fourier optics within the framework of partially coherent imaging theory.

Although Köhler and critical illumination are equivalent in principle, Köhler illumination is in practice better because of its improved illumination uniformity.

Depth of focus

The longitudinal image of a point object is found to be approximately invariant when expressed in terms of an optical coordinate u, defined as

$$u = \frac{8\pi}{\lambda} zn \sin^2 \frac{\alpha}{2} \qquad [4]$$

The axial resolution, and thus also the depth of focus, varies strongly with aperture. The FWHM for longitudinal imaging is 11.13 in optical coordinates.

Microscope objectives

Because microscope objectives often have large apertures, aberrations will be very strong, unless the objective is designed correctly. The basic design principle for a high aperture lens is the aplanatic front element (Figure 3). It is found that all rays from a point A in a sphere of refractive index n, where A is a distance r/n from its center, appear to come from a point B, distance rn, without any aberration. This principle is used by employing a front element that has one surface of the same radius of curvature as the sphere. The front surface is a sphere centered on A (Figure 3b), so that the rays are not deviated on crossing it.

A low-power achromat uses just two components with different optical dispersion to cancel chromatic aberration and correct for other aberrations (an achromatic doublet). For higher powers, two separated doublets are necessary. For the highest apertures, an aplanatic front is used. However, this can never converge the light, but only makes it less divergent. It is thus used together with an achromatic doublet. For oil immersion, the interface between the oil and the front element is not very important, so it is made planar to simplify manufacture. In practice, it is

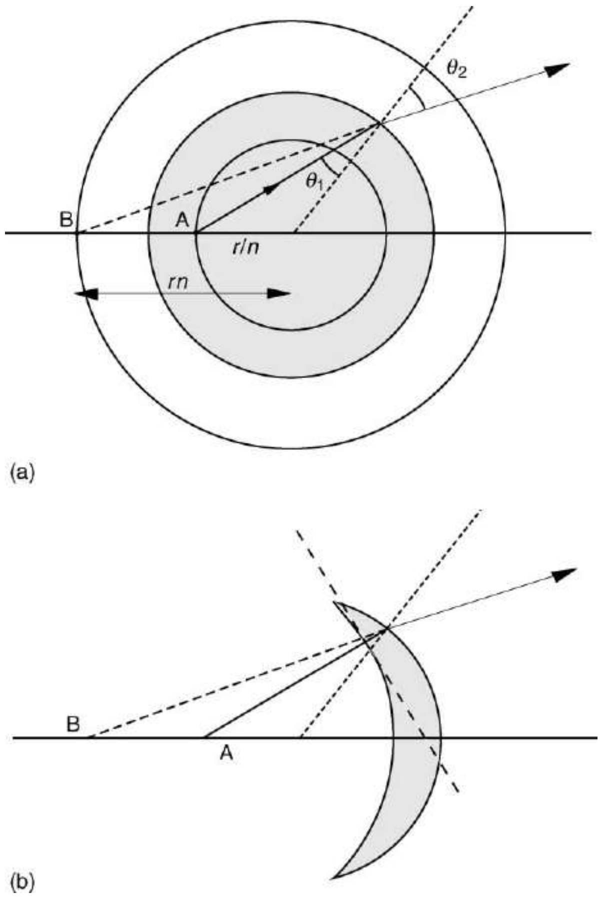

Figure 3 The principle of the aplanatic front: (a) an aplanatic surface; (b) a meniscus lens that acts as an aplanatic system.

usually made planar even for a dry lens, and the resulting aberrations cancelled out elsewhere in the objective. Apochromats, corrected for three colors, use more elements, which may include two or more stages of aplanatic front. The final objective thus consists of a number of elements that must be accurately aligned relative to one another. The objective is adjusted by the manufacturer to give a good star image (image of a point object). First a spacer is selected to optimize spherical aberration. Then a sleeve is adjusted to make the objective parfocal with others. Finally a screw is used to center the assembly to remove coma.

Conventional and Scanning Microscopes

Conventional microscope

Various different types of transmission microscope are illustrated in Figure 4. We consider image formation in a conventional microscope as illustrated schematically in Figure 4a, showing a microscope with critical illumination. A large area incoherent source is focused by the condenser lens on to the specimen, illuminating a comparatively large area of the specimen, corresponding to the whole field of view of the objective. Information from each illuminated point in the specimen is simultaneously transmitted by the objective lens to form the primary image. The objective is responsible for forming the image, with the condenser playing only a secondary role in determining the resolution of the system, through control of the coherence of the illumination. Figure 4b shows a conventional microscope in which the image is measured point by point by a detector. In practice this could be achieved by using a CCD detector.

Scanning microscope

An image can be generated by a scanning system, as illustrated in Figure 4c. A probe of light is formed by demagnification of a source, and is scanned over the object in a raster. The transmitted (or reflected) light is detected by a photodetector and thus builds up an image. The size of the probe of light limits the resolution of the system, that is, the smallest detail in the object that can be seen in the image. It should be noted that the magnification of the image is given simply by the ratio of the distance scanned in the image to the distance scanned by the probe. It is thus unrelated to the demagnification of the source. The image magnification can be altered without changing the lens and is not in any way related to the resolution.

It has been shown that the imaging properties of scanning and conventional microscopes are identical under analogous conditions, this being known as the principle of equivalence. This property is based on the principle of reciprocity, which is a very general physical law that holds even with diffraction,

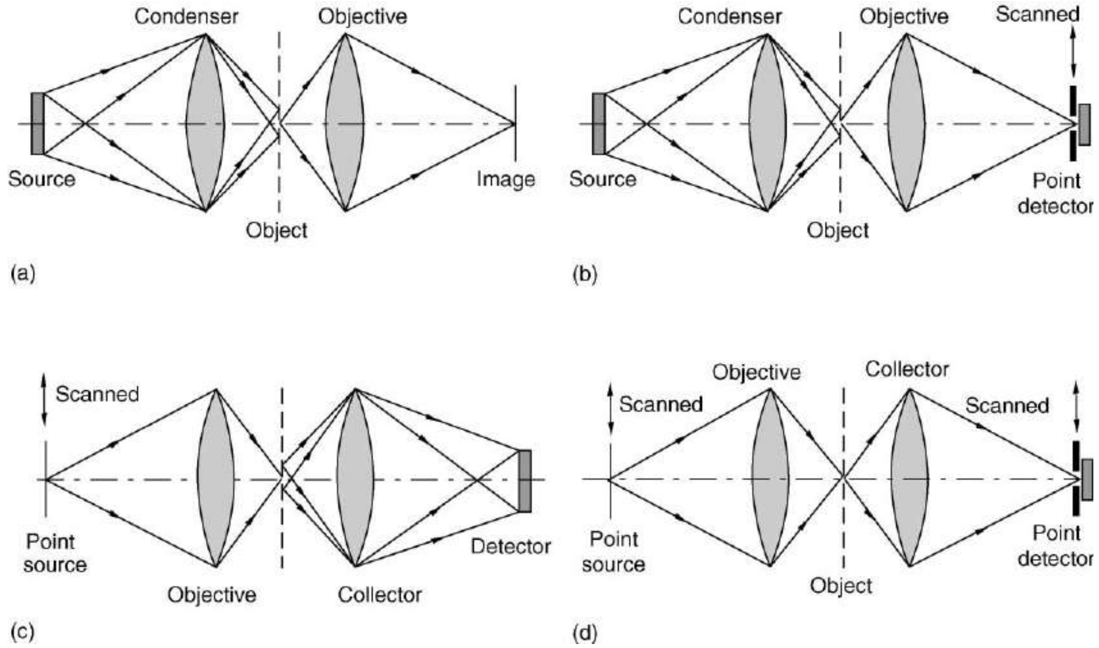

Figure 4 Different geometries of transmission microscope: (a) a conventional microscope; (b) a conventional microscope with a point detector; (c) a scanning microscope; (d) a confocal microscope.

absorption, multiple scattering, and stray light. The principle of equivalence is generally valid except if there are nonreciprocal magnetic or polarization effects, or energy losses involved (such as in fluorescence microscopy, and with inelastic scattering in electron microscopy).

Scanning systems exhibit a number of important advantages over conventional systems. Broadly, these are based on two classes of property. First, in a scanning system the image is in the form of an electronic signal. As a result, this is advantageous for quantitative measurements, as well as for image processing, including image enhancement, image restoration, and image analysis. Second, in a scanning system the object is illuminated by a focused spot, which extends the range of imaging modes available.

Confocal microscope

Finally in Figure 4d, we combine the arrangements of Figures 4b and c to give a confocal scanning optical microscope, in which a point source illuminates just a small region of the object, and a confocal point detector detects light from this illuminated region. If the point source and detector are scanned in unison, a two-dimensional image is generated. However, this system now behaves very differently from the previous ones. The confocal microscope is thus not a special case of the general partially coherent conventional imaging system. From the symmetry of Figure 4d, it is clear that the two lenses play an equal part in the imaging process. This results in an improvement in resolution. In fact, the confocal system behaves as a coherent imaging system, but with a sharper effective point spread function than in a conventional coherent microscope.

Although Figure 4 is drawn for the transmission geometry, in practice most confocal systems operate in the reflection or epi-illumination mode, in which the same objective lens is used both for illumination and detection. For a specimen placed in the focal plane, the properties of confocal transmission and reflection systems are identical. However, once the object is moved from the focal plane, certain differences arise: in particular in confocal reflection a strong optical sectioning effect occurs that allows a single section through a thick object to be imaged. This is the major advantage of the confocal microscope arrangement. Figure 4 also applies equally well to fluorescence imaging. However, in this case, because of the incoherent nature of fluorescence emissions after excitation by coherent light, imaging is then incoherent.

Confocal microscopes can be achieved by using either a point detector, in practice performed by placing a pinhole in front of a photodetector, or by using a coherent detector. Such a coherent detector, sensitive to the amplitude of the incident radiation, is not directly available for light, but is available for acoustic radiation, as in a scanning acoustic microscope. For light, a confocal effect can also be achieved by using an interferometric method to synthesize a coherent detector. Thus, in interference microscopes using the high aperture condenser and objective lenses, an optical sectioning effect arises similar to that in the confocal microscope. Another way of producing a coherent detector is to use a single-mode optical fiber.

Probe Microscopes

Scanning tunneling microscope

The scanning tunneling microscope was the first of the family of probe microscopes, relying on a fundamentally different principle from the usual forms of microscope. A physical tip in the nanometer range is brought close, also in the nanometer range, to a conducting sample. If an electric potential is applied, electrons can tunnel across between the sample and the tip. By scanning the tip mechanically across the sample, an image can be generated with a resolution in the subnanometer range. The surface topography can also be measured with subnanometer sensitivity. While interference microscopy can measure surface topography with subnanometer sensitivity, in this case the profile is averaged over the lateral resolution of the microscope, of the order of the wavelength in dimensions. The scanning tunneling microscope can image individual atoms. The signal is related to the work function of the material. Altering the bias voltage allows the band structure of the material to be investigated.

Atomic force microscope

In the atomic force microscope, the force between the surface and the tip is measured using a cantilever beam and used to build up an image. Atomic force microscopy can be performed in either a contact or noncontact mode. It can be used with insulating specimens. Optical methods, often using a position-sensitive detector, are used to measure displacements of the cantilever.

Near-field scanning optical microscope (NSOM)

In a near-field scanning optical microscope, a very small tip or aperture is scanned relative to the specimen to attain resolution greater than the classical limit set by the wavelength of the radiation. As Figure 5 shows, various different designs of near-field microscope have been proposed. It is evident that

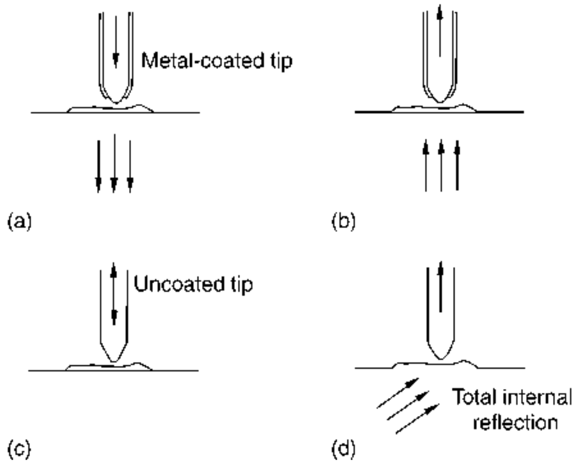

Figure 5 Different forms of near-field microscope: (a) illumination-mode NSOM; (b) collection-mode NSOM; (c) a near-field probe used for both illumination and collection in a confocal arrangement; (d) photon tunneling microscope (PTM). Reproduced with permission from Sheppard CJR (1978) The scanning optical microscope. *The Physics Teacher* 16: 648–651. © 1978 American Association of Physics Teachers.

these again form into the conventional/scanning categories described earlier. The sample can be illuminated using the near-field probe, as in Figure 5a. This is called the illumination-mode NSOM. Or the sample can be uniformly illuminated, and a signal detected using a near-field probe as in Figure 5b, giving a collection-mode NSOM. Or a confocal arrangement can be used, with a near-field probe used for both illumination and collection (c). In this case, an uncoated tip can be used; as otherwise the detected signal will be very small. Finally, Figure 5d shows the photon tunneling microscope (PTM), in which the sample is illuminated with evanescent waves produced by total internal reflection and an uncoated tip used to probe the evanescent field in the presence of the sample.

Different Types of Radiation

Electromagnetic Radiation

UV radiation and X-rays

The first microscopes used visible light, but, in fact, we do not need to use visible light, but can use any electromagnetic radiation, over a broad range of different available wavelengths. In Figure 6 we show the wavelength of the electromagnetic spectrum, illustrating how the wavelength in air varies with frequency. Instead of using visible light, shorter wavelengths allow greater resolution to be achieved. Ultra-violet light is commonly used to excite fluorescence. X-rays have also successfully been used to image biological samples.

IR and microwave radiation

In the longer wavelength region, infra-red radiation is used for observation of semiconductors or for molecular spectroscopic imaging. Imaging of emitted midinfra-red radiation can show up variations in temperature, such as hot spots in semiconductors. Microwaves have also been used, but resolution is limited by the long wavelength. For this reason microwaves (and mid-IR) have also been used for near-field microscopy.

Other Radiation

Electron microscope

The other important class of radiation that can be used for illumination is matter waves. The most common of these, of course, is electrons, the wavelength of which, as a function of acceleration voltage, is shown in Figure 7. The bend in the curve at very high voltages is caused by relativistic effects. The wavelength of the electrons is very small: even for 100 V electrons, the wavelength is only 0.1 nm. But aberrations caused by the electron lenses limit

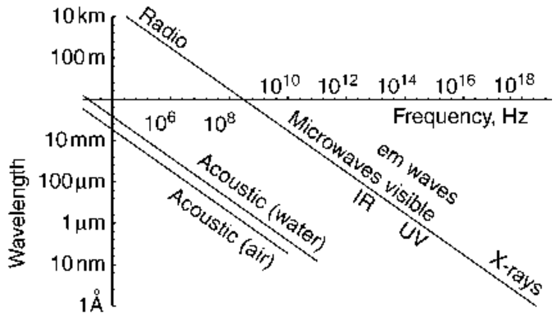

Figure 6 The wavelength of different forms of electromagnetic radiation and acoustic radiation. Reproduced with permission from Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*, pp. 1–18. New York: Wiley-Liss.

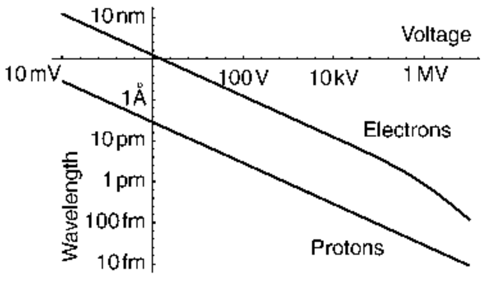

Figure 7 The wavelength of matter waves (electrons and protons) as a function of accelerating voltage. Reproduced with permission from Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*, pp. 1–18. New York: Wiley-Liss.

resolution so that at present atomic resolution is only just achievable, even with much higher voltages.

In a transmission electron microscope (TEM), a beam of electrons is produced by an electron gun. The specimen is illuminated via a condenser lens (or lenses). Usually these are magnetic lenses in which a current in a coil produces a magnetic field that focuses the electrons. The electrons that are transmitted through the specimen are then focused by a series of lenses to form a final image that can be viewed directly on a phosphor screen. An X-ray detector can be used to characterize the elements in the specimen. An electron-energy-loss spectrometer, together with a detector, can measure the energy of the electrons in a particular region of the image.

Other matter radiation

Instead of electrons, other particle beams can be used for illumination. Figure 7 shows the wavelength of a beam of protons, as function of the accelerating voltage. Because of the larger mass of protons, relativistic effects are not apparent for protons in the curve in Figure 7. Microscope images have also been formed using ions, and even neutral particles such as neutrons. Neutral atom microscopes are also under development.

Acoustic radiation

Instead of electromagnetic waves, acoustic radiation can be used to form a scanning acoustic microscope (SAM). Ultrasonic waves are focused into a water immersion medium. Mechanical scanning is again used to build up an image. Figure 6 shows the wavelength of the acoustic wave as a function of frequency. By using microwave frequencies, wavelengths, and therefore resolution, in the submicron range, can be generated. The SAM acts as a confocal microscope because the detector is sensitive to the amplitude of the acoustic wave. The acoustic microscope provides information on variations in the elastic properties (viscosity and elasticity) of the sample.

Different Contrast Mechanisms

The bright field microscope detects the intensity of the light transmitted through the object. A microscope with a condenser aperture of appreciable size behaves as a partially coherent system, and the image will, in general, depend on the phase of the light coming from the object as well as its intensity. Many objects are weakly scattering, so the contrast of this phase information is weak. In this case, the microscope effectively images variations in the transmittance or reflectance of the sample. Different designs of microscope can be used to image various properties of the object, thus providing new and complementary information. Phase contrast microscopy provides but one example of a different contrast mechanism. In fact, in general, any physical or chemical interaction of the illuminating radiation with the sample can be used as the basis for a contrast mechanism. These different contrast mechanisms can be performed in either a conventional or scanning arrangement, or in a confocal system.

Phase Contrast Microscopy

Phase contrast is a widely used technique in both biological microscopy (in transmission) and materials microscopy (in reflection). Its importance in biological microscopy stems, firstly, from the fact that contrast may otherwise be too weak to be visible, but secondly because the phase variations are caused by changes in optical thickness, related to physical density. In reflection microscopy, phase is related to surface height variations, but also depends on the optical material properties of the sample.

Polarization Microscopy

The polarization state of light transmitted through or reflected from a sample can also be changed. This change is detected in polarization microscopy, for example by placing the sample between crossed polarizers. The birefringence detected can be either material birefringence, or form birefringence (caused by the shape of the microstructure). Polarization microscopy is widely used in biological microscopy, and in mineralogy.

Fluorescence Microscopy

In fluorescence microscopy, electrons excited by illumination decay to the ground state and emit photons of light. Autofluorescence is natural fluorescence of the sample. Fluorescent dyes can be used as labels for particular biological or chemical constituents. Fluorescence microscopy is usually performed in the epi-illumination geometry to assist rejection of light from the source. Often it is performed in a confocal microscope, allowing 3D localization of the fluorescent labels.

Raman Microscopy

Raman spectroscopy can be performed on a microscopic scale to produce images of presence of particular molecular bonds. It is usually performed in a scanning system, often of the confocal type.

This is one example of the wider class of spectroscopic microscopies: in principle, any form of spectroscopy can be used as the basis for a microscope contrast mode.

Nonlinear Microscopy

Another class of contrast mechanisms is based on nonlinear optical interactions. Again, any nonlinear mechanism can be used as the basis for a contrast mechanism. These include two- (or multiphoton) fluorescence, second- (or higher-order) harmonic generation, sum or difference frequency generation, and coherent anti-Stokes Raman scattering (CARS). Probably other nonlinear optical mechanisms will be exploited in microscopy in the future.

Other Contrast Mechanisms

The contrast mechanisms described above, all rely on illumination and detection of light. But in many other modes, radiation of one form can be converted into another. These include cathodoluminescence microscopy (electrons to light), X-ray microanalysis (electrons to X-rays), photo-emission microscopy (light, usually UV, to electrons), photo-acoustic microscopy (light to ultrasonics), photothermal microscopy (light to thermal waves), optical or electron beam induced current (OBIC/EBIC). Again, in many of these, the illuminating or emitted radiation can be focused, or a combination can be used in a confocal arrangement.

See also

Microscopy: Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Interference Microscopy; Nonlinear Microscopy; Phase Contrast Microscopy.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Bradbury S and Bracegirdle B (1997) *Introduction to the Optical Microscope*. Oxford: Bios Scientific Publishers.

Inoue S and Spring KR (1997) *Video Microscopy: The Fundamentals*, 2nd edn. New York: Plenum.

Martin LC (1966) *Theory of the Microscope*. London: Blackie.

Paesler MA and Moyer P (1996) *Near-Field Optics: Theory, Instrumentation, and Applications*. New York: Wiley InterScience.

Pluta M (1988) *Advanced Light Microscopy, Volume 1: Principles and Basic Properties*. Amsterdam: Elsevier.

Sheppard CJR (1987) Scanning optical microscopy. In: Barer R and Cosslett VE (eds) *Advances in Optical and Electron Microscopy*, vol. 10, pp. 1–98. London: Academic Press.

Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances*, pp. 1–18. New York: Wiley-Liss.

Slayter EM and Slayter HS (1992) *Light and Electron Microscopy*. Cambridge: Cambridge University Press.

Wiesendanger R (1994) *Scanning Probe Microscopy and Spectroscopy*. Cambridge: Cambridge University Press.

Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

Confocal Microscopy

T Wilson, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is probably fair to say that the development and wide commercial availability of the confocal microscope has been one of the most significant advances in light microscopy in the recent past. The main reason for the popularity of these instruments derives from their ability to permit the structure of thick specimens of biological tissue to be investigated in three dimensions. It achieves this important goal by resorting to a scanning approach together with a novel (confocal) optical system.

The traditional wide-field conventional microscope is a parallel processing system which images the entire object field simultaneously, which is a severe requirement for the optical components. We can relax this requirement if we no longer try to image the whole object at once. The limit of this relaxation is to require a good image of only one object point at a time. The price that we have to pay is that we must scan in order to build up an image of the entire field.

A typical arrangement of a scanning confocal optical microscope is shown in Figure 1 where the system is built around a host conventional microscope. The essential components are some form of mechanism for scanning the light beam (usually from a laser) relative to the specimen and appropriate photodetectors to collect the reflected or transmitted light.

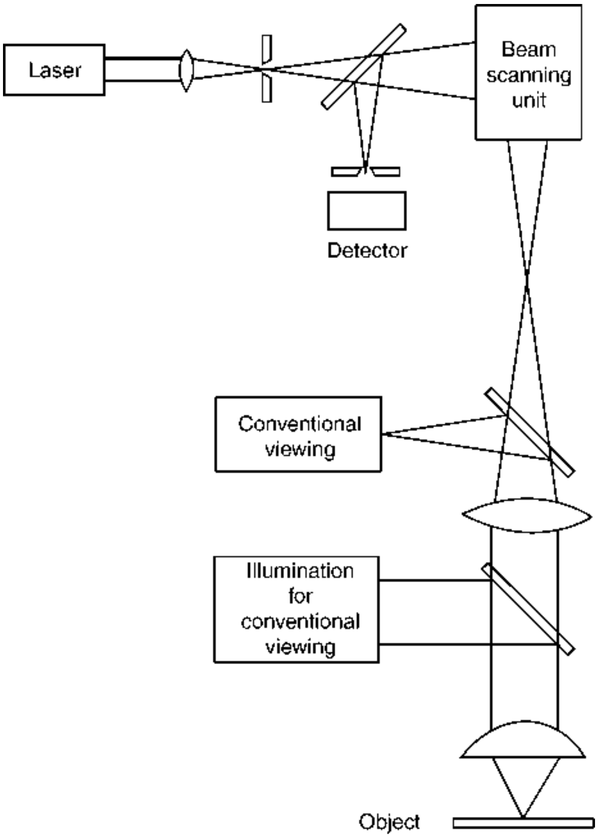

Figure 1 Schematic diagram of a confocal microscope.

Most of the early systems were analog in nature, but it is now universal, due to the serial nature of the image formation, to use a computer both to drive the microscope and to collect, process, and display the image.

In the beam scanning confocal configuration of Figure 1, the scanning is typically achieved by using vibrating galvanometer-type mirrors or acousto-optic beam deflectors. The use of the latter gives the possibility of TV-rate scanning, whereas vibrating mirrors are often relatively slow when imaging an extended region of the specimen, although significantly higher scanning speeds are achievable over smaller scan regions. We should note that other approaches to scanning may be implemented, such as specimen-scanning and lens-scanning. These methods, although not generally available commercially, do have advantages in certain specialized applications. Since this article is necessarily limited in length much additional material may be found in other sources listed in the Further Reading at the end of this article.

Image Formation in Scanning Microscopes

We will not discuss the fine detail of the optical properties of confocal systems. However, the essence is shown in Figure 1 where we see that the confocal optical system consists simply of a point source of light which is then used to probe a single point on the specimen. The strength of the reflected or fluorescence radiation from the single object point is then measured via a point, pinhole detector. The confocal – point source and point detector – optical system therefore merely produces an 'image' of a single object point and hence some form of scanning is necessary to produce an image of an extended region of the specimen. However, the use of single-point illumination and single-point detection results in novel imaging capabilities which offer significant advantages over those possessed by conventional wide-field optical microscopes. In essence, these are enhanced lateral resolution and, perhaps more importantly, result in a unique depth discrimination or optical sectioning property. It is this latter property which leads to the ability to obtain three-dimensional images of volume specimens.

The improvement in lateral resolution may at first seem implausible. However, it can be explained simply by the principle which states that resolution can be increased at the expense of field of view, which can then be increased by scanning. One way of taking advantage of this principle is to place a very small aperture extremely close to the object. The resolution is now determined by the size of the aperture rather than the radiation. In the confocal microscope we do not use a physical aperture in the focal plane but rather use the back-projected image of a point detector in conjunction with the focused point source. Figure 2 indicates the improvement in lateral resolution that may be achieved.

The confocal principle was introduced in an attempt to obtain an image of a slice within a thick specimen which was free from the distracting presence of out-of-focus information from surrounding planes. The confocal optical systems fulfills this requirement and its inherent optical sectioning or depth discrimination property has become the major

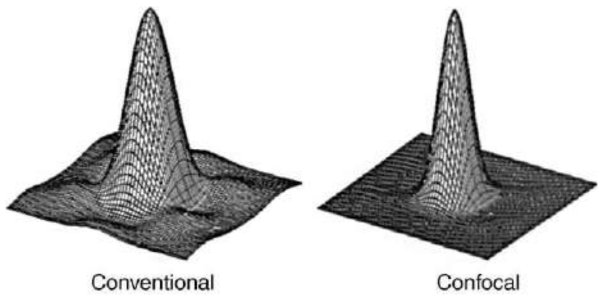

Figure 2 The point spread functions of the conventional and confocal microscopes showing the improvement in lateral resolution which may be obtained in the confocal case.

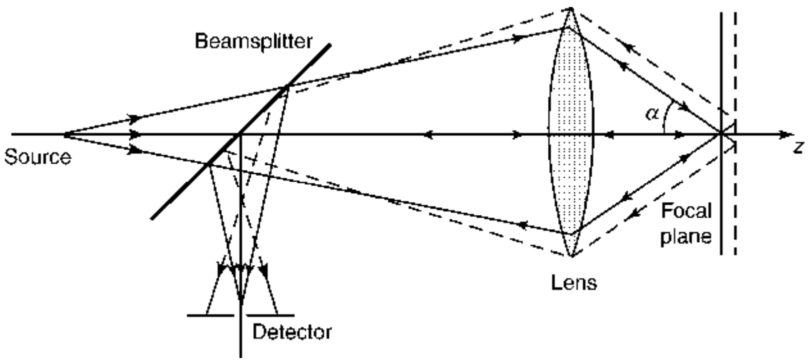

Figure 3 The origin of the optical sectioning or depth discrimination property of the confocal optical system.

motivation for using confocal microscopes, and is the basis of many of the novel imaging modes of these instruments. The origin of the depth discrimination property may be understood easily from Figure 3, where we show a reflection-mode confocal microscope and consider the imaging of a specimen with an undulating surface. The full lines show the optical path when an object feature lies in the focal plane of the lens. At a later scan position, the object surface is supposed to be located in the plane of the vertical dashed line. In this case, simple ray tracing shows that the light reflected back to the detector pinhole arrives as a defocused blur, only the central portion of which is detected and contributes to the image. In this way the system discriminates against features which do not lie within the focal region of the lens. A very simple method of both demonstrating the effect and giving a measure of its strength is to scan a perfect reflector axially through focus and measure the detected signal strength. Figure 4 shows a typical response. These responses are frequently termed the $V(z)$, by analogy with a similar technique in scanning acoustic microscopy. A simple paraxial theory models this response as

$$I(u) = \left[\frac{\sin(u/2)}{u/2} \right]^2 \qquad [1]$$

where u is a normalized axial coordinate which is related to real axial distance, z, via

$$u = \frac{8\pi}{\lambda} nz \sin^2(\alpha/2) \qquad [2]$$

where λ is the wavelength and $n \sin \alpha$ the numerical aperture. As a measure of the strength of the sectioning, we can choose the full width at half intensity of the $I(u)$ curves. Figure 5 shows this value as a function of numerical aperture for the specific case of imaging with red light from a helium neon laser. These curves were obtained using a high aperture theory which is more reliable than eqn [1]

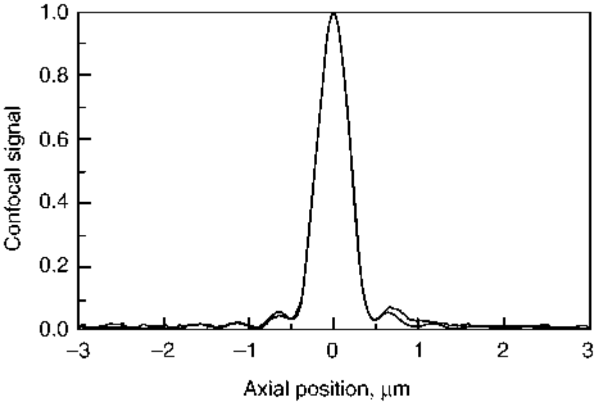

Figure 4 The variation in detected signal as a plane reflector is scanned axially through focus. The measurement was taken with a 1.3 numerical aperture objective and 633 nm radiation.

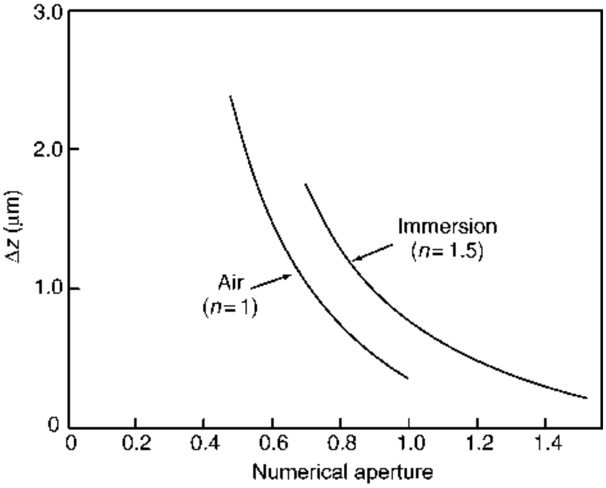

Figure 5 The optical sectioning width as a function of numerical aperture. The curves are for red light (0.6328 μm wavelength). Δz is the full width at the half-intensity points of the curves of $I(u)$ against u.

at the highest values of numerical aperture. We note, of course, that these numerical values refer to nonfluorescence imaging. The qualitative explanation of optical sectioning, of course, carries over to the fluorescence case but the actual value of the optical sectioning strength is different in the fluorescence case.

Applications of Depth Discrimination

Since this property is one of the major reasons for the popularity of confocal microscopes, it is worthwhile, at this point, to review briefly some of the novel imaging techniques which have become available with confocal microscopy.

Figure 6 illustrates the essential effect: Figure 6a shows a conventional image of a planar microcircuit which has deliberately been mounted with its normal at an angle to the optic axis. We see that only one portion of the circuit, running diagonally, is in focus. Figure 6b shows the corresponding confocal image: here the discrimination against detail outside the focal plane is clear. The areas which were out of focus in Figure 6a have been rejected. Furthermore, the confocal image appears to be in focus throughout the visible band, which illustrates that the sectioning property is stronger than the depth of focus.

This suggests that if we try to image a thick translucent specimen, we can arrange, by the choice of our focal position, to image detail exclusively from one specific region. In essence, we can section the specimen optically without having to resort to mechanical means. Figure 7 shows an idealized schematic of the process. The portion of the beehive-shaped object that we see is determined by the focus position. In this way it is possible to take a through-focus series and obtain data about the three-dimensional structure of the specimen. If we represent the volume image by $I(x, y, z)$, then by focusing at a position $z = z_1$, we obtain, ideally, the image $I(x, y, z_1)$. This, of course, is not strictly true in practice because the optical section is not infinitely thin.

It is clear that the confocal microscope allows us to form high-resolution images with a depth of focus sufficiently small that all the detail which is imaged appears in focus. This suggests immediately that we can extend the depth of focus of the microscope by adding together (integrating) the images taken at different focal settings without sacrificing the lateral resolution. Mathematically, this extended-focus image is given by

$$I_{\mathrm{EF}}(x, y) = \int I(x, y, z)\mathrm{d}z \qquad [3]$$

As an alternative to the extended-focus method, we can form an auto-focus image by scanning the object axially and, instead of integrating, selecting the focus at each picture point by recording the maximum in the detected signal. Mathematically, this might be written as

$$I_{\mathrm{AF}}(x, y) = I(x, y, z_{\max}) \qquad [4]$$

where $z_{\max}$ corresponds to the focus setting giving the maximum signal. The images obtained are somewhat similar to the extended focus and, again, substantial increases in depth of focus may be obtained. We can go one step further and turn the microscope into a noncontacting surface profilometer. Here we simply display $z_{\max}$.

It is clear by now that the confocal method gives us a convenient tool for studying three-dimensional structures in general. We essentially record the image as a series of slices and play it back in any desired fashion. Naturally, in practice it is not as simple as this, but we can, for example, display the data as an $x - z$ image rather than an $x - y$ image. This is somewhat similar to viewing the specimen from the side. As another

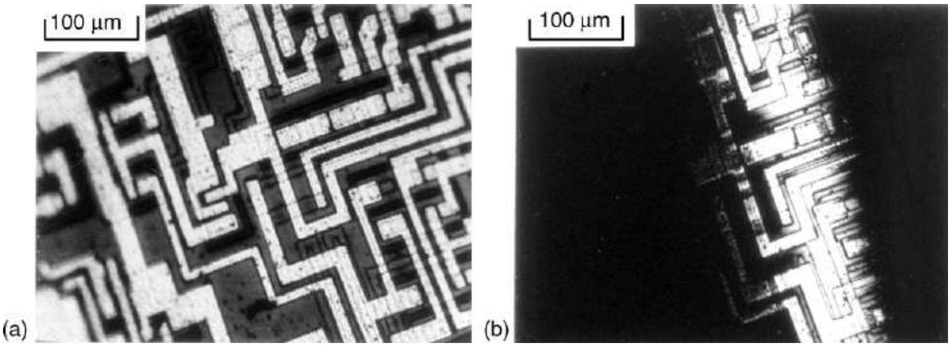

Figure 6 (a) Conventional scanning microscope image of a tilted microcircuit: the parts of the object outside the focal plane appear blurred. (b) Confocal image of the same microcircuit: only the part of the specimen within the focal region is imaged strongly.

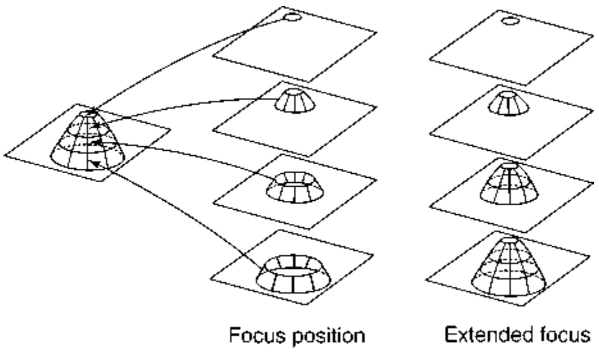

Figure 7 An idealization of the optical sectioning property showing the ability to obtain a through-focus series of images, which may then be used to reconstruct the original volume object at high resolution.

example, we might choose to recombine the data as stereo pairs by introducing a slight lateral offset to each image slice as we add them up. If we do this twice, with an offset to the left in one case and the right in another, we obtain, very simply, stereo pairs. Mathematically, we form images of the form:

$$\int I(x \pm \gamma z, y, z)\mathrm{d}z \qquad [5]$$

where γ is a constant. In practice it may not be necessary to introduce offsets in both directions to obtain an adequate stereo view.

Our discussion so far on these techniques has been by way of a simplified introduction. In particular, we have not presented any fluorescence images. The key point is that, in both brightfield and fluorescence modes, the confocal principle permits the imaging of specimens in three dimensions. The situation is, of course, more involved than we have implied. A thorough knowledge of the image formation process, together with the effects of lens aberrations and absorption, is necessary before accurate data manipulation can take place.

In conclusion to this section it is important to emphasize that the confocal microscope does not produce three-dimensional images. It essentially produces very high-quality two-dimensional images of a (thin) slice within a thick specimen. A three-dimensional rendering of the entire volume specimen may then be generated by suitably combining a number of these two-dimensional image slices from a through focus series of images.

Fluorescence Microscopy

We now turn our attention to confocal fluorescence microscopy, because this is the imaging mode which is usually employed in biological applications. Although we have introduced the confocal microscope in terms of bright-field imaging the comments we have made concerning the origin of the optical sectioning, etc., carry over directly to the fluorescence case although the numerical values describing the strength of the optical sectioning are, of course, different and we will return to this point later.

If we assume that the fluorescence in the object destroys the coherence of the illuminating radiation and produces an incoherent fluorescent field proportional to the intensity of the incident radiation, $I(v,u)$, then we may write the effective intensity point spread function, which describes image formation in the incoherent confocal fluorescence microscope, as

$$I(v,u)I\left(\frac{v}{\beta},\frac{u}{\beta}\right) \qquad [6]$$

where the optical coordinates u and v are defined relative to the primary radiation and $\beta = \lambda_2/\lambda_1$ is the ratio of the fluorescence to the primary wavelength. We note that $v = (2\pi/\lambda)rn \sin\alpha$, where r denotes the actual radial distance.

This suggests that the imaging performance depends on the value of β. In order to illustrate this we show in Figure 8 the variation in detected signal strength as perfect fluorescent sheet scans through focus. This serves to characterize the strength of the optical sectioning in fluorescence microscopy in the same way that the mirror was used in the brightfield case. We note that the half-width of these curves is

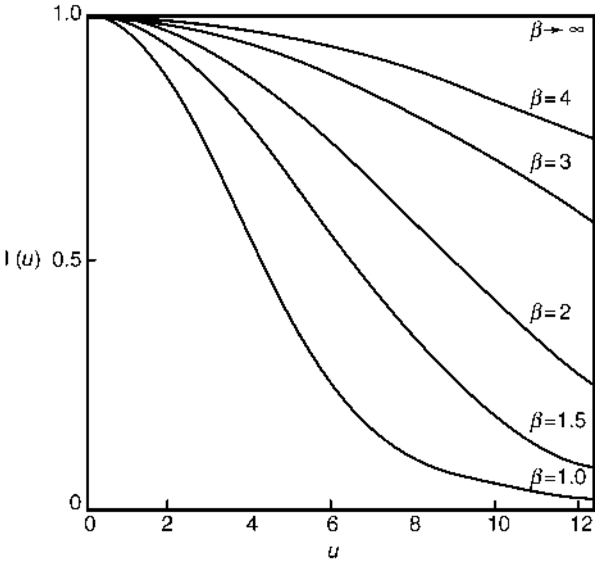

Figure 8 The detected signal as thin fluorescent sheet is scanned axially through focus for a variety of fluorescent wavelengths. We note that if we measure the sectioning by the half width of these curves, the strength of the sectioning is essentially proportional to β.

essentially proportional to β and so for optimum sectioning the wavelength ratio should be as close to unity as possible.

We have just discussed what we might call one-photon fluorescence microscopy in the sense that a fluorphore is excited by a single photon of a particular wavelength. It then returns to the ground state and emits a photon at the (slightly longer) fluorescence wavelength. It is this radiation which is detected via the confocal pinhole. Recently, however, much interest has centered on two-photon excitation fluorescence microscopy. This process relies on the simultaneous absorption of two, longer wavelength photons, following which a single fluorescence photon is emitted. The excitation wavelength is typically twice that used in the one-photon case. The beauty of the two-photon approach lies in the quadratic dependence of the fluorescence intensity on the intensity of the illumination. This leads to fluorescence emission which is always confined to the region of focus. In other words the system possesses an inherent optical sectioning property. Other benefits of two-photon fluorescence over the single photon case include the use of red or infra-red lasers to excite ultra-violet dyes, confinement of photobleaching to the focal region (the region of excitation), and the reduced effects of scattering and greater penetration. However, it should also be remembered that, compared with single photon excitation, the fluorescence yield of many fluorescent dyes under two-photon excitation is relatively low.

In order to make a theoretical comparison between the one- and two-photon modalities, we shall assume that the emission wavelength λ_{em} is the same, irrespective of the mode of excitation. Since the wavelength required for single-photon excitation is generally shorter than the emission wavelength, we may write it as $\gamma\lambda_{\text{em}}$ where $\gamma < 1$. Since two-photon excitation requires the simultaneous absorption of two photons of half the energy we will assume that the excitation wavelength may be written as $2\gamma\lambda_{\text{em}}$ which has been shown to be a reasonable approximation for many dyes.

If we now introduce optical coordinates u and v, normalized in terms of λ_{em}, we may now write the effective point spread functions in the one-photon confocal and two-photon case as

$$I_{\text{1p-conf}} = I\left(\frac{v}{\gamma}, \frac{u}{\gamma}\right) I(v, u) \qquad [7]$$

and

$$I_{\text{2p}} = I^2\left(\frac{v}{2\gamma}, \frac{u}{2\gamma}\right) \qquad [8]$$

respectively. We note that although a pinhole is not usually employed in two-photon microscopy it is possible to include one if necessary. In this confocal two-photon geometry the effective point spread function becomes

$$I_{\text{2p-conf}} = I^2\left(\frac{v}{2\gamma}, \frac{u}{2\gamma}\right) I(v, u) \qquad [9]$$

If we now look at eqns [7] and [8] in the $\gamma = 1$ limit, we find that

$$I_{\text{1p-conf}} = I^2(v, u) \qquad [10]$$

and

$$I_{\text{2p}} = I^2\left(\frac{v}{2}, \frac{u}{2}\right) \qquad [11]$$

We now see that because of the longer excitation wavelength used in two-photon microscopy the effective point spread function is twice as large as that of the one-photon confocal in both the lateral and axial directions. The situation is somewhat improved in the confocal two-photon case but it is worth remembering that the advantages of two-photon excitation microscopy are accompanied by a reduction in optical performance as compared to the single-photon case.

From the practical point of view it is worth noting that the two-photon approach has certain very important advantages over one-photon excitation, in terms of image contrast, when imaging through scattering media apart from the greater depth of penetration afforded by the longer wavelength excitation. In a single-photon confocal case it is quite possible that the desired fluorescence radiation from the focal plane may be scattered after generation in such a way that it is not detected through the confocal pinhole. Since the fluorescence is generated throughout the entire focal volume it is also possible that undesired fluorescence radiation, which was not generated within the focal region, may be scattered so as to be detected through the confocal pinhole. In either case this leads to a reduction in image contrast. The situation is, however, completely different in the two-photon case. Here the fluorescence is generated only in the focal region and not throughout the focal volume. Furthermore since all the fluorescence is detected via a large area detector – no pinhole is involved – it is not so important if further scattering events take place. This leads to high-contrast images which are less sensitive to scattering. This is particularly important for specimens which are much more scattering at the fluorescence ($\lambda/2$) wavelength than the excitation (λ) wavelength.

The Use of Structured Illumination to Achieve Optical Sectioning

An alternative approach to obtain optical sectioning is to use structured illumination in a conventional microscope. This approach is attractive since the conventional light microscope already possesses many desirable properties: real-time image capture, standard illumination, ease of alignment, etc. However, it does not produce optically sectioned images in the sense usually understood in confocal microscopy. In order to see how this deficiency may be corrected via a simple modification of the illumination system let us begin by looking at the theory of image formation in a conventional fluorescence microscope and start by asking in what way the image changes as the microscope is defocused. We know that in a confocal microscope the image signal from all object features attenuates with defocus and that this does not happen in a conventional microscope. However, when we look closely at the image formation process we find that it is only the zero spatial frequency (constant) component which does not change with defocus (Figure 9); all other spatial frequencies actually do attenuate with defocus to a greater or lesser extent. Figure 10 illustrates this by showing the image of a single spatial frequency one-dimensional bar pattern object for increasing degrees of defocus. When the specimen is imaged in focus, a good image of the bar pattern is obtained. However with increasing defocus the image becomes progressively poorer and weaker until it eventually disappears leaving a uniform gray level. This observation is the basis of a simple way to perform optical sectioning in a conventional microscope.

If we simply modify the illumination path of the microscope so as to project onto the object the image of a one-dimensional, single spatial frequency fringe pattern, then the image we see through the microscope will consist of a sharp image of those parts of the object where the fringe pattern is in focus together with an out-of-focus blurred image of the rest of the object. In order to obtain an optically sectioned image it is necessary to remove the blurred out-of-focus portion as well as the fringe pattern from the in-focus optical section. There are many ways to do this – one of the simplest involves simple processing of three images taken at three different spatial positions of the fringe pattern. The out-of-focus regions remain fairly constant between these images and the relative spatial shift of the fringe pattern allows the three images to be combined in such a way as to remove the fringes. This permits us to retrieve both an optically sectioned image as well as a conventional image in real time.

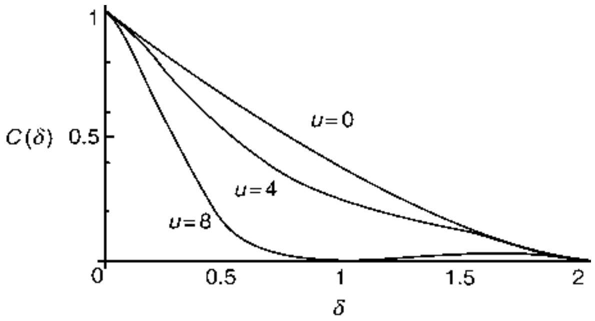

Figure 9 The transfer function for a conventional fluorescence microscope for a number of values of defocus (measured in optical units). We notice that the response for all nonzero spatial frequencies decays with defocus.

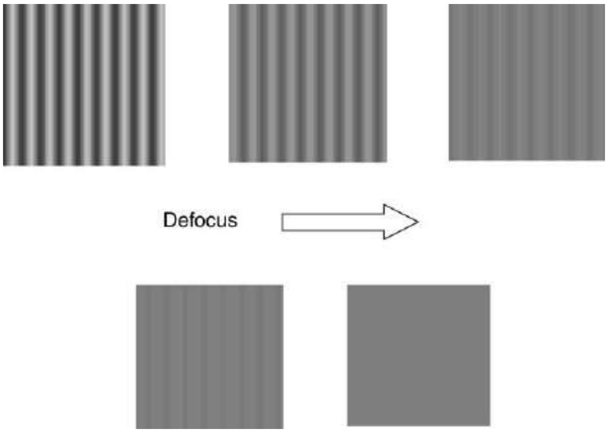

Figure 10 The image of a one-dimensional single spatial frequency bar pattern for varying degrees of defocus. We note that for sufficiently large values of defocus the bar pattern is not imaged at all.

Since the approach involves processing three conventional microscope images, the image formation is fundamentally different from that of the confocal microscope. However, the depth discrimination or optical sectioning strength is very similar and this approach, which requires very minimal modifications to the instrument, has been used to produce high-quality three-dimensional images of volume objects which are directly comparable to those obtained with confocal microscopes. A schematic of the optical arrangement is shown in Figure 11a, together with experimentally obtained axial responses in Figure 11b. It will be noted that these responses are substantially similar to those obtained in the true confocal case. As an example of the kind of images which can be obtained with this kind of microscope, we show in Figure 12 two images

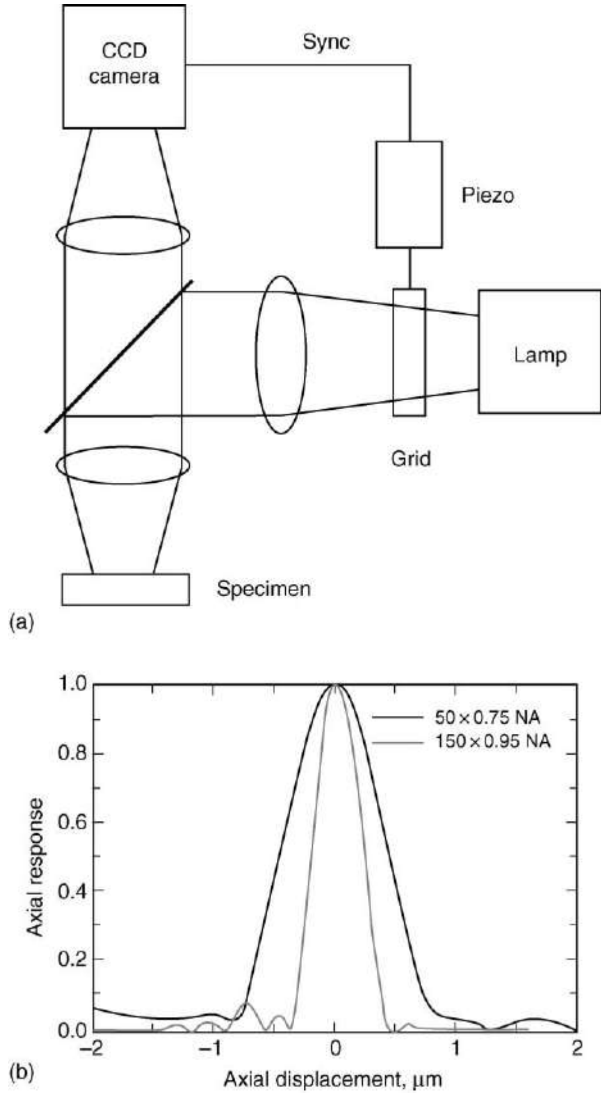

Figure 11 (a) Shows the optical system of the structured illumination microscope together with experimentally obtained axial responses in (b) which confirm the optical sectioning ability of the instrument.

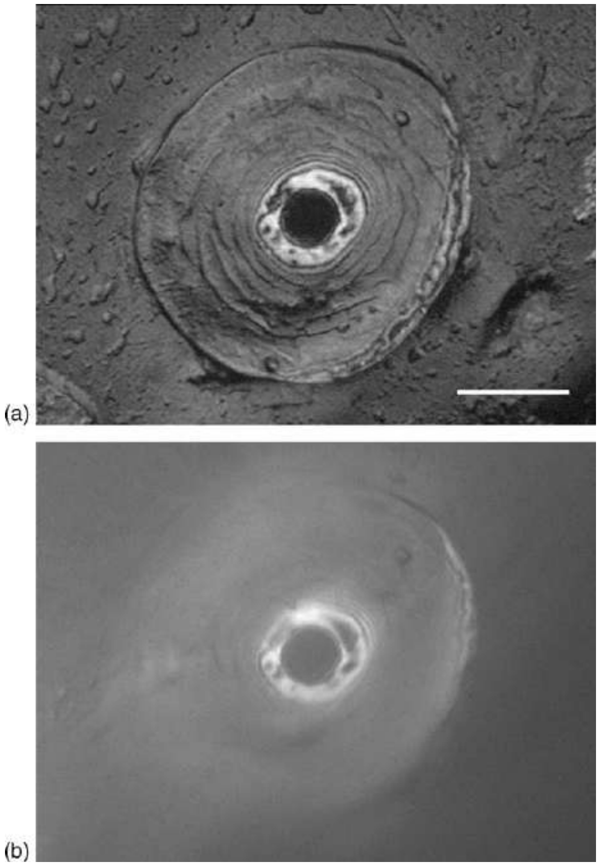

Figure 12 Two images of the region around the spiracle of a head louse. As in the previous figure, (a) is an autofocus image and (b) shows a mid-focus conventional image. The scale bar depicts a length of 10 μm.

of a spiracle of a head louse. The first is an auto-focus image of greatly extended depth of field constructed from a through-focus series of images. The second is a conventional image taken at mid-focus. The dramatic increase in depth of field is clear when compared with a mid-focus conventional image. We note that these images were taken using a standard microscope illuminator as light source. Indeed, the system is so light efficient that good-quality optical sections have been obtained or transistor specimens using simply a candle as light source!

Imaging using fluorescence light is also possible using this technique. However, an alternative approach that does not require a physical grid is possible if a laser is used as the light source. In this system the laser illumination is split into two beams which are allowed to interfere at an angle in the fluorescent specimen. This has the effect of directly ‘writing’ a one-dimensional fringe pattern in the specimen. Spatial shift of the fringe pattern is achieved by varying the phase of one of the interfering beams. As before, three images are taken from which both the optically sectioned image and conventional image may be obtained. The beauty of this approach is that no imaging optics are required at the illuminating wavelength and system.

Summary

We have discussed the origin of the optical sectioning property in the confocal microscope in order to introduce the range of imaging modes which this unique form of microscope leads to. A range of optical architectures have also been described. By far the most universal is that shown in Figure 1 where a confocal module is integrated around a conventional optical microscope. Other, more recent, real-time implementations have also been described. A number of practical aspects of confocal microscopy have not been discussed. This is because they are readily available elsewhere, such as advice on the correct choice of detector pinhole size or because they are still the focus of active research, such as the development of new contrast mechanisms for achieving enhanced three-dimensional resolution such as the stimulated emission depletion method, STED or 4-Pi which are still under development but offer great promise.

See also

Microscopy: Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Corle TR and Kino GK (1996) *Confocal Scanning Optical Microscopy and Related Imaging Systems*. New York: Academic Press.

Denk W, Strikler JH and Webb WW (1990) Two photon laser scanning fluorescence microscopy. *Science* 248: 73.

Diaspro A (2002) *Confocal and Two-Photon Microscopy*. New York: Wiley-Liss.

Gu M (1996) *Principles of Three-Dimensional Imaging in Confocal Microscopes*. Singapore: World Scientific.

Hell S and Stelzer EHK (1992) Fundamental improvements of resolution with a 4Pi-confocal fluorescence microscope using two-photon excitation. *Optics Communication* 93: 277.

Hell S and Wichmann J (1994) Breaking the diffraction limit by stimulated emission. *Optics Letters* 19: 870.

Lukosz W (1966) Optical systems with resolving power exceeding the classical limit. *Journal of the Optical Society of America* 56: 1463.

Masters BR (1996) *Confocal Microscopy*. Bellingham: SPIE.

Minsky M (1961). Microscopy Apparatus. United States Patent 3,013,467, Dec 19, 1961 (Filed Nov 7, 1957).

Neil MAA, Juskaitis R and Wilson T (1997) Method of obtaining optical sectioning by using structured illumination in a conventional microscope. *Optics Letters* 22: 1905.

Pawley JB (ed.) (1995) *Handbook of Biological Confocal Microscopy*. New York: Plenum Press.

Wilson T (ed.) (1990) *Confocal Microscopy*. London: Academic Press.

Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

Imaging Multiple Photon Fluorescence Microscopy

M Previte, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The field of light microscopy started with Ernst Abbe's introduction of the diffraction-limited objective lens in the late nineteenth century. Over the last 60 years, new imaging modalities, such as video, fluorescence, confocal, near field, and more recently multi-photon microscopy, have enabled the experimentalist to observe a plethora of new phenomena in the biological and physical sciences. Of these techniques, multi-photon microscopy has emerged not only as a valuable imaging technique but it also forms the technology base for emerging fields of optical memory and microfabrication.

The history of multi-photon microscopy (MPM) began with Göppert-Mayer's theoretical prediction of two-photon excitation in 1931. Although the theoretical groundwork was established, it took another 30 years for this process to be experimentally realized when Kaiser and Garrett observed two-photon absorption fluorescence emission from CaF_2:Eu^{+2}. Soon after this observation was made, Singh and Bradley witnessed higher-order multi-photon processes in 1964.

Historically, three-dimensional imaging was first achieved using confocal microscopy. In contrast to MPM, photo-interaction and fluorescence emission occurs throughout the excitation light path in confocal microscopy. Confocal microscopy selects the signal from the focal plane by the use of a conjugated pinhole aperture. In order to construct 3D images, successive planes must be imaged. As a result, the entire sample volume is exposed to the excitation light multiple times, and inevitably, photobleaching and photodamage occurs throughout the imaging volume. Furthermore, the pinhole aperture in the emission path causes additional signal reduction in optically thick specimens where the emission photon is scattered. Consequently, the utility of confocal microscopy is limited for imaging living biological specimens. The introduction of multi-photon microscopy, by Denk and Webb in 1990, addressed many of these limitations. By combining the methodology of multi-photon excitation with scanning microscopy, MPM offered advantages such as 3-D localized, femtoliter excitation volumes, reduced scattering and absorption of the infrared excitation light by the specimen, and a large separation between the excitation and emission wavelengths. MPM can produce excellent three-dimensional images with increased signal to noise ratios.

Theoretical Basis of Multi-Photon Microscopy

Multi-photon Excitation Selection Rules

Unlike direct ground state to excited state one-photon transitions, multi-photon transitions depend on the presence of virtual electronic intermediate states to reach an excited state (see Figure 1).

Using perturbative expansion, the probability of excitation for one-photon transitions is described by the first-order solution, and multi-photon transitions can be determined by including the higher order solutions. The interaction of an electric field vector E_λ with a molecule's dipole, $\vec{r}$, is described by the respective one-photon and two-photon transition probabilities:

$$P_{1h\nu} \approx \langle \Psi_f | \vec{E}_\lambda \cdot \vec{r} | \Psi_i \rangle^2$$
$$P_{2h\nu} \approx \left| \sum_m \frac{\langle \Psi_f | \vec{E}_\lambda \cdot \vec{r} | \Psi_m \rangle \langle \Psi_m | \vec{E}_\lambda \cdot \vec{r} | \Psi_i \rangle}{\varepsilon_\lambda - \varepsilon_{m,i}} \right|^2 \quad [1]$$

where Ψ_i, Ψ_m, and Ψ_f are the initial, intermediate, and final transition states, $\vec{E}_\lambda$ is the electric field associated with the excitation light, ε_λ is the energy difference between the ground state and excited states, $\varepsilon_{m,i}$ is the energy difference between the intermediate state m and the ground state, i, and the summation is over all possible intermediate states m. Although electronic transitions, that result from the interaction of a strong electric field with a molecular dipole, are affected by vibrational mode overlap (Franck Condon factors) and anharmonic perturbative effects, the presence of intermediate transition states creates a set of general selection rules for multi-photon absorption processes.

Molecular states have eigenfunctions with a defined parity: 'gerade' describing even or symmetric states (no change in the sign of the wavefunction when the sign changes in the spatial coordinates) and 'ungerade' describing odd or antisymmetric states (the sign of the wavefunction changes when the sign changes in the spatial coordinates). Selection rules arise from the physics inherent in these even and odd eigenfunctions and the odd dipole operator that couples transitions between ground and excited states. In one-photon excitation, transitions are allowed for ground state and excited state transitions of opposite parity. In contrast, two-photon excitation occurs between ground and excited states with the same parity, by coupling through intermediate states with opposite parity. In general, ground and excited states with the same parity have a high excitation probability for an even number of photons; ground and excited states with different parity have a high transition probability for excitations involving an odd number of photons.

In the absence of major vibrational perturbations, inductive effects, or solvent relaxation, the probability of the transition depends on the magnitude of the integral overlap between the ground, intermediate, and excited states. As a result, molecule absorption cross-sections for one-photon, two-photon, three-photon, or multi-photon excitation processes can be estimated. Choosing fluorophores with higher multi-photon cross-sections allows more efficient excitation with lower excitation intensity. Finally, it should be noted that while the absorption cross-sections for one-, two-, and multi-photon excitation of a given fluorophores are different, the molecules rapidly relax to the same vibrational level in the excited electronic state. The excited state residence time (fluorescence lifetime), and the fluorescence decay processes depend only on the molecular structure and its microenvironment. Therefore, the fluorescence

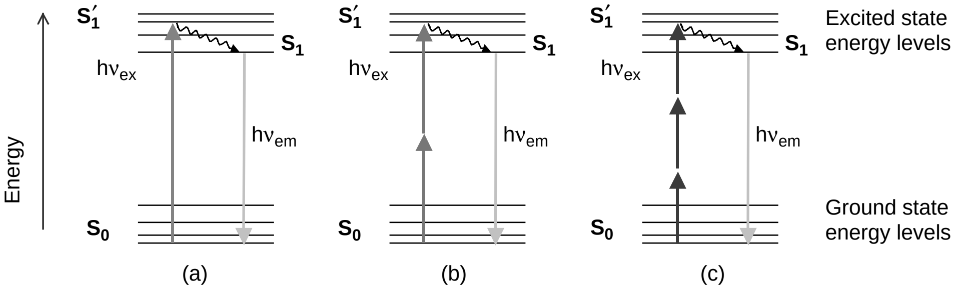

Figure 1 Jablonski diagram of a fluorescent dye molecule in one-, two- and *n*-photon excitation mode.

quantum yield and emission spectra is independent of the initial excitation process.

Enhanced Microscope Resolution Based on Multi-photon Excitation

The number of photons that are absorbed in a multi-photon process is proportional to the fluorophore concentration, $C(\vec{r},t)$, the molecule's absorption cross-section, σ_n, which depends on n, the number of photons involved in the absorption process, and the intensity of the excitation source, such that

$$N_{\mathrm{abs}}(t) = \int \mathrm{d}\vec{r} I_0^n(\vec{r},t) C(\vec{r},t)\sigma_n \qquad [2]$$

When estimating $N_{\mathrm{abs}}(t)$, the spatial and temporal intensity distribution of the excitation intensity are considered independently:

$$I_0(\vec{r},t) = S(\vec{r}) I_0(t) \qquad [3]$$

such that for an n-photon process, $I_0(t)$ is the temporal dependence of the excitation intensity at the maxima of the spatial distribution and $S(\vec{r})$ is the unitless spatial distribution of the excitation light. Combining eqns [2] and [3] gives

$$N_{\mathrm{abs}}(t) = \int \mathrm{d}\vec{r} S^n(\vec{r}) I_0^n(t) C(\vec{r},t)\sigma_n \qquad [4]$$

If it is assumed that fluorophore distribution is uniform and the system does not undergo photobleaching or ground-state depletion, $C(\vec{r},t)$ can be considered a constant and the previous equation can be rewritten as

$$N_{\mathrm{abs}}(t) = C\sigma_n I_0^n(t) \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [5]$$

By expanding the incident laser beam and overfilling the back aperture of the objective lens, the paraxial form of the normalized intensity point spread function (PSF, $h^2[u,v]$) determines the properties for a diffraction-limited lens. The dimensionless radial distance, v, from the optic axis and the distance from the in-focus plane, u, are given by

$$v = \frac{2\pi(\mathrm{NA})r}{\lambda} \qquad [6]$$

$$u = \frac{2\pi(\mathrm{NA})^2 z}{n_0\lambda} \qquad [7]$$

where $\mathrm{NA} = n_0 \sin\theta$ (θ being the half-angle of collection for the lens and n_0 is the index of refraction the medium) is the numerical aperture. Expressing the spatial distribution function with the optical units, v, u, and θ gives the following equation:

$$S(\vec{r}) = h^2[u,v] = 2\left|\int J_0(v\theta)\mathrm{e}^{-(1/2)\mathrm{i}u\theta^2}\,\theta\,\mathrm{d}\theta\right|^2 \qquad [8]$$

where J_0 is the zero-order Bessel function. If an oil immersion objective with a $\mathrm{NA} < 1.0$ is used, the paraxial approximation accurately estimates the spatial distribution, and the integral of $S(\vec{r})$ evaluates as

$$\int_{V=\infty} S^n(\vec{r})\mathrm{d}\vec{r} = a_n\left[\frac{n_0\lambda^3}{8\pi^3(\mathrm{NA})^4}\right] \qquad [9]$$

where $a_n = \int_0^\infty 2\pi v\,\mathrm{d}v \int_{-\infty}^{\infty}\mathrm{d}u[h(u,v)]^{2n}$. Since this integral cannot be solved analytically, the values of a_n are determined using numerical integrations of $h^2(u,v)$ over all space. As a result, the volume integral of the spatial distribution is always proportional to NA^{-4} and therefore independent of the number of photons involved in the excitation process. For higher NA objectives, the paraxial form of the normalized intensity PSF may no longer be an adequate approximation. The expression for temporal part of eqn [5] can be written as

$$I_0(t) = \left(\frac{\pi\mathrm{NA}^2}{\lambda^2}\right)P(t) \qquad [10]$$

where $P(t)$ is the excitation power and λ is the center wavelength of the pulsed excitation source. The temporal dependence can be described with the time average expression:

$$\langle I_0^n(t)\rangle = \left(\frac{\pi\mathrm{NA}^2}{\lambda^2}\right)^n \frac{1}{T}\int_0^T P^n(t)\mathrm{d}t \qquad [11]$$

where T is an arbitrary time interval for continuous wave lasers and $T = 1/f_{\mathrm{p}}$ for a pulsed laser with a repetition rate, f_{p}. Since multi-photon microscopy uses pulsed laser systems, the integral in eqn [11] must be evaluated accordingly. For a pulsed laser with a pulse width, τ, repetition rate, f_{p}, and averaged power, P_0, the pulse profile can be approximated as

$$P(t) = \frac{P_0}{f_{\mathrm{p}}\tau} \qquad \textit{for}\ 0 < t < \tau \qquad [12]$$

$$P(t) = 0 \qquad \textit{for} \qquad \tau < t < \frac{1}{f_{\mathrm{p}}} \qquad [13]$$

such that eqn [11] can be expressed as

$$\langle I_0^n(t)\rangle = \left(\frac{\pi\mathrm{NA}^2}{\lambda^2}\right)^n \left(\frac{P_0}{f_{\mathrm{p}}\tau}\right)^n \frac{1}{T}\int_0^T \mathrm{d}t \qquad [14]$$

and evaluated as

$$\langle I_0^n(t)\rangle = \left(\frac{\pi \mathrm{NA}^2}{\lambda^2}\right)^n (P_0)^n \left(\frac{1}{f_\mathrm{p}\tau}\right)^{n-1} \qquad [15]$$

Assuming no stimulated emission and no self-quenching, the quantity of emitted photons collected per unit time, $F(t)$, is

$$F(t) = \frac{1}{n}\phi\eta N_\mathrm{abs}(t) \qquad [16]$$

where η and ϕ are the fluorescence quantum efficiency of the dye and collection efficiency of the system, respectively. Since n photons are needed for each individual excitation event, $1/n$ is a normalization factor for N_abs which describes the total number of photons absorbed. Rewriting this expression in terms of the average number of absorbed photons, $\langle N_\mathrm{abs}(t)\rangle$, gives the average fluorescence intensity:

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta\langle N_\mathrm{abs}(t)\rangle \qquad [17]$$

where $\langle N_\mathrm{abs}(t)\rangle$ is

$$\langle N_\mathrm{abs}(t)\rangle = C\sigma_n\langle I_0^n(t)\rangle \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [18]$$

such that the measured fluorescence intensity or time averaged photon flux is expressed as

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta C\sigma_n\langle I_0^n(t)\rangle \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [19]$$

Since the time averaged fluorescence is proportional to $\langle I_0^n(t)\rangle$, eqns [9] and [15] can be substituted into eqn [19] to give a practical expression that relates average fluorescence to measurable laboratory quantities, such as P_0, τ_p, f_p, C, σ_n, n, η, and ϕ, NA, and associated numerical constants such that

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta C\sigma_n \left(\frac{\pi \mathrm{NA}^2}{\lambda^2}\right)^n (P_0)^n \left(\frac{1}{f_\mathrm{p}\tau}\right)^{n-1} \times a_n\left[\frac{n_0\lambda^3}{8\pi^3(\mathrm{NA})^4}\right] \qquad [20]$$

As eqn [19] shows, the fluorescence intensity has an nth order dependence on the spatial and temporal distribution of the excitation power. Therefore, appreciable multi-photon probability occurs only at a region of high temporal and spatial concentration of photons. Confocal pinholes can be eliminated making this technique better suited for 3D biological imaging.

The higher order (>2) dependence of the fluorescence signal upon the excitation intensity distribution results in axial confinement. Experimentally, it can be shown that over 80% of fluorescence signal emanates from a 1 μm thick region about the focal point for objectives with a numerical aperture of 1.25 under two-photon excitation. This confinement is significantly different from one-photon excitation where a spatially uniform fluorescent sample contributes equal fluorescence intensities from each z-section above and below the focal plane (assuming negligible excitation attenuation). For an excitation wavelength of 960 nm, the typical point spread function of two-photon absorption has FWHM of 0.3 μm in the radial direction and 0.9 μm in the axial direction. In comparison with confocal microscopy, the localization of the volume of excitation instead of the volume of observation, as in confocal microscope, allows 3D imaging with minimal photobleaching and damaging.

Basic Imaging Applications of Multi-Photon Microscopy

Multi-photon microscopy finds an increasingly broad range of applications, especially in biological and medical areas. For example, multi-photon microscopy has been utilized to study the physiological and biological states in various highly scattering tissues such as cornea, skin, and pancreas islet. Also, multi-photon microscopy is used in embryology to study the development of sea urchin, *Caenorhabditis elegans*, hamster, and zebra fish. Neurobiological applications include qualifying calcium dynamics dendritic spines, memory and structural plasticity, hemodynamics, and plaque evolution in Alzheimer's disease. MPM has been applied to study pharmacological distribution in tissues and in the quantification of transdermal drug delivery. Due to minimally invasive capability of multi-photon microscopy, intra-vital microscopy imaging into thick tissues of living animals is possible leading to applications in the study of cancer biology.

Although the advantages of MPM are less compelling in studying cell cultures where the specimen is optically thin, a number of uniquely exciting uses have been identified. For instance, the significantly broader two-photon absorption spectra of many fluorophores allow convenient imaging of specimens labeled with multiple color fluorophores. The localized excitation volume reduces photo-interaction with fluorophores outside the excitation volume allowing better 3D images to be obtained from cellular systems labeled with easily photobleachable fluorophores. The use of three and higher photon excitation has also allowed the excitation and imaging of important biomolecules with deep ultraviolet absorption. Another interesting application of multi-photon excitation in cellular systems is based on the ability of this method to produce localized chemical reactions. Through the photoactivation or photolysis of 'caged' molecules, their chemical structures are altered to produce biologically active molecules, fluorescent species, or other observable chemical changes. Multi-photon excitation has been used to locally uncage neurotransmitters for mapping ion channels in neurons and to uncage calcium in studying signal transduction events. In cellular systems, MPM is being used to image normally nonfluorescent biomolecules, such as serotonin in neurons, based on multi-photon induced chemical conversion and multi-photon excitation. Finally, multi-photon microscopy has also found applications in studying the conformation, aggregation, diffusion, and transport of single molecules and proteins. Two-photon fluorescence correlation spectroscopy and photon counting histograms are powerful techniques that have been applied to study protein aggregation, conformation changes, and diffusion *in vitro* with applications ranging from cell biology to high throughput drug discovery. The large separation of excitation and emission spectra of two-photon excitation also allows high sensitivity detection of fluorescence from single chromophore; however, the higher photobleaching rate at the excitation volume for some chromophores may lessen or even reverse this advantage.

Advanced Multi-Photon Microscopy Techniques

As described in the previous sections, excitation processes involving multiple numbers of photons offers considerable advantages for the imaging process in fluorescent microscopy. 4Pi Confocal Microscopy and high-speed Multifocal Multi-photon Microscopy MMM are advanced multi-photon microscopy techniques that capitalize on these advantages to improve image resolution and acquisition times in fluorescence imaging.

4Pi Confocal Microscopy

In contrast to near-field microscopy, far-field light microscopes, i.e. confocal and multi-photon microscopes, can create three-dimensional images of the interior of cells. The lateral and axial resolution of these systems is limited between 180 nm and 300 nm and 500–1000 nm, respectively. This limitation is dictated by the Abbe theory where resolution is inversely proportional to the numerical aperture of the focusing lens. When focusing with a single lens, geometric constraints limit the numerical aperture to about 1. 4Pi microscopy was developed to overcome this limitation. By utilizing two opposing objective lenses whose spherical wavefronts are coherently combined, a greater total effective numerical aperture is achieved. This results in a higher axial resolution, while the lateral resolution of the 4Pi microscope remains unchanged with respect to its confocal counterpart.

In the setup of a 4Pi microscope, the illuminating wavefronts emanate from a common point light source. The path is split by a beamsplitter, guided through two optically identical paths, and through two opposing high NA lenses that share a common focus. The fluorescence light is then collected through one lens and a confocal pinhole. In the axial direction, this arrangement yields a PSF with one main maximum lobe that is 3–4 times sharper along the optic axis than the PSF of a conventional confocal microscope, and two undesired side-lobes.

These side lobes can produce artifacts in the image and their effect needs to be suppressed for unambiguous 3D imaging. Lobe suppression can be achieved in a 4Pi system using two-photon excitation resulting in a quadratic decrease in the ratio of the side-lobe and the main lobe amplitudes. Using an excitation wavelength of 800 nm and 1.4 NA oil-immersion objectives, a FWHM along the axial direction of the main lobe is about 145 nm and side lobes are approximately 25–35% of the maximum lobe intensity. These side lobes can be further mathematically reduced by means of a linear filter after (or during) image acquisition. The filter can be applied at lobe heights below 50%, and an effective 3D PSF consisting of a single spot with a 250 nm lateral and 145 nm axial FWHM is achieved. When combined with nonlinear image restoration, 4Pi confocal microscopy achieves a 3D resolution in the 100 nm range. This resolution is similar to that obtained with near-field optical microscopy in biological applications with an added advantage of 3D imaging of the cell interior.

Since resolution improvement in 4Pi microscopy can be negated by the presence of aberrations, the usage of oil immersion lenses requires the specimen to be fixed and mounted in glycerol to minimize aberrations resulted from index mismatch. The sample thickness is further limited in the range of 20–30 μm. Due to significant index mismatches between oil and water, the imaging of live cells in aqueous medium is extremely difficult, if not impossible. In order to circumvent this difficulty, water immersion objective lenses are introduced into the 4Pi system. These lenses have a lower aperture angle $\alpha = 62°$ (NA = 1.2) than oil immersion lenses

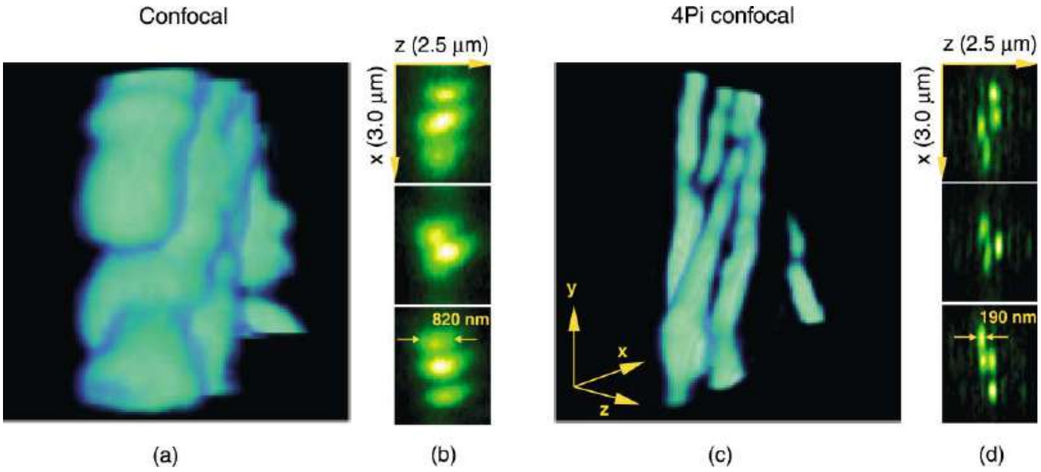

Figure 2 (a) Two-photon confocal and (c) 4Pi confocal point deconvolved images of an immunolabelled microtubule network in fixed NIH3T3 cells. The 3D rendered images in (a) and (c) were generated out of a large XZ-image-stack, partly shown in (b) and (d). As indicated in the XZ-images, the axial resolution is 820 nm and 190 nm for the two-photon confocal and the two-photon 4Pi confocal images, respectively. The images were taken with water immersion lenses (NA = 1.2). Reproduced from Bahlmann K, Jakobs S, *et al.* (2001) 4Pi-confocal microscopy of live cells. *Ultramicroscopy* 87(3): 155–164, with permission from Elsevier.

$\alpha = 68°$ (NA = 1.4) and therefore are expected to produce higher side lobes, which are reduced by 35% with careful aberration studies such that linear filtering is possible. Nonetheless, the use of water-immersion lenses solves two major problems. First, they extend the possible sample thickness to the working distance of the objective lens (~200 μm); second they allow the imaging of live cells in aqueous medium. As a result, microtubule networks (see Figure 2) and live bacteria can be imaged with a resolution of 280 nm in lateral and 190 nm in axial direction. The water immersion 4Pi microscope, utilizing two-photon excitation, is capable of the highest 3D resolution so far reported in live cell imaging to date.

Multifocal Multi-Photon Microscopy (MMM)

Improving imaging speed is critical for applying 3D microscopy in biological specimens. An MMM scans an object with an array of high-aperture foci whereas a conventional multi-photon microscope scans with a single beam. An MMM combines nonlinear fluorescence excitation with parallelized 3D imaging. The technical design consists of a densely packed microlens array; each microlens transmitting a collimated beam at a slightly different angle, overfilling the back aperture of the objective. Consequently, from the arrangement of the microlenses, an array of foci is reproduced in the image plane. The foci are scanned rapidly across the image plane to ensure full coverage of the focal plane and real-time imaging (a stack of 50 images of 30–70 μm diameter can be acquired within a few seconds). The foci's movements are either generated by rotating a microlens array that is arranged in a Lasserie pattern or by rectilinear scanning of a square microlens matrix.

MMM development was initiated because nonlinear optical damaging processes limit the usable focal peak intensity to about 200 GW/cm^2, which can be reached by a tightly focused 7–15 mW femtosecond laser beam. This power consumption makes use of only a fraction of the laser output of a typical mode-locked titanium-sapphire laser light source. The MMM better exploits the 1–2 W of the beam power by dividing the power among 25–40 foci and scanning each in parallel. With this method, typical acquisition times are cut from ~1000 ms to 10–50 ms. MMM provides an excellent method for real time 3D multi-photon imaging.

A fundamental problem in parallelized 3D microscopy is that increasing density of the foci over a certain level decreases the axial resolution. The reduction of the distance between the foci increases not only the speed or image brightness, but it also increases the interference between neighboring focal fields; especially in planes shortly before and after the focal plane. For neighboring foci in high NA objective lenses, interference effects can be significant. The relatively short laser pulses that are commonly used in multi-photon excitation provide a means to reduce this interference by using time multiplexing. Experimentally, the femtoseconds pulses of the Ti:sapphire laser, going through each lenslet, can be scheduled to arrive at the focal plane at different times by a mask of glass in front of each microlens, varying the optical path lengths of these beams. The different arrival time eliminates interference effects. The resolution of the parallelized system remains unchanged as compared with the conventional one foci system. In conclusion, the MMM enables increased image acquisition speed by a factor of the number of foci in the optical plane without compromising image resolution.

Advanced Multi-Photon Microscopy Applications

As described by the previous section, MPM technology is undergoing rapidly development with improved resolution and increased instrument throughput. Recently, MPM has also found applications in exciting new areas of optical memory and microfabrication.

Multi-Photon 3D Optical Memory

Researchers in the field of optical memory are challenged to develop instrumentation that records three-dimensional, high density, robust bitwise information in low-cost and readily available storage materials. Initially, early optical memory techniques relied on higher energy pulsed recording beams and red shifted reading beams. Aided by the development of the Ti:sapphire laser and high numerical aperture objectives that can confine laser light to a volume of the order, λ^3, the power of the focused femtosecond pulses at the focal point can reach over 10^{18} W/cm^2 with electrical field strengths reaching 2.7×10^{10} V/cm. These fields are substantial enough to locally change the optical properties of selected materials. Thus, multi-photon processes were intertwined with the field of optical memory. Recent experiments showed that with properly chosen materials and instrumentation, high density, three-dimensional data storage was possible.

Optical memory is based on photochromic material whose optical properties are changed upon irradiation such that initial and final chemical states have different absorption spectra or indices of refraction. In 1989, Parthenopoulos and Rentzepis first showed that 3D information could be stored in a polymer film doped with photochromic molecules. Although the recorded information is lost when the photo-excited compound spontaneously reverts back to the original nonfluorescent compound, Dvornikov and Rentzepis later improved the material such that the stored information was stably recorded at room temperature.

For photochromic materials, the recording and reading mechanisms often required multiple laser sources and complex optical setups. Kawata *et al.* addressed these issues with the development of an alternative optical method that used photorefractive crystals as the storage media, a Ti:sapphire as the recording device, and a phase-contrast microscope configuration for data readout of bits separated by 20 μm. Using this method, Kawata avoided the complication of different recording and read-out laser sources, thereby simplifying the instrumentation. However, in photorefractive material, poor refractive index matching, between the storage medium and the objective, results in aberration and degrades storage densities. Recent advances in objective lens design have attempted to address these limitations. Kawata and Kawata describe many of these optical memory devices in a recent review. A recent publication by Olson *et al.* describes a method for multi-photon data storage in molecular glasses and highly cross-linked polymers. These materials prove to be highly efficient and robust for long-term data storage.

Although the field of optical memory has made great progress since 1989, further advances are required before this field becomes a widely adopted technology. For example, simpler recording and reading techniques are required to insure a cost-effective and reliable data storage technology. Consequently, further progress in the production of compact ultrafast lasers and high-speed scanning instrumentation will be needed before broader acceptance of optical memory storage devices.

Multi-Photon 3D Microfabrication

Another promising emerging application of multi-photon excitation is 3D microfabrication. 3D microfabrication processes can be performed with two-photon absorption. Maruo *et al.* and Sun *et al.* provide examples of microfabricated structures generated with femtosecond pulses to induce photopolymerization. Much like the optical memory techniques, this methodology relies on photoinitiators with large two-photon cross-sections to initiate polymerization of surrounding monomer or polymer molecules. Kawata *et al.* and Sun *et al.* showed that two-photon photopolymerization could be used successfully to create intricate three-dimensional microstructures. These systems offer lower polymerization thresholds, higher polymerization rates, and wider dynamic write ranges such that rapid fabrication and low-power writing are possible. Recently, Cumpston *et al.* described a class of π-conjugated compounds with large two-photon cross-sections, which can be effectively used for two-photon polymerization initiators in three-dimensional optical data storage and microfabrication processes.

As with most of these processes, the techniques have proven to be successful, but the mechanisms behind the multi-photon initiated chemical processes are not completely understood. With further work to optimize the reaction parameters, it is possible to design experiments that will allow new materials to be synthesized or patterned based on processes such as dielectric breakdown, microexplosions, optical damage, and pyrolytic effects.

Conclusion

Relying on femtoliter excitation volumes and large separations between excitation and emission wavelengths, multi-photon microscopy has emerged as a powerful technology that allows for increased resolution and enhanced signal-to-noise ratios. It has also enabled scientists to avoid complications in imaging biological systems due to decreased sample photodamage and in some instances, photobleaching of fluorophores. Advancements in multi-photon imaging applications, such as 4Pi Confocal microscopy and multifoci multi-photon microscopy, have further enhanced the resolution of imaging technology and improved the speed of data acquistion respectively. Furthermore, the application of multi-photon technology in the fields of microfabrication and optical memory promises to address some of the difficulties involved in designing three-dimensional high-density storage devices and microfabricated components. Overall, multi-photon microscopy provides the scientific community with an invaluable tool in the study and design of biological and physical systems.

See also

Microscopy: Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Bahlmann K, Jakobs S, *et al.* (2001) 4Pi-confocal microscopy of live cells. *Ultramicroscopy* 87(3): 155–164.

Bewersdorf J, Pick R, *et al.* (1998) Multifocal multiphoton microscopy. *Optical Letters* 23: 655.

Denk W, Strickler JH, *et al.* (1990) Two-photon laser scanning fluorescence microscopy. *Science* 248(4951): 73–76.

Kawata S and Kawata Y (2000) Three-dimensional optical data storage using photochromic materials. *Chem. Rev.* 100(5): 1777–1788.

Olson CE, Previte MJR, *et al.* (2002) Efficient and robust multiphoton data storage in molecular glasses and highly crosslinked polymers. *Nature Materials* 1(4): 225–228.

Pawley JB (ed.) (1995) *Handbook of Biological Confocal Microscopy.* New York: Plenum Press.

So PT, Dong CY, *et al.* (2000) Two-photon excitation fluorescence microscopy. *Annu. Rev. Biomed. Eng.* 2: 399–429.

Xu C and Webb WW (1997) Multiphoton excitation of molecular fluorophores and nonlinear laser microscopy. In: Lakowicz JR (ed.) *Topics in Fluorescence Spectroscopy.* New York: Plenum Press.

Xu C, Zipfel W, *et al.* (1996) Multiphoton fluorescence excitation: new spectral windows for biological nonlinear microscopy. *Proceedings of the National Academy of Science of the USA* 93(20): 10763–10768.

Interference Microscopy

E Novak, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interference microscopy is a powerful noncontact metrology technique for reflection and transmission characterization of samples. The field can be divided roughly into two main categories, polarization-splitting interferometers and amplitude-splitting interferometers. In both cases, the light incident on a test sample is first divided into two beams and later recombined to form an interference pattern. This interference pattern is analyzed by one of a variety of methods to determine the difference in the optical paths the two beams traversed. This optical path difference may be used to quantify surface shape, refractive index variations, or thickness variations of the sample to a very high accuracy. This article describes the background of the two categories of interference microscopy, specifics of their implementations for different types of measurements, current state of the technologies, and examples of different measurements performed using these techniques.

Background

Interference fringes were first observed in a microscope by Robert Hooke in 1665. In the late 1800s, several practical versions of interference microscopes entered use to examine transparent specimens, though none gained widespread commercial success. Then in the 1950s, differential interference contrast (DIC) microscopy was invented to increase image contrast when examining semi-transparent samples under a microscope, particularly biological specimens. As this technique does not require staining of samples, and avoids the halo effect common in phase contrast microscopy (*see* **Microscopy:** Phase Contrast Microscopy), such systems became widely used in a variety of biological applications. In the early 1960s, Krug, Rienitz and Shulz published a comprehensive book to summarize all of the various interference microscope designs that had been produced thus far, showing the large depth and breath of research to that point.

With the advent of affordable computers and CCD cameras in the 1970s and 1980s, various commercial systems were developed using a variety of interferometric configurations to quantify the measured interference patterns from small samples. Due to their noncontact nature and ability to provide area measurements, as opposed to the two-dimensional traces produced by contact stylus techniques, such systems gained rapid acceptance in a variety of industries. Performance has steadily improved such that vertical resolution is sub-nm and accuracy on the nm-level is readily achievable. Lateral resolution is limited only by the optics of the microscope, with high numerical aperture objectives achieving resolution below 0.5 μm. Today, commercial systems are used in such measurements as index variations of biological samples, roughness of paper, microfluid channel volume, MEMS properties, hard drive head dimensions, drug coating thickness, and machine part characterization.

Polarization-Division Interference Microscopes

While the concepts behind polarization division interference microscopes have been known for well over a century, they did not gain widespread use until after Georges Nomarski invented the DIC microscope, also referred to as a polarization interference contrast microscope, in 1955. A schematic of a transmission-based and reflection-based Nomarski setup is shown in Figure 1, where the displacement of the beams has been greatly exaggerated for clarity.

The illumination in Nomarski systems is generally from an unpolarized broadband source, such as a

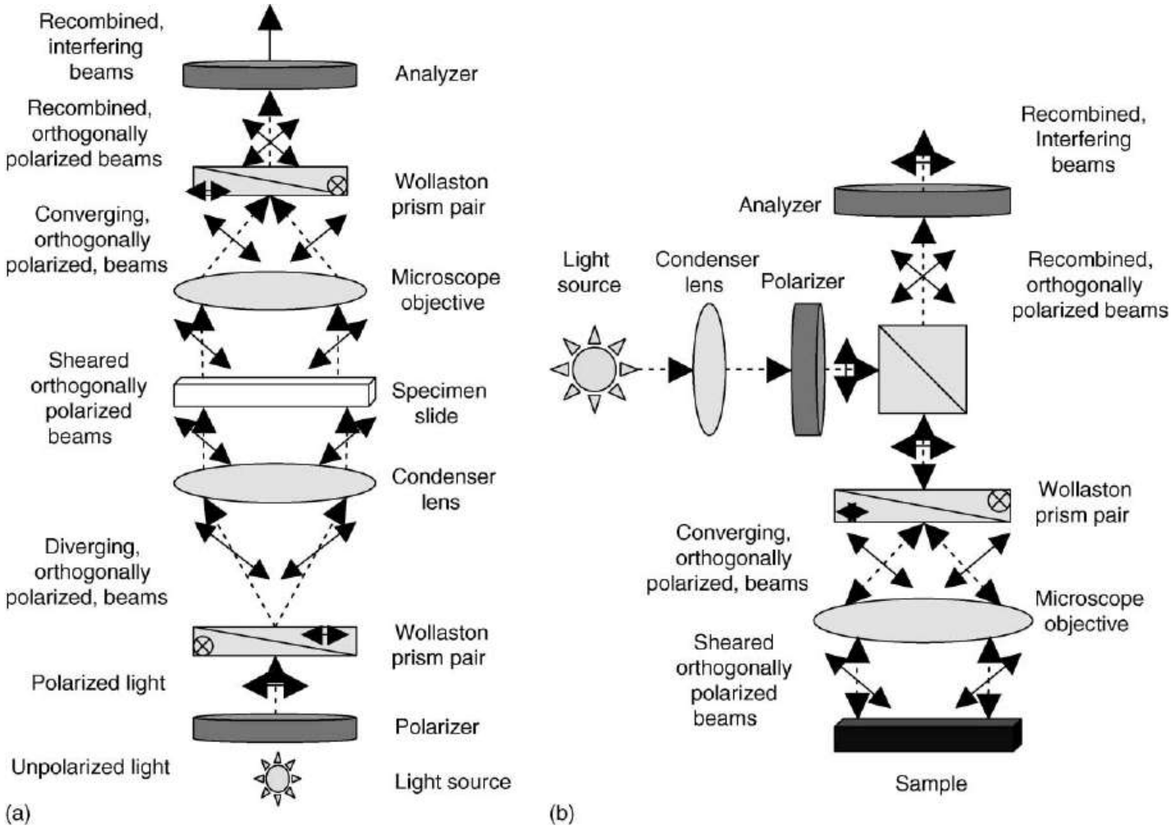
Recombined, interfering beams
Analyzer
Recombined, orthogonally polarized beams
Wollaston prism pair
Converging, orthogonally polarized, beams
Microscope objective
Sheared orthogonally polarized beams
Specimen slide
Condenser lens
Diverging, orthogonally polarized, beams
Wollaston prism pair
Polarized light
Polarizer
Unpolarized light
Light source
(a)
Recombined, Interfering beams
Analyzer
Light source
Condenser lens
Polarizer
Recombined, orthogonally polarized beams
Wollaston prism pair
Converging, orthogonally polarized, beams
Microscope objective
Sheared orthogonally polarized beams
Sample
(b)

Figure 1 (a) Schematic of transmission-based Nomarski interferometer and (b) reflection-based Nomarski interferometer.

tungsten–halogen bulb. A polarizer in the illumination path is then used to set the polarization state of the light incident on a modified Wollaston prism pair. The modified Wollaston prism consists of two birefringent wedge prisms which divide the beam in two parts, each with orthogonal polarization states. The two beams travel in slightly different directions from one another. The condenser lens brings the light to illuminate the sample, with each of the two orthogonally polarized beams slightly sheared from one another. This shear is typically much less than the resolution limit of the microscope objective, typically from tens of nanometers to several hundred nanometers, such that only one image is seen by the observer. After passing through or reflecting from the sample and the imaging microscope objective, a Wollaston prism pair is used to recombine the two polarized beams. Finally, an analyzer is used to bring both beams to a common polarization state, allowing interference between the two beams.

As the system is very sensitive to the position and optical properties of the prisms, different Wollaston prism pairs are used for each objective, usually rotating into the path automatically on a turret as the objectives are changed. In a reflection system, this prism also recombines the displaced beams after they reflect from the test piece; a transmission system will usually have only one Wollaston prism pair after the sample that can handle all objectives, instead of requiring a different recombining prism for each objective.

The observed interference pattern is an image of the gradient in the sample in the direction of shear created by the Wollaston prism pairs. In order to fully characterize the surface, the sample or optics must be rotated to capture shear images in both orthogonal directions. Essentially the technique acts as a high-pass filter, emphasizing edges and lines on the test piece.

With a white-light source, dispersion effects will cause one side of a feature to appear brighter or in a specific color range whereas the other side appears darker or in different colors. This gradient effect confers a pseudo three-dimensional appearance to the specimen. The images are quite colorful, with different colors shearing differently; by adjusting the Nomarski prisms along the optical axis, one can vary the retardation between the two beams, producing more or less contrast or color variation in the observed images.

The Nomarski system has several appealing qualities for surface and index characterization over other interferometric and polarization-based techniques. First, it is a common-path interferometer; with the exception of the slight shear, both beams pass through exactly the same optics. This means that these interferometers are mostly insensitive to vibration, as both interfering beams will be affected in the same manner. In addition, the common path nature of the system reduces the quality requirements on the optics, improving measurement quality and reducing overall system cost. In addition, Nomarski techniques do not suffer the halos associated with phase contrast imaging techniques (*see* **Microscopy:** Phase Contrast Microscopy) and provide better out of focus rejection and higher light levels since they can use the entire aperture of the microscope objective.

For reflection-based systems, the observed fringes relate to the surface slope along the direction of shear; for transmission systems, index of refraction gradients or thickness variations are observed, producing much higher contrast images of biological samples than conventional bright-field techniques. The good out-of-focus rejection mentioned earlier makes possible more accurate optical sectioning of samples, where each vertical layer of the specimen is in focus while all others are out of focus. These features, combined with the fact that high contrast images can be obtained without chemical staining, have made these systems widely used in biology. In addition, the high sensitivity to slope has made DIC systems commonly used in semiconductor inspection applications and other precision surface characterization and defect inspection applications, where automatic image processing is combined with the highly sensitive measurements to quantify both surface and subsurface defects. For instance, Figure 2 shows a bright-field image of a laser diode facet as

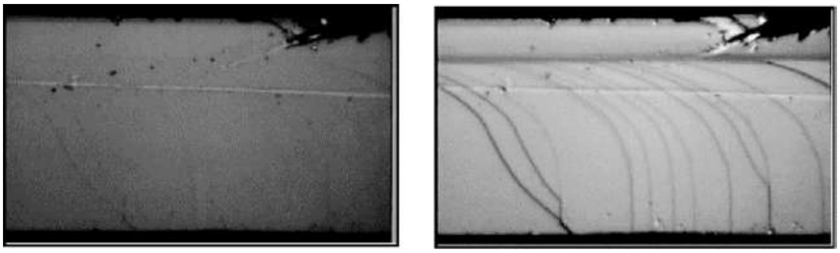

Figure 2 Bright field (left) and Nomarski (right) images of laser diode facet defects. Courtesy of Royce Instruments, Inc., Napa, CA.

well as a Nomarski image, with the facet defects greatly emphasized by the Nomarski optics.

Initial systems using differential interference contrast were strictly qualitative, with the user manually rotating the incoming polarization to produce the best contrast for a given sample and visually evaluating the resulting image. As mentioned above, visual inspection of biological specimens and defect review remain primary applications of these devices, though quantitative. By translating the Wollaston prisms laterally or rotating the analyzer in the system, one can phase shift the interference pattern. Storing multiple phase shifted images on a computer, one can obtain quantitative data about the shear using standard phase shifting interferometry equations. In addition to phase shifting through translation of the prism, quantitative data may be performed through analysis of the colors of the observed fringes or through geometric phase shifting techniques. More details on these methods can be found in the articles on **Interferometry:** Phase Measurement Interferometry; and **Interferometry:** White Light Interferometry.

Once quantitative information about the shear between the beams is obtained, the resulting data may be integrated to get back the original surface or index profile. Performing this operation on data with shear in two orthogonal directions allows the complete surface profile to be obtained. Unfortunately, since Nomarski setups are typically employed as add-on modules to standard microscope platforms rather than custom-designed metrology instrumentation, such quantitative phase shifting capability has not been widely commercialized; only a few custom or laboratory systems are in use when compared with white-light optical profilers described below. However, many research groups report nm-level surface characterization, and have developed a variety of techniques to accurately quantify thickness and index changes. The primary limitation of the quantitative techniques comes from the depth of focus of the microscope objective, which limits the vertical heights, thickness range, and index gradients that can be measured in a single measurement. Phase shifting at multiple vertical locations can be performed, but unless greatly increases the complexity of the measurement and reconstruction algorithms.

Amplitude-Division Interference Microscopes

White Light Optical Profilers

Probably the most common amplitude division interferometric microscopes are white-light interferometers, also often called optical profilers. These systems consist of three main elements: the base microscope, with illuminator and imaging optics; an interferometric element to divide and recombine the wavefront, and a scanning mechanism by which the optical path between the test and reference beams may be varied in a known manner. As the name implies, these systems typically employ spectrally broad sources, generally from a tungsten–halogen bulb. Unlike Nomarski interferometers, white light interferometers are used only in reflection-mode, as their noncommon-path nature and low-coherence source preclude interference from thick samples measured in transmission.

Figure 3 shows a schematic of a typical white light interferometer. Light from the illuminator arm, typically in a Kohler configuration, is coupled into the system with a first beamsplitter and passes through a second beamsplitter which is typically inside the microscope objective. This beamsplitter divides the beam into two parts, one of which travels to the sample and reflects from it, while the other reflects from a very high-quality reference surface. The reflected light is then recombined by the same beamsplitter and the resulting interfering light is imaged onto a camera. The static interferogram thus produced is usually by changing the optical path difference between the test and reference beams through translation of the objective with respect to the sample. The resulting signal is analyzed by one of a variety of methods as described below.

White light interferometers generally have two distinct modes of operation. In the first mode, generally referred to as phase shifting mode, a filter

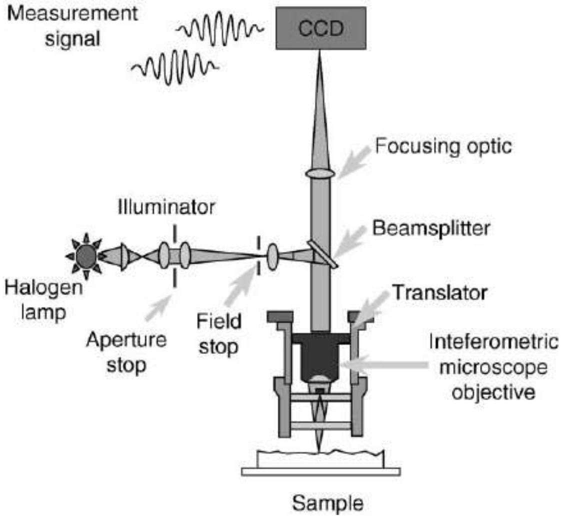

Figure 3 Schematic of white light optical profiler showing the illuminator, and translating interferometric objective.

is used to produce quasi-monochromatic light for the measurement process. This produces a large number of high-contrast interference fringes that obey the classical interference formula:

$$I_\mathrm{i}(x,y) = I_\mathrm{t}(x,y) + I_\mathrm{r}(x,y) + 2\sqrt{I_\mathrm{t} I_\mathrm{r}}(\cos(\phi(x,y) + \alpha_\mathrm{i}) \quad [1]$$

In the above $I_\mathrm{i}(x,y)$ is the irradiance observed at any one point of the image, $I_\mathrm{t}(x,y)$ and $I_\mathrm{r}(x,y)$ are the irradiances of the test and object beams, $\phi(x,y)$ is the optical path difference between the two beams, and α_i is a phase offset introduced by the interferometric instrument. By varying α_i in a known manner and collecting at least three irradiance patterns, any of the three unknowns in the above equation can be determined; specifically, the optical path difference may be solved for every point. Assuming measurements are in a medium of known and constant refractive index, such that the only contributions to the OPD are height variations rather than index variations, surface height at each point is simply:

$$H(x,y) = \frac{\lambda}{4\pi}\phi(x,y) \quad [2]$$

Phase shifting techniques produce sub-nm repeatability and very high accuracy, but suffer from a 2π phase ambiguity from the calculation of phase from irradiance data. Thus, this technique cannot handle surfaces with point-to-point steps greater than one quarter of the wavelength of the light used. Calibration methods and further information may be found in the article on **Interferometry:** Phase Measurement Interferometry.

In order to extend interference microscopy to cover large steps, steep slopes, and vertical ranges greater than the depth-of-focus of the microscope objectives, a second technique of fringe analysis may be employed, that of coherence sensing. In this method an incoherent light source is used, such as a tungsten–halogen bulb. Due to the wide spectral range of the source, as interference patterns from different wavelengths add in the observation plane, the observed interference fringes will rapidly wash out for nonzero optical path difference between the two beams. See the article on **Interferometry:** White Light Interferometry, for further details and illustrations of this effect.

With the localization of fringes only at the point of zero optical path, the instrument scans the interference objective over the desired range and then calculates the position at which maximum fringe contrast occurs for each point. Calculating this position across all points within the field creates a height map of the surface; the interference fringes serve as a tool for depth discrimination in much the same way that the pinhole in a confocal microscope does (*see* **Microscopy:** Confocal Microscopy), another technique finding wide use in surface profiling. No phase unwrapping is required, with each pixel height calculated independently of all others, avoiding ambiguities associated with phase shifting methods. The vertical range is limited only by the scan range of the measurement and the working distance of the microscope objectives; current commercial instruments can achieve vertical ranges of over 8 mm. Figure 4 shows the fringes one would observe at a given scan location during measurement of a test piece with three steps. As the OPD between the test and reference arms varies, the fringes will

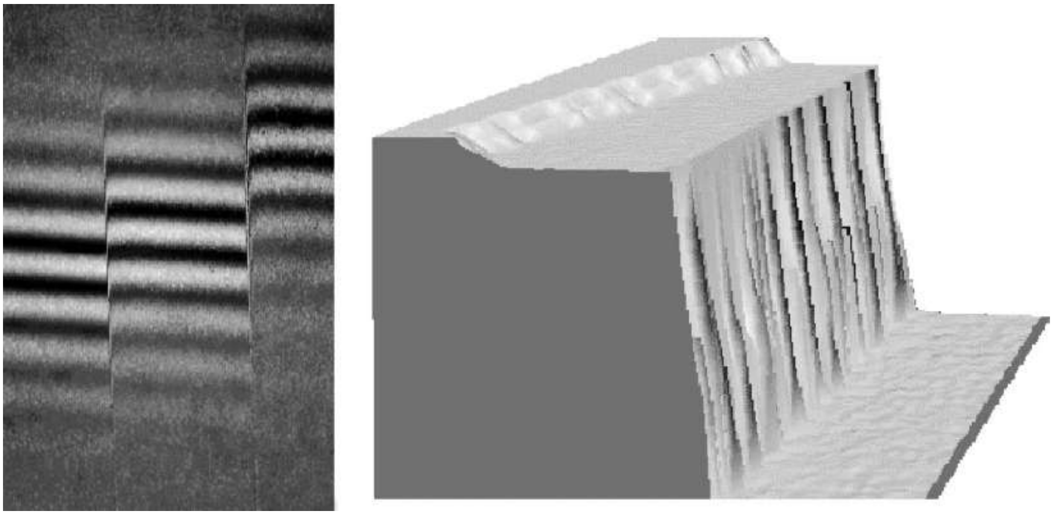

Figure 4 Localized fringes (left) of a 3-tiered surface structure (right). As the OPD is varied, the fringes will scan through the field, with the point of maximum contrast corresponding to the relative surface height of each pixel.

translate across the steps, and the surface plot at the right can be generated by calculating the scanner location of the peak contrast condition for each pixel.

There are a variety of methods for calculating the position of maximum fringe contrast using the coherence-sensing method. These include center-of-mass calculations of the peak fringe position, various frequency-space techniques and correlation of the measured fringe pattern with a theoretical pattern. Calibration of these systems generally involves repeated measurement of a known step height standard and varying the scanner motion until the step height is accurately calculated, though some systems employ separate distance-measuring interferometers to perform on-the-fly calibration for each scan. Generally speaking, coherence sensing techniques have repeatability in the range of a few nm, as opposed to the sub-nm repeatability of phase shifting techniques. This reduced sensitivity is the trade-off for avoiding the height and slope limitations of phase-shifting techniques. More details can again be found in the article on **Interferometry:** White Light Interferometry.

Recent advances in computing power have enabled an additional method of calculating surface height by combining coherence sensing with phase shifting. Coherence sensing methods are used to determine the point of best fringe contrast, and uses this information combined with the phase calculations to overcome ambiguities caused by large steps or steep slopes. Thus, one can obtain the sub-nm repeatability of phase-shifting methods without limitations on the vertical range or slope of the sample to be measured.

New fringe analysis techniques are extending the utility of these systems even further. Strobed sources are now available which effectively freeze the periodic motion for the instrument; this allows dynamic properties of samples to be measured, from surface deformation under motion to resonance frequency and switching time calculations. By analyzing the multiple interference patterns generated by transparent films deposited on a surface, film thickness, homogeneity, and differences between substrate and film surface properties may be determined.

Interferometric Microscope Objectives

One of the most critical elements in white light interferometers is the interferometric objective which splits and recombines the wavefront to form an interference pattern. Three primary interferometric configurations are used to divide and recombine the wavefront: Michelson, Mirau, and Linnik. Each of the three types of objectives is described below.

For low-magnification objectives, typically 5X or below, a Michelson interferometer configuration is typically used. A schematic of such an objective is shown in Figure 4. The illuminating beam passes through the objective and is divided by a beamsplitter placed just underneath the objective. A portion of the light travels to the test sample and the rest to a high-quality reference surface. The light reflected from each beam path is recombined by the beamsplitter and propagates to the observation plane. Note that it is important that the distance between the beamsplitter and reference surface match the distance between the beamsplitter and focus position of the sample. If the reference surface is defocused, fringe contrast will be reduced, aberrations increased, and overall system accuracy and repeatability will suffer.

This objective configuration has several advantages. First, one can add a Michelson housing to almost any standard objective with sufficient working distance for the beamsplitter, making these objectives easy to produce and relatively inexpensive. Second, the reference surface can be easily changed by the user to closely match the reflectivity of the test piece, ensuring maximum fringe contrast in the interfered beam (*see* **Interferometry:** Phase Measurement Interferometry). Third, provided there is distance between the beamsplitter and reference, one can insert a dispersion-compensating plate into the reference arm to allow measurement of parts beneath a cover glass or other dispersive medium; without such compensation, fringe contrast is generally insufficient for accurate measurements for transmissive media beyond a few hundred micrometers thickness.

As one moves towards higher magnifications, however, the working distance of the objective is typically insufficient to accommodate a large beamsplitter and still leave a reasonable distance between the objective and sample. In addition, a large cube beamsplitter will generally introduce unacceptably high aberrations in these higher numerical aperture objectives. For this reason, objectives from 10X magnification to 50X magnification are typically of the Mirau configuration, shown schematically in Figure 5.

Mirau objectives contain two additional plates compared to a standard bright-field objective: one is coated and acts as a beamsplitter, splitting the incident beam, while a second holds a small, high-quality reference surface and acts as a compensating plate so that optical paths in both arms of the interferometer are balanced (Figure 6). Such objectives are more expensive than Michelson designs, but maintain long working distances even at higher magnifications. Thermal stability becomes

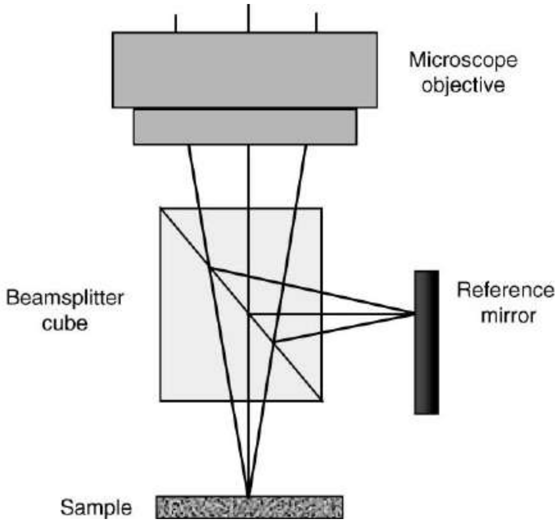

Figure 5 Schematic of a Michelson interference objective used for low-magnification applications.

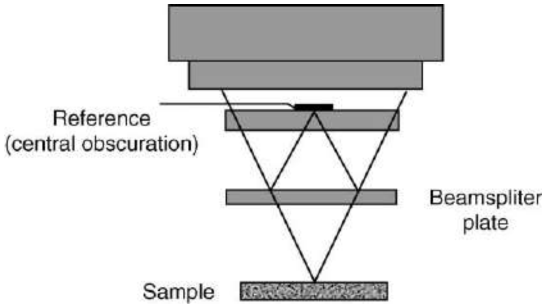

Figure 6 Schematic of a Mirau interference objective used for medium-magnification applications.

very important, as depth of focus is much reduced over lower-magnification objectives and the path from beamsplitter to reference still must be the same as from beamsplitter to in-focus test object, and most Mirau designs are therefore thermally compensated.

As one moves to even higher magnifications and numerical apertures, such as 75X or 100X, the microscope objectives do not have sufficient working distance to allow for any reasonable beamsplitter after the objective unless the numerical aperture is greatly reduced. To maintain high numerical aperture and high lateral resolution, a Linnik configuration must be used. A schematic of this configuration is shown in Figure 5. In this configuration, two separate microscope objectives are used after a beamsplitter, with one beam again reflecting from a high-quality reference and the other from the sample.

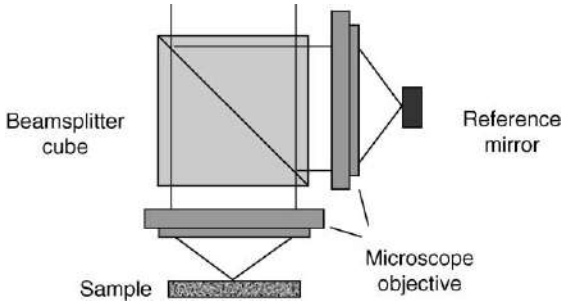

Figure 7 Schematic of a Linnik interference objective used for high-magnification applications.

A Linnik objective is generally the most difficult to use and expensive interferometric configuration (Figure 7). High numerical aperture typically limits working distance to a few hundred micrometers. In addition, depth of focus is very small, and thus these objectives must often actively adjust the reference mirror position to maintain focus over long periods of time and environmental conditions. Also, the dispersion and aberrations in the two objectives must be closely matched to ensure good fringe contrast, as there are many noncommon-path optical elements in this design; typically many objectives must be evaluated to find two that suitably match, another contributor to cost. However, these objectives do have the highest lateral resolution, and with an NA of up to 0.95, can measure the steepest slopes due to a large collection angle.

Applications

White-light interferometers are noncontact, and thus will not damage samples under test. They provide true three-dimensional surface characterization with nm-level accuracy and repeatability, and can be employed in production and laboratory environments. The various interferometric objectives allow for lateral resolutions and ranges to meet most needs; fields of view may even be stitched together to create surface maps whose extent is limited only by the range of the translation stage employed. In addition, each scan is typically quite fast, from under one second in phase-shifting mode to vertical scan rates of up to 100 μm/sec in coherence-sensing mode, with no sample preparation time. While the systems must be properly calibrated, vibration minimized, contrast maximized, and potential error sources such as dispersion and phase change on reflection effects must be managed, they remain one of the most accurate surface characterization tools available.

All of the above factors have enabled the use of white light interferometers in a variety of fields. The semiconductor and data storage industries

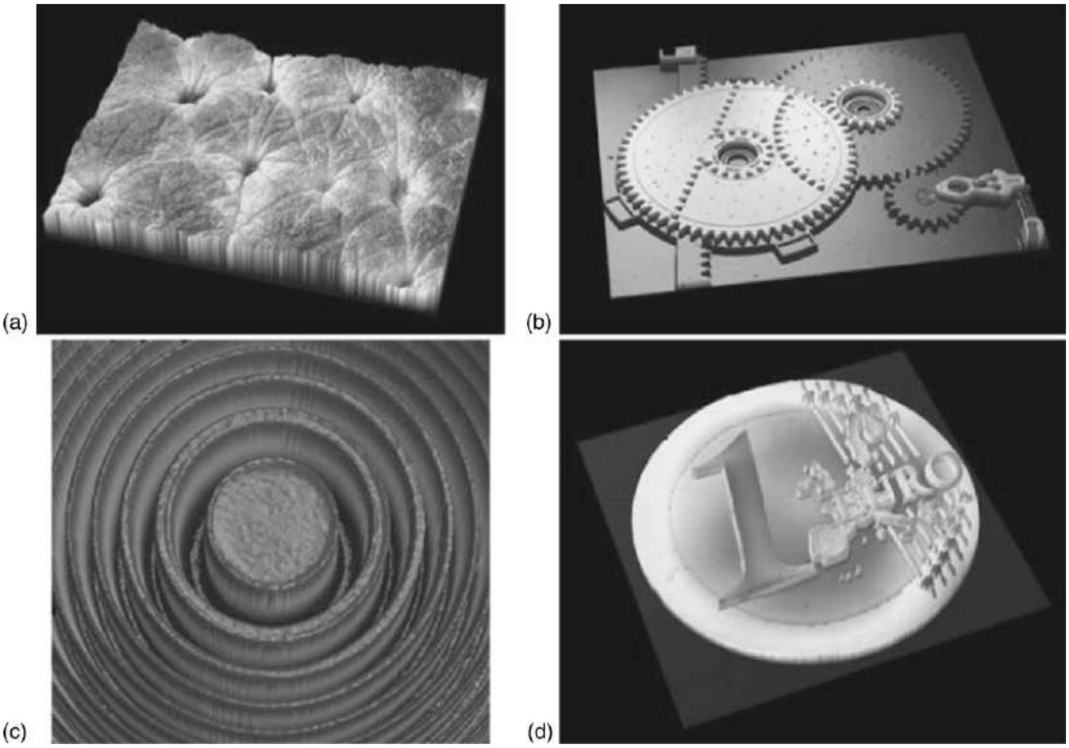

Figure 8 Surface plots from a white light optical profiler. (a) Human skin sample, (b) MEMS microgears, (c) binary optic, and (d) 1 Euro coin. Pictures courtesy of Veeco Instruments.

perform critical dimension and surface characterization of various components using these systems, some of which are configured for full wafer-handling capabilities. Systems are also used to measure precision machine parts, optics, and medical instrumentation and implants. The high sensitivity of the coherence sensing techniques allows surfaces with less than 0.1% reflectivity to be measured, so that characterization of paper, abrasives, and even different foods has been measured with these instruments. With the addition of the advanced technologies mentioned above, these systems are also now doing dynamic measurements of MEMS and thermal response characterization of precision surfaces, and also are measuring thicknesses of adhesives, drug coatings, photoresist, and a variety of other semi-transparent layers in various applications. Figure 8 shows several surface measurements, including human skin, MEMS microgears, a binary optic, and 1 Euro coin.

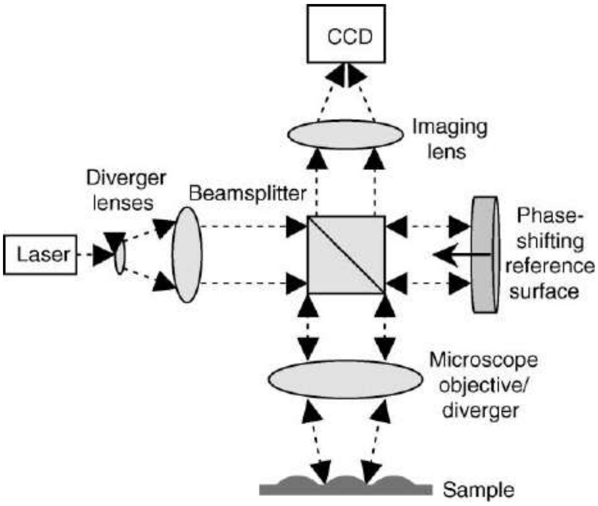
CCD
Imaging lens
Diverger lenses
Beamsplitter
Laser
Phase-shifting reference surface
Microscope objective/ diverger
Sample

Figure 9 Schematic of a micro-LUPI interferometer operating in reflection mode.

Laser-based Interference Microscopes

Another, less common type of amplitude-division interference microscope has recently been commercialized, primarily for use in examining micro-optics. These laser-based systems employ a Tywman–Green interferometric configuration combined with a microscope, and are generally called micro-LUPI (laser unequal path interferometer) systems. A schematic of a micro-LUPI interferometer is shown in Figure 9. A beamsplitter divides the incident laser beam, half of which travels to a generally flat reference surface, while the other half typically travels through a microscope objective that acts as a diverger lens and is incident on the test sample.

Due to the long-coherence length of the laser source, the two optical paths do not have to be equal in order to observe interference fringes, and samples may be measured either in reflection or transmission, depending on the desired test result. Collimated, diverging or converging illumination may be brought incident on the sample, depending on the optics in the test beam and their placement. The diverger elements are noncommon-path optics, and reduce the ultimate accuracy of the system to approximately lambda/10, similar to laser-based Fizeau interferometers.

Micro-LUPI-based systems have several advantages for testing small optics over other interference microscopes. First, the ability to measure both transmission and reflection properties of the optics, as well as different incident illumination schemes, means they can test the optics in their use configuration rather than inferring optical properties based solely on surface measurements. They can determine imaging properties, inhomogeneities and thickness variations in addition to characterizing the surfaces under test. Also, the curved test wavefront better matches spherical test surfaces, minimizing the number of observed fringes and making testing of small aspheres easier. Lastly, these systems can have curved reference surfaces as well as flat, and with the proper configuration aberrations in measurements due to the interferometric microscope can be minimized, ensuring high-quality measurements of the optic under test.

These systems have only phase-shifting measurement capability, as the coherence length of the source is too long for white-light interferometric techniques. Transmission tests necessitate a more complex part handling structure, with both the test piece and reflective return optic needing adjustment for tip/tilt and vertical and lateral translation. Therefore, such systems are more costly and not as versatile as a white light optical profiler, but do provide a valuable capability to the micro-optics community.

See also

Coherence: Coherence and Imaging. **Imaging:** Interferometric Imaging. **Interferometry:** Overview; Phase Measurement Interferometry; White Light Interferometry. **Microscopy:** Confocal Microscopy; Overview; Phase Contrast Microscopy.

Further Reading

Advanced Light Microscopy: Measuring Techniques; *Maksymilian Pluta*; North-Holland (1993).

Born M and Wolf E (eds) (1999) *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th edn. Cambridge University Press.

Creath K (1988) Phase-measurement interferometry techniques. In: Wolf E (ed.) *Progress in Optics*, vol. XXVI, pp. 349–393. North-Holland, Amsterdam.

Hariharan P (2003) *Optical Interferometry*, 2nd edn. Academic Press.

Hariharan P and Malacara D (eds) (1995). Selected Papers on Interference, Interferometry, and Interferometric Metrology. *SPIE Milestone*. MS 110.

Krug W, Rienitz J and Shulz G (1964) *Contributions to Interference Microscopy* (translated from the German edition of 1961 by J. Home Dickson): Hilger and Watts Ltd.

Malacara D (ed.) (1992) *Optical Shop Testing*. John Wiley & Sons.

Massie NA (ed.) (1988). Interferometric Metrology. *SPIE* 816.

Polarvarapu PL (ed.) (1997) *Principles and Applications of Polarization-Division Interferometry*. John Wiley & Sons.

Tolansky S (1970) *Multiple-beam Interferometry of Surfaces and Films*. Dover Publications.

Nonlinear Microscopy

S Lévêque-Fort and P Georges, Université Paris Sud, Orsay, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear microscopy has become an important tool for the investigation of biological applications where high resolution and three-dimensional imaging are essential for understanding the underlying biological function. Two-photon excitation fluorescence (TPEF) microscopy has opened the way to nonlinear techniques and is now widely available. As this new technology enables noninvasive study of biological specimens with high resolution, the number of applications in biology and medicine is still increasing. By exploiting autofluorescence of cells or tissues components, or by coupling with genetically modified fluorescent indicators or small extrinsic probes, high contrast and localization can be obtained. Different detection techniques (fluorescence lifetime imaging (FLIM), fluorescence correlation spectroscope (FCS), intensity, and spectral imaging) permits the analyze of several fluorescence properties and so obtains complementary informations.

The successful development of two-photon techniques has generated substantial interest in the scientific imaging community in further developing and exploiting optical nonlinearities for generating new image contrast. All these new developments have strongly benefited from the major advances in ultrafast lasers which have become more compact and reliable. Second- and third-harmonic generation (SHG, THG) as well as coherent anti-Stokes Raman scattering microscopy (CARS) have thus benefited from the fact that it is now easier to generate ultrashort pulses at almost any desired wavelengths. Association of these new nonlinear techniques in the same instrument (TPEF-SHG/TPEF-CARS, TPEF/THG) is commonly encountered and provides a powerful instrumental tool for probing biological samples.

Basic Principles of Two-Photon Microscopy (TPM)

Two photon microscopy is preferred over single photon method in many biological applications due to several intrinsic advantages which arise from having a nonlinear intensity depend absorption. In general, the nonlinear polarization for a material can be expressed as

$$P = \chi^{(1)}E^1 + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots \qquad [1]$$

where P is induced polarization, $\chi(n)$ is the nth order nonlinear susceptibility, and E the electric field vector of the incident light. The first term describes normal absorption and reflection of light; the second term describes SHG, hyper-Rayleigh scattering, sum and difference frequency generation, and the third terms covers multiphoton absorption, third-harmonic generation and stimulated Raman processes.

In 1931 Maria Goppert Mayer predicted that photons of lesser energy can together cause an excitation ‘usually’ produced by the absorption of a single photon of higher energy, in a process call multiphoton excitation. The hypothesis could not be confirmed until the invention of pulsed ruby lasers, in the 1970s, and was first applied to microscopy with the work of Denk *et al.* Two-photon microscopy used the simplest version of her theory: two photons of equal energy (from the same laser) can excite a molecule, and the excitation is similar to that obtained with a single photon possessing twice the energy (Figure 1). Thus, the molecule can emit a fluorescent photon just as if it was excited with a single higher energy photon (Figures 1 and 2). Of course, the absorption of the two photons must occurs nearly simultaneously ($\sim 10^{-16}$ s), resulting in a quadratic dependence on the light intensity rather than the linear dependence of conventional one photon fluorescence.

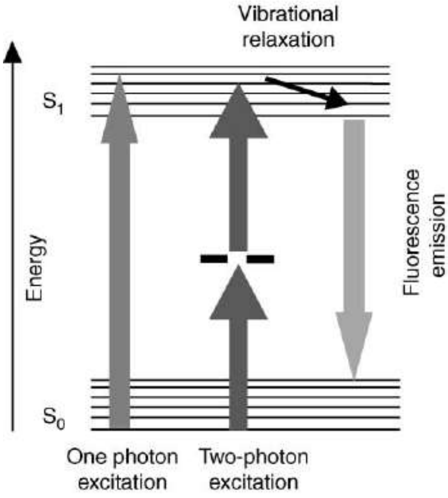

Figure 1 Jablonski diagram for one and two-photon excitation.

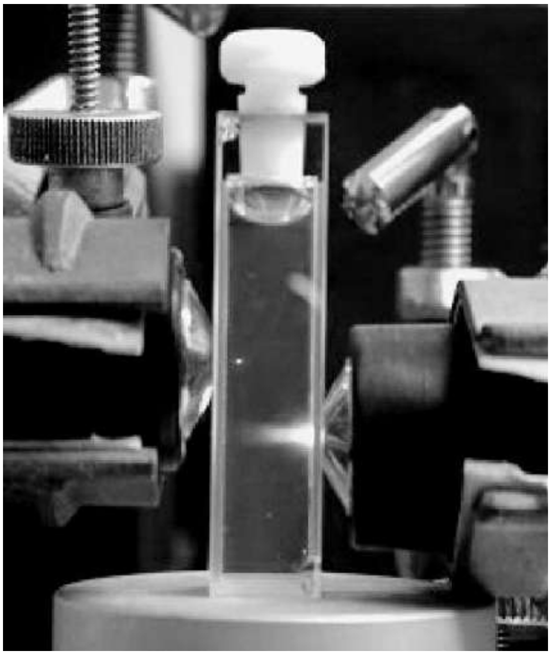

Figure 2 Localization of excitation under two photon excitation (left) compared to one photon excitation (right).

The nonlinear optical absorption property of two-photon excitation, limits the fluorochrome excitation to the point of focus. This localized excitation volume results in greatly improved axial resolution and image contrast compared with conventional microscopy, but also significantly decreases sample photobleaching and photodamage by reducing the region of photointeraction (Figure 2).

Another major advantage of TPEF is its ability to image deeper inside biological samples, owing to the reduced scattering and absorption of near infrared wavelength compared to UV and visible light. Near infrared light (700–1000 nm) is located within the 'therapeutic optical window' where the absorption of the main cells and tissue components is low.

TPEF also permits a high rejection of residual excitation light: in standard one-photon microscopy, the excitation wavelength can be spectrally close to the fluorescence emission band. Filters used to eliminate the leakage of excitation light into the detection channel often cuts off a part of the fluorescence emission. For TPEF excitation and emission band are well separated, and highly efficient filters can remove the residual excitation light without compromising the microscope sensitivity.

Under two-photon excitation, the fluorescence photon flux depends on the square of incident photon flux:

$$F \propto \sigma_{2\mathrm{P}} C I^2 \qquad [2]$$

where $\sigma_{2\mathrm{P}}$ is the two-photon cross-section (cm^4 s) with 10^{-50} cm^4 s, called a Goppert–Mayer (GM), C the fluorophore concentration, and I^2 the incident photon flux. The two-photon cross-section is a quantitative measure of the probability of two-photon absorption and as a second-order process is low. As it is difficult to directly measure two-photon cross-sections, we usually use the cross-section of two photon-induced fluorescence (σ_{TPEF}) as the product of two photon absorption cross-section ($\sigma_{2\mathrm{P}}$) with the fluorescence quantum yield $\phi_{\mathrm{F}}\, \sigma_{\mathrm{TPEF}} = \sigma_{2\mathrm{P}}\phi_{\mathrm{F}}$.

Intrinsic fluorophore, such as nicotinamide adenine dinucleotide (NADH), have low cross-sections ($<10^{-4}$ GM), usual fluorescent dye in biology ranging between 1 and 300 GM and quantum dots approaching 50 000 GM. As the probability of two-photon absorption is quiet low, high incident photon flux needs to be delivered to the sample to generate efficient absorption:

$$I \propto E \frac{1}{\Delta\tau\omega^2} \qquad [3]$$

I depends on the energy per impulsion E (mean power/pulse frequency), the pulse duration $\Delta\tau$, and the beam waist ω. The expression of I clearly demonstrates that by using ultrashort pulses and by reducing the beam waist, one can maximize the incident flux. Currently the most commonly used laser source for multiphoton microscopy is the femtosecond titanium-sapphire (Ti:Sa) system. These pulsed femtosecond lasers deliver a 100 fs pulse train at a repetition rate of ~80 MHz with a tuning range from 700 to 1000 nm. The beam waist depends on the objective numerical aperture. If we consider a Gaussian beam profile, with radial and axial parameters of σ_{r} and σ_{z} respectively and for a high numerical aperture NA > 0.8:

$$V = \left(\frac{\pi}{2}\right)^{3/2} \sigma_{\mathrm{r}}^2 \sigma_{\mathrm{z}} \qquad [4]$$

with $\sigma_{\mathrm{r}} \approx 0.52\lambda/\sin\alpha$ and $\sigma_z \approx 0.76\lambda/(1-\cos\alpha)$. where λ is the excitation wavelength, α is the one-half of the objective angular aperture ($\mathrm{NA} = n\sin\alpha$, n index of the immersion medium). So for a wavelength of 780 nm and an objective working with water immersion with a numerical aperture of 0.9, the excitation volume can be described as an obloidal shape with a radial axis of typically 500 nm and an axial axis of 1.90 μm. It corresponds to an excitation volume of ≈1 fL ou 1 μm^3. Higher resolution can be achieved using an immersion oil objective (NA = 1.3) at the cost of a higher cost and a shorter working distance.

Typical powers used to generate two-photon excitation without damaging the sample are between 1 and 10 mW for the mean incident power. For a laser beam of 1 mW, at a rate of 76 MHz with 100 fs pulse duration, and with an objective leading to a surface in the focal plane of 1 μm^2, the peak power is 0.13×10^9 W/cm^2 at the focal plane of the sample. Of course each samples has its own limits regarding incident intensity, and the key to use TPEF safely is always to reduce the excitation intensity in order to avoid photodamage.

Fluorescent Probes

Biological systems often possess endogenous fluorescent molecules that can be imaged, revealing important sample characteristics without the need of labeling (Figure 3a,b). Cells or tissues can be imaged through their autofluorescent, with the emission of several components, NADH, flavins, collagen fibers, etc. Pharmacology studies also exploit the natural autofluorescence of aromatic drugs, to

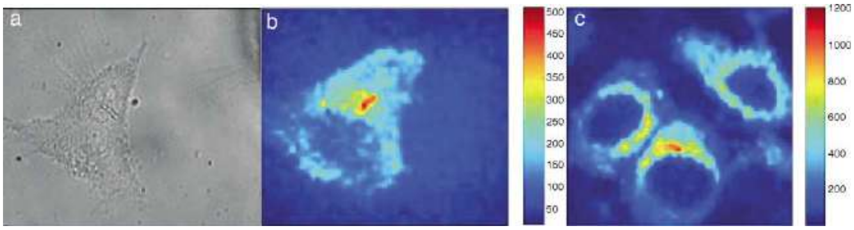

Figure 3 Applications showing various capabilities of TPEF: (a) Hela cells in transmitted image and corresponding autofluorescence image (b) under TPE (780 nm). (c) After adding 40 μM of an anti-HIV drug on similar cells, intrinsic fluorescence of the drug allows us to find its preferable localization inside the cytoplasm.

localize drug inside cells and study their reactivity without any labeling.

However, for some specific applications where the autofluorescence is too low, or if one needs to specifically identify different cells components (Figure 3g), or follow protein mobility, the samples need to be labeled your with extrinsic fluorophores. Extrinsic fluorophores are organic molecules, first developed for one-photon excitation, and are used under two-photon excitation with more or less efficiency as OPE and TPEF have different quantum mechanical selection rules, and the TPE spectrum is not always scaled to the half of OPE spectrum. Current works are focused on the development of specific markers for TPE or SHG to enhance contrast and also reduce incident laser power.

Two new generations of marker have been developed: the development of the green fluorescent protein (GFP) and its mutant introduces a convenient method for monitoring gene expression in cells and tissues; quantum dots are the latest focus in fluorescence labeling. They are promising reporters as their size (1–100 nm) induces a specific narrow emission spectrum in the visible. Absorption can occur over a wide range of short wavelengths because of the existence of multiple electronic states. Time-domain measurement has been performed with CdSe quantum dots which shows multiexponential decay with fluorescence lifetime of 3.4 ns, 16.1 ns, and 35.6 ns, much longer than autofluorescence lifetimes, hence providing greater sensitivity than intensity-based measurement. However, like GFPs, quantum dots display blinking behavior.

Two-photon Applications

Two-photon excitation has become the technique of choice for fluorescence microscopy and is applied to a variety of imaging tasks on samples ranging from single molecules to thick tissues. Owing to the high NIR penetration depth, biopsies and living tissue can be noninvasively investigated by morphological and functional fluorescence imaging of endogenous (Figure 3b) or exogenous fluorophores to depths of more than 300 μm. In genomics, multiphoton multicolour fluorescence *in situ* hybridization (FISH) are used to label various specific regions of DNA, and identified different genes and chromosome regions. In pharmacology, a higher penetration and localization of photodynamic therapy is observed, detection of low drug concentration inside cells facilitates the reactivity comprehension (Figure 3b,c). Neurobiologists investigate calcium dynamics inside brain slices and live animals, or follow neurotransmitter diffusion. Of course, these are only a few examples among the increasing number of applications, and some of them are illustrated in Figure 3.

Two-photon Instrumentation

Two-photon microscopes are now commercially available, however a large number of groups have built their own from separate components or by modifying an existing confocal microscope. Homemade microscopes present a higher configuration flexibility for excitation/detection at a much lower price.

Femtosecond, picosecond, and continuous wave (CW) laser sources have been used for TPM. By using a cw laser sources such as ArKr and NdYag, the average power increases by ~200-fold to achieve the same excitation rate obtained with a femtosecond laser, which is not highly desirable for biological samples. So as previously shown, the TPE efficiency is higher with shorter pulses, so Ti: Sa femtosecond lasers are most commonly used.

One of the drawbacks of TPM is that it is a laser scanning technique. To reconstruct an image you need either to move the sample or the laser beam. Most commonly x–y mirrors driven by a galvanometer scanner are used to scan the laser at the focal plane of the sample. Lenses are used to expand the beam to overfill the objective back aperture and so obtain a diffraction limited focal volume. The excitation light enters the modified inverted epi-fluorescence microscope (Figure 4).

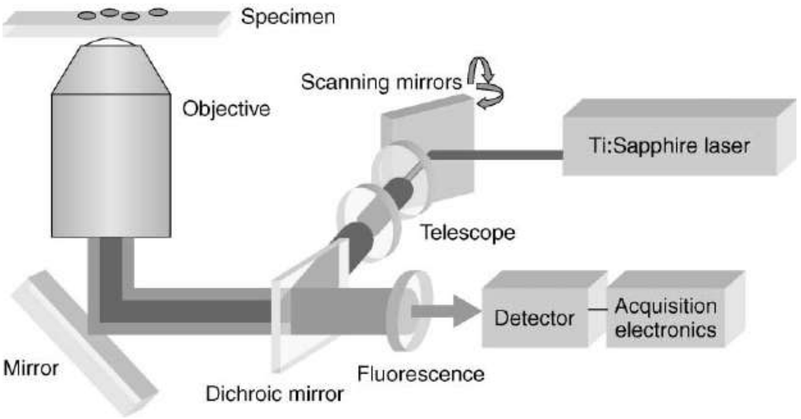

Figure 4 Schematic of a typical two-photon microscope.

A dichroic mirror transmits the excitation light and reflects fluorescence. High numerical aperture objectives are used to provide a high resolution but also maximize fluorescence collection. They are usually mounted on piezo-electric translators to provide a fast *Z*-scan. The fluorescence is collected by the same objective, reflected by the dichroic to the detector. A filter is placed in front of the detector to remove any residual excitation light. Typical detectors used are the photomultiplier tube (PMT), microchannel plate-PMT (MCP-PMT), avalanche photodiode or charge-coupled device (CCD) cameras. PMTs are the most commonly implemented because they are robust, low-cost, have large, active areas, and good sensitivity.

As TPE is intrinsically confined to the focal volume, two-photon microscopy unlike confocal microscopy, does not need emission pinholes and descanning optics to achieve axial depth discrimination Some authors have, however, added a pinhole to enhance resolution but at the cost of important sign al loss. Concerning the design of the detection pathway, depending on the size of the sample scanned, the objective magnification and the detector area, one may need to use a lens to collect the fluorescence onto the detector.

Fluorescence Detection

Contrast in fluorescence imaging can be provided through several complementary parameters such as steady-state intensity, spectral dependence (at absorption and emission), or fluorescence lifetime.

Intensity Imaging

Intensity images are obtained by summing all the incoming photons during an acquisition time (typically ~ 1 to 10 ms) which strongly depends on the fluorescence quantum yield of the sample. The whole samples intensity image is obtain by scanning the beam on the focal plane with the $X-Y$ scanner, leading to an acquisition time of 25 s for a 50×50 μm image (1 μm step).

Spectral Imaging

Spectral imaging can be a powerful way to discriminate different molecules, and to identify cellular components or perform co-localization of different proteins. It can be performed by either using an interference filter in front of the detector to extract the contribution in a specific band, or more precisely with a spectrometer. However, autofluorescent spectral bands are usually broad and sometimes it is difficult to separate the different emissions. Extrinsic trackers are thus used to labels specifically, for example, the nucleus, the plasma membrane, and mitochondria, and each tracker usually emits in a specific and well-known spectral band, allowing to easily discriminate and localize the different cellular components. Quantum dots offer narrow emission spectra and allow high spatial resolution, and with the rapid development of their functionalization, they should lead to rapid progress in intracellular imaging.

Fluorescence Lifetime Imaging (FLIM)

Imaging of fluorescence lifetimes is a useful contrast-enhancing method that provides information related to the microenvironment with spatial resolution. The fluorescence lifetime τ corresponds to the average time a fluorophore stays in its excited state is given by $\tau = 1/(k_{\mathrm{f}} + k_{\mathrm{nr}})$ with k_{f} the radiative decay and k_{nr} the nonradiative decay rate.

The radiative decay rate is a characteristic property of a molecule and therefore may be used to contrast different chemical species or fluorophores.

The nonradiative decay rate is a function of the local environment of an excited molecule, and so fluorescence lifetime measurements can provide information about environmental changes (e.g., pH, pO_2, $[Ca^{2+}]$). Furthermore, different cellular probes or tissue components with similar emission spectra but different lifetimes can be distinguished. Another advantage of lifetime measurement is that as a relative measurement of the fluorescence intensity, it does not depend on a non-uniform intensity excitation of the sample, and also does not depend on molecular concentration, allowing one to detect very low quantity of molecules.

Following a delta excitation, the fluorescence intensity de-excitation is given by$I(t) = I_0\, e^{(-t/\tau)}$, with I_0 the initial amplitude. Fluorescence lifetime measurement has been demonstrated either in frequency or time-domains. Time-resolved measurements in the frequency domain are based on the determination of demodulation M and phase shift $\Delta\varphi$ and require modulated light sources (modulation frequency $\omega = 2\pi f$).

The excitation source $I_E(t)$ and induced fluorescence $I_F(t)$ can be written as:

$$\begin{aligned} I_E(t) &= I_{E0} + I_{E\omega}\cos(\omega t + \varphi_E) \\ I_F(t) &= I_{F0} + I_{F\omega}\cos(\omega t + \varphi_F) \end{aligned} \qquad [5]$$

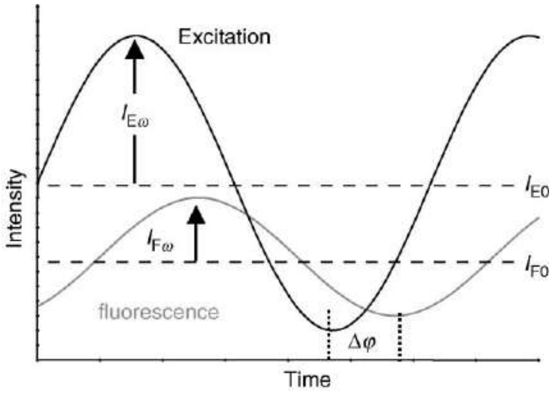

Figure 5 Time course of excitation and emission light in frequency domain fluorescence lifetime measurements techniques.

In the case of fluorescence with mono-exponential decay, the lifetime τ can deduced from the relations (Figure 5):

$$M = \frac{I_{F\omega}/I_{F0}}{I_{E\omega}/I_{E0}}. \qquad \omega\tau = \tan(\Delta\varphi) = \sqrt{\frac{1}{M^2} - 1} \qquad [6]$$

In order to translate the phase and demodulation information from the 80 MHz frequency to a lower kHz frequency, a modulated gain $G(t)$ at the frequency ω' can be applied to the detector (heterodyne detection):

$$G(t) = G_0 + G_{\omega'}\cos(\omega' t + \varphi_G) \qquad [7]$$

The detector response $S(t)$ is thus proportional to the fluorescence intensity:

$$\begin{aligned} S(t) = {} & G_0 I_{F0} + \frac{G\omega' I_{F\omega}}{2}\frac{1}{\sqrt{1+(\omega\tau)^2}} \\ & \times\cos[(\omega - \omega')t + \varphi_G - \varphi_E + \Delta\varphi] \end{aligned} \qquad [8]$$

In the time-domain, the single point detection system is called time-correlated single-photon counting (TCSPC). The principle of the TSCPC system is shown in Figure 6. For each excitation pulse a trigger signal is sent to a time-to-amplitude converter (TAC) within which a capacitor starts charging linearly with time. Upon arrival of each fluorescence photon on the detector a second trigger pulse stops the capacitor charging. The charge is proportional to the delay between the excitation pulse and the fluorescence photon. By repeating the measurement over many thousands of excitation pulse, a histogram of photon arrival can be build up. As each photon must be record individually, the incident power is reduce to avoid the detection of only photons from the beginning of the decay (pile-up) which would lead to shorter lifetimes. To match detector dead-time and TAC counting rate, the TAC is oftenly used in reverse mode: the fluorescence photon starts the TAC and the following pulse stops it.

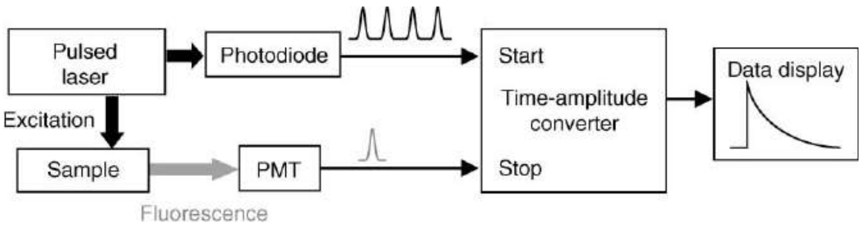

Figure 6 Schematic of time correlated single photon counting system.

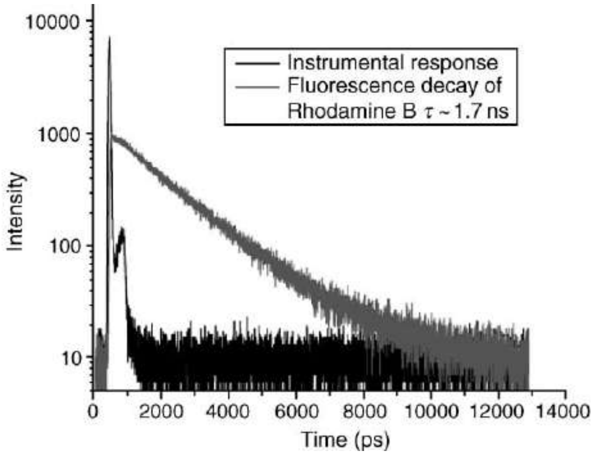

Figure 7 Typical instrumental response of MCP-PMT (black curve), and fluorescence decay curve of rhodamine B (grey curve) with $\tau \sim 1.7$ ns.

Typical detectors used are the photomultiplier tube (PMT), microchannel plate-PMT (MCP-PMT), and avalanche photodiode.

Typical instrumental response of our MCP-PMT is represented in Figure 7 (black curve) leading to a resolution of 30 ps. Depending on the fluorescence quantum yield of the sample, in order to have a good signal to noise ration, acquisition time is ranging between 1 and 30 s per point. Fluorescence decay of rhodamine B with a lifetime $\tau \sim 1.7$ ns is also presented in Figure 7.

For both detection approaches, single point lifetime measurements can be extended to lifetime imaging by synchronizing the sample scanning and the lifetime detection. However, due to the long time acquisition needed, images are only limited to small interesting areas rather than the whole sample. We will see in the recent development section below, new ways to overcome this issue.

Fluorescence Correlation Spectroscopy (FCS)

FCS experiments are performed by recording fluorescence intensity fluctuations (i.e., fluorescence bursts (Figure 8)) from the small focal volume (<1 fl). Applications include diffusion, chemical reactions, molecule concentration, or hydrodynamic flow–FCS under TPE benefits from all the advantages of TPE versus OPE.

Various experimental parameters can be extracted through temporal analysis of fluorescence intensity fluctuations by calculating the autocorrelation or cross-correlation of the fluorescence signal defined as:

$$g(\tau) = \frac{\langle I_F(t) I_F(t+\tau)\rangle}{\langle I_F^2(t)\rangle} = \frac{\int I_F(t) I_F(t+\tau)\,dt}{\int I_F^2(t)\,dt} \qquad [9]$$

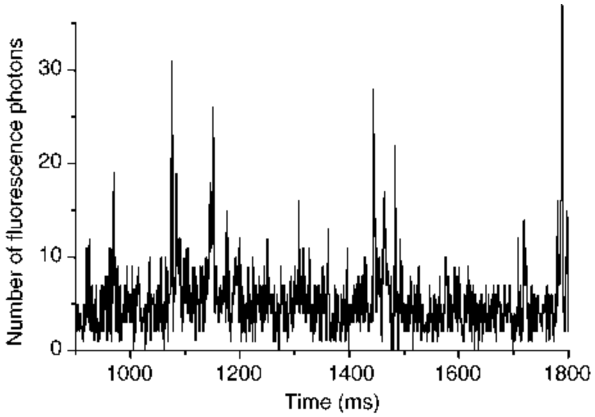

Figure 8 Fluorescence intensity raw data along the time. Bursts are observed when molecules cross the focal volume.

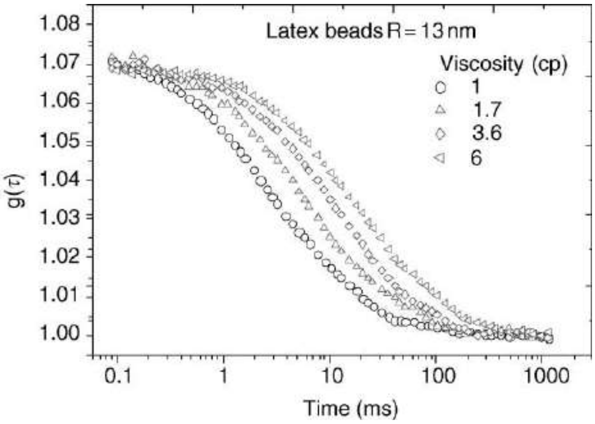

Figure 9 Correlation function observed for microspheres placed in different viscous solution.

where I_F is the time-dependent fluorescence signal. By considering that the fluorescence intensity fluctuates around a mean value:

$$I_F(t) = \langle I_F\rangle + \delta I_F(t) \qquad [10]$$

the autocorrelation function can be written as:

$$g(\tau) = 1 + \frac{\langle \delta I_F(t)\,\delta I_F(t+\tau)\rangle}{\langle I_F^2\rangle} \qquad [11]$$

FCS have been extensively used to extract diffusion coefficient of small fluorescent latex bead in solution, showing the influence of the bead radius or solvent properties. Figure 9 clearly exhibits the influence of the viscosity of the solvent on the diffusion of microspheres. Single molecule detection has also been reached. FCS as also been extended to biological applications where the low background associated with TPE and the reduced autofluorescence is critical. Diffusion of microbeads inside cells cytoplasm or bacterial biofilms have been successfully demonstrated.

Recent Developments in Two-Photon Microscopy

Towards video microscopy

All the nonlinear microscopies share an important limitation: the image build up pixel-by-pixel by scanning across the sample, leading to long-time acquisition. Depending on the sample's emission and the area studied, scanning rates between 0.5 Hz and 30 Hz can be reached for intensity images. To speed up this acquisition time, a first approach is based on a line excitation, in combination with widefield CCD detection. This provides real-time image acquisition rates at the expense of a reduced axial resolution (5 μm instead of 1 μm). A more ideal way of spreading the available excitation power over the sample and speed up image buildup, is the simultaneous use of several well separated foci, each of them generating a diffraction limited spot. In combination with whole field detector, as with a CCD camera, image acquisition rates can be strongly enhanced without compromising resolution. The spacing between the foci must be carefully chosen in order to avoid out-of-focus interference. By introducing time-delay between the different foci, larger than the pulse duration, each foci of a microlens array can be position very closed without loss of resolution. An alternative way has been proposed; a special beamsplitter has been developed to create multiple beams with a higher efficiency, furthermore this setup inherently induced time-delay between each foci. Time-domain has been demonstrated in both approach, and FLIM maps on cells imaging have been obtained in only few second with 200 ps time resolution.

Sharpen the focal point

Nonlinear microscopy facilitates 3D-imaging but the resolution is still limited to 180 nm in the focal plane and 500–800 nm along the optical axis. According to the diffraction theory, the resolution of a focusing light microscope is related to the size of its focal point. In the mid-1990s, new concepts appeared to break this diffraction theory and sharpen the focal volume. In 4-pi microscopy the sample is placed between two large NA objective lenses in a single axis at their common focus. Coherent interference of beams passing through the two lenses generate a point spread function whose central maximum is 3–4 times narrower than in the normal confocal case, but which also has side lobes. This last issue can be overcome using two-photon excitation. A two-photon 4-pi microscopy can thus result in a resolution of ~100 nm in all directions, and has been demonstrated successfully on filamentous actin and immunofluorescently labeled microtubules. However, a 4-pi microscopy requires the sample to be thin and mounted between two coverslips, unless one of the lenses is a dipping lens. Recent development on sample chambers allows one to overcome this issue, and results in live mammalian cells have been shown. A compact 4-pi commercial version has recently been proposed (Leica), achieving a seven-fold-improved axial resolution (80 nm) compared to the classical two-photon microscope.

Another technique, also developed by Professor Hell's group, is called stimulated emission depletion (STED) microscopy. The fluorophore in the fluorescent state S1 is stimulated to the ground state S0 with a doughnut-shaped beam. The depletion of the S1 saturates with increased I_{STED} intensity, and therefore establishes a nonlinear relationship between the fluorescence and the intensity applied to the STED. The saturation of S1 depletion is the essential element for breaking the diffraction limit as it confines fluorescence to the center. Spot size is not limited, but in practise depends on the quality and the saturation factor limited depletion. Using STED microscopy with wavelengths of 750–800 nm leads to a lateral resolution of up to 28 nm on single molecules experiments.

STED microscope has been successfully combined to 4-pi microscopy, and lateral resolution of 30–40 nm has been demonstrated.

Others Nonlinear Imaging Techniques

The success of two-photon microscopy has opened the way for other nonlinear microscopy. As opposite to multiphoton fluorescence absorption techniques where the induced signal is incoherent fluorescence, the signal generated by harmonic generation and stimulated Raman scattering is coherent with specific directional and polarization properties, as such fluorescence and harmonic images are derived from fundamentally different contrast mechanisms. However, both techniques share many features like intrinsic three-dimensionality, reduce photobleaching or deep sectioning, and require essentially identical equipment, allowing one to combine these two techniques as discussed further below.

Second-Harmonic Generation Microscopy

SHG is a second-order nonlinear optical process that can only arise from a medium lacking a center of symmetry, for example, an anisotropic crystal or at an interface of a membrane. SHG can be used to image tissue, muscles, cells membranes, or protein array, and has been recently reported to enhance membrane potential by using voltage sensitive dye.

SHG does not arise from an absorptive process, as the intense laser field induces a nonlinear, second-order, polarization in the assembly of molecules, resulting in the production of a coherent wave at exactly twice the incident frequency (half the wavelength). As represented in Figure 10, for an excitation at 800 nm, the SHG is observed at 400 nm and depending on the fluorescence properties of the molecule, a broadband fluorescent spectra is emitted in the visible.

Contrary to TPEF, where the emission properties depend only on the molecular properties and are independent of the laser source, the spectral and temporal characteristics of SHG are derived from the laser source. The SHG is temporally synchronous with the excitation pulse and its bandwidth is $1/\sqrt{2}$ of the laser's bandwidth. As SHG is an instantaneous process, the signal is only generated during the duration of the laser pulse (Figure 11), in contrary to TPEF where de-excitation of the molecules can occur over several nanoseconds after the pulse excitation.

A major difference between SHG and TPEF is that SHG requires a noncentrosymmetric environment to produce a signal. By examining the second term of eqn [1] $P = \chi^{(2)}E^2$ where the polarization and electric field are vector quantities and $\chi^{(2)}$ is a tensor, this condition is only satisfied for a birefringent crystal, an order array, or an interface. For example, the induced polarization from molecules in a dye cuvette (centrosymmetric sample) would be equal and opposite in all directions, leading to a vector sum zero. The SHG propagation is essentially in the forward direction, and the emission pattern is confined to twin lobes in the case of surface generation, or a cone in the case of volume generation. The off-axis of the two lobes critically depend on the tight focus of the driving field.

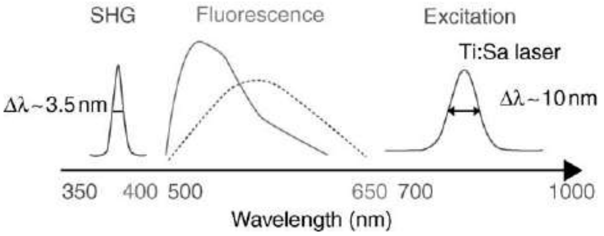

Figure 10 Wavelength domain for two-photon induced fluorescence and second-harmonic generation: excitation meets the same criteria and Ti: Sa are typically used in the range 700–1000 nm. For a typical excitation at 800 nm, fluorescence will be observed over ~450 nm, and SHG, at exactly 400 nm.

The total radiated SHG power can be derived by integrating the radiation profile from all the emission directions: $P_{\mathrm{SHG}} = \frac{1}{2}\theta_2 N^2 \sigma_2 I_0^2$ with $\sigma_2 \propto |\chi_2|^2$, I_0 the driving field intensity at the focal center, N the effective number of radiative molecules, and $\theta_2 \approx (0.1\lambda/w_0)^2$.

A schematic of a typical SHG/TPEF microscope is shown in Figure 12. The excitation source is a mode locked Ti:Sa laser which delivers ~120 fs at 76 MHz. The laser light is focused into the sample with a water immersion microscope objective (Olympus, LUMplanFl 60x) and the resultant SHG is collected in a forward direction, while TPEF is collected in the backward direction.

Figure 13 illustrates SHG and TPEF images of Di-6-ASPBS molecules under 880 nm excitation.

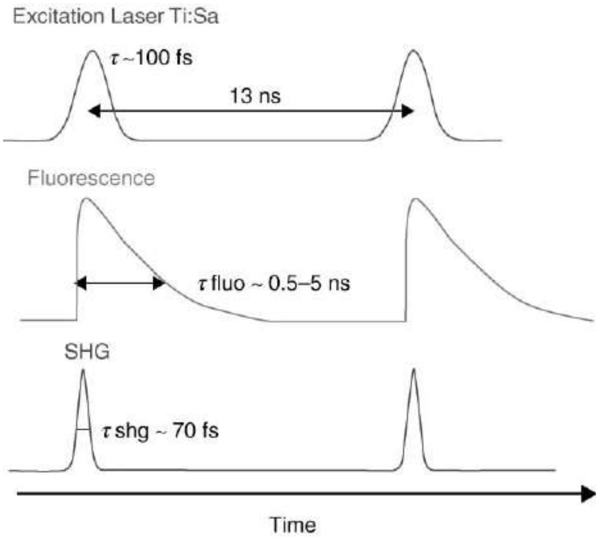

Figure 11 Typical time-response for two-photon induced fluorescence and SHG. SHG signal is an instantaneous process whereas fluorescence emission can occur several nanoseconds after the pulse excitation.

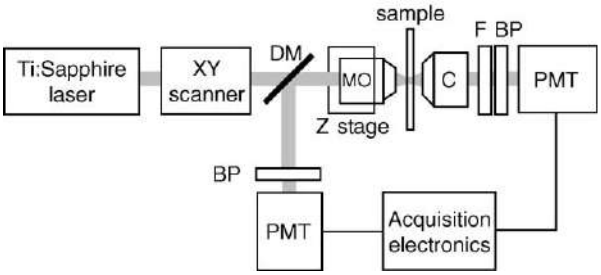

Figure 12 Schematic of SHG setup combined with TPEF extracted from (DM dichroic mirror, MO microscope objective NA 0.9 water immersion, C condenser, F colored glass filter, BP bandpass filter, PMT photomultiplier tube). Reproduced with permission from Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322, © 2000 Optical Society of America.

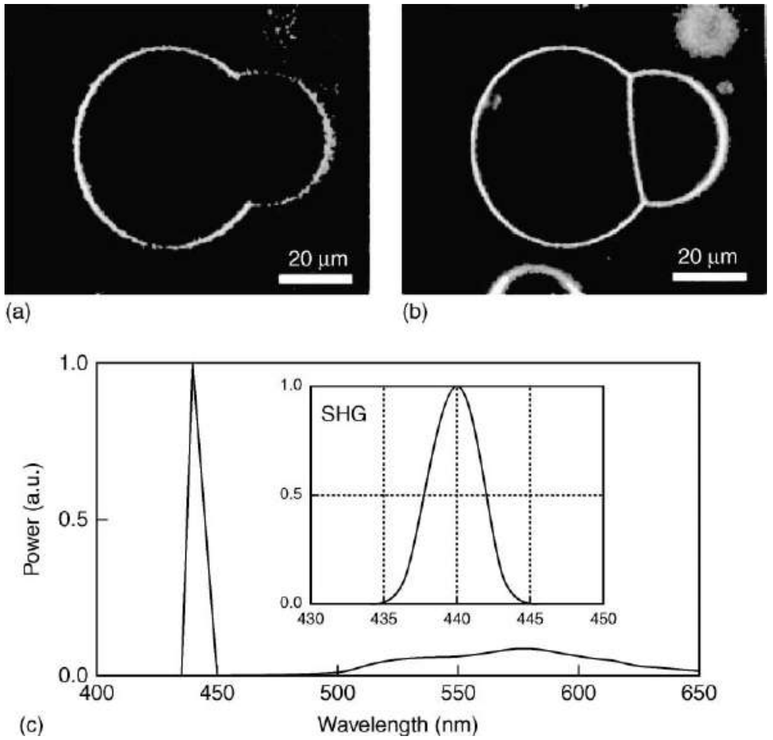

Figure 13 (a) SHG and (b) TPEF of two adhering vesicles labeled with Di-6-ASPBS (equatorial slice) excited at 880 nm. (c) The forward-detected SHG spectra (left peak) and the backward broad band fluorescence spectra (right peak). The Stokes shift in fluorescence leads to an important spectral separation between the two peaks. An expanded view of the SHG peak is presented in the inset. Reproduced with permission from Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322, © 2000 Optical Society of America.

The total acquisition time for the images was 1.5 s for an excitation power less than 1 mW. The adhesion area where the membranes are fused exhibits a centrosymmetric molecular distribution wherein TPEF is allowed, but not SHG. The corresponding forward-detected emission spectra of SHG (left peak) and TPEF (right peak) is shown in Figure 13c.

Third-Harmonic Generation (THG)

Third harmonic generation takes advantage of the discontinuity of third-order nonlinear susceptibility $\chi^{(3)}$ at material interfaces. At the focus point of an input laser beam with wavelength λ, this discontinuity gives rise to a new wave in the forward direction at the third-harmonic, $\lambda/3$. This signal can be analyzed in the transmission mode either with a single pixel detector or with a CCD camera. As with the other nonlinear techniques, THG offers a high 3D sectioning capability. Since all materials have nonvanishing third-order susceptibilities, THG microscopy can utilized as a general-purpose microscopy technique, with no need for fluorescence labeling or staining. Experimental setup for THG are very closed to SHG microscope. The need for efficient THG in the sample, together with low average laser power to avoid heating, dictates a laser source that operates with femtosecond pulses. The lasers that are often used THG microscopy are a synchronously pumped OPO at the wavelength of 1500 nm, a low repetition-rate OPA at 1200 nm and a Cr:Forsterite laser at 1240 nm. All these lasers are in the infrared above 1200 nm, thus generating THG signals in the visible range. These visible signals are compatible with the collection optics. A infrared laser undergoes low scattering, however absorption in water strongly increases.

THG have been demonstrated from subcellular to millimeter-size organisms with imaging of nerve cells, chloroplasts of a leaf, or Ca^{2+} intracellular dynamics.

Coherent Anti-Raman Stokes (CARS)

Chemical imaging by use of inherent molecular vibration signals avoids the photobleaching problem and perturbations to cell function induced by fluorophore labeling. Confocal Raman microscopy can provide 3D spatial resolution but it requires high average power due to small cross-section of Raman scattering, or long-time exposure which is not desirable in studying dynamic living system.

The above difficulties can be circumvented by multiphoton vibrational microscopy based on coherent anti-stokes Raman scattering (CARS). CARS is a four-wave mixing process in which a pump field $E_{\mathrm{p}}(\omega_{\mathrm{p}})$, a Stokes field E_{S} (ω_{S}), and a probe field $E_{\mathrm{p'}}$ $(\omega_{\mathrm{p'}})$. Interact with a sample and generate an anti-Stokes field E_{AS}, at the frequency $\omega_{\mathrm{AS}} = \omega_{\mathrm{p}} - \omega_{\mathrm{S}} + \omega_{\mathrm{p'}}$. In practise, the pump and the probe fields derive from the same laser $\omega_{\mathrm{p}} = \omega_{\mathrm{p'}}$ and $\omega_{\mathrm{AS}} = 2\omega_{\mathrm{p}} - \omega_{\mathrm{S}}$. The energy diagram of CARS is shown on Figure 14.

The CARS intensity is the square modulus of the induced polarization, $P^3 = \chi^{(3)}E_{\mathrm{p}}^2E_{\mathrm{s}}$, with $\chi^{(3)}$ the third-order susceptibility.

The application of the CARS process to microscopy opens up a new method for chemical imaging. In CARS microscopy, the signal is also dependent on the square of the pump field intensity, so the temporally and spatially overlapped pump and Stokes laser pulses tightly focused into a sample, generate a signal of a small excitation volume ($<1\ \mu\mathrm{m}^3$). As with all other nonlinear imaging techniques, the image is obtained by raster scanning the sample or the laser beams. The vibrational contrast in CARS microscopy arises from signal enhancement when $\omega_{\mathrm{p}} - \omega_{\mathrm{s}}$ is tuned to Raman-active vibrational frequency Ω. CARS signal generation needs to fulfill the phase-matching condition, $l < lc = \pi/|\Delta\mathbf{k}| = \pi/|\mathbf{k}_{\mathrm{as}} - (2\mathbf{k}_{\mathrm{p}} - \mathbf{k}_{\mathrm{s}})|$, where $\mathbf{k}_p$, $\mathbf{k}_{\mathrm{s}}$, and $\mathbf{k}_{\mathrm{as}}$ are wavevectors of the pump, Stoke and CARS fields, respectively, and $\Delta\mathbf{k}$ is the wavevector mismatch.

CARS presents several advantages similar to TPEF: high 3D sectioning capabilities, a low scattering and absorption of the excitation beams. CARS signal is higher in frequency than one-photon fluorescence, so it can be detected in the presence of a strong fluorescent background. Vibrational contrast is intrinsic to biological samples. Fluorescent probes are not necessary and photobleaching is thus circumvented. The coherent signal accumulation in the CARS process produces a strong and directional signal, making CARS microscopy much more sensitive than the conventional Raman microscopy. Consequently, CARS microscopy requires only a moderate average power tolerable by biological samples and has a high data acquisition speed. However, CARS microscopy is not without shortcomings. First, the intrinsically weak induced nonlinear polarizability requires sophisticated laser excitation sources with high peak power and moderate average power, and with low jitter. Second, unlike fluorescence imaging, CARS measurement is not background free, and the nonresonant signal limits the image contrast.

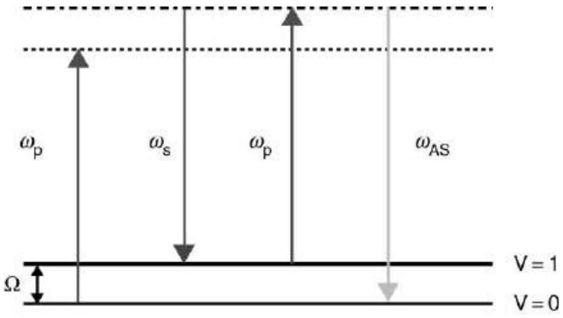

Figure 14 Energy diagram for Coherent anti-Raman spectroscopy (CARS).

The typical setup is very similar equipment compared to TPEF and SHG. The main difference arise from the excitation, two laser beams, the pump, and Stokes laser beams overlapping (in space and time) and are tightly focused into the sample using a high NA objective. The CARS signal can be detected in the forward (F-CARS) and the epi-detection (E-CARS). Unlike fluorescence emission and spontaneous Raman scattering, the radiation pattern of CARS is dependent on both the size and shape of a scatterer. The two detection geometry are thus complementary, the forward detection is suitable for imaging objects of a size comparable to or larger than the excitation wavelength.

For smaller objects, the F-CARS contrast is limited by the large nonresonant background from the solvent. Epi-detected CARS (E-CARS) microscopy provides a sensitive means of imaging objects having an axial length much smaller than the excitation wavelength because it avoids the large background from the solvent.

On the application side, CARS microscopy has been used for mapping the distribution, orientation, and diffusion of specific molecules in dynamic samples, for imaging lipids, proteins, and chromosomes in unstained live cells, and for monitoring cellular processes. CARS is a very promising technique to image small molecules such as lipids, hormones, and drug molecules, for which fluorescence labeling is prone to alter their function in cells and tissues. As for SHG, CARS microscopy is often combined with TPEF in order to obtain simultaneously the complementary contrasts.

Conclusion

Since its introduction more than a decade ago, two-photon fluorescence microscopy has quickly become a standard tool in biophotonic with applications ranging from the study of single molecules to thick tissue imaging. TPEF has opened the way to new coherent nonlinear imaging techniques such as SHG, THG, or CARS which all shared a high spatial resolution. Based on different contrasts these different approaches are often combined for a more complete diagnosis. With developments of new low cost, compact, and reliable lasers with tuning wavelength covering the spectra of a wide range of endogenous or extrinsic markers, nonlinear imaging will continue its successful implementation outside physics labs. Development of endoscopy systems in which ultrafast pulse duration is preserved, will open a new range of *in vivo* studies. Nonlinear imaging still has more to offer and will definitely be a powerful tool for biomedical applications.

See also

Microscopy: Overview. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources:** Harmonic Generation in Gases. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Cheng J and Xie XS (2004) Coherent anti-stokes Raman scattering microscopy: instrumentation, theory and applications. *Journal of Physical chemistry B* 108: 827–840.

Gratton E, Breusegem S, Sutin J, Ruan Q and Barry N (2003) Fluorescence lifetime imaging for the two-photon microscope: time-domain and frequency-domain methods. *Journal of Biomedical Optics* 8: 381–390.

Hell SW (2003) Towards fluorescence nanoscopy. *Nature Biotechnology* 21: 1347–1355.

Konig K (2000) Multiphoton microscopy in life sciences. *Journal of microscopy* 200: 83–104.

Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322.

Schwille P, Haupts U, Maiti S and Webb WW (1999) Molecular dynamics in living cells observed by fluorescence correlation spectroscopy with one- and two-photon excitation. *Biophysics Journal* 77: 2251–2265.

So PTC, Dong CY, Masters BR and Berland KM (2000) Two photon micrscopy excitation fluorescence microscopy. *Annual Review of Biomedical Engineering* 399–429.

Squier J and Muller M (2001) High resolution nonlinear microscopy: a review of sources and methods for achieving optimal imaging. *Review of Science Instruments* 72: 7.

Zipfel WR, Williams RM and Webb WW (2003) Nonlinear magic: multiphoton microscopy in the biosciences. *Nature Biotechnology* 21: 1369–1377.

Zipfel WR, Williams RM, Christie R, Nikitin AY, Hyman BT and Webb WW (2003) Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. *Proceedings of the National Academy of Science of USA* 100: 7075–7080.

Phase Contrast Microscopy

C J R Sheppard, National University of Singapore, Singapore

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many biological specimens behave as phase objects, that is, they alter the phase of the light transmitted through them, rather than its intensity. Similarly, reflection from a surface with height variations alters the phase of the reflected light. Observation of phase objects requires special imaging techniques. These include defocusing, dark field, Zernike phase contrast, Schlieren imaging, Hoffmann modulation contrast, differential phase contrast, interference microscopy, shearing interferometry, Nomarski differential interference microscopy, and digital phase retrieval. The imaging performance of these different methods are most easily appreciated and compared by consideration of the transfer function approach. In particular, we can consider weakly scattering objects and slowly varying phase objects as special cases.

When a thin unstained biological specimen is illuminated, in general the transmitted light suffers changes in amplitude, phase, and polarization. The change in polarization is related to the birefringence of the sample, which is neglected in the following. Thus the object (complex) amplitude transmittance can be written:

$$t(x, y) = a(x, y)e^{i\phi(x,y)} \qquad [1]$$

where $a(x, y)$ is real and represents the modulus (also called amplitude) of the light wave on exiting from the sample, and $\phi(x, y)$, also real, represents the phase. In a perfect microscope, we observe the intensity of the transmitted light, so the image is given by

$$I(x,y) = |a(x,y)e^{i\phi(x,y)}|^2 = a^2(x,y) \qquad [2]$$

Thus the phase information is invisible in the image. A perfect microscope does not image phase variations in the object, but results in an image that depends on the amplitude of the transmittance, called amplitude contrast. Usually the change in modulus is also small, so that the image will exhibit weak contrast, which may not be discernable. In order to render the phase variations visible it is necessary to modify the microscope in one of a number of different ways.

This conclusion is also relevant in incident light microscopy of surface structures. Then phase changes result from variations in surface height, and can be made visible by phase contrast microscopy. Quantitative measurement of the phase allows quantitative measurement of the surface profile. But care must be taken to avoid errors caused by phase change on reflection from a conducting material, or from effects of the numerical aperture of the optical system. In transmission microscopy quantitative phase measurement can be useful too, for example giving the concentration of a solute.

Some of these phase contrast methods are easily understood using the simple imaging model described above. Others are more easily appreciated by considering imaging in terms of Fourier optics, equivalent to Abbe's theory of microscope imaging. We introduce the angular spectrum of the object amplitude transmittance:

$$T(m,n) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t(x,y)\exp[-i2\pi(mx+ny)]\mathrm{d}x\,\mathrm{d}y \qquad [3]$$

where m,n are spatial frequencies in the x,y directions. We can write for the object amplitude transmittance:

$$t(x,y) = e^{b(x,y)} \qquad [4]$$

where $b(x,y)$ is in general complex. For a weak object we have

$$t(x,y) \approx 1 + b(x,y) \qquad [5]$$

and the spectrum is

$$T(m,n) = \delta(m)\delta(n) + B(m,n) \qquad [6]$$

where δ is a Dirac delta function and $|B| \ll 1$. The problem with imaging of phase arises because a change in the imaginary part of b produces only a small change in the modulus of t.

For a coherent imaging system, with coherent transfer function $c(m,n)$, the image intensity is

$$I(x,y) = \left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(m,n)T(m,n) \times \exp\left[\frac{i2\pi}{M}(mx+ny)\right]\mathrm{d}m\,\mathrm{d}n\right|^2 \qquad [7]$$

where M is the magnification. The coherent transfer function $c(m,n)$, is equal to the scaled pupil function $P(\lambda fm, \lambda fn)$, where f is the focal length of the objective. Substituting eqn [6] into eqn [7] and dropping terms of second order we obtain:

$$I(x,y) = |c(0,0)|^2 + 2\Re\left\{c^*(0,0)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \times c(m,n)B(m,n)\exp\left[\frac{i2\pi}{M}(mx+ny)\right]\mathrm{d}m\,\mathrm{d}n\right\} \qquad [8]$$

Normalizing the transfer function to unity at the origin:

$$I(x,y) = 1 + 2\Re\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(m,n)B(m,n) \times \exp\left[\frac{i2\pi}{M}(mx+ny)\right]\mathrm{d}m\,\mathrm{d}n\right\} \qquad [9]$$

For an axially symmetric aberration-free system, the coherent transfer function is real and even, so that a pure phase object is not imaged (only the Hermitian component of B is imaged). Many phase contrast methods in principle rely on appropriate modification of the coherent transfer function, either by making it complex or asymmetric.

However, this theory is an approximation as microscope systems are in practice partially coherent. For a partially coherent system the image intensity is

$$I(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(m,n;p,q)T(m,n) \times T^*(p,q)\exp\left\{\frac{i2\pi}{M}[(m-p)x+(n-q)y]\right\}\mathrm{d}m\,\mathrm{d}n \qquad [10]$$

where $C(m,n;p,q)$ is the partially coherent transfer function, sometimes called the transmission cross-coefficient, which can be calculated as an integral over the objective and condenser pupils. The partially coherent transfer function is nonzero over a finite region of $m,n;p,q$ space, thus limiting the resolution of the system. Then eqn [8] for a weak object becomes, normalizing so that $C(0,0;0,0)=1$:

$$I(x,y)=1+2\Re\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}C(m,n;0,0)B(m,n)\times\exp\left[\frac{i2\pi}{M}(mx+ny)\right]dm\,dn\right\} \quad [11]$$

where $C(m,n;0,0)$ is called the weak object transfer function.

For the image of a general object that is not necessarily weak, eqn [10] is often difficult to interpret. We can gain much insight into the imaging process by considering another special form of object, this time a slowly varying object:

$$t(x,y)=a(x,y)\exp\left[i\left(x\frac{\partial\phi}{\partial x}+y\frac{\partial\phi}{\partial y}\right)\right] \quad [12]$$

where a and ϕ are assumed slowly varying. Thus, the object spectrum, assumed now as slowly spatially varying, is

$$T(m,n)\approx a\delta\left(m-\frac{1}{2\pi}\frac{\partial\phi}{\partial x}\right)\delta\left(n-\frac{1}{2\pi}\frac{\partial\phi}{\partial y}\right) \quad [13]$$

Substituting in eqn [10], we then find:

$$I(x,y)=a^2(x,y)C(m,n;m,n)$$
$$m=\frac{1}{2\pi}\frac{\partial\phi}{\partial x},\ n=\frac{1}{2\pi}\frac{\partial\phi}{\partial y} \quad [14]$$

in which C is called the phase gradient transfer function, and is real. For an ordinary bright field system, C is an even function. Thus a phase gradient results in a change in image intensity, but for an axially symmetric system the sign of the change is independent of the sign of the phase change.

Defocused Imaging

Historically, the oldest method of phase contrast is simply to defocus the object. The transfer function is then complex, and phase detail is imaged. If the system is aberration-free, the contrast of the phase information reverses on the opposite side of focus. The strength of the contrast increases with defocus, until eventually artifacts are introduced when the imaginary part of the transfer function becomes negative. If the relative condenser aperture is too large, contrast is reduced.

Assume a pure weak phase object:

$$t(x,y)=1+id\cos(2\pi\nu x) \quad [15]$$

so that

$$T(m,n)=\left[\delta(m)+i\frac{d}{2}\delta(m-\nu)+i\frac{d}{2}\delta(m+\nu)\right]\delta(n) \quad [16]$$

Then:

$$I(x,y)=1-dC_i(\nu,0;0,0)\cos\left(\frac{2\pi\nu x}{M}\right) \quad [17]$$

where C_i is the imaginary part of C. Thus the phase information is imaged by the imaginary part of the weak object transfer function.

Dark Field Microscopy

Dark field microscopy can be used to visualize weak amplitude and phase information. The constant background in eqn [5] is eliminated, so we obtain an image of b. Note that eqns [8] and [11] are no longer valid because we cannot neglect the second-order term. For a coherent system, the background can be eliminated by use of a central stop to intercept the nondiffracted light. For a partially coherent microscope, an annular condenser aperture (larger in diameter than the objective pupil) is used (Figure 1). In both cases, in practice a range of low spatial frequencies are eliminated, resulting in the appearance of halo artifacts.

For the pure weak phase object of eqn [12], the image in a coherent microscope is

$$I(x,y)=\frac{1}{2}d^2c^2(\nu,0)\left[1+\cos\left(\frac{4\pi\nu x}{M}\right)\right] \quad [18]$$

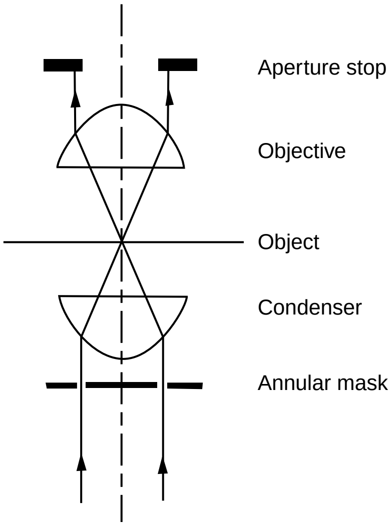

Figure 1 Schematic diagram of dark field microscopy.

so that the spatial frequency is double that in the original object. For the partially coherent microscope, we have

$$I(x,y) = \frac{1}{2}d^2\Big[C(\nu,\nu;0,0) + C(\nu,-\nu;0,0)\cos\left(\frac{4\pi\nu x}{M}\right)\Big] \qquad [19]$$

in which the two terms represent difference and sum frequency components. However, it can be shown that for the annular dark field system the transfer function for sum frequencies is zero, so the image intensity is a constant. The system images only difference frequencies of pairs of spatial frequencies, and not sum frequencies.

For a slowly varying phase object, zero intensity results from a region of constant phase, but the intensity rises quickly for small values of phase gradient, thus giving strong contrast. For larger phase gradients, the intensity falls off slowly.

Zernike Phase Contrast

The Zernike phase contrast method is similar to dark field except that the direct (nondiffracted) light has its relative phase changed by 90°, rather being eliminated. In practice this is achieved using an annular condenser aperture and an objective with a phase ring (Figure 2). The phase change can be ±90°, giving positive or negative phase contrast. The phase ring is usually only partially transmitting, which has the effect of enhancing the sensitivity. From eqns [8] or [11], imaging of a weak object is linear in phase. The halo artifact is present, as in dark field microscopy. Weak amplitude information is imaged as in dark field microscopy.

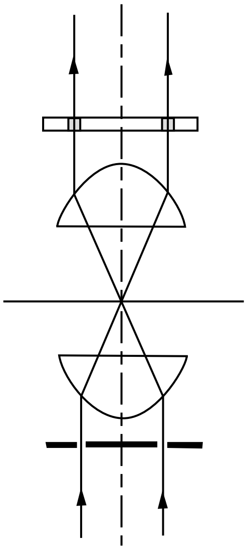

Figure 2 Schematic diagram of Zernike phase contrast microscopy.

For the weak phase object eqn [12], we then have

$$I(x,y) = 1 \pm \frac{2d}{g}C(\nu,0;0,0)\cos\left(\frac{2\pi\nu x}{M}\right) \qquad [20]$$

for positive or negative phase contrast, respectively, where g is the amplitude transmittance of the phase ring. The response of the weak object transfer function is improved as a result of the annular condenser aperture.

Schlieren Imaging, Hoffman Modulation Contrast, and Differential Phase Contrast (DPC)

In Schlieren imaging with coherent illumination, a half plane mask (a Foucault knife edge) is inserted to eliminate say the negative spatial frequencies in x. Taking now $c(0,0) = 1/2$, eqn [8] becomes, after normalization:

$$I(x,y) = 1 + 4\Re\left\{\int_{-\infty}^{\infty}\int_{0}^{\infty} c(m,n)B(m,n) \times \exp\left[\frac{i2\pi}{M}(mx+ny)\right]dm\,dn\right\} \qquad [21]$$

For the weak phase object eqn [15], we then have:

$$I(x,y) = 1 - dc(\nu)\sin\left(\frac{2\pi\nu x}{M}\right) \qquad [22]$$

The image exhibits a variation in phase quadrature to the original phase variation, that is, it is related to the x-derivative of the phase. The effect occurs because the coherent transfer function contains an odd (but real) part.

For a partially coherent microscope a similar effect can be achieved simply by offsetting the illumination system, but now the transfer function is no longer single-sided. It can be resolved into odd and even parts that result in differential phase contrast and amplitude contrast, respectively. The amplitude contrast component merely contributes to the background for a phase-only object, and can be removed simply by contrast enhancement of a digital image. The offset can be introduced in practice by inserting a half-plane mask into the condenser aperture stop. Imaging of a slowly varying object can be explained by eqn [14], in which the transfer function is not symmetrical. The image shows a pseudo-three-dimensional bas relief effect. A commercial implementation of an asymmetric transfer function is Hoffman modulation contrast, which uses offset illumination together with an amplitude mask in the objective pupil (Figure 3). Advantages of Hoffman contrast are that it uses standard bright-field objectives and avoids problems from birefringent specimens.

A similar method can be used in scanning laser microscopy. In a scanning microscope the effect of the illumination and detection systems are interchanged. Thus DPC can be achieved in a scanning laser microscope by offsetting the detector. In practice this is conveniently done by using a split detector (Figure 4), a detector divided into two semicircular elements. If the signal of one half is subtracted from that of the other, the amplitude contrast component cancels out, so that pure differential phase contrast results. For weak phase gradients, the intensity varies linearly with phase gradient, again showing the bas relief effect. It is thus bi-directional, and the difference signal can become negative. A constant value is, therefore, usually added to it before display. The sum of the two detector elements results in a conventional bright-field image. From eqn [14], both difference and sum signals are proportional to $a^2(x, y)$, so that if the ratio of these two signals is extracted, an image of the phase gradient in the x direction results. The response to either weak phase gradients or fine detail can be enhanced by modifications using annular split detectors. Use of a quadrant detector allows the phase gradients in the x and y directions to be imaged. DPC microscopy has been demonstrated to be sensitive enough to observe monomolecular films and atomic surface steps.

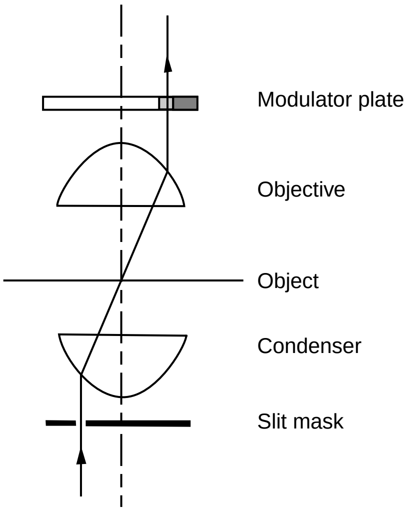

Figure 3 Schematic diagram of Hoffmann modulation contrast microscopy.

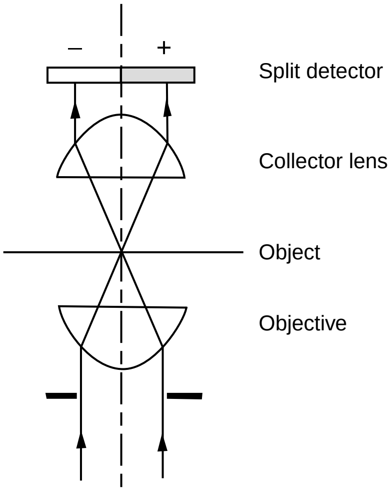

Figure 4 Schematic diagram of differential phase contrast (DPC) microscopy.

Interference Microscopy

Interference microscopy allows quantitative measurement of the object phase by interference of the object beam with a reference beam. In the transmission geometry, a Mach–Zehnder interferometer is used, which is effectively two parallel microscopes, each with matched condenser and objective lenses. In the reflection geometry we can use a Michelson, a Mirau, or a Linnik interferometer. The Michelson interferometer has an inclined beamsplitter and the reference mirror situated between the objective and object, and hence can be used only with low numerical aperture objectives. The Linnik interferometer has matched objectives in the object and reference beam paths. There is no limit to the numerical aperture that can be used, but it is not a common-path interferometer and is hence sensitive to vibrations and air currents. Probably the most practicable system for interference microscopy in the reflection geometry is based on the Mirau interferometer that uses a beamsplitter with its normal parallel to the optic axis. Until recently, interference microscopes used a small condenser aperture giving nearly coherent illumination. The image consists of three components: an object beam term (noninterference image) a reference beam term, and the interference term. The phase (and amplitude) of the object beam can be extracted using phase stepping or heterodyning techniques. The measured phase is wrapped, and requires unwrapping to obtain the absolute phase.

If the condenser aperture is opened up, the behavior is modified. The interference term is

given by

$$I_{\mathrm{int}}(x,y) = 2\Re\left\{ r^* \int\!\!\int c_{\mathrm{int}}(m,n) T(m,n) \times \exp\left[\frac{\mathrm{i}2\pi(mx+ny)}{M} \right] \mathrm{d}m\, \mathrm{d}n \right\} \quad [23]$$

where $c_{\mathrm{int}}(m,n)$ is the coherent transfer function for interference imaging, given by a convolution integral of the object path objective pupil with the product of the reference pupil and the source. The first thing to note is that for a high numerical aperture, the spacing of the interference fringes is increased by 20–40%, and depends on defocus, tilt, and central obscuration. Thus accurate measurement of phase requires careful calibration. The objective pupil exhibits a phase variation when defocused that reduces the modulus of the coherent transfer function, resulting in the fringes being modulated by an envelope the width of which becomes smaller as the aperture becomes larger. The envelope is similar to that of the optical sectioning effect in confocal microscopy. This envelope allows the zero-order fringe to be identified, although simulations have suggested that fringe hopping could occur at phase jumps as a result of diffraction effects.

Multiple Beam Interferometry

The topography of a surface can be investigated by using a Fizeau interferometer formed by bringing a coated optical flat near to the surface. Then if the surface and flat have high reflectivity, multiple beam interference occurs and the width of the bright fringes can be much reduced. The system can be used with a tilted flat, to give fringes that are substantially straight, or with an almost parallel flat, so that the entire field of view is occupied by a single fringe. In the latter case, the sensitivity can be in the subnanometer range. Multiple beam interference also enhances the contrast of weak amplitude information.

Shearing Interferometry

In conventional interferometry, the object beam is arranged to interfere with a constant reference beam. In shearing interferometry, on the other hand, it interferes with a shifted version of itself.

Lateral Shearing Interferometry

An approach that is suitable for quantitative optical path measurements is to use an interferometer that shears the microscope image. The Interphako system uses a Mach–Zehnder interferometer that can measure optical path differences as low as ~3 nm in either the reflection or transmission geometry. The interferometer allows the shear distance to be varied.

Differential Interference Microscopy (DIC)

The most common form of shearing interferometry is differential interference microscopy, usually associated with the name of Nomarski. DIC microscopy can be performed in either the transmission or the reflection geometry. A plane polarized illumination beam is split into two orthogonally polarized beams traveling at slightly differing angles by a Wollaston prism in the front focal plane of the condenser lens (Figure 5). The beams are thus sheared laterally in the object plane. The light from the sample travels through the objective and a second Wollaston prism, which recombines the two beams into one. In the reflection geometry, a single lens and Wollaston prism are used. An analyzer set at 45° allows the two beams to interfere. The shear is arranged to be smaller than the width of the point spread function. Different prisms are supplied for use with different magnification objectives. The system results in

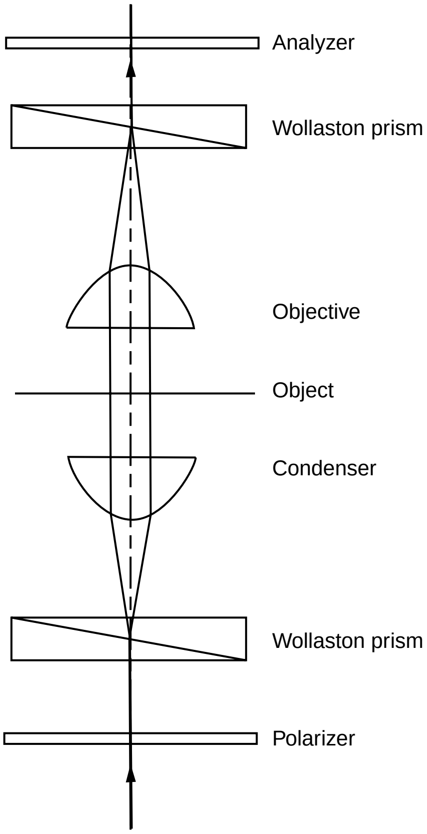

Figure 5 Nomarski differential interference contrast (DIC) microscopy.

differential phase contrast, but there are differences as compared with the DPC microscope described above.

There are two different implementations of DIC in use. In one implementation, the illumination polarizer and analyzer are crossed. The relative phase difference between the two beams (the bias retardation) is adjusted by moving the Wollaston prism laterally. It is claimed that rotation of the analyzer can be useful when observing mixed phase/amplitude objects. In the second arrangement (the Sénarmont compensator) the position of the Wollaston prism is fixed, and the analyzer rotated to change the relative phase.

A recent variation of DIC uses circularly polarized illumination. The Wollaston prism still produces two plane polarized components, but with an additional phase difference of $\pi/2$. The advantage of using circular polarization is that the prism can then be rotated to change the direction of shear.

If the shear is 2Δ, and the phase difference is $\pi - 2\phi_0$ (i.e., if $\phi_0 = 0$ then the two beams are in anti-phase), then the effective object spectrum is:

$$T_{\mathrm{eff}}(m,n) = -2\sin(2\pi m\Delta - \phi_0)T(m,n) \qquad [24]$$

Thus the effective transfer function can be written:

$$\begin{aligned} C_{\mathrm{eff}}(m,n;p,q) &= 4\sin(2\pi m\Delta - \phi_0)\sin(2\pi p\Delta - \phi_0) \\ &\quad \times C(m,n;p,q) \\ &= 2\{\cos[2\pi(m-p)\Delta] \\ &\quad - \cos 2\phi_0 \cos[2\pi(m+p)\Delta] \\ &\quad - \sin 2\phi_0 \sin[2\pi(m+p)\Delta]\} \\ &\quad \times C(m,n;p,q) \qquad [25] \end{aligned}$$

It is the final term of the three in braces that, because it is an odd function, results in differential phase contrast. We can now examine imaging of special forms of object. For a weak object, putting $p = 0$ as in eqn [11]:

$$\begin{aligned} C_{\mathrm{eff}}(m,n;0,0) = 2\{&(1-\cos 2\phi_0)\cos[2\pi m\Delta] \\ &- \sin 2\phi_0 \sin[2\pi m\Delta]\}C(m,n;0,0) \end{aligned} \qquad [26]$$

The first term in braces corresponds to an amplitude contrast background component because it is real and even. The second term results in DPC, because it is real and odd. For zero bias retardation, dark-field imaging results. As the bias retardation is increased from zero, the strength of both the amplitude contrast and the DPC components increase, but that of the DPC component increases more rapidly. For visual observation we require good contrast, and hence we choose a small value of bias retardation, just large enough so that the background is large enough to avoid artifacts in the DPC image. In digital microscopy, contrast is not so important as it can be enhanced digitally, and hence we can increase ϕ_0 to a value of $\pi/4$, which maximizes the strength of the DPC component. Further increase in bias retardation decreases the DPC signal, until for $\pi/2$ pure amplitude contrast results. The actual form of the weak object transfer function depends on the magnitude of the shear. The condition $m_0\Delta = 1/4$, where m_0 is the cut-off in spatial frequency, gives optimum performance for the imaging of weak objects.

The phase gradient transfer function is:

$$C_{\mathrm{eff}}(m,n;m,n) = 4\sin^2(2\pi m\Delta - \phi_0)C(m,n;m,n) \qquad [27]$$

For the bright-field microscope the phase gradient transfer function falls off symmetrically from its value at zero phase gradient. The effect of the premultiplying factor in eqn [27] is to increase the fall-off for positive phase gradients and reduce it for negative ones. A bas-relief effect occurs, with highlighting for negative phase gradients for some values of the parameters. Increasing the value of the shear increases the highlighting effect, but, unlike in DPC, the behavior is not anti-symmetric.

Because interference is dependent on the wavelength of the light, useful and beautiful effects can be seen with DIC microscopy using white light. Usually it is regarded as a qualitative, rather than a quantitative, imaging method. However, as from eqn [14], the intensity can be written:

$$I(x,y) = 2a^2(x,y)[1 - \cos(2\pi m\Delta - \phi_0)]C(m,n;m,n) \qquad [28]$$

By recording images at different bias retardations, conventional phase stepping methods can be used to extract the phase gradient $2\pi m = \partial\phi/\partial x$.

Axial Shearing Interferometry

Instead of shearing in the transverse direction, a shear in the axial direction can be used. In the reflection geometry, the complex amplitude from the point of observation is compared with the average over a defocused spot. This method has proved to be extremely sensitive, and can detect height changes smaller than 0.1 nm.

Phase Retrieval

Digital Phase Retrieval

The phase of a single two-dimensional (2D) measurement of intensity can be retrieved by an iterative procedure, if constraints such as nonnegativity and band-limitedness can be assumed. If the modulus of both an object and its Fourier transform are known, then the phases can also be recovered. In this case, the two 2D real data sets have information content sufficient to reconstruct the 2D complex object. This approach has been extended to the case of measurement of two defocused images, using the method of phase diversity.

Imaging Using the Transport of Intensity Equation

The transport equation of intensity can be used to recover the phase of a partially coherent wavefield directly (i.e., noniteratively) from two 2D intensity measurements. If the wavefield is represented in terms of its intensity I and phase ϕ in a plane, the transport equation can be written in the paraxial approximation and in the absence of phase singularities:

$$-k\frac{\partial I}{\partial z} = \nabla_T \cdot [I \nabla_T \phi] \qquad [29]$$

where ∇_T is a 2D operator acting in the plane. This expression can be inverted to give ϕ quantitatively, if I and $\partial I/\partial z$ are known. However, it gives the phase of the image field, rather than the phase of the object, so that obtaining quantitative values for a strong object may be difficult. The transport of intensity method is valid with partially coherent illumination: if the relative condenser aperture is too great then contrast becomes weak. Further, an effect similar to the variation in fringe spacing with aperture in interference microscopy is introduced because oblique illuminating rays traverse the sample at an angle to the axis. In practice, a series of three images are usually taken at different focus positions. Then, from the quantitative phase data, various different contrast mechanisms such as DIC, Zernike phase contrast, and so on, can be simulated.

Surface Profiling Methods: Triangulation and Coherence Domain Methods

Although strictly they are not phase imaging methods, other surface profiling methods can be used to measure optical path. These could also be adapted to measure optical path in transmission. It is important also to be aware of these methods, as they could affect measurements made using other phase contrast techniques.

Apart from the phase measurement techniques described above, surface profiling methods fall into one of two categories: triangulation and coherence domain methods. An example of a technique that relies on triangulation is confocal microscopy. Confocal microscopy exhibits an optical sectioning effect similar to that described for interference microscopy. The effect is stronger for higher numerical aperture. As a reflector is scanned through focus the intensity goes through a maximum. By locating the position of the maximum, the local surface height can be obtained. In a similar way, in a confocal transmission system, introduction of an object with some optical thickness has the effect of altering the axial position of the confocal pinhole, with a resultant drop in intensity. This effect can be used to image the optical thickness of the sample.

Coherence domain methods rely on the finite coherence length of the light. In an interferometer, this results in an envelope that multiplies that arising from the nonzero relative condenser aperture. Measurement of the peak of the envelope allows local surface height to be measured. Unlike the confocal approach, the width of this coherence envelope is independent of the numerical aperture of the system.

In reflection, both the confocal and the coherence domain methods have the advantage over direct phase imaging techniques that they allow measurement of surface height changes without errors caused by the phase change on reflection from a conducting medium.

See also

Interferometry: Phase Measurement Interferometry. **Microscopy:** Interference Microscopy. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Bennett AH, Jupnik H, Osterberg H and Richards OW (1952) *Phase Microscopy*. New York: Wiley.

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Bradbury S and Evenett PJ (1996) *Contrast Techniques in Light Microscopy*. Oxford: Bios Scientific Publishers.

Françon M (1950) *Le Contraste de Phase en Optique et en Microscopie*. Paris: Revue d'Optique.

Françon M (ed.) (1952) *Le Contraste de Phase et le Contraste par Interférences*. Paris: Revue d'Optique.

Inoue S and Spring KR (1997) *Video Microscopy: The Fundamentals*, 2nd edn. New York: Plenum.

Lessor DL, Hartman JS and Gordon RL (1979) Quantitative surface topography determination by Nomarski reflection microscopy. *Journal of the Optical Society of America* 69: 357–366.

Pluta M (1989) *Advanced Light Microscopy.* Specialized Methods, vol. 2. Amsterdam: Elsevier.
Sheppard CJR and Wilson T (1980) Fourier imaging of phase information in conventional and scanning microscopes. *Philosophical Transactions of Royal Society A* 295: 513–536.
Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy.* London: Academic Press.

MODULATORS

Contents

Acousto-Optics
Electro-Optics
Modulation and Demodulation of Optical Signals

Acousto-Optics

M Gottlieb and D Suhre, Carnegie Mellon University, Pittsburgh, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Acousto-optics deals with the interaction of light waves with sound waves and has given rise to a large number of devices related to various laser systems for display, information handling, optical signal processing, and numerous other applications requiring the spatial or temporal modulation of coherent light. The basic principles and phenomena controlling these interactions were largely understood by the mid-1930s, but had little practical significance because very high acoustic power levels were required to attain good optical efficiency and there were few good sources of well-collimated monochromatic light. During the period from 1960 to the present, several key technologies were developing rapidly, at the same time that many applications of the laser were being suggested which require high-speed, high-resolution scanning methods. These new technologies gave rise to high-efficiency, wideband acoustic transducers capable of operation to several gigahertz, high-power wideband solid-state amplifiers to drive such transducers, and the development of a number of new, synthetic acousto-optic crystals with low-drive-power requirements and low acoustic losses at high frequencies. This combination of properties makes acousto-optics feasible for many systems, and for several is the method of choice to satisfy demanding requirements. This chapter will summarize the basic features of acousto-optic interactions and the operating principles of the most common acousto-optic devices.

The Photo-Elastic Effect

The change induced in the refractive index of a transparent material by the pressure change produced by an acoustic disturbance is the underlying mechanism of all acousto-optic interactions. An acoustic wave produces periodic regions of compression and rarefaction in the material, which modulates the density. The Lorentz–Lorenz relation relates the refractive index to the density, for the simplest case of an ideal gas

$$\frac{(n^2-1)}{(n^2+2)} \propto \rho \qquad [1]$$

where n is the refractive index and ρ is the density. This relation is followed to a good approximation for most simple solid materials as well. The elastooptic coefficient is

$$\frac{\rho\,\mathrm{d}n}{\mathrm{d}\rho} = \frac{(n^2-1)/(n^2+2)}{6n} = p \qquad [2]$$

The fundamental quantity given by eqn [2], also known as the photo-elastic constant p, is related to the pressure applied by

$$p = \frac{1}{\beta}\left(\frac{\mathrm{d}n}{\mathrm{d}p}\right) \qquad [3]$$

where P is the applied pressure and β is the compressibility of the material. The photo-elastic constant of an ideal material with refractive index of 1.5 is 0.59. The photo-elastic constants of a wide variety of materials lie in the range from about 0.1 to 0.6.

The relation in eqn [3] follows from the usual definition of the photo-elastic constant:

$$\Delta(1/\varepsilon) = \Delta(1/n^2) = pe \qquad [4]$$

where ε is the dielectric constant, $\varepsilon = n^2$, and e is the acoustic strain amplitude. The induced change in refractive index, Δn, is

$$\Delta n = -n^3 pe \qquad [5]$$

Strain amplitudes typically lie in the range of 10^{-8} to 10^{-5}, with Δn in the range of about 10^{-8} to 10^{-5} (for $n = 1.5$). Devices based upon such a small change in refractive index are capable of generating large effects because these devices are configured in a way that can produce large phase changes at optical wavelengths.

The more complete relation defining the photo-elastic interaction is more complex than the simple scalar eqn [5] in which the photoelastic constant is independent of the directional properties of the material. In fact, even for an isotropic material such as glass, longitudinal acoustic waves and transverse (shear) acoustic waves cause the photoelastic interaction to assume different parameters. A tensor relation between the dielectric properties, the elastic strain, and the photo-elastic coefficient describes the interaction, particularly for anisotropic materials. The tensor equation is

$$\Delta(1/n^2) = \sum_{kl} p_{ijkl}\, e_{kl} = \Delta(1/\varepsilon)_{ij} \qquad [6]$$

where $(1/\varepsilon)_{ij}$ is a component of the optical index ellipsoid, e_{kl} are the Cartesian strain components, and p_{ijkl} are the components of the photo-elastic tensor. The crystal symmetry of the photoelastic material places limits on the possible configurations of interaction geometry.

Diffraction by Acoustic Waves

For typical acousto-optic devices the acoustic wave acts as a diffraction grating, made up of periodic changes in optical phase, moving at sound velocity. These features determine the properties of acousto-optic diffraction. Using a quantum-based model, the light and sound may be thought of as particles, photons and phonons, undergoing collisions in which energy and momentum are conserved. Either of these descriptions may be used to obtain all the important diffraction effects, but some are more easily understood on the basis of one or the other. A description of both is given here. Consider Figure 1 in which the light wave, of frequency ω and wavelength λ, is incident into the material with an acoustic wave of frequency f and wavelength Λ. If the refractive index of the medium is $n + \Delta n$ in the presence of the acoustic wave, the phase of the optical wave will be changed by an amount

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \qquad [7]$$

where L is the length of the interaction region. Some typical values of $\Delta\phi$ can be obtained by assuming $L = 2.5$ cm and $\lambda = 0.5$ μm, with Δn reaching a peak value of 10^{-5}. This yields a phase change of π rad, which is, of course, quite large. It is large because L/λ, the number of optical wavelengths, is 50 000, so that a very small Δn can still produce a sizable $\Delta\phi$. If the electric field incident on the delay line is represented by

$$E = E_0\, e^{i\omega t} \qquad [8]$$

then the field of the phase-modulated emerging light will be

$$E = E_0\, e^{(i\omega t + \Delta\phi)} = e^{i\omega t}\, e^{2\pi i (L/\lambda)\sin(ft)} \qquad [9]$$

where f is the acoustic frequency.

A detailed derivation of the resulting temporal and spatial distribution of the light field is mathematically complex, but, analogy with radio-wave modulation can be used to arrive at the resultant fields. The spectrum of a phase-modulated carrier of frequency ω consists of components separated by multiples of the modulation frequency f, as shown in Figure 2. Sidebands are produced about the carrier frequency, such that the frequency of the mth sideband is

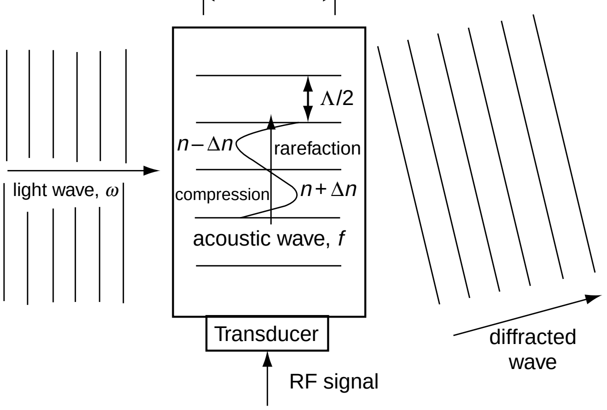

Figure 1 Interaction of light waves with acoustic waves.

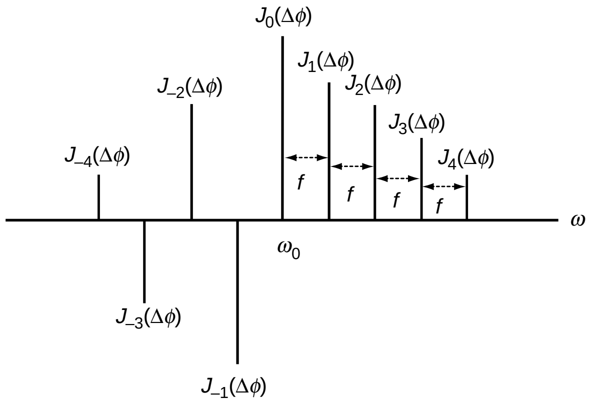

Figure 2 Spectrum of a phase-modulated wave of carrier frequency ω_0, and modulation index $\Delta\phi$.

$\omega + mf$, where m may be positive or negative. The amplitude of each of the sidebands is proportional to the Bessel function of order equal to the sideband number, and whose argument is the modulation index $\Delta\phi$. The odd-numbered negative orders are 180° out of phase with the even-numbered ones, a feature that may be useful for certain signal processing applications. The light emerging from the delay line is composed of a number of light waves whose frequencies have been shifted by mf from the frequency ω of the incident light. The relative amplitudes will be determined by the peak change in the refractive index. The two extremes of the diffraction process are simplified in the 'thin grating' and the 'thick grating' models. The ratio of the interaction length, L, to the acoustic wavelength, Λ, will determine the character of the diffraction process. The plane wave approximation in which the acoustic energy propagates as a plane wave is valid when this ratio is very large. However, when this ratio is small, the acoustic propagation can be described in terms of a sum of plane waves, the angular spectrum of which increases as the ratio decreases. In this phenomenological model the partial wave which is propagating at an angle λ/Λ to the forward direction may diffract the light second time into an angle $2\theta = 2\lambda/\Lambda$, and the frequency of this light will once again be up-shifted, for a total frequency shift of $2f$. If the spectrum of acoustic waves contains sufficient power of still higher orders, then this rediffraction process can be repeated, so that light will be multiply diffracted m times into higher-order angles, $m\theta = m\lambda/\Lambda$ each with a frequency shift mf. A similar argument holds for the negative orders, so that a complete set of diffracted light beams will appear as shown in Figure 3, where the deflection angle corresponding to the mth order is given by $\sin\theta_m = m\lambda/\Lambda$ and the frequency of the light deflected into the mth order is $\omega_m = \omega + mf$. The intensity of the carrier wave, or zero order, will be zero when the modulation index $\Delta\phi$ is equal to 2.4. The first order will be a maximum for $\Delta\phi = 1.8$, corresponding to the maximum of the first-order Bessel function B_1, and decreasing for larger modulation indices. These phenomena were described by Debye and Sears and so are often referred to as Debye–Sears diffraction. An extensive theoretical analysis of the effect was given by Raman and Nath and is alternatively referred to as Raman–Nath diffraction. As the interaction length is increased the Raman–Nath diffraction gradually weakens. The weakening begins around an interaction length $L \sim \Lambda^2/4\lambda$. This value of L is expressed in the Q parameter

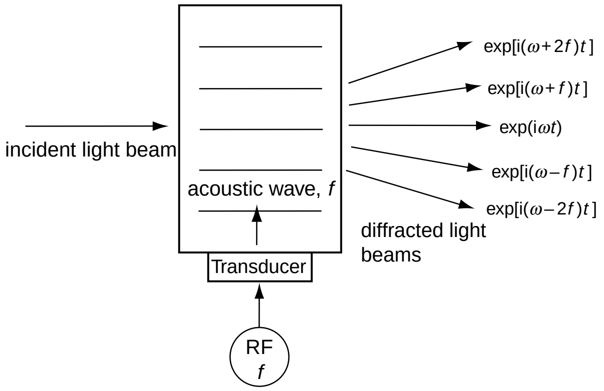

Figure 3 Diffraction of light in the Raman–Nath limit.

$$Q = 4L\lambda/\Lambda^2 \qquad [10]$$

which is known as the Raman–Nath parameter. A different regime of diffraction takes effect for values of the interaction length $Q \gg 1$.

The phenomena in this regime can be more easily understood in terms of the quantum description of the light and sound waves as collisions between photons and phonons. In this model, the dynamics of the collisions of the light and sound are governed by the laws of conservation of energy and momentum. The momentum vectors magnitudes of the light and sound, **k** and **K**, are given by the well-known expressions

$$|\mathbf{k}| = 2\pi n/\lambda \text{ and } |\mathbf{K}| = 2\pi/\Lambda \qquad [11]$$

where the acoustic wavelength, Λ, is related to the acoustic velocity, V, by $\Lambda = V/f$.

Conservation of momentum is expressed by the vector relation

$$\mathbf{k}_\mathrm{i} + \mathbf{K} = \mathbf{k}_\mathrm{d} \qquad [12]$$

the diagram for which is shown in Figure 3, where $\mathbf{k}_i$, and $\mathbf{k}_d$ represent the momentum of the incident photon and the diffracted photon, respectively. In this notation

$$\mathbf{k}_i = 2\pi n_i/\lambda, \qquad [13a]$$

and

$$\mathbf{k}_d = 2\pi n_d/\lambda \qquad [13b]$$

If the material is anisotropic (birefringent), n_i and n_d may be different.

Conservation of energy requires that

$$h\omega_d = h\omega_i + hf \qquad [14]$$

in which h is Planck's constant. Since ω_i lies in the optical frequency range, $\sim 10^{13}$ Hz, and f lies in the RF or microwave range, 10^6–10^9 Hz, then $\omega_d \sim \omega_i$. This results, for isotropic materials, in the magnitudes of $\mathbf{k}_i$ and $\mathbf{k}_d$ being equal and resulting in the isosceles momentum triangle of Figure 4. The angles of incidence and diffraction (with respect to the planar acoustic wavefronts), called the Bragg angles, are equal for this case, and are

$$\theta_B = (1/2)\lambda/\Lambda \qquad [15]$$

A schematic diagram for this process is shown in Figure 5. The interaction will be large only if the light is well-collimated, and incident at this angle. Unlike the Debye–Sears regime, there is only a single diffracted beam. The energy of the phonon may either be added to that of the incident photon, increasing its frequency to $\omega_d = \omega_i + f$, or the reverse, resulting in $\omega_d = \omega_i - f$. The sense of the momentum vectors determines which of these occurs. The Debye–Sears effect and Bragg diffraction are not different phenomena, but are the limits of the same mechanism. The Raman–Nath parameter Q determines which is the appropriate limit for a given set of values λ, Λ, and L.

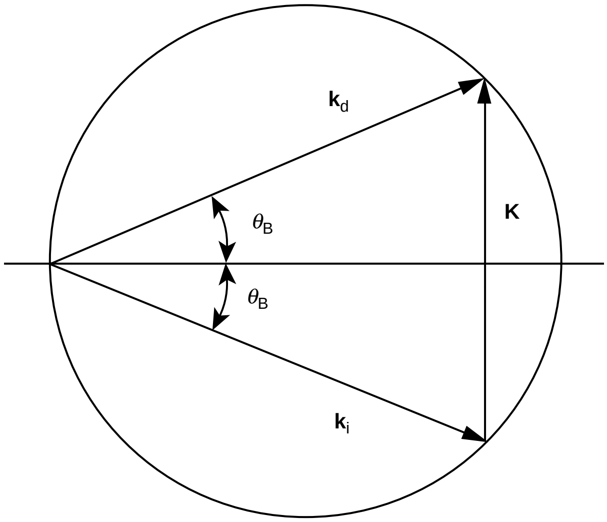

Figure 4 Momentum diagram for diffraction of light.

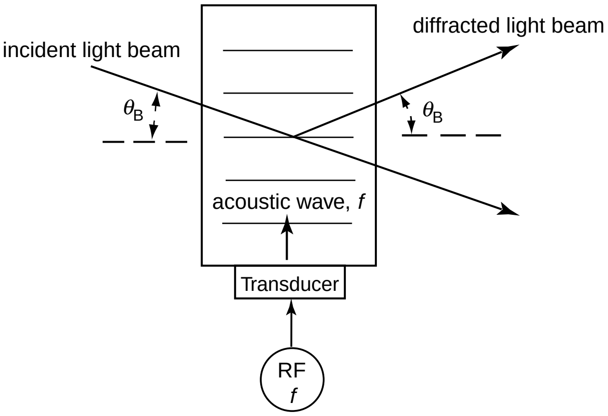

Figure 5 Acousto-optic diffraction cell.

Anisotropic Bragg Diffraction

In optically anisotropic materials, acousto-optic Bragg diffraction can occur between an ordinary and extraordinary optical beam, and vice versa. This is generally referred to as birefringent diffraction, and it offers additional design capabilities compared to the isotropic case, such as enhancing the angular aperture, and extending the aperture–bandwidth product.

The theory of diffraction of light so far described has assumed that the optical medium is isotropic. A number of important acousto-optic devices make use of the properties of birefringent materials. It is different from diffraction in isotropic media in that the magnitude of the momentum vector of the light will in general be different for different light polarization directions. Thus, the vector diagram representing conservation of momentum will no longer be the simple isosceles triangle of Figure 4. The momentum vector for ordinary polarized light will, in general, be different from the momentum vector for light that is extraordinary polarized. To understand the effect of anisotropy on diffraction, it is necessary to mention another phenomenon which occurs when light interacts with shear acoustic waves, i.e., waves in which the displacement of matter is perpendicular to the direction of propagation of the acoustic wave. A shear acoustic wave may cause the direction of polarization of the diffracted light to be rotated by 90°. The underlying reason for this is that the shear disturbance induces a birefringence which causes the plane of polarization to be rotated. The acousto-optic tunable filter (AOTF) is a particularly interesting birefringent device. It was first developed for the case of collinear optical and acoustic beams, and was used for wavelength selection in a dye laser.

Diffraction Efficiency

The acoustic power required to diffract light will be determined by the geometric parameters as well as by the properties of the medium. A simplified calculation leads to results that are useful. Referring to the spectrum of a phase-modulated wave shown in Figure 2, we can see that the ratio of the intensity in the first order to that in the zero order is

$$I_1/I_0 = [B_1(\Delta\phi)/B_0(\Delta\phi)]^2 \qquad [16]$$

By analogy this same result comes about for acousto-optically diffracted light. The acoustic power flow is given by

$$P_\mathrm{a} = \frac{1}{2}cVe^2 \qquad [17]$$

where c is the elastic stiffness constant. The elastic stiffness constant is related to the bulk modulus β, the density ρ, and acoustic velocity v through the expression

$$c = \frac{1}{\beta} = \rho V^2 \qquad [18]$$

and the acoustic power density is

$$P_\mathrm{a} = \frac{1}{2}\rho V^3 e^2 \qquad [19]$$

The optical phase modulation of the medium resulting from the change in refractive index, Δn, is

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \qquad [20]$$

Using eqn [5] for Δn, the phase modulation is related to the acoustic power density by

$$\Delta\phi = -\pi(L/\lambda)\, n^3 p\,(2P_\mathrm{a}/\rho V^3)^{1/2} \qquad [21]$$

The diffraction efficiency can now be calculated by using this value for the optical phase change in eqn [16]. The first- and second-order Bessel functions can be approximated for small modulation index by

$$\begin{aligned} &B_0(\Delta\phi) \sim \cos(\Delta\phi) \sim 1 - (\Delta\phi)^2, \quad \text{and} \\ &B_1(\Delta\phi) \sim \sin(\Delta\phi) \sim \Delta\phi \end{aligned} \qquad [22]$$

The small signal approximation to the diffraction efficiency is then

$$I_1/I_0 \sim (\Delta\phi)^2 = \left(\pi^2 L^2/2\lambda^2\right)\left(n^6 p^2/\rho V^3\right)P_\mathrm{a} \qquad [23]$$

The expression $n^6 p^2/\rho V^3$ is known as the figure of merit of the material, and is designated as M_2; it is comprised entirely of intrinsic material properties.

Acousto-optic Materials

Acousto-optic device technology has matured to the point that performance is chiefly limited by material parameters, particularly the figure of merit and acoustic attenuation. Nature has arranged that materials with high figures of merit usually have high attenuation and vice versa. The widely used acousto-optic materials are fused quartz, tellurium dioxide, and lithium niobate. Development work on new infrared materials has been reported recently. A list of commonly used acousto-optic materials is given in Table 1. For materials with a low figure of merit, a higher drive power can be used to obtain the required efficiency.

Experience has indicated that the upper limit for very small devices (active area ~0.1 mm^2) is a drive power density of 100–500 mW/mm^2, provided there is proper heat sinking to transfer the heat energy. For larger devices, sizes greater than ~1 cm^2, the limit is closer to a few W/cm^2. At the high drive power levels, the acoustic attenuation may cause significant optical distortion.

AO Devices

Resolution, bandwidth, and speed are the important characteristics of acousto-optic scanners, shared by all types of scanning devices. Depending upon the application, only one, or all, of these characteristics may have to be optimized. Consider the acousto-optic scanner in Figure 6 with a collimated incident beam of width D, diffracted to an angle θ_B at its RF bandcenter, and whose bandwidth is Δf. If the diffracted beam is focused onto a plane by a lens, or lens combination, at the scanner, the diffraction spread of the optical beam will be

$$\Delta x = F\Delta\phi \sim F - \lambda/D \qquad [24]$$

Table 1 Selected acousto-optic materials

Material	*Transmission range (μm) mission*	*Acoustic mode & propagation direction*	*v (cm/s × 10^5)*	*Acoustic attenuation (dB/cm GHz^2)^c*	*n*	*M_2 (s^3/g × 10^{-18})*
Visible–near-infrared (VIS-NIR)						
$LiNb0_3$	0.04–4.5	L[100][a]	6.57	0.15	2.20	7.0
		S[001][b]	3.59	2.6	2.29	2.92
TiO_2,	0.45–6	L[001]	10.3	0.55	2.58	1.52
α-SiO_2;	0.12–4.5	L[001]	6.32	2.1	1.54	1.48
		L[100]	5.72	3.0	1.55	2.38
TeO_2.	0.35–5	L[001]	4.20	15	2.26	34.5
		S[110]	0.616	90	2.26	793
Far IR						
Ge	2–20	L[111]	5.50	30	4.00	840
$T1_3AsS_4$ (TAS)	0.6–12	L[001]	2.5	29	2.63	510
GaAs	1–11	L[110]	5.15	30	3.37	104
ZnTe	0.55–20	L[110]	3.37	130	2.77	18
GaP	0.6–10	L[110]	6.32	60	3.31	30
$T1_3PSe_4$	0.85–9	L[100]	2.0	150	2.9	2069
Te	5–20	L[100]	2.2	60	4.8	4400
CdS	0.5–11	L[100]	4.17	90	2.44	12
GaP	0.6–10	L[110]	6.32	60	3.31	30
ZnS	0.4–12	L[001]	5.82	27	2.35	3.4
		S[001]	2.63	130	2.35	8.4

[a] Longitudinal acoustic waves propagating along the [100] crystallographic direction.

[b] Shear acoustic waves propagating along the [001] crystallographic direction.

[c] Attenuation constant at 1 GHz; the frequency dependence of the attenuation for most crystals is quadratic.

where F is the focal length of the lens. The light intensity will be distributed in the focal plane with a sinc^2 distribution by aperture diffraction. The number of resolvable spots, N, will be the angular scan range divided by the angular diffraction spread,

$$N = \frac{\Delta\theta}{\Delta\phi} \qquad [25]$$

where $\Delta\theta$ is the range of the angular scan. Differentiating the Bragg angle formula yields

$$\Delta\theta = (\lambda/V \cos\theta_0)\Delta f \qquad [26]$$

and

$$N = \Delta f(D/V\cos\theta_0) = \Delta f\tau \qquad [27]$$

where θ_0 is the Bragg angle at band center, and τ is the time that it takes the acoustic wave to cross the optical aperture. The resulting expression is the time–bandwidth product of the acousto-optic scanner, a concept applied to a variety of electronic devices as a measure of information handling capacity. The time–bandwidth product of an acousto-optic Bragg cell is equivalent to the number of bits of information which may be instantaneously processed by the system. In order to maximize the number of resolution elements, it is desirable to have as large a bandwidth as possible (i.e., large frequency range) and also as large an aperture delay time as possible. There are two factors limiting the bandwidth of an acousto-optic device: the bandwidth of the transducer structure (discussed later), and acoustic absorption in the delay medium. The acoustic absorption increases with increasing frequency; for high-purity single crystals the increase generally goes with the square of the frequency. For glassy materials, on the other hand, the attenuation will increase more slowly with frequency, often approaching a linear function. The maximum frequency is generally taken as that for which the attenuation of the acoustic wave across the optical aperture is equal to

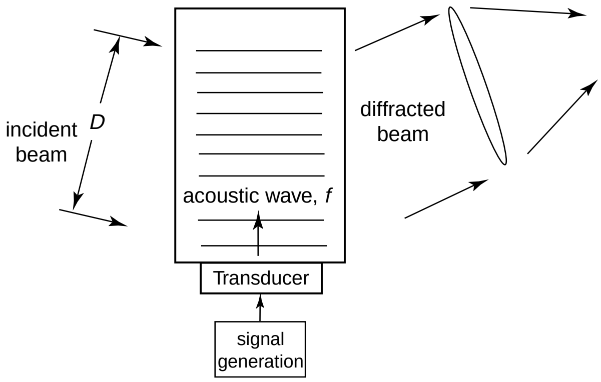

Figure 6 Acousto-optic scanner.

3 dB. A reasonable approximation of the maximum attainable bandwidth is

$$\Delta f = 0.7 f_{\max} \qquad [28]$$

Birefringent Scanners

The birefringent scanner is a significant use for anisotropic Bragg diffraction. There are a number of advantages with the birefringent scanner over the anisotropic scanner, such as a larger angular scan range along with lower frequencies. There are also disadvantages, such as lower speed due to generally lower shear wave velocities, and the particular application will dictate whether an isotropic or birefringent scanner is better. Scanners represent a fairly important aspect of acousto-optic devices due to the widespread commercial use of laser beam deflectors for displays and laser printers. For applications where the required speed is beyond that of mechanical scanners, the acousto-optic scanner is an ideal candidate. However, unlike mirrors, acousto-optic deflectors are wavelength sensitive, and can only be used with single-wavelength laser beams.

The birefringent scanner can be described with the wavevector diagram, as shown in Figure 7 for a positive uniaxial crystal where the extraordinary index of refraction is larger than the ordinary. The acoustic wavevector is tangent to the diffracted surface, which produces the largest scan angle for a given acoustic bandwidth. This is also the degenerate case, where only a single diffracted beam results, whereas two beams at two different acoustic frequencies will generally result from arbitrary input and acoustic propagation directions. The azimuthal acceptance angles (angles normal to the plane of incidence) of the acoustic and optical wavevectors can be different, although propagation and polarization directions must be selected that will produce high efficiency.

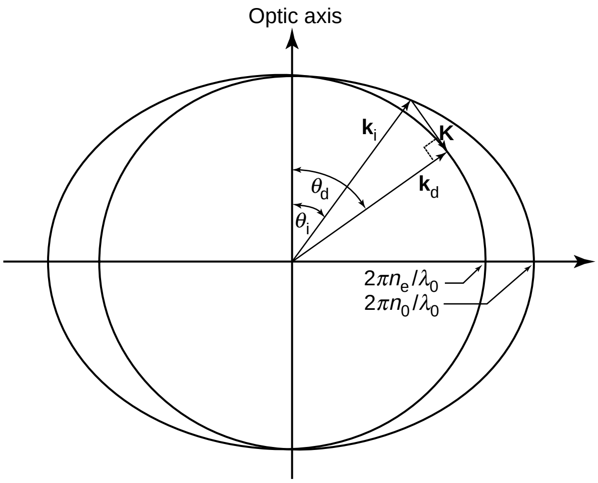

Figure 7 Wavevector diagram for birefringent scanner using a positive uniaxial crystal.

For the positive uniaxial case, the acoustic and diffracted wavevectors will be perpendicular at the design point, which allows the center frequency to be calculated from geometry as

$$f_0 = (V/\lambda)(n_\mathrm{i}^2 - n_0^2)^{1/2} \qquad [29]$$

where n_i is the index of refraction at the incidence angle. Due to the typically low velocity of shear waves, this frequency can be much lower than an isotropic scanner, and the incidence angle can also be chosen to adjust the frequency. It is also possible to use the optical activity of certain materials such as TeO_2, along with circularly polarized light to further reduce the acoustic frequency. It is important to maintain as low a frequency as possible due to acoustic attenuation, which is especially high with the soft materials typically used for birefringent applications.

The bandwidth over which the scanner can efficiently operate is fairly large due to having the acoustic and diffracted wavevectors perpendicular, and is approximately $\Delta f = 2f_0$, assuming an octave of bandwidth for the isotropic scanner. This will produce a larger scan angle than an isotropic scanner, and more spots of resolution. The number of resolution spots is determined through diffraction by the aperture size and diffraction, along with the scan angle. Since a larger aperture requires a longer time for the acoustic waves to propagate across the aperture, the response time τ to access the scanner will increase, and the product of τ and f is related to the number of spots by eqn [27]. The advantage of the birefringent scanner is that it can operate at a lower frequency f_0 with better performance. However, since other factors such as acoustic attenuation are important in designing a scanner, for some applications it might be better to operate an isotropic scanner at a higher frequency.

AOTFs

With anisotropic Bragg diffraction, the magnitude of the diffracted wavevector $\mathbf{k}_\mathrm{d}$ will differ from the incident wavevector $\mathbf{k}_\mathrm{i}$, which cannot occur for the isotropic case. This is illustrated in Figure 8 for an AOTF utilizing a negative uniaxial crystal where an extraordinary input wave propagating at an incident angle θ_i relative to the crystal axis is diffracted into an ordinary output wave at an angle θ_d. This occurs through an acoustic wave propagating at an angle θ_a

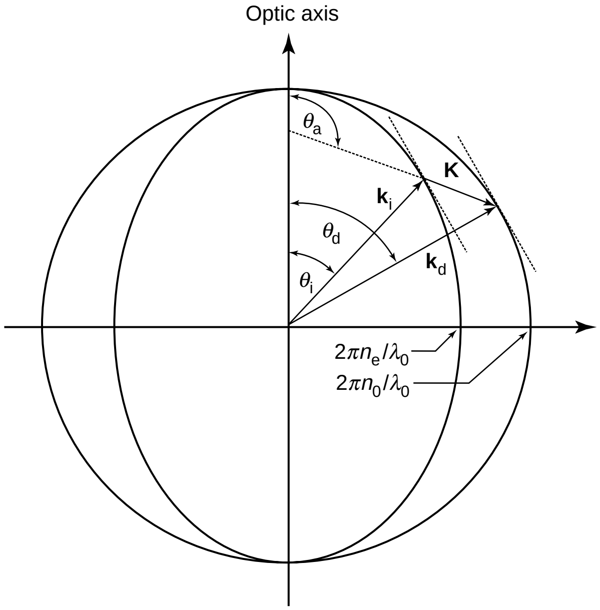

Figure 8 Wavevector diagram for an AOTF using a negative uniaxial crystal.

with wavevector $\mathbf{K}_a$, where all the wavevectors lie in a plane through the optic axis of the crystal. This diagram is identical to the index ellipsoid for the crystal, scaled by $2\pi/\lambda$, where λ is the vacuum wavelength. The acoustic wavevector is shown as adding to the incident wavevector, which increases the frequency of the optical wave by the acoustic frequency. It can also be represented in the reverse direction, which would reduce the optical frequency by the same amount.

An AOTF spectrally filters the optical input, and maintains its spectral purity over a large angular aperture. These conditions are maximized when the power flow or ray directions of the input and output beams are collinear. This produces a parallel-tangents condition, where lines drawn tangent to the wavevector surfaces and connecting the input and diffracted wavevectors are parallel. For beams at 90° to the optical axis, it is referred to as the collinear case, whereas all other parallel-tangents conditions are noncollinear. The diffracted beam is rotated by 90° during anisotropic Bragg interaction, so that crossed polarizers can be used to separate the input and diffracted beams. For the collinear case polarization separation must be used, whereas angular separation can also be used for the noncollinear case.

The AOTF design equations can be derived from the geometrical conditions imposed by the wavevector matching condition

$$\mathbf{k}_d = \mathbf{k}_i + \mathbf{K} \qquad [30]$$

along with decomposing the acoustic wave into its Fourier components which result from the finite interaction length produced by the transducer. These components allow for phase matching over a range of input angles, and they also allow for a spectral spread of the interaction. The full width half maximum (FWHM) resolution is given by

$$\Delta\lambda = \frac{1.8\lambda^2}{bL \sin 2\theta_i} \qquad [31]$$

where λ is the vacuum wavelength, b is the dispersion term (essentially $2\pi\Delta n$, where the birefringence Δn is the difference between the ordinary index n_0 and extraordinary index n_e of refraction), L is the interaction length of the input beam defined geometrically by the acoustic beam, and θ_i refers to the angle of the central or cardinal ray of the input beam, and the AOTF is designed around this angle. It is therefore a sensitive function of the incidence design angle, with the highest resolution occurring for the collinear case. The resolution is also narrower for larger birefringence.

The geometry of a noncollinear AOTF is shown schematically in Figure 9. The acoustic waves propagate in the correct direction and are generated by a transducer of length L_t, which is related geometrically to the interaction length. These waves are generally absorbed at the other side of the AOTF to prevent interfering reflections, and the sides may also be wedged to eliminate reflections into the interaction region. The optical beam enters the AOTF at the correct angle to the crystal optic axis, and either diffracts or passes through as the zero order. The input beam will have an angular spread, producing an acceptance angle that is a function of the interaction length, wavelength, and crystal parameters, and the external solid angular aperture

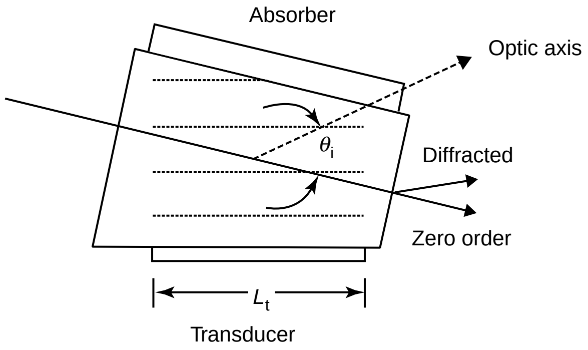

Figure 9 Noncollinear AOTF orientation of optical and acoustic beams.

is roughly given by

$$\Delta\Omega = \frac{\pi n^2 \lambda}{\Delta n L} \qquad [32]$$

where n refers to the diffracted beam index of refraction, which can be either n_0 or n_e depending upon the design. The resolution and solid angle are therefore related through the transducer length, and the product of the resolving power times the solid angle is given by

$$(\lambda/\Delta\lambda)(\Delta\Omega) = 2\pi n^2 \sin^2\theta_i / |F_\theta F_\varphi|^{1/2} \qquad [33]$$

where

$$F_\theta = 2\cos^2\theta_i - \sin^2\theta_i \qquad [34a]$$

$$F_\varphi = 2\cos^2\theta_i + \sin^2\theta_i \qquad [34b]$$

are form factors for the polar and azimuthal components of the optical field. Since all the wavectors lie in a plane containing the optic axis, the azimuthal components of the incident, diffracted, and acoustic wavevectors must always be identical under the parallel-tangents condition.

The product $(\lambda/\Delta\lambda)(\Delta\Omega)$ forms a figure of merit for spectrometers, and for the collinear case, the product is identical to a Fabry–Perot etalon having an index of n, and at other angles it is more complicated due to the angular dependence. The angular field is also symmetrical for the collinear AOTF, but nowhere else other than along the optic axis, where the resolving power becomes zero.

The great advantage of the AOTF designed under the parallel-tangents condition is the large angular aperture compared to the general case. Under this condition, the dependence of the angular aperture on resolving power is second order rather than first order, and the angular aperture can be tens of degrees for a typical resolution, which is useful for imaging applications. At the special angle of 54.7°, the second-order dependence is also zero, and the aperture can be extremely large relative to the resolving power, although this requires a high-resolution device with a corresponding narrow field, since the condition would soon be violated at larger field angles.

The efficiency of the AOTF under phase matching is given by

$$\eta = \sin^2[\pi^2 M_2 P_a L^2/(2\lambda^2)]^{1/2} \qquad [35]$$

which depends on θ_i and various material properties. Since the interaction is anisotropic, M_2 must be taken as a tensor quantity, in which both the polarization and propagation directions of the acoustic and optical waves must be accounted for. For acoustic waves, the polarization direction is the particle motion, which is perpendicular to the propagation direction for shear waves. In general, M_2 is much larger for a specific configuration of a particular material, and AOTFs are designed around this condition. The range of useful design angles is also usually limited since M_2 is generally a sensitive function of θ_i. The angular dependence on M_2 is important in designing an efficient device since it can be near zero at specific design angles for some materials, such as with TeO_2 for the collinear case.

The AOTF must be designed to optimize performance. The resolution is generally given as a system requirement, and the design angle along with the transducer length must be adjusted to optimize the throughput, or total optical power through the AOTF. This requires maximizing the efficiency, angular aperture, and aperture dimension. The maximum crystal size that can be grown ultimately limits these parameters, both in the interaction length and in the aperture size.

AOTFs have been used for a wide range of spectral filtering applications for spectroscopy and laser applications. Both collinear and noncollinear devices have electronically tuned a variety of lasers, including dye, semiconductor, and Ti:sapphire lasers. The AOTF is placed within the laser cavity, which requires high transmission on the laser line, with enough resolution to separate nearby laser transitions. A related application is the multiplexing of optical communications systems, where the AOTF is used to separate the various laser diode wavelengths. Perhaps the most widely used application is spectroscopy, and single-point detection systems have been used for a variety of biological and chemical applications in the visible and infrared. These techniques have been extended to spectral imaging, again in the visible and the infrared. Due to a variety of AOTF aberrations the image quality will be somewhat degraded. Since the AOTF is electronically tunable, it can be rapidly modulated both in amplitude and wavelength, which allows for modulation spectroscopy to detect small signals in a large background. By applying multiple frequencies, the AOTF has also been used as a rejection filter, in which all wavelengths other than that selected are allowed to pass.

Modulators

Acousto-optic modulators are used to vary and control laser beam intensity. A Bragg configuration gives a single first-order output beam, where intensity is directly linked to the power of RF control signal. The rise time of the modulator is the time required for the acoustic wave to traverse the laser beam. For minimum rise time the laser beam will be focused down, forming a beam waist as it passes through the modulator.

A Bragg configuration gives a single first-order output beam, the intensity of which is controlled by the RF signal power, and diffaction angle controlled by the drive frequency. Two deflectors can be used in series and at right angles to each other to give full two-dimensional scanning.

One requirement in the design of acousto-optic modulators is that the diffracted beam and undiffracted beam must be well separated. For an adequate extinction ratio, the Bragg angle should be at least as large as the divergence of the optical beam. This condition puts a minimum value on the center frequency. Equating the Bragg angle

$$\theta_{\mathrm{B}} = \frac{\lambda f_0}{2nV} \qquad [36]$$

and the diffraction angle of the Gaussian beam

$$\delta\theta_0 = \frac{4\lambda}{\pi(nd)} \qquad [37]$$

where d is the beam diameter at the minimum, it follows that the lower limit of the acoustic frequency is given by

$$f_0 = \frac{8}{\pi\tau} \qquad [38]$$

We may combine these expressions to determine a limit of modulator bandwidth of acousto-optic modulators, with the result

$$\Delta f \sim \frac{1}{4} f_m \qquad [39]$$

i.e., the modulation bandwidth is approximately equal to 25% of the midband acoustic frequency, f_{m}. In view of the present status of transducer technology, the modulation bandwidth of acousto-optic modulators is limited to several hundred MHz.

A frequency shifter uses the shift inherent in the acousto-optic interaction to up- or down-shift a laser's frequency. Two kinds of shifters can be distinguished: the fixed frequency shifter, and the variable frequency shifter. The frequency shift is equal to the signal frequency applied to the transducer. Various device configurations can be used to shift the laser beam, such as multiple travels inside the shifter to double or triple the frequency shift, or a combination of two frequency shifters in series. Acousto-optic frequency shifters can be used, for example, in optical heterodyning and interferometry, or in laser Doppler velocimetry for particle velocity analysis.

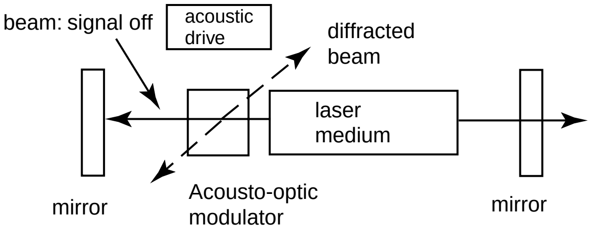

Figure 10 An acousto-optic *Q*-switch in a laser cavity.

Q-Switches

A *Q*-switch is a special modulator which introduces high repetition rate losses inside a laser cavity (typically 1 to 100 KHz). They are designed for minimum insertion loss and to be able to withstand very high laser powers. In normal use an RF signal is applied to diffract a portion of the laser cavity flux out of the cavity. This increases the cavity losses and prevents oscillation. When the RF signal is switched off, the cavity losses decrease rapidly and an intense laser pulse evolves. A diagram of a typical acousto-optic *Q*-switched laser cavity is shown in Figure 10. *Q*-switches are the preferred method of intracavity modulation where high speed, high stability, and modest cost are important design factors.

List of Units and Nomenclature

b	dispersion constant	–
c	elastic stiffness	dynes cm^{-2}
D, d	optical beam diameters	cm
e	elastic strain	–
f	acoustic frequency	hertz
F	optical focal length	cm
F_ϕ, F_θ	angular form factors	–
$\mathbf{k}$	optical momentum wavevector	cm^{-1}
$\mathbf{K}$	acoustic momentum wavevector	cm^{-1}
M_2	acousto-optic figure of merit	$\mathrm{sec}^3\,\mathrm{gram}^{-1}$
n	refractive index	
p	photo-elastic constant	–
P	pressure	dynes cm^{-2}
P_{a}	acoustic power	watts cm^{-2}
V	acoustic velocity	cm sec^{-1}
β	compressibility	$\mathrm{cm}^2\,\mathrm{dyne}^{-1}$
η	diffraction efficiency	–
ε	dielectric constant	–
ω	optical frequency	hertz
Ω	solid angle	steradians
ρ	density	grams cm^{-3}
θ	Bragg angles	radians
τ	acoustic wave travel time	sec

See also

Modulators: Acousto-Optics. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Chang IC (1976) Acousto-optic devices and applications. *IEEE Transactions: Sonics and Ultrasonics* SU-23: 2–22.

Chang IC (1981) Acousto-optic tunable filter. *Optical Engineering* 20: 824–829.

Dixon RW (1967) Photo-elastic properties of selected materials and their relevance for applications to acoustic light modulators and scanners. *Journal of Applied Physics* 38: 5149–5153.

Gordon EI (1966) A review of acousto-optical deflection and modulation devices. *Proceedings of the IEEE* 54: 1391–1401.

Gottlieb M, Ireland CLM and Ley JM (1983) *Electro-optic and Acousto-optic Scanning and Deflection*. New York: Marcel Dekker.

Goutzoulis AP and Pape DR (1994) *Design and Fabrication of Acousto-optic Devices*. New York: Marcel Dekker.

Uchida N and Niizeki N (1973) Acousto-optic deflection materials and techniques. *Proceedings of the IEEE* 61: 1073–1092.

Xu J and Stroud R (1992) *Acousto-optic Devices: Principles, Design, and Applications*. New York: Wiley.

Electro-Optics

L R Dalton, University of Washington, Seattle, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The electro-optic (EO) effect can be described as any one of a number of phenomena that occur when an electromagnetic wave in the optical spectrum (e.g., characterized by wavelengths in the range 200 to 2000 nm) interacts with an electric field, or with matter under the influence of an electric field. Two of the most important electro-optic phenomena are the Kerr effect (discovered by John Kerr in 1875) and the Pockels effect (discovered by Friedrich Pockels in 1893), in which birefringence is induced or modified in a liquid (the Kerr effect) or a solid (the Pockels effect). Birefringence is the difference in refractive indices for light of orthogonal line polarizations, one of which is parallel to the induced optical axis. The Kerr effect involves creation of birefringence in a liquid that is otherwise not birefringent. The degree of birefringence is quadratically proportional to the applied electric field strength. Hence, the Kerr effect is frequently referred to as the quadratic electro-optic effect while the Pockels effect involves a linear dependence on the applied electric field and is referred to as the linear electro-optic effect. In considering practical applications of the electro-optic effect, a commonly encountered term is that of 'electro-optic modulator', which can be described as a device wherein a signal control element is used to modulate a beam of light. The control element is typically an electric field with frequency components in the zero (DC) to hundreds of gigahertz (1 GHz = 10^9 Hz) range (or even tens of terahertz (1 THz = 10^{12} Hz). The modulation may be imposed on the phase, frequency, amplitude, or direction of the modulated optical beam.

Electro-optic effects are one class of second-order nonlinear optical phenomena. Other important second-order nonlinear optical phenomena include sum (e.g., second-harmonic) and difference frequency generation. Such phenomena derive from the second-order term in the power series expansion of macroscopic (material) polarization, P, in terms of applied electric fields:

$$P_i = \chi^{(1)}_{ij} E_j + \chi^{(2)}_{ijk} E_j E_k + \chi^{(3)}_{ijkl} E_j E_k E_l + \cdots \qquad [1]$$

where $\chi^{(1)}_{ij}$, $\chi^{(2)}_{ijk}$, and $\chi^{(3)}_{ijkl}$ are the linear, second-, and third-order optical susceptibilities, respectively. Second-harmonic generation (SHG) or frequency doubling (2ω) effects, where a beam at twice the frequency of the input beam is generated, can be seen by substituting a sinusoidal field, $E_\omega \cos(\omega t - kz)$; for E_j and E_k. After using a well-known trigonometric identity, the equation for macroscopic polarization becomes (through second-order and dropping the subscript i):

$$\begin{aligned} P = {} & \chi^{(1)} E_\omega \cos(\omega t - kz) + (1/2)\chi^{(2)} E^2_\omega \\ & \times [1 + \cos(2\omega t - 2kz)] + \cdots \end{aligned} \qquad [2]$$

The electro-optic effect can be appreciated by considering interaction of the medium with an optical field $E_\omega \cos(\omega t - kz)$ and a low-frequency field E_0. Note that although we have used the symbol zero to denote the frequency of the low-frequency electrical field, actual frequencies can extend to hundreds of gigahertz or even tens of terahertz. Again substituting (and now keeping terms to third order), the equation for polarization can be written as:

$$\begin{aligned} P &= \chi^{(1)}E_\omega\cos(\omega t-kz)+2\chi^{(2)}E_0E_\omega\cos(\omega t-kz) \\ &\quad +3\chi^{(3)}E_0^2E_\omega\cos(\omega t-kz)+(3/4)\chi^{(3)}E_\omega^3\cos(\omega t-kz) \\ &= \chi_{\text{eff}}E_\omega\cos(\omega t-kz) \end{aligned} \qquad [3]$$

The equation for the nonlinear index of refraction, n, can be expressed as

$$\begin{aligned} n^2 &= 1+4\pi\chi_{\text{eff}} \\ &= 1+4\pi[\chi^{(1)}+2\chi^{(2)}E_0+3\chi^{(3)}E_0^2+(3/4)\chi^{(3)}E_\omega^2] \end{aligned} \qquad [4]$$

Denoting the linear index of refraction as n_0, the above equation can be rewritten as

$$n^2-n_0^2=8\pi\chi^{(2)}E_0+12\pi\chi^{(3)}E_0^2+3\pi\chi^{(3)}E_\omega^2 \qquad [5]$$

which in turn leads to

$$\begin{aligned} n &= n_0+(4\pi/n_0)\chi^{(2)}E_0+(6\pi/n_0)\chi^{(3)}E_0^2 \\ &\quad +(3\pi/2n_0)\chi^{(3)}E_\omega^2 \end{aligned} \qquad [6]$$

The definition of light intensity in cgs units is $E_\omega^2=(8\pi/cn_0)I(\omega)$, which when substituted into the above equation gives

$$\begin{aligned} n &= n_0+(4\pi/n_0)\chi^{(2)}E_0+(6\pi/n_0)\chi^{(3)}E_0^2 \\ &\quad +(12\pi/cn_0^2)\chi^{(3)}I(\omega) \end{aligned} \qquad [7]$$

The index of refraction can now be written as

$$n(\omega)=n_0(\omega)+n_1(\omega)E_0+n_2(0)E_0^2+n_3(\omega)I(\omega) \qquad [8]$$

where the terms $n_1(\omega)$, $n_2(0)$, $n_3(\omega)$ correspond to the linear electro-optic effect, the quadratic electro-optic effect, and the optical Kerr effect, respectively. The above equation can also be expressed as

$$n=n_0-(1/2)n^3rE_0-(1/2)n^3sE_0^2-\cdots \qquad [9]$$

where r and s are the linear and quadratic electro-optic coefficients, respectively.

Unfortunately, a great deal of confusion exists concerning the use of the terms electro-optic effect and electro-optic modulator. First of all, the terms ‘electro-optic’ and ‘opto-electronic’ are frequently confused and used interchangeably. Opto-electronic devices function as electrical-to-optical or optical-to-electrical signal transducers. Examples of commonly used opto-electronic devices include light-emitting diodes (LEDs), photodiodes (PDs), injection laser diodes (ILDs), and integrated optical circuit (IOC) elements. An electro-optic device can also function as an electrical-to-optical signal transducer (e.g., a Mach–Zehnder interferometer or a birefringent modulator employing polarizers at the input and the output); however, the mode of operation is entirely different for electro-optic and opto-electronic devices. For example, with LEDs or modulated lasers, the applied electric field produces modulation of transmission of light by altering the excited state population of the light emitting (lasing) state of the material. With electro-optic materials, no actual excited state population is generated. Electro-optic devices typically operate in regions of the optical spectrum removed from resonant transitions, i.e., regions of relatively high transparency. The applied electric field acts only to perturb the charge distribution of the material (the spatial positions of the electrons and nuclei). The altered charge distribution interacts with the optical beam transmitting the material with the result that the speed of light in the material (i.e., the index of refraction or birefringence) is altered. The response times of opto-electronic and electro-optic phenomena (and hence bandwidths of devices exploiting these phenomena) are quite different. In the former case, response will be limited by the lifetime of the relevant (emitting) excited state while in the latter case, response will be defined by the relaxation time of the material (e.g., reorientation time of a liquid or lattice relaxation time of a solid) following removal of the perturbing electric field.

An even greater confusion can arise due to the jargon used in particular technologies, such as telecommunications. A telecommunications engineer frequently refers to ‘electro-optic switching’ when describing opto-electronic transduction of an optical signal to an electrical signal, followed by re-routing in the electrical domain, and finally opto-electronic transduction of the electrical signal back to the optical domain. Electro-optic switching (in the sense used in this article) involves quite a different operation and would be described by that same telecommunications engineer as ‘all-optical switching’ – a term reserved by physicists for the optical Kerr effect (control of one light beam by another light beam).

A second point of confusion involves distinguishing between the terms ‘electro-optic’ and ‘electro-absorptive’ modulation. Again, the terms are frequently used interchangeably. However, they involve quite different physical mechanisms. An electro-absorptive modulator is, like a modulated laser or LED, a 'resonant' device. In a device such as a gallium arsenide (GaAs) or indium phosphide (InP) quantum dot electro-absorptive modulator, the position and width of an optical absorption (resonant transition) are defined by the physical dimensions of the quantum dot. Application of an electric field shifts the spectral wavelength of the quantum dot absorption. This phenomenon can be used to dramatically change optical transmission through (or equivalently, absorption by) the material by choosing the wavelength of the propagating beam of light to be near the resonant absorption. With the electric field off, the material is reasonably transparent but becomes strongly absorbing when the electric field is applied, shifting the resonant absorption to the wavelength of the propagating beam. Obviously, the voltage required to achieve a desired change in optical transmission will depend on control of quantum dot dimensions in fabrication, which has been and continues to be a topic of research focus for electro-absorptive materials. Likewise, control of insertion losses of such devices is a concern as the wavelength used for the transmissive state of the device must be sufficiently close to the optical resonance to achieve the desired attenuation with application of a modest control voltage. Another issue to be faced with electro-absorptive devices is that of 'chirp' (optical signal distortion arising from the fact that both absorption and index of refraction – the imaginary and real parts of the optical susceptibility – change with application of the electric control field). On the positive side, electro-absorptive modulators are 'polarization-insensitive' modulators and are thus conveniently used with multimode (as well as single mode) optical transmission.

Further confusion exists because widely different types of materials can be used in electro-optic and electro-absorptive devices. As these materials are frequently competing for the same applications (signal transduction, optical switching, etc.), the advantages and disadvantages of different materials are often compared without maintaining the context that quite different phenomena are involved. Even restricting our discussion to electro-optic materials, we note that materials can range from organic liquid crystalline materials (nematic, smetic, ferroelectric, etc.) to organic electro-optic materials (both crystalline and 'macromolecular') to inorganic crystalline materials (lithium niobate, $LiNbO_3$; barium titanate, $BaTiO_3$; barium borate, BBO; potassium dihydrogen phosphate, KDP; lithium tantalate, $LiTaO_3$; zinc telluride, ZnTe, etc.). The response times of these materials to transient application of an applied electric field will be quite different, reflecting the different types of lattice motion involved. With liquid crystalline materials, particularly nematic materials, considerable molecular reorientation is involved and due to the mass that must be moved, response times are quite slow although the index change will be relatively large. With use of liquid crystalline materials, device bandwidths will typically be limited to tens of megahertz (MHz) or less. In contrast, conjugated π-electron organic materials typically exhibit response times of tens to hundreds of femtoseconds, which translates to potential device bandwidths of tens of terahertz (actual device bandwidths may be less due to factors other than material response). For conjugated π-electron organics, the response time is the phase relaxation time of the π-electron system. Because only a slight change in bond length alternation of the conjugated π-electron system occurs with application of an applied electric field and because strong electron–phonon coupling and resonance stabilization of the π-electron system act to reduce the barrier to lattice relaxation, the response times of π-electron organic materials are among the fastest observed in nature. With crystalline inorganic materials, the ionic constituents move to new locations with application of an electric field with the exact movement determined by the field strength, the charge on the ions, and the restoring force. Unequal restoring forces along the three mutually orthogonal (perpendicular) axes of the crystal lead to anisotropy in the optical properties of the material. The applied electric field will induce a change in the anisotropy (the principal refractive indices). The symmetry of the electro-optic tensor will be closely related to the symmetry of the piezoelectric tensor. The linear electro-optic effect requires that the crystal exhibit noncentrosymmetric (acentric or ferroelectric) symmetry. A centrosymmetric crystal possesses a center of symmetry defined by identical particles existing in the lattice at vectors $\mathbf{r}$ and $-\mathbf{r}$, where $\mathbf{r}$ is a position vector measured from an arbitrary origin. A centrosymmetric crystal, like an isotropic liquid or gas, can exhibit a quadratic electro-optic effect.

For the sake of completeness, it can also be noted that index of refraction changes can be induced by acoustic waves (acousto-optic modulation) and by heating (thermo-optic modulation). Also, elasto-optic and photo-elastic effects can produce index of refraction changes.

To keep this article to a manageable length, discussion is limited to solid-state electro-optic materials. Design principles being used to produce new organic electro-optic materials will be illustrated. Since inorganic materials exist as crystals, limited chemical modification of such materials is possible, although processing techniques for fabricating thin films of such crystalline materials have been developed in a number of cases. For organic macromolecular materials, development of new materials is an on-going activity. Several representative devices being fabricated, using electro-optic materials, will be discussed. Four of the most commonly encountered types of EO devices include: (i) stripline devices; (ii) prism, cascaded prism, and superprism devices; (iii) resonated devices, such as ring microresonator and photonic bandgap devices, and (iv) waveplates. Of course, devices such as stripline devices can be configured in a variety of ways to produce polarization-sensitive and polarization-insensitive electrical-to-optical signal transducers, optical switches either using simple 1 × 2 or 2 × 2 switch architectures or multimode interference (MMI) switches, optical gyroscopes, photonically controlled phased array radar systems, spectrum analyzers, optical digital signal processors, analog-to-digital (A/D) and digital-to-analog (D/A) signal converters, electromagnetic (radiofrequency to millimeter wave) signal generators, voltage sensors, etc. Prism-type devices are typically used for spatial light modulation or laser beam deflection. Ring microresonator devices afford a rich array of applications ranging from active wavelength division multiplexing (WDM) to active optical interconnect reconfiguration, to voltage-controlled wavelength tuning of laser outputs.

Basic Relationships

The effect of the applied field is to deform the index of refraction ellipsoid, which can be represented as

$$\left(\frac{1}{n^2}\right)_1 X^2 + \left(\frac{1}{n^2}\right)_2 Y^2 + \left(\frac{1}{n^2}\right)_3 Z^2 + 2\left(\frac{1}{n^2}\right)_4 YZ + 2\left(\frac{1}{n^2}\right)_5 XZ + 2\left(\frac{1}{n^2}\right)_6 XY = 1 \qquad [10]$$

leading to the corrections

$$\Delta\left(\frac{1}{n^2}\right)_i = \sum_{j=1}^{3} r_{ij} E_{0j} \qquad [11]$$

where the electro-optic tensor components are related to the tensor components of the second-order nonlinear optical susceptibility by $r_{ij} = -(8\pi/n_{0i}^2 n_{0j}^2)\chi_{ji}^{(2)}$. The full tensorial equation for change of the 'indicatrix' (index of refraction ellipsoid) can be expressed as

$$\begin{pmatrix} \Delta\left(\frac{1}{n^2}\right)_1 \\ \Delta\left(\frac{1}{n^2}\right)_2 \\ \Delta\left(\frac{1}{n^2}\right)_3 \\ \Delta\left(\frac{1}{n^2}\right)_4 \\ \Delta\left(\frac{1}{n^2}\right)_5 \\ \Delta\left(\frac{1}{n^2}\right)_6 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{pmatrix} \begin{pmatrix} E_{0X} \\ E_{0Y} \\ E_{0Z} \end{pmatrix} \qquad [12]$$

The electro-optic tensor reduces considerably for specific materials reflecting the symmetry of the particular material. For inorganic crystalline materials, the electro-optic tensor can be reasonably complex (containing many nonzero elements). It is also helpful to note that for crystalline inorganic materials, the electric-field-induced change in crystal shape leads to strain and the orientation of the indicatrix is altered, leading to an additional contribution to the change in index coefficients:

$$\Delta(1/n^2)_i = \sum_k p_{ik} S_k + \sum_j r_{ij} E_j \qquad [13]$$

where S_k is a component of the strain and p_{ik} is the elasto-optic tensor. At high frequencies, the inertia of the crystal prevents macroscopic straining so that the first term of the above equation vanishes. At low frequencies, elasto-optic effects cannot be ignored. Because the deformation leading to strain is generally caused by the inverse piezoelectric effect, it is linearly related to the applied electric field (the same dependence as the linear electro-optic effect). This can lead to a frequency dependence of the 'effective' electro-optic coefficient for crystalline materials.

A brief comment on the photo-elastic effect (the change in index coefficients produced directly by applied stress) is warranted. This effect has the form:

$$\Delta(1/n^2)_i = \sum_l \pi_{il} \sigma_1 \qquad [14]$$

where σ_l are the components of the stress and π_{il} are the piezo-optical coefficients. Note that this effect is independent of the applied electric field.

For axially symmetric 'charge-transfer type' organic chromophores prepared by deposition or electric field poling, only two unique tensor elements, r_{33} and r_{13}, survive:

$$\tilde{r} = \begin{pmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{13} & 0 \\ r_{13} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad [15]$$

The electro-optic effect for organic macromolecular materials derives from molecules with conjugated π-electrons confined within molecules (chromophores). The electro-optic tensor elements can be related to molecular first hyperpolarizability, β, and the chromophore number density, N (molecules/cm^3), by

$$\begin{aligned} r_{33} &= -2Nf(0)\beta\langle\cos^3\theta\rangle/n_0e^4 \\ r_{13} &= -Nf(0)\beta\langle\sin^2\theta\cos\theta\rangle/n_{0o}^2 n_{0e}^2 \end{aligned} \qquad [16]$$

The expressions in brackets are 'order parameters' describing the degree of noncentrosymmetric order. $\langle\cos^3\theta\rangle = 1$ corresponds to all 'dipolar' (charge transfer type) molecules pointing in the same direction (perfect noncentrosymmetric or ferroelectric order) while $\langle\cos^3\theta\rangle = 0$ corresponds to complete disorder and no electro-optic effect. The reason that three angles appear in the order parameter is that an angle is required to represent the principal symmetry axis for the chromophore, the principal symmetry axis of the optical field, and the principal symmetry axis of the electric control field. Thus, averaging (denoted by brackets) must be carried out over each of these three angles. The tensor element r_{33} corresponds to the electric control field applied along the principal axis of the ordered chromophores and the principal axis of the optical field vector, transverse magnetic (TM) polarized light. The tensor element r_{13} corresponds to the control electric field applied along the principal axis of the optical field, transverse electric (TE) polarized light, and orthogonal to the principal chromophore axis. The notation r_{33} is contracted notation for r_{333}. The factor $f(0)$ takes into account the dielectric nature of the medium into which the chromophores (molecules with large first hyperpolarizability) are embedded.

For organic electro-optic materials, elasto-optic effects do not appear to make significant contributions. Moreover, electro-optic activity can be systematically improved by improving β and by optimizing the product of chromophore number density and order parameter. Due to the presence of intermolecular electrostatic interactions, W, among chromophores, order parameters and number density are not independent, e.g., for materials that are prepared by electric field poling the order parameter relevant to the principal component of the electro-optic tensor can be expressed approximately as

$$\langle\cos^3\theta\rangle = L_3(\mu f(0)E_\mathrm{p}/kT)[1 - L_1(W/kT)^2] \qquad [17]$$

where L_3, and L_1 are Langevin functions, μ is the chromophore dipole moment, E_p is the strength of the electric poling field, k is the Boltzmann constant, and T is the Kelvin poling temperature. W is a quadratic function of chromophore number density, N. The above equation indicates that a maximum will be observed in the graph of r_{33} versus N.

Materials

With inorganic and organic crystalline materials, very little can be done to optimize electro-optic activity other than discovering new EO crystalline materials or employing isotopic (e.g., deuterium for hydrogen) or ion substitution with existing materials.

As with organic liquid crystalline materials, the electro-optic activity of macromolecular organic second-order nonlinear materials can be systematically improved by molecular modification. As shown in Figure 1, quantum mechanical calculations can provide useful guidance as to which structural modifications will lead to improvements in molecular (first) hyperpolarizability and ultimately electro-optic activity. The molecular hyperpolarizability (long wavelength limit) can be increased by a factor of two by simple variation of the acceptor structure. The calculated values of wavelengths for the interband charge transfer transitions are 390 nm, 403 nm, and 430 nm. Since dipole moments do not change significantly with structure, intermolecular electrostatic interactions should be similar for these three chromophores and the improvement in β should translate to an improvement in electro-optic activity. This theoretical prediction has been experimentally verified. This is just one example of use of quantum mechanics to guide the improvement of electro-optic activity for organic materials. This figure also illustrates the typical structure of an organic electro-optic chromophore, which consists of an electron donor region (an amine donor in the example shown), a connective conjugated π-electron bridge, and an electron deficient electron acceptor region. The theoretically predicted (by density functional theory, DFT, calculations) variation of molecular first hyperpolarizability with molecular structure has been verified for the structures shown. However, the critical point is that electro-optic activity of organic materials (currently exhibiting a maximum value of 182 pm/V at 1.3 microns wavelength) is likely to be improved for some time to come. This value is to be compared to a value of 32 pm/V for lithium niobate. Another factor influencing the value of macroscopic electro-optic activity realized for electrically poled organic materials is the effect of intermolecular electrostatic interactions on optimization of $N\langle\cos^3\theta\rangle$. Statistical mechanical calculations can provide guidance as to how this is achieved (see Figure 2 where the role of chromophore shape is illustrated). Approximating the chromophore shape by a rigid prolate ellipsoid with embedded dipole moment yields nearly quantitative reproduction of the experimental data. If the π-electron structure (dipole moment) is left unchanged but the shape of the chromophore is altered to a spherical shape, a significant improvement in electro-optic activity is predicted. The shape effect arises because there are two components to the electronic dipole–dipole intermolecular electrostatic potential function describing the many body interaction of chromophores.

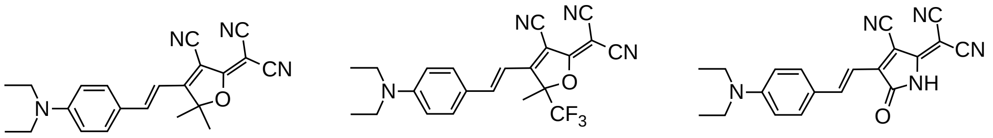

1 2 10

	1	2	10	
Dipole moment:	11.8	11.9	11.7	Debye
Hyperpolarizability	23.8	34.4	43.0	$*10^{-30}$ esu

Figure 1 Density functional calculations (DFT) of molecular first hyperpolarizability (β) and dipole moment (μ) for three chromophore structures.

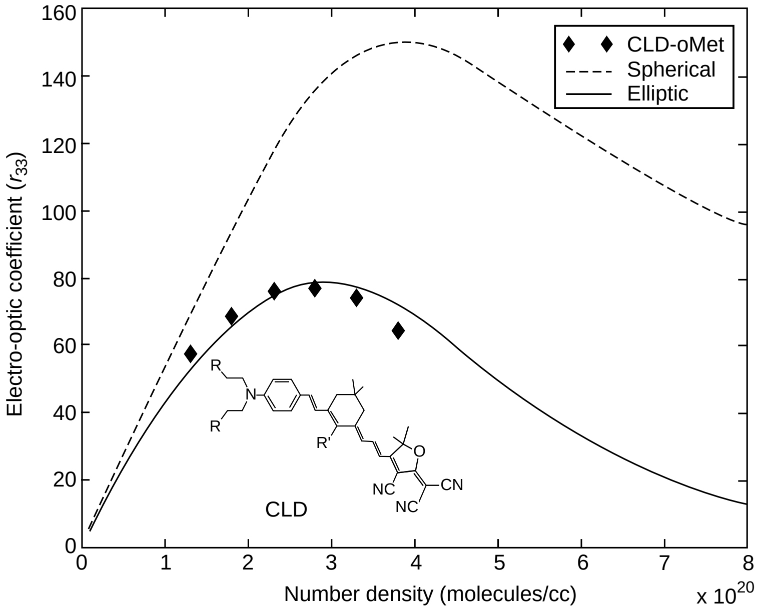

Figure 2 Two theoretical (statistical mechanical) simulations of the variation of electro-optic activity with chromophore number density are shown and compared with experimentally observed results for the CLD chromophore (structure shown) dissolved in an amorphous polycarbonate host polymer.

One component favors centrosymmetric ordering while the second component favors noncentrosymmetric ordering. Their relative weighting in defining macroscopic order depends on chromophore shape. Theoretical calculations have also shown that embedding chromophores in nanoscopic macromolecular objects (such as dendrimers and dendronized polymers) can dramatically enhance electro-optic activity. This type of engineering can also be used to control auxiliary properties such as optical loss and stability. The exceptional processability of organic electro-optic materials has permitted conformal and flexible devices to be fabricated and has permitted electro-optic circuitry to be integrated vertically on top of semiconductor VLSI (very large-scale integration) electronics. Mach–Zehnder modulators, with 3 dB operational bandwidths of 200 GHz, have been fabricated from organic electro-optic materials as have Mach–Zehnder devices with operational drive voltages of less than one volt. The major advantage that crystalline inorganic materials possess, relative to macromolecular organic EO materials, is superior thermal and photostability. A comparison of lithium niobate and the best organic EO materials, is given in Table 1. Cost is a significant factor if electro-optic materials are to compete with modulated lasers and other opto-electronic devices for applications in communications, computing, transportation, and defense. Currently, manufacturing costs for both crystalline and macromolecular materials are very high (several thousand dollars per modulator). Recently, it has been demonstrated that organic macromolecular materials can be used with soft lithography techniques to mass-produce electro-optic circuitry. This could conceivably dramatically reduce manufacturing costs. Also, the fact that organic macromolecular materials can be directly integrated with silicon photonics and VLSI semiconductor electronics, may influence future commercial utilization of electro-optic materials, which is currently dominated by lithium niobate.

Table 1 Comparison of electro-optic performance for inorganic and organic EO materials at 1.3 microns wavelength

Parameter	*$LiNbO_3$*	*NLO polymer*
Effective EO coefficient, r_{eff}, pm/V	32	182
Index of refraction (n)	2.2	1.6–1.7
Dielectric permittivity (ε)	30	3
Material Figure of Merit, $n^3(r_{\text{eff}})/\varepsilon$	10	270
Bandwidth × length product, GHz cm	7	>100
$V_\pi L$, V cm	5	1.5
Optical loss, dB/cm	0.2	0.2–1.0
Processing temperature (°C)	1000	<200
Multiple layers possible	No	Yes

Devices

Representative simple device structures are shown in Figure 3. For a simple stripline phase (birefringence) modulator (Figure 3, upper left), the phase retardation produced by application of an electric field is given by

$$\Delta\phi = 2\pi\Delta nL/\lambda = \pi n^3 rEL/\lambda = \pi n^3 rVL/\lambda d \qquad [18]$$

where L is the electrode length (the distance over which the electrical and optical fields co-propagate),

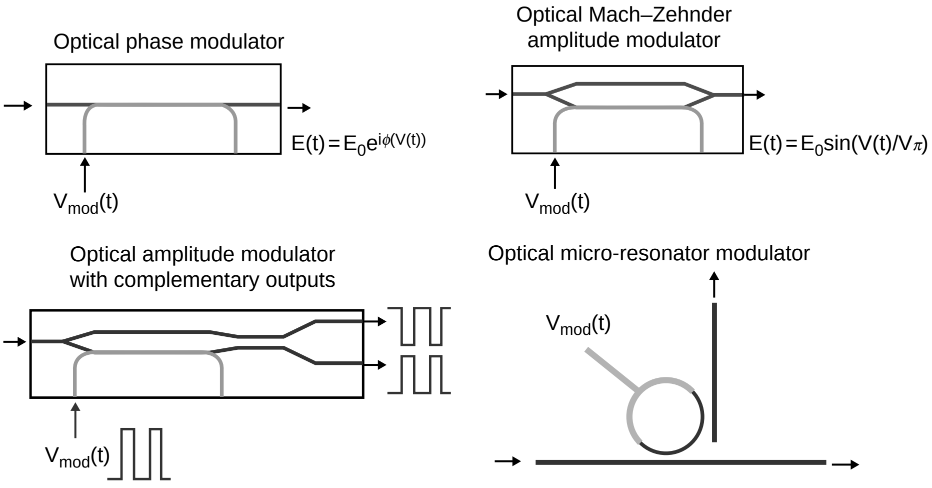
Optical phase modulator
E(t) = E0e^iφ(V(t))
Vmod(t)
Optical Mach–Zehnder amplitude modulator
E(t) = E0sin(V(t)/Vπ)
Vmod(t)
Optical amplitude modulator with complementary outputs
Vmod(t)
Optical micro-resonator modulator
Vmod(t)

Figure 3 Schematic representation of simple electro-optic device structures are shown: Upper left (birefringent modulator), upper right (Mach–Zehnder interferometer), lower left (1 × 2 optical switch), lower right (ring microresonator).

λ is the optical wavelength, $V = E/d$ is the electric field felt by the material, and d is the electrode spacing. Electro-optic device performance is frequently characterized by the voltage, V_π, required to produce a phase shift of π radians. For a TM polarized light propagating through a birefringent modulator:

$$V_\pi = \lambda d / (n_{0e}^3 r_{33} L \Gamma) \qquad [19]$$

where Γ characterizes the overlap of the optical and electrical waves.

A simple Mach–Zehnder interferometer (electro-to-optical signal transducer) is shown in the upper right of Figure 3. When such a device is constructed from electrically poled organic electro-optic chromophores and a voltage is applied to one arm, a retardation of light propagating in that arm will be effected. This, in turn, will result in a change in the constructive/destructive interference of the beams at the output of the device. The consequence of applying an electrical signal to one arm of a Mach–Zehnder interferometer is that the applied voltage is transduced onto the optical output beam as an amplitude modulation of that beam. The voltage required to achieve full wave modulation for an optical beam of TM polarization is

$$V_\pi = \lambda d / (n_{0e}^3 r_{33} L \Gamma) \qquad [20]$$

Amplitude modulation can also be achieved using a birefringent modulator by placing polarizers at the input and output of the device. If an input polarizer is used to launch TM and TE modes, the birefringent modulator functions to produce a rotation of light, which is turned into an amplitude modulation by inserting a polarizer at the output. For such a device, V_π depends upon the difference in r_{33} and r_{13}. Note that for poled organic chromophores, $r_{33} \approx 3r_{13}$, $(V_\pi)_{\mathrm{BRM}} \approx 1.5(V_\pi)_{\mathrm{MZM}}$. For the above devices, it is clear that V_π will depend on the length over which the electrical and optical fields co-propagate in phase. For low electrical field frequencies, L will be the electrode length. The electrode spacing d, that can be used, will depend on the index of refraction difference between the core (electro-optic material) and a cladding material used to prevent the optical field experiencing the metal electrodes (an event that would dramatically increase propagation loss). For poled organic macromolecular materials and standard commercial claddings, d is typically in the order of 10 microns. Electrode lengths usually range from 1 to 3 cm. Two factors limit bandwidth in stripline devices: (i) velocity mismatch between the electrical and optical waves (which relates to the difference between n_0^2 and ε, where ε is the dielectric permittivity of the material – see Table 1); and (ii) microwave and millimeter wave loss in the metal electrodes. For organic electro-optic materials, the latter defines operational 3 dB bandwidth of devices.

Optical routing switches (see Figure 3, lower left) are more complicated to understand and typically require somewhat larger V_π voltages to effect a switching operation (e.g., $(V_\pi)_{\mathrm{simple2\times2crossbar}} \approx 1.7(V_\pi)_{\mathrm{MZM}}$).

Two conditions must be satisfied for light to be coupled into a ring microresonator device (see Figure 3, lower right): (i) the circumference of the ring must be a multiple of the wavelength of light, and (ii) the velocity of light in the ring must equal the velocity of light at the input. If a ring microresonator is fabricated from an electro-optic material, then an applied electrical voltage can be used to control coupling of light into and out of the ring microresonator. The same electrode structure, that is used to achieve voltage-controlled wavelength selective filtering, can also be used to transduce electrical information onto the optical beam while it is in the ring resonator. The quality (Q) factor of a ring microresonator defines both the wavelength selectivity and the bandwidth of the device. Ring microresonator devices also have the advantage of permitting a long interaction length to be achieved in a very compact device structure. The dimensions (radii) of ring microresonator structures are defined by bending loss and thus by the difference between core and cladding materials. Typical radii can range from hundreds of nanometers to hundreds of microns. With ring microresonators, a reduction in drive voltage requirements can be achieved by trading off bandwidth. Another price paid in the use of ring microresonator device structures is that of bending loss, which is not present in the other device structure shown. Ring microresonators can be used for active wavelength division multiplexing (WDM), voltage-controlled optical signal routing, and wavelength tuning of optical sources.

The interaction length of a prism optical beam deflector can be increased by cascading prisms together. For such a device structure, the angle of deflection becomes

$$\theta = n_0^3 r_{33} (V/d)(L/h) \qquad [21]$$

where h is the height of an individual prism and L is the length of the prism array. For currently available electro-optic materials only small deflection angles (a few degrees or less) can be achieved with practical voltage levels. Superprism structures may permit large deflection angles to be achieved but such devices have yet to be demonstrated in a convincing manner.

See also

Interferometry: Overview. **Modulators:** Acousto-Optics. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Dagli N (1999) Wide-bandwidth lasers and modulators for RF photonics. *IEEE Trans. Microwave Theor. Techniques* 47: 1151–1171.

Dalton LR (2003) Rational design of organic electro-optic materials. *Journal of Physical Condensed Materials* 15: R897–R934.

Dalton L, Harper A, Ren A, *et al.* (1999) Polymeric electro-optic modulators: from chromophore design to integration with semiconductor very large scale integration electronics and silica fiber optics. *Ind. Eng. Chem. Res.* 38: 8–33.

Hornak LA (1992) *Polymers for Lightwave and Integrated Optics*. New York: Marcel Dekker.

Khazaei HR, Wang WJ, Berolo E and Ghannouchi F (1998) High speed GaAs/AlGaAs traveling wave electro-optic modulators. *Proceedings of SPIE* 3278: 94–100.

Lawetz C, Cartledge JC, Rolland C and Yu J (1997) Modulation characteristics of semiconductor mach-zehnder optical modulators. *IEEE Journal of Lightwave Technology* 15: 697–702.

Lee KS (2002). *Advances in Polymer Science*, vol. 158. Heidelberg: Springer.

Lee M, Katz HE, Erben C, *et al.* (2002) Broadband modulation of light using an electro-optic polymer. *Science* 298: 1401–1403.

Nalwa HS and Miyata S (1997) *Nonlinear Optics of Organic Molecules and Polymers*. Boca Raton, FL: CRC Press.

Noguchi K, Mitomi O and Miyazawa H (1998) Millimeter-wave Ti:LiNbO$_3$ optical modulators. *IEEE Journal of Lightwave Technology* 16: 615–619.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Shi Y, Zhang C, Zhang H, *et al.* (2000) Low (sub-1 volt) halfwave voltage polymeric electro-optic modulators achieved by controlling chromophore shape. *Science* 288: 119–122.

Wise DL, Wnek GE, Trantolo DJ, Cooper TM and Gresser JD (1998) *Electrical and Optical Polymer Systems. Fundamentals, Methods and Applications*. New York: Marcel Dekker.

Yariv A and Yeh P (1984) *Optical Waves in Crystals*. New York: John Wiley & Sons.

Modulation and Demodulation of Optical Signals

R A Minasian, University of Sydney, Sydney, NSW, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since earliest times, light has been used for communication purposes. In order to convey information, a physical variable of light, such as the optical intensity or field amplitude, must be modulated to encode a signal onto the optical carrier, and it must be demodulated or decoded after transmission. Laser sources provide the capability of extremely high modulation speeds, which result in high information-carrying capacity in optical communication systems. After transmission, semiconductor photodetectors demodulate the optical signals to recover the original information (*see* **Optical Communication Systems:** Basic Concepts).

The modulation of optical signals can be classified into two types: field modulation and intensity modulation, which is illustrated in Figure 1. In field modulation, the physical variable that is modulated is the frequency, phase, or amplitude of the optical field. These three techniques are sometimes referred to as frequency shift keying (FSK), phase shift keying (PSK) and amplitude shift keying (ASK). Figure 1b shows an example of frequency modulation of light. Field modulation requires keeping track of the phase of the carrier, which imposes strict requirements on the optical coherence (*see* **Coherent Lightwave Systems**). Because the optical carrier is at an extremely high frequency of around 2×10^{14} Hz at fiber-optic communication wavelengths of 1.5 μm, the required optical coherence is very difficult to realize, and moreover the receiver requires complex optical heterodyne methods to detect the phase of the optical field.

In intensity modulation, the physical variable that is modulated is the optical power. Figure 1c shows that it is the envelope of the light power that contains the information. The optical field oscillations at 10^{14}–10^{16} Hz in this case are not related to the

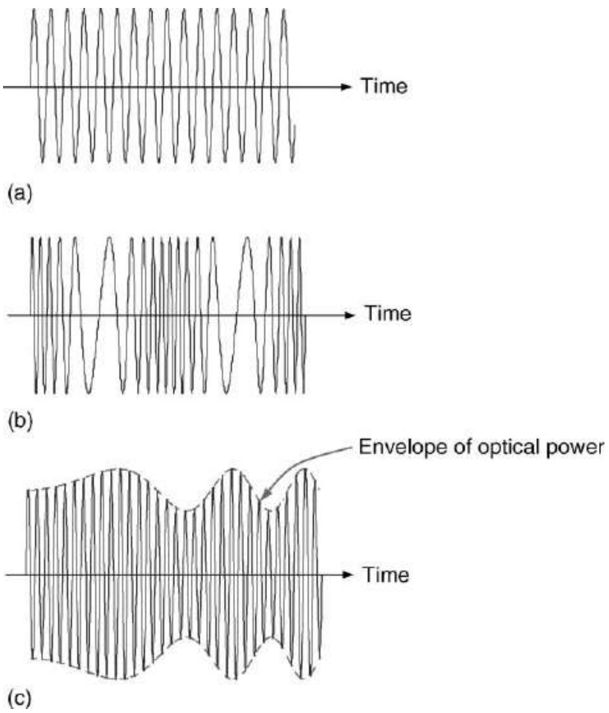

Figure 1 Types of modulation of optical signals. (a) unmodulated optical field; (b) frequency-modulated field; (c) intensity or envelope modulation of the optical power.

modulation and demodulation process, and only the optical power variation is important. This is a key advantage, and as a result intensity modulation systems are prevalent for optical communications. Here, we discuss how varying the injected current in a semiconductor laser or light emitting diode, or by operating the source with constant output and introducing an external modulator, can realize modulation of the optical power of the source. The process by which the optical power variations can be detected at the receiver with semiconductor photodiodes are also discussed, including detection noise characteristics and means of enhancing the responsivity by introducing internal gain mechanisms or wavelength selectivity. The speed capabilities of optical modulation and demodulation in semiconductor lasers and photodetectors are discussed, based on the rate equations for carrier and photon dynamics.

The Modulation of Semiconductor Lasers

Townes, Basov and Prokhorov, who shared the Nobel Prize in 1964, developed the principle of lasers. Semiconductor lasers emit light through a stimulated emission mechanism. The required optical gain for the lasing action is obtained by injecting a forward bias current into a heavily doped p-n junction, thereby injecting carriers into the active junction layer and creating a population inversion. Here we limit our discussion to the optical modulation characteristics of the laser in response to changes in the injection current.

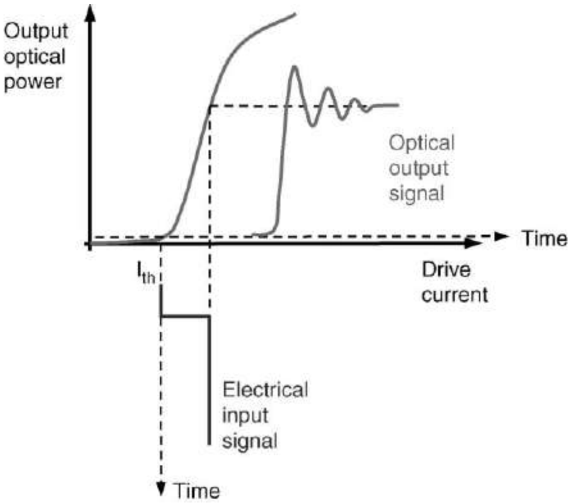

Figure 2 Light-current properties and the dynamic modulation characteristics of semiconductor lasers.

Figure 2 shows the typical light-current characteristics of semiconductor lasers. For drive currents above the threshold value I_{th}, the emitted optical power is predominantly by stimulated emission and it increases linearly with the drive current up to moderate currents. The differential responsivity of the laser, defined by the ratio of optical power increase to electric current increase above threshold, is typically around 0.5 mW/mA. At high current levels, the light output deviates from a proportional dependence on the current because of saturation effects and the differential responsivity decreases.

The dynamic response of the semiconductor laser to modulation of the injection drive current is governed by the rate equations for the carrier density and the photon density in the active junction region. Figure 3 shows the active layer under consideration (*see* **Lasers:** Edge Emitters). The electron concentration in the active layer is denoted N and the photon concentration in the active layer is denoted S. The dynamic behavior of the electron concentration and photon density is governed by the following rate equations:

$$\frac{\partial N(x)}{\partial t} = \frac{I(x)}{q v_a} + D_n \frac{\partial^2 N(x)}{\partial x^2} - \frac{N(x)}{\tau_{ns}} - G(x)\cdot S(x) \qquad [1]$$

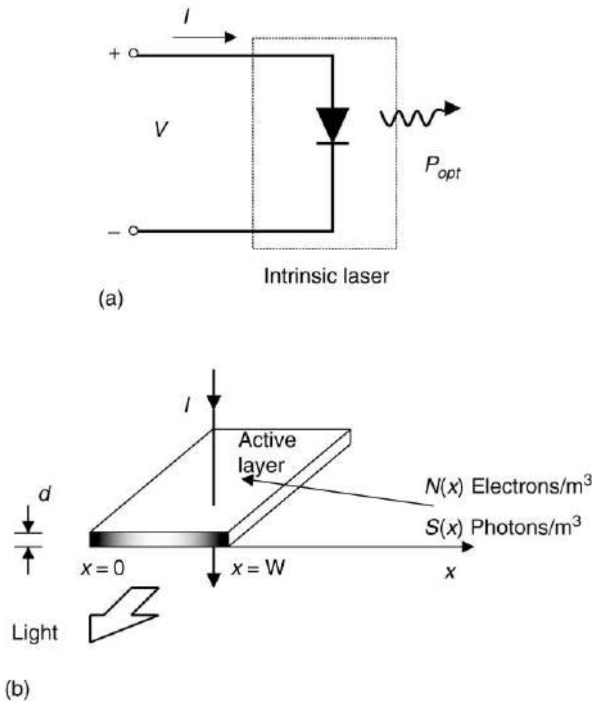

Figure 3 Semiconductor laser. (a) intrinsic device; (b) active region with injected current I, electron concentration N, and photon concentration S.

$$\frac{\partial S(x)}{\partial t} = \Gamma\cdot\beta\frac{N(x)}{\tau_{ns}} + \Gamma\cdot G(x)\cdot S(x) - \frac{S(x)}{\tau_p} \quad [2]$$

where I is the injected current, q is the electronic charge, v_a is the volume of the active layer, D_n is the diffusion constant, τ_{ns} is the carrier lifetime, Γ is the confinement factor (defined as the fraction of the optical power in the active region), β is the fraction of the spontaneous emission power coupled into the lasing mode, τ_p is the cavity or photon lifetime, and

$$G(x) = \gamma[N(x) - N_t] \quad [3]$$

where G is the net rate of stimulated emission, and N_t is the electron concentration at transparency.

Equations [1] and [2] are statements of the fundamental continuity equations for electrons and photons in the active layer. Hence, eqn [1] expresses the fact that the rate of increase of electron concentration in the active layer is given by the rate of increase of electron concentration due to injected current, the rate of increase of electron concentration due to diffusion current, the rate of decrease of electron concentration due to spontaneous recombination, and the rate of decrease in electron concentration due to stimulated photon emission. Equation [2] expresses the fact that the rate of increase of photon density in the active layer is given by the rate of increase in photon density due to spontaneous emission coupled into the lasing mode, the rate of increase of photon density due to stimulated emission coupled into the active layer, and the rate of decrease of photon density due to cavity loss mechanisms.

The actual electron and photon distribution varies across the active layer. However, for simplicity, it is convenient to use average values of N and S, and the simplified rate equations in terms of position independent variables are given by:

$$\frac{dN}{dt} = \frac{I_i}{qv_a} - \frac{N}{\tau_{ns}} - G(1 - \varepsilon S)S \quad [4]$$

$$\frac{dS}{dt} = \Gamma\cdot\beta\frac{N}{\tau_{ns}} + \Gamma\cdot G(1 - \varepsilon S)S - \frac{S}{\tau_p} \quad [5]$$

where the scaling factor $(1 - \varepsilon S)$ accounts for gain saturation arising from several phenomena such as spatial hole burning, spectral hole burning, carrier heating and two-photon absorption, and ε is a nonlinear gain parameter.

The modulation response of the laser is obtained by solving the rate eqns [4] and [5] for a time dependent input current of the form

$$I(t) = I_b + I_m x(t) \quad [6]$$

where I_b is the steady-state bias current, I_m is the modulation current amplitude and $x(t)$ represents the normalized shape of the modulation waveform.

The small-signal modulation bandwidth of the laser is obtained by considering the response to small amplitude sinusoidal current modulation at frequency ω, so that $x(t) = \sin(\omega t)$, around a steady-state bias current above lasing threshold. Under the small signal condition, the rate equations can be linearized and combined to result in a second order differential equation. The solution for the photon output exhibits a damped resonant type response. Hence the modulation response displays some peaking at the relaxation oscillation resonance frequency ω_r, and is flat for frequencies much lower than ω_r. The resonance frequency $\omega_r \propto \sqrt{S_o}$, where S_o is the photon density at the laser bias level, so the bandwidth can be increased by biasing the laser further above its threshold current. Semiconductor laser bandwidth to 30 GHz has been observed; however, it is often limited to around 10 GHz due to electrical parasitics in the current drive circuit and interface.

The large-signal modulation response of the laser is obtained by solving the differential eqns [4] and [5]. The response of the carrier-photon coupled system to a step change of drive current has the potential for damped oscillation. This is caused by a resonant type of behavior in which, for example, an increase in photon density causes a depletion in population inversion which leads to a decrease in photon density enabling the carrier density to recover, thus producing a time lag between cause and effect resulting in damped oscillations before settling to the steady state value. These dynamic characteristics are illustrated in Figure 2. For a step modulation of the laser drive current, the optical output response exhibits a delay time, a risetime, an overshoot, and damped oscillations. The risetime of the laser output is typically in the range of 100–300 ps. The turn-on time of the laser and the overshoot are improved by ensuring that the minimum drive current does not fall below threshold. The damped oscillations in the optical output occur at the relaxation oscillation frequency of the laser.

These laser pulse waveform imperfections cause deteriorations in the eye diagram of the laser transmitter that is modulated by very high speed digital bit streams (*see* **Optical Communication Systems:** Lightwave Transmitters). Figure 4 shows the eye diagram for the optical output of a semiconductor laser directly modulated by a 5 Gbit/s pseudo-random bit stream of length $2^{15}-1$. The output waveform imperfections cause a partial closure of the eye, which ultimately limits the modulation speed that can be obtained by directly modulating the laser drive current. Although the semiconductor laser optical output does not exactly replicate the applied electrical drive waveform, the deviations are small enough so that these lasers can be used for data transmission up to around 10 Gbit/s.

Another inherent consequence of direct modulation of semiconductor lasers is the production of laser frequency chirp. This arises from the fact that modulation induced changes in the carrier density in the active region of the laser cavity cause a change in the refractive index of the active layer, which results in a chirping in the output wavelength during modulation. Hence amplitude modulation of semiconductor lasers is fundamentally accompanied by phase modulation from the basic physical mechanism of carrier induced changes in the refractive index. The resulting transient changes in the output wavelength of the laser causes a chirping phenomenon as it is being modulated. The order of chirp magnitude is about 0.1–1 GHz/mA of drive current change. During digital pulse modulation, the laser wavelength shifts to the blue side near the leading edge and towards the red side near the trailing edge of the pulse. Chirp results in a broadening of the laser linewidth. This wavelength change effect in combination with the dispersion characteristics of optical fiber (*see* **Fiber and Guided Wave Optics:** Dispersion; Light Propagation) causes an important limitation to the transmission capacity of fiber optic communication systems. The chirp in semiconductor lasers can be reduced, but not entirely eliminated, by using modulation-doped strained multiquantum-well lasers, which reduce the linewidth enhancement factor.

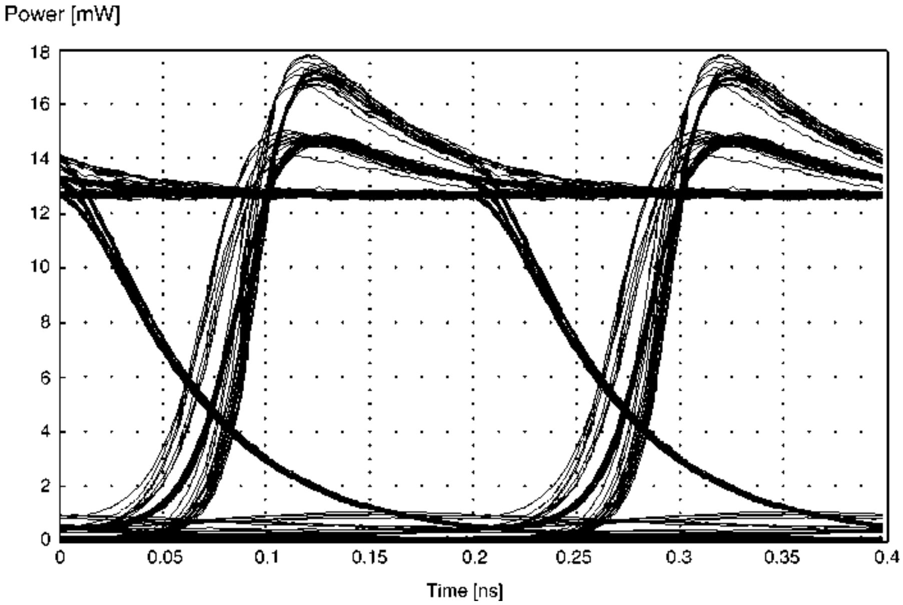

Figure 4 Eye diagram for the optical output of a semiconductor laser directly modulated by a 5 Gbit/s pseudo-random bit stream.

The Modulation of Light Emitting Diodes

The light emitting diode (LED) generates light through spontaneous emission. A semiconductor p-n junction is pumped with a forward current, so that electrons are provided in the upper energy state that can recombine radiatively to produce incoherent light by a spontaneous emission mechanism. The rate of photon generation is directly proportional to the carrier injection rate, and hence the modulating drive current.

Figure 5 shows the typical light-current properties and the modulation characteristics of a LED. The responsivity of the LED, defined as the ratio of the emitted optical power to the injected current, is typically in the range of 10 to 50 μW/mA. The optical output power generated is linearly related to the injection (drive) current up to moderate currents, and is given by

$$P = \eta \cdot h\nu \cdot \frac{I}{q} \qquad [7]$$

where η is the emission quantum efficiency (defined as the ratio of the radiative recombination rate to the total recombination rate), h is Planck's constant, ν is the optical frequency, I is the injected current, and q is the electronic charge. For larger currents, the light output deviates from a proportional dependence on the current because of saturation effects and the responsivity decreases.

The modulation response of the LED is governed by the rate equation for the carrier density N

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{I}{qv_a} - \frac{N}{\tau_c} \qquad [8]$$

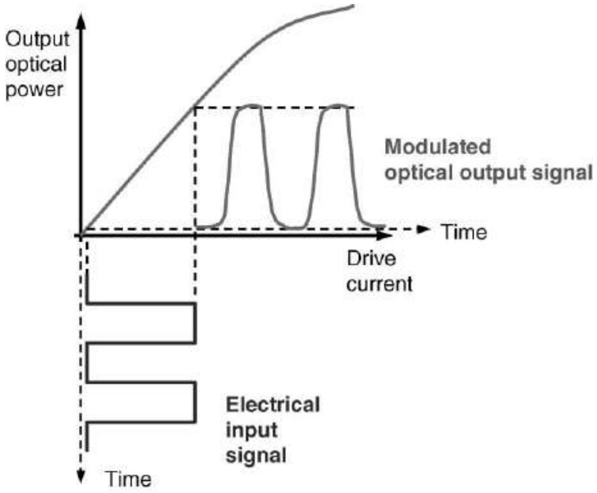

Figure 5 Light-current properties and the modulation characteristics of light emitting diodes.

where I is the injected current, q is the electronic charge, v_a is the volume of the active region, and τ_c is the carrier lifetime. Equation [8] expresses the fact that the rate of increase of carrier concentration is given by the rate of increase of carriers through current injection and the rate of carrier loss through radiative and nonradiative recombination. The small signal response arising from this first order differential equation, exhibits a simple single pole response and the resulting modulation bandwidth of the LED is given by

$$B = 1/(2\pi\tau_c) \qquad [9]$$

The pulse modulation response of the LED is shown in Figure 5. The output modulated optical waveform is not square but exhibits risetimes and falltimes determined by the carrier lifetime τ_c. This sets a limit to the maximum modulation rate that the LED can achieve. The bandwidth can be increased by decreasing the radiative lifetime, which can be controlled to a limited extent by the choice of doping level. Typical lifetimes for InGaAsP LEDs are in the range of 1–5 ns, which corresponds to attainable modulation data rates of several hundred Mbit/s.

The External Modulation of Lasers

Modulation of lasers realized by changing the injection current, as described above, is classed as direct modulation. This is the simplest method of modulation; however, it is limited by several physical phenomena. These include bandwidth limitations due to carrier dynamics in the laser active region, waveform limitations due to relaxation oscillations and overshoot effects, and chirping which causes transmission limitations in optical fibers arising from dispersion effects.

A superior modulation technique, classed as external modulation, operates the laser with continuous wave (CW) output and impresses the modulation by means of a device external to the laser. Although more complex, this approach has many modulation performance advantages. Because the laser is operated CW, this technique removes chirping effects and also enables higher power lasers to be used. Moreover, it is capable of much higher modulation bandwidths, larger on–off extinction ratios, and superior modulation spectral purity.

Electro-optic Modulation of Light

The linear electro-optic effect in materials was first described by Friedrich Pockels in 1893. The Pockels effect is the change in refractive index of the material resulting from the application of an electric field, and is given by

$$n(E) = n - \frac{1}{2} r n^3 E \qquad [10]$$

where n is the refractive index, E is the applied electric field, and r is the electro-optic coefficient. Typical values of r lie in the range 10^{-12} to 10^{-10} m/V, and the most common electro-optic materials used include LiNbO$_3$, GaAs, InP, InGaAsP, and synthetic polymers such as polymethylmethacrylate (*see* **Modulators:** Electro-optics).

If light traverses an optical waveguide of length L, to which an electric field E is applied, it undergoes a phase shift given by

$$\Delta\phi = -\frac{\pi r n^3 E L}{\lambda} \qquad [11]$$

The electric field can be produced by applying a voltage V across the waveguide of width d, hence $E = V/d$. It is therefore possible to modulate the phase of the light by varying the voltage V that is applied to the material through which the light passes. The voltage required to produce a phase shift of π is given by

$$V_\pi = \frac{\lambda d}{n^3 r L} \qquad [12]$$

The parameter V_π, known as the half-wave voltage, is an important characteristic that expresses the applied voltage required to give a phase shift of π.

Intensity modulation of light can be obtained by placing the phase modulator in one arm of an optical interferometer. The structure, known as a Mach-Zehnder modulator, is shown in Figure 6a. The optical output is the phasor sum of two lightwaves that travel through the two arms, resulting in constructive or destructive interference depending on the optical phase difference. The electric field is applied to the optical waveguide in a direction perpendicular to the direction of light propagation using the metal electrodes, see Figure 6a. Hence the refractive index of the waveguide and thus the phase of the light can be modulated by varying the voltage V. When no voltage is applied, the optical fields in the two arms of the interferometer experience identical phase shifts, and hence interfere constructively at the output. When a voltage is applied, the waveguide index changes, and the light in that arm experiences a phase shift with respect to the light in the other arm. This destroys the constructive nature of the interference at the output, and reduces the transmitted light emitted at the output. The extreme case is when the phase difference between the two arms equals π, which causes complete destructive interference resulting in no light transmission.

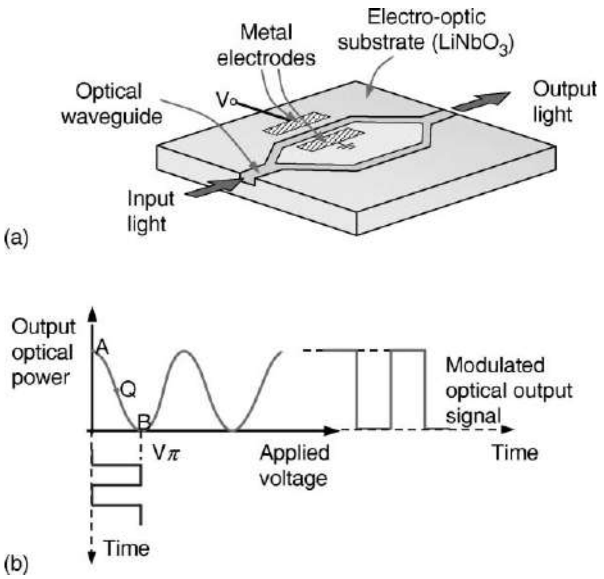

Figure 6 Mach-Zehnder electro-optic intensity modulator. (a) Structure; (b) transfer function and modulation characteristics.

The modulation transfer function is given by

$$P_{\text{out}} = \frac{1}{2} P_{\text{in}} [1 + \cos(\Delta\phi)] \qquad [13]$$

where $\Delta\phi$ is the phase difference encountered by light as it travels through the two arms. In terms of the applied voltage, the optical power transmitted through the modulator is given by

$$P_{\text{out}} = \frac{1}{2} P_{\text{in}} \left[1 + \cos\left(\pi \frac{V}{V_\pi} \right) \right] \qquad [14]$$

This transfer characteristic is shown in Figure 6b. The modulator may be operated as an analog optical modulator by operating in the nearly linear region around point Q. Alternately, it may be operated as a digital modulator by operating between points A and B, thus switching the light on and off as V is switched between 0 and V_π.

Improved efficiency of modulation, and hence a reduction in the value of V_π (typically a few volts) can be obtained by placing a second electrode on the other arm of the interferometer, and using a push-pull configuration to drive the interferometer. Hence the phase modulators in both arms are driven with modulating voltages of opposite polarities so the phase is advanced in one arm and is retarded in the other arm. This produces a net phase difference, which is double that produced in either arm.

The Pockels effect is an extremely fast physical mechanism, hence in principle external modulation is capable of achieving extremely high speeds. In practice, the modulation speed of an electro-optic modulator is limited by the capacitive effects at the electrodes and by the transit time of the light through the material. If the electric field varies significantly within the time it takes for light to transit through the waveguide, then the travelling optical wave will be subjected to different electric fields as it traverses the waveguide, and this results in a transit time limitation to the modulation bandwidth. A technique to avoid this limitation is to apply the voltage at one end of the waveguide and to design the electrodes to form a transmission line. By this means it is possible to make the velocity of the electrical wave match that of the optical wave to essentially eliminate the transit time effects. These external modulators have demonstrated speeds well in excess of 100 GHz.

Electro-absorption Modulation of Light

Electro-absorption external modulators are based on the principle of semiconductor bandgap decrease that occurs upon application of an electric field (the Franz-Keldysh and Stark effect). With no applied voltage, the semiconductor waveguide is transparent to light having a photon energy smaller than the bandgap, and such wavelengths are transmitted with very little attenuation. When a voltage is applied, the bandgap of the semiconductor decreases and the material absorbs the light once the photon energy exceeds the bandgap energy, and this produces optical attenuation. Hence by applying a modulation voltage to the semiconductor waveguide, the absorption property of the waveguide and hence the output light intensity can be modulated.

Improved modulation efficiency, and hence a reduction in the voltage required to decrease the bandgap energy, is obtained by using multiquantum-well semiconductor structures, which exhibit enhanced electro-absorption effects arising from the quantum confined Stark effect. The conventional technique of applying a transverse field across a multiquantum-well is to apply a reverse voltage across it. Modulator drive voltages of a few volts can be obtained. The on/off extinction ratio of modulation depends on the value of the field-dependent absorption coefficient at the operating voltage, and typical on/off extinction ratios are 10–15 dB.

The modulation bandwidth is determined by intrinsic factors related to carrier dynamics, such as exciton formation and dissociation times and the time taken to remove electrons and holes from the wells by thermionic emission and tunelling. It is also determined by extrinsic factors related to the capacitance of the junction diode. These modulators can have bandwidths of around 50 GHz. They also enable low chirp modulation to be obtained. The optical power handling capability of electro-absorption modulators is lower than for electro-optic modulators. However, a principal advantage of the electro-absorption modulator is that the semiconductor waveguide can be fabricated onto one substrate with the semiconductor laser, enabling the realization of monolithic integration on the same chip.

The Demodulation of Optical Signals

The process of demodulation converts the modulated intensity of an optical wave into an electrical signal to recover the information content. Semiconductor photodetectors are based on the principle of absorption of photons to generate mobile electron-hole pairs, and the transport of these mobile carriers via an applied electric field to create a flow of current.

Demodulation using PIN Photodiodes

The PIN photodiode comprises a P-doped/intrinsic/N-doped semiconductor structure. An applied reverse voltage bias creates a depletion region with a high electric field in the intrinsic layer. When the intrinsic layer is illuminated with light, photons are absorbed and electron-hole pairs are created, which are swept by the built-in electric field. The resulting flow of current is proportional to the incident optical power.

The quantum efficiency of the photodetector is defined as the ratio of electron generation rate to the photon incidence rate, and is given by

$$\eta = \frac{h\nu}{q} \cdot \frac{I_p}{P} \qquad [15]$$

where I_p is the photocurrent, P is the incident optical power, ν is the optical frequency, h is Planck's constant and q is the electron charge. If the width of the absorbing intrinsic region is large enough to capture all the light, quantum efficiency values approaching 100% can be achieved.

The current-voltage characteristics of the photodiode under optical illumination are shown in Figure 7a. The dark current I_d, which is the current generated in the photodetector in the absence of light originates from thermally generated electron-hole pairs, and is typically extremely small (<5 nA). Figure 7b shows the transfer characteristic of the

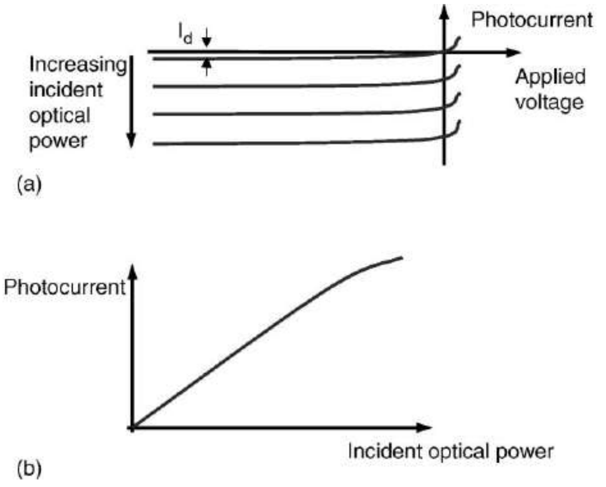

Figure 7 Semiconductor photodiode. (a) Current-light characteristics under optical illumination; (b) transfer function and demodulation photocurrent versus incident optical power.

photodetector. The current is linearly proportional to the optical power up to moderate power levels, when saturation mechanisms occur. The responsivity of the photodiode $\Re$, defined as the ratio of the generated electric current to the incident optical power, is given by

$$\Re = \frac{\eta q}{h\nu} \qquad [16]$$

The responsivity is an important parameter, and provides a measure of the photodetection efficiency of the device for demodulating optical signals. Typical responsivity values for PIN photodiodes are around 1 A/W.

The bandwidth B of demodulation of the photodetector is determined by the speed with which it responds to variations in the incident optical power. This is limited by the transit time τ_{tr} and the RC time constant τ_{RC} and is given by

$$B = \frac{2\pi}{\tau_{tr} + \tau_{\mathrm{RC}}} \qquad [17]$$

The transit time limitation arises because of the time it takes for the generated carriers to traverse the absorption region and be collected. To maximize the bandwidth, the transit time needs to be reduced, which can be achieved by decreasing the width of the absorption region. However, this requires a trade-off with the responsivity, which drops because a smaller fraction of the light can be absorbed under these conditions. The capacitance C and the resistance R of the photodetector and its circuitry form the RC time constant. To maximize the bandwidth, the photodiode capacitance needs to be reduced by decreasing the area of the photodiode. The demodulation bandwidth of PIN photodiodes can exceed 100 GHz.

Photodetection shot noise is a fundamental noise mechanism that introduces noise in the demodulation process of optical signals. This leads to fluctuations in the photocurrent, even for a constant incident optical power, because the photocurrent actually consists of a stream of electrons that are generated at random times. The photodiode current generated by a constant incident optical power is given by

$$I(t) = I_{\mathrm{avg}} + i_{\mathrm{shot}}(t) \qquad [18]$$

where I_{avg} is the average photocurrent generated, and $i_{\mathrm{shot}}(t)$ is the current fluctuation due to shot noise and which is described by a stationary random process with Poisson statistics. The mean-square noise current is given by

$$< i^2_{\mathrm{shot}}(t) >= 2qI_{\mathrm{avg}}\Delta f \qquad [19]$$

where Δf is the effective noise bandwidth of the receiver. The shot noise degrades the signal to noise ratio of the demodulated signal, and causes a fundamental limit to the sensitivity of the PIN photodetection.

Demodulation using Avalanche Photodiodes

Demodulation using avalanche photodiodes detectors (APD) has the advantage of significantly increased responsivity. This is achieved by providing an internal current gain mechanism within the device. The APD operates by generating more than one electron-hole pair per absorbed photon. This is achieved by strongly reverse biasing the diode to create a large junction electric field, so photoelectrons accelerate and acquire enough energy to produce new carriers by the process of impact ionization. These devices incorporate a modification to the PIN structure, by including an additional doped layer known as the multiplication layer, which under reverse bias provides a region of very high electric field that can produce avalanche breakdown. Hence a single primary electron generated through absorption of a photon, generates many secondary electrons and holes in the avalanche multiplication layer, and all these secondary carriers contribute to the photocurrent. As a result of this current gain mechanism, the responsivity of the APD is increased by a multiplication factor M, and is given by

$$\Re_{ADP} = \frac{M\eta q}{h\nu} \qquad [20]$$

where M can range between 5–50 depending on the bias voltage.

The bandwidth of the APD depends on the multiplication factor M. This is because the avalanche process takes additional time. The frequency response can be written

$$M(\omega) = \frac{M_0}{[1 + (\omega\tau_e M_0)^2]^{\frac{1}{2}}} \qquad [21]$$

where M_0 is the low frequency gain and τ_e is the effective transit time that depends on the ionization coefficient. The intrinsic APD bandwidth (assuming no RC limitations) is approximately given by

$$B_{\mathrm{APD}} = \frac{1}{2\pi\tau_e M_0} \qquad [22]$$

Hence the bandwidth decreases as the multiplication factor increases, and there exists a trade-off between the bandwidth and the responsivity. The demodulation characteristics of the avalanche photodiode can be expressed by means of a gain-bandwidth product $B_{\mathrm{APD}} \cdot M_0$, and this can exceed 150 GHz.

Photodetection noise in APDs has an additional contribution because the avalanche gain itself is a random variable. Hence the secondary electron-hole pairs are generated at random times and this adds a noise component to the basic shot noise associated with the generation of primary electro-hole pairs. The total shot noise is given by

$$< i^2_{\mathrm{shot,APD}}(t) >= 2qM^2F(M)I_{\mathrm{avg}}\Delta f \qquad [23]$$

where $I_{\mathrm{avg}} = \Re P$, and F is the excess noise factor expressed by

$$F = \frac{< M^2 >}{< M >^2} \qquad [24]$$

and is approximately given by $F \approx M^x$, where $0.2 < x < 1$. Hence the excess noise increases with the multiplication ratio, so there exists an optimum avalanche photodiode gain that maximizes the signal to noise ratio of the demodulated signal. Despite the fact that additional noise is introduced, because of the higher responsivity, APDs can provide higher sensitivity in demodulating optical signals than PIN photodiodes.

Demodulation using Metal-Semiconductor-Metal (MSM) Photodiodes

Demodulation using MSM photodiodes enables extremely high bandwidths to be achieved. The MSM photodiode comprises two reverse-biased metal-to-semiconductor Schottky contacts formed on a semiconductor layer that prevents flow of electrons from metal to semiconductor. Photons incident between the contacts create electron-hole pairs that are collected by the electric field to generate a photocurrent proportional to the incident optical power. The metal contacts are usually fabricated on the same side of the semiconductor to create a planar structure, which is suitable for monolithic integration. However, a drawback of the MSM device is that the electrodes block some of the light from reaching the absorption region, which results in a reduced responsivity.

The bandwidth of demodulation of the MSM photodiode is determined by the transit time of the photogenerated carriers and the RC time constant. Since the electrode spacing can be around 1 μm, the transit time is extremely small. Moreover, since in the planar structure the parasitic capacitance is inherently low, the RC time constant is also extremely small. This results in demodulation bandwidths of up to 300 GHz.

Wavelength Selective Demodulation

A technique to obtain both wavelength selective demodulation and high-speed, high-responsivity demodulation involves the use of microcavity resonant photodetectors. This is useful for very high-speed detection in wavelength multiplexed systems.

The microcavity photodiode comprises a Fabry-Perot cavity placed around the photodetector absorption region, to enhance the responsivity. The mirrors are typically formed using Bragg reflectors or dielectric mirrors implemented in the semiconductor. A resonance is built up in the cavity at the wavelengths of incoming light for which the round-trip phase shift is 2π, which corresponds to constructive interference. Hence at the resonant wavelength, the incoming light is reflected at the two mirrors and permits the absorbing layer to be reused many times by effectively recirculating the optical signal, thus enhancing the absorption and the responsivity. However, the transit length can remain extremely small, because the photogenerated carriers only have to traverse the thickness of the absorption region to be collected. This solves the problem of the fundamental responsivity-bandwidth trade-off in normal photodetectors, in which increasing the absorption layer width to increase responsivity increases the transit time thus reducing the bandwidth. Microcavity resonant photodiodes enable the simultaneous achievement of large bandwidth and high responsivity possible. Also, the resonant behavior leads to wavelength selectivity, which can be used to advantage in some wavelength division multiplexed WDM applications (*see* **Optical Communication Systems:** Wavelength Division Multiplexing). This approach can provide demodulation of optical signals with quantum efficiencies close to 100% and with bandwidths in excess of 100 GHz.

Optical Receivers – Preamplification and Clock Recovery

The front end of an optical receiver comprises the photodiode and a preamplifier to amplify the electrical signal for further processing. In order to obtain a high demodulation bandwidth, it is important to minimize the photodetector capacitance and to minimize the input resistance of the preamplifier that acts as the load resistance. However, because the thermal noise current is inversely proportional to the load resistance, minimizing the load resistance causes excessive noise and degraded demodulation sensitivity. An alternative approach is to use a high-impedance preamplifier. This reduces the noise at the price of a lower bandwidth. Thus there is a trade-off between the bandwidth and sensitivity. The trans-impedance front-end, in which the load resistor is connected as a feedback resistor around an inverting amplifier, is often used as an optimum compromise that combines high sensitivity together with large bandwidth. This enables the use of a large resistance, which minimizes noise, but at the same time the effect of negative feedback reduces the input impedance by a factor equal to the amplifier gain, thus increasing the bandwidth.

In digital optical transmission systems the data recovery section of the optical receiver comprises a clock recovery circuit and a decision circuit. Accurate timing (clock) information about the bit slot, in order to synchronize the decision process, is obtained from the amplified front-end waveform using a timing extraction circuit (*see* **All-Optical Signal Regeneration**). Typical clock recovery circuits comprise a high-Q resonant circuit which rings at the bit rate frequency or a phase locked loop. This enables precise data time sampling in the following decision circuit within the bit intervals of the original pulse train.

See also

All-Optical Signal Regeneration. **Fiber and Guided Wave Optics:** Dispersion; Light Propagation. **Lasers:** Edge Emitters. **Modulators:** Electro-Optics. **Optical Communication Systems:** Basic Concepts; Lightwave Transmitters; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*. New York: Wiley.

Coldren L and Corzine S (1995) *Diode Lasers and Photonic Integrated Circuits*. New York: Wiley.

Gowar J (1993) *Optical Communication Systems*. Englewood Cliffs: Prentice Hall.

Kazovsky L, Bendetto S and Willner AE (1996) *Optical Fiber Communication Systems*. Boston: Artech House.

Keiser GE (1991) *Optical Fiber Communications*. New York: McGraw-Hill.

Miller SE and Kaminow I (eds) (1988) *Optical Fiber Telecommunications II*. New York: Academic Press.

Personick SD (1985) *Fiber Optics: Technology and Applications*. New York: Plenum Press.

Petermann K (1991) *Laser Diode Modulation and Noise*. Boston: Kluwer Academic.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Senior J (1992) *Optical Fiber Communications*. NJ: Prentice Hall.

Siegman AE (1986) *Lasers*. Sausalito, CA: University Science Books.

N

NONCLASSICAL LIGHT

H Walther, University of Munich and
Max-Planck-Institute for Quantum Optics,
Garching, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The pioneering experiments by Hanbury-Brown and Twiss, and the subsequent development of the quantum theory of electromagnetic radiation by Glauber and Sudarshan, raised the interest in the question of photon statistics. It followed from their theories that states of the electromagnetic field which can only exist in a quantum treatment. These nonclassical states show up in photon statistics through the phenomenon of antibunching. Later, further nonclassical states of the radiation field were found in the so-called squeezed fields, which are fields with phase-sensitive quantum fluctuations being at certain phase angles smaller than those of a perfect classical coherent field. In order to achieve squeezing of laser radiation, various forms of nonlinear optical interactions can be used.

Classical Versus Quantum Waves

In classical electrodynamics, a monochromatic wave (e.g., an extended light beam) is often described using the complex representation:

$$\begin{aligned} E(r,t) &= E^{-}(r,t) + E^{+}(r,t) \\ &= \frac{1}{2}[\varepsilon(r,t)\mathrm{e}^{-\mathrm{i}\omega t} + \varepsilon^{*}(r,t)\mathrm{e}^{\mathrm{i}\omega t}] \end{aligned} \qquad [1]$$

In classical theory, this complex description is merely a tool for the simplification of some calculations. However, there is some physical significance when using the quantum description, where ε and ε^* lead to the lowering and raising operators of the field. The above field can also be represented in the following way:

$$E(r,t) = \frac{\varepsilon + \varepsilon^{*}}{2}\cos\omega t + \frac{\varepsilon - \varepsilon^{*}}{2}\sin\omega t \qquad [2]$$

with real coefficients $\varepsilon_1 = (\varepsilon + \varepsilon^*)/2$ and $\varepsilon_2 = (\varepsilon - \varepsilon^*)/2\mathrm{i}$. This description gives the state of the field in a two-dimensional phase diagram where the sin-component is plotted versus the cos-component. In this plot, the wave above is represented by one point giving phase and amplitude of the field. These phase diagrams will play an important role in characterizing the nonclassical fields in a quantum phase diagram, as discussed below.

The properties of a field are described by the first- and second-order correlation functions. The first-order correlation function is given by

$$G^{(1)}(r_1, t_1, r_2, t_2) = \langle \varepsilon^{*}(r_1, t_1)\varepsilon(r_2, t_2)\rangle \qquad [3]$$

It is obtained by multiplying the electric field amplitudes at two spacetime points and taking the ensemble average. For $r_1 = r_2$ and $t_1 = t_2$, $G^{(1)}$ is proportional to the intensity at this point and is, therefore, called the first-order in intensity. The first-order correlation function is connected to the spectrum $I(\omega)$ of the field through its Fourier transform.

The second-order correlation function is necessary to characterize the radiation field completely. It describes the intensity fluctuations, being the classical equivalent to photon statistics, and is given in the general form:

$$\begin{aligned} &G^{(2)}(r_1, t_1, r_2, t_2, r_3, t_3, r_4, t_4) \\ &\quad = \langle \varepsilon^{*}(r_1, t_1)\varepsilon^{*}(r_2, t_2)\varepsilon(r_3, t_3)\varepsilon(r_4, t_4)\rangle \end{aligned} \qquad [4]$$

If $G^{(2)}$ is considered for two spacetime points $r_1 = r_2$, the connection to intensity fluctuations becomes obvious:

$$G_{\mathrm{I}}^{(2)}(\tau) = 4\langle I(t_1)I(t_2)\rangle = \langle \varepsilon^{*}(t_1)\varepsilon(t_1)\varepsilon^{*}(t_2)\varepsilon(t_2)\rangle \qquad [5]$$

The subscript I indicates that we have to deal with intensity correlations; τ stands for $t_2 - t_1$.

In connection with the Hanbury-Brown, Twiss experiment, in general, the normalized intensity correlation is given by

$$g_I^{(2)}(\tau) = \langle I(t)I(t+\tau)\rangle / \langle I(t)\rangle^2 \qquad [6]$$

$I(t)$ stands for the intensity which is given by the mean value $\langle I(t)\rangle$ plus fluctuations $\Delta I(t)$ for which the mean value is zero. The mean value of the squared intensity is then $\langle I^2\rangle = \langle I\rangle^2 + \langle \Delta I^2\rangle$. It follows for the normalized intensity correlation function:

$$\begin{aligned} g_I^{(2)}(\tau) &= \langle I(t)I(t+\tau)\rangle / \langle I(t)\rangle^2 \\ &= \left\{ \begin{matrix} 1 + \langle \Delta I^2\rangle / \langle I\rangle^2 & \text{for } \tau = 0 \\ 1 & \text{for } \tau \to \infty \end{matrix} \right\} \end{aligned} \qquad [7]$$

For $\tau \to \infty$, the intensities are not correlated and $g_I^{(2)}(\tau)$ becomes 1. Otherwise $\langle \Delta I^2\rangle / \langle I\rangle^2 \geq 0$; this means that $g_I^{(2)}(\tau)$ is always equal to or larger than 1, and $g_I^{(2)}(\tau)$ smaller than 1 is not possible classically. The photon correlation is usually measured by the classical Hanbury-Brown, Twiss setup shown in Figure 1.

The maximum of $g_I^{(2)}(\tau)$ at $\tau = 0$ is called intensity bunching and was first observed by Hanbury-Brown and Twiss. The decrease of $g_I^{(2)}(\tau)$ as a function of τ is a measure of the coherence length of the light source.

When the experiment in Figure 1 is considered in the photon picture there is a principal difference compared to the classical intensity. In the classical case, the intensity on the two detectors is the same if the beamsplitter is dividing the light beam in equal portions. Both detectors, therefore, measure the same fluctuations. In the quantum picture, however, a photon can only be detected in one of the two detectors; this leads to the result that values for $g_I^{(2)}(\tau)$ smaller than 1 can be found. The effect of photon bunching was initially thought to be an inherent property of light owing to the boson nature of photons. Therefore, it came as a big surprise when Glauber showed theoretically that for the quantized radiation field $g_I^{(2)}(\tau)$, <1 is possible. This effect, known as photon antibunching, and its experimental observation constitutes an unambiguous proof of the quantum nature of light.

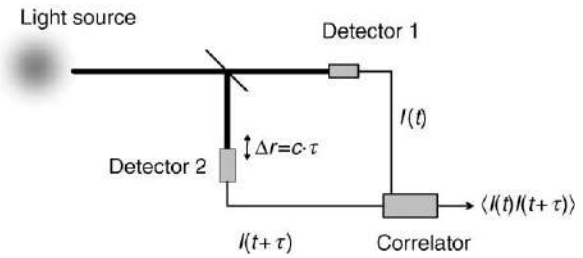

Figure 1 Experimental setup for the Hanbury-Brown, Twiss experiment.

In the quantum description of the electromagnetic field, the electric field amplitudes ε and ε^* are replaced by $\hat{a}$ and $\hat{a}^+$, the photon annihilation and creation operators.

In this case, the electric field operator is written as

$$\hat{E} \propto \frac{\hat{a} + \hat{a}^+}{2} \cos\omega t \frac{\hat{a} - \hat{a}^+}{2\mathrm{i}} \sin\omega t \qquad [8]$$

In analogy, the operators $\hat{a}_1 = (\hat{a} + \hat{a}^+)/2$ and $\hat{a}_2 = (\hat{a} - \hat{a}^+)/2\mathrm{i}$, can be defined describing the amplitude of the cosine and sine parts of the wave. $\hat{a}_1$ and $\hat{a}_2$ do not commute $[\hat{a}_1, \hat{a}_2] = (\mathrm{i}/2)1$, therefore, the expectation values of the two operators $\hat{a}_1$ and $\hat{a}_2$ underlie an uncertainty relation given by

$$\langle \Delta \hat{a}_1^2\rangle = \langle \hat{a}_2^2\rangle - \langle \Delta \hat{a}_i\rangle^2 \geq 1/16 \qquad [9]$$

with $\langle \Delta \hat{a}_i^2\rangle = \langle \hat{a}_i^2\rangle - \langle \Delta \hat{a}_i\rangle^2$. In the corresponding phase diagram, the expectation values for $\hat{a}_1$ and $\hat{a}_2$ are represented by a shaded area defined by the uncertainty (Figure 2). The state of the radiation field, where both phase and amplitude fluctuations are symmetrically minimized, $\langle \Delta \hat{a}_1^2\rangle = \langle \Delta \hat{a}_2^2\rangle = \frac{1}{4}$, best approximates the classical wave with well-defined phase and amplitude. The state with symmetrical minimum uncertainty is the coherent state describing laser radiation. It is an eigenstate of $\hat{a}$ with a photon number fluctuation given by a Poissonian distribution with $\langle \Delta n^2\rangle = \langle n\rangle$.

There are other minimum uncertainty states, however, with unsymmetrical uncertainty regions

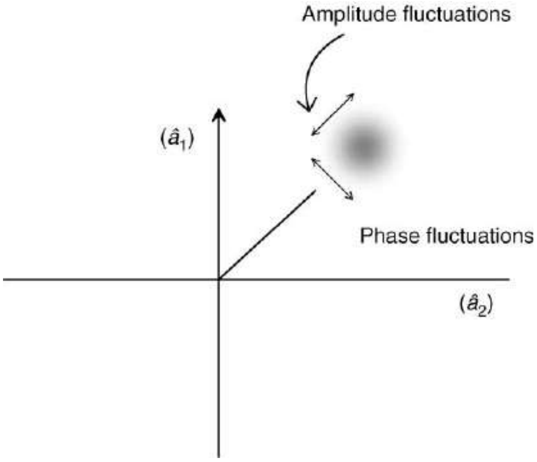

Figure 2 Phase diagram of the quantum mechanical field amplitudes $\langle \hat{a}_1\rangle$ and $\langle \hat{a}_2\rangle$ for coherent radiation.

with the uncertainty squeezed into mainly one component. These states, which are called squeezed states, are defined by $\langle \Delta \hat{a}_1^2 \rangle < 0.25$ or $\langle \Delta \hat{a}_2 \rangle < 0.25$. Squeezed state may also refer to a state which is not a minimum uncertainty state but where, nevertheless, the mean square fluctuation in one of the two components is less than 0.25.

A classical electromagnetic field consists of waves with well-defined amplitude and phase. This is not the case in a quantum treatment: fluctuations are associated with both conjugate variables. The case of a coherent state that most nearly describes a classical electromagnetic field has an equal amount of uncertainty in the two variables (when normalized to the field of a single photon). Equivalently, the field can be described in two conjugate quadrature components and the uncertainties in the two conjugate variables satisfy an uncertainty relation. The coherent state represents a minimum uncertainty state with equal uncertainties in the two quadrature components. This case is usually called the shot-noise limit; it is represented in Figure 2 as a phase diagram.

In a quantum treatment of radiation, it is also possible to generate states that are not present in the classical limit. They can show fluctuations reduced below the classical limit in one of the quadrature components, while the canonically conjugate quadrature component must display enhanced fluctuations in order to fulfill the uncertainty relation. Those states are called 'squeezed states.' An electromagnetic field with fluctuations below the standard quantum limit in one of the quadrature components has, in principle, many attractive applications, for example, in optical communication, in precision and sensitive measurements such as gravitational wave detection, or in noise-free amplification. Therefore, there has been great interest in generating squeezed radiation. Optical measurements have three characteristics that enable them to reach the quantum noise level more readily than in other fields of physics: (i) optical signals are naturally immune to external sources of noise; (ii) thermal noise at room temperature is negligible in the optical domain; and (iii) the outstanding equalities of the optical sources and detectors allow a very-low-level instrumental noise.

Squeezed states with reduced photon number fluctuations (sub-Poissonian photon statistics) correspond to a light beam in which the photons arrive more equally spaced than in a coherent beam for which $g_{\mathrm{I}}^{(2)}(0) = 1$. Therefore, sub-Poissonian number fluctuations correspond to photon antibunching. This shows that photon antibunching and squeezing are related. The state with the largest possible antibunching is the number state, represented by a circle in the phase diagram. It has large fluctuations in both components a_1 and a_2 and, therefore, is not a squeezed state. Thus, it is obvious that antibunching does not imply squeezing. On the other hand, weak squeezing of a coherent state may result in antibunching. If the state is squeezed harder the a_1 component will be better defined. The amplitude fluctuations, however, will start increasing and there will be a transition from antibunching to bunching.

The quantum calculation of the normalized second-order correlation function leads to the following result:

$$g_{\mathrm{I}}^{(2)}(0) = 1 + (\langle \Delta n^2 \rangle - \langle n \rangle / \langle n \rangle^2) \qquad [10]$$

A pure number state of the radiation field with $\langle \Delta n^2 \rangle = 0$, leads to $g_{\mathrm{I}}^{(2)}(0) < 1$, as discussed above. For a coherent state, $\langle \Delta n^2 \rangle = \langle n \rangle$ yielding $g_{\mathrm{I}}^{(2)}(0) = 1$, corresponding to the classical calculations.

It can be concluded that although antibunching and squeezing are related and nonclassical phenomena, one does not imply the other. Thus, the detection schemes necessary to determine the existence of one or the other are quite different. Antibunching or sub-Poissonian statistics can be detected via intensity correlations. Squeezing, on the other hand, is a property of either $\langle a_1 \rangle$ or $\langle a_2 \rangle$, and phase sensitive homo- or heterodyne detection is needed.

Nonclassical states of the radiation field are very sensitive to linear attenuation. Coherent and thermal light fields are the only ones that do not change their photon statistics when attenuated. This is different for nonclassical light. The change depends on the particular case and has to be carefully evaluated.

The first observation of squeezing was achieved in 1985, in an experiment of parametric generation involving four-wave mixing in sodium vapor.

Nonclassical Light from a Single Atom

Resonance fluorescence of atoms is a basic process in radiation–atom interactions, and has therefore always generated considerable interest. The methods of experimental investigation have changed continuously due to the availability of new experimental tools. A considerable step forward occurred when tunable and narrowband dye laser radiation became available. These laser sources are sufficiently intense to easily saturate an atomic transition. In addition, the lasers provide highly monochromatic light with coherence times much longer than typical natural lifetimes of excited atomic states. Excitation spectra with laser light using well-collimated atomic beams lead to a width, being practically the natural width of the resonance transition, therefore it became possible to investigate the frequency spectrum of the fluorescence radiation with high resolution and to also study the photon statistics of the fluorescent light.

Previous experiments to investigate antibunching in resonance fluorescence have been performed by means of laser-excited collimated atomic beams. The initial results obtained in 1977 showed for the second-order correlation function $g^{(2)}(t)$, a positive slope characteristic of photon antibunching, but $g^2(0)$ was larger than $g^2(t)$ for $t \rightarrow \infty$. This was due to number fluctuations in the atomic beam and to the finite interaction time of the atoms. In 1978, the analysis of the experiment was refined and in 1982, another experiment with a longer interaction time was performed. In the latter experiment, the photon correlation was also measured for very low laser intensities. The ideal experiment to study the photon statistics in resonance fluorescence is the investigation of a trapped ion since, in such an experiment, the fluctuations of the atomic number are excluded.

The fluorescence of a single ion should also display the following property. The probability distribution of the photon number recorded in a finite time interval t is narrower than Poissonian, which means, that the variance is smaller than the mean value of the photon number. This is because the single ion can only emit a single photon. Antibunching and sub-Poissonian statistics are often associated. They are, nevertheless, distinct properties and need not necessarily be simultaneously observed, as is the case in the experiment described here. Although there is evidence of antibunching in the atomic-beam experiments, the photon counts are not sub-Poissonian as a result of fluctuations in the number of atoms. In further experiments this effect was excluded by use of a special trigger scheme for the single-atom event. In the setup with a trapped ion these precautions are not necessary since there are no fluctuations in the atomic number. Figure 3 shows the photon correlation of a single trapped ion. The measurements are performed using the $3^2S_{1/2} - 3^2P_{3/2}$ transition of the $^{24}Mg^+$-ion at a wavelength of 280 nm. The laser is tuned slightly below resonance in order to Doppler-cool the secular motion of the ion.

The investigation of the photon correlation employed the ordinary Hanbury-Brown and Twiss setup with two photomultipliers and a beamsplitter (Figure 1). The time delay τ between the photomultiplier signals was converted by a time-to-amplitude converter into a voltage amplitude proportional to the time delay. A delay line of 100 ns in the stop channel allowed for the measurement of $g^{(2)}(\tau)$ for both positive and negative τ in order to check the symmetry of the measured signal. The output of the time-to-amplitude converter was accumulated by a multichannel analyzer in pulse height analyzing mode. Three typical measurements, each at small values of the laser intensity but with different detunings, are shown in Figure 3. The solid curve is a theoretical fit to the measurements. For small time delays (<20 ns), the nonclassical antibunching effect is observed, superimposed with Rabi oscillations being damped out with a time constant corresponding to the lifetime of the excited state.

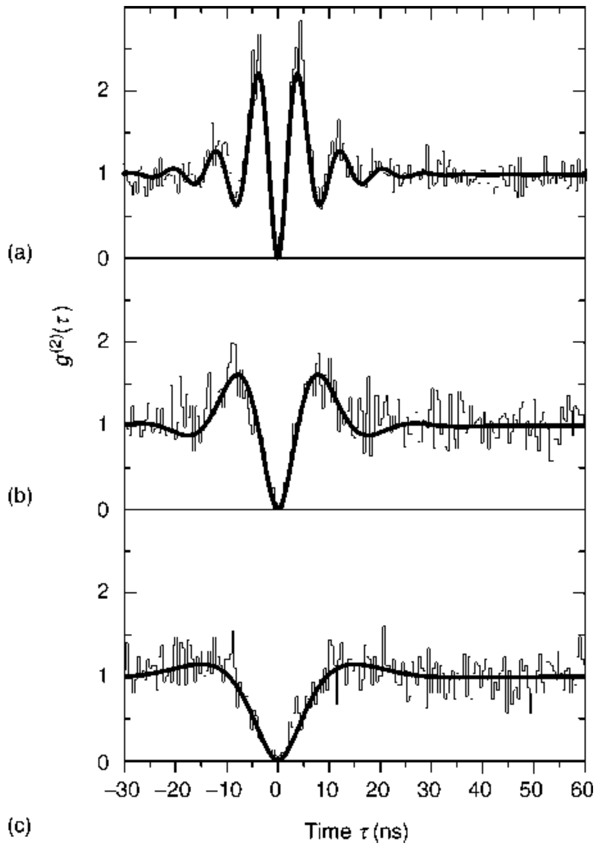

Figure 3 Antibunching signals of a single $^{24}Mg^+$-ion for different laser intensities (decreasing from (a) to (c)). At larger intensities (a) and (b) the Rabi oscillations owing to a strong coherent coupling between ground and excited state are observed.

Recently it also became possible to combine a trapped ion with a cavity. In this way the emission of a photon in the cavity mode can be achieved. This leads to a directed emission of the photon being more efficient than the fluorescence occurring in the entire 4π geometry. In this way, a deterministic single photon source is obtained.

Survey on the Experiments on Squeezing

In order to produce squeezed light from nonlinear optical phenomena, a variety of methods has been used, depending on the type of nonlinearity and the number of interacting modes involved. The choice of optical nonlinearity reduces to either three-wave or four-wave mixing, leading to quadratic or cubic nonlinear response functions, respectively. Each of these methods has advantages and disadvantages. Three-wave mixing requires a pump field at twice the frequency of the field-mode being squeezed. However, it has the advantage as any sidebands induced on the pump, owing to phase noise, have a large frequency offset from the squeezed fields. The other method, the much simpler four-wave mixing, does not require any frequency-doubling step. This was, therefore, the first method used to obtain squeezing with quadrature noise below the vacuum level.

Most of the early experiments on squeezing were performed with one or two cavity modes with the squeezed modes being transferred to propagating external modes via beamsplitters for measurement. This narrowband procedure is followed in more recent experiments commonly used for three-wave (parametric) squeezing. Also squeezing in waveguides was achieved leading to a relatively simple and broadband implementation of squeezing, with, mostly a four-wave mixing with short pulses or solitons being performed.

The main limiting factors that remain include control over the driving laser phase and intensity noise, absorption in the dielectric material and associated devices, and phase fluctuations caused both by low-frequency technical noise in components, and by intrinsic material properties of phonon interactions and refractive index fluctuations. It appears not impossible to reach squeezing levels of 10 dB or more below shot noise, over bandwidths of up to 1 THz. At this level of noise reduction, the main problem is the issue of improving detector technology to allow such low noise levels to be investigated.

Squeezing in Lasers

The quantum noise properties of the laser were studied intensively after its invention in the early 1960s. The conclusion of these studies is that the laser output approaches quite closely the noise properties of a coherent state, provided the laser is operated well above threshold. That is, the spectral variance of the photon number and phase fluctuations at high enough Fourier frequencies will be at the Poissonian or quantum noise limit (also referred to as the shot noise level). On the other hand, the phase of the laser is unconstrained and drifts as a function of time, so-called laser phase diffusion. As a result, the phase fluctuations show excess noise at low frequencies.

In the early 1980s, it was realized that if the pumping mechanism for the laser could be made sub-Poissonian then amplitude squeezing could be obtained at low frequency. A few years later it was shown that sub-Poissonian pumping could be realized in semi-conductor lasers by regularizing the pump current. An experimental demonstration of amplitude squeezing from semi-conductor lasers was soon achieved. Although still requiring rather specialized lasers, many groups now routinely obtain squeezing from semi-conductor lasers.

There are a number of disadvantages to semi-conductor lasers, including their poor beam quality and short coherence length. As a result, interest in other possible squeezing mechanisms continues. It has been found that intrinsic dynamical effects can also lead to amplitude squeezing in the laser output.

Laser squeezing is of a different character to that produced by more ‘traditional squeezers.’ For example, in parametric amplification, the de-amplification of phase fluctuations is accompanied by an equal and opposite amplification of amplitude fluctuations which preserves the uncertainty relationship between them. However, in laser squeezing the reduction in amplitude fluctuations is not linked to any increase in phase fluctuations. Thus, in laser squeezing we take advantage of the fact that the laser output is not in a minimum uncertainty state to suppress amplitude noise. This means that laser squeezing is always amplitude or intensity squeezing.

Application of Nonclassical Light

The applications of nonclassical light are numerous. The single photon sources discussed above have their main applications in quantum communication with single photons where they may be used to establish secure signal transmission. There is also a decisive advance in the field of quantum information expected from the combination of local atomic quantum processing with photonic transmission of quantum states. Conceptually, new possibilities would arise to enhance quantum computational capabilities. For example, separate local quantum processors of moderate size could be linked in a quantum network, with quantum information being exchanged by means of photons.

Although a large number of phenomena have been predicted in which the quantum nature of the squeezed vacuum plays an essential role in modifying atomic properties, only the linear dependence of the two-photon absorption rate of a three-level ladder atom has been confirmed experimentally but classical sources show a quadratic dependence. In 1998 an important series of experiments was reported which investigated the interaction of a beam of Cs atoms with squeezed light in the Fabry–Perot cavity environment. The principal phenomena examined were the modification of the transmission spectrum of a weak, tunable probe beam due to the presence of the squeezed light, and the variation of the transmission modulation with the phase of the squeezed vacuum relative to that of an applied, saturating, coherent field.

There have been many proposed practical applications of squeezed light, and some experiments have already demonstrated its potential in precision measurements and improvement of spectroscopic sensitivity. Squeezed light may be used in optical communications to enhance signal-to-noise ratio, gravitational wave detection, and quantum nondemolition measurements.

The wider applications of squeezed light are numerous. An idea of the range and variety of applications is provided by the following selection: the protective measurement of a single squeezed harmonic oscillator state, coherent nonlinear spectroscopy using number squeezed light, the improvement of quantum nondemolition measurements using a squeezed meter input, sub-shot-noise Doppler anemometry with amplitude squeezed light, noiseless transfer of nonclassical light through bistable systems, generation of few photon states from squeezed atoms, and spin squeezing in a collection of atoms illuminated with squeezed light.

Einstein–Podolsky–Rosen (EPR) Correlations and Entanglement of Photons

In 1986, a quantum nondemolition experiment confirming the quantum correlation of the quadrature phase amplitudes of two spatially separated fields, was reported. In 1987, two-mode squeezing with two spatially separated detectors was experimentally demonstrated using parametric oscillation. With the experimental realization of such two-mode squeezing and quantum nondemolition measurements, the possibility of testing for continuous variable EPR-type correlations became apparent. In these proposals, the conjugate ‘position’ and ‘momentum’ observables would be the two orthogonal quadrature phase amplitudes of the field. The high efficiency of detectors used in the quadrature phase amplitude squeezing experiments was an important issue in this connection. Experimental tests of EPR correlations at the time were based on Bohm’s version of the EPR paradox, and in fact, focused on Bell’s extension of this work. These tests involved measurements performed on individual photon pairs, and used very inefficient detectors.

A specific proposal and a criterion to demonstrate the continuous variable EPR correlations for real experiments by way of a realization of the 1935 EPR paradox, was put forward in 1989. This proposal employed the two-mode squeezed state as the EPR source. The first experimental achievement, using the parametric oscillator, of the 1989 EPR criterion, for efficient measurements with continuous variable outcomes, was reported in 1992.

The EPR fields have proved significant in enabling the experimental realization of continuous variable quantum teleportation, and may have application also to quantum cryptography. Again, because of the efficient nature of the measurement of field quadrature phase amplitudes, and the intrinsic macroscopic nature of the correlated fields, continuous variable teleportation and cryptography may provide large advantages over protocols using a finite number state basis, where signals of individual photons are sent. At the moment, several groups are working in this promising direction.

Conclusions

In this contribution the generation of nonclassical light was reviewed. The generation of nonclassical light became possible after stable and reliable laser sources were available. There are many sources of nonclassical light available now. It is clear that we will see many new applications in the future. Especially in connection with optical communication, rather interesting improvements can be expected.

See also

Coherent Lightwave Systems. **Photonic Crystals:** Electromagnetic Theory. **Quantum Optics:** Atomic Coherence Effects.

Further Reading

Caves CM and Drummond PD (1994) Quantum limits on Bosonic communication rates. *Review of Modern Physics* 44: 481–537.

Cresser JD, Häger J, Leuchs G, Rateike FM and Walther H (1982) Resonance fluorescence of atoms in strong monochromatic laser fields. *Topics in Current Physics* 27: 21–59.

Dagenais M and Mandel L (1978) Investigation of two-time correlations in photon emissions from a single atom. *Physical Review A* 18: 2217–2228.

Diedrich F and Walther H (1987) Nonclassical radiation of a single stored ion. *Physical Review Letters* 58: 203–206.

Drummond PD and Ficek Z (eds) (2004) Springer Series on Atomic, Optical, and Plasma Physics, vol. 27. Berlin, Heidelberg, New York: Springer Verlag.

Gatti A, Lugiato LA and Wiedemann H (1997) In: Giacobino E, Reynaud S, and Zinn-Justin J (eds). *Quantum Fluctuations*, p. 431. Amsterdam: Elsevier.

Georgiades NP, Polzik ES, Edamatsu K, Kimble HJ and Parkins AS (1995) Nonclassical excitation for atoms in a squeezed vacuum. *Physical Review Letters* 75: 3426–3429.

Glauber R (1963) The quantum theory of optical coherence. *Physical Review* 130: 2529–2539.

Glauber R (1963) Coherent and incoherent states of the radiation field. *Physical Review* 130: 2766–2788.

Guthöhrlein GR, Keller M, Hayasaka K and Lange W (2001) A single ion as a nanoscopic probe of an optical field. *Nature* 414: 49–51.

Heidmann A, Horowicz RJ, Reynaud S, Giacobino E, Fabre C and Camy G (1987) Observation of quantum noise reduction on twin laser beams. *Physical Review Letters* 59: 2555–2557.

Höffges JT, Baldauf HW, Eichler T, Helmfrid SR and Walther H (1997) Heterodyne measurement of the fluorescent radiation of a single trapped ion. *Optical Communications* 133: 170–174.

Hood CJ, Kimble HJ, Turchette QA, Georgiades NPh and Parkins AS (1998) Squeezed excitation in cavity QED: experiment and theory. *Physical Review A* 58: 4056–4077.

Jakeman E, Pike ER, Pusey PN and Vaugham JM (1977) Effect of atomic number fluctuations on photon antibunching in resonance fluorescence. *Journal of Physics A: Mathematical and General* 10: L257–L259.

Kimble HJ, Dagenais M and Mandel L (1977) Photon antibunching in resonance fluorescence. *Physical Review Letters* 39: 691–695.

Kimble HJ, Dagenais M and Mandel L (1978) Multiatom and transit-time effects on photon-correlation measurements in resonance fluorescence. *Physical Review A* 18: 201–207.

Kimble HJ, Carnal O, Hu Z, Mabuchi H, Polzik ES, Thompson RJ and Turchette QA (1995) *Annals of the New York Academy of Sciences* 755: 87.

Levenson MD, Shelby RM, Reid MD and Walls DF (1986) Quantum nondemolition detection of optical quadrature amplitudes. *Physical Review Letters* 57: 2473–2476.

Loudon R (2000) *The Quantum Theory of Light*, 3rd edn. Oxford, New York: Oxford University Press.

Ou ZY, Pereira SF, Kimble HJ and Peng KC (1992) Realization of the Einstein–Podolsky–Rosen paradox for continuous variables. *Physical Review Letters* 68: 3663–3666.

Reid MD (1989) Demonstration of the Einstein–Podolsky–Rosen paradox using nondegenerate parametric amplification. *Physical Review A* 40: 913–923.

Short R and Mandel L (1983) Observation of sub-Poissonian photon statistics. *Physical Review Letters* 51: 384–387.

Short R and Mandel L (1984) Sub-Poissonian photon statistics in resonance fluorescence. In: Mandel L and Wolf E (eds). *Coherence and Quantum Optics*, vol. V, pp. 671–678. New York: Plenum.

Slusher RE, Hollberg LW, Yurke B, Mertz JC and Vally JF (1985) Observation of squeezed states generated by four-wave mixing in an optical cavity. *Physical Review Letters* 55: 2409–2412.

Walls DF (1979) Evidence for the quantum nature of light. *Nature* 280: 451–453.

Wu LA, Kimble HJ, Hall J and Wu H (1986) Generation of squeezed states by parametric down conversion. *Physical Review Letters* 57: 2520–2523.

Yamamoto Y and Machida S (1987) High-impedance suppression of pump fluctuation and amplitude squeezing in semiconductor lasers. *Physical Review A* 35: 5114–5130.

NONLINEAR OPTICS AT THE CRITICAL FIELD LIMIT

G A Mourou, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is now possible to produce, with compact lasers, intensities so large that the relativistic character of the electrons dominates laser–matter interactions. This is a fundamentally new regime in optics where optics, confined to the study of eV characteristic phenomena, are now addressing phenomena with characteristic energies vastly greater in the MeV–GeV (TeV in the near future) range. With relativistic optics, the field of nonlinear optics found a number of novel applications, including fast ignition of inertially compressed fusion target, bright sources of energetic electrons, protons, ions, positrons, pions, etc., but also nuclear physics, high energy physics, astrophysics, and general relativity.

Relative optics has been one of the most active fields of optics and it is impossible to describe all the applications here. So, this new regime will be illustrated with few examples of the key relativistic mechanisms – reminiscent of conventional nonlinear optics, such as relativistic harmonics, relativistic optical rectification (wake-field acceleration), and relativistic self-focusing in refraction and reflection. We finally show that relativistic optics offers a new and efficient gateway to attosecond physics and nonlinear quantum electrodynamics (QED).

Optics is the study of how electrons respond to light. All material optical properties are the consequence of how light interacts with electrons in materials. Because, in the visible regime the photon energy is of the order of 1 eV, optics deals with phenomena with eV characteristic energy. Light is a wave composed of coupled electric and magnetic fields oscillating in synchrony at high frequencies. The electric and magnetic field oscillate perpendicular to each other, and perpendicular to the direction of propagation, k. Because of the electric field E, electrons under the action of the Lorentz force, $F = eE$, moves a distance $x = a_1 E$ transversally to the propagation direction. In classical optics, the oscillation amplitude x is small compared to the wavelength, and the electrons oscillate and radiate at the laser frequency. The radiated field amplitude and phase depend on how the electron is bound to the atom of the material. The invention of the laser in 1960 gave access to much larger intensities and electromagnetic fields. The laser field became strong enough that the electron displacement, x, in this field could no longer be considered as strictly proportional to the laser field. This departure from linearity is addressed by the addition of new terms in the Taylor series expansion of the displacement:

$$x = a_1 E + a_2 E^2 + a_3 E^3 + \cdots \qquad [1]$$

These additional terms are responsible for all the effects that form the field of nonlinear optics. Note here that the field is still not strong enough to ionize an electron from its parent atom and is therefore below the field that could irreversibly damage the material. Thus, the typical intensity must be kept to a value of less than $\sim 10^{14}$ W/cm^2, where the quiver energy E_Q starts to challenge the ionization energy of a few eV, depending on the material:

$$E_Q = \left\langle \frac{e^2 E^2}{4m\omega^2} \right\rangle \qquad [2]$$

The bracket denotes the time average over one cycle and ω is the laser angular frequency. Above this limit, the material response is dominated by multiphoton ionization or tunneling, as shown in Figure 1. During the ionization event, high harmonics are produced. Note here again that because the electron is still bound to its nucleus during the emission, we can consider the process of high harmonic generation as bound-electron nonlinear optics. As the intensity increases from 10^{14} to 10^{18} W/cm^2 (Figure 1), we are entering the regime of relativistic intensity or relativistic nonlinear optics. This is a regime that was predicted and discussed in the 1960s but that could be accessed only after the technique of chirped pulse amplification (CPA) was demonstrated, followed by significant progress in the areas of short pulse generation and pulse focusing.

Relativistic Optics

In classical optics one always assumes that during a laser oscillation period the electron mass m_0 stays constant. It is a reasonable assumption since the quiver velocity v of the electron is small compared to the speed of light c. As the intensity is increased a point is reached where the electron quiver velocity approaches the speed of light. This has several implications; first, the mass of the electron changes to become

$$m = \gamma m_0 \qquad [3]$$

where γ is the Lorentz factor given by

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1} \qquad [4]$$

Second, because v is close to c, the magnetic force component needs to be added and the Lorentz force becomes:

$$\vec{F} = e\vec{E} + \frac{\vec{v}}{c} \times \vec{B} \qquad [5]$$

Until recently, in classical optics and laser physics, the $v \times B$ component was totally ignored. However, in the relativistic regime, this component is as large as

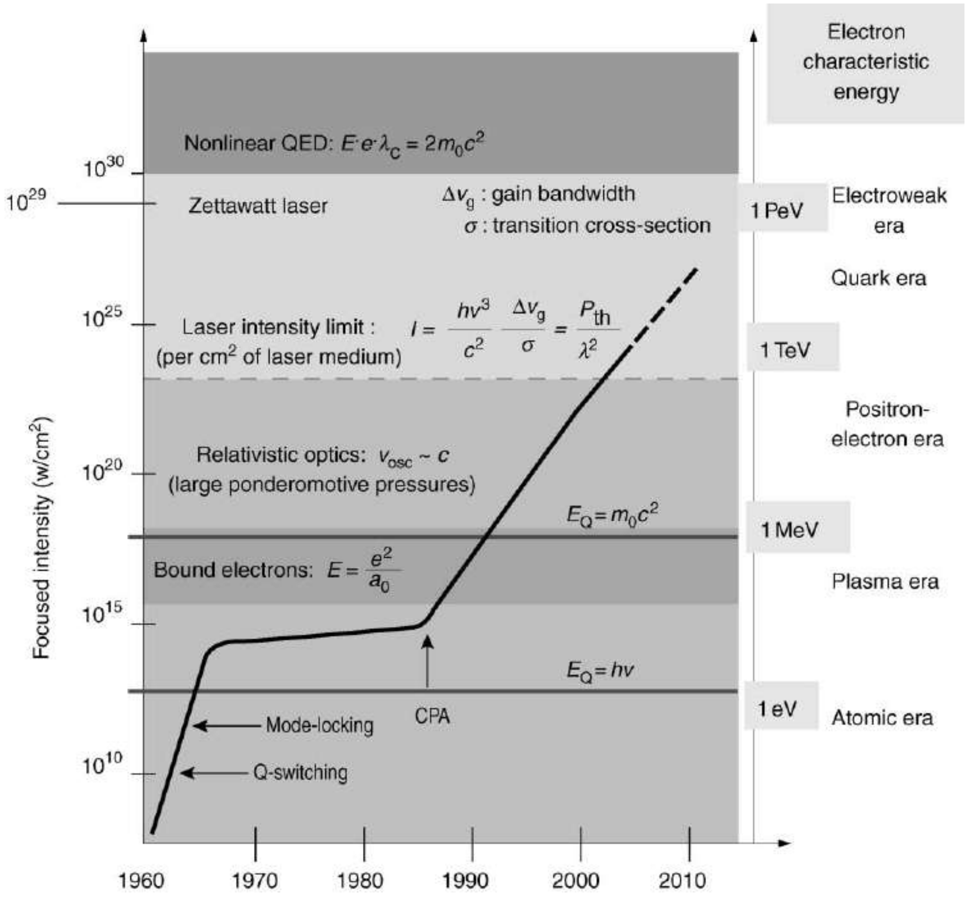

Figure 1 Laser intensity versus years. Note the very steep slope in intensities that occurs during the 1960s. This period corresponds to the discovery of most nonlinear optics effects. Since 1985 after the CPA invention followed by the successive laser short pulse improvements, we are experiencing a similar increase in intensity opening a new regime in optics at 10^{18} W/cm^2 dominated by the relativistic character of the electrons.

the electric one. It pushes the electron forward and is the tenet of the relativistic regime.

At this point, it is convenient to introduce the normalized vector potential:

$$a_0 = \frac{e\lambda E}{m_0c^2} \qquad [6]$$

which represents the energy gain by the electron over one period (λ) normalized to the electron rest mass energy. At $a_0 = 1$, corresponding to 10^{18} W/cm^2 for 1 μm wavelength light, the laser–matter interaction is strongly dominated by the relativistic character of the electrons. At these intensities, the motion of the electrons becomes relativistic and the trajectory of the electron dramatically changes. Figure 2 depicts the difference between the two regimes. It can be shown that independently of the intensity regime, the electron transverse motion is always proportional to a_0, while the longitudinal motion is always proportional to a_0^2. This means that for classical

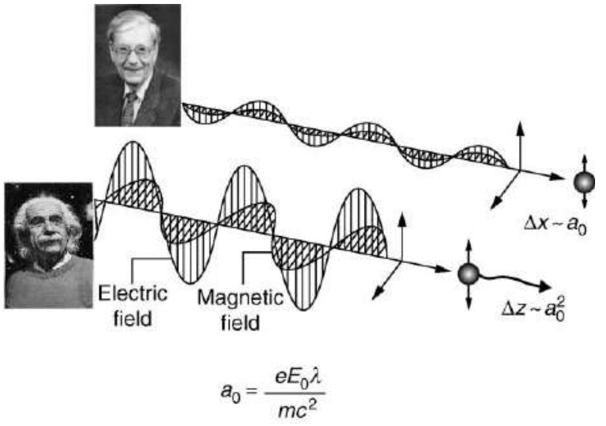

Figure 2 A comparison between classical optics and relativistic optics. The transverse electron displacement scales as a_0 and the longitudinal scales as a_0^2. (a) In classical optics $a_0 < 1$ the transverse electron displacement is proportional to a_0. It is larger than the longitudinal proportional to a_0^2. (b) In relativistic optics it is the opposite, $a_0 > 1$ the transverse displacement becomes smaller than a_0^2.

optics or even bound-electron nonlinear optics with a_0 varying from 10^{-10} to 10^{-4}, the longitudinal motion is $\sim 10^{-10}$ to 10^{-4} smaller than the transverse motion. Of course, the situation reverses completely for $a_0 > 1$, where the longitudinal motion becomes larger than the transverse motion. The departure of the electron motion from the linear regime, due to its relativistic character, gives rise to a new kind of nonlinear optical effects that constitutes the field of relativistic nonlinear optics. This area has become one of the most active in laser physics. In many cases, there are strong similarities between conventional and relativistic optics, and it is the purpose of this article to illustrate the field of relativistic optics by using some of the effects well established in bound electron nonlinear optics.

Considering the rapid progress in laser intensities that the field experienced in the 1960s, as indicated in Figure 1, scientists expected to observe the manifestation of relativistic effects in the 1970s. Formidable intensities on the order of 10^{18} W/cm^2 for 1 μm wavelength light were necessary, however, and could not be readily obtained. Such intensity levels could only be produced in the early 1980s with large-scale sources such as the Helios laser at Los Alamos National Laboratory (LANL) (Figure 3a). Work in relativistic optics virtually could not be conducted in university research laboratories, but rather only in government laboratories. The first relativistic harmonics produced by the critical surface nonlinear motion were observed in 1980 with a kJ-pulse-energy laser at mHz repetition rates (one shot/20 m).

The advent of CPA (Figure 4) revolutionized the field of relativistic optics, placing it within reach of more researchers through enormous reductions in laser size and cost. Further advances in the ultrashort-pulse duration of lasers and in laser focusing made relativistic optics more accessible, especially to universities. Relativistic intensities can be achieved with mJ energy levels at kHz repetition rates by university-size lasers (Figure 3b) instead of with large, kJ, sub-Hz-repetition-rate government-type lasers. Figures 3a and b illustrate the revolution in the relativistic laser evolution representing the Helios (LANL) system. Figure 3b depicts the relativistic λ^3 laser target chamber. The λ^3 laser reaches its relativistic intensity (10^{18} W/cm^2) with pulse energy only on the order of 1 mJ and pulse duration on the order of 10 fs, i.e., a few light periods. Its energy is concentrated on a single wavelength spot size using an f#1 paraboloid corrected by a deformable mirror. Note that all the energy is contained in a focal volume of a few λ^3, hence its name.

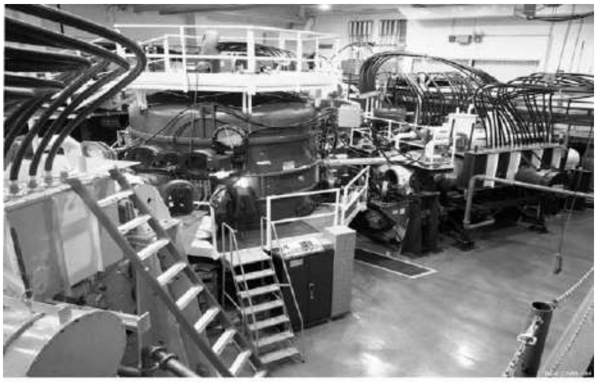
(a)

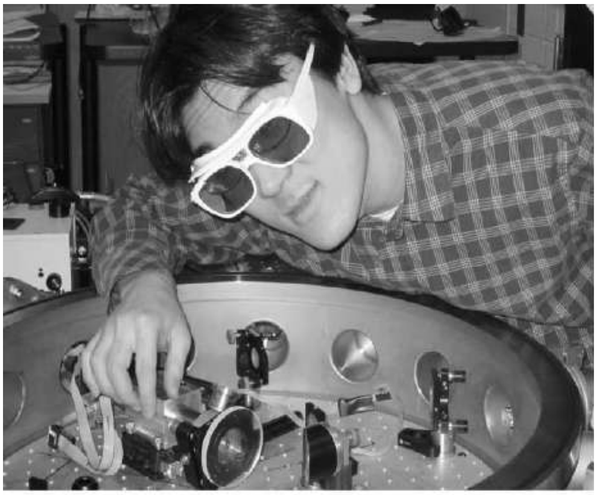
(b)

Figure 3 Two relativistic lasers. (a) Helios at LANL, was the first relativistic laser with $a_0 \sim 1$ at mHz repetition rate. (b) The λ^3 at the University of Michigan has an $a_0 \sim 1$ at kHz repetition rate.

Some Examples of Relativistic Nonlinear Optics

We will illustrate the field of relativistic optics by comparing relativistic effects with some well-known nonlinear effects.

(a) Relativistic optical rectification (laser wake-field acceleration of electrons and ions)

One well-known bound-electron optical effect coming from the second-order term in eqn [1] is optical rectification. This effect, which takes place in noncentrosymmetric crystals, gives rise to a DC field transverse to the propagation direction. In the relativistic regime, a similar effect known as wake-field acceleration will also contribute to rectification of the optical field. We have seen that in vacuum, electrons are strongly pushed forward by $v \times B$ force. In a plasma, the electromagnetic wave will also push

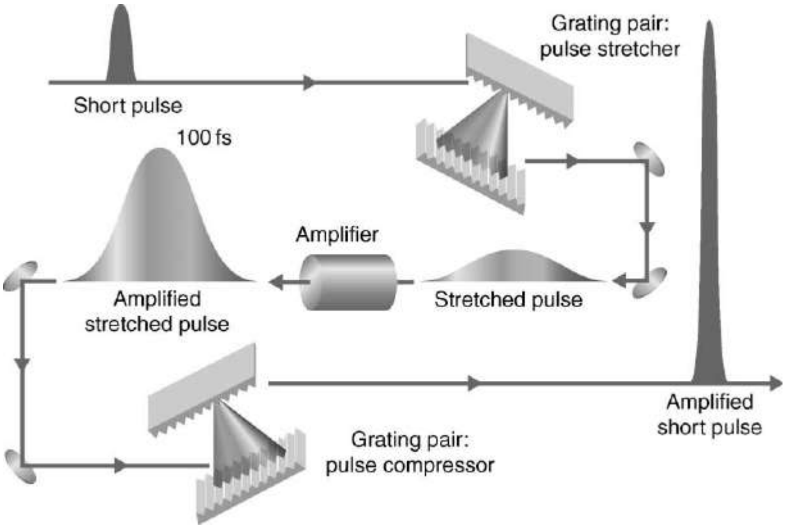

Figure 4 The concept of CPA is the key technique to get relativistic intensity. An initial short pulse is first stretched out (chirped) by a factor ~10^4, for instance by a pair of diffraction gratings. The stretched pulse intensity becomes lowered by the same factor of ~10^4. The stretched pulse intensity can then be amplified to a level 10^4 times higher. Once the laser energy has been extracted from the amplifier the stretched pulse is recompressed by the same factor (~10^4) to its initial value. This way the laser intensity has been boosted to 10^4 times the peak intensity that the amplifier could have withstood.

forward the electrons, dragging the much heavier positive ions behind, as shown in Figures 5a and b. The charge separation that ensues will produce a strong electrostatic field that can be of the order of the transverse light electric field, which is huge. Unlike in classical optical rectification, in the relativistic regime the effect can be very efficient. Of the order of 10% or more of the transverse field may be rectified. For an intensity of 10^{18} W/cm^2, the electric field can be on the order of 10^{12} V/m (CERN on one meter), or 10^5 times larger than the field that can be obtained by more conventional technology. In addition, the rectified field is flipped along the longitudinal direction. Note also that since the field is created within a plasma, the medium cannot be broken down further and can support these extremely high fields. This effect had to wait for the relativistic-intensity lasers to become a reality before being demonstrated. Today energies up to 200 MeV have been demonstrated by using a 50 TW laser focused on a gas jet. Because the electrons drag the ions, a beam of ions is also observed concomitant with the electron beam. The generation of protons with energy up to 56 MeV has been reported by LLNL.

(b) Relativistic harmonic generation

Harmonic generation in the relativistic regime can be produced in two ways, as a result of the laser plasma interaction with electron density n_e, either below or above the critical density n_c. The first is for $n_e < n_c$ and corresponds to an underdense plasma where light can propagate. Harmonics generation was predicted in the late 1960s. Even-order harmonics are observed with linear and circular polarizations. This is the signature of relativistic effects since even-harmonics cannot occur in conventional nonlinear optics in centrosymmetric media.

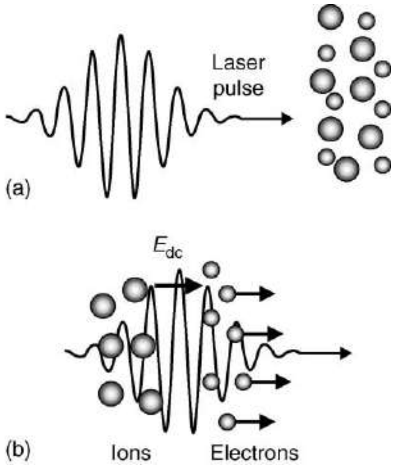

Figure 5 Relativistic rectification in plasma. (a) The high intensity pulse before entering the plasma. (b) Illustrate the $v \times B$ pushing first the plasma electrons. The electrons then drag the heavy ions behind them like a horse pulling a cart. The longitudinal electrostatic field that is generated is almost as large as the transverse laser field.

Harmonic generation can also be produced in over-critical plasmas, where $n > n_c$. In this case, the light cannot propagate and be reflected. The strong $v \times B$ oscillatory force applied to the critical surface will produce a self-modulated reflected beam. This effect was first observed using the Helios laser. The idea of the moving-mirror model was first put forward in the early 1980s. Generation of frequencies up to the 70th harmonic has been demonstrated so this method could provide an efficient way to produce high harmonics. This technique has the advantage over conventional high harmonic generation in that it is capable of accommodating a pulse of large energy, since one can focus the beam over a small area, limited only by the laser wavelength, without fear of disrupting the plasma.

(c) Relativistic self-focusing in refraction

In bound-electron nonlinear optics, the third-order term in eqn [1] is responsible for self-focusing. A similar effect also exists in relativistic optics; it is described in Figure 6a. In the relativistic regime, the electron driven by the laser field changes its mass. In turn, this mass change will modify the index of refraction according to:

$$n^2 = 1 - \left(\frac{\omega_p}{\omega}\right)^2 \qquad [7]$$

where the plasma frequency is given by

$$\omega_p^2 = \frac{e^2}{m} n_e \qquad [8]$$

where n_e is the plasma electron density, and m is the intensity-dependent mass according to eqn [1].

For $\omega > \omega_p$, the index of refraction is positive and the pulse can propagate in the plasma. For the higher-beam-intensity part of the beam, i.e., on the beam axis, the plasma frequency will be less than the frequencies corresponding to the pulse wing. This will cause a modification of the laser wavefront and the beam to self-focus.

(d) Relativistic self-focusing in reflection and isolated attosecond pulse generation

For $\omega < \omega_p$, (eqn [7]) is negative and the pulse cannot propagate and so is reflected at the critical surface. However, the large light pressure will push the plasma critical-density surface at twice the laser frequency, making it cancel out the beam spatial intensity profile. In the relativistic λ^3 regime, the laser is focused at one wavelength. If the light is linearly polarized, the electron density will be collectively moved back and forth along the critical surface. This time-dependent deformation of the critical surface will produce a deflection of each electromagnetic half-cycle in a specific direction (Figure 7), providing a simple means for generating isolated attosecond pulses efficiently. The relativistic motion of the critical surface compresses (relativistic Doppler) the incident femtosecond pulse to the attosecond regime. To understand this compression, recall that the incident pulse travels opposite to the mirror motion formed by the critical surface (for a fraction of the cycle). As we have seen in this regime the critical surface moves also at close to the speed of light, v. Since the reflected light must also travel at c, it has no choice but to be compressed, producing an electromagnetic accumulation along the critical surface. The compression factor will

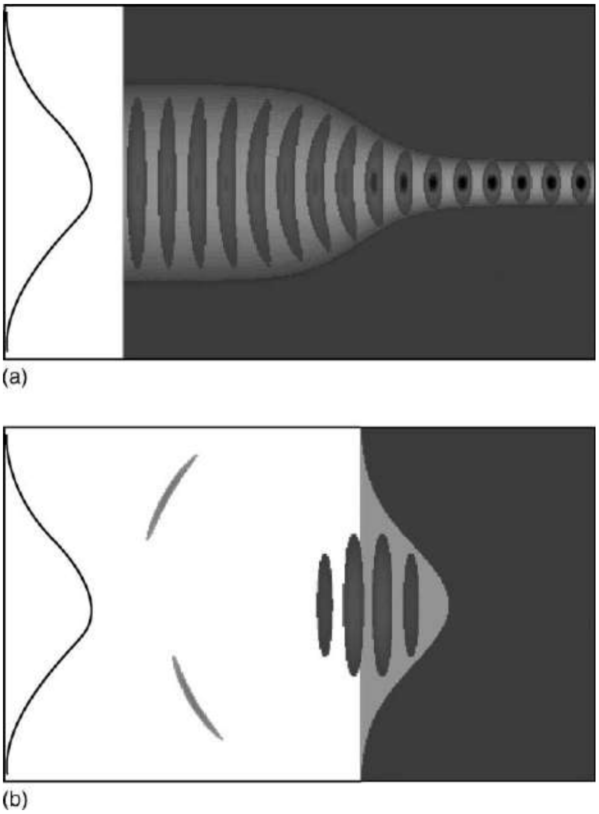

Figure 6 Relativistic self-focusing. (a) Relativistic self-focusing in refraction. Here the index of refraction is modified by the change in the electron mass. The electron mass change will in turn modify the plasma frequency across the beam and consequently the index of refraction. (b) Relativistic self-focusing in reflection showing the deformation of the critical surface by the light pressure from a strongly focused beam. The light is reflected in a series of ultrashort pulses in the attosecond domain.

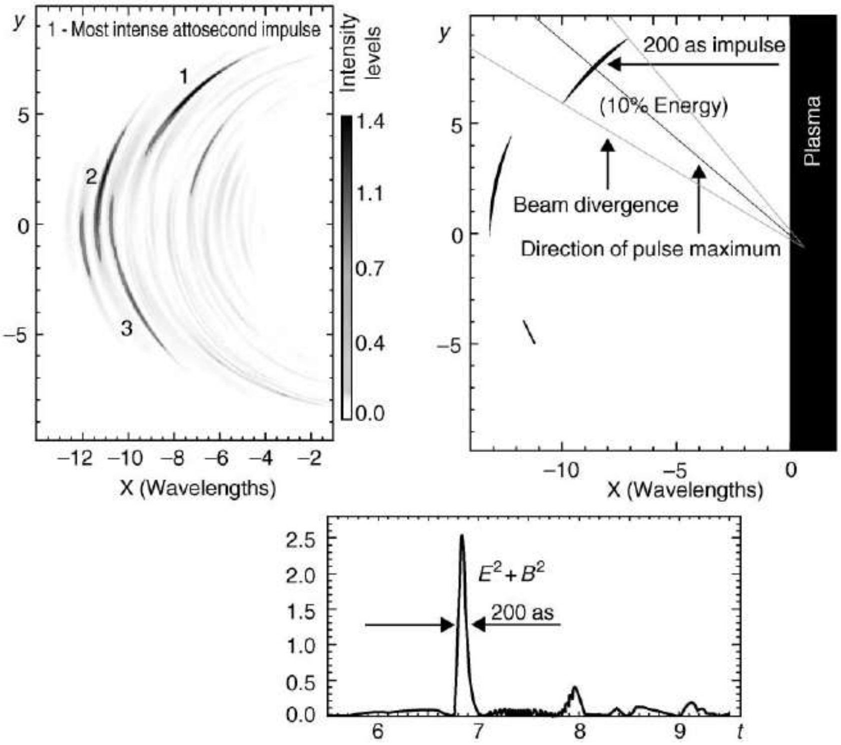

Figure 7 Compressed and deflected attosecond pulses. The combination of relativistic deflection and compression leads to 200 as isolated pulses.

depend on the critical surface velocity and is dependent on the plasma stiffness, i.e., the plasma density. Low-density plasmas are softer than high-density ones; here, n_c is the plasma critical density. The compressed pulse duration has been determined by a particles-in-cell (PIC) simulation to be given by $\tau = 600/a_0$. From the PIC simulation, it can be determined that the compressed pulse optimum duration is obtained for $n_e = a_0 n_c/2$. This technique is an interesting alternative to the one based on high-harmonic generation due to bound electrons, which has been recently demonstrated. The advantage of the relativistic optics approach is that it can be scaled up to large energies, from the mJ to the several-joule level, in contrast with demonstrated attosecond schemes that can only compress mJ pulse energies with an efficiency of 10^{-6}. For $a_0 \sim 100$, corresponding to incident intensities of $\sim 10^{22}$ W/cm^2, a 10-as pulse can be produced by relativistic compression and deflection.

(e) Nonlinear QED on a table top

Looking at Figure 1, we see that there is a regime around 10^{29} W/cm^2 where it becomes possible to produce electron–positron pairs. The condition to fulfill this is given by

$$E_s \lambda_c e = 2m_0 c^2 \qquad [9]$$

Here $\lambda_c = h/m_0 c$ is the Compton wavelength, E_s the Schwinger field, and h the Planck constant. This expression states that for pair creation the work of the laser field over a Compton wavelength must be equal to the electron–positron rest-mass energy. This field E_s, also known as the critical field, $E_s \sim 10^{16}$ V/cm, corresponds to a laser intensity of 10^{29} W/cm^2.

An alternative way to depict the process of electron–positron generation from vacuum is by considering the vacuum as a dielectric medium with a gap $E_G = 2m_0c^2$. The field to produce an electron–positron pair is the one necessary to bend the bands so that an electron can tunnel from the Dirac sea into the vacuum states (Figure 8). The probability for such an event is given by

$$W \propto \exp\left(-\frac{\pi E_s}{E}\right) \qquad [10]$$

Today, the intensity necessary to reach pair creation is more than six orders of magnitude higher

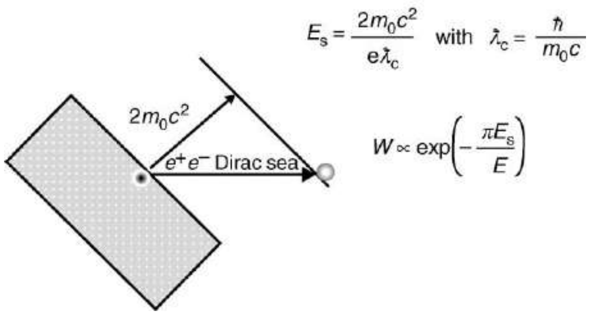
$E_s = \frac{2m_0c^2}{e\lambda_c}$ with $\lambda_c = \frac{\hbar}{m_0c}$
$2m_0c^2$
e^+e^- Dirac sea
$W \propto \exp\left(-\frac{\pi E_s}{E}\right)$

Figure 8 Pair creation. Electron tunneling from the Dirac sea into the vacuum state.

than what we could produce, $\sim 10^{22}$ W/cm^2. To bridge this enormous gap the electric field has been enhanced by colliding a short and intense pulse with a Stanford Linear Accelerator Center (SLAC) super-relativistic, 50 GeV electron beam. The Lorentz factor is $\gamma = 10^5$, and the laser field of 210^{10} V/cm (corresponding to an intensity of 10^{18} W/cm^2) can be enhanced by 10^5 to reach the critical field.

The progress in relativistic compression described above could make possible the generation of electron–positron pairs directly from the laser. We are on the verge of having a tabletop laser that could produce an intensity close to 10^{23} W/cm^2, corresponding to $a_0 > 100$. From the scaling law described previously, we should get $\tau < 6$, where the wavelength of such a compressed pulse would also be reduced by ~100. The focal volume would then be reduced by six orders of magnitude. If we assume an efficiency of 10^{-1}, an intensity close to the critical intensity of 10^{29} W/cm could be reached with a tabletop laser and boil vacuum.

Conclusion

Today, laser intensities are such that during light–matter interaction, the electron motion is dominated by relativistic effects responsible for a new type of nonlinear optics called relativistic nonlinear optics. Relativistic nonlinear optics is characterized by extreme field gradients (TV/m) and light pressure in the Gbar regime. Relativistic nonlinear optics offers an attractive and novel way for efficient attosecond pulse generation with intensities near the critical (Schwinger) field, opening optics to vast new horizons.

See also

Relativistic Nonlinear Optics.

Further Reading

Agostini P, Barjot G, Bonnal JF, Mainfray G and Manus C (1968) *IEEE Journal of Quantum Electronics* QE-4: 667.
Albert O, Wang H, Liu D, Chang Z and Mourou G (2000) *Optics Letters* 125: 1125.
Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.
Boyd RW (1992) *Nonlinear Optics*. San Diego: Academic.
Brabec T and Krausz F (2000) *Review of Modern Physics* 72: 545.
Brezin E and Itzykson C (1970) *Physics Review D* 2: 1191.
Bula C, McDonald KT, Prebys EI, *et al.* (1996) *Physics Review Letters* 76: 3116.
Bulanov SV, Naumova NM and Pegoraro F (1994) *Physics Plasmas* 1: 745.
Carman RL, Benjamin RF and Rhodes CK (1981) *Physics Review A* 24: 2649.
Chen S-Y, Maksimchuk A, Esarey E and Umstadter D (2000) *Physics Review Letters* 84: 5528.
Eberly J (1969) *Progress in Optics* 7: 361.
Franken PA, Hill A, Peters C and Weinreich G (1961) *Physics Review Letters* 7: 118.
Hentschel M, Kienberger R, Spielmann Ch, *et al.* (2001) *Nature* 414: 509.
Joshi C and Katsouleas T (2003) Plasmas accelerators at the energy frontier and on tabletops. *Physics Today* 56: 47.
Litvak AG (1969) *Soviet Physics JETP* 30: 344.
Malka V, Fritzler S, Lefebvre E, *et al.* (2002) Electron acceleration by a wakefield forced by an intense ultrashort laser pulse. *Science* 22: 298.
Max C, Arons J and Langdon A (1974) *Physics Review Letters* 33: 209.
Mourou GA, Barty CPJ and Perry MD (1998) *Physics Today* 51(1): 22.
Naumova NM, Nees JA, Sokolov IV, Hou B and Mourou GA (2004) Relativistic generation of isolated attosecond pulses in a λ^3 focal volume. *Physics Review Letters* 92: 063902.
Norreys PA, *et al.* (1996) *Physics Review Letters* 76: 1832.
Sarachik E and Schappert G (1970) *Physics Review D* 1: 2738.
Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.
Snavely RA, *et al.* (2000) *Physics Review Letters* 85: 2945.
Strickland D and Mourou G (1985) *Optics Communications* 55: 219 and 447.
Tajima T and Dawson JM (1979) *Physics Review Letters* 43: 262.
Tarasevitch A, Orisch A, von der Linde D, *et al.* (2000) *Physics Review A* 62: 023816.
Umstadter D, Chen S-Y, Maksimchuk A, Mourou G and Wagner R (1996) *Science* 273: 472.

NONLINEAR OPTICS, APPLICATIONS

Contents

Phase Matching
Pulse Compression via Nonlinear Optics
Raman Lasers
Self-Focusing and Related Effects (Solitons and Multiphoton Absorption)
Three-Dimensional Microfabrication

Phase Matching

A V Smith, Sandia National Laboratories,
Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase matching is important in any process that coherently transfers energy from one optical mode to another over a long interaction length. We will start by illustrating the principle, by imagining that the original mode is a single transverse mode of a waveguide, and that microscopic scattering centers distributed uniformly throughout the waveguide weakly scatter light into a second mode with orthogonal polarization. Then suppose that the scattering process preserves phase so the phase of the wavelet scattered at each point into mode two is the same as that of the original wave in mode one. Such a process is shown in Figure 1 where the propagation constant of the initially excited mode is k_1 and the propagation constant of the second mode is k_2. At each z location the light scattered into the second mode has the phase $\phi = k_1 z$, and after propagating to the end of the waveguide ($z = L$) with the phase velocity characteristic of the second mode, it has the phase $\phi_z = [k_1 z + k_2(L - z)] = [k_2 L + z(k_1 - k_2)]$. Coherently summing the fields scattered at all of the z positions, gives the net output field in the second mode. Figure 2 shows how this field can be constructed graphically. The rotation angle of each phasor represents its phase, ϕ_z, while the length of each phasor represents the magnitude of the field contribution from the corresponding z segment. It is apparent that the energy transfer is maximized if the contributions from all z positions have the same phase, in which case the phasors form a straight line, and the power in the second mode is proportional to L^2. This corresponds to ($k_1 - k_2 = 0$). When this condition is met the waves in the two

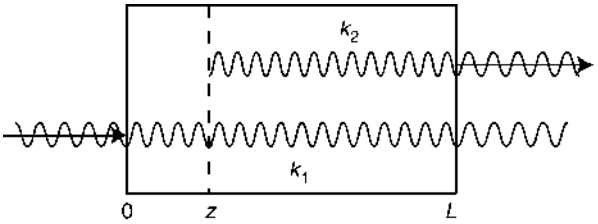

Figure 1 An input wave with propagation constant k_1 in a waveguide mode is weakly scattered into a second waveguide mode with propagation constant k_2. The phase mismatch is $\Delta k \equiv k_2 - k_1$.

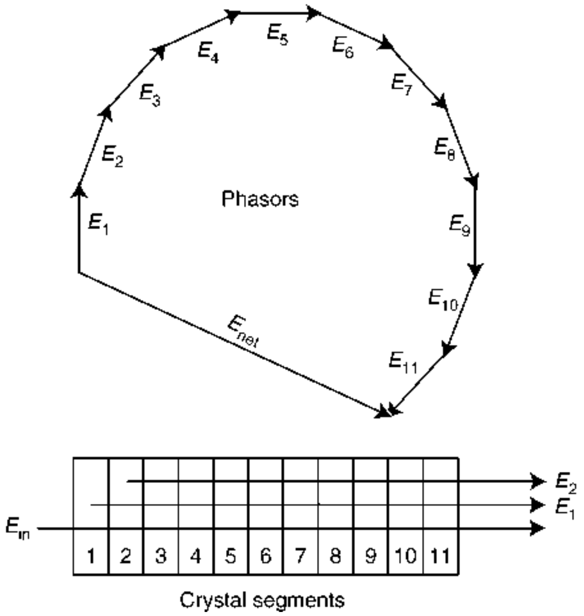

Figure 2 Phasor diagram showing the net output field contributions from segments 1–11. The net field at the output end, E_{net}, is the vector sum of the phasors. The phasors are complex quantities whose angles represent their relative phases according to $E_n = E_o \exp(i\phi_n)$ where $\phi_n = \phi_o - \Delta k(z_n - z_1)$.

modes have equal phase velocities, and the process is phase matched.

The quantity $(k_2 - k_1)$ is called Δk, the phase mismatch. If $\Delta k \neq 0$ the phasors curl to form a section of a circle, reducing the net field in mode two. If all the phasors have the same length, as they would if the wave in mode one is undepleted, the size of the mode two field at position z is

$$E(z) = Kz\frac{\sin(\Delta kz/2)}{\Delta kz/2} = Kz\,\mathrm{sinc}\left(\frac{\Delta kz}{2}\right) \qquad [1]$$

where K is a constant describing the scattering strength. The corresponding power in mode two is plotted in Figure 3 for various values of Δk. When $\Delta k = 0$ the power grows quadratically with z, but for $\Delta k \neq 0$ it is proportional to $\sin^2(\Delta kz/2)/\Delta k^2$. If we plot the normalized output power in mode two $(z = L)$ versus Δk, we have the curve shown in Figure 4 which is $\mathrm{sinc}^2(\Delta kL/2)$.

When $\Delta k = 2\pi/L$ the phasors form a complete circle so the power in mode two first grows with z but then shrinks back to zero at $z = L$. For a given value of Δk the length for which the phasors form a complete circle is known as the coherence length, $L_{\mathrm{coh}} = 2\pi/\Delta k$.

Reducing the z increments to infinitesimals, we can write the output field for mode two in integral form:

$$E_{\mathrm{net}} = \int_0^L S(z)e^{-i\Delta kz}\,\mathrm{d}z \qquad [2]$$

This shows that when Δk is independent of z, the output field is the Fourier transform of a source term, $S(z)$, transformed into Δk space. Where $S(z)$ has a constant value from $z = 0$ to $z = L$ and is zero outside that range, the Fourier transform has the form given by eqn [1], with $z = L$.

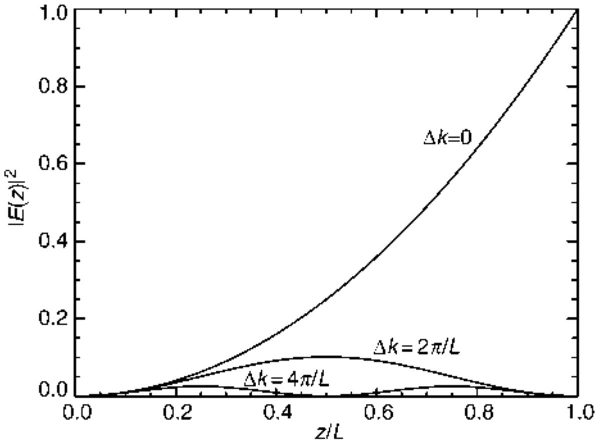

Figure 3 Plot of power in second mode versus z for different values of Δk. For $\Delta k = 0$ the power is proportional to z^2. For $\Delta k = 2\pi/L$ the phasors form a closed circle so the net field is zero at $z = L$. For $\Delta k = 4\pi/L$ the phasors trace two complete overlapped circles so again the field is zero at the $z = L/2$ and at $z = L$.

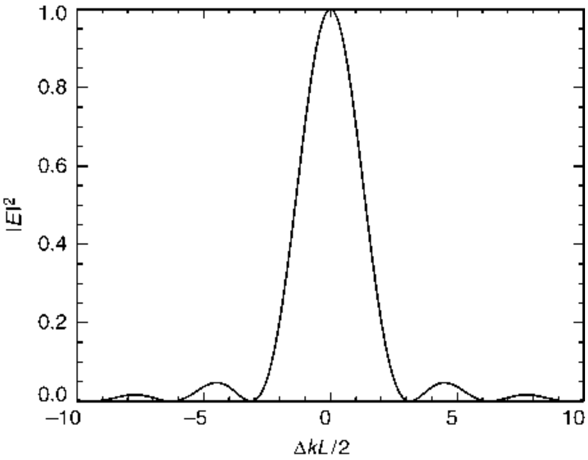

Figure 4 The plotted function indicates the relative irradiance of the output field, $|E|^2$, versus the phase mismatch Δk. The full width at half maximum is $\Delta k = 1.77\pi/L = 5.57/L$.

The $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ form for the output power is ubiquitous in functions that describe the efficiency of nonlinear optical interactions, as well as acousto-optic and electro-optic modulation of light waves. However, it should be kept in mind that $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ holds only when $S(z)$ is independent of z. If the scattering is strong enough to alter the phase or amplitude of the driving wave, or if the coupling strength varies with z for any other reason, $S(z)$ is not independent of z and the $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ dependence is modified.

Our simple example relied on scattering in a waveguide, with a source term that was linear in the field in mode one. In practice, most optical devices that couple light between modes, rely on nonlinear interactions among multiple optical fields, or between an optical field and an acoustic or an rf electric wave. These nonlinear interactions produce a polarization wave in a suitable material at a sum or difference frequency. Each point of this polarization radiates a wavelet in all directions. These wavelets interfere constructively only in directions that are nearly phase matched. Examples include traveling wave acousto-optic filters, in which an acoustic wave traveling parallel to an input light beam of one polarization in a birefringent material, couples light into the orthogonal polarization via the elasto-optic effect. Traveling wave electro-optic modulators are similar but rely on the linear electro-optic effect for the coupling. Optical frequency mixing uses the nonlinearity of various materials to mix multiple optical waves, creating polarization waves at sum and difference frequencies. A wide variety of such nonlinear interactions is discussed in textbooks on nonlinear optics, electro-optics, and acousto-optics. In all of these cases the phase velocity mismatch of interest is that between an induced polarization wave and a radiated mode, and the phase mismatch is $\Delta k = k_{\text{radiation}} - k_{\text{polarization}}$.

To find $S(z)$ and Δk for these nonlinear mixing processes, we must construct the polarization wave. We consider only monochromatic plane waves here, considering beams with spatial or temporal structure in a later section. The electric field of a monochromatic plane optical wave propagating in direction $\mathbf{r}_\text{o}$ can be described by

$$\begin{aligned} E &= E_\text{o} \cos(\omega_\text{o} t - \mathbf{k}_\text{o} \cdot \mathbf{r}_\text{o} + \phi_\text{o}) \\ &= \tfrac{1}{2}[E\, e^{-i(\omega_\text{o} t - \mathbf{k}_\text{o} \cdot \mathbf{r}_\text{o})} + E^*\, e^{i(\omega_\text{o} t - \mathbf{k}_\text{o} \cdot \mathbf{r}_\text{o})}] \end{aligned} \qquad [3]$$

The complex notation is a convenience that simplifies algebraic book-keeping. Similarly, an acoustic field propagating in direction $\mathbf{r}_\text{a}$ can be written:

$$\begin{aligned} A &= A_\text{o} \cos(\omega_\text{a} t - \mathbf{k}_\text{a} \cdot \mathbf{r}_\text{a} + \phi_\text{a}) \\ &= \tfrac{1}{2}[A\, e^{-i(\omega_\text{a} t - \mathbf{k}_\text{a} \cdot \mathbf{r}_\text{a})} + A^*\, e^{i(\omega_\text{a} t - \mathbf{k}_\text{a} \cdot \mathbf{r}_\text{a})}] \end{aligned} \qquad [4]$$

In a suitable material the polarization induced by the product of the optical and acoustic waves has a term of the form:

$$P = sEA \qquad [5]$$

P has components at both the sum and difference frequency. The sum frequency component is:

$$P_{\omega_\text{o}+\omega_\text{a}} = \frac{1}{4} s[EA\, e^{-i(\omega_\text{o}+\omega_\text{a})t + i(\mathbf{k}_\text{o}+\mathbf{k}_\text{a})\cdot\mathbf{r}} + \text{cc}] \qquad [6]$$

where ‘cc’ indicates the complex conjugate of the first term. The component at the difference frequency is

$$P_{\omega_\text{o}-\omega_\text{a}} = \frac{1}{4} s[EA^*\, e^{-i(\omega_\text{o}-\omega_\text{a})t + i(\mathbf{k}_\text{o}-\mathbf{k}_\text{a})\cdot\mathbf{r}} + \text{cc}] \qquad [7]$$

An important point is that the k vector of the sum-frequency polarization is $(\mathbf{k}_\text{o} + \mathbf{k}_\text{a})$. Phase matching requires that the radiated sum-frequency wave has the same propagation vector. The radiated difference frequency wave must have the propagation vector $(\mathbf{k}_\text{o} - \mathbf{k}_\text{a})$ to be phase matched. Phase matching for these two cases is shown in Figure 5. Note the general rule that if the frequency of a wave appears as $+\omega$ in the induced polarization, the associated k vector contribution to the polarization wave is $+\mathbf{k}$, while if it appears as $-\omega$, the k vector contribution is $-\mathbf{k}$. For example, the acoustic wave in $P_{\omega_\text{o}+\omega_\text{a}}$ is associated with $+\mathbf{k}_\text{a}$ contribution to the polarization wave, while in $P_{\omega_\text{o}-\omega_\text{a}}$ it is associated with a $-\mathbf{k}_\text{a}$ contribution.

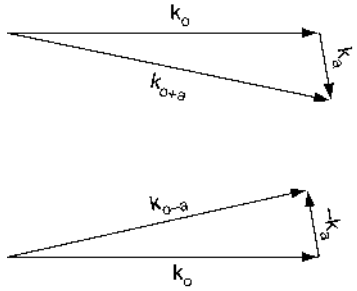

Figure 5 An optical wave with propagation vector k_o and an acoustic wave with propagation vector k_a produce a polarization wave at the sum frequency with the propagation vector k_{o+a} (top diagram), and a polarization wave at the difference frequency with the propagation vector k_{o-a} (bottom diagram).

Using the definition of k:

$$\mathbf{k} = \frac{2\pi}{\lambda}\hat{\mathbf{k}} = \frac{\omega}{v}\hat{\mathbf{k}} \qquad [8]$$

where v is the phase velocity, we can write Δk as

$$\begin{aligned} \Delta\mathbf{k} &= \mathbf{k}_{\text{o+a}} - \mathbf{k}_\text{o} - \mathbf{k}_\text{a} \\ &= \left(\frac{\omega_{\text{o+a}}}{v_{\text{o+a}}}\hat{\mathbf{k}}_{\text{o+a}} - \frac{\omega_\text{o}}{v_\text{o}}\hat{\mathbf{k}}_\text{o} - \frac{\omega_\text{a}}{v_\text{a}}\hat{\mathbf{k}}_\text{a} \right) \end{aligned} \qquad [9]$$

Clearly, if all of the waves had equal phase velocities, phase matching would be achieved for collinear interactions, where all of the propagation vectors are parallel, because the frequencies inside the parentheses sum to zero. This is seldom the case as for example, the phase velocity of an acoustic wave is much less than that of an optical wave, making an acoustic wave’s k-vector disproportionately long compared with that of an optical wave. On the other hand, an rf wave usually has a disproportionately short k-vector compared with an optical wave. Even when all of the waves are optical, dispersion of the refractive index usually makes the k-vector of the bluer waves disproportionately long compared with those of the red waves. A wide variety of methods are used separately or in combinations to compensate for these effects in order to achieve phase matching and these are discussed below.

Phase Matching Techniques

Noncollinear Phase Matching

If one of the low frequency waves in a three wave mixing process, such as the acousto-optic interaction discussed above, has a lower phase velocity than the bluest wave, then noncollinear phase matching may be possible. Acousto-optic deflectors are often phase matched with a noncollinear geometry, such as that shown in Figure 6. An input optical wave of frequency ω_o reflects from a traveling acoustic wave with frequency ω_a in an optically isotropic material to produce an output optical wave of frequency $\omega_o + \omega_a$. The Bragg condition for efficient reflection is

$$\sin\theta = \frac{k_a}{2k_o} \qquad [10]$$

which may be seen from the figure to be the same as the phase matching condition, assuming k_{out} and k_{in} have the same magnitude, which is a good approximation for typical acoustic frequencies less than a few GHz.

Stimulated Brillouin scattering is another example of noncollinear mixing of acoustic and optical waves. One possible geometry has the input optical and acoustic waves traveling in opposite directions to create a reflected optical wave at the sum frequency. The phase matching condition is then

$$k_{\text{acoustic}} \approx 2k_{\text{optical}} \qquad [11]$$

which implies

$$\omega_{\text{acoustic}} = \frac{2n\omega_{\text{optical}}}{c} v_{\text{acoustic}} \qquad [12]$$

The acoustic frequency that phase matches such counter propagating Brillouin scattering is usually a few GHz in liquids and solids.

If noncollinear phase matching in an isotropic material is not possible, as is usually the case for electro-optic modulation and for all-optical sum and difference frequency mixing, it is sometimes possible to use the properties of nonisotropic media to adjust the phase velocity of one or more of the waves. There are several possibilities, such as natural birefringence, form birefringence, photonic band edge tuning, wave guide modal propagation, and various forms of quasi phase matching. Each is briefly described below.

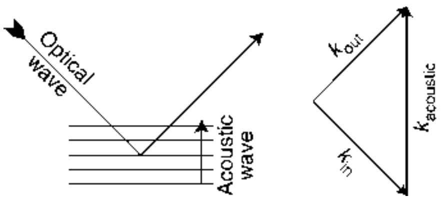

Figure 6 Phase matching in an acousto-optic deflector.

Birefringent Phase Matching

Natural birefringence

Naturally birefringent crystals can be categorized as either uniaxial or biaxial. In either case, light propagating along any direction (except along an optical axis) has two orthogonal eigen polarizations with different refractive indices. In uniaxial crystals, one of these indices is independent of the propagation direction and is called the ordinary index, n_o. The refractive index for the other polarization, known as the extraordinary polarization, is given by

$$\frac{1}{n_e^2(\theta)} = \frac{\sin^2\theta}{n_e^2} + \frac{\cos^2\theta}{n_o^2} \qquad [13]$$

where θ specifies the propagation direction measured from the z, or optical, axis, and n_e and n_o are the refractive indices for light polarized parallel and perpendicular to z. The two eigen polarizations are aligned with the lines of constant or latitude (o-wave) or constant longitude (e-wave). The variation of the two refractive indices with propagation angle θ is shown in Figure 7 for a negative uniaxial crystal which is characterized by $n_e < n_o$, and for a positive uniaxial crystal, characterized by $n_e > n_o$.

The acousto-optic tunable filter relies on natural birefringence to achieve a phase match. An optical and an acoustic wave propagate normal to the optical axis of a uniaxial crystal. If the crystal is negative uniaxial, with an input optical polarization of 'e', and an output optical polarization of 'o', the k vector of the acoustic wave must make up the difference

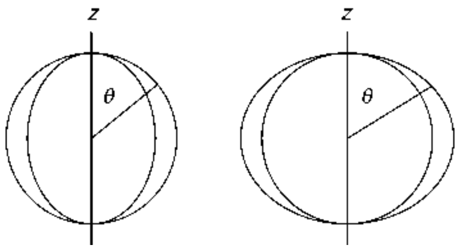

Figure 7 The circle and ellipse represent the two refractive indices for a negative uniaxial crystal (left diagram) and a positive uniaxial crystal (right diagram) for different propagation directions specified by the angles (θ, ϕ) with ϕ fixed. These curves are independent of the azimuthal angle ϕ, so in three dimensions the ordinary refractive index, n_o, forms a sphere in both cases, while the extraordinary index, n_e, forms a prolate ellipsoid for a negative uniaxial crystal, or an oblate ellipsoid for a positive uniaxial crystal.

$k_o - k_e$. This may be expressed as

$$k_a = \frac{\omega_a}{v_a} = \frac{(\omega + \omega_a)n_o}{c} - \frac{\omega n_e}{c} \qquad [14]$$

or, ignoring the difference between ω and $\omega + \omega_a$, the optical frequency that is phase matched is

$$\omega = \frac{\omega_a c}{v_a \Delta n} \qquad [15]$$

where v_a is the phase velocity of the acoustic wave, and $\Delta n = n_o - n_e$. The phase matched optical frequency is seen to be tunable with ω_a. A typical acoustic frequency for a visible filter is a few hundred MHz. The phase matching bandwidth set by the $\mathrm{sinc}^2(\Delta kL/2)$ function can be quite narrow, with a subnanometer optical bandpass in a crystal a few cm long.

Natural birefringence is also widely used to phase match all-optical sum and difference frequency mixing. To counter normal optical dispersion, the bluest wave is polarized in the direction associated with the smaller refractive index, while at least one of the two redder waves has the opposite polarization. By adjusting the propagation angle, θ, the refractive index of the extraordinary waves can be adjusted until exact phase matching is achieved. If the two redder waves both have the high index polarization the process is sometimes called type I, whereas if they have opposite polarizations it is called type II.

Figure 8 illustrates type I phase matched second harmonic generation in a negative uniaxial crystal. The fundamental wave with frequency ω is o-polarized, while the second harmonic wave at frequency 2ω is e-polarized. The circle drawn in dashes represents the line swept out by the polarization wave vector $2k_\omega$ as the propagation direction θ is swept, while the ellipse represents the loci of $k_{2\omega}$ for the same angle. Collinear phase matching is achieved at the point of intersection of the two curves. This sets the angle θ_{pm}. For angles greater than θ_{pm}, dispersion is over-compensated and $\Delta k < 0$ so noncollinear phase matching is possible. All ϕ angles have equivalent propagation properties, so the loci of phase matched propagation directions form a cone at $\theta = \theta_{\mathrm{pm}}$. Usually the propagation angle ϕ is chosen to maximize the nonlinear response of the crystal.

Biaxial crystals share with uniaxial crystals the property of two orthogonal eigen polarizations with different refractive indices, but the angular dependence of the indices is more complex. The two indices associated with propagation along direction (θ, ϕ) can be found by solving

$$\frac{\sin^2\theta\,\cos^2\phi}{1/n^2 - 1/n_x^2} + \frac{\sin^2\theta\,\sin^2\phi}{1/n^2 - 1/n_y^2} + \frac{\cos^2\phi}{1/n^2 - 1/n_z^2} = 0 \qquad [16]$$

for the two possible values of n. Here n_x, n_y, and n_z are the refractive indices for light polarized along the respective principal axes, where the convention is $n_x < n_y < n_z$. There are two optical axes lying in the x–z plane at an angle determined by the values of n_x, n_y, and n_z. The phase matching loci are no longer lines of constant latitude as they were for uniaxial crystals, but instead are curves that encircle either the optical axes or one of the principal axes. Figure 9 shows an

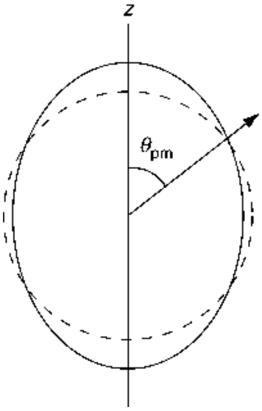

Figure 8 Collinear phase matching of second harmonic generation in a negative uniaxial crystal. The second harmonic light is e-polarized with a refractive index indicated by the solid line while the fundamental light is o-polarized as indicated by the dashed line. The crossing point of the two refractive indices determines the phase matching angle θ_{pm}.

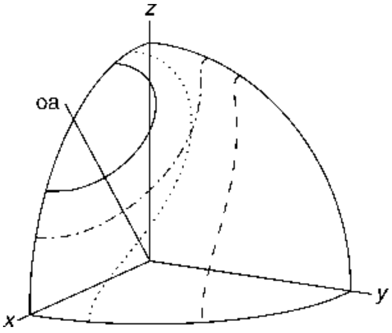

Figure 9 Calculated loci of propagation directions that achieve birefringent phase matching for type I collinear second harmonic generation of different wavelengths in the crystal CsB_3O_5. The fundamental wavelengths are 2,700 nm (dotted curve looping around the *x* axis), 3,000 nm (dashed curve looping around the *y* axis), 1,000 nm (solid curve looping around the optical axis 'oa'), and 800 nm (chained curve looping around the *z* axis). For clarity only a single octant of the sphere of propagation directions is shown. The loci in the other octants of the sphere are formed by reflections in the three principal planes.

example of each type of phase matching curve associated with second harmonic generation of different wavelengths in the crystal CsB_3O_5.

Note that because these curves always intersect at least one of the principal planes, if phase matching in biaxial crystals is possible at all, it is also possible in a principal plane. Biaxial crystals are usually phase matched in a principal plane, but phase matching outside the principal planes can in some cases maximize the nonlinear response, minimize the temperature sensitivity, or adjust the relative group velocities.

Form birefringence

Some materials, such as GaAs, have a strong nonlinear optical response but are not birefringent, so optical frequency mixing cannot be phase matched using natural birefringence. However, form birefringence can be engineered by layered growth, and it can be used to phase match. If many alternating layers of two materials with different refractive indices are stacked together with a layer thickness comparable to the wavelength of light, the material is birefringent for light propagating parallel to the layers. Light polarized with its E field parallel to the layers will have a higher refractive index than light polarized with its field perpendicular to the layers. The strength of birefringence, $(n_{\|} - n_{\perp})$, can be engineered by varying the relative thicknesses and refractive indices of the two layered materials. This has been demonstrated in GaAs/AlAs waveguide structures in which the AlAs layers were oxidized to form Al_2O_3, which has a refractive index of 1.6, compared with an index of 3.4 for GaAs. The resulting large birefringence was used to phase match infrared difference frequency mixing.

Photonic Lattice Phase Matching

The propagation constant of a light wave can be controlled by tuning its frequency near the bandgap of a Bragg structure. The Bragg structure can be a stack of layers with alternating refractive index. If the stack consists of layers of thickness and refractive index l_1, n_1 and l_2, n_2, the Bragg condition at normal incidence is

$$\phi_1 + \phi_2 = \frac{\omega_{\text{Bragg}}}{c}(n_1 l_1 + n_2 l_2) = m\pi \qquad [17]$$

where m is an integer. Frequencies near ω_{Bragg} are strongly reflected. The width of the reflective band of a structure with many layers is proportional to the refractive index contrast, $(n_1 - n_2)/(n_1 + n_2)$. At certain frequencies, on both sides of the bandgap, the reflectivity vanishes. The first transmission peak on the red side of the bandgap has an effective refractive index that is greater than the average $(n_1 + n_2)/2$, while the first peak on the blue side has an effective index less than the average. This effect can be used to adjust the effective value of Δk in the structure. As a bonus, operating on one of the transmission peaks near the bandgap enhances the optical field in the material, increasing the strength of the nonlinear interaction.

It is also possible to use a Bragg structure in reflection mode. Tuning to the red side of the band center increases the effective refractive index of the reflected wave, while tuning to the blue side decreases it. The internal fields are not enhanced in this case and the interaction length may be quite short.

Quasi-Phase Matching (QPM)

Quasi-phase matching uses spatial modulation of the polarization source with a modulation period of $L_{\text{mod}} = 2\pi/\Delta k$ to boost the efficiency of the mixing process even though $\Delta k \neq 0$. We can consider the periodic modulation as another interacting wave that is included in the creation of the polarization wave. The modulation wave has $\omega = 0$ and is stationary with period L_{mod}. Because the modulation is usually a square wave rather than sinusoidal, it is made up of multiple sinusoidal waves with periods L_{mod}/m, all of which can participate in generating a polarization wave. Like traveling waves, the stationary modulation waves can be oriented in any direction, and any Fourier component of the modulation wave with wave vector $\Lambda_m = 2\pi m/L_{\text{mod}}$ can participate in (quasi) phase matching, as shown in Figure 10.

Usually the spatial modulation is achieved by periodically reversing the sign of the nonlinear coupling coefficient. Figure 11 shows the associated phasor diagram for perfect first-order QPM. The net fields from the first two modulation periods are labeled E_1 and E_2. If the sign of the nonlinearity in the first half of the modulation period were not reversed in the second half, the phasors would form a complete circle resulting in a null net field. All following periods would also produce null net fields. However, if the sign is reversed midway through the first period, the phasors from the second half are sign reversed and the net field is E_1 as shown. The magnitude of E_1 is smaller than for a truly phase matched segment by the

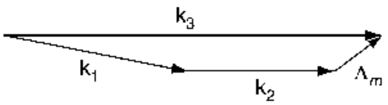

Figure 10 Quasi-phase matching vector diagram for sum frequency mixing. The mth Fourier component of the grating vector is indicated by Λm.

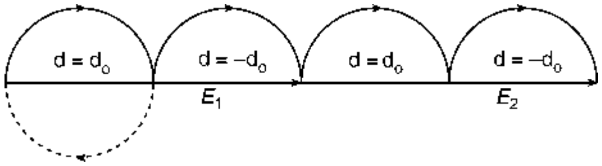

Figure 11 Phasor diagram for a quasi-phase matched interaction in a material with alternating signs of the nonlinear coefficient. The contributions from four segments, each of length $L_{\mathrm{coh}}/2$ are shown. The net field from the first QPM period is labeled E_1, and that from the second period is E_2. The dashed line shows the path the phasors would trace if the nonlinear coefficient were not sign reversed in the even numbered segments.

ratio of the diameter of the circle to its circumference, or by $2/\pi$. Coherence lengths in crystals are typically 5–50 μm so crystals that are several mm long have a large number of periods, in which case the net phasors $E_1, E_2, \ldots$ can themselves be considered elementary phasors that form the net field for the full crystal. For example, if the modulation period L_{mod} is not exactly equal to $2\pi/\Delta k$, QPM is imperfect and phasors $E_1, E_2, \ldots$ curl up to give a net field of $\mathrm{sinc}(\Delta k' L/2)$ where $\Delta k' = \Delta k - 2\pi/L_{\mathrm{mod}}$. For all practical purposes, a QPM crystal containing many modulation periods performs the same as a phase matched crystal with its nonlinear coupling reduced by $2/\pi$.

Referring to eqn [2], we note that periodic modulation of the source term $S(z)$ produces peaks in the phase matching curves at $\Delta k = m2\pi/L_{\mathrm{mod}}$ where m is an integer. If $S(z)$ is a square wave of alternating sign over $0 < z < L$, its Fourier transform is the convolution of $\mathrm{sinc}(\Delta kL/2)$ with a picket fence function with spikes at $\Delta k = m2\pi/L_{\mathrm{mod}}$. If $L_{\mathrm{mod}} \ll L$, the crystal contains many periods of the square wave and the spikes are separated by many times the width of the $\mathrm{sinc}(\Delta kL/2)$ function. This results in well separated sinc functions centered at $\Delta k = 2\pi m/L_{\mathrm{mod}}$, and any process with a coherence length of L_{mod}/m will be quasi-phase matched. That is, one of the Fourier peaks in Δk space will be centered on the corresponding value of Δk.

The magnitudes of the QPM peaks are maximized if the modulation of $S(z)$ is a sign reversal, as described above. If the positive and negative segments have unequal lengths, we can define the duty cycle as $D = l_+/(l_+ + l_-)$. The heights of the pickets in the picket fence are then

$$h_m = \frac{2}{m\pi}\sin(\pi m D) \qquad [18]$$

The largest magnitude peak is for $m = 1$ and $D = 0.5$ in which case $h_1 = 2/\pi$.

Unlike birefringent phase matching, QPM can, in principle, phase match any set of polarizations for the interacting waves. This makes it possible to use a polarization set that maximizes the nonlinear response of the crystal.

Ferroelectric poling

There are various ways to achieve QPM. By far the most popular and versatile is by ferroelectric poling. Ferroelectric crystals have an internal electric field that can be reversed by the application of a sufficiently strong external poling field. Reversing the internal field also reverses the sign of the nonlinearity for some light polarizations but does not alter the refractive indices. Crystals of $\mathrm{LiNbO_3}$ and $\mathrm{KTiOPO_4}$, with thickness up to 1 mm, can be poled with periods as short as 5 μm. This is sufficient to quasi-phase match frequency mixing of visible or near infrared light. Figure 12 shows an example of the orientation of the internal fields of patterned ferroelectric domains.

It is possible to pole a crystal aperiodically so that multiple mixing processes are quasi-phase matched simultaneously. To design the poling pattern for such a device, the Fourier transform of an initial trial pattern can be calculated and compared with the target spectrum in Δk space. Iterative algorithms can be used to modify the poling pattern, zeroing in on the target spectrum.

Ferroelectric poling can also form two-dimensional poling patterns, corresponding to a two-dimensional set of Fourier components in Δk space. It can also produce chirped gratings which have a broad phase matching peak. Chirped gratings can be used to simultaneously frequency convert and tailor the temporal shapes of the output pulse when the input pulses are frequency chirped. Poling can also yield fanned poling patterns so the poling period and thus the phase matching wavelength can be shifted by translating the grating.

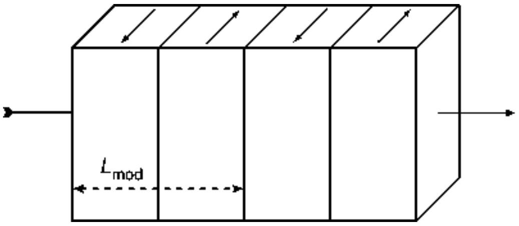

Figure 12 Diagram of a poled ferroelectric crystal. The four crystal segments represent four ferroelectric domains with internal crystal field directions indicated by the arrows. Reversal of the internal field is accompanied by reversal of the nonlinear coefficient, so this is a quasi-phase matched structure with alternating signs of d_{eff}.

Wafer stacking and patterned crystal growth

Another method of quasi-phase matching relies on stacks of thin plates oriented so the sign of the nonlinearity reverses in adjacent plates. Each plate has a thickness of $mL_{\mathrm{coh}}/2$ where m is an odd integer. If all the plates are identically prepared crystals, it is often possible to reverse the sign of their nonlinearity by reorienting them. It is sometimes possible to optically contact the plates or even to diffusion bond them, avoiding reflective losses at the interfaces. This technique is conducive to large aperture QPM crystals. However, the plates are thin so fabrication is a challenge.

Quasi-phase matched crystals of GaAs have also been epitaxially grown on patterned substrates. The patterning serves as a template for growth of zones with alternating crystal orientation and sign of the nonlinear coefficient.

Total internal reflection

Total internal reflection in a thin slab has also been exploited for QPM (Figure 13). Two effects contribute to the source modulation, one is the rephasing of the waves on total internal reflection; the other is the change in nonlinearity that may accompany the change in propagation direction. Usually plates much thicker than one coherence length are used, so the quasi-phase matching is of high order ($m \gg 1$).

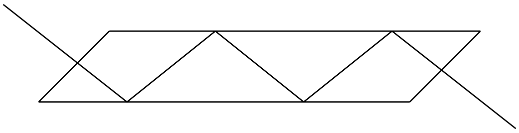

Figure 13 Diagram of quasi-phase matching based on total internal reflection. Relative phases acquired by the waves upon total internal reflection, combined with the path lengths and possible changes in the value of d_{eff} for alternating legs, can be combined to achieve quasi-phase matching.

Wave Guide Modal Phase Matching

In waveguides the size and shape of the guide influence the transverse field distributions and the propagation constants for the transverse modes. Generally the higher order transverse modes have higher phase velocities than the lower order modes. This can be used to counteract index dispersion in three wave mixing if the bluest wave propagates in a higher order mode than the redder waves. A problem frequently encountered in using this effect is that the transverse modes are orthogonal, or nearly so. Because the nonlinear response is proportional to the overlap integral

$$d_{\mathrm{eff}} \propto \iint_A d(x,y)E_1(x,y)E_2(x,y)E_3(x,y)\,dx\,dy \qquad [19]$$

where the E's are the transverse field distributions of the three waves, the nonlinear response is often quite small. If the local nonlinearity $d(x,y)$ can be engineered to maximize the integral, this problem can be alleviated. This may be thought of as transverse quasi-phase matching.

Waveguides can also be designed for Cerenkov phase matching where the phase velocity of the radiated wave is less than that of the polarization wave. Figure 14 shows an example. The emission angle is

$$\cos\theta = \frac{2k_\omega}{k_{2\omega}} = \frac{v_{2\omega}}{v_\omega} \qquad [20]$$

where the v's are phase velocities. The transverse momentum Λ must be supplied by the waveguide. In other words, the source harmonic polarization must have a transverse Fourier component with spatial frequency Λ. This is another example of transverse quasi-phase matching.

Another application of waveguide phase matching is in traveling wave electro-optic modulation, where the rf wave travels along an rf waveguide parallel to the light path. For maximum modulation bandwidth, the phase velocity of the rf wave should equal that of the light wave over the desired modulation bandwidth. Because the phase velocity of the rf wave in the bulk medium is usually faster than that of the light wave, the rf guide must be designed for a slow phase velocity.

Compensated Phase Matching (CPM)

Compensated phase matching is somewhat like QPM in that multiple crystal segments are used, but it differs by requiring that sequential crystal segments have ΔkL's of equal magnitude but opposite sign. Ideally the nonlinearity would be the same in all segments. Figure 15 shows the phasor diagram for one such arrangement, in this case with unequal nonlinearity in alternating segments. CPM is realized

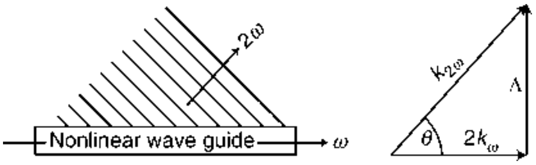

Figure 14 Cerenkov phase matched second harmonic generation. The phase velocity of the fundamental wave in the waveguide is greater than the phase velocity of the harmonic wave in the cladding.

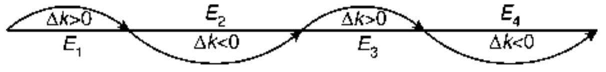

Figure 15 Phasor diagram for four segments of a device with compensated phase matching device. The magnitude of ΔkL is equal in each segment but the sign of Δk alternates. The length of the E_n field from segment n is proportional to sinc($\Delta kL/2$) so the net field ΣE_n must also be proportional to this function.

in walkoff compensated geometries where a sequence of crystals is arranged so $\Delta k = 0$ in each for a particular monochromatic plane wave. In the case of spatial walkoff compensation the walkoff direction in alternating crystals is reversed so the walkoff is exactly compensated in pairs of adjacent crystals. Tilting the input beams by a small amount changes the values of the Δk's in such a way that CPM is achieved. In the case of temporal walkoff compensation, temporal walkoff is reversed in alternating crystals, in which case tuning the input frequency leads to CPM. The advantage of CPM is that the acceptance bands, which will be discussed in the next section, are those of the individual crystal segments which are larger than those of a single crystal of equivalent length without walkoff compensation.

Spatial and Temporal Effects

Acceptance Bands

Our discussion to this point has been based on monochromatic plane waves. Real beams have spatial and temporal structure which means they can be decomposed into sets of monochromatic plane waves with tilted propagation directions relative to the nominal direction, and with detuned frequencies relative to the nominal frequency. As we showed above for monochromatic plane waves, there is a range of angles and frequencies, called the acceptance bands, over which $\Delta kL < 2\pi$. If the interacting beams' angular spectra and frequency spectra lie within these bands, the process is well phase matched for all angular and frequency components, and efficiency is maximized. This statement is equivalent to saying that if the spatial walkoff due to beam tilts and birefringence, and the temporal walkoff due to differing group velocities are small enough that the interacting beams stay well overlapped in space and time, the efficiency of energy transfer to the output channel does not suffer. Consider, for example, second harmonic generation of short pulses. The frequency acceptance bandwidth for the fundamental wave, $\Delta\omega_1$, is found using

$$\frac{\mathrm{d}(\Delta k)}{\mathrm{d}\omega_1}\Delta\omega_1 = \left(\frac{2\mathrm{d}k_2}{\mathrm{d}\omega_2} - \frac{2\mathrm{d}k_1}{\mathrm{d}\omega_1}\right)\Delta\omega_1 = \frac{2\pi}{L} \qquad [21]$$

Using the usual definition of the group velocity:

$$v_\mathrm{g} = \frac{\mathrm{d}\omega}{\mathrm{d}k} \qquad [22]$$

this can be written

$$\Delta\omega_1 = \frac{\pi}{L(1/v_{\mathrm{g}2} - 1/v_{\mathrm{g}1})} = \frac{\pi}{\tau_2 - \tau_1} = \frac{\pi}{\Delta\tau} \qquad [23]$$

where τ_1 and τ_2 are the passage times through a crystal of length L for the fundamental and harmonic pulses, and $\Delta\tau = (\tau_2 - \tau_1)$ is the temporal walkoff between the two. This shows that the acceptance bandwidth is approximately equal to the transform bandwidth of a pulse whose duration is equal to the temporal walkoff $\Delta\tau$. Input pulses shorter than $\Delta\tau$ will not convert efficiently because the fundamental and harmonic pulses do not remain overlapped as they propagate through the crystal. In the frequency domain it is said that pulses shorter than the $\Delta\tau$ have spectral widths larger than the acceptance bandwidth, meaning they are not well phase matched across their full spectra, and hence do not convert efficiently. Parallel arguments apply to spatial walkoff and angular acceptance bandwidths. Beams that do not remain spatially overlapped through the crystal do not convert efficiently. Equivalently, if the angular spectra of the beams exceed the acceptance angle, efficiency suffers.

Sometimes these notions are described in terms of an effective interaction length or an effective crystal length. For short pulses the effective interaction length is that for which temporal walkoff is equal to the pulse duration. Equivalently it is the length for which the acceptance bandwidth is equal to the spectral width of the pulse. Crystals shorter than this will have mixing efficiencies nearly equal to that expected for a phase matched process, while longer crystals have efficiencies that fall short of this expectation.

Focused beams

Focused Gaussian beams are a particularly important class of spatially structured beams. In passing through a focus, a lowest order Gaussian beam acquires a phase shift relative to an unfocused beam. The phase velocity near the focus is higher for the focused beam, leading to a phase shift of π through the focus. For second harmonic generation the phase shift of the second harmonic polarization is twice this, or 2π.

A second harmonic wave also acquires a phase shift of π, but this leaves a π phase shift between the harmonic polarization source wave and a freely propagating harmonic wave. The phasors representing harmonic radiation from symmetrically located points well before and well after the focus will therefore be 180° out of phase. Point pairs closer to the focus will have less severe cancelation, but it is not surprising that the maximum efficiency is achieved, not at $\Delta k = 0$, but for a slightly negative value Δk. The phase matching function $\mathrm{sinc}(\Delta kL/2)$ is shifted and distorted for focused beams. Its exact shape depends on the nature of the mixing process. For example, the phase matching peak is different for third harmonic generation and second harmonic generation, and it is different for mixing with small birefringence and mixing with large birefringence.

Phase Conjugation

Degenerate four-wave mixing

Some wave mixing processes are automatically phase matched. One example is phase conjugation by degenerate four wave mixing. This process creates a polarization wave that is the time reversed image of one of the input waves. Figure 16 compares the reflection from a conventional mirror with that from a phase conjugate mirror.

The mixing process shown in Figure 17 is one that can act as a phase conjugate mirror. Waves *f* and *b* are strong pump waves that are phase conjugates of one another. In their presence a weak signal field at the same frequency produces a polarization given by

$$P = sE_f E_b E_s^* \, e^{-i\omega t + i(\mathbf{k}_f - \mathbf{k}_b - \mathbf{k}_s)\cdot \mathbf{r}} \qquad [24]$$

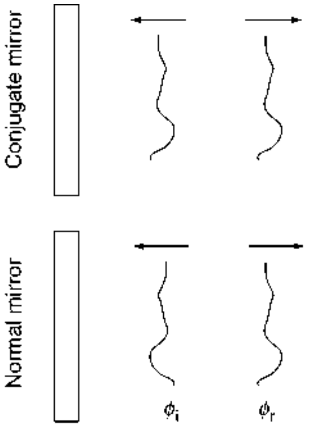

Figure 16 Comparison of a normal and a phase conjugate mirror. The wavy lines represent the phases of the incident and reflected waves.

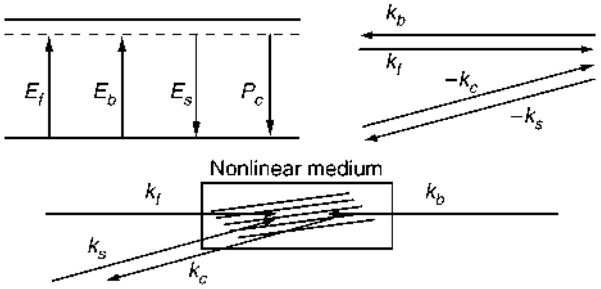

Figure 17 Phase conjugation via degenerate four wave mixing near a one photon resonance. Two counter propagating pump waves with fields E_f and E_b interact with a signal beam, E_s, to produce a polarization wave P_c where the frequencies obey $\omega_f - \omega_s + \omega_b = \omega_c$ (upper left), implying the *k*-vector of the polarization wave is given by $k_c = k_f - k_s + k_b$ (upper right). The lower diagram shows the grating induced by waves *f* and *s* that scatters wave *b* into the direction of *c*.

The two pump propagation vectors cancel one another leaving

$$P = sE_f E_b (E_s^* \, e^{-i\omega t - i\mathbf{k}_s\cdot \mathbf{r}}) \qquad [25]$$

The quantity in parentheses is the phase conjugate of the signal wave, so the polarization wave exactly phase matches a radiated conjugate wave.

Brillouin scattering

Other possibilities for automatic phase matching are processes that involve a stimulated wave in the nonlinear medium such as an acoustic wave (Brillouin scattering) or waves of coherent molecular vibration or rotation (Raman scattering). Phase conjugation, based on Brillouin scattering, is widely used to reduce phase distortions acquired on passing through an amplifier, for example. A single beam of light is focused into a material that has a high polarizability. Example media include high pressure gases Xe and SF_6, liquids acetone and freon, and fused silica. The conjugation process is not as simple to describe as it was in the case of four wave mixing because many acoustic waves can be generated. The key is that the phase conjugate wave has the highest gain under certain conditions and can dominate the Brillouin scattering process, so the reflected wave is nearly a phase conjugate of the incoming wave.

See also

Modulators: Acousto-Optics; Electro-Optics. **Nonlinear Optics, Basics:** Cascading; Four-Wave Mixing; Nonlinear Optical Phase Conjugation; $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources**: Harmonic Generation in Gases. **Optical Parametric Devices:** Overview. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed).

Phase Control: Phase Conjugation and Image Correction. **Photonic Crystals:** Nonlinear Optics in Photonic Crystal Fibers. **Scattering**: Raman Scattering.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Boyd RW (1992) *Nonlinear Optics*. New York: Academic Press.

Dmitriev VG, Gurzadyan GG and Nikogosyan DN (1999) *Handbook of Nonlinear Optical Crystals*. Berlin: Springer-Verlag.

Hanna DC, Yuratich MA and Cotter D (1979) *Nonlinear Optics of Free Atoms and Molecules*. Berlin: Springer-Verlag.

Mills DL (1991) *Nonlinear Optics*. Berlin: Springer-Verlag.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Yariv A and Yeh P (1984) *Optical Waves in Crystals*. New York: John Wiley & Sons.

Zel'dovich BY, Pilipetsky NF and Shkunov VV (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: John Wiley & Sons.

Pulse Compression via Nonlinear Optics

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ultrashort pulses have a great potential in several areas of research and application, such as measurements of ultrafast physical processes, infrared time-resolved spectroscopy, and sampling systems, but they are also invaluable for communications. With increasing demands being put on the performance of optical fiber communications, the emphasis now is on producing a source of ultrashort optical pulses that can be used in high-bit rate and long-distance optical communication systems.

Several optical pulse compression techniques are described in the literature. These techniques utilize dispersion, possibly from gratings or prisms, simultaneously to compensate for chirp and achieve pulse compression. In fact, when a chirped optical pulse propagates in a linear dispersive medium, it acquires a dispersion-induced chirp. If the initial chirp is in the opposite direction to that imposed by the group-velocity dispersion (GVD), the two tend to cancel each other, resulting in the compression of the optical pulse.

Early work on optical pulse compression did not make use of any nonlinear optical effects. Only during the 1980s, when the use of single-mode silica fibers as nonlinear media became widespread, the self-phase modulation (SPM) effect was used to achieve pulse compression. It has led, in 1987, to the production of optical pulses as short as 6 fs in the 620 nm wavelength region. Such advances were possible, only after having understood the evolution of optical pulses in silica fibers.

We can consider two main categories of pulse compressors which make use of nonlinear effects in optical fibers: grating-fiber and soliton-effect compressors. In a grating-fiber compressor, the input pulse is propagated in the normal-dispersion regime of the fiber, which imposes a nearly linear, positive chirp on the pulse through a combination of SPM and GVD. The output pulse is then sent through a grating pair where it experiences anomalous GVD and gets compressed. In the case of a soliton-effect compressor, the fiber itself acts as a compressor without the need of an external grating pair. The input pulse propagates in the anomalous-GVD regime of the fiber and is compressed through an interplay between SPM and GVD. This compression mechanism is related to a fundamental property of the higher-order solitons. These solitons follow a periodic evolution pattern such that they go through an initial narrowing phase at the beginning of each period. If the fiber length is suitably chosen, the input pulses can be compressed by a factor that depends on the soliton order.

Grating-Fiber Compressors

The grating-fiber compressor is generally used at wavelengths $\lambda < 1.3\ \mu\text{m}$. At these wavelengths, optical pulses acquire a positive chirp across its entire width during their passage through conventional silica fibers. A schematic of a grating-fiber compressor in the double pass configuration is shown in Figure 1. After the propagation in the fiber, the optical pulse is sent through a grating pair, which provides the anomalous (or negative) GVD required to get the pulse compression.

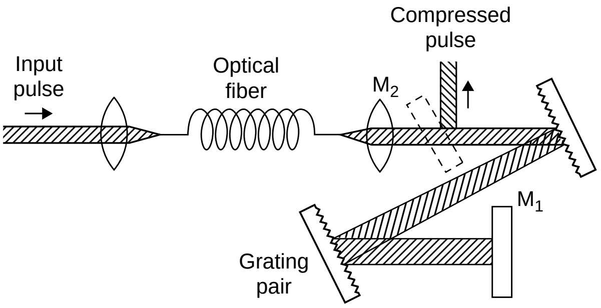

Figure 1 Schematic representation of a grating-fiber compressor in the double-pass configuration. (Reproduced with permission from Agarwal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.)

Different frequency components of an optical pulse incident at one grating are diffracted at slightly different angles. As a consequence, they experience different time delays during their passage through the grating pair, the blue-shifted components arriving earlier than the red-shifted ones. In the case of an optical pulse with a positive chirp, the blue-shifted components occur near the trailing edge of the pulse and the passage through the grating pair provides its compression.

A limitation of the grating pair is that the spectral components of the pulse are dispersed not only temporally but also spatially. As a consequence, the optical beam becomes deformed, which is undesirable. This problem can be avoided simply by using a mirror to reflect the pulse back through the grating pair. Mirror M_1, in Figure 1, performs this function. Reversing the direction of propagation not only allows the beam to recover its original cross-section, but also doubles the amount of GVD, thereby reducing the grating separation by a factor of 2. The mirror M_2 in Figure 1 is used to deflect the compressed pulse out of the compressor.

The phase shift acquired by the spectral component of the pulse at the frequency ω passing through the grating pair can be written in the form:

$$\phi_g = \phi_0 + \phi_1(\omega - \omega_0) + \frac{1}{2}\phi_2(\omega - \omega_0)^2 + \frac{1}{6}\phi_3(\omega - \omega_0)^3 + \cdots \qquad [1]$$

where ω_0 is the pulse center frequency, ϕ_0 is a constant, ϕ_1 is a constant delay, and ϕ_2 and ϕ_3 are parameters that take into account the GVD effects associated with the grating pairs. These parameters depend on the grating period (line spacing), as well as on the orientation and separation of the two gratings. In most cases of practical interest, the spectral width of the pulse satisfies the condition $\Delta\omega \ll \omega_0$ and the cubic and higher-order terms in eqn [1] can be neglected.

To achieve optimum performance from a grating-fiber compressor, it is necessary to optimize the fiber length, as well as the grating separation. Concerning the first aspect, the effects of both GVD and SPM during the propagation of the pulse inside the fiber must be considered. SPM alone determines a linear chirp only over the central part of an optical pulse. Since the grating pair compresses only this region, while a significant amount of energy remains in the wings, the compressed pulse is not of high quality in this case. The effect of GVD turns out to be important, since it broadens and reshapes the pulse, which develops a nearly linear chirp across its entire width. In these circumstances, the grating pair can compress most of the pulse energy into a narrow pulse.

The balanced effects of both GVD and SPM explain the existence of an optimum fiber length, z_{opt}, for maximum pulse compression. In fact, when the fiber length is less than z_{opt}, the SPM-induced chirp has not yet been linearized. On the other hand, when it is longer than z_{opt}, the SPM effects become negligible due to the GVD-induced pulse broadening.

If the input pulse is unchirped and presents a hyperbolic secant shape, the optimum fiber length for maximum pulse compression is well approximated by

$$z_{\mathrm{opt}} \approx (6L_{\mathrm{D}}L_{\mathrm{NL}})^{1/2} \qquad [2]$$

where

$$L_{\mathrm{D}} = \frac{T_0^2}{|\beta_2|} \qquad [3]$$

and

$$L_{\mathrm{NL}} = \frac{1}{\gamma P_0} \qquad [4]$$

are length scales associated with the GVD and the SMP effects, respectively. In eqns [3] and [4], T_0 is the input pulsewidth, β_2 is the fiber GVD parameter, γ is the nonlinearity parameter of the fiber, and P_0 is the peak power of the input pulse.

If the input pulse presents a negative chirp, this must be compensated for by the positive chirp provided by the fiber. As a result, the optimum fiber length increases. The opposite occurs in the case of an input pulse with a positive chirp.

The optimum compression factor, F_{opt}, in the case of a grating-pair compressor, is given by

$$F_{\mathrm{opt}} = \frac{T_{\mathrm{FWHM}}}{T_{\mathrm{comp}}} \approx 0.625\left(\frac{L_{\mathrm{D}}}{L_{\mathrm{NL}}}\right)^{1/2} \qquad [5]$$

where T_{FWHM} is the full width at half maximum (FWHM) of the input pulse ($T_{\mathrm{FWHM}} \approx 1.76T_0$ for a hyperbolic secant pulse) and T_{comp} is the FWHM of the compressed pulse. Equations [2] and [5] provide a good estimate even for pulse shapes other than a hyperbolic secant as long as $\frac{L_{\mathrm{D}}}{L_{\mathrm{NL}}} > 100$.

The previous results were obtained neglecting the higher-order nonlinear and dispersive effects, which is acceptable for pulsewidths $T_0 > 0.1$ ps. However, when shorter pulses are used at the input, their spectral width $\Delta\omega$ is large enough that the cubic term in the expansion [1] must be taken into account. In this case, the compression factor turns out to be smaller than that given by eqn [5].

The third-order dispersion (TOD) of the grating pair, resulting from cubic term in eqn [1], has been compensated for in one experiment using a combination of gratings and prisms. Performing such compensation, a reduction of the pulsewidth to about 6 fs has been achieved. Such a pulse at 620 nm consists of only three optical cycles.

Even though the compression factor $F_{\mathrm{c}} \propto \sqrt{P_0}$, the peak power of the input pulse must be kept below the Raman threshold to avoid the loss of pulse energy through the stimulated Raman scattering (SRS) process. If this condition is not then satisfied, besides the energy loss problem, the Raman pulse can interact with the input pulse through cross-phase modulation and deform the linear nature of the frequency chirp. As a result, the compression factor is reduced from the predictions of eqn [5].

Fiber Bragg Gratings

A uniform fiber Bragg grating (FBG) reflects light whose wavelength is within the stop-band centered at the Bragg wavelength. The GVD is anomalous on the high-frequency side of the stop-band and becomes normal on the low-frequency side. Near the stop-band edges, the grating exhibits large GVD. Since a FBG acts as a dispersive delay line, it can be used in place of the bulk-grating pair in a grating-fiber compressor, providing a compact all-fiber device. In fact, a typical 1 cm long fiber grating may provide as much dispersion as a bulk-grating pair with more than one meter spacing. However, the grating induced TOD can affect significantly the quality of the pulses when the optical frequency falls close to the edges of the stop-band.

The compression factor as well as the pulse quality can be significantly improved by using a FBG in which the grating optical period changes along its length – the so-called chirped fiber grating. In this case, different frequency components of the pulse are reflected from different regions of the grating. Such a device has been used since the mid-1990s to compensate for dispersion-induced broadening of pulses in fiber-optic communication systems, and compression factors above 100 have been achieved. In practice, the only disadvantage of a chirped fiber grating is that the compressed pulse is reflected rather than transmitted.

Soliton-Effect Compressors

Optical pulses at wavelengths exceeding 1.3 μm generally experience both SPM and anomalous GVD during their propagation in silica fibers. Such a fiber can act as a compressor by itself without the need of an external grating pair. In the case of a fundamental or first-order soliton, the effects of SPM cancel the effects of anomalous GVD perfectly, and the soliton propagates whilst preserving its hyperbolic secant shape in a lossless optical fiber. A higher-order soliton, on the other hand, changes its shape periodically as it propagates in a fiber. Such a soliton always experiences an initial pulse narrowing phase before recovering its original hyperbolic secant profile at integral multiples of the soliton period z_0, and this behavior is exploited to achieve pulse compression. Optimum compression is achieved when the higher-order soliton leaves the fiber just as it attains its narrowest width.

The evolution of a soliton of order N inside an optical fiber is governed by the nonlinear Schrödinger equation (NLSE), which can be written in the following form:

$$\mathrm{i}\frac{\partial q}{\partial \xi} + \frac{1}{2}\frac{\partial^2 q}{\partial \tau^2} + N^2|q|^2 q = 0 \qquad [6]$$

where q is the normalized amplitude, $\xi = z/L_{\mathrm{D}}$, $\tau = T/T_0$, and the parameter N is the soliton order, given by

$$N^2 = \frac{L_{\mathrm{D}}}{L_{\mathrm{NL}}} = \frac{\gamma P_0 T_0^2}{|\beta_2|} \qquad [7]$$

The soliton period, z_0, is given by

$$z_0 = \frac{\pi}{2}L_{\mathrm{D}} = \frac{\pi T_0^2}{2|\beta_2|} \qquad [8]$$

In writing eqn [6], the fiber losses were neglected, since fiber lengths employed in practice are relatively small. Even though higher-order solitons follow an exact periodic evolution only for integer values of N, eqn [6] can be used to describe pulse evolution for arbitrary values of N.

In practice, soliton effect compression is carried out by initially amplifying optical pulses up to the power level required for the formation of higher-order solitons. The peak optical power of the initial pulse required for the formation of the Nth-order soliton is given by

$$P_N = 3.11 \frac{|\beta_2| N^2}{\gamma T^2_{\mathrm{FWHM}}} \qquad [9]$$

These Nth-order solitons are then passed through the appropriate length of optical fiber to achieve a highly compressed pulse. The optimum fiber length, z_{opt}, and the optimum pulse compression factor, F_{opt}, of a soliton-effect compressor can be estimated from the following empirical relations:

$$z_{\mathrm{opt}} \approx z_0 \left(\frac{0.32}{N} + \frac{1.1}{N^2} \right) \qquad [10]$$

$$F_{\mathrm{opt}} \approx 4.1N \qquad [11]$$

These relations are accurate to within a few percent for $N > 10$.

Compression factors as large as 500 have been attained using soliton-effect compressors. In fact, considering the same input pulse, the soliton-effect compressor provides a pulse compression that is larger with a fiber that is shorter than a grating-fiber compressor. However, the pulse quality is poorer, since the compressed pulse carries only a fraction of the input energy, while the remaining energy appears in the form of a broad pedestal. From a practical point of view, the pedestal is deleterious since it causes the compressed pulse to be unstable, making it unsuitable for some applications. Despite this, soliton-effect compressed pulses can still be useful because there are some techniques that can eliminate the pedestal.

One difficulty faced, when using the soliton-effect compressor, is that pulses with high peak power are required for the formation of high-order solitons in conventional fibers. As suggested by eqn [9], the use of dispersion-shifted fibers (DSFs) with small values of β_2 at the operating wavelength can reduce the peak power required for soliton generation by an order of magnitude. However, because the soliton period z_0 is inversely proportional to $|\beta_2|$ (see eqn [8]), eqn [10], indicates that longer lengths of DSFs will be required for solitons to achieve optimum compression. As a result, the total fiber loss experienced by those solitons will be larger and the loss-induced pulse broadening will have a significant impact on the compressor global performance.

Another problem introduced by the use of DSFs in soliton-effect compressors is the TOD, which generally degrades the quality of the compressed pulse. In the case of soliton-effect compressors using conventional optical fibers, the effects of TOD become significant only when the widths of the compressed pulse become very short ($\ll$1 ps). In DSFs, however, the relative importance of TOD is increased and those effects become more pronounced because the GVD parameter, β_2, is small. In this case, TOD is detrimental in the compression, even of pulses with widths of a few picoseconds, resulting in serious degradation of the optimum compression factor F_{opt}.

The result given by eqn [11] for the optimum compression factor holds only for the case of an ideal soliton-effect compressor, when high-order nonlinear and dispersive effects, such as intrapulse Raman scattering (IRS), TOD and self-steepening, are neglected. However, these effects cannot be neglected for the highly compressed pulses because of their subpicosecond widths and high intensities.

For not-too-short optical pulses (widths >50 fs) propagating not too close to the zero-dispersion wavelength of the fiber, the dominant higher-order effect is IRS, which manifests as a shift of the pulse spectrum toward the red side – the so-called soliton self-frequency shift. As a consequence, the group velocity is reduced and there is a delay of the optical pulse. This delay affects substantially the interplay between GVD and SPM that is responsible for pulse compression.

Interestingly, the IRS effect can be used to improve the quality of the compressed pulse, by removing the pedestal mentioned above. In fact, as a consequence of the change in the group velocity induced by the IRS effect, the sharp narrow spike corresponding to the compressed pulse travels more slowly than the pedestal and separates from it. Moreover, the pedestal can be removed by spectral filtering and a red-shifted, pedestal-free, and highly compressed pulse is then produced.

Compression of Fundamental Solitons

As described in the last section, the propagation of higher-order solitons provide a rapid compression method, but it suffers from the existence of residual pedestals. Some techniques can help to reduce or even eliminate those pedestals, but energy is always wasted in this process.

A less rapid technique that provides better pulse quality is adiabatic amplification of fundamental or first-order solitons, which are of primary importance in the domain of optical communications. Solitons have fixed area, so the increased energy from amplification is accommodated by an increase in power and a decrease in width. To avoid distortion, the amplification per soliton period cannot be too great.

The same effect as adiabatic amplification can be achieved using an optical fiber with dispersion that decreases along the length of the fiber. In fact, for optical solitons, a small variation in the dispersion has a similar perturbative effect as an amplification or loss: such a variation perturbs the equilibrium between the dispersion and nonlinearity in such a way that when, for example, the dispersion decreases, the soliton pulse is compressed. It can be seen from eqn [7] that if the value of $|\beta_2|$ decreases along the fiber and to keep the soliton order N ($N = 1$ for the fundamental soliton), the pulsewidth must indeed decrease as $|\beta_2|^{1/2}$. Hence, the use of fibers with variable dispersion is viewed as a passive and effective method to control optical solitons in soliton communication systems. Dispersion-decreasing fibers (DDFs), in particular, have been recognized to be very useful for high-quality, stable, polarization-insensitive, adiabatic soliton pulse compression and soliton train generation. These fibers can be made by tapering the core diameter of a single-mode fiber during the drawing process, and hence changing the waveguide contribution to the second-order dispersion.

In a DDF of certain length, the dispersion is monotonically and smoothly decreased from an initial value to a smaller value at the end of the length according to some specified profile. Provided the dispersion variation in the DDF is sufficiently gradual, soliton compression can be an adiabatic process where an input fundamental soliton pulse can be ideally compressed as it propagates, while retaining its soliton character and conserving the energy.

The evolution of the fundamental soliton in a DDF can be described by the following NLSE:

$$\mathrm{i}\frac{\partial q}{\partial \xi} + \frac{1}{2}p(\xi)\frac{\partial^2 q}{\partial \tau^2} + |q|^2 q = 0 \qquad [12]$$

where the variable coefficient $p(\xi) = |\beta_2(\xi)/\beta_2(0)|$ takes into account the variation of dispersion along the fiber. Using the transformations $s = \int_0^{\xi} p(y)\mathrm{d}y$ and $U = q/\sqrt{p}$, eqn [12] assumes the form:

$$\mathrm{i}\frac{\partial U}{\partial s} + \frac{1}{2}\frac{\partial^2 U}{\partial \tau^2} + |U|^2 U = \mathrm{i}GU \qquad [13]$$

where

$$G = -\frac{1}{2p}\frac{\mathrm{d}p}{\mathrm{d}s} \qquad [14]$$

Equation [13] shows clearly that the effect of decreasing dispersion is mathematically equivalent to the effect of an optical amplifier, adding a gain term to the NLSE. The effective gain coefficient G is related to the rate at which GVD decreases along the fiber.

Since decreasing dispersion is equivalent to an effective gain, a DDF can be used in place of a conventional fiber amplifier to generate a train of ultrashort pulses. To achieve such an objective, a continuous wave (CW) beam with a weak sinusoidal modulation imposed on it is injected into the DDF. The sinusoidal modulation can be imposed, for example, by beating two optical signals. As a result of the combined effect of GVD, SPM, and decreasing GVD, the nearly CW beam is converted into a high-quality train of ultrashort solitons, whose repetition rate is governed by the frequency of the initial sinusoidal modulation.

Different approaches have been developed to determine the optimum GVD profile and its dependence on the width and peak power of input pulses. In the case of picosecond soliton pulse compression, direct numerical simulations of the NLSE show that linear, Gaussian, and exponential dispersion profiles may all be used effectively to provide ideal, adiabatic compression, where the input pulse energy is conserved and remains localized within the pulse. In this case, the final pulse compression factor at the end of a certain length L of DDF is equal to the ratio of the initial to the final second-order dispersions:

$$F_\mathrm{c} = \frac{T_\mathrm{FWHM}}{T_\mathrm{comp}} = \frac{\beta_2(0)}{\beta_2(L)} \qquad [15]$$

Compression factors larger than 50 are possible launching input pulses with peak powers corresponding to the fundamental soliton into a DDF whose length is about one soliton period. This technique takes advantage of soliton-effect compression but requires lower peak powers and produces compressed pulses of better quality.

In the case of subpicosecond soliton pulses, the influence of higher-order nonlinear and dispersive effects must be accounted for. Some studies show that, in this case, the linear and Gaussian dispersion profiles are the most suitable to achieve high-quality, pedestal-free, adiabatic compression of fundamental solitons. However, in the presence of higher-order effects, the final compression factor is generally lower, and after reaching a maximum at a particular length of DDF, it decreases steadily. This can be explained in terms of pulse compression stabilization, which originates from a competition between the rate of dispersion decrease in the DDF and the rate of dispersion increase due to the combined effects of soliton self-frequency shift and third-order dispersion.

List of Units and Nomenclature

F_c	compression factor
F_{opt}	optimum compression factor
G	amplifier gain
L	fiber length (m)
L_D	dispersion length (m)
L_{NL}	nonlinear length (m)
N	soliton order
P_N	peak power of the Nth-order soliton (W)
P_0	pulse peak power (W)
q	normalized pulse amplitude
T_0	input pulse (s)
T_{comp}	compressed pulse FWHM (s)
T_{FWHM}	pulse full width at half maximum (s)
z_0	soliton period (m)
z_{opt}	optimum fiber length (m)
β_2	group-velocity dispersion coefficient ($s^2\ m^{-1}$)
γ	fiber nonlinearity coefficient ($W^{-1}\ m^{-1}$)
ξ	normalized distance
τ	normalized time
ϕ_g	phase shift induced by the grating pair (rad)
ω_0	pulse center frequency (rad s^{-1})
ω	optical frequency (rad s^{-1})
$\Delta\omega$	pulse spectral width (rad s^{-1})

See also

Chirped Pulse Amplification. **Fiber Gratings**. **Scattering:** Raman Scattering. **Solitons:** Soliton Communication Systems. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

Further Reading

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. San Diego: Academic Press.

Agrawal GP (1997) *Fiber-Optic Communication Systems*, 2nd edn. New York: John Wiley & Sons.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.

Barty C, White W, Sibbett W and Trebino R (eds) (1998) Special issue on ultrafast optics. *IEEE Journal of Selected Topics on Quantum Electronics* 4: 157.

Buck JA (1995) *Fundamentals of Optical Fibers*. New York: John Wiley & Sons.

Kashyap R (1999) *Fiber Bragg Gratings*. San Diego, CA: Academic Press.

Mamyshev PV (1992) Generation and compression of femtosecond solitons in optical fibers. In: Taylor JR (ed.) *Optical Solitons – Theory and Experiment*. Cambridge, UK: Cambridge University Press.

Raman Lasers

M Santagiustina, Istituto Nazionale di Fisica della Materia, Università degli Studi di Padova, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

The Raman effect bears the name of the Indian scientist Chandrasekhara V. Raman (1888–1970) who received the Nobel Prize in the year 1930 for the discovery, published in 1928, of this scattering phenomenon.

The Raman effect is an inelastic light scattering process, occurring in gaseous, liquid, and solid media. The frequency of the scattered photon can be either down-shifted or up-shifted and the result of the process is the generation of new radiation. Frequency down-shifted and up-shifted waves are respectively called Stokes and anti-Stokes waves. The energy difference between input and output photons is either absorbed by, or released from, the atoms or molecules of the medium, which are left in different energy states; jumps can be between electronic states, but more often they are among vibrational or rotational states. The upper energy state involved in the scattering process can be either a true energy state, and this case is referred to as resonant Raman scattering, or a virtual one. The off-resonance effect has a lower probability to occur and high-power beams are needed to observe it. The emission of anti-Stokes waves requires either an inverted Raman medium or particular four-wave-mixing phase-matching conditions to be satisfied in order to occur. The second condition is referred to as coherent anti-Stokes Raman scattering. Here we will consider mainly Stokes wave generation, which is the most common in Raman lasers.

When pumping is very hard, or a seed wave is provided, stimulated Raman scattering can also occur and a high percentage of the input pump radiation can be converted into Stokes, or anti-Stokes, radiation. The stimulated process efficiency, defined as the ratio between the output Raman wave energy and the input pump wave energy, depends on the imaginary part of the third-order nonlinear susceptibility coefficient and on the pump intensity. Stimulated Raman scattering observation dates back to 1962 during the pioneering studies with lasers sources by Woodbury and Ng; their experiment with a ruby laser was actually the first example of an intracavity Raman laser.

Since its discovery, Raman scattering has become a valuable tool for spectroscopy and a practical technique to extend laser light emission to wavelengths not directly generated by laser active materials. This fundamental property stems from the fact that Raman scattering does not require an upper energy state to occur. Hence, the frequency of the emitted radiation can be any, if a suitable incident radiation wavelength is provided. New frequencies are constrained only by the available powerful pump laser sources and by Raman medium shifts.

The frequency shifts provided by some of the materials that are most often used in Raman lasers are presented in Table 1. Finally, it is important to note that Raman scattering can also occur for the newly generated radiation through an identical process in which a Stokes photon is Raman scattered. This radiation, which is generated if the Stokes radiation intensity grows large enough, is referred to as a second-order Stokes wave. For very large pump intensity, a cascading process to several Stokes orders can take place, thus extending the range of generated frequencies to the far-infrared portion of the spectrum.

We will describe the design of Raman lasers that are most commonly used, as well as their performances, also taking into consideration the Raman medium that sets the main constraint to the laser characteristics. Continuous wave as well as pulse Raman lasers will be described and the tuning capability will be discussed.

Raman Shifters

The simplest conversion scheme is obtained when a Raman medium is placed at the output end of a laser. This single-pass setup, shown in Figure 1, provides a fairly efficient frequency shifter.

Single-pass amplification of the spontaneous scattering or of an injected seed copropagating with an intense pump is usually modeled by a couple of equations whose derivation, under a plane, continuous wave approximation, can be found in the

Table 1 Measured Raman frequency shifts and steady-state gain for some of the most important Raman media

Medium	*Frequency shift [THz]*	*Gain [cm/GW]*
Calcite ($CaCO_3$)	32.58(1)	5.5(1)
Lithium iodate ($LiIO_3$)	23.1–24.66(2)	4.8(3)
Barium nitrate ($Ba(NO_3)_2$)	31.46(4)	11(5)
Potassium gadolinium tungstate ($KGd(WO_4)_2$)	27.03 or 23(6)	6(6)
Silica fiber (SiO_2)	13.2(7)	0.01(7)
Hydrogen gas (H_2) ($p = 10$ atm, $T = 293$ K)	124.65(8)	1.53(9)
Nitrobenzene	40.35(10)	3(10)

(1) Pump wavelength 694 nm; from Zverev PG, Basiev TT and Prokhorov AM (1999) Stimulated Raman scattering of laser radiation in Raman crystals. *Optical Materials* 11: 335–352.

(2) Values refer to a propagation direction respectively perpendicular or orthogonal to the crystal axis; from Guadalberto GM and Arguello CA (1974) Raman spectra from oblique phonons in powdered samples. *Solid State Communications* 14: 911–914.

(3) Pump wavelength 1064 nm; from Falk J and Moshrefzadeh (1985) Oblique Raman and polariton scattering in kithium iodate. *IEEE Journal of Quantum Electronics* QE-21: 110–113.

(4) Pump wavelength 1064 nm; from Basiev TT, Sobol AA, Zverev PG, Ivleva LI, and Osiko VV (1999) Raman spectroscopy of crystals for stimulated Raman scattering. *Optical Materials* 11: 307–314.

(5) Pump wavelength 1064 nm; from Basiev TT, Voitsekhovskii VN, Zverev PG, *et al.* (1987) Conversion of tunable radiation from a laser utilizing an LiF crystal containing F_2^- color centers by stimulated Raman scattering in $Ba(NO_3)_2$ and $KGd(WO_4)_2$ crystals. *Soviet Journal of Quantum Electronics* 17: 1560–1561.

(6) Values refer respectively to two linear and orthogonal pump beams at the wavelength 1064 nm; from Zverev PG, Basiev TT and Prokhorov AM (1999) Stimulated Raman scattering of laser radiation in Raman crystals. *Optical Materials* 11: 335–352.

(7) Pump wavelength 1064 nm; from Stolen RH (1978), Fibre Raman lasers. In Ostrowsky D (ed) *Fiber and Integrated Optics*, NATO ASI Series B41. New York, Plenum Press, 157–182.

(8) From Bischel WK and Dyer MJ (1986) Temperature dependence of the Raman linewidth and lineshift for the Q(0)–Q(1) transition in normal and para H_2. *Physical Review* A 33: 3113–3123.

(9) Pump wavelength 1064 nm; from Ottusch JJ and Rockwell DA (1988) Measurements of Raman gain coefficients in hydrogen, deuterium and methane. *IEEE Journal of Quantum Electronics* 24: 2076–2080.

(10) Pump wavelength 694 nm; from Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

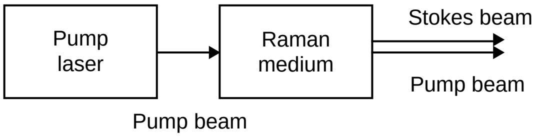

Figure 1 Experimental setup of a single-pass Raman shifter.

literature:

$$\frac{dI_S(z)}{dz} = g_R I_S(z) I_P(z) - \alpha_S I_S(z) \qquad [1]$$

I_S, I_P are the continuous wave intensities of the Stokes and pump waves, z is the propagation distance, g_R the Raman steady-state gain coefficient, and α_S, α_P are the attenuation coefficients at the Stokes and pump frequencies. The Raman gain coefficients, whose general expression as a function of the microscopic material properties can be found in many textbooks, of some of the most commonly used Raman laser media are listed in Table 1.

For short propagation lengths the medium loss can usually be neglected, and if the pump beam is well collimated, the pump intensity can be considered as a constant. For a long interaction length or very absorptive media, the evolution of the pump intensity must also be considered. Though the pump beam is depleted by the scattering process an undepleted pump approximation is often used until high conversion is reached. Under this approximation, the evolution of the pump intensity is mainly governed by diffraction spreading, by losses or by both. In bulk media, single-pass scattered radiation has a coaxial component whose divergence is dictated by the geometry of the pump beam, because Stokes radiation is close to the diffraction limit. Besides this radiation, a nonaxial component can also be observed; this phenomenon stems from the well-known effect of conical emission, which is a phase-matched parametric process. If the Raman medium is within a waveguide, losses are usually the dominating effect. In the undepleted approximation, the evolution of the pump beam intensity as a function of the propagation distance z, either dictated by diffraction or by losses, is known. Then the general solution of eqn [1] is:

$$I_S(z) = I_S(0) \exp\left[g_R \int_0^z I_P(z')dz' - \alpha_S z \right] \qquad [2]$$

where $I_S(0)$ is either the input Stokes signal intensity or an equivalent spontaneous scattering intensity. Then, the amplification gain G of the Stokes beam in the single-pass shifter, for a medium of length L, is given by

$$G = \frac{I_S(L)}{I_S(0)} = \exp\left[g_R \int_0^L I_P(z')dz' - \alpha_S L \right] \qquad [3]$$

Though single-pass Raman shifters are appealing for their simplicity, their application is hampered by the high power necessary for reaching a sufficient efficiency due to the short interaction length that can be obtained, in particular in bulk media, where diffraction dominates. Q-switched pulsed pump lasers are often required, but the propagation in the Raman medium of very powerful pulses requires special precautions to prevent other linear and nonlinear effects occurring. In fact, optical breakdown, thermal and Kerr self-focusing, four-wave mixing, and Brillouin scattering can take place, thus reducing the efficiency of the converter. Precautions may include the optimization of the pump pulse duration, of the focusing, of the interaction length, and the control of the input beam spot size. The multipass configuration has also been exploited to avoid many of the single-pass configuration shortcomings; the typical experimental setup is presented in Figure 2. Note that neither the pump nor the Stokes wave are resonant within the cavity.

The primary advantage of the multipass configuration is the lowering of the pump power required to realize an appreciable conversion. This is due to the increased total gain coefficient, that has been found to be equal to

$$G_n = G\frac{1-R^n}{1-R} + (n-1)\log R \qquad [4]$$

where G is the single-pass gain of eqn [3], R the reflectivity of the mirrors at the Stokes wavelength, and n the number of passes through the Raman medium. There are two main drawbacks of multipass Raman shifters that finally limit the conversion efficiency of this setup. The first is the simultaneous generation of many undesired and uncontrolled Stokes orders, because the first-order Stokes intensity grows rapidly during propagation. The second is the

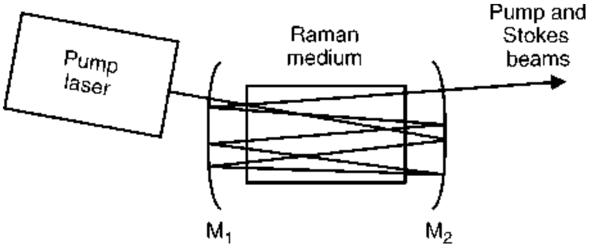

Figure 2 Experimental setup of a multipass Raman shifter. Mirrors M_1 and M_2 reflect both Stokes and pump waves.

intensity decrease associated with the dispersion of pulses, in particular if the Raman medium is a solid. These problems have limited the applications of multipass shifters to hydrogen gas.

Raman Lasers

By placing the Raman medium into an optical cavity, Stokes radiation can be selectively enhanced and Raman oscillation takes place. The two main experimental setups, either with the Raman medium placed in an external cavity or within the pump laser cavity (intracavity), are represented in Figures 3 and 4. In some cases, for the intracavity configuration, the Raman medium may also be the host of the doping laser active material.

For the external resonator Raman laser, the reflectivity of the mirrors is such that they are almost transparent for the pump but highly reflective (resonant) for the shifted radiation. The threshold condition of Raman lasers can be obtained by requiring that the net round trip gain equals the total cavity losses. The net round trip gain, under the plane wave, steady-state, undepleted pump approximation, is given by the single-pass gain G, multiplied by a factor of two to account for the forward and backward amplification. In fact, Raman scattering takes place also between the pump and the counter-propagating Stokes wave reflected by mirror M_2. Losses are given by the medium contribution, which is already accounted for in eqns [2] and [3], and by the energy leakage from the mirrors. Assuming that the mean reflectivities of the mirrors, at the Stokes wavelength, are respectively R_1, R_2, then the threshold condition is $GR_1R_2 = 1$. Through eqn [3] we obtain

$$g_R \int_0^L I_P(z')dz' - \alpha_S L = -\log(R_1R_2) \qquad [5]$$

Equation [5] can be used to evaluate the pump intensity because at threshold low depletion occurs and the approximation taken to obtain [5] is valid. Raman steady-state gain is not very different for gases, liquids and crystals, as shown in Table 1 and oscillation thresholds depend more on the interaction length and the focusing properties of the pump and Stokes beams. Nonetheless, the type of Raman medium imposes many practical constraints for laser design, which are addressed in the next paragraphs.

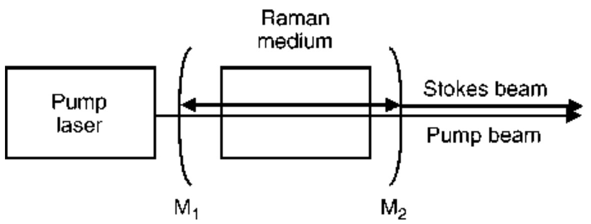

Figure 3 Experimental setup of an externally pumped Raman oscillator. Mirrors M_1 and M_2 are highly reflective for the Stokes wave but not for the pump wave.

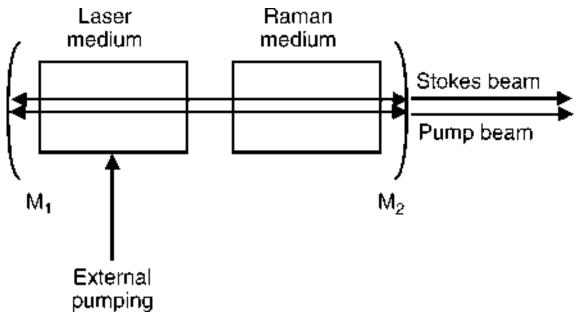

Figure 4 Experimental setup of an intracavity Raman oscillator. Mirrors M_1 and M_2 are highly reflective for both the Stokes and the pump wave.

With gas media it is hard to maintain a high pump intensity over the entire length of the cell, which must be used to contain the gas. Usually gas cells can be from 5 to 30 cm long; this means that Q-switched pump lasers are required to set intensities above 10^4 GW m^{-2}, that is the typical threshold value of Raman oscillation. Continuous wave oscillation of gas Raman lasers is, nonetheless, possible when the pump wavelength is near a medium resonance. In fact, the imaginary part of the third-order non-linear polarization coefficient, to which the Raman gain is proportional, is enhanced near energy transitions because of the resonance in correspondence of that photon energy. The main limitations of this setup are that pump wavelengths are dictated by the Raman medium transitions, and thus also Stokes wavelengths are fixed, and finally conversion efficiency is usually low (about 5% is reported) possibly because of the high absorption. Off-resonance, continuous wave operation can also be attained by using high finesse cavities in a doubly resonant cavity. In fact, by resonating both Stokes and pump wave with high reflectivity mirrors ($R \geq 0.999$), the threshold intensity is lowered and conversion efficiency is high (about 35% has been demonstrated). In spite of the relatively high Raman gain and the large Stokes shifts obtainable, the exploitation of gas cell Raman oscillators is mainly hampered by the need of a high-pressure-resistant vessel and a constant gas flow. The latter is needed because of the low thermal conductivity of gases that might cause local temperature variations and thermal-optical distortions that diminish the output beam uniformity and stability. Moreover the repetition rate of pump pulses must be small (below 50 Hz) so as not to impose a too strong thermal load on the gas.

Liquid media suffer basically the same problems as gases and, moreover, they show higher non-linear refractive indices that result in strong self-focusing. This lens effect results in filamentation, optical breakdown, and high beam divergence. Often Raman liquids are also toxic and very volatile; their use has been slowly but steadily decreasing.

Solid media, in particular crystals, do not show many of the problems listed above and present the advantages of solid-state technology such as compactness, low maintenance, and opportunity of integration. The high density and the restricted atom motion result in high Raman steady-state gains and minimal line broadening. Precise resonance enhancement, as well as low threshold intensities, can be obtained by choosing an adequate mirror reflectivity spectral response. Pump intensities as low as 10^3 GW m^{-2} were reported for the oscillation of an external cavity Raman laser whose medium was a 5-cm-long, barium nitrate (see Table 1) crystal, with mirror reflectivity $R = 0.49$ at the Stokes wavelength. Though the intensity threshold is about an order of magnitude smaller than for gas Raman lasers, Q-switched input pump pulses are still required. Solid-state Raman lasers can be easily forced to oscillate on a $\mathrm{TEM}_{0,0}$ mode with divergence close to the diffraction limit even with multimode pumps. Moreover the Stokes beam is subject to a very complex clean-up process; this can be basically explained by the fact that the Stokes beam, growing from the spontaneous emission, does not retain the pump spatial characteristics. Pulse duration is equal to or slightly less than the input one. Conversion efficiency, which depends strongly on mirror reflectivity, can reach 60%, where saturation occurs due to higher-order Stokes wave generation. Solid-state Raman lasers can be easily forced to oscillate also at these higher-order Stokes frequencies if mirror reflectivity is large enough at the relative wavelengths. Second- and third-order Stokes conversion efficiencies of 30% and 20% have been reported.

The theoretical determination of the cavity pump and Stokes fields does not have an analytical solution. For continuous wave or long Q-switched pulsed lasers, eqns [1] must be extended to include all forward and backward Stokes and pump waves and boundary conditions due to the mirrors must be imposed. This problem can only be solved numerically.

The experimental setup of intracavity Raman lasers is shown in Figure 4; both the Raman and the laser medium are placed within the cavity and sometimes the two media coincide. This is, for example, the case of potassium gadolinium tungstate which can be doped with neodymium ions (see Table 1). The cavity is usually resonant for both the pump and the Stokes waves; very high conversion efficiency (up to 80%) is attained when special dichroic coatings are provided such that pump output is suppressed (high-finesse cavity for the pump). Besides the nearly diffraction-limited property of output radiation and clean output beams that are common characteristics of all Raman lasers, enhanced brightness is also observed in this setup. This effect is probably due to the high thermal loading, the consequent linear and nonlinear lens effects, and also to a short pulse generation mechanism that increases output beam intensity. This regime can be reached by reducing the Stokes cavity length with respect to the pump cavity length. A first setup is to use an additional dichroic intracavity mirror that is transparent for the pump but highly reflective for the Stokes; an alternative setup exploits a double ring cavity configuration as depicted in Figure 5.

The principle governing the compression can be explained as follows. The optical paths of the pump and Stokes pulses are made different by an amount ΔL. Then the Stokes pulse generated within the Raman medium, once re-injected into the cavity from mirror M_{S1}, will overlap with an undepleted portion of the recirculating pump pulse. Hence, another Stokes pulse is generated at each round trip, until the pump pulse is completely depleted; the Stokes output finally consists of a train of pulses of duration $\Delta L/c$, at a repetition rate equal to the cavity round-trip. This mechanism leads to pulses as short as the limit imposed by the Raman gain bandwidth. In fact, when the pulse duration becomes comparable to the inverse of the gain bandwidth, transient effects take place because of the delay in the nonlinear medium response. Ring Raman lasers have been obtained with

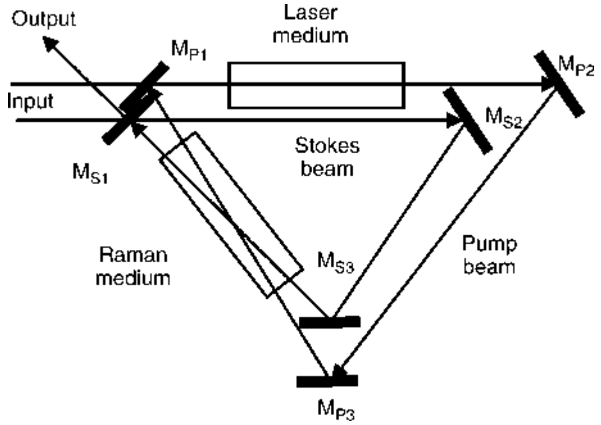

Figure 5 Experimental setup of an intracavity, double ring, Raman oscillator. Mirrors M_{S2}, M_{S3}, M_{P2}, and M_{P3} are highly reflective respectively for the Stokes and the pump wave. M_{S1}, M_{P1} are partially reflective respectively for the Stokes and pump wave.

gases (e.g., hydrogen) and crystal media (barium nitrate and lithium iodate).

Another commonly used experimental setup is that of synchronously pumped ring Raman lasers. In such devices, whose typical setup is sketched in Figure 6, a new input pump pulse is synchronized with the output Stokes pulse thus enhancing the conversion within the Raman medium.

Though electronic trigger detection is needed to lock the pump pulse time repetition rate to a multiple or submultiple of the cavity round-trip, synchronously pumped Raman lasers are relatively simple, stable, high repetition rate sources of radiation. Very compact and reliable synchronously pumped Raman lasers can be realized with optical fibers; in this case the feedback of the ring resonator can be realized by means of a fiber directional coupler. The device, besides the high integration achieved by the all-fiber technology, can also be continuously tuned over a large bandwidth.

Continuous tuning of the same laser source is a very desirable property; the Raman lasers so far described can generate radiation at wavelengths restricted to a few values. In fact, output spectral lines can be changed either by replacing the Raman medium, or the pump laser, or by exploiting cascade Raman scattering or finally resonating a different vibrational line of the same material (if there are many as for example in hydrogen gas); all these cases require substantial cavity changes. Frequency switching without setup changes is attained only in case of potassium gadolinium tungstate. For this medium there exist two independent directions, orthogonal to the crystal optical axis, for which the Raman shifts are different if the pump polarization changes. Then, the wavelength selection is just a matter of controlling the input pump polarization. Though Raman lasers can greatly expand the wavelength range of available laser sources, controlled, continuous tuning of the same Raman laser over a certain frequency band can be obtained only in a few particular cases.

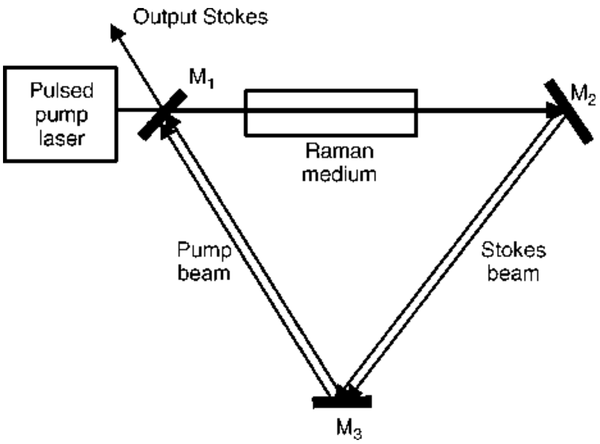

Figure 6 Experimental setup of a synchronously pumped, ring, Raman oscillator. Mirrors M_2 and M_3 are highly reflective for both the Stokes and the pump wave. M_1 is partially reflective for both the Stokes and pump wave.

Wavelength tuning can be attained by selecting the direction of propagation in polar crystals (like lithium niobate and lithium iodate). Different directions allow for wavevector matching of polariton scattering at different frequencies. Also phase matching of parametric interactions, like four-wave mixing, is a possible technique to attain wavelength tuning. However, such Raman lasers cannot operate at maximum gain, and this fact decreases the output energy, the conversion efficiency, and the stability, and increases the pump intensity threshold. Continuous tuning with an external control parameter is attained in the spin-flip Raman semiconductor laser. Semiconductors are Raman media that can be exploited to build far-infrared or even submillimeter Raman lasers, with pump intensity thresholds that are comparable with those of gas and liquid media. Though frequency shifts are fixed, under normal conditions, tuneable operation can be achieved if the semiconductor is placed in a magnetic field. Each energy level of an electron in the solid is, in fact, split into two, depending on the orientation of the electron spin with respect to the magnetic field. When there is a transition involving Raman scattering, the electron spin can change its direction. Then the wavelength of the laser can be continuously varied by changing the magnetic field applied. Tuning over a range of several micrometers in the infrared has been attained and spin-flip Raman lasers have been developed with semiconductors like InSb, $Hg_{0.77}Cd_{0.23}Te$, and $Pb_{0.88}Sn_{0.12}Te$. Use of this technique is severely limited by the need for magnetic field sources.

The simplest method to achieve tuneable radiation is through the use of a material that shows a broad Raman linewidth. Inhomogeneous broadening of the energy level is a well-known effect that takes place in noncrystalline solid media, such as glasses. Though line broadening is accompanied by a decrease in the steady-state Raman gain of about two orders of magnitude with respect to a crystalline medium, this disadvantage can be over-compensated by the increase in the Raman medium length that can be of 3 or 4 orders of magnitude. In fact, glasses are commonly exploited to make optical waveguides, in particular optical fibers, that present very low optical absorption coefficients, in the near-infrared part of the spectrum (1200–1600 nm).

Optical fiber Raman lasers based on silica telecommunication fibers are very versatile, compact, and simple to implement. In principle the setup is not very different from the corresponding external cavity or intracavity Raman lasers so far presented; however, fiber lasers show the great advantage of an integrated, all-fiber technology.

Cavity mirrors can be replaced by fiber Bragg grating reflectors (Figure 7) that can be constructed with a very well-defined spectral reflectivity, thus allowing fine enhancement or suppression of the desired wavelengths. By splicing the Raman fiber, which can be several hundred meters long, with the gratings, the net losses are very low, finally decreasing the pump intensity threshold. The pump threshold for first-order Stokes generation can be easily as low as 200 mW; this means that diode lasers can be used for pumping. Moreover, the need for cavity alignment is eliminated, the laser stability increased and the construction simplified. This simplification allows us to build, much more easily than by using bulk optics, nested multiple resonant cavities, to generate higher-order Stokes beams (Figure 8).

Continuous-wave lasing of up to six Stokes orders has been reported for a fiber Raman laser of this type, with a total injected power of about 7 W. Another key advantage of Bragg gratings is that their reflectivity can be adjusted by means of a piezoelectric stretching of the fiber support; this allows one to control the output intensity of all waves. As previously mentioned, ring cavities can also be constructed by exploiting an all-fiber technology, if feedback is provided by a fiber directional coupler (Figure 9).

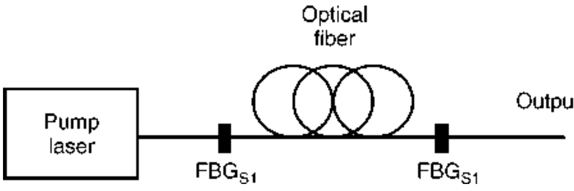

Figure 7 Experimental setup of a continuous-wave, all-fiber, Raman oscillator. FBG$_{S1}$ are fiber Bragg gratings highly reflective for both the first Stokes wave.

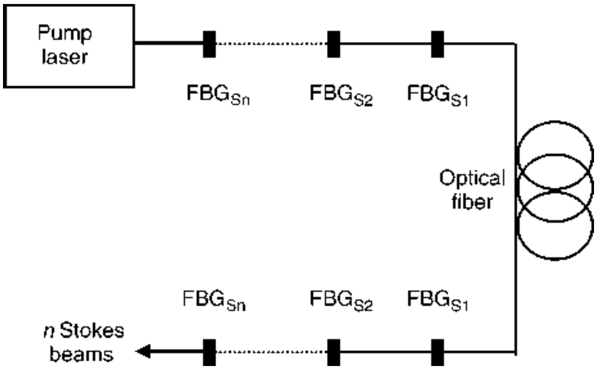

Figure 8 Experimental setup of a continuous-wave, all-fiber, *n*th-order Stokes, Raman oscillator. FBG$_{Si}$ are fiber Bragg gratings highly reflective for the *i*th Stokes wave.

Pulsed operation of fiber Raman lasers has been obtained too; due to the long interaction length, Q-switched pump lasers are no longer needed but rather mode-locked sources can be used. This means that trains of very short laser pulses, down to the hundreds of femto-second regime, can be obtained from fiber Raman lasers, the limit being represented by the Raman bandwidth, which is as large as about 7 THz, at the half maximum, in silica fibers. Propagation of short and intense laser pulses in an optical fiber may induce several other nonlinear effects, like self-, cross-phase modulation, and four-wave mixing. Besides nonlinearity, linear dispersion and velocity mismatch become important too, due to the long propagation. All these linear and non-linear effects limit the Stokes pulse quality. In more detail: self- and cross-phase modulations lead to the spectral broadening of both pump and Stokes waves; broadening is asymmetric, because of the contemporary action of group velocity mismatch. In fact during the propagation the Raman pulse walks off the pump pulse, because of velocity mismatch, and this induces an asymmetry in gain depletion and in nonlinear index modulation. The final outcome is that Stokes waves can be asymmetric both in time and in frequency, and also highly chirped. Four-wave mixing is usually negligible, if phase-matching conditions are not imposed and Brillouin scattering is not an issue for short pulses. Very symmetric and low chirped pulses are attained as soon as the fiber Raman laser is working in the soliton regime, that is when compensation between the dispersion and the nonlinear phase shift occurs. Synchronously pumped, fiber Raman, soliton lasers are able to provide 200-fs pulses, at various wavelengths and with very high conversion efficiency (50%). Intracavity fiber Raman lasers have

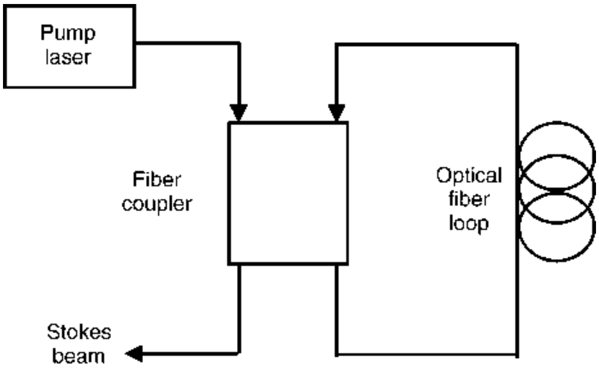

Figure 9 Experimental setup of an all-fiber, ring Raman oscillator.

also been realized by doping the fiber with neodymium, erbium, or ytterbium ions.

The extremely broad Raman spectral linewidth of the glass allows for broadband tuning of the emitted wavelength. For continuous and pulsed wave operation the change in the peak Stokes wavelength can be attained through an intracavity frequency selective element, like for example a prism or a grating. In this case tuning is not obtained in an all-fiber configuration but rather with bulk optical elements and thus additional lenses or objective are needed to collimate the Stokes wave at the fiber ends. The experimental setup is shown in Figure 10.

The laser can be continuously tuned, over about 6 THz, by adjusting the prism or grating rotation. All-fiber Raman laser tuning is possible in a double ring configuration with an inserted optical filter, whose experimental setup is shown in Figure 11. Stokes and pump waves are counterpropagating in the Raman fiber; the Stokes wave frequency is selected by the tuneable optical filter.

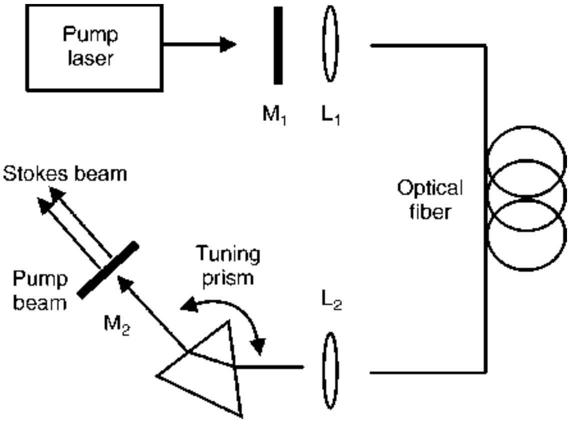

Figure 10 Experimental setup of a tuneable, fiber Raman oscillator. Mirrors M_1 and M_2 are highly reflective for the Stokes wave but not for the pump wave. L_1 and L_2 are collimating lenses or objectives. The tuning prism is rotated to control wavelength selection.

The devices of Figures 10 and 11 stop emitting Stokes radiation for frequencies at which the Raman gain has decreased to a level such that laser threshold is higher than the pump intensity. Due to the fact that Raman gain in fibers is almost constant over about 3 THz, tuning within this frequency range does not lead to any sensible decrease in conversion efficiency. For synchronously pumped ring fiber Raman lasers another technique, named time dispersion tuning, can be used. The experimental setup is shown in Figure 12.

The output fiber end is placed on a moveable mount, that controls the relative delay between the Stokes pulse and the new pump pulse at the fiber input. Since Stokes pulses are dispersed during fiber propagation, to each time slot within a pulse there corresponds a certain portion of the Stokes spectrum. By properly delaying the time of injection of the Stokes pulse back into the fiber the most powerful part of the pump pulse is overlapped to a different part of the Stokes pulse, and this causes a preferred amplification of certain frequencies with respect to others. The shift of the Stokes peak wavelength $\Delta\lambda$ is given by:

$$\Delta\lambda = \frac{\Delta l}{cLD(\lambda)} \qquad [6]$$

where Δl is the translation step, c the speed of light, L the fiber length, and $D(\lambda)$ the fiber dispersion at the Stokes wavelength λ. A total tuning range of 10 nm can be obtained with this technique that can also be used as a test measurement of fiber dispersion.

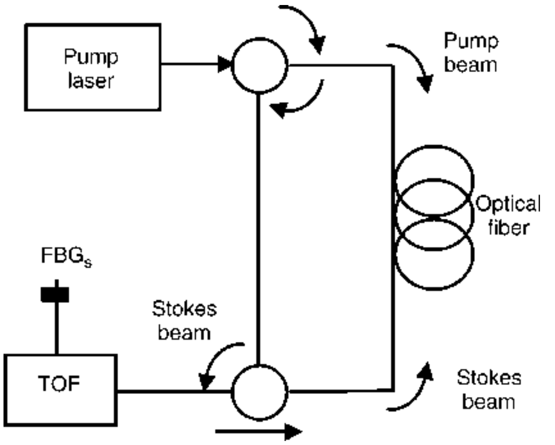

Figure 11 Experimental setup of a double-ring fiber Raman oscillator. OC_1 and OC_2 are fiber optic circulators, and TOF is a tuneable optical filter.

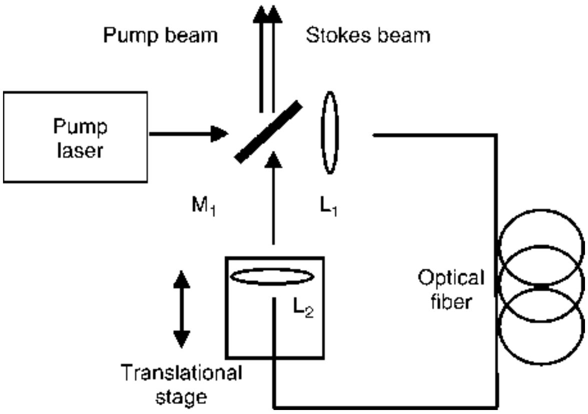

Figure 12 Experimental setup of a time dispersion tuned, ring fiber Raman oscillator. L_1 and L_2 are lenses or objectives. The output collimating lens L_2 and the fiber end are mounted on a translation stage to control wavelength selection by time delaying the re-entrant Stokes pulse.

Applications

Raman lasers have found a large variety of applications. Gas and liquid Raman lasers are fundamental tools for spectroscopic investigation. Solid-state Raman lasers have been shown to be very powerful, eye-safe sources for light detection and ranging (LIDAR) and medical applications. Fiber Raman lasers are exploited for long-haul fiber optics transmission. They can be either sources of short pulses and solitons or high-power continuous-wave sources at multiple wavelengths for subsequent link-distributed Raman broadband amplification.

List of Units and Nomenclature

Amplification gain	[dimensionless]
Efficiency	[dimensionless]
Energy	[J]
Frequency	[Hz]
Intensity	[W m^{-2}]
Length	[m]
Loss coefficient	[m^{-1}]
Power	[W]
Raman bandwidth	[Hz]
Raman gain	[m W^{-1}]
Reflectivity	[dimensionless]
Wavelength	[m]

See also

Fiber and Guided Wave Optics: Dispersion; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Overview. **Fiber Gratings**. **Optical Amplifiers:** Optical Amplifiers Long-Haul Transmission Systems. **Imaging:** Lidar. **Laser-Induced Damage of Optical Materials**. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Magneto-Optics:** Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics; Phase Matching; Self-Focusing and Related Effects (Solitons and Multiphoton Absorption). **Optical Materials:** Optical Glasses. **Solitons:** Temporal Solitons. **Scattering:** Raman Scattering; Stimulated Scattering. **Spectroscopy:** Nonlinear Laser Spectroscopy; Raman Spectroscopy.

Further Reading

Band YB, Ackerhalt JR, Krasinski JS and Heller DF (1989) Intra-cavity Raman lasers. *IEEE Journal of Quantum Electronics* QE-25: 208–213.

Basiev TT and Powell RC (eds) (1999) Special issue on solid-state Raman lasers. *Optical Materials* 11: 301–390.

Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

Eckhardt G (1966) Selection of Raman laser materials. *IEEE Journal of Quantum Electronics* QE-2: 1–8.

Grasiuk AZ and Zubarev IG (1978) High power tunable IR Raman lasers. *Applied Physics* 17: 211–232.

Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman spectroscopy. In: Arecchi FT and Schulz-Dubois D (eds) *Laser Handbook*. Amsterdam: North Holland.

Koechner W (1999) *Solid-state Laser Engineering*. Berlin: Springer-Verlag.

Pask HM (2003) The design and operation of solid-state Raman lasers. *Progress in Quantum Electronics* 27: 3–56.

Stolen RH (1978) Fibre Raman lasers. In: Ostrowsky D (ed.) *Fiber and Integrated Optics*, NATO ASI Series B41, pp. 157–182. New York: Plenum Press.

Suto K and Nishizawa J (1994) *Semiconductor Raman Lasers*. Norwood, MA: Artech House.

Weber A (1979) *Raman Spectroscopy of Gases and Liquids*. Topics in Current Physics, vol. 11, Berlin: Springer.

Self-Focusing and Related Effects (Solitons and Multiphoton Absorption)

R L Sutherland, Science Applications International Corporation, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

When an intense beam of light with axial symmetry propagates in a medium possessing a nonlinear index of refraction, it will induce its own lens in the medium and be focused. This phenomenon, first discovered just a few years after the invention of the laser, is called self-focusing. It is one of many phenomena under the general classification of self-action, which includes such effects as self-trapping, self-bending, self-phase modulation, and self-steepening. This general class of phenomena involves action on the beam of light induced by the beam itself: light causes a change in an optical property of the medium, which in turn produces an effect back on the light beam. We will discuss the mechanisms responsible for self-focusing in nonlinear optical media and the resultant action on the optical beam. In this we will see how the topics of self-trapping and self-phase modulation are intimately connected to self-focusing. This will lead to a discussion of related effects as well as many potential applications.

It is instructive to first review the properties of a conventional thin lens, illustrated in Figure 1. By 'thin', we mean that just before and just after the lens, the beam size and shape are approximately the same. However, the phase of the optical wave is changed on passage through the lens. We apply the paraxial (aberrationless) approximation where only rays that lie near the lens axis are considered. In this treatment the spherical surfaces of the lens are approximated as parabolic surfaces. For a symmetric double-convex lens (as shown in Figure 1), this phase change is

$$\Delta\varphi(r) = -k(n-1)\frac{r^2}{R} \equiv -\frac{kr^2}{2f}, \quad f^{-1} = (n-1)\frac{2}{R} \qquad [1]$$

where n is the index of refraction of the lens material, and R is the radius of curvature of the lens surfaces. The light beam has a wavevector $k = \omega/c$, where ω is the frequency and c the speed of light in a vacuum. The phase retardation is a maximum on the lens axis (z) and decreases quadratically with the radial coordinate r. In other words, the phase advances quadratically with distance from the axis and forms a spherical, converging wave that comes to focus on the z-axis a distance f (the focal length) from the lens.

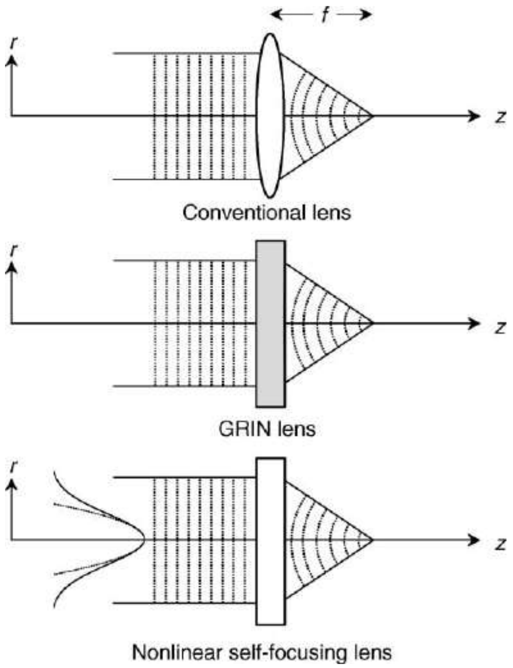

Figure 1 Focusing with a conventional lens, a graded-index (GRIN) lens, and a nonlinear self-focusing lens.

Next consider a graded-index (GRIN) lens, also illustrated in Figure 1. This is a flat slab of material with thickness L and an inhomogeneous index of refraction given by $n(r) = n_0(1 - 1/2\alpha^2 r^2)$. The index has a maximum (n_0) on the lens axis and decreases quadratically with radial distance from the axis. The constant α is a measure of the rate of the index change with r. Upon passage of light through a GRIN lens, the phase change is

$$\Delta\varphi(r) = -kn_0\alpha^2 L\frac{r^2}{2} \equiv -\frac{kr^2}{2f}, \quad f^{-1} = n_0\alpha^2 L \qquad [2]$$

Because of the parabolic nature of the index distribution, the phase retardation again decreases quadratically with r and thus forms a converging spherical wave. Notice that for both the conventional and GRIN lenses, the optical thickness of the lens is parabolic. In the conventional lens the index is constant, but the physical thickness varies continuously. In the GRIN lens, the physical thickness remains constant while the index varies continuously. Both produce the same kind of focusing action.

Finally, consider a flat slab of homogeneous material that has a nonlinear index of refraction. Light of intensity I induces a change in the refractive index given by $\Delta n = n_2 I$. The parameter n_2 is called the nonlinear refractive index coefficient and has physical units of cm^2 W^{-1}. There are other definitions of n_2 given in the literature where the index change is given in terms of the square of the electric field, which would obviously have different physical units. We will adopt the intensity form of n_2 in this discussion. We now assume the light incident on this slab is a plane wave with a Gaussian intensity distribution given by $I(r) = I_0 \exp(-2r^2/w^2)$ as shown in Figure 1. I_0 is the peak intensity along the z-axis, and w is the radius of the beam at $1/e^2$ of the peak intensity. Consistent with the paraxial approximation, we expand the exponential function, keeping only the first two terms, to yield $I(r) \approx I_0(1 - 2r^2/w^2)$. This parabolic approximation to the intensity is sketched as the dotted curve in Figure 1. The index distribution induced in the medium by the intensity approximately follows this parabolic shape, and the resultant phase change is

$$\Delta\varphi(r) = -k4n_2I_0\frac{L}{w^2}\frac{r^2}{2} \equiv -\frac{kr^2}{2f}, \qquad [3]$$
$$f^{-1} = \frac{4n_2I_0L}{w^2}$$

Note the similarity to the GRIN lens, with the association of the peak index $n_0 \rightarrow n_2I_0$, and $\alpha^2 \rightarrow 4/w^2$. We can say that the nonlinear medium acts like a GRIN lens with an intensity-dependent focal length. However, this correspondence is only accurate within the parabolic approximation, and the nonlinear lens can introduce significant aberrations to the focal properties of the full beam.

In the paraxial approximation, the conventional, GRIN, and nonlinear lenses all have the same effect on an incident beam of light. Although we have considered only plane parallel incident waves, it should be obvious that the GRIN and nonlinear lenses will produce analogous effects on converging and diverging wavefronts as the conventional lens. The effect of a stack of nonlinear lenses in an optical train can also be analyzed in the same way as a stack of conventional lenses, with the exception that the focal length of each lens will depend on the local intensity and hence size of the beam. Finally, we note that if we change the sign of R, α^2, or n_2, we obtain a negative focal length. This is a diverging lens, and in the nonlinear optics realm the effect is called self-defocusing.

Nonlinear Index of Refraction

The concept of an intensity-dependent refractive index has its roots in the power series expression of the nonlinear dielectric polarization P in terms of the electric field E of the optical wave, which in its simplest form is

$$P = \varepsilon_0\left(\chi^{(1)}E + D^{(2)}\chi^{(2)}E^2 + D^{(3)}\chi^{(3)}E^3 + \cdots\right)$$
$$= \varepsilon_0\left(\chi^{(1)} + D^{(2)}\chi^{(2)}E + D^{(3)}\chi^{(3)}E^2 + \cdots\right)E \qquad [4]$$

where $\chi^{(m)}$ $(m = 1, 2, 3\ldots)$ is called the m-th order susceptibility. The quantity $D^{(m)}$ is called the degeneracy factor and is 3 for $m = 3$ in the self-focusing case. This leads directly to the definition of a nonlinear susceptibility $\chi(E)$. All materials have first- and third-order susceptibilities. Centrosymmetric materials (e.g., gases, liquids, cubic crystals) do not possess a second-order susceptibility. Here, we will consider only centrosymmetric media and briefly discuss second-order materials later.

The refractive index is $n = [1 + \chi(E)]^{1/2}$. Consequently, the linear index (small electric field) is $n_0 = (1 + \chi^{(1)})^{1/2}$. The leading nonlinear term in $\chi(E)$ is generally small. Expanding the square root in a binomial series, the first nonlinear term in the refractive index can be approximated by $3\chi^{(3)}E^2/2n_0$. Since $I = 2\varepsilon_0 n_0 cE^2$, we see that

$$n = n_0 + n_2I \qquad [5]$$

where $n_2 = 3\chi^{(3)}/4\varepsilon_0 n_0^2 c$. Third-order (cubic) materials that exhibit an intensity-dependent refractive index described by eqn [5] are called Kerr or Kerr-like.

Several physical mechanisms produce a refractive index of the form given in eqn [5]. A list of the significant ones is presented in Table 1. We emphasize that these mechanisms lead to eqn [5] over a limited range of intensity. For example, in materials where the main mechanism is electronic state population redistribution, when the incident intensity becomes comparable with the saturation intensity, the index begins to saturate (i.e., the change of index with

Table 1 Physical mechanisms responsible for the nonlinear index of refraction

Mechanism	*Typical n_2 ($cm^2\ W^{-1}$)*	*Typical sign*	*Typical response time (seconds)*
Electronic polarization	$\sim10^{-16}$–10^{-15}	+	$\sim10^{-15}$
Molecular orientation			
Isotropic liquids	$\sim10^{-15}$–10^{-14}	+	$\sim10^{-12}$
Liquid crystals	$\sim10^{-4}$–10^{-3}	+	$\sim10^{-3}$–10^{-2}
Electrostriction	$\sim10^{-13}$–10^{-12}	+	$\sim10^{-9}$
Population redistribution			
Atomic vapors (ω below resonance)	$\sim10^{-10}$–10^{-9}	−	$\sim10^{-8}$
Atomic vapors (ω above resonance)	$\sim10^{-10}$–10^{-9}	+	$\sim10^{-8}$
Semiconductors	$\sim10^{-6}$	+	$\sim10^{-8}$
Band filling (semiconductors)	$\sim10^{-6}$–10^{-4}	±	$\sim10^{-8}$–10^{-7}
Exciton saturation (quantum wells)	$\sim10^{-4}$	−	$\sim10^{-8}$
Thermal	$\sim10^{-8}$	±	$\sim10^{-4}$

intensity becomes sublinear). The simple expression in eqn [5] no longer applies. These types of effects can be important and will be discussed later. Finally, we mention that other mechanisms exist that produce self-focusing related effects, but they are not Kerr-like and will be briefly discussed later.

We learn from Table 1 that values of n_2 range over several orders of magnitude. High-intensity lasers are required to observe the effects of small optical nonlinearities. Hence pulsed lasers are often used in self-focusing applications. Laser pulse widths can generally range from 100 s of nanoseconds to 10 s of femtoseconds. Consequently, one must keep in mind the response time of the mechanism/material when analyzing experiments. If the pulse width is long compared to the response time, a quasi-steady-state self-focusing regime results. If the pulse is too short, transient effects will appear. Also, as one goes to longer response time mechanisms, lower intensities can be used. In some cases it is even possible to use cw lasers.

Self-Lensing

Thin nonlinear media can be employed in self-lensing applications, where self-focusing is used to modify the far field properties of a beam, just as a conventional lens would do, but with effects dependent on the beam intensity. Gaussian beams are usually employed in the paraxial approximation.

A popular technique employing this effect to measure the magnitude and sign of n_2 is the Z-scan. A typical experiment is illustrated in Figure 2. A thin nonlinear medium is scanned axially between a focusing lens and an aperture in the far field. Consider a material with positive n_2. When the sample is close to the lens the intensity is too low to induce significant self-lensing. As the sample approaches the focal point of the lens, the intensity rises and a positive self-lens is induced that focuses the beam more strongly. This displaces the focus to the left, causing the beam to spread out more in the far field and less power to pass through the aperture. As the sample passes to the right of the focal point, the positive self-lens reduces the divergence of the beam so that more power is passed through the aperture. Near the focal point the self-lens is strongest, but a thin lens at this point has no effect on the beam passing through the far-field aperture. The measured aperture transmittance as a function of scan-position Z is sketched in Figure 2. The peak and valley of this dispersion-shaped curve will be interchanged for a negative n_2. For a small aperture and fixed power of the laser, the difference between the peak and valley transmittance is proportional $|n_2|$.

The same experimental configuration is also popular for measuring small absorption coefficients or thermo-optic coefficients ($\mathrm{d}n/\mathrm{d}T < 0$ for most materials). Under steady-state conditions (cw beam), the thermal n_2 is given by

$$n_2 = \frac{\alpha\tau}{\rho C}\frac{\mathrm{d}n}{\mathrm{d}T} \qquad [6]$$

where α is the absorption coefficient, τ is the thermal diffusion time, ρ is the mass density, and C is the specific heat. This technique is extremely sensitive and can easily be observed, for example, with a ~1 mW laser pointer and a 1 cm cuvette filled with minute quantities of ink or food coloring dissolved in methanol.

This combination of self-lensing and a hard aperture has also been put to use in devices. One such device is the optical power limiter. A typical configuration is similar to the Z-scan experiment in Figure 2, with the sample fixed near the lens focal point. Thick media in the focal region have also been used. At sufficiently low optical power, the medium acts like a flat window. As the input power rises, self-lensing sets in and spreads the beam out at the aperture, limiting the transmitted power. The higher the incident power gets, the more the beam spreads. The result is a transmitted power that is clamped below a preferred level. Thermal lensing has been used for cw beams, while positive Kerr

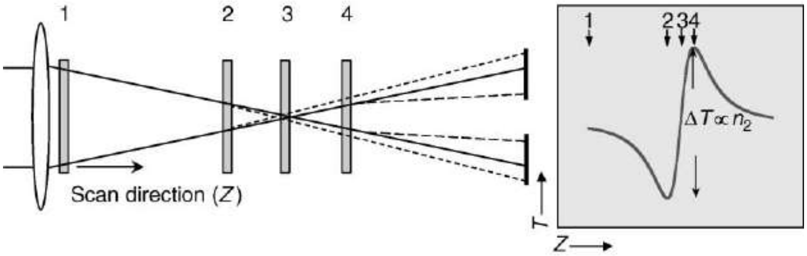

Figure 2 A Z-scan experiment. The transmittance through the aperture as a function of the position (Z) of the thin nonlinear medium is sketched at the right.

self-lensing has been employed for Q-switched lasers.

A less obvious but very important application of this configuration is for mode-locking lasers (see Figure 3). Mode-locking is a coherent phenomenon that locks the relative phase of multiple longitudinal modes in a laser having a broadband gain profile. These stabilized modes interfere to form a train of short pulses with pulsewidth inversely proportional to the gain bandwidth. A common technique for passively mode-locking a laser is to insert a saturable absorber into the laser cavity. A saturable absorber blocks low-power light, keeping the round-trip gain below the oscillation threshold, but transmits high-power light as the absorption coefficient saturates. Thus the mode-locked condition is favored for gain and oscillation since the peak power of the pulses is high. It turns out that transverse spatial modulation, due to self-focusing combined with an aperture, can produce the same type of effect in a laser cavity. The laser rod acts as both the gain medium and nonlinear lens. For example, titanium: sapphire ($Ti:Al_2O_3$) has $n_2 = 3 \times 10^{-16}\ \mathrm{cm^2\ W^{-1}}$. Thus, in a $Ti:Al_2O_3$ laser the rod produces self-lensing that modifies the beam shape propagating through an intracavity aperture. The loss due to the aperture is intensity dependent, as in the Z-scan. Therefore, the net round-trip gain of the cavity is low for low-power but high for high-power pulses, and the mode-locking condition is fulfilled. These types of lasers are called Kerr lens mode-locked (KLM) lasers. KLM lasers are generally not self-starting, but can be self-sustaining. They, therefore, require some active means of starting the mode-locking. Ultrashort pulses (<100 fs) have been obtained with a cw KLM $Ti:Al_2O_3$ laser. For pulsewidths this short, the simple quasi-steady-state self-focusing picture is not entirely adequate to describe the dynamics. Other factors, such as group-velocity dispersion (GVD) and self-phase modulation (SPM), must be considered. Moreover, since the laser rod is typically several millimeters long, it is more appropriate to model it as a thick rather than a thin nonlinear lens.

Beam Trapping

When an intense beam propagates in a thick nonlinear medium, a question arises as to the ultimate fate of the beam. A possibility that has intrigued researchers for a long time is diffractionless propagation of a beam over long distances. In other words, self-focusing exactly balances diffraction to force the beam into a nonlinear waveguiding mode. This phenomenon is called beam trapping or self-trapping. A simplified analysis of a uniform beam captures the essential physics (see Figure 4). The refractive index inside the beam is higher than outside by eqn [5]. Therefore, light incident on the 'interface' at the edge of the beam can experience total internal reflection if its angle of incidence exceeds the critical angle, dependent on the value of the nonlinear index. If the complement of the critical angle matches the diffraction angle of the beam, the beam is trapped in its own self-induced waveguide. The critical power $\mathscr{P}_{cr2}$ at which this occurs is given in Table 2. The condition for beam trapping is independent of beam size since both diffraction and self-focusing depend on beam area, and these size effects cancel.

A detailed solution of the wave equation for a Gaussian beam using the parabolic approximation yields a slightly different critical power $\mathscr{P}_{cr1}$, also given in Table 2. Numerical analyses of the wave

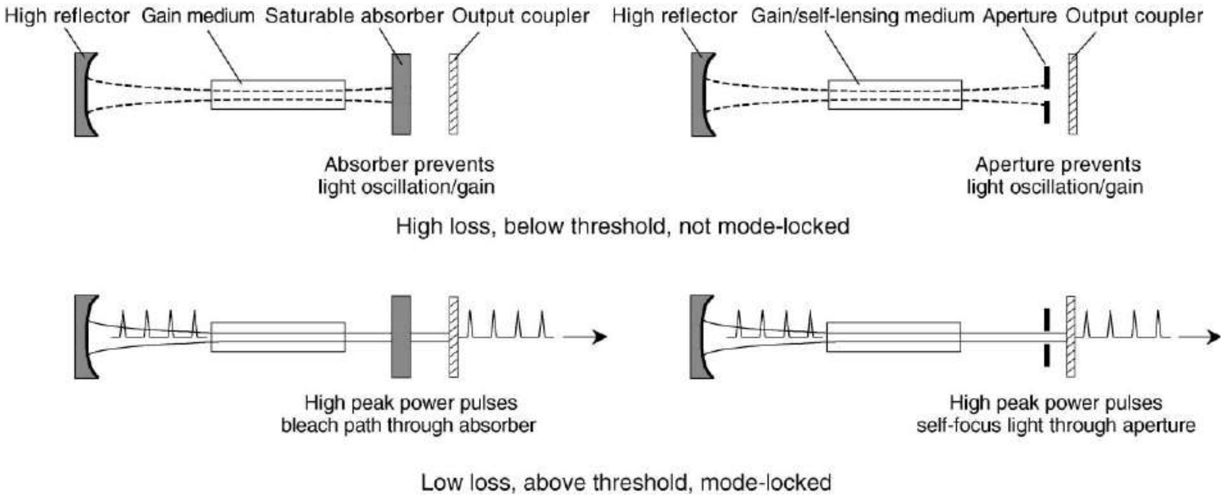
High reflector
Gain medium
Saturable absorber
Output coupler
High reflector
Gain/self-lensing medium
Aperture
Output coupler
Absorber prevents light oscillation/gain
Aperture prevents light oscillation/gain
High loss, below threshold, not mode-locked
High peak power pulses bleach path through absorber
High peak power pulses self-focus light through aperture
Low loss, above threshold, mode-locked

Figure 3 Laser mode-locking using a saturable absorber (left) and self-lensing in combination with an intracavity aperture (right).

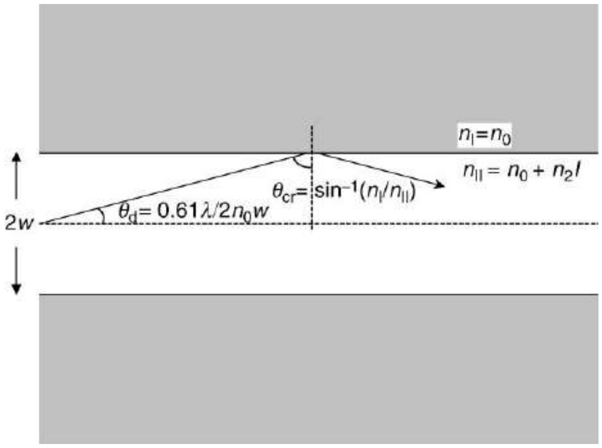
$n_I = n_0$
$n_{II} = n_0 + n_2 I$
$\theta_{cr} = \sin^{-1}(n_I/n_{II})$
$\theta_d = 0.61\lambda/2n_0w$
$2w$

Figure 4 Self-trapping conditions for a uniform beam of intensity I and width $2w$ in a Kerr medium.

Table 2 Critical power for beam trapping (collapse)

Gaussian beam (parabolic approximation)	$\mathscr{P}_{cr1} = \frac{1}{2}\frac{\lambda^2}{4\pi n_0 n_2}$
Gaussian beam ('whole beam' collapse)	$\mathscr{P}_{cr2} \approx \frac{(1.22\pi)^2}{8}\frac{\lambda^2}{4\pi n_0 n_2}$
Townes soliton, lower bound for arbitrary beam shape	$\mathscr{P}_{crT} \approx 1.86225\frac{\lambda^2}{4\pi n_0 n_2}$
Arbitrary beam (upper bound)[a]	$\mathscr{P}_{cr3} = G[f]\frac{\lambda^2}{4\pi n_0 n_2}$

[a] $G[f]$ defined in text.

equation for a whole Gaussian beam have shown for incident power such that $\mathscr{P}_{cr1} < \mathscr{P} < \mathscr{P}_{cr2}$ the beam comes to a 'soft' focus and then diffracts (partial beam focusing), but for $\mathscr{P} > \mathscr{P}_{cr2}$ the whole beam catastrophically collapses (i.e., the beam approaches a singularity). Catastrophic self-focusing has been observed experimentally, but as the beam collapses the intensity becomes so large that additional nonlinear effects, such as multiphoton absorption and stimulated Raman scattering, come into play. These effects rob the beam of energy and terminate the collapse. At this stage the beam will then spread by diffraction.

A more rigorous analysis of this problem is obtained by examining the nonlinear wave equation in the slowly varying amplitude approximation (propagation along z), but including the transverse variation of the wave. If the electric field of the optical wave is of the form $E(\mathbf{r}) = A(\mathbf{r})\exp(ikz)$, where now $k = n_0\omega/c$, then the wave equation can be written as

$$2ik\frac{\partial A}{\partial z} + \nabla_\perp^2 A + 4\varepsilon_0 n_2 ck^2|A|^2 A = 0 \qquad [7]$$

where the transverse Laplacian $\nabla_\perp^2$ is $\partial^2/\partial x^2 + \partial^2/\partial y^2$ in rectangular Cartesian coordinates, or $\partial^2/\partial r^2 + (1/r)\partial/\partial r$ in cylindrical coordinates (assuming azimuthal symmetry). This equation is known as the nonlinear Schrödinger equation (NLSE). In normalized coordinates $\xi = x/w$, $\eta = y/w$, $\rho = r/w$, $\zeta = z/2kw^2$, where w is the input beam radius, and $\psi = 2kw(c\varepsilon_0 n_2)^{1/2}A$, eqn [7] takes the normalized form of the NLSE:

$$\begin{gathered} i\frac{\partial\psi}{\partial\zeta} + \nabla_\perp^2\psi + |\psi|^2\psi = 0 \\ \nabla_\perp^2 = \frac{\partial^2}{\partial\xi^2} + \frac{\partial^2}{\partial\eta^2} = \frac{\partial^2}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial}{\partial\rho} \end{gathered} \qquad [8]$$

A waveguiding (i.e., diffractionless) solution of the form $\psi(\zeta, \rho) = \exp(i\zeta)R(\rho)$ exists for eqn [8], where $R(\rho)$ is a solution of

$$\nabla_\perp^2 R - R + R^3 = 0, \quad R'(0) = 0, \quad R(\infty) = 0 \qquad [9]$$

$R(\rho)$ is known as the Townes soliton. It has a radial profile similar to a Gaussian but differing in fine detail. Since the Townes soliton exactly balances diffraction and self-focusing, it carries a critical power $\mathscr{P}_{crT}$ given by the expression in Table 2. For comparison, $(1.22\pi)^2/8 = 1.83624$. Therefore, $\mathscr{P}_{crT}$ and $\mathscr{P}_{cr2}$ differ by only about 1%. However, the Townes soliton is unstable. An analysis of the solutions of the NLSE shows that a beam will not catastrophically collapse (i.e., its intensity will not blow up) if its power is less than $\mathscr{P}_{crT}$. Thus $\mathscr{P}_{crT}$ forms a lower bound for the critical power. An upper bound can also be found from analysis of the NLSE. It depends on the input beam profile $[\psi(\rho, 0) \propto f(\rho)]$ and is given as $\mathscr{P}_{cr3}$ in Table 2. The quantity $G[f]$ characterizes the input beam profile and is given by

$$G[f] = \frac{2\int |f|^2\rho\mathrm{d}\rho \int |\nabla f|^2\rho\mathrm{d}\rho}{\int |f|^4\rho\mathrm{d}\rho} \qquad [10]$$

For example, a Townes soliton, a Gaussian beam, and a hyperbolic secant beam yield $G = 1.86225$, $G = 2$, and $G = 1.86257$, respectively. Thus, if the beam power is such that $\mathscr{P}_{crT} \leq \mathscr{P} \leq \mathscr{P}_{cr3}$, the beam may collapse, but it will definitely collapse if $\mathscr{P} > \mathscr{P}_{cr3}$. We note that if the input beam is elliptical, with ellipticity $e = b/a$ (where a and b are the semi-major and semi-minor widths, respectively), then

$$\mathscr{P}_{crT}(e) \approx \left(0.6 + 0.2\frac{e^2+1}{e}\right)\mathscr{P}_{crT} \qquad [11]$$

This expression is accurate to within 1.5% of numerical simulations of the NLSE, and also holds well for Gaussian input beams.

The conclusion is that there is no stable self-trapped (2 + 1)-D beam (two diffracting dimensions, one propagating dimension) in a Kerr medium. The beam will diffract, blow up in a finite distance, or go through a single focusing–defocusing cycle. For an incident plane wave, the distance from the input plane to the blowup point is (assuming a Gaussian input)

$$z_{\mathrm{f}} \approx \frac{0.369\pi w^2/\lambda}{(\mathscr{P}/\mathscr{P}_{\mathrm{cr2}})^{1/2} - 0.858} \qquad [12]$$

where the approximation holds well for $\mathscr{P} > 1.2 \times \mathscr{P}_{\mathrm{cr2}}$.

Why, then, does self-trapping appear to occur in Kerr media? Time integrated photographs show what appears to be the formation of a filament, and laser damage tracks in solids follow the shape of a self-trapped filament. This can be explained in terms of the moving focus model. In the quasi-steady-state approximation, each time slice of the pulse comes to focus independently at different times and positions. Figure 5 shows a series of snapshots of a Gaussian spatial and temporal pulse (assume axial symmetry), where the peak power of the pulse is greater than $\mathscr{P}_{\mathrm{cr2}}$. Slices of the pulse with $\mathscr{P} > \mathscr{P}_{\mathrm{cr2}}$ will come to a focus at a finite distance $z_{\mathrm{f}}(\mathscr{P})$, which can be approximated by eqn [12]. The time at which that focus appears is determined by the time it takes for that slice of power $\mathscr{P}$ to propagate to the input plane ($z = 0$) plus propagate to $z_{\mathrm{f}}(\mathscr{P})$. As illustrated in Figure 5, slice 1a takes the minimum time to come to a focus. At a later time slices 2a and 2′ come to focus simultaneously but at different positions, 2a to the left and 2′ to the right of 1a, since $\mathscr{P}(2a) > \mathscr{P}(1a)$ and $\mathscr{P}(2') < \mathscr{P}(1a)$. The solid curves in the left of Figure 5 are the trajectories of the beam rays showing where each slice will come to focus. Slice 2b, having the same power as 2a, comes to focus at the same position as 2a, but later in time, and so on. The original focus thus appears to split in two, with one branch propagating forward and the other backward. The backward branch comes to a turning point when slice 4 of maximum power in the pulse center comes to focus, then propagates forward as time slices in the back part of the pulse come to focus. The motion of the foci traces out the U-shaped curve shown to the right in Figure 5. The time-integrated picture that appears like a filament is thus just the tracks of moving foci.

The question remains, do conditions exist under which stable self-trapped beams can form? The answer is yes, and these types of beams are called spatial solitons. The (1 + 1)-dimensional NLSE admits the fundamental spatial soliton solution:

$$\begin{aligned} A(x,z) &= A_0 \operatorname{sech}(x/w)\exp(iz/2kw^2) \\ |A_0|^2 &= \frac{1}{2\varepsilon_0 c n_2 (kw)^2} \end{aligned} \qquad [13]$$

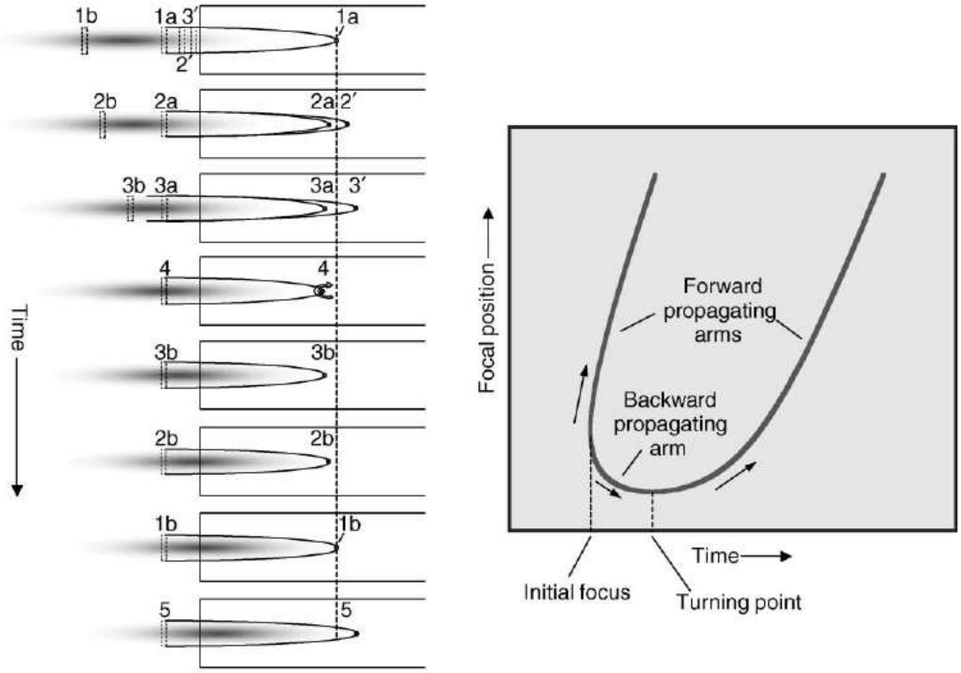

Figure 5 Moving focus model of apparent filament formation by quasi-steady-state self-focusing of a pulsed Gaussian beam. The series of snapshots on the left illustrate when and where different time slices of the pulse come to focus. The tracks of moving foci trace out the curve shown on the right.

The wave given by eqn [13] propagates along z with a constant width w along the x-axis. Such a $(1+1)$-D soliton can be formed in a slab waveguide, as illustrated in Figure 6, consisting of a nonlinear thin film deposited on a linear slab, where the wave is confined along y but can diffract along x. These types of solitons have been confirmed experimentally and shown to be quite robust. They are stable against perturbations and propagate as a self-induced waveguide. Since a soliton changes the refractive index where it propagates, it can also attract and guide other beams. The interactions or 'collisions' of solitons are an active area of current research. Intriguing potential applications include all-optical reconfigurable interconnects for telecommunications and optical computing based on the properties of soliton collisions.

Although $(2+1)$-D spatial solitons are not stable in ideal Kerr media, they can be in Kerr-like media with a saturating nonlinearity. For example, a two-level saturable absorber (e.g., an atomic vapor) at a wavelength tuned above the natural absorption frequency will exhibit a positive Kerr-like response at low intensity ($I \ll I_{\text{sat}}$), but then saturates as $(1 + I/I_{\text{sat}})^{-1}$ for $I \sim I_{\text{sat}}$ or larger, where I_{sat} is the saturation intensity. This can form a stable trapped beam, as was first reported in 1974 for a cw laser in Na vapor, because as the beam self-focuses and the intensity rises, the nonlinear index levels off asymptotically until a balance between diffraction and focusing is reached. This type of situation is stable against perturbations, but may undergo almost periodic focusing–defocusing cycles.

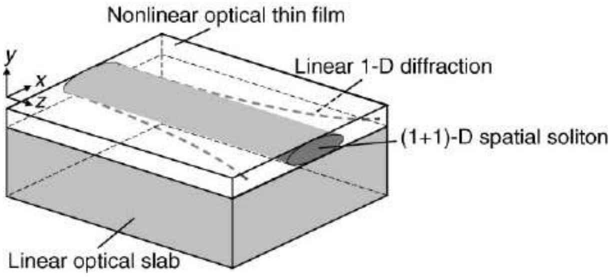

Figure 6 Spatial soliton propagation in a nonlinear slab waveguide.

Spatial solitons employing mechanisms other than the Kerr nonlinearity are also being vigorously studied. One of these is the photorefractive effect. An illustration of the generation of a slab soliton is shown in Figure 7. A Gaussian beam is focused to a line (e.g., by a cylindrical lens) and propagated as an extraordinary wave along the optic (z) axis of a uniaxial photorefractive crystal. A dc electric field is also applied along z. Charge carriers are generated in the slab beam by photo-ionization of dopant atoms and drift in the dc field. The lower resistivity in this region causes the voltage drop to be smaller and hence the field to be lower. The field induces a refractive index change by the Pockels effect ($\Delta n = -n_0^3 |r_{\text{eff}} E|/2$, where r_{eff} is the effective Pockels coefficient and n_0 is the background index). The index change is negative, but is reduced in magnitude where the field is lower. The index change depends on the local intensity, and the

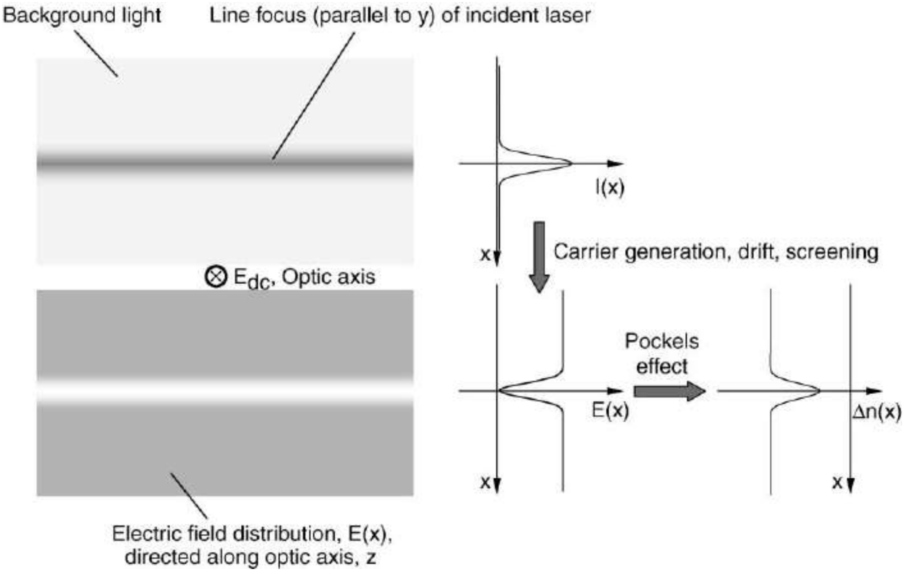
Background light
Line focus (parallel to y) of incident laser
I(x)
x
Carrier generation, drift, screening
⊗ E$_{dc}$, Optic axis
Pockels effect
E(x)
Δn(x)
x
x
Electric field distribution, E(x), directed along optic axis, z

Figure 7 Mechanism of slab soliton formation in a photorefractive material.

photorefractive mechanism produces an effective saturable self-focusing nonlinearity. (There are a limited number of charge carriers, and the field cannot be reduced below zero.) This produces a waveguide that traps the slab beam, which then propagates without spreading. Thus the photorefractive soliton is stable. This effect can be observed with milliwatt-level cw lasers, but the space-charge field takes time to build up to a steady state (typically of the order of seconds). Needle (circularly symmetric) solitons have also been observed by launching a TEM_{00} Gaussian beam in a photorefractive crystal, but the theoretical description is not as simple. Other types of spatial solitons and other mechanisms have been explored using the photorefractive and photovoltaic effects. The reader is referred to the literature for a more in-depth discussion.

Another mechanism useful for soliton formation is cascaded second-order (quadratic) effects. This includes such phenomena as second-harmonic generation (SHG) and other frequency conversion processes. Consider SHG with imperfect phase matching. (The reader is referred to other articles discussing harmonic generation and phase matching.) A fundamental (F) wave generates its second harmonic (SH), but because of dispersion $n(\omega)$, the two waves travel at different phase velocities. The SH(2ω) beats with the F(ω) wave to generate, via the second-order nonlinear polarization, another wave at ω that adds to the incident F wave. This produces a net phase shift of the F wave that accumulates continuously as the waves propagate together and periodically exchange power; this is called cascading. In the nondepleted pump approximation, this phase accumulation can be expressed as

$$\Delta\varphi = -\frac{\Delta k z}{2}\left\{1 - \left[1 + \left(\frac{2\kappa\sqrt{I}}{\Delta k}\right)^2\right]^{1/2}\right\}$$
$$\kappa = \frac{\omega\chi_{\mathrm{eff}}^{(2)}}{\sqrt{2\varepsilon_0 n_0^3 c^3}} \qquad [14]$$

where $\chi_{\mathrm{eff}}^{(2)}$ is an effective second-order susceptibility (dependent on crystal symmetry, polarization vectors, and propagation direction), $\Delta k = 2k(\omega) - k(2\omega)$ is the phase mismatch, I is the intensity of the F wave, and n_0 is the approximate refractive index for both F and SH waves. Although not due to a refractive index change, this phase change produces a self-focusing effect for a bell-shaped intensity. For a small enough pump intensity or sufficiently large phase mismatch, an equivalent n_2 can be approximated by

$$n_2 = \frac{\omega(\chi_{\mathrm{eff}}^{(2)})^2}{2\varepsilon_0 n_0^3 c^2 \Delta k} \qquad [15]$$

Note that the sign of n_2 depends on the sign of Δk. For $I \sim (\Delta k/2\kappa)^2$ or larger, eqn [15] is not valid and eqn [14] must be used. Note that the equivalent index change is then sublinear with respect to I, and soliton formation is stable. When pump depletion cannot be ignored, the F and SH coupled-wave equations must be solved numerically to analyze soliton dynamics. Figure 8 illustrates the difference between a beam diffracting normally and a diffractionless quadratic soliton. When an F beam is launched in a nonlinear medium, it does not immediately form a soliton, but evolves into one, shedding some energy in the process. Note that this soliton does not consist of a single frequency beam, but depends on the nonlinear coupling between the ω- and 2ω-waves that continuously exchange energy as the beams propagate. For this reason, these solitons are called simultons.

Self-Focusing in Space–Time

The moving focus model describing the self-focusing of optical pulses becomes inadequate when dispersion effects must be included. This is true in the case of ultrashort pulses (typically ~100 fs or less). To properly describe the propagation of a pulse with group velocity v_g, eqn [7] must be modified by $\partial A/\partial z \to [\partial/\partial z + (1/v_g)\partial/\partial t]A$. In addition, to include the effects of GVD, a term $-k_2 k(\partial^2 A/\partial t^2)$ is added, where $k_2 = (\partial^2 k/\partial\omega^2)$ is the GVD coefficient. For anomalous GVD, $k_2 < 0$. Taking this to be the case and defining another normalized coordinate $\tau = (t - z/v_g)/(|k_2|k)^{1/2}w$, the normalized NLSE (eqn [8]) becomes

$$i\frac{\partial\psi}{\partial\zeta} + \left(\frac{\partial^2}{\partial\xi^2} + \frac{\partial^2}{\partial\eta^2} + \frac{\partial^2}{\partial\tau^2}\right)\psi + |\psi|^2\psi = 0 \qquad [16]$$

Note the complete symmetry in eqn [16] between the transverse space (ξ, η) and time (τ) variables. There is thus a correlation between diffraction (space) and dispersion (time). The corresponding phenomenon to self-focusing is self-phase modulation (SPM). This effect denotes the additional phase retardation experienced by a pulse in a Kerr medium due to its time-dependent intensity: $\Delta\varphi(t) = -(\omega_0/c)n_2 I(t)z$, where ω_0 is the center (carrier) frequency of the pulse. SPM produces new frequencies ($\mathrm{d}\varphi/\mathrm{d}t$) that are red-shifted on the leading edge of the pulse and blue-shifted on the trailing edge. The frequency shifts in a self-focused filament can be quite sizeable and create a large continuum (a white light pulse) that

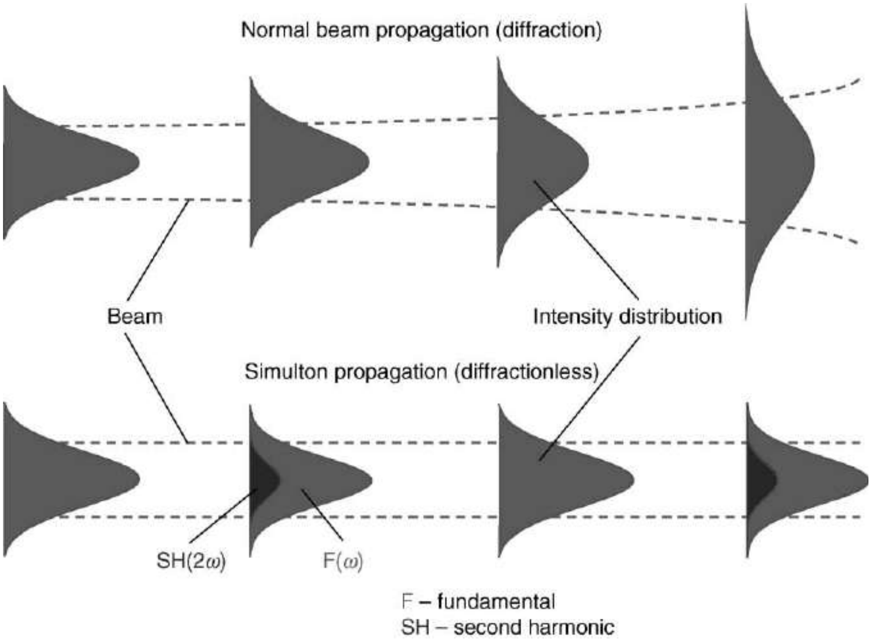
Normal beam propagation (diffraction)
Beam
Intensity distribution
Simulton propagation (diffractionless)
SH(2ω)
F(ω)
F – fundamental
SH – second harmonic

Figure 8 Comparison of normal beam propagation to simulton propagation in a quadratic nonlinear medium. In a simulton, the F and SH waves periodically exchange energy as they co-propagate in the medium.

can be used for ultrafast spectroscopy. In a medium with normal GVD the pulse spreads in time because blue frequencies in the trailing edge travel slower than red frequencies in the leading edge. However, for anomalous GVD the opposite is true, and the pulse can 'focus' or collapse in time.

The simultaneous spatiotemporal collapse of a beam of ultrashort pulses leads to the consideration of $(3+1)$-D soliton generation, where the third 'transverse' dimension is time. While the NLSE leads to unstable solitons for transverse dimensions >1, a saturating nonlinearity could possibly bound the collapse and lead to the formation of stable, diffractionless, dispersionless pulses. These types of pulses are called light bullets. Light bullets have a particle-like nature and display many interesting, exotic properties that could lead to ultrafast switching rates for optical digital logic systems in telecommunications and computing.

The saturable nature of self-focusing in quadratic nonlinear media makes them natural candidates in the search for light bullets, and $(2+1)$-D spatiotemporal simultons have been observed in crystals of lithium iodate. The drawback of ultrashort pulse simulton generation in quadratic media is the added complications of requiring anomalous GVD at both ω and 2ω, and dealing with the effects of group-velocity mismatch (GVM). The phenomenon of GVM involves the different group velocities of the F and SH pulses, which makes them walk away from one another. Without the physical overlap of these two pulses, the simulton will cease to exist. By tilting the amplitude front of the pulses with respect to their phase fronts, it is possible (over a limited distance) to compensate pulse walkoff due to GVM with spatial walkoff due to the birefringence of the quadratic medium. Accomplishing this, however, requires the beam to be in the form of a narrow stripe, and the propagation distance over which pulse overlap is good is greater for a longer stripe. This can be achieved by focusing a beam onto the medium with a cylindrical lens. The simulton then propagates without a change of width temporally or spatially (in 1-D). This is not a light bullet, however, because the confinement is not in all three dimensions.

The elliptic nature of the diffraction–dispersion operator in eqn [16] makes pulse collapse a possibility in nonlinear media with anomalous GVD. With normal GVD, the operator is hyperbolic. For such a medium, it has been predicted and observed that instead of collapsing, a self-focused pulse in a Kerr medium will split in time. Since the red-shifted frequencies generated at the front of the pulse travel faster than the blue-shifted frequencies produced at the back, the front of the pulse separates from the back part. Collapse of the pulse, as well as self-trapping, is not favored under conditions of pulse splitting.

Given this property of the NLSE, it is somewhat surprising that the reportedly first observation of light bullets was in a quadratic medium with normal GVD.

Focusing a beam of Gaussian profile in both space and time near the entrance to a lithium triborate crystal under the condition of a large positive phase mismatch, the pulses propagated without spread temporally or spatially (2D) over a distance ~20 mm (the length of the crystal). However, the pulse did not retain its shape, but evolved into a so-called 'X-wave', or X-shaped soliton. A generic X-shaped soliton (or, strictly speaking, simulton) is shown in Figure 9. The pulse has a conical (clepsydra, or hourglass-shaped) 3D structure with a central hump and slowly decaying conical tails. Interestingly, an X-shaped pulse has been shown to propagate without spreading in a linear optical medium exhibiting normal GVD. But such a pulse must be specially contrived at the input and may be impossible in many cases. In a quadratic nonlinear medium with normal GVD, the X-shaped light bullet appears to form spontaneously. The physics of this evolution lies buried in the coupled wave equations for the fundamental and second harmonic. In the strong phase-mismatched SHG process, weak perturbations in the waves at specific spatial and temporal frequencies can grow exponentially. The hyperbolic nature of the diffraction–dispersion operator with normal GVD leads to amplification of conical wave perturbations with frequencies approaching that of the X-wave. This amplification of the proper frequency components acts as a trigger that drives the evolution of the Gaussian spatiotemporal input pulse into the X-wave. This basic shape has been confirmed in experiments that outline the spatiotemporal intensity profile of the pulse exiting the rear of the crystal.

Other Related Effects (Multiphoton Nonlinearities)

At this point we should note that the susceptibility is frequency dependent. Optical waves couple principally to bound electrons, which can only occupy discrete energy levels. When the photon energy $\hbar\omega$ is nearly resonant with a difference between these energy levels, the dielectric response of the medium is enhanced. This well-known property of linear optics, accounting for absorption of light as well as normal and anomalous dispersion, is also applicable in nonlinear optics. The third-order susceptibility is actually a complex number. A sketch of its real and imaginary parts is given in Figure 10, the real part relating to n_2. The imaginary part is associated with optical loss, if positive, or gain when it is negative. In the present case the frequency dependence of the susceptibility displays a resonance (i.e., peak of the imaginary part) at a frequency $\omega = 2\omega_0$. Referring to Figure 10, when the optical frequency is significantly less than ω_0, the imaginary part of $\chi^{(3)}$ is small compared to the real part and nonlinear refraction dominates. However, when the frequency is comparable to ω_0, the resonance at $2\omega_0$ plays a major role. The real part of $\chi^{(3)}$ is resonantly enhanced, but the imaginary part signifies loss: two photons are simultaneously absorbed. This is called two-photon absorption (2PA). The intensity-dependent absorption loss per unit length is given by $\alpha_2 I$, where α_2 is called the two-photon absorption coefficient (often designated by β) and has physical units of cm W^{-1}.

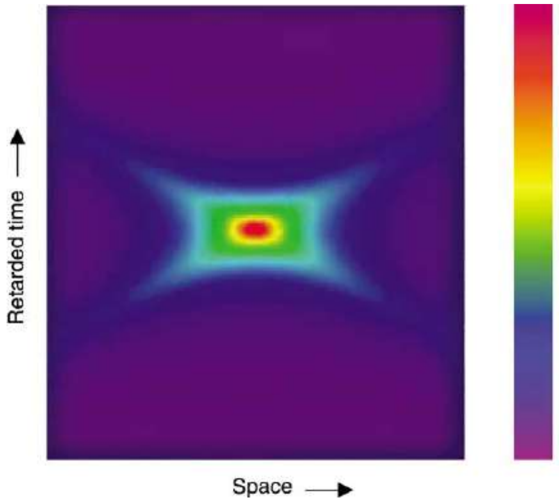

Figure 9 Generic shape of an X-wave light bullet as a function of space and retarded time ($t - z/v_g$) coordinates. The light bullet has radial symmetry.

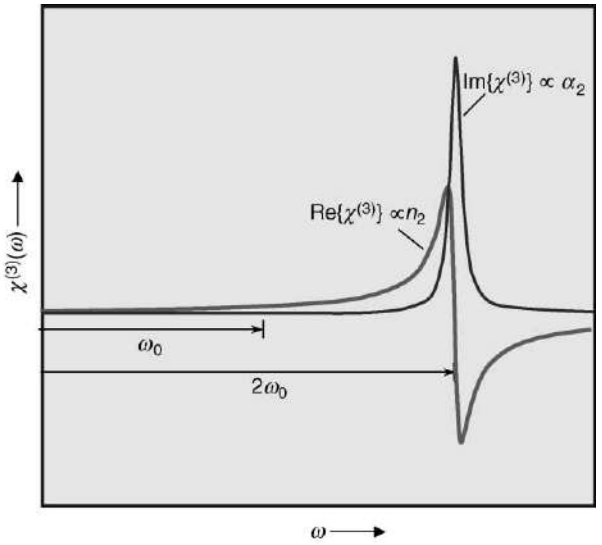

Figure 10 Sketch of the real and imaginary parts of the frequency-dependent third-order susceptibility $\chi^{(3)}$ illustrating a two-photon resonance.

In many materials at a variety of optical frequencies, 2PA will naturally accompany self-focusing. Since these two phenomena are both intensity dependent, while one decreases the intensity and the other increases it, there is an obvious interplay between the two. We illustrate the effect of 2PA in Figure 11 by comparison of the phase retardance of a Gaussian beam (in the parabolic approximation) with and without 2PA in a 'thin' medium, for which the size of an optical beam is nearly constant (diffraction ignored). Obviously, $\Delta\varphi(r)$ does not reproduce the parabolic shape when strong 2PA is present. Actually, it is possible to have multiphoton absorption (MPA) of m photons with corresponding coefficient α_m (related to $\chi^{(m)}$). The effects of 3PA and 4PA are also shown in Figure 11. 2PA, or more generally MPA, distorts the phase and modifies the focusing properties of the beam. This shows up in the Z-scan, for example, as a distortion of the dispersion-like transmittance curve: the valley is deeper and the peak is suppressed.

The reason for this phase distortion is that nonlinear absorption changes the shape of the beam: absorption is stronger where the intensity is higher. For MPA, the intensity at position z in a thin medium is

$$I(r,z) = \frac{I(r,0)}{\left[1 + (m-1)\alpha_m I^{m-1}(r,0)z\right]^{1/(m-1)}} \qquad [17]$$

The other obvious effect of MPA is that it reduces the strength of self-focusing since it robs power from the beam. The interplay of the two effects shows up more in thick media where the beam can diffract and change size. When self-focusing is dominant at the front of the medium, the intensity will start to blow up as the beam collapses, but the intensity becomes so large that MPA becomes much stronger. This robs energy from the beam and can stop or reverse the collapse. When MPA is dominant at the start, the beam will still self-focus, but MPA flattens the beam in the center causing the edges of the beam to experience dramatic diffraction effects; for example, a ring structure can develop.

The intensity of a self-focused filament can be quite high. At a level ~10^{11} W cm^{-2} or higher, it is possible for MPA to lead to multiphoton ionization (MPI) of the medium. This is illustrated in Figure 12 where the absorption of m photons promotes an electron from the ground state to the continuum in a gas, or from the valence band to the conduction band in a solid. The ionization rate is $\sigma_m I^m$, where m depends on the optical frequency and the ionization potential of the material. For example, at a wavelength of 810 nm, $m = 11$ for nitrogen and $\sigma_{11} = 5.08 \times 10^{-144}$ s^{-1} cm^{22} W^{-11}. We note that for intensities larger than ~10^{14} W cm^{-2}, the dominant ionization mechanism is tunneling, where a 'quivering' electron has sufficient energy to tunnel through the coulombic barrier. MPI can create a plasma whose optical properties depend on the density of free electrons. The refractive index of the plasma is nonuniform, and for $\omega > \omega_p = (N_e e^2/\varepsilon_0 m_e)^{1/2}$, the plasma frequency is <1; N_e, e, and m_e are the electron density, charge, and mass, respectively. Hence, plasma formation leads to defocusing. Consequently, self-focusing in air (which has a positive n_2) can lead to plasma formation by MPI, and the resultant defocusing can stabilize the formation of a self-trapped filament of light propagating over a distance of meters. These filaments have been called 'gas-induced solitons'. High peak power femtosecond pulses are typically used to generate these. The pulse width is short compared to the mean

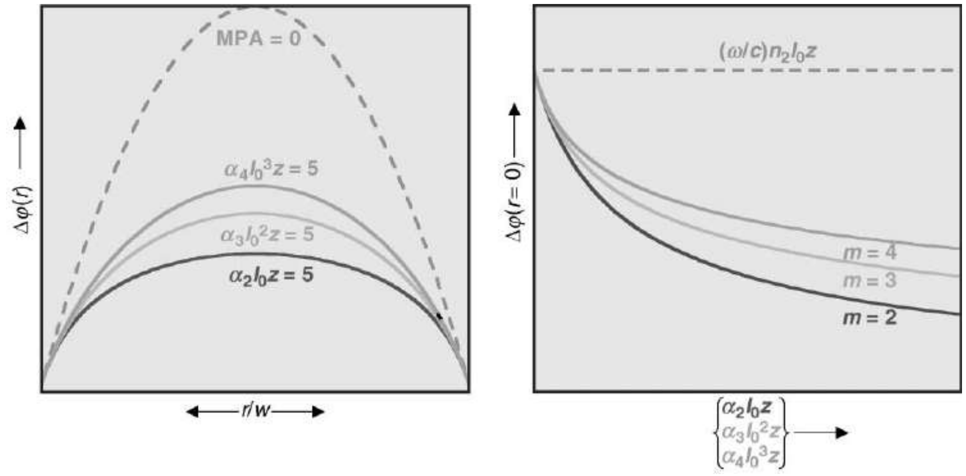

Figure 11 The effects of two-, three-, and four-photon absorption on the intensity-induced phase retardation due to a Kerr self-focusing nonlinearity. The radial dependence of $\Delta\varphi$ on MPA in the parabolic approximation is shown on the left, and the peak of $\Delta\varphi$ as a function of nonlinear absorption loss is plotted on the right.

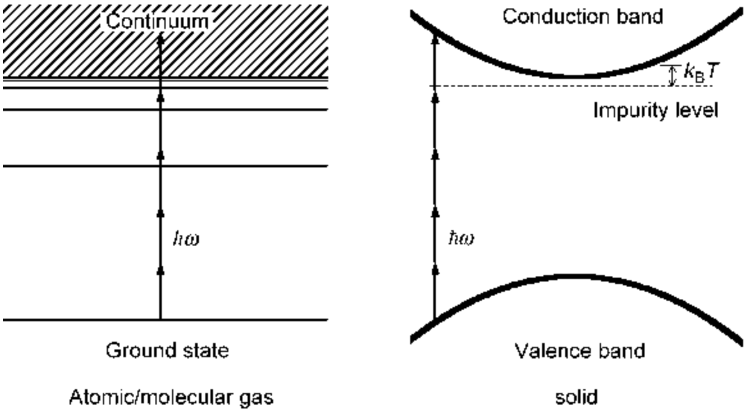

Figure 12 Multiphoton ionization in a gas (left) and a solid (right).

collision time of molecules, and the energy is low so that plasma heating is negligible.

The electron density in plasmas induced by MPI in gases is typically low enough that absorption by free electrons can be ignored. This is generally not the case in solids. MPI can provide enough 'seed' electrons to induce avalanche ionization. As illustrated in Figure 12, impurities may also supply these seed electrons by thermal ionization if they lie within an energy $\sim k_BT$ of the conduction band (k_B is the Boltzmann constant and T is the temperature). Free electrons can absorb electromagnetic energy while colliding with ions or atoms (a process called inverse bremsstrauhlung) and accelerate in a high-intensity field. They strip electrons from other atoms or molecules by impact ionization. The process multiplies free electrons and avalanche ionization ensues. A plasma forms, absorbs more energy and expands, developing a shock wave that can lead to a permanent crack in the solid. This phenomenon is called optical breakdown. Typically observed with nanosecond and picosecond pulses, permanent laser damage of the material also robs energy from the beam and generally terminates the collapse of a self-focused beam. When the electron density is sufficiently high so that $\omega_p > \omega$, the plasma absorbs and scatters light, a process that can be put to use in optical power limiters.

Many additional nonlinear phenomena play important roles when the intensity of a self-focused beam gets this large (e.g., stimulated Raman scattering, stimulated Brillouin scattering, etc.). The reader is referred to the literature on nonlinear optics for more in-depth discussions of these effects.

List of Units and Nomenclature

Critical power	[W]
Electric field	[V m^{-1}]
Four-photon absorption coefficient	[cm^5 W^{-3}]
Intensity	[W cm^{-2}]
Multiphoton (m-photon) absorption coefficient	[cm^{2m-3} W$^{-(m-1)}$]
Multiphoton (m-photon) ionization rate coefficient	[s^{-1} cm^{2m} W^{-m}]
Nonlinear refractive index coefficient	[cm^2 W^{-1}]
Pockels coefficient	[m V^{-1}]
Second-order susceptibility	[m V^{-1}]
Third-order susceptibility	[m^2 V^{-2}]
Three-photon absorption coefficient	[cm^3 W^{-2}]
Two-photon absorption coefficient	[cm W^{-1}]
Vacuum permittivity	[C V^{-1} m^{-1}]

See also

Laser-induced Damage of Optical Materials. **Materials Characterization Techniques:** $\chi^{(3)}$. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Nonlinear Optics, Basics:** Cascading; Kramers–Krönig Relations in Nonlinear Optics; Nomenclature and Units; Photorefraction; Ultrafast and Intense-Field Nonlinear Optics; $\chi^{(2)}$–Harmonic Generation. **Photonic Crystals**: Nonlinear Optics in Photonic Crystal Fibers. **Scattering:** Stimulated Scattering. **Solitons:** Bright Spatial Solitons.

Further Reading

Boyd RW (1992) *Nonlinear Optics*. Boston, MA: Academic Press.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Chiao RY, Garmire E and Townes CH (1964) Self-trapping of optical beams. *Physics Review Letters* 13: 479–482.

Crosignani B and Salamo G (2002) Photorefractive solitons. *Optics and Photonics News* 13: 38–41.
Kivshar YS and Stegeman GI (2002) Spatial optical solitons. *Optics and Photonics News* 13: 59–63.
Marburger JH (1975) Self-focusing: theory. *Progress in Quantum Electronics* 4: 35–110.
Sheik-bahae M, Said AA, Wei TH, Hagan DJ and Van Stryland EW (1989) Sensitive measurement of optical nonlinearities using a single beam. *IEEE Journal of Quantum Electronics* 26: 760–769.
Sheldon SJ, Knight LV and Thorne JM (1982) Laser-induced thermal lens effect: a new theoretical model. *Applied Optics* 21: 1663–1669.
Shen Y (1984) *The Principles of Nonlinear Optics*. New York: Wiley.
Silberberg Y (1990) Collapse of optical pulses. *Optics Letters* 15: 1282–1284.
Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn. New York: Marcel Dekker.
Torner L and Sukhorukov AP (2002) Quadratic solitons. *Optics and Photonics News* 13: 42–47.
Trillo S and Torruellas WE (eds) (2001) *Spatial Solitons*. Berlin: Springer-Verlag.
Wise F and Di Trapani P (2002) The hunt for light bullets – spatiotemporal solitons. *Optics and Photonics News* 13: 28–32.

Three-Dimensional Microfabrication

S M Kuebler, University of Central Florida, Orlando, FL, USA

M Rumi, Georgia Institute of Technology, Atlanta, GA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear optical three-dimensional microfabrication (3DM) is a photolithographic technique that enables topologically complex 3D microstructures with feature size as small as 1 μm or less to be generated in a single exposure step by nonlinear photopatterning in a material. In the broadest context, 3DM encompasses schemes based on single-beam serial patterning and parallel patterning achieved by multiple-beam interference exposure. In both cases, the photopatterning is initiated by coherent multiphoton excitation (MPE). The material may be a glass, a polymerizable resin, or even a heterogeneous composite, such as a resin containing dispersed nano-particles. The limiting resolution, or the dimensions of the smallest feature that can be generated, is determined by the optical parameters of the excitation geometry and the physical and chemical response of the material to MPE.

3DM offers great promise as a tool for generating complex microdevices, such as micro-electromechanical systems (MEMS), microfluidics, and micro-optical components. New frontiers in the technology of microdevices require that complex 3D structures can be produced in one or more materials that are optimized for a target application in terms of physical, chemical, and mechanical properties. Rapid and versatile fabrication is also essential. Extensive work has been done with conventional photolithography or electron-beam lithography, and more recently using soft-lithography approaches, such as microcontact printing. 3DM provides a new set of tools that overcomes some of the limitations inherent to other methods.

Single-Beam Three-Dimensional Microfabrication

3D Confinement

The simplest form of 3DM involves patterning a microstructure within a material using a single tightly focused laser beam that can activate the medium or one of its constituents by two-photon excitation (TPE) (Figure 1). The TPE initiates a photochemical or photophysical process (e.g., a polymerization reaction or a phase change) within the focal volume and results in an irreversible change in the material. The pattern of the target structure is impressed point by point as the focus is translated within the volume of the material. For some applications the final 3D microstructure is obtained when the unexposed material surrounding the photopatterned regions is removed. In the first report of this type of 3DM in 1990 two-photon-induced polymerization (TPIP) of an acrylate resin was used to fabricate free-standing 3D micron-scale letters on a substrate (Figure 2). The unexposed material does not have to be removed in a post-exposure process when the target structure consists of the phototransformed material supported within the matrix of the unexposed media, as in the case of 3DM of waveguide circuits embedded within a host slab.

3DM is possible because, under tight focusing, TPE and the subsequent material transformation are

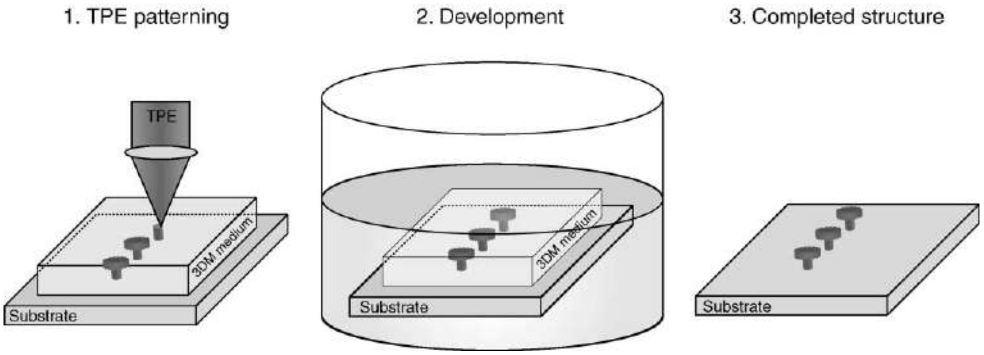

Figure 1 General scheme for single-beam 3DM based on TPE. (1. TPE patterning): The 3DM medium is supported on a substrate and translated in 3D-space relative to a tightly focused laser beam. The laser transforms the material only in the region around the focal point through a photochemical or photophysical process initiated by TPE. (2. Development): In cases where the final structure is needed free from the unexposed material, the sample is treated with a 'developer' that removes the unexposed material, for example by dissolving it in a solvent. (3. Completed structure): Following removal of the unexposed material, the final 3D structure is obtained free-standing on the substrate.

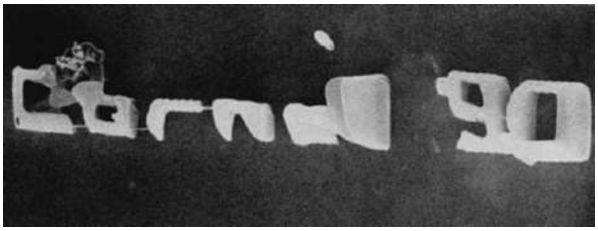

Figure 2 Scanning electron micrograph of a structure obtained by TPIP-3DM in an acrylate-ester resin. The excitation wavelength was 630 nm. The voxel dwell time was 10 ms along the letters. The thin filaments connecting some of the letters are approximately 100 nm wide. Reproduced from Strickler JH and Webb WW (1990) Two-photon excitation in laser scanning fluorscence microscopy. *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 1398: 107–118. Copyright (1990), with permission of SPIE.

confined at the focus within a volume of $\sim(\lambda/n_0)^3$, where λ and n_0 are the vacuum wavelength of the exciting radiation and the refractive index of the material, respectively. Degenerate TPE is a resonant third-order nonlinear optical process in which a species is promoted to an excited electronic state by the simultaneous absorption of two photons of equal energy. Relative to conventional one-photon excitation (OPE), TPE is typically achieved using longer-wavelength radiation (often red or near infrared), for which the combined energy of two photons is sufficient to promote the species into one of its lower lying excited electronic states (Figure 3, left). The rate of TPE, that is the number of species excited per unit volume and time, $R_{2h\nu}$, is proportional to the square of the intensity, I. For a focused Gaussian beam, the on-axis intensity attains its maximum, I_0, at the focus and decreases with distance from the focal plane, z, approximately as $I \propto z^{-2}$. For a focused Gaussian beam, the on-axis intensity is given by $I = I_0/[1 + (z/z_R)^2]$, where z_R is the Rayleigh range of the focused wavefront. Thus, $I \propto z^{-2}$ in the limit of $z \gg z_R$. Consequently, the excitation rate decreases as $R_{2h\nu} \propto z^{-4}$. Similarly, the excitation rate decreases rapidly away from the propagation axis. This results in a region of high excitation, and thus material transformation, that is tightly confined both laterally and longitudinally in space about the focal point (Figure 3, right). This situation may be contrasted with that for OPE, for which the excitation rate, $R_{1h\nu}$, is linearly proportional to I, so $R_{1h\nu} \propto z^{-2}$. The weaker dependence of $R_{1h\nu}$ on z results in material being excited appreciably throughout the irradiation volume, with less longitudinal confinement at the focus than is achieved under TPE. Similar arguments can be used to describe 3DM involving higher-order excitation, for which the excitation is confined to an even smaller volume for a given λ.

The rate for molecular n-photon excitation, $R_{nh\nu}$, represents the average number of molecules that are excited per unit volume per unit time and is given by

$$R_{nh\nu} = \frac{1}{n(h\nu)^n}\sigma^{(n)} N I^n \qquad [1]$$

where n is the order of the excitation process, h is Planck's constant, ν is the frequency of the exciting radiation, $\sigma^{(n)}$ is a molecular n-photon absorption cross-section, which has units of [$\mathrm{cm}^{2n}\,\mathrm{s}^{n-1}\,\mathrm{photon}^{-(n-1)}$], N is the number density of the photo-excitable species [cm^{-3}], and I is the intensity [$\mathrm{W\,cm}^{-2}$]. The factor of $1/n$ accounts for the fact that n photons must be absorbed to promote a single molecule to an excited state. Note that $\sigma^{(n)}$ is a

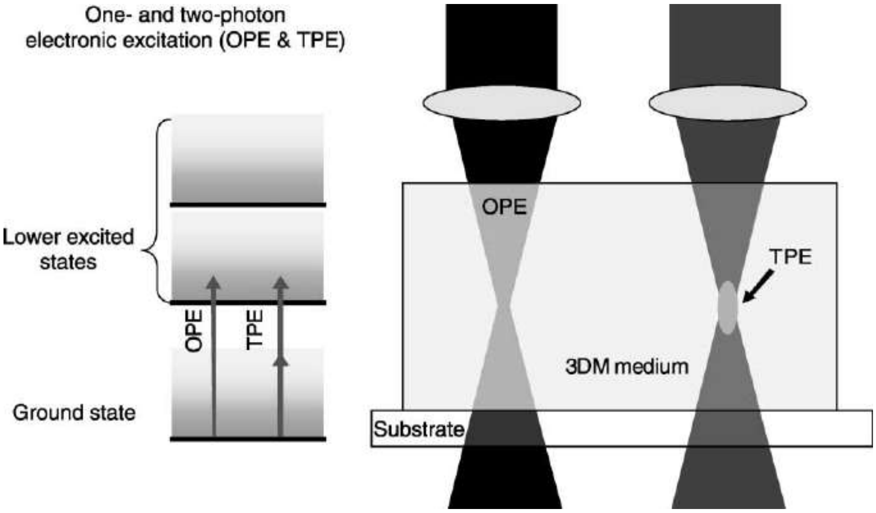

Figure 3 (Left) Jablonski representation of electronic one- and two-photon excitation (OPE and TPE, respectively). (Right) Illustration of a 3DM medium undergoing OPE and TPE using focused short- (black) and long-wavelength (dark gray) radiation, respectively. The regions that are appreciably excited are shown in light gray. Under OPE the medium is excited throughout the length of the interaction volume. Under TPE the excitation is confined in three dimensions to a small region around the focal point.

molecular parameter that is independent of the intensity. The molecular cross-sections for OPE and TPE, $\sigma^{(1)}$ and $\sigma^{(2)}$, are respectively equivalent to the parameters σ and δ that are commonly encountered in the literature.

For MPE to occur, the intensity must be sufficiently large that there is a high probability of more than one photon arriving simultaneously at the species to be excited. Typical magnitudes of the OPE and TPE cross-sections are $\sigma^{(1)} \approx 10^{-17}\,\mathrm{cm}^2$ and $\sigma^{(2)} \leq 10^{-47}\,\mathrm{cm}^4\,\mathrm{s\,photon}^{-1}$, and under conventional excitation conditions, $R_{2h\nu} \ll R_{1h\nu}$. To achieve $R2h\nu \approx R1h\nu$, $I \approx 2h\nu\sigma^{(1)}/\sigma^{(2)}$ is required. For the cross-sections given above, this condition corresponds to $I \approx 500\,\mathrm{GW\,cm}^{-2}$ at $\lambda = 800\,\mathrm{nm}$. Such peak powers can be achieved by focusing ultra-short laser pulses, so mode-locked femtosecond- or picosecond-pulse lasers are generally used for 3DM. Tight focusing in 3DM is desirable then, not only to provide a means for 3D confinement, but also to increase the intensity to a level that $R_{nh\nu}$ becomes large enough to produce a significant material change.

Opto-Mechanical Implementation

One method for implementing single-beam 3DM is illustrated in Figure 4. An excitation laser beam is directed onto a high numerical aperture (NA) objective. A beamsplitter, detector, attenuator, and shutter are placed upstream of the objective to measure and control the average power at the sample, $\langle P \rangle$. A substrate is coated with the photo-active medium and attached to a 3-axis nano/micro-positioner under the objective. The interior of the photo-active material is patterned during the exposure by translating the sample relative to the focus of the laser beam and shuttering the beam as needed. The process is automated as a computer controls the sample translation coordinates and the exposure conditions for each volume element (voxel). This configuration is well suited for patterning solid or semi-solid photo-media. An index-matching fluid must be used with high-NA objectives to achieve the smallest focal spot size. If the photo-medium is a liquid, a coverslip or other thin transparent barrier must be placed between the index matching fluid and the photomedium so that the latter does not become contaminated.

Alternatively, an inverted configuration may be used. Here, the laser is focused through the substrate into the medium. The objective and the index matching fluid are kept out of contact with the medium by the substrate. In this case, the maximum depth that can be patterned is limited by at least two conditions: (i) The working distance of high-NA objectives is small, often no more than 200 μm, so the substrate must be kept thin (~100 μm); and (ii) unless the viscosity of the medium is very high, patterning must begin at the medium/substrate interface and proceed layer by layer away from it, so that the structure remains anchored to the substrate. This could be disadvantageous, as the change in refractive index in the previously patterned layers could alter the intensity distribution

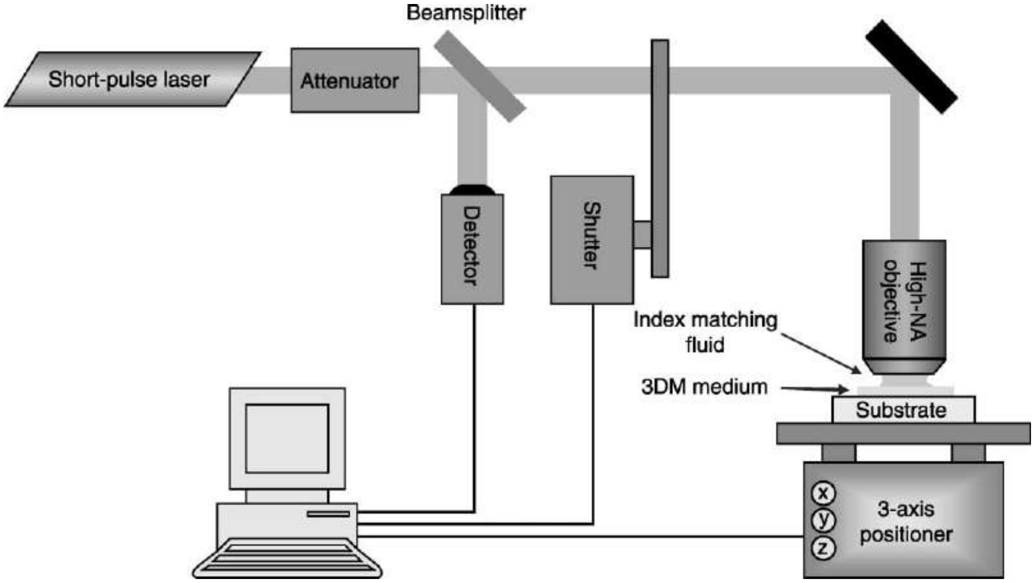

Figure 4 Optomechanical setup for single-beam 3DM.

at the focus. In all configurations, it is best to minimize the number of optical interfaces to maintain the quality of the focus.

A scanning confocal microscope system provides a convenient platform for single-beam 3DM. The nano/micro-positioner can be mounted directly on the sample stage of the microscope, and the optical components of the confocal system can be used directly for 3DM. The confocal system also enables the medium and the structure to be imaged *in situ*, using confocal reflectance or fluorescence imaging. If the material is fluorescent under MPE, a 3D image can be obtained by scanning multiphoton fluorescence microscopy. These imaging modes facilitate 3DM by providing a means for locating the focus at an absolute position within the photo-active medium.

3DM can also be realized by moving the focused beam relative to a fixed sample. One convenient implementation utilizes galvanometer-driven mirrors to scan the beam in planes parallel to the substrate (xy-plane). Motion in the z-direction is accomplished by raising and lowering the objective relative to the substrate. A galvo-based system typically enables faster fabrication, but may not afford motion as precise as that obtained using a multi-axis nano-positioner.

Excitation Sources

The most commonly used and arguably the most convenient and reliable laser for 3DM is the continuous-wave (CW) mode-locked titanium sapphire (Ti:S) laser. Ti:S lasers routinely produce sub-100 fs pulses at high repetition rates (~80 MHz) with single-pulse energies, E_p, of ~1 nJ. The Ti:S laser emission spans the range of 700–1050 nm, making it particularly well suited for two- and three-photon excitation. $\langle P \rangle$ can be as high as several watts at the peak of the gain curve ($\lambda \approx 800$ nm) with high pulse-to-pulse and time-averaged stability (RMS $\approx 1\%$). Such systems may also produce a nearly Gaussian TEM_{00} mode, which is helpful in achieving the tightest focus in the material. Note that focusing a 1 nJ, 100 fs pulse to a 1 μm-radius spot generates a focused average peak intensity of 320 GW cm^{-2}, as is needed to achieve photo-excitation rates of $R_{2h\nu} \approx R_{1h\nu}$.

For higher peak powers, amplified femtosecond (AFS) lasers can be used for 3DM. These systems can produce ~100 fs pulses with $E_p \approx 1$ mJ or higher at repetition rates of 1–100 kHz. Given that a continuous microfabricated structure should be generated from partly overlapping voxels, a low repetition rate reduces the speed with which the laser beam may be scanned within the material. An AFS laser can be used to pump an optical parametric generator to extend the output wavelength range.

Recently, some researchers have turned their attention toward identifying compact solid-state turn-key lasers for 3DM that could be more convenient and cheaper alternatives to the complex systems described above. A commercial Nd:YAG microlaser ($\lambda = 1064$ nm, $E_p = 5.4$ μJ, $\tau_p = 640$ ps, 13.6 kHz) and a mode-locked erbium-doped fiber laser ($\lambda = 780$ nm, $E_p \approx 0.5$ nJ, $\tau_p = 100$ fs, 50 MHz) have both been shown to be satisfactory for 3DM in certain acrylate-based media.

Material Systems

Structural Materials

The most widely used materials for 3DM are based on acrylates that are patterned by multiphoton induced polymerization. The process begins when an initiator chromophore **C** is excited by n-photon absorption to $\mathbf{C}^*$, with an efficiency specified by the n-photon cross-section $\sigma^{(n)}$. $\mathbf{C}^*$ then generates an initiating radical **I·**. Acrylate monomers (**M**) can react with **I·** and undergo a free-radical chain reaction that leads to high molecular weight polymer. With sufficient exposure, the starting material is converted to less soluble high-molecular weight polymer chains or densely cross-linked polymer networks. The photo-patterned structure can then be 'developed' by immersing the sample into a solvent that removes the unexposed material, leaving behind a free-standing structure that is a replica of the photopattern.

Photo-polymerization starts when the MPE exposure generates enough **I·** to react with and consume all of the inhibitor in the local volume at point $\vec{r}$ within a homogeneous medium. In this discussion we assume that the local concentration of inhibitors is not changed by diffusion on the exposure timescale. The total number of **I·** generated during an exposure at $\vec{r}$ per unit volume, $N_{\mathrm{I\cdot}}(\vec{r},t)$, is obtained from the time-integrated MPE rate:

$$N_{\mathrm{I\cdot}}(\vec{r},t) = \frac{\phi_{\mathrm{c}}^{(n)}\sigma^{(n)}N_{\mathrm{C}}}{n(h\nu)^n}\int_{-\infty}^{t} I(\vec{r},t')^n \mathrm{d}t' \qquad [2]$$

Here, N_{C} is the number density of **C**, and $\phi_{\mathrm{c}}^{(n)}$ is the n-photon chemical quantum yield, defined as the number of **I·** generated divided by the number of $\mathbf{C}^*$ present after excitation. If we identify the initial number density of the inhibitors as N_{Inh}, then polymerization occurs when $N_{\mathrm{I\cdot}}(\vec{r},t) > N_{\mathrm{Inh}}$. In multiphoton-induced polymerization, $N_{\mathrm{I\cdot}}(\vec{r},t)$ is a nonlinear function of I, which generally involves a rapidly varying time envelope, as from a mode-locked ultrashort-pulse laser.

Several different commercial acrylate systems have been used successfully for 3DM, such as Nopcocure 800 (San Nopco), NOA 72 (urethane acrylate oligomer, Norland Products), and SCR 500 (blend of urethane acrylate monomers and oligomers, Japan Synthetic Rubber Co.). Researchers have also successfully developed custom formulations using blends of commercially available monomers, oligomers, and polymer additives to create resins with tailored physical, chemical, and optical properties. This aspect of acrylates, and more generally polymer-based systems, makes them one of the most attractive classes of materials for 3DM. Conventional one-photon UV- or blue-sensitive radical generators (e.g., benzil, benzoin methyl ether, 4,4′-bis(*N*,*N*′-dimethylamino)benzophenone, to name just a few) have most often been used, and these were activated by promoting the molecule into one of its low-lying electronic states by MPE over the range of 730–800 nm.

3DM based on epoxide polymerization has also been demonstrated using commercial resins including cross-linkable small molecular weight monomers (e.g., 4-vinyl-1-cyclohexene diepoxide) and epoxide oligomers (e.g., SU-8, first developed by IBM). Epoxide polymerization is most commonly initiated by molecules that, after excitation, generate a Brønsted acid, H^+. The propagating species is a carbocation, and bases present in the medium act as inhibitors. Relative to acrylates, epoxides shrink substantially less upon polymerization, leading to less distortion of the microstructure after development.

Acrylate chemistry has also been exploited for 3DM of functional composite materials. Acrylamide resin mixtures were used to fabricate bi-layer micro-cantilevers that deflect from the surface when illuminated with UV light. This type of microstructure could form the basis of a photo-activatable MEMS. 3D gratings were fabricated from an acrylate pre-polymer containing a dendrimer functionalized with cross-linkable acrylic *L*-phenylalanine groups. The dendrimer acts as a host encapsulation site that enables a laser dye to be loaded in the material at high weight-percent (see below). A variety of complex micro-structures have also been fabricated using 'Ormocer-I', an ORganically MOdified CERamic. Ormocer-I is an inorganic-organic hybrid polymer consisting of a Si–O–Si (siloxane) backbone that is functionalized with organic moieties, including cross-linkable methacryloxypropyl groups. Ormocers are extremely promising for 3DM because of the thermal stability and chemical inertness of the backbone and their attractive physical and optical properties.

3DM could be a powerful tool for fabricating biologically active structures and devices. In this context, hydrogels and Ormocers should be useful material systems as some formulations are known to be bio-compatible. Microstructures can also be fabricated from proteinaceous material. 3D microstructures were produced by patterned MPE of solutions containing the dye rose bengal and the proteins bovine serum albumin or fibrinogen.

The material systems discussed thus far are ‘negative-tone’ media. This means that only the exposed regions of the material remain after the post-exposure development, so the final structure is a replica of the exposure pattern. In contrast, ‘positive-tone’ media are solid-state systems for which the exposed regions become soluble in a developer. The final structure is then the inverse of

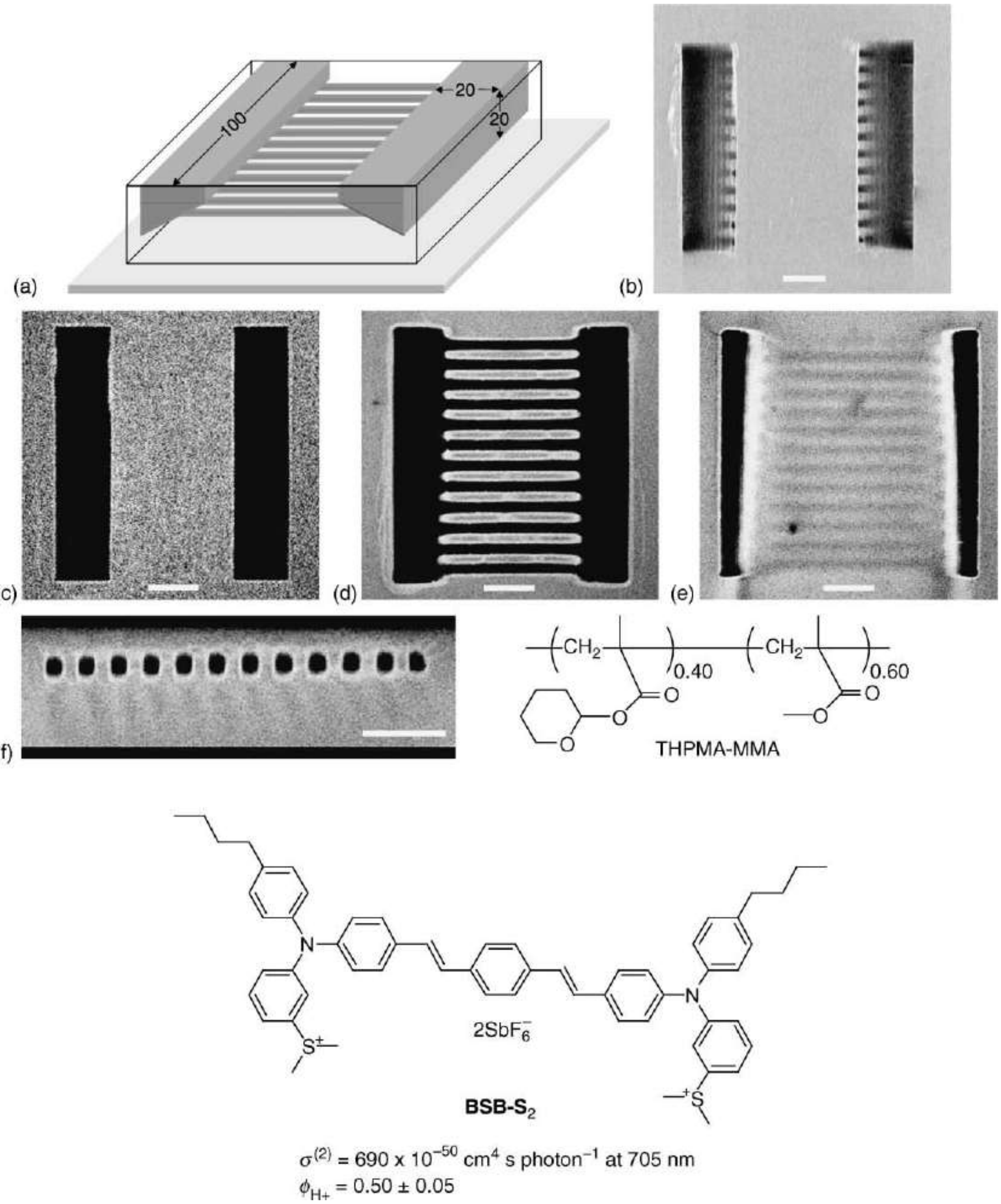

$\sigma^{(2)} = 690 \times 10^{-50}$ cm^4 s photon^{-1} at 705 nm
$\phi_{H+} = 0.50 \pm 0.05$

Figure 5 A 3D micro-channel structure fabricated by TPE-3DM using the positive-tone chemically amplified resist THPMA-MMA containing the two-photon PAG **BSB-S$_2$**. A 50 μm-thick film of the resist was exposed in the pattern of the target structure at 745 nm with tightly focused 80 fs pulses at $\langle P \rangle = 40$ μW and a linear scan speed of 50 μm s^{-1}. (a) Target structure consisting of two rectangular cavities (width: 100 μm, length: 20 μm, depth: 20 μm) with a sloped side-wall, that are connected by 12 channels (length: 50 μm; 4 μm by 4 μm cross-section) lying 10 μm below the surface and spaced apart by 8 μm (center-to-center). (b) Scanning electron micrograph of the final structure, viewed normal to the substrate. (c to e) Two-photon-fluorescence images of the final structure (viewed normal to the substrate): (c) at the surface of the film; (d) 10 μm below the surface; (e) 19 μm below the surface. (f) Two-photon-fluorescence cross-sectional image of the buried channels. The scale bar in b to f corresponds to 20 μm. Reproduced from Zhou W, *et al.* (2001) *Science* 296: 1106–1109. Copyright (2001), with permission of The American Association for the Advancement of Science.

the exposure pattern. Positive-tone 3DM was first demonstrated as early as 1992 with the generation of high aspect ratio trenches in thin layers of commercial photoresists. A positive-tone material system was designed for 3DM (Figure 5), is based on a chemically amplified resist and a high-sensitivity two-photon-activatable photo-acid generator (PAG) (discussed below). Using this medium, it was possible to fabricate subsurface microchannel structures and microgratings, using moderate laser scan speeds ($50\ \mu m\ s^{-1}$) and exceptionally low $\langle P \rangle$ ($40\ \mu W$).

Composites have been formulated that enable 3D-patterning of nanoscale metal features within a host matrix. A two-step process has been described that involves MPE of a Ag^+-containing sol–gel followed by thermal development of the latent image with a $AgClO_4$ solution. Using this system, 3D spirals of silver nano-particles embedded within the sol–gel matrix were generated. A $HAuCl_4$-doped poly(vinyl alcohol) composite was used to 3D-pattern gold nano-particles. It was shown that after TPE a particular sensitizer dye hosted in a poly(vinylcarbazole) matrix containing $AgBF_4$ and ligand-stabilized Ag nanoparticles, electron transfer from the dye to Ag^+ leads to nucleation of Ag^0 at the nanoparticles and the formation of conducting bulk metal features (Figure 6). Material systems like these should find wide application for generating micro- and nano-scale structures, having unusual and useful optical and electronic properties.

3DM in glasses is of great interest for photonic applications. Several groups have demonstrated that tightly focused infrared or visible amplified ultrashort laser pulses can be used to 3D-pattern pure and doped silica and other glasses. The patterned voxels can be generated as a result of photo-induced changes in the refractive index or extreme dielectric breakdown, resulting in pits within the material. The mechanisms for these two cases are not fully understood, but both are thought to be activated through an initial MPE event. The leading edge of the laser pulse generates a free-carrier plasma through MPE. The remainder of the pulse is increasingly absorbed by the plasma leading to avalanche ionization. The optical energy deposited into the local lattice results in cavitation, creating a pit at the focal point. Based on the electronic bandgap of silica, the process likely involves three- and six-photon absorption when 400- and 800-nm light is used, respectively. This approach has been used for the 3D-patterning fused silica and germanium doped glasses.

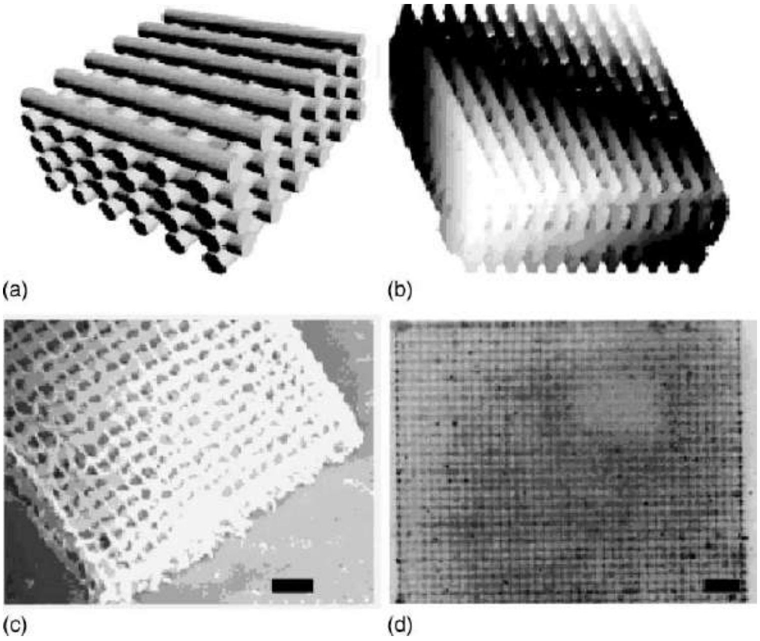

Figure 6 Metallic structures fabricated by TPE-3DM using a silver nanoparticle/polymer composite. (a) Model of the 'stack-of-logs' target structure. (b) 3D image constructed from a series of two-photon fluorescence microscopy images of the silver structure immediately following 3DM. (c) Scanning electron micrograph of the free-standing silver structure following removal of the surrounding unexposed composite. The scale bar corresponds to 10 μm. (d) Transmission optical microscopy image of the silver structure immediately following 3DM. The scale bar corresponds to 25 μm. Reproduced from Stellacci F, *et al.* (2001) *Advanced Materials* 14: 194–198. Copyright (2001), with permission of Wiley.

Multiphoton Initiators

For 3DM to develop into a widely used fabrication tool, material systems and patterning methods must be developed that enable rapid microfabrication at low cost. ‘Soft materials’, such as the polymer systems described above, are promising in this respect, as both the initiating species and the host resin can be engineered to optimize the multiphoton response and the physical and chemical properties of the patterned medium. It may also prove possible to create doped glasses that have physical properties close to that of neat silica and yet offer higher multiphoton sensitivity.

To develop this concept, let us consider 3DM in acrylates further. A high-sensitivity acrylate medium is one for which: (i) **C** strongly absorbs n-photons (large $\sigma^{(n)}$); (ii) $\mathbf{C}^*$ efficiently generates the initiating species **I·** (large $\phi_c^{(n)}$); and (iii) **I·** is an efficient initiator (high initiation rate). As criteria (i) and (ii) are predominantly determined by the molecular properties of **C**, it should be possible to engineer soft materials for 3DM by optimizing multiphoton initiators and incorporating these into pre-polymer resins that once patterned have the correct properties for the target application. The sensitivity of materials can be compared on the basis of the minimum average laser power, $\langle P \rangle_{\min}$, that is required to pattern a material using a fixed scan speed (or exposure time). Alternatively, we can use the product $\sigma(n)\phi c(n)$ as a multiphoton initiator figure-of-merit (FOM) that accounts for the strength of the multiphoton absorption and the efficiency with which **I·** is generated.

The chromophores that have been used most frequently for polymerization-based 3DM do efficiently generate highly active initiating species once excited; however, they were developed for one-photon response to UV and blue radiation and generally have low multiphoton FOMs. The use of xanthene dyes was explored as initiators for free-radical polymerization of acrylamides and it was found that the multiphoton sensitivity was low. As a result, 3DM using conventional initiators requires high laser power (typically $\langle P \rangle_{\min} \approx 10-100$ mW) and slow laser scan speeds (~50 μm s^{-1}). Under tight-focusing conditions, the resulting peak powers are often near the damage threshold of the materials, rendering the process unreliable from a manufacturing stand-point.

A series of **D–π–D** molecules was examined (Figure 7), where **D** is an electron-donating group and **π** is a conjugated π-electron bridge. This class of **D–π–D** chromophores was found to have large TPE cross-sections ($\sigma^{(2)} > 200 \times 10^{-50}$ cm^4 s photon^{-1}) and to initiate the polymerization of acrylates following TPE in the visible and near infrared, with $\langle P \rangle_{\min}$ a factor of 30–50 times lower than for conventional one-photon initiators. The sensitivity achieved with **D–π–D** initiators is so enhanced that TPIP-3DM was performed at laser scan-speeds as high as 9 cm s^{-1}, using $\langle P \rangle = 3$ mW, enabling a volume of 180 μm × 180 μm × 20 μm to be polymerized in 200 s.

D–π–A molecules have also been examined for TPIP (Figure 7), where **A** is an electron accepting functionality. Such molecules can have large TPE cross-sections, and some have been shown to activate TPIP efficiently in acrylates at $\lambda \approx 800$ nm. A derivative of Michler’s ketone was synthesized with an extended π-conjugated system (Figure 7). This molecule exhibited strong two-photon absorption over the broad range of 800–1100 nm and was also shown to be active toward acrylate polymerization at 1064 nm. A strategy was implemented for enhancing two-photon sensitivity based on up-converted fluorescence and used for 3DM of waveguide structures. The strongly two-photon-absorbing dye AF-183 was incorporated into a commercial acrylate photoresist containing a radical-generator that is activated by OPE at 315–450 nm. After AF-183 undergoes TPE at 800 nm, it fluoresces efficiently at ~465 nm. The up-converted emission then excites the one-photon initiator and radicals are generated.

The strategies discussed above are oriented toward enhancing multiphoton sensitivity by increasing $\sigma^{(2)}$ of the initiating chromophore. Little attention has been paid to the need of ensuring that the MPE couples to a mechanism for efficiently generating an active initiating species. This problem was addressed with the design of a two-photon PAG by covalently linking a **D–π–D** chromophore and sulfonium moieties, which can generate H$^+$ when reduced by photo-induced electron transfer (**BSB-S$_2$** in Figure 5). **BSB-S$_2$** features strong two-photon absorption and its quantum yield for the generation of H$^+$ per photo-excited molecule (ϕ_{H^+}) exceeds that of some commercial one-photon PAGs. Based on $\langle P \rangle_{\min}$, **BSB-S$_2$** was found to be 10 to 100 times more sensitive toward multiphoton epoxide polymerization than one-photon UV/blue-sensitive PAGs.

It should be understood that this discussion of multiphoton initiators is not comprehensive of all work that has been done in the field. Rather, it is intended as a highlight of the many outstanding examples of progress in this area. The use of inexpensive low-power lasers for 3DM should become more practical as high-sensitivity

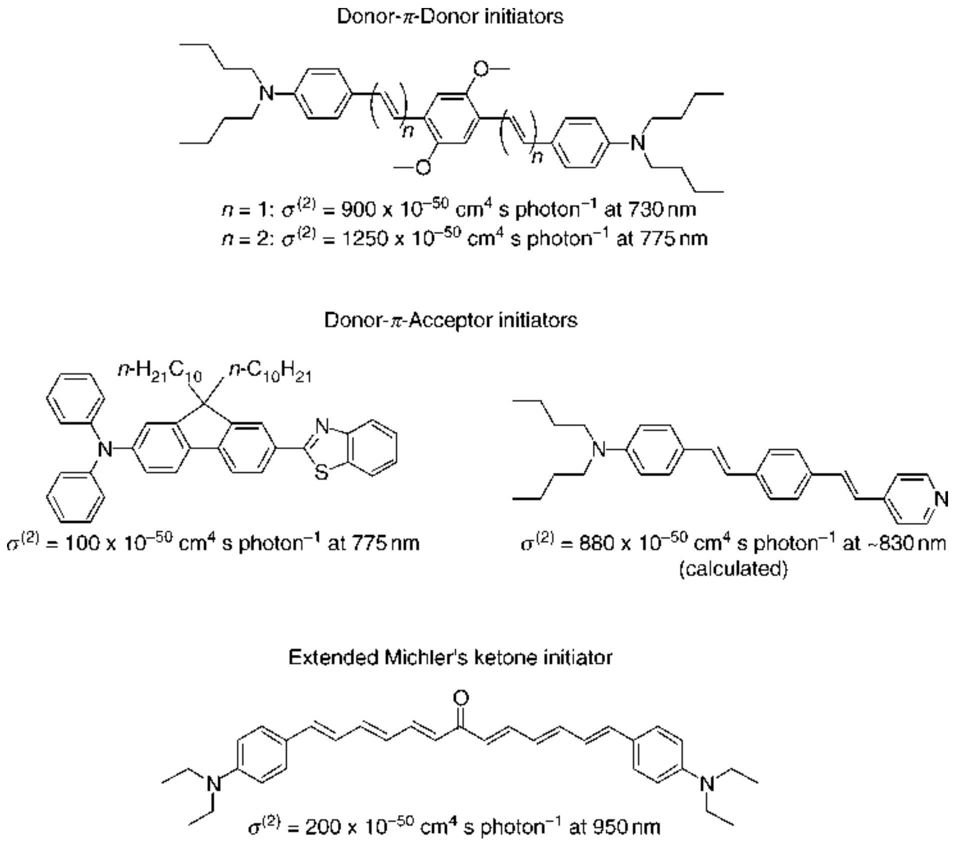
Donor-π-Donor initiators
n = 1: σ(2) = 900 x 10−50 cm4 s photon−1 at 730 nm
n = 2: σ(2) = 1250 x 10−50 cm4 s photon−1 at 775 nm
Donor-π-Acceptor initiators
n-H21C10
n-C10H21
σ(2) = 100 x 10−50 cm4 s photon−1 at 775 nm
σ(2) = 880 x 10−50 cm4 s photon−1 at ~830 nm (calculated)
Extended Michler's ketone initiator
σ(2) = 200 x 10−50 cm4 s photon−1 at 950 nm

Figure 7 Examples of high-sensitivity initiators for 3DM based on TPIP of acrylates.

3DM media are developed. This should, in turn, facilitate the widespread use of 3DM as a microfabrication tool.

Resolution

The resolution of single-beam 3DM is determined by the size of the smallest voxel that can be reliably produced. Both the optical parameters and the material response under MPE affect the size of the voxel. Given the wide range of physical and chemical mechanisms that can be involved, no single set of guidelines can be outlined regarding the effect of the material response on voxel size. Additionally, this aspect of 3DM remains relatively unexplored. However, as 3DM based on TPIP of acrylates has been examined in some detail, an overview of this material response provides some valuable insight into the overall problem of resolution. The following simplified discussion is based on the local response of the material to photo-excitation.

TPIP initiated in the vicinity of the focus could, in principle, propagate throughout the entire volume of the medium, were it not for termination by inhibitors. Let us consider further the expression for $N_{\mathrm{I}^\cdot}(\vec{r}, t)$, expanding the intensity as:

$$I(z, r, t') = i(z, r) f(t') \qquad [3]$$

Here, $f(t')$ is a time-dependent amplitude function, $i(z, r)$ is the peak intensity, and the position vector $\vec{r}$ is expressed in cylindrical coordinates (z, r) with the origin at the focus. As before, z denotes the direction in which the beam propagates (longitudinal). The radial coordinate orthogonal to the z-axis (lateral) is denoted by r. The beam is taken to have cylindrical symmetry, so I does not depend on the angular coordinate. For a fixed exposure time $t \rightarrow \tau_{\mathrm{ex}}$, we can identify a polymerization-threshold peak intensity, I_{th}, for which $i(z, r) > I_{\mathrm{th}}$ results in $N_{\mathrm{I}^\cdot}(z, r, \tau_{\mathrm{ex}}) > N_{\mathrm{Inh}}$ and polymerization occurs. The polymerized voxel consists then of all points within the volume centered at the focal point and bounded by the I_{th} isophote.

The shape of the I_{th} isophote depends upon the optical parameters of the system. The tightest focusing is achieved when the highest NA objective is used to focus parallel incident rays that overfill the entrance aperture. This situation generates a diffraction limited intensity pattern about the focal point. Within the focal plane the lateral intensity distribution follows the Airy pattern:

$$i(0, r) = [2J_1(\rho)/\rho]^2 I_0 \qquad [4]$$

where J_1 is the first-order Bessel function:

$$\rho = 2\pi r \mathrm{NA}/\lambda \qquad [5]$$

and I_0 is the peak intensity at the origin. Near the focus, the longitudinal intensity distribution is given by

$$i(z, 0) = [\sin(\zeta/4)/(\zeta/4)]^2 I_0 \qquad [6]$$

where

$$\zeta = 2\pi \mathrm{NA}^2 z/(n_0 \lambda) \qquad [7]$$

Let us define the longitudinal and lateral widths of the focal spot, Z and R respectively, to be twice the distance along the corresponding z- and r-axes over which the peak intensity of the central bright fringe decreases to $I_0/2$. If $\langle P \rangle$ is adjusted so that $I_0/2 = I_{\mathrm{th}}$, then to first order we can expect the polymerization voxel to be a solid of rotation centered about the focus having a longitudinal and lateral width of Z and R, respectively. Numerical evaluation shows that the $I_0/2$-points occur at $\rho = \pm 1.6154$ and $\zeta = \pm 5.5661$. For $\lambda = 800$ nm, NA = 1.4, and $n_0 \approx 1.5$, we find $R = 0.29$ μm and $Z = 1.1$ μm.

Several groups have demonstrated that voxels generated under conditions similar to those discussed above do indeed have dimensions comparable to those calculated, based on diffraction limited propagation. As shown in Figure 8, voxels with lateral and longitudinal widths below the diffraction limit can be fabricated by steadily reducing τ_{ex} (or I_0) so that an increasingly smaller portion of the central bright fringe exceeds I_{th}. In this way, voxels having $R \approx 100$ nm and $Z < 800$ nm have been generated, which convincingly demonstrates the potential of this technique as a nano-fabrication tool.

Post-exposure material processing can also have a profound impact on the resolution. In the case of TPIP-3DM, material near the boundary of the I_{th} isophote may not have sufficient molecular weight or cross-linking density to resist removal in a development step. In this case, the final voxel will be smaller than the I_{th} isophote defined strictly in terms of a polymerization threshold. Additionally, a solvent-based developer can swell or otherwise distort the voxels. Overall, three important factors determine the final size of the voxel: (i) the exposure and focusing conditions; (ii) the response of the material to photo-excitation; and (iii) the behavior of the photoconverted material during the development process. It should be emphasized that 3D confinement is due to the combined effects of nonlinear photo-excitation and the nonlinearity inherent in the photoconversion and post-exposure development of the material. More advanced treatments have also attempted to address the effects of radical and inhibitor diffusion on the size of the voxel.

Previously, we assumed that the voxel is cylindrically symmetric. The intriguing result of some recent work has shown that the widths of the voxel in the transverse directions can actually differ by as much as 10% when linearly polarized radiation is used because the symmetry of the electric field vectors in the x- and y-directions is broken in high-NA focusing geometries. Cylindrically symmetric voxels can be obtained by using depolarized or circularly polarized radiation. Clearly, study of the combined effects of optical parameters and material response remains an important area for further research in this rapidly advancing field. Although this discussion has centered on TPIP, similar arguments can be developed for 3DM based on other photomaterials and higher-order processes.

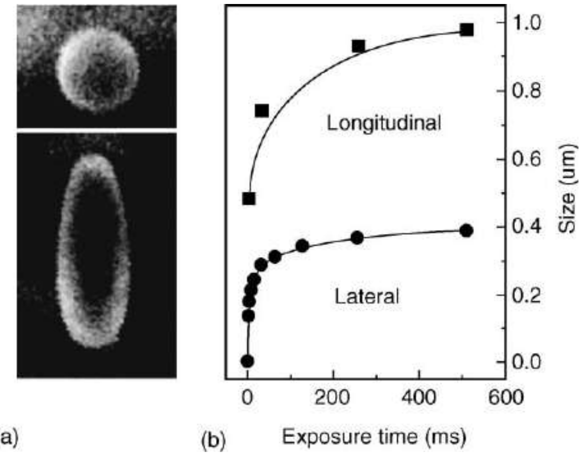

Figure 8 Single voxels produced by TPIP-3DM in an acrylate medium. (a) Scanning electron microscope (SEM) images of a single voxel showing its longitudinal (top) and lateral (bottom) width. (b) Dependence of the longitudinal and lateral widths of voxels on exposure time, as measured from SEM images following fabrication. Reproduced from Kawata S and Sun H-B (2003) Two-photon photopolymerization as a tool for making micro-devices. *Applied Surface Science* 208–209: 153–158. Copyright (2003), with permission of The American Institute of Physics.

Multibeam-Interference Three-Dimensional Microfabrication

Multibeam-interference (MBI) 3DM is an alternative approach to 3DM that enables large volumes of material to be simultaneously patterned into a 3D periodic structure. In MBI-3DM, multiple coherent laser beams interfere within a photo-active medium. A periodic structure results because the material undergoes photo-induced changes that occur maximally in the regions of peak intensity across the periodic interference pattern. Depending on the number of beams utilized and their reciprocal orientation, periodic structures in one, two, or three dimensions can be generated, and the periodicity is of the order of λ. This technique is often referred to as holographic lithography or recording.

MBI-3DM has most often been achieved using one-photon-induced polymerization; however, higher-order excitation may be used as well. In this discussion we will focus on multiphoton MBI-3DM, with the understanding that most of the concepts apply equally well to the case of linear photo-excitation. When the power of the incident beams is adjusted so that the intensity at the peaks of the interference pattern just exceeds I_{th}, polymerization only occurs in the local vicinity of these points, whereas the 3DM medium is not appreciably polymerized in regions where the intensity is below I_{th} and may be removed by dissolution in a solvent. The result is a periodic porous structure with highly interconnected parts, that extends over the whole volume where the beams overlapped.

Let us assume that m-beams are directed onto a photosensitive material. The intensity of light at a given point $\vec{r}$ in the region where the beams overlap is proportional to:

$$\begin{aligned} I(\vec{r}) \propto \vec{E}\cdot\vec{E}^* &= \left(\sum_{j=1}^{m} \vec{E}_{0j}\, \mathrm{e}^{\mathrm{i}(\vec{K}_j\cdot\vec{r}-\omega t)}\right)\cdot\left(\sum_{l=1}^{m} \vec{E}^*_{0l}\, \mathrm{e}^{-\mathrm{i}(\vec{K}_l\cdot\vec{r}-\omega t)}\right) \\ &\propto \sum_{j,l=1}^{m} \vec{E}_{0j}\cdot\vec{E}^*_{0l}\, \mathrm{e}^{\mathrm{i}(\vec{K}_j-\vec{K}_l)\cdot\vec{r}} \\ &\propto \sum_{j=1}^{m} \vec{E}_{0j}\cdot\vec{E}^*_{0j} + \sum_{j\neq l}^{m} \vec{E}_{0j}\cdot\vec{E}^*_{0l}\, \mathrm{e}^{\mathrm{i}(\vec{K}_j-\vec{K}_l)\cdot\vec{r}} \qquad [8] \end{aligned}$$

where $\vec{E}_{0j}$ and $\vec{K}_j$ are the electric field and wavevector of beam j, and ω is its angular frequency ($\omega = 2\pi\nu$). Equation [8] indicates that the intensity distribution depends on the difference between pairs of wavevectors of the interfering beams, $\vec{K}_j - \vec{K}_l$. This distribution is periodic in space and can be described as a crystal lattice with reciprocal lattice vectors equal to $\vec{K}_j - \vec{K}_l$ and with a unit cell structure determined by the pre-factors $\vec{E}_{0j}\cdot\vec{E}^*_{0l}$. By choosing the intensity of the incident beams and the exposure time appropriately, it is thus possible to generate a pattern in the medium that has the same periodicity and structure as the I_{th} isophote of the interference pattern.

The simplest implementation of MBI-3DM involves interfering two linearly polarized beams at angles $\pm\theta/2$ with respect to the normal to the film being patterned (θ is the angle between the two beams in vacuum). From eqn [8], it can be shown that the intensity is maximum if $(\vec{K}_1 - \vec{K}_2)\cdot\vec{r} = 2q\pi$, where q is an integer, or $K[\sin(\theta/2)]r' = q\pi$, where K is the magnitude of both $\vec{K}_1$ and $\vec{K}_2$, and r' is the component of $\vec{r}$ along $(\vec{K}_1 - \vec{K}_2)$. Thus, the two beams generate an interference pattern with periodicity (in one dimension) given by $\Lambda = \lambda/[2\sin(\theta/2)]$, if λ is the

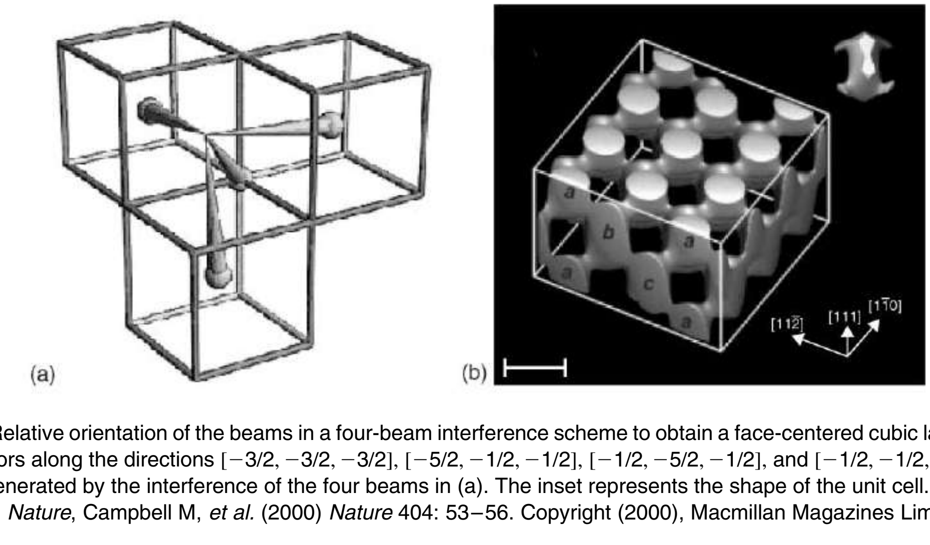

Figure 9 (a) Relative orientation of the beams in a four-beam interference scheme to obtain a face-centered cubic lattice. The beams have wave vectors along the directions [−3/2, −3/2, −3/2], [−5/2, −1/2, −1/2], [−1/2, −5/2, −1/2], and [−1/2, −1/2, −5/2]. (b) One of the isophotes generated by the interference of the four beams in (a). The inset represents the shape of the unit cell. Reproduced with permission from *Nature*, Campbell M, *et al.* (2000) *Nature* 404: 53–56. Copyright (2000), Macmillan Magazines Limited.

vacuum wavelength of the exciting beam ($\lambda = 2\pi/K$). When a film of a polymerizable material is placed at the intersection of the beams, a series of ridges is generated in the resist, spaced by Λ and with a width dependent on the beam intensity. For example, if a laser beam at 800 nm is split in two components forming an angle of 30°, $\Lambda = 1.5$ μm. The periodicity becomes smaller as the angle between the beams is increased.

By interfering three beams it is possible to obtain a two-dimensional periodic structure. For example, the interference of three beams forming an angle θ with respect to the medium normal and symmetrically arranged around it was used to generate an hexagonal lattice of rods that extends parallel to the normal. At least four beams must interfere to obtain a microstructure with 3D periodicity. The beam geometry shown in Figure 9a was used to produce a microstructure with the symmetry of a face-centered cubic lattice (Figure 9b). Beam geometries for generating interference patterns having other symmetries, such as body-centered cubic and body-centered tetragonal, have also been described.

An advantage of MBI-3DM is that it enables rapid patterning of large volumes during a single exposure. In fact, the polymerization takes place, in principle, throughout the region in which the beams overlap. When the process is activated by OPE, the beams can be attenuated significantly by absorption in the material. This can limit the thickness of the fabricated structure or result in spatial variation in the filling ratio. The use of TPIP can overcome this limitation and allow for thicker media to be patterned uniformly. Structures with the same symmetry are generated if either one- or two-photon processes are exploited to induce the material transformation. However, the structures will be different for at least two reasons: (i) if the same initiator is used in both cases, the wavelength used for TPE is approximately twice that for OPE, and

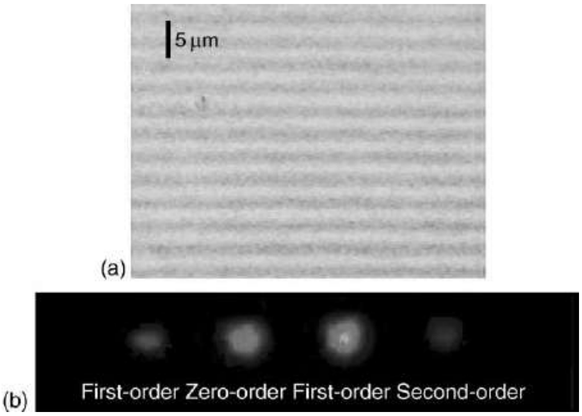

Figure 10 (a) Optical image of a diffraction grating obtained by TPIP using a two-beam interference scheme ($\lambda = 810$ nm). The spacing period is 3.4 μm and film thickness is 40 μm. The grating is made of an acrylate-based resin. (b) Diffraction pattern of the grating in (a) generated using a Helium–Neon laser (632.8 nm). The maximum diffraction efficiency was measured to be 57%. Reproduced from Guo H, *et al.* (2003) *Chemical Physics Letters* 374: 381–384. Copyright (2003), with permission from Elsevier.

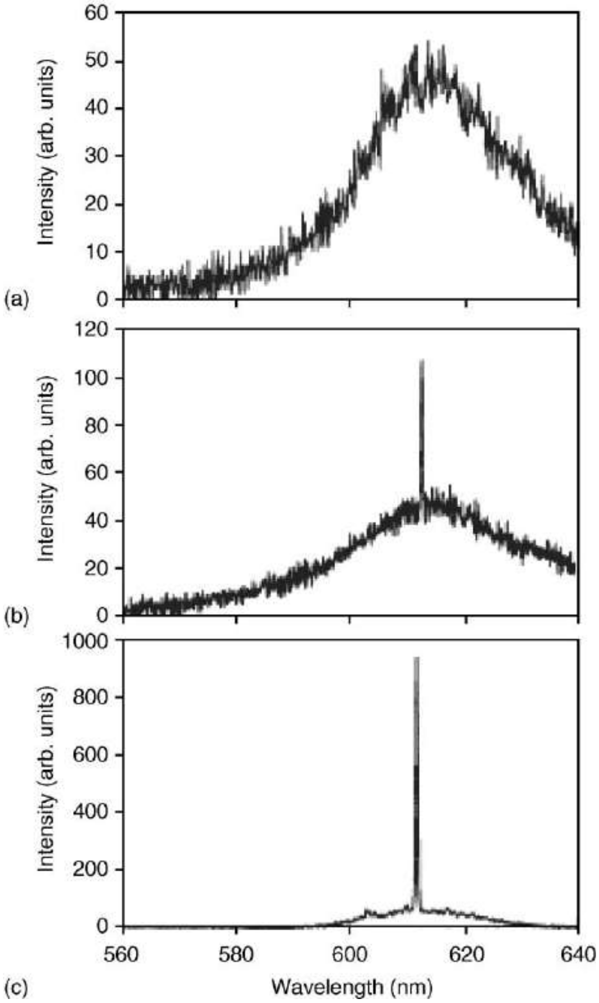

Figure 11 Emission spectra of the laser dye DCM encapsulated in a dendrimer-based resin that was patterned into a grating microstructure by single-beam 3DM. The grating was pumped at 532 nm (8 ns pulses) with excitation energies of (a) 0.03, (b) 0.05, and (c) 0.20 μJ. Starting at excitation energies as in (b), a narrow band at 612 nm appears over the spontaneous emission. The relative intensity of this lasing emission increases with excitation intensity with respect to the spontaneous emission. Reproduced from Yokoyama S, *et al.* (2003) *Thin Solid Films* 438–439: 452–456. Copyright (2003), with permission from Elsevier.

the periodicity of the structure is correspondingly different; and (ii) the I_{th} isophote, and thus the details of the polymerized structure inside each unit cell, is different under OPE and TPE because the excitation rate is proportional to I and I^2, respectively. In principle, better defined features can be obtained using two-photon or higher-order excitation because the transition between regions of low- and high-exposure is sharper than that under OPE, due to the nonlinear dependence on I. Interference fabrication techniques can enable the generation of features considerably smaller than the diffraction limit, when the beam geometry and excitation intensity are chosen appropriately.

The beams used to generate the interference pattern are usually obtained by splitting the output of a single laser source into the desired number of components. The intensity and polarization of each beam are then independently controlled. Temporal overlap of the beams at the sample can be ensured by introducing an optical delay into each of the beam paths. As fluctuations in the beam intensity can cause changes in the lattice unit cell or its filling ratio, the use of a laser with good stability is imperative in cases where exposure time extends over more than a single laser pulse.

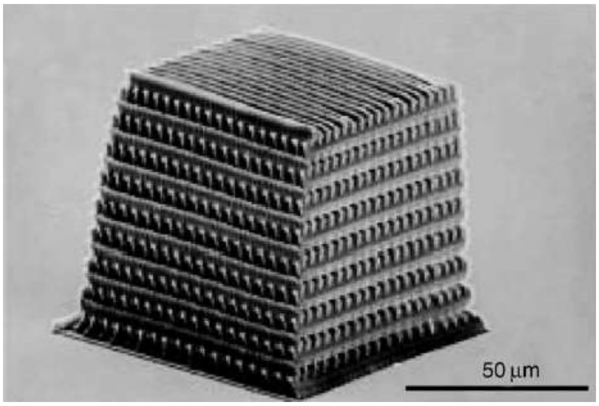

Figure 12 Example of a wood-pile PhC structure obtained by single-beam TPIP-3DM. The periodicity is 5 μm. The change in the cross-section of the structure from top to bottom, is due to shrinking of the material during development, whose effect is largest in the region farthest from the substrate. Reproduced with permission from *Nature*, Cumpston BH, Ananthavel SP, Barlow S, *et al.* (1999) Two-photon polymerization initiators for three-dimentional optical data storage and microfabrication. *Nature* 398: 51–54, Copyright (1999), Macmillan Magazines Limited.

Structures and Functional Devices

The versatility of 3DM is easily appreciated by surveying the variety of microstructures that have been fabricated over the years. These may be divided into two categories: (i) structures whose characteristics lay in their form and in the interconnectivity of their parts; and (ii) structures that incorporate an active functionality that can be exploited in a specific application. Some of these microstructures are described below.

Fabricating functional micro-optical devices is a promising application of 3DM. Figure 10 shows a grating fabricated by two-beam-interference TPIP in an acrylate with a period of 3.4 μm, whose diffraction efficiency is 28% as fabricated and 57% after annealing. The same method was used to generate a surface relief grating in a resin containing a polycationic peptide. In the presence of aqueous silicic acid, the peptide catalyzes the formation of ordered arrays of silica nanospheres in

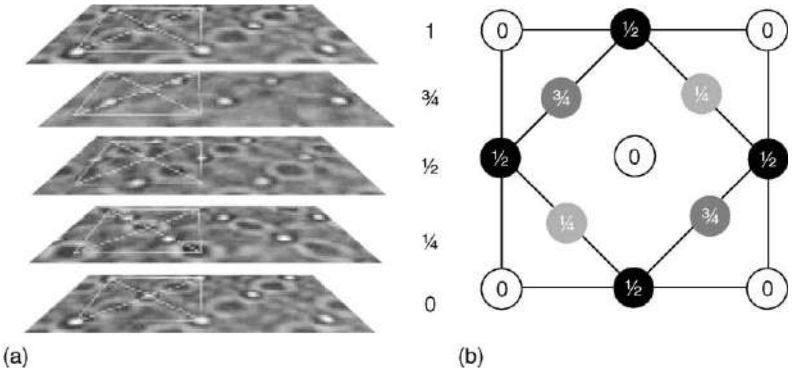

Figure 13 (a) Cross-sectional optical images of various planes of a diamond lattice fabricated in glass. The numbers to the right are the relative positions of the planes, in units of the lattice constant. (b) Diagram of the expected position of lattice points in a diamond structure. Points located in different planes have different shading. Reproduced from Mizeikis V, *et al.* (2001) *Journal of Photochemistry and Photobiology A* 145: 41–47. Copyright (2001), with permission from Elsevier.

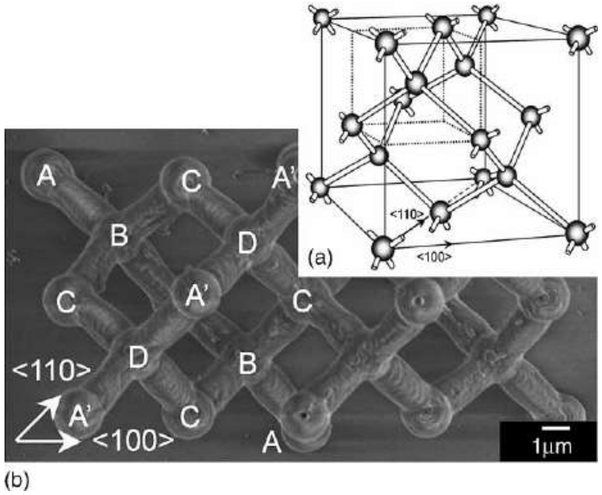

Figure 14 (a) Illustration of a diamond lattice. (b) Scanning electron micrograph of a polymeric microstructure with the symmetry of a diamond lattice, obtained by single-beam TPE-3DM. The letters A, B, C, D, A′ refer to the lattice planes at 0, 1/4, 1/2, 3/4, and 1 lattice units in the vertical direction. Reproduced from Kaneko K, *et al.* (2003) *Applied Physics Letters* 83: 2091–2093. Copyright (2003), with permission of The American Institute of Physics.

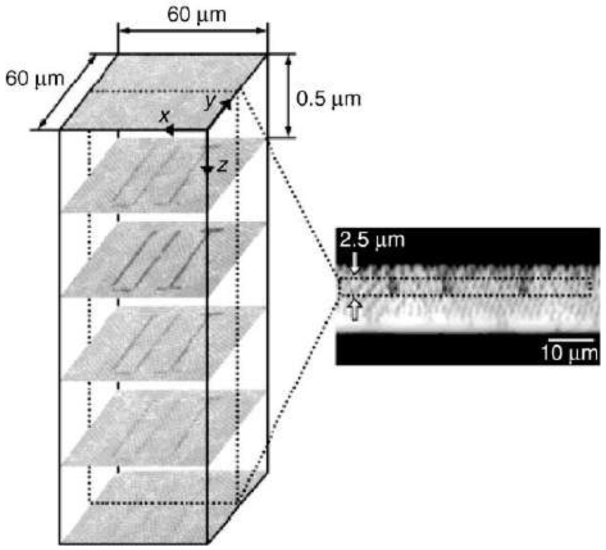

Figure 15 (Left) Images obtained by laser scanning confocal microscopy of linear defects generated by three-photon-induced polymerization of a triacrylate resin backfilled into a silica colloidal crystal. The spacing between the planes of the images is 0.5 μm. (Right) Cross-sectional image in the *xz*-plane of the same structure. Reproduced with permission from Lee W, *et al.* (2002) *Advanced Materials* 14: 271–274. Copyright (2002), with permission of Wiley.

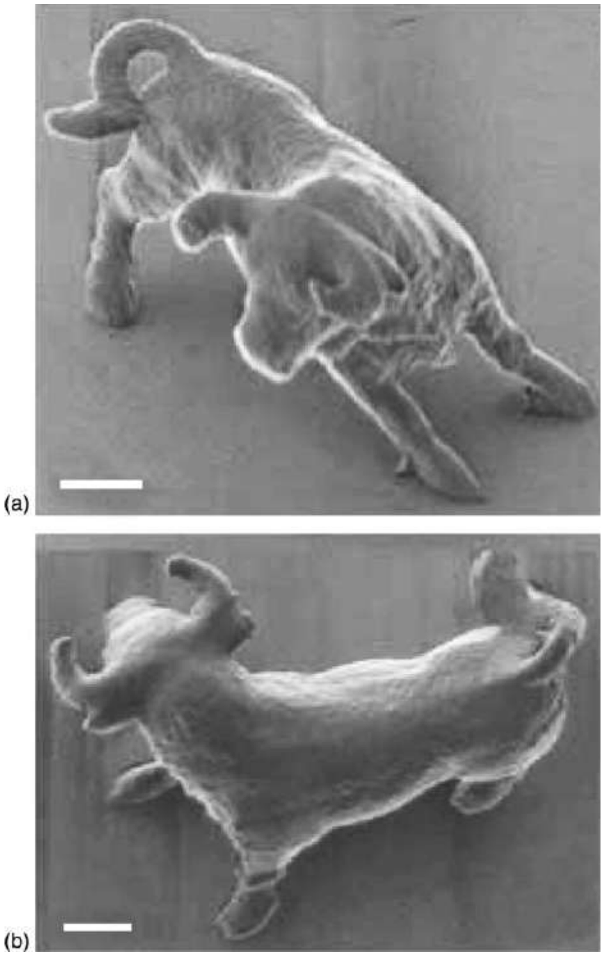

Figure 16 Microstructure of a bull fabricated by single-beam 3DM. The scale bar corresponds to 2 μm. Reproduced with permission from *Nature*, Kawata S, *et al.* (2001) *Nature* 412: 697–698. Copyright (2001), Macmillan Magazines Limited.

the troughs of the grating, which increases the diffraction efficiency 50-fold. 3DM was used to fabricate a grating structure in an acrylate resin containing the laser dye 4-(dicyanomethylene)-2-methyl-6-(4-dimethylaminostyryl)-4H-pyrane (DCM) encapsulated in a host dendrimer (see above) as an active optical medium. When pumped at 532 nm above a certain threshold, lasing emission from DCM was observed at 612 nm, with a bandwidth of 0.15 nm (Figure 11).

Several groups have explored the potential of 3DM as a tool for patterning waveguide-based optical circuitry. For example, a mixture of thermally and photocurable resins was used to generate embedded waveguides by photocross-linking within the thermally set medium. A Y-coupler was created in fused silica through the refractive index change generated by patterned MPE.

3DM provides a powerful route to 3D photonic crystals (PhC). PhCs are structures in which a basic motif, or ‘unit cell’, is repeated within the material with a given periodicity. PhCs have attracted great interest, because the modulation of the refractive

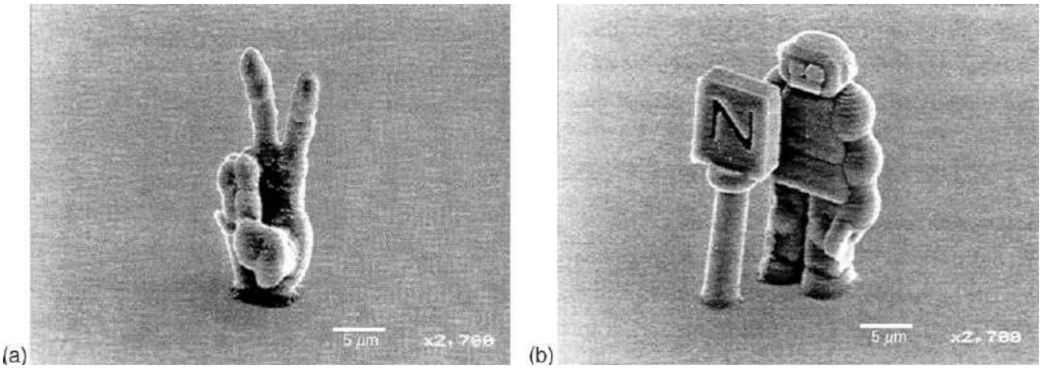

Figure 17 Scanning electron micrographs of two structures in the form of (a) a hand and (b) a robot which were fabricated by single-beam 3DM. Reproduced from Maruo S and Ikuta K (2000) *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 3937: 106–112. Copyright (2000), with permission of SPIE.

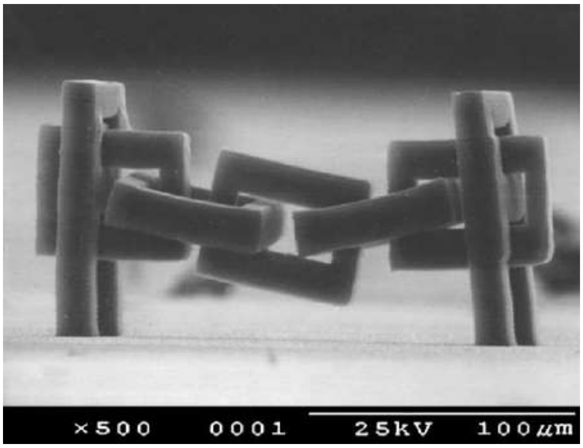

Figure 18 Scanning electron micrograph of a series of connected chain links obtained by single beam TPIP-3DM in an acrylate resin. Notice the tilt angle between the central links, indicating that each link can move with respect to the others. Reproduced from Kuebler SM, Rumi M, Watanabe T, *et al.* (2001) Optimizing two-photon initiators and exposure conditions for three-dimentional lithographic microfabrication. *Journal of Photopolymer Science and Technology* 14: 657–668. Copyright (2001), with permission of the Technical Association of Photopolymers Japan.

index between the ‘filled’ areas and the surrounding space provides a means for controlling the propagation of electromagnetic radiation within the structure. PhC with periodicities in the submicron and micron range exhibit their peculiar properties in the visible and near infrared spectral region. Optical wavelength PhCs can be fabricated in a straightforward way by 3DM, either by serial patterning using the single beam approach or by parallel patterning using the interference method, for which the number and orientation of the beams is dictated by the symmetry of the desired structure. Typical examples are woodpile structures (Figure 12), which consist of layers of rods of a material fabricated side by side with a fixed spacing, and alternate layers are rotated by 90° with respect to one another. If the rods in one layer are displaced by half the lattice constant with respect to the previous layer with the same orientation, the structure obtained has the symmetry of a face-center cubic lattice. PhCs with the symmetry of the diamond crystal have also been fabricated, either by inducing changes in a glass

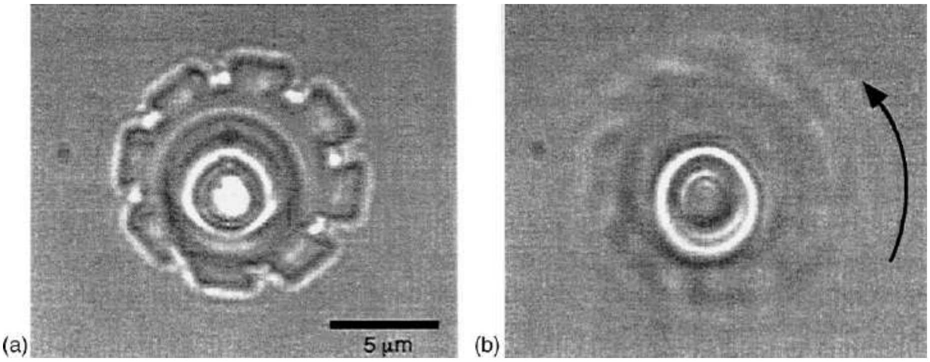

Figure 19 Optical image of a micro-wheel obtained by single-beam 3DM. (a) Structure before development. (b) When ethanol is used to remove the unpolymerized material, the wheel rotates around the axle in the solvent. Reproduced from Maruo S and Ikuta K (2000) *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 3937: 106–112. With permission of the Institute of Electrical Engineers of Japan.

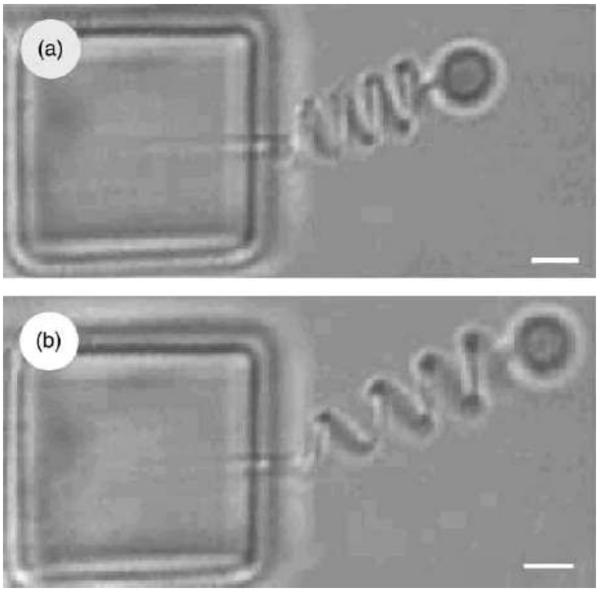

Figure 20 Optical image of a micro-spring tethered to a polymer block. The structure in (a) is at its resting point in ethanol. In (b) the spring was elongated by displacing the polymer bead at the free end using laser trapping. Reproduced with permission from *Nature*, Kawata S, *et al.* (2001) *Nature* 412: 697–698. Copyright (2001), Macmillan Magazines Limited.

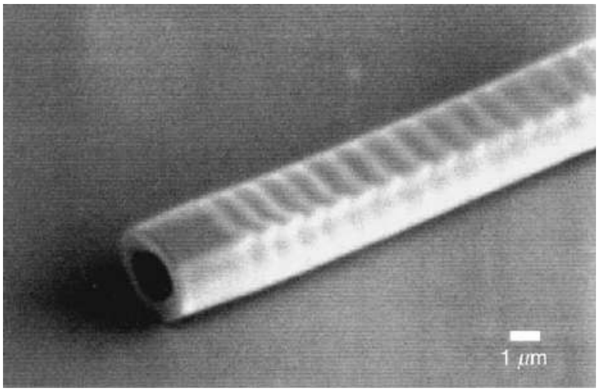

Figure 21 Hollow micro-tubule generated by single-beam TPIP-3DM in a urethane acrylate. The inner diameter is 1.8 μm. Reproduced from S Maruo and S Kawata (1998) *Journal of Microelectromechanical Systems* 7: 411–415. Copyright (1998), with permission of IEEE.

substrate at lattice points (Figure 13), or by constructing a network of polymeric rods connecting nearest neighbor points (Figure 14). It was showed that a PhC created by TPIP-3DM does exhibit a partial bandgap in the infrared spectral region. The position of the bandgap can be tuned by fabricating the structure with the appropriate periodicity.

The ability to introduce defects in PhCs in a controlled fashion is of key importance for the application of these structures in optical circuitry, as they can act as waveguides for channeling light into preferential directions. Point, line, and plane defects can be generated easily, by locally altering the fabrication pattern in the PhC. Point-by-point MPE of a Ge-doped glass was used to pattern a cubic lattice containing systematic defects that defined a Y-shaped waveguide splitter in a [100] plane. 3DM can also be used to modify a structure that is created using another fabrication method, and additional materials can be introduced into the structure in this second fabrication step. A silica colloidal crystal was backfilled with a polymerizable material and waveguides were generated by photocross-linking channels in the medium through three-photon-excitation 3DM (Figure 15).

The full potential of single-beam 3DM can be appreciated by surveying some of the structures that have been fabricated, featuring topologically complex shape, high interconnectivity, and extreme under-cut. Examples include those of Figures 16, 17, and the diamond-lattice PhC of Figure 14. Structures with freely movable parts, like those needed for MEMS, are difficult to obtain by other fabrication techniques, but can often be made by 3DM in a single exposure step. The ability to generate structures with independent but interconnected parts is exemplified by the microchain of Figure 18. A further example is the microgear in Figure 19, which is shown to spin around its axle when the sample is flooded with a liquid. Other functional microstructures have also been reported, such as the microspring in Figure 20, fabricated in a urethane acrylate resin and characterized by a core diameter of only 300 nm. When the sample is immersed in ethanol, the pendant mass can be captured and extended with optical tweezers. When the mass is released, the spring behaves like a damped oscillator.

Promising results have also been obtained in the production of structures to be used in micro-fluidics applications. Of key importance, in this case, is the ability to pattern channels, cavities, and other hollow structures that can be used to store or transfer small amounts of fluids between different parts of a device, as needed to perform, for example, analytical processes on a femtoliter scale. Examples include the hollow microtubule (Figure 21), and the submersed channels structure fabricated in a positive-tone resist (Figure 5).

List of Units and Nomenclature

A	Electron accepting molecular fragment
AFS	Amplified femtosecond
C	Chromophore
$\mathbf{C}^*$	Excited chromophore
CW	Continuous wave
D	Electron donating molecular fragment
3D	Three-dimensional
3DM	Three-dimensional microfabrication
$\vec{E}_0$	Spatial component of the electric field vector [V m^{-1}]
E_p	Laser pulse energy [J]
$f(t)$	Normalized temporal distribution of the beam intensity [dimensionless]
FOM	Figure of merit
h	Planck's constant [J s]
$i(z,r)$	Spatial distribution of the beam intensity [W cm^{-2}]
I	Intensity of a propagating beam [W cm^{-2}]
I_0	Maximum on-axis intensity for a Gaussian beam [W cm^{-2}]
I_th	Threshold intensity (minimum peak intensity at which polymerization occurs for a given exposure time) [W cm^{-2}]
I·	Initiating radical
j, l, q	Integers [dimensionless]
J_1	First order Bessel function [dimensionless]
K	Amplitude of a beam wave vector ($K = 2\pi/\lambda$) [m^{-1}]
$\vec{K}$	Wave vector of a beam [m^{-1}]
m	Number of beams in a multi-beam interference configuration [dimensionless]
M	Monomer
MBI	Multi-beam interference
MEMS	Micro-electromechanical systems
MPE	Multi-photon excitation
n	Number of photons involved in MPE [dimensionless]
n_0	Linear refractive index of the medium [dimensionless]
N	Number density (concentration) of a species [cm^{-3}]
NA	Numerical aperture of a lens [dimensionless]
N_C	Number density of chromophores present [cm^{-3}]
Nd:YAG	Neodymium:Yttrium aluminum garnet
$N_{\mathrm{I\cdot}}$	Number density of initiating radicals generated [cm^{-3}]
N_Inh	Number density of inhibitors present [cm^{-3}]
OPE	One-photon excitation
Ormocer	Organically modified ceramic
$\langle P\rangle$	Average laser power [W]
$\langle P\rangle_\mathrm{min}$	Minimum average laser power for patterning a material [W]
	Repetition rate (number of laser pulse per unit time) [Hz]
PAG	Photo-acid generator
PhC	Photonic crystal
r	Radial coordinate (distance from optical axis) [m]
r'	Component or $\vec{r}$ along a given wave vector [m]
$\vec{r}$	Position vector in three-dimensional space [m]
R	Twice the radial coordinate of the point at which the intensity is $I_0/2$ [m]
$R_{1h\nu}$	Molecular one-photon excitation rate (number of molecules excited per unit volume and time via OPE) [cm^{-3} s^{-1}]
$R_{2h\nu}$	Molecular two-photon excitation rate (number of molecules excited per unit volume and time via TPE) [cm^{-3} s^{-1}]
$R_{nh\nu}$	Molecular n-photon excitation rate (number of molecules excited per unit volume and time via n-order MPE) [cm^{-3} s^{-1}]
RMS	Root-mean-square
t	Time [s]
TEM$_{00}$	Lowest order transverse electromagnetic mode of a laser
Ti:S	Titanium:sapphire
TPE	Two-photon excitation
TPIP	Two-photon induced polymerization
z	Coordinate along the beam propagation direction [m]
z_R	Rayleigh range of a Gaussian beam; axial distance between the point of minimum beam waist and the point at which the waist has increased by a factor of $\sqrt{2}$ [m]
Z	Distance between the points of intensity $I_0/2$ along the coordinate z [m]
ζ	Longitudinal optical coordinate ($\zeta = 2\pi Z\,(\mathrm{NA})^2/n_0\lambda$) [dimensionless]
θ	Angle between a beam wave vector and the normal to a substrate [rad]
λ	Wavelength (in vacuum) [m]
Λ	Period of a one-dimensional grating [m]

ν	Frequency [s^{-1}]
$\boldsymbol{\pi}$	π-conjugated molecular fragment
ρ	Transverse optical coordinate ($\rho = 2\pi r\mathrm{NA}/\lambda$) [dimensionless]
$\sigma^{(1)}$	One-photon absorption cross-section [cm^2]
$\sigma^{(2)}$	Two-photon absorption cross-section [$\mathrm{cm}^4\ \mathrm{s\ photon}^{-1}$]
$\sigma^{(n)}$	n-photon absorption cross-section [$\mathrm{cm}^{2n}\ \mathrm{s}^{n-1}\ \mathrm{photon}^{-(n-1)}$]
τ_{ex}	Exposure time [s]
τ_{p}	Laser pulse duration [s]
$\phi_{\mathrm{c}}^{(n)}$	n-photon chemical quantum yield [(# of chemical species generated)/(# of n-photon-excited species)] [dimensionless]
$\phi_{\mathrm{H+}}$	Chemical yield for the generation of H^+ (number of H^+ generated per excited molecule) [dimensionless]
ω	Angular frequency ($\omega = 2\pi\nu$) [Hz]

See also

Chemical Applications of Lasers: Nonlinear Spectroscopies. **Diffractive Systems:** Applications of Diffractive and Micro-Optics in Lithography; Design and Fabrication of Diffractive Optical Elements. **Holography, Techniques:** Overview. **Imaging:** Volume Holographic Imaging. **Laser-Induced Damage of Optical Materials**. **Materials for Nonlinear Optics:** Organic Nonlinear Materials. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Nonlinear Microscopy; Overview. **Photonic Crystals:** Electromagnetic Theory; Nonlinear Optics in Photonic Crystal Fibers; Self-Assembled and Functionalized Photonic Crystals. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Belfield KD, Schafer KJ, Liu YU, Liu J, Ren XB and van Stryland EW (2000) Multiphoton-absorbing organic materials for microfabrication, emerging optical applications and non-destructive three-dimensional imaging. *Journal of Physical Organic Chemistry* 13: 837–849.

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. Cambridge, UK: Cambridge University Press.

Cumpston BH, Ananthavel SP, Barlow S, *et al.* (1999) Two-photon polymerization initiators for three-dimensional optical data storage and microfabrication. *Nature* 398: 51–54.

Joshi MP, Pudavar HE, Swiatkiewicz J, Prasad PN and Reinhardt BA (1999) Three-dimensional optical circuitry using two-photon-assisted polymerization. *Applied Physics Letters* 74: 170–172.

Kawata S and Sun H-B (2003) Two-photon photopolymerization as a tool for making micro-devices. *Applied Surface Science* 208–209: 153–158.

Kuebler SM, Rumi M, Watanabe T, *et al.* (2001) Optimizing two-photon initiators and exposure conditions for three-dimensional lithographic microfabrication. *Journal of Photopolymer Science and Technology* 14: 657–668.

Kuebler SM, Braun KL, Zhou W, *et al.* (2003) Design and application of high-sensitivity two-photon initiators for three-dimensional microfabrication. *Journal of Photochemistry and Photobiology A: Chemistry* 158: 163–170.

Martineau C, Anémian R, Andraud C, Wang I, Bouriau M and Baldeck PL (2002) Efficient initiators for two-photon induced polymerization in the visible range. *Chemical Physics Letters* 362: 291–295.

Odian G (1981) *Principles of Polymerization*, 2nd edn. New York: Wiley.

Pitts JD, Campagnola PJ, Epling GA and Goodman SL (2000) Submicron multiphoton free-form fabrication of proteins and polymers: studies of reaction efficiencies and applications in sustained release. *Macromolecules* 33: 1514–1523.

Ren Y, Yu X-Q, Zhang D-J, *et al.* (2002) Synthesis, structure and properties of a new two-photon photopolymerization initiator. *Journal of Materials Chemistry* 12: 3431–3437.

Strickler JH and Webb WW (1990) Two-photon excitation in laser scanning fluorescence microscopy. *Proceedings of the Society of Photo-Optical Instrumental Engineers* 1398: 107–118.

Sun H-B and Kawata S (2003) Two-photon laser precision microfabrication and its applications to micro-nano devices and systems. *Journal of Lightwave Technology* 21: 624–633.

Sun H-B, Maeda M, Takada K, Chon JWM, Gu M and Kawata S (2003) Experimental investigation of single voxels for laser nanofabrication via two-photon photopolymerization. *Applied Physics Letters* 83: 819–821.

Wu ES, Strickler JH, Harrell WR and Webb WW (1992) Two-photon lithography for microelectronic application. *Proceedings of the Society of Photo-Optical Instrumental Engineering* 1674: 776–782.

Wu P-W, Cheng W, Martini IB, Dunn B, Schwartz BJ and Yablonovitch E (2000) Two-photon photographic production of three-dimensional metallic structures within a dielectric matrix. *Advanced Materials* 12: 1438–1441.

Xu C and Webb WW (1997) Multiphoton excitation of molecular fluorophores and nonlinear laser microscopy. In: Lakowicz JR (ed.) *Topics in Fluorescence Spectroscopy*, vol. 5, pp. 471–540. New York: Plenum.

NONLINEAR OPTICS, BASICS

Contents

Cascading
$\chi^{(2)}$–Harmonic Generation
$\chi^{(3)}$–Third-Harmonic Generation
Four-Wave Mixing
Kramers–Krönig Relations in Nonlinear Optics
Nomenclature and Units
Nonlinear Optical Phase Conjugation
Photorefraction
Ultrafast and Intense-Field Nonlinear Optics

Cascading

G Assanto, University 'Roma Tre', Rome, Italy

G I Stegeman, University of Central Florida, Orlando, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Historically the field of nonlinear optics has evolved along two main axes, depending on the order of the nonlinearity used. The phenomena associated with the second-order nonlinearity $\chi^{(2)}$ have dealt with frequency conversion requiring wavevector matching (i.e., momentum conservation), such as second harmonic and optical parametric generation. On the other hand, effects associated with the third-order susceptibility $\chi^{(3)}$ have focused on inherently phase-matched phenomena at a single frequency, for example an intensity-dependent refractive index or degenerate four-wave mixing.

In cascading, multiple sequential nonlinear events based on $\chi^{(2)}$ can mimic some well-known $\chi^{(3)}$ phenomena. For example, it leads to nonlinear phase shifts $\Delta\phi^{\mathrm{NL}}$ reminiscent of those obtained through self-phase modulation over a distance L, i.e.,

$$\Delta\phi^{\mathrm{NL}}(L, I) = k_0 L n_2 I \qquad [1]$$

where I is the irradiance, k_0 the vacuum wavevector, and n_2 the Kerr coefficient or nonlinear refractive index. These can be used for all-optical switching, the generation of spatio-temporal solitons, etc. Some of these cascading effects have been known from the very earliest days of nonlinear optics, but their effectiveness was not realized until recently, when it proved possible to phase match second-order interactions with usefully large nonlinearities. Note that the medium must be $\chi^{(2)}$-active, i.e., noncentrosymmetric, which usually requires crystals rather than amorphous materials (such as glass).

Here we review the basic concepts, the characteristic interactions, some of the numbers involved and, finally, a few applications.

Physics of Cascading

Nonlinear Phase Shift in Collinear Second Harmonic Generation

The most important cascading effect, the nonlinear phase shift, is rooted in the fundamentals of the parametric interaction when two or three waves of different frequencies ($\omega_3 = \omega_1 \pm \omega_2$) are coupled by the second-order susceptibility $\chi^{(2)}(-\omega_3;\ \omega_1,\ \pm\omega_2)$. The simplest case is 'Type I' second harmonic generation (SHG), in which one eigenwave at the fundamental (FF) and one at the harmonic frequency (SH) interact with $\omega_1 = \omega_2 = \frac{1}{2}\omega_3$. When the waves have different phase velocities, this process involves simultaneously up-conversion ($\omega_1 + \omega_1 \rightarrow \omega_3$) and down-conversion ($\omega_3 - \omega_1 \rightarrow \omega_1$). By the latter the FF is regenerated from the harmonic after a characteristic propagation distance called the coherence length. This phenomenon is cyclic (see Figure 1a).

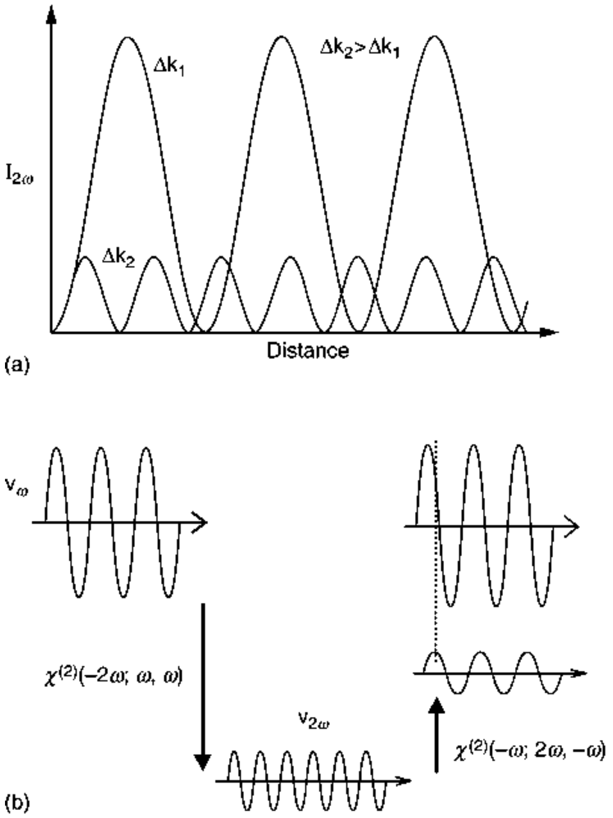

Figure 1 Phase-mismatched SHG. (a) Typical evolution of the generated second harmonic irradiance versus propagation for different mismatches. The period is twice the coherence length. (b) Intuitive picture of phase shift through cascading: the fraction of the initial energy which is first up-converted (SHG) and later down-converted (DFG), having traveled at a different phase velocity due to the wavevector mismatch, recombines with the through-FF causing an overall phase shift.

The field amplitudes in normalized form are

$$E_1(z,t) = \frac{1}{2}\sqrt{\frac{2}{c\varepsilon_0 n_1}}\, a_1(z)\mathrm{e}^{\mathrm{i}(\omega_1 t - k_1 z)} + \mathrm{c.c.}$$
$$E_3(z,t) = \frac{1}{2}\sqrt{\frac{2}{c\varepsilon_0 n_3}}\, a_3(z)\mathrm{e}^{\mathrm{i}(\omega_3 t - k_3 z)} + \mathrm{c.c.} \qquad [2]$$

where $|a_1|^2$ and $|a_3|^2$ are power densities, c is the speed of light in vacuum, and n_j and k_j are the refractive indices and wavevectors at angular frequencies ω_j $(j = 1, 3)$, respectively. As is common in nonlinear optics, the efficiency of these processes depends on the irradiance of the different waves, and on the wavevector mismatch between them, i.e., $\Delta k = 2k_1 - k_3$ for collinear beams. Since the phase velocities of the two waves, $v_{\omega_1} = \omega_1/k_1$ and $v_{\omega_3} = \omega_3/k_3$, are generally different and $\Delta k \neq 0$, their phase fronts fall out of synchronism during propagation. This leads to a periodic oscillation with distance (coherence length $= \frac{1}{2}$period $= \pi/|\Delta k|$) in the irradiances of the fundamental and harmonic waves, conserving the total energy but giving rise to a cascading phase shift, as pictured in Figure 1b.

The equations that describe the evolution of the fields stem from coupled mode theory, and are given by

$$\frac{\mathrm{d}}{\mathrm{d}z}a_3(z) = -\mathrm{i}\Gamma a_1^2(z)\mathrm{e}^{-\mathrm{i}\Delta k z}$$
$$\frac{\mathrm{d}}{\mathrm{d}z}a_1(z) = -\mathrm{i}\Gamma a_3(z)a_1^*(z)\mathrm{e}^{\mathrm{i}\Delta k z} \qquad [3]$$

where

$$\Gamma = \frac{\omega_1 \chi_{\mathrm{eff}}^{(2)}}{\sqrt{2n^3c^3\varepsilon_0}} \qquad [4]$$

$\chi_{\mathrm{eff}}^{(2)}$ is the 'effective' nonlinearity (for a specific crystal orientation and field polarization) and we assumed $n = n_1 \cong n_3$. In the limit of a large mismatch, the energy conversion to the harmonic is negligibly small, and it is possible to find an analytical expression for an effective 'n_2', $n_{2,\mathrm{eff}}$. In this limit $|a_1(z)| = |a_1(0)| =$ constant, and, in the absence of loss, solving for the small SH wave and substituting back into the equation for $\mathrm{d}a_1(z)/\mathrm{d}z$ gives

$$\frac{\mathrm{d}}{\mathrm{d}z}a_1(z) = -\mathrm{i}\frac{\Gamma^2}{\Delta k}\Big\{1 - \cos(\Delta k z) - \mathrm{i}\sin(\Delta k z)\Big\}|a_1(0)|^2 a_1(z) \qquad [5]$$

By comparing this with the corresponding equation based on n_2 $(\propto \chi^{(3)}[-\omega; \omega, -\omega, \omega])$ and describing self-phase modulation, namely

$$\frac{\mathrm{d}}{\mathrm{d}z}a_1(z) = -\mathrm{i}n_2\frac{\omega_1}{c}|a_1(z)|^2 a_1(z) \qquad [6]$$

we can define an equivalent nonlinear refractive index

$$n_2(z) = \frac{\pi\{\chi_{\mathrm{eff}}^{(2)}\}^2}{2\varepsilon_0 c n^4 \lambda_1 \Delta k}[1 - \cos(\Delta k z)] \qquad [7]$$

which turns out to be strictly nonlocal because it relies on propagation. Since the $\cos(\Delta k z)$ term oscillates rapidly in space, the above averages to

$$n_{2,\mathrm{eff}} = \frac{\pi\{\chi_{\mathrm{eff}}^{(2)}\}^2}{2\varepsilon_0 c n^4 \lambda_1 \Delta k} \qquad [8]$$

In this so-called 'Kerr limit' cascading leads to a Kerr-like nonlinear response, with either a self-focusing ($\Delta k > 0$, $n_{2,\mathrm{eff}} > 0$) or a self-defocusing nonlinearity, depending on the sign of Δk.

It is worthwhile to link the above equations to the physics of cascading with a simple argument. The evolution of the waves on propagation depends on their fields (in both amplitude and phase) at the input. When only the fundamental beam at $\omega = \omega_1$ is incident (the classical case of SHG), the harmonic is initially up-converted $\pi/2$ out of phase with the input (first of eqns [3]) via $\chi^{(2)}(-2\omega; \omega, \omega)$. If $\Delta k = 0$, the phase fronts travel at the same velocity and the relative phase is locked in at $\pi/2$. Difference frequency generation (DFG, second of eqns [3]) results in a down-converted fundamental which is an additional $\pi/2$ out of phase, i.e., exactly out of phase with the original input field. This leads to the monotonic depletion of the input and the growth of the SH with distance. If $\Delta k \neq 0$, there is only partial and periodic conversion to the harmonic, the wavefronts travel at different velocities (v_ω and $v_{2\omega}$, respectively), and the relative phase between the two waves changes with distance. As a result, the down-converted FF is no longer π out of phase with the input, thereby changing the overall phase of the total input-frequency wave by a cascaded amount $\Delta\phi^{\mathrm{NL}}$. This is the mechanism sketched in Figure 1b and buried in the better-known Maker-fringe effect represented in Figure 1a. Its magnitude depends on the irradiance and the mismatch, and its sign on that of Δk (see eqn [8]). The higher the input irradiance, the larger the harmonic conversion, and hence the larger $\Delta\phi^{\mathrm{NL}}$.

Alternatively, with a harmonic input at high irradiance and a weak seed at the FF, parametric amplification makes the fundamental grow via down-conversion. If $\Delta k = k_3 - 2k_1 \neq 0$, self-focusing (-defocusing) is obtained on the strong harmonic for $\Delta k > 0$ ($\Delta k < 0$). These phenomena can be generalized to the nondegenerate cases of sum and difference frequency generation, keeping in mind that when using three input waves their mutual phase relationship is essential in determining the details of the interaction.

The solutions to eqns [3] can be obtained numerically or written in terms of Jacobi elliptic integrals subject to boundary conditions ($a_3(0) = 0$, $a_1(0) \neq 0$) at $z = 0$, providing additional insight into the cascading phenomenon. In contrast to the Kerr case given by eqn [1], the variation in n_2 with distance implies a staircase-like increase in $\Delta\phi^{\mathrm{NL}}$ with distance for various detunings ΔkL, as shown in Figure 2a. This is consistent with our intuitive picture above, the increase in $\Delta\phi^{\mathrm{NL}}$ being maximized where the fundamental is the smallest (see Figure 2b for the oscillations in the fundamental amplitude), with the step-period corresponding to twice the coherence length. Conversely, for a fixed input irradiance, the evolution of $\Delta\phi^{\mathrm{NL}}$ exhibits two extrema with detuning ΔkL, as shown in Figure 3a. Note also that, for small ΔkL and large

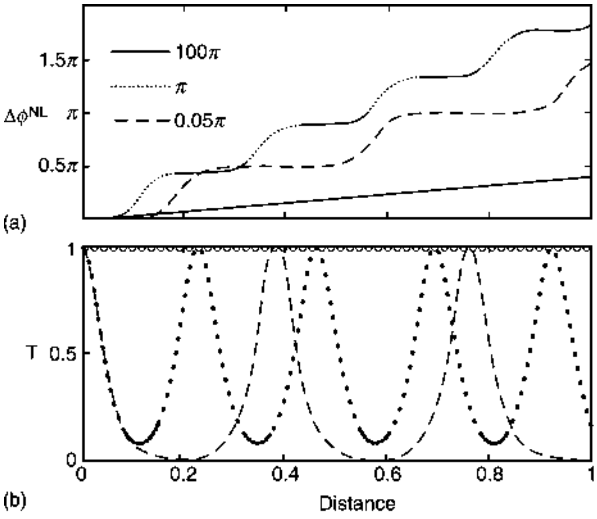

Figure 2 Calculated nonlinear phase (a) and throughput (b) of the fundamental field versus normalized propagation distance, for three values of phase mismatch.

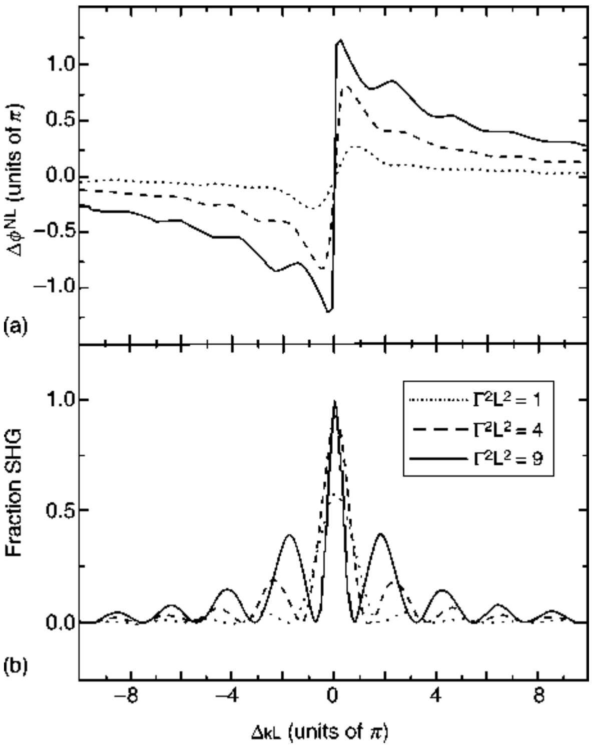

Figure 3 Calculated FF phase shift (a) and generated SH fraction (b) versus phase mismatch, i.e., the detuning curve, for three different nonlinear strengths (unitary FF irradiance).

(unrealistically high) irradiances, the increase in $\Delta\phi^{\mathrm{NL}}$ tends to become proportional to the field amplitude, a clear indication that cascading depends on a different order of nonlinearity than $\chi^{(3)}$, namely $\chi^{(2)}$. On the other hand, for large ΔkL the growth of $\Delta\phi^{\mathrm{NL}}$ is slower but linear with distance and irradiance, as in eqn [1].

There are trade-offs between the fundamental output and the net nonlinear phase shift. This is clearly visible in Figures 3a,b: the maxima in $\Delta\phi^{\mathrm{NL}}$ correspond to a reduced FF throughput. Although it is possible to operate cw at the zeros of the SHG, their location changes with the irradiances required for useful phase shifts ($>\pi$), complicating the trade-offs and the response to temporal pulses (encompassing a continuous irradiance distribution). The problem can be alleviated at the expense of reducing the net phase shift for a given irradiance and propagation length, by introducing regions with a nonuniform matching condition at both ends of the sample. The goal at the input end is to first adiabatically generate the SH required for optimum $\Delta\phi^{\mathrm{NL}}$. At the output, this harmonic is adiabatically converted back to the FF. The corresponding output for wavelengths larger than the phase-matching condition exhibits a $\Delta\phi^{\mathrm{NL}}$ with negligible fundamental depletion. This approach has been used in temperature-tuned and quasi phase matched (QPM, where a periodic modulation of the nonlinearity provides phase matching via the additional grating momentum) lithium niobate waveguides to generate phase shifts of π with less than 10% net loss of the fundamental wave to the harmonic.

While the preceding discussion has focused on Type I SHG, it is also valid for Type II SHG (i.e., with two FF input waves), provided the two fundamental inputs have equal input energy. In fact, an input imbalance between the two fundamental polarizations could give rise to a phase shift on the weaker field even when $\Delta k = 0$, leading to all-optical modulation and transistor action, as demonstrated in potassium titanyl phosphate (KTP) crystals.

Finally, due to the nature of the quadratic nonlinearity, the phase and/or amplitude of a coherent SH seed can have a large effect on the cascaded outcome, both in the phase and amplitude of the fundamental, as shown in Figure 4 for a fixed irradiance seed 1000 times weaker than the FF input. This can be used in seed-controlled optical processors.

Frequency Shifting

In the nondegenerate case of $\chi^{(2)}$ sum and difference frequency generation, two successive operations can perform the useful function of a frequency shifter (Figure 5). The idea is to shift the signal frequency ($\omega_{\mathrm{p}} - \delta$) by first generating from the pump (ω_{p}) a second harmonic ($2\omega_{\mathrm{p}}$), followed by DFG involving the latter and the signal. This leads to an output at $\omega_{\mathrm{p}} + \delta$ through $\chi^{(2)}(-(\omega_{\mathrm{p}} + \delta); 2\omega_{\mathrm{p}}, -(\omega_{\mathrm{p}} - \delta))$, i.e., a frequency shift of 2δ with an output spectrally mirror-imaged about the pump frequency. This feature is useful in WDM (wavelength division multiplexing) systems, and guarantees a protocol-insensitive wavelength shift with an additional phase conjugation useful in dispersion-compensating schemes. Notice that, since $\omega_{\mathrm{p}} + \delta \cong \omega_{\mathrm{p}} - \delta$ and $\delta << \omega_{\mathrm{p}}$, the bandwidth is large because the interaction is near the degeneracy point.

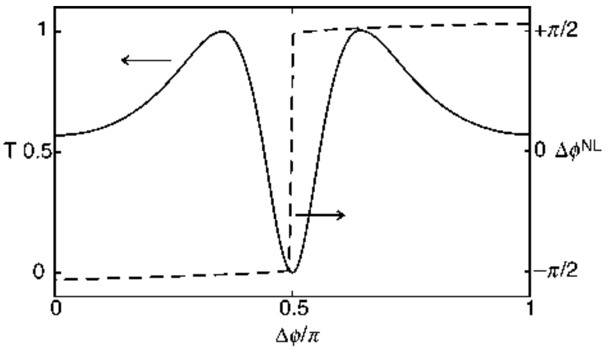

Figure 4 FF transmission (solid line) and phase shift (dashed line) versus the relative phase of a much weaker (three orders of magnitude) SH seed. Here perfect phase matching was assumed.

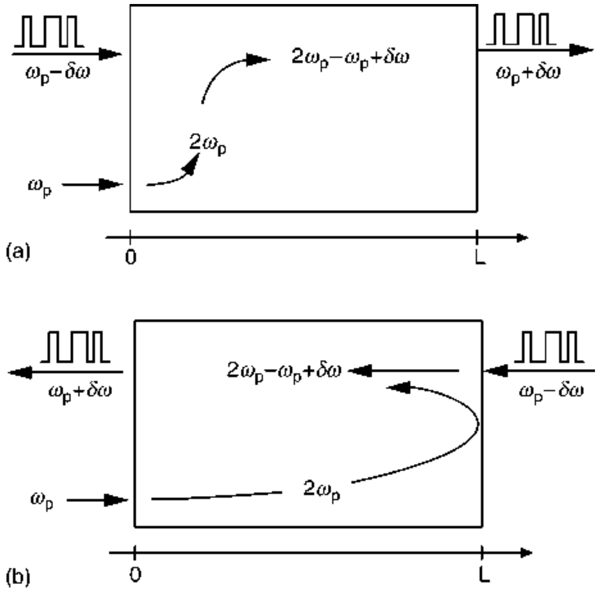
$\omega_{\mathrm{p}}-\delta\omega$
$2\omega_{\mathrm{p}}-\omega_{\mathrm{p}}+\delta\omega$
$\omega_{\mathrm{p}}+\delta\omega$
$2\omega_{\mathrm{p}}$
ω_{p}
(a)
0
L
$\omega_{\mathrm{p}}+\delta\omega$
$2\omega_{\mathrm{p}}-\omega_{\mathrm{p}}+\delta\omega$
$\omega_{\mathrm{p}}-\delta\omega$
ω_{p}
$2\omega_{\mathrm{p}}$
(b)
0
L

Figure 5 Wavelength shifter: (a) single-pass and (b) double-pass configurations. In (b) a dichroic mirror reflects the second harmonic of the pump.

Figure 5 shows two possible configurations for the same device length. The efficiency of the process can be increased in a double-pass geometry by using a back-propagating signal and a mirror-coated facet to reflect the SH, as sketched in Figure 5b. The best results to date were obtained at communication wavelengths with single-pass waveguides in periodically poled (QPM) $LiNbO_3$, including shift of multiple wavelength-channels with low cross-talk.

Multiple Beam Generation

Two successive $\chi^{(2)}$ processes can also be used to mimic three-wave mixing, a classic noncollinear $\chi^{(3)}$ interaction where two photons are taken out of one beam and one out of the second. In cascading, first a second harmonic is generated by one of the inputs, and then difference frequency mixing (DFM) between that SH and the second input leads to the signal of interest. In the simplest configuration, two equifrequency (ω) beams of wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ are incident in slightly different directions close to phase-matching and overlap inside the sample. When wave 1 is doubled (2ω) and wave 2 mixes with this harmonic via DFM, a new beam (at $2\omega - \omega \rightarrow \omega$) is generated with wavevector $2\mathbf{k}_1 - \mathbf{k}_2$. Of course the complementary output at $2\mathbf{k}_2 - \mathbf{k}_1$ is also generated via the SHG of wave 2. If these two new beams are strong enough, they can generate another pair via cascading, and so on. The generation of more than six extra beams has been demonstrated in β-barium borate with 1064 nm inputs.

This interaction can be applied to the control of the frequency chirp in pulses. For example, if wave 1 has a frequency chirp α_1 and wave 2 a chirp α_2, then the chirp of the beam scattered into the direction $2\mathbf{k}_1 - \mathbf{k}_2$ is $2\alpha_1 - \alpha_2$. This becomes especially interesting if $\alpha_2 = -\alpha_1$ so that the chirp in the new beam is $3\alpha_1$, as observed and subsequently employed in enhancing pulse compression. Note that, since a cascaded phase shift occurs on the down-conversion step back to the original fundamental (wave 1), it is not present in the scattered beam resulting in a clean transfer of the chirp.

Typical Numbers

From eqn [8] it is clear that the figure of merit (FOM) for cascaded processes is the same as for all second-order processes, namely $|\chi^{(2)}|^2/n^4$. The key is to find materials phase-matchable at the wavelengths of interest. Therefore, basically any material that is suited for frequency conversion, either SHG or parametric generation, will work. The most versatile material, which also has the largest easily usable nonlinear coefficient (18 pm/V), is lithium niobate when QPM is employed. As a reference, Table 1 gives the maximum $n_{2,\mathrm{eff}}$ for 1-cm-long samples, obtained by extrapolating eqn [8] to the ΔkL which maximizes the nonlinear phase shift. This is not strictly valid for such small ΔkL, but it is a useful estimate.

Other Implications of Cascading

A straightforward conclusion that could be drawn is that cascading with its $n_{2,\mathrm{eff}}$ and $\Delta\phi^{\mathrm{NL}}$ can generally replace the Kerr n_2. This is true for self-phase modulation and the applications stemming from it, including all-optical guided-wave switches such as directional couplers, Mach–Zehnder interferometers, etc. However, this is not necessarily true for other applications, because all the waves have to be coupled through $\chi^{(2)}$ interactions (at most three at a time) and must be nearly phase-matched for efficient effects. For example, cross-phase modulation between two input FF polarization components is feasible via Type II SHG cascading, but specific wavelengths will require different crystals, or crystal cuts. The nonlinear phase shift can also be used in all-optical devices for analog processing, such as isolators, modulators, and transistors, in laser mode-locking, etc.

In the following subsections we discuss a few additional applications. Specifically, we describe a non-reciprocal device that operates as an isolator, and the cascading of beams of finite extent in space or time. The latter can counterbalance either spatial diffraction or temporal dispersion or both and gives rise to solitons of various dimensionalities.

An Optical Diode

A nonlinear response, when combined with a device structure that is asymmetric about its midpoint, can provide nonreciprocal behavior upon excitation of the structure (crystal, waveguide) from different ends.

Table 1 Figures for cascading in various materials. The effective nonlinear coefficient is estimated under optimum conditions and a 1 cm propagation length

Material L = *1 cm*	d_{ii} *(pm/V)*	d_{ij} *(pm/V)*	$n_{2,eff}$ *(cm^2/W)*
$LiNbO_3$	36		2×10^{-11}
$LiNbO_3$		5.8	5×10^{-13}
MNA(1)	165		7×10^{-10}
NPP(2)		84	2×10^{-10}
DAST(3)	600		6×10^{-9}

(1)MNA: 2-methyl-4-nitroaniline
(2)NPP: N-(4-nitrophenyl)-(L)-prolinol
(3)DAST: dimethyl amino stilbazolium tosylate

In analogy to the standard electronic element, we can refer to such device as a power-dependent optical diode. Dealing with quadratically nonlinear waveguides, simple ways to induce a directional asymmetry are variations in wavevector mismatch, or in QPM periodicity. A mismatch profile which depends on distance from one end, or a localized defect (i.e., an isolated domain in a QPM grating), will give rise to different throughputs when the waveguide is excited from opposite sides, up to complete FF depletion in one direction, and complete transmission in the opposite one. This can be explained by the fact that the output depends on the evolution and interaction of the FF field with its harmonic. The amount and phase of the latter, however, do depend on the details of the propagation path and can substantially change the FF throughput as in the case of seeded SHG.

The sketch of an all-optical diode in a QPM waveguide with its calculated response is shown in Figure 6a,b. Clearly, total depletion is obtained via SHG in backward propagation, whereas unity FF transmission is recovered by the insertion of a phase discontinuity (and seeded interaction) in the forward path. As demonstrated at 1.55 μm in lithium niobate, different degrees of isolation are available at various excitations, with a large extinction ratio at a given power when launching FF waves from opposite ends.

Quadratic Solitons

When beams propagate in a two- (planar waveguides) or three-dimensional (bulk) medium with a quadratic response, cascading induces a nonlinear phase-front distortion which, coupled to the inherent parametric gain, can effectively counteract diffraction and lead to diffractionless propagation or 'spatial solitons' (Figure 7). In contrast to the Kerr case, such solitons are multifrequency solutions or 'simultons' with field components (FF and SH in the degenerate SHG process) bound together in propagation despite possible walk-off or transverse velocity mismatch. Moreover, since the second-order nonlinearity is inherently saturable, the solitons are stable and robust even in bulk media, and provide 'clean' filtered beam profiles at FF and SH frequencies, as has been demonstrated in KTP. Just like other cascading phenomena, quadratic solitons can be controlled by a weak seed, by polarization imbalance at the input (Type II SHG), by direction of propagation (due to birefringence), by phase mismatch, etc. Employing quadratic spatial solitons, phenomena such as collisions, soliton fusion, spatial reshaping, and transverse light localization and instabilities have been studied and constitute an active field of investigation.

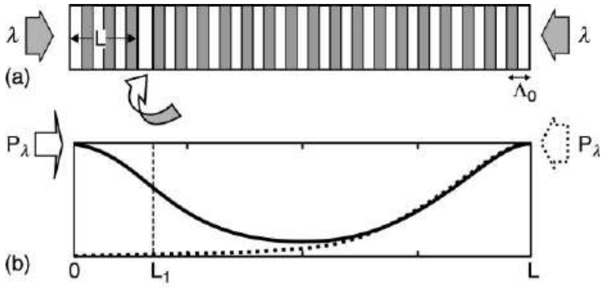

Figure 6 All-optical diode: (a) sketch of implementation with an engineered 'defect' in a QPM grating (*z*-cut periodically poled lithium niobate), and (b) example of FF transmittance *T* versus propagation for forward (solid line) and backward (dashed line) FF excitation of power P_λ.

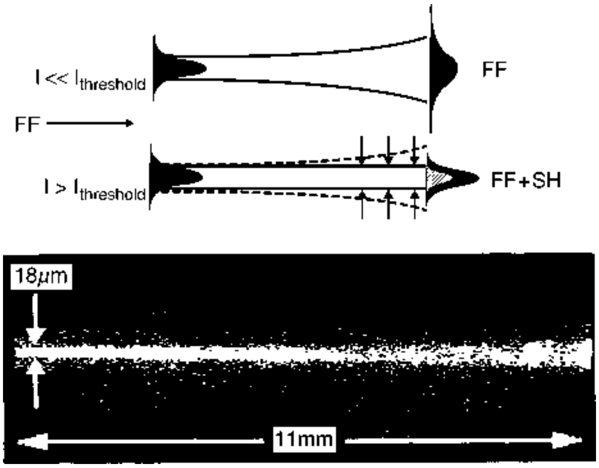

Figure 7 Top: sketch of a quadratic spatial soliton excited by an FF input: when the irradiance is high enough, linear diffraction is overcome by parametric self-trapping and a two-color simulton is obtained. Bottom: sample photograph (top view) of a quadratic soliton experimentally observed in a potassium niobate crystal excited by a 1064-nm laser beam.

When employing pulses, the cascaded phase shift can balance dispersion (material, chromatic, or photonics bandgap) leading to temporal simultons with frequency components locked together despite their individual group-velocity differences and dispersion. This feature can also substantially broaden the bandwidth available for parametric generation. Finally, in analogy to the cubic case, a combination of space and time effects is expected to lead to the formation of 'light bullets' in (3 + 1) dimensions.

List of Units and Nomenclature

Angular frequency	[s^{-1}]	ω
Dielectric constant of vacuum	[$FV^{-1}\ m^{-1}$]	ε_0

Electric field	[V m^{-1}]	E
Irradiance, intensity	[W m^{-2}]	I
Length	[m]	L
Refractive index	[ ]	n
Speed of light in vacuum	[m s^{-1}]	c
Second-order electronic susceptibility	[m V^{-1}]	$\chi^{(2)}$
Third-order electronic susceptibility	[m^2 V^{-2}]	$\chi^{(3)}$
Wavevector in medium	[m^{-1}]	$k = n\omega/c$
Wavevector in vacuum	[m^{-1}]	$k_0 = \omega/c$
Wavelength	[m]	$\lambda = 2\pi n/k = 2\pi/k_0$

See also

Solitons: Bright Spatial Solitons; Soliton Communication Systems. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Assanto G (1995) All-optical switching in integrated structures. In: DeGiorgio V and Flytzanis C (eds) *Nonlinear Optical Materials: Principles and Applications*. Amsterdam: IOS Press.

Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396.

Assanto G, Stegeman GI, Sheik-Bahae M and VanStryland E (1993) All optical switching devices based on large nonlinear phase shifts from second harmonic generation. *Applied Physics Letters* 62: 1323–1325.

Assanto G, Stegeman GI, Sheik-Bahae M and VanStryland EW (1995) Coherent interactions for all-optical signal processing via quadratic nonlinearities. *IEEE Journal of Quantum Electron* 31: 673.

Assanto G, Stegeman GI and Schiek R (1998) Thin film devices for all-optical switching and processing via quadratic nonlinearities. *Thin Solid Films* 331: 291–297.

Boardman AD (ed.) (1999) *Advanced Photonics with Second-Order Optically Nonlinear Processes*. Dordrecht, The Netherlands: Kluwer.

Boardman AD (ed.) (2001) *Soliton Driven Photonics*. Dordrecht, The Netherlands: Kluwer.

Gallo K and Assanto G (1999) Analysis of lithium niobate all-optical wavelength shifters for the third spectral window. *Journal of the Optical Society of America B* 16: 741–753.

Gallo K, Assanto G, Parameswaran KR and Fejer MM (2001) All-optical diode in a periodically poled lithium niobate waveguide. *Applied Physics Letters* 79: 314–316.

Kajzar F and Reinisch R (eds) (1997) *Beam Shaping and Control with Nonlinear Optics*. New York: Plenum Press.

Stegeman GI, Hagan DJ and Torner L (1996) $\chi^{(2)}$ cascading phenomena and their applications to all-optical signal processing, mode-locking, pulse compression and solitons. *Journal of Optics and Quantum Electronics* 28: 1691.

Stegeman GI and Assanto G (1999) Nonlinear integrated optical devices. In: Murphy EJ (ed.) *Design and Application of Integrated Optical Circuits and Components*. New York: Marcel Dekker.

$\chi^{(2)}$–Harmonic Generation

R C Eckardt, Cleveland Crystals, Inc., Highland Heights, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical harmonic generation is nonlinear conversion of laser output to light, of a frequency that is an integer multiple of that of the fundamental. It is used over a wide range of applications, from battery-operated, pencil-sized pointers emitting less than a milliwatt of light to installations that occupy large buildings and generate thousands of Joules of ultraviolet energy in nanosecond-duration pulses. The applications also range from frequency-stable continuous-wave light to generation of pulses a few femtoseconds in duration. Laser output with specific properties may only be available in narrow spectral regions, and harmonic generation is a method to convert the laser output to spectral regions of shorter wavelength. It is customary to call light of twice the frequency of the fundamental the second harmonic; three times the third harmonic, and so on.

The conversion of light by harmonic generation is a process that is driven by the nonlinear electric polarization. The total polarization $\mathbf{P}(t)$, including the linear term, is a vector quantity related to the electric field $\mathbf{E}(t)$ by the susceptibility tensors

χ of a material:

$$\mathbf{P}(t) = \varepsilon_0 \chi^{(1)} : \mathbf{E}(t) + \varepsilon_0 \chi^{(2)} : \mathbf{E}(t)\mathbf{E}(t) + \varepsilon_0 \chi^{(3)} : \mathbf{E}(t)\mathbf{E}(t)\mathbf{E}(t) + \ldots, \qquad [1]$$

where $\varepsilon_0 \cong 8.854 \times 10^{-12}$ C/N·m^2 is the permittivity of free space. The first-order tensor describes linear optical properties on the material. The second-order tensor $\chi^{(2)}$ describes second-harmonic generation and sum- and difference-frequency generation. This is the term that is of interest here. The second-order nonlinear optical tensor, equal to half the second-order susceptibility, is used widely instead of the susceptibility:

$$d_{ijk} = \chi^{(2)}_{ijk}/2 \qquad [2]$$

In calculations of harmonic generation and sum- and difference-frequency generation, the coefficient is usually reduced to a scalar quantity d_{eff}, the effective nonlinear coefficient that expresses the relationship between only the specific vector components involved in an interaction. The second-order optical nonlinearity, when present, is usually significantly larger than higher-order optical nonlinearities. When optical frequency conversion is possible, using the second-order nonlinearity, it is generally preferred. Any material that has inversion symmetry, however, will have vanishing second-order nonlinearity. Although there is continuing interest in developing other materials, such as poled polymers for future applications, materials for current practical applications are restricted to noncentrosymmetric crystals.

This article will provide a description of the process of optical harmonic generation and discuss some of the concerns in the design of harmonic generation systems. In many cases, harmonic generation can be modeled accurately by analytic or numerical techniques. Some of the theoretical development that is the basis for this modeling will be presented along with examples. There are several essential considerations that must be made in selecting a harmonic generation process. These involve both material properties and conditioning of the fundamental radiation.

General Considerations

The nonlinear optical material must be highly transmitting at the fundamental and harmonic wavelengths and must have a laser-induced-damage threshold high enough to allow optical intensities that will provide adequate conversion efficiency. Phase matching must be possible in the material. Other material characteristics must be considered for specific applications. The harmonic generation process requires that the phase velocities of the fundamental and harmonic radiation be matched. Without phase matching, power converted to the harmonic is back converted to the fundamental in a rapid oscillation, and harmonic conversion remains at low levels. The material also needs to be of good optical quality to provide optical transmission without distortion, which would destroy the phase relationship. Continuous-wave harmonic generation may involve local intensities of the order of 1 megawatt per square cm (1 MW/cm^2). Pulsed applications with Q-switched laser output of a few nanoseconds duration typically employ intensities of several hundred MW/cm^2, and harmonic-generation applications with mode-locked lasers may involve pulse durations of tens of femtoseconds with peak intensities of tens of gigawatts per square cm. Many other properties are desirable, but it is necessary to work within the limitations of available materials.

Phase matching is an essential requirement. When harmonic and fundamental waves of different eigen polarizations are coupled through the nonlinear optical tensor, it may be possible to use crystal birefringence to provide phase matching. The birefringence, if large enough, can compensate dispersion. The fundamental wave is introduced into the crystal with the polarization of the larger refractive index, and the harmonic is generated in the polarization of the smaller index. This is a description of type-I birefringent phase matching. Phase matching is achieved by selecting the direction of propagation in the crystal or by changing the temperature of the crystal. Type-II phase matching with the fundamental wave resolved into the two eigen polarizations is possible with large birefringence. Quasi phase matching involves periodically structuring the nonlinear material in a way that spatial modulation of the optical nonlinearity compensates for dispersion. Quasi phase matching is tuned by selection of the modulation period, and temperature control may provide a fine tuning. Periodic poling, one method of quasi phase matching, involves changing the orientation of ferroelectric domains after each coherence length of the harmonic generation process. These methods of phase matching are sketched in Figure 1.

The physical arrangement of the harmonic generation process depends on properties of the available fundamental light and the properties of the selected nonlinear material. The single-pass techniques, shown in Figure 1, may have the advantage of simplicity, but expanding or focusing the fundamental beam may be required. Crystal length and clear

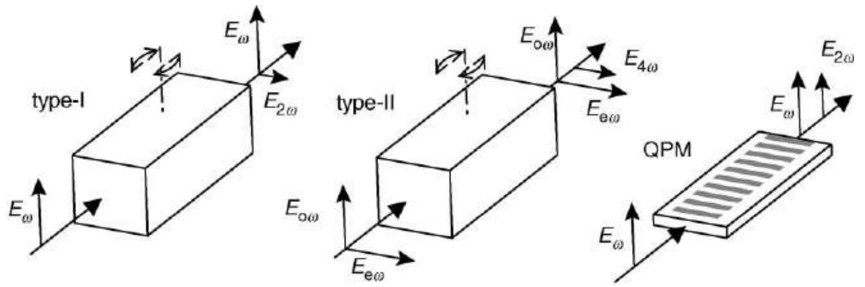

Figure 1 Different methods of phase matching optical harmonic generation are illustrated. The arrows indicate the directions of electric-field polarization and directions of propagation.

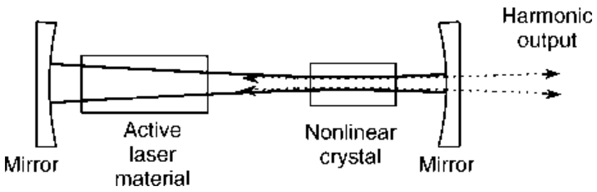

Figure 2 Placing a harmonic-generation crystal inside a laser cavity with mirrors highly reflecting at the laser wavelength is one technique of increasing the intensity of the fundamental beam.

aperture need to be selected. Intracavity harmonic generation (Figure 2), or external-resonant-cavity harmonic generation, may provide advantages in other cases. Analysis of harmonic generation provides the means to select among the options.

Basic Equations

Much of the analysis of second-harmonic generation follows from the differential equations that describe the interaction between the fundamental and harmonic fields, and it is appropriate to describe the development of these coupled equations. The wave equation:

$$\nabla^2\mathbf{E}(\mathbf{r},t) - \mu_0\sigma\frac{\partial\mathbf{E}(\mathbf{r},t)}{\partial t} - \mu_0\varepsilon_0\frac{\partial^2\mathbf{E}(\mathbf{r},t)}{\partial t^2} = \mu_0\frac{\partial^2\mathbf{P}(\mathbf{r},t)}{\partial t^2} \quad [3]$$

is obtained from Maxwell's equations. Here $\mu_0 = 4\pi\times10^{-7}$ is the permeability of free space, and σ is the conductivity, which later will be incorporated into the optical absorption coefficient $\alpha = \mu_0\sigma c/2$, and $(\varepsilon_0\mu_0)^{-1/2} = c \cong 2.998\times10^8$ m/s is the speed of light in vacuum. The electric field of a planewave, at angular frequency ω, is expressed in complex notation as the product of a slowly varying complex amplitude and a carrier wave:

$$\mathbf{E}_\omega(\mathbf{r},t) = (1/2)\{\mathbf{E}_\omega \exp \mathrm{i}(\mathbf{k}_\omega\cdot\mathbf{r} - \omega t) + \text{c.c.}\} \quad [4]$$

The intensity of this plane wave is

$$I_\omega = (nc\varepsilon_0/2)|E_\omega|^2 \quad [5]$$

Here n is the index of refraction of the material in which E is specified. The nonlinear polarization resulting from the second-order susceptibility driven by the plane wave $\mathbf{E}_\omega(\mathbf{r},t)$ is expressed as

$$\mathbf{P}_{2\omega}^{\mathrm{NL}}(\mathbf{r},t) = (1/2)\{\mathbf{P}_{2\omega}^{\mathrm{NL}} \exp \mathrm{i}(2\mathbf{k}_\omega\cdot\mathbf{r} - 2\omega t) + \text{c.c.}\} \quad [6]$$

The individual components of the polarization amplitude vector are

$$P_{2\omega,i}^{\mathrm{NL}} = \varepsilon_0\sum_{j,k} d_{ijk}(-2\omega;\omega,\omega)E_{\omega,j}E_{\omega,k} \quad [7]$$

If the nonlinear polarization is produced by two distinguishable waves, such as for sum-frequency generation of the third harmonic by the second harmonic and the fundamental, the nonlinear polarization is

$$\mathbf{P}_3^{\mathrm{NL}}(\mathbf{r},t) = (1/2)\Big\{\mathbf{P}_3^{\mathrm{NL}} \exp \mathrm{i}[(\mathbf{k}_1+\mathbf{k}_2)\cdot\mathbf{r} - (\omega_1+\omega_2)t] + \text{c.c.}\Big\} \quad [8]$$

where

$$P_{3,i}^{\mathrm{NL}} = 2\varepsilon_0\sum_{j,k} d_{ijk}(-\omega_3;\omega_1,\omega_2)E_{1,j}E_{2,k} \quad [9]$$

and

$$\omega_3 = \omega_1 + \omega_2 \quad [10]$$

A number of simplifications are used to arrive at the coupled equations for harmonic generation with monochromatic plane waves. All waves are assumed to be propagating collinearly in the z direction. Only the vector components involved in the interaction are retained, and the nonlinear optical tensor is replaced by the scalar effective nonlinear coefficient d_{eff}. Finally, the slowly varying envelope approximation is invoked and the higher-order derivatives of the amplitudes are ignored, as are time derivatives of the envelopes. The wave equation simplifies to

$$\frac{\mathrm{d}E(z)}{\mathrm{d}z}+\alpha E(z)=\mathrm{i}\frac{\mu_0 c\omega}{2n}P(z) \qquad [11]$$

where n is the index of refraction.

With three distinguishable waves, as with sum-frequency generation and type-II second-harmonic generation, there are three coupled equations:

$$\frac{\mathrm{d}E_3(z)}{\mathrm{d}z}+\alpha_3 E_3(z)=\mathrm{i}\frac{\omega_3 d_{\mathrm{eff}}}{n_3 c}E_1(z)E_2(z)\exp(-\mathrm{i}\Delta kz) \qquad [12]$$

$$\frac{\mathrm{d}E_1(z)}{\mathrm{d}z}+\alpha_1 E_1(z)=\mathrm{i}\frac{\omega_1 d_{\mathrm{eff}}}{n_1 c}E_3(z)E_2^*(z)\exp(\mathrm{i}\Delta kz) \qquad [13]$$

$$\frac{\mathrm{d}E_2(z)}{\mathrm{d}z}+\alpha_2 E_2(z)=\mathrm{i}\frac{\omega_2 d_{\mathrm{eff}}}{n_2 c}E_3(z)E_1^*(z)\exp(\mathrm{i}\Delta kz) \qquad [14]$$

Here the wave vector mismatch is

$$\Delta k=k_3-k_1-k_2 \qquad [15]$$

and

$$\omega_3=\omega_1+\omega_2 \qquad [16]$$

with the convention of $\omega_1\le\omega_2<\omega_3$. The indices of refraction of the three waves are n_1, n_2, and n_3, respectively, and α_1, α_2, and α_3 are the respective amplitude absorption coefficients.

With type-I second-harmonic generation, the fundamental field consists of only a single eigen polarization, and there are two coupled equations:

$$\frac{\mathrm{d}E_{2\omega}(z)}{\mathrm{d}z}+\alpha_{2\omega}E_{2\omega}(z)=\mathrm{i}\frac{\omega d_{\mathrm{eff}}}{n_{2\omega}c}E_\omega(z)E_\omega(z)\exp(-\mathrm{i}\Delta kz) \qquad [17]$$

$$\frac{\mathrm{d}E_\omega(z)}{\mathrm{d}z}+\alpha_\omega E_\omega(z)=\mathrm{i}\frac{\omega d_{\mathrm{eff}}}{n_\omega c}E_\omega(z)E_\omega^*(z)\exp(\mathrm{i}\Delta kz) \qquad [18]$$

Indices of refraction at the harmonic and fundamental are $n_{2\omega}$ and n_ω respectively; $\alpha_{2\omega}$ and α_ω are the respective absorption coefficients; and the wavevector mismatch is

$$\Delta k=k_{2\omega}-2k_\omega \qquad [19]$$

The equations for type-II second-harmonic generation, where $\omega_1=\omega_2$ and $\omega_3=2\omega_1$, become equivalent for those for type-I second-harmonic generation, providing intensities of the two fundamental polarizations are equal, $I_1=I_2$ and $\alpha_\omega=\alpha_{2\omega}$.

Steady-State Solutions

Three monochromatic planewave solutions for type-I second-harmonic generation are discussed. These solutions address type-I harmonic generation; first for low levels of conversion with arbitrary phase mismatch; next for high levels of conversion with perfect phase matching; and last for arbitrary levels of conversion with arbitrary amounts of phase mismatch. Near-field and focused-beam approximations also exist. Beyond that, numerical techniques, such as split-step Fourier transform methods can handle more complicated conditions. The simple approximations of monochromatic planewaves, or steady-state solutions, however, yield useful information.

The coupled equations for type-I second-harmonic generation can be solved by simple integration for low levels of fundamental depletion, where $E_\omega(z)$ is treated as constant and there is insignificant absorption. Initial conditions are $I_\omega(0)$, incident on the crystal and $I_{2\omega}(0)=0$. Integration of the equation for the harmonic electric field yields an expression for electric field after propagating a distance of l in the crystal:

$$E_{2\omega}(l)=\frac{\omega d_{\mathrm{eff}}E_\omega^2(0)}{n_{2\omega}c}\frac{1}{\Delta k}\left\{\exp\mathrm{i}(k_{2\omega}l)-\exp\mathrm{i}(2k_\omega l)\right\} \qquad [20]$$

Expressed as intensity the solution is

$$I_{2\omega}(l)=I_\omega(0)\Gamma^2 l^2\,\mathrm{sinc}^2(\Delta kl/2) \qquad [21]$$

where $\mathrm{sinc}(x)=\sin(x)/x$ and

$$\Gamma^2=2\omega^2 d_{\mathrm{eff}}^2 I_\omega(0)/(n_{2\omega}n_\omega^2 c^3\varepsilon_0) \qquad [22]$$

When $\Delta k=0$ and $\alpha_\omega=\alpha_{2\omega}=0$, conditions of perfect phase matching and no absorption, the coupled equations for type-I second-harmonic generation can be combined into a single equation. This is done using conservation of energy, which requires $E_\omega^2(z)+E_{2\omega}^2(z)=E_\omega^2(0)$, in this case with no initial harmonic intensity. The equation can then be manipulated into a standard integral that yields the inverse hyperbolic tangent. Rewritten and expressed in terms of intensity the result is

$$I_{2\omega}(l) = I_{\omega}(0)\tanh^2(\Gamma l) \qquad [23]$$

The two solutions for $I_{2\omega}(l)$ given above serve well for a number of applications. The $\mathrm{sinc}^2(\Delta k \cdot l/2)$ function is widely used to determine phase-matching acceptances. The $\tanh^2(\Gamma l)$ function is a good approximation for local conversion with large, well-collimated fundamental beams.

An exact solution to the coupled equations, however, gives insight to conditions such as high initial fundamental intensity with small amounts of phase mismatch. After extensive algebraic manipulation, a solution in the form of a Jacobian elliptic function $\mathrm{sn}(u|m)$ is obtained:

$$I_{2\omega}(l) = I_{\omega}(0)\upsilon_b^2 \mathrm{sn}^2\left(\frac{\Gamma l}{\upsilon_b}\middle|\upsilon_b^4\right) \qquad [24]$$

where

$$1/\upsilon_b = \Delta s/4 + \sqrt{1 + (\Delta s/4)^2} \qquad [25]$$

$$\Delta s = \Delta k/\Gamma$$

and Δk and Γ retain the earlier definitions. This solution can be extended to provide the phase of the fundamental and harmonic as the conversion process evolves. Expressing the complex harmonic and fundamental field amplitudes as the product of a real amplitude and a complex exponential:

$$E(z) = A(z)\exp \mathrm{i}\phi(z) \qquad [26]$$

the harmonic phase is

$$\phi_{2\omega}(z) = (\pi - \Delta k z)/2 \qquad [27]$$

The fundamental phase, however, is found to be dependent on the conversion process:

$$\begin{aligned}\phi_{\omega}(z) = {} & (1/2)\arccos[A_{2\omega}(z)\Delta k/(2\kappa A_{\omega}^2(z))] \\ & - (\pi - \Delta k z)/z \qquad [28]\end{aligned}$$

where $\kappa = \omega d_{\mathrm{eff}}/(n_{\omega}c)$. Small amounts of phase mismatch significantly change the phase of the fundamental at high levels of conversion. This phase change is used for switching applications in waveguide interferometers. It also makes phase-matching requirements more stringent at high levels of conversion. The three planewave solutions are illustrated in Figure 3. The narrowing of phase matching at higher conversion is illustrated in Figure 4.

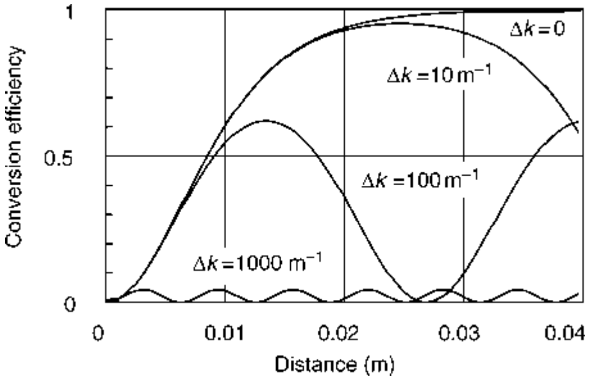

Figure 3 Monochromatic plane-wave solutions of the coupled equations for second-harmonic generation are illustrated for different amounts of wavevector mismatch. The solution takes the form $\tanh^2(\Gamma l)$ when $\Delta k = 0$ and $(\Gamma l)^2\mathrm{sinc}^2(\Delta k l/2)$ when conversion is small.

There is a Jacobian elliptic function solution for intensities with the coupled waves of type-II second-harmonic generation and sum-frequency generation. The phases of the individual waves can be recovered with elliptic integrals of the third kind. Both sets of equations, the two coupled equations for type-I phase matching and the three coupled equations for type-II phase matching, are amenable to numerical solution by Runge–Kutta methods.

Phase Matching

Phase matching is achieved in birefringent materials by balancing the dispersion of the material with the difference between the indices of refraction of the fast and slow waves. With adequate birefringence it may be possible to obtain phase matching by selecting the direction of propagation in the crystal. Equations describing the dispersion and birefringence of nonlinear crystals, necessary for calculating phase matching, along with the nonlinear optical coefficients and expressions for the effective nonlinear optical coefficients, are tabulated in handbooks. Values for a few examples are given here.

The negative uniaxial crystal potassium dihydrogen phosphate (KH_2PO_4 or KDP) is used as an illustrative example. The harmonic wave is chosen to have extraordinary polarization with the smaller index of refraction in this crystal with negative birefringence. With type-I phase matching, the fundamental consists of only an ordinary polarized wave. To have phase matching, $\Delta k = 0$, it is necessary to choose a direction of propagation that makes an angle θ with the optic axis that satisfies:

$$n_{e,2\omega}(\theta) = n_{o,\omega} \qquad [29]$$

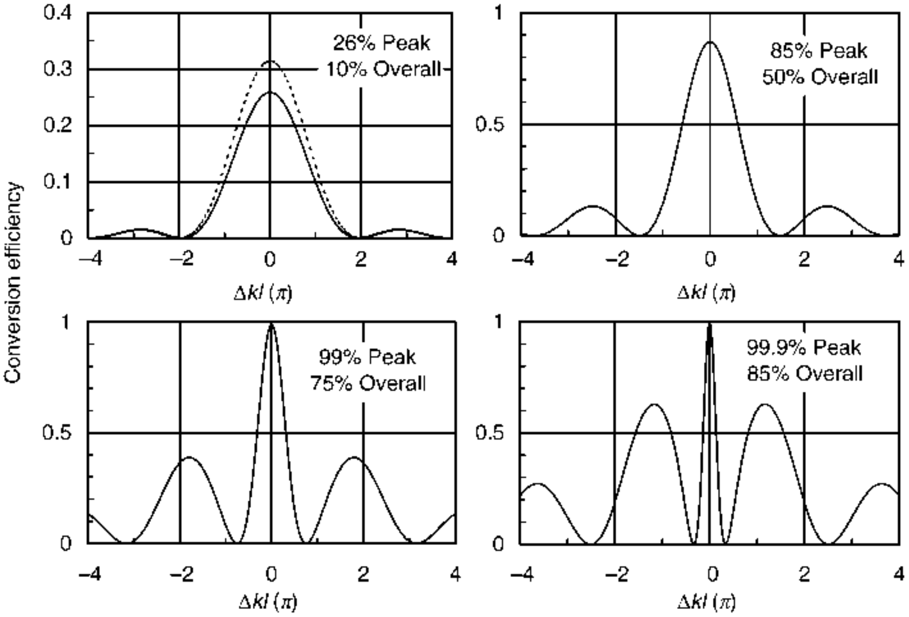

Figure 4 High levels of overall second-harmonic conversion require higher levels of conversion at peak pulse intensity. Phase-matching bandwidths narrow at the higher levels of peak conversion. The broken curve is the $(\Gamma l)^2\text{sinc}^2(\Delta kl/2)$ small conversion approximation.

The extraordinary index for propagation at an angle of θ to the optical axis in a uniaxial crystal $n_e(\theta)$ is given by

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2} \qquad [30]$$

where n_o is the ordinary index and n_e is the extraordinary index for propagation at 90° to the optic axis. The phase-matching angle is obtained by using the identity $\cos^2\theta = 1 - \sin^2\theta$ and solving for θ:

$$\theta_{PM} = \sin^{-1}\left\{\frac{n_{e,2\omega}^2(n_{o,2\omega}^2 - n_{o,\omega}^2)}{n_{o,\omega}^2(n_{o,2\omega}^2 - n_{e,\omega}^2)}\right\}^{1/2} \qquad [31]$$

where the subscripts refer to ordinary and extraordinary indices at the fundamental and second-harmonic frequencies.

The effective nonlinear coefficient is determined by crystal symmetry, the type of phase matching, the birefringence of the crystal, and the direction of propagation. For crystals of point group $\bar{4}2m$ with negative birefringence, such as KDP, the effective nonlinear coefficient for type-I phase matching is given by

$$d_{\text{eff}} = -d_{36}\sin\theta\sin 2\phi \qquad [32]$$

The azimuthal angle ϕ is measured from the crystal x-axis. The x-axis and y-axis are specified by X-ray diffraction and piezoelectric coefficients and can be determined from the growth morphology. The azimuthal orientation of a finished crystal, however, is not obvious, and requires X-ray orientation, piezoelectric characterization, or nonlinear optical measurements, if the record of crystal orientation is lost. The nonlinear coefficient d_{36} is an abbreviation for the d_{312} and d_{321} tensor components, which are equal. For harmonic generation with a 1064 nm fundamental in KDP, $d_{36} = 0.39 \times 10^{-12}$ m/V. The unit of 10^{-12} m/V is usually expressed as picometers per volt (pm/V). The effective nonlinear coefficient for type-II phase matching for this class of crystals is

$$d_{\text{eff}} = d_{36}\sin 2\theta\cos 2\phi \qquad [33]$$

Type-I and type-II crystals cut from the same material are not interchangeable. The type-II phase-matching angle for KDP can be found by iterative solution of the equation:

$$n_{e,2\omega}(\theta) = \{n_{o,\omega} + n_{e,\omega}(\theta)\}/2 \qquad [34]$$

The birefringent walkoff angle, that is the angle the Poynting vector of an extraordinary wave makes with the wavevector, is another consideration for birefringent phase matching. There will be no further harmonic generation when the extraordinary and ordinary fundamental beams of type-II phase matching no longer overlap. For type-I phase matching, birefringent walkoff can limit the interaction length, but there can be several interaction lengths in a crystal. The birefringent walkoff of an extraordinary wave propagating at an angle θ to the optic axis in a uniaxial crystal is given by

$$\rho = \tan^{-1}\left\{ \frac{n_e^2(\theta)}{2}\left[\frac{1}{n_e^2} - \frac{1}{n_o^2}\right]\sin(2\theta)\right\} \qquad [35]$$

Refractive indices for 1064 nm to 532 nm second-harmonic generation in KDP are $n_{e,2\omega} = 1.4709$, $n_{o,2\omega} = 1.5129$, $n_{e,\omega} = 1.4603$, and $n_{o,\omega} = 1.4944$. The dispersion of KDP and indices at the phase-matching angles are shown in Figure 5. For type-I phase matching $d_{eff} = 0.26$ pm/V, $\theta_{PM} = 41.2°$, and $\rho_{2\omega} = 28$ milliradians (mr), and for type-II phase matching $d_{eff} = 0.34$ pm/V, $\theta_{PM} = 59.1°$, $\rho_\omega = 20$ mr, and $\rho_{2\omega} = 25$ mr. Type-II phase matching in KDP for this application offers an advantage of larger effective nonlinear coefficient. Also the angular acceptance is larger for type-II phase matching.

The angular acceptance is calculated from dispersion equations by first evaluating the derivative $\partial\Delta k/\partial\theta$ and then finding the value of $\Delta\theta$ corresponding to the full width at half maximum (FWHM) of the $\mathrm{sinc}^2(\Delta k \cdot l/2)$ function:

$$\frac{\partial \Delta k}{\partial \theta}\Delta\theta_{FWHM} \cong 4 \times 1.39 \text{ radians}/l \qquad [36]$$

When θ is 0° or 90° for uniaxial crystals, or for propagation along one of the refractive index principal axes in a biaxial crystals, the first derivative $\partial\Delta k/\partial\theta$ becomes zero, and the angular acceptance must be calculated from the second derivative $\partial^2\Delta k/\partial\theta^2$. This case is called noncritical phase matching (NCPM), and the case for θ between 0° and 90° is called critical phase matching.

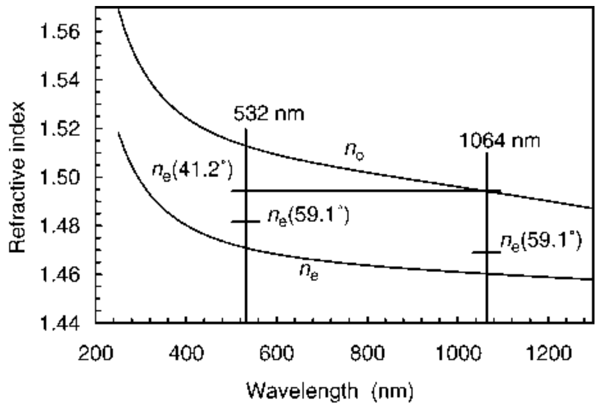

Figure 5 Dispersion of the birefringent nonlinear material KH_2PO_4 with refractive indices indicated for type-I and type-II 1064 nm to 532 nm second-harmonic generation. (Adapted from Byer RL (1977) Parametric oscillators and nonlinear materials. In: Harper PG and Wherrett BS (eds) *Nonlinear Optics*, pp. 47–159. San Francisco, CA: Academic Press.)

Second-harmonic generation of 1064 nm fundamental radiation in the crystal lithium triborate (LBO or LiB_3O_5) is an example of NCPM. At 149°C, this crystal will phase match for type-I SHG for propagation along the x-axis in the xyz frame determined by the indices of refraction of this biaxial crystal $n_x < n_y < n_z$. The following are calculated from dispersion equations for this case: $\partial\Delta k/\partial\theta = 0$; $\partial^2\Delta k/\partial\theta^2 = 2.04 \times 10^4$ (rad/cm)/rad^2; $\partial\Delta k/\partial\phi = 0$; and $\partial^2\Delta k/\partial\phi^2 = 6.45 \times 10^3$ (rad/cm)/rad^2. The FWHM angular acceptances for a 1.5 cm-long LBO crystal are $\Delta\theta_{FWHM} = 1.09°$ and $\Delta\phi_{FWHM} = 1.94°$. Angle-tuned CPM at 23°C in LBO occurs at $\theta = 90°$ and $\phi = 11°.5$. For this case $\partial\Delta k/\partial\phi = 1.32 \times 10^4$ (rad/cm)/rad, and the azimuthal angular acceptance decreases to $\Delta\phi_{FWHM} = 0.16°$. Also there is a birefringent walkoff angle of $\rho_{2\omega} = 19$ mr, which results in a displacement of 0.29 mm of the extraordinary-like harmonic beam for a 1.5 cm-long crystal. Both the larger angular acceptance and the absence of birefringent walkoff are advantages for harmonic generation. The effective nonlinear coefficient of LBO is reported as 0.85 pm/V and 1.04 pm/V by different sources.

Quasi phase matching offers the possibility, in principle, of noncritical phase matching at any wavelength in the transmission range of a material. However, it may be difficult producing the periodic structure that is required for phase matching in a particular application. The ferroelectric crystals lithium niobate ($LiNbO_3$), lithium tantalate ($LiTaO_3$), and potassium titanyl phosphate ($KTiOPO_4$ or KTP) can be processed by electric-field poling to produce domains of alternating polarity. Lithography techniques are used to apply a patterned electrode structure on the surface of a wafer or thin slab. Domains are reversed by the application of a carefully monitored and controlled electric field. Thicknesses of several mm have been achieved in periodically poled KTP (PPKTP). The higher coercive field of 21 kV/mm has limited the thicknesses that can be achieved in periodically poled lithium niobate (PPLN), particularly at shorter wavelengths for which poling periods become small. Progress is now being made with zincblende materials such as gallium arsenide. These materials have large linearity and good transmission in the infrared, but they are not birefringent. Epitaxial growth on patterned substrates is now being used to produce periodically structured GaAs on a research basis, and bonded stacks of wafers of alternating orientation also have been used. Harmonic conversion with quasi phase matching is compared with conventional phase matching and nonphase-matched harmonic generation in Figure 6.

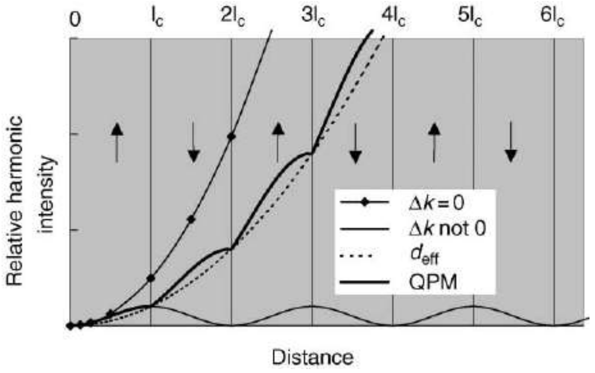

Figure 6 The growth of harmonic intensity is illustrated schematically for a quasi-phase-matched process in which the polarity of the nonlinearity is reversed after each coherence length. Conversion for $\Delta k = 0$, $\Delta k \neq 0$, and a uniform effective nonlinear coefficient are shown for comparison. (Adapted from Fejer MM (1992) Quasi-phase-matched second harmonic generation: tuning and tolerances. *IEEE Journal of Quantum Electronics* 28: 2631–2654.)

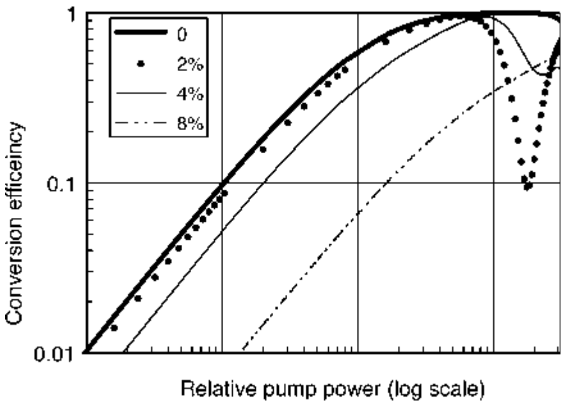

Figure 7 Random variations in domain length will lower harmonic conversion efficiency and have effect on back conversion. RMS variations in domain length are indicated in the legend for this specific example.

Quasi phase matching also provides access to tensor components that are not accessible with birefringent phase matching. This is the case for lithium niobate; the largest tensor component is $d_{333} = d_{33}$, which has a value of 25 pm/V. The effective nonlinear coefficient is the relevant Fourier component of the modulated nonlinearity:

$$d_{\text{eff}} = (2/\pi)d_{33} \approx 16 \text{ pm/V} \qquad [37]$$

This is a larger effective nonlinearity than is available with common birefringently phase matched materials in the visible and near-infrared. A more accurate treatment of second-harmonic generation involves application of the Jacobian-elliptic-function solution to each domain. After nearly complete conversion is reached, the phase perturbations on the individual domains have effect, and back conversion takes over (Figure 7). Conversion is lower and back conversion can set in sooner with random differences in the domain length. There is some experimental evidence that this type of back conversion has been observed in guided-wave harmonic conversion devices. Guided-wave nonlinear optical interaction can have long interaction lengths not limited by diffraction. Quasi phase matching is particularly useful in waveguide second-harmonic generation because the direction of propagation is fixed.

Gaussian Beams and Pulses

A radially symmetric Gaussian light beam has transverse intensity of

$$I(r) = I_0 \exp(-2r^2/w^2) \qquad [38]$$

where r is the radial coordinate and w is called the spot size. Harmonic generation in the low-conversion limit by a Gaussian beam, with a focus of spot size w_0 at the center of a crystal with optimized phase matching, is described by

$$p_{2\omega}(l)/p_\omega(0) = 2\omega^2 d_{\text{eff}}^2 p_\omega(0) lkh(B,\xi)/(\pi n^3 \varepsilon_0 c^3) \qquad [39]$$

The harmonic power after propagating distance l in the crystal is $p_{2\omega}(l)$, and $p_\omega(0)$ is the incident fundamental power. The function $h(B,\xi)$ is the Boyd and Kleinman focusing factor. It is a function of the walkoff parameter:

$$B = \rho\sqrt{lk}/2 \qquad [40]$$

and the focusing parameter:

$$\xi = l/(kw_0^2) \qquad [41]$$

The spot size w_0 of the fundamental beam at its waist is used to described the fundamental intensity distribution at the focus at the center of the crystal:

$$I_\omega(r) = (2p_\omega(0)/(\pi w_0^2))\exp(-2r^2/w_0^2) \qquad [42]$$

Values of the focusing factor for arbitrary focusing and walkoff are usually displayed in the form of a graph (Figure 8). In the case of weak focusing $\xi \ll 1$ and relatively small walkoff $w_0 > \rho l$, the focusing factor is given by

$$h(B,\xi) \approx \xi(1 - t^2/12 + t^4/120 - t^6/1344 + \ldots) \qquad [43]$$

where $t = 2B(2\xi)^{1/2}$.

An example of continuous-wave second harmonic generation is used. This calculation models a conversion experiment published in the literature. External-resonant-cavity second-harmonic generation of 6.5 W of 532 nm radiation was reported. The 1064 nm fundamental radiation with 378 W of circulating power in a resonant cavity external to the laser was focused to a 32 μm spot size in the center of a 6 mm-long LBO crystal (Figure 9). This corresponds to a focusing factor of $\xi = 0.62$ in the index $n = 1.6053$ material. The value of the focusing parameter is $h(0, 0.62) = 0.58$. The nonlinear optical coefficient for this process is reported as 0.85 pm/V and 1.04 pm/V by different sources. When these values are inserted into the earlier equation for $p_{2\omega}(l)/p_{\omega}(0)$, the higher coefficient yields a conversion efficiency of 0.0273 and 10.3 W of 532 nm output, whereas the lower coefficient yields a conversion efficiency of 0.0182 and 6.9 W of 532 nm radiation. Differences in published values of nonlinear optical coefficients are common. The reported precision of the experiment and the measured values tend to support the lower value. Practical applications are useful to check the accuracy of parameter values. With many variables and the added complication of operation at high intensity there is uncertainty that demands caution in accepting the derived parameter values. It is interesting to note that the power and beam size of the example result in a maximum intensity of 17 MW/cm^2 at the surface of this crystal. Many vendors will warranty antireflection coatings for only 1 MW/cm^2 maximum continuous-wave intensity.

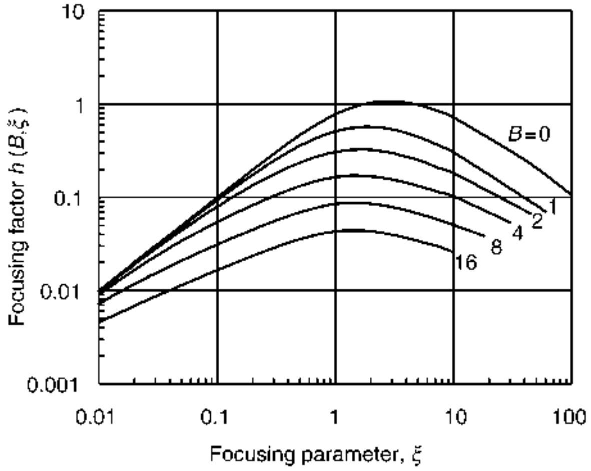

Figure 8 The Boyd and Kleinman focusing factor $h(B,\xi)$ for type-I second-harmonic generation is shown as a function of the focusing parameter ξ for different values of the walkoff parameter B. The curves shown were generated with a split-step Fourier transform calculation. (Adapted from Boyd GD (1968) Parametric interactions of focused Gaussian light beams. *Journal of Applied Physics* 39: 3597–3639.)

For a pulsed fundamental with temporal shape $p_{\omega}(t) = p_{\omega}(0)\exp(-t^2/\tau^2)$, integration over time gives a calculated energy conversion efficiency of:

$$\begin{aligned}\eta_{\text{calc}} &= U_{2\omega}(l)/U_{\omega}(0)\\ &= \sqrt{2}\omega^2 d_{\text{eff}}^2 U_{\omega}(0) lkh(B,\xi)/(\tau\sqrt{\pi^3}n^3\varepsilon_0 c^3)\end{aligned} \quad [44]$$

where $U_{2\omega}(l)$ is the harmonic energy generated after propagation distance l in the crystal and $U_{\omega}(0)$ is the incident fundamental energy. This is still a low conversion approximation. An estimate of conversion efficient at arbitrary levels of depletion η_{est} is obtained with the empirical expression:

$$\eta_{\text{est}} = \eta_{\text{calc}}/(1 + \eta_{\text{calc}}) \quad [45]$$

Returning to the example 1.064 nm to 532 nm, type-II SHG in KDP and using $l = 3$ cm, $U_{\omega}(0) = 0.3$ J, and $\tau = 3.6$ ns for a 6 ns FWHM pulse duration, the above two relationships predict a conversion efficiency of 50%. These parameters give a peak intensity of 330 MW/cm^2, a value typical for efficient harmonic conversion, while avoiding laser-induced damage in a long-term, repetitive-pulse application using KDP. Single-shot laser-induced-damage thresholds for KDP are about 10 times higher.

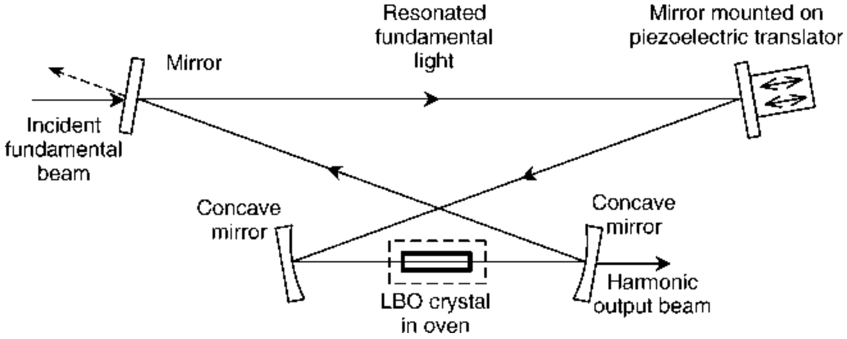
Resonated fundamental light
Mirror mounted on piezoelectric translator
Mirror
Incident fundamental beam
Concave mirror
Concave mirror
LBO crystal in oven
Harmonic output beam

Figure 9 A bow-tie cavity for externally resonant second-harmonic generation is illustrated. The reflected fundamental beam is used to develop an error signal to drive the piezoelectric translator and hold the cavity in resonance.

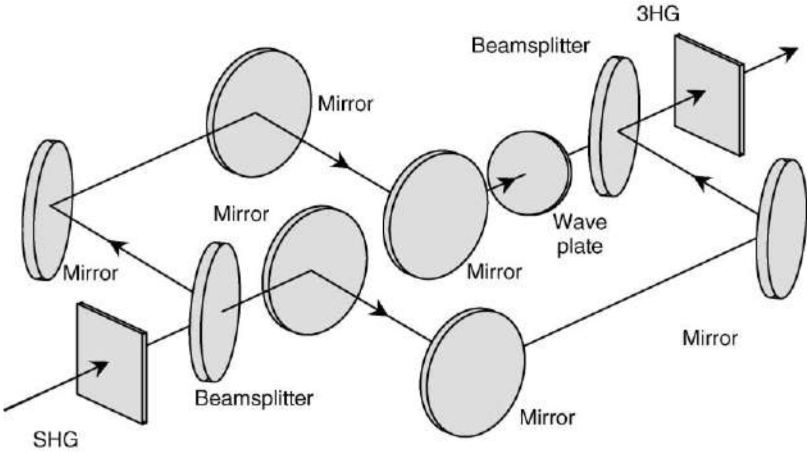

Figure 10 Short crystals less than 1 mm in length are typically used for second-harmonic generation with fundamental pulses that are only ten's of femtoseconds in duration. It is necessary to separate fundamental and second-harmonic pulses and recombine them with appropriate delay to compensate for group-velocity mismatch before sum-frequency generation to the third harmonic.

As a final example, consider second-harmonic generation of a 100 fs ($1\ \text{fs} = 10^{-15}$ s) duration pulse from a mode-locked Ti:Al_2O_3 laser operating at 800 nm. Typical laser output might consist of 10 nJ (10×10^{-9} Joule) pulses at 80 MHz repetition rate. Group velocity walkoff is a concern for this short pulse duration (Figure 10). Barium borate (BaB_2O_4 or BBO) is a common crystal for this application. The group velocity walkoff rate between the ordinary-polarization 800 nm fundamental pulse and the extraordinary-polarization 400 nm harmonic pulse is 190 fs/cm. A crystal length no longer than 0.5 mm is needed to avoid lengthening the harmonic pulse in time. Fortunately, the laser-induced-damage-threshold intensity increases approximately as the inverse of the square root of the pulse length between roughly 10 ns and 1 ps, and it is possible to focus the beam to achieve higher intensity without crystal damage and compensate for the short crystal length. However, the 68-milliradian birefringent walkoff of the harmonic will limit the size of the focus. A 20 μm spot size will allow an interaction length that roughly matches the crystal length and provides a peak fundamental intensity around 16 GW/cm^2 and a conversion efficiency of nearly 30%. Now suppose the 100 fs pulse is amplified to 100 mJ in a large chirped-pulse regenerative amplifier. A larger diameter crystal is required, and KDP is an appropriate choice. The group velocity walkoff rate in KDP is 78 fs/mm, and a crystal of 1.5 mm thickness could be used. With the beam spot size increased to 30 mm, the peak intensity will be 71 GW/cm^2, and the conversion efficiency will be about 40%. Material availability, such as the large sizes possible with KDP, often determines a choice.

Conclusion

Basic techniques for describing and modeling second-harmonic generation have been discussed. Birefringent phase matching and quasi phase matching were reviewed. Three examples were used to illustrate considerations for implementing harmonic-generation systems. There are many more parameters to be considered for other applications and for more detailed analysis. The examples presented illustrate the demands nonlinear optical frequency conversion places on materials. A few materials have been established as primary choices in common applications. But for many applications, available materials or laser system performance are limiting and it is necessary to use the materials and systems that are available. The importance of using the analysis methods is stressed. It is possible to sacrifice performance by improperly specifying a crystal. Choosing the best crystal for a harmonic-generation application with an appropriate optical design will optimize harmonic conversion.

See also

Nonlinear Optics, Applications: Phase Matching. **Nonlinear Sources:** Harmonic Generation in Gases. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Akhmanov SA (1975) Optical harmonic generation and optical frequency multipliers. In: Rabin H and Tang CL (eds) *Quantum Electronics. A Treatise, vol. 1, part B*, pp. 475–583. New York: Academic Press.

Armstrong JA (1962) Interaction between lightwaves in a nonlinear dielectric. *Physics Review* 127: 1918–1962.
Boyd GD (1968) Parametric interactions of focused Gaussian light beams. *Journal of Applied Physics* 39: 3597–3639.
Boyd RW (2003) *Nonlinear Optics*, 2nd edn. New York: Academic Press.
Byer RL (1977) Parametric oscillators and nonlinear materials. In: Harper PG and Wherrett BS (eds) *Nonlinear Optics*, pp. 47–159. San Francisco, CA: Academic Press.
Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.
Fejer MM (1992) Quasi-phase-matched second harmonic generation: tuning and tolerances. *IEEE Journal of Quantum Electronics* 28: 2631–2654.
Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.
Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn. New York: Marcel Dekker.
Yang SL (1991) 6.5 W, 532 nm radiation by cw resonant external-cavity second-harmonic generation of an 18 W Nd:YAG laser in LiB_3O_5. *Optical Letters* 16: 1493–1495.
Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: John Wiley & Sons.

$\chi^{(3)}$–Third-Harmonic Generation

B Y Soon, Thales Electro-Optics PTE Ltd., Singapore

J W Haus, University of Dayton, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The invention of the ruby laser, in 1960, provided a revolutionary new light source, which emits an intense, highly collimated polarized beam of light. The new source opened the way for developments that have enabled new technologies, thus profoundly changing our way of life. One of these new frontiers opened by the laser was the field of nonlinear optics. In 1961, Franken *et al.* were the first to report the observation of second-harmonic generation from a quartz crystal, using ruby laser light. Third-harmonic generation (THG) was soon discovered by Terhune and co-workers in 1962.

THG has found applications as a spectroscopic tool to probe optical properties of materials, but recently it has been developed as a microscopy tool and a pulse measurement technique. More efficient THG is possible by engineering the materials; for instance, one way to improve conversion efficiency is to use periodic structured dielectric materials, called photonic bandgap structures.

Polarization and the Third-Order Susceptibility

THG is a nonlinear optical phenomenon, where light of angular frequency ω is converted into light at an angular frequency 3ω that is three times higher. Light is made of particles called photons and the nonlinear medium enables the process whereby three photons of frequency ω are annihilated to create one photon of frequency 3ω. The nonlinear interactions are strongly dependent on the polarization or the dipole moment per unit volume, which is written as a series

$$P(t) = \chi^{(1)} \times E(t) + \chi^{(2)} : E^2(t) + \chi^{(3)} \vdots E^3(t) + \cdots \quad [1]$$

where χ is the susceptibility of the optical material, the superscript of χ in parentheses denotes the order of nonlinearity, t is the time variable, and E is the electric field. Both the electric field and the polarization are vectors; therefore the susceptibilities are tensors. The first-order or linear susceptibility $\chi_{ij}^{(1)}$ is a second-order tensor. This term constitutes the linear polarization of the medium denoted a P^{L}. It has nine elements, but only up to three independent elements depending on the symmetry of the lattice. The other terms represent an expansion for the nonlinear polarization of the medium, P^{NL}. Note that the superscripts 'L' and 'NL' denote linear and nonlinear polarization contributions, respectively. The third-order susceptibility $\chi_{ijkl}^{(3)}$ is a four-rank tensor, which usually has 81 elements ($3 \times 3 \times 3 \times 3 = 81$). In asymmetric crystalline solids, all 81 elements can be of nonzero value and independent. With a higher degree of symmetry in other optical materials, the 81 elements are no longer all independent, and thus the number of independent elements can be reduced to a lower number. For example, isotropic material such as glass, liquid, or vapor has only three independent elements. For illustration, we consider a general case where the applied frequencies ω are arbitrary, and the third-susceptibility is given as

$$\chi_{ijkl} \equiv \chi_{ijkl}^{(3)}(\omega_4 = \omega_1 + \omega_2 + \omega_3) \quad [2]$$

Note that the superscript has been removed. In the isotropic material, the coordinate axes must be equivalent. Thus, the susceptibility has symmetry given by

$$\chi_{1111} = \chi_{2222} = \chi_{3333} \tag{3a}$$

$$\chi_{1122} = \chi_{1133} = \chi_{2211} = \chi_{2233} = \chi_{3311} = \chi_{3322} \tag{3b}$$

$$\chi_{1212} = \chi_{1313} = \chi_{2323} = \chi_{2121} = \chi_{3131} = \chi_{3232} \tag{3c}$$

$$\chi_{1221} = \chi_{1331} = \chi_{2112} = \chi_{2332} = \chi_{3113} = \chi_{3223} \tag{3d}$$

For the isotropic material, there are only 21 nonzero elements, as given in eqns [3a]–[3d], because these elements have a Cartesian index (1, 2, or 3) that appears an even number of times. The reason for nonexistence of the elements with an index appearing an odd number of times is because, for example, χ_{1222} would give a response in the $\hat{x}_1$ direction due to a field applied in the $\hat{x}_2$ direction. This response must vanish in an isotropic material, because there is no reason why the response should be in the $+\hat{x}_1$ direction rather than in the $-\hat{x}_1$ direction. In summary, the elements in the third-order susceptibility of the isotropic material have the relationship given by

$$\chi_{ijkl} = \chi_{1122}\delta_{ij}\delta_{kl} + \chi_{1212}\delta_{ik}\delta_{jl} + \chi_{1221}\delta_{il}\delta_{jk} \tag{4}$$

where δ is the Kronecker delta function; it has a unit value, $\delta_{ij} = 1$, if and only if $i = j$, otherwise, it is zero, $\delta_{ij} = 0$. Thus, we see from eqn [4] that there are only three independent elements in the isotropic material. Crystalline materials have more independent components and they have been determined for the 32 crystals classes. The relationships between the elements for these crystalline classes are tabulated in books by Boyd and by Butcher and Cotter.

A block diagram of a material illustrates THG and its corresponding energy-level diagram is depicted in Figure 1. Figure 1a shows that fields of frequency ω excite the χ^3 optical material to generate the third-harmonic field with the residual of the fundamental harmonic field. Figure 1b shows that three photons of frequency ω are destroyed to generate a photon of frequency 3ω in the third-harmonic generation.

For third-harmonic generation, the frequency dependence of the susceptibility is expressed as $\chi_{ijkl}(3\omega = \omega + \omega + \omega)$. With the intrinsic permutation symmetry of the nonlinear susceptibility, the elements of the susceptibility tensor has a relationship such that $\chi_{1122} = \chi_{1212} = \chi_{1221}$. So we can further reduce eqn [4] for the third-harmonic generation as

$$\chi_{ijkl}(3\omega = \omega + \omega + \omega) = \chi_{1122}(3\omega = \omega + \omega + \omega) \times (\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) \tag{5}$$

Therefore, we see that the third-harmonic susceptibility tensor has only one independent element.

Wave Equation for χ^3 Third-Harmonic Generation

Theoretically, we have to use the wave equation to observe how an intense laser field interacts with the nonlinear material, so that the polarization of the medium could develop new frequency components that are not present in the incident field. Thus, these new frequency components of the polarization become the sources of the new frequency components of the electromagnetic field.

To formally describe THG, we begin with Maxwell's equations (in Gaussian units), and we have

$$\nabla \times D = 4\pi\rho \tag{6a}$$

$$\nabla \times B = 0 \tag{6b}$$

$$\nabla \times E = \frac{-1}{c}\frac{\partial B}{\partial t} \tag{6c}$$

$$\nabla \times H = \frac{-1}{c}\frac{\partial D}{\partial t} + \frac{4\pi}{c}J \tag{6d}$$

where H is the magnetic field, B is the magnetic flux density, D is the electric flux density, ρ is the charge density, and J is the current density. The desired solution is in the region of space in the nonlinear material that has no free charge ($\rho = 0$), and therefore, there is no free current ($J = 0$). We assume

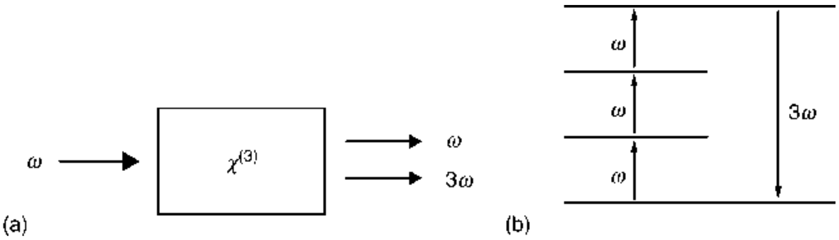

Figure 1 The third-harmonic (a) block diagram; and (b) energy level description.

that the nonlinear material is nonmagnetic, so that $B = \mu H = H$, where $\mu = 1$ in a nonmagnetic environment. In nonlinear material, the electric flux is defined as

$$D = E + 4\pi P \qquad [7]$$

so that the polarization depends nonlinearly on the strength of E. Using eqns [6c] and [6d], we have

$$\nabla \times \nabla \times E + \frac{1}{c^2}\frac{\partial^2 D}{\partial t^2} = 0 \qquad [8]$$

Substituting eqn [7] into [8], we have

$$\nabla \times \nabla \times E + \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{-4\pi}{c^2}\frac{\partial^2 P}{\partial t^2} \qquad [9]$$

Using an identity from the vector calculus, we have

$$\nabla \times \nabla \times E = \nabla(\nabla \times E) - \nabla^2 E \qquad [10]$$

In linear optics, we see from eqn [6a] that when there is no free charge in source-free isotropic material, $\nabla \times D = \nabla \times E = 0$. Thus, the first term on the right-hand side of eqn [10] has no effect. Whereas in nonlinear optics, $\nabla \times E \neq 0$ in isotropic material, but the contribution from the first term on the right-hand side of eqn [10] is negligible, therefore, we can ignore it. Thus, eqn [9] can be simplified as

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P}{\partial t^2} \qquad [11]$$

For distinction of the linear and nonlinear contributions, we modified eqn [7] to get

$$D = E + 4\pi P^{\mathrm{L}} + 4\pi P^{\mathrm{NL}}$$

where $P = P^{\mathrm{L}} + P^{\mathrm{NL}}$. Let $D^{\mathrm{L}} = E + 4\pi P^{\mathrm{L}}$, and we rewrite eqn [11] as

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 D^{\mathrm{L}}}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P^{\mathrm{NL}}}{\partial t^2} \qquad [12]$$

Equation [12] is the most general form of the wave equation seen in nonlinear optics, and could be used as our starting point for our study of third-harmonic generation.

Methods of Third-Harmonic Generation

Since nonlinear optics is a popular research topic, there are a handful of existing methods that can be used for third-harmonic generation, and possibly new methods may be developed in the future. Here, we will discuss three methods using χ^3 material: (i) photonic crystal; (ii) defect mode using photonic crystal; and (iii) homogeneous nonlinear crystal.

Photonic Crystal

After the invention of electronic bandgap structure, the optical scientists and engineers saw the possibility to extend the operating waveband into the visible, and thus we have the photonic bandgap structure. The geometry and material choices of the photonic bandgap structure create a unique bandgap in frequency spectrum that has potential applications in laser optics, telecommunication, medical imaging, optical limiting, etc. A photonic bandgap structure is also commonly known as the photonic crystal. For a one-dimensional photonic crystal, two materials of certain thicknesses each, repeat themselves a number of times. For example, the first and second materials are of thickness 'a' and 'b', respectively, so that their combined thickness is 'd', and have a periodicity of 10. The schematic of such a one-dimensional photonic crystal, with its corresponding dielectric function as a function of axial distance z, is depicted in Figure 2.

By using χ^3 materials for photonic crystal, and using method of multiple scales (MMS), we have the nonlinear polarization given as

$$P^{\mathrm{NL}} = \lambda \chi^{(3)} E^3 \qquad [13]$$

where λ is the perturbation parameter. With reference to Figure 2, assume that 'a' is equal to 'b' so that $d = 2a$, then the Fourier expansion of the dielectric

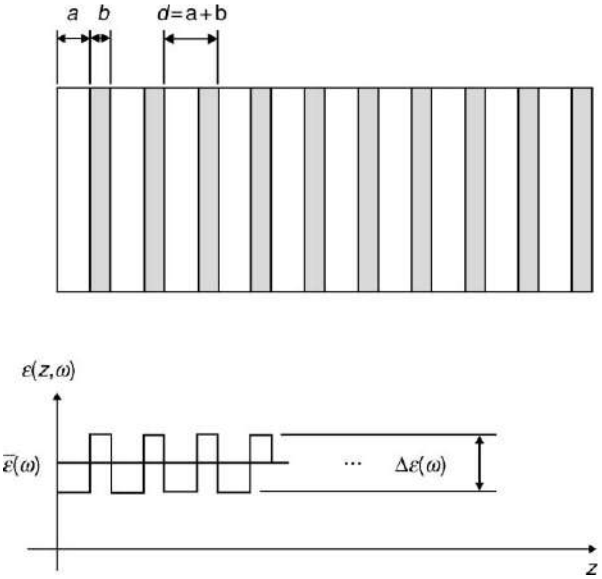

Figure 2 The schematic of the one-dimensional photonic crystal.

function is given as

$$\varepsilon(z,\omega) = \bar{\varepsilon}(\omega) + \lambda\Delta\varepsilon(\omega)\cos(2\pi z/d) - \frac{1}{3}\lambda\Delta\varepsilon(\omega)\cos(6\pi z/d) + \cdots \qquad [14]$$

where $\bar{\varepsilon}(\omega)$ is the zeroth-order coefficient, and $\Delta\varepsilon(\omega)$ is first-order coefficient of the Fourier series expansion. Without going through the mathematics of MMS, we use eqns [12] and [13] to obtain the solution of the wave equation and two sets of coupled mode equations. The solution to the wave equation in eqn [12] is given by

$$\begin{aligned} E_0 = {} & \left(\frac{1}{\sqrt{k_1}}A_{\mathrm{f}1}(z_1,t_1)\mathrm{e}^{\mathrm{i}k_1z_0} + \frac{1}{\sqrt{k_1}}A_{\mathrm{b}1}(z_1,t_1)\mathrm{e}^{-\mathrm{i}k_1z_0}\right)\mathrm{e}^{-\mathrm{i}\omega t_0} \\ & + \left(\frac{1}{\sqrt{k_3}}A_{\mathrm{f}3}(z_1,t_1)\mathrm{e}^{\mathrm{i}k_3z_0} + \frac{1}{\sqrt{k_3}}A_{\mathrm{b}3}(z_1,t_1)\mathrm{e}^{-\mathrm{i}k_3z_0}\right)\mathrm{e}^{-3\mathrm{i}\omega t_0} \\ & + \mathrm{c.c.} \end{aligned} \qquad [15]$$

where A is the amplitude of the field, k is the wavenumber such that $k_1^2 = \omega^2\varepsilon(\omega)/c^2$, and $k_3^2 = (3\omega)^2\varepsilon(3\omega)/c^2$ where c is the speed of light, and 'c.c.' refers to the complex conjugate. Note that the subscript 'f' and 'b' refer to the forward and backward modes, respectively. Also, the subscript '0' refers to the slowest scale, and '3' refers to the fastest scale chosen in the analysis using MMS. Let $a_{\mathrm{f}1} = \mathrm{N}A_{\mathrm{f}1}\mathrm{e}^{\mathrm{i}\pi\delta_1 z_1/d}$, $a_{\mathrm{b}1} = \mathrm{N}A_{\mathrm{b}1}\mathrm{e}^{-\mathrm{i}\pi\delta_1 z_1/d}$, $a_{\mathrm{f}3} = \mathrm{N}A_{\mathrm{f}3}\mathrm{e}^{\mathrm{i}\pi\delta_3 z_1/d}$, and $a_{\mathrm{b}3} = \mathrm{N}A_{\mathrm{b}3}\mathrm{e}^{-\mathrm{i}\pi\delta_3 z_1/d}$, such that

$$N^2 = \frac{2d\sqrt{3k_3}(3\omega)^2\chi^{(3)}}{k_3c^2k_1\sqrt{k_1}}$$

The coupled mode equations for the fundamental field are given as

$$\begin{aligned} \frac{\partial a_{\mathrm{f}1}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g}1}}\frac{\partial a_{\mathrm{f}1}}{\partial t_1} = {} & \frac{\mathrm{i}d}{2\pi k_1}\nabla_\perp^2 a_{\mathrm{f}1} + \mathrm{i}\delta_1 a_{\mathrm{f}1} - \mathrm{i}\kappa_1 a_{\mathrm{b}1} \\ & - \frac{\alpha_1}{2}a_{\mathrm{f}1} + \mathrm{i}a_{\mathrm{f}1}[2|a_{\mathrm{b}1}|^2 + |a_{\mathrm{f}1}|^2] \\ & + \mathrm{i}a_{\mathrm{f}1}^{*2}a_{\mathrm{f}3} \end{aligned} \qquad [16a]$$

$$\begin{aligned} \frac{-\partial a_{\mathrm{b}1}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g}1}}\frac{\partial a_{\mathrm{b}1}}{\partial t_1} = {} & \frac{\mathrm{i}d}{2\pi k_1}\nabla_\perp^2 a_{\mathrm{b}1} + \mathrm{i}\delta_1 a_{\mathrm{b}1} - \mathrm{i}\kappa_1 a_{\mathrm{f}1} \\ & - \frac{\alpha_1}{2}a_{\mathrm{b}1} + \mathrm{i}a_{\mathrm{b}1}[2|a_{\mathrm{f}1}|^2 + |a_{\mathrm{b}1}|^2] \\ & + \mathrm{i}a_{\mathrm{b}1}^{*2}a_{\mathrm{b}3} \end{aligned} \qquad [16b]$$

where, at the fundamental harmonic, $v_{\mathrm{g}1}$ is the group velocity of the homogeneous medium, δ_1 is the laser detuning, κ_1 is the coupling coefficient, α_1 is the absorption coefficient, $\nabla_\perp^2$ is the transverse Laplacian and the whole term ($\nabla_\perp^2 a_{\mathrm{f}1}$ and $\nabla_\perp^2 a_{\mathrm{b}1}$) accounts for diffraction effects, and '$*$' denote the complex conjugate of the corresponding term. Similarly, the coupled mode equations for the third-harmonic field are given as

$$\begin{aligned} \frac{\partial a_{\mathrm{f}3}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g}3}}\frac{\partial a_{\mathrm{f}3}}{\partial t_1} = {} & \frac{\mathrm{i}d}{2\pi k_3}\nabla_\perp^2 a_{\mathrm{f}3} \\ & + \mathrm{i}a_{\mathrm{f}3}\left[\delta_3 + 6\sqrt{\frac{3k_1}{k_3}}(|a_{\mathrm{f}1}|^2 + |a_{\mathrm{b}1}|^2)\right] \\ & - \mathrm{i}\kappa_3 a_{\mathrm{b}3} - \frac{\alpha_3}{2}a_{\mathrm{f}3} + \mathrm{i}a_{\mathrm{f}1}^3 \end{aligned} \qquad [17a]$$

$$\begin{aligned} -\frac{\partial a_{\mathrm{b}3}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g}3}}\frac{\partial a_{\mathrm{b}3}}{\partial t_1} = {} & \frac{\mathrm{i}d}{2\pi k_3}\nabla_\perp^2 a_{\mathrm{b}3} \\ & + \mathrm{i}a_{\mathrm{b}3}\left[\delta_3 + 6\sqrt{\frac{3k_1}{k_3}}(|a_{\mathrm{b}1}|^2 + |a_{\mathrm{f}1}|^2)\right] \\ & - \mathrm{i}\kappa_3 a_{\mathrm{f}3} - \frac{\alpha_3}{2}a_{\mathrm{b}3} + \mathrm{i}a_{\mathrm{b}1}^3 \end{aligned} \qquad [17b]$$

where at the third harmonic, $v_{\mathrm{g}3}$ is the group velocity of the homogeneous medium, δ_3 is the laser detuning, κ_3 is the coupling coefficient, and α_3 is the absorption coefficient. Note that in the limit when $k_3 \approx 3k_1$, the coupling coefficient between the fundamental and third harmonics are related by

$$\kappa_3 = 3\left(\frac{d}{\pi}\right)\frac{\Delta\varepsilon\omega_0^2}{2k_3c^2} + 3\delta_1 = \kappa_1 \qquad [18]$$

Also, note that the laser detuning between the fundamental and third harmonics are expressed as

$$\delta_3 = \left(\frac{d}{\pi}\right)\Delta k + 3\delta_1 \qquad [19]$$

where $\Delta k = k_3 - 3k_1 \times$ Equation [19] described the phase matching condition, which determines the efficiency of the third-harmonic generation. When the phase is completely matched ($\Delta k = 0$), then we can achieved the most efficient in third-harmonic generation when $\delta_3 = 3\delta_1$. Equations [16] and [17] are useful equations which are commonly used for numerically simulation of beam propagation.

Defect Mode Using Photonic Crystal

A defect mode can be generated in the photonic crystal if we introduce a defect in the geometry of the photonic crystal. The schematic of the photonic crystal with a defect of length d_{d} located at the center of the photonic crystal, where the subscript '*d*' denotes defect, is depicted in Figure 3. For the convenience of the analysis, we divide the entire photonic crystal into three regions: (i) Region I; (ii) Region II; and (iii) Region III. Since we arbitrarily

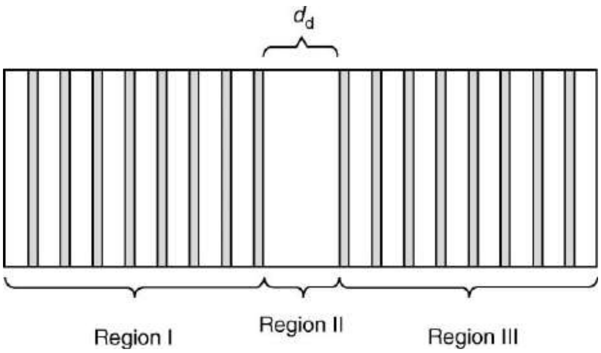

Figure 3 The schematic of the photonic crystal with a defect of length d_d.

put the defect in the center of the photonic crystal, Region I and II will be identical, thus it will be easier for us in our analysis. The strategy of our analysis is to solve the wave equation for Region I of the photonic crystal, and apply the result to the Region II analysis. Similarly, we apply the corresponding result of Region II to Region III. Thus, by breaking up the entire photonic crystal into three regions, we can then easily analyze the propagation of a field through the entire photonic crystal.

Assume that Region I and II are each of length $L/2$. Using the coupled mode equations without considering the effect of absorption, the fundamental fields in the forward and backward modes in Region I is given as

$$a_{\mathrm{f1(REG)}}(z_1) = \left(\cos(\Delta_1 z_1) - \frac{\mathrm{i}\delta_1}{\Delta_1}\sin(\Delta_1 z_1)\right)a_{\mathrm{f1(REG)}}(0) + \frac{\mathrm{i}\kappa}{\Delta_1}\sin(\Delta_1 z_1)a_{\mathrm{b1(REG)}}(0) \qquad [20a]$$

$$a_{\mathrm{b1}(REG)}(z_1) = \left(\cos(\Delta_1 z_1) + \frac{\mathrm{i}\delta_1}{\Delta_1}\sin(\Delta_1 z_1)\right)a_{\mathrm{b1(REG)}}(0) - \frac{\mathrm{i}\kappa}{\Delta_1}\sin(\Delta_1 z_1)a_{\mathrm{f1(REG)}}(0) \qquad [20b]$$

where

$$\Delta_1 = \sqrt{\delta_1^2 - \kappa^2}, \qquad \kappa = \frac{d\omega_0^2\Delta\varepsilon(\omega)}{2\pi k_1 c^2}$$

and the subscript '(REG)' denotes solution in either Region I, II, or III. The boundary conditions for the fundamental harmonic field, at the interface between Region I and II, are given as

$$a_{\mathrm{f1(II)}}(d_\mathrm{d}) = \mathrm{e}^{ikd_\mathrm{d}}a_{\mathrm{f1}(I)}(L/2) \qquad [21a]$$

$$a_{\mathrm{b1(II)}}(d_\mathrm{d}) = \mathrm{e}^{-ikd_\mathrm{d}}a_{\mathrm{b1}(I)}(L/2) \qquad [21b]$$

where $k = n\omega/c$. The field that propagates through Region II can be determined using eqns [20] and [21]. Also, the boundary conditions at the front and end faces of the Region III of the photonic crystal are given as

$$a_{\mathrm{f1(III)}}(0) = a_{\mathrm{f1(II)}}(d_\mathrm{d}) \qquad [22a]$$

$$a_{\mathrm{b1(III)}}(L/2) = 0 \qquad [22b]$$

Using eqns [20] and [22], we can determine the field that propagates through Region III. Similarly, we can use the same approach to analyze the third-harmonic field.

Homogeneous Nonlinear Crystal

If we examine Figure 2, we see that when $\Delta\varepsilon(\omega) = 0$, the photonic crystal becomes a homogeneous nonlinear crystal. In another word, $\kappa_3 = 0$ when we set the laser detuning to zero, thus there is no more coupling between the layers, and we are left with only the forward mode. Without going through the mathematics of four-wave mixing to describe the THG in a homogeneous nonlinear crystal, we can obtain the fundamental and third-harmonic coupled mode equations from eqns [16] and [17]. This will make it easier to make comparisons between the photonic crystal and the homogeneous nonlinear crystal, rather than comparing the coupled mode equation with the four-wave mixing equations among them. Since no backward wave is generated in the homogeneous nonlinear crystal when there is no coupling between the forward and backward modes, eqn [16] can remove the coupling term and the backward mode equation, and the modification is given as

$$\frac{\partial a_{\mathrm{f1}}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g1}}}\frac{\partial a_{\mathrm{f1}}}{\partial t_1} = \frac{\mathrm{i}d}{2\pi k_1}\nabla_\perp^2 a_{\mathrm{f1}} + \mathrm{i}\delta_1 a_{\mathrm{f1}} - \frac{\alpha_1}{2}a_{\mathrm{f1}} + \mathrm{i}a_{\mathrm{f1}}|a_{\mathrm{f1}}|^2 + \mathrm{i}a_{\mathrm{f1}}^{*2}a_{\mathrm{f3}} \qquad [23]$$

$$\frac{\partial a_{\mathrm{f3}}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g3}}}\frac{\partial a_{\mathrm{f3}}}{\partial t_1} = \frac{\mathrm{i}d}{2\pi k_3}\nabla_\perp^2 a_{\mathrm{f3}} + \mathrm{i}a_{\mathrm{f3}}\left[\delta_3 + 6\sqrt{\frac{3k_1}{k_3}}|a_{\mathrm{f1}}|^2\right] - \frac{\alpha_3}{2}a_{\mathrm{f3}} + \mathrm{i}a_{\mathrm{f1}}^3 \qquad [24]$$

for the fundamental and third-harmonic fields, respectively.

Applications

Generation of short wavelength radiation is possible using THG. Therefore, it becomes less expensive to generate green, blue, or ultraviolet light with the infrared and the red light. In order to improve THG conversion efficiency, a beam with a tight focus is used, which results in an electric field of highest intensity, and has the highest possible χ^3 polarization effect. Normally, due to the phenomenon of the Gouy phase shift at a tight focal point of the beam, third-harmonic generation is suppressed. However, in the presence of an interface within the focal volume, the phase is disturbed and a third-harmonic signal is generated. This phenomenon has been used to image cells in *in vivo* microscopy on a femtosecond time-scale. The ultrashort pulses have little energy and therefore little heat is deposited in the material. The cells can be continuously imaged without damage.

See also

Materials Characterization Techniques: $\chi^{(3)}$. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides.

Further Reading

Barad Y, Eisenberg H, Horovitz M and Silberberg Y (1997) Nonlinear scanning laser microscopy by third harmonic generation. *Applied Physics Letters* 70: 922–924.

Boyd RW (1992) *Nonlinear Optics*. Boston, MA: Academic Press.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge: Cambridge University Press.

Franken PA, Hill AE, Peters CW and Weinreich G (1961) Generation of optical harmonics. *Physics Review Letters* 7: 118.

Markowicz PP, Tiryaki H, Pudavar H, Prasad PN, Lepeshkin NN and Boyd RW (2004) Dramatic enhancement of third-harmonic generation in three-dimensional photonic crystals. *Physics Review Letters* 92: 083903.

Petrov GI, Shcheslavskiy V, Yakovlev VV, Ozerov I, Chelnokov E and Marine W (2003) Efficient third-harmonic generation in a thin nanocrystalline film of ZnO. *Applied Physics Letters* 83: 3993–3995.

Schins J, Schrama T, Brakenhoff G, Muller M and Squier J (2002) Determination of material properties using third harmonic generation microscopy. *Journal of the Optical Society of America B* 19: 1627.

Squier J, Mueller M, Brakenhoff GJ and Wilson KR (1998) Third harmonic generation microscopy. *Optics Express* 3: 315.

Terhune RW, Maker TD and Savage CM (1962) Optical harmonic generation in calcite. *Physics Review Letters* 8: 404.

Viswanathan R (1999). Enhanced Frequency Conversion in Periodic Structures. Ph.D. dissertation, Rensselaer Polytechnic Institute.

Four-Wave Mixing

L Canioni and L Sarger, University of Bordeaux, Talence, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Light waves propagate along straight lines in a vacuum and in materials. However, their amplitude and phase can only be affected by the intrinsic properties of the sample, such as absorption and refraction index. Experimental work from the early 1960s, has deviated from this simple approach, mainly through propagating intense light beams in crystals and glasses. Among many puzzling new effects, frequency conversion and beam coupling have triggered intense studies and opened a whole new field, theoretically and experimentally. These various phenomena have been classified according to the order of their optical nonlinearity. The second order mainly occurs in frequency generation and has already been discussed in previous articles of this encyclopedia. Therefore, this article will be devoted to the higher order, in which self focusing, coupling, and mixing of light beams will be explained. We will illustrate some simple ideas with a convenient model before describing a few applications.

The framework of electromagnetic wave interaction with matter is embedded in the Lorentz force. The oscillating electric field drives the components, ions, and electrons, but due to the large mass differences, only the electron movement is strongly perturbed. In dielectric media, these carriers are connected with static bonds such that the applied electric field modulates the spatial position of equilibrium of the charges. As a result, a dipolar moment is created and oscillates quasi-adiabatically at very high frequencies (between 10^{14} and 10^{17} Hz) for an applied optical wave. These collective oscillations of the induced dipoles compose the so-called matter polarization, proportional to the driving electric field amplitude. Magnetic field interaction is usually much weaker and so can be neglected in nonmagnetic media.

As an electromagnetic wave, lasers are a coherent source of very high intensity, i.e., large electric fields. For example, in standard pulsed lasers focused on a target, the light intensity can easily reach 10 Tw/cm^2, corresponding to an electric field of the order of magnitude of the internal bonding field (10^{10} V/m). At this level of interaction, dipolar moments are no longer strictly proportional to the applied electric field, as depicted in Figure 1. In the weak field regime, the simple perturbation approach displaces the carriers in a power series of the applied electric field. The polarization is thus developed in series, in respect to this electric field.

Formalism

Relation Between Field and Polarization: The Response Function

Although numerous works have been published with a frequency description of the material polarization, we will discuss here the study of nonlinearity in the time domain in order to follow the physical reality. In most processes, both spaces are equal and can be selected at will, as long as the entire requirements are fully understood; the electromagnetic field is classically described.

In the time domain, the polarization can be developed as

$$P(r,t) = P_{\text{Linear}}(r,t) + P_{\text{Nonlinear}}(r,t) \qquad [1]$$

where the linear part of the polarization can be written as

$$P_{\text{Linear}}(r,t) = \varepsilon_0 \int_{R^3} \int_{-\infty}^{\infty} R^{(1)}(r_1,t_1) \times E(r-r_1, t-t_1)\mathrm{d}t_1\,\mathrm{d}r_1 \qquad [2]$$

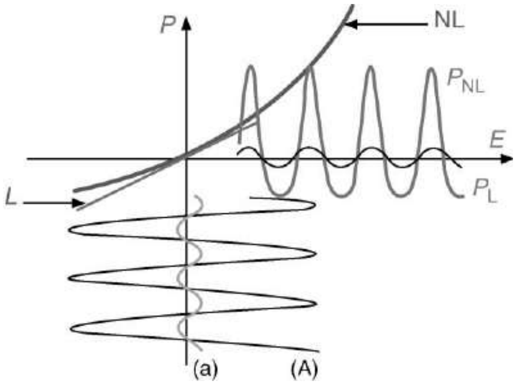

Figure 1 Polarization for various field amplitude: (a) Weak field case, linear response. (A) Strong field, distortion and nonlinearity.

This general expression is rather complex but shows that the polarization, due to collective movement of carriers in a macroscopic volume (thousands of dipoles) centered at position r, always depends on the applied electric field. This expression is also nonlocal as dipoles–dipoles interactions play an important role.

In general, for the nonlinear (NL) part of the polarization, an approximation of the local interaction is described for simplicity. For a second-order interaction, one can write:

$$P^{(2)}_{\text{NL}}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} R^{(2)}(t_1,t_2) \times E(r,t-t_2)E(r,t-t_2-t_1)\mathrm{d}t_1\,\mathrm{d}t_2 \qquad [3]$$

For the other series expansion – at third order, one obtains

$$P^{(3)}_{\text{NL}}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} R^{(3)}(t_1,t_2,t_3)E(r,t-t_3) \times E(r,t-t_3-t_2)E(r,t-t_3-t_2-t_1)\mathrm{d}t_1\,\mathrm{d}t_2\,\mathrm{d}t_3 \qquad [4]$$

This third-order polarization already combines three electric fields. This nonlinear polarization also radiates a fourth wave and is therefore the adopted framework for describing four-wave mixing experiments (Figure 2).

As the physical properties of materials depend only on time intervals between the different pulse interactions, eqn (4) can be rearranged as

$$P^{(3)}_{\text{NL}}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} R^{(3)}(t-\tau_1,t-\tau_2,t-\tau_3) \times E(r,\tau_1)E(r,\tau_2)E(r,\tau_3)\mathrm{d}\tau_1\,\mathrm{d}\tau_2\,\mathrm{d}\tau_3 \qquad [5]$$

If the sample response time is much shorter than pulse duration, the so-called adiabatic limit (as for electronic cloud deformation excited distant from

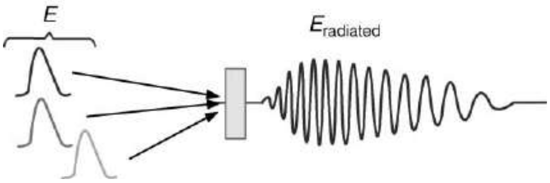

Figure 2 Three incoming beams E_1, E_2, E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4.

resonance) the polarization is written as

$$P^{(3)}_{\mathrm{NL}}(r,t)=\varepsilon_0\sigma^{(3)}E(r,t)E(r,t)E(r,t) \qquad [6]$$

We will see later how this expression is useful for third-harmonic generation and optical Kerr effects, etc. This expression is accurate for materials excited at a distance from resonances and care must be taken to reserve such a model for corresponding processes. In this case, the frequency approach is quite simple too and both spaces, time or Fourier, can be chosen at will.

For very long response time, such as orientational or structural molecular reorganization, one can use the Born–Oppenheimer approximation and present the nonlinear polarization as

$$P^{(3)}_{\mathrm{NL}}(r,t) = \varepsilon_0 E(r,t)R'(t) \qquad [7]$$

Similar is linear polarization, where the response function takes into account the changes of the molecular structure under the light intensity stress:

$$R'(t) = \int_{-\infty}^{+\infty} d^{(3)}(t-t_1)E(r,t_1)E(r,t_1)\mathrm{d}t_1 \qquad [8]$$

The material response will then be mostly driven by the field amplitude rather than the frequency.

Properties of Nonlinear Susceptibilities

The third-order nonlinear susceptibility $\chi^{(3)}$ is the Fourier transform of the response function R and has proved to be a most practical tool in optical nonlinear theory. The symmetry properties of the material lead to simplification on the tensor of rank 4 (fourth rank tensor), involved in four-wave mixing:

$$P_i^{(3)}(\omega_m)=\varepsilon_0 \sum_{n_1,\ldots,n_p}\sum_{i_1,\ldots,i_p}\chi^{(3)}_{i,i_1,\ldots,i_p}(\omega_m;\omega_{n_1},\ldots,\omega_{n_p}) \times E_{i_1}(\omega_{n_1})\cdots E_{i_p}(\omega_{n_p}) \qquad [9]$$

The nonlinear susceptibility tensor must account for all the spatial and symmetrical order of the sample. This will eventually reduce the number of independent elements from a possible 81.

Wave Equation in Nonlinear Regime

From Maxwell's equations, with a sample without free carriers and current, one can deduce the wave equation in the time domain, as:

$$\Delta E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} - \frac{1}{c^2}\frac{\partial^2 P_{\mathrm{L}}}{\partial t^2} = \frac{1}{c^2}\frac{\partial^2 P_{\mathrm{NL}}}{\partial t^2} \qquad [10]$$

In the frequency domain, this equation reduces to

$$\Delta\tilde{E}(r,\omega) + k(\omega)^2\tilde{E}(r,\omega) = -\mu_0\omega^2\tilde{P}_{\mathrm{NL}}(r,\omega) \qquad [11]$$

In order to solve these (coupled) wave equations in relation to four-wave mixing, we must develop the electric field and the polarization in a slowly varying envelope approximation, using the following conventions:

$$E_i(r,t)=\frac{1}{2}\left(A_i(r,t)\mathrm{e}^{\mathrm{i}(\omega_i t-k(\omega_i)z)}+\bar{A}_i(r,t)\mathrm{e}^{-\mathrm{i}(\omega_i t-k(\omega_i)z)}\right) \qquad [12]$$

$$P_{\mathrm{NL}}(r,t)=\frac{1}{2}\left(p_{\mathrm{nl}}(r,t)\mathrm{e}^{\mathrm{i}(\omega t-k(\omega)z)}+\bar{p}_{\mathrm{nl}}(r,t)\mathrm{e}^{-\mathrm{i}(\omega t-k(\omega)z)}\right) \qquad [13]$$

Within the framework of small perturbations, diffraction can be neglected and the wave equation can be easily reduced to one dimension:

$$\frac{\partial A}{\partial z} = \frac{\mathrm{i}\omega}{2n(\omega)\varepsilon o c}p^{\mathrm{NL}}(z,\omega)\exp(-\mathrm{i}k(\omega)z) \qquad [14]$$

where ω represents the sum or difference permutations over all the frequencies ω_i, thus reflecting the law of energy conservation.

In the case of four-wave mixing, one has to solve four coupled equations for the four interacting waves. Using the polarization eqn [13], one gets:

$$\frac{\partial A_j}{\partial z}=\frac{\mathrm{i}\omega_j}{2n(\omega_j)c}\chi^3 A_1^{(*)}A_2^{(*)}A_3^{(*)} \times\exp\left(\sum_{i=1,2,3}\exp(\pm\mathrm{i}k_i(\omega_i)z)\right)\exp(-\mathrm{i}k(\omega_j)z) \qquad [15]$$

This set of generic equations respectively represent indifferently, either the three incoming beams ($j=1,2,3$), or the radiated field E_4 at ω. The algebra used here associates at each complex conjugate field (*) with the counter propagating wave ($-k$). The amplitude variation of the j field at frequency ω_j is due to the coupling of the three other fields within the sample. The accumulated field is nevertheless always strongly dependent of the phase term and will vanish except when this phase term is zero. This is known as the phase matching condition which has to be fulfilled to ensure an efficient energy conversion. A more general approach of this condition can be found in the corresponding article of this encyclopedia.

Example and Selected Applications

When inserting the full material polarization (linear and nonlinear) into the propagation equation, even restricted at third order, a whole set of processes must be considered. All of them are described by the generic equation already presented, but only a few very important processes – with increasing complexity – will be described here.

Optical Kerr Effect

We consider one single beam incident at frequency ω on a sample. The non-linear polarization is given by

$$P^{(3)}_{\mathrm{NL}}(r,t)=\frac{\varepsilon_0}{8}\chi^{(3)}\big(E(r,t)^3+E^*(r,t)^3+3E(r,t)^2E^*(r,t)+3E(r,t)E^*(r,t)^2\big) \qquad [16]$$

The polarization at frequency ω reduces to

$$P^{(3)}_{\mathrm{NL}}(r,t)=\frac{3\varepsilon_0}{4}\chi^{(3)}|E(r,t)|^2E(r,t) \qquad [17]$$

This term has the same phase as the incident beam and, therefore, the phase matching is automatic. Using eqns [12] and [13], one can rewrite the wave equation for the optical Kerr effect in the thin sample approximation:

$$\frac{\partial A}{\partial z}=\frac{3\mathrm{i}\omega}{8nc}\chi^{(3)}|A|^2A \qquad [18]$$

The solution of this equation is

$$A(z)=A(0)\exp\left(\frac{3\mathrm{i}}{8nc}\chi^{(3)}|A(0)|^2z\right) \qquad [19]$$

Along the beam, the amplitude of the input pulse stays constant for this process but its phase is affected by a nonlinear shift proportional to power density I.

In literature, one can find the concept of n_2, the sample nonlinear index, in the following formula:

$$\psi_{\mathrm{NL}}(z)=\frac{\omega}{c}n_2Iz \qquad [20]$$

In this way of describing the total phase along the beam, the index of refraction in the nonlinear regime can be expressed more conveniently as

$$n=n_0+n_2I \qquad [21]$$

One of the most significant demonstrations of this effect is the self-focusing phenomenon. The usual Gaussian transverse structure of a laser beam presents high intensity at the center of the beam, therefore creating an index variation in the isotropic sample due to this optical Kerr effect. An induced lens can then be applied to this refractive index gradient, following the field distribution. As a result, the beam focuses itself along the propagation and can result in a complete breakdown.

Third Harmonic Generation

In the third harmonic generation, i.e., the third-order nonlinear process where the frequency of the generated wave is three times the frequency of the input wave at ω, only two waves are in the sample (at ω and 3ω). For the nonlinear wave, the master equation is then:

$$\frac{\partial A_{3\omega}}{\partial z}=\frac{3\mathrm{i}\omega}{8n(3\omega)c}\chi^{(3)}A^3_{\omega}\exp((3\mathrm{i}k(\omega)-\mathrm{i}k(3\omega))z) \qquad [22]$$

Here, propagation effects have to be carefully considered as theses waves do travel at different velocities and will eventually be out of phase. Obviously this discrepancy of phase velocities precludes the co-propagation and strongly affects the global efficiency. The only way to ensure a coherent generation of the harmonic wave throughout the material is to fulfill the relation:

$$k(3\omega)=3k(\omega) \qquad [23]$$

This is the so-called phase matching condition. Although difficult because the usual dispersion of materials is very large for such a frequency difference, this can be satisfied using either geometrical arrangement or, if possible, using the birefringent properties of the material. The efficiency of this method is poor and a better approach is implemented when two $\chi^{(2)}$ processes are cascaded (doubling and sum frequency mixing).

Degenerate Four-Wave Mixing and Phase Conjugation

One of the most popular and intriguing interactions is depicted in Figure 3. Where the sample is excited by three fields at the same frequency ω in this particular geometry: two-waves, named pump beams E_1 and E_2, are exactly counter-propagating with opposite wave vectors of $+k_1$ and $k_2=-k_1$, while the third wave (E_3) arrives at the sample with a given wave vector of k_3.

Of possible couplings, one contribution E_4 corresponds to a re-emitted beam along $k_4=-k_3$, i.e., opposite to the E_3 wave.

The third-order nonlinear polarization is given here by

$$P_{\mathrm{NL}}^{(3)}(r,t) = \frac{\varepsilon_0}{8}\chi^{(3)}(E_1(r,t) + E_1^*(r,t) + E_2(r,t) + E_2^*(r,t) + E_3(r,t) + E_3^*(r,t))^3 \qquad [24]$$

This E_4 wave (along $-k_3$) originates in this product from source terms including:

$$E_1(r,t)E_2(r,t)E_3^*(r,t) \qquad [25]$$

and thus present an evolution driven by:

$$\frac{\partial A_4}{\partial z} = \frac{3\mathrm{i}\omega}{4n(\omega)c}\chi^{(3)}A_1A_2A_3^* \qquad [26]$$

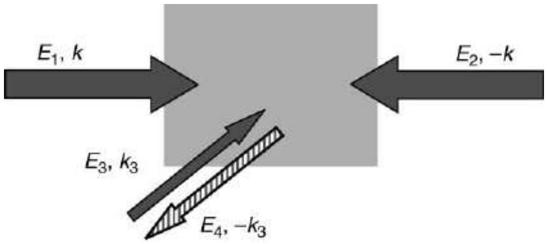

Figure 3 Two counter-propagating fields (E_1, E_2) and the third field E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4 opposite to the E_3 wave.

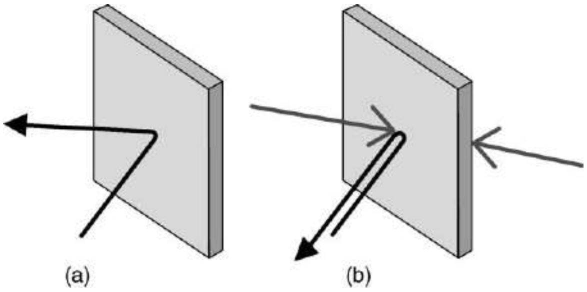

Figure 4 (a) Snell refraction in a conventional mirror; (b) Phase conjugate mirror.

where the phase matching condition is automatically fulfilled. Interesting facts can be described using this expression. First, the radiated amplitude is proportional to the complex conjugate of the E_3 field. This nonlinear process is not only able to reverse the direction of propagation but also the whole phase of an arbitrary incoming beam of light. This 'phase conjugator' can be considered as a kind of mirror with very unusual reflection properties. Unlike a conventional mirror, where a ray is redirected according to the ordinary law of reflection, a phase-conjugate mirror (PCM) retro-reflects all incoming rays back to their origin (Figure 4).

Second, the reflectivity of this mirror depends on the susceptibility and the amplitude of the first two fields, usually called pump beams. This will allow reflectivity that is much higher than the unity, unlike with conventional mirrors.

The remarkable reflection properties of the PCM have found many important applications. The most useful undoubtedly is related to distortion correction. If the image information has been distorted by the transmitting medium on its way to the PCM, then these aberrations will be corrected when the reflected signal retraces its original path through the medium. Through this amazing 'healing' property of optical phase conjugation, a high-quality optical beam can be double passed through a distorting medium, such as an imperfect imaging device or a turbulent atmosphere, without any loss of beam quality. Figure 5 shows the basic two-pass geometry for imaging and aberration correction using a PCM.

The spherical wave from the object point P is aberrated by the distorting medium. The wave-front is reversed by the PCM and a second passage through the same distorting path restores the initial field. The resulting spherical wave converges to the image point P', separated from P by the beamsplitter.

Numerous other applications for PCM and the various underlying nonlinear optical interactions have been proposed. These include effects as diverse

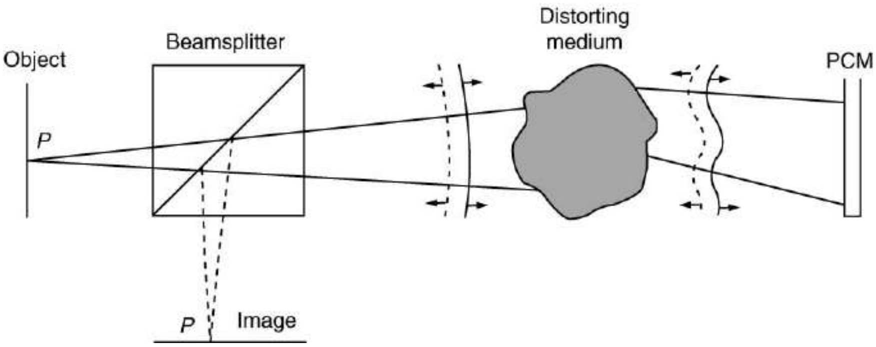

Figure 5 A basic two-pass system for imaging and distortion compensation using optical phase conjugation.

as parametric oscillation, optical tracking and pointing, spatial and temporal image processing, optical computing, optical filtering, etc. The PCM is also successfully implemented for aberration correction in high-power laser resonators.

Nondegenerate Four-Wave Mixing

If the frequencies of the impinging fields build a nonlinear polarization with an oscillating term close to absorption (or emission) frequencies of the sample, the nonlinear susceptibility increases tremendously. The radiated field amplitude changes by orders of magnitude and allows spectroscopic studies of the sample. The most popular process used to perform gas spectroscopy is CARS, an acronym for coherent anti-Stokes Raman scattering (or spectroscopy). In this case, the three incoming beams have different frequencies and the phase matching condition is eventually fulfilled in the 3D space.

Usually, two beams at the same frequency ω_P (pump beams), are mixed with a third beam at a tunable frequency ω_S, according to the energy scheme in Figure 6. As the radiated frequency ω_{CARS} is higher, the Raman spectroscopic notation is used and the subscripts 'S' hold for 'Stokes' and 'P' for 'anti-Stokes'.

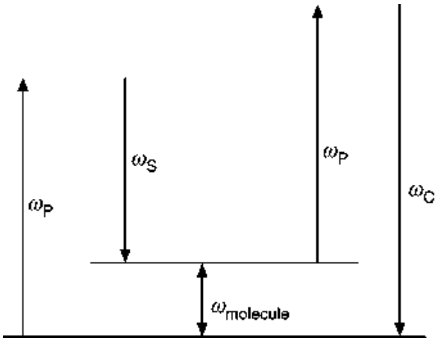

Figure 6 Energy diagram for a CARS experiment. Four-wave mixing efficiency increases when the frequency difference $\omega_P - \omega_S$ is close to a resonant frequency of the sample, $\omega_{molecule}$.

Whenever $\omega_P - \omega_S = \omega_{molecule}$, then an increase of the CARS intensity by many orders of magnitude is observed. Therefore, a Raman spectrum can be obtained by continuously changing, the Stokes laser and simultaneously recording the intensity of the CARS beam. This procedure is called scanning CARS, an interesting technique insofar as it allows taking high-resolution spectra, as well as measuring temperature of the sample while scanning over the Boltzmann distribution of the electronic ground state.

Figure 7 displays a classical geometrical arrangement where the phase matching condition fully determines the directionality of the radiated field at ω_{CARS}. Although the alignment of the setup can become tedious, this technique has been widely used, even in hostile industrial environments. The coherent aspect of this four-wave mixing process, imbedded in the phase matching condition, has a huge benefit over classical spectroscopic techniques and thus isotropic techniques, such as laser-induced fluorescence or Raman spectroscopy.

The CARS process is an interference process comparable to diffraction of a grating. The two fields at ω_P and ω_S form a laser-induced moving grating and the third field, at ω_P, undergoes a Doppler shift at ω_{CARS}. This scattered field undergoes a coherent amplitude addition in the Bragg direction.

Conclusion

In this article, we have discussed a set of optical nonlinear processes occurring at the third order of the development in series of the material polarization. At this order, wave mixing obviously shows most

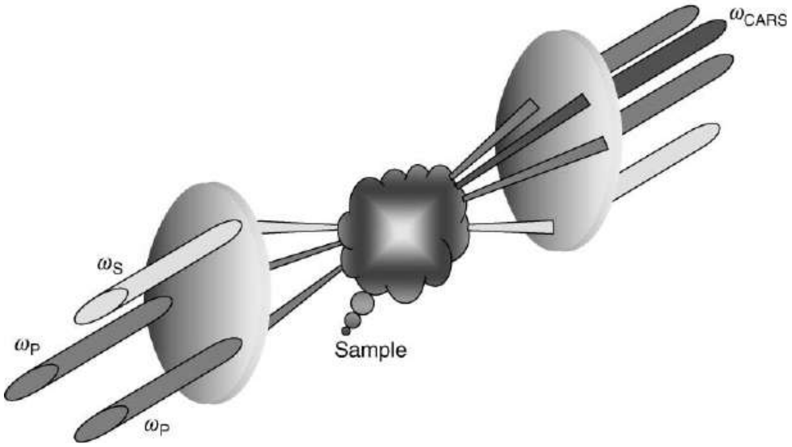

Figure 7 Layout of the CARS beams near a sample (a flame here). This particular arrangement is called folded (or 3D) Boxcars.

significant behavior and has been extensively studied and applied in many scientific areas. The general framework presented here can be further investigated using some books listed in Further Reading below.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources:** Harmonic Generation in Gases. **Optical Parametric Devices:** Optical Parametric Oscillators (Continuous Wave); Optical Parametric Oscillators (Pulsed).

Further Reading

Bloembergen N (1996) *Nonlinear Optics*. Singapore: World Scientific Pub Co.

Bloom DM and Bjorklund GC (1977) Conjugate wave-front generation and image reconstruction by four-wave mixing. *Applied Physics Letters* 31: 592–594.

Born M and Wolf E (1999) *Principles of Optics*. Cambridge, UK: Cambridge University Press.

Butcher PN and Cotter D (1991) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Eaton DF (1991) Nonlinear optical materials. *Science* 253: 281–287.

Fisher RA (1984) *Optical Phase Conjugation*. San Diego, CA: Academic.

Flytzanis C (1975) Theory of nonlinear optical susceptibilities. In: Rabin H and Tang CL (eds) *Quantum Electronics*, vol. 1, pp. 9–207. New York: Academic Press.

Hellwarth RW (1977) Generation of time-reversed wave fronts by nonlinear refraction. *Journal of the Optical Society of America* 67: 1–3.

Jackson JD (1998) *Classical Electrodynamics*. New York: Wiley.

Marburger JH (1975) Self-focusing: Theory. *Progress in Quantum Electronics* 4: 35–110.

Mittra R and Habashy TM (1984) Theory of wave-front-distortion correction by phase conjugation. *Journal of the Optical Society of America A* 1: 1103–1109.

Rockwell DA (1988) A review of phase-conjugate solid-state lasers. *IEEE Journal of Quantum Electronics* 24: 1124–1140.

Shen YR (2002) *The Principles of Nonlinear Optics*. New York: Wiley.

Yariv A (1977) Compensation for atmospheric degradation of optical beam transmission. *Optical Communication* 21: 49–50.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: Wiley.

Yariv A and Pepper DM (1977) Amplified reflection, phase conjugation, and oscillation in degenerate four-wave mixing. *Optical Letters* 1: 16–18.

Zel'dovich BYa, Pilipetsky NF and Shkunov VV (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Kramers–Krönig Relations in Nonlinear Optics

M Sheik-Bahae, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Since their introduction nearly 75 years ago, the Kramers–Krönig (KK) dispersion relations have been widely appreciated and applied in the analysis of linear optical systems. Because they are a consequence of strict causality, the KK relations apply not only to optical systems, but also to any linear, causal system such as electrical networks and particle scattering. In this article, we review the formulation and application of these relations in nonlinear optical systems. Simple logical arguments are used to derive dispersion relations that relate the nonlinear absorption coefficient to the nonlinear refraction coefficient. More general formalisms are then derived that apply to all nonlinear susceptibilities including the harmonic generating cases. Examples of recent successful application of these dispersion relations in analyzing various nonlinear materials will be presented.

The mathematical formalism of the KK dispersion relations in nonlinear optics was studied in the formative days of the field. The great usefulness of these relations was appreciated only recently, however, when they were used to derive the dispersion of the optical Kerr effect in solids from the corresponding nonlinear absorption coefficients, including two-photon absorption.

Before examining the details of KK relations in nonlinear optical systems, it is instructive to revisit the linear dispersion relations and their derivation based on the logic of causality. We will begin this task by introducing the definition of the linear as well as nonlinear susceptibilities $\chi^{(n)}$. In most nonlinear optics texts, the total material polarization (P) that drives the wave equation for the electric field (E) is expressed as

$$\begin{aligned}P_i(t) = {} & \varepsilon_0 \int_{-\infty}^{\infty} R_{ij}^{(1)}(t-t_1)E_j(t_1)\mathrm{d}t_1 + \varepsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \\ & \times R_{ijk}^{(2)}(t-t_1, t-t_2)E_j(t_1)E_k(t_2)\mathrm{d}t_1\,\mathrm{d}t_2 \\ & + \varepsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_{ijkl}^{(3)}(t-t_1, t-t_2, t-t_3) \\ & \times E_j(t_1)E_k(t_2)E_l(t_3)\mathrm{d}t_1\,\mathrm{d}t_2\,\mathrm{d}t_3 + \cdots \qquad [1]\end{aligned}$$

where $R^{(n)}$ is defined as the nth-order, time-dependent response function or time-dependent susceptibility. The subscripts are polarization indices indicating, in general, the tensor nature of the interactions. The summation over the various indices $j, k, l, \ldots$ is implied for the various tensor elements of $R^{(n)}$. Upon Fourier transformation, we obtain:

$$\begin{aligned}\vec{P}_i(\omega) = {} & \varepsilon_0 \int_{\infty}^{-\infty} \mathrm{d}\omega_1\, \chi_{ij}^{(1)}(\omega_1)\vec{E}_j(\omega_1)\delta(\omega-\omega_1) \\ & + \varepsilon_0 \int_{\infty}^{-\infty} \mathrm{d}\omega_1 \int_{\infty}^{-\infty} \mathrm{d}\omega_2 \\ & \times \chi_{ijk}^{(2)}(\omega_1, \omega_2)\vec{E}_j(\omega_1)\vec{E}_k(\omega_2)\delta(\omega-\omega_1-\omega_2) \\ & + \varepsilon_0 \int_{\infty}^{-\infty} \mathrm{d}\omega_1 \int_{\infty}^{-\infty} \mathrm{d}\omega_2 \int_{\infty}^{-\infty} \mathrm{d}\omega_3 \\ & \times \chi_{ijkl}^{(3)}(\omega_1, \omega_2, \omega_3)\vec{E}_j(\omega_1)\vec{E}_k(\omega_2)\vec{E}_l(\omega_3) \\ & \times \delta(\omega-\omega_1-\omega_2-\omega_3), \cdots \qquad [2]\end{aligned}$$

where δ is the Dirac delta-function. Here the $E(\omega)$ are Fourier transforms of the corresponding electric field. The nth-order susceptibility is defined as the Fourier transform of the nth-order response function:

$$\begin{aligned}\chi_{ijk\ldots n}^{(n)}(\omega_1, \omega_2, \ldots, \omega_m) = {} & \int_{-\infty}^{+\infty} \mathrm{d}\tau_1 \int_{-\infty}^{+\infty} \mathrm{d}\tau_2 \cdots \int_{-\infty}^{+\infty} \\ & \times \mathrm{d}\tau_n R_{ijk\ldots m}^{(n)}(\tau_1, \tau_2, \ldots, \tau_m)\mathrm{e}^{\mathrm{i}(\omega_1\tau_1+\omega_2\tau_2+\cdots+\omega_m\tau_m)} \qquad [3]\end{aligned}$$

For simplicity, we drop the polarization indices $i, j, \ldots$, and thus ignore the tensor properties of $\chi^{(n)}$ as well as the vector nature of the electric fields.

Let us for the moment concentrate on the linear polarization alone and derive the linear KK relations for the first-order susceptibility $\chi^{(1)}(\omega)$. For this, we rewrite eqn [3] for $n = 1$:

$$\chi^{(1)}(\omega) = \int_{-\infty}^{\infty} R^{(1)}(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau \qquad [4]$$

(As defined above, $\chi^{(1)}(\omega)$ and $R^{(1)}(\tau)$ are not a strict Fourier transform pair because of a missing factor of 2π.) Causality means that the effect cannot precede the cause. This can be restated mathematically as:

$$R^{(1)}(t) = R^{(1)}(t)\Theta(t) \qquad [5]$$

i.e., the response to an impulse at $t = 0$ must be zero for $t < 0$. Here $\Theta(t)$ is the Heaviside step function defined as $\Theta(t) = 1$ for $t > 0$ and $\Theta(t) = 0$ for $t < 0$. Upon Fourier transforming this equation, the product in the time domain becomes a convolution integral in frequency space

$$\begin{aligned}\chi^{(1)}(\omega) &= \chi^{(1)}(\omega)\left[\frac{\delta(\omega)}{2} + \frac{\mathrm{i}}{2\pi\omega}\right] \\ &= \frac{\chi^{(1)}(\omega)}{2} + \frac{\mathrm{i}}{2\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(1)}(\omega')}{\omega-\omega'}\mathrm{d}\omega' \\ &= \frac{1}{\mathrm{i}\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(1)}(\omega')}{\omega'-\omega}\mathrm{d}\omega' \qquad [6]\end{aligned}$$

which is the KK relation for the linear optical susceptibility. The symbol $\wp$ stands for the Cauchy principal value of the integral. The KK relation is thus a restatement of the causality condition [5] in the frequency domain. Taking the real part we have,

$$\Re e\{\chi^{(1)}(\omega)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(1)}(\omega')\}}{\omega'-\omega}\mathrm{d}\omega' \qquad [7]$$

Taking the imaginary part of eqn [6] leads to a similar relation relating the imaginary part to an integral involving the real part. It is conventional to write the optical dispersion relations in terms of the more familiar quantities of refractive index, $n(\omega)$, and absorption coefficient, $\alpha(\omega)$. For $|\chi^{(1)}| \ll 1$ then $n - 1 = \Re e\{\chi^{(1)}\}/2$ and $\alpha = \omega\Im m\{\chi^{(1)}\}/c$, and eqn [7] is transformed into

$$n(\omega) - 1 = \frac{c}{\pi}\wp\int_{0}^{\infty}\frac{\alpha(\omega')}{\omega'2 - \omega^2}\mathrm{d}\omega' \qquad [8]$$

where we additionally used the reality conditions of $n(\omega) = n(-\omega)$, and $\alpha(\omega) = \alpha(-\omega)$ to change the lower integral limit to 0. More rigorous analysis shows that eqn [8] is general and valid for any value of $|\chi^{(1)}|$. Although the KK dispersion relations and the extent of their applications in linear optics are well understood, some confusion sometimes exists about their applications to nonlinear optics. Causality clearly holds for both linear and nonlinear systems. The question is: what form do the resulting dispersion relations take in a nonlinear system? The linear Kramers–Krönig relations were derived from linear system theory, so it would appear to be impossible to apply the same logic to a nonlinear system. The key insight is that one can linearize the system. This is illustrated in Figure 1 where a linear

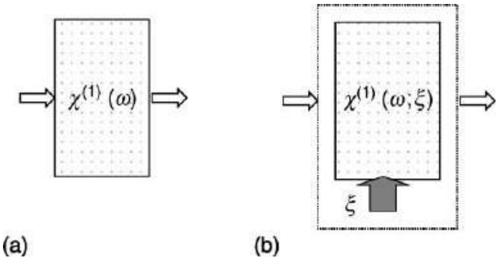

Figure 1 (a) A causal linear system obeying KK relations. (b) The system in (a) when externally perturbed by ξ. The dotted box now represents our new linear causal system whose altered $\chi^{(1)}$ obeys the KK relations.

(and of course, causal) optical material is transformed into a 'new' linear system that now contains the material and an external perturbation denoted by ξ. Although we are interested in perturbations of an optical nature, this formalism is general under any type of perturbation. It is important to appreciate the fact that our new system is causal even in the presence of the perturbation. This allows us to write down a modified form of the Kramers–Krönig relation linking the index of refraction to the absorption:

$$[n(\omega) + \Delta n(\omega;\zeta)] - 1 = \frac{c}{\pi}\wp\int_0^\infty \frac{\alpha(\omega') + \Delta\alpha(\omega';\zeta)}{\omega'^2 - \omega^2}\mathrm{d}\omega' \qquad [9]$$

which, after subtracting the linear relation between n and α leaves a relation between the changes in index and absorption:

$$\Delta n(\omega;\zeta) = \frac{c}{\pi}\wp\int_0^\infty \frac{\Delta\alpha(\omega';\zeta)}{\omega'^2 - \omega^2}\mathrm{d}\omega' \qquad [10]$$

where ζ denotes a general perturbation. An equivalent relation also exists whereby the change in absorption coefficient can be calculated from the change in the refractive index. It is essential that the perturbation be independent of frequency of observation, ω', in the integral (i.e., the excitation ζ must be held constant as ω' is varied).

Equation [10] has been used to determine refractive changes due to 'real' excitations such as thermal and free-carrier nonlinearities in semiconductors. In those cases, ζ denotes either ΔT (change of temperature) or ΔN (change of free-carrier density), respectively. In the former case, one calculates the refractive index change resulting from a thermally excited electron–hole plasma and the temperature shift of the band edge. For cases where an electron–hole plasma is injected (e.g., optically), the change of absorption gives the plasma contribution to the refractive index. In this case, the ζ parameter in eqn [10] is taken as the change in plasma density regardless of the mechanism of generation or the optical frequency.

Let us now extend this formalism to the case where the perturbation is virtual, occurring at an excitation frequency Ω that is below any material resonance. To the lowest order in the excitation irradiance I_Ω, we write

$$\Delta\alpha(\omega;\zeta) = \Delta\alpha(\omega;\Omega) = 2\alpha_2(\omega;\Omega)I_\Omega \qquad [11]$$

and

$$\Delta n(\omega;\zeta) = \Delta n(\omega;\Omega) = 2n_2(\omega;\Omega)I_\Omega \qquad [12]$$

where n_2 and α_2 are the nonlinear refractive index and absorption coefficients of the material, respectively. By definition, these coefficients are related to the third-order nonlinear susceptibility $\chi^{(3)}(\omega_1,\omega_2,\omega_3)$ via (*see* **Nonlinear Optics, Basics**: Nomenclature and Units)

$$n_2(\omega;\Omega) = \frac{3}{4\varepsilon_0 n_0(\omega)n_0(\Omega)c}\Re e\{\chi^{(3)}(\omega,-\Omega,\Omega)\} \qquad [13]$$

and

$$\alpha_2(\omega;\Omega) = \frac{3\omega_a}{2\varepsilon_0 n_0(\omega)n_0(\Omega)c^2}\Im m\{\chi^{(3)}(\omega,-\Omega,\Omega)\} \qquad [14]$$

We can therefore write the dispersion relations between α_2 and n_2:

$$n_2(\omega;\Omega) = \frac{c}{\pi}\wp\int_0^\infty \frac{\alpha_2(\omega';\Omega)}{\omega'^2 - \Omega^2}\mathrm{d}\omega' \qquad [15]$$

Note that even when the degenerate $n_2(\omega) = n_2(\omega;\omega)$ is desired (at a given ω), the dispersion relation requires that we should know the nondegenerate absorption spectrum $\alpha_2(\omega';\omega)$ at all frequencies ω'.

Let us pause here and discuss some physical mechanisms that can be involved for a given system of interest. Consider a material characterized by an optical resonance occurring at, say ω_0 (i.e., a degenerate two-level system). For a solid, this resonance can be regarded as that of the fundamental energy gap; $\omega_0 = \omega_g = E_g/\hbar$ in a two-band system. Now, let us examine how the presence of an optical excitation at $\Omega < \omega_0$ can alter the absorption spectrum (at a variable probe ω'). In the quantum mechanical picture, this gives rise to a 'new' material whose perturbed wave functions are 'dressed' by the intensity and frequency of the applied optical field.

The lowest-order correction to the absorption is given by $\alpha_2(\omega';\Omega)$ which involves three major physical processes. Recalling that $\Omega < \omega_0$, these processes include (1) two-photon absorption (2PA) when $\omega' + \Omega \rightarrow \omega_0$ and (2) Raman-induced absorption when $\omega' - \Omega \rightarrow \omega_0$, both implying an absorption of a photon at the probe frequency ω' (i.e., $\alpha_2 > 0$). The third process can be identified as resulting from the blue-shift (for $\Omega < \omega_0$) of the resonance (known as the quadratic optical Stark effect) caused by the excitation field. For our two-level system, the latter results in a decrease followed by an increase in absorption in the vicinity of ω_0. An example of the overall absorption changes due to such processes is shown in Figure 2 where $\alpha_2(\omega';\Omega)$ is qualitatively plotted for a degenerate two-level system. We should note that the relative magnitude of each contribution as well as the width and shape of the resonances are chosen arbitrarily for the purpose of illustration. Using the KK relation in eqn [15], we can now arrive at the nonlinear index coefficient $n_2(\omega;\Omega)$. The result of this transformation is also given in Figure 2. The above simple example elucidates the key concepts involving the relationship between nonlinear absorption and refraction in materials for third-order processes. These concepts, when applied more rigorously to semiconductors, have been successful in predicting the sign, magnitude, and dispersion of n_2 due to the anharmonic motion of bound electrons. This will be briefly discussed later. Returning to the mathematical foundation of KK relations, we use eqns [13] and [14] to write eqn [15] in terms of the nonlinear susceptibility $\chi^{(3)}$:

$$\Re e\{\chi^{(3)}(\omega_1, \omega_2, -\omega_2)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty} \frac{\Im m\{\chi^{(3)}(\omega', \omega_2, \omega_2)\}}{\omega' - \omega_1}\,\mathrm{d}\omega' \qquad [16]$$

The above dispersion relation for $\chi^{(3)}$ was obtained using the physical and intuitive arguments that followed the linearization scheme depicted in Figure 1. General dispersion relations can be formulated following a mathematical procedure that is similar to the derivation of the linear KK relations. In this case we apply the causality condition directly to the nth-order nonlinear response $R^{(n)}$. For example, without loss of generality, we can write

$$R^{(n)}(\tau_1, \tau_2, \ldots, \tau_n) = R^{(n)}(\tau_1, \tau_2, \ldots, \tau_n)\Theta(\tau_j) \qquad [17]$$

and then calculate the Fourier transform of this equation. Here j can apply to any one of the indices $1, 2, \ldots., n$. Following the same procedure as for a linear response, we obtain

$$\chi^{(n)}(\omega_1, \omega_2, \ldots, \omega_j, \ldots, \omega_n) = \frac{-\mathrm{i}}{\pi}\wp\int_{-\infty}^{\infty} \frac{\chi^{(n)}(\omega_1, \omega_2, \ldots, \omega', \ldots, \omega_n)}{\omega_j - \omega'}\,\mathrm{d}\omega' \qquad [18]$$

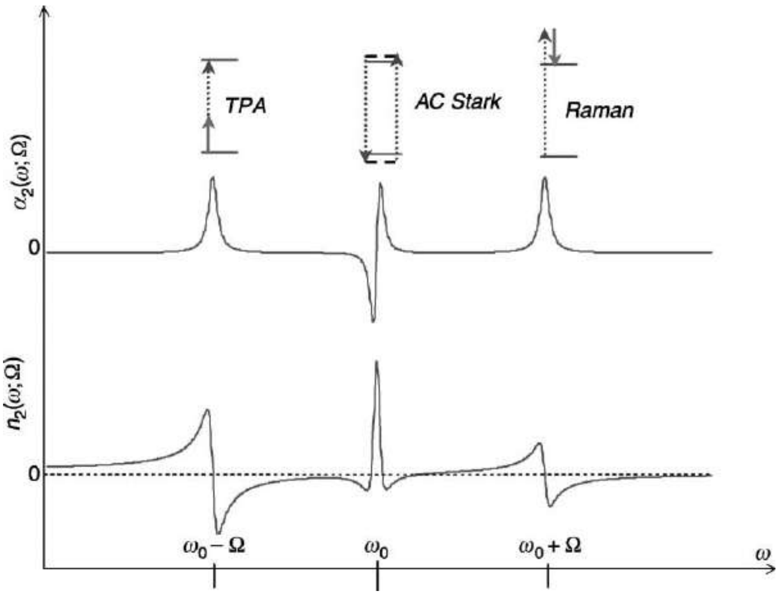

Figure 2 Upper trace: the nonlinear absorption coefficient in a fictitious 'degenerate' two-level system. Lower trace: the resulting nonlinear refractive index obtained using the KK relations. The insets show the three possible physical mechanisms involved.

By separating the real and imaginary parts of this equation, we get the generalized Kramers–Krönig relation pairs for a nondegenerate, nth-order nonlinear susceptibility:

$$\Re e\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega_j,\ldots,\omega_n)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega',\ldots,\omega_n)\}}{\omega'-\omega_j}\mathrm{d}\omega' \qquad [19]$$

and

$$\Im m\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega_j,\ldots,\omega_n)\} = -\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Re e\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega',\ldots,\omega_n)\}}{\omega'-\omega_j}\mathrm{d}\omega' \qquad [20]$$

In particular, for $\chi^{(3)}$ processes having $\omega_1=\omega_a,\omega_2\omega_b$, and ${}_3b$, this becomes identical to eqn [16].

Note that in describing the nonlinear susceptibilities, no special attention was given to the harmonic generating susceptibility $\chi^{(N)}(N\omega)\equiv\chi^{(N)}(\omega,\omega,\ldots\omega)$, i.e., the susceptibility generating the Nth harmonic at $N\omega$. It turns out that in addition to the KK relations given by eqns [19] and [20], the real and imaginary parts of $\chi^{(N)}(N\omega)$ can also be related in different sets of dispersion integrals that involve only the degenerate forms of the susceptibilities. A more general yet simple analysis gives the most general form of KK relations for any type of $\chi^{(n)}$:

$$\chi^{(n)}(\omega_1+p_1\omega,\omega_2+p_2\omega,\ldots,\omega_m+p_m\omega) = \frac{1}{\mathrm{i}\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(n)}(\omega_1+p_1\Omega,\omega_2+p_2\Omega,\ldots,\omega_m+p_m\Omega)}{\Omega-\omega}\mathrm{d}\Omega \qquad [21]$$

for all $p_1,p_2,\ldots,p_m\geq 0$. Setting $\omega_1=\omega_2=\cdots=\omega_m\equiv 0$, and $p_1=p_2=\cdots=p_m=1$ in eqn [21] yields an interesting form of the KK relations for the Nth-harmonic susceptibilities:

$$\Re e\{\chi^{(N)}(N\omega)\}=\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(N)}(N\omega')\}}{\omega'-\omega}\mathrm{d}\omega' \qquad [22]$$

These dispersion relations have allowed calculations of $\chi^{(2)}(2\omega)$ and $\chi^{(3)}(3\omega)$ in semiconductors using full band structures.

At the beginning of this article, it was noted that all the KK relations for nonlinear optics were known in the early days of the field. Their application in unifying nonlinear absorption (in particular two-photon absorption) and the optical Kerr effect (n_2) in solids came only much later. More recent work demonstrated that the KK relations are a powerful analytical tool in nonlinear optics. Following the picture of a degenerate two-level system shown in Figure 2, a simple two-band model has been used to calculate the nonlinear absorption coefficient, $\alpha_2(\omega_1;\omega_2)$, resulting from three mechanisms: 2PA, the Raman absorption process, and the ac Stark effect. The optical Kerr coefficient $n_2(\omega_1;\omega_2)$ was then calculated using eqn [15]. Of particular practical interest is the degenerate case ($\omega_1=\omega_2=\omega$), from which the 2PA coefficient $\beta(\omega)=\alpha_2(\omega;\omega)$ can be extracted. Figure 3 depicts the calculated dispersion of n_2 and β as a function of $\hbar\omega/E_g$ where E_g is the bandgap energy of the solid. The dispersion of n_2 and its sign reversal shown in Figure 3 has been observed experimentally in many optical solids.

Finally, let us discuss a related implication of causality in nonlinear optics. The KK dispersion relations are traditionally derived in terms of internal material parameters such as susceptibility, absorption coefficient, and refractive index. Similar to the case of electrical circuits, one can obtain dispersion relations that apply to an external transfer function of the system that relates an input signal to an output signal. In this case, the dispersion of the transfer function includes system structure as well as the intrinsic dispersion of the material. As an optical (and linear) example, consider a Fabry–Perot etalon. The optical transmission of this system has well-known spectral

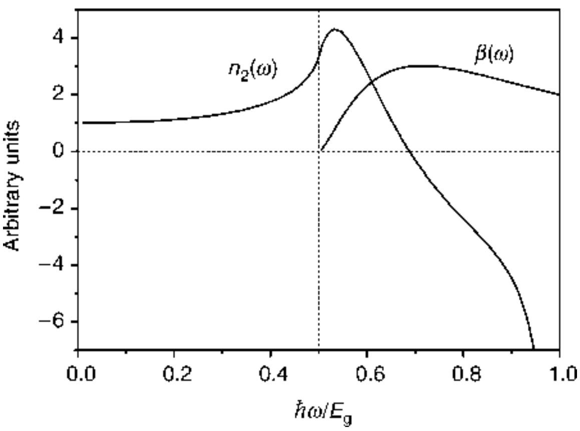

Figure 3 The two-photon absorption coefficient in semiconductors (β) calculated for a two-band model. The resultant nonlinear refractive index (n_2) obtained using a KK transformation of the calculated nondegenerate nonlinear absorption coefficient includes all major mechanisms.

features that are primarily caused by structural dispersion (i.e., interference) in addition to the intrinsic dispersion of the material. Causality still demands that the transmitted signal has a phase variation whose value and dispersion can be determined using a KK relation linking the real and imaginary parts of the transmission coefficient. In other words, the KK relations provide a spectral correlation between the real and imaginary components of the transfer function which in turn may translate to a spectral correlation between the phase and amplitude of the transmitted signal. However, the variations in phase do not necessarily imply the presence of a varying index of refraction, nor does an amplitude variation suggest the existence of material absorption (dissipation). Ultimately, this implies that any mechanism causing a variation in amplitude (including reflection, scattering, or absorption) must be accompanied by a phase variation. (One should note that the reverse of the previous statement is not necessarily true; i.e., a variation in phase does not have to be accompanied by an amplitude modulation.)

In nonlinear optics with the 'black box' approach of Figure 1, the optical perturbation ξ (with frequency Ω) can render an amplitude variation in the probe (at ω) using various frequency mixing schemes in a noncentrosymmetric material (i.e., with nonzero $\chi^{(2)}$). For instance, the probe at ω can be depleted by nonlinear conversion to $\omega_{\text{sum}} = \omega + \Omega$ via sum-frequency generation involving $\chi^{(2)}(\omega, \Omega)$ and/or to $\omega_{\text{diff}} = \omega - \Omega$ via difference-frequency generation involving $\chi^{(2)}(\omega, -\Omega)$. Such a conversion (or depletion) should be accompanied by a phase variation according to the KK dispersion relations. This type of nonlinear phase modulation is known as a $\chi^{(2)}:\chi^{(2)}$ cascaded nonlinearity. Such cascaded processes are routinely (and more simply) analyzed with Maxwell's equations governing the propagation of beams in a second-order nonlinear material. The KK relations, however, provide an interesting physical perspective of the process. We find that cascaded second-order nonlinearities are yet another manifestation of causality in nonlinear optics.

List of Units and Nomenclature

α	linear absorption coefficient
α_2	nonlinear absorption coefficient
β	two-photon absorption coefficient
$\chi^{(n)}$	nth-order nonlinear optical susceptibility
n	linear refractive index
n_2	nonlinear refractive index coefficient, coefficient of optical Kerr effect
$\Theta(t)$	step function
$\wp$	principal value
2PA	two-photon absorption
KK relations	Kramers–Krönig relations

See also

Materials for Nonlinear Optics: Liquid Crystals for NLO. **Nonlinear Optics, Basics:** Cascading; Nomenclature and Units.

Further Reading

Bassani F and Scandolo S (1991) Dispersion-relations and sum-rules in nonlinear optics. *Physical Review B–Condensed Matter* 44: 8446–8453.

Caspers PJ (1964) Dispersion relations for nonlinear optics. *Physical Review A* 133: 1249–1251.

Hutchings DC, Sheik-Bahae M, Hagan DJ and Van Stryland EW (1992) Kramers–Krönig relations in nonlinear optics. *Optical and Quantum Electronics* 24: 1–30.

Kogan SM (1963) On the electromagnetics of weakly nonlinear media. *Soviet Physics JETP* 16: 217–219.

Nussenzveig HM (1972) Causality and Dispersion Relations. New York: Academic Press.

Price PJ (1964) Theory of quadratic response functions. *Physical Review* 130: 1792–1797.

Ridener FLJ and Good RHJ (1975) Dispersion relations for nonlinear systems of arbitrary degree. *Physical Review B* 11: 2768–2770.

Sheik-Bahae M and Van Stryland EW (1999) Optical nonlinearities in the transparency region of bulk semiconductors. In: Garmire E and Kost A (eds) *Nonlinear Optics in Semiconductors I*, 58, pp. 257–318. Academic Press.

Sheik-Bahae M, Hutchings DC, Hagan DJ and Van Stryland EW (1991) Dispersion of bound electronic nonlinear refraction in solids. *IEEE Journal of Quantum Electronics* 27: 1296–1309.

Smet F and Vangroenendael A (1979) Dispersion-relations for *N*-order non-linear phenomena. *Physical Review A* 19: 334–337.

Toll JS (1956) Causality and the dispersion relation: logical foundations. *Physical Review* 104: 1760–1770.

Nomenclature and Units

M P Hasselbeck, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Nomenclature Associated with the Excitation Light

Particular care must be used when characterizing the excitation beam in nonlinear optical experiments compared to linear measurements. By definition, nonlinear optical phenomena depend on the electric field to high order. The higher the order, the more sensitive the observed behavior depends on the input. Extracting a representative nonlinear coefficient from an experiment, for example, becomes progressively more difficult as the order of the optical nonlinearity increases. In other words: the errors associated with optical beam characterization get magnified by the order of the nonlinearity under study.

There is an extensive nomenclature for characterizing the light (almost always laser light) interacting with the nonlinear optical medium. Different descriptions may be used depending on whether the excitation light is pulsed, continuous, or a continuous train of pulses. When a laser beam is constant or continuous, it is often described by its 'cw power'. 'Cw' stands for 'continuous wave', an acronym taken from the nomenclature of electronics. The cw power of a laser is determined by placing a power meter in the path of a beam. Repetitively pulsed lasers can be characterized in the same way, provided the response time of the power meter is slower than the pulse separation period. This will almost certainly be the case with a continuously pumped mode-locked laser such as a dye laser, fiber laser, or Ti:sapphire laser, which produce pulses at repetition frequencies of tens of megahertz. Cw power may also be suitable for describing pulsed flashlamp lasers, gas discharge lasers, or any laser that is excited in a periodic fashion.

When the optical output can be distinguished as individual pulses, additional metrics are used. A pulse, by definition, exists during a window of time, i.e., there is a time when light is present and a time when light is absent. How one characterizes the time light is present – the pulse duration – is critically important and not always obvious. There are a myriad of ways the pulse temporal envelope can manifest itself; common examples include square pulses, triangular pulses, Gaussian pulses, and hyperbolic secant pulses. These names refer to the mathematical waveforms that map the pulse envelope as a function of time.

A pulse waveform that exhibits symmetry about the peak in its temporal envelope is commonly characterized by its 'full-width, half-maximum', abbreviated FWHM. One obtains the full-width, half-maximum by locating the two points on the pulse profile that are at half the peak value. The temporal separation of these two points is the FWHM. One uses this measure because in a strict mathematical sense, waveforms such as the Gaussian or hyperbolic-secant exist even at times $t = \pm\infty$. A less common term is the 'half-width, half-maximum' or HWHM. As the name implies, the HWHM is exactly half the FWHM value. Not all light-pulses are temporally symmetric, however, so greater detail may be needed when giving a mathematical description of asymmetric pulses.

Light pulses are also characterized by their energy and peak power. The temporal envelope recorded by a so-called square-law detector (all laboratory detectors are square-law detectors) shows the power of the pulse as a function of time, provided the detector response time is sufficiently fast. Integrating this waveform gives the pulse energy. It makes no sense to talk about the energy of a cw beam of light, unless that beam is composed of repetitive, distinguishable pulses. Each pulse in the train carries a distinct amount of optical energy.

'Ultrafast' or 'ultrashort' laser pulses generally refer to light pulses that are of such small duration they cannot be measured directly with detectors and oscilloscopes because of bandwidth limitations. To infer the pulse duration, so-called intensity autocorrelation measurements are often made. The temporal power profile is then deduced from the autocorrelation signal with the appropriate mathematical conversion factor. Equally important is the spectrum of a short pulse. The optical bandwidth determines the lower limit of temporal compression; such ideally compressed pulses are said to be 'transform limited'. If different components of the spectrum can be distinguished at specific positions on the pulse temporal profile, then the pulse is 'chirped'. Information on the methods of ultrafast laser science can be found in **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Nonlinear optical effects take place when light is concentrated on a target – there is a cross-section or 'footprint' of the beam on the medium.

The interaction of the light and material takes place in a region called the beam area, beam cross-section, focal area, or focal volume. One must then define the appropriate interaction area or volume.

Specification of the beam area allows one to define two important quantities used frequently in nonlinear optics. The power/area is known as irradiance or power density. This quantity is commonly called the intensity, although the strict radiometric definition of intensity is power/(solid angle). The second important quantity is energy/area, which is called the energy density or fluence.

The convention in nonlinear optics is to define the optical electric field as:

$$\mathbf{E}(\mathbf{z},t) = \frac{1}{2}E_0 \exp[\mathrm{i}(\mathbf{k}\cdot\mathbf{z} - \omega t)] + \text{c.c.} \qquad [1]$$

where E_0 is the peak field, $\mathbf{k}$ is the propagation vector, ω is the angular frequency of the light, and c.c. stands for complex conjugate. This represents a forward- and backward-propagating (in the z-direction) infinite, transverse plane wave. Other definitions are also used. In SI/mks units and using the convention of eqn [1], the irradiance (I in units of W/m^2) inside a material of refractive index n is related to the electric field vector of the light ($\mathbf{E}$ in units of V/m) as follows:

$$I = \frac{1}{2}cn\varepsilon_0|\mathbf{E}|^2 \qquad [2]$$

Here c is the speed of light (2.998×10^8 m/s) and ε_0 is the permittivity of free space (8.854×10^{-12} F/m). Nonlinear optics has an unfortunate tradition of mixing mks and cgs units, resulting in a lot of confusion. Irradiance is usually expressed in units of W/cm^2. One can calculate the peak electric field (in units of V/cm) of a laser beam given its irradiance (in units of W/cm^2) by using this simple formula:

$$E_0 = 38.82\sqrt{\frac{I}{n}}\left(\frac{\mathrm{V}}{\mathrm{cm}}\right) \qquad [3]$$

The root-mean-square value of the electric field is obtained by replacing the factor 38.82 by 27.45. In Gaussian/cgs units, irradiance is expressed in units of ergs/(cm^2 sec) and is related to the field as:

$$I = \frac{cn}{8\pi}|\mathbf{E}|^2 \qquad [4]$$

where the field is in units of statvolt/cm and $c = 2.998 \times 10^{10}$ cm/sec. Conversion between mks and cgs for the electric field is 3×10^4 V/m = 1 statvolt/cm; irradiance is converted using 1 erg/(cm^{2}sec) = 1×10^{-3}W/m^2. The following point must be emphasized: depending on how the electric field is defined, there can be factors of 2 or even 4 discrepancies in the irradiance values quoted by different authors. The common definitions are used here, although they are certainly not universal.

A very common realization of the spatial irradiance profile of a laser beam is the radially symmetric Gaussian function:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \qquad [5]$$

where I_0 is the peak irradiance at the center of the beam ($r = 0$) and w is the 'spot size'. The spot size changes continuously as the beam propagates and the minimum spot size is known as the waist. One often hears the spot size referred to as the radius, but the radius of a Gaussian beam traditionally denotes the curvature of the phase front. By definition, the phase front radius is infinite at the waist. The factor of 2 appearing in the argument of the exponential in eqn [5] stems from the fact that the spot size w is conventionally defined for the electric field of the Gaussian beam. When the field is squared to obtain the irradiance, the factor of 2 appears. Using this function for the irradiance spatial profile leads to the common parameter '1/e^2 diameter'. The irradiance falls to 1/e^2 (0.1353) of its peak value when $r = w$.

The Gaussian spatial profile or 'Gaussian beam' results when the laser has been designed to operate in the lowest-order transverse mode, usually denoted TEM$_{00}$. The TEM$_{00}$ Gaussian beam is particularly convenient because the same relative power profile is maintained in the near- and far-field, whether or not the beam is collimated or focused.

The cross-sectional area of a TEM$_{00}$ Gaussian beam normally incident on a surface is found as follows. The power in the beam is obtained by integrating the irradiance profile (I) over the surface:

$$P = \iint_{\pm\infty} I\mathrm{d}A \qquad [6]$$

Referring to eqn [5], the integral becomes:

$$P = \int_0^{2\pi} \mathrm{d}\theta \int_0^{\infty} r\,\mathrm{d}r\, I_0 \exp\left(-\frac{2r^2}{w^2}\right) = I_0\frac{\pi w^2}{2} \qquad [7]$$

Hence the effective area of a TEM$_{00}$ Gaussian beam normally incident on a surface is:

$$A = \frac{\pi w^2}{2} \qquad [8]$$

There are many situations where the Gaussian formulation of eqn [5] is not suitable. Beams produced by unstable laser resonators, for example, are not Gaussian. A multitransverse mode beam profile obtained from a stable resonator does not usually lend itself to a simple mathematical characterization.

Nomenclature Associated with the Nonlinear Optical Medium

In nonlinear optics texts, the subject is often introduced by writing the macroscopic polarization in Maxwell's equations as a power series expansion in the electric field. This approach, proposed by Bloembergen and coworkers in the early 1960s, has been spectacularly successful for interpreting experiments, though it has also led to confusion in the definition of nonlinear optical coefficients. The confusion, which stems from arbitrary definitions and nomenclature used by different authors, cannot be resolved here. The reader should, however, be alert for these discrepancies. Comparison of results published by different laboratories requires that the fundamental equations for extracting the nonlinear coefficients from their data are known. As the discipline has matured, the nonlinear optics community largely recognized the ambiguities and confusion; precise definitions of terms and coefficients are now commonly provided in the research literature.

It should also be noted that the power-series framework may not always be the best formulation for modeling and characterizing nonlinear optical effects. A carefully designed optical limiter, for example, may exhibit an abrupt decrease of transmission at a specific 'threshold irradiance' or 'threshold fluence', perhaps at a given laser pulse duration. The physical phenomenon or phenomena driving this behavior may not readily succumb to characterization by an *n*th-order coefficient in a power series expansion. In this situation it may be appropriate to specify a threshold optical input parameter. Another example is a homogeneously broadened saturable absorber, in which the absorption of an optical medium decreases with increasing input irradiance. The irradiance-dependent absorption coefficient α (refer to eqn [26]) is characterized by a 'saturation irradiance' I_{sat}:

$$\alpha(I) = \frac{\alpha_0}{1 + \dfrac{I}{I_{\text{sat}}}} \qquad [9]$$

where α_0 is the linear absorption coefficient.

Nonlinear Susceptibility

The power-series expansion of the macroscopic polarization is the standard approach for modeling nonlinear optical behavior and categorizing the various phenomena. A common way to write the polarization, nonlinear in the electric field, is (mks units):

$$\mathbf{P} = \varepsilon_0\left(\chi^{(1)}\cdot\mathbf{E} + \chi^{(2)} : \mathbf{EE} + \chi^{(3)}\vdots\mathbf{EEE} + \ldots\right) \qquad [10]$$

where the polarization **P** is a time- and space-dependent vector and the terms $\chi^{(n)}$ are the various orders of the nonlinear susceptibility (the cgs equation is obtained by dropping the free-space permitivitty coefficient ε_0). A second-order effect is associated with $\chi^{(2)}$, a third-order with $\chi^{(3)}$, and so on. In general, there are distinguishable electric field vectors (in eqn [10], the fields have not been individually designated for clarity). The susceptibility terms are generally tensors, which means the medium is sensitive to the orientation of the input fields. The dots in eqn [10] indicate tensor products. The input field vectors can have different frequencies and the presence of complex conjugates in eqn [1] indicates that the frequency components will have both + and − signs, i.e. the *j*th electric field term will have associated frequency factors $\exp(\pm \mathrm{i}\omega_j t)$.

The nomenclature used here can be illustrated by example. In the case of the second-order nonlinear polarization (mks units):

$$\mathbf{P}_j = \varepsilon_0\chi^{(2)}_{jkl} : \mathbf{E}_k\mathbf{E}_l \qquad [11]$$

Note that subscripts have been introduced. The indices correspond to Cartesian space vectors. This equation says that the polarization in the *j*-direction results from the tensor product of the appropriate $\chi^{(2)}$ with the input fields at $\mathbf{E}_k$ and $\mathbf{E}_l$. The *j*th component of polarization has an oscillation frequency that is determined by mixing of the input fields. In this example, the indices *j*, *k* and *l* can each take the value of *x*, *y*, and *z*. Let each input field be at one of two possible frequencies, call them ω_1 and ω_2. There are an enormous number of permutations (81 to be exact) of eqn [11]. One term is:

$$P_x(\omega_1 + \omega_2) = \varepsilon_0\chi^{(2)}_{xzy}(\omega_1 + \omega_2)E_z(\omega_1)E_y(\omega_2) \times \exp[-\mathrm{i}(\omega_1 + \omega_2)t] \qquad [12]$$

Another possibility is:

$$P_z(\omega_2 - \omega_1) = \varepsilon_0\chi^{(2)}_{zyx}(\omega_2 - \omega_1)E_y(\omega_2)E_x(-\omega_1) \times \exp[-\mathrm{i}(\omega_2 - \omega_1)t] \qquad [13]$$

Equation [12] corresponds to sum frequency generation $(\omega_1 + \omega_2)$ and eqn [13] shows difference frequency generation $(\omega_2 - \omega_1)$.

Not explicitly written on the right-hand side of the above equations is the exponential containing the mixing of the propagation vectors. To maximize the nonlinear polarization, which is the source of nonlinear behavior in Maxwell's equations, the vector sum $\sum \mathbf{k}_j$ for *all* the interacting fields should be close to zero. Arranging the propagation vectors to accomplish this is known as 'phase matching'.

It is important to point out that there are eight more terms describing the nonlinear polarization at $P_x(\omega_1 + \omega_2)$ in addition to eqn [12]; likewise for $P_z(\omega_2 - \omega_1)$. Also note that the indices k and l may be identical, hence a nonlinear polarization driven by $\chi^{(2)}_{xyy}$ for example, is allowed. Higher-order nonlinear polarizations get progressively more complicated. The general third-order nonlinear polarization is

$$\mathbf{P}_j = \varepsilon_0 \chi^{(3)}_{jklm} \vdots \mathbf{E}_k \mathbf{E}_l \mathbf{E}_m \qquad [14]$$

Sometimes the second-order nonlinear susceptibility is written with the coefficient d instead of $\chi^{(2)}$. The relation between the d-coefficient and $\chi^{(2)}$ depends arbitrarily on how it is defined; a common definition is

$$d_{jkl} = \frac{1}{2}\chi^{(2)}_{jkl} \qquad [15]$$

but the reader is cautioned that the factor $\frac{1}{2}$ is sometimes missing. In the preceding discussion it was pointed out how there can, in principle, be a large number of tensor components involved in the expansion of the nonlinear polarization. Simplification occurs when it is realized that there is no discernible physical difference between the frequency terms ω_j and $-\omega_j$ and that ordering of the fields in eqn [14] – which suggests a time order in the arrival of the fields – is irrelevant. These symmetry arguments show that $d_{jkl} = d_{jlk}$, which reduces the number of independent d-coefficients from 81 to 18. A simpler subscript notation is then introduced that uses the integers 1–6 to represent pairs of Cartesian components k and l. This reduces the number of subscripts from three to two. An example of this notation is $d_{xzz} = d_{14}$.

In certain situations, one can take advantage of crystal symmetry to further reduce the complicated summations to a single scalar coefficient for the nonlinear susceptibility, which is referred to as 'd-effective'. For these specialized cases, the nonlinear polarization is

$$P = \varepsilon_0 d_{\text{eff}} E_1 E_2 \qquad [16]$$

In SI/mks units, the polarization is in units of coul/m^2. The second-order susceptibility (i.e. $\chi^{(2)}$, d_{jkl}) must therefore have units of m/volt. The units of $\chi^{(3)}$ are m^2/volt^2, $\chi^{(4)}$ is m^3/volt^3, and so on. In Gaussian/cgs units, $\chi^{(2)}$ has dimensions cm/statvolt, $\chi^{(3)}$ will be cm^2/statvolt^2, etc. although sometimes in the Gaussian system *all* the coefficients $\chi^{(n)}$ are discussed with shorthand 'electrostatic units' or 'esu'.

Complex Quantities

In linear optics, the susceptiblity has real and imaginary parts

$$\chi^{(1)} = \chi^{(1)}_{\text{real}} + \mathrm{i}\chi^{(1)}_{\text{imaginary}} \qquad [17]$$

The complex linear index is, in turn, written as the sum of real and imaginary components, derived from the linear susceptibility as follows (mks units):

$$\sqrt{1 + \chi^{(1)}} = n_0 + \mathrm{i}\kappa \qquad [18]$$

where n_0 is the linear refractive index and κ is the imaginary term leading to absorption of light. In general, the nonlinear susceptibilty $\chi^{(n)}$ is also a complex number:

$$\chi^{(n)} = \chi^{(n)}_{\text{real}} + \mathrm{i}\chi^{(n)}_{\text{imaginary}} \qquad [19]$$

As in linear optics, the complex notation is a bookkeeping method that conveniently accounts for what is known as 'resonant enhancement'. Nonlinear optics research has revealed that the nonlinear susceptibility $\chi^{(n)}$ can be a strong function of frequency. Specifically, this quantity will be strongly enhanced when sums and/or differences of the photon energies in the interacting light beams coincide with quantum mechanical energy resonances in the material. When a resonance condition is achieved, $\chi^{(n)}$ will be dominated by its imaginary component. Far from the resonances, $\chi^{(n)}$ behaves more like a real quantity. These complex terms directly enter the wave equations describing how light propagates in a nonlinear medium and are particularly important for $\chi^{(3)}$; the complex quantity $\chi^{(3)}$ determines whether light will be refracted or absorbed and by how much. The real part of $\chi^{(3)}$ drives nonlinear refraction while the imaginary portion characterizes nonlinear absorption (e.g., two-photon absorption) and the inverse effect – gain. This rich subject – in particular the critical importance of the wavelength dependence of the nonlinear phenomena – is considered in **Nonlinear Optics, Basics:** Kramers–Krönig Relations in Nonlinear Optics.

Nonlinear Refraction

The most common manifestation of nonlinear refraction arises from the third-order nonlinear susceptibility, the so-called optical Kerr effect. In this case, the refractive index is linearly proportional to the irradiance (I) of a monochromatic light beam. The irradiance-dependent refractive index is

$$n(I) = n_0 + n_2 I \qquad [20]$$

where n_0 is the linear, irradiance-independent refractive index and n_2 is the nonlinear refractive index coefficient. Because n is a dimensionless quantity, n_2 must be in units of area/power.

The optical Kerr effect is usually written in mks units with the expression shown in eqn [20]. In cgs units, the common (but by no means universal) convention is to write:

$$n(E) = n_0 + \frac{1}{2}\tilde{n}_2 |\mathbf{E}|^2 \qquad [21]$$

where the field is in units of statvolts/cm. The nonlinear coefficient $\tilde{n}_2$ (cgs) must therefore be in units of cm^2/statvolt^2, which is sometimes abbreviated to 'esu'. The conversion for n_2 between these two equations is:

$$\tilde{n}_2(\text{cgs}) = \frac{n_0 c}{40\pi} n_2(\text{mks}) \qquad [22]$$

where n_2(mks) is in units of m^2/W and c is in m/sec. Also useful is the relation between n_2 and $\chi^{(3)}$, written in mks units as:

$$\Delta n = n_2 I = \frac{\mathrm{Re}\left(\chi^{(3)}\right)}{4\varepsilon_0 n^2 c} I \qquad [23]$$

where 'Re' denotes the real part of $\chi^{(3)}$. In cgs units we have:

$$\Delta n = \frac{1}{2}\tilde{n}_2 |\mathbf{E}|^2 = \frac{4\pi^2 \mathrm{Re}\left(\tilde{\chi}^{(3)}\right)}{n^2 c} I \qquad [24]$$

where I is in units of ergs/(cm^2 sec) defined by eqn [4]. The reader is again advised that different authors will show discrepancies of 2, 4, or even 8 when writing these equations. The susceptibilities $\chi^{(3)}$ are related as follows:

$$\tilde{\chi}^{(3)}(\text{cgs}) = \frac{9 \times 10^8}{4\pi} \chi^{(3)}(\text{mks}) \qquad [25]$$

If the coefficients n_0 and n_2 exist, does the nomenclature imply there are terms n_1, n_3, and others? These coefficients are certainly allowed, but not often seen in discussions of nonlinear refraction. When one refers to nonlinear refraction, the common understanding is that the index depends linearly on irradiance, which is conveniently modeled by eqn [20]. But there are other physically relevant situations to consider. The coefficient n_1, for example, describes the linear electro-optic effect in which the change in index is linearly proportional to the electric field (although it is called the linear electro-optic effect, it is actually derived from the second-order nonlinear susceptibility). The electric field can be the oscillating field of a laser beam, for example, or a dc field applied to an electro-optic crystal (e.g., Pockel's cell). When other terms are added to eqns [20] or [21], the units of the coefficients must be chosen to keep the equation dimensionless.

In the preceding discussion, there is an implication that the nonlinear index is an instantaneous function of the irradiance or field. Although the material system can never respond instantaneously, this is a good approximation in many situations. Sometimes it is not. Consider a pulsed laser beam that heats the material. When the local temperature increases, the linear refractive index can change. A short, Q-switched laser can have a pulse duration far less than a microsecond, while the temperature change it induces can last many orders of magnitude longer. This means optical modification of the refractive index may persist long after the exciting pulse has vanished, i.e., when the irradiance is at zero.

Another example is when a laser beam promotes electrons from their ground state to higher energy states of the material system. These excited electrons may modify the refractive index of the material. In general, the excited electrons remain in high-energy states for some period of time before relaxing to the ground level – if this time is longer than the excitation, the change of the refractive index persists beyond the time the laser beam is present. In these situations, the model of nonlinear refraction suggested by the above equations is inaccurate. One cannot obtain the nonlinear refraction from a simple algebraic analysis. A system of time-dependent equations – describing the dynamics of excitation and relaxation – must be solved using numerical procedures. Such phenomena are sometimes loosely categorized as 'dynamic linear optical effects' or 'effective third-order nonlinearities'.

Nonlinear Absorption

Consider a single-frequency light beam passing through an optically absorbing region of length L.

For simplicity, neglect reflections caused by surfaces that may define the region of interest, i.e., ignore reflections at surfaces that may be located on the optical axis z at points $z = 0$, $z = L$, or any other point in the path. Linear absorption means that the optical power extracted from the light beam as it traverses the absorbing medium is a direct function of the power at a given point. This is described mathematically by an elementary, linear differential equation known as Beer's law:

$$\frac{\mathrm{d}}{\mathrm{d}z} I(z) = -\alpha I(z) \qquad [26]$$

The constant of proportionality is α, which is the linear absorption coefficient. Equation [23] is solved by direct integration:

$$\int_{I(0)}^{I(L)} \frac{\mathrm{d}I}{I} = -\alpha \int_0^L \mathrm{d}z \qquad [27]$$

This has the solution:

$$\frac{I(L)}{I(0)} = \exp(-\alpha L) \qquad [28]$$

which means the irradiance decreases exponentially as a function of propagation distance in a linearly absorbing medium.

In the nonlinear regime, we don't expect a classical Beer's law model to hold. By definition, the absorption will be a nonlinear function of irradiance at a given point. One makes the following power-series expansion to describe nonlinear absorption of monochromatic light:

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -\alpha I - \beta I^2 - \gamma I^3 - \cdots \qquad [29]$$

The coefficient β corresponds to a two-photon absorption process and γ is the coefficient of three-photon absorption. What about four-photon and even higher-order processes? These are rarely encountered, but certainly possible. To make these distinctions, eqn [29] is sometimes written with coefficients K_i or α_i instead of the sequential Greek alphabet:

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -K_1 I - K_2 I^2 - K_3 I^3 - K_4 I^4 - \cdots \qquad [30]$$

The units of the various absorption coefficients must maintain the dimensional consistency of eqns [29] and [30], which is irradiance/length or power/(length)^3. These are shown in Table 1.

Table 1 Dimensions of optical absorption coefficients

Process	*Coefficient*	*Units*
Linear absorption	α, K_1	(length)$^{-1}$
Two-photon absorption	β, K_2	length/power
Three-photon absorption	γ, K_3	length^3/power^2
Four-photon absorption	K_4	length^5/power^3

If there is linear absorption, the nonlinear processes at that wavelength are often (but not always) negligibly weak. This means we have the following inequalities:

$$K_1 I \gg K_i I^i, \qquad \text{where } i \geq 2 \qquad [31]$$

Nonlinear absorption is generally observed in the wavelength region where the medium is transparent to low-irradiance light, i.e., where linear absorption is negligible. It should be emphasized that there are situations where linear absorption is large and in fact a crucial component of the aggregate nonlinear effect. We will return to this point later in the discussion. For the moment, we neglect linear absorption. If the energy of the incident photons and energy levels of the system permit it, the lowest-order nonlinear process is two-photon absorption, described by the equation

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -K_2 I^2 \qquad [32]$$

which can be solved by elementary integration:

$$\frac{I(L)}{I(0)} = \frac{1}{1 + I(0) K_2 L} \qquad [33]$$

Unfortunately, the situation is rarely this convenient. The above integration has ignored the fact that nonlinear absorption may enable significant linear absorption. This is possible because nonlinear absorption promotes electrons from low- to high-energy states in the medium. The change in excited electron density associated with absorption of light (both linearly and nonlinearly) is called 'photocarrier generation'. These photocarriers (e.g., photo-electrons) may modify the linear absorption as well as the refractive properties of the material. If two-photon absorption causes the linear absorption to increase, for example, the assumption that allowed us to ignore K_1 when writing eqn [32] ceases to be valid. While eqn [32] was appropriate at the very start of the light–matter interaction, the generation of photocarriers after the passage of time may change that condition. The behavior of the system is therefore time-dependent, i.e., it is dynamic. Simple analytic solutions are almost always not possible when dealing with nonlinear absorption.

The time rate of change of the photocarrier population (N) is modeled by the following equation:

$$\frac{\partial}{\partial t}N = \frac{K_1 I}{\hbar\omega} + \frac{K_2 I^2}{2\hbar\omega} + \frac{K_3 I^3}{3\hbar\omega} + \frac{K_4 I^4}{4\hbar\omega} + \cdots - \text{recombination} - \text{diffusion} \qquad [34]$$

where ω is the angular frequency ($\omega = 2\pi f$) of the incident light. Also included are place-holders for loss processes such as recombination and diffusion, which may themselves have complicated mathematical descriptions. The coefficients K_i have exactly the same units, dimensions, and interpretation as in eqn [30] and Table 1; in the spirit of the preceding discussion, these coefficients may be time dependent. Note that the dimensions of eqn [30] describe absorption of light (irradiance/length), while eqn [34] models the photocarrier population density (length^{-3} time^{-1}). The units of these two equations are very different.

List of Units and Nomenclature

CARS	Coherent anti-Stokes Raman scattering
c.c.	Complex conjugate
CSRS	Coherent Stokes Raman scattering
DFG	Difference frequency generation
DFM	Difference frequency mixing
DFWM, FWM	(Degenerate) Four wave mixing
DRO	Doubly resonant OPO
EIT	Electromagnetically induced transparency
ESA	Excited state absorption
GVD	Group velocity dispersion
GVM	Group velocity mismatch
NLA	Nonlinear absorption
NLR	Nonlinear refraction
OFID	Optical free-induction decay
OPA	Optical parametric amplification/amplifier
OPG	Optical parametric generation/generator
OPL	Optical power limiter
OPO	Optical parametric oscillation/oscillator
2PA, TPA	Two-photon absorption
PCM	Phase conjugate mirror
QPM	Quasi phase matching
RIKES	Raman-induced Kerr effect spectroscopy
RSA	Reverse saturable absorption
SBS	Stimulated Brillouin scattering
SFG	Sum frequency generation
SHG	Second harmonic generation
SIT	Self-induced transparency
SRO	Singly resonant OPO
SRS	Stimulated Raman scattering
SRWS	Stimulated Rayleigh wing scattering
SVEA, SVAA	Slowly varying envelope (amplitude) approximation
THG	Third harmonic generation
TPF	Two-photon fluorescence

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Bloembergen N (1992) *Nonlinear Optics*. Redwood City, CA: Addison-Wesley.

Boyd RW (1991) *Nonlinear Optics*. Boston, MA: Academic Press.

Levenson MD and Kano S (1988) *Introduction to Nonlinear Laser Spectroscopy*. Boston, MA: Academic Press.

Miller A, Miller DAB and Smith SD (1981) Dynamic nonlinear optical processes in semiconductors. *Advances in Physics* 30: 697–800.

Mills DL (1998) *Nonlinear Optics: Basic Concepts*. New York: Springer.

Newell C and Moloney JV (1992) *Nonlinear Optics*. Redwood City, CA: Addison-Wesley.

Sauter EG (1996) *Nonlinear Optics*. New York: Wiley.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Yariv A (1991) *Optical Electronics*, 3rd edn. New York: Holt, Rinehart, and Winston.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: Wiley.

Nonlinear Optical Phase Conjugation

B Y Zeldovich, University of Central Florida,
Orlando, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Phase conjugation (PC) beams and their applications are best illustrated by Figure 1, which shows the passage of a coherent planewave incident upon a transparent optical element with inhomogeneities of refractive index (Figure 1a). Reversibility of light propagation implies the existence of such an 'anti-distorted' beam, which reverts back to a planewave after reversing through the same inhomogeneities (Figure 1b). A real-valued monochromatic optical wave, $E_{\text{real}}(\mathbf{R}, t)$, may be represented by the complex amplitude $E(\mathbf{R})$ via $E_{\text{real}}(\mathbf{R}, t) = 0.5[E(\mathbf{R})\exp(-i\omega t) + E^*(\mathbf{R})\exp(i\omega t)]$, where $E^*(\mathbf{R})$ represents the complex conjugate of $E(\mathbf{R})$:

$$\begin{aligned} E(\mathbf{R}) &= |E(\mathbf{R})| \exp[i\varphi(\mathbf{R})] \\ E^*(\mathbf{R}) &= |E(\mathbf{R})| \exp[-i\varphi(\mathbf{R})] \end{aligned} \qquad [1]$$

Mathematical expression of time-reversal, $E_{\text{real}}(\mathbf{R}, t) \rightarrow E_{\text{real}}(\mathbf{R}, -t)$, becomes the exchange $E(\mathbf{R}) \leftrightarrow E^*(\mathbf{R})$ for monochromatic waves. The reversibility of propagation means that the complex propagating field $E(\mathbf{R})$ (for example, planewave $\exp(i\mathbf{k}\cdot\mathbf{R})$), is a wave equation solution equivalent to $E^*(\mathbf{R})$ (in this example, $\exp(-i\mathbf{k}\cdot\mathbf{R})$). Conjugation of complex amplitude means reversal of the sign of the phase, and a mixed label PC has been coined and nowadays is firmly established. Terms such as 'wave front reversal' and 'generation of time-reversed replica of the beam' are also used to describe PC.

Figure 1 also illustrates one of the most important applications of PC. If the element in question is a laser-type amplifier, then double-passage allows the extraction of energy from the optically inhomogeneous laser medium in the form of a perfectly collimated beam of diffraction-limited divergence. A PC device may also serve as a mirror of a laser resonator, resulting in the same beam-correction purpose.

One of the most practical and robust methods of PC is based on stimulated Brillouin back-scattering (SBS). The incident beam illuminates the SBS-active transparent medium, for example, liquids (CS_2, CCl_4, acetone, etc.), compressed gases (CH_4, SF_6, etc.), or solids (fused or crystalline quartz, glass, etc.). This 'pump' beam, $E(\mathbf{R}) \equiv E(\mathbf{r}, z)$, must possess well-developed transverse inhomogeneities of intensity $|E(\mathbf{r}, z)|^2$. Here, $\mathbf{r} = \{x, y\}$ is the part of coordinate vector transverse with respect to the central direction z of the beam in question, and $\mathbf{R} = \{\mathbf{r}, z\}$. These inhomogeneities may constitute a narrow focal waist in the case of weakly distorted focused beams, speckle-inhomogeneities in cases of strong distortions, or a combination thereof; see Figure 2, where solid lines symbolize the 'rays' of the incident pump.

Spontaneous scattering of the pump results in a multitude of possible transverse profiles $S(\mathbf{r}, z = 0)$ of the signal, whose 'rays' are depicted via the dotted lines on Figure 2. The signal is amplified exponentially due to SBS processes. This exponent in a simplified form may be represented as $|S(\mathbf{r}, z)|^2 \propto \exp(\int g dz)$, with the gain $g(z)$ being the result of transverse overlapping of the intensity profiles of signal and pump:

$$g(z) \approx \text{const} \cdot \langle |E(\mathbf{r}, z)|^2 \cdot |S(\mathbf{r}, z)|^2 \rangle / \langle |S(\mathbf{r}, z)|^2 \rangle \qquad [2]$$

Thus, the similarity of the intensity profile of the signal to that of the pump is rewarded by the exponential preference in the output signal level. However, this is not enough to guarantee that the output backward-scattered signal $S(\mathbf{r}, z)$ is phase

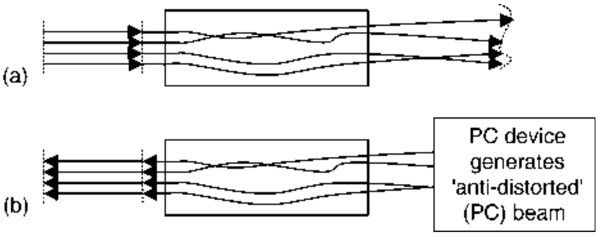

Figure 1 (a) Collimated beam is distorted by propagation through inhomogeneous medium. (b) Conjugate, a.k.a. antidistorted beam becomes collimated after backward passage through the same medium.

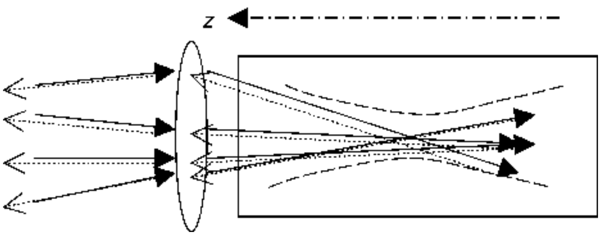

Figure 2 Stimulated Brillouin Scattering (SBS) method of PC. Solid lines symbolize the rays of incident 'beam', while dotted lines depict the rays of the signal amplified by the SBS process.

conjugate with respect to the pump. It is here that the mutual reversibility of propagation of strongly inhomogeneous signal and pump becomes crucial. Namely, good transverse overlapping is sustained along the entire interaction length z, only if $S(\mathbf{r}, z) \propto E^*(\mathbf{r}, z)$. This constitutes the 'discrimination mechanism' of SBS-PC: the 'conjugate mode' of the signal has a z-sustained advantage in exponential gain and, under appropriate conditions, suppresses all nonconjugate signal configurations in competition for pump power. Nonlinear-optical wave theory of this discrimination process has been developed further.

The holographic mechanism of PC has a different nature. It may take the form of static holography or dynamic holography; the latter essentially may be considered as a part of nonlinear optics. Here is a simplified description of one of the variants. Suppose one wants to obtain a conjugate replica of the incident monochromatic signal, whose complex profile is $S(\mathbf{R})$. A plane reference wave $A(\mathbf{R}) = \exp(i\mathbf{k}_A \cdot \mathbf{R})$ is also directed to the registering medium. If $S(\mathbf{R})$ and $A(\mathbf{R})$ are mutually coherent, then an interference pattern of intensity is produced $\{S^*(\mathbf{R})A(\mathbf{R}) + \text{compl. conj.}\}$. The medium records this pattern in the form of the modulation of refractive index and/or of the absorption coefficient, $\delta\varepsilon(\mathbf{R}) \propto (c_1 + ic_2)\{S^*(\mathbf{R})A\ (\mathbf{R}) + \text{c.c.}\}$ (Figure 3a). At the 'reconstruction' stage, the hologram is illuminated by another plane reference wave, $B(\mathbf{R}) = B_0 \exp(i\mathbf{k}_B \cdot \mathbf{R})$. In the specific case, when $\mathbf{k}_B = -\mathbf{k}_A$, the source δD_{conj} of the reconstructed wave becomes, as shown by Figure 3b:

$$\begin{aligned}\delta D_{\text{conj}} &= (c_1 + ic_2)[S^*(\mathbf{R})A(\mathbf{R})] \cdot B(\mathbf{R}) \propto S^*(\mathbf{R}) \\ &\quad \times (c_1 + ic_2)A_0B_0 \qquad [3]\end{aligned}$$

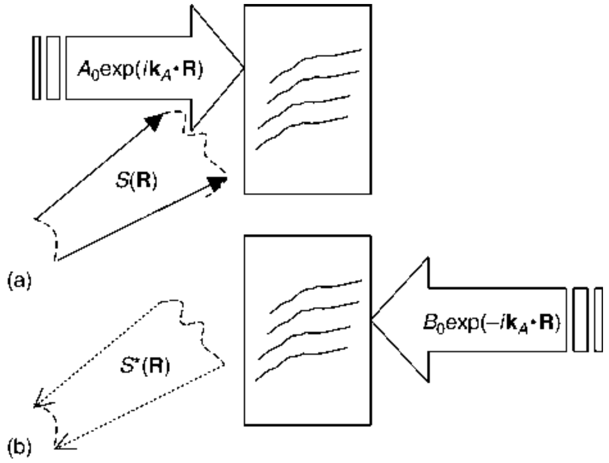

Figure 3 (a) Recording of a hologram in a photosensitive medium by a reference wave A and signal wave S. (b) Read-out of that hologram by the wave B counter-propagating to A results in generation of the conjugated signal S^*.

Reversibility of propagation laws guarantees that this source will indeed excite a conjugate wave with high efficiency. Both processes, recording of the interference pattern and reconstruction of the conjugate wave, may occur simultaneously, if one deals with the dynamic holography. The same process may also be described as nonlinear-optical four-wave mixing (FWM) via cubic nonlinearity $\chi^{(3)}$:

$$\delta D_{\text{conj}}(\mathbf{R}) = \chi^{(3)}S^*(\mathbf{R})A(\mathbf{R})B(\mathbf{R}) = \chi^{(3)}S^*(\mathbf{R})A_0B_0 \qquad [4]$$

with the four waves; A, B, S, S^*.

Important characteristics of a PC device are: fidelity of conjugation, efficiency of returning back-reflected power/energy, and reaction time or build-up time. The fidelity parameter must show how close the output of the device $E_{\text{out}}(\mathbf{r})$ is to the perfectly conjugate profile $S^*(\mathbf{R})$ of the incident signal (up to an arbitrary constant complex factor). Among other definitions, the square modulus of the normalized transverse overlapping integral of the fields is often discussed:

$$f = \frac{|\iint E_{\text{out}}(x, y)S(x, y)\mathrm{d}x\mathrm{d}y|^2}{\left(\iint |E_{\text{out}}(x, y)|^2\mathrm{d}x\mathrm{d}y \cdot \iint |S(x, y)|^2\mathrm{d}x\mathrm{d}y\right)} \qquad [5]$$

To achieve good (close to 1) fidelity of PC in the holographic or FWM scheme, one has to guarantee that the reference waves, $A_0 \exp(i\mathbf{k}_A \cdot \mathbf{R})$ and $B_0 \exp(i\mathbf{k}_B \cdot \mathbf{R})$, are exactly conjugate to each other, $\mathbf{k}_A = -\mathbf{k}_B$. That, in turn, requires the absence of any distortions in the holographic or nonlinear medium. To the contrary, there are very modest requirements on the phase inhomogeneities of the medium in the SBS scheme of PC, and for that reason it is usually labeled as a scheme of self-phase conjugation.

A hybrid scheme, Brillouin-assisted FWM, has been suggested and implemented, where mutually conjugate reference waves are produced via SBS-PC, while FWM, in a separate medium, exploits SBS nonlinearity. This scheme has produced extremely high, up to about 10^{10}, reflectivity, accompanied by an extremely low-noise detection of incoming signals.

A number of other PC schemes were first realized with the use of an important class of materials for nonlinear optics: photo-refractive crystals (PRC). These are crystals where ionization of the dopants by the incident light creates interference pattern of charge separation. The linear electro-optic (Pockels) effect transforms the resulting patterns of the electrostatic field into patterns of refractive index, thus creating dynamic holograms. A remarkable property of these crystals is that most of them act as very good electrical insulators. Therefore, the only source of conductivity, which tends to erase the hologram, is photoconductivity, the value of the latter being proportional to the intensity of incident light itself. For this reason, the steady-state strength of the hologram turns out to be independent of the intensity. It is the build-up time that must be increased, if intensity is low. One can achieve various processes in PRC, including those similar to the $\chi^{(3)}$-type optical Kerr effect, stimulated scattering, etc.

The simplest use of PRC for generation of PC waves is to devise a FWM scheme (see Figure 3). However, a multitude of other, nontrivial schemes have been devised and implemented. The most impressive use of PRC for PC is based on interactions of the beams in the vicinity of a corner of a rectangular crystal, typically $BaTiO_3$, these interactions being of the stimulated scattering type. To honor J. Feinberg's cat, whose image was reconstructed in the first demonstration, this scheme is universally called 'cat conjugator'. It turned out to be a very reliable and robust scheme of self-phase-conjugation (SPC).

Another important group of the SPC scheme is called 'tail-biting'. One of the variants of the latter is shown in Figure 4. Incident 'pump' wave $A(\mathbf{R})$, whose transverse profile one wants to conjugate, enters the medium, and then is redirected by mirrors back into the same medium, to 'bite itself'. For the purpose of discussion, this redirected pump is labeled as wave $B(\mathbf{R})$ on its second arrival into the medium. Spontaneous scattering results in a seed for the amplification of the wave $C(\mathbf{R})$ via the process of stimulated scattering (SS) $A(\mathbf{R}) \rightarrow C(\mathbf{R})$.

Wave $C(\mathbf{R})$ is also redirected into the medium by the same mirrors, and is labeled as wave $D(\mathbf{R})$ on its second arrival into the medium. Both $C(\mathbf{R})$ and $D(\mathbf{R})$ are depicted in Figure 4 by white arrows. Wave $D(\mathbf{R})$ is also amplified due to the SS process $B(\mathbf{R}) \rightarrow D(\mathbf{R})$.

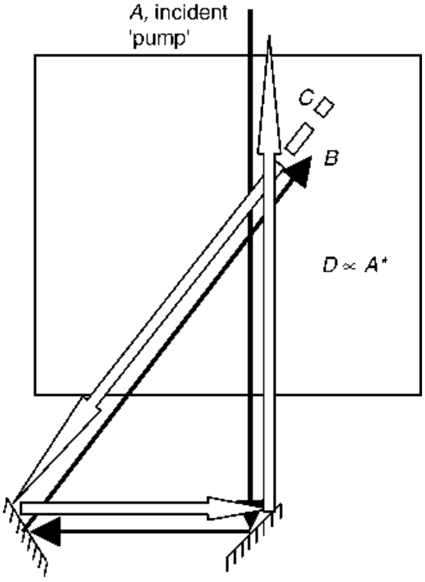

Figure 4 Tail-biting scheme of phase conjugation.

As it follows from the macroscopic description of SS process, volume gratings of refractive index are recorded in the medium, $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R}) + B^*(\mathbf{R})D(\mathbf{R})]$. Among all the transverse profiles for the seed $C(\mathbf{R})$, the one that happens to be proportional to $B^*(\mathbf{R})$ will be reflected by the mirrors into the medium in the form $D(\mathbf{R}) \propto A^*(\mathbf{R})$. The latter case is again a consequence of the reversibility of propagation laws, where both gratings have the same profile and add coherently in the form $A^*(\mathbf{R})B^*(\mathbf{R})$. What is even more important, the grating $B^*(\mathbf{R})D(\mathbf{R}) \propto B^*(\mathbf{R})A^*(\mathbf{R})$ serves to close the feedback loop for seeding the appropriate $C(\mathbf{R})$. This, and other variants of tail-biting schemes, have also proved to be reliable and robust. Quite often a laser amplifier is inserted in the path $A \rightarrow B$ and $C \rightarrow D$, thus enhancing the feedback.

Another important scheme is traditionally called 'double PC'; albeit it may be better described as 'mutual PC'. Figure 5 should help in understanding that scheme. Consider the beam $A(\mathbf{R})$ incident to a medium active with respect to the stimulated-scattering-type process $A(\mathbf{R}) \rightarrow C(\mathbf{R})$. This process originates from a random seed and results in a fan of different waves $C(\mathbf{R})$. Such a process is characteristic of PRCs and is a consequence of recording dynamic gratings of refractive index $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R})]$. Suppose another beam $B(\mathbf{R})$, typically incoherent with respect to $A(\mathbf{R})$, illuminates the medium from the other side, and also is engaged in the fanning process, this time $B(\mathbf{R}) \rightarrow D(\mathbf{R})$, with the grating of refractive index $\delta n(\mathbf{R}) \propto -i[B^*(\mathbf{R})D(\mathbf{R})]$ involved. Among all the transverse profiles for the seed grating $C(\mathbf{R})A^*(\mathbf{R})$, the one that happens to be proportional to $B^*(\mathbf{R})A^*(\mathbf{R})$ will be supported by similar profiles in the $B(\mathbf{R}) \rightarrow D(\mathbf{R})$ scattering. In this way, two fanning

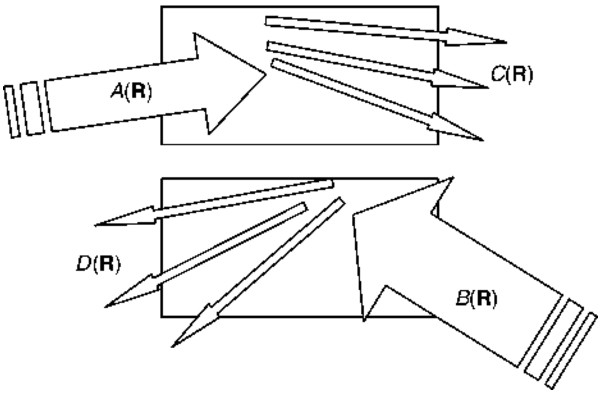

Figure 5 Towards the mechanism of double PC (mutual PC).

processes enhance each other and close the feedback loop. An important consequence is that one obtains the following output waves:

$$\begin{aligned} C(\mathbf{R}) &= \mathrm{const}_1 \times B^*(\mathbf{R}) \\ D(\mathbf{R}) &= \mathrm{const}_2 \times A^*(\mathbf{R}) \end{aligned} \qquad [6]$$

One can say that the output wave $C(\mathbf{R})$ presents a conjugate replica of $B(\mathbf{R})$-beam, while the output wave $D(\mathbf{R})$ presents a conjugate replica of $A(\mathbf{R})$-beam; hence the label 'mutual PC'. In some experiments the incident beams, $A(\mathbf{R})$ and $B(\mathbf{R})$, were of somewhat different colors (i.e., from different lasers), or their pulses did not overlap in time in the medium. It is the grating memory that 'informs' one beam about the presence of the other. The mechanism of cat conjugator is considered to be a combination of tail-biting and double PC schemes.

Pumped laser material with saturable gain often serves as an important nonlinear-optical medium for implementing PC. One of the evident advantages is that possible low efficiency of properly nonlinear process may be compensated by the gain in the same piece of material.

Considerable attention is paid nowadays to the chirp reversal of optical pulses that are used in fiber optical communications. Such chirped pulses, $S(t) = \exp(-i\omega_0 t + i\beta t^2/2 - t^2/\tau_1^2)$, with the value of the 'chirp' $\mathrm{d}\omega/\mathrm{d}t = -\beta$ and an increased pulse duration τ_1, arise as a result of propagation in a fiber which possesses group velocity dispersion (GVD). Pulse stretching and chirp are the consequences of different propagation times for the different frequency constituents of the pulse. If some device changes the sign of that chirp, then subsequent propagation of the pulse, through a piece of fiber with the same GVD, restores the duration of the pulse to original shorter value τ_0. A scheme has been suggested to use nonlinear optical FWM in a $\chi^{(3)}$-medium or three-wave mixing (ThWM) in a $\chi^{(2)}$-medium:

$$\begin{aligned} \delta D_{\mathrm{conj}}(t) &= \chi^{(3)} S^*(t) A(t) B(t) \\ \delta D_{\mathrm{conj}}(t) &= \chi^{(3)} S^*(t) C(t) \end{aligned} \qquad [7]$$

Indeed, complex conjugation of the time dependence of the signal's field is equivalent to the change of the chirp sign. The second (ThWM) expression is written assuming that the reference wave $C(t) \propto \exp(-2i\omega_0 t)$, so that is has the frequency double that of the signal. The same process of ThWM has been shown to yield the conjugation of transverse phase profile of the beams, but for a number of reasons it was not applied.

As for the applications of PC, one of them was discussed above (Figure 1: double-pass scheme of a laser amplifier). Lasers with one or two PC mirrors, or with more complicated PC scheme, promise generation of high-transverse-quality beams with the use of realistic laser media, the latter inevitably possessing thermal and other distorting inhomogeneities. Another important potential application is free-space optical communications: Earth–Earth or Earth–satellite through atmosphere, or satellite–satellite communications. An 'interrogating' beam may be sent in one direction through an imperfect optical system and/or through turbulent atmosphere. Conjugation of transverse profile of the 'interrogating' beam at the other end of the communication link leads to an almost perfect redirection of the reflected beam towards the 'beacon' source, while time modulation may carry the information.

Readout of information data from a holographic storage is almost always performed in the regime of reading the conjugate wave, since it corrects for the most part of aberrations of optical systems in question.

To conclude, one may use a citation from the two consecutive *Scientific American* popular papers on PC: 'at present the number of applications would seem to be limited only by imagination.'

Reviews of various aspects of phase conjugation may be found in monographs and *Scientific American* papers listed below. The author of the present article was introduced to PC by his colleague V. V. Ragulskii, and has greatly benefited from collaboration with V. V. Ragulskii and V. V. Shkunov.

See also

Microscopy: Phase Contrast Microscopy. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Phase Control:** Phase Conjugation and Image Correction. **Scattering:** Scattering Theory.

Further Reading

Bespalov VI and Pasmanik GA (1994) *Nonlinear Optics and Adaptive Laser Systems*. Commack, NY: Nova Science Publishers.

Feinberg J (1986) Self-pumped, continuous wave phase conjugator using internal reflection. *Optics Letters* 7: 486.

Fisher RA (ed.) (1983) *Optical Phase Conjugation*. New York: Academic Press.

Gower M and Proch D (eds) (1994) *Optical Phase Conjugation*. Berlin: Springer Verlag.

Hellwarth RW (1982) Optical beam conjugation by stimulated backscattering. *Optical Engineering* 21: 257.

Nosatch OYu, Popovichev VI, Ragulskii VV and Faizullov FS (1972) Compensation of phase distortions in an amplifying medium by a ‘Brillouin mirror’. *JETP Letters* 16: 435.
Pepper DM (1986) Applications of optical phase conjugation. *Scientific American* 254: 74–83.
Shkunov VV and Zeldovich BYa (1985) Optical phase conjugation. *Scientific American* 253: 54–59.
Yariv A (1978) Phase conjugate optics and real-time holography. *IEEE Journal of Quantum Electronics* 14: 650.
Zeldovich BYa, Popovichev VI, Ragulskii VV and Faizullov FS (1972) On relation between wavefronts of reflected and exciting radiation in stimulated Brillouin scattering. *JETP Letters* 16: 109.
Zeldovich BYa, Shkunov VV and Pilipetsky NF (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.
Zeldovich BYa, Mamaev AV and Shkunov VV (1995) *Speckle-Wave Interactions in Application to Holography and Nonlinear Optics*. Boca Raton, FL: CRC Press.

Photorefraction

M Cronin-Golomb, Tufts University, Medford, MA, USA

B Kippelen, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction and Basics of the Photorefractive Effect

In 1966 Ashkin and coworkers were pursuing research on optical devices using the electro-optic material lithium niobate. They noticed that the refractive index of lithium niobate would change when it was exposed to laser light, and upset the expected operation of their devices. They called the effect optical damage. Shortly thereafter, Chen, LaMacchia, and Fraser reported on the use of the optical damage effect for holographic data storage. Thus began the field of photorefractive nonlinear optics, which has been used in various applications such as real-time holography, optical data storage, optical image amplification, nondestructive testing, distortion compensation by phase conjugation, pattern recognition, and radar signal processing. Many inorganic and organic materials have been investigated for their photorefractive effects, including ferroelectrics, semiconductors, and sensitized polymers. The most well-known inorganic materials are lithium niobate, bismuth silicon oxide, barium titanate, and strontium barium niobate. Most organic photorefractive materials are based on polymeric photoconductors such as those used in xerography that are doped with electro-active molecules, some plasticizers, and sensitizers.

While in its broadest interpretation, the photorefractive effect occurs whenever light incident on a material causes a refractive index change, one usually applies the term to electro-optic index changes resulting from optically generated space charge fields. Materials that are photorefractive in this sense share the following properties

- high transmission at the operating wavelengths;
- linear electro-optic coefficients or orientational Kerr effects;
- charge carriers that become mobile when optically excited;
- trapping centers for these charge carriers to enable spatially non-uniform redistribution of charge.

Consider two beams from the same laser crossing inside a photorefractive material such as barium titanate. The interference pattern will have bright and dark fringes. Charge carriers are excited where the light is bright, then drift and diffuse to regions of relative darkness where they preferentially recombine into trapping centers. In this way, a net excess charge develops in the dark regions, and a net deficit of charge develops in the bright regions. The spatially varying charge distribution has an electric field associated with it and this electric field causes a spatially varying refractive index profile. Because the space charge, its field, and resulting refractive index have the same spatial periodicity as the original interference pattern we have a holographic phase grating. The diffraction efficiency of the hologram can easily approach 100% for materials such as barium titanate and strontium barium niobate which have high electro-optic coefficients. Likewise, such high diffraction efficiencies are easily obtained in 100 micrometer-thick photorefractive polymer films.

The Standard Rate Equation Model

The development of photorefractive gratings can be modeled using standard semiconductor rate equations. Figure 1 shows two beams incident

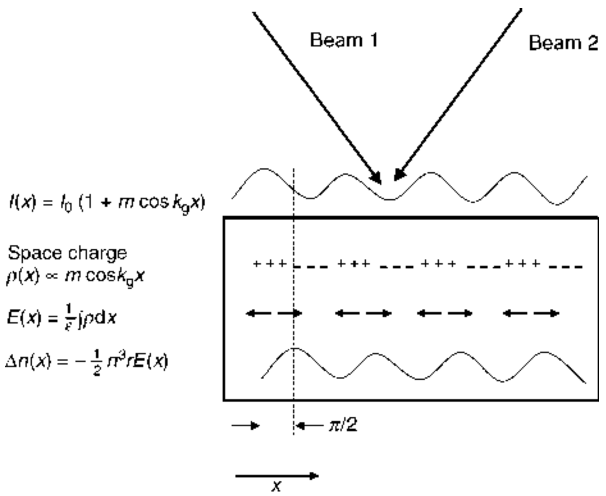

Figure 1 Diagram of two-beam coupling process showing optical interference pattern, space charge, resultant electric field, and $\pi/2$ phase shifted electro-optically induced refractive index grating.

symmetrically on a photorefractive crystal or polymer. They form an interference pattern whose intensity may be written:

$$I(x) = I_0(1 + m\cos(k_g x)) \qquad [1]$$

where x is the direction perpendicular to the interference fringes, I_0 is the average intensity, m is the modulation index, and k_g is the wavenumber of the interference pattern.

The crystal may be considered a wide bandgap semiconductor containing electron donors in the bandgap with density N_D and electron acceptors with density N_A. It is assumed that some electrons ionized from the donors permanently occupy these acceptors so that when charge in the crystal is distributed uniformly in the dark, the number density of ionized donors N_D^i is equal to N_A. Likewise, polymers contain donor and acceptor-like molecules that can be neutral or ionized. The spatially varying light distribution ionizes the donors at the following rate, assuming that $N_D^i \ll N_D$:

$$\frac{\partial N_D^i}{\partial t} = sIN_D - \gamma_R n_e N_D^i \qquad [2]$$

where s is a photoionization coefficient, γ_R is a recombination parameter, and n_e is the density of excited charge carriers, which we assume here to be electrons. The model can easily be generalized to include holes. We also use the equation of charge conservation:

$$\frac{\partial N_D^i}{\partial t} = \frac{\partial n_e}{\partial t} - \frac{1}{e}\nabla\cdot\mathbf{j} \qquad [3]$$

where e is the charge of an electron, and $\mathbf{j}$ is the current in the conduction band.

$$\mathbf{j} = \mu e n_e \mathbf{E} + k_B T \mu \nabla n_e \qquad [4]$$

where μ is the electron mobility, k_B is Boltzmann's constant, and T is the temperature. The electric field obeys Gauss's law:

$$\nabla\cdot\mathbf{E} = -e(n_e + N_A - N_D^i)/\varepsilon \qquad [5]$$

where N_A is the density of acceptors. These equations may be linearized by approximating the electron density, ionized donor density, and electric field with their first Fourier components:

$$\begin{aligned} E &= E_0 + \tfrac{1}{2}(E_1\exp(\mathrm{i}k_g x) + E_1^*\exp(-\mathrm{i}k_g x)) \\ N_D^i &= N_{D0}^i + \tfrac{1}{2}(N_{D1}^i\exp(\mathrm{i}k_g x) + N_{D1}^{i*}\exp(-\mathrm{i}k_g x)) \qquad [6] \\ n_e &= n_{e0} + \tfrac{1}{2}(n_{e1}\exp(\mathrm{i}k_g x) + n_{e1}^*\exp(-\mathrm{i}k_g x)) \end{aligned}$$

This assumption is strictly valid only when the modulation index m is much less than unity. Otherwise, a generalization to higher orders in the Fourier series is required. However, the linearized theory is sufficient to illustrate the most important features of the photorefractive effect. The solution for the space charge field E_1 for the case when the interference pattern is applied at time $t=0$ is

$$E_1 = m\frac{-\mathrm{i}E_Q(E_0 + \mathrm{i}E_D)}{E_0 + \mathrm{i}(E_D + E_Q)}(1 - \exp(-t/\tau)) \qquad [7]$$

where E_0 is an externally applied or photovoltaic field (if any), E_D is the diffusion field

$$E_D = \frac{k_B T k_g}{e} \qquad [8]$$

and E_Q is the limiting space charge field

$$E_Q = \frac{eN_A}{\varepsilon k_g} \qquad [9]$$

Some photorefractive crystals, most notably $LiNbO_3$, exhibit the photovoltaic effect in which optical illumination induces a dc field across the crystal.

The sinusoidally varying space charge field E_1 operates through the linear electro-optic effect with effective coefficient r to produce a sinusoidal variation in the refractive index n of the crystal:

$$n(x) = n_0 + \tfrac{1}{2}(n_1\exp(\mathrm{i}k_g x) + n_1^*\exp(-\mathrm{i}k_g x)) \qquad [10]$$

where

$$n_1 = -\frac{1}{2}rn_0^3E_1 \qquad [11]$$

The effective electro-optic coefficient r may be found from tensor analysis of the electro-optic tensor and the vector space charge and optical fields. Notice that there is a 90-degree phase shift between the interference pattern and the refractive index grating when E_0 is zero. The time constant τ is given by

$$\tau = \frac{N_A}{sN_D I_0} \frac{E_0 + i(E_D + E_\mu)}{E_0 + i(E_D + E_Q)} \qquad [12]$$

where E_μ is a mobility field

$$E_\mu = \frac{\gamma_R N_A}{\mu k_g} \qquad [13]$$

When E_0 is nonzero, there is an oscillatory component to the time constant.

In contrast to the case of ordinary optical nonlinearities such as the optical Kerr effect, in which the nonlinear coupling strength is proportional to the optical intensity, the steady state strength of the photorefractive effect is independent of optical intensity while the time constant is inversely proportional to intensity in the basic model described above. The time constant varies with the photoconducting performance of a given material. In the fastest polymeric and inorganic materials it is, at the time of writing, of the order of a few milliseconds at 1 W cm^{-2} of incident optical intensity.

At low intensities (below the equivalent dark intensity, 1 W cm^{-2} in barium titanate), the above equations need to be modified to include the effect of dark conductivity. Even in the dark, there will be a few mobile charge carriers in the conduction band that tend to erase the grating. This will result in the effect appearing more Kerr-like, except still with the 90-degree phase shift between the index grating and the interference pattern.

Coupled Wave Equations

The change in refractive index n_1 can be large enough to produce substantial interactions between the writing beams. The writing beams generate a phase grating that diffracts the beams into each other. The grating influences the writing beams, which in turn influence the grating. In the cases where the diffusion field dominates, for example when the externally applied or photovoltaic field E_0 is absent, one beam is amplified by in-phase diffraction of the other beam from the grating. As shown in Figure 2, this amplification results from a 90-degree phase shift due to diffraction from a phase grating coupled with a −90-degree phase shift from the spatial phase shift between the interference pattern and the index grating. The second beam is attenuated by destructive interference with the first beam diffracted by the grating. These interactions can be modeled by standard coupled wave theory. Let the electric field amplitude associated with the jth beam be

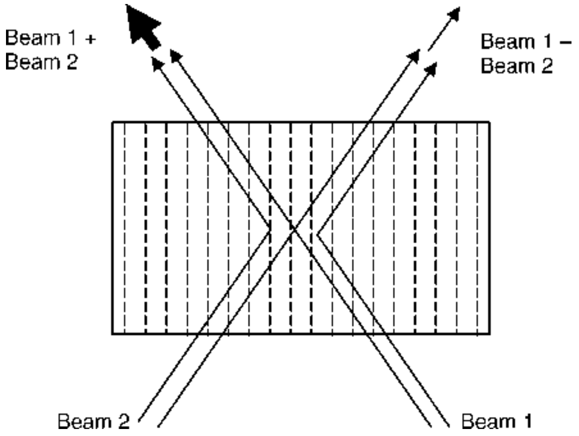

Figure 2 Two-beam coupling amplification. Beam 1 is amplified by constructive interference. Beam 2 is de-amplified by destructive interference.

$$\mathbf{E}_j(\mathbf{r}, t) = \mathbf{e}\Big[A_j(\mathbf{r}) \exp(i(\mathbf{k}_j \cdot \mathbf{r} - \omega t)) + A_j^* \exp(-i(\mathbf{k}_j \cdot \mathbf{r} - \omega t))\Big] \qquad [14]$$

where $\mathbf{e}$ is the polarization unit vector. Using the scalar wave equation

$$\nabla^2 \mathbf{E} + k^2 \mathbf{E} = 0 \qquad [15]$$

and the slowly varying envelope approximation

$$\left| \frac{d^2 A_j}{dz^2} \right| \ll k \left| \frac{dA_j}{dz} \right| \qquad [16]$$

we find

$$\begin{aligned} \frac{dA_1}{dz} &= +\gamma \frac{A_1 |A_2|^2}{I_0} \\ \frac{dA_2^*}{dz} &= -\gamma \frac{A_2^* |A_1|^2}{I_0} \end{aligned} \qquad [17]$$

with the coupling constant γ given by

$$\gamma = -\frac{i\omega \Delta n_m}{c \cos \vartheta} \qquad [18]$$

where ϑ is the half-angle between the writing beams and Δn_m is given by

$$\Delta n_m = -\frac{1}{2} n_0^3 r \frac{E_1}{m} \qquad [19]$$

Equation [17] shows that beam 1 is amplified and beam 2 is attenuated. That beam 1 is amplified instead of beam 2 is a result of the choice of crystal orientation and hence the sign of γ. These nonlinear equations can be solved for normalized intensity $I_j = |A_j|^2$:

$$\begin{aligned} I_1(z) &= \frac{I_0}{1 + (I_2(0)/I_1(0))\exp(+\Gamma z)} \\ I_2(z) &= \frac{I_0}{1 + (I_1(0)/I_2(0))\exp(-\Gamma z)} \end{aligned} \qquad [20]$$

where $\Gamma = \gamma + \gamma^*$ is the intensity coupling constant. In the limit where $I_1 \ll I_2$, the weak beam, beam 1, experiences exponential amplification. This amplification can be used to build optical oscillators with many of the same properties as conventional laser oscillators. The solution can be generalized to the case where γ has an imaginary component due to an applied or photovoltaic field E_0. Linear absorption can also be included.

Materials

Photorefractive materials may be separated into two broad classes: inorganic and organic. The first materials investigated were inorganic oxides and semiconductors. Their success led to efforts to endow more easily produced organic materials with the required photoconductivity, charge trapping centers, and electro-optic coefficients.

Inorganic Materials

The first requirement in a classic photorefractive material is a linear electro-optic coefficient, such as appears in sillenites such as bismuth silicon oxide $Bi_{12}SiO_{20}$ and bismuth germanium oxide $Bi_{12}GeO_{20}$. These were some of the first photorefractive materials used for image processing and phase conjugation applications. However, their electro-optic coefficients (a few pm/V) are not large enough to easily give rise to large diffraction efficiencies, or to photorefractive oscillators. Materials at a temperature near a phase transition generally have higher electro-optic coefficients because their crystal structures are on the verge of changing. They are extremely susceptible to the effect of any external influence such as the application of electric fields including photorefractive space charge fields. That is why ferroelectric materials are good candidates for photorefractive materials. These include barium titanate $BaTiO_3$, potassium niobate $KNbO_3$, and strontium barium niobate $Sr_xBa_{1-x}Nb_2O_6$. The mean free path of charge transport is less than that in sillenites, so they require more photons to reach a steady state. Therefore, ferroelectric materials are typically less sensitive than sillenites.

The second requirement is for photoconductivity. This requires the existence of photoexcitable charge carriers, either from valence band to conduction band or from intraband trapping centers. The latter source of photocarriers is used most often because the absorption length of light whose energy is greater than the bandgap is usually only a few micrometers. This places a severe restriction on the beams' interaction length ℓ. Thus there has been considerable research on suitable dopants, most commonly $Fe^{2+}$and Fe^{3+}. These dopants also act as trapping centers.

Lithium niobate is an example of a material with a large photovoltaic effect. When illuminated, this crystal develops a large dc field, which acts to enhance the grating strength. In those cases where the photorefractive effect is not wanted, such as in the design of lithium niobate waveguide devices, the photovoltaic effect can be greatly diminished by the addition of MgO to the crystal melt during growth.

For photorefractivity in the near infrared, semiconductors such as gallium arsenide and indium phosphide have been used with success. These materials also have the advantage that they can be grown in layered structures to produce, for example, multiple quantum wells that can be used to tailor the characteristics of optical excitation and charge transport.

Organic Materials

First-generation photorefractive polymers were designed to mimic the properties of their inorganic counterparts. Owing to their rich structural flexibility, organic synthetic materials with suitable charge transport, trapping, linear electro-optic effects and low optical absorption, were developed. This approach did build on the know-how developed previously in making photoconducting polymers for xerography and electro-optic polymers for optical modulation. Current polymers are based on an orientational photorefractive effect that leads to higher refractive index changes compared with traditional photorefractive materials. In materials with orientational photorefractivity, the refractive index change is produced by the field induced reorientation of anisotropic conjugated molecules that possess a permanent dipole moment and that have a high polarizability anisotropy. The photorefractive space charge field together with the applied field will periodically reorient these molecules and produce a periodic refractive index modulation through an orientational Kerr effect. The build-up and dynamics of this space-charge field are similar to those of traditional photorefractive crystals and can be described by eqns [2]–[7]. The time constant of the hologram is a convolution of the build-up time of the space-charge and the orientational diffusion time of the dipolar molecules in the total electric field. In contrast to crystals, polymers are nearly amorphous and the transport properties are described by charge hopping processes instead of band-type transport in crystals. Consequently, the photoconducting properties of polymers are strongly field dependent and photorefractivity is mainly observed under strong applied voltages. Numerous polymer composites have been developed using the hole transport polymer poly(N-vinylcarbazole). Several materials with a refractive index modulation amplitude of the order of a percent and two-beam coupling constants $\gamma > 100\ \mathrm{cm}^{-1}$ have been reported. New polymer composites are continuously being tested. Photorefractivity is also studied in other organic materials including organic crystals, liquid crystals, nanocomposites such as polymer-dispersed liquid crystals, or hybrid materials such as sol-gels.

Applications

Holographic Data Storage

Holographic data storage takes advantage of the Bragg selectivity of thick gratings. This allows many holograms to be superimposed in the same small volume, typically of the order of one cubic centimeter. A page of data is displayed on a spatial light modulator and a laser beam passing through the modulator is holographically recorded in the crystal with a reference beam at a specific angle. Many pages of data can be superimposed by recording many holograms with angularly multiplexed reference beams. Other multiplexing schemes are implemented by changing the shape of the wavefront of the reference beam. In principle, the upper limit of recording density is determined by the wavelength of light: one bit per cubic wavelength. If the recording wavelength is 0.5 μm, then one cubic centimeter can contain 1000 gigabytes of data. In practical circumstances, when noise is taken into consideration, the capacity is more realistically of the order of one gigabyte if a 1000 × 1000 spatial light modulator is used.

Distortion Compensation by Phase Conjugation

Photorefractive materials can be used to make high-reflectivity phase conjugate mirrors. The phase conjugate of a laser beam is produced when a hologram of the beam is read by another beam traveling in the opposite direction to the original reference beam. The phase conjugate reconstruction is a time-reversed copy of the original beam. If the original beam has passed through distorting optics, then the phase conjugate beam will retrace the path of the original beam through the distortion and emerge in its undistorted original state. In the standard nomenclature of phase conjugation, the input beam is called the signal, or probe, and the two reference beams are called the pumps. The output beam at $z = 0$ is called the phase conjugate beam and has zero amplitude at its input at $z = \ell$, where ℓ is the interaction length. Applications for phase conjugation exist, for example, in signal transmission through distortions and laser cavity design. If a phase conjugate mirror is used as a cavity mirror, then the effects of intracavity distortions are removed.

Since the photorefractive gratings can be very strong, the diffraction efficiency of the counterpropagating reference beam can be so high that the phase conjugate reflectivity can exceed unity. The simplest generalization of eqn [17] to the four-wave mixing phase conjugation case is when only transmission gratings are important, as occurs in many circumstances depending on the mutual coherence properties of the beams, and the material's properties. The coupled wave equations are

$$\begin{aligned}\frac{\mathrm{d}A_1}{\mathrm{d}z} &= \gamma\frac{(A_1A_4^* + A_2^*A_3)A_4}{I_0}\\ \frac{\mathrm{d}A_2^*}{\mathrm{d}z} &= \gamma\frac{(A_1A_4^* + A_2^*A_3)A_3^*}{I_0}\\ \frac{\mathrm{d}A_3}{\mathrm{d}z} &= -\gamma\frac{(A_1A_4^* + A_2^*A_3)A_2}{I_0}\\ \frac{\mathrm{d}A_4^*}{\mathrm{d}z} &= -\gamma\frac{(A_1A_4^* + A_2^*A_3)A_1^*}{I_0}\end{aligned} \qquad [21]$$

In the undepleted pumps approximation, $(I_1, I_2 \gg I_3, I_4)$, the equations become linear and the solution for phase conjugate reflectivity $R = I_3(0)/I_4(0)$ is

$$R = \left|\frac{\sinh[\gamma\ell/2]}{\cosh[(\gamma\ell + \ln r)/2]}\right|^2 \qquad [22]$$

where $r = I_2/I_1$ is the ratio between the intensities of the pumping beams.

The fact that the reflectivity of the phase conjugate mirror can be greater than unity means that we can build an optical oscillator bounded by a regular mirror and a phase conjugate mirror only. Not only does it not require any additional optical gain, but it also compensates for intracavity distortions. The regular mirror can have any shape, provided that it is sufficiently reflective.

Self-pumped Phase Conjugate Mirrors

If a laser beam passes through a crystal placed inside an optical cavity formed by two facing mirrors, light scattered by imperfections in the crystal can be amplified through the photorefractive effect. The cavity provides feedback and optical oscillation can build up. The oscillation beams pump the crystal as a phase conjugate mirror for the incident beam, thus forming a self-pumped phase conjugate mirror. The feedback can even be provided by total internal reflection in the crystal in which case the crystal by itself can become a phase conjugate mirror.

Pattern Recognition and Image Filtering

Photorefractive wave mixing can be used to perform pattern recognition by matched filtering. One example would be to identify a tank in a battlefield scene, another would be to identify all of the occurrences of a particular word on a page of text. Suppose the input beams contain the corresponding pictorial information, such as might be obtained by passing the beams through an image-bearing transparency or spatial light modulator. Equation [21] shows that the source for beam 3 contains a term proportional to the product of the amplitudes of the three input beams. If lenses are placed in the input beams so that the crystal receives the Fourier transforms of those beams, then the output beam at the crystal, beam 3, will be proportional to the product of the Fourier transforms of the input beams 1, 2, and 4. A lens may then be used to perform the inverse Fourier transform of the product of the Fourier transforms of the input beams, producing an output proportional to beam 1 convolved with beam 2 correlated with beam 4. If beam 1 is a point source before its Fourier transforming lens, then it will be a plane wave at the crystal. Beam 3 after inverse Fourier transformation by its lens will be the correlation of beams 4 and 2. For example, suppose we want to find all the occurrences of a particular word, say ‘optics’ in a given page of text. Then we would make a transparency of the word ‘optics’ and place that in input beam 4. We would then place an image of the page of text in beam 2. Beam 3 would then contain a field with bright spots at the places containing the word ‘optics’ in the original text.

The real-time holographic recording properties of photorefractive materials can also be exploited in medical imaging applications by performing time-gated holography. In this method, a hologram is formed by the temporal overlap in the photorefractive sample of a reference pulse and the first-arriving (ballistic photons) light from a stretched image-bearing pulse that has propagated through a scattering medium. The filtering of the useful photons from the scattered ones is achieved by reconstructing the hologram formed with the ballistic photons in a four-wave mixing geometry.

Optical Limiting, the Novelty Filter, and Laser Ultrasonic Inspection

The attenuation of beam 2 in eqn [17] can be used in several applications including optical limiting and novelty filtering. If one wants to protect a sensor from high-intensity laser radiation, then one could split a small portion of the input beam directed at the sensor and use it as beam 1 in a photorefractive recording setup with the input beam acting as beam 2. If the laser intensity is above the equivalent dark intensity such that the optically excited charge density is greater than the thermally excited charge density, the photorefractive effect will be activated and the input beam will be de-amplified by destructive interference with beam 1, thus protecting the sensor. In materials with high gain–length products ($\gamma\ell > 1$), separate provision of beam 1 is unnecessary because light scattered from the input beam by imperfections in the material and other scattering centers will be greatly amplified, often to such an extent that the input beam is almost completely attenuated. This photorefractive amplification of scattered radiation is known as the fanning effect, because the amplified scattered light often appears as a fan, or brush of light, as can be seen in Figure 3. The effect can also be used to make a novelty filter, which transmits only the moving portion of a scene. The crystal is only fast enough to

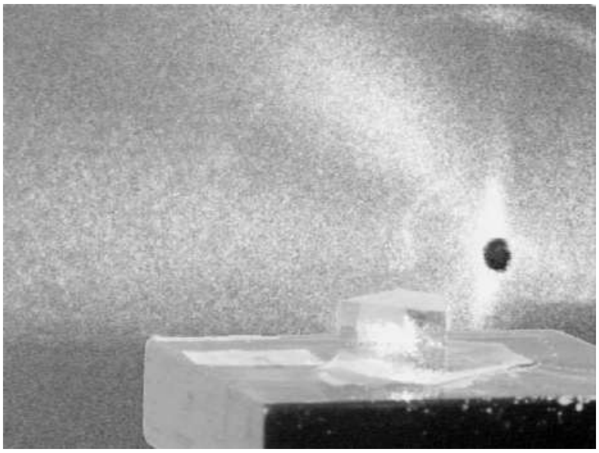

Figure 3 Photorefractive barium titanate exhibiting amplified scattering. A helium neon laser beam is incident from the lower left, passes through the crystal to a screen where it is blacked out to prevent saturation of the camera. The screen shows brushes of amplified scattering, or fanning.

respond to the slowly changing or static components of an image-bearing beam. Since the grating formed is Bragg-matched only to the slow component, the grating will only attenuate that slow component. Any rapidly changing parts of a scene pass through the crystal unattenuated. Such a filter is useful for picking out moving objects in a static cluttered background, for example a tank on a battlefield, or a micro-organism swimming against a stationary background.

A related application is in laser ultrasonic inspection. Defects in industrial material processing such as welding have characteristic ultrasonic signatures. The part under test is pinged by a pulsed laser and a probe laser is reflected from the part. As the part is shaken by the ultrasound, speckle in the reflected beam vibrates. A photorefractive recording is made of the speckle beam. Electrodes are placed on the photorefractive material, so that any optically induced currents can be detected. If the speckle pattern is not moving, or is moving more slowly than the speed of response of the photorefractive material, the photorefractive grating will be essentially at steady state: the drift currents balance the diffusion currents so there is no net current. There is no signal as the part moves through the process line. However, if the speckle beam is moving faster than the response time of the material, it will move photoexcited charge back and forth past the quasistatic grating and generate a net current for detection via the electrodes.

Adaptive Signal Processing

The relatively slow speed of photorefractive devices can be used to advantage in radio-frequency (RF) signal processing, such as signal extraction and coherent combination of signals from antenna arrays. The signal extraction application depends on grating competition in two-beam coupling. Suppose we wish to separate signals on two different RF carrier frequencies ω_1 and ω_2, respectively. The combined signal is applied to an optical carrier beam using a high-speed modulator. The resulting optical field may be represented as

$$S_1(t)\exp(i\omega_1 t) + S_2(t)\exp(i\omega_2 t) \qquad [23]$$

It is used to pump a unidirectional ring resonator so that grating competition allows oscillation only on the strongest component of the combined signal, say S_1. The output of the oscillator is proportional to the extracted component S_1. The effectiveness of the intersignal competition is enhanced by placing another photorefractive material in the cavity. A portion of the intracavity beam is picked off by a beamsplitter and used as a two-beam coupling pump in the second material. The crystal is oriented so that the photorefractive grating diffracts the picked-off beam back into the cavity. The return of the picked-off component is most effective for the stronger component S_1 thus decreasing its loss compared to that of the weaker component S_2. This beamsplitter/crystal combination is known as a reflexive coupler.

Photorefractive Solitons

Under favorable conditions, a single beam incident on a photorefractive crystal will induce a positive refractive index change at the center of the beam. This provides a self-focussing tendency that counteracts the beam's divergence due to diffraction. When these two effects balance each other, the beam can propagate with a constant diameter. Such a beam is known as a spatial soliton in analogy with temporal solitons in optical fibers and can occur when there is a component of the photorefractive response due to drift. The drift component of the photorefractive effect appears when a dc field E_0 is applied to the material. Potential applications are optically written waveguides and couplers.

List of Units and Nomenclature

Amplitude gain	[cm^{-1}]	γ
Applied dc field	[Vm^{-1}]	E_0
Current	[Am^{-2}]	$\mathbf{j}$
Dielectric permittivity	[Farads m^{-1}]	ε
Diffusion field	[Vm^{-1}]	E_D
Electron acceptor density	[m^{-3}]	N_A
Electron charge	[C]	e
Electron donor density	[m^{-3}]	N_D
Grating wavenumber	[m^{-1}]	k_g
Intensity gain	[cm^{-1}]	Γ
Interaction length	[cm]	ℓ
Ionized donor density	[m^{-3}]	N_D^i
Limiting space charge field	[Vm^{-1}]	E_Q
Mobility	[m^2 V^{-1} sec^{-1}]	μ
Mobility field	[Vm^{-1}]	E_μ
Normalized intensity	[V^2 m^{-2}]	I
Photo-ionization coefficient	[kg sec^{-2}]	s
Photorefractive time constant	[sec]	τ
Pump ratio	[1]	r
Radian frequency	[sec^{-1}]	ω
Recombination coefficient	[m^3 sec^{-1}]	γ_R
Slowly varying optical electric field	[Vm^{-1}]	A

See also

Holography, Applications: Holographic Recording Materials and Their Processing. **Holography, Techniques:** Holographic Interferometry. **Imaging:** Volume Holographic Imaging. **Information Processing:** Free-Space Optical Computing; Optical Neural Networks. **Materials for Nonlinear Optics:** Organic Nonlinear Materials. **Modulators:** Electro-Optics. **Nonlinear Optics, Basics:** Four-Wave Mixing; Nonlinear Optical Phase Conjugation. **Phase Control:** Phase Conjugation and Image Correction. **Scattering:** Stimulated Scattering. **Solitons:** Bright Spatial Solitons.

Further Reading

Coufal HJ, Psaltis D and Sincerbox GT (eds) (2000) *Holographic Data Storage*. Berlin: Springer.

Gunter P and Huignard J-P (eds) (1988) *Photorefractive Materials and their Applications I*. Berlin: Springer-Verlag.

Gunter P and Huignard J-P (eds) (1989) *Photorefractive Materials and their Applications II*. Berlin: Springer-Verlag.

Nalwa HS and Miyata S (eds) (1997) *Nonlinear Optics of Organic Molecules and Polymers*. Boca Raton: CRC Press.

Nolte DD (ed.) (1995) *Photorefractive Effects and Materials*. Kluwer.

Pepper DM, Feinberg J and Kukhtarev NV (1990) The photorefractive effect. *Scientific American* 263: 62.

Solymar L, Webb DJ and Grunnet-Jepsen A (1996) The Physics and Application of Photorefractive Materials. Oxford: Clarendon Press.

Yeh P (1993) Introduction To Photorefractive Nonlinear Optics. New York: Wiley.

Yu F and Yin S (eds) (2000) *Photorefractive Optics: Materials, Properties, and Applications*. San Diego: Academic Press.

Ultrafast and Intense-Field Nonlinear Optics

A L Gaeta, Cornell University, Ithaca, NY, USA

R W Boyd, University of Rochester, Rochester, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The tremendous development of high-powered femtosecond laser systems since the 1980s, has opened up new areas of research in nonlinear optics, plasma physics, atomic and molecular dynamics, and intense-field physics. For many of these applications, it is important to understand how ultrashort light pulses propagate through a medium under conditions in one or more of the processes in which nonlinear optics play an important role.

The starting point for the modeling of light propagation under these conditions is Maxwell's wave equation for the electric field $E(\mathbf{r},t)$ which is given in Gaussian units as

$$\nabla^2 E - \frac{\partial^2 E}{\partial t^2} = 4\pi \frac{\partial^2 P}{\partial t^2} \qquad [1]$$

where the $P(\mathbf{r},t)$ is the polarization inside the medium. Typically, the polarization is separated into a term P_l that depends linearly on the field E and into a term P_{nl} that depends nonlinearly on the applied field. The Fourier transform of the linear polarization can be expressed as $\tilde{P}_l(\mathbf{r},\omega) = \chi^{(1)}(\omega)\tilde{E}(\mathbf{r},\omega)$, where $\tilde{E}(\mathbf{r},\omega)$ is the Fourier transform of electric field $E(\mathbf{r},t)$. For the laser–matter interactions that we consider here, we assume that the linear susceptibility $\chi^{(1)}(\omega)$ is real, in which case the wave equation can be expressed as

$$\nabla^2 \tilde{E} + \left[\frac{n_l(\omega)\omega}{c}\right]^2 \tilde{E} = -\frac{4\pi\omega^2}{c^2}\tilde{P}_{\mathrm{nl}} \qquad [2]$$

where $n_l(\omega) = \sqrt{1 + 4\pi\chi^{(1)}(\omega)}$ is the frequency-dependent linear refractive index of the medium.

For light pulses that are longer than an optical period, the envelope description is valid and the electric field can be described by an amplitude envelope $A(\mathbf{r},t)$ and a carrier frequency ω_0 such that

$$E(\mathbf{r},t) = A(\mathbf{r},t)\mathrm{e}^{\mathrm{i}(k_0 z - \omega_0 t)} + c.c \qquad [3]$$

where $k_0 = k(\omega_0) = n_0\omega_0/c$ is the wavevector amplitude and $n_0 = n_l(\omega_0)$. This envelope description is advantageous for performing analytical and numerical studies of pulse propagation. However, for sufficiently short laser pulses, where the shape of the envelope function does not depend on the carrier phase of the carrier wave, such a description is no longer applicable. Nevertheless the envelope description can be made valid for pulses that are nearly as short as a single cycle or, alternatively, that have spectral bandwidths that are comparable to the central frequency ω_0. To derive an equation for the spatio-temporal evolution of the pulse envelope, the relation for the electric field is substituted into the wave equation (eqn [2]) and the following two approximations are made: i) $k_0 \partial A/\partial z \ll \partial^2 A/\partial z^2$, which signifies that the envelope varies slowly in space over a distance compared to the central wavelength; ii) $v_{\mathrm{ph}} \sim v_{\mathrm{gr}}$ where $v_{\mathrm{ph}} = c/n_0$ is the phase velocity and $v_{\mathrm{gr}} = (\mathrm{d}k/\mathrm{d}\omega)^{-1}$ is the group velocity. This latter approximation is invariably well satisfied when the frequencies contained in the laser field are highly nonresonant with any transition frequencies of the medium.

For an input pulse at $z = 0$ with a peak amplitude A_0, and characteristic widths in space and time given by w_0 and τ_p, respectively, the equation for the normalized amplitude $u(r_\perp, z, t) = A(r_\perp, z, t)/A_0$ can be expressed as

$$\frac{\partial u}{\partial \xi} = \frac{i}{4}\left(1 + \frac{i}{\omega_0 \tau_p}\frac{\partial}{\partial \tau}\right)^{-1} \nabla_\perp^2 u - i \sum_{n=2} \frac{L_{\mathrm{ds}}}{n! L_{\mathrm{ds}}^{(n)}} \frac{\partial^n u}{\partial \tau^n} + i\left(1 + \frac{i}{\omega_0 \tau_p}\frac{\partial}{\partial \tau}\right) p^{\mathrm{nl}} \qquad [4]$$

where $\tau = (t - z/v_g)/\tau_p$ is the normalized retarded time for the pulse traveling at the group velocity v_g, $L_{\mathrm{ds}}^{(n)} = \tau_p^n/\beta_n$ is the nth-order dispersion length, $\beta_n (n \geq 2)$ is the nth-order dispersion constant, $L_{\mathrm{ds}} = |L_{\mathrm{ds}}^{(2)}|$ is the dispersion length, $\zeta = z/L_{\mathrm{df}}$ is the normalized distance, $L_{\mathrm{df}} = k w_0^2/2$ is the diffraction length, p_{nl} is the normalized nonlinear polarization, and $\nabla_\perp^2$ is the transverse Laplacian. The presence of the operator $(1 + i\partial/\omega_0 \partial \tau)$ in the diffraction term $(\nabla_\perp^2 u)$ of eqn [4] leads to space–time focusing, while its presence in the nonlinear term results in self-steepening and both these terms arise from not making the slowly varying envelope approximation in time (i.e., $k_0 \partial A/\partial t \ll \partial^2 A/\partial t^2$) in deriving eqn [4]. For a self-consistent analysis of pulse propagation in the nonlinear regime, both the effects of space–time focusing and self-steepening must be included.

Nonlinear Refractive Index

For an isotropic medium in the highly nonresonant limit, the third-order term is the lowest-order contribution to the nonlinear susceptibility. This term gives rise to the nonlinear refractive index, that is, the index of refraction depends on the local intensity of the laser field. For many materials there are two contributions to the nonlinear refractive index: i) an instantaneous part that arises from the electronic response of the medium; and ii) a noninstantaneous contribution due to the nuclear motion of the molecules (i.e, the Raman contribution). For such a medium, the nonlinear polarization may be expressed as

$$p_{\mathrm{nl}}(\zeta, \tau) = \frac{L_{\mathrm{ds}}}{L_{\mathrm{nl}}}\left[(1 - f)|u(\zeta, \tau)|^2 + f\gamma_R \int_{-\infty}^{\tau} \mathrm{d}\tau' R(\tau - \tau')|u(\zeta, \tau')|^2\right] u(\zeta, \tau) \qquad [5]$$

where $L_{\mathrm{nl}} = (c/\omega_0 n_2 I_0)$ is the nonlinear length, $I_0 = n_0 c|A_0|^2/2\pi$ is the peak input intensity, f is the fraction of the Raman contribution to the nonlinear refractive index, and $R(\tau) = \{[1 + (\Omega_R \tau_R)^2]/\Omega_R \tau_R\} \exp(-\tau/\tau_R)\sin(\Omega_R \tau)$ is the Raman response function, τ_R is the characteristic Raman response time, Ω_R is the characterstic Raman frequency, and $\gamma_R = \tau_p/\tau_R$. For example, for fused silica $f = 0.15$, $\tau_R = 50$ fs, and $\Omega_R \tau_R = 4.2$. For a noble gas such as Xe, there is no Raman contribution and $f = 0$.

Self Focusing, Supercontinuum Generation, and Filamentation

The presence of the nonlinear refractive index with $n_2 > 0$ can lead to self-focusing of a laser beam. For sufficiently long pulses such that dispersion, self-steepening, and space–time focusing effects can be neglected, it is found that laser beams with input powers greater than the critical power $P_{\mathrm{cr}} = \alpha\lambda^2/4\pi n_0 n_2$ will undergo catastrophic self-focusing collapse. The dimensionless parameter $\alpha \geq 1.86$ depends on the spatial profile of the input beam and for a Gaussian input beam is given by $\alpha = 1.9$, in which case the ratio of the input power P to the critical power satisfies the relation $P/P_{\mathrm{cr}} = 1.055 L_{\mathrm{df}}/L_{\mathrm{nl}}$. Extensive studies have been made on the dynamics of laser beams undergoing self-focusing. For example, it has been shown that the shape of the collapsing beam evolves to a radially symmetric profile as it approaches the collapse point and that the total power contained in the collapsing portion always corresponds to the minimum value (i.e., $\alpha \sim 1.86$) regardless of the initial power in the beam.

For light pulses shorter than a picosecond, the role of material dispersion plays an important role and completely alters the dynamics of the self-focusing process. These dispersive effects lead to a temporal splitting of the pulse into two pulses and the arrest of its collapse. At even higher powers, other phenomena can occur, such as ‘optical shock’ formation at the rear edge of the pulse, due to self-steepening and space–time focusing. Shock formation leads to the emission of a broad spectrum of radiation extending from the ultraviolet to the mid-infrared, known as supercontinuum generation (SCG). This phenomenon was first observed in 1970, and since then it has been observed in many different solids, liquids, and gases, under a wide variety of experimental conditions.

If the peak intensity of the pulse approaches intensities of $> 10^{13}$ W/cm^2, either through self-focusing or external focusing, multiphoton ionization occurs and a plasma is formed in the medium. (See section below on Plasma Nonlinearities and Relativistic Effects.) The generation of the plasma lowers the refractive index and effectively counteracts the self-focusing process, resulting in the spatial confinement of the pulse for distances far beyond what would be allowed by ordinary linear diffraction and has been observed in gases, liquids, and solids. A striking example of this apparent self-waveguiding is the observation of 'light strings' in air which can extend more than 10 km into the atmosphere. This phenomenon was first observed with 100 fs laser pulses in the near infrared ($\lambda = 800$ nm). Researchers found that pulses with energies greater than 10 mJ undergo self-focusing collapse in air and produce a highly intense ($> 10^{13}$ W/cm^2) 100-micron-diameter light filament tens of meters long.

Multiphoton Absorption

Multiphoton absorption is a process in which an atom or molecule makes a transition from a ground state to an excited state by means of the simultaneous absorption of N photons. In the lowest order of perturbation theory, such a process can be described by means of a susceptibility of order $(2N - 1)$, that is, by $\chi^{(2N-1)}$. Alternatively, this process can be described in terms of an N-photon cross-section $\sigma^{(N)}$ defined such that the transition rate per atom is given by

$$R^{(N)} = \sigma^{(N)} I^N \qquad [6]$$

where I is the intensity of the laser field. Quantum mechanical expressions for the N-photon cross-sections are readily obtained. One finds, for instance, that

$$R^{(2)}_{\mathrm{ng}} = \left| \sum_m \frac{\mu_{\mathrm{nm}}\mu_{\mathrm{mg}}E^2}{\hbar^2(\omega_{\mathrm{mg}} - \omega)} \right|^2 2\pi\rho_{\mathrm{f}}(\omega_{\mathrm{ng}} - 2\omega)$$

$$R^{(3)}_{\mathrm{og}} = \left| \sum_{mn} \frac{\mu_{\mathrm{on}}\mu_{\mathrm{nm}}\mu_{\mathrm{mg}}E^3}{\hbar^3(\omega_{\mathrm{ng}} - 2\omega)(\omega_{\mathrm{mg}} - \omega)} \right|^2 2\pi\rho_{\mathrm{f}}(\omega_{\mathrm{og}} - 3\omega) \qquad [7]$$

In each of these expressions, the quantity ρ_{f} represents the density of final states, or equivalently the atomic lineshape function, evaluated at the N-photon transition frequency.

Optical Damage

High-intensity laser fields can produce unwanted damage to optical materials. As a point of reference, the threshold for laser damage to fused silica at a wavelength of 1.05 micrometers for a pulse of 30 ps duration corresponds to an intensity of 230 GW/cm^2 or a fluence of 7 J/cm^2. Over a wide range of pulselengths (approximately 1 ps to 1 μs), the threshold intensity for laser damage decreases with pulse length T as $T^{-1/2}$ and correspondingly the threshold fluence for laser damage increases with pulse length $T^{1/2}$. In this range of pulse durations, the dominant mechanism of laser damage is avalanche breakdown. In this process, free electrons are accelerated by the laser field until they acquire sufficient energy to impact-ionize other atoms in the sample. These additional electrons are similarly accelerated and create still more free electrons. The combined action of the breaking of chemical bonds and the deposition of heat energy leads to the fracturing of the optical material. For pulses shorter than 1 ps, processes such as multiphoton absorption and multiphoton dissociation contribute to the mechanism of optical damage. For laser pulses longer than approximately 1 μs (including continuous wave laser beams), the dominant damage mechanism is direct heating of the optical material by linear absorption.

High-Harmonic Generation

Let us consider how nonlinear optical effects are modified when excited by a super-intense pulse. Nonlinear optical effects are traditionally modeled using a power-series expansion, such as

$$P = \chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots \qquad [8]$$

but this series is not expected to converge if the laser field strength E exceeds the atomic unit of field strength $E_{\mathrm{at}} = e/a_0^2 = 2 \times 10^7$ statvolt/cm $=$ 6×10^9 V/cm. This field strength corresponds to a laser intensity of $I_{\mathrm{at}} = 4 \times 10^{16}$ W/cm^2, which constitutes the threshold intensity for exciting nonperturbative nonlinear optical response.

One of the consequences of excitation with intensities comparable to the atomic unit of intensity I_{at}, is the occurrence of high-harmonic generation. In a typical experimental arrangement, a gas jet is irradiated by high-intensity laser radiation, and all odd harmonics of the fundamental laser frequency, up to some maximum value N_{max}, are observed. The various harmonics below N_{max} are typically emitted with approximately equal intensity; such an observation is incompatible with a perturbative explanation of this phenomenon. Recent work has demonstrated harmonic generation with $N_{\max}$ as large as 341.

The phenomenon of high-harmonic generation can be understood in terms of a simple physical model. One imagines an atomic electron that has received kinetic energy from the laser field and is excited to a highly elliptical orbit. The positively charged atomic nucleus is at one focus of this ellipse, and each time the electron passes near the nucleus it undergoes strong acceleration and emits a short pulse of radiation. This radiation will occur in the form of a train of short pulses; the spectrum of the radiation will be the square of the Fourier transform of this pulse train, which will contain the odd harmonics of the oscillation frequency up to some maximum frequency, that is approximately the inverse of the time the electron spends near the atomic core. This argument can be made quantitative to show that the maximum harmonic number is given by

$$N_{\max}\hbar\omega = 3.17K + U_p \qquad [9]$$

where $K = e^2E^2/m\omega^2$ is the 'ponderomotive energy' (the kinetic energy of an electron in a laser field) and U_p is the ionization energy of the atom.

Plasma Nonlinearities and Relativistic Effects

The process of multiphoton ionization can liberate a sufficient number of free electrons to transform the optical medium into a plasma, that is, a fully or partially ionized gas. The process of plasma formation is described by the equation

$$\frac{\mathrm{d}N_\mathrm{e}}{\mathrm{d}t} = \frac{\mathrm{d}N_\mathrm{i}}{\mathrm{d}t} = (N_\mathrm{T} - N_\mathrm{i})\sigma^{(N)}I^N - rN_\mathrm{e}N_\mathrm{i} \qquad [10]$$

where N_e is the number density of electons, N_i is the number density of ions, N_T is the total number of atoms (both ionized and un-ionized) in the material, and r is the electron–ion recombination coefficient. The optical properties of plasmas are very different from those of typical dielectric materials; the plasma contribution to the dielectric constant is given by

$$\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \qquad [11]$$

where $\omega_p = \sqrt{4\pi Ne^2/m}$ is known as the plasma frequency.

Nonlinear effects can occur in the propagation of light through a plasma. One example is the nonlinear response resulting from the relativistic change in mass of the electron due to the large velocity that it can attain in the field of an intense laser beam. Detailed consideration of this effect shows that the nonlinear change in refractive index can be described as $\Delta n = n_2 I$ where

$$\begin{aligned} n_2 &= \frac{2\pi\omega_p^2 e^2}{n_0^2 m^2 c^3 \omega^4} \\ &= \frac{1}{2\pi n_0^2}\left(\frac{\omega_p}{\omega}\right)^2 1.1\times10^{-26}\,\frac{\mathrm{cm}^2}{\mathrm{W}} \end{aligned} \qquad [12]$$

Nonlinear Quantum Electrodynamics

One can imagine an electric field so intense that it could lead to the spontaneous creation of electron–positron pairs. Such a field would have a strength of the order of $E_\mathrm{QED} = mc^2/e\lambdabar_c$ where $\lambdabar_c = \hbar/mc$ is the reduced Compton wavelength of the electron. The intensity of a light beam with a peak field amplitude of E_QED is $I_\mathrm{QED} = 4\times10^{29}\ \mathrm{W/cm^2}$. Detailed consideration shows that even for fields weaker than I_QED there will be a field-induced change in the dielectric tensor given by

$$\epsilon_{ik} = \delta_{ik} + \frac{e^4\hbar}{45\pi m^4c^7}\left[2(E^2 - B^2)\delta_{ik} + 7B_iB_k\right] \qquad [13]$$

Because of the unusual tensor properties of this relation, it displays a different polarization dependence than typical optical nonlinearities. Nonetheless, to an order of magnitude one can describe the strength of this response as

$$n_2 = \frac{7}{15c}\,\frac{e^2}{\hbar c}\,\frac{1}{E_\mathrm{QED}^2} = 5.6\times10^{-34}\ \mathrm{cm^2/W} \qquad [14]$$

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Quantum Electrodynamics:** Quantum Theory of the Electromagnetic Field.

Further Reading

Agostini P, Fabre F, Mainfray G, Petite G and Rahman NK (1979) Free-free transitions following six photon ionization of xenon. *Physics Review Letters* 42: 1127.

Alfano RR (ed.) (1989) *The Supercontinuum Laser Source*. New York: Springer-Verlag.

Bloembergen N (1997) A brief history of light breakdown. *Journal of Nonlinear Optical Physics and Materials* 6: 377.

Boyd RW (2003) *Nonlinear Optics, Second Edition*, Section 12.5. Amsterdam: Academic Press.

Braun A, Korn G, Liu X, Du D, Squier J and Mourou G (1995) Self-channeling of high-peak-power femtosecond laser pulses in air. *Optical Letters* 20: 73.

Brabec T and Krausz F (2000) Intense few-cycle laser fields: Frontiers of nonlinear optics. *Reviews of Modern Physics* 72: 545.

Chang Z, Rundquist A, Wang H, Murnane MM and Kapteyn HC (1997) Generation of coherent soft X-rays at 2.7 nm using high harmonics. *Physics Review Letters* 79: 2967.

Corkum PB (1993) Plasma perspective on strong field multiphoton ionization. *Physics Review Letters* 71: 1994.

Corkum PB, Rolland C and Srinivasanrao T (1986) Supercontinum generation in gases. *Physics Review Letters* 57: 2268.

Gaeta AL (2003) Collapsing light really shines. *Science* 301: 54.

Kasparian J, Rodriguez M, Mejean G, Yu J, Salmon E, Wille H, Bourayou R, Frey S, Andŕe Y-B, Mysyrowicz A, Sauerbrey R, Wolf J-P and Woste L (2003) White light filaments for atmospheric analysis. *Science* 301: 61.

Lewenstein M, Balcou Ph, Ivanov M Yu, L'Huillier A and Corkum PB (1994) Theory of high-harmonic generation by low-frequency laser fields. *Physics Review A* 49: 2117.

Max CE, Arons J and Langdon AB (1974) Self-modulation and self-focusing of electromagnetic waves in plasmas. *Physics Review Letters* 33: 209.

Moll KD, Fibich G and Gaeta AL (2003) Self-similar wave collapse: observation of the Townes profile. *Physics Review Letters* 90: 203902.

Monot P, Auguste T, Gibbon P, Jakober F, Mainfray G, Dulieu A, Louis-Jacquet M, Malka G and Miquel JL (1995) Experimental demonstration of relativistic self-channeling of a multiterawatt laser pulse in an under-dense plasma. *Physics Review Letters* 74: 2953.

Rairoux P, Schillinger H, Niedermeier S, *et al.* (2000) *Applied Physics B—Lasers Optics* 71: 573.

Ranka JK, Schirmer RW and Gaeta AL (1996) Observation of pulse splitting in nonlinear dispersive media. *Physics Review Letters* 77: 3783.

Rothenberg JE (1992) Pulse splitting during self-focusing in normally dispersive media. *Optical Letters* 17: 583.

Sprangle P, Tang C-M and Esarey E (1987) Relativistic self-focusing of short-pulse radiation beams in plasmas. *IEEE Transactions on Plasma Science* 15: 145.

Stuart BC, Feit D, Rubenchik AM, Shore BW and Perry MD (1995) Laser-induced damage in dielectrics with nano-second to subpicosecond pulses. *Physics Review Letters* 74: 2248.

Wagner R, Chen S-Y, Maksemchak A and Umstadter D (1997) Electron acceleration by a laser wakefield in a relativistically self guided channel. *Physics Review Letters* 78: 3125.

Wood RM (1986) *Laser Damage in Optical Materials*. Bristol, UK: Adam Hilger.

NONLINEAR SOURCES

Harmonic Generation in Gases

P Villoresi, University of Padova and INFM, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The generation of laser harmonics in gases provides high-brightness ultraviolet and soft X-ray radiation induced by an intense laser pulse. This phenomenon was discovered experimentally in the late 1980s, in a simple setup as that sketched in Figure 1. The laser pulse is focused in the interaction medium, which may be a gas jet formed by a pulsed electromagnetic valve, a capillary, or a small chamber filled with the gas. The beam leaving the interaction region includes a part of the laser pulse and a number of odd harmonics.

The harmonic orders generated in this way may reach and exceed an order of 300, so that this mechanism, often referred as high-order harmonics (HHs) generation (HHG), directly up-converts low frequency radiation, usually in the near-infrared spectrum, to extreme-ultraviolet (XUV) or soft X-rays. An example of a HH spectrum is shown in Figure 2, where several tens of discrete harmonic peaks are present. The striking difference with ordinary nonlinear processes is the flatness of the photon yield versus nonlinear order, up to a sharp cut-off point.

The HHG is based on the sequence of three processes: the electric field of an intense laser pulse ionizes atoms placed near the laser focus; the liberated electrons are driven by the laser electric field in a trajectory which returns on the parent ion; and the electron recombines releasing a photon whose energy is the ionization potential plus the electron's kinetic energy. The process is periodic in time, with the period of the laser pulse, and the emitted spectrum is then expressed by discrete components, multiples of the laser frequency. The emission is in the form of a traveling wave in the direction of the laser beam and so only odd harmonics are present.

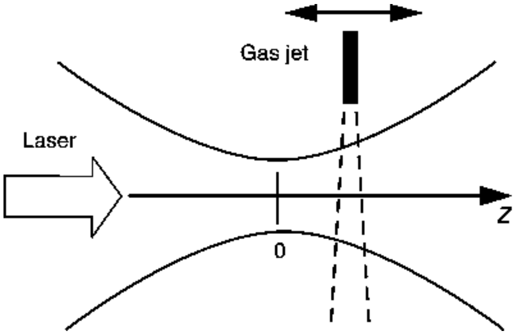

Figure 1 Scheme of the interaction between a focused laser pulse and a gas jet. *z* axis originates at laser focus.

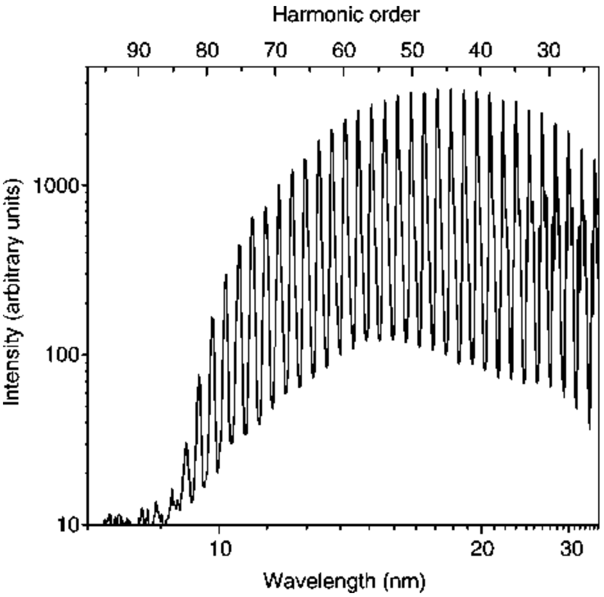

Figure 2 HHs experimental spectrum, obtained from helium gas, and a Ti:sapphire laser pulse at 790 nm, with 25 fs duration and 6×10^{14} W/cm^2 intensity.

The individual processes in HHG have highlighted the coherent response of electrons in the laser field, that show up interferences between quantum trajectories, as well as a nonlinear phase-matching issue – which have provided a simple method to investigate the interaction of single particles with the electromagnetic field, where plasma effects and bonds may be of minor influence.

Background

The interest to produce laser harmonics using gases instead of crystals was expressed by many. For the gas medium; the mass-density is three order of magnitude lower than for solids and the nonlinear susceptibility can be strongly enhanced for suitable resonance frequencies. Moreover, the phase-matching of the phase velocities may be achieved by tuning the gas pressure or by mixing diverse gases, so realizing a highly homogenous conversion region, free from walk-off and with an extension much longer than for crystals. In addition, gases have a high threshold for the intensity-induced damage, which is also non-permanent.

This research was made possible with the introduction of chirped-pulse laser amplifiers in the early 1980s, when picosecond lasers pulses reached intensities capable of directly ionizing solids and matter with their own electric field. This effect, known as optical ionization, was further enhanced when the laser pulses were reduced to subpicosecond duration, up to a few femtoseconds (10^{-15} s), close to their natural limit, which is the optical cycle. The experimental observation of HHs and of the ejection of electrons with unexpected high energy, the above threshold ionization (ATI), dramatically changed the scenario of laser matter interaction and has indicated a new regime in the coupling of laser radiation with the matter. The ordinary nonlinear optics, using the perturbative approach, was not capable of interpreting the findings and so the request was made for new models for the description of the interaction, which has been now named nonperturbative.

Typical Experimental Parameters

In the experimental investigation of the HHG, many laser wavelengths and pulse durations were used, including radiation from excimer, visible, Ti:sapphire, Nd:YAG (first and second harmonics) and CO_2 lasers, and pulses of duration from hundreds of picoseconds to a few femtoseconds. Among these, the more effective strategy in the generation of high harmonic order with high conversion efficiency, has been that of near infrared pulses with shortest duration. In order to induce optical single-ionization, the required laser intensity needs to span in the 10^{14} W/cm^2 range. Different elements are a possible choice for the gas in the interaction medium, though noble gases are used for the generation of the highest orders, since these have the higher ionization potential. Due to the very short duration of the laser pulse, the gas temperature is not relevant, while a pressure range of 10–500 mbar is usually chosen to balance, on one side, the needs of a large number of atoms that take part in the HHG and on the other side, the limit posed by the attenuation of the XUV and soft X-ray radiation after the generation, caused by the gas absorption. For the same reason, a powerful pumping system is needed, in order to reduce the ambient pressure in the apparatus below 10^{-5} mbar level.

The Models for the High Harmonics Generation

The physical processes involved in HHG are many: in the emission of the HH photons, the laser field interacts with a single atom of gas and induces the ionization, the electron's motion and its recollision with its parent ion; besides, the beam of HHs is the result of the superposition of the radiation from all the ionized atoms, as in an ordinary nonlinear medium, where the phase velocity of the laser field and of HHs is a function of the position and of the laser intensity and phase; therefore, the phase-matching issue has to be addressed in order to understand the global yield of HHG. The more relevant causes for the mismatch are the geometrical phase shift at the beam focus, which for Gaussian-profile beams is the Gouy phase, the nonlinear phase from the atomic polarization, and the effects of the plasma induced by the laser pulse itself. In addition, other effects have to be considered, such as the nonuniform density distribution of the target gas or the fact that the propagation of XUV radiation through a gas is usually influenced by a strong absorption, which is frequency dependent.

The study of HHG is first approached by introducing what is known as the simple man model, in which simplifications are introduced in order to treat the problem in a simple classical scheme, but from which the essential characteristics are derived. We then introduce the more complex quantum model based on the path integrals.

First Step: Ionization

The first step in the HHG model is the process with which the laser electric field applies to an electron, belonging to an atom or a molecule, a force that causes its detachment. This takes place only if such a field is strong enough to liberate the electron from its atomic bond. This may happen when the laser, which is taken here as linearly polarized and periodic in time, has modified the potential well so that the electrons may tunnel through it or pass over it. In fact, the sum of the atomic potential with that associated to the laser, which can be thought of as a plane that swings between a positive and negative slope at the laser frequency, gives rise to a finite sized well, through which the electron escapes by tunneling, or eventually to an aperture from which the electron is drawn away from the nucleus. In Figure 3, is shown such a well and the tunnel path for the electron.

Once freed, the electron responds to the periodic force from the laser which makes it oscillate. The average kinetic energy on such oscillations is known as the ponderomotive potential U_p, and is related to the laser intensity and wavelength, according to the following: $U_p = e^2E_0^2/4m\omega_0^2$, which reads numerically as $U_p(\text{eV}) = 9.3I_0(10^{14}\ \text{W/cm}^2)\lambda^2(\mu\text{m})$, where λ is the laser wavelength, E_0 and I_0 the laser peak electric field and intensity.

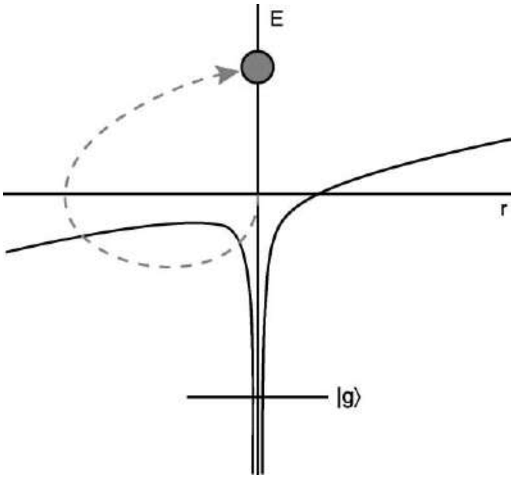

Figure 3 Electron potential when the laser pulse shines on the atom; |g⟩ indicated the level of the ground state; in red, a possible trajectory of the liberated electron after tunneling.

Second Step: Free-electron Trajectory

On the freed electron are applied both the time-varying force due to the laser field and the ionic potential. Its motion can be approximated according to the simple man model as a nonrelativistic classical-mechanics trajectory; its initial velocity is set to zero, and its initial position is by the ion; finally, the laser exerts a force much stronger than that from the ion as well as from the laser magnetic field, thus these two latter forces are cancelled out.

As mentioned above, the electron trajectories that induce HHs are those where the electron returns by the ion with strong kinetic energy. If the laser is linearly polarized, the induced trajectory takes place along a line, at which also belongs to the parent ion. Moreover, the electron final velocity is related to the actual instant of ionization, or phase of the field, with respect to the laser peak. Therefore, electron final energies span from zero to a maximum value $K_{\text{max.}}$ By considering the first semi-cycle, $K_{\max}$ occurs at about 17° of field phase angle, while for the second semi-cycle it is at 197°.

In Figure 4 the electron trajectories for different ionization in the first semi-cycle are shown: the electron displacement in the laser-field direction with

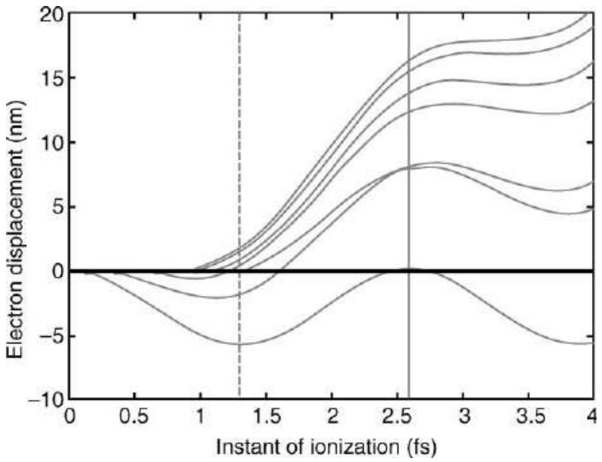

Figure 4 Electron trajectories relative to different ionization instant within the first semi-cycle; in dotted red, half-cycle, in continuous red, the whole optical cycle. Trajectories for later times do not return to the ion, do not contribute to HHG and correspond to ejected photoelectrons.

respect to the ion position, which is the initial as well as the final position, is represented versus time, where the time origin is at the laser peak, where field phase is equal to zero. The laser intensity directly influences the value of $K_{\max}$, according to the following expression: $K_{\max} = 3.17\,U_p$. In the case shown in the figure, the peak laser intensity is $I_0 = 8 \times 10^{14}$ W/cm^2 and the ponderomotive potential is $U_p = 45$ eV and $K_{\max} = 143$ eV.

Moreover, for a given final energy below $K_{\max}$, there are more than one value of the phase of the laser field where the trajectory results with the same final kinetic energy. This indicates that multiple trajectories may produce a given HH, and this is relevant in terms of the interference between them.

Third Step: Recombination

The final process which occurs in the HHG, is the recombination of the ionized electron with the parent ion. The energy conservation means that the kinetic energy and the ionization potential have to be delivered to the emitted photon. In the simple man model, the spectrum of HH then spans from I_p to the maximum values of $E_{\max} = I_p + 3.17\,K_{\max}$. In the case of Figure 2, obtained using helium gas at 6×10^{14} W/cm^2, the spectral position of cutoff can be estimated at about 9 nm, or 140 eV, which is in good agreement with the prediction of the relationship described above.

The electron recombination is not a unique process occurring at this point: the electron may be scattered by the ion and be emitted with a kinetic energy of up to 10 U_p, which is known as the rescattering plateau in ATI, or it may ionize the ion by collision, leaving a double ionized atom via sequential ionization, with a strong increase in the observed cross-section of this process.

Feynman Path-Integral Approach

The ideas in the simple man model have been expanded in order to include the quantum mechanical description of the laser-electron-ion interaction. By using the Feynman path-integral approach, the evolution of the electron driven by the laser force is associated to a probability which depends on the sum of an infinite number of amplitudes $\mathbf{f}_n$. These are complex numbers and correspond to different trajectories, or quantum orbits, $\mathbf{r}_n(t)$, followed by the electron. Their phase is proportional to the classical action of the electron motion, $S[\mathbf{r}_n(t)]$, calculated along $\mathbf{r}_n(t)$, which begins at time t_i and ends at a later time $t_f = t_i + \tau$. Here τ defines the duration of the liberated electron trajectory. In our case, the action is the sum of three terms:

$$S[\mathbf{r}_n(t)] = S_{\text{bound},n} + S_{\text{free},n} + S_{\text{final},n} \qquad [1]$$

that correspond respectively to the electron ionization:

$$S_{\text{bound},n} = -E_0 t_{i,n} \qquad [2]$$

where E_0 is the binding energy of electron in the atom ground state, $E_0 < 0$, to the free electron trajectory:

$$S_{\text{free},n} = \int_{t_i}^{t_f} [\mathbf{k} - e\mathbf{A}(t)^2]\mathrm{d}t \qquad [3]$$

where $\mathbf{k}$ is the electron momentum after ionization, directed parallel the laser electric field, and $\mathbf{A}(t)$ is the vector potential of the laser field, and to the final recombination with the parent ion:

$$S_{\text{final},n} = (\hbar\omega_q + E_0)t_{f,n} \qquad [4]$$

The condition that normally points out the trajectory made by the electron may be obtained from the principle of least action. Within the path integral approach, this principle can be used to deduce some equations which have to be obeyed by the quantum orbits, and that express the energy conservation in both the ionization and recombination process and the fulfilment of boundary conditions for electron trajectory. These are known as saddle-point equations, and reads as:

$$[\mathbf{k} - e\mathbf{A}(t_i)]^2 = 2mE_0 \qquad [5]$$

for the ionization process:

$$\mathbf{k}\tau = e\int_{t_i}^{t_f} \mathbf{A}(t)\mathrm{d}t \qquad [6]$$

which compares the initial momentum with that exchanged with the laser field during the trajectory, and

$$[\mathbf{k} - e\mathbf{A}(t_f)]^2 = 2m(\hbar\omega_q + E_0) \qquad [7]$$

for the recombination process which emits an harmonic of order q.

From eqn [5] we note that the solution for the ionization time t_i is a complex number, since the right hand side is negative. This suggests that the quantum orbits are complex as well. The reason for this is that the tunnel through the potential well is a nonclassical process.

The orbits solutions to eqns [5]–[7] have different durations, and for some of them τ is shorter than an optical cycle, while for the others the electron returns in a later cycle than the one where it was ionized. This latter solution is of minor influence, since the electron wavefunction spreads out along the trajectory, decreasing the recombination probability. Those within the initial optical cycle sum the amplitudes to give the probability of the generation of q-th harmonic, and their interference is evident from the irregular single-atom spectrum.

The time evolution of the laser pulse as well as the interaction geometry are experimental tools that influence the relevance of the different trajectories. In particular conditions it is possible to set the condition for HHG from a single trajectory, with a very simple phase structure and clear spectrum; some cases are described below.

HHs Generated by Few-Optical-Cycle Pulses

The laser electric field considered so far has been taken as stationary, relying on the hypothesis that the pulse envelope has a duration much longer than the laser optical cycle. For the analysis of the electron trajectories, that last a fraction of the laser optical cycle, this hypothesis is valid for pulses down to about 20 fs. For shorter pulses, as those generated with the hollow fiber technique, there is a noticeable variation of the electric field between a cycle and the next. The generation process is now aperiodic, and the characteristics of HH spectrum and conversion is strongly influenced by the pulse actual duration.

In order to show the striking changes in this regime, Figure 5 shows the laser electric field, the envelope and the ionization probability of helium gas in the case of a pulse of 5 fs of full width at half-maximum (FWHM) duration and a laser intensity of 9×10^{14} W/cm^2. The atom can be ionized during only three intervals, of sub-fs duration but where only one is dominant.

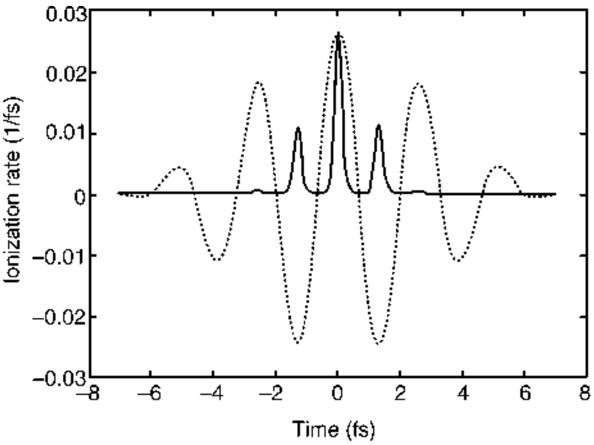

Figure 5 Ionization rate for helium gas shined by a 5 fs FWHM pulse of intensity 9×10^{14} W/cm^2. The dotted line shows the normalized evolution of electric field of the pulse. Only the three central half-cycles induce ionization.

The HHG experiment in this temporal regime have revealed that the influence of the plasma in the interaction region is progressively reduced and eventually vanishes for pulses of duration of about 5 fs. This is because the atoms are exposed directly to the maximum electric field, and so there is no pre-ionization of the gas caused by the initial part of the pulse. This effect, which is present in the case of many-cycle pulses, causes a reduction in the HHG efficiency, due to a depletion of the neutral density available at the time of maximum electric field, where the higher HHs are produced with a large dispersion, which deteriorates the phase-matching.

Moreover, the effect of the carrier-envelope-phase in the HHs spectrum has been demonstrated as an evident spectral shift of the cut-off harmonics generated with few-cycle pulses. The reason can be found in the variation of the phase of the emission from each electron after its trajectory. For the few-cycle pulses, this phase changes for every laser semi-cycle, due to the change in the actual electric field which drives the electron motion.

This portion of the spectrum is also relevant to its phase structure, since here the HHs are generated by a single electron trajectory. This condition sets a constant phase difference between the subsequent harmonic orders. In this way, similarly to the mode locking laser, this effect allows the synthesis of a pulse, from the combination of several harmonic orders, which has a duration much lower than the individual harmonics. This process has been exploited to produce a soft X-ray pulse of duration of 650 as (1 as, attosecond = 1×10^{-18} s).

HHs Beam Characteristics

The HHG is a coherent process, in that the emitted radiation has a direct relationship with the laser pump. The emitted HH beam also reflects this fact in its good directional characteristics and in its spatial and temporal coherence, as different experiments have demonstrated.

Using a 110 fs laser pulse, the interference fringes produced in the far field of two spatially separated HH beams were measured as a function of the mutual delay. The HH coherence time t_c resulted in agreement with the expected duration of the HHs, which is a fraction of that of the laser pulse. Moreover, it has been observed at a different t_c in the central and external regions of the far field of harmonic 23rd, 50 fs and 20 fs respectively. This is because different phase-matching in the HHG contribute to the on-axis and off-axis part of the HHs beam, where the more prominent quantum orbit is also different.

The measurement of the beam divergence for different harmonic orders is relevant for both the estimate of the beam brightness and the irradiance available in the case of the HH applications. Moreover, it is useful for the understanding of the phase-matching which is realized in actual interaction geometry. The experimental measurements requires an astigmatic spectrometer, as is described later. In this kind of experiment, the focalplane images of the HH spectrum are taken for different generation conditions, usually moving the gas-jet position with respect to the laser focus, the z coordinate in Figure 1. The harmonic peaks appear as elongated spots, as shown in Figure 6 for helium gas and an intensity of 4×10^{14} W/cm^2 and duration of 20 fs, whose length in a transverse direction to the dispersion plane can be related, by a optical calculation, to the beam divergence.

A comparison between divergences for HHG driven by few- or many-optical-cycle pulses has revealed another aspect of the phase-matching condition. In this case, the laser pulse was compressed in time by means of the hollow fiber technique, and so the resulting phase front is of the Bessel-beam type, truncated at the first zero. The experiment indicates a divergence for plateau harmonics of about 3 mrad FWHM of the beam, in the case of the 7 fs pulse and larger values by more than 50% using the 20 fs pulse. In both cases, the beam divergence increase moving downstream the gas jet and with increasing harmonic order. The reduced divergence can be ascribed to a more regular wavefront of the harmonic beam, which can be related to the simpler phase structure of the few-cycle case.

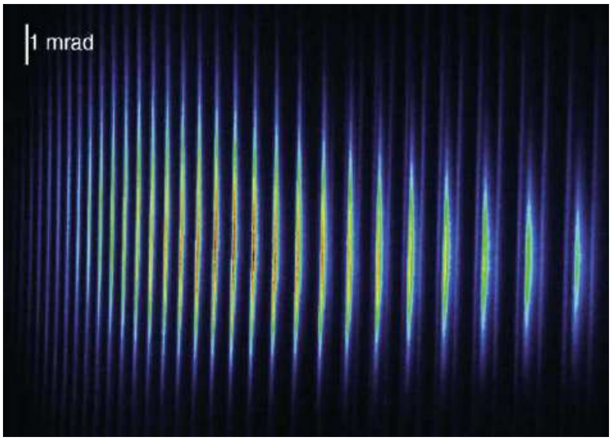

Figure 6 Focal plane image obtained from the spectrograph shown in Figure 9, relative to the HHs spectrum of helium with 20 fs and 4×10^{14} W/cm^2. The bar in upper-left corner indicates the scale-length of vertical axis corresponding to a divergence of the HH beam of 1 mrad.

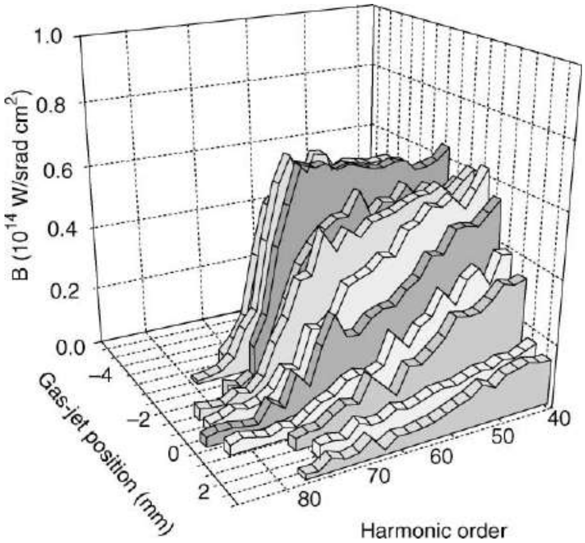

Figure 7 Peak brightness of the HHs beam determined from absolute-intensity and divergence measurements, using neon gas and a 790 nm laser pulse with 7 fs duration and 9×10^{14} W/cm^2 intensity.

By using an intensity-calibrated detector, the number of photons per harmonic order per laser shot can be measured. Combining this measure with that of the divergence, the HH beam brightness can be calculated. The source size and the HH pulse duration can be easily modeled from the experimental parameters. The result for the case of neon with 7 fs pump-pulse with intensity of 9×10^{14} W/cm^2 is shown in Figure 7, as a function of the gas-jet position, with positive values when the jet is downstream with respect to laser focus.

The large divergence reduces the brightness of downstream positions. Maximum values of about 8×10^{13} W/srad cm^2 are found close to the focus, relatively flat with respect to harmonic order.

The conversion efficiency η_{HH} from the fundamental intensity to that of the HHs can also be derived from this kind of measurement. The more convenient parameters for HHG in the range 41–71 were found using neon gas jet from a pulsed valve and a pulse duration of 7 fs, which leads to η_{HH} in the scale of 10^{-7}, while for lower orders, from 17 to 31, a value about two orders of magnitude larger has been obtained using a short hollow fiber filled with argon gas. A remarkable intracycle optimization mechanism was introduced in this case, in order to maximize the given harmonic order of chosen. This was done by tailoring the shape of the pump pulse in spectrum and time, using an adaptive component in the laser compressor. In this case, the optimal pulse shape to maximize the efficiency in the driving of the electron on a trajectory producing the required harmonic order, is sought by means of a genetic algorithm.

A strong efficiency enhancement has also been demonstrated by manipulating the laser wavefront by means of a deformable mirror (DM). In this case, the DM was controlled by a genetic algorithm, aiming to increase the HHs intensity in a given spectral interval. The purpose of this technique is the correction of the optical aberrations on the beam, introduced by both the optical components in the beam focusing line and the phase distortion caused by nonlinear effects leading to an optimum laser front. The effect is shown in Figure 8, where the spectrum obtained with the original beam is compared with that optimized by the genetic algorithm. It can be noticed there is both a remarkable extension of the HHs spectrum, of about 20 orders, as well as a strong enhancement of the plateau intensity.

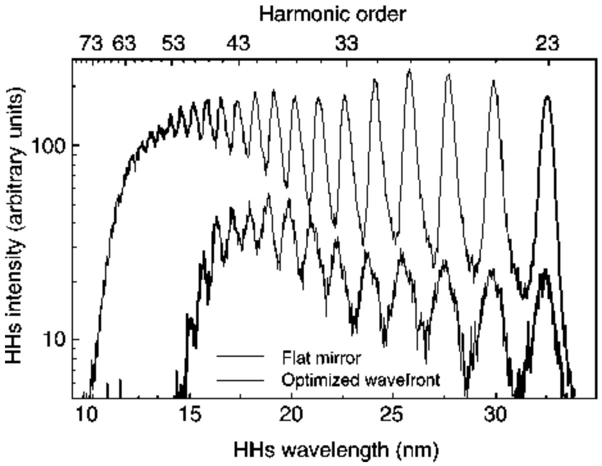

Figure 8 Comparison of the HHs spectrum obtained with and without the wavefront correction using a deformable mirror. Neon gas was used and a 790 nm laser pulse with 6 fs duration and 0.3 mJ energy.

Spectrometers and Monochromators for HHs

The extended spectrum of HHs reaches the optical domains of the extreme-ultraviolet (XUV, 10–100 nm) and soft X-ray (1–10 nm). In these regions, the optical materials have very poor or zero transmission, and so the optical design of instruments for HHs analysis uses reflective optics only. Exceptions sometimes used are the frequency-selective XUV transmission filters, whose thickness does not exceed a few tenths of a micrometer, and the diffracting zone-plates. Moreover, for the radiation of wavelength shorter than 30 nm, even the reflectivity of optics becomes poor, and this forces the use of grazing-incidence mountings or the narrow band XUV and soft X-ray multilayer mirrors.

On the other hand, HHs as optical sources have favorable characteristics, in that they are emitted in a region smaller than the laser waist focus, so usually of a few tens of micrometers, and their divergence is small, as seen above. In addition, the very short duration of the pump pulses broadens the HHs spectrum, due to time-bandwidth inequality, usually smearing details below 0.1 nm of width.

The spectrograph is then realized with the least number of components, usually combining in one or two optics of the spectral dispersion, the gathering of the HH beam and its focusing on the detector. Moreover, it is not feasible to collect with a single detector all the broad HH spectrum in once, so usually a spectral interval of interest is selected and acquired with a linear or bidimensional detector, or scanned sequentially by rotating the grating.

An example of spectrograph with 2D-detector, is shown in Figure 9, the layout of the instrument developed for the simultaneous measure of the divergence and the intensity of HHs generated from few-optical-cycle pulses.

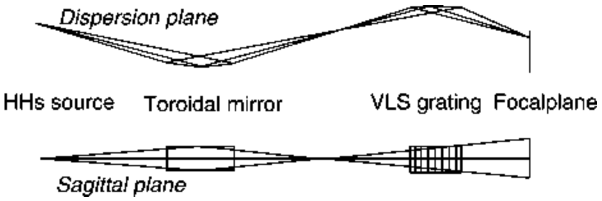

Figure 9 Optical layout of the astigmatic spectrometer for the divergence measurement.

The upper part of Figure 9 shows the optics in the dispersion plane, where the toroidal mirror gathers the HH radiation and relays it to the source point for the spherical grating. The latter has a varied line spacing (VLS) of the grooves, in order to image an erected spectrum in the focalplane. This feature is convenient in order to use a bidimensional detector. In the sagittal plane, the grating has negligible effect, and the toroidal mirror is designed to realize an astigmatic image of the source in the focalplane, from which the divergences can be extracted. The detector needs to be blind at the radiation of the laser and at the lower harmonics, and to have an extended linear response in the XUV and soft X-ray. A good solution can be obtained with an open micro-channel plate (MCP), with a suitable photocathode such as MgF_2, placed in the focalplane, coupled with a high-resolution and a high dynamic CCD camera.

For the selection of a single harmonic order or of a narrow spectral region, an XUV monochromator is used. This device is based on the spectral dispersion of the HHs beam by using a diffraction grating. The desired harmonic order is selected by scanning the wavelength which falls on the slit, usually by rotating the grating. This technique uses a single optical component, a concave grating, to reduce the reflection losses. On the other hand, the single-grating monochromator introduces a remarkable time broadening in the selected radiation. In fact, the grating diffraction is based on the optical path length difference for the rays diffracted by subsequent grating grooves. Therefore the selected beam, of wavelength λ, is composed of rays whose difference in length is $\Delta L = N_g m \lambda$, where m is the diffraction order and N_g is the number of illuminated grooves, given by the beam diameter on grating multiplied by the groove density. The time broadening of the selected pulse is then $\Delta\tau = \Delta L/c$. For a HH XUV pulse, the diffracted pulse results is 7 ps of duration, if beam size is of 3 mm and it is diffracted at 87° by a 1200 gr/mm grating.

The solution to this problem may be found in adopting a twin grating design. In this case, the dispersion induced by the first grating is used to select the desired spectral portion of the HH spectrum, and will be cancelled by the second grating. This latter grating has to be mounted in order to compensate for the different optical path length of different rays within the aperture of the HH beam. The layout of this optical scheme is reported in Figure 10. The ray tracing of this compensated monochromator indicates that for both the grazing incidence and the normal incidence regions, the residual broadening results is of the order of one femtosecond.

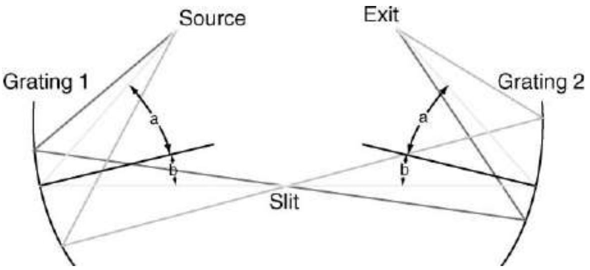

Figure 10 Scheme of the compensated XUV monochromator, for the selection of a spectral portion of the HHs spectrum without introducing temporal broadening. See also Villoresi P (1999) Compensation of optical path lengths in extreme-ultraviolet and soft X-ray monochromators for ultrafast pulses. *Applied Optics* 38: 6040.

HHs at Work: Applications and Perspectives

The advantages of HHs with respect to other XUV and soft X-ray sources, as laser-plasma, synchrotron radiation or X-ray lasers, are mainly the very short duration of their emission, their high brightness, and the source compactness and their intrinsic synchronization with a laser pulse. These aspects have been already exploited in a few experiments, based either on the detection of a photoelectron emitted in the time-dependent interaction of the matter with the HHs plus fundamental laser beams, or the observation of changes in the HHs XUV spectrum.

As examples of direct application, the interferometry of a laser-plasma has been obtained with HHs, to determine spatial distribution of the electron density. In this case, two beams of harmonic radiation were generated by splitting in equal parts the laser pulse, with one of them passing through a laser-generated plasma from a solid aluminum target. By means of a grating monochromator, a single harmonic order is selected, of both beams, and their interference pattern in the far field is acquired. From the shift in the fringes pattern, the map of the plasma electron density was estimated. Here, the broad spectrum of the HHs was exploited to operate the interference at the most convenient wavelength, far from atomic resonances and with fair fringe contrast.

The direct monitoring of a chemical reaction at the surface was achieved by pumping, with a part of the laser pulse, a platinum surface with some adsorbed molecular oxygen. By observing changes in the spectrum of the photoelectrons generated by the delayed HHs beam, the oxygen-platinum bond switch from a superoxo (O_2^-) to a peroxo (O_2^{2-}) state was observed in the time domain.

The electron bond in an atom may be dressed by a laser field, and the changes induced in its orbital are diagnosed by a HHs beam, which photo-ionizes the atom. This phenomenon have been exploited as a diagnostic of the temporal structure of a HHs beam: the sidebands in the photoelectron spectrum have been related to the phase of different harmonics. The findings indicate that HHs are locked in phase and their superposition produces a train of 250 as pulses.

By using the mixing scheme of some suitable orders of a HHs beam generated with few-optical-cycle pulses, mentioned already in section 'HHs beam characteristics,' a pulse with sub-femtosecond duration may be generated. This was used in the detailed analysis of the photoelectron spectrum in a pump-probe experiment, with time resolution of 150 as. Again using the IR and HHs radiation together, this technique has accessed the direct probe of the laser field oscillation, as demonstration by the use of HHs in the new domain of attosecond spectroscopy and metrology.

Finally, the HH radiation can also be considered as the seed for an amplifier of radiation in the soft X-ray region, as in the case of a free electron laser.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Instrumentation:** Spectrometers. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Bartels R, Backus S, Zeek E, *et al.* (2000) Shaped-pulse optimization of coherent emission of high-harmonic soft x-ray. *Nature (London)* 406: 164.

Bauer M, Lei C, Read K, *et al.* (2001) Direct observation of surface chemistry using ultrafast soft-X-ray pulses. *Physics Review Letters* 87: 025501.

Brabec T and Krausz F (2000) Intense few-cycle laser fields: frontiers of nonlinear optics. *Review Modern Physics* 72: 545.

Corkum PB (1993) Plasma perspective on strong-field multiphoton ionization. *Physics Review Letters* 71: 1994.

Descamps D, Lynga L, Norin J, *et al.* (2000) Extreme ultraviolet interferometry measurements with high-order harmonics. *Optic Letters* 25: 135.

Hentschel M, Kienberger R, Spielmann CH, *et al.* (2001) Attosecond metrology. *Nature (London)* 414: 509.

Kulander KC, Schafer KJ and Krause JL (1993) Dynamics of short-pulse excitation, ionization and harmonic conversion. In: Piraux B, L'Huillier A and Rzazewski K (eds) *Super-Intense Laser-Atom Physics*, pp. 95–110. New York: Plenum.

Lewenstein M, Balcou Ph, Ivanov M, *et al.* (1994) Theory of high-harmonic generation by low-frequency laser fields. *Physics Review A* 49: 2117.

L'Huillier A and Balcou P (1993) High-order harmonic generation in rare gases with a 1-ps 1053-nm laser. *Physics Review Letters* 70: 774.

McPherson A, Gibson G, Jara H, *et al.* (1987) Studies of multiphoton production of vacuum-ultraviolet radiation in the rare gasses. *Journal of the Optical Society of America B* 4: 595–601.

Nisoli M, Priori E, Sansone G, *et al.* (2002) High-brightness high-order harmonic generation by truncated bessel beams in the sub-10-fs regime. *Physics Review Letters* 88: 033902.

Nisoli M, Sansone G, Stagira S, *et al.* (2003) Effects of carrier-envelope phase differences of few-optical-cycle light pulses in single-shot high-order-harmonic spectra. *Physics Review Letters* 91: 213905.

Paul PM, Toma ES, Breger P, *et al.* (2001) Observation of a train of attosecond pulses from high harmonic generation. *Science* 292: 1689.

Poletto L, Tondello G and Villoresi P (2001) High-order laser harmonics detection in the EUV and soft x-ray spectral regions. *Review Scientific Instruments* 72: 2868.

Salières P, L'Huillier P, Antoine A, *et al.* (1999) Study of spatial and temporal coherence of high order harmonics. In: Bederson B and Walther H (eds) *Advances in Atomic, Molecular, and Optical Physics*, vol. 41, pp. 83. New York: Academic Press.

Villoresi P, Ceccherini P, Poletto L and Tondello G (2000) Spectral features and modeling of high-order harmonics generated by sub-10-fs pulses. *Physics Review Letters* 85: 2494.

Villoresi P, Bonora S, Pascolini M, *et al.* (2004) Optimization of high-order harmonic generation by adaptive control of sub-10 fs pulse wavefront. *Optic Letters* 29: 207.

O

OPTICAL AMPLIFIERS

Contents

Basic Concepts
Erbium Doped Fiber Amplifiers for Lightwave Systems
Optical Amplifiers in Long-Haul Transmission Systems
Raman, Brillouin and Parametric Amplifiers
Semiconductor Optical Amplifiers

Basic Concepts

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The transmission distance of a fiber-optic communication system is limited by fiber loss and dispersion. For long-haul lightwave systems, the loss limitations have traditionally been overcome by periodic regeneration of the optical signals at repeaters applying conversion to an intermediate electric signal. Because of the complexity and high cost of such regenerators, the need for optical amplifiers became obvious in the mid-1980s. The optical amplifier is ideally a transparent box that provides gain and is also insensitive to the bit rate, modulation format, power and wavelength of the signal passing through it.

Several means of obtaining optical amplification has been suggested since the 1970s, including direct use of the transmission fiber as gain medium through nonlinear effects, semiconductor amplifiers, or doping optical waveguides with an active material (rare-earth ions), that could provide gain. Due to the spectacular results on erbium-doped fiber amplifiers, which are particularly suitable in the third transmission window (around 1.5 μm), an intense worldwide research activity on optical amplifiers has developed. As a consequence, the development of erbium-doped fiber amplifiers has reached an industrial level, and commercial devices are now available.

Semiconductor amplifiers, on the other hand, have the same technical basis as semiconductor lasers. Although the strong nonlinearity of semiconductor amplifiers degrades the performances of transmission systems, the state-of-art semiconductor devices seem to be the most interesting amplifiers for transmission in the second transmission window (around 1.3 μm).

Amplifier Gain and Bandwidth

In a perfect amplifier, the amplification process would be insensitive to the bit rate, modulation format, power, state of polarization, wavelength, and optical bandwidth of the signal passing through it. On the other hand, no interaction would take place if more than one signal were amplified simultaneously. In practice, however, the optical gain depends not only on the wavelength (or frequency) of the incident signal, but also on the electromagnetic field intensity at any point inside the amplifier. Details of wavelength and intensity dependence of the optical signal depend on the amplifying medium.

We consider a case in which the gain medium is modeled as a homogeneously broadened two-level system. In such medium, the gain coefficient (i.e., the gain per unit length) can be written as:

$$g(\nu, P) = \frac{g_0}{1 + \dfrac{(\nu - \nu_0)^2}{\Delta\nu_0^2} + \dfrac{P}{P_{\text{sat}}}} \qquad [1]$$

where g_0 is the peak value of the gain coefficient determined by the pumping level of the amplifier, ν the optical frequency, ν_0 the atomic transition frequency, $\Delta\nu_0$ the 3 dB local gain bandwidth, P the optical power of the signal, and P_{sat} the saturation power, which depends on the gain medium parameters. It must be emphasized that $\Delta\nu_0$ and P_{sat} refer to the local gain. However, from the communication system point of view, it is more desirable to use the related concepts of amplifier bandwidth and amplifier saturation power that will be evaluated below.

The amplifier gain G is defined as

$$G = \frac{P_{\text{out}}}{P_{\text{in}}} \qquad [2]$$

where P_{in} is the input power and P_{out} the output power of a continuous wave (CW) signal being amplified. The amplifier gain G may be found by using the relation:

$$\frac{dP}{dz} = g(\nu, P)P \qquad [3]$$

where $P(z)$ is the optical power at a distance z from the amplifier input end.

If the signal power obeys the condition $P \ll P_{\text{sat}}$ throughout the amplifier, the gain coefficient given by eqn [1] can be considered independent of the signal power. In such a case, the amplifier is said to be operated in the unsaturated regime and works as a linear device. The gain coefficient presents in this situation a Lorentzian profile that is characteristic of homogeneously broadened two-level systems. However, the gain spectrum of actual amplifiers can deviate significantly from the Lorentzian profile.

The solution of eqn [3] in the unsaturated regime is an exponentially growing signal power, given by

$$P(z) = P_{\text{in}} \exp(gz) \qquad [4]$$

For an amplifier length L, we then find that the linear amplifier gain is

$$G(\nu) = \exp(gL) = \exp\left[\frac{g_0 L}{1 + (\nu - \nu_0)^2/\Delta\nu_0^2}\right] \qquad [5]$$

Both the amplifier gain $G(\nu)$ and the gain coefficient $g(\nu)$ are maximum when $\nu = \nu_0$. However, $G(\nu)$ decreases much faster than $g(\nu)$ with the signal detuning $\nu - \nu_0$, because of the exponential dependence of G on g. As a consequence, the amplifier bandwidth $\Delta\nu_A$, which is defined as the FWHM of $G(\nu)$, is much smaller than the gain bandwidth $\Delta\nu_0$ (Figure 1). This can result in signal distortion in the case where a broadband optical signal is transmitted through the amplifier. From eqn [5] we can obtain the following relation between the amplifier bandwidth and the gain bandwidth:

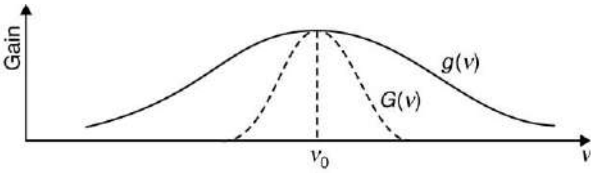

Figure 1 Gain coefficient profile $g(\nu)$ and the corresponding amplifier gain spectrum $G(\nu)$.

$$\Delta\nu_A = \Delta\nu_0 \sqrt{\frac{\ln 2}{g_0 L - \ln 2}} \qquad [6]$$

Gain Saturation

An important limitation of the nonideal amplifier is related with the power dependence of the gain coefficient given by eqn [1]. This property is known as gain saturation and it appears when the signal power ratio P/P_{sat} is non-negligible. Since the gain coefficient is reduced when the signal power P becomes comparable to the saturation power P_{sat}, the amplifier gain G will also decrease.

Assuming that $\nu = \nu_0$ and substituting g from eqn [1] in eqn [3] gives

$$\frac{dP}{dz} = \frac{g_0 P}{1 + P/P_{\text{sat}}} \qquad [7]$$

Considering the initial condition $P(0) = P_{\text{in}}$, we obtain from eqns [2] and [7] the following implicit relation for the amplifier gain:

$$(1 - G)\frac{P_{\text{in}}}{P_{\text{sat}}} = \ln\left(\frac{G}{G_0}\right) \qquad [8]$$

where $G_0 = \exp(g_0 L)$. The input saturation power P_{in}^{s} is defined as the input power for which the amplifier gain G is reduced by a factor of 2 from its unsaturated value G_0 (Figure 2). Indeed, it is obtained by using $G = G_0/2$ in eqn [8]:

$$P_{\text{in}}^{\text{s}} = \frac{2\ln(2)P_{\text{sat}}}{(G_0 - 2)} \qquad [9]$$

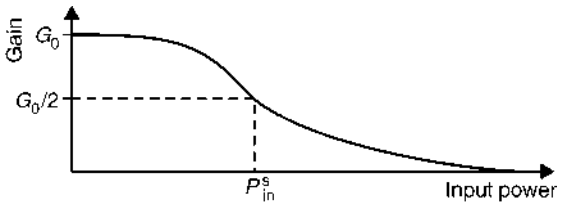

Figure 2 Saturated amplifier gain as a function of the input power.

As observed from eqn [9], the input saturation power P_{in}^s does not coincide with P_{sat}. The output saturation power is given by $P_{out}^s = G_0 P_{in}^s/2$. In practice, $G_0 \gg 2$ and P_{out}^s is found to be smaller than P_{sat}, by about 30%.

Gain saturation can be seen as a serious limitation, particularly for multichannel communication systems. However, the self-regulating effect of gain saturated amplifiers can be useful in long-haul lightwave communication systems, including many concatenated amplifiers. In fact, if the signal level in a chain of amplifiers is unexpectedly increased along the chain, the saturation effect causes a lower gain provided by the following amplifiers and vice versa for a sudden signal power decrease.

Amplifier Noise

Besides the bandwidth and gain saturation limitations, another property must be considered concerning practical optical amplifiers. In fact, optical amplifiers always add spontaneously emitted photons to the signal during the amplification process. Those photons are amplified besides the signal photons so that, at the amplifier output, an amplified spontaneous emission (ASE) noise is presented. Since spontaneous emission always takes place, ASE noise is unavoidable and does not depend on the amplifier temperature. This is one of the most important differences between optical and electrical amplifiers, where amplifier noise is of thermal origin and can be reduced by lowering the amplifier temperature.

The ASE determines a degradation of the signal-to-noise ratio (SNR). The SNR degradation is usually characterized by the amplifier noise figure, which is defined as the SNR ratio between input and output:

$$\mathrm{NF} = \frac{\mathrm{SNR_{in}}}{\mathrm{SNR_{out}}} \qquad [10]$$

The SNR is usually referred to the electrical power generated when the optical signal is converted to electrical current by using a photodetector. Therefore, the noise figure as defined in eqn [10] would usually depend on several detector parameters, which determine the shot noise and thermal noise associated with the practical detector. We will consider the case of an ideal detector, whose performance is limited by shot noise only.

The SNR of the input signal is simply determined by the detection shot noise and can be written as:

$$\mathrm{SNR_{in}} = \frac{P_{in}}{2h\nu\Delta f} \qquad [11]$$

where Δf is the detector bandwidth.

To evaluate the term $\mathrm{SNR_{out}}$, one should add the contribution of spontaneous emission to the receiver noise. The ASE noise spectral density is assumed to be constant and can be written as

$$S_{sp}(\nu) = (G-1)n_{sp}h\nu \qquad [12]$$

where G the amplifier gain and

$$n_{sp} = \frac{N_1}{N_1 - N_0} \qquad [13]$$

is known as the spontaneous emission factor or the population inversion factor. In eqn [13] N_0 and N_1 are the atomic populations for the ground and excited states, respectively.

Considering a low noise amplifier, the signal power impinging the photodetector is larger than the optical noise power and the shot noise power. As a consequence, the electrical noise, due to the signal-ASE beating, is the dominant contribution and the SNR of the amplified signal is given by:

$$\mathrm{SNR_{out}} \approx \frac{GP_{in}}{4S_{sp}\Delta f} \qquad [14]$$

Using eqns [11]–[14], the amplifier noise figure defined by eqn [10] becomes:

$$\mathrm{NF} = 2n_{sp}\frac{G-1}{G} \approx 2n_{sp} \qquad [15]$$

where the approximation holds when the gain is much higher than one. In the case of an ideal amplifier, $n_{sp} = 1$ and eqn [15] show that the SNR is degraded by 3 dB. For most practical amplifiers, NF can be as large as 6–8 dB.

Basic Amplifier Configurations

The relative importance of the different limiting factors discussed above depends on the actual

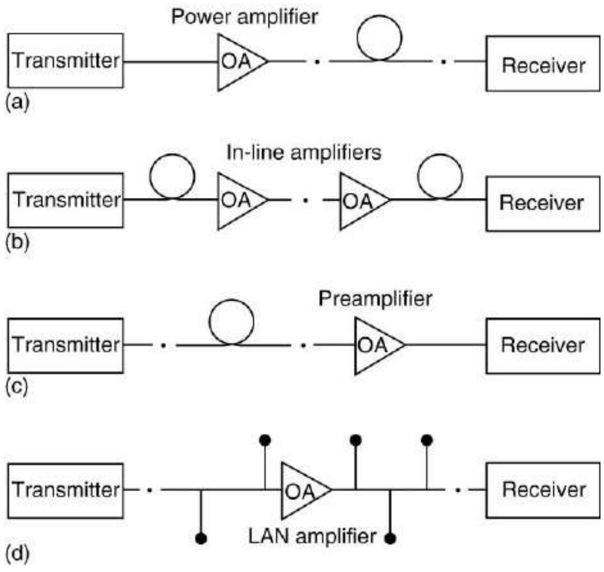

Figure 3 Four generic configurations for incorporating optical amplifiers into transmission systems: (a) as a power amplifier; (b) as in-line amplifiers; (c) as a preamplifier; (d) for compensation of distribution losses in local-area networks.

amplifier application. Figure 3 shows the four basic system configurations envisioned for the incorporation of optical amplifiers. The first configuration is to place the amplifier immediately following the laser transmitter to act as a power amplifier or booster (Figure 3a). The main purpose of such amplifiers is to boost the signal power, which can provide an increase of the transmission distance by 100 km or more. Since the signal input power is typically large (0.1–1.0 mW), the key parameter for the power amplifier will be to maximize the saturation output power and not necessarily the absolute gain.

The second configuration is to place the amplifier in-line and perhaps incorporated at one or more places along the transmission path (Figure 3b), replacing the electronic regenerators. The in-line amplifier corrects for periodic signal attenuation and may exist in a cascade form. The use of in-line optical amplifiers is particularly attractive for multichannel communication systems, since they can amplify all channels simultaneously.

The third configuration consists of using the amplifier immediately before the receiver, so it functions as a preamplifier (Figure 3c). The purpose of such an amplifier is to improve the receiver sensitivity. The main figures of merit are high gain and low amplifier noise, because the entire amplifier output is immediately detected.

In local-area networks (LANs), distribution losses often limit the number of possible nodes. The fourth application of optical amplifiers consists of using them for compensating such distribution losses (Figure 3d).

List of Units and Nomenclature

g_0	peak value of the gain coefficient at peak	m^{-1}
$g(\nu)$	gain coefficient	m^{-1}
G_0	unsaturated amplifier gain	
$G(\nu)$	amplifier gain	
L	amplifier length	m
n_{sp}	spontaneous emission factor	
P	signal power	W
P_{in}	input signal power	W
P_{in}^{s}	input saturation power	W
P_{out}	output signal power	W
P_{out}^{s}	output saturation power	W
P_{sat}	saturation power	W
$S_{sp}(\nu)$	ASE noise spectral density	J
Δf	detector bandwidth	Hz
$\Delta\nu_0$	bandwidth of the gain coefficient	Hz
ν	optical frequency	Hz
ν_0	atomic transition frequency	Hz

See also

Optical Amplifiers: Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*. New York: Wiley.
Bjarklev A (1993) *Optical Fiber Amplifiers: Design and System Applications*. Boston, MA: Artech House.
Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.
Green PE Jr (1993) *Fiber Optic Networks*. Englewood Cliffs, NJ: Prentice Hall.
Iannone E, Matera F, Mecozzi A and Settembre M (1998) *Nonlinear Optical Communication Networks*. New York: Wiley.
Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Boston, MA: Artech House.
Shimada S and Ishio H (eds) (1994) *Optical Amplifiers and their Applications*. New York: Wiley.

Erbrium Doped Fiber Amplifiers for Lightwave Systems

P Bollond, JDS Uniphase Corporation, Ewing, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The telecommunications industry has undergone a revolution since the 1980s, by using glass optical fibers for the transmission of information encoded as pulses of light. A single telecommunications-grade optical fiber has been shown to support the propagation of more than 1 Tbit per second (1×10^{12} pulses per second) over distances comparable to typical city separations (>100 km). From this technology, optical fiber links have evolved to become a network on a planet-wide scale, and form the physical backbone of the information age.

Erbium-doped fiber amplifiers (EDFAs) are an enabling technology for optical fiber communication networks. They have several important properties that make them the amplification component of choice in long-distance commercial data transport networks. Erbium ions deposited in silica-based glass allow amplification in the lowest loss region of commercial-grade optical fiber (<0.25 dB/km from approximately 1530–1620 nm). Erbium-doped fiber (EDF) is manufactured in a form that allows for low-loss fusion splicing to standard communications fiber. Compact semiconductor laser diodes are available to excite the erbium ions into an amplification state. The long lifetime of the excited state of erbium provides the ability to simultaneously amplify multiple wavelength channels without significant cross-channel interference. Multiple channel systems have been deployed with more than 100 optical channels (or wavelengths) through each EDFA, and this has allowed network capacities to be dramatically increased. The low noise properties of the EDFA also allow networks to be constructed with many amplified spans before the optical signal has to be electronically regenerated. Practically unlimited transmission distance has been demonstrated using a small number of optical soliton channels through periodically amplified EDFA lightwave systems.

The development history of the EDFA can be traced back to the first optical amplifier. In 1962, a neodymium-based fiber amplifier was invented that operated at 1064 nm. During the 1980s, the need for an optical amplifier at telecommunications wavelengths initiated research at many locations throughout the world. In 1987, the University of Southampton (UK) was first to demonstrate an EDFA that had optical gain at 1550 nm, and during the following years the design of erbium-doped fiber was optimized for this application. In 1989, a practical semiconductor laser diode became available to pump EDF, and the first compact optical fiber amplifier modules soon appeared for commercial deployment. The traditional method of signal regeneration, prior to 1990, was to use electronics to detect the optical signal after each transmission span, recover the digital signal, and then retransmit using another laser diode. The EDFA allows practical wavelength division multiplexing (WDM) of multiple optical signals with all optical signal amplification, and provides significant performance and cost advantages over electronic regeneration.

EDFAs have emerged from the laboratory to be widely deployed in communication networks. EDFAs are used to boost transmitted power of the signal lasers (booster amplifier), amplify signals in transit to compensate losses sustained in the fiber (line amplifier), or amplify signals before a receiver (pre-amplifier). Typically, the amplifier module is specifically manufactured for particular systems that are mounted on electronic circuit boards. These circuit packs are then housed in central offices (local telephone exchanges), remote ‘repeater huts’, or even in undersea ‘bottles’ as part of a transoceanic cable. The high cost of network failure requires that the manufactured EDFA modules comply with stringent reliability criteria, to provide a useful operating lifetime greater than 25 years when subject to extreme environmental conditions.

Amplifiers are constructed for incorporation into either existing fiber links as part of an upgrade, or for newly planned systems. Because of the high cost of installing new fiber into the ground and securing property rights, it has become economically desirable to upgrade many existing fiber links rather than to build new systems. Transmission cables usually have many pairs of individual optical fibers, some of which will not initially be transporting data, and these ‘un-lit’ or ‘dark fibers’ can be activated as consumer demand increases over time. Typically, each fiber of a pair is used to carry either ‘east’- or ‘west’-bound traffic. Around city areas metropolitan area networks can be expanded in this way, but for long-distance links (long haul networks with distances >1000 km) operation is designed for a larger number of channels (40 to 120 wavelengths) at higher data rates (10 or 40 GHz per channel) over specialized transmission cables containing low numbers of fiber pairs. Typically communication systems are designed to meet certain cost targets, expressed as dollars per Gbit/s per km, for total transmission distances. The total transmission distance is limited by the optical signal to noise ratio (OSNR) degradation after each fiber-amplifier link, with a smaller permissible degradation at higher bit rates. The required gain and OSNR performance for the system is then translated to an EDFA module specification.

This article discusses EDFA design and applications, and shows the elements involved in producing reliable modules for commercial lightwave systems.

Components for EDFAs

EDFAs are comprised of passive optical components, erbium-doped fiber, and pump lasers. Passive components are chosen to meet optical and environmental specifications, while the erbium-doped fiber is selected based on the optical power, gain, and noise figure requirements. Pump lasers are a key influence on the price and performance of optical amplifiers.

The amplifier module is typically connected to the transmission fiber using fiber connectors. These polished fiber connectors have a higher insertion loss and reflectivity than fusion splicing, but allow for easy deployment in the network. Internal optical components are fusion spliced together to provide low loss, low back reflection, high strength, and high reliability joins. Fusion splicing is tailored to particular fiber types, so that optical components with dissimilar fiber types are joined with the lowest loss.

The signal and pump radiation is combined with low loss using optical components called wavelength division multiplexers (see Figure 1). These components are based on fused fiber or interference filter based technology. Fused fiber WDMs offer the lowest insertion loss (<0.1 dB is commercially available) but is restricted to widely spaced wavelengths (e.g., 980 nm pump and 1550 nm signal). Interference filter based WDMs are available for closely spaced wavelengths (e.g., 1480 nm pump and 1530 nm signal) and have a very low wavelength dependent insertion loss (flatness). Interference filters can be designed to produce sharp low-pass, high-pass, or bandpass type filters suitable for combining closely spaced wavelengths, as well as broader peaks suitable for lowering the gain in particular signal wavelength regions to produce gain flattened amplifiers.

Signal reflections can cause an amplifier to act as a laser, and this detrimental effect is eliminated in EDFAs by using optical isolators. In an isolator the signal light is coupled out of the single-mode fiber through a graded index (GRIN) lens and passes through a nonlinear crystal before being coupled back into the optical fiber. The isolator consists of a birefringent rutile (TiO_2) or Yttrium OrthoVanadate (YVO_4) wedge, followed by a Yttrium Iron Garnet (YIG) Faraday rotator, followed by another birefringent wedge. The YIG crystal is surrounded by a permanent magnet that rotates the light's polarization by 45 degrees. The 45-degree polarization rotation, coupled with the two birefringent wedges, ensures that light is efficiently coupled to the output fiber but not in the reverse direction. Commercial isolators are available with low insertion losses across the signal band, with some samples below 0.35 dB. Note that the YIG crystal works well for the 1480 nm pump and 1530–1620 nm signal bands, but currently there is no suitable isolator material that covers 980 to 1550 nm and this puts some limitations on certain EDFA designs.

The isolator design has been extended to make multiport circulators. A three-port circulator has the input into port 1 and output of port 2, light entering port 2 is directed to port 3, and light entering port 3 is blocked with an isolator. The circulator allows for adding and dropping of individual channels when a

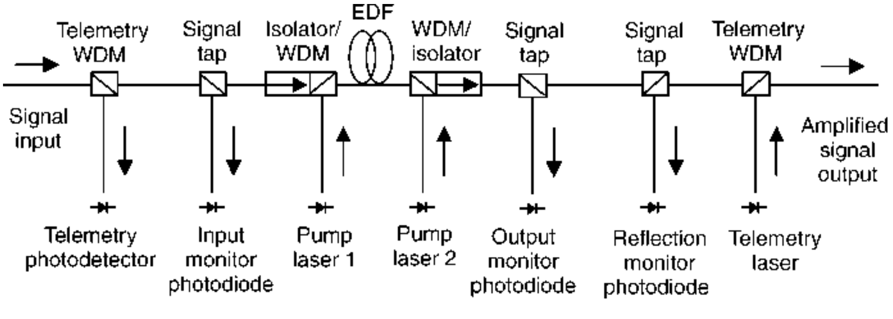

Figure 1 Single-stage EDFA with features.

narrow bandwidth reflective filter is placed on port 2. Circulators also can be used to separate co-propagating and counter-propagating traffic on a single transmission fiber before amplification is done.

Erbium-doped fiber has two strong absorption bands around 980 nm and 1480 nm that are suitable for achieving a population inversion in the erbium ion. The 980 nm wavelength allows low noise amplification, while 1480 nm lasers provide higher EDFA output. 980 nm pump lasers usually incorporate fiber Bragg gratings to stabilize the laser wavelength to match the narrow EDF absorption peak, and also have lower drive current requirements. A 1480 nm pump can provide more amplification, since there are more photons in each mW of laser power at 1480 nm than 980 nm. High-output power pump lasers incorporate thermo-electric coolers (TECs) within the pump laser package and can provide fiber coupled power greater than 500 mW. Both high-power lasers have been qualified to meet the most stringent reliability standards, with typical mean time before failure beyond 25 years. Low-cost 980 nm pump lasers, that do not have TECs, are also available in smaller packages and can supply up to ~200 mW. By using wavelength division or polarization combining techniques, it is possible to further increase the available pump power in the erbium fiber.

The EDFA module is assembled into a package that may contain passive optical components, several meters of coiled EDF, pump lasers, photo-detectors, and electronic circuit boards. In some cases, it is advantageous to thermally stabilize the components so that the amplifier performance can be maintained when exposed to extreme environmental conditions. This may occur when the central office's environmental control is compromised (e.g., air conditioning failure). In particular, EDF can exhibit undesirable spectral gain changes if the ambient temperature changes by more than 5 °C.

Pump lasers with internal TECs can dissipate more than 5 Watts of heat per laser and since cooling fans usually do not have the required reliability, passive cooling is commonly used in the module in the form of a metal heat sink. The amplifier module's size can be compact, limited by the height of optical components or pump laser diodes, or by the size of a built-in heat sink. The module is designed to survive conditions of electrostatic discharge, humidity, temperature, thermal shock, extreme vibration, and other stresses that may be inadvertently present during operation in the field. In addition, all material in the EDFA module is also qualified against problems with out-gassing (e.g., hydrogen release), combustion and chemical or biological exposure.

The Single-Stage Amplifier

A simple single-stage EDFA consists of an erbium-doped fiber spool with signal and pump combining multiplexer. The fiber is optically pumped by 980 nm and/or 1480 nm laser(s), whose light is coupled into the signal fiber by a passive component called an optical multiplexer. The pump wavelengths are readily absorbed by erbium ions embedded in silica raising them to an excited state. Amplification occurs when, stimulated by a nearby signal photon, an excited erbium ion relaxes back to the ground state producing a second, identical signal photon. The erbium ion can be approximated as a three-level atomic system that can be completely inverted by a 980 nm photon, to provide the lowest noise amplification. In contrast, the 1480 nm pump will excite the erbium ion directly into the upper laser level as a two-level system, and because of rapid spontaneous emission from this level, the maximum inversion in this case cannot exceed approximately 75%. Note that as the pump photons are absorbed, the inversion will be nonuniform along the EDF length.

There are two signal wavelength regions commonly amplified by EDFAs, the C-band (conventional band) from approximately 1528 to 1565 nm, and the L-band (long band) from approximately 1570 to 1620 nm. Amplification in the C-band readily occurs when moderate pump power is available, and relies on the erbium ion's spectral absorption and emission wavelength window that is suited to high levels of atomic inversion. L-band amplification is also achieved with moderate pump powers, but because of the lower absorption and emission cross-sections, similar gain is reached using approximately five times more EDF with a lower average inversion. The C-band amplifier is typically less costly because less EDF is used, while high-concentration erbium fibers are available specifically for L-band EDFAs.

An EDFA's most critical performance parameters are its amplified signal output power (typically stated in dBm) and its noise figure (stated in dB). Output power is mainly determined by total pump power and the amplifier internal loss. The noise figure (NF) is defined as the ratio of the signal-to-noise ratio at the input to the signal-to-noise ratio at the output.

A single-stage amplifier typically has 1 or 2 pump lasers but can have more if polarization- or wavelength-pump-combining is implemented for higher power. When the pump radiation propagates in the same direction as the signal, the amplifier is co-pumped, while counter-pumping denotes the case when the pump laser propagates against the signal. For a single pump, a co-pumping 980 nm laser minimizes the NF (suitable for a pre-amplifier) while counter-pumped 1480 nm architecture optimizes output power at some expense to the NF (suitable for a booster amplifier). Recent semiconductor pump laser improvements have enabled 980 nm pump lasers with output powers >500 mW to be commercially available, allowing most single- or dual-stage EDFAs to be energized by one laser.

Single-stage designs can be enhanced, as shown in Figure 1. To control the network an optical signal may be used as a telemetry or network supervisory channel, and this is removed with a filter. Telemetry wavelengths are usually outside of the useful EDF amplification window, and commonly range from 1500–1520 nm and from 1620–1640 nm. An isolator may be used at the input and/or output to prevent pump laser or amplified spontaneous emission from the erbium-doped fiber 'leaking' into the transmission path. Optical taps may be included to provide information about signal spectra at the input and output sides. Their feedback can be used to control pump laser biases for tuning output power, monitoring amplifier performance, or simply to trigger alarms. In addition, a reflection monitor is sometimes placed at the output to observe backwards propagating optical signals arising from reflections. Electronics in the module will continuously monitor the pump lasers and photodetectors and, for example, can place the module in an 'eye-safe' low-output power (<10 mW) mode within milliseconds if a transmission fiber break is detected through increased back-reflected optical power.

The amplifier has a signal input power of −30 to −10 dBm (1–100 μW) for each optical channel, and has a gain of 25 dB to compensate for a typical span loss of 100 km optical fiber link. This signal level allows high powers at the receiver photodetector for high-quality signal detection, yet is low enough to avoid nonlinear propagation effects in the transmission fiber. The typical total output signal power of an 80 channel (wavelength) C-band EDFA is less than 200 mW (+23 dBm) to avoid stimulated Raman scattering (SRS) in standard transmission fiber. The use of improved low nonlinearity transmission fibers and longer transmission distances can lead to specified total EDFA output powers to be greater than 400 mW (+26 dBm). To provide the best performance, the amplifier usually will allow only a single direction of propagation, with bi-directional communications systems using one transmission fiber and circulators to separate 'east' and 'west' traffic into individual EDFAs.

Figure 2 shows the results of a numerical simulation for a single-stage amplifier as a function of the input power and EDF length. The amplifier was assumed to have EDF with peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 100 mW at 980 nm. The amplifiers signal loss before and after the EDF was taken to be 0.8 and 1.2 dB, respectively. As amplifier input power increases there is a decrease in gain provided by the medium since there is a fixed amount of pump power available. The maximum gain is usually achieved near 1530 nm where the difference between the EDF's emission and absorption cross-sections is greatest. Selecting the length of EDF is critical to achieving the desired performance, and this can be examined using numerical simulation for a wide range of design options.

Undersea systems, with their long distances and large costs of network failure, place stringent design requirements on EDFAs. These systems mostly use single-stage designs with emphasis on low noise operation. This can be achieved by eliminating most of the optical components prior to the EDF, and also by co-propagating a strong 980 nm pump using a low loss fused fiber WDM. Undersea repeaters are spaced by 30 to 80 km, shorter than terrestrial networks, with each channel operated at higher power to achieve multi-thousand kilometer distances. For example, the trans-Pacific TPC-5J cable spans 8600 km from Coos Bay, Oregon (USA) to Ninomiya (Japan), using EDFAs spaced every 33 km. The tight electrical power budgets available to each repeater (powered from land) necessitate using pump lasers without TECs. During installation, the EDFA will experience large mechanical shock, as the cable and metal repeater bottles are unwound aboard ship and dropped into the ocean. Other design considerations are done for operation at the constant ambient temperature of the ocean bottom (~+2 to 4 °C) or for seasonal temperature changes on the continental shelves, where the optics and EDF will operate at 15 to 30 degrees above ambient due to heating from the electronics.

Single-stage optical amplifiers are suited for a wide range of applications such as single-channel amplifiers, simple WDM amplifiers, and low-cost amplifiers. Using high-power pump lasers or combined pump laser schemes allows such amplifiers to deliver output powers >20 dBm. However, single-stage amplifiers cannot meet the requirements of all telecommunications architectures, leading to increasing demand for multiple-stage EDFAs that are discussed below.

Multiple-Stage EDFAs

The most common implementation of a multiple-stage EDFA is to improve the noise and output power

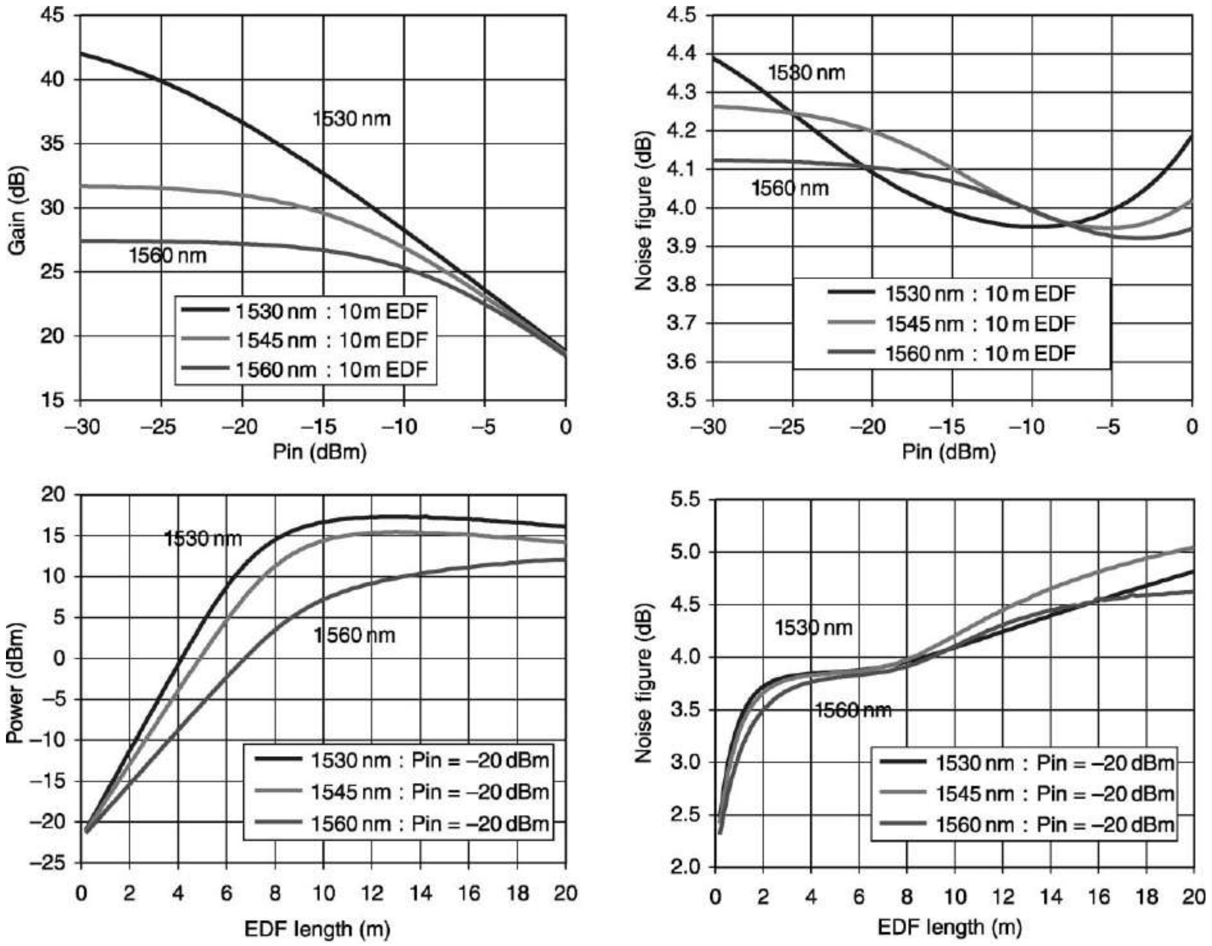

Figure 2 Typical single-channel performance of a single-stage EDFA.

characteristics by imbedding an optical isolator between two sections of EDF. The isolator blocks backward-traveling amplified spontaneous emission to improve the efficiency of the amplifier. This is shown in Figure 3 for two common single-pump designs. The first is the 'pump by-pass', design where the residual pump radiation from the first stage is redirected around the mid-stage isolator. This is often used when pumping with 980 nm since signal band optical isolators will not transmit 980 nm radiation. This design can introduce problems when using low-quality components as small signal levels can propagate around the pump by-pass fiber to create multiple path interference (MPI) effects. Although MPI effects can be eliminated by using a pump splitting coupler to directly pump each EDF section, the pump by-pass design makes the most efficient use of available pump power. The second design is called the 'pump through' design, and is used when pumping at 1480 nm since both the pump and signal band (e.g., 1550 nm) photons will transmit through commercial isolators with only small loss. Note that for cost and space constraints within the module, it is sometimes advantageous to use hybrid optical components, e.g., an isolator and WDM can be combined into one compact package.

An EDFA with multiple input signals will have a very nonuniform output gain profile. This is a consequence of the erbium ion's wavelength-dependent emission and absorption cross-sections in the host glass material. Figure 4 shows gain spectra for both C-band and L-band EDFAs with and without gain-flattening filters (GFFs). The gain spectrum is dependent on the EDF's chemical composition and EDFA design features. For an amplifier with 25 dB gain, a GFF with peak loss less than 10 dB is usually needed to correct for these gain deviations. Although many technologies are available for GFFs (e.g., thin-filter interference filters, Bragg gratings, tapered fiber filters, etc.) the basis function of these technologies is usually not identical to the EDF's gain profile, and this mismatch results in gain

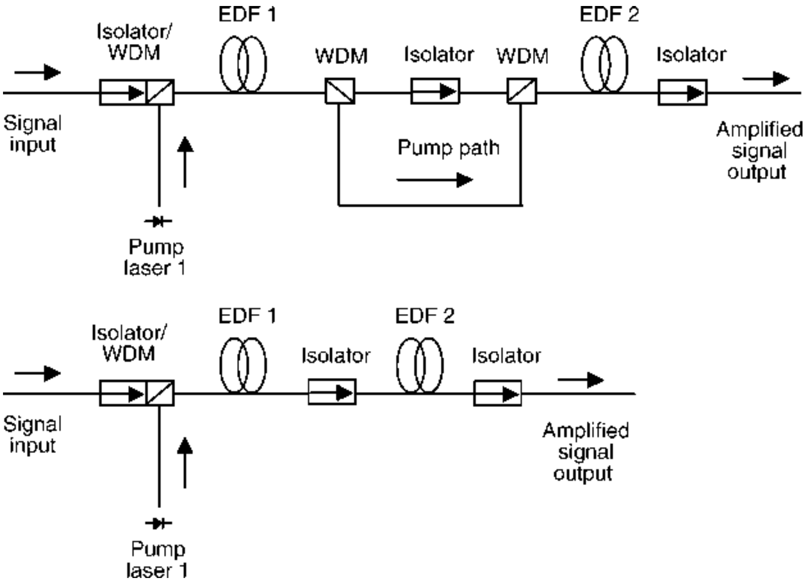
EDF 1
Isolator/ WDM
WDM
Isolator
WDM
EDF 2
Isolator
Signal input
Pump path
Amplified signal output
Pump laser 1
EDF 1
Isolator/ WDM
Isolator
EDF 2
Isolator
Signal input
Amplified signal output
Pump laser 1

Figure 3 Imbedded Isolator EDFA designs.

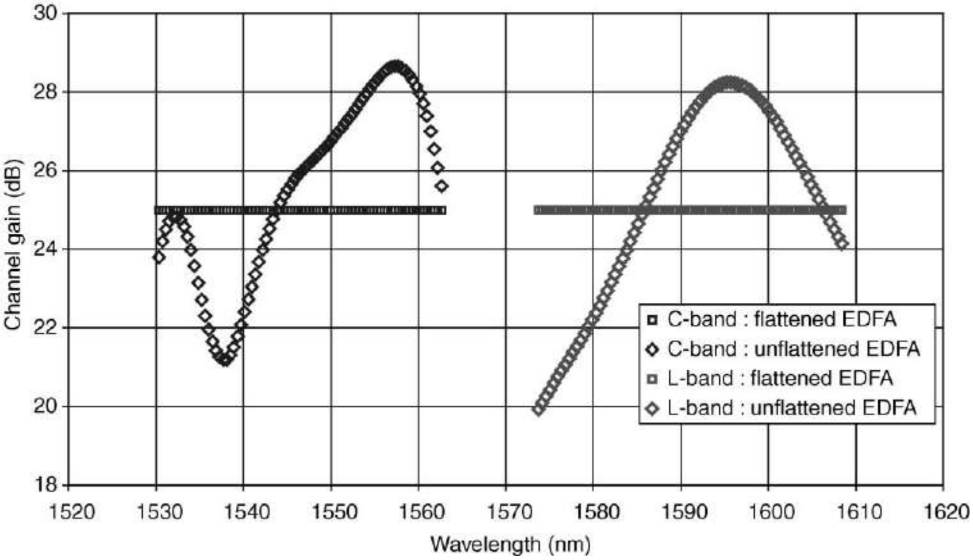

Figure 4 Gain spectrum of a multi-staged optical amplifier with and without gain flattening filters.

flatness error. However, when using these technologies, manufactured EDFAs can have a gain error (flatness) less than 1.0 dB over bandwidths greater than 35 nm.

Early EDFAs for WDM systems used 16 optical channels spaced by 200 GHz in the 1540 to 1560 nm part of the C-band. In this case, the EDF gain profile is relatively smooth and no GFFs are needed to achieve a 1.0 dB gain flatness specification. It is important to note that for a particular level of signal and pump power, a longer length of EDF in the EDFA will produce more gain at the long wavelength end of the spectrum, hence by shortening the EDF length in Figure 4 the result will be relatively flat gain from approximately 1540 to 1560 nm. As EDFA technology has matured, the EDF can be the bandwidth-limiting component in a system, and it has become necessary to use GFFs to realize useable bandwidths of up to 55 nm when using regular silica-based EDF. The most economic EDFAs operate in the C-band where relative to the L-band, pump laser efficiency is highest, EDF lengths are shortest, and the chromatic dispersion of most installed transmission fiber limits nonlinear optical effects.

In single-stage amplifiers, the gain-flattening filter is at the amplifier output, and this results in lost output power. Wideband gain-flattened amplifiers usually have a multiple EDF stage design, with the GFF component between two EDF stages. The EDF sections that follow the attenuating GFF provide additional amplification, to give high power out of the EDFA module that has large internal losses.

In addition to GFF components inside the EDFA, an attenuator can be used to compensate for input signal power changes, which will produce a spectral gain rotation or tilt. The attenuator reduces the power on all optical channels equally and allows the amplifier to operate in a ‘fixed gain’ mode. In WDM systems, optical amplifiers may also need mid-stage access where the signal is fed into an external device between the amplifier stages. Reasons for doing this include monitoring, adding or dropping of individual channels, and dispersion compensation. Since additional losses up to 10 dB are introduced in the amplification path, the amplifier design has to be optimized for those losses.

Typical two-stage gain-flattened optical amplifier architecture is shown in Figure 5. Input and output couplers and telemetry WDMs can be included if required. A single pump can be split and shared between stages to save cost. When multiple pumps are needed, the most common configuration is a 980 nm pump laser co-pumping the first stage (EDF1) for low noise and a 1480 nm laser counter-pumping the second spool (EDF2) for high gain. A significant loss element, e.g., a gain-flattening filter, add/drop module or dispersion compensation module, is situated between the stages. Note that mid-stage access can be located before or after the gain flattening or between additional EDF stages. The gain spectrum of multistage gain-flattened EDFAs is shown in Figure 6, showing the gain equalization possible using a multistage amplifier design and gain-flattening filters. Using deeper GFFs can increase the usable bandwidth of the amplifiers, but this requires higher-power pump lasers to maintain the same gain and optical signal-to-noise ratio (OSNR) performance.

The two-stage EDFA shown in Figure 5 highlights a common problem for amplifier design. Given high-grade optical components and pump lasers, what

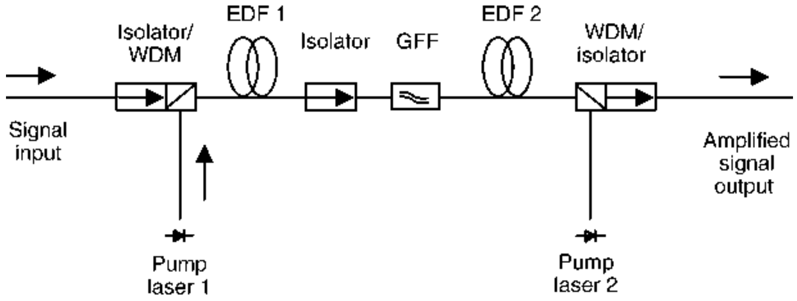

Figure 5 Typical multistage gain-flattened optical amplifier architecture.

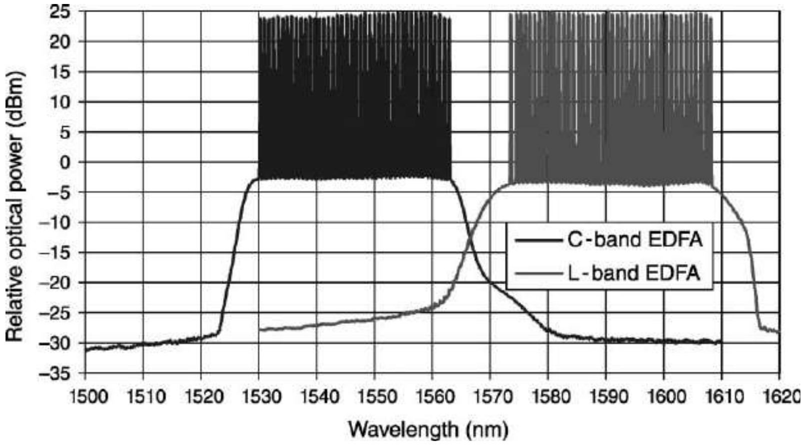

Figure 6 Measured gain and noise spectra of multistaged gain-flattened C-band and L-band EDFAs.

length of erbium-doped fiber should be used to give the best amplifier performance? This question is best answered by using the results of extensive numerical simulations of the optical amplifier.

Figure 7 shows the results of numerical simulation for a two-stage amplifier as a function of the first and second EDF stage lengths. The amplifier was assumed to have EDF with a peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 130 mW at 980 nm in the first stage and 160 mW at 1480 nm in the second stage. The total input power was assumed to be −2.5 dBm for 80 channels distributed from 1530 nm to 1563 nm. Note that each of the 80 channels is separated by 100 GHz, and will generally support a long-distance communications system with each wavelength modulated at 10 GHz.

Figure 7a shows that the optical output power is largest when both EDF stages are approximately 20 meters. For this design, EDF lengths significantly longer than 20 meters will produce less output power, since the pump radiation will have been completely absorbed in the first few meters of EDF. The spectral gain flatness is shown in Figure 7b. For this design, approximately 17 meters of EDF was needed to produce a flat gain spectrum, i.e., each optical channel had the same gain. Other combinations of EDF lengths were not matched to the particular gain-flattening filter and produced large gain variations across the band. In general as the amplifier's total EDF length increases the gain spectrum exhibits a positive tilt (e.g., longer wavelength channels will experience more gain). Figures 7c, d show the average channel and maximum channel NF, respectively. From a system design perspective, a link can be limited by the optical channel with the lowest optical signal-to-noise ratio (OSNR) or highest noise figure, and this is a critical parameter of interest when designing optical amplifiers. In practice the best amplifier design is a compromise between high-gain, low-gain flatness, and low NF. From extensive numerical simulations, for the two-stage gain flattened amplifier example it was for

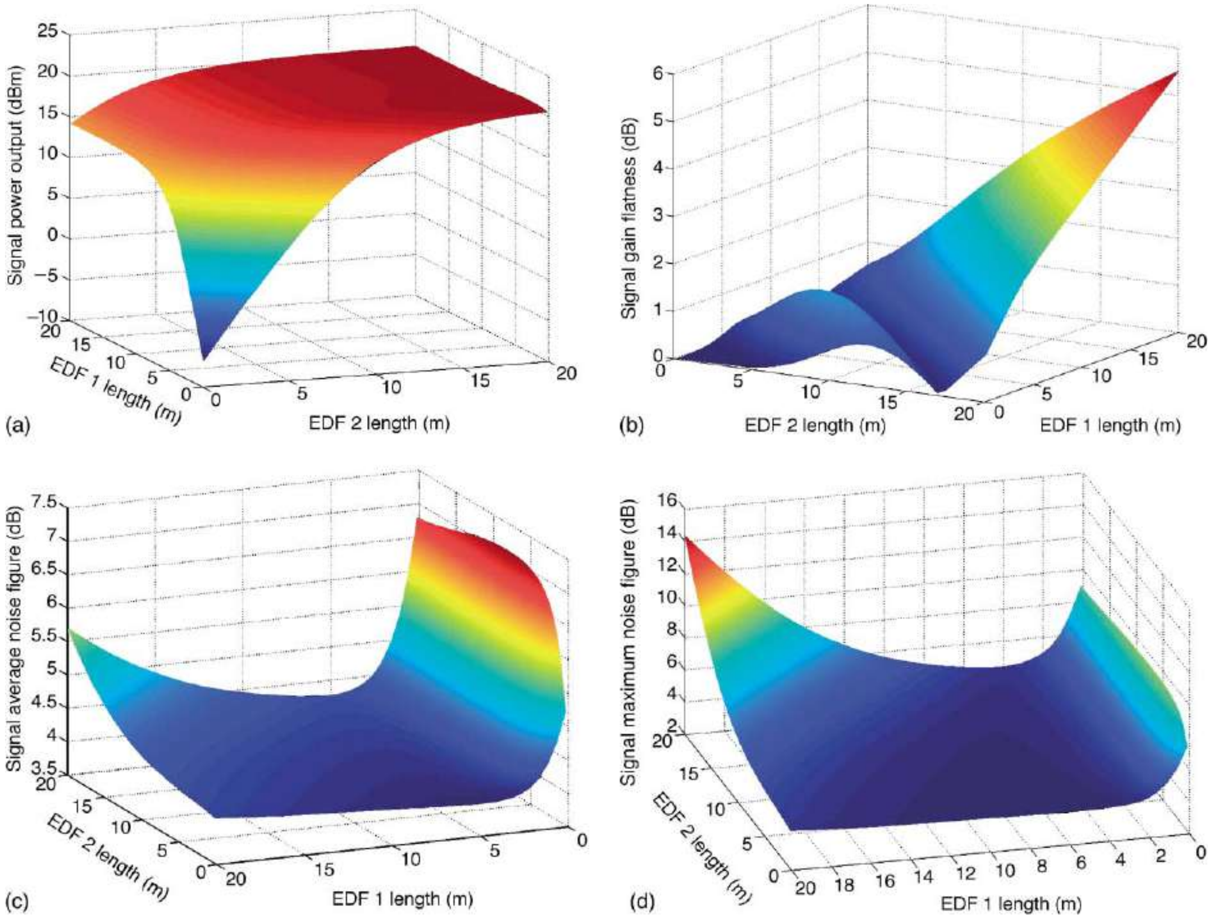

Figure 7 (a) Total output power, (b) spectral flatness, (c) average noise figure, and (d) maximum channel noise figure of a two-stage optical amplifier with a gain-flattening filter.

approximately 4 meters in EDF stage 1, and 13 meters in stage 2.

Successful commercial EDFA products are more than a single amplifier design that can accommodate all applications. In some cases an economical solution is a modular design approach, where smaller capacity, lower-cost amplifiers are initially installed in a network. Additional gain can be added to the basic amplifier for longer-distance spans or as network traffic increases over time. This ‘pay as you grow’ approach can be done using additional gain stages, or pump lasers, and can be upgraded without interruption of revenue-generating network traffic.

Low-Noise Design of EDFAs

A key design feature of commercial EDFAs is low-noise operation since the degradation of the OSNR limits the reach of the system. The NF of an EDFA can be defined as

$$\mathrm{NF} = \mathrm{SNR_{in}}/\mathrm{SNR_{out}} \qquad [1]$$

The NF can also be calculated from

$$\mathrm{NF} = \frac{1}{G} \times \left[\frac{P_{\mathrm{ASE}}(\nu_\mathrm{s}, L)}{h\nu_\mathrm{s}} + 1 \right] \qquad [2]$$

where P_{ASE} is the amplified spontaneous emission (ASE) noise power at the signal frequency ν_s for an EDFA of gain G (h = Plancks constant). An alternative equation for the NF for a single section of EDF is

$$\mathrm{NF} = \int_0^L \frac{\gamma_\mathrm{e}(\nu_\mathrm{s}, z) - \gamma_\mathrm{a}(\nu_\mathrm{s}, z)}{P_\mathrm{s}(z)} \times P_\mathrm{s}(0)\mathrm{d}z + 1 \qquad [3]$$

where γ_e and γ_a are the EDF emission and absorption factors at the signal frequency ν_s, at distance z along the amplifying fiber. Equation [3] shows that low NF can result from rapid signal gain in the initial fiber along the EDF. This result indicates that co-propagating pump and signal photons, both into the same end of the EDF, combined with the maximum inversion obtained by using a 980 nm pump, will result in low-noise amplification. The NF is typically expressed in dB units, and using a full quantum mechanical theory, the lowest possible NF for a fully inverted EDF with a large signal gain can be shown to be approximately 3 dB. Actual EDFAs will only approach the ‘3 dB quantum noise limit’ when the signal is significantly lower than the available pump power (e.g., low channel count applications) so that high inversion can be maintained along the EDF length. Optical components in the EDFA prior to the first EDF will attenuate the signal but not the noise, since ASE is only generated in the EDF, and this input stage loss will directly add to the EDFA’s NF.

The use of 1480 nm pump lasers can also result in a further NF penalty of 0.3 to 1.5 dB since complete inversion cannot be achieved when the erbium ions are operating as a two-level atomic system and more ASE is generated. The ASE, like the signal gain, has a wavelength-dependent spectral profile and this results in each channel having a different NF. Typically, the optical channel with the largest NF (lowest OSNR) is used to define the EDFA’s noise performance. EDFAs with low-input signal power (e.g., -15 dBm) and high gain (e.g., 25 dB) may be described as low noise when the NF is <4 dB. For large total input signal power (e.g., $+2.5$ dBm for high channel count operation) low noise may be $\mathrm{NF} < 5$ dB.

For an amplifier with multiple EDF sections, the noise figure can be calculated from

$$\mathrm{NF_{System}} = \frac{\mathrm{NF}_1}{L_1} + \frac{\mathrm{NF}_2}{L_1 G_1 L_2} + \ldots + \frac{\mathrm{NF}_N}{L_1 G_1 L_2 G_2 \ldots L_{N-1} G_{N-1}} \qquad [4]$$

where L_i is the signal loss prior to the gain G_i of the EDF stage i. From eqn [4], multistage amplifiers with large inter-stage losses will have numerically larger NFs, but the total module NF is dominated by the first EDF stage. Undersea EDFAs, with a single short EDF section, very small L_i and 980 nm pumping, can have NF <3.5 dB.

Advanced EDFA Functions

An important trend in optical amplifiers, and systems in general, is the move towards dynamic control. Several technologies are being considered and the primary idea is to dynamically control the intensity level of each optical channel or wavelength to correct for any gain/loss inequalities in the network. This is particularly important as channel counts and bit rates rise concurrently with the functionality demanded of optical communication networks. A common requirement in this category is fast amplifier response so that in a network where individual optical channels can be added or dropped there are no power distortions at other surviving wavelengths. In a typical drop event, in less than 1 microsecond, 31 of 32 wavelengths can be completely removed from the input to the EDFA by an add/drop module located elsewhere in the network. This event will cause a total input power decrease of 15 dB and without rapid pump power adjustment the remaining channel will be excessively amplified.

Erbium ions have an excited state lifetime of approximately 10 ms, and this sets the time-scale for the transient signal response. A network transient usually manifests as a sharp optical amplifier (OA) gain fluctuation lasting up to several tens of microseconds that will pulse large signal levels through the entire optical network. Dynamic network loading necessitates controlling temporal gain transients, and this may be achieved by using a microprocessor in the amplifier module. These intelligent EDFA modules can rapidly adjust the pump laser power based on feedback from internal photo detectors to control gain transients.

The need for dynamic control also exists in the frequency domain. Wideband amplifiers have a gain variation among the optical channels of typically up to 1 dB over all operating conditions, and since these EDFAs are concatenated over many spans the electronic receiver circuit at the end of the network will have to detect optical channels with widely varying power levels. This can limit the number of spans that the network can bridge before electronic regeneration becomes necessary. This problem can be corrected using a dynamic gain equalizing filter (DGEF) to actively attenuate individual optical channels levels. A DGEF can be embedded inside an amplifier to create a loss-less component, and can be used to equalize channel powers for an entire network.

The problems of accumulation of gain ripple become more apparent in systems that employ amplifiers based on stimulated Raman scattering. A conventional EDFA amplified system may have 10 spans transmitting 100 optical channels at 10 Gbit/s each over a distance of approximately 800 km before electronic regeneration becomes necessary. To span longer distances with higher capacity, Raman amplifiers are used to amplify the transmission fiber along with EDFAs. This system architecture has the advantage of a higher effective OSNR for the individual spans, and has been demonstrated for distances over 4000 km with 40 Gbit/s channels. However, the variability of the Raman amplifier gain, due to network loading and differing transmission fiber properties, necessitates the use of loss-less DGEFs.

High-capacity transmission uses individual channels that are each modulated up to typical 2.5, 10, or 40 GHz, that requires that the optical amplifiers have low polarization mode dispersion (PMD) and low chromatic dispersion (CD). Polarization mode dispersion distorts the temporal spread of an ultra-short pulse as different polarization states travel at different speeds through the transmission fiber and amplifier components. PMD is particularly a problem in the first generation of installed transmission fiber and through polarization components (e.g., isolators) that are not PMD compensated. Without PMD compensation, 40 Gbit/s transmission, for example, can be limited to very short distances (several km). Fortunately, PMD compensators have been developed for upgrading aged installed networks, and also amplifiers are available with very low PMD.

To achieve long-distance transmission, a chromatic dispersion map is produced for all the optical spans, and by using a dispersion compensating module in each amplifier, the net dispersion across the channel wavelength window is controlled. Although optical amplifiers have approximately 1000 times less optical fiber in them compared to the transmission span that it is amplifying, chromatic dispersion in the amplifier can also be of concern for very long-haul networks. The CD of the amplifier is primarily in the EDF, and is specific to the EDF's geometric and chemical composition. Furthermore, since the erbium ion is a resonant atomic system, there may be a pump and signal power contribution to the dispersion (resonant dispersion) that could degrade network performance. Characterization of these optical effects is typically done on manufactured 'field grade' amplifiers as part of a quality control process.

Conclusion

As optical networks evolve EDFAs will continue to be an enabling technology for higher capacity and more dynamic communications networks. EDFA technology has already advanced to accommodate multiple channels, to span several wavelength windows and to provide features such as dispersion compensation, gain-transient suppression, and dynamic gain flatness. As future network architectures are introduced, to incorporate Raman amplification and additional wavelength windows, the demands placed upon EDFA design will continue to expand.

See also

Optical Communication Systems: Architectures of Optical Fiber Communication Systems; Wavelength Division Multiplexing. **Scattering:** Raman Scattering.

Further Reading

Agrawal G (1989) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Agrawal G (1997) *Fiber Optic Communication Systems*. New York: John Wiley and Sons.

Becker PC, Olsson NA and Simpson JR (1999) *Erbium-doped Fiber Amplifiers – Fundamentals and Technology*. San Diego, CA: Academic Press.

Desurvire E (1995) *Erbium Doped Fiber Amplifiers*. New York: John Wiley and Sons.

Pedersen B, Bjarklev A, Povlsen JH, *et al.* (1991) The design of erbium-doped fiber amplifiers. *Journal of Lightwave Technology* 9(9): 1105.

Sudo S (1997) *Optical Fiber Amplifiers: Materials, Devices and Applications*. Boston, MA: Artech House.

Optical Amplifiers in Long-Haul Transmission Systems

B M Desthieux, Xtera Communications, Inc., Allen, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical amplifiers have been a key element for enabling the development of optical transmission systems, which have reached total capacity in the order of a few terabits per second. The first type of optical amplifier to see application in the telecom industry has been the erbium-doped fiber amplifier. Following its invention in the late 1980s, the mid-1990s saw its industrial application in submarine and terrestrial transmission systems.

Since then, optical transmission systems have been the predominant source for transmitting information over distances ranging from a few 100 km to more than 10 000 km, connecting cities, countries, and continents.

This article will describe the applications of optical-fiber amplifiers in long-haul transmission systems, focusing on erbium-doped fiber amplifiers and Raman amplifiers, the most popular type of optical amplifiers used in modern transmission optical systems.

Fundamentals of Long-Haul Transmission Systems

This section explains the key elements and parameters of an optical transmission system.

Key Elements

An optical transmission system consists of a transmitter, a receiver, one or more optical amplifiers, and one or more spans of transmission fiber. Figure 1 shows a simplified schematic of a bidirectional transmission system using a fiber pair.

Optical amplifiers are used to compensate for the loss of the transmission fiber and the other optical elements placed along the signal path. Boosters and pre-amplifiers refer to optical amplifiers which are located at, respectively, the input end and output end of the system. In-line amplifiers refer to optical amplifiers located between the transmission fiber spans.

The transmitter includes laser diodes to generate the signal power and optical modulators to transfer the data stream from the electrical input signals to optical signals. The optical signal is transmitted to the receiver after traveling through the transmission fiber. At the receiver are located the photodiodes, which transfer back the information into electrical output signals. In order to increase the total bandwidth of the transmission system, time-division multiplexing (TDM) and wavelength-division multiplexing (WDM) can be used. TDM techniques combine the electrical domain data stream of several electrical channels, which is carried by a single optical channel. An optical channel refers to an optical signal located at a particular wavelength and modulated with the transmission data. TDM increases, therefore, the bit rate supported by the optical channel. Typical bit rates used in modern optical transmission systems are 2.5 Gbit/s and 10 Gbit/s. With WDM, several optical channels are combined, in the optical domain, into a single optical beam, which is launched into the transmission fiber. For this technique, optical multiplexers (Mux) and de-multiplexers (Demux) are required at, respectively, the transmitter and the receiver.

Long-haul transmission systems typically operate at wavelengths around 1.5 μm, where the attenuation of fibers is the lowest. We label C-band the region located between 1530 nm and 1565 nm. Its lower bound is labeled S-band; the upper bound the L-band.

Apart from attenuation, several other effects existing in fiber can be responsible for degrading the quality of the transmission. Among those is chromatic dispersion, which distorts the optical signal as it travels along the transmission fiber. Several types of fiber have been deployed with different chromatic dispersion characteristics. The most common one is the standard single-mode fiber (SSMF), which has a zero-dispersion wavelength – wavelength for which no chromatic dispersion is observed – near 1.3 μm. The dispersion-shifted fiber (DSF) has a zero-dispersion wavelength near 1.55 μm, and the nonzero

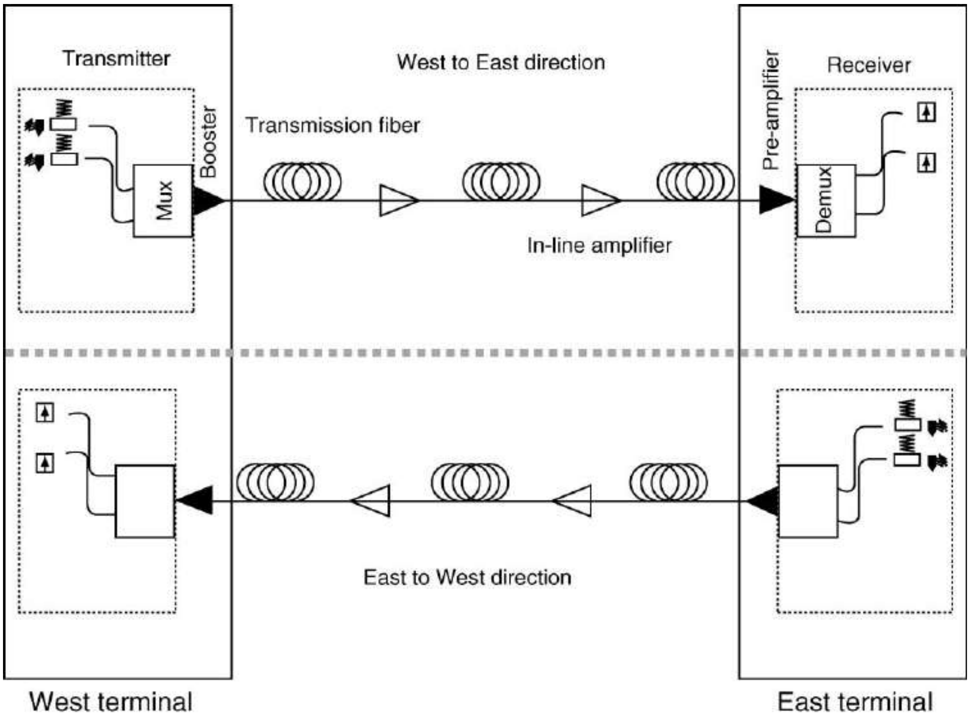

Figure 1 Schematic of a bidirectional long-haul transmission system.

dispersion shifted fiber (NZDSF) has a zero-dispersion wavelength close to 1.5 μm. To remedy this effect, dispersion compensating modules (DCM) are placed along the link. These modules – generally a special type of fiber – have chromatic dispersion with inverted characteristics with respect to those of the transmission fiber.

Key Optical-Amplifier Parameters

Optical amplifiers spontaneously generate optical noise – referred to as amplified spontaneous emission (ASE) – that is added to the modulated signal and consequently degrades the quality of the transmission. It is essential to quantify the level of ASE for a given optical transmission system.

We define the optical signal-to-noise ratio (OSNR) in the bandwidth $\Delta\nu$ as

$$\mathrm{OSNR} = \frac{P_{\mathrm{Signal}}}{P_{\mathrm{ASE}}} \qquad [1]$$

where P_{signal} is the signal power and P_{ASE} is the ASE power in the bandwidth $\Delta\nu$.

The required OSNR at a receiver, to detect the information with high quality, depends on the modulation format and on the other sources of noise or signal degradation existing in the transmission system. If we assume a nonreturn to zero (NRZ) modulation format, and signal detection limited only by ASE noise, an OSNR, in 0.1 nm bandwidth, of about 18 dB, is required at the receiver for a bit-error rate – number of erroneously transmitted bits divided by number of transmitted bits – of 10^{-12}. Introduction of other transmission degradation sources requires a higher OSNR to maintain the same bit-error rate.

The decibel unit (dB) for a given ratio, R, is defined by

$$R_{\mathrm{dB}} = 10\,\log_{10} R_{\mathrm{linear}} \qquad [2]$$

Similarly we defined the power decibel unit (dBm) as

$$P_{\mathrm{dBm}} = 10\,\log_{10} P_{\mathrm{mW}} \qquad [3]$$

where P_{mW} is the power in mW.

In order to quantify the amount of ASE generated by an optical amplifier, we define the equivalent input noise factor (n_{eq}), in a linear unit, by

$$n_{\mathrm{eq}} = \frac{P_{\mathrm{ASE}}}{h\nu\Delta\nu G} \qquad [4]$$

where P_{ASE} is the output ASE power in one polarization state in the bandwidth $\Delta\nu$, h the Planck's constant, ν the frequency of the ASE, and G the gain of the amplifier.

We also define the noise figure of the amplifier (NF) – a universal parameter to characterize amplifier noise – as

$$\mathrm{NF} = 2n_{\mathrm{eq}} + \frac{1}{G} \qquad [5]$$

If an optical component, having a loss T_{in} ($T_{\mathrm{in}} < 1$), is inserted at the input of the amplifier, the total NF of the amplifier becomes $\mathrm{NF}_{\mathrm{total}} = \mathrm{NF}/T_{\mathrm{in}}$. If an optical component, having a loss T_{out} ($T_{\mathrm{out}} < 1$), is inserted at the output of the amplifier, $\mathrm{NF}_{\mathrm{total}} = \mathrm{NF} + 1/G(1/T_{\mathrm{out}} - 1)$. When $G \gg 1$, we have $\mathrm{NF}_{\mathrm{total}} \approx \mathrm{NF}$.

The OSNR in 0.1 nm bandwidth obtained after N fiber spans of loss, $\mathrm{Loss}_{\mathrm{Fiber}}$ (in dB), each having an amplifier, is in decibels:

$$\mathrm{OSNR} = P_{\mathrm{Signal}} - \mathrm{Loss}_{\mathrm{Fiber}} - \mathrm{NF} - 10\log_{10} N + 58 \qquad [6]$$

where P_{Signal} is the signal power at the output of the amplifier.

When the optical signal power is sufficiently high, nonlinear effects also become responsible for the degradation of the system performance. Therefore, choosing the value of the signal power launched into the transmission fiber results in a trade-off between increasing the OSNR, and reducing nonlinear effects.

Optical Fiber Amplifier

An optical fiber amplifier consists of fiber with an active medium and at least one optical pump source, which is launched simultaneously with the signal into the fiber by means of a dichroic coupler.

We define as forward and backward pumping, the case where the pump is launched, respectively, in the same direction as the signal and in the opposite direction to the signal. We define as bidirectional pumping, the case where the fiber is pumped from both ends. The most commonly used pump sources are the semiconductor laser diodes. These lasers may use different technologies to meet the various performance requirements with respect to the amplifier types and applications. The most popular pump diodes are the distributed Bragg reflector (DBR), distributed feedback (DFB), or external-grating stabilized lasers.

Erbium Fiber Amplifier

Fiber doped with the rare-earth Er^{3+} ion can amplify a signal located at 1.5 μm with the appropriate fiber host material and if pumped near 0.98 μm or near 1.48 μm. Silica is the most common material used for the fiber host. Er^{3+} has three energy levels, as described in Figure 2.

0.98 μm pumping occurs between the upper level $^4\mathrm{I}_{11/2}$ and the ground level $^4\mathrm{I}_{15/2}$, and 1.48 μm pumping occurs between the metastable level $^4\mathrm{I}_{13/2}$ and the ground level.

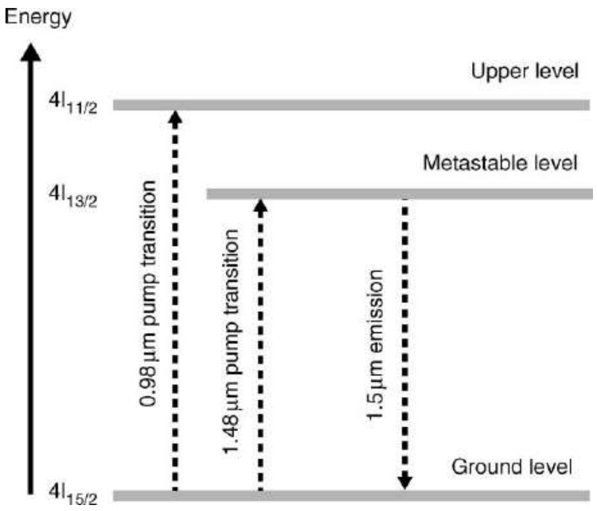

Figure 2 Energy level diagram for Er^{3+} ions.

The signal amplification takes place between the metastable level and the ground level. With 0.98 μm pumping, a rapid energy transfer occurs between the upper level and the metastable level via a phonon-assisted process. The lifetime of the upper level is much lower than the lifetime of the metastable level, about 10 ms. Therefore, with 0.98 μm pumping we can assume a two-level energy transfer, between the metastable and the ground levels, for the description of the amplification process in the Er^{3+} medium.

We define the normalized medium inversion (D) by

$$D = \frac{N_2 - N_1}{N_t} \qquad [7]$$

where N_2 is the population of the metastable level, N_1 the population of the ground level, and $N_t = N_2 + N_1$ the total population. $D = +1$ when the metastable level is fully populated and $D = -1$ when only the ground level is populated.

A simple approach for calculating the gain (G) with erbium-doped fiber amplifiers (EDFA) is given, in decibel units, by

$$G(\lambda) = \left((\alpha(\lambda) + g(\lambda)\right)\frac{D+1}{2} - \alpha(\lambda)L \qquad [8]$$

where λ is the wavelength of the signal, α the absorption coefficient in dB unit length at the signal wavelength, g the emission coefficient in dB unit length at the signal wavelength, and L the length of the erbium-doped fiber (EDF). Here D is the normalized path-average medium inversion of the EDF.

The gain per unit length, $G(\lambda)/L$, is equal to $g(\lambda)$ for $D = +1$ and to $-\alpha(\lambda)$ for $D = -1$.

The absorption and the emission coefficients depend on the EDF host material and the codoping element used for the realization of the EDF. With silicate EDF the most commonly used codoping elements are germanium, aluminum, and phosphorus.

In Figure 3 we show the dependence of the gain per unit length on the normalized path-average medium inversion, for a signal wavelength ranging from 1500 to 1560 nm.

For a more comprehensive description of the EDFA characteristics we use the standard confined-doping model. This model, which is also based on the two-level system, uses the following differential equations:

$$\frac{dP_S^+(z)}{dz} = \left(\frac{N_2}{N_t}(\alpha_S + g_S) - \alpha_s\right)P_S^+(z) \qquad [9]$$

$$\begin{aligned}\frac{dP_{ASE}^+(z)}{dz} &= \left(\frac{N_2}{N_t}(\alpha_S + g_S) - \alpha_s\right) \\ &\quad \times P_{ASE}^+(z) + \frac{N_2}{N_t} g_S 2h\nu_S \Delta\nu \end{aligned} \qquad [10]$$

$$\begin{aligned}\frac{dP_{ASE}^-(z)}{dz} &= \left(\alpha_s - \frac{N_2}{N_t}(\alpha_S + g_S)\right) \\ &\quad \times P_{ASE}^-(z) - \frac{N_2}{N_t} g_S 2h\nu_S \Delta\nu \end{aligned} \qquad [11]$$

$$\frac{dP_P^+(z)}{dz} = \left(\frac{N_2}{N_t}(\alpha_P + g_P) - \alpha_P\right)P_P^+(z) \qquad [12]$$

$$\frac{dP_P^-(z)}{dz} = \left(\alpha_P - \frac{N_2}{N_t}(\alpha_P + g_P)\right)P_P^-(z) \qquad [13]$$

where $P_S(z)$, $P_{ASE}(z)$, and $P_P(z)$ in W are, respectively, the signal power, the ASE power, and the pump power along the length (z-axis) of the EDF; $^+$ and $^-$ indicate, respectively, the backward-traveling and the forward-traveling directions; α_s and α_P in m^{-1} are the absorption coefficients at, respectively, the signal and pump wavelengths, g_s and g_p in m^{-1} are the emission coefficients at, respectively, the signal and pump wavelengths, ν_s and ν_p are the frequencies of, respectively, the signal and the pump beams; $\Delta\nu$ is the frequency bandwidth of the ASE; and h is the Planck's constant. N_2/N_1 is the medium inversion along the z-axis of the EDF, and is given by

$$\frac{N_2}{N_t} = \frac{\sum_k \dfrac{P_k(z)\alpha_k \tau}{h\nu_k B_{\text{eff}}^2 N_t}}{1 + \sum_k \dfrac{P_k(z)(\alpha_k + g_k)\tau}{h\nu_k B_{\text{eff}}^2 N_t}} \qquad [14]$$

where the index k represents all the beams, i.e., signal, ASE, and pump, traveling backward and forward along the EDF, τ is the lifetime of the metastable level, and B_{eff} is the effective erbium-doping area. For a typical silicate EDF, having a peak absorption coefficient of $1\ \text{dBm}^{-1}$, we have $\tau = 10$ ms, $B_{\text{eff}} = 12\ \mu\text{m}^2$, and $N_t = 5 \times 10^{23}\ \text{m}^{-3}$.

Using the standard confined-doping model we calculate the dependence of the EDFA noise figure on the normalized path-average medium inversion, as shown in Figure 4. We vary the path-average normalized medium inversion by changing the length of the EDFA, while adjusting the pump power so that the signal gain is kept constant.

The noise figure decreases when the medium inversion increases, and approaches the quantum limit value of 3 dB when the path-average normalized medium inversion is close to unity. Near full inversion

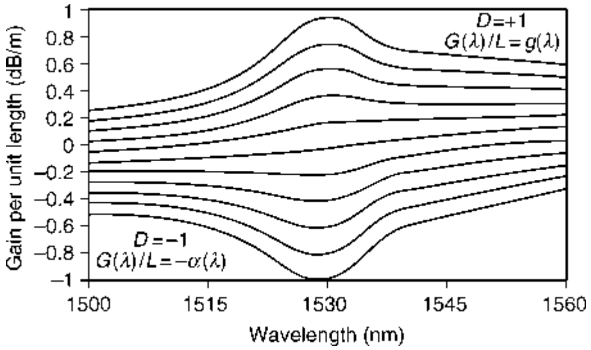

Figure 3 Dependence of the gain per unit length on the normalized path-average medium inversion, *D*. *D* ranging from −1 to +1 with a 0.2 step (alumino-silicate EDF; peak absorption of 1 dB m^{-1}; peak emission of 0.93 dB m^{-1}).

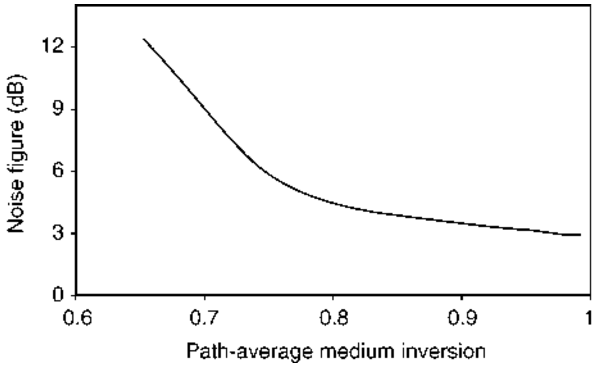

Figure 4 Dependence of the noise figure on the normalized medium inversion (alumino-silicate EDF with peak absorption of 1 dB m^{-1}; EDF lengths ranging from 16 m to 60 m; gain equals to 15 dB; input signal power of −40 dBm at 1530 nm; backward pumping at 980 nm).

may be obtained with a pump located at 0.98 μm because the emission coefficient is null at this pump wavelength. However, this is not the case at 1.48 μm where the ratio g_p/α_p is about 0.3. The maximum medium inversion that can be obtained with 1480 nm is $D = 0.8$. Therefore, higher noise figures are obtained with 1.48 μm pumping than with 0.98 μm pumping. The pumping direction also has an effect on the noise figure, in particular with 0.98 μm. Since at 0.98 μm the emission coefficient is null, when backward pumping configuration is used the 0.98 μm pump is quickly absorbed along the EDF length. This leads to a low-medium inversion at the input end of the EDF, resulting in high noise figures. However, in order to achieve forward pumping, a dichroic coupler must be placed at the input end of the EDF, thus inducing an additional loss at the input of the amplifier. This increases the noise figure and, therefore, to some extent, counterbalances the NF improvement associated with forward pumping.

Raman Fiber Amplifier

Stimulated Raman scattering, an optical-phonon based nonlinear effect present in silica fiber, is used to convert energy from a pump located at a frequency ν_p, to a signal located at a lower frequency ν_s. The efficiency of the so-called Raman amplification depends on the frequency shift $\Delta\nu_r = \nu_p - \nu_s$. Figure 5 shows the dependence of the Raman gain on the frequency shift in standard single-mode fibers. The exact Raman gain spectrum depends on the doping elements used during the fabrication process of the fiber. With standard telecommunications fibers the Raman gain is maximum for $\Delta\nu_r$ close to 13 THz.

Raman amplification can be achieved within the transmission fiber of the system – distributed amplification – or within a specific fiber, such as a dispersion compensation fiber – discrete amplification. Contrary to EDFA, Raman amplification has a very fast response time – in the order of a few hundred femtoseconds – and is strongly dependent on the polarization state of the signal and the pump, the highest gain efficiency being obtained when the polarization states of the signal and the pump are parallel. Because of the nonpolarization-maintaining nature of standard telecommunications fibers, depolarized pump sources are preferable for Raman amplification. To avoid pump fluctuation being transferred to the signal, Raman amplifiers generally use backward pumping, which virtually averages the fluctuation. Forward pumping can also be employed with Raman amplifiers but with careful control of the pump characteristics.

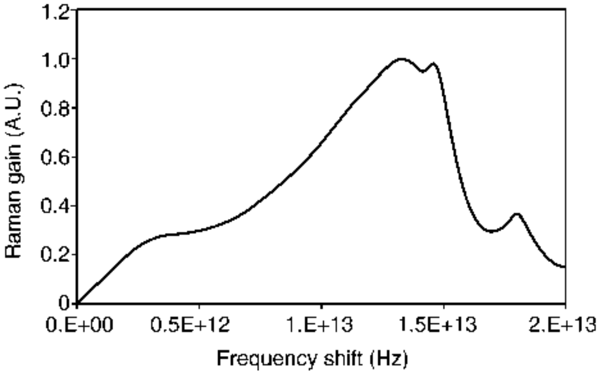

Figure 5 Raman gain spectrum in standard single mode fiber.

If the signal is sufficiently low the pump depletion can be neglected. In that case the gain of the Raman amplifier is given by

$$G = \exp\left(\frac{g_R P_P}{A_{\mathrm{eff}} \alpha_p} (1 - \exp^{-\alpha_p L}) - \alpha_S L \right) \qquad [15]$$

where P_p is the launched pump power, L is the length of the fiber, g_R is the Raman gain coefficient, and A_{eff} is the effective area of the fiber. We assume for the simulation examples given in this article that the peak value of g_R is equal to 6×10^{-14} m/W.

A more comprehensive model for Raman amplification is given by the following set of differential equations:

$$\frac{\mathrm{d}P_S^+(z)}{\mathrm{d}z} = -\alpha_s P_S^+(z) + \left(\frac{g_R}{A_{\mathrm{eff}}}\right)(P_p^+(z) + P_p^-(z))P_S^+(z) + r_S P_S^-(z) \qquad [16]$$

$$\frac{\mathrm{d}P_S^-(z)}{\mathrm{d}z} = \alpha_s P_S^-(z) - \left(\frac{g_R}{A_{\mathrm{eff}}}\right)(P_p^+(z) + P_p^-(z))P_S^-(z) - r_S P_S^+(z) \qquad [17]$$

$$\frac{\mathrm{d}P_{\mathrm{ASE}}^+(z)}{\mathrm{d}z} = -\alpha_s P_{\mathrm{ASE}}^+(z) + \left(\frac{g_R}{A_{\mathrm{eff}}}\right)(P_p^+(z) + P_p^-(z)) \times [P_{\mathrm{ASE}}^+(z) + 2h\nu_S\Delta\nu] + r_S P_{\mathrm{ASE}}^-(z) \qquad [18]$$

$$\frac{\mathrm{d}P_{\mathrm{ASE}}^-(z)}{\mathrm{d}z} = \alpha_s P_{\mathrm{ASE}}^-(z) - \left(\frac{g_R}{A_{\mathrm{eff}}}\right)(P_p^+(z) + P_p^-(z)) \times [P_{\mathrm{ASE}}^-(z) + 2h\nu_S\Delta\nu] - r_S P_{\mathrm{ASE}}^+(z) \qquad [19]$$

$$\frac{\mathrm{d}P_p^+(z)}{\mathrm{d}z} = -\alpha_p P_p^+(z) - \left(\frac{\nu_p}{\nu_S}\right)\left(\frac{g_R}{A_{\mathrm{eff}}}\right)P_p^+(z) \times (P_S^+(z) + P_S^-(z) + P_{\mathrm{ASE}}^+(z) + P_{\mathrm{ASE}}^-(z)) + r_P P_p^-(z) \qquad [20]$$

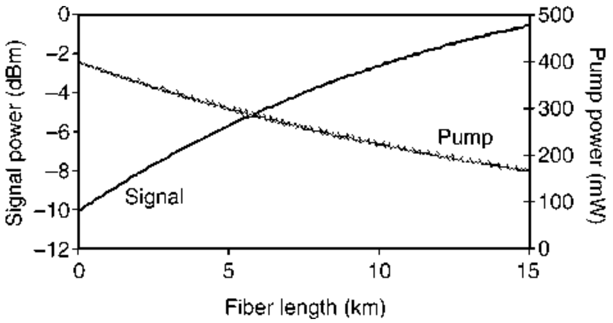

Figure 6 Evolution of the signal power and pump power along fiber length with Raman amplification (standard single-mode fiber; fiber length of 15 km; input signal power of –10 dBm at 1550 nm; forward pumping with 400 mW at 1450 nm; total amplifier gain of 9.5 dB).

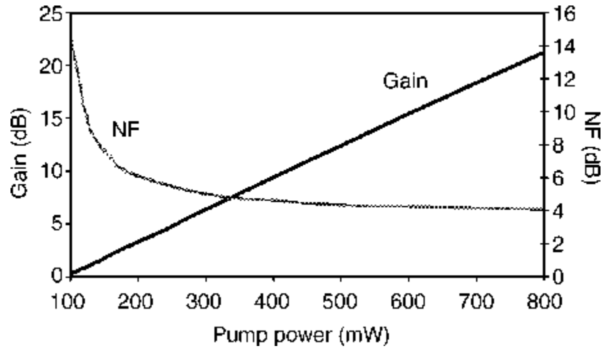

Figure 7 Dependence of the Raman amplifier gain and the noise figure on the pump power (standard single-mode fiber; fiber length of 15 km; input signal power of −10 dBm at 1550 nm; forward pumping at 1450 nm with power ranging from 100 mW to 800 mW).

$$\frac{\mathrm{d}P_{\mathrm{p}}^{-}(z)}{\mathrm{d}z} = \alpha_{\mathrm{p}} P_{\mathrm{p}}^{-}(z) + \left(\frac{\nu_{\mathrm{p}}}{\nu_{\mathrm{S}}}\right)\left(\frac{g_{\mathrm{R}}}{A_{\mathrm{eff}}}\right) P_{\mathrm{p}}^{-}(z) \times \left(P_{\mathrm{S}}^{+}(z) + P_{\mathrm{S}}^{-}(z) + P_{\mathrm{ASE}}^{+}(z) + P_{\mathrm{ASE}}^{-}(z)\right) - r_{\mathrm{P}} P_{\mathrm{p}}^{+}(z) \qquad [21]$$

where r_{S} and r_{p} are the Rayleigh backscattering coefficients at, respectively, the signal wavelength and the pump wavelength.

An example of the evolution of the signal power and pump power along fiber length is given by Figure 6.

Figure 7 shows the dependence of the Raman amplifier gain and noise figure on the pump power.

Distributed and Discrete Amplification

A system uses distributed amplification when the transmission fiber is used as the medium for generating the optical signal amplification. It uses discrete amplification when an optical amplifier is located between the transmission fiber spans.

Discrete Amplification

Both EDFA and Raman amplifiers can be used to provide discrete amplification. Discrete amplifiers can be used as booster, pre-amplifier, or in-line amplifier. They can also be used to compensate for the loss of additional elements such as DCM, which may be located between the fiber spans. In Figure 8 we show the path of the signal when a discrete amplification is used to compensate for the loss of the transmission fiber and for the loss of a DCM. In this example, we assume that the discrete amplifier is made of two amplifiers. The first-stage amplifier compensates mainly for the loss of the transmission fiber, while the second-stage amplifier compensates mainly for the loss of the DCM.

If a ‘lossy’ element is inserted between two amplifiers, the noise figure of the composite amplifier increases the total noise figure of the dual-stage amplifier.

The total noise figure of a dual-stage amplifier, $\mathrm{NF}_{\mathrm{total}}$, is given by

$$\mathrm{NF}_{\mathrm{total}} = \mathrm{NF}_1 - \frac{1}{G_1} + \frac{\mathrm{NF}_2}{G_1 T_{\mathrm{mid}}} \qquad [22]$$

where NF_1 is the noise figure of the first-stage amplifier, G_1 is the gain of the first-stage amplifier, NF_2 is the noise figure of the second-stage amplifier, and T_{mid} is the loss of the optical element inserted between the two amplifiers ($T_{\mathrm{mid}} < 1$).

Distributed Amplification

When fibers are deployed it is generally assumed, at least for terrestrial networks, that it can be used with different types of transmission systems, which may or

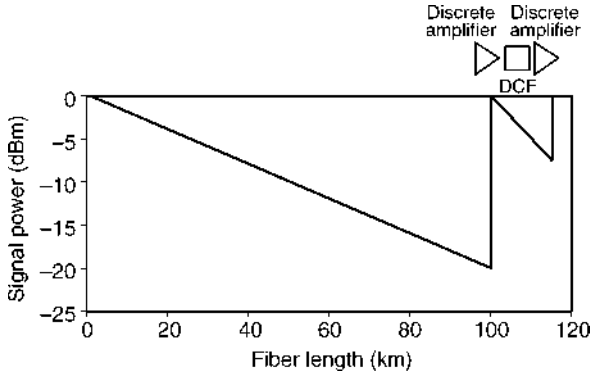

Figure 8 Schematic of a transmission system using discrete amplification to compensate for the loss of the transmission fiber and the dispersion compensating module.

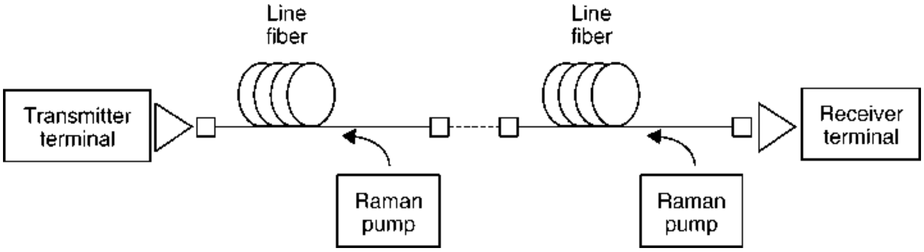

Figure 9 Schematic of a transmission system using distributed Raman amplification.

may not use distributed amplification. Doping the transmission fiber with erbium is not a practical solution since the absence of a pump would result in high fiber loss. On the other hand, Raman amplification can be achieved with conventional fibers, and is consequently the preferred solution to generate distributed amplification. Raman amplification is then obtained by launching a pump signal into the transmission fiber. Figure 9 shows the schematic of a transmission system using distributed Raman amplification with backward pumping.

Even if distributed amplification can be used to compensate fully for the loss of the transmission fiber, discrete amplifiers are generally required to compensate for the loss of the transmitter and receiver terminals. In order to compare distributed amplification with discrete amplification we define the equivalent noise figure as follows:

$$\mathrm{NF}_{\mathrm{Equ}} = \frac{1}{G_{\mathrm{On/Off}}}\left(\frac{\mathrm{Pase}}{h\nu\Delta\nu} + 1\right) \qquad [23]$$

where Pase is the total ASE power at the output end of the transmission fiber and where the on/off gain, $G_{\mathrm{On/Off}}$ is given by

$$G_{\mathrm{On/Off}} = \frac{P_{\mathrm{signal}}^{\mathrm{On}}}{P_{\mathrm{signal}}^{\mathrm{Off}}} \qquad [24]$$

where $P_{\mathrm{signal}}^{\mathrm{On}}$ and $P_{\mathrm{signal}}^{\mathrm{Off}}$ are the signal power levels at the output end of the transmission fiber when the distributed amplifier is, respectively, turned on and off.

To illustrate the meaning of the on/off gain, we show in Figure 10 the evolution of the signal power along a fiber with and without Raman pump power. In this case, the on/off gain is 20 dB.

The equivalent noise figure, which has no real physical meaning, is equal to the noise figure of an imaginary discrete amplifier that would be located at the output end of the transmission fiber and would lead to the same OSNR. Figure 11 shows the dependence of the equivalent noise figure on the Raman gain, with no Rayleigh backscattering and with a Rayleigh backscattering level of $-74\ \mathrm{dB\ m^{-1}}$.

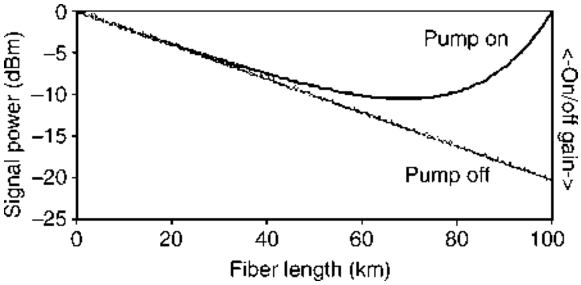

Figure 10 Evolution of signal power along the fiber length with and without pump power (standard single-mode fiber; fiber length of 100 km; fiber attenuation of 0.2 dB/km at signal wavelength; input signal power of 0 dBm at 1550 nm; backward pumping 400 mW at 1450 nm).

Once the on/off gain is higher than ~10 dB the equivalent noise figure becomes negative and it decreases when the on/off gain further increases. However, once the on/off is higher than 20 dB, Rayleigh backscattering limits the minimum equivalent noise figure that can be achieved. The lowest equivalent noise figure is observed for an on/off gain of ~33 dB. In addition, Rayleigh backscattering causes the reflected portion of the signal to be superimposed onto the transmitted signal, leading to multipath interference (MPI), which level increases with the gain. This results in an additional source of noise for the signal, and consequently degrades the system performance.

Hybrid Amplification

Discrete and distributed amplification can be used simultaneously in a same-transmission system. An example of the evolution of the signal power with hybrid amplification is represented in Figure 12, for the case where distributed Raman amplification and a dual-stage discrete amplifier are used to compensate for the loss of the transmission fiber and for the loss of a DCM.

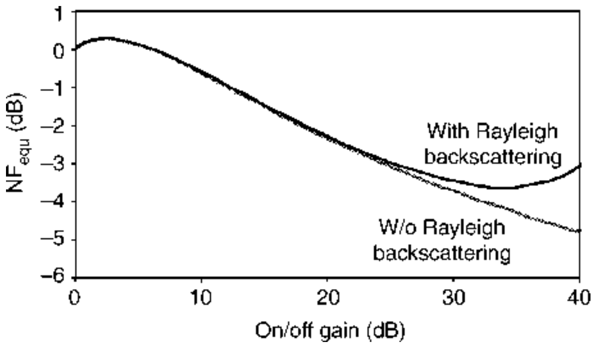

Figure 11 Equivalent noise figure versus distributed on/off gain (standard single-mode fiber; fiber length of 250 km; fiber attenuation of 0.2 dB/km at 1550 nm; input signal power of 0 dBm at 1550 nm; backward pumping ranging from 0 to 1 W at 1450 nm).

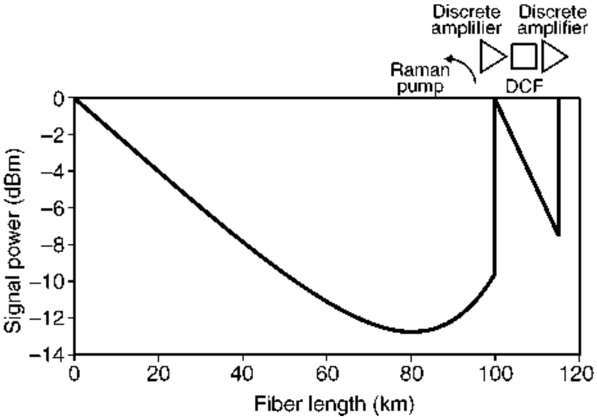

Figure 12 Evolution of the signal power using discrete and distributed amplification (standard single-mode fiber of 100 km; fiber loss of 20 dB at signal wavelength; DCM loss of 8 dB; input signal power of 0 dBm at 1550 nm; Raman backward pumping with 200 mW at 1450 nm; Raman on/off gain of 10 dB).

Similar to the case of distributed amplification, the total equivalent noise figure of a hybrid amplifier can be defined by

$$\mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Total}} = \mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Distributed}} + \frac{\mathrm{NF}^{\mathrm{Discrete}} - 1}{G_{\mathrm{On/Off}}} \qquad [25]$$

where $\mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Distributed}}$ is the equivalent noise figure of the distributed amplifier and $\mathrm{NF}^{\mathrm{Discrete}}$ is the noise figure of the discrete amplifier.

In Figure 13 we show the dependence of the total equivalent noise figure on the on/off gain for a hybrid amplifier with fixed discrete noise figure. The dependence of the distributed equivalent noise figure on the on/off gain is also plotted on the same graph.

Since the equivalent noise figure of the distributed amplifier is relatively low, the total equivalent noise figure is lower than the noise figure of the discrete amplifier once distributed amplification is used. Moreover, as the on/off gain becomes large, the total equivalent noise figure of the hybrid amplifier converges toward the distributed equivalent noise figure.

Wavelength-Division Multiplexing

The transmission of WDM signals over long distances requires meticulous control of the spectral characteristics of the amplifier gain. The amplifier gain excursion induces a power excursion between the channels of the WDM signal, which increases with the number of spans. Consequently, the available bandwidth for WDM transmission narrows with the total distance of the link. This is illustrated in Figure 14, where the power excursion observed after an amplifier chain is plotted for different numbers of amplifiers, a fiber span being included between each amplifier. With this example, the 1 dB-bandwidth drops from 9 nm after one span to 1.5 nm after 50 spans.

For a comprehensive analysis of the spectral gain characteristic of EDFA and Raman amplifiers we generalize the single-channel model described in the amplifier section of this article to the multichannel case. The generalization of the models is obtained by adding a frequency dimension to each beam propagating in the amplifier, as described in Table 1, where each k element represents the signal, the ASE noise or the pump, λ is the wavelength of the kth beam, $\Delta\nu$ is the frequency separation between the pump beam p, and the signal beam s.

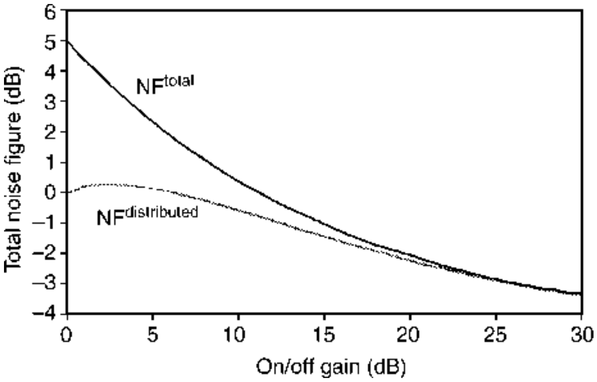

Figure 13 Total noise figure of the hybrid amplifier and noise figure of the distributed amplifier versus on/off gain (noise figure of the discrete amplifier equals to 5 dB; standard single-mode fiber of 100 km; fiber loss of 20 dB at signal wavelength; input signal power of 0 dBm at 1550 nm; Raman backward pumping ranging from 0 mW to 600 mW at 1450 nm).

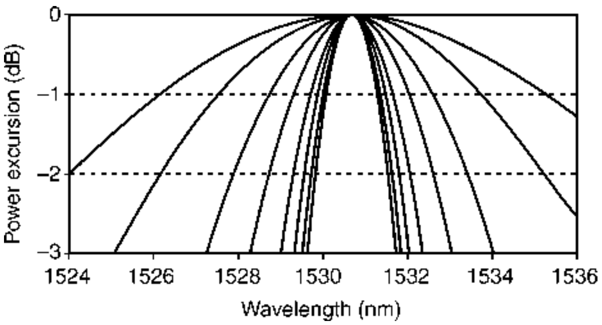

Figure 14 Dependence of the power excursion on the number of fiber spans. Number of spans equals, respectively, to 1, 2, 5, 10, 20, 30, 40, and 50 (amplifier with 15 m long alumino-silicate EDF; peak absorption of 1 dB m^{-1}; $D = 0.6$; fiber loss of 8.5 dB; flat fiber loss spectrum).

Table 1 Correspondence between parameters used for the EDFA and Raman amplifier models with the single-channel and the multichannel approaches (the multichannel approach is also used to describe the power evolution of pump beam when WDM pump scheme is used)

	Single-channel model	*Multichannel model*
Power	$P_k^{+,-}(z)$	$P_k^{+,-}(\lambda, z)$
Sum of power	$P_k^{+}(z) + P_k^{-}(z)$	$\sum_\lambda P_k^{+}(\lambda, z) + P_k^{-}(\lambda, z)$
Absorption coefficient	α_k	$\alpha(\lambda)$
Emission coefficient	g_k	$g(\lambda)$
Raman gain coefficient	g_R	$g_R(\Delta\nu_{p,s})$

With EDFAs, the main factors in the gain flatness are the path-average medium inversion, the EDF codopant composition, and fiber host material. Depending on the bandwidth and the wavelength of the WDM signal, the way the EDF is optimized may vary.

In the following example, we focus on the path-average medium inversion for the case of C-band applications. Figure 15 shows the output spectrum of a WDM signal amplified by an alumino-silicate EDFA having, respectively, a medium inversion of $D = +0.4$ and $D = 0$. The peak-to-peak gain excursion is, respectively, 2 dB and 14 dB for $D = +0.4$ and $D = 0$.

Application with C-band WDM signals requires amplifiers operating with high-medium inversion. As can be extrapolated from Figure 15, if the WDM signal is centered around 1560 nm, a lower medium inversion is more suitable. Indeed, as shown in Figure 3, when the medium inversion is close to zero the peak gain wavelength changes from 1530 nm to 1560 nm and higher; the optical window for L-band applications.

Regarding the doping elements, we show in Figure 16 the gain spectrum for a germano-silicate and an alumino-silicate EDF, respectively. The peak-to-peak gain excursion is 8 dB with the germano-silicate EDF compared to 2 dB with the alumino-silicate EDF.

With Raman amplifiers, the gain flatness can be adjusted by using wavelength-multiplexed pumps. In Figure 17 we show the power spectrum when a Raman amplifier is pumped with three wavelength-multiplexed pumps located, respectively, at 1430 nm, 1455 nm, and 1480 nm. The maximum power excursion between the channels is 2 dB.

For comparison, we also show in Figure 18 the power spectrum of the Raman amplifiers when only

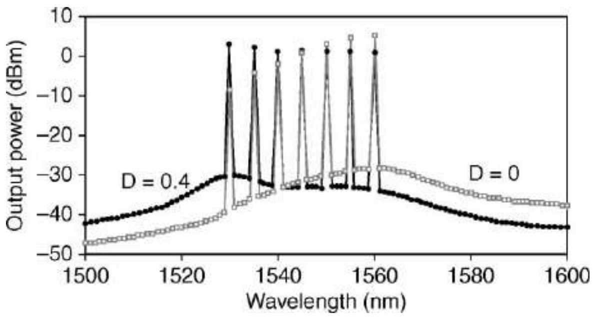

Figure 15 Gain spectrum of an alumino-silicate EDF for $D = 0.4$ and $D = 0$ (C-band WDM signal made of 7 channels equally spaced ranging from 1530 nm to 1560 nm; flat input power of −10 dBm per channel; gain of channel located at 1545 nm of 10 dB for $D = 0.4$ and $D = 0$; EDF with peak absorption of 1 dB m^{-1}; EDF length of 30 m for $D = 0.4$; EDF length of 100 m for $D = 0$).

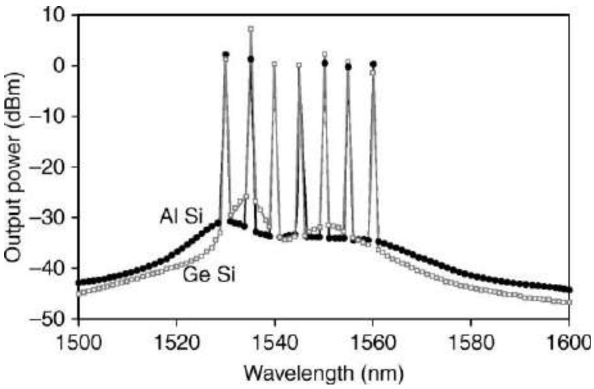

Figure 16 Gain spectrum of an alumino-silicate EDF and of a germano-silicate EDF for $D = 0.4$ (C-band WDM signal made of 7 channels equally spaced ranging from 1530 nm to 1560 nm; flat input power of −10 dBm per channel; gain of channel located at 1545 nm of 10 dB for $D = 0.4$ and $D = 0$; alumino-silicate EDF of 30 m with peak absorption of 1 dB m^{-1}; germano-silicate EDF of 40 m with peak absorption of 1 dB m^{-1}).

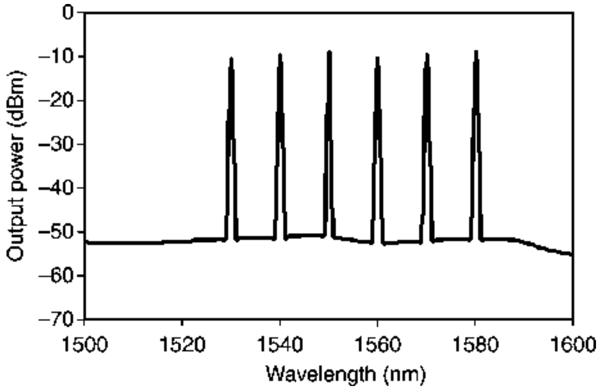

Figure 17 Output power spectrum of a Raman amplifier pumped with three wavelength-multiplexed pumps (standard single-mode fiber of 100 km; WDM signal made of 6 channels equally spaced ranging from 1530 nm to 1580 nm; pumping in forward direction with 100 mW per pump located at 1430 nm, 1455 nm, and 1480 nm).

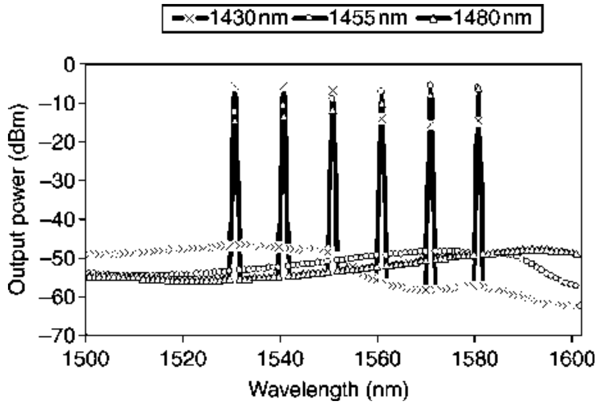

Figure 18 Output power spectrum of Raman amplifiers pumped at three different wavelengths (standard single-mode fiber of 100 km; WDM signal made of 6 channels equally spaced ranging from 1530 nm to 1580 nm; pumping in forward direction with 300 mW at, respectively, 1430 nm, 1455 nm, and 1480 nm).

one of the three multiplexed pumps is used, with the same total pump power. The power excursion becomes 10 dB for the 1430 nm pump, 7 dB for the 1455 nm pump, and 8.5 dB for the 1480 nm pump, which are much higher values than are obtained with the multiplexed pump scheme.

To further reduce the gain excursion of the amplifiers an external gain equalizing filter can be used. The filter, which is located next to the amplifier, is manufactured so that its attenuation spectrum matches the gain spectrum of amplifier, as illustrated in Figure 19. Gain equalizing filters are generally made with thin film filters or fiber Bragg gratings (FBG).

In order to obtain amplification over ultra-wide bandwidths, amplifiers of smaller bandwidths, e.g., C-band and L-band EDFAs, can be operated in parallel by means of Mux and Demux. Another option is to use a Raman amplifier with multiple pumps having their wavelengths distributed over a very large bandwidth.

Long-Haul Transmission System Types

Long-haul optical transmission systems are appropriate for different types of applications. We distinguish here three types of transmission system: the submarine systems; the terrestrial systems; and the unrepeatered systems. Each of these applications requires a careful design of the system, with particular attention to the optical amplifiers. In this section we focus on the key points of their design.

Submarine Systems

Submarine systems are used to transmit information across the oceans, i.e., over distances of up to several thousand kilometers. The transmission fiber and the in-line optical amplifiers, which are packaged in the repeater modules, are deployed undersea at the same time. The drastic constraints imposed on the reliability of the repeaters lead to their high cost, and consequently minimizing their number in a link is essential. The optimum span length, and therefore the number of in-line amplifiers, depends on the total distance of the transmission system. To illustrate this point, we derive from eqn [6] which number of spans of a particular length can be used for a given received OSNR, launched signal power, amplifier noise figure, and span loss per km. Based on the number of spans and span length, we then calculate the maximum distance that can be reached. The result is represented in Figure 20.

With the assumption made here regarding the amplifier characteristics, in order to reach distances of more than 5000 km, span lengths close to 50 km are required.

Terrestrial Systems

Terrestrial systems are used to connect traffic from cities to cities and from cities to submarine-system landing sites. Their total distances vary from a few hundred kilometers to several thousand kilometers. Generally with terrestrial systems, the deployments of the transmission fibers and the transmission systems are done at different times. During its lifetime, the same fiber plant might be used with a different generation of transmission system. Furthermore, because of the geographical constraints that can be imposed on the

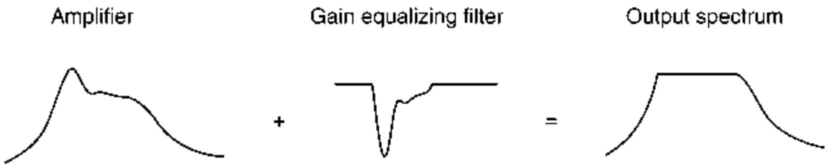

Figure 19 Combination of optical amplifier and gain equalizing filter to provide a flat gain amplifier.

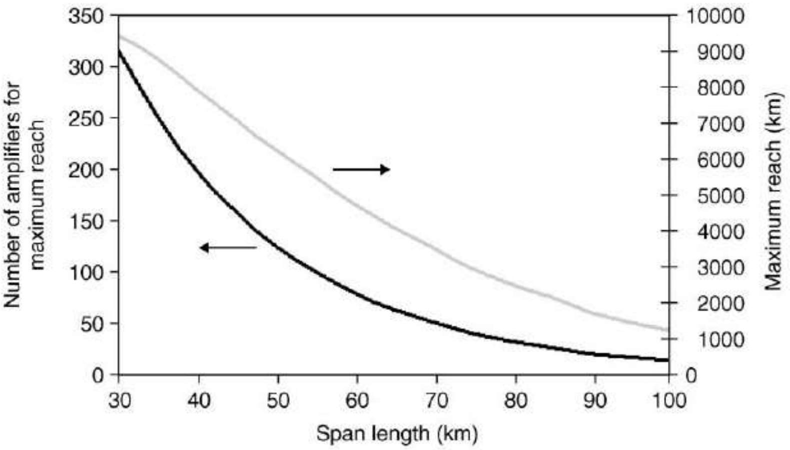

Figure 20 Dependence of the maximum distance and span number of a system on its span length (signal power at the output of the amplifier of −2 dBm; amplifier noise figure of 5 dB; fiber loss of 0.2 dB/km; required OSNR of 20 dB/0.1 nm).

locations of the terminals and the in-line optical amplifiers, the different span lengths may greatly vary in a given link. Consequently, an important characteristic of the terrestrial systems is their capability to operate with different span parameters.

For terrestrial systems using large numbers of WDM channels, maintaining the gain excursion of the optical amplifiers as little as possible under different operating conditions is crucial. Approaches to limit the gain excursion under these variations can be taken. With EDFAs of a given length, the medium inversion depends on its operating gain. The EDFA can be forced to operate at constant gain, independent of the span loss, by adding an adjustable attenuator to the amplifier, e.g., in the middle of a dual-stage amplifier, such that the sum of the attenuator loss and the span loss remains constant. With Raman amplifiers, the control of the gain excursion can be realized by adjusting the power of the pumps when a wavelength-multiplexed pumping scheme is used. Another possibility is to insert dynamically adjustable gain flattening devices along the link. The automatic adjustment of the gain flatness requires the need of a feedback generally provided by an optical channel power monitor.

Unrepeatered Systems

Unrepeatered systems are used to connect mainland sites separated by a water area, such as islands or cities located on the coast. Their main characteristic is that they do not use any electronic component along the transmission fiber; therefore optical amplifiers cannot be placed under water. The length of such systems is typically of a few hundred kilometers, meaning that, even for those systems where very low loss fiber is used, the loss of the transmission fiber is in the order of several tens of dB. To allow sufficiently high OSNR at the receiver, several options can be used. Among them are the use of very high-power boosters, distributed Raman amplification, and remotely pumped discrete amplification. Remote amplification is obtained by inserting, along the submerged fiber cable, a piece of EDF. A 1.48 μm pump is launched into the EDF from the mainland via the transmission fiber or via a dedicated fiber running along the transmission fiber. Figure 21 shows a schematic of a possible configuration for unrepeatered systems when a very high-power booster, distributed Raman amplification, and remote discrete amplification are used simultaneously.

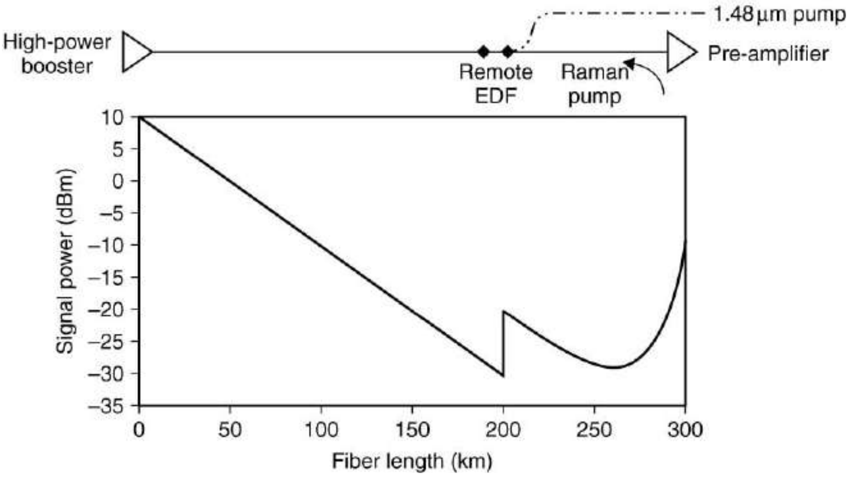

Figure 21 Schematic of unrepeatered systems using high-power booster, remote EDFA, and distributed Raman amplification.

The total equivalent noise figure, NF_{Equ}^{Total}, of an unrepeatered transmission system, using remote discrete amplification and distributed amplification, is given by

$$NF_{Equ}^{Total} = T NF^{Remote} + \frac{1}{G_{Remote}} \times \left(NF_{Equ}^{Distributed} + \frac{NF^{Discrete} - 1}{G_{On/Off}} - T \right) \qquad [26]$$

where NF^{Remote} and G_{Remote} are, respectively, the noise figure and the gain of the remote amplifier; T is the loss of the transmission fiber between the remote amplifier and the pre-amplifier ($T < 1$); $NF_{Equ}^{Distributed}$ and $G_{On/Off}$ are, respectively, the equivalent noise figure and on/off gain of the distributed amplifier; and $NF^{Discrete}$ is the noise figure of the pre-amplifier.

Assuming that the gain and noise figures of the remote EDFA are, respectively, 10 dB and 7 dB, that the remote EDFA is located at 100 km from the pre-amplifier, that the fiber loss per km is 0.2 dB/km, that the Raman distributed amplifier provides 30 dB of on/off gain with an equivalent noise figure of −3 dB, and that the pre-amplifier noise figure is 5 dB, the equivalent total noise figure of the unrepeatered system is −10 dB.

List of Units and Nomenclature

A_{eff}	effective area
ASE	amplified spontaneous emission
B_{eff}	effective erbium doping area
D	normalized medium inversion
dB	ratio in decibel
dBm	power in decibel
DBR	distributed Bragg reflector
DCM	dispersion compensating modules
Demux	optical de-multiplexers
DFB	distributed feedback
DSF	dispersion-shifted fiber
EDF	erbium-doped fiber
EDFA	erbium-doped fiber amplifiers
FBG	fiber Bragg gratings
g_R	Raman gain coefficient
G	gain
$G_{On/Off}$	on/off gain
h	Planck's constant
Length	meter (m)
MPI	multi-path interference
Mux	optical multiplexers
n_{eq}	equivalent input noise factor
NF	noise figure of the amplifier
NRZ	non-return to zero
NZDSF	non-zero dispersion shifter fiber
N_2/N_1	medium inversion
OSNR	optical signal-to-noise ratio
Power	watt (W)
SSMF	standard single-mode fiber
TDM	time-division multiplexing
Time	second (s)
WDM	wavelength-division multiplexing
τ	lifetime

See also

Coherent Lightwave Systems. **Dispersion Management**. **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics). **Optical Amplifiers:** Basic Concepts; Erbium Doped Fiber Amplifiers for Lightwave Systems; Raman, Brillouin and Parametric Amplifiers.

Optical Communication Systems: Architectures of Optical Fiber Communication Systems; Basic Concepts; Historical Development; Lightwave Transmitters; Optical Time Division Multiplexing; Wavelength Division Multiplexing. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal G (2001) *Nonlinear Fiber Optics*. London: Academic Press.

Chesnoy J (2002) *Undersea Fiber Communication Systems*. London: Academic Press.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers, Principles and Applications*. New York: John Wiley & Sons.

Desurvire E, Bayart D, Desthieux B and Bigo S (2002) *Erbium-Doped Fiber Amplifiers, Device and System Developments*. New York: John Wiley & Sons.

Emori Y, Kado S and Namiki S (2002) Broadband flat-gain and low-noise Raman amplifiers pumped by wavelength-multiplexed high-power laser diodes. *Optical Fiber Technology* 8: 107.

Hansen PB and Eskildsen L (1997) Remote amplification in repeaterless transmission systems. *Optical Fiber Technology* 3: 221.

Islam MN (2004) *Raman Amplifiers for Telecommunications*. New York: Springer-Verlag.

Kaminow I and Li T (2002) *Optical Fiber Telecommunications IV*. London: Academic Press.

Keiser G (2000) *Optical Fiber Communications*. New York: McGraw-Hill.

Raman, Brillouin and Parametric Amplifiers

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Glass fibers for optical communications are made of fused silica, an amorphous material, to which dopant materials of various kinds can be added to produce changes in refractive index. A number of third-order nonlinear processes can occur; these can grow to appreciable magnitudes over the long lengths available in fibers, even though the nonlinear coefficients in the materials are relatively small. The effects are particularly important in single-mode fibers, in which the small mode field dimensions result in substantially high light intensities with relatively modest input powers.

Nonlinear optical effects, such as stimulated Raman scattering (SRS), stimulated Brillouin scattering (SBS), and four-wave mixing (FWM), can cause some problems such as nonlinear loss, pulse distortion, and cross-talk between signals in wavelength division multiplexed (WDM) transmission systems. However, these effects are also useful for efficient amplification and frequency conversion of injected signals, as well as for generation of new frequencies. In this chapter the main properties of fiber Raman amplifiers (FRAs), fiber Brillouin amplifiers (FBAs), and fiber parametric amplifiers (FPAs) will be reviewed, regarding their applications in the field of fiber optic communications.

Fiber Raman Amplifiers

Stimulated Raman scattering is likely to occur in fibers, producing appreciable amplification for down-shifted signals when moderate pump power levels are used. This is largely because the process is phase-matched, so that the long interaction lengths available in fibers will enhance the process. Additionally, the small mode field cross-section yields high intensities for relatively modest power levels.

In the SRS process one has simultaneously the absorption of a photon from the pump beam at frequency ω_p and the emission of a photon at the signal (Stokes) frequency, stimulated by the incident signal beam at frequency ω_s (Figure 1a). The difference in energy is taken up by a high energy

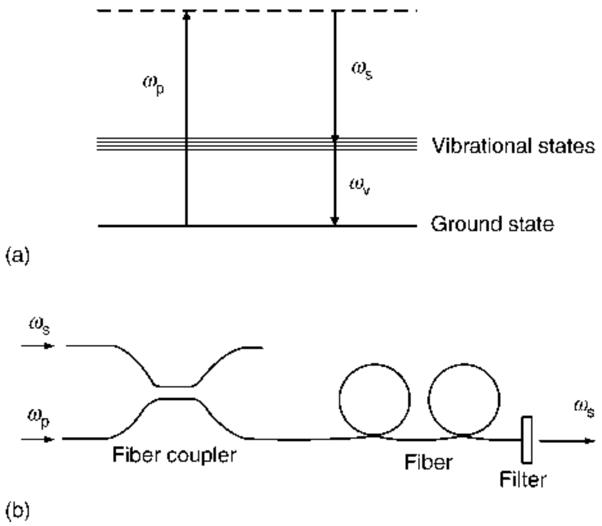

Figure 1 (a) Energy-level illustration of the process of stimulated Raman scattering and (b) schematic diagram of a fiber Raman amplifier.

phonon (molecular vibration) at frequency ω_v. Thus, SRS provides for energy gain at the signal frequency at the expense of the pump.

Figure 1b shows schematically a fiber Raman amplifier. The pump and the signal beams are injected into the fiber through a WDM fiber coupler. The case illustrated in Figure 1a shows the two beams co-propagating inside the fiber, but the counter-propagating configuration is also possible. In fact, the SRS can occur in both directions, forwards and backwards. It was confirmed experimentally that the Raman gain is almost the same in both cases.

Raman Gain and Bandwidth

Figure 2 shows the Raman gain coefficient for fused silica as derived from the spontaneous Raman spectrum. At the peak, the Raman gain coefficient for silica fibers is 9.4×10^{-14} m W^{-1} for a pumping wavelength $\lambda_p = 1.0\ \mu$m and varies as λ_p^{-1}. At this wavelength, a shift of 15 THz corresponds to 50 nm. The most significant feature of the Raman gain in silica fibers is that $\bar{g}_R$ extends over a large frequency range (up to 40 THz) with a broad dominant peak near 13 THz. This behavior is due to the amorphous nature of silica glass, whose molecular vibrational energy levels merge together to form a band.

The use of oxide glasses as dopants for fiber fabrication is suitable to enhance the Raman gain coefficient. For example, in the case of pure GeO_2, the Raman gain coefficient is about 9.2 times that for pure SiO_2 glass. In fact, doping the core of an SiO_2 Raman fiber amplifier with GeO_2 has two beneficial effects. First, the index difference Δ is raised, which increases the effective waveguiding and in turn reduces the effective cross-sectional area A_e. As a consequence, the rate of SRS is increased. Secondly, the Raman gain coefficient increases above that of pure SiO_2 in proportion to the concentration GeO_2, and is given by

$$g_R = (1 + 80\Delta)\bar{g}_R \qquad [1]$$

where $\bar{g}_R$ is the Raman coefficient for pure SiO_2.

In spite of the positive aspects mentioned above, doping with GeO_2 also increases the fiber losses, due to the concomitant rise in Rayleigh scatter and in the number of dopant-dependent loss centers. Thus, greatly increasing the GeO_2 concentration will not be worthwhile if the increased fiber losses outweigh the improvement in amplifier gain.

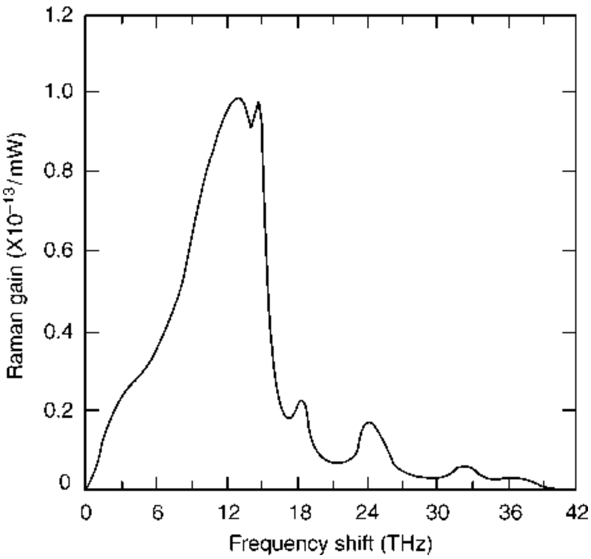

Figure 2 Measured Raman gain spectrum for fused silica at a pump wavelength $\lambda_p = 1.0\ \mu$m. The Raman gain scales inversely with λ_p. Reproduced with permission from Stolen RH and Ippen EP (1973) *Applied Physics Letters* 22: 276. © 1973 American Institute of Physics.

Gain Saturation

The interaction between the pump and Stokes waves is governed by the following coupled equations:

$$\frac{dP_s}{dz} = -\alpha_s P_s + \frac{g_R}{A_e} P_p P_s \qquad [2]$$

$$\frac{dP_p}{dz} = -\alpha_p P_p - \frac{\omega_p}{\omega_s}\frac{g_R}{A_e} P_s P_p \qquad [3]$$

where A_e is the effective cross-sectional area of the fiber mode, α_s and α_p are absorption coefficients which account for the fiber loss at the signal and pump frequencies, respectively, and the signal wave is considered to be co-propagating with the pump wave. Assuming that $\alpha_s = \alpha_p = \alpha$ (which is a reasonable approximation around the 1.5 μm wavelength region in a low loss fiber), the following approximate solutions of eqns [2] and [3] can be written when $P_p(0) \gg P_s(0)$:

$$P_p(z) = \frac{P_p(0)\exp(-\alpha z)}{1 + F(z)} \qquad [4]$$

$$P_s(z) = \frac{P_s(0)\exp(\Gamma(1 - e^{-\alpha z}) - \alpha z)}{1 + F(z)} \qquad [5]$$

where

$$\Gamma = \frac{g_R P_p(0)}{\alpha A_e} \qquad [6]$$

and

$$F(z) = \frac{\omega_p P_s(0)}{\omega_s P_p(0)} \exp(\Gamma(1 - e^{-\alpha z})) \qquad [7]$$

The dimensionless parameter $F(z)$ accounts for the effects of pump depletion and the attendant saturation of the gain seen by the signal.

Figure 3 shows the output signal power $P_s(L)$ versus the input signal power $P_s(0)$ for a Raman amplifier with length $L = 8$ km and several values of the input pump power. Typical values $g_R = 6.7 \times 10^{-14}\ \text{m W}^{-1}$, $\alpha = 0.2\ \text{dB km}^{-1}$, $A_e = 30 \times 10^{-12}\ \text{m}^2$, $\lambda_p = 1.46\ \mu\text{m}$ and $\lambda_s = 1.55\ \mu\text{m}$ were assumed. The transfer characteristics are linear for $P_p(0) = 0.1$, 0.2, and 0.4 W, but the effects of pump depletion become discernible for input signal levels $P_s(0) > -15$ dB m when $P_p(0) = 0.6$ W. For a pump power $P_p(0) = 0.8$ W the linear behavior of the Raman amplifier cannot be observed, even for signal power levels as low as -25 dB m.

Since, in the absence of Raman amplification, the signal power at the amplifier output would be $P_s(L) = P_s(0)\exp(-\alpha L)$, the amplifier gain is given by

$$G_R = \frac{P_s(L)}{P_s(0)\exp(-\alpha L)} = \frac{\exp(g_R P_p(0) L_e / A_e)}{1 + F(L)} \qquad [8]$$

where

$$L_e = \frac{1}{\alpha}[1 - \exp(-\alpha L)] \qquad [9]$$

is the effective interaction length and L is the actual fiber length. The gain given by eqn [8] is seen to be a function of the input signal power (i.e., a saturation nonlinearity) through the term $F(L)$.

When $F(L) \gg 1$ we have from eqn [5], that

$$P_s(L) = \frac{\omega_s}{\omega_p} P_p(0)\exp(-\alpha L) \qquad [10]$$

In this case, the amplifier output reaches the pump level irrespective of the input signal level. This implies that any spontaneous Raman scatter in the fiber will be amplified up to power levels comparable to that of the pump. This is obviously inadmissible in an amplifier application, thus this operating regime must be avoided.

On the other hand, when $F(L) \ll 1$ we have, from eqn [5], that

$$P_s(L) = P_s(0)\exp(g_R L_e P_p(0)/A_e - \alpha L) \qquad [11]$$

In this case, the amplifier gain is

$$G_R = \frac{P_s(L)}{P_s(0)\exp(-\alpha L)} = \exp(g_R L_e P_0 / A_e) \qquad [12]$$

From eqn [12], the Raman gain in decibels is expected to increase linearly with the pump input power. Raman gains were experimentally confirmed to be almost the same for forwards and backwards pumping.

Figure 4 illustrates the variation of the amplifier gain (full curves, left scale) and the saturation parameter $F(L)$ (dashed curves, right scale) with the pump power $P_p(0)$ for a fixed input signal power $P_s(0) = 0$ dBm and amplifier lengths $L = 2$, 4, and 8 km. The full curve corresponding to $L = 2$ km shows a linear dependence of the gain on the pump power. However, the other two full curves show how the pump depletion limits the achievable gain. The black point on these curves marks the value of the pump power at which the depletion parameter F

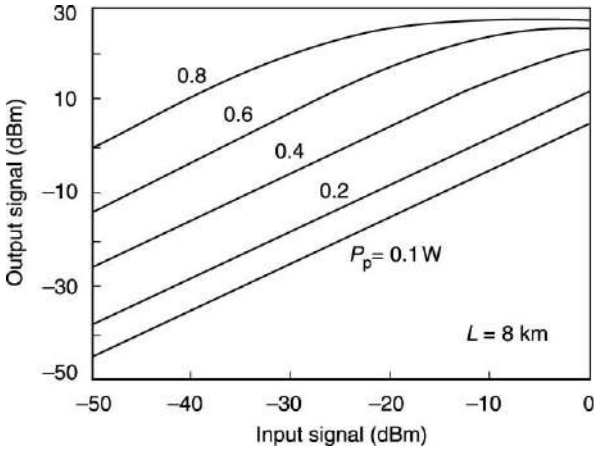

Figure 3 Transfer characteristics for a Raman amplifier with a length $L = 8$ km and several values of the input pump power.

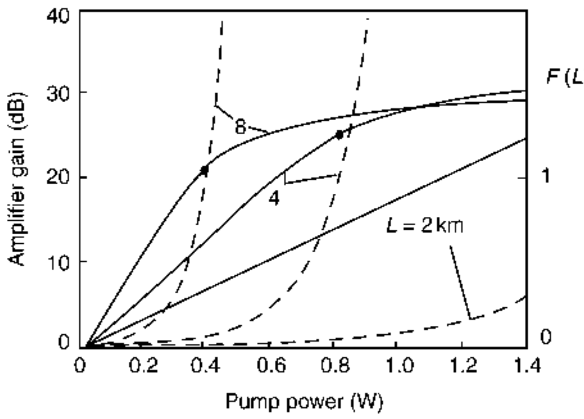

Figure 4 Raman amplifier gain (full curves, left scale) and parameter $F(L)$ (dashed curves, right scale) against input pump power for three values of the amplifier length.

becomes equal to unity. Only operating at pump powers sufficiently below each point ensures linear operation of the amplifier, whilst choosing to operate the amplifier significantly above that point will mean $F \gg 1$ and spontaneous scatter will be amplified to intolerable levels.

Noise Light Power

The dominant noise light in fiber Raman amplification is due to the amplified spontaneously Raman scattered light. In fact, a part of the pump energy is spontaneously converted into Stokes radiation extending over the entire bandwidth of the Raman gain spectrum and is amplified together with the signal. The output thus consists not only of the desired signal but also of background noise extending over a wide frequency range (~10 THz or more).

It has been shown that, for both forwards and backwards pumpings, the noise light power is equivalent to a hypothetical injection of a single-photon per unit frequency at the fiber input end for forwards pumping, and at some distance away from the fiber output for backwards pumping. Assuming that $G_R \gg 1$ and $\alpha_L \gg 1$, the noise light powers for forwards pumping, $P_{\text{fasp}}(L)$, and backwards pumping, $P_{\text{basp}}(L)$, are approximately given by

$$P_{\text{fasp}}(L) \approx h\nu_s \Delta\nu_R (G_R - 1)\exp(-\alpha L) \qquad [13]$$

$$P_{\text{basp}}(L) \approx h\nu_s \Delta\nu_R (G_R - 1)/\ln G_R \qquad [14]$$

where $\Delta\nu_R$ is the Raman gain bandwidth, h is Planck's constant, and ν_s is the signal frequency. We can observe from eqn [13] that the noise light power in the case of forward pumping decreases as the fiber length is increased. However, in the case of backwards pumping, it depends mainly on the Raman gain, being nearly independent of fiber length and loss.

FRA Performance and Applications

One main limitation affecting the operation of fiber Raman amplifiers is the requirement of a relatively high-power laser as a pump source. The experiments near 1.55 μm are usually carried out using color-center lasers for this purpose. However, such lasers are too bulky for communication applications and the use of compact high-power laser diodes (LDs) becomes desirable. This objective has recently been achieved with the advent of high-power multimode LDs. In some cases, the output lights from various LDs are combined to attain the necessary pump power.

In a Raman amplifier, SRS has to compete frequently with other nonlinear effects. For example, since the intrinsic gain coefficient for SBS is two orders of magnitude larger than that for SRS, in some cases the SBS may occur at lower pump powers and may affect significantly the SRS process. In particular, the Raman gain may not only be reduced but also become unstable. However, the SBS can be suppressed taking into account its narrow linewidth, which is typically less than 100 MHz. Therefore, if a pump laser with linewidth broader than the Brillouin bandwidth is employed, the SBS effect would not be significant. When a semiconductor laser is used as pump source, its spectral linewidth can be artificially increased by direct frequency modulation, resulting in effective SBS suppression.

Considering its broad bandwidth (>5 THz), fiber Raman amplifiers can be used to amplify several channels simultaneously in a multichannel communication system. The same characteristic also makes such amplifiers suitable for amplification of short optical pulses.

Considering a repeaterless transmission system, the FRA can be used in three different configurations: forwards pumping, backwards pumping, and bi-directional pumping. In the first case, the signal amplification occurs near the fiber input, where the signal power is relatively high. This configuration for the Raman amplification does not generally affect the transmission quality and can compensate for the transmission loss by approximately the value of the Raman gain. For example, 10 dB Raman gain provides an increase in transmission distance of 50 km when the fiber attenuation is 0.2 dB km^{-1}. In the case of backwards pumping, the signal amplification occurs near the fiber output end before photodetection, where the signal power may be very weak, comparable with the Raman noise power. This scheme for the Raman amplification can considerably reduce the minimum detectable signal power, which results in a corresponding increase of the transmission distance. However, the transmission quality is affected due to backwards noise light. The bi-directional pumping configuration uses both forwards and backwards pumping simultaneously and its performance can be estimated from those for the respective unidirectional schemes considered above.

The FRAs can also be used in nonregenerative all-optical multirepeater systems. In this case, the transmission loss is compensated for by both forwards and backwards Raman gain providing the periodically injected pump light. This scheme can be used to overcome the fiber loss in soliton-based communication systems. In this case, the pulse broadening due to chromatic dispersion of the fiber is also eliminated, which provides the possibility of achieving ultralong distance distortionless transmission.

Fiber Brillouin Amplifiers

The process of SBS can be described as a classical three-wave interaction involving the incident (pump) wave of frequency ω_p, the Stokes wave of frequency ω_s, and an acoustic wave of frequency ω_a. The pump creates a pressure wave in the medium owing to electrostriction, which in turn causes a periodic modulation of the refractive index. This process is illustrated schematically in Figure 5. Since the acoustic wavefronts are moving away from the incident pump wave, the scattered light is shifted downward in frequency to the Stokes frequency:

$$\omega_s = \omega_p - \omega_a \qquad [15]$$

The response of the material to the interference of the pump and Stokes fields tends to increase the amplitude of the acoustic wave. Therefore, the beating of the pump wave with the acoustic wave tends to reinforce the Stokes wave, whereas the beating of the pump wave and the Stokes waves tends to reinforce the acoustic wave. This explains the appearance of the stimulated Brillouin scattering process.

In spite of the apparent similarity between SBS and SRS, these processes differ in three important aspects: i) Brillouin amplification occurs only when the pump and signal beams counter-propagate inside the fiber; ii) the Stokes shift for SBS is smaller by three orders of magnitude compared with that of SRS; and iii) the Brillouin gain spectrum is extremely narrow, with a bandwidth <100 MHz. These characteristics significantly affect the performance and the fiber Brillouin amplifiers.

Brillouin Gain and Bandwidth

Physically, each pump photon in the SBS process gives up its energy to create simultaneously a Stokes photon and an acoustic phonon. Since both the energy and the momentum must be conserved in this process, the frequencies of the three waves must satisfy eqn [15], whereas the wave vectors are related by

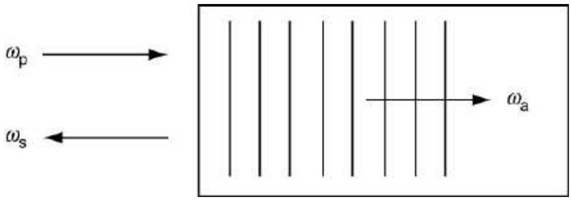

Figure 5 Schematic illustration of the stimulated Brillouin scattering process.

$$\vec{k}_p = \vec{k}_s + \vec{k}_a \qquad [16]$$

where $\vec{k}_p$ and $\vec{k}_s$ are the wave vectors of the pump and signal beams, respectively. The frequency ω_a and the wave vector $\vec{k}_a$ of the acoustic wave satisfy the dispersion relation:

$$\omega_a = |\vec{k}_a| v_a = 2v_a |\vec{k}_p| \sin(\theta/2) \qquad [17]$$

where v_a is the acoustic velocity, θ is the angle between the pump and signal waves, and $|\vec{k}_p| \approx |\vec{k}_s|$ was used. Equation [17] shows that ω_a vanishes in the forward direction ($\theta = 0$), whereas it is maximum in the backward direction ($\theta = \pi$). The frequency shift ν_B in the backward direction is given by

$$\nu_B = \frac{\omega_a}{2\pi} = \frac{2nv_a}{\lambda_p} \qquad [18]$$

where eqn [17] was used with $|\vec{k}_p| = 2\pi n/\lambda_p$, λ_p is the pump wavelength, and n is the refractive index. Considering $n = 1.45$ and $v_a = 5.96\ \mathrm{km\ s^{-1}}$ as typical values for silica fibers, we obtain $\nu_B = 11.1$ GHz at $\lambda_p = 1.55\ \mu\mathrm{m}$.

Assuming a monochromatic pump, the spectrum of the scattered light depends on the uniformity of the fiber parameters. For a uniform medium the spectrum is Lorentzian with a width $\Delta\nu_B$ determined by the acoustic attenuation. For 1.5 μm light the width is about 17 MHz. However, both the frequency shift ν_B and the gain bandwidth $\Delta\nu_B$ can vary from fiber to fiber because of the guided nature of light and the presence of dopants in the fiber core. In particular, the inhomogeneities in the core section along the length determine an increase of the amplifier bandwidth, which can exceed 100 MHz, although typical values are 50–60 MHz for $\lambda_p \approx 1.55\ \mu\mathrm{m}$.

The peak value of the Brillouin gain coefficient is given by

$$g_B(\nu_B) = \frac{2\pi n^7 p_{12}^2 \gamma}{c\lambda_p^2 \rho_0 v_a \Delta\nu_B} \qquad [19]$$

where p_{12} is the longitudinal elasto-optic coefficient, ρ_0 the average material density, n the refractive index, c the vacuum speed of light, and γ is a numerical factor that depends on the polarization properties of the optical fiber, which assumes the value $\frac{1}{2}$ for long lengths of nonpolarization-preserving optical fiber. By using the parameter values typical of fused silica, g_B is estimated to be about $2.5 \times 10^{-11}\ \mathrm{m\ W^{-1}}$, which is two orders of magnitude larger than the Raman gain coefficient at $\lambda_p = 1.55\ \mu\mathrm{m}$.

The expression for the Brillouin gain in eqn [19] assumes that the spectral width of the pump beam ($\Delta\nu_p$) is much narrower that the Brillouin linewidth ($\Delta\nu_B$). When such condition is not verified, the Brillouin gain is reduced according with the relation:

$$\tilde{g}_B(\nu_B) = \frac{\Delta\nu_B}{\Delta\nu_B + \Delta\nu_p} g_B(\nu_B) \qquad [20]$$

For this reason, fiber Brillouin amplifiers need to be pumped by narrow-linewidth semiconductor lasers ($\Delta\nu_p < 10$ MHz).

Gain Saturation

The slowly varying steady-state Fourier amplitudes of the Stokes (E_s) and pump (E_p) electric fields are related by the following coupled-wave equations:

$$\frac{\partial E_s}{\partial z} = \left[j\frac{K_1 K_2 |E_p|^2}{\delta - j\Gamma} + \frac{\alpha}{2} \right] E_s \qquad [21]$$

$$\frac{\partial E_p}{\partial z} = -\left[j\frac{K_1 K_2 |E_s|^2}{\delta + j\Gamma} + \frac{\alpha}{2} \right] E_p \qquad [22]$$

where Γ^{-1} is the acoustic phonon lifetime resulting in a spontaneous Brillouin scattering linewidth $\Delta\nu_B = \Gamma/\pi$, $\alpha = \alpha_s \approx \alpha_p$ is the absorption coefficient, and the coupling constants are

$$K_1 = \frac{K_2 \rho_0 n^2 \varepsilon_0}{2\nu_a}, \qquad K_2 = \frac{n^3 p_{12} \omega_s (\hat{e}_p \times \hat{e}_s)}{2c\rho_0} \qquad [23]$$

ε_0 being the free-space permittivity and $\hat{e}_p$ and $\hat{e}_s$ the unit vectors in the directions of the pump and Stokes optical fields, respectively. In eqns [21] and [22] it is assumed that the pump wave is launched at $z = 0$ and propagates in the $+z$ direction, whereas the signal wave is launched at $z = L$ and propagates in the $-z$ direction. The parameter $\delta \equiv \omega_p - \omega_s - \omega_a$ takes into account a possible detuning of the signal from the SBS gain line center.

Equations [21] and [22] can be written in terms of optical power as follows:

$$\frac{dP_s}{dz} = \alpha P_s - \frac{g_B(\delta)}{A_e} P_p P_s \qquad [24]$$

$$\frac{dP_p}{dz} = -\alpha P_p - \frac{g_B(\delta)}{A_e} P_s P_p \qquad [25]$$

where A_e is the effective core area and $g_B(\delta)$ is the Brillouin gain coefficient, given by

$$g_B(\delta) = \frac{\Gamma^2}{\delta^2 + \Gamma^2} g_B(\nu_B) \qquad [26]$$

where $g_B(\nu_B)$ is given by eqn [19].

Considering that $a \ll 1$, the following approximate solutions of eqns [24] and [25] can be written:

$$P_s(z) = P_s(0) D(z) e^{\alpha z} \qquad [27]$$

$$P_p(z) = P_p(0) D(z) H(z) e^{-\alpha z} \qquad [28]$$

where

$$D(z) = \frac{[P_p(0) - P_s(0)]}{[P_p(0)H(z) - P_s(0)]} \qquad [29]$$

and

$$H(z) = \exp\left\{ g_B [P_p(0) - P_s(0)][1 - \exp(-\alpha z)]/(A_e \alpha) \right\} \qquad [30]$$

The Brillouin amplifier gain is given from eqn [27] as

$$G_B = \frac{P_s(0)}{P_s(L) e^{-\alpha L}} = \frac{1}{D(L)} \qquad [31]$$

Figure 6 shows the output signal power $P_s(0)$ versus the input signal power $P_s(L)$ for a Brillouin amplifier with length $L = 8$ km and several values of the input pump power. Typical values $g_B(0) = 2.5 \times 10^{-11}$ m W^{-1}, $\alpha = 0.2$ dB km^{-1}, $A_e = 30 \times 10^{-12}$ m^2 were assumed. The transfer characteristics are linear for $P_p(0) = 0.2$ mW, but the effects of pump depletion become discernible

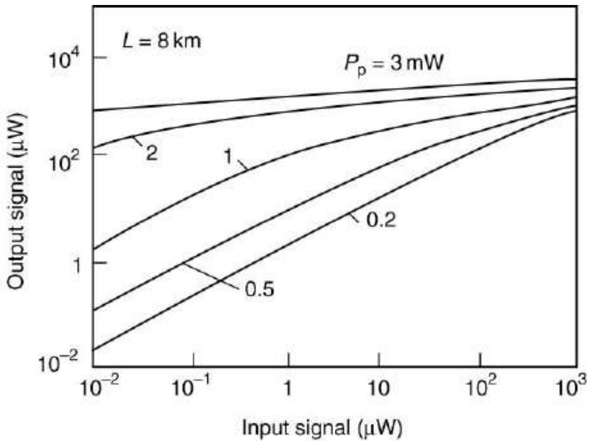

Figure 6 Transfer characteristics for a Brillouin amplifier with a length $L = 8$ km and several values of the input pump power.

for input signal levels $P_s(0) > 2\ \mu\mathrm{W}$ when $P_p(0) = 1$ mW. For a pump power $P_p(0) = 2$ mW, linear behavior of the Brillouin amplifier cannot be observed even for signal powers as low as 10 nW.

The dependence of the Brillouin amplifier gain on the input pump power is illustrated in Figure 7 for several amounts of the signal detuning and two values of the signal power. The fiber length is $L = 5$ km and the gain bandwidth is $\Delta\nu_B = 30$ MHz, whereas the other parameter values are equal to those assumed in Figure 6. The saturated gain shows a significant dependence on the input signal power. An input signal of 1 nW, which is comparable to the original value of spontaneous emission, can be amplified by about 60 dB for a pump power $P_p(0) \approx 5$ mW. However, the amplifier gain can be considerably reduced when the signal is detuned from the gain peak. For example, for an input signal power $P_s(L) = 1$ nW and a pump power $P_p(0) \approx 3$ mW, a detuning of 15 MHz (which corresponds to a reduction of the gain coefficient to half of its maximum value) determines a reduction of about 25 dB on the amplifier gain.

Substituting eqn [28] for $P_p(z)$ in eqn [21], the following result is obtained for the Stokes field at the amplifier output:

$$E_s(0) = E_s(L)\sqrt{G_B}\exp\{-\alpha L/2 - j\phi\} \qquad [32]$$

where

$$\phi = \frac{\delta}{2\Gamma}\ln(G_B) \qquad [33]$$

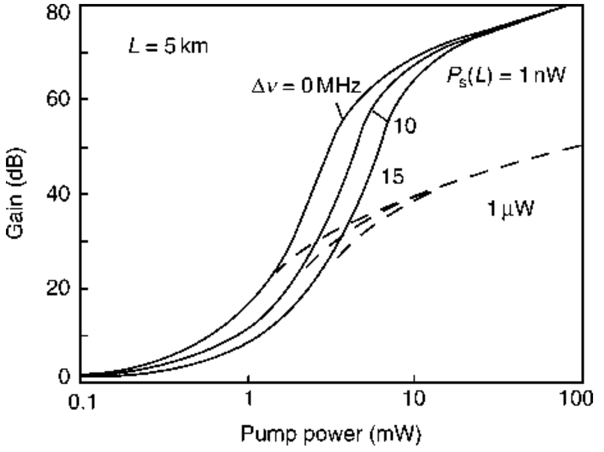

Figure 7 Brillouin amplifier gain against the input pump power for an amplifier length $L = 5$ km and different amounts of detuning from the gain center. The full curves and the dashed curves correspond to input signal powers $P_s(L)$= 1 nW and $P_s(L)$= 1 μW, respectively. Reproduced with permission from Ferreira MFS, *et al.* (1994) *Optical and Quantum Electronics* 26: 35, with permission from Kluwer Academic Publisher.

represents a nonlinear phase change, which depends on the SBS gain, input signal power, and frequency detuning. This nonlinear phase shift can impose a stringent limit on the pump- and signal-frequency stability when FBAs are used in some phase-sensitive detection schemes.

Noise Light Power

To describe the spontaneous emission noise we must add a spontaneous emission term to the equations describing the evolution of the Stokes and pump waves. The amplified spontaneous scattered power per unit frequency is then obtained in the form:

$$P_{\mathrm{asp}}(\nu, z) = (h\nu/A_e)g_B(\delta)(N+1)G(\nu, z) \times \int_z^L P_p(z')[G(\nu, z')]^{-1}\mathrm{d}z' \qquad [34]$$

where

$$G(\nu, z) = \exp\left\{\int_z^L [P_p(z)g_B(\delta)/A_e - \alpha]\mathrm{d}z\right\} \qquad [35]$$

is the gain function, h is Planck's constant, and N is the thermal equilibrium number of acoustic phonons, given by

$$N = \frac{1}{\exp(h\nu_B/kT)^{-1}} \approx kT/h\nu_B \approx 500 \qquad [36]$$

where k is the Boltzmann constant and T is the absolute temperature. The total amplified signal and spontaneous powers are given by integration of $P_s(\nu, z)$ over the gain profile.

Similarly to the behavior of the gain, the noise power depends significantly on the signal magnitude in the saturation regime and becomes particularly high for low signal powers. This high noise level, which is about 20 dB above that of an ideal amplifier, imposes some limitations on the use of FBA as a receiver preamplifier.

FBA Performance and Applications

Fiber Brillouin amplifiers can provide high gains at low pump powers. However, the gain bandwidth is small and the amplified spontaneous emission noise is substantially larger than that observed with other amplifiers, leading to a noise figure that is quite large (>15 dB). This fact limits their usefulness as a preamplifier. Use as in-line amplifiers at high signal levels is possible in some configurations. Some problems in this application are concerned with the narrow bandwidth and the small saturation power of the FBAs. In particular, the Brillouin linewidth strictly limits the bandwidth of data signals that can be amplified in a FBA.

The Brillouin gain bandwidth can be intentionally extended by more than one order of magnitude by applying frequency modulation to the pump laser. Of course, the bandwidth enhancement is accompanied by a reduction of the peak gain. Consequently, a higher pump power will be necessary to achieve the same gain.

The narrow bandwidth of the FBAs can be used to advantage in coherent lightwave systems to amplify selectively the carrier of a phase- or amplitude-modulated wave, thus allowing homodyne detection, using the amplified carrier as local oscillator. The scheme should work well at bit rates $>100\ \text{Mb s}^{-1}$ because the modulation sidebands then fall outside the amplifier bandwidth, and the optical carrier can be amplified selectively. However, an inevitable accompaniment of this narrow-band amplification is the nonlinear phase shift induced by the pump on the signal, given by eqn [33]. This phase shift imposes the most stringent limit on pump- and signal-frequency stability when FBAs are used for self-homodyne detection. In the case of amplitude-shift keyed (ASK) signals, for example, a phase stability of about 0.1 rad for the amplified signal carrier may be required, which corresponds to a pump-signal frequency offset $<100\ \text{kHz}$. The nonlinear phase shift can be used with advantage in some self-homodyne schemes which require some specific adjustment of the carrier phase. In the case of a self-homodyne coherent receiver for phase-shift keyed (PSK) signals, for example, a quadrature phase correction is required. Typically, this requires a detuning of only a few MHz from the SBS gain peak, well within the Brillouin gain bandwidth.

Another application of the narrow linewidth associated with the Brillouin gain, is as a tunable narrow-band filter for channel selection in a densely packed multichannel communication system. A channel can be selectively amplified through Brillouin amplification by launching a pump beam at the receiver end so that it propagates inside the fiber in a direction opposite to that of the multichannel signal. Different channels can be selectively amplified by tuning the pump laser. The adjustable bandwidth and high out-of-band rejection can be used to advantage in this case. In another approach, the fiber Brillouin amplifier can be used to selectively amplify multiple subcarrier-multiplexed (SCM) channels carried by a single optical carrier within a WDM signal, or to amplify a backward-propagating supervisory signal.

Fiber Parametric Amplifiers

In both stimulated Raman scattering and stimulated Brillouin scattering, the optical fiber plays an active role through the participation of molecular vibrations or acoustic phonons. However, in many other nonlinear phenomena, the fiber plays a passive role except for mediating the interaction among several optical waves. Since these phenomena involve the light-induced modulation of some medium parameter, they are known as parametric processes.

The parametric processes can be classified as second-order or third-order processes, depending on whether the second-order susceptibility or the third-order susceptibility is responsible for them. Concerning the second-order parametric processes, they are not significant in silica fibers, due to the small value of the second-order susceptibility. In contrast, some third-order parametric processes, namely third-harmonic generation, four-wave mixing, and parametric amplification, can exhibit relatively high conversion efficiencies. The parametric gain in optical fibers can be used to make fiber parametric amplifiers, which find some interesting applications for squeezing and optical fiber communications.

Phase-Matching in Parametric Processes

The efficiency of the four-wave mixing process depends on the relative phase among the interacting optical waves. A high efficiency requires matching of the frequencies as well as of the wave vectors. The latter requirement is often referred to as phase-matching. In contrast with SRS and SBS, where phase-matching is automatically fulfilled as a result of the active participation of the nonlinear medium, the phase-matching condition in parametric processes requires a specific choice of the frequencies and the refractive indices.

Considering the case in which two photons at energies ω_1 and ω_2 are annihilated with simultaneous creation of two photons at frequencies ω_3 and ω_4, the matching of the frequencies imposes the condition:

$$\omega_1 + \omega_2 = \omega_3 + \omega_4 \qquad [37]$$

On the other hand, the phase-matching is given by the condition $\Delta k = 0$, where

$$\begin{aligned}\Delta k &= k_3 + k_4 - k_1 - k_2 \\ &= (\tilde{n}_3\omega_3 + \tilde{n}_4\omega_4 - \tilde{n}_1\omega_1 - \tilde{n}_2\omega_2)/c\end{aligned} \qquad [38]$$

where the $\tilde{n}_i$ are the effective refractive indices of the fiber modes. It is relatively easy to satisfy the phase-matching condition when $\omega_1 = \omega_2$. In this case, a strong pump at $\omega_1 = \omega_2 \equiv \omega_\mathrm{p}$ generates a low-frequency sideband at ω_3 and a high-frequency sideband at ω_4, when we assume $\omega_4 > \omega_3$. These sidebands are referred as the Stokes and anti-Stokes bands, respectively, which are also often called the signal and idler bands. The frequency shift of the two sidebands is given by

$$\Omega_s = \omega_p - \omega_3 = \omega_4 - \omega_p \qquad [39]$$

The effective indices $\tilde{n}_i$ in eqn [38] can be written as

$$\tilde{n}_i = n_i + \Delta n_i \qquad [40]$$

where Δn_i is the change of index refraction due to waveguiding. As a consequence, considering the eqn [40] and the case $\omega_1 = \omega_2$, eqn [38] can be written in the form:

$$\Delta k = \Delta k_m + \Delta k_w \qquad [41]$$

where

$$\Delta k_m = (n_3\omega_3 + n_4\omega_4 - 2n_1\omega_1)/c \qquad [42]$$

results from the material dispersion and

$$\Delta k_w = (\Delta n_3\omega_3 + \Delta n_4\omega_4 - (\Delta n_1 + \Delta n_2)\omega_1)/c \qquad [43]$$

results from the waveguide dispersion. We must note that although the two pump waves have the same frequency, Δn_1 and Δn_2 are, in general, different if these waves propagate in different modes.

The material contribution Δk_m can be expressed in terms of the frequency shift Ω_s given by eqn [39], if we use an expansion of Δk_m about the pump frequency $\omega_p = \omega_1 = \omega_2$. By retaining up to terms quadratic in Ω_s in this expansion, one obtains

$$\Delta k_m = \beta_2 \Omega_s^2 \qquad [44]$$

where $\beta_2 = d^2k/d\omega^2$ is the group-velocity dispersion (GVD) coefficient at the pump frequency. In the description of the evolution of the Stokes and anti-Stokes waves, an effective propagation constant appears, given by

$$\kappa = \Delta k + \Delta k_{NL} = \Delta k_m + \Delta k_w + \Delta k_{NL} \qquad [45]$$

In eqn [45], Δk_{NL} represents a nonlinear contribution, given by

$$\Delta k_{NL} = 2\gamma P_p \qquad [46]$$

where $P_p = P_1 = P_2$ is the power in each pump wave and

$$\gamma = \frac{\bar{n}_2 \omega}{c A_e} \qquad [47]$$

is the nonlinearity coefficient, $\bar{n}_2$ being the nonlinear-index coefficient.

Parametric Gain

The evolution of the Stokes and anti-Stokes waves are governed by the following coupled equations:

$$\frac{dE_3}{dz} = 2j\gamma P_p \exp(-j\kappa z) E_4^* \qquad [48]$$

$$\frac{dE_4^*}{dz} = -2j\gamma P_p \exp(j\kappa z) E_3 \qquad [49]$$

The general solution of eqns [48] and [49] is given by

$$E_3(z) = (c_1 \exp(g_P z) + c_2 \exp(-g_P z)) \exp(-j\kappa z/2) \qquad [50]$$

$$E_4^*(z) = (c_3 \exp(g_P z) + c_4 \exp(-g_P z)) \exp(j\kappa z/2) \qquad [51]$$

where the coefficients c_i are determined from the boundary conditions and the parametric gain g_P is given by

$$g_P = \sqrt{(2\gamma P_p)^2 - \left(\frac{\kappa}{2}\right)^2} \qquad [52]$$

In the range

$$-6\gamma P_p \le \Delta k \le 2\gamma P_p \qquad [53]$$

the radicand in eqn [52] is positive, so a gain occurs. The maximum gain is achieved when $\kappa = 0$ and it is given by

$$g_{P\max} = \Delta k_{NL} = 2\gamma P_p = \bar{g}_P \frac{2P_p}{A_e} \qquad [54]$$

where

$$\bar{g}_P = \frac{\omega_p \bar{n}_2}{c} \qquad [55]$$

Considering that $\bar{n}_2 = 3.2 \times 10^{-20}\ m^2\ W^{-1}$ and assuming a pump wavelength $\lambda_p = 1\ \mu m$, we find that the gain coefficient $\bar{g}_P$ is larger by a factor of 2 than the peak of the Raman gain coefficient g_R. Consequently, the threshold for the parametric process is lower than with the Raman effect. However, the Raman effect dominates in general for long fibers, since it is difficult to maintain phase matching because of fluctuations of the core diameter. One can define a coherence length, given by

$$L_{coh} = 2\pi/\kappa_{max} \qquad [56]$$

where κ_{max} is the maximum tolerable phase mismatch. Significant four-wave mixing occurs only if $L < L_{coh}$.

In the case of a monomode fiber, Δk_w is very small and can be ignored. For λ_p below the zero of dispersion $\lambda_D \approx 1.3\ \mu m$ ($\beta_2 > 0$), one has to keep both Δk_m and Δk_{NL} small by using small frequency shifts Ω_s and low pump powers. When Δk_m dominates, we have

$$L_{coh} \approx \frac{2\pi}{|\Delta k_m|} = \frac{2\pi}{\beta_2 \Omega_s^2} \qquad [57]$$

In the visible range, one has $\beta_2 \approx 50 - 60\ ps^2\ km^{-1}$, and $L_{coh} \approx 1$ km for $f_s = \Omega_s/2\pi \approx 50$ GHz. For a pump wavelength above the zero dispersion ($\beta_2 < 0$), the negative value of Δk_m must be compensated by Δk_{NL}. For phase matching one has

$$\kappa = 0 \approx \Delta k_m + \Delta k_{NL} = \beta_2 \Omega_s^2 + \Delta k_{NL} \qquad [58]$$

From this, we obtain

$$\Omega_s = \sqrt{2\gamma P_p / |\beta_2|} \qquad [59]$$

which coincides with the frequency shift characteristic of the modulation instability.

Amplifier Gain and Bandwidth

Four-wave mixing can be used, for example, in a fiber-optical parametric amplifier. In fact, the main difference between the four-wave mixing experiments and the parametric-amplification experiments is whether a signal at the phase-matched frequency is co-propagated with the pump or not. In the absence of a signal, the signal and idler waves are generated from amplifications of noise.

Let us consider the situation in which a signal wave with a power P_3 is launched at $z = 0$, and simultaneously an idler wave appears, with $P_4(z = 0) = 0$. In this case, considering the solution given by eqs [50] and [51] and taking into account eqn [52], we can write the following results for the signal and idler powers:

$$P_3(z) = P_3(0)\left[1 + \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P z)\right] \qquad [60]$$

$$P_4(z) = P_3(0)\left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P z) = P_3(z) - P_3(0) \qquad [61]$$

The single-pass gain of the parametric amplifier becomes

$$\begin{aligned} G_P &= \frac{P_3(L)}{P_3(0)} = 1 + \frac{P_4(L)}{P_3(0)} = 1 + \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P L) \\ &= 1 + \frac{(2\gamma P_p)^2}{g_P^2}\sinh^2(g_P L) \approx \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P L) \end{aligned} \qquad [62]$$

The approximation holds for $g_P L \gg 1$. According to eqn [52], amplification ($g_P > 0$) occurs only in the case:

$$|\kappa| < 4\gamma P_p \qquad [63]$$

Since $|\kappa|$ must be small, this also means good phase-matching. If one has complete phase-matching ($\kappa = 0$) and still $g_P L \gg 1$,one obtains exponential behavior for the amplifier gain:

$$G_P = \sinh^2(g_P L) \approx \frac{1}{4}\exp(2g_P L) = \frac{1}{4}\exp(4\gamma P_p L) \qquad [64]$$

From Eqs [44] and [45], we have that the phase mismatch κ and the frequency shift Ω_s are related by

$$\kappa = \beta_2 \Omega_s^2 + \Delta k_w + \Delta k_{NL} \qquad [65]$$

From eqn [65] we have

$$\frac{\Delta\kappa}{\Delta\Omega_s} = 2|\beta_2|\Omega_s \qquad [66]$$

In order to estimate the amplification bandwidth $\Delta\Omega_s$, we set

$$\Delta\kappa = \kappa_{max} = 4\gamma P_p \qquad [67]$$

From Eqs [66] and [67] we obtain

$$\Delta\Omega_s = \frac{2\gamma P_p}{|\beta_2|\Omega_s} = \frac{2}{|\beta_2|\Omega_s L_{NL}} \qquad [68]$$

where $L_{NL} = (\gamma P_p)^{-1}$ is the nonlinear length. Considering typical values $|\beta_2| = 20\text{–}60\ ps^2\ km^{-1}$, $\Omega_s/2\pi = 10\text{–}100$ THz, and $L_{NL} \sim 1$ m, we obtain a bandwidth in the range ~10–100 GHz. This bandwidth lies between that of a fiber Raman amplifier (~5 THz) and a fiber Brillouin amplifier (~100 MHz).

For $|\kappa| > 4\gamma P_p$ the parametric gain becomes imaginary and there is no longer amplification, but rather a periodical power variation of the signal and idler waves. In particular, if $|\kappa| \gg 4\gamma P_p$, then

$g_P \approx j\kappa/2$ and the parametric amplifier gain is

$$G_P = (2\gamma P_p L)^2 \frac{\sin^2(\kappa L/2)}{(\kappa L/2)^2} \qquad [69]$$

A convenient measure for the width in this case corresponds to a wavevector mismatch $\Delta\kappa = 2\pi/L$, where L is the fiber length. Using eqn [66], we then have

$$\Delta\Omega_s = \frac{\pi}{L|\beta_2|\Omega_s} \qquad [70]$$

Amplifier Performance and Applications

Parametric amplification can induce considerable crosstalk in WDM communication systems. Such crosstalk affects the system performance, particularly for channel spacings less than 10 GHz. Considering a channel spacing of 5 GHz and −20 dB allowed crosstalk, calculations show that the input power per channel should be smaller than 1 mW for a 15 km-long monomode fiber, and 0.1 mW for 100 km-long fiber. However, this effect can be avoided by an appropriate frequency separation of the channels.

Besides the limitation mentioned above, parametric amplification may be useful in the field of optical fiber communication systems. In fact, parametric amplifiers have a bandwidth of ~100 GHz, which is smaller that the bandwidth of fiber Raman amplifiers (~5 THz) but still sufficiently large for many applications. On the other hand, the frequency shift in FPAs can be as large as ~100 THz, which must be compared with the pump-signal frequency shift of ~13 THz of FRAs. These characteristics of FPAs allow a considerable flexibility in the choice of the pump source.

Parametric amplifier gains above 40 dB, with a 30-m-long fiber for input signal powers less than 1 mW and a pump power $P_p \approx 70$ W, were already measured. Such large gain values show clearly the potential of fiber parametric amplifiers provided phase-matching can be achieved. In fact, this condition puts stringent limits on the control of frequency shift Ω_s between the pump and signal waves. One method to satisfy the phase-matching requirement consists in using large birefringence of polarization-preserving fibers. Several parametric amplifiers using this kind of fiber have been demonstrated. Moreover, the possibility to realize a parametric laser was also demonstrated, by placing the fiber inside a Fabry–Perot cavity.

Fiber-optic parametric amplifiers have also attracted considerable attention in the context of a phenomenon known as squeezing. In contrast to coherent light containing equal minimum noise fluctuations in cophasal and quadrature components, in squeezed states one of these components acquires a noise fluctuation below the quantum-noise level at the expense of the other component. This phenomenon can occur in a parametric amplifier for certain values of the relative phase between the signal and idler waves. In fact, spontaneous emission at the signal and idler frequencies generates photons with random phases. Four-wave mixing increases or decreases the number of specific signal-idler photon pairs, depending on their relative phases. The noise reduction below the quantum-noise level can be verified using a phase-sensitive detection scheme in which the phase of the local oscillator is conveniently adjusted. The phase-sensitive nature of parametric amplifiers can also be usefully exploited to reduce the timing jitter in soliton communication systems.

List of Units and Nomenclature

A_e	effective fiber core area, m^2
c	vacuum speed of light, m s^{-1}
E_p	amplitude of the pump wave electric field, V m^{-1}
E_s	amplitude of the signal wave electric field, V m^{-1}
g_B	Brillouin gain coefficient, m W^{-1}
G_B	Brillouin amplifier gain
g_P	parametric gain, m^{-1}
G_P	parametric amplifier gain
g_R	Raman gain coefficient, m W^{-1}
G_R	Raman amplifier gain
$\vec{k}_a$	wave vector of the acoustic wave, rad m^{-1}
$\vec{k}_p$	wave vector of the pump wave, rad m^{-1}
$\vec{k}_s$	wave vector of the signal wave, rad m^{-1}
L	amplifier length, m
L_{coh}	coherence length, m
L_e	effective amplifier length, m
L_{NL}	nonlinear length, m
n	refractive index
$\bar{n}_2$	nonlinear-index coefficient, m^2 W^{-1}
p_{12}	longitudinal elasto-optic coefficient
P_{asp}	amplified spontaneous scattered power, W
P_p	pump power, W
P_s	signal (Stokes) power, W
v_a	acoustic velocity, m s^{-1}
ν_B	Brillouin frequency shift, s^{-1}
α_p	absorption coefficient at the pump frequency, m^{-1}

α_s	absorption coefficient at the signal frequency, m^{-1}
β_2	group-velocity dispersion coefficient, $s^2\,m^{-1}$
γ	nonlinearity coefficient, $W^{-1}\,m^{-1}$
Γ	inverse acoustic phonon lifetime, s^{-1}
δ	frequency detuning from Brillouin gain peak, s^{-1}
$\Delta\nu_B$	Brillouin gain bandwidth, s^{-1}
$\Delta\nu_p$	pump linewidth, s^{-1}
$\Delta\nu_R$	Raman gain bandwidth, s^{-1}
ε_0	free-space permittivity, $F\,m^{-1}$
ρ_0	average material density, $kg\,m^{-3}$
ϕ	nonlinear phase shift, rad
ω_a	acoustic wave frequency, rad s^{-1}
ω_p	pump wave frequency, rad s^{-1}
ω_s	signal (Stokes) wave frequency, rad s^{-1}
Ω_s	sideband frequency shift, rad s^{-1}

See also

Optical Communication Systems: Wavelength Division Multiplexing. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Scattering:** Raman Scattering.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*, Ch. 8. New York: John Wiley & Sons.

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn, Ch. 8–10. San Diego, CA: Academic Press.

Boyd RW (1992) *Nonlinear Optics*, Ch. 8–10. San Diego, CA: Academic Press.

Buck JA (1995) *Fundamentals of Optical Fibers*, Ch. 7 and 8. New York: John Wiley & Sons.

Cotter D (1987) Fiber nonlinearities in optical communications. *Optical Quantum Electronics* 19: 1–17.

Cotter D (1993) Nonlinearity in optical fiber communication. In: Eason RW and Miller A (eds) *Nonlinear Optics in Signal Processing*, Ch. 9. London: Chapman & Hall.

Forghieri F, Thach RW and Chraplyvy AR (1997) Fiber nonlinearities and their impact on transmission systems. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications – IIIA*, Ch. 8. San Diego, CA: Academic Press.

Sauter EG (1996) *Nonlinear Optics*, Ch. 5–11. New York: John Wiley & Sons.

Schubert M and Wilhelmi B (1986) *Nonlinear Optics and Quantum Electronics*, Ch. 11 and 12. New York: John Wiley & Sons.

Shen YR (1984) *The Principles of Nonlinear Optics*, Ch. 9–11, 14, 26. New York: John Wiley & Sons.

Semiconductor Optical Amplifiers

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

There has been rapid growth in the deployment and capacity of optical fiber communication networks over the past 25 years. This growth has been made possible by the development of new optoelectronic technologies that can be utilized to exploit the enormous bandwidth of optical fiber. Today, systems are operational which operate at bit rates in excess of 100 Gb/s. Optical technology is the dominant carrier of global information. It is also central to the realization of future networks that will have the capabilities demanded by society. These capabilities include virtually unlimited bandwidth to carry communication services of almost any kind, and full transparency that allows terminal upgrades in capacity and flexible routing of channels. Many of the advances in optical networks have been made possible by the optical amplifier.

Optical amplifiers can be divided into two classes: the optical fiber amplifier and semiconductor optical amplifier (SOA). The former dominates conventional system applications such as in-line amplification to compensate for optical link losses. However, due to advances in optical semiconductor fabrication techniques and device design, the SOA is showing great promise for use in evolving optical communication networks. It can be utilized as a general gain element but also has many functional applications, including optical switching and wavelength conversion. These functions, where there is no conversion of optical signals into the electrical domain, are required in transparent optical networks.

We will review SOA basics, technology (materials and structures), signal transmission performance (pattern effects, crosstalk, and ultrashort pulse amplification), and some important functional applications (optical switching and wavelength conversion).

Basic Principles

The SOA is based on similar technology as a laser diode. Optical gain is achieved by electrically pumping a suitable semiconductor material, such that a population inversion occurs between the material conduction and valence bands. An incoming photon can then be amplified when the resulting stimulated emission exceeds losses due to stimulated absorption and internal material losses. SOAs can be designed to operate in either the 1300 nm or 1550 nm optical communication windows. The principle of operation of an SOA with low residual reflectivities ($R \approx 0$) is shown in Figure 1. An input optical power experiences a single-pass gain $G = \exp(gL)$ after traveling through the SOA active waveguide of length L. The net gain coefficient $g = \Gamma g_m - \alpha_{int}$, where Γ is the optical confinement factor (the fraction of the propagating signal power confined to the SOA waveguide), g_m the material gain, and α_{int} the optical loss coefficient. g_m is a function of the injected carrier (electron) density and wavelength. The single-mode active waveguide can support two orthogonal polarization modes: transverse electric (TE) and transverse magnetic (TM).

The amplification process also adds broadband noise, amplified spontaneous emission (ASE) to the output signal. The amplifier noise figure (NF) is a measure of the signal-to-noise ratio (SNR) degradation of the signal after amplification. It is defined as the ratio of the SNR of the input to that of the output when the input noise is shot noise limited. Expressions for the output SNR are usually calculated assuming that the SOA is followed by a narrowband optical filter and an ideal photodetector, such that the main source of detector noise is the beat noise between the signal and the ASE falling within the filter bandwidth. In this case the SOA noise figure is given by

$$\mathrm{NF} = 2n_{sp}K \qquad [1]$$

with the excess noise factor K given by

$$K = \frac{(1 + RG)(G - 1)\Gamma g_m}{(1 - R)G(\Gamma g_m - \alpha_{int})} \qquad [2]$$

where n_{sp} is the effective population inversion parameter. To achieve a low NF, the internal losses must be small and a value of n_{sp} close to one is required. A favorable value of n_{sp} is achieved by operating the SOA at high gain.

A low R is required to prevent the SOA from oscillating at high gains. The residual reflectivity is manifest as ripples in the amplifier gain spectrum and ASE spectrum.

SOA Structures

The key parameters required for a practical SOA are:

- low reflectivity ($<10^{-4}$) to ensure low gain ripple (<0.5 dB);
- low optical loss coefficient to achieve a high gain;
- high material gain to allow low drive current operation;
- low polarization sensitivity (<0.5 dB) because the polarization state of the optical signal coming from a link fiber is usually random;
- high saturation output power (P_{sat}), defined as the output power at which the gain is reduced by 3 dB; and
- low fiber-to-chip coupling losses (<3 dB per facet).

A schematic diagram of a commercial SOA chip is shown in Figure 2. The active waveguide consists of a 0.2 μm thick InGaAsP bulk active layer sandwiched between 0.1 μm thick InGaAsP separate confinement heterostructure (SCH) layers. The central section of the active waveguide is 600 μm long with a constant width of 1.4 μm. The mode-expanding active waveguide tapers are 150 μm long with a width that changes linearly from 1.4 μm to 0.4 μm at the tip. The tapers allow optical coupling to an underlying passive waveguide enabling efficient coupling to the input and output optical fibers. The structure provides a high confinement factor because of the refractive index mismatch between the layers in the gain section. The p-n junction formed by the p-InP

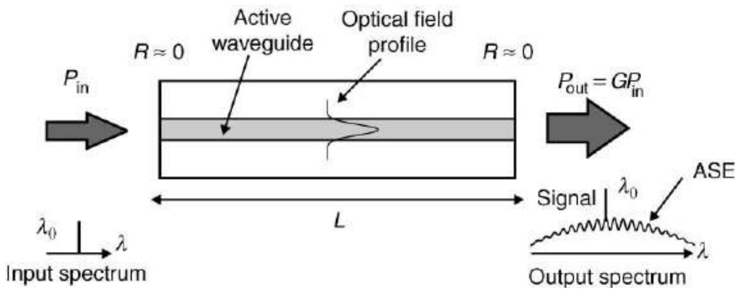

Figure 1 SOA basic structure.

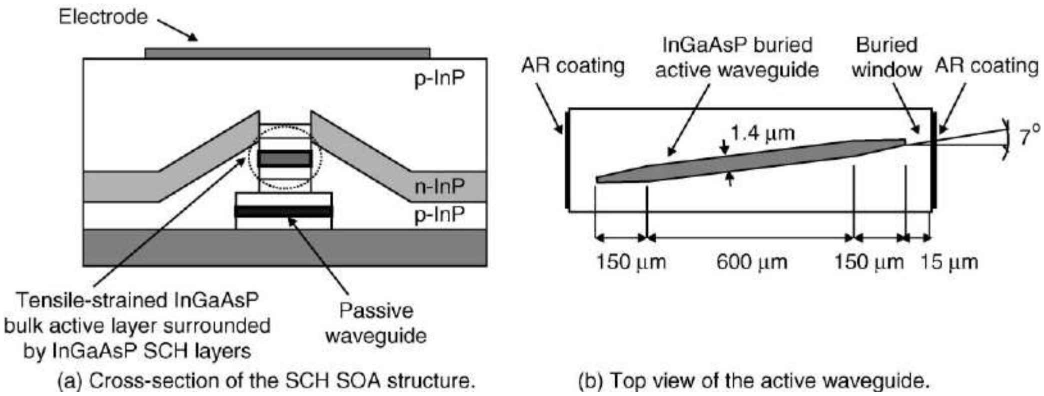
Electrode
p-InP
n-InP
p-InP
Tensile-strained InGaAsP bulk active layer surrounded by InGaAsP SCH layers
Passive waveguide
AR coating
InGaAsP buried active waveguide
Buried window
AR coating
1.4 μm
7°
150 μm
600 μm
150 μm
15 μm

(a) Cross-section of the SCH SOA structure.

(b) Top view of the active waveguide.

Figure 2 SCH SOA with tensile-strained bulk active layer.

Table 1 Typical specifications of a commercial SOA

Wavelength of peak gain	1550 nm
Peak fiber-to-fiber gain	22 dB
Noise figure	6.5 dB
Saturation output power	10 dBm
Polarization sensitivity	0.2 dB
Gain ripple	0.1 dB
3 dB bandwidth	40 nm
Drive current	200 mA

and n-InP layers acts as a current block, thereby providing good confinement of the injected carriers from the drive current to the active layer. Very low-reflectivity ($<10^{-6}$) is obtained by combining buried windows with antireflection-coated tilted facets (7° tilt angle).

Because of the active waveguide asymmetry, the TE confinement factor is larger than the TM confinement factor. g_m is isotropic in bulk material. The introduction of tensile strain, due to the lattice mismatch between the active layer and SCH layers, causes the bulk material band structure to change in such a way that the TM material gain is greater than the TE material gain. The introduction of the correct amount of tensile strain compensates for the difference in the TE and TM confinement factors leading to low polarization dependence.

Other SOA structures that can achieve low polarization dependency are based on active waveguides with near square cross-section having almost the same TE and TM confinement factors or the combination of compressively strained quantum wells (higher TE gain) and tensile strained quantum wells (higher TM gain). Typical specifications and characteristics of a commercial SOA are listed in Table 1 and shown in Figure 3.

As Figure 3 shows, gain saturation in conventional SOAs is manifest at output signal powers well below P_{sat}. Because the gain recovery time (carrier lifetime) in SOAs is fast (typically 0.1–1 ns), this can lead to pattern effects in single-channel systems and severe crosstalk in wavelength division multiplexed (WDM) systems. This problem can be greatly reduced by the use of a gain-clamped SOA (GC-SOA). In a GC-SOA, lasing action is produced, at a wavelength remote from the operating wavelength range, by the introduction of wavelength specific feedback. Once lasing begins, the carrier density in the SOA active layer is clamped at a fixed value. Changes in the input signal power lead to opposing changes in the lasing mode power. This has the effect of keeping the carrier density fixed (i.e., clamped), thereby making the signal gain relatively insensitive to changes in the total input power. A common method of providing this feedback is the use of distributed Bragg reflectors (DBRs) as shown in Figure 4. A typical GC-SOA gain versus output power characteristic is shown in Figure 5.

Basic Applications of SOAs in Optical Communication Systems

The three basic applications of SOAs in optical communication systems are: booster amplifier, in-line amplifier, and preamplifier. The main requirements of SOAs for such applications are listed in Table 2.

A booster amplifier is used to increase the power of a high power signal. Boosting laser power in an optical transmitter can be used to overcome external modulator losses, compensate for splitting and tap losses in optical distribution networks and to increase the power budget of optical links. The most critical requirement of a booster amplifier is a high P_{sat}, to obtain a high output signal power and minimize pattern effects. SOAs with P_{sat} in excess of 10 dBm are now available commercially.

The sensitivity of a conventional optical receiver is limited by thermal noise. The sensitivity is the

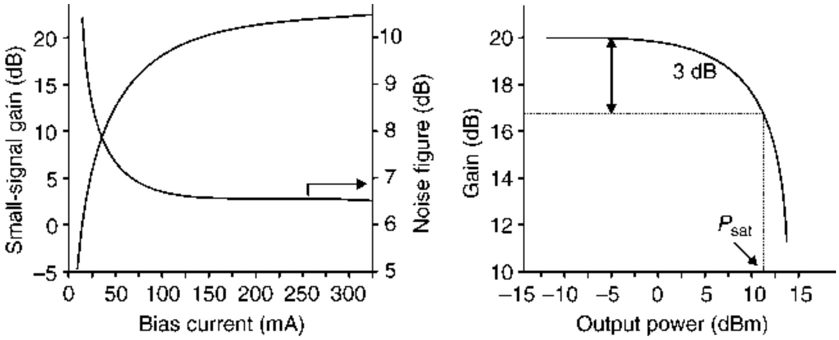

Figure 3 Typical commercial SOA small-signal gain and noise figure versus bias current and gain versus output power characteristics.

minimum signal power required at the receiver input to achieve a desired bit-error-rate (typically 10^{-9}). An optical preamplifier can be used to increase the power level of an optical data signal prior to detection and demodulation, leading to an increase in sensitivity. The performance of a preamplified optical receiver is dependent on the detector signal-to-noise ratio, $\text{SNR} = i_{\text{sig}}^2 / \bar{i}_{\text{noise}}^2$. The signal photocurrent i_{sig} is proportional to the amplified signal power. The mean square noise current $\bar{i}_{\text{noise}}^2$ includes circuit noise (thermal noise and detector dark current noise), signal shot noise, spontaneous emission shot noise, signal-spontaneous beat noise, and beat noise between the spectral components of the spontaneous emission. The best improvement in SNR occurs when the SOA is operated in the signal-spontaneous beat noise limit. In this regime, the signal power is sufficiently large such that the dominant receiver noise is the signal-spontaneous beat noise that falls within the signal bandwidth. This usually requires that the spontaneous emission from the SOA be reduced using a narrowband optical filter. The noise figure is a critical parameter in this application and should be as low as possible.

In long-haul optical transmission systems, in-line optical amplifiers can be used to compensate for link losses thereby increasing the spacing between optical regenerators. The main advantages of in-line SOAs are transparency to data rate and modulation format (unsaturated operation), bidirectionality, WDM capability, simple mode of operation, low power consumption, and compactness. The latter two advantages are important for remotely located optical components. To avoid noise accumulation in the link, it may be necessary to follow each SOA by a narrowband optical filter, but this can prevent capacity enhancements using WDM.

An example of a single channel optical transmission system utilizing booster, in-line, and

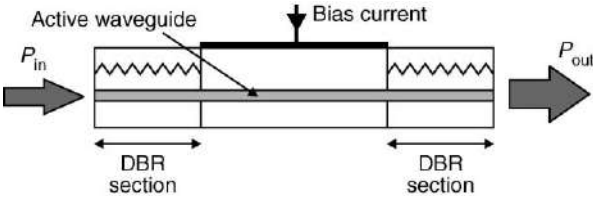

Figure 4 Schematic structure of a GC-SOA.

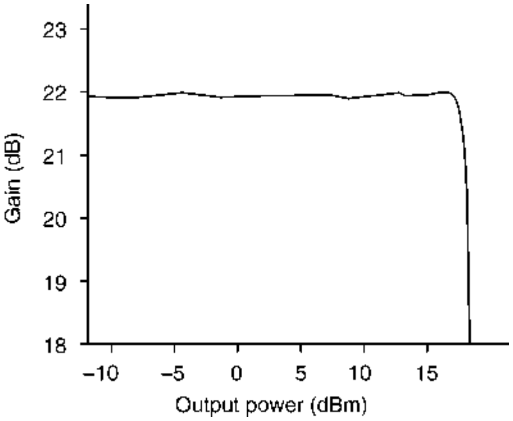

Figure 5 Typical GC-SOA gain versus output power characteristic.

preamplifier SOAs is shown in Figure 6. The transmitter comprised a 1309 nm gain-switched laser diode directly modulated with a 10 GHz sinusoid to produce a train of 40 ps wide pulses at a repetition rate of 10 GHz. At 1309 nm fiber dispersion is small and the maximum transmission distance is mainly limited by the link losses. The laser output was connected to an external modulator, driven by a 2^{31}-1 pseudo-random bit sequence (PRBS) to produce an optical data stream with an extinction ratio of 13 dB. A booster SOA was used to increase

Table 2 Requirements of basic SOA applications

Requirement	*Booster amplifier*	*In-line amplifier*	*Preamplifier*
High gain	Yes	Yes	Yes
High P_{sat}	Yes	Yes	Not critical
Low noise figure	Not critical	Yes	Yes
Low polarization sensitivity	Not critical	Yes	Yes
Optical filter	Not necessary	Not critical	Yes

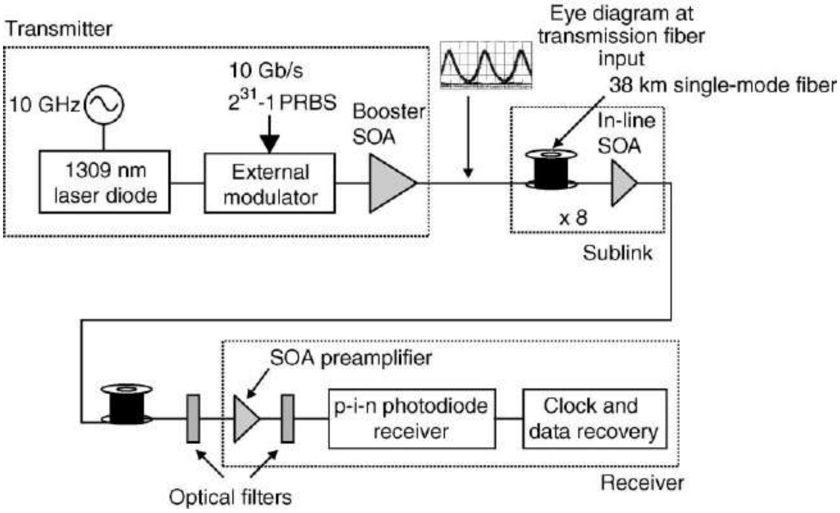

Figure 6 10 Gbit/s transmission experiment utilizing booster, in-line and preamplifier SOAs. (Adapted from Kuindersma *et al.* (1996) 10 Gbit/s RZ transmission at 1309 nm over 420 km using a chain of multiple quantum-well semiconductor optical amplifier modules at 38 km intervals. European Conference on Optical Communications 1996.)

the average transmitted power to between 0 and 2 dBm (7–9 dBm peak power). The transmission fiber length was 420 km with 12 in-line SOAs, used to compensate for fiber loss, spaced at 38 km intervals. At the receiver the signal was passed through a 1 nm bandpass filter, to reduce the accumulated spontaneous emission, amplified by an SOA preamplifier and filtered by a 1 nm bandpass filter. The signal was then detected by a p-i-n photodiode followed by an electronic clock and data recovery circuit. The SOA preamplifier and filter increased the receiver sensitivity from −14 dBm to −31 dBm for a BER of 10^{-10}. The receiver penalty after 420 km was 5 dB. In this experiment, the main limitation on transmission distance was the accumulation of spontaneous emission within the optical filter bandwidth.

Pattern Effects and Crosstalk

When an SOA is operated in the unsaturated regime, the amplifier gain is independent of the number of input signals and signal data rate. Outside this regime, the SOA will cause distortion because at high input powers, the gain saturates and compresses. Dynamic gain saturation occurs on the same time scale as the gain recovery time. This leads to pattern effects where the power of an incoming bit affects the gain experienced by subsequent bits. This is particularly important when the bit rate is of the same order as the inverse of the gain recovery time as shown in Figure 7. In WDM systems cross-gain modulation (XGM) between the channels can lead to severe interchannel crosstalk.

A further complication in WDM systems using SOAs is interchannel crosstalk caused by four-wave mixing (FWM). FWM is a coherent nonlinear process between two optical fields within the SOA, resulting in gain modulation at the beat frequency between the fields, and in the process generating new sidebands. In WDM systems with equally spaced wavelengths, the net effect of FWM is the generation of crosstalk terms interfering with the desired signals. Although the crosstalk can be relatively low in power terms, it can produce significant power penalties due to the coherent beat noise phenomena. The level of FWM crosstalk increases as the channel spacing decreases and the channel output power increases.

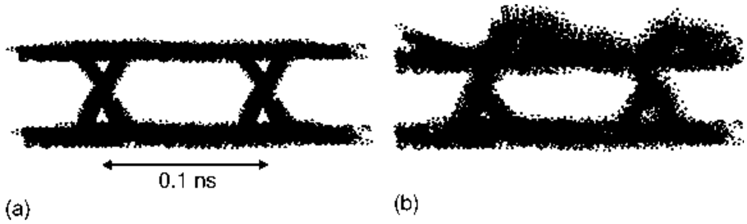

Figure 7 Typical eye diagrams of 10 Gb/s nonreturn to zero data (a) before and (b) after amplification by an SOA. ASE noise is present in the output eye diagram.

FWM crosstalk is of particular importance in dense WDM systems where the inter channel spacing is less than 100 GHz. Figure 8 shows a typical SOA output spectrum for an 8-channel multiplex with FWM-generated crosstalk signals.

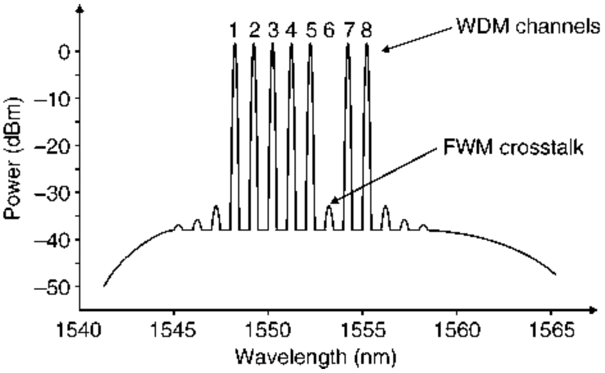

Figure 8 Typical SOA output spectrum for an 8-channel WDM multiplex with 100 GHz spacing. Channel 6 has been dropped to show the FWM crosstalk signal caused by channels 5 and 7.

Ultrashort Pulse Amplification

Optical time division multiplexing is an efficient method for increasing the bit rate of optical transmission systems. This requires the time interleaving of short optical pulses. The transmission distance is limited by the fiber group velocity dispersion, which is proportional to the pulse spectral width. Soliton transmission is also possible where the pulse shape is preserved as it propagates in the fiber.

Because an SOA has a very large bandwidth (typically 5 THz) it is capable of amplifying pulses as short as 100 fs. An input pulse can be amplified without significant distortion if the pulse energy is much less than the saturation energy E_{sat} of the SOA. Typical SOA saturation energies are of the order of a few pico Joules. As the pulse energy approaches E_{sat}, considerable spectral broadening and distortion can result. For an input pulse width τ_{p} (full width at half maximum) of the order of 10–100 ps spectral broadening is primarily due to self phase modulation (SPM). SPM is caused by gain saturation, which leads to intensity-dependent changes in the SOA active layer refractive index in response to carrier density variations. The degree of spectral broadening and distortion also depends on the input pulse shape. If τ_{p} is much less than the carrier lifetime, the output pulse power $P_{\text{out}}(t)$ and phase $\phi_{\text{out}}(t)$ are given approximately by

$$P_{\text{out}}(t) = P_{\text{in}}(t)\exp[h(t)] \qquad [3]$$

$$\phi_{\text{out}}(t) = \phi_{\text{in}}(t) - \frac{1}{2}\alpha h(t) \qquad [4]$$

with

$$h(t) = \left[1 - \left(1 - \frac{1}{G_0}\right)\exp\left(\frac{-U_{\text{in}}(t)}{E_{\text{sat}}}\right)\right]^{-1} \qquad [5]$$

$$U_{\text{in}}(t) = \int_{-\infty}^{t} P_{\text{in}}(\tau)\mathrm{d}\tau \qquad [6]$$

where $P_{\text{in}}(t)$ and $\phi_{\text{in}}(t)$ are the input pulse power and phase, respectively. G_0 and α are the amplifier unsaturated gain and linewidth enhancement factor, respectively. Typical values for α are in the range 2–10 and depend on the amplifier active region material and operating conditions. The output pulse spectrum can be obtained from

$$S(\omega) = \left|\int_{-\infty}^{\infty}[P_{\text{out}}(t)]^{1/2}\exp[i\phi_{\text{out}}(t) + i(\omega - \omega_0)t]\mathrm{d}t\right|^2 \qquad [7]$$

where $\nu_0 = \omega_0/2\pi$ is the pulse optical frequency. The output pulse chirp (frequency variation) is given by

$$\Delta\nu_{\text{out}}(t) = -\frac{1}{2\pi}\frac{\partial\phi_{\text{out}}}{\partial t} = \Delta\nu_{\text{in}}(t) + \frac{\alpha}{4\pi}\frac{\partial h}{\partial t} \qquad [8]$$

where $\Delta\nu_{\text{in}}(t)$ is the input pulse chirp. Using the above theory the shape, chirp and spectrum of an amplified zero-chirp (transform limited) Gaussian pulse of input power

$$P_{\text{in}}(t) = \frac{E_{\text{in}}}{\tau_0\sqrt{\pi}}\exp\left(-\frac{t^2}{\tau_0^2}\right)$$

is shown in Figure 9. E_{in} is the pulse energy and $\tau_p \approx 1.665\tau_0$. The amplified pulse is asymmetric because the leading edge of the pulse experiences a larger gain than the trailing edge. The amplified pulse spectrum is broader than the input pulse and also develops a multipeak structure. This is due to the SPM-induced frequency chirp imposed on the pulse as it propagates through the amplifier. In this case the chirp is linear around the peak of the pulse power spectrum. It can be compensated by transmitting the pulse through an optical fiber with anomalous group-velocity dispersion. In practice optical pulses can be far from Gaussian and can also have an initial chirp. In this case the induced chirp and resulting pulse spectrum can be more complex and more difficult to compensate.

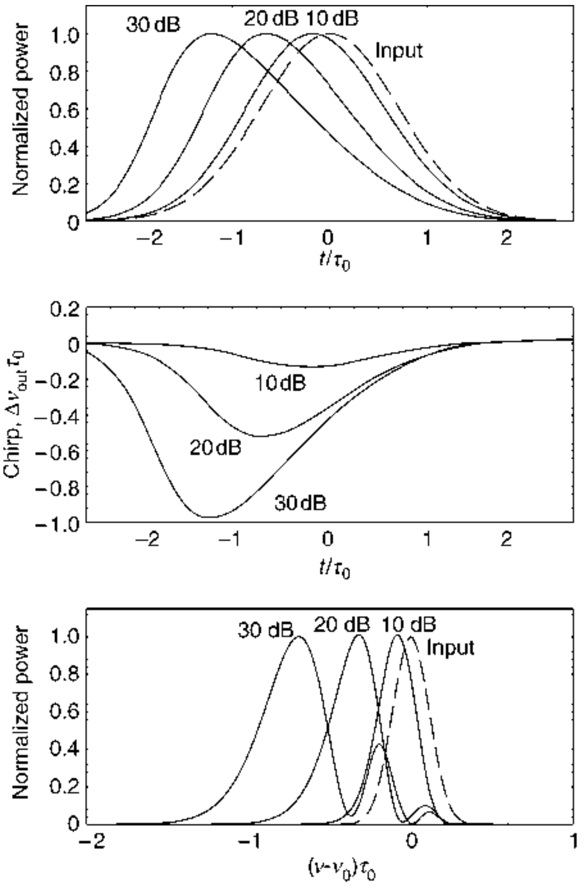

Figure 9 Amplified pulse shape, frequency chirp and spectrum for a transform-limited Gaussian input pulse with $E_{\text{in}}/E_{\text{sat}} = 0.1$ and $\alpha = 5$. ν is optical frequency. The parameter is the amplifier unsaturated gain.

When the input pulse width is less than ~10 ps, the above theory is no longer adequate and other nonlinear effects within the SOA such as carrier heating and spectral hole burning must be taken into account. The resulting pulse power and spectral distortions induced by an SOA on such pulses can be very complex.

Functional Applications

SOAs can be used to perform functions that are particularly useful in optically transparent networks. Recent advances in photonic integrated circuit and optoelectronic device packaging technology have made the deployment of SOA functional elements feasible. Two of the most important applications of SOAs are optical switching and wavelength conversion.

An example of a 2 × 2 SOA switch module used for optical routing is shown in Figure 10. Larger switch matrices can be constructed using this basic element. The module is based on a hybrid structure consisting of four integrated SOAs mounted on a silicon submount. The SOA-array is aligned to input and output polymer waveguides through V-grooves and alignment indentations. An incoming data packet can be routed to any output port by switching on the appropriate SOA. Switching times of the order of 1 ns are possible.

Ultrafast switching (<100 fs) can be achieved by incorporating SOAs in non-linear loop mirror structures, such as the terahertz optical asymmetric demultiplexer (TOAD) shown in Figure 11. Switching is achieved by placing an SOA offset from the center of an optical fiber loop mirror and injecting data into the loop via a 50:50 coupler. The two

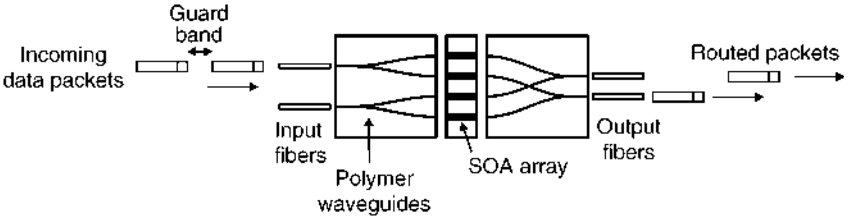

Figure 10 2 × 2 hybrid SOA switch module.

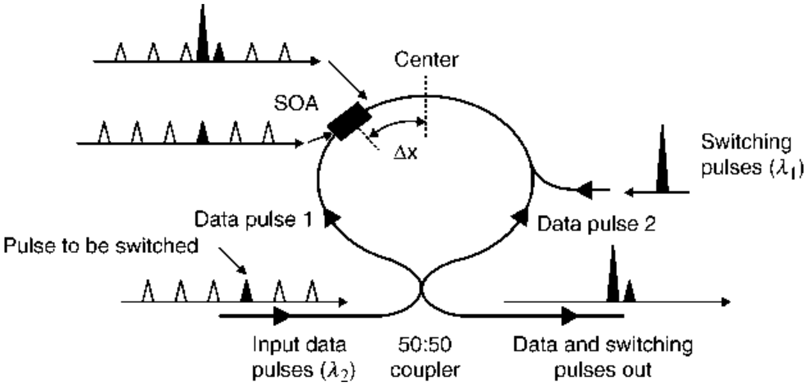
Center
SOA
Δx
Switching pulses (λ_1)
Data pulse 1
Data pulse 2
Pulse to be switched
Input data pulses (λ_2)
50:50 coupler
Data and switching pulses out

Figure 11 Ultrahigh speed optical switching using a TOAD.

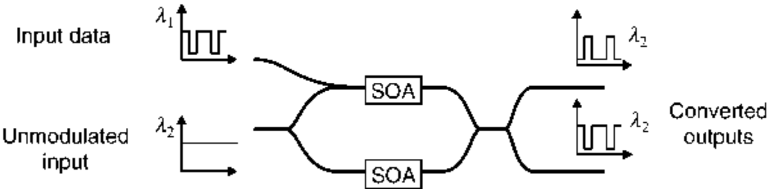

Figure 12 SOA Mach–Zehnder wavelength converter.

counter-propagating data pulse streams arrive asynchronously at the SOA. A switching pulse is timed to arrive after one data pulse but just before its replica. The switching pulse power is adjusted to impart a phase change of π radians onto the replica, so the data pulse is switched out when the two counter-propagating components interfere on their return to the coupler. The TOAD can also be used to demultiplex high-speed time division multiplexed pulse streams.

All-optical wavelength converters will play an important role in broadband optical networks. Their most important function will be to avoid wavelength blocking in optical cross-connects in WDM networks. Wavelength converters increase the flexibility and capacity of a network using a fixed set of wavelengths and can be used to centralize network management. In packet switching networks, tunable wavelength converters can be used to resolve packet contention and reduce optical buffering requirements.

Wavelength conversion in SOAs can be achieved using XGM, cross-phase modulation (XPM), or FWM. The design of an SOA wavelength converter based on XPM requires that one or more SOAs be incorporated into an interferometer, an example of which is the Mach–Zehnder interferometer (MZI) shown in Figure 12. An input data signal at wavelength λ_1 is used to modulate the upper SOA refractive index, which controls the phase shift experienced by a second unmodulated input signal at λ_2 in one of the interferometer arms. When the input data signal is low (logic 0), the arms are in phase and the λ_2 signal appears at the top output. When the input pump signal is high (logic 1) it produces an extra π phase shift between the two arms, causing them to be totally out of phase and the λ_2 signal to appear at the bottom output. Both outputs contain a copy of the original data signal imposed onto the new wavelength. Up and down conversion is possible. The lower SOA is used to equalize the gain experienced by the λ_2 signal in each of the arms. An optical filter is required at the output to remove λ_1. Such wavelength converters can operate at very high bit rates (>10 Gb/s). An important advantage of the MZI structure is that it also gives 2R regeneration (re-amplification and reshaping) of the input data signal with wavelength conversion. This is because the interferometer nonlinear response increases the converted signal extinction ratio compared to the input data. All-optical 3R (2R + retiming) regenerators are of great interest in future optical communication networks to restore degraded transmission signals. SOA-based interferometric structures can be used to achieve 3R (2R + retiming) regeneration at data rates >80 Gb/s.

See also

Coherent Lightwave Systems. **Lasers:** Semiconductor Lasers. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems; Optical

Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Optical Time Division Multiplexing; Wavelength Division Multiplexing. **Semiconductor Physics:** Band Structure and Optical Properties.

Further Reading

Agrawal GP and Olsson NA (1989) Self phase modulation and spectral broadening of optical pulses in semiconductor laser amplifiers. *IEEE Journal of Quantum Electronics* 25: 2297–2306.

Connelly MJ (2002) *Semiconductor Optical Amplifiers*. Boston, MA: Kluwer Academic.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.

Durhuus T, Mikkelsen B, Joergensen C, Lykke Danielsen S and Stubkjaer KE (1996) All-optical wavelength conversion by semiconductor optical amplifiers. *Journal of Lightwave Technology* 14: 942–954.

Ghafouri-Shiraz H (1995) *Fundamentals of Laser Diode Amplifiers*. New York: Wiley.

Shimada S and Ishio H (1992) *Optical Amplifiers and their Applications*. New York: Wiley.

OPTICAL COATINGS

Contents

Anti-Counterfeiting and Decorative Coatings
Diamond Optical Devices and Coatings
Laser Damage in Thin Film Coatings
Optical Black Surfaces
Thin Film Optical Coatings
X-Ray Coatings

Anti-Counterfeiting and Decorative Coatings

R W Phillips and R L Bonkowski, Flex Products, Inc., Santa Rosa, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Scientists are well known for the diversity of their interests. One topic that captures the interest of virtually all scientists, and people in general, is money. Not only are people engaged in obtaining money, but they are also fascinated about the manufacture of money and the protection of its value.

The relationship between manufacturing techniques and the protection of value documents (currency, stock shares, bonds, etc.) is not accidental. Technology has always been employed to counter the contemporary threats of counterfeiters. These threats vary, depending on the skill set and resources available to the criminal, as well as the counterfeiting strategy adopted. While we refer to banknotes repeatedly in this chapter, the information applies, as well, to passports, stamps, tax banderoles, identification cards, visas, licenses, checks, and bankcards. Banknotes can be considered the most challenging application for optical anticounterfeiting technologies. Stamps, banderoles, and checks are single-use items, and do not have to be as durable as currency. Identification cards, passports, and the like are used many times but are generally in protective covers and are examined by trained personnel. Currency must endure the rigors of commerce and travel, while still being easily authenticated by the average citizen.

The ever-increasing availability of low-cost computer hardware and software, along with color photocopiers and printers, has stimulated the use of optical and electro-optical technologies that were previously applied in a limited fashion or not at all. These optical techniques include: transmitted light devices, such as watermarks and threads; luminescent materials; reflective, interference, and diffraction schemes; modulation techniques (Moiré, for example); and light polarization effects.

Each of these concepts either has been or is likely to be applied to value documents to prevent counterfeiting. Each has its own set of potential benefits and application challenges. Also, some of these effects have been successfully applied to commercial decorative products. It is important to understand the difference between commercial and security uses. Effects aimed at commercial uses are mainly decorative or aesthetic, whereas security devices are functional and integrated into a security strategy. To ensure economic success, decorative products must be easily manufactured and widely distributed. In contrast, security products are carefully controlled in both their availability and the diffusion of manufacturing expertise.

Historical Perspective

Paper currencies (banknotes) have been printed for hundreds of years. Their history has been marked by a succession of technical changes in an effort to prevent counterfeiting. Sophisticated manufacturing techniques, such as intaglio printing and watermarks, along with special papers, made it difficult for counterfeiters to produce acceptable imitations. Intaglio is the centuries old method of printing, using incised or engraved metal plates. The tremendous pressure of the intaglio press forces the paper into an ink-coated plate engraving, forming patterns of ink on the paper.

In the eighteenth and nineteenth centuries, banknote printers often used unique engravings of guilloche patterns (intricate, overlapping curves), vignettes, and portraits as anticounterfeiting devices. These sophisticated patterns effectively discouraged early, relatively unskilled, counterfeiters. The development of photography and lithographic printing in the late nineteenth century created new opportunities for counterfeiting. Photochemical technology made it possible to produce suitable reproductions using print shop facilities throughout the world.

Photochemical techniques had significant limitations, however. The printed notes were passable but not high-quality imitations of genuine intaglio notes. Making lithographic plates required specific skill, considerable equipment, and left behind meaningful evidence (films, plates, press blankets, etc.) for police to find.

Until the 1980s, counterfeiting detection methods by governmental officials were relatively uncomplicated, as three general perpetrators could be classified in fairly predictable ways (Table 1).

Until the advent of home computers, optical scanners, color printers, and copiers, the challenge for law enforcement was difficult but focused. Amateur counterfeiting was virtually nonexistent. The effort required to make a few notes was daunting, the risk high, and potential profit meager. This meant that anticounterfeiting police efforts could be concentrated on criminal counterfeiters. Government-sponsored counterfeiting was certainly known (both the Germans and the British attempted to de-stabilize their opponents' national economies during World War II), but this was more an issue of foreign policy, diplomacy, and espionage and not the main province of national police.

The situation changed quickly and radically in the late 1980s. Color photocopiers had improved in cost, quality, and size to the point that commercial establishments were buying hundreds of machines. The 1990s introduced numerous new technologies to seduce the casual counterfeiter. These included home computers, graphics editing software, color printers, and low-cost scanners.

The amateur counterfeiter now had an array of sophisticated tools that could copy, edit, and reproduce virtually any kind of document. Banknotes made by the techniques of 'digifeiting' (a term coined by Dr. Sara Church of the US Treasury Department) were not of high quality, but passable in small quantities in low risk situations. It was now possible, economical, and relatively safe to make one or a dozen bills, especially in low denominations.

Police efforts were now severely challenged as the proliferation of amateur counterfeiters made investigation and prosecution costly. Especially in the United States, average citizens were lured by the prospect of augmenting their incomes by printing one or two notes a week. When apprehended, effective prosecution and punishment of such small-stakes criminals was expensive and unpopular.

Large-production criminal counterfeiting was augmented by new digital and electro-optical

Table 1

Class of threat	*Typical volume of notes*	*Resources and skills (1980)*
Amateur	Small (<100)	Retail paper, ink. Household chemistry
Criminal	Medium to large (1000 to 1 000 000)	Litho press, darkroom, plate making, cutting equipment. Skilled printing personnel
State-sponsored	Large (millions)	Intaglio press, currency paper. Skilled engravers and printers, Major distribution networks

breakthroughs. Lithographic plates could now be made without cameras or chemistry. Laser ablated plates made directly on the printing press (such as Heidelberg/Presstek) accurately duplicated features as small as 10 or 15 microns. Further, computer-to-plate technology with erasable plates presented an unusual opportunity for counterfeit production in small press shops.

The use of new electronic and optical tools by amateur and criminal counterfeiters coincided with a rise in state-sponsored counterfeiting. The United States, in particular, was victimized in the early 1990s by high-quality, intaglio-printed notes. One such note, dubbed the 'super $100' was allegedly produced in huge volumes in the Middle East and circulated widely in Europe.

Various means of detecting counterfeits were promoted but remain largely ineffective today. The so-called 'detector pen' applies a small amount of iodine solution onto a banknote. The iodine reacts with starch (often used to stiffen paper to give it a banknote 'feel'), turning the pen mark from light brown to black. Unfortunately, the presence or absence of starch in the paper is not a foolproof criterion. US currency paper contains no starch, while the currencies of other countries, such as Canada, have a starch component.

Concurrent with the availability of new high-tech counterfeiting tools, optical science has been hard at work to create protection schemes, utilizing the latest optical effects and manufacturing technologies. These new technologies have been largely outside the skill set and resources of amateur, criminal, and even state-sponsored counterfeiters.

An analysis of all the types of optical materials used for anticounterfeiting purposes as well as for decorative purposes shows that all are based on the control of electromagnetic radiation in the ultraviolet (UV), visible, and infrared. Exploiting various degrees of light transmission, reflectance, absorption, or scatter controls this radiation. In particular, all anticounterfeit devices fall into one or more of five basic categories: light patterning, light re-emission, light diffraction, polarization, or light interference. Characteristics of an effective anticounterfeiting solution include:

- quick verification by the average person;
- ability to function in a variety of lighting situations;
- deterrence of scanning and electronic photocopying;
- compatibility with existing banknote production schemes and equipment;
- ability to be verified by machines;
- durable and reliable despite strenuous handling conditions;
- tight control of materials and manufacturing processes.

Devices Based on Light Patterning

Watermarks

Images can be created in the paper substrate during the manufacturing process. Watermarks have been in use for at least 500 years and are produced by the arrangement of the paper fibers in the cylinder-mold paper machine or by impression in wet paper in the Fourdrinier paper machine. Because they are not on the surface of a document, watermarks are not readily visible in reflected light, but become apparent when light is transmitted through the paper substrate. Credible counterfeits, however, have been produced using images created by imprinting paper with waxy or oily substances. The fatty 'ink' is not readily detected on the surface of the document and serves to change the light transmission characteristics of the paper.

Watermarks serve chiefly to authenticate the paper substrate and to deter the practice of 'raising' notes, that is, bleaching a genuine low-denomination note and re-printing it as a higher value. The most effective watermarks are produced by the cylinder-mold process and use an image that is a duplicate of that image printed in ink (Figure 1) elsewhere on the document. This gives the examiner a reference point to compare the watermark as, for example, on all high-denomination US currency.

Watermarks are an effective counterfeiting deterrent when the general population is educated about their presence and their function. The limited opportunity to expose banknotes to a source of transmitted light (outdoors or in a dimly lit restaurant, for example) somewhat reduces their effectiveness.

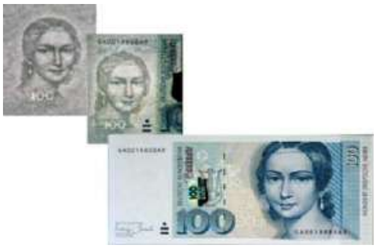

Figure 1 Watermark portrait of Clara Schumann, famous pianist and wife of composer Schumann: far left – portrait in transmitted light; center – portrait in reflected and transmitted light; and right – banknote in reflection. Courtesy of Papierfabrik Louisenthal.

Threads

Thin strips of polymer or metallized polymer can be imbedded in the paper substrate during the paper-making process. Such strips are approximately 12 microns thick and 1–5 millimeters wide. These strips are known as threads because Crane & Company developed the original concept in the 1860s, using colored silk thread. Typically, threads are installed border-to-border in the narrow direction of the banknote. Because they are inside the paper, they are not seen in reflected light, but are readily visible when light is transmitted through the substrate. Such threads are effective deterrents against counterfeiting achieved by scanning or photocopying, as the scanner uses only reflected light and cannot duplicate the buried thread. Threads have evolved into several varieties:

- De-metallized threads: Thin polymer coated with an opaque metal. The metal is selectively removed to create text or images.
- 'Window' threads: Metallized polymer that is 'woven' in and out of the paper to permit observation in both reflected and transmitted light (Figure 2).
- Fluorescent threads: Polymer impregnated with ultraviolet inks that emit a specific color under UV irradiation (Figure 3). See the section 'Devices Based on Light Re-emission' later in this article.
- Diffractive threads: Metallized polymer strips that have been embossed with diffractive or holographic patterns and attached by hot-stamping or adhesive to the surface of the note. See the section 'Devices Based on Light Diffraction.'

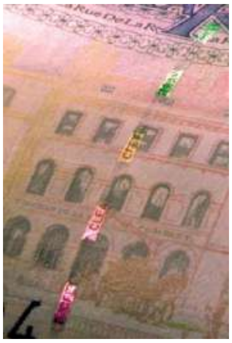

Figure 2 Windowed thread. Reproduced with permission from De La Rue International Limited.

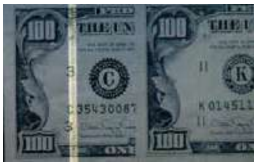

Figure 3 Fluorescent thread embedded into US $100 note. Reproduced with permission from Crane and Co., Inc.

Potential drawbacks to the use of threads are durability and the temptation of the general public to try to extract them from the banknote. Also, some threads can be simulated by printing or by attaching a polymer film to the surface of the banknote.

Laser Perforation

A modern 'watermark' concept is laser perforation of the banknote substrate. Precise patterns of microscopic holes (ca. 100 microns diameter) can be drilled using a high-power laser apparatus. The 200-Franc Swiss banknote features denomination numbers that are not readily visible in reflected light. However, when a light source is held behind the note, the numerals fabricated by laser perforation are clearly visible (Figure 4). In addition to being easily authenticated by non-expert observers with no special viewing apparatus, laser perforation has forensic advantages. The penetrations made by laser beams have a particular shape, spacing, and dimension, which are not easily mimicked. By clever adaptation of laser scanning software, holes of different shapes can be made and inter-mixed for additional security complexity. As a result, manual

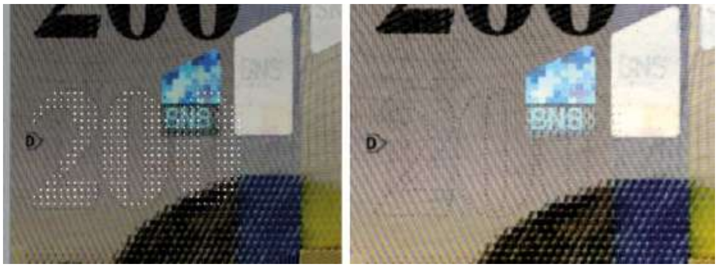

Figure 4 Microperf® laser holes in banknote. Left image in transmitted light; right image in reflected light. Reproduced with permission from Orell Füssli Security Printing Ltd.

pin pricks and mechanically pierced holes can be discriminated as forgeries.

Reflective Foils

One of the earliest defenses against photocopying and digital scanning was the use of reflective metallic foils affixed to the surface of the document. Such foils yield >80% reflection during scanning and the resultant scanned image is usually a black field.

Attachment of foils to currency paper or documents is carried out by applying a thermoset adhesive under heat and pressure and releasing the optical film from its carrier web by means of a separate polymeric release layer. This process is known as hot-stamping. Alternatively, an ultraviolet curing adhesive can be used to affix the foil to the paper substrate using a cold roll-on process.

Foiling may impede the amateur counterfeiter; since decorative foiling of stationery and packaging is well known, it presents little challenge to the professional perpetrator.

Planchettes

Planchettes are small flakes or fragments a few millimeters in diameter and 10–25 microns thick. These are often colorful or reflective materials and can feature a variety of effects including iridescence, diffraction, fluorescence, magnetic properties, or thermochromic properties. These fragments are attached to the banknote substrate by an adhesive (Figure 5). Planchettes discourage scanning and photocopying, since the copier cannot duplicate their special properties. Unfortunately, planchettes have been found to be non-durable and easy to simulate. The generally random placement makes it difficult for an examiner to authenticate a genuine note.

Thermochromic Inks

Inks that react to temperature changes have been employed on checks and identity cards. Thermochromic inks fluctuate from one color to another when the temperature of the printed image is raised. While it is possible for body temperature to activate this change, the effect is more certain if infrared heat is used. Since the process requires several seconds to activate (depending on the heat source), thermochromic materials have not been widely used on banknotes where immediate authentication is necessary.

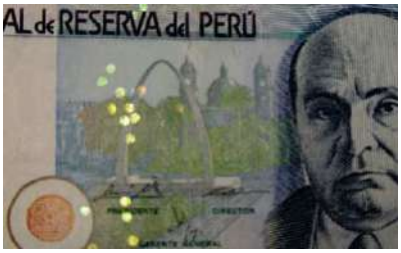

Figure 5 Peruvian 100 Nuevo Soles note showing iridescent planchettes. Reproduced with permission from Flex Products, Inc.

Colored Fibers

Extremely fine fibers (ca. 15 microns diameter) have been incorporated into currency papers for decades. The United States, for example, uses a few hundred grams of red and blue fibers per ton of paper in making currency. While of some forensic interest, the amount and size of the fibers makes them nearly undetectable by the average citizen.

Alternatively, fibers may be impregnated with fluorescent materials, or with absorbers that selectively absorb the wavelengths of light typically used in copy machines, producing obvious forgeries when copied.

Moiré Effects

Images can be created with very fine details that produce aliasing errors when digitized. Aliasing is the creation of new and spurious patterns through the superposition of two patterns whose periodicities are unequal. It represents interaction between the spatial frequency of the genuine document and the sampling frequency of the copying device. Aliasing is commonly known as moiré in the printing industry. Typical moiré generating patterns are constructed as closely spaced concentric circles, often found around the eyes in portraits on banknotes. When scanned digitally, the resulting copy will contain spurious images (scan traps) that are easily identified as fraudulent.

For example, if a genuine document is designed in a 1000 dpi (dots per inch) printing process it can later be scanned on a 1000 dpi scanner and, theoretically, all the data can be reproduced accurately. If, however, the scanner and the document are not in synchrony, the scanner will miss pieces of data, depending on the registration error, resulting in a digitized image that does not resemble the original.

In the scanned section of a US $5 note (Figure 6), one can see patterns of curved lines intersecting in the background, whereas the authentic image consists of fine parallel lines that do not intersect. The curvature and intersections are artifacts of the moiré created by digital scanning.

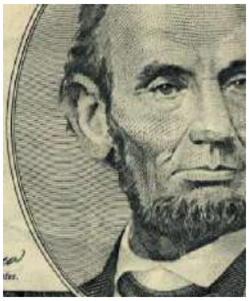

Figure 6 Scanned section of US $5 note showing moiré effect. Authentic image has no intersecting lines. Courtesy of Flex Products, Inc.

Modulation Techniques

Images produced by offset lithography are based on patterns (screens) of fine dots. By varying dot density, an array of shadings and tones can be created. More progressive modulation techniques, such as amplitude modulation (varying dot sizes) and frequency modulation (varying dot spacing) are often used to produce high-resolution images. To deter counterfeiting, banknotes can be printed using pixels, which are lines of equal length but assorted angular orientation. This technique is known as screen angle modulation (SAM).

SAM techniques can be used to defeat digital scanners by preventing the scanning of all available information due to under-sampling, or by concealing hidden patterns which are not demodulated by the human eye, but can be detected with suitable filters (Figure 7). A variant of this technology, called Scrambled Indicia®, has been used to protect US postage stamps (Figure 8).

Devices Based on Light Re-emission

Ultraviolet Fluorescent Inks

Ultraviolet (UV) inks are widely used on banknotes as semicovert deterrents. Images printed in UV inks are challenging to detect in normal lighting situations but are readily visible under UV (black light) illumination. Such printed images require a special light source for authentication and are subject to degradation over time from the handling and chemical exposure of banknotes. Typically, these products are sensitive to either short UV at 254 nanometers or long UV at 366 nanometers.

UV inks provide good protection against photocopying but can be readily simulated by more advanced counterfeiters.

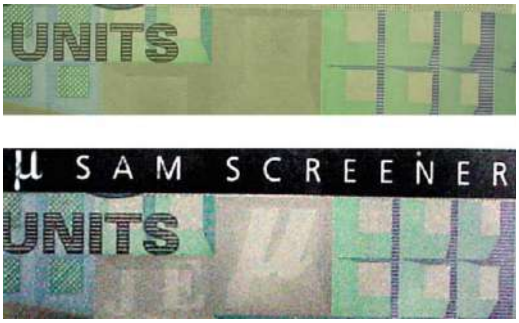

Figure 7 Top image – hidden image as normally seen; Bottom image – Image decoded with visible line screen. Courtesy of Joh Enschedé Security Printing Ltd.

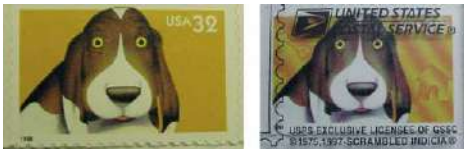
USA 32
UNITED STATES POSTAL SERVICE
USPS EXCLUSIVE LICENSEE OF GSSC
©1975,1997 SCRAMBLED INDICIA®

Figure 8 Hidden image in 32¢ US postage stamp decoded by invisible line screen (left image – normal view; right image – with decoder in place). Courtesy of Flex Products, Inc.

UV Fibers

Several countries mix minute quantities of UV luminescent fibers into the paper substrate. In most cases, these are randomly placed and invisible to the naked eye.

Devices Based on Light Diffraction

Diffractive structures range from simple periodic surface relief structures (textured surfaces) to multilevel holographic structures that form a continuum of products ranging from decorative coatings to anticounterfeit devices. Figure 9 shows a generalized structure for all known types of reflective-based relief structures.

Most devices are based on reflection; however, in a few instances where clear polymer is used, devices may also be based on diffractive light transmission. The sizes, shapes, and orientation of the grooves vary according to the exact device. The opaque aluminum layer insures that the design works in reflected light. Most of these structures exhibit a rainbow effect where the color moves from short to long wavelength as the viewing angle increases. This rainbow effect can be found for all types of structures with grooves that have sizes near that of light wavelengths. The color shifts in a direction opposite to that of a thin film interference structure. Further, the rainbow effect can only be seen in a direction perpendicular to the groove direction. In the special case where the groove dimensions are less than the wavelength of light, the

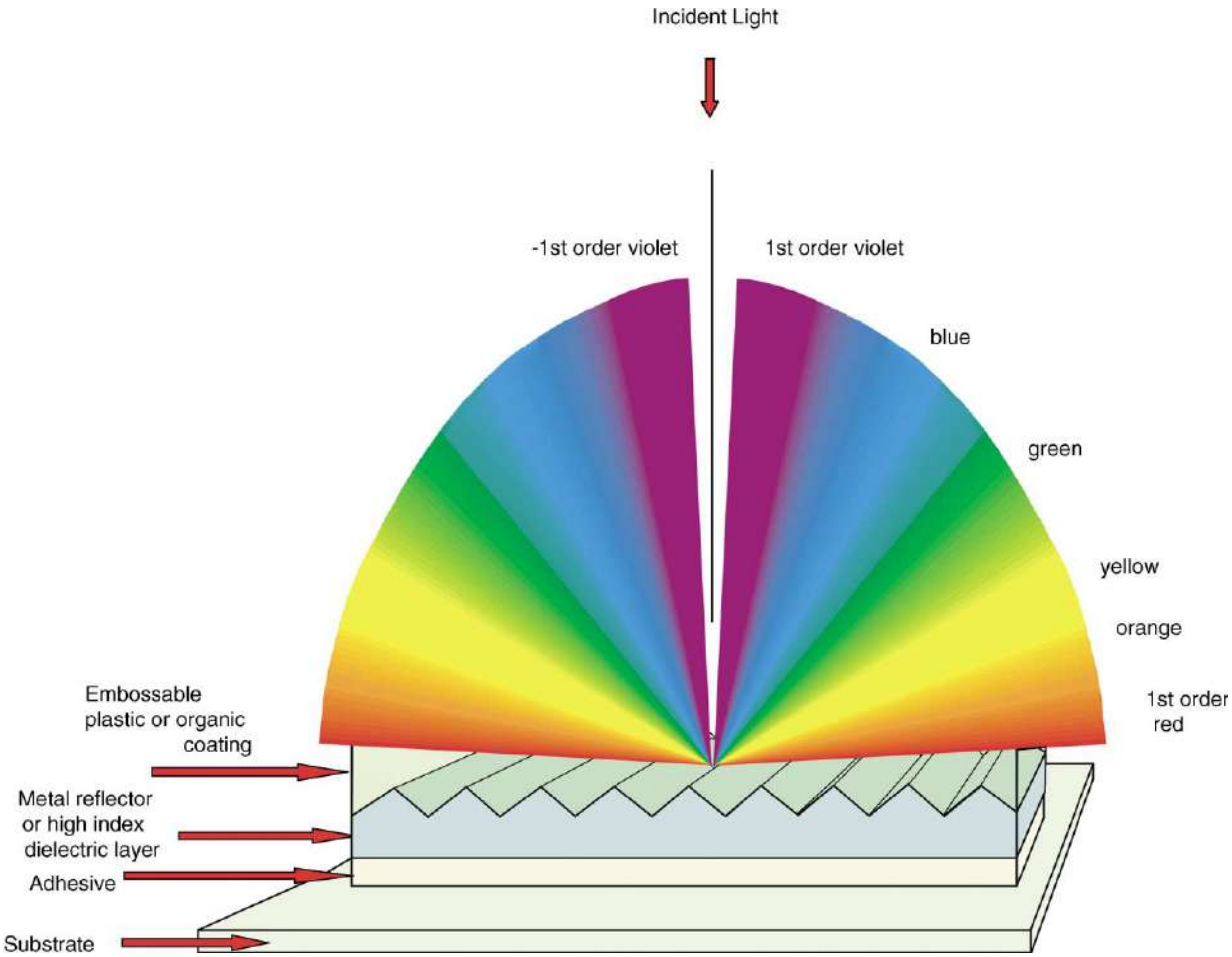

Figure 9 Cross-section of diffraction device. Courtesy of Flex Products, Inc.

rainbow effects disappear and are replaced by another type of diffraction (zero-order diffraction), where only specific colors are seen in reflection. Groove height and width for all diffraction products range from 0.1 to 10 microns.

For a unique set of discrete angles and for a given spacing d, the diffracted light from each facet is in phase with the light diffracted from any other facet, so they combine constructively:

$$Gm\lambda = \sin\alpha + \sin\beta \qquad [1]$$

where $G = 1/d$ is the groove density or pitch, α is the angle between the incident light and the normal to the grating, β is the angle between the diffracted beam and the normal to the grating and m is an integer called the diffraction order.

For $m = 0$, $\beta = -\alpha$ for all λ, the grating acts as a mirror and the wavelengths are no longer separated. This is called specular or zero-order reflection.

In contrast to grating structures (linear grooves) that only modify light intensities, holographic structures are complex surface contours that modify both the phase and intensity of the reflected light. Light reflecting off a metallized holographic structure is capable of producing an image in either two dimensions (2D) or three dimensions (3D). Heat and pressure may emboss the relief structure into a metallized polymeric film by hot-stamping. Alternatively, an embossing metal roller presses the relief image into a thermoplastic film or thin ultraviolet curable coating. The embossed structure is subsequently metallized with opaque aluminum or with a high-index dielectric material such as zinc sulfide or zirconium dioxide. The latter layers allow printed information on the substrate to show, while at the same time provide a holographic image. However, such high-index holograms are not as bright, since some of the reflective light is lost in transmission. Polymethylmethacralate (PMMA), polycarbonate (PC), polyvinyl chloride (PVC), and polyester (type G) are typical materials that can accept such an embossing.

Thick metallized embossed plastic films of ca. 50 microns are used for decorative labels. Thin metallized embossed hardcoats of ca. 7 microns, which can be released from carrier sheets or embossed thin plastic films, are used for the security field.

Holograms used for high-security applications have one characteristic that distinguishes them from ordinary holograms or other diffractive devices used for decorative purposes. This characteristic finds itself in the complication of images, multilevel images at different focal depths, different images that appear at various viewing angles, images of 3D objects superimposed with synthetic 2D images, micro-images, or images that appear to move as the device is tilted back and forth. Specific types of diffractive structures are described below. The list starts with simple holograms used in decorative markets and progresses to high-security devices.

Diffractive grating: A structure comprised of square or sinusoidal linear grooves that breaks light into its component colors. An assemblage of abutting islands, each randomly oriented and with its own parallel grooves, gives the viewer a twinkling color effect (Figure 10).

2D hologram: A relief structure that, upon viewing with a point source of light, shows a 2D image of objects, text, logos, or lines that emerge over a background. Such holograms are often referred to as 'cartoon images' since they are images of original drawings (Figure 11). Because of their simplicity and frequent use as decorative packaging, these holograms are generally not used in the security market.

Multilevel 2D hologram: A relief structure that shows a simulated 3D effect. This is accomplished by creating an image with two or more layers of flat artwork, with each layer at a slightly different focal depth (Figure 12). These images are well suited for labels, cards, and other decorative products. They have also been used on coupons, gift certificates, and tickets.

3D hologram: A relief structure that reconstructs a 3D image from a real or computer-synthesized 3D object (Figure 13). A 3D dove has been used on the VISA credit card as a security device for many years, though many now consider it more of a marketing logo than a security feature.

2D/3D hologram: A relief structure that reconstructs an image showing both a 3D and a 2D object, created from a 3D sculpture and flat art drawings (Figure 14). The artwork is created in such a way that the images appear at different depths of view.

Stereo-hologram: A relief structure that reconstructs a 3D image such that upon tilting from side to side one can see around the object, and additional features can then be viewed. At any given angle, it appears to be a 3D hologram. The image is created through multiple exposures of a 3D object during the holographic mastering process.

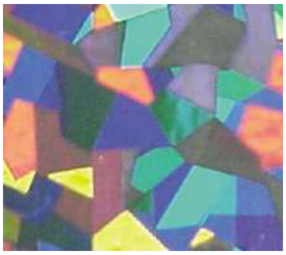

Figure 10 Picture of a typical diffractive assemblage of gratings. Reproduced with permission from Amagic Holographics Inc./K-Laser Technology, Inc.

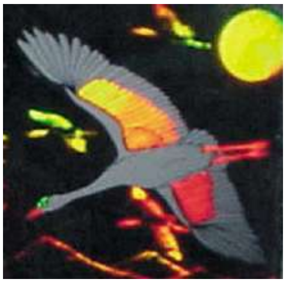

Figure 11 2D hologram. Courtesy of American Bank Note Holographics, Inc.

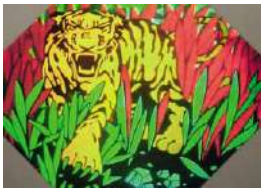

Figure 12 Multilevel, 2D hologram. Courtesy of American Bank Note Holographics, Inc.

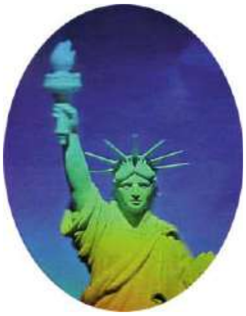

Figure 13 3D hologram. Courtesy of American Bank Note Holographics, Inc.

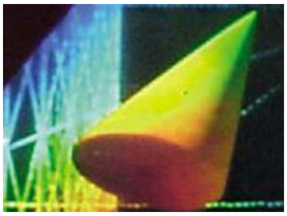

Figure 14 2D/3D hologram. Courtesy of American Bank Note Holographics, Inc.

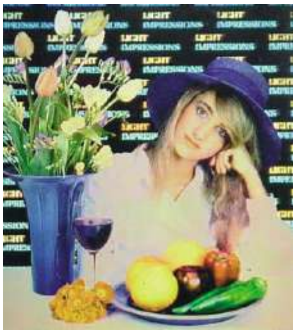

Figure 15 True color hologram. Reproduced with permission from Light Impressions International, Inc.

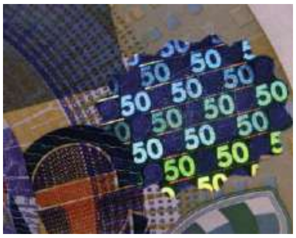

Figure 16 High-definition security hologram. Number 50 appears in various colors in first angle of view. Colored shells appear at another angle of view. Reproduced with permission from DeLaRue International Limited.

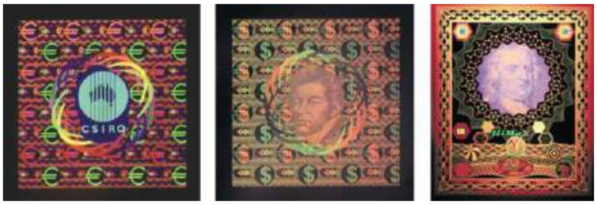

Figure 17 Left and center images – Exelgram™ at two different viewing angles. Reproduced with permission from CSIRO. Right image – HiMax® OVD high security device showing portrait of Sir Isaac Newton. Reproduced with permission from Light Impressions International, Inc.

True color hologram: In contrast to the holograms described above, a true color hologram is created by diffraction angle control of the colors red, green, and blue from a 3D object or color transparency. The relief structure, when viewed in white light, reveals a picture in true color, as opposed to the rainbow effect seen in regular 2D and 3D holograms (Figure 15).

High-definition security hologram: A relief structure that produces an image composed of multiple elements, such as fine line text and complex patterns that simulate movement as the device is tilted back and forth. Different images also appear at different angles (Figure 16). These images occur because different images are recorded in the holographic mastering process at different angles. The complicated imagery and the use of multiple images is the major holographic technique in use today for anticounterfeiting.

High-definition diffraction gratings: The relief structure in Figure 17 shows a multitude of elements, each with colors that change upon tilting. It consists of a very fine line grating structure, 1000 lines or greater per millimeter. These gratings are sometimes combined with holographic images that change depictions upon tilting, thus providing a high level of security.

Kinegrams®: This grating structure, based on computer-generated graphic images, is composed of light diffracting microstructures of less than one micron, with blaze (saw-tooth) contours that when viewed display optokinematic movements. This security device was first used in 1987 on Austrian 5000 Schillings. When the orientation of the blaze grating changes across the device, two distinct images are possible, depending on the view with respect to the blaze angles. The images do, however, require an aluminum reflective layer deposited onto the relief structure in order to generate the highly reflective colored images.

A 180° rotation of the device can lead to a marked change in the diffraction properties and, consequently, the image properties. This stands in contrast to holographic symmetrically based structures that yield identical images. An example of a Kinegram® security device is shown in Figure 18. The letter ‘K’ on the front facet of the cube is filled with an asymmetric structure, whereas the background is realized by using the same structure rotated by 180°. When viewed from one side, the letter ‘K’ appears bright on a dark background. Upon rotation in its plane by 180°, the contrast between the letter ‘K’ and the background is reversed.

Zero-order diffraction grating: This relief structure is quite different from holographic or normal grating structures, as it is based on an abrupt change in the

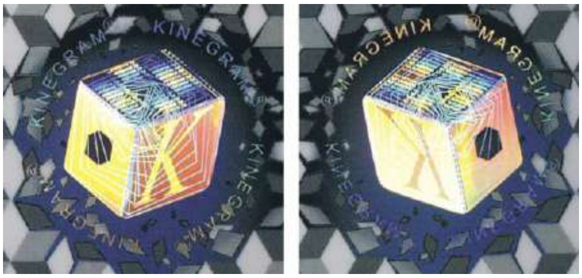

Figure 18 The Kinegram® security device as seen in two views differing from each other by a rotation of 180° in its plane, while illumination and viewing conditions remain unchanged. Reproduced with permission from OVD Kinegram Corporation.

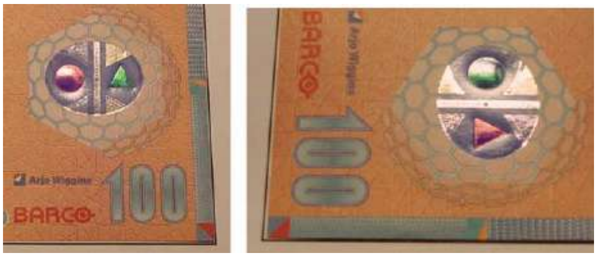

Figure 19 Zero-order diffraction images in a DID® security device: two views 180° apart in planar rotation. Reproduced with permission from Hologram Industries.

refractive index from approximately 2.3 to 1.5. Rather than employing an aluminized layer for reflection, the structure uses a thin, high-index layer ($n = 2.3$, for example) embedded in an organic polymer with a refractive index of about 1.5. The high-index layer appears as a parallel series of lines, approximately 0.2 microns wide, with another parallel series of lines beneath the first set, but oriented under the spacing of the first set of lines at a depth of about 0.2 microns.

This structure not only changes color as the device is tilted back and forth but also as it is rotated. In rotation, the color changes from red to green (Figure 19). Further, the light is polarized as it reflects. When a linear polarizer is inserted between the eye and the device, the color changes to black. The colors seen are based on specular zero-order reflection, unlike standard diffraction gratings where the reflected color is based on first-order reflection. The preeminent effect is when the device is placed over a black background. Over a white background the colors are subtler. Viewing reflection in zero order allows one to see the reflected colors in both point and diffuse lighting. The combination of polarization effects and color in rotation is ideally suited to high-security applications.

Devices Based on Light Interference

Thin-Film Foils

There are two types of thin-film interference devices. The first is based on an all-dielectric filter, and the second is based on a Fabry–Perot resonance cavity structure of the type absorber/dielectric/reflector. The all-dielectric filter uses pairs of high and low refractive index materials at a quarter-wave thickness. A typical filter is the one used initially on Canadian currency that employed the design: ZrO_2 1QW/SiO_2 3QW/ZrO_2 1QW/SiO_2 3QW/ZrO_2 1QW, where the QW (quarter-wave optical thickness) is centered at 585 nm. This design exhibits a gold-to-green color shift with viewing angle.

A Fabry–Perot filter also has a gold-to-green color shift and typically uses the design Cr 5 nm/MgF_2 4QW @ 604 nm/Al (opaque). This filter has slightly higher color saturation than the all-dielectric filter. The design is opaque and, therefore, the color observed is independent of the substrate color onto which it is affixed. This stands in contrast to the all-dielectric filter that must be laid over a black background for optical color saturation. If the all-dielectric filter is placed on a white background, the transmitted light passing through the optical filter will reflect off the white background and combine with the filter's reflection to give white light. However, with a varied colored background, the reflected color will vary depending on the combination of reflected color from the substrate and the reflected color from the filter.

The gamut of reflected colors from a metal–dielectric foil covers all four-color quadrants in the CIE Lab color space. Brightness of the foil depends on the materials used and on the design. Figure 20 shows the theoretical gamut of colors as plotted on the a^* b^* plane for the design, using the materials Cr, MgF_2 and Al.

Foils based on thin film interference layers contain a reflectance peak property moving from long wavelength to short wavelength as the viewing angle is increased. This is true across the UV, through the visible and into the near-infrared. To the human eye, the color moves from red to blue. The anti-counterfeit foil shown in Figure 21 exemplifies this effect.

Shifts from long to short wavelength occur because the path light difference (see the bold line in Figure 22) between the rays of light that reflect from the surface and the reflector are a function of the cosine of the incidence angle.

The path length difference is shown by the term $2d \cos \theta$, where θ is the angle of incidence within the

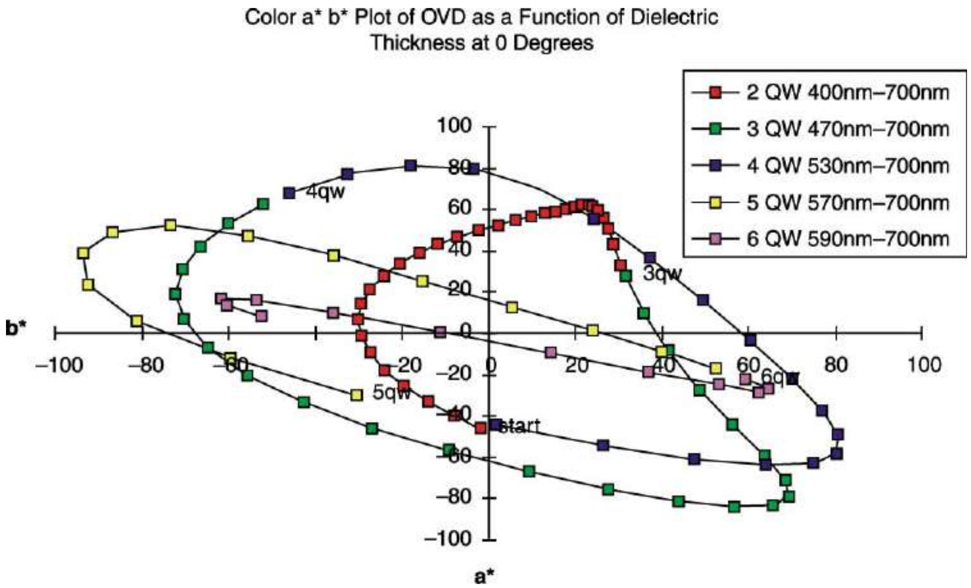

Figure 20 Gamut of colors at 0° incidence viewing under Illuminant C produced from Fabry–Perot designs as a function of quarter-wave optical thickness. Courtesy of Flex Products, Inc.

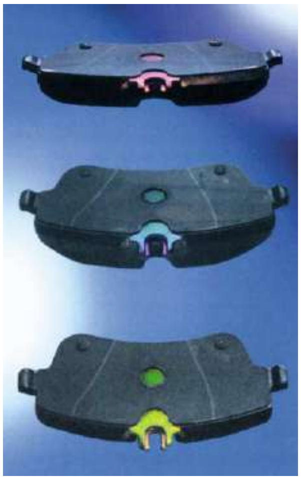

Figure 21 Anticounterfeit interference foil on brake shoes. Reproduced with permission from November AG.

filter and d is the distance between the surface of the filter and the reflector. The angle θ and the angle ϕ are related by Snell's law: $n_0 \sin \phi = n_1 \sin \theta$.

Light interference occurs where the wavefronts superimpose on one another. The fractional number of wavelengths that can fit into this path length difference is given by $\Psi = 2n_1 d \cos(\theta)/\lambda$, where λ is the wavelength of light. Waves reflected from these two surfaces will interfere constructively when Ψ is an integral number of wavelengths and interfere destructively when Ψ is at multiples of half wavelengths. As the viewing angle (ϕ) increases, the angle θ also increases. Thus, constructive interference translates into a reflectance peak that moves toward shorter wavelengths as the angle of viewing is increased due to its cosine dependence.

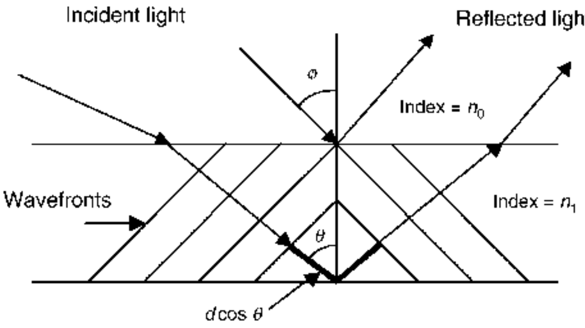

Figure 22 Diagram showing the principle of optical shift with viewing angle. Courtesy of Flex Products, Inc.

The Fabry–Perot filter can use any transparent dielectric (colored or clear), a reflective metal, although optical metals are preferable since they have the highest reflectance (e.g., Ag, Al, Ni, Cu), and an absorber that is partially transmissive and partially reflective. Thin metals, absorbing dielectrics, and metal nitrides are typical materials used as an absorbing layer.

Light Interference Inks

Inks based on interference pigments have been developed over the last thirty years, starting with pigments based on high-index thin film coated micas (typically TiO_2 coated micas) and culminating with

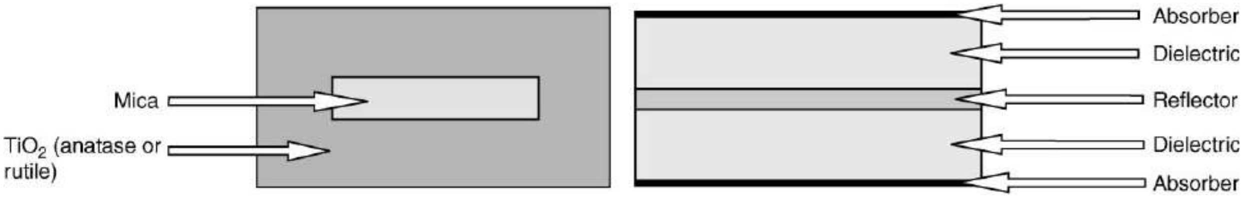

Figure 23 Left image – cross-section of interference mica; right image – cross-section of optically variable pigment (OVP®). Courtesy of Flex Products, Inc.

OVP® pigments based on the Fabry–Perot structure (Figure 23).

Inks created from micas are known as pastel or pearlescent inks, whereas inks formulated from OVP® are known as optically variable inks (OVI®). The latter inks are used as a primary defense against counterfeiting, defeating color copying and color printing from computer images. The difficulty with using interference micas for security features is that they can be readily purchased in the commercial world. In contrast, OVP® is a tightly controlled material and cannot be purchased by the general public.

OVI® security ink has been formulated into intaglio, silkscreen, flexographic and gravure applications. Its use on security documents encompasses over 75 banknotes worldwide, including the newly issued Euro notes, post-1996 \$10, \$20, \$50, and \$100 US banknotes, and banknotes for China. With the use of such inks, postage stamps, tax stamps, identification cards, and other security documents have also been effectively protected against counterfeiting. Color control for OVI® is assured because OVP® can be blended according to additive color theory to produce a desired color and color shift. Two examples of OVI®– the 100 Pula Botswana note and the new 100 Euro note – can be seen respectively in Figures 24 and 25.

OVI® inks have been extremely successful for banknote protection and might even be a *de facto* standard, because they meet the seven characteristics of an effective anticounterfeit solution as described earlier.

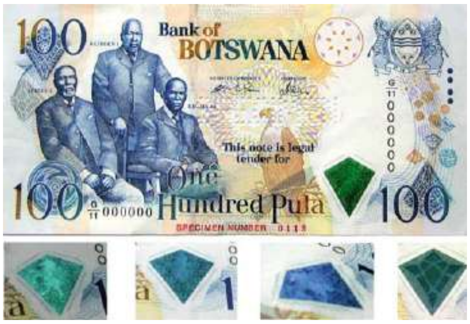

Figure 24 100 Pula Botswana note showing green-to-blue color shift OVI® with latent images. Reproduced with permission from SICPA.

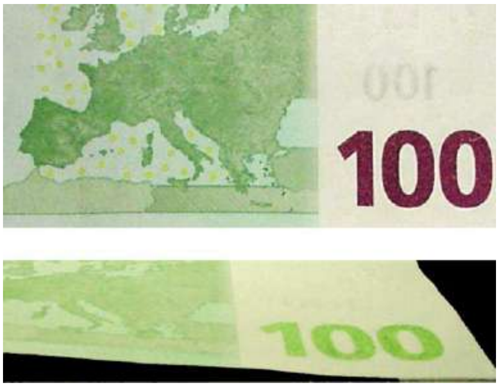

Figure 25 100 Euro note showing magenta-to-green OVI®. Courtesy of Flex Products, Inc.

Devices Based on Light Polarization

The use of light polarization as an anticounterfeit feature has taken several avenues, one as coatings or pigments, the other as invisible images that are detected with a linear polarizer. With the exception of the zero-order diffraction device described earlier, polarization devices are based on cholesteric liquid crystals, using either nematic (planar oriented) layers or helical twisted layers.

Liquid crystal pigments are constructed from cholesteric polysilozanes that have been oriented and cross-linked in place with photopolymerization. With this cross-linking, pigments are relatively stable to temperature changes. These materials can form helical structures in either right or left twist orientation. Such materials reflect 50% of a band of wavelengths, centered at wavelength, λ_0, of the incoming light as circularly polarized light and 50% of the band centered at λ_0 as transmitted, circular polarized light, but with the opposite twist. The remaining wavelengths pass through the structure with no interaction. The amount and type of the chiral phase (components that inherently orient into a helical structure) in the polymer determine the pitch of the helical structure. This, in turn, determines the wavelength of reflection, i.e. $\lambda_0 = nP$ where n is the average refractive index and P is the pitch length of the material (pitch length determined by 360° rotation of the helix). At tilted viewing angles, the color changes according to the formula $\lambda = \lambda_0 \cos\theta$ where θ is the viewing angle relative to the vertical, and λ_0 is the reflected wavelength seen perpendicular to the surface. By the judicious combination of left- and right-handed helical based layers, or by inserting a half-wave retarder layer between two layers having the same helical structure, the resulting pigments reflect nearly 100% at λ_0.

When these materials are properly processed and ground (particles 2–50 microns thick, with an aspect ratio of 3:1 in lateral/thickness dimensions), particles with the helical structure perpendicular to the planar surfaces can be realized. Such pigments, or liquid crystal polymeric coatings, when printed, have color-shifting properties with viewing angles that defeat color copiers and the like, since toners have no color-shifting properties (Figure 26). These pigments also have an additional security feature, namely, that when viewed through a circular polarizer having the opposite twist, the color-shifting pigments appear black.

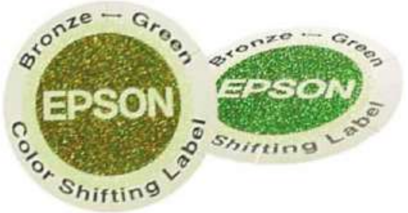

Figure 26 Anticounterfeit label showing liquid crystal coating viewed straight-on and at tilt angles. Courtesy of Flex Products Inc.

A special technology known as linear photopolymerization (LPP) is a process to produce photo-patterned liquid crystal polymer (LCP) $\lambda/4$ phase retarders, which, when combined with a series of layers involving a reflector, and a linear polarizer, shows positive and negative images. Without the diagnostic linear polarizer, however, the image is invisible (Figure 27).

Adding a nematic liquid crystal layer, which exhibits a helical twist configuration, extends this technology. The latter technology can produce colored images when viewed with a linear polarizer. Such devices can be coated or hot-stamped onto a substrate containing viewable information.

Combination Devices

Beginning in 2000, combinations of holographic structures with liquid crystal or interference thin films were developed, which could make counterfeiting even more difficult. Such structures emerge as a hologram floating on a color shifting background. Magnetic layers, integrated into part of the optical interference stack, allow spatial orientation and 3D effects during printing.

Decorative Coatings

Coatings Based on Diffractive Pigments

When pigments based on diffractive effects are created, individual flakes have surface contours of groove-like structures of 2000–3500 lines mm^{-1}. In spray-out coatings using such pigments, diverse interplay of subtle rainbow colors results (Figure 28). The effect is particularly striking in full sunlight as, for example, on curved surfaces found on cars and motorcycles. Some manufacturers have successfully used finely shredded diffraction gratings

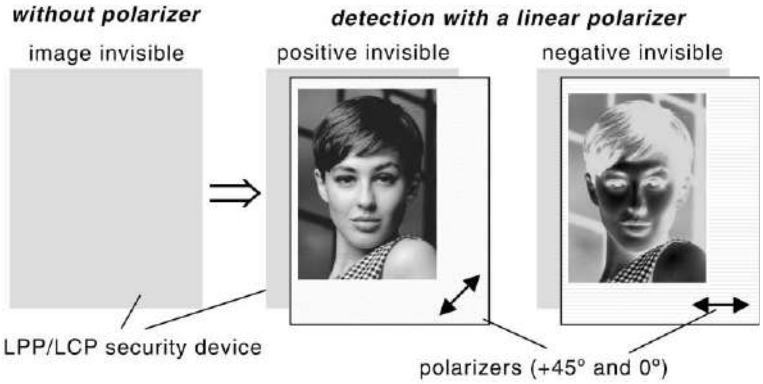

Figure 27 Reflective optical LPP/LCP security device showing a complex LPP/LCP retarder image visualized with one polarizer; rotation of the polarizer changes the image from positive to negative. Reproduced with permission from Rolic Research, Ltd.

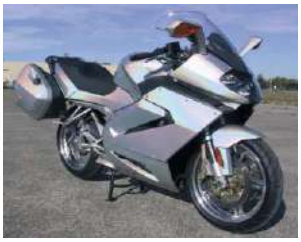

Figure 28 Motorcycle painted with diffraction-based pigments. Courtesy of Flex Products, Inc.

(ca. 25 microns thick) as flakes mixed in paints for decorative purposes. The use of such thick flakes seriously limits the applicability and uniformity of the coating.

Coatings Based on Interference Pigments

Pigment-based interference effects are widely known and have been used for decades in the paint, plastics, and cosmetic industries. These structures are based on TiO_2 coated mica, where the mica thickness is between 300 and 600 nm with planar diameters ranging from 1 to 200 microns, depending on the application.

Depending on the thickness of the TiO_2, different reflective colors of silver, gold, red, blue, and green are achievable. For maximum color reflection, a black substrate or viewing at a particular angle is required so as to prevent dilution of the reflected color from light reflected from the substrate surface. Colored mica pigments are platelets of mica coated with colored oxides. Metallic-like colors are also formed from iron oxide coated micas. In the first instance, two colors may be seen: the platelet color (due to absorption) and the interference color at one angle. Only the interference color may be seen at another viewing angle. If the substrate is colored, the color switches from the interference color to the substrate color as the viewing angle changes. More recently, laminar pigments with an all-dielectric optical stack have been produced with a low-index central core, encapsulated with high-index layers that cause interference.

Interference pigments based on absorber/dielectric/metal/dielectric/absorber have also been introduced into the decorative marketplace. These structures are based on aluminum as the central layer, a dielectric spacer layer, and either a wavelength selective absorber layer such as iron oxide (Fe_2O_3) or a nonwavelength selective absorber such as chromium (Cr). In Fe_2O_3 based pigments, the face-on colors are mainly found in the yellows and reds whereas for the Cr-based pigments, a full color gamut is possible.

Diffractive Foils

These foils are based on diffractive and grating surfaces where the contours are generally less than 2000 lines mm^{-1}. These materials show a glittering effect, i.e., multiple points of color, as the viewing angle is changed.

Interference Polymeric Films

These films are made from multilayer extrusions of pairs of polymers of 50 or more layers. Each pair is composed of two different polymers with a slightly different refractive index. Each layer is approximately one-quarter wave in thickness. Due to the small refractive index differences between layers, the resultant film is approximately 25 microns thick. Uniform color at a given viewing angle is generally not possible due to the lack of thickness control. Hence, they tend to look like oil slicks, with each color changing as the viewing angle changes (Figure 29). Interference polymeric films have found use as decorative gift wrapping material.

Holographic Foils

As indicated earlier, holograms designed for non-security markets may range from simple 2D holograms to more complex ones that function for either decorative or brand protection purposes. Another application of holographic foils is in holographic fabrics, whereby small holograms approximating several millimeters in diameter are affixed in a pattern to the cloth (Figure 30).

Aluminum Flake Based Inks and Coatings

These inks are composed of leafing or nonleafing aluminum powders and a polymeric vehicle. The use of leafing pigments produces images and coatings of high brilliance, whereas nonleafing pigments produce a matte or satin-like finish.

Figure 29 Visual effect from multilayer extruded polymeric film. Reproduced with permission from Englehard Corporation.

Summary and Challenges for the Future

Optical scientists have shown tremendous creativity in discovering and employing clever technical devices in the fight against counterfeiting of valuable documents. Optical solutions have been created that meet the functional as well as the economic needs of the document security industry. A cadre of research institutions and manufacturing companies has accepted the challenge of protecting the instruments of our economic and physical security.

This vigilance will have to continue in the future. Nearly as fast as new defenses are developed, criminals are illicitly using new tools to defeat the protection schemes. Financing for sophisticated counterfeiting efforts is likely to increase, whether sourced by international terrorists or drug dealers. Likewise, the more we move toward an information society, the more we depend on machines to count, sort, and verify documents and money. This machine-driven trend will mandate new features, especially in banknotes and cards.

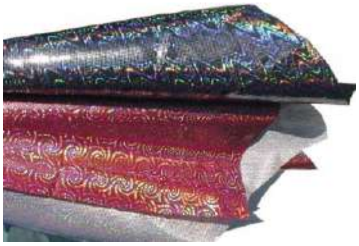

Figure 30 Fabrics showing holographic effects. Reproduced with permission from Amagic Holographics, Inc./K-Laser Technology, Inc.

It is likely that the future of optical anticounterfeiting technology will be strongly influenced by:

- Active, rather than passive, devices. That is, devices that can be applied to a document and later stimulated or interrogated by a machine to elicit a specific effect or signal.
- Schemes that may be merged with new biometric technologies, especially for individualized documents, such as passports and identification cards. In such applications, it will be necessary to confirm that a document is genuine, and that it is in the possession of its rightful owner.
- The need to clearly differentiate security applications from decorative or aesthetic uses. Effective security products must always be carefully controlled and their availability limited.
- The willingness of the public to accept regular changes in the appearance of value documents. Currency will have to undergo redesign on a three-to-six year cycle, rather than the long cycles typical in the past.

This chapter was intended only as a survey of the issues involved in anticounterfeiting protection. To learn more about the optical methods discussed here, or to learn more about the science of anti-counterfeiting, please consult the "Further Reading" section and the Internet sites listed in Table 2.

Table 2 Internet sites: Banknote Design & Protection

Sponsor	*Web Address*
Amagic Holographics, Inc.	http://www.amagic-usa.com/
American Bank Note Holographics, Inc.	http://www.abholographics.com/
Bank of England	http://www.bankofengland.co.uk/banknotes/index.htm
Bundesdruckerei GmbH	http://www.bundesdruckerei.de/en/products/index.html
Crane & Co.	http://www.crane.com/specialty_papers/default.asp
Engelhard Corporation	http://www.engelhard.com/
European Central Bank	http://www.ecb.int/
Flex Products, Inc.	http://www.ocli.com/markets/flex_products.html
Giesecke & Devrient	http://www.gdm.de/eng/main/home/index.php4
Graphic Security Systems Corp.	http://www.graphicsecurity.com
Hologram Industries	http://www.hologram-industries.com/pages/index.htm
Joh Enschedé	http://www.joh-enschede.nl/divisies/jess-engels.html
K-Laser Technology, Inc.	http://www.klaser.com.tw
Light Impressions International, Inc.	http://www.lightimpressions.com/
November AG	http://www.november-ag.de
OVD Kinegram Corporation	http://www.kinegram.com/
Papierfabrik Louisenthal GmbH	http://www.louisenthal.de/
Rolic Research, Ltd.	http://www.rolic.com/
SICPA Security Inks	http://www.sicpa.com/default.asp?lang=e
Thomas De La Rue	http://www.delarue.com/
United States Bureau of Engraving	http://www.bep.treas.gov/

The authors express their gratitude to Ms. Sue Spadini for the technical editing of this work, and to the many individuals and companies who provided exemplars and illustrations. Please see the figure captions and the reference lists for their names.

List of Units and Nomenclature

2D	two-dimensional
3D	three-dimensional
CIE	Commission Internationale de l'Eclairage; International Commission on Illumination
CIELAB color space	(L^* = Lightness; $a^*.b^*$ = chroma coordinates)
CMYK	cyan-magenta-yellow-black
QW	quarter wave

See also

Holography, Techniques: Color Holography; Holographic Interferometry; Holographic Recording Materials and their Processing; Sandwich Holography and Light in Flight. **Microscopy:** Interference Microscopy. **Optical Coatings:** Thin-film Optical Coatings. **Polarization:** Introduction; Matrix Analysis.

Further Reading

Committee on Next-Generation Currency Design, National Materials Advisory Board (1993) *Counterfeit Deterrent Features for the Next Generation Currency Design*. Washington, DC: National Academic Press.

Demus D, Goodby J, Gray GW, Spiess HW and Vill V (eds) (1998) *Handbook of Liquid Crystals*. Weinheim, Germany: Wiley-VCH.

Fagan WF (1990) *Optical Security and Anticounterfeiting Systems*. Volume 1210. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Faris SM (1994) US Patent 5,364,557: *Aligned Cholesteric Liquid Crystal Inks*.

Glausch R, Kieser M, Maisch R, *et al.* (eds) (1998) *Special Effect Pigments*. Hanover, Germany: Vincentz Verlag.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Phillips RW and Bonkowski RL (2000) *Security Enhancement of Holograms with Interference Coatings*. Volume 3973. SPIE – The International Society for Optical Engineering. Bellingham, WA.

Phillips RW and Nofi M (1999) *Colors by Chemistry or by Physics?* Society of Vacuum Coaters 42nd Annual Technical Conference Proceedings, pp. 494–499, Chicago, IL.

Phillips RW, Nofi M and Slusser R (1994) *Color Effects From Thin Films Designs*. Eighth International Conference on Vacuum Web Coating, pp. 270–284. Las Vegas, NV.

Proceedings of the International Conference on Security Documents for the 21st Century (1987) San Diego, CA. Prepared by Oak Ridge National Laboratories, CONF-87040175, Contract DE-A-C05-840R2140, Oak Ridge, TN 37831.

Rancourt JD (1987) *Optical Thin Films, Users' Handbook*. New York: Macmillan.

US House of Representatives (1984) *The Currency Design Act*. Hearing Before the Sub-Committee on Banking, Finance, and Urban Affairs, Second Session on H.R. 6005, Serial No. 98-98. Washington, DC: US Government Printing Office.

Van den Assem B, Brongers D, Rath J, *et al.* (1994) *Practical Guide for the Security Documents Printing Sector.* The Hague: SDU Uitgeverij Koninginnegracht.

Van Renesse RL (1994) *Optical Document Security.* London: Artech House.

Van Renesse RL (1996) *Optical Security and Counterfeit Deterrence Techniques*. Volume 2659. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Van Renesse RL (1998) *Optical Security and Counterfeit Deterrence Techniques II*. Volume 3314. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Van Renesse RL and Vliegenthart WA (2000) *Optical Security and Counterfeit Deterrence Techniques III*. Volume 3973. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Diamond Optical Devices and Coatings

D M Aslam, Michigan State University, East Lansing, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

What do a human, a plant, the photoresist used in microfabrication, and a polymer have in common? Why has carbon become a crucial part of all known living systems? The answer to these questions may be embedded in carbon's hybrid bonding structure (sp, sp^2, and sp^3 C–C bonds). Carbon has 4 electrons in its outermost shell (the valence shell), which is the second shell. At first sight, one would expect that these 4 electrons are distributed among its one 2s and three 2p orbitals. However, it is well known that, in the diamond lattice, each carbon atom bonds with its four neighbors. Are the electrons in 2s and 2p orbitals responsible for the four bonds? One possible answer is that the 2s, $2p_x$, $2p_y$, and $2p_z$ orbitals of the carbon atom form bonds with their neighbors. There are two problems with this answer. First, these bonds will have different strengths (s and p orbitals), which is in contradiction with the experimental fact indicating an equal strength for all the four C–C bonds in the diamond. Second, the computed strength of such bonds will be lower than that found experimentally. What is then the nature of C–C bonds found in organic compounds?

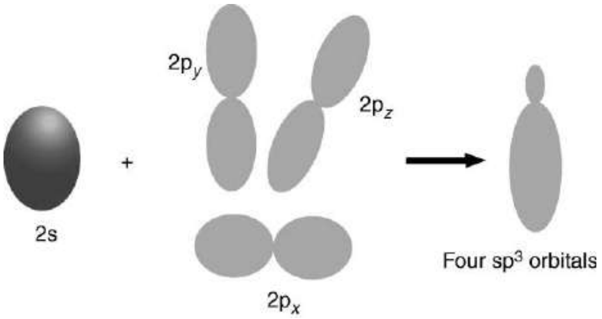

Figure 1 Hybridization of s and p orbitals in carbon to form sp^3 orbitals, which form the so-called sigma bonds responsible for the strong diamond lattice.

The answer to this difficult question was provided in 1931 by Linus Pauling (1901–1994), the winner of two Nobel prizes. With the help of quantum mechanical calculations, Pauling showed how one s orbital and three p orbitals can mix, or hybridize, to form four equivalent atomic orbitals with tetrahedral orientation. The resulting hybrid orbitals are called sp^3 orbitals as shown in Figure 1. The sp^2 or sp^1 orbital results if two or one p orbital combine with the s orbital, respectively, to form the hybrid. While sp^3 C–C bonds lead to a diamond lattice, the sp^2 C–C bonds lead to graphite, and a material which contains both sp^2 and sp^3 C–C bonds is called diamond-like carbon (DLC). The optical properties, such as the bandgap, of DLC depend on the sp^3/sp^2 ratio. If a sheet consisting of a monolayer of graphite, which consists of sp^2 bonds, is wrapped into a pipe, the resulting structure is called a single-wall nanotube (SWNT), as shown in Figure 2a. A multiwall nanotube (MWNT) has many such layers (Figure 2b–c). Other elements such as oxygen (in H_2O) and nitrogen (in NH_3) also form sp^3 orbitals but carbon seems to be only one which is capable of forming sp, sp^2, or sp^3 bonds, depending upon the bonding environment (Figure 3). This may explain why nature has selected carbon

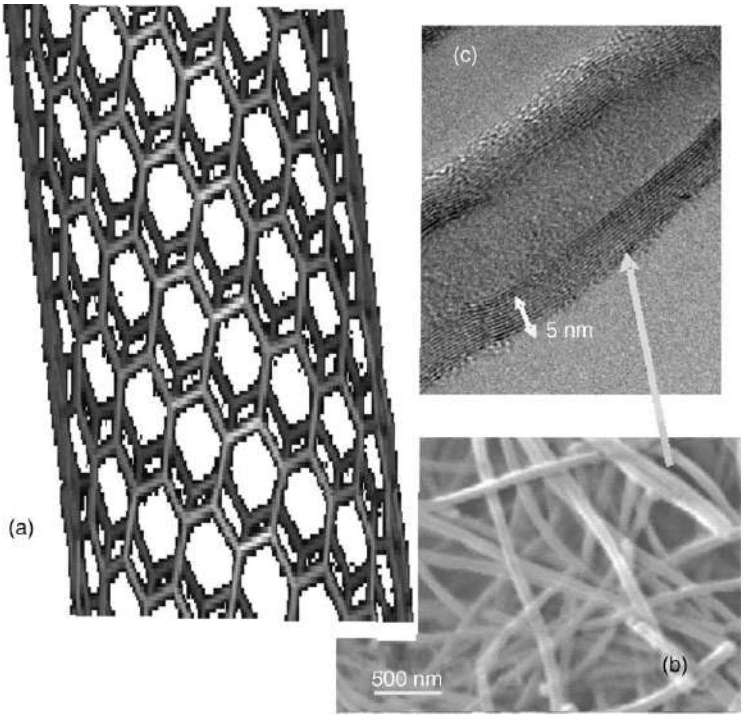

Figure 2 CNT: (a) drawing of a single wall CNT showing sp^2 bonded layer; (b) a mat of multi-wall CNTs; and (c) a TEM of MWNT showing 13 sp^2 bonded layers.

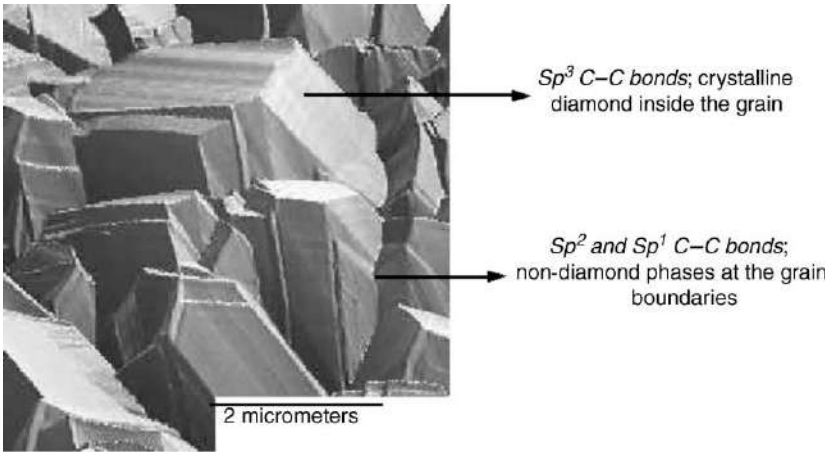
Sp^3 C–C bonds; crystalline diamond inside the grain
Sp^2 and Sp^1 C–C bonds; non-diamond phases at the grain boundaries
2 micrometers

Figure 3 Polycrystalline diamond film; crystalline regions separated by grain boundaries containing non-diamond phases.

Table 1 Comparison of semiconductor properties

Property	*Diamond*		*Si*	*GaA*	*β*-SiC
	Poly-C	*Crystalline*			
Fig. of merit					
Johnson, 10^{23} W Ω s^{-2}		73 856	9	62.5	10 240
Keyes, 10^2 W cm^{-1} s^{-2} °C		444	13.8	6.3	90.3
Sat. electron vel., 10^7 cm s^{-1}		2.7	1	1	2.5
Carrier mobilities, (cm^2 V^{-1} s^{-1})					
Holes	1–165	1600	600	400	50
Electrons		2200	1500	8500	400
Resistivity, Ω cm	10^{-2}–10^8	10^{-3}–10^{13}	10^{-4}–10^8	?–10^8	150
Breakdown field, MV cm^{-1}	0.1–1	10	0.3	6	4
Band gap, eV		5.45	1.12	1.42	3
Dielectric constant	6.7	5.5	11.7	12.5	9.7
Thermal expan. coeff., 10^{-6} °C^{-1}	2.6	1.1	2.6	5.9	4.7
Thermal cond., W cm^{-1} K^{-1}	20	4–22	1.5	0.5	5
Lattice constant, Å		3.57	5.43	5.65	4.36
Density, g/cm^3		3.52	2.32	5.31	3.215
Melting point, °C		4000[a]	1412	1240	2540
Hardness, kg mm^{-2}		10 000	1000	600	3500
Poisson ratio	0.11	0.15	0.23	0.315	
Young's modulus, 10^{12} Pa	0.8–1.2	1.1–1.2	0.155	0.085	0.7

[a] Diamond may convert to graphite well below 4000 °C depending upon the ambient and it oxidizes at 650 °C to form CO_2/CO in O_2.

(perhaps due to its bonding flexibility, to form long C–C chains and rings) as the material of choice for building living systems.

The sp^3 bonding in a diamond lattice makes it a very unique material with a combination of properties not matched by any other known materials (Table 1). In fact, the C–C sp^3 bond in diamond is responsible for its remarkable properties leading to current and potential applications as shown in Figure 4. Although diamond (for centuries) and its bonding structure (for over 70 years) has been known for a long time, the growth of diamond films using chemical vapor deposition (CVD) took a long time to mature. Again, this delay relates to the difficulty with which the diamond lattice with sp^3 C–C bonds is fabricated.

Diamond Growth by CVD

First evidence of diamond growth by CVD, by Eversole in 1952–1953, led to the use of H_2 and CH_4 in the hot filament CVD (HFCVD) to grow diamond on diamond substrates (homoepitaxial growth) by Angus in 1971. The inexpensive CVD polycrystalline diamond (poly-C) was grown on non-diamond substrates by Deryagin in 1976, Spitsyn in 1981, and by Matsumoto *et al.* in 1983. Currently,

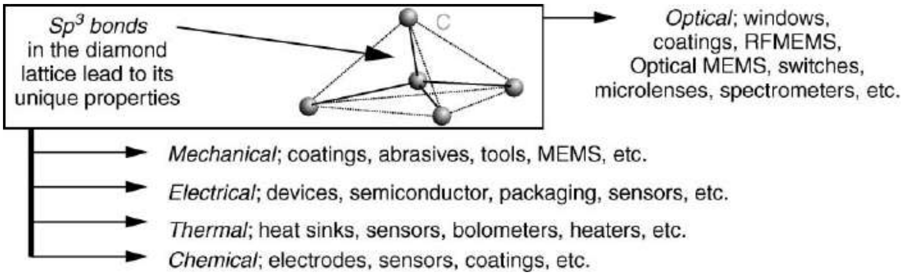
Sp^3 *bonds* in the diamond lattice lead to its unique properties
C
Optical; windows, coatings, RFMEMS, Optical MEMS, switches, microlenses, spectrometers, etc.
Mechanical; coatings, abrasives, tools, MEMS, etc.
Electrical; devices, semiconductor, packaging, sensors, etc.
Thermal; heat sinks, sensors, bolometers, heaters, etc.
Chemical; electrodes, sensors, coatings, etc.

Figure 4 Sp^3 bonds in diamond make it a unique material.

Table 2 Different diamond deposition methods

Methods	*HFCVD*	*MPCVD*	*DC-arc Jet CVD*	*Combustion synthesis*	*Multiple pulsed laser*
Deposition rate (μm/hr)	0.1–10	0.1–10	30–150	4–40	3600
Substrate temperature (°C)	300–1000	300–1200	800–1100	600–1400	50
Deposition area (cm^2)	5–900	5–100	<2	<3	N/A
Advantages	Simple, large area	Quality, stability	High rate, good quality	Simple, high rate	Ultra-high rate
Disadvantages	Contamination, fragile filament	Rate	Contamination, small area	Small area	Expensive

HFCVD, hot filament CVD.

it is believed that, during the CVD of diamond, the CH_3 is responsible for deposition of C as diamond and non-diamond phases. The atomic hydrogen, present in the growth environment, removes the non-diamond phases leaving behind the diamond phase. Near the end of the 1990s, basic science of CVD diamond was well understood, and today diverse plasma and thermal techniques have been developed to produce poly-C films several mm thick and over 12 inches in diameter. Optically smooth 300-micrometer thick undoped poly-C wafers are available on the market. Although there are some reports of n-type poly-C and crystalline diamond growth, the well-established techniques exist only for *in situ* doping of p-type diamond.

A number of diamond growth techniques are currently available for the growth of diamond and carbon nanotubes. Microwave plasma CVD (MPCVD) can be used for the growth of both diamond and carbon nanotubes. As shown in Table 2, the deposition rate of diamond shows a large variation (0.1–3600 micrometers/hour) depending on the growth technique. For applications of diamond in optical coatings, optoelectronic devices and optical systems, the study of transmission, reflection and absorption of light for diamond is very important. These properties of diamond are responsible for its beauty and glamor. For applications in optoelectronic materials, devices and systems, the micro-fabrication and -machining technologies are crucial, particularly in the case of inexpensive poly-C films.

Absorption, Reflection, and Transmission in Diamond

The optical properties, which have been studied for hundreds of years, of naturally occurring single crystal diamond have led to the gem commerce. The fundamental concept involved in the tremendous historical fame of diamond is its high refractive index, n_d (in the range of 2.38–2.64 depending upon the wavelength of light). The refractive index of a material is a measure of decrease of velocity of light (which is also true for all electromagnetic waves) when the wave enters the material. Mathematically, n_d is defined by

$$n_d = V_{vac}/V_d \qquad [1]$$

where V_{vac} and V_d are velocities of light in vacuum and diamond, respectively.

When a beam of light of intensity I_0 in air is incident normally on the surface of diamond, the reflected part of the beam, characterized by the reflection coefficient R, depends on n_d, and R is given by

$$R = \left(\frac{n_a - n_d}{n_a + n_d}\right)^2 \qquad [2]$$

where n_a is refractive index of air and n_d is given by Sellmeier dispersion equation:

$$n_d = \sqrt{1 + \frac{4.3\lambda^2}{\lambda^2 - 0.0121} + \frac{0.33\lambda^2}{\lambda^2 - 0.031}} \qquad [3]$$

where λ is the wavelength of light expressed in micrometers. When the beam enters diamond, part of it is absorbed in diamond. The absorption, characterized by the absorption coefficient α, is caused by: (i) impurities (such as nitrogen, boron, hydrogen, etc.); and/or (ii) intrinsic defects (such as dangling bonds, sp^2 bonds, and sp bonds). However, if the energy of light is close to or above the bandgap of diamond (5.5 eV), the absorption can be caused by electronic transitions from valence to conduction bands. The intensity $I(x)$ at a distance x below the diamond surface is given by

$$I(x) = I_0(1 - R)\exp(-\alpha x) \qquad [4]$$

The fundamental expressions for α are different for indirect and direct type semiconductors. However, the modeling of α, for energies close to the bandgap of diamond, can be accomplished empirically using Urbach's rule according to the following equation:

$$\alpha = \alpha_0 \exp\left[\frac{\sigma(T)(E - E_0)}{kT}\right] \qquad [5]$$

where α_0 and σ are Ubach tail parameters, and E_0 (6.5 eV) is direct gap energy of diamond. K is Boltzmann constant and T is the temperature. As at $T = 300$ K, $\alpha_0 = 4.23 \times 10^{11}$ cm^{-1} and $\sigma = 0.585$, eqn [5] can be simplified:

$$\alpha = 4.23 \times 10^{11}\, e^{22.6(E-6.5)} \qquad [6]$$

where the units for α and E are in cm^{-1} and eV, respectively. Using the expression $E = 1.24/\lambda$, where E is in eV and λ is in micrometers, eqn [6] can be written as

$$\alpha = 4.23 \times 10^{11}\, e^{27.9(1/\lambda - 5.24)} \qquad [7]$$

where λ is in micrometers. The variations of α and n_d with wavelength and energy of light are shown in Figure 5, which are plotted using eqns [3] and [7]. Impurities and intrinsic defects in diamond strongly affect α, n_d, and its other properties.

Impurities in Diamond

A pure diamond with perfect lattice structure is transparent to ultraviolet (UV), visible, infrared (IR), millimeter and microwave regions of the electromagnetic spectrum. However, even an impurity-free diamond has some absorption in the infrared due to phonon excitations and an ultraviolet cutoff due to band-to-band absorption. The defects, impurity-related and intrinsic, in single crystal diamond have been extensively studied because they strongly affect its properties. In fact, diamonds are classified according to the type of defects that are dominant.

Impurities in diamond give rise to additional energy levels in its bandgap. If the density of these impurities is sufficiently high, they can cause absorption of electromagnetic radiation (including light) on one hand and break the lattice symmetry on the other. The breaking of the lattice symmetry can lead to one-phonon IR absorption, which is forbidden in pure diamond. The fact that impurities affect the optical properties of diamond led to its classification into types Ia, Ib, IIa, and IIb which correspond to pale yellow, golden yellow, no color, and blue, respectively. The transmission properties of these various types of diamond are depicted in Figure 6. The dominant impurities are nitrogen

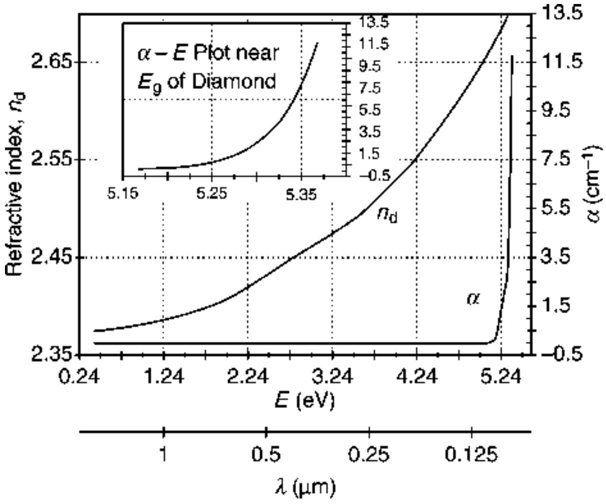

Figure 5 Wavelength and energy dependence of absorption coefficient and refractive index of diamond.

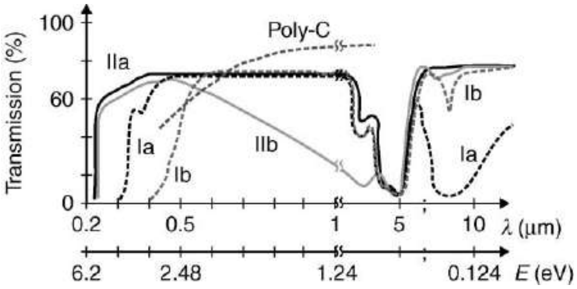

Figure 6 Transmission spectra of natural diamond and CVD polycrystalline diamond (poly-C).

(5–2500 ppm atomic) in type I and boron (10 ppm atomic) in type II. It may be pointed out that boron-doped diamond is a semiconductor with p-type conductivity measured at room temperature. Nitrogen-doped diamond does not show n-type conductivity at room temperature as the donor levels are too deep to be ionized.

Polycrystalline CVD Diamond

Due to the lack of a reliable hetero-epitaxial diamond growth technique, the crystalline diamond films are very expensive (require single crystal diamond substrates for their growth) and, consequently, their widespread commercial use is severely limited. Polycrystalline diamond (poly-C) grown by CVD, which can be as inexpensive as any other material if mass produced, is very attractive for optical and other applications (Figure 2). The properties of high-quality (with high density of sp^3 C–C bonds) poly-C films are comparable to those of crystalline films.

As shown in Figure 6, the transmission properties of poly-C are comparable to or better than those of type I and II diamonds for part of the energy range. Free-standing commercial poly-C films are typically polished and have thicknesses in the range of 50–300 micrometers (the least expensive are 300 μm thick). White poly-C films transparent to visible light are also available. Such films, with optically smooth surface, can be considered free of scattering effects. These films can be considered as a transparent plate in air and, for normal incidence of light with energy less than the bandgap, the transmission coefficient is given by

$$T = \frac{(1 - r_1^2)(1 - r_2^2)}{1 + r_1^2 r_2^2 + 2r_1 r_2 \cos(2\delta)} \qquad [8]$$

where the Fresnel coefficients are given by

$$r_1 = \frac{n_a - n_d}{n_a + n_d}, \qquad r_2 = \frac{n_d - n_a}{n_d + n_a} \qquad [9]$$

where n_a and n_d are refractive indexes of air and poly-C, respectively, and δ is phase change for a poly-C film with a thickness of t:

$$\delta = \frac{2\pi}{\lambda} n_d t \qquad [10]$$

Using $n_a = 1.0003$, $n_d = 2.4$, and $t = 100$ μm, eqn [8] can be written as

$$T = \frac{1}{1.49 - 0.493 \cos\left(\frac{3015}{\lambda}\right)} \qquad [11]$$

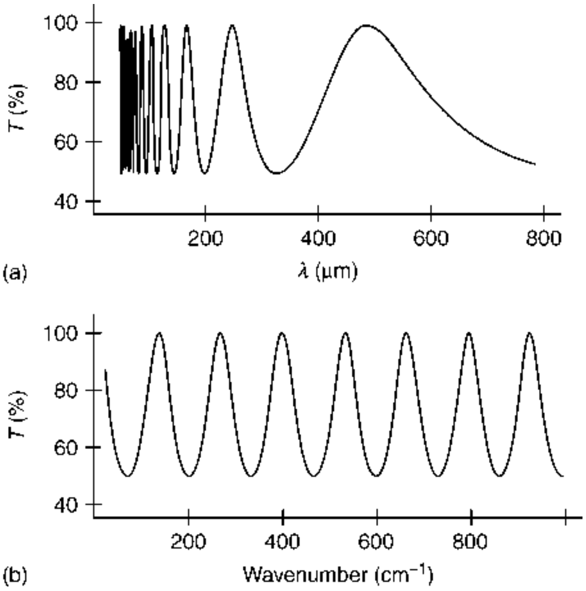

Figure 7 Transmission coefficient of a polished poly-C film with a thickness of 100 μm as a function of (a) wavelength and (b) wavenumber.

or

$$T = \frac{1}{1.49 - 0.493 \cos(0.048\, k)} \qquad [12]$$

where λ is in μm and k, the wavenumber, is in cm^{-1}. Variation of T as a function of λ and k, is shown in Figure 7. A similar behavior is observed for non-polished poly-C films if the wavelength is larger than the surface roughness of the films. For a sample with a very rough surface, scattering effects need to be included in eqn [8].

Electroluminescence in Poly-C

In a luminescence process, the electron must first be excited to higher energy, which can be accomplished by light (photoluminescence), electron beam (cathodoluminescence) or electrical energy (electroluminescence). A subsequent electron transition to lower energies leads to emission of light. A common example of electroluminescence is a forward biased p–n junction made out of direct bandgap type semiconductor such as GaAlAs. The light emitting diodes (LED) and LASERS are based on electroluminescence. Red, green, and blue LEDs are made from GaAlAs, GaP, and SiC/GaN, respectively. Indirect bandgap type semiconductors, such as diamond, typically do not exhibit electroluminescence. The fact that electroluminescence has experimentally been observed in crystalline diamond and poly-C seems to be, at first sight, a contradiction of its indirect bandgap. In fact, defects in diamond are believed to be responsible for their luminescence properties. All types of diamond are found to have a number of defects related to vacancies and impurities. Due to a unique combination of properties of diamond (Table 1), its electroluminescence properties offer a very unique application potential in optical microelectromechanical systems (MEMS), displays and communication devices in IR, visible and UV regions.

Electroluminescence (EL) in optoelectronic devices is produced by excitation of electrons by electronic current. Yellow-green EL has been observed in Schottky diodes made of boron-doped poly-C. Blue-green and blue-violet EL has been observed in free-standing poly-C films, with thicknesses in the range of 15–400 micrometers, when the samples were subjected to applied fields in the range of 2 V/μm. The EL peaks are observed at energies in the range of 3–3.85 eV, with the main peak occurring at 3 eV.

Recently, a very interesting type of EL has been observed in poly-C and carbon nanotubes, which seems to have an application potential in displays and optical MEMS. Under applied electric fields in the range of 20–40 V/μm, electrons are known to tunnel out of a poly-C film into vacuum and the process is called field emission. EL has been observed during the field emission process, and this new effect is called field emission EL (FEEL). A typical FEEL spectra is shown in Figure 8. It is believed that FEEL results from direct type defects in poly-C, which are populated during the field emission process. The electrons in these defects, instead of being field emitted, make transitions to lower energy levels leading to light emission. The main FEEL peak is observed at 2.6 eV which is in the visible range.

Poly-C Applications

The current applications of diamond, which include optical windows, temperature sensors, heat sinks, cutting tools, surgical instruments, and optical coatings, are based on natural, synthetic, and CVD diamond. The cost of producing diamond has always been prohibitive in widespread diamond applications. Fortunately, over the past 20 years, the development of CVD diamond technologies has eradicated the fears of diamond being too expensive for its widespread applications. As the CVD poly-C can be as inexpensive as any other electronic material, widespread applications of diamond are expected for poly-C. In the area of micro- and nanostructures, the poly-C can provide a unique opportunity for both polycrystalline and crystalline microdevices. Using the present fabrication techniques, the grain size in poly-C can be varied in the range from a few tens of nanometers (nanocrystalline diamond) to hundreds of micrometers (commercial free-standing poly-C films). This leads to a whole range of device structures ranging from inter- to intra-grain devices. An intra-grain device made in poly-C will be an inexpensive single crystal device with properties better that those found in natural/synthetic diamonds, because an intra-grain poly-C device will have a higher degree of purity (because it is grown by CVD).

The applications of poly-C in microsystems (current, in-progress, and future) include optical MEMS, biological/protective/hard coatings, sensors,

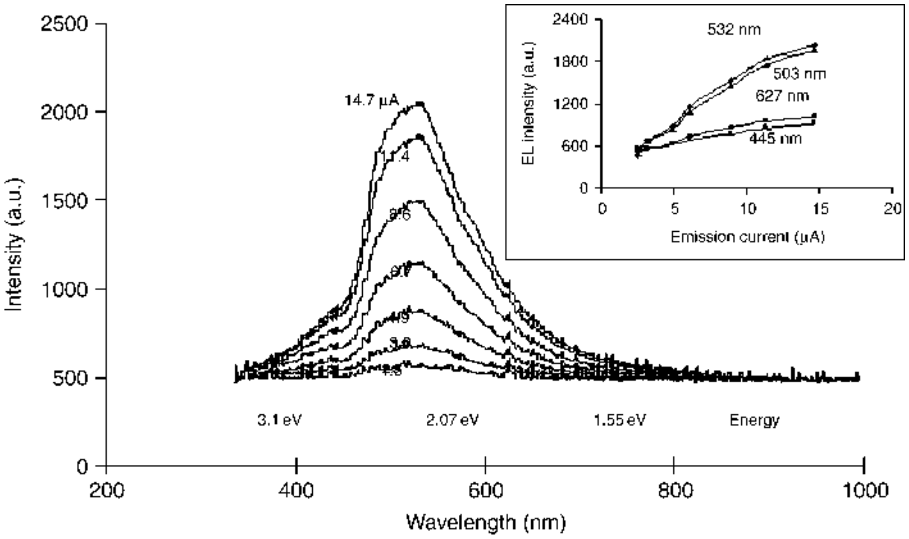

Figure 8 FEEL spectra in poly-C as function of wavelength and emission current (inset).

field emission displays, and WIMS (wireless integrated microsystems). Intra-grain microlenses, RFMEMS, heat pipes, sensors, microabrasives, and spectrometers may lead the way for unique poly-C applications.

Optical MEMS

A basic MEMS structure, with electrostatic actuation mechanism, is shown in Figure 9. Currently, most of the Optical MEMS and RFMEMS (radio frequency MEMS) devices are made from poly-Si (polycrystalline Si) using surface or bulk micro-machining fabrication techniques. The ability to integrate micro-optical elements with movable structures and micro-actuators has opened up many new opportunities for optical and optoelectronics systems, as MEMS allow a more effective manipulation of optical structures than the conventional methods. Optical MEMS have applications in displays, sensing, data storage, and communication systems. Some of the examples are optical switches, optical cross-connects, wavelength division add/drop mixers, tunable filters/lasers/detectors, dispersion components, polarization dispersion components, and spectrometers.

Material-related limitations put some constraints on the current optical MEMS technologies and applications. Diamond is an excellent material for optical MEMS and can extend the application range and reliability. The use of poly-C in optical MEMS can also remove the traditional cost constraints associated with the use of single crystal diamond. Although the poly-C optical MEMS are in their infancy, recent developments are very exciting. Figure 10 shows a poly-C resonator structure fabricated using a poly-C fabrication technology compatible with Si processing techniques. Such structures, if used as filters and switches, can lead to a superior performance as compared to any other currently used material. Poly-C on-chip spectrometers are also expected to operate in a large wavelength range.

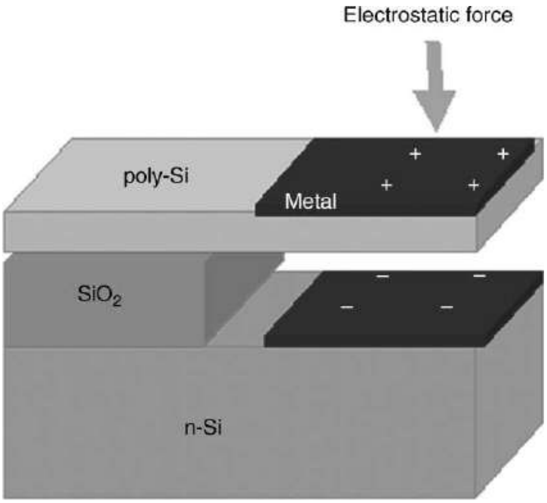

Figure 9 A simple MEMS structure with electrostatic moving of the cantilever beam.

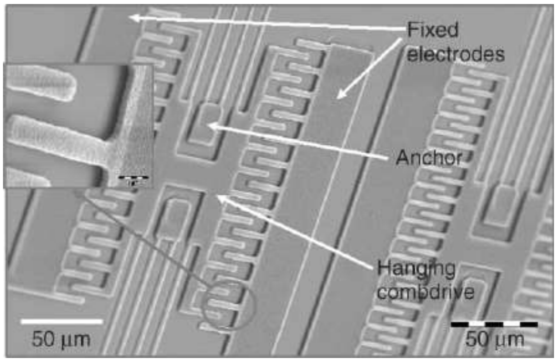

Figure 10 A poly-C MEMS combdrive for high-quality factor resonators, inertial sensors, optical MEMS, etc. Poly-C is superior to poly-Si as its Young's modulus is 4 times higher than that of poly-Si.

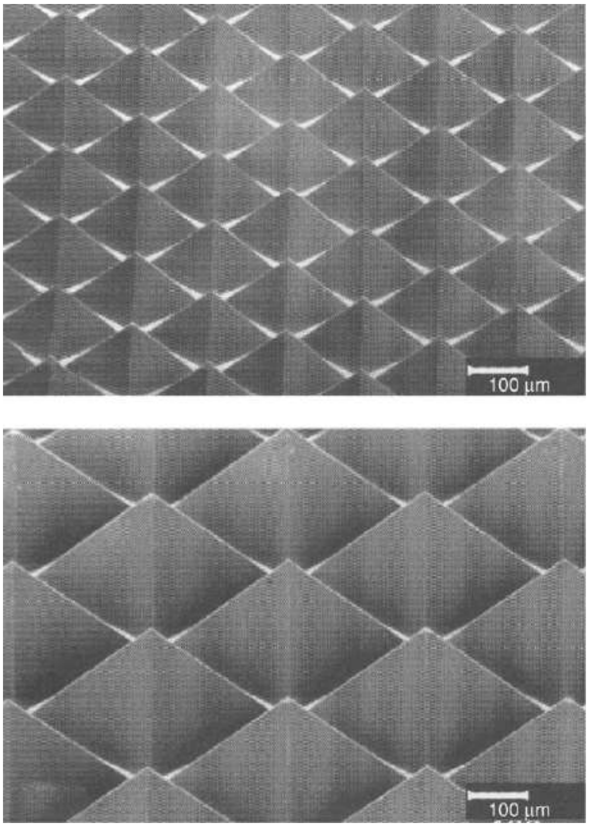

Figure 11 Poly-C microabrasive structures two different densities of pyramids. (Courtesy of Dr. H. Bjorkman of Upsala University.)

Microabrasives

Poly-C is also a very promising material for microabrasives. Micromachined poly-C MEMS structures have recently been fabricated for possible use for microabrasives. Figure 11 shows micromachined MEMS structures which are expected to enhance the life of a tool. Such abrasives by design are inherently cost-effective because they are batch fabricated.

See also

Light Emitting Diodes.

Further Reading

Asmussen J and Reinhard D (2002) *Diamond Handbook*. New York: Marcel Dekker.

Kim U and Aslam DM (2003) Field emission electroluminescence on diamond and carbon nanotube films. *Journal of Vaccum Scientific Technology B* 21(4): 1291–1296.

McMurry J (1996) *Organic Chemistry*. New York: Brooks/Cole.

Pan LS and Kania DR (1996) *Diamond: Electronic Properties and Applications*. Boston, MA: Kluwer.

Spear KE and Dismukes JP (1994) *Synthetic Diamond: Emerging CVD Science and Technology*. New York: Wiley.

Laser Damage in Thin Film Coatings

D Ristau, Laser Zentrum Hannover e.V., Hannover, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the very beginning of laser technology, Laser-induced Damage Thresholds (LIDT) of optical components were always an obstacle for the development of laser systems operating at high power levels. In such systems, the surface or the bulk of the optical components can be damaged by laser radiation of sufficiently high power or energy density resulting in an uneconomical standstill of the laser and its application. In the early days of high power laser development, mainly inclusions in laser rod materials were discussed as a major complication for an augmentation of the output power in solid state laser systems. Nowadays, in the course of the development of optical materials with excellent quality and power handling capability, the problem of laser-induced damage has shifted from the bulk to the surface of the optical component. The optical surface is subjected to various production steps and environmental influences, which modify its structure and composition. Especially, the thin film coating, which is deposited on the optical surface to adapt its reflectance and transmittance to the application, contributes predominantly to the reduction of the LIDT values. As a consequence, the measurement and optimization of the power handling capability of thin films is considered as one of the primary research areas in modern optics technology and is supported by an extensive scientific community.

In the following, a brief review will be given on selected fundamental damage mechanisms in thin films, considering also the scaling of LIDT values for different operation conditions. Also, current standards for the measurement of LIDT will be described, and examples illustrating some practical aspects of high power optical coatings will be presented. Finally, a summary of the present state of the art will be given and recent trends in laser technology will be discussed in respect to research in laser-induced damage.

Fundamental Damage Mechanisms in Thin Films

During thirty years of intense research in laser damage of thin films, a broad spectrum of different damage and degradation mechanisms had been discovered and studied for a variety of laser operation conditions and wavelengths. For prominent high-power laser systems often applied in industrial production fields or fundamental research, specific models for laser-induced breakdown have been developed, which are also of practical relevance. In many experiments, damage sites with the morphology illustrated in Figure 1 are observed. The coating is delaminated from the surface of the optical component in the center of the laser beam area. Obviously, mechanical stress has built up in the coating by heating, which is induced by the absorbed laser power. At a certain stress level, exceeding the adhesion strength of the coating to substrate surface, the thin film cracks and may even delaminate from the component. In other cases of this absorption-induced damage effect, the film reaches its melting point prior to the mechanical threshold and

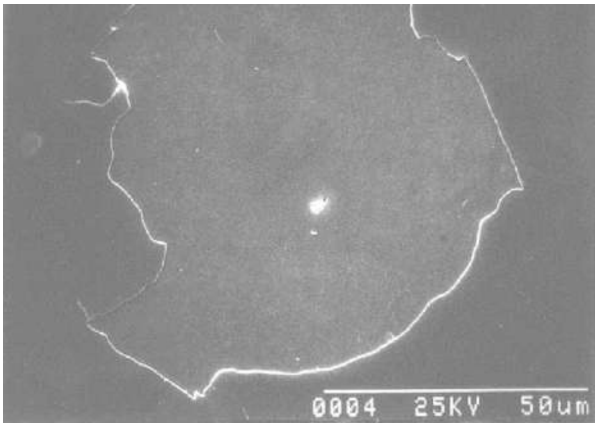

Figure 1 SEM-picture of a damage site on an alternating coating of SiO_2/HfO_2. The system was irradiated by a Nd:YAG-laser with an energy density of 45 J/cm^2. As a dominant damage mechanism, a delamination of the coating by extreme mechanical stress is suggested. Reproduced with permission from Ristau D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.49.

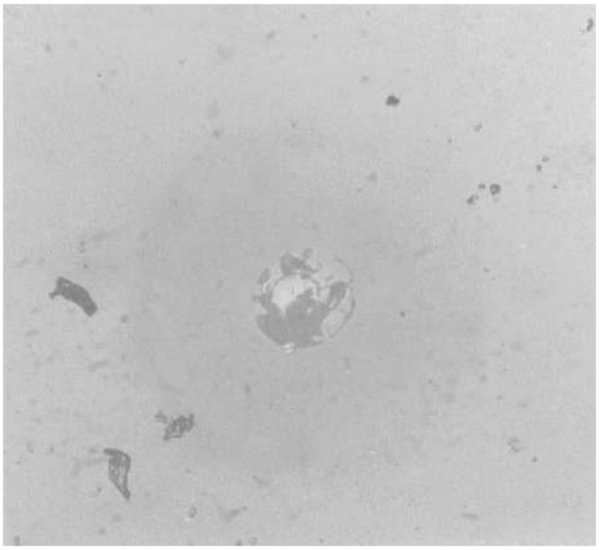

Figure 2 Nomarski micrograph (magnification 250×) of a damage site on an antireflecting coating of SiO_2/Ta_2O_5. The system was irradiated with an energy density of 51 J/cm^2 at a wavelength of 1.064 μm. A delamination of the coating in the center of the beam is surrounded by a recrystallization zone.

evaporates or changes its crystalline structure. The morphology of such damage sites exhibits a discoloring or an increased surface roughness in the center of the laser beam (see Figure 2). Absorption-induced damage is dominated by the instantaneous heating of the coating material in the area of interaction with the laser beam and can be described on the basis of the heat diffusion equation. In this approach, the source term in the heat diffusion equation is given by the heat coupled into the system by the absorbed laser power, and the boundary conditions are determined by the geometry and the structure of the layer system as well as the spatial laser beam profile. Apparently, absorption-induced damage has to be considered only for components with significant absorptance at their operation wavelength. Typical examples for coating materials and wavelength regimes, where absorption-induced damage is frequently observed, are the wavelength 10.6 μm of the CO_2-laser with ZnSe, ZnS, and fluorides as coating materials, or the DUV/VUV-wavelengths dominated by excimer lasers and fluoride coating materials. The effect of absorption-induced damage in optical materials is also of practical relevance for cw- and long-pulse operation conditions of the laser system corresponding to long thermal diffusion lengths. For cw-irradiation, the temperature rise in the component can be calculated by numerical methods involving finite elements or differences. For boundary conditions with circular symmetry and a Gaussian beam profile (beam diameter w, power P), an analytical expression can be derived for the temperature rise ΔT in the center of an irradiated circular component

$$\Delta T = \frac{2\beta_s P}{\pi^{3/2} k w} \tan^{-1}\left(\frac{16\kappa t_I}{w^2}\right)^{1/2} \qquad [1]$$

In this model the temperature rise is dependent on the thermal properties (k: thermal conductivity, κ: thermal diffusivity), on the surface absorption of the component (β_S), and on the beam diameter. For irradiation times t_I, long compared to the typical heat diffusion time w^2/κ, eqn [1] reduces to the asymptotic dependence

$$\Delta T \rightarrow \frac{\beta_s P}{\pi^{1/2} k w} \qquad t_I \rightarrow \infty \qquad [2]$$

Obviously, the temperature rise ΔT scales with the P/w for these long-time irradiation conditions compared to the short pulse regime, where the scaling of the temperature rise is proportional to P/w^2. For the application of optical components in high power cw- and long-pulse lasers, this P/w scaling law has to be considered carefully, because the power handling capability depends critically on the size of the irradiated area. Also, the onset of damage may be delayed in respect to the start of irradiation by time constants in the range of several 10 seconds.

In the course of the development of improved coating processes, optical absorptance could be reduced to very low levels in the near infrared and visible spectral region. For example, high reflecting mirrors for the wavelength of the Nd:YAG-laser with total losses below 1 ppm can be produced with advanced ion beam sputtering processes nowadays. Therefore, another principal damage mechanism, which is based on inclusions or defects in the coatings, is often found in the near infrared spectral range. A detail of a damage site representing inclusion dominated damage induced by a pulsed Nd:YAG-laser in a coating of HfO_2/SiO_2 is shown in Figure 3. In the laser beam area numerous craters are located, which extend from circular voids with different diameters in a range below 1 μm. At the high magnification of the employed electron microscope, the multilayer structure of the coating system can be resolved indicating an origin of damage approximately in the same depth of the layer structure at the substrate surface. The underlying damage mechanism is initiated by inclusions or other defects with an absorption, which is significantly higher than for the surrounding thin film material. As a consequence, the inclusion is rapidly heated during the interaction with the laser beam and explodes removing the covering layer structure. This inclusion breakdown can be modeled by calculating the temperature behavior $T(r)$ at the radial position r of a sphere (radius g, thermal conductivity k_E, diffusivity κ_E) which is embedded in a medium with defined properties (see eqn [3]). The interaction of the sphere, which has a size in the range of the wavelength λ and a refractive index n', with the laser irradiation, is described by the Mie absorption coefficient Q. For the instant of damage, a critical temperature at the outer surface of the sphere ($r = g$) is assumed, which is often related to the melting point of the layer material

$$T(r) = \frac{3QI}{4\pi k_E g}\left[\frac{k_E}{3k} + \frac{1}{6}\left(1 - \frac{r^2}{g^2}\right) - \frac{2gb}{r\pi}\int_0^\infty dy \exp\left(-\frac{y^2 t_p}{\tau_1}\right) \times \frac{(\sin y - y\cos y)(\sin ry/g)}{y^2\left[(c\sin y - y\cos y)^2 + b^2y^2\sin^2 y\right]}\right] \qquad [3]$$

$$Q = \pi g^2\left[1 + \frac{2\exp(-8\pi n'g/\lambda)}{8\pi n'g/\lambda} + \frac{\exp(-8\pi n'g/\lambda) - 1}{(8\pi n'g/\lambda)^2}\right]$$

$$\tau_1 = \frac{g^2}{\kappa_E} \quad b = \frac{k}{k_E}\left(\frac{\sqrt{\kappa_E}}{\sqrt{\kappa}}\right) \quad c = 1 - \frac{k}{k_E}$$

Besides the thermal properties of the host material (thermal conductivity k, diffusivity κ), the properties and the statistic distribution of the inclusions have to be known for a quantitative description of inclusion breakdown.

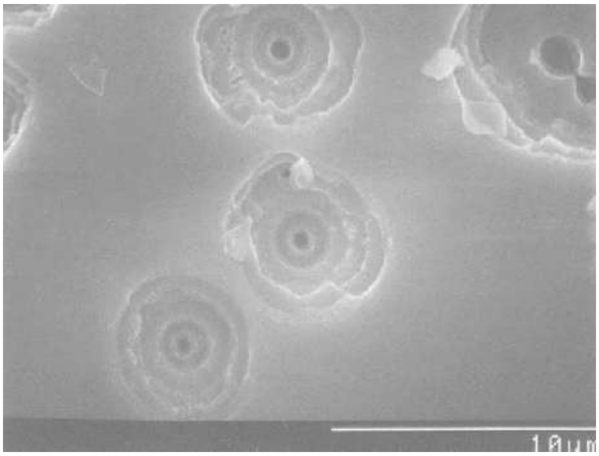

Figure 3 SEM-picture of damage site on a high reflecting coating of SiO_2/HfO_2. This inclusion-dominated breakdown mechanism was initiated at an energy density of approximately 130 J/cm^2 with a Nd:YAG-laser.

In contrast to the nearly deterministic behavior of absorption induced damage, defect dominated damage mechanisms often exhibit a statistical nature. Since the inclusions are not homogeneously distributed in the layer structure, a variation of the threshold value is observed for different damage sites on a single sample. The assessment of the inclusion parameters is extremely difficult in practice. Therefore, the inclusion model is mainly employed for the prediction of tendencies for the LIDT as a function of the beam parameters and the properties of the layer materials.

Different aspects related to the substrate polishing, cleaning, and deposition process have to be considered for an investigation in the various origins of inclusions or defects in a layer structure. Besides contamination of the substrate surface prior to the deposition process, particles may be generated during the deposition process by mechanical abrasion of material from moving parts in the plant or by delamination of material coated on the inner walls. Also, the deposition source may contribute to the formation of defects by sputtering of particles or droplets from the melt or the target. Especially in low energetic thermal deposition processes, defects are often the origin for the formation of nodules. These conical faults in the layer structure reduce the LIDT because they are less resistant against intensive laser radiation than the bulk of the coating system.

Besides models, which are based on the transformation of laser radiation into heat, direct electronic excitation has to be considered as a breakdown mechanism for short-pulse lasers. In these intrinsic breakdown models for dielectric materials, the generation of a high electron density in the conduction band is assumed. This carrier generation can be explained by high field ionization of electrons in the valence band of the dielectric material. Many theoretical concepts introduce the avalanche effect as an important mechanism contributing to the generation of free electrons. Electrons excited by strong field ionization can interact with the electric field and accumulate excitation energy. By collisions with the lattice, this energy can be also transferred to the electrons in the valence band. Thus, when the energy of the free electrons exceed a critical value, other electrons can be excited to the conduction band, and a carrier generation like an avalanche arises. Another theoretical approach is the multiphoton model which is formulated on the basis of multiphoton absorption for the generation of free electrons. In most intrinsic models it is assumed, that laser-induced damage breakdown occurs when the electron density in the conduction band reaches a critical value of approximately 10^{21} $1/\mathrm{cm}^3$. At this high electron density, a plasma state is reached and the impinging radiation is coupled efficiently into the layer structure resulting in its destruction.

Intrinsic damage is often characterized by a deterministic damage behavior with a well defined threshold value which is also characteristic for damage in bulk materials. As a prominent example, damage induced by ultra-short pulses with a duration below 1 ps can be often attributed to intrinsic mechanisms, because the diffusion length of thermal effects can be neglected in comparison to the intensive interaction of the laser radiation with the electrons of the coating materials. In this case, the generation rate $\mathrm{d}n/\mathrm{d}t$ of electrons in the conduction band can be described on the basis of the impact ionization rate and multiphoton excitation

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \alpha n(t) q I(t) + \beta_m \left[q I(t) \right]^m \qquad [4]$$

In this model $I(t)$ is the power density of the laser radiation, and q is a correction factor representing interference effects in the coatings, which may result in a local power exceeding the power density of the radiation impinging on to the layer system. The electron excitation mechanisms are described by the impact ionization coefficient α and the m-photon absorption coefficient β_m. The relative contributions of both excitation mechanisms to laser-induced breakdown are depending on the pulse duration and the bandgap energy of the layer material.

In a rough approximation, the damage threshold increases with the bandgap energy of the materials as a consequence of the photon absorption coefficient β_m decreasing with its order m. For short pulse durations below a few 10 fs, multiphoton processes contribute significantly in the range of some 10% to the free-carrier generation. In most materials, avalanche ionization develops to the dominant generation mechanism for pulse durations above 100 fs. These tendencies could be demonstrated for single layers of selected oxide material deposited by ion beam sputtering on quartz substrates, which were tested in a pulse duration regime from 20 fs to 1 ps. According to the theoretical model, the expected increase of the damage also influencing values with the pulse duration was observed in the damage experiments.

In practice, the power handling capability of optical components is also often limited by imperfections like scratches, digs, and areas with high roughness on the optical surface. In most deposition techniques applied in optical technology, the layer system tends to replicate or even enhance surface imperfections of the substrate. In the application, additional weak points may be introduced by improper handling or contamination of the optical surfaces. If the contaminants are not removed by the impinging high-power laser radiation, they increase the surface absorption and act as initiation points for laser-induced damage. Voids, grooves, pores, or scratches reduce the power capability of the optical element, because they act as concentrators for the electric field.

Units and Scaling of Laser-Induced Damage Threshold

According to the theoretical models of laser-induced damage, the appropriate units of measurement for LIDT values are mainly given by the dominant damage mechanism and the irradiation time. For pulsed laser radiation and dielectric breakdown effects, the laser-induced damage threshold is usually reported in $\mathrm{W/cm}^2$. In the case of absorption-induced damage or inclusion dominated breakdown and a pulse duration short compared to the thermal diffusion time in the layer structure, LIDT values are often scaled in $\mathrm{J/cm}^2$. The unit of linear power density (W/cm) is indicated for the LIDT of long-pulse, cw-lasers or other sources, which induce a temperature increase in the optical component with rise times in the regime of seconds.

Laser-induced damage thresholds of optical thin films are critically dependent on the operation conditions of the applied laser system and on the design of the layer structure. For example, even the damage mechanism can change if the wavelength or the pulse duration is varied. Therefore, LIDT values should be scaled with the irradiation parameters only for small intervals, in cases where the fundamental damage mechanism is clarified. In all other situations, the extrapolation of threshold values is extremely difficult and may lead to inaccurate results. Especially, an overestimation of the LIDT value holds severe dangers in practical applications, where the replacement of a damaged component and the repair of other parts impaired by the damage event causes high expenses. If the damaged components consist of toxic materials (e.g., ZnSe, GaAs, CdTe, ThF_4, chalcogenides, Be, Cr, etc.) severe health hazards may occur, and an expensive decontamination of the environment may be necessary. Therefore, the determination of the laser-induced damage threshold using the actual laser parameters under controlled environmental conditions is always recommended for unclear cases.

In the present state of research in laser-induced damage mechanisms, only few tendencies are generally accepted as confident for scaling of LIDT values. As a function of the pulse duration, laser-induced damage thresholds increase for longer pulses. For inclusion and absorption dominated breakdown, a $t_p^{1/2}$-law is often used in the pulse duration regime between 10^{-10} to 10^{-8}s. This dependency can be extended to other pulse regimes up to pulse lengths in the ms-range, if the exponent is replaced by a value between 0.5 to 2. In respect to the laser wavelength, a decrease of the LIDT value with decreasing wavelength is observed for most materials and operational conditions. Investigations in the influence of the beam diameter have been performed by many research groups, indicating a decrease of the LIDT values for increasing beam diameters. Especially for inclusion dominated breakdown, the event of damage for a certain laser irradiated site will be dependent on the distribution of inclusions at that position. If the beam size is small, the probability for interrogating a defect vulnerable to damage is low. By increasing the spot diameter, this probability will asymptotically reach unity, because at a certain beam diameter, at least one defect will always be covered by the beam. Therefore, the onset of laser-induced damage is not dependent on the beam diameter for inclusion-dominated breakdown, which is often encountered in conventional optical coating systems. Another special case is the scaling of the cw-damage threshold with the beam diameter. Since the cw-LIDT is expressed in linear power density, the power handling capability increases more slowly with the beam diameter for the cw-lasers than expected from the normal pulsed operation. For example, extrapolating the LIDT value of 100 W at a beam diameter of 1 mm for a laser mirror to a beam diameter of 10 mm, results in a maximum power load for the component of 1 kW according to the correct P/w-law. The threshold power would be extremely overestimated to approximately 3.2 kW, if the general P/w^2-dependence is applied.

Measurement of Laser-Induced Damage Thresholds

As a consequence of the complicated relation between the laser damage mechanism and a broad spectrum of thin film properties and laser parameters, laser-induced damage threshold measurements have to be performed under well-defined conditions. To investigate the comparability of LIDT measurements, an extended international round-robin experiment has been conducted on coated optics for the wavelength of 1.064 μm at the beginning of the 1980s. This experiment indicated the need for a clearly specified LIDT-measurement procedure, and conceptual work was initiated to develop a corresponding ISO-Standard series. During recent years, an International Standard (ISO 11254) has been adopted covering testing conditions relevant for most typical laser applications. In the first part of ISO 11254 1 on 1-testing of optical surfaces, in respect to laser damage, is described. The fundamental approach of the standard measurement procedure is illustrated in Figure 4. A laser source operating in transversal and longitudinal single mode is employed for the irradiation of the sample surface. The beam

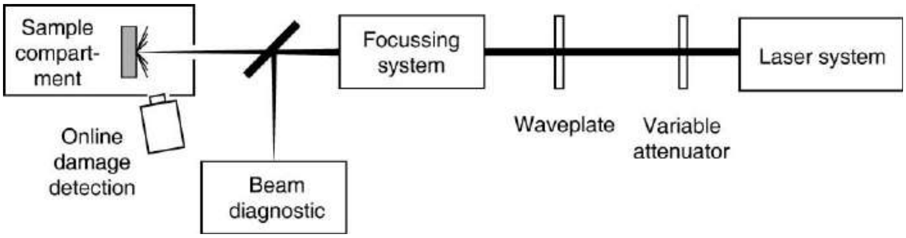

Figure 4 Fundamental setup for the measurement of laser-induced damage thresholds.

parameters of the laser are assessed by a beam diagnostic system which monitors the spatial and temporal profile as well as the energy of the laser radiation in the target plane. In order to achieve the high energy density levels necessary to destroy the surface of the specimen, a well characterized focusing system is installed. For adjustment of the laser energy in the target plane, an attenuating system is employed. In the 1 on 1 measurement protocol, each site on the sample surface is subjected to a single laser pulse once only. In a test sequence, the test surface is examined by pulses of different energies, covering low values without damage and high values causing damage. After the test, the specimen will be inspected with a Nomarski interference contrast or a darkfield microscope at a magnification of 150× or higher to identify the damaged sites. For the evaluation of the damage test, the damage probability method is recommended. In this evaluation scheme, the ratio of the number of damaged sites to the total number of sites objected to a certain energy or power level is interpreted as the damage probability. The plot of these damage probability values as a function of energy or power, which is called the survival curve of the optical component, provides an insight into the damage mechanisms involved. The damage threshold is given by the highest quantity of laser radiation for which the extrapolated probability of damage is zero. A typical example for a 1 on 1 damage test at the wavelength 1.064 μm and the extrapolated LIDT value is depicted in Figure 5.

In most catalogs of optics manufacturers, the 1 on 1 LIDT values are used to illustrate the power handling capability of their products, even though 1 on 1 data are of limited importance for practical applications, where an optical component is always subjected more than one time to a laser beam. This customary situation is covered by Part 2 of the ISO-LIDT-Standard which describes a damage test procedure for a series of pulses (S on 1-tests). For an assessment of the reliability, the concept of the characteristic damage curve is introduced by this standard. This curve is deduced directly from the S on 1 test data by plotting the energy density for a selected damage probability as a function of the number of pulses (see Figure 6). By an extrapolation of the characteristic damage curve to high pulse numbers in the order of 10^9 to 10^{12} shots, the lifetime of the optical component can be roughly estimated. For a certification of optical components in respect to their power handling capability, a third part of ISO 11254 is under development, which is concentrated on different testing protocols of a defined surface fraction at power levels expected in the application. The fundamental approach of these tests is a simulation of conditions at the upper limits of operation parameters encountered in practice.

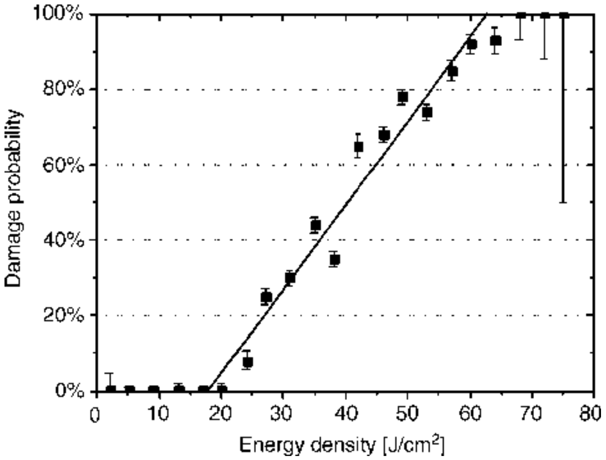

Figure 5 Survival curve of an anti-reflective layer system (two layer V-coating, SiO_2/Ta_2O_5) on fused silica substrate for the Nd:YAG-laser wavelength. The measurement was performed at a wavelength of 1.064 μm, a beam diameter of 420 μm, and a pulse duration of approximately 15 ns, respectively.

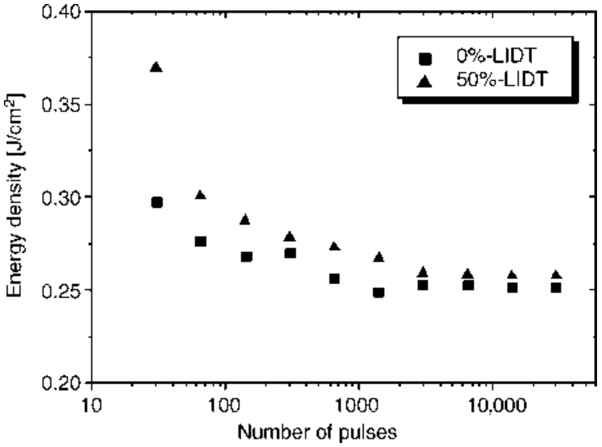

Figure 6 Characteristic damage curve of a high reflecting mirror of SiO_2/TiO_2 for fs-laser systems. The layer structure was tested with a high repetition (1 kHz repetition rate) fs laser system operating at a pulse duration of 150 fs.

Optical Coatings for High Power Lasers

For the development and application of coating systems with high LIDT values, several major aspects have to be considered. In the first approach, appropriate materials and processes have to be selected which deliver coatings with sufficient power resistance. In many studies, a correlation of the LIDT values of the constituent single layers to the power handling capability of a layer system was observed. Therefore, single layers of potential deposition materials were investigated at the wavelengths of prominent laser systems from the VUV- to the FIR-spectral region. For specific deposition processes, often a clear ranking of the coating materials in respect to their power handling capability could be demonstrated. An example for a Nd:YAG-laser with a pulse duration of 15 ns is illustrated by Table 1. With the exception of the ZrO_2-layer, which exhibited severe inhomogeneity and a high inclusion density, a correlation can be assumed between the LIDT value and the melting point of the bulk materials under these operation conditions. This dependency can be attributed to absorption-induced damage as well as to the inclusion model, and therefore, the damage morphology has to be studied in each case to identify the dominating damage mechanism. In the fs-regime or the UV/VUV-spectral range, often a dependence of the power handling capability on the bandgap of the materials is found. According to the relationships of the avalanche and other dielectric breakdown models, materials with highest bandgaps are frequently encountered at the top of an LIDT ranking.

For the influence of the coating process on the laser-induced breakdown of optical coatings, no clear tendencies can be detected. For example, Ion Beam Sputtering (IBS), which is considered as the deposition process for coatings with extremely low losses and contamination, cannot always surpass conventional thermal deposition processes in laser stability. In general, ion or plasma assisted deposition techniques produce coatings with lower LIDT values than conventional coating processes for most operation conditions. Sol-gel processes, which were developed for the deposition of removable coatings on large optics in laser fusion, can reach superior power handling capabilities. In general, as a consequence of the complicated relation between the production parameters and laser-induced breakdown, production processes for optical coatings have to be optimized separately for different wavelength regimes and irradiation conditions.

Table 1 Laser induced damage thresholds of selected single-layer coatings in relation to the melting points of the corresponding bulk materials. In the last column, LIDT values of anti-reflective coatings (AR-Coating) composed of the high index material and SiO_2 are compiled to illustrate the correlation between single LIDT values and the damage threshold of layer systems

Material	*Melting point (°C)*	*LIDT value (J/cm^2) single layer*	*LIDT value (J/cm^2) AR-coating*
TiO_2	1775	13 ± 1	14 ± 1
Ta_2O_5	1918	28 ± 2	32 ± 5
HfO_2	2758	41 ± 3	46 ± 6
ZrO_2	2700	34 ± 4	28 ± 1
Al_2O_3	2072	39 ± 1	
SiO_2	1723	34 ± 7	

Besides the fundamental production parameters, the design of the coating system is of similar importance for the achievement of high LIDT values, because damage is directly driven by the electric field strength in the layer structure. A standing wave field pattern is depicted in Figure 7 for a high reflecting stack of Ta_2O_5 and SiO_2. The design consisting of layers with an optical thickness of 1 QWOT is typical for most standard laser mirrors. The power density reaches extreme values always at the interfaces between the layers and even exceeds the incoming irradiation power density (100%). Interfaces between the layers can be considered as weak points in a coating system, because additional contamination may occur during the switching of the material in the production process. Also, the adhesion between the adjacent layers may be reduced, and mechanical stress may be built up by the different materials. To improve damage thresholds of laser mirrors, the thickness of the first few layer pairs can be adjusted to shift the points of maximum field strength into the bulk of the layer with higher damage resistivity (see Figure 8). Also, a thick layer can be attached to the system in order to stabilize the outer layer pair in respect to thermal or mechanical stress. Another technique, which can be applied to enhance the stability of the interface is the codeposition of materials, resulting in a gradual interface with a mixing zone of materials between the layers. This region of codeposited material exhibits a higher resistance against mechanical stress resulting in an improvement of LIDT values of up to 20%.

The effect of the internal electric field strength distribution on the power handling capability of a layer system should be most apparent for intrinsic damage mechanisms, which are dominant in the ultra short pulse regime below 1 ps. According to the fundamental model (see eqn [4]), a direct relationship of the damage threshold to the maximum field strength value within the layer system is expected. As an example, an investigation in the threshold behavior of ion beam sputtered coating systems with different field strength values is illustrated in Figure 9. In this experiment, the maximum field strength in the last low index layer of SiO_2 has been adjusted to factors between 0.4 and 1.6 of the impinging field strength (see upper diagram in Figure 9) by depositing different designs on a basic 1-QWOT layer stack of TiO_2/SiO_2. In the lower part of Figure 9, the S on 1 LIDT values, measured with a fs-laser, are depicted for selected pulse numbers N. For all layer systems,

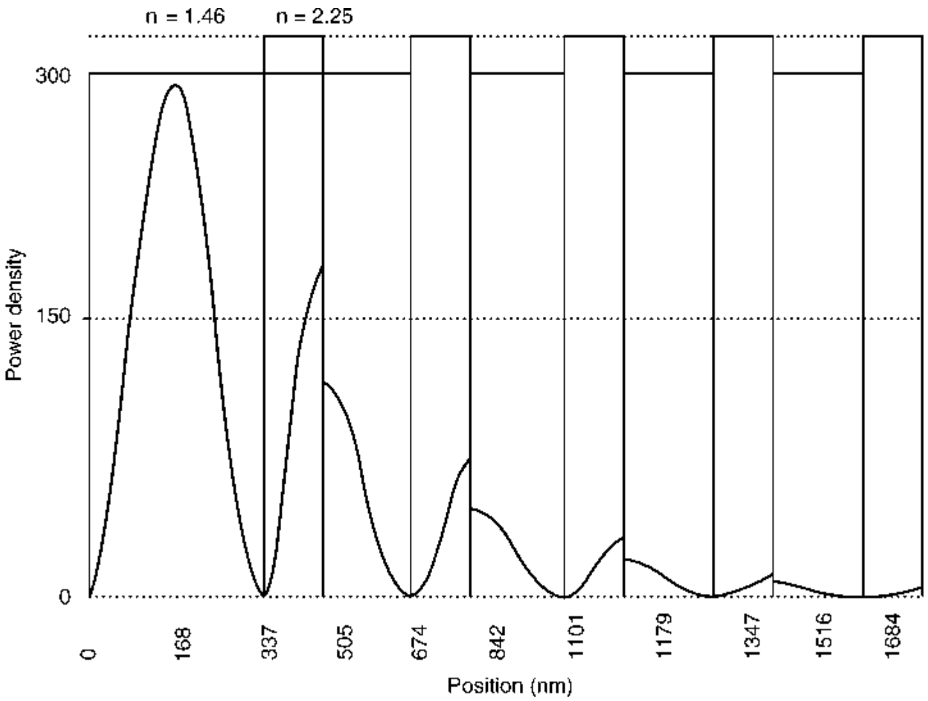

Figure 7 Distribution of the power density in a high reflecting layer stack of SiO_2/Ta_2O_5. The first nine layers of the system and a final thick stabilizing layer of SiO_2 next to the air interface are depicted. The incoming energy density is calibrated to 100%. Reproduced with permission from Ristan D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.53.

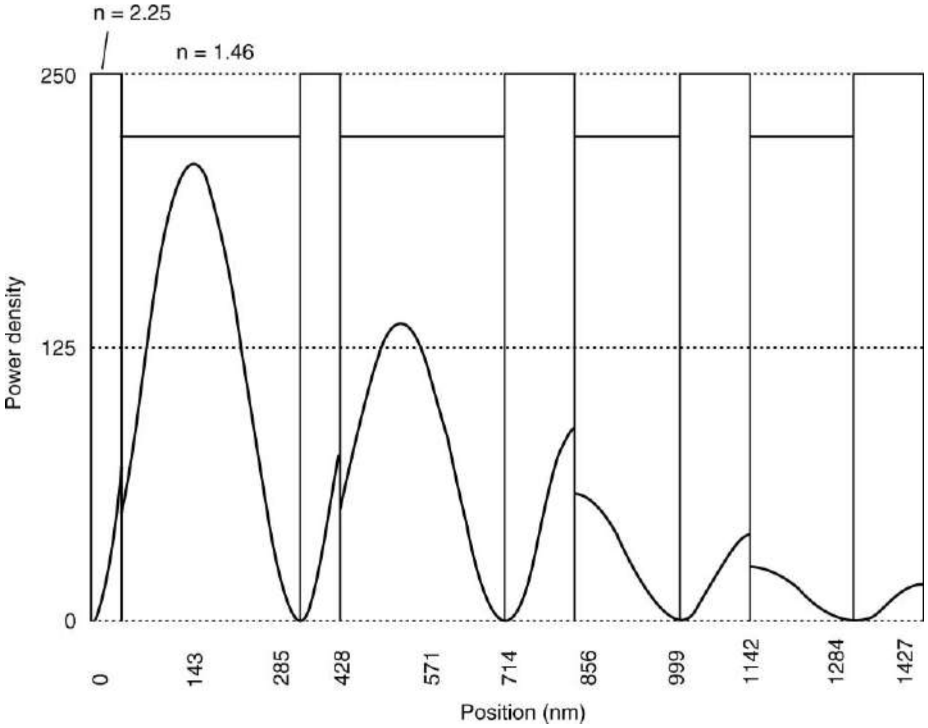

Figure 8 Distribution of power density in a high reflecting layer stack of SiO_2/Ta_2O_5. The first nine layers of the system, which is designed for reduced power density at the interface in the first layer pair are depicted. Reproduced with permission from Ristan D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.53.

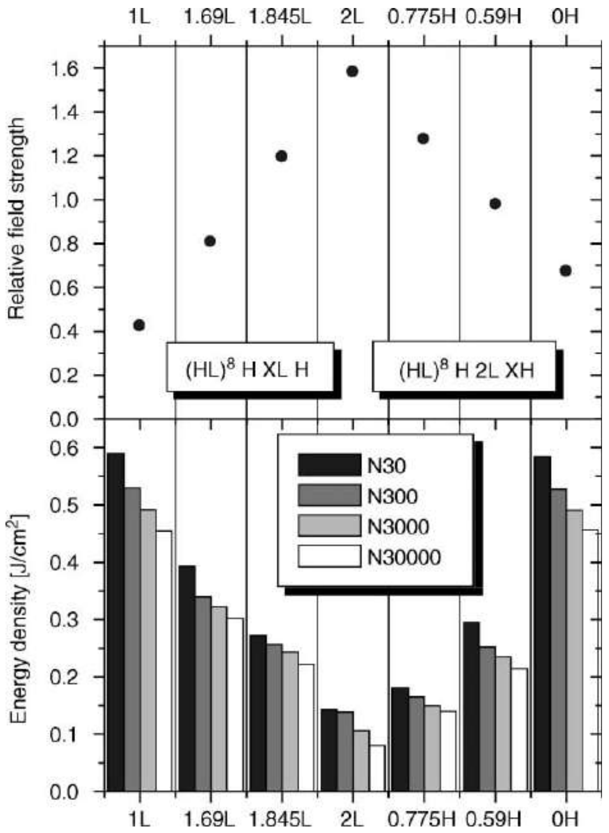

Figure 9 Damage thresholds (lower diagram) and internal field strength relative to the field strength of the impinging wave (upper diagram). The damage thresholds were measured with an ultra-short pulse laser operating at a repetition rate of 1 kHz, a pulse duration of 150 fs, and a beam diameter of 100 μm on the sample surface. Damage thresholds for selected numbers between $N = 30$ to $N = 30\,000$ of pulses are indicated by columns. The plotted maximum field strength in the last SiO_2-layer is adjusted on the basis of design variation.

which were deposited in separate deposition runs, a strong correlation between the maximum field strength and the laser threshold is observed. This experiment clearly demonstrates the role of the electric field strength distribution in high-power coatings and the potential of advanced thin film design strategies.

Besides the properties of the coatings, the quality of the substrate has to be considered in respect to surface imperfections and to the polishing procedure. For example, substrates for the production of high-quality optics with enhanced power handling capability in the VIS/NIR-spectral range should be polished to a surface roughness of less than 1 nm_{rms} with a surface imperfection value $5/1 \times 0,010$ according to ISO 10110. For an illustration of the effect of the polishing grade on the LIDT values, selected results are compiled in Table 2 for substrates of BK7-glass and high reflecting mirrors.

Table 2 Laser-induced damage threshold values of samples polished using compounds of different grain size. Besides the LIDT values of the uncoated surfaces, also data for anti-reflective coatings of selected materials on these surfaces are summarized

Coating type	*LIDT value (J/cm^2) polishing type I*	*LIDT value (J/cm^2) polishing type II*
Bare surface	76 ± 1	68 ± 12
Ta_2O_5/SiO_2	44 ± 2	38 ± 1
HfO_2/SiO_2	47 ± 16	35 ± 5
Nd_2O_3/MgF_2	34 ± 6	24 ± 8

Substrates with the polishing Type I and Type II were processed with powder of different grain diameter (2 μm and 3 μm) in conventional pitch polishing. For the LIDT values of the bare substrate and the coatings, a clear relation to the surface roughness can be observed. Especially for optical coatings with significant transmittance, subsurface damage in the substrate has to be taken into account. As a consequence of the chemical and mechanical interaction of the surface with the polishing compounds, impurities and dislocation are introduced in the surface structure resulting in a reduction of the damage threshold.

Summary

In the course of the rapid development of laser technology, a large background in high power optical coatings had been built up during the last three decades. Nowadays, the corresponding experience in the production of high power coatings is mainly located at industrial companies and a few research institutes, which are also involved in the characterization of optical coatings. For an illustration of the present state in thin film technology, damage threshold values of advanced optical components are summarized in Table 3 for laser systems and operation conditions often applied in industrial production environments.

Besides the spectacular experiments in laser fusion, isotope separation, and fundamental physics, industrial applications of lasers in material processing, medicine, information technology, and semiconductor lithography are considered as major pacemakers for the progress of high power optics. Therefore, trends in these technology fields will dominantly govern the future development of thin film technology. For example, in semiconductor lithography at a wavelength of 157 nm, which would open the way to feature sizes well below 100 nm, optical coatings with extended lifetimes are still on rank 6 of the list of challenges to achieve an effective lithographic production facility. In next generation lithography, optics

Table 3 Laser-induced damage threshold values of selected optical coating systems for laser applications (Types: HR: high reflecting mirror, AR: antireflective coating, th: thermal evaporation, IBS: ion beam sputtering)

Laser system wavelength	*Type*	*Laser-induced damage threshold ISO 11254*
193 nm, ArF-excimer	AR/th	1–2 J/cm^2 (1 on 1, 20 ns)
	HR/th	2–4 J/cm^2 (1 on 1, 20 ns)
248 nm, KrF-excimer	AR/th	10 J/cm^2 (1 on 1, 30 ns)
	HR/th	$>$20 J/cm^2 (1 on 1, 30 ns)
	HR/IBS	$>$3 J/cm^2 (1 on 1, 30 ns)
1,064 μm, Nd:YAG	AR/th	$>$60 J/cm^2 (12 ns, 0,25 mm)
	HR/th	$>$100 J/cm^2 (12 ns, 0,25 mm)
	HR/IBS	$>$80 J/cm^2 (12 ns, 0,25 mm)
10,6 μm, CO$_2$-Laser	AR/th	$>$20 J/cm^2 (100 ns, 1,4 mm)
		$>$2 kJ/cm^2 (1,2 ms, 250 μm)
		$>$3 kW/mm (cw, 100 μm)
	HR/th	$>$25 J/cm^2 (100 ns, 1,4 mm)
		$>$2 kJ/cm^2 (1,2 ms, 250 μm)

of even smaller wavelengths, around 13 nm, have to be optimized to achieve the 60 nm node. Ultra-short pulse lasers gain importance as innovative tools for material processing, laser medicine, and biology, as well as the analysis and control of chemical reactions. For the development of fs lasers needed in these applications, special high-power broadband coating systems are required. In addition to high LIDT values, these coatings have to fulfil demands with respect to their group delay dispersion for the compensation of dispersion effects in the laser systems. Other challenges have to be expected from applications, where a combination of the high-power handling capability with additional properties has to be achieved. As a typical example, high-power laser coatings with improved mechanical or chemical stability for applications in laser medicine can be considered.

List of Units and Nomenclature

dn/dt	generation rate of electrons in the conduction band
g	radius of a spherical inclusion; (cm)
$I(t)$	power density of laser radiation,
k	thermal conductivity; (W/(cm °C))
k_E	thermal conductivity of a spherical inclusion; (W/(cm °C))
LIDT	laser induced damage threshold (J/cm^2; W/cm^2; W/cm)
n'	refractive index of a spherical inclusion; (1)
P	output power of a laser; (W)
q	correction factor representing interference effects in coatings
Q	Mie absorption coefficient of a spherical inclusion;
QWOT	quarter wave optical thickness: unit for the thickness of the layer
t_I	irradiation time for a cw-laser; (s)
$T(r)$	temperature in a spherical inclusion at the radial position r; (°C)
w	diameter of a laser beam with Gaussian profile; (cm)
α	impact ionization coefficient; (1)
β_m	m-photon absorption coefficient; (1)
β_S	surface absorption of an optical component; (1)
ΔT	temperature rise in the center of an irradiated circular spot; (°C)
κ	thermal diffusivity; (cm^2/s)
κ_E	thermal diffusivity of a spherical inclusion; (cm^2/s)
λ	wavelength of a laser; (cm)

See also

Optical Coatings: Thin-Film Optical Coatings.

Further Reading

Bloembergen N (1973) Role of cracks, pores and absorbing inclusions on laser damage threshold of transparent dielectrics. *Applied Optics* 12(4): 661–664.

Genin FY and Stolz CJ (1996) *Morphologies of laser induced damage in hafnia-silica multilayer mirror and polarizer coatings*, Third International Workshop on Laser Beam and Optics Characterization. In: Morin M and Giesen A (eds). *Proc. SPIE 2870*, 439–448.

Guenther AH, *et al. Proceedings of the Symposium on Laser Induced Damage in Optical Materials, NBS (NIST) Special Publication 1969–1989*, SPIE vols. 1989–2003. CD-Rom version of collected papers from 1969 to 1998 available: Washington: SPIE.

Guenther KH, Ebert J, Kiesel E, *et al.* (1984) 1.06-m laser damage of thin film optical coatings: a round-robin experiment involving various pulse lengths and beam diameters. *Applied Optics* 23: 3743.

ISO 11254 (2001) *Test methods for laser induced damage threshold of optical surfaces*. Part 1: 1 on 1-test, 2000, Part 2: S on 1 test.

ISO 11254-3 (2002) *Laser and laser-related equipment – Determination of laser-induced damage threshold of optical surfaces – Part 3: Assurance of laser power handling capabilities*. Draft Standard. International Organization of Standardisation.

Jasapara J, Nampoothiri AVV, Rudolph W, Ristau D and Welsch E (2001) Femtosecond laser pulse induced breakdown in dielectric thin films. *Physical Review B* 63(4): 045117.

Puttick K, Holm R, Ristau D, *et al.* (1997) *Continuous wave CO$_2$ laser induced damage thresholds in optical* *components*, Proceedings of the Symposium on Laser Induced Damage in Optical Materials. SPIE vol. 3244, 188–197.

Rudolph W, Mero M, Liu J, *et al.* (2003) *Femtosecond pulse damage and pre-damage behavior of dielectric thin films*, Proceedings of the 34th Annual Symposium on Optical Materials for High Power Lasers. SPIE vol. 4932, 202–215.

Walker TW, Vaidyanathan A, Guenther AH and Nielsen P (1979) *Impurity breakdown in thin films*, Proceedings of the Symposium on Laser Induced Damage in Optical Materials. NBS Special Publication, No.568, 479–495.

Wood R (1990) *Laser Damage in Optical Materials*. Bristol and New York: Adam Hilger. Reprint with corrections.

Optical Black Surfaces

S M Pompea, National Optical Astronomy Observatory, Tucson, AZ, USA

S H McCall, Stellar Optics Research International Corporation (SORIC), Thornhill, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Black surfaces play an important role in many, if not most, optical systems. Optical instruments and systems of all types often rely on black surfaces to help minimize the effect that stray or off-axis light can have in degrading optical system performance. The strategic selection and placement of appropriate black surfaces can often limit significant detrimental stray light effects in an optical systems, thereby dramatically improving system performance. The high emissivity of black surfaces also gives them an important role in the design and construction of black bodies, calibration surfaces, and radiator surfaces.

The characterization and selection of black surfaces is an important field of optics and optical system design. However, the selection of black surfaces is a specialized undertaking requiring careful study to ensure the proper selection of particular surfaces for each optical system. A black surface that works well in one application may not be at all appropriate for a different application, where different systems performance goals are desired. The wavelength of operation of the system is an important consideration since a surface which has low reflectance at one wavelength band may have different reflective properties at another wavelength or range of wavelengths. Specialized measuring devices called goniophotometers are often used to characterize the reflectance and scatter of black surfaces for a given wavelength or set of wavelengths. Goniophotometric measurements can be used to create a specialized functional description of reflectance and scatter off a surface, called the Bidirectional Reflectance Distribution Function (BRDF) or Bidirectional Scatter Distribution Function (BSDF). In practice, having the BRDF or BSDF of a surface is extremely valuable for surface selection and is more useful than other reflectance and scatter measurement measurements or descriptions. Comparing the measured BRDF or BSDF of several surfaces helps to distinguish the general optical properties of these surfaces and allows a direct comparison of the value of these different surfaces in an optical system as they contribute to overall system performance.

The BRDF distinguishes, at a particular wavelength, how a surface specularly reflects or how it scatters light in different directions when incident light from a particular direction relative to the surface (at a given angle of incidence) interacts with the surface. By understanding the BRDF of a surface, the optical designer can understand how the surface behaves in a given system location when illuminated from a specific direction. The BRDF of a black surface is also important because it can be used as a mathematical function in stray light analysis programs to predict system performance. The BRDF and other properties of surfaces can be input into stray light codes so that the propagation of light from black surfaces can be modeled through Monte Carlo-based propagation models. Using these models, the stray light can be directed through the placement of black surfaces into directions less likely to degrade system performance. Thus a measure of stray light control can be gained through absorption by black surfaces or by redirection of the stray light into less critical directions.

Black surfaces may be considered a subset of a class of surfaces known as spectrally selective surfaces. The term ‘spectrally selective surface’ indicates that the spectral properties of many surfaces are different when examined in various spectral regions. For applications such as the design of blackbodies, space radiators, and baffles, surfaces with unique spectral emissive and reflection properties are needed. Large databases of spectrally selective surfaces are available to allow the optical designer flexibility in choosing surfaces for applications where specific optical reflectance and optical scatter properties are desired. Spectrally selective surfaces are important for applications that range from solar energy to telescope design. In some applications, it is desirable for a surface to be highly absorbing at one wavelength and more reflective at another, rather than to be highly absorbing of light over a wide range of wavelengths. An analysis of the spectral properties of a large number of surfaces allows the optical designer to choose surfaces optimized for a particular application.

Black surfaces come in many types, but may be generally characterized as being more shiny (specular) or more diffuse (Lambertian) in their reflective and scattering properties, depending on the form of their BRDF curve. Surfaces are often described in computer models as a combination of their respective specular and Lambertian properties for a given wavelength. Many black surfaces have significant reflectivity at some region of the electromagnetic spectrum, even if they are highly absorbing at one wavelength. Some black surfaces are constructed to be remarkably nonreflecting over a range of wavelengths; these surfaces are often used as baffle surfaces for telescopes and instruments that must operate over a wide range of wavelengths. No black surface can work well in all optical systems. Optical designers must understand the specific optical properties of a black surface to determine if the surface can work in their particular system. To characterize a surface, it is helpful to know at least the BRDF of the surface at the wavelength of interest as well as basic materials properties of the surface. The specific BRDF at the most appropriate wavelength for the system can be used with the optical system stray light model to determine if the surface can fulfill the designer's, expectations of it for the system performance. The placement and choice of black surfaces in an optical system are often critical decisions with highly significant effects on overall system performance.

For this reason, optical designers rely greatly on databases of the reflection and scatter properties of black surfaces in order to make the right choice of surface. These databases have been created to provide the optical engineer with the information needed to understand a wide variety of candidate surfaces and select the most appropriate ones. These databases not only describe the optical properties of surfaces but also elucidate material properties so that the selection process can examine the heat resistance, outgassing rates, fade resistance, durability, and a myriad of other surface or material properties that may be relevant.

The reflectance and scatter data on black surfaces are often taken using a variety of reflectance-measuring instruments and the data format often varies greatly. When published and unpublished data sets are compiled, they are often confusing and sometimes even contradictory. This has made the comparison of surfaces a difficult proposition. In recent years, there have been successful attempts to gather existing large amounts of existing data so that comparisons among black surfaces can be made in a data-rich environment so that anomalous data can be ignored. A significant amount of reflectance data on black surfaces has been created by private or governmental organizations such as NASA and NIST – the US National Institute for Standards and Technology. These data now appear in databases of BRDF and reflectance measurements. Designers are relying increasingly on databases of BRDF surfaces rather than the individual measurements that can be found in the optical physics and applied optics published literature.

Uses of Black Surfaces

Table 1 lists some of the major areas in which black surfaces or black coatings may be used in instruments and optical systems. The optical properties of desirable surfaces may be different for the specific categories in the table. Plausible surfaces for these different applications can be found in materials databases such as the SOLEXIS™ database which contains a useful compilation of data from private

Table 1 Possible uses for black surfaces and black coatings in optical systems

Apertures	Baffles	Barrels	Calibration screens	Choppers
Cold shields	Detectors	Dewars	Diffusion screens	Domes for telescopes and instruments
Enclosures and curtains for rooms and instruments	Integrating sphere enclosures	Lens edges	Light traps	Mounts
Radiators	Radiometers	Solar collector absorbers and concentrators	Standards and references of reflectance	Targets

and governmental data compilations from the US, Canada, and the former Soviet Union.

Selection of Optical Black Surfaces

The selection of appropriate black surfaces for a given application becomes progressively harder as the number of optical requirements (e.g., reflectivity, scatter, emissivity, etc.) and materials requirements (e.g., stiffness, coefficient of thermal expansion, etc.) increase. Because of this, the selection process for specialized surfaces for space applications can be quite involved, since there are a large number of criteria, each of which may be regarded as an essential requirement.

The optical system designer must consider at least the following aspects of the system in selecting a surface:

- the wavelength(s) and waveband(s) of interest;
- the position of the surface in optical system;
- the general robustness required of the surface;
- the environment of the optical system;
- the cost including inspection, handling, and assembly; and
- the production and delivery schedule, including spares.

The selection process must carefully address the system needs as well. Poor decisions are often made when only a cursory study of potential black surfaces is made or when databases and the literature on the subject are neglected. Therefore, an important part of the selection process is the management of the timing of the selection, in order to give the designer time to access available information and to influence the system design in a timely manner.

If black surfaces are used for stray light control, then the use of these surfaces must be planned at the beginning of the optical system design process. If the placement of black surfaces and the array of potential surface treatments are examined early in the program, the optimal choices can often be determined in a careful manner that enhances the system performance. If the selection is saved until the very end of the program, the choice of surfaces will be more limited and improvements in system performance may be minimal, even with the most excellent black surfaces. This is because the location of the surfaces in the system often proves to be as critical as the choice of surface. If the surface is in the wrong location, the best optical black may not be able to improve the stray light performance. For best results, the layout of the system and the placement of baffles must be considered first, before the baffle surfaces are selected.

An optimal flowchart of good system design for black surface selection is given in Figure 1. Figure 2 describes a work flowchart where black surface selection is neglected until the end of the program, often with dire consequences for system performance, cost, and schedule.

For space-based systems, the effect of atomic oxygen in a low-earth orbit environment is an important consideration. For space systems, the designer must also examine how black surfaces may be affected by or contribute to contamination and outgassing. Other considerations are the generation of particles during launch, the effect of solar

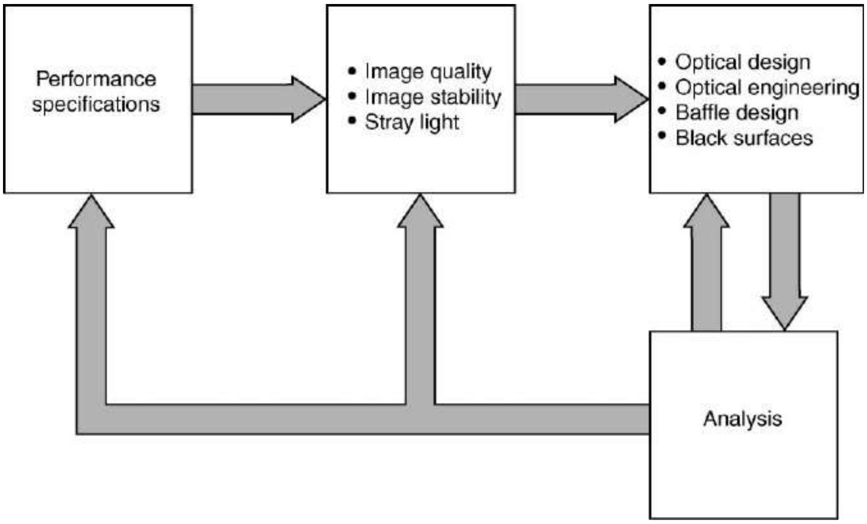

Figure 1 Diagram of the system design process which treats black surface selection as an important part of basic system design, to be done early in the design process. This gives the designer an opportunity to conduct system tradeoffs to optimize system performance.

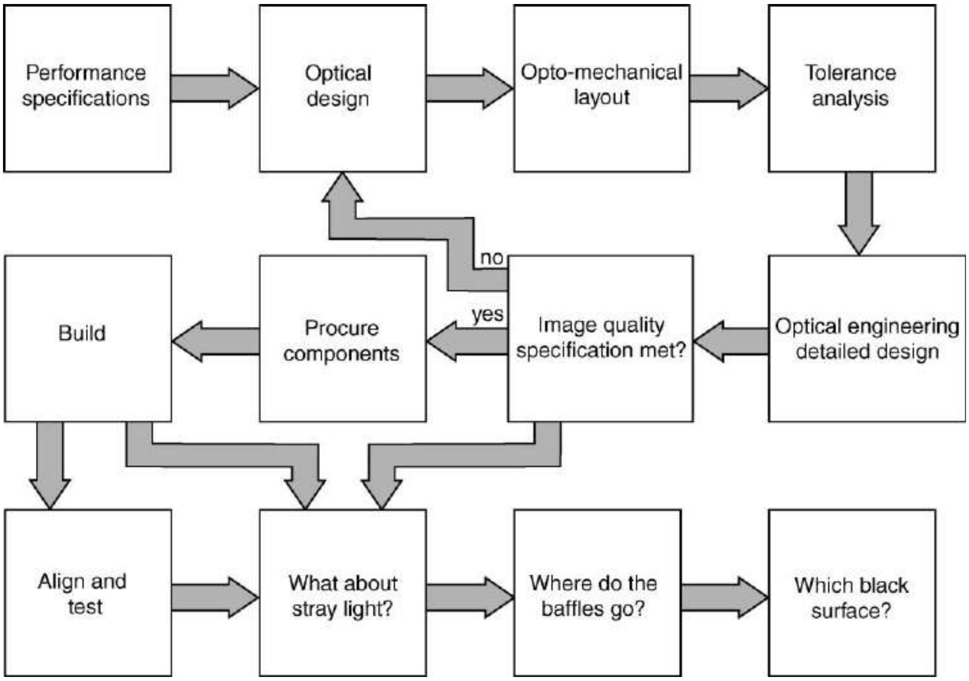

Figure 2 Diagram of a poor system design process. This process addresses the selection of black surfaces late in the system design process, where selection is significantly more difficult, expensive, and risky. Even with a late selection, there may be severe impacts on schedule.

ultraviolet radiation on black surfaces, and how spacecraft charging may affect conductive, dielectric, or insulating black surfaces. Since black surfaces may have a large surface area in an optical system, the designer must ensure that the black surfaces do not contribute significantly to the risk of system failure through some degradation process that may contaminate other optical surfaces. For space-based systems, data on surface outgassing and the past performance of a surface in a specific space environment is essential.

There are a variety of resources available to assist in the selection of black surfaces. The best databases allow the user to create 2-dimensional and 3-dimensional BRDF plots for extensive lists of potential black surfaces. For most surfaces, with no preferred directionality in the structure of the surface, a 2-dimensional BRDF plot is adequate. Most of these databases provide information on a wide variety of spectrally selective materials including black, white, and reflective surfaces for thermo-optical applications for ground and space-based instrumentation. The BRDF plots can be examined by the user or fed into existing, commercially available optical design and analysis software packages. The results of these models indicate whether the chosen surfaces give adequate performance or whether more specialized coatings are necessary to achieve the desired level of performance.

A significant problem for many designers is that the formulations of many paints and other surface treatments often change. Sometimes the name of the coating remains the same even as the optical properties change, due to a change in the compounds or methods used to prepare the coating. Sometimes the name changes with or without a change in the formulation. These often unpublished changes increase the risk of using older plots from the published literature as definitive optical modeling data. Experts in the optical blacks area continuously investigate these changes and their implications and can advise optical designers of the implications of recent changes in the paint and coating arena. These experts are often aware of situations where coating failures have occurred or where mission performance was degraded by unanticipated changes in coatings, even if reports of this incidents were never published.

In summary, the selection process, by its very nature, must be a deliberate and careful effort to examine the best surface options. Some of the better surfaces which were considered laboratory standards (e.g., 3M Black) are no longer easily available and replacements must be found. For the most systematic decision-making process, access is needed to the best optical black surface databases, often commercial databases with hundreds to thousands of BRDF curves on several hundred materials. The use of these optical surface databases can substantially reduce the costs, time, and scope of a measurement program, as well as streamline the selection process. Additionally it allows the designer to consider materials used successfully in other programs which may not be familiar to the designer. Finally, the designer may avoid using materials that have been rendered unreliable through changes in the manufacturing process or in the coating chemistry.

Characterization of Black and Spectrally Selective Surfaces

Optical Characterization

An understanding of black surfaces relies heavily upon an understanding of the meaning and use of the BRDF. While the optical properties of surfaces can be characterized through specular and diffuse reflectance measurements, the use of BRDF measurements has become extremely important. BRDF measurements are equally adept at characterizing highly specular and highly diffuse surfaces and the myriad of surfaces with both characteristics. The BRDF is often used in general to describe the angle-dependent optical scatter from these surfaces. This is important as the angular distribution of scatter can be used to calculate if the scatter from a black surface will be a limiting factor in the resolution or noise level of the system. BRDF measurements can even be used in computer graphics visualization programs that illustrate what a surface may look like under various illumination conditions. For example, the BRDF is used in computer graphics models to render a certain color for an automobile or a countertop in a kitchen. For some black surfaces, the angular distribution of scatter can be used to calculate a variety of surface parameters such as the surface roughness, given certain assumptions about the surface. For opaque black surfaces, BRDF measurements have proved to be one of the best overall descriptors of the optical properties of the surface.

BRDF is defined as the ratio of scattered radiance (watts/(cm^2 sr)) to surface irradiance (watts/cm^2); its units are inverse steradians. Radiance is used in the definition, in order to make the BRDF independent of the parameters of the measuring instrument, such as the detector aperture and distance to the detector. BRDF measurements are made by a variety of instruments for a number of wavelengths. For example, BRDF measurements of surfaces have been made in wavelengths ranging from the vacuum ultraviolet to the far infrared. The most common BRDF measurements made on black surfaces are made at laser wavelengths of 0.6328 and 10.6 micrometers. However, these measurements may be of limited value if the system is operating at substantially different wavelengths, where the surface's BRDF is significantly different.

To make a BRDF measurement requires a light source, a sample mounted and illuminated by that source, a receiver (detector) to measure the scatter from the sample, and the computer/electronics package to accurately record the detection of the scattered light. Additionally the sample and/or light source must be able to be moved so that the sample can be illuminated from nearly every angle, and scattered light from the surface can be detected in a hemisphere around the sample. The geometry of the BRDF measurement is illustrated in Figure 3.

To make a BRDF measurement on black surfaces, appropriate sources, detectors, and low-scatter optics must be used. This becomes more difficult for wavelengths less than about 0.25 micrometers. At longer, infrared wavelengths, diffraction effects become important. Despite the instrumental difficulties, it is often better to measure the amount of scatter at a specific wavelength λ and angle θ_s, rather than try to predict it from theoretical models. This is because the extrapolation to new angles or wavelengths using scaling theories is not always accurate or even possible. Methods for wavelength scaling (extrapolating from BRDF data for one wavelength to data for any other wavelength) or angle scaling (where data taken at particular incident and scatter angles are extrapolated to other angles) will not be discussed as BRDF curves are now widely available for many surfaces at a variety of wavelengths and angles.

The definitions of BRDF are given in Table 2.

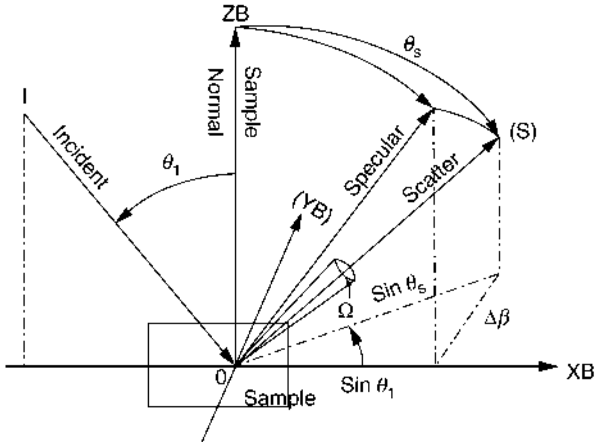

Figure 3 Geometry of BRDF measurement and the angle conventions accepted by the American Society for Testing Materials (ASTM). The plane of incidence is the I-0-ZB plane and the scatter plane is the S-0-ZB plane.

Table 2 BRDF term definitions

Bidirectional reflectance distribution function BRDF	The sample radiance divided by the sample irradiance. The procedures given in this practice are correct only if the field of view (FOV) determined by the receiver field stop is sufficiently large to include the entire illuminated area
Scatter	The radiant flux that has been redirected (by interaction with the sample) over a range of angles
Sample coordinate system	A coordinate system is fixed to the sample and used to a specify position on the sample surface for the measurements. The sample coordinate system may be application and sample specific. A Cartesian coordinate system is recommended for flat samples, with the origin at the geometric center of the sample face with the Z axis normal to the sample
Plane of incidence, PLIN	The plane defined as containing the central ray of the incident flux and the sample normal
Beam coordinate system, XB, YB, ZB	A Cartesian coordinate system defined with the origin on the central ray of the incident flux at the sample surface, the XB axis in the plane of incidence (PLIN) and the ZB axis normal to the surface. The angle of incidence, scatter angle, and incident and scatter azimuth angles are defined with respect to the beam coordinate system
Incident direction	The central ray of the incident flux specified by θ_i and φi in the beam coordinate system
Angle of incidence, θ_i	Angle between the central ray of the incident flux and the ZB axis
Incident azimuth angle, φ_i	The fixed 180° angle from the XB axis to the projection of the incident direction onto the XB–YB plane. It is convenient to use a beam coordinate system in which $\varphi_i = 180°$ since this makes φ_s the correct angle to use directly in the familiar form of the grating equation. Conversion to a sample coordinate system is straight forward, provided the sample location and rotation are known
Scatter direction	The central ray of the scattered flux specified by θ_s and φ_s in the beam coordinate system
Scatter angle, θ_s	Angle between the central ray of the scattered flux and the ZB axis
Scatter azimuth angle, φ_s	Angle from XB axis to the projection of the scatter direction onto the XB–YB plane
Specular direction	The central ray of the reflected flux that lies in the PLIN with $\theta_i = \theta_s$ and $\varphi_s = 0$
Delta theta, $\Delta\theta$	The angle between the specular direction and the scatter direction
Delta beta, $\Delta\beta$	The projection of $\Delta\theta$ onto the XB–YB plane, that is, the $\Delta\theta$ angle measured in direction cosine space. For scatter in the PLIN, $\Delta\beta = \sin\theta_s - \sin\theta_i$. For scatter out of the PLIN, the calculation of $\Delta\beta$ becomes more complicated
Receiver solid angle, Ω	The solid angle subtended by the receiver aperture stop from the sample origin
Sample radiance, L_e	A differential quantity that is the reflected radiant flux per unit projected receiver solid angle Ω per unit sample area. L_e is often an average calculated from scattered power, P_s collected by the projected receiver solid angle $\Omega\cos\theta_s$, from the illuminated area, A. The receiver aperture and distance from the sample determines Ω and the angular resolution of the instrument
Sample irradiance, E_e	The radiant flux incident on the sample surface per unit area. In practice, E_e is an average calculated from the incident power, P_i, divided by the illuminated area, A. The incident flux should arrive from a single direction; however, the acceptable degree of collimation or amount of divergence is application specific and should be reported

Table adapted from ASTM E1392-90 "Standard Practice for Angle Resolved Optical Scatter Measurements on Specular or Diffuse Surfaces," American Society for Testing Materials, 1990.

Given these definitions, BRDF, is mathematically described by

$$\text{BRDF} = \frac{L_e}{E_e} = \frac{(P_s/\Omega A\cos\theta_s)}{(P_i/A)} = \frac{P_s}{P_i\Omega\cos_s}\ [\text{sr}^{-1}]$$

The BRDF is a function of many measurable quantities. It depends on the angle of incidence of the incident flux and the wavelength of this radiation. It also depends on the direction in which the scattered light is measured as well as the wavelength and polarization of the incident and scattered flux. The BRDF can be calculated for a given angular configuration of incident source and detector from the average radiance divided by the average irradiance. The BRDF is therefore referenced by the incident power. The cos θ_s term in the formula accounts for the foreshortening of A as the angle θ_s increases, so that true radiance is being used in the calculation. Because of this, the BRDF is constant for a surface with Lambertian (uniformly diffuse) properties. The detector must have a field of view so that all of A can be viewed and such that P_i includes the entire beam. In practice, P_i is measured by removing the sample from the measurement position. Then the detector is moved into the straight-through position. Often a reference detector is used to monitor the incident beam power so that any changes after the initial V_i, measurement can be recorded.

For BRDF measurements to be meaningful, the instrument signature must be examined and understood. This is particularly important in the measurement of low-scatter surfaces. The instrument signature is the BRDF when the sample is removed and there is no sample scatter. In this case the stray light and scatter signature of the device will be apparent. Angular positioning errors and electronic noise can also be important. For example, as θ_s approaches 90° the accuracy of measuring θ_s becomes critical due to the $1/\cos\,\theta_s$ term in the BRDF equation. Ideally the instrument signature is a small fraction of the BRDF being recorded. An understanding of the instrument signature provides the needed information to decide where the BRDF data loses its meaning.

The BRDF is most often displayed in a plot, where the BRDF value (vertical axis) and the angle of scatter (horizontal axis) are plotted on logarithmic scales. In these ‘log–log’ plots it is easy to examine the scatter distribution close to the specular direction which, for shiny surfaces, represents the bulk of the scattered light. For surfaces that are diffuse in nature, a log–log plot of the BRDF generally yields a straight line. The slope of this plot is characteristic of the surface of the process which is used to produce the surface. This straight-line plot is useful in comparing samples made with the same process. Deviations from the slope of the line or the value of the points on the line indicate that there may be problems with a sample. Thus the simplest scattering instrument can be used for quality control of black samples. For a fixed angle of incidence and a fixed detection angle, variations in the BRDF value may indicate changes in the materials or the processes which created their surfaces. The quality control process can be performed by looking for variations in the BRDF among the different samples.

The BRDF can be used as a measure of surface roughness – an important attribute for a diffuse black surface. As random surface roughness features increase on a surface, the scatter from the surface also increases. Because of this, the BRDF can be used to characterize surface roughness parameters. If the surface roughness is all in the form of small, micro-rough surface features, the BRDF value then is very much proportional to the surface power spectral density function (PSD), which is a measure of the power plotted against spatial frequency. The characterization of surface roughness by BRDF scatter measurements only works if the scattered energy is due to surface micro-roughness. If the scatter is from larger surface features or from surface contamination, then it will be difficult to accurately characterize the surface roughness. Thus only smooth, mirror-like black surfaces can be used for this simple leap from BRDF values to surface roughness parameters. Some sample BRDFs of black surfaces are found in Figures 4–7.

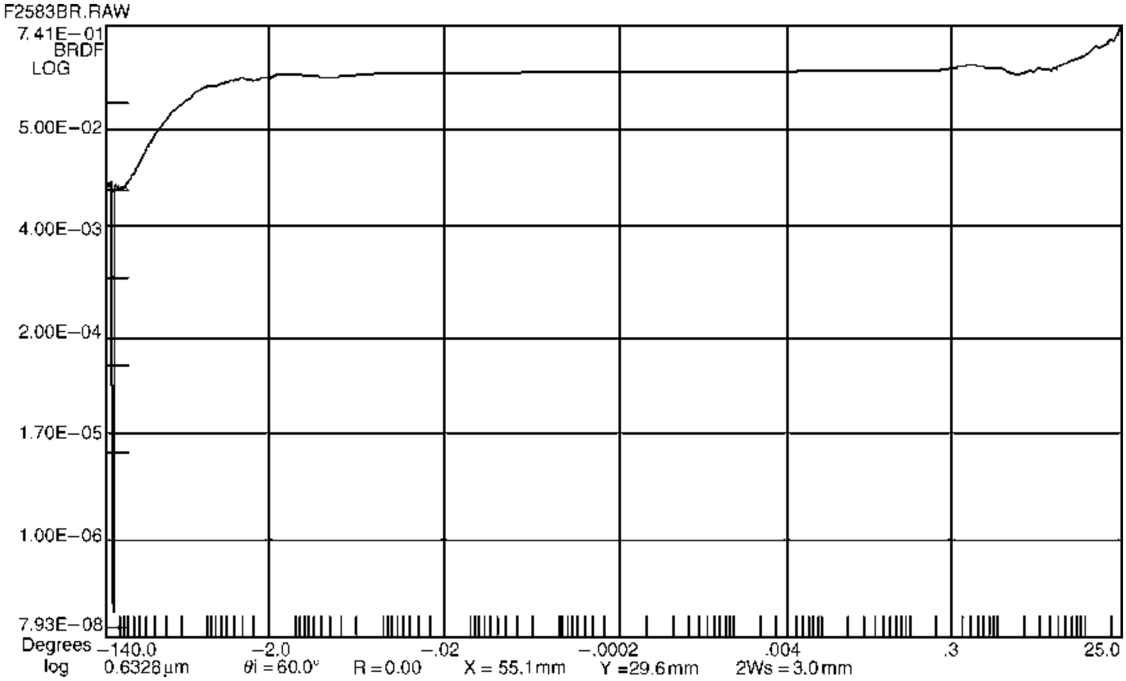

Figure 4 BRDF at wavelength of 0.633 micrometers and 60° incidence angle of black anodize sample. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

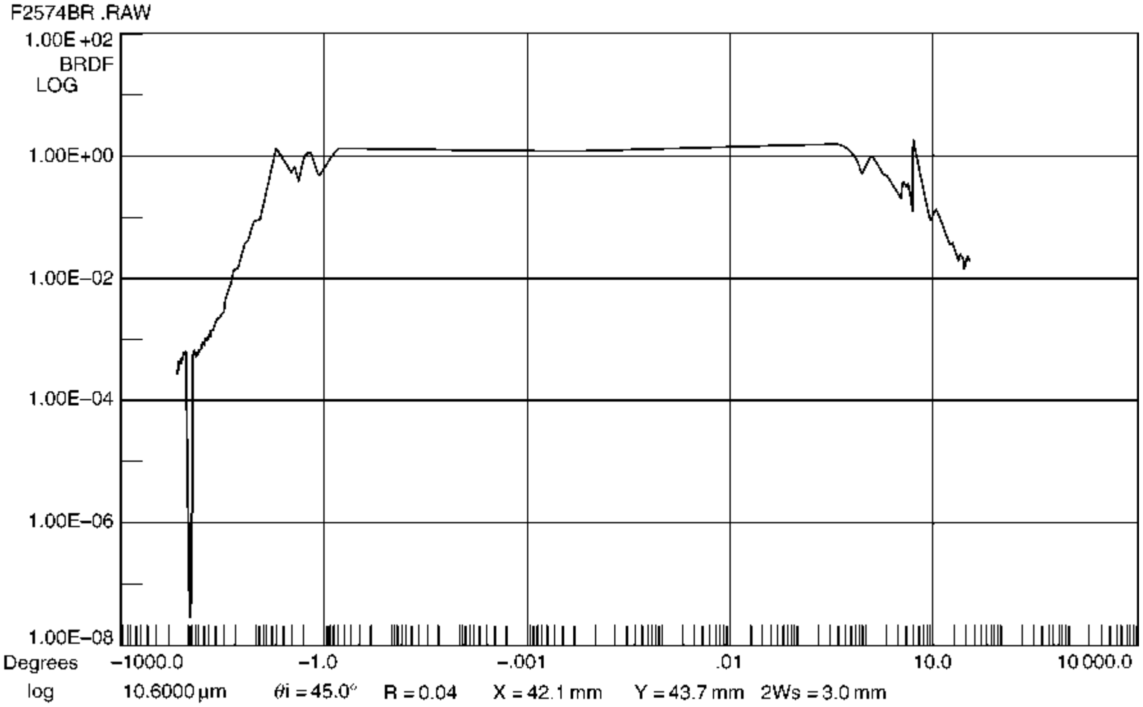

Figure 5 BRDF at wavelength of 10.6 micrometers of black anodize at 45° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

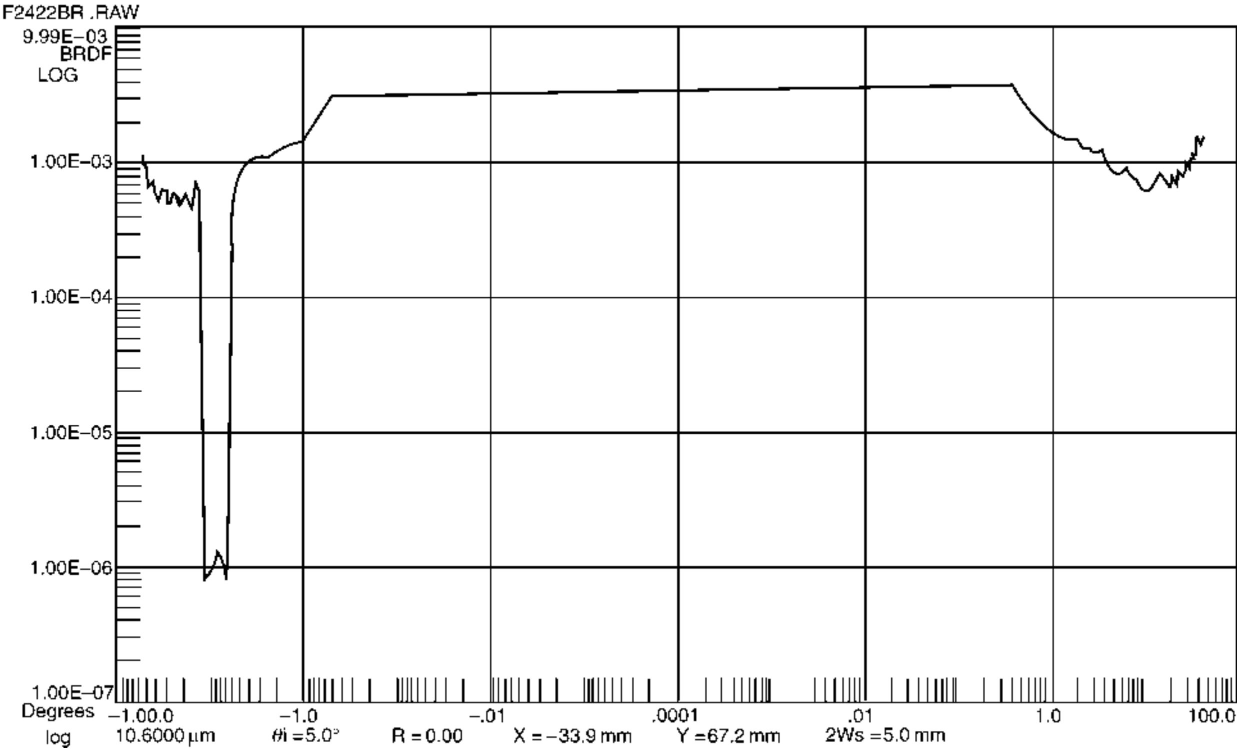

Figure 6 BRDF at a wavelength of 10.6 micrometer of Martin Optical Black at 45° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

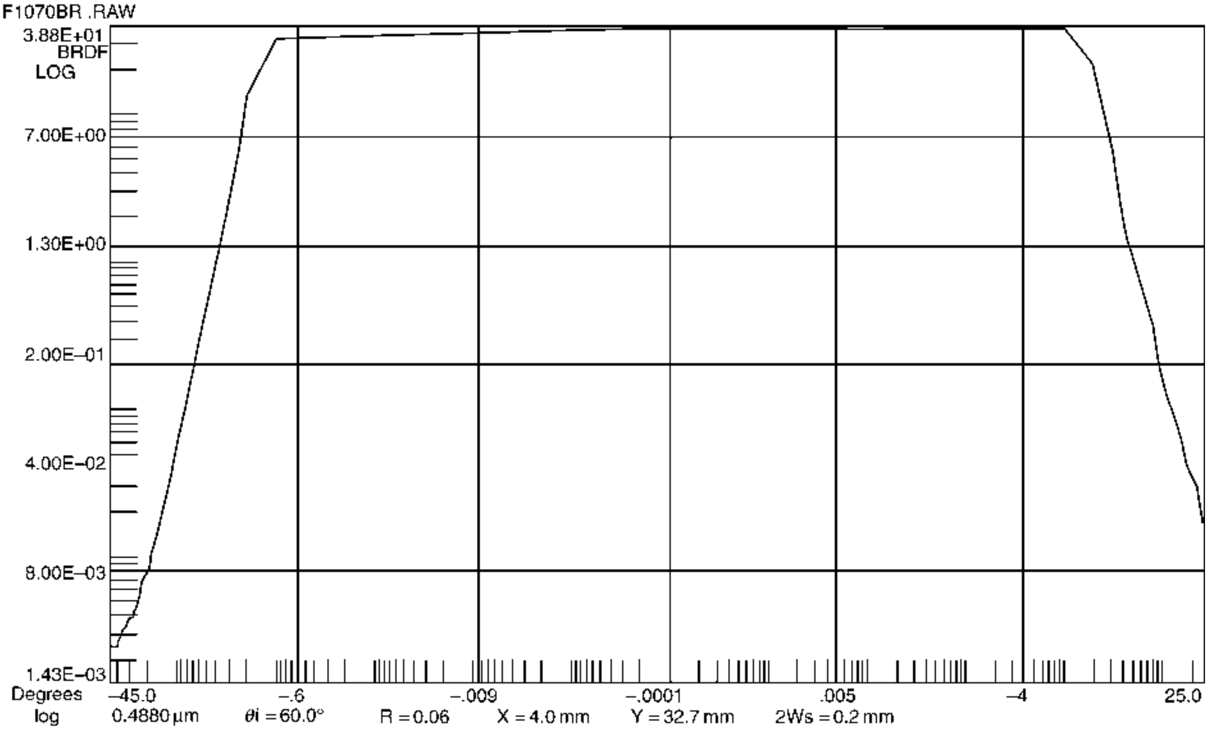

Figure 7 BRDF at a wavelength of 0.488 micrometers of a glossy black surface at 60° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

Examples of Black Surfaces

Black surfaces appear in many forms. They may be either diffuse or specular in nature, or both. Black surfaces may be created by painting a surface, by altering a surface (such as through an anodization process), or by any number of other surface treatment effects. For example, absorbing compounds may be added to paint or to a surface to decrease reflection. Adding large (compared to the wavelength) craters or holes to a surface may result in multiple reflections within the surface leading to decreased reflection. Surfaces may be modified to increase their light-scattering ability, with the hope that reflection in a specific direction will be reduced. Sometimes surfaces may be created in such a way as to enhance thin-film interference effects, leading to reduced reflection. These and other techniques may be combined together in an attempt to reduce reflection and make the material appear blacker in a particular wavelength range of interest or in a particular direction.

Some sample black surfaces are given in Table 3, although databases are much more current on current names and manufacturers.

Many black surfaces use surface texture to enhance their light-absorbing ability. Scanning electron micrographs of surfaces reveal the surface morphology and how multiple reflections can be used for light trapping. Surface chemistry, changes in the index of refraction, and the use of particles of different sizes are also used to create better absorbers. A variety of other proprietary methods are used as well (Figures 8–11).

Future Developments

There are many new kinds of black surfaces being developed for specialized applications. Black surfaces can be created using a variety of new techniques and materials. For example, new surfaces can be created using laser ablation or chemical microetching of surfaces. Surfaces can also be created using new materials such as carbon nanotubes or by using the technology used to create flocked materials. These new kinds of surfaces generally have good absorption at a variety of wavelengths. Even with the creation of these new surfaces, the greatest gain in an optical system will probably be attained through a concerted effort to choose the appropriate black surfaces early in the program and to place them properly in the optical system. An early selection of surfaces and an analysis of their performance will lead to a higher level of system performance than a search for an exotic, ultrablack surface at a late stage

Table 3 Selected optical black surfaces

Name	*Manufacturer*	*Notes*	*Type*
Aeroglaze L300	Lord Corporation Erie, PA	Formerly called Chemglaze L300	Paint
Aeroglaze Z004	Lord Corporation Erie, PA	Formerly called Chemglaze Z004	Paint
Aeroglaze Z302	Lord Corporation Erie, PA	Formerly called Chemglaze Z302	Paint
Aeroglaze Z306	Lord Corporation Erie, PA	Formerly called Chemglaze Z306	Paint
Aeroglaze Z306 with microspheres	Lord Corporation Erie, PA	Formerly called Chemglaze Z306 with microspheres	Paint
Aeroglaze Z307	Lord Corporation Erie, PA	Formerly called Chemglaze Z307	Paint
Aeroglaze Z313	Lord Corporation Erie, PA	Formerly called Chemglaze Z313	Paint
Ames 24E (Ames 24E2)	NASA Ames Research Center Moffett Field, CA		Paint
Cardinal 6450	Cardinal Industrial Finishes South Elmonte, CA	Formerly 'Cardinal 6550'	Paint
Cornell Black	Cornell University		Paint
DeSoto Flat Black	Courtaulds		Paint
Electrically Conductive Black Optical Paint	Jet Propulsion Laboratory Pasadena, CA	Has no trade name	Paint
IITRI Bone Black D-111 (IITRI D111)	Illinois Institute of Technology Chicago, IL		Paint
LMSC Black	Lockheed Martin Palo Alto		Painted multilayer coating
MH21-I	Illinois Institute of Technology Chicago, IL		Paint
MH55	Illinois Institute of Technology Chicago, IL		Paint
MH2200	Illinois Institute of Technology Chicago, IL	Formerly 3M's ECP 2200 paint, but sold to IIT	Paint
Solarchem	Eastern Chem Lac Corporation Malden, MA		Paint
463-3-8	Akzo Coatings, Inc. Orange, CA	Formerly called 'Cat-a-lac 463-3-8' diffuse black paint	Paint
443-3-8	Akzo Coatings, Inc. Orange, CA	Formerly called 'Cat-a-lac 443-3-8' glossy black	Paint
443-3-17	Akzo Coatings, Inc. Orange, CA	Formerly called 'Sikkens 443-3-17'	Paint
Infrablack	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Martin Black	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Martin Black, Enhanced	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Infrablack	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Black Chrome Type Surfaces	Lockheed Martin Denver, CO	For many kinds of metal substrates	Electrodeposition process
Black Cobalt Type Surfaces (Cobalt Black)	Lockheed Martin Denver, CO and many more companies	For many kinds of metal substrates. References to Black Copper, Black Steel etc. are covered by Black Cobalt	Electrodeposition process, and can be followed by chemical or thermal oxidation
Black Nickel	−NIST	For many kinds of metal substrates	Deposition and etching process
(NBS Black)	−Ball Aerospace		
(Ball Black)	−Lockheed Martin and others	Ball modified the process developed by NBS	
Black Kapton™ Foil			Foil
Black Tedlar™ Foil			Foil
Boron Black	Lockheed Martin Denver, CO	For many kinds of metals	Plasma spray deposition process
Boron Carbide	Lockheed Martin Denver, CO	For Ti substrates only	Proprietary process
Silicon Carbide	Lockheed Martin Denver, CO and other companies	For many kinds of metals	Chemical vapor deposition
Textured Surfaces	−NASA Ames Research Center Moffat Field, CA −SPIRE Corporation Bedford, MA −Optics MODIL Oak Ridge National Laboratory	For many kinds of metal substrates	Either: −sputter coated −ion beam etched −sputter coated then etched

Note that finish names and manufacturers change regularly. Table adapted from SOLEXIS™ database, Stellar Optics Research International Corporation (SORIC), 78 Normark Dr., Thornhill, Ontario, Canada L3T 3R1.

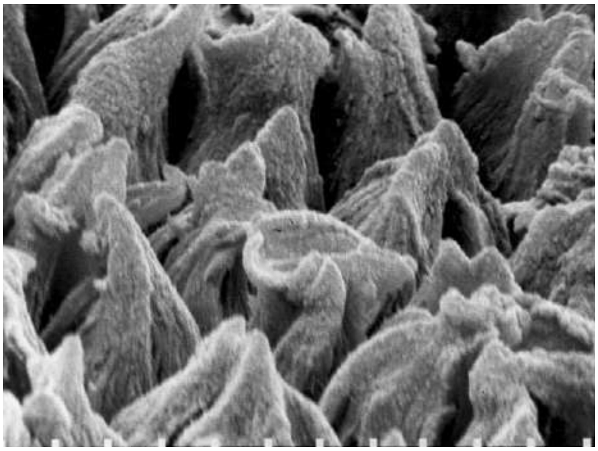

Figure 8 Scanning electron micrograph of Martin Optical Black surface on 6061T6 aluminum. This surface is a good absorber at ultraviolet, visible, and near infrared wavelengths. (Courtesy of Stephen Pompea and Donald Shepard.)

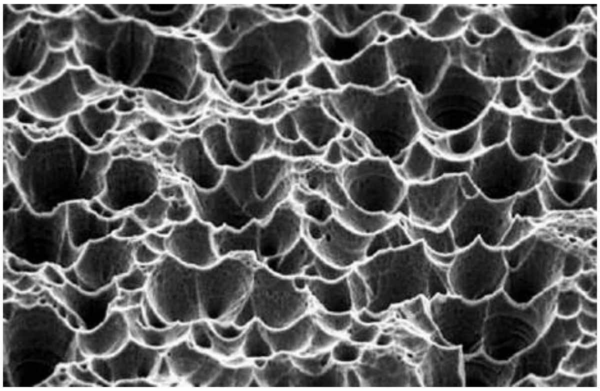

Figure 9 Scanning electron micrograph of phosphoric acid etched electroless nickel surface. This surface is a good absorber in the visible wavelength region. (Courtesy of Stephen Pompea and Donald Shepard.)

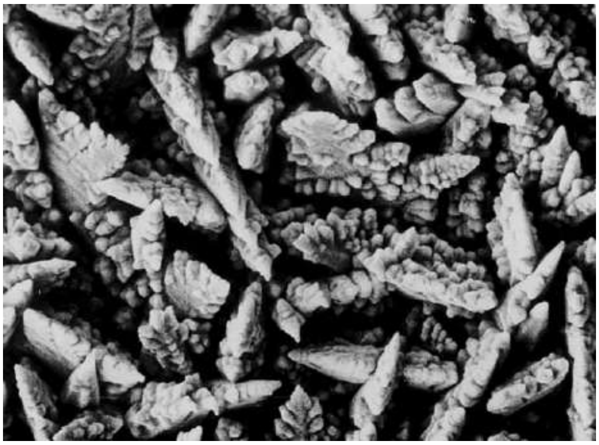

Figure 11 Scanning electron micrograph of tungsten hexafluoride surface. This surface is a good absorber at visible and near infrared wavelengths. (Courtesy of Stephen Pompea and Donald Shepard.)

in the program. Although better black surfaces are constantly being developed, the most important future development in the area of black surfaces may be a wider realization of the importance of an early and comprehensive surface selection process. The value of starting the selection process early using black surface databases has been confirmed repeatedly by optical designers.

See also

Instrumentation: Photometry; Telescopes. **Scattering:** Scattering from Surfaces and Thin Films.

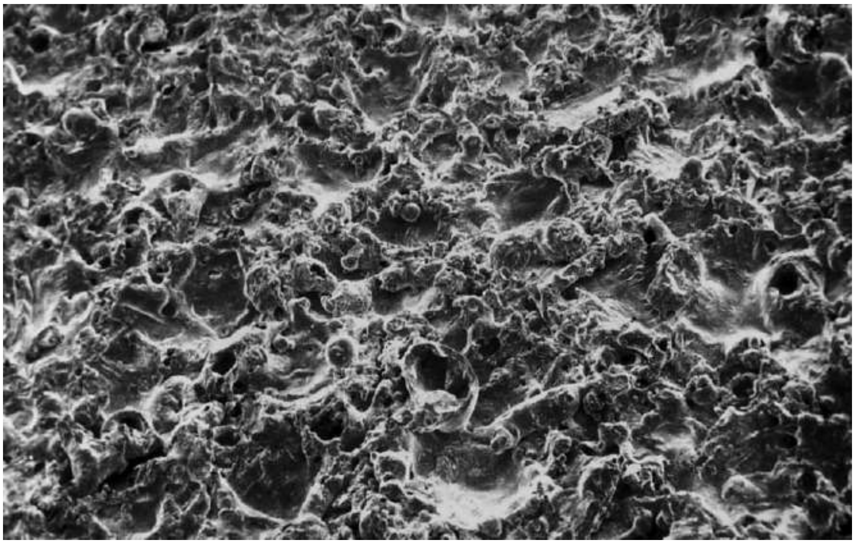

Figure 10 Scanning electron micrograph of Aeroglaze Z-306 painted surface with embedded microspheres. This surface is used primarily at visible wavelength. (Courtesy of Stephen Pompea and Donald Shepard.)

Further Reading

American Society for Testing Materials (ASTM) (1990) ASTM E1392-90. *Standard Practice for Angle Resolved Optical Scatter Measurements on Specular or Diffuse Surfaces*. West Conshohocken, PA: ASTM.

Bennett JM and Mattsson L (1989) *Introduction to Surface Roughness and Scattering*. Washington, DC: Optical Society of America.

Bohren CF and Huffman DF (1983) *Absorption and Scattering of Light by Small Particles*. New York: John Wiley and Sons.

Breault RP (1995) Control of stray light. In: *Handbook of Optics*, 2nd edn. New York: Optical Society of America.

Church EL and Takacs PZ (1995) Surface scattering. In: Bass M (chief ed.) *Handbook of Optics*, 2nd edn. New York: Optical Society of America.

Granqvist CG (1989). *Spectrally Selective Surfaces for Heating and Cooling Applications*. Volume TT 1. Bellingham, WA: SPIE Optical Engineering Text.

Gunderson JA (1977) *Goniometric Reflection Scattering Measurements and Techniques at 10.6 Micrometers*. Thesis, University of Arizona.

Henninger JH (1984) Solar absorptance and thermal emittance of some common spacecraft thermal control coatings. *NASA Reference Publication 1121*.

Jelinsky P and Jelinsky S (1987) Low reflectance EUV materials: a comparative study. *Applied Optics* 26(4): 613–615.

Persky MJ (1999) Review of black surfaces for space-borne infrared systems. *Rev. Sci. Instruments* 70(5): 2193–2217.

McCall SHCP, Pompea SM, Breault RP and Regens NL (1992) Reviews of properties of black surfaces for ground and space-based optical systems. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Radiation II*, vol. 1753.

Pompea SM and Breault RP (1995) Optical black surfaces. In: *Handbook of Optics*, 2nd edn. New York: Optical society of America.

Pompea SM, McCall SHCP (1992) Outline of selection process for black buffle surfaces in optical systems. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Radiation in Optical Systems* II, vol. 1753.

Pompea SM, Shepard DF and Anderson S (1988) BRDF measurements at 6328 Angstroms and 10.6 micrometers of optical black surfaces for space telescopes. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Light and Contamination in Optical Systems*, vol. 967, pp. 236–248.

Smith SM (1980) Far infrared (FIR) optical black bidirectional reflectance distribution function (BRDF). *Proceedings of the SPIE: Radiation Scattering in Optical Systems* 362: 161–168.

SOLEXIS™ Database of BRDF Measurements (2003) Stellar Optics Research International Corporation (SORIC), Ontario, Canada.

Stover J (1995) *Optical Scattering: Measurement and Analysis*, 2nd edn. Bellingham, WA: SPIE Optical Engineering Press.

Wolfe WL (1978) Optical materials. In: Wolfe WL and Zissis GJ (eds) *The Infrared Handbook*. Ann Arbor, Michigan: ERIM.

Thin Film Optical Coatings

D Ristau, Laser Zentrum Hannover e.V., Hannover, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Thin-film optical coatings are employed to adapt the properties of an optical surface to its specific application. Commonly, optical coatings comprise a sequence of single layers of different transparent materials with a thickness in the range of the operation wavelength. Depending on this coating design, the spectral characteristics of an optical surface can be tuned to nearly any required transfer function including a suppression of reflection for a broad spectral range by an antireflective coating or even the achievement of highest reflectance values up to 99.999% using a high reflecting layer stack. Besides these basic functions, optical coatings can be employed to realize a variety of transfer spectra specified for multiple wavelengths or wavelength ranges to manipulate radiation according to even extremely sophisticated demands in optical systems for laser technology, communication, or fundamental research. Moreover, the desired optical properties can be combined with additional specifications, for example, concerning the mechanical or chemical stability of an optical surface by selecting adapted coating materials. As a consequence of this enormous versatility, thin-film optical coatings can be found in nearly every product and development of modern optical technology today. Aside from the large market of consumer optics including ophthalmology, camera objectives, binoculars, and optical systems for data storage, progress in optical coating technology is predominantly driven by demanding applications in many future technology fields. As one prominent pace-setter, semiconductor lithography imposes new demands on optical coatings extending the operation wavelength of stepper systems in chip production to the extreme ultraviolet (EUV)-spectral range. Other examples can be found in communication technology, where thin-film filters with a bandwidth below

0.5 nm are required, or also in laser material processing, where coatings with high power handling capability and lifetime are indispensable for the development of cost-efficient laser tools.

In this article a concentrated review will be given on the design, production, and on specific properties of optical coatings. After an introduction to the fundamental principles of operation of coating systems, basic expressions for the optical transfer function of a single layer will be derived and extended to layer structures on the basis of the matrix method. Selected conventional and ion processes for the production of coatings will be discussed with respect to the achievable optical and structural properties of the coatings. Also, the thickness control of the coatings during production will be considered introducing the crystal monitoring technique as well as optical measurement strategies at a single wavelength or in a broad spectral range. The article will conclude with a review of the present state of optical coatings and an analysis of future trends.

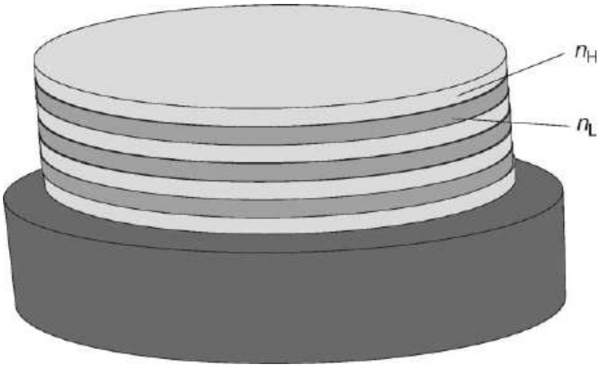

Figure 1 Basic structure of a thin-film system. Transparent layers of at least two different materials (n_H, n_L) are deposited on a substrate with refractive index n_T.

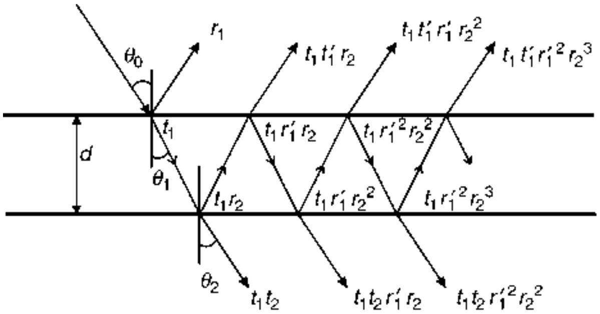

Figure 2 Schematic path of a plane wave incident onto a single layer with refractive index n_1. The reflection and transmission coefficients for the amplitudes at the interfaces $m = 1, 2$ are denoted by t_m and r_m, respectively. A prime added to the coefficients indicates a transfer of the wave opposite to the direction of incidence.

Design of Optical Coatings

The colorful appearance of optical coatings and microstructures, which can be observed on some butterflies and other insects, was present in nature long before mankind invented them. The beginning of the technical evolution of optical surfaces may be dated back to Greek culture, where polished metal surfaces were used as mirrors. Considering the application of transparent coatings, the investigations of Joseph Fraunhofer in the reflectivity of aged glass surfaces during the year 1819 should be mentioned as a first approach towards an antireflective coating. On the basis of controlled etching, this technique was further developed for different glass materials by Dennis Taylor in 1891 to realize the first antireflective coatings on lenses for applications in optical devices. In 1899, the Fabry–Perot theory was developed as a milestone in the theoretical description of multilayer structures. This theory describes the interference of partial waves reflected from two parallel optical surfaces, which can be considered as the fundamental arrangement of a multilayer structure (see Figure 1). In an optical thin-film system, layers of different transparent materials are composed on a stack deposited on the surface of the optical component (substrate refractive index n_T). In general, to adjust the spectral characteristics of the layer system to the demanded specifications, at least two layer materials with a high refractive index n_H and a low index n_L have to be selected. The choice of the thin-film materials is dependent on the application wavelength and the desired stability of the coating system as well as on their optical properties essential for the specific design and the targeted optical quality. Similar to the Fabry–Perot principle, the optical function of the layer structure is based on the interference of partial waves reflected and transmitted at the interfaces between the layers. The trace of the incoming wave and the induced partial waves are depicted in Figure 2 for a single layer, which can be considered as the basic unit for coating systems even with highest complexity. For a theoretical description of the transfer function of this single layer, the contributions of the individual partial waves have to be calculated and accumulated to form the total transmitted and reflected wave. In this classic approach, the incoming wave can be described by the function of a plane wave with amplitude E_0, wavenumber $k = 2\pi/\lambda$ incorporating the wavelength λ, and frequency ω:

$$E(x, t) = E_0 \exp[-\mathrm{i}(kx - \omega t)] \qquad [1]$$

At the first interface, a part of this plane wave is directly reflected with an amplitude of $A_0 = E_0 r_1$

(order zero), which can be calculated using the reflection coefficient r_1 given by the Fresnel formulas. After passing the first interface, a reflection and transmission of the incoming wave at the second interface has to be considered. Also a phase shift has to be taken into account, which is introduced by the path of the wave from the front to the rear interface of the single layer. This phase shift δ_i is dependent on the thickness d_i of the layer and the refractive index n_i:

$$\delta_i = \frac{2\pi n_i d_i \cos\theta_i}{\lambda} \qquad [2]$$

For oblique incidence also the angle of incidence θ_i in the layer, which has to be calculated by Snell's law, has to be taken into account. The index i has been chosen in this consideration for convenience to account for the layer number in coating systems, which will be discussed later. Following the wave reflected the first time inside the layer, a further phase shift for the path from the rear side to the front side has to be taken into account. This partial wave is then transmitted by the front surface resulting in an amplitude A_1 of the first-order wave leaving the layer:

$$A_1 = t_1 t'_1 r_2 E_0 \exp(-i2\delta) \qquad [3]$$

where t'_1 is the transmission coefficient for the amplitude of the wave passing the layer at the interface to the ambient. By following the wave further through the layer, the amplitudes A_k of each partial wave of order k can be determined:

$$A_k = t_1 t'_1 r_2 E_0 \exp(-i2\delta) * (r'_1 r_2 \exp(-i2\delta))^{k-1} \qquad [4]$$

Obviously the contributions of the partial waves follow the rule of a geometric expansion, and therefore the total amplitude of the wave reflected by the layer, which is the sum of the amplitudes of the partial waves, can be expressed in closed form:

$$A_{\text{tot}} = \left[r_1 + \frac{t_1 t'_1 r_2 \exp(-i2\delta)}{1 - r'_1 r_2 \exp(-i2\delta)} \right] E_0 \qquad [5]$$

The reflection coefficient r_S of the single layer is given by the ratio between the total amplitude of the reflected wave and the amplitude of the incoming wave:

$$r_S = \frac{r_1 + r_2 \exp(-i2\delta)}{1 + r_1 r_2 \exp(-i2\delta)} \qquad [6]$$

In this equation, the coefficients r'_1 and t'_1 have been replaced by using the relations

$$t_1 t'_1 = (1 - r'_1)(1 - r_1) = (1 - r_1)(1 + r_1) = 1 - r_1^2 \qquad [7]$$

In principle, this method of compiling the contributions of the partial waves reflected at the layer interfaces can be extended also to the calculation of systems with more than one layer. But, considering the enormously increasing number of partial beams, the resulting equations become extremely complicated for multilayer structures. Therefore, the matrix formalism, which can be deduced from the boundary conditions of the electric and magnetic field at the interfaces, is usually employed for the calculation of thin-film systems. The major advantage of this approach, which was developed in the 1940s, is the representation of each layer by a single matrix M_i containing all specific parameters of the film. The matrix M_i of a thin film at position i within the layer system relates the electric (E_{i-1}) and magnetic (H_{i-1}) field strength at the front face to the field strength values (E_i and H_i) at the rear face of the layer:

$$\begin{pmatrix} E_{i-1} \\ H_{i-1} \end{pmatrix} = \mathbf{M}_i \begin{pmatrix} E_i \\ H_i \end{pmatrix} = \begin{pmatrix} \cos\delta_i & \dfrac{i}{n_i}\sin\delta_i \\ i n_i \sin\delta_i & \cos\delta_i \end{pmatrix} \begin{pmatrix} E_i \\ H_i \end{pmatrix} \qquad [8]$$

By considering the boundary conditions for the field strengths at the interfaces, the matrix components given in eqn [8] can be determined. On the basis of the matrix formalism, the transfer function of a layer system can be simply accomplished by multiplication of the matrices M_i representing the constituent single layers. Thus, the transfer matrix M_S of a stack, which is formed by a number K of single layers with layer 1 located at the first interface with respect to the incoming wave, is given by

$$\mathbf{M}_S = \mathbf{M}_1 \mathbf{M}_2 \mathbf{M}_3 \ldots \mathbf{M}_i \mathbf{M}_{i+1} \ldots \mathbf{M}_K \qquad [9]$$

For calculating the reflection coefficient r_{SK} of the entire arrangement including the substrate (index of refraction n_T) and the ambient medium (index of refraction n_0), again the ratios of the amplitudes have to considered:

$$r_{SK} = \frac{n_0 \mathbf{M}_{11} + i n_0 n_T \mathbf{M}_{12} - i\mathbf{M}_{21} - n_T \mathbf{M}_{22}}{n_0 \mathbf{M}_{11} + i n_0 n_T \mathbf{M}_{12} + i\mathbf{M}_{21} + n_T \mathbf{M}_{22}} \qquad [10]$$

As a consequence of its perceptible structure, the matrix method can be readily transferred into computer codes, which cover the calculation of all important properties of even extremely complicated layer systems. In particular, considering the fundamental relations of the matrix method, also the field strength distribution within the layer structure can be determined. In software packages, which are commercially available as standard tools for optical thin-film technology, often the film thickness D_i of the layers is expressed in the units of QWOT (quarter wave optical thickness) at the design wavelength λ_Z:

$$D_i = \frac{4n_i d_i}{\lambda_Z} \qquad [11]$$

In the case of normal incidence, the matrices M_i reduce to simple expressions for layers of integer QWOT values, because the corresponding phase shift terms are multiples of $\pi/2$. Clearly, an integer QWOT value of optical thickness also represents an extreme value in the spectral behavior at the design wavelength. This fact is often exploited for a precise thickness control by optically monitoring the reflectance or transmittance of the growing layer at the design wavelength λ_Z: by recording the transfer behavior as a function of time during production, extreme values and accordingly thickness values of multiple QWOT can be accurately determined. Thus, if a layer structure is composed only of multiple QWOT layers, switching between the layer materials can always be precisely performed at a minimum or maximum read-out of the optical monitor tuned to the design wavelength. In respect of this advantage in production, most coating designs are based on QWOT layers indicating the practical importance of the unit QWOT. For an efficient description of the designs, often a notation in capital letters is used, which indicate a 1 QWOT layer of a certain coating material. For example an alternating QWOT-stack with 11 layers of a high-index material (H, refractive index n_H) and a low-index material (L, refractive index n_L) is described by the sequence HLHLHLHLHLH or even more condensed by $(HL)^5H$. The QWOT stack is the fundamental design for dielectric high reflecting mirrors and represents the basis for a variety of functional coatings in optical thin-film technology. By applying the matrix method, an approximate expression for the reflectance R_S at the design wavelength can be derived for such a stack:

$$R_S \approx 1 - 4n_0 n_T \frac{n_L^{2N}}{n_H^{2(N+1)}} \qquad [12]$$

This expression is only valid for a large number N of QWOT layers and underlines the role of the materials for the practical production of optical coatings. Highest reflectivity with lowest number of layers can be achieved only by selecting materials with a high contrast of their refractive indices. This contrast also influences the spectral broadness of the high-reflecting region of the stack (Figure 3).

Besides highly reflecting mirrors, antireflective coatings are another important class of functional coatings for modern optics and laser technology. Antireflective coatings often involve non-QWOT layer structures of more than two materials in order to accomplish the demands of lowest residual reflectivity for a broad spectral range. Eye glasses with antireflective coatings covering the entire visible spectral range can be considered as a prominent example for consumer optics, which impose requirements on the coatings also in respect to their stability against cleaning and other environmental influences during normal use. In this context, the introduction of optical plastic materials is an additional challenge in consumer optics and ophthalmology, which has to be mastered by adapted designs and production processes. In contrast to the typical demands of consumer optics, antireflective coatings on laser windows are

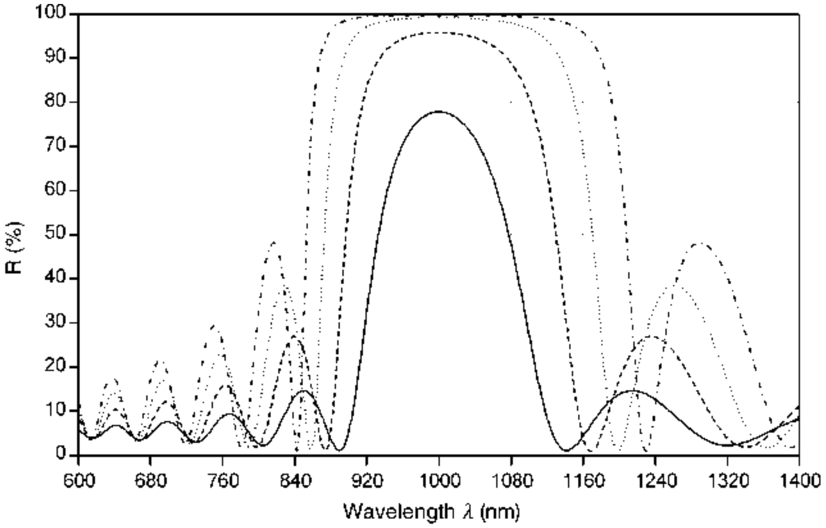

Figure 3 Spectra calculated for high-reflecting QWOT stacks (15 layers, substrate index $n_T = 1.52$) composed of a low-index material ($n_L = 1.46$) and different materials with high indices of refraction varying between $n_H = 1.7$ (solid), 1.9 (dashed), 2.1 (dotted) and 2.3 (dash-dotted). The spectral broadness and the reflectivity increase with the ratio n_H/n_L.

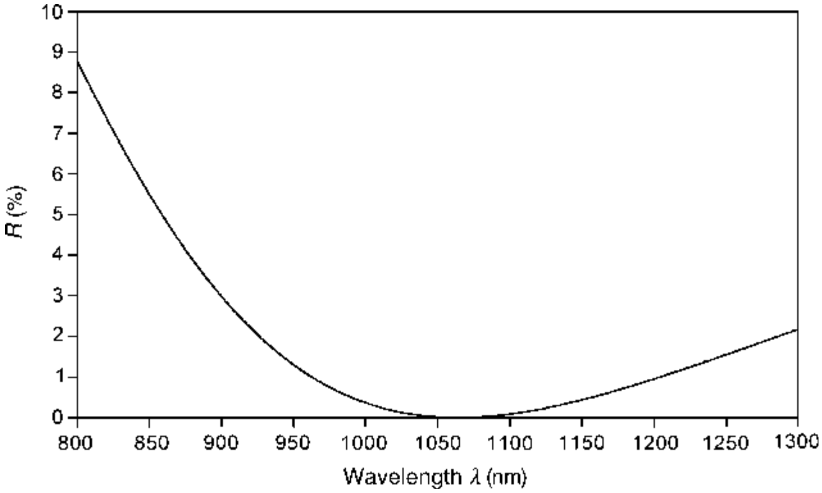

Figure 4 Spectrum of an antireflective coating for the wavelength 1.064 μm, which is designed as a V-coating with two layers of Ta_2O_5 ($n_H = 2.03$; $d_H = 0.383$ QWOT) and SiO_2 ($n_L = 1.46$; $d_L = 1.308$ QWOT) for a quartz substrate (substrate index $n_T = 1.46$).

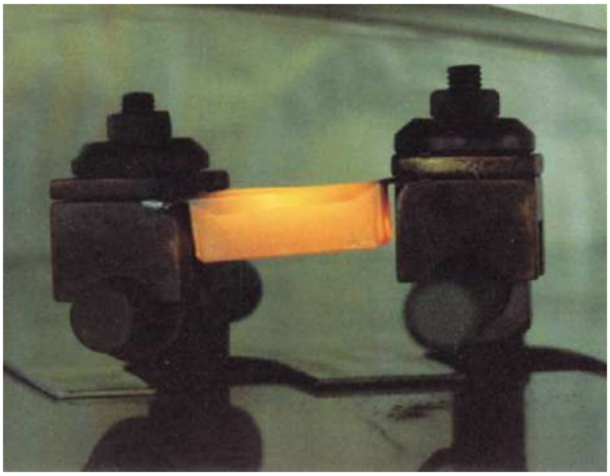

Figure 5 Boat evaporation source in operation. The deposition material is put into the boat which is made of a metal with high melting point. A high electric current is conducted through the boat for heating to temperatures up to 2000 °C.

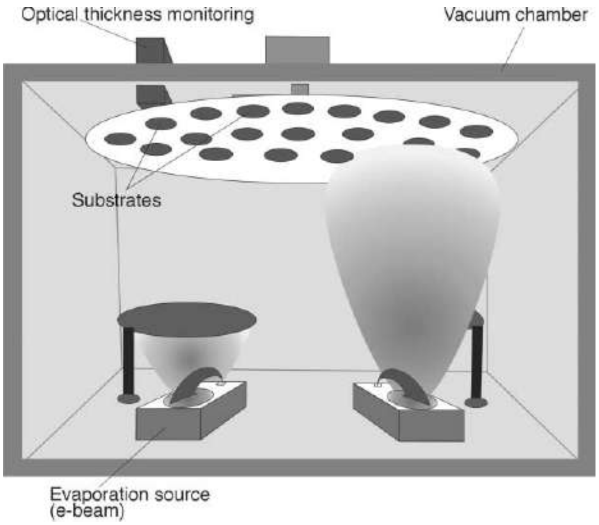

Figure 6 Outline of a modern box coater used for the production of optical coatings with e-beam deposition processes.

often restricted to one wavelength, but have to fulfill extreme requirements with respect to optical losses and power handling capability. Therefore, most antireflective coatings in laser technology are composed only of two thin layers to minimize the total thickness, which is often a direct measure of the expected losses of the layer system. With such a coating, which is often called V-coating according to its spectral characteristic (Figure 4), the reflectivity of the window surface can be suppressed ideally for one wavelength.

Production of Optical Coatings

The industrial production of optical coatings started in the 1940s on the basis of glass bell jar systems equipped with diffusion pumps and boat evaporators for the deposition of single layers of metals for mirrors and fluoride materials for antireflective coatings. Metal containers are used as boat evaporation sources. These containers are filled with the deposition material and heated by electric currents up to several 100 A (Figure 5). In the course of the rapid improvement of vacuum systems and especially the development of the electron beam evaporator over the next three decades, new materials with improved optical and mechanical quality could be combined to form more complicated dielectric coating systems. The layout of an e-beam deposition plant of these times, which can still be considered as fundamental for the advanced deposition systems of today, is illustrated in Figure 6. In such a box coater, the substrates are installed in a calotte which is rotated to achieve a uniform distribution of the layer thickness. The deposition materials are put in water-cooled crucibles and are subjected to an electron beam operated at currents in the range between 0.1 A to more than 10 A and voltage levels of several kilovolts. As a consequence of the interaction with the energetic electrons, the deposition material can be heated to high temperatures above 5000 K allowing the evaporation of most of the desired deposition materials. These achievable high temperatures in conjunction with the good controllability are the major advantages of e-beam sources (Figure 7) in comparison with the classical boat evaporation process. In particular, e-beam evaporation opened the way for the deposition of oxide materials and other optical materials with high evaporation temperatures, which exhibit optical properties as well as a high mechanical and chemical stability superior to the material classes accessible with boat evaporation. However, the cost efficient boat evaporation technique is still applied today for the deposition of metals or temperature-sensitive materials for coating systems in the vacuum ultraviolet (VUV)/UV and mid-infrared (MIR) spectral ranges. If the shutter above the deposition source is opened, the evaporated material reaches the substrates and forms the layer by condensation. This condensation process is sensitively dependent on the environmental conditions of the substrates and the deposition sources. Sufficient microstructural properties of the growing layer can generally be achieved only under fine vacuum below 100 Pa and heating of the substrates to temperatures in the range of some 100 °C. Even under well-optimized conditions, the typical microstructure of such thermally evaporated coatings exhibits specific deficiencies (Figure 8) which influence the optical and mechanical properties of the layer systems. Often, microcrystallites with diameters in the range of a few nm are observed, which are arranged in columns of a typical size growing from the substrate interface to the outer surface of the layer. For a simulation of the growing mechanisms for this low energetic deposition process, uncomplicated models with a few input parameters already lead to satisfactory results. Assuming a kinetic energy of 0.1–0.3 eV, which is typical for thermal evaporation according to the Boltzmann distribution, the impinging atoms can move only a small distance or even adhere directly at their position of incidence on the growing layer. As a consequence of this small mobility, the atoms are not able to reach their energetically optimum positions in the layer, and they even obstruct the path of the following atoms by shadowing. On the basis of these assumptions, computer simulations of the growing process result in structures which are comparable to the columnar growth observed in practice. The packing density of such a columnar structure is significantly below the bulk values of the corresponding deposition materials, and it contains also a high number of defects and voids, which reduce the optical quality of the coating. In particular, the coating tends to

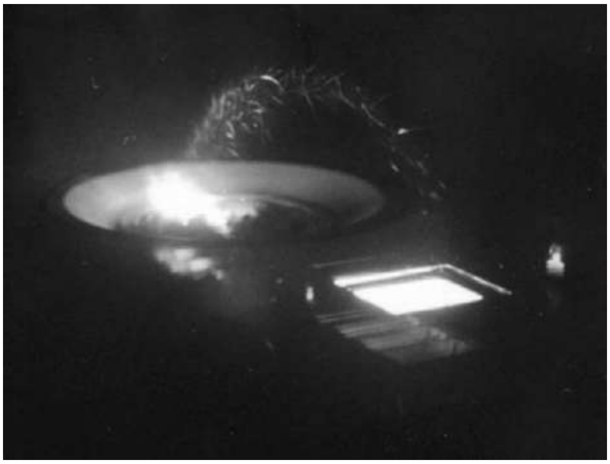

Figure 7 Electron beam source in operation. The deposition material is put in a water-cooled crucible and subjected to an e-beam. Since the zone heated by the e-beam is entirely embedded in the surrounding deposition material, very low contamination levels and temperatures above 5000 °C can be achieved. For the generation of the e-beam an arrangement similar to the system of a cathode ray tube is employed: the electrons emitted by a hot filament are focused by an aperture and accelerated by a high voltage.

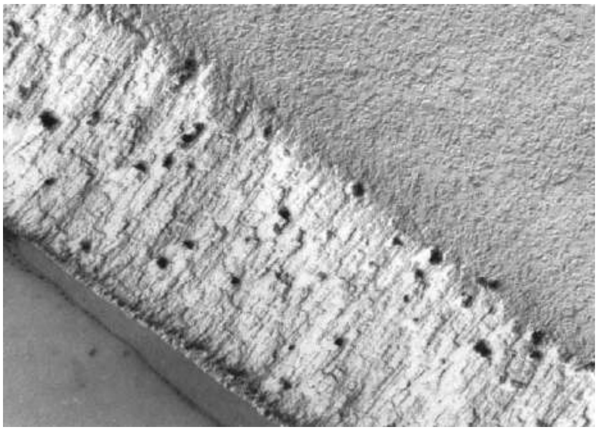

Figure 8 Microstructure of a MgF_2 single layer (geometric thickness approximately 780 nm) deposited by conventional boat evaporation. For the preparation of the TEM picture, the coating has been fractured and replicated.

adsorb water from the normal atmospheric environment into its internal structure leading to a shift of the spectral characteristics as a function of the ambient conditions. Besides this loss of spectral stability, the adsorbed water also induces an increased optical absorption of the coatings in the VUV and the MIR spectral ranges.

The increasing demand in laser technology and other optical applications for coatings far exceeds the quality level of conventionally deposited coatings and this was the driving force for the development of advanced concepts for the production of optical coatings during the last three decades. Considering the fundamental growth mechanism of conventional optical coatings, two basic approaches for an improvement of the microstructure are available. On one hand, the mobility of the atoms on the surface of the growing layer can be increased by the insertion of additional energy from other sources. This concept leads to the development of ion-assisted deposition processes, where the additional energy is transferred to the adatoms by collisions with ions generated directly in the plant by an ion source at energy levels of approximately 30–300 eV (Figure 9). Especially, by the deposition of oxide materials, which can be assisted directly with oxygen ions for balancing their stoichiometry, coating systems with high environmental stability and low optical losses can be achieved for the VIS up to the MIR spectral range. On the other hand, the energy of the adatoms can be directly increased in sputter processes which deliver deposition energies orders of magnitude higher than the thermal energy range. Besides adapted activated reactive magnetron sputter processes, particularly ion beam sputtering (IBS) was developed for a coating process delivering high-end coatings mainly for laser technology with an unsurpassed optical quality (Figure 10). For example, mirrors with total losses below one ppm can be produced for the Nd:YAG laser (wavelength 1.064 μm) with adapted IBS processes.

In addition to stable deposition processes, precise control of the layer thickness with an accuracy below 1 nm is a further important step towards the realization of complex optical coatings. Most deposition plants are equipped with quartz crystal monitors, which measure the optical thickness on the basis of the resonance frequency shift of an oscillating quartz crystal induced by the increase in mass of the deposited coating. The relative accuracy of these monitors ranges from 2% to 10% and is well suited for the regulation of the deposition rate. In order to achieve the required precision for high-quality coatings, optical monitors are often employed, which record the transmittance of the growing layer system at a single wavelength or even in a broad spectral range. Since these systems account directly for the optical properties of the deposited layer system, an efficient and accurate control of the film thickness is possible allowing the production of extremely sensitive filters with bandwidths below 1 nm for telecommunications or mirrors with defined group delay dispersion for fs-laser systems (Figure 11).

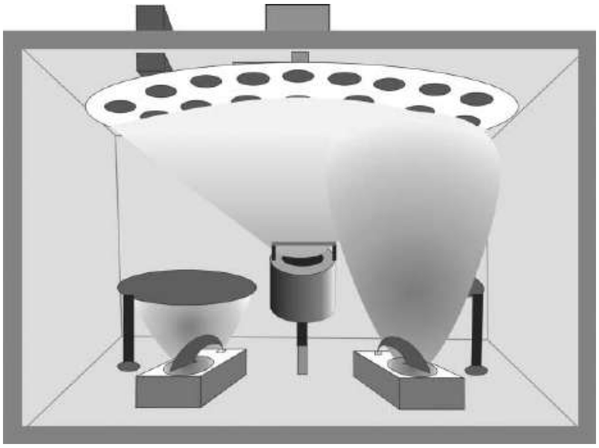

Figure 9 Ion assisted deposition. An additional ion source is installed, which delivers inert or reactive ions impinging on the growing layer. The energy transfer from the ions to the adatoms results in improved microstructure of the produced coatings with higher packing density.

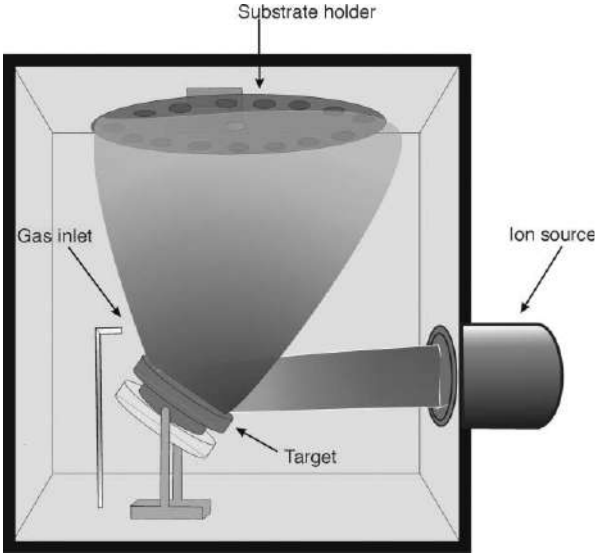

Figure 10 Principle of the ion beam sputtering (IBS) process. Inert ions with an energy in the range 1–2 KeV are generated in an ion source and directed onto the target which consists of the desired deposition material. Material is sputtered from the target and reaches the substrates. As a consequence of the high energy of the adatoms (5–30 eV) the formed layer exhibits a superior microstructure.

Quality Parameters of Optical Coatings

Besides the spectral characteristics, a variety of specific parameters has to be considered for each application of optical coating systems (Table 1). For laser systems the laser-induced damage threshold and the optical losses have to be evaluated in the context of technical and also economic requirements. Absorption is a loss channel, which transforms a fraction of the impinging radiation energy into heat, inducing a temperature increase in the coating. This generated temperature profile influences the imaging properties of the optical component via thermal expansion and temperature-induced alterations of the refractive indices of the coating as well as the substrate materials. Despite numerous technical problems, absorption always implies the loss of laser energy which is produced with high economic expense. Absorption losses are related to defects and impurities, which are introduced during the production process into the layer by a variety of contamination mechanisms. Furthermore, an unbalanced stoichiometry with a higher metal content is often observed as a cause for increased absorption in dielectric optical coatings, which are mainly composed of binary metal compounds. Absorptance values of optical components are determined by laser calorimetric techniques described in ISO 11551. In this measurement scheme, the temperature increase in the sample is monitored during irradiation with a well-characterized laser beam. After this heating phase, the laser beam is blocked and the temperature curve during the cool-down cycle is also recorded. On the basis of the measured temperature development, the absorptance can be calculated taking into account the laser power and the heat capacity of the specimen.

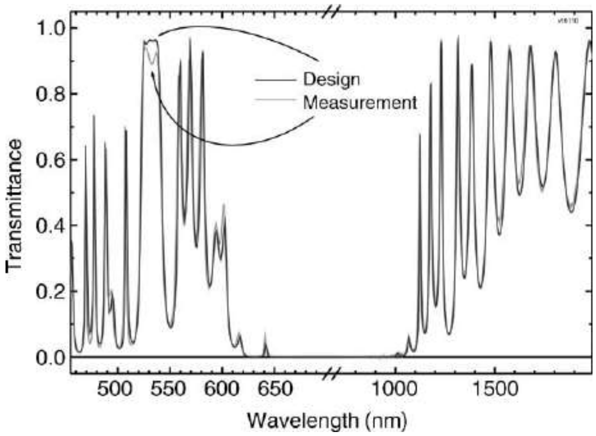

Figure 11 Example of a complicated layer system deposited under the control of an advanced optical broadband monitoring device: double chirped mirror (56 non-QWOT layers of SiO_2/TiO_2). Only small deviations between the calculated spectrum and the spectrum of the produced coating can be observed.

All mechanisms, which deflect radiation from its specular direction, contribute to optical scattering representing the second major loss channel in optical coatings. Since scatter losses are mainly attributed to microstructural imperfections and the surface roughness of the layers, a scaling with wavelength can be deduced from the corresponding theoretical models. Especially in the VUV/UV spectral range, high scatter values are apparent and may dominate the losses of the optical components. For longer wavelengths, the contribution of scattering is decreased to a point in the MIR spectral range where it can be neglected in comparison to the absorption losses. Scatter effects

Table 1 Selected quality parameters of optical coatings and surfaces in conjunction with corresponding ISO standards and measurement principles

Specification	*Parameter/Unit*		*Standard/Measurement principle*
Laser induced damage threshold (LIDT)	cw-LIDT	W/cm	ISO 11254-1: cw-laser irradiation
	1 on 1-LIDT	J/cm^2	ISO 11254-1: irradiation with single pulses
	S on 1-LIDT	J/cm^2	ISO 11254-2: repetitive irradiation with pulses
	Assurance	J/cm^2	DIS 11254-3: irradiation sequence
Optical losses	Absorptance	ppm	ISO 11551: laser calorimetry
	Total scattering	ppm	ISO 13696: integration of scattered radiation
Transfer function	Reflectance	%	CD 13697: precise laser ratiometric method
	Transmittance	%	DIS 15368: spectrophotometry
Surface quality	Scratch/Digs		ISO 10110: 13 parts containing different types of imperfections
	Roughness		
Stability	Abrasion		ISO 9211: different test methods
	Environmental Stability		ISO 9022: 21 parts containing a variety of conditioning methods

may impair the imaging properties or the resolution, and they reduce the efficiency of an optical system. Besides the economic aspect of each loss effect, scattering can also be considered as relevant for safety in high-power laser systems, where some 100 ppm of scattered radiation carry enough power to injure or blind the operator. In the ISO 13696 standard, measurement procedures based on the collection or integration of the total radiation scattered by the sample are described.

Depending on their application, coatings have to withstand also manifold mechanical, chemical, and environmental influences. Severe mechanical and chemical requirements are imposed onto coatings for ophthalmology or consumer optics, which are often cleaned with abrasive and aggressive cleaning solvents and tissues. To qualify coatings for these applications, various testing procedures have been developed and are routinely employed in industrial quality management. Coatings for laser systems and precision optics are predominantly certified according to conditioning methods described in the ISO standard series 9211. Test procedures with respect to environmental influences are extensively outlined in the more than 20 parts of ISO 9022, which is dedicated to a variety of climatic, mechanical, and chemical conditioning cycles.

Summary and Outlook

Optical coatings are the major tool for engineering the properties of functional optical surfaces for an extensive spectrum of applications. On the basis of the matrix algorithm, a comprehensive theoretical model for the spectral transfer characteristics of optical coatings is available, which has been transferred to commercial computer programs. Now, most design problems can be solved with these computer codes, whereas the reproducible and precise production of optical coatings with high quality could only be partially achieved until today (Table 2). Besides conventional thermal processes, new ion-assisted deposition processes and sputtering concepts are gaining importance in optical thin-film technology. For the quality management of optical coatings, a variety of often standardized testing procedures has been developed during the last two decades including the determination of the power handling capability, the measurement of the optical losses and a broad spectrum of conditioning methods for optical coatings.

Demanding future challenges will be imposed on optical coatings by the development of semiconductor lithography in the EUV spectral range to achieve nodes well below 100 nm in chip production. Another important trend is directed by new laser applications in material processing, medicine and communication technology, operating at extremely high repetition rates and power levels, or with ultrashort pulses in the fs-regime. Other requirements will be defined by the rapid development of optical technology towards a broad range of special products with short innovation cycles obligating significant progresses in thin-film technology in the direction towards flexible and reproducible process concepts

Table 2 Selected quality parameters of optical coating systems for laser applications

Laser, wavelength	*Type*	*Absorption ISO 11551*	*Total scattering ISO 13696*	*Laser induced damage threshold, ISO 11254*
157 nm, F_2 excimer	HR/th		1–4%	
193 nm, ArF excimer	AR/th	0.7–2.5%	0.2–0.5%	1–2 J/cm^2 (1on1, 20 ns)
	HR/th	0.4–2.0%	0.2–2.5%	2–4 J/cm^2 (1on1, 20 ns)
248 nm, KrF excimer	AR/th		<0.025%	10 J/cm^2 (1on1, 30 ns)
	HR/th	<500 ppm	<0.2%	>20 J/cm^2 (1on1, 30 ns)
	HR/IB		<0.1%	>3 J/cm^2 (1on1, 30 ns)
633 nm, HeNe laser	HR/th	<30 ppm	<30 ppm	–
	HR/IB	<5 ppm	<5 ppm	–
1.064 μm, Nd:YAG	AR/th	<20 ppm	<100 ppm	>60 J/cm^2 (12 ns, 0.25 mm)
	HR/th	<50 ppm	<100 ppm	>100 J/cm^2 (12 ns, 0.25 mm)
	HR/IB	<1 ppm	<1 ppm	>80 J/cm^2 (12 ns, 0.25 mm)
10.6 μm, CO_2 laser	AR/th	<0.16%	–	>20 J/cm^2 (100 ns, 1.4 mm) >2 kJ/cm^2 (1.2 ms, 250 μm) >3 kW/mm (cw, 100 μm)
	HR/th	<0.10%	–	>25 J/cm^2 (100 ns, 1.4 mm) >2 kJ/cm^2 (1.2 ms, 250 μm)

Types: HR: highly reflecting mirror; AR: antireflective coating; th: thermal evaporation; IB: ion beam sputtering.

for coatings with increasing complexity. In all fields of modern optics, optical thin films will represent an enabling technology which will play a key role for many future applications and products.

List of Units and Nomenclature

A_0	amplitude of partial wave of order zero
A_1	amplitude of first-order partial wave
A_k	amplitude of the partial wave of order k
d_i	thickness of layer number i
D_i	film thickness expressed in units of QWOT
E_0	amplitude of a plane wave
E_i, H_i	electric field strength and magnetic field strength at the rear front of layer number i
IAD	ion-assisted deposition
IBS	ion beam sputtering
k	wavenumber of a plane wave $k = 2\pi/\lambda$
M_i	matrix of layer number i
M_S	composite matrix of a layer stack of K single layers
n_0	index of refraction of the ambient medium
n_H	refractive index of a high-index material
n_i	refractive index of layer number i
n_L	refractive index of a low-index material
n_T	refractive index of the substrate
QWOT	quarter wave optical thickness: unit for the thickness of the layer
r_1, r'_1	reflection coefficients of the ambient–layer interface
r_2	reflection coefficient at the layer–substrate interface
r_S	reflection coefficient of a single layer
r_{SK}	reflection coefficient of a layer stack of K single layers
R_S	reflectance of a QWOT stack
t_1, t'_1	transmission coefficients of the ambient–layer interface
δ_i	phase shift of a plane wave in layer number i
λ	wavelength of a plane wave
λ_Z	design wavelength
θ_i	angle of incidence in layer number i
ω	frequency of a plane wave

See also

Optical Coatings: Laser Damage in Thin Film Coatings.

Further Reading

Guenther AH (ed.) (2002) *International Trends in Applied Optics*. Bellingham, Washington: SPIE Press.

ISO 11551 (1997) *Test Method for Absorptance of Optical Laser Components*. International Organization for Standardization.

ISO 13696 (2002) *Test Method for Radiation Scattered by Optical Components*. International Organization for Standardization.

ISO 9022 (1996, 1997) *Optics and Optical Instruments – Environmental Test Methods – Parts 1–21*. International Organization for Standardization.

ISO 9211 (1994–2001) *Optical Coatings, Parts 1–4*. International Organization for Standardization.

ISO/FDIS 15368 (2003) *Measurement of Reflectance of Plane Surfaces and Transmittance of Plane Parallel Elements*. International Organization for Standardization.

Jacobson MR (1989) *Deposition of Optical Coatings*. SPIE Milestone Series, Volume MS 6. Bellingham, Washington: SPIE Press.

Kaiser N and Pulker HK (2003) *Optical Interference Coatings*. Berlin: Springer.

Liddell HM (1980) *Computer-Aided Techniques for the Design of Multilayer Filters*. Bristol, UK: Adam Hilger.

Macleod HA (2001) *Thin Film Optical Filters*. Bristol, UK: Institute of Physics Publishing.

Rancourt JD (1996) *Optical Thin Films*. Bellingham, Washington: SPIE Press.

Thelen A (1988) *Design of Optical Interference Coatings*. New York: McGraw-Hill.

X-Ray Coatings

P Dhez, Université Paris Sud, Orsay, France

© 2005, Elsevier Ltd. All Rights Reserved.

In all spectral ranges, there exist two principal options for coating an optical element. This may be a simple layer, as for example a metallic layer is frequently used in the visible spectral region. The second option is to use a multilayer coating, making use of the optical properties of two or more materials to achieve enhanced reflectivity over a limited spectral range. In both cases, a knowledge of the refractive index of the materials (real or complex) is required, and of course one must choose materials that can be coated in a practical deposition process. A frequent side-advantage of coating is to make it possible to choose the material of the substrate for its mechanical or cooling properties without regard to the optical properties of the coatings.

Originally (at the beginning of the twentieth century) the progress of X-ray optics was hampered by a poor understanding of the nature of these rays, and in particular by the observation that they are very weakly reflected and refracted. After the electromagnetic nature of these rays had been confirmed, the problem of producing polished surfaces that were sufficiently smooth on an atomic scale to reflect X-rays still remained. Optical coatings gave little hope of improvement due to the growth of roughness during the coating process. Thus early attempts to make X-ray mirrors produced poor results and little progress was made until the 1960s. However, in the past 30 years, much technical progress has been made, and now many optical principles used in the visible spectral region have been extended to very short wavelengths. Even refractive and diffractive focusing lenses (not treated here) have been implemented in the X-ray region.

In the first section we outline the history of the discovery of the properties of X-rays, which were achieved mainly in experiments using crystals. This allows us to understand how their electromagnetic nature was discovered, and how they interact with materials. We then describe the phenomenon of the 'total reflection' of X-rays, which occurs only at very high angles of incidence ('grazing incidence') and its main applications in X-ray optics. Finally, we outline the principles of mirrors based on interference, their applications in the X-ray region, their advantages and their limitations.

X-rays Compared to the Other Electromagnetic Spectral Regions

Infrared radiation was discovered by W. Herschel in 1800, and ultraviolet rays by J. Ritter in 1801. These were the only 'invisible rays' known before Maxwell's contribution to the understanding of the electromagnetic spectra through the development of his equations in 1869. The discovery of radio waves by H. Hertz in 1887 confirmed that our eyes are blind to most of the electromagnetic spectrum.

In 1895, W. Roentgen discovered some previously unknown 'rays' that had the surprising property of being able to penetrate low-density material opaque to visible light. This discovery was a side-effect of his studies of low-pressure electrical discharges, fashionable at the time. The nature of these unknown or 'X'-rays as Roentgen named them, remained unknown for 12 years or so. Were they electromagnetic waves or particles? Roentgen used lenses, prisms and mirrors to try to detect a reflective, refractive or focusing effect, with inconclusive results. Although the electromagnetic nature of the X-rays was favored, the magnitude of the frequency or the refractive index was unknown. We note that natural radioactivity and gamma rays were discovered a year after X-rays, and the same questions concerning their nature were raised. However, in neither case did the ignorance of the nature of the radiation impede their use in medical radiography or in cancer therapy.

In 1906, the experiments of C. Barkla provided the first evidence of the electromagnetic nature of X-rays. In scattering X-rays from blocks of a light element (in this case, carbon), he was able to show that the scattered radiation was polarized, and therefore had the nature of a transverse wave. This kind of experiment had been performed almost a century previously to demonstrate the transverse wave nature of visible light. In 1916, following a suggestion by M. von Laue that X-rays might be diffracted from a grating of period comparable to their wavelength, Friedrich and Knipping succeeded in diffracting X-rays from the atomic planes of a natural crystal. This experiment established definitively the electromagnetic nature of X-rays and also the scale of their wavelength.

In 1915 W. L. Bragg summarized the diffraction rules for crystals through his equation [1], now known as the simplified Bragg law:

$$2d_i \sin \theta_i = N\lambda_i \qquad [1]$$

This equation relates the diffracted X-rays wavelength λ_i to a set of parallel atomic planes separated by a distance d_i and to the incidence angle θ_i. N is an integer indicating the order of diffraction, as in the grating equation. In using this equation, care is needed because of two peculiarities compared to the usual mirror or grating equations in the visible range. In the Bragg equation, the angle of incidence θ is measured from a given atomic plane to the incident rays. However, inside a crystal several different atomic planes exists, not only the planes parallel to the surface as indicated in Figure 1. In addition, θ is not measured from the normal as is usual in optics; hence the angle θ is referred to as the glancing or grazing angle of incidence.

We note that the Bragg equation is not restricted to X-rays, but is a coherent scattering phenomenon which is common to all cases where waves, such as

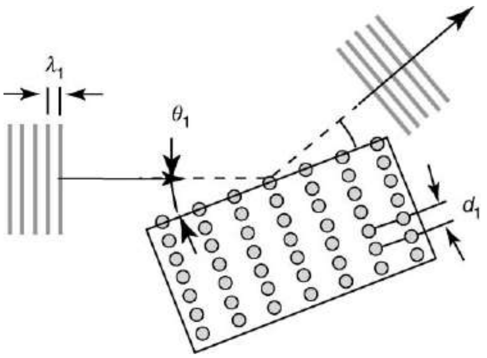

Figure 1 The geometry for the conditions leading to selective reflection on a periodic medium.

sound waves or ripples on water, are diffracted by regular arrays of scattering centers.

In order to visualize the phenomena, we refer to Figure 1 which is a schematic representation of a plane wave incident upon a two-dimensional array of scattering centers (electrons around nucleus, in the case of X-rays), shown here as dots. As the sine function has a maximum value of 1, the maximum wavelength λ that can be reflected from a spacing d is given by $\lambda = 2d_{\max}$. As the largest value of $2d$ that can be found in natural crystals is about 2 nm (in natural muscovite mica, for example), X-rays of 2 nm, with an equivalent energy of about 620 eV, are the 'softest' (least energetic) that can be reflected in this manner.

The model shown in Figure 1, similar to that proposed by M. von Laue, does not take into account the detailed interaction of the atoms with the X-rays, and therefore ignores effects such as absorption and refraction, and discrete atomic transitions. A.H. Compton attempted the first treatment of the interaction of matter with X-rays from an atomic viewpoint. In 1917, he developed a classical theory in which the electrons bound to the nucleus are treated as harmonic oscillators. In this way, he was able to derive an expression for the atomic scattering coefficients and to describe resonance effects at frequencies corresponding to specific atomic transitions. A few years later, he was able to establish a link between these 'atomic scattering factors' and the macroscopic concept of refractive index. This new formulation allowed the use of the Fresnel equations, widely used in other spectral regions, to calculate the intensity of X-rays reflected from smooth, polished surfaces. These surfaces may be amorphous as well as crystalline.

The refractive index can be written in several ways; since it is close to unity over the X-ray range, the most common form is complex:

$$n = 1 - \delta - \mathrm{i}\beta \qquad [2]$$

where δ is the refractive index decrement, and the imaginary part β the absorption index. For photon energies corresponding to X-rays, both have small negative values, typically 10^{-3}–10^{-6}, and are correcting factors to the vacuum index 1. As in other photon energy ranges, the absorption and the refraction coefficient variations are linked through the Kramers–Krönig relations. In the visible range the index values are usually larger than 1, being below 1 only around specific optical thresholds. In the UV region, the index values become smaller than unity at energies above about 30 eV, which corresponds to the energy range for which atoms can be ionized.

At an early stage, Roentgen demonstrated X-ray absorption by sheets of materials. The demonstration of refraction took much longer and was more difficult. The first direct X-ray refraction index measurement was achieved in 1924, using a prism and an X-ray tube. Figure 2 is the scheme of the experiment, which is easy to perform in the visible range to measure the refraction of a transparent material. The footprint of the direct beam on the photographic plate is made before inserting the prism. In the visible range for which $n > 1$, the beam is refracted at larger angle than the glancing incidence angle, and so below the direct beam direction as shown in Figure 2a. At the contrary, as indicated in Figure 2b, the X-ray beam is refracted toward the prism surface and such an experiment effectively demonstrated that the index was slightly inferior to unity, as suggested by A.H. Compton. We will return

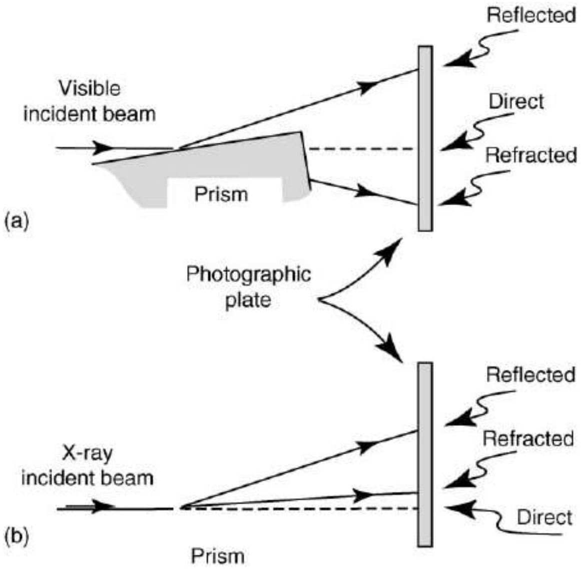

Figure 2 The geometry of the experiment used to evaluate the magnitude and the sign of the refractive index in the X-ray and the visible ranges.

to this peculiarity in the following section, when the total reflection phenomenon is described.

More details about the successive steps in our understanding and the pioneering works on X-rays can be found in standard textbooks.

Total Reflection X-Ray Mirrors

Using Fresnel's equations and tabulated values of the refractive index one can verify that, on any material, grazing incidence is necessary to reflect X-rays. Reflectivity calculations over the range from EUV to X-rays (30 eV to 10 keV), with approximations to allow calculations for rough surfaces, can be carried out using the web site of the Center for X-ray Optics of Lawrence Berkeley Laboratory http://www-cxro.lbl.gov, and curves similar to those presented below can be reproduced.

In the case of normal incidence, Fresnel's equation indicates that for any material the X-ray reflectivity $R^2 = |(n-1)/(n+1)|^2$ is very low, typically below 10^{-4}. This is a direct consequence of the X-ray index values, which are close to unity, as mentioned above.

In the visible range, normal incidence reflectivity from the surface of a transparent medium is low but can be increased just by going to glancing incidence. This is also the case in the X-ray region but in addition as the real part of the X-ray refractive index is lower than that of the vacuum, we have an 'external' total reflection phenomenon. From Figure 2 one can see how, for a sufficiently small incidence angle, the refracted wave becomes parallel to the interface. This is well known in the visible range (in which case it is the 'internal' reflection) and the phenomenon is used, for example, in fiber optics.

Figure 3 shows examples of calculated reflectivity variations versus grazing incidence angles, for perfectly smooth glass (SiO_2), nickel, and platinum surfaces. Curves are drawn for the 8 keV energy photons, corresponding to the Cu Kα line produced by conventional X-ray tubes. One observes that for this energy the reflectivity is low beyond a grazing angle around 1°, then rises quickly at low angles. This is typical of the total reflection phenomenon. The grazing angle corresponding to a 50% reflectivity is called the critical angle Θ_c, its value being roughly proportional to the refractive index decrement at the relevant photon energy. Notice that weaker absorption materials, lower Z, lead to more abrupt reflectivity variation but with smaller critical angles.

Such reflectivity variations at grazing incidence were systematically studied with X-ray tubes as early as the 1930s. The penetration depth of X-rays at grazing incidence was also studied. The evanescent wave inside the material, under total reflection conditions, is confined to a surface layer of a few tens of wavelengths. Hence, in the total reflection of X-rays, a coating having a thickness of a few tenths of a micron, on any substrate, is all that is required.

Figure 4 shows examples of the variation in nickel reflectivity calculated over the energy range 0.3–10 keV, for several grazing incidence angles. Nickel is one of the coatings commonly used for X-ray mirrors. The abrupt reflectivity variations and the dips correspond to the energies of the nickel atomic thresholds L (0.85 keV) and K (7.5 keV) for which the index tables predict a rapid change in absorption and refraction coefficients. Choosing an incidence angle provides a practical way of limiting the reflected spectral range. The energy cut-off of the

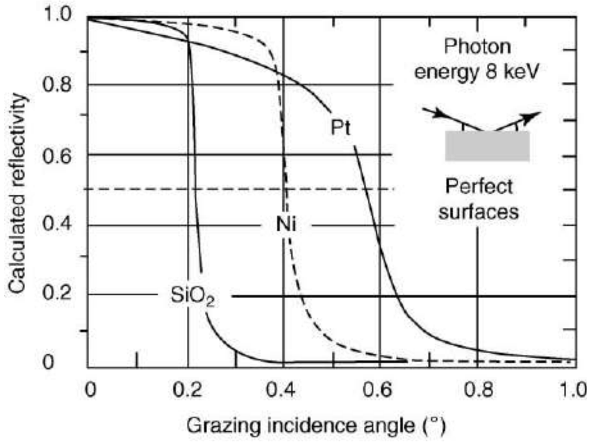

Figure 3 Examples of calculated reflectivity variations versus the incidence angle in the case of 8 keV X-rays on perfect platinum, nickel and glass surfaces.

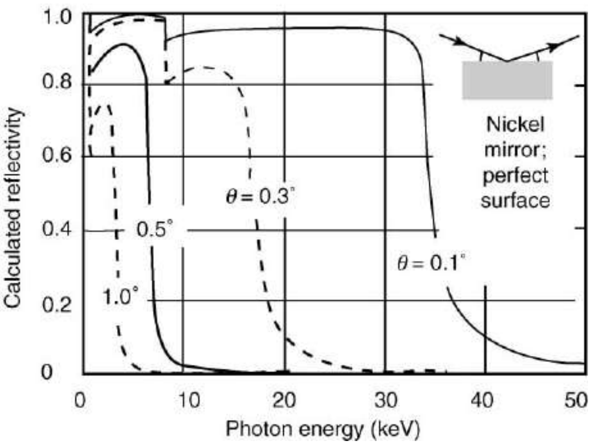

Figure 4 Calculated reflectivity variations versus photon energy, on a perfect nickel surface, for different incidence angles.

X-ray mirror acts as a filter for the higher-energy photons. That is effectively the method of suppressing the short wavelengths from an incident spectrum if they are not desired in an experiment.

When using highly divergent sources such as X-ray tubes, the restriction to operation at grazing incidence leads to a strong limitation of the collected solid angle, and thus of the beam size and angular aperture. These constraints are less stringent on highly collimated beams emitted by synchrotron sources. Thus, most synchrotron beam lines use total reflection optics in the form of curved mirrors, to refocus the beams and remove the undesired high X-ray energies. Recently, X-ray imaging optics, consisting of a series of parabolic and hyperbolic nested grazing incidence mirrors to increase the collected aperture, have been launched on satellites.

The reflectivity curves in Figures 3 and 4 are calculated for perfect surfaces. Real surfaces, even polished surfaces, have roughness created by defects over a large and continuous range of spatial frequencies, ranging from several centimeters down to the atomic scale. Coating a surface replicates the large period defects. Eventually new defects are added at near-atomic scale by nucleation of the coating material. Interferometric methods have been developed to study the range of defects from centimeter to micron spatial periods, whereas the atomic force microscope (AFM) is able to measure roughness from the micron down to the 0.1 nm scale. As waves are scattered most efficiently by asperities of a size comparable to their wavelength, scattering measurements are used in the X-ray region to probe defects of atomic scale. The prevalence of scattering explains why mirrors that reflect well in the visible range were disappointing when used in early X-ray experiments. The recent development of adaptive polishing methods has almost solved this problem.

The main effect of scattering by an optical system is not to decrease the number of photons coming out, but to modify their angular distribution. Scattering measurements are experimentally studied to test and evaluate the spatial defect distribution. In X-rays, one first characterizes the incident beam angular distribution in measuring its angular variation intensity after a scanning slit. Performing the same measurement after the optic, across the reflected beam in the case of mirrors, provides the modified angular distribution. From such a measurement, a few theoretical models adapted to the X-ray range can be used to deduce the mirror spatial defect distribution and so to check progress in polishing methods. The simplest model, coming from early crystal studies, is describing defects by a single parameter named roughness and denoted σ. Figure 5 shows an example of calculated reflectivity; on nickel surfaces versus photon energy and for increasing roughness; it shows how many photons have deviated from the symmetrical direction. Corresponding angular distributions can also be calculated.

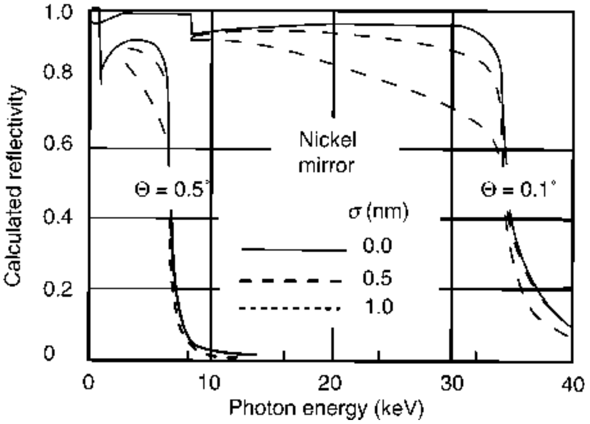

Figure 5 Calculated reflectivity variations versus photon energy, on nickel surfaces having different roughness, for different incidence angles.

The coating systems used for X-ray mirrors are material dependent. Deposition methods are similar to those used for visible optics coatings, but specific effort has been devoted to avoid nucleation growth to achieve uniform smooth layers. In general, one tries to give the evaporated particles, ions or clusters a kinetic energy lying within a specific range, which promotes mobility on the surface but is below the energy needed for implantation. In the case of electron beam deposition the energy distribution is thermal, but an auxiliary ionic bombardment can be used during deposition to provide the necessary energy to the previously deposited atomic layer (ion assisted deposition). For each material, the specific sputtering deposition and ionic bombardment conditions need to be carefully studied.

Until the 1970s, Bragg reflection by crystals and the total reflection phenomenon was the only way to reflect X-rays.

Multilayer X-Ray Mirrors

In the visible range, optics are frequently coated with stacks of dielectric multilayers obtained by periodic deposition of layers of two amorphous transparent materials. Such stacks are highly selective reflectors, or bandpass filters. They operate on the principle of constructive interference between all the waves partially reflected by the successive interfaces. They can be regarded as analogous to two-dimensional Bragg reflectors, in which the interfaces play the same role as atomic planes in crystals. Hence the Bragg equation [1] can be applied, the relevant period *d* being the thickness of an individual bilayer, if the stack is truly periodic.

The concept of using artificially deposited stacks of thin films, of alternating high and low electron density, to reflect and diffract X-rays dates back at least to the 1920s. It was not until 1944, however, that DuMond and Youtz were able to demonstrate significant X-ray reflection from such an artificially layered structure. Their work used alternating layers of copper and gold, laid down by evaporation. Unfortunately these materials diffuse into one another, reducing the electron density contrast and hence the X-ray reflectivity. In 1967 Dinklage and Frerichs succeeded in making a structure of alternating layers of iron and magnesium, which do not interdiffuse quickly, and thus produced the first stable X-ray reflecting multilayer.

Since then, enormous progress has been made in the fabrication and application of such multilayer reflectors, principally because of the efforts of two pioneers. Troy Barbee, while at Stanford University, developed open loop sputtering methods for controlling the thickness of the many thin (~nm) layers required to make up such a stack, while Eberhard Spiller at IBM developed an evaporation method using *in situ* X-ray monitoring to achieve similar thickness control. Almost all of the many multilayer deposition systems in use around the world are based on one of these two methods.

As for the total reflection mirror, only recent progress in methods for deposition and controls has made this new kind of X-ray optics practical. Present applications range from the extreme ultra violet (30 eV) region to hard X-rays having energies of a few tens of keV.

Prediction of the reflectivity variations can be simulated by calculations on the same web site mentioned in the preceding section. Multilayer mirrors are based on interference, so the reflectivity coefficient and the width of the reflected bandpass are determined by the amount of absorption inside the stack and the partial reflectivity at each interface. The more efficient pairs of materials have the lowest X-ray absorption and the highest refractive index difference at the required energy. The selection of real materials is constrained by the coating conditions and the need to achieve smooth interfaces. In particular, a low interdiffusion coefficient between the successive material pairs is a requirement for temporal stability. Epitaxial deposition methods have been considered but no useful pair of materials has yet been found.

An example of a flourishing application is multilayer-coated normal incidence X-ray optics. Optical schemes used in the visible region for telescopes and microscopes have excellent optical performance and high numerical aperture but a very low reflectivity in the X-ray region if coated with a single layer. However, if coated with multilayers they can be used, providing that extremely low-roughness surfaces can be produced. The interface qualities of the deposited thin layers presently limit normal incidence optics to the soft X-ray energy range, typically below 300 eV so around the carbon edge. This corresponds to periods of about 2 nm, or a few atomic layers, and the resulting roughness seriously decreases the reflectivity.

Figure 6 is a reflectivity curve versus energy for a quasinormal incidence Mo/Si multilayer mirror optimized to get the first-order peak at 93 nm. On Mo/Si mirrors, precise control of the layer period and the quality of the interfaces allow the manufacture of mirrors with reflectivity only a few percent below the theoretical limit shown here. Considerable effort has been invested in such mirrors, with a view to applications to EUV projection lithography.

It is possible to build polarizer mirrors over the soft X-ray range. An incidence angle of 45° corresponds to the Brewster angle, because the X-ray index is close to unity. X-ray multilayer and semitransparent selective beam splitters now permit the development of Michelson interferometers for newly developed soft X-ray lasers.

Figure 7 is an example of a tungsten/carbon (W/C) reflectivity curve for a multilayer adapted to the 8 keV energy of Cu Kα, mentioned before. The curve

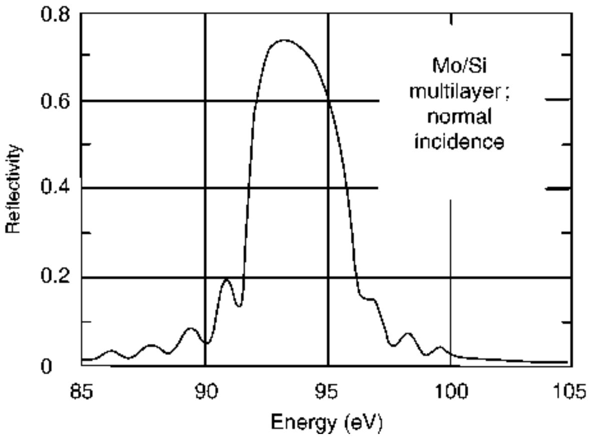

Figure 6 Calculated reflectivity versus incidence angle for a typical Mo/Si multilayer, optimized around 93 eV, under quasinormal incidence.

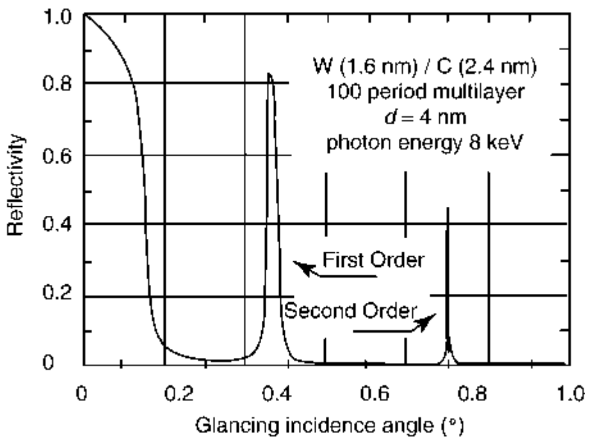

Figure 7 Calculated reflectivity versus incidence angle for a typical W/C multilayer adapted to reflect 8 keV photons.

of reflectivity versus glancing angle has been calculated from the total reflection range to the second-order peak. This kind of multilayer is being used in X-ray tubes to premonochromatize the beam and make it parallel. In such an application, the period of the coating is continuously graded laterally to satisfy locally the Bragg law at each point of the curved mirror.

As a final example of the new possibilities offered by X-ray multilayers, we mention the possibility of widening the reflected bandpass by varying continuously the period from the bottom to the top of the deposited stack. This is the equivalent of multilayer neutron 'supermirrors'. Such interference mirrors are usable at quite high X-ray energies, and give high reflectivity and a larger incidence angle than total reflection mirrors working over the same energy range.

The continuous fast development of coating technologies and X-ray multilayer applications can be followed through the proceedings of dedicated meetings at, for example, on a web site like http://PXRM.

Summary and Perspectives

After the discovery of X-rays in 1895, almost 25 years were necessary to understand their characteristics compared to the visible range. Only in 1923 did F. Holweck and R.A. Millikan demonstrate the continuity of electromagnetic radiation from X-rays to ultraviolet. The selective Bragg reflection by natural crystals is limited to energy commonly larger than about 2 keV photons. Below this energy the total reflection mirror was the only optical device to reflect and focus wavelengths shorter than the far UV until the development of adapted interferential mirrors.

During the 1960s, the EUV and X-UV ranges gained interest due to the spread of fusion research and in the astrophysics community studying the Sun and stars. For both communities, the continuum and the lines emitted at short wavelengths by ionized matter were a convenient tool to evaluate high plasma temperatures. At about the same time, the development of synchrotron radiation sources increased the need for X-ray optics. In addition, this last source offered the possibility to attempt new optical solutions with convenient intense and parallel beams from UV to X-rays. Finally, the concomitant progress in various microtechnologies afforded the necessary tools to produce new adapted optical devices for short wavelengths. At the present time, adapted optics are under development to take advantage of various applications from microscopy to interferometry over the very large electromagnetic energy X-ray range.

As a matter of fact, it is now possible to tailor a choice of mirror coatings adapted to each X-ray photon energy and application. Such coatings are specific because they must be optimized to take into account the inevitable absorption energy dependence.

Progress achieved in blank polishing and evaporation methods has led to efficient and high-quality X-ray mirrors even able to keep the spatial coherence of synchrotron sources. Such technological advances are also steps to master surfaces and interfaces at the atomic scale. They are now becoming applicable to the other spectral ranges and to various micro- and nanotechnologies.

List of Units and Nomenclature

Crystal period	d	[nm]	1 nm = 10^{-9} m
Incidence Angle	Θ	[degree]	
Photon Energy	E	[eV]	$(h/2\pi)\omega.\lambda = hc =$ 1239.842 eV.nm
Wavelength	λ	[nm]	1 nm = 10^{-9} m

See also

Diffraction: Fresnel Diffraction. **Geometrical Optics:** Lenses and Mirrors. **Incoherent Sources:** Synchrotrons. **Optical Coatings:** Thin-film Optical Coatings.

Further Reading

Attwood D (1999) *Soft X-Rays and Extreme Ultraviolet Radiation*. Cambridge: Cambridge University Press.

Ditchburn RW (1963) *Light*. Appendix XIX, 2nd edn. London: Blackie.
Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.
Henke BL, Gullikson EM and Davies JC (1993) *Atomic Data and Nuclear Data Tables* 54: 181.
James RW (1982) *The Optical Principles of the Diffraction of X-Rays*. Woodbridge, CT: Ox Bow Press.
Michette A and Pfauntsch S (1997) *X-Rays: The First Hundred Years*. New York: John Wiley.
Mould RF (1993) *A Century of X-Rays and Radioactivity in Medicine*. Bristol, UK: Institute of Physics Publishing.
Paratt LG (1954) Surfaces studies of solids by total reflection of x-rays. *Physical Review* 95: 359.
Spiller E (1972) Low-loss reflection coatings using absorbing materials. *Applied Physics Letters* 20(9): 365–367.
Underwood JH and Attwood DT (1984) The renaissance of x-rays optics. *Physics Today, April*: 44–51.

OPTICAL COMMUNICATION SYSTEMS

Contents

Basic Concepts
Historical Development
Architectures of Optical Fiber Communication Systems
Free Space Optical Communications
Lightwave Transmitters
Local Area Networks
Optical Time Division Multiplexing
Wavelength Division Multiplexing

Basic Concepts

S Lee and A E Willner, University of Southern California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since modern lightwave communication started in the 1970s, extensive efforts have been made to increase data rate and transmission distance. Recently, the bit rates for single-channel and multichannel systems have exceeded 40 and 2 Tb/s on single fibers respectively, for transoceanic distances. One revolutionary development in lightwave systems has been the deployment of erbium doped fiber amplifiers (EDFAs). These amplifiers facilitate transmission of an optical signal over long distances, by providing periodic analog-like amplification rather than digital-like regeneration. The wide bandwidth provided by EDFAs has made possible the increased use of wavelength division multiplexing (WDM) in which multiple lightwave signals, each having a different wavelength, are co-propagated on a single fiber. EDFAs and WDM techniques can enhance the lightwave system capacity, both in terms of obtainable transmission distance, and total number of data rates (see Figure 1).

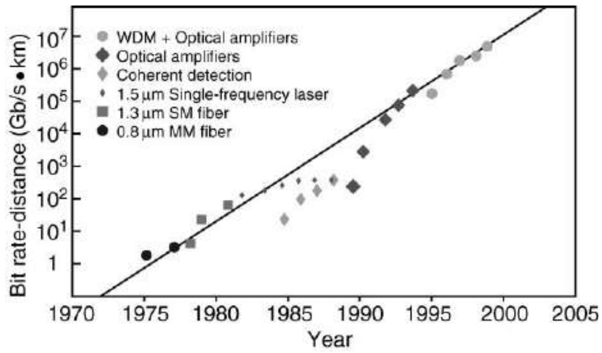

Figure 1 Bit rate product. Data from T. Li (private communication).

Optical Fiber

Most telecommunication fibers are made of silica, therefore following the attenuation of silica material. Figure 2 shows the silica fiber attenuation as a function of wavelength. The downward slope towards 1.6 μm and upward slope away from 1.6 μm, are the theoretical limits due to Rayleigh scattering and absorption of silica material, respectively. The absorption peak near 1.4 μm is due to water molecule absorption. There are two low attenuation regions: ~1.3 μm and 1.55 μm, both of which are used for communications in general, and some other wavelengths are also used for shorter distance communications. Figure 2 also shows the comparison between gain bandwidth of EDFA (~3 THz) and the low attenuation window around 1.55 μm (~25 THz). The typical values of attenuation at 1.3 μm and 1.55 μm for single-mode fiber (SMF) are ~0.35 dB/km, and 0.2 dB/km, respectively.

When an electromagnetic wave interacts with the bound electrons of a silica fiber, the medium response depends upon optical frequency (ω). This property, referred to as material dispersion, manifests itself through the frequency dependence of the refractive index $n(\omega)$. Fiber material dispersion plays a critical role in propagation of short optical pulses since different spectral components associated with the pulse travel at different speeds, given by $c/n(\omega)$. Consequently, the optical pulse at the output of the fiber is broadened in time. The commonly used system parameter is called the dispersion parameter D (the negative D value is called normal dispersion, and the positive D value is called anomalous dispersion), and is measured in ps/(nm·km); D for SMF is ~+ 17 ps/(nm · km) at 1.5 μm. However, because of dielectric waveguiding, the effective mode index is slightly lower than the material index $n(\omega)$, with the reduction itself being ω-dependent (waveguide dispersion). This results in a waveguide contribution that must be added to the material contribution to obtain the total dispersion (see Figure 3).

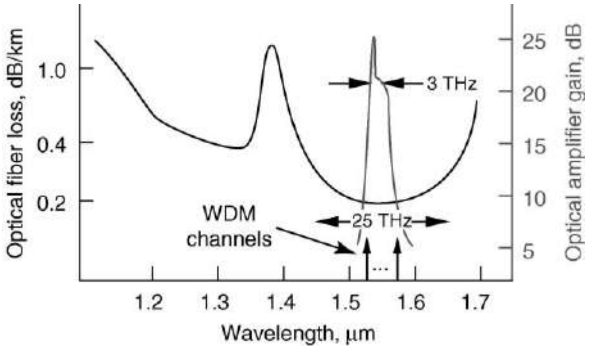

Figure 2 Silica fiber attenuation versus wavelength.

Generally, the waveguide contribution to dispersion D is to shift zero dispersion wavelength (λ_0) to longer wavelengths. Furthermore, the waveguiding can be tailored to shift the λ_0 from 1.3 μm to 1.5 μm, the wavelength at which the fiber has the minimum attenuation loss. The fiber having λ_0 in the neighborhood of 1.5 μm, is called dispersion-shifted fiber (DSF); D for DSF is usually between −2.5 and +2.5 ps/nm·km, at a wavelength of 1.5 μm. The typical dispersion parameter, D, as a function of a wavelength for both SMF and DSF, is shown in Figure 4.

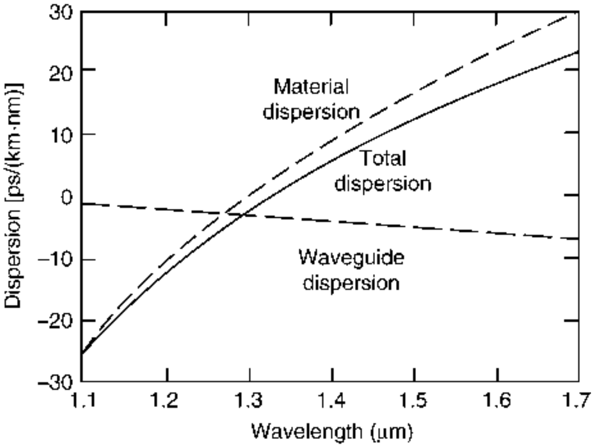

Figure 3 Total dispersion (D) and contribution of material dispersion and waveguide dispersion in conventional SMF.

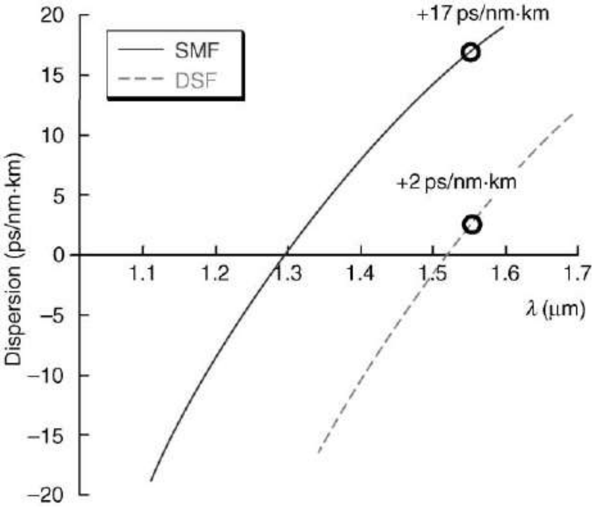

Figure 4 Dispersion versus wavelength for conventional SMF and DSF.

In addition to chromatic dispersion, optical fiber has another dispersion properties related to polarization. An optical wave of arbitrary polarization can be represented as the superposition of two orthogonally polarized modes. In an ideal fiber, these two modes are indistinguishable, and have the same propagation constants owing to the cylindrical symmetry of the waveguide. However, in real fibers there is some residual anisotropy due to unintentional circular asymmetry, usually caused by noncircular waveguide geometry or asymmetrical stress around the core, as shown in Figure 5.

In either case, the loss of circular symmetry gives rise to two distinct orthogonally-polarized modes with different propagation constants (differential phase velocity) responsible for polarization mode dispersion (PMD) in the fiber, and can be related to the difference in refractive indices (birefringence) between the two orthogonal polarization axes. The differential phase velocity indicated is accompanied by a difference in the group velocities for the two polarization modes, therefore the pulse at the end of the transmission fiber is broadened by this differential group delay (DGD). DGD is usually expressed in units of ps/km for a short length (1 m to 1 km) of birefringent fiber. However, DGD does not accumulate along a long fiber link in a linear fashion. Instead, because of random variations in the perturbations along a fiber span, the DGD in one section may either add to or subtract from another section of the fiber. As a result, average DGD in long fiber spans accumulates in a random-walk-like process that leads to a square root of transmission-length dependence. Therefore, average DGD is expressed in ps/km$^{1/2}$ in long fiber spans, referred to as the PMD of the fiber, and the typical PMD parameter has a value of 0.01 to 10 ps/km$^{1/2}$. Because of the many perturbations that act on a real-world fiber (e.g., temperature, vibration, etc.), transmission properties typically vary with time. Therefore, the PMD of the fiber link fluctuates randomly, thus causing random fluctuations in system performance.

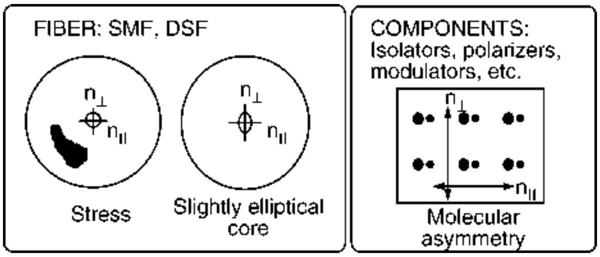

Figure 5 Origin of PMD.

It is well known that optical fibers show nonlinear behavior under conditions of high power and long interaction length. The power-times-distance products for amplified transmission systems can be large enough to make fiber nonlinear effects the dominant factor in the design of long-distance systems.

There are two categories of fundamental optical nonlinear effects: stimulated scattering effects and refractive index effects. Stimulated scattering effects arise from parametric interactions between light and acoustic or optical phonons in the fiber. Two nonlinear effects fall into this category: stimulated Brillouin scattering (SBS) and stimulated Raman scattering (SRS). The main difference between the two is that optical phonons participate in SRS, while acoustic phonons participate in SBS. In a simple quantum-mechanical picture, a photon of the incident field is annihilated to create a photon at a longer wavelength. The new photon is co-propagated and counter-propagated along the original signal in SRS, while counter-propagated in SBS. Refractive index effects are caused by modulation due to changes in light intensity. There are three types of refractive index effects: self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). SPM introduces the change of optical phase by its own intensity, and, this leads to frequency chirp of the pulse, depending on the relative position from the peak. When this nonlinear frequency chirp interacts with the fiber dispersion, the pulse either broadens or compresses, depending on the sign and the amount of dispersion along the fiber. In XPM, the phase of the signal in one wavelength channel is modulated by the intensity fluctuation of the other wavelength channels. In a WDM system, XPM imposes far more damaging effects than SPM because XPM is stronger by a factor of two than SPM when the channel power is the same, and large number of WDM channels can contribute to XPM. FWM is the generation of modulation sideband at new frequencies, due to the phase modulation of channels between lights at different frequencies in multichannel system. FWM causes penalties in a WDM system if the newly generated frequency is either equal to or close to the frequency of existing WDM channels.

Fiber nonlinearities impose significant degradation in optical transmission system, and limit the system on allowable number of channels, channel power, and channel spacing (see Figure 6). But, the effects can be well suppressed by carefully designing the fiber dispersion profile (i.e., dispersion map) in the transmission system, and controlling the optical launch power for each fiber span.

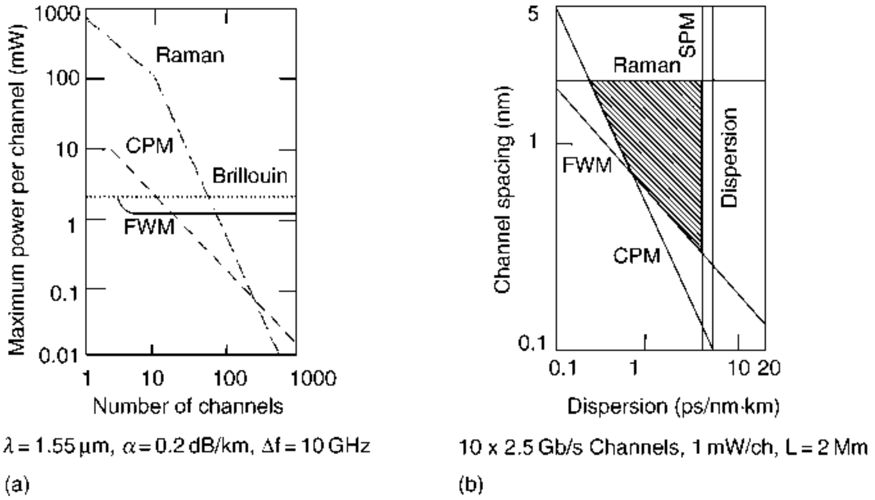

Figure 6 Limitation due to fiber nonlinearities in (a) maximum channel power and number of channels, and (b) channel spacing and dispersion. Reproduced with permission from Chraplyvy AR (1990) Limitation on lightwave communications imposed by optical-fiber nonlinearities. *IEEE Journal of Lightwave Technology* 8:1548–1557. Copyright © 1990 IEEE.

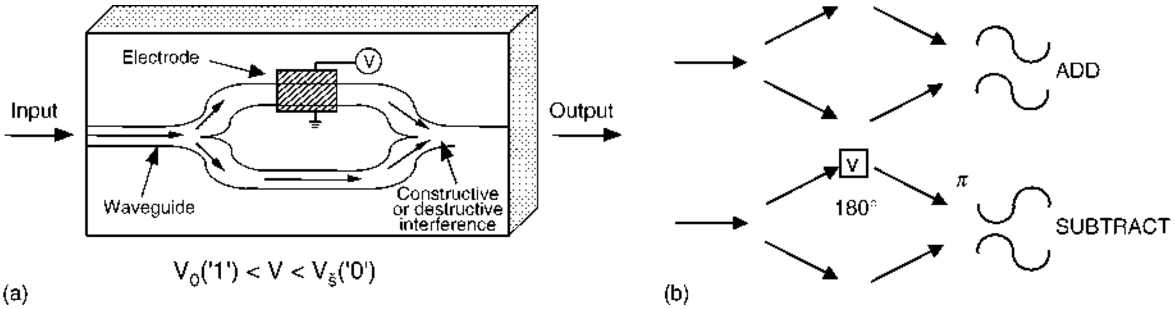

Figure 7 (a) EOM: MZ interferometer on LiNbO2, (b) operation of EOM; constructive and destructive interference resulting in intensity modulation.

Communication Components

Distributed feedback laser (DFB) is widely used as a light source for metro, long-haul, and undersea applications, due to its narrow spectral width, and wavelength stability. Fabry–Perot (FP) lasers, and vertical cavity surface emitting lasers (VCSEL) are used for local area networks (LAN), and for access applications, because these are cheaper than DFB lasers. For an application, where the cost is the most important parameter with low data modulation speed, light-emitting diode (LED) is used as a light source.

In optical data modulation, the simplest and easiest technique is turning ON or OFF the laser, depending on the binary logic '1' or '0' (direct modulation). This type of modulation is simple and cheap compared to external modulation technique. But, its application is limited by the bandwidth of modulation, and the induced frequency chirp (i.e., difference in optical frequency at the turn ON state and just before the turn OFF state of laser) that limits the transmission distance. Two different types of external modulators are widely used for digital optical data transmission. Electro absorption modulator (EAM) uses the Franz–Keldysh effect, where it is observed that the wavelength of the optical absorption edge in a semiconductor is lengthened by applying an electric field. Electro-optic modulator (EOM) utilizes the linear electro-optic effect, called the Pokels effect, which changes the refractive index of the material caused by and proportional to an applied electric field, so therefore changes the phase of the optical signal. This phase change is used to modulate the intensity of a lightwave through a Mach–Zehnder (MZ) interferometer (see Figure 7). Because the Pockels effect exists only in crystal, lithium niobate crystal ($LiNbO_2$) or electro-optic polymers are used for EOM.

EDFA is a wideband optical amplifier that has merits in that: (i) erbium ions (Er^{3+}) emit light in the 1.55 μm loss-minimum band of optical fiber, (ii) a circular fiber-based amplifier is inherently compatible with a fiber optics system; (iii) it provides

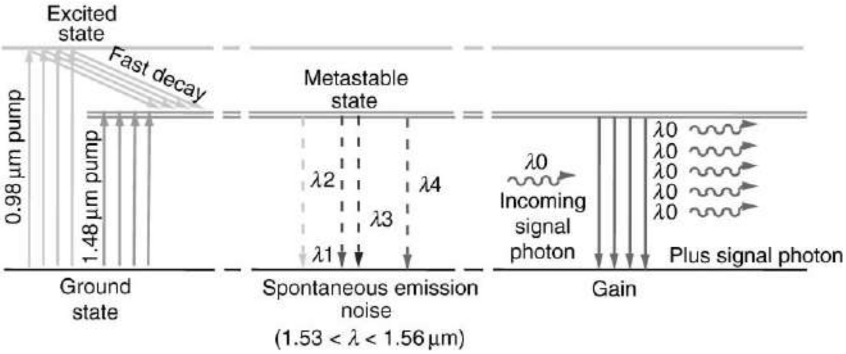

Figure 8 Energy level and transition diagram of EDFA. Reproduced with permission from Willner AE (1997) Mining the optical bandwidth for a terabit per second. *IEEE Spectrum* Apr.: 32–41. Copyright © 1997 IEEE.

amplification independent of bit-rate, modulation format, power, and wavelength; and (iv) it has low distortion and low noise during amplification. EDFA contains a gain medium (i.e., erbium-doped fiber) that must be inverted by a pump source. A signal initiates stimulated emission, resulting in gain, and spontaneous emission occurs naturally, which results in noise.

Figure 8 shows the energy level and transition diagram of EDFA. A 0.98 µm pump photon is absorbed and excites a carrier (Er^{3+} ions) into higher excited states, and the excited carrier decays rapidly to the first excited state that has a very long lifetime of ~10 ms (metastable state). In contrast, a 1.48 µm pump photon is absorbed and excites a carrier into a metastable state. This long metastable state is one of the key advantages of EDFA, and intersymbol distortion and interchannel crosstalk are negligible as a result of these slow dynamics. Depending on the external optical excitation signal, this carrier will decay in a stimulated or spontaneous fashion to the ground state and emit a photon. Both absorption and emission spectra have an associated bandwidth depending on the spread in wavelengths that can be absorbed or emitted from a given energy level. This is highly desirable because (i) a pump laser does not need to be an exact wavelength; and (ii) the signal may be at one of several wavelengths, especially in a WDM system. Figure 9 shows the bandwidth in the 1.48 µm absorption and 1.55 µm fluorescence spectrum of a typical erbium-doped amplifier.

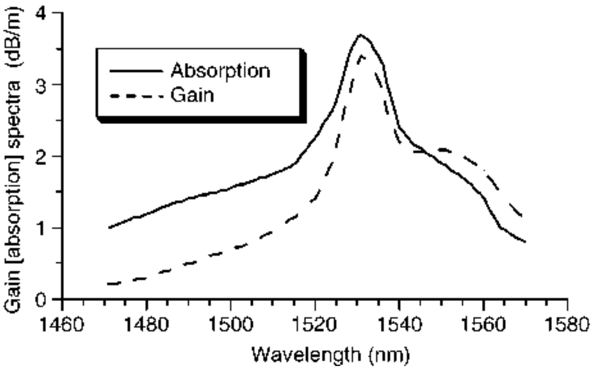

Figure 9 The absorption and fluorescence spectra for erbium near 1.5 µm. Reproduced with permission from Miniscalco WJ (1991) Erbium-doped glasses for fiber amplifiers at 1550 nm. *IEEE Journal of Lightwave Technology* 9: 234–250. Copyright © 1991 IEEE.

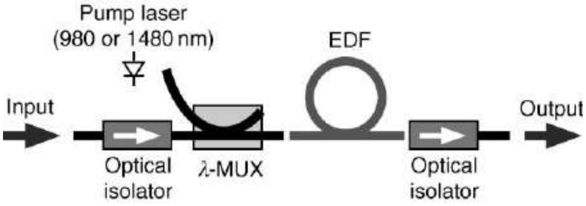

Figure 10 Typical EDFA diagram.

When the noise characteristics of EDFA are considered, noise figure (NF) is an important parameter, which is defined by $NF \equiv SNR_{in}/SNR_{out}$, where SNR_{in} (SNR_{out}) is the electrical equivalent signal-to-noise ratio (SNR) of the optical wave going into (coming out) of the amplifier, if it were be detected. The typical value of NF in commercially-available EDFA is 4.5–6 dB. The importance of NF in optical communication systems is presented later in this chapter, when the required optical SNR (OSNR) for error-free optical data transmission is considered. Figure 10 shows an example of EDFA diagram.

Recent development of EDFA in longer wavelength regimes or L-band (1570–1620 nm) has doubled the useful transmission bandwidth over the conventional band or C-band (1525–1265 nm). L-band gain of EDFA is generally achieved by proper arrangement of pump lasers through longer lengths of erbium-doped fiber than conventional C-band EDFA. Figure 11a shows the concept of wideband EDFA. The incoming

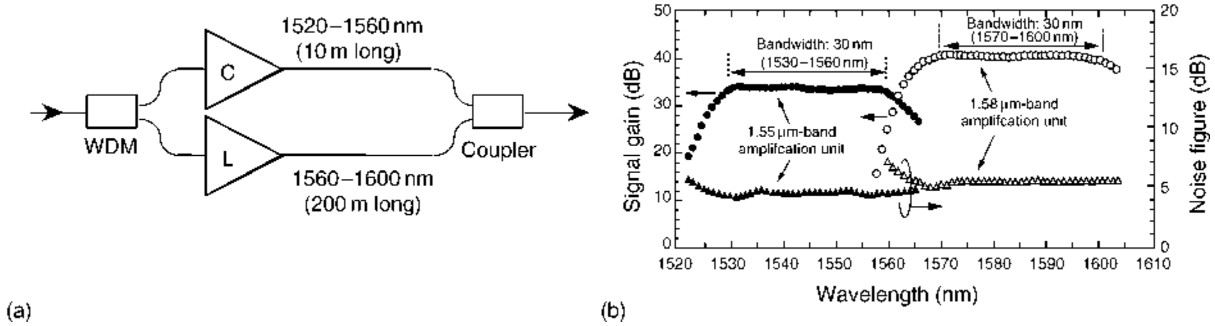

Figure 11 Wide band EDFA. Reproduced with permission from Yanada M (1997) *Electronics Letters*. Copyright © 1997 IEEE.

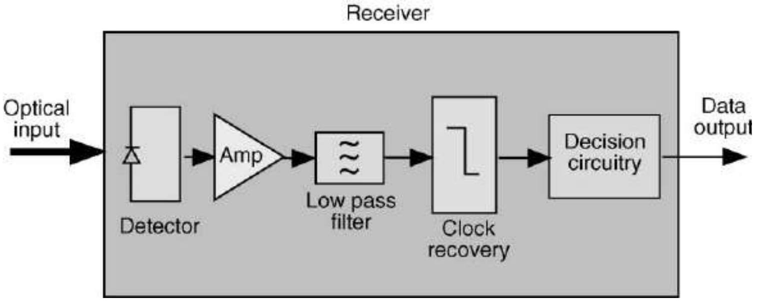

Figure 12 Block diagram of functions performed in advanced receiver packages.

WDM signal is split into two bands (C- and L-bands), amplified separately, and combined at the output. Gain and NF characteristics of this wideband EDFA is shown at Figure 11b.

Because SRS transfers the energy of a lower wavelength channel into the higher wavelength channel in the optical fiber, the transmission fiber itself can be used as a gain medium on specific channel (or channels) when the proper pump signal (or signals) are provided in the fiber. Even though Raman gain is much lower than the gain of EDFA (i.e., <0.1 dB gain/mW pump in Raman amplifier (RA) compared to a few dB gain/mW pump in EDFA), RA has several advantages over EDFA (i) RA has the capability to provide gain at any signal wavelengths, (ii) the amplification window is expandable by combining multipump wavelengths; and (iii) RA offers improved noise performance because the signal is amplified over transmission fiber (i.e., distributed amplifier). Especially improved noise performance increases system OSNR, which can be used to extend system reach or create more wavelength channels on the system. The Raman pumps are typically located ~100 nm shorter than the wavelength of the signal to be amplified, and a few 100s mW pump lasers are used to get a 10–15 dB gain.

In general, telecom receiver (RCVR) consists of photo diode (PD) with transimpedence amplifier (TIA), followed by limiting amplifier (LA) or automatic gain control amplifier (AGC), low pass filter, and clock data recovery circuit (CDR) (see Figure 12).

Vertically illuminated positive-intrinsic-negative (PIN) photodiode and avalanche photo diode (APD) are the most common PD that is used in RCVR. In PIN, incoming light is absorbed in the n^- (low doped or intrinsic) region and generates an electron-hole pair, which drifts to a depletion region and collected n^+ and p^+ regions, therefore generating electric current. APD is used for high-sensitivity detection up to 10 Gb/s. It is structured as absorption, grading, and multiplication layers. The multiplication layer of APD provides additional transit time delay compared to PIN, therefore the bandwidth of APD is smaller than PIN in general. TIA is a broadband low noise amplifier that is used to convert and amplify the weak current from PD to the desired output voltage. A low pass filter is used to suppress the noise and data distortion from high frequency components with a typical cut-off frequency ~0.7 bitrate. LA and AGC are amplifiers that limit the output voltage as a constant. LA is a nonlinear amplifier that makes a decision either to make logic ‘1’ or ‘0’ based on input voltage level, and limits output voltage as fixed values. AGC is a linear amplifier that limits the output voltage by adjusting the gain of the amplifier based on input voltage level. CDR consists of a phase lock loop (PLL) with a voltage controlled oscillator (VCO)

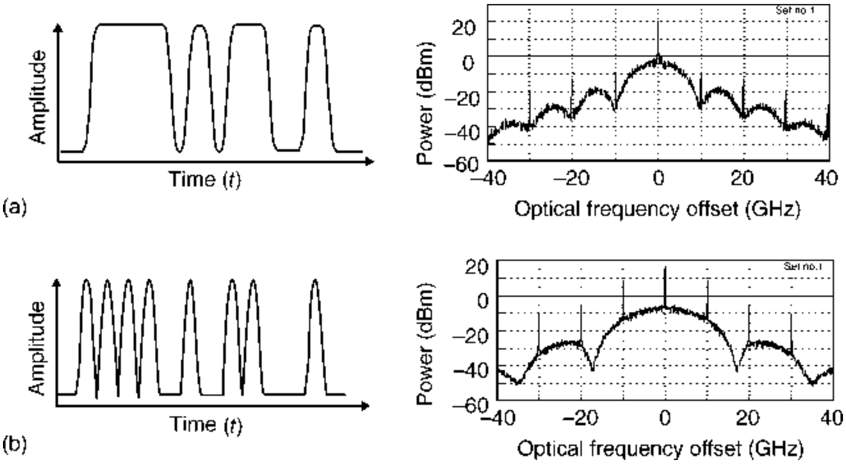

Figure 13 Amplitude modulation (a) nonreturn-to-zero (NRZ), and (b) return-to-zero (RZ).

and data flip-flop circuit. The PLL is used to synchronize the VCO to the incoming high-speed bit stream, therefore recovering the system clock. The recovered clock is used in D-FF to retime and regenerate the data.

Data Modulation Formats

In digital optical communications, numerous modulation formats have been developed and utilized, in order to achieve higher spectral efficiency (i.e., ratio of individual channel bit rate to dense WDM (DWDM) separation), and to improve system performance over fiber impairments. A simple amplitude modulation of lightwave is still the most widely used method for optical communication. The easiest amplitude modulation technique is to using a nonreturn-to-zero (NRZ) format. Figure 13a shows an example of NRZ binary sequence in time domain, and the corresponding frequency domain spectrum. In NRZ, data '1' and '0' correspond to the ON and OFF state of a lightwave transmitter, respectively, occupying entire bit time (i.e., binary symbol time). When the ON or OFF state of a lightwave transmitter does not occupy the entire bit time (T_B) (i.e., when the data pulse duration (T_d) is less then (T_B), it is called the return-to-zero (RZ). Figure 13b shows a binary sequence of RZ, and its frequency spectrum when $T_d \approx 0.5\ T_B$. The Fourier transform of the square wave is the Sinc function. Therefore, the frequency spectrum of an ideal NRZ and RZ follows the pattern of a Sinc square function. Note that the first null of frequency spectrum of RZ is about twice as high as the null of NRZ.

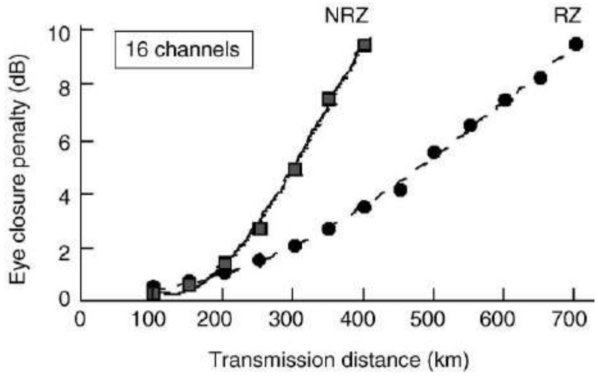

Figure 14 System performance of different modulation format (NRZ vs. RZ).

Even though NRZ is better in spectral efficiency than RZ, RZ has an advantage over NRZ in tolerance of fiber impairments. Figure 14 shows an example on eye closure penalty comparison between NRZ and RZ data formats at a 16-channel WDM system at 40 Gb/s data rate. It is clearly seen that the penalty of RZ is much less severe than the penalty of NRZ, as the transmission distance increases. (Note, nonlinearities and dispersion increases as the transmission distance increases.)

There has been much effort to increase the spectral efficiency, to improve the system performance over fiber impairments, or to achieve both simultaneously. Figure 15 shows some of examples of these. Chirped RZ (CRZ) is the modulation format inducing frequency chirp on conventional RZ. CRZ is robust on fiber impairment at the cost of losing spectral efficiency. In carrier-suppressed RZ (CS-RZ), a strong un-modulated optical carrier is removed from conventional RZ, therefore suppressing the nonlinear effects. Duo-binary is a three-level signal modulation format generated by a series of delay and added

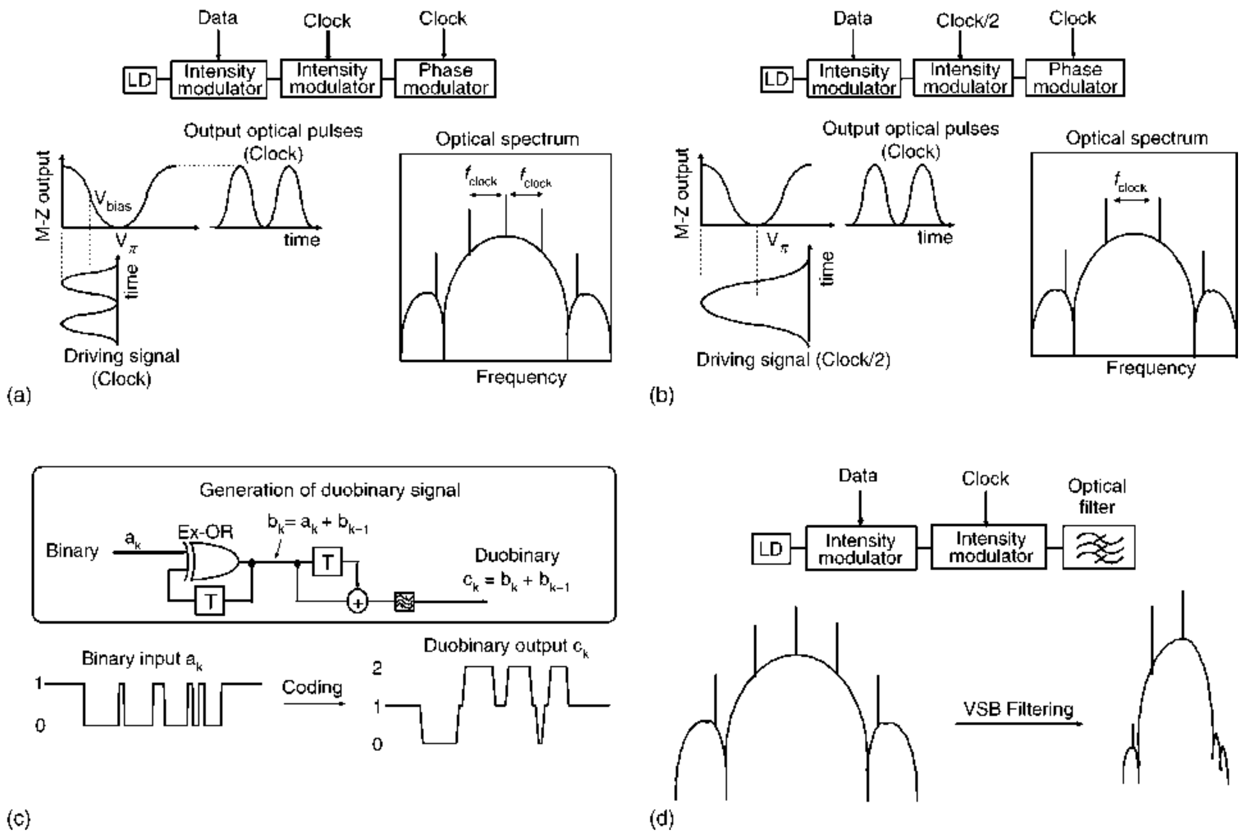
Data
Clock
Clock
LD
Intensity modulator
Intensity modulator
Phase modulator
M-Z output
Output optical pulses (Clock)
V_{bias}
V_π
time
time
Driving signal (Clock)
Optical spectrum
f_{clock}
f_{clock}
Frequency
(a)
Data
Clock/2
Clock
LD
Intensity modulator
Intensity modulator
Phase modulator
M-Z output
Output optical pulses (Clock)
V_π
time
time
Driving signal (Clock/2)
Optical spectrum
f_{clock}
Frequency
(b)
Generation of duobinary signal
Binary
a_k
Ex-OR
$b_k = a_k + b_{k-1}$
T
T
+
Duobinary
$c_k = b_k + b_{k-1}$
Binary input a_k
1
0
Coding
Duobinary output c_k
2
1
0
(c)
Data
Clock
Optical filter
LD
Intensity modulator
Intensity modulator
VSB Filtering
(d)

Figure 15 Examples of bandwidth efficient modulation formats. (a) Chirped RZ (CRZ) generation, (b) carrier suppressed-RZ (CS-RZ) generation, (c) duo-binary generation, and (d) vestigial single side band (VSB) RZ generation.

circuits in the transmitter. The advantage of duo-binary is that its bandwidth is reduced to one half of conventional NRZ, therefore having enhanced spectral efficiency, and has a high tolerance in chromatic dispersion, and suppression of fiber nonlinearites.

Single-sideband (SSB) or vestigial singleside band (VSB) can be generated into either NRZ or RZ, by either optically utilizing optical filtering with a sharp cut-off filter, or electrically applying tapped-delay-line filter approximation on both amplitude and phase modulators. SSB (or VSB) has about a two times higher spectral efficiency and higher tolerance on chromatic dispersion than conventional modulation formats.

Data Multiplexing

At their most basic, optical networks can imitate electrical networks in which time division multiplexing (TDM) is overwhelmingly used for digital data transmission. A fiber can carry many time-multiplexed channels, in which each channel can transmit its data in an assigned time slot. A typical TDM link is shown in Figure 16, in which N transmitters are sequentially polled by a fast multiplexer to transmit their data. The time-multiplexed data are sequentially and rapidly demultiplexed at the receiving node.

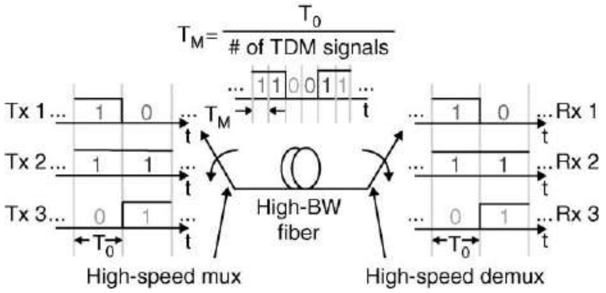
$T_M = \frac{T_0}{\text{\# of TDM signals}}$
... 1 1 0 0 1 1 ...
Tx 1... 1 0 ...
Tx 2... 1 1 ...
Tx 3... 0 1 ...
T_0
t
T_M
High-BW fiber
High-speed mux
High-speed demux
... 1 0 ...Rx 1
... 1 1 ...Rx 2
... 0 1 ...Rx 3
T_0

Figure 16 Concept of bit-interleaving TDM.

Time multiplexing and demultiplexing functions can be performed either electrically with an electrical time (DE)MUX switch and an optical transmitter/receiver, or optically with multiple optical transmitters/receivers and an optical time (DE)MUX switch. The major advantage of TDM is that there is no output–port contention problem (each data bit occupies its own time slot and there is only a single high-speed signal present at any given instant). The disadvantage of TDM is that the scheme requires a ultra-high-speed switching component if the individual signals are themselves high-speed and if there are many users.

In WDM, multiple wavelength channels are transmitted through a single fiber, therefore enabling the fiber to carry more throughput. By using wavelength-selective devices for the ON and OFF ramps, independent signal routing also can be accomplished. Figure 17a shows a simple point-to-point WDM system in which several channels are multiplexed at one node, the combined signals transmitted across a distance of fiber, and the channels demultiplexed at a destination node. As shown in Figure 17b, the wavelength becomes the signature address for either the transmitter or the receivers, and the wavelength will determine the routing path through an optical network.

Many interesting challenges face the eventual implementation of WDM systems and networks. Several of these challenges involve the control and management of the data through this novel high-speed network. Figure 18 shows a small subset of critical component technologies typically required for a WDM system, including multiple-wavelength transmitters, multiport star couplers, passive and active wavelength routers, EDFAs, and tunable optical filters.

Some generic goals that a WDM-device technologist aims to achieve include; large wavelength tuning range; multi-user capability; wavelength selectivity and repeatability; low cross-talk; high extinction ratio; minimum excess losses; fast wavelength tunability; high-speed modulation bandwidth; low residual chirp; high finesse; low noise; robustness; high yield; potential low cost. Depending on system requirements, device availability and cost, WDM technologies divide into two; course-WDM (CWDM) and dense-WDM DWDM. Figure 19 shows the wavelength allocation for CWDM and

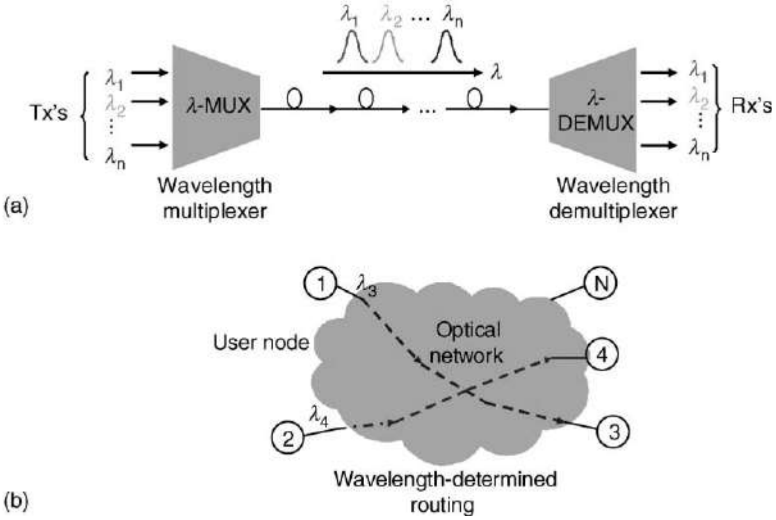
λ_1 λ_2 ... λ_n
λ
Tx's
λ_1
λ_2
λ_n
λ-MUX
λ-DEMUX
Rx's
Wavelength multiplexer
Wavelength demultiplexer
(a)
1
λ_3
N
User node
Optical network
4
λ_4
2
3
Wavelength-determined routing
(b)

Figure 17 (a) A simple point-to-point WDM transmission system; (b) a generic multiuser network in which the communication links and routing paths are determined by the wavelengths used within the optical switching fabric.

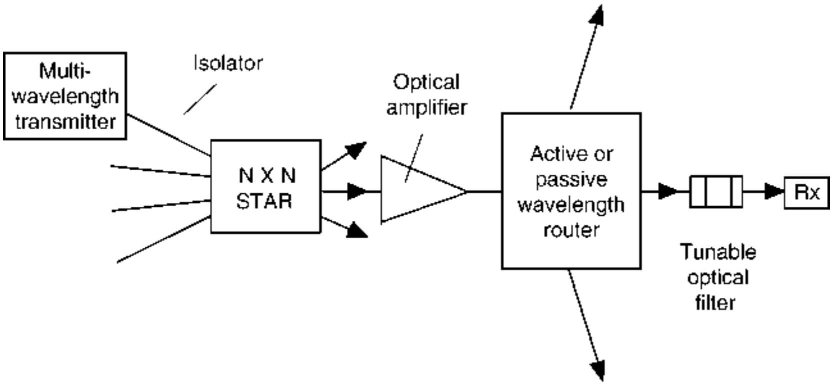
Multi-wavelength transmitter
Isolator
Optical amplifier
N X N STAR
Active or passive wavelength router
Rx
Tunable optical filter

Figure 18 Schematic of a small subset of enabling device technologies for a WDM system.

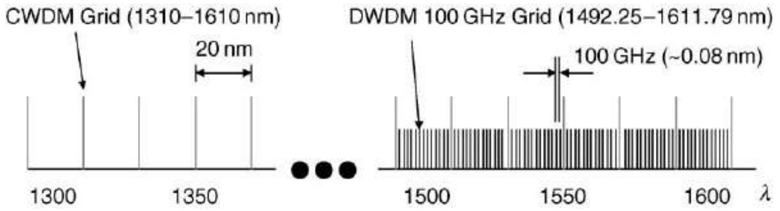
CWDM Grid (1310–1610 nm)
20 nm
DWDM 100 GHz Grid (1492.25–1611.79 nm)
100 GHz (~0.08 nm)
1300
1350
1500
1550
1600
λ

Figure 19 WDM wavelength allocation.

DWDM. CWDM technology uses an International Telecommunication Union (ITU) standard 20 nm spacing between the wavelengths from 1310 nm to 1610 nm. Also DWDM technology uses an ITU standard 100 GHz or 200 GHz between wavelengths, arranged in several bands at around 1500–1600 nm.

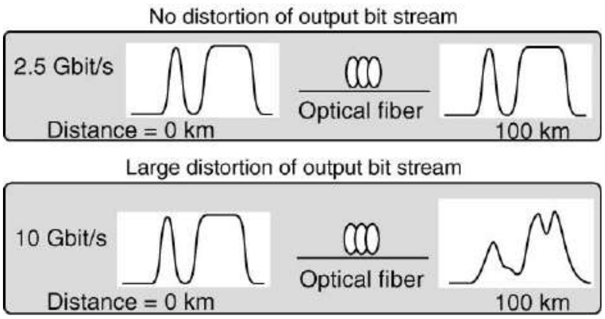
No distortion of output bit stream
2.5 Gbit/s
Distance = 0 km
Optical fiber
100 km
Large distortion of output bit stream
10 Gbit/s
Distance = 0 km
Optical fiber
100 km

Figure 20 Pulse broadening at two different data rates (2.5 Gbit/s, and 10 Gbit/s) as a result of quadratic nature of chromatic dispersion.

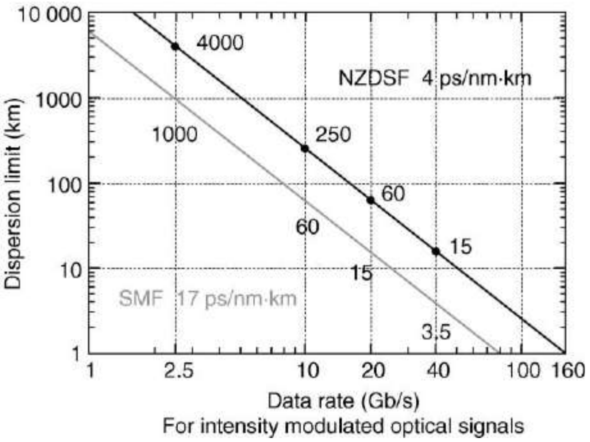

Figure 21 Dispersion-limited (uncompensated) transmission distance in single-mode fiber (SMF) as a function of data rate for intensity-modulated optical signals. Courtesy of L.D. Garrett.

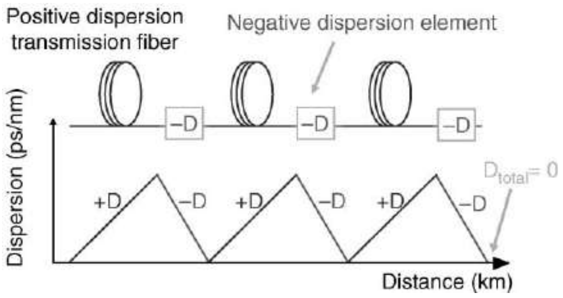
Positive dispersion transmission fiber
Negative dispersion element
–D
–D
–D
Dispersion (ps/nm)
+D
–D
+D
–D
+D
–D
$D_{total} = 0$
Distance (km)

Figure 22 Typical static management of chromatic dispersion.

Dispersion Management and Compensation

As mentioned previously, chromatic dispersion is one of the most basic characteristics of fiber, although it is possible to manufacture fiber that induces zero chromatic dispersion. But such fiber is incompatible with the deployment of a WDM system since harmful nonlinear effects are generated, therefore chromatic dispersion must exist, and must be compensated for. The effect of chromatic dispersion is cumulative and increases quadratically with the data rate (see Figure 20).

The quadratic dependence of dispersion with the data rate is a result of two effects. First, a doubling of the data rate will double the Fourier-transformed frequency spectrum of the signal, thereby doubling the effect of dispersion. Second, the same doubling of the data rate makes the data pulse only half as long in time and therefore twice as sensitive to temporal spreading due to dispersion. The conventional wisdom for the maximum distance over which data can be transmitted is to consider a broadening of the pulse equal to the bit time period. For a bit period B, a dispersion value D and a spectral width $\Delta\lambda$, the dispersion-limited distance is given by $L_D = 1/(D{\cdot}B{\cdot}\Delta\lambda)$ (see Figure 21).

In theory, compensation of chromatic dispersion for high-speed or long-distance systems can be fixed in value if each link's dispersion value is known. A simple yet elegant solution is to create a dispersion 'map', in which the designer of a transmission link alternates elements that produce positive and then negative dispersion (see Figure 22). This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, effectively

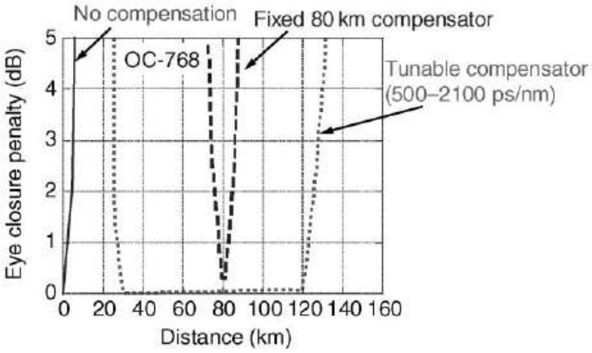

Figure 23 Effect of tunable compensation at 40 Gb/s (OC-768); it allows a wide range of transmission distance. On the contrary, the 80 km fixed compensator allows only small range of transmission around 80 km.

eliminating FWM and XPM, but the total accumulated dispersion at the end of the fiber link is zero, so that minimal pulse broadening is induced. The specific system design, as to the periodicity of management, depends on several variables, but a typical number for SMF as the embedded base is a compensation at every 80 km in a 10 Gb/s system. Dispersion-compensating fiber (DCF) has been generally used as a dispersion compensating element.

However, there are several important aspects of optical systems and networks that make tunable dispersion compensation solutions attractive, including: (i) it significantly reduces the inventory of different required types of compensation modules; (ii) it tunes to adapt to routing path changes in a reconfigurable network, (iii) it tracks dynamic changes in dispersion due to environment, and (iv) it achieves a high degree of accuracy necessary for 40 Gb/s channels (see Figure 23).

One brute-force method to achieve tunable dispersion compensation is to build a module with optical switches used to add or remove sections of fixed DCF to achieve a discrete set of dispersion compensation. Many other elegant, yet viable, approaches have been developed for tunable dispersion compensation, and these approaches include linearly chirped FBG with nonuniform heating, nonlinearly chirped Fiber Bragg grating (FBG) with a simple mechanical stretcher, virtually imaged phase array, electronic tap delay filter with weights, multiple stage all pass filter, etc.

System Performance Parameters

An eye diagram provides a simple and useful tool to visualize intersymbol interference between data bits. Figure 24a shows a perfect eye diagram. A square bit stream (i.e., series of symbol ‘1’s and ‘0’s) is sliced into sub-bit stream with predetermined eye intervals (i.e., several bit periods), and displayed through bit analyzing equipment (e.g., digital channel analyzer), overlapping the sliced sub-bit stream in order to obtain the eye diagram. When a perfect transmitter and receiver (i.e., infinite receiver bandwidth with zero noise characteristics and jitter), and a perfect transmission media (i.e., no dispersion and nonlinearites) are used, the received eye diagram is shaped as a perfect rectangular. In reality, the transmitter and receiver have a limited bandwidth with noise and jitter, and the transmission media (i.e., optical fiber) has dispersion and nonlinearites. Therefore, the eye diagram deviates from the perfect rectangular shape. Figure 24b shows the eye diagram close to a real situation. The shape of the eye is generally broadened and distorted (i.e., eye is closed) due to limited bandwidth and fiber impairments, and noise and timing jitter are added onto this broadened and distorted eye shape.

The Q-factor (q) is also an important system parameter widely used in long-distance optical transmission system design. It is defined as the electrical signal-to-noise ratio before the decision circuit at receiver. The Q-factor is directly related to bit-error-rate (BER: the percentage of bits that has an error relative to total number of bits received in a specific time) by: $\mathrm{BER} = 0.5 \cdot \mathrm{erfc}(q/2^{0.5})$, where $\mathrm{erfc}(x)$ is the complementary error function. The Q-factor is an unitless linear ratio, and is express in dB by $20 \cdot \log(q)$. As a conventional relationship, Q-factor of 15.6 dB (i.e., linear ratio 6) is required to achieve $\mathrm{BER} = 10^{-9}$. Figure 25 shows the relationship of eye diagram, Q-factor, and BER.

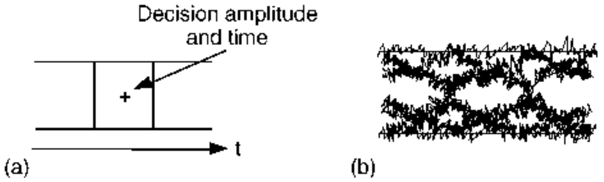

Figure 24 (a) A perfect square eye diagram, (b) a closed eye diagram due to bandwidth, fiber impairments, noise, and timing jitter.

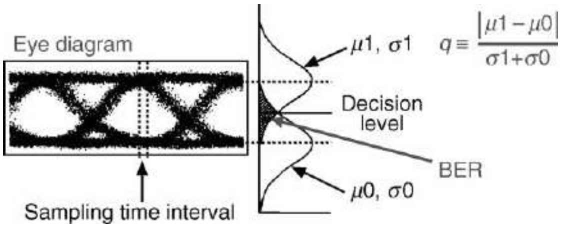

Figure 25 The relationship of eye diagram, Q-factor, and BER ($\mu 1$, $\mu 2$: mean levels of logic ‘1’ and ‘0’, $\sigma 1$, $\sigma 2$: standard deviation of logic ‘1’ and ‘0’).

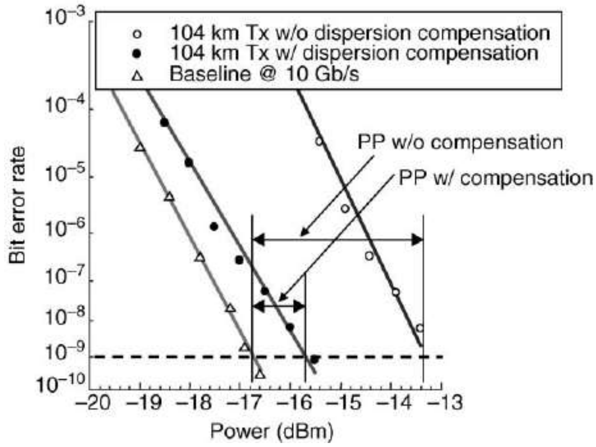

Figure 26 BER curve and power penalty.

Power penalty (PP) is one of the most important system parameters, and is defined as the received optical power difference in dB with and without signal impairments at a specified BER (conventionally 10^{-9}), from measured BER versus optical power curve. Therefore, 1 dB PP means that a system with signal impairments requires 1 dB more optical power at the receiver in order to achieve the same BER performance compared to the system without signal impairments. Figure 26 shows the typical BER curve as a function of received optical power and PP. Note that PP reduces from ~3.5 dB to ~1 dB with dispersion compensation.

In long-haul and undersea transmission system with many EDFA chains, OSNR is an important system parameter to design and characterize the optical transport system. OSNR is defined as the ratio of the power of signal channel to the power of ASE in a specified optical bandwidth (conventionally 0.1 nm). The OSNR (in dB) of a signal channel at the end of the system is approximated by: $\text{OSNR (dB)} = 58 + P_{\text{out}} - L_{\text{span}} - \text{NF} - 10\log(N_{\text{amp}})$, when the system has N_{amp} fiber spans, span loss L_{span} (in dB) followed by an optical amplifier with output power P_{out} (in dBm) per channel launched into the span and noise figure (NF) (in dB). As an example, a typical OSNR requirement for $\text{BER} = 10^{-9}$ is about 17 dB at 10 Gb/s data rate. The required OSNR should be increased by 6 dB when the data rate is increased by a factor of 4. Improving the amplifier's NF can increase OSNR, and the improved OSNR can be used to increase the system reach, reducing the channel power in order to suppress the nonlinearities, etc.

See also

Fiber and Guided Wave Optics: Light Propagation. **Imaging:** Multiplex Imaging. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Leonid K, Sergio B and Alan W (1996) *Optical Fiber Communication Systems*. Norwood, MA: Artech House Inc.

Govind PA (1995) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Emmanuel D (1994) *Erbium-Doped Fiber Amplifiers*. New York: Wiley.

Ivan K and Tingye L (2002) *Optical Fiber Telecommunications IV A Components*. San Diego, CA: Academic Press.

Ivan K and Tingye L (2002) *Optical Fiber Telecommunications IV B Systems and Impairments*. San Diego, CA: Academic Press.

Historical Development

G Keiser, PhotonicsComm Solutions, Inc., Newton Center, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article describes the development of optical fiber communication systems. After describing some of the motivations for using optical fiber communications and the advantages of this technology, the key milestones and the principal people involved in developing optical fibers and compatible light sources are presented. Following this, the article looks at the evolution of fielded systems and the use of optical fiber links in undersea applications.

One of the principal needs of people since antiquity has been to communicate. This need created interests in devising communication systems for sending messages from one distant place to another. Many forms of such systems have appeared over the years. The basic motivations behind each new one were either to improve the transmission fidelity, to increase the data rate so that more information could be sent, or to increase the transmission distance between relay stations. Prior to the nineteenth century, all communication systems operated at a very low information rate and involved only optical or acoustical means, such as signal lamps or horns. One of the earliest known optical transmission links, for example, was the use of a fire signal by the Greeks in the eighth century BC for sending alarms, calls for help, or announcements of special events.

The invention of the telegraph by Samuel FB Morse in 1838 ushered in the era of electrical communications. In the ensuing years an increasingly larger portion of the electromagnetic spectrum, shown in Figure 1, was utilized for the conveying of information from one place to another. The reason for this trend is that, in electrical systems, the data usually are transferred over the communication channel by superimposing the information onto a sinusoidally varying electromagnetic wave, which is known as the carrier. When reaching its destination the information is removed from the carrier wave and processed as desired. Since the amount of information that can be transmitted is related directly to the frequency range over which the carrier operates, increasing the carrier frequency theoretically increases the available transmission bandwidth and, consequently, provides a larger information capacity. Thus the trend in electrical communication system developments was to employ progressively higher frequencies (shorter wavelengths), which offer corresponding increases in bandwidth or information capacity. This activity led to the birth of communication mechanisms such as radio, television, microwave, and satellite links.

Another important portion of the electromagnetic spectrum encompasses the optical region shown in Figure 1. In contrast to electrical communications, transmission of information in an optical format is carried out, not by frequency modulation of the carrier but by varying the intensity of the optical power. Similar to the radio-frequency spectrum, two classes of transmission medium can be used: an atmospheric channel or a guided-wave channel. Whereas transmission of optical signals through the atmosphere was done thousands of years ago, the use of a guided-wave optical channel such as an optical fiber is a fairly recent application.

Fiber optic communication systems have a number of inherent advantages over their copper-based and radio-transmission counterparts. Fiber optic cable can transmit at a higher capacity, thereby reducing the number of physical lines and the amount of equipment needed for a given transmission span. The lower weight and smaller size of optical fibers offer a distinct advantage over heavy, bulky copper cables in crowded underground cable ducts, in ceiling-mounted cable trays, and in mobile platforms such as aircraft and ships. Their dielectric composition is an especially important feature of optical fibers, since this ensures freedom from electromagnetic interference between adjacent fibers, eliminates ground loops, and results in extremely low fiber-to-fiber crosstalk. In addition, an optical fiber affords a high degree of data security since the optical signal is well confined within the optical waveguide.

Optical Fiber Development

Since an atmospheric channel requires a line-of-sight link and because it can be adversely affected by weather conditions, a guided-wave channel is the preferred approach in most cases for communication system applications. A challenge in using an optical fiber channel is to have a flexible, low-loss medium that transfers the optical signal over long distances

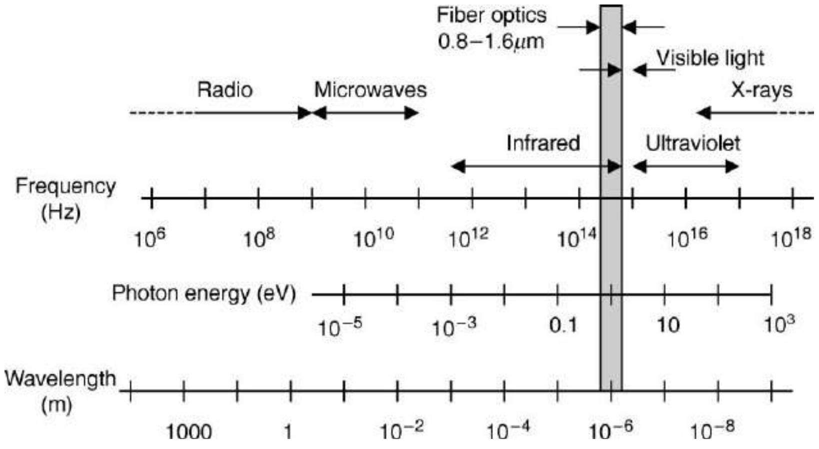

Figure 1 Electromagnetic spectrum for electrical and optical communications.

without significant attenuation and distortion. Glass is an obvious material for such applications. The earliest known glass was made around 2500 BC and glass already was drawn into fibers during the time of the Roman Empire. However, such glasses have very high attenuations and thus are not suitable for communication applications. One of the first known attempts of using optical fibers for communication purposes was a demonstration in 1930 by Heinrich Lamm of image transmission through a short bundle of optical fibers for potential medical imaging. However, no further work was done beyond the demonstration phase since the technology for producing reasonably low-loss fibers with good light confinement was not yet mature.

Further work and experiments continued on using optical fibers for image transmission and by 1960 glass-clad fibers had attenuations of about one decibel per meter. This attenuation allowed fibers to be used for medical imaging, but it was still much too high for communications. Optical fibers had attracted the attention of researchers at that time, because they were analogous in theory to plastic dielectric waveguides used in certain microwave applications. In 1961, Elias Snitzer demonstrated this similarity by drawing fibers with cores so small they carried light in only one waveguide mode. He published a classic theoretical description of single-mode fibers in May 1961. However, to be useful for communication systems, optical fibers would need to have a loss of no more than 10 or 20 decibels per kilometer.

As a reminder, decibels measure the ratio of the output power to the input power on a logarithmic scale. The abbreviation for a decibel is 'dB.' The power ratio in decibels is given by the expression '10 log (power out/power in).' As an example, a power loss of 20 dB over a 1-km distance in an optical fiber means that only 1% of the light injected into the fiber comes out of the other end. Table 1 gives some typical examples of various power ratios and their decibel equivalents.

In the early 1960s Charles Kao pursued the idea of using a clad glass fiber for an optical waveguide, building on optical waveguide research being done at the Standard Telecommunication Laboratories in England. After he and George Hockman painstakingly examined the transparency properties of various types of glass, Kao made a prediction, in 1966, that losses of no more than 20 dB/km were possible in optical fibers. In July 1966, Kao and Hockman presented a detailed analysis for achieving such a loss level. Kao then went on to actively advocate and promote the prospects of fiber communications, which generated interest in laboratories around the world to reduce fiber loss. It took four years to reach Kao's predicted goal of 20 dB/km, and the final solution was different from what many had expected.

Table 1 Examples of some optical power ratios and their decibel equivalents

Power ratio (P_{out}/P_{in})	*Decibel equivalent (dB)*
0.01	−20
0.10	−10
0.50	−3
1	0
2	3
10	10
100	20

To understand the process of making a fiber, consider the schematic of a typical fiber structure shown in Figure 2. A fiber consists of a solid dielectric (glass or plastic) cylinder of refractive index n_1 called the core. This is surrounded by a dielectric cladding which has a refractive index n_2, that is less than n_1, in order to achieve light guiding in the fiber. This structure is then encapsulated by a buffer coating to protect the fiber from abrasion and outside contaminants. The first step in making a fiber is to form the clear glass rod or tube called a preform. This normally is done by a vapor-phase oxidation process. The preform has two distinct regions that correspond to the core and cladding of the eventual fiber. Fibers are made from the preform by precision feeding it into a circular furnace that softens the end of the preform to the point where it can be drawn into a very thin filament which becomes the optical fiber.

Most researchers had tried to purify the compound glasses used for standard optics, which are easy to melt and draw into fibers. A different approach was taken at the Corning Glass Works where Robert Maurer, Donald Keck, and Peter Schultz started with fused silica. This material can be made extremely pure, but has a high melting point and a low refractive index. Silica has the approximate attenuation versus wavelength characteristic shown in Figure 3. Note that for early silica fibers there are regions of low attenuation around 850, 1310, and 1550 nm, which the literature refers to as the first, second, and third windows, respectively. The large attenuation spikes in the 1000- and 1400-nm spectral

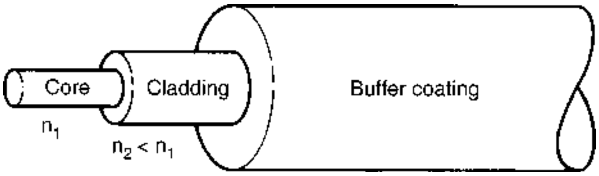

Figure 2 Typical optical fiber structure.

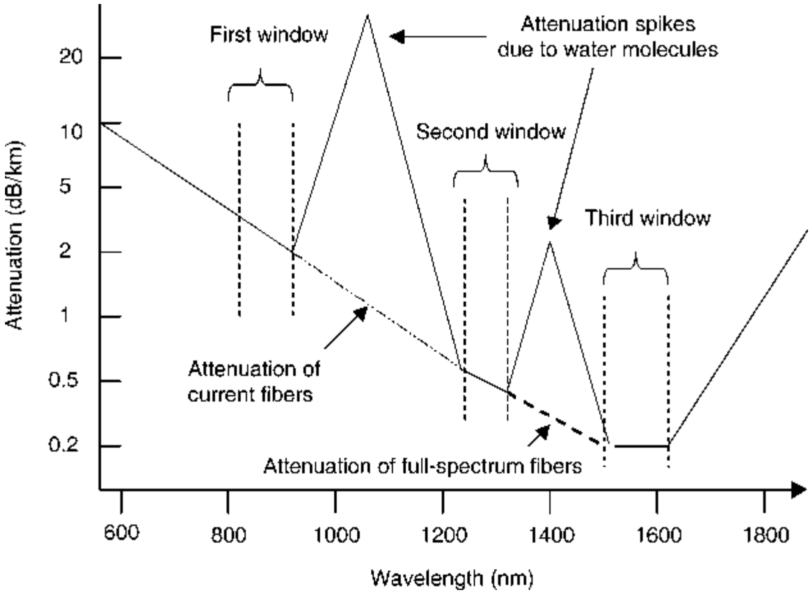

Figure 3 Attenuation versus wavelength approximation for silica fibers.

regions are due to absorption by residual water molecules in the glass. Although ultra-pure material processing techniques can eliminate these spikes to produce what are known as full-spectrum or low-water-peak fibers, most installed fibers still have a relatively large attenuation between 1350 and 1500 nm.

Note that since the attenuation spikes no longer are present in various types of fibers, the idea of operating windows has been replaced by the concept of spectral bands. The International Telecommunications Union (ITU) has designated six spectral bands for use in intermediate-range and long-distance optical fiber communications within the 1260 to 1675-nm region. The regions, which are known by the letters O, E, S, C, L, and U, are defined as follows:

- Original Band (O-Band): 1260 to 1360 nm
- Extended Band (E-Band): 1360 to 1460 nm
- Short Band (S-Band): 1460 to 1530 nm
- Conventional Band (C-Band): 1530 to 1565 nm
- Long Band (L-Band): 1565 to 1625 nm
- Ultra-long Band (U-Band): 1625 to 1675 nm

The Corning team made cylindrical preforms by depositing purified materials from the vapor phase. To produce a fiber that has light guiding properties they carefully added controlled trace levels of titanium to the core to make its refractive index slightly higher than that of the cladding without raising the attenuation significantly. In September 1970, they announced the fabrication of single-mode fibers with an attenuation of 17dB/km at the 633-nm helium-neon line. The fibers were fragile, but independent tests at the British Post Office Research Laboratories facility in Martlesham Heath, England confirmed the low loss.

This dramatic breakthrough was the first among the many developments that opened the door to fiber optic communications. The ensuing years saw further reductions in optical fiber attenuations. By mid-1972, Maurer, Keck, and Schultz at Corning had made multimode germania-doped fibers with a 4-dB/km loss and much greater strength than the earlier brittle titania-doped fibers. In order to couple a sufficient amount of optical power into a fiber, early applications used multimode fibers with a refractive-index gradient between core and cladding of around 2% and core diameters of 50 or 62.5 micrometers.

Single-mode fibers have much smaller core diameters on the order of 9 micrometers in order to allow only one propagation mode. This type of fiber has a much higher transmission capacity since the effect of modal dispersion is eliminated. The first single-mode fibers were optimized to have a zero dispersion value at 1310 nm, since the silica material used at that time exhibited a low loss within a spectral band around this wavelength. Figure 4 shows the dispersion characteristic of this type of fiber as a function of wavelength in the S-, C-, and L-bands. As a result of its widespread use in early single-mode transmission systems, this fiber design has been standardized by the International Telecommunication Union (ITU-T) under the designation Recommendation G.652.

Standard G.652 silica fibers provide the lowest attenuation at 1550 nm, but have a much larger signal

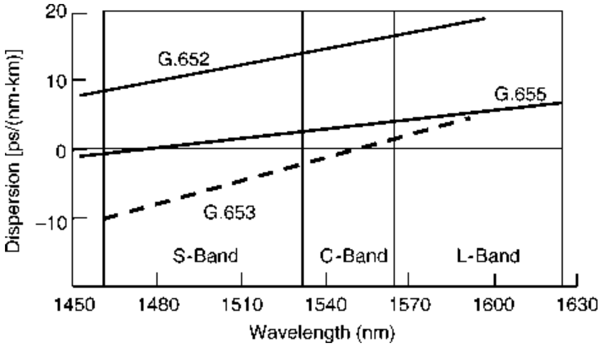

Figure 4 Dispersion characteristics of three types of standard optical fibers.

dispersion at this wavelength than at 1310 nm. Since system designers wanted to use fibers at a 1550-nm wavelength, in order to transmit high-speed data over longer distances, fiber manufacturers overcame the larger signal dispersion limitation by creating the so-called dispersion-shifted fibers. This was done through a clever manipulation of the core and cladding designs that allowed the zero-dispersion point to shift to longer wavelengths. In particular, the ITU-T created a specification for operation at 1550 nm which is designated as recommendation G.653. Figure 4 also shows the dispersion characteristic of this type of fiber as a function of wavelength.

Although the G.652 and G.653 fibers work well for single-wavelength operation, a different type of fiber, having non-zero dispersion within a broad spectral range, is needed when implementing systems that use many independent light sources simultaneously within a particular wavelength band. This led to the specification of the non-zero dispersion shifted fiber (NZDSF) in the ITU-T Recommendation G.655 that are designated to operate around 1550 nm. The main purpose of having a positive dispersion value over the entire operating spectrum is to mitigate a nonlinear optical effect called four-wave mixing (FWM), which is analogous to intermodulation distortion in electrical systems. Figure 4 shows the dispersion characteristic of the G.655 fiber in the S-, C-, and L-bands.

Light Sources

The fiber by itself is not practical unless there is a compatible optical source for launching light signals into it. The most suitable device for this is a semiconductor light-emitting diode (LED) or a laser diode. In the 1960s, a great deal of effort took place to achieve laser action in pn-junction diodes. The early devices were GaAs and GaAsP lasers that operated at a temperature of 77 K, emitted at a wavelength around 850 nm, and had high lasing threshold current densities. To make devices that were more application-friendly by operating at room temperature, structures consisting of sandwiched layers of AlGaAs and GaAs were investigated. Finally, in 1970, researchers at Bell Laboratories and a team at the Ioffe Physical Institute in Leningrad made the first semiconductor diode lasers based on a layered AlGaAs/GaAs/AlGaAs structure that were able to emit continuous-wave light in the 850-nm region at room temperature.

Major improvements in laser diode performance and reliability followed this achievement during the next decade. In addition, around 1976, room-temperature laser diodes operating at longer wavelengths, in the 1100 to 1600-nm region, started to appear. Of particular interest were GaInAsP/InP-based laser diodes emitting in the 1310-nm and 1550-nm low-loss windows of optical fibers. The progressive development of ever-improving devices during the 1980s and 1990s, included single-frequency emission with narrow linewidths under continuous operation, low levels of chirp under direct modulation, high output power, and the ability to tune specially constructed laser diodes over a wavelength range of up to 30 nm.

Fielded Systems

The bit rate-distance product B × L, where B is the transmission bit rate and L is the repeater spacing, measures the transmission capacity of optical fiber links. Since the inception of optical fiber communications in the mid-1970s, the link transmission capacity has experienced a tenfold increase every four years. Several major technology advances spurred this growth. Among the technology developments are laser diodes emitting over an extremely narrow spectral band, optical amplifiers, fibers with low losses and low dispersions, and the concepts of wavelength division multiplexing.

Some of the initial telephone-system field trials in the USA were carried out in 1977 by GTE in Los Angeles and by AT&T in Chicago. These transmission links were largely for the trunking of telephone lines, which are digital links consisting of time-division-multiplexed 64-kb/s voice channels. Similar demonstrations were carried out in Europe and Japan. Applications ranged from 45 to 140 Mb/s with repeater spacings around 10 km.

With the advent of high-capacity fiber optic transmission lines in the 1980s, service providers established a standard signal format called synchronous optical network (SONET) in North America and synchronous digital hierarchy (SDH) in other parts of the world. These standards define a synchronous frame structure for sending multiplexed digital traffic over optical fiber trunk lines. The basic building block and first level of the SONET signal hierarchy is called the synchronous transport signal–level 1 (STS-1), which has a bit rate of 51.84 Mb/s. Higher-rate SONET signals are obtained by byte-interleaving N STS-1 frames, which are then scrambled and converted to an optical carrier–level N (OC-N) signal. Thus the OC-N signal will have a line rate exactly N times that of an OC-1 signal. For SDH systems, the fundamental building block is the 155.52-Mb/s synchronous transport module–level 1 (STM-1). Again, higher-rate information streams are generated by synchronously multiplexing N different STM-1 signals to form the STM-N signal. Table 2 shows commonly used SDH and SONET signal levels.

The development of optical sources and photodetectors capable of operating at 1310 nm allowed a shift in the transmission wavelength from 850 nm to 1310 nm. This resulted in a substantial increase in the repeaterless transmission distance for long-haul telephone trunks, since optical fibers exhibit lower power loss and less signal dispersion at 1310 nm. Intercity applications first used multimode fibers, but in 1984 switched exclusively to single-mode fibers, which have a significantly larger bandwidth. Bit rates for long-haul links typically range between 155 and 622 Mb/s (OC-3 and OC-12), and in some cases up to 2.5 Gb/s (OC-48), over repeater spacings of 40 km. Both multimode and single-mode 1310-nm fibers are used in local area networks, where bit rates range from 10 Mb/s to 100 Mb/s over distances ranging from 500 m to tens of kilometers.

In the next step of system evolution, links operating in the low-loss window around 1550 nm attracted much attention for high-capacity, long-span terrestrial and undersea transmission links. These links routinely carry traffic at 2.5 Gb/s over 90-km repeaterless distances. By 1996, advances in high-quality lasers and receivers allowed single-wavelength transmission rates of 10 Gb/s (OC-192).

In 1989, the introduction of optical amplifiers gave a major boost to fiber transmission capacity. Although there are GaAlAs-based solid-state optical amplifiers for the first window and InGaAsP amplifiers for the second window, the most successful and widely used devices are erbium-doped fiber amplifiers (commonly called EDFAs) operating in the 1530 to 1560-nm range and Raman fiber amplifiers that are used for operation in the 1560 to 1600-nm region.

During the same time period, impressive demonstrations of long-distance high-capacity systems were made using optical soliton signals. A soliton is a nondispersive pulse that makes use of nonlinear dispersion properties in a fiber to cancel out chromatic dispersion effects. As an example, solitons at rates of 10 Gb/s have been sent over a 12 200-km experimental link using optical amplifiers and special modulation techniques.

Entrance of Wavelength Division Multiplexing

The use of wavelength division multiplexing (WDM) offers a further boost in fiber transmission capacity. The basis of WDM is to use multiple sources operating at slightly different wavelengths to transmit several independent information streams over the same fiber. Although researchers started looking at WDM in the 1970s, during the ensuing years it generally turned out to be easier to implement higher-speed electronic and optical devices than to invoke the greater system complexity called for in WDM. However, a dramatic surge in WDM popularity started in the early 1990s, as electronic devices neared their modulation limit and high-speed equipment became increasingly complex.

Figure 5 shows the concept of implementing many closely spaced wavelengths within a spectral band centered around 1552.524 nm. This scheme is referred to as dense WDM or DWDM. Conceptually, the DWDM scheme is the same as frequency

Table 2 Commonly used SONET and SDH transmission rates

SONET level	*Electrical level*	*Line rate (Mb/s)*	*SDH equivalent*	*Common rate name*
OC-1	STS-1	51.84	–	
OC-3	STS-3	155.52	STM-1	155 Mb/s
OC-12	STS-12	622.08	STM-4	622 Mb/s
OC-24	STS-24	1244.16	STM-8	
OC-48	STS-48	2488.32	STM-16	2.5 Gb/s
OC-96	STS-96	4976.64	STM-32	
OC-192	STS-192	9953.28	STM-64	10 Gb/s
OC-768	STS-768	39,813.12	STM-256	40 Gb/s

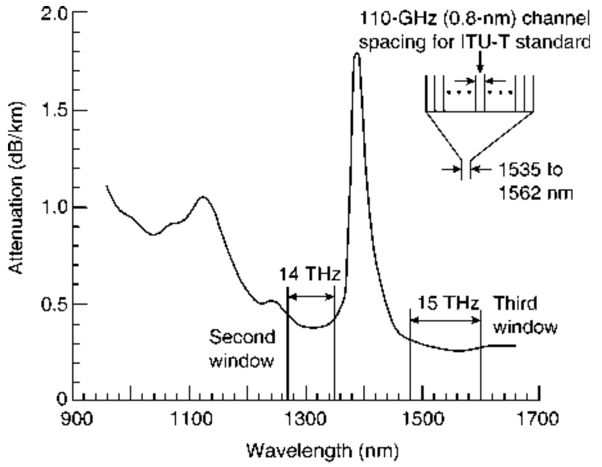

Figure 5 Wavelength division multiplexing (WDM) concept.

division multiplexing (FDM) used in microwave radio and satellite systems. Just as in FDM, the wavelengths (or optical frequencies) in a DWDM link must be properly spaced to avoid interference between channels. In an optical system this interference may arise from the fact that the center wavelength of laser diode sources and the spectral operating characteristics of other optical components in the link may drift with temperature and time, thereby giving rise to the need for a guard band between wavelength channels.

Since WDM is essentially frequency division multiplexing at optical carrier frequencies, the ITU developed DWDM standards that specify channel spacings in terms of frequency. The ITU-T Recommendation G.694.1, which is entitled 'Dense Wavelength Division Multiplexing (DWDM),' specifies WDM operation in the S-, C-, and L-bands for high-quality, high-rate metro area network (MAN) and wide area network (WAN) services. It calls out for narrow frequency spacings of 100 to 12.5 GHz (or, equivalently, 0.8 to 0.1 nm at 1550 nm). This implementation requires the use of stable, high-quality, temperature-controlled and wavelength-controlled (frequency-locked) laser diode light sources.

With the production of full-spectrum (low-water-content) fibers, the development of relatively inexpensive optical sources, and the desire to have low-cost optical links operating in metro and local area networks, came the concept of coarse WDM (CWDM). In 2002, the ITU-T released a standard aimed specifically at CWDM. This is Recommendation G.694.2, which is entitled 'Coarse Wavelength Division Multiplexing (CWDM).' The CWDM grid is made up of 18 wavelengths defined within the range 1270 nm to 1610 nm (O-through L-bands) spaced by 20 nm with wavelength-drift tolerances of ± 2 nm. This can be achieved with inexpensive light sources that are not temperature-controlled. The targeted transmission distance for CWDM is 50 km on single-mode fibers, such as those specified in ITU-T Recommendations G.652, G.653, and G.655.

Wavelength tunability of a source is an important property of WDM systems. Obviously it is not desirable or practical to maintain an inventory of dozens of lasers that emit at different wavelengths for WDM applications. The ideal tunable laser should be adjustable to emit at a specific wavelength across a broad spectral range. One such device is a distributed Bragg reflector (DBR) laser diode that can be tuned over a 10 to 20 nm spectral range. Work on perfecting such devices are still underway.

Starting in the mid-1990s, a combination of EDFAs and WDM was used to boost fiber information capacity to even higher levels and to increase the transmission distance. A major system consideration in these super-high capacity links is to ensure that there is appropriate link and equipment redundancy, so that alternate paths are available in case of disruptions in communications resulting from cable ruptures (for example, caused by errant digging from a backhoe) or equipment failures at an intermediate node. Such disruptions otherwise could have a devastating effect on a large group of users.

Undersea Optical Cable Systems

The first transoceanic fiber optic cable systems were installed in the Atlantic and Pacific Oceans in 1988 and 1989. Initially these systems operated at 280 Mb/s per fiber pair using 1310-nm lasers and single-mode fibers. The links consisted of a series of point-to-point optical fiber segments between electronic-based undersea regeneration points that were located nominally 60 km apart. Later the transmission capacity of these links was upgraded to 2.5 Gb/s and the regenerator spacing was increased to 100 km by converting the 1310-nm multiple-frequency light sources to 1550-nm single-frequency laser diodes. Later, the regenerator spacing was increased to 140 km.

Although these cable systems significantly improved the quality of the international telephone service, the optical-to-electrical conversion process at each regeneration point remained a capacity bottleneck. The introduction of erbium-doped optical fiber amplifiers (EDFA) eliminated this bottleneck from undersea lightwave systems by amplifying signals directly in the optical domain. Since EDFAs operate over a 30-nm wavelength band, they are well suited for use with undersea WDM links, which have provided a further capacity increase. These undersea optical amplifiers are typically spaced about 45 km apart.

One example of the many worldwide installations of optically amplified WDM networks is the SEA-ME-WE-3 Cable System. This undersea network runs from Germany to Singapore, connecting more than a dozen countries in between. Hence the name SEA-ME-WE, which refers to Southeast Asia (SEA), the Middle East (ME), and Western Europe (WE). The network has two pairs of undersea fibers with a capacity of eight STM-16 wavelengths per fiber.

See also

Lasers: Optical Fiber Lasers. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Wavelength Division Multiplexing.

Further Reading

Bergano NS (1997) Undersea amplified lightwave system design. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIA, pp. 302–335. New York: Academic.

Goralski WJ (2002) SONET/SDH, 3rd edn. New York: McGraw-Hill.

Hecht J (1999) *City of Light*. New York: Oxford University Press.

Joyner CH (1997) Semiconductor laser growth and fabrication technology. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIB, pp. 163–199. New York: Academic.

Kaiser P and Keck DB (1988) Fiber types and their status. In: Miller SE and Kaminow IP (eds) *Optical Fiber Telecommunications II*, pp. 29–54. New York: Academic.

Kao KC and Hockman GA (July 1966) Dielectric-fibre surface waveguides for optical frequencies. *Proceedings IEE*, vol. 113, 1151–1158.

Keiser G (2003) *Optical Communications Essentials*. New York: McGraw-Hill.

Keiser G (2000) *Optical Fiber Communications*, 3rd edn. Burr Ridge, IL: McGraw-Hill.

Mollenauer LF, Gordon JP and Mamyshev PV (1997) Solitons in high bit rate long-distance transmission. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIA, pp. 373–460. New York: Academic.

Nyman B, Farries M and Si C (Oct 2001) Technology trends in dense WDM demultiplexers. *Optical Fiber Technology* 7(4): 255–274.

Ramaswami R and Sivarajan KN (2002) *Optical Networks*, 2nd edn. San Francisco, CA: Morgan Kaufmann.

Snitzer E (May 1961) Cylindrical dielectric waveguide modes. *Journal Optical Society of America* 51: 491–498.

Trischitta PR and Marra WC (1998) Applying WDM technology to undersea cable networks. *IEEE Communication Magazine* 36(2): 62–66.

Architectures of Optical Fiber Communication Systems

G Keiser, PhotonicsComm Solutions, Inc.,
Newton Center, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Simple Optical Fiber Links

The most basic optical fiber communication system architecture is a point-to-point link consisting of the elements shown in Figure 1. The key sections are a transmitter consisting of a light source and its associated drive circuitry, a cable offering mechanical and environmental protection to the optical fibers contained inside, and a receiver consisting of a photodetector plus amplification and signal-restoring circuitry. Very long links may include periodic optical amplifiers for boosting the level of the optical signal. In networks that use only simple point-to-point links, the optical fiber is used principally as a transmission medium and all the switching and processing of information bits is done by electronics.

At the transmitting end of an optical fiber link, information bits consisting of electronic pulses are used to modulate the light output of the source. The resulting light pulses are coupled into an optical fiber. As they travel along the fiber these signal pulses become progressively weakened and distorted due to attenuation and dispersion mechanism that are characteristic of an optical fiber. At the end of the fiber link a photodetector converts the optical pulses back to an electrical format for switching to another link or for processing at the end station.

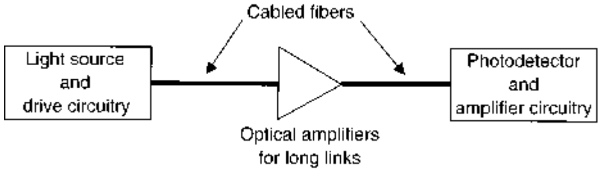
Cabled fibers
Light source and drive circuitry
Photodetector and amplifier circuitry
Optical amplifiers for long links

Figure 1 Basic elements of a point-to-point fiber optic link.

The cabled fiber is one of the most important elements in an optical fiber link, since it determines operational parameters such as transmission distance and achievable data rates. In addition to protecting the glass fibers during installation and service, the cable may contain copper wires for powering optical amplifiers or signal regenerators, which are needed periodically in long-distance links for amplifying and reshaping the signal.

Analogous to copper cables, the installation of optical fiber cables can be either aerial, in ducts, undersea, or buried directly in the ground. As a result of installation and/or manufacturing limitations, individual cable lengths will range from several hundred meters to several kilometers. Practical considerations such as reel size and cable weight determine the actual length of a single cable section. The shorter segments tend to be used when the cables are pulled through ducts. Longer lengths are used in aerial, direct-burial, or undersea applications. Splicing together individual cable sections forms continuous transmission lines for these long-distance links. For undersea installations, the splicing and repeater-installation functions are carried out on board a specially designed cable-laying ship.

One of the principal characteristics of an optical fiber is its attenuation as a function of wavelength, as shown in Figure 2. Early technology used the 800 to 900 nm wavelength band, since in this region the fibers made at that time exhibited a local minimum in the attenuation curve, and optical sources and photodetectors operating at these wavelengths were available. This region originally was referred to as the 'first window'. By reducing the concentration of hydroxyl ions and metallic impurities in the fiber material, in the 1980s manufacturers were able to fabricate optical fibers with very low loss in the 1100 to 1600 nm region. Two windows are defined here, these being the 'second window' centered around 1310 nm and the 'third window' centered around 1550 nm.

In 1998 a new ultra-high purifying process patented by Lucent Technologies eliminated virtually all water molecules from the glass material. The resulting fiber is designated as either a low-water-peak or a full spectrum fiber. By dramatically reducing the water-attenuation peak around 1400 nm, this process opens the transmission region between the second and third windows to provide around 100 nm more bandwidth than in conventional single-mode fibers. This particular fiber allows wider spectral tolerances on optical components, which reduces costs for moderate-distance metropolitan network applications.

Since the attenuation of low-water-peak fibers makes the designation of transmission windows obsolete, the concept of operational spectral bands arose for the 1260–1675 nm region. As shown in

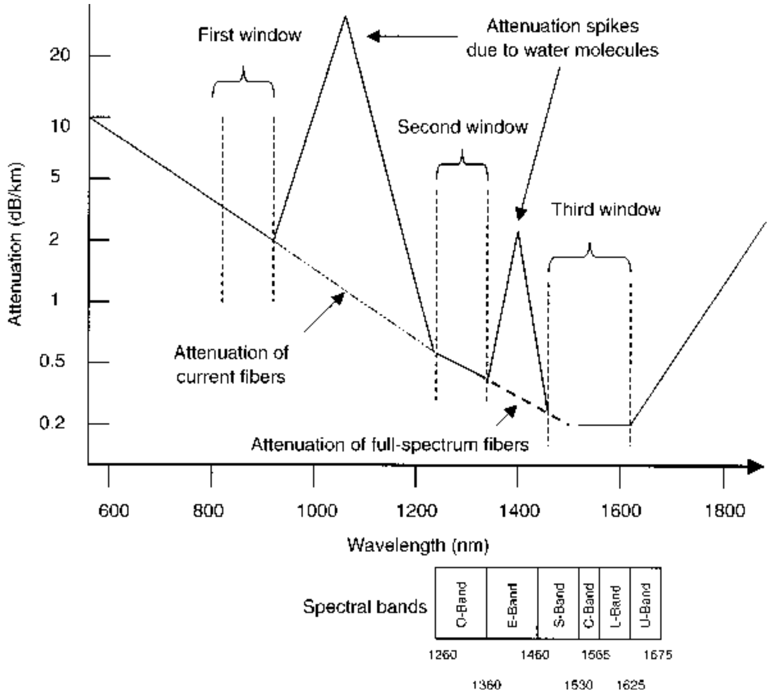

Figure 2 Attenuation versus wavelength for silica fibers and the concept of spectral bands.

Figure 2, the regions are known by the letters O, E, S, C, L, and U, which are defined as follows:

- Original band (O-Band): 1260–1360 nm;
- Extended band (E-Band): 1360–1460 nm;
- Short band (S-Band): 1460–1530 nm;
- Conventional band (C-Band): 1530–1565 nm;
- Long band (L-Band) 1565–1625 nm;
- Ultra-long band (U-Band): 1625–1675 nm.

Once the cable is installed, a light source that is dimensionally compatible with the fiber core is used to launch optical power into the fiber. Semiconductor light-emitting diodes (LEDs) and laser diodes are suitable for this purpose, since their light output can be modulated rapidly by simply varying the bias current at the desired transmission rate, thereby producing an optical signal. The electric input signals to the transmitter circuitry for the optical source can be either of an analog or digital form. For high-rate systems (nominally greater than 1 Gb/s), direct modulation of the source can lead to unacceptable signal distortion. In this case, an external modulator is used to vary the amplitude of a continuous light output from a laser diode source.

After an optical signal is launched into a fiber, it will become progressively attenuated and distorted with increasing distance because of scattering, absorption, and dispersion mechanisms in the glass material. At the receiver a photodiode will detect the weakened optical signal emerging from the fiber end and convert it to an electric current. The design of an optical receiver is inherently more complex than that of the transmitter, since it has to interpret the content of the weakened and degraded signal received by the photodetector. The principal figure of merit for a receiver is the maximum optical power necessary at the desired data rate to attain either a given error probability for digital systems or a specified signal-to-noise ratio for an analog system. The ability of a receiver to achieve a certain performance level depends on the photodetector type, the effects of noise in the system, and the characteristics of the successive amplification stages in the receiver.

Networks of Links

With the advent of fiber optic transmission lines, the next step in the evolution of the digital multiplexing scheme used by telecommunication providers was a standard signal format called synchronous optical network (SONET) in North America and synchronous digital hierarchy (SDH) in other parts of the world. In the mid-1980s, efforts started on developing a standard that would allow network engineers to interconnect fiber optic transmission equipment from various vendors through multiple-owner trunk networks. This resulted in a series of standards for SONET by the American National Standards Institute (ANSI) and a series of recommendations for SDH from the International Telecommunications Union (ITU). Examples of particular interest are the ANSI T1.105.06 standard and the ITU-T G.957 recommendation. Although there are some implementation differences between SONET and SDH, all SONET specifications conform to the SDH recommendations. A key characteristic of SONET and SDH is that they are usually configured as a ring architecture. This is done to create loop diversity for uninterrupted service protection purposes in case of link or equipment failures. The SONET/SDH rings commonly are called self-healing rings, since the traffic flowing along a certain path can be switched automatically to an alternate or standby path following failure or degradation of the link segment.

The basic building block and first level of the SONET signal hierarchy is called the Synchronous Transport Signal – Level 1 (STS-1), which has a bit rate of 51.84 Mb/s. Higher-rate SONET signals are obtained by byte-interleaving *N* STS-1 frames, which are then scrambled and converted to an Optical Carrier – Level *N* (OC-N) signal. Thus the OC-N signal will have a line rate exactly *N* times that of an OC-1 signal. For SDH systems the fundamental building block is the 155.52 Mb/s Synchronous Transport Module – Level 1 (STM-1). Again, higher-rate information streams are generated by synchronously multiplexing *N* different STM-1 signals to form the STM-N signal.

Three main features, each with two alternatives, classify all SONET/SDH rings, thus yielding eight possible combinations of ring types. First, there can be either two or four fibers running between the nodes on a ring. Second, the operating signals can travel either clockwise only (which is termed a unidirectional ring) or in both directions around the ring (which is called a bidirectional ring) Third, protection switching can be performed either via a line-switching or a path-switching scheme. Upon link failure or degradation, line switching moves all signal channels of an entire OC-N channel to a protection fiber. Conversely, path switching can move individual payload channels within an OC-N channel (e.g., an STS-1 channel within an OC-12 channel) to another path.

Of the eight possible combinations of ring types, the following two architectures have become popular for SONET and SDH networks:

- two-fiber unidirectional path-switched ring (two-fiber UPSR);
- two-fiber or four-fiber bidirectional line-switched ring (two-fiber or four-fiber BLSR).

The common abbreviations of these configurations are given in parentheses. They also are referred to as unidirectional or bidirectional self-healing rings (USHR or BSHR).

Figure 3 shows a two-fiber unidirectional path-switched ring network. By convention, in a unidirectional ring the normal working traffic travels clockwise around the ring, as indicated by the heavy arrows. For example, the connection from node 1 to node 3 uses links 1 and 2, whereas the traffic from node 3 to node 1 traverses links 3 and 4. Thus, two communicating nodes use a specific bandwidth capacity around the entire perimeter of the ring. If nodes 1 and 3 exchange information at an OC-3 rate in an OC-12 ring, then they use one-fourth of the capacity around the ring on all the primary links. In a unidirectional ring the counter-clockwise path is used as an alternate path for protection against link or node failures. To achieve this, the signal from a transmitting node is dual-fed into both the primary and protection fibers. This establishes a designated protection path on which traffic flows counter-clockwise, i.e., from node 1 to node 3 via links 5 and 6 (in that order), as shown in Figure 3b. A heavy line and a dashed line indicate the primary and protection paths, respectively.

Consequently, two identical signals from a particular node arrive at their destination from opposite directions, usually with different delays, as denoted in Figure 3b. The receiver normally selects the signal from the primary path. However, it continuously compares the fidelity of each signal and chooses the alternate signal in case of severe degradation or loss of the primary signal. Thus, each path is individually switched based on the quality of the received signal. For example, if path 2 breaks or equipment in node 2 fails, then node 3 will switch to the protection channel to receive signals from node 1.

Figure 4 illustrates the architecture of a four-fiber bidirectional line-switched ring. Here two primary fiber loops (with fiber segments labeled 1p through 8p) are used for normal bidirectional communication, and the other two secondary fiber loops are standby links for protection purposes (with fiber segments labeled 1s through 8s). In contrast to the two-fiber UPSR, the four-fiber BLSR has a capacity advantage since it uses twice as much fiber cabling and because traffic between two nodes is only sent partially around the ring.

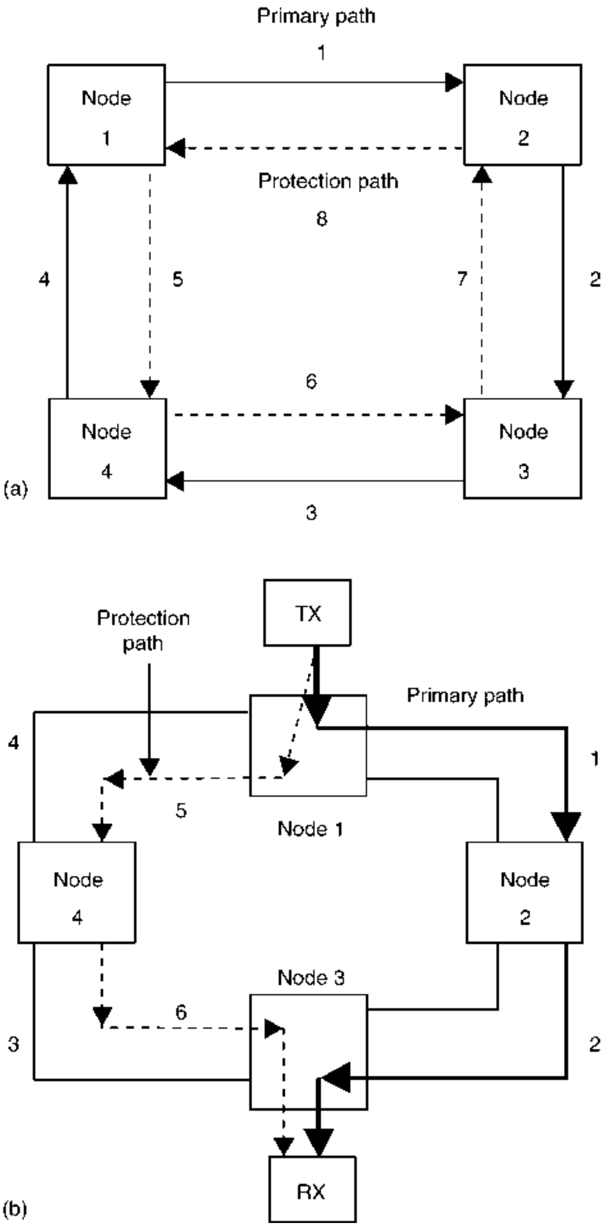

Figure 3 Schematic of a two-fiber unidirectional path-switched ring network.

To see the function and versatility of the standby links in the four-fiber BLSR, consider first the case where a transmitter or receiver circuit card used on the primary ring fails in either node 3 or 4. In this situation the affected nodes detect a loss-of-signal condition and switch both primary fibers connecting them to the secondary protection pair, as shown in Figure 5. The protection segment between nodes 3 and 4 now becomes part of the primary bidirectional loop. The exact same reconfiguration scenario will

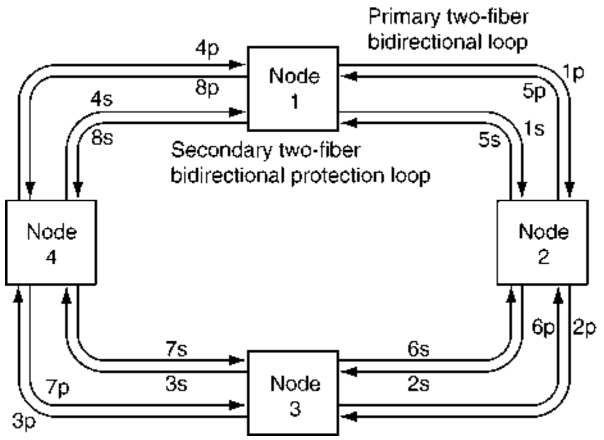
Primary two-fiber bidirectional loop
4p
8p
4s
8s
Node 1
1p
5p
1s
5s
Secondary two-fiber bidirectional protection loop
Node 4
Node 2
6p
2p
7s
6s
Node 3
7p
3s
2s
3p

Figure 4 The architecture of a four-fiber bidirectional, line-switched ring.

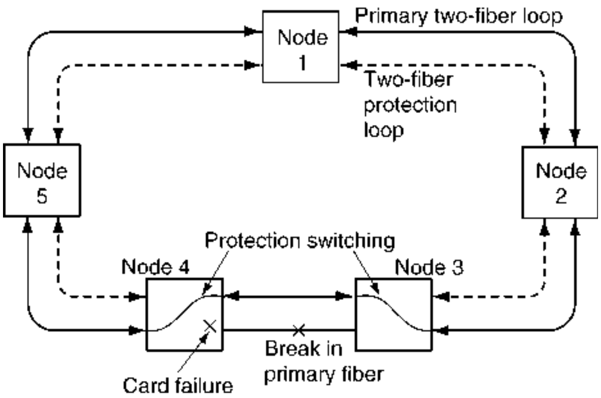
Node 1
Primary two-fiber loop
Two-fiber protection loop
Node 5
Node 2
Protection switching
Node 4
Node 3
Break in primary fiber
Card failure

Figure 5 Reconfiguration of a four-fiber BLSR under transceiver failure.

occur when the primary fiber connecting nodes 3 and 4 breaks. Note that in either case the other links remain unaffected.

Now suppose an entire node fails, or both the primary and the protection fibers in a given span are severed, which could happen if they are in the same cable duct between two nodes. In this case the nodes on either side of the failed span internally switch the primary-path connections from their receivers and transmitters to the protection fibers, in order to loop traffic back to the previous node. This process again forms a closed ring, but now with all of the primary and protection fibers in use around the entire ring, as shown in Figure 6.

Single-Span WDM Links

An interesting and powerful aspect of an optical communication link is that many different wavelengths can be sent along a fiber simultaneously in the 1300–1600 nm spectrum. The technology of combining a number of wavelengths onto the same fiber is known as wavelength division multiplexing or WDM. Figure 7 shows the basic WDM concept. Here *N* independent optically formatted information streams, each transmitted at a different wavelength, are combined with an optical multiplexer and sent over the same fiber. Note that each of these streams could be at a different data rate. Each information stream maintains its individual data rate after being multiplexed with the other streams, and still operates at its unique wavelength. Conceptually, the WDM scheme is the same as frequency division multiplexing (FDM) used in microwave radio and satellite systems.

To see the potential of WDM, consider the characteristics of a high-quality optical source such as a distributed feedback (DFB) laser, which has a very narrow frequency spectrum on the order of 1 MHz, which is equivalent to a spectral linewidth of 10^{-5} nm. When using such a source, a guard band of 0.4–1.6 nm is typically employed. This is done to take into account possible drifts of the peak wavelength due to aging or temperature effects, and to give both the manufacturer and the user some leeway in specifying and choosing the precise peak emission wavelength. With such spectral band widths, simplex systems make use of only a very

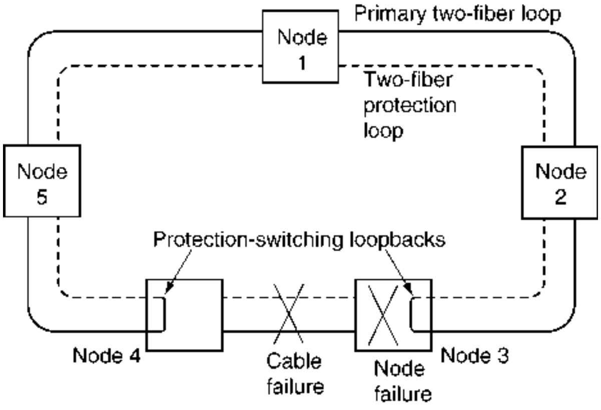
Node 1
Primary two-fiber loop
Two-fiber protection loop
Node 5
Node 2
Protection-switching loopbacks
Node 4
Cable failure
Node failure
Node 3

Figure 6 Reconfiguration of a four-fiber BLSR under node or fiber-cable failure.

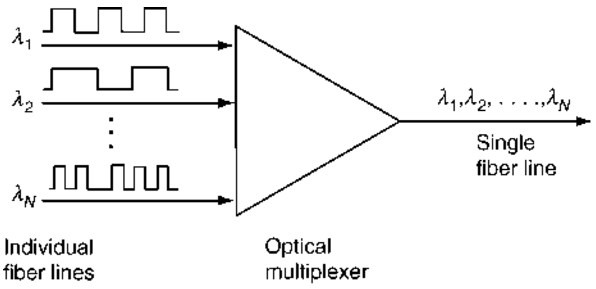
λ_1
λ_2
λ_N
$\lambda_1, \lambda_2, \ldots, \lambda_N$
Single fiber line
Individual fiber lines
Optical multiplexer

Figure 7 The basic WDM concept.

small portion of the transmission bandwidth capability of a fiber. This can be seen from Figure 2, which depicts the attenuation of light in a silica fiber as a function of wavelength. The curve shows that the two low-loss regions of a single-mode fiber extend over the wavelengths ranging from about 1260 to 1360 nm (the O-band) and from 1460 to 1625 nm (the S-, C-, and L-bands).

These regions can be viewed either in terms of spectral width (the wavelength band occupied by the light signal and its guard band) or by means of optical bandwidth (the frequency band occupied by the light signal). To find the optical bandwidth corresponding to a particular spectral width in these regions, consider the fundamental relationship $c = \lambda\nu$, which relates the wavelength λ to the carrier frequency ν, where c is the speed of light. Differentiating this yields $\Delta\nu = (c/\lambda^2)\Delta\lambda$, where the deviation in frequency $\Delta\nu$ corresponds to the wavelength deviation $\Delta\lambda$ around λ. Thus the optical bandwidth is $\Delta\nu = 14$ THz for a usable spectral band $\Delta\lambda = 80$ nm in the O-band. Similarly, $\Delta\nu = 15$ THz for a usable spectral band $\Delta\lambda = 120$ nm in the region spanning the S-, C-, and L-bands.

Since the spectral width of a high-quality source occupies only a narrow optical bandwidth, the two low-loss windows provide many additional operating regions. By using a number of light sources, each emitting at a different peak wavelength that is sufficiently spaced from its neighbor so as not to create interference, the fidelities of the independent messages from each source are maintained for subsequent conversion to electrical signals at the receiving end.

Since WDM is essentially frequency division multiplexing at optical carrier frequencies, the WDM standards developed by the ITU specify channel spacings in terms of frequency. A key reason for selecting a fixed frequency spacing, rather than a constant wavelength spacing, is that when locking a laser to a particular operating mode it is the frequency of the laser that is fixed. The ITU-T Recommendation G.692 specifies selecting the channels from a grid of frequencies referenced to 193.100 THz (1552.524 nm) and spacing them 100 GHz (0.8 nm at 1552 nm) apart. Suggested alternative spacings include 50 GHz and 200 GHz.

Historically the term dense WDM (DWDM) was used somewhat loosely and generally referred to the spacings denoted by Recommendation G.692. In 2002 the ITU-T released an updated standard aimed specifically at DWDM. This is Recommendation G.694.1, which specifies WDM operation in the S-, C-, and L-bands for high-quality, high-rate metro area network (MAN) and wide area network (WAN) services. It calls out for narrow frequency spacings of 100 to 12.5 GHz (or, equivalently, 0.8 to 0.1 nm at 1550 nm).

Alternatively, by using the larger spectral window available in low-water-content fibers, one can relax the spectral tolerances on the optical components and space the individual sources much further apart. This is the basis of coarse WDM (CWDM), which is useful for metropolitan area networks where variations in attenuation between different wavelength channels is not critical over short transmission distances. In 2002 the ITU-T released Recommendation G.694.2, which is aimed specifically at CWDM. The CWDM grid is made up of 18 wavelengths defined within the range 1270–1610 nm (O- through L-bands) spaced by 20 nm with wavelength-drift tolerances of ±2 nm.

A key feature of WDM is that the discrete wavelengths form an orthogonal set of carriers which can be separated, routed, and switched without interfering with each other. This holds as long as the total optical power intensity is kept sufficiently low to prevent nonlinear effects such as stimulated Brillouin scattering and four-wave mixing processes from degrading the link performance.

The implementation of WDM networks requires a variety of passive and/or active devices to combine, distribute, isolate, and amplify optical power at different wavelengths. Passive devices require no external control for their operation, so they are somewhat limited in their application in WDM networks. These components are mainly used to split and combine or tap off optical signals. The performance of active devices can be controlled electronically, thereby providing a large degree of network flexibility. Active WDM components include tunable optical filters, tunable sources, and optical amplifiers.

Passive Optical Networks

Another step towards realizing the full potential of optical fiber transmission capacity is the concept of all-optical WDM networks to extend the versatility of communication networks beyond architectures such as those provided by SONET. These networks can be classified as either broadcast-and-select or wavelength-routing networks. In general, broadcast-and-select techniques employing passive optical stars, buses, or wavelength routers are used for local network applications, whereas active optical components form the basis for constructing wide-area wavelength-routing networks. Broadcast-and-select networks can be categorized as single-hop or multihop networks. Single-hop refers to networks where information transmitted in the form of light reaches its destination without being converted to an electrical form at any intermediate point. On the other hand, intermediate conversion can occur in a multihop network.

Figure 8 shows two alternate physical architectures for a WDM-based local network. Here *N* sets of transmitters and receivers are attached either to a star coupler or a passive bus. Each transmitter sends its information at a different fixed wavelength. All the transmissions from the various nodes are combined in a passive star coupler or coupled onto a bus and the result is sent out to all receivers.

Each receiver sees all wavelengths and uses a tunable filter to select the one wavelength addressed to it. In addition to point-to-point links, this configuration can also support multicast or broadcast services, where one transmitter sends the same information to several nodes. An interesting point is that these passive WDM networks are protocol transparent. This means that different sets of communicating nodes can use different information-exchange rules (protocols) without affecting the other nodes in the network. This is analogous to current time-division-multiplexed telephone lines in which voice, data, or facsimile services are sent in different time slots without interfering with each other.

Although the architectures of single-hop broadcast-and-select networks are fairly simple, there needs to be careful dynamic coordination between the nodes. For example, if the network is set up so that a transmitter sends its information at a unique fixed wavelength, then the destined receiver needs to be informed when a message is being sent to it, so that it can tune its selective filter to that wavelength. Also conflicts need to be resolved in cases where two stations transmitting at different wavelengths want to send information to the same recipient simultaneously.

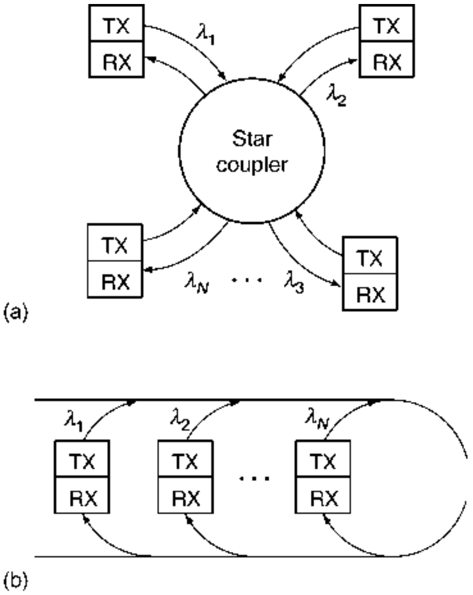

Figure 8 Two alternate physical architectures for a WDM-based local network.

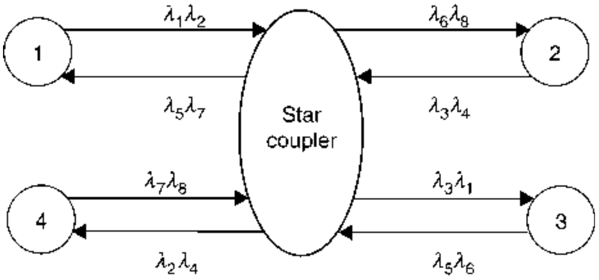

Figure 9 Example of a four-node broadcast-and-select multihop network.

A drawback of single-hop networks is the need for rapidly tunable lasers or receiver optical filters. The designs of multihop networks avoid this need. Multihop networks generally do not have direct paths between each node pair. Each node has a small number of fixed-tuned optical transmitters and receivers. Figure 9 shows an example of a four-node broadcast-and-select multihop network where each node transmits on one set of two fixed wavelengths and receives on another set of two wavelengths. Stations can send information directly only to those nodes that have a receiver tuned to one of the two transmit wavelengths. Information destined for other nodes will have to be routed through intermediate stations.

The flow of traffic can be seen from Figure 9. If node 1 wants to send a message to node 2, it first transmits the message to node 3 using λ_1. Then node 3 forwards the message to node 2 using λ_6. In contrast to single-hop networks, with this scheme there are no destination conflicts or packet collisions in the network, since each wavelength channel is dedicated to a particular source–destination link. However, for *H* hops between nodes, there is a network throughput penalty of at least $1/H$.

Wavelength-Routed Optical Networks

Wavelength-routed networks overcome the limitations of passive optical networks through wavelength reuse, wavelength conversion, and optical switching. The physical topology of a wavelength-routed network consists of optical wavelength routers interconnected by pairs of point-to-point fiber links in an arbitrary mesh configuration, as illustrated in Figure 10. Each link can carry a certain number of wavelengths, which can be directed independently to different output paths at a node. Each node may have logical connections with several other nodes in the network, where each connection uses a particular wavelength. Provided the paths taken by any two connections do not overlap, they can use the same wavelength. Thereby the number of wavelengths is greatly reduced. For example, the connection from node 1 to node 2 and from node 2 to node 3 can both be on λ_1, whereas the connection between nodes 4 and 5 requires a different wavelength (λ_2).

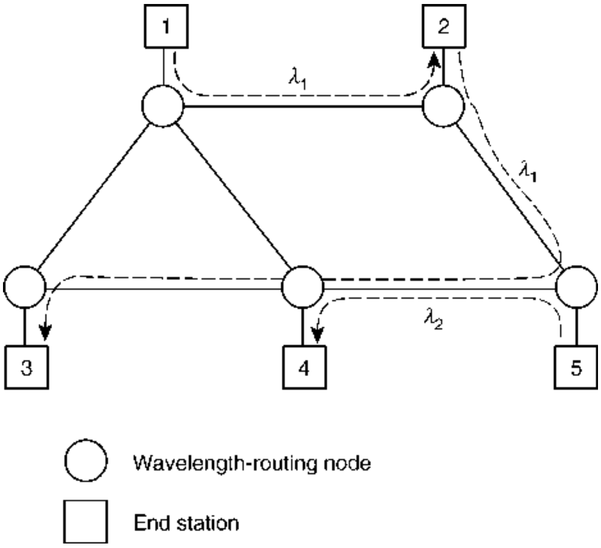

Figure 10 The physical topology of a wavelength-routed network.

A high degree of path modularity, capacity scaling, and flexibility in adding or dropping channels at a user site can be achieved by introducing the concept of an optical cross-connect architecture in the physical path structure (the so-called path layer) of an optical network. These optical cross-connects (OXCs) operate directly in the optical domain and can route very high-capacity WDM data streams over a network of interconnected optical paths.

Contentions could arise when channels having the same wavelength but traveling on different input fibers enter the OXC and need to be switched simultaneously to the same output fiber. This could be resolved by assigning a fixed wavelength to each optical path throughout the network, or by dropping one of the channels and retransmitting it at another wavelength. However, in the first case, wavelength reuse and network scalability (expandability) are reduced, and in the second case the add/drop flexibility of the OXC is lost. These blocking characteristics can be eliminated by using wavelength conversion at any output of the OXC.

As an example consider the 4 × 4 OXC shown in Figure 11. Here two input fibers are each carrying two wavelengths. Either wavelength can be switched to any of the four output ports. The OXC consists of three 2 × 2 switch elements. Suppose that λ_2 on input fiber 1 needs to be switched to output fiber 2 and that λ_1 on input fiber 2 needs to be switched to output fiber 1. This is achieved by having the first two switch elements set in the bar state (the straight-through configuration) and the third element set in the cross state, as indicated in Figure 11. Obviously without wavelength conversion there would be wavelength contention at both output ports.

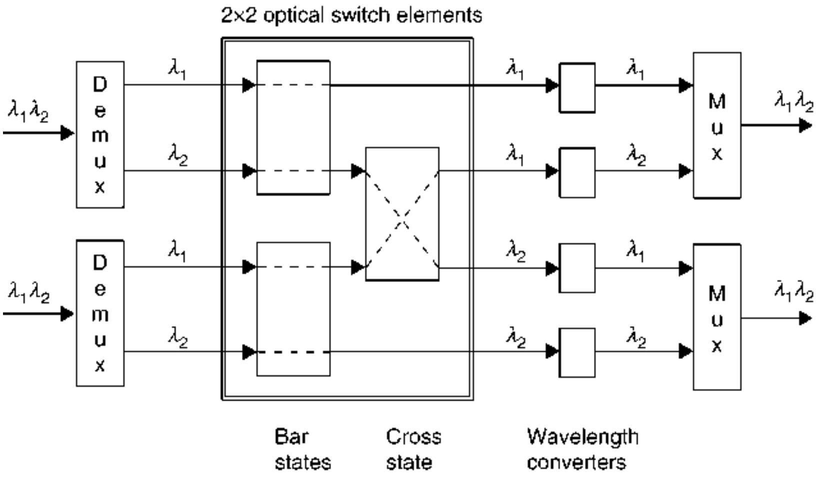

Figure 11 Example of a 4 × 4 optical cross-connect (OXC).

By using wavelength converters, the cross-connected channels can be converted to noncontending wavelengths.

See also

Lasers: Optical Fiber Lasers. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Blumenthal DJ, Prucnal PR and Sauer JR (1994) Photonic packet switches: architectures and experimental implementations. *Proceedings of the Institute of Electrical and Electronics Engineering* 82(11): 1650–1667.

Forouzan BA (2001) *Introduction to Data Communications and Networking*, 2nd edn. Burr Ridge, IL: McGraw-Hill.

Goralski WJ (2002) *SONET/SDH*, 3rd edn. New York: McGraw-Hill.

Gumaste A and Antony T (2003) *DWDM Network Designs and Engineering Solutions*. Upper Saddle River, NJ: Prentice Hall.

Hecht J (2002) *Understanding Fiber Optics*, 4th edn. Upper Saddle River, NJ: Prentice Hall.

Iannone E and Sabella R (1996) Optical path technologies: a comparison among different cross-connect architectures. *Journal of Lightwave Technology* 14(10): 2184–2196.

Keiser G (1999) A review of WDM technology and applications. *Optical Fiber Technology* 5(1): 3–39.

Keiser G (2000) *Optical Fiber Communications*, 3rd edn. Burr Ridge, IL: McGraw-Hill.

Keiser G (2002) *Local Area Networks*, 2nd edn. Burr Ridge, IL: McGraw-Hill.

Keiser G (2003) *Optical Communications Essentials*, New York: McGraw-Hill.

Ramaswami R and Sivarajan KN (2002) *Optical Networks*, 2nd edn. San Francisco: Morgan Kaufmann.

Special Issue (1995) *Undersea Communications Technology. AT&T Technical Journal* 74(1): 4–92.

Trischitta PR and Marra WC (1998) Applying WDM technology to undersea cable networks. *IEEE Communications Magazine* 36(2): 62–66.

Zang H, Jue JP, Sahasrabuddhe L, *et al.* (2001) Dynamic lightpath establishment in wavelength-routed WDM networks. *IEEE Communications Magazine* 39: 100–108.

Free Space Optical Communications

R Martini, Stevens Institute of Technology,
Hoboken, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Free space optical (FSO) communication is the wireless transmission of data via a modulated optical beam directed through free space, without fiber optics or other optical systems guiding the light. The fundamental idea goes back to ancient times, as light (or smoke) signals were used to transmit information. From a more modern point of view, Graham Bell's patent on the photophone may mark the onset of modern FSO techniques, as it transmitted audio signals (i.e., voice) via the modulation of sunlight. A renaissance of FSO systems started with the availability of lasers, light sources with high output power and high coherence, which allowed the accurate direction of the light beams over long distances. During the 1970s and 1980s the main proposed application of FSO systems was for secure and long distance (50–1000 km) communication, mainly targeted for ground–satellite or satellite–satellite communication. This focus changed drastically over the last decade as a new market for FSO grew in the establishment of high bandwidth data link and their integration over a locally restricted area.

The main competitors in this market are the fiber-based optical network, the RF communication system, as well as the low bandwidth copper cable-based system. In comparison to the closely related wireless radio frequency transmission, the higher frequency of the optical carrier ($\sim 10^{14}$–10^{15} Hz) thereby allows for much higher transmission rates, comparable to those of typical fiber optic networks. On the other hand, the use of an optical carrier also results in much more directed beam propagation, which requires an undisturbed line-of-sight between emitter and detector. This restricts the application of most FSO systems to a range between a few hundred meters up to several kilometers, which are still favorable for distribution of high bandwidth networks over a locally restricted area. This makes FSO a highly attractive candidate for the 'last-mile' distribution of high-bandwidth Ethernet to the individual homes.

The simplicity of setup of FSO links, as well as their modularity, is thereby their biggest advantage, compared to fiber-based networks. It makes them not only highly cost efficient, but it eases the maintenance or allows for fast and easy upgrades, as it does not require any extensive and time-consuming installation – in contrast to the installation of fiber optic cables. This advantage makes FSO highly attractive for temporary installations (emergency or short-time high broadcast situations, such as the Olympic Games), as well as to overcome geometrical restraints (river, seas, etc.). As most detector and emitter systems are typically based on the same electro-optical components used for fiber optic networks, they seamlessly integrate and expand an existing network without complications in data handling.

Standard FSO System

Figure 1 shows schematically the setup for a typical FSO system connecting two separate local networks. It consists of: an electronic data input; a small but powerful light source which can be modulated; emitter optics which shape the emitted beam to a highly directed beam; the atmosphere as transmitting device; detector optics which receive the transmitted light and focus it onto a photodetector; and an electronic amplifier which serves as data output.

Compared to a fiber optical communication system, the main difference is the use of the atmosphere as the transmission medium, in contrast to guided optical propagation. The transmission properties are thereby determined by the atmospheric conditions, which can lead to degradation and/or redirection of the beam. The main influences which have to be taken into account are atmospheric scattering, molecular absorption, turbulence effects (including beam wander and fading), as well as the impact of low-visibility weather situations like rain, snow, or fog. These depend strongly on wavelength and linewidth of the used emitter system, distance between emitter and receiver, as well as the environmental conditions. As a consequence, the characteristics of the atmospheric channel determine the layout of the other FSO components – and have to be considered first.

Atmospheric Losses

The overall losses of the atmospheric transmission are determined by the Beer–Lambert absorption law:

$$I = I_0 \exp[-\alpha L] \qquad [1]$$

where I is the intensity at the receiver, I_0 the intensity of the transmission beam after the emitter optics, and L is the range (i.e., distance between emitter and detector optics). The total extinction coefficient $\alpha = \alpha_{\text{Rayleigh}} + \alpha_{\text{Mie}} + \alpha_{\text{Absorption}}$ is based on three different contributions: Rayleigh scattering, Mie scattering, and absorption.

In contrast to absorption, the losses due to scattering originate from redirection of light and will be discussed first. Depending on the size of the scattering partner, two separate regimes have to be taken into account. For scatter smaller than the wavelength, the effect can be described based on Rayleigh's classical formula. The scattering cross-section of an element with mass m, charge q, natural oscillation wavelength λ_0, is then:

$$\sigma_s = \frac{f q^4 \lambda_0^4}{6\pi\varepsilon_0^2 m^2 c^4} \frac{1}{\lambda^4} \qquad [2]$$

where ε_0 is the dielectric constant, c the speed of light, and λ the wavelength of the incident light, leading to an extinction coefficient of $\alpha_{\text{Rayleigh}} = \sigma_s N_s$, where N_s is the concentration of the scattering elements. The λ^{-4} dependence in the scattering predicts stronger scattering for shorter wavelengths and Rayleigh scattering is therefore dominant for short and visible wavelengths (which gives the sky its blue color). For typical FSO

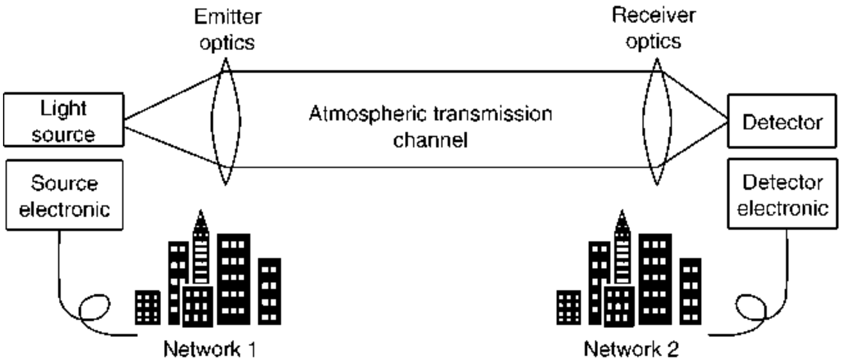

Figure 1 Typical components of a FSO system.

wavelengths in the region of near-infrared radiation (NIR, 800–2000 nm) or mid-infrared radiation (MIR, 2000–20 000 nm), the contribution of Rayleigh scattering is typically relatively low and can be neglected.

On the other hand, Mie scattering is, in most cases, the strongest contribution for FSO systems – but also the hardest to predict. It is due to resonant scattering effects, where the particle size is of the wavelength of the scattered light. For the typical FSO wavelengths these particles include water and oil droplets, dust, smog, fly ash, and aerosols (which can originate from pollution). A correct calculation of the Mie scattering coefficient is possible, but requires knowledge of the exact particle concentration as well as their size distribution. As these numbers are not only difficult to measure, but also fluctuate continuously depending on weather situation, time of day, season, temperature, climate, pollution, and other factors, a numerically correct evaluation is not possible.

Historically, most systems rely on an empirical formula, which estimates the attenuation coefficient to

$$\sigma_{\mathrm{Mie}} = \frac{3.91}{V}\left(\frac{\lambda}{550}\right)^{-p} \qquad [3]$$

where V denotes the visibility and the value of p ranges from 1.6 for high visibility ($V > 50$ km) and 1.3 for average visibility (50 km $> V >$ 6 km) down to 0.585 $V_{1/3}$, respectively for low visibility. The formula is based on data collected in the 1950s and 1960s. More accurate measurements show that the actual scattering cross-sections differ quite strongly, especially for low-visibility cases. Especially, the predicted dependence on wavelength is not correct. The failure of such a general model is not surprising, as the actual scattering properties are strongly dependent on the weather situation and the environmental conditions. It is the focus of ongoing research to find more accurate models describing the scattering condition in multiple environments and conditions to provide the background for more accurate empirical models.

Apart from the light redirection due to scattering, absorption either by molecules or aerosols is nearly as important for the transmission calculation of a FSO system. The strongest molecular absorptions are typically due to water, carbon dioxide, as well as ozone. A typical absorption spectrum can be easily calculated using standard software together with the publicly available HITRAN database, both available from ONTAR. The upper part of Figure 2 shows a low-resolution transmission spectrum for the interesting wavelength region for a path length of 1 km.

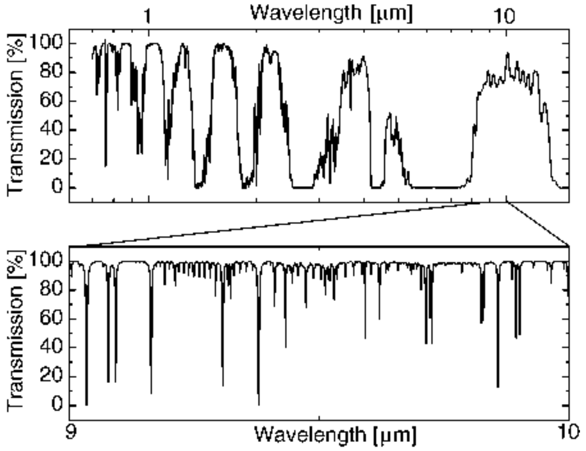

Figure 2 Upper: Low-resolution transmission spectrum of the atmosphere showing molecular absorption for a path length of 1 km. Lower: high-resolution spectrum of the atmospheric windows around 9.5 µm.

Within this spectrum, several well-known atmospheric windows exist, which are naturally preferred for FSO applications. These include the region around 830 nm, 1550 nm, and from 8 µm to 12 µm. For the exact selection of a wavelength, a closer look is needed. As an example, a high-resolution analysis of the atmospheric windows around 9.5 µm is shown in the lower part of Figure 2, which identifies additional absorption lines as well as highly transparent wavelength ranges within the window. Such substructure has to be taken carefully into account for the FSO wavelength selection, especially if lasers are employed, which have a linewidth comparable to the width of the absorption line.

Besides the molecules, aerosol particles also contribute to the absorption spectrum. Calculation of their contribution can also be performed using available software package, such as Air Force's MODTRAN and LOWTRAN program as well as OPAC from the LMU University in Munich. Figure 3 shows a typical low-resolution transmission spectrum for 50% humidity as well as 99% humidity, where by scattering as well as absorption effects have both been included in the calculation. The occurrence of sharp absorption features, as well as the increasing losses for wavelengths above 10 µm, is typically due to the aerosol absorption, whereas the scattering dominates the function for wavelengths below 3 µm. Losses due to the aforementioned Rayleigh scattering are also shown in the same diagram, demonstrating its low influence at longer wavelengths.

MODTRAN and LOWTRAN also allow the calculation of transmission spectra under rain and snow conditions, where rain and snow consist of

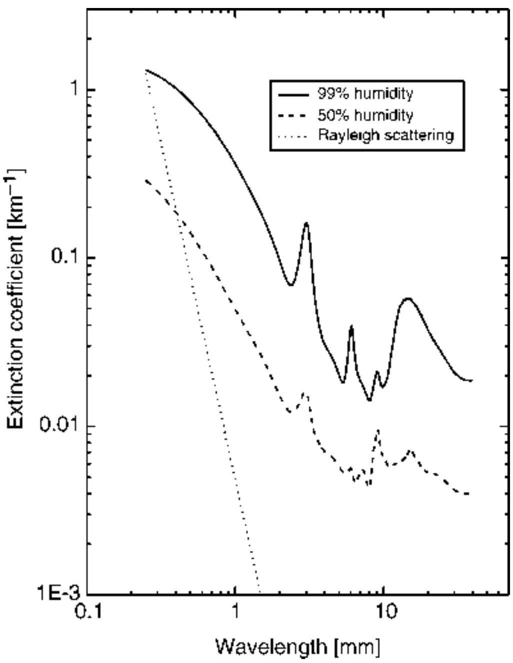

Figure 3 Typical low-resolution transmission spectrum calculated based on absorption and scattering losses from aerosols for 50% humidity (dashed line) and 99% humidity (solid line). The dotted line shows additionally the losses due to Rayleigh scattering.

clusters of water and aerosol with diameters from 200 μm to 2 mm and above. Depending on the severity of the precipitation, the losses can range from ~0.5 dB/km for light rain up to 13 dB/km for cloudbursts, requiring an additional power margin in the preparation of the FSO link. Significantly stronger is the impact of fog, as typically particles with much lower sizes are present. The losses inflicted by a moderate fog (visibility 500 m) range above 20 dB/km, surpassing the effect of a cloudburst; for thick fog the attenuation can even grow beyond 200 dB/km. However, depending on the size distribution within the fog itself, the losses can be highly wavelength dependent. Recent experiments show that long wavelength FSO systems operating in the MIR wavelength region have a much stronger stability under the influence of thick fog than one operating at NIR wavelengths.

Turbulence and Scintillation

In contrast to a static medium, the atmosphere is a dynamic mixture which undergoes continuous changes due to wind, temperature, and other influences. This creates a constant change in the parts of the atmosphere which are contributing to the FSO channel and therefore in the optical properties of the beam path. The index of refraction can differ, not only in its total value, but also within subsections of it, creating a random fluctuation in the beam paths. Such localized changes in the refractive index of the air are schematically shown in Figure 4, where turbulences of different sizes modify the beam path and disturb the beam front. The turbulence in the atmosphere can be quantified by the refractive-turbulence structure coefficient C_n^2, which can range from $10^{-13}\,\mathrm{m}^{-2/3}$ in the middle of the afternoon (highest value) to $10^{-17}\,\mathrm{m}^{-2/3}$ one hour after sunrise or sunset. As the turbulence effect is typically due to strong temperature difference, C_n^2 is strongest near hot surfaces (e.g., tar roofs or asphalt in the summer) and decreases with altitude. The effects of such turbulences on a directed light beam (such as used in FSO systems) include mainly beam wander and scintillation effects.

Beam wander is thereby based on large size turbulences which act as a lens on the propagating beam. Consequently, a narrow focused beam could be randomly steered away from the detector area, creating a fluctuation in the detected intensity.

Stronger impact for FSO links has scintillation, which is better known as the twinkling of a star or the heat shimmer of the horizon on a hot day. It is based on the continuous variation of the beam front which can lead to intensity fluctuations (like the twinkling) and loss in the image resolution (heat shimmer). Strong intensity fluctuations can thereby cause signal loss as well as receiver saturation, both resulting in downtime in the link. Such local intensity disturbances can occur up to sizes of $\sqrt{\lambda L}$, reaching, for example, 4 cm for a 1.55 μm link after 1 km. For small fluctuations, the intensity follows a classical stochastic behavior and can be described by the irradiance variation (normalized to its mean value) as

$$\sigma^2 = \exp\left[0.5\left(\frac{2\pi}{\lambda}\right)^{6/7} L^{11/6} C_n^2\right] - 1 \qquad [4]$$

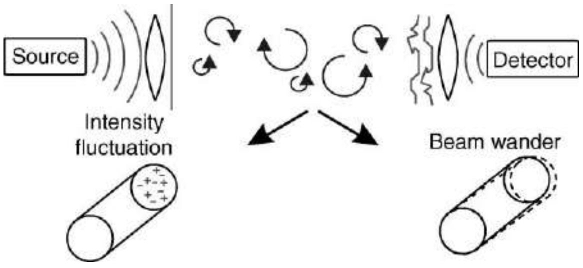

Figure 4 Schematic depiction of atmospheric turbulences with different sizes in FSO systems. Shown below are illustrations of its main effects on a directed beam causing intensity fluctuations and beam wander.

with usual values in the range from 5 to 70%. Nonetheless, it should be pointed out that actual turbulence effects can differ quite strongly from these predictions, due to local inhomogeneities in temperature distribution. Consequently, test measurements are necessary for the correct evaluation of a specific installation site.

What makes scintillation effects difficult to handle for communications systems is the relatively slow frequency of fluctuations in the order of a few hundred Hz or less. The corresponding downtimes for a link can range from 1 ms to 10 ms, or even up to seconds. This is magnitudes higher than typical downtimes in standard communications links and cannot be compensated by standard forward error correction systems. One possibility to reduce the influence of scintillation is the method of aperture averaging, whereby multiple beams originating from slightly different positions are deployed and targeted at the same receiver. Assuming that the variation in each of the separate beam paths is random and uncorrelated, the signal-to-noise ratio will decrease by the square root of the number of beams. Recent experiments have impressively demonstrated this effect and allow reduction of intensity fluctuation to an acceptable level.

Another alternative to decreasing the scintillation effects is in the application of longer wavelength FSO systems. Whereas beam wander is relatively independent from the wavelength, scintillations effects scale with $\lambda^{7/6}$, as turbulences smaller than the wavelength cannot obscure its beam front. Using a 2 times longer wavelength, for example, should have the same impact as signal averaging with 4 different beams.

Beam Spreading and Pointing

Based on optical diffraction theory, the beam leaving the emitter optics will be divergent and its beam diameter will consistently increase during transmission through the optical channel. Assuming an optimal alignment and using standard geometrical optics, the losses due to beam spreading can estimate the ratio of emitted power (P_E) to received power (P_R):

$$R_{BS} = \frac{P_R}{P_E} = \frac{D_R}{(D_E + L\theta)} \qquad [5]$$

where D_R is the diameter of the receiving optics, D_E the diameter of the emitting optics, and θ the divergence of the beam. The minimal divergence of the beam is typically determined by the light source and is given by $\theta_L = \lambda/D_L M^2$, where D_L is the initial size of the laser beam and M^2 the laser-mode structure parameter, for a typical semiconductor laser θ_L below 0.1 mrad. In many cases, the divergence of the emitted beam is commonly enlarged by slight defocusing of the emitting optics – if the power budget allows for the additional losses. The beam size at the receiver is typically several times larger than the receiving aperture, which not only eases the alignment of the system, but also reduces pointing losses. This includes the effect of beam wander due to turbulence as well as swaying of the buildings or the mount. If those effects are solely compensated for by an increased beam spread, divergence can easily reach values of 2–20 mrad for short-distance links. Obviously this is not an option for longer-distance FSO links, as the losses will increase dramatically. In these cases, active tracking systems are incorporated in the emitter optics allowing for continuously adjustment to achieve and keep an optimal alignment. Typical tracking devices are thereby based on servo-driven Gimbal mounts, steering mirrors, or recently, MEMS devices.

Laser Eye Safety

As the atmospheric transmission channel of most FSO systems is not completely isolated, there exists the possibility of a human person intercepting the beam. To allow for this, most FSO systems are designed to be eye-safe following the voluntary ANSI standard, which limits the output irradiance to a certain value depending on the deployed wavelength. As radiation below 1400 nm is still being focused on the retina, it causes a higher irradiance, which results in a lower damage threshold. For 800 nm laser systems, the minimum permissible exposure (MPE) level for a 10 s long exposure is therefore recommended not to exceed 1 mW cm^{-2}, whereas for 1.55 μm or longer wavelength systems, an MPE of 100 mW cm^{-2} is suggested. The recommendations for incoherent light sources such as LEDs are thereby less stringent as they cannot be focused quite as strongly as laser light.

The Light Source

Following the discussion above, the optimal light source for FSO system should be emitting at a wavelength away from any absorption line (or should be much broader then the absorption line), have a relatively long wavelength as well as a relative low divergence, be eye-safe, and allow for high modulation speed. In reality, most FSO systems use either GaAsAl laser diodes or LEDs emitting at 0.8 μm, or the typically more expensive 1.5 μm laser diodes based on the InGaAsP semiconductor system. The output power of both laser systems is typically in the range of 1 mW to 100 mW, which for the longer wavelength case can be boosted to 1–10 W using erbium-doped fiber amplifiers (or even higher for specific applications). The longer wavelength system has the additional advantage of higher ANSI standards as discussed above, allowing stronger eye-safe emission power, but is typically far more expensive.

Both laser types were developed and optimized for standard fiber-based communication systems allowing direct modulation up to 10 Gbit/s. To achieve higher bandwidth in FSO systems, standard techniques known from fiber-based networks can easily be deployed, like wavelength division multiplexing which increases the bandwidth up to state-of-the-art levels of fiber-based networks.

Contrary to semiconductor lasers, LEDs provide a much less expensive alternative, but allow only for modulation up to 10–100 MHz. Their broad emission spectrum is thereby an advantage, as it is less sensitive to sharp absorption lines which – as discussed earlier – can have a drastic impact on the FSO link performance. Consequently, LEDs have been successfully implemented in several commercial FSO systems targeted at the short-range market.

Recently developed sources, which have also been applied for FSO systems, include vertical-cavity surface-emitting lasers as well as quantum cascade lasers (QCL) which emit in the mid-infrared regime. The latter devices have been predicted to have a strong impact on the FSO market, as they combine high-output power (up to 1 W) together with fundamentally high modulation bandwidth (2.5 Gbit/s recently demonstrated and higher rates predicted) at wavelengths ranging from 3.5 μm to 12 μm and beyond. They can access the long-wavelength atmospheric window atmospheric for the FSO systems (typically reserved for CO_2 lasers) with the convenience of direct semiconductor lasers like direct electrical modulation, small size, and low power consumption. Recently, first FSO experiments, employing experimental QCLs, demonstrated the predicted higher stability of the link in low-visibility situations, including strong fog as well as high dust concentration. Nonetheless, to date the high cooling requirement for QCLs restricts more general use of them, but their very fast development, recent reports of cw-operation of room-temperature QCLs as well as FSO links with TE-cooled QCLs in pulse operation, give hope for a solution to this problem.

The Detector System

Once the light reaches the receiver, it is typically collimated using a short focal length telescope or a single lens system. To optimize efficiency without sacrificing ease of alignment a compromise has to be made between a wide aperture for greater light collection and a short focal length. The concentrated light then hits the detector, which, depending on the wavelength, is either Si or InGaAs-photodiodes. Si-based devices can detect up to wavelengths of 1100 nm and InGaAs systems are typically used for the 1.5 μm region, respectively. To enhance the sensitivity of the detectors, avalanche photodiodes are preferred which internally enhance the detected signal by a factor ranging from 20 to 100. To ensure high bandwidth operation (10 MHz–10 GHz), the size of the detecting device has to be kept relatively small (20 to 100 μm) avoiding parasitical capacitance effects – but raising the requirement on the detecting optics.

The sensitivity of these NIR detectors is typically limited by the shot-noise, which is due to statistical fluctuations of the received phonons and scales with the square root of its number. For practical purposes the internal noise of the detector is quantified using the corresponding amount of incident optical power leading to a comparable signal. The values for this noise equivalent power (NEP) can range from microwatts to tens of nanowatts. Other sources for detector noise can include a background current from the detector–amplifier combination, typically referred to as Johnson noise.

Mid-infrared detectors are additionally objected to strong background radiation due to thermal emission, which peaks around the 10 μm at temperature of 300 K. The use of additional cold filters and high-resolution spectral filters can help to overcome this restriction, but still the sensitivity of those detectors is lacking compared to NIR systems.

For communication system the noise is more commonly quantified as bit error rate (BER), quantifying the average percentage of wrongly received bits (submitted ‘1’, detected ‘0’ or vice versa). A BER greater than 10^{-9} is usually referred to as an error-free system, as the residual errors can be corrected employing forward error correcting algorithms. In contrast, the performance of FSO systems can more easily be measured in its signal-to-noise ratio (SNR). Following standard communication theory, a peak SNR of 12 and an average SNR of 6 is required to achieve a BER of 10^{-9}, guaranteeing error-free communication. Consequently range limits for FSO links are typically based on SNR = 6.

Link Budget and FSO Range Equation

Classical link budget estimation is a commonly used tool to access the performance of FSO systems. To take into account the losses due to atmospheric absorption, scattering, turbulence, and beam spreading, as well as optical losses in the used receiver and emitter optics, they are deducted from the starting power emitted from the FSO light source. The estimated intensity at the detector is then compared with the intensity needed to establish error-free communication (for example $\mathrm{SNR} = 6$), the calculated difference is the link margin of the specific system, i.e., the amount of additional losses under which the FSO link will still be working.

Comparing this number to losses from nonpredictable fluctuations (like heavy rain or fog) gives an insight under which implications the link will fail. Using weather station data, the frequency of such extremely low-visibility occurrences can be obtained and a number for the reliability of the FSO link be estimated. Typical well-installed FSO systems estimate very high reliabilities, between 99.9% and 99.999%, which have been obtained in installation and verified in long-term measurements.

Another way to estimate the performance for a given system is based on the FSO range equation, where the received power on the detector (P_R) is calculated to

$$P_\mathrm{R} = P_\mathrm{E} \frac{D_\mathrm{R}}{(D_\mathrm{E} + L\theta)^2} TK \exp[-\alpha L] \qquad [6]$$

T denotes the optical losses of the transmission and receiver system, and K denotes additional losses for incoherent lights sources and is given by the ratio of LED emitting area to detector area, and is $K = 1$ for laser light sources. This equation can also be used to estimate losses for different weather situations and obtain the maximum range L of a specific FSO system.

Recent Research and Future Considerations

Existing FSO systems have impressively demonstrated their versatility and stability, as well as their seamless integration into existing networks, on multiple occasions. A prime example is the re-establishment of communication links between Merrill–Lynch Brokerage and Wall Street in New York, in the aftermath of the terrorist attack on September 11th, which destroyed one of the central fiber optic hubs in downtown Manhattan. The installation of a commercial FSO system by Terabeam ensured a fast replacement and a stable continuation of their work.

The demonstrated performance of such links compares to standard fiber-based links. The inclusion of wavelength division multiplexing as well as other standard or nonstandard techniques in FSO systems proves the strengths of a system, which is based on identical hardware for emitter and detectors as fiber-based networks. Nonetheless, FSO allows a multitude of applications not possible in fibers, which current research projects are targeted at. This includes, for example, long wavelength FSO communication systems, continuously ground-satellite links and free-space quantum communication for extreme secure data transmission – to list only a few of the possibilities. One very obvious example is extreme wavelength division multiplexing, which uses one link operating in all atmospheric windows at the same time. Such a FSO system would not only allow very high data rates and extreme stability against weather effects, but could also employ new kinds of wavelength-based modulation and encoding schemes increasing the security of the data link. Concluding the actual state of the FSO system as well as ongoing research, free space optical communication has all the ingredients needed to be one of the mayor key ingredients of future networks – and based on the increasing demand of bandwidths they will contribute strongly.

List of Units and Nomenclature

c	speed of light
C_n	refractive-turbulence structure coefficient
D_E	diameter of the emitting optics
D_R	diameter of the receiving optics
I	intensity
L	range (i.e., distance between emitter and detector optics)
m	mass
M^2	laser-mode structure parameter
N_s	concentration
P_E	emitted power
P_R	received power
q	charge
T	optical losses of the transmission and receiver system
V	visibility
α	extinction coefficient
ε_0	dielectric constant
θ	divergence of the light beam
λ	wavelength
λ_0	natural oscillation wavelength
σ_s	scattering cross-section

See also

Environmental Measurements: Doppler Lidar; Optical Transmission and Scatter of the Atmosphere. **Imaging:** Infrared Imaging; Lidar.

Further Reading

Andrews LC and Phillips RL (1998) *Laser Beam Propagation through Random Media*. Bellingham, CA: SPIE Optical Engineering Press.

Begley DL (ed.) (1991) *Selected Papers on Free-Space Laser Communications*. SPIE Milestone Series. Bellingham, CA: SPIE Optical Engineering Press.

IEE Proceedings Optoelectronics, *Special Issue on Free Space Optical Communications*. December 1996.

Jha AR (2000) *Infrared Technologies*. New York: John Wiley & Sons.

Killinger D (2002) *Free Space Optics for Laser Communications Through the Air*. Washington: Optics and Photonics News, Optical Society of America, p. 36.

Measures R (1984) *Laser Remote Sensing*. New York: John Wiley & Sons.

Mecherle GS (ed.) (2003) *Free-Space Laser Communication Technologies XV*. Bellingham, CA: SPIE Optical Engineering Press.

Weichl H (1990) *Laser Beam Propagation in the Atmosphere*. Bellingham, CA: SPIE Optical Engineering Press.

Willebrand H and Ghuman BS (2001) *Free-Space Optics: Enabling Optical Connectivity in Today's Networks*. Indianapolis, IN: SAMS.

Wolfe WL and Zissis GJ (1978) *The Infrared Handbook*. Ann Arbor: Enviromental Research Institute of Michigan.

Zuev VE (1988) *Laser Beams in the Atmosphere*. New York: Consultant Bureau.

Lightwave Transmitters

J G McInerney, National University of Ireland-Cork, Cork, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical fiber telecommunications has changed human society forever, providing the capacity for affordable and ubiquitous communication. It provides data rates and transport economies far in excess of those available using purely electronic means. It is doubtful whether the Internet and worldwide web, or the vast infrastructure of wireless cellular telephony, would be viable without optical fiber technology. Its rapid development and acceptance has been driven largely by contemporaneous successes in manufacturable low-loss optical fiber cables, sensitive pin-FET photoreceivers, and reliable high-performance laser diode-based transmitters. The earliest developments in silica glass fibers at Standard Telecommunication Laboratories, UK, (then part of the US-based ITT Corporation) and Corning Glass Works (USA) in the 1960s and 1970s, were pivotal in determining the directions of modern fiber technology, as were later key inventions of rare-earth doped fiber amplifiers at the University of Southampton (UK) and AT&T Bell Laboratories (USA) in the 1980s and early 1990s. However, perhaps the greatest contributors to the success of fiber optics have been semiconductor injection lasers, or laser diodes, which in various forms have made up the vast majority of all fiber optic transmitters, and the entirety of long-haul transmitters, as well as pump sources for the doped-fiber amplifiers which enable modern networks. The fascinating story of the development of fiber optics is to a large extent driven by the dramatic progress in the development of high-performance, reliable semiconductor lasers. These emitted at wavelengths near 1300 and 1550 nm, which respectively correspond to the dispersion and attenuation minima of single-mode silica optical fibers.

Semiconductor Laser Principles

Semiconductor lasers were first demonstrated in research laboratories at General Electric, IBM Corporation and the MIT Lincoln Laboratory (all USA) as early as 1962, although it took almost two decades for the basic science and engineering of materials, fabrication, reliability, and performance design issues to be developed sufficiently for their use in practical communication systems. Early successes were obtained using the GaAs/AlGaAs lattice-matched material system, in which a lightly doped GaAs or ternary compound AlGaAs active layer is sandwiched between two lattice-matched heterojunctions with n- and p-doped AlGaAs layers with higher Al fractions than the active layer, so that their refractive indices are lower and their bandgaps higher, providing optical and charge-carrier confinement. In all cases the material is epitaxially grown, meaning that the entire laser chip forms a single crystal. The gain, and hence the laser emission, occurs at the direct bandgap of the active layer, at wavelengths in the range 750–870 nm depending on the Al fraction.

Later evolutions used the quaternary compound GaInAsP clad by InP guiding layers, where the In and P fractions are chosen to match the lattice constant to the InP substrate.

In general, optical gain in a laser diode is generated by creating an electron-hole plasma in the vicinity of the forward-biased junction, a situation corresponding to population inversion in a conventional laser, and with sufficient forward current the material becomes transparent, in that the gain exactly equals the absorption and scattering losses, at a particular wavelength. When this arrangement is enclosed within a suitable optical cavity, and the forward current is increased further, the system begins to lase, that is to oscillate continuously to produce coherent optical radiation at a wavelength which satisfies two basic conditions: it must be within the set of wavelengths at which the material produces sufficient gain, and it must be close to one of the electromagnetic modes at which the optical cavity is resonant. These conditions, which define lasing threshold, are characterized by the equality of the optical gain and the total losses, that is the material losses (absorption, scattering, and free-carrier plasma) and the cavity losses which include the light output.

Early laser diodes in both the GaAs/AlGaAs and GaInAsP/InP systems were of the edge-emitting type, in which the light is emitted from the edge of the laser chip, perpendicular to the growth direction in the plane of the wafer. Although this geometry complicates laser production flow, particularly the testing function which requires that the wafer be scribed into bars to define the laser output facets, edge-emitting lasers (EELs) make up the vast majority of optical fiber transmission sources. An alternative laser diode structure is the vertical cavity surface-emitting laser (VCSEL), in which emission occurs in the growth direction perpendicular to the plane of the wafer. VCSELs offer major advantages in spectral stability, beam quality, manufacturability and cost, but their output powers are low (a few mW) and they are not yet available at the key telecommunication windows around 1300 nm and 1550 nm, due to difficulty in forming the necessary high-reflectivity Bragg mirrors in the GaInAsP/InP material system. Major initiatives are currently underway to produce long-wave VCSELs in the GaAs/AlGaAs system using InAs quantum dots, or dilute nitrides such as InGaAsN, or hybrid approaches such as InP-based gain media fused to high-reflectivity mirrors using GaAs/AlGaAs, dielectric coatings, air gaps, and others. At present, however, all long-haul optical fiber transmission is based on edge-emitting laser diodes fabricated in GaInAsP/InP.

General Structure and Requirements of Optical Fiber Communication Systems

The first optical fiber communication links were simple point-to-point affairs, consisting essentially of series-connected transmitters, fibers, and receivers. The transmitters were simple Fabry–Perot laser diodes, onto which data were encoded by direct digital modulation of the injection current at rates ~100 Mbit/s. The first commercial optical fiber link was built in 1976, a single fiber cable linking two switching centers of the Illinois Bell Telephone Co just 2.5 km apart in the Chicago metropolitan area, using 850 nm GaAs laser technology. In 1988, the first transatlantic fiber link, AT&T's TAT-8, was completed linking endpoints in New Jersey, England, and France with three fiber pairs carrying signals generated by 1300 nm GaInAsp single mode lasers. In these early systems, data were regenerated frequently by repeaters each consisting of a receiver, signal conditioner and transmitter. Each fiber in TAT-8 carried a single optical frequency modulated at 280 Mbit/s and the construction cost of the system was ~$50 k/km, resulting in an economic figure of merit of over $600 k/Mbit/s. Twenty-five years later, current systems cost the same per kilometer to build in absolute dollars despite inflation, a major effective cost reduction enabled largely by replacement of the expensive repeaters by doped-fiber optical amplifiers. Moreover, each fiber is now highly multiplexed, with the potential for hundreds of wavelengths each carrying 10 Gbit/s signals generated by 1550 nm laser transmitters, so that the unit cost has plummeted to ~$200/Mbit/s, a reduction of three and a half orders of magnitude.

In the remainder of this article, we will describe transmitters for both short- and long-haul systems where the channel spacing is only tens of GHz in the optical carrier frequency. Modern short-haul systems either feed client networks as part of bi-directional transceivers, or drive metro and local access networks connected directly by photonic switches to terminals of the long-haul systems.

DWDM Transceivers

Modern practice uses modular units containing transmitters and receivers allowing bi-directional data flow. In the send mode, data are applied to a long-haul transmitter at the local (client) source and coupled through the external fiber network to the remote long-haul receiver. In the receive mode, data arrive from a remote long-haul transmitter and are coupled (via a regenerator including signal conditioning, clock recovery, and error correcting steps) to a local short-haul transmitter leading to the client's internal network. Each transceiver card therefore contains a long-haul transmitter and receiver set (known collectively as dense wavelength division multiplexing (DWDM) transport devices) and a short-haul transmitter and receiver set (known as client interfaces or local transport devices).

Transmitter Requirements

Short-haul transmitters may be simple Fabry–Perot edge-emitting laser diodes, or vertical cavity laser diodes, or even superluminescent LEDs for sub-Gbit/s data rates. They are generally not required to be at one of the minimum-loss telecom windows near 1550 nm, as the transmission distances are short. However, as client networks grow into large metropolitan networks, it is likely that 1300 nm sources will be used to minimize dispersion. Power requirements are modest, a few mW peak, at data rates ~1–10 Gbit/s. Single mode laser emission and polarization control are not required by present day systems, although polarization-mode dispersion may limit transmission at higher data rates. Both return-to-zero (RZ) and nonreturn-to-zero (NRZ) coding may be used, and direct modulation of the laser amplitude is the norm.

For long-haul transmitters the situation is dramatically different. Operation in one of the standard erbium-doped fiber amplifier (EDFA) bands near 1550 nm is mandatory: the C (conventional) band extends from 1530–1565 nm, the S (short-wave) band from 1460–1530 nm) and the L (long-wave) band from 1565–1625 nm. In the C-band, output powers ~10 mW are generally sufficient, and powers above ~30 mW are limited by nonlinear effects such as self-phase modulation and four-wave mixing in the fiber. In the S- and L-bands, where the available gain from EDFAs is less, lasers may need to operate close to the nonlinear limit, several tens of mW, although various schemes (differential pumping, gain-flattening filters, etc.) are being evaluated to flatten the global EDFA gain spectrum.

In current practice, a single fiber utilizing DWDM, using only the C-band, can carry up to 1 Tbit/s comprising 100 channels at 10 Gbit/s on each channel, the current standard. For such performance a channel spacing of 50 GHz in optical frequency is required. Using all three bands with similar density would enable ~3 Tbit/s, which could be doubled by using 25 GHz channel spacings. Ultimately, even higher channel densities and hence overall system data rates are possible, with a practical limit ~10 Tbit/s. The channel spacing is a key parameter in that it determines the laser design, specifically its spectral stability and linewidth, and the overall transmitter design in that laser chirp (modulation induced dynamic spectral shift) must be less than half the channel spacing. In practice, only single-wavelength and coarse wavelength division multiplexing (CWDM) systems may be modulated directly. All modern DWDM systems require external modulation, for example using electro-absorption or Mach–Zehnder interferometric devices outside the laser cavity. For the latter cases, stabilization of the laser wavelength is required for example by on-chip gratings to form distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, or by external fiber gratings.

Directly Modulated Lasers

Direct modulation of semiconductor lasers is convenient and effective: by modulating the laser injection current, one modulates the carrier density and hence the optical gain, resulting in modulation of the laser output up to ~10 Gbit/s. The actual modulation bandwidth is determined by interactions between photons and carriers, whose respective decay lifetimes, τ_s and τ_p, are determined by cavity losses and total recombination rates. The simplest form of such interactions is described by the photon and carrier rate equations for the injected carrier density N and photon density S in a single lasing mode:

$$\mathrm{d}N/\mathrm{d}t = J/ed - gNS - N/\tau_s \qquad [1]$$

$$\mathrm{d}P/\mathrm{d}t = gNS + \beta N/\tau_s - S/\tau_p \qquad [2]$$

where J is the injected current density, e the electronic charge, d is the thickness of the active region (hence J/ed is the rate of injection of carriers per unit volume), g is the gain coefficient per unit length and β is the fraction of the spontaneous emission coupled into the lasing mode, typically $\sim 10^{-2}$–10^{-5} for edge-emitting lasers (product of geometric and spectral overlap factors). These equations simply state that the rate of change of the carriers or photons is given by the rate of generation less the loss rate. Additional terms are required in the presence of optical feedback or coupling (in which case the photon density is replaced by the complex amplitude and phase of the optical electric field) or for extremely rapid modulation where the traveling wave nature of the disturbances in the photon fields is significant. These so-called traveling-wave rate equations for the photons are of the form:

$$\mathrm{d}S_+/\mathrm{d}t + c\mathrm{d}S_+/\mathrm{d}z = gNS_+ + \beta N/\tau_s \qquad [3]$$

$$\mathrm{d}S_-/\mathrm{d}t - c\mathrm{d}S_-/\mathrm{d}z = gNS_- + \beta N/\tau_s \qquad [4]$$

with the same carrier rate equation as [1] using the total photon number S made up of the forward-traveling component S_+ and the backward-traveling component $S_- : S = S_+ + S_-$.

Both sets of rate equations are based on approximations such as homogeneous gain broadening, neglect of transverse field effects, and diffusion. These are valid in most situations but care is required when analyzing lasers with large transverse apertures, or when ultrashort (picosecond) pulses are involved.

Small-Signal Modulation

Small-signal modulation may be analyzed by a linear perturbation analysis of the rate equations, leading to a conjugate pair of poles in the response function, with damped relaxation oscillations at frequency:

$$f_R = (1/2\pi)\surd(gS_0/\tau_p) \qquad [5]$$

or, in terms of the injected current density J (assuming $S \sim G(J - J_{th})/ed$ with Γ the electromagnetic confinement factor of the mode):

$$f_R = (1/2\pi)\surd[\Gamma g(J - J_{th})/ed]$$

f_R can be regarded as the resonant frequency for the interaction between the carriers and photons. As a practical matter, lasers can be modulated up to $\sim 2f_R$ but the modulation response rolls off rapidly with increasing frequency above f_R. Using explicit expressions for the threshold gain, we can write f_R as:

$$f_R = (1/2\pi)\surd[(GN_T g\tau_p + 1)(J/J_{th} - 1)/\tau_s\tau_p] \qquad [6]$$

where N_T is the transparency carrier density $\sim 10^{18}\,\mathrm{cm}^{-3}$. f_R can take values in excess of 10 GHz for a well-designed laser. It should be noted, however, that high modulation bandwidth almost always requires high differential gain $\mathrm{d}g/\mathrm{d}N$ and short τ_p. Increasing $\mathrm{d}g/\mathrm{d}N$ requires p-doping (which increases optical loss) and/or some special quantum-confining structure such as wells, wires, or dots. Decreasing τ_p inevitably leads to greater optical loss and thus a higher threshold.

The damping rate of the relaxation oscillations is also important in limiting modulation bandwidth, and also (when increased) in reducing the sensitivity of the laser to optical feedback. For bulk or quantum well lasers, the condition for critical damping is $A^2 = 4B$ with $A = (S/g\tau_s) - \sigma[N - N_T]/g\tau_p$ and $B = \sigma[1 + (S_0/g\tau_s) - (1 - \beta)(N - N_T)]$. For quantum dot lasers there are additional damping terms due to carrier transport and thermalization which may restrict small-signal modulation bandwidths to a few GHz.

Experimentally, small-signal modulation properties may be determined using a simple sampling oscilloscope, low-noise tunable signal generator and spectrum analyzer. A combination of these elements with automated frequency sweeping may be found in a scalar network analyzer acting as an *s*-parameter test set. The modulated laser is connected to port 1 and a high speed photodetector connected to port 2, then a swept-frequency measurement of the transfer characteristic s_{21} is performed. In addition to the limitations on modulation response due to the carrier-photon resonance, practical lasers are limited by RC parasitics in the laser chip (junction impedance) and its package. For the highest modulation speeds, packages need to be designed as microwave transmission line components with effective impedance matching and low back reflections, as characterized by the voltage standing wave ratio (VSWR).

Large-Signal Modulation

Large-signal direct modulation can be analyzed by numerical integration of the rate equations. In general, its results are beyond the scope of this review, but for effective high-speed response, digital modulation of laser diodes should be carried out with a pre-bias close to threshold. Modulation bandwidth also increases with laser power, which is limited for systems considerations by fiber nonlinearities, and for reasons of laser reliability. In pulse code modulation, care must be taken that the laser is near critical damping, to minimize thermal patterning effects due to long strings of ones (high power) or zeros (low power). Experimentally, large signal modulation is analyzed by constructing eye diagrams, in which pulse traces on a fast oscilloscope are continuously overlaid when the laser is modulated using pseudorandom binary pulses. When the signal quality is high, the high and low digital levels are easily distinguishable, resulting in an open ‘eye’ in the accumulated traces.

Requirements for Externally Modulated Lasers

The requirement for external modulation occurs when the laser chirp exceeds half the desired channel spacing in the communication system. Chirp occurs due to changes in the refractive index, and hence the optical phase, during modulation. This dynamic phase shift then results in an instantaneous frequency shift.

For direct modulation we have instantaneous wavelength

$$\lambda(t) = \lambda(0)(\mu(t)/\mu(0)) \qquad [7]$$

Where $\mu(t)$ is the refractive index at time t. μ shifts with carrier density N due to combinations of several factors (free carrier plasma dispersion, thermal bandgap shrinkage, bandgap renormalization, dynamic Burstein–Moss effects) which give $\mathrm{d}\mu/\mathrm{d}N \sim -3\times10^{-20}\,\mathrm{cm}^3$, which for large signal digital modulation can result in tens of GHz. The actual degree of chirp depends on the so-called antiguiding or linewidth enhancement factor a given by

$$\alpha = -2k_0(\mathrm{d}\mu/\mathrm{d}N)/(\mathrm{d}g/\mathrm{d}N) \qquad [8]$$

where $k_0 = 2\pi/\lambda_0$ is the free-space wavenumber. α as defined is positive, typically in the range 2–7 depending on the material, operating carrier density, and wavelength relative to the gain peak. Distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, which use integrated gratings to control the lasing wavelength, can have lower chirp by detuning from the gain peak but therefore are likely to require careful temperature stabilization. Quantum dot lasers, based on 3D nanoclustered active regions, have the potential for very low chirp, narrow linewidths, and reduced wavelength shifts, so that directly modulated DWDM transmitters may be possible.

Integration of laser diodes with electro-absorption or Mach–Zehnder type modulators is achieved in the transmitter optical subassembly and may in some cases be accomplished by monolithic integration on the same chip. Using separate modulators allows each device to be optimized independently but requires optical assembly, alignment, and retention. Mach–Zehnder modulators in X-cut lithium niobate, for example, have essentially zero chirp and can operate to 40 Gbit/s and above, with dynamic extinction ratios (ratio between fully on and fully off) of ~20 dB. Integrated semiconductor M-Z modulators have chirp an order of magnitude less than those of directly modulated lasers, so that channel spacings ~25 GHz are possible in DWDM systems with suitable temperature and wavelength controls. Although the details of such modulators are beyond the scope of this article, the laser requirements are to generate ~10 mW of continuous power with stable center wavelength and narrow linewidth. Such lasers are usually mounted in hermetic packages (e.g., the current 14-pin butterfly standard) with integrated thermo-electric cooler, power monitor photodiode, an optical isolator to suppress optical backreflections which cause instabilities and self-pulsing, and optional wavelength locking optics.

Reliability of Lasers in Fiber Optic Systems

From the early days, when laser diode lifetimes were measured in seconds even at cryogenic temperature, enormous progress has been made in achieving materials purity and reducing crystalline defects. Today's fiber optic laser transmitters have projected lifetimes of decades. In common with electronic devices, diode lasers fail at a rate given by a 'bathtub' curve, that is the failure rate $r(t)$ defined as the probability of failure per unit time at time t, has relatively high values at low t (early failures) and high t (wearout failures) and very low values in between. In terms of the population of lasers $n(t)$ we have:

$$r(t) = (-1/n(t))\,\mathrm{d}n(t)/\mathrm{d}t \qquad [9]$$

If Δn is the number of samples which fail in time Δt, then assuming Δt begins at $t = 0$, the effective or average failure rate r_{eff} over the interval is

$$r_{\mathrm{eff}}(t) = \Delta t \Delta n/n(0) \qquad [10]$$

In reliability science it is customary to define r_{eff} in FITs (failures integrated in time) with the time-span chosen to yield statistically significant numbers. For electronic and photonic devices, it is customary to select the time interval $\Delta t = 10^9$ h (1 billion operating hours), so that if 1% of the devices fail in 10 years ($\sim 10^5$ hours) we have $r_{\mathrm{eff}} \sim 100$ FITs, which is the order of magnitude required by modern fiber optic laser sources.

In terms of actual statistical models, failure rates can be estimated using failure probability density functions $f(t)$. This is related to $r(t)$ by $f(t) = r(t)S(t)$ where $S(t)$ is the cumulative probability of surviving until t. When $r(t)$ is nearly constant, as in early failures due to material defects or process errors, the statistics are approximately exponential:

$$f(t) + (1/\tau)\exp(-t/\tau) \qquad [11]$$

or Weibull-distributed (an exponential is a Weibull distribution with $\beta = 1$):

$$f(t) = (\beta/t^{\beta})t^{b-1}\exp[-(t/\tau)^b] \qquad [12]$$

but this description is not suited to wearout failures for which $f(t)$ shows a significant increase as the population ages. For such cases, the lognormal distribution is often applicable:

$$f(t) = (1/\sigma t\sqrt{(2\pi)})\exp[-(\ln t - \ln\tau)^2/2\sigma^2] \qquad [13]$$

where in each case τ is approximately the mean time to failure (MTTF).

Laser diodes undergo standardized qualification tests to demonstrate sufficient reliability prior to being used in commercial systems. Operationally, failure statistics must be tabulated and the best fit obtained. Burn-in procedures are used to screen early failures. Chip and module failures should be distinguished clearly, as the latter includes many additional factors such as thermal and power management, light coupling, mechanical or chemical integrity.

Given that effective lifetimes of decades are required, it is clearly necessary to accelerate aging to produce statistically meaningful failure rates in reasonable times $\sim 10^3$ h. For the most common modes of laser chip degradation, due to recombination-induced aggregation of defects in the crystalline epitaxial material, these are thermally activated and current driven, so that it is customary to use a modified Arrhenius law at temperature T:

$$\tau \sim (1/J^n)\exp(E_a/k_B T) \qquad [14]$$

where J is the operating current density, n is an index ~ 2, E_a is the activation energy, and k_B is Boltzmann's constant. In terms of thermal acceleration, we have a factor $F = \tau(T_A)/\tau(T_B)$ with:

$$F = \exp[(E_a/k_B)(1/T_A - 1/T_B)] \qquad [15]$$

Typical values of E_a for laser diodes are ~ 1 eV so that acceleration factors $\sim 10^3$ are possible.

Finally, it should be noted that laser diodes are generally subject to catastrophic failure in the event of overdriving or static discharge, the failure mode being thermal facet damage at $\sim 1-10$ MW/cm^2 for continuous-wave operation, the exact value depending on the material, surface preparation and specifically surface state absorption and its temperature coefficient.

Conclusions

Laser diodes are ideal sources for optical fiber communication systems and have propelled the development of fiber optics from its origins in the 1960s to the present day. They are compact, rugged, efficient, and reliable sources of light at key wavelength ranges such as 1300–1310 nm (short haul high speed links) and 1500–1600 nm (long haul amplified systems). They are capable of direct modulation at gigabit rates for simple systems but require external modulation and careful wavelength control for dense wavelength division multiplexed terabit systems. Transmitter lasers for real systems must satisfy stringent reliability conditions.

See also

Lasers: Edge Emitters; Semiconductor Lasers. **Semiconductor Physics:** Quantum Wells and GaAs-based Structures.

Further Reading

Telcordia Technologies (formerly Bellcore): document family FR-796 on Reliability and Quality Generic Requirements for telecommunications equipment. Latest issue is 003 dated Jul 2003. The specific document related to laser reliability is GR-468 – Generic Reliability Assurance Requirements for Optoelectronic Devices used in Telecommunications Equipment. Standards for other network photonic, electronic and mechanical components may be found in the document family FR-2063 Network Equipment-Building System Family of Requirements, issue 003 dated Feb 2003 and specifically FR-357 – Network Equipment-Building System Family of Requirements: Component Design for Reliability.

Adams MJ and Thomas B (1977) Detailed calculations of transient effects in semiconductor injection lasers. *IEEE Journal of Quantum Electronics* QE-13: 580.

Agrawal GP (2002) *Fiber Optic Communication Systems*, 3rd edn. New York: Wiley.

Alferov ZhI, Andreev VM, Portnoi EL and Trukan MK (1970) AlAs-GaAs heterojunction injection lasers with a low room-temperature threshold. *Soviet Physics – Semiconductors* 3: 1107.

Basov NG, Krokhin ON and Popov YuM (1961) Production of negative-temperature states in p-n junctions of degenerate semiconductors. *Soviet Physics – JETP* 13: 1320.

Bernard MGA and Duraffourg G (1961) Laser conditions in semiconductors. *Physica Status Solidi* 1: 699.

Bhattacharya P (1996) *Semiconductor Optoelectronic Devices*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Cheng J and Dutta NK (eds) (2000) *Vertical Cavity Surface-Emitting Semiconductor Lasers: Technology and Applications*. London: Taylor & Francis.

Chuang S-L (1995) *Physics of Optoelectronic Devices*. New York: Wiley.

Coldren LA and Corzine SW (1995) *Laser Diodes and Photonic Integrated Circuits*. New York: Wiley.

Dupuis RD (1987) An introduction to the development of the semiconductor laser. *IEEE Journal of Quantum Electronics* QE-23: 651.

Fukuda M (1991) *Reliability and Degradation of Semiconductor Lasers and LEDs*. Norwood, MA: Artech House.

Fukuda M (1998) *Optical Semiconductor Devices*. Wiley.

Guekos G (ed.) (1998) *Photonic Devices for Telecommunications*. New York: Springer.

Hall RN, Fenner GE, Kingsley JD, Soltys TJ and Carlson RO (1962) Coherent light emission from GaAs junctions. *Physical Review Letters* 9: 366.

Hayashi I, Panish MB, Foy PW and Sumski S (1970) Junction lasers which operate continuously at room temperature. *Applied Physics Letters* 17: 109.

Holonyak N Jr and Bevacqua SF (1962) Coherent (visible) light emission from Ga(As1-xPx) junctions. *Applied Physics Letters* 1: 82.

Ito R, Nakashima H, Kishino S and Nakada O (1975) Degradation sources in GaAs-AlGaAs double-heterostructure lasers. *IEEE Journal of Quantum Electronics* QE-11: 551.

Kaminow IP and Li T (eds) (2002) *Optical Fiber Telecommunications IV (two parts)*. Amsterdam: Elsevier.

Kaminow IP and Koch TL (eds) (1997) *Optical Fiber Telecommunications III (two parts)*. London: Academic Press.

Kapon E (1998) *Semiconductor Lasers, (two parts)*. London: Academic.

Kartalopoulos SV (1999) *Introduction to DWDM Technology*. New York: Wiley.

Kartalopoulos SV (1999) *Understanding SONET/SDH and ATM*. IEEE Press.

Kasap SO (2001) *Optoelectronics and Photonics: Principles and Practices*. Englewood Cliffs, NJ: Prentice-Hall.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Norwood, MA: Artech House.

Keiser G (1991) *Optical Fiber Communications*. New York: McGraw-Hill.

Kressel H and Lockwood HF (1974) A review of gradual degradation phenomena in electroluminescent diodes. *Journal de Physique* 35: C3–223.

Lasher G and Stern F (1964) Spontaneous and stimulated recombination radiation in semiconductors. *Physical Review* 133: A553.

Milonni PW and Eberly JH (1988) *Lasers*. New York: Wiley.

Morthier G and Vankwikelberge P (1997) *Handbook of Distributed Feedback Laser Diodes*. Norwood, MA: Artech House.

Nathan MI, Dumke WP, Burns G, Dill FH Jr and Lasher G (1962) Stimulated emission of radiation from GaAs p-n junctions. *Applied Physics Letters* 1: 62.

Numai T (2004) *Fundamentals of Semiconductor Lasers*. New York: Springer.

Paoli TL (1977) Changes in the optical properties of cw (AlGa) as junction lasers during accelerated aging. *IEEE Journal of Quantum Electronics* QE-13: 351.

Paoli TL and Ripper TE (1970) Direct modulation of semiconductor lasers. *Proceedings of IEEE* 58: 1457.

Pearsall TP (2003) *Photonics Essentials: An Introduction with Experiments*. New York: McGraw-Hill.

Piprek J (2003) *Semiconductor Optoelectronic Devices: Introduction to Physics and Simulation*. London: Academic.

Quist TM, Rediker RH, Keyes RJ, Krag WE, Lax B, McWhorter AL and Zeiger HJ (1962) Semiconductor maser of GaAs. *Applied Physics Letters* 1: 91.

Reinhart FK, Hayashi I and Panish MB (1971) Mode reflectivity and waveguide properties of double-heterostructure injection. *Lasers* 42: 4466.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Senior JM (1992) *Optical Fiber Communications*, 2nd edn. Engelwood cliffs, NJ: Prentice-Hall.

Siegman AE (1986) *Lasers*. University Science Books.

Verdeyen JT (1994) *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: Wiley.

Zachos TH and Ripper JE (1969) Resonant modes of GaAs junction lasers. *IEEE Journal of Quantum Electronics* QE-5: 29.

Local Area Networks

E Wong, University of Melbourne, Melbourne, VIC, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Local area networks (LANs) are privately owned networks that are used to carry internal traffic loads within an organization. A LAN may exist on many levels categorized by its application and speed. This is illustrated in Figure 1 whereby, at the simplest requirement, a low-speed LAN may be used to interconnect a cluster of personal computers and workstations. Such a scenario may exist in a department within a university campus, or a building within a technology park. If need be, these individual departmental or building LANs may be interconnected using a moderately high-speed backbone LAN. The backbone LAN may also provide interconnection to servers which support mass storage devices that are shared by the individual lower speed LANs. In larger organizations, a much higher speed LAN is linked to the backbone LAN to support bulk data transfer between mainframes, supercomputers, and other office equipment.

The common transmission mediums for traditional low-speed LANs are coaxial cable and twisted pairs. High-speed backbone LANs advocate the use of the optical fiber. The continuing progress in fiber-optic technology and devices, and the availability of appropriate standards have given rise to LANs that can span distances of hundreds of kilometers or more, and that operate at data rates in the order of hundreds to thousands of megabits per second (Mb/s).

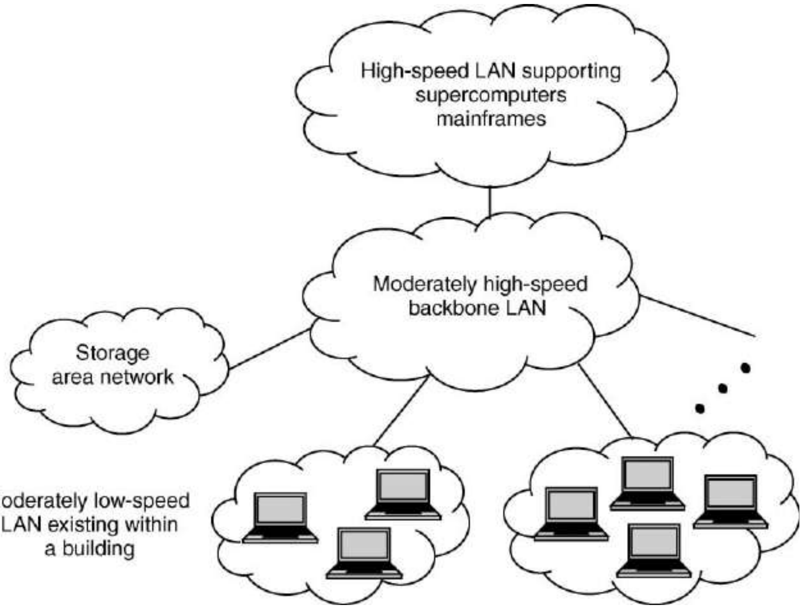

Figure 1 Interconnection of local area networks (LANs).

The increase in demand, to extend LAN capabilities towards higher data rates and across larger geographical areas, is fueled primarily by the need to provide new bandwidth intensive services such as multimedia teleconferencing, image processing, and streaming video, to a significantly higher number of end-users.

This article focuses on three existing fiber-optic LAN standards, namely Ethernet, Fiber Distributed Data Interface (FDDI), and Distributed Queue Dual Bus (DQDB). While the current uses of FDDI and DQDB are not as widespread as Ethernet, both standards have significantly contributed to the initial deployment of fiber-optics for LAN applications, with standard-based products still commercially available. In general, standards are required to ensure interoperability between components and equipment from different vendors and suppliers. Two key features are addressed by the standards: the physical topology, and the medium access control (MAC) protocol. A network's physical topology defines the way in which the users, or nodes, are interconnected to each other and to the transmission medium. Common optical LAN topologies are the star (e.g., Ethernet), bus (e.g., DQDB and Ethernet), and ring (e.g., FDDI). In comparison, the MAC protocol controls network access such that its capacity is efficiently and fairly shared amongst all nodes.

Access control can be either centralized or decentralized. In the former, a node wishing to transmit must wait until it receives permission from a controller, whereas in the latter, nodes collectively determine the transmission order. Further, capacity allocation to each node can be either synchronous or asynchronous. The synchronous approach guarantees a specific capacity to each node, e.g., using time division multiple access (TDMA), whereby nodes are allowed to transmit only during a designated time slot. In contrast, the asynchronous approach allows capacity to be dynamically allocated to each node. The asynchronous approach can be further categorized into round robin, reservation, and contention schemes. With round robin, nodes take it in turn to transmit in a cyclic manner and a node may transmit up to a maximum amount of data during its turn. With the reservation scheme, time on the medium is divided into slots, and a node wishing to transmit must reserve future slots by transmitting additional control information into the network. With the contention scheme, nodes randomly access the network. The Ethernet standard, discussed in the following section, is based on the contention scheme.

Ethernet

Overview

The IEEE (Institute of Electrical and Electronics Engineers) 802.3 standards committee first standardized 10 Mb/s Ethernet in 1985. It is now the most widely used technology for high-speed LANs, with standards extended to cater for operating speeds of 100 Mb/s, 1 Gb/s and 10 Gb/s, and encompassing the use of the optical-fiber transmission medium. Low-speed 10 Mb/s and 100 Mb/s Ethernet LANs use the carrier-sense multiple access with collision detection (CSMA/CD) MAC protocol, whereas a switched technique is used for the 1 Gb/s and 10 Gb/s options.

In CSMA/CD, a node wishing to transmit an Ethernet frame first listens to the transmission medium by detecting the presence of signals on the channel (i.e., carrier-sense). This process allows the node to determine if another transmission is in progress. If the medium is in use, the node continues to listen until the medium is idle, and then transmits immediately. While transmitting, the node continues to listen to the medium. It may happen that two or more nodes may attempt to transmit at the same time. If this occurs, there will be a collision, and data from both transmissions will become garbled.

A collision can be indicated by a combined signal that exceeds a predefined CD threshold voltage at the carrier-sense receiver, taking into account the collision between frames from nodes that are located furthest apart. If a collision is detected during transmission (i.e., collision detection), the transmitting node ceases transmission and proceeds to transmit a brief jamming signal to notify all nodes of the collision. After transmitting the jamming signal, the node must back off before attempting to transmit the same data frame again. The back-off period is random, and is determined by a binary exponential back-off algorithm. Every time a node attempts to retransmit, the mean value of the random back-off period is doubled until the tenth attempt. After that the node is allowed to retransmit a further six times with the same mean random delay before discarding the frame and reporting an error to higher control layers.

The transmission medium remains unusable when frames collide but only until the collision is detected. As such, the minimum transmission time of a frame must be longer than the maximum propagation delay of the network so that collision is detected prior to the end of transmission. If shorter frames are used, then collision detection does not occur, and CSMA/CD thus exhibits the same performance as its less efficient predecessor protocol, CSMA. Accordingly, the CSMA/CD standards specify a minimum frame length and maximum network size, which place fundamental limitations on the practical application of a pure CSMA/CD LAN.

IEEE 802.3 10 Mb/s Ethernet

In 1993, the IEEE 802.3 committee extended the 10 Mb/s Ethernet standard to include the optical-fiber transmission medium. The standard notation of this option is 10BASE-F. Each transmission link comprises a pair of graded-index 62.5 μm core multimode optical fibers (MMF), one for each direction of transmission. The transmission wavelength is 850 nm. 10BASE-F can be further categorized into three specifications. The first, denoted 10BASE-FP (fiber passive star), specifies a broadcast star topology that can interconnect up to 33 nodes and repeaters. Each of these devices is attached to a passive star coupler via two optical-fiber links with distances up to 500 m. The passive star coupler splits the optical power of any incoming frame equally amongst its output links, allowing all nodes or repeaters in the network to receive the frame, albeit at a lower optical power level.

The second and third specifications, denoted 10BASE-FL (fiber link) and 10BASE-FB (fiber backbone) respectively, define point-to-point links over distances of up to 2 km. 10BASE-FL supports both nodes and repeaters, whereas 10BASE-FB supports only repeaters. A repeater is a device that detects the optical signals on its incoming link, performs electrical signal regeneration, and then optically retransmits the signals on its output link. Repeaters used in 10BASE-FB have an additional function in which the detected signals are retimed with a local clock before being retransmitted. This prevents the propagation of timing distortions, allowing distances of up to 15 cascaded repeater links to be implemented. The characteristics of the three specifications for 10BASE-F, along with the 100 Mb/s and 1 Gb/s specifications to be discussed in the following sections, are summarized in Table 1.

IEEE 802.3 100 Mb/s Ethernet (Fast Ethernet)

Fast Ethernet refers to the standard developed by the IEEE 802.3 committee to provide low-cost, Ethernet-compatible LANs operating at 100 Mb/s. The optical-fiber transmission medium option is denoted 100BASE-FX. Originally defined for FDDI (covered in the section on FDDI below), 100BASE-FX specifies the use of two 62.5 μm MMF links, each operating at a unidirectional rate of 100 Mb/s. While 10 Mb/s Ethernet can only operate in half-duplex mode, whereby a station can either transmit or receive a frame but cannot do both simultaneously, Fast Ethernet can operate in either half- or full-duplex modes. A node operating in full-duplex mode can simultaneously transmit and receive data frames.

In order to deploy full-duplex Fast Ethernet, all repeaters must be replaced with switching hubs. Figure 2a illustrates a repeater located in the central point in an active star topology. Each node is attached to the repeater via two optical links. The repeater retransmits frames received from any one of its input

Table 1 Summary of IEEE 802.3 physical and MAC layer specifications for 10 Mb/s, 100 Mb/s and 1 Gb/s Ethernet

	10BASE-FP	*10BASE-FL*	*10BASE-FB*	*100BASE-FX*	*1000BASE-SX*	*1000BASE-LX*
Data rate	10 Mb/s	10 Mb/s	10 Mb/s	100 Mb/s	1 Gb/s	1 Gb/s
Transmission wavelength	850 nm	850 nm	850 nm	850 nm	850 nm	1310 nm
Transmission medium	62.5 μm MMF pair	62.5 μm MMF pair	62.5 μm MMF pair	62.5 μm MMF pair	50 μm or 62.5 μm MMF pair	10 μm SMF, 50 μm or 62.5 μm MMF pair
MAC protocol	CSMA/CD (half duplex)	CSMA/CD (half duplex)	CSMA/CD (half duplex)	CSMA/CD (half duplex) Switched (full duplex)	CSMA/CD (half duplex) Switched (full duplex)	CSMA/CD (half duplex) Switched (full duplex)

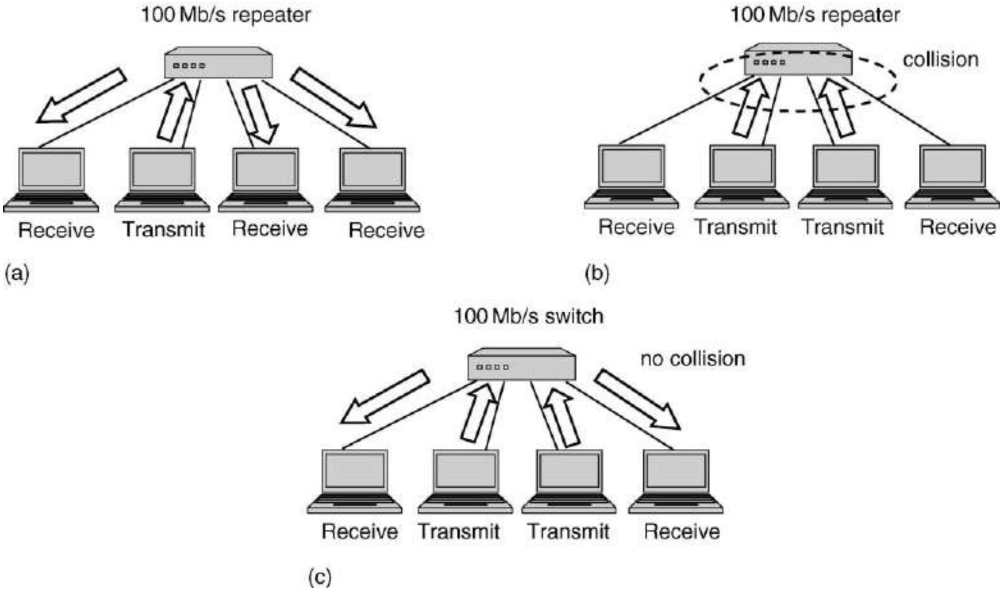

Figure 2 (a) Shared LAN repeater. (b) Collision at a shared LAN repeater due to simultaneous transmission of frames. (c) LAN switch (adapted from William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education).

links to all of its output links. If more than one node transmits at the same time, then collisions will occur at the repeater, as illustrated in Figure 2b. A switching hub, on the other hand, switches an incoming frame from a particular input to an appropriate output line. The same switching function can be performed simultaneously on other input–output links without interference to the frames involved, as illustrated in Figure 2c. As such, each node has a dedicated capacity equal to that of the transmission rate of the LAN, and the switching hub can support mixed capacity links, e.g., 10 Mb/s and 100 Mb/s. With dedicated access to the medium and appropriate buffering at the switch to avoid contention of the same output link, there is no need to implement the CSMA/CD access algorithm. Nonetheless, the conventional Ethernet frame format is used to ensure inter-operability between existing 10 Mb/s and 100 Mb/s LANs.

IEEE 802.3 1 Gb/s Ethernet (Gigabit Ethernet)

The IEEE 802.3 1 Gb/s Ethernet standard, commonly referred to as Gigabit Ethernet, was finalized in 1998. An example of a Gigabit Ethernet backbone LAN is shown in Figure 3, whereby a Gigabit Ethernet backbone switch connects central servers to high-speed workgroup switches. In turn, each workgroup switch supports gigabit high-performance workgroup servers, as well as 100 Mb/s workstations and hubs. As in the 100 Mb/s standard, Gigabit Ethernet uses the conventional Ethernet frame format and allows both half-duplex and full-duplex operation. For half-duplex operation, the CSMA/CD access protocol is

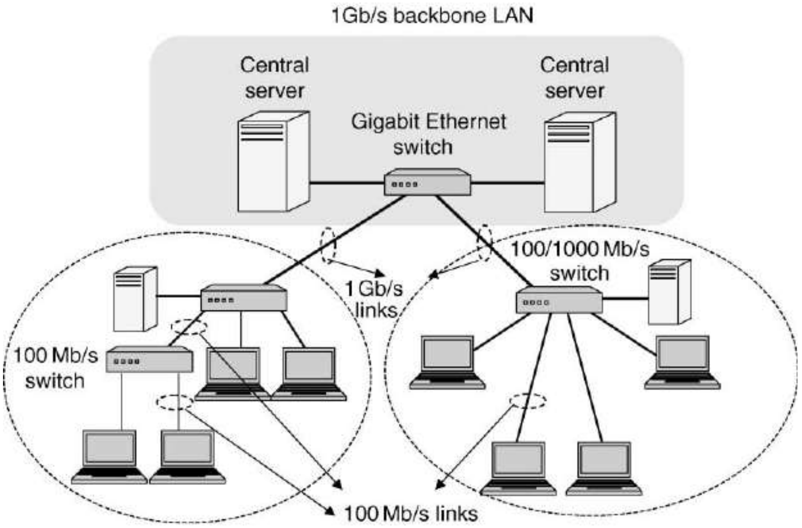

Figure 3 Gigabit Ethernet backbone network (adapted from William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education).

equipped with two additional features. The first feature, known as carrier extension, extends the propagation time of short frames by adding nondata symbols to the end of frames. The minimum length of a frame for Gigabit Ethernet is 512 bytes as opposed to the specified minimum of 64 bytes in the preceding two standards. The second feature, known as frame bursting, allows a node to transmit multiple consecutive frames, up to a specified maximum, without having to undergo the contention process for each transmission. Both features improve the efficiency of the protocol.

There are two optical-fiber transmission medium specifications for Gigabit Ethernet. The short wavelength option, denoted 1000BASE-SX, is for 850 nm transmission. Specified maximum link distances are 275 m using 62.5 μm MMF, and 550 μm using 50 μm MMF. The long wavelength option, denoted 1000BASE-LX, is for 1310 nm transmission. Achievable duplex link distances are 550 m using 62.5 μm or 50 μm MMF, and 5 km using 10 μm SMF.

IEEE 802.3 10 Gb/s Ethernet (10GbE)

The Ethernet standard was extended to incorporate 10 Gb/s (10GbE) in 2002, fueled by ever-increasing Internet traffic and bandwidth intensive applications. 10GbE fulfills not only LAN specifications, but also metropolitan and wide-area network specifications. It is envisioned that with the emergence of 10GbE metropolitan and wide-area networks, the deployment of 10GbE LANs will enable a seamless transport of Ethernet frames across different network boundaries, thus simplifying network management and optimizing operational costs.

Unlike its predecessor standards, the physical layer options for 10GbE include only the optical-fiber transmission medium and allow only for full-duplex operation. Four specifications are defined for 10GbE. 10GBASE-S is designed for 850 nm transmission on 63.5 μm MMF, while 10BASE-L and 10BASE-LE use 10 μm single mode fiber (SMF) for 1310 nm transmission and 1550 nm transmission, respectively. The corresponding maximum achievable distances are 300 m, 10 km, and 40 km, respectively. An alternative specification, 10GBASE-LX4, is designed for wavelength division multiplexed (WDM) transmission links. The specification caters for four 2.5 Gb/s transmission channels within the 1270–1355 wavelength region on either 10 μm SMF, 50 μm MMF, or 62.5 μm MMF. Single-mode transmissions can reach distances of up to 10 km, while multimode transmissions can reach distances up to 300 m. The characteristics of these physical layer specifications are summarized in Table 2.

Fiber Distributed Data Interface (FDDI)

Overview

The Fiber Distributed Data Interface (FDDI) standard was developed by the Accredited Standards Committee X3T9.5, and approved by both ANSI (American National Standards Institute) and ISO (International Standards Organization) in 1989. While not as

Table 2 Summary of IEEE 802.3 physical and MAC layer specifications for 10 Gb/s Ethernet

	10GBASE-S	*10GBASE-L*	*10GBASE-E*	*10GBASE-LX4*
Data rate	10 Gb/s	10 Gb/s	10 Gb/s	4 × 10 Gb/s
Transmission wavelength	850 nm	1310 nm	1550 nm	1270 nm to 1355 nm
Transmission medium	62.5 μm MMF pair	10 μm SMF pair	10 μm SMF pair	10 μm SMF or 50 μm, 62.5 μm MMF pair
MAC protocol	Switched (full duplex)	Switched (full duplex)	Switched (full duplex)	Switched (full duplex)

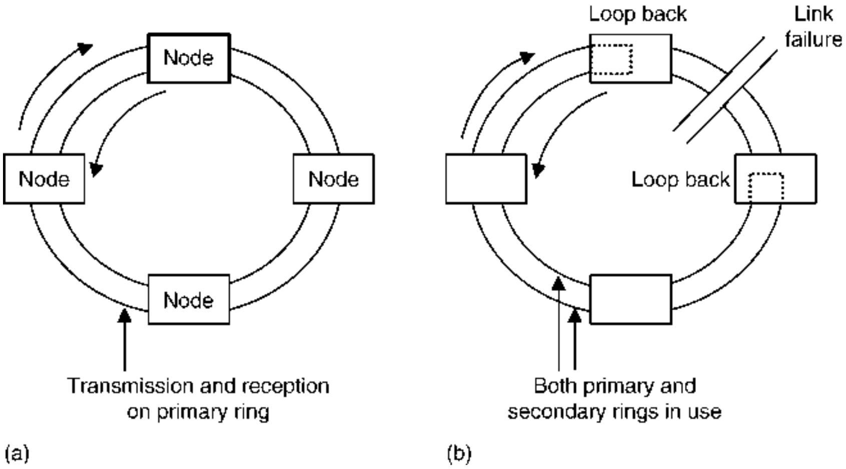
Node
Node
Node
Node
Transmission and reception on primary ring
(a)
Loop back
Link failure
Loop back
Both primary and secondary rings in use
(b)

Figure 4 (a) FDDI ring network. (b) Use of secondary ring for protection against link failures.

widely deployed as the Ethernet family of LANs, due to the high cost of equipment and devices, it remains as the first to be standardized for high-speed fiber-optic LAN applications and is now mainly used to interconnect existing FDDI-compatible LANs. Extensions have been made to the FDDI standard to incorporate single-mode fiber operation for increased signal reach.

FDDI specifies a dual counter-propagating ring network as shown in Figure 4a. Each ring operates at the speed of 100 Mb/s, and can reach a maximum length of 100 km. The dual ring topology is self-healing against fiber-cuts and node failures. Under normal operation, frame transmission and reception are performed on a primary ring. The secondary ring is utilized to form a closed loop when a node or link failure occurs, as illustrated in Figure 4b. In FDDI, the interface between each node and the optical-fiber transmission medium is a repeater. Data frames, each of which carries its source and a destination address in a defined field in the frame header, are transmitted sequentially around the ring from one repeater to another. Each frame can vary in length by up to 4550 bytes. Each repeater copies the destination address field of all incoming frames and, upon recognizing the destination address as its own, additionally copies the remainder of the frame. The repeater attached to the source node also facilitates the removal of frames it has transmitted after a round-trip propagation around the ring.

FDDI employs a distributed timer-controlled, token-passing mechanism, referred to as timed token rotation protocol, to control access to the shared optical fiber ring. While Ethernet-based LANs are highly suited to support throughput sensitive services which can sustain reasonable delay in return for high throughput, FDDI provides an additional capability to support delay sensitive services such as packetised voice and video. For such services, frames must be received within a certain time interval for quality of service requirements. A node connected to an FDDI network can therefore transmit both classes of services: the former referred to as burst traffic and the latter as stream traffic. In addition, the timed token rotation protocol can limit the amount of frames that can be transmitted by a particular node for each class of service, ensuring efficient and fair network access amongst all nodes and services. FDDI is thus well-suited to carry the mixture of burst and stream traffic expected in a backbone network which supports a number of smaller LANs.

MAC Protocol

Token-passing mechanism

In FDDI, access to the shared transmission medium is governed by a timed token rotation protocol. A token is a special frame created during network (re)-initialization by a pre-assigned node. The token circulates the ring network and gives the node that holds it the right to transmit. Figure 5 gives an example of a single frame transmission from Node A to Node C, and from Node B to Node A. Node A has a frame to transmit to Node C and thus waits for a token (Figure 5a). Upon recognizing that a token frame is passing through, it aborts the token transmission without regenerating the token bits. After Node A has completely received the token, the node begins transmitting frame FA addressed to Node C, and appending a new token upon completion (Figure 5b). Meanwhile, Node B regenerates and retransmits FA as it circulates past. When FA reaches Node C, the node copies FA in addition to regenerating and retransmitting FA. Node B captures the token issued by A and transmits FB to Node A (Figure 5c). Eventually FA recirculates back to its originating node, Node A, whereby it is removed from the ring (Figure 5d). At the same time, Node B finishes transmitting FB and releases the token. Nodes C and D, with no frames to transmit or receive, regenerate and retransmit both FB and the token. When FB reaches Node A, it is copied by the node. Node A also regenerates and retransmits both FB and the token. FB is finally removed from the ring by Node B (Figure 5e), and the token remains in the ring to be received by the next node with a frame to transmit (Figure 5f).

Aside from source and destination address fields, each dataframe also contains a frame status field which comprises error detected (E), address recognized (A), and frame copied (C) indicator bits. Each node can check passing frames for errors and can set the E indicator if an error is detected using the standard cyclic redundancy check procedure. If a node detects its own address, it sets the A indicator, and upon copying the frame, the station may also set the C indicator. When a dataframe recirculates back to its source node, the indicator bits are examined. If both A and C are not set, then the destination node does not exist or is inactive. On the other hand, if the frame has not been copied by an active destination node or an error has been detected, higher control layers are notified for future retransmission.

Time-controlled mechanism

FDDI uses a time-controlled mechanism to prevent nodes from dominating the network capacity. During initialization, a target token rotation time (TTRT) is established based on the shortest allowable time between token arrivals at each node. The same TTRT value is then stored at all nodes. In addition, each node has a maximum capacity guaranteed for the transmission of its stream traffic. The total transmission time of this stream traffic and the latency of the ring (i.e., total round-trip propagation time, the transmission time of a maximum length frame and the token frame) must never exceed the TTRT. During network operation, each node measures the

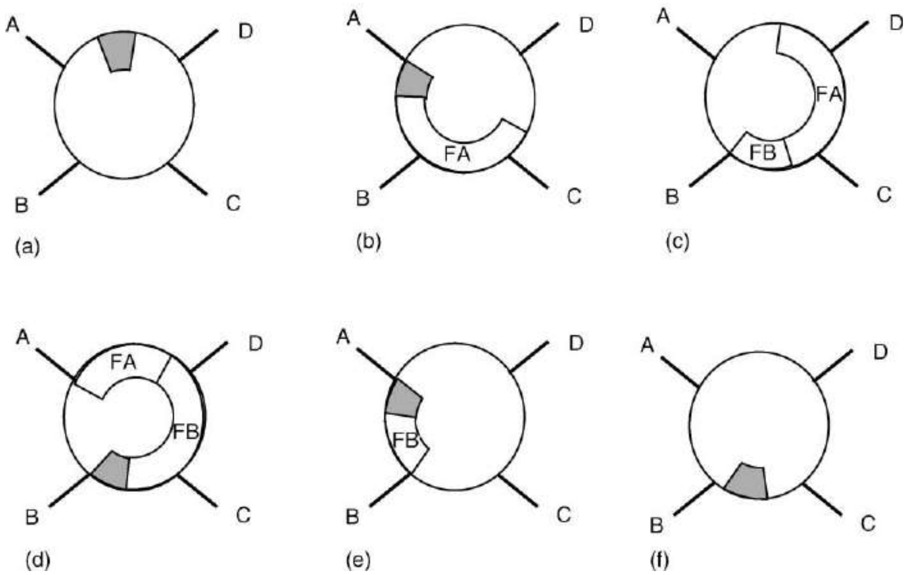

Figure 5 Fiber Distributed Data Interface (FDDI) token passing mechanism.

time interval between successive token arrivals, using a token-rotation timer (TRT). The TRT is enabled only upon a token arrival, and expires if the counter reaches the TTRT value. In turn, a late counter (LC) is incremented and TRT is reset to zero. The number of TRT expirations since a token was last received is, therefore, counted by LC. Any token that arrives at a node with a nonzero LC value is referred to as a late token. Likewise, an early token is one which arrives at a node with a zero LC value.

The action of a node when it receives a token depends on whether the token is early or late. When a late token is received, LC is reset to zero but TRT continues to run. The node then transmits for the period of time which corresponds to the predetermined bandwidth for steam traffic transmission. The node is not allowed to transmit any bursty traffic frames. In the event of an early token reception, the node transfers the current value of TRT to a second timer, the token-holding timer (THT) but does not enable it. TRT is reset to zero and is enabled to track the next rotation of the counter. The node then transmits stream traffic frames for the allocated period of time. After that, THT is enabled and bursty traffic frames are transmitted until THT reaches the TTRT value. The THT is then reset and the node releases the token into the ring. The transmission of bursty frames may be further controlled by two additional mechanisms. The restricted-token mechanism allows only two nodes in the network to use all the available bursty bandwidth, and restricts all other nodes to their stream bandwidth. This is facilitated by the use of a restricted token which is issued by one of the two nodes. The priority mechanism provides each station with the capability to transmit bursty frames according to eight priority levels. Each priority level corresponds to a different time threshold in the THT.

Note that the time-controlled mechanism of the MAC protocol guarantees a maximum response time for the ring because, in the worst case, the time between the arrival of two successive tokens will never exceed twice the value of TTRT. Most multimedia applications, such as video conferencing, real-time video, and entertainment video can be supported on FDDI with stream service and a small TTRT. However, since the TTRT cannot be less than the ring latency, applications requiring time bounds less than twice the ring latency cannot be supported by FDDI. Long and continuous isochronous services which must adhere to strict periodic access require FDDI-II.

FDDI-II

FDDI-II provides the additional circuit-switching capability to support constant bit-rate telecommunication applications. Like FDDI, the transmission bit-rate of FDDI-II is 100 Mb/s. FDDI-II can operate in either complete isochronous mode, complete packet-switching mode, or in hybrid mode in which both isochronous as well as bursty and stream services are supported. Strict periodic access required by isochronous services are accomplished with transmission opportunities repeated every 125 μs. At 100 Mb/s, time on the ring is divided into several frames, each 1562.5 bytes long. Out of the 1562.5 bytes, 1560 bytes are used for data transmission while the remaining 2.5 bytes are used as interframe gap. In turn, the 1560 bytes are divided into 16 wideband channels of 96 bytes each. Each wideband channel thus provides a bandwidth of 6.144 Mb/s, sufficiently adequate for television broadcast and multiple voice channels. For operation in the hybrid mode, some of the wideband channels are allocated for packet mode transmission. Likewise, if complete packet-switching operation is required, all wideband channels are allocated for packet switching transmission.

IEEE 802.6 Distributed Queue Dual Bus (DQDB)

Overview

The IEEE 802.6 Distributed Queue Dual Bus (DQDB) standard, finalized in 1990, specifies a dual bus network with an operating speed of 150 Mb/s. DQDB is now mainly used for MAN applications, although it can be readily deployed for LAN backbone interconnection. Figure 6 illustrates the basic architecture of a DQDB network with access nodes, Node 1 to Node *N*, distributed along two counter propagating buses, BUS A and BUS B. Each access node is considered to be connected to a lower speed LAN. At the head end of each bus, a slot generator continuously generates fixed-size time slots which propagate along its respective bus. The contents of each time slot are regenerated at each access node it passes, and upon reaching the end of the bus, a sink terminal receives and removes the time slots from the network. Data are transported across the network from one access node to another in the form of fixed-size segments. A node transmits a segment by changing the contents of a passing time slot. Each segment needs only be transmitted on one of the two buses, the choice depending on the locations of the transmitting and receiving nodes. For example, referring to Figure 6, a downstream transmission from Node i to Node $i + 1$ requires the use of BUS A. On the contrary, an upstream transmission from Node i to Node $i - 1$ uses BUS B.

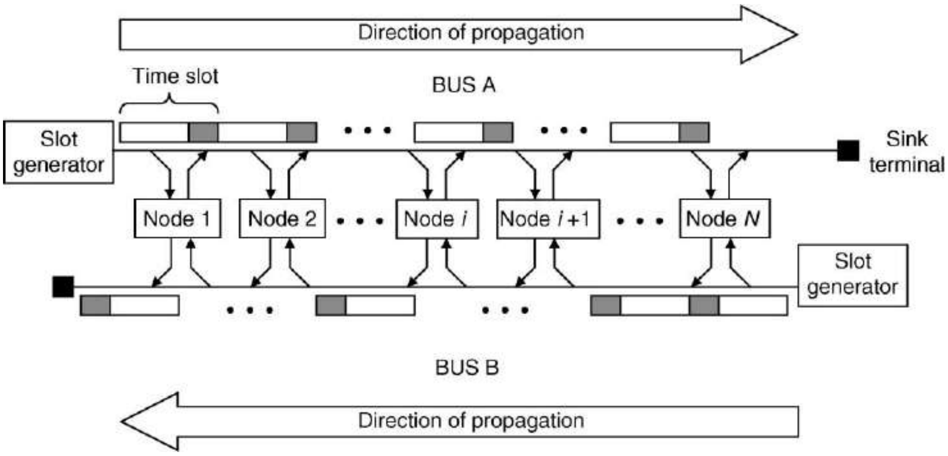
Direction of propagation
Time slot
BUS A
Slot generator
Sink terminal
Node 1
Node 2
Node *i*
Node *i* +1
Node *N*
Slot generator
BUS B
Direction of propagation

Figure 6 Distributed Queue Dual Bus (DQDB) architecture.

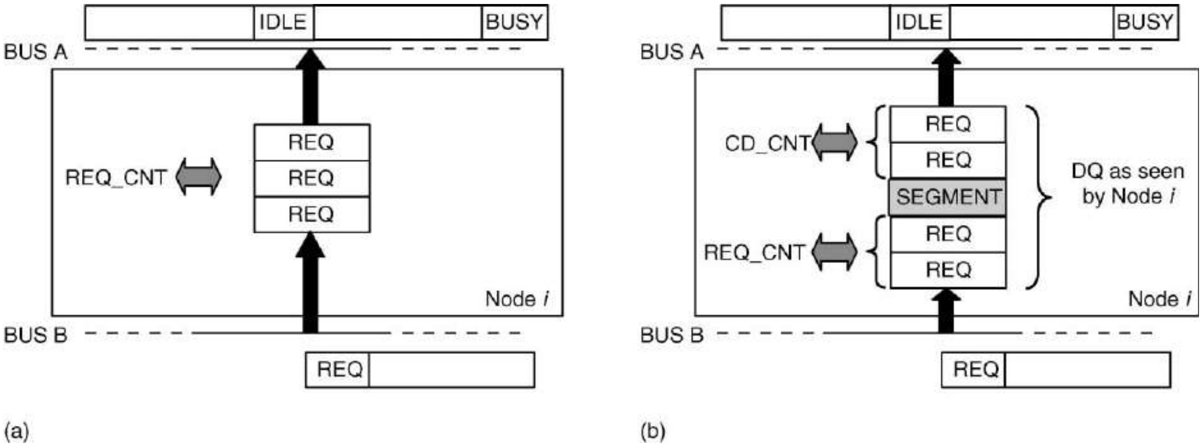
IDLE
BUSY
BUS A
REQ_CNT
REQ
REQ
REQ
Node *i*
BUS B
REQ
(a)
IDLE
BUSY
BUS A
CD_CNT
REQ
REQ
SEGMENT
REQ_CNT
REQ
REQ
DQ as seen by Node *i*
Node *i*
BUS B
REQ
(b)

Figure 7 Behavior of Node *i* during (a) idle state; and during (b) countdown state.

MAC Protocol

In DQDB, permission to transmit segments into the network is determined by a global distributed queueing algorithm. The algorithm allows each node to continuously monitor both buses, counting the requests and available time slots that pass by. In addition to maintaining a global first-in-first-out queue (DQ) for each bus, the algorithm relies on information in the header of time slots, namely the BUSY and REQ fields, and counters at the access nodes to coordinate transmissions. A ‘BUSY = 0’ field represents the availability of a time slot for the transport of a segment. For example, time slots generated at the head end of each bus have default ‘BUSY = 0’ and ‘REQ = 0’ fields. The BUSY field is set to 1 when a segment is transmitted into an available time slot.

The REQ field is used in conjunction with the counters to inform nodes of an impending transmission from downstream. To simplify the explanation, consider only packets propagating on BUS A. A node wishing to transmit a segment downstream needs to queue the segment in the appropriate DQ. The node with an impending transmission must then notify all upstream nodes that it has inserted a segment in the DQ so that an available time slot can be reserved for the transmission. This is accomplished by setting the first time slot propagating in the opposite bus, BUS B, with REQ = 0 to REQ = 1. As such, the node maintains the outstanding segments in the DQ belonging to downstream nodes via the value in the REQ field.

To illustrate, Figure 7a shows an arbitrary node, Node *i*, in an idle state, whereby the node has no segments to transmit downstream on BUS A. With every ‘REQ = 1’ observed on BUS B, a request counter, REQ_CNT is incremented by one. However, with every empty slot, i.e., ‘BUSY = 0’, observed on

Table 3 Comparison of standardized local area networks

	Ethernet	*Fast Ethernet*	*Gigabit Ethernet*	*10GbE*	*FDDI*	*DQDB*
Data rate	10 Mb/s	100 Mb/s	1 Gb/s	10 Gb/s	100 Mb/s	150 Mb/s
MAC protocol	CSMA/CD	CSMA/CD or Switched technique	CSMA/CD or Switched technique	Switched technique	Timed token rotation	Distributed queue
Standard	IEEE 802.3	IEEE 802.3	IEEE 802.3	IEEE 802.3	ANSI-FDDI	IEEE 802.6

the BUS A, REQ_CNT is decremented by one. When Node *i* has a segment to transmit, it enters the countdown state as illustrated in Figure 7b. The value in REQ_CNT is transferred into an independent countdown counter, CD_CNT, and REQ_CNT is reset to zero. Node *i* also sets 'REQ = 1' in the opposite bus, BUS B. As in the idle state, each 'REQ = 1' observed on BUS B increments REQ_CNT while each 'BUSY = 1' observed on BUS A decrements it. When CD_CNT reaches zero, Node *i* transmits its segment in the first time slot with 'BUSY = 0' on BUS A and sets 'BUSY = 1'. The node returns to the idle state if there are no other segments to transmit. The algorithm allows the node to determine the number of time slots reserved by downstream nodes and the time to transmit a segment into an available time slot. Note that both the idle and countdown states are replicated for transmissions on BUS B.

The DQDB standard also specifies a limited support of isochronous services. Nodes which provide and receive such services are reserved a certain amount of time slots which are indicated with additional header information. The remaining time slots are used by all other access nodes for the transport of packet-switched traffic, and access to these time slots is managed by the global distributed queueing algorithm. A priority mechanism based on using multiple DQs, each with a different priority level, may also be implemented to guarantee access to certain segments in the network.

Summary

Due to the explosive demand for high-bandwidth applications, fiber-optics has become an essential technology in LANs. Table 3 compares the various fiber-optic LAN standards discussed in this article. The Ethernet family remains the most prevalent of LAN technologies. Its widespread use is contributed by the availability of low-cost chipsets, mature and familiar management and analysis tools. The future of fiber-optic LANs operating beyond 10 Gb/s will depend upon WDM solutions. The current usage of one transmission channel per network is impeded primarily by the limitation of the speed of electro-optic conversion, and thus does not fully exploit the massive transmission bandwidth of the optical fiber. With WDM, multiple transmission channels are concurrently transmitted on a single optical fiber, thereby greatly increasing the capacity of an optical LAN. Ongoing research efforts have been directly towards advancing WDM component technology and addressing prevailing issues such as polarization control, chromatic dispersion, and four-wave mixing among others.

List of Units and Nomenclature

CD_CNT	countdown counter
CSMA/CD	carrier-sense multiple access with collision detection
DQ	global first-in-first-out queue
DQDB	distributed queue dual bus
FDDI	fiber distributed data interface
LAN	local area network
LC	late counter
MAC	medium access control
MAN	metropolitan area network
MMF	multimode fiber
REQ_CNT	request counter
SMF	single mode fiber
TDMA	time division multiple access
THT	token-holding timer
TRT	token rotation timer
TTRT	target token rotation timer
WDM	wavelength division multiplexing

See also

Information Processing: Optical Neural Networks.

Further Reading

IEEE Std 802.6 (1990) IEEE standards for local and metropolitan area networks: distributed queue dual bus (DQDB) subnetwork of a metropolitan area network (MAN).

IEEE Std 802.3 (2002, Revision of IEEE Std 802.3, 2000 ed.) IEEE Standard for Information Technology – Telecommunications and Information Exchange Between Systems – Local and Metropolitan Area

Networks – Specific Requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications.
ISO 9314-1,2 (1989) Information processing systems – Fibre Distributed Data Interface (FDDI) – Part 1: Token Ring Physical Layer Protocol (PHY) and Part 2: Token Ring Media Access Control (MAC).
Marco C, Enrico G and Luciano L (1997) *Metropolitan Area Networks*. New York: Springer-Verlag.
Mills A (1995) *Understanding FDDI*. Englewood Cliffs, NJ: Prentice Hall.
William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education.

Optical Time Division Multiplexing

L P Barry, Dublin City University, Dublin, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

As the demand for bandwidth continues to grow, driven by the massive increase in Internet usage, so will the necessity to have communications networks that can handle very high data rates. The use of optical fiber networks is the clear choice for such systems given the huge available bandwidth of the fiber transmission medium. However, in a basic optical communication system comprising a laser transmitter, an optical fiber transmission medium, and a receiver, the capacity is essentially limited by the speed at which light can be modulated at the transmitter. To overcome this limitation, which is basically due to the speed of available electronics, it is necessary to use optical multiplexing techniques, and one such technique, known as optical time division multiplexing (OTDM), will be the subject of this article.

Principles of Time Division Multiplexing

Time division multiplexing (TDM) has long been the traditional method for electrically combining information channels. If we take the most fundamental data rate to be that of a simple voice call at 64 kbit/s, then the transmission of data at higher bit rates is achieved by electrically multiplexing a large number of 64 kbit/s channels in the time domain. With the evolution of standards, a number of different data rates have been specified as standard transmission rates. In Europe the synchronous digital hierarchy (SDH) has a basic data rate of 155.52 Mbit/s, which is known as the synchronous transport module – Level 1 (STM-1). This particular data rate is essentially obtained by electrically multiplexing over 2000 voice calls in the time domain (with some of the capacity required for overhead information). By subsequently multiplexing a number of STM-1 channels together we can obtain transmission at the higher standard data rates of STM-4 (622 Mbit/s), STM-16 (2.48 Gbit/s), and STM-64 (9.88 Gbit/s), as outlined in Table 1.

Figure 1 illustrates how basic electrical TDM is used in standard optical communication systems. Clearly as we approach the higher data rates of STM-64, a serious level of electrical multiplexing is required, and as the data rates increase so does the cost and complexity of the electrical equipment at the transmitter and receiver. Indeed the present state of the art in electronics seems to suggest 40 Gbit/s as the limit for electrically multiplexed communication systems. However, as we stated earlier, communications traffic has been growing explosively over the last decade and will continue to do so. In order to meet this demand for capacity, and better exploit the massive available bandwidth of optical fiber, it is necessary to use optical multiplexing techniques for communications systems. The two main optical multiplexing techniques available are wavelength division multiplexing (WDM) and optical time division multiplexing (OTDM). WDM essentially involves transmitting data at a number of different wavelengths on the same fiber. Although electrical multiplexing may be limited to data transmission rates of around 40 Gbit/s using one laser, by multiplexing together N different wavelength channels each carrying 40 Gbit/s, we can achieve overall data rates up to and beyond a terabit/s.

The second of these optical multiplexing techniques, OTDM, is the subject of this article. Whereas WDM multiplexes optical data channels in the

Table 1 Standard data rates for SDH transmission systems

SDH standard	*Data rate (Mbit/s)*
STM-1	155.52
STM-4	622.08
STM-16	2488
STM-64	9953
STM-256	39 813

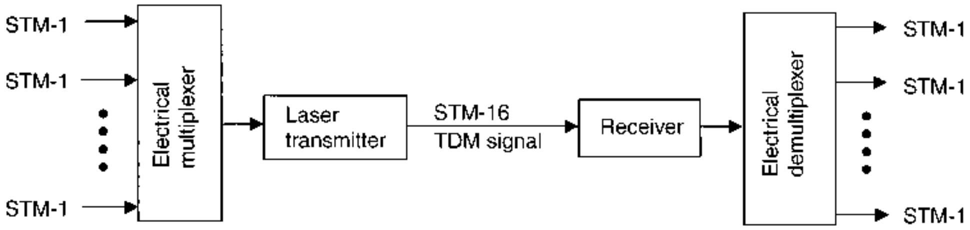
STM-1
STM-1
STM-1
Electrical multiplexer
Laser transmitter
STM-16
TDM signal
Receiver
Electrical demultiplexer
STM-1
STM-1
STM-1

Figure 1 Electrical time division multiplexing (ETDM) in a basic optical communication system.

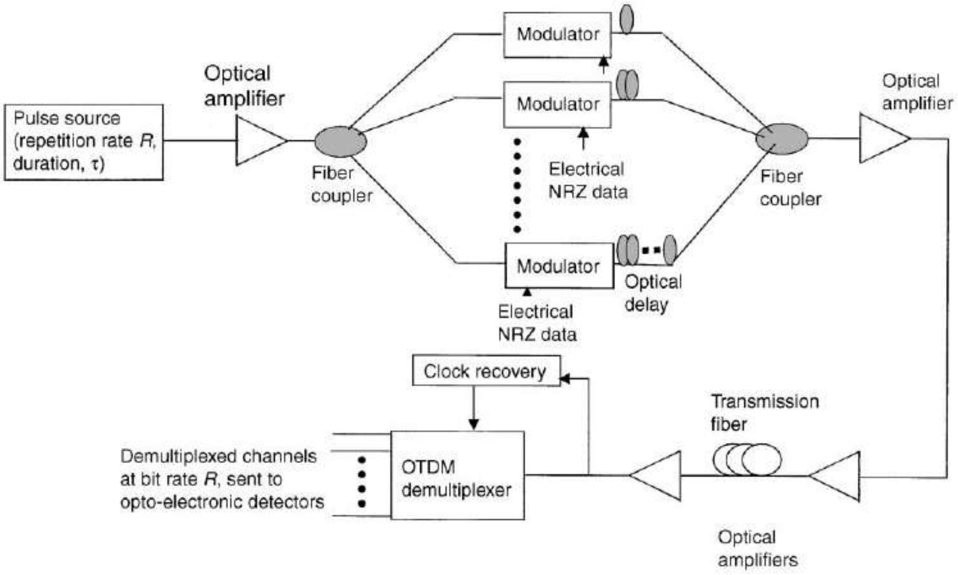
Pulse source (repetition rate R, duration, τ)
Optical amplifier
Fiber coupler
Modulator
Modulator
Modulator
Electrical NRZ data
Electrical NRZ data
Optical delay
Fiber coupler
Optical amplifier
Clock recovery
Transmission fiber
Demultiplexed channels at bit rate R, sent to opto-electronic detectors
OTDM demultiplexer
Optical amplifiers

Figure 2 Basic configuration for an OTDM transmission system comprising transmitter, transmission fiber, and receiver.

wavelength domain, the basic principle of OTDM communications is to increase the system capacity by multiplexing optical data channels in the time domain. This multiplexing is usually achieved by a process known as bit-interleaving. This process can be explained by considering Figure 2, which shows a basic schematic of an OTDM based transmission system. The main component of the overall system is a source of ultrashort optical pulses (pulse duration τ) at a certain repetition rate, R. The optical pulse source is initially split into N channels using a passive fiber coupler, and each pulse train is subsequently modulated by electrical data which is at a data rate of R. The resulting output from each modulator is essentially an optical data channel where the data are represented using ultrashort optical pulses (return-to-zero data format). The data from each modulator then passes through a fixed fiber delay line which delays each channel by a time equal to $1/RN$ relative to its adjacent channel (as shown in Figure 2). The N modulated and delayed optical data channels are then recombined in another passive fiber coupler to form the OTDM data signal. We can consider that the delay lines essentially assign each data channel to a specific bit slot (of width $1/RN$) in the overall multiplexed signal. The multiplexed data signal may then be transmitted over optical fiber before arriving at the receiver which is responsible for demultiplexing the optical signal into its discrete channels.

The duration of the optical pulse source is extremely important in determining the maximum overall data rate which can be achieved. The overall data rate is basically defined by the temporal separation between channels in the multiplexor, but in order to avoid cross-channel interference, this separation must be significantly greater than the pulse duration. Thus to increase the overall data rate we must use shorter optical pulses. However, as we reduce the optical pulse width to raise the data rate we need to take into account the problems that may be encountered as this high-speed data signal propagates over optical fiber, and also the difficulty in demultiplexing a high-speed OTDM data signal. The following section will look in greater detail at these key elements of the OTDM communication system, namely the optical pulse source, the transmission of the ultrahigh capacity signal, and the demultiplexing at the receiver.

Key Elements of an OTDM Communication System

Ultra-Short Optical Pulse Sources

As stated earlier, the optical pulse source is a key element in any OTDM-based communication system. The important characteristics of the pulse source that will affect its usefulness in an OTDM system are the pulsewidth, the spectral width, and the temporal jitter. The pulse duration clearly has to be short enough to support the desired overall transmission rate. For example, if we wish to design an OTDM system with an aggregate data rate of 100 Gbit/s (which will have a delay of 10 ps between each channel in the multiplexed signal), then the pulsewidth should normally be less the 30% of the channel spacing to avoid cross-channel interference (i.e. around 3 ps). For terabit/s OTDM systems we would require pulsewidths less than 0.5 ps. The spectral width of the pulse source is also important, as it will have a major impact on how the pulse will evolve during propagation in the fiber. A standard figure of merit which is employed is the time–bandwidth product, and ideally we require the pulse source to be transform limited, which implies that the spectral width is as small as possible for the associated pulsewidth. The impact of temporal jitter on the pulses can be easily understood by considering the above example of a 100 Gbit/s system, employing 3 ps pulses, with the multiplexed channels spaced by 10 ps. Obviously if the jitter on the pulses becomes significant in relation to the channel spacing, then this can also lead to interference between adjacent channels in the overall OTDM system. A large number of techniques have been employed to develop ultrashort pulse sources suitable for use in OTDM systems, but the three main methods which will be described below are active mode-locking, gain-switching, and external modulation of a CW light signal.

Active mode-locking of laser diodes normally involves modulating the amplitude of the optical field inside the laser cavity at a frequency which is equal to the mode spacing of the laser. This can be achieved by applying an electrical sinusoidal signal at the correct frequency, and results in the generation of optical pulses at the repetition rate of the applied signal. This technique has been successfully employed in the generation of subpicosecond pulses at repetition rates up to and beyond 40 GHz, with excellent spectral and temporal jitter characteristics. However, an inherent problem in all mode-locked pulse sources is the difficulty in synchronizing the mode-locked frequency to a specific SDH standard data rate.

Gain switching of semiconductor laser diodes is probably the simplest and most reliable technique to generate optical pulses. The technique, which is presented in Figure 3, involves applying a high-power electrical pulse (or electrical sinusoidal signal) to the laser in conjunction with a certain bias current. By ensuring that the electrical pulse signal and the bias signal have the correct level, the relaxation oscillation phenomenon of the laser results in the production of optical pulses with durations between 10 and 30 ps, at the repetition rate of the applied electrical signal. The frequency of the electrical signal applied to the laser is essentially arbitrary (provided it is not larger than the modulation bandwidth of the diode), thus making it straightforward to synchronize the optical pulse train to a SDH line-rate. The main problem with this technique is that the spectral width of the pulses generated is such that the pulses are far from transform-limited, which would affect their subsequent propagation in the fiber. In addition, temporal jitter on gain-switched pulses can also be a problem. However, by using novel arrangements such as external injection into the gain-switched laser, this difficulty can be overcome.

The third pulse generation technique mentioned above involves external modulation of a cw light signal with an electro-absorption modulator. The experimental configuration for this pulse generation technique is shown in Figure 4. By biasing the modulator around its null point, and applying an electrical sinusoidal signal to it, the cw light passing through the modulator becomes shaped into optical pulses. The optical pulse train which is generated due

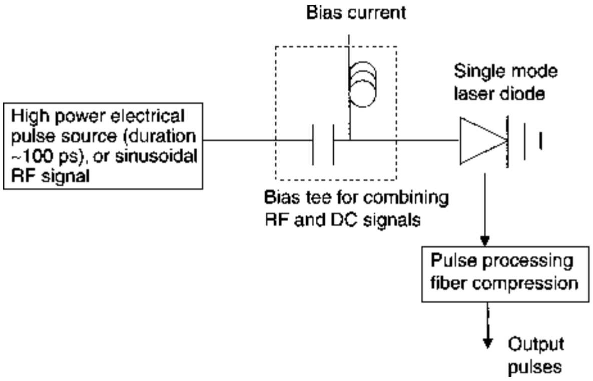
Bias current
Single mode laser diode
High power electrical pulse source (duration ~100 ps), or sinusoidal RF signal
Bias tee for combining RF and DC signals
Pulse processing fiber compression
Output pulses

Figure 3 Pulse generation using the gain-switching technique followed by pulse compression.

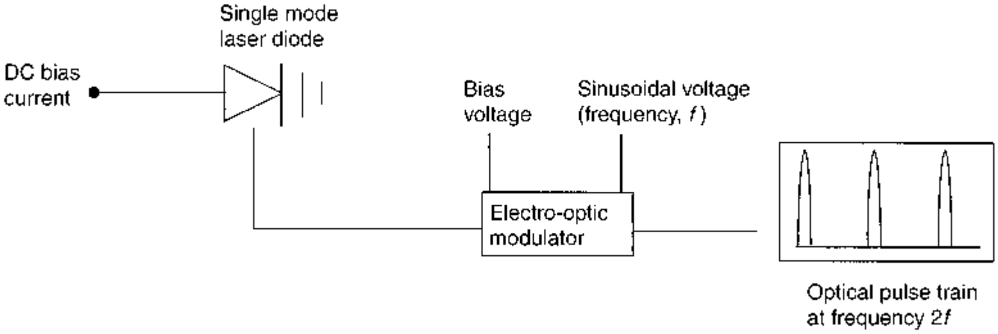

Figure 4 Pulse generation based on shaping of cw light using the nonlinear response of an external modulator biased about its null point.

the nonlinear response of the electro-absorption modulator is at a repetition rate of twice the applied electrical signal. The pulses generated are normally transform limited with extremely low temporal jitter, and the repetition rate is arbitrary as in the gain-switching technique (with the limit being ultimately determined by the modulation bandwidth of the modulator). This method may be readily used to generate pulses at repetition rates up to 40 GHz with pulsewidth around 5 ps, and it has the advantage that the optical source and the modulator can be integrated in a single device to form a compact pulse source suitable for OTDM communication systems.

It should also be noted that in addition to the various techniques that have been outlined, it is possible to use pulse compression in order to reduce the pulse width. The main issues concerned with pulse compression are the shape and spectral width of the optical pulses after compression. One of the most attractive methods of achieving pulse compression involves using nonlinear compression in dispersion-decreasing fiber, with the main advantage of this technique being the ability to maintain a transform limited pulse after compression. By employing this compression scheme, optical pulse sources at 10 GHz with pulsewidths below 200 fs have been developed, and such pulsewidths would be suitable for use in Tbit/s OTDM systems.

Transmission of an OTDM Signal over Fiber

The transmission performance of an OTDM data signal is vital in determining the distance over which the data can be transmitted successfully. The main fiber parameters, which will affect the signal transmission, are attenuation and dispersion. If we consider OTDM systems operating at a wavelength of around 1550 nm (minimum loss wavelength), to overcome the fiber attenuation and maintain a suitable optical power budget for the system, optical amplifiers are normally employed. If we thus assume that the amplifiers overcome the fiber loss problem, then the maximum transmission distance will be limited by the fiber dispersion. In a very basic OTDM system operating at 1550 nm, with transmitter and receiver linked using standard fiber (dispersion parameter of about 16 ps km^{-1} nm^{-1}), the maximum transmission distance will be limited by the overall data rate, and the pulsewidth and spectral width of the optical pulse source. For example, consider a 40 Gbit/s OTDM data signal which is formed using 8 ps optical pulses with a spectral width of 40 GHz (0.32 nm). From the spectral width we can calculate the signal broadening due to dispersion to be around 5 ps km^{-1}, and as the pulses broaden and spread into adjacent channels of the OTDM signal then this interference will make it increasingly difficult to correctly detect the signal at the receiver. In this case after propagation through 5 km of fiber, the signal pulses will have already dispersed to around 25 ps duration, the same value as the temporal bit slot into which each channel is placed. This will clearly result in serious loss of signal integrity. A straightforward possibility to increase the transmission distance of OTDM systems is to employ dispersion shifted fiber; the dispersion parameter is now around 1–2 ps km^{-1} nm^{-1}, at an operating wavelength of 1550 nm. This reduction in dispersion will obviously increase the allowed transmission distance by about an order of magnitude for the example described above. However, to develop ultrahigh-speed, long-haul OTDM communications we require more complex transmission schemes. Two possibilities for this include dispersion compensation, and soliton transmission techniques.

Dispersion compensation basically involves compensating for the dispersion that has been encountered during transmission by using some fiber with a total dispersion of opposite sign but equal magnitude to the transmission fiber. Dispersion compensation may also be achieved using a suitably designed fiber grating. For a long-haul OTDM system, a dispersion compensator may be used every 50 or 60 km, in conjunction with the optical amplifier, thus allowing us to compensate both fiber loss and attenuation periodically along the link. The main limitation, however, with the dispersion compensation technique is caused by the dispersion slope of the fiber, as for ideal compensation it is necessary to compensate completely for the dispersion slope in addition to the overall fiber dispersion. The second technique that may be employed to greatly extend the transmission distance of high-speed OTDM communication systems is the use of soliton transmission. The basic principle of soliton transmission is to use optical data pulses with a particular shape, pulsewidth, and peak power, such that as the pulse propagates, the effects of fiber dispersion and nonlinearity counterbalance to allow the signal to propagate undistorted. By using optical amplifiers to ensure that the optical pulse peak power does not vary too much along the transmission link, OTDM transmission at data rates greater than 100 Gbit/s, over distances approaching 500 km have been demonstrated.

Demultiplexing of OTDM Signal at Receiver

For OTDM communication systems with data rates above 40 Gbit/s, it is not feasible to use electrical switching. Indeed for ultrahigh-speed systems operating at 100 Gbit/s and beyond, the only solution for demultiplexing is to use all-optical switching in which optical control signals are used to switch the OTDM data signal. All-optical switches normally use nonlinear optical effects either in optical fiber or semiconductors. A typical scenario (as presented in Figure 5) involves the injection of both the OTDM data signal and an optical control signal into the nonlinear device. The control signal consists of high-power ultrashort pulses at the repetition rate of the individual channels within the temporally multiplexed signal, and by synchronizing it with one of the OTDM channels it is possible to demultiplex this channel from the high-speed OTDM signal.

Two of the most popular all-optical demultiplexers are the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD). The NOLM is based on the nonlinear refractive index of optical fiber. It essentially consists of a 2 × 2 fiber coupler with its two outputs joined using a certain length of fiber. When an OTDM signal is injected into an input port of the coupler, it splits into two counterpropagating components in the fiber loop, and when these components recombine and interfere at the coupler, the overall signal is output through its initial input port. If we now inject a high-power control signal directly into the loop, such that it propagates unidirectionally, and is

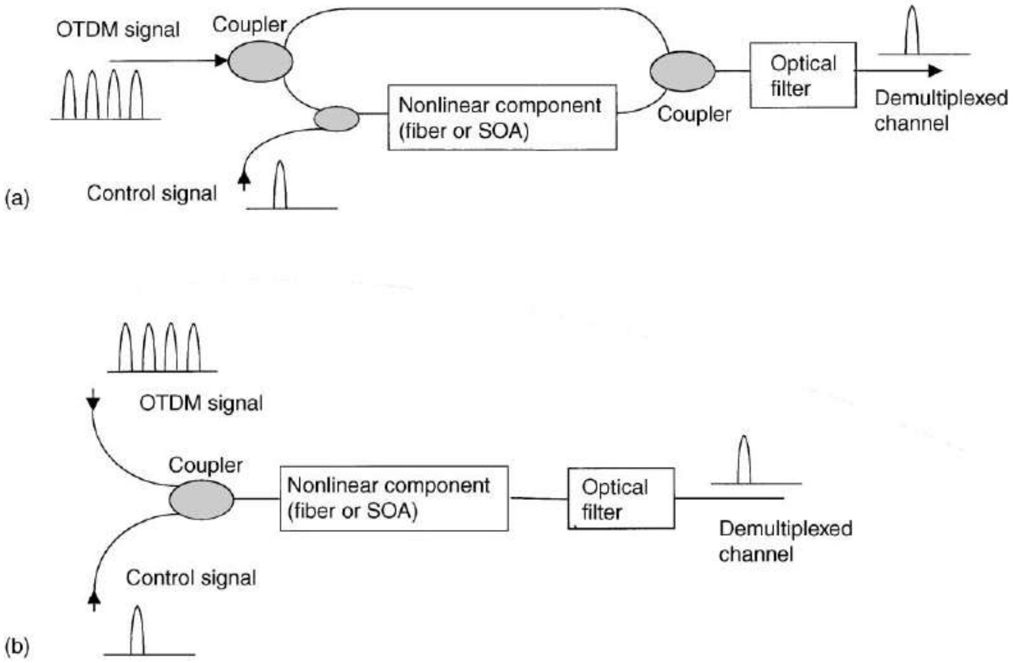

Figure 5 Configuration of OTDM demultiplexers using (a) an interferometer based on nonlinear phase shift in fiber or semiconductor optical amplifiers, and (b) four-wave mixing in fiber or semiconductor optical amplifiers.

synchronized with one of the OTDM data channels, then the phase shift induced by the control on the copropagating signal channel (via the nonlinear refractive index of the fiber) results in that particular channel being switched out to the second input port of the coupler. The remaining data channels of the OTDM signal are once again output through the same port that they entered the coupler. As this particular nonlinear effect in fiber (known as the Kerr effect) has a femtosecond response time, it should be possible to realize extremely high-speed switching with the NOLM. Demultiplexing at data rates in excess of 100 Gbit/s has already been achieved using this technique. One disadvantage of this method is that the nonlinear index coefficient of standard fiber is very small, so in order to achieve the required phase shift from the control pulse, it is necessary to use a fiber loop of around 1 km in length (depending on the power of the control pulse used). This length may affect the overall stability of the demultiplexer and its ability to be readily integrated into high-speed OTDM systems. However, it should be noted that recent developments in the design of high-nonlinearity fibers may reduce this problem.

The second all-optical switch mentioned above is known as a TOAD. Whereas the NOLM is based on the nonlinear refractive index of the fiber, the TOAD is based on a nonlinear optical effect in semiconductor optical amplifiers (SOA). The overall setup for a TOAD is very similar to that for a NOLM, except that a semiconductor amplifier, as shown in Figure 5, replaces the length of fiber in the loop. As for the NOLM, the OTDM data signal is injected into the TOAD using one input of the coupler, such that it splits into two counterpropagating signals, while the control signal propagates unidirectionally in the loop containing the SOA. By synchronizing the control with one channel in the OTDM signal, this particular channel is switched out. Although the response time of a TOAD-based switch may not be as fast as the NOLM, it does have the major advantage that it can be developed into a very compact demultiplexer, suitable for deployment in OTDM-based communication systems.

In addition to the NOLM and the TOAD structures, which are both interferometric-based switches, it is also possible to use four-wavemixing (FWM) in optical fiber or semiconductor optical amplifiers to carry out the demultiplexing of OTDM signals. In this case, the high-power control pulse is synchronized with one of the OTDM channels such that the wavelength of this channel is shifted as it passes through the nonlinear device. Optical filtering may then be used to select out the demultiplexed channel.

As we discussed in the section on demultiplexing, whatever specific all-optical switching technique is employed, we need to generate an optical clock signal for use as the control pulses. This implies that it is necessary to extract a clock at the base rate of the individual data channels from the high-speed OTDM signal. Depending on the base data rate, which may be anywhere from 2.5 to 40 Gbit/s for typical OTDM systems, different clock recovery techniques have been demonstrated. These range from basic electronic clock recovery schemes to more advanced techniques based on self-pulsating laser diodes and optical phase-locked loops.

OTDM Networking Issues

The previous section has predominantly dealt with the basic elements required to implement a high-speed OTDM transmission link. However, if we are to use OTDM technology for high-speed networking, then additional elements will be required. One of the most important of these is an add/drop multiplexer (ADM). The ADM allows us to switch out one channel from the OTDM signal, and then insert another data channel for transmission in the vacant bit slot. If we consider employing OTDM in a ring configuration network, then a typical architecture for the network will be as shown in Figure 6. The high-speed OTDM data signal propagates around the closed fiber loop, and each node in the network is responsible for switching out the data it requires, in addition to inserting the data it wishes to transmit onto the fiber ring. Add/drop multiplexing is thus imperative in developing such an OTDM ring network.

The configuration for each node in an OTDM ring network is shown in Figure 7. At each node it is initially necessary to recover the base rate clock signal. The recovered clock is then used in conjunction with the OTDM signal to switch out the required data channel from the multiplexed signal. When an add/drop multiplexer is used for this purpose we obtain two outputs, one being the demultiplexed data, and the other being the OTDM signal less the switched-out data channel. A simple fiber coupler can subsequently be used to allow this particular node to insert the data it wishes to send into the free bit slot of the overall OTDM signal. This obviously requires control of the relative delay between the OTDM signal and the local data channel to be inserted onto the fiber ring, to ensure that the local data are inserted into the vacant bit slot. A number of techniques have been used for implementing add/drop multiplexers in

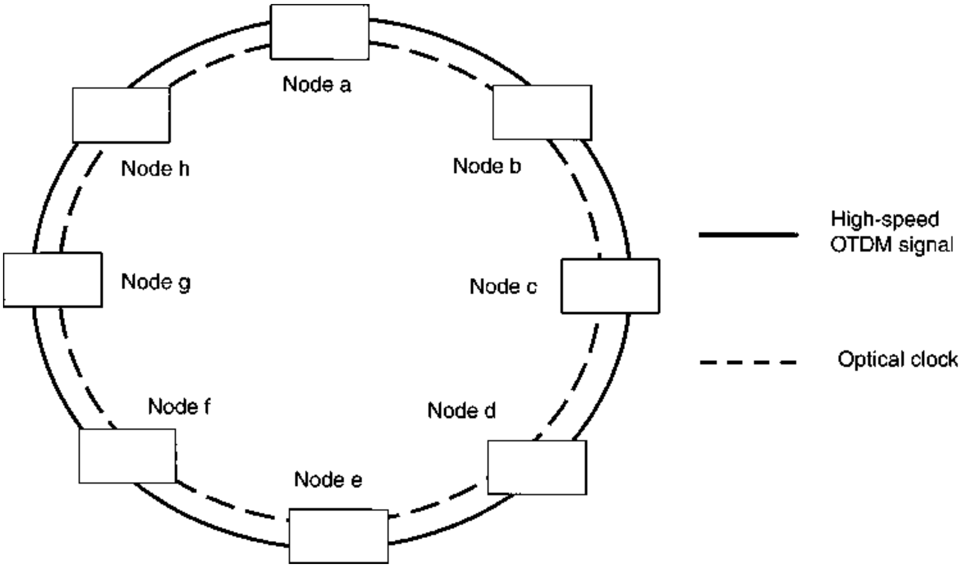
Node a
Node b
Node c
Node d
Node e
Node f
Node g
Node h
High-speed OTDM signal
Optical clock

Figure 6 Schematic representation of a ring network employing OTDM with a separate fiber for clock distribution.

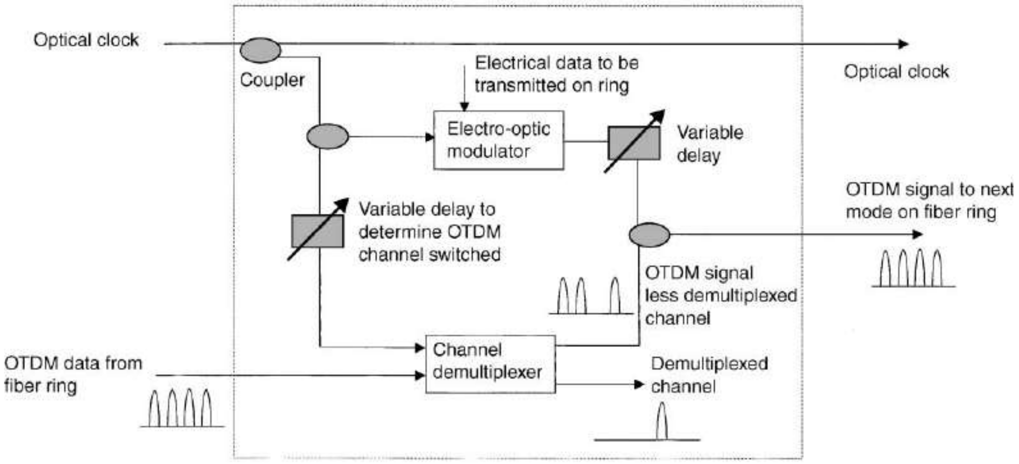
Optical clock
Coupler
Electrical data to be transmitted on ring
Electro-optic modulator
Variable delay
Optical clock
Variable delay to determine OTDM channel switched
OTDM signal to next mode on fiber ring
OTDM signal less demultiplexed channel
OTDM data from fiber ring
Channel demultiplexer
Demultiplexed channel

Figure 7 Configuration of an add/drop multiplexing node for use in OTDM-based communication systems.

OTDM-based networks. These have used either fiber interferometers, or interferometers using semiconductor optical amplifiers, in which we obtain both the demultiplexed data signal and the OTDM signal less the switched-out channel (to which the local data channel is then added). The continuing development and enhancement of ADM devices for OTDM systems will be vital for the future implementation of OTDM communication systems.

Conclusion

In this article we have explained the main ideas involved in building OTDM communication systems, and also introduced the principal technologies which are required to do so. OTDM is an extremely wavelength-efficient technique for delivering high-capacity data signals, and it may be used both for long-haul transmission and networking around metropolitan areas. In addition, unlike WDM, it does not require a different wavelength source for each channel in the multiplexed signal. Despite these and other advantages, OTDM is still way behind WDM when it comes to commercial maturity, with no OTDM systems currently available in the telecommunications market. However, the technologies required to implement high-speed OTDM systems (outlined in Table 2), including ultrashort pulse sources, clock extraction circuitry, and demultiplexing devices, are developing rapidly. It therefore seems

Table 2 Main elements of an OTDM communications system

Device	*Principal function*	*Technical implementation*
Optical pulse source	Generate ultrashort optical pulses suitable for transmission in high-speed OTDM systems	Gain-switching, mode-locking, or external modulation techniques
Multiplexer	Multiplex individual pulse data channels in the temporal domain using appropriate delays to achieve an OTDM signal	Passive couplers with fixed fiber delay lines, or planar waveguide circuits (for accurate control of delays)
Optical amplifier	Amplify the OTDM signal during fiber transmission to overcome fiber loss	Fiber doped with rare earth materials (e.g. erbium) in conjunction with pump laser
Transmission fiber	Transmit the OTDM data signal from transmission site to required receiver	Standard single-mode fiber, dispersion shifted fiber, or other specialty fiber may be used
Clock recovery	Extract clock (at base rate of individual channels) from OTDM data signal at receiver, for use in demultiplexing	Electronic or optical phase-locked loops, self-pulsating lasers, or mode-locked ring laser techniques
Demultiplexer	Demultiplex one of the channels from the overall OTDM signal	Interferometric or four-wave mixing techniques based on fiber or semiconductor nonlinearity
Add/drop demultiplexer	Extract one channel from OTDM data signal, and insert a local channel into the empty time slot for transmission	Similar technical implementations to demultiplexer described above

likely that in the future, OTDM technology may be used for enhancing the overall capacity of optical communication systems.

One possibility for employing OTDM systems would be for upgrading installed WDM networks. The current method for improving WDM networks involves using more wavelength channels, but it may also be feasible to increase the overall capacity by developing hybrid WDM/OTDM communications systems. In such a system, each wavelength channel may carry aggregate data rates in excess of 100 Gbit/s, achieved using OTDM. The design of these hybrid networks would permit switching of the optical data signals to be carried out at two different levels in the overall network. The WDM signals could be switched coarsely using passive filtering devices, and the OTDM data channels could be switched with fine granularity using one of the all-optical switching techniques available.

List of Units and Nomenclature

Dispersion	[ps km^{-1} nm^{-1}]
Transmission data rate	[bit s^{-1}]
ADM	Add drop multiplexer
CW	Continuous wave
FWM	Four-wave mixing
NOLM	Nonlinear optical loop mirror
OTDM	Optical time division multiplexing
SDH	Synchronous digital hierarchy
SOA	Semiconductor optical amplifier
STM	Synchronous transport module
TOAD	Terahertz optical asymmetric demultiplexer
WDM	Wavelength division multiplexing

See also

All-Optical Signal Regeneration. **Optical Communication Systems:** Basic Concepts; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1997) *Fibre Optic Communications*, 2nd edn., pp. 329–335. New York: Wiley.

Cole M (2000) *Introduction to Telecommunications: Voice, Data, and the Internet*, chapter 5. Englewood Cliffs, NJ: Prentice Hall.

Islam M (1992) *Ultrafast Fiber Switching Devices and Systems*. Cambridge, UK: Cambridge University Press.

Kawanishi S (1998) Ultrahigh-speed optical time-division-multiplexed transmission technology based on optical signal processing. *IEEE Journal of Quantum Electronics* 34: 2064–2079.

Midwinter JE and Guo YL (1992) *Optoelectronics and Lightwave Technology*, pp. 148–178. Chichester, UK: Wiley.

Mukherjee B (1997) *Optical Communications Networks*, pp. 493–506. New York: McGraw-Hill.

Nakazawa M, Kubota H, Suzuki K, Yamada E and Sahara A (2000) Ultrahigh-speed long distance TDM and WDM soliton transmission technologies. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 363–396.

Prucnal PR, Santoro MA and Sehgal SK (1986) Ultrafast all-optical synchronous multiple access fiber networks.

IEEE Journal of Selected Areas of Communication 4: 1484–1493.

Spirit DM, Ellis AD and Barnsley PE (1994) Optical time division multiplexing: systems and networks. *IEEE Communications Magazine* 32(12): 56–62.

Spirit DM, Blank LC and Ellis AD (1995) Optical time division multiplexing for future high capacity network applications. In: Smith DW (ed.) *Optical Networking Technology*, pp. 54–77. London: Chapman and Hall.

Tucker RS, Einstein G and Korotky SK (1988) Optical time-division multiplexing for very high bit rate transmission. *IEEE Journal of Lightwave Technology* 6: 1737–1749.

Wavelength Division Multiplexing

J Bowers and H F Chou, University of California, Santa Barbara, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Wavelength division multiplexing (WDM) has enabled a revolution in communications technology. This article describes the technology, critical components of WDM systems, and transmission impairment and limits to transmission distance and capacity.

Multiplexing

There are three techniques used in fiber optics to multiplex the information that is transmitted from one point to another. The first is space division multiplexing (SDM), which simply consists of having multiple fibers in a cable or conduit. In a typical installation, there may be between 8 and 1000 fibers in parallel. The second approach is time division multiplexing (TDM) (*see* **Optical Communication Systems:** Optical Time Division Multiplexing), which consist of interleaving multiple data streams into one higher bit rate stream. For example, 672 voice channels (DS0) at 64 kbit/s are multiplexed into one T1 data stream at 1.544 Mbit/s. Eighty-four of these T1 channels are then multiplexed into a SONET (Synchronous Optical NETwork) OC-3 signal at 155 Mbit/s. The third multiplexing approach is wavelength division multiplexing (WDM), which consists of sending multiple signals at different wavelengths. These signals can be used to fill up the entire low loss transmission band of an optical fiber (Figure 1). The low loss transmission band extends from 1200 nm to 1600 nm, which corresponds to over 40 THz. One approach is to pack the signals densely together, as in dense wavelength division multiplexing (DWDM). This is a relatively expensive, but very high capacity approach. The other approach is to space the signals widely apart, (coarse WDM or CWDM) resulting in very low cost components, but with fairly limited transmission capacity.

Dense Wavelength Division Multiplexing (DWDM)

Most DWDM systems use multiple beams spaced at 100 GHz spacing centered at 193.1 GHz as defined by an International Telecommunications Union (ITU) standard (Figure 2a). Some higher capacity systems use finer spacing, 50, 25, or even 12.5 GHz spacing to pack more information into the fiber. The 100 GHz spacing used in DWDM systems corresponds to 0.8 nm spacing at 1.55 mm. In principle, 40 THz of information can be sent down an optical fiber. If high spectral efficiency (~1 bit/Hz) is maintained across the band, then this would correspond to 40 Tbit/s. Practical DWDM systems are limited by the bandwidth of erbium-doped fiber amplifiers (EDFAs) and by signal impairments discussed below. Consequently, most systems utilize a single band, typically the C-band extending from 1530 nm to 1565 nm. Additionally some transmission systems use multiplexers to combine other bands, typically C and L bands for additional transmission capability. The wavelength definitions are shown in Figure 1. Subbands are often used within a band to facilitate adding and dropping a group of wavelengths within a band. Typically 40 10 Gbit/s signals spaced at 100 GHz are used to fill the C-band. Higher capacity

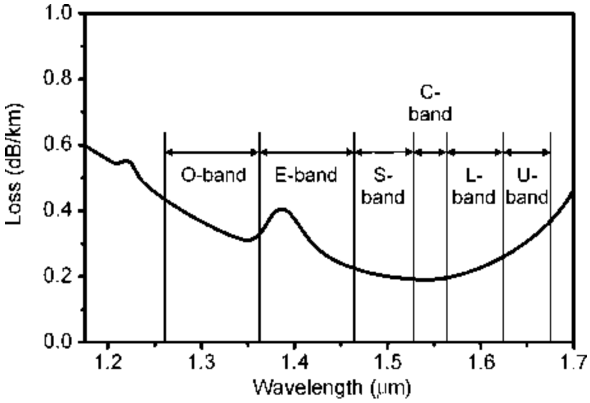

Figure 1 Transmission spectrum of optical fiber. Also shown are the standard transmission bands.

systems with 80 to 200 wavelengths have also been deployed commercially.

The historical evolution of optical transmission systems can be seen in Figure 3. The lower line shows the increasing bit rate demonstrated in experimental TDM systems, which began in 1978 with 45 Mbit/s systems and has now reached 1.28 Tbit/s. The upper line shows what has been demonstrated with DWDM with over 10 Tbit/s transmission demonstrated.

Coarse Wavelength Division Multiplexing (CWDM)

The fine frequency requirements of DWDM systems result in significant cost for transmitters, multiplexers and other components. A low-cost approach is to space the colors widely apart. The ITU standard is 20 nm spacing from 1271 to 1611 nm (Figure 2b). Note that the spacing is uniform in wavelength, but varies in frequency because the frequency spacing is related to the wavelength spacing by

$$\Delta f = -\frac{c}{\lambda^2}\Delta\lambda \qquad [1]$$

The wide spacing of signal wavelengths means that temperature stabilization of elements is not needed and the yield of components is high, resulting in a lower-cost system.

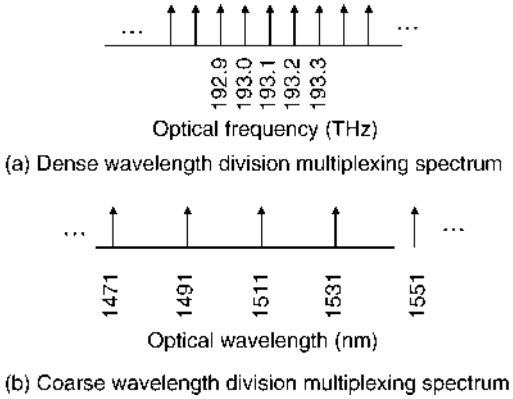
...
...
192.9
193.0
193.1
193.2
193.3
Optical frequency (THz)
(a) Dense wavelength division multiplexing spectrum
...
...
1471
1491
1511
1531
1551
Optical wavelength (nm)
(b) Coarse wavelength division multiplexing spectrum

Figure 2 (a) DWDM spectrum (b) CWDM spectrum.

Point-to-Point DWDM Links

Point-to-point DWDM links were widely deployed during the 1990s. The essential elements are shown in Figure 4, namely single frequency lasers, optical multiplexers and demultiplexers, optical amplifiers, optical add drop multiplexers (OADM), and optical receivers. These will be described in more detail in subsequent sections. Essential parts of the design of these links are managing the signal-to-noise ratio and managing the pulse broadening. With periodic amplification in EDFAs and dispersion compensation, links of 1000 km are common. The widespread use of forward error correction (FEC) circuits allows error-free transmission over a link with limited signal-to-noise ratio. With careful design, ultra-long-haul

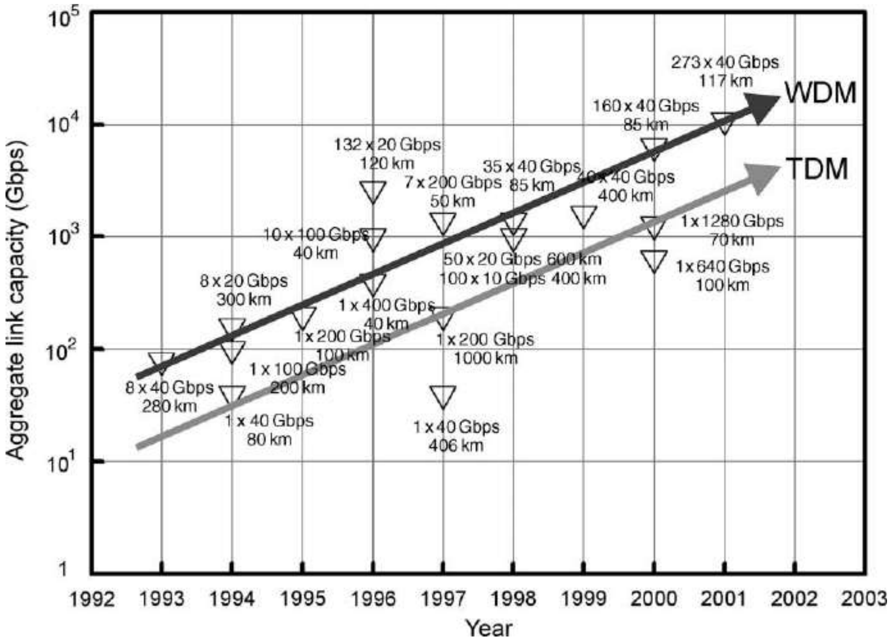

Figure 3 Historical evolution of transmission systems.

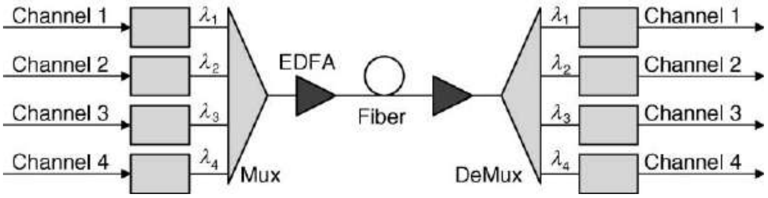
Channel 1
Channel 2
Channel 3
Channel 4
λ_1
λ_2
λ_3
λ_4
Mux
EDFA
Fiber
DeMux
λ_1
λ_2
λ_3
λ_4
Channel 1
Channel 2
Channel 3
Channel 4

Figure 4 Point-to-point WDM link.

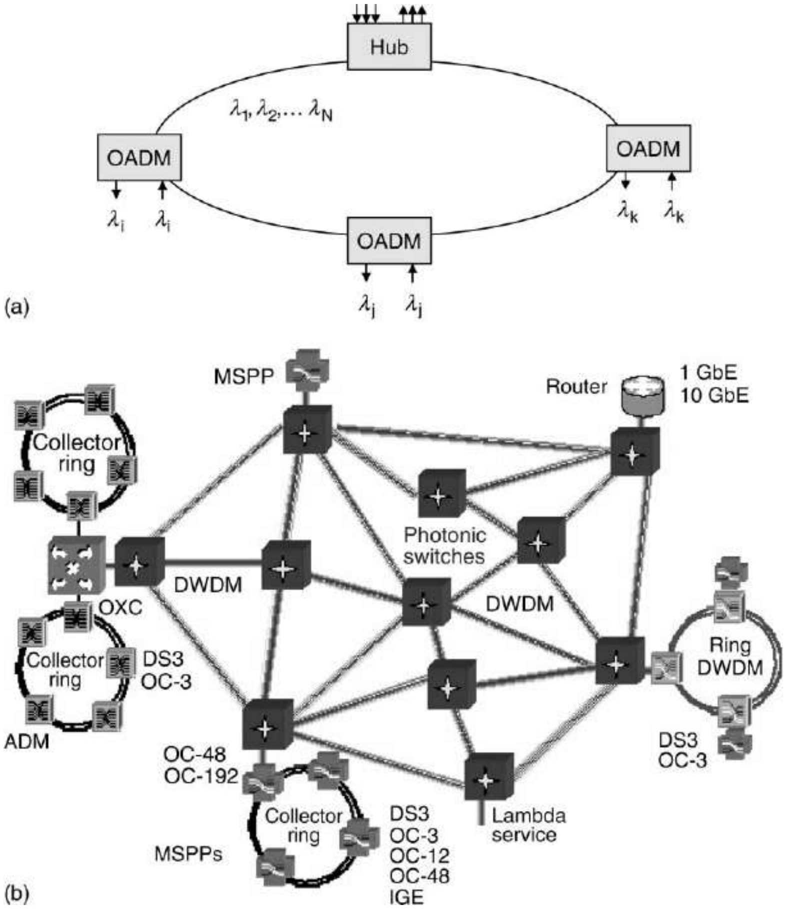
Hub
$\lambda_1, \lambda_2, \ldots \lambda_N$
OADM
λ_i λ_i
OADM
λ_k λ_k
OADM
λ_j λ_j
(a)
MSPP
Router
1 GbE
10 GbE
Collector ring
Photonic switches
DWDM
OXC
DWDM
Collector ring
DS3
OC-3
ADM
Ring DWDM
DS3
OC-3
OC-48
OC-192
Collector ring
MSPPs
DS3
OC-3
OC-12
OC-48
IGE
Lambda service
(b)

Figure 5 (a) WDM ring network. (b) DWDM mesh networks (after Basch B, Gringeri S, Goudreault R and Ravinkumar S (2002) Evolution of the photonic core within metropolitan and regional networks. *NFOEC*. Technical Proceedings).

transmission over 5000 km is possible, even with many 10 Gbit/s channels.

DWDM Networks

Point-to-point DWDM links are widely deployed in opaque networks with electronic regeneration and switching at nodes. The next stage in optical networks is the deployment of DWDM transparent ring and mesh networks. An example of a simple ring network is shown in Figure 5a where most of the light signals transit nodes transparently. The OADM elements drop particular wavelengths at each node. As more wavelengths are deployed in these ring networks, it becomes necessary to dynamically reallocate wavelengths and resources. Reconfigurable optical add/drop multiplexers (ROADMs) and optical switches have been developed to allow for dynamic transparent mesh networks (Figure 5b). This allows service providers to provision their networks remotely and quickly adapt to changing demand and restore vital connections when service interruptions occur.

Critical WDM Components

The previous sections described WDM networks. This section describes in more detail the critical components used in these networks.

Multiplexers/Demultiplexers

Multiplexers (MUXs) are used to combine multiple wavelength channels into a single fiber. They are usually wavelength selective and utilize wave phenomena such as diffraction and interference. A simple cascade of 3 dB couplers is not a practical multiplexer since the output power decreases exponentially with channel count. Due to the reciprocity of electromagnetic waves, most MUXs can also serve as a DEMUX. For high channel count systems, a multistage design may be necessary to overcome the limitations of a single MUX/DEMUX. This approach also provides a more flexible upgrading plan. In this case, wavelength channels can be grouped into bands or interleaved. Diffraction gratings and arrayed waveguide gratings (AWGs) are among the available technologies for MUX/DEMUX.

In a diffraction grating, narrow slits or micro structures are spaced evenly on the grating plane and serve as a secondary light source for the input light. The transmitted or reflected light (depending on the design) forms a diffraction pattern on the image plane. The locations of peak intensity on the image plane depend on the wavelength and the grating pitch. Therefore, wavelength channels can be separated in space and fibers can be positioned properly to couple individual channels. An AWG is a planar integrated and digitized form of diffraction grating as shown in Figure 6. It is usually fabricated on a silicon substrate with silica waveguides (silica-on-silicon). The input light is coupled into an array of waveguides with incremental difference in length. These tributaries are then combined in a free-propagating output coupler to several output waveguides. Due to the interference of these tributaries from the arrayed waveguides, different wavelength channels are focused to different output waveguides.

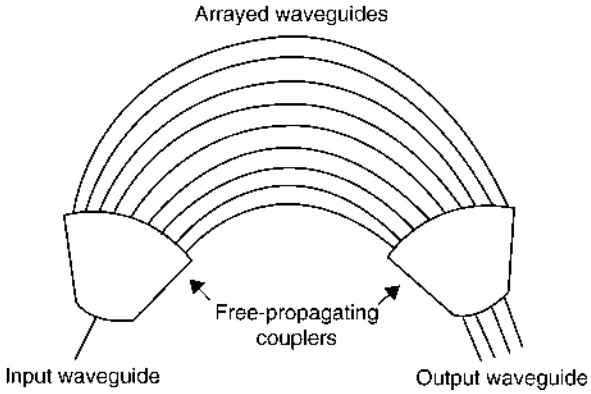

Figure 6 Schematic diagram of an AWG.

Optical Add/Drop Multiplexer

An optical add/drop multiplexer (OADM) provides critical networking capability for WDM systems. In an OADM, some of the wavelength channels are dropped from the input and are either terminated or routed. New channels at these wavelengths are then added with the through channels. This function can be done with the combination of a MUX and a DEMUX of any kind as shown in Figure 7a. However, if only one wavelength channel needs to be processed while others pass through, a simpler configuration can be used where a circulator drops the reflected wavelength from the fiber Bragg grating and a new channel is added with a coupler (Figure 7b).

The two approaches described above are termed static since the wavelength processed is fixed. To enable a more flexible network, reconfigurable OADM (ROADM) are required, where the wavelengths to be added or dropped can be dynamically specified. Figure 7c shows a configuration of ROADM based on a high port-count optical switch and a set of MUX/DEMUX. A tunable laser is highly desirable in this scenario since the wavelength of the added channel can be reconfigured to resolve wavelength contention in the network.

Optical Amplifiers

Erbium-doped fiber amplifiers (EDFA) (*see* **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems) provide simultaneous amplification of wavelength channels in the entire C-band (1530 to 1565 nm) and part of the L-band (1565 to 1605 nm). EDFAs were a key enabling technology for the deployment of WDM systems, especially for long-haul transmission where costly regenerators were eliminated. Both 980 nm and 1480 nm pumps can be utilized to achieve gain in erbium-doped fibers. The 980 nm pumps provide low-noise and high-gain operation while the 1480 nm pumps are more favorable for high-output power. One of the most significant advantages of an EDFA over a semiconductor optical amplifier (SOA) is that crosstalk among wavelength channels is much smaller in EDFA due to a millisecond-scale spontaneous emission lifetime. One critical issue for EDFAs is that the gain is not flat over the operation wavelength range and peaks around 1532 nm. However, this can be equalized by specially designed filters.

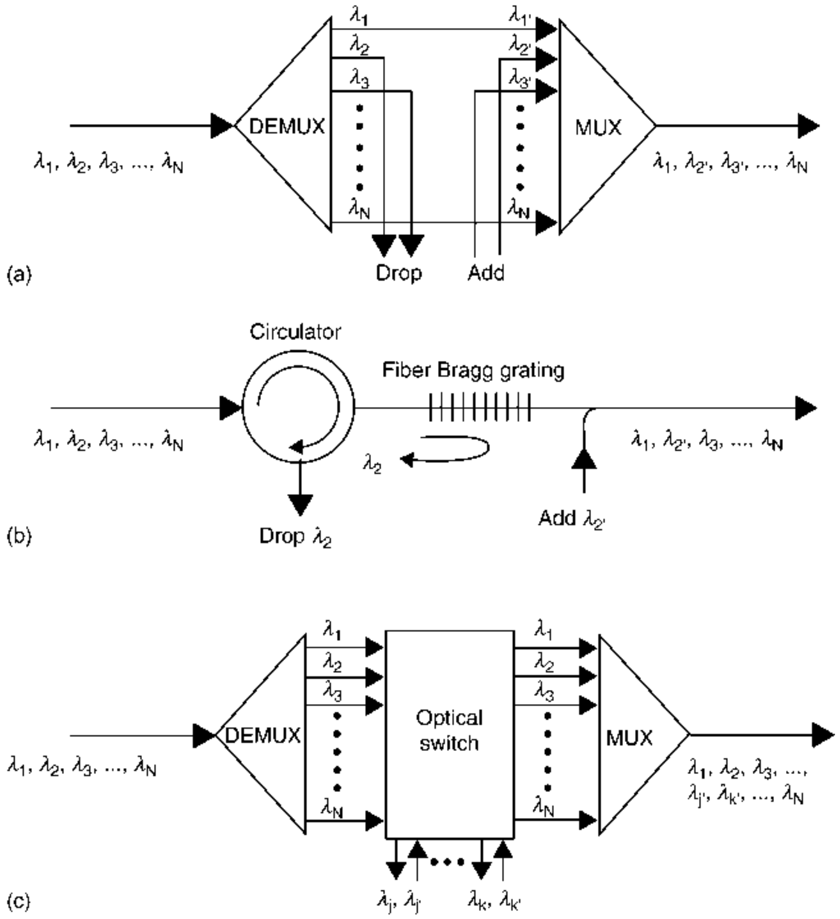
λ_1
λ_2
λ_3
λ_N
$\lambda_{1'}$
$\lambda_{2'}$
$\lambda_{3'}$
λ_N
DEMUX
MUX
$\lambda_1, \lambda_2, \lambda_3, ..., \lambda_N$
$\lambda_1, \lambda_{2'}, \lambda_{3'}, ..., \lambda_N$
Drop
Add
(a)
Circulator
Fiber Bragg grating
$\lambda_1, \lambda_2, \lambda_3, ..., \lambda_N$
λ_2
$\lambda_1, \lambda_{2'}, \lambda_3, ..., \lambda_N$
Drop λ_2
Add $\lambda_{2'}$
(b)
λ_1
λ_2
λ_3
λ_N
DEMUX
Optical switch
MUX
$\lambda_1, \lambda_2, \lambda_3, ..., \lambda_N$
$\lambda_1, \lambda_2, \lambda_3, ..., \lambda_{j'}, \lambda_{k'}, ..., \lambda_N$
$\lambda_j, \lambda_{j'}$
$\lambda_k, \lambda_{k'}$
(c)

Figure 7 (a) A parallel OADM architecture which allows multiple simultaneous add's and drop's; (b) a linear OADM architecture for processing a single channel; (c) an ROADM which enables reconfigurable processing of multiple wavelength channels.

Outside the gain spectrum of EDFAs, Raman amplifiers can be used, where the gain peak is about 13 THz (100 nm) below the frequency of the pump signal. Also, Raman amplification occurs in the same fiber used for transmission so that it can provide a distributed gain over the span of the fiber, which results in a better signal-to-noise ratio. This is particularly important for ultra-long-haul transmission. A counter-propagating pump scheme is preferable to reduce the influence of pump noise and depletion. The biggest challenge for Raman amplifiers is the availability of high-power pump lasers at the required wavelengths.

Optical Filters

Filters can find many applications in WDM systems, such as in multiplexers, demultiplexers, OADM, receivers and optical amplifiers. Fiber Bragg grating, Fabry–Perot filter, and multilayer thin-film filters are among the most widely used filters and utilize the interference property of optical waves. There are some general requirements for filters in WDM systems. First, the top of the pass-band must be flat enough so that cascading of filters in the network does not lead to a significant decrease of bandwidth. In addition, the roll-off of the pass-band must be sharp so that cross-talk from other channels is minimized.

A fiber Bragg grating has a periodic perturbation of refractive index made by exposing a photosensitive fiber to interfering ultraviolet lights. These perturbations reflect the input light as it propagates. At a specific wavelength, these reflected beams add in phase, which results in a strong overall reflection. The reflections at other wavelengths do not add up and these wavelengths will eventually pass through the grating. In short, the fiber Bragg grating acts as a wavelength-selective reflector and is usually used with a circulator or isolator to manage the reflected light. A Fabry–Perot filter is composed of two partially transmitting interfaces and a cavity in between. The underlying mechanism is similar to the fiber Bragg grating but the transmission is periodic in optical frequency. The spacing is called the free spectral range, which depends on the optical length of the cavity. The bandwidth of the passband is a strong function of the reflectivity of the partially transmitting interfaces. A multilayer thin-film filter is an extended form of Fabry–Perot filter where several layers of dielectric reflectors are formed with dielectric cavities in between. This allows for more tuning dimensions in the design of the passband shape in order to meet the requirements of WDM systems and the temperature tolerance is better than other kinds of filter.

Laser Sources

Semiconductor lasers (*see* **Lasers:** Semiconductor Lasers) are widely used in optical communication because they are compact, low cost and efficient. As a light source for transmission, single frequency operation is necessary for high-capacity WDM systems. Multilongitudinal mode lasers such as Fabry–Perot lasers have excess bandwidth which causes excessive penalty due to dispersion unless the transmission distance is short. They also result in crosstalk in DWDM systems. The most widely used single-frequency laser in WDM systems is the distributed-feedback laser (DFB), where a corrugated grating structure is fabricated on top of the gain region and only the wavelength which satisfies the Bragg condition can oscillate in the laser cavity and all other modes are suppressed.

As WDM networks advance, tunable lasers are strongly desired to reduce the excessive inventory of fixed wavelength lasers for DWDM and also to provide reconfigurability in the network. One possible structure is the distributed Bragg reflector (DBR) laser, in which the Bragg grating is decoupled from the gain region and the Bragg wavelength can be tuned independently by isolated current injection. More complicated variations of this concept such as the three-section DBR laser and the sampled grating DBR laser can provide more continuous and wider tuning in wavelength.

Photonic Switches

Photonic switches are necessary in transparent networks to allow reconfigurability in networks, and allow rapid provisioning and fault recovery. They provide connectivity between rings and between point-to-point links without requiring the DWDM signal to be broken into individual wavelengths and to be electronically regenerated. Photonic switches can be used to switch fibers, wavebands, or individual wavelengths. They are inherently transparent to bit-rate and data format. A number of different technologies have been developed.

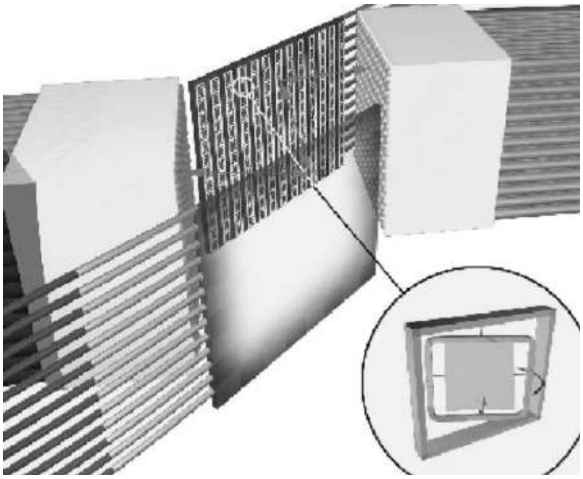

Figure 8 3D MEMS optical switch.

Microelectromechanical system (MEMS) switches are widely used for 2 × 2 protection switches, wavelength selective switches and for nonblocking crossbar switches with sizes from 8 × 8 to 1024 × 1024. The technology usually used for large switches is called 3D MEMS switches (Figure 8), in contrast to 2D technologies where the optical beams are confined to a two-dimensional plane.

Transmission Issues

Chromatic Dispersion

Chromatic dispersion (*see* **Fiber and Guided Wave Optics:** Dispersion) means that the group velocity, which is the propagation speed of an optical signal, changes with wavelength. Since every real pulse has a finite bandwidth, each slice of wavelength within this bandwidth propagates with a different speed and results in the distortion of the pulse as it propagates along the fiber. Depending on the sign of the dispersion and the initial phase distribution, the pulse can either be broadened or compressed. However, if the fiber is long and nonlinear effects are small, the pulses will eventually be broadened and result in intersymbol interference. Dispersion compensating fibers (DCFs), which have an opposite sign of dispersion, can be used to reverse this process and restore the original pulse shape. In general, it is advantageous to minimize the length of DCF by keeping the dispersion low. Dispersion-shifted fibers (DSF) have a zero-crossing of dispersion around 1.5 μm and hence a reduced dispersion for the C-band.

For a WDM system, a very low dispersion will result in more severe nonlinear impairments since the distance pulses can overlap with each other along the fiber is long. Nonzero-dispersion-shifted fibers

(NZ-DSF), which introduce a suitable amount of dispersion, can be used instead to solve this issue. Chirped fiber Bragg grating can also be used, where different wavelengths are reflected at different positions in the grating so as to compensate the dispersion caused by transmission. It is capable of compensating a given amount of dispersion with a much shorter length when compared to DCF. For high channel count WDM systems, the dispersion experienced by each channel may be different and dispersion compensation needs to be implemented on a per-channel basis after a demultiplexer.

Polarization Mode Dispersion

Polarization mode dispersion (PMD) means that the group velocity dispersion of the two orthogonal polarizations in the fiber are different and cannot be simultaneously compensated by a single DCF. It is usually manifested as the splitting of a pulse at the receiver even after chromatic dispersion compensation and causes intersymbol interference. PMD can originate from the imperfections in the fiber or environmental factors like temperature and mechanical pressure. Therefore, the amount of PMD can vary with time for a given link, which means that dynamic compensation is necessary. Traditional electrical equalization after the optical signal is detected can be used for compensation but the highest bit-rate achievable is limited by the speed of electronics. Optical compensation is another choice. The optical signal is split into fast and slow components and the fast component is delayed properly before being recombined with the slow component.

Nonlinear Impairments

The most important nonlinear impairments for WDM systems are the cross-phase modulation (CPM) and the four-wave mixing (FWM). They are both third-order nonlinear effects, where the nonlinear dielectric polarization is proportional to the third power of the electric field of the signal. In CPM, the phase of a wavelength channel can be modified by other channels, which increases the chirp and results in more dispersion penalty. On the other hand, FWM can generate new waves at other wavelengths, causing unwanted crosstalk if the generated wave happens to coincide with another channel. The existence of dispersion in fiber can help to reduce the effects of both impairments since the pulses at different wavelengths propagate at different speeds and walk away from each other so that nonlinear interactions are reduced. This is why NZ-DSF is preferred over DSF. Self-phase modulation (SPM) is another third-order nonlinearity, where the phase distribution of a pulse is altered by its intensity profile. It will lead to the distortion of the pulse. Other nonlinear effects include stimulated Brillouin scattering and stimulated Raman scattering. Both originate from the interactions of phonons and photons in the fiber and cause power fluctuations at different wavelengths.

Forward Error Correction

To reduce the errors of a transmission link, extra information relating to the data can be sent by the transmitter and used by the receiver to detect and correct errors in the received data. This technique is called forward error correction (FEC) and imposes redundancy to the transmission which means that the bit-rate transmitted will be higher than the actual data rate. A measure of performance for FEC is the coding gain, which is defined as the decrease in receiver sensitivity at a given bit-error-rate. Reed–Solomon codes are one of these codes. With a 7% redundancy, up to 6 dB of coding gain can be achieved. This could correspond to more than 9 orders of magnitude improvement in bit-error-rate. Therefore, the use of FEC can increase the spacing between regenerators in ultra-long-haul transmission and reduces the cost. It is also useful to overcome error floor problems due to crosstalk in a WDM system, where a further increase in received power cannot reduce the bit-error-rate. Crosstalk among wavelength channels can happen due to nonlinear impairments during fiber transmission or due to imperfect channel isolation in components like multiplexers, demultiplexers, filters, and switches. The use of FEC can greatly increase the system margin.

Conclusions

WDM technology has enabled a huge increase in the transmission capacity of optical fibers and a huge reduction in the cost of fiber optic transmission since the cost of the installed fiber is shared between the signals traveling on different wavelengths. As new components have been developed, WDM technology has expanded simple point-to-point links to transparent networks. Great progress has been made over the past 20 years to solve transmission impairments and extend the reach and bandwidth of these systems. This is likely to continue to meet the increasing communication needs of our society.

List of Units and Nomenclature

Distance (km, m, μm)
Bit rate (bit/s)
Wavelength (nm)

See also

Color and the World. **Information Processing:** All-optical Multiplexing/Demultiplexing. **Fiber and Guided Wave Optics:** Optical Fiber Cables; Passive Optical Components. **Lasers:** Edge Emitters; Semiconductor Lasers. **Modulators:** Modulation and Demodulation of Optical Signals. **Optical Amplifiers:** Basic Concepts; Optical Amplifiers in Long-Haul Transmission Systems; Semiconductor Optical Amplifiers; **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Basic Concepts; Historical Development; Lightwave Transmitters; Optical Time Division Multiplexing. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2002) *Fiber Optic Communication Systems*, 3rd edn. New York: John Wiley & Sons Inc.

Basch B, Gringeri S, Goudreault R and Ravinkumar S (2002) Evolution of the photonic core within metropolitan and regional networks. *NFOEC*. Technical Proceedings.

Green PE (1993) *Fiber Optic Networks*. Upper Saddle River, NJ: Prentice Hall.

Ramaswami R and Sivarajan KN (2002) *Optical Networks: A Practical Perspective*, 2nd edn. San Francisco, CA: Morgan Kaufmann Publishers.

OPTICAL MATERIALS

Contents

Color Filter and Absorption Glasses
Heterogeneous Materials
Lightweight Mirrors
Measurement of Optical Properties of Solids
Optical Glasses
Plastics
Sculptured Thin Films
Smart Optical Materials
Sol-Gel Materials

Color Filter and Absorption Glasses

J E Shelby, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical filter and absorbing glasses can be categorized by either the optical behavior of the filter, for example, bandpass, shortpass or longpass, neutral density, or interference filters – or by the method used to filter the light, i.e., glasses which absorb light themselves and those which serve as high quality, transparent substrates for coatings which either absorb or reflect light. The term 'filter' usually refers to glasses used in optical systems, to either transmit or prevent transmission of specific wavelengths. The term 'absorber' is usually reserved for glasses which are used for controlling light to provide protection from damage to humans, for example, for eye protection, as in sunglasses, or welder and laser goggles, or to materials exposed to light, for example, plastics and other components in the passenger compartment of automobiles. Manufacturing technology is normally based on the method used to produce the device, such as control of absorbing species in the glass or method of application of optical coatings. Terminology for optical applications is almost always based on optical behavior and not on the method used to produce that behavior.

Bandpass Filters

The first glasses used to filter light were colored by inclusion of ions from the first row of transition metals. These colored glasses still provide most of

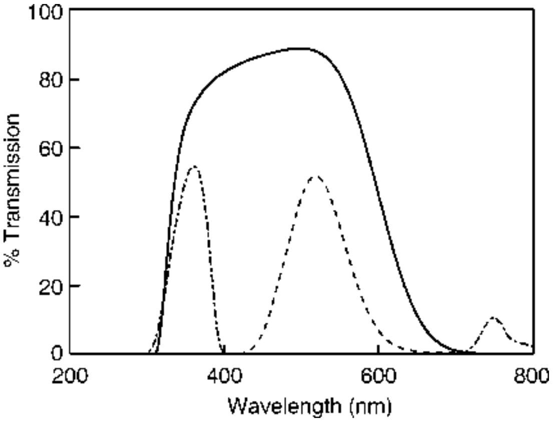

Figure 1 Transmission spectra for bandpass filters with a large bandwidth (solid line), medium bandwidth (dashed line), and narrow bandwidth (dot dash line).

the glasses used to control the wavelength of light. They are widely used as absorbers in windows in buildings and automobiles, to control the entrance of ultraviolet and infrared light into these environments. These filters allow the passage of light over a significant range of wavelengths, while absorbing wavelengths outside that range. They are designated by the peak transmission wavelength, or central wavelength, and the bandwidth, or wavelength range, over which light is transmitted with at least some designated transmission value. Filters based on absorption by transition metal or rare earth ions usually have larger bandwidths and somewhat diffuse cutoffs at the edges of the transmitted spectral region (Figure 1).

The most common bandpass filters absorb radiation in a desired wavelength region. Historically, the first bandpass filters, which are still the most common, relied on absorption of light as a result of ligand field (also known as crystal field) effects and/or charge transfer processes within a glass, to create absorption bands in the spectrum.

Ligand field effects will produce optical absorption bands in glasses which contain transition metal (usually row 3d) or rare earth (lanthanide, or 4f) ions. The differences in the electronic structures of the transition metal and rare earth ions result in very different absorption spectra. Absorption spectra, due to transition metal ions, are characterized by broad bands and are frequently accompanied by charge transfer bands, due to the same ions, while the bands due to absorption of light by rare earth ions are very sharp and are not accompanied by charge transfer absorption bands.

Iron oxide was probably the first, and is still the most widely used additive for production of glasses which primarily transmit light in the visible portion of the spectrum. A large, broad absorption band due to ferrous ions is centered at ≈1050 nm in common soda-lime-silica glasses. The tails of this band extend into the visible portion of the spectrum, allowing passage of the blue-green portion of the spectrum, but absorbing much of the red light, and to wavelengths in excess of 2000 nm, absorbing much of the near infrared spectrum.

Ferric iron produces relatively small ligand-field absorption bands in the region from 380 to 440 nm. These bands remove most of the blue part of the spectrum, allowing passage of light with longer wavelengths. Both ferrous and ferric ions also display charge transfer bands in the ultraviolet, effectively eliminating transmission of light with a wavelength <300 nm. Glasses containing enough iron oxide, to effectively eliminate both ultraviolet and near infrared transmission, exhibit various shades of green, depending on the ratio of ferrous to ferric ions present. Melting under strongly reducing conditions yields blue-green glasses, whereas melting under strongly oxidizing conditions yields yellow-green glasses. Glasses containing $\geqq 0.5$ wt% iron oxide are used in windows of automobiles to minimize the amount of ultraviolet and infrared radiation entering the passenger compartment.

Copper oxide is also used to remove a portion of the near infrared spectrum from sunlight and to produce glasses which transmit in the green and blue region. The primary ligand-field absorption band of copper is due to cupric ions and is located near 800 nm. This band in similar in shape to that due to ferrous ions, with tails extending into the visible and near infrared regions of the spectrum. The shift in band position toward the visible, however, results in less absorption in the infrared and more in the visible than for ferrous iron. As a result, glasses containing copper are more strongly colored than those containing ferrous iron and are less effective filters for infrared light. Glasses colored with cupric ions can be used to produce a transmission window from about 350 to 600 nm in soda-lime-silica glasses. Cuprous ions do not create any significant absorption bands.

Glasses containing hexavalent chromium, which must be produced in a strongly oxidizing environment, effectively remove ultraviolet light between 300 and 400 nm, but allow transmission of ultraviolet light in a narrow range centered around 300 nm. A small tail of the charge transfer band due to Cr^{6+}extends into the visible, producing a yellow-green glass. The presence of a set of much weaker ligand-field bands, centered around 650 nm due to trivalent chromium, also contributes to the green color of these glasses. Combining chromium with copper ions in a glass yields a filter with a small transmission window centered at about 520 nm.

Combination of chromium with iron results in suppression of both hexavalent chromium and ferrous iron, yielding a green glass without either the strong ultraviolet absorption of hexavalent chromium or the strong near infrared absorption of ferrous iron.

Manganese oxide can be used to produce bandpass filters which block most of the visible, but allow transmission of ultraviolet and infrared light. Addition of manganese ions to a glass results in a set of absorption bands with a maximum intensity at about 500 nm and a tail extending to around 900 nm. Since these ions do not absorb in the ultraviolet, the transmission spectra of these glasses has a window between about 300 and 400 nm, which overlaps the transmission window due hexavalent chromium. The absorption bands due to manganese ions are relatively weak, so that concentrations of manganese oxide in excess of 1 wt% are needed to effectively block visible transmission.

Cobalt oxide is the strongest colorant of the oxides of 3d transition metals. A small addition of cobalt oxide to silicate glasses creates strong absorption bands between about 450 and 700 nm. The optical window of these glasses in the ultraviolet is centered near 380 nm and is somewhat broader than that of glasses containing manganese oxide, which have similar spectra. Since cobalt oxide is such a strong absorber, black glasses are easily produced by use of cobalt oxide in combination with other oxides. If cobalt oxide is combined with nickel oxide, the resulting glass retains a bandpass in the ultraviolet, while combining cobalt oxide with iron oxide, which has a complementary absorption spectrum, yields a black glass which also effectively eliminates transmission in the ultraviolet and near infrared.

The optical spectra of glasses containing most rare earth ions are characterized by sharp absorption bands. Some rare earth ions create spectra with many absorption bands, while others display only a few bands. The band structures which result from rare earth ions are not particularly useful for creation of bandpass filters, although Yb^{3+} ions do effectively remove light around 1100 nm, without absorbing light in the rest of the spectrum.

While not commonly used for filters, glasses containing rare earth ions have been used for wavelength calibration standards for uv-vis spectrophotometers. A common wavelength standard called a ‘didymium’ glass contains a combination of neodymium and prasedymium oxides. Since the optical spectrum from each of these ions contains many sharp bands, these glasses provide known wavelengths for absorption bands over a wide region of the spectrum. Neodymium ions, which have absorption bands near the D line of sodium, are used in goggles for eye protection for glassblowers, to remove the intense yellow light resulting from the sodium in the glass.

Interference Filters

Interference filters are a special case of bandpass filters which are created by application of dielectric films to glass substrates. The dielectric films are deposited on a glass substrate in alternating layers of high and low refractive index materials. The combination of refractive indices and thickness of the dielectric films determines the wavelength regions where constructive and destructive interference occur, i.e., those wavelength regions which will transmit or reflect the incident light. Since the choice of materials will control the indices, production of filters for different wavelengths usually relies on deposition of layers of differing thicknesses.

These filters can be produced with very narrow bandwidths (±10 nm) or with broader bandwidths (commonly ±50 or ±80 nm). The peak transmission frequently does not occur at the midpoint of the bandwidth. Since the wavelengths for constructive and destructive interference depend upon the path length through the dielectric layers, these filters are extremely dependent upon the incident light angle, with a shift toward lower wavelengths for incident angles other than normal to the filter surface. Interference filters typically reduce throughput of light as compared to absorption-based bandpass filters, but provide for much narrower transmission windows. If the dielectric layers are deposited in such a manner that their thickness varies along the length of the substrate, interference filters can be produced with a continually varying central transmission wavelength along the length of the substrate. While convenient for applications, which might otherwise require several filters of different central wavelengths, these linearly variable interference filters are much more expensive than those designed to block specific wavelengths.

Dispersion Filters

Dispersion filters are based on the passage of light through a medium consisting of a matrix material and dispersed, very small particles of a material which has a very different dispersion curve. The original version of these filters, called a Christiansen filter, consists of small glass particles with a controlled particle size suspended in a liquid. The dispersion curve for the glass is much less wavelength-dependent than that of the liquid. Transmission of light occurs at the wavelength where the refractive index of the glass and the liquid are equal, while light of other wavelengths is scattered. Since the refractive index of a typical liquid is highly temperature sensitive, the transmitted, or peak, wavelength can be shifted over a wide region by controlling the temperature of the mixture. The bandwidth of dispersion filters can be very narrow (<± 10 nm).

The large temperature dependence of the traditional dispersion filter can be useful in that it provides an optical filter with an adjustable peak wavelength. On the other hand, this sensitivity to temperature also results in a serious control problem, since a temperature change of only a few degrees can shift the transmission window by up to 100 nm. This problem can be decreased or even eliminated by making a solid dispersion filter using two glasses of different dispersions, with one glass distributed in the other, or a glass dispersed in some other transparent solid such as a polymer with a similar thermal expansion coefficient. Since the refractive index of solids tends to vary with thermal expansion coefficient, these solid state dispersion filters can be almost temperature independent.

Neutral Density Filters

Neutral density filters are used to reduce the throughput of light to some desired value over a wide spectral region. Gray glasses are used to reduce the intensity of light from a bright source such as the sun. Black glasses are used to completely block light in the visible region. Neutral density filters can be produced using a combination of transition metal ions or by use of optical coatings. They are characterized by their percent transmission, range of wavelength where the transmission is constant, and the 'ripple', or spectral flatness, within the designated spectral region (Figure 2).

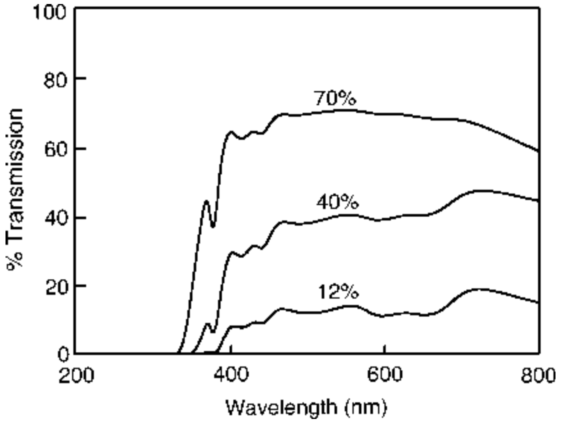

Figure 2 Transmission spectra for neutral density filters with different % transmission.

An approximation to a neutral density filter for visible light is often created by combining the absorptions in the ultraviolet and near infrared from ferrous and ferric ions, with the multiple absorption bands centered in the region of 500 to 600 nm due to divalent cobalt ions. While the resulting absorption spectrum displays small ripples throughout the region from 350 nm to beyond 1200 nm, the transmission of the glass can be controlled to within a few percent by the proper balance of ferric and ferrous ions combined with the desired concentration of CoO. These glasses are generally gray, or, if enough of iron and cobalt oxides are present, black. Combination of CoO with NiO can produce similar filters covering a smaller wavelength range which maintain transmission in the ultraviolet. Filters with different transmission values can be produced by either varying the composition of the glass or by use of different glass thicknesses. A continuously graded neutral filter can be produced from a single piece of glass by etching or polishing to form a wedge. With more effort, a step-graded filter can be produced by etching or polishing to form sharply delineated thickness steps along the length of a glass plate.

Neutral density filters can also be produced by coating a glass substrate with a metal film. Coated neutral density filters have less ripple than absorptive filters based on iron and cobalt oxides dissolved in the glass. In order to produce a constant transmission throughout the desired wavelength region, the metal used (commonly inconel) should have a constant reflectance across the desired wavelength region. Use of a metal with a varying wavelength dependence in its reflectivity, such as silver, will result in a wavelength dependent transmission. Continuously graded neutral filters can be produced by varying the thickness of the film along the length of a strip of glass.

Longpass/Shortpass Filters

Longpass filters are designed to pass light with wavelengths longer than the specified cutoff value. They are frequently used to block ultraviolet light while transmitting visible light or to block visible light while transmitting infrared light. Shortpass filters provide the opposite effect, i.e. they pass light with wavelengths shorter than the specified cutoff value, while blocking light with longer wavelengths. Shortpass filters are often used to block passage of heat (infrared light), while allowing transmission of visible light (Figure 3).

Ultraviolet blocking longpass filters can be produced by varying the basic composition of a glass or

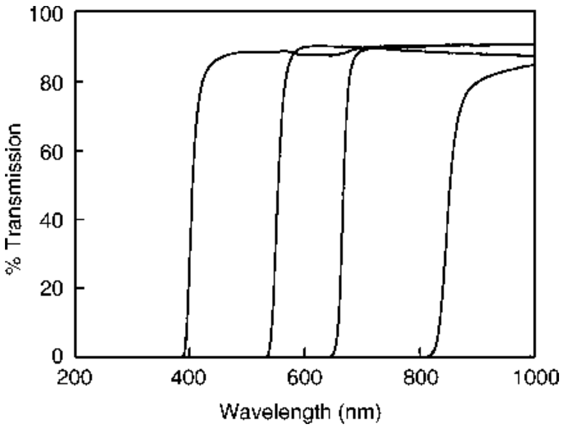

Figure 3 Transmission spectra for longpass filters with different cutoff wavelengths.

by addition of specific ions which absorb strongly in the ultraviolet. Varying the primary glassforming oxide can alter the ultraviolet cutoff frequency of a glass. Vitreous silica, for example, has an ultraviolet cutoff beyond 200 nm, so that it is a very effective transmitter of ultraviolet light. Vitreous germania, on the other hand, has an ultraviolet cutoff near 400 nm, so that it acts as an effective longpass filter for most ultraviolet light. While vitreous silica is widely used as an optical material, the poor chemical durability of vitreous germania prevents its use in most optical applications.

The ultraviolet cutoff wavelength of silicate glasses is shifted toward the visible by addition of alkali and alkaline earth oxides to silica. The addition of these oxides creates nonbridging oxygens in the vitreous network, which shift the ultraviolet cutoff toward the visible. These glasses are effectively opaque beyond their cutoff wavelength, creating longpass filter glasses with cutoff wavelengths which vary with the nonbridging oxygen concentration. Impurities such as iron or copper can also reduce the ultraviolet transmission of silicate glasses due to the creation of charge transfer bands, producing longpass filters with cutoffs in the range of 250 to 300 nm.

Longpass filters with tightly controlled ultraviolet cutoff wavelengths are often made by adding either cerium or a combination of cerium and titanium oxides to silicate glasses. Additions of cerium oxide result in a strong absorption band near 320 nm, which provides protection from ultraviolet light. Cerium also is routinely used to reduce radiation-induced coloration of glasses exposed to high energy radiation. An absorption band due to trivalent titanium also serves to cut off ultraviolet light, while producing a yellow glass. The use of a combination of cerium and titanium produces a very effective ultraviolet filter which eliminates transmission of almost all ultraviolet light.

Long bandpass absorption filters can also be based on precipitation of colloids of metals or semi-conductors in an otherwise colorless glass. The glasses are initially prepared in a colorless state, with the materials which will form the colloids dissolved as ions. The glasses are then reheated into the glass transformation region, where the desired ions are reduced to the atomic state and precipitated as colloids. Since this process is known as ‘striking’, these filters are known as striking filters. Colloids can be produced by striking from several elements, including silver, gold, copper, arsenic, antimony, lead, and bismuth, and from a number of compounds, including the cadmium chalcogenides and ZnS.

Formation of silver colloids in common soda-lime-silicate glasses results in a very strong optical absorption band centered near 400 nm. Glasses containing silver colloids are yellow, with a bandpass window in the ultraviolet. Since these glasses absorb so strongly at a wavelength commonly used to irradiate photosensitive polymers in the production of electrical circuit boards, they are frequently used for the production of beam masks. A pattern of silver colloids is easily produced in the surface of common float glasses by depositing metallic silver through a mask onto the surface of the glass and then heating the glass to 500 °C for a short time.

Copper can also be used to form longpass filters with a cutoff in the red portion of the spectrum by a similar process to that used to form the silver colloid filters. In this case, it is necessary to use a copper salt in place of the metallic silver, but the remainder of the process is essentially the same (Figure 4).

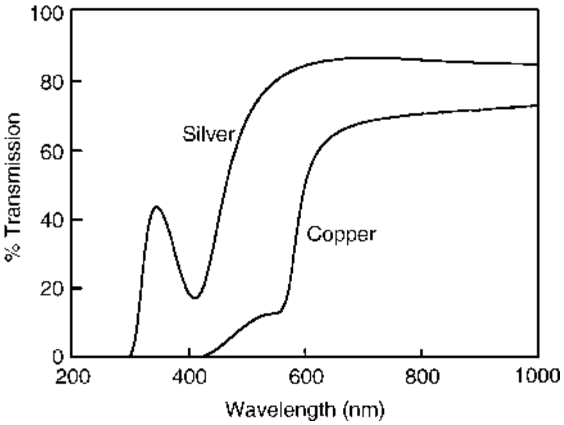

Figure 4 Transmission spectra for longpass filters containing silver or copper colloids.

Precipitation of semi-conducting colloids containing CdS, CdSe, and CdTe in silicate glasses results in sharp cutoffs in the range from the near ultraviolet to the near infrared. The cutoff is the direct result of the bandgap of these crystallites. Since these materials form a broad series of solid solutions, the desired cutoff wavelength can be obtained by use of an appropriate mixture of two of these compounds in the glass batch. Glasses containing only CdS, for example, have cutoffs in the range of 400 to 500 nm. Combining CdS with CdSe will shift the cutoff toward the infrared, with cutoffs in the range of 500 to 700 nm. Combining CdS with CdTe produces glasses with cutoffs near the edge of the visible region. Use of CdTe alone results in filters with cutoffs around 800 nm.

While the most common explanation for the shift in the cutoff in glasses containing the cadmium chalcogenides is based on solid solution arguments, it has been shown that the size of the particles also strongly affects the position of the cutoff for glasses containing CdSe. In this case, the energy gap, which determines the cutoff wavelength, varies inversely with the square of the particle size.

Glasses containing CdS and ZnS can be heat-treated to form filters with cutoffs in the range from 340 to 518 nm. Heat treatment at lower temperatures yields CdS colloids, with a cutoff around 500 nm. At higher heat treatment temperatures, the absorption edge shifts toward the ultraviolet, approaching the value of 340 nm representative of the lattice absorption of ZnS. This behavior has been interpreted as indicating that CdS crystals precipitate at lower temperatures, while at higher temperatures, solid solution crystals of ZnS and CdS are formed.

Colloids of arsenic, antimony, bismuth, and lead form less interesting filter glasses. The optical properties of these metals are such that there is no sharp cutoff wavelength. Growth of colloids produces wavelength-dependent scattering, with a long tail extending into the visible or infrared. These glasses can be produced in colors ranging from light tan to black by progressive heat treatments of the originally colorless glasses in hydrogen. The hydrogen serves to reduce the ions to the atomic state. The atoms then diffuse to form colloids. The growth process creates a sharp interface between the colored and colorless region of the glass, which can be used to produce windows for visible light of highly transmitting glass surrounded by essentially opaque glass.

Absorber Glasses

Glass filters are used to remove unwanted wavelengths from light. When the application is less demanding than those of complex optical devices or systems, the glasses are commonly called 'absorbers' rather than filters. These applications usually involve protection of humans or materials from damage by light, for example, sunglasses and goggles for welders and glassworkers, or to enhance the environment by controlling the entrance or exit of ultraviolet and infrared light. Most of these applications are variations on the filters discussed above.

The most common use of glass as an absorber for both protection from damage and enhancement of the environment is found in the soda-lime-silicate glasses used as windows in buildings and automobiles. The presence of a relatively large concentration of iron oxide provides absorption in both the ultraviolet and infrared regions of the spectrum. Since glasses containing iron do absorb some light in the visible, a compromise between total absorption of unwanted, nonvisible light and the desired visible light must be made. As a result, these glasses usually contain between 0.5 and 2 wt% iron oxide. Changes in melting conditions which alter the ratio of ferrous to ferric ions in the glass are used to tune the glass for specific applications. While these glasses are technically broad bandpass filters, their common applications are based on the favorable location and bandwidth of their transmission window, which roughly corresponds to the visible region.

Iron is also used for protection from ultraviolet and infrared radiation in welder goggles. These glasses contain much more iron (up to 8 wt% or more) than those used for windows, since the demand for protection from the intense light created in welding is much greater. The high absorption coefficient for infrared light results in considerable heating of the glass during lengthy use. A thin film of gold is often applied to the glass to reflect infrared radiation and reduce heating of the glass.

Other absorber applications are more demanding in terms of absorption wavelength. As mentioned earlier, glasses containing neodymium have a strong absorption band at the wavelength of the yellow sodium emission line. These glasses transmit visible light at other wavelengths, allowing their use by glassblowers, who need to be able to see the unheated part of the glass and their immediate surrounding area, while simultaneously reducing the intense yellow light given off by the sodium in the glasses normally used for glassworking.

Protection of the eye from laser radiation is even more demanding due to the very high intensity over very narrow wavelength ranges. Common ultraviolet longpass filters are usually satisfactory for absorption of light from ultraviolet lasers. Specific ions must be added to glasses to provide protection from many visible and near infrared lasers, since the glass must simultaneously permit transmission of visible light of other wavelengths and total absorption at the laser wavelengths. Light from lasers which operate in the far infrared is readily absorbed by most glasses, since the infrared cutoff of oxide glasses lies around 5 micrometers. Great caution should always be used in selecting glasses for eye protection from lasers. No assumptions regarding the degree of protection should ever be made without insuring that the glass is designed for the conditions under which it will be used.

See also

Optical Coatings: Thin-Film Optical Coatings. **Optical Materials:** Optical Glasses.

Further Reading

Bamford CR (1977) *Colour Generation and Control in Glass*. Amsterdam: Elsevier Scientific Publishing Company.

Deeg EW (1986) Optical glasses. In: Boyd DC and MacDowell JF (eds). *Commercial Glasses*, pp. 1–34. Columbus, OH: The American Ceramic Society, Inc.

Edmund Industrial Optics Catalog N017 (2001) pp. 68–87. Barrington, NJ: Edmund Industrial Optics.

Hillmann E and Neuroth N (1998) Glasses for eye protection. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 370–379. Heidelberg, Germany: Springer-Verlag.

Kolberg U (1998) Coloured glasses. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 350–370. Heidelberg, Germany: Springer-Verlag.

Neuroth N (1998) Transmission and reflection. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 82–95. Heidelberg, Germany: Springer-Verlag.

Paul A (1982) *Chemistry of Glasses*, pp. 204–276. London: Chapman and Hall.

Vogel W (1991) Optical properties of oxide glasses. In: Ullmann DR and Kreidl NJ (eds). *Optical Properties of Glass*, pp. 1–36. Westerville, OH: The American Ceramic Society, Inc.

Weyl WA (1951) *Coloured Glasses*. Society of Glass Technology: Sheffield, England.

Heterogeneous Materials

U Kreibig and M Quinten, Aachen University of Technology, Aachen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The term heterogeneous materials – being a quite diffuse classification of matter with discrete building units of different kinds, irrespective of their sizes – is defined here by scaling these inhomogeneities in relation to optical light wavelengths. Matter consisting of atoms in a regular structural order (or, alternatively, in fully statistical disorder) is classified as homogeneous since the inhomogeneities are too small to effect optical fields. This applies to crystalline, liquid, and glassy or amorphous matter likewise.

All geometrical/topological and chemical heterogeneities on a larger length-scale than the atomic scale, i.e., single crystallites, organic macromolecules, domains, extended lattice defects, clusters, and nanoparticles of chemical compositions differing from their host material, etc., result in a macroscopic material that is heterogeneous, in optical properties and optical experiments. Examples are multilayer-structures, heterogeneous in one dimension, three-dimensional heterogeneous composites, i.e., isometric building units distributed in some homogeneous host (photographic systems, drugs, cosmetics, color pigments in binders, etc.), and three-dimensional heterogeneous densely packed systems (poly- or nanocrystalline materials, ceramics, or nanoceramics, etc.).

The materials most intensively investigated, in the field of nanoscience, represent a subclass of heterogeneous materials.

Most of the common materials in our everyday environment have heterogeneous structures, often on larger characteristic size-scales. More generally, heterogeneous materials are so widespread that they are much more frequent than homogeneous materials. Some more examples, relevant for their optical properties, may illustrate this: pigment colors, photovoltaic composites, aerosol-systems, hydrosol-systems, many minerals, photonic crystals, island film devices, nanosensors, special biosystems, and polymer-systems. Most interesting optical features arise when these inhomogeneities are below or of the order of light wavelengths. Recent research has concentrated on a narrow size region on the lower nanometer scale where many-atoms-systems change from molecular to solid-state behavior with increasing sizes. This ‘transition’ is accompanied by drastic changes of almost all physical and chemical properties, hence these properties develop strong dependences on particle size in that size region, and materials with uncommon properties can be developed.

In contrast to the widespread applications and research perspectives, the basic theoretical understanding of heterogeneous material is generally still poor. One reason is that, since the beginning of the development of optics as a science, the homogeneous materials like glass, water, etc. were preferentially treated since they were easier to describe theoretically. Another reason is the extremely wide variability of existing heterogeneous structures usually based upon low symmetries. This variability is too large, and the resulting optical properties too complex and multifarious to enable a general and uniform description of heterogeneous matter on a quantitative level, like the one successfully developed for the much simpler topology of homogeneous crystals.

Qualitative general features, however, can be obtained from investigations of model systems, with especially simplified topological structures. In the following we describe the model system of nanoparticle/host-matter (or, synonymously, cluster-matter), which is based upon nanoparticles/clusters packed closely in a host material, thus forming macroscopic or microscopic heterogeneous many-particle systems. Even this model system exhibits broad variability of topologies and of chemical compositions, as shown in Figure 1.

The restriction in the present contribution to this topologically simple model system appears to be justified by the fact that the optical properties discussed below for this system are also qualitatively characteristic of other arbitrary heterogeneous materials. We also admit that the nanos may consist of a nucleus and a surrounding shell of different structure and/or different material.

Heterogeneous matter can be defined by its heterogeneous building units and by the borders of these units and by the host material and the borders between building units and host, i.e., the internal interfaces.

It is a general feature that structures are visible because of property changes at their terminating borders: borders create structures, and structures require borders. The more refined the structures are, the more increases the topological extent of the borders. This means that when sizes of building units, for example, in cluster matter decrease, the relative amount and the effects of surface/interface regions increase. Such interfaces in heterogeneous matter are not two-dimensional (as suggested by the term 'interface') but consist of three-dimensional

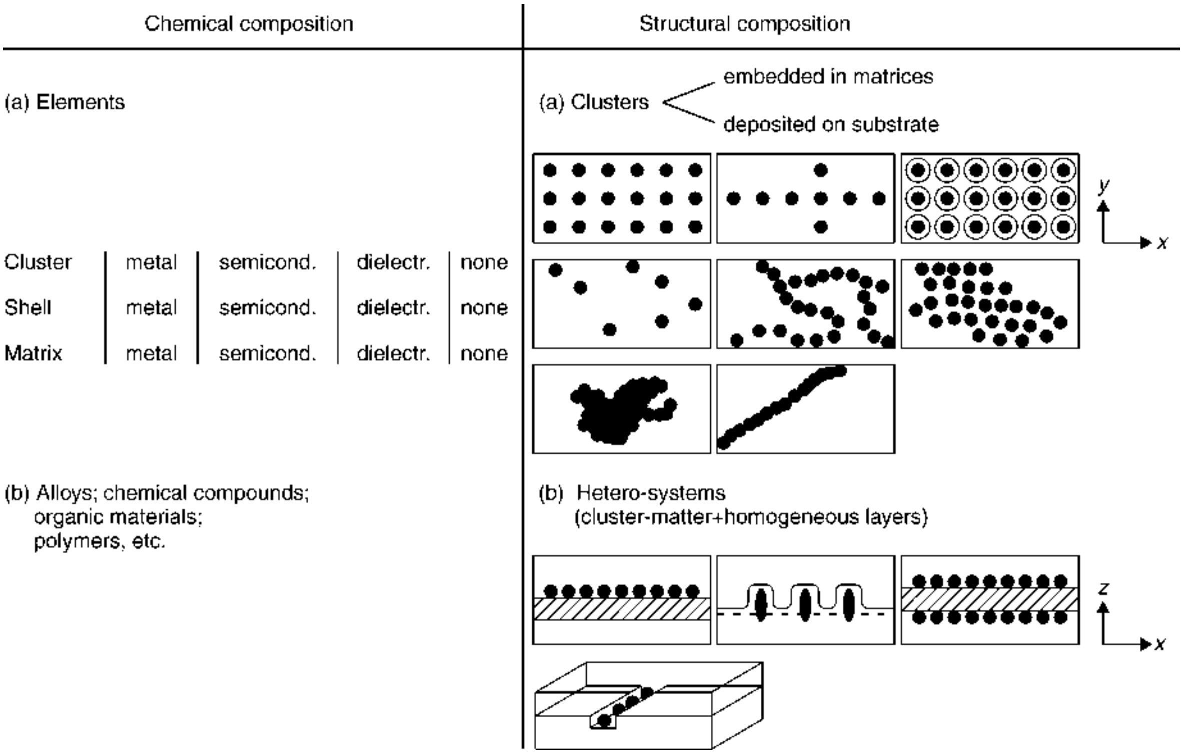

Figure 1 Examples of heterogeneous nanoparticle/host materials. Left: possible material combinations of the three building units: cluster, shell, matrix. Right: some examples of sample topologies.

interlayers, the atomic and electronic structure and the chemical composition of which are usually in complex disorder. In the case of very small particles the interface volume can easily exceed the volume of the building units themselves. There are, however, also examples where the atomic arrangement forms regular interfaces, such as at coherent grain boundaries.

Hence, heterogeneous materials are characterized by various building units, homogeneous matrices and the interface regions. To describe them, we will focus on our introduced model system of nanoparticle/host matter, where we further specialize to spherical shapes of the particles.

The investigation and description of optical properties of heterogeneous materials have to deal with the interior of the building units, e.g., characteristic optical size effects, and with the interfaces, e.g., optical interface effects. The importance of both increase with decreasing characterizing structural lengths. The role of the host materials, when present, can be different, depending on their concentration: in dilute particle systems they may act as bulk-like homogeneous matter (e.g., the electrolyte in colloidal liquid systems); at low amounts, the host may also develop size effects (e.g., the water layers in the opal).

General Optical Properties

Among the experimental characterization methods most frequently applied to heterogeneous matter, the optical ones are essential, ranging from spectroscopy in the frequency domain (from FIR to UV), to femtosecond analysis in the time domain.

Optics can mean physical optics, i.e., the science of optical material properties, including light emission (the 'active' properties) and light absorption (the 'passive' properties), and instrumental optics, i.e., special experimental techniques based upon propagation of light to analyze such samples.

The most usual experiments for the investigation of linear optical response are refraction, absorption, reflection, and elastic scattering. Also, there are the emission responses and the broad field of nonlinear optical responses.

Refraction, reflection, and absorption are typical of homogeneous matter. Elastic light scattering, on the other hand, can be used to define optical heterogeneous matter, since elastic scattering is a direct consequence of the existence of interior interfaces forming deviations from the homogeneous structure. While the grainy structure on the atomic level in homogeneous matter does not give rise to disturbances of the shape of a plane interacting optical wave, this does occur with heterogeneous matter. Each inhomogeneity gives rise to a wavefront deformation and, hence, to light scattering. Scattering spectra and intensities provide important information about the heterogeneous material, so such optical methods are frequently used for its identification.

Besides scattering, the electrical charging of the interior interfaces induced by the light field, can lead to spectacular multipolar optical absorption features, unseen in related macroscopic homogeneous materials, such as Mie resonance, discussed below.

The optical properties of our model system are composed of various contributions which all act together in a complex way, which can be described as:

1. the optical properties of the single building units. In the case of solid state units, they are characterized by the optical material properties, frequency dependent and, generally, complex dielectric function $\varepsilon(\omega)$ or, alternatively, the (generally complex) refraction index $n(\omega)$, which, in the nanometer-size region, both develop strong variations with size.
2. the optical interactions among all building units in the sample. There are different effects:
   - interference effects among the scattering radiation of the units;
   - electromagnetic coupling effects;
   - the contributions of host materials, if present; and
   - the interactions between heterogeneous building units and host.

The Optical Nanoparticle Resonances (Mie Resonances)

Metallic and semiconductor nanoparticles can exhibit special, spectrally selective absorption and scattering bands which do not have counterparts in the molecular state, in thin film structures, or in the extended bulk. They represent the most spectacular optical properties of matter with nanosized heterogeneities. Therefore they will now be treated in more detail in the frame of our model material of spherical inclusions in a homogeneous embedding medium. We begin with a comparison to thin film structures.

The interband transitions (in Ag below 325 nm) are almost equal in both spectra shown in Figure 2, of a thin Ag film and spherical Ag nanoparticles, respectively, while above 325 nm the flat relaxator spectrum of the conduction electrons in the metal film differs strongly from the oscillator spectrum of the same electrons in the clusters. The inset demonstrates the excitation of the dipolar surface oscillator mode by

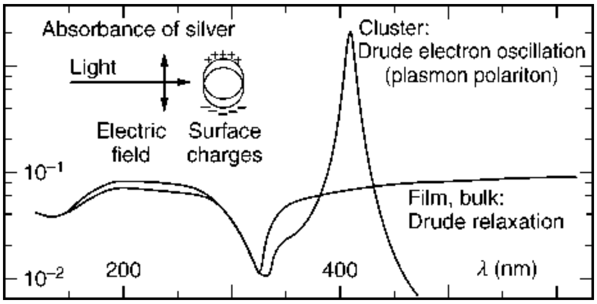

Figure 2 Comparison of absorbances of a thin Ag film and spherical small Ag nanoparticles.

incident polarized light. This figure exemplifies the above-mentioned effect that charging of interfaces by the incident electromagnetic lightwave can give rise to electromagnetic multipole resonances, the Mie-resonances.

The monotonic conduction electron absorption spectrum of the thin film, ranging from UV to zero frequency (the dc current), is thereby compressed in the nanoparticles into the narrow Mie-absorption band with peak position in the blue, containing almost the whole oscillator strength of the conduction electrons. In particular, the Ag-spheres are therefore transparent in the low frequency region, and if we assume the conductivity according to the Drude model as being definitive for a metal, we must assume that these particles are no longer metallic. The conduction electrons no longer follow the Drude relaxator behavior; instead the narrow peak indicates the properties of an oscillator.

The origin of this oscillator and its excitation can be described, being a quantum phenomenon; due to the incident electrical field, the electron plasma, the 'electronic particle', is shifted against the fixed ionic background, the 'ionic particle', causing surface charges modulated with the light frequency. The surface charges produce restoring forces to the plasma, following the periodic field strength, which is the condition required for oscillator behavior. As every free oscillator the particle develops resonances at special eigenfrequencies, and in the case of our metal clusters these are close to the peak frequency of the Mie-absorption, the oscillations of which not being free but forced by the incident field. Analogous to the mechanical case, the resulting resonance frequencies of free and forced oscillations differ slightly. To distinguish them, Mie resonances in metallic particles, coupled to the external electromagnetic field, are called plasmon polaritons.

In the frame of classical electrodynamics each single particle in the heterogeneous material acts as a spherical nano-antenna. This acts as receiving antenna by absorption and, as with every other kinds of antenna, simultaneously becomes a transmitting antenna by re-radiating elastically scattered light.

In the case of Cu, Ag, and Au, and many other metals, these absorption frequencies are in the visible, hence beautiful colors appear. However, there are a lot of metals where this oscillator is overdamped due to low electron mobility. Under particular conditions, such resonances also occur in semiconductor nanoparticles, but due to the lower charge densities, in the IR. At weak fields, the optical response of the oscillator is linear. At higher fields it becomes nonlinear, and thus second and higher harmonics can be excited.

Such unique collective excitations are not restricted to electrons but also spherical surface phonon polaritons can be excited if the particles are nonmetallic (in this case, free electron excitations do not cover the spectral region of phonon excitations). Only in the most simple case of particles which are extremely small compared to wavelength ($R/\lambda \leq 1/20$), the dipolar mode of excitation described so far (i.e., homogeneous polarization of the whole particle) comprises the whole response. For larger particle sizes, a complicated, nonhomogeneous polarization of the whole particle occurs, which can be described by an orthogonal multipole expansion, i.e., multipolar oscillations. The according size (R-) dependent multipolar contributions, which are excited simultaneously and sum up to the total optical response. Further complications arise in larger particles by retardation-induced phase shifts of the lightwaves along the particles.

A complete description of this complex multipole excitation was given by Gustav Mie on the basis of classical electrodynamics. Mie's theory is the exact solution within the frame of Maxwell theory for spherical symmetry, as are the Fresnel-formulae for the planar geometry. However Mie's theory does not describe exactly the conditions which we observe in real heterogeneous systems, hence several extensions of Mie's theory will be presented in the section entitled *Beyond Mie's Theory,* below.

Mie's theory also gives some qualitative insight into the optical behaviour of other kinds of heterogeneous matter. A highly abridged summary of this theory will be presented below.

The Theory of Mie

In the following, an outline of the Mie theory for one spherical particle is given by listing the mathematical steps:

1. Introduction of spherical coordinates r, θ, ϕ. Particle radius R;

2. Incidence of a plane, monochromatic, harmonic electromagnetic wave (circular frequency ω, wavelength λ);
3. Solution of Maxwell's equations for the incident and scattered fields in the surrounding dielectric host medium with (real) dielectric constant $\varepsilon_{\text{matrix}}$, and in the interior of an arbitrary spherical particle of (complex) dielectric function $\varepsilon(\omega)$.

   There are two solutions for transverse fields: one with the radial component of the electric field $E_r = 0$, called 'transverse electric (TE) solution', and one with the radial component of the magnetic field $H_r = 0$, called 'transverse magnetic (TM) solution'. These solutions can be obtained in terms of vector spherical harmonics which can be derived from corresponding scalar potentials Π_E and Π_H. These are solutions of the Helmholtz-equation:

   $$\begin{aligned} &\nabla^2 \Pi_{E,H} + K^2_{\text{transverse}} \times \Pi_{E,H} = 0, \\ &K^2_{\text{transverse}} = \varepsilon_{\text{transverse}} \omega^2 / c^2 \end{aligned} \qquad [1a]$$

   In spherical coordinates, the Helmholtz equation is solved by a separation of variables ansatz and a multipole expansion, yielding an infinite number of independent partial solutions with multipolar order L, the TM partial waves or 'electric partial waves', and the TE partial waves or 'magnetic partial waves'.

   A third solution, the potential Γ, that only affects the electric field, can be found as the solution of

   $$\nabla^2 \Gamma = 0, \qquad \varepsilon_{\text{longitudinal}} = 0 \qquad [1b]$$

   and represents longitudinal excitations in the particle.
4. Boundary conditions at $r = R$:

   for the transverse fields:

   $$\begin{aligned} &E_\theta^{\text{incident}} + E_\theta^{\text{scattered}} = E_\theta^{\text{interior}} \\ &E_\phi^{\text{incident}} + E_\phi^{\text{scattered}} = E_\phi^{\text{interior}} \end{aligned} \qquad [2a]$$

   $$\begin{aligned} &H_\theta^{\text{incident}} + H_\theta^{\text{scattered}} = H_\theta^{\text{interior}} \\ &H_\phi^{\text{incident}} + H_\phi^{\text{scattered}} = H_\phi^{\text{interior}} \end{aligned} \qquad [2b]$$

   for the longitudinal fields:

   $$E_r^{\text{incident}} + E_r^{\text{scattered}} = E_r^{\text{interior}} \qquad [3]$$

   (Sauter–Forstmann condition)

   The linear response function of the particle material used here, is ε, where $\varepsilon = \varepsilon_1 + i\varepsilon_2$ is averaged over the particle volume, including the surface–interface region.

   The magnetic permeability μ is set 1 for the investigated high frequency regions.
5. Computation of cross-sections for extinction, scattering, and absorption from Poynting's law:

   - Extinction, i.e., the sum of absorption and scattering losses:

     $$\sigma_{\text{ext}} = \frac{2\pi}{k^2} \sum_{L=1}^{\infty} (2L+1) \operatorname{Re}\{a_L + b_L\} \qquad [4]$$

   - Elastic scattering:

     $$\sigma_{\text{sca}} = \frac{2\pi}{k^2} \sum_{L=1}^{\infty} (2L+1) \left(|a_L|^2 + |b_L|^2 \right) \qquad [5]$$

   - Absorption, i.e., the energy dissipation (production of heat and thermal radiation):

     $$\sigma_{\text{abs}} = \sigma_{\text{ext}} - \sigma_{\text{sca}} \qquad [6]$$

     L indicates the multipolar order of the Mie resonances.

The coefficients a_L and b_L, following from the Maxwell boundary conditions for transverse fields, are

$$a_L = \frac{m\psi_L(mx)\psi'_L(x) - \psi'_L(mx)\psi_L(x)}{m\psi_L(mx)\eta'_L(x) - \psi'_L(mx)\eta_L(x)} \qquad [7a]$$

$$b_L = \frac{\psi_L(mx)\psi'_L(x) - m\psi'_L(mx)\psi_L(x)}{\psi_L(mx)\eta'_L(x) - m\psi'_L(mx)\eta_L(x)} \qquad [7b]$$

with $m = 2x/n_{\text{Matrix}}$, where m denotes the complex-valued index of refraction of the particle $\varepsilon = n'^2$ and n_{matrix} the real-valued index of refraction of surrounding medium. k is the wavevector and $x = k \cdot R$, the size parameter. $\Psi_L(z)$ and $\eta_L(z)$ are Riccati–Bessel functions of the first and third kind. The prime indicates differentiation with respect to the argument in parentheses. Including longitudinal fields, eqn [7] will be slightly modified.

Applying Lambert–Beer's law, these cross-sections are related to the absorption-, scattering-, and extinction-constant $A(\omega)$, $S(\omega)$, and $E(\omega)$, respectively, and the resulting intensity loss ΔI of a many-particle system of number-density Z of the (noninteracting) particles, is

$$\begin{aligned} &\Delta I = I_0 \left(1 - \exp(-Z \cdot \sigma_{\text{ext}} \cdot d) \right), \\ &d = \text{sample thickness} \end{aligned} \qquad [8]$$

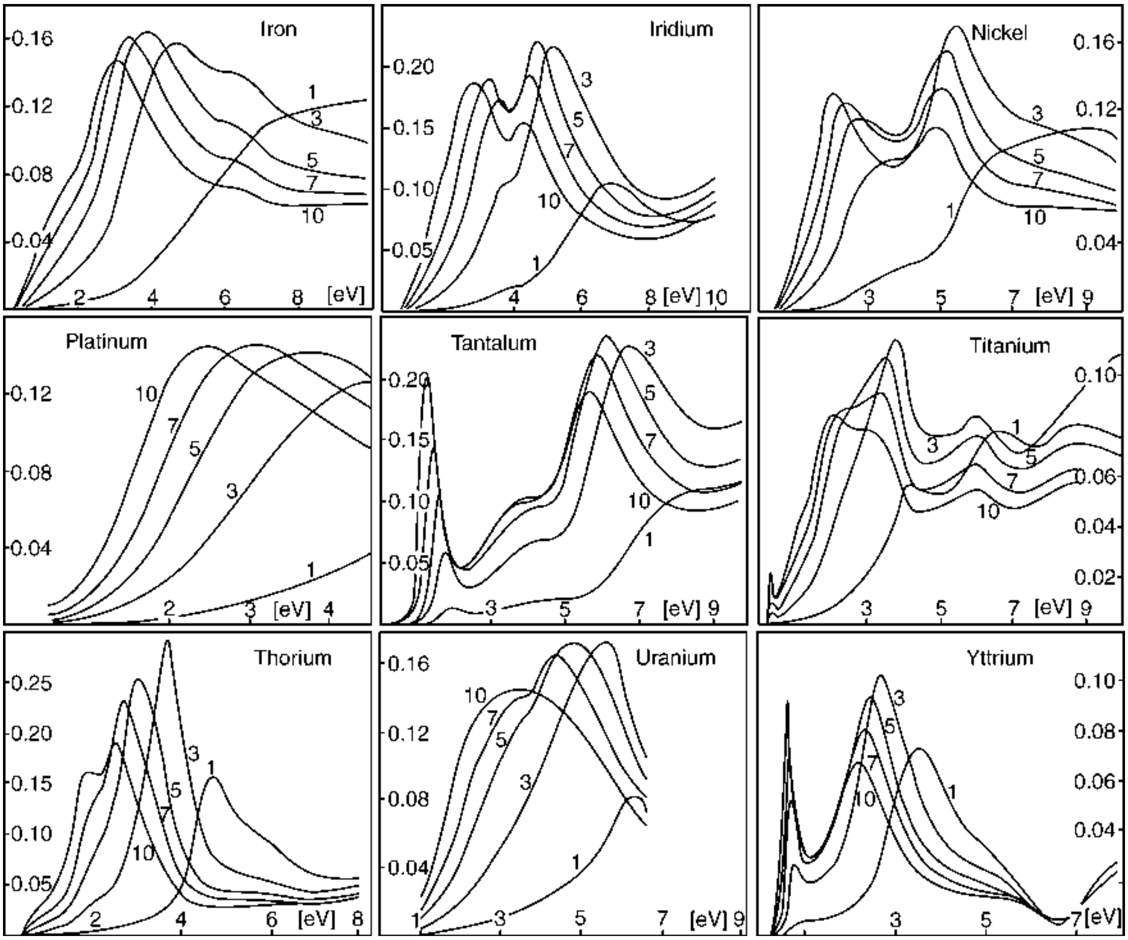

Figure 3 Mie extinction spectra of various particle materials. $2R = 10\,\text{nm}$. The clusters are embedded in matrix materials with dielectric constants between 1 and 10, which are indicated for each spectrum. Obviously, the peak structure of the spectra strongly depends on the matrix.

Figure 3 gives a compilation of some spectra of general interest determined by evaluating Mie's theory.

Beyond Mie's Theory

Application of Mie's theory to experimental results is disappointing because, in most cases, quantitative coincidence is not obtained. One reason is that the simple assumptions incorporated either in Maxwell's theory or in Mie's derivation are not met by reality.

Mie's theory does not apply, if:

1. particle sizes in a diluted cluster matter sample are not uniform;
2. particle shapes differ from the sphere;
3. $\varepsilon_{\text{matrix}}$ varies with frequency, $\varepsilon_{\text{matrix}}(\omega)$
4. cluster-dependent dielectric functions differ from the according functions of the bulk;
5. the (step-function like) Maxwellian boundary conditions at the particle-matrix interface are not applicable;
6. dielectric functions of particles and/or matrix are nonlinear;
7. surrounding/embedding matrix material is absorbing;
8. particle structures are heterogeneous (core-shell structures, multi-grain clusters, etc); and
9. particles form close-packed aggregates instead of being well-separated, and hence, electromagnetic coupling among particles is important.

Of course, this list may be continued. Regarding realistic samples, it is obvious already from these deviations that quantitative correspondence between experiment and Mie theory usually cannot be expected. But, over time, each of these extra effects has been treated and included into extensions of Mie's theory, which should therefore be applied, instead of the original theory. In the following several of these effects will be treated in some detail.

Particle Shapes Deviating From the Sphere

Even metals of high structure-symmetry, like the fcc-Au, can yield nanoparticles with low symmetry habits. Figure 4 shows Au nanoparticles with exotic shapes ranging from triangle and hexagonal platelets over fivefold symmetry multitwin-particles to rodlike shapes. The majority of particles are multigrain structures of irregular isometric shapes.

Multishell Particles

In experiments, particle surfaces often react chemically with the surrounding media, thereby forming compound shells. Core/shell structures can be clearly visible in the optical absorption spectra. Figure 5 shows, as an example, the absorption spectra of nanoparticles of Ag core/Au shell and of Au core/Ag shell, respectively with varied shell thicknesses.

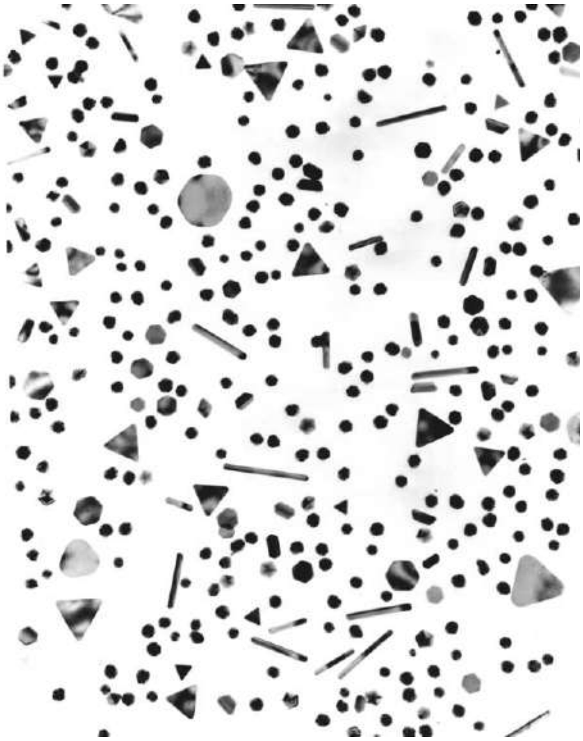

Figure 4 Chemically (Esiguondy method) produced Au clusters with extraordinary variety of different shapes. (Of special interest are the small particles of five-fold symmetry.) Mean sizes of the irregular, almost spherical particles: 36 nm.

Resonances will also occur, e.g., in metal-coated dielectric particles. An example is shown in Figure 6, with extinction spectra of silver-coated titanium-dioxide particles, computed by an extension of Mie's theory.

Aggregates of Nanoparticles with Electrodynamic Particle–Particle Coupling

The optical extinction and scattering spectra of many-particle systems with dense packing forming aggregates (see Figure 4) are strongly influenced by electromagnetic interactions: each particle feels the

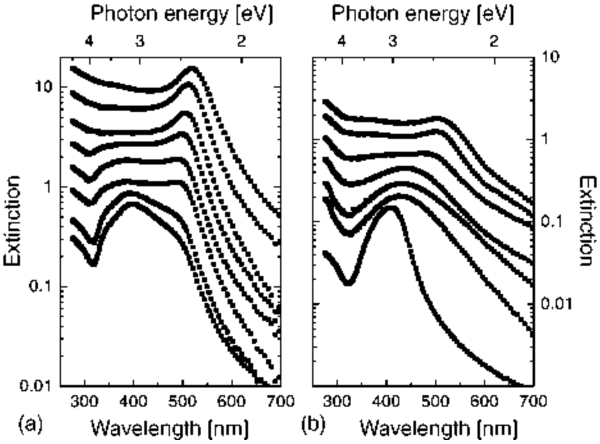

Figure 5 Measured extinction spectra of (a) silver-coated gold particles (gold core $2R = 17$ nm, shell thickness varying from $d = 0$ to $d = 3.6$ nm from top to bottom), and (b) gold-coated silver particles (silver core $2R = 18$ nm, shell thickness varying from $d = 0$ to $d = 5$ nm from bottom to top).

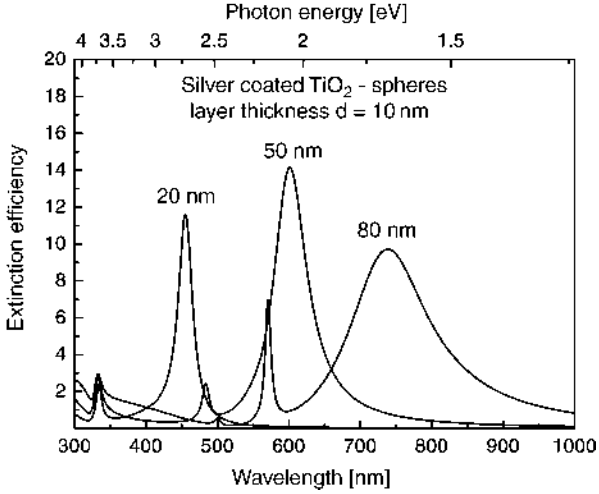

Figure 6 Calculated extinction spectra of spherical silver-coated titanium dioxide particles. The shell thickness is kept constant at $d = 10$ nm, the core sizes amount to $2R = 30$ nm, 50 nm, and 80 nm, respectively, as indicated in the graph.

scattering near fields of all its neighbors, which give rise to splitting, broadening, and shifts of resonances. This is clearly demonstrated with colloidal Ag and Au particles in Figure 7.

The splitting, broadening, and shifts can be assigned to the shape and size of the resulting many-particle aggregates. An example for the influence of the topology of the many-particle system is given in Figure 8 with extinction and scattering spectra of silver particle aggregates of $N = 5$ identical silver particles with $2R = 40$ nm, which were computed by the Generalized Mie Theory (GMT). As a rough rule-of-thumb, such electrodynamic coupling effects of the oscillators only can be ignored if the smallest next neighbor center-to-center distance d_{cc} exceeds $d_{cc} \approx 5 \cdot R$.

Size Effects

The extraordinary practical success of Mie's theory is based upon the fact that this example of classical electrodynamics enables us to compute numerically the optical response for arbitrary realistic particle materials. In contrast quantum theoretical calculations are restricted, beyond the molecule-solid state transition, to jellium and jellium-related hypothetical material.

Classical electrodynamics, being a phenomenological theory to describe light propagation, does not yield information about optical material properties. These enter via the dielectric functions inserted into the Maxwell boundary conditions. They must be taken from elsewhere, such as from experiments or from quantum solid state theory model calculations. This twofold basis of Mie's theory is illustrated in Figure 9.

There are two different sources for particle size dependences of the optical extinction spectra:

1. The electrodynamics: the larger the particles, the larger are the number of contributing multipoles and the influences of retardation effects, both influencing strongly the optical spectra;
2. The optical material properties describing the particle material which, in the nanoscale size region, exhibit size dependences. These size dependences can be treated by corrections entering the bulk material function.

In general, the dielectric function $\varepsilon(\omega)$ of a metal or semiconductor is the sum of susceptibility contributions of the lattice, of the conduction electrons, and of the interband (electron–hole) excitations:

$$\begin{aligned}\varepsilon(\omega) &= \varepsilon_1(\omega) + i\varepsilon_2(\omega)\\ &= 1 + \chi^{\text{lattice}}(\omega) + \chi^{\text{free electrons}}(\omega)\\ &\quad + \chi^{\text{interband}}(\omega)\end{aligned} \qquad [9]$$

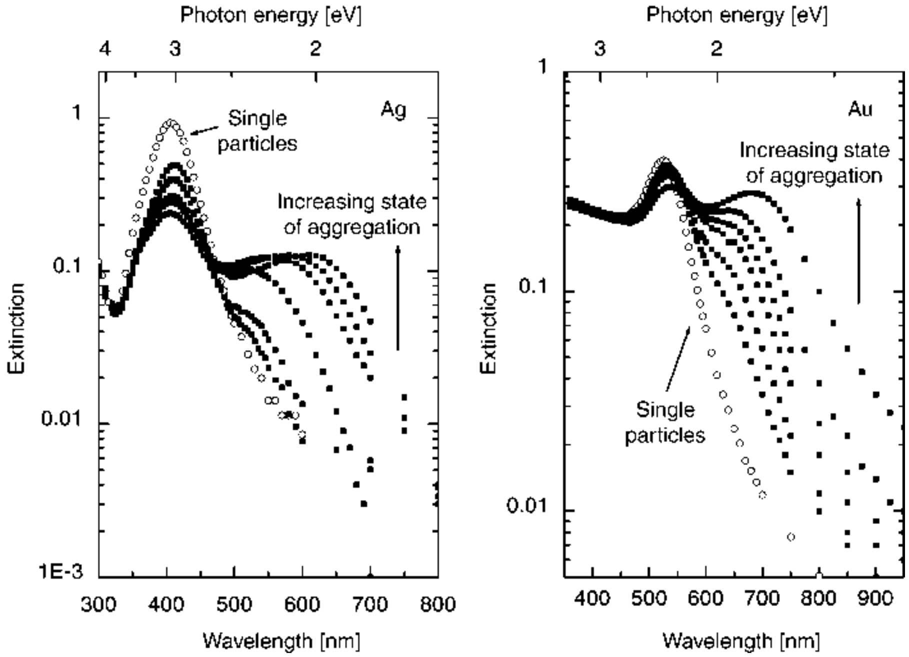

Figure 7 Measured extinction spectra of aggregated samples of (a) silver particles with $2R = 28$ nm, and (b) gold particles with $2R = 38$ nm. The samples contain random distributions of aggregates of various topologies, but with increasing state of aggregation from bottom to top.

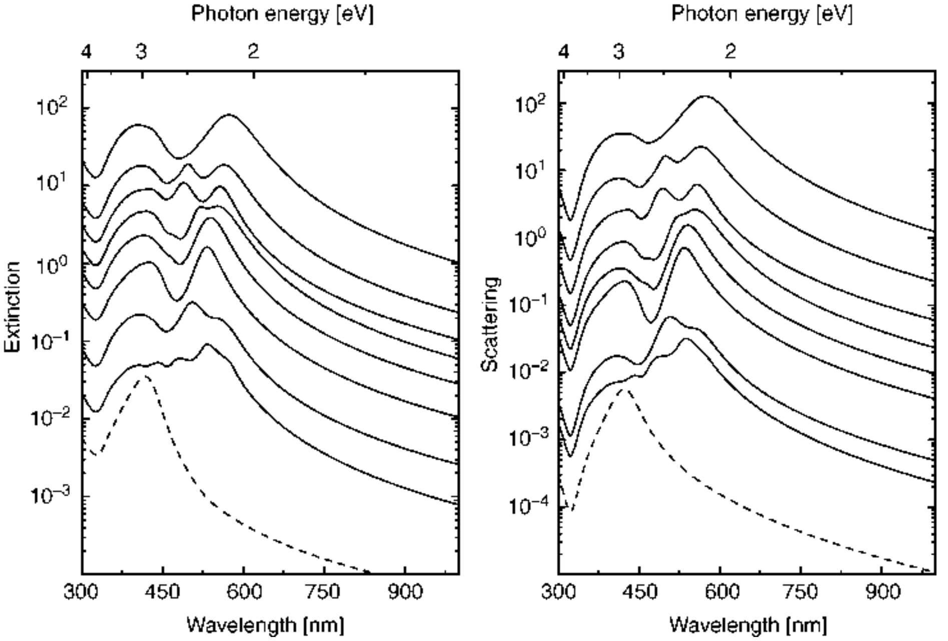

Figure 8 Calculated extinction and scattering spectra of aggregates with $N = 5$ almost touching identical silver spheres of $2R = 40$ nm. From top to bottom, the spectra belong to the following particle arrangements in the aggregate: . The dashed line represents the single particle spectra.

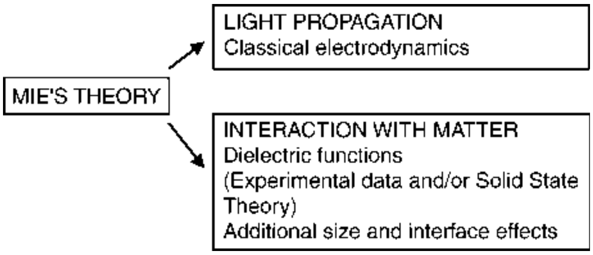
MIE'S THEORY
LIGHT PROPAGATION
Classical electrodynamics
INTERACTION WITH MATTER
Dielectric functions
(Experimental data and/or Solid State Theory)
Additional size and interface effects

Figure 9 Scheme of the basics of Mie's theory.

A detailed expression was given by Bassani (here, the lattice contributions, mainly important in the IR, were disregarded):

$$\varepsilon(\omega) = 1 - \underbrace{\frac{(ne_0^2/\varepsilon_0 m_{\text{eff}})}{\omega^2+i\gamma\omega}}_{\text{conduction electrons}} + \underbrace{\chi^{\text{interband}}(\omega)}_{\text{interband transitions}} \qquad [10]$$

with

$$\begin{aligned}\chi^{\text{interband}} &= \frac{2\hbar^2 e^2}{\varepsilon_0 m_{\text{eff}}^2} \sum_{i,f} \int_{\text{BZ}} \frac{2d^3k}{(2\pi)^3} |\mathbf{e} \times M_{\text{if}}(\mathbf{k})|^2 \\ &\times \left\{ \frac{1}{[E_f(\mathbf{k}) - E_i(\mathbf{k})][(E_f(\mathbf{k}) - E_i(\mathbf{k}))^2 - \hbar^2\omega^2]} \right. \\ &\left. + i\frac{\pi}{2\hbar^2\omega^2}\delta(E_f(\mathbf{k}) - E_i(\mathbf{k}) - \hbar\omega) \right\} \qquad [11]\end{aligned}$$

In this equation, the quantity n is the conduction electron density, $\mathbf{k}$ the wavevector of the electrons with norm k, and BZ denotes the Brillouin Zone. The vector $\mathbf{e}$ is the polarization vector of the incident light. M_{if} is the transition matrix element with i = initial and f = final states.

To which degree of approximation a single dielectric function can describe the optical response of a nanoparticle consisting of inner core and surface with differing local optical properties (i.e., polarizabilities), and how this function has to be modelled as an average, is still a problem. An easy, but only approximate answer is to identify terms in the eqns [10] and [11], which may change if the cluster size is reduced down to the nanometer scale, and to introduce correction terms.

In the case of the conduction electrons of eqn [10], these are the electron density n, the effective electron mass m_{eff}, and the relaxation frequency γ. In the terms of the interband transitions (index inter), these are the transition matrix elements M_{if} and the band structure energies of the initial and final states E_i and E_f. All of these quantities change, when the particle size is reduced, in different size regions and to different amounts, and detailed quantum theoretical investigations are required to derive proper correction terms.

The application of the continuum-based dielectric functions and of Mie's theory fail when the molecule-solid state transition region is reached.

The effects of size-dependent dielectric functions upon cluster absorption spectra are demonstrated in the series of absorption spectra of Figure 10, measured on highly diluted heterogeneous systems containing spherical Ag, Au, and Cu particles which were embedded in a glass matrix. If $\varepsilon(\omega)$ were independent of size, then the spectra should be identical, in the investigated particle size region. So, their differences reflect clearly that $\varepsilon = \varepsilon(R)$.

To describe the size effects of the conduction electrons, the relaxation frequency γ of eqn [10] has been modified introducing a parameter that is called A-parameter which also contains size effects and also particle–host interface effects. To explain the electronic size effects in detail, quantum effects have to be considered taking the discretization of the bandstructure into account, but they can also be treated by a simple classical approach: this is the mean free path effect (FPE), assuming the mean flight time of the effective conduction electrons (i.e., those with Fermi energy E_F) between subsequent surface collisions amount to

$$\tau\gamma_{\mathrm{FPE}} = A_{\mathrm{FPE}}\frac{v_{\mathrm{Fermi}}}{R} \qquad [12]$$

The A-parameter is a measure of the effectiveness of the single collision in view of relaxation. Complete 'memory loss' of the drift momentum in each collision would lead to $A_{\mathrm{FPE}} = 1$. Taking electron correlation effects and energy relaxation into account, A-values larger than 1 may also be realistic. Values larger than unity also occur in multigrain particles, due to grain boundary 'collisions'.

Various quantum mechanical models of electronic particle size effects result in relations, formally coinciding with the $1/R$-dependence in eqn [12]. Most relevant is the value of A for particles *in vacuo* with clean surface. This value has been determined theoretically and experimentally to $A \approx 0.25$ for Ag particles.

Properties of Internal Interfaces

Structures of Surfaces/Interfaces

Only mathematical interfaces in geometry can, in fact, be two-dimensional planes separating two different volume elements. This is assumed in the frame of Maxwell boundary conditions entering Mie's theory. In real nanostructured matter, however, the interface regions are three-dimensional layers formed by disturbances of atomic and/or of electronic structures, and/or of chemical compositions. In addition, due to the finite wavelengths of the conduction electrons, the according electron density 'spills out' and, as a consequence, we find electron density beyond the geometric surface. It should be kept in mind that these effects hold, with proper changes, for interfaces in nanostructured materials in general. Hence, an understanding of the interface regions is of key importance for all nano-material.

Electronic Interface States

Up to now, this discussion has been restricted to particles with uncovered, 'clean' surfaces. In most practical examples of heterogeneous materials however, the building units are embedded in a host matrix, deposited upon the surface of a substrate, or are in close contact to neighboring crystallites, etc., to stabilize the samples. Then, special electronic

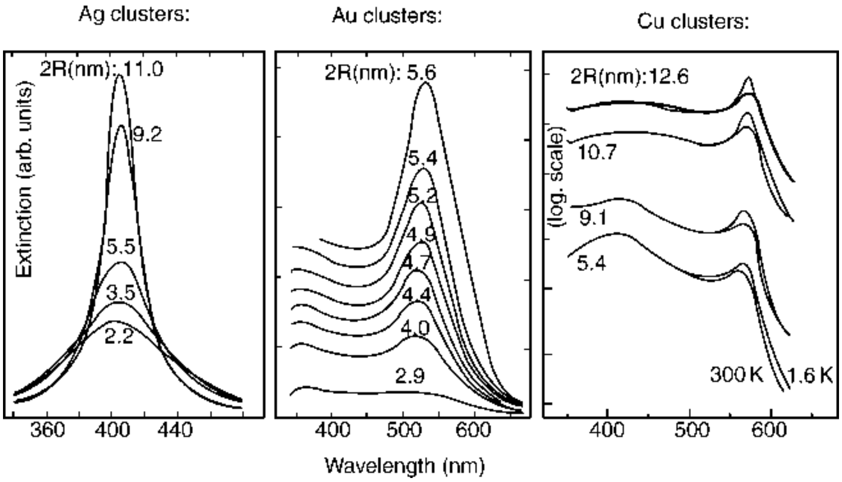

Figure 10 Measured extinction spectra of Ag, Au, and Cu clusters of various mean sizes, grown in glass matrix. In the case of Cu-particles, measurements at 1.5 K are added.

interface states are created by contributions of particle electrons and host electrons, which are (disregarding their density on the surface) similar to adsorbate states.

One might assume that the broad knowledge about these states in surface science, obtained from samples with extended crystallographic planes, should be very useful here. However, the surface of a cluster, in principle, differs from the planar surface of a single crystal by containing a multitude of different local surface structures, atomic sites, and varying atomic coordination numbers. Even if the surface structure is not irregular (as is usual in nanoparticles), the surface consists of several facets of different orientation separated by edges and corners. The coordination numbers of surface atoms, thus, vary strongly (e.g., between 5 at a corner and 9 on a plane), indicating surface sites to have different energies.

We have shown that, if the particles are metallic, the 'particle/host'-interface states give rise to two different electronic charge transfer processes of particle conduction electrons, the static and the dynamic charge transfer. This will be demonstrated qualitatively for metal particles by the example presented in Figure 11.

The Static Interface Charge Transfer and its Effects Upon Mie Resonance Positions

At the left side of Figure 11, the conduction band of the metal cluster close to its surface is shown, which is occupied with electrons up to the Fermi energy E_F. Chemisorption of a free atom (one energy level E_a of which is plotted at the right-hand side) is assumed to occur by approaching the atom to the cluster surface. The resulting energy states are also shown. Let us assume the marked electronic level E_a in the free atom is nonoccupied. During the chemisorption process, this level is shifted and broadened, developing density of nonoccupied states below the particle Fermi energy, which is localized at the three-dimensional interface region. After complete the process is, such a state will be occupied permanently by electrons of the cluster (static charge transfer), and the nanoparticle will have lost electrons, thus its electron density being reduced. The driving impact of this process is the equilibration of the chemical potential in the whole cluster-matrix system.

Now assume we embed a cluster into some surrounding material. Then all those atoms of the latter, which are situated directly at the interface, become chemisorbed at the cluster. Hence remarkable amounts of charges can be transferred into the interlayer region and an electrical double layer is created. As an example, a very strong amount of static charge transfer was observed experimentally when applying solid Fullerite (C_{60} molecules) as embedding matrix for Ag-nanoparticles: here, each C_{60} in direct contact with the Ag particle attracts one electron, forming (C_{60}^{-}). Hence, in this experiment with 500 atoms large particles, in total about 20% of the conduction electrons of the Ag-particle were transferred into the interlayer. Strictly speaking, such a nanoparticle of the heterogeneous material is no longer a silver particle because of the remarkable changes of electronic properties by static charge transfer.

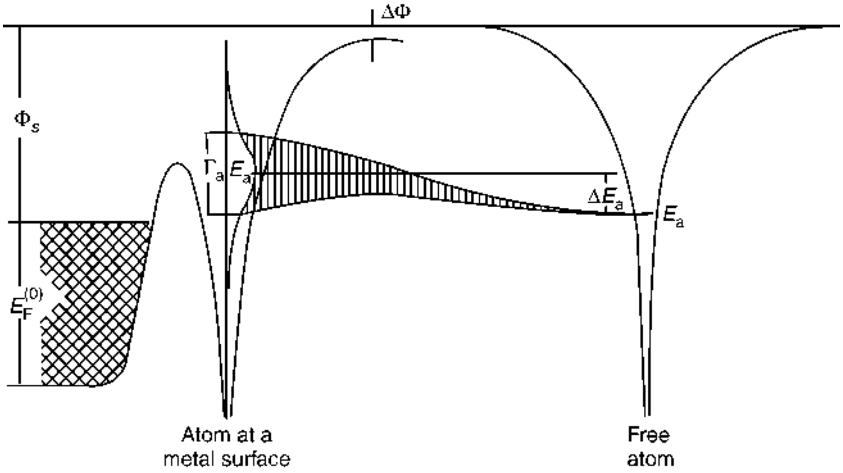

Figure 11 Electron energy scheme for atomic adsorbation at a metal particle surface. The free atom (energy spectrum: right) approaches the particle surface (left; conduction band with Fermi-level E_F). Its energy level E_a is thereby shifted and broadened (including ensemble broadening of many atoms deposited at many clusters). The atoms are assumed to be separated from the metal by a tunnel barrier (see text).

Some consequences are listed in the following:

In a first step, let us assume that the electronic bandstructure is not altered by the charge transfer. Then we expect: (i) changes of level occupation; (ii) changes of Fermi energy and Fermi velocity; (iii) changes of effective masses close to E_F; (iv) changes of surface 'spill out'; (v) additional adsorbate levels and special surface states; and (vi) changes of the optical interband transition edge.

In a second step, we may consider possible changes of electronic bandstructure (e.g., energies) and, finally, also changes of atomic structures (i.e., lattice structures and/or surface structures).

Since plasmon frequencies depend on the conduction electron density, the resulting Mie resonance can monitor and give quantitative information of this effect. We find shifted by the static charge transfer:

$$\Delta\omega_{\text{resonance}} \approx (\sqrt{n_2} - \sqrt{n_1}) \times \sqrt{\frac{e_0^2}{\varepsilon_0 m_{\text{eff}}(2\varepsilon_{\text{matrix}} + 1 + \chi_1^{\text{interband}})}} \qquad [13]$$

with n_1 being the conduction electron density in the cluster *in vacuo*, and n_2 being the electron density after contact with foreign material (adsorbate, substrate, or host-matrix). This peak shift of the Mie resonance, however, is superimposed in the experiment by a much larger shift, following simply from the change of Maxwell boundary conditions by the altered dielectric constant of the particle surrounding $\varepsilon_{\text{matrix}}$, when going from vacuum to a matrix material. Hence, the experimental determination of static charge transfer processes is rendered more difficult.

The Dynamic Interface Charge Transfer and its Effects Upon Mie Resonance Widths

After the static charge transfer in Figure 11 has taken place, electrons close to E_F can fluctuate between the cluster and empty adsorbate levels, if present at E_F. The residence times in these levels, which are usually separated from the Fermi sea of the metal cluster by some tunnel barrier, vary statistically. If a plasmon is excited, i.e., mainly a collective, phase ordered motion of all conduction electrons superimposed to their statistical motion, we may assume that, after such electron transfer back and forth, the reminiscence to the primary drift momentum, and the phase of this excitation, has been lost. Thus, the phase relaxation of the plasmon is increased.

The consequences of this dynamic charge transfer effect ('chemical interface damping') have been investigated in detail. Here, we only state that the directly observable main consequence is the broadening of Mie resonances. If the particles are sufficiently small to neglect radiation damping (this is the case for Ag clusters smaller ≈15 nm), band broadening is only due to internal energy dissipation and the band width can be modeled quantitatively with the modified relaxation frequency γ in eqn [10] which yields an additive correction term:

$$\Gamma(A,R) \approx \Gamma_0 + \frac{2ne_0^2}{\varepsilon_0 m_{\text{eff}}\omega^3} \times \frac{A}{\sqrt{(\partial\varepsilon_1/\partial\omega)^2 + (\partial\varepsilon_2/\partial\omega)^2}} \frac{v_{\text{Fermi}}}{R} \qquad [14]$$

with the A-parameter extended to the additional electronic excitations in the particle/host interlayer:

$$A = A_{\text{size effects}} + A_{\text{interface effects}} \qquad [15]$$

The A-parameter of eqn [15] now expresses both the strengths of the particle size effects and of the chemical interface damping, due to statistically varying occupation of interface states near E_F.

An important feature of this model is that the dependence on the kinds of materials is described by specific material parameters ($A_{\text{size effects}}$ and $A_{\text{interface effects}}$), and the direct size dependence effects are given by an $1/R$-dependence, for both the direct size effects and the interface effects. For one fixed cluster size R, the magnitude of $A_{\text{interface effects}}$ directly characterizes the probability of the single temporal charge transfer, including the number and energies of contributing interface states; hence, important properties of the adsorption process can be investigated by determining the A-parameter.

Experimental Results

In the experiments presented in the following, we deal with the interfaces between clusters and a more or less homogeneous solid embedding medium. We selected very small silver nanoparticles of about 2 to 4 nm size with accordingly large ratio of surface- to volume-atoms, in order to optimize the sensitivity of the measurements. Silver is unique in having the most pronounced Mie resonance of all known materials.

In Figure 12 two classes of matrix materials are presented, fluorides and oxides. The absorption spectrum of the free particles with uncovered 'clean' surface (measured in a cluster beam in UHV) is added in both diagrams which is only influenced by size effect. General features are, that embedding in a host medium causes:

- drastic Mie resonance peak shifts toward IR, depending on the dielectric constant of the matrix

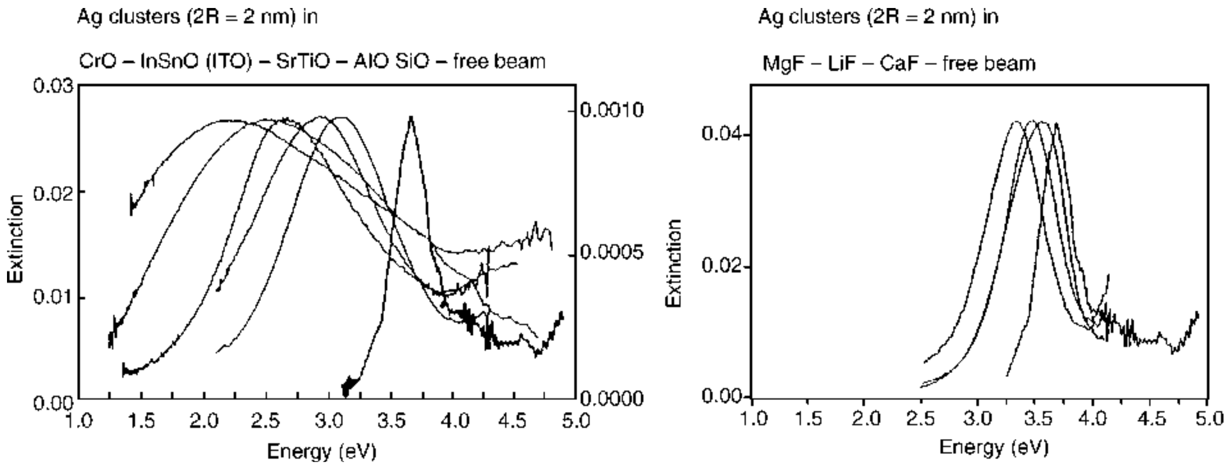

Figure 12 Measured optical extinction spectra of 2 nm Ag nanoparticles, embedded in various different embedding materials. For comparison, the spectrum of the free nanoparticle beam is added. Left side: oxidic materials, right side: fluoridic materials.

$\varepsilon_{\mathrm{matrix}}$, which classically follow from Mie's theory and by comparingly small additional shift due to static charge transfer;

- drastic broadening (exceeding, in some cases, all known size effects), which according to eqn [14], can be described by characterizing values of the *A*-parameter. Recently, these parameters have been experimentally determined for Ag nanoparticles embedded in a broad variety of host materials. Their interpretation is rendered difficult, however, due to inhomogeneous broadening of varying magnitude.

We can immediately judge from Figure 12 that both shift and broadening are larger in the case of oxidic matrices than of fluoridic ones, i.e., oxides possess more interface states close to E_F of the material. In Figure 13, observed peak positions $\hbar\omega_{\max}$ and broadenings expressed in the resulting *A*-parameters, are summarized for a broad variety of different embedding materials.

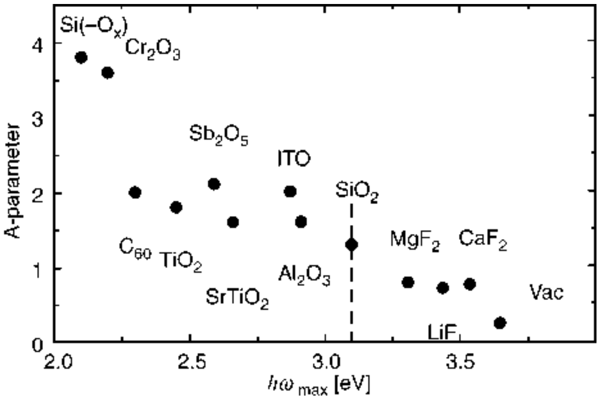

Figure 13 Silver nanoparticles ($2R = 2$ nm) embedded in various solid host matrices: measured Mie peak positions and resonance halfwidths expressed by the *A*-parameters.

Comparison of Plasmon Polariton Lifetimes with Femtosecond Experiments

In the last section, experimental examples were presented to show how fundamental physical properties are changed by forming heterogeneous materials. Because the Mie resonance widths are mainly a consequence of finite plasmon polariton life times τ, (in addition, there are interband contributions, in larger particles radiation damping, and dispersion influence), the latter can be evaluated from measured optical absorption spectra. This supposes the bandwidth to be essentially homogeneous. From a width analysis fitting Mie's theory numerically we obtained for Ag particles, free flying in UHV with uncontaminated free surfaces and 2 nm in size, a plasmon lifetime of $\tau = 6$ fs. Under these experimental conditions, the measured bandwidth is the homogeneous one.

Concerning the size dependences, we can extrapolate to larger particles. The longest lifetimes are expected for 15 nm particles, for which $\tau = 15$ fs was obtained by extrapolating the experimental data for 2 nm particles to this size. The same width analysis was performed for the nanoparticles embedded in matrix materials, and as a result, the lifetimes drastically decrease down to 1–2 fsec. The observed increase of the *A*-parameter, as shown in the previous section, is thus directly correlated to the measured decrease of the plasmon polariton lifetime. Hence, we conclude that the decrease of the lifetime is mainly due to increased phase relaxation caused by the dynamic charge transfer effect.

These results can be compared with recent femtosecond spectroscopy experiments which have recently given much impact on the experimental investigation of the dynamics of Mie resonances in silver and gold nanoparticles. In these experiments data between 6 and 10 fsec were obtained for deposited Ag and Au particles of different sizes and different substrates. Regarding the complexity of the different experiments the correspondence of the data appears convincing.

Obviously the lifetimes evaluated in the frequency domain, from our spectra of particles *in vacuo*, coincide with the lifetimes determined in the time domain. The latter, though, were measured with larger particles deposited on substrates (i.e., not from free particles but including interactions with substrates).

As a concluding remark, it may be pointed out again, that the optical properties and the heterogeneity effects presented here for the model system of nanoparticle/host-matter, are also to be found in other examples of the large field of heterogeneous materials.

List of Units and Nomenclature

Absorption bandwidth	Γ
circular frequency [s^{-1}]	ω
dielectric constant/dielectric function [–]	ε
effective electron mass [kg]	m_{eff}
electric field [V m^{-1}]	E
(electron) relaxation frequency [s^{-1}]	γ
electron state energies for initial and final state [J]	E_i, E_f
elementary charge [C]	e_0
extinction, scattering, absorption cross section [m^2]	σ_{ext}, σ_{sca}, σ_{abs}
Fermi energy [J]	E_{F}
Fermi velocity of electrons [m s^{-1}]	v_{Fermi}
intensity [W m^{-2}]	I
Kronecker energy distribution [J^{-1}]	$\delta(E)$
magnetic field [A m^{-1}]	H
norm of wave vector of electrons [m^{-1}]	k
number density of conduction electrons [m^{-3}]	n
Riccati-Bessel functions of first kind [–]	$\psi_n(z)$
particle radius [m]	R
(phase-) life time [s]	τ
Planck constant [$\mathrm{J_s}$]	$\hbar$
Riccati-Bessel functions of third kind [–]	$\eta_n(z)$
scalar electromagnetic potentials [V]	Π, Γ
susceptibility [–]	χ
transition matrix element [kg m s^{-1}]	M_{if}
vacuum permittivity [CV^{-1} m^{-1}]	ε_0
vacuum velocity of light [m s^{-1}]	c
vacuum wavelength of light [m]	λ
wave vector of electrons [m^{-1}]	$\mathbf{k}$

See also

Scattering: Scattering Theory. **Semiconductor Physics:** Polarons.

Further Reading

Bassani F and Parravicini G (1975) *Electronic States and Optical Transitions in Solids*. Oxford: Pergamon.

Bohren CF and Huffman DR (1983) *Absorption and Scattering of Light by Small Particles*. New York: Wiley.

Born M (1960) *Optik*. Berlin: Springer-Verlag.

Duncan M (ed.) (1998) Advances in Metal and Semiconductor Clusters, vol. IV. London: JAI Press Inc.

Haberland H (ed.) (1994) *Clusters of Atoms and Molecules*. Berlin: Springer Series in Chemical Physics 56.

Nalwa HS (ed.) (2001) *Handbook of Surfaces and Interfaces of Materials*, vol. 3. San Diego: Academic Press.

Hölzl J, Schulte F and Wagner H (1979) *Solid Surface Physics*. Springer Tracts in Modern Physics 85. Heidelberg: Springer.

Hummel R and Wissmann P (eds) (1997) *Handbook of Optical Properties*, vol. II: Optics of Small Particles, Interfaces and Surfaces. Boca Raton, FL: CRS Press.

Jena P, Khanna SN, Rao BK (eds) (2000) *Cluster and Nanostructure Interfaces*. New Jersey: World Scientific.

Kerker M (1969) *The Scattering of Light and Other Electromagnetic Radiation*. New York: Academic Press.

Kreibig U and Vollmer M (1995) *Optical Properties of Metal Clusters*. Berlin: Springer Series in Material Science 25.

Nanometer Scale Science and Technology, Internatl. School of Physics 'Enrico Fermi' (2001) CXLIV. Amsterdam: IOS Press.

Quinten M (1998) *Physical and Chemical Properties of Surfaces in Nano-Particle Systems*. Aachener Beiträge zur Physik kondensierter Materie, vol. 24. Mainz, Aachen: Wissenschaftsverlag.

Ruhbahn HG (2002) *Nanophysik und Nanotechnologie*. Stuttgart: Teubner.

Schmid G (1994) *Clusters and Colloids*. Weinheim: VCH Verlag.

van de Hulst HC (1957) *Light Scattering by Small Particles*. New York: Wiley.

von Plessen G (1998) *Ultrafast Dynamics of the Surface Plasmon in Gold Particles*. Advances in Solid State Physics 253. Berlin: Vieweg.

Lightweight Mirrors

J W Bilbro, George C Marshall Space Flight Center, AL, USA

Published by Elsevier Ltd.

The desire to probe the heavens is as old as humanity; the realization that the human eye could be aided in these investigations is somewhat more recent. In May 1609, Galileo Galilei, while visiting in Venice, heard about the new ‘telescope’ invented by the Dutch lens-maker Hans Lipershay and decided that it would be useful in his astronomical investigations. He returned to his home in Padua and made his own telescope comprised of 4.2 cm diameter plano-convex and plano-concave lenses. With this telescope, Galileo discovered the satellites of Jupiter; got tenure; had his salary doubled – and the race was on! It did not take too long to realize that ‘bigger’ was ‘better’, and from this point forward, there has been a continuous striving to increase collecting area and resolution, thereby enabling the astronomer to ‘see’ farther and farther into the heavens.

The limitations of refractive optics soon prompted investigations into reflecting optics – mirrors. In 1668, Isaac Newton produced the first reflecting telescope having a 2.5 cm diameter mirror made of speculum (copper (6), tin (2), and arsenic (1)). A little over 100 years later, in 1789, William Herschel constructed his ‘Great Telescope’, having a 1.22 m diameter mirror made of bronze. Of course as telescopes grew, so did mirror weight. In 1845, William Parsons’ ‘Leviathan’ was built. It had a 1.84 m diameter mirror made of speculum that weighed 3600 kg – a whopping 1354 kg m^{-2}.

Problems with corrosion in speculum and bronze mirrors led to a renewed interest in glass, and the development of highly reflective coatings soon made metal mirrors obsolete. In 1917, the mirror for the Hooker telescope arrived on Mount Wilson. Made by the Saint Gobain glassworks in France from the same material used for wine bottles, the mirror was 2.54 m in diameter, 33 cm thick and weighed over 4000 kg (789 kg m^{-2}).

The increasing weight of larger and larger mirrors produced practical problems for telescope makers and astronomers. Herschel’s telescope was so unwieldy that he found it almost impossible to use and returned to smaller telescopes for most of his observations, and the 2.54 m Hooker telescope required a 100 ton mount and drive system. Obviously, if telescope apertures were to increase, steps had to be taken to make mirrors lighter. The first major success in this arena was accomplished for what was to become the Hale Telescope commissioned at the Palomar Observatory in 1948. In 1936, Corning cast a mirror made of a new borosilicate glass blend called Pyrex. The blank was lightweighted by placing hexagonal shaped obstructions in the mold which created corresponding voids in the glass. The finished mirror (Figure 1) was 5.08 m in diameter, 61 cm thick at the edge and weighed approximately 13 000 kg (641 kg m^{-2}). Even with this degree of ‘lightweighting’ the mounting and drive mechanisms still weighed around 182 000 kg.

Although the desire for larger ground-based telescopes provided the initial impetus for developing lightweight mirrors, the advent of the space program and the desire to put telescopes in space became the main driver in the pursuit of ‘serious’ lightweighting techniques. The cost of launching

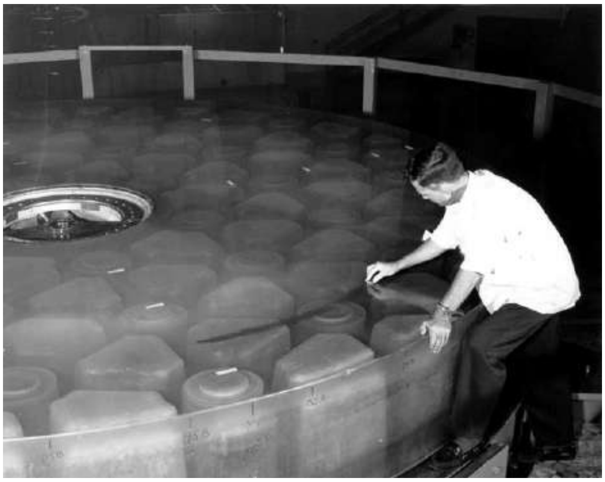

Figure 1 5.08 m Pyrex™ mirror made by Corning/Caltech for the Hale Telescope. Reproduced with permission from the Palomar Observatory/California Institute of Technology.

large telescopes into space quickly becomes prohibitive as aperture size increases. Consequently, the current emphasis is on developing ultra-lightweight, deployable telescopes that permit ever-increasing apertures to be placed in orbit. Fortunately, the zero gravity environment of space permits satisfactory operation of such ultra-lightweight telescopes and therefore the extent to which a mirror can be lightweighted is primarily dependent upon the mirror material and the manufacturing techniques utilized.

Manufacturing

Mirror manufacturing can, in some sense, be divided into two general classes, one that removes material (direct generation) and one that deposits material (replication). The two classes have significantly different limitations relative to producing lightweight mirrors. Direct generation encompasses the more traditional methods of mirror manufacturing, blank fabrication, and optical surface generation, i.e., grinding and polishing. Many innovative techniques have been developed over the past decade that have not only made significant strides in improving fabrication times, optical figure, and surface finish, but have also made possible the reduction of weight to unprecedented levels.

Although extremely lightweight mirrors are being fabricated using direct generation, in the end, there are limitations in the fabrication processes that are not present in replication. This process offers the fascinating potential to create mirrors that are only a few molecules thick, thereby bringing a new meaning to the term ‘lightweight optics’.

Direct Generation

Mirror blank fabrication

Casting lightweight glass mirror blanks has evolved substantially since 1936. Four 8.2 m diameter Zerodur™ mirror blanks were spun cast for the European Southern Observatory’s Very Large Telescope Interferometer. The largest mirror cast to date was the 8.4 m borosilicate mirror cast in a spinning oven for the Large Binocular Telescope on Mount Graham. Although not exactly lightweight by space standards – the blank weighs 15 455 kg (266 kg m^{-2}) – it is substantially less than a conventionally cast solid blank.

It has been postulated that an areal density of around 64 kg m^{-2} probably represents the limit of lightweight glass castings. Therefore, as the demand for lighter and lighter mirrors increased, methods other than casting had to be developed. Investigations in the 1960s indicated that mirrors could be lightweighted to about 180 kg m^{-2} by first fabricating a ‘honeycomb’ core and then fusing plates on the front and back of the core using a high-temperature fusion process. This process was used to produce the Hubble Space Telescope mirror (Figure 2). Hubble had a 2.5 cm thick face-sheet and back-sheet and a honeycomb interior constructed of individual pieces 0.5 cm thick. These pieces were assembled and placed in a large furnace where the temperature was increased until they fused together. The resulting 2.4 m diameter mirror weighed 828 kg (183 kg m^{-2}).

During the 1980s, Frit bonding technology was developed. Unlike the fusion process, Frit allowed the bonding of glass elements at a lower temperature. This allowed the cross-sections of the glass elements to be significantly reduced. The thicknesses of face-plates were typically reduced to around 7 mm and areal densities in the 40–50 kg m^{-2} range became routine. New advances were also made in core manufacturing. Diamond tooling was developed along with computer controlled systems to machine cores from solid blanks. High-pressure water jets were developed that allowed cores to be made much faster than with diamond tooling. In addition, the water jet cutting provided the ability to create very flexible geometries optimized to minimize weight while retaining stiffness. Water jets could machine core walls that could be held to a thickness of around 1.5 mm over a core depth of 20 cm. Computer controlled diamond machining has created even thinner cores (less than 1 mm in thickness) but the process requires longer generation times.

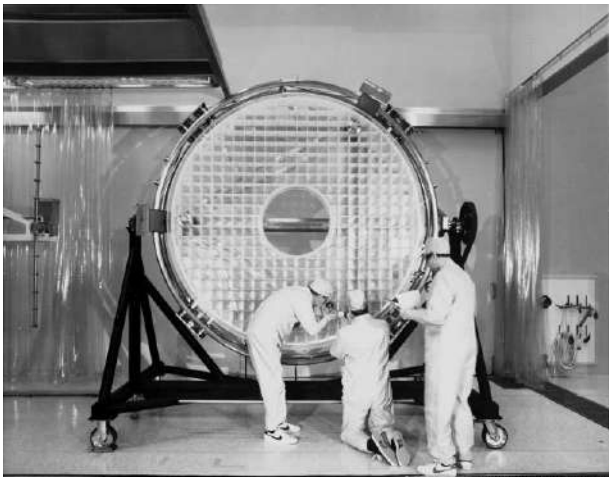

Figure 2 2.4 m ULE™ mirror made by Corning/Perkin Elmer for the Hubble Space Telescope. Photography courtesy of NASA.

Material removal

The evolution in blank manufacturing required a corresponding evolution in the process of material removal. Many new grinding and polishing techniques were developed over the years to more accurately and rapidly generate precision optical surfaces. These included adaptive force grinding, electrolytic in-process dressing (ELID), water jet figuring, computer controlled polishing, stress lap polishing, magneto-rhelogical figuring (MRF), plasma-assisted chemical etching (PACE), ion figuring, and vacuum lap figuring. Most of these processes induce localized forces on the face-sheet in the process of material removal, and for a 'honeycomb' mirror, the face-plate deflects away from the grinding/polishing tool as the tool reaches the center of the cell. This means that less material will be removed in the center of the cell than at the edge of the cell where the core wall holds the face-sheet against the tool. This creates a 'quilted' pattern in the surface that must be removed in subsequent steps. As face-sheets become thinner, quilting becomes more pronounced and the process of removing 'quilting' becomes exacerbated. For a number of years, this was the limiting factor in the development of high-precision lightweight mirrors.

Ion figuring is a noncontact process that has eliminated the problems associated with mechanical material removal. It bombards the surface of the optic with neutralized ions causing atoms to be dislodged. The process creates no deflection of the surface and therefore can remove quilting and other irregularities without generating further distortion in the surface. This has permitted 2–3 mm thick face-plates to be manufactured without concern for quilting effects. Primary limitations now being encountered are in face-plate generation and the various associated handling processes. Great care must be taken to avoid fracture. Spin casting ultra-thin face-plates is being investigated as a means of further reducing problems in the generation process. The 0.75 m mirror made for IKONOS, the high-resolution imaging satellite, represents the state-of-the-art in all-glass lightweight mirrors (28 kg m^{-2}).

Replication

Replication techniques are quite common in low-cost production mirrors and lenses for commercial purposes. However, there are also a number of techniques that hold great potential for precision mirrors as well. Chemical vapor deposition (CVD), plasma vapor deposition (PVD), plasma spray deposition, and electro-forming are all being investigated for producing lightweight precision mirrors. This class of mirror generation does not start with a mirror blank *per se*, rather it replicates a 'mandrel' that has the appropriate surface finish and the inverse of the shape desired in the final mirror. The mandrel can be made very rigid and therefore unperturbed by either gravitational forces or the localized forces generated by grinding and polishing. Extremely thin mirrors with areal densities on the order of 0.1 kg m^{-2} are currently being manufactured using replication techniques. These 'membrane' mirrors are currently being used for solar energy collection; however, recent advances in membrane quality and wavefront correction techniques hold out the promise of producing high-resolution, membrane imaging systems.

The Future?

Theoretically, single-molecule thick mirrors could be manufactured in space using self-replicating nanotechnology systems that 'grow' mirrors according to a mathematical description of the desired optical surface. These 'smart' mirrors could self-correct in the presence of physical disturbances and self-heal when damaged. They could be 'grown' to hitherto unimagined size, being limited only by the availability of raw material. Such mirrors, however, are definitely of the future and should merit re-examination 20 years hence.

Materials

A variety of materials have been developed over the years that have led to lighter and lighter mirrors. Recently, metals are coming back into play as means of fabricating stiff, lightweight mirrors. These metals include beryllium, nickel, nickel alloys, aluminum, and aluminum alloys. Composite materials are also playing a large role in lightweight mirror fabrication, both as a support structure and also as a reflecting surface (with appropriate coatings). Hybrid mirrors comprised of a combination of glass and composites have also been developed, as well as combinations of glass and beryllium. Various forms of silicon carbide have been extensively investigated and lightweight mirrors of pure silicon are being manufactured. Glass still plays a significant role in large lightweight mirrors, and various plastics are under investigation for the production of ultra-thin membrane mirrors.

Glass, Glass Ceramics, and Hybrids

The most common glass/glass ceramic materials used in large telescope mirrors are fused silica, ULE™ (Ultra-Low Expansion) glass, and Zerodur™, a glass ceramic. As was mentioned earlier, there are a variety

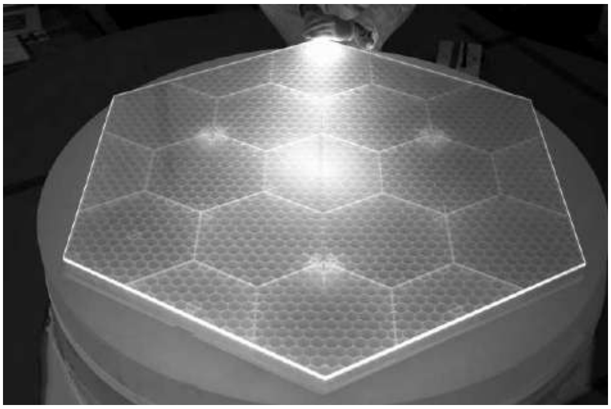

Figure 3 1.5 m ULE™ mirror made by Corning/Kodak for NASA's Next Generation Space Telescope Advanced Mirror System Demonstrator technology program. Photograph courtesy of NASA.

of methods available for the manufacture of glass mirrors depending upon the extent of lightweighting desired. Most of the current effort in lightweight mirrors is centered around hybrid mirrors. One method bonds very thin (2–3 mm) glass face-sheets to composite cores to form 'sandwich' mirrors. Another method attaches face-sheets to composite back-plane structures via actuators. In the latter case, the actuators can be used to actively control mirrors to correct for distortions. These mirrors were developed as part of the James Webb Space Telescope technology program, and typically have areal densities on the order of 15 kg m^{-2} (including back-plane structures). An all glass face-sheet and core is shown in Figure 3.

Silicon Carbide

Silicon carbide has long been a promising material for lightweight mirrors. It has many of the desirable properties of beryllium and is nontoxic. Until recently difficulties were encountered in producing mirrors larger than ~0.5 m in diameter. Of late, several manufacturers have produced large (<1 m) lightweight mirrors using different processes; sintering, liquid silicon infiltration and chemical vapor deposition.

The sintering process consists of preparing a finely ground silicon carbide powder mixed with organic binders and additives and then compacting the powder by isostatically pressing it at high pressure. The resulting compact is then machined to the desired shape and weight and the blank heated to cause agglomeration of the particles. The resulting material is known as reaction-bonded silicon carbide; it can be polished to around 30 angstroms and can be ion figured. If smoother surfaces are required, then a layer of CVD or PVD silicon carbide can be added. A 1.35 m diameter mirror has been fabricated as a demonstration for the European Space Agency's Herschel Space Observatory. The mirror was fabricated in nine segments that were then brazed together to form a unified 26 kg m^{-2} blank. This mirror was not designed to test the limits of lightweight and, consequently, there is no reason to believe that this process could not achieve an areal density comparable to the 7.8 kg m^{-2} of Cesic™ described in the following paragraph. The size of mirrors produced using this technique is currently limited to around 4 m due to the availability of the furnaces required for segment brazing.

The liquid infiltration process is used to create a biphase carbon, silicon carbide material known as Cesic™. This material is manufactured by creating a block of chopped carbon fibers embedded in a phenolic resin. The block is machined to the desired shape and then put into a furnace where the temperature is elevated until the resin is turned to carbon. The temperature is further elevated and liquid silicon infiltrated to produce silicon carbide. Bare Cesic™ can be polished to approximately 100 angstroms. In a subsequent step, a silicon carbide slurry can be applied to the blank that can be polished to around 30 angstroms. It is also possible to add a CVD or PVD silicon carbide layer to the blank that can be polished to obtain angstrom-level surfaces. A 1.04 m diameter Cesic™ mirror has been constructed for the NASA program, Solar Lite. The mirror weighs approximately 60 kg for an areal density of 56 kg m^{-2}. Again, this mirror was not designed to test the limits of lightweighting, but in a separate demonstration, an areal density of 7.8 kg m^{-2} was achieved. The size of mirrors produced using this technique is also currently limited to around 4 m due to the availability of existing furnaces required for infiltration.

The CVD process produces a SiC material that is highly polishable, yielding surface finishes on the order of angstroms. It has been used quite successfully as a coating layer, and has the potential to provide a means of replicating very lightweight mirrors by depositing a thin layer of SiC onto a mandrel. In one approach, the mandrel has the appropriate shape, but not the surface finish. In this case, once the SiC has been deposited, the mandrel is burned away and a SiC mirror blank is left. The blank is then ground and polished to produce the desired figure and finish. If the blank is too thin, this can present a problem in achieving a high-performance mirror. A second approach involves the deposition of SiC onto a figured and polished mandrel. In principle, both the figure and surface finish can be replicated. Success in this approach, however, has been somewhat limited, and although 20-angstrom surface finishes have been replicated, they have not been repeatable. A 1 m diameter blank was manufactured using the first approach, but it too has not been repeated and the lack of success appears to be due to high stress developed in the deposition process. The residual stress in the blank causes large-scale deformation and frequently fracture.

Beryllium

The extremely high stiffness of beryllium makes it a highly desirable candidate for lightweight mirrors. Unfortunately, it is a highly toxic material and there are only a very few companies involved in the manufacture of beryllium mirrors and even fewer that produce the raw material. Blanks are manufactured by isostatically compressing beryllium powder at high temperature and pressure. The blanks are then machined and subsequently ground and polished. Early beryllium mirrors often exhibited interesting, mysterious changes in figure after seemingly benign handling events. Unpredictable changes in figure also occurred as they were cycled between ambient and cryogenic temperatures. These problems appear to have been resolved with the production of spherical powder. The older powder consisted of irregularly shaped particles and when the powder was compressed, small, irregularly shaped voids were created throughout the material. It is hypothesized that these voids changed in size and shape as the blank was stressed, thereby causing unpredictable dimensional changes in the blank. The spherical powder permits uniform, high-density packing yielding homogeneous blanks that do not appear to exhibit unpredictable dimensional changes.

The lightest-weight beryllium mirror to date ($11.8\ \text{kg m}^{-2}$) was made for NASA's James Webb Space Telescope technology program. This technology was selected for the flight mirror. The weight reduction was achieved by machining the blank to create an open back honeycomb structure with an integral face-sheet. The ability to fabricate large beryllium mirrors is limited by the availability of facilities capable of isostatically pressing the material at the requisite pressures and temperatures. Currently, blank sizes are limited by available isostatic presses to around 1.2 m in width and 1.6 m in length. The expense of construction of larger facilities combined with limited need make it highly unlikely that any larger mirrors will be manufactured in the near term. A brazing process, however, has been demonstrated for joining segments to form a large blank. This process could, in principle, permit the fabrication of mirrors several meters in diameter. Bare beryllium can be polished to around 30 angstroms surface finishes. If smoother surfaces are required, beryllium can be vacuum deposited on the blank. Nickel may also be deposited on the blank; this has the added benefit that it can be polished without the special facilities required for the hazardous beryllium polishing operations. The drawback of using nickel is that of the bimetallic effect if the mirror is to be operated over a wide temperature range. The Vavilov Institute in Russia has developed an alternative approach to obtaining good surfaces by creating a glass that matches the coefficient of thermal expansion (CTE) of beryllium over a 50° temperature range. The glass is then fused to the beryllium blank. There have been numerous attempts over the years to produce replicated mirrors using a CVD process to deposit beryllium on a mold. So far these attempts have been unsuccessful, the process producing a material structure that no longer has high stiffness.

Silicon

Silicon as a mirror material has the advantage of being a very homogeneous material that can be polished to very smooth surfaces (angstrom level) very rapidly. A lightweight mirror has been fabricated using a silicon face-sheet brazed to a central core of silicon foam. This mirror 'sandwich' has demonstrated an areal density of $<15\ \text{kg m}^{-2}$. Technology has been developed to braze segments together in a manner that has been shown to have no adverse effects on either polishability or cryogenic performance (i.e., no distortion across the braze). Consequently, it should be possible to manufacture lightweight silicon mirrors of 1 m diameter or larger.

Composites

There have been many attempts to manufacture mirrors from composite materials, beginning initially with graphite/epoxy and more recently with graphite/cyanate-ester. These mirror are produced by laying up graphite sheets impregnated with epoxy, or cyanate-ester on a near net shape mold and then curing the material at elevated temperatures. The optical surface is typically created by coating another mold (this time with the desired optical surface) with a passivation layer and then with epoxy and placing the near net shape composite substrate on the epoxy. When the epoxy cures, the completed mirror is separated from the mold and the epoxy surface ion milled to remove residual errors such as fiber print-through. These mirrors can be very lightweight ($\sim 5\ \text{kg m}^{-2}$), but typically have problems with large-scale distortion due to inhomogeneities due to the lay-up process. The epoxy surface replication can also be a problem if the mirror is subjected to temperature extremes. A 3.5 m diameter composite mirror was made as a demonstration for the FIRST program. It has suitable performance for submillimeter and infrared wavelengths, but not for shorter wavelengths. At this time, the accuracy required for large (>1 m) visible imaging systems has not yet been achieved.

Gossamer Mirrors

The term 'gossamer' refers to that category of mirrors that in and of themselves have no structural rigidity beyond spatial scales of fractions of millimeters. Several materials are currently being used for these types of materials including Kapton™ and various other polyimide films. The initial shape of the membrane is formed by a variety of techniques including vacuum forming, pressure forming, spin casting, hot forming, and others. Early mirrors were manufactured by joining together triangular-shaped segments (gores) to form the desired shape. When deployed, these mirrors require some means of applying pressure to achieve the desired shape. Some methods have involved applying force at the edge to put the mirror in tension; some involve using a clear membrane as a face-sheet and then pressurizing the combination to create a 'lens' with a reflecting inner surface. Most recently, techniques have been demonstrated involving various forms of electrostatic deflection to achieve the desired shape after deployment. At present the primary applications are in the area of solar collection, power beaming, and thermal propulsion. The mirrors are approaching the quality required for imaging telescopes, particularly for small sizes. Techniques for correcting these mirrors with 'downstream' active optics have reached the point where it is not unrealistic to imagine that membrane optics could soon reach the accuracy required for imaging. There have been two space-based demonstrations, one by the Russians – a 20 m reflector – and the other by the USA – a 14 m inflatable antenna (Figure 4).

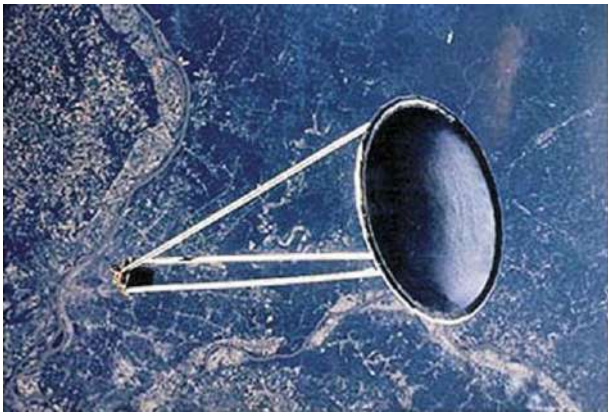

Figure 4 15 m antenna made by L'Garde for NASA's Inflatable Antenna Experiment. Photograph courtesy of NASA.

Grazing Incidence Mirrors

Broadband X-ray telescopes make use of the reflection property of materials at angles of grazing incidence. The most common design is based on the X-ray microscope objective designs of Wolter in the early part of the twentieth century. These mirrors are hyperboloids and paraboloids of revolution. This configuration provides a structural rigidity not present in normal incidence mirrors; consequently, X-ray mirrors can be made much lighter. The largest mirror of the Chandra X-ray Observatory is made from Zerodur™ (Figure 5). It is a 1.2 m diameter, 0.9 m long cylindrical mirror with a wall thickness of only 24 mm. Such a thickness in a normal incidence mirror of this size could be easily deformed.

Lightweight approaches in grazing incidence optics are almost all concentrated in various forms of replication. Recent developments in electroformed nickel alloy mirror technology have resulted in a 0.5 m diameter 0.6 m long mirror that weighs only 1.2 kg (Figure 6). A similar-sized mirror in the Chandra telescope weighs over 100 kg. Extremely lightweight cylindrical mirrors have been made from graphite/epoxy and from CVD SiC. Various investigations are underway in producing segments of cylinders that are then mounted to form a complete

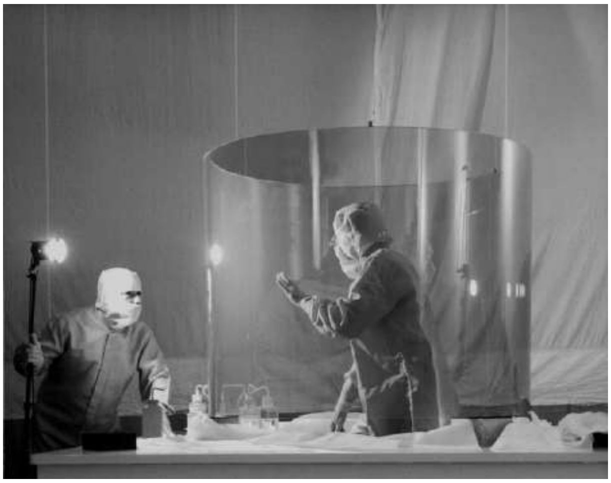

Figure 5 1.2 m Zerodur™ made by Schott Glass Werken/Hughes Danbury Optical Systems for NASA's Chandra X-ray Observatory. Photograph courtesy of NASA.

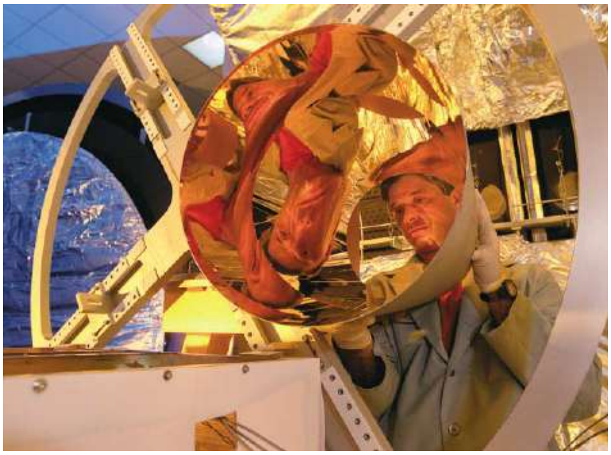

Figure 6 0.5 m electro-formed nickel–cobalt technology mirror made by NASA's George C. Marshall Space Flight Center for NASA's Constellation X-ray Mission technology program. Photography courtesy of NASA.

cylindrical mirror. These techniques involve a variety of materials including aluminum foil, and beryllium plates. In segment replication, the mandrel is passivated and then coated with the desired reflecting material of the final optic (the passivation process prevents the coating from adhering too strongly to the mandrel.) After the mandrel has been coated with the reflecting material, an epoxy coating is applied and a preformed shell segment is placed on the epoxy. After the epoxy cures, the segment can be removed from the mandrel along with the desired reflective surface (typically gold). In this case, the preformed shell does not have to have as precise a figure, or surface as the mandrel, since the epoxy will accommodate minor differences.

See also

Geometrical Optics: Lenses and Mirrors. **Imaging:** Adaptive Optics; Wavefront Sensors and Control (Imaging through Turbulence). **Instrumentation:** Astronomical Instrumentation; Telescopes. **Optical Materials**: Smart Optical Materials.

Further Reading

Bilbro J (2001) Optics in orbit. *SPIE OE Magazine* (Aug. issue).

Breckinridge JB and Jakobsen P (eds) (2000) UV, Optical, and IR Space Telescopes and Instruments. *Proceedings of SPIE* 4013.

Chen PC, Bowers CW, Content DA, *et al.* (2000) Advances in very lightweight composite mirror technology. *Optical Engineering* 39(9).

Ealey MA (ed.) (1995) Silicon carbide materials for optics and precision structures. *Proceedings of SPIE* 2543.

Engelhaupt DE, Ramsey BD, O'Dell SL, *et al.* (2000) New alloys for electroformed replicated X-ray optics. *Proceedings of SPIE* 4138: 22.

Goodman WA and Jacoby MT (2000) Dimensionally stable ultra-lightweight silicon optics for both cryogenic and high-energy laser applications. *Proceedings of SPIE* 4198: 260–270.

Jenkins C (vol. ed.) (2001) Gossamer spacecraft: membrane and inflatable structures technology for space applications. In: Zarchan P (series ed.) *Progress in Astronautics and Aeronautics* 191. Reston, VA: American Institute of Aetronautics and Astronautics.

Kendrick SE, Reed T and Streetman S (2001) In-process status of the 1.4-m beryllium semi-rigid Advanced Mirror System Demonstrator (AMSD). In: Stahl HP (ed.) *Optical Manufacturing and Testing,* vol. IV. *Proceeding of SPIE* 4451: 58–66.

Mathews G (2001) Semi-rigid active mirror technology advancements. *Proceedings of SPIE* 4451.

Müller C, Papenburg U, Goodman WA and Jacoby M (2000) C/SiC high precision lightweight components for optomechanical applications. *Proceedings of SPIE* 4198: 51–57.

Truemper JE and Aschenbach B (eds) (2000) X-Ray optics, instruments, and missions III. *Proceedings of SPIE* 4012.

Measurement of Optical Properties of Solids

P Lucas, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The interaction of light with solids takes place through different mechanisms, depending on the type of material and the range of wavelength investigated. Insulators or dielectrics are typically transparent to visible light while most semiconductors are opaque to visible light yet transparent to infrared radiations; in contrast metallic solids appear shiny because they reflect all wavelength up to the ultraviolet region. The optical properties of a solid depends on its chemical composition and its structural properties and vary for every material, though, one parameter, the complex refractive index **n**, is sufficient to characterize entirely the optical properties of a specific material.

In this article we will briefly review the physics underlining the nature of the refractive index **n**, in order to understand its relationship to the observable quantities such as absorption, reflection, and transmission routinely measured in optics experiments. A description of the experimental techniques, as well as a physical interpretation of the data, will also be presented.

Origin of Light Interaction with Matter

Light is described as a transverse electromagnetic wave consisting of an electric and magnetic field oscillating perpendicular to the direction of propagation at optical frequencies within $10^{13} \rightarrow 10^{17}$ Hz. This range of frequencies comprises the ultraviolet, visible, and infrared domain. On the other hand, solid mediums are composed of charged particles, negative electrons, and positive ions producing electric dipoles, which can be polarized under the action of an electric field. Hence, when electromagnetic radiation impinges upon a material it interacts by polarizing the molecular units, producing oscillating dipole moments. This interaction results in several observable optical phenomena such as reflection, transmission, absorption, or scattering, which will be described in more detail in a following section. The classical model of light propagation assumes that the oscillating electric field can interact with several different types of dipole oscillators within the material. Different dipoles are usually accessed by light wave from different frequency range depending on their mass.

Electronic Oscillators

Lorentz originally proposed that the electrons bound to the nucleus of an atom behave as oscillators. In this representation, electrons behave as if they were held by springs and subjected to a linear restoring force, therefore oscillating up and down (Figure 1a). This representation of an atom might seem to contradict the conventional picture of electrons orbiting around the nucleus; however, it is valid as far as light–matter interactions are concerned. The resonant frequency of such an oscillator is proportional to the inverse of its reduced mass μ defined as

$$\frac{1}{\mu} = \frac{1}{m_e} + \frac{1}{m_N} \qquad [1]$$

where m_N and m_e are the mass of the nucleus and electrons, respectively. The resonant frequency ω_0 is then related to μ according to

$$\omega_0 \propto \sqrt{\frac{1}{\mu}} \qquad [2]$$

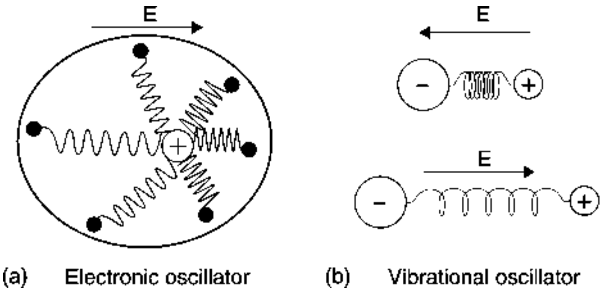

Figure 1 (a) Classical oscillator model of a polyelectronic atom. The electrons depicted as black dots are bound to the positive nucleus by springs, which represent the electrostatic restoring forces between the charged particles. The electronic oscillators are distorted under the effect of an electric field. (b) Classical model of a diatomic vibrational oscillator. The chemical bond is represented as a spring and the charged ions vibrate around their equilibrium position under the effect of an oscillating electric field.

Considering that $m_N \ll m_e$ we can assume that $\mu = m_e$ and show that the smaller mass of the electron determines the resonant frequency of the oscillator. Hence, resonant electronic oscillations are accessed at higher frequencies in the ultraviolet and visible region of the spectrum. In solids, the quantum theory shows that these resonances correspond to excitation of electrons across the gap between a full valence band and an empty conduction band.

Vibrational Oscillators

The second type of dipoles contained in the medium are based on pairs of charged atoms vibrating around their equilibrium position in the solid. Every atom has a different electronegativity, hence the electronic density on atomic pairs is distorted and result in a charge imbalance, which can couple to the electric field of light. This generates oscillating dipoles, which involves the motion of atomic nucleus (Figures 1b). The nuclear mass being several orders of magnitude larger than the electronic mass, we can show by analogy with eqns [1] and [2] that the resonant vibration will occur at lower frequencies in the infrared region. In solids, the quantum theory shows that the lattice atoms vibrate cooperatively and generate quantized lattice waves called phonons.

Free Electron Oscillators

In contrast with the bound electrons from the electronic oscillator model, free electrons can move without being subjected to a restoring force. These correspond to the conduction electrons present in metals and to a smaller extent in doped semiconductors. The free electron oscillator model is therefore

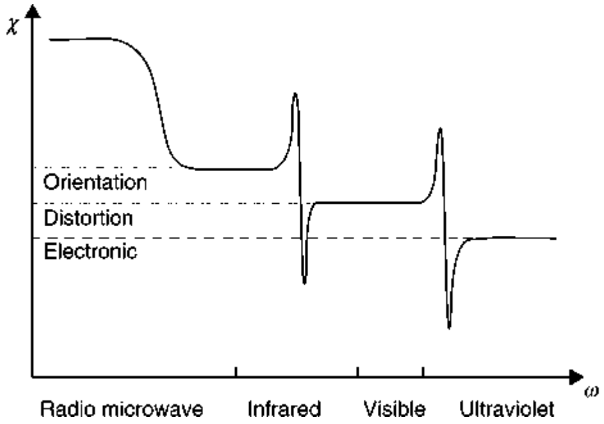

Figure 2 Contributions to the polarizability of a medium as a function of light frequency. At optical frequencies, the molecules cannot re-orient fast enough to follow the reversing electric field and the contribution of dipole orientation is lost.

principally used to describe the optical properties of metals such as reflectivity. The free electrons are well described as conventional oscillator with a natural resonant frequency $\omega_0 = 0$.

The relationship between the applied electric field E and the resulting polarization of the medium P is expressed through the dielectric susceptibility χ as

$$P = \varepsilon_0 \chi E \qquad [3]$$

where ε_0 is the permittivity of vacuum. We can generally recognize three contributions to the dielectric susceptibility or the polarizability of an optical material (Figure 2). Applied fields in the radio or microwave region oscillate slowly enough that polar molecular units can reorient and align in the direction of the field. The characteristic time for orientational motion is fairly long and this contribution is quickly lost when the molecules cannot follow the field oscillating at higher frequency. When the field reaches optical frequency, the two remaining contributions are the vibrational and electronic dipoles mentioned above. The vibrational dipoles involve a distortion of the chemical bond and are the slowest process occurring only in the infrared region. This contribution is lost at visible frequencies and leaves only the fast electronic polarization processes.

If we treat it rigorously, the polarization P is actually expressed as a function of higher order terms according to

$$P = \varepsilon_0 \left(\chi E + \chi^{(2)} EE + \chi^{(3)} EEE + \ldots \right) \qquad [4]$$

This gives rise to nonlinear effects such as self-focusing and second or third harmonic generation. However, the higher order susceptibilities $\chi^{(n)}$ are very small and these effects are only significant for very high intensity light source such as lasers. In the following text we will only consider linear optical processes.

The Dipole Oscillator Model (Lorentz Oscillator)

In the Lorentz model, the bound electrons behave as harmonic oscillators and are constrained to vibrate at their natural resonant frequency ω_0. These oscillators are also submitted to a damping force corresponding to a resistance to the motion due to collisional processes. The equation of motion is then

$$m\left(\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + \gamma \frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x \right) = qE \qquad [5]$$

where x is the displacement along E, γ is the damping factor, m is the mass of the electron and q its charge. The passage of an electromagnetic wave through a medium exerts an oscillating force on the electrons causing them to vibrate up and down. In other words these electronic dipoles are submitted to a forced oscillation induced by an electric field varying periodically with time. The electric field felt by the atomic dipoles is written as

$$E = E_0 e^{i\omega t} \qquad [6]$$

where ω is the frequency of the light.

The electric field of the light wave then drives the dipole oscillations at its own frequency ω and the dipoles displacement oscillates according to

$$x = x_0 e^{i\omega t} \qquad [7]$$

Solving eqn [5] for x in terms of E gives the expression for the displacement of the electrons

$$x = \frac{q/m}{\omega_0^2 - \omega^2 + i\gamma\omega} E \qquad [8]$$

The atomic dipole moment p induced by the electrons displacement is therefore

$$p = qx = \frac{q^2/m}{\omega_0^2 - \omega^2 + i\gamma\omega} E \qquad [9]$$

And for a medium with N atoms per unit volume the total polarization P is

$$P = N \frac{q^2/m}{\omega_0^2 - \omega^2 + i\gamma\omega} E \qquad [10]$$

Our result in eqn [10] only contains one natural frequency ω_0 and therefore only accounts for one single type of oscillator in the medium. However, normal optical mediums are composed of different kinds of atoms, each with several characteristic resonant frequencies ω_j. In our model, we can simply treat these multiple resonances by considering that each oscillator acts separately and add the contributions of all oscillators.

$$P = \left(\frac{q^2}{m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j \omega} \right) E \qquad [11]$$

Comparing eqns [3] and [11] then gives an expression for the polarizability χ of the medium

$$\chi = \frac{q^2}{\varepsilon_0 m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j \omega} \qquad [12]$$

Additionally, the polarization P is associated with Maxwell's displacement vector, D, through the relationship

$$D = \varepsilon_0 E + P \quad \text{and} \quad D = \varepsilon E \qquad [13]$$

where ε is the permittivity of the medium. The permittivity ε is normally expressed relative to the permittivity of vacuum ε_0 in terms of the dimensionless quantity $\varepsilon_r = \varepsilon/\varepsilon_0$. Where ε_r is the familiar dielectric constant of a material, which is related to the refracted index n as $n = \sqrt{\varepsilon_r}$.

Combining eqns [3] and [13] we obtain an important relationship linking χ to ε_r, and in turn χ to n.

$$\varepsilon_r = 1 + \chi \quad \text{or} \quad n^2 = 1 + \chi \qquad [14]$$

Finally, combining eqns [12] and [14] we obtain the expression for n in terms of the oscillator model.

$$n^2 = 1 + \frac{q^2}{\varepsilon_0 m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j \omega} \qquad [15]$$

The Refractive Index

Two important points should be made concerning the result of eqn [15]. The refractive index is a wavelength-dependent quantity and is a complex quantity. The complex refractive index **n** is indeed usually expressed as

$$\mathbf{n} = n + i\kappa \qquad [16]$$

where n is the real refractive index also defined as the ratio of the wave velocity in vacuum to the velocity in the medium $n = c/v$. And κ is the extinction coefficient, which is directly related to the absorption coefficient α as we shall see in the next section.

As we mentioned in the introductory section, the complex refractive index **n** is sufficient to characterize the optical properties of the solid. The real part of the index describes the change in velocity or wavelength of a wave propagating from a vacuum into a medium while the imaginary part is a measure of the dissipation rate of the wave in the medium.

By inspection of eqn [15] we can predict the form of the variation of n with wavelength. At frequencies ω below the resonant frequency ω_j, the terms ω^2 and $i\gamma_j\omega$ are much smaller than ω_j^2 and to a first approximation n is constant. When ω reaches a resonant frequency ω_j, the term $(\omega_j^2 - \omega^2)$ goes to zero and since the term $i\gamma_j\omega$ is very small, the value of n increases very sharply and exhibits a resonance line as depicted in Figure 3.

On closer inspection of eqn [15] however, we can see that n is not exactly constant in between resonance lines. As ω rises, the denominator slightly decreases and therefore n rises slowly with frequency. The refractive index is then slightly larger for wave of higher frequency. This is why a glass prism bends blue light more strongly than red light. This phenomenon is called dispersion as waves of different frequencies are 'dispersed' by the prism.

Another important point raised by eqn [15] is that because $(\omega_j^2 - \omega^2)$ goes to zero at a resonant frequency, the damping term $i\gamma_j\omega$ dominates and n becomes almost entirely imaginary. **n** is then mostly governed by the extinction coefficient κ. The damping associated with the coefficient γ_j corresponds to a friction force causing a loss of energy, which becomes the dominant effect during resonance. The resonance region is then associated with strong attenuation or

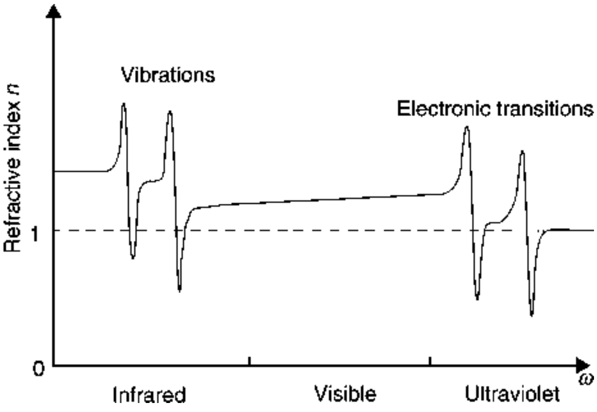

Figure 3 Variation of the refractive index of a hypothetical dielectric solid in the optical frequency range. At high frequency, ω is larger than all resonant frequencies and the refractive index n reaches unity.

absorption of the wave. In contrast, between resonant frequencies, the refractive index is almost entirely real as the imaginary part $i\gamma_j\omega$ is negligible in comparison with $(\omega_j^2 - \omega^2)$. Absorption is then very small and that region corresponds to the transparency domain of the material. In fact, we can assume that $\mathbf{n} = n$ in the transparent region, that is why tables of optical constants of transparent optical materials only list the real part of the refractive index n. However, at resonant frequencies on either side of the transparency domain, absorption becomes significant and the imaginary part of the refractive index must be taken into account so that $\mathbf{n} = n + i\kappa$.

The refractive index $\mathbf{n}$ being a complex quantity it ensues that ε_r must also be complex, since $n = \sqrt{\varepsilon_r}$. Hence we also define the complex dielectric constant as

$$\varepsilon_r = \varepsilon_1 + i\varepsilon_2 \qquad [17]$$

Both parameters describe equivalently the optical properties of a solid so that $\mathbf{n}$ or ε_r are called the optical constants of the material. Note that it is an unfortunate misnomer since they vary distinctly with wavelength (Figure 3).

Using equation [17] and $n = \sqrt{\varepsilon_r}$, we can establish the relationships between the real and imaginary parts of $\mathbf{n}$ and ε_r.

$$\varepsilon_1 = n^2 - \kappa^2 \qquad [18]$$

$$\varepsilon_2 = 2n\kappa \qquad [19]$$

We can see that $\mathbf{n}$ and ε_r are not independent variables but that the real and imaginary part of ε_r can be calculated, knowing the real and imaginary part of $\mathbf{n}$ and vice versa.

Moreover, it can be shown that the real and imaginary parts of each individual parameter are not independent either but can be computed from one another with use of the following Kramers–Kronig relations.

$$\varepsilon_1(\omega) = 1 + \frac{2}{\pi}\int_0^\infty \frac{\omega'\varepsilon_2(\omega')}{\omega'2 + \omega^2}\mathrm{d}\omega' \qquad [20]$$

$$\varepsilon_2(\omega) = -\frac{2\omega}{\pi}\int_0^\infty \frac{\varepsilon_1(\omega')}{\omega'2 + \omega^2}\mathrm{d}\omega' \qquad [21]$$

In these equations, only the principal value of the integral is calculated so that ε_1 can be computed if ε_2 is known over all frequencies and vice versa. The same equation connect n and κ together.

This analysis shows that it is only necessary to know one real or one imaginary part over a wide frequency range to determine all others. The interaction of light with matter is in fact characterized by a single independent real quantity.

Absorption

All electromagnetic phenomena are governed by Maxwell's equations. The form of an electromagnetic wave propagating through a medium is obtained by solving Maxwell's equation. For a wave propagating along z through a homogeneous, optically isotropic medium the form of the oscillating electric field is given by:

$$E = E_0 e^{i(kz - \omega t)} \qquad [22]$$

where k is the wave vector and ω the angular frequency of the light. For a wave traveling into a transparent medium of refractive index n, the relationship between k and ω is

$$k = \frac{n\omega}{c} \qquad [23]$$

However, for an ordinary optical material there are regions of absorption and the refractive index is therefore complex. We should then rewrite eqn [23] as

$$k = (n + i\kappa)\frac{\omega}{c} \qquad [24]$$

By substituting this value of k in eqn [22] we obtain an expression with two exponential terms

$$E = E_0 e^{-\frac{\omega\kappa z}{c}} e^{i\left(\frac{\omega n z}{c} - \omega t\right)} \qquad [25]$$

The term $e^{i\left(\frac{\omega n z}{c} - \omega t\right)}$ represents a wave traveling at a speed $v = c/n$, while the term $E_0 e^{-(\omega\kappa z/c)}$ represent the amplitude of this wave which decays exponentially with distance z. The intensity I of the wave is proportional to the square of the amplitude so that

$$I \propto e^{-\frac{2\omega\kappa z}{c}} \qquad \text{or} \qquad I \propto e^{-\alpha z} \qquad [26]$$

where α is the familiar absorption coefficient routinely measured by absorption spectroscopy. Hence we obtain an important relationship relating the observable quantity α to the imaginary part of the refractive index κ according to

$$\alpha = \frac{2\kappa\omega}{c} \qquad [27]$$

Local Field Correction

It should be pointed out that the derivation of our refractive index expression eqn [15] assumes that only the electric field of the light affects the polarization of the dipole oscillators. However, this assumption is not correct in dense materials because of the contribution from the electric field of neighboring atomic dipoles. The field generated by each dipole will affect all other dipoles in the vicinity. It is possible to approximate this contribution by accounting for the effect of dipoles within a local sphere. The result of this approximation is known as the Clausius–Mossotti equation, which is accurate for isotropic medium such as glasses and cubic crystals

$$\frac{n^2-1}{n^2+2} = \frac{\chi}{3} \qquad [28]$$

Measurable Optical Parameters

Four things can happen to a beam of light propagating through a thick slab of optical material (Figure 4). Some of the light can be reflected at the surface of the solid, some can be absorbed by the sample, some can be scattered in different directions, and some of the light can be transmitted through the sample. The effect on the light beam, resulting from these phenomena, can be quantified by a number of optical coefficients, which characterize the macroscopic properties of the material.

Reflection of light at the surface of a solid is described by the reflectance R defined as the ratio of the reflected intensity I to the incident intensity I_0. For a beam falling perpendicularly on a flat surface, the reflection is called specular and is governed by the complex refractive index according to the Fresnel equation

$$R = \frac{I}{I_0} = \frac{(n-1)^2+\kappa^2}{(n+1)^2+\kappa^2} \qquad [29]$$

This provides us with a second important formula relating a measurable quantity (R) to the optical constant of the material. For measurements performed within the transparency region of a solid, the value of κ is much less than n and eqn [29] reduces to the more familiar form

$$R = \frac{(n-1)^2}{(n+1)^2} \qquad [30]$$

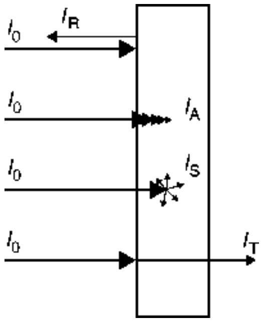

Figure 4 Representation of the four main optical processes happening to a beam as it propagates through a slab of transparent material.

Absorption occurs when the light frequency reaches the natural resonance of some dipole oscillators in the medium. The energy of the electromagnetic wave is partly transferred to the material and generally dissipated in the form of heat. The light intensity is therefore attenuated as it propagates through the material and the attenuation efficiency is quantified by the absorption coefficient α. The intensity decreases exponentially with pathlength z according to Beer's law

$$I(z) = I_0 e^{-\alpha z} \qquad [31]$$

where I_0 is the incident intensity. The imaginary part κ of the optical constant can then be directly obtained from measurements of α using eqn [27].

Scattering is the process whereby light is redirected in different directions due to microscopic density fluctuation in the sample produced by defects, impurities, or structural inhomogeneities. The intensity of light propagating in the forward direction is attenuated by the scattering event and it can be quantified in a way equivalent to absorption. The intensity has an exponential dependence on pathlength z analogous to Beer's law

$$I(z) = I_0 e^{-Sz} \qquad [32]$$

where S is the scattering coefficient. When the scattering center is smaller than the wavelength of light this phenomenon is called Rayleigh scattering and the scattering coefficient S vary with the inverse fourth power of the wavelength

$$S(\lambda) \propto \frac{1}{\lambda^4} \qquad [33]$$

By measuring light attenuation we cannot tell the difference between absorption and scattering and the total attenuation is $\alpha_{\mathrm{Tot}} = \alpha + S$. However, the scattering contribution is generally much weaker than the absorption and can be neglected so that $\alpha_{\mathrm{Tot}} = \alpha$.

Transmission occurs when the light is neither reflected, absorbed, or scattered. The beam is then transmitted through the sample and the fraction of light exiting the back surface is quantified as the transmittance T. If we disregard scattering in comparison to absorption then conservation of energy require that

$$R + T + A = 1 \qquad [34]$$

where A is the fraction of light absorbed. The transmittance can then be compiled from R and α for an incident beam I_0 traveling across a sample of thickness l. In this case we must consider the reflection of the light on the front surface as well as on the back surface when it exits the transparent medium. The expression for the transmittance accounting for dual reflection is

$$T = (1 - R)^2 e^{-\alpha l} \qquad [35]$$

It should be noted that a rigorous treatment of eqn [35] should account for multiple reflections between the front and back surface in the interior of the sample. In this case, interference effects complicate the determination of **n** and a modified version of eqn [35] can be computed by summing up the intensities due to the contribution of the multiple reflections.

Transmission and reflection measurement are readily obtained using conventional spectrometers. Two types of spectrometers must be used to cover the entire optical spectral region. FTIR (Fourier Transform Infra Red) spectrometers use a glow bar as the light source, which enable it to cover the infrared and near infrared domain. The UV-VIS (ultraviolet visible) double beam spectrometers use a tungsten and deuterium lamp and cover the spectrum from the ultraviolet down to the near infrared. Using eqns [27], [29] and [35], the reflection and transmission measurements thus obtained are used to calculate the real and imaginary component of the optical constant over the entire optical frequency range.

An example of transmission measurement is shown in Figure 5, for a dielectric and a semiconductor. The two spectra have the same principal features. At short wavelength the edge of the transparency window is due to absorption of energy by the valence electrons. In classical terms this corresponds to the resonance of electronic oscillators, and in quantum mechanical terms to the excitation of electrons from the valence to the conduction band. This edge is due to a sharp increase of κ, which extends over a significant range of frequency corresponding to the domain of opacity or high absorption. The wavelength at the edge determines the minimum energy of the photon necessary to promote an electron across the bandgap E$_g$. Insulators have a large bandgap and appear transparent to the human eye because visible light is not energetic enough to promote an electron across E$_g$. In contrast, semiconductors have a smaller bandgap and appear black (opaque) because all the visible light is absorbed to induce electronic transitions. On the low frequency side the transparency window is limited by atomic vibrations in the solid structure. The wavelength of the vibrations increases with the mass of the atomic oscillators. Semiconductor compounds have higher atomic masses and therefore transmit at longer wavelength in the infrared while insulator have lower atomic mass and exhibit their vibrational absorption edge at shorter wavelength.

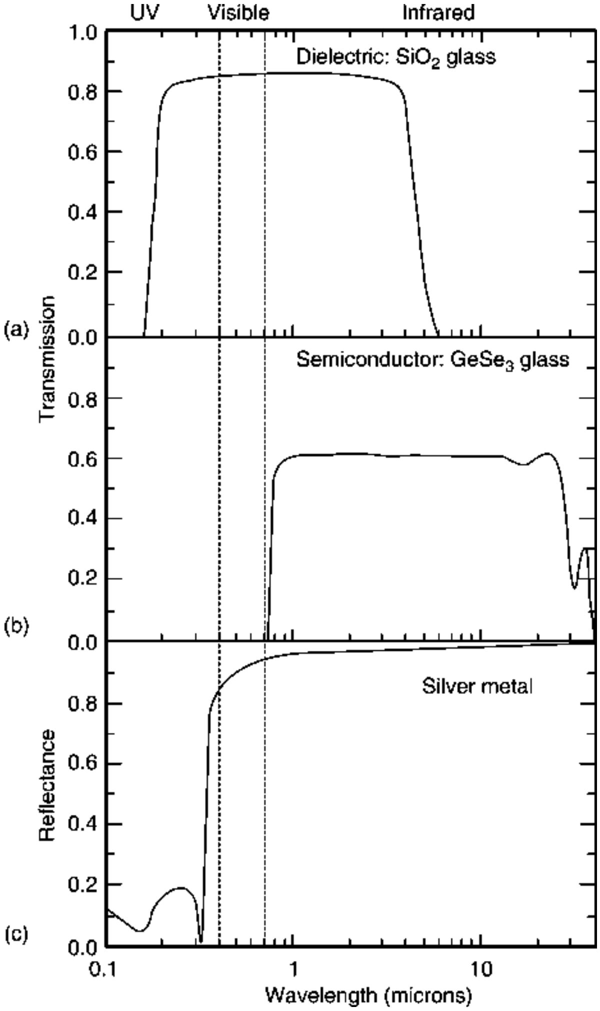

Figure 5 (a) Transmission spectrum of SiO$_2$ glass, a standard dielectric solid. (b) Transmission spectrum of GeSe$_3$ glass, a semiconductor (E$_g$ = 1.6 eV). (c) Reflectance spectrum of silver metal.

In between the two absorption regions, κ is very small and the material is transparent. The refractive index is almost entirely governed by the real part n. Eqn [15] shows that n is proportional to the number N of electronic oscillators. The refractive index will then be higher for heavy atoms with a high electronic

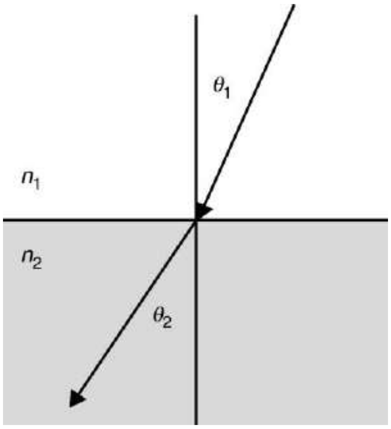

Figure 6 Refraction of a beam propagating across the interface between two mediums of different refractive index. The incident angle and the refraction angle are related to the index of refraction through Snell's Law.

density. Consequently, semiconductors exhibit a maximum transmission of only 60% because of significant surface reflectivity (eqn [35]) resulting from their higher refractive index n (eqn [30]).

Most applications of materials in optics are only concerned with the transparency domain so that the useful optical constant is reduced to the real part n. Hence a number of experimental techniques have been developed to obtain n, often by measuring angles of refraction with the use of Snell's law

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad [36]$$

where θ_1 is the angle of the beam incident in the medium of index n_1, and θ_2 the angle of the beam refracted at the interface with a medium of index n_2 (Figure 6). The basic principle is to compare the sample with a standard glass of known refractive index and measure the refraction angle at their interface. The sample's index can then be obtained following eqn [36]. Since n varies with wavelength, these techniques are normally performed with monochromatic light and provide n at a single wavelength. Among these methods are: the *Abbe refractometer*, using a glass hemisphere as a standard, the *V-block refractometer*, using a V shaped prism as a standard, and the *prism goniometer* based on the relative deviation between a known glass prism and a prism shaped sample.

Ellipsometry is another technique widely used to measure the index n. This technique is normally used on thin films deposited on substrate, as it allows to simultaneously measure the thickness and the refractive index. However, the method can also be applied to bulk samples. The principle of ellipsometry is based on measuring the change in polarization of a beam reflected off the sample surface as a function of incidence angle. The reflected light is elliptically polarized to an extent depending on n. This technique is especially useful to measure n at frequency range above the absorption edge where the sample is highly absorbing.

Metals

The optical properties of metals are mainly characterized by their very high reflectivity, which causes their shiny appearance. All metals reflect light in the infrared and visible region up to a cutoff frequency in the ultraviolet. This critical frequency is called the plasma frequency ω_p. The physical significance of the plasma frequency can be understood using the Drude–Lorentz model of the free electron oscillator. As mentioned previously, the free electron can be described accurately as dipole oscillators with a resonant frequency $\omega_0 = 0$. If we consider a system with negligible damping, we can introduce $\gamma = 0$ and by assuming all free electrons to be equivalent we can consider only one type of oscillators so that eqn [15] reduces to

$$n^2 = 1 - \frac{\omega_p^2}{\omega^2} \qquad [37]$$

where

$$\omega_p^2 = \frac{Nq^2}{\varepsilon_0 m_e} \qquad [38]$$

Equation [37] means that in the low frequency domain $\omega < \omega_p$, the refractive index n must be imaginary ($n^2 < 0$) and in the high frequency domain $\omega > \omega_p$, n is a real positive number. The low frequency region is therefore dominated by the extinction coefficient κ and the reflectance R, given by eqn [29], is essentially unity. In the high frequency limit, R decreases and n is real. These features are shown in Figure 5c. Metals are opaque and highly reflective below the plasma frequency, while they become transparent in the ultraviolet.

List of Units and Nomenclature

A	absorbance
c	light velocity in vacuum
D	electric displacement
E	electric field
I	intensity of light (power per unit surface)
k	wave vector
m_e	mass of electron
m_N	mass of nucleus
$\mathbf{n}$	complex refractive index
n	real refractive index

p	dipole moment
P	polarization
q	charge of electron
R	reflectance
S	scattering coefficient
T	transmittance
v	phase velocity
α	absorption coefficient
ε_0	permittivity of vacuum
ε_r	relative dielectric constant
κ	extinction coefficient
λ	wavelength
μ	reduced mass
χ	dielectric susceptibility
ω	angular frequency
ω_0	resonant frequency
ω_p	plasma frequency

See also

Instrumentation: Ellipsometry; Spectrometers. **Scattering:** Scattering from Surfaces and Thin Films.

Further Reading

Abeles F (1972) *Optical Properties of Solids*. North Holland Publishing Company.

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge University Press.

Dragoman D and Dragoman M (2002) *Optical Characterization of Solids*. Springer.

Feynman RP (1970). *Lectures on Physics*, vol. I–III. Addison–Wesley.

Fox M (2001) *Optical Properties of Solids*. Oxford University Press.

Hecht E (1998) Optics, 3rd edn. Addison–Wesley.

Klingshirn CF (1997) *Semiconductor Optics*. Springer.

Nazarewick W, Rolland P, Dasilva E and Balkanski M (1962) ABAC chart for fast calculation of the absorption and reflection coefficients. *Applied Optics* 1(3): 369–370.

Palik ED (1985) *Handbook of Optical Constants of Solids*. Academic Press.

Savage JAR (1990) Signals radar. *Optical Materials* 1:323–419.

Simmons CJ and El-Bayoumi OH (1993) *Experimental Techniques of Glass Science*. The American Ceramic Society.

Simmons JH and Potter KS (2000) *Optical Materials*. Academic Press.

Optical Glasses

J E Shelby, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

While advances continue in this field, some optical glasses have histories which can be traced to the seventeenth century. These glasses are produced by direct melting of raw materials of relatively high purity. Modern, or nontraditional, optical glasses are more commonly used to transmit light outside the visible region and are often made by nontraditional methods, including many variations on chemical vapor deposition, reactive melt processing, and sol–gel techniques. These glasses are used for ultraviolet and infrared lenses and other optical components, for optical fibers, for glass lasers, and for photonic devices such as optical isolators. Purity requirements and production demands for these glasses are significantly greater than those for traditional optical glasses.

Traditional Optical Glasses

Traditional optical glasses are primarily used as components such as lenses, prisms, or windows for optical systems operating in the visible region. These applications require glasses of high homogeneity, with precisely known refractive indices and dispersions. Glasses, as defined in this article, are nonmetallic, inorganic materials which have no long-range atomic structure and which display a temperature region of time-dependent properties.

Crown and Flint Glasses

Traditional optical glasses are designated as either crown or flint glasses. Crown glasses were originally based on the soda-lime-silica ternary system and were developed for use as window glass. The crown process involved spinning a spherical gob of molten glass to form a flattened disk which was much thicker in the center than at the edges. The edges of these disks were used for window panes, while the ‘crown’ in the center was often used as a crude lens. Crown glasses have low refractive indices and low to moderate dispersions.

The first flint glasses were potash-lead-silicate glasses prepared using English ‘flint’ as the source of silica. The presence of lead in these glasses results in higher refractive indices and dispersions than those of crown glasses.

The designations of crown and flint have since been expanded to include glasses from many other compositional systems. Glasses with lower refractive indices and dispersions are designated as crowns, while those with higher refractive indices and dispersions are designated as flints. Modern crown glasses are based on silicate, phosphate, and borosilicate systems. Additions of fluorine are particularly useful for production of low refractive index, low dispersion glasses, and for adjusting these parameters to meet specific needs. The lowest combinations of refractive index and dispersion are found for glasses based on BeF_2, while the highest combinations of refractive index and dispersion are found in rare earth borate systems. Other common optical glasses contain barium or lanthanum, which can be used to produce either crown or flint glasses, depending on the other components of the glass.

The need for glasses with different combinations of refractive index and dispersion can be traced back to the development of the telescope. Use of lenses made of a single glass made it difficult to obtain sharp images. Newton recognized that this problem was due to divergence of light of different wavelengths while passing through a lens. He felt that this problem was inherent to lenses and could not be eliminated. Others eventually recognized that a compound lens containing a positive crown glass and a negative flint glass would result in cancellation of the divergences and produce a sharper image. While this development did not totally eliminate the colored fringes in telescope images due to the presence of the secondary spectrum, it did lead to major improvements in optical imaging. Binary compound lenses are still fashioned from a combination of crown and flint glasses, with much greater flexibility today due to development of glasses with much more varied combinations of refractive index and dispersion.

Notation for Optical Glasses

Since the refractive index and the dispersion are the two most important characteristics of an optical glass for use as a lens, a notation system has been developed to designate any glass by a six digit number. This notation system requires that the dispersion be expressed by the Abbe number, ν_d, which is defined as

$$\nu_d = (n_d - 1)/(n_F - n_c) \qquad [1]$$

where d, F, and c indicate the value of the refractive index measured at the yellow helium line at 587.6 nm, the blue hydrogen line at 486.1 nm, and the red hydrogen line at 656.3 nm, respectively. Since the Abbe number is a reciprocal dispersion, a large Abbe number indicates a low dispersion, while a small Abbe number indicates a larger dispersion. A glass is then designated by the expression

$$1000(n_d - 1)(10\nu_d) \qquad [2]$$

where each term is rounded to 3 significant digits. A common crown glass known as BK7, for example, which has a n_d of 1.516 800 and ν_d of 64.17 is designated as 517 642. A very high index flint glass, with $n_d = 1.807\,410$ and $\nu_d = 31.61$ would be designated as 807 316. This system can be used for all but the exceptional glasses with $n_d \geq 2.00$. Some tabulations are based on the green mercury line, or e line, at 546.1 nm instead of the yellow helium (d) line. The choice of the d or e line results in a small difference in the reported refractive index and Abbe number for any given glass (Figure 1).

Dispersion Formulae for Optical Glasses

The combination of a refractive index near the middle of the visible region of the spectrum and the Abbe number describe the contribution of an optical glass to the refractive power and chromatic aberration of a lens. This information was sufficient for many early applications of lenses and is still sufficient for the design of optical devices of modest expectations. More sophisticated optical devices require more information for proper design. The simple Abbe number is often insufficient for expressing the dispersion, since it is necessary to provide a

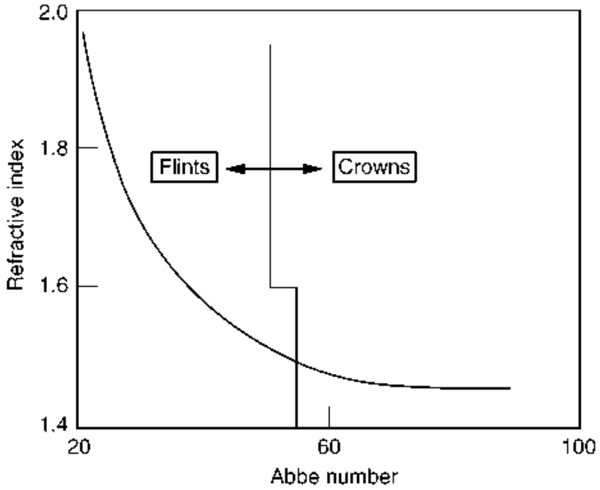

Figure 1 Relation between the refractive index and Abbe number of optical glasses. Commerical optical glasses lie in the regions labeled flint and crown above the curve.

detailed set of values of the refractive index as a function of wavelength across the entire visible region. The refractive index can be listed at a number of wavelengths, or described by an equation covering the entire wavelength range of interest. Although there are a number of expressions for the wavelength dependence of the refractive index, the most commonly used are the Cauchy dispersion formula, or

$$n = a + b\lambda^{-2} + c\lambda^{-4} \qquad [3]$$

where a, b, and c are constants derived from a best fit of the data to this expression, and variations of the Sellmeier dispersion formula, which is often simplified to the expression

$$n^2 = A_0 + A_1\lambda^2 + A_2\lambda^{-2} + A_3\lambda^{-4} + A_4\lambda^{-6} + A_5\lambda^{-8} \qquad [4]$$

where the six A_n constants are tabulated. The Cauchy dispersion formula is accurate to $\approx 10^{-4}$, while the modified Sellmeier dispersion formula is accurate to $\approx 3 \times 10^{-6}$. The validity of the latter expression can be extended further into the ultraviolet by adding a λ^{-10} term and into the infrared by adding a λ^4 term.

Annealing of Optical Glasses

Much of our knowledge of the behavior of glasses in the glass transformation region is based on studies of the annealing of optical glasses. Glass-forming melts have much higher thermal expansion coefficients than those of solid glasses of the same composition. During cooling of a melt, the density increases rapidly so long as the melt is relatively fluid. As the viscosity increases, the rate of structural rearrangement of the melt slows and eventually becomes fixed at some effective, or fictive, temperature. If the melt is cooled rapidly, the structure of the glass will be representative of a high temperature melt. If the melt is cooled more slowly, the structure will have time to equilibrate at lower temperatures and the glass formed will have the structure of a lower temperature melt and hence will be more dense. Since the refractive index of a solid is related to its density, the refractive index will be less for the glass produced by rapid cooling than for one produced by slower cooling.

There are two ramifications of the cooling effect on refractive index. First, the refractive index will be different for different cooling rates, even though the composition is identical. Statements regarding the refractive index of a glass are only applicable for a known thermal history. Second, but less obvious, the outside of a piece of glass will cool more rapidly than the inside due to the kinetics involved with removing heat from the glass. As a result, a refractive index gradient will exist in the glass, with a lower refractive index at the surface. If the glass is ground and polished to form a lens, the lens will not perform as expected if the shape is based on the assumption that the refractive index is constant through the entire lens. While cooling-induced differences in refractive index between the surface and center of a large block of glass are usually of the order of 0.000 1, they can be ≥ 0.001 for a rapidly cooled block.

Control of the cooling rate is essential for the production of a block of glass of uniform refractive index. The cooling rates used to reduce stresses to acceptable levels in the production of nonoptical glasses are much too fast for the production of optical glasses. Use of a constant cooling rate throughout the glass transformation region will not yield a block of uniform refractive index. Ideally, the cooling rate should be decreased gradually as the temperature decreases throughout the glass transformation region. The production of a fine annealed glass requires a very slow cooling process, with a gradually decreasing cooling rate with decreasing temperature within the temperature region where permanent changes in the structure of the glass can occur.

Athermal Lenses

While the thermal expansion coefficient of many glasses is relatively low, it is not zero. As a result, the dimensions and refractive index of a glass are functions of temperature. Temperature changes in an optical system therefore result in changes in optical path length. While changes in optical path length of this magnitude may be unimportant for optical systems used within a narrow temperature range, those used over ranges exceeding 10 to 20 K must be designed to minimize thermal effects. Production of athermal optical systems, i.e., systems of constant power over a range of temperatures, requires use of glasses where the thermal changes in refractive index are exactly offset by the changes in dimensions, i.e., the optical path length is independent of temperature. This condition is met if the temperature coefficients of the index and the thickness are equal and of opposite sign.

The overall effect of changes in temperature on the optical path length are tabulated in terms of the thermo-optical constant, G, given by

$$G = \alpha(n - 1) + \mathrm{d}n/\mathrm{d}T \qquad [5]$$

where α is the linear thermal expansion coefficient, n is the refractive index, and $\mathrm{d}n/\mathrm{d}T$ is the thermal coefficient of the refractive index. Perfect compensation between changes in index and dimensions are obtained when $G = 0$. Since α is positive for most optical glasses, with values in the range of 3 to 15 ppm/K, this condition requires a negative dn/dT. Negative values of dn/dT are found primarily for glasses containing some fluorine, such as fluoride glasses and those in which oxygen is partially replaced by fluorine, such as fluorosilicates, fluoroborates, etc., or for some phosphate glasses. Silicate glasses rarely, if ever, have a negative value of dn/dT. For the same glassformer, e.g., P_2O_5, use of modifier ions with large diameters and low field strengths favor increasingly negative values of dn/dT.

The value and sign of dn/dT is determined by the competition between the decrease in refractive index induced by an increase in molar volume and the increase in the refractive index induced by an increase in polarizability of the ions in the glass with an increase in temperature. Equation [5] indicates that the absolute value of the refractive index is also important in determining the value of G. If the refractive index is high, then dn/dT must be more negative to yield $G = 0$. As a result, most athermal glasses are crown glasses. The value of dn/dT is also wavelength-dependent, generally decreasing with increasing wavelength. In particular, dn/dT becomes more positive at wavelengths near the UV edge of the glass. An athermal lens is only perfectly athermal at the design wavelength.

A few special glasses and glass-ceramics used for optical applications, e.g., vitreous silica, titanium-doped silica, and some aluminosilicate glass-ceramics, have thermal expansion coefficients which approach zero or are slightly negative. Vitreous silica has a very low refractive index and thermal expansion coefficient and can be used to produce lenses which, while not truly athermal, have relatively low temperature sensitivity. These low thermal expansion materials are often used as substrates for coated reflective optics (mirrors), where dimensional stability is far more important than any change in the refractive index of the substrate.

Production of Traditional Optical Glasses

Quality requirements for optical glasses are much greater than those for other common glasses. A constant refractive index throughout the glass requires an extremely high degree of homogeneity. The glass must be free from defects such as striae (regions of refractive index variation due to density/composition variations) and stones (undissolved material) and the bubble content must be as low as possible. The thermal history of the glass must be such that the refractive index variation, due to differing fictive temperatures, is minimized. Internal stresses must be minimized to yield a low birefringence. The glass must also be free of any species which contribute to absorption in the desired spectral region.

The quality demands of optical glasses require use of much higher quality, and thus more expensive, raw materials. Concentrations of third-row transmission metal ions must be held to rigid specifications, with maximum values usually in the ppm, or even in the ppb range, for the highest quality glasses. Defects are reduced by melting in platinum containers instead of the oxide containers used in the past. Stirring the melt, using platinum wherever possible, is essential for obtaining acceptable homogeneity. Control of the melting atmosphere may be essential for controlling the redox state of the melt. Rigorous annealing schedules must be used to produce glasses with the desired thermal history so that the problems of stress-induced birefringence and refractive index gradients are reduced to acceptable levels.

Non-Traditional Optical Glasses

While most traditional applications of optical glasses involve transmission in the visible, many modern applications require high transmission in the ultraviolet, near-infrared, or infrared regions of the spectrum. Transmission in each of these regions is determined by both the bulk composition of the glass and by the concentration of impurities present. Most nontraditional applications place even more stringent requirements on the purity of the materials used in manufacturing the glass. An extreme example of the demand for purity is found in the glasses used for telecommunication optical fibers, where concentrations of some impurities must be held to the low ppb level.

Ultraviolet-Transmitting Glasses

The ultraviolet transmission of a glass is determined by the intrinsic electronic transitions of the material, by charge transfer absorption bands due to impurities, and by surface and bulk scattering. The first of these is determined by the electronic bandgap of the material, while the other two are controlled by the care taken during production of the glass. These glasses are used in spectroscopy for windows, lenses, UV-lamps, and sample cuvets, for microlithography, and for lenses in excimer laser optical systems.

Vitreous silica is widely used as an ultraviolet transmitting glass for transmission optics at wavelengths well below 200 nm. The lowest energy absorption band for this glass occurs at 10.2 eV, or 122 nm. Many UV–VIS spectrometers, which normally are capable of measurements extending to 185 to 190 nm, use vitreous silica as optical elements. Additions of alkali or alkaline earth oxides to silica produce nonbridging oxygens in the network. Nonbridging oxygens do not bind the excitable electrons as tightly as bridging oxygens, which shifts the absorption edge to lower energies, i.e., longer wavelenths. These glasses are incapable of high transmission beyond 5 eV, or about 250 nm, and are not used for applications requiring good ultraviolet transmission.

Addition of elements which eliminate nonbridging oxygens favors good ultraviolet transmission. Alumina and/or boric oxide are frequently used in conjunction with alkali or alkaline earth oxides to counter the formation of nonbridging oxygens and preserve the ultraviolet transparency of silicate glasses. This approach can be taken to its logical conclusion by eliminating silica from the composition and producing alkali/alkaline-earth-aluminoborate glasses with good ultraviolet transparency.

In theory, fluoride glasses should provide better ultraviolet transmission than oxide glasses due to the larger bandgaps of fluorides. This assumption is true for single crystals, where fluorides such as CaF_2 are under development for transmission to higher energies than those obtainable with oxide crystals or glasses. In practice, the superiority of fluoride glasses for ultraviolet transmitting glasses has proven difficult to achieve. Vitreous BeF_2 and related glasses have superior ultraviolet transmission when compared to silicate glasses. Unfortunately, BeF_2 is very volatile and quite toxic, which restricts production of glasses based on BeF_2. Some fluoroaluminate and fluorophosphate glasses also have excellent ultraviolet transparency, but difficulties in producing these glasses in large sizes and of optical quality has limited their application.

Impurities are particularly important in production of ultraviolet transmitting glasses. The third-row transition metals, platinum, and cerium are all known to absorb strongly in the region from 200 to 300 nm when dissolved in glasses. Ferric iron has such a strong absorption band at $\approx$230 nm that concentrations of ferric ions must be kept in the low ppm range in glasses used for ultraviolet transmission. Since iron is a very common impurity in the sands used to produce silicate glasses, special care must be taken to minimize the Fe^{3+} content of ultraviolet-transmitting silicate glasses. Since ferrous iron has lower absorption in the ultraviolet, the effect of iron can be reduced by melting under reducing conditions. Platinum can enter melts from the melting unit and can be present as either ions or metallic particles. The ionic form of platinum absorbs ultraviolet light, while the metallic form scatters the same light.

Near-Infrared-Transmitting Glasses

Most glasses have good intrinsic optical transmission in the near-infrared. Absorption in this region is primarily due to impurities. Ferrous iron is especially detrimental to transmission in this region as a result of an extremely broad ligand field absorption band centered at about 1050 nm and extending into the visible and out to over 2000 nm. Reduction of absorption due to ferrous ions is obtained by careful control of raw materials to minimize the amount of iron oxide present and by melting under oxidizing conditions, which converts the ferrous ions to ferric ions. Conversion from ferrous to ferric ions increases the ultraviolet absorption of the glass. If good transmission in both the ultraviolet and near infrared regions is required, the iron content must be held to very low levels.

Although many rare earth ions also have absorption bands in this region, these elements are rarely found as impurities in sufficient levels to impair transmission in the near-infrared region for lenses and other common optical components. A more serious problem occurs for glasses used as optical fibers, where the greater optical path length magnifies the effect of impurities which may be undetectable in the spectra of samples only a few mm thick. The same can be said of absorptions of other ions which have very weak absorption bands in this region. Even higher order overtones/combination bands of infrared vibration absorption bands due to hydroxyl become important in optical fibers. While the hydroxyl absorption in infrared-transmitting glasses used in mm thicknesses is insignificant for concentrations less than 1 ppm, the hydroxyl concentration of optical fibers must be reduced to levels of a few ppb to reduce the effect of the overtone/combination bands to acceptable levels.

Infrared-Transmitting Glasses

Infrared transparency is limited by the intrinsic vibration, or multiphonon, edge of the glass and by the presence of impurities, primarily hydroxyl and carbon dioxide, which absorb in the infrared. The position of the infrared edge is determined by the force constant, which is proportional to the bond strength, and the masses of the ions. The best infrared transmission is found for materials with weak bonds and heavy ions. Infrared transmitting glasses are used in infrared spectrometers, lenses for pyrometers, infrared transparent domes for aircraft and missiles, and optical components for many systems using infrared lasers.

The fundamental absorptions of B_2O_3, SiO_2, P_2O_5, and GeO_2 lie in the region from 7.5 to 12 μm. These bands have such strong absorptions that the first overtone, at half the wavelength of the fundamental vibration, effectively determines the infrared edge for samples of more than a few hundred micrometers in thickness. The best infrared-transmitting oxide glasses do not contain significant amounts of B_2O_3, SiO_2, or P_2O_5. The lower fundamental vibrational frequency of GeO_2 allows production of some glasses of limited use for infrared transmission to about 5 μm. Comparable transmission is obtained for calcium aluminate glasses. Better transmission has been obtained for glasses combining the oxides of lead, bismuth, or cadmium with gallium oxide. These gallate glasses are termed heavy metal oxide, or HMO, glasses and transmit to ≈7 μm in mm thicknesses. The HMO glasses transmit into the visible, with a cutoff at 400 to 500 nm. Since these glasses are transparent to visible light, they offer advantages over many other infrared-transmitting glasses which are opaque in the visible.

Recently, heavy metal halide (HMH) glasses have been found to exist in a wide range of compositional systems. HMH glasses transmit much further into the infrared than the best oxide glasses, with cutoff wavelengths extending beyond 20 μm. These glasses are much more difficult to prepare than oxide glasses and are highly prone to crystallization, less durable in water, and mechanically weaker than the HMO glasses. Improvements in infrared transmission correlate with degradation in the other properties needed for most applications. Oxide impurities severely degrade the infrared transmission of these glasses. While much greater infrared cutoffs can be obtained in the laboratory, glasses currently produced commercially transmit to ≈8 μm.

Chalcogenide glasses contain either S, Se, or Te as the glassforming anion and are free of oxygen. Commercial compositions transmit to 16 μm, while laboratory samples transmitting beyond 20 μm can be made. These glasses have refractive indices in the infrared in the range of 2 to 3, very low glass transformation temperatures, which limit their temperature range of application, and are, with few exceptions, opaque in the visible. Toxicity is also a problem for Se and Te. Oxygen impurities must be kept below the 1 ppm level to avoid unacceptable absorption bands.

All infrared-transmitting glasses suffer from transmission interference in the region from 2.5 to 6 μm, due to hydroxyl in the glass. Since hydroxyl readily forms in any melt exposed to an atmosphere containing water vapor, great care must be used to reduce the hydroxyl content to acceptable levels. Glasses can be melted in sealed containers, under reactive atmospheres containing fluorine or chlorine, or under vacuum. Some glasses are prepared under dry conditions and then remelted under a vacuum to reduce the hydroxyl content. Similar care must be used to exclude CO_2 from HMH glasses, where a band at 4.25 μm has been assigned to the asymmetric stretching vibration of the dissolved molecules (Figure 2).

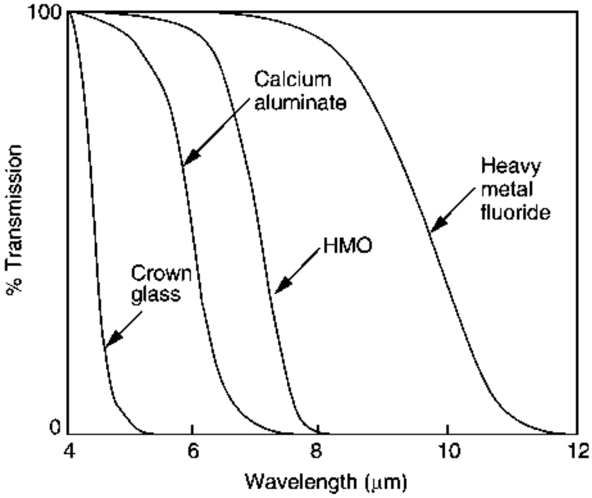

Figure 2 Infrared transmission spectra of several optical glasses showing the effect of glass composition on the infrared cutoff frequency.

Optical Isolators

Optical isolators based on the Faraday effect contain high concentrations of rare earths for paramagnetic devices and lead for diamagnetic devices. The highest concentrations of rare earth ions are obtained in aluminosilicate, aluminoborate, and aluminogermanate systems, where excellent quality glasses can be produced, while the best diamagnetic isolators are obtained using the lead bismuth gallate glasses.

See also

Optical Materials: Color Filter and Absorption Glasses. **Magneto-Optics:** Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping.

Further Reading

Deeg EW (1986) Optical glasses. In: Boyd DC and MacDowell JF (eds) *Commericial Glasses*, pp. 1–34. Columbus, OH: The American Ceramic Society, Inc.

Fanderlik I (1983) *Optical Properties of Glass*. Amsterdam: Elsevier Scientific Publishing Company.
Hoffmann HJ (1998) Differential changes in the refractive index. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 96–123. Heidelberg, Germany: Springer-Verlag.
Lentes FT (1998) Refractive index and dispersion. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 19–58. Heidelberg, Germany: Springer-Verlag.
Liepmann MJ and Neuroth N (1998) Ultraviolet-transmitting glasses. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 288–297. Heidelberg, Germany: Springer-Verlag.
Liepmann MJ and Neuroth N (1998) Infrared-transmitting glasses. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 297–308. Heidelberg, Germany: Springer-Verlag.
Pfander HG (1996) *Schott Guide to Glass*. London: Chapman & Hall.
Rawson H (1980) *Properties and Applications of Glass*. Amsterdam: Elsevier Scientific Publishing Company.
Shelby J (1997) *Introduction to Glass Science and Technology*. Cambridge, UK: Royal Society of Chemistry.
Vogel W (1991) Optical properties of oxide glasses. In: Ullmann DR and Kreidl NJ (eds) *Optical Properties of Glass*, pp. 1–36. Westerville, OH: The American Ceramic Society, Inc.

Plastics

T Bauer, Wahl Optoparts GmbH, Triptis, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Injection molded precision plastic optics in high volumes were first produced during the 1960s. After the development of sophisticated measuring and manufacturing methods in the late 1960s, precise aspheric surfaces were as easy to make as spheric contours in plastics. Today plastic is a widely used low-cost option compared to glass with even more degrees of freedom for optical and component design. Polymer optical systems are used in sensor applications, visual systems, cameras, mobile phones, video-conferencing cameras, scanners, security systems, and so on.

The physical and chemical properties of plastic materials are very different from those of glass. Generally speaking, glass materials are harder, more durable, and more stable against temperature and humidity than plastics. The variety of optical glasses comprises hundreds of different materials. Compared with this the choice for plastic materials is limited only to about 10 different materials (and even less optical parameter variations). However, plastic optics offers other design freedoms that are not achievable with glass optics.

The manufacturing technologies for glass and plastic optics are totally different. Glass lenses are made by a grinding and polishing process whereas precision plastic lenses are usually made by injection molding, compression molding, or casting. Because of the material's characteristics and the manufacturing process, plastic optics have some unique advantages.

High production numbers at low costs

Injection molding is ideal for high-volume production with low-unit costs. Moderate raw material costs and multicavity mold designs (up to 32 cavities) allow large production volumes at a reasonable unit price. In spite of considerable tooling costs, these costs compared to glass design versions, can be relatively low.

Lightweight and hardiness

For a given volume glass is much heavier than plastic (by a factor of 2.3 to 4.9). However, plastic materials are relatively shatter and impact resistant. These features are important for head-mounted systems and other weight-sensitive applications (mobile phones, cameras, etc.).

Design potentials

Injection molding makes it economical to produce sophisticated optical shapes such as aspheres, diffractive optical elements, or even freeform surface structures. From the design point of view, the more sophisticated surface shapes reduce costs or obtain better performance (Fresnel structures, lens arrays, diffractive optical elements, etc.).

Optical systems and component assembly

For typical optical system designs optical components (mirrors, lenses, prisms, etc.) must be fixed in a mounting. With plastic optics it is possible to mold mounting elements, posts, or alignment notches integrally with the optical component. This can reduce part and assembly costs considerably.

Technologies adapted to plastic materials, such as ultrasonic and laser welding, gluing and integrated snap-in structures, allow fast and cost-efficient automated and manually operated assembling solutions.

History of Transparent Plastic Materials

A Material With No Use

In 1840, the Austrian chemist Redtenbacher (1810–1870), a student of Justus Liebig, discovered during a chemical experiment a malodorous transparent liquidity, the acrylic acid. It was an unwanted byproduct which was not further analyzed at that time. Nevertheless, that discovery can be considered to be the origin of today's plastic optics.

The development of plastics began in the US in 1860 as people were searching for an alternative for ivory in the production of billiard balls.

In 1869 John Wesley Hyatt developed from cellulose nitrate and camphor the first plastic material: celluloid. At first it was used for dental plates and afterwards for shirt collars and as a film base material. Despite its flammability and its lack of light resistance, it was a first economical success.

In the following decades the first synthetical plastics were developed, for example, the material with the trade name Bakelite in 1906.

Otto Röhm's Doctorate

The basis of acrylates was rediscovered in the records of Redtenbacher in 1900. From 1901 onwards, Prof. Pechmann (1850–1902) asked his postgraduates to do more research into the acrylic acid. The acrylic acid is an acetylene compound which can be extracted from carbon, petroleum, or lime.

In 1901 Otto Röhm, a student of Pechmann, wrote his dissertation 'About polymerisation products of the acrylic acid' – the basis of the methacrylate chemistry.

Hermann Staudinger (1881–1965), head of the Institute of Chemistry in Freiburg started theoretical investigations into the structure and natural characteristics of natural and synthetical polymers in 1920. He assumed that the molecules of plastic materials consist of numerous small molecule units. Experiments as proof of the theory led to a rapid development of scientific research and to significant breakthroughs in plastic chemistry. Staudinger won the Nobel Prize for Chemistry for his work in the field of macromolecules in 1953.

In 1928 methyl methacrylate was syntheticized for the first time and patented. Acrylic resin was used as a binder in multi-layer glass.

In the 1920s and 1930s a number of plastic materials were developed, including cellulose acetate CA (photo film, synthetic fiber); polyvinyl chloride PVC (pipes, coatings, isolations); urea formaldehyde resin (dishes, electrical equipment).

Polystyrene resins were commercially produced for the first time in 1937. Polytetrafluoroethene (PTFE) was produced for the first time in 1938, and distributed as 'Teflon' since 1943. Another key development was the synthesis of Nylon, the first technical high-performance plastic material.

Plexiglas Captures the World

The industrial production of methacrylic acid methyl ester began in 1934. With its polymerization to a hard and transparent plastic material, polymethylmethacrylic (PMMA or acrylic) was developed. It was patented in 1936 and caused, under the registered trade name Plexiglas®, a sensation at the world exhibition in Paris in 1936, almost 100 years after the discovery of acrylic acid.

Plates of Plexiglas were manufactured by means of effusion and polymerization between glass plates. At first it was used for watch glass, cockpits, and the construction of street lighting and luminous advertising. People already experimented with Plexiglas injection moulding in 1935. Arthur Kingston, the founder of Combined Optical Industries Limited (COIL) had invented and patented the first plastic lens in 1933.

The First Plastic Lenses in Cameras

The first camera, incorporating plastic optical parts, to be mass produced was the 'Purma Special' introduced by RF Hunter Ltd., in 1937, using 'Perspex' (another brand name for Acrylic), for the direct-vision viewfinder.

Due to the scarcity of natural raw materials at the beginning of World War II, the plastic industry became a source of outstanding substitute materials. This development had a lasting effect far into the post-war period. Results were mainly achieved with technical plastics, such as polycarbonates, acetates, and polyamides. Other plastics were developed as a substitute for metals as well as for hardhats, for high-temperature constructions, and for products which are resistant against acids and bases.

The German chemist Karl Ziegler discovered polyethylene (Polyethen, PE) in 1953 and the Italian chemist Giulio Natta Polypropylene (PP) in 1954.

The Eastman Kodak company is credited with giving birth to the modern-day plastic Fresnel lens. In 1946 Kodak developed a tooling and manufacturing process to mass produce plastic Fresnel lenses.

The circular Fresnel lenses made at that time had spiral grooves like music records.

It took another 10 years of improvement in injection molding techniques to the first plastic lens camera – the Kodak Brownie 44A from 1959.

Between 1953 and 1959 the Polycarbonate polymer was developed by Bayer and General Electrics. Currently acrylics and polycarbonates are still the most important materials for plastic optics.

But not only thermoplastic materials were developed over this time. In May 1940 research work resulted in the monomer allyl diglycol carbonate (ADC). In the following years more than 180 different compounds of this monomer were investigated and examined.

During World War II Columbia Southern Chemical Company (a subsidiary of Pittsburgh Plate Glass) began research into ways of obtaining nonthermoplastic materials. The project was given the name 'Columbia Resins'. The 39th tested (CR 39) turned out to be the most significant due to its unique properties and was used since 1947.

At first, solely the low-refractive CR 39 was available for organic eyeglasses. Only in the mid-1980s did the development begin in the field of high-refractive nonthermoplastic materials.

Properties of Plastics for Optics

As already mentioned, the choice for plastic optical material is limited only to about 10 different types of material. Optical properties (abbe value, refractive index, transparency) as well as mechanical, thermal, and humidity boundary conditions are decisive for the material choice. Different to glass during plastic processing (injection molding, casting, etc.) the process affects not only the geometry but also inner properties like refractive index, transparency, and birefringence.

Optical Properties

Fundamental optical properties are defined by optical transmission, refractive index, and dispersion. For plastic optics birefringence is an important parameter too.

Although the total number of plastic materials has increased in recent years, the range of refractive index and dispersion characteristic is limited to almost two major groups – crown-like materials such as Acrylic (PMMA), Polyolefin (COC, COP) and flint-like materials such as Polystyrene (PS), Polycarbonate (PC) and Styrol-Acrylnitril-Copolymer (SAN).

This strongly limited variety of plastic optical materials significantly restricts the optical design freedom (Figures 1 and 2).

Physical Properties

Important physical properties are weight, impact, and abrasion resistance – and thermal properties like temperature resistance and thermal expansion. Already, during system design, both mechanical and thermal properties have to be taken into consideration. The thermal expansion of optical plastics is approximately ten times higher than that of glass materials. In an optical system this effect has to be compensated for by optical design or mounting.

Typically most optical plastics can withstand temperatures up to 90°C. The maximum service temperature of Polycarbonate and Poly-Olefin materials reaches 120°C.

The specific gravity of plastic optical materials ranges from 1 to 1.3.

Polycarbonate has the highest impact resistance of all optical plastics and so is used for windshields and crash helmets. Acrylic has the best abrasion resistance (Table 1).

Common Materials

Thermoplastics

Thermoplastics are polymers with a linear molecule structure – with or without side chains – which can be deformed reversibly without any modification of their thermoplastic characterization by means of the impact of heat. Optical elements can be manufactured by injection molding or hot embossing.

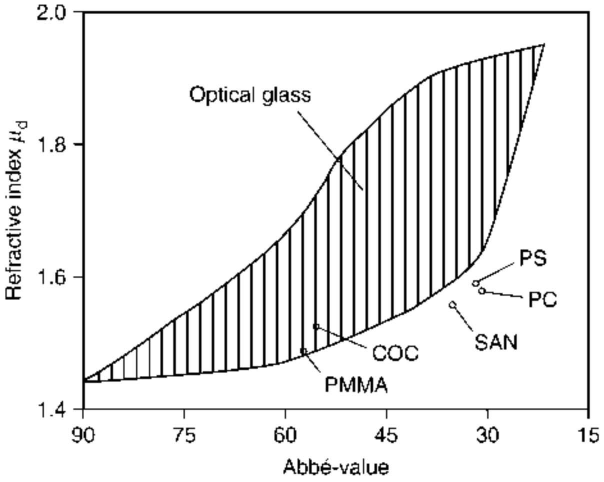

Figure 1 Abbe diagram.

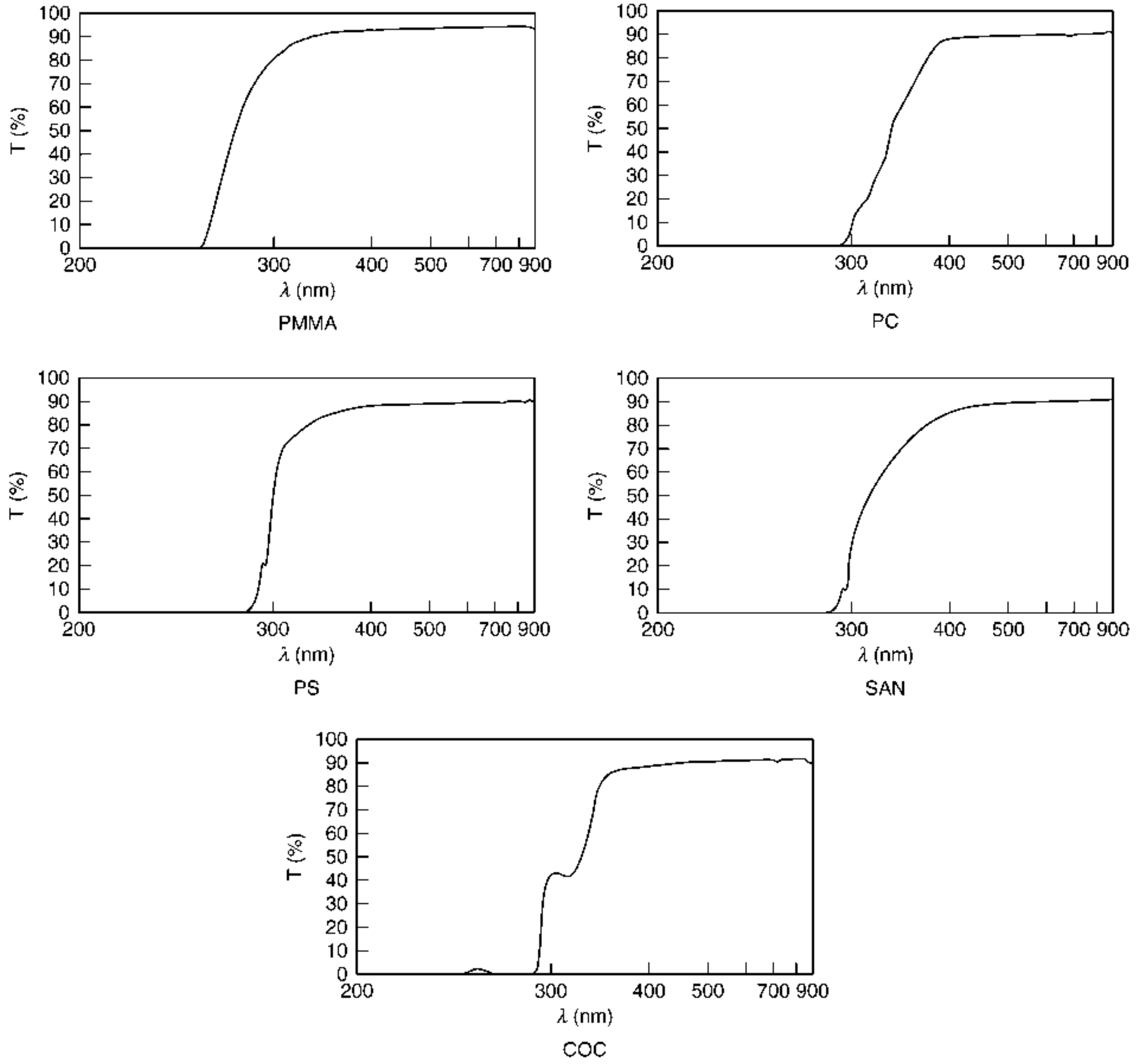

Figure 2 Transmission characteristics.

Typical optical thermoplastic polymers are listed below.

Acrylic or PMMA (Polymethyl methacrylate)

Acrylic is the most commonly used optical plastic. It is moderately priced, easy to mold, scratch resistant, and not very water absorptive. Its transparency is greater than that of most optical glasses throughout the visible range of the spectrum (organic glass). PMMA can replace glass in every application as long as the service temperature is less than 90°C and a low chemical resistance is required. Additives to acrylic (as well as to several other plastics) considerably improve its ultraviolet transmittance and stability.

PC (Polycarbonate)

Polycarbonates are linear polyesters of the carbonic acid which combine many characteristics of metal, glass, and plastics.

Polycarbonate is very similar to styrene in terms of optical properties such as transmission, refractive index, and dispersion. Polycarbonate, however, has a much broader operating temperature band of up to more than 120°C. It is used as the flint material for systems that have to withstand severe thermal conditions.

Another advantage is the high-impact resistance of polycarbonate. Safety glasses and systems requiring durability often consist of polycarbonate. Because of its high ductility polycarbonate is not easily machined.

Table 1 Material properties

Material	*Characteristics*	*Acrylic (PMMA)*	*Polystyrene (PS)*	*Polycarbonate (PC)*	*Styrol-Acrylnitril (SAN)*	*Cyclic Olefin Polymer (COP)*	*Cyclic Olefin Copolymer (COC)*	*Polyether-sulfone (PES)*	*Acrylnitril-Butadien-Styrolco-polymere (ABS)*	*CR39*	*Optical glass (B7)*
Optical	Spectral Passing Band (nm)	390–1600	400–1600	360–1600	395–1600	300–1600					
	Refractive index at 587 nm and 20 °C	1.4918	1.5905	1.5855	1.5674	1.5261		1.6600	1.5380	1.501	1.517
	Abbe value $(nD - 1)/(nF - nC)$	57.2 (53)	30.8	34 (29.9)	34.8	55.8	58	19.4		58	64.4
	Transmittance (%) thickness 3.2 mm	92	87–92	85–91	88	92	92	80	85		
	Haze% thickness 3.2 mm	1.3	2–3	1.7	1.5	1.5	1.5	1.5			
Physical	Specific Gravity (g/cm)	1.18	1.06	1.25	1.07	1.01	1.03	1.37	1.05	1.32	2.53
	Max. Service Temperature (°C)	92	82	124	95	123	130	200	90	130	400
	Linear Expansion Coefficient $(1/K)$	6.8×10^{-5}	8.0×10^{-5}	6.6×10^{-5}	7.0×10^{-5}	7.0×10^{-5}	7.0×10^{-5}	5.5×10^{-5}	8.5×10^{-5}		1.1×10^{-4}
	Abrasion Resistance (1–10)	10	4	2		6	6				
Environmental	$\mathrm{d}n/\mathrm{d}T\ (\times 10^{-5})$	−8.5	−12	−11.8…−14.3		−8	−10.1				
	Sensitivity to Humidity	high	low	low	medium	low		high	medium		
	Water Absorbtion (weight %) 23 °C, ISO 61	0.30	0.20	0.15	0.30	<0.01	<0.01	0.70	0.45		0
Manufacturability	Processability	excellent	good	poor	excellent	good					
	Birefrigence	low	high	high		low					
Chemical	Resitance to Alcohol	limited	good	limited				good			
Costs	approx. Material costs (EUR/kg)	3.3	2.5	4.4	4.4	27.1		21.0	3.5		25.0

PS (Polystyrene)

Polystyrene or Styrene is a low-cost material with excellent molding properties. Styrene has a higher index and a lower numerical dispersion value than other plastics. It is often used as the flint element in color-corrected plastic optical systems.

Compared with acrylic, styrene has lower transmission in the UV range of the spectrum. It has a lower resistance to UV than acrylic but is more easily scratched than acrylic.

Because its surface is less durable, styrene is more typically used in nonexposed areas of a lens system. A styrene lens, when paired with an appropriate acrylic lens, offers an effective achromatic solution.

SAN (Styrol-Acrylnitril-Copolymer)

SAN is a copolymer of certain fractions of Polystyrene and Acrylic (typically 70/30), which allows adjusting refractive indexes. It is a glassy polymer with a low tendency to stress cracks. SAN has slightly more chemical durability than PS and is a low-cost material with excellent molding properties.

ABS/MABS (Acryl-Butadien-Styrol-Copolymer)

ABS is developed from polymerization of styrene and acrylnitrit on polybutadien during the emulsion process. Process variations lead to different material characteristics. Compared to PS, the chemical resistance and temperature stability of ABS is considerably better. ABS has a natural color so it cannot be used for every optical application.

PES (Polyethersulfon)

PES shows a remarkable temperature stability, a low transmission against UV-light, a limited transmission with a natural yellow color, and cannot be used for every optical application. Because of its excellent temperature stability, PES is suitable for assembling, even on electronic motherboards by reflow soldering. This process causes temperature stress up to 220°C (Figure 3).

COP and COC (Cyclic olefin polymer and copolymer)

Cyclic olefin (co-)polymer provides a high-temperature alternative to acrylic. Its refractive index and transmittance are similar but the heat distortion temperature is about 30°C higher than for acrylic. It has a remarkably low water absorption value of less than 0.01% (compared with Polycarbonate 0.2% and PMMA 0.3%).

Cyclic Olefin Polymer (COP) is a family of amorphous plastic resins with low native stress birefringence properties. Zeonex is a resin developed by the Nippon Zeon Company – one grade, E48R, has a heat distortion temperature of about 122°C. Zeon Chemicals has another class of polymers called Zeonor, that can be provided as a lower-cost alternative to Zeonex.

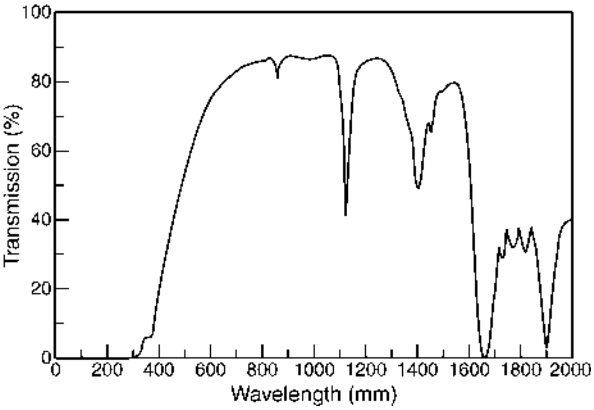

Figure 3 PES.

Cyclic Olefin Copolymer (COC), with the trade name of Topas, is another high-temperature alternative to acrylic. Generally, it is a lower-cost alternative to COP. Particularly favorable is the high flowability (good injection moldability) of these thermoplastics.

PMP (Polymethylpenten – TPX®)

PMP has a structure similar to PE, except that only the methyl groups are replaced with isobutyl groups. The chemical resistance can be compared to PP, although it tends to stress crack under the impact of ketones or chlorinated solvents. The main advantages of PMP are its excellent transparency and its good mechanical characteristics, even at high service temperatures of up to 150°C.

PE-HD (high density polyethylene)

If polymerization is controlled catalytically, molecules with a low number of side chains will be received. Compared to PE-LD (low density), the molecules are quite compact with a higher stability and chemical resistance and service temperature can be up to 105°C. PE shows a bad transparency (milky appearance) in the visible spectral range. However, this material is permeable for infrared beams and so is frequently used for motion detectors.

Elastomere

Elastomeres are plastics with loose and networked molecules which are rubber elastic at a normal temperature. The most popular elastomeres are natural rubber and silicone rubber. These materials are optically used as flexible light guides.

Duromeres

Organic glass is a fully synthetic plastic material available in a vitreous state. It consists of macromolecular organic compounds which do not follow any principle of periodic arrangement and are hence amorphous.

Duromeres are plastics with spacial closely networked molecules which are very hard and refractory at a normal temperature. Impact of heat causes an irreversible induration. In most cases, duromers are used to produce plastic lenses made of organic glass. Once they have been thermally treated after production, their shape can no longer be changed.

The well-known plastic CR 39 is one of the organic materials used for plastic lenses.

CR39 plastic is made by polymerization of the allyl diglycol carbonate (ADC). It is a transparent, thermosetting resin which combines an exceptional range of qualities which are not available in other plastic transparent materials. The CR39 plastic is colorless and completely transparent to the visible light and almost completely opaque in the infrared and ultraviolet region of the spectrum. For these reasons, it is largely used for the production of sunglass lenses. They can be colored by surface dyeing or bulk tinting, have high abrasion resistance, and high-quality optical properties. Their weight is about half as much as glass, they keep their excellent optical properties despite long-term exposure to chemicals and resist heat distortion up to 100°C. These qualities, combined with the exceptional optical characteristics, make CR39 the best choice for applications where severe conditions of use exclude all other optical materials.

Additives and Colors

The characteristics of many plastics, particularly of thermoplastic materials, can be changed by adding so-called additives.

UV-filter materials are frequently added to improve the stability of materials in the sunlight. Sometimes nonstabilized plastics tend to yellow (transmittance decreases). Color filters can also be generated by coloring. Infrared coloring permits the production of optical elements which appear black in the visible spectral range. These materials are particularly used for sensors, optical scanners, or remote control front windows. Even characteristics, such as temperature stability, can be influenced by additives.

Manufacturing Methods

The manufacturing processes for glass and plastic optics are totally different. Glass lenses are made by a grinding and polishing process. By contrast, typical manufacturing methods for precise technical plastic optical parts are diamond turning of plastic blocks, injection, and compression molding of granulates. Other methods, such as casting, have their application in special fields like ophthalmics (eyeglasses). In any case, processing transparent plastic materials for optical purposes should take place in a clean environment.

Diamond Point Turning

Diamond point turning is an ultra-precision machining process carried out on special high-performance, ultra-precision diamond CNC-machining systems.

In combination with vibration isolation systems, digital signal processor-based machine control and integrated measuring systems with nanometer programming resolution manufacturing of optical surfaces is possible. This method can be expanded to full 3D-milling systems. With such an arrangement the generation of free-form surface profiles is possible.

Because of the long production time and the high machine costs, this technology is used for prototyping in plastics and nonferrous metals, mold inserts, and series production plastic parts – which cannot be manufactured by injection molding (because of size or precision demands).

Injection Molding

Injection molding produces several parts per shot (molding cycle) in single or multicavity tools. This production method is used for most of the plastic optic parts. Other fabrication techniques are used only when molding is inappropriate.

A plastic injection-molding machine consists of a fixed platen, a moving platen, a clamping unit, and an injection unit. Molding of optical parts requires special machine configurations and auxiliary equipment.

Compression Molding

Compression molding is used in the making of Fresnel lenses or other micro-optical structures. The material is pressed between heated platens with accurately defined temperature cycling during pressing. Mold inserts are formed by electroplated copies, replicated from master structures. These masters can be, for instance, diamond turned structures or diffraction gratings made by holographic methods.

The compression molding process allows realizing small structures with high aspect ratios and tight angular and positioning tolerances.

Manufacturing Process

Optical and System Design

On the one hand, optic design for plastic optics in principle uses the same mathematical algorithms as optic design for glass. On the other hand, designing plastic optics requires a profound understanding of the material properties and the manufacturing processes. Knowledge of production technologies, material characteristics, and assembling methods, together with design expertise, are needed to fully exhaust what precision plastic optics can be. Simply substituting the indices of refraction and re-optimizing, the design will not succeed. Expert design assistance is essential at this stage.

Designing plastic optics with modern design tools allows greater design freedom. The advantage of combining integral mounting structures with the optical surfaces to create mounting flanges, alignment, and snap features, provides the ability to automated assembly. Aspheric, cylindric, or toroidal surfaces are as easy to realize as spherical ones. Microstructures such as diffractive optical elements can be integrated too.

Prototyping

After designing a plastic optical system, lens prototypes can be made by diamond point turning in various plastic materials. The best surface finishes can be obtained with PMMA. Materials such as polycarbonate, pleximid, or Zeonex do not yield smooth surface finishes.

A major problem is the availability of semi-finished plastic blocks in various materials. For PMMA, a wide variety of bars or plates are available. For other materials or colored options, semi-finished items have to be made by injection molding. Because of the high manufacturing costs, diamond point turning is only recommended for making a limited number of prototypes to verify functionality of the optical design and to perform first tests.

In this stage optical systems are normally assembled from single elements. Housing parts often are made from aluminum. The resulting surface quality and system performance cannot be a validation of the manufacturability by injection molding. To get reliable knowledge about this, making a molded prototype from a single-cavity prototype mold is recommended.

The Injection Mold

A high-quality injection mold is obviously essential for precise plastic optic parts. The parts can never be better than the tool – but good tooling, however, does not guarantee good parts. A strong understanding of the whole manufacturing process is the key to producing precision plastic components.

Any injection mold consists of three main parts. The upper half, which is affixed to the injection-side platen, the lower mold half, which is affixed to the ejector-side platen, and the mold-ejection mechanism. Guide pins and taper locks ensure proper alignment of the mold halves.

Production volume and precision required from a tool influences the selection of mold materials and built-in maintenance features.

Single-cavity and multi-cavity molds from two to eight, 16, or even 32 cavitys are used for plastic optic parts.

Thermoplastic materials shrink during cooling in the mold. This geometric effect has to be compensated for in the injection mold. Exact shrink rates can be calculated and the tool can be modified.

For manufacturing optical plastic parts by injection molding, the optical surfaces (plano or aspherical shapes, diffractive, conical, lenticular, and cylindrical surfaces) are generated as separate inserts in the tool.

Aspheric inserts are manufactured by two steps. First a best-fit curve is generated on a stainless steel substrate. The substrate is then subjected to a nickel-plating process (electroless nickel) that deposits a thin layer of nickel (up to 500 μm). In the second step, single-point diamond turning produces the final aspheric or diffractive curve in the nickel. Because the hardness of a nickel-plated insert is less than that of a steel spherical insert, it will be more susceptible to scratch defects (Figure 4).

Pre-Production

A pre-production stage is recommended to check the manufacturing process. Typically this is done by using a single-cavity prototype mold. This mold can be used to find optimal molding conditions.

Optical and mechanical design can be verified with real molded components and design revision is possible to affordable conditions. Often the prototype molds is used to start limited production since production tooling may take much more time.

Series Production

For series production of high volumes, multicavity production molds are required. Depending on quality,

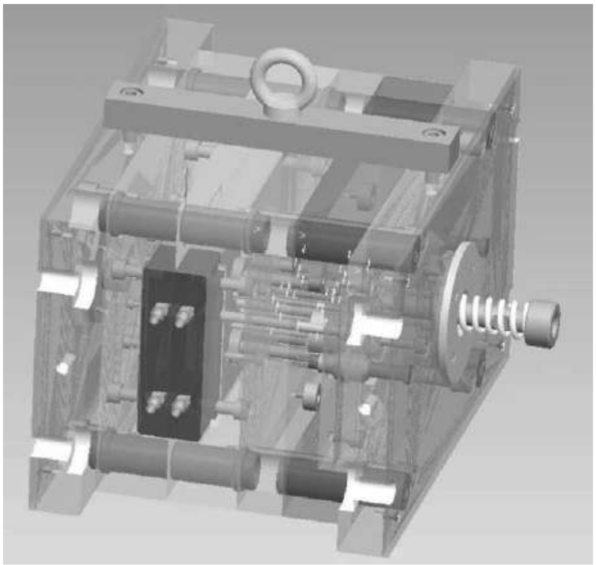

Figure 4 Injection mold.

volume, throughput, and cost, the production tool may have 2, 4, 8 or even up to 32 cavities. The production molds function for at least several hundred of thousands of injection cycles.

Coating and Component Assembly

Coating

Because of their limited temperature and UV resistance, plastic lenses must not be coated in an elevated temperature and radiation environment. During the deposition of thin films onto plastic, the coating chamber temperature is significantly lower than that for glass optics. This requires deposition techniques such as ion-assisted deposition to apply antireflective, conductive, mirror, and beamsplitter coatings.

Today, multi-layer dielectric coatings are routinely deposited on plastic components. Typical broadband antireflection coatings reduce reflection to about 0.5% per surface across the entire visible spectrum. Narrowband, multilayer antireflection coatings can achieve surface reflectances less than 0.2%. Multilayer dielectric coatings can be modified easily to scratch-resistant designs for front lenses and windows.

Several front and back surface reflector coatings are available for plastic substrates. Standard coating metals include aluminum, silver, and gold. Aluminum coatings provide surface reflectances greater than 88% across the visible spectrum and gold coatings greater than 95% for the near-infrared region.

Component Assembly

Many advantages of using plastic optics will come out in manufacturing components. The integration of optical, mechanical, and electronic elements allows building low-cost polymer optical components such as small camera lenses or scanner heads.

Plastic optical assemblies are usually carried out in a clean room environment by manual, semi-, or fully automated processes. The components have to be designed to ensure ease of assembly. Snap-on features, UV-cementing, heat-staking, and ultrasonic and laser welding, can also be employed with plastic materials.

Since most optical tolerances are additive, it is essential to establish quality check points in the manufacturing process to sort out nonconforming sub-assemblies.

Automatic in-line optical performance, monitoring such as MTF testing, can be implemented. SPC techniques should be used here to ensure the process is not drifting out of the controllable range.

Summary

For centuries, glass proved to be the material of choice for all optical applications. During the twentieth century, applications for glass elements and complex optical systems greatly expanded. Early uses of plastic molded elements included toy objectives, gauge windows, and watch crystals. Throughout the 1970s and up to the present day, high-grade polymers were developed specifically for optical applications. These advancements in materials, coupled with improved mold design have enabled plastic optics to replace glass optics in a wide and growing number of applications.

Plastic optic technology permanently expands their traditional limits and fields of application. On the one hand, plastic optic is still limited by material properties but, on the other hand, design and assembling freedoms allow new approaches.

See also

Diffraction: Fresnel Diffraction. **Geometrical Optics:** Lenses and Mirrors. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Color Filter and Absorption Glasses.

Further Reading

Corning Precision Lens Inc. (2000) *The Handbook of Plastic Optics*, 2nd edn. Cincinatti, OH: Corning Precision Lens Inc.

OPTICAL MATERIALS

Sculptured Thin Films

K Robbie, Queen's University, Kingston, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical materials, like all materials, ultimately derive their physical properties from their chemical composition and physical form. Aluminum behaves differently from copper because of a chemical difference, while graphite behaves differently from diamond because of a crystallographic difference. For materials that are not infinite perfect crystals, an important contribution to their electromagnetic (optical) response comes from the morphological form of the material. Morphological form describes the physical structure of the material: a perfect crystal is one type of morphological structure, a porous fibrous bundle is another. An example of a morphological difference is that between snow and ice. Snow and ice are chemically identical, composed of solid water, but they exhibit different optical properties because of a difference in morphology. The polished glasses and large perfect crystals that comprise most optical components have morphologies more like ice than snow; their morphologies are dense and crystalline, with few and small disordered 'defects'. Unlike ice, snow is rich with morphological structure (see Figure 1) arising from the complex growth of each flake in the clouds, and the aggregate growth of snow on the ground. The morphology of snow scatters all colors of visible light, giving it an opaque white appearance, while ice is transparent and nearly colorless.

Sculptured thin films (STFs) are materials with controlled and designed morphology that exhibit a rich behavior unlike that of the constituent materials in bulk; they are the snow to most optical materials' ice. Unlike snow, sculptured thin film morphology is designed to scatter light in a controlled manner to produce a desired optical response. A sculptured thin film could be a coating applied to an optical element where the coating converts polarizations of light, circular or linear, but is composed entirely of an optically isotropic material (which doesn't usually affect the polarization of light). The film morphology is designed to create the desired optical properties. Sculptured thin films are also useful for magnetics, where an example of a sculptured thin film is the metal film coated onto plastic tape for video, audio, and data recording. The film is deposited under conditions that impose an elongated shape onto the magnetic domains, altering the magnetic properties. The morphological form of a sculptured thin film defines the way it responds in electromagnetic, chemical, and biological interaction, and there is hope that these unique materials will someday form the core of technologies that make the world a better place.

Structure in Thin Films

As chemical composition and morphology determine the electromagnetic response of materials, it would be wonderful to have a technique where we could build materials atom by atom, placing each exactly where we wished, and varying the element at will. With this level of control it would be possible to design and fabricate materials for almost any conceivable optical application. In some sense this is what is done now to fabricate lasers and optical switches. Microelectronic and micromachining processes have produced fantastic optical devices by controllably mixing materials (mostly semiconductors) and patterning designed structures. These techniques typically employ a repeated series of thin-film depositions and photolithographic patterning. Examples of thin-film coatings include antireflection coatings on eyeglasses, hard coatings on grocery store barcode scanners,

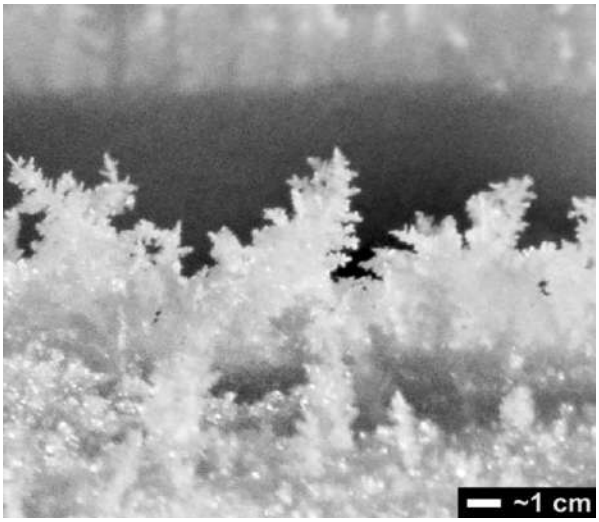

Figure 1 Photograph of snow accumulating at −20 °C. Photo courtesy of Chelsea Elliott.

ultraviolet blocking aluminum inside potato chip bags, and all integrated electronic circuits. Thin films are usually produced by vapor deposition in vacuum, and can be made with most elements and inorganic compounds (and a growing list of organics). Most thin films grow as uniform, dense, two-dimensional (2D) layers, and three-dimensional structure (3D) is produced by removing portions of the film with photolithography and etching. Photolithography is a process where ultraviolet light is projected through a mask and microscope system onto a photosensitive polymer layer atop the thin film. The light hardens areas of the polymer in the pattern of the mask, preventing them from dissolving in a chemical developer solution. Exposed portions of the thin film can then be etched away with chemical or other means. Repeating as desired, many steps can create complex 3D designed structures. Computer chips and other integrated circuits are semiconductor crystals covered with a layered structure of conductors (aluminum, copper, platinum) and insulators (metal and semiconductor oxides) patterned with thin film deposition and lithography. Semiconductor lasers and many other optical devices are made this way as well, usually employing compound semiconductors (GaAs, AlGaAs, GaAsP, InP, etc.). The micrometer-scale patterning made possible with thin-film deposition and etching techniques has produced a wealth of semiconductor optical devices.

Optical lithography, however, is limited in size scale to dimensions comparable to the wavelength of light used to expose the photomask. Modern UV lithography can pattern 100–200 nm structures. Visible light has wavelengths from 400 nm (violet) to 700 nm (red). Shrinking the dimensions of physical structures is a great benefit for optical devices as it allows light to be controlled coherently without strong random scattering. Structures with dimensions comparable to the wavelengths of light strongly scatter light waves; scratches on optical components or imperfections in materials with dimensions comparable to the wavelength are considered defects, and are usually deleterious. However, materials with morphology much smaller than the wavelength (nanometer structure, or nanostructure) are seen by propagating light as homogeneous, with polarization and dispersion effects characteristic of the underlying structure. Sculptured thin films are materials where complex nanostructure has been created by controlling growth parameters dynamically during deposition of a thin film. The simplest sculptured thin film might be one where the temperature is periodically varied during thin-film deposition (although this is a very simple example, lacking the anisotropy central to the sculptured thin-film concept). Variation in the growth process with temperature affects film density, producing a porosity variation (periodic inhomogeneity) as a function of film thickness. Varying porosity produces varying refractive index, and this film would be an optical interference filter. The largest limitation of sculptured thin films is that all attainable structures must be characteristic of the growth process. As all morphological control is accomplished by varying film growth parameters, structure that is not characteristic of thin film growth is not possible. Fortunately, thin film deposition and growth is sufficiently complex to allow the fabrication of a large range of useful structures. As with lithographic patterning, the structural control is two dimensional, but now the third dimension is constructed by continuing the deposition, not by repeating the deposition and lithography cycle. Complex structures can be produced on the scale of tens or hundreds of atoms (approximately tens of nanometers) with controlled 3D morphology. This reduction in size scale compared to lithographic patterning is a huge benefit. Materials can now be designed and created with subwavelength structure to coherently scatter light. Because the morphological structure (<100 nm) is much smaller than the wavelength of visible light (~500 nm), light traveling in a sculptured thin film sees a medium with averaged properties. The material can be designed to control refractive index through a wide range, and birefringence and circular polarization effects are controllable.

The morphological structure of thin films is a vast topic. Thin films are produced with a nonequilibrium growth process, somewhat like the growth of clouds, and the mechanics of the growth determine the structure, which is often very different from that seen in a bulk sample of the same substance. Crystal and morphological structure vary with deposition process conditions, and vary during deposition to record a history of growth much like the rings of a tree, or stratified rock. If chemical composition is varied during the deposition, that change will be reflected in layers in the film. If vapor energy or temperature is varied, variations in density and crystal structure will result. Thin film growth is an aggregation process (the film accumulates over time), and exhibits many of the features of similar processes, sometimes producing materials with structures similar to broccoli or clouds. This self-affine (self-similar) structure is observed over a large range of sizes under some film deposition conditions. Under other deposition conditions, crystallographic effects dominate growth and films become layers of faceted crystals. In sculptured thin films, these growth characteristics are controlled and exploited in a functional manner to design material properties. For example, magnetic video and audio tapes that have the symbol ME, abbreviating 'metal evaporated', employ a structured magnetic thin film that has superior magnetic response for recording. Similar effects are observed in optical materials, and sculptured thin films appear promising for improving existing optical devices or enabling new ones to be developed.

The most studied form of sculptured thin films is that produced by an asymmetric deposition process. In deposition by vacuum evaporation, a thin film is fabricated by evaporating some source material in vacuum, and condensing a vapor onto a substrate. Eyeglasses, mirrors, television phosphor screens, etc. are coated this way. In most thin films the vapor arrives straight down (perpendicular) to the plane of a flat substrate. The film structure reflects the symmetry of the vapor, and is typically composed of a dense layer with some defects or voids oriented perpendicular to the substrate surface. A thin film exhibiting this morphology is shown in the scanning electron microscope image of Figure 2. When vapor arrives asymmetrically onto a substrate, however, that asymmetry is reflected in the growth, producing structural asymmetry in the resulting film, and altered physical properties. A second scanning electron microscope image is shown in Figure 3, of a sculptured thin film where the orientation of the substrate relative to the vapor is changed twice, producing three layers with different morphologies. The thin uniform gray at the absolute bottom of the image is the edge of the silicon crystal substrate which has been cleaved to reveal an interior section of the film. The bottom and first film layer is inclined to the right, toward the vapor source for the duration of this first part of the deposition. The middle and second film layer has a twisted structure created by slowly rotating the substrate relative to the vapor, while keeping the tilt fixed. The top layer has a vertical structure created by rapidly rotating the substrate. The entire film was produced with the vapor arriving at an angle of 45°, measured from either the plane of the substrate or its perpendicular. Varying the tilt angle has a profound effect on the morphology of thin films. The void structure in these films records the geometry of the deposition. When we cleave a film we are able to observe a 2D slice of this rich morphology. At large tilt angles, glancing angle deposition (GLAD) produces films with highly porous and complex structure. At these angles, the film structure becomes isolated columns of material where growth variations are now reflected in variations in the shape of individual columns. At less oblique vapor angles, growth variations are reflected in the shape of networks of atomic-scale voids within a largely dense film. Indeed, for the best optical devices fabricated to date (with the serial bideposition technique described later) it is difficult to distinguish any variation in structure even using a high-resolution scanning electron microscope. The film deposition process can be designed to control the morphology of the growing film, and to vary the structure as a function of time and thickness. The morphology of these films determines their optical properties; optical birefringence, second harmonic generation, filter bandwidth and attenuation, etc. can all be controlled in sculptured thin films.

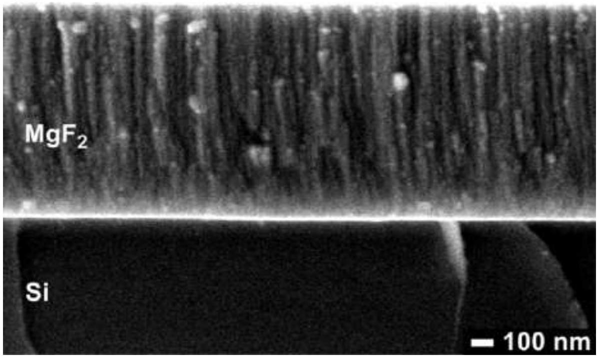

Figure 2 Micrograph of MgF_2 film deposited onto a 75 mm silicon wafer substrate, then fractured to reveal the edge shown.

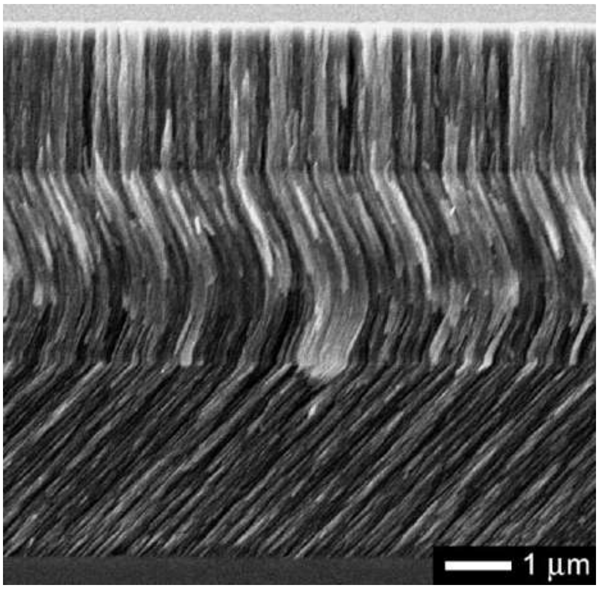

Figure 3 Micrograph of SiO film deposited with a tilt angle of 45° with one layer of no rotation, one layer with one rotation, then one layer with rapid rotation.

Optical Thin Films

Light can be described as a traveling wave with electric and magnetic fields oscillating perpendicular to each other, and to the direction of propagation. If the orientation of the electric field remains constant as the wave propagates, the light is said to be linearly polarized. If the orientation of the electric field rotates through a circular path, the light is circularly polarized, and can be left or right handed. In the most general description, the electric field traces an ellipse during propagation and the light is said to be elliptically polarized. The experimental study of light polarization is called ellipsometry. When light propagates through vacuum or air, the state of polarization remains constant. When light propagates through some types of materials, however, crystallographic or morphological asymmetry (or more correctly, anisotropy) results in different propagation conditions for different polarizations. Upon passage through these materials, the orientation and form of the state of polarization is changed, for example from left circularly polarized (LCP) to right circularly polarized (RCP). When multiple layers of anisotropic materials are combined into an optical circuit, as in the sculptured thin film concept, highly controlled optical response characteristics can be created.

The discussion of the optical properties of thin films begins with definitions of isotropic, uniaxial, and biaxial materials. Isotropic materials have optical properties that are independent of the direction that the light propagates through them. Vacuum and air are isotropic, as is glass. Rotation of the optical material relative to the light propagation direction has no effect on how the light propagates, and the state of polarization remains constant. Uniaxial materials have properties that vary with direction, but with one preferred direction (or axis) where response is independent of rotation. Biaxial materials have properties that are even more dependent on propagation direction, and require two axes to describe their optical properties. For biaxial materials there is no propagation direction where the optical response is independent of rotation. The mineral calcite is an example of a uniaxial material where the refractive index, and therefore the speed of light propagation, varies with direction relative to the crystal structure. This is a direct consequence of the trigonal crystal structure of calcite, and is an example of linear birefringence. Birefringence is the effect caused by anisotropic materials whereby light of different polarizations, even within one ray of light, travel with different speeds. When light propagates in a birefringent material, the orientation of the light's electric field relative to the material's optical axis determines the speed of propagation. Light with different polarization (orientation) propagates at different speeds, and a beam's polarization state is altered on passage through the material. Thin films can be isotropic, uniaxial, or biaxial, depending on the conditions under which the film is deposited. Sculptured thin films are films fabricated in a manner that controllably enhances the anisotropic optical response to produce materials with designed optical properties based on control of their morphological structure. As in calcite where the anisotropic crystal structure yields a uniaxial optical response, anisotropic morphological structure in sculptured thin films yields biaxial optical response. A plan view scanning electron microscope image of an anisotropic thin film is shown in Figure 4. The film was deposited with two vapor sources with the substrate placed between and the vapor arriving from both sides at an angle of 70° from the substrate perpendicular. This film is similar to those produced with the serial bideposition technique, and is optically birefringent. Polarization filters and other optical components can be constructed by designing the morphological structure of the thin film.

As discussed above, when a thin film is deposited from a vapor arriving perpendicular to a flat substrate surface, a structure composed of small (tens of nanometers diameter) columns of film material grows, with the columns oriented perpendicular to the substrate (Figure 2). This film is optically uniaxial, with the preferred axis perpendicular to the substrate, and parallel to the vapor flux. Light propagating along this optical axis (perpendicular to the substrate) travels at one speed regardless of polarization, and independent of rotations of the substrate around this axis. Light propagating along other paths will travel at a speed that depends on the polarization of the light

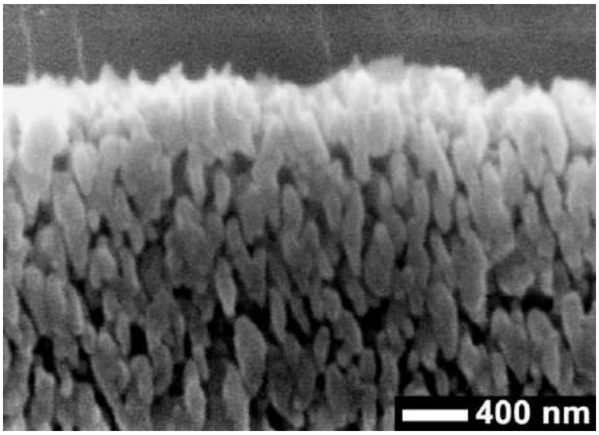

Figure 4 Micrograph of MgF_2 film, viewed from above, deposited with simultaneous growth from two vapor sources.

(orientation of the electric field relative to the propagation direction). The effect is small, and arises from the morphological structure of the film. More pronounced optical response is produced when the vapor arrives onto the substrate at an angle from the substrate perpendicular, producing optically biaxial materials, and enabling exciting polarization and filtering effects. If a thin film of magnesium fluoride is deposited onto a polished piece of glass with the vapor arriving at 45° (relative to an axis perpendicular to the substrate), a slanted columnar film is produced. The morphology of the film is tightly packed columns of polycrystalline magnesium fluoride with diameters between 50 and 200 nanometers (nm), inclined toward the vapor source, and penetrated by a network of atomic-scale voids. Because of the anisotropic nature of the deposition, the voids are anisotropic, and the film is optically biaxial. If the substrate is then held at that tilt angle and is rotated during deposition, the anisotropy in the structure rotates with the deposition, and an important type of sculptured thin film is produced. The electromagnetic description of this type of film is a 'helicoidal bianisotropic medium', describing both the helix shape of the rotation of the structure and the local bianisotropic (biaxial) physical, and optical, structure. Films with helicoidal symmetry are ideally suited for studies involving circular polarizations of light, and many demonstrations of sculptured thin films have involved these films and circularly polarized light. Like many organic molecules and some mineral crystals, these film are chiral, existing in two mirror image forms, called enantiomers. A right-handed structure rotates clockwise traveling away from an observer; a left-handed structure rotates counterclockwise. Also like some organic molecules and mineral crystals, helicoidal films are optically active – they affect the polarization of light. The observed effect depends on the wavelength of light relative to the physical helicoidal structure. If these two lengths match, a strong resonance effect is observed where the light wave coherently scatters from the anisotropic structure. Filters to produce and convert circular polarized light have been constructed with sculptured thin films of this type.

The range of structural control over the morphology of thin optical films limits the ability to tailor optical response for desired applications. While optical systems must always be treated as a whole to determine response properties, the description of some basic elements is useful. A film deposited at normal incidence will be uniaxial with the axis perpendicular to the substrate. For light propagation along this axis (common for optical systems), the material exhibits no birefringence. This structure is often called an isotropic element even though the columnar structure is slightly optically anisotropic. A truly isotropic layer could be produced by ion bombardment or elevated temperature during deposition, either of which would have the effect of eliminating the columnar structure and reducing anisotropy in the film. The second elemental structure is a slanted columnar film that is commonly described as a matchstick morphology. It is produced by changing the vapor incidence away from the substrate normal, leading to an inclination of the column structure toward the vapor source, as described above. Further structures can include: helicoidal structure where the anisotropy rotates through the thickness of the film, and periodically bent nematic structure where the anisotropy tilts in an s-shaped pattern through the film (fabricated with a complex pattern of rotations and depositions). These four elementary microstructures are shown in Figure 5. Electrical, optical, and magnetic anisotropy can also be increased by depositing obliquely from two sides instead of one, and this technique (called serial bideposition) can be used to increase the anisotropy in sculptured thin films. At extremely oblique deposition angles, a regime called glancing angle

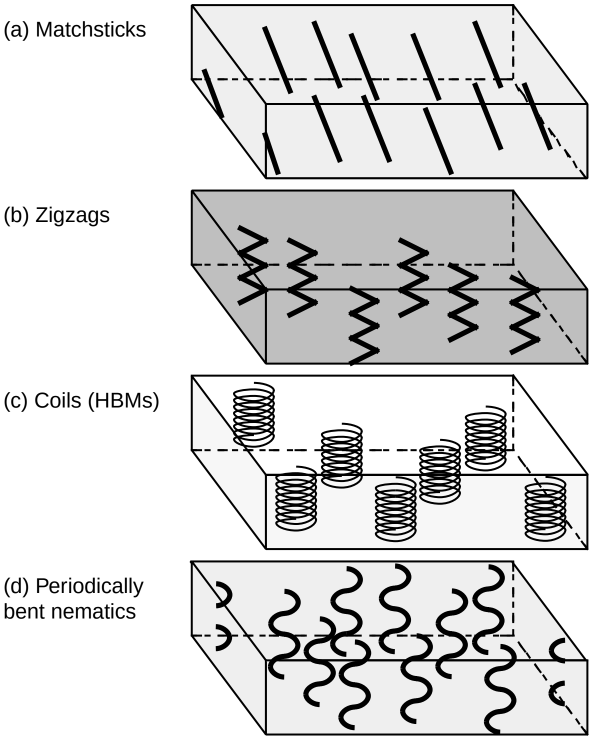

Figure 5 Elementary microstructures of sculptured thin films. From Lakhtakia A, Messier R, Brett MJ and Robbie K (1996) Sculptured thin films (STFs) for optical, chemical and biological applications. *Innovations in Materials Research* 1(2): 165–176, reproduced with permission.

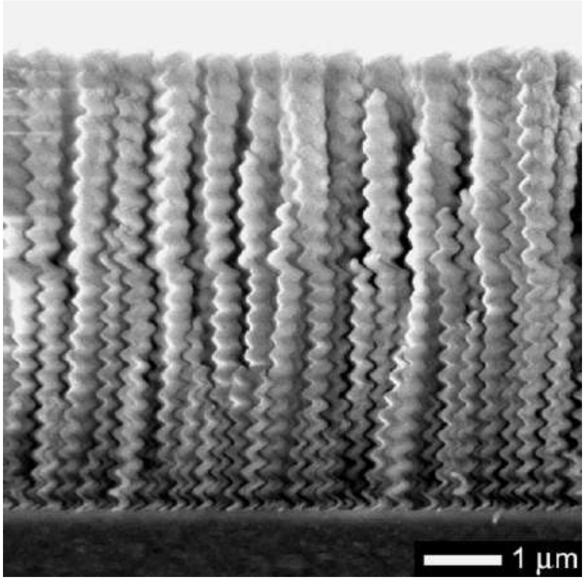

Figure 6 Micrograph of MgF_2 film deposited with a tilt angle of 85° and substrate rotation clockwise for the first half and counterclockwise for the second.

deposition, a rapid increase in porosity with approximately constant column tilt results from a competition between deposition processes. Films deposited here are highly porous and strongly scatter light, limiting their usefulness for conventional optical devices, though some application has been seen in these highly porous films when impregnated with liquid crystals, and switched similarly to liquid crystal displays. An example of a highly porous optical filter with circular polarization effects is shown in Figure 6, where two layers of helicoidal porous films are deposited with different handedness and pitch. Sculptured thin films for most optical devices are deposited at less oblique angles to reduce porosity and therefore diffuse light scatter (haze) as well as sensitivity to humidity and other environmental changes.

Combining the concepts of film growth and optical response to morphology leads to the hoped-for vision of sculptured thin films: optical circuits in a chip. Combining the elementary microstructures described above, and employing powerful computer growth modeling and optical design programs, researchers are currently defining and producing sculptured thin films with increasingly complex optical response characteristics. There are substantial steps required to move from the fairly simple optical filters fabricated today to large-scale devices with complex functionality, but continuing success and increasing interest suggest that sculptured thin films could become a dominant optical technology in years to come.

Current Research

While thin films deposited at off-perpendicular incidence angles have been studied for many years, the last decade has produced significant new insights into these materials, leading to the concept of sculptured thin films. Understanding that morphologically rich materials can be used to produce new optical effects has yielded some impressive demonstrations. Research is also underway to explore these materials for chemical, magnetic, biological, and other applications.

Significant research has been conducted in the past few years to develop and extend the concepts of sculptured thin films. Optical studies, both theoretical and experimental, have revealed a range of remarkable properties including circular birefringence and Bragg reflection, liquid crystal alignment and switching, optical interconnection, and even suggestions of photonic bandgaps. Studies of optical sensing of gases and other materials is progressing, and simulation and modeling of the growth processes is ongoing. The following are two examples of optical materials that have been fabricated with sculptured thin-film techniques; one is a polarization discriminatory light-handedness inverter and the second is a liquid crystal switching cell that can be used to electrically modulate the transmission of polarized light. The handedness inverter converts left circularly polarized light to right circularly polarized within a spectral band, and blocks right circular polarized light. It is constructed by combining a layer with vertical columnar, or matchstick, morphology together with a layer with helicoidal morphology. In keeping with the sculptured thin film concept, the films are deposited sequentially during the same deposition, and optical properties are monitored during growth. The transmission spectrum for the devices is shown in Figure 7, where T_{RL} and T_{LR} are the light transmittances of right circularly polarized light when the incident is LCP, and of LCP when the incident is RCP, respectively. LCP light is almost entirely converted to RCP upon transmission, and RCP light is blocked. The bandwidth of the devices is about 50 nm, determined by matching of the circularly polarized light with the layer with helicoidal morphology.

When films are deposited with helicoidal morphology at glancing angles, i.e., glancing angle deposition, highly porous structures are produced where the columns stand alone and are individually shaped by the deposition process. Optical devices can be constructed by filling these porous films with gases or liquids, and there have been demonstrations of display-type switching with liquid crystals, and discussions of chemical sensing possibilities.

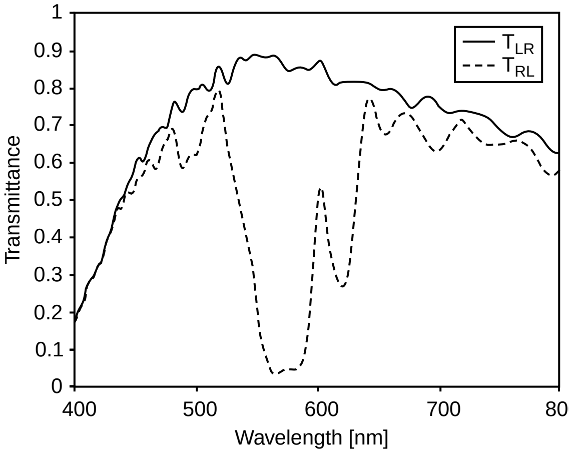

Figure 7 Transmission spectrum for visible light through a polarization-discriminatory light-handedness inverter. Adapted with permission from Hodgkinson IJ, Lakhtakia A and Wu Q-H (2000) Experimental realization of sculptured-thin-film polarization-discriminatory light-handedness inverters, *Optical Engineering*, **39**: 2831–2834.

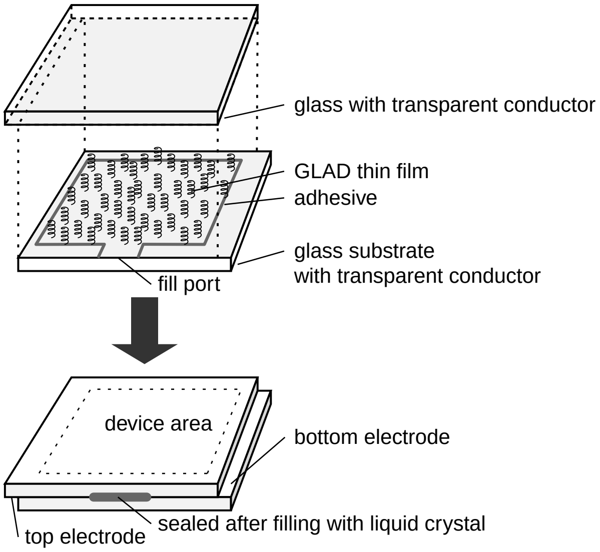
glass with transparent conductor
GLAD thin film
adhesive
glass substrate
with transparent conductor
fill port
device area
bottom electrode
top electrode
sealed after filling with liquid crystal

Figure 8 Liquid crystal switching cell schematic with highly porous glancing angle deposited (GLAD) film between glass plates coated with transparent electrodes. Reproduced with permission from Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391, www.tandf.co.uk/journals.

The device shown in Figure 8 is a liquid crystal switching cell where the space between the electrodes, usually composed solely of liquid crystalline material, is now filled with a highly porous helical structured film. The film structure is somewhat similar to helicoidal film produced for the handedness inverter discussed above, but glancing angle deposition has produced a film of individual formed columns instead of a more dense film with a helicoidal void morphology. When the porous portions of the film are filled with simple nematic liquid crystalline material, the orientation of the rod-like molecules is strongly influenced by the helical columns. An orientational phase occurs in the liquid crystals mirroring the helical structure of the columns in the film, and producing optical properties similar to chiral nematic liquid crystals. The discrimination of left and right circularly polarized light increases relative to the unfilled film. When an electric field is applied across the liquid crystal cell, dipole moments in the liquid crystal molecules force them to align with the field, and chiral optical properties vanish (see Figure 9). Used as an electro-optic filter for circularly polarized light, this cell can switch light transmission on and off. Initial work has not shown clear benefits of this system over conventional liquid crystal display technology, but control of liquid crystal orientation throughout the cell instead of just at the glass plates appears promising. Possibly more complex optical circuits, incorporating layers of porous films filled with liquid crystals, will be useful.

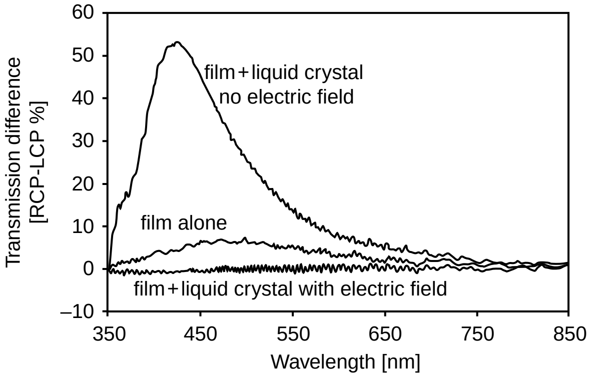

Figure 9 Transmission spectrum for visible light through a liquid crystal cell where the difference between right- and left-circularly polarized light transmission is eliminated with applied electric field. Adapted with permission from Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391, www.tandf.co.uk/journals.

Theoretical and experimental study of sculptured thin films is in its infancy, and many exciting uses for these materials are expected in coming years. Continuing study of thin film growth dynamics will improve and expand the range of structures attainable. Theoretical investigations will aid in the design and understanding of the optical response.

List of Units and Nomenclature

Aluminum (Al)	white metal used in airplanes, mirrors, and microelectronic circuits

Anisotropic	having properties that vary with orientation (rotation)
Biaxial	having anisotropic optical response characteristics that require two axes to describe
Chiral	having a shape that exists in two mirror image forms, called enantiomers
Diamond	carbon crystallographically ordered in a face-centered-cubic lattice, and used in cutting tools and jewelry
Electromagnetic waves	traveling waves of energy that compose light, x-rays, radiowaves, etc.
Epitaxy	growth of a thin film that preserves the crystal structure of the underlying film or substrate
Etching	removing material with chemical or physical atomic scale process
Graphite	carbon crystallographically ordered in hexagonal sheets, and used in pencils and lubricants
Inhomogeneity	varying with translation (position)
Isotropic	having properties that are identical in all directions
Lithography	process where patterns are produced in a sensitive layer by irradiating the layer through a mask with the desired pattern
Magnesium fluoride	ceramic (MgF_2) that is used in thin-film or single-crystal form in many optical systems
Morphology	microstructural shape, or physical arrangement of atoms in a material
Nanometers (nm)	one billionth of a meter, 10^{-9} m
Nematic liquid crystal	optical material composed of rod-shaped organic molecules that usually exhibits uniaxial optical properties
Pitch	physical period for one full rotation of a helical structure
Polarization	orientation and phase of the electric and magnetic fields in a traveling electromagnetic wave; can be linear, circular, or generally, elliptical
Refractive index	the ratio of the speed of light in vacuum to the speed of light in a material
Substrate	object that a thin film is deposited onto, often polished glass or semiconductors
Ultraviolet (UV)	electromagnetic radiation (light) with wavelength shorter than visible light and longer than X-rays
Uniaxial	having optical response characteristics that are anisotropic, but are rotationally symmetric about one axis
Vacuum	absence of matter, typically produced inside sealed glass or steel chambers by removing air with pumps
Vapor flux	moving gas-like collection of material, emitted from a source such as an evaporator or effusion cell, that will condense to form a thin film

See also

Coherence: Overview. **Coherent Lightwave Systems**. **Fiber and Guided Wave Optics:** Overview. **Imaging:** Imaging Through Scattering Media. **Information processing:** All-optical Multiplexing/Demultiplexing; Coherent Analogue Optical Processors; Optical Neural Networks. **Instrumentation:** Ellipsometry. **Optical Processing Systems**. **Optical Coatings:** Anti-Counterfeiting and Decorative Coatings; Optical Black Surfaces; Thin-Film Optical Coatings. **Optical Materials:** Measurement of Optical Properties of Solids. **Polarization:** Introduction; Matrix Analysis. **Scattering:** Scattering from Surfaces and Thin Films; Scattering Theory. **Semiconductor Materials:** Band Structure Engineering. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Ball P (1997) *Made to Measure*. Princeton, NJ: Princeton University Press.

Barabasi A-L and Stanley HE (1982) *Fractal Concepts in Surface Growth*. Cambridge, UK: Cambridge University Press.

Collins PJ (1990) *Liquid Crystals*. Princeton, NJ: Princeton University Press.

de Gennes PG and Prost J (1993) *The Physics of Liquid Crystals*. Oxford, UK: Clarendon Press.

Hodgkinson IJ and Wu Q-H (1997) *Birefringent Thin Films and Polarizing Elements*. Singapore: World Scientific Publishing.

Hodgkinson IJ, Lakhtakia A and Wu Q-H (2000) Experimental realization of sculptured-thin-film polarization-discriminatory light-handedness inverters. *Optical Engineering* 39: 2831–2834.

Hodgkinson IJ, Wu Q-H, Thorn KE, *et al.* (2000) Spacerless circular-polarization spectral-hole fitters using chiral sculptured thin films: theory and experiment. *Optics Communications* 184: 57–66.

Lakhtakia A, Messier R, Brett MJ and Robbie K (1996) Sculptured thin films (STFs) for optical, chemical and biological applications. *Innovations in Materials Research* 2: 165–176.

Liu F, Umlor MT, Shen L, *et al.* (1999) The growth of nanoscale structured iron films by glancing angle deposition. *Journal of Applied Physics* 85: 5486–5488.

Messier R, Venugopal VC and Sunal PD (2000) Origin and evolution of sculptured thin films. *Journal of Vacuum Science and Technology* 18: 1538–1545.

Ohring M (1992) *The Materials Science of Thin Films*. New York: Academic Press.

Robbie K, Brett MJ and Lakhtakia A (1996) Chiral sculptured thin films. *Nature* 384: 616.

Singh ON and Lakhtakia A (eds) (2000) *Electromagnetic Fields in Unconventional Materials and Structures*. New York: John Wiley.

Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391.

Toader O and John S (2001) Proposed square spiral microfabrication architecture for large three-dimensional photonic band gap crystals. *Science* 292: 1133–1135.

Wang L and Clancy P (2001) Kinetic Monte Carlo simulation of the growth of polycrystalline Cu films. *Surface Science* 473: 25–38.

Smart Optical Materials

P M Martin, Pacific Northwest National Laboratory, Richland, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Smart optical materials are those materials whose optical properties (transmittance, reflectance, absorptance, emittance, fluorescence) can be controlled by external stimuli; either by an applied electric field or voltage, mechanical stress, incident light or electromagnetic field intensity, temperature variation, or time duration. This includes electrochromic, photochromic, thermochromic, nonlinear optical, light emitting, fluorescing, emissive, and piezochromic materials. These materials can be in various states, such as solids, gels, composites, and mesophases. Electrochromic and photochromic materials, however, are used most extensively.

The mechanism for adaptive control of light must be reversible, have a well-defined response time, and have an activation threshold, otherwise, the material is considered passive. The primary mechanisms that can cause variations in optical properties are crystal phase transitions, nonlinear effects in the polarizability, optical energy band transitions and band bending, compositional changes, defect and color center formation, and lattice strain. Crystal phase transitions change the bonding morphology of a material, which affects the energy bandgap. Most optical materials have several possible crystalline phases, depending on temperature. The optical properties of the anatase and rutile phases of TiO_2, for example, are very different, and result from different energy bandstructures (recall that the energy bandstructure depends on lattice spacing and crystalline plane orientation). The crystalline phase of several thermochromic materials changes from the martensite structure to the austenite structure with change in temperature, and an associated change in optical properties. Compositional changes are caused by chemical or physical reactions in the material, such as a metal ion bonding to an atom in the lattice, or loss of an atom in the lattice. The reaction can be caused by an energetic photon or ion driven by an electric field; electrochromic materials fit into this category. If their energy is high enough, incident photons can create defect states in the bandgap of materials, which in turn form color centers that absorb light. Modulated electric and electromagnetic fields can cause changes in the linear and nonlinear polarizability of the atoms in a lattice. Nonlinear optical and piezochromic effects are based on this mechanism.

Photochromic Glass

Photochromic glass is one of the most widely used smart optical materials, and definitely the most widely marketed. It has been around for over 30 years with hundreds of associated patents. The transmission of photochromic materials adjusts to the light level of their environment, but these materials give no output information for data processing or communications. They do, however, require no other external control mechanism other than the intensity of light (characteristics of a true smart material). Photochromic materials darken reversibly when exposed to sunlight or a specific optical (UV–LWIR) wavelength band. Absorption in specific narrow or broad wavelength bands is due to the formation of color centers, or point defect states in the lattice. The kinetics of darkening has been described by a two-color center model with two rate constants. Energy (usually in the ultraviolet wavelength range) of the incident photons, ionizes atoms in the glass, thus forming 'F' centers. F centers are highly absorptive and the absorption generally occurs in a broad wavelength band. The earliest photochromic glass developed utilized a dispersed and sensitized colloidal silver halide as the darkening agent. Copper oxide was a typical sensitizer and AgCl is also a fairly common impurity used in PC glass. This material darkens by the equation

$$\mathrm{Ag^{I} + Cu^{I} \rightleftharpoons Ag^{0} + Cu^{II}} \qquad [1]$$

This type of reaction is fairly typical of these materials. The defect state is created in the silver and copper, creating optical absorption. Photochromic glass can darken down to 22% transmittance in sunlight and become nearly clear indoors, transmitting 85% of visible light. Figure 1 shows the transmittance of the darkened and normal states. Temperature, however, does have an effect on the lenses. The glass will not darken completely at temperatures above 32 °C, and at cold temperatures the lens will get extremely dark in a shorter period of time. Recently, $BaO{-}R_2O_3{-}SiO_2$ (R = Al, B, La) nanoparticles have been introduced to enhance photochromicity.

Thermochromic Materials

Thermochromic materials darken and lose transmittance when heated above a critical temperature. They can be as sophisticated as a multilayer thin film optical interference coating or as mundane as a cereal bowl that changes color when heated.

Thermochromic products are used to darken windows of automobiles, for example, when the internal temperature gets above a specific value. The conductivity of these materials also decreases with darkening in many cases because additional electrons are excited into the conduction band due to the increased number of defect states in the bandgap. The most common inorganic thermochromic thin film materials are transition metal oxides, and among these, the vanadium oxide family (V_2O_5, VO_2) shows the best performance. Figure 2 shows the transmittance of a VO_2 thin film before and after heating. Here T_t is the transition temperature. When heated above T_t, the film darkens. Stoichiometric VO_2 darkens at a temperature near $T_t = 68\,°C$ due to a semiconductor to metal transition. The resistivity of VO_2 decreases by three orders of magnitude and the emittance can be increased up to 90% at the transition temperature. This material has promising applications for thermal control of space structures, where surfaces are exposed to intense solar radiation and periods of darkness at very low temperatures. The transition temperature can be engineered (increased or decreased) by doping with metals such as W and Ta.

Loss of transmission can be achieved by absorption or increased reflectance. Reflectance increases with heating are observed in substoichiometric WO_3 films and absorptance increases in stoichiometric films. The reflectance increase is attributed to a decrease in W/O ratio and an associated increase in the refractive index (n) and extinction coefficient (k).

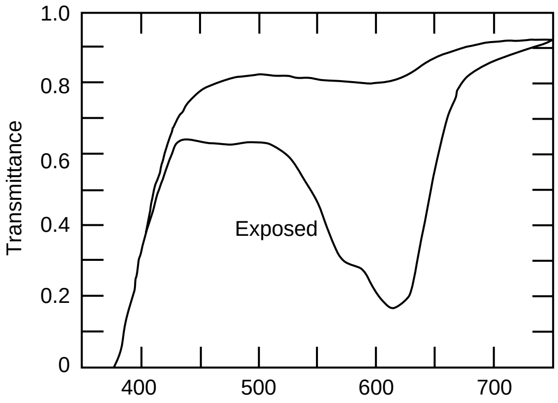

Figure 1 Transmittance of photochromic glass in darkened and normal states. (Reproduced with permission from Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.)

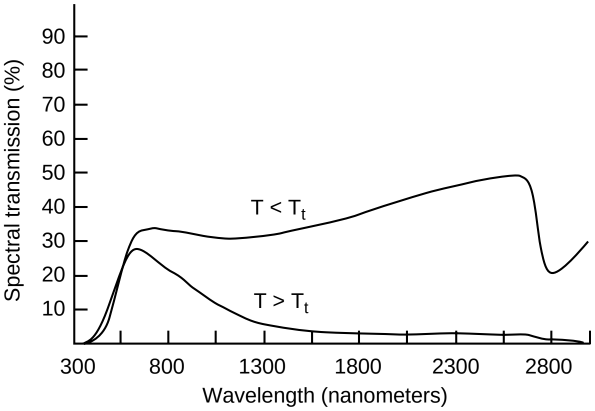

Figure 2 Change in transmittance of a thermochromic VO_2 film with heating. $T < T_t$ for normal transmission and $T > T_t$ when heated. (Reproduced with permission from Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.)

Electrochromic Materials

Electrochromic materials, also called switchable materials, are usually used in the form of thin film devices. Electrochromic windows, or smart windows, have finally reached maturity and promise to be the next major advance in energy-efficient window technology. These devices are used in architectural glazings, automotive mirrors, rear view mirrors, sunroofs, sunglasses, and other high-end applications. Electrochromic devices are well suited for both solar and thermal control. A window can be controllably darkened during periods of bright sunlight and switched back to transparency during the evening or during cloudy periods. Windows with multilayer thin film electrochromic coatings change color and darken with an applied voltage, and bleach with a reverse bias. The magnitude of the applied voltage determines the degree of coloring and bleaching. In addition to a change in visible transmission, the NIR transmission and reflectance, and the associated thermal properties, can also be controlled. The major advantages of electrochromic devices are: (1) power is used only during switching; (2) switching voltages are small (1–5 V); (3) reflectance and transmittance is always specular; (4) gray scale variations are possible; (5) polarization is minimal (reduces birefringence and distortion); and (6) the memory is adjustable in many cases.

The electrochromic effect occurs in inorganic compounds by dual injection (cathodic) or ejection (anodic) of ions (M) and electrons (e^-). A typical reaction for a cathodic coloring material is

$$WO_3(\text{colorless}) + yM^+ + ye^- \rightleftharpoons MyWO_3(\text{blue}) \quad [2]$$

where M is H^+, Li^+, Na^+, Ag^+, etc.

A typical anodic reaction is

$$Ir(OH)_x(\text{colorless}) \rightleftharpoons IrO_x(\text{black}) + XH^+ + xe^- \quad [3]$$

Movement to the right creates color or darkening, and movement to the left results in bleaching. Figure 3 shows the spectral data for the switching of a tungsten oxide film over the solar spectrum; the spectral response of the coating in the bleached and colored (or 'switched') states, and Figure 4 shows the bleached and darkened states of an actual electrochromic window.

Figure 5 shows the design of a simple electrochromic coating. The basic electrochromic coating consists of a transparent top electrode (usually indium tin oxide – ITO), which also injects electrons into the electrochromic layer beneath. By far the most studied and most promising electrochromic material is tungsten oxide (WO_3), which can have either an amorphous or crystalline microstructure. WO_3 has a high coloration efficiency and capacity. Other inorganic materials that are presently being developed for solar control applications are various forms of WO_3, NiO, $WMoO_3$, MoO_3, and IrO_x. Many of these are doped with a conduction ion such as Li^+ or H^+. The ion conductor layer is used to transport protons (H^+), which are supplied by the ion storage or counter-electrode layer, into the electrochromic material. The most promising ion conductors are certain immobile solvent polymer systems, ionic glasses, and open-channel metal oxide structures such as perovskites. Oxides of Ta and Nb also show promise. The bottom layer is another transparent conducting coating. It is possible to inject both electrons and protons into the electrochromic material, thereby inducing a strong absorption band in a given region, e.g., the visible spectrum, that results in a color change. In an

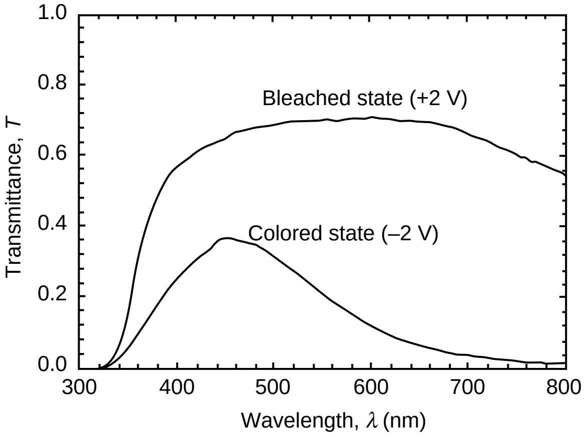

Figure 3 Spectral switching of an electrochromic window for darkened (−2V bias) and bleached (2V bias) states. (Reproduced from Kre J, Topic M, Smole F, *et al.* (2002) *Solar Energy Materials and Solar Cells* 71: 387–395, with permission from Elsevier.)

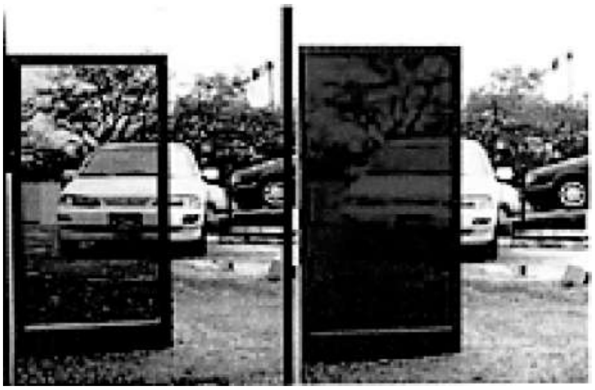

Figure 4 Bleached (left) and darkened (right) states of an electrochromic window.

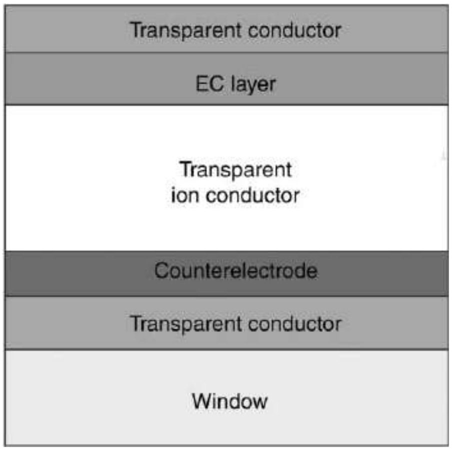

Figure 5 Basic layer design of electrochromic coating for use on a window.

Table 1 Electrochromic device structures

Electrochromic layer	*Ion conductor or electrolyte*	*Ion storage*
a-Li_xWO_3	$LiClO_4$ + PC	Redox Cple/NiO
a-WO_3	Ta_2O_5	$Ir_xSn_yO_2$:F
a-Li_xWO_3	$LiClO_4$ + PC	Prussian blue
a-H_xWO_3	Polymer	Polyaniline
a-H_xWO_3	SiO_2/metal	WO_3
Viologen	PMMA + organic	none
a-WO_3	Ta_2O_5	a-IrO_2
c-WO_3	Li–B–SiO Glass	IC/$Li_xV_2O_5$
a-WO_3	Li–PEO	CeO_x
a-Li_xWO_3	PPG–$LiClO_4$–MMA	$Li_yV_2O_5$
a-H_xWO_3	Ta_2O_5	NiO
a-WO_3	Li–PEO	NiO
NiO	a-PEO copolymer	Nb_2O_5, WO_3:Mo
a-WO_3	a-PEO copolymer	Polymer
Polyaniline	HCl	–

a, amorphous; c, crystalline; IC, ion conductor; MMA, methyl methacrylate; PC, polycarbonate; PEO, poly (ethylene oxide); PMMA, poly (methyl methacrylate); PPG, poly (propylene glycol).

electrochromic device, an externally applied electric field is used to control the injection process (voltage is applied across the two transparent electrodes). The coloration remains for some time, even after the external field is removed. The system returns to the initial state upon reversing the polarity of the external field. It is interesting to note that electrochromic windows being developed by NREL now employ solar cells to facilitate optical switching.

There are two major categories of electrochromic materials: transition metal oxides (including intercalated compounds) and organic compounds (including polymers). Table 1 shows a few of the most common electrochromic materials. Organic electrochromic materials are based on the viologens, anthraquinones, diphthalocyanines, and tetrathiafulvalenues. With organic compounds, coloration is achieved by an oxidation–reduction reaction, which may be coupled to a chemical reaction. The viologens are the most studied of the organic electrochromics. Originally, organic electrochromics tended to suffer from problems with secondary reactions during switching, but recently more stable organic systems have been developed. However, their reliability in harsh environments is questionable, and protective encapsulation may be required.

Electrochromic nickel oxide (NiO) is also being developed. NiO has been combined with manganese oxide, cobalt oxide, and niobium oxide in all alkaline devices.

Nonlinear Optical Materials

Nonlinear optical materials have evolved slowly over the past 20 years. At relatively low light intensities, the properties of most optical materials are independent of intensity of illumination. There is no interaction between the light waves in the medium. However, if the intensity of illumination gets high enough, nonlinear optical effects can occur and the optical properties become intensity dependent. When an electromagnetic field oscillating at optical frequencies ($\sim 10^{13}$–$\sim 10^{17}$ Hz) is applied to the lattice of a dielectric material, the electrical charges in the lattice become polarized and form electric dipoles. The positive ion cores do not move significantly, but the electrons are displaced in a harmonic motion. The polarization $\mathbf{P}$ is defined as $-Ne\mathbf{X}$, where N is the number of electric dipoles, $\mathbf{X}$ is the displacement of the electrons, and e is the charge on an electron. Here $\mathbf{X}$ is a complex number. Now, $\mathbf{P}$ is related to the electric field E by $\mathbf{P} = \varepsilon_0 \chi E$, where χ is the susceptibility and ε_0 is the permitivity of free space.

The linear dielectric constant is $\mathrm{Re}(1 + \mathbf{X})^{1/2}$, which is just the index of refraction. The imaginary part causes optical absorption.

When the intensity of illumination is high, the elastic limit of the electron's displacement is exceeded and the harmonic restoring force becomes nonlinear in $\mathbf{X}$, which adds quadratic and higher order terms to the expression for $\mathbf{X}$ and $\mathbf{P}$. In addition to the fundamental resonance, the dipoles now also oscillate at twice and thrice their fundamental resonant frequency (second and third harmonics). The polarization then is also nonlinear in displacement and can be expressed by the well-known relation

$$\mathrm{P} = \varepsilon_0(\chi^{(1)}\mathrm{E} + \chi^{(2)}\mathrm{E}^2 + \chi^{(3)}\mathrm{E}^3 + \ldots) \qquad [4]$$

where $\chi^{(2)}$ and $\chi^{(3)}$... are called the nonlinear susceptibilities. The second-order polarization $P^{(2)}$ causes optical mixing phenomena, involving sums and difference frequencies. When mixed with a DC field, the refractive index of the $\chi^{(2)}$ material can be controllably varied by varying the DC field. This is the well-known Pockels effect and is used in optical modulators. Figure 6 shows one method that is used to modulate the refractive index by the photorefractive effect. Here two laser beams (A_1 and A_2) are mixed in the $\chi^{(2)}$ material with the resulting intensity $I = I_o(1 + m\cos(kx))$. The material is polarized in phase with the mixed beams $\rho = \rho_o(1 + m\cos(kx))$, which results in a modulated electric field E_{sc}. The refractive index (Δn in the figure) is modulated in phase with the mixing signal. The $\chi^{(2)}$ materials are also used for frequency doubling.

Barium titanate ($BaTiO_3$), lithium niobate ($LiNbO_3$), KH_2PO_3 (KDP), lithium tantalite ($LiTaO_3$) and potassium niobate ($KNbO_3$) are $\chi^{(2)}$ optical switching and frequency doubling materials that have been studied for decades. Note that because they can be easily polarized, these are also excellent piezoelectric/ferroelectric smart materials.

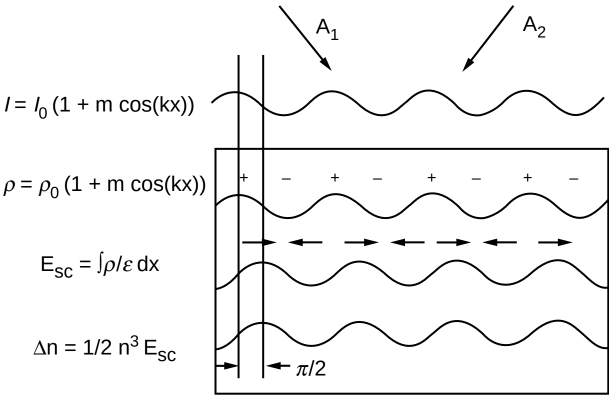

Figure 6 Modulation of the refractive index of a nonlinear optical material with incident laser modulation. From top to bottom: modulated laser intensity I, density ρ, electric field E_{sc}, refractive index Δn, and phase angle $\pi/2$. (Reproduced with permission from Bass M (ed.) (1995) *Handbook of Optics II*. New York: McGraw-Hill.)

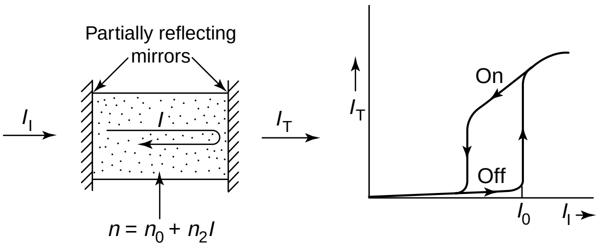

Figure 7 Optical switching behavior of a Fabry–Perot filter. I_I and I_T are the incident and transmitted intensities respectively. (Reproduced with permission from Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.)

The $\chi^{(3)}$ effects are sought after for applications such as laser eye protection, optical sensor protection, optical switching, optical limiting, and signal processing. In this case the refractive index (real and imaginary) depends on the intensity of the incident radiation. Figure 7 shows how this nonlinear effect is used in an optical switching device. It is all based on the change in the optical path length of a Fabry–Perot filter. At normal intensities, the refractive index of the spacer layer is n_0 and the filter is in the 'off' state. The refractive index becomes $n_0 + n_2 I$ and the optical thickness (nt) of the spacer layer changes at high illumination intensities, which in turn shifts the resonant wavelength (frequency) of the Fabry–Perot filter and creates the optical switch ('on' state).

There are a large number of $\chi^{(3)}$ and $\chi^{(2)}$ organic materials, including dyes, dimethylamino nitrostilbene, methyl nitroanaline, poly-BCMU, polydiacetylenes, and urea. Because they are essentially chains, many organic molecules can be easily polarized and possess higher order susceptibilities. The polarizability of organic materials is often enhanced by the mobility of delocalized π electrons in C–C bonds in aromatic rings. Hybrid nonlinear materials have even been formed by combining organic and inorganic components such as metallophthalocyanines which display very strong excited state absorptions. Most of these materials are deposited by spin casting.

Emissive Materials

Two of the newest applications of smart optical materials are emissive displays and electronic ink. This encompasses light-emitting polymers and small molecules, light-emitting diodes, and quantum cascade lasers. Organic light-emitting device (OLED) technology is leading the revolution in the flat panel display industry. This device was first invented by Tang and Van Slike and first reported in *Applied Physics Letters* in 1987. Figure 8a shows the layer structure of the OLED and Figure 8b shows a blue OLED pixel. The device is deposited onto an indium tin oxide (ITO)-coated glass or plastic substrate, with the ITO acting as the anode. The active layers can be either light-emitting polymers or small molecules. About 500 Å of a hole transport material (hole transport layer, HTL), here a tertiary amine, is deposited over the ITO. An aluminum trisquinalate (Alq), which transports electrons (electron transport layer, ETL) is deposited next. The device is topped off with

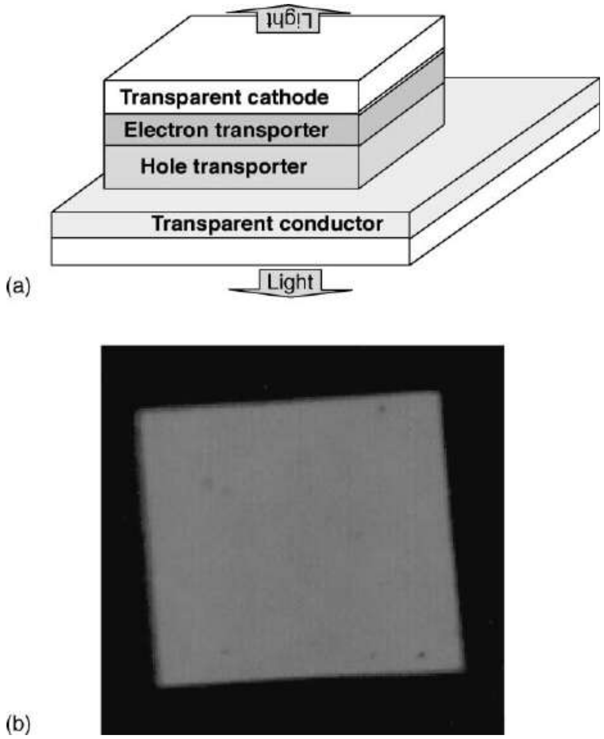

Figure 8 (a) Layer diagram of an OLED. (b) Blue OLED pixel.

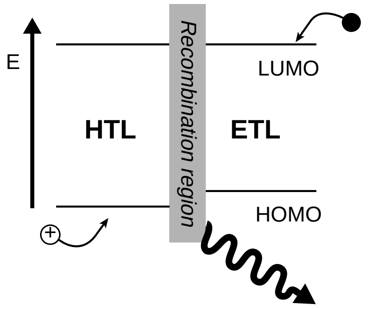

Figure 9 Simple energy level diagram of an OLED showing the relationship between the ETL and LUMO, and HTL and HOMO.

a low work function cathode, usually Ag:Mg or Ca. The entire device thickness is about 1000 Å. Even a dust particle can ruin the operation of an OLED by creating inactive regions or shorting out the electrodes.

The OLED emits light in a blue, green, or red wavelength band when activated by an electric current, i.e., electron and hole injection. The device functions by injecting holes from the ITO into the HTL. As shown in the simplified energy band diagram (Figure 9) for an OLED, the hole moves into the HOMO (highest occupied molecular orbital), which loosely corresponds to the valence band of a semiconductor. Electrons are injected from the cathode into the ETL, where it is transported into the LUMO (lowest unoccupied molecular orbital), which loosely corresponds to the conduction band of a semiconductor. When a voltage is applied between the anode and cathode, an electron flows and the holes and electrons combine in the ALQ layer and form excitons. Light is emitted when the excitons decay back a lower energy state. The color of the emitted light (green, red, or blue) can be changed by changing the HTL and ETL material composition. Light is emitted by fluorescence processes, from both singlet and triplet states if the material chemistry is correct, which results in 100% internal quantum efficiency. Twenty percent external quantum efficiency results from using all available exciton states. Photons can be lost through absorption in the layers, scattering, reflection between layers, and waveguide modes. These factors must be minimized to maximize the external quantum efficiency.

OLED video displays now have outputs near 100 cd/m (20 lm/W), which is about twice the normal video brightness. Red, green, and blue OLEDs now emit with very high efficiencies. For comparison, a 100 W light bulb emits at 17 lm/W and the best fluorescent light emits at 80 lm/W. Today, OLEDs are already sitting at intensities of 50–60 lm/W, and could easily push 150 lm/W in the near future. Another advantage of OLED displays is their wide viewing angle, up to 160°.

There are several types of OLED display configurations. Figure 10 shows a passive matrix display geometry and placement of the OLED layers. Here the OLED is sandwiched between crossed cathode and anode contacts. By injecting current between sets of crossed anode and cathode rows, the OLED will light up, each contact point forming a pixel. Displays can typically have as many as 3 million pixels. OLED displays seem brighter and livelier to the eye because the colors are additive, rather than subtractive as with LCD displays. For a full-color passive matrix display green, red, and blue pixels must be located very close together. Sony has developed an active matrix 13-inch diagonal display with 800 × 600 pixels (∼150 000 functioning OLEDs).

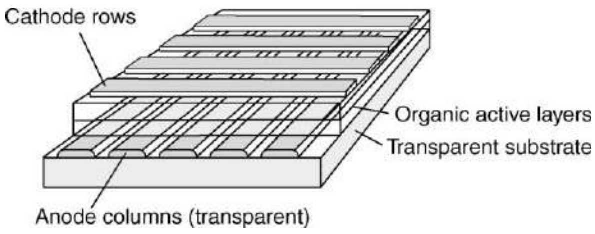

Figure 10 Geometry of a passive matrix display.

Among the more promising pixel geometries is the SOLED, the stacked OLED, where, red, green, and blue pixels are stacked on top of each other. It is mandatory that the pixels and the electrical contacts be transparent for this design, or light will not be transmitted through each layer. Very thin Mg contacts are used with a top sputtered ITO cathode. The emitted color is simply changed by applying current across the pixels, with white being obtained when all pixels are activated.

The future of OLED displays is plastics, or the flexible OLED (FOLED). Plastic substrates are lightweight, low cost, and rugged. However, they have an inherent problem of being very permeable to oxygen and water vapor, both of which severely degrade performance of the OLED by reacting with the anode layer. For displays to be practical, they must have lifetimes in excess of 10 000 hr and preferably 20 000 hr. Barrier coatings are now being developed for plastics to prevent oxygen and water ingress and extend the life of the OLED to this goal.

Electronic ink is a display medium that can be formatted using a passive matrix. It consists of millions of tiny negatively charged microcapsules, about 100 microns in diameter, floating inside a liquid medium. Each microcapsule is filled with hundreds of pigmented chips. One half of each microcapsule is white and the other half is dyed blue. When an electric charge is applied to the microcapsules, the chips are drawn either to the bottom or top of the microsphere. The viewer sees dark areas when they are drawn to the top. In another version being developed by Xerox Corp., half the chip is white and half the chip is dark. When an electric charge is placed on the chip, the dark side is oriented up and the white side down, creating dark regions where charge is placed. This can be easily erased and reformatted into text, symbols, signs, and pictures.

Future smart optical materials will simulate biological systems such as an insect's eye and will employ quantum structures to achieve optical switching, emission, and control of optical properties.

List of Units and Nomenclature

Absorptance (A)	The amount of light absorbed by an optical element divided by the incident intensity. $A = 1 - T - R$.
Active matrix display	A display geometry in which the voltage to the pixel is controlled individually by a transistor.
Candela (cd)	Formerly known as a candle. The luminous intensity equal to 1/60 of the luminous intensity per square meter of a black body radiator operating at the temperature of freezing platinum.
Exciton	An electron–hole pair, usually generated by a photon.
FOLED	Organic light emitting device on a flexible plastic substrate.
LCD	Liquid crystal diode.
Lumen (lm)	A unit of luminous flux equal to the luminous flux emitted in a solid angle of one steradian for a point source having an intensity of one candela.
NIR	The near-infrared spectral region, consisting of wavelengths between 700 and 2500 nm.
Passive matrix display	A display geometry consisting of a grid of horizontal and vertical wires. At the intersection of each grid is an active element (LCD or OLED) which constitutes a single pixel, either letting light through or blocking it.
Piezoelectric	Capable of generating a voltage when a force or stress is applied.
Polarization (coulomb/m^3)	A vector quantity equal to the electric dipole moment per unit volume.
Quantum efficiency	The number of electrons, holes, or electron–hole pairs generated per incident photon at a given optical wavelength.
Reflectance (R)	The intensity of light reflected from an optical element divided by the incident intensity. $R = 1 - T - A$.
Susceptibility (no units)	A scalar quantity equal to the ratio of the electric polarization of a material to the electric field strength.
Transmission (T)	The intensity of light transmitted through an optical element divided by the incident intensity. Note that transmission + reflection + absorption = 1 ($T + R + A = 1$).
UV	The ultraviolet spectral region consisting of wavelengths between 200 and 400 nm.
Visible	The spectral region consisting of wavelengths between 400 and 700 nm.
Wavelength (μm or nm)	The length of one oscillation of a light wave.

See also

Materials for Nonlinear Optics: Organic Nonlinear Materials. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Optical Glasses.

Further Reading

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Chemla DS and Zyss J (1987) *Nonlinear Optical Properties of Organic Molecules and Crystals*, vol. 1. Orlando, FL: Academic Press.

Chemla DS and Zyss J (1987) *Nonlinear Optical Properties of Organic Molecules and Crystals*, vol. 2. Orlando, FL: Academic Press.

Clark DT (1994) Second European Conference on Smart Structures and Materials, *SPIE Proceedings*, vol. 2361. Bellingham, CA: SPIE.

Crowson A (1996) Smart Materials Technology and Biomimetics, *SPIE Proceedings*, vol. 2716. Bellingham, CA: SPIE.

Gobin P-F and Friend CM (2000) Smart Structures and Materials, *SPIE Proceedings*, vol. 4073. Bellingham, CA: SPIE.

Janocha H (1999) *Adaptronics and Smart Structures, Basics, Materials, Designs, and Applications*. Berlin: Springer.

Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.

Goff AH-L, Granqvist C-G and Lampert CM (1992) Optical Materials Technology for Energy Efficiency and Solar Energy Conversion XI: Chromogenics for Smart Windows, *SPIE Proceedings*, vol. 1728. Bellingham, CA: SPIE.

McDonuch A, Gardiner APT and McEwen KS (1994) In: Culshaw B (ed.) Second European Conference on Smart Structures and Materials, *SPIE Proceedings*, vol. 2361. Bellingham, CA: SPIE.

Zerbi G (ed.) (1993) *Organic Materials for Photonics*, European Materials Research Society Monographs, V6. Amsterdam: North-Holland.

Sol-Gel Materials

Lisa C Klein, Rutgers the State University of New Jersey, Piscataway, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Sol-gel optics represents a methodology for fabricating optical materials in fiber, thin film, and bulk geometries. Defined broadly, sol-gel optics encompasses a variety of optical devices that contain sol-gel processed components. The optical device may be based wholly on sol-gel processing or incorporate a sol-gel material, along with conventional optical materials. Principally, the optical devices prepared by sol-gel processing or assembled with sol-gel materials are traditional optics or, in some cases, opto-electronics. Among the current sol-gel optical materials, by far the most common are oxides. Further, the majority of sol-gel optics is for transparent components, which typically means glasses, and, consequently, silicates.

In fact, most sol-gel materials have been developed to duplicate conventional silica optical materials. While glass melting technology is highly advanced for making quality lenses, an equivalent material can be made using a sol-gel route. Similarly, ultralow-loss optical fiber can be made by chemical vapor deposition (CVD). Now sol-gel processing can be used to prepare comparable fibers for telecommunications. In cases where sol-gel processing produces an exact duplicate of optical materials produced other ways, the perceived advantage is that sol-gel precursors have a higher purity and better level of mixing than powder raw materials. Perhaps, a more significant advantage is that sol-gel formulations can be designed to promote a particular geometry in the final material, essentially making near-net shape materials.

While sol-gel processing is not designed to produce new optical phenomena, it does offer new ways to assemble materials to produce new optical designs, such as filters, Bragg gratings, amplifiers, and beam splitters. One noted case of sol-gel materials in nonlinear optics is the case where semiconductor particles are trapped in nanometer-sized pores. This leads to so-called quantum dots. The microporosity of the gel, no doubt, assists in this effect, but it is not the gel itself that has produced the optical effect. The optical effect is derived from the nanometer size of the particles. The gel serves as a transparent host that restricts the size of the particles and provides for their uniform distribution.

Sol-Gel Chemistry of Oxide Optical Materials

The sol-gel process for oxides is the name given to any one of a number of processes involving a solution or sol that undergoes a sol-gel transition. Solution refers to a single-phase liquid, while a sol is a kinetically stabilized suspension of colloidal particles. At the transition, the solution or sol becomes a rigid, porous mass by destabilization, precipitation, or supersaturation. Once the solution or sol goes through a sol-gel transition, it is a rigid two-phase system of solid and solvent filled pores. The sol-gel transition is not reversible.

In using a sol-gel process, it is important to be familiar with the reagents, the intermediates and the products. The most important reagent in this case is a hydrolyzable organometallic which is called a metal alkoxide, $M(OR)_x$ where (OR) is an alkoxy group. Alkoxides are the precursors for SiO_2, GeO_2 or B_2O_3, among others. Hydrolysis occurs when the metal alkoxide and water are mixed in a mutual solvent.

The first step is choosing the right reagents. Since most of the glasses are silicates, we start with the silica precursor. Of the available silicon alkoxides, tetraethylorthosilicate (TEOS) is used most often, because it reacts slowly with water and comes to equilibrium as a complex silanol. The colorless liquid $Si(OC_2H_5)_4$ has a density of about 0.9 g/cm^3, is easy to handle safely and is extremely pure when distilled. Most metal alkoxides are prepared by the reaction of halogenated species or pure metals with alcohol.

The two other ingredients in a typical formulation are alcohol and distilled water. In the case of TEOS, ethanol serves as the mutual solvent for TEOS and water. Once TEOS is dissolved in ethanol to make it soluble in water, the chemical reactions hydrolyzation and polymerization begin. The chemical reactions are approximately:

Hydrolyzation

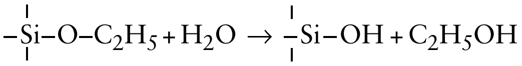

$$-\overset{|}{\underset{|}{Si}}-O-C_2H_5+H_2O \rightarrow -\overset{|}{\underset{|}{Si}}-OH+C_2H_5OH \qquad [1]$$

Polymerization

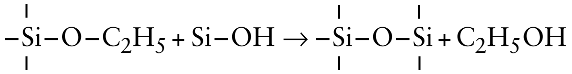

$$-\overset{|}{\underset{|}{Si}}-O-C_2H_5+Si-OH \rightarrow -\overset{|}{\underset{|}{Si}}-O-\overset{|}{\underset{|}{Si}}+C_2H_5OH \qquad [2]$$

An acid is usually added to control the rates of these reactions. The temperature is often ambient, and mixing is continued for between 1 and 3 hours. Complete hydrolysis does not occur. Instead, condensation occurs between silanol and ethoxy groups to give bridging oxygens or siloxane groups. The intermediates that remain soluble in the alcohol–water medium are silanols, ethoxysilanols, and polysiloxanes.

The equipment needed for laboratory-scale production consists of a narrow-necked glass flask with magnetic stir bar on a stirring hot plate. The reactants are added volumetrically to the flask in the order: solvent, TEOS, water, and acid. When everything is dissolved and the solution is clear, the hot plate may be turned on to accelerate reactions or a reflux condenser may be attached.

The parameters that influence the chemical reactions are temperature, pH, amount of water, solvent, and precursor. More complex compositions are easier to mix in longer chain alcohols, and longer chain organic groups give slower reaction rates. Low pH favors a polymerization scheme that gives linear molecules and the solution remains clear. High pH favors cluster growth and the solution may become cloudy. At high pH or high water, solutions produce cross-linked polymers or branched clusters. A small volume fraction of branched clusters restricts flow, when the same volume fraction of linear polymers does not. At a higher volume fraction, the linear polymers become tangled, and then low water solutions will gel. The result of these reactions under all conditions is an increase in the molecular weight of the oxide polymer. Eventually, the solution reacts to a point where the molecular structure is no longer reversible, and the sol-gel transition has been reached.

Shapes, Compositions and Precursor Concentrations

A sampling of compositions and precursors is given in Table 1. Some compositions such as SiO_2 and TiO_2-SiO_2 have been prepared in bulk, to duplicate materials made by melting for lenses and mirror blanks. Other compositions such as $B_2O_3-SiO_2$ have been prepared for thin-film applications. The distinguishing aspect between the solutions formulated for bulk objects and those formulated for thin films is the oxide concentration in the solution. An indication of the solution concentration is the molar ratio of water to alkoxide. The higher ratios are used when bulk objects are desired.

Table 2 gives a sampling of the geometries for sol-gel silica. For each geometry, an example of a shape is given, with an application for each shape and the approximate dimensions of the sol-gel material.

In terms of near-net shape processing, when the material goes through its sol-gel transition, it proceeds to dry and shrink. This phase is characterized either by isotropic shrinkage or anisotropic shrinkage. Monoliths fall into the first category with isotropic shrinkage following molding. Thin films on substrates fall into the second category with anisotropic shrinkage following dipping. Fibers also show

Table 1 Typical compositions of sol-gel optical materials

Oxides	*Precursors*	*Compositions in mol %*	*R*[*]
Single oxide			
SiO_2	$Si(OC_2H_5)_4$	100 SiO_2	16
Two oxides			
$B_2O_3-SiO_2$	$Si(OC_2H_5)_4$ $B(OCH_3)_3$	80 SiO_2–20 B_2O_3	1.5
GeO_2-SiO_2	$Si(OCH_3)_4$ $Ge(OC_2H_5)_4$	90 SiO_2–10 GeO_2	4
TiO_2-SiO_2	$Si(OC_2H_5)_4$ $Ti(OC_3H_7)_4$	94 SiO_2–6 TiO_2	50
$P_2O_5-SiO_2$	$Si(OC_2H_5)_4$ H_3PO_4	90 SiO_2–10 P_2O_5	16
Three oxides			
$Li_2O-Al_2O_3-SiO_2$	$Si(OC_2H_5)_4$ $Al(OC_4H_9)_3$ $LiNO_3$	82 SiO_2–3 Al_2O_3–15 Li_2O	8
$Na_2O-B_2O_3-SiO_2$	$Si(OC_2H_5)_4$ $B(OCH_3)_3$ $NaNO_3$	82 SiO_2–12.4 B_2O_3–5.6 Na_2O	10

[*] Approximate ratio of water to alkoxide, e.g., $R = 16$ means 16 moles of H_2O per mole of $Si(OC_2H_5)_4$.

Table 2 Geometries and approximate dimensions for selected applications of sol-gel silica

Geometry	*Shape*	*Dimensions*	*Application*
Monolith	Flat disk	10 cm diameter 1 cm thick	Lens
	Cylinder	0.5 cm diameter 2 cm long	Rigid host for chromophores
Thin film	Film on Si wafer	500 nm thick	Antireflection
	Film on Hollow tube	~10 000 nm thick	Solar collector panel
Fiber	Rod	0.8 cm diameter 25 cm long	Preform for drawing
	Filament	150 micron diameter/100 m length	Lightguide

anisotropic shrinkage, though it depends on whether or not the fiber is drawn from a preform many times its diameter or produced from a filament that is closer to its diameter.

Sol-Gel Monolith Processing

Monoliths formed from an alkoxide solution are achieved only in dilute solutions. Therefore, alkoxide gels are somewhat difficult to dry because of their small pores (<10 nm). To avoid cracking in alkoxide gels, different treatments have been tried. Aerogels are dried in an autoclave by hypercritical techniques. That is, the solvent is removed above its critical point. The resulting gel is about 10% dense and shows no shrinkage. Xerogels are dried by natural evaporation. Xerogels are 60% dense and have reductions 40–60% in volume.

Alternatively, monoliths can be formed from colloidal solutions. These particles coalesce at gelling. Colloidal gels are easily dried, and they are less likely to crack than alkoxide xerogels. While alkoxide gels have small pores, colloidal gels have larger pores or voids between particles. For the same reason, colloidal gels densify at higher temperatures than alkoxide xerogels. The higher temperature for colloidal gels may lead to undesirable crystallization.

By either route, a serious problem for sol-gel processing is the presence of water. Especially when monoliths are used for fiber drawing, residual OH ions have a deleterious effect on optical properties. That is why non-aqueous solvents, as well as fluorination or chlorination treatments, have been investigated to reduce the OH content.

Sol-Gel Fiber Processing

While monolithic pieces need high water content, fibers can be drawn out of low water content solutions. The sol-gel process allows one to bait and draw a string of gel about the same diameter as the desired fiber directly from the solution. This has been shown where the water to alkoxide ratio was controlled to give the property of spinnability. The time and manner in which the system loses fluidity depend on composition and catalyst. Low water solutions show a gradual increase in viscosity before losing fluidity, while high water solutions lose fluidity more abruptly from a lower viscosity. A problem with this method is the shape of the fiber, which tends not to be cylindrical.

Another approach to making near-net shape fibers from a sol-gel process involves coating a shell or concentric shells of sol-gel glass on a filament of textile fiber. This approach is referred to as the volatile host method, where a plastic filament is used as a template. Multiple dipping or adjustment of the viscosity can lead to different thicknesses. Upon exposure to atmospheric moisture, the films gel. After gelling, a rigid, continuous shell runs the length of the filament.

Once the shell has hardened, solvent and water can escape through interconnected pores. The shell shrinks in the radial dimension, but remains continuous along its axis. During heat treatment and desiccation, little if any densification occurs at temperatures well below the glass transition. The interconnected pore structure remains open until the volatile host is eliminated and undesirable gases are removed. This method is capable of producing optical fibers, but is more suitable for hollow fibers or ion exchange media.

Sol-Gel Thin Film Processing

Like sol-gel fibers, sol-gel thin films are prepared from low water content solutions. The process is exceedingly simple. A solution containing the desired oxide precursors is applied to a substrate by dipping or spinning.

At this time, the majority of sol-gel coatings are applied by dipping. This is an approach that allows the properties of the solution to control the deposition. A substrate is lowered into a vessel containing the solution. A meniscus develops at the contact of the liquid and the substrate. As the substrate is withdrawn, the meniscus generates a continuous film on the substrate. The process is able to apply a coating to the inside and outside of complex shapes simultaneously.

The equipment is inexpensive, especially in comparison to any deposition techniques that involve vacuum. Coatings can be applied to metals, plastics, and ceramics. Typically, the coatings are applied at room temperature, although most need to be calcined and densified with heating. Dip films are typically one micron, uniform over large areas and adherent. The time-to-gel is especially important when it comes to coatings because film formation, drying, and creation of pores must be rapid. Optimum film formulations correspond to those solutions that lose tackiness quickly.

In most cases, the effect of the viscosity is far greater than the surface tension. Two simple relationships are that the film thickness increases with increasing withdrawal rate and, for a given withdrawal speed, the film thickness increases with an increase in oxide content. The process is frequently a batch process, certainly on the laboratory scale. However, the batch process can be scaled up. Repeated dipping builds up a thicker film.

For one-side coating, a technique that is often used is spin coating. In this case, the substrate, usually a silicon wafer, is placed on a spinner, rotated at perhaps 200 rpm while solution is dripped on the center of the substrate. As with dip coatings, a film between 50 and 500 nm results.

After the Sol-Gel Transition

Having selected a shape and designed the formulations accordingly, there are several further steps common to monoliths, fibers, and films. First of all, the gels must be dried. Using the example of silica, TEOS is less than 30% by weight silica and the solution is even less. From the time the solution is applied to the time it gels, there is about 50% weight loss. Then, as the gel dries there is another reduction in weight by one-half. To go along with the weight loss there is about 70% volume reduction.

Films can be dried quickly in air because of the one thin dimension. It has been shown repeatedly that all shrinkage is taken up in the thin dimension and not in the plane of the substrate. Yet, the films remain adherent and continuous and maintain complete surface coverage. To go from a tacky film to a hard film may take only a few minutes.

Following drying and shrinkage comes the heat treatment. It would be inaccurate to say that glass has been formed at room temperature by simply reacting, gelling, and drying the solutions. In this state, the materials have many of the characteristics of the corresponding glass, but they are more or less porous. Water and solvent escape through interconnected pores that remain open at the surface until the gels are fired to temperatures well above 600°C. When needed to fulfill the requirements of an application, the hard gel is heated to various degrees of collapse.

The microporosity in silica xerogels is not removed entirely until 1000°C.

Applications of Sol-Gel Optical Coatings

Sol-gel processing has a special relevance to optical coatings. In most cases, the applications are substitutions of a sol-gel coating for a coating obtainable by other means. Most optical coatings achieve their effect by selective absorption and interference. By depositing 1–14 or more layers of dielectric thin films, using two or more different compositions, a variety of antireflection (AR), half-mirror, hot-mirror and cold-mirror coatings can be designed. An early design was for rear-view mirrors for cars, where the coatings consisted of titania-silica-titania interference filters that gave the effect of total reflectance and replaced conventional metallizing. Another early application was solar reflecting films for windows. These coatings consisted of Pd-containing titania films which show selective absorption.

A sampling of optical coatings is given in Table 3. The application, the mechanism, and the oxide composition are listed. Generally, these coatings can be prepared by other deposition techniques. Those applications which are specific to sol-gel coatings relate to the microporous nature of the sol-gel processed films. In Table 3, there are antireflective coatings of several types. Both borosilicates and titania silicates have been developed for this application. Changes in the composition, changes in the microstructure, and changes in the porosity can be used to grade the index of refraction. Antireflection in the ultraviolet range has been accomplished with thoria and hafnia films. Broadband antireflection has been used for laser optics. The sol-gel coating technique provides control over composition and deposition thickness. The layers, being quarter wavelength in thickness, require precise control.

Table 3 Thin-film applications

Optical application	*Mechanism*	*Typical composition*
Color filters	Selective Absorption	Cr_2O_3–SiO_2
Mirror	Interference	14-layer TiO_2: SiO_2
Antiglare	Absorption	CoO–SiO_2
Antireflection		
Narrow	Graded porosity	SiO_2
Broadband	Interference	Three-layer
	1/4 wavelength	SiO_2–TiO_2:ZrO_2: SiO_2
	Graded porosity	SiO_2
IR reflector	Interference	Three-layer TiO_2:Ag:TiO_2

Sol-Gel Rigid Hosts

An exciting way of looking at sol-gel materials is to use the gel as a host. The important point to note is that the sol-gel materials are transparent. While most of the materials are porous, the porosity is on a scale that does not scatter visible light. Therefore, most light is transmitted. The appeal of the bulk shape host is that it can be molded to near-net shape and can be cut and polished.

Hybrids

Several approaches have been taken to prepare so-called hybrid gels incorporating organic and inorganic components. Those methods that mix the components on the nanometer scale include the infiltration of previously formed oxide gels with monomers, reacting of alkoxide and organic monomers, or organic polymers with alkoxides in mutual solvents. The products of these processes are variously called ormosils (organically modified silicates), ceramers (ceramic polymers), interpenetrating networks (IPNs) and nanocomposites.

The original route to hybrids was one where a gel was formed first as a rigid matrix. Monomer was infiltrated into the open porosity of the shaped gel and polymerized inside the gel using UV curing, for example. Since silica gel is the easiest gel to prepare in monolithic form, one combination that has been investigated widely is silica gel/polymethyl methacrylate (PMMA).

Silica and PMMA, which both have good visible light transmission and similar index of refraction, have been used in hybrids, although it turns out that the match in index of refraction is less critical on the nanoscale. They are also glassy materials with similar strength. PMMA has much higher fracture energy than silica and is often used as glazing in applications where impact resistance is a priority. However, PMMA has a much lower hardness than silica and is less resistant to abrasion. Ideally, silica–PMMA hybrid nanocomposites have the strength and impact resistance of PMMA, and the hardness of silica in one. One obvious application is transparent, lightweight windows.

Encapsulation

Sol-gel hosts have been used to encapsulate a variety of organic and inorganic molecules, chromophores, and biological agents. Whatever is incorporated into the gel can act in two ways, to give information about the gel host or to be influenced by the gel in its interaction with light. For example, spiropyrans have been incorporate for photochromics, CdS for nonlinear optics, Nd for tunable lasers, and various enzymes for sensors. Early on, silica gel hosts were used to encapsulate the electro-optically active organic compounds 2-methyl-4-nitroaniline (MNA). In comparison to conventional optical cells, MNA doped silica hosts retained optical quality at higher temperature and were more abrasion resistant. Since this discovery, silica hosts, in particular, and sol-gel processing, in general, has been used extensively for encapsulating molecules and species that cannot be heated, but need to be protected in a rigid matrix.

Future Directions

As more and more is learned about the sol-gel process, the greater the number of possibilities for useful optical devices. In the selected examples given above, a key feature is that the sol-gel material is transparent. In addition, the sol-gel material is suitable for application to a variety of substrates. Also, the process is a low-temperature process that is compatible with polymers and organic chromophores. In summary, sol-gel processing has had a number of successes in its application to optical materials, and its use is likely to expand.

See also

Optical Coatings: Thin-Film Optical Coatings.

Further Reading

Andrews MP and Najafi SI (eds) (1997) *Sol-Gel and Polymer Photonic Devices*. Bellingham, CA: SPIE Critical Reviews, CR68-03.

Brinker CJ and Scherer GW (1990) *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing*. Boston, CA: Academic Press.

Dunn B and Zink JI (1991) Optical properties of sol-gel glasses doped with organic molecules. *Journal of Material Chemistry* 1: 903–913.

Klein LC (1993) Sol-gel optical materials. *Annual Reviews of Material Science* 23: 437–452.

Klein LC (ed.) (1994) *Sol-Gel Optics: Processing and Applications*. Dordrecht: Kluwer Academic.

Najafi SI (ed.) (1998) *Selected Papers on Sol-Gel for Photonics*. Bellingham, CA: SPIE Milestones Series, MS148.

Riesfeld R and Jorgenson CK (1992) Optical properties of colorants or luminescent species in sol-gel glasses. *Structure and Bonding* 77: 208–256.

Sol-Gel Optics. SPIE vols. 1328, 1758, 2288 and 3943, 1990–2000. Bellingham, CA: SPIE.

OPTICAL MICROLENSES

H Ottevaere and H Thienpont, Vrije Universiteit Brussel, Brussel, Belgium

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In today's world of information processing the role of optics and opto-electronics is expected to become increasingly important as the performance of communication, processing, sensing, and display technologies is continuously evolving. Making these photonics technologies faster and smaller requires, at the same time, the introduction of massive parallelism and miniaturization. As a result, high-quality, high-precision, and low-cost microlens arrays are becoming indispensable components. Moreover, the introduction of micro-optics provides new degrees of freedom for the system design and, therefore, also for the functionality. It allows solutions which are not possible with purely conventional optics such as the integration of different micro-optical components and systems, hence reducing the additional mechanical assembly cost as required in conventional optics.

Since the beginning of the 1990s, several research groups and industrial research labs have been focusing their attention on the development of fabrication techniques for refractive and diffractive microlenses. Microlenses and microlens arrays are used for beam shaping purposes such as collimation and focusing (e.g., in combination with laserdiode arrays, detector arrays, and fibers), for illumination (e.g., in display systems and projection systems), and for imaging purposes (e.g., in photocopiers).

In this article we will mainly focus on refractive microlenses and we intend to provide an introductory guide to those basic geometrical and optical parameters that are used to characterize refractive microlenses. Indeed, some elementary principles of geometrical optics (the geometrical laws for optical propagation without the inclusion of diffraction) and wave optics (in wave optics the imaging is studied as a result of the propagation of a wavefront to the image plane; diffraction effects can be explained on the basis of this wave theory) are required to understand the characterization and performance evaluation of these lenses. We, therefore, bring forward the basic nomenclature of refractive micro-optics and will highlight how the performance of an optical lens is affected when reducing its physical dimensions.

We start this article by introducing the imaging properties of both thick and thin lenses in the paraxial approximation. In the section on lens aberrations, we discuss the different primary aberrations that influence the imaging properties of a lens. Combined effects of aberrations and diffraction give rise to some frequently used lens performance criteria which we discuss in the section on diffraction. Different aspects have to be considered when discussing the quality of a lens, because each type of microlens application requires specific performances. It is, therefore, not possible to come up with a single figure of merit which satisfactorily evaluates the overall quality of a microlens. Moreover, one needs specific information about the characteristics of the light sources and detectors that are used in the system and also details about the imaging system geometry and the required imaging contrast. We review them in the section on quality criteria.

In the section on interference, we describe these phenomena and focus on two-beam interference that forms the basis of the photonic instrumentation that can be used to characterize discrete microlenses and microlens arrays. For a complete evaluation of a microlens it is essential to measure both its geometrical properties (diameter, sag, surface profile, etc.) and its optical performance (focal length, wave aberrations, etc.). In the final section of this article we introduce three interference-based instruments which are crucial to the full quantitative characterization of refractive microlenses: a noncontact optical profiler based on a Mirau interferometer, a transmission Mach–Zehnder interferometer, and a Twyman–Green interferometer.

Geometrical Optics

Introduction

Within the paradigm of geometrical optics or ray-optics, light travels from the source along straight lines or rays. When a light ray traverses an optical system consisting of several homogeneous media in sequence, the optical path is a chain of straight-line segments. Discontinuities in the line segments occur each time the light is refracted or reflected. The laws of geometrical optics that describe the direction of the rays are described below.

The law of refraction, also called Snell's law

The refraction of a ray at a 'well-polished' surface between two isotropic media, with indices of refraction n_1 and n_2, is given by

$$n_1 \sin \varphi_1 = n_2 \sin \varphi_2 \qquad [1]$$

where φ_1 and φ_2 are the angles between, respectively, the incident and the refracted ray and the normal to the surface.

The law of reflection

When a ray is reflected on a flat 'polished' interface dividing two uniform media, the reflected ray remains within the plane of incidence and the angle of reflection equals the angle of incidence.

Geometrical wavefronts can then be constructed from the optical rays propagating through the system. Locally they are orthogonal to the rays and the optical path lengths along all rays from one wavefront to another are equal. In regions where the wavelength is considered to be negligible, as compared to the dimensions of the relevant components of the optical system, geometrical wavefronts are close approximations of the physical wavefronts.

The Paraxial Approximation

The basic optics of image location and magnification is demonstrated by the application of paraxial or first-order optics, also called Gaussian optics. In the paraxial approximation, only those points and rays will be considered which lie in the immediate vicinity of the optical axis, so that all angles φ are small ($\varphi \leq 15°$). This assumption is true in lens systems with small object and image field sizes and small apertures of pupils. The terms involving squares and higher powers of the off-axis distances or of the angles, which the rays make with the optical axis, will not be discussed here. This leads to:

$$\sin \varphi = \tan \varphi = \varphi + O_{\geq 2}(\varphi) \approx \varphi \qquad [2]$$

With this assumption we can write the refraction law by the linear relation:

$$n_1 \varphi_1 = n_2 \varphi_2 \qquad [3]$$

Refraction at a spherical surface

In Figure 1 a spherical surface with a radius of curvature R forms an interface between two materials with indices of refraction n_1 and n_2. Two rays are shown emanating from an axial object point S. One is an axial ray, normal to the surface of the lens at its vertex V. The vertex of a lens is defined as the point on the refracting surface at the center of its free aperture. This assumes that the aperture is circular

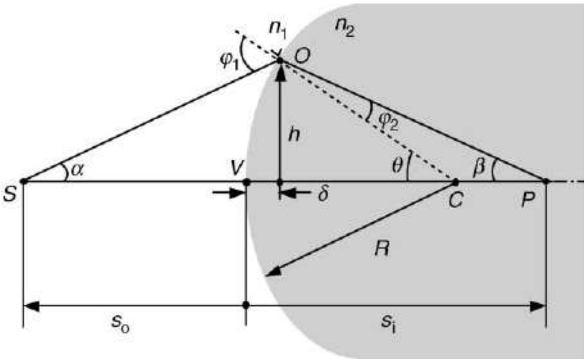

Figure 1 Refraction at a spherical surface.

and centered and the surface is spherical. More generally, if the surface is not spherical but has rotational symmetry, the vertex is the point where the axis of symmetry intersects the optical surface. It is refracted without any change in direction. The other ray is an arbitrary ray incident at O and refracting there according to Snell's Law.

We can rewrite the paraxial refraction law (eqn [3]) and derive a relation between the object and image distances, s_o and s_i, respectively.

From the triangles SOC and POC, we have:

$$\varphi_1 = \alpha + |\theta| \qquad |\theta| = |\beta| + \varphi_2 \tag{4}$$

where $|\theta|$ and $|\beta|$ represent the absolute values of θ and β.

We also have:

$$\tan\alpha = \frac{h}{|s_o| + \delta} \qquad \tan|\beta| = \frac{h}{s_i - \delta}$$
$$\tan|\theta| = \frac{h}{R - \delta} \tag{5}$$

As discussed above, for paraxial rays we may approximate the sine and tangent of an angle by the angle itself and one can show that the small distance δ may be ignored. In accordance with the sign convention the object distance, being to the left of the refracting surface, is to be taken as negative. Inserting eqns [4] and [5] into [3] results in:

$$\frac{n_2}{s_i} - \frac{n_1}{s_o} = \frac{n_2 - n_1}{R} \tag{6}$$

This equation does not contain the angle α, so the image distance s_i is the same for all paraxial rays starting from S.

The spherical surface power $\wp$ is now defined as:

$$\wp \equiv \frac{n_2 - n_1}{R} \tag{7}$$

We can relate this power of the surface $\wp$ to the object and image distance, by combining eqns [6] and [7]:

$$\frac{n_2}{s_i} - \frac{n_1}{s_o} = \wp \tag{8}$$

The distance from the object point to the vertex of the refractive surface is called the front focal length f_{FFL}, if the image is at infinity ($s_i = \infty$). Equations [6] and [8] then give:

$$f_{\mathrm{FFL}} = -\frac{n_1}{n_2 - n_1} R = -\frac{n_1}{\wp} \tag{9}$$

Similarly, if the image is placed at the back focal length f_{BFL}, such that the object distance equals infinity ($s_o = \infty$), then:

$$f_{\mathrm{BFL}} = \frac{n_2}{n_2 - n_1} R = \frac{n_2}{\wp} \tag{10}$$

Thick lenses

A thick lens can be treated as a component consisting of two spherical refracting surfaces separated by a distance d between their vertices V_1 and V_2, as shown in Figure 2. The determination of the focal length as a function of the object and image distance and of the radii of curvature can also be solved using the matrix formalism.

The first and second focal points or the object and image foci, F_o and F_i, can be measured from the two vertices or poles and are called the front and the back focal length denoted by f_{FFL} and f_{BFL}, respectively (Figure 2). The incident and the emerging rays will virtually meet at points that belong to two curved surfaces. These surfaces, approximating planes in the paraxial region, are termed the primary and secondary principal planes, PP_1 and PP_2. Points where the primary and secondary principal planes intersect the optical axis are known as the first and second principal points, H_1 and H_2, respectively.

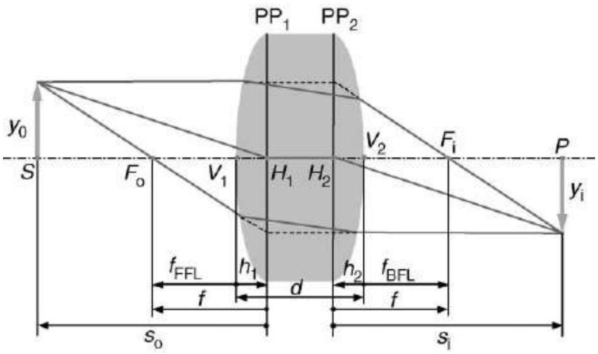

Figure 2 The thick lens geometry.

They provide a set of very useful references from which to measure several of the system parameters. To fully describe a system by its cardinal points, two more points are needed in addition to the two focal points, F_o and F_i and the two principal points, H_1 and H_2. These points are called the nodal points. However, they coincide with the principal points when the lens is surrounded on both sides by the same medium, for example, air. Note that h_1 and h_2 determine the positions of the principal points relative to the vertices V_1 and V_2, while f determines the focal point position relative to the principal points. One can show that these distances are given by:

$$\begin{aligned} h_1 &= \frac{-R_1 n_m d}{n_l(R_2 - R_1) + (n_l - n_m)d} \\ h_2 &= \frac{R_2 n_m d}{n_l(R_2 - R_1) + (n_l - n_m)d} \\ f &= \frac{n_m}{(n_l - n_m)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{d(n_l - n_m)}{n_l R_1 R_2}\right]} \end{aligned} \qquad [11]$$

We now derive the Gaussian formulae that describe imaging through two surfaces which are rotationally symmetric about the same axis. We assume here that the lens, with refractive index n_l, faces the same medium of refractive index n_m on both sides (Figure 3). We can correctly use this hypothesis since the microlenses which we discuss throughout this work are mainly surrounded by air. Let R_1 and R_2 be the radii of curvature of the surfaces at their axial points, measured as positive when the surface is convex towards the incident light.

The image of a given object, formed by refraction at the first surface with radius of curvature R_1, becomes the object for refraction at the second surface. The image formed by the second surface with R_2 as radius is then the final image due to the action of the thick lens. We know from eqn [6] that the paraxial rays issuing from S at a distance s_{o1} from V_1 will meet at P', at a distance which we call s_{i1}. This gives:

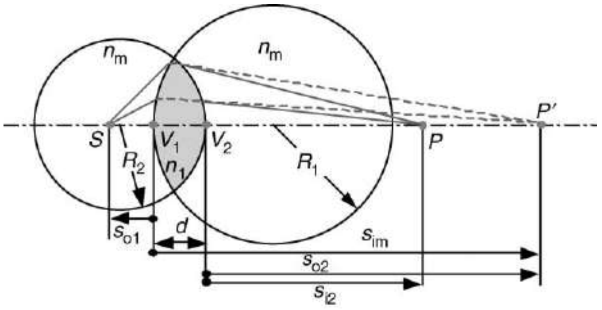

Figure 3 Refraction at both spherical surfaces of a thick lens.

$$\frac{n_l}{s_{i1}} - \frac{n_m}{s_{o1}} = \frac{n_l - n_m}{R_1} \qquad [12]$$

Thus, at the second spherical surface, eqn [6] becomes:

$$\frac{n_m}{s_{i2}} + \frac{n_l}{(s_{i1} - d)} = -\frac{n_m - n_l}{R_2} \qquad [13]$$

As discussed above, we have:

$$\begin{aligned} -\frac{1}{f_{FFL1}} &= \frac{n_l - n_m}{n_m}\frac{1}{R_1} \\ \frac{1}{f_{BFL2}} &= -\frac{n_m - n_l}{n_m}\frac{1}{R_2} \end{aligned} \qquad [14]$$

where f_{FFL1} is the front focal length of the first refracting surface and f_{BFL2} the back focal length of the second refracting surface.

Eliminating s_{i1} in eqns [12] and [13] and inserting [14] yields:

$$d = \frac{n_l}{n_m\left(\frac{1}{s_{o1}} - \frac{1}{f_{FFL1}}\right)} - \frac{n_l}{n_m\left(\frac{1}{f_{BFL2}} - \frac{1}{s_{i2}}\right)} \qquad [15]$$

where s_{o1} is measured with respect to V_1 and s_{i2} is measured with respect to V_2. After some calculations this equation may also be put in the form:

$$\begin{aligned} s_{o1}s_{i2} &- \frac{f_{BFL2}(n_l f_{FFL1} + d n_m)}{n_l(f_{FFL1} + f_{BFL2}) + d n_m} \\ s_{o1} &- \frac{f_{FFL1}(n_l f_{BFL2} + d n_m)}{n_l(f_{FFL1} + f_{BFL2}) + d n_m} \\ s_{i2} &+ \frac{d n_m f_{FFL1} f_{BFL2}}{n_l(f_{FFL1} + f_{BFL2}) + d n_m} = 0 \end{aligned} \qquad [16]$$

Obviously eqn [16] is much too cumbersome to find image locations. However, we can transform and reduce the latter equation to the form:

$$\frac{1}{s_{i2} - h_2} - \frac{1}{s_{o1} - h_1} = \frac{1}{f} \qquad [17]$$

where h_1, h_2 and f are the unknowns.

Indeed, rewriting eqn [17] gives:

$$s_{o1}s_{i2} - (f + h_2)s_{o1} + (f - h_1)s_{i2} + f(h_1 - h_2) + h_1 h_2 = 0 \qquad [18]$$

Equations [16] and [18] will be identical if the coefficients of their corresponding terms are equal;

that is when:

$$f+h_2=\frac{f_{\mathrm{BFL2}}(n_l f_{\mathrm{FFL1}}+dn_{\mathrm{m}})}{n_l(f_{\mathrm{FFL1}}+f_{\mathrm{BFL2}})+dn_{\mathrm{m}}}$$

$$f-h_1=-\frac{f_{\mathrm{FFL1}}(n_l f_{\mathrm{BFL2}}+dn_{\mathrm{m}})}{n_l(f_{\mathrm{FFL1}}+f_{\mathrm{BFL2}})+dn_{\mathrm{m}}} \qquad [19]$$

$$f(h_1-h_2)+h_1h_2=\frac{dn_{\mathrm{m}}f_{\mathrm{FFL1}}f_{\mathrm{BFL2}}}{n_l(f_{\mathrm{FFL1}}+f_{\mathrm{BFL2}})+dn_{\mathrm{m}}}$$

These three equations may be solved for the three unknowns h_1, h_2, and f:

$$f=\frac{n_{\mathrm{m}}}{(n_l-n_{\mathrm{m}})\left[\frac{1}{R_1}-\frac{1}{R_2}+\frac{d(n_l-n_{\mathrm{m}})}{n_lR_1R_2}\right]}$$

$$h_1=\frac{-R_1n_{\mathrm{m}}d}{n_l(R_2-R_1)+(n_l-n_{\mathrm{m}})d} \qquad [20]$$

$$h_2=\frac{R_2n_{\mathrm{m}}d}{n_l(R_2-R_1)+(n_l-n_{\mathrm{m}})d}$$

Thus we see that the relation between the object position and the image position (eqn [16]) is expressed in the more convenient form of eqn [17] with the values of h_1, h_2, and f given by eqn [20]. Here h_1, h_2 and f depend on the constants of the system where, as shown in eqn [11], h_{i} represents the distance between the vertex V_{i} and the principal point H_{i}, and f is the focal length measured from a principal point on the axis and not from the vertex of the lens. If we now assume that:

$$s_{\mathrm{o}}=s_{\mathrm{o}1}-h_1, \qquad s_{\mathrm{i}}=s_{\mathrm{i}2}-h_2 \qquad [21]$$

we can write the thick lens equation as:

$$\frac{1}{s_{\mathrm{i}}}-\frac{1}{s_0}=\frac{1}{f} \qquad [22]$$

where the object and image distance, s_{o} and s_{i}, respectively, are measured from the principal points of the lens. We can conclude that the formula that describes imaging through two surfaces may be simplified to the thick lens formulation given in eqn [22] by choosing the principal points H_{i} instead of the vertices V_{i}, from which S and P are to be measured.

Thin lenses

The preceding eqn [20] takes a particularly simple form when the lens is so thin that the axial thickness of the lens d may be neglected. But obviously there exists no real lens with a thickness equal to zero. A thin lens is only a concept which is an extremely useful tool in optical system design. When a lens or an optical system has a zero thickness, the object and image calculations can be greatly simplified. In this case, the two principal points are coincident and are located where the thin lens is situated.

As a further simplification, we assume that the surrounding medium of the lens is air which means that $n_{\mathrm{m}} \approx 1$. When $n_l = n$, we find:

$$\frac{1}{f}=\frac{1}{s_{\mathrm{i}}}-\frac{1}{s_{\mathrm{o}}}=(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right) \qquad [23]$$

These equations are often called the thin-lens equation and the lensmaker's formula, respectively, because the latter predicts the focal length of a lens, fabricated with a given refractive index and radii of curvature. This formula is very useful in the optical design process as a first-order approach to calculate the features of the thin lenses and results in a set of lens powers (or focal lengths) for the different lenses and in the spacings between them. Once the thin lens optical system is solved and the primary aberrations are corrected for, one can design the thick lens system. The corresponding thick lens system must have the same component powers and the lenses must be spaced apart by the same distances as for the thin lens design, but the thick lens spacings must be measured from the principal points of the lenses instead of from the vertices of the lenses.

Other lens characteristics

Another basic parameter of a refractive lens is its focal number $f\#$. The f-number, also called the speed of the lens, is defined as the ratio of its effective focal length and the lens aperture diameter D (Figure 4):

$$f\#=\frac{f}{D} \qquad [24]$$

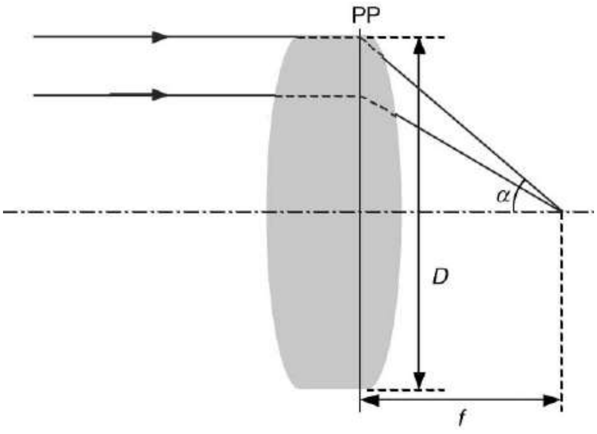

Figure 4 A lens with an aperture diameter D, a focal length f, and a semi-angle α.

The effective focal length f of a lens or often called the focal length is defined from a principal point on the axis of the lens and not from its vertex. Typical focal numbers for refractive microlenses range between 0.8 and 30.

The numerical aperture (NA), is another way of defining the same characteristic. It is described as the sine of the angular semi-aperture α in the object space multiplied by the refractive index n of the object space (Figure 4):

$$\mathrm{NA} = n \sin \alpha \qquad [25]$$

Both the focal number ($f\#$) and the NA are related in the paraxial approximation by

$$f\# = \frac{n}{2\mathrm{NA}} \qquad [26]$$

Lens Aberrations

Introduction to Third-Order Monochromatic Aberrations

An exact ray trace or a measurement on a real lens will certainly disclose inconsistencies with the corresponding paraxial description. Such deviations from the idealized Gaussian optics are known as aberrations. For monochromatic light there are five third-order or Seidel aberrations: three of them, namely spherical aberration, coma, and astigmatism, deteriorate the image by blurring it, while the remaining two, curvature of field and distortion, deform the image. An additional aberration, chromatic aberration, results from the wavelength dependence of the imaging properties of an optical system. We will not discuss the latter as this aberration can be minimized by selecting a lens material with a refractive index that only varies by a negligible amount over the wavelength range of interest.

For a brief discussion of aberrations we consider the case of a perfect lens focusing a collimated light beam. In this case, the lens transforms the incoming planewave into a spherical wave, also called the reference sphere. For a lens with aberrations, however, the real wavefront behind the lens shows deviations from the ideal spherical shape (Figure 5). The optical path length [AB] may be called the wavefront aberrations $\psi(x, y)$ and can be calculated as:

$$\psi(x, y) = W(x, y) - R(x, y) \qquad [27]$$

where $W(x, y)$ represents the spherical wavefront in the Gaussian approximation and $R(x, y)$ depicts the real wavefront. The wave aberration is given by the optical path difference along a ray between the reference sphere and the real wavefront behind the lens. Alternatively, the aberrations can be described in terms of the ray aberrations. In this case, not all the rays intersect the paraxial image plane at the same point and the distribution of the ray intercepts around the paraxial image point can be used to measure the amount of light blur. These two ways of quantitatively describing the aberrations are apparently not independent of each other, the ray aberrations being the derivate of the wavefront aberrations.

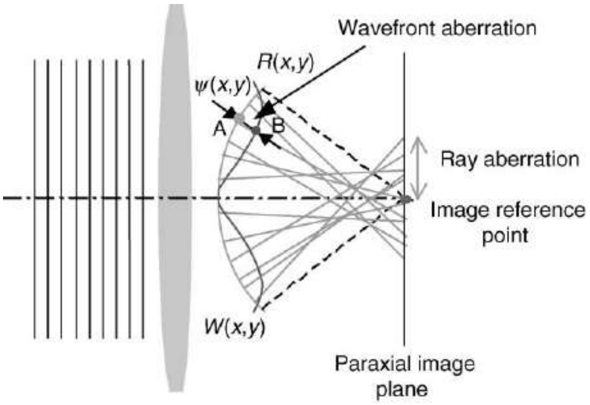

Figure 5 Relation between wavefront and ray aberration.

There are several sets of suitable polynomials over which the wavefront aberrations can be expanded. The Zernike polynomials will be further used in this work (see, for example, the Mach–Zehnder and Twyman–Green interferometers in the section on Microlens testing).

For rotational symmetric optical systems, the most common approach is to expand the aberration function ψ to the fourth order as a power series in polar pupil coordinates. This corresponds to the third-order ray aberrations or primary aberrations. The different terms in the aberration polynomial ψ are to the fourth power of the object coordinate h' and the pupil coordinate r. In that case, the wavefront aberration polynomial ψ, as a function of the object coordinate h' and the polar pupil coordinates r and θ (Figure 6), is given by

$$\psi(h', r, \theta) = {}_0C_{40}r^4 + {}_1C_{31}h'r^3 \cos\theta + {}_2C_{22}h'^2r^2\cos^2\theta + {}_2C_{20}h'^2r^2 + {}_3C_{11}h'^3 r\cos\theta \qquad [28]$$

The ${}_iC_{jk}$ coefficients are subscripted by the numbers that specify the powers of the term dependence on h', r and $\cos\theta$, respectively. These terms comprise the five monochromatic or Seidel aberrations as shown in Table 1.

Each aberration is characterized by its dependence on the deviation from the optical axis h', on the aperture of the refracting surface r, and on the

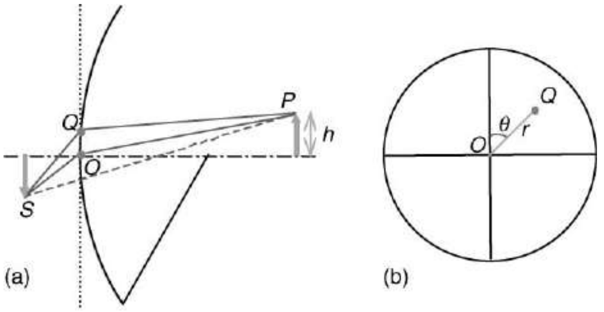

Figure 6 (a) Imaging of an off-axis point *P*; (b) frontal view of a portion of the wavefront.

Table 1 The five monochromatic or Seidel aberrations

r^4	Spherical aberration
$h'r^3 \cos\theta$	Coma
$h'^2 r^2 \cos^2\theta$	Astigmatism
$h'^2 r^2$	Curvature of field
$h'^3 r \cos\theta$	Distortion

symmetry around the optical axis θ. Equation [28] can be used to estimate how the aberration blurs change if the aperture or field of view of a lens is changed. In the following section we briefly describe each of these aberrations in terms of their visual effects.

Spherical Aberration

The first term in eqn [28], ${}_0C_{40}r^4$, is the only term in the third-order wave aberration that does not depend on the deviation h' from the optical axis. Thus, spherical aberration (SA) is the only monochromatic aberration that occurs on the optical axis and that exists even for axial object and image points. As shown in Figure 7a, SA results from the rays that enter via different zones of the aperture and are focused at different distances from the lens. It is evident from Figure 7a that the best focus is not at the paraxial plane. The best focus, however, is found at the location where the circular image blur is the same over the entire field of view and is called the circle of least confusion.

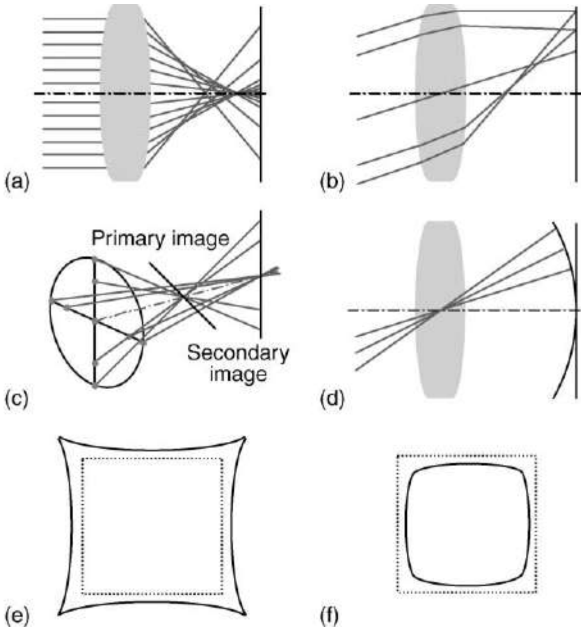

Figure 7 The primary aberrations: (a) spherical aberration – negative SA: the rays through the outer zones of the lens focus closer to the lens than the rays through the central zones; (b) coma – positive coma: the rays through the outer zones of the lens form a larger image than those going through the center; (c) astigmatism – negative astigmatism: the tangential or vertical fan of rays is focused to the left of the horizontal or sagittal fan; (d) curvature of field – negative curvature of field: images farther from the axis focus nearer to the lens than the on-axis images; (e) positive distortion – the magnification increases as the field angle increases; (f) negative distortion.

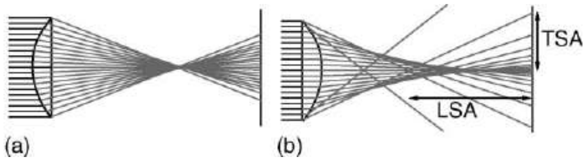

Figure 8 Ray diagram of a planoconvex lens ($D = 200\ \mu$m; $f = 249\ \mu$m) with planewave illumination incident on (a) the curved surface and (b) on the plane surface. LSA (longitudinal SA) represents the difference between the paraxial and the marginal focus while TSA (transversal SA) shows the radius of the spot in the paraxial focal plane.

Because most applications consist of on-axis illuminated planoconvex microlenses, we will firstly focus our attention on how microlenses have to be positioned to minimize SA. When we consider a planoconvex lens with the curved surface towards the infinite conjugate for an incident parallel beam, both surfaces are contributing approximately equal amounts of ray bending. Figure 8a shows the ray diagram for a planoconvex lens as obtained with the ray-tracing program Solstis. Solstis is a commercial photonics design software, consisting of different modules, each of which is dedicated to a certain optical modeling approach. If, however, the flat surface is turned toward the incident rays, then the first surface does not bend the rays and all the ray-bending work is done by the second surface (Figure 8b). We can conclude here that the best orientation is that which distributes the ray bending equally between the two surfaces of the lens.

Because the ray blur varies with the third power of the lens aperture, lowering the aperture diameter will cause a rapid decrease in the extent of the spherical

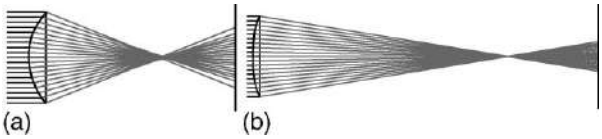

Figure 9 Ray diagram of a planoconvex lens (a) with a focal length *f* and (b) with a focal length *2f*.

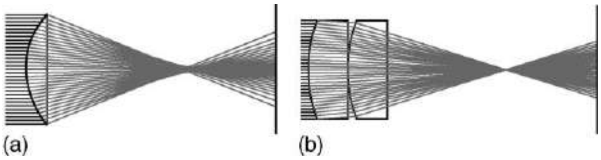

Figure 10 (a) Ray diagram for a planoconvex lens ($D = 200\ \mu\text{m}$; $f = 249\ \mu\text{m}$). (b) The spherical aberration is reduced with a factor 4 by splitting the optical power of the original lens and using a doublet ($D = 200\ \mu\text{m}$; $f = 621\ \mu\text{m}$ for both lenses).

aberration blur. For a given aperture of the lens, the spherical aberration is a function of the focal length of the lens and increases with a decreasing focal length or an increasing NA (Figure 9).

A change in SA, contributing to the wave aberration, for a variation of the lens diameter D or the focal length f can be estimated with the following rules:

$$\text{Spherical aberration} \div D^4 \quad \text{for a constant } f$$

$$\text{Spherical aberration} \div 1/f^3 \quad \text{for a constant } D$$

However, by preserving the focal length and hence the optical power of the lens, the SA can be changed by varying the radii of the surfaces. This would be the same as physically bending a lens made of flexible plastic. In Figure 10 we have reduced the spherical aberration with a factor of four by splitting a lens into two elements, each with a longer focal length, keeping the total optical power of the elements the same as the original lens. This means that, in practice, high NA microlenses can be replaced by a number of microlenses reducing the SA significantly.

However, in optical systems, where only a small number of surfaces are allowed, aspherical surfaces can be used to correct the SA. In Figure 11 we show that making the lens shape elliptical while maintaining the other characteristics of the microlens ($D = 200\ \mu\text{m}$; $R = 150\ \mu\text{m}$) results in a reduction of the SA. The most common form of an aspheric surface is a rotationally symmetric surface with the sag, defined as:

$$z_{\text{conic}} = \frac{R^{-1}(x^2 + y^2)}{1 + \sqrt{1 - (1 + K)R^{-2}(x^2 + y^2)}} \qquad [29]$$

where R is the radius of curvature, K the conic parameter, and x and y the coordinates in the substrate plane. The meaning of K is given in Table 2.

In eqn [28] we have shown that the wavefront aberrations can be written as the sum of the Seidel or third-order aberrations. Furthermore, we have seen that only the SA contributes to the wavefront aberration when we illuminate the microlens on-axis ($h' = 0$). When we limit our experiments to on-axis illumination during the optical characterization of the microlenses, the wavefront aberrations are a measure for the SA. To prove this, we have calculated and plotted in Figure 12, the SA and the wavefront aberrations as a function of the focal length for a 200 μm lens ($n = 1.48$).

Both the SA and wavefront aberration values are obtained with the ray-tracing software Solstis. The wavefront aberration is calculated here as the optical path difference [BA] between a real wavefront surface and a spherical reference surface. From these results we can observe that for microlenses with an $f > 330\ \mu\text{m}$ or an NA lower than 0.3 only the spherical aberration contributes to the root-mean-square (RMS) wavefront aberration. The RMS or standard deviation σ of a number N of data x_i ($i = 1, \ldots, N$) with a mean value X can be written as $\sigma = \sqrt{\sum_{i=1}^{N} (X - x_i)^2 / N}$. For smaller focal length microlenses, however, we notice that higher-order aberrations appear and compensate the SA, resulting in a lower wavefront aberration compared to the SA. We can conclude that the wavefront aberration values represent the optical performance of the microlenses with a main contribution from the SA for on-axis illumination.

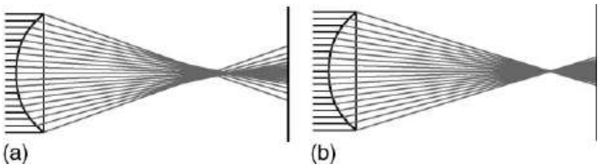

Figure 11 (a) Ray diagram of a spherical planoconvex lens ($D = 200\ \mu\text{m}$; $R = 150\ \mu\text{m}$; $K = 0$) with planewave illumination. (b) The spherical aberration is reduced by making the lens shape elliptical ($K = -0.57$).

Table 2 The surface type for different conic parameters K

Conic constant K	*Surface type*
$K = 0$	Sphere
$K < -1$	Hyperboloid
$K = -1$	Paraboloid
$-1 < K < 0$	Ellipsoid
$K > 0$	Oblate ellipsoid

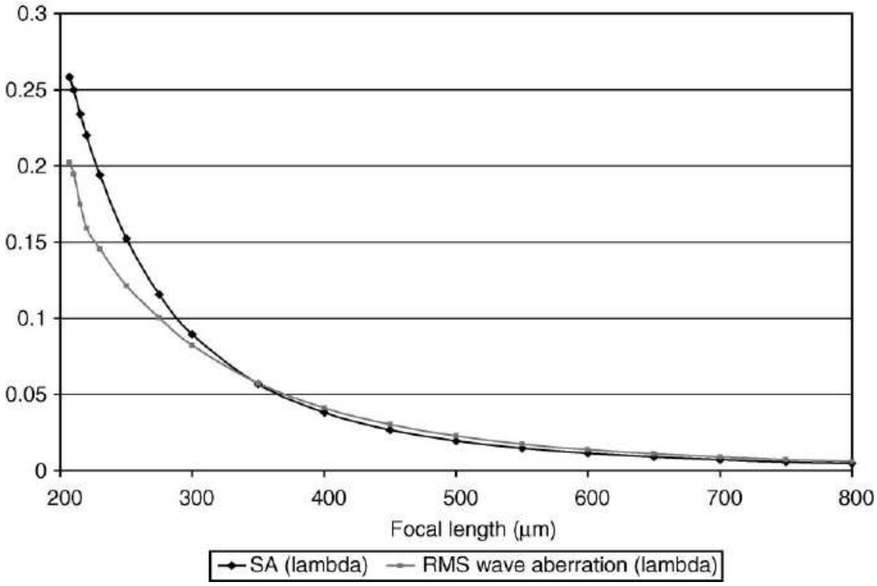

Figure 12 The spherical aberration SA and the RMS wave aberration as a function of the focal length for a 200 µm lens ($n = 1.48$).

Other Seidel Aberrations

Coma is an image-degrading aberration that increases rapidly with the lens aperture r. It indicates an off-axis aberration ($h' \neq 0$) that is nonsymmetrical around the optical axis ($\cos\theta \neq$ constant). Figure 7b illustrates the aberration due to a vertical fan of parallel rays refracted by a spherical lens. It results from a lens having different magnifications for rays passing different zones of its aperture. Each circular zone of the lens forms a circular image called the comatic circle. Every other fan of rays forms images that complete the comatic circle. The combination of all these circles causes a comet-shaped flare.

Astigmatism is the imaging of an off-axis point as two perpendicular lines at different distances. In this aberration the rays from a point object converge at a certain distance from the lens to a line which is called the primary image. This line lies in the sagittal plane and is perpendicular to the plane defined by the optical distance and the object point. At a somewhat different distance from the lens they converge to a second line, called the secondary image, which is parallel to the tangential plane. This effect is shown in Figure 7c. The elimination of astigmatism requires that the tangential and sagittal surfaces be made to coincide. The resulting surface is then called the Petzval surface. Although in that case astigmatism has been eliminated, the associated aberration, called curvature of field, remains. So off-axis images are focused on a curved saucer-shaped surface as drawn in Figure 7d, instead of on an ideal flat image surface.

The last of the five primary aberrations is distortion and causes straight lines which do not intersect the axis to be imaged as curved lines. This effect is not related to a lack of sharpness of the image but results from the fact that the magnification varies across the field and induces the image of a square object to be bowed outward (Figure 7e) or sagged inward (Figure 7f).

We have shown that three of the five Seidel aberrations (namely SA, coma, and astigmatism) are responsible for a lack of sharpness of the image. The other two (i.e., curvature of field and distortion) are related to the position and the form of the image. In general, it is impossible to design a system which is free from all the primary as well as the higher-order aberrations, and therefore, a suitable compromise as to their relative magnitude has to be made. In some cases, the effects of the Seidel aberrations are reduced by balancing them against aberrations of higher orders. In other cases, one has to eliminate certain aberrations completely, even at the price of introducing aberrations of other types. So far we have studied the aberration effects on the basis of geometrical optics only. However, if the aberrations are very small (wave aberrations of the order of a wavelength or less), diffraction starts to play an important role. That is why, in the next section, we will discuss the role of diffraction on the imaging quality of a lens.

Diffraction

Introduction

So far we have used simple geometrical optics to evaluate the effects of aberrations – deviations of the wavefront behind the lens from the ideal shape – that influence the quality of an image generated by a lens. However, we know that even for an aberration-free lens, blur of the spot in the image plane occurs. This means that diffraction effects at the lens aperture can also influence the spot size and the point spread function (PSF), certainly because diffraction gains importance when miniaturizing optical components. Therefore, we have to introduce diffraction before we can bring forward the most important figures of merit for microlens performance and study its effects on imaging quality.

Fraunhofer and Fresnel Diffraction

In certain regions of space occupied by our optical lens system, the simple geometrical model of energy propagation is inadequate. In particular, deviations from this model can be expected in the immediate vicinity of the boundaries of shadows and in regions where a large number of rays meet. These regions are of great practical interest, as they include that part of the image space in which the optical image is situated. These deviations are manifested by the appearance of diffraction fringes. Also, the phenomena involving the superposition of coherent wavefronts cannot be treated with a geometrical model of energy propagation. Thus a more accurate treatment is necessary where the wave nature of light and diffraction are taken into account. Wave optics is such an approach with which the actual light intensity, that will be observed as a result of the propagation of a wavefront to the image plane, can be constructed. According to Huygens' construction, every point of a wavefront may be considered as a center of secondary disturbance, which gives rise to spherical wavelets. The wavefront at any later instance may be regarded as the envelope of these wavelets. Fresnel was able to account for diffraction by supplementing Huygens' construction with the postulate that the secondary wavelets mutually interfere. We discuss diffraction here, on the basis of wave theory, by the application of Huygens' construction together with the principle of interference. The latter principle, also called the Huygens–Fresnel principle, sufficiently describes the propagation of light in free space.

This Huygens–Fresnel principle may be regarded as an approximate form of the integral theorem of Helmholtz and Kirchhoff. Kirchhoff's theory makes an approximation to the diffraction of scalar waves. This scalar theory is adequate enough for the treatment of problems in instrumental optics where polarization effects are of minor importance and can be neglected. If we consider a strictly monochromatic electric field $\varepsilon(x,y,z,t) = E(x,y,z)\cdot e^{-i\omega}$ in any ideal isotropic dielectric, the space-dependent part will satisfy the homogeneous time-independent wave equation: $(\nabla^2 + k^2)E = 0$, also known as the Helmholtz equation where $k = \omega/c$. This theorem expresses the solution E of the Helmholtz equation at an arbitrary field point P, in terms of its solution E and its first derivatives $\partial E/\partial n$ at all points on an arbitrary closed surface surrounding P, provided the medium is a homogeneous dielectric:

$$E(P) = -\frac{1}{4\pi}\oint\!\!\!\oint_S \left[E\overline{\nabla}\left(\frac{e^{-iks}}{s}\right) - \frac{e^{-iks}}{s}\overline{\nabla}E \right]\bar{n}\,dS \qquad [30]$$

Here, s is the distance from point P to the point on surface S, n is the outward normal to S, and $k = \omega/c$ is the wave number in vacuum.

When we consider a monochromatic wave, from a point source P_o propagating through an aperture Ω in a plane opaque screen, and P is the point at which the light disturbance has to be calculated, the theorem can be reduced to an approximate but much simpler form (Figure 13a) on condition that we can assume that the linear dimensions of the aperture Ω, although large compared to the wavelength, are small compared to the distances of both P_o and P from the screen.

As a result of Kirchhoff's theorem we obtain:

$$E(P) = \frac{iA}{2\lambda}\int\!\!\int_\Omega \frac{e^{-ik(r+s)}}{rs}[\cos(n,r) - \cos(n,s)]dS \qquad [31]$$

where A is a constant, λ the wavelength, and r and s are the distances from element dS on Ω to P_o and P, respectively. Equation [31] is known as the Fresnel–Kirchhoff diffraction formula. In case a more general field E_o is incident at the aperture Ω, replacing the point source Ae^{-ikr}/r, an equivalent formulation, known as the Huygens integral, can be found:

$$E(P) = \frac{i}{\lambda}\int\!\!\int_\Omega E_o(x_o,y_o,z_o)\frac{e^{-iks}}{s}\cos(n,s)dS \qquad [32]$$

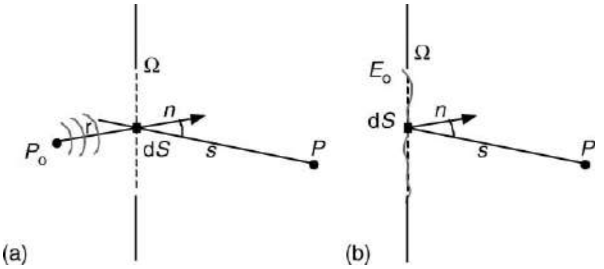

Figure 13 Geometry for evaluating (a) the Fresnel–Kirchhoff diffraction formula and (b) the Huygens' integral.

In the case of the Huygens integral an alternative Green's function is chosen, namely $G(P) = \exp(-\mathrm{j}ks)/s - \exp(-\mathrm{j}kr)/r$ instead of $G(P) = \exp(-\mathrm{j}ks)/s$ used in the Fresnel–Kirchhoff diffraction formula. This then results in an obliquity factor $\cos(n,s)$ instead of $\cos(n,s) - \cos(n,r)$. If the distances of the points P_o and P are large compared to the dimensions of the aperture, the factor $[\cos(n,r) - \cos(n,s)]$ in eqn [31] will not vary appreciably over the aperture and may be replaced by 1. In this case, the optical beams are close to the optical axis and this is considered as the paraxial diffraction theory. In the phase factors, we can expand $r(x,y,z)$ and $s(x,y,z)$ as power series. If the quadratic and higher-order terms may be neglected in the series expansion $(r,s \gg \lambda)$, one speaks of Fraunhofer diffraction, when the previous assumption is not fulfilled $(r,s \geq \lambda)$, the quadratic terms cannot be neglected and one speaks of Fresnel diffraction. The regions where the Fresnel and Fraunhofer diffraction integral may be applied are called the near-field (NF) and far-field (FF) diffraction regions respectively. The Rayleigh range z_R marks the boundary between the NF and FF regions for the beam emerging from the aperture (Figure 14).

As an example, we discuss the diffraction of a uniform planewave by a circular aperture of diameter $2a$. The Rayleigh range for this circular aperture is given by

$$z_R = \frac{\pi a^2}{\lambda} \qquad [33]$$

The FF diffraction pattern $(z \gg z_R)$ for a uniformly illuminated circular aperture is the well-known Airy disk pattern and is given by

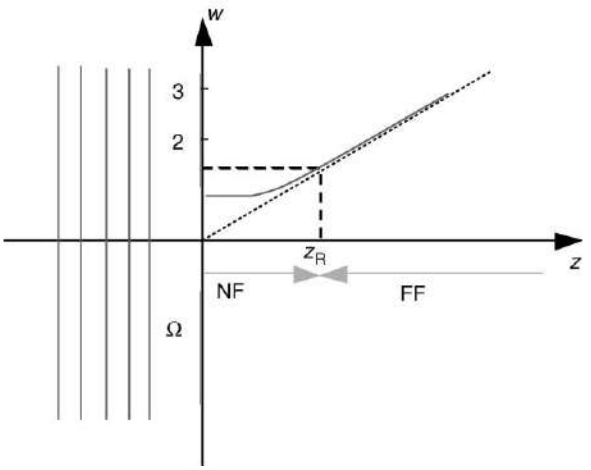

Figure 14 The Rayleigh range z_R marks the boundary between the NF and FF regions for a beam emerging from the aperture Ω.

$$E(r,z) \sim \frac{2J_1(2\pi a\theta/\lambda)}{(2\pi a\theta/\lambda)} \qquad [34]$$

where J_1 is the first-order Bessel function. This pattern has a single dominant central lobe surrounded by a series of increasingly weaker circular rings. The first zero of this pattern occurs at a half angle θ_1 or a radius r_1 in the FF, given by

$$\theta_1 = \frac{r_1}{z} \approx \frac{1.22\lambda}{2a} \qquad [35]$$

The NF or Fresnel diffraction patterns for a uniformly illuminated circular aperture $(z < z_R)$ consist of a series of circular rings modulating a constant-amplitude background. When moving even closer to the aperture, the beam profile becomes more and more square and the frequency of the Fresnel ripples increases.

We now know that light is diffracted at the apertures of optical elements (lenses) and that this diffraction can affect its performance. Ideally, the focus of a planewave should be infinitely small, being the image of a point source located at infinity. But even in the ideal case, when no geometrical aberrations are introduced, diffraction will limit the lens performance and the focus of such diffraction-limited lenses will have a finite extension.

Quality Criteria for Lens Performance

In this section we discuss some of the most important figures of merit for diffraction-limited lens performance. This survey raises the problems related to this issue. A more detailed overview can be found in the literature.

Diffraction in a Lens System: The Point Spread Function (PSF)

Diffraction at the lens aperture D causes a blur of the focus. The light distribution in the focal plane of an aberration free lens essentially arises from Fraunhofer diffraction at the aperture of the lens. From eqn [34] we know that the diffraction image for an aberration-free lens with a circular aperture D is the scaled Airy disk, given by

$$E(r,z) \sim \frac{\pi J_1(rD/\lambda f)}{(rD/\lambda f)} \qquad [36]$$

where λ denotes the wavelength of the illuminating light beam and f the focal length.

The normalized intensity of this pattern is called the diffraction-limited (PSF) of the lens:

$$\mathrm{PSF}(r,z) = |E(r,z)|^2 \qquad [37]$$

Although irradiance replaces nowadays the word intensity in optics, we will still use intensity to nominate the flow of energy per unit area per unit time. We can define the irradiance as $I \equiv \frac{1}{2}\varepsilon_0 c\langle \mathbf{E}^2\rangle$, with ε_0 and c the electric permittivity in vacuum and the speed of light, respectively. The radius of this pattern is known as the Rayleigh diffraction limit or Rayleigh resolution and is compared to the measured spot size in the image plane.

To include the effects of aberrations in the diffraction computation, we have to include both the phase profile and the amplitude of the real wavefront in the pupil of the lens. The pupil function $T(x,y)$ may be written as:

$$T(x,y) = E(x,y)\mathrm{e}^{\mathrm{i}k\psi(x,y)} \qquad [38]$$

where $E(x,y)$ represents the amplitude distribution over the exit pupil and $\psi(x,y)$ is the wavefront deformation on this pupil. Starting from the Huygens–Fresnel approach, the diffraction integral can be rewritten in the form of a Fourier transform:

$$E(r,z) \sim FT(\mathrm{e}^{\mathrm{i}k\psi(x,y)}) \qquad [39]$$

The PSF at the focal position can then be found by calculating the Fourier transform of the wavefront error ψ or the pupil function $T(x,y)$ of the lens and taking the complex square of the latter. Figure 15 shows the shape of the PSF of an ideal or diffraction-limited lens. When the lens aperture D decreases, the extension of the PSF increases proportionally, which means that it becomes easier to achieve diffraction-limited performance. Since the extension of the diffraction-limited PSF becomes larger, the constraints on the shape of the phase profile become less stringent. This has important consequences on the scaling behavior of lenses. In the extreme case, the microlens diameter becomes so small, that diffraction-limited imaging is possible without any specific phase profile in the lens pupil. This is also one of the reasons why small water droplets often form relatively good lenslets. For systems including micro-optic lenses, we therefore will have to find better criteria than the PSF of a diffraction-limited lens to evaluate or compare these optical components.

The Rayleigh and Maréchal Criterion

The widely used Rayleigh criterion states that a system is diffraction-limited, as long as the maximum wavefront deviation $|\psi_{\max}(x)|$ is less than one quarter of a wavelength (Figure 16). It was shown by Rayleigh that for a system which suffers primarily from SA such that the wavefront deviation is less than a quarter wavelength, the intensity at the Gaussian focus is diminished by less than 20%; a loss of light that can usually be tolerated. Also in the presence of other commonly occurring Seidel aberrations is the quality of the image not seriously affected when the deviation is less than λ/4. This criterion, which determines the amount of aberration that can be tolerated in an image-formation system is, of course, only a rough guideline since the light distribution in the image does not only depend on the maximum deformation but also on the shape of the wavefront, and even more on the type of aberration. Moreover, the loss of light that can be tolerated depends on the particular use to which the optical component is put and more stringent tolerances have to be imposed for certain applications.

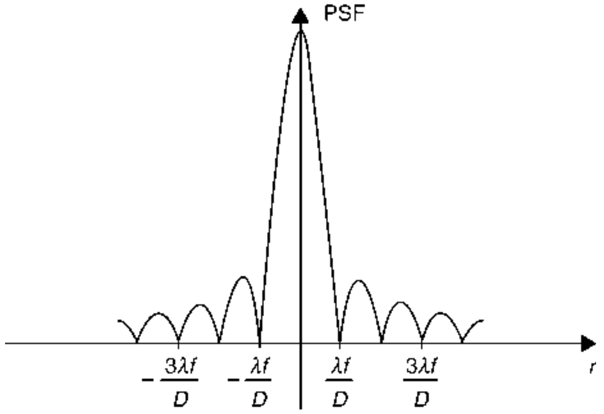

Figure 15 The 1D point spread function of an ideal lens.

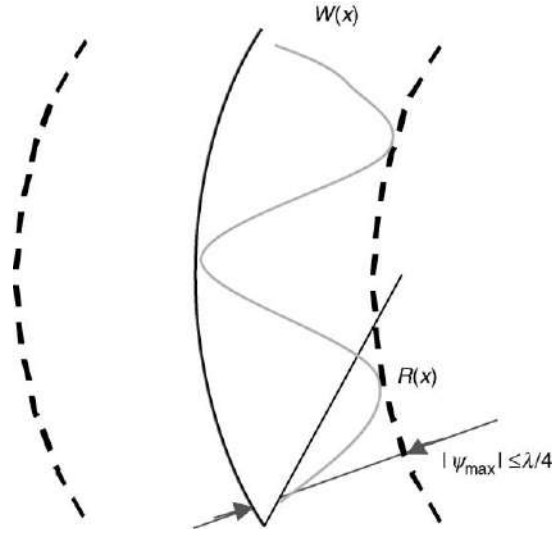

Figure 16 Illustration of the Rayleigh criterion.

Low-order aberrations merely cause a shift of light from the main lobe of the PSF to the side lobes. Higher-order aberrations additionally cause a broadening of the PSF and the correlation between the image quality and the maximum value of the wavefront aberration is usually lost. In the latter case, it seems more appropriate to formulate tolerance criteria restricting the average amount of aberration. Criteria of this type were considered by Maréchal, who used the relation that exists between the intensity at the center of the reference sphere and the RMS deviation of the wavefront from a spherical form.

This Maréchal criterion uses the following RMS wavefront aberration ψ_{RMS} to define the lens quality:

$$\psi_{\mathrm{RMS}} = \sqrt{\int |\psi(x)|^2 \, \mathrm{d}x - \left[\int |\psi(x)| \mathrm{d}x\right]^2} \leq \frac{\lambda}{14} \qquad [40]$$

This criterion is sensitive to statistical noise on the measured phase profile, while a criterion such as the Rayleigh criterion is little affected by it.

Strehl Ratio

For some applications not the imaging quality but the maximum deposited energy is the most essential feature. In such a situation, the Strehl ratio S is a suitable figure of merit. The Strehl ratio is defined as the normalized peak intensity of the PSF of the lens:

$$S = \frac{I_{\mathrm{real}}(0,0)}{I_{\mathrm{ideal}}(0,0)} = \left| \int \int \mathrm{e}^{ik\psi(x,y)} \mathrm{d}x \, \mathrm{d}y \right|^2 \qquad [41]$$

$I_{\mathrm{real}}(0,0)$ and $I_{\mathrm{ideal}}(0,0)$ denote the intensities at the center of the real point image and the ideal PSF without aberrations, respectively. Even for small aberrations, which do not affect the extension of the PSF, the light intensity in the center peak can drop significantly (Figure 17). A Strehl ratio $S \geq 0.8$ is generally considered to correspond to diffraction-limited performance.

By a Taylor expansion of eqn [40], the relation between the Strehl ratio S and the Maréchal criterion can be written as:

$$S = 1 - \left(\frac{2\pi}{\lambda} \psi_{\mathrm{RMS}}\right)^2 \qquad [42]$$

If only low-order aberrations are taken into account, Rayleigh's $\lambda/4$ criterion also results in a Strehl ratio of $S = 0.8$. For large aberrations ($S \leq 0.3$), however, the behavior of the Strehl ratio becomes uncorrelated to the image quality due to interference effects.

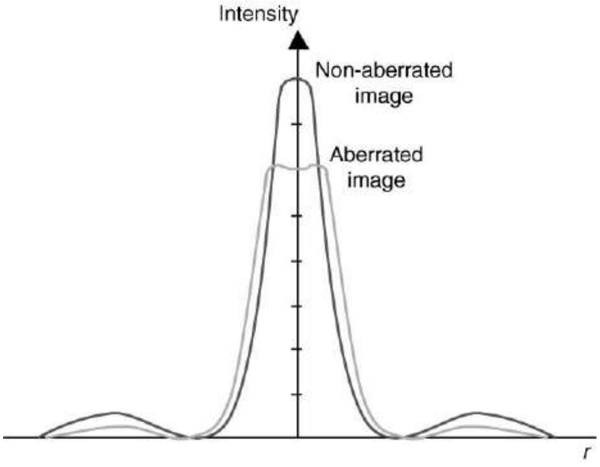

Figure 17 Strehl ratio criterion.

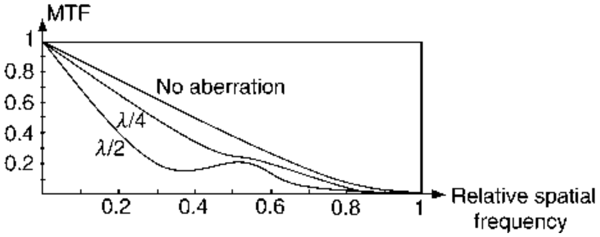

Figure 18 Modulation transfer function for a perfect lens and for small amounts of spherical aberration (λ/2 and λ/4).

Transfer Functions

Transfer functions are most commonly used for evaluating the image quality in lens design. As discussed above, the PSF is the Fourier transform of the wavefront error ψ. For simplicity we can assume that the modulus of the pupil function is 1. The PSF then follows from the Fourier transform:

$$PSF(\mu, \nu) = |\mathrm{FT}(\mathrm{e}^{ik\psi(x,y)})|^2 \qquad [43]$$

The optical transfer function (OTF) is calculated from the PSF data by means of an inverse Fourier transformation:

$$\mathrm{OTF}(\vartheta, \rho) = \mathrm{FT}^{-1}[\mathrm{PSF}(\mu, \nu)] \qquad [44]$$

The OTF is a spatial frequency-dependent complex quantity whose modulus is the modulation transfer function (MTF) and whose phase is, of course, the phase transfer function (PTF). The former is a measure for the reduction in contrast from the object to the image over the whole spectrum. Figure 18 shows the normalized MTF for both a perfect lens and a lens subjected to some small amounts of spherical aberration. The PTF is, in most cases, not so conclusive concerning the optical behavior.

Therefore the wave aberrations, the PSF, the MTF, and the Strehl ratio are the characteristics that are usually provided as an output of lens design programs.

Interference

Many of the processes for the manufacturing of microlenses are not fully understood nor perfectly controlled. It occurs that one generates a microlens array and then measures its performance, hoping that by accurately repeating the processing conditions, a consistent result can be achieved. To achieve optimum conditions, it is therefore crucial to measure the geometrical and optical parameters of the microlenses. Most of the microlens measurement instruments are based on two-beam interference, the phenomena where light from a source is divided into two beams which are later on superposed again. The intensity in the region of superposition is found to vary from point to point between maxima which exceed the sum of the beam intensities and minima which may be zero.

Two-Beam Interference

We consider the interference of two waves, a reference and an object wave, represented by $\mathbf{E}_{\text{ref}}$ and $\mathbf{E}_{\text{obj}}$ respectively, where we take into account the vector properties of the electric fields. In the case of interference, both waves typically originate from a single source and recombine after traveling along different paths. For the moment we will consider only monochromatic linearly polarized waves of the form:

$$\mathbf{E}_{\text{ref}} = \mathbf{E}_{0\text{ref}}\, \mathrm{e}^{\mathrm{i}\Phi_{\text{ref}}(r)-\mathrm{i}\omega t}, \quad \mathbf{E}_{\text{obj}} = \mathbf{E}_{0\text{obj}}\, \mathrm{e}^{\mathrm{i}\Phi_{\text{obj}}(r)-\mathrm{i}\omega t} \qquad [45]$$

The total electric field intensity $\mathbf{E}$ at the output of the interferometer, according to the principle of superposition, is given by:

$$\mathbf{E} = \mathbf{E}_{\text{ref}} + \mathbf{E}_{\text{obj}} \qquad [46]$$

The total intensity at the output, being the time average of the square of the wave amplitude, can be written as

$$I = I_{\text{ref}} + I_{\text{obj}} + I_{\text{ref/obj}} \qquad [47]$$

with $I_{\text{ref}} \sim \langle \mathbf{E}^2_{\text{ref}} \rangle$; $I_{\text{obj}} \sim \langle \mathbf{E}^2_{\text{obj}} \rangle$ and $I_{\text{ref/obj}} \sim 2\langle \mathbf{E}_{\text{ref}} \cdot \mathbf{E}_{\text{obj}} \rangle$. Notice that if both $\mathbf{E}$-fields are perpendicular to one another, the interference term $I_{\text{ref/obj}}$ vanishes, while if we assume that both fields are parallel, we can restrict our discussion to scalar theory. Under the latter condition we can write the total intensity as:

$$I = I_{\text{ref}} + I_{\text{obj}} + 2\sqrt{I_{\text{ref}} I_{\text{obj}}} \cos(\Phi_{\text{obj}} - \Phi_{\text{ref}}) \qquad [48]$$

where the presence of the third term is indicative of the wave nature of light which can produce enhancement or diminution of the intensity through interference.

We can rewrite eqn [48] in Michelson's manner of writing:

$$I = I_0(1 + V\cos(\Phi_{\text{obj}} - \Phi_{\text{ref}}))$$
$$\text{where } I_0 = I_{\text{ref}} + I_{\text{obj}}, \quad V = \frac{2\sqrt{I_{\text{ref}} I_{\text{obj}}}}{I_{\text{ref}} + I_{\text{obj}}} \qquad [49]$$

with V the visibility or fringe contrast indicating the degree of modulation.

We can conclude that the measured intensity of a two-beam interferogram depends on the mean intensity I_0, the visibility V, and the phase difference $\Phi_{\text{obj}} - \Phi_{\text{ref}}$. There are a number of techniques to measure the three unknowns: the mean intensity, the visibility, and the interesting spatial phase variation. In the optical instrumentation, further discussed in this article, phase shifting interferometry (PSI) is used which requires a measurement of the intensity in each point of the aperture for at least three different reference phase values, in order to obtain three equations for the three unknowns (I_0, V, and $\Phi_{\text{obj}} - \Phi_{\text{ref}}$). In practice, the reference phase values are obtained by a piezodriven transducer (PZT) mirror and should be equally spaced over one phase cycle of 2π. It has been shown that the five-phase algorithm is the optimal choice as the calculation still remains simple and the second-order nonlinear response of the detector is eliminated. A charge coupled device (CCD) sensor then records the intensity of the resulting interference pattern for different relative phase shifts and converts the intensity data to wavefront phase data through integration.

Microlens Testing

In microlens (lens with a diameter lower than several millimeters) testing, it is essential to measure in a first step the shape or the surface profile of the microlenses or in cases where gradient-index lenses are concerned, their refractive index distribution. In addition, the paraxial parameters, such as the focal length, and the more complex properties like the wave aberrations and surface quality have to be characterized. As discussed in the introduction, massive parallelism and miniaturization ask for the fabrication of microlenses and microlens arrays. The most critical features of these lenslet arrays are the uniformity of the microlenses and their mutual positioning accuracy. As a variation of these properties deteriorates the optical performance in most applications, dedicated test instrumentation is needed.

In the following, we discuss different measurement instruments designed to characterize all the geometrical and optical parameters of microlenses. First, we introduce an optical noncontact profiler which can be used to determine the diameter and the sag (the maximum height of the spherical lenscap) of the surface profile microlenses. In addition, it allows us to quantify the surface roughness on the top of the microlenses. With a transmission Mach–Zehnder (MZ) interferometer the wave aberrations and the focal length of the microlenses can be measured, while a Twyman–Green (TG) interferometer can be used for the measurement of the surface deviations of a microlens from an ideal spherical shape and for the measurement of the radius of curvature of the surface.

Surface Profile Measurements

Optical noncontact profiler

Principle. The surface profile of a microlens can be measured with a Wyko NT2000 noncontact optical profiler (Figure 19a). This surface profiler system uses two approaches to measure a wide range of surface heights. The PSI mode allows measuring fairly smooth and continuous surfaces ($0.1\ \text{nm} <$ heights < 160 nm) while the vertical scanning interferometry (VSI) mode can measure rough surfaces and heights ranging between 160 nm and 2 mm. We only discuss here the vertical scanning mode because the principle on which PSI is based has been explained above. As shown in Figure 19b, this noncontact profiler is based on a Mirau interference microscope where a white light beam passes through a beamsplitter which reflects half of the incident beam to a reference surface and transmits the other half onto the sample. The light reflected from the sample and from the reference surface then recombine at the beamsplitter to form interference fringes. The system measures the degree of the fringe modulation or the fringe contrast. Because white light has a short coherence length, interference fringes are present only over a very shallow depth for each focus position. The coherence length can be described here as the optical path distance (OPD) between two wavefronts of an interferometric system for which the phase remains correlated. Or in other words it is a measure of the range of heights over which the instrument will be able to obtain measurable interference fringes. Fringe contrast at a single sample point reaches a peak as the sample is translated through focus. During a measurement, the reference arm containing the interferometric objective moves vertically to scan the surface at varying heights. A linearized PZT precisely controls this motion. As the system scans downward, an interference signal for each point on the surface is recorded. Finally, a series of advanced computer algorithms is used to demodulate the envelope of the fringe signal and to extract the surface information.

The vertical resolution is approximately 3 nm RMS for a single measurement. The lateral resolution is a function of the total magnification, of the cursor position accuracy, and of the detector array size.

Measuring the geometrical lens dimensions. This profiler provides us with information on geometrical dimensions of microlenses. For each microlens we can derive from the contour plot (Figure 20) the sag and the diameter of the microlenses along the x and the y direction. As an example, we show in Figure 21, the

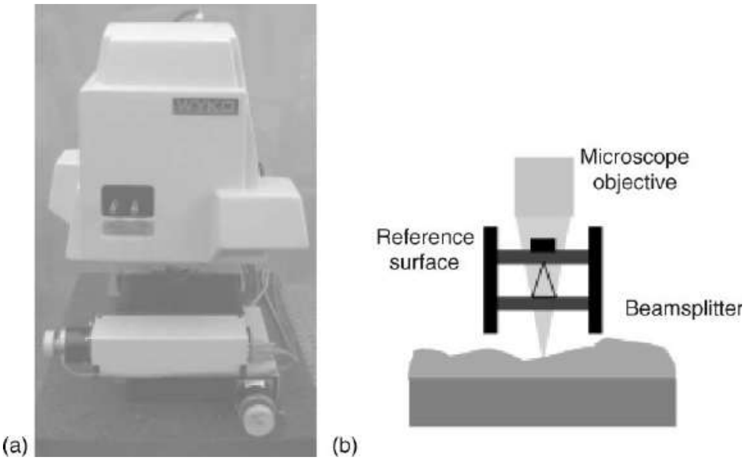

Figure 19 (a) Vertical scanning noncontact profiler (WYKO NT2000); (b) Working principle.

profiles of the cross-section of a 200 μm fused silica microlens ($NA = 0.05$; $D = 200$ μm) along the x and y directions.

The surface roughness on top of the microlens can also be derived from measurements with the noncontact profiler. As an example, we show the fringes on the vertex area (65 × 61 μm^2) of a 200 μm diameter lens as well as the average and RMS roughness on the vertex of the lens (Figure 22).

Measurement of the Optical Lens Performance

To investigate the optical properties of discrete microlenses one of the most accurate and reliable techniques is interferometry. It is well known that microlenses can be tested in transmitted light using instruments with a single pass geometry (e.g., MZ interferometer) as well as in reflected light interferometers using a double pass configuration (e.g., TG interferometer). The interferometers proposed in this section have been especially designed and constructed for microlens testing.

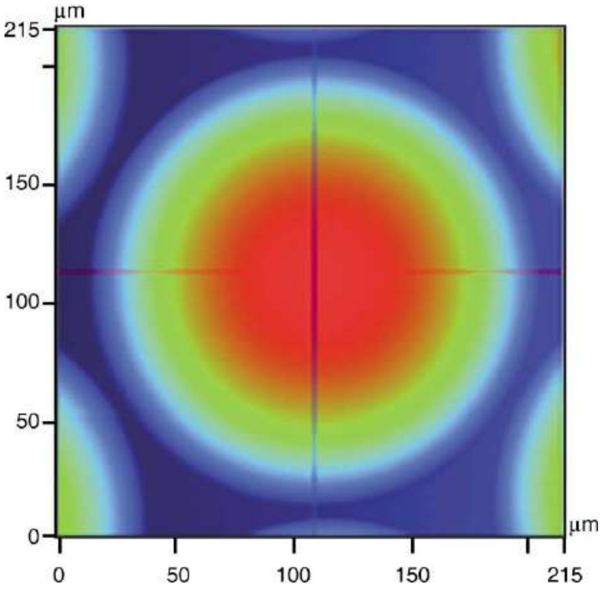

Figure 20 Contour plot of a 200 μm fused silica microlens as obtained with the noncontact profiler.

Mach–Zehnder (MZ) interferometer for transmitted light measurements

Principle. A MZ interferometer, constructed at the Erlangen–Nürnberg University, is used in this work for the measurement of microlenses in transmitted light. The interferometer is installed in a commercial Carl Zeiss Jena microscope, while fiber-optics is used for beam delivery. The scheme is shown in Figure 23. A polarizing beamsplitter (PBS) cube splits the beam of a HeNe laser while a halfwave plate at the entrance enables the balancing of the two interfering beams. The fibers are polarization preserving in order to keep the polarization in the two beams of the interferometer. The planewave of the object arm is either focused by an illumination microscope objective so that a spherical wave impinges onto the microlenses (this is the so-called spherical wave illumination) or the first microscope objective is removed so that a planewave illuminates the object (this is the so-called planewave illumination). By using a telescope consisting of the imaging microscope objective and an achromatic doublet, the expanded planewave then enters a second beamsplitter and is superimposed on a reference planewave generated by the second fiber and an expansion lens. This second beamsplitter is a nonpolarizing one. To force the two beams to interfere, the plane of polarization of the reference beam must be rotated by 90° which is achieved by twisting the fiber end accordingly. This solution has been chosen to avoid a polarizer in front of the CCD camera as it has been shown that the latter would decrease the optical quality. In addition one of the mirrors in the object

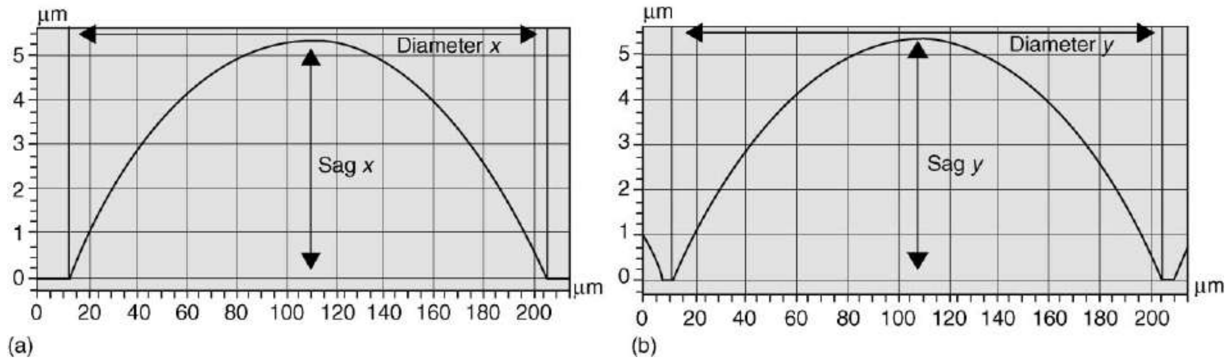

Figure 21 (a) 2D lens profile of the microlens shown in Figure 20 along the x direction and (b) along the y direction.

Roughness values	(nm)
Average roughness R_1	1.55
RMS roughness R_q	1.94

(a)

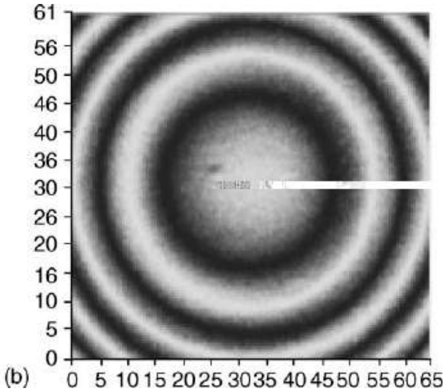
(b)
61
56
50
46
40
36
30
26
20
16
10
5
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65

Figure 22 (a) Different roughness values on the vertex area of a 200 μm lens; (b) Fringes in this vertex area ($65 \times 61\ \mu m^2$).

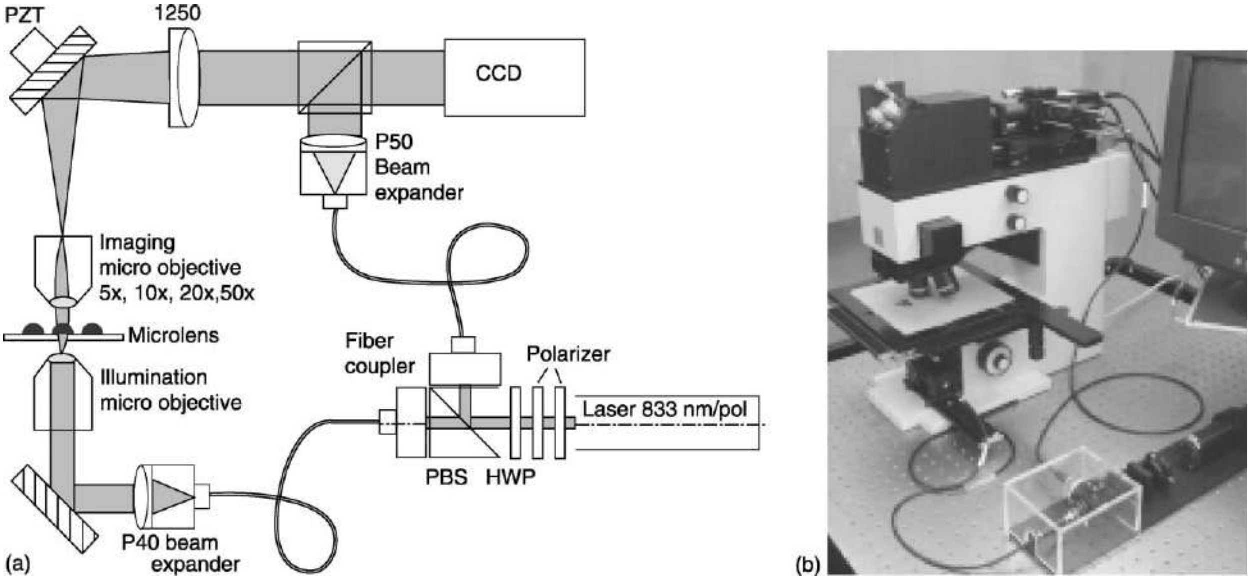
PZT
1250
CCD
P50
Beam
expander
Imaging
micro objective
5x, 10x, 20x,50x
Microlens
Illumination
micro objective
Fiber
coupler
Polarizer
Laser 833 nm/pol
PBS
HWP
P40 beam
expander
(a)
(b)

Figure 23 (a) Scheme and (b) picture of the Mach–Zehnder interferometer for measurements in transmitted light.

arm of the MZ is mounted on a PZT driver to enable phase shifting interferometry.

Focal length and array uniformity through plane-wave illumination. The MZ interferometer with planar wavefront illumination offers for slow lenses (NA < 0.1) the opportunity to measure also the total phase lag introduced by the lenses, as shown in Figure 24. This makes it possible to determine the focal length and hence the NA of the microlenses.

For certain applications of arrays of microlenses it is also important to know the uniformity of the lens array or at least of a part of the lens array. There are several different test methods (e.g., Smartt test, shearing test) which can provide a measure of the uniformity. One can even obtain the total phase retardation for a subsection of the microlens array (see Figure 25b).

Measuring the aberrations with spherical wave illumination. Wave aberrations of refractive microlenses can be determined with a spherical wavefront illumination (Figure 26). The measurement of the wave aberrations is carried out by using a PSI algorithm. This requires a measurement of the intensity in each point of the aperture for five different reference phase values. The result of a PSI evaluation is the phase distribution $\Phi(x, y)$ in the exit pupil:

$$\Phi(x, y) = \frac{2\pi}{\lambda}\psi(x, y) \qquad [50]$$

where $\psi(x, y)$ are the wave aberrations of the lens under test.

From these data, the software calculates the unwrapped phase distribution and by subtracting the tilt and defocus, the deviation from an ideal wavefront (phase/lambda) can be found. Fitting this deviation to a fourth degree Zernike polynomial then gives the aberrations (phase/lambda). Besides the aberrations we also derive from the measured phase distribution other optical lens characteristics such as the modulation transfer function and the PSF.

Finally, we emphasise that the measurements should preferably be done in a cleanroom environment as the smallest dust particle on the lens surface decreases the optical performance.

Twyman–Green interferometer

Principle. For most applications microlenses have to feature a spherical or nearly spherical shape. The measurement of the deviations of the microlens surface from an ideal sphere is therefore very important. The basic principle comes down to a planewave which is focused by a high NA microscope objective in such a way that the focus coincides with the center of curvature of the microsphere under test. The light rays are reflected, and behind the microscope objective a nearly planewave carrying the surface deviations and misalignment errors, results. The vertex of the test surface is then imaged onto a CCD camera and the nearly plane object wave interferes with the plane reference wave. The evaluation here is also done with phase shifting interferometry by axially shifting the reference mirror with a piezo transducer.

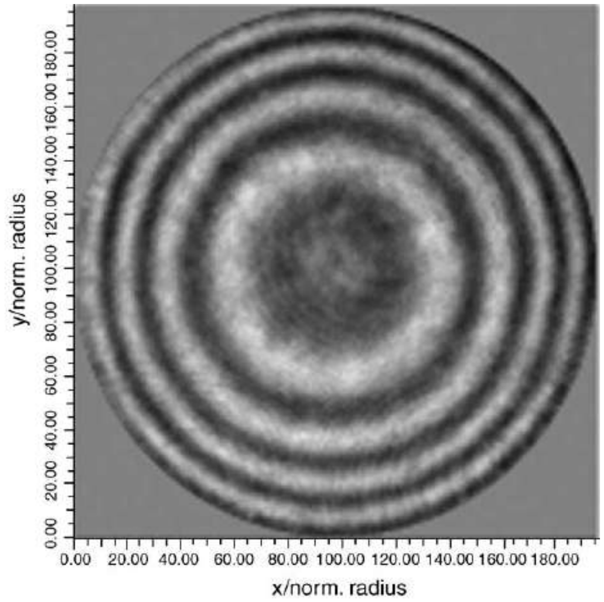

Figure 24 Interferogram of a 200 μm lens with an NA of 0.05 under planewave illumination.

The use of laser illumination allows us to obtain high contrast fringes. However, the high spatial and temporal coherence of the laser light also has important drawbacks, such as speckle noise and spurious fringes across the field. It is convenient to divide coherence into two classifications, temporal and spatial. The former relates directly to the finite bandwidth of the source, the latter to its finite extent in space. Therefore, polarization beamsplitter optics is commonly used to reduce the contribution of spurious reflections. The polarizing beamsplitter group additionally allows balancing the intensity between the object and the reference arm of the interferometer. Even so, the remaining scatter leads to poor and rough interference fringes. To further increase the fringe quality, the spatial coherence has to be reduced while the temporal coherence of the laser is preserved. This can be done by introducing a rotating scatterer in the light beam which gives a smoothing effect of the interference fringes.

Measurement of the deviations from sphericity. To determine the surface deviation of a microlens from a perfect sphere we make use of a TG

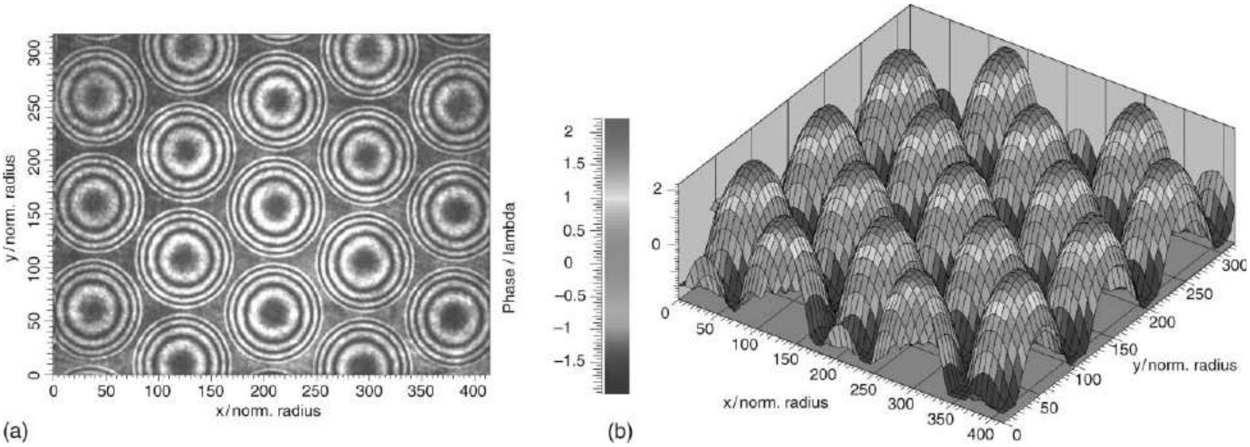

Figure 25 (a) Interferogram of an array of 200 μm diameter microlenses with an average NA of 0.05 and illuminated with a plane wavefront; (b) 3D representation of the phase lag (RMS 1.26λ, *PV* 4.37λ) for the same microlens array.

interferometer, constructed by the Erlangen-Nürnberg University (Figure 27). In this interferometer, the light of a HeNe laser is imaged onto a rotating scatterer by a microscope objective and the scattered light is collimated by an achromat. The collimated laser beam then passes a field stop which is sharply imaged by a 4f system onto the lens under test. The image of the field stop can be focused onto the surface under test by moving the field stop along the optical axis. Once the image of the field stop is sharply imaged onto the CCD array, one can translate the reference mirror until the reference image of this stop in its turn becomes sharp onto the CCD camera. Thus, two conditions have to be met to achieve high contrast fringes: the surface under test and the image of the field stop of both arms of the interferometer have to be sharp on the detector. The half wave plate in front of the PBS adjusts the splitting ratio at the PBS while the polarizer after the beamsplitter is necessary to force the two orthogonally polarized beams to interfere. Finally, the optics at the output of the interferometer gives a sharp image of the intermediate image of the lens surface onto the CCD camera. Analogous to the MZ interferometer, five phase images are grabbed and subjected to different algorithms, leading to the desired information.

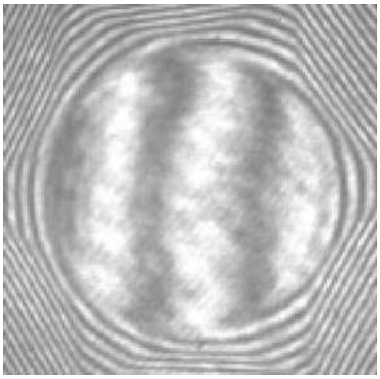

Figure 26 Interferogram of a microlens under spherical wavefront illumination: the resulting planewave is a measure for the wave aberrations.

A condensor objective transforms the radius of curvature of the spherical wavefront such that it matches the radius of curvature of the microlens. Hence, the wavefront impinges everywhere perpendicularly on the lens surface. This means that the sensitivity to surface deviations is independent on the surface position leading to high accuracy measurements. The TG interferometer allows us to test the sphericity of microlenses with diameters ranging

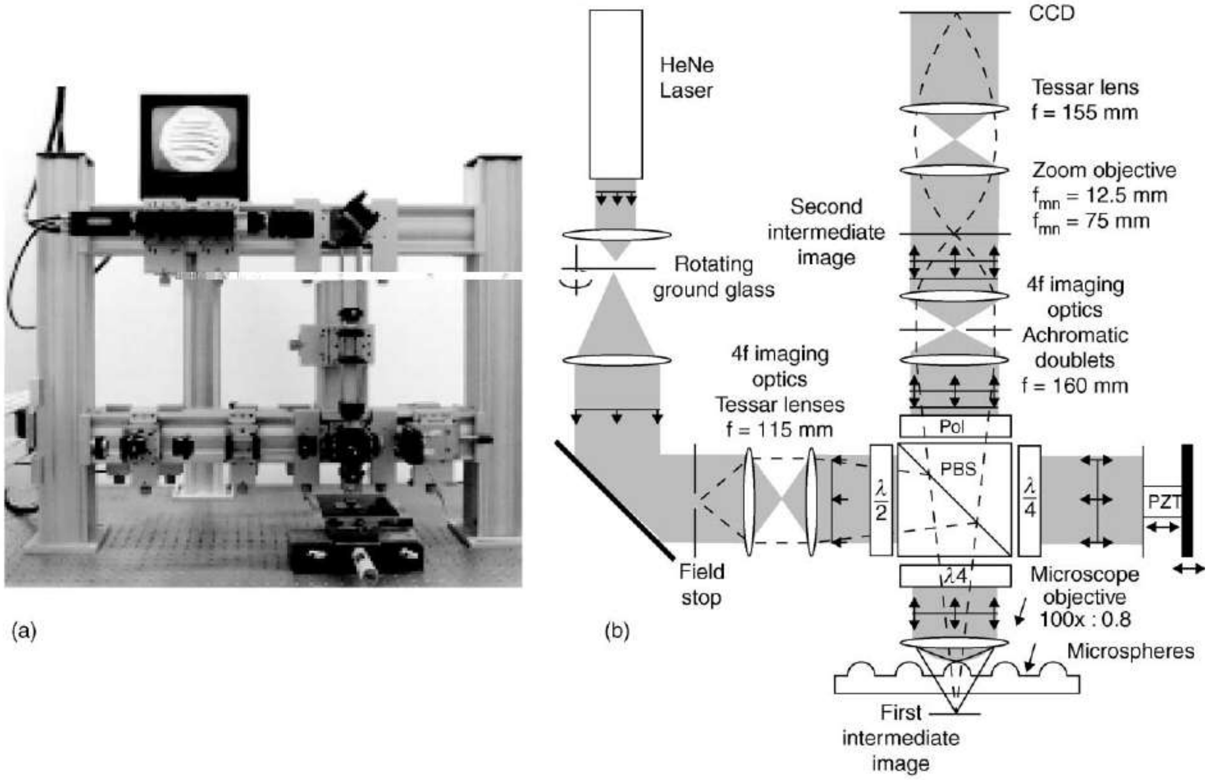
(a)
(b)
HeNe Laser
Rotating ground glass
4f imaging optics Tessar lenses f = 115 mm
Field stop
Second intermediate image
CCD
Tessar lens f = 155 mm
Zoom objective f_{mn} = 12.5 mm f_{mn} = 75 mm
4f imaging optics
Achromatic doublets f = 160 mm
Pol
PBS
$\lambda/2$
$\lambda/4$
PZT
$\lambda/4$
Microscope objective 100x : 0.8
Microspheres
First intermediate image

Figure 27 (a) Twyman–Green interferometer setup. (b) Detailed schematic overview of this interferometer.

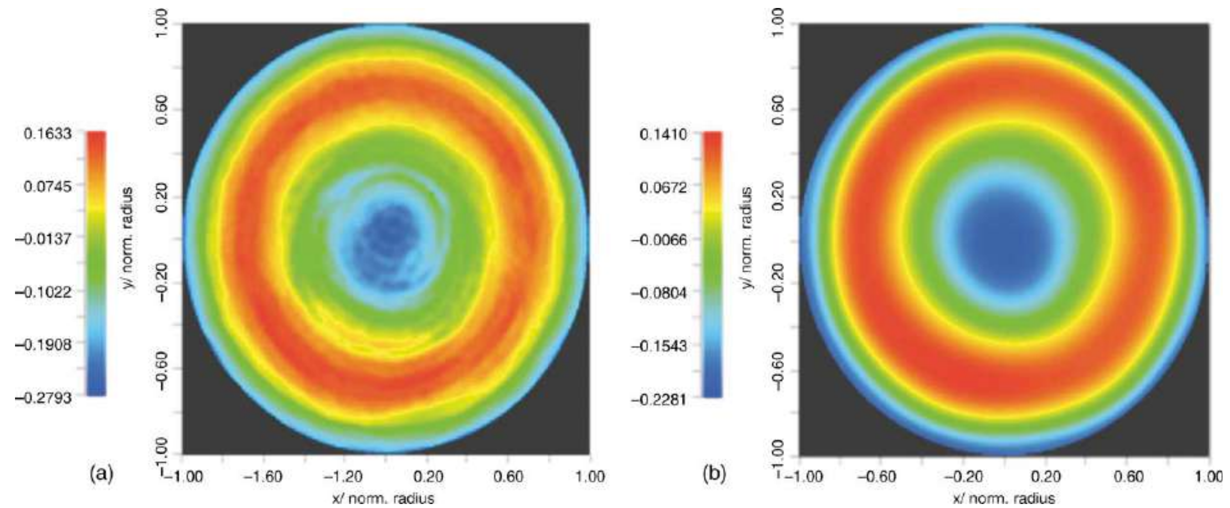

Figure 28 (a) Measured deviation from an ideal sphere (phase/lambda) for a 200 µm lens with an NA of 0.05 as obtained with a Twyman–Green interferometer. (b) Polynomial fitted surface deviation (phase/lambda) of this microlens.

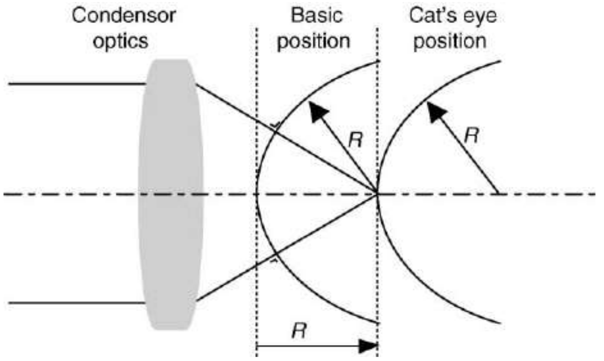

Figure 29 Necessary adjustments of the spherical microlens for radius measurements.

from 50 to 1000 µm and with a maximum radius of curvature of 1960 µm.

An example of a 200 µm fused silica microlens, with an NA of 0.05, is shown in Figure 28. The deviation from an ideal sphere can be derived from the unwrapped phase after subtracting the tilt and defocus. Fitting the deviation to a fourth degree Zernike polynomial then gives the surface deviation.

Measuring the radius of curvature. Using the TG interferometer it is also possible to determine the radius of curvature R of a microlens. To do this, the test object has to be shifted axially from the basic position, where the light rays impinge perpendicularly onto the surface and the focal length of the condensor objective coincides with the center of the sphere under test, to the so-called cat's eye position, where the focus of the impinging spherical wave is at the vertex of the sphere (Figure 29). For both positions straight, parallel, and equidistant interference fringes are obtained, only perturbed by aberrations due to the lens surface and the condensor objective. This forms the criterion for the adjustment of the two positions. The difference between these two positions then gives the radius of the microsphere. It should be mentioned that the cat's eye position causes an inversion of the object wavefront with respect to the reference wavefront, which results in a vanishing contrast with the partially coherent illumination. The latter can only be avoided by providing the necessary coherence. This implies that the rotating scatterer will have to be removed.

Summary of the Interferometric Instrumentation

Figure 30 summarizes the various instruments we discussed in this section: a noncontact optical profiler, a MZ and a TG interferometer. For each of these instruments, we highlight the measurands, the assumptions that are made, and the resulting optical lens characteristics. From this figure, it is clear that one can obtain a measure for the focal length and hence the numerical aperture and the focal number of a refractive microlens, with all three instruments. However, only the focal length obtained with the MZ interferometer is a direct measurement. The focal length obtained with the optical profiler is derived from the sag of the lens under the assumption of an ideal sphere while the radius of curvature measured with the TG interferometer allows to calculate the focal length as well. Nevertheless, the three values make an interesting comparison possible.

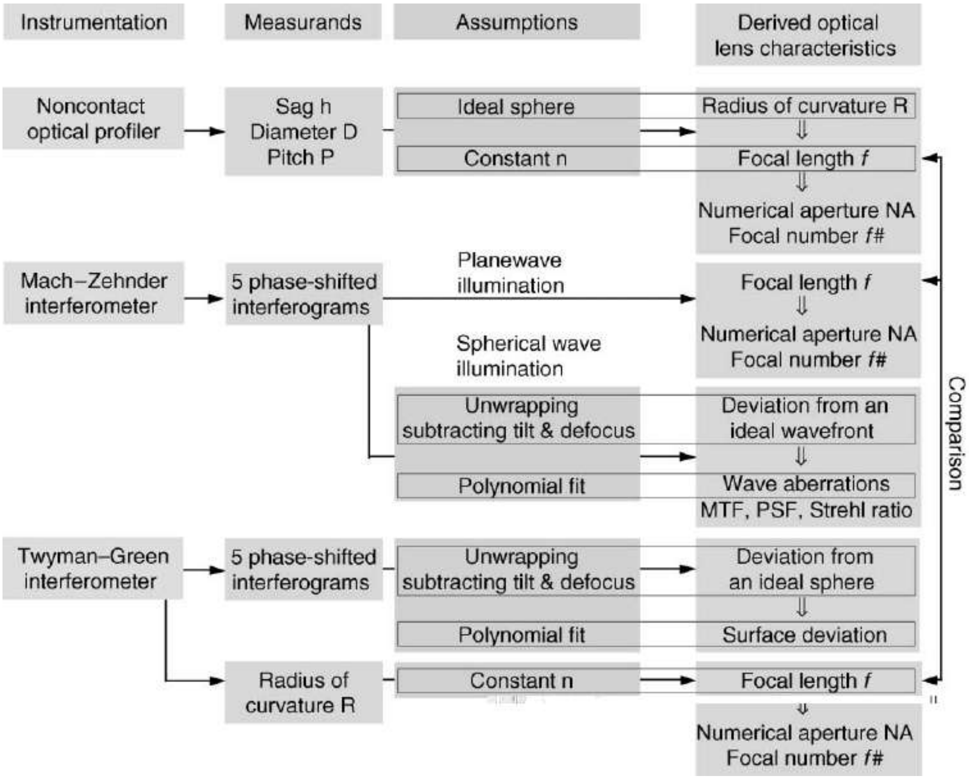
Instrumentation
Measurands
Assumptions
Derived optical lens characteristics
Noncontact optical profiler
Sag h
Diameter D
Pitch P
Ideal sphere
Constant n
Radius of curvature R
Focal length *f*
Numerical aperture NA
Focal number *f#*
Mach–Zehnder interferometer
5 phase-shifted interferograms
Planewave illumination
Focal length *f*
Numerical aperture NA
Focal number *f#*
Spherical wave illumination
Unwrapping subtracting tilt & defocus
Polynomial fit
Deviation from an ideal wavefront
Wave aberrations
MTF, PSF, Strehl ratio
Comparison
Twyman–Green interferometer
5 phase-shifted interferograms
Unwrapping subtracting tilt & defocus
Polynomial fit
Deviation from an ideal sphere
Surface deviation
Radius of curvature R
Constant n
Focal length *f*
Numerical aperture NA
Focal number *f#*

Figure 30 Overview of the instruments that we will use to quantitatively characterize the refractive microlenses: the measurands and the resulting optical lens characteristics.

Conclusions

In this article we introduced geometrical optics to discuss the imaging properties and the aberrations of refractive lenses, while wave optics was used to explain the effects of diffraction on the imaging properties of lenses. Next we put forward some commonly used quality criteria to assess lens performance and we pointed out the unique properties of microlenses. Finally, interference was briefly introduced to clarify the working principle of various instruments to quantitatively characterize refractive microlenses and microlens arrays: a noncontact profiler, a transmission MZ interferometer and a TG interferometer.

We can conclude that micro-optics is a key technology with many applications in various areas of tomorrow's information society. It enables new functionalities and applications previously unachievable with conventional optics and improves product performance, stability and robustness while reducing cost, volume, and weight. While in conventional optics discrete lenses are mounted to form optical systems using mechanical alignment equipment, for microlenses new integration approaches are available. For most of the applications, the integration of micro-optical components into fully functional systems is very important.

Thus micro-optics is a generic technology that allows the manipulation of light with micron and submicron scale structures and components. Micro-optics is therefore the technology to help provide the necessary tools to interface the macroscopic world we live in with the microscopic world of opto- and nano-electronic data processing. Micro-optic processing and fabrication technologies can be adapted to a wide variety of specialty optical materials such as plastics, polymers, semiconductors, sol-gels, and doped glasses. The application of these modified materials permits the widening of the field of photonic functionalities to virtually any region of the optical spectrum and beyond, such that a wealth of novel optical and photonic applications can be realized. Micro-optic processing technologies can be made compatible with scalable industrial mass-manufacturing and replication techniques.

Acknowledgment

This research has been made possible through the financial support of the Fund for Scientific Research-Flanders (FWO-Vlaanderen), the IAP Photon

Network, the Bilateral Scientific and Technological Cooperation Flanders-Poland, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the IWT GBOU project ‘Generic Technologies for Plastic Photonics’, the Research Council (OZR) of the Vrije Universiteit Brussel and the EC Network of Excellence NEMO.

List of Units and Nomenclature

[AB]	optical path length
c	speed of light
d	lens thickness
D	lens aperture diameter
$\mathbf{E}$	electricfield
f	focal length
f_{BFL}	back focal length
f_{FFL}	front focal length
$f\#$	focal number
F_{i}	image focal point
F_o	object focal point
h_1	position of the first principal point relative to the vertex V1
h_2	position of the second principal point relative to the vertex V2
h'	deviation from the optical axis
H_1	first principal point
H_2	second principal point
I	intensity
k	wave number in vacuum
n	index of refraction
n_{i}	index of refraction of medium i
O	object point
P	arbitrary field point
P	image point
P	lens pitch
$\wp$	spherical surface power
r	aperture of the refracting surface
R	radius of curvature
$R(x, y)$	real wavefront
S	axial object point
S	Strehl ratio
s_0	object distance
s_i	image distance
V	lens vertex
V	visibility or fringe contrast
$W(x, y)$	spherical wavefront in the Gaussian approximation
y_{o}	object height
y_{i}	image height
z_{R}	Rayleigh range
α	angular semi-aperture
θ	symmetry around the optical axis
λ	wavelength of light
σ	standard deviation
φ_1	angle between the incident ray and the normal to the flat surface
φ_2	angle between the refracted ray and the normal to the flat surface
$\Phi_{\mathrm{obj}} - \Phi_{\mathrm{ref}}$	phase difference
$\Phi(x, y)$	phase distribution
$\psi(x, y)$	wavefront aberration
ω	angular frequency
Ω	aperture in a plane opaque screen
BFL	Back focal length
CCD	Charge coupled device
FF	Far-field
FFL	Front focal length
MTF	Modulation transfer function
MZ	Mach–Zehnder
NA	Numerical aperture
NF	Near-field
OPD	Optical path difference
OTF	Optical transfer function
PBS	Polarizing beamsplitter
PP	Principal plane
PSF	Point spread function
PSI	Phase shifting interferometry
PTF	Phase transfer function
PZT	Piezo-electric transducer
RMS	Root mean square
SA	Spherical aberration
VSI	Vertical scanning interferometry

See also

Diffraction: Fraunhofer Diffraction. **Diffractive Systems:** Aberration Correction with Diffractive Elements. **Geometrical Optics:** Aberrations; Lenses and Mirrors. **Interferometry:** Overview; Phase Measurement Interferometry.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge, UK: Cambridge University Press.

Hecht E (1987) *Optics*. San Francisco, CA: Addison-Wesley Publishing Company.

Herzig HP (1997) *Micro-Optics: Elements, Systems and Applications*. London: Taylor & Francis.

Jenkins FA and White HE (1981) *Fundamentals of Optics*, 4th edn. New York: McGraw-Hill.

Malacara D and Malacara Z (1994) *Handbook of Lens Design*. New York: Marcel Dekker.

Malacara D, Servin M and Malacara Z (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Dekker.

Mouroulis P and Macdonald J (1997) *Geometrical Optics and Optical Design*. New York: Oxford University Press.

Pedrotti FL and Pedrotti LS (1993) *Introduction to Optics*, 2nd edn. Englewood Cliffs: Prentice-Hall International.

Shannon RR (1997) *The Art and Science of Optical Design*. Cambridge, UK: Cambridge University Press.

Siegman AE (1986) *Lasers*. Sausalito: University Science Books.

Sinzinger S and Jahns J (1999) *Microoptics*. Weinheim, Germany: Wiley-VCH GmbH.

Smith WJ (1997) *Practical Optical System Layout and Use of Stock Lenses*. New York: McGraw-Hill.

Smith W (2002) Engineering an optical system. *SPIE OE Magazine* 7: 49–54.

Young HD (1992) *Physics*, 8th edn. San Francisco, CA: Addison-Wesley Publishing Company.

OPTICAL PARAMETRIC DEVICES

Contents

Overview
Optical Parametric Oscillators (Continuous Wave)
Optical Parametric Oscillators (Pulsed)

Overview

B J Orr, Macquarie University, Sydney, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many forms of optical parametric device (OPD) play significant roles as three-wave nonlinear-optical sources of tunable coherent light for laser-based science and technology. Any such device will fall into one of three categories, in increasing order of sophistication: optical parametric generator (OPG); optical parametric amplifier (OPA); and optical parametric oscillator (OPO). This article concentrates largely on nanosecond-pulsed OPOs, with particular emphasis on designs that facilitate their wavelength control and their optical power- or energy-efficiency. The two other important temporal manifestations of OPOs (ultra-fast pulsed and continuous-wave) are briefly outlined. OPDs serve as versatile sources of tunable, intense, coherent radiation covering the infrared, visible, or ultraviolet spectral regions. Their most obvious uses are associated with laser wavelength-conversion and/or tunability, resulting in a variety of spectroscopic applications. Four-wave nonlinear-optical parametric contributions, such as those involved in stimulated Raman scattering and highly nonlinear optical fibers, are also briefly considered.

What is an Optical Parametric Device?

An optical parametric device (OPD) is a source of coherent, laser-like radiation that is typically intense and tunable over a wide range of wavelengths. Its mechanism invariably depends on nonlinear optics, usually through a three-wave mixing process mediated by the nonlinear-optical susceptibility $\chi^{(2)}$ in a noncentrosymmetric crystalline medium. Three forms of OPD are illustrated in Figure 1, and also illustrated is a closely related (but distinct) nonlinear-optical device. Coherent light waves are represented by arrows, with their associated optical frequency ω_j and wave vector $\mathbf{k}_j$ (which are defined below). In Figure 1, input and output waves are shown as arrows on the left and right, respectively, with their breadth intended to indicate typical relative intensities.

An optical parametric generator (OPG) is the simplest form of OPD, with a single input wave (pump, P) from a laser generating two coherent output waves (signal, S; idler, I; $\omega_S \geq \omega_I$), as depicted in Figure 1a. The nonlinear-optical process itself is initiated by low-intensity parametric spontaneous emission/noise/fluorescence processes that occur naturally, effectively 'splitting' a pump photon into two new photons. Once a signal and/or idler wave has been generated, it can be coherently amplified by passing it through an optical parametric amplifier (OPA) together with input pump radiation, as depicted in Figure 1b. A further order of sophistication is reached in an optical parametric oscillator (OPO), as depicted in Figure 1c, where the

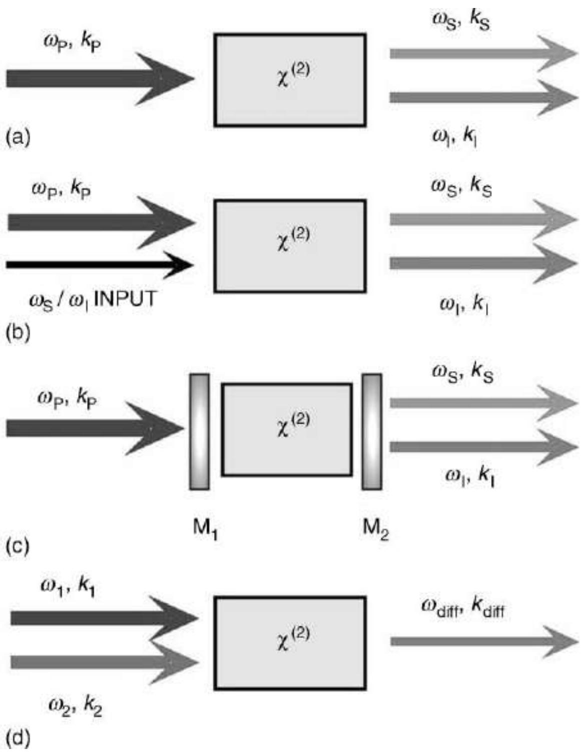

Figure 1 Schematic diagrams of three forms of optical parametric device (OPD): (a) optical parametric generator (OPG); (b) optical parametric amplifier (OPA); (c) optical parametric oscillator (OPO). Note that, by convention, optical frequencies of the signal (S) and idler (I) output waves are defined such that $\omega_S \geq \omega_I$. Also shown is a fourth closely related device: (d) difference-frequency generator (DFG). Nonlinear-optical media are denoted by their susceptibility $\chi^{(2)}$. Arrows are used to represent input and output waves, together with their optical frequencies ω_j and wave vectors $\mathbf{k}_j$. An OPO requires an optical resonator, comprising at least two aligned reflectors (M_1, M_2).

functions of an OPG and an OPA are combined by multipassing the optical waves involved inside a resonant optical cavity, formed by two or more appropriately aligned reflectors (M_1, M_2).

A difference frequency generator (DFG), as depicted in Figure 1d, is not an OPD, although the DFG source term is central to the nonlinear-optical mechanism of an OPD. In a DFG, two input waves (labeled 1 and 2) are combined coherently to generate a third output wave at the difference frequency of the two input waves: there are now two relatively high-power driving waves, rather than one, and photons are effectively ‘joined’, rather than ‘split’ as in the OPD case. Nevertheless, the outcome and utility of a DFG can be similar to that of an OPD. For instance, if coherent radiation is required at a particular infrared (IR) wavelength, it can be generated either as the idler wave of an OPD, with frequency $\omega_I = (\omega_P - \omega_S)$ *or* as the output wave of a DFG, with frequency $\omega_3 = |\omega_1 - \omega_2|$.

Many of the desirable attributes of OPDs in general, and OPOs in particular, arise from the fact that any such instrument is a nonlinear-optical device, not a laser. This allows versatile, flexible design features, such as methods of temporal and wavelength control that are not possible with lasers. The latter generally depend on population inversion of an optical gain medium, with associated optical lifetime and saturation limitations. On the other hand, optical parametric gain, oscillation and amplification are more amenable to modular system design because (as is explained below) they depend on nonlinear-optical coefficients and phase-matching conditions.

Nonlinear-Optical Origins of OPDs

Nonlinear optics involves interaction of a number ($\sigma > 2$) of optical waves in a medium with a nonlinear-optical susceptibility tensor $\chi^{(\sigma-1)}$. For inelastic optical processes, these waves (each with an angular frequency $\omega_1, \omega_2, \ldots, \omega_\sigma$) obey two conservation conditions, one for energy (or frequency):

$$\omega_1 + \omega_2 + \ldots + \omega_\sigma = 0 \qquad [1]$$

The other conservation condition (expressed in terms of wave vectors $\mathbf{k}_i$, with $i = 1, 2, \ldots, \sigma$, that have magnitudes $k_i = n_i\omega_i/c = 2\pi n_i/\lambda_i$, where n_i is the refractive index at vacuum wavelength λ_i and c is the speed of light) is for momentum:

$$\mathbf{k}_1 + \mathbf{k}_2 + \ldots + \mathbf{k}_\sigma + \Delta\mathbf{k} = 0 \qquad [2]$$

where $\Delta\mathbf{k}$ is the phase-mismatch between the σ interacting waves. Each frequency component and wavevector is ascribed a positive or negative sign, according to their phase relationships. Equation [2] defines phase-matching conditions in which $\Delta\mathbf{k}$ must be minimized to optimize the efficiency of the nonlinear-optical process of interest.

We shall consider below the two specific three-wave nonlinear-optical processes that are particularly relevant to this article, namely, those for an OPD (i.e., OPG, OPA, or OPO) or a DFG. Each of these is mediated by the nonlinear-optical susceptibility tensor $\chi^{(2)}$, which is nonzero in a crystalline medium only if it is noncentrosymmetric. Many such crystals are available. For example, lithium niobate ($LiNbO_3$) has been popular since the early days of pulsed tunable OPOs. Subsequent interest and activity in OPD technology have been stimulated by the availability of new nonlinear-optical materials such as BBO (β-barium borate, BaB_2O_4) and KTP (potassium titanyl phosphate, $KTiOPO_4$).

Additional recent impetus has come from quasi-phase-matched (QPM) nonlinear-optical media, such as periodically poled lithium niobate (PPLN) and PPKTP, tailored for specific wavelengths by periodic optical structuring of domains. QPM media offer compact, efficient, low-threshold alternatives to conventional birefringently phase-matched (BPM) media. Characteristics of BPM and QPM nonlinear-optical crystalline media are accessible *via* Arlee Smith's SNLO software package, accessible by internet at www.sandia.gov/imrl/XWEB1128/xxtal.htm.

For a three-wave OPD, which is of principal interest in this article, the energy and momentum conservation conditions of eqns [1] and [2] become:

$$\omega_P - \omega_S - \omega_I = 0, \qquad \mathbf{k}_P - \mathbf{k}_S - \mathbf{k}_I - \Delta\mathbf{k} = 0 \qquad [3]$$

where a laser input wave ('pump', frequency ω_P, wave vector $\mathbf{k}_P$) yields two coherent output waves ('signal', frequency ω_S, wave vector $\mathbf{k}_S$; 'idler', frequency ω_I, wave vector $\mathbf{k}_I$; $\omega_S \geq \omega_I$). Note that the idler frequency ω_I equals the difference $|\omega_P - \omega_S|$ between pump and signal frequencies. Equation [3] should be viewed in the context of sections (a–c) of Figure 1.

Equation [2] and the second half of eqn [3] apply strictly only to the conventional case of BPM media. In such media, the phase-matching condition $\Delta\mathbf{k} \approx 0$ is achieved by adjusting the angle and/or temperature of a birefringent nonlinear-optical crystal *via* its ordinary- and extraordinary-ray refractive indices. Such adjustments are used to optimize OPD conversion efficiency and thereby control the output signal and idler wavelengths, λ_S and λ_I. If it is assumed that the three waves are collinear and that $\Delta\mathbf{k}$ is exactly zero, then the signal frequency/wavelength is given simply in terms of the pump frequency/wavelength and the refractive indices n_i $(i = P, S, I)$ as:

$$\begin{aligned} \omega_S &= \omega_P(n_P - n_I)/(n_S - n_I); \\ \lambda_S &= \lambda_P(n_S - n_I)/(n_P - n_I) \end{aligned} \qquad [4]$$

An alternative to birefringent phase matching is the quasi-phase-matching approach first suggested by Nobel Laureate Nicolaas Bloembergen and co-workers in 1962. This approach has been realized experimentally only in the last few years, *via* QPM nonlinear-optical media such as PPLN or PPKTP. Equation [2] and the second half of eqn [3] need to be slightly modified to include QPM grating contributions.

In the corresponding case of a DFG (which, we repeat, is not an OPD), the energy and momentum conservation conditions of eqns [1] and [2] become:

$$\begin{aligned} |\omega_1 - \omega_2| - \omega_{\mathrm{diff}} &= 0, \\ \mathbf{k}_1 - \mathbf{k}_2 - \mathbf{k}_{\mathrm{diff}} - \Delta\mathbf{k} &= 0 \end{aligned} \qquad [5]$$

where two laser input waves (frequencies ω_1, ω_2; wave vectors $\mathbf{k}_1$, $\mathbf{k}_2$) yield one coherent output wave at the difference frequency $\omega_{\mathrm{diff}} = |\omega_1 - \omega_2|$ (with wave vector $\mathbf{k}_{\mathrm{diff}}$), as depicted in Figure 1d. As above, phase matching is defined by $\Delta\mathbf{k} \approx 0$ for BPM media and, in the case of QPM media, there is an additional grating contribution in the second half of eqn [5].

Finally, we mention briefly two other important forms of nonlinear-optical device arising from four-wave mixing processes that are mediated by the nonlinear-optical susceptibility $\chi^{(3)}$ which can be nonzero, even in isotropic or centrosymmetric media such as gases, liquids, optical fibers, and all classes of crystal. Optical parametric processes of this type contribute to stimulated Raman scattering (SRS), involving an optical medium with Raman-active resonance frequencies ω_R that coincide with the difference between two optical frequencies. This is particularly useful in providing a relatively straightforward source of coherent radiation, Raman-shifted at discrete intervals from the frequency ω_L of an input pump laser (either tunable or fixed-wavelength). Four-wave parametric processes are involved, such as that in eqn [1] with $|\omega_1| = |\omega_3| = \omega_L, |\omega_2| = \omega_L - \omega_R$, and $|\omega_4| = |\omega_1| + |\omega_3| - |\omega_2| = \omega_L + \omega_R$, where ω_R is the characteristic Raman frequency of the nonlinear-optical medium used in the Raman shifter (e.g., high-pressure hydrogen, deuterium or methane gas). These Raman-shifted intervals, both added to (anti-Stokes) and subtracted from (Stokes) the laser frequency ω_L, are integer multiples of ω_R. Other nonlinear-optical Raman parametric processes give rise to various forms of nonlinear Raman spectroscopy, such as CARS (coherent anti-Stokes Raman spectroscopy), and to Raman fiber-optical amplifiers, used in optical telecommunications.

Another promising area of developing OPD technology entailing $\chi^{(3)}$ nonlinearities concerns OPGs, OPAs, and OPOs based on highly nonlinear optical fibers, with either pulsed or cw pump lasers. Such processes use two pump waves (P) to generate tunable signal (S) and idler (I) output waves, such that $\omega_I = 2\omega_P - \omega_S$.

Nanosecond-Pulsed OPOs – Design and Wavelength Control

Here we consider pulsed OPOs that operate on nanosecond (ns, 10^{-9} s) time-scales, and design

Table 1 Some elements of OPO applications to spectroscopy

Key properties	*Options*	*Comments on instrument/technique*
Wavelength range: what forms of spectra need to be measured?	UV/visible (0.2–0.7 μm) Near IR (0.7–4.0 μm) Longwave IR (>4.0 μm)	Many nonlinear-optical OPO materials are available, but less well developed for longwave IR applications
OPO phase matching: BPM or QPM?	Birefringent phase matching (BPM) Quasi-phase matching (QPM)	BPM is well established and preferred for high-power operation. Various low-threshold QPM media are now available for both pulsed and cw OPOs
Temporal: continuous-wave (cw) or pulsed?	Pulsed for power and timing Cw for narrowest bandwidth	Ultrafast (ps, fs) OPO output pulses are relatively broadband, owing to the Fourier transform limit
Optical bandwidth: broad or narrow?	Broadband (free-running) Single longitudinal mode	The Fourier transform limit is 44 MHz (0.0015 cm^{-1}) for an ideal 10 ns pulse Cw OPOs offer even lower $\Delta\nu$. Ultrafast-pulsed OPOs are inevitably broadband: e.g., $\Delta\nu \geq 0.44$ THz ($\geq$15 cm^{-1}) for a 1 ps pulse
Mode of recording of spectra: continuous tuning or multiplex?	Scan narrowband signal/idler OPO output wavelength Free-running OPOs operate broadband in multiplex case	Wavelength control yields continuously tuned narrowband spectra. Multiplex spectra use dispersed detection or multi-wavelength spectroscopic tailoring
Wavelength control: intracavity elements or injection seeding?	Intracavity gratings or étalons Injection seeding of signal or idler by a tunable low-power coherent light source	Intracavity-element designs yield broad tunability but can be complicated. Injection seeding facilitates narrowband, mode-hop-free spectra and tailored multi-wavelength experiments

The above options determine ways to use OPOs for spectroscopic applications, such as: (1) linear absorption (e.g., with a multipass cell); (2) cavity ringdown (CRD) absorption spectra; (3) high-resolution spectra ($\Delta\nu \approx$ MHz or kHz); (4) nonlinear-optical (e.g., coherent Raman); (5) atmospheric remote sensing (e.g., DIAL); (6) fast (μs, ns) and ultrafast (ps, fs) processes.

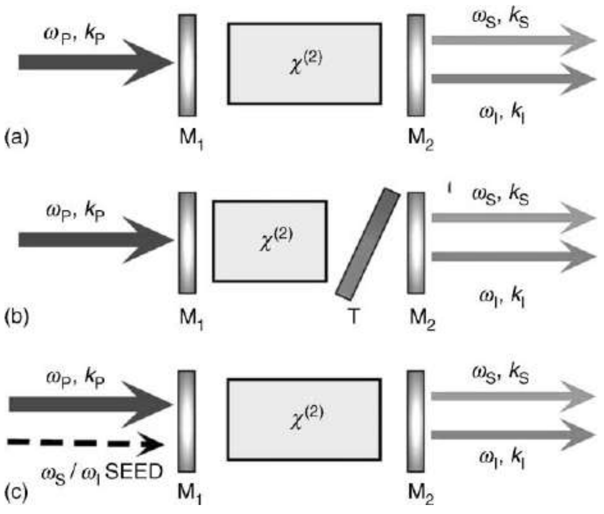

Figure 2 Schematic diagrams of three forms of optical parametric oscillator: (a) free-running OPO (with no active wavelength control), similar to Figure 1c; (b) OPO with an intracavity tuning element (T); (c) injection-seeded OPO.

features that make them fit for one of the principal applications of OPDs, namely, spectroscopy. Options that determine ways to use OPOs for spectroscopic applications are summarized in Table 1. Some design and wavelength-control features used in ns-pulsed OPOs are depicted schematically in Figure 2. Some of these applications and features are shared with other forms of OPO, notably the ultrafast-pulsed OPOs and cw OPOs that are considered much more briefly in later sections.

The first OPO, demonstrated by Giordmaine and Miller in 1965, was ns-pulsed; it was based on $LiNbO_3$, tunable over the range ~0.96–1.16 μm, spanning a signal- and idler-wavelength range of ±0.1 μm on either side of the degeneracy point defined by the 529 nm pump radiation (from a frequency-doubled, Q-switched $Nd:CaWO_4$ laser). This advance occurred soon after lasers were discovered and the potential of nonlinear optics had been realized through harmonic-generation processes such as frequency doubling. Pioneering work by Stephen Harris, Robert Byer and their co-workers at Stanford University soon established ns-pulsed OPOs as practical sources of tunable coherent light for significant applications such as spectroscopic sensing of chemical processes, in industrial or environmental diagnostics and in basic science. Such ns-pulsed OPOs had an impressive impact in the 1970s, followed by sporadic progress in much of the 1980s, and a dramatic resurgence of interest and activity since the late 1980s.

This irregular rate of progress is attributable to the availability of suitable high-quality nonlinear-optical crystals (as mentioned above), as well as high-performance pump lasers and advanced tunable OPO system designs.

In a typical ns-pulsed OPO, the pump laser delivers sufficient optical energy to enable parametric gain to build up from noise during the pump–pulse duration, exceeding the threshold for oscillation. To maximize gain, the parametric (signal and/or idler) waves are amplified by multipassing them during the pump pulse in an optical resonator, as depicted in Figures 1c and 2. Light travels ~3 m during a 10 ns pump pulse, so that the parametric waves can make ~15 round trips of a 10 cm linear OPO cavity.

A simple free-running OPO, comprising an optical cavity with input and output mirrors M_1 and M_2 but with no wavelength-selective elements, is depicted in Figure 2a. The output radiation from such an OPO generally has a relatively broad optical bandwidth – typically ~5 cm^{-1} (~150 GHz) or more – depending on several factors: refractivity, dispersion, and absorption of the OPO medium; wavelengths λ_P, λ_S, and λ_I; type of phase matching (BPM or QPM, whether collinear or not); crystal dimensions and orientation; cavity reflectivity and effective number of passes of the resonated wave; optical bandwidth, divergence, pulse duration and pulse energy of the pump radiation. Free-running pulsed OPOs represent one extreme of operational simplicity, yielding relatively broadband output radiation suitable for low-resolution or multiplex spectroscopy. Additional OPO wavelength-control measures are usually necessary for higher-resolution spectroscopic applications.

At the other extreme of operational complexity, intracavity wavelength-selective elements, such as gratings and/or étalons, provide a traditional way to control OPO output wavelength and to achieve narrow optical bandwidths. One such approach is depicted in Figure 2b, where a tuning element T (in this case, a tilted étalon or filter; in other designs, an intracavity diffraction grating replacing the output cavity mirror M_2) is inserted in the cavity. Such approaches were used in early pulsed $LiNbO_3$ OPO designs that were continuously tunable in the near IR with an optical bandwidth of ~0.1 cm^{-1} (~3 GHz), but these were often difficult to operate and were damage-prone (with intracavity losses from gratings and étalons causing the operating threshold to approach the damage threshold of OPO nonlinear-optical materials such as $LiNbO_3$).

Continuous narrowband tunability, preferably on a single-longitudinal-mode (SLM) basis, is a desirable performance characteristic of pulsed OPOs amenable to high-resolution spectroscopy. A commercially viable approach to this ideal was achieved by Walter Bosenberg and Dean Guyer in their advanced KTP OPO/OPA system, which includes a master OPO stage with a grazing-incidence grating. This is continuously tunable under computer control in the near IR (1.3–4 μm) with narrow optical bandwidth – ~0.02 cm^{-1} (~600 MHz) or better. Such a high-performance tunable coherent IR source has been used for various forms of laser spectroscopy (cavity ringdown, degenerate four-wave mixing, long-path absorption, etc.), including investigations of chemically reactive media, combustion diagnostics, and studies of processes in molecular beams.

A popular alternative approach to OPO wavelength control is injection seeding by a low-power tunable coherent source such as a tunable diode laser (TDL). This approach is depicted in Figure 3; in practice, a ring cavity is often used to avoid

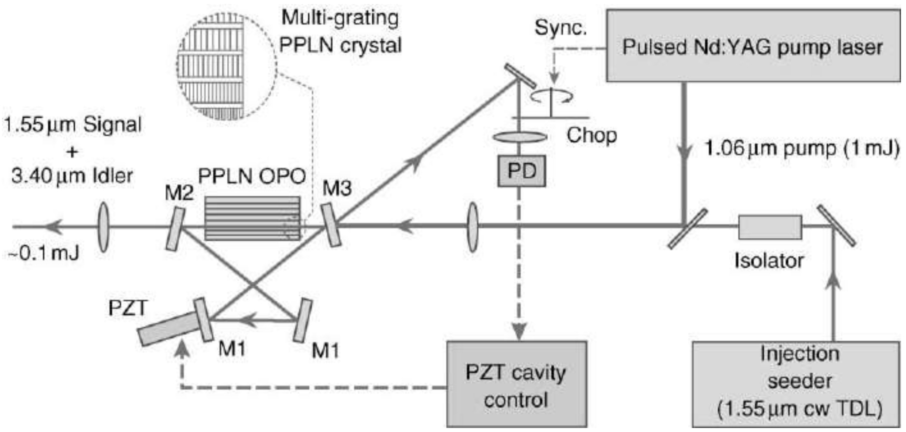

Figure 3 Schematic diagram of an injection-seeded tunable PPLN OPO system with active cavity control, developed in the author's laboratory. Legend: PD = photodetector, PZT = piezoelectric translator, TDL = tunable diode laser, M_{1-3} = cavity reflectors; the inset shows the QPM multi-grating structure of the PPLN nonlinear-optical crystal.

feedback from the OPO to the seed laser. A significant advantage of injection seeding is that OPO construction is simplified by putting the wavelength-control function into a module that is effectively separate from the optical generation and amplification functions. For example, a high-performance, TDL-seeded PPLN OPO system developed in the author's laboratory is illustrated in Figure 3. It features a multigrating QPM gain medium that is temperature-adjustable up to 250°C, and is pumped at a wavelength of 1.064 μm by a ns-pulsed, SLM Nd:YAG laser. An active cavity control scheme devised by Yabai He enables the SLM near-IR signal and idler outputs from this OPO (at ~1.5 μm and ~3.5 μm, respectively) to be tuned continuously without mode hops over optical frequency ranges of $\sim250\,\text{cm}^{-1}$ (~7.5 THz) with a SLM optical bandwidth of $<0.004\,\text{cm}^{-1}$ (<120 MHz), as has been confirmed in assorted applications to molecular spectroscopy. The PPLN OPO can by itself generate only moderate output pulse energy (~0.1 mJ, approximately 70% of which is signal radiation), since the 1.064 μm Nd:YAG pump laser pulse energy is limited to ~1 mJ to avoid the optical damage threshold of the PPLN crystal. For higher-power applications, a further portion of the Nd:YAG laser energy (~200 mJ/pulse) can be used to pump an OPA stage based on BPM $LiNbO_3$, thereby generating SLM signal and idler output pulse energies of ~1.3 mJ and ~0.7 mJ, respectively.

An alternative approach to injection-seeded tuning of ns-pulsed OPOs (used extensively by the author's group in the context of BPM media such as BBO and $LiNbO_3$) entails passive control of the OPO cavity by slightly misaligning one of its reflectors. This facilitates continuous tuning of the injection-seeded OPO signal and idler outputs by decreasing the effective finesse of the OPO cavity, so that it is then not necessary to lock the OPO cavity length to the seed wavelength. This mode of operation is simpler optically and electronically than active control of the OPO cavity. The method depends on the OPO cavity having a series of high-order transverse modes to smooth out the sharp, widely separated resonances that occur when the OPO cavity is well aligned; a resulting disadvantage is that the multiple transverse modes tend to cause some degradation of output beam quality. Nevertheless, this approach has proved useful for many applications of tunable OPOs, with seeding by either pulsed dye lasers or SLM TDLs.

This passive, misaligned-cavity approach to injection seeding of ns-pulsed OPOs is particularly well suited for spectroscopic applications requiring a coherent source that simultaneously generates two or more adjustable output wavelengths. Such a TDL-seeded, dual-wavelength BBO OPO has been used in the author's laboratory for coherent-Raman spectroscopic measurements of nitrogen (N_2) in furnace air; the relative intensity of the two spectroscopic features provides an instantaneous, turbulence-immune estimate of the temperature of gas in the furnace, within a single shot of the OPO's pump laser (i.e., within ~10 ns). This is an example of so-called spectroscopic tailoring, in which an OPO generates a structured set of two or more discrete output wavelengths, each of which is set to be on- or off-resonance with characteristic features in spectra of molecular target species of interest.

Dual-wavelength injection seeding of ns-pulsed OPOs is relevant to atmospheric remote sensing by DIAL (differential absorption lidar), with TDL-seeded OPO output switched between on- and off-resonance IR wavelengths on alternate shots of the pump laser; range-resolved measurements have been made of key atmospheric species such as methane (CH_4) and water vapor (H_2O). Figure 4 illustrates a proposed extension of the spectroscopic tailoring concept, applicable to dual- and multi-wavelength remote sensing applications by long-path absorption or DIAL. This multiplex system comprises a multi-wavelength passive-cavity pulsed OPO, injection-seeded by a set of single-mode TDLs and a computer-controlled fiber-optic switch to code and decode the resulting multiplex spectroscopic signals.

A significant frontier for narrowband pulsed OPOs entails extending the spectral output range beyond the present readily attainable near-IR limit of ~4 μm (as provided by materials such as BBO, KTP, $LiNbO_3$, PPLN, and PPKTP). This extension yields high-performance tunable coherent spectroscopic sources at longwave-IR wavelengths where many of the strongest molecular fundamental IR absorption bands are located. OPOs that target the longwave IR depend on availability of suitable nonlinear-optical materials and coherent pump sources at sufficiently long wavelengths (typically >2 μm). One of the most promising longwave-IR OPO materials is zinc germanium phosphide ($ZnGeP_2$), pumped at 2–3 μm by solid-state crystal lasers such as those based on holmium or erbium. An étalon-tuned narrowband ($\sim0.1\,\text{cm}^{-1}$) $ZnGeP_2$ OPO, itself pumped at 2.55 μm by a broadband ($\sim15\,\text{cm}^{-1}$) Nd:YAG-pumped $LiNbO_3$ OPO, is able to cover the wavelength range 3.7–8 μm and has been used to record spectra of H_2O vapor in the 6.2 μm region. Pulsed OPOs based on gallium arsenide (GaAs) also have great potential; they can be pumped at wavelengths as short as 1 μm and

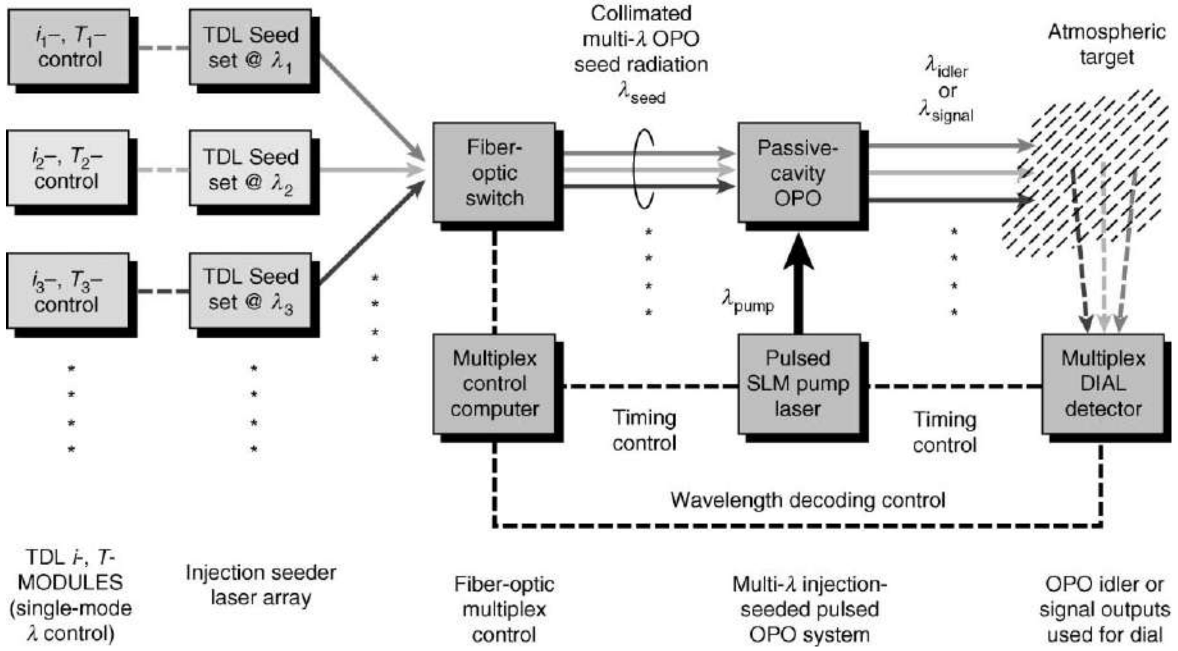
i1–, T1– control
TDL Seed set @ λ1
i2–, T2– control
TDL Seed set @ λ2
i3–, T3– control
TDL Seed set @ λ3
Fiber-optic switch
Collimated multi-λ OPO seed radiation λseed
Passive-cavity OPO
λidler or λsignal
Atmospheric target
Multiplex control computer
λpump
Pulsed SLM pump laser
Multiplex DIAL detector
Timing control
Timing control
Wavelength decoding control
TDL i-, T-MODULES (single-mode λ control)
Injection seeder laser array
Fiber-optic multiplex control
Multi-λ injection-seeded pulsed OPO system
OPO idler or signal outputs used for dial

Figure 4 Schematic diagram of a multi-wavelength pulsed OPO system employing spectroscopically tailored injection seeding. This design has been proposed by the author for atmospheric remote sensing by DIAL (differential absorption lidar) or long-path absorption.

offer (so far unrealized) tunability out to ~16 μm. Despite its cubic structure and lack of birefringence, GaAs serves as a QPM nonlinear-optical medium in the form of orientation-patterned films.

Another approach to operating continuously tunable OPOs is to vary the wavelength of the pump laser, while fixing other phase-matching parameters such as temperature of the gain medium and crystal angle or QPM grating pitch. Suitable tunable pump laser sources that have been used to pump ns-pulsed OPOs include Ti:sapphire and tunable fiber lasers.

The various wavelength-control strategies outlined above illustrate some of the distinctive attributes of OPOs. It should be recognized that the signal and idler wavelengths of an OPO are distinctly different from that of the pump laser itself, thereby opening up fresh regions of the electromagnetic spectrum. Even if an OPO is injection-seeded (e.g., by a cw TDL) at its signal or idler wavelength, it still generates complementary idler or signal radiation at a new wavelength and (in the case of a pulsed OPO) it also pulse-amplifies the original seed light. Moreover, the nonlinear-optical parametric relationships between the pump, signal, and idler waves in an OPO allow various wavelength-control approaches (e.g., spectroscopic tailoring by multi-wavelength injection seeding) that are simply not attainable in tunable lasers.

Ultrafast-Pulsed OPOs

Given the relatively high peak power of very short (sub-ns) pulses of coherent laser light, it might seem quite easy to use such light to pump the gain medium of an OPD and to exceed its OPO threshold. However, such a task is not at all straightforward, because an ultrafast light pulse, on the time-scale of picoseconds (ps, 10^{-12} s) or femtoseconds (fs, 10^{-15} s), does not stay in one place long enough to enable a coherent parametric wave to build up from noise in the nonlinear-optical medium. Light travels a distance of only ~0.3 mm in 1 ps, which does not allow it to make multiple traversals of the nonlinear-optical medium and build up coherent signal and idler waves. This problem is overcome by synchronous pumping (also used in some ultrafast lasers), in which a train of many consecutive ultrafast pulses from a pump laser interact sequentially with a single signal or idler pulse circulating within the OPO cavity. The mode-locking interval of the pump laser (essentially the round-trip transit time in the laser cavity) must therefore equal the round-trip time of the down-converted (signal or idler) pulse in the OPO cavity. The situation is aggravated when sub-ps (e.g., ~100 fs) pump pulses are used, because the pump and signal (or idler) waves have sufficiently different group velocities that they undergo 'temporal walk-off', becoming separated in time after traversing a relatively short length (typically 1–10 mm for 100-fs pulses) of the nonlinear-optical medium.

Practical ultrafast OPOs were introduced around 1990 and their utility grew as improved pump lasers, nonlinear-optical materials (e.g., KTP, PPLN), and OPO cavity designs became available. Research on ps- and fs-OPOs is now a hot topic, and numerous broadly tunable commercial systems are on the market. Ultrafast OPO-based spectroscopy is becoming extremely important in applications to biology and medicine, where many key processes occur on ps- and fs-time-scales.

Continuous-Wave OPOs

The first cw OPO, reported in 1968, operated in the near-IR and was based on barium sodium niobate ($Ba_2NaNb_2O_5$ – known popularly as 'banana'); it was soon followed by visible cw OPOs based on $LiNbO_3$. However, there was not much more substantial progress on practical cw OPOs for another 20 years, until several technological advances had been made: low-threshold QPM nonlinear-optical gain media, such as MgO:$LiNbO_3$ and KTP; efficient, wavelength-selective cavity designs; all-solid-state cw pump lasers (e.g., diode-pumped crystal lasers or MOPA diode systems). More recently, QPM media (such as PPLN and PPKTP) have provided a further stimulus to efficient, compact cw OPOs.

Most ns-pulsed OPOs use singly resonant (SROPO) designs in which the cavity resonates at either the signal or idler wavelength (but not both), doubly resonant (DROPO) schemes in which both signal and idler resonate are often needed to meet the more demanding pump threshold requirements of cw OPOs. Continuous tuning of a DROPO without longitudinal mode hops needs multiparameter approaches, either dual-cavity (with separate signal- and idler-resonant arms) or concerted tuning of pump and signal/idler wavelengths. Efficient cavity designs for cw OPOs include those that are pump-enhanced (i.e., the cavity is resonant at the pump wavelength as well as the signal and/or idler) and intracavity arrangements in which the laser and OPO gain media share a common cavity. Progress is also being made on cw OPOs based on highly nonlinear optical fibers. With such developments, cw OPOs now have great potential as tunable spectroscopic sources, through their narrow optical bandwidth and their compactness.

OPGs and DFGs – Dispensing with the Optical Cavity

Most of this article has focused on OPOs, in which parametric processes are generated by pumping a nonlinear-optical medium in an optically resonant cavity. As already explained, such a cavity is used to enhance the optical parametric gain by multipassing the signal and/or idler radiation. In some cases, the cavity also contributes to wavelength control, temporal characteristics, beam quality, etc. However, not all optical parametric and related devices need a cavity, as is evident from the OPG, OPA, and DFG schematics in Figures 1a, b, and d.

The use of one or more OPA stages, to supplement the intracavity amplification already present in an OPO itself, is widespread in tunable OPO/OPA systems. Moreover, in some applications (e.g., spectroscopic measurements where optical power is not at a premium) of pulsed OPDs, it is found that a single pass of a (relatively long) nonlinear-optical crystal results in a useful OPG (without an OPO cavity). For instance, Scott Bisson, Thomas Kulp, Peter Powers and co-workers have developed and used several OPD systems of this type. A cavity-free OPG/OPA system that is pumped by a 1 kHz Nd:YAG laser, based on 'fan' PPLN (tuned by continuous transverse variation of the QPM grating pitch), and spectrally filtered by a high-finesse étalon, generates ~10 μJ/pulse of idler radiation. This OPG/OPA radiation has been used to measure narrowband ($\Delta\nu < 0.1\ \mathrm{cm}^{-1}$) cavity-ring-down spectra of CH_4 gas.

Likewise, various tunable DFG devices (closely related, but not identical, to OPDs) have been used in many spectroscopic applications. For instance, a DFG system developed by Frank Tittel and colleagues at Rice University took the near-IR output of two room-temperature TDLs (at 1.55 μm and 1.32 μm, respectively), mixed them in a BPM nonlinear-optical medium such as $AgGaSe_2$, and generated 8.7 μm mid-IR radiation for long-path absorption studies of sulfur dioxide (SO_2), an important atmospheric pollutant. The same group has also used QPM media (e.g., PPLN) for DFG mixing of TDL outputs in the 0.8 μm–1.6 μm range, yielding output at ~3.5 μm to monitor gases such as H_2O, CH_4, and formaldehyde (H_2CO); such DFG devices have been used in NASAs Lunar-Mars Life Support Test program.

Relative to OPGs and DFGs, OPOs and OPAs offer higher output power and coherence, both of which are advantageous in some applications (e.g., remote spectroscopic sensing of the atmosphere). Nevertheless, OPGs and DFGs still have a useful role to play, as indicated above.

Concluding Remarks – When to Use an OPD?

Coherent light sources based on optical parametric generation, amplification, and oscillation represent a significant application of nonlinear optics. Such devices can in turn address as many applications as lasers themselves, at least in principle. However, the fact that OPDs need to be pumped by a laser tends to limit their use to situations that take advantage of their wavelength-conversion and tunability capabilities. The most obvious areas of application of an OPO or any other OPD are therefore those in which the wavelength of the output radiation is critical (e.g., spectroscopy, as indicated in various parts of the above text).

From time to time, there have been suggestions that OPOs would largely displace established tunable lasers such as dye and Ti:sapphire lasers. In the early 1990s, one major laser manufacturer ventured to advertise its new OPO-based instrument with the slogan, "Goodbye to Ti: and Dye". However, such a projected monopoly for OPOs has not eventuated and most laser spectroscopy facilities now depend upon a balanced combination of OPOs together with various forms of tunable laser (including semiconductor lasers, which are almost ubiquitous).

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$-harmonic Generation; $\chi^{(3)}$ – Third-harmonic Generation; Four-wave Mixing. **Optical Parametric Devices:** Optical Parametric Oscillators (Continuous Wave); Optical Parametric Oscillators (Pulsed). **Scattering:** Stimulated Scattering. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Barnes NP (1995) Optical parametric oscillators. In: Duarte FJ (ed.) *Tunable Lasers Handbook*, pp. 293–348. San Diego, CA: Academic Press.

Byer RL (1975) Optical parametric oscillators. In: Rabin H and Tang CL (eds) *Quantum Electronics: A Treatise*. Nonlinear Optics, part B, chap. 9, vol. 1, pp. 587–702. New York: Academic.

Byer RL and Herbst RL (1977) Parametric oscillation and mixing. In: Shen Y-R (ed.) *Nonlinear Infrared Generation*, pp. 81–137. New York: Springer-Verlag.

Dmitriev VG, Gurzayan GG and Nikogosyan DN (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn, pp. 339–362.

Dunn MH and Ebrahimzadeh M (1998) Optical parametric oscillators. In: Finlayson DM and Sinclair BD (eds) *Advances in Lasers and Applications*, pp. 61–82. Bristol and Philadelphia: Institute of Physics Publishing.

Ebrahimzadeh M and Dunn MH (2001) Optical parametric oscillators. In: Bass M, Enoch JM, Van Stryland EW and Wolfe WL (eds) *Handbook of Optics, Fiber Optics and Nonlinear Optics*, vol. 4, pp. 22.1–22.73. New York: McGraw-Hill.

Harris SE (1969) Tunable optical parametric oscillators. *Proceedings of the IEEE* 57: 2096–2113.

Orr BJ (2000) IR Lidar applications in air monitoring. In: Myers RA (ed.) *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation, Applications of Instrumental Methods*, vol. 3, pp. 2007–2032. Chichester: Wiley.

Orr BJ, Johnson MJ and Haub JG (1995) Spectroscopic applications of pulsed tunable optical parametric oscillators. In: Duarte FJ (ed.) *Tunable Laser Applications,* chap 2, pp. 11–82. New York: Marcel Dekker.

Sorokina IT and Vodopyanov KL (eds) (2003) *Solid-State Mid-Infrared Sources*. New York: Springer-Verlag.

Sutherland RL (1996) Optical parametric generation, amplification, and oscillation. *Handbook of Nonlinear Optics,* chap. 3, pp. 111–206. New York: Marcel Dekker.

Optical Parametric Oscillators (Continuous Wave)

S Schiller, Heinrich-Heine Universität Düsseldorf, Düsseldorf, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A continuous-wave optical parametric oscillator (cwOPO) is a frequency conversion device in which a continuous-wave (cw) laser field (pump) of frequency ω_p is partially converted into two cw waves of lower frequencies ω_s, ω_i (historically denoted by signal and idler) inside an optical cavity. The three frequencies are related by $\omega_i + \omega_s = \omega_p$. The cavity contains a nonlinear medium with a second-order susceptibility and the mirror reflectivities are chosen such that at least one of signal and idler resonates in the cavity. The conversion efficiency from pump power to signal or idler power or both can be larger than 10%. The frequencies of signal and idler waves can be tuned by means similar to conventional lasers, i.e., by changing the spectral dependence of the cavity loss and by tuning the cavity length. Additional tuning parameters specific to OPOs are the pump frequency and the refractive indices of the nonlinear medium for pump, signal and idler. CwOPOs provide a unique combination of features that makes them suitable sources for a variety of spectroscopic applications. Today's cwOPOs can exhibit the following features:

- spectral coverage from the near-infrared (0.7 μm) to the mid-infrared (5.2 μm); for a single device two optical octaves can be covered without change of optical components;
- high output power (up to several W);
- high pump power conversion efficiency (>30%) and good wall-plug efficiency;
- nearly diffraction-limited output beams (modes are close to lowest-order transverse electromagnetic (TEM_{00}) modes)
- single-frequency output, absence of frequency jumps (mode hops);
- narrow free-running linewidth (<6 kHz);
- high absolute frequency stability (<100 MHz drift per hour);
- continuous frequency tuning exceeding 10 GHz;
- rapid, stepwise tuning over several THz; and
- transportability.

cwOPOs complement their pulsed relatives (*see* **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed)) nanosecond and synchronously pumped pulsed parametric oscillators (*see* **Optical Parametric Devices:** Overview).

Basic Principle

The operation of a cwOPO is most simply described for a singly resonant OPO (SRO) (see Figure 1 and Figure 2a).

Without loss of generality we take the signal wave to be the resonant wave. When a higher-frequency pump wave (power P_p, angular frequency ω_p) and a lower-frequency (signal) wave (power P_s, angular frequency ω_s) traverse a $\chi^{(2)}$ medium (*see* **Fiber and Guided Wave Optics:** Nonlinear Optics), a transfer of power from the pump wave to the signal wave occurs, described by a power gain:

$$\Delta P_s = GP_s = EP_pP_s \qquad [1]$$

This parametric amplification expression is valid as long as $\Delta P_s \ll P_s, P_p$, i.e., in the limit of small gain and small pump depletion. The (unsaturated) gain coefficient E is a function of: the medium's indices of refraction, orientation and length L, the frequencies

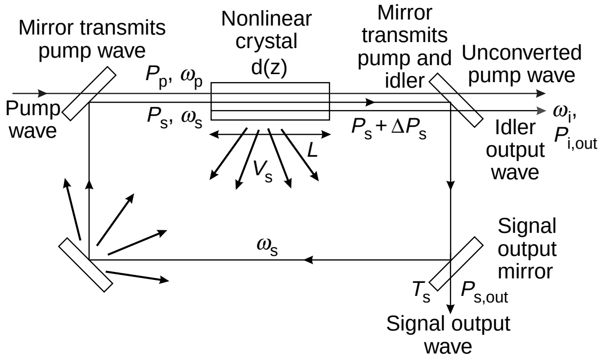

Figure 1 Schematic of a singly resonant cwOPO (SRO) in ring configuration. The signal wave (ω_s) is resonated, the idler wave (ω_i) is not. In the nonlinear crystal, the pump wave overlaps with the signal and idler waves; here they are shown displaced with respect to each other for clarity. The signal wave's attenuation coefficient per roundtrip, S_s, is the sum of loss inside the nonlinear crystal, V_s (indicated by arrows), and transmission through mirrors, here represented by a single mirror of transmission T_s. Scatter and absorption losses at mirrors (indicated by arrows) are included in V_s.

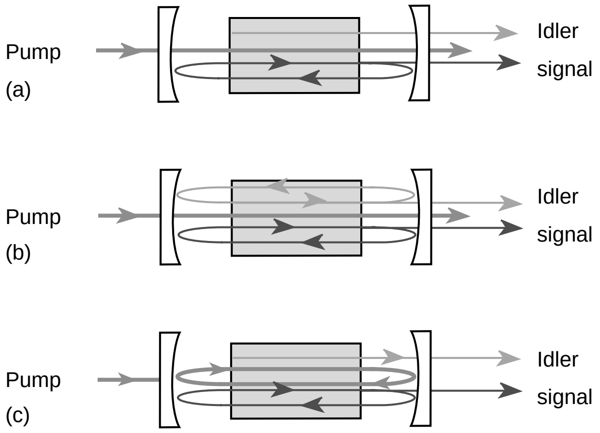

Figure 2 Common types of cwOPOs. (a) basic SRO. (b) Doubly resonant OPO (DRO). Standing-wave cavities are shown. (c) SRO with enhancement of the pump wave (PR-SRO); a common-cavity configuration is shown.

ω_s, ω_p, and the spatial modes and overlap of the two beams.

If the wave ω_s is made to circulate inside a cavity and the power attenuation per roundtrip, S_sP_s is smaller than the power gain per roundtrip (eqn [1]), a sustained oscillation is possible. This is satisfied when the input pump power P_p exceeds the threshold:

$$P_{p,in}^{th} = S_s/E \qquad [2]$$

As the pump power P_p is increased above $P_{p,in}^{th}$, conversion of pump power to idler power occurs, implying that the pump power decreases within the medium. Qualitatively, the average pump power $\langle P_p \rangle$

along the medium remains at the level $P^{\mathrm{th}}_{\mathrm{p,in}}$, so that the roundtrip fractional power gain remains constant, $G \simeq E\langle P_{\mathrm{p}}\rangle \simeq EP^{\mathrm{th}}_{\mathrm{p}} \simeq S_{\mathrm{s}}$. This ensures that the device reaches a stationary state. The circulating signal power P_{s} increases steadily with increasing pump power, and so does the signal power emitted from the OPO, $T_{\mathrm{s}}P_{\mathrm{s}}$, which is equal to the 'useable' fraction of the roundtrip power gain, $(T_{\mathrm{s}}/S_{\mathrm{s}})GP_{\mathrm{s}}$.

The parametric process that allows the signal wave to experience gain (eqn [1]), also generates an idler wave (difference frequency generation). Its power at the end of the medium is given by the Manley–Rowe relation:

$$P_{\mathrm{i}} = \frac{\omega_{\mathrm{i}}}{\omega_{\mathrm{s}}}\Delta P_{\mathrm{s}} \qquad [3]$$

An OPO also requires a process that starts the oscillation. Spontaneous parametric fluorescence provides this trigger. Due to this quantum effect, a pump wave traversing a $\chi^{(2)}$ medium will generate a flux of signal photons even if there is no signal wave present at the crystal input face. In the photon picture, a tiny fraction (on the order 10^{-9} for a 1 W pump) of the pump photons spontaneously annihilate, creating pairs of signal and idler photons (photon splitting). The idler photons have a frequency $\omega_{\mathrm{i}} = \omega_{\mathrm{p}} - \omega_{\mathrm{s}}$, so that photon energy conservation holds. The process may be thought of as being stimulated by the vacuum fluctuations of the electromagnetic field at the signal and idler frequencies. Some of the signal photons are emitted into the cavity mode, so that after a roundtrip they can serve as a seed for further (stimulated) photon splitting. The process repeats, and the signal wave builds up exponentially in time to a macroscopic level if the pump power is above threshold. The spontaneous (as well as the parametric amplification) process is strongest when momentum conservation is essentially fulfilled, $k_{\mathrm{p}} \simeq k_{\mathrm{s}} + k_{\mathrm{i}}$, where $k = \omega n(\omega)/c$.

In the photon picture, the relation (eqn [3]) is a statement that signal and idler photons are always created in pairs.

Types of OPOs

OPO types may be classified according to the number of waves that resonate within cavities. Typically, a realizable roundtrip loss coefficient is on the order of $S_{\mathrm{s}} = 1\%$ (to which the linear absorption of the nonlinear crystal and mirror transmissivities contribute) and the gain coefficient $E \sim 0.1\%/\mathrm{W} - 1\%/\mathrm{W}$ for Gaussian waves, depending on the pump wavelength, crystal type, etc. This leads to SRO threshold powers on the order of 1 to 10 W. Until the early 1990s, solid-state pump lasers of this power level and single-frequency output were unavailable. This led to the interest in and development of cwOPOs in which more than one of the three waves is resonated (Figure 2). The corresponding external threshold powers are

$$\begin{aligned} P^{\mathrm{th}}_{\mathrm{p,in}} &\simeq S_{\mathrm{s}}S_{\mathrm{i}}/4E \text{ for a doubly-resonant OPO} \\ &\qquad \text{(DRO, Figure 2b)} \\ P^{\mathrm{th}}_{\mathrm{p,in}} &\simeq S_{\mathrm{s}}S_{\mathrm{p}}/E \text{ for a pump-resonant SRO} \\ &\qquad \text{(PR-SRO, Figure 2c, see eqn [35])} \\ P^{\mathrm{th}}_{\mathrm{p,in}} &\simeq S_{\mathrm{s}}S_{\mathrm{i}}S_{\mathrm{p}}/E \text{ for a triply-resonant OPO (TRO)} \end{aligned} \qquad [4]$$

Here $S = T + V$ is the fractional power attenuation per roundtrip, arising from the mirror transmission T and absorption/scatter loss V. While the mirror transmission maintains power in a given spatial mode, the loss V removes optical power completely and is the fundamental cause for the nonunity efficiency of the frequency conversion process. Typically, each additional resonant enhancement lowers the threshold power by one to two orders.

Examples for threshold values reported in the literature are: 1.9 W for a 920 nm-pumped SRO, 4 mW for a 0.5 μm-pumped DRO, and 140 mW for a 1 μm-pumped common-cavity PR-SRO. DROs can therefore be pumped even by low-power diode lasers. TROs (where pump, signal, and idler are resonated) are not of importance due to their higher complexity.

Output Power

As an example, the signal and idler output powers for a PR-SRO that resonates the signal are (for a derivation, see Appendix B):

$$P_{\mathrm{s,out}} \simeq T_{\mathrm{s}}P_{\mathrm{s}}(0), \qquad P_{\mathrm{i,out}} \simeq (\omega_{\mathrm{i}}/\omega_{\mathrm{s}})S_{\mathrm{s}}P_{\mathrm{s}}(0) \qquad [5]$$

where the circulating signal power is

$$P_{\mathrm{s}}(0) \simeq \frac{\omega_{\mathrm{s}}}{\omega_{\mathrm{p}}}\frac{4T_{\mathrm{p}}}{S_{\mathrm{s}}S_{\mathrm{p}}}P^{\mathrm{th}}_{\mathrm{p,in}}\left(\sqrt{\frac{P_{\mathrm{p,in}}}{P^{\mathrm{th}}_{\mathrm{p,in}}}} - 1\right) \qquad [6]$$

T_{p} is the input mirror transmission for the pump wave, and T_{s} is the output mirror transmission of the signal wave. Signal and idler output powers have the same dependence on pump input power and increase monotonically with $P_{\mathrm{p,in}}$. The conversion efficiencies reach a maximum four times above threshold, i.e., when $P_{\mathrm{p,in}} = 4P^{\mathrm{th}}_{\mathrm{p,in}}$. There, the signal and idler photon (quantum) conversion efficiencies $\omega_{\mathrm{p}}P_{\mathrm{s,out}}/P_{\mathrm{p,in}}\omega_{\mathrm{s}}$ and $\omega_{\mathrm{p}}P_{\mathrm{i,out}}/P_{\mathrm{p,in}}\omega_{\mathrm{i}}$ exceed 25% and 50%, respectively, if the mirror transmissions T_{p}, T_{s} for pump and signal are chosen larger than the respective loss V_p, V_s. This can be achieved in practice, of course, at the expense of a higher threshold.

Nonlinear Materials

While the fundamental developments in cwOPOs were performed using birefringently phase-matched nonlinear crystals (MgO:LiNbO$_3$, LBO, KTP, KNbO$_3$) (*see* **Materials for Nonlinear Optics:** Liquid Crystals for NLO), and substantial spectral coverage is possible with these crystals, most cwOPOs are now operated with quasi-phase-matched (QPM) nonlinear crystals (*see* **Nonlinear Optics, Application:** Phase Matching). The most commonly used material is periodically poled (PP) LiNbO$_3$ (PPLN), other crystals used are PPKTP and PPLiTaO$_3$, crystals. Their use has led to a substantial widening of the spectral coverage, opening in particular the wavelength range $\lambda > 1.6$ μm, a reduction of the required pump powers thanks to their much higher nonlinearities compared to birefringently phase-matched crystals, and a simplification of (gross) tuning through the use of multigrating structures (containing several sections of differing Λ periods side by side within the same crystal). The above PP crystals exhibit a loss of 0.1%/cm at wavelength around 1 μm, and can be coated with broadband dielectric antireflection coatings, in order to maximize the overall transmission. Another important aspect is that the crystals withstand long irradiation times with high-power focused cw light. For PPLN, operation temperature above 100°C is chosen in order to prevent photorefractive effects.

Spectral Coverage

An example of particularly wide spectral coverage reported in the literature is a cwOPO pumped by a frequency-doubled Nd:YVO$_4$ laser ($\lambda_p = 532$ nm) using a multigrating PPLN crystal. With grating periods $\Lambda = 6.5-9.6$ μm, the range 660–1030 nm (signal wave) and 1100–2800 nm (idler wave) was covered. A cwOPO with a Nd:YAG pump source emitting at the fundamental wavelength ($\lambda_p = 1064$ nm) and a multigrating PPLN crystal ($\Lambda = 25.5-31.2$ μm) achieved a range of 1.45–1.99 μm (signal) and 2.3–4.0 μm (idler).

For idler wavelengths beyond 4 μm, idler absorption by the crystals types above becomes relevant, leading to a reduction in the gain G (this effect is not considered in the treatment given in the Appendix) and to an increase of the threshold. To a certain extent, this can be compensated with powerful cw pump lasers. For example, a 1064 nm pumped PPLN-SRO achieved idler emission in the range 3.6–4.7 μm at 11 W pump power. Also, an 850 nm pumped PR-SRO covered an idler range 4.3–5.3 μm when pumped with 750 mW.

In the future, cwOPOs will certainly profit from the development of novel QPM materials with wider IR transparency range.

Frequency Control

The small-signal gain $G(\omega_s)$ of a parametric interaction in a cwOPO is analogous to the homogenously broadened gain in a laser medium. In contrast to it, however, spatial hole-burning cannot occur. Therefore, mode competition is generally strong in a cwOPO, which will oscillate on a single longitudinal mode: the one for which the difference (or ratio) between 'unsaturated' parametric gain G and total loss S is largest (Figure 3). This is equivalent to the condition of minimum threshold.

The functions $G(\omega_s)$ and $S(\omega_s)$ are therefore important characteristics of a cwOPO. Oscillation can only occur for a frequency ω_s that is very close to a cavity mode frequency $\omega_q = 2\pi qc/OPL(\omega_q)$, where q is the integer mode number, and $OPL(\omega_s)$ is the round-trip optical path length at the frequency ω_s.

To tune a cwOPO to a particular signal/idler frequency pair, a coarse tuning is initially performed

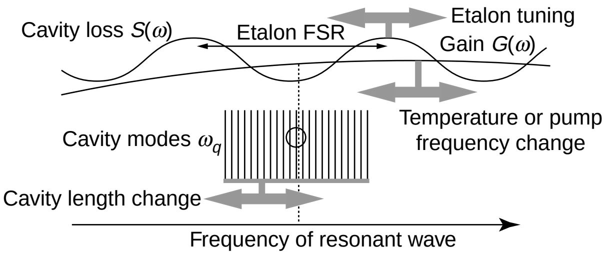

Figure 3 Frequency selection and tuning in a SRO (or PR-SRO) containing an etalon. Oscillation occurs on the signal cavity mode (circled) closest to the frequency where the difference between the small-signal gain G and the loss S is largest (dashed line). The frequency of this mode or the mode number q can be changed by changing the parameters indicated near the thick double arrows. The parameter changes shift the line(s) 'attached' to the corresponding double arrow. Changing the temperature leads to a rough setting of the oscillation frequency. Tilting the etalon allows mode-hop tuning of signal and idler frequencies. Fine tuning by a small amount can be done by changing the cavity length. The frequency of the idler is not shown here; it is always given by the difference between pump frequency and the frequency of the oscillating mode. Thus, tuning of the pump frequency generally leads to a tuning of the idler frequency. For modest pump tuning the amount of signal tuning however depends on SRO type: no signal tuning in a basic SRO and in a dual-cavity PR-SRO, while in a common-cavity PR-SRO pump and signal frequencies are correlated and thus pump tuning leads to signal tuning. FSR: free spectral range.

by tuning the corresponding phase mismatch ΔkL (L is the crystal length) approximately to zero. Taking into account a modulation of the nonlinear susceptibility of the crystal with period Λ, the quantity $\Delta k = k_p - k_s - k_i \pm 2\pi/\Lambda$ describes the deviation from momentum conservation in the photon splitting process. When using a multigrating QPM crystal, the first tuning consists in the selection of a grating of appropriate period Λ, which is moved into the pump wavepath. This is followed by tuning the photon momenta via the refractive indices of the nonlinear medium, which is usually accomplished by temperature tuning, or alternatively by angle-tuning of the crystal. These steps place the maximum of $G(\omega_s)$ to the vicinity of the desired frequency. Options for tuning the emission exactly to the desired frequency are described below for the different OPO types.

Because external perturbations change parameters that enter in $G(\omega_s)$, $S(\omega_s)$, and ω_q, the value of the oscillating frequency will also change in time. The change will be continuous over a certain range of the parameters, but if their perturbation is large enough, a mode hop can occur: the oscillation on the initial mode q will eventually cease and another mode q', for which $G(\omega_{q'}) - S(\omega_{q'})$ is larger, will oscillate instead. A variety of parameters is affected by perturbations: pump power, pump beam direction, cavity length, crystal temperature, and pump frequency.

When such perturbations are applied on intentionally, the OPO output frequencies can be tuned, scanned, or adjusted. The tolerable change in the various parameters before a mode-hop occurs, and the sensitivity of the output frequencies to changes in the parameters depends on the particular cwOPO type. One important approach to suppress mode-hops consists in adding elements into the cavity, such as an etalon, that modify the function $S(\omega)$, typically by modulating it spectrally.

SRO

In the following, we assume that the signal wave is resonated. The SRO is the simplest cwOPO type. The oscillation nominally occurs at the cavity frequency that is closest to the gain maximum, as described above. In practice, environmental disturbances will cause mode-hops on a time-scale of minutes. For this reason, an etalon is added to the cavity. Mode-hop-free oscillation, for several hours, is then achieved.

Typical nonlinear crystals exhibit parametric gain with spectral widths of hundreds of GHz. In contrast, the cavity mode spacing (free spectral range, FSR) c/OPL is on the order of 1 GHz. The tuning range achievable by a cavity length change (using, e.g., a piezo translator) is therefore, at most, one free spectral range (equal but opposite for signal and idler); then a mode-hop to the neighboring cavity mode occurs, with a signal/idler frequency jump by one FSR in the opposite direction to the previous tuning. This process repeats itself as the cavity length is tuned further. It is, therefore, not possible to easily and reliably access a large frequency range by using the cavity length as tuning parameter.

As shown in Figure 3, if an etalon of linewidth smaller than the gain bandwidth is added to the cavity, it becomes the dominant influence (etalon FSRs on the order of one hundred to a few hundred GHz have been used). This means that shifting the etalon's transmission maximum (i.e., the cavity loss S) in frequency space (by tilting a solid etalon or changing the spacing of an air-spaced etalon) as indicated by the double-headed arrows in Figure 3, forces the signal to mode-hop to an adjacent mode, with a corresponding idler frequency hop in the opposite direction. As the etalon is tuned further this occurs repeatedly. This discontinuous 'mode-hop tuning' is useful for tuning the SRO output frequencies over a significant range (e.g., 100 GHz) and allows to reach a desired frequency within one cavity free spectral range.

Continuous tuning of both signal and idler waves is obtainable, if the pump laser is not frequency tunable, by changing the cavity length and synchronously tuning the etalon. This can be achieved by a feedback or feedforward system that controls the etalon angle or spacing. Up to 38 GHz tuning for signal and idler were achieved in this way.

When a tunable pump is available, one can tune the pump frequency while keeping the cavity length and the etalon angle fixed. This keeps the signal cavity mode fixed, and thus the nonresonant wave's frequency (usually the idler) is tuned by an amount equal to the pump frequency change. Continuous tuning ranges exceeding 50 GHz have been obtained in this way. For such pump tuning ranges, the shift of the gain curve, indicated by a double-arrow in Figure 3, is usually sufficiently small that it can be neglected. The use of a widely tuneable pump laser is also possible. A (discontinuous) idler tuning of 9 THz around 3.3 μm has been demonstrated, covered in just 0.3 ms thanks to a large pump tuning rate. In such a case of large pump tuning range, the gain shift is important, causing a significant tuning also of the signal.

The advantages of the basic SRO as compared to other cwOPOs are mainly the simpler tuning behavior, the possibility of continuous tuning without necessitating a tunable pump laser, and the possibility to achieve a very wide tuning range without optics change, since only the spectral range covered by one of the two parametric waves must be provided as a high reflectivity coating on the cavity mirrors. Moreover, in the simplest implementation, it is possible to operate without any electronic servo control of cavity elements, at the expense of a reduced power and frequency stability. In terms of spectral properties (linewidth, frequency stability, time between undesired mode-hops) the SRO does not offer particular advantages compared to other cwOPO types, even if actively stabilized (see below). The disadvantage of the basic SROs based on Gaussian modes is the high threshold, necessitating expensive high-power (several W) cw lasers. This disadvantage is likely to disappear with progress in laser development. The alternative of waveguide SROs is mentioned below.

PR-SRO

In a PR-SRO, the pump wave is also resonated, leading to a substantial reduction of threshold. For example, it is possible to generate mid-IR radiation (3 μm range), even with much more affordable 1 W-level Nd:YAG pump lasers.

One implementation is the common-cavity PR-SRO (Figure 4). The cavity length is either locked to the pump laser frequency or vice-versa. If a frequency-stable pump laser is used, the first solution is favorable, since then both signal and idler waves have a stability comparable to that of the pump laser. With an etalon (coated so as to be transparent for the pump wave) inside the cavity, mode-hop-free operation for several hours has been demonstrated.

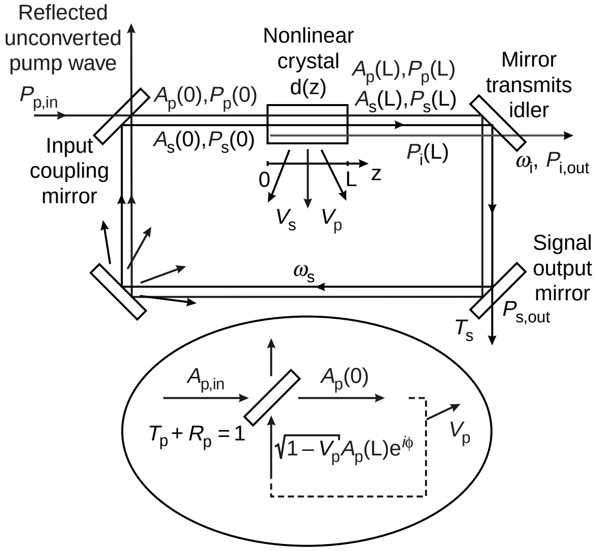

Figure 4 Schematic of a pump-resonant SRO with a common cavity for pump and signal waves. The inset shows the pump wave fields incident on and leaving the input coupling mirror. The arrow symbolizes pump wave loss during the round trip.

Mode-hop tuning in a PR-SRO with cavity-to-laser lock can be achieved:

(i) by tuning the etalon at constant pump frequency: the signal mode-hops from mode to mode, the idler tunes equally but in opposite direction;
(ii) by tuning the pump frequency at constant etalon position: the cavity length must follow and therefore the signal tunes by $\Delta\omega_s = (\omega_s/\omega_p)\Delta\omega_p$, while the idler frequency tunes by $\Delta\omega_i = (\omega_i/\omega_p)\Delta\omega_p$. This continues until the signal must mode-hop back, causing an equal but opposite frequency jump in the idler. The pump tuning is thus completely transferred to the idler.

Continuous signal and idler tuning is performed by tuning the pump frequency and synchronously adjusting the etalon angle. Thus, a limitation of the common-cavity PR-SRO concept is that continuous tunability necessitates a tunable pump laser.

In a dual-cavity, PR-SRO pump and signal waves resonate in separate cavities. The pump cavity is usually stabilized to the pump frequency and an etalon is placed into the signal cavity for better stability against mode-hops. This configuration combines the tuning simplicity of a basic SRO with a low-threshold power. For example, a threshold for a 1 μm-pumped PPLN dual-cavity PR-SRO at the 0.4 W level was obtained.

Mode-hop tuning (in opposite directions) of the frequencies of both parametric waves in steps of the cavity's FSR can be performed by changing the etalon angle (Figure 3). A range exceeding 50 GHz (at constant output power) has been achieved. Continuous signal and idler wave tuning (by equal but opposite amounts) is possible by changing the signal cavity length (Figure 3). However, the range is limited to the FSR of the cavity, typically on the order of 500 MHz, then the frequencies mode-hop back. It is expected that wider continuous tuning should be possible by synchronous etalon and cavity length scan. If a tuneable pump laser is employed, continuous tuning of only the idler frequency is possible by keeping the etalon and signal wave cavity length fixed; the pump tuning range is fully transferred to the idler.

DRO

In contrast to an SRO, this device type requires both parametrically generated waves to be resonant in the cavity. Here we consider only DROs with a common cavity for both signal and idler waves (dual-cavity DROs have also been demonstrated).

Because of the added resonance condition, mode-hops of the oscillating mode pair occur already upon a very small change (~1 nm) in the cavity length. Under normal environmental conditions such changes are unavoidable and therefore a 'free-running' DRO does not produce frequency-stable output. However, by active control of one additional parameter of the OPO system the mode-hops can be effectively suppressed. A practical solution is to actively control the length of the cavity via the position a of one of the cavity mirrors so that a particular mode pair, q_s, q_i, continues to oscillate; this is achieved by maximizing the OPO output power or by minimizing the detuning of one of the two resonant waves from its respective cavity mode frequency. This control system ensures that energy conservation and the two resonance conditions are simultaneously satisfied:

$$\begin{aligned} &\omega_s + \omega_i = \omega_p \\ &\omega_s = 2\pi c q_s/\mathrm{OPL}(d, \omega_s, T) \\ &\omega_i = 2\pi c q_i/\mathrm{OPL}(d, \omega_i, T) \end{aligned} \qquad [7]$$

Experimentally, mode-hop-free oscillation exceeding 10 h was achieved by such active stabilization.

Another system parameter may then be used to tune both signal and idler frequencies. The tuning coefficient for the crystal temperature T is small, but tuning the pump frequency is an efficient way. The tuning coefficient is approximately $\Delta\omega_s = (\omega_s/\omega_p)\Delta\omega_p$, and accordingly for ω_i.

Continuous tuning of signal and idler frequencies over 10 GHz by pump tuning was achieved. This range was limited by the available continuous tuning range of the pump laser. Mode-hop tuning can be induced by mechanical perturbations: by tapping on the DRO cavity, mode-hop tuning over 220 GHz has been demonstrated.

Linewidth and Frequency Stability

For high-resolution spectroscopy a narrow linewidth for signal or idler is desirable. One contribution to the linewidth of a cwOPO arises from the unavoidable spontaneous parametric fluorescence, similar to the Schawlow–Townes limit in lasers due to spontaneous emission. This contribution is negligible if the OPO operates at power levels of practical interest. Thus, other sources of classical noise dominate the linewidth of cwOPOs. As discussed above, the frequency of cwOPOs depends (on a fine scale) on pump frequency and cavity optical path length. Both parameters fluctuate: the pump laser has a finite linewidth as well as frequency jitter, and the cavity optical path length fluctuates due to acoustic noise affecting the mirror positions, to air pressure fluctuations, temperature fluctuations of the crystal and of the cavity structure.

With careful construction and shielding, many of these noise sources can be minimized. Furthermore, it is highly favorable to use pump lasers of narrow linewidth and high intrinsic frequency stability (e.g., diode-pumped monolithic solid-state lasers), or actively frequency-stabilized lasers. Note that this special requirement on the pump source does not exist for lasers.

In an SRO, the linewidth and frequency stability of the signal are mostly determined by the stability properties of the cavity. The idler spectral properties then follow entirely from the signal and pump spectral properties through $\omega_i = \omega_p - \omega_s$. Since the idler is usually the wave of interest, this implies that the SRO cavity and the pump laser must both have a stability appropriate to the spectroscopic application.

If the SRO cavity is not sufficiently stable (e.g., because of insufficient temperature stability of the oven containing the nonlinear crystal), the frequency fluctuation/drift of the signal and idler can be substantial, tens to hundreds of MHz within minutes being typical.

To avoid mode-hops also on long time-scales, the cavity length (or the etalon angle) can be actively stabilized so as to maximize the idler's output power. This ensures that the cavity mode position and the frequency of maximum difference between gain and loss are kept equal.

If a frequency-stable pump wave is used, it is possible to transfer these characteristics by stabilizing the SRO cavity length to the pump wave frequency. In an SRO without resonant pump one can make use of the residual reflection of the pump wave by the cavity mirrors that leads to interference with the pump wave reflected from the pump input mirror. This was implemented in a SRO pumped by a high-power diode-pumped internally doubled Nd:YVO$_4$ laser (532 nm) of high intrinsic frequency stability. Signal and idler drifts of less than 50 MHz/h were achieved. At the same time, a very small 20 kHz linewidth of the (resonant) signal radiation was obtained. Similar levels were obtained for a PR-SRO with a common cavity for pump and signal wave.

In a PR-SRO with dual-cavity configuration, the same requirements of minimization of disturbances to the parametric wave's cavity arise as in a basic SRO. Using a highly frequency-stable kHz linewidth monolithic pump laser, together with the above-mentioned idler power maximization control, frequency drift values for both parametric waves at the level of 30 MHz/h were obtained and a short-term linewidth below 6 kHz.

In a single-cavity DRO, the linewidth and frequency stability of the parametric waves are essentially determined by the pump wave properties (see discussion in the section on DRO above). Here, too, the use of narrow-linewidth, high-frequency-stability pump lasers has led to excellent signal/idler characteristics. For example, less than 40 kHz linewidth and <40 MHz per hour drift were obtained with a DRO pumped by an externally doubled monolithic Nd:YAG laser of 10 kHz linewidth.

If further narrowing of linewidth or reduction of frequency drift is desired, the OPO output frequencies can be stabilized to a stable cavity, atomic, or molecular transition. Locking to a cavity has been demonstrated for a DRO.

Applications

CwOPOs are most interesting as sources of radiation in the IR range beyond 2 μm, especially in the 3–5 μm range, where vibrational transitions of many molecules occur. In comparison to other sources of coherent radiation (color center laser, quantum cascade laser, lead-salt diode laser, CO overtone laser, difference frequency generation), cwOPOs provide a unique combination of desirable features.

A simple application of cwOPOs would be as a source providing a power level/output wavelength/mode quality combination not available from other sources. Possible applications are vibrational spectroscopy in the condensed phase, where a narrow linewidth is not required. The excellent beam quality provides a high spatial resolution.

Concerning high-resolution spectroscopy (mostly in the gas phase), methods demonstrated in conjunction with cwOPOs include:

- Absorption spectroscopy;
- Doppler-free saturation spectroscopy;
- Photo-acoustic spectroscopy (PAS);
- Hole-burning spectroscopy;
- Coherent atomic spectroscopy;
- Cw cavity ring down spectroscopy (cw-CRDS).

Trace gas detection using PAS or cw-CRDS is a particular application where cwOPOs have reached excellent sensitivity (minimum detectable concentrations as low as 1 part in 10^{11}). The high output power and/or the narrow linewidth are specific features that make such sensitivity possible.

Quantum-Optical Properties

The pairwise production of signal and idler photons is at the heart of the parametric process. This perfect quantum correlation is to a certain extent also maintained in the output waves of a cwOPO. In particular, the intensity fluctuations of signal and idler waves emitted by a DRO are strongly (albeit never perfectly) quantum correlated if their output coupler transmissions significantly exceed their respective roundtrip losses. For these ‘twin beams’ the spectral density of the fluctuations of their power difference can be less than the spectral density of the fluctuations of the power of an individual wave (squeezing). Making use of this correlation permits to increase the obtainable sensitivity in spectroscopic measurements for a given power level of the wave interacting with the sample.

Device Development Trends

CwOPOs using solid-state pump lasers have become commercially available. In the near future, some likely developments will be miniaturization and optimization in order to reduce the influence of perturbations, further extension of continuous and mode-hop tuning ranges and the wider use of diode lasers as pump lasers, leading to cost reduction. To access a larger range of molecules in spectroscopy, extension of the emission range well beyond 5 μm is also of significant interest. Novel nonlinear materials and pump lasers of wavelength longer than the often employed Nd:YAG lasers will become of importance in this development.

Two specific lines of development complementary to the bulk cwOPOs described above are waveguide and intracavity cwOPOs (ICSRO). If pump, signal and idler waves are confined in a waveguide (*see* **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides) the parametric gain coefficient is a few orders higher than for free-space Gaussian waves, due to the small mode cross-section and the scaling with the square of the crystal length, compared with the linear scaling in case of Gaussian beams (see eqn [36]). Values are on the order of $E \sim 1/\mathrm{W}$ at $\lambda_p = 1.5\ \mu\mathrm{m}$ leading to low thresholds even for SROs, with pump lasers of comparatively long wavelength. For example, in a SRO consisting of a 9 cm long titanium:PPLN waveguide, a threshold below 300 mW for a 1.5 μm pump was achieved. The idler emission range was 3.1–3.4 μm. The potential for mass production and, therefore, cost minimization is clear.

An ICSRO consists of an SRO internal to the cavity of a cw laser, where the high circulating laser power is taken advantage of for pumping. An ICSRO is thus similar to a PR-SRO but there is no need to implement a control system for the frequency lock between pump laser and pump cavity nor to mode-match the pump wave to the SRO cavity mode.

Appendix A

Three-Wave Mixing in a $\chi^{(2)}$ Medium with Focused Waves

The wave equation in a nonlinear medium reads

$$\Delta\vec{E} - \mu_0\mu_r\frac{\partial^2}{\partial t^2}\vec{D}_{\mathrm{L}} = \mu_0\mu_r\frac{\partial^2}{\partial t^2}\vec{P}_{\mathrm{NL}} \qquad [8]$$

The displacement field $\vec{D}_{\mathrm{L}}$ is related to the electric field $\vec{E}$ by linear response. $\vec{P}_{\mathrm{NL}}$ is the nonlinear polarization. We deal with nonmagnetic media ($\mu_r = 1$) and choose the principal axis system as coordinate system. The Fourier transform of, for example, the x-component of the wave equation then reads:

$$\begin{aligned}&\Delta E_x(\omega,\vec{r}) + \mu_0\omega^2\varepsilon_0\varepsilon_x(\omega)E_x(\omega,\vec{r})\\ &\quad = -\mu_0\omega^2 P_{\mathrm{NL},x}(\omega,\vec{r})\end{aligned} \qquad [9]$$

In the following, we take into account only second-order nonlinear effects, so the nonlinear polarization reduces to $P_{\mathrm{NL}} = P^{(2)}$. The fields to be considered are monochromatic, propagate in the z-direction and are linearly polarized along x. The notation can then be simplified to:

$$E_x(\omega,\vec{r}) = E(\vec{r})\mathrm{e}^{\mathrm{i}kz}, \quad P_{\mathrm{NL},x}(\omega,\vec{r}) = P^{(2)}_{\omega}(\vec{r}) \qquad [10]$$

with the wavevector $k = \omega n(\omega)/c$ and the refractive index for the polarization under consideration $n_\omega^2 = n(\omega)^2 = E_x(\omega)$. The electric field is obtained as $\vec{E}(t,\vec{r}) = \frac{1}{2}\mathrm{Re}(E(\vec{r})\mathrm{e}^{\mathrm{i}(kz-\omega t)})\hat{x}$.

Since the parametric gains are small in cw conversion, the derivative of the envelope $E(\vec{r})$ changes only little over a propagation distance on the order of a wavelength $2\pi/k$. Inserting eqn [9] into eqns [6] and [8] and neglecting the term $\sim \mathrm{d}^2E(\vec{r})/\mathrm{d}z^2$, yields the paraxial or slowly varying envelope wave equation:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + 2\mathrm{i}k\frac{\partial}{\partial z}\right)E(\vec{r}) = -\mu_0\omega^2 P^{(2)}_{\omega}(\vec{r})\mathrm{e}^{-\mathrm{i}kz} \qquad [11]$$

The electric field envelope can be expanded in a complete set of Gaussian TEM$_{mn}$ – mode functions ψ_{mn}:

$$E(\vec{r}) = \sqrt{\frac{\omega}{n_\omega}}\sum_{m,n=0}^{\infty} A_{mn}(z)\psi_{mn}(\vec{r}) \qquad [12]$$

These mode functions are defined to satisfy the paraxial wave equation in absence of nonlinearity (vanishing right-hand side in eqn [11]) and the orthonormality properties

$$\iint_{-\infty}^{+\infty}\psi^*_{mn}(\vec{r})\psi_{m'n'}(\vec{r})\mathrm{d}x\,\mathrm{d}y = \delta_{mm'}\delta_{nn'} \qquad [13]$$

A particular mode function set is characterized by position (along z) and size of the waist, which can be chosen arbitrarily. The functions $A_{mn}(z)$ are slowly varying amplitude functions. The power in a particular mode can be calculated from:

$$P_{mn}(z) = \frac{\omega|A_{mn}(z)|^2}{2\mu_0 c} \qquad [14]$$

where c is the speed of light *in vacuo*. Inserting eqn [12] into eqn [11], multiplication with $\psi^*_{m'n'}(\vec{r})$, and integration over x, y yields:

$$\frac{\mathrm{d}A_{m'n'}(z)}{\mathrm{d}z} = \mathrm{i}\frac{\mu_0 c}{2}\sqrt{\frac{\omega}{n_\omega}}\mathrm{e}^{-\mathrm{i}kz}\iint_{-\infty}^{\infty} P^{(2)}_{\omega}(\vec{r})\psi^*_{m'n'}(\vec{r})\mathrm{d}x\,\mathrm{d}y \qquad [15]$$

If all electric fields are linearly polarized, the Fourier amplitudes of the nonlinear polarization can be written as:

$$\begin{aligned}P^{(2)}_{\omega_{\mathrm{s}}}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_{\mathrm{p}}}(\vec{r})E^*_{\omega_{\mathrm{i}}}(\vec{r})\\ P^{(2)}_{\omega_{\mathrm{i}}}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_{\mathrm{p}}}(\vec{r})E^*_{\omega_{\mathrm{s}}}(\vec{r})\\ P^{(2)}_{\omega_{\mathrm{p}}}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_{\mathrm{s}}}(\vec{r})E_{\omega_{\mathrm{i}}}(\vec{r})\end{aligned} \qquad [16]$$

where d is the tensor element appropriate to the particular combination of linear polarizations of the three waves. (In eqn [16], $\omega_{\mathrm{i}} + \omega_{\mathrm{s}} = \omega_{\mathrm{p}}$.)

The mode amplitude equations for the three interacting waves follow from eqns [10], [15], and [16] as:

$$\begin{aligned}\frac{\mathrm{d}A_{mn}^{\left\{\begin{smallmatrix}\mathrm{p}\\ \mathrm{s}\\ \mathrm{i}\end{smallmatrix}\right\}}(z)}{\mathrm{d}z} &= \mathrm{i}\frac{d(z)}{c}\left\{\begin{matrix}\sqrt{\omega_{\mathrm{p}}/n_{\mathrm{p}}}\mathrm{e}^{-\mathrm{i}\Delta kz}\\ \sqrt{\omega_{\mathrm{s}}/n_{\mathrm{s}}}\mathrm{e}^{+\mathrm{i}\Delta kz}\\ \sqrt{\omega_{\mathrm{i}}/n_{\mathrm{i}}}\mathrm{e}^{+\mathrm{i}\Delta kz}\end{matrix}\right\}\iint \mathrm{d}x\,\mathrm{d}y\\ &\quad\times\left\{\begin{matrix}\mathrm{E_s}(\vec{r})\mathrm{E_i}(\vec{r})\\ \mathrm{E_p}(\vec{r})\mathrm{E_i}(\vec{r})^*\\ \mathrm{E_p}(\vec{r})\mathrm{E_s}(\vec{r})^*\end{matrix}\right\}\psi_{mn}^{\left\{\begin{smallmatrix}\mathrm{p}\\ \mathrm{s}\\ \mathrm{i}\end{smallmatrix}\right\}}(\vec{r})^{*+\infty}\end{aligned} \qquad [17]$$

The quantity $\Delta k = k_{\mathrm{p}} - k_{\mathrm{s}} - k_{\mathrm{i}}$ is the wavevector mismatch. Here, the envelope of each wave has an expansion of the form in eqn [12] with its respective (independent) set of mode functions.

Appendix B

Steady-State Description of the cwOPO

The general eqns [17] are applicable to any type of cwOPO. Below and at threshold at least two of the three waves have given modes, i.e., the corresponding function $E(\vec{r})$ is known up to a slowly z-dependent amplitude: the resonant parametric wave has a mode defined by the cavity geometry and mirrors, while the pump wave mode is defined by the used focusing optics. This knowledge allows to obtain an analytic expression for the threshold of any OPO type and for an arbitrary cavity.

Above threshold two cases can be distinguished:

(i) in a basic SRO the pump mode itself becomes a function of pump power, due to its depletion by signal and idler wave, and only a numerical solution can be given; and
(ii) if two or more waves resonate in one or more cavities (DRO, PR-SRO, TRO), an analytic solution is possible even above threshold.

In the following, the PR-SRO is treated. Figure 4 shows a schematic of a common cavity PR-SRO and the notation; the treatment below also holds for the dual cavity PR-SRO. Without loss of generality, the signal wave is taken to be the signal. For the sake of simplicity, below a further assumption about the spatial mode of the idler will be made.

The pump and signal waves are given by TEM_{00} modes of the cavity (or cavities):

$$E_\mathrm{p}(\vec{r}) = \sqrt{\frac{\omega_\mathrm{p}}{n_\mathrm{p}}} A_{00\mathrm{p}}(z)\psi_{00\mathrm{p}}(\vec{r})$$
$$E_\mathrm{s}(\vec{r}) = \sqrt{\frac{\omega_\mathrm{s}}{n_\mathrm{l}}} A_{00\mathrm{s}}(z)\psi_{00\mathrm{s}}(\vec{r}) \qquad [18]$$

In the case of a common-cavity PR-SRO, pump and signal resonate in the same cavity and therefore their modes have the same waist position and Rayleigh ranges $z_\mathrm{R} = kw^2/2$, (w is the waist). In the case of a dual-cavity PR-SRO the modes are in principle independent.

Although the idler wave is in general not purely TEM_{00}, it may be approximated by a TEM_{00} mode, whose mode parameters are determined by an optimization criterion. We can then introduce the overlap function:

$$O(z) = \int\int \mathrm{d}x\, \mathrm{d}y (\psi_{00\mathrm{p}}(\vec{r}))^* \psi_{00\mathrm{s}}(\vec{r})\psi_{00\mathrm{i}}(\vec{r})^{+\infty}_{-\infty} \qquad [19]$$

Under these assumptions, the infinite set of mode eqns [17] reduces to just three coupled equations ($A \equiv A_{00}$):

$$\frac{\mathrm{d}A_\mathrm{p}(z)}{\mathrm{d}z} = -\mathrm{i}\xi(z)A_\mathrm{s}(z)A_\mathrm{i}(z)\mathrm{e}^{-\mathrm{i}\Delta kz}O(z) \qquad [20]$$

$$\frac{\mathrm{d}A_\mathrm{s}(z)}{\mathrm{d}z} = -\mathrm{i}\xi(z)A_\mathrm{p}(z)A_\mathrm{i}^*(z)\mathrm{e}^{\mathrm{i}\Delta kz}O(z)^* \qquad [21]$$

$$\frac{\mathrm{d}A_\mathrm{i}(z)}{\mathrm{d}z} = -\mathrm{i}\xi(z)A_\mathrm{p}(z)A_\mathrm{s}^*(z)\mathrm{e}^{\mathrm{i}\Delta kz}O(z)^* \qquad [22]$$

where

$$\xi(z) = \frac{d(z)}{c}\sqrt{\frac{\omega_\mathrm{p}\omega_\mathrm{s}\omega_\mathrm{i}}{n_\mathrm{p}n_\mathrm{s}n_\mathrm{i}}} \qquad [23]$$

Two phasematching cases have to be considered: for birefringent phasematching, $d = \mathrm{const}$ and ξ is independent of z. In QPM, the d-coefficient is modulated periodically along z with period Λ. It suffices to consider only the particular term of its Fourier series expansion that gives the smallest wavevector quasi-mismatch. When only the first term is relevant, this is referred to as first-order QPM and $d(z) \simeq \frac{1}{2}\hat{d}(g_1\mathrm{e}^{\mathrm{i}2\pi z/\Lambda} + \mathrm{c.c.})$. This case can be taken into account by replacing $d(z)$ in eqn [23] by $d_\mathrm{eff} = \frac{1}{2}\hat{d}g_1$, and in eqns [20]–[22] the relevant wavevector mismatch becomes $\Delta k = k_\mathrm{p} - k_\mathrm{s} - k_\mathrm{i} \pm 2\pi/\Lambda$, where, for usual materials, the minus sign applies.

We solve eqns [20]–[22] by expanding in powers of the nonlinearity ξ. This is appropriate due to the small gain and implies that the resonant pump and signal wave amplitudes do not change substantially (relative to the respective average amplitude) along the crystal length and around the cavity.

Integrating eqn [22] after approximating the amplitudes on the right-hand side by their values at $z = 0$ yields:

$$A_\mathrm{i}(z) = -\mathrm{i}\xi z A_\mathrm{p}(0)A_\mathrm{s}^*(0)I^*(z)$$
$$I(z) = \frac{1}{z}\int_0^z \mathrm{d}z' O(z')\mathrm{e}^{-\mathrm{i}\Delta kz'} \qquad [24]$$

We insert this result into eqns [20] and [21] and approximate $A_\mathrm{s}(z)$ and $A_\mathrm{p}(z)$ by their values at $z = 0$. The result is

$$A_\mathrm{p}(z) = A_\mathrm{p}(0) - \frac{1}{2}(\xi z)^2 D^*(z)A_\mathrm{p}(0)|A_\mathrm{s}(0)|^2 \qquad [25]$$

$$A_\mathrm{s}(z) = A_\mathrm{s}(0) + \frac{1}{2}(\xi z)^2 D(z)A_\mathrm{s}(0)|A_\mathrm{p}(0)|^2 \qquad [26]$$

$$D(z) = \frac{2}{z^2}\int_0^z dz' z' I(z') O(z')^* \, e^{i\Delta k z'} \qquad [27]$$

This completes the calculation of the propagation through the nonlinear medium. Now we proceed to calculate the propagation around the cavity.

Since the signal wave is resonant with the cavity, we can immediately write the self-consistency equation for the roundtrip:

$$|A_s(0)| = \sqrt{R_s(1-V_s)}|A_s(L)| \qquad [28]$$

The roundtrip cavity reflectivity and roundtrip loss for the signal wave are denoted by R_s and V_s. At the input mirror (see lower part of Figure 4), characterized by a transmission $T_p = 1 - R_p$, the interference of the injected pump wave $A_{p,in}$ and the internally circulating pump wave must be considered:

$$A_p(0) = \sqrt{T_p}A_{p,in} + \sqrt{R_p(1-V_p)}A_p(L)e^{i\phi} \qquad [29]$$

V_p is the pump roundtrip loss. Evaluating eqn [26] at $z = L$ and comparing with eqn [28] yields:

$$|A_p(0)|^2 = \left[\frac{1}{\sqrt{R_s(1-V_s)}} - 1\right]\frac{1}{\frac{1}{2}(\xi L)^2 \mathrm{Re}\, D(L)} \qquad [30]$$

Here the imaginary part of $D(z)$ was neglected since $|\xi L A_p(0)|^2 \ll 1$. Expression [30] shows that the intracavity pump power is clamped at a fixed value, equal to the value at threshold, independent of the input power.

In eqn [29] we can assume $A_{p,in}$ to be real without loss of generality, yielding the relation:

$$\sqrt{T_p}\frac{A_{p,in}}{A_p(0)} = 1 - \sqrt{R_p(1-V_p)}e^{i\phi}\frac{A_p(L)}{A_p(0)} \qquad [31]$$

where $\phi = \omega_p \mathrm{OPL}(\omega_p, T)/c$. The pump wave is assumed to be exactly resonant; this means that $A_p(0)$ is maximum. The phase ϕ must therefore be such that:

$$e^{i\phi}\frac{A_p(L)}{|A_p(0)|} \in \mathbb{R} \qquad [32]$$

Together, with eqn [25], we obtain:

$$\left|\frac{A_p(L)}{A_p(0)}\right|^2 = \left|1 - \frac{1}{2}(\xi L)^2 D^*(L)|A_s(0)|^2\right|^2 \qquad [33]$$

The right-hand side is evaluated neglecting the contribution proportional to ξ^4; inserting the result in eqns [31] and [32] gives:

$$\sqrt{T_p}\frac{A_{p,in}}{|A_p(0)|} = 1 - \sqrt{R_p(1-V_p)} + \sqrt{R_p(1-V_p)}\frac{1}{2}(\xi L)^2 \mathrm{Re}\, D|A_s(0)|^2 \qquad [34]$$

Equations [30] and [34] represent the steady-state solution of the PR-SRO: they give the circulating pump and signal powers, from which all desired quantities can be computed.

The external threshold is found by setting $A_s(0) = 0$ and assuming small losses, T_p, $V_p \ll 1$

$$P_{p,in}^{th} \simeq \frac{T_p + V_p}{T_p}\cdot(T_p + V_p)\cdot\frac{(T_s + V_s)}{E} \qquad [35]$$

with the nonlinearity

$$E = \frac{2\mu_0 d^2}{\pi c^2}\frac{\omega_s \omega_i}{n_p n_s n_i}L^2 \,\mathrm{Re}\, D(L) \qquad [36]$$

Note that $\mathrm{Re}\, D(L) = |I(L)|^2$. The second factor in eqn [35] is responsible for the reduction of threshold due to pump resonance. The last factor is the internal pump threshold and constant circulating pump power level:

$$P_p(0) \simeq \frac{(T_s + V_s)}{E} \qquad [37]$$

which is also equal to the threshold of a basic SRO. The nonlinearity may be written as

$$E \simeq \frac{4\mu_0 d^2}{\pi c^2}\frac{\omega_s^2 \omega_i^2}{n_p^2 \omega_p}L h_{PA} \qquad [38]$$

with the dimensionless focusing function h_{PA}. A detailed calculation shows that the value of h_{PA} cannot exceed 1.1 and is close to its maximum when the phase mismatch $\Delta k L$ is approximately zero and the pump and signal focusing by the cavity is such that their Rayleigh range $z_R \simeq L/2$.

The OPO will select the signal cavity mode frequency ω_s and idler spatial mode that minimizes the threshold eqn [35]. Note that this is not necessarily equivalent to a maximization of E, since T_s and V_s may also be dependent on signal frequency, for example, because of the presence of an etalon inside the cavity.

The circulating signal power as derived from eqns [30] and [34], is given in eqn [6].

The signal power extracted from the cavity is $P_{s,out} = T_s P_s(L) \simeq T_s P_s(0)$, while the (fully extracted) idler power is given by:

$$P_{i,out} = P_i(L) = \frac{\omega_i}{\omega_s}E P_s(0) P_p(0) \qquad [39]$$

See also

Nonlinear Optics, Applications: Phase Matching. **Optical Parametric Devices:** Overview.

Further Reading

Breitenbach G, Schiller S and Mlynek J (1995) 81% conversion efficiency in frequency-stable continuous-wave parametric oscillation. *Journal of the Optical Society of America B* 12: 2095–2101.

Ebrahimzadeh M and Dunn MH (2000) Optical parametric oscillators. In: Bass M, Enoch JM, van Stryland EW and Wolfe WL (eds) *Handbook of Optics IV*, pp. 2201–2272. New York: McGraw-Hill.

Ebrahimzadeh M, Turnbull GA, Edwards TJ, Stothard DJM, Lindsay ID and Dunn MH (1999) *Journal of the Optical Society of America B* 16: 1499–1511.

Gibson GM, Dunn MH and Padgett MJ (1998) Application of a continuously tunable, cw optical parametric oscillator for high-resolution spectroscopy. *Optics Letters* 22: 40–43.

Klein ME, Laue CK, Lee D-H, Boller K-J and Wallenstein R (2000) Diode-pumped singly resonant continuous-wave optical parametric oscillator with wide continuous tuning of the near-infrared idler wave. *Optics Letters* 25: 490–493.

Klein ME, Gross P, Boller K-J, *et al.* (2003) Rapidly tunable continuous-wave optical parametric oscillator pumped by a fiber laser. *Optics Letters* 28: 920–923.

Müller F, Popp A, Kühnemann F and Schiller S (2003) Transportable, highly sensitive photoacoustic spectrometer based on a continuous-wave dual-cavity optical parametric oscillator. *Optics Express* 11: 2820–2825.

Schiller S, Schneider K and Mlynek J (1999) *Journal of the Optical Society of America B* 16: 1512–1524.

Schreiber G, Hofmann D and Grundkötter W, *et al.* (2001) Nonlinear integrated optical frequency converters with periodically poled Ti:LiNbO$_3$ waveguides. *Proceeding SPIE* 4277: 144.

Special issue (1998) Continuous-wave optical parametric oscillators: materials, devices, and applications. *Applied Physics B* 66(6).

Strössner U, Meyn J-P, Wallenstein R, *et al.* (2002) Single-frequency continuous-wave optical parametric oscillator system with an ultrawide tuning range of 550 to 2830 nm. *Journal of the Optical Society of America B* 19: 1419–1424.

Yariv A and Yeh P (2002) *Optical Waves in Crystals*. Hoboken, NJ: Wiley-Interscience.

Optical Parametric Oscillators (Pulsed)

H Giessen and X P Zhang, University of Bonn, Bonn, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Advantages and Applications of Ultrashort Pulsed OPOs

The advantages of ultrashort pulsed optical parametric oscillators (OPOs) are twofold: on the one hand, it is possible to generate tunable laser radiation with a pulse duration in the femtosecond and picosecond time domain; on the other hand, the short pulses give rise to a high peak intensity of the electric field of the pump wave inside the nonlinear medium. Due to the fact that the OPO single-path gain coefficient reads as:

$$\eta = \frac{P_{\text{Signal}}}{P_{\text{Pump}}} = \frac{8\pi^2 d_{\text{eff}}^2 L^2 I_{\text{P}}}{\varepsilon_0 n_{\text{P}} n_{\text{S}} n_{\text{I}} c \lambda_{\text{S}}^2} \frac{\sin^2(\Delta k L/2)}{(\Delta k L/2)^2} \qquad [1]$$

where I_{P} is the pump intensity, Δk is the phase mismatch between pump, signal, and idler, n_{P}, n_{S}, n_{I}, are the refractive indices at the pump, signal, and idler wavelengths, respectively, and L is the length of the nonlinear medium with an effective nonlinear coefficient d_{eff}, the efficiency grows proportional to the intensity of the pump.

Ultrashort pulsed OPOs can provide widely tunable femtosecond or picosecond pulses, covering the spectral range from the visible to the mid-infrared, which have many applications, such as time-resolved spectroscopy, pump-probe measurements, semiconductor analysis, photochemistry, optical ranging, data transfer, and data processing with high carrier frequency, white-light generation, etc.

Synchronous Pumping

The key issue in operating ultrashort-pulse OPOs is synchronization between the pump pulses and the signal pulses. Consider the following sequence: a pump pulse enters the OPO cavity and generates spontaneously downconverted light at the signal wavelength (see Figure 1, inset). These signal photons travel inside the OPO cavity, entering the nonlinear medium again after exactly one cavity round trip. This light will act as seed light for another downconversion process if a second pump pulse enters the nonlinear medium at exactly the same instance (see Figure 1). The presence of the signal pulse will lead to a much more efficient stimulated downconversion process.

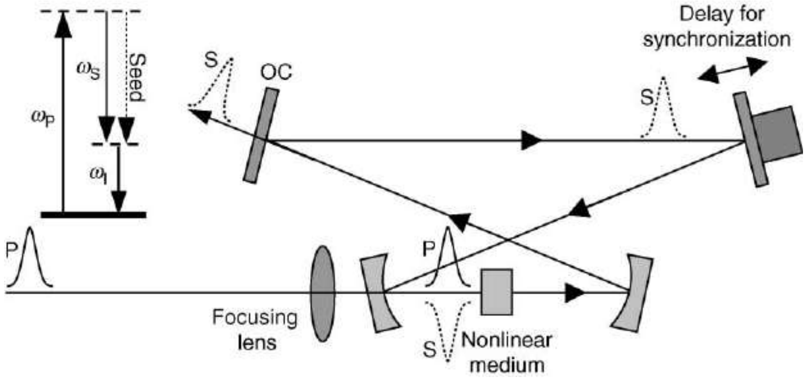

Figure 1 Synchronous pumping of a typical ring-cavity OPO: P, pump pulses; S, signal pulses; OC, output coupler; ω_P, ω_S, ω_I, frequencies of the pump, the signal, and the idler, respectively. Inset: energy level diagram for optical parametric oscillation.

This requires the repetition rate of the pump oscillator and the OPO to be synchronized precisely. Therefore, this pump scheme is called *synchronized pumping*. If the temporal separation between the pump pulses does not exactly equal the roundtrip time of the signal pulses inside the OPO cavity, the generation efficiency will decrease dramatically as the temporal pulse overlap is reduced.

Dispersion in Nonlinear Media and Optical Glasses

Maintaining synchronicity between the pump pulses and the signal pulses is a necessary, but not sufficient prerequisite for successful operation of an ultrashort-pulse OPO. The signal pulses will be temporally broadened during their repeated trip through the nonlinear medium. This will lead to problems, as the newly generated signal pulses are still rather short (as they have been generated by a short pump pulse), but the previously generated signal pulses, which should act as seed pulses and ideally have a similar pulse duration as the pump pulses, have already been broadened temporally, due to dispersion in the nonlinear medium and other intracavity optical elements. Meanwhile, temporal walk-off between the signal (idler) and the pump pulses will take place, resulting in reduced conversion efficiency and broadened signal pulses. Group velocity dispersion (GVD) and group velocity mismatch (GVM) are normally used to characterize the dispersion in nonlinear medium.

The group velocity of the laser pulses propagating in a nonlinear optical medium is defined as:

$$v_G = \frac{c}{n - \lambda \dfrac{dn}{d\lambda}} \qquad [2]$$

where c is the velocity of light in vacuum, n is the refractive index of the nonlinear medium at the center wavelength λ of the laser pulses. Then the GVD can be written as:

$$\frac{dv_G}{d\lambda} = \frac{\lambda v_G^2}{c} \frac{d^2 n}{d\lambda^2} \qquad [3]$$

In most cases, we evaluate the GVD of the optical materials using the group velocity dispersion parameter (k''), which is defined as:

$$k'' = \frac{d^2 k}{d\omega^2} = \frac{\lambda^3}{2\pi c^2} \frac{d^2 n}{d\lambda^2} \qquad [4]$$

where $k = n\omega/c$ is the wave vector, ω is the frequency of the incident wave, and consequently:

$$\frac{dv_G}{d\lambda} = \frac{2\pi c v_G^2}{\lambda^2} \frac{d^2 k}{d\omega^2} \qquad [5]$$

Due to GVD in the nonlinear medium, the spectral components in the broadband spectrum of an ultrashort pulse will travel at different group velocities and will consequently be delayed differently in the crystal, therefore the pulses will become temporally broadened or become chirped. This linear chirp induced pulse broadening effect can be characterized quantitatively by the following formula:

$$\tau_P = \tau_{P0} \sqrt{1 + \left(\frac{z}{z_d} \right)^2} \qquad [6]$$

where τ_{P0} is the incident pulse duration, τ_P is the pulse duration after the pulse propagates in the nonlinear medium over a distance of z, and z_d is called the characteristic length and defined as:

$$z_d = \frac{\tau_{P0}^2}{4 \ln 2 |k''|} \qquad [7]$$

In frequency conversion processes, such as second-harmonic generation (SHG), sum-frequency generation (SFG), difference frequency generation (DFG), optical parametric oscillation (OPO), or optical parametric amplification (OPA), three waves interact with each other while propagating through the nonlinear medium. Due to dispersion of the material, different waves propagate at different group velocities. This effect is generally called group velocity mismatch. In cases of interaction on the femtosecond scale, the interacting pulses may get separated after propagating some distance in the medium, which means that there is a reduced effective interaction length. The GVM is a quantitative evaluation of this effect, which is defined as:

$$\Delta v_{\mathrm{G}} = \left(\frac{1}{v_{\mathrm{G,i}}} - \frac{1}{v_{\mathrm{G,j}}}\right)^{-1} \qquad [8]$$

where $v_{\mathrm{G,i}}$ and $v_{\mathrm{G,j}}$ are the group velocities of the two interacting waves i and j. The effective interaction length L_{eff} for wave i and wave j can thus be calculated by the following formula:

$$L_{\mathrm{eff}} = \tau_{\mathrm{P}}|\Delta v_{\mathrm{G}}| = \tau_{\mathrm{P}}\left|\left(\frac{1}{v_{\mathrm{G,i}}} - \frac{1}{v_{\mathrm{G,j}}}\right)^{-1}\right| \qquad [9]$$

where τ_{P} is the incident pulse duration at full width at half maximum (FWHM). It is assumed that wave i and wave j have a similar pulse duration. This temporal walk-off will lead to two effects. First, the interacting pulses will get separated and the conversion efficiency will be reduced. The second effect is the broadening of the signal pulses. In the case of sub-100 fs operation, to restrict the broadening in pulse length to less than ~10%, the crystal length should be smaller than $L_{\mathrm{eff}}/2$.

Figure 2 shows the GVD of some typical OPO crystals, including KTP, KTA, RTA, CTA, PPLN, PPKTP, and LBO. Their Sellmeier equations are as summarized in Appendix A. In addition to the wavelength of the incident beam, the GVD is also dependent on the polarization of the incident beams and the type of phase matching.

We give the GVM values between the pump and the signal (P–S), the signal and the idler (S–I), and between the pump and the idler (P–I) for a KTP crystal in Figure 3.

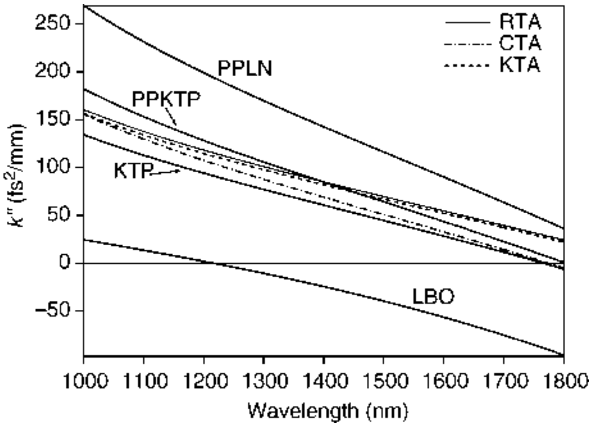

Figure 2 Group velocity dispersion characteristics of some typical OPO crystals.

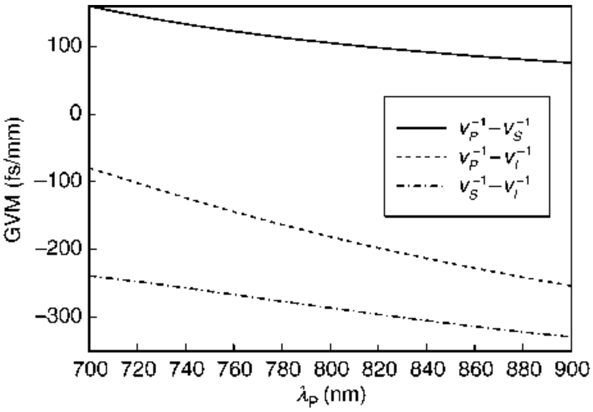

Figure 3 GVM between the interaction waves of a KTP OPO versus pump wavelength (λ_{P}).

Intracavity Dispersion Compensation and Extra Cavity Pulse Compression

In order to tackle the problem of intracavity dispersion, corrective elements have to be placed inside the cavity in order to compensate the chirp. There have been several solutions to this problem, namely the combination of intracavity prism pairs, chirped mirrors, chirped grating periods (for quasi-phase matching) with oppositely chirped pump pulses, and glass plate dispersion compensators. Certainly, a shorter crystal length will yield a smaller GVM and GVD, but will also give a smaller nonlinear conversion efficiency due to the shorter crystal length (see eqn [1]).

Pump Sources

As pump pulses, mode-locked femtosecond or picosecond lasers are used. The requirements on pulse durations are determined by the desired pulse durations of the OPO signal and by the phase-matching bandwidth of the correspondingly used crystal. Shorter pump pulses normally lead to higher pump power thresholds due to shorter temporal overlap between the interacting pulses. It can also lead to higher instability of the OPO operation due to its higher sensitivity to the cavity length fluctuations. Generally, larger pump bandwidth means larger bandwidth of the signal. Consequently, a potentially shorter signal pulse duration can be achieved if the intracavity dispersion is perfectly compensated. Sometimes it is advantageous for stability to operate the OPO in the slightly negative total GVD regime. In a quasi-phase-matched (QPM) OPO, the phase-matching bandwidth is so large that only a narrow pump bandwidth is required to phase-match a large bandwidth of the signal. Therefore, a pump depletion hole can be commonly observed with the oscillation of the OPO.

Design of Ultrashort Pulsed OPOs

Design of an OPO device may involve the design of the cavity length, cavity mirrors, focal length for focusing the pump beam, intracavity dispersion compensation scheme, frequency stabilization module, crystal length, and some other design depending on the special configuration of the OPO, for example, a heating module (an oven) is required for a PPLN OPO and temperature control module is required for a temperature phase-matched LBO OPO. For an ultrashort pulsed OPO, the most important designs include the design of the crystal length, the design of intracavity dispersion compensation components, and the frequency stabilization control.

Crystal Length

Several factors have to be taken into account when designing a crystal length. The design of a crystal length should be a compromise of all of these considerations. What is special for a femtosecond or picosecond OPO is the dispersion and temporal walk-off between the interacting pulses in the crystal:

1. Effective interaction length: Pump pulse duration, estimated signal pulse duration, temporal walk-off between the pump and the signal give a limit of the crystal length that is characterized by eqn [9].
2. Pulse broadening due to GVM and GVD from the OPO crystal and other intracavity optical glass elements should be estimated. It is reasonable that shorter crystal length is always favorable for ultrashort pulse generation. The crystal length should be as short as possible while maintaining an acceptable conversion efficiency. When designing the crystal length, the available dispersion compensation device should be considered simultaneously.

Dispersion Compensation

For a crystal with a given material and length, the intracavity dispersion can be calculated. The compensation device should introduce an amount of negative GVD such that the total intracavity GVD is equal to or slightly smaller than zero at the desired center wavelength of the signal pulses. Meanwhile, the compensation should be relatively uniform (for instance having a GVD fluctuation of $<50\ \mathrm{fs}^2/\mathrm{mm}$) over the bandwidth of the expected signal spectrum. Then a compensation method (as described above) should be chosen and corresponding device should be designed.

Piezo-Driven Mirror and Frequency Stabilization

To achieve fine tuning of the cavity length, one of the cavity mirrors should be mounted onto a piezoelectric translation stage. When feeding back the signal output of the OPO to this piezo-driven mirror, forming a closed loop to compensate for the drift of the cavity length, a stabilization of the OPO repetition rate is achieved. If the signal is dispersed by a grating, a change in cavity length will also lead to a change in the output spectrum. Using a split photodiode, the difference signal of the dispersed spectrum can be utilized as the measured signal in the feedback loop.

Tuning Characteristics and Methods

Tunability is one of the most important advantages of the OPO over other ultrashort pulsed lasers. A large variety of tuning methods make the OPO more practical and versatile in many applications. For birefringently phase-matched OPOs, the tuning of the signal wavelength can be achieved through changing the pump wavelength (pump wavelength tuning), using an intracavity wavelength selection element (such as a birefringent filter), changing the temperature of the crystal when its refractive index is sensitive to the temperature change (temperature tuning), or changing the orientation of the OPO crystal (angle tuning). Some more methods can be utilized for a QPM OPO, for example through changing the poling period of the crystal (grating period tuning). A special method for an ultrashort pulsed OPO is the so-called cavity length tuning.

As has been discussed above, in synchronous pumping, the OPO must have the same cavity length as the pump laser, so that the signal pulse arrives simultaneously with the pump pulse at the crystal to get the most efficient interaction. Because of dispersion of the nonlinear crystal, the signal will be delayed differently at different wavelengths. In other words, signal pulses have different round-trip times in the cavity at different center wavelengths. When the cavity length of the OPO is changed slightly, the signal has to change its wavelength to keep constant optical path length or constant round-trip time. Cavity length tuning can be achieved easily by tuning the voltage on a piezoelectric transducer (PZT), onto which one of the cavity mirrors is mounted. Therefore, this method is the most common and frequently used to tune the signal wavelength of a synchronously pumped OPO. This is also the basis for stabilizing the operation of a synchronously pumped OPO.

High-Repetition-Rate (HRR) Operation

Synchronously pumping an OPO at high repetition rate produces a high-repetition-rate signal pulse train which extends the application fields of the OPO. Using high-repetition-rate laser pulses (e.g., >1 GHz) as the pump is a straightforward method to run the OPO at a high repetition rate. But a high repetition rate means an extremely short cavity length. For instance, the cavity length is about 30 cm at 1 GHz repetition rate. Such a short cavity length leads to a large mode of the signal beam waist in the crystal. Mode matching between the signal and the pump beams requires that the transverse mode of the pump beam should be slightly larger than that of the signal beam within the interaction zone to maintain stable and efficient oscillation of the OPO. Large modes of the signal and the pump beams mean low interacting intensities, increasing the pump threshold significantly. Using smaller curvature radius of the curved mirrors and shorter focal length of the focusing lens might reduce the mode sizes at the focus in the nonlinear medium, but this will introduce extremely serious astigmatism when the signal beam is reflected from a strongly tilted curved mirror with a quite small curvature radius. In addition to exact synchronous pumping, there are at least two methods to realize high-repetition-rate operation of an OPO. First, pumping at low repetition rate while using a shorter cavity length of the OPO (L_{OPO}) than the pump laser (L_{P}) with $L_{\mathrm{OPO}} = L_{\mathrm{P}}m/N$, where m and N are both integers and $m < N$, as demonstrated in Figure 4a, which is called lower-order synchronous pumping. However, applying this technique, the signal pulses experience more round-trip losses, also resulting in a much higher pump threshold and periodically varying intensities of the output pulses. Second, pumping the OPO at high repetition rate while using a much longer cavity length of the OPO than the pump laser with $L_{\mathrm{OPO}} = L_{\mathrm{P}}N/m$, where m and N are both integers and $m < N$, is called higher-order pumping, as demonstrated in Figure 4b. In higher-order synchronous pumping, the signal pulse train has the same repetition rate as the pump, while the mode-matching problem is removed through a long OPO cavity.

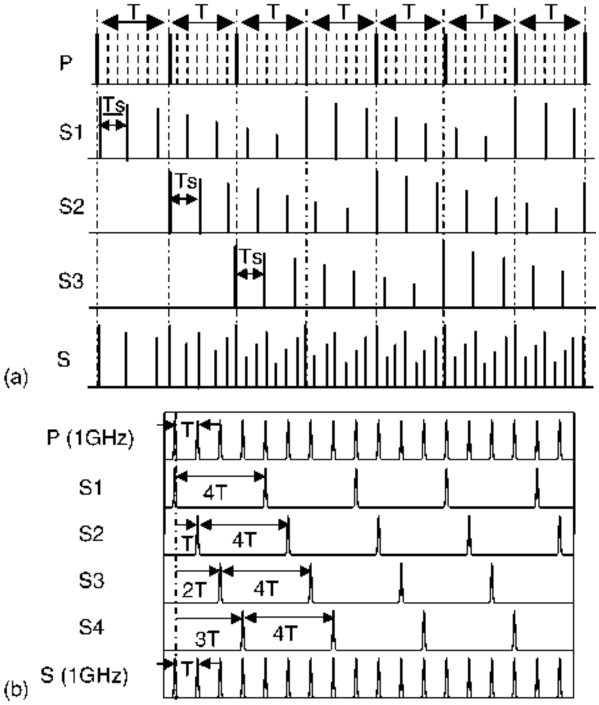

Figure 4 (a) 3/7-order synchronous pumping. Three signal pulse trains (S1–S3) are produced, each having a repetition period (Ts) 3/7 that of the pump pulse train (P): Ts = 3/7*T*. The pump repetition rate is multiplied by a factor of 7 in the signal pulse train (S). (b) Fourth-order synchronous pumping at 1 GHz. The signal pulse train (S) is a combination of four pulse trains (S1–S4), each having a repetition period (4*T*) 4 times that of the pump pulse train (P).

See also

Optical Parametric Devices: Optical Parametric Oscillators (Continuous Wave); Overview.

Appendix A

Sellmeier Equations for some Typical Crystals and Optical Glasses

In the equations, n denotes the refractive index; λ is the wavelength of incident light in μm; $A, B, C, D, A_1, A_2, A_3, A_4, A_5, A_6, B_1, B_2, B_3, B_4, \ldots$, are the Sellmeier coefficients; the subscripts X, Y, Z, stand for the directions of the principal axes of the biaxial crystals.

KTP

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	*B*	*C*	*D*
n_X	2.1146	0.89188	0.20861	0.01320
n_Y	2.1518	0.87862	0.21801	0.01327
n_Z	2.3136	1.00012	0.23831	0.01679

KTA

$$n^2 = A + \frac{B}{\lambda^2 - C} - D\lambda^2$$

	A	B	C	D
n_X	3.1413	0.04683	0.04055	0.01023
n_Y	3.1593	0.04828	0.04710	0.01049
n_Z	3.4435	0.06571	0.05435	0.01460

RTA

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	B	C	D
n_X	2.04207	1.17785	0.20157	0.01035
n_Y	2.14941	1.09267	0.21462	0.01067
n_Z	2.18962	1.30103	0.22809	0.01390

CTA

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	B	C	D
n_X	2.34498	1.04863	0.22044	0.01483
n_Y	2.74440	0.70733	0.26033	0.01526
n_Z	2.53666	1.10600	0.24988	0.01711

LiNbO$_3$

$$n_e^2 = A_1 + B_1 f + \frac{A_2 + B_2 f}{\lambda^2 - (A_3 + B_3 f)^2} + \frac{A_4 + B_4 f}{\lambda^2 - A_5^2} - A_6\lambda^2$$

$$f = (T - T_0)(T + T_0 + 546) \text{ and } T_0 = 24.5\,^\circ\text{C}$$

Coefficients	*Values*
A_1	5.35583
A_2	0.100473
A_3	0.20692
A_4	100
A_5	11.34927
A_6	1.5334×10^{-2}
B_1	4.629×10^{-7}
B_2	3.862×10^{-8}
B_3	-0.89×10^{-8}
B_4	2.657×10^{-5}

LBO

$$n^2 = A + \frac{B}{\lambda^2 + C} + D\lambda^2$$

	A	B	C	D
n_X	2.4543	1.1413×10^{-2}	-9.4981×10^{-3}	-1.3900×10^{-2}
n_Y	2.5382	1.2830×10^{-2}	-1.1387×10^{-2}	-1.7034×10^{-2}
n_Z	2.5854	1.3065×10^{-2}	-1.1617×10^{-2}	-1.8146×10^{-2}

$$\frac{dn_X}{dT} = 2.0342 \times 10^{-7} - 1.9697 \times 10^{-8} T - 1.4415 \times 10^{-11} T^2$$

$$\frac{dn_Y}{dT} = -1.0748 \times 10^{-5} - 7.1034 \times 10^{-8} T - 5.7387 \times 10^{-11} T^2$$

$$\frac{dn_Z}{dT} = -8.5998 \times 10^{-7} - 1.5476 \times 10^{-7} T + 9.4675 \times 10^{-10} T^2 - 2.2375 \times 10^{-12} T^3$$

BBO

$$n_o^2 = 1.7018379 + \frac{1.0357554\lambda^2}{\lambda^2 - 0.01800344} + \frac{1.2479989\lambda^2}{\lambda^2 - 91}$$

$$n_e^2 = 1.5920433 + \frac{0.7816893\lambda^2}{\lambda^2 - 0.016067891} + \frac{0.8403893\lambda^2}{\lambda^2 - 91}$$

Sapphire

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - B_1} + \frac{A_2\lambda^2}{\lambda^2 - B_2} + \frac{A_3\lambda^2}{\lambda^2 - B_3}$$

A_1	A_2	A_3	B_1	B_2	B_3
1.023798	1.058264	5.280792	0.00377588	0.0122544	321.3616

Quartz

$$n_{o,e}^2 = A + \frac{B\lambda^2}{\lambda^2 - C} + \frac{D\lambda^2}{\lambda^2 - E}$$

	A	B	C	D	E
n_o	1.28604141	1.07044083	0.0100585997	1.10202242	100
n_e	1.28851804	1.09509924	0.0102101864	1.15662475	100

Fused Silica, BK7, SF2, SF5, SF10, SF11, SF18

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - B_1} + \frac{A_2\lambda^2}{\lambda^2 - B_2} + \frac{A_3\lambda^2}{\lambda^2 - B_3}$$

	A_1	A_2	A_3	B_1	B_2	B_3
Fused silica	0.6961663	0.4079426	0.8974794	0.0046791	0.01351206	97.9340025
BK7	1.03961212	0.23179234	1.01046945	0.0060007	0.0200179	103.560653
SF2	1.40301821	0.23176750	0.93905659	0.0105796	0.04932270	112.405955
SF5	1.46141885	0.24771302	0.94999583	0.0118261	0.05085946	112.041888
SF10	1.61625977	0.25922933	1.07762317	0.0127535	0.05819840	116.607680
SF11	1.73848403	0.31116897	1.17490871	0.0136069	0.06159605	121.922711
SF18	1.56441436	0.29141358	0.96030788	0.0121864	0.05355680	111.451201

Further Reading

Agrawal GP (1989) *Nonlinear Fiber Optics*. Boston, San Diego: Academic Press.

Albrecht TF, Sandmann JHH, Feldmann J, Stolz W, Göbel EO, Hillmer H, Lösch R and Schlapp W (1995) Design and application of a femtosecond optical parametric oscillator for time-resolved spectroscopy of semiconductor heterostructures. *Applied Physics* B 60: 459.

Arbore MA, Marco O and Fejer MM (1997) Pulse compression during second-harmonic generation in aperiodic quasi-phase-matching gratings. *Optics Letters* 22: 865–867.

Arigas D and Reid DT (2002) Efficient femtosecond optical parameteric oscillators based on aperiodically poled nonlinear crystals. *Optics Letters* 27: 851–853.

Bass M, Enoch JM, Van Stryland EW and Wolfe WL (2001) *Handbook of Optics: Volume IV, Fiber Optics and Nonlinear Optics*. New York: McGraw-Hill.

Beddard T, Ebrahimzadeh M, Reid TD and Sibbett W (2000) Synchronously pumped optical parametric oscillator driven by a femtosecond mode-locked fiber laser. *Optics Letters* 25: 1052.

Bierlein JD and Vanherzeele H (1997) Potassium titanyl phosphate: properties and new applications. *Journal of the Optical Society of America B* 6: 622.

Burr KC, Tang CL, Arbore MA and Fejer MM (1997) High-repetition-rate femtosecond optical parametric oscillator based on periodically poled lithium niobate. *Applied Physics Letters* 70: 3341.

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena: Fundamentals, Techniques, and Applications on a Femtosecond Time Scale*. San Diego, CA: Academic Press.

Dmitriev VG, Gurzadyan GG and Nikogosyan DN (1991) *Handbook of Nonlinear Optical Crystals*. New York: Springer-Verlag.

Dunn MH and Ebrahimzadeh M (1999) Parametric generation of tunable light from continuous-wave to femtosecond pulses. *Science* 286: 1513.

Fenimore DL, Schepler KL, Zelmon D, Kuck S, Ramabadran, Von Richter P, and Small D (1996) Rubidium titanyl arsenate difference-frequency generation and validation of new Sellmeier coefficients. *Journal of the Optical Society of America B* 13: 1935.

Ghost G (1995) Temperature dispersion of refractive indices in β-BaB_2O_4 and LiB_3O_5 crystals for nonlinear optical devices. *Journal of Applied Physics* 78: 6752.

Ghost G (1999) Dispersion-equation coefficients for the refractive index and birefringence of calcite and quartz crystals. *Optics Communications* 163: 95.

Hebling J, Giessen H, Linden S and Kuhl J (1997) Mirror-dispersion-compensated femtosecond optical parametric oscillator. *Optical Communications* 141: 229.

Jundt DH (1997) Temperature-dependent Sellmeier equation for the index of refraction, ne, in congruent lithium niobate. *Optics Letters* 22: 1553.

Kato K, Umemura N and Tanaka E (1997) 90° phase-matched mid-infrared parametric oscillation in undoped $KTiOAsO_4$. *Japanese Journal of Applied Physics 2* 36(4A): L403.

Loza-Alvarez P, Reid DT, Faller P, Ebrahimzadeh M and Sibbett W (1999) Simultaneous femtosecond-pulse compression and second-harmonic generation in aperiodically poled KTiOPO4. *Optics Letters* 24: 1071.

Matuschek N, Gallmann L, Sutter DH, Steinmeyer G and Keller U (2000) Back-side-coated chirped mirrors with ultra-smooth broadband dispersion characteristics. *Applied Physics B* 71: 509.

Mokhtari A, Fini L and Chesony J (1987) Efficient frequency mixing of a cw femtosecond laser synchronously pumped by a frequency doubled Nd:YAG laser. *Optics Communications* 61: 421.

Nebel A, Fallnich C, Beigang R and Wallenstein R (1997) Noncritically phase-matched continuous-wave mode-locked singly resonant optical parametric oscillator synchronously pumped by a Ti:sapphire laser. *Journal of the Optical Society of America B* 10: 2195.

Schubert M and Wilhelmi B (1978) *Nonlinear Optics and Quantum Electronics*. New York: John Wiley & Sons.

Seres J and Hebling J (2000) Nonstationary theory of synchronously pumped femtosecond optical parametric oscillators. *Journal of the Optical Society of America B* 17: 741.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker Inc.

Szipöcs R and Kohazi-Kis A (1997) Theory and design of chirped dielectric laser mirrors. *Applied Physics B* 65: 115.

Tang Y, Cui Y and Dunn MH (1995) Thermal dependence of the principal refractive indices of lithium triborate. *Journal of the Optical Society of America B* 12: 638.

Tomaru T and Petek H (2000) Femtosecond Cr4+ YAG laser with an L-fold cavity operating at a 1.2-GHz repetition rate. *Optics Letters* 25: 584.

Tomaru T and Petek H (2000) Thermal dependence of the principal refractive indices of lithium triborate. *Journal of the Optical Society of America B* 18: 388.

Velsko SP, Webb M, Davis L and Huang C (1991) Phase-matched generation in lithium triborate (LBO). *IEEE Journal of Quantum Electronics* 27: 2182.

Yariv A (1984) *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: John Wiley & Sons.

Yariv A (1989) *Quantum Electronics*. New York: John Wiley & Sons.

Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: John Wiley & Sons.

Zhang XP, Hebling J, Kuhl J, Rühle WW, Palfalvi L, Giessen H (2002) Femtosecond near-IR optical parametric oscillator with efficient intracavity generation of visible light. *Journal of the Optical Society of America B* 19: 2479.

Zhang XP, Kuhl J, Hebling J, Bartels A, Nau D, Rühle XP, Giessen H (2002) 1-GHz repetition-rate femtosecond optical parametric oscillator. *Applied Physics Letters* 80: 1873.

OPTICAL PROCESSING SYSTEMS

D Arbel and N S Kopeika, Ben-Gurion University of the Negev, Beer Sheva, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modern optical systems find many applications in sensing, recording, storage, and transmission of data. These applications stimulate the development of many new optical devices, components, materials, and applied technologies. The basic element of an optical system is the light. The physical implications of the light in optical systems are very analogous to those of electricity, hence, light signals are quite similar to electric signals. High-speed spatial light components are being developed for more stringent applications in coherent optical processing systems, where they may be used in the input plane or in the Fourier transform plane. Research in this area is directed towards the design of robust and compact correlators.

Novel transforms for information processing (also known in electronics) include Fourier and wavelet transforms. Such optical transformations serve as the basic platform for linear optical system processing. Optical devices and systems for information processing include spatial light modulators, which are used to implement optical information processing systems. There are holographic applications in information processing, display systems, memory systems, optical computing systems, and imaging systems. Such devices and systems are based on applications of novel materials for information processing. Optical and digital techniques are used for optical system design, optical computing, image processing, and even for encryption and security systems.

The imaging system is an optical system and integrates hardware as well as software. Figure 1 presents a general model of an extensive imaging system. The starting point of the system is the object (or the light source). An imaging device, such as camera or optical sensor detects the energy radiated in electromagnetic waves from the light source. The image (or series of images – video) formed in the imaging device depends on the received irradiance of the image propagated through the imaging path and is correspondingly diminished from that in the object plane. The properties of the optical medium (transmission path) such as atmospheric properties, background light, motion, and vibration in both the object and the image planes, and the quality of the imaging device, all affect the quality of the images.

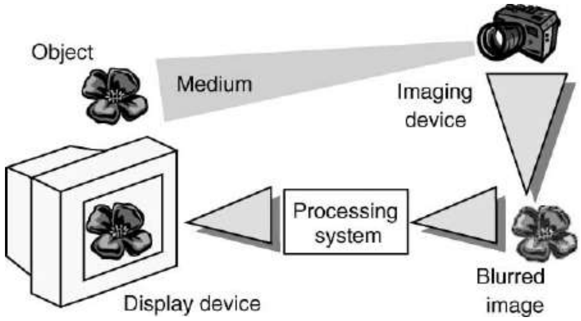

Figure 1 General description of an imaging system.

There are atmospheric phenomena that give rise to attenuation of the irradiance of the propagating image, thus reducing the contrast of the final image, as well as atmospheric phenomena that cause blurring of detail. Both types of phenomena degrade image quality.

Image blurring caused by vibration is a factor whose influence on resolution is often significant in imaging systems that involve mechanical motion. In order to obtain a well processed imaged, as similar as possible to the ideal image (the source), image processing techniques such as image enhancement, restoration, and filtering are implemented in an optical processing system which is a combination of software (algorithms) and hardware (electronic and optical components). The processing procedure involves mathematical and especially optical transformations. The output of the processing system is presented by using a display device or hardcopy. The development of imaging systems (imaging devices and displays) which include integrated optical components and processing units, assist in improving image quality at early stages, and as a result reduce the need of unique processing systems and may serve as the standalone device.

The key issues which must be addressed in the design, evaluation, and implementation of practical systems for optical processing, are provided here in a detailed review. Thus, the review is focused on imaging where considerations such as optical transformations, optical computing, effects of diffraction, scatter, absorption, and design on the performance of optical systems, imaging and the optical transfer function (OTF) for image restoration, and filtering, are all discussed in detail.

Digital Processing and Control

The Fourier transform properties of Fraunhofer diffraction permit data processing operations carried out optically that can, in many situations, be more advantageous than digital signal processing. The primary advantages of optical techniques over electronic analog and digital signal processing, are best realized when the information to be processed has two degrees of freedom. Optical systems correspondingly can have two degrees of freedom. Thus, for handling pictorial data, they can be intrinsically superior to electric or electronic systems, which have only time as an independent variable. This two-dimensional (2D) processing property can be particularly valuable in radar applications where previously velocity and position data had to be processed separately in order for sufficient accuracy in the estimation of each set of data to be obtained.

Many of the basic operations that can be performed by optical computers can also be done with electronic systems. However, if large quantities of wideband data are to be processed, the electronic techniques can often be prohibitively expensive and time consuming. The optical computer is faster than a digital computer because it operates on signals that are functions of positions, not time. The input to an optical data processor is a pattern of light irradiance that varies over some area, not light irradiance that varies as a function of time.

Thus, optical data processing can be very appropriate for such lengthy operations as the calculation of complicated Fourier transforms and the processing of side-looking radar signals. In addition to mathematical operations such as differentiation, integration, and producing Fourier transforms, the two classes of applications for which optical signal processing is most often used are filtering and correlation.

Optical Computing

Optical computing (OC) means the use of light as a primary means for carrying out numerical calculations, reasoning, artificial intelligence, etc. The history of OC is linked to that of radar systems. OC systems received a great push from the invention of laser in 1960. In the 1960s, the first schemes for all-optical digital computers were proposed. Since the 1990s, researchers have been pursuing the development of computers based on optical technology, and emphasis has shifted to optical interconnection of arrays of semiconductor smart pixels. OC provides the first in-depth review of the possibilities and limitations of optical data processing. The partial listing of those points that make OC appealing are direct image processing, massive parallelism and connectivity, immunity to electromagnetic interference (EMI), speed, size, and cost. There are three distinct trends in OC: special purpose analog optical systems, general purpose digital optical systems, and hybrid optical/electronic systems.

Innovations such as the electro-optic spatial filter, the 2D spatial light modulator, and the variable electro-optic mirror provide the fundamental building blocks for both analog optical computing (AOC) and digital optical computing (DOC) technologies. Optical techniques can provide solutions and a number of ways of extending the information processing capability of electronics. The optical processors have to be compatible with existing electronic systems. Free space digital optics is one direction that provides some valuable solution. DOC requires the use of nonlinear optics.

All-optical processing refers to optically based processors employed nonlinear optical resources either of liquid crystal spatial light modulators (SLMs) or nonlinear interference filters (NLIFs). Electro-absorption devices are a bridge between all-optical demonstrators and optically interconnected smart pixels (logic and local on-chip interconnection is electrical whereas chip to chip interconnection is optical). Optically interconnected smart pixels have several advantages like faster data acquisition and low power consumption. A major breakthrough in nonlinear optics is essential to increase the use of optics within computing and the manufacture of optical digital computers.

An optically bistable device used for photonic switching is the self electro-optic effect device (SEED). It is the analogous component for the transistor in optical processing. A SEED consists of multiple elements, biased by an external voltage, which creates an external field that shifts the wavelength of the onset of absorption, thus varying the intensity of the transmitted light. SEEDs rely on changes in the optical transmission of a semiconductor induced by an applied electric field. SEEDs can be configured in pairs (connected electrically) so that a beam of light switching one device can cause a complementary switch in the transmission of the other. Hence, a small change in the intensity of one beam can cause a large change in intensity of the other. This kind of configuration is known as symmetric SEEDs, and acts like an electronic flip-flop and permits logic operations NAND and NOR to be carried out on pairs of optical input signals. With symmetric SEEDs, the higher the input optical power, the faster the switching speed. Cellular logic image processor (CLIP) is implemented using the symmetric SEEDs. CLIP computer architecture is designed to permit parallel information processing, in which logic operations are performed on each element of the array simultaneously.

Holography

Holography is a photographic method of recording information (two mutually coherent beams) about an object, in which one beam is the object beam and the other a reference beam, which enables the construction of objects in three dimensions. Holography relies on the encoding of the object formation in a set of complex interference fringes formed by the interaction of a plane coherent reference wave with the wavelets diffusely scattered by the object.

The microscopic interference fringes recorded on a high resolution photographic plate after development and processing become a hologram. Hologram applications assist the study of small deformations of objects, to replace conventional optical elements such as lenses and beamsplitters, for data/information storage, and use of photo-refractive materials in analog computing, object recognition, and correlation, etc.

The purpose of recording a hologram is not optical data processing but rather a lensless method of forming an image. The uniqueness of a hologram as compared to an ordinary photographic image, for example, is that the latter contains amplitude information only. A hologram, because it involves interference of the complex field amplitude (CFA) of a given scene with a reference beam, records both amplitude and phase of the CFA of the given scene as seen within its field of view (FOV) by a square-low photosensitive device such as a film.

The film is exposed within its region of linearity whereby transmission is proportional to irradiance for a given time exposure. The transparency resulting from the film is the hologram. To view the holographic image, the hologram is illuminated with a reference wave, which causes the CFA of the given scene (both amplitude and phase) to be reconstructed. Since hologram reconstruction involves the CFA of the given scene in its entirety, depth information is included as well. Furthermore, it is even possible to view the holographic scene from different physical points of view and to see around objects in the scene. This ability, however, is limited by the solid angle subtended by the object when the hologram is recorded. Hologram reconstruction cannot supply more information than that recorded in the hologram itself.

Image Processing

Figure 1 demonstrates a general behavior of an imaging process. The ‘Processing System’ block in Figure 1 is expanded and presented in more detail by Figure 2. Figure 2 demonstrates in a flow chart, a general process occurring within the processing system. The input to the processing system is a degraded image (or set of images or video signal) such as blurred image, noisy image, image with clutters, multi-spectral image, etc. The input depends on exterior (environmental conditions) and interior (optical, electronic, and mechanical) effects of the imaging system. It may be generated as an analog signal or digital signal. As a result, the output of the imaging system is expected to be a processed (filtered, enhanced, estimated, restored, cleared, etc.) image.

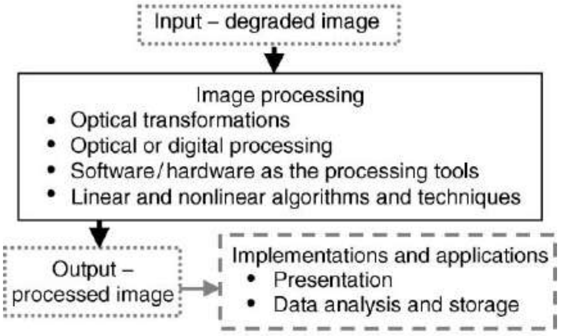
Input – degraded image
Image processing
• Optical transformations
• Optical or digital processing
• Software/hardware as the processing tools
• Linear and nonlinear algorithms and techniques
Output – processed image
Implementations and applications
• Presentation
• Data analysis and storage

Figure 2 Image processing.

Implementations and applications depend on the output obtained by the imaging system. The output can be presented by a display device and/or hardcopy. The output can be stored for future assignments and may also serve for data analysis such as detection, recognition, and identification of patterns and other details of interest (mapping, target acquisition, defects and imperfections in the human body, meteorology, security, intelligence, etc.).

Spatial units are reciprocal to time, and are cycles/seconds or Hz. The spatial units are also reciprocal to the domain of the information being sought. Spatial signals are 2D or even 3D on occasions. 2D spatial transforms may actually exist physically in the spatial plane. Novel transforms for information processing include Fourier and wavelets transforms. Such optical transformations serve as the basic platform for linear optical system processing. Image processing techniques are based on the image spatial frequencies. In most techniques the degraded image is transformed using an optical transformation, as a primary stage before any other process is being conducted. Optical (inverse) transformations also serve as the last processing tool before the output. Practically, optical transformations are involved and integrated in most processes related to image processing.

Image processing refers to both digital and optical image processing, and also considers algorithms and techniques, which are implemented by an integration of software and hardware. Image processing is an extremely broad field. It considers digital image processing, 2D and 3D image processing and pattern recognition, real-time processing, neural networks, image enhancement/improvement, image restoration, estimation, filtering for noise removal, linear and nonlinear filtering techniques for variety of implementations, optimum filtering, etc. Image processing also include varieties of algorithms such as linear and nonlinear algorithms, algorithms for pattern recognition, linear and nonlinear techniques for pattern recognition, neural networks based algorithms, etc. Both optical and electronic implementations (hardware) are still under investigation. Nevertheless, image processing based on software presents reliable results and hence, can be transformed into hardware components, which may become an integrated part of the processing hardware.

Optical Transformations

Mathematical transformations are applied to signals to obtain further information from the signal that is not readily available in the raw signal. Most of the signals, in practice, are time-domain signals in their raw format. That is, whatever that signal is measuring, is a function of time. In many cases, the most important information is hidden in the frequency content of the signal. The frequency spectrum of a signal is basically the frequency components (spectral components) of that signal. The frequency spectrum of a signal shows what frequencies exist in the signal.

There are many transforms that are widely used, often by engineers and mathematicians, in solving problems in science and engineering. Fourier, Laplace, Hilbert, Wigner distributions, Radon, and wavelet transforms, constitute only a small portion of a huge list of transforms that are available. Every transformation technique has its own area of application, with advantages and disadvantages.

Fourier Transform

The Fourier transform (FT) is probably the most popular linear transform being used. The FT, a pervasive and versatile tool, is used in many fields of science as a mathematical or physical tool to alter a problem into one that can be more easily solved. It is used in linear systems analysis, antenna studies, optics, random process modeling, probability theory, quantum physics, and boundary-value problems and has been successfully applied to restoration of astronomical data. Some scientists understand Fourier theory as a physical phenomenon, not simply as a mathematical tool. In some branches of science, the FT of one function may yield another physical function. The FT, in essence, decomposes or separates a waveform or function into sinusoids of different frequency which sum to the original waveform. It identifies or distinguishes the different frequency sinusoids and their respective amplitudes. There are functions for which the Fourier transform does not exist; however, most physical functions have a FT, especially if the transform represents a physical quantity. Other functions can be treated with Fourier theory as limiting cases. Many of the common theoretical functions are actually limiting cases in Fourier theory.

Wavelets

Wavelets were developed independently in the fields of mathematics, quantum physics, electrical engineering, and seismic geology. Interchanges between these fields have led to many new wavelet applications including image processing. Wavelets are mathematical functions that cut up data into different frequency components, and then study each component with a resolution matched to its scale. Wavelet transforms comprise an infinite set. The different wavelet families make different trade-offs between how compactly the basis functions are localized in space and how smooth they are. Within each family of wavelets are wavelet subclasses distinguished by the number of coefficients and by the level of iteration.

The wavelet transform (WT), which is also a linear operation, provides the time-frequency representation. Often a particular spectral component occurring at any instant can be of particular interest. In these cases it may be beneficial to know the time intervals at which these particular spectral components occur. WT is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the signal. The WT was developed as an alternative to the short-time FT (STFT). The WT was developed to overcome some resolution related problems of the STFT. The frequency and time information of a signal at some certain point in the time-frequency plane cannot be known. What spectral component exists at any given time instant is a mystery. Therefore, it is best to investigate what spectral components exist at any given interval of time. This is a problem of resolution, and it is the main reason why researchers have switched to WT from STFT. STFT gives a fixed resolution at all times, whereas WT gives a variable resolution as follows: higher frequencies are better resolved in time, and lower frequencies are better resolved in frequency. This means that a certain high frequency component can be located better in time (with less relative error) than a low frequency component. On the contrary, a low frequency component can be located better in frequency, compared to high frequency component.

Wavelet versus Fourier Transforms

Both transforms are linear and the mathematical properties of the matrices involved in the transforms are similar. The inverse transform matrix for both transforms is the transpose of the original. Wavelets have advantages over traditional Fourier methods in analyzing physical situations where the signal contains discontinuities and sharp spikes. The most interesting dissimilarity is that individual wavelet functions are localized in space, Fourier sine and cosine functions are not. This localization feature, along with wavelets localization of frequency, makes many functions and operators using wavelets ‘sparse’ when transformed into the wavelet domain. This sparseness, in turn, results in a number of useful applications such as data compression, detecting features in images, and removing noise from time series. One way to see the time-frequency resolution differences between the FT and the WT is to look at the basis function coverage of the time-frequency plane. In order to isolate signal discontinuities, some very short basis functions are needed. At the same time, in order to obtain detailed frequency analysis, some very long basis functions are needed. A way to achieve this is to have short high-frequency basis functions and long low-frequency ones. This happy medium is obtained with WT. FT utilizes just the sine and cosine functions (single set of basis functions). Therefore, its new domain contains basis functions that are sines and cosines. On the other hand, WT includes an infinite set of possible basis functions. Hence, its new domain contains more complicated basis functions called wavelets, mother wavelets, or analyzing wavelets. The basis functions are localized in frequency, making mathematical tools such as power spectra and scalegrams useful at picking out frequencies and calculating power distributions. Thus wavelet analysis provides immediate access to information that can be obscured by other time-frequency methods such as Fourier analysis.

Physical Properties and Analog/Digital Mitigation Processing in Nonideal Systems

The weakest link in most imaging systems is time varying distortions, such as the Earth’s atmosphere, motion, or vibrations. To obtain clear imaging through time varying distortions, it is useful to characterize the physical degradation phenomena using an OTF. It is often important to use the appropriate filter based on the OTF magnitude, modulation transfer function (MTF), to correct (restore) the degraded image (using software or hardware). On the other hand, it may be useful to incorporate real-time phase correction into the imaging system such as adaptive optics.

Optical Transfer Function for Incoherent Imaging

In electronics, the transfer function is redefined as the Fourier transform of the impulse response. The OTF is defined similarly, but is normalized to its own maximum value which normally occurs at zero spatial frequency, that is:

$$
\begin{aligned}
\mathrm{OTF} &= \tau(\omega) = \tau(\omega_x, \omega_y) \\
&= \frac{\iint_{-\infty}^{\infty} s(x', y') \exp[-j(\omega_x x' + \omega_y y') \mathrm{d}x' \, \mathrm{d}y']}{\iint_{-\infty}^{\infty} s(x', y') \mathrm{d}x' \, \mathrm{d}y'} \\
&= \frac{S(\omega_x, \omega_y)}{S(0,0)} = \mathrm{MTF} \exp(j\mathrm{PTF}) \qquad [1]
\end{aligned}
$$

where $S(\omega_x, \omega_y)$ is the FT of the spread function or impulse response $s(x', y')$. Ordinarily, for artificial imaging systems, the OTF is maximum at zero frequency. The magnitude of the OTF is called the MTF. The name refers to spatial modulation of light irradiance. If there is no such modulation, irradiance is uniform and there is no image. Thus, spatial modulation is related to image quality and MTF is a measure of the ability of an imaging component or system to transfer such spatial modulation from input (object plane) to output (image plane).

There is also a phase transfer function (PTF) whose importance in resolution is not nearly as widespread as that of MTF but which, nevertheless, cannot be neglected. Spatial phase determines image position and orientation, rather than amount of detail. If a target is displaced bodily in the image plane such that each part of the target image is displaced by the same amount, the target image is not distorted. However, if portions of the target image are displaced more or less than other portions, then the target image is distorted. This information is contained in the PTF.

The physical implications of the OTF are very analogous to those of electronics transfer functions. Both permit determination of output for any given input. In both cases, since the FT of the delta function is a constant (unity) amplitude, a delta function input permits indication of the frequency response of the system, that is, which frequencies pass through the system unattenuated and which are attenuated and by how much. Perfect fidelity or resemblance between input and output requires infinite bandwidth so that an impulse function is obtained at output for impulse function input. In imaging, an impulse formation at output is a point image. This means that, to obtain a point image for a point object, infinite spatial frequency bandwidth is required of the imaging system. Physically, this means that because of diffraction effects optical elements such as lenses, field stops, and mirrors must be of infinite diameter so as to eliminate diffraction at the edges. This, of course, is not attainable.

Pixel Modulation Transfer Function

To measure the point spread function (PSF), it is important to determine which size object can be considered a point object. This is affected primarily by instrumentation pixel size. The pixel is the smallest element recordable in image space. The brightness value represents average irradiance over that small portion of the image scene. Pixel size is often related to detector size. If only a portion of the detector is illuminated, the output current is equivalent to that obtained for the same total radiant power absorbed by the detector but averaged over the entire detector area. No detail smaller than a pixel can be resolved in the image. In practice, dead space between pixels violates isoplanaticism (spatial stationarity), which is a requirement for linear systems. Nevertheless, dead space is small and such effects are often neglected in image system characterization. However, it is possible to overcome distortions in MTF measurements arising from this lack of isoplanaticism by using a ‘white’ spatial noise random object. This gives rise to a white noise random image. The ratio of the image power spectral density to object power spectral density is equal to the square of the system MTF. In this way the lack of isoplanaticism is overcome by the spatial randomness of the object and image.

Pixel size strongly affects system MTF. If a pixel represents a point image, then pixel size and shape are minimum spread function. A best case OTF for such an imaging system is then a normalized FT of the pixel shape. For example, a square pixel yields a two-dimensional sinc MTF whose width increases as pixel size decreases.

Optical Transfer Functions for Image Motion and Vibration

Image blurring caused by vibrations is a factor whose influence on resolution is often significant in imaging systems that involve mechanical motion. A stabilized system, which obtains image motion corrections by moving optical elements so as to counteract sensor motions, has a residual error typically on the order of 2500 μrad/s involving pitch, roll, yaw, and forward motion compensation errors. The greatest errors usually result from rotational vibration. A stabilized mount-type system has a combined error typically on the order of 100 μrad/s. Even where no motion is involved, as in large fixed-position astronomical telescopes, vibrations deriving from thermal gradients in the walls and other parts of the telescope can be the limiting factor in image resolution.

As a consequence of such image motion, in many high-resolution systems, despite the use of high-quality sensors, resolution is limited by image motion and, as a result, the high-resolution capability of the sensor may be wasted. It can be a significant waste of means to use an expensive high-resolution sensor in such a situation unless the blur can be corrected with image restoration. In imaging system design, a convenient engineering tool is OTF. The overall system OTF is limited by the OTF of the weakest link. The formulation of such image motion blur into an OTF format is thus very convenient for system design and system analysis purposes, as well as for image restoration.

OTFs can describe image quality for many forms of image motion such as linear (uniform) motion, sinusoidal vibrations at high vibration frequencies, sinusoidal vibrations at low vibration frequencies, acceleration, etc. The OTF for any type of image motion can be calculated numerically in real time provided the motion function is known. This is usually determined easily via a gyroscope or microelectronic accelerometer.

In general, the motion of interest is not the actual relative motion between object and imager, but rather the relative motion between imager and image plane. The latter motion is that observed in the image plane. Both of these motions are related to each other by the image system magnification. Hence, image plane relative motion is usually much less than the actual relative motion. However, any motion during an exposure gives rise to blurring of image plane detail. One question to be considered is how much blur can take place as a result of motion and still not be noticeable because of inherent blur deriving from the PSF of the rest of the imaging system. In other words, assuming that image motion and vibration blur cannot be removed completely, how much is tolerable before resolution and target acquisition capability are affected? This depends on focal length and system zoom.

A general method of OTF calculation is obtained by the line spread function (LSF) derived from image motion transverse to the optical axis. The MTF is derived as the modulus of the MTF or FT of the LSF, and the PTF is derived as the phase of the OTF. The LSF is the probability density function (PDF) of the histogram of the relative transverse displacement between object and sensor beginning at time t_s and ending at time $t_s + t_e$ where t_s is measured from the instant the sensor is first exposed and t_e is exposure time.

Recently a method of calculating motion OTF analytically has been developed for any type of motion. Also, methods to calculate motion OTF without a motion sensor have been derived for a sequence of consecutive images and even from a single motion-blurred image. All three methods have been used successfully in image restoration.

Optical Transfer Functions for the Atmosphere

Many properties of the atmosphere affect the quality of images propagating through it. There are atmospheric phenomena that give rise to attenuation of the irradiance of the propagating image, thus reducing the contrast of the final image. There are also atmospheric phenomena that cause blurring of detail. Both types of phenomena prevent small detail from being resolved in the final image, thus degrading image quality.

Electromagnetic (EM) wave absorption and scattering by the constituent gases and particulates of the atmosphere and airborne particulates give rise to attenuation. Scattering of photons by airborne particulates is manifested as deflections of the photons to directions other than that of the original propagation. If such scattering causes the deflected photons to miss the imaging receiver, then the scattering is manifested as attenuation. The received irradiance of the image propagated through the atmosphere is correspondingly diminished from that in the object plane. However, if the light scattering is at very small angles with respect to the original direction of the propagation, and several such small-angle scattering events take place, then forward-scattered radiation can take round-about paths and still be received by the imaging system together with the unscattered radiation. The net effect is image blurring caused by a multitude of angles of arrival at the imaging receiver of radiation emanating from the same point in the object plane. Many multiple-scattering paths give rise to a relatively large blurred-point image rather than a fine sharp-point image. Adjacent image plane points can then appear for a single object plane point, thus degrading image resolution. Such aerosol scatter blur is known as the adjacency effect.

Another effect of light scattering, particularly at large angles, is increased path radiance. The atmospheric background irradiance at wavelengths less than 2–4 μm, consists through the day of scattered background light such as sunlight. This atmospheric background radiation is imaged over the same image space as the received irradiance from the target plane. The result is decreased contrast of the target plane scene. At wavelengths larger than 4 μm, most of the path radiance is not scattered sunlight, but rather thermal emission of the atmospheric constituents, gases in particular. This thermal path radiance decreases contrast in thermal imaging systems, particularly in the longwave infrared (8–13 μm).

Turbulence results from variations in the atmospheric refractive index, which are random in both time and space. These refractive index fluctuations are caused by local fluctuations in atmospheric temperatures, pressure, humidity, etc., all of which are affected by wind speed, which in itself is also random in time and space. Generally, the larger the temperature and humidity gradients, the more serious the image resolution degradation caused by turbulence. The lower the wind speed, the stronger the impact of turbulence. Scattering and turbulence are basically two very different physical mechanisms, each exhibiting quite different properties. Time dynamic properties such as scintillations and image dancing derive from turbulence, while aerosol scatter light changes only slowly with time.

A point that must be strongly emphasized is that there is no such thing as a purely turbulent, purely scattering, or purely absorption atmosphere. All three properties occur simultaneously in the real-world atmosphere. Modeling and interpretation of experimental results must relate to all three processes. Any attempt to pretend otherwise is not realistic. In general, blur is caused by turbulence and small-angle scatter of light by aerosols and particles. Attenuation is caused by large-angle scatter and by absorption.

Generally turbulence blur is dominant for images near the ground, where ground heating gives rise to the turbulence. Aerosol blur is usually dominant at higher elevations, even as little as several meters above ground, depending on weather conditions and wavelength.

The product of both aerosol and turbulence MTFs (in the spatial frequency domain, after FT) yields approximately atmospheric MTF. The atmospheric MTF can be predicted according to atmospheric (meteorological) parameters. Aerosol MTF calculation is based on optical depth, which includes parameters such as scattering and absorption coefficients through the atmosphere over the imaging path. Turbulence MTF calculation is based on the refractive index structure coefficient, which includes parameters such as temperature, humidity, wind speed, etc.

By referring to the edge response in a given scene the atmospheric MTF may be calculated. The edge separating the wide white and black bars is imaged from the scene. The image of the edge is not a step function because of atmospheric blur. The gradient of the edge image is the LSF, from which the overall atmospheric MTF can be obtained after the FT of the LSF is divided by the hardware MTF.

Spatial Filtering

Spatial filtering is a process by which we can alter properties of an optical image by selectively removing certain spatial frequencies that make up an object, for example, filtering video data received from satellite and space probes, or removal of raster from a television picture or scanned image. In digital image processing, the term image refers to a two-dimensional light-intensity function, where amplitude at spatial coordinates (x, y) gives the intensity (brightness) of the image at that point.

Digital image processing techniques are oriented to transform, enhance, restore, or encode an image. Digital image restoration is on the basis of the OTF or even MTF alone describing the image degradation. Resolution is limited usually not by electronics or optics, but by image motion and vibration. For long-range imaging, the image quality may be limited by atmospheric effects.

By image restoration filters we refer primarily to spatial frequency filters placed in the spatial frequency plane in the optical data processor. Although data processing is done optically, the same mathematical concept is usually carried out with a digital computer. It is also possible to do restoration in the spatial domain. The simplest image restoration filter is the inverse filter. When noise becomes a problem, there are often used three classes of minimum mean square error estimations (MMSEE), which are the noncausal Wiener filtering, causal Wiener filtering, and Kalman filtering. Kalman and Wiener filters are proposed to obtain the optimum solution for the MMSEE problem in linear systems. Wiener filters are applied in linear time invariant systems, and Kalman filters in linear systems which may be either time invariant or time variant. A novel area of application for Wiener filtering is image restoration. On the other hand, if the estimation processes are partial and an ideal image is not obtainable, a Kalman filter may present better restoration results than a Wiener filter.

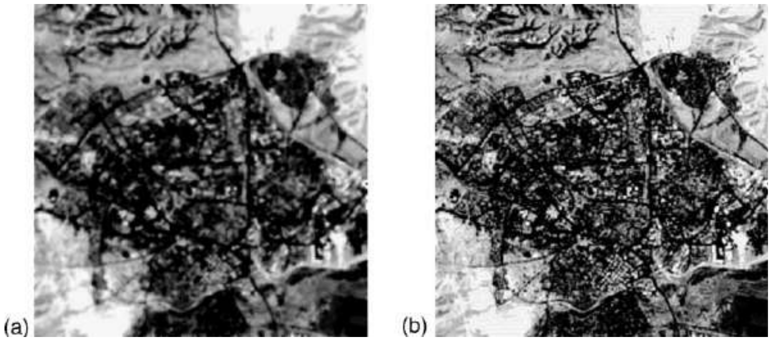

Figure 3 An example for satellite image restoration. (a) Original image of Beer-Sheva, Israel, from Landsat 7 satellite; (b) restored image after Kalman filter using aerosol and turbulence MTFs. Reproduced with permission of the American Society for Photogrammetry and Remote Sensing from Arbel D, Cohen E, Citroen M, Blumberg DG and Kopeika NS (2004) Landsat TM satellite image restoration using Kalman filters. *Photogrammetric Engineering and Remote Sensing* (PE&RS) 70(1): 91–100.

Figure 3 demonstrates the restoration result for a satellite image. Figure 3a presents an original Landsat 7 enhanced thematic mapper plus (ETM+) image of the city of Beer-Sheva, Israel. The image was recorded from 750 km satellite elevation. The image was recorded using an ETM+ sensor with 15 m ground resolution, and its wavelength range is 520–900 nm (Panchromatic band). Figure 3a is of 256 × 256 pixel size, which is a segment defined as a region of interest (ROI) in the original image. The scene was recorded on August 7, 1999 and was degraded (blurred) as a result of the meteorological conditions, which were measured and recorded (as data) at the same time. Figure 3b shows the restored image using the Kalman filter, which was based on the atmospheric MTF, which is obtained by using aerosol MTF with turbulence MTF. The restored image (Figure 3b) is more detailed. For example, roads and buildings were deblurred and became more accurate than in the original image (Figure 3a).

See also

Environmental Measurements: Optical Transmission and Scatter of the Atmosphere. **Fourier Optics**. **Holography, Techniques:** Computer-Generated Holograms. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Arbel D, Greenberg S, Hadar O and Kopeika NS (2003) Criteria for satellite image restoration success. *Optical Engineering* 42(9): 2607–2613.

Arbel D, Cohen E, Citroen M, Blumberg DG and Kopeika NS (2004) Landsat TM satellite image restoration using Kalman filters. *Photogrammetric Engineering and Remote Sensing (PE&RS)* 70(1): 91–100.

Daniels A, Boreman GD and Ducharme AD (1995) Random transparency targets for modulation transfer function measurement in the visible and infrared. *Optical Engineering* 34(3): 860–868.

Gabor D (1948) A new microscope principle. *Nature* 161: 777.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceeding of the Royal Society A* 197: 454.

Gabor D (1951) Microscopy by reconstructed wavefronts: II. *Proceeding of the Royal Society B* 64: 449.

Hadar O, Dror I and Kopeika NS (1994) Image restoration limits resulting from mechanical vibrations, part IV: real-time numerical calculation of optical transfer functions and experimental verification. *Optical Engineering* 33: 566–578.

Hadar O, Robbins M, Novogrozky Y and Kaplan D (1996) Image motion restoration from a sequence of images. *Optical Engineering* 35: 2898–2904.

Kopeika NS (1998) *A System Engineering Approach to Imaging*. Bellingham, WA: SPIE Press.

O'Donnell E (1987) Image steadiness: a critical factor in system performance. In: Wright R (ed.) *Electron Optical Imaging System Integration. Proceedings SPIE* 762: 51–56.

Stern A and Kopeika NS (1995) Analytical method to calculate optical transfer functions for image motion and vibrations using moments. *Journal of the Optical Society of America A* 14: 388–396.

Yitzhaky Y and Kopeika NS (1997) Identification of blur parameters from motion blurred images. *CVGIP: Graphical Models and Image Processing* 59(5): 321–332.

OPTICAL TWEEZERS

A Gajraj and J C Meiners, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The concept of radiation pressure or pressure from the propagation of light is not new. In the early seventeenth century, Johannes Kepler proposed that radiation pressure from the sun accounted for the fact that tails of comets always point away from the sun. In 1873, James Maxwell, on the basis of electromagnetic theory, showed that pressure should result from the absorption or reflection of a beam of light. However, early studies in the beginning of the twentieth century that attempted to detect and quantify the effects of radiation pressure were unsuccessful, mainly frustrated by thermal forces from temperature gradients that tended to obscure the weak effects of radiation pressure. It was only in the 1960s, that the advent of lasers provided a light source with the power to generate sufficient radiation pressure to overcome forces due to thermal effects.

As lasers became generally available in the early 1970s, Arthur Ashkin from the Bell Labs pioneered the development of optical tweezers. He showed that the forces of radiation pressure from focused laser beams could be used to affect the dynamics of small transparent micrometer-sized particles. According to Ashkin, his early experiments were motivated by simple back of the envelope calculations of the magnitude of the radiation pressure force of light on a totally reflecting mirror. Let us consider such a calculation: a photon of frequency ω has a momentum of $\hbar\omega/c$. For a total power of I, we have a flux of $I/\hbar\omega$ photons per second incident on the mirror. When the photons are reflected, they transfer a momentum of $2I/c$ to the mirror. The force exerted on the mirror for $I = 1$ W, is approximately 10^{-9} N. If the mirror is a 1 μm size particle weighing 10^{-15} kg, this force will induce an acceleration of $10^6\, g$!

In the following decade, Ashkin demonstrated that by focusing laser light using a high numerical aperture microscope objective, strong intensity gradients could be created to optically trap small dielectric particles such as microspheres. Ashkin named this trapping scheme a ‘single-beam gradient force trap’, now more commonly called optical tweezers.

Today optical tweezers are well-established and extremely versatile tools for the physical and life sciences. Using optical tweezers, forces up to 200 pN can be applied with a resolution of 100 aN on particles with dimensions on the order of the wavelength of light. Even particles as small as 5 nm can be trapped, rotated, positioned, and arrayed. Optical tweezers have been used in single molecule studies of biopolymers (DNA, RNA), in the manipulation of various types of cells and in cell sorting, in microsurgical procedures, as well as in viscoelastic studies of membranes, and protein. Optical tweezing has also been the key tool in the understanding of the workings of molecular motors such as myosin, kinesin, RNA polymerases, etc.

Theory

Optical tweezers trap dielectric particles using radiation pressure generated by a tightly focused beam of light from a laser. Depending on the size of the particle in relation to the wavelength of the trapping beam, there are two different mechanisms responsible for trapping small particles in a single beam gradient trap. For particles in the Mie regime, where the particle size is large compared to the laser wavelength, a simple geometric-optics model serves best. Figures 1a–c, illustrate possible cases of a Mie particle in an optical trap. Here, the particle is assumed to be a sphere that is transparent with a refractive index higher than that of the surrounding medium. The particle behaves as a convergent lens refracting and deflecting the beam when its center is displaced from the laser focus. In Figure 1a, the sphere is axially displaced from the focus, away from the microscope objective. Ray A is then refracted toward the surface normal as it enters and exits the sphere, causing a change in the momentum of the ray. Due to conservation of momentum, the sphere reacts with a force F_a in the opposite direction. The situation is similar for ray B. The gradient force, F, obtained from the vectorial addition of F_a and F_b, is directed toward the light source. At the same time, the scattering force, arising from back and scattered reflections, is directed away from the light source. It can be seen that above the focal point, the scattering and gradient forces act in the same direction, pushing the particle away from the light source (Figure 1b). For a particle in the focal plane, but transversely displaced from the trap center, the gradient force acts in a restoring direction pushing the bead into the trap focus (Figure 1c). At a location just below the focus, the scattering force balances the gradient force and stable optical trapping is achieved.

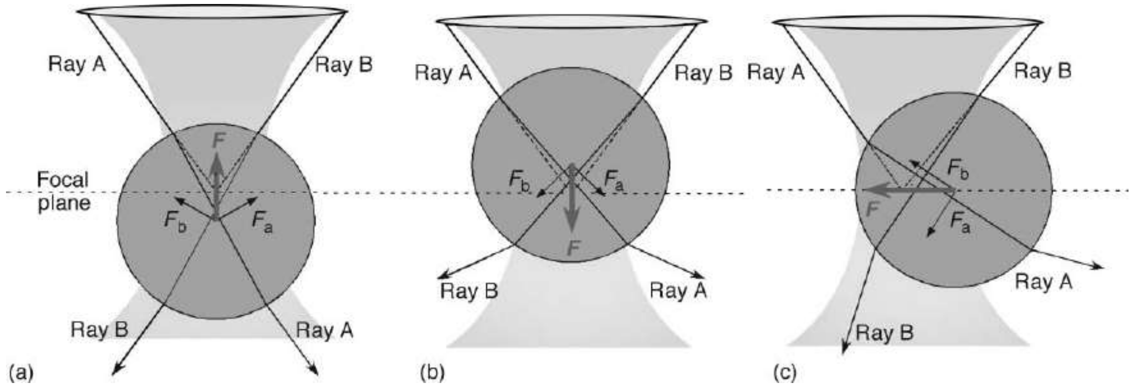

Figure 1 Ray diagram sketches showing three cases of particle trapping in the Mie regime. Rays A and B border the cone of light focused by a high numerical aperture microscope objective. The refraction of these rays gives rise to forces F_a and F_b, the resultant force F tends to restore axial and transverse displacements of the particle from the trap focus (at the intersection of dotted rays A and B).

This simple ray optics model can also be used to estimate the trapping force, an example of such a calculation is shown in Box 1.

For dielectric particles in the Rayleigh regime, where the size of the particle is small compared to the wavelength of the incident light, the particle acts as a simple dipole. The electromagnetic light wave polarizes the particle, inducing a set of oscillating dipoles. The dipoles respond to the electric field of the incident light in the direction of the gradient, resulting in the movement of the particle towards the focal point. Additionally, there is also a scattering force, acting in the direction of beam propagation, from light that is scattered or absorbed. While the scattering force is proportional to the electromagnetic field, the gradient force is proportional to the gradient of the electromagnetic field. Therefore, stable optical trapping can be achieved for a field with a gradient large enough to overcome the scattering force. This can be obtained by using a microscope objective with a high numerical aperture.

Stable trapping is achieved when the energy associated with the trapping significantly exceeds the thermal energy of the trapped particle, $\frac{1}{2}k_B T$ per degree of freedom. Since the optical gradient is weakest in the axial direction of the trap, the trap stiffness in this direction limits its stability. Furthermore, the trap is weakened in the axial direction by the scattering force, which pushes the microsphere away from the objective. To overcome this, two counterpropagating laser beams can be used, not only to increase the axial gradient of the trap, but also to provide a compensating scattering force in the opposite direction.

Force Measurement

Optical tweezers can be used as picotensiometers or highly sensitive force sensors. Measuring the displacement of a particle in a trap with known stiffness k, yields the force F given by

$$\mathbf{F} = -k \cdot x \qquad [1]$$

The displacement can be measured in different ways, ranging from video microscopy to light-scattering techniques. Some of these methods are discussed in the instrument design section below.

The trap stiffness is generally not known *a priori* as it depends on the shape and intensity of the trap, as well as the relative refractive index, shape, and position of the object in the trap. Since theoretical approximations are usually not accurate enough, the trap stiffness must be empirically determined and calibrated.

Various force calibration strategies have evolved over the years, however, they can be broadly divided into two types: (i) schemes that rely on the interaction of the optical trap with viscous forces; and (ii) those that exploit the thermal fluctuations of the trapped object.

Force Calibration by Viscous Drag

The conceptual basis for calibration by viscous drag is simple. A known force is exerted on a trapped object and the deviation in its position is measured as a function of the applied force. The trap stiffness can then be calculated from eqn [1]. Force can be exerted by either flowing liquid past the trapped object, translating the microscope stage while keeping the trap fixed, or translating the optical trap while keeping the sample fixed.

In a situation where the Reynolds number is small, inertial forces can be neglected and Stokes' Law gives the viscous drag force on a sphere of known radius r:

$$F_d = (6\pi\eta r)v = \beta v \qquad [2]$$

Box 1

Optical Tweezers in the Ray Optics Model

In a simple ray optics model, the trapped microsphere acts like a small lens on the incident laser beam. If the microsphere is not centered on the beam axis, the beam will be refracted. The corresponding change in the momentum of the laser beam gives rise to a restoring force on the microsphere.

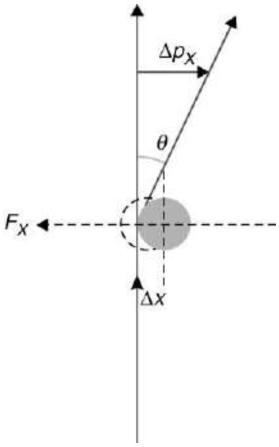

Figure B1

A photon of frequency ω in a medium of refractive index n_1 that is deflected by an angle θ, as shown in Figure B1, acquires a lateral momentum of

$$\Delta p_x = \frac{\hbar \omega n_1}{c} \sin \theta \qquad [B1]$$

Momentum conservation requires a counteracting force on the microsphere, which is for a laser beam of intensity I

$$F_x = \frac{I n_1}{c} \sin \theta \qquad [B2]$$

Assuming that the deflection angle θ is given by the deflection of a beam that is incident on biconvex lens with curvature R and refractive index n_2 at an off-axis displacement Δx, we find for the restoring force on the microsphere:

$$F_x = \frac{2I}{Rc}(n_2 - n_1)\Delta x \qquad [B3]$$

which describes a linear spring with a stiffness of the optical trap of

$$k = \frac{2I}{Rc}(n_2 - n_1) \qquad [B4]$$

Here, η, the viscosity of the sample fluid, is known and immediately allows the determination of the friction coefficient β of the sphere. The fluid velocity, v, can be measured in the case of external fluid flow by either tracking the trapped object when the trap is switched off or measuring the speed of some other marker in the sample. If the stage is moved to generate the fluid flow, the velocity of the fluid surrounding the trapped object can be estimated from the motion of the stage. In either case, the drag coefficient, from eqn [2] may need to be corrected if the trapped object is close to the walls of the sample cell. This correction is well known as Faxen's Law, which accounts for hydrodynamic interactions between the trapped particle and the wall. These effects are particularly significant when the center of the trapped particle is less than a diameter away from the wall. The advantage of these fluid-flow calibration methods is that the trap stiffness at regions far from the trap center can be measured. The drawback is that, in general, only transverse trapping forces can be measured and the trapped objects have to be approximated as spherical or ellipsoidal.

Thermal Motion Methods

These methods measure the Brownian motion of a trapped particle to deduce the trap stiffness. A trapped particle experiences three forces: (i) the thermal forces due to collisions with surrounding solvent molecules; (ii) forces from viscous drag; and (iii) the restoring force of the trap. For low Reynolds numbers, where the inertial effects can be neglected for the microsphere, the equation of motion is

$$\beta\dot{x} + kx = F(t) \qquad [3]$$

Where β is the viscous drag, k is the trap stiffness, and $F(t)$ is a randomly fluctuating Langevin force due to thermal fluctuations. This motion of the microsphere in the trap at a temperature T is characterized by a Lorentzian power spectrum of the form:

$$x_f^2 = \frac{k_\mathrm{b}T}{2\pi^3\beta(f_0^2 + f^2)} \qquad [4]$$

where $f_0 = k/(2\pi\beta)$ is the corner frequency, above which the motion of the particle is Brownian.

To calibrate the optical trap by the corner frequency method, the motion of a trapped particle is observed and its power spectrum, which is the Fourier transform of the autocorrelation function, is calculated. Fitting the power spectrum to a Lorentzian function will yield the corner frequency f_0, which can be directly related to the trap stiffness k, provided that the drag coefficient β can be estimated.

Alternatively, the trap stiffness can be obtained directly from the time constant of the exponential decay of the autocorrelation function of the particle motion. The spring constant is the ratio of the drag coefficient to the time constant of the trap.

Another approach known as the mean square displacement method does not require the estimation of the drag coefficient but does require a well-calibrated position detector. In this method, the mean square displacement $\langle x^2 \rangle$ of the fluctuations of a trapped particle is computed from position data and then by the Equipartition theorem:

$$\langle x^2 \rangle = k_\mathrm{b}T/k \qquad [5]$$

This enables a simple, straightforward calculation of k, the trap stiffness.

Design

Optical tweezing stations broadly fall into three categories: (i) stand-alone systems; (ii) systems incorporating an optical microscope; and (iii) completely configured commercially available systems. Stand-alone systems feature optical tweezers built around a microscope objective and suitable optical components. These systems provide the maximum flexibility and can be easily reconfigured to adapt to different experimental needs. Furthermore, they also tend to provide the best mechanical stability, if configured properly. Figure 2a shows a picture of a stand-alone optical tweezing setup from our laboratory. This system is capable of the basic tasks of trapping, holding, and positioning microscopic objects. The most common optical tweezing stations are integrated into commercial light microscopes as sketched in Figure 3. The advantage of these hybrid systems is that the commercial microscope provides a flexible, optimized optical imaging platform, in which the optical trapping system can be integrated with only moderate difficulty. Figure 2b shows a picture of a typical optical tweezing station built around a commercial inverted microscope and mounted on a vibration–isolation table. This system includes an acousto-optic modulator that enables beam scanning for line-tweezing experiments. At the time of this writing, optical tweezing systems range from less than $10 000 for basic demonstration stations to between $50 000 to $100 000 for more sophisticated hybrid systems.

Commercial fully integrated systems are available from P.A.L.M. Microlaser Technologies AG (Bernried, Germany), Cell Robotics, Inc (Albuquerque, NM, USA), and Arryx (Chicago, IL, USA). Figure 2c shows a commercial optical tweezers set-up available from P.A.L.M. Microlaser Technologies.

When designing an optical tweezing station, the first item to consider is the choice of laser system. A simple single beam tweezing system can be built around a laser with less than 100 mW of power. Even though the minimum power required for trapping is a few milliwatts in the laser focus, there is considerable power loss in the tweezing and microscope optics. Also, the trap stiffness and thus the maximum trapping force decreases linearly with intensity, therefore most optical tweezers typically use lasers with an output power of 1 W. In some special applications, when beam scanning or multiple traps are required, the power requirement can be significantly larger. The wavelength of the laser also affects the capability of the trap in that the optimum wavelength depends on the size of the objects that are to be trapped. In general, the strongest traps occur

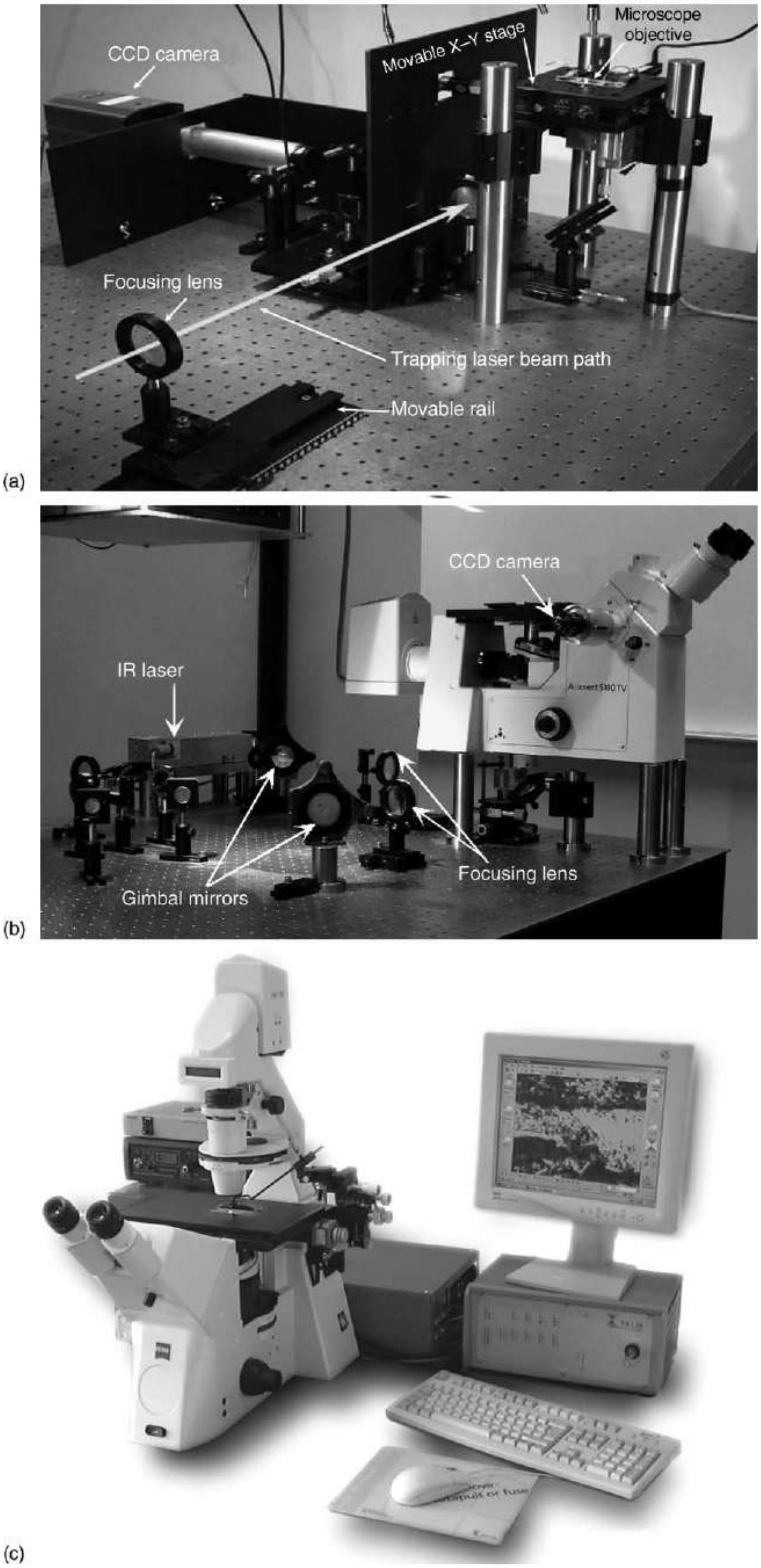
Microscope objective
Movable X–Y stage
CCD camera
Focusing lens
Trapping laser beam path
Movable rail
(a)
CCD camera
IR laser
Axiovert S100TV
Gimbal mirrors
Focusing lens
(b)
(c)

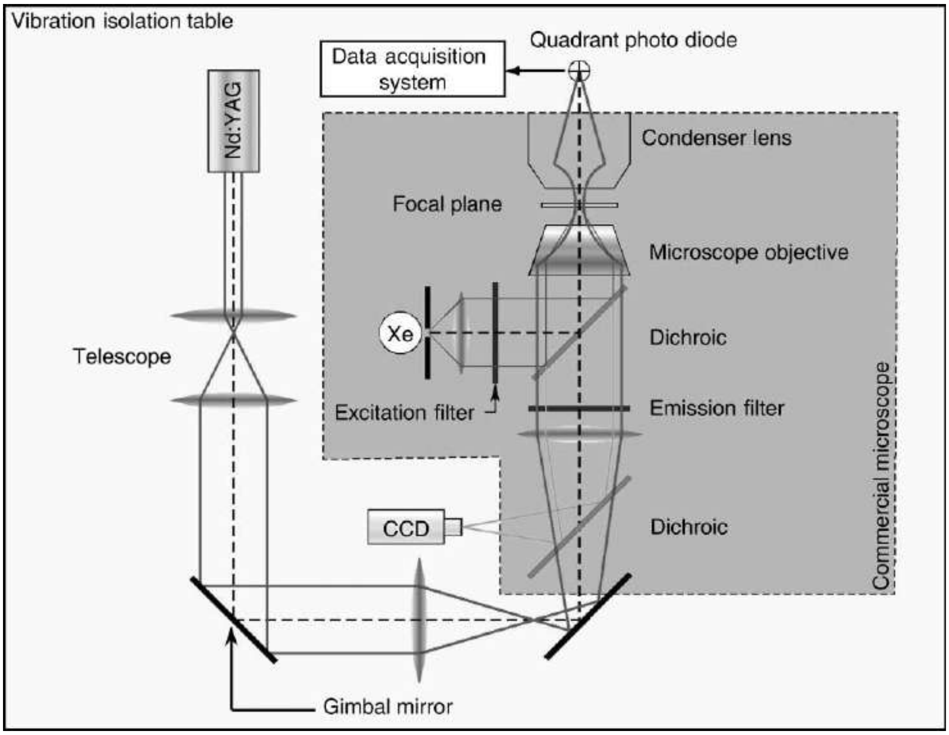

Figure 3 A sketch showing the instrumentation of a typical optical tweezing system incorporated into a commercial microscope.

for particles of size comparable to the wavelength of the laser. Another important concern in the choice of wavelength is possible damage to the sample from absorption and heating. In general, biological material is more transparent to the longer wavelengths of the IR spectrum. Studies on living cells have shown that generally, a range of wavelengths of 700 nm to 1100 nm, termed 'the optical window' of cells and tissue, is most suitable. Importantly, at higher wavelengths there is a reduced possibility of damage from multiphoton effects. DNA (260 nm) and proteins (280 nm) are strong absorbers in the ultraviolet range. Due to multiphoton effects, wavelengths between 700–800 nm have been seen to affect cell metabolism and lead to cell death within seconds. However, at wavelengths greater than 1000 nm, the simultaneous absorption of four photons required is unlikely at typical laser intensities used in trapping experiments.

The beam quality of the laser also restricts the choice of laser system. Diode lasers, in particular, have poor beam quality since the beam is often elliptical in shape and astigmatic, leading to a larger inefficient trap. Over the years, Nd-YAG lasers, at 1064 nm, have emerged as the lasers of choice, providing the best balance between price, power, and quality.

To achieve stable trapping it is essential to create an optical potential in which the gradient forces dominate the scattering force as discussed in the Theory section above. This suggests that the higher the numerical aperture (NA) of the microscope objective the stronger and more stable the trap will be. Oil or water-immersion microscope objectives with NA > 1.2 are commonly used. Further, this requirement also implies that the back aperture (rear entrance) of the microscope objective needs to be completely filled or, in the case of a Gaussian beam, overfilled to a certain degree to achieve optimal trapping efficiency. Since most commercial objectives are corrected for achromatic and spherical aberrations in the visual region of the spectrum, only

Figure 2 (a) A stand-alone tweezing station built around a microscope objective mounted on a vibration isolation table. (b) A typical tweezing station incorporated into a standard commercial microscope. This setup features dual-steerable optical traps and includes a CCD camera. Position detecting instrumentation can be mounted on the existing microscope stage. (c) The commercially available PALM® Combisystem utilizes an IR laser (wavelength 1064 nm) and a UV laser (wavelength 377 nm) coupled via the epi-fluorescence path to a research microscope. The system advertises automated and noncontact laser microdissection as well as trapping, sorting, and positioning. (**Figure 2:** (c), courtesy of P.A.L.M. Microlaser Technologies, Bernried, Germany.)

a small fraction of the working distance is usable for viable trapping. Normally this is limited to within 40 μm of the cover slip.

In addition to providing a beam large enough to overfill the back aperture, there are two other criteria that the external optics have to satisfy: first, the optical trap should be parfocal with the specimen plane. This enables the visualization of trapped objects. Second, the laser beam should be able to pivot around the back aperture without moving laterally to maintain the same degree of overfilling. This allows for a stable two-dimensionally movable trap. In practice, these requirements are fulfilled by using a telescopic system of lenses and Gimbal mounted mirrors. The latter are special mounts that enable a mirror to be tilted around its center position without lateral displacement. The telescope expands the incoming beam to fill the back aperture of the microscope objective and also images the back aperture onto the center of the Gimbal mounted mirror. As seen in Figure 4, since the back aperture of the microscope objective and the Gimbal mounted mirror are in conjugate planes, a small tilting of the mirror will only result in a change in the angle of the laser beam entering the microscope objective. The position of the beam will be unchanged, as will the degree of overfilling.

More sophisticated optical tweezer designs use acousto-optic modulators and galvanometer scanning mirrors to create scanning line tweezers and/or multiple traps. The basic approach is to deflect the laser beam to create multiple traps or to scan a single trap to create a one-dimensional potential well. Additionally, a new generation of holographic optical tweezers, capable of creating three-dimensional arrays of trapped particles, has emerged. Computer generated holograms that modify the phase and amplitude of the trapping laser are reverse engineered from the desired two- or three-dimensional optical trapping patterns. Furthermore, constructing such holograms from liquid crystal light modulators allows for dynamic control of the trapping pattern, yielding a fully computerized reconfigurable trapping system. An interesting example of the application of holographic optical tweezers to colloidal transport is shown in Figure 5, the first demonstration of 'kinetically locked-in states' where particles traveling across a potential mesh display a preference for certain paths.

It is to be noted that the optical tweezer is fundamentally a constant-extension technique, that is, it attempts to maintain the extension of a stretched molecule whereas the force is varied. However, it is of interest to study the action of molecular motors under constant force conditions. A basic optical tweezers setup can be modified to include an electronic feedback loop to provide constant force conditions. More recently, scanning-line optical tweezers have been adapted to implement all-optical constant force schemes that dispense with the need for electronic feedback.

An important aspect of any force measurement is the ability to accurately image and measure the position of a trapped particle. Standard digital video microscopy techniques yield a resolution of ~5 nm but have limited bandwidth and thus cannot be used for systems that employ feedback control of laser intensity or trap position. In these systems, the forward- or back-scattered trapping laser light from the trapped particle can be imaged onto a calibrated quadrant photodiode to provide sensitive position detection at high bandwidths (~1 kHz). If the photodiode is fixed in a plane conjugate to the back focal plane of the microscope objective, the trap can

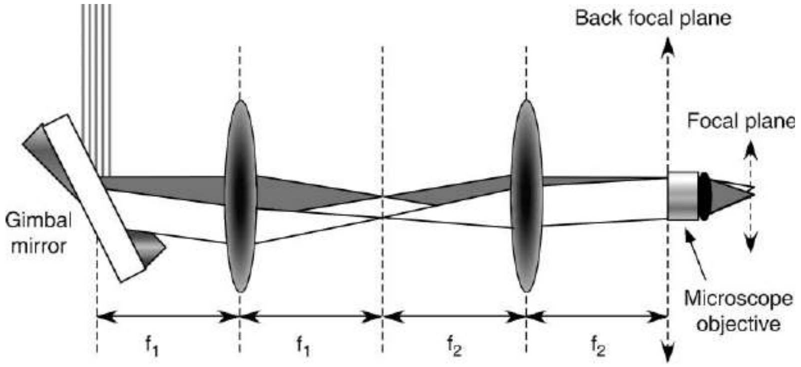

Figure 4 A sketch showing the working of a gimbal mounted mirror in two positions (shown as shaded and as unshaded). For small tilts of the mirror, the beam is seen to 'pivot' about the back aperture of the microscope objective resulting in the displacement of the optical trap in the focal plane. Note that for both beams the degree of overfilling of the back aperture of the microscope objective is unchanged, a crucial requirement for movable traps.

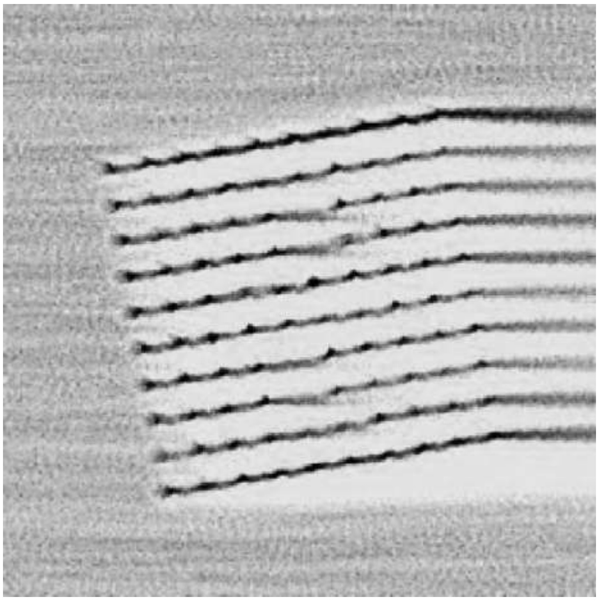

Figure 5 Data showing the superimposed paths of approximately 18 000 micron-sized particles flowing across a 10 × 10 grid of optical traps created by holographic optical tweezers. The particles are deflected laterally assuming a path dictated by the potential energy landscape and display kinetically locked-in states. This technique can be used to sort particles based on size, refractive index or charge. (Picture courtesy of D. Grier.)

be moved without having to move or recalibrate the photodiode detector. Using the quadrant photodiode method 2- and 3-dimensional position sensing is possible. A further improvement is found in schemes that use an additional laser dedicated to position sensing. The most sensitive method, however, for position sensing is interferometry. In this scheme, the trapping laser beam is split into two orthogonally polarized beams that are recombined after passing through the microscope objective and the sample. For a particle perfectly centered in the trap, the resulting combined beam has a circular polarization. In other cases, the bead displacement is sensitively reflected in the ellipticity of the polarization.

Applications

Colloids

Optical tweezers have been widely used to probe the interaction in colloids and other complex fluids. In fact, an optical tweezer apparatus was included in the instrumentation on the International Space Station for use with colloidal crystallization experiments in space. Various aspects of colloidal interactions have been studied using optical tweezers including mutual- and self-diffusion coefficients, the effects of hydrodynamic and electrostatic interactions, and depletion layer effects. More sophisticated scanning-line tweezing setups have been used to probe entropically driven self-assembly and interactions of colloids in polymer suspensions.

Blinking optical tweezers provide an elegant method for positioning colloidal particles in configurations of interest in a solution. When the tweezers are switched off the colloids respond to their environment, reacting to thermal forces and interacting either with other colloids or interfaces. The trajectories of the released colloids can be tracked and analyzed in 2D using digital video microscopy or 3D if combined with an evanescent illumination scheme. Switching on the tweezers again results in a return to the initial configuration, yielding a reproducible and tightly controlled experiment to examine colloidal behavior.

Biopolymers

The study of biopolymers has benefited from the development of optical tweezers. By chemically attaching a trappable latex microsphere to a biopolymer of interest, the molecule can be stretched, positioned, or moved. Measuring the response of the biopolymer to these external forces yields useful information such as elasticity, rigidity, and domain unfolding characteristics. This information obtained at a single molecule level has provided invaluable insight into the relationships between the structure and function of biopolymers. For example, in an early tweezing experiment, characterizing the intrinsic elasticity of λ-phage DNA, it was discovered that the double-stranded DNA undergoes a reversible structural transition when stretched at forces greater than 65 pN, extending to about 1.7 times its crystallographic contour length. These studies complement well-established structural techniques, such as X-ray crystallography and NMR, which provide information in a force-free environment. Furthermore, by anchoring one strand of DNA to a surface and pulling on the other strand, unzipping forces for DNA have been measured. Optical tweezers have also been combined with other mechanical tools, such as a micropipette, and used as sensitive tensiometers. Protein folding and unfolding experiments are generally harder, as their contour length is shorter than that of typical DNA fragments. Nevertheless, the folding and unfolding behavior of single giant muscle protein, titin has been probed by attaching a different latex bead to either end of the molecule. The first bead is held by a movable micropipette and the second bead is optically trapped. As the micropipette is moved, the position of the optically trapped bead provides information about the forces required to unfold the membrane domains of the protein.

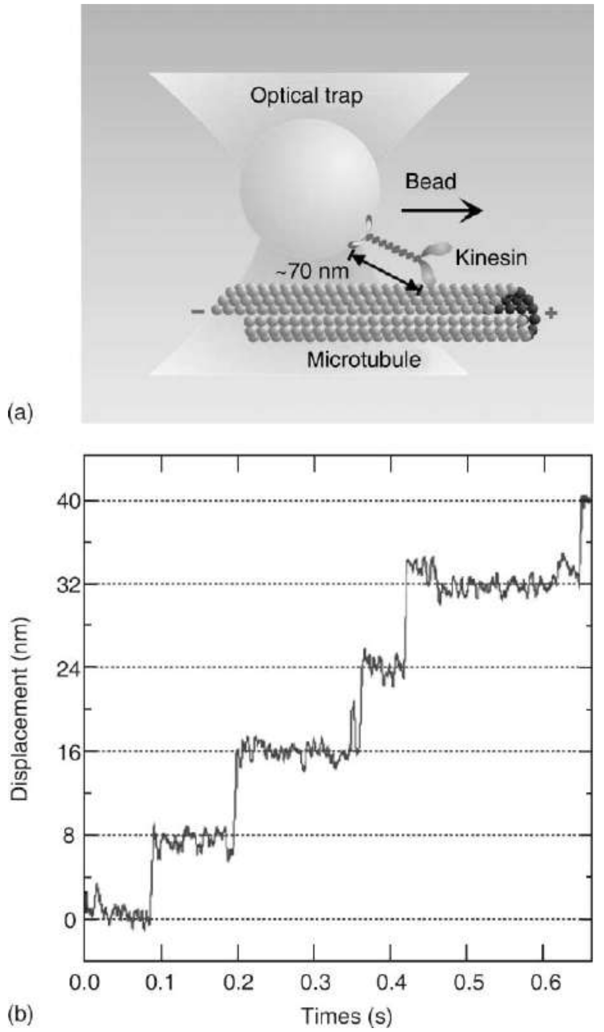

Figure 6 Observation of the stepping motion of the molecular motor kinesin. (a) A microsphere attached to a kinesin molecule is optically trapped; as the kinesin steps along a microtubule track which is immobilized onto a glass substrate, it pulls on the microsphere, whose motion is followed by an electronic feedback mechanism. (b) Motion of a single kinesin motor stepping against a 5 pN force. The data shows the discrete 8 nm steps of kinesin. (Figures courtesy of S. Block.)

Molecular Motors

Molecular motors (proteins that transduce chemical potential energy in to mechanical motion) have been the subject of intense attention since the early 1990s. Optical tweezers have played a crucial role in furthering the understanding of the mechanisms of molecular motors. The characteristics of molecular motor systems that have been studied with optical tweezers include speed, step-size, processivity, ATPs consumed per stroke, stall force, and static force. In fact, many of the advances in optical tweezers technology have been motivated by the study of molecular motors and have lead to spectacular results. An impressive example is the observance of the stepping motion of kinesin, as shown in Figure 6. Other experiments on molecular motors with optical tweezers have measured the stall force of motors as they move along their polymer track, or observed the transcription of DNA and RNA polymers, yielding interesting observations of the pauses during the transcription process.

Cells

The application of optical tweezers to cell biology was first demonstrated by Ashkin and co-workers in the late 1980s, by successfully trapping living objects: large viruses and bacteria. Since then, optical tweezers have been widely used to study various aspects of cellular biology. These include cell reproduction, growth, adhesion, mobility, and membrane elasticity. Additionally, the noninvasive nature of optical tweezers enables manipulations in the interior of an unopened object, allowing for the study of intracellular mechanisms by subcellular manipulations.

Optical tweezers have also been combined with various techniques to yield application-oriented tools for cell biologists. For instance, the combination of optical tweezers and state-of-the-art image recognition algorithms has been used to develop microfluidic cell sorters. In reproductive medicine, two of the major problems are penetration of the egg and sperm motility. A promising new solution combines laser zona drilling to pierce the egg envelope and optical tweezing to transport the sperm cells to the egg.

See also

Modulators: Acousto-Optics.

Further Reading

Ashkin A (2000) History of optical trapping and manipulation of small-neutral particle, atoms and molecules. *IEEE Journal of Selected Topics in Quantum Electronics* 6(6): 841–856.

Greulich KO (1999) *Micromanipulation by Light in Biology and Medicine: The Laser Microbeam and Optical Tweezers*. Basel; Boston; Berlin: Birkhauser.

Grier DG (2003) A revolution in optical manipulation. *Nature* 424: 810–816.

Lang MJ and Block SM (2003) Resource letter: LBOT-1: Laser-based optical tweezers. *American Journal of Physics* 71(3).

Sheetz MP (1998) *Methods in Cell Biology. Volume 55: Laser Tweezers in Cell Biology*. San Diego, CA: Academic Press.

Svoboda K and Block SM (1994) Biological applications of optical forces. *Annual Review of Biophysics Biomolecular Structure* 23: 247–285.

P

PHASE CONTROL

Contents

Phase Conjugation and Image Correction
Wavefront Coding

Phase Conjugation and Image Correction

E N Leith, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase conjugation of a wave is a process that has long been recognized. However, the technique for producing high-quality phase conjugation is relatively new. Also, the principal use of phase conjugation, the correction of phase errors in waves that have propagated through phase-distorting media, is relatively new, having originated in holography in the 1960s. Today, phase conjugation is an important area of modern optics. This article describes the conjugation of a wavefield by the process of holography and the holographic phase conjugation method of removing phase distortions from a wavefield, thereby permitting high-quality images to be produced from phase-distorted waves.

The basic meaning of a conjugated wave is explained with the aid of Figure 1. Figure 1a shows a wavefield propagating to the right. For simplicity, we show only one wavefront, i.e., one contour of constant phase. It is understood that a wavefield consists of a sequence of such wavefronts, with each wavefront differing in phase by 2π from the adjacent one, and separated by the wavelength λ. Of the two points, A and B, on the wavefront, we say that the wavefront at A lags that at B, since when the wavefront crosses a plane normal to the direction of travel, point A of the wavefront crosses that plane behind point B. The wavefield produces, at the plane P, a field $ae^{i\phi}$, where a is the amplitude and ϕ is the phase of the wavefront at the plane P.

We place a phase conjugator in the path of the wave. The wave passes through the conjugator, emerging as a phase conjugated wave, in which the point B, which formerly led point A in phase, now lags behind it by a corresponding amount. This illustrates the basic meaning of the term 'phase conjugation'. This device could be considered as a forward phase conjugator, since the phase conjugated wave travels in the same direction as the original wave.

Next (Figure 1b) we describe a backward phase conjugator. The conjugate wave travels in a direction opposite to that of the incident wave. The conjugate wave emitted by the phase conjugator seems to be

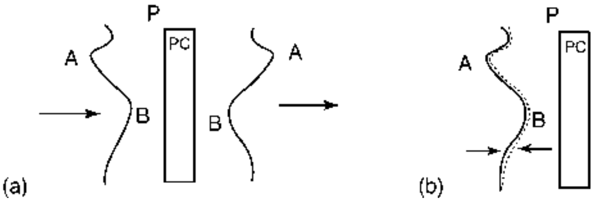

Figure 1 A phase conjugator. (a) shows a forward phase conjugator; (b) shows a backward phase conjugator. For clarity, in (b) the conjugated wave is shown with dotted lines. PC is phase conjugator, and P is the entrance plane of the conjugator.

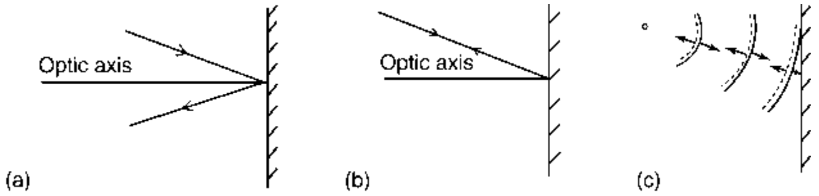

Figure 2 Light reflected from mirrors. (a) Light falling on a conventional mirror and being reflected; (b) light reflected from a conjugating mirror; (c) as in (b), except that illustration is with wavefronts instead of rays.

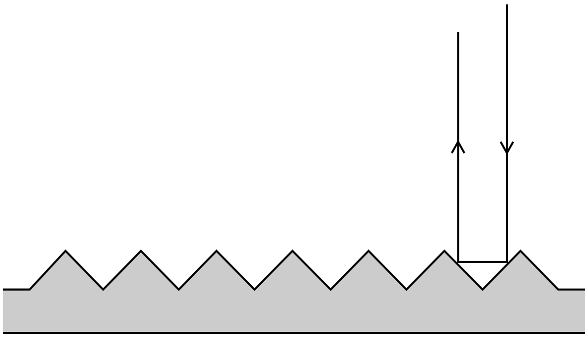

Figure 3 A ray of light reflected from a retroreflector (or corner cube).

identical to the incident wave; we show the two waves to be in coincidence, or more precisely, for purposes of clarity, almost in coincidence. The points A on the two wavefronts coincide, as do the points B. As before, on the incident wave, point A lags behind point B. On the exiting wave, since the wave is traveling in the opposite direction, point B is now lagging point A, i.e., is lagging in phase. In either case, forward or backward conjugation, the wave emerging from the conjugator has the complex amplitude $a(x, y)e^{-i\phi(x,y)}$.

A phase conjugator (backward variety) is like a mirror, in that it obeys the basic mirror law, that the angle of reflection equals the angle of incidence, but in a rather different way, as Figure 2 shows. In fact, the reflected ray travels back on itself. A simple way to generate a conjugate wave, and one that has been known for a very long time, is with the use of an array of retroreflectors, or corner cubes. As shown in Figure 3, each ray undergoes two reflections and emerges traveling in exactly the opposite direction, although with some displacement. If we considered the third dimension, there could be three reflections. The amount of displacement depends on where the ray strikes; if it strikes near the apex of the corner cube, the ray displacement will be slight; if at the edge, the displacement is greatest. Thus, the image formed in reflection is degraded. This geometrical degradation is greater for larger corner cubes. Thus, why not make the cubes very small? Then, however, the image is degraded by diffraction effects, since each corner cube constitutes a small aperture. Thus, there is an optimum size for the cubes; if too large, geometrical factors degrade the image; if too small, diffraction effects degrade the image.

Phase Conjugation by Holography

With holography, however, came a method of conjugating a wavefront that yields high-resolution imagery. The holographic process inherently produces a phase conjugate image, as was noted by Gabor in his early papers on holography. The holographic process is illustrated in Figure 4. For simplicity, we assume the object to be a planar transparency $s(x, y)$. A coherent light wave, generally a plane or spherical wave, passes through the transparency and a diffracted field u is produced at the recording plane P, as in Figure 4a. The field u is typically the Fresnel diffraction pattern of the object s, but if the propagation distance is sufficiently great, u could be the Fraunhofer pattern of s. Of course, lenses can be placed between the object and recording planes, thus modifying the diffraction pattern.

The field u is combined with a reference wave u_o, which is generally an off-axis planar or spherical wave. The two beams produce an interference pattern with intensity

$$I = |u_o + u|^2 = |u_o|^2 + |u|^2 + 2Re\, u_o^* u \qquad [1]$$

where the final term can be written as $u_o^* u + u_o u^*$; the * denotes complex conjugate. The process of forming the interference of the two beams, i.e., forming the intensity of their sum, has generated the conjugate term u^*. When the intensity is recorded, as on photographic film, for example, both terms are stored on the film. Suppose for simplicity that u_o is a plane wave $a_o e^{i2\pi f_o x}$, where a_o is a constant, and $f_o = (\sin\theta)/\lambda$, where θ is the angle that the reference wave makes with the optic axis and λ is the wavelength of the light. Also, u can be

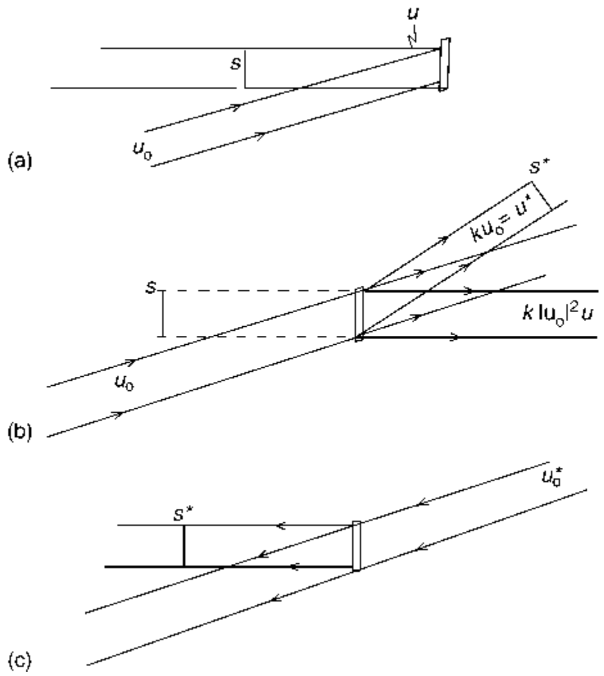

Figure 4 The holographic process. (a) Making the hologram; (b) reading out the hologram, using a readout beam that duplicates reference beam; (c) reading out with reversal of reference beam.

written as a product of an amplitude term a and a phase term $e^{i\phi}$, or $u = ae^{i\phi}$. Let the photographic plate, after exposure and development, have a transmittance t proportional to the intensity I, or $t = kI$, with k a constant; this is an oversimplification of the recording process, but preserves the essence of the holographic process. Now let the photographic plate, which after development becomes a hologram, be illuminated with a readout wave that, for simplicity, we let be u_o, a duplicate of the reference wave. Then, the field emerging from the hologram is

$$u_e = u_o k\left[|u_o|^2 + |u|^2 + u_o^* u + u_o u^*\right] \qquad [2]$$

Only the last two terms are of interest. These can be written as $k|u_o|^2 u$ and $ku_o^2 u^*$, respectively, or alternatively, as $ka_o^2 ae^{i\phi}$ and $ka_o^2 ae^{i(4\pi f_o x - \phi)}$. The term $k|u_o|^2 u$ represents a wave that is, to within a constant, identical to the object wave u; the other term represents a wave that is, to within a constant and a linear phase term, a conjugate of the original wave. The factor $e^{i4\pi f_o x}$ has the physical meaning that the wave is propagating in a direction 2θ given by $\sin 2\theta/\lambda = 2f_o$. The readout waves are shown in Figure 4b.

The so-called true wave u and the conjugate wave u^* thus propagate angularly away from each other and from the direct transmitted beam, which carries the uninteresting components $|u_o|^2$ and $|u|^2$.

However, a more suitable way to form the conjugate wave is to illuminate the hologram with a readout wave that is, again, a duplicate of the original reference wave, but propagating in the opposite direction, as in Figure 4c. The conjugate wave is now propagating along the z axis, along the path of the original object wave. It is readily shown that this counterpropagating wave is identical in shape to the original object wave recorded by the hologram, except that it travels in the opposite direction, exactly the situation portrayed in Figure 2. This wave propagates so as to form an image that is, to within a constant, identical to the original object, except for the counterpropagation, which produces a conjugate image $s^*(x, y)$. The image intensity, however, is $|s|^2$, which is identical to the original image intensity.

This conjugate wave was, in the early days of holography, a detriment, since without the use of the off-axis method (reference wave introduced at an angle to the object wave) the two images were inseparable, and the true image could be observed only against the background of the defocused conjugate image. Thus, an out-of-focus image overlaid the true image, making it noisy.

In 1962, the same year that the off-axis reference wave for separating the two images was introduced, a possible use for the conjugate image was suggested. If the original propagation path of the object wave were aberrated, as would occur if the wave propagated through an irregular medium, then the conjugate wave propagating through the medium would undergo a correction and would arrive at the original object position perfectly corrected, resulting in a sharp image, a process sometimes called wavefront healing. The idea lay dormant for 3 years, until in 1965, experimental demonstration of this idea was reported. In one case, the object was a transparency containing opaque lettering against a transparent background and the inhomogeneity was a piece of frosted glass, which resulted in extremely severe aberrations, so that the object wave, having traversed the diffuser, was no longer capable of forming an image. However, when the distorted wave was recorded holographically, conjugated, and sent back through the same diffuser, the result was a highly legible image.

The principle of wavefront healing is illustrated in Figure 5, when the inhomogeneity is an extended, or three-dimensional, structure. A point object O sends a spherical wave into the inhomogeneity. As the wavefront propagates, it becomes distorted, and loses its sphericity. When it emerges from the inhomogeneity, the wavefront is highly distorted.

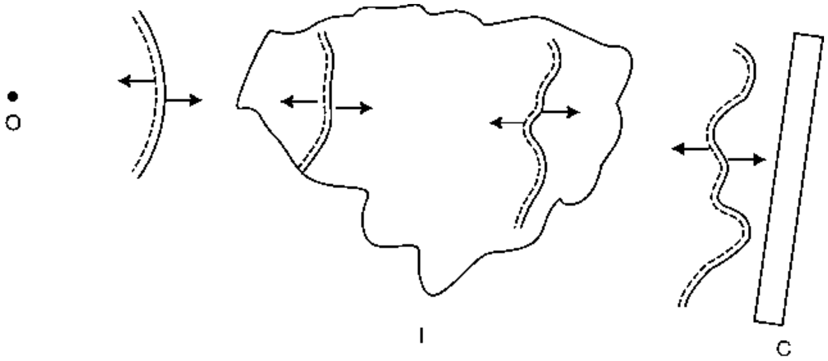

Figure 5 The wavefront healing process. C is the conjugator, I is the inhomogeneity, and O is a point object.

The wave is conjugated at the conjugator C and emerges as a back-propagating wave that retraces the original wave path through the inhomogeneity, gradually losing its aberrations and looking more like a spherical wave. When it emerges from the inhomogeneity, it is a perfect spherical wave that converges to a point image that coincides with the original object. At each plane throughout the inhomogeneity, the back-propagating conjugate wave (dashed lines) has the same shape as did the original wave when it crossed that plane (solid line).

This interesting wavefront healing property of holographic phase conjugation was the beginning of a new branch of optics, called phase conjugation, giving rise to a number of applications. However, the holographic method had the problem that it did not occur instantaneously, but had to await the processing of the photographic film on which the hologram was recorded. This delay limited the range of potential applications. Later, in the 1970s, nonlinear crystals became available that did the equivalent of holography, but in real time, or near real time – a process known as four wave mixing. There are yet other types of real-time phase conjugators with nonlinear devices.

Pseudoscopic Image

As long as the object is only two-dimensional, a function of x and y only, with no extent along the axial dimension, z, the intensity distribution of the true and conjugate images is identical. The conjugateness, being a pure phase phenomenon, disappears when the observable is the image intensity $|s|^2$, and not the complex amplitude s. Since the intensity is always the observable, whether the image is viewed by eye, or with a camera, the conjugateness generally is of no significance.

However, when the object for holography is three-dimensional, the situation is different, and the conjugate image exhibits some strange properties. These arise from conflicting cues pertaining to the three-dimensionality of the image. To explain, we review some basics about how a three-dimensional world is perceived by a viewer. First, there is the stereo effect, resulting from the viewer having two eyes, with each eye seeing an object from slightly different positions, a phenomenon called binocular disparity. The mind uses this stereo effect to perceive three-dimensionality. But there are other cues that also contribute to perception of three-dimensionality. Another is parallax; as the viewer moves his head, objects farther from the observer change their angular position more slowly than objects close by. The observer can sense the three-dimensionality of an object, even with only one eye, by moving the head, thereby inducing parallax. Also, there is interposition, whereby near object portions block the viewing of object portions lying behind them, again giving a sense of depth. There are still other cues, more subjective, such as perspective, parallel lines converging toward infinity, shadowing, etc. With these observations in mind, we consider the hologram under illumination and the two images that it forms (Figure 6). As depicted in Figure 6, these two images lie in mirror symmetrical positions relative to the hologram. Both images have the same side facing the hologram. It must be this way, because when the hologram was recorded, it was side **a** that faced the recording plane. And both images must therefore have side **a** closer to the hologram.

When a viewer observes the true image (a virtual image), he views it looking through the hologram, and he sees the image at the position where the original object was when the hologram was recorded, and it is, in its visual properties, indistinguishable from the actual object; it has all of the three-dimensional and the parallax properties that the original object would exhibit. However, when the viewer observes the conjugate image, he must view it

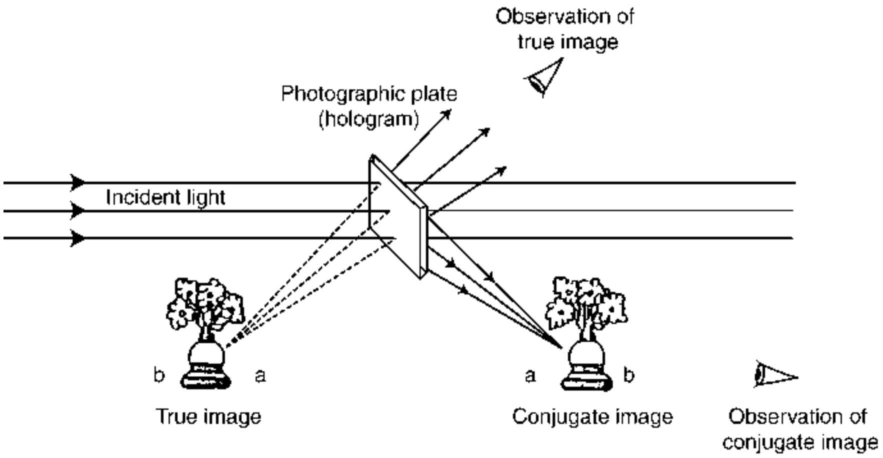
Observation of true image
Photographic plate (hologram)
Incident light
b
a
True image
a
b
Conjugate image
Observation of conjugate image

Figure 6 Formation of the true and conjugate images from a hologram.

from the side of the image opposite from the hologram, i.e., he views the image from the back side. When a viewer observes the conjugate image, he sees it from the back side (side **b**). All of the above cited cues are in agreement about the orientation of the image, viz., that side **b** is closer to the observer than is side **a**. That is, all cues but one. The conjugate image is formed as the hologram recording plate saw it, with side **a** being closer. Thus, all interpositions are formed with the same relation that the object had with respect to the recording plate. The side **a** is the part of the object that was closer to the hologram, so that when interpositions occur, side **a** blocks side **b**. Since the observer thus sees side **a** partially obscuring side **b**, he must conclude that side **a** is closer than is side **b**. However, this observation is in direct conflict with the other cues, especially the stereo and the parallax cues, which are themselves strong cues, as is the interposition cue. The conflict of these cues, with two of them interpreting the image as being formed with one depth orientation, the other with the opposite, gives a number of strange effects. For example, as the viewer moves his head so as to induce parallax, the object appears to rotate. If the image is that of a human head, for example, the observed effect is that as the viewer changes his position, the head appears to rotate in synchronism with the observer, as if the image had its gaze fixed on the viewer, an effect that is often perceived as being eerie.

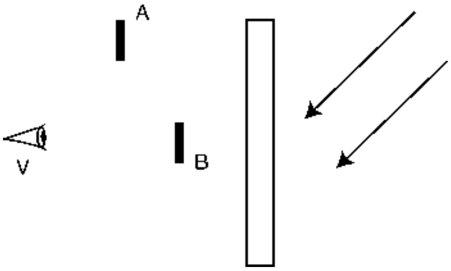

Figure 7 Pseudoscopic conjugate image formed from an object consisting of two plates. The viewer V perceives plate A as being closer than B and there are no conflicting cues, since there is no interposition. However, if the observer moves his head upward, there is partial interposition, thus conflicting cues, and the sense of three-dimensionality is lost.

Similarly, if the object were two flat plates, (Figure 7) in the situation depicted, the planes have no overlap, i.e., no interposition, as seen by the observer. And the observer unequivocally perceives plate A as being closer than plate B. However, if the viewer moves his head to a position where interposition occurs, the conflict of cues arises, A becomes partially blocked by B, which is quite correct as observed at the position of the hologram, but is quite incorrect from the viewer's vantage point. The result of this conflict is that the viewer, who previously perceived the three-dimensionality, now sees the three-dimensionality perception suddenly collapse, and the sense of three-dimensionality is lost. Other strange perception anomalies can also be produced. Such images, with conflicting depth cues, are called pseudoscopic, as opposed to orthoscopic images, where the cues do not conflict.

One Way Phase Conjugation

A significant limitation of phase conjugation aberration correction is that the image and object are on the same side of the inhomogeneity. One would usually want them to be on opposite sides. In a modified form

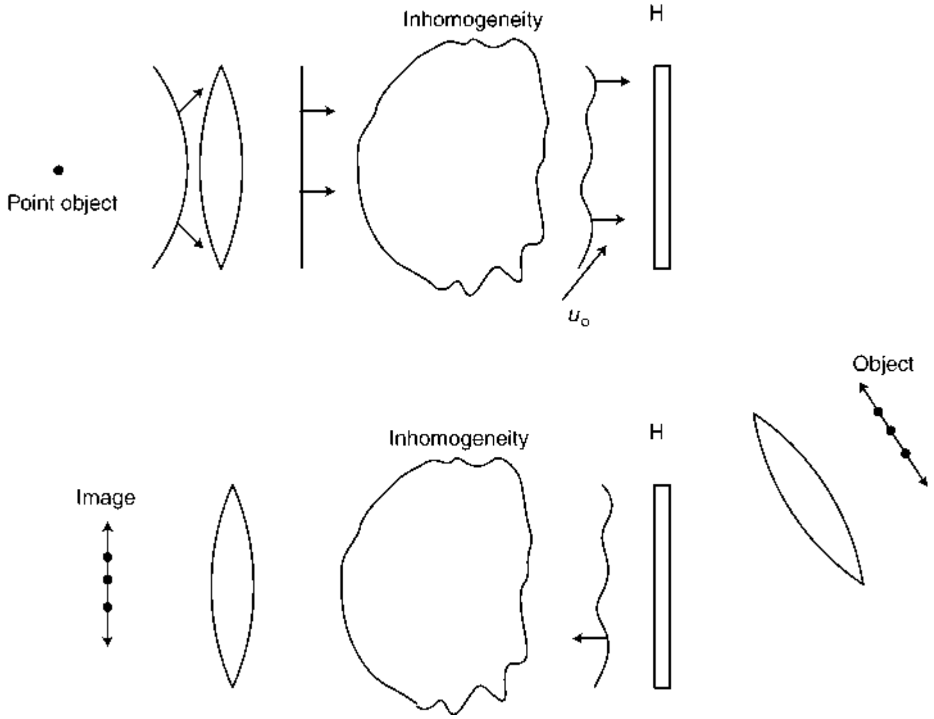

Figure 8 One-way phase conjugation. Top: Making the hologram. Bottom: The viewing process. We show three points on the object. The center point forms a well-corrected image, but the other two points produce imperfectly corrected images.

of the phase conjugation process, shown in Figure 8, this result is achieved. The first step of the process is carried out by passing a plane wave, instead of the object wave, through the inhomogeneity. This plane wave originates from a point source located at the focal plane of a collimating lens. The plane wave thus carries information about the inhomogeneity to the recording plate, where it is combined with a reference wave and a hologram of the distorted plane wave is recorded. Thus, information about the phase defect is stored on the hologram.

Next, the hologram is read out, not with a counterpropagating replica of the reference wave as before, but instead with an object-bearing wave that passes through the hologram. This wave emerges from the hologram and now has on it the recorded aberration information. As it counterpropagates along the path of the original object wave, it passes through the inhomogeneity, whereupon the aberration is cancelled and a corrected image forms at the focal plane of the collimating lens, where previously the point source had originated. The goal is achieved; the object and the corrected image are now on opposite sides of the inhomogeneity.

This correction is only approximate. To explain, we think of the object as being a collection of many points. The radiation from each point is converted into a plane wave by the lens, with different object points producing plane waves propagating at different angles. These plane waves then impinge on the hologram and emerge with each one bearing the aberration stored on the hologram. They continue on their paths and enter the inhomogeneity. The plane wave component, whose path direction matches that of the reference wave used in recording the hologram, will emerge from the inhomogeneity as a perfectly corrected wave, thus forming a well-corrected point image at the position of the original point radiator. However, another plane wave component, traveling in an incorrect direction, will be laterally displaced from the proper position for entering the inhomogeneity, and the aberration correction will be imperfect, the degree of imperfection being in direct relation to this displacement error. The displacement error is the product of the angular error and the propagation distance between hologram and inhomogeneity. If the inhomogeneity is slowly varying as a function of x and y, and if the object does not have a large extent, so that the range of angles is small, the image degradation will be small.

If the last two conditions do not apply, two options are available. First, the distance between inhomogeneity and hologram can be reduced, either by bringing the hologram closer to the inhomogeneity, or by imaging the hologram onto the inhomogeneity. The correction is now valid over a wider extent of the image, especially if the inhomogeneity can be treated as planar rather than extending over a volume. Alternatively, a scanning process can be carried out on the object beam falling on the hologram, so that in sequence, each plane wave component assumes the proper angle to counter-propagate along the path for perfect correction, and will form a well-corrected image point. This output spot then falls on a pinhole positioned to separate it from other, imperfectly corrected points. Then, over time, all points of the object will form well-corrected image points, and the entire object will thus be well-imaged. The one way phase conjugation method was first used to correct the spherical aberration of a lens, and the result was quite successful.

Phase conjugation has become an important branch of modern optics. The process is often carried out using photographic materials. More often nowadays, it is carried out using nonlinear crystals. Some of the crystal methods are essentially holography in real time, but others, such as phase conjugation by Brillouin scattering, go beyond holography.

See also

Nonlinear Optics, Basics: Nonlinear Optical Phase Conjugation. **Phase Control:** Wavefront Coding.

Further Reading

Caulfield HJ (ed.) (1979) *Handbook of Optical Holography*. New York: Academic Press.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw-Hill, chap. 9. This chapter is devoted to holography, with section 9.12.3 dealing with phase conjugation aberration correction.

Kogelnik H (1965) Holographic image projection through inhomogeneous media. *Bell Systems Technology Journal* 48: 2451–2452.

Leith EN and Upatnieks J (1965) Holograms: their properties and uses. *SPIE Journal* 4: 3–6. This is the earliest publication demonstrating the principle of wavefront healing by holographic phase conjugation.

Smith HM (1975) *Principles of Holography*, 2nd edn. New York: Wiley.

Wavefront Coding

W T Cathey, University of Colorado, Boulder, CO, USA

E R Dowski, CDM Optics, Inc. Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We describe a new paradigm for designing imaging systems that have performance that is otherwise unobtainable. These are hybrid systems that use the optics of the system to acquire an ‘intermediate’ image, which is processed to produce the final image. The optical components of the system are designed jointly with signal processing to optimize overall performance. It is possible to extend the depth of field of an imaging system, for example, by a factor of eight to ten. On the other hand, by concentrating on the increase of the depth of focus in the detector plane, it is possible to use that increase in the depth of focus to make the system invariant to focus-related aberrations. By doing this, it becomes possible to design a single-lens imaging system with small f-number and large field of view. Another example of the sort of new imaging system that can be designed is a two-lens zoom lens that images with a large field and a very short overall length.

First, the term ‘Wavefront Coding’ is defined as employed with imaging systems, and a description is given of how such coding is achieved. Examples of means of applying Wavefront Coding are given along with the analysis and design tools that are useful in designing hybrid optical/digital imaging systems. Finally, some potential future applications of Wavefront Coding in imaging systems are suggested.

Background and Basic Concepts

Some early suggested modifications to imaging systems can be viewed as wavefront modification, even though they were not described in that manner at the time. These include the work of Ojeda-Castañeda that described masks to block portions of the lens to extend the depth of field. In addition, it has long been known that the depth of field is slightly greater in lenses that have spherical aberration. In both of these cases, the wavefront is modified, in one case by changing the wave amplitude, and in the other by a deviation from the traditional ideal phase front. However, in both cases, the increase in the depth of field was only a few percent. Häusler extended the depth of field by requiring that the focus be continuously changed during the exposure time of the camera. This resulted in a modulation transfer function (MTF) for the imaging system that was the incoherent superposition of a spread of MTFs for the different foci during the exposure. To form an acceptable image, signal processing was used.

The use of Wavefront Coding here refers to new optical systems arrived at by designing the modulation of the wave and signal processing to achieve previously unobtainable imaging modalities. These new imaging systems can have more than an order of magnitude increase in depth of field, for example. The coding of signals so that particular information is optimally conveyed is not new. For example, in radar, coding of the pulses is done to optimally provide information concerning a target's range. At the radar receiver, signal processing is required to extract the range information. In a similar manner, an imaging system can have a coded wavefront in order to provide range or misfocus information in a manner such that signal processing can easily extract the desired information in the scene. Alternatively, the wavefront can be modified so that the imaging system is invariant to object depth or to misfocus of the system, so that, after processing, good resolution can be obtained over a large image depth.

A primary feature of Wavefront Coding in imaging is that the optical image-gathering portion of the system can be designed in combination with the design of the signal processing. This is normally not done in the design of imaging systems. It is done, however, in tomography, coded aperture imaging, and sometimes, interferometric imaging. In 1984, a group pointed out the potential of increasing the performance of imaging systems by jointly designing the optics and the signal processing. Since that time, major advances have been made in doing this. Now, several analysis and design tools are available, and are described in the following section.

Theory and Applications of Wavefront Coding

Wavefront Coding refers to a modification of the imaging system by placing specialized phase plates (or phase and amplitude masks) in the aperture stop, or in an image of the aperture stop, of the imaging system. This could be a phase plate as a separate optical element, or, if a surface of a lens is close to the aperture stop, the Wavefront Coding surface can be superimposed onto the normal lens surface. Processing to decode the image that is formed by the detector array (the intermediate image) is related to the coding that is impressed on the wavefront. The term 'signal processing' also includes the spatial integration provided by the width and height of the pixels. The final image is hence a function of the optics that form the intermediate image and the signal processing that produces the final image.

In the following section, we discuss the Woodward function as used in analysis and design of hybrid imaging systems, and show visualizations of the effect of Wavefront Coding on the image performance of two dramatically different types of systems. Then, other design techniques are discussed for imaging systems using Wavefront Coding and several examples of systems are given where the depth of field of imaging systems is extended by at least an order of magnitude. By concentrating on the increase in the depth of focus, rather than the depth of field, very simple imaging systems can be designed where Wavefront Coding makes the imaging system invariant to focus-related aberrations, and the shape of the lens is used to minimize the other aberrations.

Wavefront Coding Theory

The Woodward function

P. M. Woodward showed that the capability of a radar system to estimate both the range and the velocity of a target can be described by a mathematical function that became known as the Ambiguity Function. The Ambiguity Function shows how well systems that use different types of radar pulses can measure the transit time of the pulse (hence the range to the target) and the Doppler shift in the frequency of the returned pulse (hence the velocity of the target). Because the parameters, range, and Doppler, are measured with different accuracy, there is ambiguity in the measurement, and the Ambiguity Function is a display of the tradeoffs that are possible, Brenner, Lohmann, and Ojeda-Castañeda showed that the same mathematical function, with differently defined variables, can represent the Optical Transfer Function (OTF) of a one-dimensional imaging system.

The Ambiguity Function for a radar system is given by

$$A(u, v) = \int P(x + u/2) P^*(x - u/2) \exp(i2\pi v x) dx \quad [1]$$

where $i = \text{sqrt}(-1)$, * denotes complex conjugate, u is time, v is frequency, x is a dummy variable, and $P(x)$ is a complex function representing the transmitted waveform of the radar system. The OTF of a one-dimensional imaging system for different values of misfocus can be written as

$$H(u,\psi)=\int\left\{P(x+u/2)\exp(j(x+u/2)^2\psi)\right\}\times\left\{P^*(x-u/2)\exp(-j(x-u/2)^2\psi)\right\}\mathrm{d}x \qquad [2]$$

where $P(x)$ represents the complex pupil function of the imaging system. The misfocus parameter ψ can be shown to be equal to the physical parameters:

$$\psi=(\pi L^2/(4\lambda))(1/f-1/d_\mathrm{o}-1/d_\mathrm{i}) = 2\pi W_{20}/\lambda = kW_{20} \qquad [3]$$

where L is the one-dimensional length of the aperture, and λ is the center wavelength of the narrowband illumination. The distance d_o is measured between the object and the first principal plane of the lens, and d_i is the distance between the second principal plane of the lens and the image plane. The focal length of the lens is given by f. The wavenumber is given by k and the traditional misfocus aberration constant is given by W_{20}. A comparison of the form of the ambiguity function and the OTF shows that:

$$H(u,\psi)=A(u,u\psi/\pi) \qquad [4]$$

The mathematics of the Woodward function as used for the Ambiguity Function has been studied extensively. Hence one can apply the theory that was worked out for the Ambiguity Function to the study of imaging systems. This gives a tool that is, in many ways, more powerful than other forms of three-dimensional optical transfer functions. For example, it can be shown that the squared volume under the Woodward function is a constant. This means that whatever one does to modify the Woodward function, there is a constraint on what can be done. Another constant is the power under a vertical cut of the Woodward function at a particular value of spatial frequency. This means that there is constant power for all misfocus values at any given spatial frequency. It can be bunched up about the normal in-focus position, or it can be spread out as shown below, but the squared area does not change.

Examples of the Woodward function for different imaging systems

A display of the Woodward function for a high-quality, one-dimensional lens is shown in Figure 1a. The horizontal line and the slanted line represent in-focus and out-of-focus slices through the Woodward function. A projection onto the horizontal axis gives

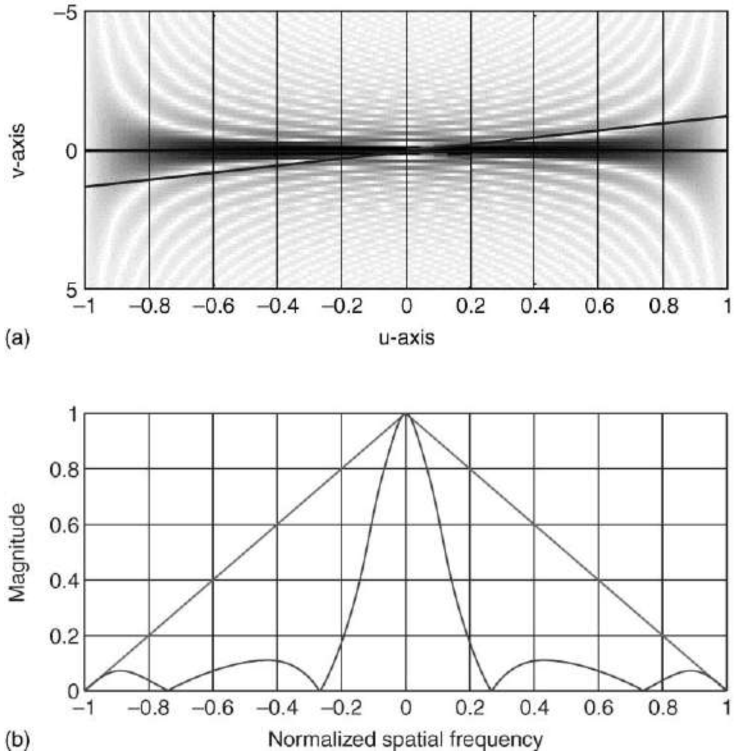

Figure 1 (a) The Woodward function for a high-quality one-dimensional lens with slices that represent in-focus and out-of-focus imaging. (b) The corresponding values of the modulation transfer function for those two cases.

the OTFs for those two focal conditions. The magnitudes of those two OTFs, the MTF, is shown in Figure 1b. Note that the out-of-focus MTF goes through a zero, where there is a phase shift, which gives a contrast reversal in the image.

Extended depth of field. Figure 2a shows the corresponding display of the Woodward function for a one-dimensional lens with a cubic phase function, $\phi = \alpha x^3$ added in the aperture stop. The appearance of the Woodward function immediately shows that this imaging system will have an increased depth of field, because the Woodward function shows that the energy is spread over a larger misfocus region. The slices in the Woodward function are taken for the same misfocus values as were used for the images in Figure 1. Notice that the MTFs of Figure 2b are very similar, and that the MTF for the misfocused case does not go through a zero. This implies that the imaging system will have an extended depth of focus.

The analogy to the Ambiguity Function of radar was key in selection of the cubic function for early demonstrations of extending the depth of field of an imaging system. Work in the radar field had shown that a radar pulse, where the carrier frequency increases quadratically with time, has an Ambiguity Function that is identical to the distribution in Figure 2a. Because a quadratic in time is a cubic in phase, the cubic phase function:

$$\phi = \alpha x^3 \qquad [5]$$

was selected as one that will provide the same Woodward function, and hence an increase in the depth of field. Many families of phase distributions have now been found that extend the depth of focus of an imaging system.

Antialiasing. When the resolution of the optics of an imaging system exceeds the resolution of the image detector array, the high spatial frequencies cannot be detected. They are not lost, however. When a spatial frequency is sampled below the Nyquist sampling rate, the high spatial frequency shows up, or aliases, as a lower spatial frequency. This aliased frequency is equal to the sampling frequency (determined by the pixel spacing) less the difference between the high frequency that has been undersampled and the sampling frequency. This results in an image with broad bands (low spatial frequencies) across the image as is often seen in television images of patterns with high spatial frequency content, such as striped shirts. Because the signal power at a given spatial

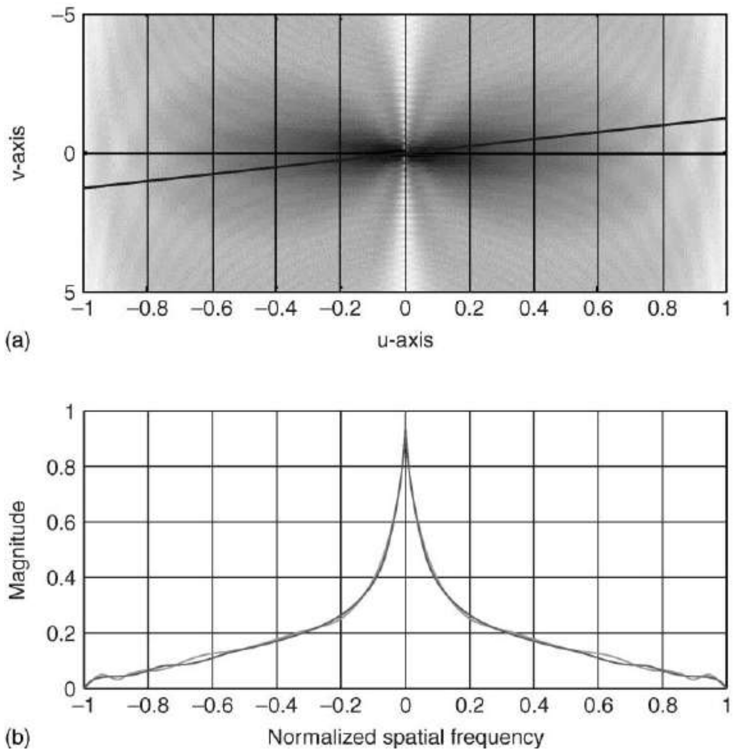

Figure 2 The Woodward function and the modulation transfer functions for a high-quality one-dimensional lens that has a phase plate with a cubic profile in the aperture stop. (a) The Woodward function showing a broadened pattern that is indicative of an increased depth of field. (b) The MTFs that are associated with the slices that represent in focus and out of focus.

frequency over all misfocus (the squared area under a vertical slice of the Woodward function at that spatial frequency) is constant, these frequency components cannot be removed. They can, however, be reduced at a given focus value by spreading the Woodward function vertically. Figure 3 shows a Woodward function that would be ideal. Below the cutoff spatial frequency, all of the signal power is in the focal plane, and beyond the cutoff frequency, the power is spread as much as possible over a large misfocus region. There is no assurance that a phase distribution can be found to give a desired Woodward function, however. Figure 4 shows the effect of one phase function, given by

$$S(r, \theta) = r^3 \cos(3\theta) \qquad [6]$$

upon the MTF of the imaging system. The dashed vertical line shows the cutoff spatial frequency that is due to the sampling rate of the pixels. Note that in Figure 4, the magnitude of the Wavefront Coded spatial frequencies that are aliased have lower values than do those that are passed by a traditional full aperture F/2.5 system. Stopping the traditional system down to F/25 reduces the aliased MTF to an amount close to that obtained with the wide aperture and Wavefront Coding, but it also reduces the amount of light by a factor of 100.

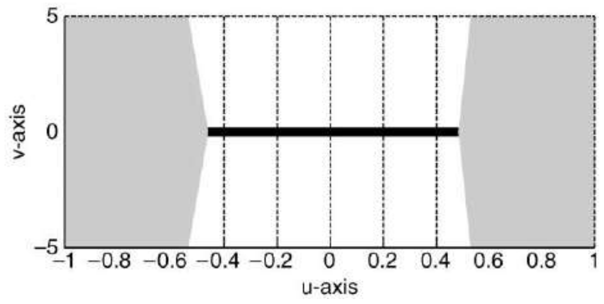

Figure 3 The Woodward function for an ideal antialiasing filter.

Wavefront Coding in Imaging Systems with Extended Depth of Field or Focus

Figure 2 shows the Woodward function for one phase distribution (cubic in x and y for a two-dimensional imaging system) that increases the depth of field of an imaging system. The cubic function was selected by using analogies to the Ambiguity Function of radar. Even though it is doubtful that this phase distribution could have been found by using conventional lens design techniques, it is instructive to show the ray traces and point spread functions that are obtained when this phase distribution is used in the aperture stop.

Ray trace analysis

Figure 5 shows the ray trace from a point object through a conventional (one-dimensional) lens and through a lens that has been modified with the cubic profile of eqn [5]. Figure 5a shows the rays through a conventional lens coming to a focus, and Figure 5b shows the corresponding rays as they go through a lens with a cubic profile added. In the latter case, the

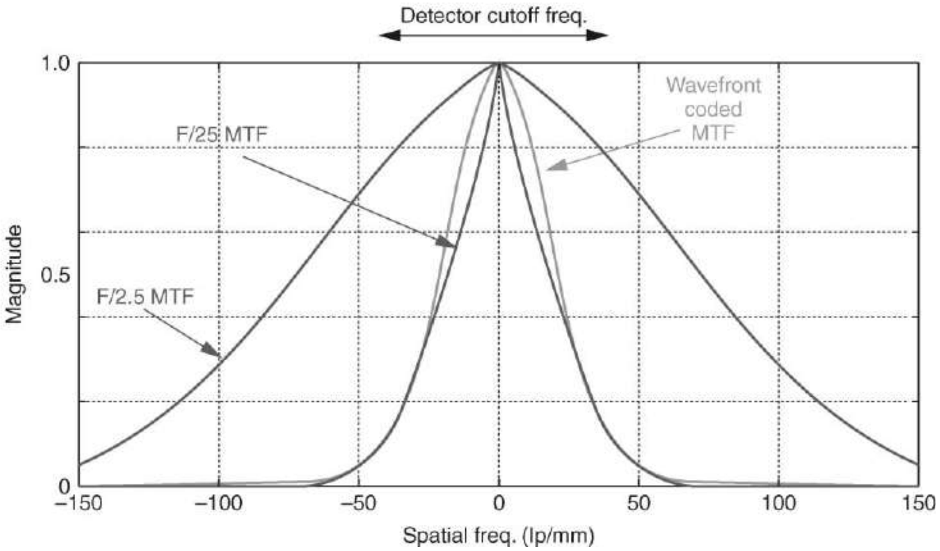

Figure 4 Normalized MTFs for traditional optics and Wavefront Coding. The full aperture F/2.5 system has a large amount of optical power beyond the detector cutoff frequency of 50 lp/mm. With Wavefront Coding the aliased optical power can be greatly reduced. The traditional system aperture needs to be reduced to F/25 to compare to the antialiasing performance of the Wavefront Coded system. This reduction reduces the light level by a factor of 100 in comparison with the F/2.5 system.

rays cross, but have a relatively constant ray density in the region of what was the focal region. Figure 6 shows experimentally acquired point spread functions (PSF) for both cases. Note that the PSF changes very little over the region of misfocus. After signal processing, the PSF can be brought back to that of the conventional imaging system when in focus. Figure 7 illustrates the effect of the signal processing upon the intermediate image that is acquired by the optics. The object is a tilted bar pattern, so that a portion is in focus, and the remainder of the bar pattern is so far out of focus that there is a contrast reversal in the image. With the system that has a coded wavefront to give an extended depth of field, the entire image appears to have the same blur, no matter which part of the tilted bar pattern is observed. This is because the same PSF is convolved with the object distribution for all regions of misfocus. Signal processing removes the effect of the modified PSF, resulting in the image of Figure 7b. Figure 7c shows a trace through the image of Figure 7b. The amount of the misfocus that can be tolerated is a design parameter, and is discussed in the next section.

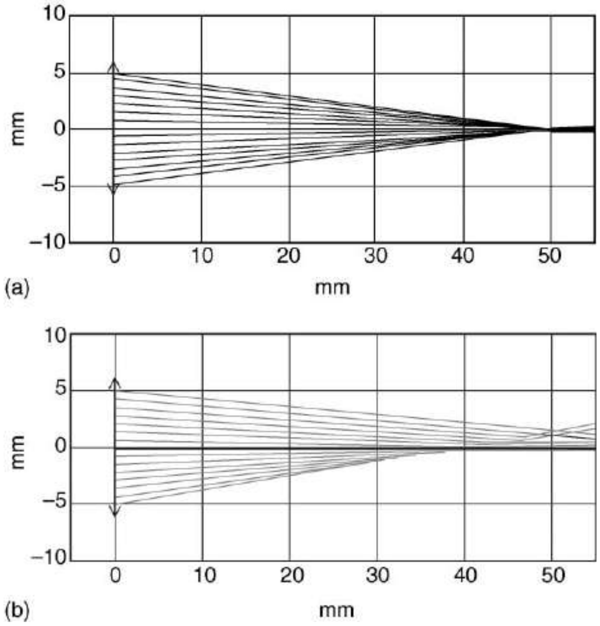

Figure 5 Ray traces from a point object through (a) a conventional one-dimensional lens, and (b) a lens that has been modified to have a cubic profile on the surface.

Signal processing requirements

It is necessary to do signal processing when using an imaging system that employs Wavefront Coding; the signal processing is needed to ‘decode’ the image that is formed by the detector array. This intermediate image is not a clear image. To obtain an intermediate image, the object distribution is convolved with a PSF, or impulse response, that has a shape similar to that

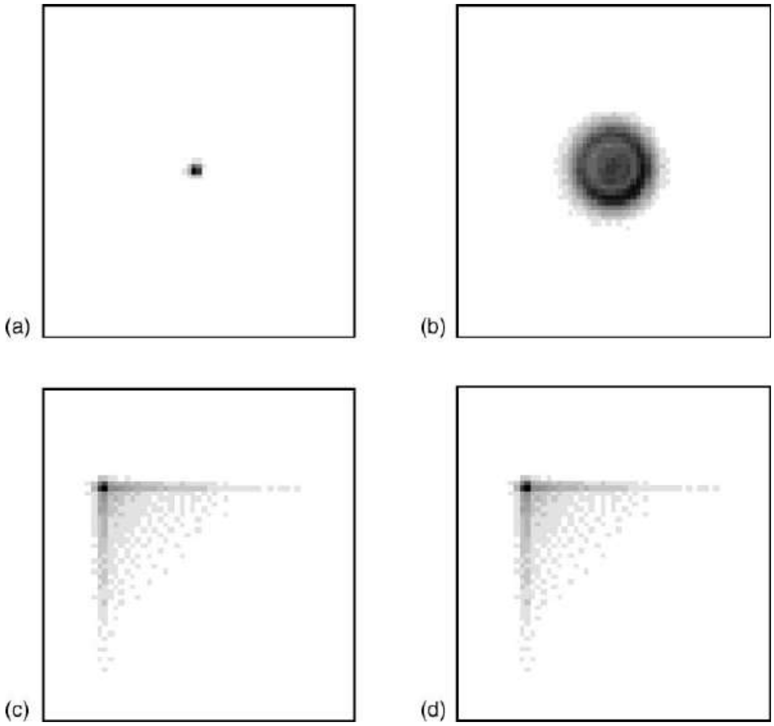

Figure 6 Point spread functions for a conventional lens in focus (a) and with large misfocus (b), and a lens that has been modified to have a cubic profile on the surface when in focus (c) and with a large misfocus (d).

of Figure 6c and d. To obtain the final image, the intermediate image must go through a deconvolution with the PSF. Because of the square law detection that occurs in the detector array, this deconvolution does not simply undo the effects of the convolution with the PSF that describes image formation. A straightforward digital convolution is required in most cases, and has been shown to run at video frame rates in software.

Parameter trade offs in imaging systems with extended depth of field

Depth of field and signal-to-noise ratio. The greater the phase modulation by the phase plate (with a cubic profile in this example), the greater the extension of the depth of field. The greater the phase modulation for extending the depth of field, the larger the PSF, and the greater the drop in the MTF, as seen in Figure 2b, which show the MTF of the imaging system. The use of the phase function to extend the depth of field causes portions of the MTF to sag as compared with the MTF of the in-focus conventional system. This reduces the signal-to-noise ratio. If there is a greater phase alteration, the MTF sags more.

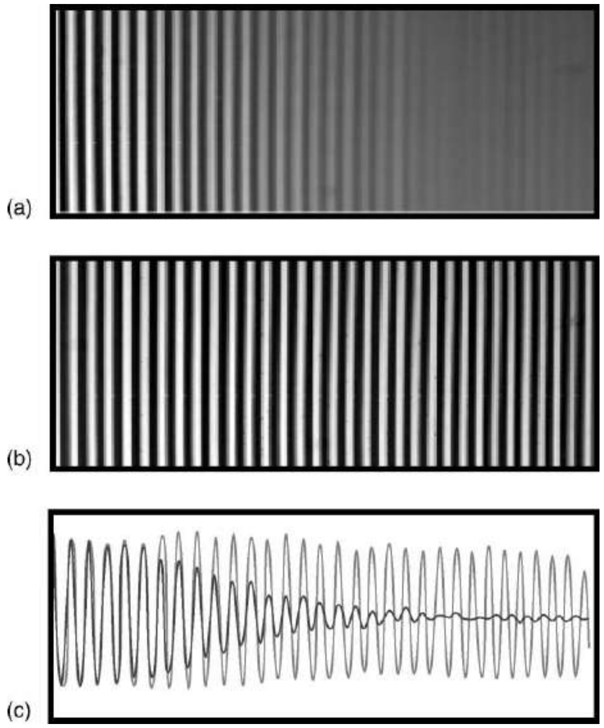

Figure 7 Image of a tilted bar pattern with an extended depth of field imaging system. (a) Image with a conventional imaging system. (b) Final image from the imaging system with Wavefront Coding. (c) Traces through the images to give quantitative comparisons of the image quality.

Focus budget and focus-related aberrations. Another set of trade offs can be seen by considering the imaging systems from the detector plane. Here, we see an increase in the depth of focus of the system. All of this increase in the focal depth can be used to increase the depth of field if the optics of the system are of high quality. Otherwise, the increase in the depth of focus can be used to compensate for focus-related aberrations. For example, chromatic aberration occurs when different colors focus to different planes. Figure 8 shows that if the depth of focus is extended, then there is a region where red, blue, and green, for example, are in focus. Figure 9 shows the red, blue, and green images of a US Air Force test pattern, as formed with a two-element imaging system with severe chromatic aberration, and with the same system when a Wavefront Coding element is used to extend the depth of focus.

Compensation can also be seen for the case of curvature of field, where the image of a plane object falls onto a curved surface. If an extension of the depth of focus causes the focal region to resemble a curved bar instead of a curved line, then it is possible to place a flat detector array in the curved region, and therefore obtain a good image after signal processing. Similar arguments can be made for other focus-related aberrations. This leads to the concept of a focus budget, where a portion of the increased depth of focus can be allocated to extend the depth of field by some amount, and other portions of the depth of focus allocated to provide invariance to focus-related aberrations.

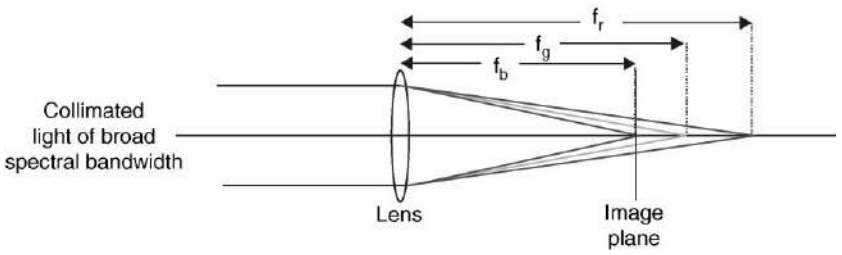

Figure 8 Ray traces showing different focal planes for red, blue, and green, and the effect of extending the depth of focus for each color.

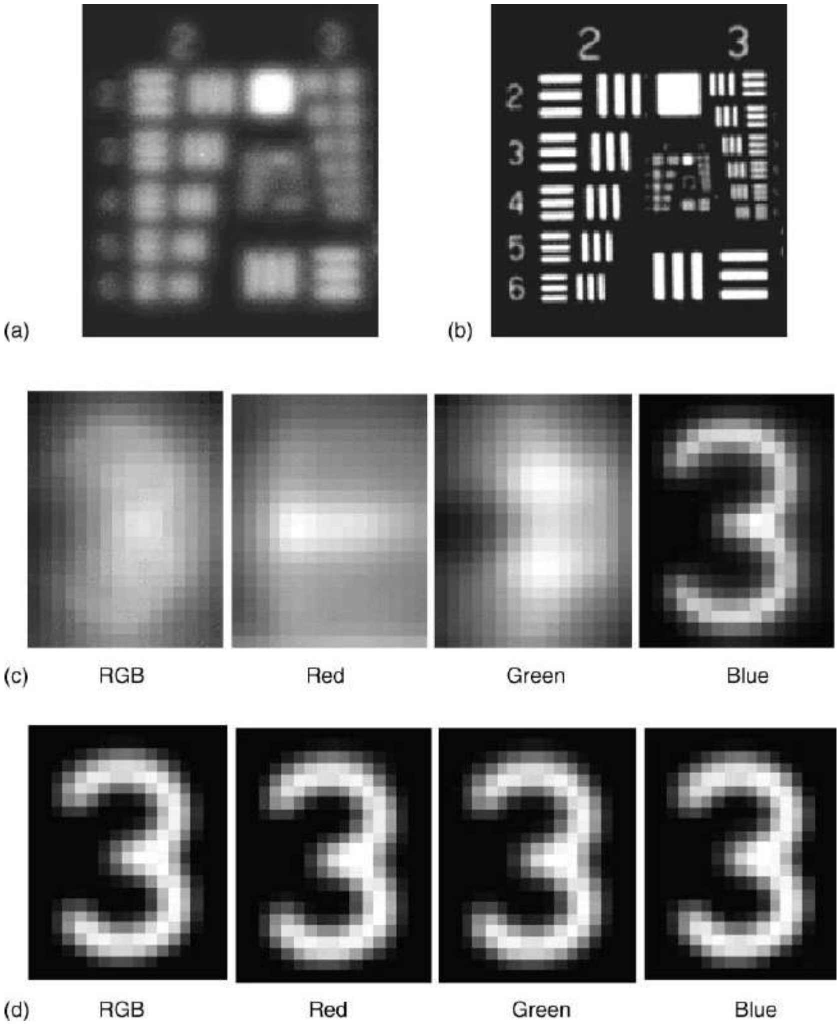
(a)
(b)
2
3
2
3
4
5
6
(c)
RGB
Red
Green
Blue
(d)
RGB
Red
Green
Blue

Figure 9 Color images of a US Air Force test pattern. (a) Images from a two-element lens of the center portion of the test pattern. (b) Images from an extended-depth-of-focus imaging system of the center portion of the test pattern. (c) Images from a two-element lens of a small number 'three' on the test pattern. (d) Images from an extended-depth-of-focus imaging system of a small number 'three' on the test pattern. Some residual effects remain due to different magnifications at different wavelengths.

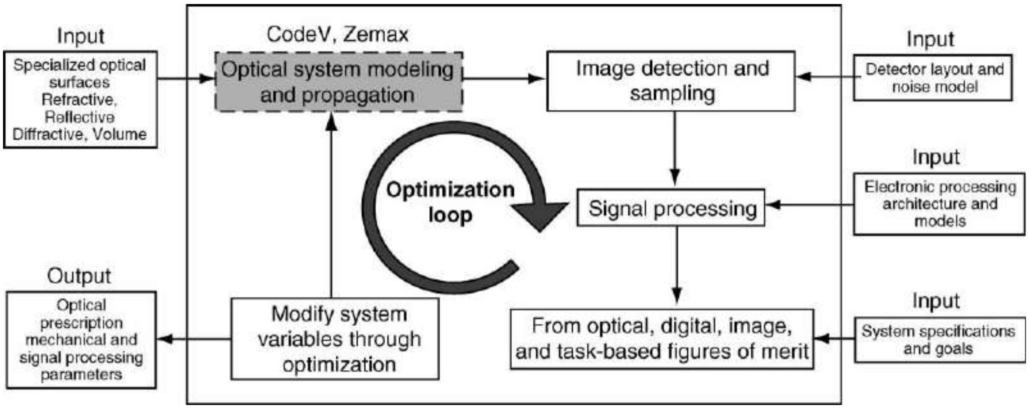
Input
Specialized optical surfaces Refractive, Reflective Diffractive, Volume
CodeV, Zemax
Optical system modeling and propagation
Image detection and sampling
Input
Detector layout and noise model
Optimization loop
Signal processing
Input
Electronic processing architecture and models
Output
Optical prescription mechanical and signal processing parameters
Modify system variables through optimization
From optical, digital, image, and task-based figures of merit
Input
System specifications and goals

Figure 10 Schematic of a tool for designing imaging systems with Wavefront Coding. A ray trace determines the distribution in the exit aperture, and software determines the effect of the detector array and the signal processing. Then the ray trace and signal processing can be modified to optimize the entire system.

New design tools for imaging systems with wavefront coding

The Woodward function is a very good tool for analyzing imaging systems with a coded wavefront and is a good guide for someone who is expert in designing such systems. However, another approach is better for those designers who are more familiar with conventional lens design tools. If one combines a conventional ray-tracing tool with custom software, it is possible to jointly design the optical and the signal processing portions of the system. A schematic of such a system is shown in Figure 10. The ray-tracing tool is used to trace the rays and determine the distribution in the exit aperture. Various software tools then are used to take into account the effect of pixel size and spacing, and the signal processing. The criteria for optimizing the design can include the size of the digital filter, acceptable noise gain, and other parameters related to signal processing as well as the parameters of the optical elements.

Example applications of hybrid imaging systems with an extended depth of field or focus

Barcode and label readers. The images of Figure 7 show that image-based bar code reading is a good application for imaging systems that are based on Wavefront Coding. A bar code or label reader need not be positioned nearly as accurately as without the

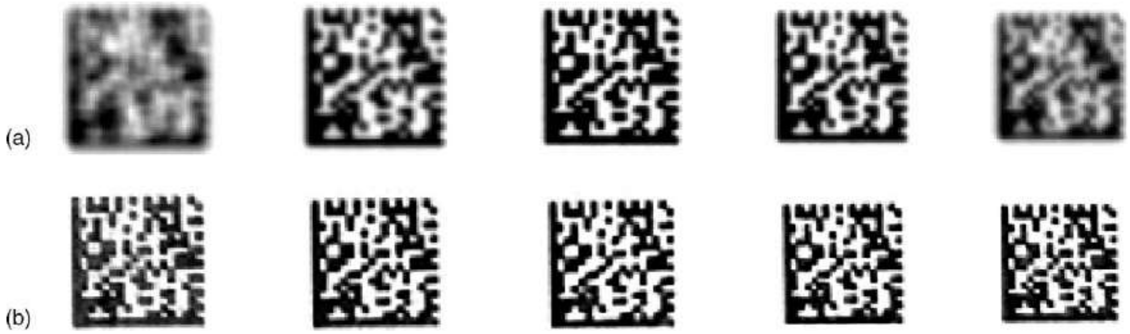

Figure 11 Images of a 2D bar code as the object position is varied for (a) conventional images and (b) Wavefront Coded images after signal processing for the same object positions as in (a).

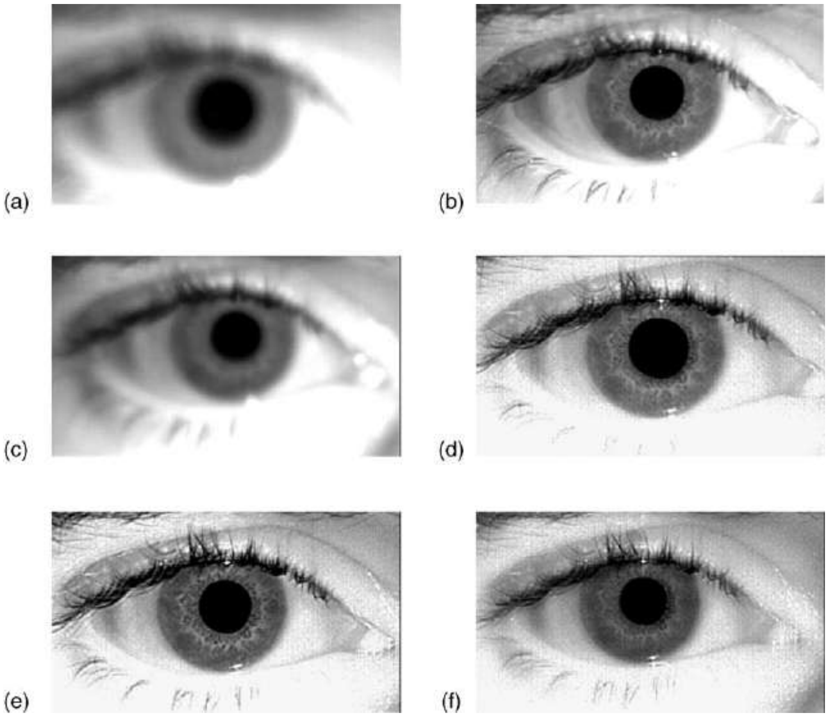

Figure 12 Images of an iris taken with infrared illumination. The top row shows images with a traditional CMOS camera. The lower row is taken with a camera using wavefront coding to extend the depth of field. The misfocus distance in the right and left columns is the same.

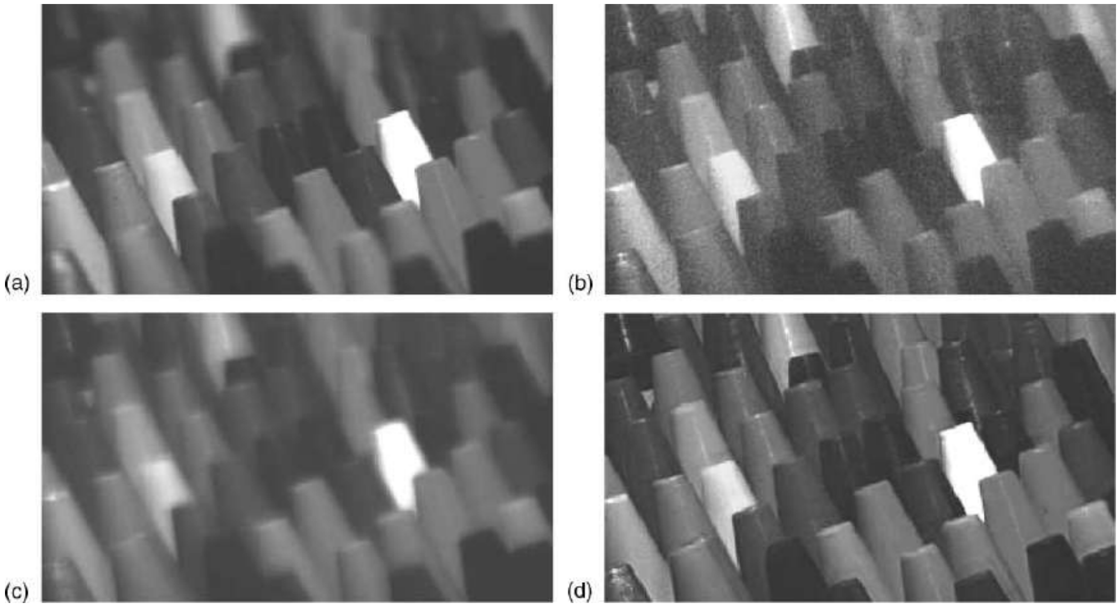

Figure 13 Images from a digital camera (a) Conventional camera with the lens wide open and focused on the white crayon. (b) Conventional camera with the lens stopped down to increase the depth of field, with the resulting increase in noise due to a lower light level. (c) Intermediate image of a camera that was modified with Wavefront Coding to have an extended depth of field. (d) Final image with the modified camera.

extended depth of field. Images of 2D bar codes imaged with and without Wavefront Coding are shown in Figure 11. Images in Figure 11a show the image of a 2D bar code formed by a conventional imaging system as the position of the object is varied. As with most imaging systems with fast optics and moderate magnification, the depth of field of the system is smaller than the application would prefer. Images in Figure 11b show the Wavefront Coded images of a 2D bar code after signal processing for the same object positions as (Figure 11a). These images are sharp and clear over a very large change in the distance to the bar code.

Biometric imaging systems. Biometric imaging applications, such as fingerprint or iris recognition, provide the means to eliminate passwords, secure E-commerce, and securely identify humans or animals. In traditional imaging technology, quite often the trade-off between light throughput and depth of field makes conventional biometric systems difficult to use or not sufficiently reliable. By using Wavefront Coding, the depth of information gathered can greatly increasing the ease of use and possibly reducing the cost of the systems.

Figures 12a,b, and c shows example images of an iris taken with near IR illumination and a traditional CMOS imaging system at three different planes of focus. At best focus (Figure 12b) the detail of the iris is sharp and clear. This iris detail is used by iris recognition algorithms to accept or reject the person. With slight movement towards (Figure 12a) or away (Figure 12c) from the imaging system the detail of the iris is lost. Figures 12d,e, and f show example iris images taken with the same near-IR illumination and a CMOS Wavefront Coded imaging system. At best focus (Figure 12e), the iris detail is as high as that from the traditional system. Movement towards (Figure 12d) or away (Figure 12f) from the imaging system results in images with essentially a constant level of iris detail. The size of the iris changes through magnification change. Increasing the range at which clear images can be made extends the useful image capture volume, enabling a more flexible system that is easier to use.

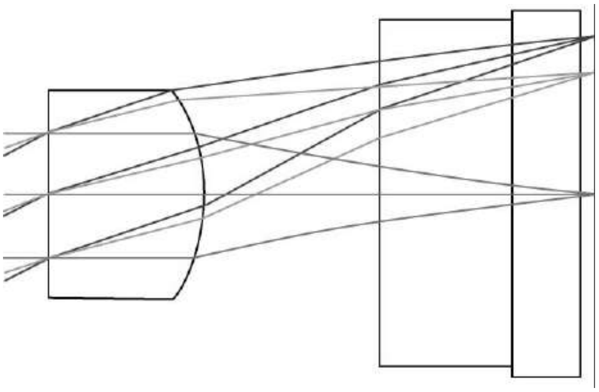

Figure 14 A schematic of a fast single-lens imaging system that has good performance over a wide field of view and the full visible range.

Color images with digital cameras. A conventional digital camera has a small depth of field as shown in Figure 13a, where the camera was focused on the white crayon. In a conventional imaging system, the only way to increase the depth of field is to stop down the lens. As seen in Figure 13b, this reduces the light that is available, and results in a very noisy image. The intermediate image of an imaging system that has an extended depth of field is shown in Figure 13c, where it is seen that the result of convolving the object distribution with the modified PSF is an intermediate image that appears to have the same blur over the entire image. After signal processing of the intermediate image to obtain the final image, a clear image is seen with good signal-to-noise ratio.

Inexpensive imaging systems with high-quality imaging characteristics. By exploiting the extension of the depth of focus that is made possible with Wavefront Coding techniques, one can design imaging systems that have fewer optical elements than traditional designs that do not use Wavefront Coding. In high-volume applications, reduction in the number of optical elements can lead to greatly reduced costs. Figure 14 shows a schematic of a fast, single-lens Wavefront Coded imaging system that is very compact and has the performance of a longer lens that traditionally requires two or more optical elements. In this example, the entire focus budget was used to compensate for focus-related aberrations, and the shape of the lens was used to compensate for distortion and coma. Even zoom systems can be greatly simplified with the new design techniques. Figure 15 shows a fast two-element Wavefront Coded zoom lens that images over a wide field with performance that is traditionally obtained only with three or more optical elements. Both of these imaging systems use non-rotationally symmetric forms of phase functions and are different from the ones of eqns [5] and [6].

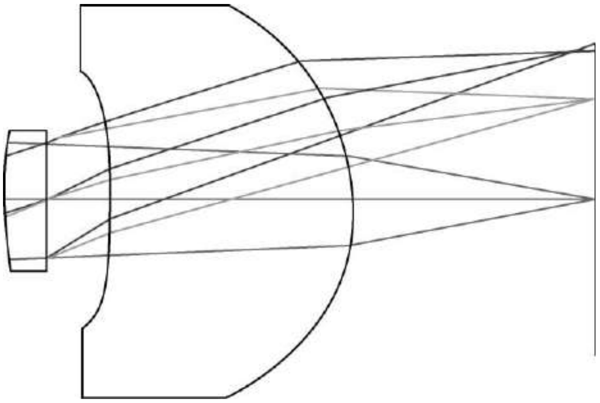

Figure 15 A schematic of a fast two-element zoom system with good performance over a 2.5 × zoom and a wide viewing angle.

Large depth of focus microscopes. A microscope can be modified to employ Wavefront Coding by either modifying the surface of one of the lenses in the objective, or by inserting a phase plate into the system at some appropriate position. One easy means of doing the modification is to place a Wavefront Coding phase plate in the slider position of a microscope, as shown in Figure 16. In this case, the optics is of high quality, and the entire focus budget was allocated to increasing the depth of field. A modified Zeiss microscope was used to obtain the images shown in Figure 17. An image of a diatom in Figure 17 shows the comparison with a conventional microscope in (a) and as with a Wavefront Coded microscope in (b). Notice that the greater depth of field gives more of an appearance of depth in the image.

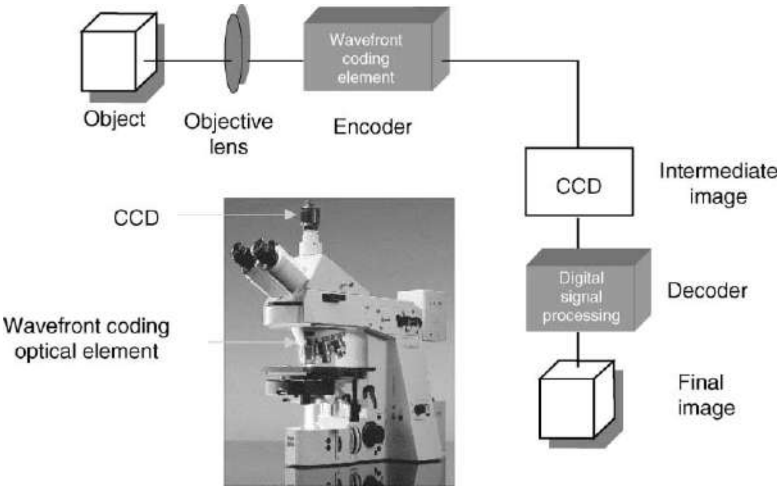

Figure 16 Schematic of a modified Zeiss microscope to employ Wavefront Coding to extend the depth of field.

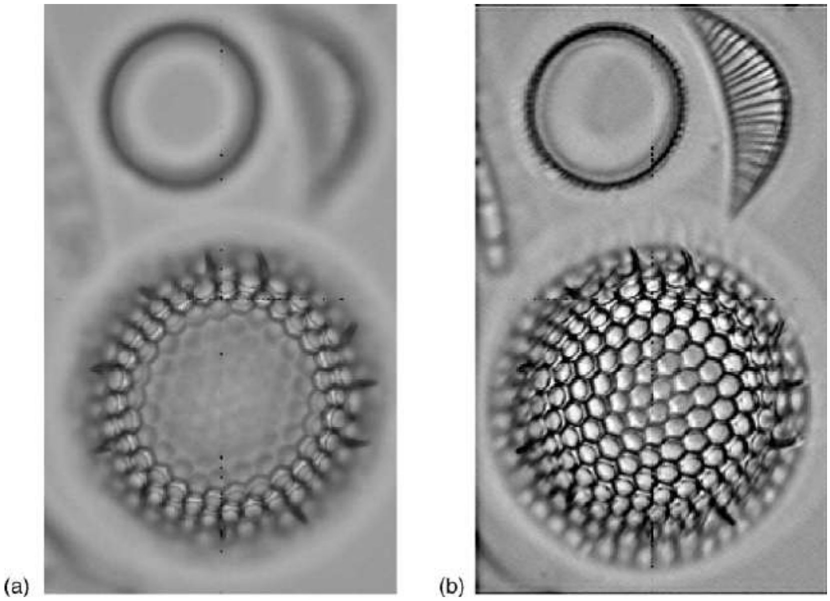

Figure 17 A 100 × image of diatoms made with (a) conventional and (b) modified Zeiss microscope to give an extended depth of field.

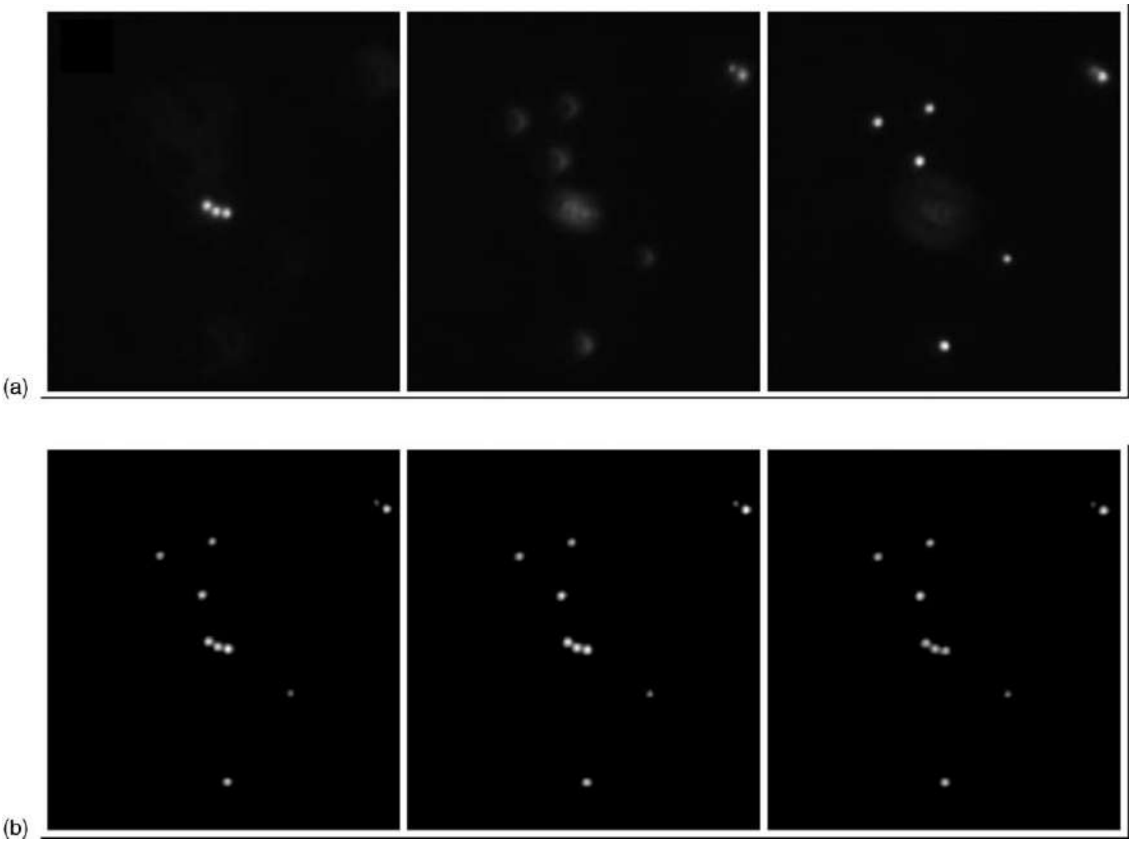

Figure 18 Imaging 0.5 micron diameter fluorescent beads at different sample distances. The images were taken with a 63×, NA = 1.4 imaging system in one micron steps. (a) Three image planes taken using a traditional system. They show that individual beads are sharp only in a single plane. The images in (b), which were taken with a Wavefront Coding system at the same image planes, show all beads are imaged clearly at each plane.

Other modes of microscopy are also possible. In fluorescence microscopy, ultraviolet light illuminates objects that have fluorescing compounds in them. The images are formed using the fluorescence that is excited by the ultraviolet illumination. Figure 18 shows images of fluorescing beads having a 0.5 micron diameter. The beads are at various depths within the sample. The traditional system, operating at 63 × with an immersion objective having a NA of 1.4, has a depth of field that is too small to image all of them. The images in Figure 18a were taken with the same microscope but with a wavefront coding phase element placed in the slider and signal processing of the captured images. All beads are clearly seen at all three of the focus positions.

Potential Future Applications

Wavefront Coding should be useful in many imaging system where signal processing is to be done anyway; in that case, optimizing the optics and the signal processing to the purpose of the imaging system. Examples where an extended depth of field may be desired are endoscopes, video cameras, and lithography. In designing an imaging system to take advantage of a large depth of focus to produce inexpensive digital cameras, high-quality images can be formed with fewer optical elements and much lighter cameras. In lithography, a larger process window is possible.

Because an increase in the depth of focus can permit wide spectral imaging, wide band infrared images can be formed. In systems that are used to provide images for machine interpretation, it is possible to design the optics and signal processing so that the desired process can be optimized globally, with the optics and the signal processing. For example, if the object is to classify cells as cancerous or noncancerous, it may be that the optics should be modified so that the information of primary importance is acquired and transmitted to the detector array. This modification would be done in conjunction with the optimization of the signal processing.

Conclusion

The mathematics and concepts that underlie a new approach to lens design were summarized. The technique has thus far been useful in enhancing the performance of some imaging systems such as microscopes, and in making simpler systems for inexpensive cameras. This new approach requires that some signal processing be done, but in digital cameras, some signal processing is being done already.

List of Units and Nomenclature

Modulation transfer function	The magnitude of the optical transfer function of an imaging system.
Optical transfer function	The response of an imaging system as a function of the spatial frequency of a sinusoidal object distribution.
Spatial frequency	A measure of spatial variation in the form of sinusoids in cycles per millimeter.

See also

Imaging: Wavefront Sensors and Control (Imaging Through Turbulence). **Microscopy:** Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Barrett HH and Swindell W (1981) *Radiological Imaging: The Theory of Image Formation, Detection, and Processing*. New York: Academic Press.

Brenner K, Lohmann A and Ojeda-Castañeda J (1983) The ambiguity function as a polar display of the OTF. *Optics Communications* 44: 323–326.

Cathey WT and Dowski ER (2002) New paradigm for imaging systems. *Applied Optics* 41: 6080–6092.

Cathey WT, Frieden BR, Rhodes WT and Rushforth CK (1984) Image gathering and processing for enhanced resolution. *Journal of the Optical Society of America A* 1: 241–249.

Cushman SC (2001) *Finite-Dimensional Matrix Treatment of Partially Coherent Optical Systems*. University of Colorado thesis, Boulder, Colorado.

Dowski ER and Cathey WT (1995) Extended depth of field through wavefront coding. *Applied Optics* 34: 1859–1866.

FitzGerrell AR (1993) *Applications of the Ambiguity Function to Optics*. Thesis, University of Colorado, Boulder, Colorado.

Häusler G (1972) A method to increase the depth of focus by two step image processing. *Optics Communications* 6: 38–42.

Ojeda-Castañeda J, Tepichin E and Diaz A (1989) Arbitrary high focal depth with quasioptimum real and positive transmittance apodizer. *Applied Optics* 28: 2666–2670.

Papoulis A (1974) Ambiguity function in Fourier optics. *Journal of the Optical Society of America* 64: 779–788.

Rihaczek AW (1969) *Principles of High Resolution Radar*. New York: McGraw-Hill.

Van der Gracht J, Dowski ER, Taylor MG and Deaver DM (1996) Broadband behavior of an optical-digital focus-invariant system. *Optics Letters* 21: 919–921.

Wach HB, Dowski ER and Cathey WT (1998) Control of chromatic focal shift through wave-front coding. *Applied Optics* 37: 5359–5367.

Woodward PM (1953) *Probability and Information Theory with Applications to Radar*. New York: Pergamon Press.

PHOTON PICTURE OF LIGHT

S J Bentley, Adelphi University, Garden City, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The photon nature of light is the central topic in the field of quantum optics, which has developed into a leading area of research. The possible technological impacts of this field to many areas, such as imaging, computing, and lithography, have greatly expanded interest into the nature of the photon. A photon description of light goes well beyond the particle nature of light, to include any optical phenomena that cannot be adequately described by classical physical optics. This article will introduce the nature of the photon through a discussion of many such phenomena, covering topics from the beginning of quantum mechanics, including black-body radiation and Compton scattering, to areas of modern interest, such as squeezed light and entangled photons.

The Origins of Quantum Optics

One of the earliest theories in the development of quantum physics was the description of black-body radiation by Max Planck (1858–1947). Using the model of a radiating body as a collection of oscillators, he postulated that the amount of energy emitted or absorbed was proportional to its frequency, ν, and that these oscillators could only have energies corresponding to integer multiples of such a packet of energy, where one unit of energy was given by

$$E = h\nu \qquad [1]$$

where $h = 6.626 \times 10^{-34}$ Js is the universal constant, now known as Planck's constant. He then assumed a Maxwell–Boltzmann distribution to describe the probability of such oscillator states being occupied:

$$P(n) = P_0 e^{-nh\nu/kT} \qquad [2]$$

where $k = 1.381 \times 10^{-23}$ JK^{-1} is Boltzmann's constant and T is the temperature of the blackbody. The energy density in the radiation field, at a given frequency, is given by the oscillator density (at that frequency) times the average energy of the oscillators. Previous models had not only failed to quantize the oscillator energies, but in doing so also assumed equal occupation probabilities that led to a prediction of infinite energy densities at short wavelengths, the so-called ultraviolet catastrophe. The oscillator density is given by

$$n(\nu) = 8\pi\nu^2/c^3 \qquad [3]$$

where $c = 3 \times 10^8$ m s^{-1} is the speed of light. This value can be found by a mode density calculation within a box, as is often done in an introductory electromagnetics text. The average energy of the oscillators is found by calculating the sum of the occupation probability times the state energy and dividing by the sum of the occupation probabilities:

$$E_{\text{av}} = \sum_{n=0}^{\infty} [P(n)h\nu n] \Big/ \sum_{n=0}^{\infty} P(n) \qquad [4]$$

Evaluating the above expression for E_{av} and multiplying by $n(\nu)$, one finds the Planck radiation law:

$$I(\nu) = \frac{8h\pi\nu^3}{c^3(e^{h\nu/kT} - 1)} \qquad [5]$$

The black-body curve described by Planck's law is shown in Figure 1. By introducing quantization to solve a problem that classical physics had failed to explain, Planck paved the way for quantum physics.

Albert Einstein (1879–1955) made use of Planck's quantization for two important results that will be described here. The first is the photoelectric effect, a process by which electrons are released from a metal surface when exposed to light. The process has several traits that could not be explained by

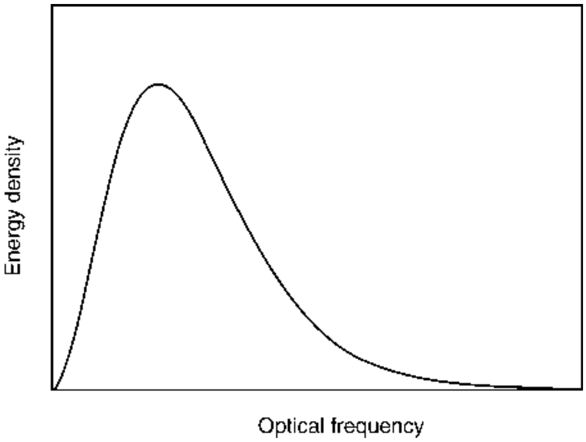

Figure 1 The spectral density emitted by a blackbody at a given temperature as given by Planck's law.

classical physics. A given metal will only release electrons if the incident light is above a certain frequency, regardless of the intensity. However, once sufficient frequency components are present, electrons are immediately released, with their number dependent on the intensity of the light and independent of its frequency. Contrarily, the maximum potential energy of the released electrons depends only on the maximum frequency of light present, independent of intensity. Einstein was able to satisfy all of these conditions by saying that the light energy comes in quantized amounts as described by Planck, and that the electrons in the metal require a minimum energy, or work function W, to be ejected. Then for each quantum of light, which we now call a photon, with energy greater than W that strikes an electron, the electron will be released with kinetic energy given by

$$K = h\nu - W \qquad [6]$$

It is historically important to note that, although Planck introduced the idea of a quantized release of radiation from oscillators, he still believed the light field itself to be continuous and classical. Planck's approach of quantizing energy levels of matter while treating light classically has since become known as the semiclassical method, and is adequate for solving many problems. Thus, it was also in quantizing light itself that Einstein took the next step, and thus can be viewed in many respects as the father of the modern photon, although the term photon was not used until two decades later.

The second area in which Einstein used the idea of energy quantization is in predicting radiative transition probabilities. A radiative transition is a process by which a photon is either absorbed or emitted to satisfy conservation of energy when an atom moves to a higher or lower energy state. Einstein assumed three such processes existed, as shown in Figure 2: absorption of a photon raising the atom to a higher energy level; spontaneous emission of a photon, where the atom emits a photon in going to a lower energy level; and stimulated emission, a process in which a photon interacts with the atom and causes it to emit a second photon (in phase with the first) and thereby go to a lower energy level. For each of these processes he defined a corresponding coefficient, now known as the Einstein A and B coefficients. They are labeled B_{12} for absorption, A_{21} for spontaneous emission, and B_{21} for stimulated emission. Note that there is no A_{12} as absorption by its very nature is a stimulated process. Einstein further stated that the rate of spontaneous emission was only dependent on his material dependent coefficient and the number of atoms in the upper energy state, with the rate given by

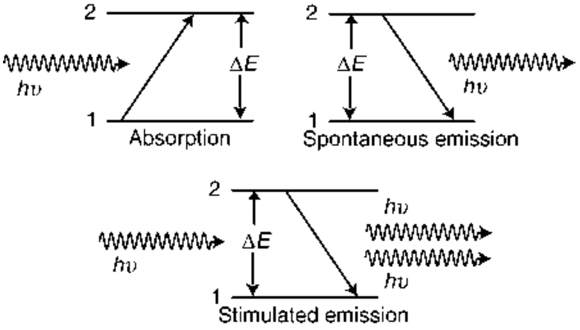

Figure 2 Energy level diagrams representing the processes of absorption, spontaneous emission, and stimulated emission.

$$R_{\mathrm{sp}} = A_{21}N_2$$

However, he said that the rates of the stimulated processes were also dependent upon the strength of the optical field seen by the atoms, $I(\nu)$, and were given by

$$R_{\mathrm{ab}} = I(\nu)B_{12}N_1 \qquad [7]$$

for absorption and

$$R_{\mathrm{st}} = I(\nu)B_{21}N_2 \qquad [8]$$

for stimulated emission. The full absorption rate is given by R_{ab} as defined above, whereas the full emission rate is the sum of the spontaneous and stimulated rates, given by

$$R_{\mathrm{em}} = [A_{21} + B_{21}I(\nu)]N_2 \qquad [9]$$

Einstein was primarily interested in a system in equilibrium, for which $R_{\mathrm{ab}} = R_{\mathrm{em}}$. If one then looks at the limit where the optical field is very large such that spontaneous emission is negligibly small and the population is equally divided between the upper and lower state, the relationship $B_{12} = B_{21}$ is found. Applying the same Maxwell–Boltzmann occupation statistics that Planck used to find his black-body expression, one can find that in thermal equilibrium the ratio of probability of finding an atom in state 1 to finding an atom in state 2 is

$$N_1/N_2 = e^{(E_2-E_1)/kT} = e^{h\nu/kT} \qquad [10]$$

Note that for a two-level system in thermal equilibrium, the population in the upper level can at most equal, never exceed, the population in the lower level, regardless of the strength of the applied field. This is important in showing that the B coefficients are equal, and will also be a crucial point when discussing the theory of the laser later in this article.

Appropriately combining the above expressions leads to

$$I(\nu) = A_{21}/[B_{21}(e^{h\nu/kT} - 1)] \qquad [11]$$

Equating this with Planck's radiation law yields the relation:

$$A_{21}/B_{21} = 8\pi h\nu^3/c^3 \qquad [12]$$

Thus, to fully specify the transition probabilities for a given system, one needs only to find A or B. It is rather straightforward using only atomic quantization to calculate B (such a derivation can be found in many introductory quantum mechanics textbooks). Direct calculation of A is more involved, requiring quantization of both the optical field as well as the atomic system.

Arthur Compton (1892–1962) used the idea of quantized radiation to explain the presence of longer wavelength components present when X-rays are scattered from matter, an effect which now bears his name. The amount of scattering is independent of the scattering material and the wavelength of the source, depending only on the angle of scattering, θ. If the scattering is treated as a collision between a particle-like photon and a massive particle (taken to be initially at rest) as shown in Figure 3, the result can then be found simply by applying conservation of energy and momentum, where the photon energy is as given above and the photon momentum is given by

$$p = E/c = h/\lambda \qquad [13]$$

Recalling that the momentum is a vector quantity, the conservation laws provide a set of three coupled equations. Solving them directly one finds that

$$\Delta\lambda = (h/c)(1 - \cos\theta)(2K)/[p^2 - (K/c)^2] \qquad [14]$$

where K and p correspond to the final kinetic energy and momentum of the massive particle. It is important to allow for the possibility of very high velocities for the massive particle, thus implying that relativistic expressions for K and p are needed. In terms of the kinetic energy, the relativistic expression for momentum is

$$p = \sqrt{2Km + (K/c)^2} \qquad [15]$$

where m is the mass of the scattering particle. Using this relation, the above expression can be simplified, giving the wavelength shift to be

$$\Delta\lambda = (1 - \cos\theta)h/mc \qquad [16]$$

The scaling factor is known as the Compton wavelength:

$$\lambda_c = h/mc \qquad [17]$$

For the case of an electron, the Compton wavelength has the value 0.02426 angstroms. For a massive particle, the Compton wavelength is often used when wanting to discuss and analyze wavelike properties. As a final note on this process, if the calculation would have been done nonrelativistically, there would have been an extra factor of $[1 - 0.25(v/c)^2]^{-1}$, where v is the velocity of the particle. The appearance of the factor (v/c) in any result implies the need for a relativistic treatment, though often (as in this case) taking the limit of $v/c = 0$ yields the proper result.

Another area in which the photon nature of light is revealed is a form of inelastic scattering of light now known as the spontaneous Raman effect. Discovered by CV Raman (1888–1970), this is a process in which a material illuminated by a light source scatters light into frequency components not found in the source. The frequency shift is due to the energy difference, ΔE, between two energy states in the material. There are two possible cases, each of which is represented in Figure 4. In the first, known as Stokes scattering, an atom in the lower state absorbs a source photon, raising it to the upper state via a virtual level, during which it emits a photon at the Stokes frequency, ν_s, as represented by the relation

$$h\nu_i = h\nu_s + \Delta E \qquad [18]$$

where ν_i is the frequency of the incident photon from the source. The second case, anti-Stokes scattering, has an atom in the upper state absorb the source photon and simultaneously fall to the lower state, again through the virtual level during which it emits a photon at the higher anti-Stokes frequency, ν_a,

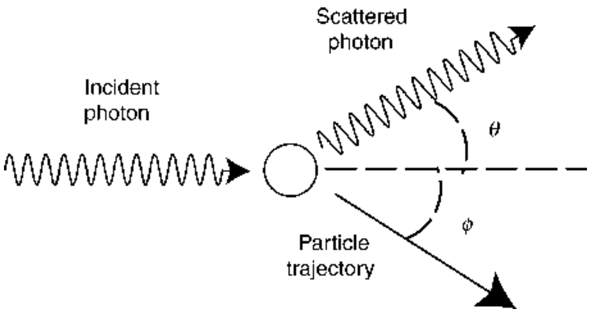

Figure 3 Schematic representation of Compton scattering.

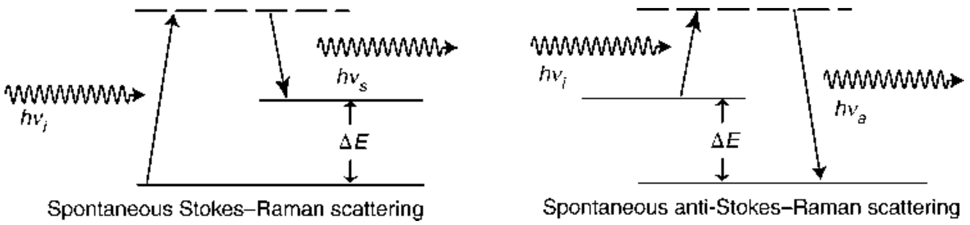
$h\nu_i$
$h\nu_s$
ΔE
Spontaneous Stokes–Raman scattering
$h\nu_i$
ΔE
$h\nu_a$
Spontaneous anti-Stokes–Raman scattering

Figure 4 Energy level diagrams for Stokes and anti-Stokes Raman scattering.

given by

$$h\nu_i = h\nu_a - \Delta E \qquad [19]$$

Since the anti-Stokes scattering requires the atom to be initially in the upper state, the probability of such scattering is generally much less than that for Stokes scattering. This can be quantified by the same occupation statistics used above, giving the ratio of probabilities as

$$P_{as}/P_s = e^{-h\Delta E/kT} \qquad [20]$$

It is important to note that the energy difference can either be based on the electronic structure of an atomic system as has been discussed thus far, or other energies such as the vibrational energies in a molecular system. The ability to probe a variety of energy structures without having to tune the source to a particular wavelength has led to wide use of Raman scattering for spectroscopy. A related effect, known as stimulated Raman scattering, will be discussed later.

The Photon in Modern Physics

As an indication of the widespread importance of the results discussed thus far, they led to four Nobel prizes in physics: Planck in 1918; Einstein in 1921; Compton in 1927; and Raman in 1930. Our initial understanding of the photon comes from these results, but it was with the development of modern quantum mechanics by Erwin Schrödinger (1887–1961), Werner Heisenberg (1901–1976), and Paul Dirac (1902–1984); quantum electodynamics by Richard Feynman (1918–1988) and others; and the invention of the laser by Charles Townes (1915–) and others, that the field of quantum optics and our modern view of the photon began to fully take shape. In this section, the basics of modern quantum optics will first be introduced. A brief discussion of the theory of the laser as related to photon transitions will then be given, followed finally by examples of important areas of experimental quantum optics that have been made accessible by the laser.

It is now common to treat the photon with a very similar formalism as that originally created for massive particles in the development of quantum mechanics. That is, one describes the photon by a wave function, $|\Psi\rangle$, which contains all the information about the photon and must satisfy the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}|\Psi\rangle = H|\Psi\rangle \qquad [21]$$

where $\hbar = h/2\pi = 1.054\times10^{-34}$ Js and H is the Hamiltonian for the photon. Just as in classical electromagnetics, the Hamiltonian is related to an electric and magnetic field. Clearly, for this treatment to be in agreement with the well-established classical electromagnetic theory, these fields must satisfy Maxwell's equations. There are several possible representations of the Hamiltonian, but for the purposes of this discussion a convenient choice is

$$H = h\nu(\hat{n} + 0.5) \qquad [22]$$

where $\hat{n}$ is called the number operator as it returns the average number of photons in the state, $\langle\hat{n}\rangle$, represented by the operation

$$\langle\hat{n}\rangle \equiv \langle\Psi|\hat{n}|\Psi\rangle \qquad [23]$$

Explanations of the notation used can be found in any elementary quantum mechanics text. Note that in defining the Hamiltonian, a single frequency ν was used. This implies a single mode field. For a more general case, one would simply sum over the modes corresponding to various frequencies.

In addition to obeying the Schrödinger equation, the photon is also restricted by the Heisenberg uncertainties as for massive matter, given in a general form as

$$\Delta q \Delta p \geq \hbar \qquad [24]$$

where q and p are any set of canonically conjugate variables, such as position and momentum or energy and time. Note that, depending on the exact definitions used for q and p, there can be extra factors such as 0.5, and sometimes normalized variables are used such that the inequality is with respect to 1. However, it is interesting to note that, unlike for massive particles that were considered well localized in classical theory, the idea of uncertainty for light had been accepted in some form long before the development of quantum mechanics. Neglecting the scaling factor $\hbar$, which is purely quantum in origin, an uncertainty principle can be derived for any wavelike phenomena simply using the relationship between Fourier transform pairs.

Using the basic ideas of a photon state and the uncertainty principle, some important properties of light will now be explored. There are several possible forms the wave function can take. It is often convenient to write the wave function in terms of states containing a certain number of photons. Such states are known as Fock or number states and are represented as $|n\rangle$, with n being the number of photons in the state and are said to be eigenstates of the number operator. The energy of a Fock state is given by

$$E_n = h\nu(n + 0.5) \qquad [25]$$

Notice that this predicts a background, or vacuum, energy for the electromagnetic field, even when the average number of photons in the field is zero. This vacuum energy is given by 0.5 $h\nu$. Correspondingly, due to the uncertainty principle, there must be some uncertainty in this energy as well as with any energy. The fluctuations in the vacuum energy are known as vacuum fluctuations. Many processes, such as spontaneous emission that cannot be adequately described by a simple semiclassical approach, where matter is treated quantum mechanically and light is treated classically, can be treated by the addition of vacuum fluctuations, without requiring a full quantum mechanical treatment of the electromagnetic field. Such an approach is used in many texts, but there are several important physical phenomena that do require the fully quantized approach discussed here.

The state most resembling a classical field is known as the coherent state, represented here as a sum of Fock states:

$$|\Psi\rangle = |\alpha\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle \qquad [26]$$

where α is a parameter describing the field, with the average number of photons in the field given by

$$\langle \hat{n} \rangle = |\alpha|^2 \qquad [27]$$

One key property of coherent states is that they exhibit minimal uncertainty. That is, the uncertainty relations become equalities. Also, the uncertainty is equally divided between quadrature components.

Before applying the above principles to explore some of the dramatic results of quantum optics, a brief discussion of the laser will be given due to its importance in the field. The theory of operation of the laser is based directly on Einstein's transition probabilities. For the transition of interest, one wants to generate a large number of coherent photons (that is, photons in phase with one another). Recall that in the process of stimulated emission, one photon leads to two coherent photons, whereas spontaneous emission leads to photons which are not phase referenced to anything else in the system. Also, if a significant number of the photons are to escape the region where they are generated, it is important to minimize the absorption. Thus, a laser requires that the stimulated emission rate exceed both the spontaneous emission rate and the absorption rate. The first condition places a lower limit on the optical field density present at the lasing frequency. This density can be reasonably large, but presents no fundamental difficulties. In practice, mirrors are placed around the emission region to confine the photons and counteract losses, thereby building up the energy density until the above condition is satisfied. The second condition, however, can be written as

$$N_2 > N_1 \qquad [28]$$

and is the more restrictive requirement. Such a condition is referred to as inversion and, as was previously mentioned, cannot be sustained in a two-level system once equilibrium is reached. Therefore, to be able to maintain lasing, additional levels are needed. Two typical laser energy diagrams are shown in Figure 5. In such multilevel systems, it is possible to create an inversion in the levels of interest and thus allow lasing. Further details on lasers (both in theory and practice) are widely available, though it should be noted that the output of a laser is a good approximation of a coherent state described above.

The invention of the laser revolutionized optics, as well as the whole of modern physics, by making readily available electromagnetic fields of a vastly different nature than that from any previous source of light. This quickly led to the birth of a new field known as nonlinear optics. Nonlinear optics describes interactions between light and matter which have properties dependent on the field strength. Although some nonlinear processes can be described adequately without the use of a photon picture, most nonlinear process can be best described as some combination of annihilation (absorption) and creation (emission) of photons. Two such

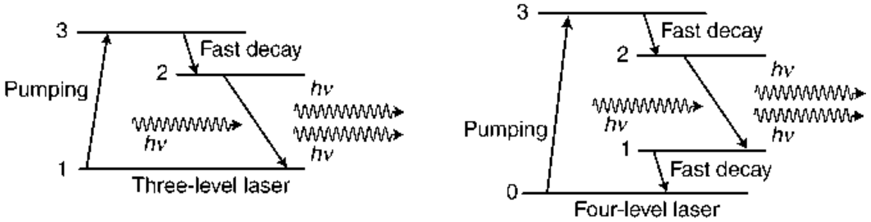

Figure 5 Energy level diagrams for three-level and four-level lasers.

processes, that are of great interest in quantum optics, are stimulated scattering and parametric down conversion.

Earlier, the process of spontaneous Raman scattering was discussed. For such a process, it is assumed that the incident field driving the process is relatively weak such that the scattering process generates very low photon densities in the various modes of the electromagnetic field. However, just as there are both stimulated and spontaneous emission processes in the radiative transitions discussed previously, in the presence of strong fields there can also be stimulated scattering. The total Raman scattering rate can be expressed as

$$R = \eta I_i(I_R + 1) \qquad [29]$$

where η is a scaling factor, including the gain of the process for a given material, I_i is the strength of the incident optical field (generally a laser field to be sufficiently strong to cause stimulated scattering), and I_R is the Raman field strength. Note that for I_R, the rate is just linear with the applied field, representing the spontaneous process. However, with a strong pump, the generated field will quickly become large enough such that the stimulated process will dominate. The spontaneous scattering is not directional dependent, but since the stimulated process grows along the length of the scattering material in the path of the incident beam, it will tend to be generated along the path of the incident laser. Due to linear and nonlinear dispersion effects, however, the light will form in a narrow cone about the pumping laser rather than exactly collinear. Unlike the spontaneous scattering, a significant percentage of the incident photons can be converted to the Raman frequencies, making stimulated Raman scattering an extremely powerful spectroscopic tool.

Possibly the most important nonlinear optical process for quantum optics is parametric down conversion. This process is interesting both in the nature of the process itself, as well in the properties of the generated light. In parametric down conversion, shown schematically in Figure 6, one photon (known as a pump photon) of frequency ν_p is annihilated and two new photons (generally referred to as signal and idler) are created with frequencies ν_s and ν_i respectively, given by the conservation of energy relationship:

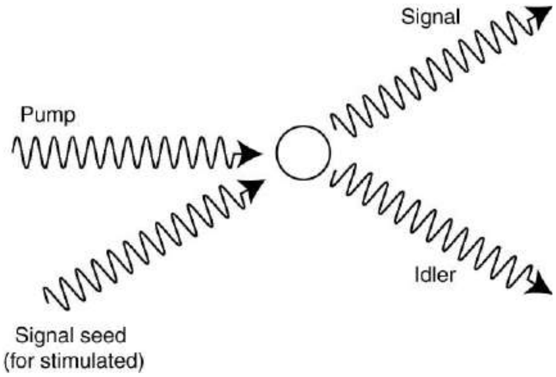

Figure 6 Schematic representation of parametric down conversion.

$$h\nu_p = h\nu_s + h\nu_i \qquad [30]$$

In the case where $\nu_s = \nu_i$, the process is referred to as degenerate. Down conversion can either be spontaneous, in which only pump photons are externally applied, or stimulated, in which both pump photons and signal photons are applied. In the stimulated case, the signal is amplified and idler photons are generated as a byproduct. Parametric down conversion is one of the leading sources of nonclassical states of light. The term nonclassical is used rather than quantum, because all states of light can be accurately described by a quantum picture, but while most can also be described very well with a classical picture the states referred to as nonclassical cannot. Two of the most common nonclassical states of light are entangled states and squeezed states.

Entangled states are multiphoton states in which the state cannot be written as a product of the states of the individual photons. For simplicity, only two-photon entangled states (commonly referred to as EPR states for Einstein, Podolsky, and Rosen) will be discussed. More extensive and general material, including discussions of higher-order entanglement,

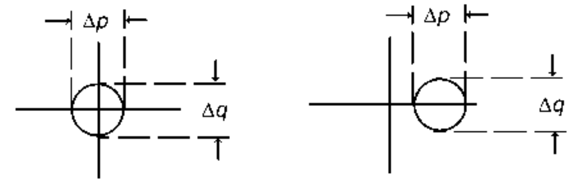

Unsqueezed vacuum state

Unsqueezed coherent state

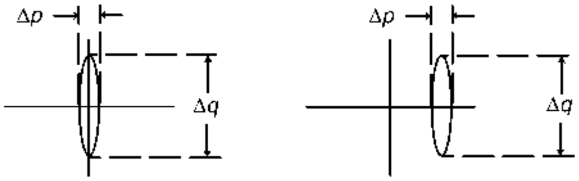

Squeezed vacuum state

Squeezed coherent state

Figure 7 Uncertainty diagrams for squeezed and unsqueezed vacuum and coherent states.

such as the three-photon entangled states (commonly referred to as GHZ states for Greenberger, Horne, and Zeilinger), can be found in a variety of textbooks on modern quantum optics. One of the simplest ways to generate a two-photon entangled state is by the process of spontaneous parametric down conversion described above. If the product of the pumping intensity, generation length, and generation gain is of an appropriate level, individual pairs of photons are emitted. The states of these photon pairs are entangled, with the position, momentum, wavelength, polarization, and all other properties of the photons completely correlated and inseparable. Such states have been a powerful tool for the experimental verification of many of the predictions of quantum mechanics and quantum optics. Entangled states can also act as a source for many of the newer technologies in quantum optics, including cryptography and teleportation.

Another important form of nonclassical light is known as squeezed light. The principle behind squeezed light is that, unlike for coherent states where the uncertainty is equally distributed, the uncertainty in one quadrature can be reduced at the expense of increased uncertainty in the other quadrature. This is shown schematically in Figure 7. As opposed to entangled states which tend to be very weak, containing only a few photons, squeezed states of light can be quite intense. This makes squeezed light more desirable for many applications.

In conclusion, it should be noted that while the photon description of light can explain a wide variety of physical processes and is now leading to many exciting technological developments, much is still unknown about the photon at the fundamental level. Any accurate description of the photon must be in agreement with experimental evidence, including that discussed in this article, but one should be aware that the accepted picture of the photon nature of light is still constantly evolving and growing.

See also

Nonclassical Light. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Quantum Optics:** Entanglement and Quantum Information. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Boyd RW (2002) *Nonlinear Optics*, 2nd edn. San Diego, CA: Academic Press.

Cohen-Tannoudji C, Diu B and Laloë F (1977) *Quantum Mechanics*. New York: Wiley-Interscience Publications.

Hecht E (2001) *Optics*, 4th edn. Reading, MA: Addison-Wesley Publishing.

Heisenberg W (1949) *The Physical Principles of the Quantum Theory*. New York: Dover Publications.

Lipson SG, Lipson H and Tannhauser DS (1995) *Optical Physics*, 3rd edn. Cambridge, UK: Cambridge University Press.

Loudon R (2000) *The Quantum Theory of Light*, 3rd edn. Oxford, UK: Oxford University Press.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Milonni PW and Eberly JH (1988) *Lasers*. New York: Wiley-Interscience Publications.

Roychoudhuri C and Roy R (eds) (2003) *OPN Trends – The Nature of Light: What is a Photon?* Washington, DC: Optical Society of America.

Schwinger J (ed.) (1958) *Selected Papers on Quantum Electrodynamics*. New York: Dover Publications.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: John Wiley & Sons.

PHOTONIC CRYSTALS

Contents

Atomic Physics
Electromagnetic Theory
Microwave Photonic Crystals
Nonlinear Optics in Photonic Crystal Fibers
Photonic Crystal Lasers, Cavities and Waveguides
Self-Assembled and Functionalized Photonic Crystals

Atomic Physics

G Kurizki and A G Kofman, Weizmann Institute of Science, Rehovot, Israel

D Petrosyan, Institute of Electronic Structure and Laser, Heraklion, Greece

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fabrication and study of dielectric structures whose refractive index is periodically modulated on a micron or submicron scale, known as photonic crystals (PCs), are attracting considerable interest at present. One-, two-, or three-dimensional (1D-, 2D-, or 3D-) periodic PCs exhibit photonic band gaps (PBGs), analogously to electronic band gaps in ordinary crystals, and can incorporate defects, designed to form localized narrow-linewidth (high-Q) modes at PBG frequencies. The advent of such structures opens up new perspectives in atomic physics and quantum optics, since they are expected to allow an unprecedented control over the spectral density of modes (DOM) and the spatial modulation of narrow-linewidth (high-Q) modes, in the microwave, infrared, and optical domains. An interesting situation arises when foreign atoms or ions – dopants – with transition frequencies within the PBG, are implanted in the PC. Then light near one of these frequencies resonantly interacts with the dopants and is concurrently affected by the PBG dispersion. Consequently, highly nonlinear processes with a rich variety of unusual PC-related features are anticipated. Such nonlinear optical processes, involving near-resonant transitions in PCs, undergo basic modifications as compared to the corresponding processes in free space, which are attributed to the strong suppression of the DOM within PBGs, to sharp bandedges and to intra-gap narrow lines associated with high-Q defect modes.

Results are detailed below for spontaneous emission of radiation in PCs and for the resonant interaction of atoms with the field of a single high-Q defect mode. These results stem from the failure of perturbation theory and the onset of strong field-atom coupling near sharp bandedges or narrow defect-mode lines in 2D and 3D PCs. 3D-PBGs are needed, in order to extinguish spontaneous emission in all possible directions of propagation and dipole orientations. If only one polarized atomic transition is involved in the spontaneous emission, 2D-PCs suffice for its suppression. For controlling strictly unidirectional field propagation, it is sufficient to resort to PBGs in 1D-periodic structures (Bragg reflectors or dielectric multi-layer mirrors).

As a further example of field–atom interactions in doped PCs, we will show that two ultraweak electromagnetic fields (or photons), coupled to appropriate transitions of the dopants in a PC, can mutually induce large phase shifts or drastic changes in absorption. Appreciable photon–photon correlations can then be established, which can be utilized in both classical and quantum optical communications.

Radiative Decay and Photon–Atom Binding in a PC

The interaction of a two-level atom with an arbitrary field-mode continuum can be described by the following second-quantized Hamiltonian in the rotating-wave approximation (RWA)

$$H = \hbar\omega_{\mathrm{a}}|e\rangle\langle e| + \hbar\sum_{\vec{k}} \omega_{\vec{k}} a_{\vec{k}}^{\dagger} a_{\vec{k}} + \hbar\sum_{\vec{k}} [\kappa(\omega_{\vec{k}}) a_{\vec{k}}^{\dagger} |g\rangle\langle e| + \text{hermitian constant}] \qquad [1]$$

Here $|e\rangle$ and $|g\rangle$ are the excited and ground atomic states, respectively, ω_{a} is the atomic transition frequency, $a_{\vec{k}}^{\dagger}$ and $a_{\vec{k}}$ are the creation and annihilation operators of a field mode with wavevector $\vec{k}$, and frequency $\omega_{\vec{k}}$, and $\hbar\kappa(\omega_{\vec{k}})$ is the resonant coupling energy of this mode with the atomic dipole. The $\vec{k}$-summation can be replaced by an integral over the continuum DOM in the frequency domain, $\sum_{\vec{k}} \to \int_0^{\infty} \mathrm{d}\omega\rho(\omega)$, avoiding, for the sake of simplicity, the study of direction-(angle-)dependent DOM effects.

We consider a two-level atom that is excited at time $t = 0$ in an empty, periodic dielectric structure. The wavefunction of the combined system field + atom can be cast in the general form:

$$|\Psi(t)\rangle = \alpha(t)|e, \{0_{\omega}\}\rangle + \int \beta_{\omega}(t)|g, 1_{\omega}\rangle\rho(\omega)\mathrm{d}\omega \qquad [2]$$

where $\{0_{\omega}\}$ signifies the completeness of field modes in the vacuum state and $|1_{\omega}\rangle$ denotes single-photon occupation of the ω-mode. The corresponding Schrödinger equation can be solved with the initial condition $|\Psi(0)\rangle = |e, \{0_{\omega}\}\rangle$ by means of the continuum spectral response:

$$G(\omega) = |\kappa(\omega)|^2\rho(\omega) \qquad [3]$$

The analysis of eqn [2], with $G(\omega)$ appropriate for photonic bandstructures, is aimed at revealing the prominent features of the atomic excitation decay $\alpha(t)$ and the corresponding emission spectrum.

In what follows, we consider a photonic bandstructure with several PBGs, separated by allowed bands. Each PBG is labeled by index i and has lower and upper cutoff frequencies $\omega_{\mathrm{L}i}$ and $\omega_{\mathrm{U}i}$, respectively. Then $G(\omega) = 0$ for $\omega_{\mathrm{L}i} < \omega < \omega_{\mathrm{U}i}$. Naively, one might expect that an excited atom would either be stable against single-photon decay, if ω_{a} is within a PBG, or decay completely at $t \to \infty$, if ω_{a} is anywhere in an allowed band. However, both statements turn out to be inaccurate.

Incomplete decay of $\alpha(t \to \infty)$ occurs if there is a stable eigenvalue (energy level) $\hbar\omega_i$ of the total (field-atom) Hamiltonian eqn [1]. Such an energy level is possible only if $G(\omega_i) = 0$, i.e., for ω_i in a PBG. It must satisfy:

$$\omega_i = \omega_{\mathrm{a}} + \Delta(\omega_i) \qquad [4]$$

$$\Delta(\omega_i) = \int_0^{\omega_{\mathrm{L}i}} \frac{G(\omega')}{\omega_i - \omega'}\mathrm{d}\omega' + \int_{\omega_{\mathrm{L}i}}^{\infty} \frac{G(\omega')}{\omega_i - \omega'}\mathrm{d}\omega' \qquad [5]$$

Here the integrals are the frequency shifts of the atomic resonance $\hbar\omega_{\mathrm{a}}$ induced by the spectral parts of the reservoir situated, respectively, below and above the ith PBG (Figure 1). If eqn [4] holds, then the corresponding term in $\alpha(t)$ is proportional to $\exp(-i\omega_i t)$, and does not decay. Physically, such a stable level can be interpreted as representing the binding of the photon to the atom (photon-dressed atom), without the ability to leave the atomic vicinity, due to Bragg reflection in the PBG.

Assuming that ω_i occurs in the ith PBG, we observe that the first shift is positive whereas the second shift is negative, i.e., each part of the continuum repels the atomic level from its PBG edge. Equation [4] has a solution if ω_{a} falls between the minimum and maximum values assumed by its left-hand side in the ith PBG, i.e., if

$$\omega_{\mathrm{L}i} - \Delta(\omega_{\mathrm{L}i}) < \omega_{\mathrm{a}} < \omega_{\mathrm{U}i} - \Delta(\omega_{\mathrm{U}i}) \qquad [6]$$

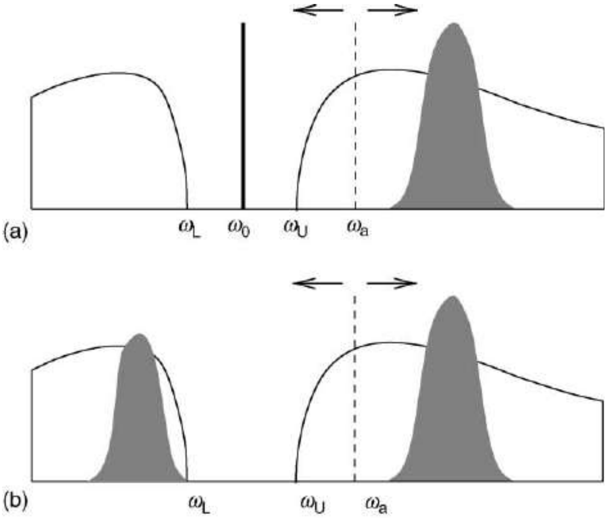

Figure 1 Frequency shifts of the atomic resonance ω_{a} (dashed line) due to 'repulsion' (arrows) by asymmetric spectral parts of the continuum below and above the PBG: (a) Incomplete decay for ω_{a} in allowed band. The interaction with the continuum splits the state into a superposition of a stable part in the PBG (solid line) and a decaying part above ω_{a} (shaded peak). (b) Complete decay for same ω_{a}, due to larger shifts that split the state into decaying parts in two allowed bands.

Incomplete decay occurs if this condition holds at least for one PBG. There may be, at most, one stable state in a single PBG (Figure 1). However, a two-level atom in a periodic structure can have several stable dressed states simultaneously, when conditions hold for several PBGs. It is quite extraordinary that the conditions (eqn [6]) for incomplete decay can be fulfilled with ω_a, not only in a PBG, but even in an allowed band, e.g., when $\Delta(\omega_{Ui})$ is negative or when $\Delta(\omega_{Li})$ is positive (Figure 1a,b).

Equally counterintuitive is the converse possibility, of complete decay for ω_a within a PBG. When there is a single PBG and one of the inequalities in eqn [6] is violated, complete decay will occur for $\omega_{Li} < \omega_a < \omega_{Li} - \Delta(\omega_{Li})$, if $\Delta(\omega_{Li}) < 0$, or $\omega_{Ui} - \Delta(\omega_{Ui}) < \omega_a < \omega_{Ui}$, if $\Delta(\omega_{Ui}) > 0$.

The possibility that one or several discrete, stable, states exist yields the corresponding wavefunction in the form:

$$|\Psi(t)\rangle = \sum_i \sqrt{c_i}|\psi_i\rangle e^{-i\omega_i t} + |\Psi_c(t)\rangle \qquad [7]$$

Here the summation is over all discrete atomic photon-dressed states with energies $\hbar\omega_i$, $|\psi_i\rangle$ is a discrete-(dressed-)state eigenfunction of the Hamiltonian (eqn [1]) normalized to unity, and weighted by the amplitude $\sqrt{c_i} = [1 + \int_0^\infty d\omega G(\omega)/(\omega - \omega_i)^2]^{-1/2}$. The dressed state $|\psi_i\rangle$ consists of an excited-state component and a photon-bound ground-state component.

The population of the excited state for long times:

$$|\alpha(t)|^2 = \sum_{i,i'} c_i c_{i'} \cos(\omega_i - \omega_{i'})t \ \text{ for } t \to \infty \qquad [8]$$

is a nonzero time-constant if there is one discrete stable state, whereas in the case of several such states, it undergoes beats at the frequencies corresponding to their energy differences. The splitting of an excited state $|e\rangle$ into superposed stable states, oscillating at bandgap frequencies ω_i, whose amplitudes c_i, and eigenfrequencies ω_i are controllable by the atomic transition detuning from cutoff, constitutes spontaneous coherence control.

If, however, ω_a is far from the cutoff, then eqn [2] results in an exponentially decaying amplitude:

$$\alpha(t) \approx e^{-(\gamma_a + i\tilde{\omega}_a)t} \qquad [9]$$

where $\tilde{\omega}_a = \omega_a + \Delta_a$, Δ_a, and γ_a being the effective spectral shift and width of the decaying atom. This regime holds for a locally smooth $G(\omega \approx \omega_a)$ such that

$$|\gamma'_a|, |\Delta'_a| \ll 1; \ |\gamma'_a \Delta_a| \ll \gamma_a; \ \gamma_a, |\Delta_a| \ll |\omega_a - \omega_g| \qquad [10]$$

where ω_g is the bandedge frequency nearest to ω_a and the primes denote differentiation with respect to ω_a.

We now apply the foregoing general results to a model DOM distribution. This distribution is derived on keeping the lowest term in the Taylor expansion of the dispersion relation $\omega(\vec{k})$ near a photonic bandedge ω_U (the effects of the further PBG edge are neglected). This yields the effective mass approximation:

$$\omega \approx \omega_U + \sum_{i=x,y,z} (k - k_U)_i^2/m_i \qquad [11]$$

with $1/m_i = (1/2)(\partial^2\omega/\partial k_i^2)|_{\omega=\omega_U}$. In a structure with period L, k_U satisfies the Bragg condition $k_U = \pi/L$. The corresponding DOM in a 3D-periodic structure with an allowed symmetry may be approximated as

$$\rho(\omega) \sim (\omega - \omega_U)^{(1-D)/2}\theta(\omega - \omega_U) \qquad [12]$$

where θ is the step function and D is the dimension of the Brillouin-zone surface spanned by bandedge modes with vanishing group velocity. In realistic photonic structures $D \leq 2$, $D = 2$ corresponding to completely isotropic dispersion (spherical Brillouin-zone surface) and $D = 0$ corresponding to an anisotropic three-dimensional Brillouin zone. Both cases can be represented, respectively, as the limits $\varepsilon \to 0$ and $\varepsilon \to \infty$ of the function:

$$G(\omega) = \frac{C}{\pi} \frac{\sqrt{\omega - \omega_U}}{\omega - \omega_U + \varepsilon} \theta(\omega - \omega_U) \qquad [13]$$

where ε is the 'cutoff-smoothing' parameter and C is the continuum coupling constant. Depending on whether $C^{2/3}/\varepsilon$ is greater or less than one, the continuum is close to the case $\varepsilon = 0$ ($D = 2$) or $\varepsilon \to \infty$ ($D = 0$), respectively.

Using the properties of the model DOM (eqn [13]), we can infer the criteria for the two regimes discussed above:

(i) The conditions for incomplete decay, eqn [6], are now (Figure 2) $\Delta_c \equiv \omega_a - \omega_U < C\varepsilon^{-1/2}$. Hence, the abrupt, singular cutoff of the DOM with $D = 2$ ($\varepsilon \to 0$) implies the existence of a discrete state for any ω_a, either inside or outside the PBG. The energy of the discrete state, $\hbar\omega_0$, which must lie in the PBG, is found to be a real and positive root of the equation $\omega_a - \omega_0 = C/(\sqrt{\omega_U - \omega_0} + \sqrt{\varepsilon})$.

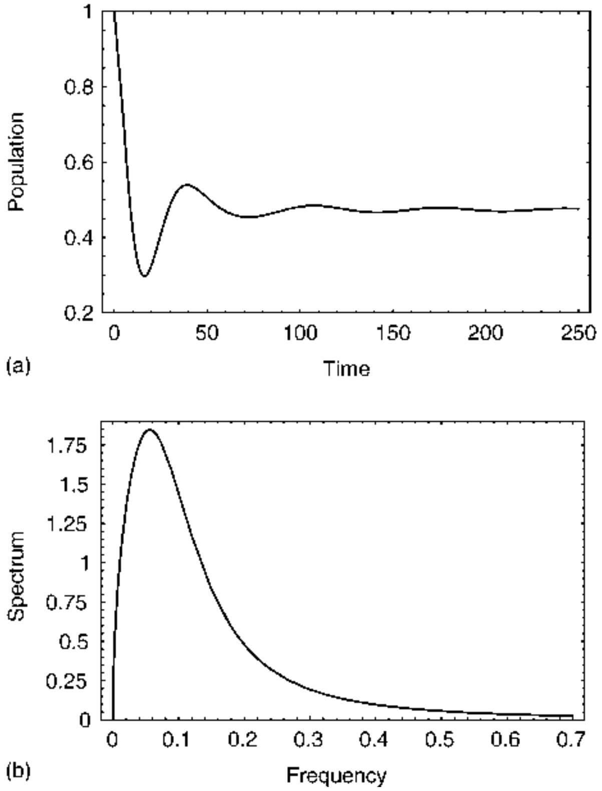

Figure 2 (a) Incomplete decay of the excited state population as a function of $\gamma_c t$, where $\gamma_c = C\varepsilon^{-1/2}$, for $\varepsilon = 10^{-3}$, at cutoff $\Delta_c = 0$. The beat period is $2\pi/(\omega_U - \omega_0)$ and the nondecaying probability is $c_0^2 = 4/9$. (b) The corresponding spectrum. The frequency is normalized to γ_c.

(ii) The conditions for nearly exponential decay, eqn [10], can be shown to reduce now to the requirement that ω_a be in the allowed zone sufficiently far from cutoff, $\Delta_c \gg \min\{C^{2/3}, C\varepsilon^{-1/2}\}$. Under this condition, the intermediate-time exponential decay of the excited state (Figure 3) is given to first approximation by eqn [9], with $\gamma_a = C\sqrt{\Delta_c}/(\varepsilon + \Delta_c)$ and $\Delta_a = -C\sqrt{\varepsilon}/(\varepsilon + \Delta_c)$.

The resulting atomic frequency shift Δ_a is negative in the present model, vanishing for $\varepsilon \to 0$ $(D = 2)$. The long-time behavior of $\alpha(t)$ can be shown to exhibit a tail decaying as $t^{-3/2}$ (or $t^{-1/2}$ at $\Delta_c = C\varepsilon^{-1/2}$) and oscillating at the cutoff frequency ω_U (Figure 3).

With the increase of ε, the smoothing inhibits the decay more strongly for ω_a at cutoff, $\Delta_c = 0$, because $G(\omega_a \simeq \omega_U)$ is now weaker and the stable-state probability c_0 (eqn [7]) is correspondingly larger. By contrast, at a large detuning from cutoff Δ_c, the large-ε smoothing allows the decay to become complete, and the corresponding spectrum is entirely Lorentzian.

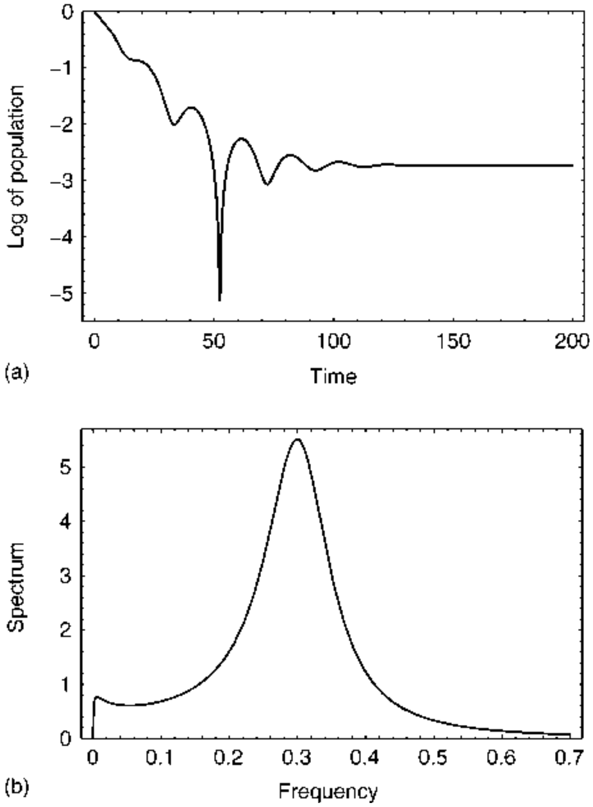

Figure 3 As in Figure 2, for a large detuning from the sharp cutoff, $\varepsilon = 10^{-3}$, $\Delta_c = 0.3$. (a) $\log_{10}|\alpha|t|^2$ is plotted as a function of *i*. The nearly complete decay ($c_0^2 \sim 10^{-3}$) is modulated by beats with frequency Δ_c. A power tail obtains for $t \gg \gamma_a^{-1} \approx \Delta_c^{1/2}/C$, decreasing as $t^{-1/2}$ for $t \ll \Delta_c^2/C^2$ and as $t^{-3/2}$ for $t \gg \Delta_c^2/C^2$. (b) The nearly Lorentzian spectrum, w(ω), has a small peak near ω_U, due to the sharply peaked DOM near the cutoff.

A defect in the periodic structure can produce a narrow-linewidth local mode in the PBG, whose spectral response is describable by a Lorentzian:

$$G_d(\omega) = \frac{\gamma_d}{\pi} \frac{\Gamma_d^2}{\Gamma_d^2 + (\omega - \omega_d)^2} \qquad [14]$$

where γ_d characterizes the coupling strength of the atomic dipole with the defect field, whereas ω_d and Γ_d stand for the line center and width, respectively. The presence of a nonvanishing DOM in the PBG, due to a defect, causes spontaneous emission in the PBG spectral range. This broadens the discrete state ω_0, which becomes metastable (see Figure 4). The oscillator strength of the line at ω_0 is then proportional to c_0.

EIT and Cross-Coupling of Photons in Doped PCs

Nonlinear effects, whereby one light beam influences another, require strong fields or else light confinement in a high-Q cavity. The analysis above has shown that by choosing an appropriate detuning of an atomic transition from the PBG cutoff, the spontaneous coherence is established between the states of an initially excited atom. The ability to control this coherence, by varying the detuning, opens interesting perspectives for optical processes in PCs containing multilevel atoms with a resonant transition near a bandedge. From among such processes, we discuss here electromagnetically induced transparency (EIT) and its applicability to nonlinear photon switching or giant cross-phase modulation.

Let us examine the nonlinear coupling of two weak (single-photon) optical fields E_a and E_b with the frequencies ω_a and ω_b, respectively, propagating along the z-axis in a PC dilutely doped with identical four-level atoms (**Figure 5a**). These fields interact with the atoms via the transitions $|1\rangle \to |2\rangle$ and $|3\rangle \to |4\rangle$, respectively, while the transition $|2\rangle \to |3\rangle$ is coupled to the structured mode-continuum $\rho(\omega)$ in the PC (**Figure 5b**). Initially the atoms are in the ground state $|1\rangle$ and the continuum is in the vacuum state $\{0_\omega\}$. Then the wavefunction of the system at the position z_l of the lth atom reads:

$$|\Psi(z_l,t)\rangle = \alpha_1|1,\{0_\omega\}\rangle + \alpha_2|2,\{0_\omega\}\rangle + \int \beta_{3,\omega}|3,1_\omega\rangle\rho(\omega)\mathrm{d}\omega + \int \beta_{4,\omega}|4,1_\omega\rangle\rho(\omega)\mathrm{d}\omega \qquad [15]$$

The consecutive terms in eqn [15] denote the atom being in states $|1\rangle$, $|2\rangle$, $|3\rangle$, or $|4\rangle$, with zero $\{0_\omega\}$ or one $|1_\omega\rangle$ photon in mode ω whose DOM is $\rho(\omega)$, and α_1, α_2, $\beta_{3,\omega}$, or $\beta_{4,\omega}$ are the corresponding probability amplitudes. Upon making the weak-field linear-response approximation, we can set $\alpha_1 \simeq 1$ and solve the set of equations for the slowly varying (compared to an optical cycle) probability amplitudes α and β, using the perturbation theory. Under these conditions, we effectively obtain a free-space propagation of the E_b field. By contrast, the evolution of the E_a field in the slowly varying envelope approximation, is given by $E_\mathrm{a}(z,t) = E_\mathrm{a}(0, t - z/v_\mathrm{g})\exp(ipz)$. We thus see that the real part of the macroscopic complex polarizability p is responsible for the phase shift ϕ_a of the E_a field, $\phi_\mathrm{a} = \mathrm{Re}(p)z$, while the probability of the absorption $\mathscr{A}$ of the field depends on the imaginary part of p, $\mathscr{A} = 1 - \exp[-2\mathrm{Im}(p)z]$. The polarizability is expressed by

$$p = a_0 \frac{i\gamma_2/2}{\gamma_2/2 - i\Delta_\mathrm{a} + I(\Delta_\mathrm{a})} \qquad [16]$$

where $a_0 \equiv \sigma_0 N$ is the linear resonant absorption coefficient on the atomic transition $|1\rangle \to |2\rangle$, with σ_0 the resonant absorption cross-section and N the density of doping atoms, γ_2 is the radiative width of

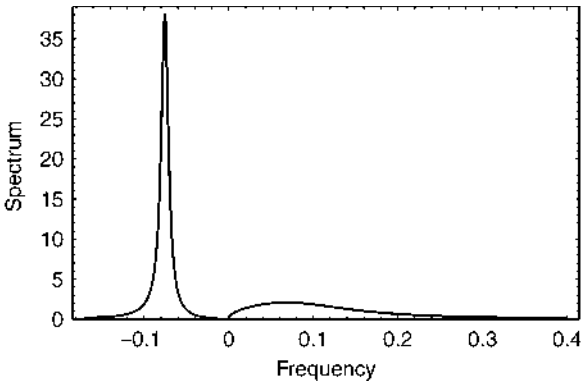

Figure 4 Spectrum w(ω)(same units as in **Figures 2** and **3**) for ω_a at cutoff, $\Delta_\mathrm{c} = 0$, in the presence of a defect in the PBG, $\omega_\mathrm{d} = -10$, $\gamma = 0.1$, $\Gamma_\mathrm{d} = 3$, $\varepsilon = 10^{-3}$. The oscillator strength is now roughly equally shared between the distribution above cutoff (same as in **Figure 2b**) and the defect peak at ω_0.

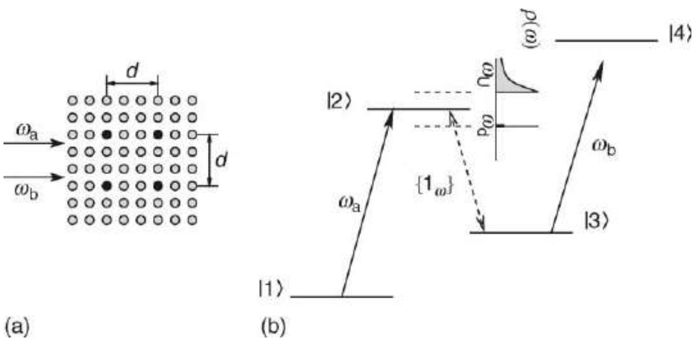

Figure 5 (a) Photonic crystal dilutely doped with atoms located at black dots. (b) Four-level atom coupled to a structured continuum $\rho(\omega)$ near the band-edge or defect mode frequencies (DOM plotted) via the intermediate transition $|2\rangle \to |3\rangle$ and interacting with two weak fields E_a and E_b at the sideband transitions $|1\rangle \to |2\rangle$ and $|3\rangle \to |4\rangle$, respectively.

state $|2\rangle$, $\Delta_a = \omega_a - \omega_{21}$ is the detuning from the atomic resonance ω_{21}, and $I(\Delta_a)$ is the integral of the saturation factor over the structured DOM. The group velocity v_g is expressed as $v_g = \partial_{k_a}\omega_a = [n_a/c + \partial_{\omega_a}\mathrm{Re}(p)]^{-1}$, where n_a is the (averaged) refraction index at the frequency ω_a.

To calculate $I(\Delta_a)$, we assume the isotropic PBG model, eqn [13] with $\varepsilon \to 0$, with the atoms doped at the positions of the local defects in the PC separated by a distance d from each other. These defect modes in the PBG are localized around each atomic site in a volume $V_d \simeq (rL)^3$ of several $(r)^3$ lattice cells L^3, with $L \simeq \pi c/\omega_{23}$, and serve as effective high-Q cavities, eqn [14] with $\Gamma_d \ll 1$. Assuming $|\Delta_b| \gg |\Delta|, |\Delta_a|, \gamma_4$, where $\Delta_b = \omega_b - \omega_{43}$ and $\Delta = \omega - \omega_{23}$, the integration leads to

$$I(\Delta_a) = \frac{B_d^2}{\gamma_{31} - i(\Delta_a - \Delta_d - s_3)} - \frac{B_U^{3/2}}{\sqrt{i\gamma_{31} + (\Delta_a - \Delta_U - s_3)}} \qquad [17]$$

where $\Delta_{d,U} = \omega_{d,U} - \omega_{23} \ll \omega_{23}$ are the detunings of the defect-mode and PBG-edge frequencies from the atomic resonance ω_{23}, γ_{31} the $|1\rangle \leftrightarrow |3\rangle$ decoherence rate, $s_3 = (\mu_{34}^2/\hbar^2\Delta_b)|E_b|^2$ is the E_b field-induced ac Stark shift of level $|3\rangle$ (μ_{ij} is the atomic dipole matrix element on the transition $|i\rangle \to |j\rangle$), and B_d and B_U are the coupling constants of the atom with the structured reservoir, whose main contributions are near ω_d and ω_U.

To illustrate the results of the foregoing analysis, we plot in Figure 6 the imaginary and real parts of the polarizability (eqn [16]). Consider first the case of one incident field $E_a(E_b = 0)$. Clearly, two frequency regions, $\Delta_a \sim \Delta_d$ and $\Delta_a \sim \Delta_U$, where the absorption vanishes and, at the same time, the dispersion slope is steep, are of particular interest. One can see that there is, however, a substantial difference between the spectra in the foregoing frequency regions, for the following physical reasons. First, in the vicinity of Δ_d, the atom interacts with the defect mode as in a high-Q cavity. This strong interaction 'dresses' the atomic states $|2\rangle$ and $|3\rangle$, thereby splitting the spectrum around $\Delta_a \simeq \Delta_d$ by the amount equal roughly to $2B_d$. Near the two-photon Raman resonance $\Delta_a = \Delta_d$, the two alternative transition paths $|1\rangle \to |2\rangle$ (the direct transition) and $|1\rangle \to |2\rangle \to |3\rangle \to |2\rangle$ (transition via state $|3\rangle$) interfere distractively with each other, cancelling thus the absorption of the E_a field and the medium becomes transparent to the radiation. This effect has been widely studied in atomic vapors, where the transition $|2\rangle \to |3\rangle$ is strongly driven by a coherent laser field, and is called electromagnetically induced transparency (EIT). The transparency window is rather broad and is given by the inverse Lorentzian (see the first term on the right-hand side of eqn [17]). Due to the steepness of the dispersion curve, the corresponding group velocity is much smaller than the speed of light, $v_g \simeq 2B_d^2/(\gamma_2 a_0) \ll c$, which leads to a large delay time $T_{del} = \zeta/v_g$ at the exit $z = \zeta$ from the medium. One has to keep in mind, however, that the absorption-free propagation time is limited by the EIT decoherence time $T_{del} < \gamma_{31}^{-1}$, which imposes a limitation on the length ζ of the active PC medium. Second, in the vicinity of Δ_U, the strong interaction of the atom with the continuum near the bandedge ω_U causes the Autler–Townes splitting of level $|2\rangle$ into a doublet with a separation equal roughly to B_U. One component of this doublet is shifted out of the PBG, while the other one remains

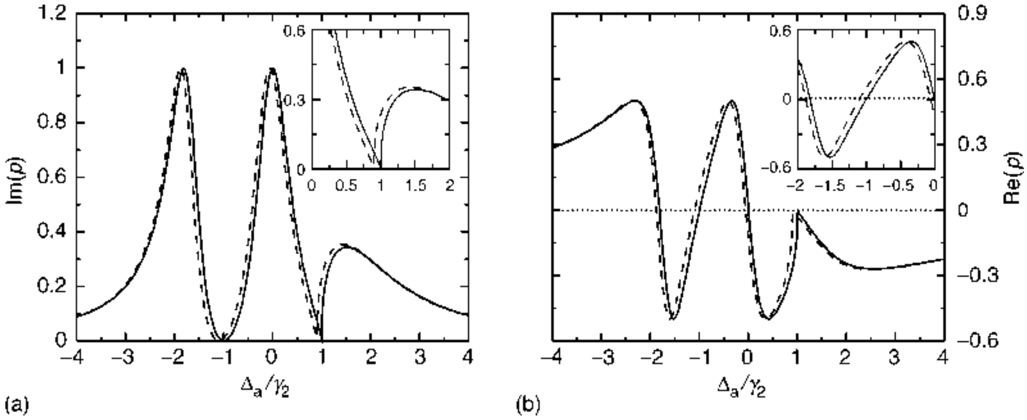

Figure 6 (a) Imaginary and (b) real part of the complex polarizability p as a function of the detuning Δ_a for the case $E_b = 0$ (solid lines) and $E_b \neq 0$ (dashed lines). The parameters (normalized by γ_2) are: $\Delta_d = -1$, $\Delta_U = 1$, $\gamma_{31} = 0.001$, $B_d = B_U = 1$, $s_3 = -0.1$, and $a_0 = 1\ \mathrm{cm}^{-1}$. The insets magnify the important frequency regions.

within the gap and forms the photon–atom bound state. Consequently, there is vanishing absorption and rapid variation of the dispersion at $\Delta_a \simeq \Delta_U$. Since the transparency region is very narrow with a width $\delta\omega \sim \gamma_{31}(\ll \gamma_2, B_U)$, for an absorption-free propagation of the E_a pulse, its temporal width τ_a should satisfy the condition $\tau_a > \pi/\delta\omega$. Simultaneously, a small deviation from the condition $\Delta_a = \Delta_U$ will lead to a strong increase in the absorption of the E_a field.

Let us now switch on the E_b field. As seen from eqn [17], its effect is merely to shift the spectrum by the amount equal to s_3 (Figure 6). This shift, however, will have different implications in the two frequency regions distinguished above: if $s_3 \ll B_d$, i.e., the Stark shift is smaller than the width of the EIT window at Δ_d, the medium will still remain transparent for an E_a field with the detuning $\Delta_a = \Delta_d$, but its phase will experience an appreciable nonlinear shift ϕ_a, given by $\phi_a = Re(p)z \simeq -\gamma_2 a_0 s_3 z/(2B_d^2)$. On the other hand, for an E_a field with the detuning $\Delta_a = \Delta_U$, the medium, which is transparent for $E_b = 0$, $\mathrm{Im}(p) \ll a_0$, will become highly absorptive (opaque) even for such a small frequency shift as s_3 (provided $s_3 < 0$ and $|s_3| > \Delta\omega$), $\mathrm{Im}(p) \simeq \gamma_2 a_0 \sqrt{|s_3|}/(2B_U^{3/2})$ and thus acting as an ultrasensitive, effective switch.

The remaining question is how to maximize the interaction between the E_a pulse, which propagates with a small group velocity, and the E_b pulse, which propagates with a velocity close to the speed of light. The interaction between the fields is maximized if: they enter the medium simultaneously; the transverse shapes of their envelopes overlap completely; and the pulse length l_b of the E_b field satisfies the condition $(l_b + \zeta)/c \leq \zeta/v_g$. Then the E_b pulse leaves the medium not later than the E_a pulse. The effective interaction length between the two pulses is, therefore, $z_{\mathrm{eff}} \sim l_b v_g/c \leq \zeta$, after which the two pulses slip apart. Thus, the presence of the E_b (control) field induces either strong absorption or a large phase shift of the E_a (signal) field, depending on the frequency region employed.

The effects surveyed above reveal unusual features of spontaneous emission and photon–atom binding in PCs, atomic interaction with the field of a high-Q defect mode, and nonlinear coupling of two fields via four-level dopant atoms in PCs. These effects are of fundamental interest. In addition, they can serve as the basis for highly efficient optical communications and data processing, in either the classical or the quantum domain, by providing two key elements: ultrasensitive nonlinear phase-shifters and photon switches.

List of Units and Nomenclature

Atomic or photonic eigenstates [dimensionless]	$\vert\ldots\rangle$
Complex polarizability [cm^{-1}]	p
Coupling constant [s^{-1}]	κ, $C^{2/3}$, B
Decay rate [s^{-1}]	γ
Density of modes [s]	ρ
Frequency [s^{-1}]	ω, Δ, s
Phase [dimensionless]	ϕ
Probability amplitudes [dimensionless]	α, β, c
Reservoir spectral response [s^{-1}]	G
Resonant absorption coefficient [cm^{-1}]	a_0
Spontaneous emission spectrum [s]	W
Wave vector [m^{-1}]	k

See also

Electromagnetically Induced Transparency. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides.

Further Reading

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton, NJ: Princeton University Press.

John S and Wang J (1991) Quantum optics of localized light in a photonic band gap. *Physics Review B* 43: 12772–12789.

Kofman AG, Kurizki G and Sherman B (1994) Spontaneous and induced atomic decay in photonic band structures. *Journal of Modern Optics* 41: 353–384.

Lambropoulos P, Nikolopoulos GM, Nielsen TR and Bay S (2000) Fundamental quantum optics in structured reservoirs. *Reports on Progress in Physics* 63: 455–503.

Loudon R (1983) *The Quantum Theory of Light*. Oxford, UK: Clarendon Press.

Meystre P and Sargent M III (1991) *Elements of Quantum Optics*. Berlin: Springer Verlag.

Petrosyan D and Kurizki G (2001) Photon–photon correlations and entanglement in doped photonic crystals. *Physics Review A* 64: 023810-1-6.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. Springer Series in Optical Sciences, 80. Berlin: Springer Verlag.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Electromagnetic Theory

S G Johnson and J D Joannopoulos,
Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystals are periodically structured electromagnetic media, generally possessing photonic bandgaps: ranges of frequency in which light cannot propagate through the structure. This periodicity, whose lengthscale is proportional to the wavelength of light in the bandgap, is the electromagnetic analog of a crystalline atomic lattice, where the latter acts on the electron wavefunction to produce the familiar band gaps, semiconductors, etc., of solid-state physics. The study of photonic crystals is likewise governed by the Bloch–Floquet theorem, and intentionally introduced defects in the crystal (analogous to electronic dopants) give rise to localized electromagnetic states: linear waveguides and point-like cavities. The crystal can thus form a kind of perfect optical 'insulator', which can confine light around sharp bends, in lower-index media, and within wavelength-scale cavities, among other novel possibilities for control of electromagnetic phenomena. Below is introduced the basic theoretical background of photonic crystals in one, two, and three dimensions (schematically depicted in Figure 1), as well as hybrid structures that combine photonic-crystal effects in some directions with more-conventional index guiding in other directions. (Line and point defects in photonic crystals are discussed in another article.)

Electromagnetic wave propagation in periodic media was first studied by Lord Rayleigh in 1887, in connection with the peculiar reflective properties of a crystalline mineral with periodic 'twinning' planes (across which the dielectric tensor undergoes a mirror flip). These correspond to one-dimensional photonic crystals, and he identified the fact that they have a narrow bandgap prohibiting light propagation through the planes. This bandgap is angle-dependent, due to the differing periodicities experienced by light propagating at non-normal incidences, producing a reflected color that varies sharply with angle. (A similar effect is responsible for many other iridescent colors in nature, such as those of butterfly wings and abalone shells.) Although multilayer films received intensive study over the following century, it was not until 100 years later, when Yablonovitch and John, in 1987, joined the tools of classical electromagnetism and solid-state physics, that the concepts of omnidirectional photonic bandgaps in two and three dimensions was introduced. This generalization, which inspired the name 'photonic crystal', led to many subsequent developments in their fabrication, theory, and application, from integrated optics to negative refraction to optical fibers that guide light in air.

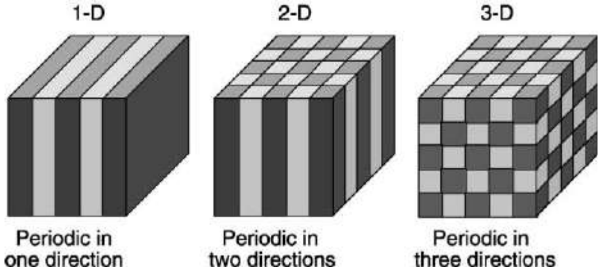
1-D
2-D
3-D
Periodic in one direction
Periodic in two directions
Periodic in three directions

Figure 1 Schematic depiction of photonic crystals periodic in one, two, and three directions, where the periodicity is in the material (typically dielectric) structure of the crystal. Only a 3d periodicity, with a more complex topology than is shown here, can support an omnidirectional photonic bandgap.

Maxwell's Equations in Periodic Media

The study of wave propagation in three-dimensionally periodic media was pioneered by Felix Bloch in 1928, unknowingly extending an 1883 theorem in one dimension by Gaston Floquet. Bloch proved that waves in such a medium can propagate without scattering, their behavior described by a periodic envelope function multiplied by a planewave. Although Bloch studied quantum mechanics, leading to the surprising result that electrons in a conductor scatter only from imperfections and not from the periodic ions, the same techniques can be applied to electromagnetism by casting Maxwell's equations as an eigenproblem in analog with Schrödinger's equation. By combining the source-free Faraday's and Ampere's laws at a fixed (angular) frequency ω, i.e., time dependence $e^{-i\omega t}$, one can obtain an equation in only the magnetic field $\vec{H}$:

$$\vec{\nabla} \times \frac{1}{\varepsilon} \vec{\nabla} \times \vec{H} = \left(\frac{\omega}{c}\right)^2 \vec{H} \qquad [1]$$

where ε is the dielectric function $\varepsilon(x, y, z)$ and c is the speed of light. This is an eigenvalue equation, with eigenvalue $(\omega/c)^2$ and an eigen-operator $\vec{\nabla} \times (1/\varepsilon)\vec{\nabla} \times$ that is Hermitian (acts the same to the left and right) under the inner product $\int \vec{H}^* \cdot \vec{H}'$ between two fields $\vec{H}$ and $\vec{H}'$. (The two curls correspond roughly to the 'kinetic energy' and $1/\varepsilon$ to the 'potential' compared to the Schrödinger Hamiltonian $\nabla^2 + V$.) It is sometimes more convenient to write a generalized Hermitian eigenproblem in the electric field $\vec{E}$, $\vec{\nabla} \times \vec{\nabla} \times \vec{E} = (\omega/c)^2 \varepsilon \vec{E}$, which separates the kinetic and potential terms. Electric fields that lie in higher ε, i.e., lower potential, will have lower ω; this is discussed further in the context of the variational theorem of eqn [3].

Thus, the same linear-algebraic theorems as those in quantum mechanics can be applied to the electromagnetic wave solutions. The fact that the eigen-operator is Hermitian and positive-definite (for real $\varepsilon > 0$) implies that the eigenfrequencies ω are real, for example, and also leads to orthogonality, variational formulations, and perturbation–theory relations that are discussed further below. An important difference compared to quantum mechanics is that there is a transversality constraint: one typically excludes $\vec{\nabla} \cdot \vec{H} \neq 0$ (or $\vec{\nabla} \cdot \varepsilon \vec{E} \neq 0$) eigensolutions, which lie at $\omega = 0$; i.e., static-field solutions with free magnetic (or electric) charge are forbidden.

Bloch Waves and Brillouin Zones

A photonic crystal corresponds to a periodic dielectric function $\varepsilon(\vec{x}) = \varepsilon(\vec{x} + \vec{R}_i)$ for some primitive lattice vectors $\vec{R}_i$ ($i = 1, 2, 3$ for a crystal periodic in all three dimensions). In this case, the Bloch–Floquet theorem for periodic eigenproblems states that the solutions to eqn [1] can be chosen of the form $\vec{H}(\vec{x}) = e^{i\vec{k}\cdot\vec{x}} \vec{H}_{n,\vec{k}}(\vec{x})$ with eigenvalues $\omega_n(\vec{k})$, where $\vec{H}_{n,\vec{k}}$ is a periodic envelope function satisfying

$$\left(\vec{\nabla} + i\vec{k}\right) \times \frac{1}{\varepsilon}\left(\vec{\nabla} + i\vec{k}\right) \times \vec{H}_{n,\vec{k}} = \frac{\omega_n(\vec{k})}{c}^2 \vec{H}_{n,\vec{k}} \qquad [2]$$

yielding a different Hermitian eigenproblem over the primitive cell of the lattice at each Bloch wavevector $\vec{k}$. This primitive cell is a finite domain if the structure is periodic in all directions, leading to discrete eigenvalues labeled by $n = 1, 2, \ldots$. These eigenvalues $\omega_n(\vec{k})$ are continuous functions of $\vec{k}$, forming discrete 'bands' when plotted versus the latter, in a 'band structure' or dispersion diagram – both ω and $\vec{k}$ are conserved quantities, meaning that a band diagram maps out all possible interactions in the system. (Note also that $\vec{k}$ is not required to be real; complex $\vec{k}$ gives evanescent modes that can exponentially decay from the boundaries of a finite crystal, but which cannot exist in the bulk.)

Moreover, the eigensolutions are periodic functions of $\vec{k}$ as well: the solution at $\vec{k}$ is the same as the solution at $\vec{k} + \vec{G}_j$, where $\vec{G}_j$ is a primitive reciprocal lattice vector defined by $\vec{R}_i \cdot \vec{G}_j = 2\pi\delta_{i,j}$. Thanks to this periodicity, one need only compute the eigensolutions for $\vec{k}$ within the primitive cell of this reciprocal lattice – or, more conventionally, one considers the set of inequivalent wavevectors closest to the $\vec{k} = 0$ origin, a region called the first Brillouin zone. For example, in a one-dimensional system, where $R_1 = a$ for some periodicity a and $G_1 = 2\pi/a$, the first Brillouin zone is the region $k = -\pi/a \ldots \pi/a$; all other wavevectors are equivalent to some point in this zone under translation by a multiple of G_1. Furthermore, the first Brillouin zone may itself be redundant if the crystal possesses additional symmetries such as mirror planes; by eliminating these redundant regions, one obtains the irreducible Brillouin zone, a convex polyhedron that can be found tabulated for most crystalline structures. In the preceding one-dimensional example, since most systems will have time-reversal symmetry ($k \rightarrow -k$), the irreducible Brillouin zone would be $k = 0 \ldots \pi/a$.

The familar dispersion relations of uniform waveguides arise as a special case of the Bloch formalism: such translational symmetry corresponds to a period $a \rightarrow 0$. In this case, the Brillouin zone of the wavevector k (also called β) is unbounded, and the envelope function $\vec{H}_{n,k}$ is a function only of the transverse coordinates.

The Origin of the Photonic Bandgap

A complete photonic bandgap is a range of ω in which there are no propagating (real $\vec{k}$) solutions of Maxwell's eqn [2] for any $\vec{k}$, surrounded by propagating states above and below the gap. There are also incomplete gaps, which only exist over a subset of all possible wavevectors, polarizations, and/or symmetries. Both sorts of gaps are discussed in the subsequent sections, but in either case, their origins are the same and can be understood by examining the consequences of periodicity for a simple one-dimensional system.

Consider a one-dimensional system with uniform $\varepsilon = \bar{\varepsilon}$, which has planewave eigensolutions $\omega(k) = ck$, as depicted in Figure 2(left). This ε has trivial periodicity a for any $a \geq 0$, with $a = 0$ giving the usual unbounded dispersion relation. One is free, however, to label the states in terms of Bloch envelope functions and wavevectors for some $a \neq 0$, in which case the bands for $|k| > \pi/a$ are translated (folded) into the first Brillouin zone, as shown by the dashed lines in Figure 2(left). In particular, the $k = -\pi/a$ mode in this description now lies at an equivalent wavevector to the $k = \pi/a$ mode, and at the same frequency; this accidental degeneracy is an artifact of

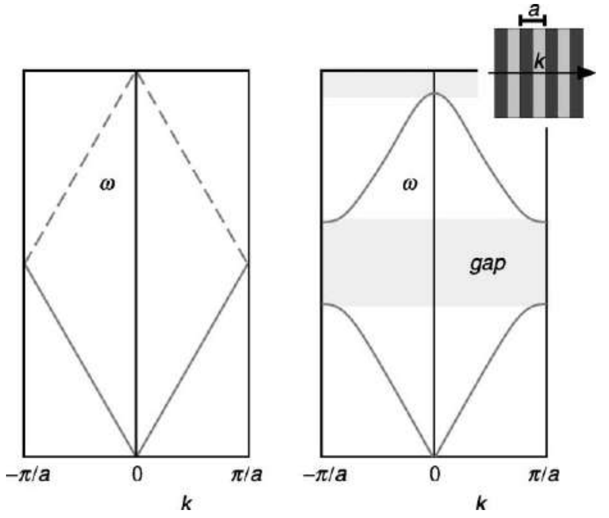

Figure 2 Left: Dispersion relation (band diagram), frequency ω versus wavenumber k, of a uniform one-dimensional medium, where the dashed lines show the 'folding' effect of applying Bloch's theorem with an artificial periodicity a. Right: Schematic effect on the bands of a physical periodic dielectric variation (inset), where a gap has been opened by splitting the degeneracy at the $k = \pm\pi/a$ Brillouin-zone boundaries (as well as a higher-order gap at $k = 0$).

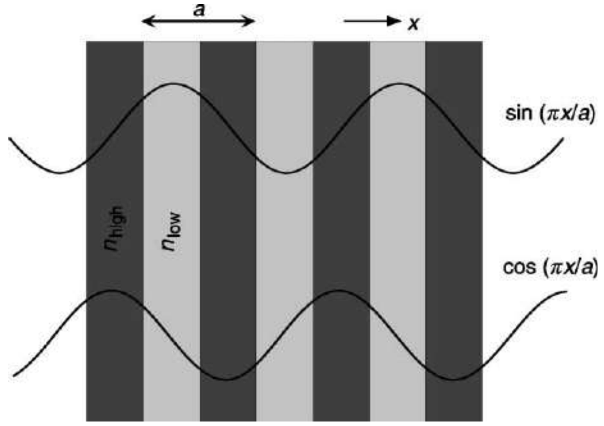

Figure 3 Schematic origin of the band gap in one dimension. The degenerate $ka = \pm\pi/a$ planewaves of a uniform medium are split into $\cos(\pi x/a)$ and $\sin(\pi x/a)$ standing waves by a dielectric periodicity, forming the lower and upper edges of the bandgap, respectively – the former has electric-field peaks in the high dielectric (n_{high}) and so will lie at a lower frequency than the latter (which peaks in the low dielectric).

the 'artificial' period that has been chosen. Instead of writing these wave solutions with electric fields $\vec{E}(x) \sim e^{\pm i\pi x/a}$, one can equivalently write linear combinations $e(x) = \cos(\pi x/a)$ and $o(x) = \sin(\pi x/a)$ as shown in Figure 3, both at $\omega = c\pi/a\sqrt{\bar{\varepsilon}}$. Now, however, suppose that one perturbs ε so that it is nontrivially periodic with period a; for example, a sinusoid $\varepsilon(x) = \bar{\varepsilon}\cdot[1 + \Delta\cdot\cos(2\pi x/a)]$, or a square wave as in the inset of Figure 2(right). In the presence of such an oscillating 'potential', the accidental degeneracy between $e(x)$ and $o(x)$ is broken: supposing $\Delta > 0$, then the field $e(x)$ is more concentrated in the higher-ε regions than $o(x)$, and so lies at a lower frequency. This opposite shifting of the bands away from the mid-gap frequency $\omega \cong c\pi/a\sqrt{\bar{\varepsilon}}$ creates a bandgap, as depicted in Figure 2(right). (In fact, from the perturbation theory described subsequently, one can show that for $\Delta \ll 1$ the bandgap, as a fraction of mid-gap frequency, is $\Delta\omega/\omega \cong \Delta/2$.) By the same arguments, it follows that any periodic dielectric variation in one dimension will lead to a bandgap, albeit a small gap for a small variation; a similar result was identified by Lord Rayleigh in 1887.

More generally, it follows immediately from the properties of Hermitian eigensystems that the eigenvalues minimize a variational problem:

$$\omega^2_{n,\vec{k}} = \min_{\vec{E}_{n,\vec{k}}} \frac{\int \left|\left(\vec{\nabla} + i\vec{k}\right) \times \vec{E}_{n,\vec{k}}\right|^2}{\int \varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2} c^2 \tag{3}$$

in terms of the periodic electric field envelope $\vec{E}_{n,\vec{k}}$, where the numerator minimizes the 'kinetic energy' and the denominator minimizes the 'potential energy' Here, the $n > 1$ bands are additionally constrained to be orthogonal to the lower bands:

$$\int \vec{H}^*_{m,\vec{k}} \cdot \vec{H}_{n,\vec{k}} = \int \varepsilon \vec{E}^*_{m,\vec{k}} \cdot \vec{E}_{n,\vec{k}} = 0 \tag{4}$$

for $m < n$. Thus, at each $\vec{k}$, there will be a gap between the lower 'dielectric' bands concentrated in the high dielectric (low potential) and the upper 'air' bands that are less concentrated in the high dielectric: the air bands are forced out by the orthogonality condition, or otherwise must have fast oscillations that increase their kinetic energy. (The dielectric/air bands are analogous to the valence/conduction bands in a semiconductor.)

In order for a complete bandgap to arise in two or three dimensions, two additional hurdles must be overcome. First, although in each symmetry direction of the crystal (and each $\vec{k}$ point) there will be a bandgap by the one-dimensional argument, these bandgaps will not necessarily overlap in frequency (or even lie between the same bands). In order that they overlap, the gaps must be sufficiently large, which implies a minimum ε contrast (typically at least 4/1 in 3d). Since the 1d mid-gap frequency $\sim c\pi/a\sqrt{\bar{\varepsilon}}$ varies inversely with the period a, it is also helpful if the periodicity is nearly the same in different directions – thus, the largest gaps typically arise for hexagonal lattices in 2d and fcc lattices in 3d, which have the most nearly circular/spherical Brillouin zones. Second, one must take into account the vectorial boundary conditions on the electric field: moving across a dielectric boundary from ε to some $\varepsilon' < \varepsilon$, the inverse 'potential' $\varepsilon|\vec{E}|^2$ will decrease discontinuously if $\vec{E}$ is parallel to the interface ($\vec{E}_\parallel$ is continuous) and will increase discontinuously if $\vec{E}$ is perpendicular to the interface ($\varepsilon\vec{E}_\perp$ is continuous). This means that, whenever the electric field lines cross a dielectric boundary, it is much harder to strongly contain the field energy within the high dielectric, and the converse is true when the field lines are parallel to a boundary. Thus, in order to obtain a large bandgap, a dielectric structure should consist of thin, continuous veins/membranes along which the electric field lines can run – this way, the lowest band(s) can be strongly confined, while the upper bands are forced to a much higher frequency because the thin veins cannot support multiple modes (except for two orthogonal polarizations). The veins must also run in all directions, so that this confinement can occur for all $\vec{k}$ and polarizations, necessitating a complex topology in the crystal.

Ultimately, however, in two or three dimensions there are only rules of thumb for the existence of a bandgap in a periodic structure, since no rigorous criteria have yet been determined. This made the design of 3d photonic crystals a trial and error process, with the first example by Ho *et al.* of a complete 3d gap coming three years after the initial 1987 concept. As is discussed by the final section below, a small number of families of 3d photonic crystals have since been identified, with many variations thereof explored for fabrication.

Computational Techniques

Because photonic crystals are generally complex, high index-contrast, two- and three-dimensional vectorial systems, numerical computations are a crucial part of most theoretical analyses. Such computations typically fall into three categories: time-domain 'numerical experiments' that model the time-evolution of the fields with arbitrary starting conditions in a discretized system (e.g., finite-difference); definite-frequency transfer matrices wherein the scattering matrices are computed in some basis to extract transmission/reflection through the structure; and frequency-domain methods to directly extract the Bloch fields and frequencies by diagonalizing the eigenoperator. The first two categories intuitively correspond to directly measurable quantities such as transmission (although they can also be used to compute e.g., eigenvalues), whereas the third is more abstract, yielding the band diagrams that provide a guide to interpretation of measurements as well as a starting-point for device design and semi-analytical methods. Moreover, several band diagrams are included in the following sections, and so the frequency-domain method used to compute them is briefly outlined here.

Any frequency–domain method begins by expanding the fields in some complete basis, $\vec{H}_{\vec{k}}(\vec{x}) = \sum_n h_n \vec{b}_n(\vec{x})$, transforming the partial differential eqn [2] into a discrete matrix eigenvalue problem for the coefficients h_n. Truncating the basis to N elements leads to $N \times N$ matrices, which could be diagonalized in $O(N^3)$ time by standard methods. This is impractical for large 3d systems, however, and is also unnecessary – typically, one only wants the few lowest eigenfrequencies, in which case one can use iterative eigensolver methods requiring only $\sim O(N)$ time. Perhaps the simplest such method is based directly on the variational theorem eqn [3]: given some starting coefficients h_n, one iteratively minimizes the variational 'Rayleigh' quotient using e.g., preconditioned conjugate–gradient descent. This yields the lowest band's eigenvalue and field, and upper bands are found by the same minimization while orthogonalizing against the lower bands ('deflation'). There is one additional difficulty, however, and that is that one must at the same time enforce the $(\vec{\nabla} + i\vec{k}) \cdot \vec{H}_{\vec{k}} = 0$ transversality constraint, which is nontrivial in three dimensions. The simplest way to maintain this constraint is to employ a basis that is already transverse, for example planewaves $\vec{h}_{\vec{G}} e^{i\vec{G}\cdot\vec{x}}$ with transverse amplitudes $\vec{h}_{\vec{G}} \cdot (\vec{G} + \vec{k}) = 0$. (In such a planewave basis, the action of the eigen-operator can be computed via a fast Fourier transform in $O(N \log N)$ time.)

Semi-analytical Methods: Perturbation Theory

As in quantum mechanics, the eigenstates can be the starting point for many analytical and semi-analytical studies. One common technique is perturbation theory, applied to small deviations from an ideal system – closely related to the variational eqn [3], perturbation theory can be exploited to consider effects such as nonlinearities, material absorption, fabrication disorder, and external tunability. Not only is perturbation theory useful in its own right, but it also illustrates both old and peculiarly new features that arise in such analyses of electromagnetism compared to scalar problems such as quantum mechanics.

Given an unperturbed eigenfield $\vec{E}_{n,\vec{k}}$ for a structure ε, the lowest-order correction $\Delta\omega_n^{(1)}$ to the eigenfrequency from a small perturbation $\Delta\varepsilon$ is given by

$$\Delta\omega_n^{(1)} = -\frac{\omega_n(\vec{k})}{2} \frac{\int \Delta\varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2}{\int \varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2} \qquad [5]$$

where the integral is over the primitive cell of the lattice. A Kerr nonlinearity would give $\Delta\varepsilon \sim |\vec{E}|^2$, material absorption would produce an imaginary frequency correction (decay coefficient) from a small imaginary $\Delta\varepsilon$, and so on. Similarly, one can compute the shift in frequency from a small $\Delta\vec{k}$ in order to determine the group velocity $d\omega/dk$; this variation of perturbation theory is also called $k \cdot p$ theory. All such first-order perturbation corrections are well known from quantum mechanics, and in the limit of infinitesimal perturbations give the exact Hellman–Feynman expression for the derivative of the eigenvalue. However, in the limit where $\Delta\varepsilon$ is a small shift Δh of a dielectric boundary between some ε_1 and ε_2, an important class of geometric perturbation, eqn [5] gives a surface integral of $|\vec{E}|^2$ on the interface, but this is ill-defined because the field there is discontinous. The proper derivation of perturbation theory in the face of such discontinuity requires a more careful limiting process from an anisotropically smoothed system, yielding the surface integral:

$$\Delta\omega_n^{(1)} = -\frac{\omega_n(\vec{k})}{2}\,\frac{\int\int \Delta h\left(\Delta\varepsilon_{12}\left|\vec{E}_{\|}\right|^2 - \Delta\varepsilon_{12}^{-1}\left|D_{\perp}\right|^2\right)}{\int \varepsilon\left|\vec{E}_{n,\vec{k}}\right|^2}$$

where $\Delta\varepsilon_{12} = \varepsilon_1 - \varepsilon_2, \Delta\varepsilon_{12}^{-1} = \varepsilon_1^{-1} - \varepsilon_2^{-1}$, and $\vec{E}_{\|}/D_{\perp}$ denotes the (continuous) interface parallel/perpendicular components of the unperturbed electric/displacement eigenfield. A similar expression is required in high index-contrast systems to employ, e.g., coupled-mode theory for slowly-varying waveguides or Green's functions for interface roughness.

Standard perturbation–theory techniques also provide expressions for higher-order corrections to the eigenvalue and eigenfield, based on an expansion in the basis of the unperturbed eigenfields. This approach, however, runs into immediate difficulty because the eigenfields are also subject to the transversality constraint, $(\vec{\nabla} + i\vec{k}) \cdot \varepsilon\vec{E}_{\vec{k}} = 0$, and this constraint varies with ε and $\vec{k}$ – the eigenfields $\vec{E}_{\vec{k}}$ are not a complete basis for the eigenfields constrained at a different ε or $\vec{k}$. For ε perturbations, this problem can be eliminated by using the $\vec{H}$ or $\vec{D}$ eigenproblems, whose constraints are independent of ε. For $\vec{k}$ perturbations, one can employ a transformation by Sipe to derive a corrected higher-order perturbation theory (for e.g., the group-velocity dispersion), based on the fact that all of the non-transverse fields lie at $\omega = 0$. Such completeness issues also arise applying the variational eqn [3], as was noted in the previous section: in order for a useful variational bound to apply, one must operate in the constrained (transverse) subspace.

Two-Dimensional Photonic Crystals

After the identification of one-dimensional bandgaps, it took a full century to add a second dimension, and three years to add the third. It should therefore come as no surprise that 2d systems exhibit most of the important characteristics of photonic crystals, from nontrivial Brillouin zones to topological sensitivity to a minimum index contrast, and can also be used to demonstrate most proposed photonic–crystal devices. The key to understanding photonic crystals in two dimensions is to realize that the fields in 2d can be divided into two polarizations by symmetry: *TM* (transverse magnetic), in which the magnetic field is in the (xy) plane and the electric field is perpendicular (z); and *TE* (transverse electric), in which the electric field is in the plane and the magnetic field is perpendicular.

Corresponding to the polarizations, there are two basic topologies for 2d photonic crystals, as depicted in Figure 4(top): high index rods surrounded by low index (top) and low-index holes in high index (bottom). Here, a hexagonal lattice is used because, as noted earlier, it gives the largest gaps. Recall that a photonic band gap requires that the electric field lines run along thin veins: thus, the rods are best suited to TM light ($\vec{E}$ parallel to the rods), and the holes are best suited to TE light ($\vec{E}$ running around the holes). This preference is reflected in the band diagrams, shown in Figure 4, in which the rods/holes (top/bottom) have a strong TM/TE band gap. For these diagrams, the rod/hole radius is chosen to be $0.2a/0.3a$, where a is the lattice constant (the nearest-neighbor periodicity) and the high/low ε is 12/1. The TM/TE bandgaps are then 47%/28% as a fraction of mid-gap frequency, but these bandgaps require a minimum ε contrast of 1.7/1 and 1.9/1, respectively. Moreover, it is conventional to give the frequencies ω in units of $2\pi c/a$, which is equivalent to a/λ (λ being the vacuum wavelength) – Maxwell's equations are scale-invariant, and the same solutions can be applied to any wavelength simply by choosing the appropriate a. For example, the TM mid-gap ω in these units is 0.36, so if one wanted this to correspond to $\lambda = 1.55\ \mu\text{m}$ one would use $a = 0.36 \cdot 1.55\ \mu\text{m} = 0.56\ \mu\text{m}$.

The Brillouin zone (a hexagon) is shown at left-center, with the irreducible Brillouin zone shaded (following the sixfold symmetry of the crystal); the corners (high symmetry points) of this zone are given canonical names, where Γ always denotes the origin $\vec{k} = 0$, K is the nearest-neighbor direction,

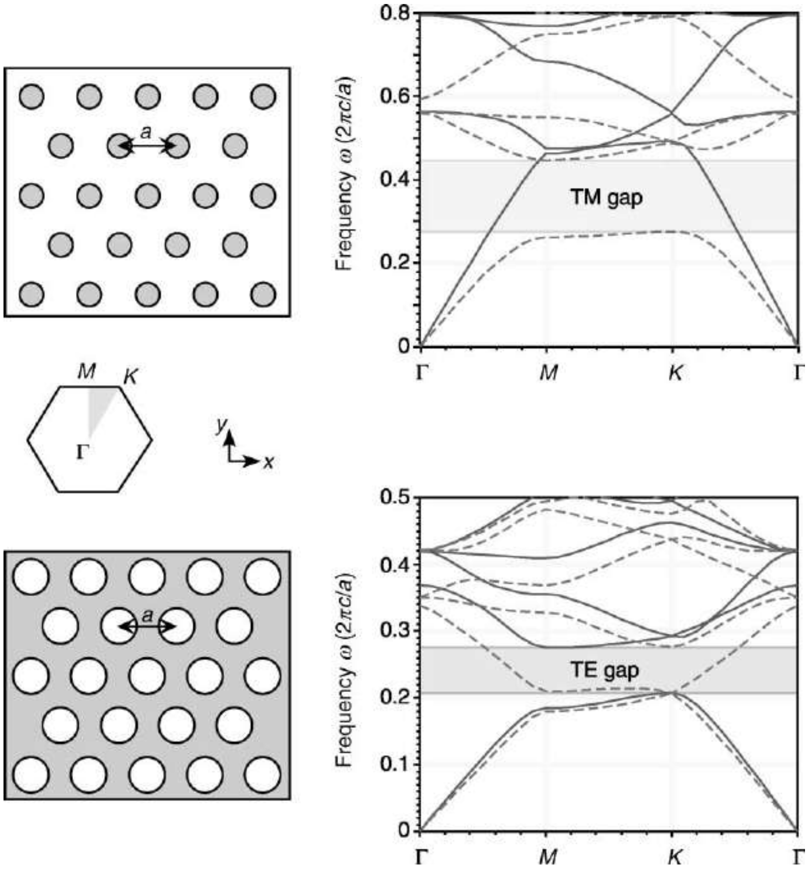

Figure 4 Band diagrams and photonic band gaps for hexagonal lattices of high dielectric rods ($\varepsilon = 12$, $r = 0.2a$) in air (top), and air holes ($r = 0.3a$) in dielectric (bottom), where a is the center–center periodicity. The frequencies are plotted around the boundary of the irreducible Brillouin zone (shaded triangle, left center), with solid/dashed lines denoting TE/TM polarization (electric field parallel/perpendicular to plane of periodicity). The rods/holes have a gap in the TM/TE bands.

and M is the next-nearest-neighbor direction. The Brillouin zone is a two-dimensional region of wavevectors, so the bands $\omega_n(\vec{k})$ are actually surfaces, but in practice the band extrema almost always occur along the boundaries of the irreducible zone (i.e., the high-symmetry directions). So, it is conventional to plot the bands only along these zone boundaries in order to identify the bandgap, as is done in Figure 4.

Actually, the hole lattice can display not only a TE gap, but a complete photonic bandgap (for both polarizations) if the holes are sufficiently large (nearly touching). In this case, the thin veins between nearest-neighbor holes induce a TE gap, while the interstices between triplets of holes form 'rod-like' regions that support a TM gap overlapping the TE gap.

Photonic-Crystal Slabs

In order to realize 2d photonic-crystal phenomena in three dimensions, the most straightforward design is to simply fabricate a 2d-periodic crystal with a finite height: a photonic-crystal slab, as depicted in Figure 5. Such a structure can confine light vertically within the slab via index guiding, a generalization of total internal reflection – this mechanism is the source of several new tradeoffs and behaviors of slab systems compared to their 2d analogs.

The key to index guiding is the fact that the 2d periodicity implies that the 2d Bloch wavevector $\vec{k}_\parallel$ is a conserved quantity, so the projected band structure – all states in the bulk substrate/superstrate (the uniform regions far below/above the slab) versus their in-plane wavevector component (projected wavevector) – creates a map of which states can radiate vertically. If the slab is suspended in air, for example, then the eigensolutions of the bulk air are $\omega = c\sqrt{|\vec{k}_\parallel|^2 + k_\perp^2}$, which when plotted versus $\vec{k}_\parallel$ forms the continuous light cone $\omega \geq c|\vec{k}_\parallel|$, shown as a shaded region in Figure 5. Because the slab has a higher ε (12) than the air (1), and frequency goes as $1/\sqrt{\varepsilon}$, discrete guided bands are 'pulled down' in frequency from this continuum – these bands, lying beneath the light cone, cannot couple to any vertically

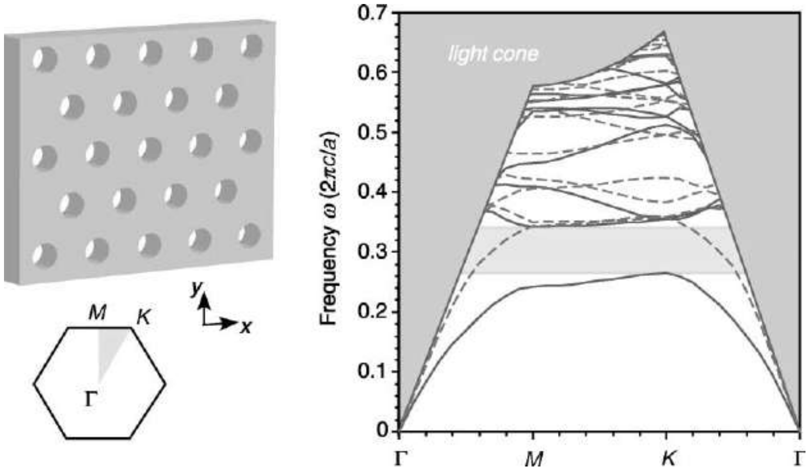

Figure 5 Projected band diagram for a finite-thickness (0.5a) slab of air holes in dielectric (cross section as in Figure 4 bottom), with the irreducible Brillouin zone at lower left. The shaded region is the light cone: the projection of all states that can radiate in the air. Solid/dashed lines denote guided modes (confined to the slab) that are even/odd with respect to the horizontal mirror plane of the slab, whose polarization is TE-like/TM-like, respectively. There is a 'bandgap' (region without guided modes) in the TE-like guided modes only.

radiating mode by the conservation law and so are confined to the slab (exponentially decaying away from it). If the horizontal mid-plane of the slab is a mirror symmetry plane, then just as there were TM and TE states in 2d, here there are two categories of modes: even (TE-like) and odd (TM-like) modes under reflections through the mirror plane (which are purely TE/TM in the mirror plane itself). Because the slab here is based on the 2d hole crystal, which had a TE gap, here there is a 26% 'bandgap' in the even modes: a range of frequencies in which there are no guided modes. It is not a complete photonic bandgap, not only because of the odd modes, but also because there are radiating (light cone) modes at every ω. The presence of these radiating modes means that if all in-plane translational symmetry is broken by a localized change in the structure, say a waveguide bend or a resonant cavity, then vertical radiation losses are inevitable; there are various strategies to minimize the losses to tolerable levels, however. On the other hand, if only one direction of translational symmetry is broken, as in a linear-defect waveguide, ideally lossless guiding can be maintained.

Photonic-crystal slabs have two new critical parameters that influence the existence of a gap. First, it must have vertical mirror symmetry in order that the gaps in the even and odd modes be treatable separately – such symmetry is broken by the presence of a substrate that is not the mirror image of the superstrate, but in practice the symmetry breaking can be weak if the index contrast is sufficiently high (so that the modes are strongly confined in the slab). Second, the height of the slab must not be too small (or the modes will be weakly confined) or too large (or higher-order modes will fill the gap); the optimum height is around half a wavelength λ/n_{eff} (relative to an average/effective index n_{eff} that depends on the polarization). In Figure 5, a height of 0.5a is used, which is near the optimum (with holes of radius 0.3a and $\varepsilon = 12$ as in the previous section).

Three-Dimensional Photonic Crystals

Photonic-crystal slabs are one way of realizing 2d photonic-crystal effects in three dimensions; an example of another way, lifting the sacrifices imposed by the light cone, is depicted in Figure 6. This is a 3d-periodic crystal, formed by an alternating hole-slab/rod-slab sequence in an ABCABC stacking of bilayers – equivalently, it is an fcc (face-centered cubic) lattice of air cylinders in dielectric, stacked and oriented in the 111 direction, where each overlapping layer of cylinders forms a rod/hole bilayer simultaneously. Its band diagram is shown in Figure 6 along the boundaries of its irreducible Brillouin zone (from a truncated octahedron, inset), and this structure has a >21% complete gap ($\Delta\omega$ as a fraction of mid-gap frequency) for $\varepsilon = 12/1$, forbidding light propagation for all wavevectors (directions) and all polarizations. Not only can this crystal confine light perfectly in 3d, but because its layers resemble 2d rod/hole crystals, it turns out that the confined modes created by defects in these layers strongly resemble the TM/TE states created by corresponding defects in two dimensions. One can therefore use this crystal to directly transfer designs from two to three dimensions while retaining omnidirectional confinement. Its fabrication, of

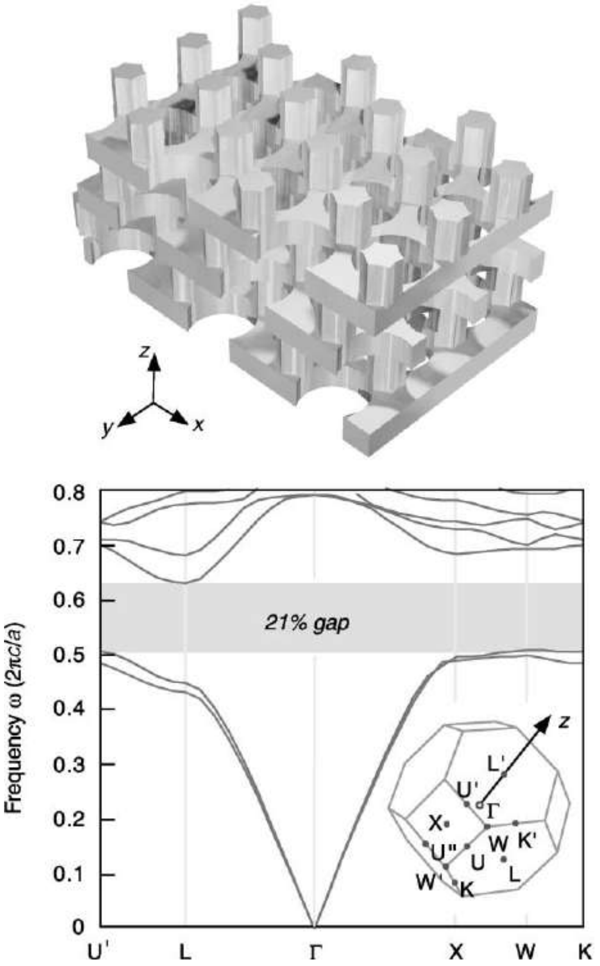

Figure 6 Band diagram (bottom) for 3d-periodic photonic crystal (top) consisting of an alternating stack of rod and hole 2d-periodic slabs (similar to Figure 4), with the corners of the irreducible Brillouin zone labeled in the inset. This structure exhibits a $\Delta\omega/\omega_{\text{midgap}} = 21\%$ omnidirectional bandgap.

course, is more complex than that of photonic-crystal slabs (with a minimum ε contrast of 4/1), but this and other 3d photonic crystal structures have been constructed even at micron (infrared) lengthscales, as described below.

There are three general dielectric topologies that have been identified to support complete 3d gaps for $\varepsilon = 12/1$ (e.g., Si:air) contrast: diamond-like arrangements of high dielectric 'atoms' surrounded by low dielectric, which can lead to $>20\%$ gaps between the 2nd and 3rd bands; fcc 'inverse opal' lattices of nearly-touching low dielectric spheres (or similar) surrounded by high dielectric, giving gaps around 10% between the 9th and 10th bands; and cubic 'scaffold' lattices of rods along the cube edges, giving $\sim$7% gaps between the 2nd and 3rd bands. It is notable that the first two topologies correspond to fcc lattices, which have the most nearly spherical Brillouin zones in accordance with the rules of thumb given above. Many variations on these topologies continue to be proposed – for example, the structure of Figure 6 is diamond/graphite-like – mainly in conjunction with different fabrication strategies, such as the following three successful approaches. First, layer-by-layer fabrication, in which individual crystal layers (typically of constant cross-section) are deposited one-by-one and etched with a 2d pattern via standard lithographic methods (giving fine control over placement of defects, etc.); Figure 6 can be constructed in this fashion (as well as other diamond-like structures with large gaps). Second, colloidal self-assembly, in which small dielectric spheres in a fluid automatically arrange themselves into close-packed (fcc) crystals by surface forces – these crystals can be back-filled with a high-index material, out of which the original spheres are dissolved in order to form inverse–opal crystals with a complete gap. Third, holographic lithography, in which a variety of 3d crystals can be formed by an interference pattern of four laser beams to harden a light-sensitive resin (which is then back-filled and dissolved, as with colloids, to achieve the requisite index contrast). The second and third techniques are notable for their ability to construct large-scale 3d crystals (thousands of periods) in a short time.

See also

Photonic Crystals: Atomic Physics; Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Ashcroft NW and Mermin ND (1976) *Solid State Physics*. Philadelphia: Holt Saunders.

Cohen-Tannoudji C, Din B and Laloë F (1977) *Quantum Mechanics*. Paris: Hermann.

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton: Princeton University Press.

Johnson SG and Joannopoulos JD (2002) *Photonic Crystals: The Road from Theory to Practice*. Boston: Kluwer.

Johnson SG, Ibanescu M, Skorobogatiy M, Weisberg O, Joannopoulos JD and Fink Y (2002) Perturbation theory for Maxwell's equations with shifting material boundaries. *Physical Review E* 65: 066611.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. Berlin: Springer.

Sipe JE (2000) Vector $k \cdot p$ approach for photonic band structures. *Physical Review E* 62: 5672–5677.

Microwave Photonic Crystals

D F Sievenpiper, HRL Laboratories LLC, Malibu, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Photonic Crystals

Photonic crystals are materials that have a periodically varying dielectric constant that results in a bandgap, or range of frequencies where electromagnetic waves are unable to propagate. A simple one-dimensional example is the Bragg mirror, or quarter-wave stack, in which periodic layers of alternating dielectrics reflect photons within a frequency band determined by the thickness and refractive index of the layers. Photonic crystals are the extension of this concept into two or three dimensions. They are constructed as a lattice of two different dielectrics, and examples include a triangular lattice of circular rods, or a diamond lattice of spheres. Certain geometries will produce a bandgap, within which propagation is impossible in any direction. These materials are the photonic analog of semiconductors, in that their periodic dielectric function provides a bandgap for photons in much the same way that the periodic atomic potential of a crystal lattice can provide a bandgap for electrons.

While most photonic crystal research has focused on optical wavelength structures, many of the first experimental studies were performed on scale models at microwave frequencies, where fabrication and measurements were simpler. Operating at microwave frequencies also allowed a new class of materials to emerge, based on metals and lumped circuit elements. The nearly infinite dielectric contrast provided by the use of metals, and the ability to make complex subwavelength scale structures at radio frequencies, allows photonic crystals with ultra-wide bandgaps to be produced. Furthermore, the use of lumped circuit elements allows the electromagnetic properties to be easily tuned, resulting in unique microwave structures.

Wire Mesh Photonic Crystals

At optical frequencies, conductive losses preclude the use of metals for many applications. However, at microwave frequencies, metals are nearly ideal materials, since they have low loss and are easily formed into complicated geometries. The simplest metallic photonic crystal is a three-dimensional wire mesh, shown in Figure 1. This example resembles a diamond crystal lattice, in which metal wires have been substituted for atomic bonds. It can be constructed using zig-zag-shaped metal strips that fit together to form a continuous conductive network.

The electromagnetic properties of this wire mesh can be understood by considering a traditional microwave analog: an array of metal waveguides. Open tunnels permeate the mesh in various directions, and it behaves as an interpenetrating fabric of waveguides, whose walls are defined by the wires. Shorter wavelengths, that fit between the wires, can propagate through the crystal, while longer waves see the mesh as a continuous metal surface, and are reflected. As for a conventional waveguide, the cutoff frequency occurs approximately where one half wavelength fits between the rows of wires. Similarly, for a diamond lattice of wires, the cutoff frequency is

$$\omega_{\text{cutoff}} = \frac{\pi c}{n a} \qquad [1]$$

where n is the refractive index of the material between the wires, and a is the length of the unit cube of the diamond lattice. Because the cutoff frequency is primarily determined by the diameter of the open tunnels, and it is largely independent of the detailed arrangement of the wires, other mesh geometries can

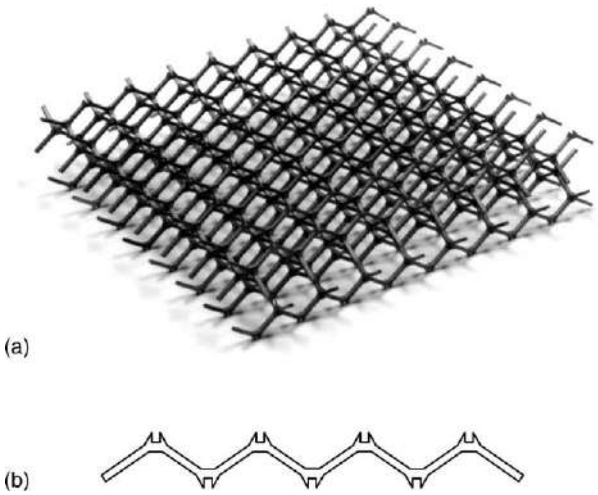

Figure 1 A wire mesh photonic crystal with diamond geometry. (a) The mesh forms continuous conductive structure that is permeated with open tunnels. (b) A diagram of one of the interlocking metal parts that fit together to form the crystal. (Reproduced with permission from Sievenpiper D, Sickmiller M and Yablonovitch E (1996) 3D wire mesh photonic crystals. *Physics Review Letters* 76: 2480. Copyright (1996) by the American Physical Society.)

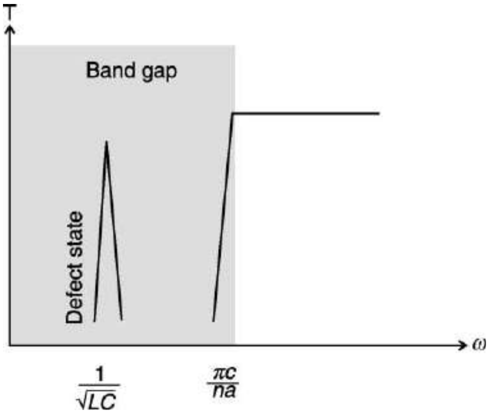

Figure 2 The transmission magnitude through a conductive wire mesh photonic crystal as a function of frequency, ω. Shorter wavelengths can propagate through the open tunnels, while longer wavelengths see the mesh as a continuous block of metal. The cutoff frequency is determined largely by the lattice constant, a, and the refractive index, n, of the material between the wires. A cut wire forms a defect state within the band gap.

be expected to give similar results. However, the cutoff frequency can vary by a constant of order unity, depending on the lattice and the unit cell chosen.

The transmission magnitude, as a function of frequency for such a mesh, is depicted in Figure 2. Experimentally, it is found that a typical wire mesh structure attenuates roughly 6 dB per layer for frequencies below cutoff. However, because the exact transmission magnitude above cutoff depends on details of the measurement, such as impedance matching at the crystal boundary, the vertical scale shown here is arbitrary, and the intent is simply to give the qualitative behavior of these and similar crystal structures. Waves just below the cutoff frequency can penetrate a crystal with finite thickness, as depicted by the finite slope of the transmission edge.

We can better understand the behavior of waves within the crystal by examining the dispersion diagram, shown in Figure 3, in which we plot the allowed frequencies as a function of wave vector $k = 2\pi/\lambda$, the spatial frequency of waves inside the material. A band of allowed frequencies is bounded by upper and lower band edges, corresponding to high and low cutoff frequencies, beyond which propagation is forbidden. The reflective nature of the crystal below cutoff is shown as a bandgap that extends down to zero frequency. The slope of the dispersion curve corresponds to the group velocity of waves within the allowed band, and states at the bandedges represent standing waves. The upper edge of this band occurs where the wavelength matches the periodicity of the crystal, and for a diamond mesh, it is approximately

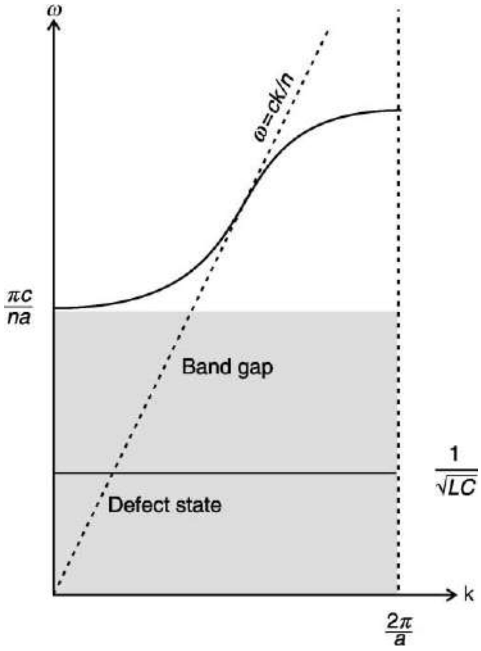

Figure 3 The dispersion diagram for a wire mesh photonic crystal, plotted as frequency, ω, versus wave vector k. The bandgap extends to zero frequency. The lower bandedge of the allowed band is determined largely by the lattice constant. Like other photonic crystals, the wire mesh also supports higher frequency bands, which are not shown. A point defect such as a cut wire creates a mode at a single frequency, which contains all wave vectors.

$$\omega_{\text{lattice}} = \frac{2\pi c}{na} \qquad [2]$$

Like traditional dielectric photonic crystals, these wire mesh structures also have higher order bands and gaps that occur at higher frequencies. The nature of the higher band structure depends on the detailed geometry of the mesh to a much greater extent than the lowest order gap, which is our primary interest for radio frequency materials.

Defect States

As in semiconductor crystals, we can alter the behavior of photonic crystals by adding defects to create localized states within the bandgap. For the wire mesh photonic crystal, a defect may be a cut wire, or a lumped circuit element inserted into the mesh. For example, replacing one of the wires with a capacitor creates a parallel resonant circuit, which acts as a small electromagnetic cavity within the mesh. Its resonance frequency is given by

$$\omega_{\text{defect}} = \frac{1}{\sqrt{LC}} \qquad [3]$$

where C is the capacitance of the defect, and L is the inductance of the surrounding wires. The inductance depends logarithmically on the wire diameter, so it is largely fixed for a given mesh lattice. However, the capacitance can be easily changed, since it is a simple lumped circuit element. Thus, by adding capacitive defects we can create tunable local states within the bandgap. The presence of a defect is seen as a narrow transmission band within the bandgap, shown in Figure 2. Because a point defect contains all wave vectors, it is shown as a horizontal line in the dispersion diagram in Figure 3.

If two defects are located in close proximity, their modes overlap, and they interact with each other. For example, two neighboring defects with identical capacitance will exhibit a split pair of resonances, corresponding to a lower frequency symmetric mode, and a higher frequency anti-symmetric mode. The modes are highly localized at the defects, and this interaction is limited to defects that are within one or two lattice constants of each other. If one member of a pair of nearby defects is tuned with respect to the other, the frequencies the two modes exhibit are an avoided crossing behavior, as is typical in many systems of coupled oscillators. The resonance frequencies for more complex arrangements of defects can be calculated by assigning an interaction energy to each neighboring pair, and solving the resulting Hamiltonian matrix to find the energy levels of the complete system. In this way, the interaction among defects in small clusters is analogous to atoms in simple molecules, and the resulting energy level structure is similar.

As more defects are added periodically throughout the mesh, they form a superlattice, and their states merge into an additional band within the original bandgap. The upper frequency of this band depends on the LC resonance frequency of the defects. The lower frequency depends on the density of the remaining wires, since waves at the lower bandedge must be able to fit through the open regions in the mesh. In the extreme case, where every wire is cut and replaced by a capacitor, the mesh is no longer a continuous conductive network. These structures have an allowed band that extends to zero frequency, and they will be explored in greater detail in the next section. Not every wire must be cut to reach this limit, and the conditions for this to occur are the subject of percolation theory.

Capacitive Mesh Photonic Crystals

In a capacitive mesh photonic crystal, small isolated metal islands are distributed on a lattice, and they are capacitively coupled to their neighbors. This represents the extreme case of the wire mesh photonic crystal, where every wire has been cut, and replaced by a capacitor. An example of such a structure, built as a stack of printed circuit boards, is shown in Figure 4. A diamond lattice can be approximated using planar circuit boards, with printed tetrapodal structures consisting of two orthogonal metal strips connected by a metal plated via. Each metallic island couples to its four nearest neighbors through their overlapping capacitor plates.

The capacitive mesh photonic crystal is transmissive at low frequencies because it is not traversed by conducting paths, like the continuous wire mesh version. The crystal supports a lower band that extends from zero frequency, up to an LC resonance frequency:

$$\omega_{\text{lower}} = \frac{1}{\sqrt{LC}} \qquad [4]$$

which marks the lower edge of the bandgap. This is determined by the capacitance between the metal islands, and the inductance associated with each island. At higher frequencies, the capacitors appear as shorts, and the crystal behaves similarly to the conductive wire mesh structure described previously, with an upper edge given by

$$\omega_{\text{upper}} = \frac{\pi c}{na} \qquad [5]$$

At this frequency, the waves can again fit through the open areas between the metallic islands. The transmission magnitude, as a function of frequency for the capacitive mesh photonic crystal, is shown in Figure 5, and the additional band is illustrated in the dispersion diagram in Figure 6. Like the conductive wire mesh, the capacitive mesh also has a rich upper band structure, that is outside the scope of this article.

With these capacitive mesh structures, it is possible to create materials with finite, but arbitrarily large bandgaps, because the upper and lower edges of the bandgap are controlled independently by the lattice constant and the capacitance, respectively. In conventional dielectric photonic crystals, the width of the bandgap depends on the dielectric contrast between the constituent materials. Waves at the upper edge primarily reside in the lower dielectric material, and waves at the lower edge inhabit the higher dielectric material. This reasoning can also be applied to capacitive mesh structures. The lower dielectric is the substrate material that fills the open spaces between the metal islands, and modes at the upper edge of the gap primarily fill these open spaces. However, for modes at the lower edge of the gap, the electric field is primarily concentrated within the capacitors. The metal regions and the capacitors can

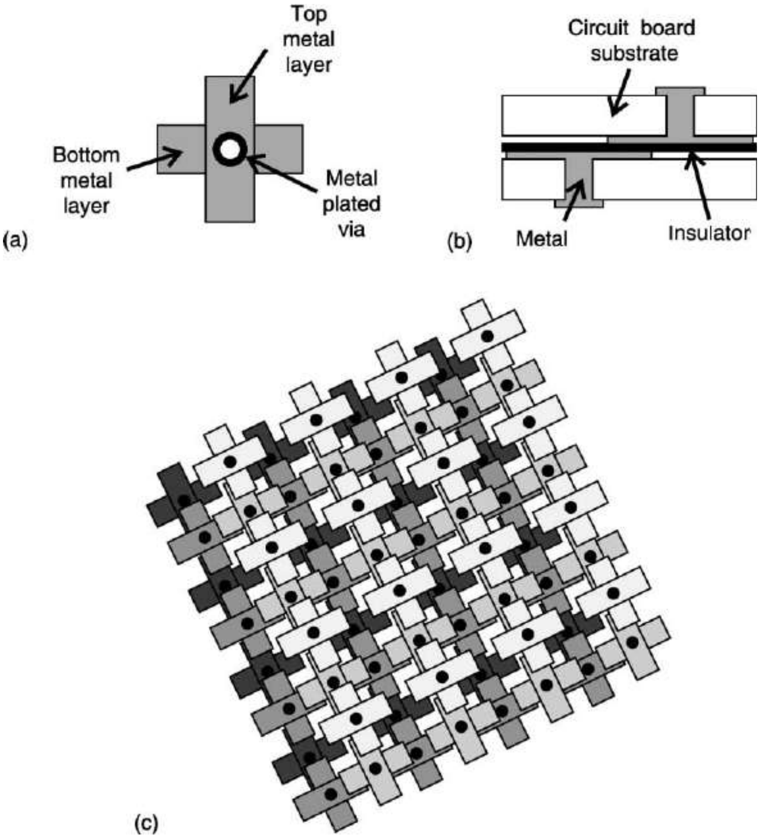

Figure 4 A capacitive mesh photonic crystal, built as a diamond lattice of capacitively coupled metallic islands. It can be built by stacking printed circuit boards. (a) One 'atom' of the crystal, which is made of two small orthogonal metal strips connected by a metal-plated via. (b) A side view, showing how the metal strips overlap to form a capacitor. (c) Several periods of the completed crystal structure, shown without the substrate. (Reproduced with permission from Sievenpiper D, Yablonovitch E, Winn J, et al. (1998) 3D metallo-dielectric photonic crystals with strong capacitive coupling between metallic islands. Physics Review Lett. 80: 2829. Copyright (1998) by the American Physical Society.)

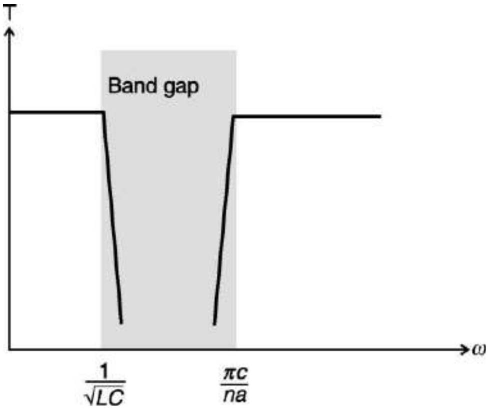

Figure 5 The transmission magnitude through a capacitive mesh photonic crystal, as a function of frequency. The crystal is transparent at low frequencies, because the conducting paths are broken by capacitors. At higher frequencies, the capacitors appear as shorts, and crystal behaves like a solid wire mesh.

be considered as an effective medium, or artificial dielectric whose properties depend on the density, arrangement, and values of the capacitors. By using large capacitors, the dielectric contrast can be made arbitrarily large, and the bandgap arbitrarily wide.

Within the bandgap, these materials can support surface waves, which are modes that are bound to the interface between the crystal and free space. Surface waves propagate along the crystal surface, and they have their own band structure. In the following sections, we examine these waves, and we will introduce periodically textured surfaces that suppress them, much like three-dimensional photonic crystals suppress internal plane waves.

Surface Waves

Electromagnetic waves that are bound to the interface between two materials are known as surface waves. They exist on ordinary metals, and at optical frequencies they are often called surface plasmons. They also exist on photonic crystals within the

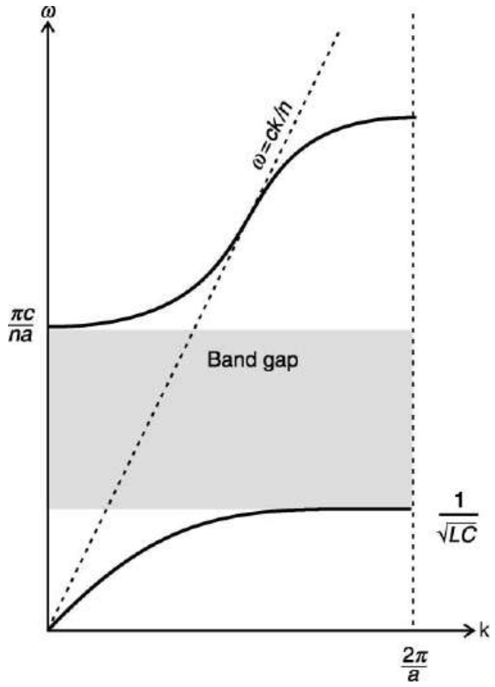

Figure 6 The dispersion diagram of the capacitive mesh photonic crystal. The two edges of the bandgap can be adjusted independently by varying the lattice constant and the capacitance.

bandgap region, as well as on other impenetrable materials. They cannot propagate into the crystal because of Bragg reflection, yet they are prevented from escaping into free space by total internal reflection. They typically interact strongly with the top layer of the crystal, and they can often be manipulated or suppressed by adjusting the geometry of that layer. At radio and microwave frequencies, these waves are usually called surface currents, because the associated fields are only loosely bound to the surface, and extend a great distance into free space. Their presence can be seen in antenna radiation patterns and other phenomena, and by suppressing them we can make a variety of unique radio frequency devices. However, we must first understand the nature of these waves.

To understand the nature of these surface currents, we begin by examining surface waves on metals, where we assume a wave confined to a dielectric interface, and then insert the complex dielectric constant of a metal. Consider a surface in the YZ plane, where $X < 0$ is filled with dielectric ε, and $X > 0$ is filled with vacuum. A wave propagates in the Z direction with propagation constant k. For transverse magnetic (TM) polarized waves, which are illustrated in Figure 7, the transverse electric field $E_y = 0$. The electric field has the form

$$E_1 = (\hat{x}E_{1x} + \hat{z}E_{1z})e^{i\omega x - ikz - \alpha x} \qquad [6]$$

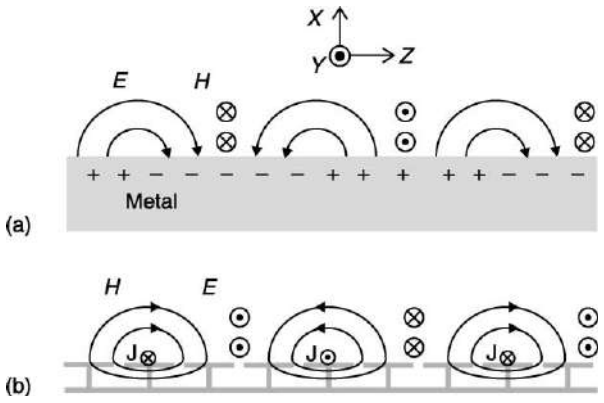

Figure 7 Surface waves can be classified into two polarizations. (a) A TM wave, shown on a metal surface; (b) A TE wave, shown on a textured surface. TM waves consist of electric charges on the metal surface that oscillate along the direction of propagation. They have nonzero electric field components both perpendicular to the surface, and parallel to the direction of propagation. TE waves are associated with transverse currents in the surface. They are similar to TM waves, but with the electric and magnetic fields exchanged.

for $X > 0$ and

$$E_2 = (\hat{x}E_{2x} + \hat{z}E_{2z})e^{i\omega x - ikz + \gamma x} \qquad [7]$$

for $X < 0$, where α and γ are the attenuation constants into free space and dielectric, respectively. The fields of such a wave are shown in Figure 7. We can solve the wave equation for the electric field:

$$\nabla \times \nabla \times \vec{E} = \varepsilon_r \frac{\omega^2}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} \qquad [8]$$

by matching E_x and $\varepsilon_r E_z$, at the interface to yield

$$k = \sqrt{\frac{\varepsilon}{1+\varepsilon}}\,\frac{\omega}{c} \qquad [9]$$

$$\alpha = \sqrt{\frac{-1}{1+\varepsilon}}\,\frac{\omega}{c} \qquad [10]$$

$$\gamma = \sqrt{\frac{-\varepsilon^2}{1+\varepsilon}}\,\frac{\omega}{c} \qquad [11]$$

A solution for transverse electric (TE) waves, where $H_y = 0$, can be found by applying the principle of duality, by exchanging the electric and magnetic fields, and substituting μ for ε.

For surface waves on a metal surface, we insert the dielectric constant of a metal, which is

$$\varepsilon = 1 - \frac{i\sigma}{\omega\varepsilon_0} \qquad [12]$$

The conductivity, λ, is

$$\sigma = \frac{nq^2\tau/m}{1 + i\omega\tau} \qquad [13]$$

where τ is the mean electron collision time, q is the electron charge, m and n are the effective mass and the density, respectively, of the conduction electrons. For frequencies much lower than $1/\tau$, which includes microwave and radio frequencies, we can substitute eqns [12] and [13] into eqn [9] to find

$$k \approx \frac{\omega}{c} \qquad [14]$$

Thus, surface waves on metals propagate at nearly the speed of light in vacuum, and they travel for many wavelengths with little attenuation. We can also solve for the decay constants into free space using eqn [10]:

$$\alpha = (1 - i)\frac{\omega}{c}\sqrt{\frac{\omega\varepsilon_0}{2\sigma}} \qquad [15]$$

and into the metal using eqn [11]:

$$\gamma \approx (1 + i)\sqrt{\frac{\omega\mu_0\sigma}{2}} = \frac{(1 + i)}{\delta} \qquad [16]$$

where δ is recognized as the skin depth. Thus, we have derived the skin depth of a metal from the penetration depth of surface waves.

From the skin depth, we can derive the impedance of a metal sheet. Assuming that E_0 is the electric field at the surface, the surface current is given by

$$J_z(x) = \sigma E_z(x) = \sigma E_0 e^{-x(1+i)/\delta} \qquad [17]$$

The magnetic field at the surface is found by integrating along a path that surrounds the surface current:

$$H_0 = \overset{-\infty}{\underset{0}{\delta}} J_z(x)\mathrm{d}x = \frac{\sigma\delta}{1 + i}E_0 \qquad [18]$$

The surface impedance of the metal is therefore

$$Z_s = \frac{E_z}{H_y} = \frac{1 + i}{\sigma\delta} \qquad [19]$$

We can see that the resistance of a metal surface is accompanied by an equal amount of inductance, because the real and imaginary parts of the surface impedance are equal, and have the same sign.

The part of the wave that extends into vacuum is very loosely bound to the metal surface, but the currents in the metal occur in a well-defined skin depth. For example, copper has a surface impedance of $0.03(1 + i)$ ohms/square at 10 GHz. Using eqns [15] and [16], a surface wave at that frequency extends about 70 meters, or 2300 wavelengths into free space, but less than 1 micron into the metal. For this reason, surface waves are often described simply as surface currents at microwave and radio frequencies, although they are the same phenomenon as the more tightly bound surface plasmons that occur at optical frequencies. Nonetheless, their presence can be seen in a variety of phenomena at RF frequencies, such as antenna radiation patterns.

The above discussion can be generalized to a simple impedance surface, and it can be shown that TM waves are supported on inductive surfaces having an impedance of

$$Z_s(\mathrm{TM}) = \frac{i\alpha}{\omega\varepsilon} \qquad [20]$$

while TE waves are supported on capacitive surfaces, having an impedance of

$$Z_s(\mathrm{TE}) = \frac{-i\omega\mu}{\alpha} \qquad [21]$$

Thus, while ordinary metals, which are inductive, support TM waves, they do not support TE waves, which require a capacitive surface.

Surface waves on metals can be suppressed over a finite frequency band by applying various kinds of periotic textures. For example, at optical frequencies, a lattice of small bumps on a metal surface can create a surface wave bandgap. Surface waves are scattered by the bumps, and form standing waves when one-half wavelength fits between the rows of bumps. Two modes occur, in which the wave nodes lie either on the bumps or between them. These two modes have slightly different frequencies, separated by a narrow bandgap.

At microwave frequencies, more elaborate structures can be built to control the surface properties. One example is a corrugated metal surface, shown in Figure 8. The corrugations can be considered as transmission lines running into the slab, and shorted at the bottom. They are typically one-quarter wavelength deep, so that the short circuit at the bottom is transformed into an open circuit at the top. Because of these resonant structures built into the surface, it has high impedance for waves that are polarized with the electric field transverse to the corrugations, but it has low impedance for waves that are polarized in the orthogonal direction, with the electric field parallel to the metal ridges. For waves propagating transverse to or parallel to the corrugations, the surface can be described as hard or soft, respectively, by analogy to the acoustic surfaces of the same names. Both of these boundary conditions are used in antennas to control scattering properties or

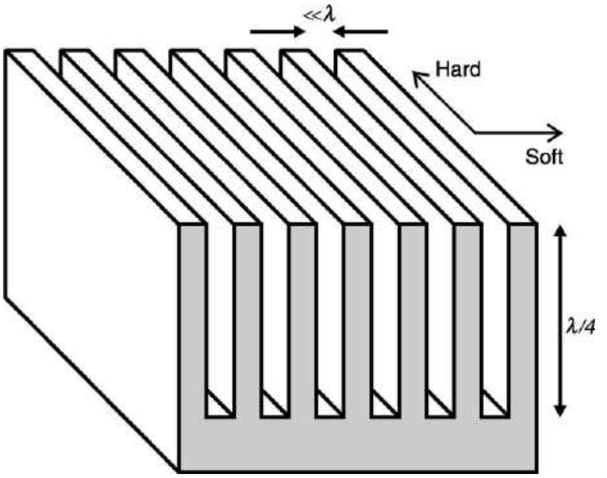

Figure 8 A corrugated metal surface containing quarter-wave resonant grooves presents a soft boundary condition to waves propagating transverse to the grooves, and a hard boundary condition to waves propagating along the grooves.

Table 1 Four basic boundary conditions in electromagnetics, and the surface impedance for each polarization

Boundary condition	*Z_s(TM)*	*Z_s(TE)*
Electric conductor	0	0
Soft surface	∞	0
Hard surface	0	∞
Magnetic conductor	∞	∞

field profiles. These two surfaces, along with the electric and magnetic conductors, describe four basic boundary conditions for electromagnetics, which are summarized in Table 1. Like the soft and hard surfaces, the magnetic conductor does not exist in nature, but it can be produced artificially over a finite frequency band, using a kind of two-dimensional photonic crystal known as a high-impedance surface.

High-Impedance Surfaces

Just as three-dimensional photonic crystals forbid internal plane waves, two-dimensional photonic crystals can be designed to suppress surface waves. Like the capacitive mesh described earlier, they can be constructed with lumped circuit elements. They can be understood as corrugated surfaces, in which the resonant corrugations have been folded up into resonant LC circuits, and distributed on a two-dimensional lattice. An example of such a structure is shown in Figure 9. It can be built using printed circuit boards, in which flat metal patches are connected to a common ground plane by metal plated vias.

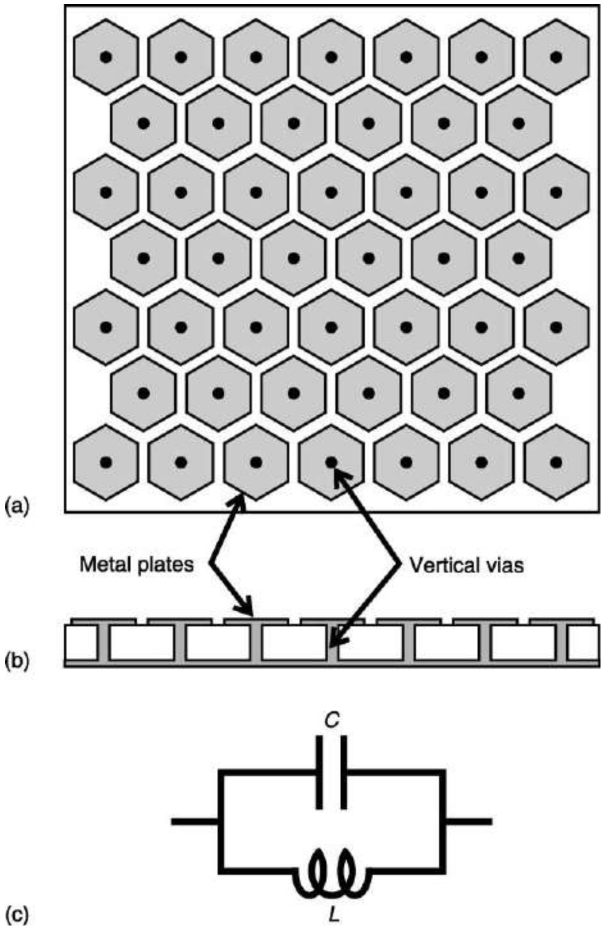

Figure 9 A high-impedance surface, constructed as a printed circuit board. (a) The top view shows an array of hexagonal metal patches; (b) the side view shows metal-plated vias that connect the patches to the ground plane; (c) the effective surface impedance can be modeled as a parallel resonant LC circuit.

When the plates are small compared to the wavelength, we may treat the surface as an effective medium, with a surface impedance defined by a parallel resonant LC circuit. The proximity of the neighboring metal plates provides capacitance, while the current paths, through the vias and the ground plane, provide inductance. The surface impedance is given by

$$Z_s = \frac{i\omega L}{1 - \omega^2 LC} \qquad [22]$$

and the resonance frequency is

$$\omega_0 = \frac{1}{\sqrt{LC}} \qquad [23]$$

The sheet capacitance, C, is equal to the value of the individual capacitors, multiplied by a geometrical factor that is related to their arrangement. It can be shown that for nonmagnetic materials, the sheet inductance L is equal to

$$L = \mu_0 t \tag{24}$$

where t is the total thickness of the high-impedance surface. The surface is inductive for frequencies below resonance, and capacitive for frequencies above resonance. Near the resonance frequency, the magnitude of the surface impedance is very high ($Z_s >> \eta$, where $\eta = 377\Omega$ is the impedance of free space).

For a normally incident wave, the reflection coefficient of the surface is

$$R = \frac{Z_s - \eta}{Z_s + \eta} \tag{25}$$

The phase $\phi = \mathrm{Im}(Ln(R))$ of the reflection coefficient is plotted as a function of frequency in Figure 10. While a low impedance surface such as a conventional electric conductor reflects with a phase shift of π, a high-impedance surface or artificial magnetic conductor reflects with a phase shift of 0. This can be used to build low-profile antennas, because radiating elements can lie very near to a high-impedance ground plane without being shorted out. For example, a thin wire antenna can be placed within a small fraction of a wavelength of the high-impedance surface, yet retain high radiation efficiency. While image currents in a metal ground plane would normally oppose the antenna current, the currents in a high-impedance ground plane reinforce the antenna current because of the opposite reflection phase. This effect occurs when the reflection phase falls between $\pi/2$ and $-\pi/2$, and the corresponding frequency range defines the practical bandwidth of a low-profile antenna on such a surface. Experimentally this bandwidth is found to be roughly equal to the surface wave bandgap, to be described later.

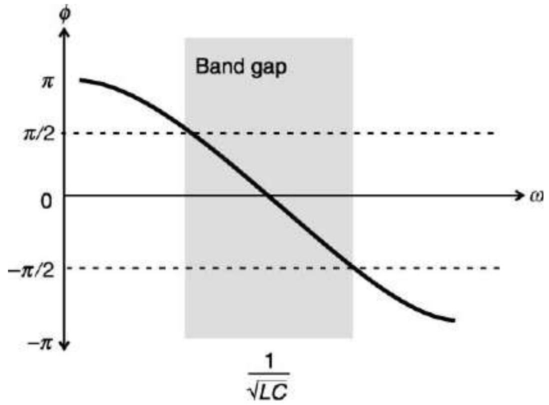

Figure 10 The reflection phase of a high-impedance surface crosses through 0 at the resonance frequency, where the surface appears as an artificial magnetic conductor. The reflection phase falls between $\pi/2$ and $-\pi/2$ for frequencies corresponding roughly to the edges of the surface wave bandgap.

The reflection phase properties of these materials can also be used to build unique microwave reflectors. Since the phase depends on the resonance frequency, one can vary the local resonance frequency in order to tune the reflection phase as a function of position on the surface. For example, a linear phase gradient can be used for beam steering, or a parabolic phase function can be used for focusing. The resonance frequency can be adjusted by tuning the capacitance of the individual unit cells, using varactors or mechanical motion, to build actively controlled microwave reflectors.

The other important property of these textured ground planes is their surface wave bandgap, which can be used to control the radiation pattern of antennas near the surface. We examine the surface wave properties by considering a wave that decays into free space, as in eqn [6]. From Maxwell's equations, we can obtain

$$k^2 = \mu_0 \varepsilon_0 \omega^2 + \alpha^2 \tag{26}$$

which can be combined with eqs [20] and [21] to obtain the dispersion relations for TM and TE waves on a general impedance surface, respectively. For TM waves, which occur where $\omega < \omega_0$, we have

$$k(\mathrm{TM}) = \frac{\omega}{c}\sqrt{1 - \frac{Z_s^2}{\eta^2}} \tag{27}$$

and for TE waves, which occur where $\omega > \omega_0$ we have

$$k(\mathrm{TE}) = \frac{\omega}{c}\sqrt{1 - \frac{\eta^2}{Z_s^2}} \tag{28}$$

These functions describe, to first order, the dispersion of waves on a high-impedance surface. However, they do not predict the bandgap, which can only be obtained by including the effects of radiation damping.

Consider an oscillating sheet of current that sees the impedance of the textured surface on one side, and the impedance of free space on the other side. At very low, or very high frequencies, the current is shunted through the surface. Near the resonance frequency, where the surface impedance is high, power is delivered to free space through radiation damping. The half power bandwidth of this radiation is defined by the frequencies where the magnitude of the surface impedance is equal to the impedance of free space:

$$\left|\frac{i\omega L}{1-\omega^2 LC}\right| = \eta \qquad [29]$$

We solve for λ to yield the frequencies of the two bandedges:

$$\omega^2 = \frac{1}{LC} + \frac{1}{2\eta^2 C^2} \pm \frac{1}{\eta C}\sqrt{\frac{1}{LC} + \frac{1}{4\eta^2 C^2}} \qquad [30]$$

The terms in $1/\eta^2 C^2$ are typically small compared to the terms in $1/LC$, so we neglect them to obtain

$$\omega \approx \omega_0\sqrt{1 \pm \frac{Z_0}{\eta}} \qquad [31]$$

where

$$Z_0 = \sqrt{\frac{L}{C}} \qquad [32]$$

is the characteristic impedance of the surface. For structures that are thin compared to the wavelength, the characteristic impedance is small compared to η. We can expand the square root in eqn [31], to approximate the fractional bandwidth of the radiation as

$$\frac{\Delta\omega}{\omega_0} \approx \frac{Z_0}{\eta} = \frac{\sqrt{L/C}}{\sqrt{\mu_0/\varepsilon_0}} \qquad [33]$$

Using eqn [23], this can be rearranged to obtain

$$\frac{\Delta\omega}{\omega_0} \approx t\frac{2\pi}{\lambda_0} \qquad [34]$$

where λ_0 is the free space wavelength at the resonance frequency. This defines a surface wave bandgap that is centered at the resonance frequency, between a lower-frequency TM band and a higher-frequency TE band. From eqn [34], we see that the bandwidth of a high-impedance surface, or of an antenna on that surface, is limited by its thickness with respect to its free space wavelength at resonance. The term $\lambda_0/2\pi$ is recognized as the radian length, described by Wheeler, who showed that the fractional bandwidth limit of a small resonant antenna is given by its volume with respect to a notional sphere of this radius. Equation [34] represents the equivalent bandwidth limit for a two-dimensional resonant surface.

The surface wave bandgap can be measured experimentally with a pair of small probes placed near the surface. A surface wave probe can be constructed from the end of a coaxial cable with a short section of the inner conductor exposed. Fields from the probe excite surface waves when they are placed near the surface, and the transmission magnitude between the probes can be used to analyze the surface wave band structure. For a high impedance surface, the bandgap can be seen as a range of frequencies with a significant reduction in surface wave transmission magnitude, shown in Figure 11. Within the bandgap, waves are not bound to the surface, and currents on the surface readily radiate energy into free space. The upper edge of the bandgap

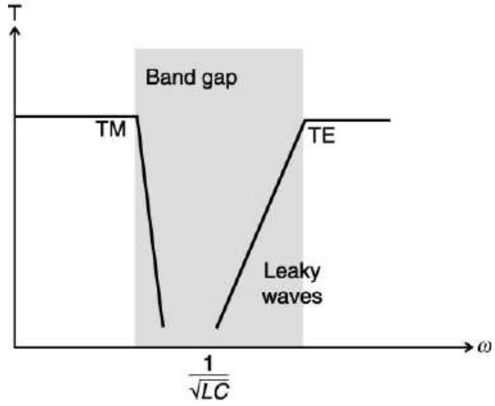

Figure 11 The surface wave transmission magnitude as a function of frequency for a high-impedance surface. TM waves are supported at low frequencies, and TE waves are supported at high frequencies. Leaky TE waves, which are damped by radiation, can occur within the bandgap.

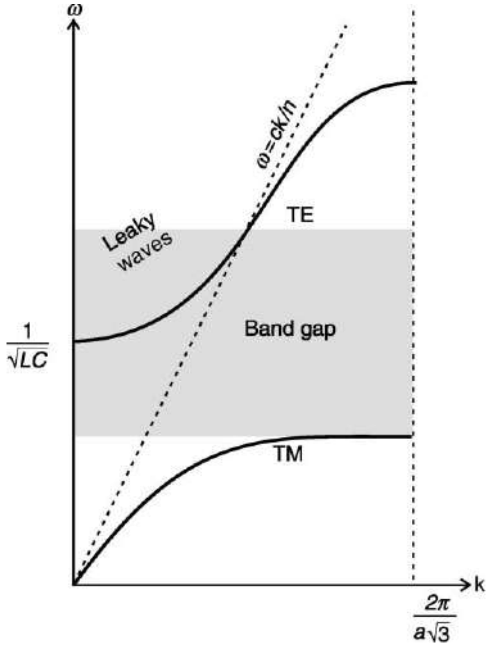

Figure 12 The dispersion diagram for surface waves on a high-impedance surface. The bandgap spans from the upper edge of the TM band, to the point where the TE band crosses the light line. Within the bandgap, electromagnetic waves are not bound to the surface. Modes to the left of the light line represent leaky waves.

is typically less abrupt than the lower edge, due to the presence of leaky TE waves.

The dispersion diagram for surface waves on a high-impedance surface is illustrated in Figure 12. The TM band begins at zero frequency, and bends over flat below the resonance frequency. The TE band begins at the resonance frequency, and slopes upward through the light line. The bandgap, which is the region where bound surface waves are not supported, extends from the upper edge of the TM band to the point where the TE band crosses the light line. This boundary, where $\omega = ck/n$, represents waves in free space propagating parallel to the surface. TE modes, that lie to the right of the light line, are bound to the surface while those that lie left of the light line lose energy to radiation damping. They are not bound to the surface, and are described as leaky waves. The approximate dispersion functions of eqns [27] and [28] are similar to the curves shown in Figure 12, but because they neglect radiation damping, they do not describe these leaky modes. Furthermore, like other photonic crystals, high-impedance surfaces also have higher-order bands and gaps that are not explained by the effective medium model.

The suppression of propagating surface modes can be used for a variety of applications, such as antennas and other microwave structures. One example is to change the radiation pattern of an antenna by covering its ground plane with a textured surface. Figure 13 shows a simple monopole antenna, along with the radiation patterns produced by the monopole when it is placed on a flat or textured ground plane. These plots show the relative magnitude of radiation from the antenna as a function of angle. TM surface currents on the finite metal ground plane form standing waves, which scatter from the edges and corners of a finite ground plane. This causes ripples in the radiation pattern as radiation from the edges interferes with the direct radiation from the antenna. Scattering of the surface currents from the ground plane edges also allows the antenna to radiate power into the backward hemisphere, which would otherwise be shadowed by the ground plane. On the high-impedance ground plane, the suppression of surface

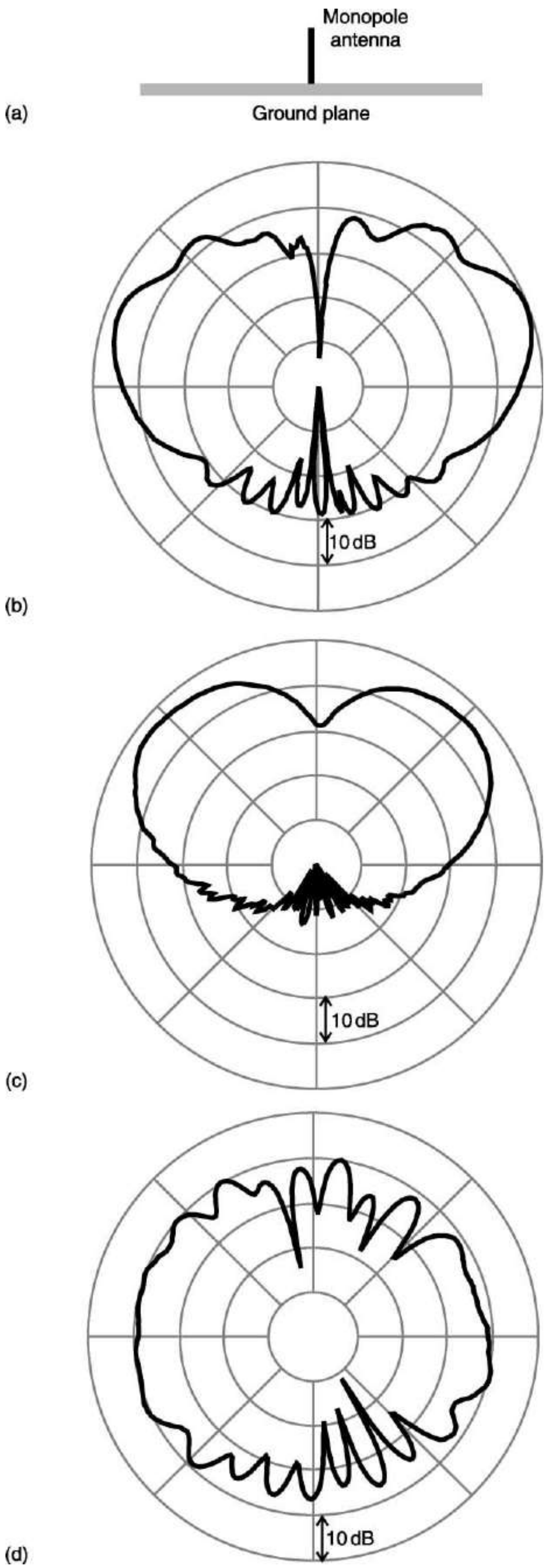

Figure 13 (a) A monopole antenna on a ground plane, used for testing the effect of a textured surface on the radiation pattern. (b) The E-plane radiation pattern of the monopole on a flat metal ground plane shows ripples due to standing surface waves. (c) On a high-impedance ground plane, within the bandgap, surface waves are forbidden, and the radiation pattern is smooth. (d) Outside the bandgap, in the TM band, there is a high density of surface modes.

currents within the bandgap results in a smooth radiation pattern in the forward hemisphere, and very little power in the backward hemisphere. For frequencies outside the bandgap, such as within the TM band, the high density of TM modes can be seen as fine ripples in the pattern, and significant power in the backward hemisphere.

The monopole is useful for illustrating the suppression of TM surface modes, but many other antenna designs are possible. For example, the suppression of both TM and TE surface waves can be used to build patch antennas or aperture antennas with a high degree of symmetry between the *E* and *H* planes. Because these materials behave as artificial magnetic conductors, they can also be used to design a variety of thin antennas that would be impractical on conventional metal ground planes. Finally, by incorporating active elements into the surface, tunable or steerable antennas are possible.

List of Units and Nomenclature

a	lattice constant	meters
c	speed of light in vacuum	meters/second
C	capacitance	Farads
E	electric field	Volts/meter
H	magnetic field	Amps/meter
i	unit imaginary number	–
J	current density	Amps/meter^2
k	wave vector	1/meters
L	inductance	Henrys
m	effective electron mass	kilograms
n	refractive index	–
n	electron density	number/meter^3
q	electron charge	Coulombs
r	radius	meters
R	reflection coefficient	–
t	thickness	meters
T	transmission coefficient	–
Z	surface impedance	Ohms/square
α	decay constant into free space	1/meters
γ	decay constant into material	1/meters
δ	skin depth	meters
ε	electric permittivity	Farads/meter
η	impedance of free space	Ohms
λ	wavelength	meters
μ	magnetic permeability	Henrys/meter
π	diameter/circumference ratio	–
σ	electrical conductivity	1/Ohms
τ	mean electron collision time	seconds
ω	angular frequency	radians/second
∇	vector differential operator	–
∂	scalar differential operator	–
∞	infinity	–

See also

Babinet's Principle. **Diffractive Systems**: Diffractives in Animals; Microstructure Fibers; Omnidirectional Surfaces and Fibers. **Photonic Crystals:** Atomic Physics; Electromagnetic Theory; Nonlinear Optics in Photonic Crystal Fibers; Photonic Crystal Lasers, Cavities and Waveguide; Self-Assembled and Functionalized Photonic Crystals.

Further Reading

Ashcroft N and Mermin N (1976) *Solid State Physics*. Orlando, FL: Saunders College Publishing.

Brillouin L (1953) *Wave Propagation in Periodic Structures: Electric Filters and Crystal Lattices*, 2nd edn. New York: Dover Publications.

Brown E, Parker C and Yablonovitch E (1993) Radiation properties of a planar antenna on a photonic-crystal substrate. *Journal of the Optical Society of America B* 10: 404.

Collin R (1991) *Field Theory of Guided Waves*, 2nd edn. New York: IEEE Press.

Fan S, Villeneuve P and Joannopoulos J (1996) Large omnidirectional band gaps in metallodielectric photonic crystals. *Physics Review B* 54: 11245.

Feynman R, Leighton R and Sands M (1964) *The Feynman Lectures on Physics*, vol. 3. Reading, MA: Addison-Wesley.

Joannopoulos J, Meade R and Winn J (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton, NJ: Princeton University Press.

Kildal P-S (1990) Artificially soft and hard surfaces in electromagnetics. *IEEE Transactions on Antennas and Propagation* 38: 1537.

Kurizki G and Haus J (eds) (1994) Photonic band structures. *Journal of Modern Optics* 41, special issue.

Lee S, Zarrillo G and Law C (1982) Simple formulas for transmission through periodic metal grids or plates. *IEEE Transactions on Antennas and Propagation* 30: 904.

Pendry J, Holden A, Stewart W and Youngs I (1996) Extremely low frequency plasmons in metallic mesostructures. *Physics Review Letters* 76: 4773.

Ramo S, Whinnery J and Van Duzer T (1984) *Fields and Waves in Communication Electronics*, 2nd edn. New York: John Wiley and Sons.

Sievenpiper D (1999) *High-Impedance Electromagnetic Surfaces*. Ph.D. Thesis, Los Angeles: University of California.

Sievenpiper D, Sickmiller M and Yablonovitch E (1996) 3D wire mesh photonic crystals. *Physics Review Letters* 76: 2480.

Sievenpiper D, Yablonovitch E, Winn J, *et al.* (1998) 3D metallo-dielectric photonic crystals with strong capacitive coupling between metallic islands. *Physics Review Letters* 80: 2829.

Sigalas M, Chan C, Ho K-M and Soukoulis C (1995) Metallic photonic band-gap materials. *Physics Review B* 52: 11744.

Smith D, Schultz S, Kroll N, et al. (1994) Experimental and theoretical results for a two-dimensional metal photonic band-gap cavity. *Applied Physics Letters* 65: 645.

Soukoulis C (ed.) (1996) Photonic band gap materials. *Proceedings of the NATO ASI on Photonic Band Gap Materials, Elounda, Crete, Greece, June 18–30, 1995.* The Netherlands: Kluwer Academic Publishers.

Wheeler H (1947) Fundamental limitations of small antennas. *Proceedings of the Institute of Radio Engineers* 35: 1479.

Yablonovitch E (1987) Inhibited spontaneous emission in solid-state physics and electronics. *Physics Review Letters* 58: 2059.

Nonlinear Optics in Photonic Crystal Fibers

J E Sharping, Cornell University, Ithaca, NY, USA

P Kumar, Northwestern University, Evanston, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystal fibers (PCFs) are very similar to normal optical fibers in that they consist of a core surrounded by cladding, such that light is guided within the core of the fiber. The primary difference between PCF and standard optical fibers is that PCFs feature an air–silica cross-section, whereas standard optical fibers have an all-glass cross-section. An electron micrograph of a typical PCF is shown in Figure 1. The air holes extend along the axis of the fiber for its entire length and the core of the fiber is formed by a defect, or missing hole, in the periodic structure. The core is formed of solid glass, whose refractive index is that of pure silica (or whatever other glass is chosen), and the cladding is formed by the air–glass mixture, whose effective refractive index depends on the ratio of air-to-glass, also known as the air-fill fraction, that comprises the structure. The resulting effective-index of the cladding will be lower compared with that of the core and, as such, will provide the refractive index variation necessary to support total internal reflection at the core-cladding boundary, and guide light in a manner similar to that of standard optical fibers. The fiber design (i.e., size, shape, and the air-fill fraction) dictates solutions to Maxwell's equations for light propagating within the fiber. Valid solutions are referred to as 'modes' which propagate along the fiber in a known manner, and have a well-defined shape in the transverse direction (i.e., they have a well-defined transverse mode structure).

Nonlinear-optical effects in fibers result from the interaction of optical fields with the glass via the $\chi^{(3)}$, or Kerr nonlinearity. The phenomenon of nonlinear refractive index is a manifestation of a light–material interaction mediated by $\chi^{(3)}$. The magnitudes of the components of the third-order susceptibility tensor in glass, $\chi^{(3)}$, are generally quite small compared with the analogous second-order ($\chi^{(2)}$) terms for materials exhibiting such nonlinearities (e.g., lithium niobate, beta-barium borate (BBO), etc.). The relatively small $\chi^{(3)}$ nonlinearity in optical fibers makes them ideal for wavelength-division multiplexed optical communication where light propagation subject to a minimum of nonlinear effects is critical. Nonlinearity does, however, eventually become an issue in wavelength-division multiplexed systems as the launched optical power increases and as the channel spacing decreases. On the other hand, one can utilize nonlinear-optical effects in soliton communication systems and to build useful photonic devices. Despite the weak $\chi^{(3)}$

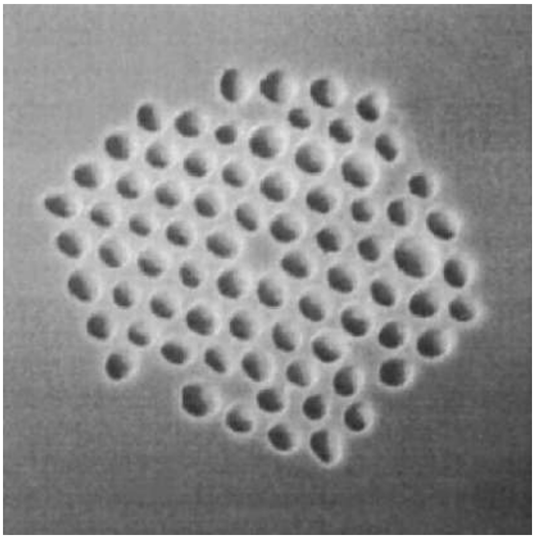

Figure 1 An electron micrograph showing the periodic microstructure of a typical PCF. The core is formed by the 'missing hole' in the center of the microstructure. (Reproduced with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America, courtesy of OFS.)

nonlinearity, the net nonlinear-optical effect in fibers can be large due to the ability to tightly confine intense fields within the core of an optical fiber and maintain the interaction over a long distance as the guided fields propagate through the fiber.

The study of nonlinear-fiber optics has benefited from dramatic improvements in optical fiber and fiber-optic device fabrication. The importance of understanding nonlinear-fiber optics is driven by the need to develop fiber-integrated devices, and also by the need to understand and mitigate the problems that these nonlinearities cause in optical communication systems.

This section introduces the unique linear- and nonlinear-optical properties of PCFs in order to understand the reasons why nonlinear-optical effects are often enhanced in such fibers. These discussions pertain to PCFs which are 'highly nonlinear'. It is essential to clarify that 'highly nonlinear' in this context does not mean that the $\chi^{(3)}$ is any larger than that of standard telecommunication fibers, rather that the effect of this nonlinearity is enhanced due to the fiber's very small core.

PCF Properties

Photonic crystal fibers feature a variety of interesting properties. From the standpoint of nonlinear-fiber optics there are four very useful fundamental properties of PCFs:

- a mechanically robust optical fiber can be fabricated with an extremely small core (a few μm^2);
- a fiber can be made to guide in a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- there are new degrees of freedom that allow one to manipulate the fiber's group-velocity dispersion (GVD) properties; and
- many, but not all, PCFs are polarization maintaining as a result of form birefringence present in the core.

The fact that small-core PCFs can be fabricated is clear from Figure 1 by taking note of the fact that the center defect region which comprises the core is about 1.7 μm in diameter. Photonic crystal fibers with even smaller cores have been fabricated.

Transverse Mode Structure

A widely accepted model used to describe the transverse modal behavior of PCFs is called the effective-index model. The effective-index model can be used to understand why some PCFs are 'endlessly single mode', meaning that the fiber guides in a single transverse mode over an exceptionally wide wavelength range (370 nm–1600 nm). In the effective-index model, the refractive index of the core, $n_{co}(\lambda)$, is that of glass, and the refractive index of the cladding, $n_{eff}(\lambda)$, assumes a value in between that of glass and air. In the context of PCFs, one makes a modification to the standard expression describing single-mode behavior in step-index fibers:

$$V_{eff} = \frac{2\pi\Lambda}{\lambda}\sqrt{n_{co}(\lambda)^2 - n_{eff}(\lambda)^2} < V_{cutoff} \qquad [1]$$

where Λ is the spacing between air holes, λ is the wavelength of light, and V_{cutoff} is the cutoff condition for the PCF. A similar expression for the V parameter is commonly used to understand the modal behavior of standard fibers where the larger the V is, the more transverse modes are supported within the fiber. In standard fibers the cutoff condition below which only a single mode can propagate within the core of a fiber is given by $V_{cutoff} < 2.405$. In the case of PCFs, a numerical method should be used to determine V_{cutoff}. Mechanically robust PCFs can be fabricated where the dispersion in $n_{eff}(\lambda)$ (i.e., the variation of n_{eff} with λ) offsets dispersion in $n_{co}(\lambda)$ and compensates for the $2\pi\Lambda/\lambda$ coefficient in eqn [1]. Therefore, the light within the fiber propagates in a single, Gaussian-like mode because for all wavelengths $V_{eff} < V_{cutoff}$. A graph of the variation of V_{eff} with Λ/λ is shown in Figure 2, where d represents the size of an air hole.

Conceptually, the effective index model can be understood by noting that at short wavelengths the mode field is confined well within the all-silica core, but as λ increases the mode field extends further into the air–glass cladding and V_{eff} and $n_{eff}(\lambda)$ both decrease.

Dispersion in PCF

Other critical differences between PCFs and standard optical fibers lie in the dispersion properties. When light propagates through a fiber its behavior depends on the light's optical frequency:

$$E(t,z) = A(t,z)e^{i[\omega t - \beta(\omega)z]} \qquad [2]$$

Equation [2] describes the mode as it propagates through the fiber. It is decomposed into a slowly varying envelope, $A(t,z)$, and a rapidly varying exponential component where ω is the frequency of the mode, t is time, z is the position along the length of the fiber, and $\beta(\omega)$ is called the mode-propagation constant. The general term used in describing the

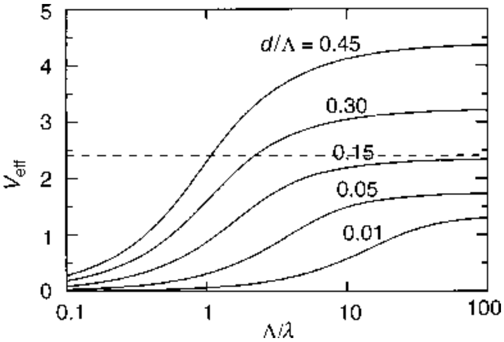

Figure 2 Variation of V_{eff} for different relative hole diameters d/Λ. The calculation assumes a fiber with an air–glass cross-section where the refractive index of air and glass was taken to be 1 and 1.45, respectively. The dashed line marks $V_{eff} = 2.405$, the cutoff value for a step-index fiber. (Reproduced with permission from Birks TA, Knight JC and Russell PStJ (1997) Endlessly single-mode photonic crystal fiber. *Optical Letters* 22: 961–963. ©1997 Optical Society of America.)

frequency dependence of β is chromatic dispersion, which includes contributions from the material as well as the waveguide. Other types of dispersion present in optical fibers include multi-modal (arising from multiple guided transverse modes) and polarization-mode dispersion.

One way to understand the chromatic dispersion of a mode propagating through an optical fiber is to study the Taylor series expansion of the mode-propagation constant, β, about the center frequency of the field, ω_0:

$$\beta(\omega) = \beta_0 + \beta_1(\omega - \omega_0) + \frac{1}{2}\beta_2(\omega - \omega_0)^2 + \cdots \qquad [3]$$

where $\beta_i = \mathrm{d}^i\beta/\mathrm{d}\omega^i$. The physical significance of the various β_i in eqn [3] are as follows: the phase-fronts of the electric field move at a speed given by $\omega/\beta_0 = v_p$, and the envelope, $A(t,z)$, moves at a group velocity given by $1/\beta_1$. A GVD term, which governs temporal spreading of the envelope, is given by β_2. Higher-order β terms are usually negligible for propagation of pulses of ≥1 ps duration in optical fibers and are lumped into the category of 'higher-order chromatic dispersion'.

The notation 'β_2', as defined above, is often used in the literature with dimensions of $\mathrm{ps}^2/\mathrm{km}$. However, another expression is frequently used because of its direct relationship to measured quantities. It is straightforward to measure the relative delay, T, between two pulses having different center wavelengths. Choosing a particular wavelength as a reference, one can then measure relative delay as a function of an injected pulse's center wavelength. The first derivative with respect to λ of the relative delay curve gives the GVD according to

$$D = \frac{\partial\left(\frac{1}{v_g}\right)}{\partial\lambda} = \frac{1}{L}\frac{\mathrm{d}T}{\mathrm{d}\lambda} = -\frac{2\pi c}{\lambda^2}\beta_2 \qquad [4]$$

where v_g is the group velocity, and L is the length of the fiber under test. The dimension commonly used for D is ps/(nm km).

Chromatic dispersion in single-mode optical fibers results from two different wavelength-dependent fiber parameters. The medium itself, glass in this case, has a wavelength-dependent refractive index. This 'material' contribution has the same magnitude regardless of the various parameters associated with the waveguide. A second contribution has to do with the design of the optical fiber. This 'waveguide' contribution to dispersion arises from the fact that the wavelength-dependent mode depends on the properties of the waveguide (i.e., the core size and refractive index contrast between the core and cladding). Empirical models can be used to describe the material, waveguide, and total GVD for standard communication fibers. Such a set of curves is given in Figure 3 where it can be seen that it is possible to have positive, negative, or zero values for D. For historical reasons, the regions where D is negative (β_2 is positive) exhibit 'normal GVD', while those where D is positive (β_2 is negative), exhibit 'anomalous GVD'. The wavelength corresponding to $D = 0$ is referred to as the zero-dispersion wavelength (λ_0), which for most silica glass fibers is about 1,300 nm.

In contrast with standard optical fibers, where the waveguide contribution to D is always less than zero, small-core PCFs can be fabricated where the waveguide contribution to GVD is positive and quite large. As such, in PCFs, λ_0 can be shifted to wavelengths shorter than the intrinsic dispersion zero of glass. Control over the GVD is essential for phase matching certain nonlinear-optical interactions involving light

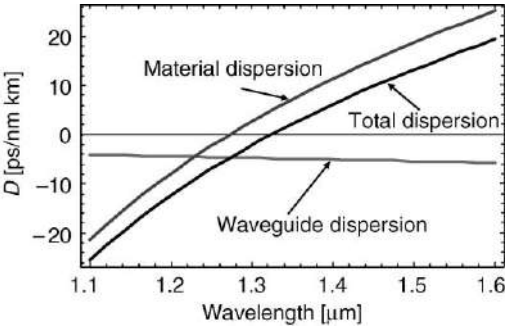

Figure 3 Plots of the theoretical dispersion coefficient, D, as a function of wavelength for a standard optical communication fiber.

of different colors co-propagating within a fiber. Indeed, several exciting applications of nonlinear optics in PCF require a fiber with a $\lambda_0 \sim 800$ nm. The GVD is also of great importance when working with pulsed light in PCF, because GVD results in temporal pulse broadening. It also governs pulse temporal walkoff effects, limiting the effective interaction length between pulses of different colors. This new flexibility to manipulate the GVD curve, by varying waveguide design parameters, is a key advantage associated with using PCFs for nonlinear optics.

Birefringence in PCF

The core of an optical fiber often exhibits some amount of anisotropy. The core may be elliptical in form (shape) which leads to a phenomenon referred to as form birefringence. Since mode propagation depends on the fiber structure, a fiber with an elliptical core will exhibit mode propagation that depends on the electric field's polarization with respect to the axes of the elliptical core. As a result of birefringence, the polarization of the mode varies as it propagates through the fiber (unless care is taken to align the polarization of the injected light with respect to a principal axis of the birefringence).

Polarization-maintaining (PM) fibers are designed to include birefringence in a particular axis of the fiber. By including a well-defined birefringence throughout the length of a fiber that is larger than that induced by external perturbations, fast and slow axes of the optical fiber are created for all guided wavelengths, giving two orthogonal 'polarization modes'. If light is injected into one of the polarization modes (i.e., with its linear polarization along one of the axes) it remains linearly polarized along that axis as it propagates along the fiber. The two polarization modes generally have different group velocities, so pulsed light in each mode will take a different amount of time to propagate through a given segment of fiber. Most PCFs exhibit strong birefringence due to a slightly elliptical core combined with a large core-cladding index difference, and so they behave similarly to PM fibers. Special care must be taken when working with PCF to be sure that the polarization of the light launched into the fiber is aligned with one of the birefringent axes.

In practice, there are a few other features of PCF that are of importance when discussing nonlinear-optic interactions:

- propagation losses are generally larger in PCFs than in standard optical fibers; and
- free-space coupling and splicing are difficult and can result in large coupling losses.

Nonlinear Phenomena

The basic principles determining nonlinear effects in PCFs are the same as those for standard optical fibers (*see* **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics)). It is the new flexibility in PCFs to obtain transverse-modal and GVD behavior different from that of standard optical fibers that makes PCFs truly interesting for nonlinear optics. The relevant nonlinear-optical effects are: self-phase modulation (SPM); cross-phase modulation (CPM); third-harmonic generation (3HG); four-wave mixing (FWM); Raman scattering; and Brillouin scattering.

Self-phase modulation (also known as the optical Kerr effect) refers to the self-induced phase shift experienced by an optical field as it propagates through a fiber. It becomes particularly important for the case of pulses of light propagating through optical fibers. In small core PCFs, SPM is enhanced due to the high-intensity light propagating within the core. Self-phase modulation can lead to substantial spectral broadening of pulsed light propagating along an optical fiber.

When a pulse of light experiences normal GVD (i.e., $D < 0$) as it propagates, the longer-wavelength components travel faster than the shorter-wavelength components. Anomalous GVD (i.e., $D > 0$) leads to the opposite, short-wavelength components traveling faster than the long-wavelength components. Group-velocity dispersion generally leads to temporal broadening of pulses as they propagate along a fiber. Under ideal conditions, however, SPM in combination with anomalous GVD, leads to pulses which propagate without any temporal or spectral broadening. These self-sustaining pulses are called 'optical solitons'.

When waves of light having different wavelengths co-propagate along a fiber, CPM can occur. It can be understood as a phase shift induced on one wave, due to the presence of the other wave. Cross-phase modulation also leads to spectral broadening and solitonic pulse propagation.

In 3HG and FWM, one or more photons are destroyed and others are created. In 3HG, three 'fundamental' photons are destroyed to create one with three times the energy of the fundamental photons. In FWM, two fundamental photons are destroyed while two others are created. While it is straightforward to conserve energy in 3HG and FWM, these interactions must be 'phase-matched', meaning that the interacting waves must be made to propagate in-phase over a meaningful length. Such phase-matching conditions need to be carefully considered when studying 3HG and FWM. Nevertheless, 3HG and FWM can be used to obtain frequency shifts and all-optical amplification. In comparison with other types of optical fibers, PCFs are particularly useful for 3HG and FWM applications. The small core of the PCF allows interactions to occur at much lower input powers, and the new flexibility associated with the GVD properties permits phase matching in cases which are not possible using standard optical fibers. Finally, the fiber's endlessly single-mode behavior permits very good transverse mode overlap between interacting waves having widely different center wavelengths.

Self-phase modulation, CPM, 3HG, and FWM are photon–photon interactions wherein no energy is exchanged with the medium itself. In contrast, Raman scattering and Brillouin scattering result from photon–phonon interactions. The differences between Raman and Brillouin scatterings lie in the energy of the phonons involved and the direction in which the interactions occur. Raman scattering is an interaction between a photon and an optical–phonon mode of the molecules making up the material. In the case of Raman scattering in glass, the energy shift, associated with molecular vibrational (Raman) modes, corresponds to frequencies of 1–12 THz. Raman scattering occurs in the forward and backward directions. With pulsed light, stimulated Raman scattering can occur when the lower-frequency spectrum of the pulse overlaps with the spectrum of the Raman resonances excited by the higher-frequency spectrum. When this happens, energy can be efficiently shifted in spectrum towards the peak of the Raman resonance. In the forward direction, this 'Raman self-frequency shift' builds up. Additionally, if the Raman self-frequency shift occurs in the presence of anomalous dispersion, a 'Raman soliton self-frequency shift' can result. As the injected power is increased, the spectral shift between the injected pulse and the resulting Raman soliton increases. The principal advantage associated with PCF is the ability to generate Raman solitons for a broader range of wavelengths than was previously possible.

For Brillouin scattering, interaction with the acoustical phonons results in frequency shifts of about 10 GHz and the interaction only occurs in the backward direction. Brillouin scattering is generally a nuisance in fiber-based devices, leading to intensity noise and other problems. The interesting feature of Brillouin scattering in PCFs is that the threshold intensity where problems begin to occur is higher for PCFs than for standard optical fibers. The higher threshold permits further optimization of fiber-based devices wherein Brillouin scattering limits the performance.

Experiment Examples

In the following subsections, a selection of experiments, demonstrating a few of the relevant nonlinear-optical phenomena, are briefly described.

Supercontinuum Generation

One of the most exciting demonstrations of nonlinear optics in PCF is that of supercontinuum generation. In a typical experiment, 100 femtosecond pulses from a mode-locked Ti:Sapphire laser operating at a wavelength of 800 nm were injected into the PCF. As the injected power was increased, a broad continuum of spectrum was generated from wavelengths of 400 nm up to 1,600 nm. Typical data showing the input and output optical spectra are given in Figure 4, where it can be seen that well over an octave of frequency spectrum is generated from an input pulse whose spectral width is only about 10 nm.

It is widely accepted that supercontinuum generation results from a combination of linear and nonlinear optical effects conspiring to generate the broad spectrum. As the pulse propagates through the PCF, SPM, FWM, and Raman scattering are all likely to occur with relative efficiencies depending heavily on the input pulse's spectral and temporal characteristics, as well as the fiber's properties. Indeed, efficient supercontinuum generation can occur in PCFs for pump wavelengths lying near the zero-dispersion point in the normal or anomalous dispersion regime of the PCF, and for fibers as short as a few millimeters in length.

Spectral broadening, due to SPM, is common in standard optical fibers, but what makes this

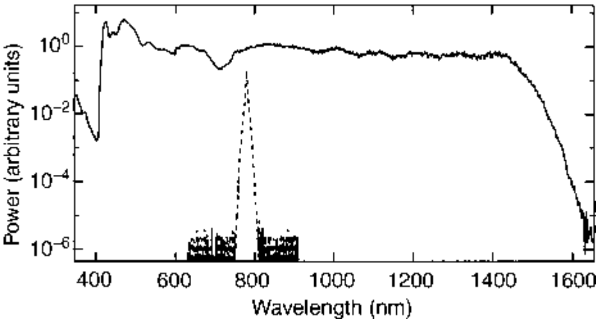

Figure 4 In supercontinuum generation one observes a broad continuum generated after short pulses of light from a Ti:Sapphire laser propagate through a 75 cm section of PCF. The spectrum of the input pulse is shown as a dashed curve, while the output is a solid curve. (Reproduced with permission from Ranka JK, Windeler RS and Strentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America.)

particular experiment interesting is the remarkable width of spectrum generated with a short piece of fiber (~1 m) and with comparatively small optical powers.

Optical Switching in PCF

Cross-phase modulation can be used to build an all-optical switch. Such a switch can be implemented as a three-port device where the output port for a given optical bit (the signal pulse), is determined by the presence of an optical control (the pump pulse). Switching can then be achieved by dividing the signal pulse equally on two arms of an interferometer and injecting the strong pump pulse only on one arm. Because the pump pulse co-propagates with only one of the two signal pulses, there exists a CPM-induced phase difference $\phi_{\mathrm{NL}} = 2\gamma P_{\mathrm{p}} L$ at the output of the interferometer, where P_{p} is the peak power of the pump pulse and L is the interaction length. By varying the intensity of the pump pulses one can vary the magnitude of this phase difference. If a π-phase shift is achieved, one can switch the interference from destructive to constructive, or vice-versa, thus realizing an all-optical switch.

Figure 5 shows an experimental setup used to observe switching near 1,550 nm (a similar apparatus using bulk-optic rather than fiber-optic components can be used to conduct experiments near 780 nm). The pump and signal are synchronous few-ps-duration pulses with a tunable wavelength separation

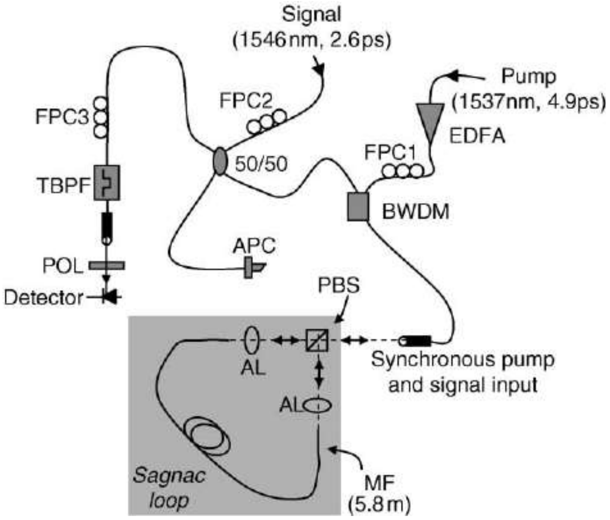
Signal
(1546nm, 2.6ps)
Pump
(1537nm, 4.9ps)
FPC3
FPC2
EDFA
50/50
FPC1
TBPF
BWDM
APC
POL
Detector
PBS
AL
Synchronous pump
and signal input
AL
Sagnac
loop
MF
(5.8m)

Figure 5 Experimental setup used to demonstrate all-optical switching near 1,550 nm. (EDFA, erbium-doped fiber amplifier; BWDM, bandpass wavelength-division multiplexor; PBS, polarization beamsplitter; FPC, fiber polarization controller) (Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE).

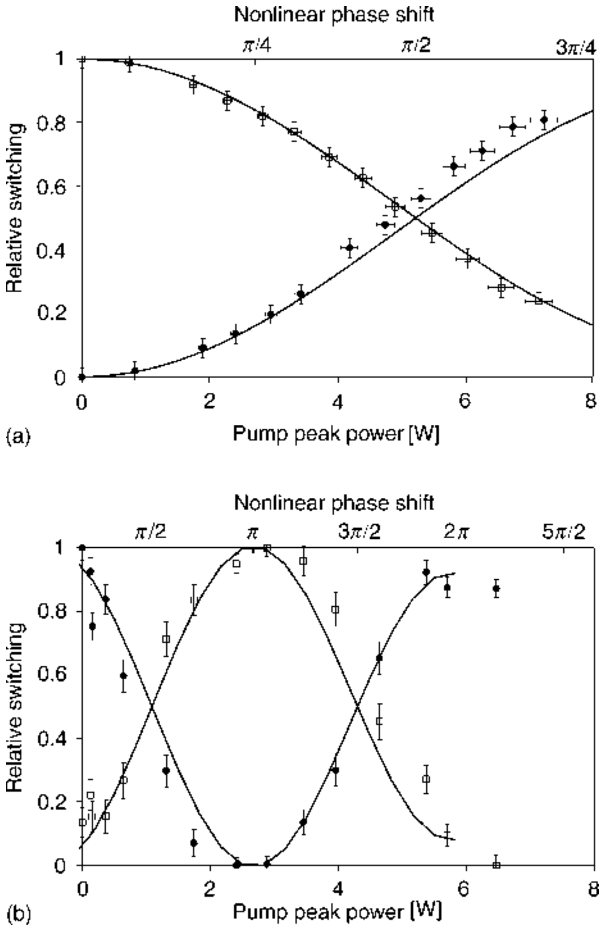

Figure 6 Switching curves (open boxes and filled circles), showing the relative power measured in each port of the switch vs. the pump peak power, for experiments conducted near (a) 1,550 nm and (b) 780 nm. The curves accompanying the data are generated from numerical solutions of coupled wave equations for CPM. (Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE.)

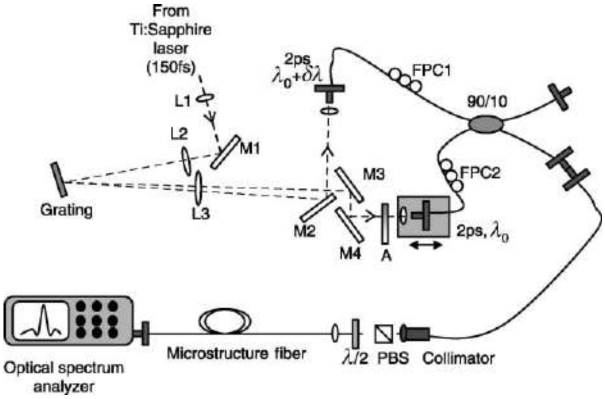
From
Ti:Sapphire
laser
(150fs)
L1
2ps
$\lambda_0+\delta\lambda$
FPC1
90/10
L2
M1
Grating
L3
M3
FPC2
M2
M4
A
2ps, λ_0
Optical spectrum
analyzer
Microstructure fiber
λ/2
PBS
Collimator

Figure 7 A schematic of the experimental setup used to investigate FWM in PCFs. (Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.)

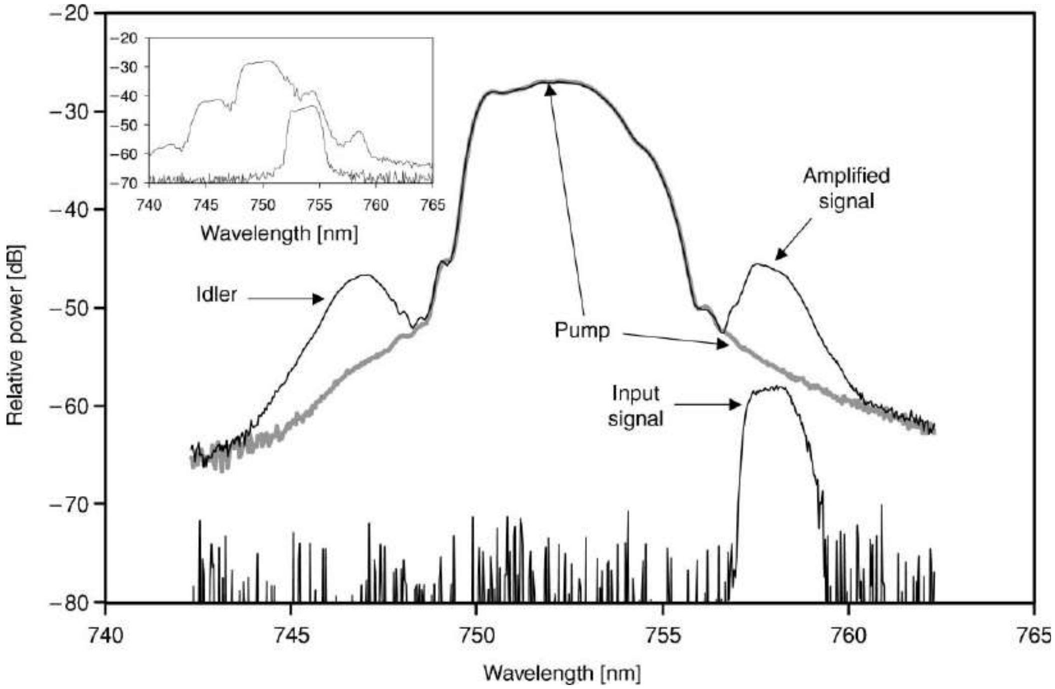

Figure 8 A typical FWM spectrum observed at the output of the microstructure fiber. The inset shows a spectrum where higher-order cascaded mixing is evident. (Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.)

of 5–15 nm. The switching characteristics for this implementation are shown in Figure 6. The apparatus has the advantage of requiring short fiber lengths, low switching powers, and allows switching of weak pulses. It demonstrates the feasibility of using nonlinear optics in PCF to perform essential functions in high-speed all-optical processing.

Parametric (Mixing) Processes

The first set of experiments with controlled FWM in a PCF achieved nondegenerate parametric gains over a 30 nm range of pump wavelengths near the λ_0 of the PCF. The experiments also confirmed the wavelength dependence of the GVD coefficient of the PCF near λ_0. Since the dispersion characteristics of these fibers can be adjusted during the fabrication process, the experiments demonstrate the potential for the use of PCFs in broadband parametric amplifiers, wavelength shifters, and other optical communication devices.

The experimental setup used to demonstrate phase-matched FWM in PCF is shown in Figure 7. The pump and the input signal are two synchronous pulsed beams having 3–5 nm wavelength separation with the center wavelength tunable over a 720–850 nm range. The maximum peak power of the pump pulses is $\simeq$12 W. The two synchronous beams are then combined and injected into the PCF. The pump and signal's optical paths are adjusted to obtain temporal overlap in the PCF and their polarizations are aligned by fiber polarization controllers.

Figure 8 shows a typical FWM optical spectrum at the output of a 6.1 m long PCF. Here the strong pump beam and the weak signal beam have wavelengths of 753 nm and 758 nm, respectively. The spectrum shows the undepleted pump, the amplified signal, and the generated idler at 747 nm. The spectra in Figure 8 show that large gain is achievable for a pump-to-signal spacing of 5 nm. Gain values of more than 20 (13 dB) were obtained.

Conclusion

In summary, the advantages of using photonic crystal fibers for demonstrating nonlinear-fiber optical effects arise from four novel properties:

- the nonlinear coefficient is enhanced in small-core PCFs (core area of a few μm^2);
- PCFs can support a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- PCF design parameters allow one to manipulate the fiber's GVD properties; and
- nonlinear interactions are enhanced due to the polarization maintaining properties of PCFs which result from form birefringence present in the core.

These four properties combine to allow efficient interactions to occur in PCFs which are either inefficient or not possible at all in standard optical fibers.

See also

Optical Amplifiers: Raman, Brillouin and Parametric Amplifiers. **Photonic Crystals:** Electromagnetic Theory. **Scattering:** Raman Scattering.

Further Reading

Agrawal GP (2000) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.

Birks TA, Knight JC and Russell PStJ (1997) Endlessly single-mode photonic crystal fiber. *Optical Letters* 22: 961–963.

Hansryd J, Andrekson PK, Westlund M, Li J and Hedekvist P (2002) Fiber-based optical parametric amplifiers and their applications. *IEEE Journal of Selected Topics in Quantum Electronics* 8: 506–520.

Monro TM, Richardson DJ, Broderick NGR and Bennett PJ (2000) Modeling large air fraction holey optical fibers. *Journal of Lightwave Technology* 18: 50–56.

Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27.

Russell PStJ (2003) Photonic crystal fibers. *Science* 299: 358–362.

Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050.

Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79.

Photonic Crystal Lasers, Cavities and Waveguides

J O'Brien and W Kuang, University of Southern California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic bandstructure engineering, in which the electromagnetic dispersion relations are intentionally modified, can be accomplished by creating one-, two-, and three-dimensionally periodic dielectric materials. These periodic materials, called photonic crystals, can be used to confine and guide light. This article describes the formation of optical resonant cavities and waveguides in these materials. Most of the discussion will focus on two-dimensional photonic crystals. One-dimensionally periodic systems are treated elsewhere in this volume. This article will begin with a short review of the relevant electromagnetic properties of photonic crystals. The formation and properties of resonant cavities will follow this, and the article will conclude with waveguides formed by linear defects in two-dimensional photonic crystals.

Electromagnetics of Photonic Crystals

A two-dimensional photonic crystal consists of a dielectric material in which the dielectric constant is periodic in two dimensions. The period of these materials is on the order of a half-wavelength of the operating optical wavelength. In this section, we briefly review the electromagnetic properties of two-dimensional photonic crystals. For more details, see the article entitled **Spectroscopy:** Raman Spectroscopy in this volume.

There are five Bravais lattices in two dimensions. Most of the research on photonic crystal resonant cavities and waveguides has focused on square and triangular lattices, and in this article we will consider examples based on the triangular lattice. Since these structures are usually defined lithographically however, there is, in principle, no need to confine the investigation to naturally occurring Bravais lattices. Due to the limitations of current nanofabrication technology, much of the research has focused on photonic crystals that have a finite thickness. The important feature of photonic crystals is that the bandgaps in the electromagnetic spectrum opened up by all of the Bragg planes overlap spatially and spectrally, so that a frequency region exists in which no electromagnetic propagation is allowed. These electromagnetic bandgaps can be used to confine optical modes in very small-volume resonant cavities and in waveguides. In high-contrast dielectric systems, such as a semiconductor/air periodic system, only a few lattice periods may be necessary to confine an electromagnetic mode. In the case of finite thickness photonic crystals, this periodicity has the effect of limiting the bandgap formed to the in-plane directions and in the case of a two-dimensional photonic crystal in a high-index dielectric slab to the formation of a bandgap in the guided modes of the slab. In this case there are still radiation modes of the slab.

Consider a two-dimensional photonic crystal that is periodic in the $x-y$ plane. To start, we can consider materials that are uniform and infinitely extended in z. Electromagnetic fields propagating in the $x-y$ plane are Bloch states. These can be written as:

$$\vec{H}_{k_z,k_\parallel}(\vec{\rho}, z, t) = \mathrm{e}^{\mathrm{i}(\omega t-(\vec{k}_\parallel \cdot \vec{\rho}+k_z z))}\vec{u}_{k_z,k_\parallel}(\vec{\rho}) \qquad [1]$$

where $\vec{\rho}$ labels the in-plane coordinates and $\vec{k}_\parallel$ labels the in-plane wavevectors. $\vec{u}_{k_z,k_\parallel}$ is a periodic function in the $x-y$ plane. Modes propagating in the $x-y$ plane in such a system can be classified as transverse electric (TE) waves or transverse magnetic (TM) waves. TE waves have nonzero E_x, E_y, and H_z field components, and TM waves have nonzero E_z, H_x, and H_y field components. A triangular lattice of air holes patterned into a high refractive index dielectric can be used to create a bandgap for both the guided TE and TM modes of the membrane for a range of hole radii. The TE and TM bandgaps that are formed can be over different spectral ranges, however. For laser applications this is unimportant since in practice the emission occurs from electron-hole recombination in a semiconductor quantum well, and for unstrained or compressively strained quantum well materials this emission is TE polarized. In most cases, waveguides have also been designed to work for a single optical polarization. It is generally true that a photonic lattice that consists of a connected high dielectric region is likely to exhibit a TE bandgap, while a lattice formed by disconnected high dielectric regions is more likely to exhibit a TM bandgap. The bandgap in the TE modes is formed between the first and second bands. The first band is called the dielectric band because the field at the Brillouin zone boundary is a standing wave with its intensity concentrated in the high dielectric regions. The second band is called the air band because at the Brillouin zone boundary this field is a standing wave with its intensity localized in the low dielectric regions. Over a reasonable range of lattice parameters, the bandwidth of the TE bandgap can be changed by changing *r*/*a* where *r* is the radius of the holes and *a* is the lattice constant of the triangular lattice. As *r*/*a* gets larger the dielectric band moves up in frequency. This can be thought of as being due to the fact that this mode has a decreasing effective index as *r*/*a* increases. The air band frequency increases as *r*/*a* increases. Since more of the electric field of the air band is located in the low dielectric regions than the field of the dielectric band, the air band increases in frequency with increasing *r*/*a* faster than the dielectric band and the bandgap therefore increases with increasing *r*/*a*. This trend holds in photonic crystals of finite thickness as well.

For photonic crystals of a finite thickness, the ability to simply classify modes as either TE or TM is lost. In the most simple finite thickness photonic crystals, that of a high index dielectric slab in which a two-dimensional photonic crystal has been patterned surrounded above and below by the same low index material which may be air of sapphire, or silicon dioxide, the modes can be classified as being either even or odd modes with respect to the mid-plane of the high index slab. In this mid-plane, however, modes can be classified as being either TE or TM, with each containing the same nonzero vector field components as in the infinite case. Away from this mid-plane, however, modes have in general six nonzero electromagnetic field components. The photonic bandgap that is created in this case is in the guided mode spectrum of the high index dielectric slab. There is no gap formed in the radiation modes of the slab. The radiation modes consist of modes with a ray vector that is not totally internally reflected at the high index slab/low index cladding interfaces. Figure 1 shows the dispersion relation, plotted over the irreducible Brillouin zone, of the even modes for dielectric slab with a refractive index of 3.4 in which a two-dimensional triangular lattice of holes has been patterned. The surrounding material is air. The figure shows an electromagnetic bandgap

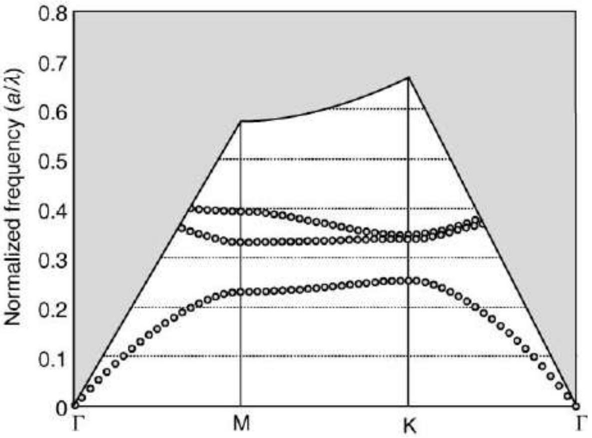

Figure 1 The photonic band diagram for a photonic crystal slab with a triangular lattice of air holes perforating a high-index dielectric slab. The slab, which assumes a refractive index $n = 3.4$, is suspended in the air with a thickness of $d/a = 0.6$. The air holes have radii of $r/a = 0.3$, where a is the lattice constant. Only the lowest three eigenmodes are shown in the plot.

formed between the first and the second guided modes of the slab. Radiation modes of the slab have real valued propagation vectors in the z-direction. These modes lie above the in-plane dispersion relation for air and exist inside the shaded region in the figure. This dividing line between the guided modes and the radiation modes is referred to as the light line.

If the upper and lower cladding layers of the high index slab are different and therefore have different indices of refraction, then the ability to classify modes as being either even or odd about the mid-plane of the slab disappears and it is possible to lose the bandgap in the guided modes altogether. In cases where the refractive index of the bottom cladding is not significantly different from that of the top cladding, such as a case where air serves as the top cladding layer and a material such as sapphire or silicon dioxide forms the bottom cladding, it is found that the even and odd modes describe the field reasonably accurately. To emphasize the approximation involved in the model, these modes are most often referred to in the literature as even-like or odd-like. Overall, there are three primary effects of having asymmetric cladding layers such as an air top cladding and sapphire bottom cladding case. These are a reduction in the effective bandgap width, an increase in the area of the radiation modes on the dispersion relation, and the loss of a rigorous bandgap in the guided modes of the slab. The first two of these effects are the most serious for device designers.

Photonic Crystal Resonant Cavities and Lasers

Photonic crystal resonant cavities can be formed at frequencies inside the bandgap or at the bandedges. Modes in the bandgap are formed as a result of a defect in the lattice. Modes at the bandedge are also sometimes used because the fields with wavevectors at Brillouin zone boundary are standing waves. These standing waves are analogous to the resonant modes that are formed in distributed feedback (DFB) laser structures. Here we will focus on modes confined in the bandgap. Allowed modes in electromagnetic bandgaps in photonic crystal slabs can be introduced by introducing one or more defects into the lattice. By perturbing a single lattice site, we can permit a localized mode or modes that have a frequency in the gap. This phenomenon is similar to the formation of deep levels in an electronic bandgap of a solid due to impurities and identical to the formation of resonant modes in vertical cavity surface emitting lasers (VCSELs) and phase-shifted DFB lasers. In each of these last cases, a defect in the periodicity of a precise thickness occurs between two one-dimensional distributed Bragg reflectors. Defect modes in photonic crystal resonant cavities can be engineered to radiate in a particular direction by trading-off the in-plane versus out-of-plane losses, and the resonant frequency is determined by the parameters of the lattice. Very small mode optical mode volumes can also be obtained in these resonant cavities leading to the possibility of modifying spontaneous emission.

Simple models exist which illustrate the formation of localized modes inside the bandgap as a result of defects in the otherwise periodic lattice. One of these was applied to the formation of deep levels in solids. This model can also be applied to the electromagnetics of defects in periodic structure. For a simple one-dimensional case, assuming that the defect exists at a single unit cell and that the dispersion of the band is dominated by nearest neighbor interactions, this can be solved exactly. Figure 2 shows the allowed frequencies of a ten unit cell chain with a single defect located at the center unit cell. Notice that one mode, called a donor mode, drops out of the band for positive values of $\Delta\varepsilon$ and one mode, called an acceptor mode, rises out of the band for negative values of $\Delta\varepsilon$. This is true in general. A perturbation of the lattice in which the defect has a higher index than the unperturbed lattice will cause a mode to drop out of the air band while a perturbation that has a lower dielectric constant than the background lattice will cause a mode to rise into the bandgap out of the dielectric band. In practice, resonant modes and frequencies in photonic crystals are modeled numerically.

A great deal of variety exists in resonant modes formed by defects in two-dimensional photonic crystal slabs. The simplest examples consist of a single missing hole in a two-dimensional square or

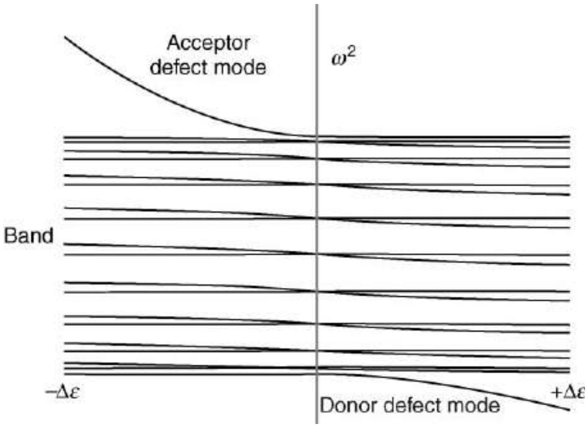

Figure 2 A one-dimensional Slater–Koster model for a single defect in a 10-unit-cell chain, showing the allowed states as a function of the defect perturbation.

triangular lattice patterned into a high-index dielectric slab. However, the resonant mode size, shape, frequency, and polarization can be engineered by tailoring the local dielectric in the region of the resonant mode. Losses in these cavities can be conceptually separated into losses in-plane through the photonic crystal and radiation losses out-of-plane. The in-plane losses are a result of having only a finite number of lattice periods surrounding the localized mode. A finite number of periods results in a finite loss due to tunneling of the fields through the lattice. This loss can in principle be made arbitrarily small by increasing the number of lattice periods surrounding the resonant mode. For a fixed number of lattice periods, the in-plane radiation loss will be reduced as the bandwidth of the photonic bandgap is increased because the magnitude of the in-plane decay constant increases with increasing bandgap width. A resonant mode formed by a defect in a truly three-dimensional photonic crystal will have losses limited by the number of achievable lattice periods and the bandwidth of the photonic crystal bandgap.

Photonic crystal cavities formed by two-dimensional photonic crystals of finite thickness suffer from an out-of-plane radiation loss. Optical confinement in the direction perpendicular to the two-dimensional photonic crystal is due to total internal reflection of the mode that occurs in the high index photonic crystal slab surrounded by lower index materials. Out-of-plane losses are due to the presence of wavevector components in the resonant mode that lie above the light line inside the radiation cone of the cladding. Figure 3 shows the Fourier transform of the magnetic field at the mid-plane of photonic crystal slab in which the resonant mode is formed by a single missing hole in a two-dimensional triangular lattice. In this cavity geometry, a doubly degenerate pair of modes is introduced into the photonic bandgap. The figure shows the Fourier transform of the field for one mode of this double degenerate pair. The boundary of the radiation cone in the figure is marked by the solid circle. Components of the mode inside this circle are not confined to the slab and contribute to out-of-plane radiation losses. These losses can be reduced by engineering the mode so that it has a smaller overlap with the radiation cone. This can be accomplished by allowing the mode to expand in real space in directions in which the wavevectors of the mode extend into the radiation cone. This reduces the spread of the Fourier transform of the mode and reduces the overlap of the mode with the radiation cone.

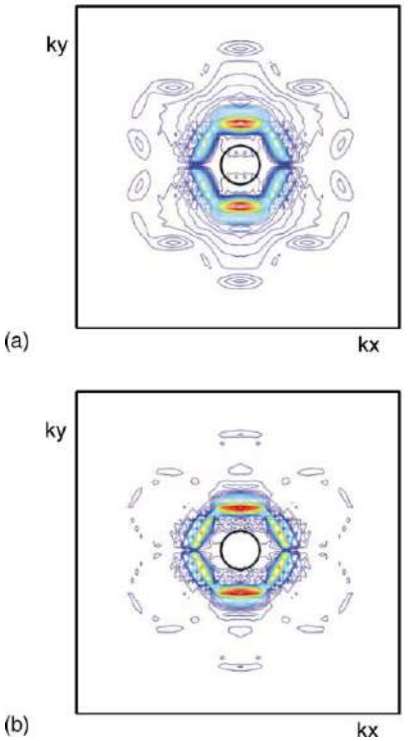

Figure 3 (a) The spatial Fourier transform of the magnetic field component, Hz, of a defect cavity mode at the mid-plane of a triangular-lattice suspended membrane single-defect photonic crystal microcavity. The inner circle indicates the light cone inside which mode components radiate vertically. (b) The spatial Fourier transform of a modified defect cavity mode. The overlap of the mode with light cone has been reduced by modifying the defect cavity.

Symmetry can also be used to reduce the coupling to the radiation fields of the slab. Reducing the resonant frequency of the mode can also reduce out-of-plane radiation losses. This has the effect of reducing the size of the radiation cone in *k*-space. One way to accomplish this is to reduce *r*/*a* in the photonic crystal lattice. This will generally reduce frequencies of all of the photonic crystal modes. It will also, however, generally be accompanied by a reduction in the bandwidth of the photonic bandgap which will increase the in-plane losses. As a result, care must be taken in designing photonic crystal resonant cavities. Nevertheless, in the smallest mode volume photonic crystal resonant cavities, the optical losses will generally be dominated by out-of-plane radiation loss.

Quality factors have been predicted to exceed 100 000 in carefully designed photonic crystal resonant cavities in which the optical mode volume is on the order of a few cubic half-wavelengths in the material. Larger resonant cavities formed by introducing multiple defects in the lattice are often less well confined in the real space lattice leading to a reduction in the components overlapping the radiation cone in *k*-space and a reduction in out-of-plane radiation loss. These cavities also are much more likely to support multiple resonant modes for a given index perturbation.

In analyzing resonance modes of photonic crystal resonant cavities, one of the most important parameters is the quality factor Q of the cold cavity. Practically, a theoretical prediction of the quality factor requires the use of numerical methods and finite-difference time-domain and finite element techniques are commonly used. There are basically three numerical approaches that can be used to calculate the quality factor of the cavity modes. Two of these methods calculate the quality factor from the time domain fields while the third is a frequency domain approach. The first method is to calculate the slope of the exponential decay of the energy of a given cavity mode with time. This decay of energy from the cavity is described by $\exp(-t/t_{\mathrm{ph}})$ where t_{ph} is the photon lifetime which is related to the quality factor Q, by $Q = \omega t_{\mathrm{ph}}$. This method is most useful for relatively low Q modes where the slope of energy decay is visibly greater than zero.

Another method is to calculate the ratio of the full width at half magnitude (FWHM) of the cavity resonance in the frequency domain, $\Delta\omega$, to the center frequency, ω_0. However, distortion to the spectrum is often introduced because the numerical simulation terminates before the impulse response is fully evolved. This has the effect of viewing the true time-domain response through a rectangular window, which translates mathematically into the convolution of the true spectrum with a sinc function. This must be accounted for in the determination of the quality factor.

A third method calculates the ratio of cycle-averaged power absorbed in the boundary to the total energy in the cavity mode. This last method has the advantage of by being able to separate the radiation losses into different directions:

$$\begin{aligned} \frac{1}{Q} &= \frac{\omega P}{U} = \frac{\omega(P_{\perp} + P_{\|})}{U} \\ &= \frac{1}{Q_{\perp}} + \frac{1}{Q_{\|}} \end{aligned} \qquad [2]$$

in which, the effective vertical quality factor $Q_{\perp}$ is given by the ratio of power lost to the absorber at the top and bottom $P_{\perp}$ to the total cavity energy $U(t)$, and the effective in-plane quality factor $Q_{\|}$ is similarly given by the ratio of in-plane power loss $P_{\|}$ to the total cavity energy.

Photonic crystal lasers and resonant cavities have been demonstrated experimentally. The first demonstration of lasing in a two-dimensional photonic crystal laser occurred in a cavity formed by a single defect in a triangular lattice that was patterned into an InGaAsP membrane. The laser operated under optically pumped conditions at 143 K. Room temperature pulsed operation of photonic crystal defect lasers was reported shortly thereafter. Figure 4 shows an electron micrograph of such a resonant cavity. In the figure, the resonant cavity is formed in a suspended membrane in which a triangular lattice photonic crystal with a defect has been patterned. The fabrication of these devices involves pattern definition by electron beam lithography and pattern transfer by a series of wet and dry etching steps. The observed quality factors are, of course, sensitive to fabrication imperfections. Many other cavity designs and demonstrations have also been reported.

To operate a photonic crystal laser continuously at room temperature, it is necessary to design a high-Q resonant cavity that dissipates heat well. Poor heat dissipation is a major drawback of suspended membrane resonant cavities. One strategy for improving the heat dissipation in photonic crystal laser cavities is to form the two-dimensional photonic crystal membrane cavity on top of a low-index high-thermal-conductivity substrate. Figure 5 shows an electron micrograph of a photonic crystal cavity formed by leaving out 37 holes from a triangular lattice patterned into an InGaAsP membrane. This membrane is bonded to a sapphire substrate, which facilitates heat dissipation. This cavity is capable of room temperature continuous wave operation.

It is believed that these lasers are capable of electrically-pumped room temperature operation. This is based on the fact that large quality factors have been demonstrated in optically-pumped photonic crystal lasers and some quality factor reduction due to free carrier absorption and absorption at metal contacts can be tolerated and still lead to lasing. A reasonable assumption would be that the

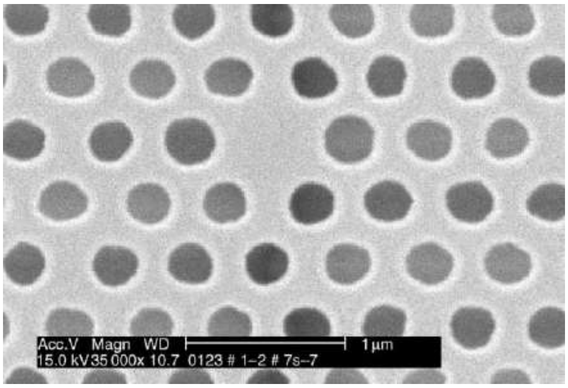

Figure 4 The scanning electron micrograph of a triangular-lattice suspended membrane photonic crystal single defect cavity.

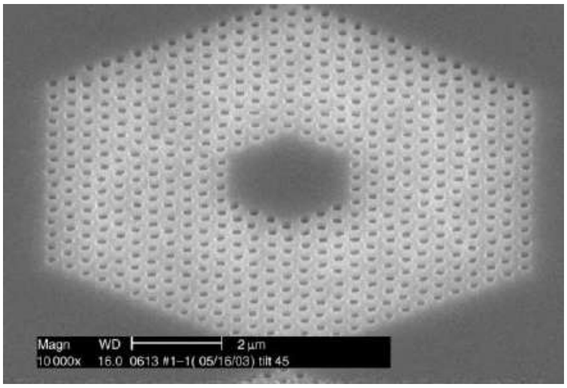

Figure 5 The scanning electron micrograph of a triangular-lattice, 37-missing-hole, photonic crystal defect cavity, bonded on sapphire which serves as lower cladding to improve heat dissipation. CW operation of the laser cavity was demonstrated.

performance of these electrically-pumped photonic crystal lasers would have much in common with the performance of VCSELs. Both VCSELs and photonic crystal cavities have small mode volumes with the cavity formed in some directions by distributed Bragg reflection. The emission direction of photonic crystal lasers can be engineered to be vertical or in-plane, but the basic resonant cavities do have important similarities. The free carrier absorption loss in photonic crystal laser cavities may be expected to be larger than in VCSELs since a photonic crystal laser mode lacks the standing wave behavior in the vertical direction that allows doping at the nodes of the standing wave to reduce the absorption loss. If free carrier absorption loss leads to larger internal losses for photonic crystal lasers than for VCSELs, then some reduction in the achievable slope efficiencies will also occur for photonic crystal lasers.

Photonic Crystal Waveguides

Another important optical element for integrated photonic circuitry is a linear waveguide. Conventional dielectric waveguides confine propagating beams by an index of refraction difference between the waveguide core and the waveguide cladding. Photonic crystal waveguides are most often formed by a linear defect, which consists of a row of modified unit cells of the lattice, patterned into a high index dielectric membrane. The guiding in this case is due to a mixture of total internal reflection at the high index membrane/low index cladding interfaces and distributed reflections from the photonic crystal in-plane. Because this confinement does not exclusively dependent on total internal reflection, the transmission loss through small bending radius waveguide branches and bends can be made very small. Predictions and experimental demonstrations exist which show that photonic crystal waveguides are capable of supporting small bending radius waveguide bends with almost total transmission. This makes photonic crystal waveguides a candidate for waveguides in densely integrated photonic circuits.

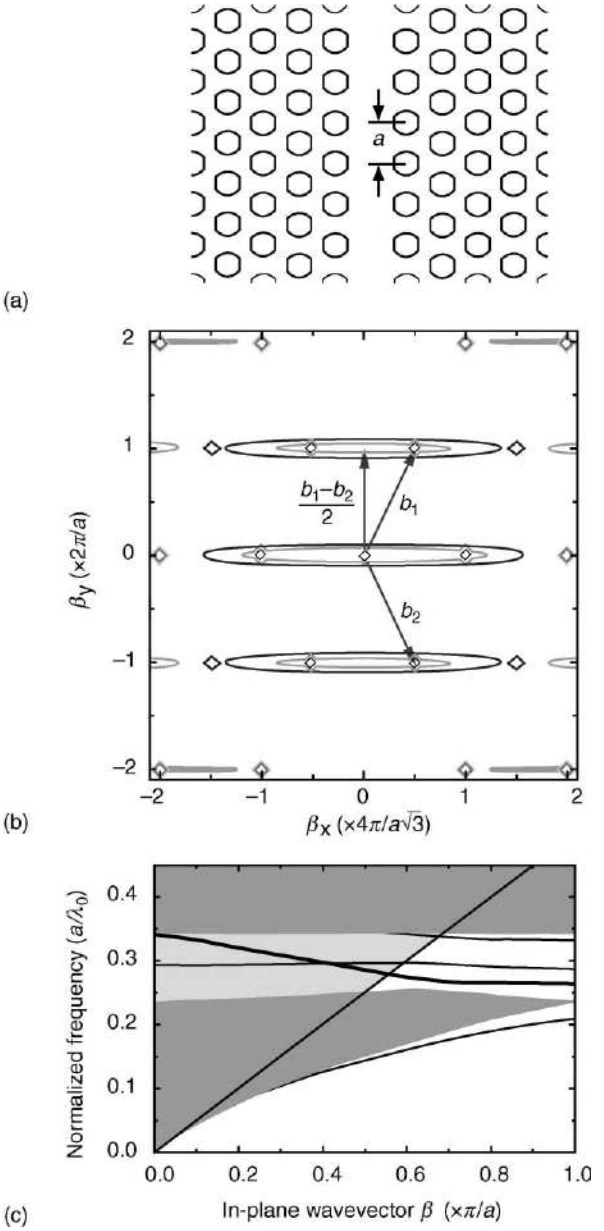

Figure 6 (a) The top view of a triangular lattice photonic crystal single-line defect waveguide with air hole radius $r/a = 0.3$. (b) The 2D spatial Fourier transform of the waveguide dielectric distribution. (c) The photonic band diagram for a suspended membrane single-line photonic crystal waveguide. The membrane has a refractive index of $n = 3.4$ and its thickness is $d/a = 0.6$. The surrounding material is air. The light gray and dark gray area indicate the regimes where field radiates through the air and photonic crystals, respectively.

Figure 6a shows an illustration of the top view of a two-dimensional photonic crystal waveguide.

This is a triangular lattice in which one row of unit cells has been modified. Shown in Figure 6b is the Fourier transform of index of refraction for the situation in which the photonic crystal consists of low index holes in a high index semiconductor. The presence of the linear defects modifies and broadens the reciprocal lattice from a simple reciprocal triangular lattice, before the creation of linear defect, to the distribution shown in the figure. This Fourier transform has components along the direction of propagation at $(\vec{b}_1 - \vec{b}_2)/2$ as shown in the figure where $\vec{b}_1$ and $\vec{b}_2$ are reciprocal lattice unit vectors. These reciprocal lattice components at $(\vec{b}_1 - \vec{b}_2)/2$ couple a planewave with wavevector $\vec{\beta}$ to other planewaves with wavevectors $\beta \pm (\vec{b}_1 - \vec{b}_2)/2$. The result of this coupling is that the electric field of the waveguide mode is a Bloch wave. In a photonic crystal waveguide in which the waveguide axis is the z-direction, the Bloch wavefield in the waveguide can be written in the form:

$$\vec{E}(x,y,z,t) = \sum_n c_n \vec{f}_\beta(x,z) \times \exp\left[i\left(\omega t - \left[\beta + n\left(\frac{\vec{b}_1 - \vec{b}_2}{2} \right) \right] y \right) \right] \qquad [3]$$

In this expression the periodic function of the Bloch wave has been explicitly expanded as a Fourier series and each term in the Fourier series is called a spatial harmonic. Note that these fields differ from that of a photonic crystal fiber because the cladding of a photonic crystal waveguide has a periodicity along the direction of a waveguide. A photonic crystal fiber does not.

Most of the realizations of photonic crystal waveguides occur in two-dimensional photonic crystals formed in high index dielectric slabs. The membrane is usually located a few microns above a high index substrate. Because the fields of the guided modes decay exponentially outside the membrane, the effect of a substrate several microns away is very nearly negligible. The photonic band structure for $r/a = 0.3$, normalized membrane thickness $d/a = 0.6$ and refractive index of $n = 3.4$ is shown in Figure 6c. Only the modes with even symmetry along the mid-plane of the membrane are shown. A bandgap opens for modes with odd symmetry only for large low index filling factors in the triangular lattice. Filling factors large enough to create a bandgap for the odd modes are often impractical so that the odd modes are rarely considered. The even and odd modes are orthogonal, so that plotting the dispersion relation for the even modes only is justified. In reality, even and odd modes may be weakly coupled as a result of fabrication imperfections. The horizontal axis of the dispersion relation covers the in-plane wavevector β along the propagation direction of the irreducible Brillouin zone. The vertical radiation light cone and transverse radiating region of the photonic crystal are mapped as light gray and dark gray areas, respectively. They correspond to the regimes in which light radiates through the air cladding and photonic crystal cladding, respectively. Waveguide dispersion relations can be calculated using two-dimensional approaches in which case the guiding due to the high index slab is accounted for by using an effective index of refraction for the mode. The effective index is the ratio of the propagation coefficient to the free space wavevector of the guided mode of the slab. The two-dimensional

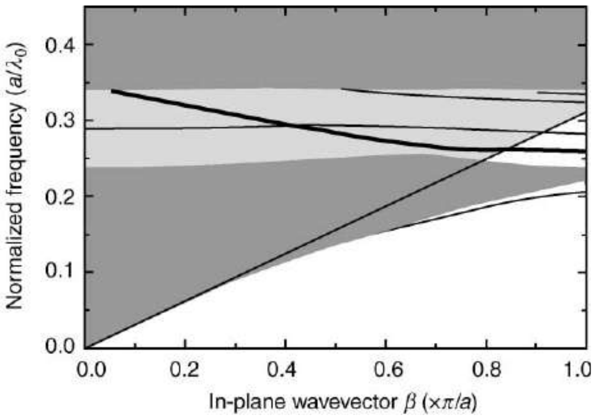

Figure 7 The photonic band diagram for a triangular-lattice sapphire-bonded single-line photonic crystal defect waveguide. The dielectric membrane has a refractive index of $n = 3.4$ and thickness of $d/a = 1.0$. The index of the sapphire bottom cladding is assumed to be 1.6.

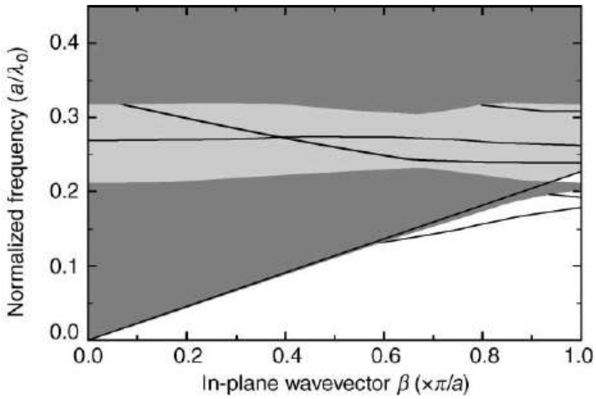

Figure 8 The photonic band diagram for a triangular-lattice deeply etched single-line photonic crystal defect waveguide. The top cladding, guiding membrane, and bottom cladding have an index of refraction of 3.0, 3.4, and 3.0, respectively. Their normalized thicknesses d/a are 1.0, 1.0, and 6.0, respectively.

calculation then uses the effective index of the mode as the high index of material in the photonic crystal. To obtain accurate results, the effective index for each mode calculated separately and the two-dimensional calculation of the photonic crystal waveguide dispersion must be repeated for each waveguide mode.

Examples of photonic crystal waveguide dispersion relations are shown in Figures 7 and 8 for waveguides formed by photonic crystals in dielectric slabs on high index lower cladding layers. Figure 7 is the dispersion relation for an oxide lower cladding layer. The area of the light gray region, which is the projection of the oxide cladding light cone onto waveguide propagation direction, increases as a result of higher index of refraction. Figure 8 shows the dispersion relation for a waveguide with an even larger lower cladding index. In this case the air holes of the two-dimensional photonic crystal are patterned all the

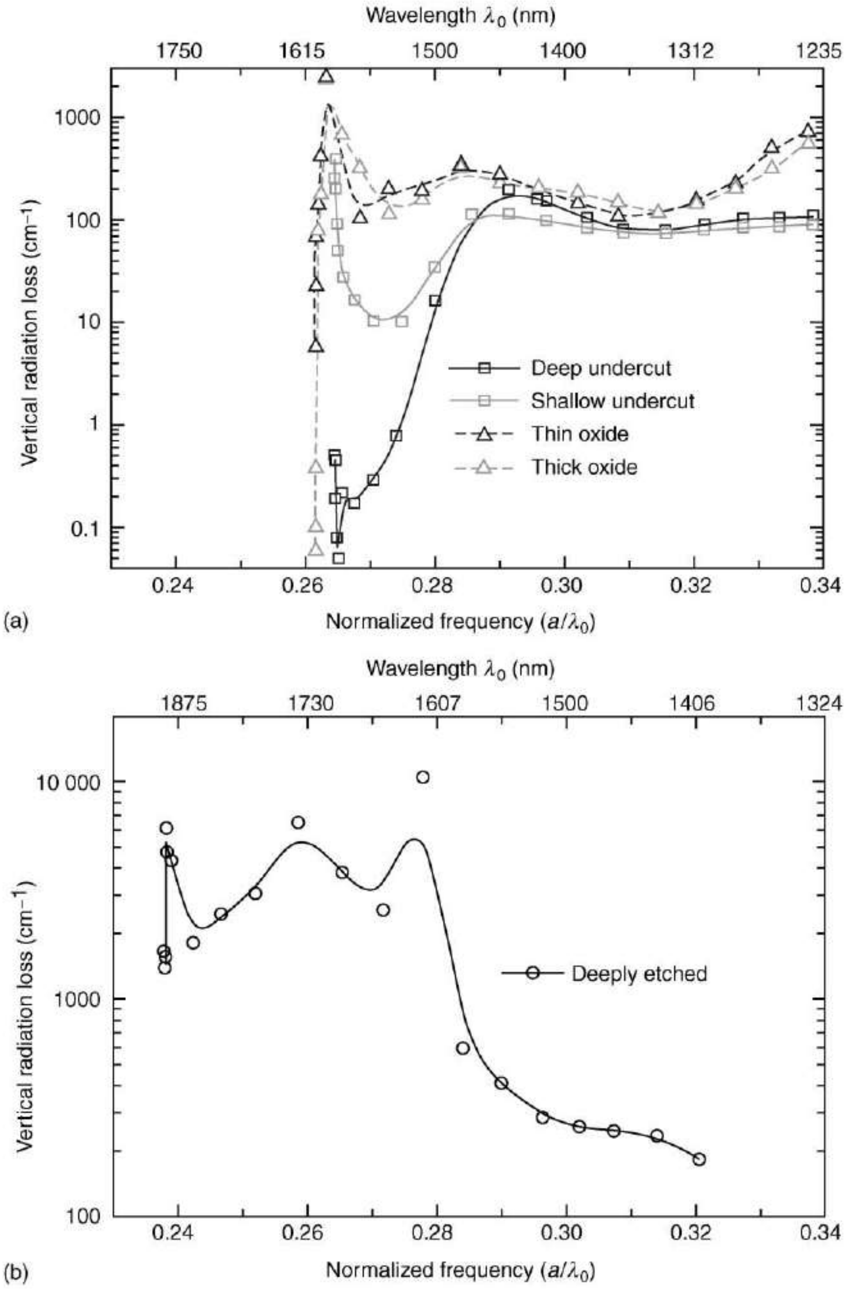

Figure 9 The out-of-plane radiation loss as a function of normalized frequency for the photonic crystal defect slab waveguides illustrated in Figures 6–8. The lattice constant $a = 450$ nm for deeply etched waveguide and 420 nm for the rest of the cases. Points are calculated values and lines are B-spline curve fits.

Table 1 Photonic crystal defect slab waveguides considered in the calculations

Waveguide structures	*Suspended membrane*		*Oxidized lower cladding*		*Deeply etched*
	Deep undercut	*Shallow undercut*	*Thin oxide*	*Thick oxide*	
Layer description: Material (normalized thickness *d*/*a*)			Air (>3.0)		
	–	–	–	–	AlGaAs (1.0)
	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (1.0)
	Air (3.0)	Air (1.0)	Al_xO_y (2.0)	Al_xO_y (5.0)	AlGaAs (6.0)
			GaAs substrate (>3.0)		

Refractive index of 3.4 is assumed for GaAs, 3.0 for AlGaAs, 1.6 for Al_xO_y, and 1.0 for air.

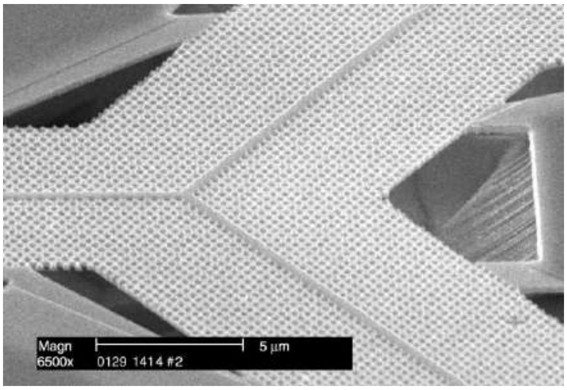

Figure 10 The scanning electron micrograph of a photonic crystal Y-branch fabricated on InP material system.

way through the lower cladding layer, with an index of 3.0, on a high index substrate. The defect modes lying within the in-plane photonic bandgap are all above the cladding light line in this case. Propagation losses are predicted to increase by over two orders of magnitude from the structure with very low index cladding layers shown in Figure 6 to the high index lower cladding structure of Figure 8.

Figure 9a and b shows the results of a fully three-dimensional calculation for the propagation loss of the three structure shown in Figures 6–8. The details of these structures are shown in Table 1. This calculated loss is due to coupling between the guided mode and the radiation modes of the high index substrate. A wavelength scale is also included using $a = 420$ nm in Figure 9a and 450 nm in Figure 9b. The suspended membrane suffers the least vertical radiation loss among all of the waveguides considered and has a minimum loss around 0.2 cm^{-1}. Low-loss waveguides formed on high index cladding layers may be possible, but care should be taken to eliminate coupling between the guided modes and the radiation modes of the cladding. In this case, the design strategy for reducing the optical loss is to reduce the magnitude of the Fourier component of the electromagnetic field inside the light cone. It should also be noted that deeply-etched waveguides formed by removing multiple rows of holes have been demonstrated with significantly lower radiation loss than is predicted for the single line defect waveguides. However, these waveguides are multi-moded at all frequencies. It is also important to remember that the nanofabricated waveguides will suffer additional losses due to fabrication imperfections that were not included in the numerical model. Figure 10 shows a nanofabricated photonic crystal waveguide. The image is an electron micrograph of a y-branch component of a Mach–Zehnder interferometer formed in a two-dimensional photonic crystal waveguide patterned into a suspended InGaAsP membrane.

Finally, it is worth noting that the group velocity of photonic crystal waveguide modes, where the group velocity is given by the gradient of the dispersion surface:

$$v_{\mathrm{g}} = \nabla_{\vec{\beta}}\,\omega \qquad [4]$$

is very small near the zone boundary. This can be seen from the dispersion relations shown in Figures 6–8. In fact, the slope of the dispersion relations is likely to be a parameter which can be engineered by careful waveguide design. This property may allow photonic crystal waveguides to find applications in optical processing functions such as delay lines.

See also

Photonic Crystals: Atomic Physics; Electromagnetic Theory.

Further Reading

Baba T (1997) Photonic crystals and microdisk cavities based on GaInAsP–InP system. *IEEE Journal of Selected Topics in Quantum Electronics* 3: 808–830.

Joannopoulos JD, Meade RD and Winn JD (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton: Princeton University Press.

Mekis A, Chen JC, Kurland I, Fan S, Villeneuve PR and Joannopalos JD (1996) High transmission through sharp bends in photonic crystal waveguides. *Physics Review Letters* 77: 3787–3790.

Painter O, Lee RK, Yariv A, *et al.* (1999) Two-dimensional photonic crystal defect laser. *Science* 284: 1819–1821.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. New York: Springer.

Slater JC (1967). *Insulators, Semiconductors, and Metals, Vol. 3*. New York: McGraw-Hill Book Company.

Srinivasan K and Painter O (2002) Momentum space design of high-Q photonic crystal optical cavities. *Optics Express* 10: 670–684.

Villeneuve PR, Fan S and Joannopoulos JD (1996) Microcavities in photonic crystals: mode symmetry, tenability, and coupling efficiency. *Physics Review B* 54: 7837–7842.

Yariv A and Yeh P (1984) *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: John Wiley & Sons.

Self-Assembled and Functionalized Photonic Crystals

S Bhandarkar, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystals mimic the role of the atomic structure of a semiconductor in manipulating electrons, in this case for light. Hence, called semiconductors of light, these crystals are not generally found in nature and have to be designed and fabricated. Research and development in this field is taking place at a feverish pace to both understand the various nuances of this concept, as well as to evolve new methods to fabricate such crystals in the laboratory and ultimately on a commercial scale. The driving force behind this is the tremendous potential of these systems. They promise to revolutionize the way light can be used for future devices: from forming perfect, omnidirectional mirrors to the smallest ever lasers, and ultimately, complex optical chips with a wide range of functionality arranged in three-dimensional (3D) circuits.

This article seeks to elaborate on a specific technique to make photonic crystals called colloidal self assembly. Here, we will review the fundamental concepts underlying photonic crystals and see how the above process can lead to a versatile platform in making such systems. We will discuss the basic aspects of this method and look at the various innovative manifestations that have been used to fabricate many types of photonic bandgap materials.

Photonic Bandgap

A key attribute of a semiconductor is the formation of a bandgap, a band of energy states that the electrons cannot acquire. They may exist in the so-called valence band where their energy levels are less than that of the bandgap. Or they may be energized to states above the bandgap, i.e., to the conduction band, but the bandgap is the forbidden region. The continuous band of states that constitutes this bandgap is a result of the long-range order imposed by the crystal structure of the semiconductor, where the atomic lattice presents a periodic potential to a propagating electron. The tremendous usefulness of semiconductors comes from their ability to perform switching and logic functions, which in turn is a result of controlling the availability of electrons above and below the bandgap.

The propagation of light as an electromagnetic wave is influenced by the dielectric properties of the medium. For a waveguide aligned in the z direction, the field is the sum of the two propagating transverse modes, each with the form $E(x) = \Phi(x, y)e^{ik_z z}$, where $\Phi(x, y)$ is the field distribution and k_z is the propagation constant. The wave equation for the transverse mode reduces to

$$\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Phi(x, y) + n(x, y)^2\Phi(x, y) = (\bar{n})^2\Phi(x, y) \qquad [1]$$

Here, $\bar{n}$ is referred to as the effective refractive index and is given by $\bar{n} = k_z\lambda/2\pi$.

Equation [1] has the same form as the Schrödinger equation, which describes the wave functions of quantum particles such as electrons. This equivalence helps us to understand the similarity between the attractive potential well structures that bind electrons to the wells created by a local increase in the refractive index that confines light into optical modes.

The creation of a structure with a 3D periodic change in the refractive index (Figure 1) opens a gap, due to Bragg scattering off the dielectric interfaces whereby a set of frequencies of light becomes forbidden to propagate in any direction – a photonic bandgap. To understand the formation of the

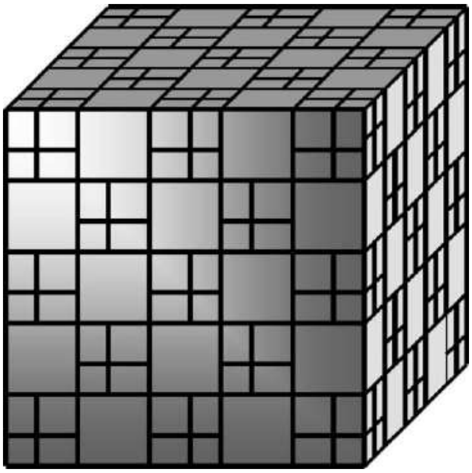

Figure 1 Schematic representation of a three-dimensional photonic crystal. The structure formed is a periodic variation in the dielectric properties of the matrix. Here, the two different dielectric regions are represented by □ and ⊞.

bandgap, one needs to consider the complex pattern of overlapping beams that results from the refractions and reflections from the numerous interfaces. This pattern is determined by the characteristics of lightwaves (wavelength and direction of travel) and those of the periodic dielectric variation (pitch and magnitude of the contrast). Thus, for a select band of wavelengths, complete cancellation through destructive interference occurs in all directions.

One way to breach this restriction is by creating a defect or a disturbance in the periodicity, which would allow localized photonic states in the gap. This defect can be a point defect to form a microcavity, or have a 1D trace that behaves as a waveguide or a planar defect. Thus, a photonic crystal with a specific defect structure can be used to manipulate light in unique ways that can be used for a variety of applications, including lasers, filters, switches, waveguides, and for 3D integrated optics circuits. These structures are also interesting because they have been predicted to exhibit enhanced optical nonlinearities.

Photonic Crystals

The lattice constant for the photonic crystal must be comparable to the wavelength of light, typically in the range of 0.5 to 2 μm. This range is three orders of magnitude greater than the lattice constants found in solids and hence photonic crystals need to be synthesized. A common example of such a structure is a diffraction grating in a waveguide, which serves as a 1D photonic crystal. There are also examples of photonics crystal structures found in nature, such as the iridescent colors seen in certain butterflies and opal gemstones. The structure of opal is a 3D photonic crystal consisting of small spherical particles of hydrated silica (SiO_2). These spheres tend to be irregular in size but in rare precious opals, they are monosized. Diffraction from these spheres leads to the remarkable colors seen in opals.

Fabrication of 3D photonic crystal structures, with two materials with sufficiently different dielectric properties, has been extensively investigated over the past several years. Given the requirement of a tightly controlled periodicity, one obvious means of doing this is to use the microlithographic patterning processes that are so well developed for very large-scale integration (VLSI). Since these are planar and not volumetric techniques, several iterations of layer-by-layer deposition, lithography, and etching are necessary. A popular second method is that of colloidal self assembly.

The Self-Assembly Technique

The term ‘self assembly’ is usually used to describe the act of spontaneous and controlled agglomeration of some basic moieties. These basic units could be molecules, supramolecules, nanoclusters, or particles. It is similar to what is seen in living systems, where organic and inorganic materials are assembled into useful configurations mediated by various proteins. Since there is no intrinsic restriction on the number of such building blocks that can spontaneously organize into periodic structures, this process is readily scalable to large dimensions, and is the main attraction of this process.

The self-assembly technique is well suited for making periodic structures such as photonic crystals. Starting with particles of the appropriately uniform size (in the range 0.5 to 2 μm), one can conceive of organizing them into a periodic 3D array. Particles, rather than molecules or nanoclusters, are the preferred starting materials for photonic crystals because their larger size provides periodicity on the length-scale of optical wavelengths. Large block copolymer molecules have also been used instead, but they are limited in their scope.

The following sections will discuss the salient details of using self assembly to form photonic crystals. The sequence of the discussion follows the path increasing process complexity. To begin, we will examine the simplest of photonic crystals made by neatly packing uniform particles called a colloidal lattice. Given the limited range of monosized particles available, such lattices are inadequate as useful photonic crystals but can be used as templates into which other materials can be incorporated. We shall discuss one particularly attractive incorporation technique, namely electrochemical deposition, in more detail. Finally, since the process is dominated by surface interactions, diverse classes of particles can be generated by simply functionalizing the particle surface. Looking ahead to the future, this is an important aspect that can vastly increase the capability of a colloidal assembly.

Formation of Colloidal Lattices

Colloidal crystallization has been studied for several decades but has received renewed attention due to both the availability of more uniform powders as well as new applications such as photonic crystals. Fine particles are unwieldy in their dry state due to high solid–gas interfacial energy and are typically dispersed in a liquid. Particle concentration is kept low to avoid agglomeration. Further stabilization of the particles may be achieved by either charging the particles' surface (electrostatic stabilization) or by adsorbing a neutral shielding molecule (steric stabilization). A preferred approach is to induce monosized spherical particles to settle into a tightly packed lattice. The settling rate, v, of particles of diameter d_p is given by the well-known Stokes equation:

$$v = \frac{d_\mathrm{p}^2 a \Delta\rho}{18\mu} \qquad [2]$$

where a is the acceleration due to gravity (or a centrifugal field), $\Delta\rho$ is the density difference between the particle and the liquid, and μ is the viscosity of the liquid. Whereas settling is a popular means of self assembling, other methods such as centrifugation, filtering, pressing, or dip coating can be used (Figure 2). For charged particles, electrophoresis, where an electric field induces flow, is applicable.

The typical lattice types in these domains are hexagonal close-packed planes (hcp) as shown in Figure 3 or face-centered cubic (fcc) structures (Figure 4). In fcc crystals, the layer stacking follows the order ABCABC; in hcp it is ABAB. In both stacking sequences, hard spheres occupy a volume fraction of 0.7404 and can be shown to have little entropic differences. However, it is experimentally observed that the preferred ordering of spheres is the fcc close-packed arrangement, which is also the case in high-quality opals.

This method clearly has the advantage of being a simple, cost-effective platform to make 3D photonic crystals that can be readily scaled to various levels. There are no repetitive steps as would be the case in layer-wise lithographic pattering. However, there are still some outstanding challenges. The primary one is the difficulty in ensuring that the organization of the particles does not incur the formation of unwanted defects and domains. Also, introduction of intentional defects such as waveguides is not trivial, although this is being demonstrated in different laboratories. Another shortcoming is the

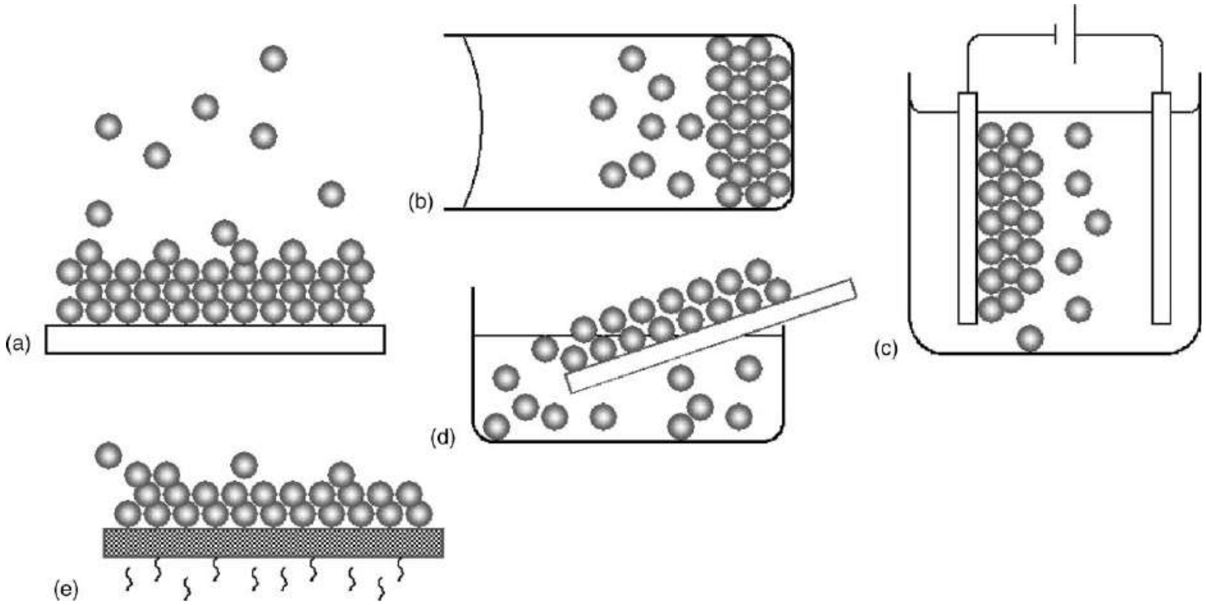

Figure 2 Diagram showing some of the techniques to use self assembly of spherical particles to make colloidal lattices. These include settling (a); centrifuging (b); electrophoresis (c); dip-coating (d); and filter-pressing (e).

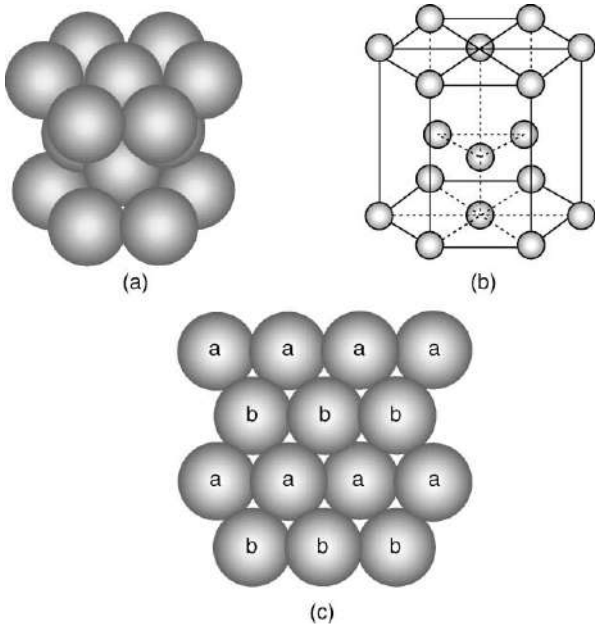

Figure 3 The arrangement of spheres in a hexagonal closed packed (hcp) arrangement (a). The unit cell is shown in (b) whereas the stacking *abab* sequence can be seen in (c).

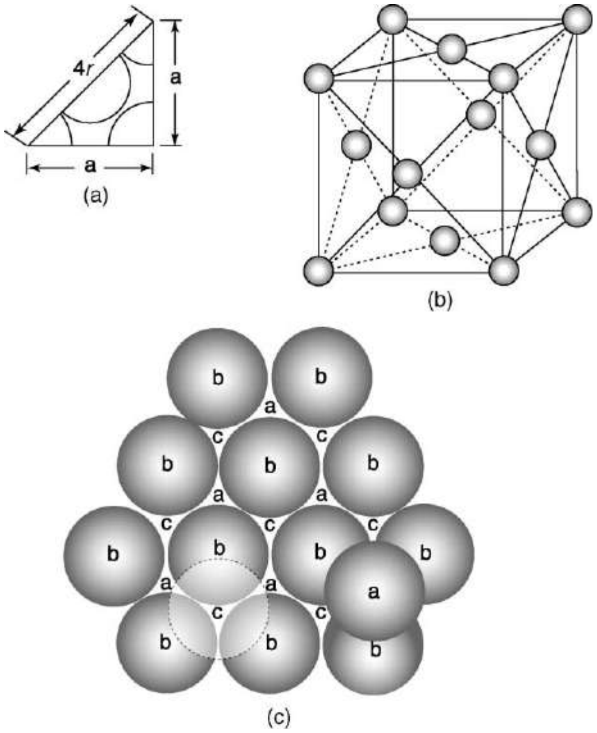

Figure 4 The arrangement of spheres in a face centered cubic (fcc) arrangement with spheres of radius *r* (a). The unit cell is shown in (b). The stacking follows an *abcabc* order. The *a* and *c* sites relative to spheres populating the *b* sites are seen in (c).

unavailability of uniformly sized particles of various different chemical compositions.

The formation of monosized spherical particles is a complex problem in its own right. Usually, such particles can be made from specific polymeric materials such as polystyrene or latex, whereas SiO_2 is the best-known inorganic system. These particles are not perfectly uniform but the typical deviations from the mean size of a few percent are acceptable for most applications. Given the attractiveness of photonic crystals for a wide array of applications in optics, research in this field is burgeoning and more breakthroughs are bound to come. Already first-generation photonic bandgap devices, such as fiber gratings or photonic crystal optical fiber, are commercially available.

Colloidal Lattices as Photonic Crystals

Due to the limited range of composition of the particles available, the particle-to-air refractive index contrast provided by straightforward colloidal lattices is small. This results in low diffraction efficiencies and hence colloidal lattices do not have photonic bandgap in all directions. However, these structures can then be used as templates for various additional modifications, as seen later.

On the other hand, a simple colloidal lattice can be more easily modeled. An analytical solution can be achieved for this system using dynamical diffraction theory for a low diffraction efficiency approximation. For a close-packed stacking of spheres in a fcc array, there exists a minimum in the transmission band at a wavelength λ_c which is given by the Bragg–Snell equation:

$$\lambda_c = 2d_s\sqrt{n_{\mathrm{eff}}^2 - \sin^2\theta} \qquad [3]$$

where d_s is the d-spacing between the planes in the $\langle 111 \rangle$ direction, n_{eff} is the effective refractive index of the assembly and θ is the angle between the incident beam and normal to the (111) plane. This minimum in the transmission is referred to as the primary stop-band.

Photonic crystals are characterized using a Michelson or Mach Zehnder interferometer to obtain the phase and amplitude of the transmission (or reflection) as a function of the wavelength of light (Figures 5 and 6). Ideally, the whole band structure can be calculated from these data. In addition to λ_c, the close-packed structures also exhibit a transmission threshold at a wavelength λ_t, as seen in Figure 5. Both λ_c and λ_t depend on the particle diameter and increase with particle size. In a close-packed structure, the lattice d-spacing d_s is related to the particle diameter d_p as $d_s = d_p\sqrt{2/3}$. So when the incident light is orthogonal to the structure, eqn [3] can be simplified to:

$$\lambda_c = 2\sqrt{2/3}(d_p n_{\mathrm{eff}}) \qquad [4]$$

Thus, n_{eff} can be obtained from the slope of a plot of the stop-band wavelength λ_c as a function of the particle diameter d_p. The angular dependence of the reflectance spectra is illustrated in Figure 7. These results depend upon the polarization state of the incident light and hence need to be obtained separately for s and p polarized light.

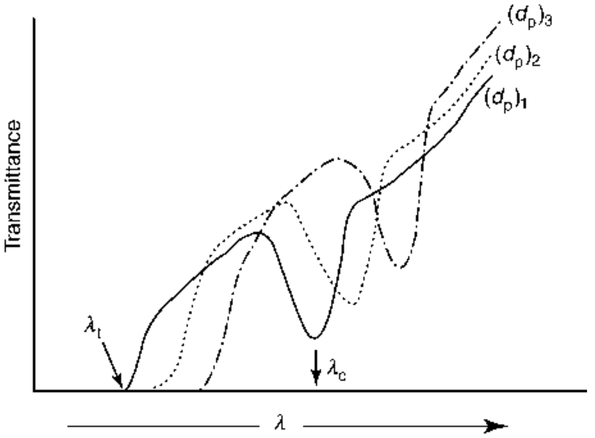

Figure 5 The typical profiles observed from transmission measurements on a photonic crystal. The stopband wavelength (λ_c) increases with d_p, where $(d_p)_1 > (d_p)_2 > (d_p)_3$. The polarization of light is fixed at s or p polarized.

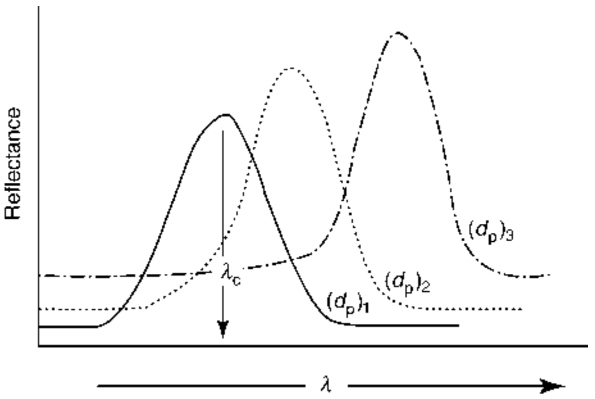

Figure 6 The typical profiles of the reflectance peaks corresponding to the transmission plot seen in Figure 5.

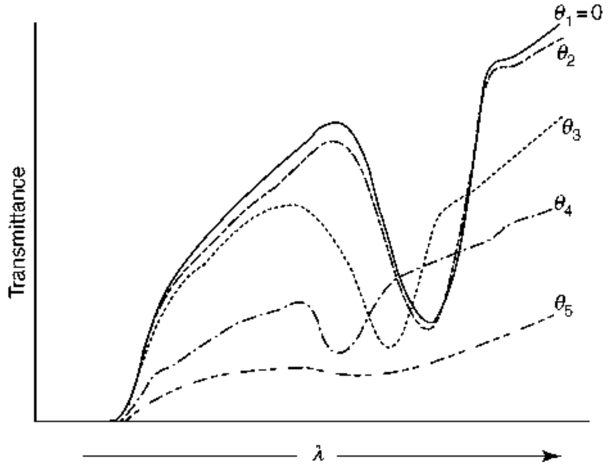

Figure 7 The transmittance profiles as a function of the angle of incidence of light of a fixed polarization. Typical values of θ would be $\theta_1 = 0$, $\theta_2 = 10°$, $\theta_3 = 30°$, $\theta_4 = 50°$, $\theta_5 = 60°$.

More Complex Colloidal Lattices

While the use of SiO_2 and polymer microspheres is ubiquitous, they are optically passive materials. New methods are being developed for making uniform spherical particles from other more optically active systems. The zinc oxide (ZnO) system is one such example. The higher refractive index (~2.2) and the light emission ability of this system make it particularly interesting. Monosized particles in the size range of 100 to 600 nm can be made using a two-stage reaction method, which allows extra control over the particle formation. Other optically active systems include rare-earth doped SiO_2, dye doped PMMA spheres, or composite spheres comprised of core–shell configuration with SiO_2 or zinc sulfide (ZnS) cores and gold (Au) or SiO_2 shells, respectively.

Whereas the above discussion focuses on monosized particles, binary crystals can be made using spheres of two different sizes. The strategy here is to vary the particle size during the deposition of every alternate layer. For instance, a first layer of closely packed spheres is followed by the deposition of layers of smaller particles. The resultant crystal structure depends primarily on the concentration of the smaller particles in the dispersion. As seen in Figure 8, the second layer can vary from having three to six particles around every underlying larger particle. Thus the large-small (LS) stoichiometry of these binary crystals can be varied from LS to LS_2 to LS_3 by increasing the concentration of the smaller particles in the suspension phase. Such crystals open up interesting new avenues for complex photonic crystals.

Colloidal Lattices as Templates

By using the colloidal lattice as porous host for a second material, a greater range of photonic crystals can be synthesized. Here, the role of the colloidal array

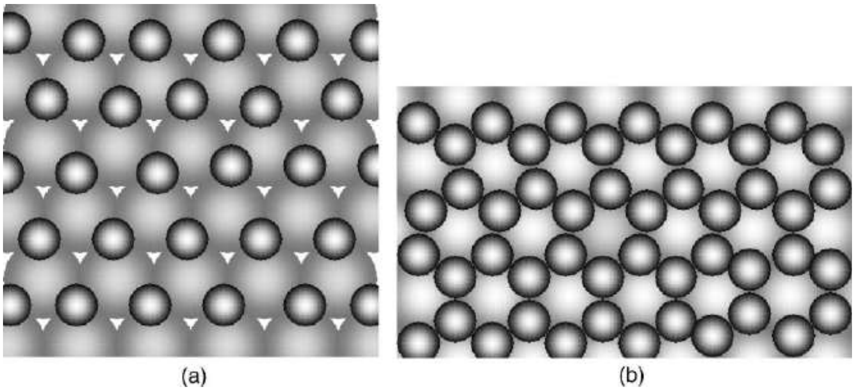

Figure 8 Binary crystals formed using spheres of two different sizes. The large-small (LS) ratio in (a) is 1:1 whereas in (b) it is 1:2.

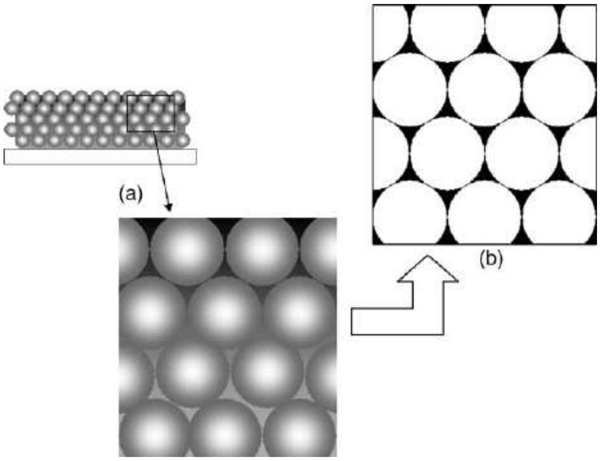

Figure 9 Schematic representation of the templating process using infiltration (a). Upon removal of the particles, a macroporous photonic crystal is formed (b). This is the inverted opal configuration.

is simply as a scaffolding structure. The second material to be infiltrated may be either a liquid (in the form of a dispersion, a solution, or a molten liquid) or a vapor. Both eventually deposit a solid material inside the template. Subsequent selective removal of the template yields a 3D photonic crystal in the form of an inverted opal (Figure 9). This is one of the best and most versatile means of achieving high index contrast, which is required for true photonic bandgap behavior.

It must be noted that to use self-assembled particle arrays as a template, the whole structure has to be often heat treated to confer greater mechanical strength. As assembled, the particles are normally attached by short-range Van der Waals forces and hydrogen bonding, which in the case of SiO_2 is silanol (Si–OH) mediated. Both of these are relatively weak bonds. By heating the assembly, a greater number of siloxane (Si–O) bonds can be formed which additionally strengthens and stabilizes the particulate network. This is important to retain the periodic structure of the assembly during the exposure to high stresses during the infiltration step, for instance. At temperatures around 750 °C, there is significant generation of siloxane bonds but minimal sintering so that the lattice does not change dimensions and retains its original open porosity. With latex particles, the annealing is done just above the glass transition temperature T_g of the polymer for no more than a critical period of time. Longer treatments can destroy the ordering of the spheres.

A template thus strengthened by heat treatment is now ready for infiltration. For example, a template comprised of polystyrene spheres can be infiltrated with a hydrolyzed sol of titanium alkoxide to ultimately form titania in the interstitial regions. Upon standing, the alkoxide sol gels in the interstitial regions leading to a polymeric network called a xerogel. This entire composite can then be further heated to high temperatures (>300 °C). All organics are removed by decomposition and combustion so as to leave behind a titania matrix with periodic voids (in place of the polystyrene spheres). The templating spheres may also be dissolved, rather than combusted, in solvents such as toluene. The refractive index of titania depends on the crystalline phase that results after heat treatment – anatase is ~ 2.5 while rutile is ~ 2.9. High indices can also be achieved with semiconducting materials such as selenium (Se) with an index of ~ 2.5 or cadmium selenide (CdSe), ~ 2.9. These chalcogen-derived materials have low absorption in the visible and near infrared, so they are suitable candidates.

When SiO_2 particles are used for templating, they are eventually removed by etching, typically with dilute, buffered hydrofluoric acid at low temperatures. Given the thermal and chemical stability of the SiO_2 particles, molten Se (or other low melting metals) can be used to rapidly fill the interstitial space. Vapor phase deposition has also been used to fill the porous templates. Examples include the backfilling with silicon (Si) using chemical vapor deposition (CVD), or indium phosphide (InP) using metal-organic chemical vapor deposition (MOCVD).

A variety of structured, macroporous carbon matrices has been made with remarkable optical properties. The templates are artificial opals assembled using SiO_2 microspheres, which are filled with carbon by either: (a) infiltrating with phenolic resin followed by anaerobic pyrolysis of the resin to glassy carbon; (b) CVD of graphitic carbon; and (c) growing diamond-seeded CVD graphite from a plasma. The resultant carbonaceous 'inverse opals' are highly conductive, show intense opalescence due to the ordered arrays of holes, and have a photonic bandgap in the infrared region.

Electrochemical Deposition

One particularly effective backfilling technique is electrodeposition, which has been successfully used to deposit most of the different classes of materials – semiconductors, metals, and polymers. Here, a conducting substrate serves as the appropriate electrode where deposition occurs. By placing a colloidal lattice template on this electrode, the electrochemical deposit is used to fill the interstitial voids in the template (Figure 10). This method has been successfully applied for a host of different metals such as gold (Au), nickel (Ni), platinum (Pt), and alloys such as tin–cobalt (Sn–Co). Likewise, conducting polymers can be efficiently deposited by this method. Although these polymers suffer from inherently low indices, they have favorable electro-optic properties that can be tuned by controlling the composition. In the case of II–VI semiconductors, galvanostatic or potentiostatic deposition can be used: for instance, interstitial CdSe has been deposited (Figure 10) using an acidic solution of electrolytes such as cadmium sulfate ($CdSO_4$) and selenium dioxide (SeO_2).

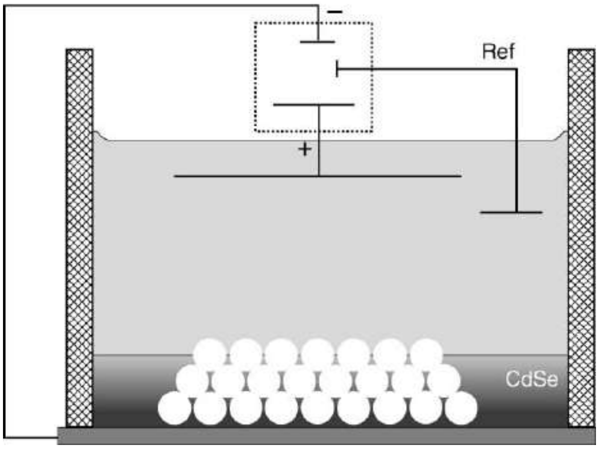

Figure 10 An electrochemical cell for deposition of CdSe from an electrolytic solution on a conducting electrode. The solution contains sulfuric acid (H_2SO_4), $CdSO_4$, and SeO_2. A potentiostat is used to generate the electric fields.

Electrochemically deposited structures have minimal porosity leading to densities close to bulk density. Hence, they have a higher refractive index and are generally more robust than those made by simple infiltration. Also, they exhibit little change in dimensions and a lower tendency toward cracking during the various physicochemical processes such as drying and sintering that occur en route to forming the photonic crystal. Hence, it is a preferred technique in many cases.

Functionalization

The self-assembly process, whether used directly or as a template to synthesize a photonic crystal, is dominated by surface interactions between particles starting from the organization process through to the thermal stabilization step. These interactions can be tailored by specially treating the surface of the particles with functional groups. Functionalized colloidal templates can enhance the infiltration of the high index material and its properties such as smoothness, density, or optical nonlinearity. Latex spheres can be surface functionalized with sulfate and carboxylic acid groups to increase their surface charging capability, which can lead to more robust self-organization. These spheres can also be coated with a layer of SiOH, which gives them a SiO_2-like appearance on the surface and allows them to be used in conjunction with SiO_2 microspheres. The infiltration of metal (such as Au) nanoparticles into a template crystal of SiO_2 microspheres can be significantly enhanced by functionalizing the SiO_2 surface. This can be achieved using long-chain organic molecules with polar groups on one end that bond with silanols on the SiO_2 surface and thiol groups on the other that strongly latch onto the gold nanoparticles. Simulations of light interaction with ordered three-dimensional arrays of these metallodielectric spheres, show that they are promising starting points for the construction of full photonic bandgap materials in the visible part of the optical spectrum. Also, in these systems, the plasmon resonance of the Au shell particles can be controlled over the visible and infrared region by varying the thickness of the Au shell.

The site-selective attachment ability of biological moieties can also be used effectively here by attaching them onto particle surfaces. DNA oligomers are well known to form conjugate specific bonds – adenine to thymine and guanine to cytosine. Thus particles functionalized with these oligomers can be made to organize in specific sequences that are differentiated, depending on the surface group. These particles can have variable sizes or compositions but their surface properties are controlled largely by the attached biomolecules. Likewise, the antigen–antibody attachment ability can be exploited to improve surface bonding.

Positioning Defects in Photonic Crystals

Much of the work on photonic crystals involves the making of a long-range periodic microstructure with high dielectric contrast because the underlying materials engineering aspects are very challenging. However, the usefulness of photonic crystals depends on the successful placement of defects in precise locations within the crystal. The placement of a point defect leads to a photonic microcavity and that of a line defect, a waveguide for propagation of a lightwave (Figure 11). However, there are certain unique advantages to this style of propagation, such as the possibility of bending light around sharp corners. Line defects in the form of air cores have been introduced in a photonic crystal using a combination of lithography and self assembly. First, ridges are defined in a photoresist on top of the substrate by optical lithography (Figure 12). Upon forming a colloidal lattice on top, the ridges are converted to linear voids by removing the photoresist using combustion or etching. Another technique is to infiltrate a photopolymerizable monomer into a colloidal crystal and preferentially polymerize the liquid by focusing a light beam inside the crystal. Thus the locus of the focused beam defines the waveguiding region that is formed upon removal of the unpolymerized liquid.

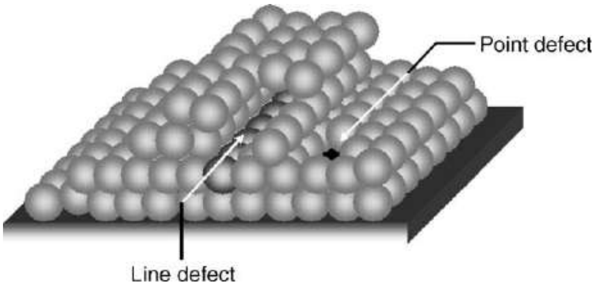

Figure 11 Two common defects in colloidal lattices.

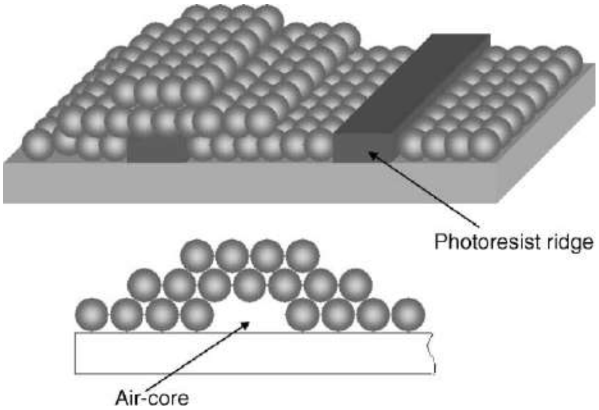

Figure 12 Formation of intentional defect using a lithographic process to define an air-core. This structure can also be used as a template for backfilling a high dielectric material.

Tunable Photonic Crystals

The holy grail in many photonic devices is tunability, and photonic crystals are no exception. Given the wide spectral range that constitutes the domain of light, the ability of devices to switch or translate their performance to another wavelength through an external stimulus is indeed very attractive. The two parameters that can be externally tuned are the dielectric contrast, which would result in a change in the n_{eff} of the crystal and the periodicity of the grating. One technique to tune the former is to form an interference pattern that organizes liquid crystals dispersed in a polymer matrix into droplets to form the photonic crystal. Applying an electric field that changes the refractive index of the droplets can therefore alter the transmission spectrum of the photonic crystal.

To control the lattice parameter, the size of or the spacing between the spheres need to be varied. This concept has also been demonstrated using self-assembled hydrogel-based nanoparticles, which are made of a thermo-responsive material such as poly-*N*-isopropylacrylamide. These particles are soft and conformable and their sizes can be tuned depending on the water content of the system. This ordered matrix is in the form of a gel that can be made to swell and compress upon changing the temperature as water is exchanged between the gel and the ambient. Wavelength tunability of one nanometer has been achieved this way. Tunability has also been sought using methods such as straining the lattice. Calculations show that a 2 or 3% shear strain ought to create a 52–73% distortion of the bandgap. In one novel approach, SiO_2 or polystyrene microspheres were strongly surface charged and ordered with a less than dense packing within a poly-acrylamide

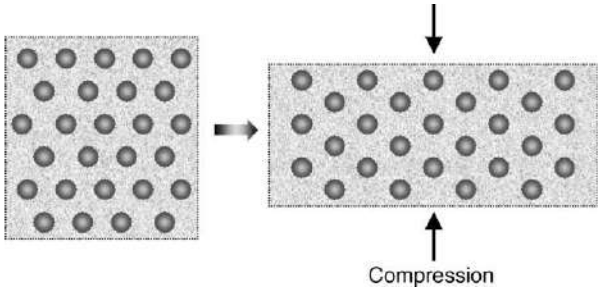

Figure 13 Tunable photonic crystal. Spherical particles are assembled in a compressible media like a hydrogel, which deforms upon applying a stress. The deformation can be used to tune the lattice parameter and hence the optical properties of the photonic crystal.

hydrogel housing (Figure 13). Due to the strong charging the arrangement tends to be body centered cubic (bcc) with the lattice spacing that depends on the concentration of the particles. When the hydrogel is deformed by applying mechanical stress, the diffraction peaks were shown to be tunable over almost the whole visible light wavelength spectrum.

Photonic crystals made from self-organized hydrogels have been used as sensors. Here, they are functionalized with chemical recognition agents that cause the gels to swell upon binding to the analyte. The resultant change in the lattice parameters causes a measurable change in the color. Specific examples include the detection of the toxic organophosphate, parathion, using acetylcholinesterase or of glucose using boronic acid and polyethylene glycol. Utilizing the same principle, even metal ions such as copper can be detected using a hydroxyquinoline-based hydrogel. These serve as a good illustration of the multifaceted potential of photonic crystals.

Summary

Photonic crystals have aroused tremendous interest with their promise to lift the field of photonics to a new level of capability. These crystals are conceptually analogous to electronic semiconductors in their function and hence are called photonic bandgap materials. The big challenge lies in synthesizing them artificially without defects. One of the few methods to synthesize these crystals is by self assembly of micron-sized building units. Self assembly leads to formation of a uniform array of monodispersed particles, typically silica or latex. This process can be further embellished by permeating a second, high-index material using the particulate assembly as a template. Useful applications of photonic crystal require them to have specifically positioned defects and tunability. These aspects, which currently constitute the state-of-the-art research in this field, promise to deliver unique devices and applications.

List of Units and Nomenclature

a	Acceleration (typically due to gravitational or applied centrifugal force)
d_p	Diameter of the particle
d_s	d-spacing or distance between lattice planes
E	Electric field
k_z	Propagation constant for travel in z direction
n	Refractive index of the medium
$\bar{n}$	Modal effective refractive index
n_{eff}	Effective refractive index of the colloidal assembly
v	Settling velocity
$\Delta\rho$	Density difference between the particle and the fluid in which it is suspended
θ	Angle between the incident beam and normal to the lattice plane in question
λ	Wavelength of the lightwave
λ_c	Stopband wavelength
λ_t	Threshold wavelength for the photonic crystal
μ	Viscosity of the suspending fluid
Φ	Electric field distribution

See also

Photonic Crystals: Atomic Physics; Electromagnetic Theory; Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Born M and Wolf E (1981) *Principles of Optics: Electromagnetic Theory of Propagation Interference and Diffraction of Light*. Oxford, UK: Pergamon Press.

Braun PV and Wiltzius P (2002) Macroporous materials – electrochemically grown photonic crystals. *Current Opinion in Colloid and Interface Science* 7: 116–123.

Graf C and Van Blaaderen A (2002) Metallodielectric colloidal core-shell particles for photonic applications. *Langmuir* 18(2): 524–534.

Iwayama Y, Yamanaka J, Takiguchi Y, *et al.* (2003) Optically tunable gelled photonic crystal covering almost the entire visible light wavelength region. *Langmuir* 19(4): 977–980.

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals*. Princeton, NJ: Princeton University Press.

Johnson SG and Joannopoulos JD (2002) *Photonic Crystals: The Road from Theory to Practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kazuaki S (2001) *Optical Properties of Photonic Crystals*. New York: Springer-Verlag.

Koenderink AF, Johnson PM, Galisteo López JF and Vos WL (2002) Three-dimensional photonic crystals as a cage for light. *Comptes Rendus Physique* 3: 67–77.

Soukoulis CM (ed.) (2001) *Photonic Crystals and Light Localization in the 21st Century*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stein A (2001) Sphere templating methods for periodic porous solids. *Macroporous and Mesoporous Materials* 44–45: 227–239.

Velev OD and Lenhoff AM (2000) Colloidal crystals as templates for porous materials. *Current Opinion in Colloid and Interface Science* 5: 56–63.

Xia Y, Gates B, Yin Y and Lu Y (2000) Monodispersed colloidal spheres: old materials with new applications. *Advanced Materials* 12(10): 693–713.

Yablonovitch E (1987) Inhibited spontaneous emission in solid-state physics and electronics. *Physics Review Letters* 58(20): 2059–2062.

PHYSICAL APPLICATIONS OF LASERS

Contents

Free-Electron Lasers in Physics
Industrial Applications
Sum-Frequency Generation at Surfaces

Free-Electron Lasers in Physics

T Dekorsy, Forschungszentrum Rossendorf, Dresden, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The topic of this article is a survey on the main applications of free-electron lasers (FELs) in physical sciences. The merits of FELs are discussed with respect to other available laser sources. For the wavelength ranges, where FELs contributed significantly to advances in physical sciences, specific examples of the research with FELs are given.

The basic physics of FELs, their historic development, and the state of the art of these laser sources are described in detail in this Encyclopedia (*see* **Lasers:** Free Electron Lasers). The present article covers the applications of FELs in physical sciences, but excludes all phenomena involved in the physics of FELs themselves. From the application point of view, one should regard an FEL as any other laser, as a source of coherent electromagnetic radiation with high power, narrow bandwidth, and a diffraction limited beam. Thus, the radiation produced by an FEL significantly differs from that generated by synchrotrons (*see* **Incoherent Sources:** Synchrotrons). The specifications of the FEL, such as wavelength, temporal structure of the output intensity, and peak power, determine the possible applications of the laser. One might now ask, why it is worthwhile to write – and read – an article on its own devoted to the applications of FELs. The answer to this question is the uniqueness of the radiation produced by FELs, which enables certain classes of experiments and applications in specific wavelength ranges not covered by other laser sources, such as gas, solid state, or semiconductor lasers.

Since the construction and operation costs of an FEL facility are usually orders of magnitude higher than those for the equipment of a conventional laser spectroscopy laboratory, FEL facilities have been built, and are planned for the future, in different places all over the world as dedicated-user facilities. These user facilities allow researchers from universities and industry to exploit the unique properties of FEL radiation for their research and provide the basic infrastructure and support for setting up the experiments. Usually potential users apply using a standard procedure for beam time at an FEL by submitting a short proposal to the facility. The proposals are judged by a panel which distributes the available beam time to the different users. So researchers active in the field of optics should keep in mind that these facilities exist and are available to anyone who has specific needs which can be fulfilled by an FEL facility. A list of all major FEL facilities is given in Table 1.

Table 1 FEL user facilities. Facilities marked with an asterisk are under construction at the time of publication

Facility name	*Wavelength ranges*
iFEL (Osaka, Japan)	230 nm–1.2 μm 1–6 μm, 5–22 μm 20–60 μm, 50–100 μm
Duke University Free Electron Laser Laboratory (US)	193–400 nm 2–9 μm
Vanderbilt University Free Electron Laser Center (US)	2.1–9.8 μm
FELIX–FOM (Rijnhuizen, Netherlands)	4.5–250 μm
Stanford Picosecond FEL Center (US)	3–15 μm 15–65 μm
CLIO–LURE (Orsay, France)	3–90 μm
Jefferson Lab (US)	3–6.2 μm
FEL–SUT–Science University of Tokyo (Japan)	5–16 μm
UCSB Center for Terahertz Science and Technology (US)	63–338 μm 338 μm–2.5 mm
ENEA Compact FEL (Frascati, Italy)	2–3.5 mm
FELBE (Dresden, Germany)*	5–150 μm
SCSS SPring-8 Compact SASE Source (Japan)*	X-ray
Institute for Plasma Research (India)*	sub-mm
BESSY FEL (Germany)*	deep UV–X-ray
Brookhaven–NSLS SDL DUV–FEL (US)*	deep UV
University of Maryland–MIRFEL (US)*	far-infrared
University of Hawaii (US)*	UV–far-infrared
SLAC SSRL X-ray FEL–LCLS (US)*	X-ray
TESLA X-ray FEL, DESY (Germany)*	X-ray
Daresbury 4GLS (UK)*	IR, X-ray
MIT Bates Lab X-ray FEL (US)*	X-ray
INFN SPARX X-FEL (Italy)*	X-ray
Tel Aviv University (Israel)*	far-infrared
Center for Advanced Technology (India)*	far-infrared

The basic principles of an FEL are briefly summarized here, while for an in-depth description the reader is referred to the article on free electron lasers (*see* **Lasers:** Free Electron Lasers). As with any other laser, the FEL consists of a laser cavity and a gain medium. The gain media are free electrons, which travel with a velocity close to the speed of light in vacuum through a periodic array of magnets called an undulator or wiggler. The magnets are arranged in a way that the electrons feel an alternating magnetic field. Due to the Lorentz force, the electrons perform an undulating or wiggling motion perpendicular to the magnetic field, which gives rise to spontaneous electromagnetic radiation into the forward direction. The radiation, emitted in due course of the passage of the electrons through the undulator, superposes constructively. The FEL may operate in such a way that the gain from one passage of an electron bunch is high enough to produce a coherent light pulse. This regime is known as self-amplified spontaneous emission (SASE) regime which is especially desirable for short wavelength operation in the UV and X-ray region, where the manufacturing of adequate resonator optics is difficult. At longer wavelengths, a laser cavity provides the feedback for the build-up of the coherent light field. The laser starts lasing when the amplification exceeds the losses of the cavity, including the light intensity coupled out of the cavity.

One of the distinct properties of FELs, in comparison to other laser sources, is their continuous tunability. The emission wavelength of an FEL is given by the following relation (*see* **Lasers:** Free Electron Lasers):

$$\lambda = \frac{d}{2\gamma^2}\left(1 + \frac{K^2}{2}\right) \qquad [1]$$

with d the period of the magnet array, the γ-parameter is the relativistic factor related to the electron velocity u via $\gamma = 1/(1 - u^2/c^2)^{1/2}$, and the K-parameter is given by d, and the magnetic field B as $K = 0.934 \times B[\mathrm{T}] \times d[\mathrm{cm}]$. From eqn [1] it becomes obvious that the emission wavelength of the FEL can be continuously tuned via the electron energy and the magnetic field. This is in distinct contrast to other laser sources, where the emission wavelength is fixed by the energy separation of electronic levels in atoms or by the bandgap energy of semiconductors.

The main advantage of FELs concerns the high power, both average and peak, which can be obtained. In addition, an FEL provides short optical pulses due to the fact that the electron beam passing through the undulator consists of short electron bunches. The temporal width of these electron bunches can be below one picosecond (10^{-12} s), so that optical pulses in the sub-picosecond to picosecond range are obtained. These pulses are extremely adjuvant for performing nonlinear optical experiments (*see* **Spectroscopy:** Nonlinear Laser Spectroscopy) and pump-probe experiments with high temporal resolution (*see* **Chemical Applications of Lasers**: Pump and Probe Studies of Femtosecond Kinetics). Typically the optical pulses are Fourier transform limited, i.e., the product of the spectral bandwidth $\Delta\nu$ times the pulse duration $\Delta\tau$ equals a constant, which depends on the specific shape of the pulse. For a pulse of Gaussian shape in frequency and time, this so called time-bandwidth product is given as $\Delta\nu\,\Delta\tau \geq 0.441$. From this equation the minimum spectral width (temporal duration) can be calculated for a given temporal duration (spectral width) of the laser pulse. When the electron beam driving the FEL is produced by a superconducting radio frequency linear electron accelerator, the repetition rate of the pulses is in the range of some tens of MHz. The energy per optical pulse is obtained by dividing the average power of the FEL through the repetition rate. The peak power within an optical pulse is obtained by dividing the energy per pulse through the pulse duration, which can be in the picosecond to sub-picosecond range. Typical values for FELs are energies per pulse in the μJ to mJ range, pulse lengths from some 100 femtoseconds to a few picoseconds, and accordingly peak powers in the range of kW to GW. Depending on the type of electron accelerator employed, the pulse structure of an FEL may have further peculiarities. Superconducting radio-frequency accelerators provide a quasi continuous-wave (CW) train of electron pulses, thus enabling the FEL to emit a continuous optical pulse train. Electrostatic accelerators provide macropulses of electrons at some Hz repetition rates with a duration of several μs. These macropulses consist of micropulses with MHz repetition rates and picosecond or sub-picosecond duration. Hence the FEL output has the same temporal characteristics with the FEL turning on at the beginning of each macropulse.

Another important development is the synchronization of table-top short-pulse laser systems to FELs. This is accomplished by stabilizing the repetition rate of the table-top laser system via a movable intracavity mirror accurately to the repetition rate (or a higher harmonic of the FEL) of the FEL. Using a high-frequency electronic phase-locked loop, a timing jitter, down to 500 femtoseconds between the FEL pulses and those from a mode-locked titanium-sapphire laser, have been obtained. Nonlinear frequency conversion of the near-infrared pulses from the synchronized titanium-sapphire allows one to perform multi-color pump-probe measurements in a broad spectral range in combination with the FEL.

In the following, FELs will be discussed with respect to the wavelength range which they cover. For each wavelength range alternative laser sources are given and compared to FELs. It should be noted that these alternative sources evolve very rapidly and thus pose challenges on the justification of FELs in certain wavelength and power ranges. For each wavelength region selected scientific achievements obtained from experiments performed with FELs are given.

FELs in the Far Infrared Region

The far infrared region is roughly defined as the wavelength (energy) range from 20 μm to 300 μm (62 meV to 4 meV). At the long wavelength end (wavelengths larger than 220 μm) radiation can be obtained from pure electronic devices, so-called backward wave oscillators (BWO). Furthermore, optically pumped gas lasers operate at certain discrete wavelengths in this range where there exists one fundamental problem for realizing a semiconductor laser based on transitions between the conduction and the valence bands. Auger recombination, where an electron and hole combine by transferring another electron high into the conduction band, becomes increasingly important for smaller bandgaps. Recent advancements in this field are based on quantum cascade lasers, which use inter-sub-band transitions of quantum structures for achieving lasing in the wavelength range from 105 μm to 3.4 μm. Another possibility to generate coherent radiation in this wavelength range is optical rectification or difference frequency mixing of femtosecond laser pulses at near-infrared wavelengths. However, none of these sources outperform FELs in terms of peak or average power, making FELs the most important tunable laser source in this wavelength range.

In the far-infrared, several scientific opportunities exist. In condensed matter many important elementary excitations exist in this frequency range, e.g., phonon-polaritons, plasmons, polarons, and magnons. The quantum confinement in semiconductor heterostructures leads to the formation of discrete electronic levels with energy separation typically in the far-infrared. The associated radiative transitions, so-called inter-sub-band transitions, can be arbitrarily tuned by varying the materials and the dimensions of the confining potential. In surface science, intramolecular vibration of molecules adsorbed to surfaces are of large interest. They can be studied by resonant nonlinear optical experiments such as sum-difference frequency generation or second harmonic generation, giving information about the energy transfer between the molecules and the surface. In biophysics, low-frequency modes of large bio-molecules lie in the far-infrared, whose dynamics can also be studied.

Important scientific results have been obtained with FELs in the following fields, which exclude the important fields of molecular science, chemistry, biology, and medical research.

Scanning Near-Field Optical Microscopy (SNOM)

Increasing the spatial resolution is an important task in spectroscopy. Free-space focusing of a laser beam is diffraction-limited to a spot with the dimensions of the wavelength. Especially in the far-infrared region this would allow to resolve objects larger than 10 μm only. However, the imaging of elementary excitations in solid state or of collective excitation in biomolecules on a sub-μm scale, would be highly desirable. SNOM allows to circumvent the diffraction limit. Several different SNOM techniques have been introduced in recent years, most of them working with apertures smaller than the wavelength. Imaging is performed by scanning the aperture over the sample or vice versa. An evanescent electromagnetic field is present at the aperture which can couple to the specimen brought into the proximity of the aperture. Since the aperture is smaller than the wavelength, the light intensity drops by order of magnitudes. It is here that the high power available from FELs becomes advantageous. Usually, the end of a fiber is tapered and metalized, leaving a small aperture at the tip of the fiber. In the far-infrared region the problem exists in the absorption of the fiber material. Chalcogenide fibers are capable of transmitting in the 2–12 μm range, while hollow glass capillaries coated with metals inside work also at longer wavelengths. Recent approaches try to work without apertures, e.g., by coupling the light locally via a nanometer metal tip to the sample. Another interesting approach is to use the focus of visible laser light to create a small highly reflective spot on a material. By creating free-carriers with the visible laser light in the material, the reflectivity of the far-infrared light of the FEL can be strongly increased, thus a reflecting spot is created which is diffraction-limited by the wavelength of the visible light rather than that of the infrared light. The techniques described here are, of course, not limited to the far-infrared region but may also be applied successfully, for example, in the mid-infrared region.

Rabi Oscillations

Rabi oscillations are a fundamental phenomenon in a quantum mechanical two-level system driven by a light field. The light field promotes electrons from the lower to the upper state and back, depending on the duration and intensity of the light field, i.e., the population oscillates between the lower and upper state of the two-level system with the oscillation period depending on the intensity of the light field. This phenomenon has been studied extensively in atoms. In semiconductors, scattering processes strongly aggravate the observation of Rabi oscillations. By using the Santa Barbara FEL Rabi oscillations between the ground and excited state of a hydrogenic, impurities in GaAs could be observed. From the excited state of the impurity, electrons are transferred to the conduction band, thus changing the photoconductivity of the crystal, which is measured as a function of time delay. Overdamped oscillations with frequencies changing with the laser intensity could be clearly resolved.

Optical Nonlinearities in Semiconductors

An important concept of nonlinear optics is based on the nonlinear susceptibility $\chi^{(n)}$ of a material, where n denotes the order of the nonlinearity. The nonlinear susceptibility has tensorial character and determines, for example, the efficiency of second harmonic generation and difference frequency mixing ($\chi^{(2)}$), third harmonic generation, and the Kerr nonlinearity ($\chi^{(3)}$), etc. Hence the knowledge about the dispersion and the absolute value of $\chi^{(n)}$s is both of fundamental and practical importance. Due to the lack of tunable laser sources, these values are not well known in the far-infrared. Especially close to fundamental lattice resonances in the far-infrared, a strong dispersion is theoretically expected. Here FELs play an important role in the determination of these values. For example, in GaAs, the dispersion of the $\chi^{(2)}$ has been recently determined from second harmonic generation experiments below the fundamental optical phonon resonance at 8 THz, i.e., wavelengths larger than 37 μm, where a resonance enhancement at half the phonon frequency could be observed.

Relaxation Kinetics in Semiconductor Quantum Structures

The investigations of inter-sub-band transitions in semiconductor heterostructures have opened the pathway for the realization of the quantum cascade laser. In these devices, the radiative and nonradiative transition rates between different electronic quantum levels are of prime importance for a working device. The relevant time-scales have been obtained from pump-probe experiments with FELs. In semiconductor quantum dots, the strong coupling of phonons to inter-sublevel transitions could be observed and the polaron lifetime could be determined.

Bloch Oscillations in Semiconductor Superlattices

Bloch oscillations are an intriguing concept of solid state physics. When an electron experiences the force of an electric field in a periodic potential, the resulting dynamics is oscillatory in the absence of scattering events. However, since in real solid state the scattering rates are much higher than the expected oscillation period, such oscillations cannot be observed and the application of an electric field leads to a net electric current. However, in artificial semiconductor superlattices, the Bloch oscillation period can be tailored to be smaller than the scattering rate, leading to the observation of Bloch oscillations and the associated THz emission. The inverse effect of a Bloch oscillator, i.e., the generation of a photocurrent by tuning an FEL to the Bloch oscillation frequency, could be observed.

Cluster Physics

FELs can be an important tool for studying the dynamics of clusters and molecules in the gas phase by infrared resonance enhanced multiphoton ionization (IR-REMPI). In IR-REMPI, the species of interest are irradiated by pulsed, tunable IR radiation from a free electron laser. When the radiation is resonant with a vibrational mode of the molecule or cluster, the absorption of many – up to several hundred – photons can take place. When the internal energy is high enough, the thermal emission of an electron can take place. The resulting ions can be detected mass selectively in a time-of-flight mass spectrometer. Monitoring the mass specific ion yield as a function of IR wavelength then yields the IR spectrum of that species. This technique has given important information on the IR absorption spectra of interstellar dust.

FELs in the Mid-Infrared Region

This wavelength (energy) range extends from 20 μm to 2 μm (62 meV to 0.62 eV). Alternative sources in this wavelength range are nonlinear optical parametric generators and amplifiers (OPGs/OPAs) which compete against FELs in terms of peak power. However, there still exists a large number of interesting phenomena where FELs are better suited, e.g., due to their higher repetition rate in the MHz range as opposed to the kHz repetition rate of OPGs/OPAs and their higher average power exceeding several W up to kW as compared to mW for OPGs/OPAs. These high-power levels find applications in material processing and recently in surgery. Since these applications cover such a wide variety of different materials, we refer the reader to publications from users of the Jefferson Laboratories FEL and the Vanderbilt University FEL (see Table 1). The advantage of FELs in this context is that their wavelength can be tuned exactly to a resonance of the material under preparation. The following significant spectroscopic application was performed with FELs.

Vibrational Relaxation of Defects in Semiconductors

The heat dissipation in submicron electronic devices is of prime importance for the failure behavior of such devices. As the device size shrinks to the nanometer scale, atomic defects play an important role. In silicon, hydrogen passivation of defects is a standard process step in the production of metal-oxide-semiconductor electronic devices. Hence detailed knowledge about the stability and heat dissipation of hydrogen defects is very important. By employing nonlinear optical methods such as four-wave mixing or a transient bleaching technique, the vibrational lifetime of hydrogen vibrational modes in silicon (around a wavelength of 5 μm to 6 μm) could be determined and important information on the dependence of the local atomic arrangement and the vibrational lifetime could be obtained.

FELs in the Near-Infrared to the UV Region

The wavelength (energy) range from 2 μm to 200 nm (0.62 eV to 6.2 eV) is the realm of solid state lasers. Especially diode pumped solid state lasers have made tremendous improvements in terms of power and stability in the past years. Nonlinear frequency conversion of ultrashort optical pulses generated from titanium-sapphire lasers and amplifiers cover almost all of this wavelength range. Hence there is no specific need for operating FELs in this wavelength range for applications in physical sciences.

FELs in the Extended UV to X-ray Region

In the wavelength (energy) range shorter than 200 nm (larger than 6.2 eV) the availability of laser sources is very scarce. In this wavelength range, solid state lasers have the problem of the reabsorption of the laser light by the crystals' hosts. Also nonlinear frequency conversion processes are hampered by the absorption of the nonlinear crystals in this wavelength range. The generation of ultrashort pulses at the wavelength of the Cu K_α line has been obtained on the basis of amplified femtosecond titanium-sapphire pulses, thus enabling the study of the lattice vibrations in the time domain via a modulation of the X-ray diffraction. Since these sources are limited in their wavelength and output power, the development and operation of FELs in this wavelength range is truly important.

Only recently the first experimental verification of the feasibility of a SASE FEL at short wavelengths was obtained at the DESY facility. The lasing wavelength was 110 nm. Since several of these sources are presently under construction worldwide, there are as yet no experimental results published. However, the future development of short-wavelength FELs will enable new types of time-resolved experiments in atoms, solid state physics, chemistry, and biology. One prominent example is protein crystallography with X-ray pulses. The high intensity of FELs in this wavelength range should allow one to obtain a diffraction pattern of a single protein or other important bio-molecules in a single shot on a subpicosecond time-scale, i.e., before the protein is fragmented. The short-pulse duration opens the pathway to study the dynamical behavior of such structures in pump-probe experiments with solid-state lasers synchronized to the FEL. In solid state physics, such experiments could yield important insights into the dynamics of phase transitions.

Other Developments

Beyond the use of an FEL as a light source in the above described wavelength ranges, there exists the possibility of using FEL facilities in a broader way. One example is the development of a high-intensity γ-ray source which can be accomplished by Compton backscattering of FEL photons from the relativistic electrons driving the FEL. This work has been performed at a storage ring FEL (OK-4/Duke University), where 12.2 MeV γ rays were obtained by backscattering 379.4 nm free-electron laser photons from 500 MeV electrons circulating in the storage ring. The advantage of these γ rays is that they are highly monochromatic and linearly polarized, thus making them an ideal source for nuclear spectroscopy and investigations in cancer therapy.

See also

Incoherent Sources: Synchrotrons. **Lasers:** Free Electron Lasers. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Andruszkow J, Aune B, Ayvazyan V, *et al.* (2000) First observation of self-amplified spontaneous emission in a free-electron laser at 109 nm wavelength. *Physical Review Letters* 85: 3825–3828.

Benson SV (2003) Free electron lasers push into new frontiers. *Optics and Photonics News* 14: 20–25.

Budde M, Lupke G, Chen E, *et al.* (2001) Lifetimes of hydrogen and deuterium related vibrational modes in silicon. *Physical Review Letters* 87: 145501–145503.

Cole BE, Williams JB, King BT, Sherwin MS and Stanley CR (2001) Coherent manipulation of semiconductor quantum bits with terahertz radiation. *Nature* 410: 60–62.

Colson WB, Johnson ED, Kelley MJ and Schwettman HA (2002) Putting free-electron lasers to work. *Physics Today* 55: 35–41.

Dekorsy T, Yokovlev VA, Seidel W, Helm M and Keilmann F (2003) Infrared-phonon-polariton resonance of the nonlinear susceptibility in GaAs. *Physical Review Letters* 90: 055508.

Helm M (ed.) (2000) *Long Wavelength Infrared Emitters Based on Quantum Wells and Superlattices*. Amsterdam: Gordon and Breach.

Knippels GMH, van de Pol MJ, Pellemans HPM, *et al.* (1998) Two-color facility based on a broadly tunable infrared free-electron laser and a subpicosecond-synchronized 10-fs-Ti sapphire laser. *Optical Letters* 23: 1754–1756.

Litvinenko VN, Burnham B, Emamian M, *et al.* (1997) Gamma-ray production in a storage ring free-electron laser. *Physical Review Letters* 78: 4569–4572.

Murdin BN, Knippels GMH, Vandermeer AFG, *et al.* (1994) Excite-probe determination of the intersubband lifetime in wide GaAs/AlGaAs quantum wells using a far-infrared free-electron laser. *Sem. Sci. Techn.* 9: 1554–1557.

Palanker DV, Knippels GMH, Smith TI and Schwettman HA (1998) IR microscopy with a transient photo-induced near-field probe. *Opt. Comm.* 148: 215–217.

Sauvage S, Boucaud P, Lobo RPSM, *et al.* (2002) Long polaron lifetime in InAs/GaAs self-assembled quantum dots. *Physical Review Letters* 88: 177402.

Sorokina IT and Vodopyanov KL (eds) (2003) *Solid-State Mid-Infrared Laser Sources*. Berlin: Springer.

Unterrainer K, Keay BJ, Wanke MC, *et al.* (1996) Inverse Bloch oscillator: strong terahertz-photocurrent resonances at the Bloch frequency. *Physical Review Letters* 76: 2973–2976.

von Helden G, Tielens ACGM, van Heijnsbergen D, *et al.* (2000) Titanium carbide nanocrystals in circumstellar environments. *Science* 288: 313–316.

World Wide Web Virtual Library (2004) *Free Electron Laser Research and Applications*. Available online at http://sbfel3.edu/www/UL_fel.html

Industrial Applications

I P Mercer, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lasers are seeing increasing application in industry, driven by emerging market demand and continuing enabling advances in laser technology. Core laser technology advances are based on robustness, performance, and cost. A combination of gas and solid-state lasers currently see application across the spectrum, from the extreme ultraviolet (EUV) to the infrared (IR). Industrial applications continue to emerge across the spectrum and are being extended to the extreme ends of the spectrum at terahertz (THz) and X-ray frequencies (see Figure 1).

Where a step in level of performance is offered by a laser based tool, that is robust enough and cost-effective for industrial application, it can be anticipated that new applications and markets will emerge. For this reason, many established laser manufacturers and new entries in the 1980s and 1990s were driven in part by the exploitation of laser technology advances, as opposed to market application demand. As markets mature for lasers applied to materials processing and diagnostics, so the driver across industry is beginning to shift from being predominantly set by the capability of emerging laser technology, to a combination of laser technology drive and application demand.

Figure 2 shows the current and emerging picture of laser technologies across a core range of wavelengths and pulse durations. Current technologies are shown along with emerging technologies expected to see industrial application in this decade, opening up new applications and markets. Currently established technologies include: CO_2 gas lasers at 10 μm; excimer and KrF gas lasers emitting in the ultraviolet at 248 and 193 nm, respectively; direct diode lasers emitting in the near infrared and visible; and solid state lasers emitting in the near infrared and visible. Current technology developments are centered on extending the range of performance of both gas and solid-state lasers: gas lasers are being extended to the ultraviolet with gases such as F_2; diode lasers are being extended to the ultraviolet with materials such as nitride-based semiconductors and to the infrared with the incorporation of quantum cascade technology; and solid-state diode pumped lasers are being extended to the infrared and ultraviolet via nonlinear conversion. In essence, achievable laser parameter space is seeing continued expansion, not only in terms of pulse duration and wavelength, but also in terms of average power, pulse energy, beam quality, robustness, and cost. Laser technology has now arrived as a major driver for industrial and social change through this decade and into the next.

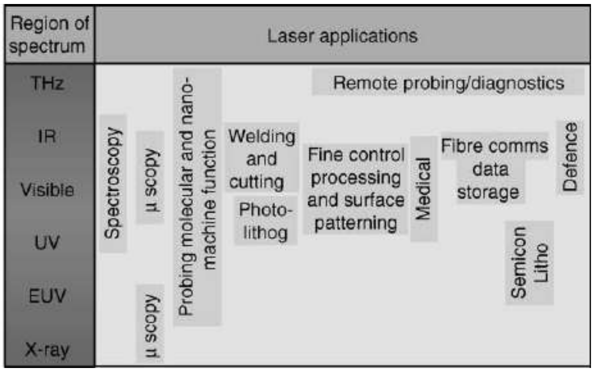

Figure 1 Industrial laser applications across the spectrum.

Infrared wavelengths directly stimulate nuclear motion, coupling to heat without necessary recourse to electronic excitation. Visible wavelengths couple to electronic transitions and at high laser intensities can ionize via multiphoton absorption. Ultraviolet wavelengths couple to electronic transitions and for wavelengths in the region of or below 200 nm, ultraviolet radiation can directly result in bond breaking. For materials processing applications, ultraviolet radiation can have advantages in reducing the laser threshold required for ablation, translating to reduced thermal load imparted to the material.

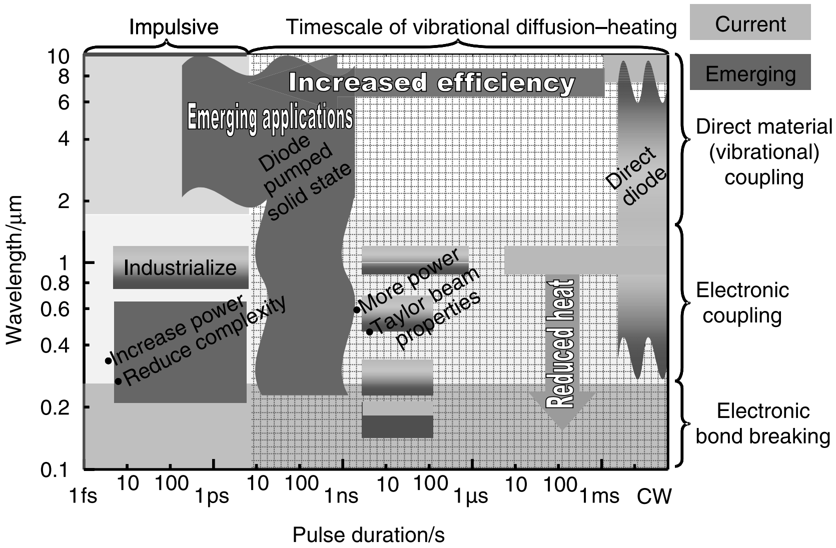

Figure 2 Current and emerging laser technologies across the spectrum as a function of laser pulse duration. Currently established technologies (light gray) are shown with technologies expected to emerge during this decade (dark gray).

Nuclear motions in materials cover a range of timescales of dynamics from 100 ps (0.3 cm^{-1}) to 10 fs (3000 cm^{-1}). As the laser pulse duration is reduced below 1 μs, the effects of thermal diffusion are mitigated in solids and as the pulse duration is further reduced below 100 ps, the interaction between radiation and matter is seen to adopt impulsive behavior. Reduction in pulse duration below the microsecond regime can lead to increased efficiency for material ablation combined with improved quality.

Materials Processing – Photochemical

Lithography in Semiconductor Manufacture

Laser technology's role in semiconductor and microelectronics fabrication is growing as manufacturers move towards producing smaller, more powerful devices. The overriding commercial driver for a generation has been to increase the number of integrated components per area of silicon. This has been achieved by reducing the feature size of a transistor on silicon and this trend is anticipated to remain in place through to 2010.

Features on silicon are chemically etched following lithographic exposure. To meet the increasing demand in resolution and throughput, lasers have been applied with a trend in development towards shorter (ultraviolet) wavelengths, with increasing power and reliability.

The semiconductor industry is currently in transition from 248 nm (XeCl gas) to 193 nm (ArF gas) ultraviolet (UV) pulsed laser technology to expose silicon wafers of diameter 30 cm. A typical laser power available at this time at 193 nm, is 40 W at a repetition rate of 4 kHz. To continue the reduction of feature sizes by a factor of two every four to five years, the industry is considering moving to 157 nm (F_2 gas) pulsed laser technology. As wavelengths reach the deep ultraviolet (DUV), transmission through gas is significantly reduced, requiring a move to vacuum-based systems. Associated technologies relating to translation (stepper), mask, and image relay optics also require significant advancement. A technology also being considered at this time is to image the mask to silicon through a medium with refractive index greater than unity, thereby reducing the minimum feature size afforded by diffraction theory, whilst maintaining the same laser wavelength and numerical aperture.

It is estimated that at a feature size of approximately 80 nm, the above technologies will be reaching their resolution limit. This will necessitate a switch to electron beam projection lithography or extreme ultraviolet (EUV, 13 nm) technology. Electron beam projection carries a disadvantage in terms of projected rates of wafer throughput and as such, emphasis has been placed by the industry on the current development of EUV technology. EUV radiation can be produced by laser interaction with a gaseous volume.

Computer to Plate Printing

A common advantage where lasers are replacing current technologies is to reduce the number of associated processes and thereby reduce the cost of process ownership. Computer to plate (CTP) printing involves the laser exposure of a printing plate without the use of a film intermediate. This technology is finding fast traction within the printing industry and is applicable to drum or flat bed platforms. CTP first gained serious attention in the 1990s and has since substantially advanced, with a wide range of technology suppliers (Figure 3).

Traditional methods have involved offset printing plates, imaged used traditional film, and high-power

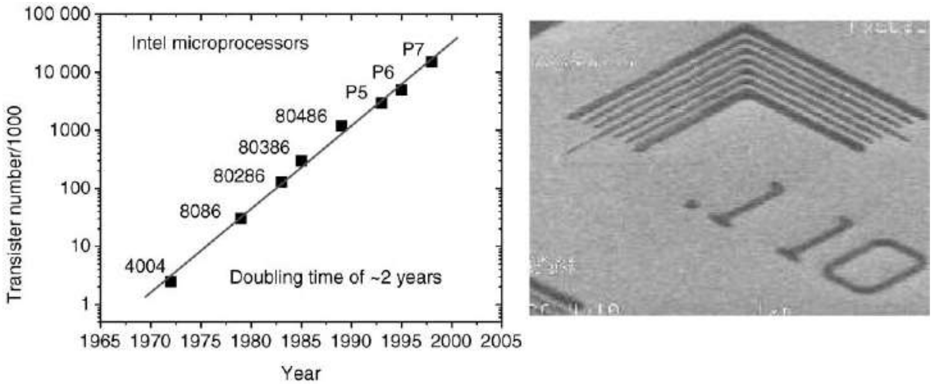

Figure 3 Moore's Law: continuing growth in computing power, showing a doubling in the number of transistors per unit area of silicon every 2 to 2.5 years. Also shown are 110 nm dense lines on silicon (state of the art in 2002). Source: International Sematech.

UV lamps. In contrast, CTP exposes the plate directly with laser radiation. The major saving reported by users of CTP is a faster turnaround time. For color printers, CTP gives almost instant color register on the press and there is a reduced requirement to stop the press to delete dust marks and scratches, as was the case with film. As the industry sees transition to reduced press run lengths, so the importance of CTP is increasing.

Digital plates generally use a high-quality, grained, and anodized aluminum substrate, as for analog plates (see Figure 4). High tolerance electrochemical graining and anodizing ensure stable ink/water balance and good-quality printed results and translates to predictable press behavior. Multiple coatings are applied for CTP, in contrast to a single coating for offset printing, and these take the form of a photopolymer or silver halide and a thermally responsive layer.

On the press, ink and water rollers are continually rolled over the plate. The grained and anodized aluminum substrate serves as a hydrophyllic surface volume to attract water. Ink is oil based and is repelled from areas of water absorbed into the substrate, but is attracted to the hydrophobic dot material.

Silver halide is rated for runs up to 350 000, proving a popular choice for commercial and newspaper applications. It is straightforward to sensitize for a range of writing wavelengths and is the fastest to write.

Thermal plates offer high resolution, equivalent to silver halide and can be baked to enable the longest application lifetimes. Typically the plates are exposed to 830 nm radiation on the external surface of a drum. This technology has served the industry well, but has struggled to find full application in the newspaper business, due to its disadvantage in writing speed.

The most common CTP technology used for newspapers uses a photopolymer coating, which is fast to write and durable. Resolution has traditionally been limited, with Agfa, for example, rating their N91 photopolymer plate to a maximum of 175 lines per inch (LPI), whilst thermal and silver plates are able to image at 200 LPI and above. However, recent advances in shorter wavelength violet diode lasers and violet sensitive polymers have improved resolution. Launched in 2000, violet technology is seeing widespread takeup. This technology is a spinoff from the DVD industry and with laser lifetimes estimated at 10 to 20 years, cost of ownership is one of the drivers. Both silver halide and photopolymers can be sensitized to violet. Current typical laser parameters are for: silver halide, 2–3 μJ/cm^2 at 405 nm; photopolymer, 0.2 mJ/cm^2 at 488 or 532 nm; and thermal, 0.1 J/cm^2 at 830 nm.

The chemical composition of the CTP surface layers is balanced to determine the response to different laser wavelengths, as well as performance characteristics in different print applications. The plate coatings undergo a photochemical change during laser exposure. This is a critical stage since it is important that the plate surface is not under- or over-exposed.

Chemical post processing after exposure is generally required for CTP to date, with the exception of ablative CTP, to remove unexposed material. After imaging, thermal plates go through a pre-bake step, during which the system heats the plate to 130 °C (266 °F) for about 30 seconds. The energy from this

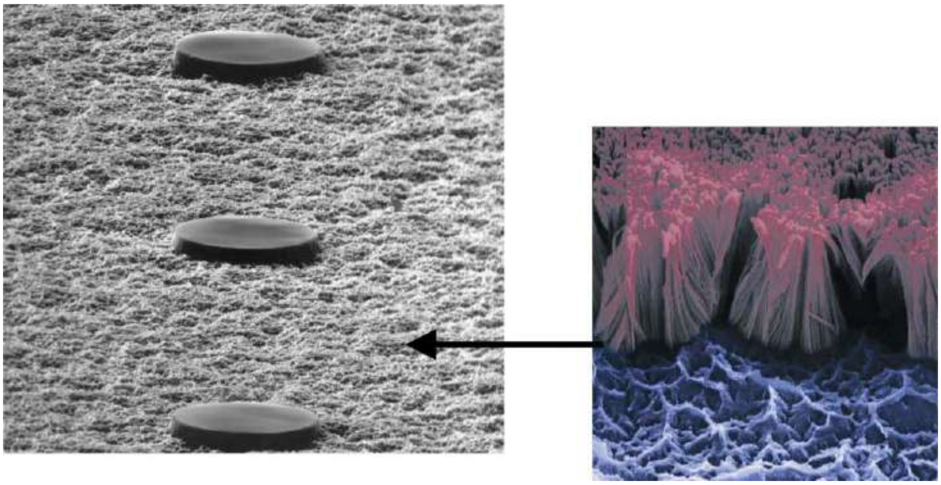

Figure 4 Three typical polymer dots of a few microns depth on an anodized aluminum substrate. Also shown is a further magnified electron microscope image of the substrate material, with the anodized layer cut away for presentation. The porous structure increases water retention. Oil-based ink is repelled from areas other than the dots. Figure courtesy of Agfa (Tony King, tony.king1@agfa.com).

heating stage completes cross-linking of polymer chains in the exposed portions of the emulsion. Chemical processing consists of processing the plate through an alkaline developer solution where the polymer that is not cross-linked washes away. Following the wash is a rinse and application of a coating. After chemical processing, some plates require baking to 200 °C (392 °F) for 35 minutes. Afterwards, the plate continues on to rinsing.

The machine tools required for these processing steps measure 30 feet long and take up to 7 minutes to process a plate after imaging. For both polymer and thermal plate processing, the temperatures are carefully controlled to prevent the aluminum substrate from deforming during the baking stages.

The effluent from thermal and other photopolymer plate processors is fairly benign and is typically directly disposed without significant environmental concern. Spent photopolymer chemistry can, however, pose problems with waste flow, it being a plastic sludge that can solidify and over time, clog plumbing systems.

Silver diffusion plate processors are relatively compact and tend to process plates in less than a minute. Environmental precautions are required with the disposal of accumulated silver particulates. Relatively simple systems are available, which neutralize the pH of the effluent and filter out silver until concentration levels are below 3 parts per million. Under these circumstances, most jurisdictions allow drain disposal.

Ablative plates do not require post processing, other than a rinse to remove residues. This is an emerging technology that has not taken hold at this time in the printing industry and requires more powerful lasers to ablate a layer from the plate. These plates use layers of ink-repellent silicon on top of an aluminum substrate to provide the non-inking portion of the plate. Variations of this technology use a laser to ablate the silicon from the substrate in the image areas of the plate, or to transfer material from a donor layer to the aluminum substrate. A weakness of this technology has been the limited durability of the plates in press. There is continuing development to produce ablative plates with other surface layers that are more durable on press.

Materials Processing – Heating

One way in which lasers can be applied to materials processing is as a directed source of heating. With material quickly raised in temperature to the point of melting and vaporization, applications lie in drilling, welding, cutting, and scintering.

The exploited characteristics of a laser source are its directionality and discrete color. Directionality enables the formation of a restricted area for heating as well as an extended working distance, whilst a discrete color enables absorption to be optimized for the material being worked. Laser technologies typical in application to materials processing through heating, range from diodes at 808 and 940 nm to flashlamp or diode pumped solid state lasers at 1 μm, and CO_2 gas lasers at 10 μm. A recent entry technology is in diode pumped fiber lasers, reaching output powers in the kilowatt regime at a wavelength of 1 μm and typically being continuous in emission or pulsed with a pulse duration in the millisecond regime.

Direct diode lasers (see Figure 5) have advantages in simplicity and robustness and with the minimum of additional assembled optical components, costs can be expected to continue to reduce, fed by Moore's law for the cost of semiconductor manufacture. Typical powers range from 1 to 10 kW and light can be fiber delivered to the work piece. A disadvantage of direct diode emission is the reduced directionality in comparison with alternative laser sources. For example, a current typical numerical aperture of emission from a diode bar array emitting a power of 60 W, is 10 and 60 degrees in directions parallel and perpendicular, respectively, to the array. These divergences can be reduced to a numerical aperture of approximately one degree with the addition of micro-optics, at a device cost increment of approximately 20%. For an array of diode bars with emitting powers in the kilowatt regime and with emitting dimensions sided 2 cm, a typical optical quality is one thousand time diffraction limited, translating to working distances are on the order of centimeters for a laser spot size at a work piece that is on the order of 1 mm.

CO_2 gas lasers and solid state lasers offer the advantage of reduced divergence and corresponding greater working distance, combined with a smaller laser spot size at the work piece for finer processing.

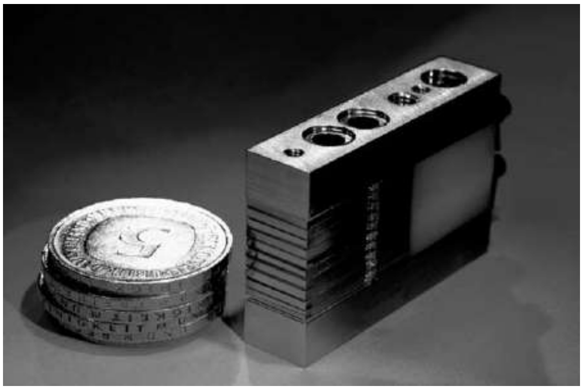

Figure 5 Diode bar array, with an output power exceeding 500 W. Courtesy of Rofin.

Both laser technologies can operate in the kilowatt regime, for application to sheet metal cutting and welding.

With the interaction process involving high temperatures and the ejection of material, a high-pressure gas bathes the area of illumination to remove smelt and either reduce exothermic chemical reactions, or where process acceleration is required, to utilize an assist gas to increase heating through exothermic reaction. Oxygen is typically used as an assist gas for the processing of low-alloy steels, with the process being typically known as laser flame cutting. The gas is, in general, injected coaxially with the laser beam. Laser flame cutting is used in industry for material up to 4 cm thick, albeit the width of the cut becomes significant at this thickness. Both CO_2 and Nd:YAG lasers are applied for this application, the choice of which is dependant on the cut geometry, the cycle time, and the material. Typical cutting speeds are, for example, in the region of 10 meters per minute for 1 mm, 3 meters per minute for 6 mm, and 1 meter per minute for 15 mm thick material.

In the case of high-alloy steels and aluminum, in particular, the material is bathed with an inert gas such as nitrogen or argon. The material is heated solely by the energy in the laser beam, and is typically referred to as laser fusion cutting (see Figure 6). The laser power required is higher than for laser flame cutting. Laser fusion cutting reduces oxidization to the cut edges, which is particularly important where welding is the next process step after cutting.

Laser fusion cutting is used in industry for material up to 25 mm thick, albeit the width of the cut becomes significant at this thickness.

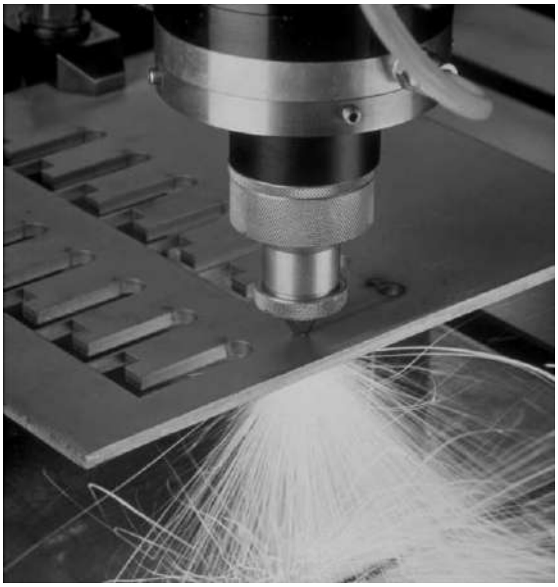

Figure 6 Laser fusion cutting. Courtesy of Rofin.

As with laser flame cutting, both CO_2 and Nd:YAG lasers are suitable for this application, the choice of which is influenced by the geometry of the cut, the cycle time, and the material. Typical cutting speeds in steel are, for example, in the region of ~8 meters per minute for 1 mm, 4.5 meters per minute for 3 mm and 1.5 meters per minute for 8 mm thick material.

Materials Processing – Localized Heating

As a material is heated in a localized volume to the point of vaporization, so also is the surrounding material, which can melt and experience cracking or changes in phase. Such heating is mainly considered deleterious, reducing process tolerances and reducing the tensile strength of the surrounding material.

As the pulse duration is reduced for a localized laser heating source, so are the dimensions that experience heating around the interaction volume (the heat affected zone), translating to improved part quality, with the length for thermal diffusion, l_{th} given by:

$$l_{\mathrm{th}} = \sqrt{Dt} \qquad [1]$$

where D is the thermal diffusivity and t, the time of diffusion. As the pulse duration is reduced, the pulse energy required to ablate material is also reduced, with the laser pulse energy required to raise the material temperature to the point of vaporization being proportional to the volume heated. For an example such as silver, a pulse duration of 200 fs is associated with a diffusion length of 6 nm. Extrapolating, a pulse duration of 100 ps, 100 ns, and 100 μs are associated with diffusion lengths of 0.1, 4, and 130 μm, respectively. For typical metals and dielectrics, the process quality and pulse energy threshold for ablation become decoupled from variation in pulse duration for pulse durations less than a few picoseconds. In this regime, the volume for heating is dictated more by the depth for optical absorption and further, the physical mechanism for absorption changes to that of material breakdown and multiphoton ionization (Figure 7).

Additional advantages of operating in the short-pulse regime are a sharpening of threshold to ablation, enabling the processing of features of dimension less than that of the laser spot size and an increase in the number of materials that can be processed by the same color laser source.

Given the advantages of the short-pulse regime, it should be noted that this is currently juxtaposed with the complexity and cost of the required laser sources. Solid-state laser technology is, however, advancing, significantly reducing the number of optical components required to deliver short pulses at multi-kHz

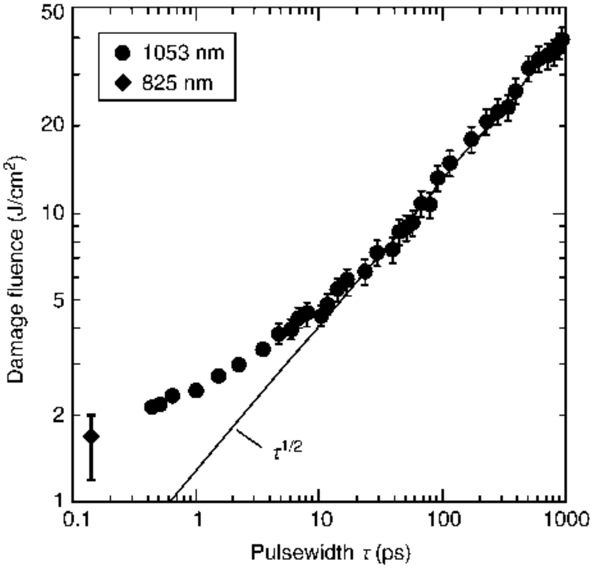

Figure 7 Pulse duration dependence of threshold damage fluence for fused silica. Reproduced with permission from Stuart BC, Feit MD, Herman S, Rubenchik AM, Shore BW and Perry MD (1996) Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. *Physics Review B* 53: 1749. Copyright (1996) by The American Physical Society.

repetition rates and multi-mJ pulse energies. At this time of evolution in laser technology, solid-state, diode-pumped high repetition rate (>10 kHz), high pulse energy (>40 mJ) nanosecond (1–200 ns) lasers at a wavelength of 1 μm are seeing increasing application. Such laser sources offer a combination of robustness and cost-effectiveness, combined with a process quality improvement over sources with longer pulse duration.

In the nanosecond regime, the nature of light-matter interaction remains dependent on the wavelength of the laser source and the material properties. In particular, wavelengths in the UV exhibit reduced pulse energy threshold and can offer improved process quality. UV wavelengths less than 300 nm are currently produced by excimer and ArF gas lasers, emitting at 248 and 193 nm, respectively, or solid-state sources that are wavelength converted with nonlinear optical parametric materials to emit at 354 or 256 nm. Significant laser development effort is underway to improve optical parametric materials for nonlinear conversion, to enable greater robustness for solid-state UV sources that then offer reduced overall laser tool cost and complexity.

Semiconductor Lithography

The advantage of nanosecond and shorter pulsed lasers for materials processing extends from localization of heating. With heating being localized, a greater temperature can be achieved. This behavior is fundamental to the production of EUV radiation for application to next-generation lithographic processing.

EUV lithography is considered as an attractive candidate to succeed deep DUV-based optical lithography within this decade. EUV wavelengths around 13 nm will allow the reduction of structure sizes to less than 80 nm in semiconductor devices.

For such a high-power light source that emits at a wavelength of 13 nm, solutions are currently being pursued by multiple industry players in electrical plasma discharge and laser produced plasmas (LPP) (Figure 8). Electron beam and synchrotron sources could, in principle, satisfy the application requirements, but suffer from restriction relating to cost, size, and throughput. The choices of technologies remain to be finalized. It is, however, accepted across the industry that Moore's Law should remain intact and that a solution is required that is scalable to high throughput rates for wafer exposure.

For a laser-produced plasma, the peak wavelength of the light emitted from a plasma is determined by its temperature, T, where according to the blackbody formula:

$$T \sim E_{\mathrm{ph}}/3 \qquad [2]$$

where E_{ph} is the photon energy associated with the wavelength of maximum emission. Liner photopolymerization (LPP) schemes use pulsed laser radiation to heat a target to the optimum temperature required for efficient EUV production. For plasma-emitted light which peaks at a wavelength of 13.5 nm, the optimum temperature is approximately 3.5×10^5 Kelvin (30 eV).

A laser is focused onto a target, transferring its energy into the target in such a way that a plasma is formed and heated. A range of targets have been investigated for use in LPP schemes ranging from gaseous and cluster targets, to liquid targets and macroscopic solid targets. The interaction begins with the production of free electrons via multiphoton ionization, which is strongly dependent on the peak intensity of the laser. As heating progresses, there is further ionization from inelastic collisions between heated electrons and ions to reach a pseudo-equilibrium state, where cooling competes with heating on nanosecond timescales. Given this time-scale for cooling and requirements on source size and temperature, the optimum laser pulse duration lies in the regime of nanoseconds and the optimum laser pulse energy is hundreds of millijoules. Laser technology

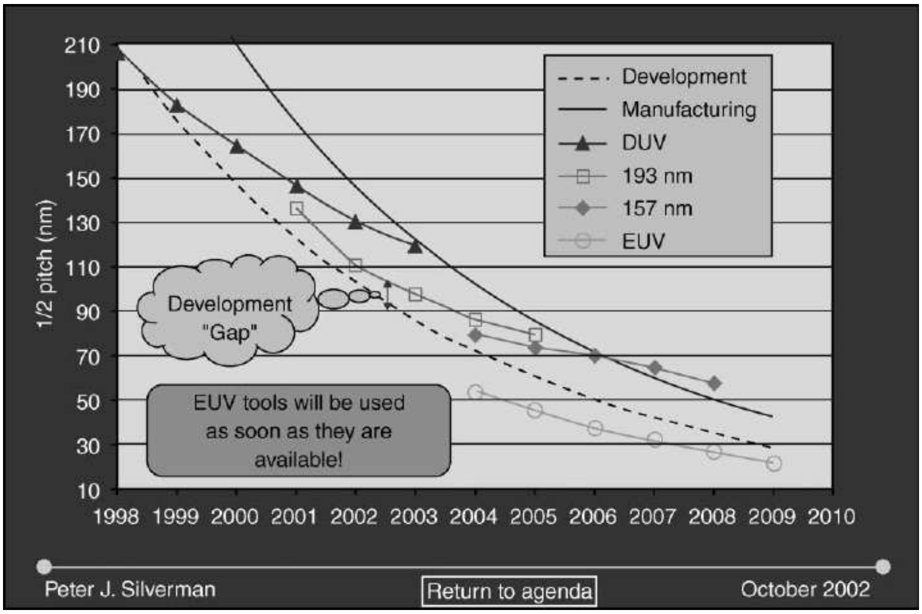

Figure 8 Light source roadmap to 2010. Courtesy of Intel Corporation.

suited to this application is diode-pumped Nd:YAG, emitting at a wavelength of 1 μm and pulsed at multikilohertz repetition rates.

Fine Control Material Removal

Localized heating afforded with the application of lasers with pulse durations in the nanosecond regime and shorter, is seeing increasing takeup to replace processes ranging from wet etching and electro-discharge machining (EDM) in applications such as scribing, drilling, and differential layer removal. Laser methods are seeing advantage in direct ablation of material, for feature sizes on the order of microns to hundreds of microns. Large areas on the order of 2 m^2 are now becoming cost-effective to process with direct laser ablation, with the advent of high average power laser sources of hundreds of watts of average power, combined with nanosecond regime or shorter pulse durations. Typical sources are gas excimer or diode-pumped solid-state.

The finest processing is advantaged by the application of pulse durations in the picosecond regime. A restriction in application of this technology has been robustness and cost of the laser sources to date; however, such technology is continuing to see rapid improvement and may be expected to see wider takeup in industry by 2010.

Also seeing application, are wavelengths in the UV regime combined with nanosecond pulse durations. UV radiation in general offers a reduced threshold of ablation, translating to reduced heating and improved process quality. High absorption at UV wavelengths in most materials, combined with photon energies that can exceed bond energies of the material, contribute to the removal of material with the minimum of heating. Excimer, KrF, and ArF gas lasers are applied here (Figure 9), but it is expected that solid-state lasers will emerge by 2010 to capture a large part of the UV market. CO_2 gas lasers are used for fine processing where the material being ablated is of low thermal diffusivity, which is typical for dielectrics or ceramics. Materials of low thermal diffusivity are associated with a low heat affected zone (see eqn [1]).

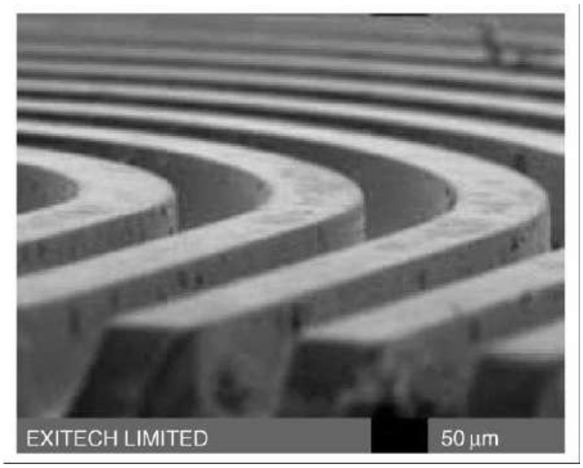

Figure 9 KrF laser micromachined microfluidic channels in polyester for application to lab-on-a-chip technologies or sensing devices. Courtesy of Exitech Ltd.

Scribing is required for dicing of wafers such as silicon and other low-*k* materials, both for computing device manufacture and for solar cell manufacture. Nanosecond and shorter pulsed lasers have seen recent application in this area, enabling higher aspect ratio scribing and a reduced material affected zone. This application has seen laser wavelengths applied from the UV to near-infrared. Short pulse laser scribing can see advantages of high throughput yet reduced or eliminated chipping, delamination, and film peeling.

Fine drilling is required for a wide variety of applications and markets ranging from constructing vias for through-layer interconnects in printed circuit board (PCB) fabrication, to ink jet nozzles (Figure 10) and high aspect ratio holes in hard materials, such as steel or alloys for application to combustion engines.

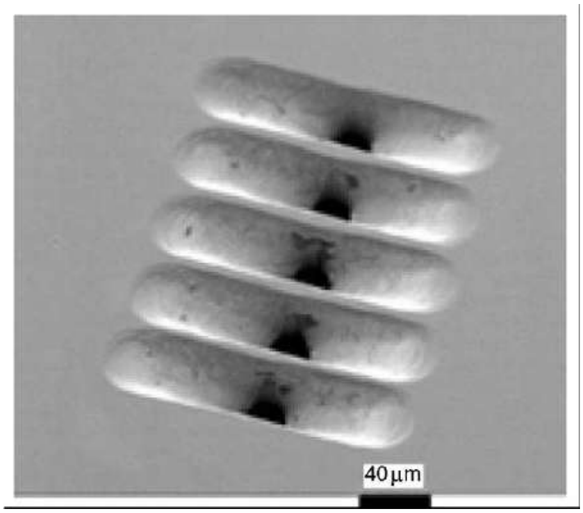

Figure 10 A laser drilled hole in steel and an array of shaped holes in polyamide for inject nozzle application. Courtesy of Exitech Ltd.

Excimer and CO_2 laser drilling has seen industrial application in the drilling of via holes for electronics interconnection packages, since early investigations in the 1980s. More recently, high repetition rate diode-pumped solid-state lasers, nonlinearly converted to emit in the UV, are seeing application. In parallel with the increasing density of transistors in silicon processors, the density of packaging of multichip modules (MCM) and PCB components has also been increasing, requiring the drilling of blind vias on the order of 100 microns in diameter to enable interconnects between multiple conducting layers (Figure 11). Also, as processor power increases, so does the number of connections required to devices. Microvias have enabled the application of two-dimensional grid array connections to the underside of devices, reducing overall package size, including processor electrical connections, by a factor of five.

Laser drilling of microvias in PCBs is expected to increase as packing density increases. PCBs remain cheaper than MCMs and traditionally have been mechanically drilled. As via sizes are reduced below 100 μm, however, it is expected that pulsed CO_2 and Nd:YAG lasers will be applied.

Thin layer ablation is required for applications ranging from surface cleaning to display manufacture and for polymers, dielectrics, and metals. For example, manufacture of plasma displays currently incorporates lithography combined with wet etching, to pattern layers of thin conducting oxide onto a substrate. Multiple process steps can be replaced by application of direct laser ablation and increased laser average power combined with pulse durations in the nanosecond regime are allowing millions of cells to be processed per minute, with the accommodation of process areas currently greater than 2.5 m^2 (see Figure 12).

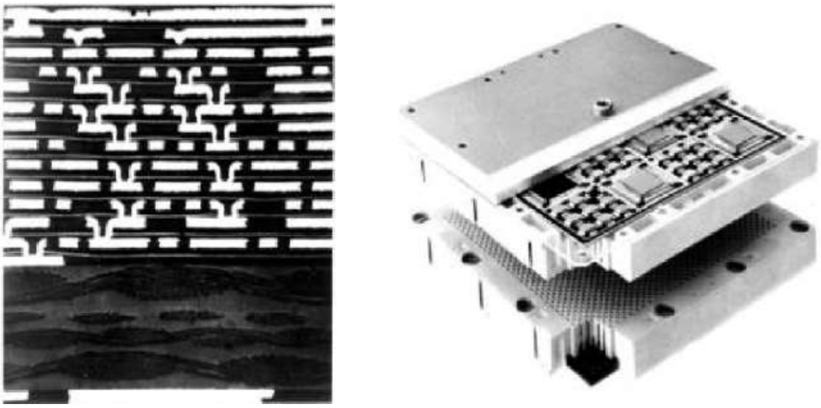

Figure 11 Cross-section of 14-layer copper plated MCM board showing 80 μm diameter microvias drilled by a KrF laser in acrylic resin-polyimide and, fully assembled MCM package. Courtesy of Siemens AG.

Laser-Based Diagnostics

With a continuing increase in reliability, performance, and range of specifications, lasers are seeing growing industrial application for process monitoring and diagnostics, ranging from gas phase detection of molecular specific concentrations to *in vitro* microscopy in the life sciences. Markets are emerging, driven in a large part by advances in laser technology. Chemical sensing with application of lasers is a vast field, which is difficult to cover here completely. Following, are some applications seeing current industrial application and growth.

Microscopy

Confocal microscopy has seen large uptake, in particular, for the application to imaging of cell and membrane function, with more than fifteen manufacturers now supplying confocal microscope tools including Zeiss, Biorad, Leica, Olympus, and Nikon. Confocal microscopy enables increases in resolution and contrast compared to standard microscopy. Techniques such as fluorescence resonant energy transfer (FRET), fluorescence recovery after photobleaching (FRAP), fluorescence loss in photobleaching (FLIP), fluorescence lifetime imaging (FLIM), or fluorescence correlation spectroscopy (FCS), where married with confocal microscopy and, in particular, where relevant molecular structures are known in advance, can offer the additional advantages of specific molecular and structural specificity through the incorporation of tracer fluorophores. This is seen as color coding of the resultant image, with sophisticated microscopes currently offering differential imaging stimulated by three laser colors and with six selectable colors for detection.

The confocal microscope operates with advantage over a standard microscope, by illuminating only a small area in the sample with dimensions of the resolution limit of the objective and re-imaging emission from that area through a pinhole spatial filter (Figures 13 and 14). Structures outside of the illuminated area or at a depth outside that of the focus, are suppressed from detection. Fluorophores (tracers) can be inserted into protein sequences at predetermined locations and are stimulated by either single or two-photon absorption of the stimulating laser light.

Two photon absorption has the additional advantages of increased resolution and detected contrast over single photon absorption. Laser sources applied are generally modelocked oscillators operating near 800 nm, with a pulse duration on the order of 100 fs. This technology is continuing to improve in robustness, with fiber and direct diode pumped solid-state sources available from multiple vendors.

FRAP and FLIP monitor diffusion or transport of fluorescent labeled molecules into or out of the laser focal volume, respectively. FCS monitors the fluctuations in detected signal from emissions within the laser focal volume, to determine the rate of diffusion with sensitivity achievable to the single molecule level (Figure 15).

FRET is seeing application for the probing of protein interactions in cells or at cell membranes. FRET involves the nonradiative transfer of excitation energy from one fluorophore to another, which subsequently emits at a different wavelength. The fluorophores are inserted into molecules, the

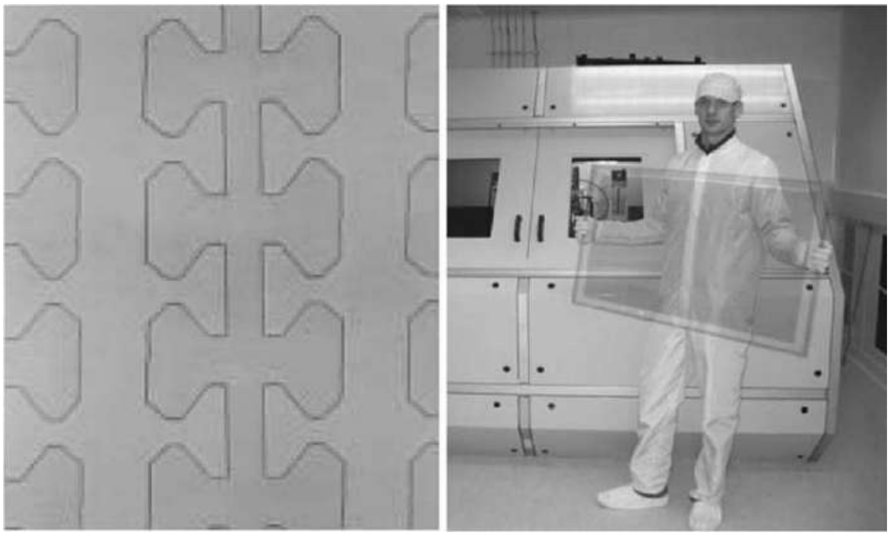

Figure 12 Plasma display manufactured by direct laser ablation of thin-film indium tin oxide (ITO). Courtesy of Exitech Ltd.

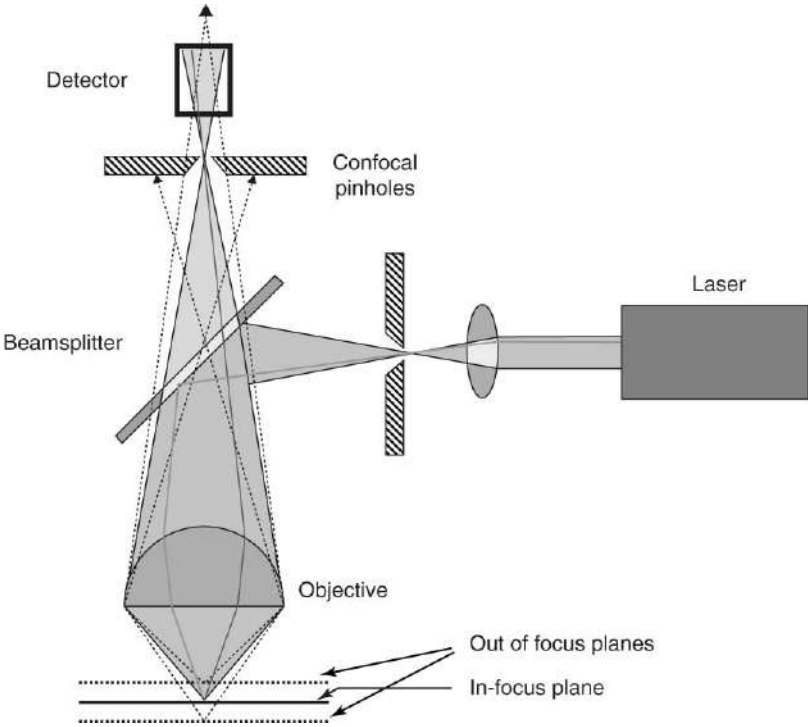

Figure 13 Confocal microscope principle of operation. Courtesy of Leica Microsystems Heidelberg GmbH.

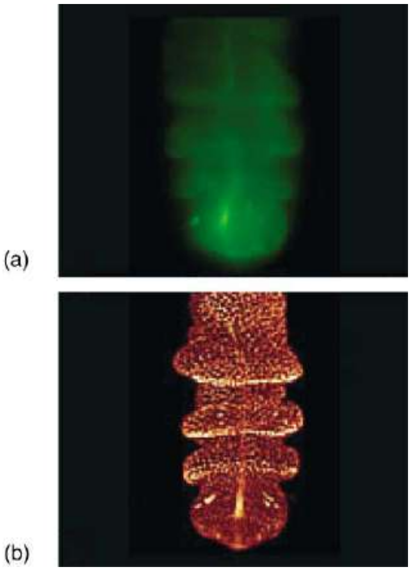

Figure 14 Image (a): *Drosophila melanogaster*, fluorescence nonconfocal configuration and Image; (b): *Drosophila melanogaster*, fluorescence confocal 3D projection. Courtesy of Leica Microsystems Heidelberg GmbH.

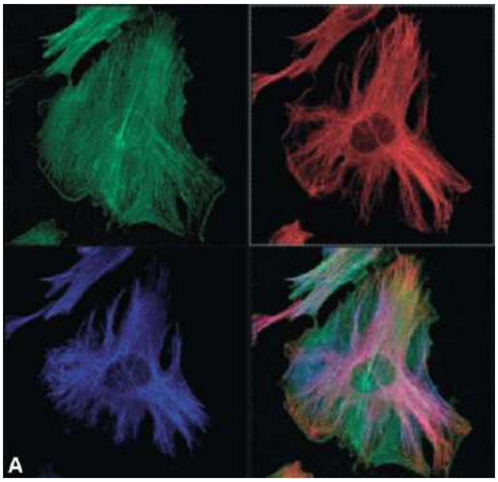

Figure 15 Spectral selective imaging with three fluorescence channels showing triple labeled mouse fibroblast: green: actin fibers, red: microtubules, blue: vimentin. Courtesy of Leica Microsystems Heidelberg GmbH.

interaction of which is being investigated. FRET effectively enables sensitivity to relative position at fluorophore separations less than 10 nm; a significant advance over the 200 nm resolution provided by a standard optical microscope. Concentrations of binding events can be monitored over time (Figure 16).

Raman Methods and Microscopy

Raman spectroscopy sees broad application in chemistry in the analysis of chemical composition and molecular structure, being used, for example, in pharmaceuticals, polymer materials science, bioclinical science, and forensic science.

The history of Raman scattering stems from theoretical predictions in the early 1920s,

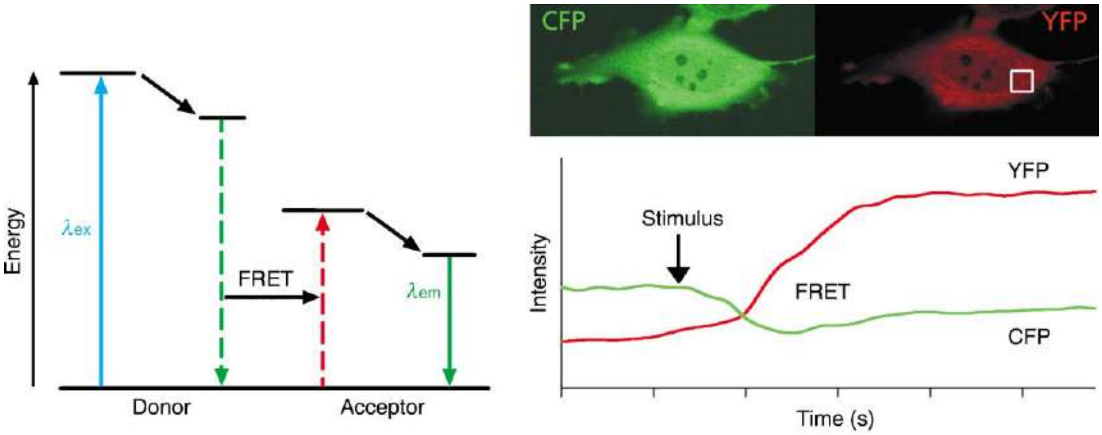
Energy
λ_{ex}
FRET
λ_{em}
Donor
Acceptor
CFP
YFP
Intensity
Stimulus
YFP
FRET
CFP
Time (s)

Figure 16 Fluorescence resonant energy transfer mechanism and example relating to cytoplasmic Ca^{2+} concentration. Cyan Fluorescent Protein (CFP) and Yellow Fluorescent Protein (YFP) are acceptor and donor, respectively, for the FRET. Courtesy of Carl Zeiss, Germany.

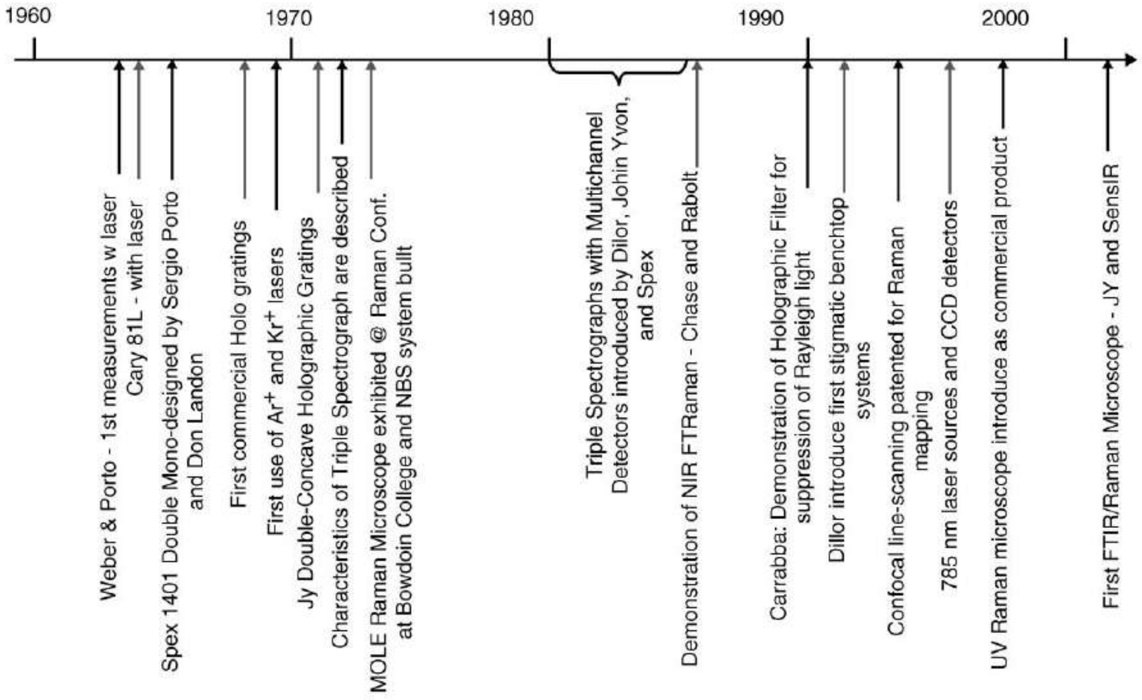
1960
1970
1980
1990
2000
Weber & Porto - 1st measurements w laser
Cary 81L - with laser
Spex 1401 Double Mono-designed by Sergio Porto and Don Landon
First commercial Holo gratings
First use of Ar^+ and Kr^+ lasers
Jy Double-Concave Holographic Gratings
Characteristics of Triple Spectrograph are described
MOLE Raman Microscope exhibited @ Raman Conf. at Bowdoin College and NBS system built
Triple Spectrographs with Multichannel Detectors introduced by Dilor, Johin Yvon, and Spex
Demonstration of NIR FTRaman - Chase and Rabolt
Carrabba: Demonstration of Holographic Filter for suppression of Rayleigh light
Dillor introduce first stigmatic benchtop systems
Confocal line-scanning patented for Raman mapping
785 nm laser sources and CCD detectors
UV Raman microscope introduce as commercial product
First FTIR/Raman Microscope - JY and SensIR

Figure 17 Timeline of Raman spectroscopy from the first incorporation of a laser source. Courtesy of Jobin Yvon.

accompanied also in the 1920s by early measurements, notable being those of C.V. Raman for which he was awarded the Nobel prize, soon followed in the 1930s by hundreds of chemistry papers describing this technique. Raman used for his early experiments filtered sunlight, a prism spectrometer, and visual observation. This was followed by application of filtered Hg emission and photographic plates. Lasers first saw application to Raman scattering in the 1960s, shortly after their innovation (see Figure 17).

In the simplest form of application of the Raman effect, UV or visible radiation illuminates a sample, coupling to excite electronic transitions near or on resonance. Spontaneous emission is shifted to a longer wavelength than that of the excitation (Stokes emission) and this is filtered and detected. In general, a broad spectrum of fluorescence is superimposed by weaker, narrow features that originate from vibration-electronic coupling (Raman emission). Inclusion of laser excitation increases the rate of data collection and also enables nonresonant interaction of the stimulating radiation with the sample whilst maintaining data collection rates. Nonresonant interaction preferentially reduces electronic emission of fluorescence relative to that of the narrow featured Raman emission. Lasers applied to spontaneous Raman spectrometers in general emit in the visible or near-infrared running nonpulsed, such as Ar-ion, Nd:YAG or He:Ne.

Further enhancements are made to deliver spatial as well as vibration-specific information, through tight focusing of the stimulating radiation on to the sample in a confocal microscope arrangement, delivering spatial resolution down to 200 nm (Figure 18).

A limitation of standard Raman spectrometers and microscopes is the rate of signal collection, requiring concentrated samples for interrogation. The restricted signal level derives from a typical cross-section for the spontaneous emission of Raman radiation being $10^{-30}\,\mathrm{cm}^2$; 10^{10} to 10^{14} times less than that typical for spontaneous electronic fluorescence. Enhancements in the level of signal and the ratio of signal to background noise may be made through coherent laser stimulation of anti-Stokes Raman emission, known as coherent anti-stokes Raman scattering (CARS) (Figures 19 and 20).

It can be anticipated that bond-specific imaging will see wide industrial application in future, employing Raman as well as resonant infrared techniques. Application of CARS to microscopy involves two laser colors. A vibration transition is exited at the beat frequency of the two laser frequencies and the shorter laser wavelength is subsequently scattered from the sample to a yet shorter wavelength of emission. Such a coherent process has the advantage of signal scaling proportional to the square of emitters at the sample, thereby increasing the signal over that of spontaneous Raman scattering. Typical laser pump and Stokes powers range from less than 1 mW to 10s of mW, with laser repetition rates ranging from multi-kHz to greater than 100 MHz and typical pulse durations from 10s of picoseconds to 100 fs.

Infrared Methods

Infrared absorption spectroscopy provides a wealth of information on bond-specific sample concentration

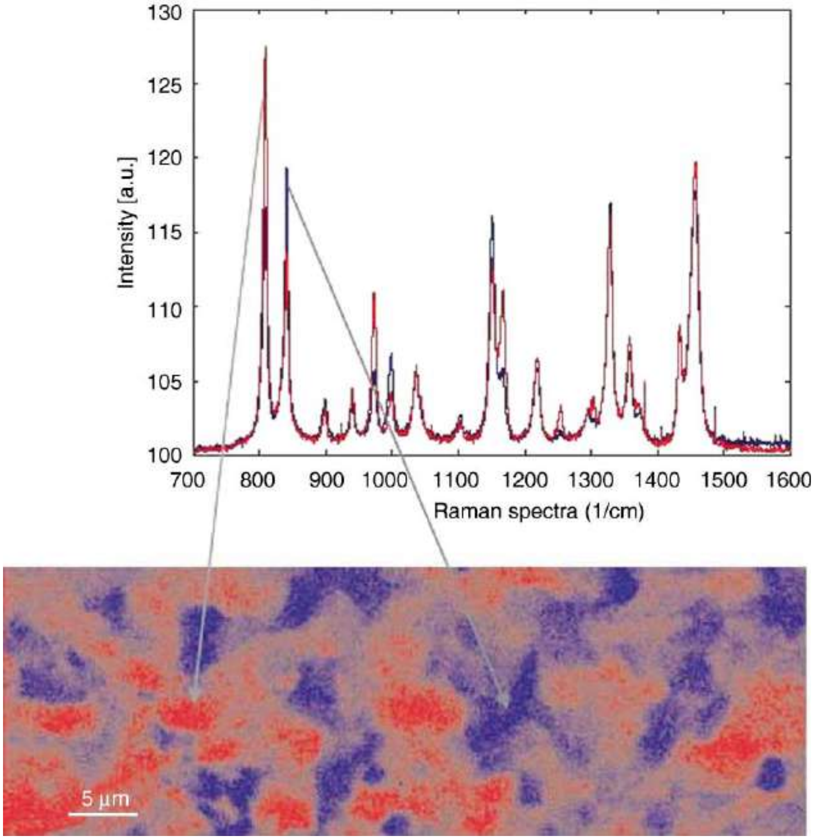

Figure 18 Confocal Raman image of semicrystalline polypropylene film. Two typical Raman spectra are shown. The ratio of the Raman lines at 809 and 841 cm increases with the degree of crystalline order, delivering a two-dimensional image (Hendra *et al.* 1995). Courtesy of Witec GmbH.

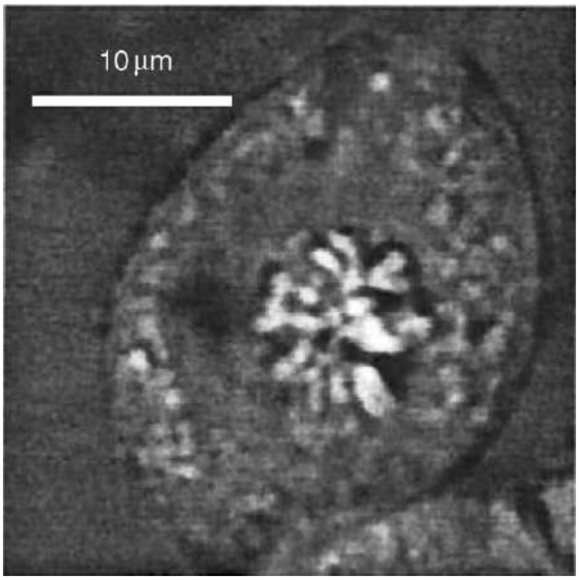

Figure 19 CARS image of a NIH 3T3 cell in metaphase, where Raman vibration stimulation is tuned to the PO_2^- symmetric stretching vibrational frequency at 1090 cm^{-1}. Reproduced from Cheng J-X, Kevin JYK, Zheng G and Xie X-S (2002) Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology. *Biophysical Journal* 83: 502–509, with permission from The Biophysical Society.

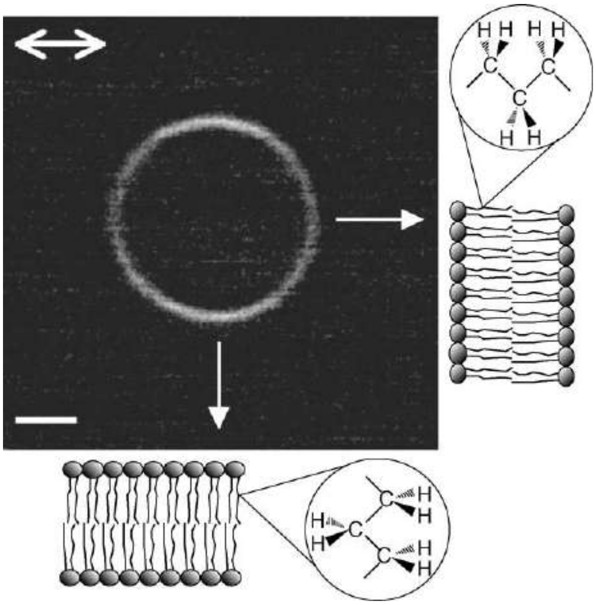

Figure 20 CARS microscope image of an erythrocyte ghost, consisting of a single lipid bilayer, with a Raman shift of 2845 cm^{-1}. The scale bar is of length 2 μm. Reproduced from Potma EO and Xie X-S (2003) Detection of single lipid bilayers with coherent anti-Stokes Raman scattering (CARS) microscopy. *Journal of Raman Spectroscopy* 36: 642–650, with permission from Wiley InterScience.

and has become a standard tool for the chemistry laboratory. Interaction between the radiation and sample is via resonant vibrational absorption and is otherwise analogous to the operation of a standard optical absorption spectrometer. Applied to a microscope arrangement, resolution on the order of 10 μm is standard. One restriction associated with infrared absorption spectroscopy applied to solution, is that absorption lineshapes are broadened and the spectra lose detail at room temperature, hindering identification of the contributing species.

A technique showing early promise for application in bond-specific molecular detection in solution, is coherent resonant infrared wavemixing. Here, cross coupling between vibrational modes are detected, promising information more detailed than that of infrared absorption. Figure 21 shows an example of triply vibrationally enhanced four-wave mixing, where two infrared laser beams mix to deliver cross peaks representing vibration mode coupling. Here, the 2D spectrum contains two sets of cross-peaks corresponding to the couplings of the two strong IR modes at 1943 and 2002 cm^{-1} of $Ni(CO)_2(PPh_3)_2$ and 1996 and 2068 cm^{-1} of $Ni(CO)_3PPh_3$. Concentrations in solution are on the order of 10 mM. Such methods provide a means sensitive to chemical structure without a requirement for fluorophores.

Current application requires complex laser equipment delivering microjoule pulse energies, tunable between 3 and 10 microns, with pulse durations on the order of 1 ps. However, following the example of other laser diagnostic techniques in application

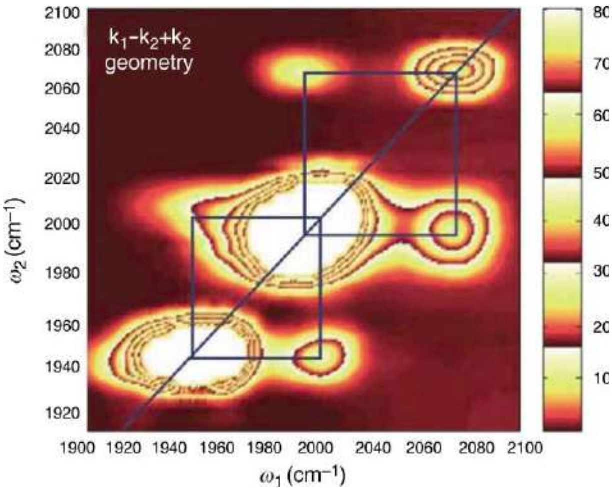

Figure 21 Triply vibrationally enhanced four-wave mixing of $Ni(CO)_2(PPh_3)_2$ and $Ni(CO)_3PPh_3$. Reproduced from Meyer KA, Besemann DM and Wright JC (2003) Coherent two dimensional spectroscopy with triply vibrationally enhanced infrared four-wave mixing. *Chemical Physics Letters* 381: 642–649, with permission of Elsevier.

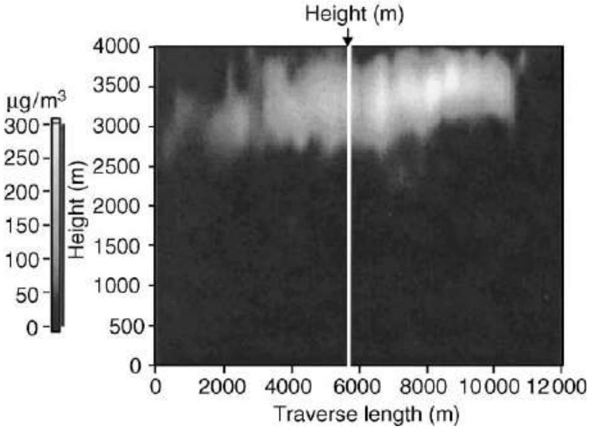

Figure 22 DIAL trace of sulfur dioxide distribution from Mount Etna. Reproduced with permission from Weibring P, Edner H and Svanberg S (2003) Versatile mobile lidar system for environmental monitoring. *Applied Optics* 42: 3583–3594. © 2003 Optical Society of America.

across industry today, as laser technology advances over this decade, we can expect such methods of probing chemical systems to see significant uptake, in particular in the bio-industries.

Laser Diagnostics and Treatment of Tumors

For clinical medicine, fluorescent detection is applied in the visualization of malignant tumors. Important natural chromophores emitting fluorescence in the blue-green region are elastin, collagen, NADH, and NAD^+, whilst administered agents such as porphyrins and phthalocyanines yield sharp peaks in the dark red.

Excitation is via UV or violet light. Where emission is monitored in both the blue-green and red regions of the spectrum and the ratio taken, an enhanced tumor differentiation is possible from that of normal tissue whilst simultaneously reducing sensitivity to geometry, illumination, and detection efficiency. Fiber optic probes are generally applied for delivery and collection of radiation.

Such detection can be combined with techniques of laser photodynamic therapy (PDT) to preferentially remove cancerous tissues. Typical laser wavelengths for photodynamic therapy are in the visible region of 600 to 700 nm with a pulse fluence of tens of joules/cm^2, a pulse duration in the nanosecond regime and a repetition rate from single shot to 10 Hz. Where combined with diagnostic detection, the PDT can be carried out by exciting the same administered sensitizer, with subsequent transfer of ground-state triplet oxygen to a toxic singlet state.

Remote Chemical Sensing

The atmosphere can be monitored by absorption and laser-induced fluorescence (LIF). Three-dimensional mapping of chemical distribution can be achieved with light detection and ranging (LIDAR), using a pulsed laser as a transmitter and an optical telescope as a receiver. Information collated includes pollution concentration, temperature, humidity, and wind velocity. Strong quenching prohibits LIF at atmospheric pressures; however, it does become practical to monitor for example, Li, Na, K, and Ca layers in the mesosphere. Pollution monitoring in the troposphere is most frequently performed with differential absorption LIDAR (DIAL). Range resolved optical transients are taken as a ratio of two laser wavelengths; on and just off resonance for the species of interest. Typical ranges for detection of SO_2, O_3, NO_2, NO, and Hg are 500 m to 5 km, making this technique suitable for urban and industrial monitoring. Sulfur dioxide is the most common pollutant emitted when fossil fuels are burned or sulfide ores are roasted and this technique sees application for the monitoring of industrial emissions. Figure 22 shows a DIAL trace recorded from Mount Etna, from which emissions of SO_2 are estimated at 75 tonnes/h. Laser wavelengths used are 300.02 and 299.30 nm for on and off resonance for absorption, respectively.

Important to the widespread application of LIDAR and LIF for atmospheric chemical monitoring, has been the advance in robustness and flexibility of the associated laser technologies. The laser technology applied for the above measurement is based on solid-sate Nd:YAG, nonlinearly converted via optical parametric amplification to wavelengths tunable between 220 nm and 4 µm, delivering up to 20 mJ in the mid-infrared at a repetition rate of 20 Hz and with bandwidths of less than 0.2 cm^{-1}, in a pulse duration of 4 ns.

See also

Imaging: Infrared Imaging. **Lasers:** Carbon Dioxide Lasers. **Microscopy:** Confocal Microscopy. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Allen SD, Bass M and Teisniger ML, Comparison of pulsed Nd:YAG and pulsed CO2 lasers for hole drilling in printed circuit board materials. CLEO 1982 Conference Summary.

Cheng J-X, Kevin Jia YK, Zheng G and Xie X-S (2002) Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology. *Biophysical Journal* 83: 502–509.

Cymer Inc., http://www.cymer.com

Cymer Inc., ASML, TRW, Xtreme Technologies GmbH, Fraunhoffer Institut fur Lasertechnik, Philips Extreme UV GmbH, Royal Institute of Technology Sweden, PLEX LLC.

Disco Corporation, http://www.disco.co.jp/

Exitech Ltd., http://www.exitech.co.uk/

Lamda Physik, http://www.lambdaphysik.de/

Meyer KA, Besemann DM and Wright JC (2003) Coherent two dimensional spectroscopy with triply vibrationally enhanced infrared four-wave mixing. *Chemical Physics Letters* 381: 642–649.

Perkin Elmer, www.perkinelmer.com

Phillips D (1997) Chemical mechanisms in photodynamic therapy with phthalocyanines. *Progress in Reaction Kinetics* 22(3–4): 175–300.

Potma EO and Xie X-S (2003) Detection of single lipid bilayers with coherent anti-Stokes Raman scattering (CARS) microscopy. *Journal of Raman Spectroscopy* 34: 642–650.

Powerlase Ltd., http://www.powerlase.com

Presstek Inc., Hudson, NH, http://www.presstek.com

Pronko PP, Dutta DK, Squier J, Rudd JV, Du D and Mourou G (1995) Machining of sub-micron holes using a femtosecond laser at 800 nm. *Optics Communications* 114: 106–110.

Reiser A (1989) *Photoreactive Polymers – The Science and Technology of Resists*. John Wiley & Sons.

Silverman PJ (2002) Intel Corporation. Presented at 1st Intl EUV Lithography Symposium, 15–17 October 2002, Dallas, USA.

Stuart BC, Feit MD, Herman S, Rubenchik AM, Shore BW and Perry MD (1996) Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. *Physics Review B* 53(4): 1749.

Svanberg S (1996) Chemical sensing with laser spectroscopy. *Sensors and Actuators B* 33: 1–4.

Wang I, Bendsoe N, Klinteberg CAF, *et al.* (2001) Photodynamic therapy vs. cryosurgery of basal cell carcinomas: results of a phase III clinical trial. *British Journal of Dermatology* 144(4): 832–840.

Watson MN (1984) Laser drilling of printed circuit boards. *Circuit World* 11: 13.

Waynant RW (ed.) (2002) *Lasers in Medicine*. CRC Press.

Weibring P, Edner H and Svanberg S (2003) Versatile mobile lidar system for environmental monitoring. *Applied Optics* 42(18): 3583–3594.

WITEC, WITec Wissenschaftliche, Instrumente und Technologie GmbH, Hörvelsinger Weg 6, D-89081 Ulm, Germany, www.WITec.de

Yvon J, www.jobinyvon.com

Sum-Frequency Generation at Surfaces

M B Raschke and Y R Shen, University of California, Berkeley, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Under sufficiently intense illumination the optical properties of a material system depend nonlinearly on the strength of the electromagnetic field. Related to this nonlinear optical response is a large number of phenomena and fundamental processes which are discussed in various articles of this encyclopedia. This article will focus on second-harmonic generation (SHG) and sum-frequency generation (SFG) for the investigation of surfaces and interfaces. The intrinsic surface sensitivity of these techniques allows for investigations of surface properties not readily accessible by other spectroscopies. Here, the basic principles of these optical processes as well as their experimental implementation are discussed, and a summary of the applications to different material systems is given.

Among the various techniques employed for the characterization of surfaces and interfaces, those using light are particularly attractive. They are applicable *in situ* to all interfaces accessible by light, are nondestructive, and offer unprecedented time resolution. However, the penetration depth of optical radiation in matter is generally of the order of a wavelength, which makes isolation of the surface or interface contribution to the optical response from the bulk contribution difficult. In contrast, for reasons of symmetry, SHG and SFG are intrinsically surface sensitive in media with inversion symmetry, and hence the signal generated mainly originates from the topmost surface layer where the inversion symmetry is broken. By means of electronic or vibrational SHG or SFG spectroscopy, information on surface structure, chemical composition and bond or molecular orientation at solid and liquid interfaces can be deduced. To date, SFG and SHG have been well established as important tools for the investigation of surfaces and interfaces of solids ranging from metals and semiconductors to insulators and magnetic materials, liquids, polymers, biological membranes, and other systems. The studies are motivated by fundamental interests as well as applications in many areas such as heterogeneous catalysis, electrochemistry, device fabrication, epitaxial growth, and environmental science.

Theory

The nonlinear optical response of a material, in the electric-dipole approximation, can often be described by an induced polarization in the form of a power series,

$$\begin{aligned} \boldsymbol{P}(\omega) = \epsilon_0[\chi^{(1)}(\omega)\boldsymbol{E}(\omega) \\ + \chi^{(2)}(\omega = \omega_1 + \omega_2)\boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2) \\ + \chi^{(3)}(\omega = \omega_1 + \omega_2 + \omega_3)\boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2)\boldsymbol{E}(\omega_3) + \ldots] \end{aligned} \qquad [1]$$

where $E(\omega_i)$ is the optical field with frequency ω_i and $\chi^{(n)}$ is the nth-order linear ($n = 1$) or nonlinear ($n > 1$) susceptibility tensor. The second term, being the lowest-order nonlinear optical response, is responsible for sum-frequency generation and second-harmonic generation (with $\omega_1 = \omega_2$). In the electric-dipole approximation, all even-order terms ($\chi^2, \chi^{(4)}$, etc.) are forbidden in media with inversion symmetry. At the surface or interface, however, the inversion symmetry is necessarily broken and hence $\chi^{(2)} \neq 0$. This makes SHG and SFG surface-sensitive and specific. While electric-quadrupole and magnetic-dipole contributions from the bulk may not be totally negligible, in many cases it has been shown that the surface contribution clearly dominates.

Surface SFG and SHG is best described by radiation from a surface polarization $\boldsymbol{P}_s^{(2)}(\omega)$ induced in a thin sheet with dielectric constant ε' sandwiched between two linear media (1) and (2) as shown schematically in Figure 1. This surface polarization has an in-plane wavevector component equal to the sum of the in-plane wavevector components of the incident fields: $\boldsymbol{k}_\parallel(\omega) = \boldsymbol{k}_\parallel(\omega_1) + \boldsymbol{k}_\parallel(\omega_2)$. The ratio of $\boldsymbol{k}_\parallel(\omega)$ to $\boldsymbol{k}(\omega)$ then determines the direction of the sum-frequency output that appears in both transmission and reflection. The reflection geometry is often used in order to minimize the bulk contribution.

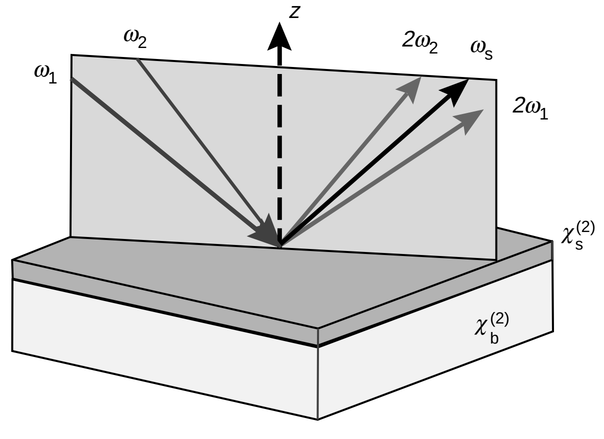

Figure 1 Schematic of SHG and SFG geometry in reflection and transmission from an interface.

With $\boldsymbol{P}_s^{(2)}$ as the source term, the sum-frequency output can be obtained from the solution of the wave equation. With the proper boundary conditions the sum frequency (SF) intensity is given by:

$$I(\omega) = \frac{8\pi^3 \omega_s^2 \sec^2\theta_\omega}{\hbar c^3 \sqrt{\epsilon_1(\omega)\epsilon_1(\omega_1)\epsilon_1(\omega_2)}} \times |\boldsymbol{e}^\dagger(\omega)\cdot\boldsymbol{\chi}_s^{(2)} : \boldsymbol{e}(\omega_1)\boldsymbol{e}(\omega_2)|^2 I_1(\omega_1) I_2(\omega_2) \qquad [2]$$

In this expression, $\chi_s^{(2)}$ is the surface nonlinear susceptibility defined by $\boldsymbol{P}_s^{(2)} = \varepsilon_0 \boldsymbol{\chi}_s^{(2)} : \boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2)$; the quantities $\boldsymbol{e}(\omega_i)$ correspond to the unit polarization vectors after the appropriate Fresnel correction, $\boldsymbol{e}(\omega_i) \equiv \boldsymbol{F}(\omega_i)\hat{\boldsymbol{e}}(\omega_i)$, with $\boldsymbol{F}(\omega_i)$ being the transmission Fresnel factor and $\hat{\boldsymbol{e}}(\omega_i)$ the unit polarization vector of $\boldsymbol{E}(\omega_i)$; θ_ω denotes the SF output angle with respect to the surface normal, $\varepsilon(\omega_i)$ is the dielectric constant at frequency ω_i, and I_i is the input pump intensity at ω_i.

The surface nonlinear susceptibility $\chi_s^{(2)}$ is a third-rank tensor. In Cartesian coordinates it has 27 tensor elements ($\chi_{ijk}^{(2)}$), many of which could vanish or depend on others due to the surface structural symmetry. As an example, $\chi_{ijk}^{(2)}(\omega = \omega_1 + \omega_2)$ for an isotropic surface, with the z-direction defined by the surface normal, has only four independent nonvanishing elements: $\chi_{xxz}^{(2)} = \chi_{yyz}^{(2)}, \chi_{xzx}^{(2)} = \chi_{yzy}^{(2)}, \chi_{zxx}^{(2)} = \chi_{zyy}^{(2)}$, and $\chi_{zzz}^{(2)}$. Different combinations of input and output beam polarizations in SFG measurements are often used to deduce values for the non-vanishing elements. Such measurements then allow the determination of surface symmetry or surface molecular orientation. Being a third-rank tensor, $\chi_s^{(2)}$ can reflect up to three-fold rotational symmetry. $\chi_{s,ijk}^{(2)}$ is related to the molecular nonlinear polarizability or hyperpolarizability $\alpha_{\xi\eta\zeta}^{(2)}$, where $\hat{\xi}$, $\hat{\eta}$, and $\hat{\zeta}$ define the molecular coordinates, through the coordinate transformation

$$\chi_{s,ijk}^{(2)} = N_s \left\langle (\hat{\boldsymbol{i}}\cdot\hat{\boldsymbol{\xi}})(\hat{\boldsymbol{j}}\cdot\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{\xi}}) \right\rangle \alpha_{\xi\eta\zeta}^{(2)} \qquad [3]$$

The angular brackets denote an average over the molecular orientational distribution, and N_s is the surface molecular density. For simplicity, microscopic local-field correction is neglected in eqn [3]. Knowing $\chi_{s,ijk}^{(2)}$ and $\alpha_{\xi\eta\zeta}^{(2)}$ thus permits deduction of information on the orientational distribution.

An explicit expression for the nonlinear optical polarizability is obtained from a second-order quantum mechanical perturbation calculation:

$$\alpha^{(2)}_{s,ijk}(\omega_s) = -\frac{e^3}{\hbar^2}\sum_{g,n,n'}\left[\frac{\langle g|r_i|n\rangle}{\omega_s-\omega_{ng}+\mathrm{i}\Gamma_{ng}} \times\left(\frac{\langle n|r_j|n'\rangle\langle n'|r_k|g\rangle}{\omega_2-\omega_{n'g}+\mathrm{i}\Gamma_{n'g}} + \frac{\langle n|r_k|n'\rangle\langle n'|r_j|g\rangle}{\omega_1-\omega_{n'g}+\mathrm{i}\Gamma_{n'g}}\right)+\cdots\right]\rho_g^{(0)} \qquad [4]$$

This expression shows how the nonlinear polarizability or susceptibility depends on material parameters such as the dipole transition moments $\langle n|r_i|g\rangle$ and energy levels. It contains a sum over eight resonant terms. The quantities ω_{ng} and Γ_{ng} are the frequencies and half-widths for the transitions between quantum states $|n\rangle$ and $|g\rangle$, and $\rho_g^{(0)}$ denotes the population in $|g\rangle$. It can be seen that $\alpha^{(2)}$, and hence the SF output, should be resonantly enhanced when the pump frequency ω_1 or ω_2 and/or the sum frequency $\omega_1 + \omega_2$ approach resonance. The resonant enhancement provides spectral information and makes SHG and SFG effective spectroscopic tools. The resonances could be electronic or vibrational, or more generally, any surface characteristic transition.

Experimental

Here we deal with basic considerations concerning the optical setup for SHG or SFG experiments in reflection from a surface as is shown schematically in Figure 2. In general, the pump radiation from pulsed laser sources is directed onto the surface. High-power tunable light pulses can be obtained from optical parametric generation and amplification (OPG/OPA) together with harmonic, sum- or difference-frequency generation, preferably pumped by picosecond or femtosecond lasers with high repetition rates. By this means radiation tunable from near UV at ~200 nm to mid IR at ~18 μm can be generated. For SFG the input pulses with frequencies ω_1 and ω_2 are directed to overlap spatially and temporally on the sample. The sample could be the surface of a liquid or solid in air, a single crystalline surface in ultrahigh vacuum, buried interfaces, etc. Polarizers and half-wave plates allow for tuning of incident power and polarizations at the sample. Together with the polarizer in the output path, different input/output polarization combinations can be chosen. In the detection system the signal has to be discriminated against reflected and scattered pump light. This is achieved by spatial and spectral filtering with apertures, dielectric mirrors, and interference and color glass filters. An optional monochromator can also be used for additional stray-light suppression. The signal is then detected by a photomultiplier tube and gated electronics are used for signal integration or photon counting.

From eqn [2] the expected signal strength for SHG/SFG can be estimated. With a typical value of $\chi_s^{(2)}$ of 10^{-21} m^2 V^{-1}, a single 1-μm pump pulse, incident at $\theta = 45°$ having pulse energy $E = I A\tau = 100$ μJ, beam cross-section $A = 1$ mm^2, and pulse duration $\tau = 10$ ps, can generate about 10^3 photons per pulse of SHG. By means of photon counting a minimal count rate of 10^{-3} photons/pulse

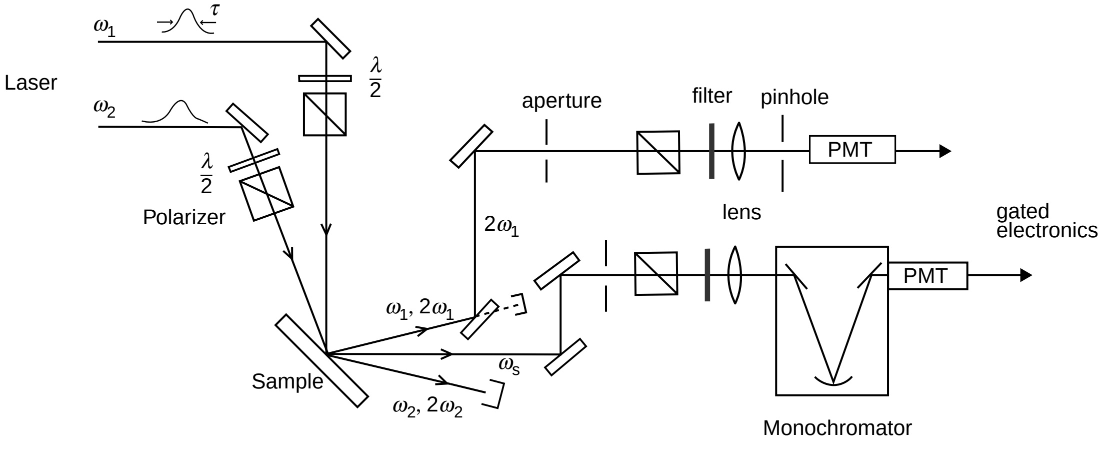

Figure 2 Schematic of experimental setup for second-harmonic and sum-frequency generation.

can be achieved which allows the detection of $\chi_s^{(2)}$ as small as 10^{-24} m^2 V^{-1}. The spectral resolution is limited by the spectral width of the pump pulses. In order to obtain accurate spectral data over a wide tuning range or to obtain absolute values for $\chi_{s,ijk}^{(2)}$ of the material investigated, a reference material such as z-cut quartz with known values of χ is to be measured for comparison.

The simplest kind of experiments are SHG at a fixed pump frequency. From the signal variation in response to surface modification one can probe kinetics and dynamics of adsorption, desorption, diffusion, surface melting, phase transitions, etc. By tuning the pump wavelength, second-harmonic spectroscopy can give information on surface electronic states. For surface vibrational spectroscopy, however, IR-vis SFG must be employed. In the latter case, tunable IR input is mixed with visible input to yield a SF signal in the visible region.

Surface Specificity

In many cases, the surface contribution to SFG or SHG from a centrosymmetric medium clearly dominates over the bulk contribution. This generally occurs when the surface or interface is composed of a polar oriented molecular layer. An example is shown in Figure 3, where the vibrational spectra in the C–H stretch region of three buried solid–liquid interfaces are presented. The spectrum for hexadecane ($C_{16}H_{34}$)/fused silica shows no C–H peaks. This is in spite of the fact that liquid hexadecane has strong infrared and Raman activity in this spectral range. The result indicates that the bulk signal contribution is negligible. In contrast, strong resonances are observed from the hexadecane/OTS/silica system, where OTS is a monolayer of octadecyltrichlorosilane ($CH_3(CH_2)_{17}SiCl_3$) chemisorbed on the fused silica plate. A similar spectrum is obtained for the CCl_4/OTS/silica interface. This clearly indicates that it is the contribution from the OTS monolayer that dominates the spectra.

Another example is described in Figure 4 where the SFG spectrum in reflection from a free liquid water surface is shown. The resonant features between 3000 and 3600 cm^{-1} are due to the O–H stretches of hydrogen-bonded OH groups. The sharp peak at 3700 cm^{-1} can be identified with the free OH bonds. Since no dangling OH bonds can exist below the surface its presence indicates that the spectrum must originate from the topmost layer of water molecules. Furthermore it can be shown that the surface water molecules are oriented with one of the OH bonds directed out of the liquid. In addition to the above examples, the surface sensitivity and specificity has been demonstrated and successfully applied to a large number of systems as will be shown below.

A note on the definition of surface or interface is in order. In SHG and SFG, the surface or interface layer refers to a thin layer between two adjacent bulk media that has a different structure from the bulk media and lacks inversion symmetry. If molecules in the surface layer are polar-ordered, then SHG and

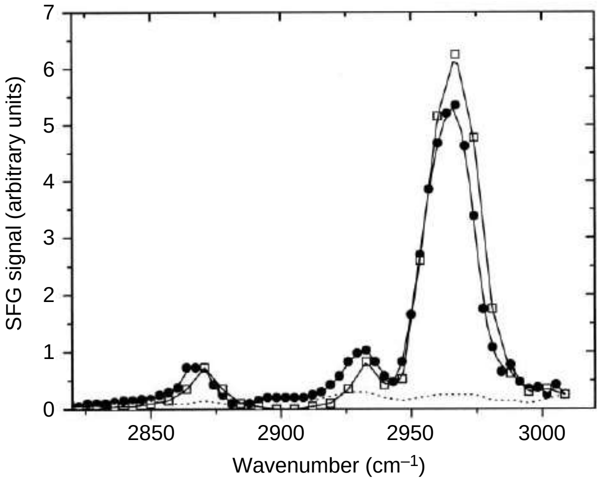

Figure 3 SFG spectra of different interfaces obtained with *ssp* polarization combination: hexadecane/silica (dashed), CCl_4/ OTS/silica (solid circles), and hexadecane/OTS/silica (open squares). OTS refers to a monolayer of octadecyltrichlorosilane.

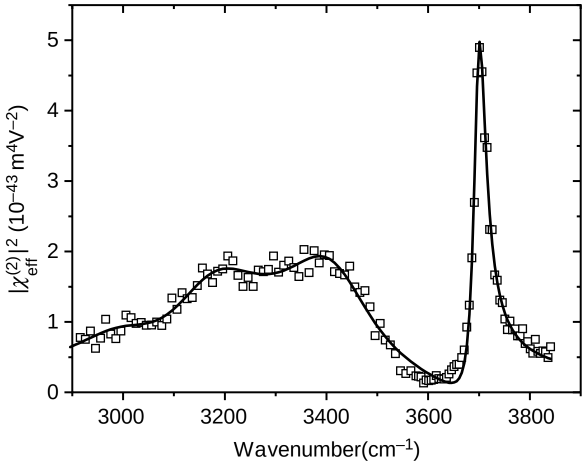

Figure 4 SFG spectrum of the liquid–vapor interface of pure water (*ssp* polarization combination). The mode at 3700 cm^{-1} is due to the dangling OH bonds at the surface.

SFG are often dominated by contributions from the surface layer. However, the bulk contribution is not strictly zero even if the media have inversion symmetry, as $\chi^{(2)}$ can arise from electric-quadrupole and magnetic-dipole contributions. Separation of surface and bulk contributions in SHG and SFG is a subtle problem in general. Experimentally, how surface modification affects the signal can be used to evaluate whether or not the surface contribution dominates.

A special case concerns materials that lack bulk inversion symmetry, for example, III–V and II–VI semiconductors such as GaAs and ZnSe. In this case, the bulk contribution to SHG and SFG is electric-dipole allowed and may be very significant. With an appropriate choice of polarization combinations and surface orientation, it can nevertheless be well suppressed, leaving the surface contribution dominant.

Applications

Both SFG and SHG have been applied to a wide variety of surface and interfacial systems. To illustrate their capabilities and versatility as surface spectroscopic tools, the focus of this summary will be on examples where the information deduced cannot readily be obtained by other techniques. (For details, readers are referred to the review articles cited in Further Reading and the original papers cited therein.)

SFG and SHG have been used for the study of adsorbates at surfaces even under ambient conditions. The adsorbate alters electronic or structural properties of the substrate and the signal is thus related to the surface density of the adsorbate and its orientation.

Elementary surface reactions, in particular adsorption, dissociation, or desorption of simple molecules such as CO, hydrogen, oxygen, or water on metal and semiconductor single-crystal surfaces under ultrahigh-vacuum conditions, have been investigated. Sensitivity into the sub-percent monolayer regime is achieved and allows the identifictation of competing reaction channels and adsorbate dynamics on surfaces. Such studies can help to understand epitaxial growth or heterogeneous catalysis because the techniques are applicable at any gas pressure and surface temperature. Related to this is the study of surface diffusion, which can be probed following the temporal decay of a diffracted SHG signal off a submonolayer adsorbate grating formed by laser-induced desorption.

Making use of the spectroscopic capabilities of SHG, the surface electronic structure of selected metals and semiconductors, in particular silicon, were studied with an emphasis on the effects of temperature, crystallographic orientation and different adsorbates. Among buried interfaces, the Si/SiO$_2$ interface has been the subject of intense investigations and the sensitivity of SHG to static electric fields at the interface and to inhomogeneous strain was also demonstrated.

Different in-plane surface symmetries associated with different crystallographic orientations or surface reconstructions yield different SHG and SFG responses. This allowed for the investigation of surface reconstructions (e.g., Si(111)$2 \times 1 \rightarrow 7 \times 7$), order–disorder transitions (e.g., Si(111)$7 \times 7 \rightarrow 1 \times 1$, Au/Si(111)) and surface melting (e.g., single-crystal ice, Si(111), Ge(111)), providing information on latent heat, superheating, and its dynamics using pump-probe techniques. Magnetization-induced SHG was also used for probing surface and interface magnetism of ferromagnetic metals and bimetallic systems.

Many unique applications of spectroscopic SFG have been developed based on its capability to study large molecules, in particular under ambient conditions. Surface vibrational spectra with different input/output polarization combinations and sample geometries provide information on orientation, conformation and alignment of surface molecules and the composition and structure of the surface layer. The study of self-assembled monolayers (e.g., OTS on quartz) to deduce the orientation and conformation of the alkyl chains in the monolayer is an example. The effect of molecular density and temperature on the spectrum gives information on intermolecular interactions and phase transitions. Another example concerns the adsorption of liquid crystal molecules on nanostructured polymer substrates. Both the chain orientation at the polymer surface and its effect on the alignment of adsorbed liquid crystal molecules can be addressed.

Surface structures of liquids, especially those of pure liquids, are of great interest in science and technology, and SFG is a unique spectroscopic tool for probing liquid surfaces and interfaces. For example, the SFG vibrational spectrum for the water–vapor interface in Figure 4 gives information on the density and orientation of the surface water molecules at the water–vapor interface. The surface layer was found to be more ordered compared to the bulk as was also observed for various other organic liquids studied with SFG. In other cases soluble and insoluble organic molecules at a liquid–vapor, liquid–liquid, or liquid–solid interface have also been studied. They appear as adsorbed monolayer films and their structure and phase behavior are relevant to many applications. Identification of the adsorbed molecular species at solid–liquid interfaces is also important for understanding electrochemical processes. Various studies with SFG have addressed bonding of, e.g., CO, CN^-, and SCN^- at electrodes and its variation with electrode potential.

SFG spectroscopy also finds unique applications in studies of polymer surfaces and interfaces most relevant to modern science and technology. The observed surface vibrational spectra of polymers provide information on surface composition, molecular orientation, and conformation of neat and blended polymers. The effect of environment and surface treatment on the surface structure can also be studied. Investigation of the biological functions of complex molecular systems has attracted increasing interest, and SHG and SFG have been applied successfully to selected systems ranging from the isomerization of retinal – a molecule involved in the vision process – to functional aspects of model membranes.

Both SHG and SFG can be combined with microscopy techniques for surface microscopic studies. Microscopic imaging can be achieved by rastering the probe beam or, preferably, the sample position. Near-field SFG/SHG spectroscopy has recently been developed.

Performing surface SHG and SFG experiments in the time domain can provide information on surface dynamics. Pump-probe experiments allow for measurements of energy relaxation and phase coherence of excitations such as surface electronic states and surface phonons and vibrations.

Outlook

Although SHG and SFG were established as surface analytical tools more than 15 years ago, it is still an active field of research with much potential that has not yet been explored. As an example, it has only recently been demonstrated that doubly resonant infrared–visible SFG, as a two-dimensional spectroscopy, can give access to couplings between vibrational modes and surface electronic transitions. With the commercial availability of suitable laser sources, the techniques can be extended to a much wider range of applications. This includes possible investigations of ultrafast surface dynamics, nanostructures, and biological systems.

Further Reading

Bennemann KH (ed.) (1998) *Nonlinear Optics in Metals*. Oxford, UK: Clarendon Press.

Bloembergen N (1977) *Nonlinear Optics*. New York: Benjamin.

Boyd RW (1993) *Nonlinear Optics,* 2nd edn. San Diego, CA: Academic Press.

Dumas P, Weldon MK, Chabal YJ and Williams GP (1999) Molecules at surfaces and interfaces studied using vibrational spectroscopies and related techniques. *Surface Review & Letters* 6(2): 225–255.

Guyot-Sionnest P and Harris AL (1995) Surface vibrational dynamics probed by sum frequency generation. In: Ho W and Dai HL (eds) *Laser Spectroscopy and Photochemistry of Metal Surfaces*, pp. 405–458. Singapore: World Scientific.

Guyot-Sionnest P, Superfine R, Hunt JH and Shen YR (1988) Vibrational spectroscopy of a silane monolayer at air/solid and liquid/solid interfaces using sum-frequency generation. *Chemical Physics Letter* 144: 1–5.

Heinz TF (1991) Second-order nonlinear optical effects at surfaces and interfaces. In: Stegeman GI and Ponath H-E (eds), *Nonlinear Surface Electromagnetic Phenomena*, pp. 353–416. Amsterdam, The Netherlands: North-Holland.

Matthias E and Träger F (eds) (1999) *Applied Physics B, Special Issue: Nonlinear Optics at Interfaces*, Volume 210.

McGilp JF (1999) Optical second-harmonic generation as a semiconductor surface and interface probe. *Physica Status Solidi (a)* 175: 153–167.

Miranda PB and Shen YR (1999) Liquid interfaces: A study by sum-frequency vibrational spectroscopy. *Journal of Physical Chemistry B* 103: 3292–3307.

Reider GA and Heinz TF (1995) Second-order nonlinear optical effects at surfaces and interfaces: recent advances. In: Halevi P (ed.) *Photonic Probes of Interfaces*, 413. Amsterdam, The Netherlands: Elsevier.

Shen YR (1984) The Principles of Nonlinear Optics. New York: Wiley.

Shen YR (1994) Nonlinear optical studies of polymer interfaces. *International Journal of Nonlinear Optical Physics* 3: 459–468.

Shen YR (1994) Surface spectroscopy by nonlinear optics. In: Ḧansch TW and Inguscio M (eds) *Frontiers in Laser Spectroscopy*, pp. 139–165. Amsterdam, The Netherlands: North-Holland.

Shen YR (2001) Vibrational spectroscopy of water at interfaces. In: Morra M (ed.) *Water in Biomaterials Surface Science*, Chapter 8, pp. 215–244. Chichester, UK: Wiley.

Wei X and Shen YR (2001) Motional effect in surface sum-frequency vibrational spectroscopy. *Physical Review Letter* 86: 4799–4802.

POLARIZATION

Contents

Introduction
Matrix Analysis

Introduction

J M Bennett, Michelson Laboratory, China Lake, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Polarization is one property of light waves and will be defined after a brief overview of the properties of light. Light is part of the electromagnetic spectrum of waves that have both particle-like and wave-like properties. Light waves carry energy in the form of photons which act like particles; the photon energy increases in proportion to the frequency of the wave. The particle-like properties of light and other electromagnetic waves are described by quantum mechanics. Light also acts like transverse waves that travel in straight lines in air and vacuum and can be described by classical electromagnetic theory. Polarization deals with the wave-like properties of light, and is described mathematically by Maxwell's equations, the key relations in electromagnetic theory. Polarization effects occur when light interacts with matter. In order to understand polarization, there will be a brief introduction to the subject of electromagnetic waves.

For a more in-depth treatment of the material in this article, the reader is directed to the Further Reading list at the end of this article and in particular to the two chapters on 'Polarization' and 'Polarizers' written by the author in Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chapter 5 and vol. II, chapter 3. New York: McGraw-Hill Inc.

Electromagnetic waves have both electric and magnetic fields associated with them. These are vibrations in directions perpendicular to the direction the wave is traveling, i.e., the direction of propagation. $\vec{E}$ represents the vector of the electric field and $\vec{H}$, perpendicular to $\vec{E}$, represents the magnetic field vector. Both these vectors are complex and have real and imaginary parts. Polarization effects are always associated with the $\vec{E}$ vector. Specifically, the plane of polarization is defined as the plane in which the $\vec{E}$ vector is vibrating. Waves having different amplitudes, phases, or angular orientations (azimuths) of their electric or magnetic vectors can be combined by conventional vector addition methods. Also the $\vec{E}$ vector of a particular vibration can be resolved into two components in mutually perpendicular directions that are vibrating in phase.

If a light source such as the sun, a candle flame, or an electric light bulb is considered on a microscopic scale, each vibrating atom or molecule emits linearly polarized light (see the definition below). But the individual atoms or molecules do not act together, so their vibrations have no fixed phase relationships to each other and they cannot be added into a single linearly polarized beam. Thus, we call light from these sources unpolarized. In an unpolarized light beam, the $\vec{E}$ vector vibrates in all directions perpendicular to the direction of propagation. If a snapshot is taken at a particular instant of time, different parts of the beam will have $\vec{E}$ vectors vibrating with different amplitudes and phases at different angles to each other, but all in a plane perpendicular to the direction of propagation. In the most common convention used in optics, the wave travels in the $+z$ direction in a right hand coordinate system and the $\vec{E}$ vectors are all vibrating at various angles in the $x-y$ plane. The angle of vibration is measured from the positive x axis in a counterclockwise direction when the observer is looking against the direction of propagation of the light beam.

For linearly polarized or plane polarized light, if a snapshot of a light beam is taken at a particular instant, the $\vec{E}$ vector will be vibrating at a certain angle in the $x-y$ plane. As time (or position on the traveling wave) varies, the amplitude of the $\vec{E}$ vector will vary in a sinusoidal manner, but the vibration will remain at the same angle in the $x-y$ plane. As an example, the real part of the electric vector of a linearly polarized beam that is vibrating in the $+x$ and $-x$ directions and traveling in the $+z$ direction is given by the relation:

$$\vec{E}_x(z,t) = \hat{\imath} E_0 \cos(kz - \omega t) \qquad [1]$$

where t is the time, $\hat{\imath}$ is a unit vector in the $+x$ direction, E_0 is the amplitude of the vibration, $\vec{k}$ is the propagation vector in the direction the wave is traveling, i.e., the z direction (its magnitude is $2\pi/\lambda$), and $\omega = 2\pi f$ where f is the vibration frequency of the wave. In free space, $\omega = 2\pi c/\lambda$ where c is the velocity of light and λ is the wavelength of the vibration.

Light can also be circularly or elliptically polarized. Circularly polarized light is produced by adding the electric vectors of two waves, each having the same amplitude but which are 90° out of phase; the resulting vibration sweeps out a circle in the x–y plane. When the first wave is vibrating in the $+x$ and $-x$ directions (eqn [1]), the second wave is vibrating in the $+y$ and $-y$ directions:

$$\vec{E}_y(z,t) = \hat{\jmath} E_0 \sin(kz - \omega t) \qquad [2]$$

and is added to the first wave. The resultant wave is the sum of eqns [1] and [2]:

$$\vec{E} = E_0[\hat{\imath} \cos(kz - \omega t) + \hat{\jmath} \sin(kz - \omega t)] \qquad [3]$$

The direction of vibration of this wave will rotate in a circle as either the time increases or as the distance along the wave increases, but the amplitude of the vibration will not change. As time increases, the vibration will make a circle in the clockwise direction (to negative angles). This light is defined to be right circularly polarized. If the + sign between the two parts of eqn [3] is changed to a minus sign, as time increases, the vibration will make a circle in the counterclockwise direction (to positive angles) and the light is now defined to be left circularly polarized. These definitions are for conventional (traditional) optics. However, the opposite definitions with right (left) circularly polarized light defining circles in the counterclockwise (clockwise) direction are also in use. In modern physics, there is still another convention that defines right (left) circularly polarized light as having negative (positive) helicity.

Right and left elliptically polarized light beams have the same angle conventions as for circularly polarized beams but are produced by adding two electric vectors that have different amplitudes. In eqns [1] and [2], the amplitude terms will be, for example, E_1 and E_2 instead of E_0 and eqn [3] can no longer be used. There is now a major axis and a minor axis of the ellipse. If $E_1 > E_2$, the major axis will be along the x axis; for $E_1 < E_2$, the major axis will coincide with the y axis.

The preceding discussion has dealt entirely with the electric vector of the electromagnetic field. However, one cannot observe the $\vec{E}$ vector. The irradiance (energy per unit area per unit time), $\vec{E}\vec{E}^*$, is what can be observed visually and measured by electronic detectors. Thus, measurements of the polarization are irradiance measurements. Light transmitted by a polarizer is called the transmittance of the polarizer; similarly, light reflected from a polarizer is called the reflectance of the polarizer.

One of the most important parts of the subject of polarization is how to produce linearly polarized light starting with unpolarized light. This is done by using polarizers, as discussed in the section on polarizers below. Certain materials have special properties that make them able to polarize light. Depending on the application, different kinds of polarizers are preferred. Optics textbooks by Hecht and Guenther discuss the most important polarizers and two articles by Bennett describe many kinds of polarizers including special ones (see the Further Reading at the end of this article for full references). Only the basic principles will be described here.

Sunlight scattered by air molecules in the atmosphere (Rayleigh scattering) is also partially linearly polarized. The air molecules act like tiny dipoles and vibrate when they absorb sunlight. They emit radiation that is polarized in certain directions relative to the vibration direction. When the viewer is at a 90° angle with respect to the sun, the polarization of the skylight is a maximum. Rayleigh scattering is the subject of several books and will not be further discussed here.

Retarders are used to change linearly polarized light into circularly or elliptically polarized light and compensators, which are a form of retarders, can analyze an unknown type of polarized light and determine its composition. They are discussed below.

Polarimetry and ellipsometry are closely related techniques that are used to determine the optical properties of a material by shining a known type of polarized light on the material at non-normal incidence and analyzing the polarization properties of the light after it has been reflected from the material. These subjects are treated in other articles in this encyclopedia and in several references at the end of this article.

Changes can be produced in the optical properties of some materials by applying an electric field, a magnetic field, an acoustic field, or another form of variable pressure. The materials changed in these ways are said to be electro-optic, magneto-optic, or piezo-optic. The changes in the optical properties modify some parameter of a light wave passing through a material or reflecting from it.

The amplitude, phase, frequency, polarization, or direction of the light wave can be modified. In this article, we are only concerned with modifications that produce changes in the polarization; these are discussed towards the end of this article.

Matrix methods have been developed to handle problems involving polarization and there is also a visual representation of the matrix algebra that is based on the Poincaré sphere. Both topics are covered at the end of this article.

Polarizers

Basic Relations for Polarizers

A linear polarizer is anything which, when placed in an incident unpolarized beam, produces a beam of light whose electric vector is vibrating primarily in one direction with only a small component vibrating in the direction perpendicular to it. The transmittance T of the linear polarizer is

$$T = \tfrac{1}{2}(T_1 + T_2) \qquad [4]$$

where T_1, the principal transmittance of the polarizer, is $\gg T_2$, the transmittance of the polarizer at 90° to the principal transmittance. Thus, a perfect polarizer would transmit only 50% of an incident unpolarized beam.

If a linear polarizer is placed in a linearly polarized beam and is rotated about an axis parallel to the beam direction, the transmittance will vary between a maximum value T_1 and a minimum value T_2 according to the law:

$$T = (T_1 - T_2)\cos^2\theta + T_2 \qquad [5]$$

where θ is the angle between the plane of vibration of the principal transmittance and the plane of vibration of the electric vector in the incident beam.

The extinction ratio ρ_P of a polarizer is defined as

$$\rho_P = \frac{T_2}{T_1} \qquad [6]$$

and the degree of polarization of a polarization P is

$$P = \frac{T_1 - T_2}{T_1 + T_2} \qquad [7]$$

When two identical polarizers are placed in an unpolarized beam, and the directions of their principal transmittances, T_A and T_B, are inclined at an angle θ to each other, the transmittance of the pair will be

$$T_\theta = \tfrac{1}{2}(T_1^2 + T_2^2)\cos^2\theta + T_1 T_2 \sin^2\theta \qquad [8]$$

Thus, when the directions of the principal transmittances are aligned, $T_\parallel = \tfrac{1}{2}(T_1^2 + T_2^2)$, and when they are perpendicular, $T_\perp = T_1 T_2$.

Birefringent Materials (Calcite)

The majority of high-quality polarizers are made from calcite. This is a birefringent (doubly refracting) crystalline material that is uniaxial (i.e., there is one preferred direction in the crystal). A birefringent material acts differently for light going in different directions through the crystal. For example, if an unpolarized light beam passes through the crystal in a certain direction, it will be split into two spatially separated beams that are parallel but are linearly polarized at right angles to each other. A uniaxial crystal has an optic axis (i.e., a certain direction in the crystal). When light rays travel parallel to the optic axis, they travel at the same velocity and there is no difference between them. When light passes through the crystal in other directions, the ray whose vibration direction is perpendicular to the optic axis is governed by the ordinary laws of geometrical optics (the same as for isotropic materials) and is called the ordinary ray. The ray whose vibration direction is parallel to the optic axis does not follow the normal geometrical optics laws and is called the extraordinary ray. One ray travels faster than the other, so there is a phase retardation for one ray relative to the other. This is the principle of a retarder or retardation plate (see the next section).

Calcite is a negative uniaxial crystal which means that the refractive index for the ordinary ray is larger than the refractive index for the extraordinary ray. When the ordinary ray enters a block of calcite from air at non-normal incidence, it is bent more than the extraordinary ray.

Calcite can be easily cleaved along three distinct planes, making it possible to produce rhombs of the form shown in Figure 1. The optic axis, going in the HI direction, makes equal angles with all three faces at point H. Any plane, such as DBFH, which contains the optic axis and is perpendicular to the two opposite faces of the rhomb ABCD and EFGH is called a principal section. If the plane of incidence of light on the rhomb coincides with a principal section, the light entering the crystal will be split into two components polarized at right angles to each other which travel in slightly different directions and leave the crystal as two beams slightly displaced but parallel to each other.

The large birefringence of calcite and its excellent transmission through the visible spectral region and into the ultraviolet and infrared regions has made it possible to make excellent high extinction ratio polarizers with different designs. Some of these are shown in Figure 2. There are two main types: Glan types with rectangular shapes and Nicol types with rhombohedral shapes. The polarizers are made of two pieces of calcite cemented together. Glan types have their optic axes in the plane of the entrance face. In the Nicol types, the principal section is perpendicular to the entrance face, but the optic axis is neither parallel nor perpendicular to the face. The two halves of conventional polarizing prisms are cemented together with cement that has a refractive index intermediate between the ordinary and extraordinary refractive indices. This enables one ray (generally the extraordinary, or *e* ray) to be transmitted and the other to be reflected at the cut, so that only one ray exits from the prism in the direction of the incoming ray. The Glan types are also used without cement with only an air space between the two halves. They can be used at shorter wavelengths in the ultraviolet where the cement absorbs. The extinction ratio can be very high for air-spaced prisms. For example, for a Glan–Foucault prism (an air-spaced Glan–Thompson prism, Figure 2), the extinction ratio can be better than 1×10^{-5} and the prism is usable from about 0.214 μm in the ultraviolet to 2.3 μm in the infrared. However, air-spaced polarizers have very small field angles, so they are mainly used with laser sources where the beam is parallel.

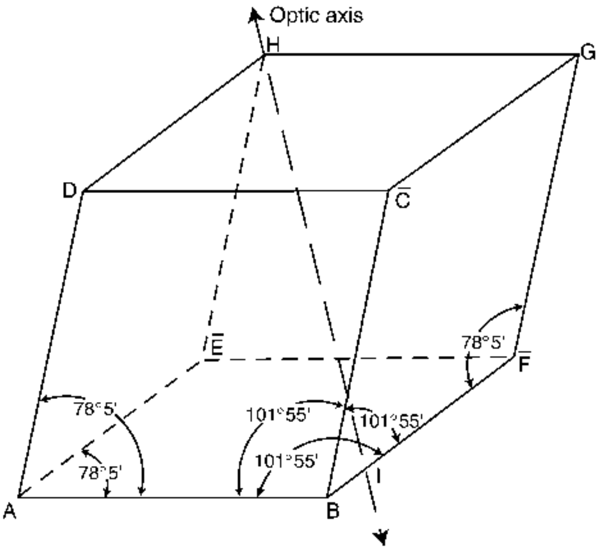

Figure 1 Schematic representation of a rhombohedral calcite crystal showing the angles between faces. The optic axis passes through corner H and point I on side BF. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

The extreme paths of light through a cemented Glan–Thompson prism are shown in Figure 3. This prism has a length that is three times the width of the entrance aperture (i.e., an *L*/*A* ratio of 3). The optic axis is perpendicular to the plane of incidence which is in the plane of the paper. In the first half of the polarizer, the paths of the ordinary and extraordinary rays, both of which follow the conventional law of refraction (Snell's Law, eqn [8] above) nearly coincide. Ray A is incident on the entrance face of the polarizer at an angle such that the angle of incidence on the cut is the smallest angle for which the *o* ray is totally internally reflected. Ray B is incident at an angle such that the angle of refraction in the first half of the prism is essentially equal to the cut angle, S, so that the *e* ray just passes through the cut. The field angle of the polarizer is twice the smaller of angles θ_A and θ_B. Thus, all rays having angles of incidence between rays A and B will be polarized

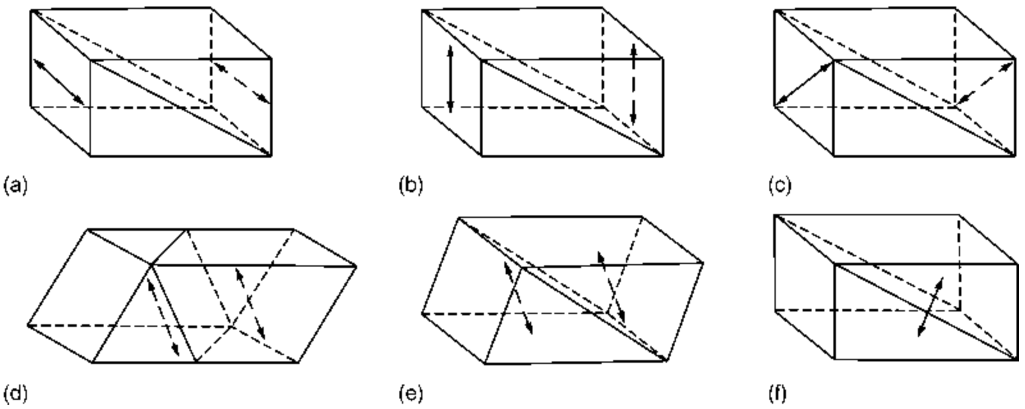

Figure 2 Types of polarizing prisms. Glan types: (a) Glan-Thompson, (b) Lippich, and (c) Frank-Ritter; Nicol types: (d) conventional Nicol, (e) Nicol, Halle form, and (f) Hartnack-Prazmowsky. The optic axes are indicated by the double-pointed arrows. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

when exiting the prism. The paths of similar rays can be traced through the other prism designs.

In addition to conventional polarizing prisms, there are also polarizing beamsplitter prisms and Feussner prisms. In polarizing beamsplitter prisms, the two beams that are polarized at right angles to each other both emerge from the prism but are separated spatially. Figure 4 shows a diagram of several types of polarizing beamsplitter prisms and Figure 5 shows the paths of rays through side views of these prisms.

A Feussner prism is made of isotropic material but the film separating the two halves is birefringent. These prisms have the advantage that much less birefringent material is required but they have a more limited wavelength range when calcite or sodium nitrite (another birefringent material) is used because the transmitted ordinary ray has a shorter transmission range than the extraordinary ray which is transmitted by the conventional and air-spaced polarizing prisms.

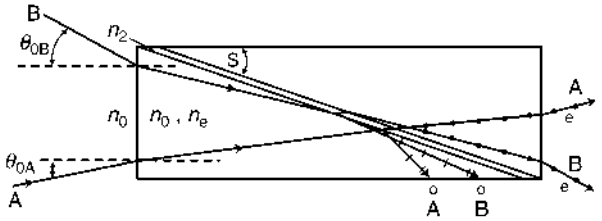

Figure 3 Extreme rays (A and B) that can pass through a cemented Glan-Thompson prism. Both rays change to ordinary and extraordinary rays in the calcite. Both ordinary rays are reflected at the boundary between the two halves of the prism and the extraordinary rays exit in the directions indicated. Rays entering the prism between rays A and B would be transmitted by the prism. The field angle of the prism is twice the smaller of the two angles of incidence (ray A).

Dichroic Absorbers

A dichroic material is one that absorbs light polarized in one direction more strongly than light polarized at right angles to that direction. Dichroic materials are different from birefringent materials because the latter usually have negligible absorption coefficients for both the ordinary and extraordinary rays. Stretched polyvinyl alcohol sheets treated with absorbing dyes or polymeric iodine are the most common type of dichroic absorbers and are primarily sold under the tradename Polaroid. These polarizers do not have as good an extinction ratio as the calcite prism polarizers (see above), but they are inexpensive, come in large sizes, are easily rotated, and produce negligible beam deviation. Also, they are thin, lightweight, and can be made in any desired shape. One of their main advantages is that they are insensitive to the degree of collimation of the beam and can be used in strongly convergent or divergent light.

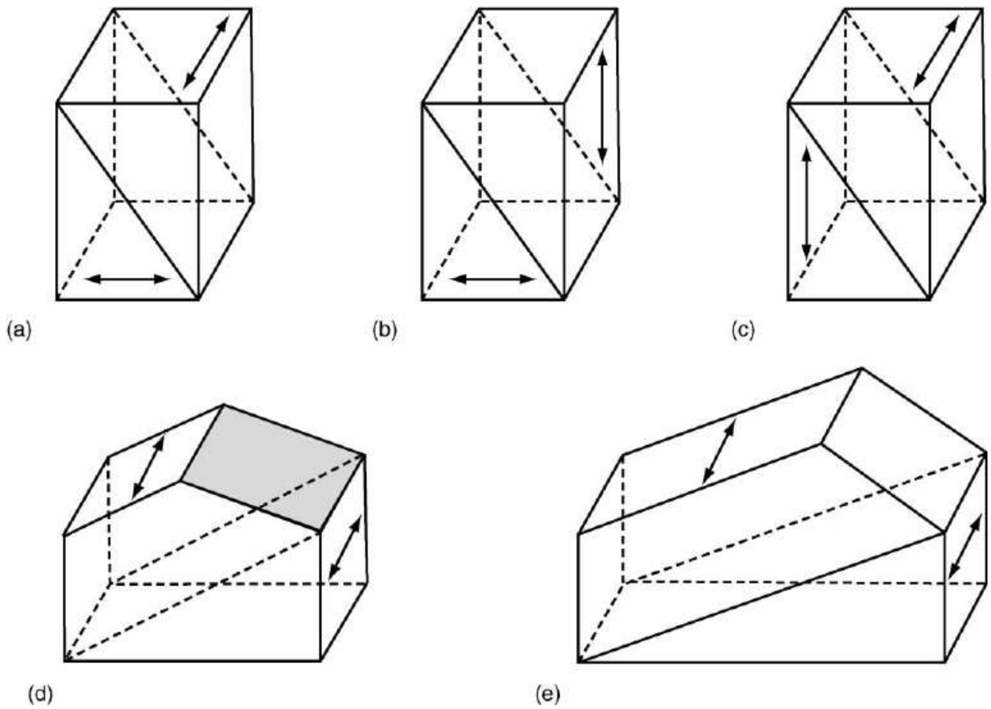

Figure 4 Three-dimensional views of various types of polarizing beamsplitter prisms: (a) Rochon; (b) Sénarmont; (c) Wollaston; (d) Foster (shaded face is silvered); and (e) beamsplitting Glan-Thompson. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

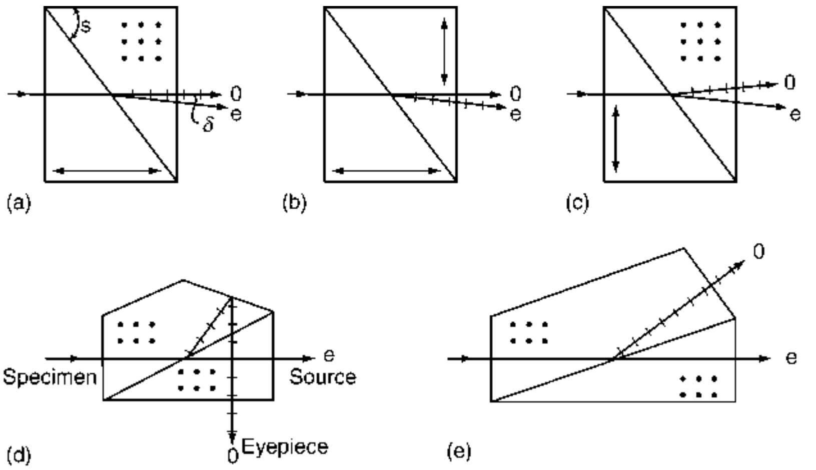

Figure 5 Side views of the polarizing beamsplitter prisms in Figure 4. The directions of the optic axes are indicated by the dots and the heavy double-pointed arrows. When the Foster prism is used as a microscope illuminator, the source, specimen, and eyepiece are in the positions indicated. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

Depending on the density and type of absorbing dye used to make the polarizer, the transmission can be maximized for the visible or near infrared spectrum. The extinction ratio of the Polaroid HN-22 sheet compares favorably with that of the Glan–Thompson prisms throughout the visible spectral region, but the transmission of the Glan–Thompson prism is superior. As the dichroic polarizer transmission increases, the extinction ratio becomes worse. Transmission and extinction ratio curves for various types of dichroic polarizers are shown in Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

A wire grid polarizer is another kind of dichroic absorber. It is made of a series of equally spaced conducting bars or wires that are either free standing or deposited onto a transparent substrate (backing plate). Energy that is polarized parallel to the length of the bars is absorbed out of the incoming wave by inducing oscillations in the electrons in the metal. Thus, only light polarized perpendicular to the bars will be transmitted. In order to have an appreciable degree of polarization, the wavelength must be at least twice the spacing between grids in the polarizer. Because of the difficulty of making wire grids with small enough spacings, these polarizers are limited to the mid- and long-wavelength infrared spectral region, beyond about 5 μm. The polarizer will have the best extinction ratio if the substrate has a low refractive index; if a high refractive index substrate such as silicon or germanium is used, it must be covered with an antireflection coating before the grid is deposited.

Most of the wire grid polarizers have been used with microwaves, so the theory is in the form for transmission lines. Similar transmission line theory has been applied to infrared polarizers in the form of bars and wires.

Reflection and Transmission

Light can be polarized by reflecting it from the flat surface of a material inclined at non-normal incidence to the light beam or by transmitting it through a transparent plate at non-normal incidence. These polarizers work because light has different reflectances when the electric vector is linearly polarized parallel and perpendicular to the plane of incidence. The plane of incidence contains both the incoming light beam (incident beam) and the reflected light beam. The angles for both the incident and reflected light beams are measured relative to an axis perpendicular to the surface (the surface normal). The reflection coefficients are given by the Fresnel equations and, for nonabsorbing materials, are:

$$R_s = r_s^2 = \frac{\vec{E}_s^2(\text{reflected})}{\vec{E}_s^2(\text{incident})} = \frac{\sin^2(\theta_0 - \theta_1)}{\sin^2(\theta_0 + \theta_1)} \qquad [9]$$

$$R_p = r_p^2 = \frac{\vec{E}_p^2(\text{reflected})}{\vec{E}_p^2(\text{incident})} = \frac{\tan^2(\theta_0 - \theta_1)}{\tan^2(\theta_0 + \theta_1)} \qquad [10]$$

where R_s and R_p are called the reflectances (previously called the intensity reflection coefficients), and r_s and r_p are the amplitude reflection coefficients. $\vec{E}_s$ ($\vec{E}_p$) is the incident or reflected electromagnetic wave vibrating perpendicular (parallel) to the plane of incidence and θ_0 and θ_1 are the angles of incidence and refraction, respectively. The angle of refraction is the angle of the light beam in the material, measured from the surface normal. For a nonabsorbing material, it can be obtained in terms of the refractive index n_1 of the material from Snell's Law:

$$\frac{\sin\theta_0}{\sin\theta_1} = \frac{n_1}{n_0} \qquad [11]$$

where n_0 and n_1 are the refractive indexes of the incident medium (usually air with $n_0 = 1$) and the material. Sometimes the refractive index of the material is expressed as a ratio measured relative to the refractive index of the incident medium: $n = n_1/n_0$. The reflectances of absorbing materials are similar to eqns [9]–[11] but involve complex refractive indexes and angles of refraction.

Figure 6 shows curves for R_s and R_p as a function of angle of incidence for four nonabsorbing materials that have different refractive indexes. In all cases the reflectance is higher for the *s* component than for the *p* component except at 0° and 90° angles of incidence where they are the same. At the so-called Brewster angle, θ_B, $R_p = 0$, $\tan\theta_B = n_1/n_0$, and incident unpolarized light is now completely linearly polarized perpendicular to the plane of incidence (s-polarized light). Note that high refractive index materials produce more intense polarized beams (i.e., have higher reflectances) than low refractive index materials.

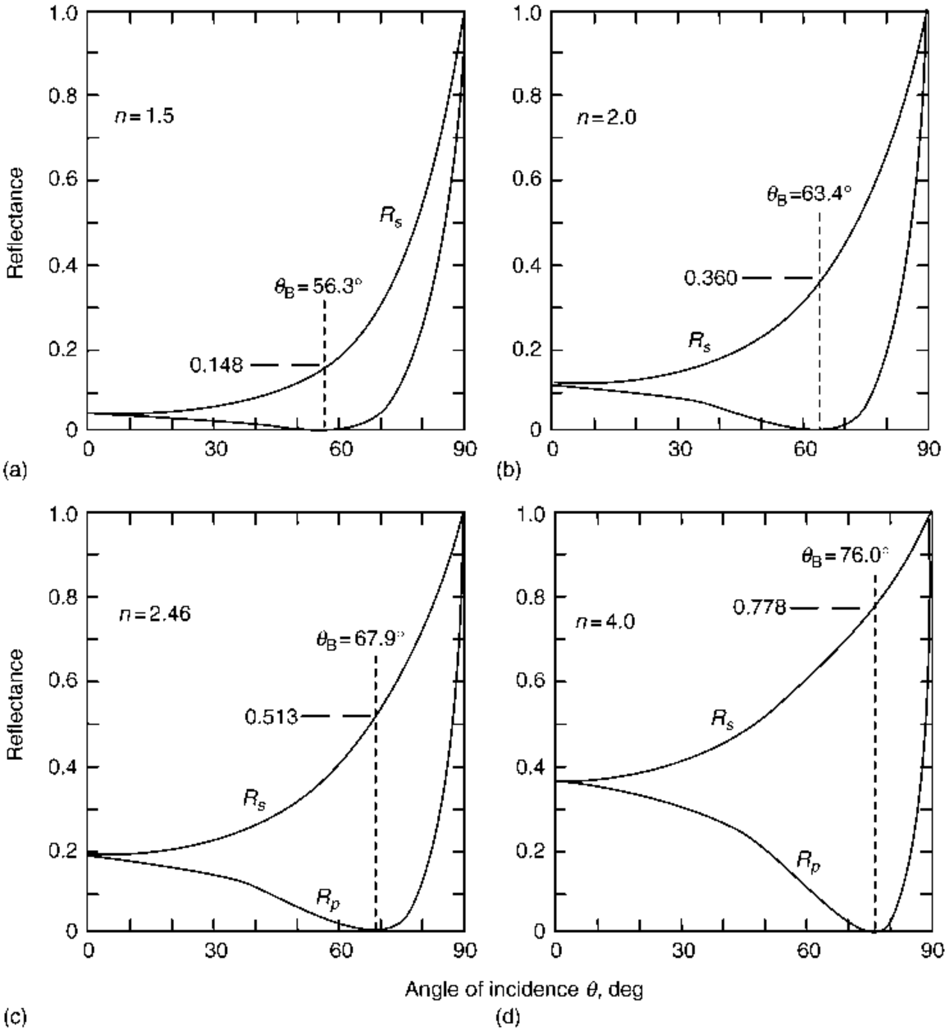

Figure 6 Reflectance of light polarized parallel R_p and perpendicular R_s to the plane of incidence from materials of different refractive index n as a function of angle of incidence: (a) $n = 1.5$ (alkali halides in the ultraviolet, glass (approximately) in the visible, and sheet plastics in the infrared), (b) $n = 2.0$ (AgCl in the infrared), (c) $n = 2.46$ (Se in the infrared), and (d) $n = 4.0$ (Ge in the infrared). The Brewster angle θ_B (at which R_p goes to 0) and the magnitude of R_s at θ_B are also indicated. (Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

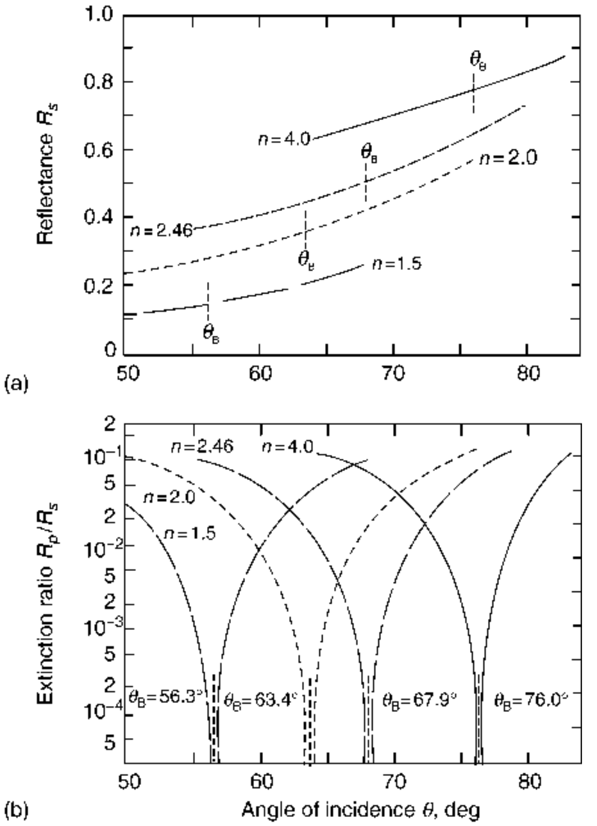

Figure 7 (a) Reflectance R_s and (b) extinction ratio R_p/R_s for materials of different refractive index at angles near the Brewster angle θ_B. A single surface of the material is assumed. (Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

Figure 7 shows the reflectances and extinction ratios for materials having different refractive indexes. Each curve is for a single reflection. Since plates have two surfaces, there will be two reflections from an actual reflection polarizer so the reflectance and extinction ratios will increase. If the plates are thick, the reflection from the back surface of the plate will be displaced from the reflection from the front surface of the plate. A reflection polarizer is normally used close to the Brewster angle because a high degree of polarization (i.e., a very small extinction ratio) is desired. Because the extinction ratio changes rapidly for angles of incidence close to the Brewster angle, the incident beam must be well collimated to obtain a high degree of polarization. A main disadvantage of reflection polarizers is that the reflected beam is no longer parallel to the incident beam and two additional reflections are required to align the beam.

Light transmitted through a plate will only be partially linearly polarized at any angle of incidence including the Brewster angle because both s- and p-components are partially transmitted. Transmission polarizers are thus made of several plates to increase the degree of polarization. Figure 8 shows the transmittance and extinction ratio for a stack of four plane parallel plates assuming multiple incoherent reflections within each plate and no reflections between plates. At the Brewster angle the p-transmittance is theoretically 1 (assuming that there is no absorption within the material) but the extinction ratio greatly depends on the refractive index of the material. A stack of low refractive index plates ($n = 1.5$) has a poor extinction ratio, while a pile of high refractive index plates ($n = 4.0$) has a much better extinction ratio. The plates are often inclined at small angles to each other so the light multiply reflected between plates (which decreases the polarization) is removed from the transmitted beam. The sides of each plate are plane parallel to increase the polarization, but the plates are too thick for the amplitudes of the multiple internally reflected beams to add or subtract. If the amplitudes of the beams could be added, there would be interference effects and the transmittance would vary with the thickness of each plate and with the wavelength. These are so-called interference polarizers and are mentioned below.

Transmission polarizers do not have the angle deviation problem of the reflection polarizers, but the transmitted beam may be slightly displaced parallel to the incident beam if the plates are thick. The main problem with transmission polarizers is that the light is not completely polarized even with several plates in a stack. If the plates are slightly absorbing, the transmission of the polarizer is reduced.

Miscellaneous Types

There are numerous other types of polarizers that mostly use thin films. Interference effects in thin films can increase the polarization efficiency at certain wavelengths depending on the thicknesses of the films and the design of the multilayer coating. Many of the applications involve non-normal incidence designs. For example, a single high refractive index film or a multilayer coating evaporated onto a low refractive index substrate increases the degree of polarization of reflection and transmission polarizers. Polarizing beamsplitters also use multilayer dielectric coatings. Free-standing films of different thicknesses were formerly used for infrared polarizers before wire grid polarizers became available.

Dichroic absorbers have been made from two-phase lamelar eutectics of thin needles of a conducting material embedded in a transparent matrix. Thin sheets of pyrolytic graphite material also act

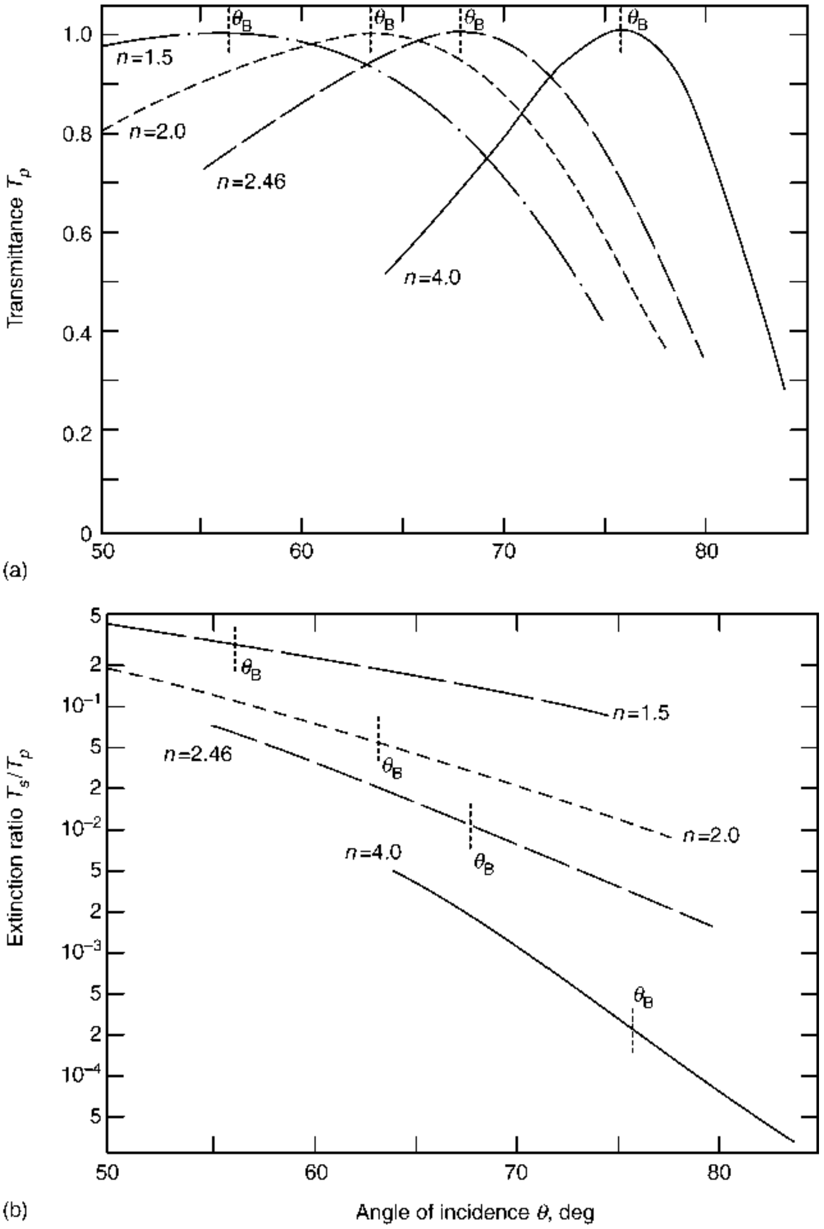

Figure 8 (a) Transmittance and (b) extinction ratio of four plane-parallel plates of refractive index n as a function of angle of incidence, for angles near the Brewster angle. Assumptions are multiple reflections but no interference within each plate and no reflections between plates. (Reproduced with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

as dichroic absorbers. There are also numerous miniature polarizer designs for fiber optics and nanotechnology applications.

Retarders or Retardation Plates

Retarders, or retardation plates, are devices that can change linearly polarized light into circularly or elliptically polarized light. They can also rotate the plane of polarization of linearly polarized light. A retardation plate is made from a uniaxial crystal that has an optic axis (as discussed above); the light travels in a direction perpendicular to the optic axis, as described below and shown in Figure 9. The ordinary and extraordinary rays travel at different velocities through the crystal depending on their refractive indexes: $n = c/v$, where c is the velocity of light in a vacuum and v is the velocity of light in the material. The larger the refractive index, the slower is the velocity of light in the material. The ordinary ray with a refractive index n_o and the extraordinary ray with a refractive index n_e have mutually perpendicular vibration directions. If the ray has a vibration direction at another azimuth in the crystal, the refractive index is intermediate between n_o and n_e. For a positive uniaxial crystal $n_e > n_o$ so the extraordinary ray travels slower than the ordinary ray through the crystal. Thus, the designation 'fast axis' is often used for the ordinary ray and 'slow axis' is used for the extraordinary ray, as shown in Figure 9. In a negative uniaxial crystal, $n_e < n_o$, i.e., the velocities along the two axes are reversed.

Because there is a velocity difference between the ordinary and extraordinary rays traveling through a uniaxial crystal, they get out of phase. Depending on their velocities and the optical thickness of the material (the refractive index times the physical thickness), the addition of their amplitudes results in a wave whose vibration direction is either: (a) perpendicular to the original vibration direction; (b) rotates in a circle (right or left circularly polarized light); or (c) rotates in an ellipse (right or left elliptically polarized light). The path difference N between two rays in the crystal, measured in terms of the wavelength of the light, is $N\lambda = \pm d(n_e - n_o)$, where d is the physical thickness of the material. The corresponding phase difference δ between the two rays is $\delta = 2\pi N = \pm(2\pi d/\lambda)(n_e - n_o)$. The quantity N is important because if beams of light vibrating along the fast and slow axes of a crystal get out of step by one quarter of the wavelength of the light, the device is called a quarter wave (or $\lambda/4$) retardation plate, or simply a quarter-wave plate. There are also half-wave ($\lambda/2$) plates that rotate the plane of polarization, and full-wave (λ) plates that rotate the plane of polarization through 180°. If light passes through a full-wave plate, theoretically it is indistinguishable from the original beam. However, materials are normally temperature sensitive so that as the temperature changes, the retardation is no longer exactly one wave, but may be slightly less than or greater than one wavelength. Other thicknesses of retardation plates produce elliptically polarized light.

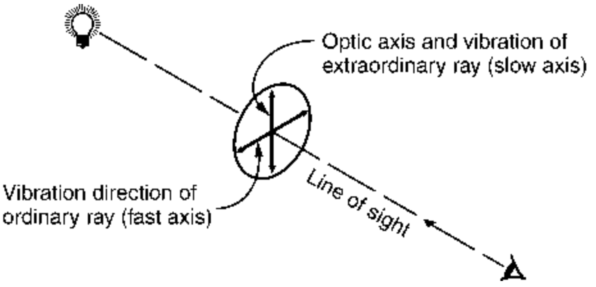
Optic axis and vibration of extraordinary ray (slow axis)
Line of sight
Vibration direction of ordinary ray (fast axis)

Figure 9 Light incident normally on the front surface of a retardation plate showing the vibration directions of the ordinary and extraordinary rays. In a positive uniaxial crystal, the fast and slow axes are as indicated in parentheses; in a negative uniaxial crystal, the two axes are interchanged. (Adapted with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

Figure 10 shows what happens to a beam of linearly polarized light when the axis of vibration is at 45° to the fast and slow axes of a positive

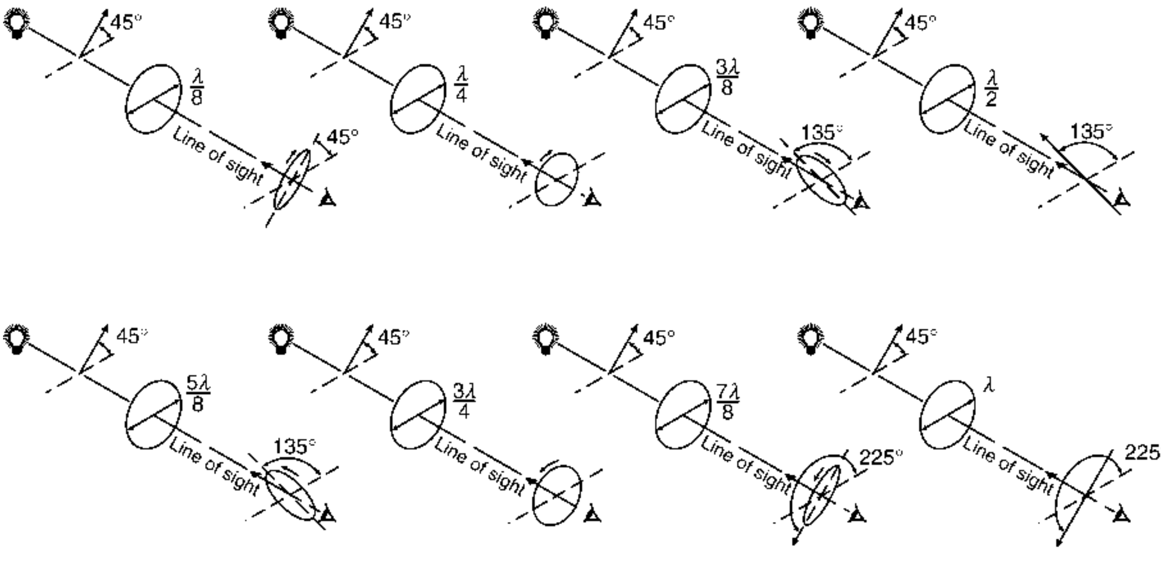
45°
$\frac{\lambda}{8}$
Line of sight
45°
45°
$\frac{\lambda}{4}$
Line of sight
45°
$\frac{3\lambda}{8}$
Line of sight
135°
45°
$\frac{\lambda}{2}$
Line of sight
135°
45°
$\frac{5\lambda}{8}$
Line of sight
135°
45°
$\frac{3\lambda}{4}$
Line of sight
45°
$\frac{7\lambda}{8}$
Line of sight
225°
45°
λ
Line of sight
225°

Figure 10 State of polarization of a light wave after passing through a crystal plate whose retardation is indicated in fractions of a wavelength (phase retardation $2\pi/\lambda$ times these values) and whose fast axis is indicated by the double arrow. In all cases the incident light is linearly polarized at an azimuth of 45° to the direction of the fast axis. (Adapted with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

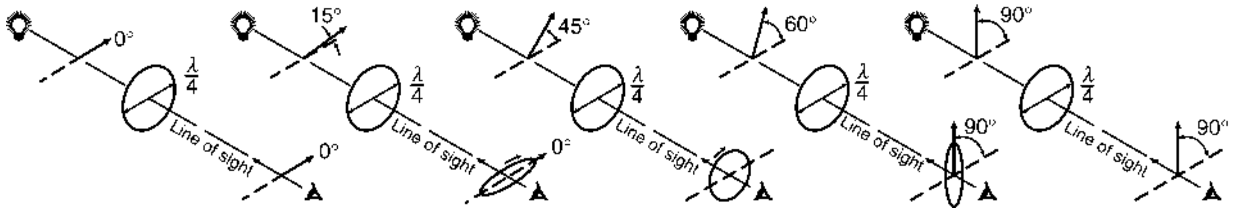

Figure 11 State of polarization of a light wave after passing through a λ/4 plate (whose fast axis is indicated by the double arrow) for different azimuths of the incident linearly polarized beam. (Adapted with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

uniaxial crystal. The fast axis is in the horizontal direction, the same as in Figure 9.

The main purpose of a quarter-wave plate is to change linearly polarized light into circularly polarized light. However, as Figure 11 shows, the state of polarization of the light will be quite different depending on the orientation of the plane of polarization of the incident beam relative to the fast axis of the quarter-wave plate.

A half-wave plate is used to rotate the plane of polarization of a linearly polarized beam. The plane of polarization is always rotated through an angle that is twice the angle the initial plane of polarization makes with the fast axis of the uniaxial crystal. A linearly polarized beam always remains linearly polarized.

Retardation plates are often made of mica, stretched polyvinyl alcohol, or crystal quartz, although they can also be made of other stretched plastics, sapphire, magnesium fluoride, and other materials.

Variable Retardation Plates and Compensators

Variable retardation plates are devices whose retardation can be varied in a variety of ways. They can be used to modulate or vary the phase of a beam of linearly polarized light or to analyze a beam of unknown polarization (often elliptically polarized light) such as might be produced by transmission through a birefringent material or by reflection from a metal or film-covered surface. The term compensator is often applied to a variable retardation plate since it can be used to compensate for the phase retardation produced by a material. Common types of compensators include the Babinet and Soleil compensators, in which the total thickness of a birefringent material in the light path is changed, the Sénarmont compensator which consists of a fixed quarter-wave plate and rotatable analyzer to compensate for varying amounts of ellipticity in a light beam, and tilting-plate compensators, which change the thickness of birefringent material in the light beam by changing the angle of incidence. Electro-optic and piezo-optic modulators can also be used as high-frequency variable retardation plates since their birefringence can be changed by varying the electric field or pressure (see next section).

The Babinet compensator, shown schematically in Figure 12 consists of two crystal quartz wedges, each with its axis in the plane of the face but with the axes at right angles to each other. One wedge is stationary, and the other can be moved in the direction indicated by the arrow, so that the total amount of quartz through which the light passes can be varied. The total retardation is proportional to the difference in thickness between the two wedges. This type of compensator was used extensively when the light source was a vertical slit and visual measurements were made of the state of polarization of a beam. However the bands representing different phase retardations were too narrow to be used effectively with a photoelectric detector, so the Babinet compensator was replaced by a Soleil compensator (Figure 13). This device, sometimes called a Babinet–Soleil compensator, is similar to the Babinet compensator except that the field of view has a uniform tint if the compensator is constructed correctly. This is because the ratio of the thicknesses of the two crystal quartz blocks (a movable wedge and a fixed wedge attached to a plate with the two axes perpendicular to each other) is the same over the entire field. The Soleil compensator will produce light of varying ellipticity depending on the position of the movable wedge. It is used in the same way as a Babinet compensator with the uniformly dark field of the Soleil compensator corresponding to the black zero-retardation band in the Babinet compensator.

It is sometimes necessary to accurately measure the azimuth of a beam of linearly polarized light using a photoelectric detector with a rotatable analyzer (i.e., a polarizer) directly in front it. The obvious method is to rotate the analyzer until the detector signal is a minimum and then read the analyzer angle, which equals the extinction angle (perpendicular to the azimuth of the linearly polarized beam). However, the

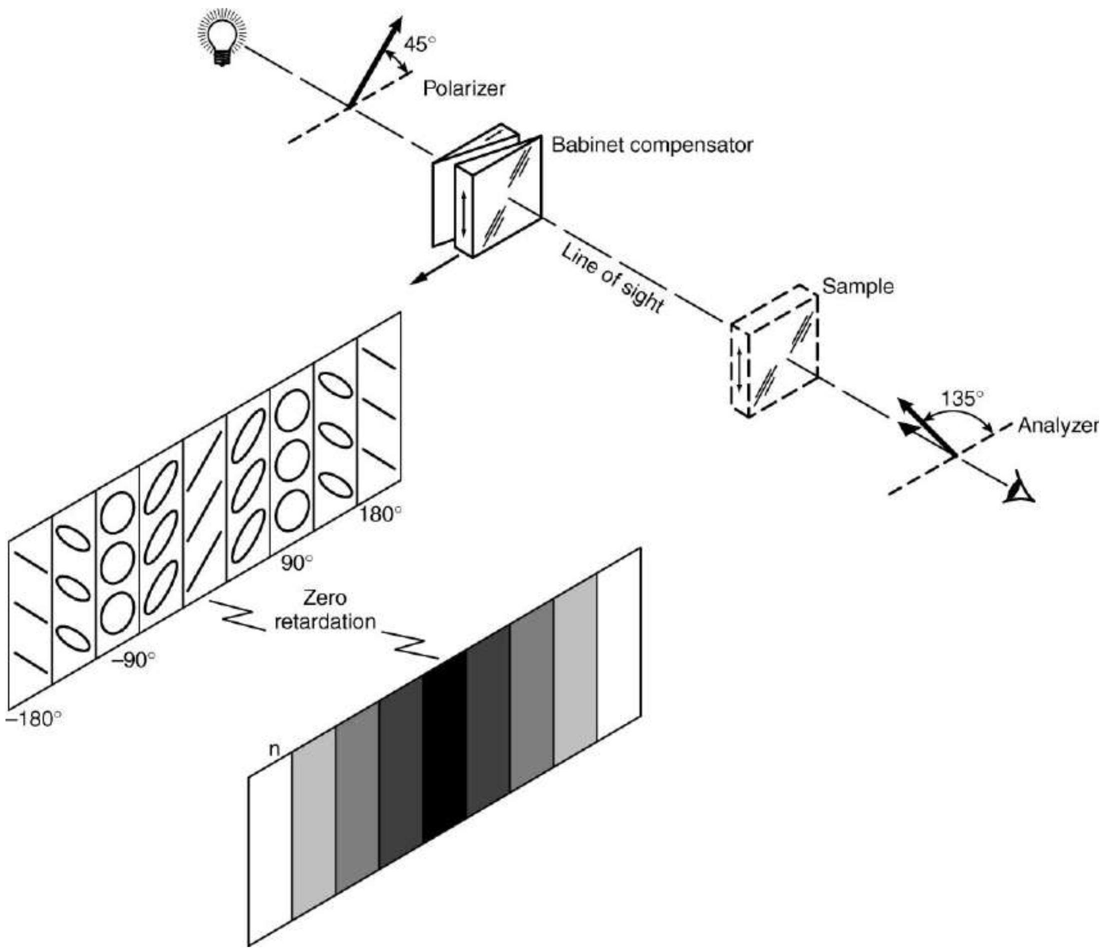

Figure 12 Arrangement of a Babinet compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. Retardations are indicated for alternate regions. After the beam passes through the analyzer, the field is crossed by a series of dark bands, one of which is shown to the left of the analyzer. (Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

angle can be determined more precisely if the analyzer is offset by a small angle from the extinction angle and the transmittance noted. Then the analyzer is offset by a small angle on the other side of the extinction angle at the angle where the transmittance is the same. One half the difference between these two azimuthal angles gives a more accurate value of the extinction angle than can be obtained by measuring it directly.

Before the days of photoelectric detectors, half-shade devices were extensively used to determine the azimuth of a linearly polarized beam. The device consisted of two polarizers with their axes inclined at a small angle to each other. As the device was rotated, one part of the field became darker and the other part became lighter. At the match position, both parts of the field appeared equally bright. There were a variety of these devices as well as ellipticity half-shade devices that could detect very small amounts of ellipticity in a nominally linearly polarized beam and hence could verify when a compensator had completely converted elliptically polarized light into linearly polarized light.

Electro-Optic, Magneto-Optic, and Piezo-Optic Devices

The state of polarization of light can be rapidly altered by passing it through a material that has electro-optic, magneto-optic or piezo-optic properties. The voltage, magnetic field, or pressure are varied to make the material birefringent (see above). Some materials are isotropic without the applied field, i.e., the refractive index is the same in all directions. Other materials have an optic axis

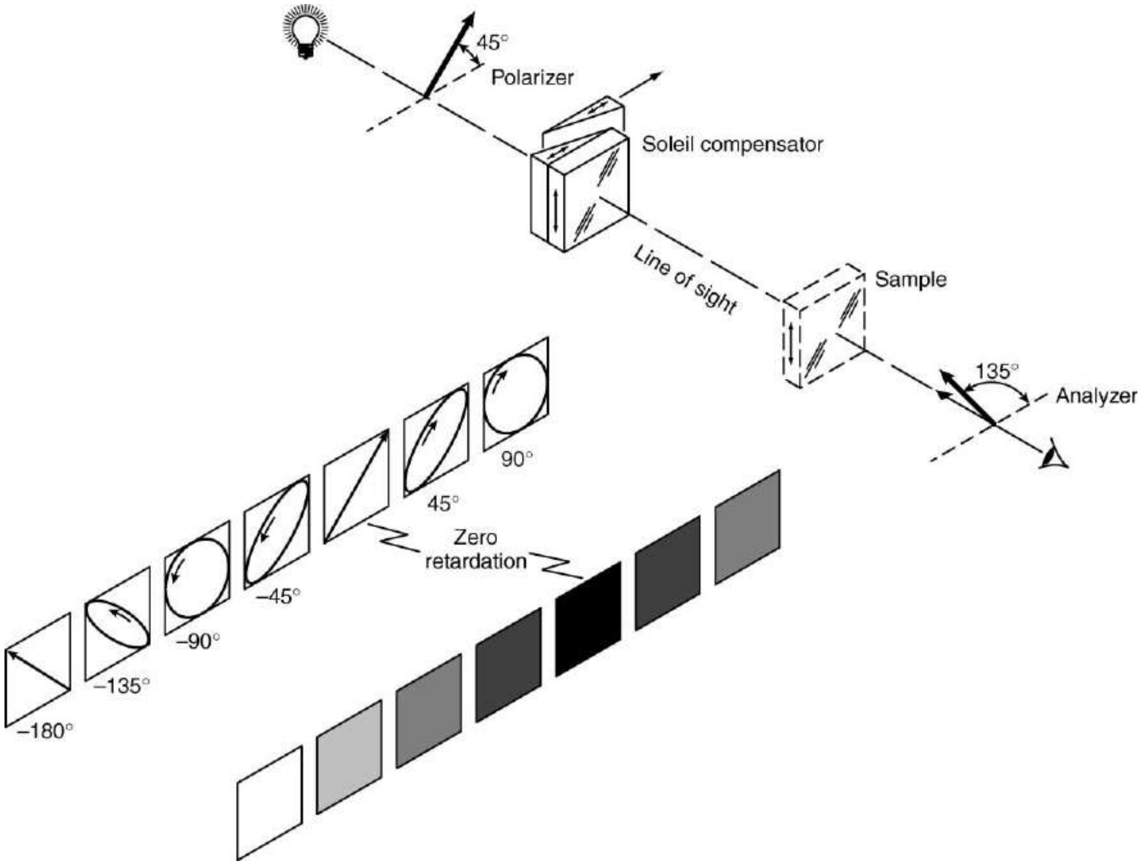

Figure 13 Arrangement of a Soleil compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. After the beam passes through the analyzer, the field appears as one of the shades of gray shown to the left of the analyzer. (Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

(uniaxial materials) or two optic axes (biaxial materials). In these cases, the applied field creates further asymmetries in the material.

The most common electro-optic, magneto-optic and piezo-optic effects are the Pockel's effect (depending on the electric field), the Kerr effect (depending on the square of the electric field), the Faraday effect (depending on the magnetic field), the Cotton–Mouton and Voigt effects (depending on the square of the magnetic field), and stress birefringence or the photoelastic effect (depending on pressure changes). These effects are shown in Table 1 along with order of magnitude strengths needed to produce changes in the refractive index, and materials that most strongly exhibit the effects. The mathematical descriptions of the effects involve tensors and are too involved to present here. However, simple physical descriptions will be given.

If a varying electric field acts on an electro-optic material, electrons, ions, or permanent dipoles in the material are made to reorient, inducing an electric polarization. The induced polarization creates birefringence that modifies the optical polarization of a light beam passing through the material. The electric field strength determines how the polarization is changed (see Figure 10).

Linearly polarized light passing through a magneto-optic material will be rotated in a direction parallel to the direction of the applied magnetic field. This phenomenon, called the Faraday effect, or Faraday rotation, is similar to what happens in optically active materials (see the next paragraph) except that the Faraday effect depends on the direction of the magnetic field, but not on the direction the light is traveling.

There are naturally optically active materials such as camphor, nicotine, sugar solutions, and crystal quartz that can rotate the plane of linearly polarized light passing through the material. For example, the Si–O bonds in crystal quartz form a helical path around the optic axis. This crystal structure interacts with an incoming linearly polarized beam traveling parallel to the optic axis and rotates it in a clockwise direction. A 1 mm-thick piece of quartz will rotate a

Table 1 Electro-optic, magneto-optic, and piezo-optic effects

Effect	*Device*	*Effect proportional to*	*Strength* (Δn)	*Materials*
Pockels (electro-optic)	Pockels cell	E	$10^{-12}E$	$BaTiO_2$, $LiNBO_3$
Kerr	Kerr cell	E^2	$10^{-19}E^2$	Nitrobenzene, Benzonitrile
Faraday (magneto-optic)	Faraday rotation	H	$10^{-14}H$	Zns, GaAs, CS_2
Cotton–Mounton	[a]	H^2	$10^{-25}H^2$	Chloroform, acetone
Photoelastic effect; stress birefringence	[b]	Pressure	–	Fused silica, polystyrene, Ge, KDP, ruby

[a] Too small to be of technological importance.

[b] Depends on the mounting or the processing of the material.

Adapted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

linearly polarized beam by 21.7°. Other materials have asymmetric structures that have mirror images. A mirror image structure cannot be obtained by simply rotating the group of atoms in space. These materials are also optically active. The amount of rotation produced by an optically active material is proportional to the thickness (optical density) of the material that the light passes through. Dextrose (a form of glucose) and levulose (also known as fructose) rotate the plane of linearly polarized light in clockwise and counterclockwise directions, respectively. Many other organic molecules have D- (right handed, dextro-rotary) and L- (left handed, levo-rotary) mirror image forms.

An isotropic material can become anisotropic when stress or an induced strain is applied because of an elasto-optic interaction known as stress birefringence, or the photoelastic effect. This effect is usually bad because it reduces the performance of optical components and devices by introducing phase distortions caused by improper mounting or unequal thermal expansion between the mounts and components. Another source of distortion is strain that has been frozen into optical components during manufacture. French curves are excellent examples that show colored strain birefringence patterns when viewed between crossed polarizers. In one application, strain birefringence has been used constructively to produce a variable phase retarder made from crystal quartz and fused silica. A block of electro-optic crystal quartz, is cemented to a block of isotropic fused silica. When a variable electric field is applied to the crystal quartz, its length changes at the resonant frequency of the block. This produces strain in the fused silica block which in turn produces a variable retardation of light passing through the strained fused silica. Depending on the magnitude of the varying electric field, the device can act like a variable quarter-wave plate, a variable half-wave plate, or have other applications.

Matrix Methods for Computing Polarization

An optical system containing various polarizing and retarding devices can be modeled using matrices. In general, there is an incident beam (in matrix form), that has some state of polarization. It interacts with a device called an instrument (also in matrix form) that alters the state of polarization, so that the exiting beam (a third matrix, the product of the first and second matrices) has another state of polarization. There can be more than one instrument matrix acting on the same incident matrix to produce the final matrix. The two most common matrix methods are the Mueller calculus and the Jones calculus.

There is also a visual representation of this process, a Poincaré sphere, on which vectors represent different states of polarization. The polarization instrument moves the vector representing the incident polarization around the sphere in a prescribed manner. The vectors are called Stokes vectors and are used in the Mueller calculus. The various states of polarization are represented on the Poincaré sphere as follows: The equator represents various forms of linear polarization, the poles represent right- and left-circular polarization, and other points on the sphere represent elliptically polarized light. Every point on the sphere corresponds to a different polarization form. The radius of the sphere represents the incident irradiance of the light beam (which is usually assumed to be unity). The effects of various polarization instruments are determined by displacements on the sphere. Partially polarized light or absorption may be dealt with approximately by ignoring the irradiance factor, since the state of polarization is generally the quantity desired. The Poincaré sphere is most useful for visualizing problems involving nonabsorbing materials, various polarization instruments including polarizers, retarders, compensators, half-shade devices, and depolarizers, and has also been applied to ellipsometric problems and stress-optical measurements.

The Poincaré sphere is used with Stokes vectors, which are often designated S_0, S_1, S_2, and S_3. S_0 is the incident irradiance of the light beam, corresponding to the radius of the Poincaré sphere. S_1 is the difference in irradiances between the horizontal and vertical polarization components of the beam; for example, when S_1 is positive, the preference is for horizontal polarization. S_2 indicates preference for $+45°$ or $-45°$ polarization depending on whether it is positive or negative, and S_3 gives the preference for right or left circular polarization. Thus, the Stokes vectors S_1, S_2, and S_3 are simply the three Cartesian coordinates of a point on the Poincaré sphere. S_1 and S_2 are perpendicular to each other in the equatorial plane, and S_3 points toward the north pole of the sphere. Any state of polarization of a light beam can be specified by these three vectors. The irradiance vector S_0 is related to the other three by the relation $S_0^2 = S_1^2 + S_2^2 + S_3^2$ when the beam is completely polarized. If the beam is partially polarized, $S_0^2 > S_1^2 + S_2^2 + S_3^2$.

In the Mueller calculus, the incident beam is represented by the four-component Stokes vector, written as a column vector. This vector has all real elements and gives information about irradiance properties of the beam. Thus it is not able to handle problems involving phase changes or combinations of two beams that are coherent. The instrument matrix is a 4×4 matrix with all real elements.

In the Jones calculus, the Jones vector representing the incident beam is a two-component column vector that generally has complex elements. It contains information about the *amplitude* properties of the beam and hence is well suited for handling problems involving the phases of light waves. However, it cannot handle problems involving depolarization, as the Mueller calculus can. The Jones instrument matrix is a 2×2 matrix whose elements are generally complex. The Jones calculus is well suited to problems involving a large number of similar devices arranged in series in a regular manner and makes it possible to obtain a result expressed explicitly in terms of the number of such devices. The Jones instrument matrix of a train of transparent or absorbing nondepolarizing polarizers and retarders contains no redundant information. The matrix contains four elements, each of which has two parts, so that there are a total of eight constants, none of which is a function of any other. The Mueller instrument matrix of such a train contains much redundancy; there are 16 constants but only 7 of them are independent. More information about the Poincaré sphere and the Mueller and Jones calculus is available in other references.

See also

Electromagnetically Induced Transparency. **Magneto-optics:** Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping; Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering.

Further Reading

Auton JP (1967) Infrared transmission polarizers by photolithography. *Applied Optics* 6: 1023–1027.

Azzam, Rasheed MA and Bashara NM (1977) *Ellipsometry and Polarized Light*, chaps. 1 and 2. Amsterdam: North-Holland Publishing Co.

Azzam, Rasheed MA (1995) Ellipsometry. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 27. New York: McGraw-Hill.

Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill.

Bennett M (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill.

Bohren CF and Huffman DR (1983) *Absorption and Scattering of Light by Small Particles*. New York: John Wiley.

Born M and Wolf E (1999) *Principles of Optics*, 7th edn, chap. 1. Cambridge: Cambridge University Press.

Chipman RA (1995) Polarimetry. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 22. New York: McGraw-Hill.

Clark JR (1956) New calculus for the treatment of optical systems. VIII. Electromagnetic theory. *Journal of the Optical Society of America* 46: 126–131.

Guenther RD (1990) *Modern Optics*. chaps. 2, 13, and 14. New York: John Wiley and Sons.

Hass M and O'Hara M (1965) Sheet infrared transmission polarizers. *Applied Optics* 4: 1027–1031.

Hecht E and Zajac A (1997) *Optics*, 3rd edn, chap. 8. Reading, MA: Addison-Wesley Publishing Co.

Jackson JD (1999) *Classical Electrodynamics*, 3rd edn, chap. 7. New York: John Wiley and Sons, Inc.

Jasperson SN and Schnatterly SE (1969) An improved method for high reflectivity ellipsometry based on a new polarization modulation technique. *Reviews of Scientific Instruments* 40: 761–767.

Kerker M (1969) *The Scattering of Light and Other Electromagnetic Radiation*. New York: Academic Press.

Kliger DS, Lewis JW and Randall CE (1990) *Polarized Light in Optics and Spectroscopy*, chaps. 1–5. San Diego: Academic Press.

Maldonado TA (1995) Electro-optic modulators. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 13. New York: McGraw-Hill, Inc.

Markuvitz N (ed.) (1951) *Waveguide Handbook*, pp. 218–229, 285–289. New York: McGraw-Hill Book Co.

Shurcliff WA (1962) *Polarized Light, Production and Use*. Cambridge, MA: Harvard University Press.

van de Hulst HC (1957) *Light Scattering by Small Particles*. New York: John Wiley.

Matrix Analysis

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hooke postulated, in the seventeenth century, that light waves must be transverse but his idea was forgotten. Young and Fresnel put forward the same idea in the nineteenth century and accompanied their postulation with a theoretical description of light based on transverse waves. Forty years later, Maxwell proved that light must be a transverse wave and that **E** and **H**, for a plane wave in an isotropic medium with no free charge and no currents, are mutually perpendicular and lie in a plane normal to the direction of propagation, **k**.

Convention requires that we use the electric vector to label the direction of the electromagnetic wave's polarization. The direction of the displacement vector is called the direction of polarization and the plane containing the direction of polarization and the propagation vector is called the plane of polarization. The selection of the electric field is not completely arbitrary, except in relativistic situations, when $v \approx c$, the interaction of the electromagnetic wave with matter will be dominated by the electric field.

Assume that a plane wave is propagating in the z-direction. In complex notation, the plane wave is given by

$$\tilde{\mathbf{E}} = \mathbf{E}_0\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi)} = \mathbf{E}_0\, e^{i(\omega t - kz + \phi)} \qquad [1]$$

The procedure used to decompose an arbitrary polarization into components parallel to two axes of a Cartesian coordinate system, is a technique used extensively in vector algebra to simplify mathematical calculations. According to the mathematical formalism associated with this technique, the polarization is described in terms of a set of basis vectors, $\mathbf{e}_i$. An arbitrary polarization would be expressed as

$$\mathbf{E} = \sum_{i=1}^{2} a_i \mathbf{e}_i \qquad [2]$$

The set of basis vectors, $\mathbf{e}_i$, are orthonormal, i.e.:

$$\mathbf{e}_i \mathbf{e}_j^* = \delta_{ij} = \begin{Bmatrix} 1 & i = j \\ 0 & i \neq j \end{Bmatrix} \qquad [3]$$

where we have assumed that the basis vectors could be complex. We mention this mathematical formalism because an identical formalism is encountered in a description of spin.

In a Cartesian coordinate system, the $\mathbf{e}_i$'s are the unit vectors **i**, **j**, **k**. The summation in eqn [2] extends over only two terms because the electromagnetic wave is transverse, confining **E** to a plane normal to the direction of propagation (according to the coordinate convention we have selected, the **E** field is in the x, y plane).

Polarization Ellipse

Following the formalism of eqn [2], a polarized wave can be written in terms of the x and y components of $\mathbf{E}_0$

$$\tilde{\mathbf{E}} = \mathbf{E}_{0x}\, e^{i(\omega t - kz + \phi_1)}\hat{\mathbf{i}} + \mathbf{E}_{0y}\, e^{i(\omega t - kz + \phi_2)}\hat{\mathbf{j}} \qquad [4]$$

We will use only the real part of **E** for manipulation, to prevent errors. We divide each component of the electric field by its maximum value, so that the problem is reduced to one of the following two sinusoidal varying unit vectors:

$$\begin{aligned} \frac{E_x}{E_{0x}} &= \cos(\omega t - kz + \phi_1) \\ &= \cos(\omega t - kz)\cos\phi_1 - \sin(\omega t - kz)\sin\phi_1 \\ \frac{E_y}{E_{0y}} &= \cos(\omega t - kz)\cos\phi_2 - \sin(\omega t - kz)\sin\phi_2 \end{aligned} \qquad [5]$$

When these unit vectors are added together, the resulting equation will describe the path taken by the tip of the resultant vector. The path will create a Lissajous figure.

To obtain the equation for the Lissajous figure, we eliminate the dependence of the unit vectors on $(\omega t - kz)$. First, multiply the equations by $\sin\phi_2$ and $\sin\phi_1$, respectively and then subtract the resulting equations. Second, multiply the two equations by $\cos\phi_2$ and $\cos\phi_1$, respectively and then subtract the new equations. These two operations yield the following pair of equations:

$$\begin{aligned} &\frac{E_x}{E_{0x}}\sin\phi_2 - \frac{E_y}{E_{0y}}\sin\phi_1 \\ &\quad = \cos(\omega t - kz)[\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2] \end{aligned} \qquad [6]$$

$$\frac{E_x}{E_{0x}}\cos\phi_2-\frac{E_y}{E_{0y}}\cos\phi_1$$

$$=\sin(\omega t-kz)[\cos\phi_1\sin\phi_2-\sin\phi_1\cos\phi_2] \qquad [7]$$

The term in brackets can be simplified using the trig identity

$$\sin\delta=\sin(\phi_2-\phi_1)$$

$$=\cos\phi_1\sin\phi_2-\sin\phi_1\cos\phi_2 \qquad [8]$$

After replacing the term in brackets by sin δ, the two equations are squared and added yielding the equation for the Lissajous figure:

$$\left(\frac{E_x}{E_{0x}}\right)^2+\left(\frac{E_y}{E_{0y}}\right)^2-\left(\frac{2E_xE_y}{E_{0x}E_{0y}}\right)\cos\delta=\sin^2\delta \qquad [9]$$

The trig identity

$$\cos\delta=\cos(\phi_2-\phi_1)=\cos\phi_1\cos\phi_2+\sin\phi_1\sin\phi_2$$

was also used to further simplify eqn [9].

Equation [9] has the same form as the equation of a conic

$$Ax^2+Bxy+Cy^2+Dx+Ey+F=0$$

Geometry defines the conic as an ellipse because, from eqn [9]

$$B^2-4AC=\frac{4}{E_{0x}^2E_{0y}^2}(\cos^2\delta-1)<0$$

This ellipse is called the polarization ellipse. The orientation of the ellipse, with respect to the x-axis, is

$$\tan 2\theta=\frac{B}{A-C}=\frac{2E_{0x}E_{0y}\cos\delta}{E_{0x}^2-E_{0y}^2} \qquad [10]$$

The tip of the resultant electric field vector obtained from eqn [4] traces out the polarization ellipse in the plane normal to **k**, as predicted by eqn [9]. The ratio of the length of the minor to the major axis of the ellipse is equal to the ellipticity, φ, i.e., the amount of deviation of the ellipse from a circle

$$\tan\varphi=\pm\left(\frac{E_m}{E_M}\right)$$

$$=\frac{E_{0x}\sin\phi_1\sin\theta-E_{0y}\sin\phi_2\cos\theta}{E_{0x}\cos\phi_1\cos\theta+E_{0y}\cos\phi_2\sin\theta} \qquad [11]$$

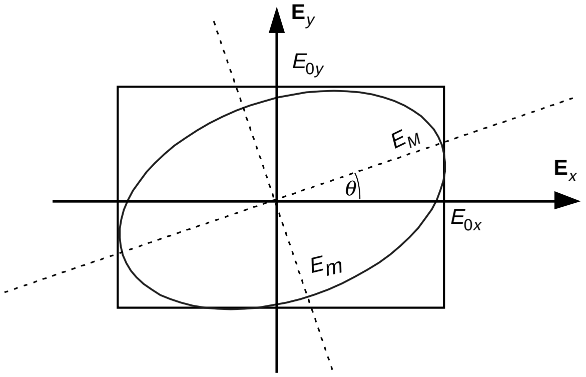

Figure 1 General form of the ellipse described by eqn [9]. Reproduced with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

In particle physics, the light would be said to have a negative helicity if it rotated in a clockwise direction. If we look at the source, the electric vector seems to follow the threads of a left-handed screw, agreeing with the nomenclature that left-handed quantities are negative. However, in optics the light that rotates clockwise, as we view it traveling toward us from the source, is said to be right-circularly polarized. The counterclockwise rotating light is left-circularly polarized. The association of right-circularly polarized light, with 'right-handedness' in optics, came about by looking at the path of the electric vector in space at a fixed time, then $\tan\psi=\tan(\phi-kz)$ (Figure 1).

Stokes Parameters

In the formalism associated with eqn [2], the expansion coefficients, a_i, can be used to form a 2−2 matrix which, in statistical mechanics, is called the density matrix and in optics, the coherency matrix (Table 1). The elements of the matrix are formed by the rule

$$\rho_{ij}=\mathbf{a}_i\mathbf{a}_j^* \qquad [12]$$

The matrix is Hermitian, so that $\rho_{ij}=\rho_{ji}^*$. We will not develop the theory of polarization using the coherency matrix, but simply use the coherency matrix to justify the need for four independent measurements to characterize polarization. There is no unique set of measurements required by the theory, but normally measurements made are of the Stoke's parameters which are directly related to the parameters of the polarization ellipse of eqn [9].

The Stokes parameters of a light wave are measurable quantities, defined as:

$s_0 \Rightarrow$ Total flux density
$s_1 \Rightarrow$ Difference between flux density transmitted by a

Table 1 Typical Stokes vectors

Horizontal polarization	$\begin{bmatrix}1\\1\\0\\0\end{bmatrix}$	Vertical polarization	$\begin{bmatrix}1\\-1\\0\\0\end{bmatrix}$
+45° polarization	$\begin{bmatrix}1\\0\\1\\0\end{bmatrix}$	−45° polarization	$\begin{bmatrix}1\\0\\-1\\0\end{bmatrix}$
Right circular polarization	$\begin{bmatrix}1\\0\\0\\1\end{bmatrix}$	Left circular polarization	$\begin{bmatrix}1\\0\\0\\-1\end{bmatrix}$

linear polarizer oriented parallel to the x-axis and one oriented parallel to the y-axis. The x and y-axes are usually selected to be parallel to the horizontal and vertical directions in the laboratory.

$s_2 \Rightarrow$ Difference between flux density transmitted by a linear polarizer oriented at 45° to the x-axis and one oriented at 135°.

$s_3 \Rightarrow$ Difference between flux density transmitted by a right-circular polarizer and a left-circular polarizer.

If the Stokes parameters are to characterize the polarization of a wave, they must be related to the parameters of the polarization ellipse. It is therefore important to establish that the Stokes parameters are variables of the polarization ellipse (eqn [9]).

In its current form, eqn [9] contains no measurable quantities and thus must be modified if it is to be associated with the Stokes parameters. The time average of the Poynting vector is the quantity observed when measurements are made of light waves. We must, therefore, find the time average of eqn [9] if we wish to relate its parameters to observable quantities.

The time average eqn [9] can now be written as

$$\frac{\langle E_x^2\rangle}{E_{0x}^2} + \frac{\langle E_y^2\rangle}{E_{0y}^2} - 2\frac{\langle E_x E_y\rangle}{E_{0x}E_{0y}}\cos\delta = \sin^2\delta \qquad [13]$$

where time average is denoted by $\langle\ \rangle$

$$\langle E_x^2\rangle = \frac{1}{T}\int_{t_0}^{t_0+T} E_{0x}^2[\cos(\omega t - kz)\cos\phi_1 - \sin(\omega t - kz)\sin\phi_1]^2 \mathrm{d}t \qquad [14]$$

With the time averages, eqn [13] can be written as

$$(E_{0x}^2 + E_{0y}^2)^2 - (E_{0x}^2 - E_{0y}^2)^2 - (2E_{0x}E_{0y}\cos\delta)^2 = (2E_{0x}E_{0y}\sin\delta)^2 \qquad [15]$$

Each term in this equation can be identified with a Stokes parameter:

$$s_0 = \langle E_{0x}^2\rangle + \langle E_{0y}^2\rangle \qquad s_1 = \langle E_{0x}^2\rangle - \langle E_{0y}^2\rangle$$
$$s_2 = \langle 2E_{0x}E_{0y}\cos\delta\rangle \qquad s_3 = \langle 2E_{0x}E_{0y}\sin\delta\rangle \qquad [16]$$

Equation [15] can now be written as

$$s_0^2 - s_1^2 - s_2^2 = s_3^2 \qquad [17]$$

For a polarized wave, only three of the Stokes parameters are independent. This agrees with the requirement placed upon elements of the Hermitian, coherency matrix, introduced above.

With this demonstration of the connection between the Stokes parameters and the polarization ellipse, the Stokes parameters can be written in terms of the parameters of the polarization ellipse.

$$s_1 = s_0\cos 2\varphi\cos 2\theta$$
$$s_2 = s_0\cos 2\varphi\sin 2\theta \qquad [18]$$
$$s_3 = s_0\sin 2\varphi$$

Mueller Calculus

Mueller pointed out that the Stokes parameters can be thought of as elements of a column matrix or a 4-vector (Table 2)

$$\mathcal{S} = \begin{pmatrix} s_0\\ s_1\\ s_2\\ s_3\end{pmatrix} \qquad [19]$$

Each element of the Stokes vector represents a measurable intensity. The vector can represent not only polarized light, but unpolarized or partially polarized light.

In 1929, Soleillet discovered that an optical device performed a linear transformation on the input wave and, in 1942, Perrin put this fact into a matrix formalism involving the Stokes vectors:

$$\mathcal{S}_1 = \mathcal{M}\cdot\mathcal{S}_0 \qquad [20]$$

Mueller used experimentally derived 4 × 4 matrices, the $\mathcal{M}$ in eqn [20], to describe the effect of an optical device on a light wave's polarization. The $\mathcal{S}$'s in eqn [20] are column matrices whose elements

Table 2 Mueller matrices for polarizers

Polarizer	*Transmission axis*	*Mueller matrix*
Linear	Horizontal	$\frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	Vertical	$\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	+45°	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	−45°	$\frac{1}{2}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
Circular	Right	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$
	Left	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$

Table 3 Mueller matrices for retarders

Retarder	*Transmission axis*	*Mueller matrix*
Quarter-wave plate	Horizontal	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
	+45°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$
	−45°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$
Half-wave plate	0° or 90°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$
	±45°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$

are the Stokes parameters (eqn [16]). The matrices are based upon an assumed linear relationship between the incident and transmitted beams (Table 3).

The analysis of the effect of a number of polarizers and retarders is made easier by the use of the Mueller–Stokes matrix calculus, coupled with the use of the Stokes vectors. To determine the Mueller matrices, one must measure the effect a device has on unpolarized, horizontally polarized, linearly polarized at + 45°, and right-circularly polarized light. It is then an algebraic exercise to generate the elements of the matrix. A few optical devices and their Mueller matrix are listed in Tables 2 and 3.

The Mueller matrix contains 16 parameters but there are only seven independent parameters. The matrix contains no information about absolute phase but it handles partially polarized and unpolarized light without modification (Table 4).

Jones Vector

There is one other representation of polarized light, complementary to the Stokes vector, developed by Clark Jones in 1941 and called the Jones vector. It is superior to the Stokes vector in that it handles light of a known phase and amplitude with a reduced number of parameters. However, it is inferior to the Stokes vector in that, unlike the Stokes representation, which is experimentally determined, the Jones representation cannot handle unpolarized or partially polarized light. The Jones vector is a theoretical construct that can only describe light with a well-defined phase and frequency. The density matrix formalism can be used to correct the shortcomings of the Jones vector, but then the simplicity of the Jones representation is lost.

Assuming the coordinate system is such that the electromagnetic wave is propagating along the z-axis, it was shown earlier that any polarization could be decomposed into two orthogonal **E** vectors, i.e., parallel to the x and y directions. The Jones vector is defined as a two-row, column matrix consisting of the complex components in the x and y direction:

$$\mathbf{E} = \begin{bmatrix} E_{0x}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_1)} \\ E_{0y}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_2)} \end{bmatrix} \qquad [21]$$

Table 4 Jones vectors

Horizontal polarization	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Vertical polarization	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
+45° polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	−45° polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$
Right circular polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}$	Left circular polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix}$

If absolute phase is not an issue, then we may normalize the vector by dividing by that number (real or complex) that simplifies the components but keeps the sum of the square of the components equal to one. For example, assume that $E_{0x} = E_{0y}$, then

$$\mathbf{E} = E_{0x}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_1)} \begin{bmatrix} 1 \\ e^{i\delta} \end{bmatrix} \qquad [22]$$

The normalized vector would be the terms contained within the bracket, each divided by $\frac{1}{\sqrt{2}}$.

The general form of the Jones vector is

$$\mathbf{E} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \qquad \mathbf{E}^* = [\mathbf{A}^*\mathbf{B}^*] \qquad [23]$$

Some examples of Jones vectors are shown in Table 4.

Jones Calculus

The Jones calculus is complementary to the Mueller calculus and operates on the Jones vector, (eqn [23]), similar to the way the Mueller matrix operates on the Stokes vector (Table 5).

The Jones matrix contains eight independent parameters with no redundancy, making it simpler than the Mueller calculus. However, the Jones calculus only applies to polarized light. The Jones calculus can be extended, using the density matrix formalism to allow manipulation of unpolarized

Table 5 Jones matrices for retarders

Retarder	*Transmission axis*	*Jones matrix*
Quarter-wave plate	Horizontal	$\begin{bmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{bmatrix}$
	Vertical	$\begin{bmatrix} e^{-i\pi/4} & 0 \\ 0 & e^{i\pi/4} \end{bmatrix}$
	+45°	$\frac{1}{2}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$
	−45°	$\frac{1}{2}\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}$
Half-wave plate	0° or 90°	$\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$
	±45°	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Table 6 Jones matrices for polarizers

Polarizer	*Transmission axis*	*Jones matrix*
Linear	Horizontal	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
	+45°	$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
	−45°	$\frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$
Circular	Right	$\frac{1}{2}\begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$
	Left	$\frac{1}{2}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}$

light, but with a loss of simplicity. The matrix equation for Jones calculus is

$$\mathbf{E}_{\text{out}} = \mathcal{W}\mathbf{E}_{\text{in}} \qquad [24]$$

A few optical devices and their Jones matrices, $\mathcal{W}$, are listed in Tables 5 and 6.

For every matrix in Jones calculus, there is a matrix in Mueller calculus, but the converse is not true. For example, it is possible to construct a depolarizer, by using a thick piece of opal glass in the visible or by using gold covered sandpaper in the infrared. Such a device can be described in Mueller calculus, but there is no matrix for such a device in Jones calculus (Tables 5 and 6).

See also

Polarization: Introduction.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Clark Jones R (1941) A new calculus for the treatment of optical systems. *Journal of the Optical Society of America* 32: 488–503.

Clark Jones R (1942) A new calculus for the treatment of optical systems. *Journal of the Optical Society of America* 31: 486–493.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Kliger DS, Lewis JW and Randall CE (1990) *Polarized Light in Optics and Spectroscopy*. Boston, MA: Academic Press.

McMaster WH (1954) Polarization and the Stokes parameters. *American Journal of Physics* 22: 351–362.

Mueller H (1948) The foundation of optics. *Journal of the Optical Society of America* 38: 661.

Q

QUANTUM ELECTRODYNAMICS

Contents

Quantum Theory of the Electromagnetic Field
Cavity QED
Cavity QED in Semiconductors

Quantum Theory of the Electromagnetic Field

I Bialynicki-Birula and Z Bialynicka-Birula,
Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

The quantization of the free electromagnetic field is explained by analogy with the quantization of a collection of harmonic oscillators.

Brief History

The first step in the formulation of the quantum theory of the electromagnetic field (EMF) was the 'quantum hypothesis' put forward in 1900 by Planck. According to this hypothesis the exchange of energy between the EMF and the cavity walls occurs in portions $h\nu$, where h – the quantum of action – is a new fundamental constant (the Planck constant) and ν is the frequency of the material oscillator in the wall. With the help of this hypothesis Planck explained the spectrum of the electromagnetic radiation in thermodynamic equilibrium with the cavity walls. The second step was made in 1905 by Einstein, who postulated that the quantum hypothesis of Planck should be applied directly to the EMF. The assumption that the energy of the electromagnetic radiation is carried by light quanta (named photons by Lewis in 1926), each of energy $h\nu$, enabled Einstein to explain the puzzling properties of the photoelectric effect, namely, the gross disparity between the tiny classical energy flux delivered to an electron and the amount of energy needed to eject the photo-electron from the metal. After the formulation of quantum mechanics, it became possible to complete the quantum theory of the EMF (Dirac 1927) by applying the general rules of quantization to the classical Maxwell theory. However, this theory ran into difficulties when it was applied to the interaction of the EMF with charged matter.

The lightest carriers of the electric charge are the electrons and their dynamics determines to a very large extent the electromagnetic properties of ordinary matter. The main principles of quantum electrodynamics (QED) – the theory unifying the quantum theory of the EMF and the relativistic theory of electrons (Dirac 1928) – have been set forth by Heisenberg and Pauli already in 1929 but the proper tool (renormalization theory) to perform highly accurate calculations was developed by Feynman and Schwinger in 1949. In 1967 QED was unified by Weinberg and Salam with the theory of weak interactions. Later, after the unification with the theory of strong interactions, QED became part of the so called 'standard model' – the present day theory of elementary ingredients of matter: quarks, leptons, gluons, intermediate bosons, and photons. Therefore, the modern quantum theory of the EMF is not a self-contained theory but part of a much more encompassing theory. Often it requires some input from a more general theory or from phenomenology. However, the quantum theory of the EMF with only a very crude description of its coupling to the rest of the world (for example, in the form of boundary conditions or constitutive relations) is still not only a good starting point to a more elaborate theory but also a source of important information about the behavior of photons under realistic conditions. The main reason why treating photons as free particles gives a reasonably good approximation is the smallness of the Sommerfeld fine structure constant $\alpha = e^2/4\pi\varepsilon_0\hbar c \approx 1/137$ that measures the strength of the electromagnetic interactions between photons and charged particles.

Electromagnetic Field as a Collection of Harmonic Oscillators

In order to apply the same quantization procedure as in quantum mechanics, it is useful to formulate classical electrodynamics in a way that exhibits its close relationship with mechanical systems. Many properties of the electromagnetic field may be described and understood in the simplest case of the free field. Therefore, in this article we restrict ourselves to the quantum description of the free EMF.

It was discovered by Rayleigh and Jeans that the electromagnetic field may be viewed as a collection of harmonic oscillators. This analogy enables one to apply well-established methods of simple quantum mechanics to describe the much more complicated problem of the quantum field. The dynamics of a single, one-dimensional harmonic oscillator is governed by the Hamiltonian:

$$H(p, q) = \frac{p^2}{2m} + \frac{m\omega^2 q^2}{2}$$

The equations of motion for the position and momentum of a harmonic oscillator have the same form in classical and in quantum theory:

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t} = \frac{p(t)}{m}, \qquad \frac{\mathrm{d}p(t)}{\mathrm{d}t} = -m\omega^2 q(t)$$

The solutions of these equations describe harmonic oscillations with the frequency ω

$$q(t) = q_0 \cos(\omega t) + (p_0/m\omega)\sin(\omega t)$$

$$p(t) = -m\omega q_0 \sin(\omega t) + p_0 \cos(\omega t)$$

A complete description of a multidimensional harmonic oscillator may be obtained in terms of normal modes. The transformation to normal modes leads to a decomposition of the Hamiltonian of the system into the sum, one term for each mode,

$$H = \sum_k \left(\frac{p_k^2}{2m} + \frac{m\omega_k^2 q_k^2}{2} \right)$$

Thus, each mode may be viewed as a separate system – a one-dimensional harmonic oscillator with characteristic frequency ω_k – noninteracting with the remaining oscillators.

The Maxwell equations for the free EMF read

$$\partial_t \mathbf{B}(\mathbf{r}, t) = -\nabla \times \mathbf{D}(\mathbf{r}, t)/\varepsilon, \ \nabla \cdot \mathbf{B}(\mathbf{r}, t) = 0$$

$$\partial_t \mathbf{D}(\mathbf{r}, t) = \nabla \times \mathbf{B}(\mathbf{r}, t)/\mu, \ \nabla \cdot \mathbf{D}(\mathbf{r}, t) = 0$$

In the vacuum, the constants ε and μ must be replaced by ε_0 and μ_0. The Maxwell equations lead to the wave equation for the vector $\mathbf{D}(\mathbf{r}, t)$

$$\left(\frac{1}{c^2}\partial_t^2 - \Delta \right) \mathbf{D}(\mathbf{r}, t) = 0$$

and to the same equation for the vector $\mathbf{B}(\mathbf{r}, t)$. The solution of this equation obtained by the separation of variables has the form

$$\mathbf{D}(\mathbf{r}, t) = a(t)\mathbf{u}(\mathbf{r})$$

where $a(t)$ satisfies the harmonic equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} a(t) + \omega^2 a(t) = 0$$

and $\mathbf{u}(\mathbf{r})$ satisfies the Helmholtz equation

$$(\Delta + (\omega/c)^2)\mathbf{u}(\mathbf{r}) = 0$$

Without going into mathematical details we shall assume that there exist a complete set of vector functions $\mathbf{u}_k(\mathbf{r})$, called the mode functions, satisfying the Helmholtz equations

$$(\Delta + (\omega_k/c)^2)\mathbf{u}_k(\mathbf{r}) = 0$$

the divergence condition $\nabla \cdot \mathbf{u}_k(\mathbf{r}) = 0$, the orthonormality conditions

$$\int \mathrm{d}^3 r \mathbf{u}_k(\mathbf{r}) \cdot \mathbf{u}_l(\mathbf{r}) = \delta_{kl}$$

and also the appropriate boundary conditions. In most cases of physical interest such a set does exist.

Given the set of mode functions $\mathbf{u}_k(\mathbf{r})$ one may expand the vector field $\mathbf{D}(\mathbf{r}, t)$ at each t into the series

$$\mathbf{D}(\mathbf{r}, t) = -\sum_k p_k(t)\mathbf{u}_k(\mathbf{r})$$

The minus sign is purely conventional; it makes the correspondence with the harmonic oscillator a closer one. The same procedure may be repeated for the vector **B** with a different complete set of functions $\mathbf{v}_k(\mathbf{r})$, resulting in the decomposition

$$\mathbf{B}(\mathbf{r},t) = \sum_k q_k(t)\mathbf{v}_k(\mathbf{r})$$

The Maxwell equations are satisfied if the coefficient functions $p_k(t)$ and $q_k(t)$ obey the equations of motion for harmonic oscillators

$$\frac{\mathrm{d}q_k(t)}{\mathrm{d}t} = \frac{p_k(t)}{\varepsilon}, \quad \frac{\mathrm{d}p_k(t)}{\mathrm{d}t} = -\varepsilon\omega^2 q_k(t) \qquad [1]$$

and the functions $\mathbf{v}_k(\mathbf{r})$ are related to $\mathbf{u}_k(\mathbf{r})$ by the formulas

$$\mathbf{v}_k(\mathbf{r}) = \nabla \times \mathbf{u}_k(\mathbf{r})$$

Note that the electric constant ε plays the same role as the mass for a mechanical oscillator. The calculation of the Hamiltonian, the total field energy,

$$H = \int \mathrm{d}^3r\left(\frac{\mathbf{D}^2}{2\varepsilon} + \frac{\mathbf{B}^2}{2\mu}\right)$$

completes the reduction of the electromagnetic field to a collection of harmonic oscillators. Upon inserting the representation of the field vectors in terms of the complete set of functions $\mathbf{u}_k(\mathbf{r})$ into the integral, one obtains the following formula

$$H = \sum_k \left(\frac{p_k^2}{2\varepsilon} + \frac{\varepsilon\omega_k^2 q_k^2}{2}\right)$$

which differs from its mechanical counterpart only by the replacement $m \to \varepsilon$. The time dependence has been omitted here because the Hamiltonian is a constant of motion.

Mode Functions

The simplest example of a complete set of mode functions is that obtained for a rectangular metallic cavity. In this case the functions $\mathbf{u}_k(\mathbf{r})$ may be expressed in terms of trigonometric functions. In order to write them down, it is convenient to replace one label k by a collection of three numbers $\boldsymbol{\kappa} = \{\kappa_x, \kappa_y, \kappa_z\}$ and an additional index λ taking on the values 1 and 2. The mode functions $\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$ satisfying the proper boundary conditions have the following components

$$\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r}) = N\begin{pmatrix} a_x^{(\lambda)} \cos(\kappa_x x)\sin(\kappa_y y)\sin(\kappa_z z) \\ a_y^{(\lambda)} \sin(\kappa_x x)\cos(\kappa_y y)\sin(\kappa_z z) \\ a_z^{(\lambda)} \sin(\kappa_x x)\sin(\kappa_y y)\cos(\kappa_z z) \end{pmatrix}$$

where $N = \sqrt{8/(L_xL_yL_z)}$, the coefficients $\{\kappa_x, \kappa_y, \kappa_z\}$ have the form $\{\kappa_x, \kappa_y, \kappa_z\} = \{\pi n_x/L_x, \pi n_y/L_y, \pi n_z/L_z\}$ and all the n's are natural numbers. The components of both vectors $\mathbf{a}^{(\lambda)}$ are subjected to the conditions $\boldsymbol{\kappa}\cdot\mathbf{a}^{(\lambda)} = 0$ (vanishing divergence) and orthonormalization $\mathbf{a}^{(\lambda)}\cdot\mathbf{a}^{(\lambda')} = \delta_{\lambda\lambda'}$. For each set $\{\kappa_x, \kappa_y, \kappa_z\}$ one must choose two independent solutions of these conditions $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$. These two vectors define two states of linear polarization. The mode function $\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$ is a solution of the Helmholtz equation for the value of ω/c equal to the length of the vector $\boldsymbol{\kappa}$. The expansion of the EMF vectors into the mode functions in a rectangular cavity has the form

$$\mathbf{D}(\mathbf{r},t) = -\sum_{\boldsymbol{\kappa}\lambda} p_{\boldsymbol{\kappa}\lambda}(t)\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$$

$$\mathbf{B}(\mathbf{r},t) = \sum_{\boldsymbol{\kappa}\lambda} q_{\boldsymbol{\kappa}\lambda}(t)\nabla \times \mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$$

where the summation extends over all allowed values of $\boldsymbol{\kappa}$ and two values of λ. One may explicitly write down mode functions also for a cylindrical and for a spherical cavity. They involve Bessel functions and are obtained by solving the Helmholtz equation in cylindrical and spherical coordinates, respectively.

One often uses another set of mode functions that does not correspond to any physical boundary conditions but is very helpful in the theoretical analysis. It is obtained by imposing the so-called periodic boundary conditions that were first introduced by Born in the study of crystal lattices. The periodic boundary conditions require that the electromagnetic field is periodic in space; it does not change when the coordinates are displaced by the 'period of the lattice' in all three directions. Assuming, for simplicity, that this period is the same, equal to L, in all directions, one obtains the following conditions for the mode functions

$$\mathbf{u}_k(\mathbf{r}) = \mathbf{u}_k(\mathbf{r} + L\mathbf{i}_x) = \mathbf{u}_k(\mathbf{r} + L\mathbf{i}_y) = \mathbf{u}_k(\mathbf{r} + L\mathbf{i}_z)$$

where $\mathbf{i}_i$ are the unit vectors in the directions of the coordinate axes. The complete set of solutions of the Helmholtz equation satisfying the periodic boundary conditions is labeled by the wavevector $\mathbf{k} = \{2\pi n_x/L, 2\pi n_y/L, 2\pi n_z/L\}$, where all the n's are arbitrary integers, and by the polarization index λ. The value of ω/c in the Helmholtz equation is equal in this case to the length of the vector k. Each member of the set $\mathbf{u}_{\mathbf{k}\lambda}$ has the form

$$\mathbf{u}_{\mathbf{k}\lambda}(\mathbf{r}) = \frac{1}{\sqrt{V}}\mathbf{e}_{\mathbf{k}\lambda}\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r}) \qquad [2]$$

where $V = L^3$. The polarization vectors $\mathbf{e}_{\mathbf{k}\lambda}$ must be orthogonal to the wave vector $\mathbf{k}$. It is convenient to impose the unitary orthonormality conditions $\mathbf{e}_{\mathbf{k}\lambda}\cdot\mathbf{e}^*_{\mathbf{k}\lambda} = \delta_{\lambda\lambda'}$ and also to assume that $\mathbf{e}^*_{\mathbf{k}\lambda} = \mathbf{e}_{-\mathbf{k}\lambda}$. The expansion of the EMF vectors into the mode functions obeying the periodic boundary conditions has the form

$$\begin{aligned} \mathbf{D}(\mathbf{r},t) &= -\frac{1}{\sqrt{V}}\sum_{\mathbf{k}\lambda} P_{\mathbf{k}\lambda}(t)\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \\ \mathbf{B}(\mathbf{r},t) &= \frac{1}{\sqrt{V}}\sum_{\mathbf{k}\lambda} Q_{\mathbf{k}\lambda}(t)\mathrm{i}\mathbf{k}\times\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \end{aligned} \qquad [3]$$

where the sum extends again over all allowed values of the vectors $\mathbf{k}$ and over two values of λ. The expansion coefficients $P_{\mathbf{k}\lambda}$ and $Q_{\mathbf{k}\lambda}$ are now complex. In order to guarantee that the field vectors are real, they must obey the conditions

$$P^*_{\mathbf{k}\lambda} = P_{-\mathbf{k}\lambda},\ Q^*_{\mathbf{k}\lambda} = Q_{-\mathbf{k}\lambda}$$

These conditions are satisfied if $P_{\mathbf{k}\lambda}$ and $Q_{\mathbf{k}\lambda}$ are parametrized by the real (but otherwise arbitrary) variables $p_{\mathbf{k}\lambda}$ and $q_{\mathbf{k}\lambda}$

$$P_{\mathbf{k}\lambda} = \frac{p_{\mathbf{k}\lambda} + p_{-\mathbf{k}\lambda}}{2} + \frac{\varepsilon\omega_{\mathbf{k}}(q_{\mathbf{k}\lambda} - q_{-\mathbf{k}\lambda})}{2i}$$

$$Q_{\mathbf{k}\lambda} = \frac{q_{\mathbf{k}\lambda} + q_{-\mathbf{k}\lambda}}{2} + \frac{\mathrm{i}(p_{\mathbf{k}\lambda} - p_{-\mathbf{k}\lambda})}{2\varepsilon\omega_{\mathbf{k}}}$$

The energy of the EMF splits again into a sum of independent contributions from all the modes

$$H = \sum_{\mathbf{k}\lambda}\omega_{\mathbf{k}}\left(\frac{p_{\mathbf{k}\lambda}p_{\mathbf{k}\lambda}}{2\varepsilon\omega_{\mathbf{k}}} + \frac{\varepsilon\omega_{\mathbf{k}}q_{\mathbf{k}\lambda}q_{\mathbf{k}\lambda}}{2}\right)$$

An important property of the periodic mode functions [2] is that not only the Hamiltonian but also the total momentum of the field splits into a sum of independent contributions. The total momentum $\mathbf{P}$ of the EMF is

$$\mathbf{P} = \int \mathrm{d}^3 r\,\mathbf{D}\times\mathbf{B}$$

Upon substituting into this expression the expansions [3] of the field vectors, one obtains, with the use of the orthonormality conditions,

$$\mathbf{P} = \sum_{\mathbf{k}\lambda}\mathbf{k}\left(\frac{p_{\mathbf{k}\lambda}p_{\mathbf{k}\lambda}}{2\varepsilon\omega_{\mathbf{k}}} + \frac{\varepsilon\omega_{\mathbf{k}}q_{\mathbf{k}\lambda}q_{\mathbf{k}\lambda}}{2}\right)$$

Canonical Quantization

The canonical quantization in quantum mechanics consists in replacing the c-numbers representing the classical canonical variables q_k and p_k by the operators $\hat{q}_k$ and $\hat{p}_k$. These operators are assumed to obey the canonical commutation relations

$$[\hat{q}_k, \hat{p}_l] = \mathrm{i}\hbar\delta_{kl} \qquad [4]$$

The canonical quantization of the EMF may proceed along the same lines as in quantum mechanics. All the variables q_k and p_k appearing in the decomposition into normal modes are replaced by the operators. When this is done, the field vectors $\mathbf{B}$ and $\mathbf{D}$ become the field operators $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ and the expression for the energy of the field becomes the Hamiltonian operator $\hat{H}$

$$\hat{H} = \sum_k \frac{\hat{p}_k^2}{2\varepsilon} + \frac{\varepsilon\omega_k^2\hat{q}_k^2}{2}$$

The only difference between quantum mechanics and quantum field theory is that in the second case the number of canonical variables is infinite but this may cause complications only when one wants to give the quantum theory of the EMF a firm mathematical foundation. In physical applications one may think of the number of field oscillators as being very large but finite.

Time evolution of a quantum system is described by associating the time dependence either with the operators (the Heisenberg picture) or with the state vectors (the Schrödinger picture). In both pictures the time dependence is governed by the system Hamiltonian. In the Heisenberg picture the time dependence of the operators is given by the commutators with $\hat{H}$

$$\frac{\mathrm{d}\hat{q}_k(t)}{\mathrm{d}t} = \frac{1}{\mathrm{i}\hbar}[\hat{q}_k(t), \hat{H}] = \frac{\hat{p}_k(t)}{\varepsilon}$$

$$\frac{\mathrm{d}\hat{p}_k(t)}{\mathrm{d}t} = \frac{1}{\mathrm{i}\hbar}[\hat{p}_k(t), \hat{H}] = -\varepsilon\omega_k^2\hat{q}_k(t)$$

These equations have the same form as in the classical theory (cf. eqn [1]). Therefore, in the Heisenberg picture the field operators will also obey the same equations as in the classical theory – the Maxwell equations:

$$\partial_t\hat{\mathbf{B}}(\mathbf{r},t) = -\nabla\times\hat{\mathbf{D}}(\mathbf{r},t)/\varepsilon,\quad \nabla\cdot\hat{\mathbf{B}}(\mathbf{r},t) = 0$$

$$\partial_t\hat{\mathbf{D}}(\mathbf{r},t) = \nabla\times\hat{\mathbf{B}}(\mathbf{r},t)/\mu,\quad \nabla\cdot\hat{\mathbf{D}}(\mathbf{r},t) = 0$$

Of course, the average values of the field operators evaluated in both pictures are the same and for every quantum state they obey the classical Maxwell equations.

The commutation relations [4] for the canonical operators and the completeness of the mode functions imply the following equal-time commutation relations for the field operators

$$[\hat{B}_i(\mathbf{r},t),\hat{D}_j(\mathbf{r}',t)] = \mathrm{i}\hbar\sum_k \varepsilon_{ijk}\partial_k\delta^{(3)}(\mathbf{r}-\mathbf{r}')$$

These relations do not depend on the choice of the mode functions. They summarize in a universal manner the canonical properties of the electromagnetic field operators.

Annihilation and Creation Operators

The canonical variables q_k and p_k are convenient for bringing home the analogy between mechanical oscillators and wave fields. However, there exists also a complementary, corpuscular aspect of the quantum theory of the EMF and that is most easily expressed in terms of photons. The best way to describe photons in quantum field theory is in terms of annihilation and creation operators. These operators were introduced by Schrödinger in the context of the quantum-mechanical harmonic oscillator but they have become really indispensable in quantum field theory. Formally, the annihilation $\hat{a}_k$ and the creation $\hat{a}_k^\dagger$ operators are defined as the following complex combinations of canonical operators

$$\begin{aligned}\hat{a}_k &= \sqrt{\frac{\varepsilon\omega}{2\hbar}}\left(\hat{q}_k + \mathrm{i}\frac{\hat{p}_k}{\varepsilon\omega}\right)\\ \hat{a}_k^\dagger &= \sqrt{\frac{\varepsilon\omega}{2\hbar}}\left(\hat{q}_k - \mathrm{i}\frac{\hat{p}_k}{\varepsilon\omega}\right)\end{aligned} \qquad [5]$$

Annihilation and creation operators are dimensionless, one is the Hermitian adjoint of the other, and they obey very simple commutation relations:

$$[\hat{a}_k,\hat{a}_l^\dagger] = \delta_{kl} \qquad [6]$$

The Hamiltonian expressed in terms of annihilation and creation operators reads

$$\hat{H} = \sum_k \hbar\omega_k\hat{a}_k^\dagger\hat{a}_k + \frac{1}{2}\sum_k\hbar\omega_k \qquad [7]$$

This Hamiltonian determines the following time dependence of the annihilation and creation operators

$$\hat{a}_k(t) = \hat{a}_k\exp(-\mathrm{i}\omega_k t),\ \hat{a}_k^\dagger(t) = \hat{a}_k^\dagger\exp(\mathrm{i}\omega_k t)$$

The field operators expressed in terms of annihilation and creation operators corresponding to the periodic boundary conditions are

$$\hat{\mathbf{D}}(\mathbf{r},t) = \mathrm{i}\sum_{\mathbf{k}\lambda}\sqrt{\frac{\hbar\varepsilon\omega_{\mathbf{k}}}{2V}}\hat{a}_{\mathbf{k}\lambda}\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + \mathrm{H.c.}$$

$$\hat{\mathbf{B}}(\mathbf{r},t) = \mathrm{i}\sum_{\mathbf{k}\lambda}\sqrt{\frac{\hbar}{2V\varepsilon\omega_{\mathbf{k}}}}\hat{a}_{\mathbf{k}\lambda}\mathbf{k}\times\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + \mathrm{H.c.}$$

where H.c. stands for the Hermitian conjugate of the first term. In the limiting case for infinite empty space, when $V \to \infty$, the Fourier sum becomes the Fourier integral:

$$\begin{aligned}\hat{\mathbf{D}}(\mathbf{r},t) = \mathrm{i}\sum_\lambda\int \mathrm{d}^3k\sqrt{\frac{\hbar\varepsilon\omega_{\mathbf{k}}}{2(2\pi)^3}}&\Big(\hat{a}_\lambda(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\\ &- \hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_{\mathbf{k}}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\Big)\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{B}}(\mathbf{r},t) = \mathrm{i}\sum_\lambda\int \mathrm{d}^3k\sqrt{\frac{\hbar}{2\varepsilon\omega_{\mathbf{k}}(2\pi)^3}}&\Big(\hat{a}_\lambda(\mathbf{k})\mathbf{k}\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\\ &- \hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{k}\times\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_{\mathbf{k}}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\Big)\end{aligned}$$

In this case, the commutation relations involve the (three-dimensional) Dirac delta function

$$[\hat{a}_\lambda(\mathbf{k}),\hat{a}_{\lambda'}^\dagger(\mathbf{k}')] = \delta_{\lambda\lambda'}\delta^{(3)}(\mathbf{k}-\mathbf{k}') \qquad [8]$$

The field operators may also be expressed in terms of the vector potential operator $\hat{\mathbf{A}}(\mathbf{r},t)$

$$\hat{\mathbf{D}}(\mathbf{r},t) = -\varepsilon\partial_t\hat{\mathbf{A}}(\mathbf{r},t),\ \hat{\mathbf{B}}(\mathbf{r},t) = \nabla\times\hat{\mathbf{A}}(\mathbf{r},t)$$

$$\begin{aligned}\hat{\mathbf{A}}(\mathbf{r},t) = \sum_\lambda\int \mathrm{d}^3k\sqrt{\frac{\hbar}{2\varepsilon\omega_{\mathbf{k}}(2\pi)^3}}&\Big(\hat{a}_\lambda(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\\ &+ \hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_{\mathbf{k}}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\Big)\end{aligned}$$

The vector potential is needed for the description of the interaction with charged particles.

Space of Quantum States

The standard basis in the space of quantum states consists of the state vectors of stationary states – the eigenvectors of the Hamiltonian.

Vacuum State

The lowest energy state of the EMF – the ground state of the system – is often called the vacuum state. The state vector describing the vacuum state is denoted by $|0\rangle$. The vacuum state is the common ground state of all harmonic oscillators making up the EMF. Since the ground state of a single harmonic oscillator has the energy $\hbar\omega/2$, the energy of the ground state for an infinite number of oscillators is infinite (last term in eqn [7]). This infinity has no physical consequences because only the energy differences are observable. In the presence of obstacles (mirrors, resonators, beamsplitters, etc.), owing to the change in the mode structure, there will be a change in the energy of the new ground state relative to the energy in empty space. The energy differences between various ground states do have physical significance.

Fock Basis

It follows from the commutation relations [6] and from the form [7] of the Hamiltonian that the operator $\hat{a}_k$ decreases the energy of the EMF by the Planck quantum of energy $\hbar\omega_k$. Therefore, all operators $\hat{a}_k$ must annihilate the ground state vector, $\hat{a}_k|0\rangle = 0$, because there is no state with a lower energy. Each creation operator $\hat{a}_k^\dagger$ acting on any state vector adds one excitation with the energy $\hbar\omega_k$. These excitations of the electromagnetic field with characteristic energies $\hbar\omega_k$ are interpreted as photons. The state vectors of the form $\hat{a}_k^\dagger|0\rangle$ describe one-photon states, the state vectors of the form $\hat{a}_k^\dagger\hat{a}_l^\dagger|0\rangle$ describe two-photon states, etc. All the vectors generated from the vacuum state vector by acting with the creation operators form a basis, called the Fock basis. The general state vector of this basis is characterized by a sequence of natural numbers $\{n_1, n_2, \ldots\}$ and is generated from the vacuum state according to the formula

$$|n_1, n_2, \ldots\rangle = (n_1!n_2!\cdots)^{-1/2}(\hat{a}_1^\dagger)^{n_1}(\hat{a}_2^\dagger)^{n_2}\ldots|0\rangle \qquad [9]$$

where the prefactor is needed for normalization. The number n_k tells us how many photons of type k are present in this state. The spatial structure of a field excitation created by $\hat{a}_k^\dagger$ is described by the mode function $\mathbf{u}_k(\mathbf{r})$. Thus, this mode function plays the same role as the wavefunction $\psi(\mathbf{r})$ of a massive particle and the set of indices labeling the mode functions is the analog of quantum numbers for $\psi(\mathbf{r})$. In particular, for the mode function [2], the photons have a definite momentum $\hbar\mathbf{k}$. In a cylindrical waveguide the photons may have a definite momentum only along the axis of the cylinder. In a spherical cavity the photons may have the definite value of square of the total angular momentum and its projection on a given axis. The Fock states are highly nonclassical; all average values of the EMF in these states vanish.

All states of the EMF are either pure or mixed. Pure states are described by all linear superpositions of the basis states. In particular, by superposing photon states with a range of energies, one may form nonmonochromatic photon wavepackets. All superpositions are allowed since photons do not have any absolutely conserved quantum numbers (like, e.g., charge). Mixed states are described by the density operators. Among the mixed states the prominent role is played by the thermal state describing the EMF in equilibrium with a reservoir at temperature T.

Coherent States

The one-mode coherent states $|\alpha\rangle$ are obtained from the vacuum state by the action of the unitary displacement operator $\hat{D}(\alpha, \alpha^*)$

$$|\alpha\rangle = D(\alpha, \alpha^*)|0\rangle = \mathrm{e}^{\alpha\hat{a}^\dagger - \alpha^*\hat{a}}|0\rangle = \mathrm{e}^{-|\alpha|^2/2}\mathrm{e}^{\alpha\hat{a}^\dagger}|0\rangle$$

where $\hat{a}$ and $\hat{a}^\dagger$ are the annihilation and creation operators corresponding to a given mode. Coherent states are superpositions of an infinite number of Fock states. The multimode coherent state is generated by the product of displacement operators

$$|\alpha_1, \alpha_2, \ldots\rangle = \mathrm{e}^{\sum_k (\alpha_k\hat{a}_k^\dagger - \alpha_k^*\hat{a}_k)}|0\rangle$$

In reality, the coherent states are generated by the action of an external (strong) time-dependent electric current. In that case the coefficients α_k are the projections of the current on the mode functions. The average values of the field operators $\langle\hat{B}_i(\mathbf{r}, t)\rangle$ and $\langle\hat{D}_j(\mathbf{r}, t)\rangle$ in coherent states are the solutions of the Maxwell equations in the presence of an external electric current.

Entangled States

The states of the EMF describing photons in different modes may exhibit correlations. The simplest such state is a two-photon state of the form $|\phi\rangle = 2^{-1/2}(\hat{a}_1^\dagger\hat{a}_2^\dagger + \hat{a}_3^\dagger\hat{a}_4^\dagger)|0\rangle$. In this case, when one photon is found, say, in the state 1, the other will be found with the probability 1 in the state 2. The states that exhibit correlations between photons belonging to different modes are called the entangled states.

Correlations occur also in the classical theory, but only for statistical mixtures when there is a loss of information. In quantum theory the correlations may arise even without a loss of information, i.e., for pure states. A striking feature of entangled states is that the decomposition of such a state into a sum of Fock states is not unique. For example, the state vector $|\phi\rangle$ may also be written in the form $|\phi\rangle = 2^{-1/2}(\hat{a}^{\dagger}_{1'}\hat{a}^{\dagger}_{2'} + \hat{a}^{\dagger}_{3'}\hat{a}^{\dagger}_{4'})|0\rangle$, where $\hat{a}^{\dagger}_{1'} = 2^{-1/2}(\hat{a}^{\dagger}_{1} - \hat{a}^{\dagger}_{3})$, $\hat{a}^{\dagger}_{2'} = 2^{-1/2}(\hat{a}^{\dagger}_{2} - \hat{a}^{\dagger}_{4})$, $\hat{a}^{\dagger}_{3'} = 2^{-1/2}(\hat{a}^{\dagger}_{1} + \hat{a}^{\dagger}_{3})$, $\hat{a}^{\dagger}_{4'} = 2^{-1/2}(\hat{a}^{\dagger}_{2} + \hat{a}^{\dagger}_{4})$. Entangled states play a crucial role in quantum information theory, especially in quantum cryptography.

Quantum Fluctuations and the Uncertainty Relations

It follows from the basic principles of quantum theory that there are no states of the system for which both canonical variables q and p have sharply defined values. The limits on the quantum fluctuations (variances) of these values are determined by the uncertainty relation. The most stringent form of this relation obtained from the commutation relations [4] (for a single mode) reads

$$(\Delta q)^2(\Delta p)^2 \geq \frac{\hbar^2}{4} + \frac{\langle \delta\hat{q}\delta\hat{p} + \delta\hat{p}\delta\hat{q}\rangle^2}{4} \qquad [10]$$

where $\delta\hat{q} = \hat{q} - \langle\hat{q}\rangle$, $\delta\hat{p} = \hat{p} - \langle\hat{p}\rangle$, $(\Delta q)^2 = \langle(\delta\hat{q})^2\rangle$, and $(\Delta p)^2 = \langle(\delta\hat{p})^2\rangle$. The states for which this inequality is saturated play a distinguished role in quantum theory. Since the uncertainty relation expresses the limitations on the quantum fluctuations of q and p, the best one can do in approaching the classical regime is to achieve equality in the uncertainty relation. That is why the quantum states that saturate [10] are viewed as the counterparts of the classical states. In the quantum theory of the EMF, the uncertainty relations express the complementarity between the magnetic and electric fields. The better one knows the magnetic field, the more uncertain is the electric field, and vice versa. However, the projections of the field operators $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ on the same direction commute (cf. eqn [5]). Therefore, there are no limitations on the magnetic and electric field values in a given direction and only the values in different directions are subjected to the quantum uncertainty relation.

In the vacuum state the variances of q and p for each mode are

$$(\Delta q)^2 = \frac{\hbar}{2\varepsilon\omega}, (\Delta p)^2 = \frac{\hbar\varepsilon\omega}{2} \qquad [11]$$

and they saturate the uncertainty relation. These fluctuations are called the shot noise. Very often one uses dimensionless quadrature operators $\hat{X}_1 = \hat{q}\sqrt{2\varepsilon\omega/\hbar}$ and $\hat{X}_2 = \hat{p}\sqrt{2/\varepsilon\omega\hbar}$, whose fluctuations in the vacuum state are equal to 1. All coherent states have the same fluctuations as the vacuum state. They may be viewed in the phase space as the vacuum state displaced from the origin (corresponding to the vacuum state) by the mean values of q and p.

Coherent states are not the only states that saturate the uncertainty relations. The other states with this property are called the squeezed states. The fluctuations of X_1 and X_2 for squeezed states are different from 1 but their product is still equal to 1, $(\Delta X_1)^2(\Delta X_2)^2 = 1$.

Fock states have well-defined energies and numbers of photons since they are the eigenstates of the Hamiltonian. However, the phases of the EMF are maximally uncertain; they are evenly spread over the whole range from 0 to 2π. That is why the average values of the field vectors vanish in all Fock states.

See also

Photon Picture of Light. **Quantum Electrodynamics:** Cavity QED.

Further Reading

Bialynicki-Birula I (1996) Photon wave function. In: Wolf E (ed.) *Progress in Optics*, vol. 36. Amsterdam, The Netherlands: Elsevier.

Bialynicki-Birula I and Bialynicka-Birula Z (1975) *Quantum Electrodynamics*. London: Pergamon.

Loudon R (1973) *The Quantum Theory of Light*. Oxford, UK: Clarendon Press.

Louisell W (1990) *Quantum Statistical Properties of Radiation*. New York: Wiley.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Power EA (1964) *Introductory Quantum Electrodynamics*. London: Longman.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Cavity QED

H Walther, University of Munich and
Max-Planck-Institute for Quantum Optics, Garching,
Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Modification of the Spontaneous Transition Rate in Confined Space

In order to understand the modification of the spontaneous emission rate in an external cavity-like structure, it must be remembered that in quantum electrodynamics this rate is proportional to the density of modes of the electromagnetic field, i.e., the vacuum field fluctuations above the atomic transition frequency ω_0. As a consequence the spontaneous emission rate is increased if the atom is surrounded by a cavity tuned to the transition frequency. Conversely, the decay rate decreases when the cavity is mistuned. This fact was already recognized in the pioneering days of nuclear magnetic resonance by Purcell.

The spontaneous decay rate of the atom in the cavity, γ_c, will then be enhanced in relation to that in free space, γ_f, by a factor given by the ratio of the corresponding mode densities

$$\frac{\gamma_c}{\gamma_f} = \frac{\rho_c(\omega_0)}{\rho_f(\omega_0)} = \frac{2\pi Q}{V_c \omega_0^3} = \frac{Q\lambda_0^3}{4\pi^2 V_c}$$

where V_c is the volume of the cavity and Q is the quality factor of the cavity which expresses the 'sharpness' of the mode. For low-order cavities $V_c \approx \lambda_0^3$ this means that the spontaneous emission rate is increased by roughly a factor of Q. As mentioned already, when the cavity is detuned, the decay rate will decrease. In this case the atom cannot emit a photon, since the cavity is not able to accept it, and the energy will stay with the atom, i.e., its decay is inhibited.

To change the decay rate of an atom, in principle no resonator has to be present; any conducting surface near the radiator affects the mode density and, therefore, the spontaneous radiation rate. Parallel conducting planes can alter the emission rate somewhat but can only reduce the rate by a factor of 2 owing to the existence of modes, independent of the plate separation with an electric field perpendicular to the planes.

In order to demonstrate experimentally the modification of the spontaneous decay rate, it is not necessary to go to single-atom densities. The experiments where the spontaneous emission is inhibited can also be performed with large atom numbers. However, in the opposite case, when the increase of the spontaneous rate is observed, a large number of excited atoms may disturb the experiment by induced transitions. The first experimental work on inhibited spontaneous emission was done by Drexhage, Kuhn and Schäfer in 1974. The fluorescence of a thin dye film near a mirror was investigated. A reduction of the fluorescence decay by up to 25% was observed. Later experiments with microcavities filled with dye solutions were performed by de Martini *et al.* These experiments demonstrated that thresholdless lasing is possible in such systems.

Inhibited spontaneous emission was observed by Gabrielse and Dehmelt. In these neat experiments with a single electron stored in a Penning trap they observed that cyclotron orbits show lifetimes which are up to 10 times longer than that calculated for free space. The electrodes of the trap form a cavity which decouples the cyclotron motion from the vacuum field leading to the longer lifetime.

A new stage in experiments on cavity QED was reached when it was recognized that highly excited alkaline atoms are very suitable for these experiments. The main quantum numbers are in the range $n = 20-60$. Those so-called Rydberg atoms couple very strongly to the radiation field as will be discussed later. The transitions to neighboring states are in the microwave region. The cavities can therefore be built as low-order cavities with dimensions on the order of the wavelength of the transitions being in the mm or cm regions. The spontaneous rate of these transitions in free space is small owing to the low value of their transition frequency, therefore an enhancement is possible. Experiments with Rydberg atoms on the inhibition of spontaneous emission have been performed by Kleppner and coworkers and by Haroche and coworkers. In the latter experiment a 3.4 μm transition of the Cs atom was suppressed.

The first observation of enhanced atomic spontaneous emission in a resonant cavity was published by Haroche *et al.* This experiment was performed with Rydberg atoms of Na excited in the 23s state in a niobium superconducting cavity resonant at 340 GHz. Cavity tuning-dependent shortening of the lifetime was observed. The cooling of the cavity had the advantage of totally suppressing the black-body field. The latter effect is completely absent if optical transitions are observed, however, in this case it is more difficult to obtain low-order cavities. The first experiments on optical transitions were performed by Feld and collaborators. They succeeded in demonstrating the enhancement of spontaneous transitions even in higher-order optical cavities.

In modern semiconductor devices both electronic and optical properties can be tailored with a high degree of precision. Therefore, electron–hole systems producing recombination radiation analogous to radiating atoms can be localized in cavity-like structures, e.g., in quantum wells. Thus optical microcavities of half or full wavelength size are obtained. Both suppression and enhancement of spontaneous emission in semiconductor microcavities were demonstrated in experiments by Yamamoto and collaborators. Yablonovitch *et al.* proposed the use of photonic band gap structures in semiconductors. In those systems spontaneous emission is forbidden in certain frequency regions owing to the special material properties.

Energy Shift in Confined Space

In the previous section we focused on changes of radiation rates of atoms near conducting walls or in cavity-like structures. Next we will discuss the more subtle phenomenon of energy shifts. While radiation rates are modified by the influence of the vacuum field on the real emission of photons the energy shifts are caused by the modification of a virtual photon interaction.

Resonant and nonresonant phenomena have to be distinguished. The resonant self-energy shift of a decaying atomic dipole in the vicinity of a conducting wall can be determined from the average polarization energy produced by the image dipole field. For distances comparable to the wavelength the near-field condition is satisfied, resulting in the z^{-3} dependence of the static dipole–dipole interaction characteristic for the van der Waals energy; under far-field conditions the distance dependence is given by z^{-1}. The polarization of the atom by the nonresonant parts of the broadband electromagnetic field causes energy shifts, of which the Lamb shift is the most prominent. In the sense of the nonrelativistic treatment by Bethe the major contribution of that shift can be described as being a result of the emission and reabsorption of virtual photons. It is plausible that just as the real emission of a photon is modified in confined space, so also is the virtual process. The latter ‘real’ radiation energy shift is thus a consequence of vacuum fluctuations only. It is identical with the energy shift predicted by Casimir and Polder and is analogous to the better-known result of Casimir on the force between two plane neutral conducting plates.

The question of modification of atomic energies in confined space has recently found considerable interest and many calculations of the phenomenon have been performed. Direct application to the energy shift of Rydberg atoms, which are of special interest for experimental studies, was performed by Barton in 1987. He calculated the direct electrostatic interaction with a conducting wall and the radiation induced (retarded) effects. The result is that in the case of two parallel plates the electrostatic effect is dominant when the distance L between the conducting plates is small, $L < n^3 a_0/\alpha$ (n is the principal quantum number of the Rydberg atom, a_0 the Bohr radius, and α the fine-structure constant), and the radiative effect plays the major role when large distances are used, $L > n^3 a_0 \alpha$.

Experiments to measure the van der Waals force of conducting planes have been performed by Hinds and coworkers. They could clearly demonstrate the z^{-3} dependence of the van der Waals force. The retarded effects were not detectable at large distances but at small distances from the wall a deviation from the z^{-3} dependence was found which was attributed to the retarded QED potential.

The energy shift of rubidium Rydberg atoms in confined space has been measured by Marrocco *et al.* in an experiment with extreme high-resolution using the Ramsey experiment two-field method. The atoms are excited in s-Rydberg states ($n \approx 30$) by two-photon transitions using the light of an ultrastable dye laser, the linewidth of which was less than 10 Hz. The laser intensity was enhanced in a folded cavity locked to the laser frequency. Both interaction zones necessary for the Ramsey method were enclosed in this cavity. Between the two interaction regions the atoms pass through a pair of conducting plates, the distance between which can be changed. The shift of the Ramsey interference was measured as a function of the plate distance. Using the Ramsey method has the advantage that the shift can be determined without a direct probing of the atoms in the space between the plates. A level shift on the order of about 150 Hz could be measured, in very good agreement with theory.

Maser Operation

If the rate of atoms crossing a cavity exceeds the cavity damping rate ω/Q, the photon released by each atom is stored long enough to interact with the next atom. Here ω stands for the cavity frequency and Q for the quality factor of the cavity. The atom–field coupling becomes stronger and stronger as the field builds up and evolves to a steady state. Using Rydberg atoms with a large field–atom coupling constant leads to a new kind of maser which operates with exceedingly small numbers of atoms and photons.

The photons corresponding to transitions between neighboring Rydberg atoms are in the microwave region at about 20–100 GHz. Atomic fluxes as small as a few atoms per second have generated maser action, as could be demonstrated by Walther *et al.* of the University of Munich in 1985 using a superconducting cavity. For such a low flux there is never more than a single atom in the resonator – in fact, most of the time the cavity is empty of atoms. It should also be mentioned that in the case of such a setup a single resonant photon is sufficient to saturate the maser transition.

The scheme for this one-atom maser or micromaser is shown in Figure 1. The setup represents the simplest system in radiation–atom interaction: a single atom is interacting with a single mode of the radiation field. This device was used to verify the complex dynamics of a single atom in a quantized field predicted by the Jaynes–Cummings model. All of the features are explicitly a consequence of the quantum nature of the electromagnetic field: the statistical and discrete nature of the photon field leads to new characteristic dynamics such as collapse and revivals in the exchange of a photon between the atom and the cavity mode. The frequency of this exchange is usually called the Rabi flopping frequency. The field in the cavity is measured through the number of atoms in the lower state of the maser transition. Due to the strong coupling between the atom and maser field both are entangled. The strong coupling can also be used to entangle subsequent atoms.

The steady-state field of the micromaser shows sub-Poisson statistics. This is in contrast to regular masers and lasers where coherent fields (Poisson statistics) are observed. The reason for nonclassical radiation being produced is that a fixed interaction time of the atoms is chosen, leading to careful control of the atom–field interaction dynamics.

Under steady-state conditions, the photon statistics $P(n)$ of the field of the micromaser are essentially determined by the pump parameter $\Theta = N_{\mathrm{ex}}^{1/2}\Omega t_{\mathrm{int}}$, denoting the angular rotation of the pseudospin vector of the interacting atom. Here, N_{ex} is the average number of atoms that enter the cavity during the decay time of the cavity τ_{cav}, Ω the vacuum Rabi flopping frequency, and t_{int} is the atom–cavity interaction time. The normalized photon number of the maser $\langle \nu \rangle = \langle n \rangle / N_{\mathrm{ex}}$ shows the following generic behavior

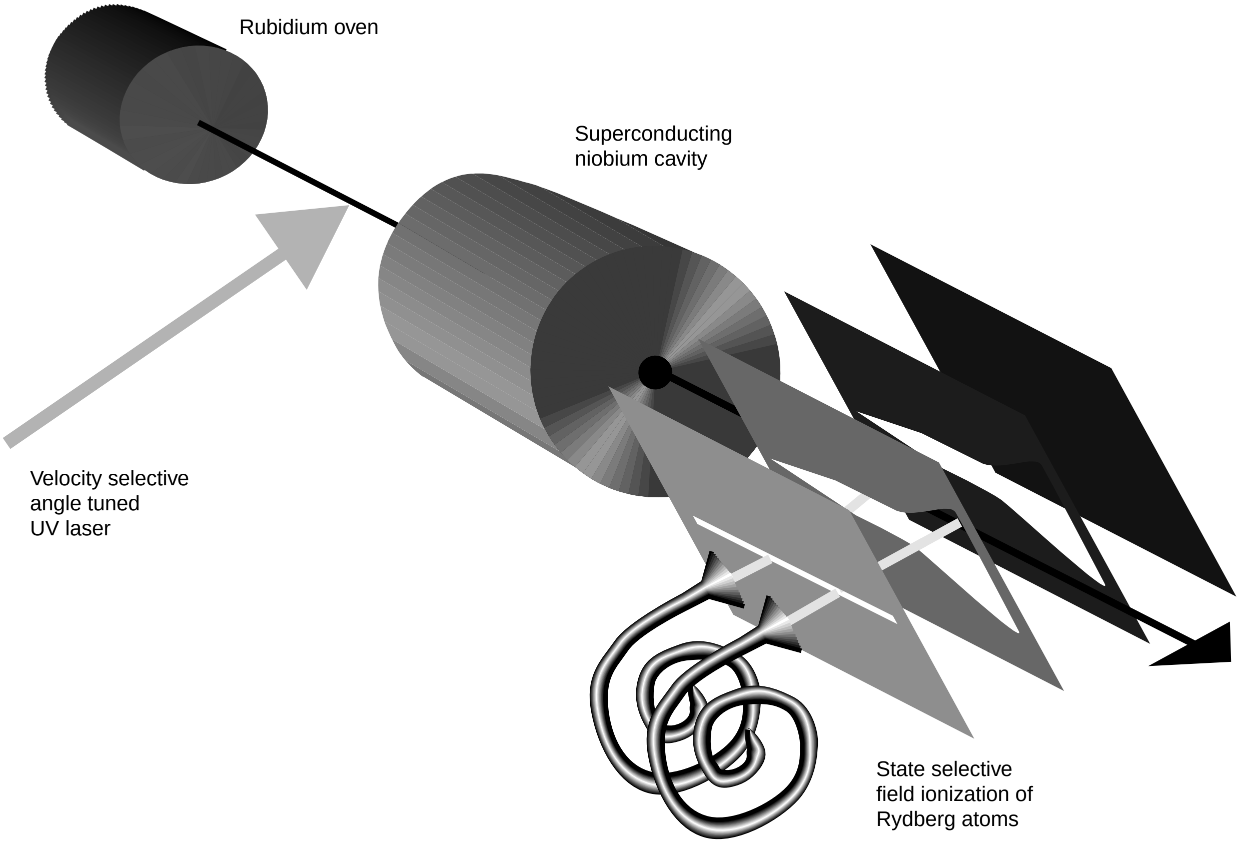
Rubidium oven
Superconducting niobium cavity
Velocity selective angle tuned UV laser
State selective field ionization of Rydberg atoms

Figure 1 Micromaser setup of rubidium Rydberg atoms in the 63p state. The velocity of the atoms is controlled by exciting a velocity subgroup of atoms in the atomic beam. The atoms in the upper and lower maser levels are selectively detected by field ionization. The number of photons deposited in the cavity is determined through the number of atoms in the lower state (61d).

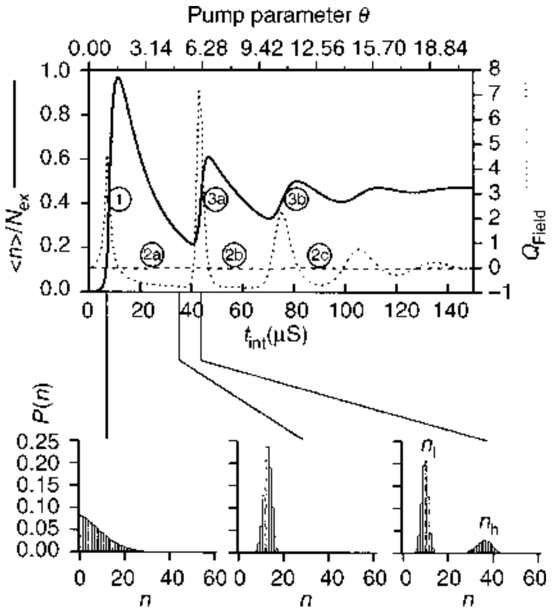

Figure 2 One-atom maser or micromaser characteristics. The upper part of the figure shows the average photon number versus interaction time (solid curve) and the photon number fluctuations represented by the *Q*-factor (dotted line). Both curves are determined by the photon exchange dynamics between atom and field. The pump parameter gives the angular rotation of the pseudospin vector of the interacting atom (e.g., 2π corresponds to a full rotation, i.e., the atom is again in the upper level). In the lower part of the figure the steady-state photon number distribution $P(n)$ is shown for three values of t_{int}. The distribution on the left side corresponds to the maser threshold, that on the right gives an example of the double-peaked distribution associated with the quantum jump behavior. In this situation, the atom is back in the excited state and can again emit, leading to a higher steady-state photon number n_{h}. With increasing t_{int} this part will grow and n_{l} will decrease and disappear. A new jump occurs in the region 3b.

(see Figure 2). It suddenly increases at the maser threshold value $\Theta = 1$ and reaches a maximum for $\Theta \approx 2$ (denoted by 1 in Figure 2). The behavior at threshold corresponds to the characteristics of a continuous phase transition. As Θ increases further, the normalized averaged photon number $\langle \nu \rangle / N_{\text{ex}}$ decreases to a minimum slightly below $\Theta \approx 2\pi$, and then abruptly increases to a second maximum (3a in Figure 2). This general type of behavior recurs roughly at integer multiples of 2π, but becomes less pronounced with increasing Θ. The reason for the periodic maxima of the average photon number is that for integer multiples of $\Theta = 2\pi$ the pump atoms perform an integer number of full Rabi flopping cycles, and start to flip over at a slightly larger value of Θ, thus leading to enhanced photon emission. The periodic maxima for $\Theta = 2\pi, 4\pi$, and so on may be interpreted as first-order phase transitions.

The photon statistics of the maser radiation is usually characterized by the Q-parameter introduced by Mandel:

$$Q_{\text{field}} = \left\{ \left(\langle n^2 \rangle - \langle n \rangle^2 \right) / \langle n \rangle \right\} - 1$$

It can be seen that $Q_{\text{field}} = 0$ corresponds to a Poissonian photon distribution. Q_{field} for the micromaser is plotted as a dotted line in Figure 2. The value drops below zero in the region 2a, 2b, etc. This shows the highly sub-Poissonian character of the one-atom-maser field being present over a large range of parameters.

The reason for the sub-Poissonian atomic statistics is the following. A changing flux of atoms changes the Rabi frequency via the stored photon number in the cavity. Adjusting the interaction time allows the phase of the Rabi nutation cycle to be chosen such that the probability of the atoms leaving the cavity in the upper maser level increases when the flux, and hence the photon number in the cavity, is raised. This leads to sub-Poissonian atomic statistics since the number of atoms in the lower state decreases with increasing flux and photon number in the cavity. This feedback mechanism can be neatly demonstrated when the anticorrelation of atoms leaving the cavity in the lower state is investigated. Measurements of this anticorrelation phenomenon could be made.

The fact that anticorrelation is observed shows that the atoms in the lower state are more equally spaced than expected for a Poissonian distribution of the atoms in the beam. This means, for example, that when two atoms enter the cavity close to each other, the second one performs a transition to the lower state with reduced probability.

The interaction with the cavity field thus leads to an atomic beam with atoms in the lower maser level showing number fluctuations which are up to 40% below those of a Poissonian distribution found in usual atomic beams. This is interesting because atoms in the lower level have emitted a photon to compensate for cavity losses inevitably present. Although this process is induced by dissipation giving rise to fluctuations, the atoms still obey sub-Poissonian statistics.

Generation of Number States (Fock States) of the Radiation Field

When the micromaser is operated at low temperatures, a very interesting feature is observed in the inversion of the population of the two maser levels called trapping states. They are a steady-state feature of the maser field; they occur in the micromaser as a direct consequence of field quantization in a cavity. At low cavity temperatures the number of black-body photons, n_{th}, in the cavity mode is reduced and trapping states begin to appear; at higher temperatures they are washed out by the thermal photons. The trapping states show up when the atom–field coupling Ω and the interaction time t_{int} are chosen such that in a cavity field with n_{q} photons the atoms undergo an integer number k of Rabi cycles. This is summarized by the condition

$$\Omega t_{\mathrm{int}}\sqrt{n_{\mathrm{q}}+1}=k\pi \qquad [1]$$

for the trapping state, denoted by (n_{q}, k). When eqn [1] is satisfied, the cavity photon number is left unchanged after interaction of an atom and hence the photon number is 'trapped'. This will occur regardless of the atomic pump rate N_{ex}. The trapping state is therefore characterized by the upper-bound photon number n_{q} and the number of integer multiples of full Rabi cycles k. In this situation the field is stabilized. Whenever a photon disappears, e.g., due to dissipation, the next incoming atom experiences a changed Rabi nutation frequency and emits a photon with high probability. At the trapping condition a quantum nondemolition situation is present. Through the dynamics of the Rabi nutation the field is measured, without any net change of the field.

The build-up of the cavity field can be seen in Figure 3, where the emerging atom inversion is plotted against the interaction time and pump rate. At low atomic pump rates (low N_{ex}) the maser field cannot build up and the maser exhibits Rabi oscillations due to the interaction with the vacuum field. At the positions of the trapping states, the field increases until it reaches the trapping-state condition. This is manifested as a reduced emission probability and hence as a dip in the atomic inversion. Once in a trapping state, the maser will remain there regardless of the pump rate.

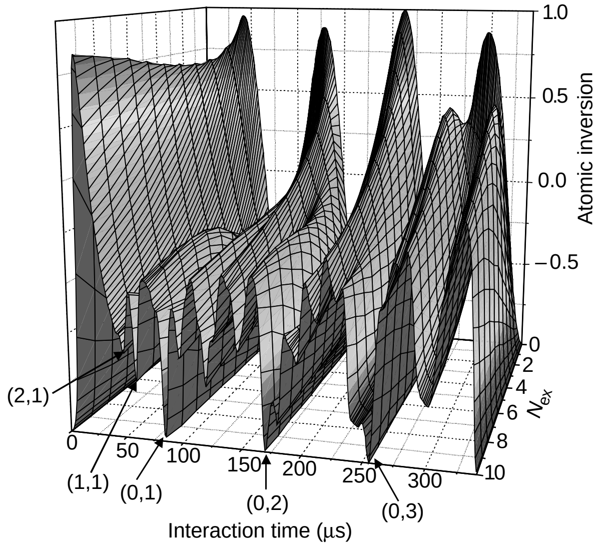

Figure 3 Behavior of the micromaser at low temperature showing the trapping states, manifested as valleys along the N_{ex} axis. For the designation of the trapping states see text.

Under ideal conditions the micromaser field in a trapping state is a Fock state, but when the micromaser is operated in a continuous wave mode, the field state is very fragile and highly sensitive to external influences and experimental parameters. However, in contrast to continuous-wave operation, in pulsed operation trapping states are more stable and more practical and can be used over a much broader parameter range than in continuous-wave operation.

To demonstrate the principle, Figure 4 shows a simulation of a sequence of 20 pulses of the pumping atoms in which an average of seven excited atoms per pulse are present. Under the trapping condition only a single emission event occurs, producing a single lower state atom which leaves a single photon in the cavity. Since the atom–cavity system is in the trapping condition, the emission probability is reduced to zero and the photon number is stabilized. Consequently, excited state atoms following the emitting atom stay in the upper maser level. The variation of the time when an emission event occurs during the atom pulses in Figure 4 is due to the Poissonian spacing of upper-state atoms entering the cavity and the stochasticity of the quantum process.

The lower part of Figure 4 shows the photon number distribution resulting from this process. It was demonstrated experimentally by Brattke *et al.* that a single photon number state could be generated with a success rate of 85%. By improving the experimental parameters one can expect to prepare single-photon Fock states in 98% of the pulses by this method.

Other Cavity Experiments

Besides the micromaser, there are also other experiments using Rydberg atoms in cavities and the strong interaction of these atoms with cavities which lead to interesting applications. One example is an experiment by S. Haroche, J. M. Raimond *et al.*, who succeeded in realizing a Schrödinger cat state in a cavity. Studies on the decoherence of this state could be conducted. The experiments are quite important in exploring the boundaries between the quantum and classical worlds. Another example is the quantum nondemolition detection of a single photon in a cavity by S. Haroche employing the dispersion level shift of probing atoms.

There is an interesting equivalence between an atom interacting with a single-mode field and a quantum

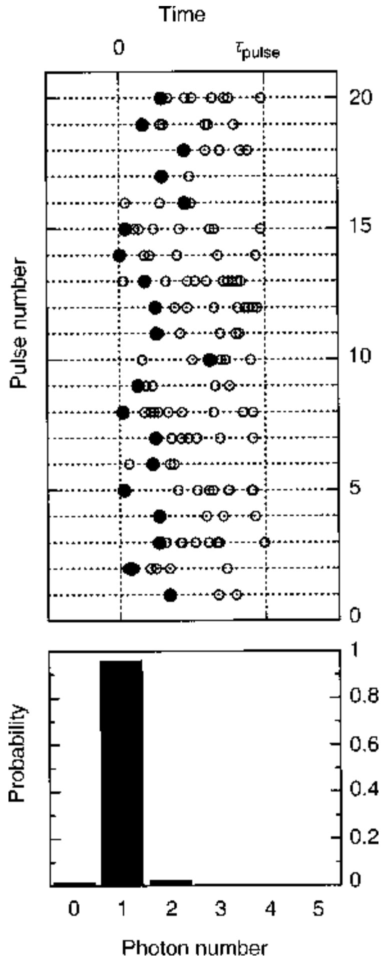

Figure 4 A simulation of a subset of 20 subsequent atom bunches generated by pulsed laser excitation with the associated probability distribution for photons in the cavity or lower-state atom production (filled circles represent lower-state atoms and open circles represent excited-state atoms). The start and end of each pulse is indicated by the vertical dotted lines marked 0 and τ_{pulse}, respectively. For details, see Battke S, Varcoe BTH and Walther H (2001) Generation of photon number states on demand via cavity, quantum electrodynamics. *Physical Review Letters* 86: 3534–3537.

particle in a harmonic potential, as was first pointed out by D. F. Walls and H. Risken; this connection results from the fact that the radiation field is quantized on the basis of the harmonic oscillator. The Jaynes–Cummings dynamics can therefore also be observed with trapped ions, as recently demonstrated in a series of beautiful experiments by D. Wineland *et al.* They also produced a Schrödinger cat state by preparing a single trapped ion in a superposition of two spatially separated wavepackets which are formed by coupling different vibrational quantum states in the excitation process.

Besides the experiments in the microwave region, a single-atom laser emitting in the visible range has also been realized by M. Feld *et al.* Furthermore, cavity quantum electrodynamic effects have been studied in the optical spectral region by J. L. Kimble *et al.*

Today's technology in microfabrication of semiconductor diode structures allows the realization of low-order cavity structures for diode lasers. In these systems the spontaneous emission is controlled in the same way as in the micromaser. Since spontaneous decay is a source of strong losses, control of this phenomenon leads to highly efficient laser systems. Quantum control of spontaneous decay thus has important consequences for technical applications. This topic will be briefly described in the next section.

Microlasers

The simplest approach to fabricating an optical microcavity is to shrink the spacing between the mirrors of a Fabry–Perot resonator to λ/n (where n stands for the refractive index) while reducing the lateral dimensions to a range of the same order of magnitude. This structure provides a single dominant longitudinal field mode that radiates into a narrow range of angles around the cavity axis.

The first optical microcavity experiments used dye molecules between high-reflectivity dielectric mirrors in the Fabry–Perot configuration. Because spontaneous emission is a major source of energy loss, speed limitations, and noise in lasers, the capability to control spontaneous emission is expected to improve laser performance. If the fraction of spontaneous emission coupled into the lasing mode is made close to one, the 'thresholdless' laser is obtained, in which the light output increases almost linearly with the pump power instead of exhibiting a sharp turn-on at the pump threshold.

Semiconductor microcavities provide high-Q Fabry–Perot cavities for both basic studies and potential applications. Molecular beam epitaxy or organometallic chemical vapor deposition techniques are used to deposit high-reflectivity mirrors consisting of alternating quarter-wavelength layers of lattice-matched semiconductors. For example, 15–20 pairs of quarter-wave layers of $Al_{0.2}Ga_{0.8}As$ and AlAs result in a reflectivity greater than 99% and Q values greater than 500. The optically active layer in such a microcavity is typically a GaAs quantum well located at the midplane of the cavity, where the field strength is maximum.

Since the microcavities have an extremely low threshold, their efficiency will also be high. Low-cost, high-density light source arrays and photonic circuits are made possible by the small size and low power consumption of such resonators. It will be possible to produce entire wafers containing millions of microlasers with a multipole arrangement instead of having to cleave each individual semiconductor laser as at present. This improvement will lead to higher yields and lower cost per element. Another advantage of surface-emitting microcavity sources is the efficient optical coupling of their stable symmetric mode patterns into optical fibers or waveguides. These lasers are thus certain to spark a revolution in optical communication in the future.

Conclusion

This article reviewed experiments in cavity quantum electrodynamics. The experiments shed new light on our understanding of the radiation interaction of atoms. The achievable strong coupling between atoms and radiation leads to the possibility to generate entanglement between the generated radiation field and the atoms, presenting a basis for interesting applications in connection with, e.g., quantum computing and quantum information processing. Furthermore Fock states of the radiation field can be generated. The control of spontaneous decay finally leads to interesting new laser systems offering high efficiency.

See also

Quantum Electrodynamics: Cavity QED in Semiconductors; Quantum Theory of the Electromagnetic Field. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Berman PR (1994) Cavity quantum electrodynamics. In: Bederson B and Bates D (eds). *Advances in Atomic, Molecular, and Optical Physics*, Supplement 2. Boston, MA: Academic Press.

Englert B-G, Löffler M, Benson O, *et al.* (1998) Entangled atoms in micromaser physics. *Fortschritte der Physik* 46: 897–926.

Haroche S (1992) Cavity quantum electrodynamics. In: Dalibard J, Raimond J-M and Zinn-Justin J (eds) *Fundamental Systems in Quantum Optics*. pp. 767–935. Amsterdam: North-Holland.

Haroche S and Kleppner D (1989) Cavity quantum electrodynamics. *Physics Today* 42: 24–30.

Kimble HJ, Carnal O, Georgiades N, *et al.* (1995) Quantum optics with strong coupling. In: Wineland DJ, Wieman CE and Smith SJ (eds) *Atomic Physics*, vol. 14, pp. 314–335. New York: American Institute of Physics.

Marrocco M, Weidinger M, Sang RT and Walther H (1998) Quantum electrodynamic shifts of Rydberg energy levels between parallel metal plates. *Physical Review Letters* 81: 5784–5787.

Meschede D (1992) Radiating atoms in confined space: from spontaneous emission to micromasers. *Physics Reports* 211: 201–250.

Meystre P (1992) Cavity quantum optics and the quantum measurement process. In: Wolf E (ed.) *Progress in Optics*, vol. 30. Amsterdam: North-Holland.

Walther H (1992) Experiments on cavity quantum electrodynamics. *Physics Reports* 219: 263–281.

Cavity QED in Semiconductors

M Kira, W Hoyer and S W Koch,
Philipps-University, Marburg, Germany

G Khitrova and H M Gibbs,
University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

According to classical electromagnetism, propagating light fields obey Maxwell's equations under which electric and magnetic fields exchange their strength in an oscillatory manner. If a propagating field is followed at a single reference point, it is convenient to represent the electric field as a complex number $|E(t)|\exp(-i\omega t)$ where $|E(t)|$ is the magnitude of the field and the phase of the field $\exp(-i\omega t)$ oscillates in time t with the optical frequency ω. Figure 1 illustrates how the classical field is determined by one single point in the complex plane; consequently, the phase and amplitude of the field are known exactly. According to quantum mechanics, this simple picture has to be altered because the Heisenberg uncertainty relation states that complementary quantities like phase and amplitude cannot be determined precisely at the same time. As a result, the real and imaginary parts of the field can be determined only with accuracy $\Delta\mathrm{Re}[E]$ and $\Delta\mathrm{Im}[E]$, respectively, and the Heisenberg uncertainty principle determines the best possible accuracy $\Delta\mathrm{Re}[E]\,\Delta\mathrm{Im}[E] \geq 1$. To incorporate the fundamental inaccuracy, the electric field has to be defined by using complex-valued distributions or equivalently wavefunctions as shown in Figure 1. This quantization procedure leads to the Schrödinger equation for light analogous to that for particles. The field called quantum optics investigates the quantum electrodynamics (QED) features of light.

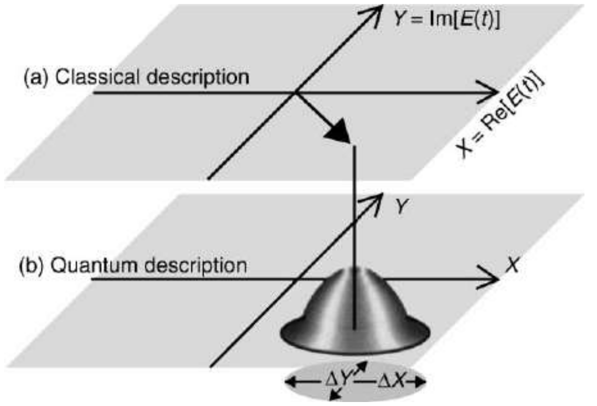

Figure 1 (a) The classically described light field $E(t)$ is a single point in the complex XY-plane; the corresponding (X, Y) vector has a definite phase and angle. (b) Quantum mechanically described light is a distribution with fluctuations ΔX and ΔY described by the shaded circle.

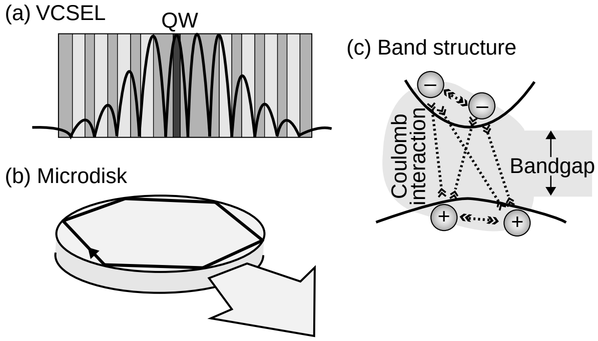

Figure 2 (a) A typical vertical-cavity surface-emitting laser (VCSEL) structure; a stack of quarter-wavelength layers with different refractive index provides a mirror such that the light intensity (solid line) is strongly confined inside the cavity. A quantum well (QW) is positioned at the field maximum. (b) In a microdisk laser, light encounters ideal total reflection when the edge of the disk is reached. The so-called whispering gallery mode emits (leaks) light from the edge. (c) A schematic illustration of semiconductor band structure, bandgap, and Coulomb interaction for an electro–hole system.

In general, light does not propagate freely in space, but it interacts strongly with the surrounding matter. For example, light can be absorbed, and its energy can be converted into excitation of the matter. As a result, the light may be slowed down, reflected, scattered, or diffracted. In order to explain the implications of the light–matter interaction in detail, one obviously has to apply a quantum mechanical description also for the matter. This approach determines the so-called eigenstates of the matter, so that the matter may occupy only certain discrete states with discrete energies. The Rydberg series of a hydrogen atom is a typical example. Similar discrete energy levels are also found for the quantized light; these levels are often referred to as photons which heuristically describe the particle properties of the light. The lowest-order light–matter interaction consists of processes where one photon is absorbed (emitted) while the matter is simultaneously excited (de-excited) from one eigenstate to another. This generic type of interaction leads to a microscopic description of the optical effects mentioned above. The magnitude of these effects depends on the strength of the interaction. By implementing different cavity configurations, the reflected light can be forced to travel across the same matter many times. As a result, the light–matter interaction is enhanced by a factor proportional to the multiple passes of the light. Thus, a cavity can be efficiently used to amplify optical phenomena. Figure 2 shows cavity configurations commonly used for semiconductors. Semiconductor cavity QED investigates quantum optical effects in semiconductor systems by enhancing them with cavity mirrors. In general, quantum optical features produce classically unexpected effects which typically stem from: (a) the discrete nature of eigenstates of light and matter; (b) the superposition principle of quantum mechanics; and (c) the Heisenberg uncertainty principle. In this article, we review both theoretical and experimental advances made so far to predict, observe, and control quantum optical effects in semiconductors with respect to effects (a)–(c).

During the last few decades, atomic quantum optics has already developed toward QED investigations, and semiconductor optics is developing rapidly into the same regime. Advanced atomic QED theories have successfully explained and guided intriguing experimental developments like laser cooling, atom condensation, and photon teleportation. However, the atomic approach can mostly be applied to describe dilute and only weakly interacting atomic gases since relatively simple models of few-level systems are used to describe the material. The elementary electronic excitations in semiconductors consist of electrons and holes (missing electrons) lifted into the conduction and valence bands, respectively. The corresponding eigenstates form continuous energy bands with the band-gap energy separation indicated in Figure 2. Since electrons and holes have opposite charges, they experience Coulomb attraction whereas bare electrons or holes repel each other. Under favorable conditions, the attractive Coulomb interaction prevails and atom-like bound electron–hole pairs (excitons) may be formed. However, since the Coulomb force has an infinite range and an electron–hole system is typically dense, excitons cannot be treated as weakly interacting quasiparticles. As a result, the atomic QED approaches cannot be used to describe quantum optics of semiconductors in general. Thus, this article concentrates mostly on semiconductor QED theory which fully includes the interaction of charge carriers, i.e., electrons and holes, as well as quantum features of light.

The development work on quantum devices and processes is vital for advancements in several key technologies such as computers and telecommunications. The continuous decrease in component dimensions leads to a microscopic structure size (less than 1 μm); at the same time, increase in the device performance is accompanied by ultrafast operation time (faster than 1 ps). Due to these development trends, the properties of components and processes are becoming more and more quantum mechanical. Although the description of the quantum processes is complicated, the microscopic behavior offers entirely new operational functionality such as massively parallel quantum computers. These possibilities are based on the controlled manipulation of the quantum mechanical state of the light and matter. Due to rapid recent developments, optically coupled semiconductor devices have a great potential to become technologically successful and commercially viable quantum components.

Quantized States: Observation via Strong Light–Matter Interaction

In order to enhance the light–matter interaction, we choose a system where the semiconductor is placed inside a cavity in the position where the field intensity is maximum. For an empty cavity, the field intensity has a strong dependence upon frequency with a strong resonance at ω_R when the cavity length is an integer number of light wavelengths in the material. The width of the resonance is directly related to the quality of the cavity as determined by the number of back-and-forth reflections of light inside the cavity. If one chooses a vertical-cavity surface-emitting laser (VCSEL) structure, the most efficient coupling is obtained by placing a thin planar semiconductor structure at the field maximum as shown in Figure 2. If the semiconductor is planar and thin, the structure is called a quantum well because electrons and holes are confined in one direction in analogy to the standard particle-in-a-box problem of fundamental quantum mechanics. As a result, the energy levels of a quantum well are discrete in the confinement direction. If the structure is narrow enough, the system dynamics is confined to the lowest quantum-well level such that the carriers are quasi-two-dimensional due to the free in-plane motion. The optical response of such a system to a weak classical probe beam has been successfully analyzed with the so-called semiconductor Bloch equations. For the quantum-well system alone, the absorption spectrum may have a sharp resonance; this is often referred to as an excitonic resonance since it is located below the fundamental bandgap energy at a position corresponding to the exciton binding energy. Obviously, the light–matter coupling becomes large when the exciton and cavity resonances coincide. However, since these resonances are coupled, the optical response is altered; we observe a splitting into two absorption peaks instead of the original degenerate resonances. This phenomenon is commonly referred to as normal-mode coupling; in general, it is a quite common feature in quantum mechanics that degenerate states split due to an additional interaction. Figure 3 shows the first experimental observation of normal-mode coupling in semiconductors.

Since a semiconductor is a strongly interacting many-body system, several effects alter the specific character of an excitonic resonance when electrons and holes are excited. The attractive interaction between electrons and holes becomes weaker for

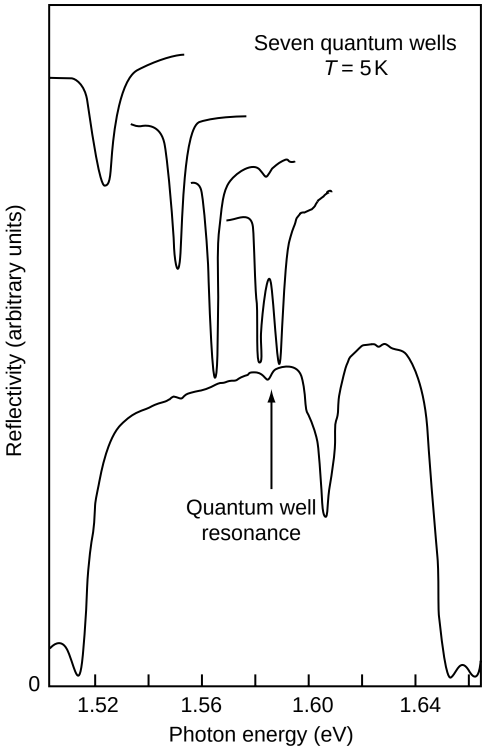

Figure 3 Probe reflectivity of a seven-quantum-well microcavity structure. The various curves correspond to the energy difference between exciton and cavity resonances. When these resonances become degenerate, the reflectivity shows two resonances corresponding to the normal-mode splitting. From Weisbuch C, Nishioka M, Ishikawa A and Arakawa Y (1992) Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Physical Review Letters* 69: 3314. Copyright (2004) by the American Physical Society.

increased carrier density since the surrounding charges screen the bare Coulomb interaction. A particular exciton may also experience scattering from nearby electrons, holes, or other excitons via the Coulomb interaction. Furthermore, electrons and holes are Fermions, which requires that any given carrier state can only be occupied once. This poses fundamental limits on how many carriers can coexist in a specific volume; above a certain limit, additional occupation is prevented, and so-called Pauli blocking is observed. Since an exciton consists fundamentally of Fermionic constituents, eventually Pauli blocking effects become important even for excitons. Due to a combination of these many-body effects, the excitonic resonance is weakened by an increasing carrier density. Figure 4 shows a comparison between theory and experiment of the excitonic absorption spectrum for different carrier densities. For elevated densities, the exciton resonance is broadened and its height reduced. When the same investigations are repeated in a microcavity, we observe that the normal-mode peaks decrease in height but their separation is roughly unchanged for moderate densities. In general, the normal-mode spitting is proportional to the oscillator strength of the excitonic absorption, i.e., the area under the resonance. Thus, the observed constant splitting suggests an unchanged oscillator strength for moderate densities. The invariant oscillator strength was unexpected and can only be explained by using a theory which includes the Coulomb interaction of carriers microscopically. When the carrier density is increased even further, the exciton resonance is completely bleached, and the microcavity transmission has only a single peak, at the cavity mode energy; this is commonly referred to as the weak-coupling regime, in contrast to the nonperturbative strong coupling with two transmission peaks. The strong coupling regime provides several intriguing phenomena; for example, parametric amplification of emission has been demonstrated by applying simultaneously multiple light beams to the sample.

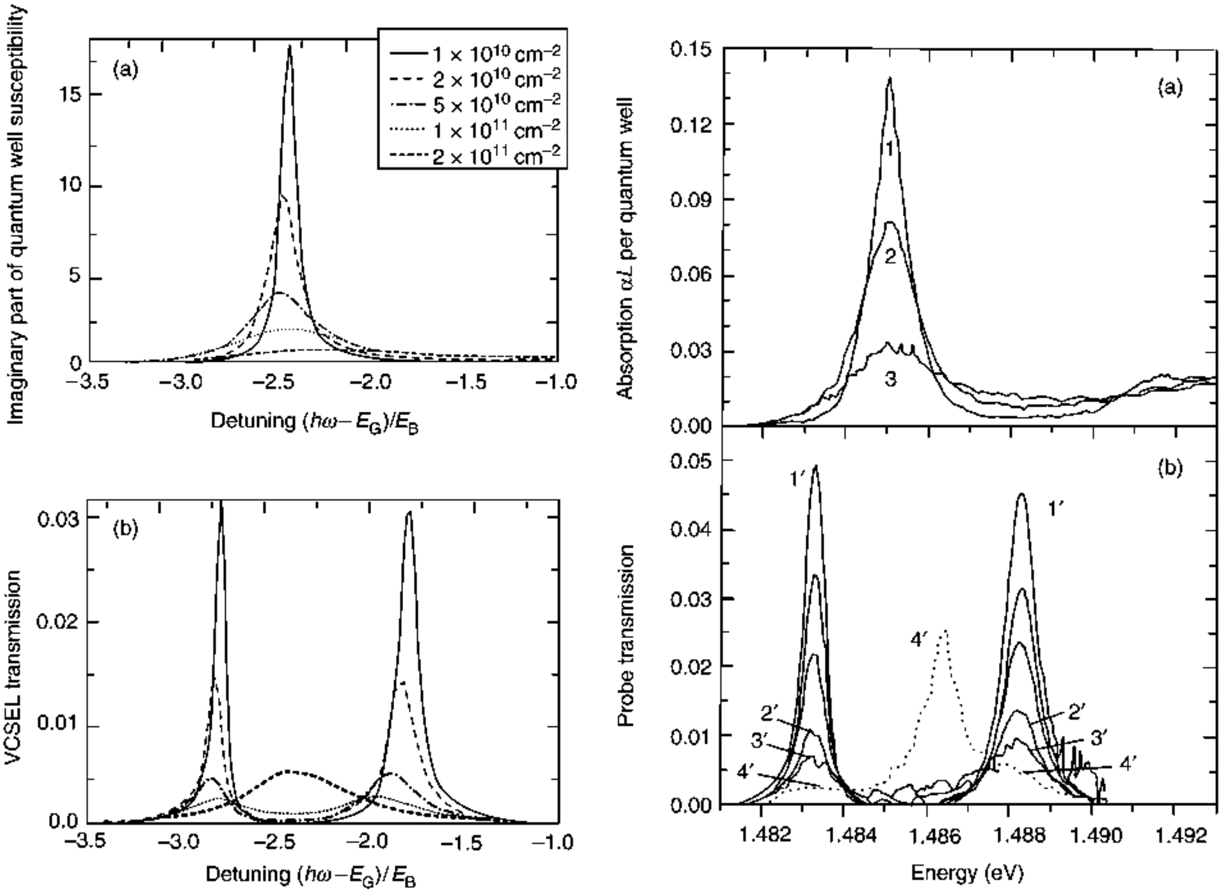

Figure 4 Left column: (a) microscopically computed bare quantum-well absorption spectrum, i.e., the imaginary part of the susceptibility, as a function of carrier density. (b) Calculated transmission of the quantum-well microcavity for corresponding densities. The right column shows experimentally determined (a) absorption and (b) microcavity transmission for the same conditions as in the calculations. From Jahnke F, Kira M, Koch SW, *et al.* (1996) Excitonic nonlinearities of semiconductor microcavities in the nonperturbative regime. *Physical Review Letters* 77: 5257. Copyright (2004) by the American Physical Society.

The ultimate cavity QED limit of normal-mode coupling follows when the light–matter interaction is so strong that even a single photon leads to splitting (genuine strong coupling). We briefly outline some aspects of the quantum statistical limit known from atomic physics. This situation can be analyzed with the so-called Jaynes–Cummings model where only two states of the atom are included together with a single light mode. In this case, the interaction between a single photon and an atom leads to a normal-mode splitting g_0. If two photons interact with one atom, the splitting increases to $2^{1/2}g_0$; more generally coupling between n photons and an atom leads to a splitting of $n^{1/2}g_0$. In the QED limit, the matter response is very nonlinear because one can detect the quantized nature of light directly from the energy splitting. This limit has already been reached in atomic cavity QED, whereas this regime is hard to approach with semiconductors because they behave more like multi-atom systems. If one has N atoms in a cavity, the first photon excites one atom, but there are N different ways of doing this, so the splitting is $N^{1/2}g_0$. If a second photon arrives, it interacts with the remaining unexcited atoms rather than the excited one, so that the splitting remains $N^{1/2}g_0$ as long as the number of photons is much smaller than N. Thus, in an N-atom system, or equivalently in semiconductors, the quantized light effects are much harder to observe than for a single atom. Currently, the normal-mode splitting in a semiconductor microcavity is explainable by the classical features of light, so that only the quantized nature of the matter is observed. In order to approach the ultimate cavity QED limit, one obviously has to reduce N and increase the light–matter coupling. In the future, this objective might be possible in quantum-dot/nanocavity systems where the semiconductor is confined in all spatial directions.

When an excited semiconductor is not under any influence of external classical light fields, it can still emit light via spontaneous emission resulting from the recombination of electron–hole pairs. The resulting light emission is called photoluminescence. Since the emission process takes place in a strongly interacting many-body system, one has to systematically include the Coulomb interaction and Fermionic features. The emission properties can be consistently described by the so-called semiconductor luminescence equations. When the microcavity photoluminescence spectrum is investigated, one observes a similar normal-mode splitting as for the transmission studies. However, this normal-mode coupling is not in the ultimate cavity QED limit even though the quantum fluctuations of the light field trigger the spontaneous emission. Nevertheless, the luminescence shows an interesting new transition as a function of the excitation level. Figure 5 contains a comparison between theory and experiment of normal-mode peak heights and positions as a function of carrier density. For low densities, the high-energy peak is lower in height but it overtakes the low-energy peak for moderate carrier densities that are still below lasing threshold. This threshold-like overtaking was attributed to Boser action, involving exciton formation, final-state stimulation, and Bose condensation. The inset shows an estimate of just how Bosonic the excitons are, i.e., how much they actually behave as independent atoms; the value unity corresponds to the fully Bosonic situation. The nonlinear luminescence transition takes place at a density regime where the underlying Fermionic electron and hole contributions become important (commutator is roughly 0.5). A more detailed analysis with the semiconductor luminescence equations shows that Fermionic carrier nonlinearities in a detuned cavity are responsible for the experimental observations. This example of a misinterpretation based on a Bosonic analysis stresses how important it is to include the Coulomb interaction and Fermion character of carriers when analyzing the properties of semiconductors.

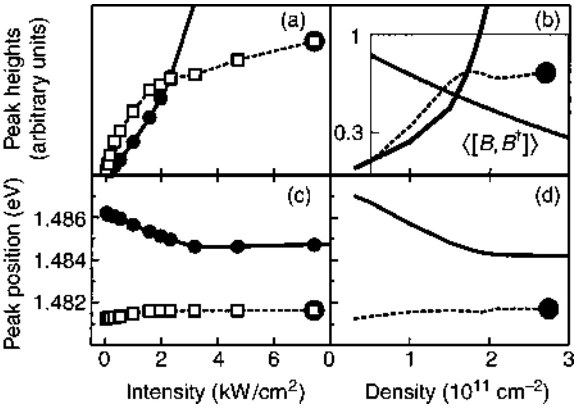

Figure 5 Measured microcavity luminescence intensities (a) and peak energies (c) versus excitation intensity for the high-energy peak (solid line) and the low-energy peak (dashed line). The results of the microscopic theory are shown in (b) and (d). From Kira M, Jahnke F, Koch SW, *et al.* (1997) Quantum theory of nonlinear semiconductor microcavity luminescence explaining “boser” experiments. *Physical Review Letters* 79: 5170. Copyright (2004) by the American Physical Society.

Superposition Principle: Observation of Quantum Optical Correlations

The above normal-mode-coupling investigations have shown interesting nonlinear effects, but they also revealed that the ultimate QED limit – demonstrating discrete states of light directly – has not yet been reached. Alternatively, a QED investigation may concentrate on other fundamental features of quantum mechanics. One possibility is to study the implications of the superposition principle which states that the joint properties of a light–matter system can always be expressed as a superposition that combines light and matter wavefunctions. The most dramatic consequence can be observed in so-called entangled wavefunctions which cannot be factorized into light and matter parts. In a factorized wavefunction, the light and matter parts are independent whereas entangled wavefunctions conditionally connect light and matter wavefunctions. To elaborate the subtle details of entanglement, we first analyze a simple example. Assume that light can be in two different polarization states, $|\sigma^+\rangle$ or $|\sigma^-\rangle$, and the matter part is either excited $|\mathrm{up}\rangle$ or de-excited $|\mathrm{down}\rangle$. A wavefunction, $[|\mathrm{up}\rangle + |\mathrm{down}\rangle][|\sigma^+\rangle + |\sigma^-\rangle]$, is clearly a superposition of the fundamental states, and at the same time the light and matter parts are completely factorized. However, the wavefunction $[|\mathrm{up}\rangle\ |\sigma^+\rangle + |\mathrm{down}\rangle|\sigma^-\rangle]$ is entangled, since the light and matter parts cannot be separated. In the entangled state, measurement on the light state will conditionally determine a definite state of the matter, whereas the factorized state has no such conditionality. The entanglement has far-reaching consequences which cannot be understood with a classical analysis. For example, the principles of teleportation and quantum computing follow directly from the controlled manipulation of different parts of the entangled wavefunction. For atomic systems, wavefunction entanglement has been demonstrated and utilized in several experiments. For semiconductors, the direct manipulation and detection of the wavefunction seems difficult due to the overwhelming complexity of the many-body wavefunctions. Once again, the entanglement and the wavefunction are simplest for low-dimensional structures; direct entanglement effects have been demonstrated recently in a single quantum dot and between a pair of dots. Also for more complicated semiconductor systems, such as a quantum well in a microcavity, the entanglement can be observed as correlations between light and matter. In this case, the existence of QED correlations basically means that light and matter properties depend conditionally on each other. In general, the direct entanglement and QED correlation investigations have a large development potential for semiconductors.

When a semiconductor is excited with an external light pulse, entanglement-related correlations couple the classical and luminescence emission dynamics. The resulting dynamic interplay between the semiconductor Bloch and luminescence equations is mainly mediated by the correlation between photons and electron–hole densities. In the following, we analyze a microcavity configuration where a strong pump pulse generates large QED correlations. The effect of the correlations is then measured by the response of a weak probe beam. Figure 6 shows a comparison of theory and experiment of probe reflection in a configuration where the pump and probe do not have any spectral overlap. The measured probe reflection displays long-living oscillations as a function of time delay, i.e., phase difference, between the pump and the probe. Only by including the QED correlation in the theory can the oscillatory probe reflection be explained. When the QED contributions are omitted from the theory, the phase difference does not have any effect on the probe reflection. Thus, this experiment–theory example represents a direct observation of the cavity QED effects in semiconductors. Figure 6 also shows more pronounced oscillations when two phase-locked pump pulses provide the excitation; again the full QED theory explains the enhanced oscillation features.

The QED features can also alter the normal-mode coupling characteristic of a weak probe beam. Figure 7 shows a comparison of theory and experiment for a situation where the pump spectrum is

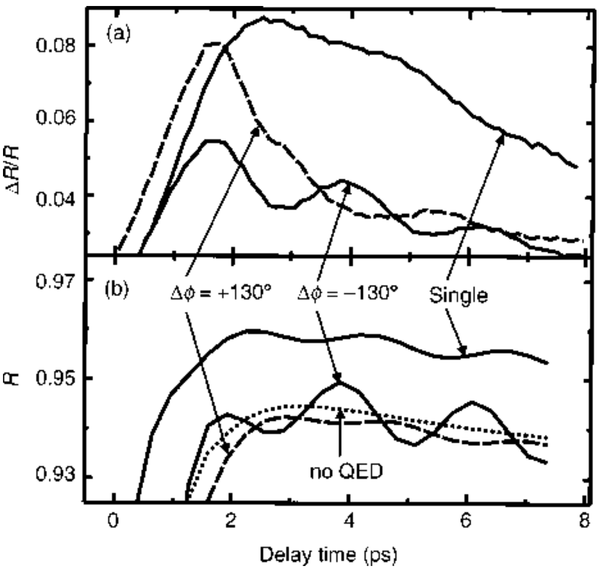

Figure 6 (a) Measured differential reflectivity and (b) computed reflection probability of the probe pulse as function of probe delay with respect to the excitation pulse. For two-pulse excitations, ±130° relative phase shifts are used. The dotted line is computed without the QED correlations. From Lee Y-S, Norris TB, Kira M, *et al.* (1999) Quantum correlations and intraband coherences in semiconductor cavity QED. *Physical Review Letters* 83: 5338. Copyright (2004) by the American Physical Society.

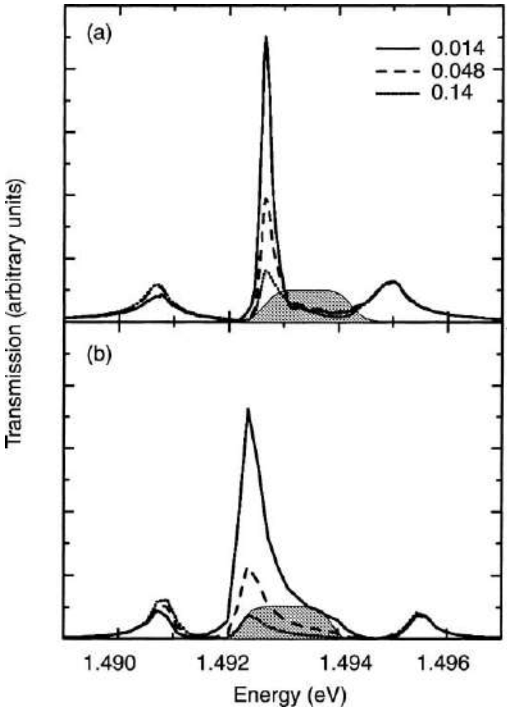

Figure 7 Dependence of probe transmission on probe intensity (nJ cm^{-2}) at a constant pump intensity of 330 nJ cm^{-2} for the microcavity. (a) Experiment, and (b) theory. The shaded regions indicate the pump spectrum. From Ell C, Brick P, Hübner M, *et al.* (2000) Quantum correlations in the nonperturbative regime of semiconductor microcavities. *Physical Review Letters* 85: 5392. Copyright (2004) by the American Physical Society.

located between the two normal-mode coupling resonances. Both the theory and experiment show a new third resonance which follows the energetic position of the pump. The third peak is a direct consequence of the QED correlations. Since the third peak is generated from the QED field–carrier correlations generated by the pump pulse, its magnitude follows the intensity of the pump. As the probe intensity is decreased the relative effect of the QED correlations is increased so that the third peak grows in the probe transmission.

In these QED investigations, the cavity plays an important role since the mirror feedback leads to a significant amplification of the QED effects through the enhanced light–matter coupling. From a practical point of view, normal-mode coupling also provides well-separated spectral features which can be used as classical-emission reference points.

Heisenberg Uncertainty Principle: Squeezing of Light Emission

Quantized light effects can also be investigated by measuring the light fluctuations of the field directly. To maximize the quantum effects, we analyze the field fluctuations in the emission directions where the classical field vanishes. This situation can be realized in planar quantum-well structures which are nearly free of disorder. In such systems, a light pulse propagating perpendicular to the structure leads to transmission and reflection of classical light only along the excitation axis. If the detection is performed at an angle, the so-called secondary emission is purely quantum mechanical. However, the classical and quantum emissions are coupled in the same way as for the QED correlation study. The resulting fluctuations of secondary emission obey the Heisenberg uncertainty principle; in the following, the special quantum features of these fluctuations are investigated.

In the full analysis, we determine the variance ΔX and ΔY of the emission as shown in Figure 8 (see also Figure 1). The Heisenberg uncertainty principle requires that the quadrature fluctuations obey $\Delta X \Delta Y \geq 1$. For a specific quadrature, the minimum uncertainty limit is usually defined to be $\Delta X = 1$ or $\Delta Y = 1$. If ΔX and ΔY are different, the observed light field is squeezed; and if the variance in one quadrature is less than one, the field is squeezed below the minimum uncertainty limit. In both cases, the field has a strong quantum nature; in particular, squeezing below unity suggests that measurements can be more accurate than the standard quantum limit for that quadrature.

To illustrate the behavior of the quantum fluctuations, we excite the quantum well resonantly with a relatively strong pulse. The emission is detected at an angle of 45° away from the excitation axis. Since the excitation pulse is relatively strong, the carrier density starts to oscillate during the pulse because Fermionic carriers can be excited only once; thus, further excitation actually leads to de-excitation. In other words, when these states become almost fully excited, the pulse can no longer excite the system, and we observe periodic de-excitation and excitation analogous to the Rabi-flopping of a strongly excited two-level system. Figure 8 shows the exciting pulse and corresponding quadrature fluctuations during the excitation process. As long as the pulse is present, the squeezing is at the 4% level and oscillates with the Rabi-flopping frequency. The quadrature fluctuations clearly show squeezing below the minimum uncertainty limit. Hence, the field has obvious quantum properties.

Similar squeezing and quantum characteristics statistics have been predicted and observed in the photon statistics of resonance fluorescence from two-level atoms subjected to an intense coherent light beam. The quantum properties of the scattered light

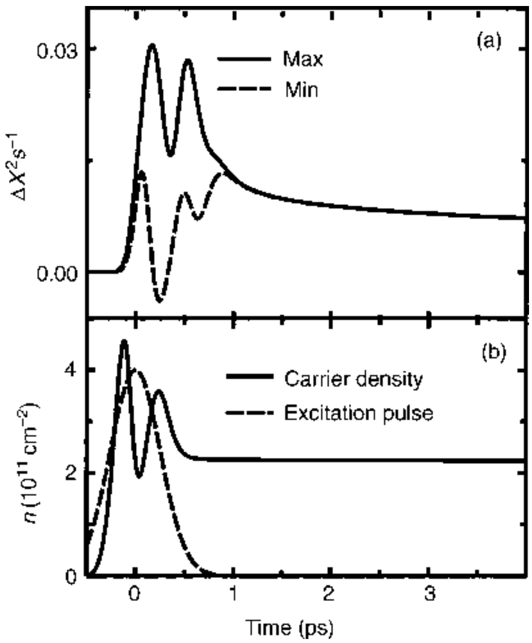

Figure 8 (a) Time-resolved maximum (solid line) and minimum (dashed line) deviations of the mode quadratures from the vacuum value of unity. (b) Corresponding time-resolved carrier density (solid line) and excitation pulse (dashed line). From Kira M, Jahnke F and Koch SW (1999) Quantum theory of secondary emission in optically excited semiconductor quantum wells. *Physical Review Letters* 82: 3544. Copyright (2004) by the American Physical Society.

in both quantum-well and atomic emission are enhanced when the driving field forces the system to oscillate between the excited and de-excited state. Just as described above, when the system becomes de-excited, it can no longer emit an additional photon; this inhibition manifests itself as sub-Poissonian photon statistics and squeezing in the mode quadratures. When the excitation pulse re-excites the system, such restrictions are no longer imposed. Thus, the field properties show quantum features oscillating with the Rabi-flopping frequency. Squeezing effects can also be observed with excitation by an electric current; for example, amplitude squeezing of diode laser emission has been demonstrated by controlling the electron statistics of the current flow.

The squeezing investigations demonstrate that some QED effects are expected to be seen without a cavity. Figure 9 shows the quadrature fluctuations when the cavity is added to the squeezing analysis. We observe qualitatively similar squeezing, but now the level of squeezing is enhanced up to 30% compared to the minimum uncertainty. This is once again a demonstration how QED effects can be amplified by using an increased light–matter coupling provided by the cavity.

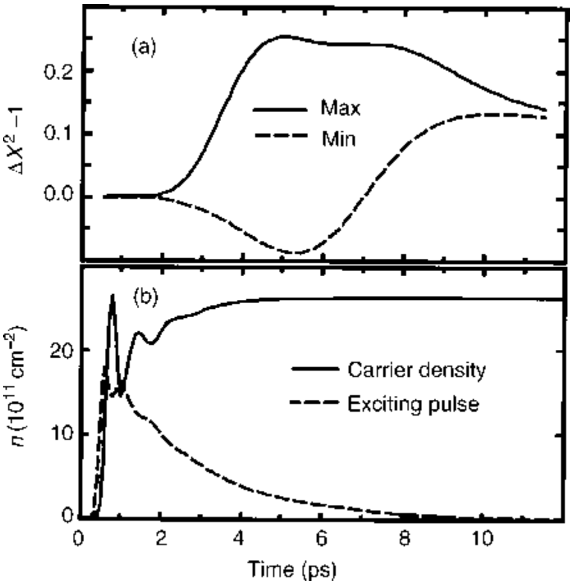

Figure 9 Squeezed light from a quantum well inside a microcavity: (a) the time evolution of maximum and minimum of the quadrature fluctuations; (b) the corresponding density and light field intensity at the quantum well.

Summary

This article overviews some of the most intriguing features of cavity quantum electrodynamics in semiconductors. Even though the related research has started only recently, several important quantum phenomena have already been predicted and measured: (a) the discrete quantum mechanical states of light and/or matter can be measured with strong coupling; (b) the consequences of the superposition principle have been detected as light–matter entanglement; and (c) the Heisenberg uncertainty principle has been tested via the squeezing of light. In all cases, the cavity enhances light–matter coupling leading to more pronounced quantum effects. Compared to atomic systems, the semiconductor has strong Coulomb correlation effects due to the relatively high density, which makes the analysis challenging. However, the semiconductor can also provide new mechanisms like the photon/carrier-density correlations which trigger new unexpected quantum phenomena.

The field of semiconductor quantum optics is very active and is developing rapidly. It benefits both from advances in computer capabilities and in semiconductor crystal growth. The increasing computer capacity allows more profound, accurate, and realistic modeling of semiconductor structures. At the same time, advances in crystal growth technology provide us with improved samples which are almost disorderless. In particular, the growth of quantum wells with narrower exciton linewidths and quantum dots with larger dipole moments and reduced dephasing rates may be achieved, which is certain to make quantum features increasingly apparent and unavoidable.

All these research efforts eventually focus on producing devices utilizing quantum mechanical principles. One of the main objectives endeavors to develop quantum logic components for building blocks of quantum computers. Similar expanding possibilities can be expected, e.g., for accuracy of detection, device efficiency, and component design in general. Considering all of this, we are almost guaranteed to see new astonishing advancements; in this bright future, semiconductor cavity QED research will most likely be a pre-eminent element.

See also

Quantum Electrodynamics: Cavity QED.

Further Reading

Allen L and Eberly JH (1987) Optical Resonance and Two-Level Atoms, 1st edn. New York: Dover.

Banyai L and Koch SW (1993) Semiconductor Quantum Dots, 1st edn. Singapore: World Scientific.

Berman PR (ed.) (1994) Several articles in *Cavity Quantum Electrodynamics*, 1st edn. Boston, MA: Academic Press.

Cohen-Tannoudji C, Dupont-Roc J and Grynberg G (1989) Photons and Atoms, 3rd edn. New York: Wiley.

Haug H and Koch SW (2004) Quantum Theory of the Optical and Electronic Properties of Semiconductors, 4th edn. Singapore: World Scientific.

Khitrova G, Gibbs HM, Jahnke F, Kira M and Koch SW (1999) Nonlinear optics of normal-mode-coupling semiconductor microcavities. *Reviews of Modern Physics* 71: 1591–1639.

Kira M, Jahnke F, Hoyer W and Koch SW (1999) Quantum theory of spontaneous emission and coherent effects in semiconductor microstructures. *Progress in Quantum Electronics* 23: 189–279.

Scully MO and Zubairy MS (1999) Quantum Optics, 1st edn. Cambridge, UK: Cambridge University Press.

Walls DF and Milburn GJ (1994) Quantum Optics, 1st edn. New York: Springer-Verlag.

QUANTUM OPTICS

Contents

Atom Optics
Atomic Coherence Effects
Entanglement and Quantum Information
Laser Cooling of Ions
Quantum Computing with Atoms
Squeezed Phonons in Solids

Atom Optics

A D Cronin and D E Pritchard, Massachussetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

In 1923 Louis de Broglie suggested that particles of matter propagate as waves with a wavelength

$$\lambda_{dB} = h/p \qquad [1]$$

where h is Planck's constant and p is the particle's momentum. This breakthrough idea motivated Schrödinger's equation, which is the equation of motion of quantum mechanics.

Schrödinger's equation is a wave propagation equation, and therefore implies the existence of a whole new type of optics – matter wave optics – in which electron waves, neutron waves, atom waves, and entire molecule waves can be manipulated coherently. Such optics enable new measurement technologies and devices that depend on matter wave interference.

Guided by a knowledge of light waves, one can understand that the de Broglie wavelength of a particle sets the scale for focusing, diffraction, and interference. However, there are differences between matter and light. For example, the vacuum is tremendously dispersive for matter waves. In other words, the group velocity for a particle depends on its de Broglie wavelength, while for light waves the group velocity is independent of wavelength.

To illuminate the mathematical connection with light optics, we note the wave equation for light is

$$\nabla^2 \vec{E} = \mu\varepsilon \frac{\partial^2 \vec{E}}{\partial t^2} \qquad [2]$$

where $\vec{E} = \vec{E}(\vec{x}, t)$ is the optical electric field, and the parameters μ and ε may vary throughout space, describing an index of refraction. The wave equation for matter waves is Schrödinger's equation:

$$\left(-\frac{\hbar^2\nabla^2}{2m} + V\right)\psi = i\hbar\frac{\partial}{\partial t}\psi \qquad [3]$$

where $\psi = \psi(\vec{x}, t)$ is the probability amplitude, or wavefunction, for a particle of mass m. Here, the potential energy V can vary throughout space, and thus control the phase and amplitude of propagating atoms. Just as density for photons is proportional to $\vec{E}(\vec{x})^2$, the probability density for particles is given by $|\psi(\vec{x})|^2$.

In a vacuum $V = 0$ and $(\mu_0\varepsilon_0)^{-1/2} = c$, and if we assume that both ψ and E have a time dependence described by $e^{-ik\omega}$, the wave equations become

$$\left[\nabla^2 + \frac{\omega^2}{c^2}\right]\vec{E} = 0 \qquad [4]$$

and

$$\left[\nabla^2 + \frac{2m\omega}{\hbar}\right]\psi = 0 \qquad [5]$$

which look similar, but have a critical difference stemming from the single time derivative in Schrödinger's equation. To see how this makes the vacuum dispersive for matter waves, but not for light, consider a propagating plane wave of either ψ or $\vec{E}$ described by $e^{i(kx-\omega t)}$ (for either kind of wave, $k \equiv 2\pi(\lambda)^{-1}$). Using the wave equation for light waves, the dispersion relation is

$$\omega = ck \qquad [6]$$

where c is the phase (and group) velocity of light, which is independent of k. The same plane wave substituted into Schrödinger's equation for matter waves yields the dispersion relation

$$\omega = \frac{\hbar k^2}{2m} \qquad [7]$$

This is quadratic for matter waves, which means the group velocity $v_g = d\omega/dk$ is twice the phase velocity $v_p = \omega/k$, and both still depend on k.

From an engineering perspective, matter wave optics consists of lenses, mirrors, beam-splitters, and other components of an optics toolkit. The fact that electrons and neutrons can be transmitted through solid material, and that electrons can be reflected or refracted by static electromagnetic fields, has enabled optical elements to be created for these kinds of matter waves. For atoms, however, finding suitable materials for lenses or mirrors is a daunting task, since with rare exceptions (e.g., cold H on liquid He) atoms thermalize on surfaces rather than elastically bouncing off or passing through them. This was the major obstacle which delayed the blossoming of atom optics until the mid-1980s.

This obstacle has now been circumvented, but only with difficulty. Electromagnetic fields (whose spatial configurations are severely constrained by Maxwell's equations) can change the potential energy of neutral atoms, but since an atom's net charge is zero, provide only a weak and highly dispersive index of refraction based on the atom's polarizability or magnetic moment. However, these effects are tremendously enhanced if the electromagnetic fields oscillate near a transition frequency of the atom. Hence, propagating waves and standing waves of single-frequency laser light have become standard tools in atom optics. Standing light waves are often referred to as a light crystal, because the periodic potential can diffract atom beams.

An alternative source of atom optical components is based on nanofabrication. Thin sheets of material with patterned holes serve as absorptive atom optics by transmitting atom waves only through the nanometer scale openings. A variety of techniques have been used to make physical diffraction gratings, zone plates, and holograms for atom waves. Methods such as electron beam lithography and ultraviolet photolithography both rely on sophistocated etching procedures to make the final free-standing structures. Thus, in either approach – near-resonant light, or miniature physical structures – current technologies such as the tunable laser or nanofabrication techniques have been required for the development of useful atom optics.

While coherent manipulation of atom waves is often the goal, incoherent manipulations are possible also, e.g., if photons are spontaneously scattered. Dissipative processes offer an additional class of possibilities such as increasing the brightness of an atom source, and are therefore an important part of atom optics. This article excludes atom slowing and trapping, and focuses instead on atom diffraction – a coherent process which can be used to make beam splitters for atom interferometers.

Nanofabricated Atom Optics

Because atoms stick to surfaces, mechanical structures are absorptive atom optics. Yet the de Broglie waves transmitted through slots of a mask are nearly unaffected, provided the mask is thin enough. This is because the van der Waals potential energy of interaction with nearby surfaces falls off rapidly for neutral atoms.

Consider, for example, an atom wave packet being diffracted by a mechanical grating made with physical bars and slots as shown in Figure 1. Initially, let the atom wave packet have a slowly varying spatial envelope, $f = f(x, y, z)$, and momentum only in the $\hat{x}$-direction, $\vec{p} = \hbar\vec{k}_0 = h(\lambda_{dB})^{-1}\hat{x}$, so the wavefunction can be written as

$$\psi = fe^{i(k_0x - wt)} \qquad [8]$$

Immediately after the grating, only the portions of the wavefunction passing through slots remain nonzero, as depicted in Figure 2. Here, the wavefunction is now periodic both due to the grating lattice vector $k_g\hat{y}$ and the initial wavevector $k_0\hat{x}$. According to the Schrödinger equation, a periodicity in any direction implies a momentum in that direction. Thus, variation of ψ in the $\hat{y}$ variation gives rise to a new momentum distribution.

The square wave modulating the envelope of the wave packet can be described by a Fourier sum times the incident wavefunction

$$\psi = fe^{i(k_0x - \omega t)} \sum_{n=-\infty}^{\infty} a_n \cos(nk_g y) \qquad [9]$$

where $k_g = 2\pi(d_g)^{-1}$ is the grating lattice vector. The only other physical parameter of the grating is the open fraction, γ, and this determines the Fourier amplitudes a_n

$$a_n = \frac{\sin(n\pi\gamma)}{n\pi} \qquad [10]$$

From the principle of superposition (which is applicable here because the Schrödinger equation is linear) the wavefunction can be written as a sum of traveling waves with wavevectors $\vec{k}_n$:

$$\psi = f \sum_{n=-\infty}^{\infty} a_n e^{(i\vec{k}_n \cdot \vec{r} - wt)} \qquad [11]$$

$$\vec{k}_n = \frac{k_0\hat{x} + nk_g\hat{y}}{N} \qquad [12]$$

where, just as in standard optics, the factor N serves to keep $|\vec{k}_n|$ on a spherical shell, i.e., the grating does not add energy to the atom.

Inserting this expression for the wavefunction into Schrödinger's equation we get the dispersion relation

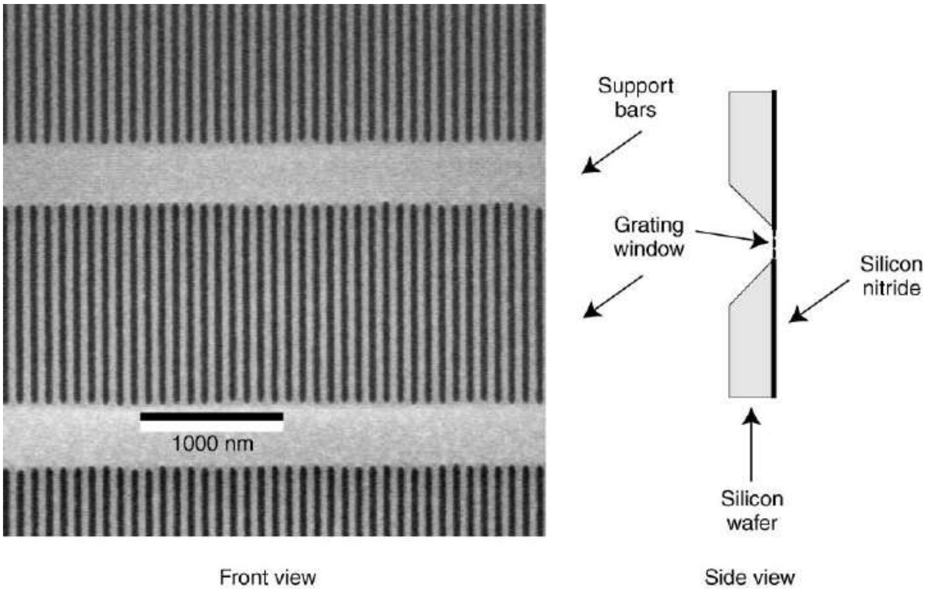

Front view

Side view

Figure 1 Scanning electron microscope image of a silicon nitride diffraction grating for atom waves. The bars (light) absorb, and the slots (dark) transmit atom waves. The grating period is 100 nm, the open fraction is 0.67, and the free-standing bars are supported every 5 μm with a thicker strip of silicon nitride material. The grating was fabricated by Tim Savas and Henry I Smith at MIT NanoStructures laboratory.

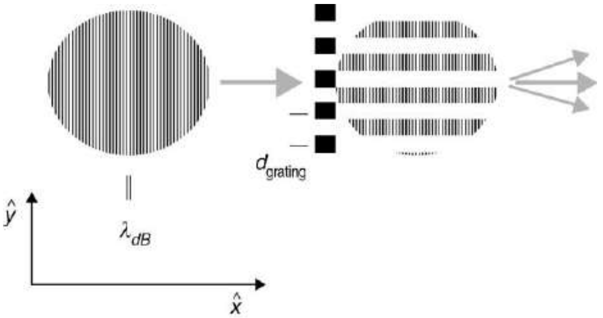

Figure 2 Schematic diagram of an atom wavepacket passing through an absorbing grating, as seen from above. The atom wavepacket is depicted before and after passing through the grating from left to right.

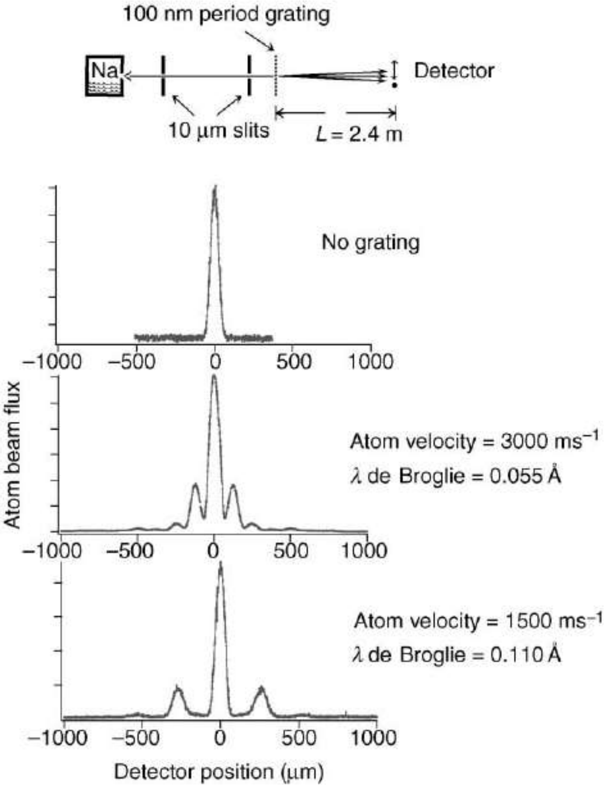

Figure 3 Diffraction data from a beam of sodium formed by passing atoms through two collimating slits followed by a nanofabricated diffraction grating. The momentum of the atoms determines their de Broglie wavelength, and thus their diffraction angle $\sin(\Theta_D) = \lambda_{dB}/d_{\text{grating}}$. These data were obtained at MIT by the authors.

for matter waves and a different group velocity, v_n, for each k_n

$$\hbar\omega = \frac{\hbar^2 k_n^2}{2m} \qquad [13]$$

$$v_n = \frac{d\omega}{dk_n} = \frac{\hbar k_n}{m} = p_n/m \qquad [14]$$

Thus, diffraction through a grating puts an atom into a coherent superposition of different momenta, each separated by $\sim\hbar k_g \hat{y}$. Some time later, the atom will be in a coherent superposition of multiple locations, and the far-field probability density will be peaked at integer multiples of the diffraction angle

$$\Theta_D = \frac{k_g}{k_0} = \frac{\lambda_{dB}}{d_{\text{grating}}} \qquad [15]$$

where d_{grating} is the lattice spacing of the grating and the small-angle approximation for $\sin(\theta)$ has been used. Exactly as in light optics, these diffraction gratings are good momentum spectrometers, i.e., the diffraction angle is a measure of longitudinal momentum. (Contrary to light optics, group velocity can also be used to measure matter wave momentum.) Coherence between the diffraction orders remains until a perturbation occurs which would (even in principle) be sufficient to determine which path the atom took. This assertion has been tested with atoms in an interferometer (discussed below).

Atom flux in the nth diffraction order is given by

$$I_n = \left(\frac{\sin(n\pi\gamma)}{n\pi}\right)^2 \qquad [16]$$

To resolve the diffracted orders the initial transverse momentum distribution of the atoms must be smaller than k_g, and the time until the far-field observation must be sufficiently large given the beam width. At intermediate times, the atom waves are in the Fresnel diffraction region where the Talbot effect can be observed. This effect causes the atom waves to form a replica of the grating in free space at integer multiples of the Talbot length, $Z_T = 2d^2_{\text{grating}}\,\lambda_{dB}^{-1}$.

Figure 3 shows the far-field diffraction pattern from a beam of sodium (^{23}Na) atoms using a material grating with a period of 100 nm. The atom beam velocity peaked at 1500 ms^{-1} corresponds to a de Broglie wavelength of 0.11 Å; thus, $\Theta_D \approx 10^{-4}$ rad. (The beam of 3000 ms^{-1} atoms diffracts at half this angle.) The spread in longitudinal velocity of $\sigma_v/v \approx 6\%$ can be deduced from the data because it slightly increases the width for higher order diffraction peaks. The open fraction of the grating, $\gamma \approx 60\%$, can also be determined from the diffraction data, because it determines the relative intensity of the orders. Effects of the atoms' finite size and interaction with the grating bars make small corrections to this simple theory.

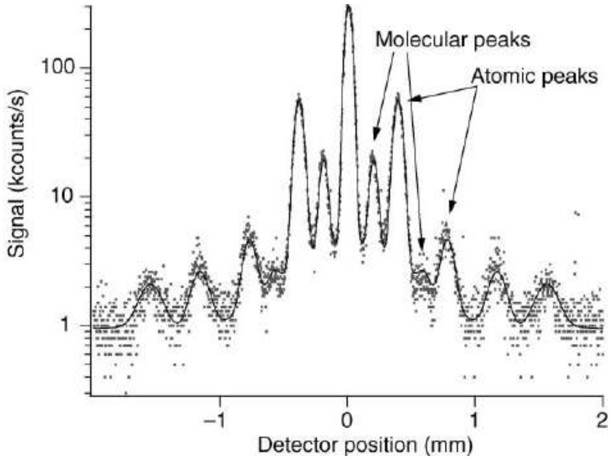

Figure 4 Diffraction data using a mixed beam of sodium atoms (Na) and sodium dimer molecules (Na_2). From this source molecules have the same speed as atoms, thus twice the momentum. Thus, molecule waves are diffracted at half the diffraction angle compared to atom waves. These data were obtained at MIT by the authors.

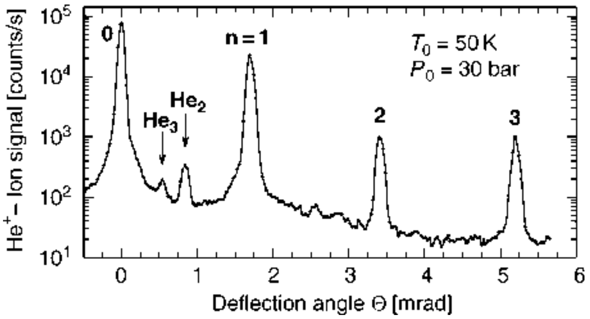

Figure 5 Diffraction of helium atoms and helium molecules through a nanofabricated grating. These data, reproduced with permission from Wieland Schoellkopf, were obtained at the Max-Planck-Institute in Göttingen.

Nanofabricated atom optics differ from those based on light forces in a number of ways: they are amplitude structures (with corresponding loss of transmission intensity), they are species-independent, their scale size can be several times smaller than attainable with light, and they can be arbitrarily patterned since they are fabricated by electron beam lithography. Insensitivity to species is demonstrated in Figure 4, where a beam of Na_2 molecules and Na atoms is diffracted and separated by a nanofabricated diffraction grating. The flexibility of electron beam lithography has allowed fabrication not only of diffraction gratings, but also of spherical and cylindrical zone plates, as well as a combination of lens and a hologram that generates a focused atom image. One stunning application laid to rest a long-standing argument concerning whether a stable bound state of the $^4He_2$ dimer exists. For this a diffraction grating was used to separate and resolve $^4He_2$ and more massive clusters (Figure 5). Subsequently a nano-sieve was used to estimate the size of the $^4He_2$ dimers to be 62 ± 10 Å.

Standing Waves of Light

In contrast to nanofabricated structures, standing waves of near-resonant light are phase gratings for atoms. The optical electric field dynamically polarizes the atoms and causes a shift to the ground state energy (the ac-Stark shift). For optical electric fields oscillating above (below) the atomic resonance frequency, the ac-Stark shift increases (decreases) potential energy for atoms and in either case perturbs the de Broglie wave phase.

In many respects, the interaction of an atom with a standing light wave is richer than the more familiar topic of light–atom interactions in a traveling light wave. Part of this richness reflects two ways in which a standing wave can be considered, either as two counter-propagating traveling waves or as a single, stationary standing wave. The standing wave–atom interaction is capable of transferring momentum in well-determined quantities, coherently splitting atomic wave packets, and generating forces much larger than possible with spontaneously scattered light. Not surprisingly, this interaction has many distinct facets that only appear in different regimes of the interaction parameters (intensity, detuning, and pulse duration) and atomic parameters (mass, initial momentum with respect to the standing wave, and excited state natural lifetime).

We begin with the Bragg scattering of atoms from a standing wave light grating. Although it can be difficult to realize the physical conditions that assure its occurrence, pure Bragg scattering is simpler than intermediate cases involving spontaneous decay (from light too close to resonance) or Kapitza–Dirac diffraction (from shorter interaction times).

Consider a standing wave light grating formed by two counter-propagating plane waves (traveling parallel to the z-axis) of equal amplitude, E_0, wavevector, k, frequency, ω, polarization vector, $\hat{e}$, and temporal envelope function, $f(t)$:

$$\vec{E}(z,t) = E_0 f(t)\sin(kz - \omega t)\hat{e} + E_0 f(t)\sin(kz + \omega t)\hat{e} \qquad [17]$$

$$= 2E_0 f(t)\sin(kz)\cos(\omega t)\hat{e} \qquad [18]$$

We would like to work with momentum states as our atomic basis, thus it is easiest to consider the description of the electric field driving the transitions in terms of two counter-propagating traveling waves of definite momentum (eqn [17]), as opposed to the single standing wave they jointly form (eqn [18]).

Momentum is transferred by paired stimulated absorption and emission processes, resulting in a transfer of photons between the traveling waves. An N_B^{th} order diffraction process transfers N_B photons from one traveling wave to the counter-propagating traveling wave and changes the atomic momentum by $2N_B\hbar k\hat{z}$. Furthermore, atomic population is transferred only between $|g, -N_B\hbar k\rangle$ and $|g, +N_B\hbar k\rangle$, where $|g(e), \pm N\hbar k\rangle$ denotes a two-level atom in its ground (excited) state with momentum $\pm N\hbar k$ parallel to the standing wave axis. The excited state remains nominally unpopulated so long as the temporal envelope function, $f(t)$, does not have strong frequency components near the laser detuning

$$\delta = \omega - \omega_0 \qquad [19]$$

where ω_0 is the unperturbed frequency of the atomic transition. Furthermore, for a given initial state ($|g, -N_B\hbar k\rangle$) the uniqueness of the final state ($|g, +N_B\hbar k\rangle$) occurs because of the fundamental assumption that makes Bragg scattering so simple to describe; the uncertainty in the photon energy driving the transitions is small compared to the energy separation between neighboring momentum states. A quantitative discussion of the validity of this assumption will be given later.

We now calculate the probability, $P_1^B(\tau)$, of the first order ($N_B = 1$) Bragg process. Scattering transfers an atom from $|g, -\hbar k\rangle$ to $|g, +\hbar k\rangle$ when the atoms interact with a constant light intensity for a time τ (i.e., in eqns [17] and [18] $f(t)$ is a square wave of unit amplitude and duration τ). The geometry for Bragg diffraction from a standing wave of light is shown in Figure 6, and the energy/momentum states involved in the transition are depicted in Figure 7.

Considering the electronic degree of freedom, we rewrite the Schrödinger equation as

$$(\mathcal{H}_0 + \mathcal{H}_{\rm int})|\psi\rangle = i\hbar\frac{\partial}{\partial t}|\psi\rangle \qquad [20]$$

where the total Hamiltonian $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_{\rm int}$ is a matrix. The wavefunction is a vector of coefficients c_i in the basis $\{|g, -\hbar k\rangle, |e, 0\rangle, |g, +\hbar k\rangle\}$, with normalized population amplitudes for each component:

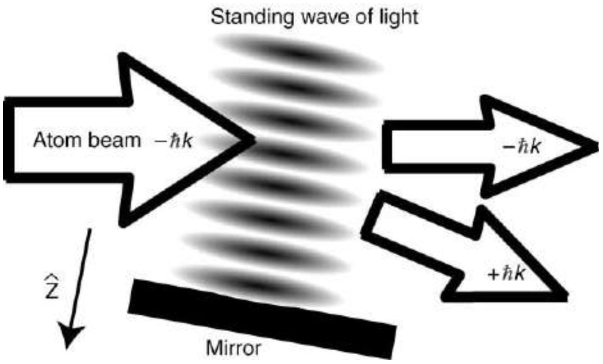

Figure 6 Two counter-propagating running waves of light superimpose to form a standing wave of light. This 'light crystal' acts as a diffraction grating for atoms, shown here in the configuration for Bragg diffraction.

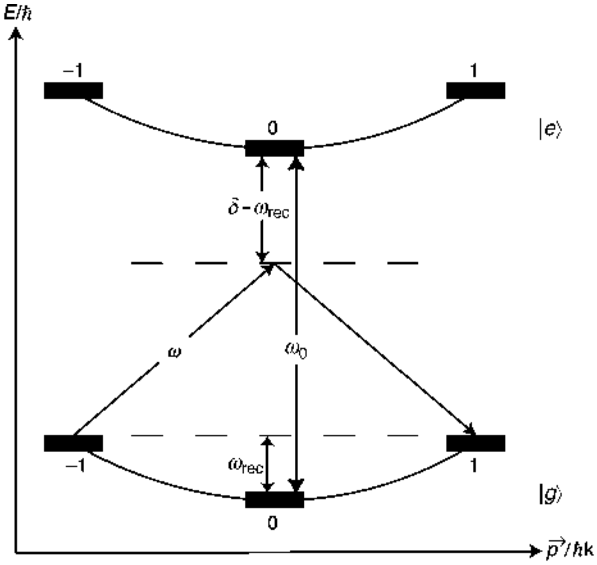

Figure 7 Energy vs. momentum diagram describing atomic states involved in Bragg diffraction. The total energy is due to the electronic state ($|g\rangle$, $|e\rangle$) plus the kinetic energy associated with the recoil from photon emission (or absorption). In Bragg diffraction, atoms make a coherent transition from $|g, -\hbar k\rangle \leftrightarrow |g, +\hbar k\rangle$. Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

$$|\psi\rangle = \begin{pmatrix} c_{-1}(t) \\ c_0(t) \\ c_{+1}(t) \end{pmatrix} \qquad [21]$$

In the electric dipole approximation, the interaction Hamiltonian is $\mathcal{H}_{\rm int}(t) = -\vec{\mu}\cdot\vec{E}(t)$, where $\mu = \langle e|q\vec{r}|g\rangle \cdot \hat{e}$ is the electric dipole matrix element connecting the ground ($|g\rangle$) and excited ($|e\rangle$) states of the atom. $\hat{e}$ is the polarization vector of the light and q is the charge of the electron. By momentum conservation, only the plane wave traveling in the $\pm\hat{z}$ direction couples the $|e,0\rangle \leftrightarrow |g,\mp\hbar k\rangle$ transitions. By using this argument, we are effectively viewing the electric field as a quantum mechanical operator. Expanding the sinusoidal variation of the electric field in terms of complex exponentials and treating the spatially dependent complex exponential terms as quantum mechanical momentum translation operators yields the interaction Hamiltonian:

$$\mathcal{H}_{\text{int}} = \frac{\hbar\omega_R}{2}\begin{pmatrix} 0 & ie^{i\omega t} & 0 \\ -ie^{-i\omega t} & 0 & ie^{-i\omega t} \\ 0 & -ie^{i\omega t} & 0 \end{pmatrix} \qquad [22]$$

For example, $e^{\pm ik\hat{z}}|g,n\hbar k\rangle = |e,(n\pm 1)\hbar k\rangle$ describes absorption, and for stimulated emission, $e^{\mp ik\hat{z}}|e,n\hbar k\rangle = |g,(n\pm 1)\hbar k\rangle$. This procedure includes the rotating wave approximation. In addition, we have neglected any frequency components associated with the sudden switch on of the fields and the finite duration of the light–atom interaction. In formulating $\mathcal{H}_{\text{int}}$, the strength of the light–atom interaction is parameterized by the single-photon Rabi frequency:

$$\omega_R = \frac{\mu E_0}{\hbar} \qquad [23]$$

Without loss of generality, we will take μ and hence ω_R to be positive, real-valued quantities.

The total Hamiltonian follows simply by including the electronic and kinetic energy terms:

$$\mathcal{H}_0 = \hbar\begin{pmatrix} \omega_{\text{rec}} & 0 & 0 \\ 0 & \omega_0 & 0 \\ 0 & 0 & \omega_{\text{rec}} \end{pmatrix} \qquad [24]$$

where the single-photon recoil energy, E_{rec}, of an atom of mass m is given by

$$E_{\text{rec}} = \hbar\omega_{\text{rec}} = \frac{\hbar^2 k^2}{2m} \qquad [25]$$

Making the *ansatz* for the solution wavefunction as

$$|\psi\rangle = \begin{pmatrix} c_{-1}(t)e^{-i\omega_{\text{rec}}t} \\ c_0(t)e^{-i\omega_0 t} \\ c_{+1}(t)e^{-i\omega_{\text{rec}}t} \end{pmatrix} \qquad [26]$$

and substituting into the Schrödinger equation yields three coupled first-order differential equations:

$$\dot{c}_{\pm 1}(t) = \mp\frac{\omega_R}{2}e^{i\Delta t}c_0(t), \qquad [27]$$

$$\dot{c}_0(t) = \frac{\omega_R}{2}e^{-i\Delta t}\big(c_{+1}(t) - c_{-1}(t)\big), \qquad [28]$$

where $\Delta = \delta + \omega_{\text{rec}} = (\omega - \omega_0) + \omega_{\text{rec}}$. Differentiating eqn [27] and substituting eqn [28] into the result yields two coupled second-order differential equations:

$$\ddot{c}_{\pm 1}(t) - i\Delta\dot{c}_{\pm 1}(t) + \frac{\omega_R^2}{4}\big(c_{\pm 1}(t) - c_{\mp 1}(t)\big) = 0 \qquad [29]$$

With the initial conditions:

$$c_{-1}(0) = 1 \qquad [30]$$

$$c_0(0) = 0 \Rightarrow \dot{c}_{\pm 1}(0) = 0 \qquad [31]$$

$$c_{+1}(0) = 0 \qquad [32]$$

and the assumption $\Delta^2 \gg \omega_R^2$, the solutions to eqn [29] are

$$c_{-1}(t) = e^{-i\frac{\omega_R^{(2)}}{2}t}\cos\left(\frac{\omega_R^{(2)}}{2}t\right) \qquad [33]$$

$$c_{+1}(t) = ie^{-i\frac{\omega_R^{(2)}}{2}t}\sin\left(\frac{\omega_R^{(2)}}{2}t\right) \qquad [34]$$

where the two-photon Rabi frequency is

$$\omega_R^{(2)} = \frac{\omega_R^2}{2\Delta} \rightarrow \frac{\omega_R^2}{2\delta}, \quad |\delta| \gg \omega_{\text{rec}} \qquad [35]$$

with both transitions driven at equal single-photon Rabi frequencies, ω_R.

Substituting the solution of eqn [33] or [34] into eqn [27] yields an expression for the excited state amplitude:

$$c_0(t) = -i\frac{\omega_R}{2\Delta}e^{-i\Delta t}e^{-i\omega_R^{(2)}t} \qquad [36]$$

This will be important in calculating the rate of spontaneous emission events later.

The solutions for $c_{-1}(t)$ and $c_{+1}(t)$ oscillate with the interaction duration, τ, yielding the result for the $|g,-\hbar k\rangle \rightarrow |g,+\hbar k\rangle$ transition probability:

$$P_1^B(\tau) = |c_{+1}(\tau)|^2 = \sin^2\left(\frac{\omega_R^{(2)}}{2}\tau\right) \qquad [37]$$

Thus, the system oscillates between the two momentum states $|g,-\hbar k\rangle$ and $|g,+\hbar k\rangle$ in a manner analogous to the Rabi oscillation of atomic population between two resonantly coupled states. This solution with oscillatory probabilities for the two Bragg coupled states is known as the Pendellösung and has been observed for atoms, neutrons, and X-rays (Figure 8). The nice feature here is that the strength of the grating can be actually controlled by the intensity of the light.

Viewing Bragg scattering as a two-photon transition from the initial ground state to the final ground state with opposite momentum, illuminates the close connection with a Raman transition between two internal sub-states of the ground state manifold, each with its own external momentum state. The formalism describing the Raman transition is basically the same as that presented here, except the two transitions can be driven at different single-photon Rabi frequencies, ω_{R1} and ω_{R2}, so that the generic two-photon Rabi frequency is given by $\omega_R^{(2)} = \omega_{R1}\omega_{R2}/2\Delta$, where Δ is the detuning from the intermediate state.

An N_B^{th} order Bragg process (similar to a $2N_B$-photon Raman process) is a coherent succession of N_B two-photon transitions from $|g, -N_B\hbar k\rangle$ to $|g, +N_B\hbar k\rangle$ with $2N_B - 1$ intermediate states of the form:

$$|e, (-N_B + 1)\hbar k\rangle, |g, (-N_B + 2)\hbar k\rangle \ldots$$

$$\ldots |g, (N_B - 2)\hbar k\rangle, |e, (N_B - 1)\hbar k\rangle$$

Such a process is characterized by a $2N_B$-photon Rabi frequency given by

$$\omega_R^{(2N_B)} = \frac{[\omega_R]^{2N_B}}{2^{2N_B-1}\Delta_1\Delta_2\cdots\Delta_{2N_B-1}} \qquad [38]$$

where Δ_n is the detuning from the nth intermediate state. Figure 9 shows what this process would look like for an N_B^{th} order Bragg transition where the intermediate state detunings are given by

$$\Delta_n = \left\{ \begin{array}{ll} \delta + (2N_B n - n^2)\omega_{\text{rec}} : & n \text{ odd} \\ (2N_B n - n^2)\omega_{\text{rec}} : & n \text{ even} \end{array} \right\} \qquad [39]$$

Substituting these detunings into eqn [38] yields the N_B^{th} order Bragg transition $2N_B$-photon Rabi frequency, $\omega_R^{(2N_B)}$:

$$\omega_R^{(2N_B)} = \frac{[\omega_R]^{2N_B}}{2^{4N_B-3}[(N_B - 1)!]^2 \delta^{N_B} \omega_{\text{rec}}^{N_B-1}} \qquad [40]$$

where we have assumed $|\delta| \gg N_B^2 \omega_{\text{rec}}$.

To ensure that the system truly undergoes Bragg scattering, and validate the assumption that only states of equal kinetic energy and opposite

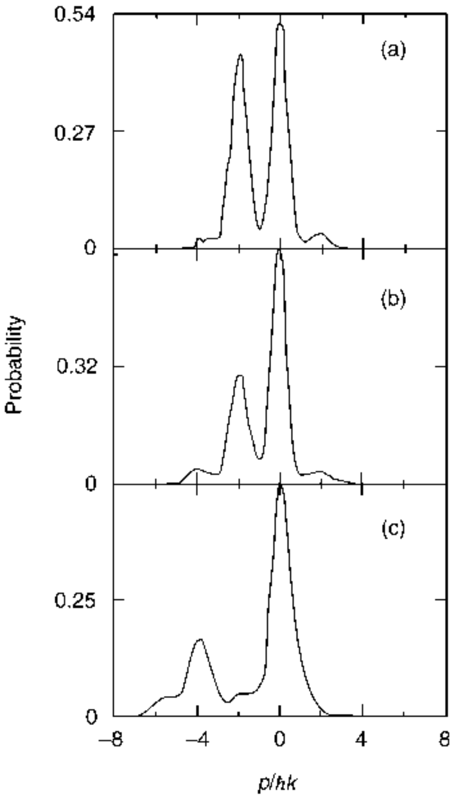

Figure 8 Bragg scattering of an atom beam from an optical standing wave. First-order Bragg scattering at (a) lower power, (b) higher power of the laser beams (Pendellösung has set in, increasing the amplitude of the undeflected peak), (c) second-order Bragg scattering. Reprinted with permission from *Physical Review Letters* 60(6): 515–518. Copyright (1998) by the American Physical Society.

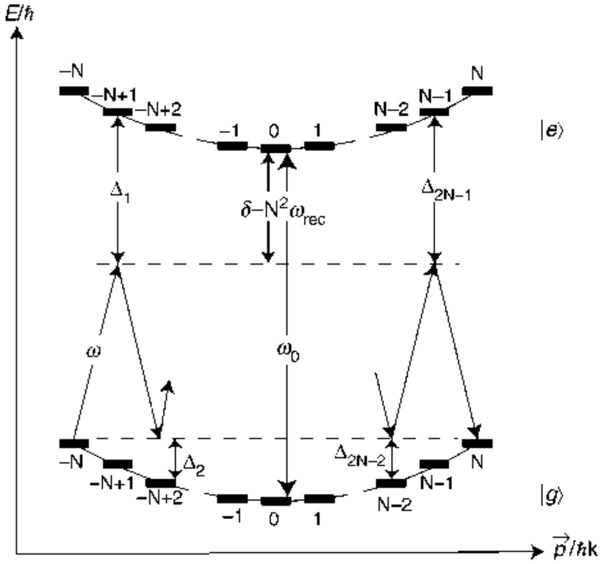

Figure 9 Energy vs. momentum diagram describing atomic states involved in higher-order Bragg diffraction. Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

momentum are coupled, the overall exposure time, τ, of the atoms to the fields must be limited both from below and above.

The lower bound on τ is necessary to resolve the final momentum state ($|g, N_B\hbar k\rangle$) from neighboring momentum states ($|g, (N_B \pm 2)\hbar k\rangle$) two photon recoil momenta away. Avoiding population transfer into these states requires that the energy separation be resolvable, or

$$\tau(\delta E) \gg h \qquad [41]$$

For first-order Bragg scattering processes ($N_B = 1$) the nearest lying momentum states that may be mistakenly populated are $|g, \pm 3\hbar k\rangle$, for which

$$\delta E = \frac{p_{n=3}^2}{2m} - \frac{p_{n=1}^2}{2m} = 8\hbar\omega_{\text{rec}} \qquad [42]$$

Therefore

$$\tau \gg \frac{h}{8\hbar\omega_{\text{rec}}} = \frac{\pi}{4\omega_{\text{rec}}} \qquad [43]$$

For sodium atoms (^{23}Na) in 590 nm light, $\tau \gg 6\ \mu$s is the lower bound required for first-order Bragg scattering.

For all higher-order ($N_B > 1$) Bragg scattering processes, the nearest momentum states are $|g, \pm(N_B - 2)\hbar k\rangle$, which limits the interaction time to

$$\tau \gg \frac{\pi}{2(N_B - 1)\omega_{\text{rec}}} \qquad [44]$$

which for $N_B \gg 1$ reduces to

$$\tau \gg \frac{\pi}{2N_B\omega_{\text{rec}}} \qquad [45]$$

The upper bound on the interaction duration is necessary to avoid spontaneous emission. The interaction duration must be short enough so that the expected number, N_s, of spontaneous emission events per atom during the time τ is negligible. N_s is simply given by the product of the excited state fraction (eqn [36]) and the probability of spontaneous decay given that the atom is in the excited state:

$$N_s = |c_0(t)|^2\Gamma\tau = \frac{\omega_R^2}{4\Delta^2}\Gamma\tau \qquad [46]$$

where Γ is the natural decay rate of the excited state. Avoiding spontaneous emission ($N_s \ll 1$) while still having a significant probability for transitions ($\omega_R^{(2N_B)}\tau \simeq \pi$) forces one to work in the regime where $\Delta \gg \Gamma$, which is a practical requirement to avoid spontaneous emission and maintain coherence. For sodium atoms on the 590 nm transition, $\Gamma = 6.3 \times 10^7\ \text{s}^{-1}$, and a typical choice in Bragg diffraction experiments is $\Delta = 50\Gamma$, or $\Delta/2\pi = 500$ MHz.

Briefly, to relate Rabi frequency to optical intensity, the transition rate W on resonance with a Lorentzian spectral line is given by Fermi's golden rule:

$$W = \frac{\omega_R^2}{\Gamma} = \frac{\lambda_{\text{photon}}^2}{2\pi}\frac{I}{\hbar\omega} \qquad [47]$$

where the last term can be regarded as the cross section for absorption multiplied by intensity in units of photon flux. Hence

$$\omega_R^2 = \Gamma\frac{\lambda_{\text{photon}}^2}{2\pi}\frac{I}{\hbar\omega} \qquad [48]$$

For Na atoms in 12 mW(cm)$^{-2}$ resonant light, $\omega_R = \Gamma = 6.3 \times 10^7\ \text{s}^{-1}$. Both the probability of Bragg scattering and the spontaneous emission increase linearly with $\omega_R^2\tau\Gamma^{-1}$. However, spontaneous emission decreases as Δ^2 while the Bragg scattering probability decreases as Δ. Thus, as shown in Figure 10, at sufficient detuning and laser power, Bragg scattering occurs with no spontaneous emission. For reference, 12 mW(cm)$^{-2}$ × 10 µs corresponds to $\omega_R^2\tau\Gamma^{-1} = 630$ for sodium atoms.

Bragg scattering of atoms from a standing light wave was first observed at MIT in 1988. A supersonic atomic beam was diffracted from a standing wave of near-resonant laser light as depicted in Figure 6. The angle between the atomic beam (of thermal wavelength λ_{dB}) and the light grating (of periodicity $\lambda_L/2$) was tuned to the appropriate Bragg angle, θ_B, where:

$$\lambda_{dB} = \lambda_L \sin(\theta_B) \qquad [49]$$

Population transfer corresponding to both first- and second-order Bragg scattering was observed (Figure 7). The experiment required a sub-recoil transverse momentum spread of the atomic beam in order to resolve the different momentum states in the far field and limit the final state to only one

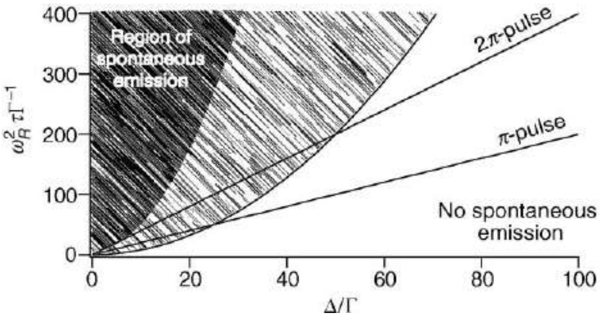

Figure 10 The conditions for which $N_s > 0.02(0.10)$ are hashed in light(dark) gray. ω_R^2/Γ is proportional to optical intensity. The line labeled π-pulse indicates the parameters which cause complete first-order Bragg diffraction. The line labeled 2π-pulse indicates conditions where population is completely transferred back to the initial state (Pendellösung oscillation).

diffracted order. The Pendellösung was observed as an oscillation in population transfer as a function of standing wave intensity, $I \propto \omega_R^{(2)}$, for a fixed interaction time, τ.

Atomic beam diffraction from an optical standing wave is a continuous-wave (CW) experiment in which the selectivity needed for the Bragg process is imposed by good angular resolution of the particle beam and a high degree of parallelism between the light crystal planes. This ensures that of the various final Bragg orders allowed by momentum conservation, only one conserves energy (energy conservation is exact in a CW experiment). For atoms scattering from a light crystal, parallelism of the crystal planes requires highly parallel photon momentum that implies a minimum width of the standing wave (the diffraction limit for the collimated photons). The transit time, τ, of the atoms across this width then exceeds the lower bound given in eqn [44].

The excellent collimation required of the atomic beam to ensure resolution of the Bragg scattered atoms reduces the intensity of the source by many orders of magnitude. A Bose–Einstein Condensate (BEC) is an attractive alternative source of atoms because its momentum spread is typically an order of magnitude below a single-photon recoil momentum. To Bragg diffract atoms initially in a stationary BEC, it is easier to move the light crystal than to accelerate the condensate. This is done by simply frequency shifting one of the traveling waves so that the resultant standing wave formed by its interference with the unshifted traveling wave moves with the proper velocity ($N_B\hbar k/m$) relative to the stationary atoms to impart the necessary momentum. The Bragg scattered atoms will then have momentum $2N_B\hbar k$ in the laboratory frame. The resonance condition thus becomes a condition on relative detuning, δ_{N_B}, between the two laser beams forming the diffraction grating. For N_B^{th}-order Bragg diffraction, the relative detuning is given by

$$\delta_{N_B} = \frac{2N_B\hbar k^2}{m} = 4N_B\omega_{\text{rec}} \qquad [50]$$

The first demonstration of Bragg scattering in a BEC was at NIST in 1999. They used Bragg scattering mainly as a tool to manipulate the momentum of the BEC, observing up to sixth-order processes. At MIT the interaction time was lengthened (to ≈100 times the lower bound of eqn [44] for first-order Bragg scattering), creating a new type of spectroscopy called Bragg Spectroscopy. It is a spectroscopic measurement of the shift of the Bragg scattering frequency due to the Doppler shift ($k \cdot v$) from the atoms' motion together with any mean field interaction. (These effects can be separated by going to higher-order Bragg scattering to enhance the Doppler shift.)

Bragg Spectroscopy was used to observe the momentum distribution of a BEC in a magnetic trap. The width of the Bragg resonance curve was primarily due to a 2 kHz Doppler-broadening that yielded the momentum distribution of the condensate. The spread in the corresponding velocity distribution was very small ($\approx 0.5\ \text{mm s}^{-1}$), even smaller than allowed by the Heisenberg uncertainty limit for a particle confined in the ground state of a harmonic trap. This reflects the increase in size of the condensate due to the mean field repulsion (^{23}Na, used in the MIT experiment, has a positive scattering length). The distribution was Heisenberg uncertainty limited at the observed size of the BEC, establishing for the first time that the coherence length of the condensate was equal to its size. In addition, the narrow Bragg resonance was shifted by the repulsive interactions within the condensate, resulting in a spectroscopic measurement of the mean-field energy.

Kapitza–Dirac Scattering

Diffraction of neutral atoms from a standing wave of near-resonant light with a short interaction time ($\tau_{KD} \ll 1/\omega_{\text{rec}}$) has come to be called Kapitza–Dirac scattering, in honor of their pioneering suggestion. In 1933, Kapitza and Dirac predicted that an electron beam propagating in a standing light wave would undergo stimulated Compton scattering and be reflected. This process has a tiny cross-section, equal to the classical electron radius squared: $\sigma_{\text{Compton}} = 8\pi/3(e^2/mc^2)^2 \approx 6 \times 10^{-25}\ \text{cm}^2$, and has only recently been observed using extremely high laser intensities. If the electrons are replaced by atoms, however, the scattering cross-section for resonant light $\sigma_{\text{atom}} = (1/2\pi)\lambda_{\text{photon}}^2 \approx 4 \times 10^{-10}\ \text{cm}^2$ is 15 orders of magnitude larger. This large cross-section, together with the ready availability of tunable lasers, has allowed stimulated scattering, both in the Kapitza–Dirac and Bragg regimes, to become the primary tool for the coherent manipulation of atoms.

In Kapitza–Dirac scattering, atomic motion during the interaction time is small compared to the characteristic dimensions of the interaction potential. This is equivalent to the eikonal approximation for scattering or the thin-lens approximation in optics. The idea is that the phase of the incident particle changes along each classical trajectory, but not the amplitude. Thus, there may be momentum transfer perpendicular to the trajectory, but the trajectory is not significantly displaced (until after the interaction is over). Mathematically, this regime can be treated by neglecting the atomic kinetic energy term in the Hamiltonian (the Raman–Nath approximation).

The standing wave interaction may be treated by considering the standing wave (ac-Stark shift) potential resulting from the applied fields given in eqn [18]:

$$V(z,t) = \frac{\hbar\omega_R^2}{\delta} f^2(t)\sin^2(kz) \qquad [51]$$

where we have assumed $\delta \gg \Gamma$, because the constraint $N_s \ll 1$ (eqn [46]) holds in this regime as well, where N_s is the expected number of spontaneous emission events per atom during the time τ. Although we neglect the kinetic energy term in the Hamiltonian, we continue to will treat z in eqn [51] as an operator.

Given the initial atomic wavefunction in momentum space, $\psi(p_0)$, the atomic wavefunction immediately after the interaction is given by

$$\psi(p) = e^{-\frac{i}{\hbar}\int^{t} dt' U(z,t')}\psi(p_0) \qquad [52]$$

$$= e^{-i\frac{\omega_R^2}{2\delta}\tau} e^{i\frac{\omega_R^2}{2\delta}\tau\cos(2kz)}\psi(p_0) \qquad [53]$$

where $\tau = \int dt' f^2(t')$ and the integral is over the interaction duration. With the use of the identity for Bessel functions of the first kind, $e^{i\alpha\cos(\beta)} = \sum_{n=-\infty}^{\infty} i^n J_n(\alpha)e^{in\beta}$, the atomic wavefunction can be written as

$$\psi(p) = e^{-i\frac{\omega_R^2}{2\delta}\tau}\sum_{n=-\infty}^{\infty} i^n J_n\left(\frac{\omega_R^2}{2\delta}\tau\right)e^{i2nkz}\psi(p_0) \qquad [54]$$

$$= e^{-i\frac{\omega_R^2}{2\delta}\tau}\sum_{n=-\infty}^{\infty} i^n J_n\left(\frac{\omega_R^2}{2\delta}\tau\right)\psi(p_0 + 2n\hbar k) \qquad [55]$$

States with $2N\hbar k$ of momentum are populated with the probability:

$$P_N = J_N^2(\theta), \quad N = 0, \pm 1, \pm 2, \ldots \qquad [56]$$

where

$$\theta = \frac{\omega_R^2}{2\delta}\tau = \omega_R^{(2)}\tau \qquad [57]$$

is the pulse area. This leads to a transverse rms momentum of the diffracted atoms that is linearly proportional to the pulse area:

$$p_{\mathrm{rms}} = \sum_{n=-\infty}^{\infty}(n\hbar k)^2 P_n = 2^{1/2}\theta\hbar k \qquad [58]$$

The maximum and minimum bounds on the interaction time, τ, and the amount of momentum transfer to the atoms are discussed in the next section.

Kapitza–Dirac diffraction of atoms was first observed at MIT in 1986. Diffraction of a well-collimated (subrecoil) supersonic atomic beam was observed after passage through the tightly focused waist of a near-resonant standing wave (Figure 11). Significant diffraction into momentum states $|g, \pm 10\hbar k\rangle$ was observed (Figure 12).

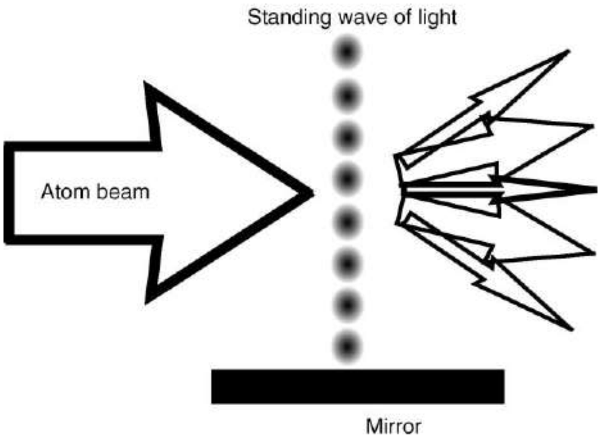

Figure 11 Two counter-propagating running waves of light superimpose to form a standing wave of light. The more narrow beam shown here represents the configuration for Kapitza–Dirac diffraction.

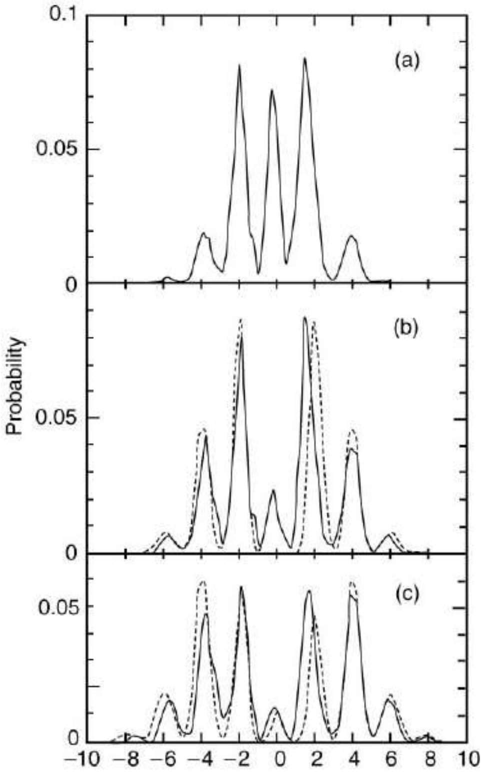

Figure 12 Kapitza–Dirac diffraction of an atomic beam from a standing wavelight grating. The solid lines are experimental data, the dashed lines are theoretical curves. (a) $\theta = 1.69$, (b) $\theta = 2.33$, (c) $\theta = 2.84$. Reprinted with permission from *Physical Review Letters* 56(8): 827–830. Copyright (1986) by the American Physical Society.

Even higher diffracted orders should be observable in the future using laser beams directed at small BECs for somewhat longer times.

Comparison of Atom-Standing Wave Interactions

The atomic motion induced by interaction with a standing light wave varies qualitatively with the interaction parameters. Figure 13 summarizes the various interaction regimes of an atom with a standing wave in terms of the two most basic parameters: the two-photon Rabi frequency, $\omega_R^{(2)} = \omega_R^2/2\delta$, and the duration of the interaction, τ. To render the plot independent of atomic species, $\omega_R^{(2)}$ is given in units of the single-photon recoil frequency, ω_{rec}, and τ is given in units of the inverse single-photon recoil frequency, ω_{rec}^{-1}. Thus, for ^{23}Na the point (1,1) on Figure 13 corresponds to (6.4 μs, $2\pi \times 25$ kHz). The product of the two-photon Rabi frequency and the duration of the interaction is the pulse area, $\theta = \omega_R^{(2)}\tau$ (eqn [57]), which is likewise the product of the pair of coordinates forming a point on the plot.

The lines KD1(KD10) on Figure 13 show where the first maxima of $P_{1,10} = J_{1,10}^2(\theta)$ (eqn [56]) occur, corresponding to the maximum possible population transfer into the first and tenth Kapitza–Dirac diffracted order respectively. Since the momentum distribution of diffracted atoms depends only on the pulse area (θ), all Kapitza–Dirac orders are parallel to each other and further offset from the origin in increasing order number.

The Kapitza–Dirac regime ends at large interaction times where the Raman–Nath approximation fails due to motion of the atoms down the slope of the standing wave potential. We show these lines dashed as the interaction time approaches the beginning of the classical oscillation regime, and terminate these lines where the oscillation produces its first focus, at the first focus line. This line corresponds to the atoms completing a quarter period of oscillation in the standing wave potential, $\tau = \tau_{\text{osc}}/4$, where the oscillation time is derived from the approximation that the potential $\langle V(z,t)\rangle_t$ (eqn [51]) is harmonic

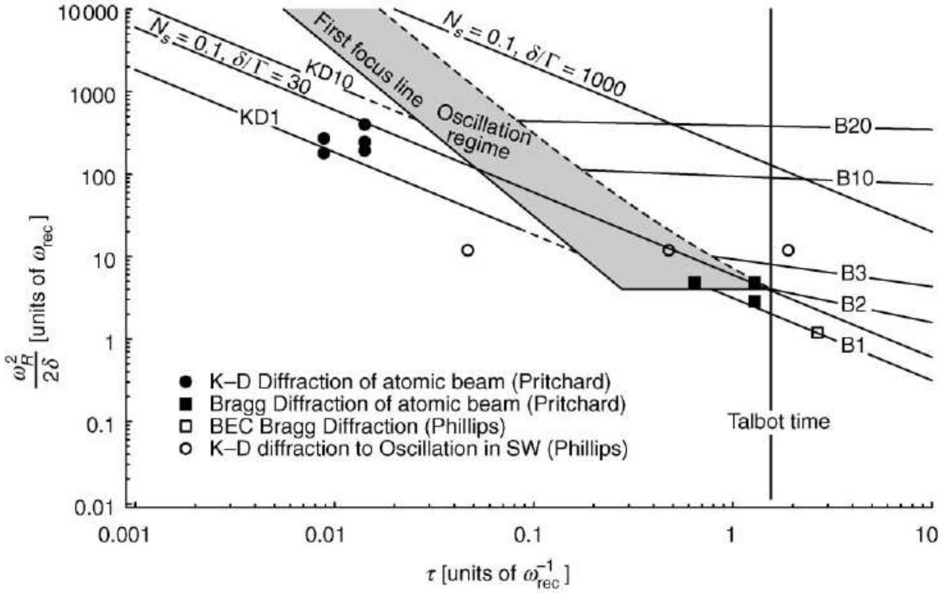

Figure 13 Atomic diffraction from a standing light wave. The vertical axis is the two-photon Rabi frequency, $\omega_R^2/2\delta$, in units of the single-photon recoil frequency, ω_{rec}, and the horizontal axis is the pulse duration, τ, in units of the inverse single-photon recoil frequency, ω_{rec}^{-1}. The scaling is chosen to eliminate the atomic species dependence of the plot. All coherent momentum transfer processes are destroyed by spontaneous decay, which occurs with probability $N_s = 0.1$ along the lines labeled accordingly and parameterized by the given ratio δ/Γ. Lines KD1 and KD10 show conditions for the maximum transfer into the first and tenth Kapitza–Dirac diffracted order respectively. As the interaction time is increased the Raman–Nath approximation is violated (termination of Kapitza–Dirac curves) and the atoms enter the oscillatory regime, executing at least a quarter period of oscillation above the first focus line (shaded area). Cures B1, B2, B3, B10, and B20 correspond to conditions that generate complete Bragg reflection in the first-, second-, third-, tenth-, and twentieth-order respectively. Experimental conditions are shown as points. Filled circles: Kapitza–Dirac diffraction of an atomic beam (Pritchard 1986). Filled squares: Bragg diffraction of an atomic beam (Pritchard 1988). Open squares: Bragg diffraction of a BEC (Phillips 1999). Open circles: transition from Kapitza-Dirac diffraction to oscillation of a BEC in a standing wave light pulse (Phillips 1999). Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

about the minimum:

$$\tau_{\text{osc}} = \frac{\pi}{\omega_R}\left(\frac{|\delta|}{\omega_{\text{rec}}}\right)^{1/2} \qquad [59]$$

In the shaded region of Figure 13, the atoms will oscillate classically about the potential minimum, causing a periodic focusing of the atoms alternately in position and momentum space. However, anharmonicities of the potential away from its minimum will degrade the quality of the focusing effects.

The maximum diffracted order that can be significantly populated by the light–atom interaction is limited by energy conservation. Classically, the maximum momentum transfer to the atoms due to the sudden switch on of the standing wave is delivered to atoms that convert the full height of the standing wave potential ($\hbar\omega_R^2/\delta$), into kinetic energy. This gives

$$\frac{p_{\max}^2}{2m} \simeq \frac{\hbar\omega_R^2}{\delta} \Rightarrow p_{\max} \simeq \omega_R\left(\frac{2\hbar m}{\delta}\right)^{1/2} \qquad [60]$$

Equating the maximum absorbed momentum, $p_{\max}$, to an integer number of two-photon recoil momenta yields the maximum expected diffracted order, $N_{\max}$:

$$p_{\max} = 2N_{\max}\hbar k \Rightarrow N_{\max} \simeq \frac{\omega_R}{2}(\omega_{\text{rec}}\delta)^{-1/2} \qquad [61]$$

This is just half of the square-root of the standing wave potential height measured in units of the single-photon recoil energy:

$$N_{\max} = \frac{1}{2}\left(\frac{\frac{\omega_R^2}{\delta}}{\omega_{\text{rec}}}\right)^{1/2} = \frac{1}{2}\left(\frac{\omega_R^{(2)}}{\omega_{\text{rec}}}\right)^{1/2} \qquad [62]$$

As the two-photon Rabi frequency is reduced, $N_{\max}$ falls below unity. As a result, there is no longer time for the higher-order multi-photon processes to generate significant amplitudes in diffracted orders with $N > 1$ before dephasing due to the kinetic energy term (eqn [24]) becomes significant and the higher-order processes become negligible. Only a small population is ever transferred to the $N = \pm 1$ orders and it does not oscillate at ω_{rec}. Therefore the classical oscillation regime does not extend below $N_{\max} \simeq 1$ where it is terminated on Figure 13.

Classical oscillation would result in atoms with momentum from zero up to the maximum allowed by energy conservation. Therefore, the classical focusing of atoms must be manifest as an atomic population distribution over many of the quantized momentum states (separated by $2\hbar k$) allowed by energy conservation (eqn [61]). However, as the interaction time lengthens and extends into the Bragg regime, the populated momentum states become restricted by energy conservation until only the oscillatory Pendellösung into and back from only one final state remains. Therefore, we have ended the classical oscillation regime where the Bragg condition (eqn [44]) is satisfied. The Bragg regime presupposes a smooth light pulse shape. For a pulse with sharp edges, classically oscillatory behavior can still be observed at longer times than included in the shaded region of Figure 13, which is why we show the large τ boundary of the classical oscillation regime as dashed.

In the Bragg regime, transfer of population is restricted to (and back from) only one final momentum state. The allowable final states are restricted by limiting the frequency bandwidth (i.e., energy uncertainty) of the light fields in the atomic rest frame. This is accomplished by lengthening the interaction time and smoothing the rise and fall of the electromagnetic fields. The parameters for a first-order Bragg transition (Table 1) are typically $\delta/\Gamma \simeq 50$, $\Gamma\tau \simeq 1$, and $\omega_R/\Gamma \simeq 10$ giving $N_s \simeq 10^{-2}$ and $\omega_R^{(2)}\tau \simeq 1$. Obtaining significant population transfer with higher-order Bragg processes requires larger intensities. Various orders (1, 2, 3, 10, and 20) of Bragg diffraction are shown as lines on Figure 13 corresponding to $\omega_R^{(2N_B)}\tau = \pi$, where $\omega_R^{(2N_B)}$ is given in eqn [40]. The lines are terminated at the appropriate interaction time determined from the final momentum state resolution condition (eqn [44]). The Bragg regime extends indefinitely to larger interaction times, which might be termed the region of Bragg spectroscopy. In experiments in this regime with ^{23}Na, atomic velocity resolution below 1 mm s^{-1} was obtained at $\tau = 80\omega_{\text{rec}}^{-1}$. The study of adiabatically expanded BECs would require larger interaction times to resolve their smaller velocity spread and weaker mean field shifts.

The Kaptiza–Dirac and Bragg regimes assume a different initial atomic momentum (in the rest frame of the standing wave) parallel to the standing

Table 1 Natural parameters and typical experimental parameters involved in standing wave diffraction. The relevant frequencies together with the corresponding times ($1/\omega$) are tabulated. The parameters for ^{23}Na have been used for system dependent quantities

ω	$\omega/2\pi$	$1/\omega$
ω_0	500 THz	0.3 fs
δ (Bragg)	500 MHz	0.3 ns
ω_R (Bragg)	100 MHz	1.6 ns
ω_R (K–D)	100 MHz	1.6 ns
Γ	10 MHz	16 ns
ω_{rec}	25 kHz	6.4 μs
ω_{osc}	1 MHz	160 ns

wave axis. Kapitza–Dirac scattering assumes no component of the initial atomic momentum along this axis. The efficiency of Kapitza–Dirac diffraction falls rapidly as the initial momentum in units of the photon momentum approaches $1/\omega_{\mathrm{rec}}\tau$. To observe N_B^{th}-order Bragg scattering, the initial atomic momentum along the standing wave axis must have (nonzero) magnitude $N_B\hbar k$. Without adhering to this constraint, no final momentum state will be energetically degenerate with the initial state and the atomic sample will not respond to the presence of the standing wave, even if the interaction parameters are appropriate ($\omega_R^{(2N_B)}\tau \simeq 1$) for N_B^{th}-order Bragg scattering.

The condition $N_s = 0.1$ (eqn [46]), where N_s is the expected number of spontaneous emission events per atom during the light–atom interaction, is drawn on Figure 13 for two ratios (30 and 1000) of standing wave detuning, δ, to the atomic excited state natural lifetime, Γ. For fixed δ/Γ, we see that $N_s \propto \theta$, thus conditions with an increased pulse area (above and to the right of the $N_s = 0.1$ lines) result in a proportionally increased N_s. These $N_s = 0.1$ lines are extended across all regimes of Figure 13 because restricting spontaneous emission is required for both Kapitza–Dirac and Bragg scattering. As N_s approaches unity, the correct optical potential describing the light–atom interaction no longer is eqn [51], and scattering of atoms becomes an incoherent process which can lead to new possibilities such as a complex valued optical potential. Thus, while the standing wave fields in pure Bragg or Kapitza–Dirac scattering behaved as phase gratings for atoms, light gratings on resonance can effectively become amplitude diffraction gratings.

Atom Interferometry

In the 19th century, Fizeau (1853), Michelson (1881), Rayleigh (1881), and Fabry and Perot (Fabry 1899) exploited the interference properties of light waves to create the light interferometer which has since resulted in many beautiful experiments and precise measurements. Using technologies invented since the Second World War, the initial ideas of de Broglie and Schrödinger have evolved into construction of interferometers for neutrons, electrons, and atoms. These new interferometers are proving to be valuable tools for probing fundamental physics, for studying quantum mechanical phenomena, and for making inertial measurements.

The scientific value of interferometry with atoms and molecules has long been recognized. In fact, the concept of an atom interferometer was patented in 1973 and it has been extensively discussed since. Atom interferometry offers great richness stemming from the varied internal structure of atoms, the wide range of properties possessed by different atoms (e.g., mass, magnetic moment, absorption frequencies, and polarizability), and the great variety of interactions between atoms and their environment (e.g., static E–M fields, radiation, and other atoms).

Light interferometers are generally based either on achromatic beamsplitters such as half-silvered mirrors or on other semi-transparent membranes whose structure is small compared to the wavelength of the wave they are splitting. Lacking material structures that are either transparent to atoms or smaller than their de Broglie wavelength, diffraction gratings have been pressed into service both as beamsplitters and mirrors for atom waves. This means that atom interferometers are constrained to designs which somehow compensate for the dependence of diffraction angle on the wavelength of the individual atoms. In spite of this challenge, a surprising variety of atom and molecule interferometers have been built since 1991. A majority have used the three-grating configuration in which the first grating splits the incident beam, the second reverses the differential momenta given by the first, and the third recombines the two beams at the location where they overlap. Both material gratings and standing waves of light have been used in the Raman–Nath, Bragg, and adiabatic regimes to obtain interference fringes. Some designs render the interference fringes in position space, others in internal state space.

Figure 14 shows the MIT setup for a three-grating atom interferometer which features sufficient separation between interfering paths to accommodate an interposed metal foil. This allows different fields or media to act on atom waves only in one of the two

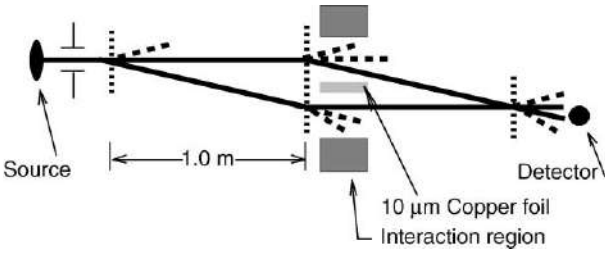

Figure 14 An interferometer for atoms built with three nanofabricated gratings. Each atom propagates in a superposition of two paths, shown in bold. The beam of atoms is well enough collimated, and the diffraction angle is large enough, that a metal barrier can be inserted between the two paths inside the interferometer. The interference fringes are observed by translating one grating transverse to the atom beam. See for example: Berman PR (ed.) (1997) *Atom Interferometry*. San Diego, CA: Academic Press.

arms of the interferometer. Such an interaction region has been used to measure phase shifts due to an electric field applied to only one of the interfering paths. The phase difference between the two paths is then given by the Feynman path integral:

$$\phi = \frac{-i}{\hbar}\int V(x)dt \qquad [63]$$

where

$$V = -\frac{1}{2}\alpha E^2 \qquad [64]$$

is the potential energy due to the electric filed E, and α is the electric polarizability of the atoms. The resulting phase shift of the interference fringes (Figure 15) has been used to measure the polarizability of sodium atoms to unprecedented accuracy.

Phase shifts have also been measured due a dilute target gas in one arm of the interferometer. In this case there is a complex index of refraction for the forward scattered atoms (some attenuation, and some phase shift) which depends on matter wave wavelength. Recent measurements reveal glory oscillations in the matter wave index of refraction due to passing through a dilute gas, and are a sensitive probe of interatomic potentials.

In the domain of inertial sensing, atom interferometers have been used to the local gravitational acceleration, $\vec{g}$, Newton's constant, G, gravity gradients, $\nabla\vec{g}$, and rotations Ω, each with sensitivity rivaling if not exceeding any other method. The Sagnac phase shift, due to rotations, for a two-path interferometer is

$$\phi_S = \frac{4\pi}{\lambda_{dB} v}\vec{\Omega}\cdot\vec{A} \qquad [65]$$

where v is the velocity of atoms, $\vec{A}$ is the enclosed area of the interferometer, and $\vec{\Omega}$ is the rotation rate. The two factors in the denominator, λ_{dB} and v, are both $\sim 10^5$ times larger for light than for atoms, which is why atom interferometers have the potential to be 10^{11} times more accurate than optical interferometers. In practice, atom interferometers only outperform laser gyros by a small amount because they have dramatically smaller enclosed areas and particle flux than optical interferometers.

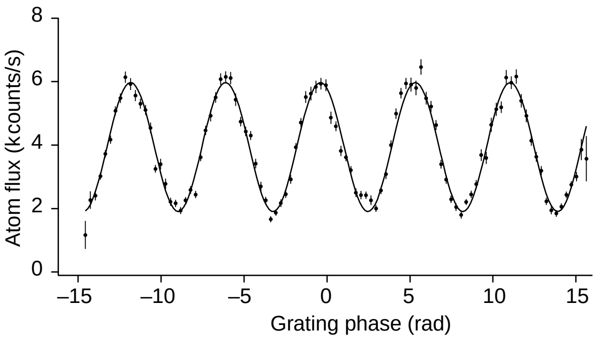

Figure 15 Interference fringes from atoms. 2π of grating phase corresponds to translating the grating one grating period, or 100 nm. These data were obtained at MIT.

Since the first atom interferometers for Na and He* began working in 1991, others have been made for Li, Ar*, Ca, Cs, K, Mg, Ne*, Rb, and Li_2, Na_2, I_2, C_{60} and C_{70} molecules, interferometers starting with trapped atoms have been made for Cs, Ca, He*, Mg, and Rb. Interferometers using Bose–Einstein condensates have been demonstrated with Na and Rb. These lists are still growing. Diffraction from a single grating has been observed for molecules as large as C_{70} buckyballs, which may someday be used in a three-grating interferometer.

Conclusion

The manipulation of atoms using light forces from standing light waves is a rich subject. The seminal suggestion of Kapitza and Dirac laid dormant for 50 years due to lack of experimental technology. However, in the 1980s this suggestion was realized with atomic sources and considerably extended both experimentally and theoretically. In the early 1990s, coherent standing wave manipulation and nonofabricated atom optics became the two major routes to making atom interferometers. As the new century begins, these techniques are being refined further with improved sources, new detection schemes, and better atom optics. Replacing thermal atomic beams by Bose–Einstein condensates (BECs), as sources of atoms, will revolutionize the field of atom optics just as lasers did in light optics.

Conversely, stimulated light scattering and nanostructures provide new ways to study atomic and molecular properties, such as Bragg spectroscopy for BEC characterization. As progress continues, exciting developments should be forthcoming, such as definitive measurements of the fine structure constant, α, atom gyroscopes that are far superior to the best laser gyroscope technology, and improved studies of BECs. Another area that is on the brink of spectacular development is the confinement of coherent matter waves in atom waveguides and the development of the scientific and technological opportunities that these represent, in analogy to fiber optic waveguides for light.

While we have described the application of nanofabrication techniques to atom optics, it is possible to imagine technology transfer in the other direction. The fundamental problem of fabricating ever smaller structures might be tremendously advanced by atom optics because the de Broglie wavelength of atoms is so much smaller than that of light, permitting them to be focused to directly deposit much smaller features than possible with photolithography.

See also

Interferometry: Overview. **Scattering:** Scattering Theory.

Further Reading

Cohen-Tannoudji C, Diu B and Laloe F (1977) *Quantum Mechanics*. New York: Wiley and Sons.

Hecht E (1987) *Optics*. Reading, MA: Addison-Wesley.

Berman PR (ed.) (1997) *Atom Interferometry*. San Diego, CA: Academic Press.

Mandel L and Wolf EM (1995) *Optical Coherence and Quantum Optics*. London: Cambridge University Press.

Balykin VI and Letokhov VS (1995) *Atom Optics with Laser Light*. Chur, Switzerland: Harwood Academic Publishers.

Meystre P and Sargent M III (1999) *Elements of Quantum Optics*. Heidelberg, Germany: Springer Verlag.

Rauch H (2000) *Neutron Interferometry*. Oxford, UK: Oxford University Press.

Meystre P (2001) *Atom Optics*. New York: Springer-Verlag.

Campargue R (ed.) (2001) *Atomic and Molecular Beams, the State of the Art 2000*. Berlin: Springer.

Atomic Coherence Effects

A Belyanin, G R Welch and M O Scully,
Texas A&M University, College Station, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

History and Perspective

Quantum optics and laser physics have entered a new era. This started with the invention of the laser and the associated theoretical developments such as the theory of photon correlations and statistics, and the quantum theory of the laser. Later, new and novel lasers such as the free electron laser and tunable dye laser fueled many new research directions. With the development of the micromaser, quantum optics became a beautiful new test bed for the study of nonequilibrium statistical physics of photons interacting with matter in a controlled environment. The realization that a deeper understanding was needed to understand the measurement process led to the utilization of the laser as a probe of gravity, and to insight into nonclassical processes such as quantum nondemolition measurements and the production of squeezed states of light.

Effects of atomic coherence in quantum electronics have a long history. One of the first examples was the Hanle effect: excitation by polarized light creates atomic coherence that can evolve in a magnetic field, causing the change of polarization of the resulting fluorescence. Another important example is the interference of decay processes, first suggested by Fano. Modern schemes utilizing quantum coherence in atomic vapors or solid media usually involve strong coupling laser fields (laser drives). In each case, the effect of quantum coherence can be understood as interference of multiple (usually two) transition paths between quantum states of a system.

In the past decade, the concepts of atomic coherence and interference have led to many new and surprising innovations in laser physics such as electromagnetically induced transparency (EIT), refractive index enhancement, lasers without population inversion, ultra-slow light group velocity, storage of the quantum states of light, teleportation of quantum states, and nonlinear optics at an ultra-low intensity level.

Electromagnetically Induced Transparency

One of the most vivid manifestations of quantum coherence effects is the reduction of absorption in a three-level system that is driven by a coherent laser field. This effect is known as electromagnetically induced transparency, or EIT. We consider in more detail the case when the probe and drive fields are in a so-called Λ-configuration (Figure 1a) named after the appearance of the level scheme. In the conventional usage, the two lower levels are coupled to the upper level with two lasers. One of the lasers may be strong and is called the driving or coupling laser, and the other is weak and is called the probe laser.

Consider a quasi-monochromatic laser field E_d (called the 'driving' field, or 'coupling' field) and probe field E_p interacting with a three-level medium of Figure 1a. This interaction is described by a density matrix ρ_{ij}, where i, j take the values a, b, c. Within the

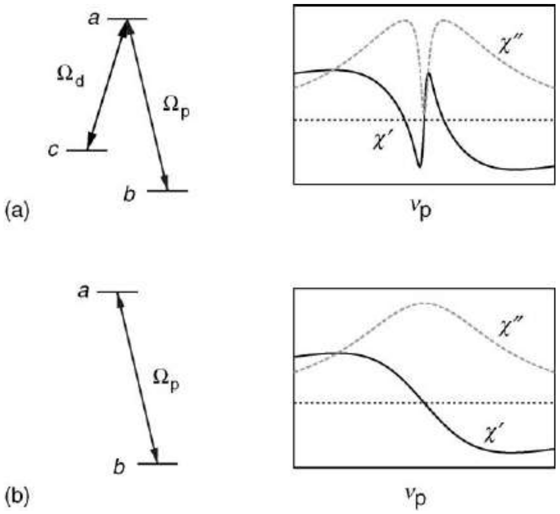

Figure 1 Real and imaginary parts of the susceptibility as functions of the frequency of the probe laser for (a) three-level Λ–type atoms, and (b) two-level atoms.

density matrix approach, quantum coherence in the system has the precise meaning of an off-diagonal element of the density matrix at the two-photon transition c–b. To have strong quantum coherence in a Λ-system, one must excite a large value of ρ_{cb}.

It is convenient to simplify the equations from the very beginning by using the rotating-wave approximation. We introduce slowly varying complex amplitudes of fields and polarizations and neglect their second derivatives. For example, for the probe field applied at the $a \to b$ transition we write

$$E_p(z,t) = \frac{1}{2}\varepsilon_p(z,t)\exp(-i\omega_p t + ik_p z) + \text{c.c.} \qquad [1]$$

After introducing the complex Rabi frequency $\Omega_p(z,t) = d_p\varepsilon_p(z,t)/2\hbar$, the wave equation for waves propagating in the z-direction can be written as

$$\frac{\partial\Omega_p}{\partial z} + \frac{n_p}{c}\frac{\partial\Omega_p}{\partial t} + \kappa_p\Omega_p = \frac{2\pi i\nu_p d_p^2 N}{\hbar c n_p^2}\sigma_{ab} \qquad [2]$$

Here d_p is the dipole moment of the probe transition a–b, N the total volume density of atoms in the active region, κ_p represents cavity losses, and $n_p = k_p c/\omega_p$ is the refraction index. The variable σ_{ab} is the slowly varying amplitude of the corresponding off-diagonal element $\rho_{ab} = \sigma_{ab}\exp(-i\nu_p t)$ of the density matrix. The same representation is assumed for the drive field E_d, with parameters n, d, ν, κ having index 'd'.

The functions σ_{ab}, σ_{ac}, and σ_{cb}, which determine the polarization response of the medium at the frequency of the external field, should be found from the density matrix equations

$$\frac{d\sigma_{cb}}{dt} + \Gamma_{cb}\sigma_{cb} = -i\Omega_p\sigma_{ac}^* + i\Omega_d^*\sigma_{ab} \qquad [3]$$

$$\frac{d\sigma_{ab}}{dt} + \Gamma_{ab}\sigma_{ab} = i\Omega_p n_{ba} + i\Omega_d\sigma_{cb} \qquad [4]$$

$$\frac{d\sigma_{ac}}{dt} + \Gamma_{ac}\sigma_{ac} = i\Omega_d n_{ca} + i\Omega_p\sigma_{cb}^* \qquad [5]$$

where

$$\begin{aligned}\Gamma_{cb} &= \gamma_{cb} + i(\nu_{cb} - \nu_p + \nu_d),\\ \Gamma_{ab} &= \gamma_{ab} + i(\nu_{ab} - \nu_p),\\ \Gamma_{ac} &= \gamma_{ac} + i(\nu_{ac} - \nu_d)\end{aligned} \qquad [6]$$

Here the γ's are relaxation rates of the polarizations at the corresponding transitions (the so-called homogeneous broadening). The population differences $n_{ik} = \rho_{ii} - \rho_{kk}$ are defined from eqns [2]–[6] together with the three equations for $\rho_{ii}, i = a, b, c$, with phenomenological rates of population relaxation and pumping:

$$\begin{aligned}\frac{d\rho_{bb}}{dt} &= -2\text{Im}[\Omega_p^*\sigma_{ab}] + r_{cb}\rho_{cc} + r_{ab}\rho_{aa} - r_{ba}\rho_{bb}\\ \frac{d\rho_{cc}}{dt} &= -2\text{Im}[\Omega_d^*\sigma_{ac}] + r_{ac}\rho_{aa} - r_{cb}\rho_{cc}\\ \frac{d\rho_{aa}}{dt} &= 2\text{Im}[\Omega_p^*\sigma_{ab}] + 2\text{Im}[\Omega_d^*\sigma_{ac}] + r_{ba}\rho_{bb}\\ &\quad - (r_{ab} + r_{ac})\rho_{aa}\end{aligned} \qquad [7]$$

Here we have included only the pumping term r_{ba} between levels b and a for simplicity. Later, we will consider the effect of incoherent pumping on gain, but in this section we put $r_{ba} = 0$. The sum of populations is evidently conserved in this model. Analysis of open systems with, e.g., flow of atoms through a cavity, leads to qualitatively similar conclusions.

Suppose that the probe field is weak, $|\Omega_p|^2 \ll \gamma_{cb}\gamma_{ac}$, and the drive-field intensity is constant in a sample. The complex susceptibility at the probe-field frequency, which is proportional to $\sigma_{ab}/\varepsilon_p$, is given by

$$\chi = \frac{i d_p^2 N}{\hbar c}\left\{\frac{1}{\Gamma_{ab} + |\Omega_d|^2/\Gamma_{cb}}\left(n_{ba} - \frac{|\Omega_d|^2 n_{ca}}{\Gamma_{cb}\Gamma_{ac}^*}\right)\right\} \qquad [8]$$

This expression can be further simplified if there is no pumping to the states $|c\rangle$ and $|a\rangle$, so that we can assume that $n_{ca} = 0$ and drop the second term in the brackets. The behavior of χ as a function of detuning from resonance $\nu_{\mathrm{p}} = \nu_{ab}$ is shown in Figure 1a. In the absence of the drive field Ω_{d}, we retrieve the usual two-level susceptibility, shown in Figure 1b.

There are two remarkable features in the susceptibility of Figure 1a. The first one is the transparency window at the point of two-photon resonance. The absorption coefficient at resonance is

$$K'' = \frac{2\pi \nu_{\mathrm{p}} d_{\mathrm{p}}^2 N}{\hbar c(\gamma_{ab} + |\Omega_{\mathrm{d}}|^2/\gamma_{cb})} \qquad [9]$$

Clearly, absorption can be arbitrarily small when $\gamma_{cb} \ll \gamma_{ab}$ and $|\Omega_{\mathrm{d}}|^2 \gg \gamma_{cb}\gamma_{ab}$, since it is proportional to $\gamma_{cb}/|\Omega_{\mathrm{d}}|^2$. For any value of $|\Omega_{\mathrm{d}}|^2$, the absorption is much smaller than that resulting from simple Stark splitting of two Lorentzians. Moreover, in the most interesting case

$$\gamma_{cb}\gamma_{ab} \ll |\Omega_{\mathrm{d}}|^2 \ll \gamma_{ab}^2$$

the Stark splitting is negligible while the deep and narrow EIT feature remains, with width $\gamma_{cb} + |\Omega_{\mathrm{d}}|^2/\gamma_{ab}$.

Note that the intensity of the drive field required for the coherent bleaching of the medium is a factor $r_{ab}/\gamma_{cb} \gg 1$ smaller than the intensity needed for usual saturation bleaching.

EIT is closely related to the phenomenon of coherent population trapping (CPT). This can be illustrated by transforming from states $|a\rangle$, $|b\rangle$, $|c\rangle$ to a new set of states $|a\rangle$, $|B\rangle$, $|D\rangle$, where

$$|D\rangle = (\Omega_{\mathrm{d}}|b\rangle - \Omega_{\mathrm{p}}|c\rangle)/\sqrt{|\Omega_{\mathrm{d}}|^2 + |\Omega_{\mathrm{p}}|^2}$$

$$|B\rangle = (\Omega_{\mathrm{d}}|b\rangle + \Omega_{\mathrm{p}}|c\rangle)/\sqrt{|\Omega_{\mathrm{d}}|^2 + |\Omega_{\mathrm{p}}|^2}$$

It can be easily found that under the condition of two-photon resonance $\nu_{\mathrm{p}} - \nu_{\mathrm{d}} = \nu_{cb}$ the dipole matrix element from $|D\rangle$ to $|a\rangle$ goes to zero. This means that the population in state $|D\rangle$ does not interact with the coherent radiation field and does not contribute to absorption. That is why this state is called a non-absorbing or dark state. Under the condition $|\Omega_{\mathrm{d}}|^2/\gamma_{cb}\gamma_{ab} \gg 1$ practically all the population of states $|b\rangle$ and $|c\rangle$ is optically pumped into the dark state, and only a small fraction $\propto \gamma_{cb}$ remains in the absorbing state $|A\rangle$ and contributes to absorption.

The second remarkable feature of a driven Λ-system is sharp variation of the real part of χ near the resonance, which can be used for manipulation of the group velocity of light, as discussed in the next section.

Slow, Ultra-Slow, Stored, and Frozen Light

Introduction

There are five useful definitions for the velocity of a traveling wave:

- the phase velocity, which is the speed at which the zero crossings of the carrier wave move;
- the group velocity, at which the peak of a wave packet moves;
- the energy velocity, at which energy is transported by the wave;
- the signal velocity, at which the half-maximum wave amplitude moves;
- the front velocity, at which the first appearance of a discontinuity moves.

Although these can all be different, in most cases (linear, passive, dispersive media) the last four coincide, and are usually less than the phase velocity. By using the very steep frequency dispersion in the vicinity of the narrow resonance associated with electromagnetically induced transparency (EIT), it has been shown that the group velocity of light can be reduced by approximately 100 million compared with its velocity in vacuum.

In order to understand the concept of group velocity, it is useful to consider the superposition of two waves of the same amplitude E_1 and E_2 where $E_i = E_0 \cos(k_i z - \nu_i t)$ with $i = 1, 2$. Addition of these waves gives rise to the modulation shown in Figure 2:

$$E = E_0(\cos(k_1 z - \nu_1 t) + \cos(k_2 z - \nu_2 t))$$

$$= 2E_0 \cos(\Delta k z - \Delta \nu t)\cos(kz - \nu t)$$

where $\Delta k = (k_1 - k_2)/2$, $\Delta\nu = (\nu_1 - \nu_2)/2$, $k = (k_1 + k_2)/2$, and $\nu = (\nu_1 + \nu_2)/2$. This superposition

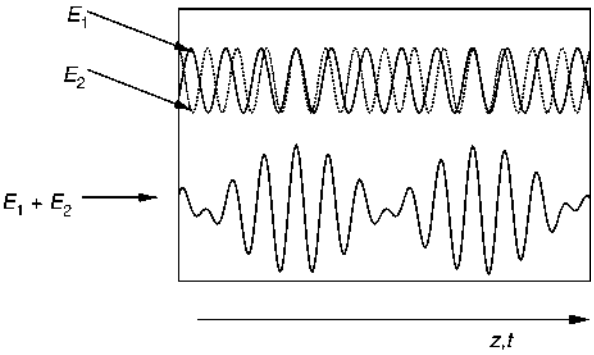

Figure 2 Interference of two monochromatic waves with different frequencies results in a wave modulated in time and space.

displays an interference pattern consisting of a rapid oscillation propagating with the so-called phase velocity,

$$v_{\text{phase}} = \frac{\nu}{k} \qquad [10]$$

and a slowly varying envelope propagating at the group velocity,

$$v_{\mathrm{g}} = \frac{\Delta\nu}{\Delta k} \qquad [11]$$

Usually, the group velocity of wave packets consisting of more than two harmonics is given by transforming the ratio $\Delta\nu/\Delta k$ into $\mathrm{d}\nu/\mathrm{d}k$.

Slow and Ultra-Slow Light

In ordinary absorption and refraction of light, it is sufficient to consider only two atomic energy levels interacting with the light. However, in multilevel systems, coherent excitation of more than one transition can dramatically alter the dispersion and hence the group velocity. The first experiments of this effect demonstrated a group velocity of $c/13$. However, this was progressively lowered in later experiments as the steepness of the dispersion was increased, culminating in a group velocity as low as $c/300\,000\,000$ (1 m/s) in a Bose–Einstein condensate of ultra-cold sodium atoms. Slow light has been observed in a wide range of systems, from the Bose–Einstein condensates just mentioned, to hot optically dense gases of atoms and molecules, and in solids ranging from rare-earth doped crystals to solid hydrogen.

Most of these experiments occurred in a Λ system. The susceptibility of this medium has very steep dispersion without absorption, as shown in Figure 1.

The minimum possible group velocity slowed by the steep dispersion in EIT is given by

$$v_{\mathrm{g}} = \frac{8\pi\Omega^2}{3\lambda^2 N\gamma} > 0 \qquad [12]$$

where $\gamma = \gamma_{ab} = \gamma_{ac}$. The driving field Ω should be strong enough to meet the condition $\Omega^2 \gg \gamma\gamma_{bc}$ for coherent effects. It can be shown that the group velocity approaches its minimum $v_{\mathrm{g\,min}} = 8\pi\gamma_{bc}/(3\lambda^2 N)$ when $\Omega^2 \approx \gamma\gamma_{bc}$.

Bringing Light to a Halt: Frozen Light

Consider a laser pulse propagating through a medium with ultra-slow group velocity while the medium itself is moving with some velocity v in the opposite direction to the direction of laser pulse propagation. Let us denote the group velocity of the light in the frame co-moving with the atoms as $\tilde{v}_{\mathrm{g}}$. In this frame, atoms are at rest, and hence, there is no spatial dispersion. The Galilean transformation to the laboratory frame, $k = \tilde{k}$, $\nu = \tilde{\nu} - \tilde{k}v$, where v is the atomic velocity, yields the group velocity $v_{\mathrm{g}} = Re(\mathrm{d}\nu/\mathrm{d}k) = \tilde{v}_{\mathrm{g}} - v$. This simple transformation shows that the light pulse is 'dragged' by the moving atoms. If the velocity of the atoms with respect to the laboratory frame is exactly equal and opposite to the group velocity of light in the frame where atoms are at rest, then we find that the group velocity in the laboratory frame is zero. We refer to this as 'freezing' of light. We may even make the group velocity negative so that the propagation direction is opposite to the wavevector (that is, that $v_{\mathrm{g}} < 0$). The simplest example of spatial dispersion is the so-called drift dispersion corresponding to a mono-velocity atomic beam or moving sample with velocity v.

Slow light experiments in Doppler broadened systems such as warm gases allow group velocities that are much lower than the mean thermal speed (v_{T}) of the atoms in the thermal sample (Figure 3). This means that freezing of light can occur without actually moving the atomic sample. When the experimentally obtained group velocity is less than or about equal to the mean thermal speed of the atoms, a pulse of light can be stopped in a stationary cell. In this case the delay time for a pulse passing through the medium would tend to infinity (contrary to the case of a finite atomic beam) in the sense that this pulse would never leave the cell.

This is accomplished by using a single velocity group from the Maxwellian thermal distribution of atomic velocities, and adjusting the frequency of the driving field to be resonant with that velocity group (see Figure 4). In this case it is mainly the atoms in this single velocity group that support the ultra-slow group velocity as a slow EIT polariton, and the atoms

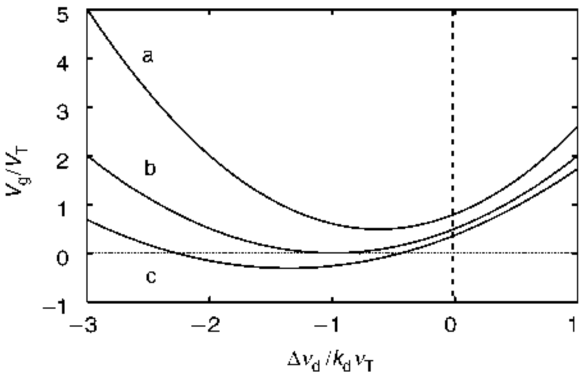

Figure 3 Ultra-slow and negative group velocity of EIT polariton versus detuning of drive laser; $\Omega = 0.25\gamma$, $k_d v_T = 100\gamma$, $\gamma_{cb} = 0.001\gamma$, (a) $N = 0.6 N_{\mathrm{cr}}$; (b) $N = N_{\mathrm{cr}}$; (c) $N = 1.5 N_{\mathrm{cr}}$.

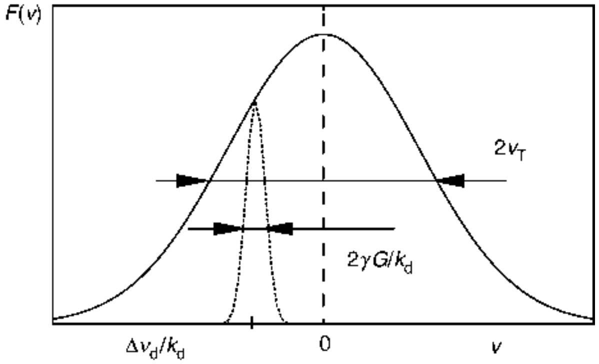

Figure 4 The velocity distribution of atoms in a cell (solid line). Effective drifting beam (dotted) selected by drive laser.

then act as an effective atomic beam. If the driving field is on resonance with the velocity group of atoms moving into the light with speed equal to $\tilde{v}$, we will have $v_g = \tilde{v}'_g - v_d$, where $\tilde{v}'_g$ is the group velocity of the light defined by the effective atom beam if those atoms were at rest. This requires that the intensity of the drive is strong enough to provide EIT for the resonant velocity group of atoms, i.e., $\Omega^2 > \gamma_{bc}\gamma$ (see Figure 4) but at the same time weak enough to avoid interaction with off-resonant atoms moving with the 'wrong' velocities (i.e. $\Omega < kv_T\sqrt{\gamma_{bc}/\gamma}$).

Storing and Retrieving Quantum Information

Photons are potentially the best particles for use in quantum computing and quantum information processing. They are the fastest, and perhaps the simplest and most robust carriers of quantum information, but their storage and manipulation is difficult without destroying their quantum state. However, for application to quantum cryptography, quantum teleportation, and quantum computation it is necessary to achieve a strong coupling between electromagnetic waves and a large nonlinear susceptibility with small losses. For this task, it may be resonant coherent media that offer the promise of success.

Recently, new coherent and reversible methods based on EIT have been found for the transfer and manipulation of photons and their quantum information. Quantum states of photons are trapped in coherently driven atomic media in which the group velocity is reduced adiabatically toward zero. The methods allow for an ideal transfer of quantum correlations between light fields and metastable states of matter. Currently, interesting applications such as quantum state memories and quantum information processing including transporting and time reversing the state of light are being investigated.

Although experiments have not yet been conducted with single quanta, proof-of-principle experiments using classical laser fields have demonstrated the technique. In these experiments control and signal light pulses propagate in a gas of three-level Λ-type atoms and excite a spatial profile of a long-lived coherence of ground spin states. This spin coherence profile contains quantum information about the 'writing' pulses. When the control field is switched back on, the signal pulse is restored via Raman scattering on the atomic coherence. Ideally, the scattered pulse is in the same quantum state as the input signal pulse, preserving entanglement, quantum statistics, etc.

Similar experiments have also shown that different reading pulses (rather than switching back on the same control pulse) can be used to obtain new tools for quantum information storage and processing. Time reversing, teleportation, color switch and multiplexing of the signal pulse of light have all been demonstrated. These new operational possibilities arise as a result of the reading of the stored information with a light pulse that is spatially separated from, has different frequency from, and propagates in the opposite direction to the signal beam.

To demonstrate these effects, one pair of control and signal pulses (the 'writing' fields) prepares a spin coherence of the lower level (Figure 5). A second control pulse (which can be thought of as a 'reading' field) scatters from this coherence resulting in the generation of a new, or 'recovered', field. This recovered field acquires some properties of the reading field because it is a Raman scattered component of the reading pulse. If the reading field pulse is centered about a frequency other than that of the writing fields, and propagates in the opposite direction to the writing fields, then the recovered field propagates in the same direction as the reading field and has a different frequency from the incident writing fields. Even if the writing and reading pulses are separated in space as well as time the scattering effect will persist. This allows the transportation of the coherence grating to another point in the atomic cell via the atomic center-of-mass motion.

In the ideal case the signal pulse profile and quantum statistics are exactly stored by the atomic coherence grating. Thus, reading by a backward propagating pulse results in a generated pulse that is also backward propagating. That is it represents an exactly time-reversed copy of the signal pulse. Similarly, atomic motion in the direction transverse to the light propagation moves the coherence grating to another spatial position. Reading the light by a spatially shifted laser results in the equivalent of

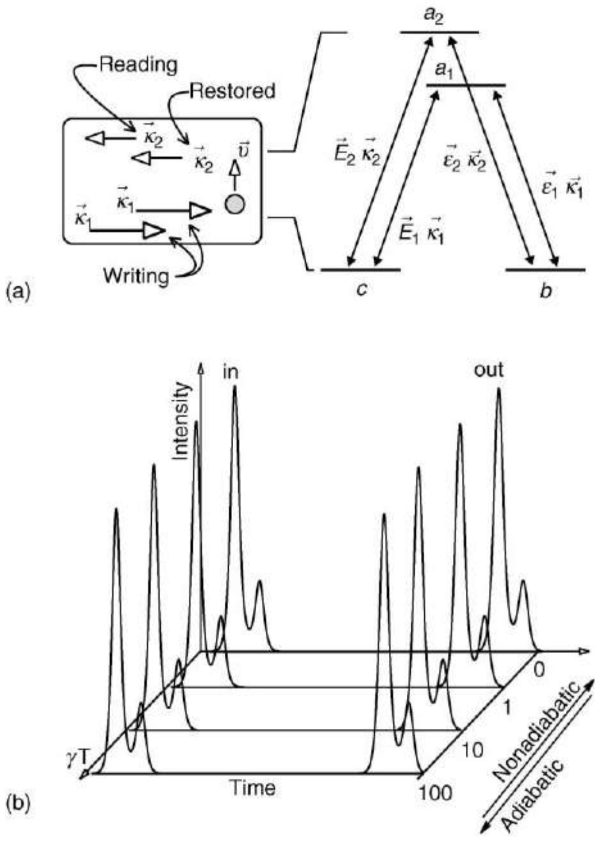

Figure 5 (a) Energy level scheme used in experiments for multiplexing, transporting, and time-reversing stored light. (b) 'Storage of light': intensity of the signal field versus time for different switching times (T) of the reading and writing fields.

'teleportation' of light by atoms in the same spirit as the teleportation of the quantum state of light by atoms that has been discussed recently. It is also possible for the frequency of the 'restored' pulse to be easily shifted by switching the frequency of the reading pulse. This means that the use of a frequency comb of reading pulses propagating in different spatial modes should result in multiplexing of the signal light.

Novel Laser Sources Based on Quantum Coherence

Introduction

In the previous sections we have seen how quantum coherence can be used to overcome resonant absorption on a transition of interest, leading to EIT. A natural question arises whether it is possible to achieve gain without population inversion at this transition, i.e. lasing without inversion (LWI). This subject has been intensively investigated since the early 1990s, and the first proof-of-principle experiments have been successful.

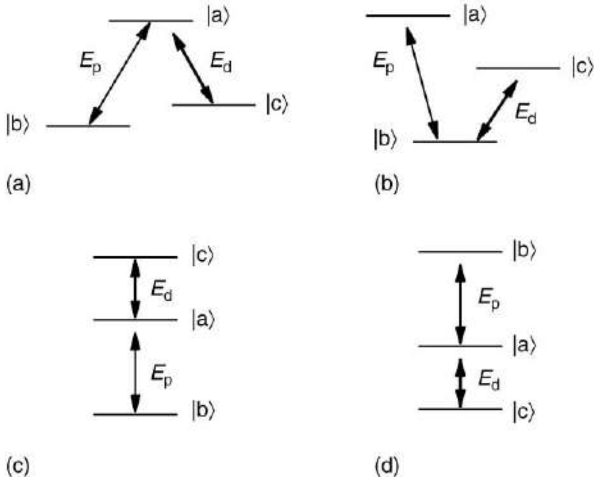

Figure 6 LWI schemes in a three-level system coherently driven by a laser field.

Besides the fundamental physical beauty of this phenomenon, the research effort has been motivated by potentially important applications. It is clear that LWI could be particularly useful in situations where a population inversion is difficult to maintain due to, e.g., very fast relaxation or inefficient pumping. The obvious examples are high-frequency UV, X-ray and gamma-ray lasers, or semiconductor mid/far-infrared lasers based on intersub-band transitions. Another promising application of LWI is efficient frequency up- or down-conversion. Furthermore, if resonant absorption can be eliminated, resonant nonlinear optical interactions become a reality. Since all nonlinear coefficients are maximized at resonance, highly efficient nonlinear optical generators of light can be created.

In this section we discuss the simplest possible schemes of inversionless lasing which allow one to clarify the physical origin of LWI and to see the shortcomings which have so far prevented LWI-based schemes from reaching the level of practical laser devices.

Consider a three-level system, coherently driven by a strong quasi-monochromatic field nearly resonant with one transition, and probed by a field that is nearly resonant with an adjacent transition. Some possible configurations are shown in Figure 6. We consider in detail the case when the two fields are in a Λ-configuration (Figure 6a). This is probably the most popular and well-studied scheme since it can provide probe-field gain for a very low pumping rate, when most of the population is at the lowest energy level. However, we will see that the gain condition for the Λ-scheme is actually rather restrictive and does not allow one to realize large frequency up-conversion. For the latter purpose, ladder-type schemes (Figure 6c,d) turn out to be more favorable.

Weak-field gain in the Λ-scheme

From the general formalism presented earlier, it is straightforward to get the amplification condition

$$K'' = \mathrm{Re}\left\{ \eta_{\mathrm{p}} \frac{1}{\Gamma_{ab} + |\Omega_{\mathrm{d}}|^2/\Gamma_{cb}} \left(n_{ba} - \frac{|\Omega_{\mathrm{d}}|^2 n_{ca}}{\Gamma_{cb}\Gamma^*_{ac}} \right) \right\} + \kappa_{\mathrm{p}} < 0 \qquad [13]$$

where

$$\eta_{\mathrm{p}} = \frac{2\pi\omega_{\mathrm{p}} d_{\mathrm{p}}^2 N}{\hbar c n_{\mathrm{p}}^2}$$

The physical meaning of the two terms in brackets in eqn [13], which originate from the two terms on the right-hand side of eqn [4] for σ_{ab}, can be easily understood. The first term is just the resonant one-photon absorption proportional to the population difference n_{ba}. It can be strongly reduced in the presence of a drive field Ω_{d}, as explained earlier. However, it can change sign and give rise to the field amplification only if $n_{ba} < 0$, meaning there is a population inversion!

The second term, originating from the product $\Omega_{\mathrm{d}}\sigma_{cb}$ in eqn [4], is due to the mixing of the drive field Ω_{d} and polarization σ_{cb} where the latter is parametrically excited by a two-photon process: $\sigma_{cb} \propto e_{\mathrm{p}}\sigma^*_{ac} \propto e_{\mathrm{p}}\Omega^*_{\mathrm{d}} n_{ca}$. It is this term that can overcome one-photon absorption to provide amplification without inversion in the Λ-scheme. As we see, the possibility of LWI depends crucially on our ability to excite a sufficiently large polarization σ_{cb} at the transition resonant with the beat frequency of the probe and drive fields. This conclusion is universal for all LWI schemes with a coherent drive. We see that the sometimes loosely defined term 'quantum coherence' has a precise meaning here as the polarization σ_{cb} at a two-photon transition c–b. It is clear that the effects of quantum coherence are maximized when γ_{cb} is small, i.e., the transition c–b is long-lived.

We are of course interested in the case of weak incoherent pumping. The typical dependence of gain on detuning from two-photon resonance is shown in Figure 7. At resonance, where the gain is maximized, the amplification condition $K'' < 0$ yields (in the absence of nonresonant losses):

$$r_{ac} > 2\gamma_{cb}(r_{ab}/r_{ba} - 1) \geq r_{ab} - r_{ba} \simeq r_{ab} \qquad [14]$$

because $\gamma_{cb} \geq r_{ba}/2$. Here we assumed that the incoherent pumping is weak, $r_{ba} \ll r_{ab}$, so that it does not produce population inversion at the probe transition.

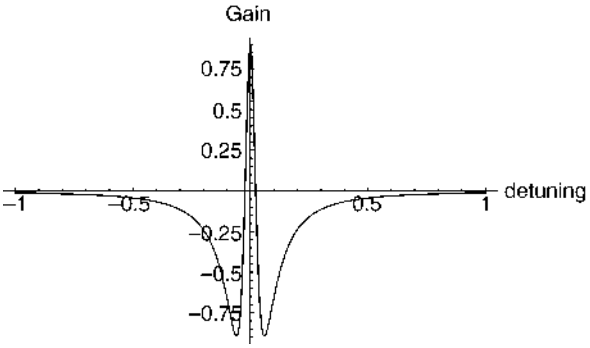

Figure 7 Gain without inversion in Λ-scheme.

Equation [14] is a simple and very important result, which is valid for all three-level laser schemes in the absence of inversion at both probe and drive transitions: the relaxation rate at the drive transition should exceed that at the probe transition. This result is an immediate consequence of the fact that the long-lived coherence σ_{cb} is inevitably destroyed by incoherent pumping, and γ_{cb} is now bounded from below. One can follow how the EIT regime is transformed into lasing as follows. With increasing incoherent pumping rate r_{ba}, EIT is actually destroyed due to the growth of $\gamma_{cb} \simeq r_{ba}/2$. In terms of CPT states, this means increase of population of the absorbing state $|B\rangle$. At the same time, the resonant Raman term in [13] also grows since n_{ac} is proportional to r_{ba}. With increase in the incoherent pumping rate, the contribution to gain overcomes the contribution to absorption if $r_{ac} > r_{ab}$.

Propagation effects

Inequality [14] has important implications for the problem of co-propagation of probe and drive fields. Indeed, to provide a significant amplification of the probe field, the absorption length $L_{\mathrm{d}} = \gamma_{ac}/(\eta_{\mathrm{d}} n_{ca})$ of the drive field should be larger than the amplification length $L_{\mathrm{p}} \simeq \gamma_{ac}/(\eta_{\mathrm{p}} n_{ca})$ of the probe. This requirement gives simply

$$\eta_{\mathrm{p}}/\eta_{\mathrm{d}} > 1 \qquad [15]$$

which in the absence of nonradiative relaxation can be rewritten as

$$\frac{r_{ab}}{r_{ac}} \frac{\nu_d^2}{\nu_p^2} > 1$$

In view of the gain condition given in eqn [14], the above requirement can be satisfied only if $\nu_\mathrm{d} > \nu_\mathrm{p}$, i.e., up-conversion is impossible.

An interesting solution to this problem is to realize population inversion at the driving transition. Then up-conversion can be realized with a lower relaxation rate at the drive transition as compared with probe transition. In this case in the Λ and V-schemes the two-photon term contributes to absorption at resonance, and gain is only possible far off-resonance. At the same time, in the ladder schemes shown in Figure 6c and d the gain is achieved at resonance, as discussed below in more detail.

It should also be noted that to excite laser oscillations in a high-Q cavity, large one-pass gain is not required at all, so that a Λ-scheme could still be employed for up-conversion.

Now, consider a probe field of arbitrary intensity. An important question is where the energy for amplification of the probe field comes from. It is known that for parametric processes in a transparent medium the Manley–Rowe relations are satisfied. Therefore, it might seem that the drive field is also the ultimate energy source for LWI. However, this is the case only for very large detunings of the drive and probe fields from the corresponding transition frequencies. In that case, we can neglect one-photon absorption, and the wave equations take the form typical for stimulated Raman scattering. They possess the first integral

$$\frac{|\varepsilon_\mathrm{d}|^2}{\nu_\mathrm{d}} + \frac{|\varepsilon_\mathrm{p}|^2}{\nu_\mathrm{p}} = \mathrm{const} \qquad [16]$$

which expresses the conservation of the total number of quanta in both waves (Manley–Rowe conditions).

However, in the case of exact resonance, one can show that one-photon absorption of the drive field is always important, and the Manley–Rowe-type conservation law does not exist at all. The optimal condition for the probe field amplification is $\eta_\mathrm{p}/\eta_\mathrm{d} \gg 1$. In this case the stage of exponential amplification of the probe is maintained over a long distance, and the probe field intensity can grow to much larger values than the sum of initial intensities of the drive and probe fields. Apparently, the energy for the probe field amplification comes mainly from the medium, namely from the upper-level population ρ_{aa} supported by incoherent pumping.

V-Scheme

The V-scheme, shown in Figure 6b, is qualitatively similar to the Λ-scheme. An important difference is that there is now no long-lived coherence at the mixing transition c–a. Indeed, even if the transition $a \rightarrow c$ is strongly forbidden, the coherence γ_{ac} is destroyed by fast relaxations at the c–b and a–b transitions. We can immediately suggest that for the V-scheme inversionless lasing takes place at large driving field intensities exceeding the saturation value. However, one must be careful: although it appears that the reduction of absorption in the absence of pumping cannot be greater than a factor of 2, this turns out to be incorrect. Rather, almost complete transparency can be achieved, as confirmed by calculations and recent experiments. Analysis of the laser gain condition shows that the V-scheme shares with Λ-scheme the same negative feature: relaxation at the driving transition should be much faster than at the probe transition.

Schemes with Population Inversion at the Driving Transition

The above analysis illustrates the common problem of LWI schemes: the requirement of rapid relaxation at the driving transition and the resulting strong absorption of the drive field. This difficulty can be overcome in schemes with population inversion at the driving transition. Then, of course, all this makes practical sense only if the relaxation at the drive transition is much slower than that on the probe transition, so that it is easy to provide population inversion for the drive transition as compared with the probe transition. Luckily, such schemes do exist in principle, and seem to have many potential applications. The simplest possibility is the so-called ladder (or cascade) schemes shown in Figure 6c,d. LWI in ladder schemes is usually studied in the regime when there is no inversion at both probe and drive transitions. In this case gain is achieved at large detunings from two-photon resonance and under the same unfortunate condition of fast relaxation at the drive transition as compared with the probe transition. In the opposite case, population inversion at the driving transition is required for LWI at the probe transition. In ladder schemes the maximum gain of the probe field is achieved when both fields are at resonance with atomic transitions, while in V and Λ-schemes the resonance point corresponds to maximum absorption, and the gain is achieved in sidebands.

Note that population inversion at the driving transition means that the drive field can be amplified. This immediately puts forward the possibility to use a drive field which is self-generated in the same active medium. In other words, the system supports lasing with inversion at the drive frequency and LWI at the probe frequency. This situation is especially beneficial for dense gases or solids where the problems of external drive absorption, reflection, and inhomogeneity may become otherwise insurmountable. One specific example considered in the literature is LWI on rapidly decaying UV and soft X-ray transitions in gas lasers in the presence of lasing with inversion on long-lived adjacent transitions in the visible range (as in Figure 6c). Another example is LWI on rapidly decaying mid/far-infrared intersub-band transitions in semiconductor quantum wells in the presence of lasing with inversion on the interband transition, as in Figure 6d. These interband transitions have 10^4 times longer lifetime than the intersub-band transitions.

Experiments

Experimentally, amplification without inversion has been observed in rarefied atomic vapors in the early and mid-1990s. In particular, amplification of a probe beam in the Λ-configuration was realized in atomic samarium, sodium, and cadmium, while the V-scheme was successfully implemented in atomic rubidium. Inversionless laser oscillation was observed in a V-scheme in rubidium vapor and a Λ-scheme in a sodium atomic beam. In all of these proof-of-principle experiments, the frequencies of the probe and drive fields were very close to each other, and the coherence was excited between two hyperfine states or Zeeman sublevels. One-pass gain of a few per cent was observed.

Nonlinear Optical Sources Based on Quantum Coherence

Excitation of a strong coherence has a profound impact on nonlinear optical interactions. Strong scattering on the polarization wave excited at the beat frequency of two or more laser fields allows one to achieve huge nonlinear coefficients and conversion efficiencies. A particularly impressive example is ultra-broadband collinear Raman generation in a resonant Λ-scheme, that has been observed experimentally. A pair of strong laser pulses propagating in molecular deuterium excites a strong polarization wave at a beat frequency close to a vibronic transition in deuterium. As a result, a comb of Raman sidebands from the infrared to the deep-UV region of the spectrum has been observed.

Another example is an extremely efficient simultaneous blue light and mid-infrared generation in a four-wave mixing experiment in rubidium vapor. Here, all four fields are resonant with corresponding transitions.

A final example illustrates the use of self-generated drive fields for nonlinear generation. It has been suggested that one can use two intracavity generated laser modes in a semiconductor quantum-well laser for the difference frequency, sum frequency, and second-harmonic generation in fully resonant three-wave mixing processes. This idea has already been successfully realized in two-wavelength quantum-cascade lasers, and experiments with other types of semiconductor lasers are under way.

The whole field of resonant nonlinear optics, stemming from original ideas of LWI in systems with quantum coherence, is now in explosive development.

Summary

Atomic coherence effects have already found many important applications in metrology and low-intensity nonlinear optics. Novel applications in quantum nonlinear optics, laser physics, and quantum information processing seem to be feasible. The coherence allows the storage of information about the probe light, its transportion in space, time reversing, and multiplexing the light. Moreover, the coherence allows an increase of the coupling between light fields such that it becomes possible to study the interaction between single photons. New coherent radiation sources based on the ideas of LWI and coherent nonlinear optics are at the stage of active research. There are surely many other applications ahead.

See also

Electromagnetically Induced Transparency. **Non-Classical Light**. **Nonlinear Optics, Applications:** Raman Lasers. **Nonlinear Optics, Basics:** Cascading; Four-Wave Mixing. **Quantum Optics:** Entanglement and Quantum Information.

Further Reading

Belyanin AA, Bentley C, Capasso F, *et al.* (2001) Inversionless lasing with self-generated driving field. *Physical Review A* 64: 013814.

Harris SE (1997) Electromagnetically induced transparency. *Physics Today* 50(9): 36–42.

Kocharovskaya O (1992) Amplification and lasing without inversion. *Physics Reports* 219: 175–190.

Mompart J and Corbalán R (2000) Lasing without inversion. *Journal of Optics B: Quantum and Semiclassical Optics* 2: R7–R24.

Rostovtsev Y, Kocharovskaya O, Welch GR and Scully MO (2002) Slow, ultra-slow, stored and frozen light. *Optics Photonics News*, June.

Scully M and Zubairy S (1997) Quantum Optics. Cambridge, UK: Cambridge University Press.

Entanglement and Quantum Information

P G Kwiat, University of Illinois at
Urbana-Champaign, Urbana, IL, USA

D F V James, Los Alamos National Laboratory,
Los Alamos, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Quantum information is an emerging field of technology that encompasses the application of fundamental quantum mechanical phenomena, such as entanglement, to tasks in information processing and communications. The importance of optical technology in quantum information is not surprising, given the success of quantum optics in the investigation of these fundamental phenomena and the pre-eminent role of optics in modern communications. This article describes some of the important advances in photonic quantum information and discusses the prospects for the future.

Basic Notions: Qubits and Entanglement

Qubits

Quantum information is usually confined to two-level quantum systems, which are referred to as quantum bits, or qubits. The qubits form the register of a quantum computer or carry the information in a quantum communications channel. Physical examples of qubits under active study include the spin degrees of freedom of an electron or spin-1/2 nucleus, two well-isolated energy levels of an atom or trapped ion, and the polarization degrees of freedom of a photon; here we concentrate on the latter.[1] The two states of each qubit are generically denoted by $|0\rangle$ and $|1\rangle$ (although, because of our concentration on optical polarization, we shall use the horizontal and vertical states $|H\rangle$ and $|V\rangle$ interchangeably with $|0\rangle$ and $|1\rangle$ when describing experiments). An arbitrary state of a single qubit, $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, is specified by the complex probability amplitudes α and β associated with the two possible states, constrained by $|\alpha|^2 + |\beta|^2 = 1$. States for which α and β are deterministic quantities are referred to as pure states; when they are stochastic, the state is mixed, and must be described by a density matrix, e.g., $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, where p_i is the probability of the pure state $|\psi_i\rangle$. Pure states are a useful but unattainable ideal, just like a perfectly coherent laser beam; in reality even the best-prepared quantum states have some level of mixedness, which can be quantified by the entropy: $S \equiv -Tr(\rho \ln\rho) = -\sum_i \lambda_i \ln\lambda_i$, where λ_i are the eigenvalues of the matrix ρ.

Multiqubit States and Entanglement

For two qubits, the possible pure states of the system are specified by four probability amplitudes:[2] $|\psi_2\rangle = \alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle$, with $|\alpha|^2 + |\beta|^2 + |\gamma|^2 + |\delta|^2 = 1$, and, e.g., $|00\rangle \equiv |0\rangle \otimes |0\rangle$ denoting the state in which the first and the second qubits are both in state $|0\rangle$. In general, 2^n amplitudes are needed to characterize a pure state of n qubits. Immediately we can see one advantage of quantum systems for information storage: if each distinct probability amplitude is regarded as a data register, the size of the memory grows exponentially with the number of qubits. Further, if one were to flip just one qubit, for example the second, the state would be transformed into $|\psi'_2\rangle = \beta|00\rangle + \alpha|01\rangle + \delta|10\rangle + \gamma|11\rangle$, where all of the amplitudes have been affected. This simple example demonstrates the notion of quantum parallelism, one of the most powerful properties of quantum information processors.

If the two-particle state can be written as a product of two single-particle states, i.e., if $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle = (A|0\rangle + B|1\rangle) \otimes (C|0\rangle + D|1\rangle)$, then these two particles can be considered as separate, unconnected entities: the state is said to be separable.[3] When the state cannot be written in this form, it is called an entangled state. Entanglement is one of the most fascinating fundamental properties of quantum systems: Erwin Schrödinger described it as '*the* characteristic trait of quantum mechanics … that enforces its entire departure from classical lines of thought.' Consider performing a measurement on the second of the two qubits, to establish which of its two states the particle was in, leaving the first qubit untouched. The result would collapse the state of the first qubit into one of two possibilities: either $|\phi_0\rangle = (\alpha|0\rangle + \gamma|1\rangle)/\sqrt{P_0}$ if the outcome of the measurement were 0 (which occurs with probability $P_0 = |\alpha|^2 + |\gamma|^2$) or $|\phi_1\rangle = (\beta|0\rangle + \delta|1\rangle)/\sqrt{P_1}$ if it were 1 (with probability $P_1 = |\beta|^2 + |\delta|^2$). Thus the state of the first qubit is instantaneously projected into a specific state by performing a measurement on the second qubit, which can be in a completely different location. This example demonstrates the nonlocality of quantum mechanics, which has been confirmed experimentally in some depth by a number of elegant quantum optics experiments over the last 30 years.

---

[1] Recently some attention has been devoted to quantum information based on continuous variables, such as those arising in connection with 'squeezed light'. However, here we will focus on discrete systems, i.e., qubits.

[2] Qubits are assumed to be distinguishable particles, and so the wavefunctions describing their states need not be symmetrized under exchange of particles.

[3] When mixture is also present, the state is separable if it can be written as $\rho = \sum_i p_i \rho_i^A \otimes \rho_i^B$, i.e., as a sum of product states for systems A and B.

The fidelity $F = |\langle\Phi|\Psi\rangle|^2$ characterizes the overlap of two states.[4] The fidelity of $|\phi_0\rangle$ and $|\phi_1\rangle$ is a way to quantify the amount of entanglement in the initial two-qubit state. If the final two states are identical, the fidelity will be unity, and the initial state is separable; conversely if the two final states are orthogonal and occur with equal probability, then in a sense the initial state is maximally entangled, since the largest possible nonlocal influence would occur due to the measurement. Mathematically, the fidelity of the two final states is $|\langle\phi_0|\phi_1\rangle|^2 = 1 - C^2/4P_0P_1$, where $C = 2|\alpha\delta - \beta\gamma|$ is called the concurrence of the state. If $C = 0$, the state is separable; if $C = 1$, the state is maximally entangled.[5]

As discussed below, entanglement forms the heart of a number of quantum information protocols, such as dense coding, teleportation, and one type of quantum key distribution (also known as quantum cryptography), and large-scale entangled states of many qubits seem to be a requirement for the more ambitious goal of practical quantum computing.

Creating Entangled States Experimentally

Entangled states can currently be created in a controlled manner using technologies such as ion traps, cavity quantum electrodynamics, and optical spontaneous parametric down-conversion (SPDC). Down-conversion is a nonlinear optical process by which an incident ‘pump’ photon can be split (or down-converted) into a pair of longer-wavelength daughter photons (historically called ‘signal’ and ‘idler’) in a crystal possessing a $\chi^{(2)}$ nonlinearity, such as beta-barium borate (BBO). Mathematically, the process is described using the creation and annihilation operators of the field modes $(\hat{a}_\lambda^\dagger, \hat{a}_\lambda)$, thus:

$$|\Phi_{\text{out}}\rangle = \exp\left[\; -i\sum_{p,s,i} c_{p,s,i}(\hat{a}_p^\dagger \hat{a}_s \hat{a}_i + \hat{a}_i^\dagger \hat{a}_s^\dagger \hat{a}_p)\right]|\Phi_{in}\rangle$$

where the triple sum is over all field modes (subscripts p, s, and i refer to pump, signal, and idler modes, respectively) and $c_{p,s,i}$ is a coefficient linearly dependent on the second-order nonlinear susceptibility $\chi_{ijk}^{(2)}$ and also on the birefringent properties of the crystal. Further, $c_{p,s,i}$ will be negligibly small unless both total photon energy and momentum are conserved (i.e., $\omega_p = \omega_s + \omega_i$ and $\vec{\kappa}_p = \vec{\kappa}_s + \vec{\kappa}_i$, where $\vec{\kappa}$ is the intracrystal momentum).

One particularly efficient method for using this phenomenon to create entangled states is as follows. For a specific geometry (type-I phase matching), the daughter photons emerge from the crystal with identical polarizations (perpendicular to the parent polarization and the crystal optic axis) on opposite sides of a cone that is centered about the pump beam (Figure 1a). Because each photon is in a definite state of polarization, the two photons are not in an entangled state. Two crystals, aligned with their axes of symmetry oriented at 90° to each other, as shown in Figure 1b, can then be used to create an entangled state.[6] With crossed crystals, two processes are possible: the parent photon can down-convert in the first crystal to yield two vertically polarized photons, or it can down-convert in the second crystal to yield two horizontally polarized photons. Because it is impossible to distinguish which of these processes has occurred, the state of the daughter photons is a coherent quantum-mechanical superposition of the states that would arise from each crystal alone; the output of the crossed crystals is photons in the maximally entangled state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|HH\rangle + |VV\rangle)$.[7]

Figure 2 shows how this basic source can be adapted to produce any pure quantum state of two photons by placing rotatable half- and quarter-wave plates (which can be used to transform the

---

[4] The generalized fidelity between mixed states ρ_A and ρ_B is $F = |Tr\sqrt{\sqrt{\rho_A}\rho_B\sqrt{\rho_A}}|^2$.

[5] The concurrence can also be generalized to mixed states, although it has a much more complicated expression. The quantification of entanglement for mixed states with more than two subsystems is presently an active area of research.

[6] Another widely used method employs type-II phase matching in a single crystal. The down-conversion photons are emitted with perpendicular polarizations along a pair of cones. Pairs emitted along particular directions are in the maximally polarization-entangled state $|\psi_-\rangle = \frac{1}{\sqrt{2}}(|HV\rangle - |HV\rangle)$. Such a source was used in the first demonstrations of superdense coding and quantum teleportation.

[7] Note that in an arbitrary basis $|\theta\rangle \equiv \cos\theta|H\rangle + \sin\theta|V\rangle$ and $|\theta^\perp\rangle \equiv -\sin\theta|H\rangle + \cos\theta|V\rangle$, this maximally entangled state has the form $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|\theta\theta\rangle + |\theta^\perp\theta^\perp\rangle)$, demonstrating that the nonlocal correlations are present regardless of the bases used to represent the state.

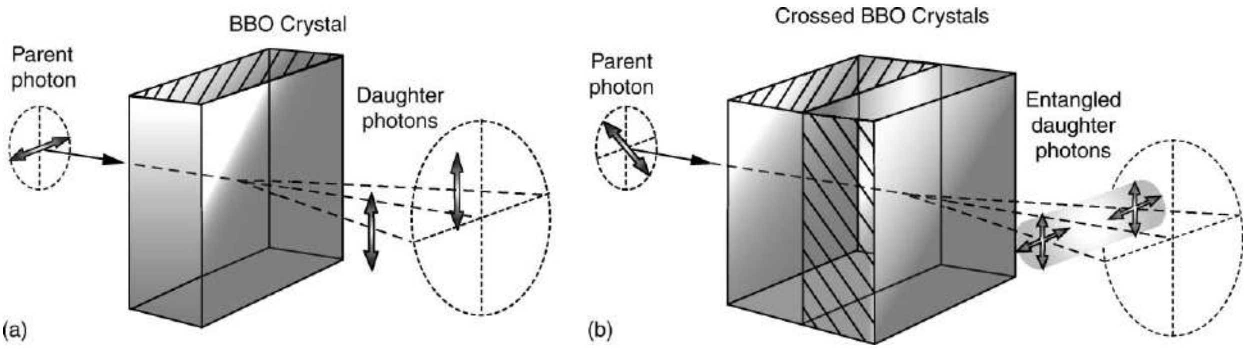

Figure 1 An entangled-photon source. (a) For a given orientation of the nonlinear crystal, a horizontally polarized parent photon produces a pair of vertically polarized daughters. The daughters emerge on opposite sides of an imaginary cone. The cone's axis is parallel to the original direction taken by the parent photon. The two daughter photons are not in an entangled state. Reorienting the BBO crystal by 90° will produce a pair of horizontally polarized daughters if a vertically polarized pump beam is used. (b) Passing a photon polarized at 45° through two crossed BBO crystals can produce two photons in an entangled state. Because of the Heisenberg uncertainty principle, there is no way to tell in which crystal the parent photon 'gave birth,' and so a coherent superposition of two possible outcomes results: a pair of vertically polarized photons or a pair of horizontally polarized photons. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

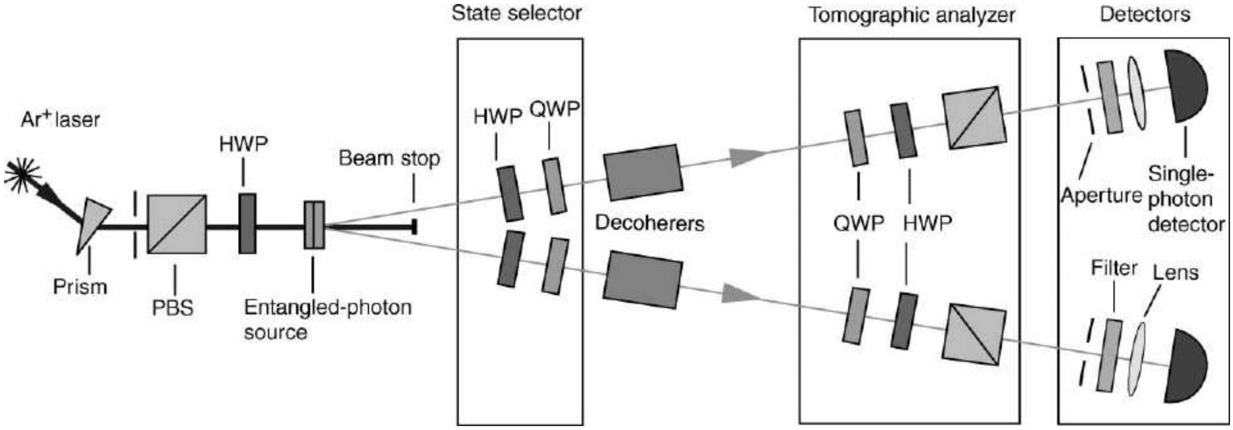

Figure 2 Creating and measuring two-photon entangled states. The 'pump' photons are created, for example, in an argon ion laser and are linearly polarized with a polarizing beamsplitter (PBS). The half-wave plate (HWP) rotates the polarization state before the photon enters the pair of nonlinear crystals that constitute the entangled-photon source; the initial angle of pump polarization controls the entanglement of the pair produced. Each photon's polarization state can be altered at will by the subsequent HWP and quarter-wave plate (QWP). The decoherers following the state selection allow the production of mixed photon states. The optical elements (QWP, HWP, and PBS) in the tomographic analyzer allow the measurement of each photon in an arbitrary basis. Coincidence measurements of photons allow the quantum state to be determined. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

polarization state of a single photon) before the crystal pair and in the paths of the two daughter photons. To create mixed states, a long birefringent crystal can be used to delay one polarization component with respect to the other. If the delay is longer than the coherence time of the photons, the horizontal and vertical components are effectively decohered; that is, the phase relationship between the different states is destroyed.

The figure also shows schematically the apparatus required for measuring the quantum state. In classical optics, four Stokes parameters are required to specify the polarization of a single beam (i.e., an ensemble of uncorrelated photons); for a pair of photons, 16 projective measurements, each with different wave plate settings, is required. From these 16 measurements, all of the elements of the 4×4 density matrix describing the (in general, mixed) state of the photon pairs can be deduced. This is an example of quantum state tomography, a technique that has found application to a number of quantum optical systems.

Quantum Key Distribution

Two parties, historically known as Alice and Bob, want to have a secret conversation. A generic, classical encryption protocol would begin when Alice and Bob convert their messages to separate binary streams of 0's and 1's, which are then encrypted and decrypted with a set of secret 'keys' known only to the two. Each key is a random string of 0's and 1's that is as long as the binary string comprising each message. To encrypt, Alice (the sender) sequentially adds each bit of the key to each bit of her message, using addition modulo 2. She then sends the encrypted message over a public channel to Bob, who decrypts it by simply repeating the addition modulo 2 of the key to the message. This type of encryption, known as a one-time pad, is currently the only provably secure encryption protocol. But the one-time pad is effective only if Alice and Bob never reuse the key, and more obviously, if the key remains secret. A potential eavesdropper, Eve, cannot be allowed to glean any part of the bit stream that makes up the key. Therein lies a central problem of cryptography – how can secret keys be created and then securely distributed?

Quantum key distribution (QKD) exploits the fundamentally indivisible nature of photons to perform this task. There are a variety of QKD protocols; here we describe one that employs entangled photon pairs (see Figure 3). Alice and Bob use a source such as the one described above to produce maximally entangled photons in the state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|HH\rangle + |VV\rangle) = \frac{1}{\sqrt{2}}(|45°,45°\rangle + |-45°,-45°\rangle)$. One photon goes to Alice and the other to Bob. For each pair, Alice and Bob randomly and independently analyze their respective photons (using a polarizing beamsplitter) in the H/V or the 45°/−45° basis. They record a bit value of 0 for all H or 45° results, and a 1 for all V or −45° results. After a sufficient number of measurements (dictated by the length of the key), Alice and Bob have a public discussion, e.g., over the internet. For each detected photon, they announce which basis they used for the measurement, but not the actual measurement result. Whenever they made the same basis choice (50% of the time), the correlations of the entangled state ensure their measured bit values agree. By contrast, they discard the results when they used different bases, because their measurements are completely uncorrelated (see Table 1).

An eavesdropper (Eve) cannot tap the line, as she might with conventional communications, because of the indivisibility of individual photons and the fact

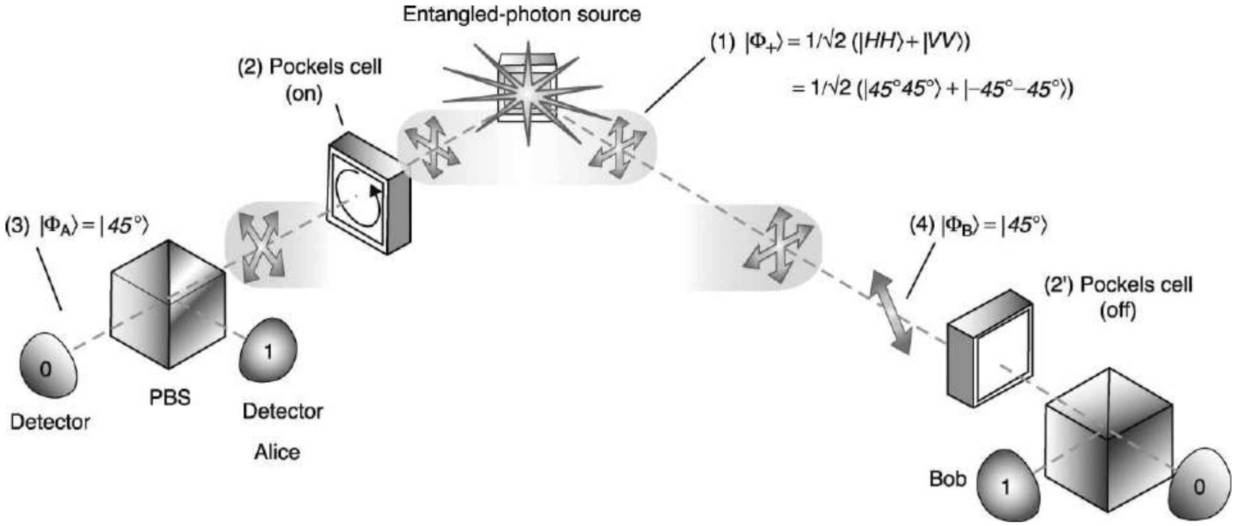
Entangled-photon source
(1) $|\Phi_+\rangle = 1/\sqrt{2}\,(|HH\rangle + |VV\rangle)$
$= 1/\sqrt{2}\,(|45°45°\rangle + |-45°-45°\rangle)$
(2) Pockels cell (on)
(3) $|\Phi_A\rangle = |45°\rangle$
(4) $|\Phi_B\rangle = |45°\rangle$
(2′) Pockels cell (off)
0
Detector
PBS
1
Detector
Alice
Bob
1
0

Figure 3 Quantum cryptography using entangled photons. Entangled photons can be used to create a pair of identical cryptographic keys. One member of an entangled pair (1) is sent to Alice, and the other to Bob. Alice and Bob each randomly and independently analyze their respective photon in one of two linear polarization bases with a polarizing beamsplitter (PBS); the basis can be actively chosen using a Pockels cell before the PBS, as shown, to rotate the polarization (alternatively, the 'choice' could be made by directing the photon onto a nonpolarizing 50/50 beamsplitter; in the transmitted path, one analysis basis is used, in the reflected path, the other is used). In the example shown, Alice used the 45°/−45° basis (2), and measured 45° polarization (3), thus projecting Bob's photon into the identical state (4). Since he chose the H/V analysis basis (2′), he is equally likely to detect a 0 or a 1. By subsequent public discussions Alice and Bob determine the events for which they used the same analysis basis (and discard the other events). For these events Alice and Bob will have obtained identical measurement results, which they may interpret as raw key material. After classical error correction and privacy amplification techniques are applied, the remaining string is the sought-after shared secret key. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

Table 1 Polarization entanglement-based quantum cryptography protocol

Alice's analysis basis	H/V	H/V	45°/−45°	H/V	45°/−45°	45°/−45°	H/V	45°/−45°
Alice's measurement result[a]	*H*	–	*−45°*	*V*	*45°*	*45°*	*V*	*−45°*
Alice's bit value	0	–	1	1	0	0	1	1
Bob's analysis basis	H/V	45°/−45°	45°/−45°	45°/−45°	H/V	45°/−45°	H/V	HV
Bob's measurement result	*H*	*−45°*	*−45°*	*45°*	*H*	–	*V*	*H*
Bob's bit value	0	1	1	0	0	–	1	0
Public discussion: Both photons detected and same basis used?	yes	no	yes	no	no	no	yes	no
Remaining secret key	0		1				1	

[a] In some cases Alice or Bob may not detect a photon due to loss or detector inefficiency. These events simply do not contribute to the key data.

that arbitrary quantum systems cannot be accurately cloned.[8] If Eve steals Bob's photon (a 'denial-of-service' attack), the photon's information never becomes part of the key. Thus, although a wiretap would reduce the rate of the transmission, it would not jeopardize the security of the key. Eve can try to intercept the photon, measure it, and send another one to Bob. But any measurement Eve would make to determine the photon's polarization state would perturb the photon and collapse the entangled state. The photon she sends to Bob would therefore only be 'classically' correlated with Alice's photon. Consequently, Eve's intervention necessarily induces additional errors into Bob's key, which Alice and Bob can detect by publicly revealing a small subset of their actual key. Unfortunately, even with no eavesdropper, the encryption keys created by any real-world quantum cryptography system typically possess a few-percent errors. To make sure their key is secure, Alice and Bob ascribe all errors to Eve and then estimate the maximum amount of information available to the eavesdropper. They then use a classical privacy amplification protocol to reduce Eve's knowledge of the secret key to less than one bit by reducing the length of the key. It has been proven that if the initial error probability per bit is greater than ~15%, no secret bits will remain after error detection and privacy amplification.

Researchers are working to make entanglement-based quantum cryptography more practical. Also, a number of longer distance quantum cryptography demonstrations using weak pulses have been carried out, over tens of kilometers in fibers and in free space. In fact, the first commercially available systems have recently been announced (in Europe and the United States).

Table 2 Method to convert the initial state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ into any of the four Bell states

Alice's operation	*Polarization transformation*	*Resultant two-qubit state*
$\hat{I}$, the identity	H → H; V → V	$|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$
$\hat{\sigma}_x$	H → V; V → H	$|\Psi_+\rangle = \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$
$i\hat{\sigma}_y$	H → V; V → −H	$|\Psi_-\rangle = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)$
$\hat{\sigma}_z$	H → H; V → −V	$|\Phi_-\rangle = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)$

Quantum Superdense Coding

It is possible for Alice to send Bob two bits of classical information using a single qubit in the quantum superdense coding protocol. Suppose that Alice and Bob share one qubit each of an entangled pair in the maximally entangled state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, where we have returned to the generic labeling scheme for the qubit states. Alice has two classical bits of information, which is equivalent to one of four choices. She can encode this information by applying one of four possible transformations on her qubit; a suitable set of transformations are the three Pauli matrices and the identity (i.e., do nothing). This set of operations can be performed experimentally on photon qubits quite easily, e.g., using waveplates, as indicated in Table 2. The four resultant two-qubit states are all maximally entangled and are orthonormal. They form a special basis for the two-qubit states, the Bell basis, which is of particular use in analyzing entanglement.

Alice now sends her qubit to Bob, e.g., through an optical fiber. Bob can then perform a Bell state analysis[9] on his qubit pair, i.e., a projective measurement of the two-qubit state in the Bell basis. The result immediately reveals the choice of operation Alice made, and in effect, two bits of classical information have been encoded on a single qubit.

---

[8] The no-cloning proof is very simple. If we have a copying operation such that $|0\rangle|c\rangle \to |0\rangle|0\rangle$ and $|1\rangle|c\rangle \to |1\rangle|1\rangle$ ($|c\rangle$ is the initial state of the copier), then by linearity, a superposition becomes an entangled state, instead of two copies: $\frac{(|0\rangle + |1\rangle)}{\sqrt{2}}|c\rangle \to \frac{(|0\rangle|0\rangle + |1\rangle|1\rangle)}{\sqrt{2}} \neq \frac{(|0\rangle + |1\rangle)}{\sqrt{2}}\frac{(|0\rangle + |1\rangle)}{\sqrt{2}}$.

[9] Complete discrimination of all four Bell states is presently an unsolved technical problem: for photons one needs either a nonlinear interaction (which is typically very weak) or to exploit so-called 'hyper-entanglement' involving other entangled degrees of freedom of the photons.

Quantum Teleportation

Another application of entanglement is quantum state teleportation (Figure 4), in which the infinite amount of information contained in an arbitrary qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is transferred by communication of two bits of classical information between Alice and Bob, if they also share two qubits in the maximally entangled state $|\Psi_-\rangle$. The three-photon initial state (i.e., the input photon plus the two entangled photons) is

$$|\psi_0\rangle = (\alpha|H\rangle + \beta|V\rangle) \otimes \frac{1}{\sqrt{2}}(|HV\rangle - |VH\rangle),$$

which can be rewritten with the first two photons (the input plus the first half of the entangled pair) represented by the Bell state basis:

$$\begin{aligned} |\psi_0\rangle = \ & \tfrac{1}{2}\{|\Phi_+\rangle \otimes (-\beta|H\rangle + \alpha|V\rangle) \\ & + |\Phi_-\rangle \otimes (\beta|H\rangle + \alpha|V\rangle) \\ & + |\Psi_+\rangle \otimes (-\alpha|H\rangle + \beta|V\rangle) \\ & + |\Psi_-\rangle \otimes (-\alpha|H\rangle - \beta|V\rangle)\}. \end{aligned}$$

Now suppose that a Bell state measurement is performed on the first two photons. The third photon

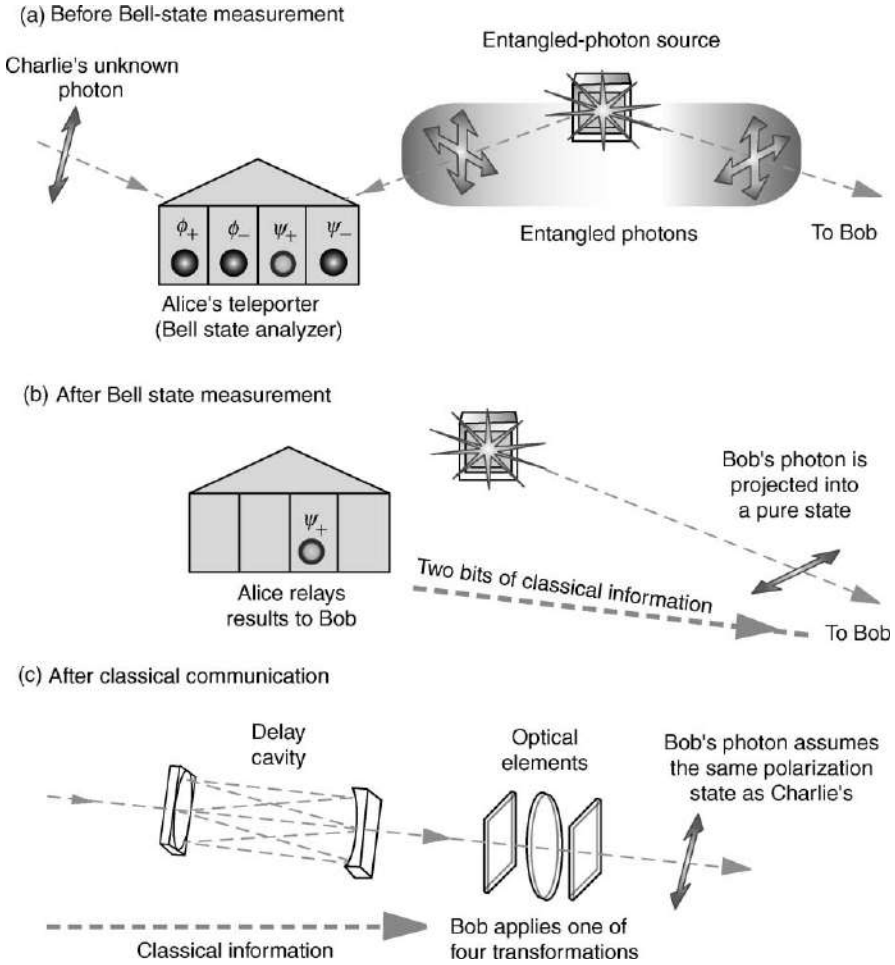
(a) Before Bell-state measurement
Charlie's unknown photon
Entangled-photon source
ϕ_+
ϕ_-
ψ_+
ψ_-
Entangled photons
To Bob
Alice's teleporter (Bell state analyzer)
(b) After Bell state measurement
ψ_+
Bob's photon is projected into a pure state
Alice relays results to Bob
Two bits of classical information
To Bob
(c) After classical communication
Delay cavity
Optical elements
Bob's photon assumes the same polarization state as Charlie's
Classical information
Bob applies one of four transformations

Figure 4 Quantum state teleportation. (a) Teleportation requires an entangled photon source, a Bell state analyzer and a classical communications channel. One entangled photon goes to Bob and the other to Alice, who also receives a photon of unknown polarization state. (b) Alice performs a joint polarization measurement of the two photons and relays the result to Bob using two classical bits of information. The photon going to Bob is projected into a pure state as a result of Alice's measurement. (c) Upon receiving Alice's classical information, Bob performs a simple transformation on his photon (which he has been storing), such as a rotation of the polarization vector. The resulting state of his photon is then identical to the unknown state Alice wished to teleport. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

is immediately projected into one of four possible states, which can be transformed back into the state of the original input photon by a simple operation, e.g., with a waveplate. For example, if the Bell state measurement produced the result $|\Psi_+\rangle$, the third photon is immediately 'collapsed' into the pure state $|\psi_3\rangle = -\alpha|H\rangle + \beta|V\rangle$. By applying a π-phase shift to the horizontal polarization (relative to the vertical), $|\psi_3\rangle$ can be transformed into the original input state: $|\psi'_3\rangle = \alpha|H\rangle + \beta|V\rangle$. Alice (who does the Bell state measurement) need only communicate to Bob (who wants to receive the teleported photon) which of the four Bell states she measured. Note that during the entire teleportation procedure, neither Alice nor Bob can obtain any idea of the values of the parameters α and β, which specify the state. Also, because Alice's measurement collapses the unknown state, there is only a single copy at the end of the protocol.

Quantum state teleportation was first demonstrated experimentally by Zeilinger and coworkers (University of Innsbruck, Austria). The group was able to determine two of the four Bell states unambiguously (the other two states gave the same experimental signature) and proved for those cases that the state of the input photon could indeed be transferred to Bob. Other experiments have realized modified forms of quantum teleportation in different systems. For example, Kimble's group (California Institute of Technology) teleported the coherent state of an optical mode using squeezed (rather than polarization-entangled) light. Researchers have recently suggested how teleportation might form the basis of a distributed network of quantum communication channels, and how it might enable quantum computing in all-optical systems.

Other Application of Optical Entanglement

Quantum Computing

The most challenging and powerful application of quantum information is large-scale quantum computing. Two features that make quantum information processing potentially powerful are the exponentially large Hilbert space, which gives quantum registers very large capacities, and quantum parallelism, which means that data processing tasks can be performed very efficiently. One fundamental drawback, which severely constrains useful applications of quantum computers, is that the final measurement can only produce a number of classical bits equal to the number of qubits. Thus, quantum computers are limited to performing tasks in which a small amount of information is meant to be gleaned from processing a large amount of data; examples include searching an unstructured database for a specific entry (Grover's algorithm) or finding the periodicity of a function (the quantum Fourier transform). This second task is central to Shor's factor-finding algorithm, the most famous quantum computing algorithm to date.

A practicable quantum computer technology must have at least the following features, first identified by DiVincenzo:

- A set of well-characterized, distinguishable qubits to form a quantum data storage register.
- The ability to initialize the qubits of the register in a simple fiducial state.
- Decoherence times that are much longer than the time needed to perform logical operations.
- The ability to perform any single qubit operation on any qubit in the register, and the ability to perform two-qubit conditional logic gates (such as the CNOT gate: $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle \rightarrow \alpha|00\rangle + \beta|01\rangle + \gamma|11\rangle + \delta|10\rangle$). Together, these operations constitute a universal set of quantum gates, from which all other gates can be synthesized.
- The ability to measure each qubit.

All-optical schemes have been used to implement small quantum algorithms. However, most of the approaches are not scalable, due to the limitations of non-linear optics to perform a CNOT gate at the single-photon level. Recent proposals have suggested that very high efficiency single-photon detectors, along with sources of single photons 'on-demand,' allow scalable quantum computing with only linear optics; preliminary two-qubit gates have been experimentally demonstrated. A number of other promising candidate technologies that meet DiVincenzo's five requirements are being pursued vigorously. For example, the ability to create multi-qubit entanglement and perform reliable measurements on trapped ions cooled and manipulated by lasers has recently been demonstrated. Solid state systems offer the possibility of scalability, and a number of schemes, such as quantum dots, isolated impurities with nuclear spins, and superconducting quantum interference devices (SQUIDs), are being investigated. It is possible that the final quantum computing technology may take the form of a hybrid between various present approaches.

Lithography

Lithography, in which a pattern is optically imaged onto some photoresistive material, is the primary method of manufacturing microscale or nanoscale electronic devices. An inherent limitation of this process is that details smaller than a wavelength of light cannot be written reliably. However, quantum state entanglement might circumvent this limitation. Under the right circumstances, the interference pattern formed by beams of entangled photons can have half the classical fringe spacing. Quantum lithography requires two beams of photons, a coherent superposition consisting of the state in which two photons are in beam A while none are in B, and the state in which no photon is in beam A while two photons are in B. Such number-entangled states can be made in the laboratory, and the predictions about fringe spacings have been verified. However, other obstacles must be overcome to surpass current classical-lithography techniques.

Two-Photon Imaging and Microscopy

At present, two-photon microscopy is widely used to produce high-resolution images, often of biological systems. However, the classical light sources (lasers) used for the imaging have random spreads in the temporal and spatial distributions of the photons, and the light intensity must be very high if two photons are to intersect within a small enough volume to cause a detectable excitation. Such high intensity can damage the system under investigation. Because the temporal and spatial correlations may be much stronger between members of an entangled photon pair, much weaker light sources could be used, which would be much less damaging to the systems being observed. The development of such systems is currently an active area of research.

List of Units and Nomenclature

Entanglement: a property of quantum systems consisting of two or more distinct subsystems, often separate particles. When transformations on one system do not affect the other, the state is said to be separable; when this is not the case, and the quantum state of the overall system cannot be resolved into separate states of the individual pieces, the system is entangled.

Pure and mixed states: when the probability amplitudes specifying a quantum state are deterministic, the state is said to be pure; when the amplitudes are random quantities, the state is mixed. The distinction is similar to that of coherent and partially coherent optical fields.

Quantum computer: an information processing device that exploits quantum mechanical phenomena to greatly enhance computational power for certain problems. Some mathematical problems thought to be intractable on conventional computers can, in theory, be performed efficiently on a quantum computer.

Quantum crytography (also known as quantum key distribution): a quantum information protocol by which two parties acquire a shared series of random numbers (a cryptographic key) by exploiting the quantum nature of light. Absolute security can be proven by virtue of the indivisibility and uncopy-ability of individual quanta.

Quantum state tomography: a means by which quantum states can be determined experimentally by performing a series of appropriate measurements on multiple identically prepared systems. From such measurements, the elements of the density matrix, which fully specifies the state, may be inferred.

Quantum superdense coding: A quantum information protocol by which two bits of classical information may be communicated by a single quantum bit, initially part of an appropriate entangled quantum state.

Quantum teleportation: a quantum information protocol by which the unknown quantum state of one particle can be transferred to another distant particle, using a pair of entangled particles, a projective measurement, and exchange of two bits of classical information.

Qubit (quantum bit): a two-level quantum mechanical system, which constitutes the building blocks of quantum information processing devices. Examples include a spin-1/2 particle, an atom with two well-isolated levels, or the polarization of a photon.

See also

Nonlinear Optics, Applications: Three-Dimensional Microfabrication. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Berman GP, Doolen GD, Mainieri R and Tsifrinovitch V (1998) *Introduction to Quantum Computers*. Singapore: World Scientific.

Bouwmeester D, Ekert AK and Zeilinger A (eds) (2000) *The Physics of Quantum Information*. Berlin: Springer-Verlag.

Braunstein SL and Lo HK (eds) (2001) *Scalable Quantum Computers: Paving the Way to Realization*. Berlin: Wiley-VCH.

DiVincenzo DP (2000) The physical implementation of quantum computation. *Fortschrcitte der Physik* 48: 771–783.

Gisin N, Ribordy G, Tittel W and Zbinden H (2002) Quantum cryptography. *Review of Modern Physics* 74: 145–195.

Kok P, Braunstein SL and Dowling JP (2002) Quantum lithography. *Optics and Photonics News* 13(9): 24–27.

Lo HK, Popescu S and Spiller T (eds) (1998) *Introduction to Quantum Computation and Information*. Singapore: World Scientific.

Macchiavello C, Palma GM and Zeilinger A (eds) (2000) *Quantum Computation and Quantum Information Theory*. Singapore: World Scientific.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Nielsen MA and Chuang IL (2000) *Quantum Computation and Quantum Information*. Cambridge, UK: Cambridge University Press.

Williams CP and Clearwater SH (1998) *Explorations in Quantum Computing*. Santa Clara: Telos.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Laser Cooling of Ions

H Walther, University of Munich and
Max-Planck-Institute for Quantum Optics, Garching,
Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fact that light can exert a mechanical force on atoms by the transfer of linear momentum was recognized as long ago as the 1930s, when Frisch demonstrated the deflection of an atomic beam using resonant light. With the advent of tunable laser light in the early 1970s, a much larger excitation rate for atoms was achieved, leading to a much stronger deflection. A little earlier it was also observed that in addition to the force resulting from photon recoil, a dipole force is also produced in a focused light beam so that even small solid particles can be trapped and manipulated. The original proposal for laser cooling of atomic gases was put forward in 1975 and later laser cooling of trapped ions was proposed. The principle of laser cooling is very simple. The ion is excited by laser light below a resonance frequency; in order that a transition can be made, the residual energy has to be taken from the kinetic energy of the particle, leading to cooling of the particle. In this article, we will discuss this cooling process in more detail.

Laser Cooling of Ions in a Radiofrequency Trap

Of the various proposed and realized methods for laser cooling of atoms or trapped ions, sideband cooling is conceptually the simplest and most efficient. It allows a trapped particle to be cooled to the quantum-mechanical ground state of a harmonic trap potential. The process can be regarded as anti-Stokes Raman scattering of the pseudomolecule formed by the ion and the trap. If the linewidth of an optical transition in this atom is smaller than the vibrational frequency of the trapped ion, both electronic levels of the transition split into a ladder of resolved vibrational levels. The absorption spectrum consists of a carrier at the resonance of the free atom and sidebands at multiples of the vibrational frequency. With a laser tuned to the m-th lower frequency sideband, each absorption of a photon will reduce the vibrational energy of the atom by m quanta. Since the subsequent spontaneous re-emission will, on average, not change the vibrational excitation, repeated photon scattering processes lead to efficient cooling of the atom. This cooling will continue until the absorption probability for the low-frequency sidebands decreases when approaching the vibrational ground state of the ion in the trap. The final temperature is limited by the heating process caused by photon recoil. However, after the ion is cooled to the vibrational ground state this heating process disappears, when the vibrational energy of the ion is larger than the recoil energy of a photon. In this case, the recoil is taken up by the entire trap setup, as it is the case in the Mössbauer effect.

Despite these obvious advantages of sideband cooling, relatively few experiments have been made using this method. This is because the regime, where the oscillation frequency ω is larger than the natural linewidth Γ of an optical transition (the so-called 'strong binding regime'), is not easily accessible. Vibrational frequencies of ions in standard electromagnetic traps are of the order of 1 MHz, whereas linewidths of electric dipole transitions in positive ions are in the range 20–50 MHz. In this 'weak binding regime' with $\Gamma > \omega$, laser cooling of trapped ions is essentially analogous to Doppler cooling of free atoms. The minimum temperature is given by the Doppler limit $k_B T = \hbar\Gamma/2$, and the mean vibrational quantum number is $\langle n \rangle = \Gamma/2\omega > 1$. The first experiments with cooling on resolved sidebands, therefore, used a strongly forbidden optical transition or stimulated Raman transitions. In these cases, however, sideband cooling is inevitably slow and it was used only to extract a few vibrational quanta from Doppler-precooled ions. In both experiments the vibrational ground state was reached to a very good approximation (mean quantum number $\langle n \rangle \approx 0.05$). Raman sideband cooling has also recently been applied to neutral atoms in optical traps.

Laser cooling of trapped ions was motivated by their possible use for optical frequency standards where the relativistic Doppler shift, as well as the Stark shift due to the trap potential, are eliminated at low temperatures. Cooling to the ground state also opens up the possibility of studying a well-controlled mechanical quantum oscillator experimentally. This led to the realization of nonclassical states of motion, investigation of quantum decoherence phenomena, and the demonstration of a quantum logic gate for quantum computing.

Sideband Cooling

Most theoretical treatments of sideband cooling consider a static harmonic trap, whereby the time dependence of the potential of the radio-frequency (rf) ion trap is neglected. In such a model a running laser wave $\propto \exp(\mathrm{i}kr)$ of wave number k induces transitions between the eigenstates $|n\rangle$ and $|n'\rangle$ of a harmonic oscillator. The relevant transition matrix elements are

$$U_{n'n} = \langle n' | \exp(\mathrm{i}kr) | n \rangle \qquad [1]$$

$$U_{n'n} = \exp(-\eta^2/2)\sqrt{\frac{n_<!}{(n_< + \Delta n)!}}(\mathrm{i}\eta)^{\Delta n} L_{n_<}^{\Delta n}(\eta^2) \qquad [2]$$

Here $\eta = kx_0 = k\sqrt{\hbar/2m\omega}$ is the Lamb–Dicke parameter for the spatial extension x_0 of the ground state, $n_<$ the smaller of the two numbers n and n', $\Delta n = |n - n'|$, and L_n^m denotes the generalized Laguerre polynomial. In the case of high vibrational excitation $n, n' \gg 1$, $U_{n',n}$ can be approximated as

$$U_{n+\Delta n,n} \approx \mathrm{i}^{\Delta n} J_{\Delta n}(2\eta\sqrt{n}) \qquad [3]$$

with the Bessel function of n-th order J_n. The argument of the Bessel function is the product of the wave number and the classical vibration amplitude of the highly excited state. This is the result one would expect for the frequency-modulated spectrum of a classical oscillator. For arguments β among the Bessel functions $J_n(\beta)$, those of order $n \approx \beta$ have the largest modulus. Consequently, during the initial stages of cooling it is most efficient to tune the laser so that it induces transitions with $\Delta n \approx 2\eta\sqrt{n} \gg 1$.

For the final stages of cooling one can usually assume that the ion is in the Lamb–Dicke regime, i.e., that it is localized within dimensions of the optical wavelength. A detailed study of the limits of sideband cooling in a static trap was carried out in 1984. The final distribution of population over the vibrational states is thermal and can be characterized by a mean occupation number $\langle n \rangle$ or a temperature $T = \hbar\omega/k_B \ln(1 + 1/\langle n \rangle)$. The number $\langle n \rangle$ is a rational function of the four frequencies involved, i.e., the natural linewidth Γ, the oscillation frequency in the trap ω, the laser detuning δ, and the Rabi frequency ω_R. For certain limiting cases, simple expressions for the mean vibrational excitation $\langle n \rangle$ can be derived from the general result. The lowest vibrational excitation is achieved in the case of well-resolved sidebands ($\omega \gg \Gamma$) and under weak laser excitation ($\omega_R \ll \Gamma$), tuned to the first sideband ($\delta = -\omega$):

$$\langle n \rangle = \left(\alpha + \frac{1}{4}\right)\frac{\Gamma^2}{4\omega^2} \qquad [4]$$

The constant α depends on the geometry of excitation and photon re-emission and is of the order of 1. In the case of strong saturation ($\omega_R \gg \Gamma$) of the first sideband ($\delta = -\omega$), the mean vibrational quantum number is proportional to the laser intensity:

$$\langle n \rangle = \frac{\omega_R^2}{8\omega^2} \qquad [5]$$

This case is of practical relevance if one wants to detect a fluorescence signal on the cooling transition, for example, to measure the internal state of the ion in a double-resonance experiment, while keeping the ion cooled. For the choice of parameters $\omega_R = \omega = -\delta$, the ion is still predominantly in the vibrational ground state ($\langle n \rangle \approx 1/8$), but photons are scattered from the wing of the saturation-broadened carrier at a rate $P = \Gamma\omega_R^2/(\Gamma^2 + 2\omega_R^2 + 4\delta^2) \approx \Gamma/6$, which is already one-third of the maximum scattering rate. Finally, in the case of large laser detuning ($\delta \gg \Gamma$, ω, ω_R), the energy is proportional to the detuning because the lower vibrational levels are no longer depleted by the laser:

$$\langle n \rangle \approx \frac{\alpha + 1}{4}\frac{\delta}{\omega} \qquad [6]$$

If the Lamb–Dicke parameter satisfies $\eta \ll 1$, the strength of the sidebands can be calculated from a power series expansion of the matrix elements of eqn [2] and the thermal distribution of population over the oscillator levels. The $U_{n',n}$ are evaluated to first-order in η only. In this approximation, the height of the first lower-frequency sideband in the absorption spectrum relative to the carrier is $\eta^2\langle n\rangle$ and that of the first higher-frequency sideband $\eta^2(\langle n\rangle+1)$. This information can be used to determine experimentally the vibrational quantum number from the absorption spectrum. The strengths of the higher-order sidebands contain higher powers of η^2 and therefore decrease rapidly.

Let us now briefly review some facts about the motional spectrum of an ion in a Paul trap, which is not just a static system but possesses an explicitly time-dependent potential. The equation of motion for a classical particle of charge e and mass m in the quadrupole potential of the trap $\phi(r,z) = (V_0 - U_0 \cos \Omega t)(r^2 - 2z^2)/2r_0^2$ is given by the Mathieu differential equation:

$$\frac{\delta^2 r}{\delta \tau^2} + (a_r - 2q_r \cos 2\tau)r = 0 \qquad [7]$$

with dimensionless parameters:

$$\tau = \frac{\Omega}{2}t, \qquad a_r = \frac{4eV_0}{m\Omega^2 r_0^2}, \qquad q_r = \frac{2eU_0}{m\Omega^2 r_0^2} \qquad [8]$$

Here a static voltage V_0 and an alternating voltage with amplitude U_0 and frequency Ω are applied to a Paul trap with the radius of the ring electrode r_0. The equation of motion for the z coordinate is obtained by multiplying the parameters a and q by the factor -2. The general solution of the Mathieu equation in the first stability region of the (a,q) parameter space can be written by using the Floquet ansatz:

$$r(\tau) = A\sum_{j=-\infty}^{\infty} c_j \cos(2j+\beta_r)\tau + B\sum_{j=-\infty}^{\infty} c_j \sin(2j+\beta_r)\tau \qquad [9]$$

where A and B are constants determined by the initial conditions, and the c_j and β_r depend on a and q. In the first stability region, where most of the experiments are carried out, the value of β_r is between 0 and 1. According to eqn [9] the motional spectrum of the ion has resonances at

$$\omega_{r,j} = \left(j \pm \frac{\beta_r}{2}\right)\Omega, \qquad j = 0, 1, 2\ldots \qquad [10]$$

In the adiabatic approximation, which is valid for $a, q^2 \ll 1$, the motion of the ion can be separated into an oscillation at the driving frequency Ω, the so-called micromotion, and a slower oscillation, called the secular motion, describing the oscillation in the time-averaged pseudopotential. This pseudopotential describes the effect of the ponderomotive force of the oscillating trapping field that drives the ion to the field minimum at the center of the trap. Since, in the quadrupole trap the pseudopotential is harmonic, the solution to the equation of motion can be approximated as

$$r(\tau) = r_0 \cos\beta_r\tau\left(1 + \frac{q_r}{2}\cos 2\tau\right) \qquad [11]$$

The approximate expression for β_r is $\beta_r = \sqrt{q_r^2/2 + a_r}$ and the secular frequencies are given by

$$\omega_r = \frac{\Omega}{\sqrt{8}}\sqrt{q_r^2 + 2a_r}, \qquad \omega_z = \frac{\Omega}{\sqrt{2}}\sqrt{q_r^2 - 2a_r} \qquad [12]$$

which are the lowest-order resonances $\omega_{r,z} = \beta_{r,z}\Omega/2$ in the Floquet ansatz, eqn [9]. From eqn [11] it is seen that the amplitude of motion at the resonances $\Omega \pm \omega$ is a factor $q_r/4$ smaller than those at the resonance ω. It is noteworthy that the frequency Ω of the driving field does not appear directly in the motional spectrum, but only in combination with the secular frequency ω. This changes, however, if there is an additional static force acting on the ion, displacing it from the center of the quadrupole. The approximate solution to the equation of motion then becomes:

$$r(\tau) = r_1 + r_0 \cos\beta\tau\left(1 + \frac{q_r}{2}\cos 2\tau\right) \qquad [13]$$

and a resonance at Ω appears. The force that displaces the ion by r_1 may be due to static electric stray fields or the presence of a second ion in the trap. These effects can have a strong influence on sideband laser cooling of the trapped ions. A number of theoretical studies have been published on the quantum-mechanical treatment of the Paul trap. The explicitly time-dependent potential does not allow stationary states with time-independent energy eigenvalues. One can, however, as in the case of the pseudopotential model, distinguish between two time-scales and construct wave functions of the harmonic secular oscillator that show some additional breathing motion at the trap frequency. A quantum-mechanical theory of sideband cooling in the time-dependent potential of the Paul trap thus far exists only for special cases; in 1994 there was a case where the ion was localized in the Lamb–Dicke regime at the node of a standing laser wave and time averaged kinetic energies were calculated. Minima of the energy that agree with the results of the theory for the static harmonic trap to within a factor of 2 were found. The strongest modifications to the predictions of the cooling theory for the static trap arise in the dependence of the energy on the laser detuning. The influence of the micromotion gives rise to heating for certain regions of negative laser detuning (on the high-frequency side of a low-frequency micromotion sideband) and to cooling for some positive detunings (on the low-frequency side of a high-frequency micromotion sideband).

Trapped LASER-cooled ions have a variety of applications in spectroscopy and other fields.

In the following, a phenomenon will be discussed where observations during the cooling process of ions in traps is interpreted as phase transitions between chaotic and ordered ion structures during the cooling process.

Generation of Chaos During the Cooling Process and Phase Transitions to Ordered Configurations

Using a Paul trap, it is possible to study the physics of a few-body phase transition of ions. The ordered state of the trapped ions can be observed directly by monitoring their resonantly scattered light. The transition to an ordered or 'crystalline' configuration is expected when the Coulomb coupling constant which corresponds to the ratio of Coulomb energy to the kinetic energy of a particle, is larger than 1. The ordered structures represent configurations of maximum density for the ions, they are determined by the trap potential pushing the ions towards the trap center, and by the Coulomb repulsion between the ions acting in the opposite direction, so that minimum energy configurations are finally adopted. The system corresponds to an ordered one-component plasma or a strongly coupled plasma.

In the following, computer simulations of the phase transition in the Paul trap will be reviewed. In the trap the ions are subjected to five different forces. These are the force $\mathbf{F}^{(\text{trap})}$, arising from the dynamical rf trapping field, the Coulomb interaction between the ions $\mathbf{F}^{(\text{Coul})}$, the laser cooling force $\mathbf{F}^{(\text{laser})}$, a random force $\mathbf{F}^{(\text{rand})}$ arising from the recoil of the spontaneously emitted photons, and a parasitary force $\mathbf{F}^{(\text{cont})}$, arising from a contact potential, which may be generated unintentionally by locally coating the electrodes with the trapped atoms.

The potential generated by the rf voltage applied to the hyperbolic electrodes of the trap is given by

$$\begin{aligned} \phi(\mathbf{r},t) &= \varphi(t)[x^2+y^2-2z^2] \\ \varphi(t) &= \frac{U_0+V_0\cos(\Omega t)}{r_0^2+2z_0^2} \end{aligned} \qquad [14]$$

where U_0 and V_0 are the dc and ac components of the voltages applied to the trap. From eqn [1] we obtain the trapping force and the resulting equation of motion of a single ion in the trap:

$$\mathbf{F}^{(\text{trap})} = m\ddot{\mathbf{r}} = \mathbf{e}\nabla\phi(\mathbf{r},t) = -2\mathbf{e}\varphi(t)[\mathbf{r}-3z\mathbf{e}_z] \qquad [15]$$

where m, $\mathbf{r}$, and $z\mathbf{e}_z$ denote the mass of the ion, the position, and the $z\mathbf{e}_z$-component of the position vector, respectively.

The Coulomb interaction between the particles results in the force:

$$\mathbf{F}_i^{(\text{Coul})} = \frac{e^2}{4\pi\varepsilon_0}, \sum_{m\neq 1}\frac{\mathbf{r}_i-\mathbf{r}_m}{[\mathbf{r}_i-\mathbf{r}_m]^3} \qquad [16]$$

In the process of laser cooling, every scattered photon changes the momentum of an ion on the average by an amount $\hbar k$.

The ensemble average of the resulting laser cooling force acting on the i-th ion is thus:

$$\mathbf{F}_i^{(\text{laser})} = \hbar k R[\mathbf{r}_i(t);\,\dot{\mathbf{r}}_i(t)] \qquad [17]$$

The number $N_i(l)$ of spontaneously emitted photons from the i-th ion during the l-th cycle of period $T = 2\pi/\Omega$ is

$$N_i(l)\int_{(l-1)T}^{lT} R[\mathbf{r}_i(t);\,\dot{\mathbf{r}}_i(t)]\mathrm{d}t \qquad [18]$$

In the simulations, it is assumed that these N_i spontaneous photons are emitted one by one, at the end of each rf cycle, rather than emitting them one by one, according to the appropriate photon-statistics during the cycle. The orientation of the unit vector $\hat{\mathbf{u}}_j$, which points in the direction of the j-th emitted photon, is chosen at random, but weighted with a $\cos^2$ distribution (dipole characteristic) with respect to the laser polarization axis. This way, only the linearly polarized spontaneously emitted photons have been simulated in the calculations. The circularly polarized ones, which occur with a $\sin^2$ distribution, have been neglected. The random force $\mathbf{F}_i^{(\text{rand})}$, is thus:

$$\mathbf{F}_i^{(\text{rand})}(t) = \hbar k\sum_{l=-\infty}^{\infty}\delta(t-lT)\sum_{j=1}^{N_i}\hat{\mathbf{u}}_j \qquad [19]$$

In some of the calculations, a contact potential has been taken into account which is generated experimentally by a coating of the quadrupole electrodes by the trapped particles. Close to the center of the trap, the contact potential gives rise to an approximately homogeneous electric field. Altogether, the motion of the i-th particle is governed by the sum of the above forces, that is:

$$m\ddot{\mathbf{r}}_i = \mathbf{F}_i^{(\text{trap})} + \mathbf{F}_i^{(\text{Coul})} + \mathbf{F}_i^{(\text{laser})} + \mathbf{F}_i^{(\text{rand})} + \mathbf{F}_i^{(\text{cont})} \qquad [20]$$

Given appropriate initial conditions, the eqns of motion [20] are integrated forward in time to obtain the trajectories of the n-ion system. These trajectories provide all the information necessary to find the quantities of interest, such as the occupation probability density of the i-th ion in phase space (i.e., the probability density of finding the i-th ion at a given position, and with a given velocity), the (total) kinetic energy, and the fluorescence intensity.

Some of the results of the described simulations will be summarized below. Figure 1 shows the excitation spectrum of the ions as a function of laser detuning. The experiment confirms the results of the simulations.

Owing to the Coulomb interaction of the ions at particular phases of the cooling process, chaos in the ion cloud is observed leading to a heating process of the ions by the applied field of the trap. This heating mechanism was investigated in the simulations in detail and the summary of the results is given in Figure 2.

For zero laser power and large r, there is no net heating of the ions. This is confirmed by the experiments, in which, even in the absence of a cooling laser, large clouds of ions can be stored in a Paul trap over several hours without being heated out of the trap. The ions are far apart, the Coulomb force is small, and on short time-scales the ions behave essentially like independent singly stored ions. For this reason, this part of the heating diagram is called the 'Mathieu regime' as it is well described by the Mathieu equation [7]. Turning on a small laser power, the rms radius r reduces drastically due to laser

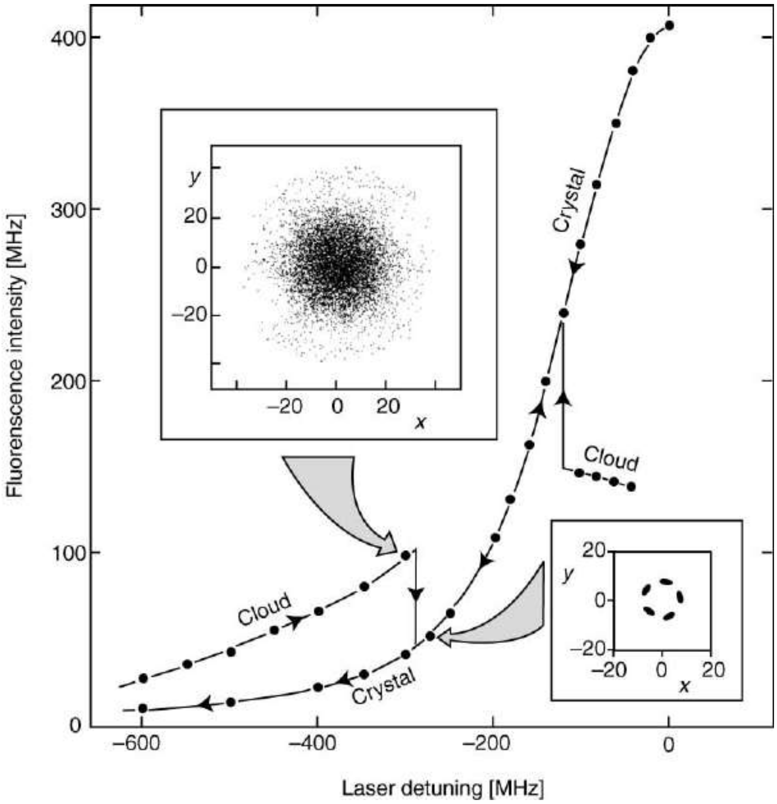

Figure 1 Excitation spectrum of five ions as a function of laser detuning. When the laser detuning is changed the cooling rate of the ion is varying. Two phase transition points (and bistability) are apparent. Insets show the results of the molecular dynamics calculations shortly before and after the jump, which demonstrates directly the existence of a transition from a cloud phase to a crystalline state. The units given at the insets are μm.

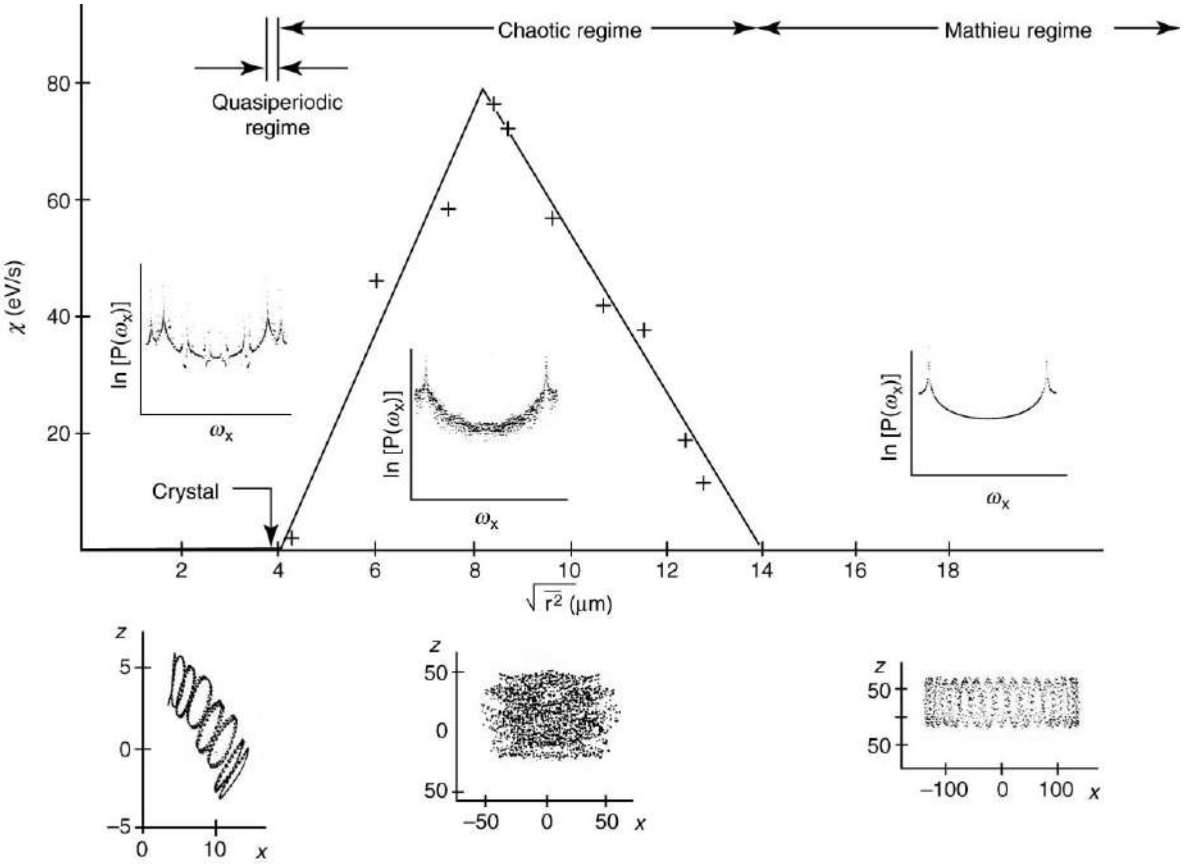

Figure 2 Average heating rate of five ions in a Paul trap versus the rms radius of the ion configuration. The insets show the power spectrum and the corresponding stroboscopic Poincaré sections in the x–z plane of relative separation for two ions in three characteristic domains: the quasiperiodic regime, the chaotic regime and the Mathieu regime. All length scales are in units of micrometers. Units gives at the x–z axis of the insets are in μm.

cooling, but comes to a halt at about 14 μm where chaotic rf heating sets in and balances the cooling power of the laser. Increasing the laser power results in an even smaller cloud. The smaller cloud produces more chaotic rf heating, as seen clearly by the negative slope of the heating curve (Figure 2) in the range $8 < r < 14$ μm.

In the range 4 μm $< r <$ 8 μm, there is still chaotic heating, but the slope of the heating curve is positive. As a consequence of the resulting triangular shape of the heating curve, at about 8 μm, the chaotic heating power can no longer balance the cooling power of the laser, and the cloud collapses into the 'crystalline' state located at $r \approx 3.8$ μm; a 'phase transition' has occurred. Due to this collapse, the behavior of the heating rate in the range 3.8 μm $< r <$ 8 μm cannot be studied by equilibrating laser cooling and rf heating. In this case, we start out from the crystal state and slightly displace the ions to explore the 'vicinity' of the crystal. We observe no heating for 3.8 μm $< r <$ 4 μm, but quasiperiodic motion, and thus name this regime the 'quasi-periodic' regime. We call the upper edge of the quasiperiodic regime ($r \approx 4$ μm) the 'chaos threshold'. An initial condition beyond the chaos threshold, i.e., satisfying $r > 4$ μm, leads to heating and expansion of the ion configuration and numerical data relevant for the shape of the heating curve can be taken during this 'explosion phase'. The laser power P is set to zero for this type of experiment. We conjecture that, apart from the trivial case of a single stored ion, the heating curve is 'universal', i.e., its qualitative shape, including the existence of the chaotic regime, does not depend on the number of simultaneously trapped ions.

For the quasiperiodic, the chaotic and the Mathieu-regime, respectively, the corresponding type of power spectrum are shown in the insets above the abscissa of Figure 2. The data were actually taken from the case of two ions, but would look similar in the five-ion case. A discrete spectrum in the quasi-periodic regime and a complicated

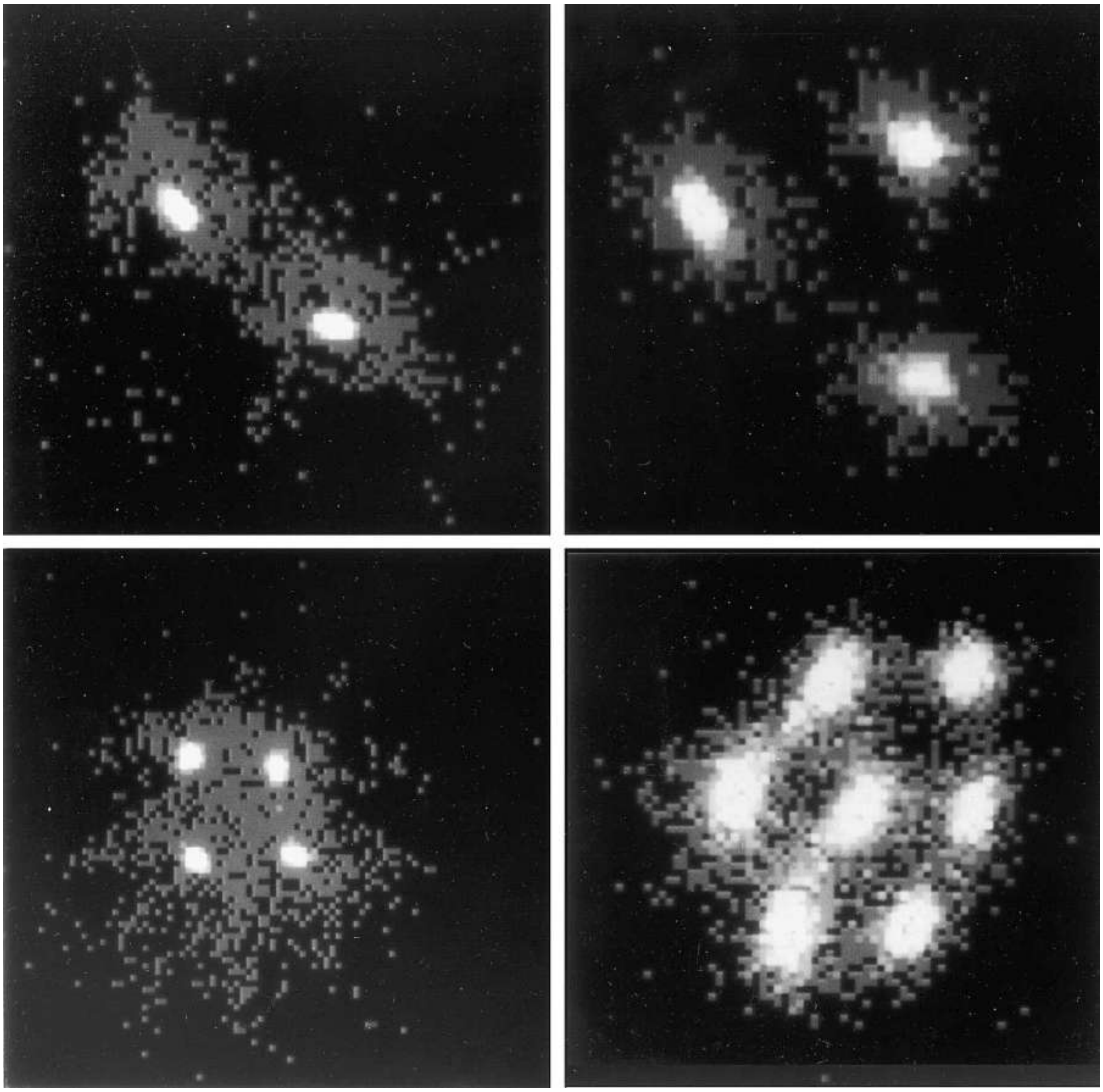

Figure 3 Ordered ion configuration of two, three, four, and seven laser-cooled Mg^+ ions confined in a Paul trap and observed using resonantly scattered light. The average ion separation is 20 μm. The structures of four and seven ions are slightly deformed owing to contact potentials in the trap.

noisy spectrum in the chaotic regime is obtained. The spectrum in the Mathieu regime is again quite simple and dominated by the secular motion frequency. Stroboscopic pictures of the locations of two ions in the x–y plane of the trap characterizing the three regions, are shown on the insets below the abscissa of Figure 2.

Some of the ordered ion structures observed experimentally are shown on Figures 3 and 4. The structures in Figure 3 were obtained in a Paul trap; the others are measured in a storage ring leading to quasi-linear structures of the ion configurations, depending on the symmetry of the trap. Ordered ion structures have also been observed in Penning traps. Due to the symmetry of this trap rotationally symmetric structures are observed.

Conclusion

Laser cooling of ions in traps provides an excellent tool for numerous experiments, varying from tests of basic physics to important and new applications in quantum information processing. The linear ion structure shown on Figure 4 is presently investigated in many laboratories in order to realize gates for

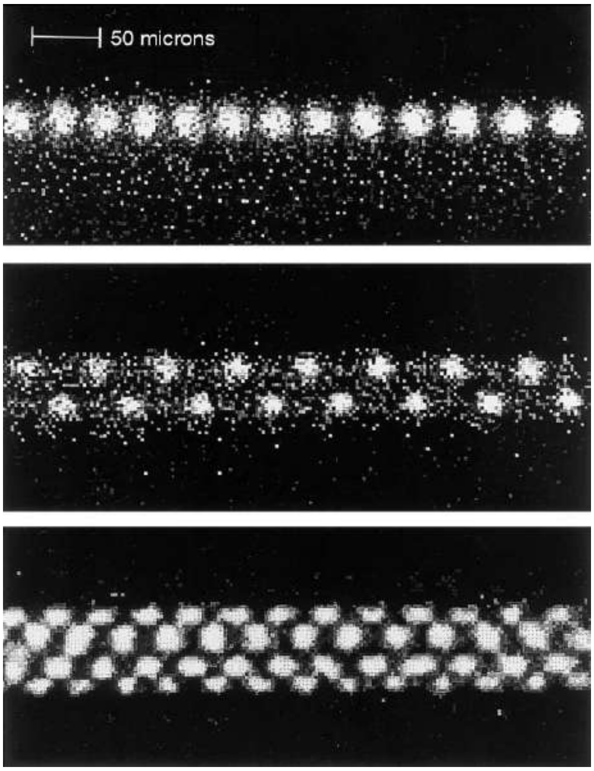

Figure 4 Crystalline structures of laser-cooled $^{24}Mg^+$ ions in the quadrupole storage ring. At a low ion density the ions form a string along the field axis (upper). Increasing the ion density transforms the configuration to a zigzag (middle). At still higher ion densities, the ions form ordered helical structures on the surface of a cylinder, e.g., three interwoven helices (lower). As the fluorescent light is projected onto the plane of observation in this case the inner spots are each created by two ions seated on opposite sides of the cylindrical surface, resulting in a single, bright area. (Reproduced by permission of *Nature* from Birkl G, Kassner S and Walther H (1992) Multiple shell structures of laser cooled $^{24}Mg^+$ ions in a quadrupole storage ring. *Nature* 357: 310–313.)

quantum computing. Single trapped ions in cavities may be used as deterministic light sources for secure single photon communication links. Furthermore, in many laboratories single laser-cooled ions are investigated as the basis for new time and frequency standards, being several orders of magnitude more accurate than our present atomic clock.

See also

Chaos in Nonlinear Optics. **Quantum Optics:** Entanglement and Quantum Information; Quantum Computing with Atoms.

Further Reading

Arimondo E, Philips WD and Strumia F (eds) (1992) Laser manipulation of atoms and ions. *Proceedings of the International School of Physics Enrico Fermi, Course CXVIII*. Amsterdam, Oxford, New York, Tokyo: North Holland.

Ashkin A (1970) Atomic-beam deflection by resonance-radiation pressure. *Physical Review Letters* 25: 1321–1324.

Birkl G, Kassner S and Walther H (1992) Multiple shell structures of laser cooled $^{24}Mg^+$ ions in a quadrupole storage ring. *Nature* 357: 310–313.

Blümel R, Chen JM, Peik E, Quint W, Schleich W, Shen YR and Walther H (1988) Phase transitions of stored laser-cooled ions. *Nature* 334: 309–313.

Blümel R, Kappler C, Quint W and Walther H (1989) Chaos and order of laser-cooled ions in a Paul trap. *Physical Review A* 40: 808–823.

Bollinger JJ, Kriesel JM, Mitchell TB, King LB, Jensen MJ, Itano WM and Dubin DHE (2003) Laser-cooled ion plasmas in Penning traps. *Journal of Physics B: Atomic Molecular Optical Physical* 36: 499–510.

Cirac JI, Garay LJ, Blatt R, Parkins AS and Zoller P (1994) Laser cooling of trapped ions: the influence of micromotion. *Physical Review A* 49: 421–432.

Dehmelt H (1967 and 1969) In: Bates DR and Estermann I (eds) *Advances in Atomic and Molecular Physics*, vols. 3 and 5. New York: Academic Press.

Dehmelt HG (1983) Stored ion spectroscopy – three lectures. In: Arecchi FT, Strumia F and Walther H (eds) *Advances in Laser Spectroscopy*. New York: Plenum.

Diedrich F, Peik E, Chen JM, Quint W and Walther H (1987) Observation of a phase transition of stored laser-cooled ions. *Physical Review Letters* 59: 2931–2934.

Glauber R (1991) In: Arimondo E and Phillips W (eds) *Proceedings of the International School of Physics 'Enrico Fermi,'* vol. 118. Amsterdam: North Holland.

Hänsch TW and Schawlow AL (1975) Cooling of gases by laser radiation. *Optics Communications* 13: 68–69.

Javanainen J, Lindberg M and Stenholm S (1984) Laser cooling of trapped ions: dynamics of the final stages. *Journal of Optical Society of America B* 1: 111–115.

Leibfried D, Blatt R, Monroe C and Wineland D (2004) Quantum dynamics of single trapped ions. *Review of Modern Physics* 75: 281–324.

Peik E, Abel A, Becker Th, von Zanthier J and Walther H (1999) Sideband cooling of ions in radio-frequency traps. *Physical Review A* 60: 439–449.

Schieder R, Walther H and Wöste L (1972) Atomic beam detection by the light of a tunable dye laser. *Optics Communications* 5: 337–340.

Wineland D and Dehmelt H (1975) Proposed 10–14 laser fluorescence spectroscopy on Tl^+ mono-ion oscillator. *Bulletin of the American Physical Society* 20: 637.

Quantum Computing with Atoms

S F Huelga, University of Hertfordshire, Hatfield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is a well-known fact that the speed of commercial microprocessors is doubling, roughly, every 18 months (Moore's law). This speedup is achieved mainly at the expense of reducing the size of the microchip. If this tendency is to continue, it seems unavoidable that quantum mechanical effects will eventually become important.

Remarkably, far from being a technological nuisance, the fact that information carriers are truly quantum mechanical systems offers an entirely new paradigm for computation. As opposed to classical bits, with logical values 0 and 1, quantum bits (qubits) can be prepared in arbitrary, normalized superpositions of the logic values. In addition, quantum states of multiqubit systems can be entangled, i.e., cannot be written as tensor products of states of the individual components. The theoretical description of entanglement requires a huge number of parameters which, in fact, grows exponentially with the number of particles that are involved. This exponential number of new degrees of freedom is responsible for the computational potential that quantum systems offer. Recently it has been demonstrated theoretically how we can harness this potential in the form of quantum algorithms, including Shor's factorization algorithm and Grover's protocol for data base search.

From a theoretical point of view, the implementation of quantum computation is greatly simplified by the fact that any unitary transformation involving an arbitrarily large number of qubits can be decomposed into two building blocks: single qubit operations and two-qubit controlled-NOT gates (CNOT-gates). We only need to identify two-level systems whose interaction can be mediated in a way that allows for conditional dynamics, that is, the second system undergoes certain operation (a NOT operation in our case), conditional on the state of the first qubit. However, it is easy to convince ourselves that from the experimental point of view, the challenge is going to be formidable. The following (DiVincenzo)-checklist summarizes the necessary requirements for the potential physical realization of a quantum computer:

- The system needs to provide a well-defined qubit, with two robust levels to represent the logical values 0 and 1. Moreover, the system should be scalable.
- Both the initial preparation and the final readout of the system has to be done with sufficient accuracy.
- The system has to allow the implementation of a universal set of gates (single-qubit rotations and CNOTs gates) with sufficient high fidelity. Implicit in this requirement is the ability to address individual qubits.

From the above list it is clear that the practical realization of a quantum computer relies on the ability to achieve a delicate equilibrium: the system needs to interact strongly with selected external agents, typically laser light, both during the gate operations and the preparation and readout steps, but should be shielded from any unwanted interaction if the desired dynamics is to be preserved. Perfect, selective, shielding from the environment is not possible in practice and unwanted interactions with the surroundings will unavoidably result in some form of decoherence. In this article we will review the fundamentals of the implementation of a universal set of quantum gates using (i) atomic system coupled to a resonant field sustained in a microwave cavity and (ii) trapped ions confined in an electromagnetic trap. Atomic and ionic systems provide a good means for storing quantum information in long-lived electronic states. In addition, both neutral atoms and ions can be selectively detected with high efficiency. While none of the presented schemes is expected to be scalable to a very large number of qubits, they constitute the first prototypes for quantum registers and are expected to play an important role in applications where only a moderate number of qubits are required. These applications range from the development of new frequency standards to quantum communication protocols. We will discuss in the following sections how a set of universal gates can build within these systems and will present the state of the art of the experiments performed to date.

Quantum Computation: Basic Definitions

As stated before, the basic unit of quantum information is a qubit, an arbitrary quantum mechanical two-level system. We will associate logical values 0 and 1 with quantum states $|0\rangle$ and $|1\rangle$, the eigenvectors of an observable such as the energy, and refer to such states as the computational basis. Any unitary operation on a qubit will be called a single qubit gate. Joint operations on more than one qubit will be denoted as multi-qubit gates. Using the fact that any two-level system is isomorphic to a spin-1/2 particle (e.g., an electron), we can always think of a qubit in terms of a certain pseudo-spin whose dynamics can be represented as rotation of the associated Bloch vector within the Bloch sphere.

For instance, as depicted in Figure 1, a qubit initially prepared in state $|1\rangle$, after experiencing a single qubit gate will be represented by certain Bloch vector $\vec{S}$, whose location on the Bloch sphere is a function of the gate parameters. Note that a classical system can only assume two positions (upwards and downwards), while a quantum system can explore the whole surface of the sphere.

It is also quite standard in quantum computation to use a diagramatic representation in terms of quantum circuits. Quantum circuits are built up of wires to denote the quantum bits and boxes to represent the gates, or quantum operations. Figure 2 shows two basics elements of a quantum circuit. The element on the left is the so-called Hadamard transformation, a particular example of single-qubit gate which transform each element of the computational basis as follows:

$$\begin{aligned} |0\rangle &\rightarrow \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right) \\ |1\rangle &\rightarrow \frac{1}{\sqrt{2}}\left(|0\rangle - |1\rangle\right) \end{aligned} \qquad [1]$$

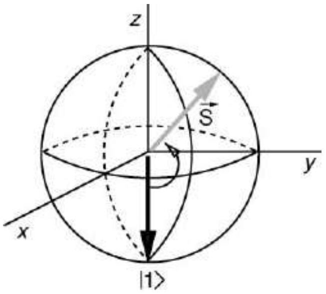

Figure 1 Geometrical representation of qubit systems using the Bloch sphere formalism. See the text for details.

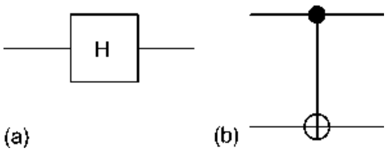

Figure 2 Examples of single qubit gates (a) Hadamard transformation and two qubit gates (b) CNOT gate as customarily represented in quantum circuits diagrams.

Part (b) of Figure 2 depicts a CNOT-gate. This is now an operation involving two qubits. The first one, the control qubit, remains unaltered after the gate operation while the second qubit, referred to as the target, undergoes a spin flip (logical NOT operation) only when the control is in the state $|1\rangle$. Specifically, the action of a CNOT-gate is defined by the following transformation rule:

$$\begin{aligned} &\text{CNOT} \\ |00\rangle &\rightarrow |00\rangle \\ |01\rangle &\rightarrow |01\rangle \\ |10\rangle &\rightarrow |11\rangle \\ |11\rangle &\rightarrow |10\rangle \end{aligned} \qquad [2]$$

We have already mentioned that single qubit rotations and CNOT gates provide a universal set of gates. Our aim now is to show how such operations can be built when the so far generic qubit are realized in terms of real atomic systems. In the two experimental scenarios we are going to analyze here, the CNOT operation will be achieved by means of a different type of gate, a controlled-phase gate supplemented by two one-qubit Hadamard gates on the target qubit. The arrangement is showed in Figure 3. The controlled-phase gate is again a two-qubit operation defined by the following transformation rule:

$$\begin{aligned} &\text{CNOT} \\ |00\rangle &\rightarrow |00\rangle \\ |01\rangle &\rightarrow |01\rangle \\ |10\rangle &\rightarrow |10\rangle \\ |11\rangle &\rightarrow -|11\rangle, \end{aligned} \qquad [3]$$

i.e., a global phase equal to π is acquired if, and only if, the two basis states are in the logical state 1. It is easy to check that when the target bit is rotated by a Hadamard gate prior and subsequently to the action

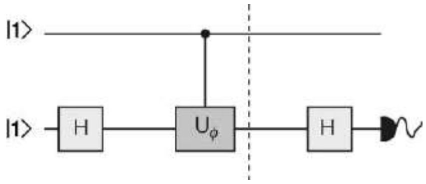

Figure 3 Controlled-phase gate with both inputs, control, and target qubits, in state $|1\rangle$. The first Hadamard gate rotates the target to state $(|0\rangle - |1\rangle)/\sqrt{2}$. The action of the controlled-phase gate with $\phi = \pi$ yields the joint state $|1\rangle + (|0\rangle - |1\rangle)/\sqrt{2}$ and a final Hadamard rotation on the target leaves the system in the product state $|1\rangle|0\rangle$.

of the joint phase gate, the resulting state is tantamount to the action of a single CNOT gate. Therefore, to prove that our atomic system allows for the implementation of a universal set of gates, it suffices to show that both single qubit rotations and controlled phase gates can be implemented with sufficient accuracy.

The next section analyzes a typical cavity QED setup and presents the basic physical ideas to understand how single and two-qubit gates can be implemented within atomic systems interacting with radiation fields. Subsequent sections will rely heavily in this analysis.

Quantum Computing with Atoms I: Cavity QED

A simplified cavity-QED arrangement for performing both single- and two-qubit gate operations is depicted in Figure 4. This configuration is well-known to quantum opticians and it is often referred to as a Ramsey interferometric setup. For concreteness, we will focus here in the experimental set up currently in operation at the ENS in Paris. There, the atomic system is a highly excited Rydberg atom initially prepared in a long-lived circular state. The relevant atomic states are represented in Figure 5, with the qubit being stored in levels $|i\rangle$ and $|g\rangle$ which will play the role of the general basis states $|0\rangle$ and $|1\rangle$.

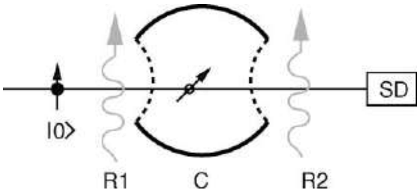

Figure 4 Cavity QED setup for single-qubit and two-qubit operations. Rotations of the atomic pseudo spin are performed via interation with classical radiation (Zone R), while the interaction with a resonant cavity field sustained in C allows the realization of a phase gate where the cavity photon acts as the control qubit.

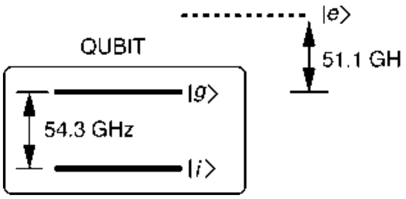

Figure 5 Circular Rydberg levels involved in the ENS experimental setup. The qubit is stored in levels $|i\rangle$ (logical value 0) and $|g\rangle$ (logical value 1). The radiative lifetime of these levels lies in the millisecond range. Dissipation can then be neglected along the 20 cm path inside the apparatus.

Regions denoted by R represent the interaction of the atomic system with classical microwave radiation. As detailed below, this interaction will allow arbitrary rotations of the atomic pseudo spin. Region C represents the interaction of the atomic system with a single mode cavity field resonant the atomic transition $g - e$. The highest Q value reported to date is of the order of 10^8, which corresponds to a photon storage time of a few milliseconds. Therefore, the relaxation time of the cavity field is much longer than the atom-cavity interaction time, which defines the two-qubit gate operation time and is of the order of a few tens of μs. The cavity field can be prepared in either of the Fock states $|0\rangle_\gamma$ or $|1\rangle_\gamma$ and will play the role of the control qubit for the realization of a phase gate, as detailed below. This set up is completed with a detection scheme (SD in the figure) that allows the selective detection of the atomic electronic state.

We will start our discussion showing how to perform an arbitrary single-qubit gate by means of irradiating the atomic system with classical radiation. We will keep the analysis in very general terms with the aim of using it later in the description of the implementation of quantum logic with trapped ions.

Single-Qubit Rotations

Let our qubit be a two-level atomic system with frequency separation ω_0. We will show that an arbitrary rotation can be achieved by means of irradiating the system with classical light, e.g., microwave radiation or a laser field, with tunable frequency ω_L. The radiation field will be specified by two controllable parameters, its Rabi frequency Ω and its phase ϕ. The atom-field Hamiltonian can be written as

$$\begin{aligned} H = {} & \frac{\hbar\omega_0}{2}\Big(|1\rangle\langle 1| - |0\rangle\langle 0|\Big) \\ & + \hbar\Big(g e^{i\omega_L t}|0\rangle\langle 1| + g^* e^{-i\omega_L t}|1\rangle\langle 0|\Big) \end{aligned} \qquad [4]$$

where the first term corresponds to the atomic internal energy and the second describes the coupling with the external field. The coupling constant g can be written in terms of the field parameters as

$$g = \frac{i\Omega}{2} e^{-i\phi} \qquad [5]$$

In order to write down a time-independent Hamiltonian, we will move to an interaction picture with respect to $H_0 = \hbar\omega_L/2(|1\rangle\langle 1| - |0\rangle\langle 0|)$. In this frame,

$$H_{\text{int}} = \frac{\hbar\delta}{2}\Big(|1\rangle\langle 1| - |0\rangle\langle 0|\Big) + \hbar\Big(g|0\rangle\langle 1| + g^*|1\rangle\langle 0|\Big) \quad [6]$$

The interaction Hamiltonian can be readily expressed in terms of the Pauli spin operators

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad [7]$$

as

$$H_{\text{int}} = \frac{\hbar\delta}{2}\sigma_z + \frac{\hbar\Omega}{2}(\sin\phi\,\sigma_x - \cos\phi\,\sigma_y) \quad [8]$$

where $\delta = \omega_L - \omega_0$ denotes the detuning from the atomic transition and we have taken into account the explicit form of g as given by eqn [5]. We can now easily derive the time evolution operator by noting that H_{int} can be written in a compact form as $H_{\text{int}} = \vec{\sigma}\cdot\vec{n}$, where the vector $\vec{n}$ has cartesian coordinates $\vec{n} = (\Omega/2\sin\phi, -\Omega/2\cos\phi, \delta/2)$. Note that $|\vec{n}| \equiv \Omega_g = 1/2\sqrt{\Omega^2 + \delta^2}$. The time evolution operator can then be written as follows:

$$U = e^{-iH_{\text{int}}t} = \cos(|\vec{n}|t)11 - i\sin(|\vec{n}|t)\vec{\sigma}\cdot\vec{n}/|\vec{n}| \quad [9]$$

In matrix form:

$$U = \begin{pmatrix} \cos\left(\frac{\Omega_g t}{2}\right) - i\frac{\delta}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right) & \frac{\Omega}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right)e^{-i\phi} \\ -\frac{\Omega}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right)e^{i\phi} & \cos\left(\frac{\Omega_g t}{2}\right) + i\frac{\delta}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right) \end{pmatrix} \quad [10]$$

In the limit of large Rabi frequency, where $\Omega \gg \delta$, we can write the following approximate expressions for the time evolution of the computational basis states

$$\begin{aligned} U|0\rangle &= \cos\left(\frac{\Omega_g t}{2}\right)|0\rangle - \sin\left(\frac{\Omega_g t}{2}\right)e^{i\phi}|1\rangle \\ U|1\rangle &= \sin\left(\frac{\Omega_g t}{2}\right)e^{-i\phi}|0\rangle + \cos\left(\frac{\Omega_g t}{2}\right)|1\rangle \end{aligned} \quad [11]$$

The transformed states by U are eigenstates of the spin operator along the direction $\vec{n}$ which is a function of the external field parameters. Therefore, by adjusting the Rabi frequency and the field's phase, it is possible to transform the basis states into spin eigenstates along any arbitrary direction, and to perform an arbitrary rotation of our qubit, we simply have to tailor appropriately the parameters of the driving field. For example, when the Rabi frequency and the interaction time are chosen such that $\Omega t/2 = \pi/4$, and the field's phase is taken equal to π, we get

$$\begin{aligned} U|0\rangle &= \frac{1}{2}\Big(|0\rangle + |1\rangle\Big) \\ U|1\rangle &= \frac{1}{2}\Big(|0\rangle - |1\rangle\Big) \end{aligned} \quad [12]$$

This relationship defines the Hadamard transformation H introduced before. Alternatively, using quantum optics jargon, we have performed a $\pi/2$-pulse. Note that using the Bloch sphere representation, we can visualize the action of a Hadamard gate on the computational basis states as a rotations that brings them to the equatorial plane of the Bloch sphere, the transformed states by H now lying in opposite directions along the x-axis.

The next step is to show that our system also allows the implementation of conditional dynamics and, in particular, allows to build a controlled-phase gate.

Conditional Dynamics: Controlled-Phase Gate

The interaction of the atomic system with the cavity field sustained by the microwave cavity is described by the Jaynes–Cummings (JC) Hamiltonian. We can see this Hamilton operator as a generalization of the operator defined by eqn [4], where the radiation field is now quantized and the terms $e^{\pm i\omega_L t}$ are replaced by the operators a and $a^{\dagger}$ corresponding to the annihilation and creation of a cavity photon of frequency ω. The interaction with the quantized cavity field couples levels $|e\rangle$ and $|g\rangle$, while the far detuned level $|i\rangle$ will now play the role of a passive spectator. Remember that our atomic qubit is stored in levels $|i\rangle$ and $|g\rangle$. To avoid confusion with the notation, we will write the Hamilton operator in terms of the atomic states e and g. Explicitly:

$$\begin{aligned} H_{\text{JC}} &= \frac{\hbar\omega_0}{2}\Big(|e\rangle\langle e| - |g\rangle\langle g|\Big) + \hbar\omega a^{\dagger}a \\ &\quad - i\frac{\hbar\Omega}{2}\Big(a|g\rangle\langle e| - a^{\dagger}|g\rangle\langle e|\Big) \end{aligned} \quad [13]$$

where the coupling constant g has been taken to be $g = -i(\hbar\Omega/2)$, proportional to the Rabi frequency averaged over the mode structure, Ω. Using the same techniques employed in the previous section, we can derive the corresponding JC time evolution operator, which yields, for exact resonance, the following transformation rules:

$$\begin{aligned}
&|g\rangle|0\rangle_\gamma \rightarrow |g\rangle|0\rangle_\gamma \\
&|g\rangle|1\rangle_\gamma \rightarrow \cos\frac{\Omega t}{2}|g\rangle|1\rangle_\gamma - \sin\frac{\Omega t}{2}|e\rangle|0\rangle_\gamma \\
&|e\rangle|0\rangle_\gamma \rightarrow \cos\frac{\Omega t}{2}|e\rangle|0\rangle_\gamma + \sin\frac{\Omega t}{2}|g\rangle|1\rangle_\gamma
\end{aligned} \qquad [14]$$

where states $|\rangle_\gamma$ refer to the cavity field. Note that:

- The cavity field will play the role of our control qubit in the performance of a phase gate, the target qubit being the atomic system with computational states defined as before. The computational basis for the photonic qubit is provided by the Fock states $|0\rangle_\gamma$ and $|1\rangle_\gamma$.
- Strictly speaking, the time t entering in the equations above corresponds to an effective interaction time which accounts for the spatial variation of the coupling along the field mode.
- When the atom enters the cavity in state $|i\rangle$, it will acquire a global phase proportional to the ratio Ω^2/δ_i, where δ_i denotes the detuning between the atomic transition $e \leftrightarrow i$ and the mode frequency ω. In conditions of resonance, or quasi-resonance, this factor can be safely neglected, which justifies the term *passive spectator* used before.

We now have all the ingredients to understand how we can perform a controlled phase gate. Consider the situation where $\Omega t = 2\pi$, so the system experiences a full Rabi oscillation. According to the previous discussion, the basis states of two-qubit system atom-cavity field will transform as follows:

$$\begin{aligned}
&|i\rangle|0\rangle_\gamma \rightarrow |i\rangle|0\rangle_\gamma \\
&|i\rangle|1\rangle_\gamma \rightarrow |g\rangle|0\rangle_\gamma \\
&|g\rangle|0\rangle_\gamma \rightarrow |g\rangle|0\rangle_\gamma \\
&|g\rangle|1\rangle_\gamma \rightarrow -|g\rangle|1\rangle_\gamma
\end{aligned} \qquad [15]$$

That is, whenever the atomic system enters the cavity in state $|i\rangle$, it will be left unaffected, irrespective of the photon state. However, if the atom enters the cavity in state $|g\rangle$, the system will undergo a phase shift equal to π conditional to the photon number. And that is precisely the dynamics required for a controlled phase gate as defined earlier.

A detailed account of the experiments in cavity QED, showing coherent gate operation, can be found in Further Reading. These experiments prove the working principle of a universal set of quantum logic gates with neutral atoms and provide the basic elements for the development of a cavity QED quantum register in the near future.

Quantum Computing with Atoms II: Ion Traps

Charged atomic particles, ions, provided the first atomic system where conditional dynamics was demonstrated experimentally. The theoretical proposal in 1995, by Cirac and Zoller, for performing a two-qubit controlled NOT gate between trapped ions cooled to their motional ground state, was followed a few months later by a pioneering experiment at NIST, in Colorado. While the physics involved in the description of laser cooled and trapped ions is conceptually more involved than the dynamics ruled by a purely JC Hamiltonian, we can review the rudiments of the Cirac–Zoller scheme on the basis of the techniques developed in the previous sections. Nevertheless, as discussed briefly at the end of this section, we should point out that the exact procedure followed in the actual experiments departs from the analysis presented here.

Figure 6 depicts a linear ion trap where a string of ions is confined in an essentially 1-D structure. Quantum bits are stored in two long-lived hyperfine states which can be driven with laser light of controllable Rabi frequency Ω and phase ϕ. Ideally, the separation d between ions exceeds the wavelength of the external field so that ions can be addressed individually. Detection of the atomic state relies on the well mastered quantum jump technique which provides an efficiency close to 100%. From our previous discussion, it is clear that we already have all the ingredients for performing arbitrary single qubits gates, but we still need to provide a procedure to achieve two-qubit gate operations. As detailed in the next section, this will be achieved using the motional degrees of freedom of the ions as a bus for information transfer between an arbitrary pair of ions.

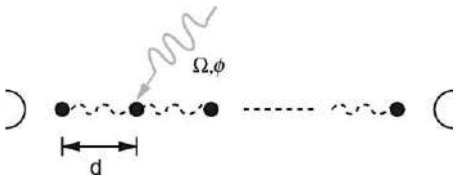

Figure 6 Schematic representation of a linear ion trap. Ions are individually addressed by external laser fields to perform single qubit operations, while two-qubit gates are mediated by the coupling to the motional degrees of freedom, depicted here as mechanical strings linking the chain of ions in the trap. The exact duration of a two-qubit gate operation and the characteristic decoherence time depend very much on the trap design, but orientative values are in the range of 50 μs and 100 − 1000 μs respectively.

The Cirac and Zoller Quantum Computer

At sufficiently low temperatures, the process of absorption and emission of photons produces a non-negligible recoil of the atomic system and the ion's motion can no longer be treated classically. In this regime, we can think of the string of ions as a system of coupled quantum harmonic oscillators whose normal mode configurations correspond to the collective motional states of the ions. The lowest energy solution corresponds to an oscillation where all ions are in phase (center of mass (CM) mode) and will be the one considered in our description. Figure 7 depicts the level scheme for ions initially cooled down to the motional ground state, illustrating both internal levels, $|e\rangle$ and $|g\rangle$, and motional (vibrational) states, denoted by $0, 1, \ldots$ Laser radiation can be tuned to be resonant with the $|g\rangle|0\rangle_{cm} \leftrightarrow |e\rangle|0\rangle_{cm}$ transition (carrier transition), but may also be tuned to drive transitions with different vibrational states (sideband transition), as represented by the external arrows in the figure. The realization of single qubit gates simply involves tuning the laser towards the carrier transition. The interaction is ruled by essentially the same Hamiltonian discussed above. Therefore, choosing appropriately the laser parameters, it is possible to induce an arbitrary rotation of the ion pseudo-spin.

To prove that the system allows the implementation of a universal set of gates, we will now show that it is possible to operate a controlled phase gate between two arbitrary ions. For that we need to couple internal and motional degrees of freedom which will be done by means of tuning the laser light to a sideband transition. The corresponding Hamiltonian will not be derived here in detail but it can be understood as describing a Jaynes–Cummings form of dynamics where the quantized electromagnetic field within a cavity is replaced by the quantum mechanical oscillator describing the vibrational states of the ion. Explicitly:

$$H = \frac{\eta\Omega}{2}\left(c|e\rangle\langle g| + c^{\dagger}|g\rangle\langle e|\right) \qquad [16]$$

where c and $c^{\dagger}$ denote the annihilation and creation operators of vibrational quanta and η is the so-called Lamb–Dicke parameter, a measure of the recoil energy. The gate operation involves the following sequence:

- Mapping the internal state of the control ion into the collective motional state of the chain,
- performing an operation on the target ion conditional to the excitation number of the vibrational state and
- reverse the initial operation restoring the vibrational state to its ground state.

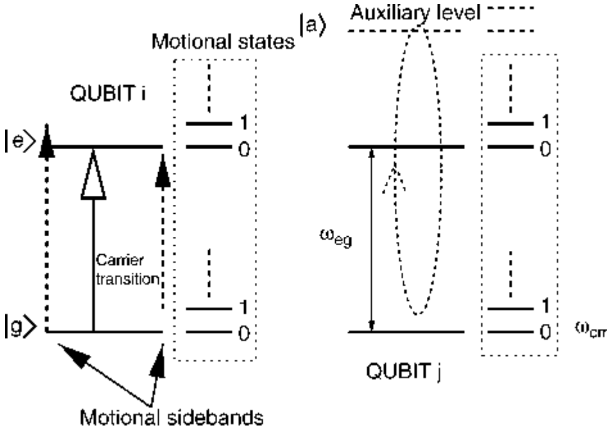

Figure 7 Relevant internal and motional atomic levels for the realization of a controlled NOT gate between two arbitrary ions within a chain. See the text for details.

Imagine ion i, which will play the role of the control, and j, the target, initially prepared in arbitrary superpositions of the corresponding internal states, with the whole system cooled down to their motional ground state $|0\rangle_{cm}$. The joint state of the two particles can be written as a product state of the form:

$$\left(\alpha_i|g\rangle_i + \beta_i|e\rangle_i\right) \otimes |0\rangle_{cm} \otimes \left(\alpha_j|g\rangle_j + \beta_j|e\rangle_j\right) \qquad [17]$$

When ion i is irradiated by a laser tuned to the first motional side band $|g\rangle|1\rangle_{cm} \leftrightarrow |e\rangle|0\rangle_{cm}$, and the timing is such that the system undergoes a π-Rabi rotation, the internal state of the ion i is mapped onto the vibrational state (Note the transformation rules given in eqn [14]. As a result, the joint state of the system is still factorizable as

$$\left(|g\rangle_i \otimes \alpha_i|0\rangle_{cm} + \beta_i|1\rangle_{cm}\right) \otimes \left(\alpha_j|g\rangle_j + \beta_j|e\rangle_j\right) \qquad [18]$$

The second step of the protocol consists of applying an operation on ion j conditional on the vibrational state of the joint system. For that we apply a laser pulse resonant with the transition $|e\rangle_j|1\rangle_{cm} \leftrightarrow |a\rangle_j|0\rangle_{cm}$ with an auxiliary level $|a\rangle_j$ so that ion j undergoes a full Rabi oscillation. This results in acquiring a phase shift equal to π only if the system is state $|e\rangle_j|1\rangle_{cm}$ (see eqn [15]). This step entangles the internal and vibrational degrees of freedom of ion j, leading to a joint state of the form:

$$|g\rangle_i \otimes \Big(\alpha_i|0\rangle_{cm}(\alpha_j|g\rangle_j + \beta_j|e\rangle_j) + \beta_i|1\rangle_{cm}(\alpha_j|g\rangle_j - \beta_j|e\rangle_j)\Big) \qquad [19]$$

Applying a final laser pulse on ion i identical to the initial one, the state of the vibrational mode is transferred back to ion i and the atomic systems are left in the, generally, entangled state

$$\alpha_i\alpha_j|g\rangle_i|g\rangle_j + \alpha_i\beta_j|g\rangle_i|e\rangle_i + \beta_i\alpha_j|e\rangle_i|g\rangle_j - \beta_i\beta_j|e\rangle_i|e\rangle_j, \qquad [20]$$

which completes the controlled-phase gate operation.

A simplified version of this scheme shortly followed the theoretical proposal by Cirac and Zoller, proving the working principle of the ions trap quantum processor. Since then, both theoretical and experimental progress have allowed the development of schemes for implementing faster gates as well as protocols which do not require cooling to the motional ground state. We will briefly revise the current state of the art and future prospects of this technology in the last section of this article.

Frequency Estimations as a Quantum Computation: Building Frequency Standards with Entangled States

Recently, it has been proposed to make use of entangled states of trapped ions for performing high-resolution spectroscopy beyond the shot noise limit. This boundary is the ultimate limit achievable when performing a frequency measurement using uncorrelated particles. Consider n atomic two-level systems with frequency separation ω_0 initially prepared in their ground state $|0\rangle$. The aim is to lock a certain oscillator of frequency ω_L to the atomic frequency. To do that, the currently employed protocols use the so-called Ramsey method of separate oscillating fields, where each atomic system undergoes two $\pi/2$-pulses (or Hadamard transformations) separated by a free time evolution of duration t, much larger than the duration of the pulse. Looking at the Hamiltonian describing the interaction between the ion and the field, eqn [1], one can easily see that the free time evolution can be understood as a controlled phase gate in the following sense. If we set $g = 0$ in eqn [1] and for simplicity re-scale the ion internal energies so that $H = \hbar\delta|1\rangle\langle 1|$, the free evolution operator can be written as $U = -\hbar\delta|1\rangle\langle 1|$, so that the atomic basis states evolve as:

$$U|0\rangle = |0\rangle$$
$$U|1\rangle = e^{-i\delta t}|1\rangle \qquad [21]$$

In other words, the ground state remains invariant, whereas the upper levels pick up a phase which is proportional to the detuning from the atomic transition. The state of each particle after the time evolution can then be written as

$$\frac{1}{\sqrt{2}}\Big(|0\rangle + e^{-i\delta t}|1\rangle\Big) \qquad [22]$$

After the second pulse, which again mixes populations of levels $|0\rangle$ and $|1\rangle$, one measures the population of, say, level $|1\rangle$. The signal is therefore a sinusoidal function of the form

$$P_1 = \frac{1}{2}\Big(1 + \cos(\delta t)\Big) \qquad [23]$$

which is symmetrical with respect to $\delta = 0$. This procedure is repeated a number of times to accumulate a sufficient data. Let us call T the total duration of the procedure, so that we accumulate $N = nT/t$ experimental data. Finally, whenever statistics in the populations are different from 50:50 at half maximum are detected, the oscillator frequency is steered to achieve exact resonance. The ultimate resolution $\Delta\omega_0$ achievable with this technique is limited simply by projection noise and can be written as

$$\Delta\omega_0 = \frac{1}{\sqrt{nTt}} \qquad [24]$$

After time T, the oscillator's frequency is steered towards the resonance and new sequence starts. We will now illustrate how the use of entanglement can yield an uncertainty in the measurement beyond the shot noise limit eqn [24]. Consider the quantum network depicted in Figure 8. The n particles are now initially prepared in a maximally entangled state. This is achieved by means of subjecting the first ion to

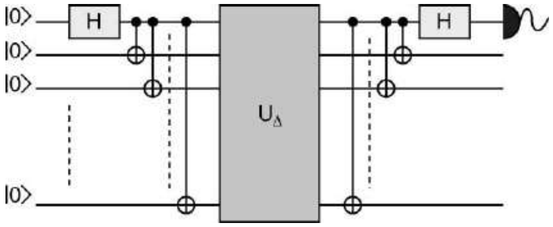

Figure 8 Quantum circuit representation for an experimental setup for a frequency estimation using maximally entangled states.

a Hadamard transformation followed by a cascade of CNOTs between the first ion, which acts as a control, and each of the remaining $(n-1)$ particles.

The resulting state is of the form

$$\frac{1}{\sqrt{2}}\Big(|0\ldots0\rangle + |1\ldots1\rangle\Big) \qquad [25]$$

The systems now evolve freely for a time t and, in light of the discussion above, we can represent this evolution in terms of a quantum gate which will leave state $|0\rangle$ invariant while introducing a phase factor proportional to the detuning if the system is in state $|1\rangle$. As a result, the joint state of the n particle after the free time evolution is given by

$$\frac{1}{\sqrt{2}}\Big(|0\ldots0\rangle + e^{-in\delta t}|1\ldots1\rangle\Big) \qquad [26]$$

with a phase factor amplified by a factor n as compared with the single particle state given by eqn [22]. The second part of the network disentangles the n particles, leaving the first ion in the state $(1/\sqrt{2})(|0\rangle + e^{-in\delta t}|1\rangle)$. If we now subject it to an additional Hadamard gate, the probability for the ion to be excited is

$$P_{1,ent} = \frac{1}{2}\Big(1 + \cos(n\delta t)\Big) \qquad [27]$$

which oscillates n times faster than the signal eqn [23]. The corresponding uncertainty for the estimation of the atomic frequency is now

$$\Delta\omega_0, ent = \frac{1}{n\sqrt{Tt}} \qquad [28]$$

This corresponds to an increase in the resolution of $\sqrt{n}$ beyond the shot noise bound. In practice, the achievable resolution using entangles states is limited by the decoherence time of the system t_{dec}, which sets the condition $T < t_{\text{dec}}/2$ for the procedure to be valid. Remarkably, entanglement is still advantageous in a situation where T is chosen such that it exceeds the decoherence time. However, maximally entangled states should now be replaced by certain partially entangled ones to overcome the precision achievable with uncorrelated particles.

The application of entangled states in precision spectroscopy is a nice illustration of the potential of the ideas developed in quantum information beyond the ultimate task of building a large-scale quantum computer. The first experimental results showing the enhanced performance of entangled states have been reported by the NIST group in Boulder.

Conclusion and Perspectives

We have revised the fundamental concepts for the practical realization of quantum computation using atomic systems. We have showed how single qubit and two-qubit controlled-phase gates can be implemented using atoms interacting with the cavity field sustained by a microwave resonator and ion traps. We have illustrated how a very simple quantum computation can be used for the development of more precise frequency standards, emphasizing the potential of the ideas developed in quantum information beyond the implementation of large scale quantum computation.

Cavity QED technology has so far allowed the generation of entangled states of two and three particles and has demonstrated coherent gate operation. Ion trap technology has pushed controllable multi-particle entanglement to four particles and has showed the possibility of implementing forms of error correction. The combination of these two technologies offers interesting prospects for the near future. Surrounding traps by optical cavities and networking these devices via optical fibers is expected to allow the construction of the first prototypes for a quantum communication network.

See also

Quantum Optics: Entanglement and Quantum Information.

Further Reading

Cirac JI and Zoller P (1995) Quantum computations with cold trapped ions. *Physical Review Letters* 74: 4091.

Deutsch D, Barenco A and Ekert AK (1995) Universality in quantum computation. *Proc. Roy. Soc. London A MAT* 449: 669–677.

Grover L (1999) Quantum mechanics helps in searching for a needle in a haystack. *Physical Review Letters* 79: 325–328.

Huelga SF, *et al.* (1997) Improvement of frequency standards with quantum entanglement. *Physical Review Letters* 79: 3865.

Jaynes ET and Cummings FW (1963) Comparison of quantum and semiclassical radiation theories with application to the beam maser. *Proc. IEEE* 51: 89.

Kielpinski D, Meyer V, Rowe MA, *et al.* (2001) A decoherence-free quantum memory using trapped ions. *Science* 291: 1013–1015.

Monroe C, Meekhof DM, King BE, *et al.* (1995) Demonstration of a fundamental quantum logic gate. *Physical Review Letters* 75: 4714–4717.

Raimond JM, Brune M and Haroche S (2001) *Colloquium*: manipulating quantum entanglement with atoms and photons in a cavity. *Review Modern Physics* 73: 565 and references therein.

Sackett CA, Kielpinski D, King BE, *et al.* (2000) Experimental entanglement of four particles. *Nature* 404: 256–259.

Shor P (1994) *Proc. of the 35th Annual Symposium on the Foundations of Computer Science, IEEE Computer Society, Los Alamitos*. Santa Fe, NM: IEEE Computer Society Press.

Sorensen A and Molmer K (1999) Quantum computation with ions in thermal motion. *Physical Review Letters* 82: 1971.

Wineland DJ, Monroe C, Itano WM, Leibfried D, King BE and Meekhof DM (1998) Experimental issues in coherent quantum-state manipulation of trapped atomic ions. *J. Res. Natl. Inst. Stand. Tech.* 103: 259 and references therein.

Squeezed Phonons in Solids

A V Bragas, Universidad de Buenos Aires, Buenos Aires, Argentina

R Merlin, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The reduction of noise is crucial for measurements requiring high accuracy and, often, for the observation of new phenomena. Among noise sources, noise inherently related to the quantum nature of an object persists even after all the classical sources of noise are removed. Quantum noise is a fundamental limitation, loosely related to the Heisenberg uncertainty principle, which imposes a lower limit to the fluctuations of any pair of conjugated variables, like position and momentum. For conjugated variables A_1 and A_2 we have

$$\Delta A_1 \Delta A_2 \geq \frac{\hbar}{2} \qquad [1]$$

where

$$\Delta A_i = \sqrt{\langle A_i^2 \rangle - \langle A_i \rangle^2} \qquad i = 1, 2 \qquad [2]$$

While quantum mechanics does not forbid reducing the noise in one of the conjugated observables down to zero, the noise for the other variable will increase according to eqn [1]. Squeezing refers to the reduction of noise of a quantum variable at the expenses of its conjugated variable. The term originates from quantum optics. Nonlinear interactions of the electromagnetic field with passive and active atomic media have been successfully used to generate squeezed photons, opening new possibilities for essentially noiseless optical communications and precision measurements. The electromagnetic field is not the only quantum object that can be squeezed. By performing appropriate manipulations of the internal quantum state of a system, it is, in principle, possible to squeeze an arbitrary variable. In the past few years, vibrational states have also been experimentally squeezed in molecules and solids, by exciting the system with ultrafast optical pulses. The concept of squeezing can also be extended to an ensemble of classical objects for which the fluctuations are of thermal origin.

Squeezing applies in particular to situations where the relevant variables are the position Q and momentum P of a harmonic oscillator. Consider the Hamiltonian of a particle of mass m in a harmonic potential of frequency Ω

$$H_0 = \frac{P^2}{2m} + \frac{m}{2}\Omega^2 Q^2 \qquad [3]$$

The ground state wavefunction is

$$\varphi_0 = \frac{1}{\left[2\langle Q_0^2 \rangle \pi\right]^{1/4}} \exp\left(-\frac{Q^2}{4\langle Q_0^2 \rangle}\right) \qquad [4]$$

where $\langle Q_0^2 \rangle = \hbar/2m\Omega$. The stationary probability, $|\varphi_0|^2$, is represented in Figure 1a. This state is also a minimal uncertainty state in that $\Delta Q \Delta P \equiv \hbar/2$. $\langle Q_0^2 \rangle$ gives a measure of the quantum noise and it is known as the vacuum noise level.

Let us now discuss the behavior of a harmonic oscillator that interacts with an external source. There are two types of quantum states for which the Gaussian shape of the wavefunction is preserved after interacting with an external driving force. These are the coherent state and the squeezed state. If the interaction places the system in a state of the same form as eqn [4] but displaced with respect to the center, the wavepacket will oscillate periodically, with period $\tau = 2\pi/\Omega$, without changing its shape as shown in Figure 1b. This quantum state is known as the coherent state. It has the same uncertainty as the ground state but the expectation value $\langle Q(t) \rangle$ depends on the time following the classical trajectory in the harmonic potential. If, instead, we introduce a sudden change in the spring constant at $t = 0$, the wavepacket will begin to 'breathe' for $t > 0$, so that its width will vary periodically with a period given by $\pi/2$. This is the squeezed state. Here the noise of one of the variables is periodically reduced below the level of the vacuum noise (Figure 1c). A state that is a combination of a coherent and a squeezed state is shown in Figure 1d. The wavepacket follows the classical trajectory and breathes at the same time.

The results in Figure 1 apply only to harmonic oscillators. While the concept of squeezing can be extended to anharmonic potentials, there are additional features to consider such as, for example, wavepacket revivals.

Dynamics of a Parametrically Excited Crystal

Harmonic oscillators describe a wide range of physical systems. In particular, the dynamics of a crystal lattice can be described in terms of a set of independent harmonic oscillators, the quantum nature of which is described in terms of phonons. In this section we discuss interactions of a harmonic oscillator with external sources that produce squeezing.

The normal modes in a crystal obey equations of the form:

$$\ddot{Q}_q + \Omega_q^2 Q_q = 0 \qquad [5]$$

where Q_q and Ω_q are the phonon coordinate and frequency, and q is the wavevector.

As briefly mentioned in the previous section, a sudden change in the spring constant produces a squeeze state. To do this, we need what is known as time-dependent parametric coupling. As in the experiments on phonon squeezing, using ultrashort optical pulses, we consider an impulsive driving force proportional to Q. The relevant equation is:

$$\ddot{Q}_q + \Omega_q^2 Q_q = \lambda_q Q_q \delta(t) \qquad [6]$$

where $\delta(t)$ is the Dirac delta function, and λ accounts for the strength of the external impulsive force. This impulsive excitation induces a sudden change in momentum, proportional to the normal coordinate Q at $t = 0$, as can be shown by integrating eqn [6]. Explicitly, $\Delta\dot{Q}_q = \Delta P_q = \lambda_q Q_q(0)$. $\Delta P_q = P_q(0^+) - P_q(0^-)$, where $0^-(0^+)$ is the time immediately before (after) the excitation. This result is illustrated in Figure 2. The main effect of the interaction is to change the width of the Gaussian packet or variance of the distribution. This leads to a time-dependence of the noise defined in eqn [2], $\Delta Q^2(t) = \langle Q^2(t)\rangle$ reflecting the 'breathing' of the wavepacket. Since P_q is the conjugated variable of Q_q, $\langle P_q^2(t)\rangle$ shows the same time-dependent behavior as $\langle Q_q^2(t)\rangle$ but in quadrature with it, so as to fulfill eqn [2] at all time. Assuming an oscillator initially in its ground state, the solution of eqn [6] is

$$\langle Q_q^2(t)\rangle = \frac{\hbar}{2m\Omega_q}\left[1 + \left(\frac{\lambda_q}{\Omega_q}\right)\sin(2\Omega_q t) + \left(\frac{\lambda_q}{\Omega_q}\right)^2 \sin^2(\Omega_q t)\right] \qquad [7]$$

It is then clear that the fluctuations of Q_q periodically fall below the vacuum quantum noise level, determined by $\langle Q_0^2\rangle$, and that the oscillations have a frequency that is twice the frequency of the oscillator.

The considerations above apply to a quantum harmonic oscillator at temperature $T = 0$. In the classical case, we need to consider an ensemble of oscillators in equilibrium with a thermal bath at temperature T. It can be shown that eqn [7] also applies to the classical case, provided $\hbar/2m\Omega$ is replaced by $k_B T/\Omega^2 m$, where k_B is the Boltzman constant. In classical squeezing, the external perturbation is applied simultaneously to the ensemble, leading to a sudden change of the momentum of all the oscillators. The jump in momentum is proportional to the position of the particle at $t = 0$. After the interaction, the noise of the ensemble will fall periodically below the thermal noise at temperature T given by $k_B T/\Omega^2 m$. A pictorial representation of classical squeezing is shown in Figure 3.

To describe a more realistic situation, it is necessary to take into account effects due to dissipation. Theoretically, this is usually accomplished by considering the coupling with a continuum of external modes simulating a thermal bath. The solution to this type of problem is nontrivial even for the simplest case of linear coupling to the bath.

The impulsive excitation of eqn [6] is a parametric excitation containing all the frequencies. A more common way to parametrically excite an oscillator is to use a monochromatic source, as described by the equation

$$\ddot{Q} + [\Omega^2 + \varepsilon\cos(\Omega_0 t)]Q = 0 \qquad [8]$$

where the parametric excitation is characterized by the amplitude ε and the modulation frequency Ω_0. Notice that the excitation induces a time-dependent frequency $\omega^2(t) = \Omega^2 + \varepsilon\cos(\Omega_0 t)$, and, hence, it

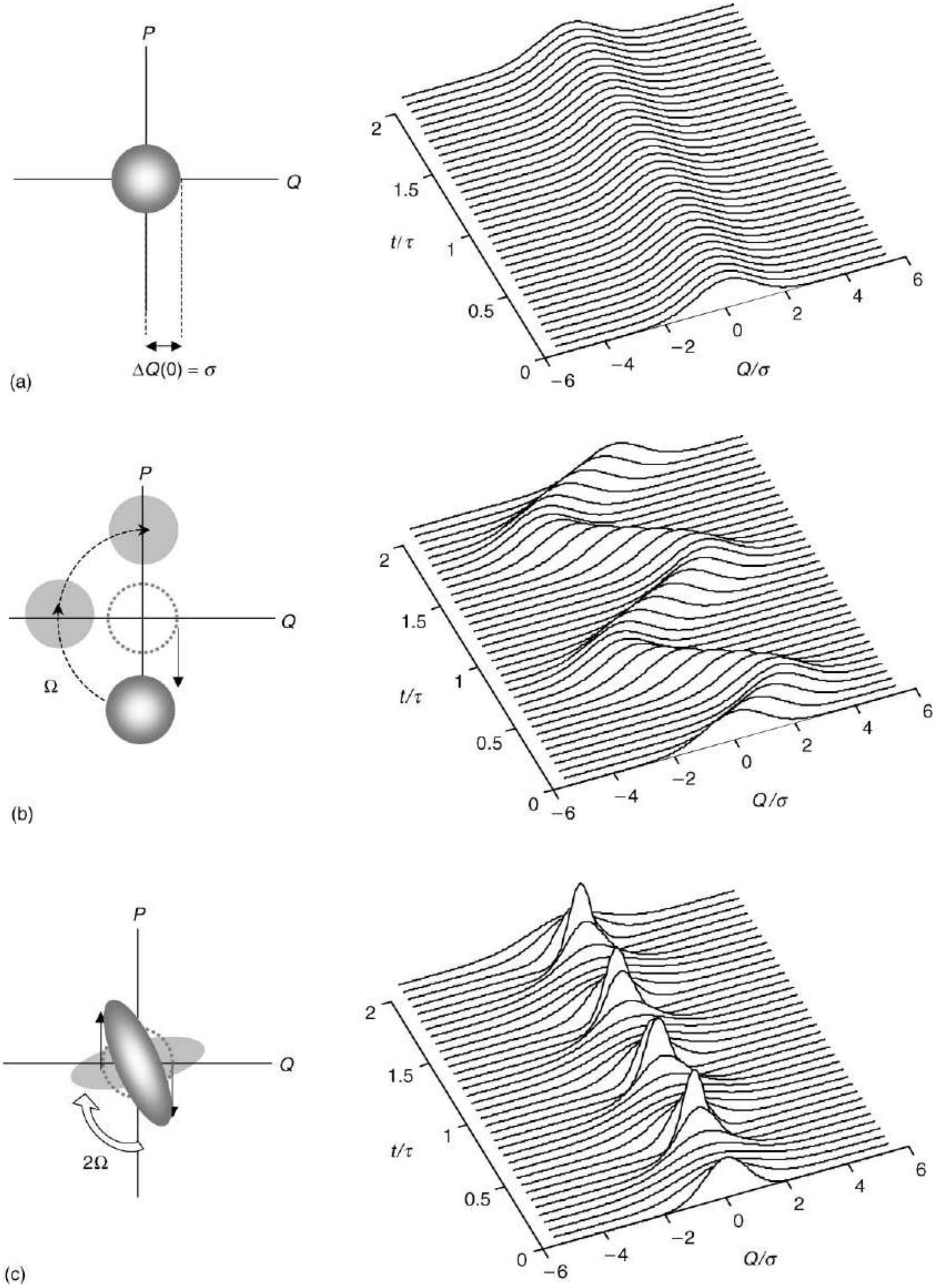

Figure 1 (a) Ground state probability density $|\varphi_0|^2$ for a quantum harmonic oscillator of frequency Ω. Q is the position, P is the momentum and σ is the standard deviation of the Gaussian wavepacket, or the square root of $\langle Q_0^2 \rangle$. The phase-space diagram on the left shows the uncertainty region for the ground state. (b) Coherent state. The uncertainty $\langle Q^2 \rangle$ does not change in time and it is always equal to $\langle Q_0^2 \rangle$, but the wavepacket oscillates with the period τ of the harmonic oscillator. The phase diagram shows the evolution of the uncertainty region, starting from the ground state (dashed line). A sudden constant change in momentum at time $t = 0$ produces a change in the equilibrium position and the wavepacket begins to oscillate. (c) Squeezed state. The wavepacket evolves by changing its width periodically with period $\tau/2$. The uncertainty, $\langle Q^2(t) \rangle$, takes values smaller than $\langle Q_0^2 \rangle$ for a fraction of the cycle. Due to the sudden change on momentum, proportional to Q, at $t = 0$, the distribution becomes ellipsoidal and rotates with frequency 2Ω. (d) Coherent-squeezed state. It is a combination of both (b) and (c). The wavepacket oscillates and breathes at the same time.

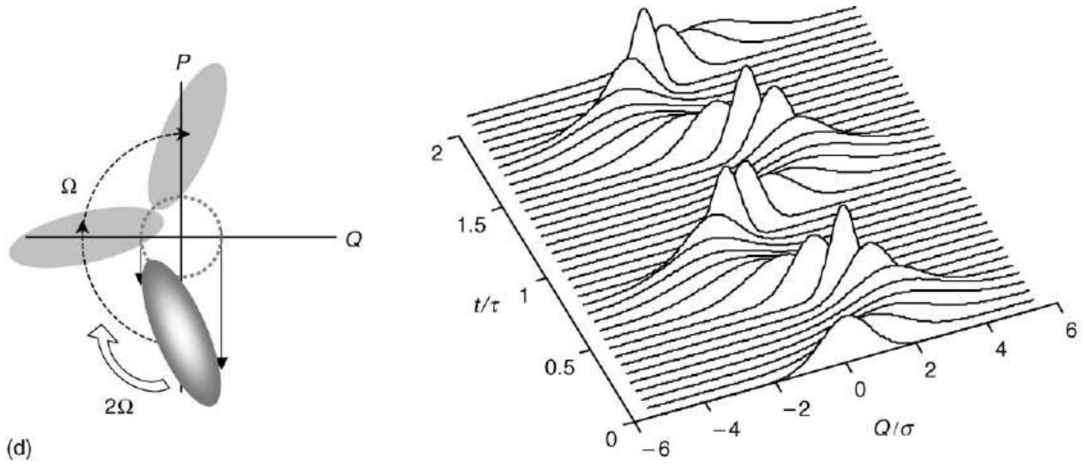

Figure 1 *Continued.*

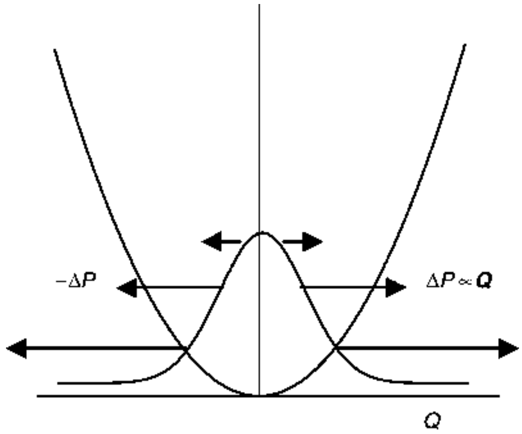

Figure 2 Schematic representation of the effect of an impulsive excitation on the ground state of the harmonic oscillator. At $t = 0$, the pulse induces a change in momentum P, proportional to the position Q. As a result, the wavepacket width in momentum space changes suddenly and starts to oscillate.

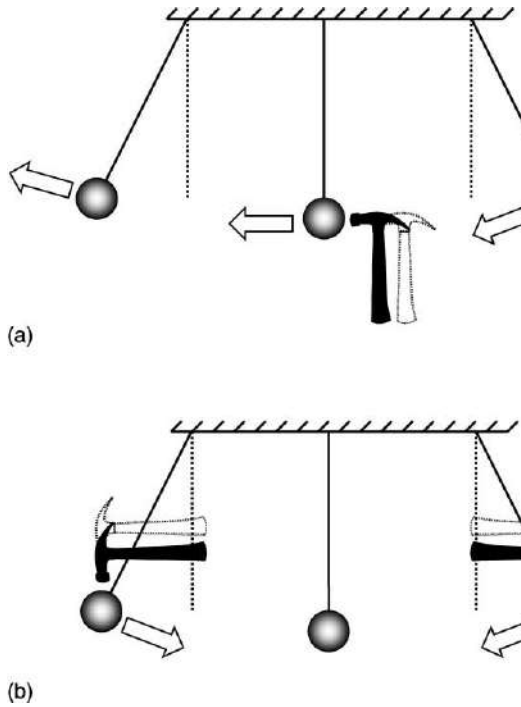

Figure 3 (a) Classical representation of a coherent phonon. An impulsive horizontal force acts on all members of a thermal distribution of pendula, displacing the group in phase space along the momentum axis. The ensemble precesses with an average position $\langle x \rangle \sim \sin(\Omega t)$. (b) Classical representation of a squeezed phonon. An impulsive vertical force produces an impulse proportional to the pendulum's displacement from equilibrium. Those pendula farthest from equilibrium receive the largest kick. The phase space distribution becomes elliptical and precesses with an ensemble averaged variance $\langle x^2 \rangle \sim \sin(2\Omega t)$.

shows squeezing effects. The complete set of solutions with and without dissipation has been obtained by Zerbe and Hanggi for the quantum and classical cases.

Physical Interactions Giving Phonon Squeezing

The considerations above are rather general and specific to any physical system that behaves like a set of harmonic oscillators. Here, we consider proposals for phonon squeezing.

Equations [6] and [8] show that we need a driving force proportional to the phonon coordinate Q to achieve squeezing. One way to do this is for the external interaction to produce time-dependent correlations between two modes. Prior to the experimental achievement of phonon squeezing in 1997, using stimulated second-order Raman scattering, models based on (a) three-phonon parametric amplification and (b) polaritons had been proposed.

These proposals have not yet been implemented experimentally.

The scheme proposed in (a) is based on three-phonon parametric amplification, similar to photon squeezed states that are generated by optical parametric amplification. Specifically, the proposal relies on a two-step process. First the generation by external excitation with two lasers of a 'pump' longitudinal optical phonon (LO) and, second, the decay of the optical phonon into two longitudinal acoustic phonons (LA), which will be squeezed. This is the kind of parametric oscillator described by eqn [8], for which the parametric control is driven by the coherent optical phonon field.

In case (b) it has been shown that mixed polarization modes known as phonon polaritons are intrinsically squeezed. These hybrid objects, sometimes referred to as dressed photons, are a mixture between phonons and photons that cannot be separated once they are formed.

Squeezing of vibrational states in molecules was confirmed experimentally earlier than in solids. Here, a molecule initially in the ground electronic and vibrational state is excited by an ultrafast optical pulse tuned with an electronic resonance. After the pulse arrives, the molecule is placed in a excited electronic state and the nuclei in a new (excited) potential with a new set of discrete vibrational modes. In Figure 4 we show, as an example of a diatomic molecule, the potential of a sodium dimer. For the sake of simplicity, we consider a single vibrational mode in the excited state and we disregard effects due to anharmonicity. If the optical pulsewidth is much shorter than any vibrational frequency, and the frequencies are different in the ground and excited state, the sudden change in frequency produces a vibrational squeezed state. On the other hand, even if the frequencies in the ground and excited state are equal, but the laser pulse is finite or chirped (frequency sweep on time), the final vibrational state may be squeezed. The formal demonstration of this is not trivial. Analytic solutions for the vibrational wavefunction for an arbitrary external field excitation are known for a very limited number of cases.

Squeezing of a wavepacket in an excited sodium diatomic molecule was achieved by Dunn and coworkers, measuring the time and frequency-resolved spectrum of the spontaneous emission. In this experiment, the molecule can return at any time to its electronic ground state by emitting light, the wavelength of which depends on the internuclear separation at that time. As shown in Figure 4, this is because the equilibrium position for the potential surfaces in the ground and excited states are different.

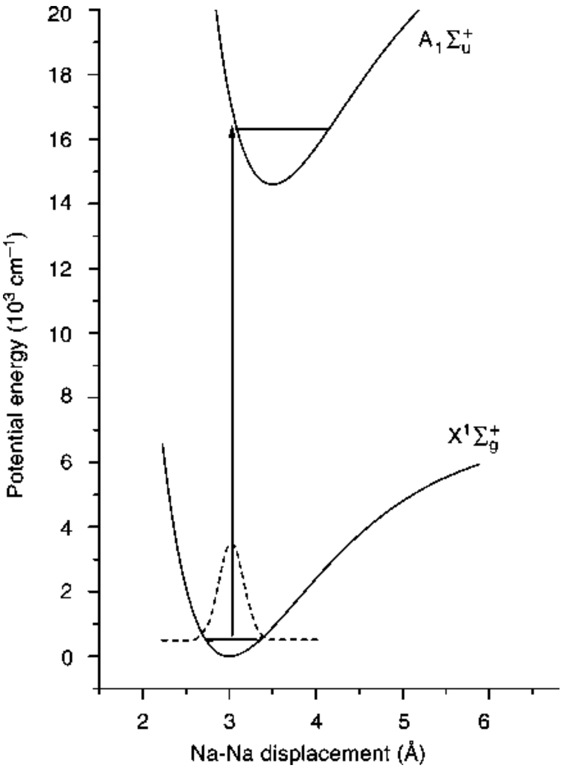

Figure 4 Potential curves for a Na_2 molecule, showing the ground and first excited states. The dashed line represents the ground state wave function. The equilibrium position of both potentials are shifted and the vibrational frequencies are different.

Thus, the time-dependent average wavelength of the emitted light oscillates as the atom vibrates, giving a signature of the dynamics of the wavepacket. Dunn observed oscillations in the signal strength that are consistent with a wavepacket expanding and contracting.

Experimental Generation and Detection of Squeezed Phonons by Second Order Raman Scattering

The first experimental demonstration of phonon squeezing was performed in 1997 by exciting a $KTaO_3$ (potassium tantalate) crystal with an ultrafast pulsed laser and coupling the light to the phonons through second order Raman Scattering. The squeezed phonon field produced by an intense optical pulse, referred to as the 'pump', is detected by a second low intensity pulse, known as the 'probe'. Phonon squeezing was also reported for $SrTiO_3$ (strontium titanate). In addition, correlations between pairs of phonons of different frequencies was reported in a closely related experiment on $KTaO_3$.

The simplest Hamiltonian describing this experiment contains the harmonic contribution from the lattice and a Raman-type interaction term. Raman scattering (RS) refers to light-crystal interaction in which the initial quantum of radiation energy, $\hbar\omega_i$, is divided into a vibrational excitation of the crystal with energy $\hbar\Omega$ and a scattered photon with energy $\hbar\omega_f$. Here, the total energy and momentum are conserved. The interaction energy is proportional to the second-order electronic susceptibility which can be expanded in terms of the atomic displacement Q. The nth term of the Taylor expansion is associated with RS of order n. Two phonons participate in the second order RS process. Although the absence of first-order RS is not required for achieving squeezing, $KTaO_3$ and SrTiO3 (above ~100 K) show no first-order Raman active modes. As a result, the leading contribution is

$$H_{\text{int}} = -\frac{1}{4}|E(t)|^2 \sum_q \chi_2(q) \qquad [9]$$

with $\chi_2(q) = (\partial^2\chi/\partial Q_q^2)Q_q^2$, where $\partial^2\chi/\partial Q_q^2$ is the second order Raman tensor, $E(t)$ is the magnitude of the light electric field, and χ is the linear electronic susceptibility. The Hamiltonian eqn [9], reflecting contributions from pairs of modes at $\pm q$, represents a change in the phonon frequency $\Delta\Omega_q(t) \approx -(1/4\Omega_q)(\partial^2\chi/\partial Q_0^2)|E(t)|^2$ that results in a time-dependent variance $\langle Q_q^2(t)\rangle$ and thus, squeezing. We note that, while individual scattering events produce a two-phonon state, the coherent ensemble of these correlated two-phonon states modulates the variance of Q. Notice that second-order RS excites phonons throughout the whole Brillouin zone instead of just the zone center at $q \approx 0$, as in first-order RS. This means that the light couples to a continuum of modes. The strength of the scattering at any particular frequency is proportional to the number of modes, i.e., the density of states, which is inversely proportional to the gradient $\nabla_q\Omega_q$ and thus, exhibits singularities at points where the gradient vanishes. These so-called critical points lead to associated singularities in the density of states referred to as van Hove singularities. $KTaO_3$ and $SrTiO_3$ show a sharp peak in the density of states at frequencies associated with certain van Hove singularities and, therefore, the continuum of modes becomes quasimonochromatic.

In the experiments, the period of the relevant phonons, $2\pi/\Omega_q$, is large compared with both the pulsewidth and the time the pulse takes for traversing the sample. Then, the time-dependent electric field can be approximated by the delta function $|E(t)|^2 \cong (4\pi I_0/nc)\delta(t)$, where I_0 is the pulse energy density, n the refractive index and c the speed of the light. Within this approximation, the dynamics of the system at $T = 0$ is described by eqn [6]. To lowest order in I_0, we have

$$\begin{aligned}\Delta Q_q &= \langle Q_q^2(t)\rangle^{\frac{1}{2}} \\ &\approx \langle Q_q^2(0)\rangle^{\frac{1}{2}}\left[1 + \left(\frac{\lambda}{\Omega_q}\right)\sin(2\Omega_q t)\right]\end{aligned} \qquad [10]$$

with $\lambda = 2\pi(I_0/nc)(\partial^2\chi/\partial Q_q^2)$.

As before, the variance drops below the quantum limit $\langle Q_q^2(0)\rangle$ for some fraction of the cycle. This zero-temperature description is only accurate for temperatures such that $k_B T$ is very small compared with the phonon energies. For $T \neq 0$, the conditions overcoming the thermal contribution and produce quantum squeezing at low intensities are

$$n_q \ll 1 \quad \text{and} \quad \left(\frac{\lambda}{\Omega_q}\right) > 2n_q \qquad [11]$$

where $n_q = 1/[\exp(\hbar\Omega_q/k_B T) - 1]$ is the Bose factor.

Next we discuss the detection of the squeezed field. As mentioned earlier, the pump pulse creates the squeezed phonon state and the probe is affected by the changes in the material produced by the pump. The scattering of the probe by the squeezed phonon is also described by the nonlinear polarization of the material $P^{NL} = (2V)^{-1}E'(t)\sum_q\chi_2(q)$, where $E'(t)$ is the probe field, and V is the scattering volume. If the probe pulse is Gaussian-shaped of width τ_0 and centered at $\omega = \omega_0$, the total normalized change in the probe transmission ΔT is

$$\begin{aligned}\frac{\Delta T}{T}(\omega) \approx &-\frac{l\pi\tau_0^2\omega(\omega-\omega_0)}{ncV}\sum_q \frac{\partial^2\chi}{\partial Q_q^2}|E'(t)|^2 \\ &\times \lambda\langle Q_q^2(0)\rangle\cos(2\Omega_q\Delta t)\exp(-2\Omega_q^2\tau_0^2)\end{aligned} \qquad [12]$$

where l is the length of the sample. Equation [12] can be obtained from well-known results in coherent phonon detection. Comparing eqn [12] with eqn [10] it follows that the change in probe transmission is proportional to the derivative of the total time-dependent variance $\langle Q^2(t)\rangle = \sum_q\langle Q_q^2(t)\rangle$. Accordingly, the integral of $\Delta T/T$ probes $\langle Q^2(t)\rangle$.

$KTaO_3$

The perovskite $KTaO_3$ was the first material for which phonon squeezing was achieved. This compound has no first-order Raman active modes. The time-domain data in Figure 5a shows behavior

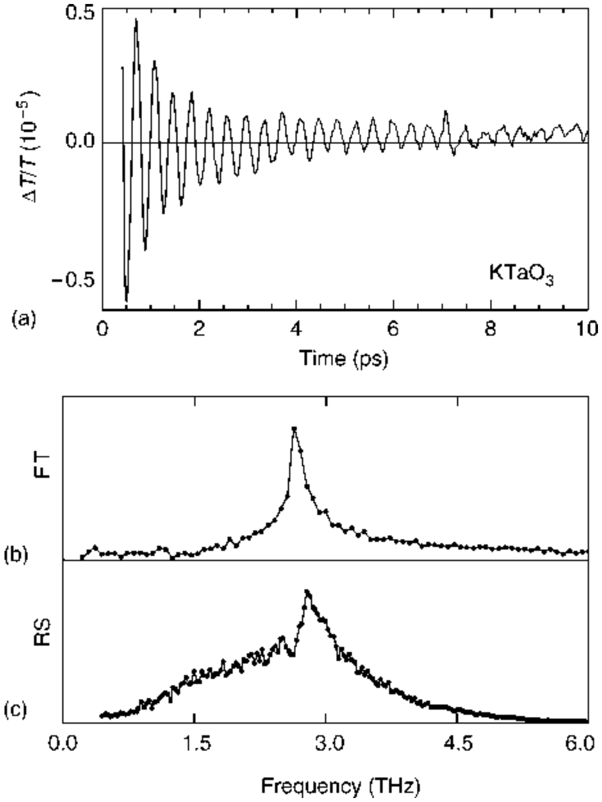

Figure 5 (a) The normalized transmitted probe intensity as a function of the probe delay for KTaO$_3$ at $T = 10$ K. Reprinted with permission from Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science*, 275(5306): 1638–1640. Copyright 1997 American Association for Advancement of Science. (b) Fourier transform of the time domain data in (a). The narrow peak appears very close to twice the frequency of the TA (transverse acoustic) mode at the X point of the Brillouin zone. (c) Second order Raman cross section.

consistent with eqn [12]. The Fourier transform $\mathcal{F}_{\mathrm{FT}}$ in Figure 5b, is dominated by a narrow peak, strongly dependent on temperature, that appears very close to twice the frequency of the TA (transverse acoustic) mode at the X point of the Brillouin zone. The comparison with the spontaneous Raman scattering data, Figure 5c, indicates that the $\mathcal{F}_{\mathrm{FT}}$ peak is due to the 2TA overtone. The squeezing factor defined as $S = 1 - [\langle Q^2(t)\rangle/\langle Q^2(0)\rangle]^{1/2}$, is depicted in Figure 6. Notice that, whereas the squeezing factor in a quantum optics experiment can reach values as large as 70%, phonons in a solid can be squeezed by no more than, say, 0.01%. This is because larger values may lead to collapse of the solid since, according to the Lindemann criterion, solids melt when the root-mean-square displacement of an atom is a fraction of 0.2–0.25 of the lattice parameter.

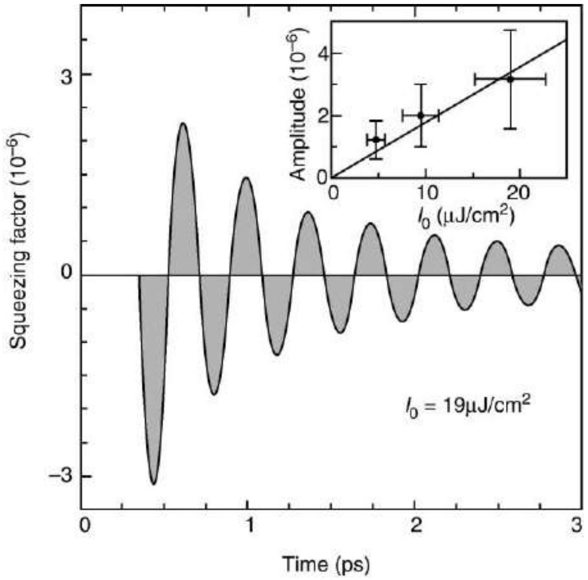

Figure 6 Experimental time dependence of the squeezing factor $S = 1 - [\langle Q^2(t)\rangle/\langle Q^2(0)\rangle]^{1/2}$. I_0 is the pulse intensity. Inset: the amplitude of S as a function of the integrated intensity of the pulse, showing the linear behavior predicted by eqn [12]. Reprinted with permission from Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science* 275(5306): 1638–1640. Copyright 1997 American Association for the Advancement of Science.

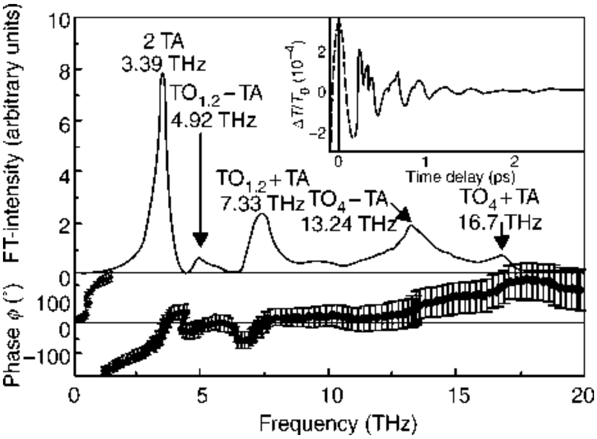

Figure 7 Fourier transform of the time domain data (inset) of the probe differential transmitivity of KTaO$_3$ at room temperature. The excited phonon combinations are indicated in the figure. Reproduced with permission from Bartels A, Dekorsy T and Kurz H (2000) Impulsive excitation of phonon-pair combination states by second-order Raman scattering. *Physical Review Letters* 84(13): 2981–2984. Copyright 2000 by the American Physical Society.

In KTaO$_3$ it was also possible to induce a time-dependent covariance between two different independent modes by using second order RS to couple the two oscillators. In Figure 7 we show

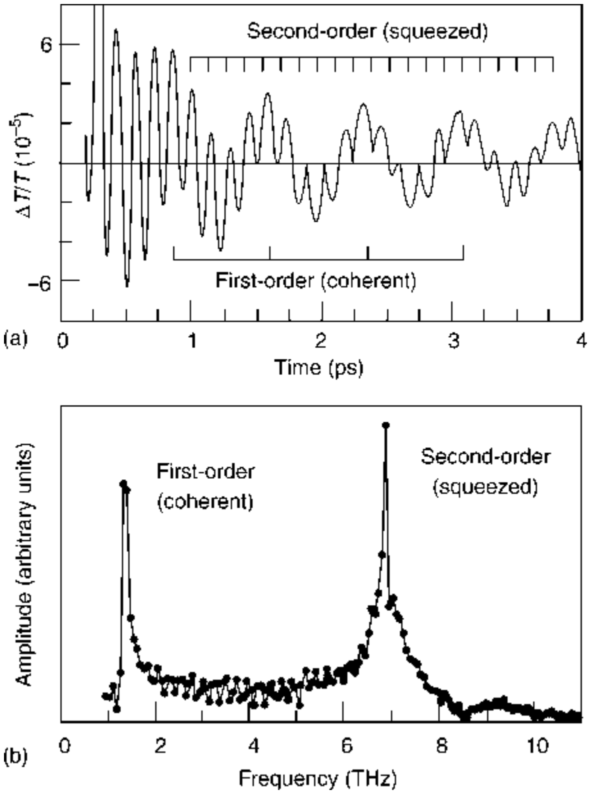

Figure 8 (a) The normalized transmitted probe intensity as a function of the probe delay for $SrTiO_3$ at $T = 7$ K. (b) Fourier transform of the time domain data. The 1.3 THz peak is the soft A_{1g}-phonon (coherent), and the 6.9 THz is the 2 TA overtone (squeezed). The second order feature is dominated by a sharp peak very close to twice the frequency of TA modes at X and M points of the Brillouin zone.

data revealing correlations between transverse optical (TO) and transverse acoustical (TA) modes.

$SrTiO_3$

This material undergoes a structural phase transition, called antiferro-distortive, at $T_C \approx 110$ K. Second-order RS is allowed at all temperatures, but first-order RS is only allowed for temperatures below T_C. Figure 8 shows data at low temperatures, revealing both first and second order features. From the comparison with spontaneous RS, the peak appearing in the $\mathcal{F}_{FT}$ at ~1.3 THz is due to first-order RS, while the peak at ~6.9 THz is due to second order RS (2TA overtone). The quantum state created by the ultrafast pulse is then a mixture of a coherent and a squeezed state. Hence the phonon wavefunction will evolve qualitatively as in the picture shown in Figure 1d.

List of Units and Nomenclature

Gaussian units

I_0	energy density [Jcm^{-2}]
E	electric field [statvolt cm^{-1}]

See also

Scattering: Stimulated Scattering; Raman Scattering. **Semiconductor Physics:** Light Scattering. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Dunn TJ, Sweetser JN, Walmsley IA and Radzewicz C (1995) Experimental determination of the dynamics of a molecular nuclear wave-packet via the spectra of spontaneous emission. *Physical Review Letters* 74(6): 884–887.

Garraway BM and Suominen KA (1995) Wave-packet dynamics: new physics and chemistry in femto-time. *Reports on Progress in Physics* 58(4): 365–419.

Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science* 275(5306): 1638–1640.

Hayes W and Loudon R (1978) *Scattering of Light by Crystals*. New York: John Wiley & Sons, Inc.

Lindemann F (1910) Über die Berechnung molekularer Eigenfrequenzen. *Zeitschrift Für Physik* 11: 609–612.

Walls DF and Milburn GJ (1994) Generation and Applications of Squeezed Light. *Quantum Optics*, pp. 137–174. Berlin Heidelberg: Springer Verlag.

Zerbe C and Hanggi P (1995) Brownian parametric quantum oscillator with dissipation. *Physical Review E* 52(2): 1533–1543.

RELATIVISTIC NONLINEAR OPTICS

D P Umstadter, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Immediately after the invention of the laser in 1960, light could be focused to sufficient intensity to cause nonlinear optical effects in atomic media. In conventional nonlinear optics, where the electric field of a light wave E is much smaller than an atomic field, $E_{\mathrm{at}} = 3 \times 10^9$ V/cm, various nonlinear phenomena, such as self-focusing, harmonic generation, and Raman scattering, arise due to the anharmonic motion of electrons in the combined fields of atom and laser. Approximate analytical solutions can be obtained by means of perturbation expansion methods, using E/E_{at} as the expansion parameter. At higher light fields, when E approaches E_{at}, this method breaks down and the medium becomes photo-ionized, creating a plasma, as illustrated by Figure 1. Further increases in light intensity enabled nonlinear optical effects of even these free plasma electrons (see Figure 2). The nonlinearity arises, in this case, because the electrons oscillate at relativistic velocities in laser fields that exceed 10^{11} V/cm, resulting in relativistic mass changes exceeding the electron rest mass and the light's magnetic field becomes important. The work done by the electromagnetic field (E) on an electron ($eE\lambda$) over the distance of a laser wavelength (λ), then approaches the electron rest mass energy ($m_e c^2$), where e is the elementary charge of an electron, m_e is the electron rest mass, and c is the speed of light. Effects analogous to those studied with conventional nonlinear optics – self-focusing, self-modulation, harmonic generation, and so on – are all found, but based on this entirely different physical mechanism. Thus, a new field of nonlinear optics, that of relativistic electrons, has been launched, as illustrated by Figure 1.

One outcome of accessing this new optical regime is the generation of frequency-shifted light in a spectral region where there are no other compact sources. Another is the acceleration of other types of particles, such as positrons, ions, and neutrons. These novel radiation sources have properties (femtosecond duration, micron source size, MeV energy) that make them suitable for numerous applications in imaging and spectroscopy in basic research, as well as medical diagnostics, cancer therapy, energy production, and space propulsion. Rapid advancement is underway and new research tools, subfields, and commercial products are on the horizon, e.g., compact and ultrashort pulse duration laser-based electron accelerators and X-ray sources.

Another physical regime will be encountered at even higher intensities ($I\lambda^2 \simeq 10^{24}$ W/cm^2), when even protons will quiver relativistically, i.e., the work done on a proton over the distance of a laser wavelength approaches its rest mass energy. This might be called the nuclear regime of laser–plasma interactions, because of the fusion and fission reactions and the generation of pions, muons, and neutrinos that should occur as nuclei collide in such energetic plasmas.

The recent dramatic increase in light intensity was partly made possible by the development, in the last decade, of compact lasers that have the ability to amplify shorter light pulses. For instance, solid-state lasers use the technique of chirped-pulse amplification to generate femtosecond duration pulses. To accomplish this, an ultrashort, low-energy laser pulse is first stretched in time before it is amplified and then recompressed. Gas or dye lasers, using solid-state switches, have produced picosecond duration pulses. Present day advanced laser systems now have multi-terawatt peak powers and, when focused to micron spotsizes with adaptive optics, can produce electromagnetic intensities $I\lambda^2 \simeq 10^{21}$ W/cm^2, as illustrated by Figure 3. An example of a modern ultra-high-power (0.9 PW) solid-state (Ti:Sapphire) laser

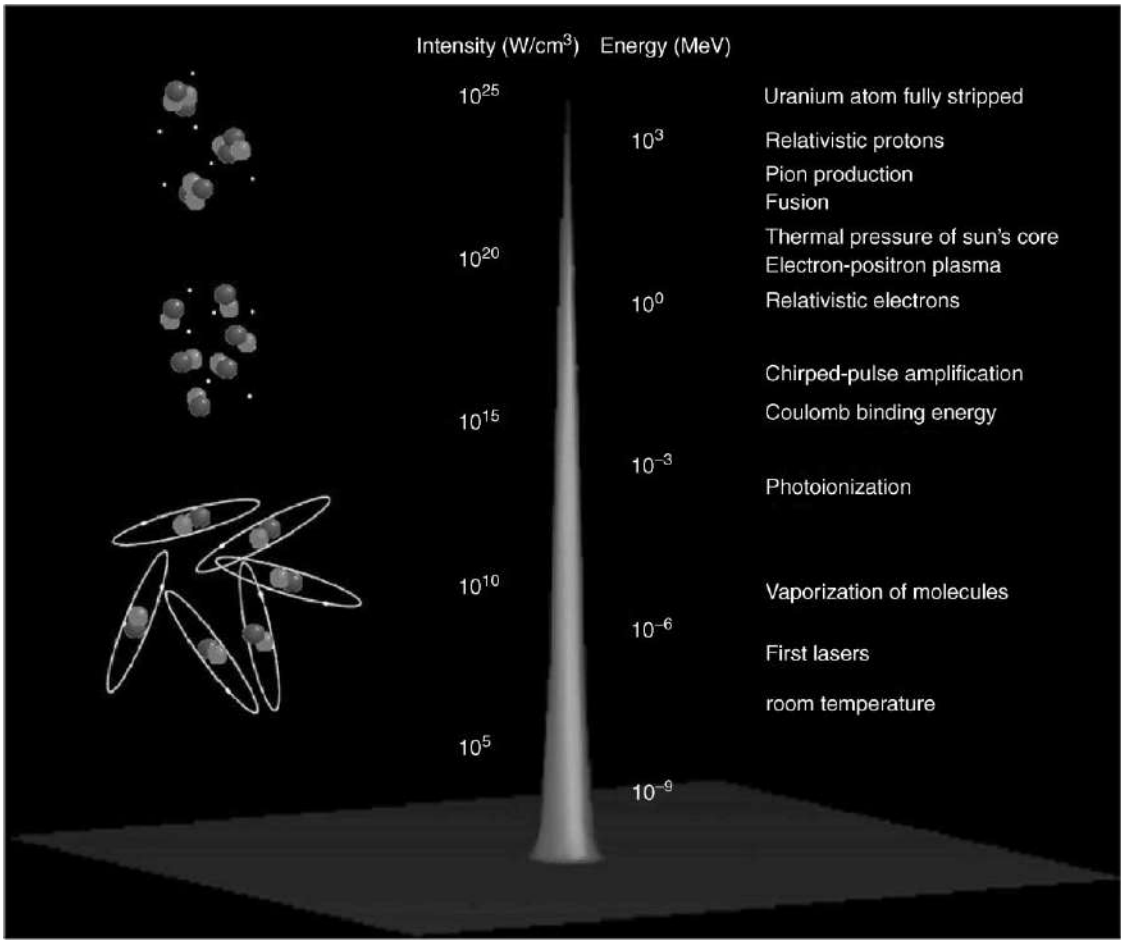
Intensity (W/cm^3)
Energy (MeV)
10^{25}
10^{20}
10^{15}
10^{10}
10^5
10^3
10^0
10^{-3}
10^{-6}
10^{-9}
Uranium atom fully stripped
Relativistic protons
Pion production
Fusion
Thermal pressure of sun's core
Electron-positron plasma
Relativistic electrons
Chirped-pulse amplification
Coulomb binding energy
Photoionization
Vaporization of molecules
First lasers
room temperature

Figure 1 The various regimes of laser–matter interactions, represented by the ideal laser pulse. As the intensity of laser light increases, so does the energy of electrons accelerated in the light field and the regime of conventional nonlinear optics with electrons bound to atoms is replaced by the regime of relativistic nonlinear optics with free electrons in relativistic plasmas. At the highest intensities, even protons become relativistic, giving rise to what might be called the regime of nuclear optics, in which various nuclear processes, such as fusion, can take place. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

system, located at the Advanced Photon Research Center, Kansai Research Establishment, Japan Atomic Energy Research Institute (JAERI), can be seen in the photograph shown in Figure 4.

This paper will discuss the fundamental concepts underlying what might be referred to more broadly as the field of high-field science. For more detailed descriptions of recent progress in experiment or theory, several review papers have been published on related topics: (i) relativistic nonlinear optics; (ii) high-intensity laser development; (iii) laser accelerators; (iv) intense laser–plasma interactions; and (v) relativistic scattering.

Definitions

For an infinite plane wave for the light wave propagating in the $\hat{z}$ direction (shown in Figure 5), the fields may be described by

$$\mathbf{E} = \hat{e}_x E_0 \exp[\mathrm{i}(kz - \omega t)] = \hat{e}_x E_0\, \mathrm{e}^{\mathrm{i}\psi} \qquad [1]$$

and

$$\mathbf{B} = \hat{e}_y B_0 \exp[\mathrm{i}(kz - \omega t)] = \hat{e}_y B_0\, \mathrm{e}^{\mathrm{i}\psi} \qquad [2]$$

where ω and k are the frequency and wavenumber of the laser light, respectively. For the ultrashort pulse durations of the lasers typically used to produce high fields, it is assumed that the ions do not have time to move because of their greater inertia. In this case, $m(\mathrm{dv}/\mathrm{d}t) \approx -e\mathrm{E}$ and the term v/c can be used as an expansion parameter. At even higher fields, electrons can be accelerated to nearly the speed of light as they quiver in the oscillating fields of the light wave. In this case, the force due to the magnetic field can no longer be ignored and the mass of the electron

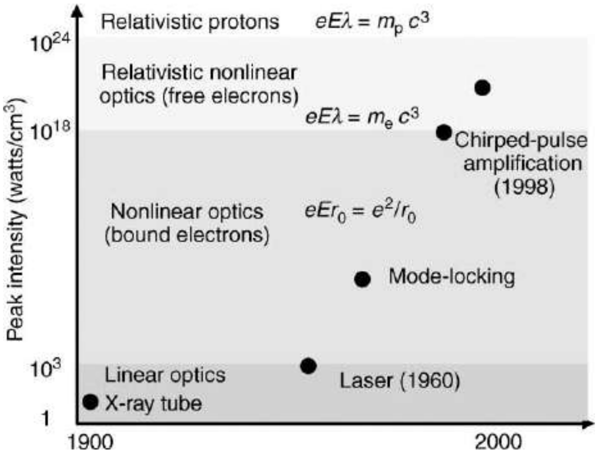

Figure 2 History of light sources over the last century. Each advance in laser power enables a new regime of optics. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

can increase as $m = \gamma m_0$, where m_0 is the electron rest mass and the relativistic gamma factor is given by

$$\gamma = \frac{1}{\sqrt{1 - \left(\frac{v}{c}\right)^2}} \qquad [3]$$

The threshold for the relativistic regime of optics is crossed when the electron momentum p exceeds the product of the electron rest mass m_0 and c.

For problems involving charged particle motion in intense electromagnetic fields, it is often convenient to introduce the vector potential: $\mathbf{A}(\mathbf{r}, t) = \mathrm{Re}\{\mathbf{A}_0\, \mathrm{e}^{\mathrm{i}\psi}\}$. From Maxwell's equation $\nabla \cdot B = 0$, it can be seen that without loss of generality, the magnetic field can be written as the curl of a potential $B = \nabla \times A$. Further, from Faraday's law in vacuum:

$$\nabla \times E = -\frac{1}{c}\frac{\partial B}{\partial t} = -\frac{1}{c}\frac{\partial(\nabla \times A)}{\partial t} \qquad [4]$$

it is clear that $\mathbf{E} = -(1/c)\partial \mathbf{A}/\partial t$. To include the space-charge field of the plasma, as we will do in the section on collective plasma response below, we introduce the potential Φ, and the definition:

$$\mathbf{E} = -\nabla\Phi - \frac{1}{c}\frac{\partial}{\partial t}\mathbf{A} \qquad [5]$$

Assuming the Coulomb gauge:

$$\nabla \cdot \mathbf{A} = 0 \qquad [6]$$

or $A_z = 0$ in 1D, then $\mathbf{A}_\perp$ represents the laser field.

If we neglect Φ for now, then, by Fourier analysis:

$$\mathbf{E} = \mathrm{Re}\left\{\frac{\mathrm{i}\omega}{c}\mathbf{A}_0\, \mathrm{e}^{\mathrm{i}\psi}\right\} \qquad [7]$$

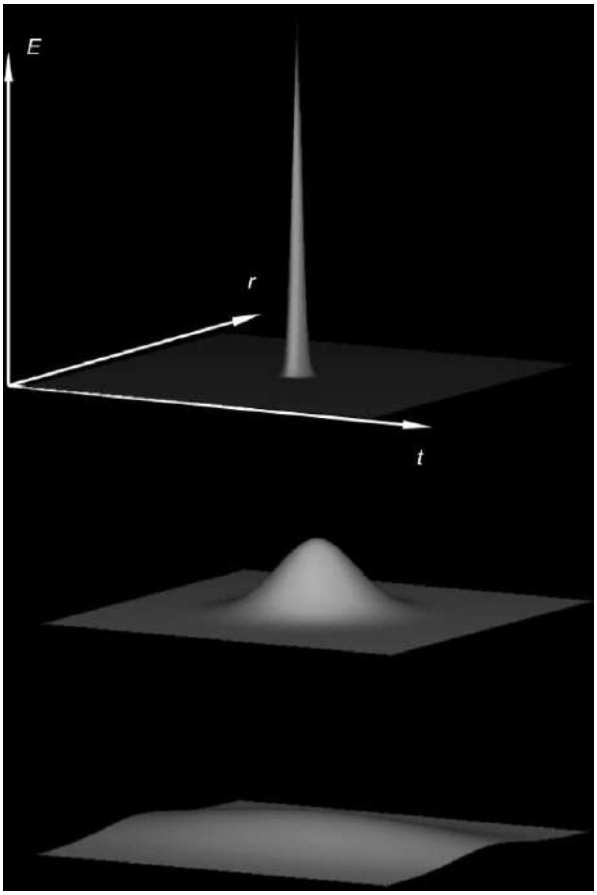

Figure 3 Laser light has become concentrated to ever-smaller regions of space (*r*) and time (*t*), dramatically increasing the peak electric field (*E*) at the laser focus. Prior to the development of chirped-pulse amplification (CPA), the energy of light was produced in long-duration pulses, as shown in the pulse of the bottom figure. After CPA, the pulse duration decreased dramatically as shown in the pulse in the middle. The latest improvement in laser technology has been the use of deformable mirrors, which has allowed lasers to be focused to a spatial dimension that is as small as the temporal dimension, a few laser wavelengths, as shown in the pulse on top. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

$$\mathbf{B} = \mathrm{Re}\{\mathrm{i}k\hat{z} \times \mathbf{A}_0\, \mathrm{e}^{\mathrm{i}\psi}\} \qquad [8]$$

The light intensity is given by the time-averaged Poynting's vector, which can be written for linear (LP), $\mathbf{A}_0 = A_0\hat{e}_y$ and circular polarizations (CP), $\mathbf{A}_0 = A_0(\hat{e}_x \pm \hat{e}_y)$ as:

$$I = \langle|\mathbf{S}|\rangle = \left\langle\left|\frac{c\mathbf{E} \times \mathbf{B}}{8\pi}\right|\right\rangle = \frac{\omega k}{16\pi}A_0^2 \qquad [9]$$

It is often the combination $I\lambda^3$ that matters, which has units of power and for linear polarization can be

Figure 4 Photograph of the 33-fs duration 0.9-PW-peak power laser system at JAERI. (Reproduced with permission from Yamakawa K (2004) Table-top lasers create ultrahigh peak powers. *OYOBUTURI* 73: 186–193.)

written as

$$I\lambda^2 = \frac{\omega k \lambda^2}{16\pi} A_0^2 = \frac{\pi}{2} c A_0^2 = \left[1.37 \times 10^{18} \frac{\mathrm{W}}{\mathrm{cm}^2} \mu\mathrm{m}^2\right] a_0^2 \qquad [10]$$

where the normalized vector potential is defined as

$$a_0 = \frac{eA_0}{mc^2} \qquad [11]$$

Relativistic effects become significant when a_0 approaches unity, which, as can be seen from eqn [10] occurs for 1 μm wavelength light at a laser intensity of $\sim 1.37 \times 10^{18}$ W/cm^2, corresponding to a electric field strength of 6×10^9 V/cm, which is four orders of magnitude greater than E_{at}.

Single-Particle Motion

The electron orbit, subject to a linearly polarized electromagnetic wave propagating in the $+z$ direction, is governed by the Lorentz equation:

$$m \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -e\left(\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B}\right) \qquad [12]$$

where v is the electron velocity, and E and B are the light's electric and magnetic field.

A zeroth order solution to eqn [12] is found by setting $v/c \ll 1$, which allows the term $(v/c) \times B$ to be neglected and γ from eqn [8] is set to unity. Integrating eqn [12] in this limit once over time (t) yields for the velocity:

$$\mathbf{v} = \mathrm{Re}\left\{\frac{e\mathbf{E}}{\mathrm{i}m\omega}\right\} = \frac{a_0}{c}\begin{cases} \hat{e}_y \cos\psi, & \mathrm{LP} \\ (\hat{e}_x \cos\varphi \mp \hat{e}_y \sin\psi), & \mathrm{CP} \end{cases} \qquad [13]$$

An electron in low-intensity light oscillates with this velocity in a straight line along the polarization vector ($\hat{e}_y$), which when normalized to c, is the normalized vector potential a_0. Integrating again yields for the

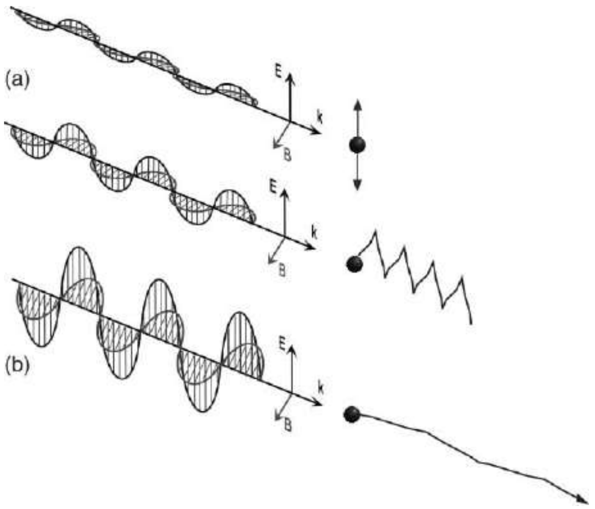

Figure 5 Classical optics versus relativistic optics. (a) In classical optics, the amplitude of the light wave is small, electrons oscillate in the direction of the electric field at the light's frequency, and there is no displacement along the light's propagation direction. Note that only the *E* field acts on the electron and the electron-oscillation velocity is very small compared with the speed of light. (b) In relativistic optics, the amplitude of the light wave is very large, the light's magnetic field becomes important and the combined action of the electric and magnetic fields push the electron forward. In this case, the electron velocity becomes close to the speed of light and its motion is a figure-eight superimposed upon a steady drift. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

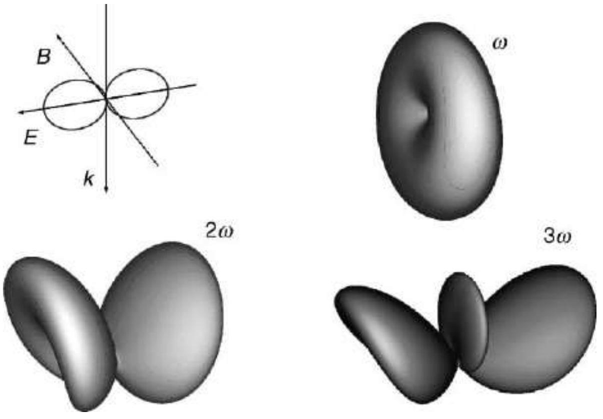

Figure 6 Harmonics driven by relativistic Thomson scattering as the electrons in high-intensity laser fields (a_0^2) undergo figure-eight motion display unique angular distributions. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

transverse displacement:

$$\mathbf{x} = \mathrm{Re}\left\{\frac{e\mathrm{E}}{m\omega^2}\right\} = \frac{a_0}{\omega c}\begin{cases} \hat{e}_y \sin\psi, & \text{LP} \\ (\hat{e}_y \sin\varphi \mp \hat{e}_z \cos\psi), & \text{CP} \end{cases} \qquad [14]$$

Thus, for the $a_0 \sim 1$, the electron excursion during its oscillation is approximately λ.

A first-order approximation for the electron motion can be found by substituting the zeroth order velocity eqn [13] into the $v/c \times B$ term of eqn [12]. The latter then becomes proportional to

$$E \times B \propto \frac{a_0^2}{2}[1 + \cos(2\omega_0)]\hat{e}_z \qquad [15]$$

In the frame in which the electron is on the average at rest, the relativistic motion of an electron is thus described by a figure eight, oscillating twice in the $\hat{e}_z$ or $\hat{k}$ direction for every once in the polarization direction ($\hat{e}_x$); i.e., a figure of eight motion. This originates from the fact that $v \times B \propto E \times B \propto E^2\hat{k}$, which is a product of two functions that vary sinusoidally at frequency ω and thus varies itself at frequency 2ω. In the lab frame, this transverse motion is superimposed upon a steady drift in the ($\hat{e}_z$) direction (shown in Figure 5), originating from the DC term in eqn [15]. The next order approximation would include the mass shift $m = \gamma m_0$. As the field strength increases ($a_0^2 \gg 1$), the longitudinal motion ($\propto a_0^2$) begins to dominate the transverse motion ($\propto a_0$), as shown in Figure 5. This is described more formally below.

In the regime $a_0 \lesssim 1$, electrons radiate photons at harmonics of a modified laser frequency ω_0, with each harmonic order having its own unique angular distribution, as shown in Figure 6. The radiation at the fundamental is the usual donut pattern, with a maximum in the direction perpendicular, and a minimum along, the polarization vector ($\hat{e}_x$). The second harmonic has two emission lobes with maxima pointing at an angle between $\hat{e}_z$ and $\hat{e}_x$. An additional lobe is added for each additional harmonic order. This is referred to as nonlinear or relativistic Thomson scattering. The unique angular distributions of the second and third harmonics emitted from nonlinear relativistic Thomson scattering were observed experimentally and are shown in Figure 7. As will be shown in detail in the section on radiation from relativistic electrons, below, the situation is more complex for $a_0 \gtrsim 1$, and the scattered light is no longer simply harmonic.

Constants of the Motion

Several constants of the motion can be found from an exact treatment for the motion. The starting point is the relativistically correct Lagrangian, which is written as

$$L(\mathbf{r}, \mathbf{v}, t) = -mc^2\sqrt{1 - \frac{v^2}{c^2}} + \frac{q}{c}\mathbf{v}\cdot\mathbf{A} + q\phi \qquad [16]$$

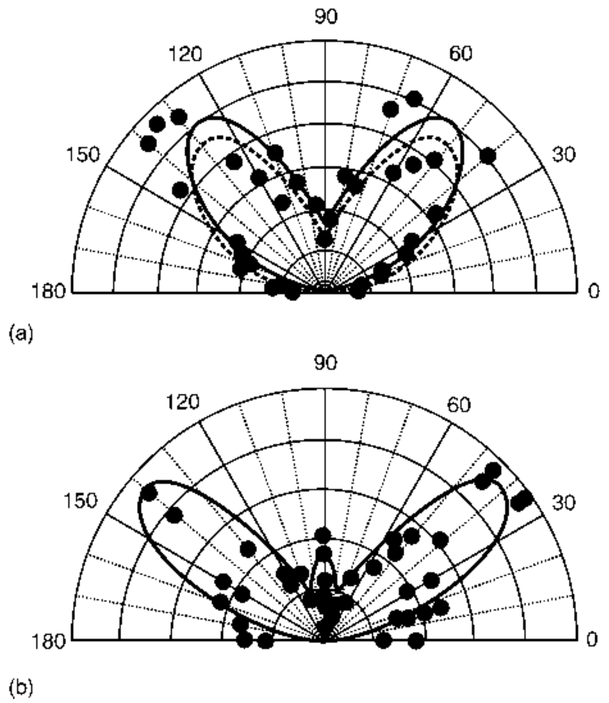

Figure 7 Angular pattern of higher-order harmonic light. Shown are polar plots of the intensity of the second-harmonic light (top) and third harmonic (bottom) as a function of azimuthal angle. Filled circles, experimental data; solid and dashed lines, theoretical results. (Reproduced with permission from Chen S-Y, Maksimchuk A and Umstadter D (1998) Experimental observation of relativistic nonlinear Thomson scattering. *Nature* 396: 653.)

where ϕ is the scalar potential associated with electrostatic fields. The Lorentz equation, eqn [12], is derived from the Euler–Lagrange equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \mathbf{v}} - \frac{\partial L}{\partial \mathbf{r}} = 0 \qquad [17]$$

For an infinite plane wave, in which the Lagrangian is independent of space in the transverse direction, eqn [17] yields the conservation of transverse canonical momentum:

$$\frac{\partial L}{\partial \mathbf{v}_\perp} = \mathbf{p}_\perp + \frac{q}{c}\mathbf{A}_\perp = \text{constant} \qquad [18]$$

The sum of the transverse momentum and field strength remain constant. The next constant follows from $\mathrm{d}H/\mathrm{d}t = -\partial L/\partial t$, yielding the relation:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{\partial L}{\partial t} = c\frac{\partial L}{\partial z} = c\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial v_z} = c\frac{\mathrm{d}p_z}{\mathrm{d}t} \qquad [19]$$

where E is the time-dependent energy of the particle, which yields our second constant of the motion:

$$E - cp_z = \text{constant} \qquad [20]$$

For a particle initially at rest, the kinetic energy E_{kin} becomes

$$E_{\mathrm{kin}} = E - mc^2 = p_z c \qquad [21]$$

which yields

$$E_{\mathrm{kin}} = \frac{p_\perp^2}{2m} = p_x c = mc^2(\gamma - 1) \qquad [22]$$

Thus, the electron scattering angle θ is related to the transverse and longitudinal momenta or the kinetic energy by

$$\tan^2\theta = \left(\frac{p_\perp}{p_z}\right)^2 = \frac{2mE_{\mathrm{kin}}}{(E_{\mathrm{kin}}/c)^2} = \frac{2}{\gamma - 1} \qquad [23]$$

The angle of electrons produced by photo-ionization with intense lasers has been shown experimentally to obey the conservation of canonical momentum, as demonstrated in experiments that studied the angular distribution of relativistic electrons emitted from barrier-suppression ionization of atoms in intense laser fields (shown in Figure 8). The transverse normalized vector potential, $\mathbf{a} = e\mathbf{A}_\perp/mc^2$, is thus equal to the transverse normalized momentum:

$$\hat{\mathbf{p}}_\perp = \frac{\mathbf{p}_\perp}{mc} = \mathbf{a} = (0, a_y, a_x) \qquad [24]$$

The normalized kinetic energy can also be written in terms of a:

$$\hat{E}_{\mathrm{kin}} = \frac{E_{\mathrm{kin}}}{mc^2} = \gamma - 1 = \hat{p}_z = \frac{\hat{p}_\perp^2}{2} = \frac{a^2}{2} \qquad [25]$$

This can be rewritten as

$$\gamma = 1 + \frac{a^2}{2} \qquad [26]$$

From eqn [25], the normalized momenta in each direction can be found:

$$\hat{p}_z = \gamma\beta_z = \frac{\gamma}{c}\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{a^2}{2} \qquad [27]$$

$$\hat{p}_y = \gamma\beta_y = \frac{\gamma}{c}\frac{\mathrm{d}y}{\mathrm{d}t} = a_y \qquad [28]$$

$$\hat{p}_x = \gamma\beta_x = \frac{\gamma}{c}\frac{\mathrm{d}x}{\mathrm{d}t} = a_x \qquad [29]$$

Note that the momentum in the longitudinal ($\hat{e}_z$) direction is proportional to a^2 while in the two

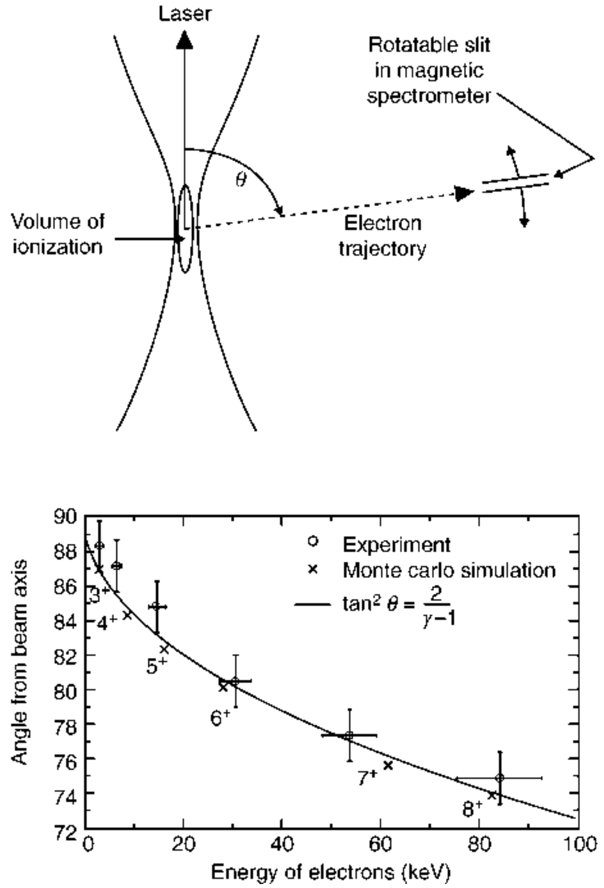

Figure 8 Top: Experiment to measure the angle of electron ejection during photo-ionization. Bottom: Kinetic energy versus angle during the production of 3^+ through 8^+ of neon. (Reproduced with permission from Moore CI, Knauer JP and Meyerhofer DD (1995) Observation of the transition from Thomson to Compton scattering in multiphoton interactions with low-energy electrons. *Physics Review Letters* 74: 2439. Copyright (1995) by the American Physical Society.)

transverse directions it is proportional to a. Equations [26]–[29] indicate that in the limit of large a, the normalized velocities are given by

$$\beta_z = \frac{a^2/2}{1+a^2/2} \to 1, \qquad \beta_y = \frac{a_y}{1+a^2/2} \to 0, \qquad \beta_x = \frac{a_x}{1+a^2/2} \to 0 \qquad [30]$$

This indicates that in the deeply relativistic regime, the electron is accelerated to near c in the longitudinal direction, while in the transverse directions its velocity goes to zero. It follows that, as a increases and the electron moves in phase with the light wave, relativistic Thomson scattering will be significantly reduced. Integrating eqn [29] in the variable $\tau = t - x(t)/c$ gives

$$\gamma\frac{\mathrm{d}}{\mathrm{d}t} = \gamma\frac{\mathrm{d}\tau}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}\tau} = \gamma\left(1 - \frac{1}{c}\frac{\mathrm{d}z}{\mathrm{d}\tau}\right)\frac{\mathrm{d}}{\mathrm{d}\tau} = \left(1 + \frac{a^2}{2} - \frac{a^2}{2}\right)\frac{\mathrm{d}}{\mathrm{d}\tau} = \frac{\mathrm{d}}{\mathrm{d}\tau} \qquad [31]$$

which allows the equations for the displacement to be written simply as

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = c\frac{a^2}{2} \qquad [32]$$

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = ca_y \qquad [33]$$

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = ca_x \qquad [34]$$

For a box-shaped pulse, $a_x = a_0\cos(\omega\tau)$ for $0 < \tau < N(2\pi/\omega)$, $a_z = 0$, $a^2 = a_x^2\cos^2(\omega\tau)$, the trajectory for an electron initially at rest at $x = y = z = 0$ at $\tau = 0$ is then

$$x(\tau) = \frac{ca_0^2}{2}\int_0^\tau \cos^2(\omega\hat{\tau})\mathrm{d}\hat{\tau} = \frac{ca_0^2}{4}\left[\tau + \frac{1}{2\omega}\sin(2\omega\tau)\right] \qquad [35]$$

$$y(\tau) = ca_0\int_0^\tau \cos(\omega\hat{\tau})\mathrm{d}\hat{\tau} = \frac{ca_0}{4}\sin(\omega\tau) \qquad [36]$$

Thus, as we saw from the first-order appoximation above, eqn [15], the motion consists of a drift along $\hat{x}$:

$$x_\mathrm{d}(t) = \frac{a_0^2}{a_0^2+4}ct \qquad [37]$$

and a figure-eight motion in the drift frame of reference:

$$ky = a_0\sin(\omega\tau), \qquad k(z - z_\mathrm{d}) = \frac{a_0}{8}\sin(2\omega\tau) \qquad [38]$$

In the strongly relativistic regime, because the electron is strongly accelerated in the direction of the light wave, the harmonic emission from relativistic Thomson scattering loses its dependence on the laser frequency, ω, but instead depends only on the direction of observation, the amplitude a_0 and the initial phase of the electron, relative to the light wave. Electrons that do not begin their trajectories at zero or π phase with respect to the light wave can drift in the transverse direction. For pulses with finite spatial width, such as a Gaussian pulse with field strength peaked on axis, the ponderomotive force of the light can also act to expel the electrons transversely from the axis.

For circular polarization:

$$\mathbf{a}(\mathbf{r}, t) = \mathrm{Re}\{a_0(\hat{e}_y \pm \mathrm{i}\hat{e}_x)\mathrm{e}^{-\mathrm{i}\omega\tau}\} \qquad [39]$$

the electron drifts as it did in the case of linear polarization and described in eqn [37] but orbits in a helical trajectory:

$$y(\tau) = \frac{ca_0}{\sqrt{2}\omega}\sin(\omega t/\gamma) \qquad [40]$$

$$x(\tau) = \mp\frac{ca_0}{\sqrt{2}\omega}\cos(\omega t/\gamma) \qquad [41]$$

Rather than oscillate in amplitude, the relativistic factor remains constant:

$$\gamma = 1 + \frac{a_0^2}{4} \qquad [42]$$

In either polarization case, the light pulse eventually overtakes the electron, leaving the latter at rest. However, net acceleration in vacuum can be obtained only under special circumstances, described by the Lawson–Woodward criterion, such as in the near field of metallic surfaces; with high-order focusing modes instead of infinite plane waves; or with mixed frequencies, or with optics to terminate the light but not the electrons. Regions of a tightly focused Gaussian pulse, where the accelerating field propagates at sub-luminal velocities also exist, but they are effective for only extremely large values of *a* and thus only very short acceleration lengths.

Role of Initial Phase

If we wish to understand the role of the electron's initial phase, the relativistic Lorentz equation (eqn [12]) may also be written, for an $\hat{x}$-polarized wave traveling in the $\hat{z}$ direction, as

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma\beta) = (\hat{x} - \hat{z}\times\beta)a\cos(t-z) \qquad [43]$$

where we normalize time by $1/\omega_0$, velocity by c, and distance by c/ω_0. In eqn [43], $a = eE_0/m\omega_0 c$ is the dimensionless parameter measuring the electric field strength, $\gamma = (1-\beta_x^2-\beta_y^2-\beta_z^2)^{-1/2}$ is the relativistic mass factor, and $\beta = (\beta_x, \beta_y, \beta_z)$ is the electron velocity (in units of c). The electron orbit, subject to the following general initial conditions at time $t = 0$:

$$x = 0, \qquad y = 0, \qquad z = z_{\mathrm{in}} \qquad [44]$$

$$\beta_x = \beta_{x0}, \qquad \beta_y = \beta_{y0}, \qquad \beta_x = \beta_{z0} \qquad [45]$$

$$\beta_{x0} = \beta_{y0} = 0 \qquad [46]$$

has a closed form solution when it is expressed parametrically: $t = t(\theta)$, $r = r(\theta)$, $\beta = \beta(\theta)$, where θ is the phase of the wave, defined by

$$\theta = t - z \qquad [47]$$

Note that $\beta_0 = (\beta_{x0}, \beta_{y0}, \beta_{z0})$ is the unperturbed velocity of the electron ($a = 0$ limit) and that the initial phase that the electron sees is $\theta_{\mathrm{in}} = -z_{\mathrm{in}}$ according to eqns [44] and [47]. This phase can be important in the ionization of the gas by an intense laser.

For the special case $\beta_{x0} = 0$, $\beta_{y0} = 0$, one finds $\beta_y = 0$, and the orbital equation yields the following closed form solution:

$$\gamma = \gamma_0 + \frac{a^2(\sin\theta - \sin\theta_{\mathrm{in}})^2}{2\gamma_0(1-\beta_{z0})} \qquad [48]$$

$$\gamma\beta_z = \gamma - \gamma_0(1-\beta_{z0}) \qquad [49]$$

$$\gamma\beta_x = \frac{1}{\gamma_0(1-\beta_{z0})}a(\sin\theta - \sin\theta_{\mathrm{in}}) \qquad [50]$$

$$x = \frac{a[(\cos\theta_{\mathrm{in}} - \cos\theta) - (\theta-\theta_{\mathrm{in}})\sin\theta_{\mathrm{in}}]}{\gamma_0(1-\beta_{z0})} \qquad [51]$$

$$t = \frac{(\theta-\theta_{\mathrm{in}})}{1-\beta_{z0}}\left[1 + \frac{a^2}{2}\left(\frac{1}{2} + \sin^2\theta_{\mathrm{in}}\right)\right] + \frac{a^2(1+\beta_{z0})}{2(1-\beta_{z0})}\left[-\frac{\sin 2\theta}{4} + 2\cos\theta\sin\theta_{\mathrm{in}} - \frac{3\sin 2\theta_{\mathrm{in}}}{4}\right] \qquad [52]$$

In eqns [48]–[52], $\gamma_0 = (1-\beta_{z0}^2)^{-1/2}$ and β_{z0} may either be negative (counter-propagating against the laser), or zero (Thomson scattering), or positive (co-propagating with the laser). Note that the velocity components β_x and β_z are given as explicit functions of θ according to eqns [49] and [50], upon using eqn [48]. They are periodic functions of θ of period 2π. The period, T, of this periodic (Figure 8) motion is thus equal to the increase in t as θ increases by 2π. Thus, we obtain from eqn [53]:

$$T = \frac{2\pi}{1-\beta_{z0}}\left[1 + \frac{a^2}{2}\left(\frac{1}{2} + \sin^2\theta_{\mathrm{in}}\right)\right] \qquad [53]$$

The parametric solution for the z-coordinate of the electron orbit is given by $z = t - \theta$, in which t is given by eqn [52]. Over one orbital period, T, the electron undergoes a net displacement $r_0 = (x_0, 0, z_0)$ where x_0 is given by the increase in eqn [51] as θ increases by 2π, and z_0 is simply $T - 2\pi$:

$$\begin{aligned} z_0 &= T - 2\pi \\ &= \frac{2\pi}{1-\beta_{z0}}\left[\beta_{z0} + \frac{a^3}{2}\left(\frac{1}{2} + \sin^2\theta_{\text{in}}\right)\right] \end{aligned} \qquad [54]$$

$$x_0 = \frac{-2\pi a \sin\theta_{\text{in}}}{\gamma_0(1-\beta_{z0})} \qquad [55]$$

Note that the electron trajectory depends on a, β_{z0}, and θ_{in} in a rather complicated manner. Accordingly, the fundamental frequency ω_1 of the radiation spectrum depends on these three quantities. For backscattered radiation ($\hat{n} = -\hat{z}$), the dependence on electron beam and on the laser becomes decoupled.

Following is a useful formula relating the change of t with respect to θ:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = 1 - \beta_z = \frac{\gamma_0(1-\beta_{z0})}{\gamma} \qquad [56]$$

which may be verified from eqns [47] and [49]. Equation [56] is also valid for arbitrary values of β_{x0}, β_{y0}, β_{z0}, and θ_{in}, in which case γ_0 is the electron's relativistic mass factor in the absence of the laser.

Radiation from Relativistic Electrons

An electron with displacement $r(t)$ and velocity $v(t)$ carries a current density $J(r,t) = ev(t)\delta r - r(t)]$, whose Fourier transform $J_{\text{F}}(k,\omega)$ may easily be obtained. The total work done, W, in ergs, performed by the current $\mathbf{J}(\mathbf{r},t)$ on the electric field $\mathbf{E}(\mathbf{r},t)$ is given by

$$\begin{aligned} W &\equiv \int \mathrm{d}\omega\,\mathrm{d}\Omega \frac{\mathrm{d}^2 W}{\mathrm{d}\omega\,\mathrm{d}\Omega} \\ &= \int \mathrm{d}\omega\,\mathrm{d}\Omega \frac{\omega^2}{c^2}\left[|\mathbf{J}_{\text{F}}(\mathbf{k},\omega)|^2 - |\mathbf{n}\cdot\mathbf{J}_{\text{F}}(\mathbf{k},\omega)|^2\right] \end{aligned} \qquad [57]$$

where the $\mathbf{k}$-space differential volume $\mathrm{d}^3\mathbf{k} = k^2\,\mathrm{d}k\,\mathrm{d}\Omega$ is expressed in terms of the solid angle (Ω) in the direction of the unit vector $\mathbf{n} = c\mathbf{k}/\omega$. We immediately obtain from eqn [56], under the far-field approximation:

$$\frac{\mathrm{d}^2 W}{\mathrm{d}\Omega\,\mathrm{d}\omega} = \frac{e^2\omega^2}{4\pi^2 c}|\mathbf{n}\times[\mathbf{n}\times\mathbf{F}(\omega)]|^2 \qquad [58]$$

$$\mathbf{F}(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\,\beta(t)\mathrm{e}^{\mathrm{i}\omega[t-\mathbf{n}\cdot\mathbf{r}(t)/c]} \qquad [59]$$

Equation [58] gives the energy radiated by the electron in the direction of the unit vector, per unit solid angle, per unit frequency ω. Radiation damping is ignored throughout.

Let us consider the simplest case where the electron orbit is strictly a periodic function of time with period T and a net displacement $\mathbf{r}_0$ per period. Thus, we have for all integers m (positive, negative, or zero):

$$\beta(t+mT) = \beta(t), \quad \mathbf{r}(t+mT) = m\mathbf{r}_0 + \mathbf{r}(t) \qquad [60]$$

Equation [59] may then be written as, because of Thomson scattering the energy of the scattered photon is much less than the electron rest mass in the average rest frame of the electron, i.e.:

$$\frac{\hbar\omega_{\text{L}}}{mc^3} \ll 1 \qquad [61]$$

$$\begin{aligned} \mathbf{F}(\omega) &= \sum_{m=-\infty}^{\infty}\int_{mT}^{(m+1)T} \mathrm{d}t\,\beta(t)\mathrm{e}^{\mathrm{i}\omega[t-\mathbf{n}\cdot\mathbf{r}(t)/c]} \\ &= \sum_{m=-\infty}^{\infty} f(\omega)\mathrm{e}^{\mathrm{i}m\omega[T-\mathbf{n}\cdot\mathbf{r}_0/c]} \end{aligned} \qquad [62]$$

where

$$f(\omega) = \int_0^T \mathrm{d}t\,\beta(t)\mathrm{e}^{\mathrm{i}\omega[t-\mathbf{n}\cdot\mathbf{r}(t)/c]} \qquad [63]$$

and we have used eqn [60]. Upon using $\sum_m \mathrm{e}^{\mathrm{i}mx} = \sum_m 2\pi\delta(x - 2m\pi)$ in the last infinite sum in eqn [64] and the property of the Dirac delta function, $\delta(ax) = (1/a)\delta(x)$, we obtain from eqns [62] and [63], the following expression for the spectrum:

$$\mathbf{F}(\omega) = \sum_{m=-\infty}^{\infty} \mathbf{F}_m\delta(\omega - m\omega_1) \qquad [64]$$

$$\omega_1 = \frac{2\pi}{T - \mathbf{n}\cdot\mathbf{r}_0/c}, \qquad [65]$$

$$\mathbf{F}_m = \frac{\omega_1}{2\pi}\int_0^T \mathrm{d}t\,\bar{\beta}(t)\mathrm{e}^{\mathrm{i}m\omega_1[t-\mathbf{n}\cdot\mathbf{r}(t)/c]} \qquad [66]$$

Note from eqn [64] that the radiation spectrum is discrete, for strictly periodic motion of the electron. The base frequency of this spectrum, ω_1, depends on the periodicity (T) of the electron, on the electron's net displacement ($\mathbf{r}_0$) in one such period, and on the direction ($\hat{n}$) in which the radiation is observed. Thus, the radiation spectrum is, in general, not at the harmonic frequency of the laser (nor at the harmonics of the electron orbital frequency, $2\pi/T$). It would thus be wrong to simply insert $\omega = n\omega_0$ in eqn [63] and to replace the electron's orbital period T there by the laser's optical period $2\pi/\omega_0$ and consider the resultant value of that integral to give the spectral amplitude of the radiation at the nth harmonic of the laser frequency. Erroneous conclusions regarding generation of high laser harmonics have appeared in the literature based on such an intuitive (but incorrect) substitution.

The power, p_m (in erg/s), radiated at the harmonic frequency $\omega = m\omega_1$ per unit solid angle in the direction of the unit vector $\hat{n}$ is then given by (from eqns [58] and [64]):

$$p_m = \frac{e^2 m^2 \omega_1^2}{4\pi^2 c} |\mathbf{n} \times \mathbf{F}_m|^2 \qquad [67]$$

where the dimensionless spectral amplitude $\mathbf{F}_m$ is given in eqn [66]. It is easy to show from eqns [60] and [65] that the integrand in eqn [66] is a periodic function of t of period T. Integrals of this type are readily evaluated by the Romberg method. The radiation spectrum observed *exactly* in the forward direction of the laser ($\hat{n} = \hat{z}$) always has only one discrete frequency, $\omega = \omega_1 = \omega_0$, which is easily shown from eqn [65] upon using the first equality of eqn [53]. This statement is true regardless of the velocity of the electron or the laser intensity, and may easily be deduced from eqn [66] for this case. However, for an energetic electron beam that is almost co-propagating with the laser, such as that produced by the laser itself, high harmonics at the laser frequency may be observed in the direction just slightly off the laser direction.

In the backscattering direction of the laser ($n = -z$), if we set $\theta_{\text{in}} = 0$, one obtains the following expressions for ω_1 and p_m, the backscatter power at $\omega = m\omega_1$ (from eqn [67]):

$$\frac{\omega_1}{\omega_0} = \left(\frac{2}{2+a^2}\right)\left(\frac{1-\beta_{z0}}{1+\beta_{z0}}\right) = \left(\frac{2}{2+a^2}\right)\gamma_0^2(1-\beta_{z0})^2 \qquad [68]$$

$$p_m = \frac{A}{\gamma_0^2(1-\beta_{z0})^2}\left(\frac{\omega_1}{\omega_0}\right)^4 s_m \qquad [69]$$

where $A = e^2\omega_0^2/4\pi^2 c = 0.69[\lambda(1\ \mu\text{m})]^{-2}$ erg/s, $s_m = 0$ for $m = 0, \pm 2, \pm 4, \ldots$, and for $m = \pm 1, \pm 3, \pm 5, \ldots$,

$$s_m = (a\pi)^2 m^2 \left[J_{\frac{(m-1)}{2}}(m\kappa) - J_{\frac{(m+1)}{2}}(m\kappa)\right]^2 \qquad [70]$$

$$\kappa = \frac{a^2}{2(a^2+2)} \qquad [71]$$

The $J_\nu(x)$ above are Bessel functions of the first kind of order ν. They also appear in the quantity '$[JJ]$' or '$F_m(K)$' in the FEL/synchrotron literature, where K is the undulator/wiggler parameter K. Setting $K = a$, one finds $s_m = \pi^2(1 + K^2/2)^2 F_m(K)$. It is easy to show that $s_m = s_{-m}$ for all odd integers m. Note that the relative spectral shape of s_m depends only on a, and is independent of the electron beam energy. The discrete spectrum for small a approaches a continuum for $a^2 \gg 1$.

The maximum values of s_m, occurring at $m = M$ with a value s_M, are shown in Figure 9. Note that the frequency component $\omega = M\omega_1$ contains the highest backscattered power. In terms of the laser frequency, the frequency component $\omega = N\omega_0$ would contain the highest backscatter power, where $N = M\omega_1/\omega_0$. The total backscatter power, P_{T} (in ergs/s), per unit solid angle in the $\mathbf{n} = -\mathbf{z}$ direction is then given by $P_{\text{T}} = \Sigma p_m$, where the sum is taken over all odd values of m:

$$P_{\text{T}} \approx \frac{\gamma_0^6(1-\beta_{z0})^6}{(\lambda/1\ \mu\text{m})^3} \times \begin{cases} 13.7a^3, & (a < 0.3) \\ 11.1/a, & (a \gg 1) \end{cases} \qquad [72]$$

To summarize, due to the rapid acceleration of electrons in the direction of the light wave and decrease of the oscillation frequency in the deeply relativistic regime, the photon energy from relativistic nonlinear scattering scales only linearly with laser field strength, $\sim a$. However, the relativistic motion of the electron results in a reduction in the angle of the scattered light, such that the harmonics are generated in a low-divergence-angle forward-propagating beam. There is also a relativistic Doppler shift ($\sim\gamma^2$). For instance, this mechanism will be used in high-energy physics experiments to cleanly make constituent particles in the gamma–gamma ($\gamma\gamma$) collider, in which gamma rays with energy 200 GeV will be generated by Compton scattering 1 eV photons from 250 GeV energy conventionally accelerated electron beams ($\gamma = 5 \times 10^{11}$).

Using a much smaller, laser-based accelerator (discussed in the following section), a 1 eV photon can be upshifted by this Doppler shift to an energy of 50 keV, corresponding to subatomic spatial resolution and of interest in medical diagnostics, by an electron beam with only $\gamma = 200$ (100 MeV). In this case, the maximum efficiency is obtained for laser fields a of order unity. When compared with conventional light sources based on cm-wavelength magnetostatic wigglers, the electromagnetic wigglers

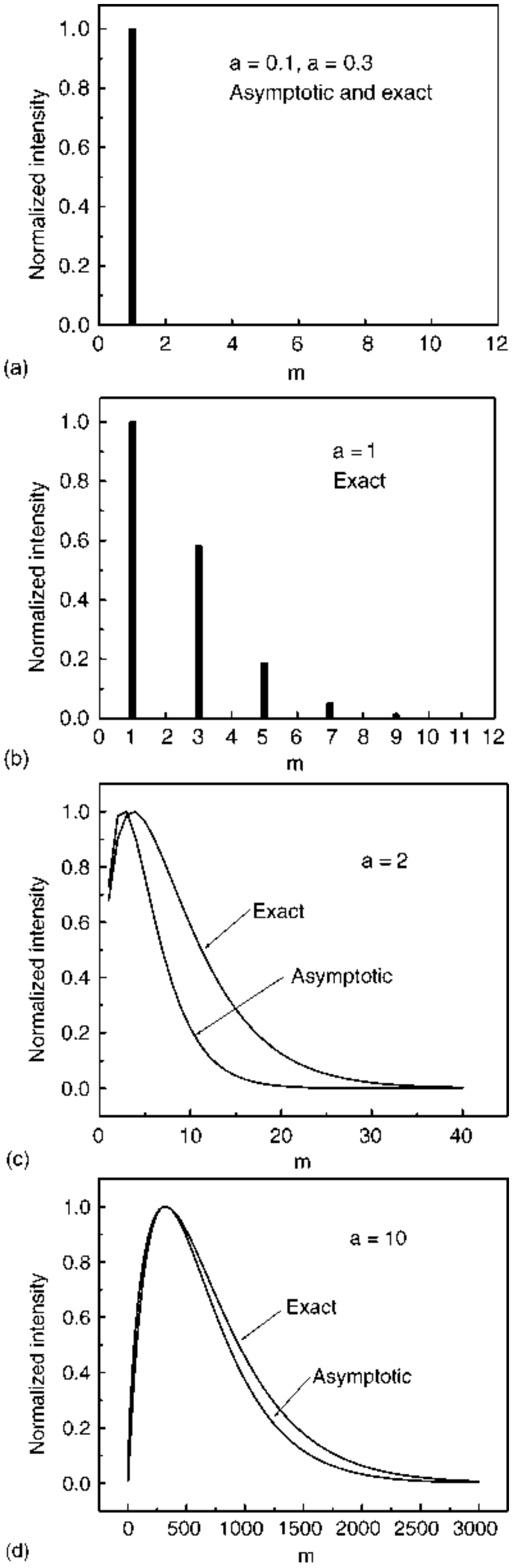

Figure 9 Normalized spectral distribution of s_m, at frequency $\omega = m\omega_1$. Here, s_m is normalized with respect to the maximum value s_M, occurring at $m = M$. (Reproduced from Lau Y, He F and Umstadter D (2003) Nonlinear Thomson scattering: A tutorial. *Physics Plasmas* 10: 2155, with permission from the American Institute of Physics.)

of such all-optical-laser-based EUV sources have ten-thousand times shorter wavelength (micron-scale). Thus, the total length of the wiggler region is correspondingly smaller (only mm in length). Another consequence of this is that the frequency upshift required to reach a given output wavelength is also ten thousand times smaller. Also, given that the required electron energy scales as the square root of the upshift, the required electron energy can be one-hundred-times lower (10–100 MeV). It follows from this, and the fact that the field gradients of these accelerators can be ten thousand times higher (1 GeV/cm) than convention RF-based accelerators, that the size of the accelerating region can, in principle, be a million times smaller (only mm in length). Besides its small size, this EUV source can produce femtosecond duration pulse and be synchronized with a relatively low jitter with another femtosecond light pulse having a different wavelength (by virtue of the possibility of deriving the two pulses from the same laser pulse); this is advantageous for the study of ultrafast pump-and-probe photo-initiated processes. The exceptionally low transverse emittance of laser-accelerated electron beams may even make it possible to generate coherent XUV radiation by means of the self-amplified spontaneous emission (SASE) free-electron lasing (FEL) mechanism.

Collective Plasma Response

The above equations apply to electron motion in vacuum. In the case of electrons in plasmas, the collective plasma electron response needs to be considered. A theory for the 1D nonlinear interaction of intense laser fields with electrons has been developed.

The variables are normalized such that

$$\mathbf{a} = e\mathbf{A}/mc^2, \qquad \phi = e\Phi/mc^2 \qquad [73]$$

$$\mathbf{u} = \mathbf{p}/mc, \qquad \beta = \mathbf{v}/c, \qquad \mathbf{u} = \gamma\beta \qquad [74]$$

$$\gamma = (1 + u^2)^{1/2} = (1 - \beta^2)^{-1/2} \qquad [75]$$

The relativistic force equation (Lorentz equation) is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u} = \nabla\phi + \frac{\partial}{\partial t}\mathbf{a} - \bar{\beta}\times(\nabla\times a) \qquad [76]$$

The energy equation ($\mathbf{u}\cdot$force equation) is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\gamma = \beta\cdot\left(\frac{\partial}{\partial z}\phi + \frac{\partial}{\partial t}\mathbf{a}\right) \qquad [77]$$

The transverse force equation (1D) is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}_\perp = \left(\frac{\partial}{\partial t} + \beta_z \frac{\partial}{\partial z}\right)\mathbf{a}_\perp \qquad [78]$$

Since $\mathbf{a} = \mathbf{a}(z, t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{a}_\perp = \left(\frac{\partial}{\partial t} + \beta_z \frac{\partial}{\partial z}\right)\mathbf{a}_\perp \qquad [79]$$

hence:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{u}_\perp - \mathbf{a}_\perp) = 0 \qquad [80]$$

This is just conservation of transverse canonical momentum. Assuming $\mathbf{u}_\perp = 0$ prior to the laser interaction gives

$$\mathbf{u}_\perp = \mathbf{a}_\perp \qquad [81]$$

which is the same as eqn [24].

The electron response to the normalized scaler and vector potentials of the form $\phi = \phi(z - ct)$ and $\mathbf{a} = \mathbf{a}(z - ct)$, which are a function of only $\xi = z - ct$, is completely described by the following constants of the motion:

$$\beta_\perp - a_\perp/\gamma = 0, \qquad [82]$$

$$\gamma(1 - \beta_z) - \phi = 1, \qquad [83]$$

$$n(1 - \beta_z) = n_0, \qquad [84]$$

where $\gamma = (1 - \beta^2)^{-1/2}$ is the relativistic factor and n is the electron density. This allows the various electron quantities to be specified solely in terms of the potentials:

$$\beta_\perp = \frac{2(1+\phi)a_\perp}{(1 + a_\perp^2) + (1+\phi)^2}, \qquad [85]$$

$$\beta_z = \frac{(1 + a_\perp^2) - (1+\phi)^2}{(1 + a_\perp^2) + (1+\phi)^2}, \qquad [86]$$

$$\gamma = \frac{(1 + a_\perp^2) + (1+\phi)^2}{2(1+\phi)}, \qquad [87]$$

$$\frac{n}{n_0} = \frac{(1 + a_\perp^2) + (1+\phi)^2}{2(1+\phi)^2} \qquad [88]$$

In the single particle limit, $\phi \equiv 0$, eqns [26] and [30] are retrieved. For a long pulse interacting with a dense plasma, $\tau_\mathrm{L} \gg \omega_\mathrm{p}^{-1}$ (i.e., neglecting wakefield effects) it can be shown that $(1 + \phi) \simeq (1 + a_0^2)^{1/2}$ for circular polarization and $(1 + \phi) \simeq (1 + a_0^2/2)^{1/2}$ for linear polarization.

Propagation

To understand the propagation of high-intensity light in plasma, we need to understand how the dielectric properties of a plasma medium are affected by the relativistic electron mass change. A wave equation is found by taking the curl of Faraday's law:

$$\nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} \qquad [89]$$

and using Ampere's law:

$$\nabla \times \mathbf{B} = 4\pi \mathbf{J}/c + \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} \qquad [90]$$

gives

$$\begin{aligned}\nabla \times \nabla \times \mathbf{E} &= -\nabla^2 \mathbf{E} + \nabla(\nabla \cdot \mathbf{E}) \\ &= \frac{1}{c^2}\left[4\pi \frac{\partial \mathbf{J}}{\partial t} - \frac{\partial^2 \mathbf{E}}{\partial t^2}\right]\end{aligned} \qquad [91]$$

where $\mathbf{J} = \sum n_i q_i v_i$ and the summation over the index i is done over all charged species of electrons and ions. Using Coulomb's law:

$$\nabla \cdot \mathbf{E} = 4\pi\rho \qquad [92]$$

where $\rho = \sum n_i q_i$, and assuming a uniform plasma, the second term on the left vanishes. Substituting eqn [13] for the velocity, we find

$$\left(-\frac{\partial^2}{\partial t^2} + c^2\nabla^2 - \omega_\mathrm{p}^2\right)\mathbf{E} = 0 \qquad [93]$$

where

$$\omega_\mathrm{p} = (4\pi n_i q_i^2/m_i)^{1/2} \qquad [94]$$

is the plasma frequency. Assuming again plane waves and Fourier analyzing, yields the well-known dispersion relation for electromagnetic waves in plasma:

$$\omega^2 = \omega_\mathrm{p}^2 + c^2k^2 \qquad [95]$$

The index of refraction can be written fully relativistically as

$$\eta^2 = \frac{c^2k^2}{\omega^2} = \frac{c^2}{v_\phi^2} = 1 - \frac{\omega_\mathrm{p}^2}{\omega^2} \qquad [96]$$

where v_ϕ is the phase velocity of the light wave. Assuming infinitely massive ions, the plasma frequency can be written as

$$\omega_\mathrm{p} = \frac{\omega_{\mathrm{p}0}}{\gamma^{1/2}} = \left(\frac{4\pi n_\mathrm{e} e^2}{\gamma m_0}\right)^{1/2} \qquad [97]$$

where ω_{p0} is the plasma frequency in a quiescent plasma, e is the electron charge, m_0 is the electron rest mass, and n_e is the plasma electron density. A change in mass changes the plasma frequency, which in turn modifies the index of refraction and the velocity of the light wave. The light's phase velocity then depends on the laser intensity. This can be seen clearly if we expand the phase velocity for small field strength ($a_0 \ll 1$):

$$v_\phi = \frac{c}{\eta} \sim c\left[1 + \frac{\omega_{p0}^2}{2\omega^2}\left(1 - \frac{\langle a_0^2 \rangle}{2}\right)\right] \qquad [98]$$

An on-axis minimum of the phase velocity [i.e., $v_\phi(r) > v_\phi(0)$] can be created by a laser beam with an intensity profile peaked on axis, such as with a Gaussian beam, causing the wavefronts to curve inward and the laser beam to focus, as shown in Figure 10. When this focusing effect just balances the defocusing caused by diffraction, the laser pulse can propagate over a longer distance than it could in vacuum, while maintaining a small cross-section. This mechanism is referred to as relativistic self-guiding.

Relativistic Self-Focusing

The threshold for relativistic self-focusing can be obtained using the fully relativistic formalism with the vector potential, as in our previous discussion of electron motion in vacuum. In this case, we can rewrite eqn [91] as

$$\nabla^2 \mathbf{a} - \frac{1}{c^2}\frac{\partial^2 \mathbf{a}}{\partial t^2} = \frac{\omega_p^2}{c^2}\frac{n_e \mathbf{a}}{n_0 \gamma} \qquad [99]$$

where $n_0 = n_i$ is the uniform background density. For a Gaussian beam focused in vacuum, where

$$\mathbf{a}(\mathbf{r}, t) = \mathrm{Re}\left\{\mathbf{a}_0(\mathbf{r}, t)\exp[i(\mathbf{k}\cdot\mathbf{r} - \omega t)]\right\} \qquad [100]$$

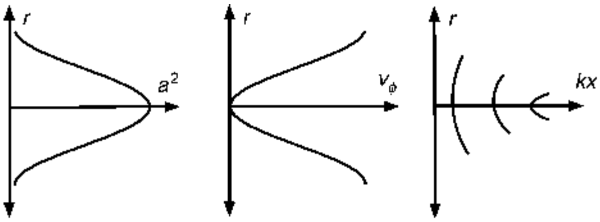

Figure 10 Relativistic self-focusing occurs when an on-axis peak in light intensity (left) produces an on-axis dip in the phase velocity (middle), which acts like a positive lens to cause the phase fronts to curve inward (right). (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

and the amplitude $\mathbf{a}(\mathbf{r}, t)$ varies much less less in r and t than does the phase:

$$|\partial a_0/\partial t| \ll |\omega a_0|, \qquad |\partial a_0/\partial z| \ll |k a_0| \qquad [101]$$

then

$$a_0(r,z) = \frac{e^{-r^2/(r_0^2(1+z^2/L_r^2))}}{\sqrt{1+z^2/L_R^2}}\exp\left\{-\mathrm{i}\arctan\left(\frac{z}{L_R}\right) + \mathrm{i}\left(\frac{r}{r_0}\right)^2 \frac{z/L_R}{1+z^2/L_R^2}\right\} \qquad [102]$$

satisfies the envelope equation (assuming second deriviatives are small compared to first derivatives), where L_R is the Rayleigh range and r_0 is the focal radius. In plasma, eqn [99] can then be written in this envelope approximation as

$$\left[\nabla_\perp^2 + 2ik\frac{\partial}{\partial z}\right]a = -\frac{\omega_p^2}{c^2}\frac{|a|^2}{4}a \qquad [103]$$

The power at which the first and last terms of eqn [103] are in balance is

$$P = \frac{\pi R^2 I_0}{2} \qquad [104]$$

An equation for the beam radius R can thus be found:

$$\frac{d^2R(z)}{dz^2} = \frac{4}{k^2 r_0^2}\left[1 - \frac{1}{22}\frac{\omega_p^2}{c^2}|a_0|^2 r_0^2\right] \qquad [105]$$

Self-focusing occurs when the two terms in the brackets in eqn [105] are equal, which gives for the critical power:

$$P_{\mathrm{crit}} \simeq 17.4\cdot(\omega/\omega_p)^2\ \mathrm{GW} \qquad [106]$$

This corresponds to 1 TW for 1 μm light focused into a gas with electron density of $10^{19}/\mathrm{cm}^3$. Note that this is a power threshold, not an intensity threshold, because the tighter the focusing, the greater the diffraction (the first term in eqn [103]). Numerous recent experiments have confirmed this focusing mechanism when $P > P_{\mathrm{crit}}$.

Any spatial variation of the laser intensity will act to push an electron to regions of lower intensity through the so-called ponderomotive force:

$$\mathbf{F} = \nabla\gamma = \nabla\sqrt{1+a^2} = (2\gamma)^{-1}\nabla a^2 \qquad [107]$$

$$\mathbf{F}_{\text{pond}} = -\frac{m_0 c^2}{8\pi} \frac{\nabla a_0^2}{\sqrt{1+\dfrac{a_0^2}{2}}} \cdot \left[\left(\frac{2}{a_0^2} + 1 \right) E(2\pi, \kappa) - \frac{2}{a_0^2} F(2\pi, \kappa) \right] \qquad [108]$$

where

$$\kappa = \frac{a_0^2/2}{1 + a_0^2/2} \qquad [109]$$

and F and E are the elliptic integrals of the first and second kind, respectively. In the low intensity limit, i.e., when $a_0^2 \ll 1$;

$$\mathbf{F}_{\text{pond}} = -\frac{m_0 c^2}{4} \nabla a_0^2 \qquad [110]$$

That is, the laser ponderomotive force is roughly proportional to the gradient of laser intensity.

A Gaussian-shaped laser's ponderomotive force will tend to expel electrons radially from the region of the axis, so called 'electron cavitation.' If the ponderomotive force is high enough for long enough, the charge displacement due to expelled electrons (in either the lateral or longitudinal direction) will eventually cause the ions to move as well through the Coulomb electrostatic force, forming a density channel. Because $n_e(0) < n_e(r)$, and thus $v_\phi(0) < v_\phi(r)$, this enhances the previously-discussed relativistic self-guiding and can itself guide a laser pulse.

For plasmas created by photo-ionization of a gas by a Gaussian laser pulse, the density will be higher on the axis than off the axis. If we instead expand the phase velocity in terms of changes in density:

$$\frac{v_\phi}{c} \sim 1 + \frac{\omega_{p0}^2}{\omega^2} \frac{\Delta n_e}{n_e} \qquad [111]$$

The phase velocity will thus be higher on axis, which will tend to defocus the light and increase the self-guiding threshold. In order to avoid this, gases with low atomic number and thus fewer available electrons, such as H_3, are commonly used as targets.

Raman Scattering, Plasma Wave Excitation and Electron Acceleration

The local phase velocity, described in eqns [98] and [111], can also vary longitudinally if the intensity and/or electron density does. Local variation in the index of refraction can 'accelerate' photons, i.e., shift their frequency, resulting in photon bunching, which in turn bunches the electron density through the ponderomotive force (F), and so on. When the laser pulse duration is longer than an electron plasma period, $\tau \gg \tau_p = 2\pi/\omega_p$, this photon and electron bunching grows exponentially, leading to the stimulated Raman scattering instability. Energy and momentum must be conserved when the electromagnetic wave $(\omega_0, \mathbf{k}_0)$ decays into a plasma wave $(\omega_p, \mathbf{k}_p)$ and another light wave $(\omega_0 - \omega_p, \mathbf{k}_0 - \mathbf{k}_p)$.

From an equivalent viewpoint, the process begins with a small density perturbation, Δn_e, which, when coupled with the quiver motion, eqn [13], drives a current $J = \Delta n_e e v_e$. This current then becomes the source term for the wave equation (eqn [91]), driving the scattered light wave. The ponderomotive force, due to the beating of the incident and scattered light wave, enhances the density perturbation, creating a plasma wave and the process begins anew. In three dimensions, a plasma wave can be driven when transverse self-focusing and stimulated Raman scattering occur together, a process called the self-modulated wakefield instability.

Two conditions must be satisfied for self-modulation to occur in the plasma. First, the laser pulse must be long compared to the plasma wave, $L \gg \lambda_p$. This allows the Raman instability time to grow, and it allows for feedback from the plasma to the laser pulse to occur. Second, the laser must be intense enough for relativistic self-focusing to occur, $P > P_c$, so that the laser can be locally modified by the plasma. Under these conditions, the laser can form a large plasma wave useful for accelerating electrons.

As the long laser pulse enters the plasma, it will begin to drive a small plasma wave due to either forward Raman scattering or the laser wakefield effect from the front of the laser pulse. This small plasma wave will have regions of higher and lower density with both longitudinal and radial dependence. That is, the plasma wave will be three-dimensional in nature with a modulation along the propagation direction of the laser and a decay in the radial direction to the ambient density (see Figure 11). The importance of this lies with how it affects the index of refraction in the plasma. In the regions of the plasma wave where the plasma density is lower, the radial change in the index of refraction is negative, $\partial n(r)/\partial r < 0$. This means that this part of the plasma acts like a positive lens and focuses the laser. Whereas regions of the plasma wave where the density is higher, $\partial n(r)/\partial r > 0$, the opposite occurs and the laser defocuses. This has the effect of breaking up the laser pulse into a series of shorter pulses of length $\lambda_p/2$ which will be separated by the plasma period. The instability occurs because of how the plasma responds to this. Where the laser is more tightly focused, the ponderomotive force will be

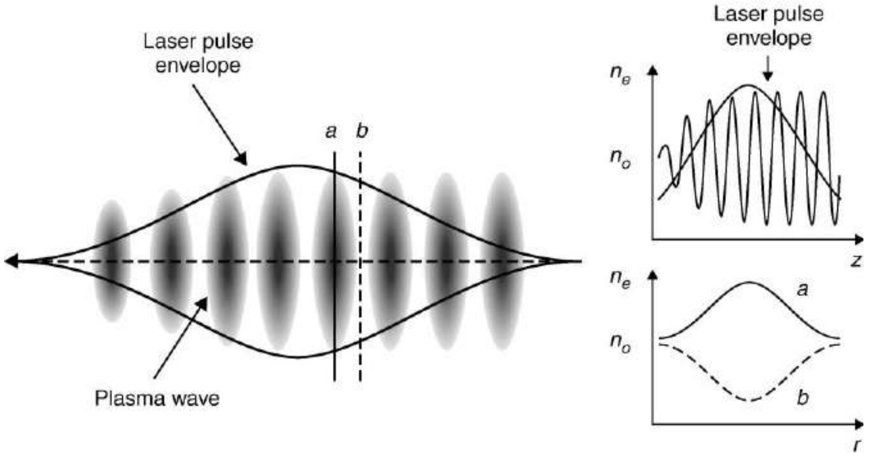

Figure 11 The plasma wave generated by a SMLWFA is three-dimensional in nature. Note that the darker regions correspond to areas of higher plasma density. The graphs to the right represent lineouts of the plasma density longitudinally and radially at the indicated points. (Reproduced with permission from Wagner R (1998) Laser–plasma electron accelerators and nonlinear relativistic optics. PhD thesis, University of Michigan.)

greater and will tend to expel more electrons. This decreases the density in these regions even further, resulting in more focusing of the laser. This feedback rapidly grows, hence the instability.

The phase velocity of the plasma wave in the case of forward scattering is equal to the group velocity of the beat wave, which for low-density plasma is close to the speed of light, as can be seen from the relation:

$$v_\phi = \omega_p/k_p = \Delta\omega/\Delta k = v_g = c\eta \sim c \qquad [112]$$

where eqn [94] and $\omega_p^2 \ll \omega^2$ were used to show that η is close to unity. Such relativistic plasma waves can also be driven by short pulses ($\tau \sim \tau_p$). In this case, the process is referred to as laser–wakefield generation, referring to the analogy with the wake driven by the bow of a boat moving through water, but the mechanism is similar (except it has the advantage that the plasma wave is driven linearly instead of as an instability).

In either case, the resulting electrostatic plasma wave can continuously accelerate relativistic electrons with enormous acceleration gradients. The gradient can be estimated from eqn [92] and the fact that because

$$\nabla \cdot \mathbf{E} \sim E\omega_p/c \propto E\sqrt{n_e} \qquad [113]$$

then

$$E \sim \sqrt{n_e}\ \mathrm{eV/cm} \qquad [114]$$

corresponding to 1 GeV/cm for $n_e = 10^{18}\ \mathrm{cm}^{-3}$. Because this gradient is four orders of magnitude greater than achieved by conventional accelerators (based on fields driven by radio-frequency waves pumped into metal cavities), laser-driven plasma accelerators have received considerable recent attention. They have been shown to accelerate an amount of electron charge (100 nC) comparable to that from conventional accelerators and to have superior transverse geometrical emittance (product of divergence angle and spotsize, similar to the *f*/# in light optics). However, their longitudinal emittance is currently much inferior, energy spreads of 100%. They have been shown to be useful for much of the same applications: radio-isotope production, radiation chemistry, as well as X-ray, proton, and neutron generation. Once the longitudinal emittance can be reduced, they may be advantageous for, among other applications, injectors (especially of short-lived unstable particles) into larger conventional accelerators for high-energy physics research and light sources, and, as discussed in the section on radiation from relativistic electrons about, as stand-alone all-optically driven ultrashort-pulse duration X-ray sources.

The SIMLAC code has been used to study wakefield generation and laser propagation in the limit $a^2 \ll 1$. It draws from nonlinear optics models and treats propagation in the group velocity frame. In this idealized model (which assumes perfect Gaussian beams), the pulse and wake are maintained over long enough propagation distances to accelerate an electron to GeV energy, as shown in Figure 12. A three-dimensional envelope equation for the laser field was derived that includes nonparaxial effects, wakefields, and relativistic nonlinearities.

The resonant wakefield has been characterized experimentally by temporal interferometry, as shown in Figure 13. However, this was done only for the tight-focusing case in which the laser spotsize is much

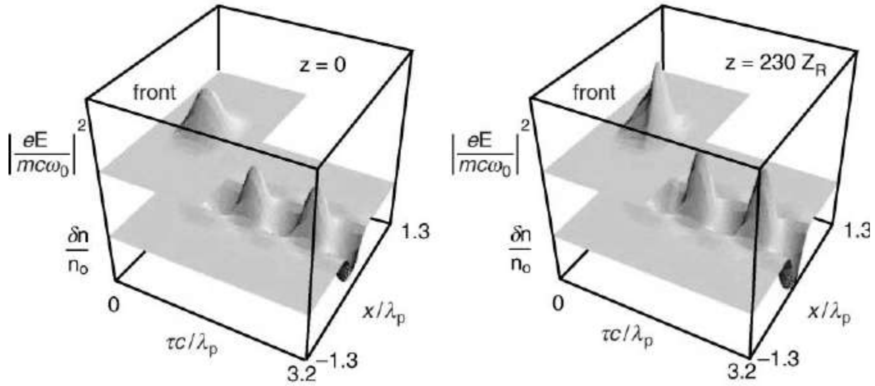

Figure 12 The 'standard' resonant wakefield simulated with SIMLAC, a code that moves at the light pulse's group velocity. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774.)

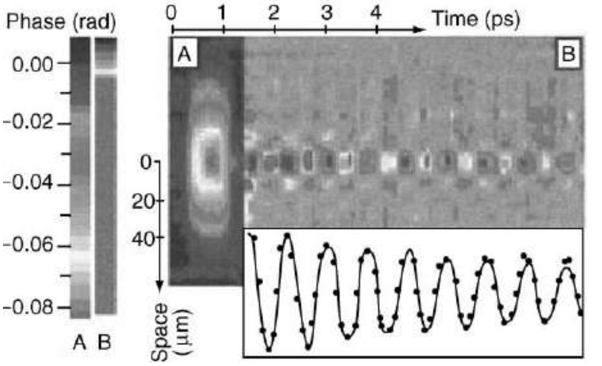

Figure 13 Typical result of a phase shift measurement to study a resonantly excited laser wakefield plasma wave by means of time-domain interferometry. Parts (A) and (B) have different color scales. The bottom graph is a line out of part (B) along the laser axis. (Reproduced with permission from Marquès JR, Dorchies F, Audebert P, Ceindre JP, Amiranoff F, Gauthier JC, Hammoniaux G, Antonetti A, Chessa P, Mora PTM and Antonson J (1997) Frequency increase and damping of nonlinear electron plasma oscillations in cylindrical symmetry. *Physics Review Letters* 78: 3463. Copyright (1995) by the American Physical Society.)

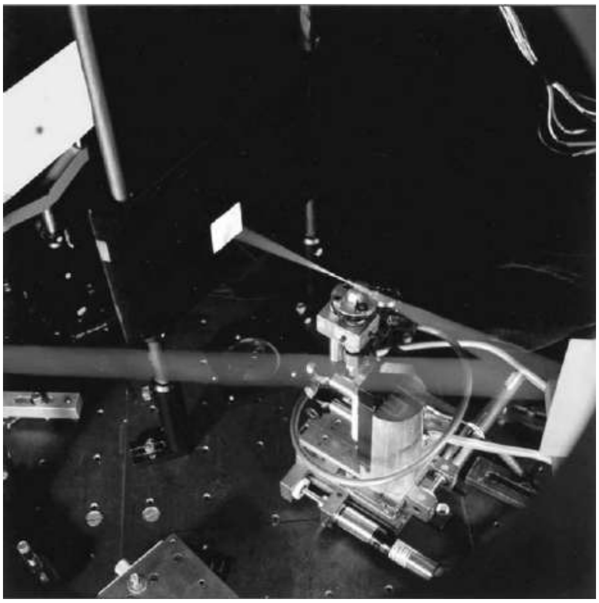

Figure 14 Artistically enhanced photograph of the acceleration of an electron beam by a laser interacting with a gas jet inside a vacuum chamber. The laser crosses the picture from left to right and is focused by a parabolic mirror (right side of the picture). The supersonic nozzle (shown in the middle of the picture) is positioned with micron accuracy with a 3-axis micropositioner. The e-beam makes a small spot on a white flourescent (LANEX) screen, shown in the upper left-hand corner of the picture. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics.)

smaller than the plasma wave wavelength ($r_l \ll \lambda_p$) and thus the transverse wakefield was much greater than the longitudinal wakefield.

A typical experimental setup, used to study electron acceleration, is shown in Figure 14.

Dramatic reduction of the angular divergence of a laser accelerated electron beam was observed with increasing laser power above the relativistic self-focusing threshold, as shown in Figure 15.

Interactions with Solid-Density Targets

The generation of electrons by high-intensity laser light interacting with solid targets can generate energetic X-rays, accelerate other types of particles and induce nuclear reactions, as illustrated in Figure 16. For instance, high-order harmonics have been generated by the oscillation of the critical density surface, in the so-called moving mirror model. Bright X-rays, originating from Bremstrahlung caused by electron collisions with high-*Z* atoms in solid targets, have created isotopes by means of photofission. Laser-accelerated electron energies and angular distributions have been inferred from analyzing (γ, n) and $(\gamma, 2n)$ reactions in composite Pb/Cu targets and in Ta/Cu targets. Positrons were created by colliding laser-accelerated electrons with a tungsten target.

When electrons are heated to high temperatures or accelerated to high energies, they can separate from plasma ions. Such charge displacement creates an electrostatic sheath, which eventually accelerates the ions. The ions are pulled by the charge of the electrons and pushed by the other ions' unshielded charges (similar to the 'Coulomb explosion' that can occur during the ionization of atoms). When the charge displacement is driven by thermal expansion, as in long-pulse (low-power) laser–plasma experiments, the maximum ion energies are limited to less than 100 keV. However, when the charge displacement is driven by direct laser heating, as in short-pulse high-power laser–plasma experiments, multi-MeV ion energies are possible. This was first shown with gas jet targets, in which case the ions were accelerated radially into 2π, and then later with thin solid-density-films, in which case the ions were accelerated into collimated beams. In the latter case, hydrocarbons and water on the surface of the film can become ionized and provide a source of protons to be accelerated.

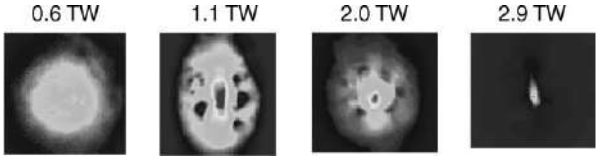

Figure 15 Images of the spatial profiles of the electron beam measured by a ccd camera imaging a LANEX screen at a distance of 15 cm from the gas jet for various laser powers. The divergence angle of the beam decreases to a value of $\Delta\theta = 1°$ at a power of 2.9 TW, corresponding to a transverse geomtrical emittance of just $\varepsilon_{\perp} \lesssim 0.06\pi$ mm-mrad. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics.)

An intense laser can ponderomotively heat electrons. If the laser contrast is high, vacuum heating can occur in the following manner. When light encounters a sharp interface between vacuum and solid density, the electromagnetic field becomes evanescent in the region above the critical density. The instantaneous '$v \times B$' force can push electrons in the direction of the light's propagation vector; it also has a frequency twice that of the pump and a magnitude proportional to the square of the normalized vector potential, a_0^3. Thus, electrons can only complete half of their figure-eight orbits, on the vacuum side, gaining relativistic energies; they move through the overdense region without the electromagnetic field to pull them back. An electrostatic sheath can thus form, which will accelerate the ions left behind. Another important heating mechanism is stochastic heating, which occurs when the light that is reflected from the critical surface beats with the incoming wave to create a standing wave. The motion of electrons in such a wave can become chaotic, resulting in a large increase in electron temperature (>100 keV).

As the heated electrons propagate through a solid, they can instantaneously field-ionize the neutral atoms of the solid. This will both modify the solid's

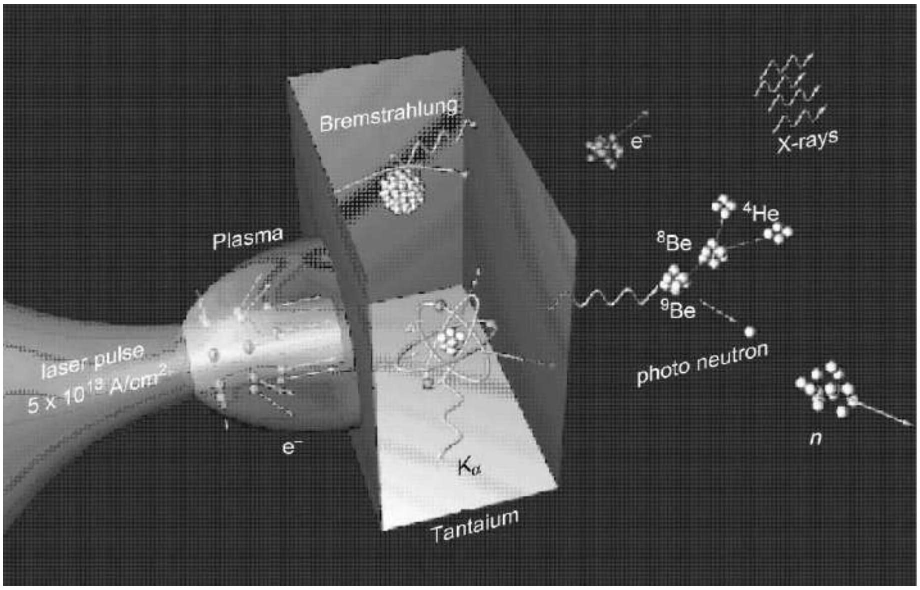

Figure 16 Illustration of the various mechanisms of particle acceleration and X-ray generation in high-intensity laser interactions with solid-density targets. (Courtesy of R. Sauerbrey.)

conductivity and provide a source of protons on the rear-side of the target. If the film is thin enough, the electrons can pass through, and create a sheath on the rear-side of the target. This latter mechanism has been dubbed the target normal sheath acceleration mechanism. The ions from thin foils have been claimed to originate from both the front and rear-side of the foil.

Several groups have reported the observation of ions originating from thin-film solid-density targets. Unlike previous long-pulse experiments, the ions were accelerated along the direction normal to the side of the target, that is opposite to that upon which the laser was incident. The ions generally originate from water or hydrocarbons on the surface of the material. The acceleration results from several different mechanisms, which may be occurring simultaneously. Charge-displacement is again common to all, with the electrons being heated ponderomotively, such as by Brunel, $J \times B$ or stochastic heating. In one case, the electrostatic sheath is formed at the backside of the ionization layer formed on the side of the target upon which the laser is incident (front side). In another case, the electrostatic sheath is formed by field ionization of the ion layer on the opposite side of the thin film target (back-side), the target normal sheath acceleration (TNSA) hypothesis. Numerical simulations show evidence for both front- and back-side acceleration.

The results of these experiments indicate that a large number of protons ($10^{13}\,p$) can be accelerated, corresponding to source current densities (10^8 A/cm^2) that are nine orders of magnitude higher than produced by cyclotrons, but with comparable, or even lower, transverse emittances ($\varepsilon_\perp \leq 1.0\pi$ mm-mrad). Proton energies up to 60 MeV have been observed in experiments at intensities exceeding 10^{20} W/cm^2 (using a petawatt power laser). The high end of the proton spectrum typically has a sharp cut-off, but is a continuum. In one experiment, protons were observed to be emitted in ring patterns, the radii of which depend on the proton energy, which was explained by self-generated magnetic fields.

A typical experimental setup used to observe the acceleration of ions is shown in Figure 17. The production of radionuclides have been used as an ion energy diagnostic. In another example of a nuclear reaction initiated by an intense laser, neutrons

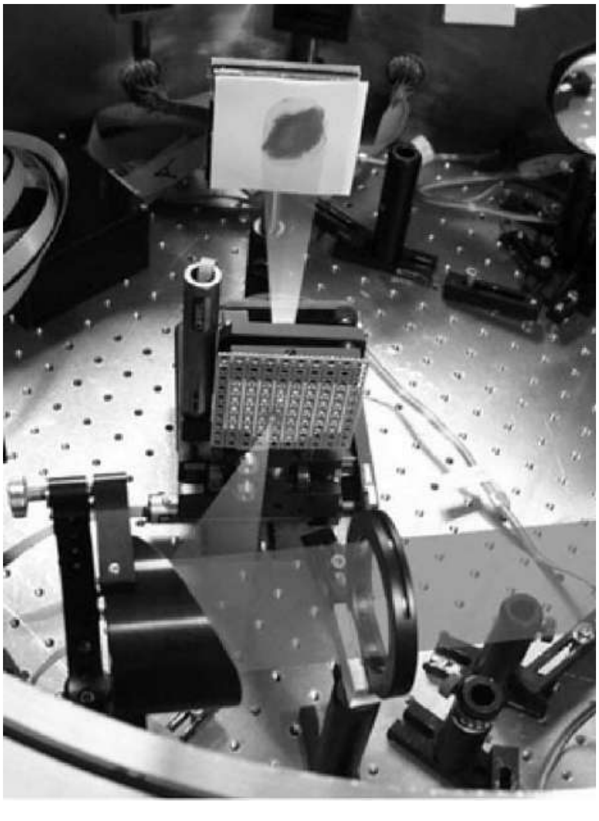

Figure 17 Artistically enhanced photograph of the typical setup used to observe the acceleration of ions. The laser (shown in the foreground) is focused with an off-axis parabola onto a thin-foil, held by a mesh that is positioned by a 3-axis micropositioner. CR-39, a nuclear track detector (shown in the background) is used to detect the ions. An actual proton-produced pattern is shown in red. (Reproduced from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics).

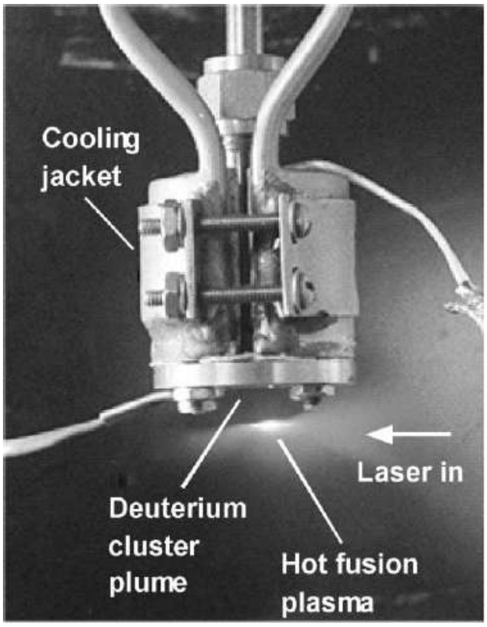

Figure 18 Photograph of a laser interacting with a deuterium cluster jet emanating from cooled gas nozzle. Fusion neutrons are produced with relatively low laser intensities. (Reproduced with permission from Ditmire T (2002) Laser fusion on a tabletop. *Optics and Photonics News* 13: 28, with permission from IEEE.)

have been produced by the He fusion reaction in the focus of 200 mJ, 160 fs Ti:sapphire laser pulses on a deuterated polyethylene target. Optimizing the fast electron and ion generation by applying a well-defined prepulse led to an average rate of 140 neutrons per shot. Neutrons have also been generated from cluster plasmas, which were produced by the cooled-nozzle depicted in Figure 18, but with significantly lower laser intensities than required with planar solid targets.

Concluding Remarks

The field of nonlinear optics, with electrons bound to atoms, has over the last forty years given rise to many scientific discoveries and technologies that are now commonplace. The relatively young field of relativistic nonlinear optics has already begun to do the same. Some of the exciting phenomena and applications that have already been identified have been discussed in this article. Others await discovery as ever higher laser intensities are reached, with no foreseeable limit.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Scattering:** Raman Scattering.

Further Reading

Chen S-Y, Maksimchuk A and Umstadter D (1998) Experimental observation of relativistic nonlinear Thomson scattering. *Nature* 396: 653.

Cowan TE, Hunt AW, Phillips TW, Wilks SC, Perry MD, Brown C, Fountain W, Hatchett S, Johnson J, Key MH, Parnell T, Pennington DM, Snavely RA and Takahashi Y (2000) Photonuclear fission from high energy electrons from ultraintense laser–solid interactions. *Physics Review Letters* 84: 903.

Ditmire T (2002) Laser fusion on a tabletop. *Optics and Photonics News* 13: 28.

Eberly JH (1986) Fundamentals of laser interactions, vol. 229 of lecture notes in physics. *Journal Optical Society of America B* 3: 1324.

Esarey E, Sprangle P, Krall J and Ting A (1996) Overview of plasma-based accelerator concepts. *IEEE Transactions on Plasma Science* 24: 252.

Gibbon P and Förster E (1996) Short-pulse laser–plasma interactions. *Plasma Phys. Controlled Fusion* 38: 769.

Hartemann F (2001) *High-Field Electrodynamics*. Boca Raton: CRC Press.

Joshi CJ and Corkum PB (1995) Interactions of ultra-intense laser light with matter. *Physics Today* 48: 36.

Lau Y, He F and Umstadter D (2003) Nonlinear Thomson scattering: A tutorial. *Physics Plasmas* 10: 2155.

Ledingham KWD, Spencer I, McCanny T, Singhal RP, Santala MI, Clark E, Watts I, Beg FN, Zepf M, Krushelnick K, Tatarakis M, Dangor AE, Norreys PA, Allott R, Neely D, Clark RJ, Machacock AC, Wark JS, Cresswell AJ, Sanderson DCW and Magill J (2000) Photonuclear physics when a multiterawatt laser pulse interacts with solid targets. *Physics Review Letters* 84: 899.

Luther-Davies B, Gamaly EG, Wang Y, Rode AV and Tikhonchuk V (1992) Matter in ultrastrong laser fields. *Soviet Journal Quantum Electronics* 22: 289.

Marquès JR, Dorchies F, Audebert P, Ceindre JP, Amiranoff F, Gauthier JC, Hammoniaux G, Antonetti A, Chessa P, Mora PTM and Antonson J (1997) Frequency increase and damping of nonlinear electron plasma oscillations in cylindrical symmetry. *Physics Review Letters* 78: 3463.

Meyer-ter-Vehn J, Pukhov A and Sheng Z-M (2001) Relativistic laser plasma interaction. In: Batani D (ed.) *Atoms, Solids and Plasmas in Super-Intense Laser Fields*, pp. 167. New York: Kluwer Academic–Plenum.

Moore CI, Knauer JP and Meyerhofer DD (1995) Observation of the transition from Thomson to Compton scattering in multiphoton interactions with low-energy electrons. *Physics Review Letters* 74: 2439.

Mourou G and Umstadter D (1992) Development and applications of compact high-intensity lasers. *Physics Fluids B* 4: 2315.

Mourou GA and Umstadter D (2002) Extreme light. *Scientific America* 81.

Mourou GA, Barty CPJ and Perry MD (1998) Ultrahigh-intensity lasers: Physics of the extreme on a tabletop. *Physics Today* 22.

Siders CW, Blanc SPL, Fisher D, Tajima T, Downer MC, Babine A, Stepanov A and Sergeev A (1996) Laser wakefield excitation and measurement by femtosecond longitudinal interferometry. *Physics Review Letters* 76: 3570.

Sprangle P, Peñano JR, Hafizi B, Hubbard RF, Ting A, Gordon DF, Zigler A and Antonsen TM (2002) GeV acceleration in tapered plasma channels. *Physics Plasmas* 9: 2364.

Tajima T and Mourou G (2002) Zettawatt-exawatt lasers and their applications in ultrastrong-field physics. *Physics Review STAB* 5: 031301.

Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774.

Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165.

Umstadter D and Norris TB (eds) (1997) Nonlinear optics with relativistic electrons. *IEEE Journal Quantum Electronics* 33: 1878.

Wagner R (1998) Laser–plasma electron accelerators and nonlinear relativistic optics. PhD thesis, University of Michigan.

Wagner R, Chen S-Y, Maksimchuk A and Umstadter D (1997) Electron acceleration by a laser wakefield in a relativistically self-guided channel. *Physics Review Letters* 78: 3125.

Yamakawa K (2004) Table-top lasers create ultrahigh peak powers. *OYOBUTURI* 73: 186–193.

S

SCATTERING

Contents

Raman Scattering
Scattering from Surfaces and Thin Films
Scattering Phenomena in Optical Fibers
Scattering Theory
Stimulated Scattering

Raman Scattering

F Kannari, Keio University, Yokohama, Japan

© 2005, Elsevier Ltd. All Rights Reserved.

When an electromagnetic wave passes over a small elastically bound charged particle, the particle will be set into motion by the electric field. Scattering takes place for frequencies not corresponding to the natural frequencies of the particles. If the particle is bound by a force obeying Hooke's law, this vibration will have the same frequency and spatial direction as that of the electric force in the electromagnetic wave. The phase, on the other hand, of the forced vibration will differ from that of the incident wave, which causes the basis of dispersion. If the charged oscillation is bound by a force which does not obey Hooke's law, but some more complicated law, it will reradiate not only the impressed frequency, but also various combinations of this frequency with the fundamental and overtone frequencies of the oscillator.

The Raman effect was discovered by C.V. Raman (1888–1971). Ordinary Raman spectroscopy is used as a tool for studying the vibrational energy levels of molecules and of lattice optical branch vibration in crystals. A spectroscopic measurement of the scattered light reveals the existence of frequencies shifted down by increments corresponding to vibrational frequencies of the material. This down-shifted scattering is referred to as Stokes scattering. Frequencies shifted up by the vibrational frequencies are also present in the scattered radiation, which are the so-called anti-Stokes scattering.

Stimulated amplification of scattered Stokes radiation occurs through a third-order nonlinear process. To introduce stimulated Raman scattering (SRS), first we describe the steady-state stimulated scattering theory.

In general, nonlinear optical processes are described by the classical Maxwell equation with the nonlinear polarization of a material $\mathbf{P}^{\mathrm{NL}}$:

$$\nabla \times \nabla \times \mathbf{E} + \varepsilon\mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\mu_0 \frac{\partial^2 \mathbf{P}^{\mathrm{NL}}}{\partial t^2} \qquad [1]$$

where ε and μ_0 are the linear dielectric constant of the material and the magnetic permeability of the vacuum. We consider a relatively intense excitation radiation $\mathbf{E}_{\mathrm{L}}$ (frequency ω_{L}) and a weak signal radiation $\mathbf{E}_{\mathrm{S}}$, having the carrier frequency ω_{s} corresponding to the scattered light frequency ($\omega_{\mathrm{s}} = \omega_{\mathrm{L}} - n\omega_0$);

$$\mathbf{E}_{\mathrm{L}} = \frac{1}{2}\{\bar{\mathbf{E}}_{\mathrm{L}}(\mathbf{r}, t) \exp \mathrm{i}(\mathbf{k}_{\mathrm{L}} \times \mathbf{r} - \omega_{\mathrm{L}} t) + \mathrm{c.c.}\} \qquad [2]$$

$$\mathbf{E}_{\mathrm{S}} = \frac{1}{2}\{\bar{\mathbf{E}}_{\mathrm{S}}(\mathbf{r}, t) \exp \mathrm{i}(\mathbf{k}_{\mathrm{S}} \times \mathbf{r} - \omega_{\mathrm{S}} t) + \mathrm{c.c.}\} \qquad [3]$$

where $\bar{\mathbf{E}}_L(\mathbf{r},t), \bar{\mathbf{E}}_S(\mathbf{r},t)$ are a slowly varying envelope of the complex electric fields. Assuming that the variation of the electric field during the time-scale shorter than the decay time of the induced oscillation in the material is negligible, nonlinear polarization induced by the excitation radiation is described as follows:

$$\mathbf{P}_S^{NL} = \frac{1}{2}\{\bar{\mathbf{P}}_S^{NL}(\mathbf{r},t)\exp i(\mathbf{k}_S \times \mathbf{r} - \omega_S t) + \text{c.c.}\} \qquad [4]$$

$$\bar{\mathbf{P}}_S^{NL} = \varepsilon_0 \chi^{NL}(-\omega_S, -\omega_L, \omega_L, \omega_S)\bar{\mathbf{E}}_L^* \bar{\mathbf{E}}_L \bar{\mathbf{E}}_S \qquad [5]$$

where ε_0 is the vacuum dielectric constant and $\chi^{NL}(-\omega_S, -\omega_L, \omega_L, \omega_S)$ is the third-order nonlinear susceptibility, of which the real part corresponds to the variation of the dielectric constant depending on the optical radiation, whereas the imaginary part describes the stimulated scattering. In steady-state approximation, substituting eqns [3]–[5] into [1], we obtain

$$\frac{d\bar{\mathbf{E}}_S}{dz} = \frac{1}{2}\left(-a + i\frac{\omega_S}{cn_S}\chi^{NL}|\bar{\mathbf{E}}_L|^2\right)\bar{\mathbf{E}}_S \qquad [6]$$

where we ignore the higher-order differential of amplitude. n_S is the refractive index at ω_S, and a is the linear absorption coefficient of the signal radiation. When ignoring the decay of the excitation radiation along the z direction in the material, for making the right-hand term of eqn [6] positive, the next condition must be satisfied:

$$I_L = \frac{1}{2}c\varepsilon_0 n_L|\bar{\mathbf{E}}_L|^2 \geq I_L^{th} = \frac{c^2\varepsilon_0 n_S n_L}{2\omega_S}\frac{a}{-\text{Im}\chi^{NL}} \qquad [7]$$

The signal radiation is then amplified along the z direction at an amplification constant of g,

$$G = g(I_L - I_L^{th}) \qquad [8]$$

$$g = \frac{\omega_S}{c^2\varepsilon_0 n_S n_L}(-\text{Im}\,\chi^{NL}) \qquad [9]$$

Stimulated Raman Scattering

In SRS, the frequency difference $\omega_L - \omega_S$ corresponds to the resonance frequency of vibrational oscillation ω_V. Therefore, the vibrational oscillation mode is excited in the material as well as the generation of scattered radiation. In the following section, we consider the SRS by optical phonon. In general, nonlinear polarization under the optical excitation is described as follows:

$$\mathbf{P}^{NL} = N_0\alpha' q\mathbf{E}, \qquad [10]$$

$$\ddot{q} = \Gamma_V\dot{q} + \omega_V^2 q = \frac{1}{2}\alpha'\mathbf{E}^2 \qquad [11]$$

where q is the basic coordinate of molecular vibration, Γ_V is the decay rate of the vibrational oscillation, $\alpha' = (\partial\alpha/\partial q)_0$ is the Raman polarizability of a molecule, and N_0 is the molecule density. We can obtain the steady-state nonlinear polarization from eqns [4], [5], [10], and [11] as

$$\chi^{NL} = (-\omega_S, -\omega_L, \omega_L, \omega_S) = \frac{N_0(\alpha')^2}{4\varepsilon_0}\frac{1}{\omega_V^2 - \omega^2 + i\omega\Gamma_V} \qquad [12]$$

where $\omega = \omega_L - \omega_S$. Thus, the gain coefficient can be obtained from eqns [9] and [12] under the assumption of $\omega_V \gg \Gamma_V$, as follows:

$$g_R(\omega_S) = \pm\frac{N_0(\alpha')^2\omega_S}{2c^2\varepsilon_0 n_S n_L \omega_V \Gamma_V} \times \frac{(\Gamma_V/2)^2}{(\omega_L - \omega_S \mp \omega_V)^2 + (\Gamma_V/2)^2} \qquad [13]$$

In general, Raman scattering is characterized by the scattering cross-section $d\sigma/d\Omega$, which is defined for one of the optical polarization directions in terms of the optical power density:

$$g_R(\omega_L - \omega_V) = \frac{16\pi^2 c^2 N_0}{\hbar n_S^2 \omega_S^3 \Gamma_V}\frac{d\sigma}{d\Omega}\left\{1 - \exp\left(-\frac{\hbar\omega_V}{k_B T}\right)\right\} \qquad [14]$$

The value of $(d\sigma/d\Omega)\Gamma_V^{-1}$ can be characterized as a peak intensity in the measured spontaneous Raman scattering spectrum.

When the conversion to the first Stokes wave is large, we must consider both pump depletion and conversion to higher-order Stokes waves (Figure 1a). Pump depletion in first Stokes generation can be accounted for by solving the following coupling

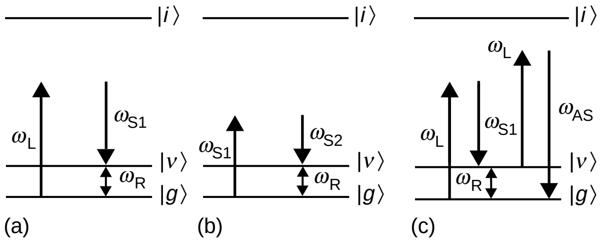

Figure 1 Energy diagrams of stimulated Raman scattering: (a) first Stokes scattering; (b) cascaded second Stokes scattering; and (c) first anti-Stokes scattering.

equations between pump power and Stokes power:

$$\frac{\partial I_L}{\partial z} + \frac{n_L}{c}\frac{\partial I_L}{\partial t} = -g_L I_L I_S \qquad [15]$$

$$\frac{\partial I_S}{\partial z} + \frac{n_S}{c}\frac{\partial I_S}{\partial t} = g_R I_L I_S \qquad [16]$$

where

$$g_L = \frac{\omega_L}{\omega_S} g_R$$

In SRS the power generated in the first Stokes serves as a pump for a second Stokes wave (Figure 1b). Cascade shifts of this type, in which the Stokes wave of order n serves as a pump for the generation of the Stokes wave order $n+1$, can result in the generation of radiation at several frequencies, each one shifted from the original pump frequency by a multiple of the vibrational mode frequency. In addition to the cascade higher-order Stokes generation, multiple Stokes orders can also be generated by four-wave mixing interactions (Figure 2), of the form:

$$\omega_{S,n+1} = 2\omega_{S,n} - \omega_{S,n-1} \qquad [17]$$

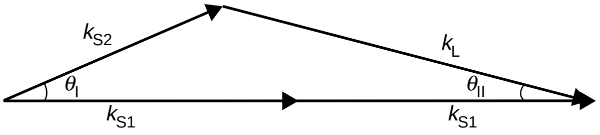

Figure 2 Vector diagram of multiple Stokes generation through four-wave mixing in positively dispersive media.

These interactions are important when the wavevector mismatch among the various waves is small. In most media exhibiting positive dispersion, the phase matching of the various interactions involves off-axis components of the generated waves. Thus, these four-wave mixing interactions are stronger for tightly focused beams than for collimated pump beams.

SRS can be realized in a variety of configurations, including single pass Raman generation, or using a variety of extra-cavity or intra-cavity Raman resonator. Variation of the Raman cross-section for the wavelength is shown in Figure 3 for various gaseous Raman media. The Raman gain coefficient simply increases with decreasing wavelength, as indicted by eqn [13]. Therefore, higher conversion efficiencies are obtainable with UV lasers such as rare-gas halide excimer lasers in a single pass configuration. Note that in some gases, such as H_2, the Raman gain becomes independent of pressure at higher pressures since the increase in linewidth, due to pressure broadening, cancels the increase in gain originated from number density. Typical SRS gains for some gas, liquid, and solid media are listed in Table 1.

Although SRS in high-pressure gases has been widely employed for frequency conversion, SRS in crystals is currently a growing area of research activity. Table 2 lists SRS frequency shifts of various crystals. Since a very high pump laser power ~1 GW/cm^2, which is already close to the laser damage threshold of many crystals, solid-state Raman lasers are

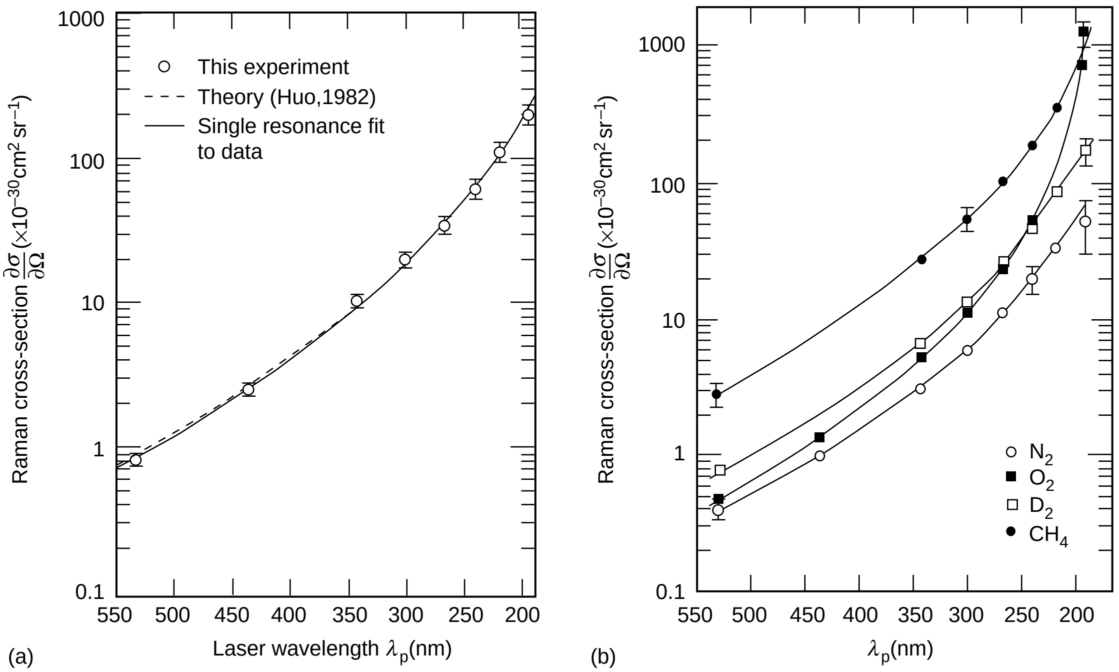

Figure 3 Variation of the Raman cross-section with pump wavelength for various gases. Source: AIP Conference Proceedings No. 100 (Excimer Laser – 1983), 181.

Table 1 Typical Raman media and their parameters

Materials	*Stokes shift (cm^{-1})*	*Raman linewidth (cm^{-1})*	$N\frac{d\sigma}{d\Omega}$ *(10^{-6} m^{-1} str^{-1})*	*Raman gain (10^{-5} m/MW)*
H_2 (100 atm)	4155	0.20	3.0×10^3	1.5
H_2O	3420	176	0.51	0.07
Acetone CH_3COCH_3	2921	18	5.4	1.2
Cyclohexane C_6H_{12}	2852	11	3.7	1.2
Liquid N_2	2326.5	0.067	0.29 ± 0.09	17.5 ± 5
Liquid O_2	1552	0.117	0.48 ± 0.14	14.5 ± 4
Nitrobenzene $C_6H_5NO_2$	1345	6.6	6.4	2.1
Toluene C_7H_8	1003	1.94	1.1	1.2
C_6H_5Cl	1002	1.6	1.5	1.9
C_6H5Br	1000	1.9	1.5	1.5
C_6H_6	992	2.15	3.06	2.8
CS_2	655.6	0.5	7.55	23.8
CC_{l4}	459	5	2.3	0.85
Diamond	1332	2.04	17	6.9
Calcite	1086	1.1	2.9	4.4
Si	521	0.8	305	190
Quartz	467	6.7	3.1	0.8
Li^7TaO_3	201	22	238	4.4
	215	12	167	10
Li^7NbO_3	256	23	381	8.9
	258	7	262	28.7
	637	20	231	9.4
	643	16	231	12.6
InSb ($n_e \sim 2 \times 10^{22}$ m^{-3})	100	2	30	16.7

designed with an intra-cavity, coupled cavities, or an extra-cavity. Efficient Raman lasers have been demonstrated using external resonator configuration when the pumping pulse is intense enough. For example, ~40% conversion efficiency to S1 was reported with a Q-switched frequency doubled Nd:YAG laser in a $Ba(NO_3)_2$ crystal using a pump fluence of 1.4 J cm^{-1} and a pulse repetition rate of 30 Hz. The resonator mirror characteristics can be designed to enhance the desired Stokes-order radiation by completely confining the lower-order Stokes radiation inside the cavity. For a lower power pump source, such as cw-pumped Q-switch lasers, a Raman crystal is placed inside a pump laser resonator to reduce the effective threshold for SRS. The resonator configuration has a significant influence on the Stokes output performance. The optimum spot size in the resonator and power density in the Raman crystal is a compromise between maximizing the conversion efficiency and avoiding optical damage. It is noteworthy that the thermal lens in the Raman crystal is caused by the energy dissipation of the phonon and thus depends on the power density of the Stokes field.

A fiber Raman laser is an optical fiber with pump light focused into one end and SRS light appearing at the output. By injecting a weak frequency-shifted signal light together with the pump light, the signal light experienced substantial amplification along the length of the fiber. This fiber Raman amplifier has been actively developed to extend the wavelength band for optical fiber communication. The fiber Raman lasers have utilized fused silica (SiO_2), germanosilicate, or phosphosilicate optical fibers for which the Raman frequency shifts are ~440–490 cm^{-1}. In contrast to Erbium-doped fiber amplifier (EDFA), since the single mode fiber itself can act as an amplifier, the Raman amplifier can be distributed over optical transmission lines in lengths of >20 km, which can keep the transmission power always nearly constant and consequently reduce the signal noise. Moreover, by preparing multipumping wavelengths, the number of signal wavelengths can increase without the limitation of an amplification gain bandwidth. High-power light sources around 1480 nm developed for EDFA can be utilized by the fiber Raman amplifier. By tailoring the power levels among multiwavelength pump lights, a flat amplification gain spectral width of ~100 nm is obtainable.

Transient Effects

In the steady-state SRS analysis, we assume that the pump pulse duration is much longer than the vibrational mode dephasing time ($\Delta\nu_p \ll \Gamma_V$). In a transient regime, when the pump pulse duration is shorter than the dephasing time ($\Delta\nu_p \gg \Gamma_V$), the effect of the transient response of the medium reduces the gain and delays the Stokes pulse relative to the pump pulse. With no pump depletion, the analytical expression for the Stokes intensity for large

Table 2 Typical solid-state Raman materials and parameters

Materials	*Stokes shift (cm^{-1})*	*Raman linewidth (cm^{-1})*	*$\frac{d\sigma}{d\Omega}$ (arb.u.)*
Diamond	1332.9	2.7	100
SO_2	464.5	7.0	2.2
Nitrates and calcite			
$Ba(NO_3)_2$	1048.6	0.4	21
$NaNO_3$	1069.2	1.0	23
$CaCO_3$	1086.4	1.2	6.0
Tungstates			
$CaWO_4$	910.7	4.8	47
$SrWO_4$	921.5	3	–
$BaWO_4$	926.5	2.2	–
$NaY(WO_4)_2$	918	15	–
$KGd(WO_4)_2$	901	5.4	54
$KGd(WO_4)_2$	901	5.4	43
$KGd(WO_4)_2$	768	6.4	19
$KGd(WO_4)_2$	768	6.4	65
$KY(WO_4)_2$	905.6	7	50
$KY(WO_4)_2$	905.6	7	45
$KY(WO_4)_2$	767.4	8.4	20
$KY(WO_4)_2$	767.4	8.4	64
$KYb(WO_4)_2$	908	7.4	48
$KYb(WO_4)_2$	908	7.4	48
$KYb(WO_4)_2$	757	15	25
$KYb(WO_4)_2$	757	15	70
Molybdates			
$CaMoO_4$	879.3	5.0	64
$SrMoO_4$	887.7	2.8	55
$BaMoO_4$	892.4	2.1	52
Iodate and niobates			
$LiIO_3$	821.6	5.0	54
$LiNO_3$	872	21.4	44
$LiNbO_3$	632	27	166
$LiNbO_3$	250	28	–
$LaNbO_4$	805	9	22
Phosphates			
$Ca_2(PO_4)_3F$	964.7	2.8	3.4
$Sr_5(PO_4)_3F$	950.3	2.8	3.4

amplification gain is given as

$$I_S(\ell) \propto I_S(0)\exp\left(-\frac{\tau_p}{T_V}\right) \times \exp\left\{2\left[I_L\tau_p\ell\frac{16\pi^2c^2N_0}{\hbar n_S^2\omega_S^3}\left(\frac{d\sigma}{d\Omega}\right)\right]^{1/2}\right\} \qquad [18]$$

Here, T_V is the dephasing time ($T_V = \pi\Gamma_V^{-1}$), and τ_p is the pump pulse width. One can see that the Raman gain is proportional to the square root of the multiplication of the pump pulse energy $I_L \times \tau_p$, the crystal length ℓ, and the total integral Raman scattering cross-section ($d\sigma/d\Omega$), but does not depend on the Raman linewidth Γ_V.

Anti-Stokes Raman Scattering

From eqn [13] the gain coefficient is negative at $\omega = \omega_L + \omega_S$. However, anti-Stokes scattering (Figure 1c) is stimulated through a four-wave mixing interaction of the form:

$$\omega_A = 2\omega_L - \omega_S \qquad [19]$$

The equations describing anti-Stokes Raman scattering in the absence of pump depletion are:

$$\frac{d\bar{E}_S}{dz} + \frac{a}{2}\bar{E}_S = i\frac{\omega_S^2}{2c^2\varepsilon_0 k_{Sz}}\bar{P}_S^{NL} \qquad [20]$$

$$\frac{dE_A^*}{dz} + \frac{a}{2}\bar{E}_A^* = -i\frac{\omega_A^2}{2c^2\varepsilon_0 k_{Az}}\bar{P}_A^{NL*} \qquad [21]$$

$$\bar{P}_S^{NL} = \varepsilon_0\chi^{NL}\left\{|\bar{E}_L|^2\bar{E}_S + \bar{E}_L^2\bar{E}_A^*\exp(i\Delta kz)\right\} \qquad [22]$$

$$\bar{P}_A^{NL*} = \varepsilon_0\chi^{NL}\left\{\bar{E}_L^{*2}\bar{E}_S\exp(-i\Delta kz) + |\bar{E}_L|^2\bar{E}_A^*\right\} \qquad [23]$$

Here, Δk is the wavevector mismatch, given by

$$\Delta k = k_S + k_A - 2k_L \qquad [24]$$

Then, power gain per unit length is obtained to give

$$G \cong -a \pm 2\,\mathrm{Im}\left\{\left(\frac{\Delta k}{2}\right)^2 - \frac{\omega_S}{2cn_S}\chi^{NL}|\bar{E}_L|^2\Delta k\right\}^{1/2} \qquad [25]$$

The interaction of the anti-Stokes wave with the other waves is stronger when the wavevector mismatch is small. The exact value of the phase-matching angle depends on the dispersion of the medium. At the tightly focused pumping condition, strong anti-Stokes radiation is generated so that the wavevector mismatch is minimum. Just as in Stokes scattering, multiple anti-Stokes orders can be generated (Figure 4).

So far, Stokes and anti-Stokes Raman scattering have been used to generate radiation over much of the UV and VUV, ranging from 138 to 400 nm. The shortest wavelength achieved in this manner is at 138 nm, generated by a seventh-order anti-Stokes shift in H_2 using 139 nm radiation of an ArF laser. In general, the pump intensity used in experiments with rare-gas halide excimer lasers was in the range of 100 MW/cm^2 to 5 GW/cm^2. The conversion efficiency and the distribution of power among the various Stokes and anti-Stokes

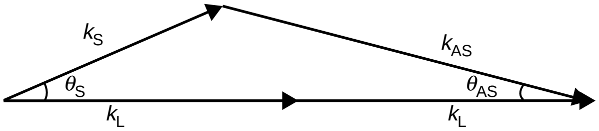

Figure 4 Vector diagram of anti-Stokes Raman scattering in positively dispersive media.

components depend on the focal geometry, and the extent of the spectrum could be varied by changing the gas pressure.

See also

Magneto-Optics: Cyclotron Resonance, Interband Magnetoabsorption, Spin Flip Raman Scattering. **Nonlinear Optics, Applications:** Raman Lasers.

Further Reading

Basiev TT, Sobol AA, Zverey PG, Ivleva LI, Osiko VV and Powell RC (1999) Raman spectroscopy of crystals for stimulated Raman scattering. *Optical Matters* 11: 307.

Bischel WK and Blach G (1983) New lines in the UV: SRS of excimer laser wavelengths. *AIP Conference Proceedings No. 100 (Excimer Lasers-1983)*, 181.

Loree TR, Sze RC, Barker DL and Scott PB (1979) *IEEE Journal of Quantum Electronics* QE-15: 337.

Reinties JF (1985) Coherent ultraviolet and vacuum ultraviolet sources. In: Laser Bass M and Stitch ML (eds) *Laser Handbook*. Amsterdam: North-Holland Physics Publishing.

Weber MJ (ed.) (2001) Raman, Brillouin and Soliton lasers. In: *Handbook of Laser*, p. 291. Boca Raton, FL: CRC Press.

Scattering from Surfaces and Thin Films

A Duparré, Fraunhofer Institut IOF, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The quality of optical components can be critically limited by light scattering. The majority of these components, such as lenses, mirrors, beamsplitters, and polarizers are covered with thin film optical coatings. To reduce scattering losses, their main sources and mechanisms must be known and, whenever possible, be controlled. Moreover, light scattering techniques have been widely recognized as a powerful tool for roughness and defect analysis as well as for general quality assessment.

Scattering can be divided into surface/interface scattering (SIS) and volume scattering (VS), attributed to the following scatter sources:

- roughness and surface defects of the substrate: SIS,
- surface/interface roughness and defects of the thin film coating: SIS,
- imperfections in the bulk substrate material: VS,
- intrinsic thin film morphology and defect structures: VS,
- surface contaminations: SIS.

Even though in particular cases and applications, VS can constitute the main scattering source, SIS largely dominates both substrate and thin film scattering losses. Studies into light scattering have widely focused on SIS effects from surface and interface micro-roughness. The theoretical outline given in the next section will also follow this concept in describing scattering from surface and interface roughness. It is emphasized, however, that volume scattering theories are accessible in the literature.

Light Scattering Models

Scattering from Rough Surfaces

A randomly rough surface can be considered as a Fourier series of sinusoidal waves with different amplitudes, periods, and phases. Following the grating equation, a single grating with spacing D causes scatter into the angle Θ according to

$$\sin\Theta = \lambda/D \qquad [1]$$

where λ is the wavelength of light. D represents one spatial wavelength in the Fourier series. Accordingly, $f = 1/D$ represents one single spatial frequency. A randomly rough surface contains many different spatial frequencies. This is quantitatively expressed by the power spectral density (PSD), giving the relative strength of each roughness component of a surface microstructure as a function of spatial frequency:

$$\mathrm{PSD}(\mathbf{f}) = \lim_{A\to\infty} \frac{1}{A}\left|\int_{\mathbf{A}} \zeta(\mathbf{r})\exp(-2\pi i\mathbf{f}\cdot\mathbf{r})\mathrm{d}\mathbf{r}\right|^2 \qquad [2]$$

where $\zeta(\mathbf{r})$ represents the height of the surface roughness profile, $\mathbf{r}$ is the position vector, and $\mathbf{f}$ is the spatial frequency vector in the $x-y$ plane. A is the area of the measured region $\mathbf{A}$. We confine our discussion to isotropic surfaces, which represent the majority of cases in optical surface and thin film scattering studies. Thus, a $\mathrm{PSD}(f)$ independent of the surface direction Φ of vector $\mathbf{f}$ is obtained by averaging the two-dimensional function $\mathrm{PSD}(\mathbf{f})$ over all surface directions after transformation into polar coordinates:

$$\mathrm{PSD}(f) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{PSD}(f,\Phi)\mathrm{d}\Phi \qquad [3]$$

Well-established vector scattering theories developed, for instance, by Bousquet *et al.* or Elson provide the link between the PSD and the scattering intensity per solid angle of a surface:

$$\frac{\mathrm{d}P}{P_0\,\mathrm{d}\Omega} = F(\lambda, n, \Theta, \Phi)\mathrm{PSD}(f) \qquad [4]$$

This theory is valid for surfaces whose rms roughness is small compared to the wavelength. $\mathrm{d}P/(P_0\,\mathrm{d}\Omega)$ denotes the differential power scattered into the direction (Θ, Φ) per unit solid angle $\mathrm{d}\Omega = \sin\Theta\,\mathrm{d}\Theta\,\mathrm{d}\Phi$ divided by the incident power P_0. Θ and Φ are the polar and azimuthal angles of scattering, respectively. The optical factor F contains all information on the corresponding perfect surface (without the roughness properties), i.e., the refractive index n, wavelength, and the conditions of illumination and observation. Both backscattering and forward scattering can be expressed by eqn [4], according to the illumination and observation conditions chosen. Without loss of generality, all formulas have been written here for normal incidence. The formalism, however, allows consideration of all possible cases, including oblique angles of incidence and arbitrary polarization properties. $\mathrm{d}P/(P_0\,\mathrm{d}\Omega)$ is called angle resolved scattering (ARS) which is related to the well-established term BRDF/BTDF (bidirectional reflectance/transmittance distribution function) by multiplication with $\cos\Theta$:

$$\frac{\mathrm{d}P}{P_0\,\mathrm{d}\Omega} = \mathrm{ARS} = \mathrm{BRDF\ (or\ BTDF)}\cdot\cos\Theta \qquad [5]$$

Total scattering (TS), which is defined as the power P scattered into the backward or forward hemisphere divided by the incident power P_0 (see also section on scattering measurement below), is obtained by integrating eqn [4] over the forward or backward hemisphere:

$$\mathrm{TS} = 2\pi\int_0^{\frac{\pi}{2}}\left(\frac{\mathrm{d}P}{P_0\,\mathrm{d}\Omega}\right)\sin\Theta\,\mathrm{d}\Theta \qquad [6]$$

If the correlation length of surface roughness is much larger than the wavelength, scalar scattering theories like the one from Carniglia can also be employed. Moreover, in this case, the well-known simple approximate formula for total backscattering can be derived from both vector and scalar theories:

$$\mathrm{TS_{back}} = R_0\left(\frac{4\pi\sigma}{\lambda}\right)^2 \qquad [7]$$

where σ is the rms roughness and R_0 the specular reflectance.

It must be emphasized that this formula is only valid if the above-mentioned condition is met and as long as only single surfaces without coatings are considered.

Scattering from Dielectric Thin Films

For surfaces coated with a dielectric single layer or multilayer, ARS for a system of N layers is given by

$$\frac{\mathrm{d}P}{P_0\,\mathrm{d}\Omega} = \sum_{i=0}^{N}\sum_{j=0}^{N} F_i F_j^{*}\mathrm{PSD}_{ij}(f) \qquad [8]$$

where F_i is the optical factor at the i-th interface and F_j^{*} is the conjugate complex number of the optical factor at the j-th interface. The optical factors include both the conditions of illumination and observation and the characteristics of the ideal multilayer (refractive indices, film thickness). PSD_{ij} are the power spectral densities of the corresponding interfaces. For $i \neq j$, they describe the roughness cross-correlation between two interfaces.

From eqn [8] it becomes obvious that the scattering mechanism of coated surfaces is considerably more complex as compared to single surfaces.

The amplitudes of the scattered fields from all individual interfaces add up to the total scatter. In particular, the cross-correlation properties between the interface roughness profiles significantly determine the overall total scatter. So even for coatings having the same statistical roughness, the scatter losses can differ drastically when the cross-correlation between the interfaces is different. An example is demonstrated in a model calculation of total backscattering for $\lambda = 632.8$ nm in Figure 1.

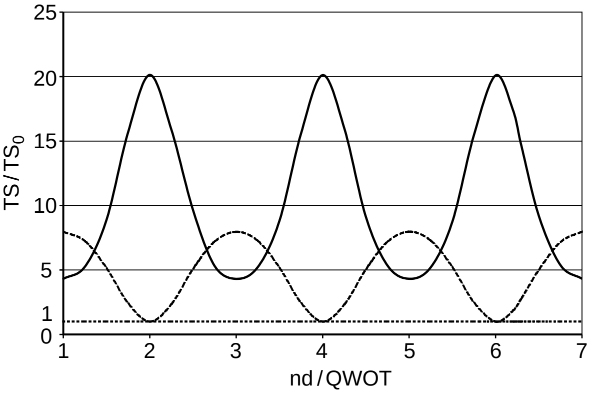

Figure 1 Calculated ratio of the total backscattering of a single high index layer ($n = 2.4$) on BK7 substrate ($n = 1.52$) to the scattering of the bare substrate as a function of optical thickness nd. (QWOT: quarter wave optical thickness). rms roughness was 1 nm, correlation length 1 μm for both the substrate–film and film–air interfaces. Full line: full interface cross-correlation. Dashed line: zero interface cross-correlation. Lower dashed line: bare substrate.

A thin film with refractive index $n = 2.4$ (typical for Titania films), and continuously increasing thickness, was modeled onto a BK 7 substrate ($n = 1.52$). Gaussian autocovariance functions with an rms roughness $\sigma = 1$ nm and a correlation length $\tau = 1$ μm were chosen for both the substrate–film and film–air interfaces. Only the cross-correlation function was varied. The two extreme cases of fully uncorrelated and fully correlated interfaces are depicted in the figure and related to the scattering of the bare substrate (TS_0). As a result of scattering interference effects, in both cases the scattering varies periodically with optical film thickness. The appearance of maxima and minima as well as the total amount of scattering crucially depends on the type of cross-correlation.

These effects have to be taken into careful deliberation whenever scattering in dielectric thin film coatings is considered. Otherwise, misinterpretation of the scatter loss origin is possible and may result in unsuitable technological attempts to minimize scatter losses.

Instrumentation for Light Scattering Measurement

Types of Measurements

Light scattering experiments on optical surfaces and thin films are, in most cases, either performed as ARS measurements using goniophotometers, or the scattering is collected over the backward and forward hemispheres yielding total integrated scattering (TIS) or TS. The latter can be measured either by using a Coblentz sphere or an integrating sphere (see schematic picture in Figure 2).

As Coblentz spheres image the scattered radiation directly onto the detector, they are much less sensitive to scattering by air particles than integrating spheres. So if super-smooth samples with scattering levels in the visible range of 1 ppm and lower shall be measured, integrating spheres require operation in vacuum or He-gas atmosphere whereas Coblentz spheres can still be used under usual clean room conditions.

Angle Resolved Scattering (ARS)

Sophisticated instruments for ARS measurements have been established in a number of laboratories. Most frequently the systems are operated at the He–Ne laser wavelength 632.8 nm, but also at 10.6 μm (CO_2 laser), 325 nm (He–Cd laser), and yet other wavelengths. These techniques are versatile and powerful, but rather unsuitable for routine measurements. For single surfaces, PSD can be calculated from such measurements. A standard procedure for ARS measurements is defined in ASTM standard E 1392 and was successfully verified in various round-robin experiments at different wavelengths. Figure 3 shows the photograph of a typical instrumentation.

The main parts of this setup are the illumination system and a double goniometer. The sample holder and the detector are each mounted on a precision goniometer. The detector head is located on the outer goniometer and can be revolved 360 degrees around the sample with an angular resolution of 0.01°. The performance of a scatterometer depends also considerably on the quality of the illumination system. Laser radiation passes several optical elements, such as a spatial filter, diaphragms, polarizers, and mirrors providing the high beam quality necessary for precision measurement.

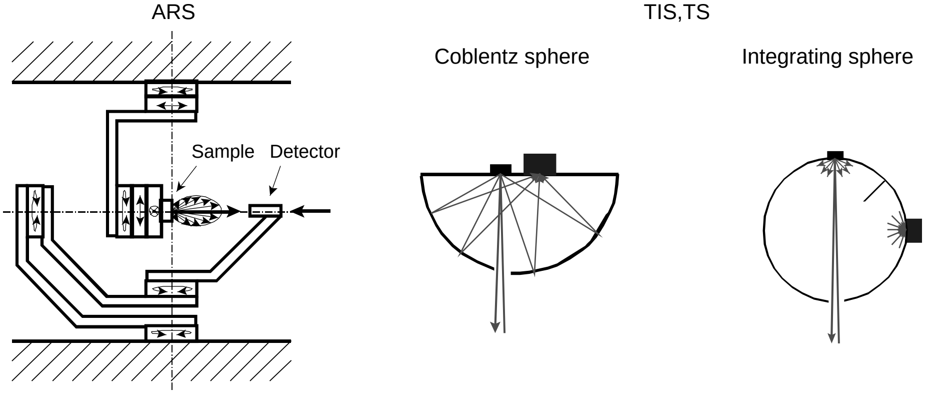

Figure 2 Schematic picture of types of scattering measurement. ARS: angle resolved scattering, TIS, TS: total integrated scattering, total scattering.

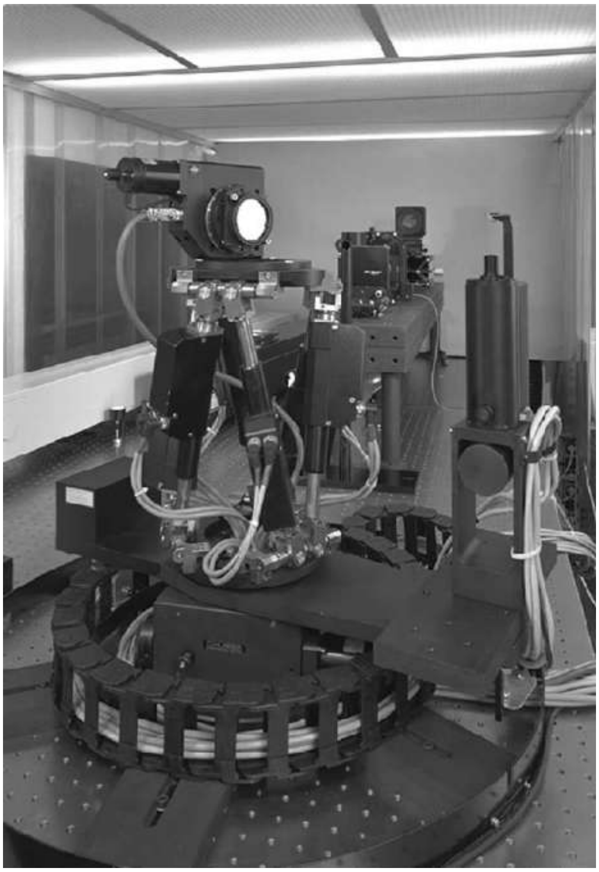

Figure 3 Instrumentation for angle resolved scattering measurement (courtesy of Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany).

Total Scattering (TS)

TIS measurements collect the light scattered into the backward hemisphere and are defined as the backscattered radiation divided by the total reflectance. The TIS measurement procedure is prescribed in ASTM standard F 1048 which applies to opaque reflective surfaces. Transparent or semi-transparent samples such as substrates, AR coatings, and beamsplitters cannot be measured without ambiguous additional assumption. The new international standard ISO 13696 defines TS as the backscattered and forward scattered radiation P divided by the incident radiation P_0. TS is hence equivalent to the scattering loss of the component. Opaque as well as transparent surfaces and coatings can be measured both in the backward and forward directions. TS and TIS can, however, be converted into one another, if the reflectance of the sample is known. This standard procedure was proved in an international round-robin experiment at 632.8 nm.

An example of a facility for TS measurements, in a wide range of wavelengths, is given in Figure 4. The apparatus for total backscattering and forward scattering measurements is operated from 193 nm to 10.6 μm by using several lasers. The setup is based on a Coblentz sphere imaging the light scattered into the backward or forward hemisphere within an angular range from 2° to 85° onto the detector unit. This angular range complies with the specification given in ISO 13696.

A special arrangement allows easy change from the backscatter to forward scatter operation modus. The detector unit consists of the detector (photomultiplier in the UV, VIS, NIR, and HgCdTe-element in the IR) and a small integrating sphere. The latter is used for homogeneously illuminating the detector. At all wavelengths, the light beam is modulated by a chopper and passes a beam-cleaning element. The incident radiation hits the sample surface at nearly zero degrees and the specular beam is guided back through the entrance/exit aperture of the Coblentz sphere. While performing the measurement, the sample is scanned across its surface by a positioning system, yielding one- or two-dimensional scattering diagrams. Calibration is performed with a commercial diffusing (Lambertian) standard. Filters are used for attenuating the beam during measurement of high-scatter samples and diffuse reflectance standards. Background levels smaller than 0.1 ppm at 632.8 nm were achieved without the necessity of He-gas flow or operation in vacuum.

Whereas in the past the majority of such facilities were designed for the visible, infrared, and near-ultraviolet spectral regions, the increasing demands for low-scatter optical components in the vacuum ultraviolet (VUV) region for application in photolithography has driven the development of scattering facilities for wavelengths as short as 157 nm. Because of the enhanced technical effort needed for the development of VUV scattering measurement instrumentation, only few setups exist that can be operated at 157 nm.

Applications

Fields of Applications

There exists a diversity of purposes for light-scattering investigations of surfaces and thin films, which requires proper selection of the particular method. For a number of purposes, routine measurements without detailed understanding of the scatter sources can be sufficient, while other problems need deeper insight into the scatter mechanisms and hence, require the combination of measurement with theoretical modeling. It is up to the engineer or scientist to decide in each particular case, to which level the

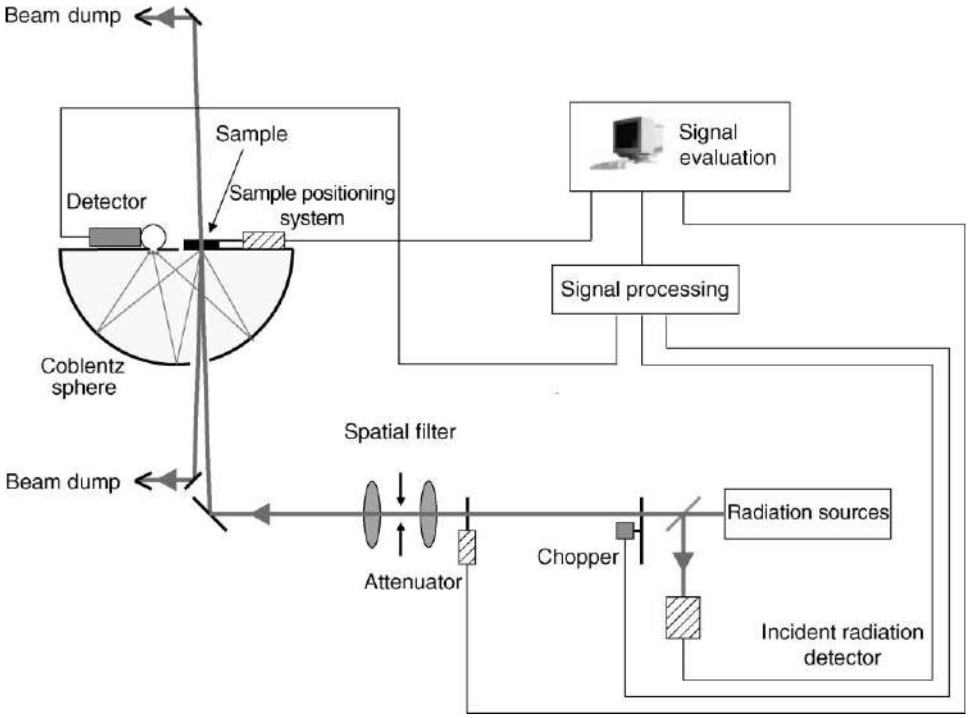
Beam dump
Sample
Sample positioning system
Detector
Coblentz sphere
Signal evaluation
Signal processing
Spatial filter
Beam dump
Radiation sources
Chopper
Attenuator
Incident radiation detector

Figure 4 Schematic picture of an instrument for total scattering measurement (courtesy of Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany).

light-scattering study has to be extended. Examples for typical application tasks arising in industry and institutes are:

- determination of total scattering as an optical loss limiting the performance of a thin film component;
- surface finish assessment (roughness, defects) of optical substrates;
- study into the relation between scattering and the morphology and roughness of optical thin films as a precondition for optimizing deposition processes;
- assessment of cleanliness of coated components;
- detection of degradation effects in coatings and substrate materials;
- separation of interface scattering from the film and substrate;
- distinction of volume scattering from interface scattering in substrates and films.

Scattering measurements are noncontact, do not require sample preparation, enable rapid mapping of large sample areas, and are highly sensitive. This makes them usable for a large variety of applications extending from super-smooth substrates with surface roughness below 0.01 nm, thin film structures in the nanometer range, and small and large defect features up to rough engineering surfaces with micron roughness.

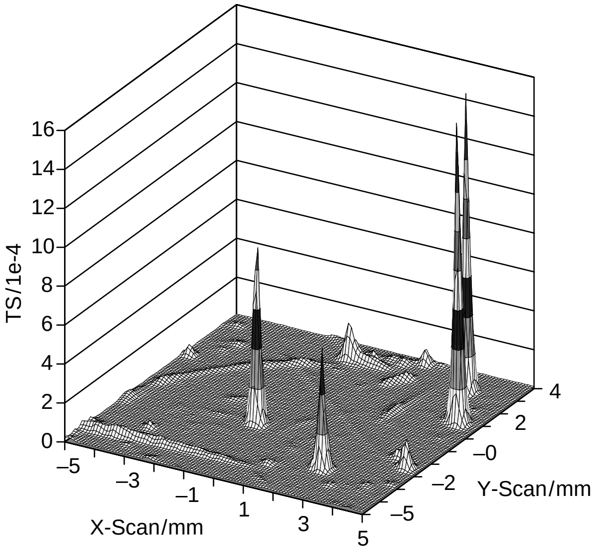

Figure 5 TS measurement (backscattering) on conventionally polished fused silica at 632.8 nm (2D-mapping) for surface quality assessment. Measurement performed with the TS instrument shown in Figure 4.

Examples of Measurements

Surface finish assessment

The result of a scattering measurement (TS, backscattering) at the He–Ne laser wavelength (632.8 nm), on conventionally polished fused silica, is shown in Figure 5. Symmetric defects, as well as scratch-like features, can be recognized in this diagram. The minimum and averaged rms-roughness determined from the minimum (defect-free) and averaged (including defects) scattering levels are 0.51 and 0.72 nm, respectively.

Scattering from thin film components

Figure 6 shows a TS measurement (backscattering) performed on a polished CaF_2 sample. Half of the area of one surface of this sample was coated with a fluoride multilayer mirror system designed for 248 nm. The area mapping of the TS measurement at 248 nm delivers information of both the bare substrate quality and the increase in scattering after coating. According to the explanations in the first section, this increase has to be interpreted as a result of two influences: the increased reflectance (i.e., changed optical factors) and the roughness growth during deposition of the multilayer (i.e., changed PSD).

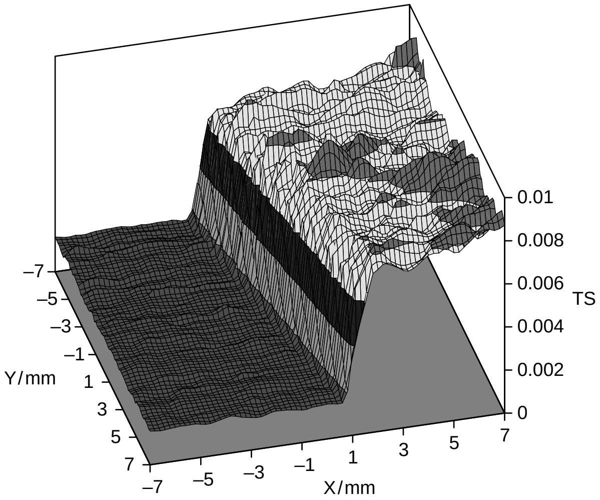

Figure 6 TS measurement (backscattering) at 248 nm on a CaF_2 substrate half of which was coated with a fluoride multilayer mirror system designed for 248 nm. Measurement performed with the TS instrument shown in Figure 4.

On the basis of such measurements, detailed interpretation can be accomplished by additional roughness analysis using scanning force microscopy, subsequent scattering modeling employing the formulas given above, and comparison of the modeled and measured results.

ARS techniques can, for example, be employed to compare the angular distribution of the scattering of a coating to that of the bare substrate. Figure 7 displays ARS measurements in the backward hemisphere at 325 nm (He–Cd laser) on a fused silica substrate before and after deposition (magnetron sputtering) of a Ta_2O_5/SiO_2 quarterwave multilayer mirror (325 nm design wavelength). As with the example discussed above, the higher scatter of the multilayer is caused by both the optical factors and increased roughness. The slight bump at large scatter angles in particular reveals high spatial frequency roughness attributed to the film morphology.

Figure 8 displays TS measurements at 157 nm. The diagram contains one-dimensional scans of forward scattering from an uncoated CaF_2 substrate and an antireflective (AR) coating on CaF_2. The background signal level is also included.

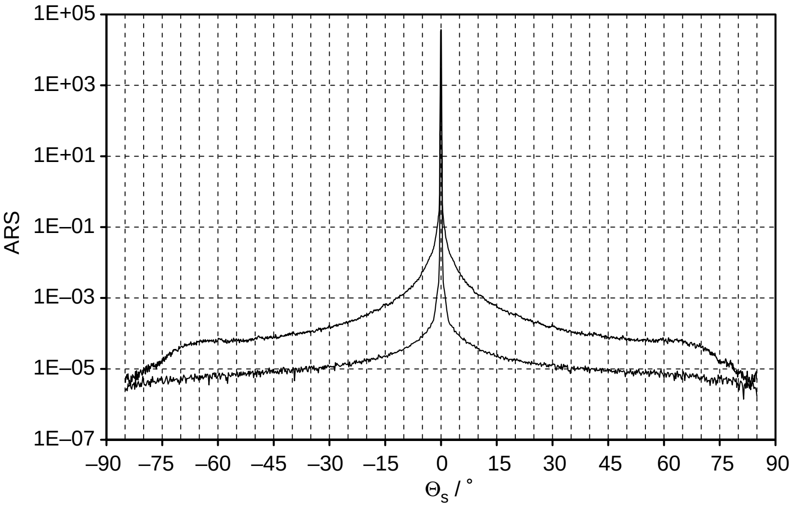

Figure 7 ARS measurement at 325 nm (backward hemisphere) on a fused silica substrate before (lower curve) and after (upper curve) deposition of a Ta_2O_5/SiO_2 quarterwave multilayer mirror with 325 nm design wavelength. Measurements performed with the ARS instrument shown in Figure 3.

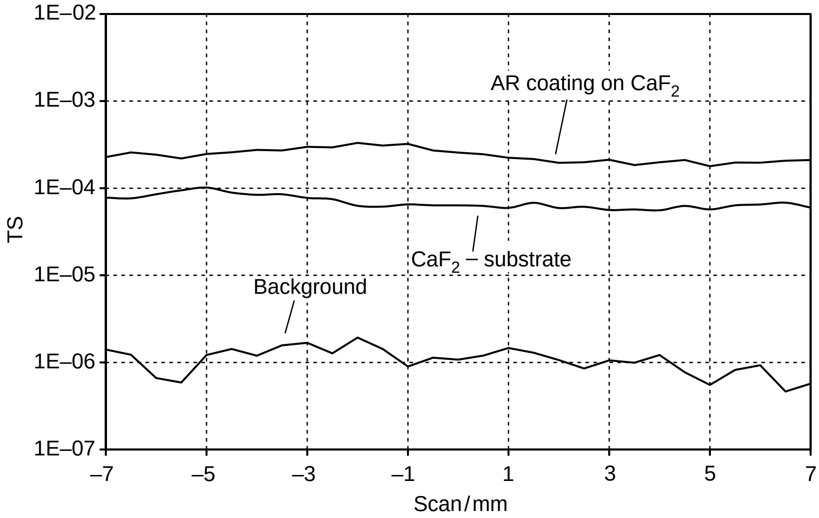

Figure 8 TS measurements at 157 nm (forward scattering): bare CaF$_2$ substrate, AR coating on CaF$_2$ and background scattering level. Measurements performed with the VUV total scattering instrumentation at the Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany.

List of Units and Nomenclature

Angle resolved scattering	[nondimensional]	ARS
Azimuthal scattering angle	[degree]	Φ
Correlation length	[nm]	τ
Differential scattered power	[nondimensional]	$\mathrm{d}P/(P_0\,\mathrm{d}\Omega)$
Grating spacing	[μm]	D
Height of surface roughness profile	[nm]	ζ
Incident power	[W]	P_0
Optical factor	[nm^{-4}]	F
Polar scattering angle	[degree]	Θ
Power spectral density	[nm^4]	PSD
Refractive index	[nondimensional]	n
Rms roughness	[nm]	σ
Spatial frequency	[μm^{-1}]	f
Specular reflectance	[nondimensional]	R_0
Total integrated scattering	[nondimensional]	TIS
Total scattering	[nondimensional]	TS
Wavelength	[nm]	λ

See also

Optical Coatings: Thin-Film Optical Coatings. **Semiconductor Physics:** Light Scattering.

Further Reading

Baeumer S, Duparré A, Herrmann T, *et al.* (1999) SLIOS – a contribution to standard procedures in stray light measurement. *Proceedings SPIE* 3739: 414–421.

Bennett JM and Mattsson L (1999) *Introduction to Surface Roughness and Scattering*, 2nd edn., pp. 28–35, 62–70. Washington, DC: Optical Society of America.

Bousquet P, Flory F and Roche P (1981) Scattering from multilayer thin films: theory and experiment. *Journal of Optical Society of America* 71: 1115–1123.

Carniglia CK (1979) Scalar scattering theory for multilayer optical coatings. *Optical Engineering* 18: 104–115.

Church L, Jenkinson HA and Zavada JM (1979) Relationship between surface scattering and microtopographic features. *Optical Engineering* 18: 125–136.

Duparré A (1995) Light scattering of thin dielectric films. In: Hummel RE and Günther KH (eds) *Thin Films for Optical Coatings*, pp. 273–303. Boca Raton, FL: CRC Press.

Elson JM (1984) Theory of light scattering from a rough surface with an inhomogeneous dielectric permittivity. *Physical Review B* 30: 5460–5480.

Elson JM and Bennett JM (1979) Vector scattering theory. *Optical Engineering* 18: 116–124.

Herman IP (1996) *Optical Diagnostics for Thin Film Processing*, pp. 522–529. London: Academic Press.

Kadkhoda P, Müller A, Ristau D, *et al.* (2000) International round-robin experiment to test the international organization for standardization total-scattering draft standard. *Applied Optics* 39: 3321–3332.

Kassam S, Duparré A, Hehl K, Bussemer P and Neubert J (1992) Light scattering from the volume of optical thin films: theory and experiment. *Applied Optics* 31: 1304–1313.

Leonhard TA and Rudolph P (1995) BRDF round robin test of ASTME 1392. *Proceedings SPIE* 1995: 285–293.

Macleod HA (2001) *Thin-Film Optical Filters*, 3rd edn., pp. 12–52. Bristol, UK and Philadelphia: Institute of Physics Publishing.

Rönnow D and Veszelei E (1994) Design review of an instrument for spectroscopic total integrated light scattering measurements in the visible wavelength region. *Review of Scientific Instruments* 65: 327–334.

Stover JC (1995) *Optical Scattering, Measurement and Analysis*, 2nd edn., pp. 133–176. Bellingham, CA: SPIE Press.

Van de Hulst HC (1957) *Light Scattering by Small Particles*, pp. 28–31, 63–84. London: Chapman and Hall Ltd.

Whitehouse DJ (1994) *Handbook of Surface Metrology*, pp. 49–60, 512–517. Bristol, UK and Philadelphia: Institute of Physics Publishing.

Scattering Phenomena in Optical Fibers

P Tchofo-Dinda and G Millot, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear phenomena that result from interaction of intense light beams with dielectric media can be broadly divided into two main categories: namely, parametric effects and scattering phenomena. Parametric effects arise whenever the state of the matter is left unchanged by the interaction, whereas scattering phenomena imply transitions between vibrational or rotational energy levels in the medium. The two dominant scattering phenomena in optical fibers are stimulated Raman scattering (SRS) and stimulated Brillouin scattering (SBS), which are nonlinear processes in which a part of the energy propagating at a given frequency is converted into one or several down-shifted Stokes beams, through interaction with the vibrational modes of the materials. Although SBS and SRS phenomena are beneficial for developing all-optical fiber lasers and amplifiers for long-haul fiber transmissions, they are detrimental to transmission of ultrashort light pulses in silica fibers. In this article we describe the fundamental aspects of SBS and SRS.

General Features

Three well-known kinds of scattering phenomena in optical fibers are Rayleigh, Raman and Brillouin scattering. Although these three phenomena manifest themselves in qualitatively different ways, all of them originate fundamentally from the effects of a light wave on the atomic and molecular charges of the dielectric medium. The Raman and Brillouin scattering phenomena involve the vibrational modes of the material that are associated with the optical and acoustic branches of the dispersion curve of the material, respectively. Such vibrational modes, called phonons, are schematically represented in Figure 1 with the help of typical dispersion curves of a one-dimensional atomic lattice.

In standard silica fibers (SiO_2), the phonons associated with the optical branch correspond to intramolecular vibrations, which give rise to the Raman scattering phenomenon, whereas those associated with the acoustic branch correspond to intermolecular vibrations, which are responsible for Brillouin scattering. In either of these two scattering processes, part of the energy of the incoming light, which propagates at a given frequency ω, is converted into downshifted (Stokes) or upshifted (anti-Stokes) light waves, at frequencies $\omega - \Omega$ and $\omega + \Omega$, respectively. The frequency shift Ω is determined by the vibrational modes involved in the scattering process. Figure 2 represents schematically

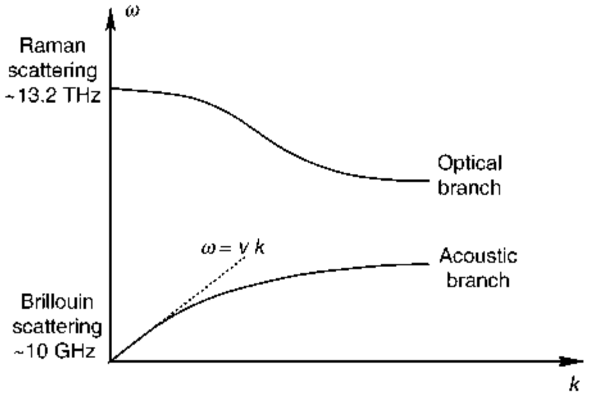

Figure 1 Schematic representation of the dispersion curves for material waves.

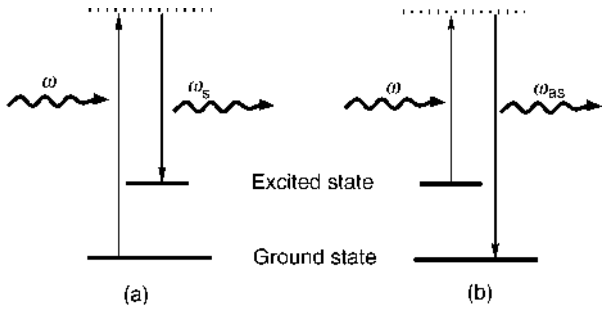

Figure 2 Schematic representation of the diagrams of energy transfer between incoming photons with frequency ω and the material, leading to (a) generation of the Stokes radiation ω_s, and (b) generation of the anti-Stokes radiation ω_{as}.

these energy conversion processes with the help of only two vibrational states, to simplify matters.

As Figure 2a shows, the material, which is initially in the ground state, absorbs a fraction $\hbar\Omega$ of the energy $\hbar\omega$ of incoming photons, to move up to the excited state. As a result, the frequency of the photons involved in this process is downshifted to $\omega_s = \omega - \Omega$. Those photons that are generated at frequency ω_s are commonly called 'Stokes radiation'. On the other hand, when a sufficiently large number of molecules is excited, they can interact with the incoming photons and return back to the ground state. As Figure 2b shows, in this process the energy is transferred from the material to the incoming photons, which thereby generates an upshifted anti-Stokes radiation at frequency $\omega_{as} = \omega + \Omega$. The Raman (Brillouin) frequency shift Ω is of the order of $\Omega_R \approx 13\,\text{THz}$ ($\Omega_B \approx 10\,\text{GHz}$).

In fact, the Raman and Brillouin effects may lead to so-called spontaneous processes, which occur in the limit of low-amplitude for all waves other than the incoming wave. For example, when a laser beam is injected in the fiber with a sufficiently low intensity, one can observe a spontaneous scattering process as schematically represented in Figure 3a.

One can clearly identify in Figure 3a the Raman and Brillouin radiation, as well as radiation scattered at the same frequency as the incoming wave, that is, Rayleigh radiation. Now, when the incoming wave propagates together with light waves whose intensity is no longer negligible with respect to that of the incoming wave, the Raman and Brillouin effects can lead to stimulated processes, in which energy is continually converted from higher to lower frequency photons. As a result, the number of Stokes photons is continually amplified at the expense of anti-Stokes photons, as schematically represented in Figure 3b. The stimulated process can be obtained by either launching the incoming wave ω together with a probe wave at frequency $\omega - \Omega$ (which corresponds to the situation represented in Figure 3b), or by using a sufficiently intense incoming light beam. Indeed, above a certain power threshold (which depends strongly on the fiber length and the amount of dopants in the fiber), for which the gain corresponding to the Stokes amplification compensates for the linear fiber losses, the energy conversion from the incoming wave to the scattered waves can become sufficiently important for the scattering process to become stimulated. This process is commonly referred to as a self-stimulated process, as it does not require the use of a probe wave. Although the SRS and SBS exhibit some common general features as those mentioned above, there exist however some major differences between these two phenomena, from a fundamental point of view as well as for their practical applications to optical communication systems. We discuss separately these specific features for SRS and SBS.

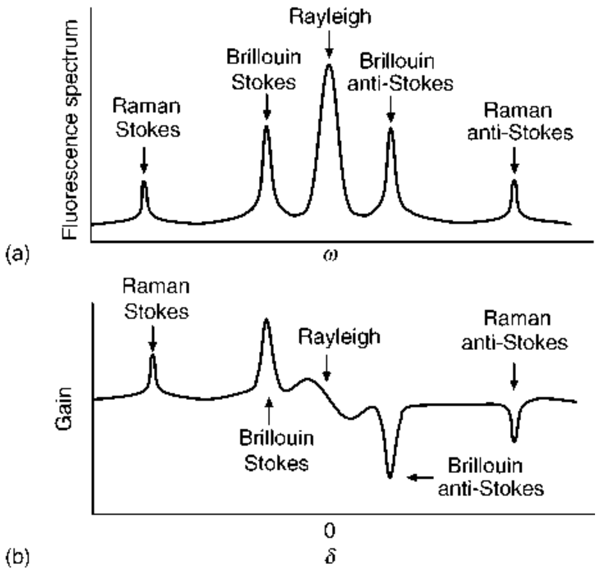

Figure 3 (a) Schematic representation of the frequency spectra for spontaneous and stimulated Raman and Brillouin scattering processes. (a) Fluorescence spectrum; (b) gain spectrum for stimulated Raman and Brillouin scattering processes. A pump and a probe waves are injected together in the fiber. δ is the frequency spacing between the pump and the probe wave.

Stimulated Raman Scattering

Under adiabatic following by the electrons of the optical-field-induced nuclear motions, the third-order nonlinear polarization in optical fibers may be

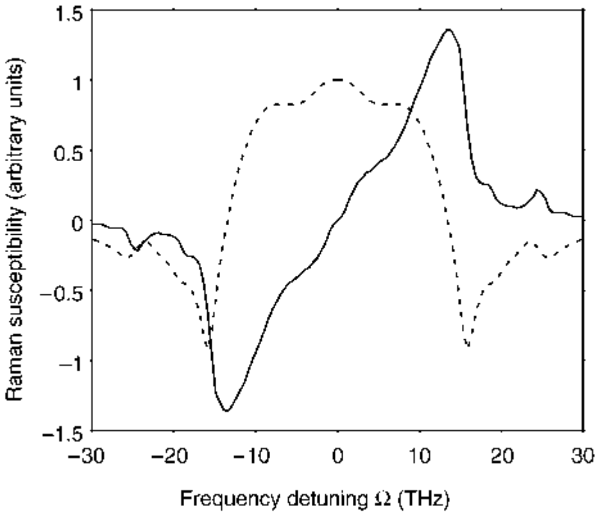

Figure 4 Schematic representation of the Raman susceptibility in an optical fiber. The solid curve represents the imaginary part of the susceptibility, that is, the Raman gain. The dashed curve represents the real part of the susceptibility.

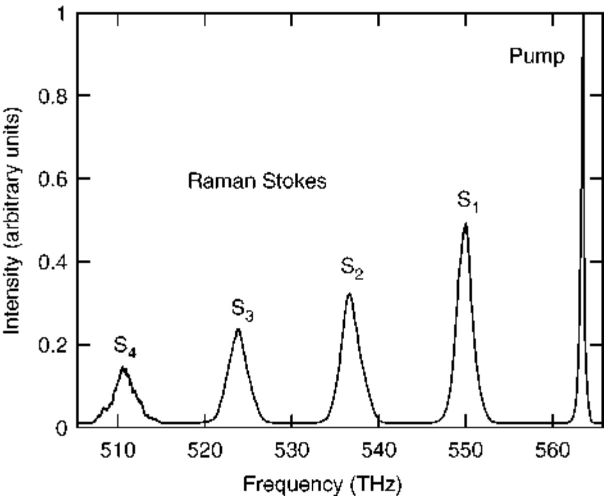

Figure 5 Self-stimulated Raman spectrum (i.e., amplified spontaneous noise), as recorded for a peak pump power $P_0 = 150$ W and a fiber length $L = 10$ m.

written as

$$P(t) = \sigma|E(t)|^2E(t) + E(t)\int_{-\infty}^{t} \mathrm{d}s\sigma_R(t-s)|E(s)|^2 \quad [1]$$

where σ and σ_R represent the instantaneous (Kerr) and noninstantaneous (Raman) parts of the fiber response, respectively. Here, we consider linearly polarized waves, for simplicity. Figure 4 shows schematically the Raman susceptibility of the fused silica, $\chi_R(\Omega) = \sigma_R(\Omega)/2$, which is quite close to that of standard silica fibers.

In Figure 4, the real and imaginary parts of χ_R are represented in solid and dashed curves, respectively. The imaginary part of the Raman susceptibility, which corresponds to the Raman gain $G_R = \mathrm{Im}(\chi_R)$, is an experimentally measurable quantity. For example, at a pump wavelength $\lambda = 1\,\mu$m, the maximum value of the Raman gain in standard silica fibers is $\approx 10^{-13}$ mW^{-1}. In fact the Raman gain $G_R(\Omega)$ depends on the type and amount of dopants used in the fabrication of the fiber. In contrast to crystalline silica where the Raman gain occurs at a well-defined frequency of 13.2 THz, the Raman gain of silica fibers extends over a large frequency range, as can be seen in Figure 4, with a broad prominent peak around 13.2 THz. In the absence of an input signal the broadband noise is amplified within the Raman Stokes components, which are centered about the frequency detuning of 13.2 THz.

If the power of the Stokes wave becomes large enough, the Stokes wave serves as a pump to generate a second-order Stokes wave. This process of cascade SRS can generate multiple Stokes waves whose number depends on the input pump power and fiber length. Figure 5 displays a typical experimental Raman spectrum generated from a Q-switched, frequency-doubled Nd:YAG laser, emitting 5 ns duration (30 GHz spectral width) pulses at 563.63 THz (532.26 nm). The laser output was injected along one birefringence axis of a 14 m length of polarization-maintaining single-mode fiber. At an estimated injected peak power of 200 W, several higher-order Raman Stokes orders S_j ($j = 1, 4$) (separated by 13.2 THz) are generated between 510 and 550 THz (588 and 545 nm). The spectral profile of each resulting Stokes line is near-Lorentzian with a full width at half maximum ranging from about 2 THz for the first Stokes to about 3.5 THz for the fourth Stokes. The broadening of the Raman Stokes wave, as the Stokes order increases, is due to several competing nonlinear processes.

This broadband gain is due to the amorphous nature of fused silica, in which the molecular vibrational frequencies spread out into bands which overlap and create a continuum. This outstanding feature has been fully exploited to develop wide-band amplifiers for multichannel optical communications. Recent work demonstrated the achievement of a Raman amplifier with a bandwidth of 100 nm, by use of a set of height pumps and an appropriate choice of the frequency separations between the pumps. Another important application of the Raman gain in optical fibers lies in the development of fiber-Raman lasers, which can be tuned over a wide frequency range (≈10THz). Such lasers can be obtained by inserting a piece of single-mode fiber inside a Fabry–Perot cavity formed by two partially reflecting mirrors. A rotating output mirror and an intracavity prism allow one to select the Stokes light that provides the laser wavelength. Another example is that the Raman gain can be used to assist non-phase-matched parametric interactions.

A large power-gain enhancement was recently demonstrated for nonphase-matched waves in a three-wave mixing interaction. The Raman-assisted three-wave mixing can be used to achieve wide-bandwidth frequency conversion processes without having to satisfy the strict conditions imposed by phase-matching conditions. The above discussion clearly illustrates some useful effects of the SRS in optical fibers. On the other hand, the SRS may represent a drawback for some optical communication systems, as we discuss below.

Let us consider a pump wave injected in the fiber at frequency ω_0. The total electric field in the fiber may be written as follows:

$$E = \frac{1}{2}\mathcal{A}(z,t)\exp[\mathrm{i}(k_0 z - \omega_0 t)] \qquad [2]$$

where $\mathcal{A}$ is a slowly varying field envelope. By expanding the wavevector k_0 of the electric field in a Taylor series around ω_0, the amplitude $\mathcal{A}$ of the electric field is found to satisfy the generalized nonlinear Schrödinger equation (NLSE):

$$\begin{aligned}\frac{\partial\mathcal{A}}{\partial z} &+ \frac{\mathrm{i}}{2}\beta\frac{\partial^2\mathcal{A}}{\partial t^2}\\ &= \mathrm{i}\gamma(1-\rho)|\mathcal{A}|^2\mathcal{A} + \mathrm{i}\gamma\rho\mathcal{A}\int_0^\infty \chi_R(s)|\mathcal{A}|^2(t-s)\mathrm{d}s\\ &= \mathrm{i}\gamma|\mathcal{A}|^2\mathcal{A} + \mathrm{i}\gamma\rho C_1\mathcal{A}\frac{\partial|\mathcal{A}|^2}{\partial t} + \frac{\mathrm{i}}{2}\gamma\rho C_2\mathcal{A}\frac{\partial^2|\mathcal{A}|^2}{\partial t^2}\\ &\quad + \frac{\mathrm{i}}{6}\gamma\rho C_3\mathcal{A}\frac{\partial^3|\mathcal{A}|^2}{\partial t^3} + \cdots \end{aligned} \qquad [3]$$

In eqn [3], $C_1 \approx -7.9\times10^{-3}\,\mathrm{ps}$, $C_2 \approx -1.28\times10^{-4}\,\mathrm{ps}^2$, $C_3 \approx 9\times10^{-6}\,\mathrm{ps}^3$, and the parameter ρ, which measures the fractional contribution of Raman effects to the total nonlinearity, is ≈ 0.18 for standard silica fibers.

When the wave propagation involves only a relatively small frequency range around the pump frequency (less that 1 THz) then only the term proportional to C_1 in eqn [3], i.e., the leading term of the Raman contribution, plays a significant role. Most high-capacity transmission systems using a single channel fall into this situation, as long as picosecond pulses are used. The term proportional to C_1 induces a continuous frequency shift of the pulse frequency, as a result of a small transfer of energy from higher- to lower-frequency waves. This phenomenon, known as the soliton self-frequency shift, constitutes one of the undesirable effects of SRS in optical communications.

On the other hand, when the wave propagation involves a relatively large frequency region (of the order of a few THz), then higher-order terms that are proportional to C_2 and C_3 begin to play a significant role. In particular, the term proportional to C_3 indicates a strong transfer of energy from higher- to lower-frequency waves. In optical communications, this effect causes higher-wavelength channels of wavelength-division multiplexing systems to be amplified while depleting the lower-wavelength channels. This effect can be reduced by using filters and more amplification to compensate for the depletion at lower wavelengths (but this is achieved at the expense of a degradation of the signal-to-noise ratio).

Stimulated Brillouin Scattering

In optical fibers, the SBS occurs at input power levels that are much lower that those needed for SRS, and manifests itself through the generation of a backward-propagating Stokes wave. Although SBS is not (strictly speaking) a parametric process, the generation of the Brillouin Stokes radiation is commonly described as a parametric interaction between the pump wave, the Stokes wave and an acoustic wave (material wave). Indeed, one assumes that the annihilation of a pump photon creates simultaneously a Stokes photon and an acoustic phonon, with the following conservation laws for the energy and the momentum:

$$\Omega_\mathrm{B} = \omega_0 - \omega_\mathrm{s} \qquad [4]$$

$$K_\mathrm{B} = k_0 - k_\mathrm{s} \qquad [5]$$

where ω_0 and ω_s are the frequencies, and k_0 and k_s are the wavevectors of the incoming and Stokes waves, respectively. Then, assuming $|k_\mathrm{s}| \approx |k_0|$, the acoustic wave is found to satisfy the following dispersion relation

$$\Omega_\mathrm{B} = K_\mathrm{B}v = 2v|k_0|\sin(\theta/2) \qquad [6]$$

where v is the phase velocity of the acoustic wave, and θ is the angle between the pump and Stokes wavevectors. As eqn [6] shows, the frequency shift

Ω_B depends on the scattering angle. Ω_B is a maximum for $\theta = \pi$ (i.e., in the backward direction) and vanishes for $\theta = 0$ (forward direction). Thus, although eqn [6] predicts that Brillouin scattering should not occur in the forward direction, spontaneous Brillouin scattering can occur in the forward direction in optical fibers. But this phenomenon is quite negligible. In optical fibers, SBS occurs mainly in the backward direction with a frequency shift given by

$$\nu_B = \Omega_B/(2\pi) = 2v|k_0| = 2n_0 v/\lambda_0 \qquad [7]$$

where n_0 is the refractive index and λ_0 is the pump wavelength (incoming wave). In silica fibers, $v \approx 6\,\text{km s}^{-1}$ and $n_0 \approx 1.45$; which leads to $\nu_B \approx 10\,\text{GHz}$ at $\lambda_0 = 1.55\,\mu\text{m}$. Thus, the Brillouin frequency shift is smaller by three orders of magnitude compared with the Raman frequency shift ($\approx$13 THz). The amorphous nature of fused silica leads to a Brillouin-gain coefficient $G_B(\nu)$ which extends over a frequency range of $\approx$ 10 MHz, with a peak value at $\nu = \nu_B$. The Brillouin gain can be approximated by a Lorentzian profile given by

$$G_B(\nu) = \frac{(\Delta\nu_B/2)^2}{(\nu - \nu_B)^2 + (\Delta\nu_B/2)^2} G_B(\nu_B),$$
$$G_B(\nu_B) = \frac{2\pi n_0^7 p_{12}^2}{c d_0 \lambda_0^2 v \Delta\nu_B} \qquad [8]$$

where $\Delta\nu_B$ is the full width at half maximum, p_{12} is the elasto-optic coefficient and d_0 is the material density.

The Brillouin gain has been exploited to develop fiber-Brillouin lasers. Such lasers are obtained by inserting a piece of fiber inside a (ring or Fabry–Perot) cavity. Another useful application of the Brillouin gain of an optical fiber lies in the amplification of a weak signal. Fiber-Brillouin amplifiers allow one to amplify signals with a frequency shift that corresponds to the Brillouin frequency shift ν_B. Fiber-Brillouin amplifiers can provide up to 40 dB gain at pump powers of a few milliwatts. But the bandwidth of such amplifiers is relatively narrow.

Conclusion

In fact scattering phenomena can be beneficial as well as detrimental for optical communication systems. If signals are transmitted in a communication channel with a power above the threshold of a given scattering phenomenon, this phenomenon will induce a transfer of energy from the signal to the Stokes radiation. The resulting signal depletion would then require more amplification. In multichannel communications, the SBS (SRS) can induce cross-talk between two bidirectional channels if their frequency separation happens to be close to the Brillouin (Raman) shift, $\approx$10 GHz ($\approx$13 THz). However, those undesirable effects can be avoided by in-line control techniques for pulse transmission in fiber-optics communication systems. On the other hand, the Raman and Brillouin effects are useful for generating light at new frequencies as well as for amplifying weak light signals. In particular Raman amplifiers are one of the key devices for high-speed telecommunication systems.

List of Units and Nomenclature

Group velocity dispersion coefficient β_2 [$\text{ps}^2\,\text{km}^{-1} = 10^{-27}\text{s}^2\,\text{m}^{-1}$]
Nonlinear parameter γ [$\text{W}^{-1}\,\text{m}^{-1}$]
NLS (E): nonlinear Schrödinger (equation)
Optical intensity I [$\text{GW cm}^{-2} = 10^9\,\text{W cm}^{-2}$]
SBS: stimulated Brillouin scattering
SRS: stimulated Raman scattering
WDM: wavelength division multiplexing

See also

Fiber and Guided Wave Optics: Light Propagation; Nonlinear Effects (Basics). **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2001) Nonlinear Fiber Optics. San Diego, CA: Academic Press.

Agrawal GP (2001) Applications of Nonlinear Fiber Optics. San Diego, CA: Academic Press.

Bloembergen N (1977) Nonlinear Optics. Reading, MA: Benjamin.

Dianov EM, Mamyshev PV, Prokhorov AM and Serkin VN (1989) Nonlinear Effects in Optical Fibres. Switzerland: Harwood Academic Publishers.

Hellwarth R (1977) Third-order optical susceptibilities of liquids and solids. *Progress in Quantum Electronics* 5: 1–68.

Pocholle JP, Papuchon M, Raffy J and Desurvire E (1990) Non linearities and optical amplification in single mode fibres. *Revue Technique Thomson-CSF* 22: 187–268.

Shen YR (1984) Principles of Nonlinear Optics. New York: Wiley.

Scattering Theory

Y A Eremin, Moscow State University, Moscow, Russia

© 2005 Published by Elsevier Ltd.

Lord Rayleigh (John William Strutt) (1842–1916)

In the nineteenth century, Lord Rayleigh offered the first explanation for the sky's blue color (on a clear day!). From Rayleigh's initial theory, representing a simplification of Maxwell's electromagnetic theory for the case of particles very small compared to the wavelength of the incident light, the theory was extended to describe the scattering of light by larger macromolecules in solution.

Scattering theory has played a central role in twentieth-century mathematical physics. From Rayleigh's explanation of why the sky is blue to modern medical application of computerized tomography, has attracted both scientists and mathematicians for over a hundred years. Scattering theory is concerned with the effect an inhomogeneous medium has on an incident wave. In particular, if the total field is viewed as sum of an incident field and a scattered field, then the direct scattering problem is to determine the scattered field from a knowledge of the incident one, obstacle properties, and the equations governing the wave motion. In contrast, the inverse scattering problem consists of reconstruction of the inhomogeneity from the knowledge of asymptotic behavior of the scattered field. An everyday example is human vision: from the measurements of scattered light that reaches our retinas, our brain reconstructs a detailed three-dimensional map of the world around us. In summary, the task of direct scattering theory is to determine the relation between the incident and scattered waves. The task of inverse scattering theory is to determine properties of the obstacle.

Light propagating can be described in terms of an electromagnetic field, which is composed of two related vector fields, the electric and the magnetic fields. This means that the vectors of electric (**E**) and magnetic (**H**) field have values defined at each point in space and time. The vector fields **E** and **H** interact through the material parameters: electric permittivity ε and magnetic permeability μ. The behavior of electromagnetic fields can be described by Maxwell's equations.

Maxwell Equations. James Clerk Maxwell (1831–1879)

Consider the electromagnetic wave (light) propagation in a homogeneous, isotropic medium with electric permittivity $\varepsilon > 0$, electric conductivity $\sigma \geq 0$, and magnetic permeability $\mu > 0$, containing a local obstacle. We assume that $\sigma = 0$ outside an obstacle. If J is a current density, the electric field $\vec{E}(x,t)$ and magnetic field $\vec{H}(x,t)$ satisfy the Maxwell equations:

$$\begin{aligned} \nabla \times \vec{E}(x,t) + \mu \frac{\partial \vec{H}}{\partial t}(x,t) &= 0 \\ \nabla \times \vec{H}(x,t) - \varepsilon \frac{\partial \vec{E}}{\partial t}(x,t) &= \vec{J}(x,t) \end{aligned} \qquad [1]$$

Also, in an isotropic conductor, the electric field satisfies Ohm's law:

$$\sigma \vec{E}(x,t) = \vec{J}(x,t) \qquad [2]$$

The speed of wave propagation for the electric and magnetic fields is $c = 1/\sqrt{\varepsilon\mu}$ and in free space $c = 1/\sqrt{\varepsilon_0\mu_0} = 2.99792458 \times 10^8\ \mathrm{m\ s^{-1}}$. Assume that the electromagnetic field is time-harmonic, i.e., of the form:

$$\begin{aligned} \vec{E}(x,t) &= \frac{1}{\sqrt{\varepsilon_0}} \mathbf{E}(x) e^{-j\omega t} \\ \vec{H}(x,t) &= \frac{1}{\sqrt{\mu_0}} \mathbf{H}(x) e^{-j\omega t} \end{aligned} \qquad [3]$$

where $\omega > 0$ is the frequency. Many phenomenon of light scattering can be described based on monochromatic wave propagation. Let us consider a local obstacle occupying a bounded domain D. Inside the obstacle we introduce relative complex permittivity:

$$\varepsilon = \frac{1}{\varepsilon_0}\left(\varepsilon + j\frac{\sigma}{\omega}\right) \qquad [4]$$

Then $\mathbf{E}(x)$ and $\mathbf{H}(x)$ obey the time-harmonic Maxwell equations:

$$\nabla \times \mathbf{E}(x) = jk\mathbf{H}(x), \quad \nabla \times \mathbf{H}(x) = -jk\varepsilon\mathbf{E}(x) \qquad [5]$$

where $k^2 = \varepsilon_0\mu_0\omega^2$.

Helmholtz Equation. Herman Ludwig von Helmholtz (1821–1894)

A wave equation for vector **E** can be readily derived from the Maxwell equations. We obtain:

$$(\nabla^2 + k^2)\mathbf{E}(x) = 0 \qquad [6]$$

This is known as the homogeneous Helmholtz equation. A general planewave solution of the equation can be written as

$$\mathbf{E}(x) = \mathbf{E}_0\, e^{j\vec{k}\cdot\vec{x}} \qquad [7]$$

here $\mathbf{E}_0$ is a constant vector independent of spatial coordinates and $\vec{k}\cdot\vec{x}$ is a scalar product of the vectors $\vec{x} = \{x_1, x_2, x_3\}$ and $\vec{k} = \{k_1, k_2, k_3\}$, the so called wavevector. Equations [6] and [7] yield the dispersion relation:

$$k^2 = \varepsilon_0 \mu_0 \omega^2 = k_1^2 + k_2^2 + k_3^2 \qquad [8]$$

For a given $\vec{k}$ vector, a constant phase front is determined by $\vec{k}\cdot\vec{x} = \text{const}$, which indicates that the front is perpendicular to the wavevector. So, the phase front is a plane and the amplitude of the electric field on the plane is a constant.

The Maxwell equations for the plane wave solution become:

$$\begin{aligned} &\vec{k} \times \mathbf{E}(x) = k\mathbf{H}(x), \quad \vec{k} \times \mathbf{H}(x) = -k\mathbf{E}(x) \\ &\Rightarrow \vec{k}\cdot\mathbf{E} = 0, \quad \vec{k}\cdot\mathbf{H} = 0 \end{aligned} \qquad [9]$$

Therefore, vectors $\mathbf{E}$ and $\mathbf{H}$ belong to the constant phase plane perpendicular to vector $\vec{k}$ and are orthogonal each other.

Poynting's Vector. John Henry Poynting (1852–1914)

For time-dependent fields, the Poynting vector is introduced as

$$\vec{S}(x,t) = \vec{E}(x,t) \times \vec{H}(x,t) \qquad [10]$$

It is interpreted as the power flow density with the dimension of watts/m^2. The time-average Poynting's vector power density is given by

$$\langle S \rangle = \frac{1}{T} \int_0^T \vec{S}(x,t)\mathrm{d}t \qquad [11]$$

For the time-dependent planewave fields:

$$\begin{aligned} \vec{E}(x,t) &= \mathbf{E}_0 \cos(\vec{k}\cdot\vec{x} - \omega t) \\ \vec{H}(x,t) &= \mathbf{H}_0 \cos(\vec{k}\cdot\vec{x} - \omega t) \end{aligned} \qquad [12]$$

The vector Poynting's power density shown in eqn [10], accepts the form:

$$\vec{S}(x,t) = |\mathbf{E}_0|^2 \cos^2(\vec{k}\cdot\vec{x} - \omega t) \qquad [13]$$

The time-average vector power density gives

$$\langle \vec{S}(x,t) \rangle = \frac{1}{2} |\mathbf{E}_0|^2 \qquad [14]$$

which shows that Poynting's power density is also propagating in the direction of the wave vector $\vec{k}$.

Boundary Conditions

Let us consider a bound obstacle D_i with a smooth closed surface S. The Maxwell eqn [5] has been written in differential form. It must be supplemented with boundary conditions enforced at the obstacle surface. The boundary conditions can be derived from the integral form of Maxwell equations. The field vectors $\mathbf{E}$, $\mathbf{H}$ are assumed to be finite but may be discontinuous across the obstacle surface S. In cases where there is no surface current density (which is so for a perfect conductor) the boundary conditions accept the following form:

$$\vec{n} \times (\mathbf{E}_\mathrm{i} - \mathbf{E}_\mathrm{e}) = 0, \quad \vec{n} \times (\mathbf{H}_\mathrm{i} - \mathbf{H}_\mathrm{e}) = 0 \qquad [15]$$

where subscripts i and e correspond to the total fields inside and outside the obstacle, respectively, and $\vec{n}$ is unit outward normal to S. Essentially the boundary conditions state that tangential components for electromagnetic fields are continuous across the boundary.

Silver–Müller Radiation Conditions

Analysis of a direct scattering problem on a local obstacle involves Maxwell eqn [10] and boundary conditions as in eqn [15] and some infinity conditions, which provide the uniqueness of the scattering problem solutions. Let us assume that a local obstacle is excited by an incident electromagnetic field $\{\mathbf{E}_0, \mathbf{H}_0\}$, given as entire solution to the Maxwell equations. Total field outside the obstacle can be represented as

$$\mathbf{E}_\mathrm{e} = \mathbf{E}_0 + \mathbf{E}_\mathrm{s}, \quad \mathbf{H}_\mathrm{e} = \mathbf{H}_0 + \mathbf{H}_\mathrm{s} \qquad [16]$$

where scattered field $\{\mathbf{E}_\mathrm{s}, \mathbf{H}_\mathrm{s}\}$ in the exterior domain $D_\mathrm{e} = \mathbb{R}^3/\bar{D}_\mathrm{i}$ satisfies the Silver–Müller radiation conditions:

$$\left.\begin{aligned} &\frac{x}{|x|} \times \mathbf{H}_\mathrm{s} + \mathbf{E}_\mathrm{s} = o\left(\frac{1}{|x|}\right) \\ &\frac{x}{|x|} \times \mathbf{E}_\mathrm{s} - \mathbf{H}_\mathrm{s} = o\left(\frac{1}{|x|}\right) \end{aligned}\right\} \quad |x| \to \infty \qquad [17]$$

uniformly for all directions $x/|x|$.

Scattering Amplitude

The following result is valid: every radiating solution $\{\mathbf{E}_s, \mathbf{H}_s\}$ to the Maxwell equations has the following asymptotic form:

$$\begin{aligned}\mathbf{E}_s(x) &= \frac{e^{ik|x|}}{|x|}\left\{\mathbf{E}_\infty(\varpi) + O\left(\frac{1}{|x|^2}\right)\right\} \\ & \qquad\qquad |x| \to \infty \\ \mathbf{H}_s(x) &= \frac{e^{ik|x|}}{|x|}\left\{\mathbf{H}_\infty(\varpi) + O\left(\frac{1}{|x|^2}\right)\right\}\end{aligned} \qquad [18]$$

where the vector functions $\mathbf{E}_\infty$ and $\mathbf{H}_\infty$ defined on the unit sphere $\varpi \in \Omega = \{0 \le \vartheta \le \pi; 0 \le \varphi \le 2\pi\}$ are known as the electric far field pattern (scattering amplitude) and magnetic far field pattern, respectively. They satisfy:

$$\mathbf{H}_\infty = \vec{\nu} \times \mathbf{E}_\infty, \quad \vec{\nu}\cdot\mathbf{E}_\infty = \vec{\nu}\cdot\mathbf{H}_\infty = 0 \qquad [19]$$

with the unit outward normal $\vec{\nu}$ on Ω. The latter means that $\mathbf{E}_\infty$ and $\mathbf{H}_\infty$ have only tangential components at the unit sphere.

Far field pattern represents real value analytic functions on a unit sphere. Reillich's lemma establishes one-to-one correspondence between radiating electromagnetic field and their far field pattern. It means that there exists only one scattered field $\{\mathbf{E}_s, \mathbf{H}_s\}$ corresponding to a given far field pattern $\mathbf{E}_\infty$.

Reciprocity Relations

Examine the set of far field pattern corresponding to the scattering problem including: Maxwell eqn [5], boundary conditions in eqn [15], and radiation conditions in eqn [17] at the infinity for the scattered fields. Let the incident electromagnetic field is given by the planewave:

$$\mathbf{E}_0(x) = \vec{e}_0\, e^{jk\vec{x}\cdot\vec{\kappa}}, \quad \mathbf{H}_0(x) = \vec{\kappa} \times \vec{e}_0\, e^{jk\vec{x}\cdot\vec{\kappa}} \qquad [20]$$

where $\vec{\kappa} \in \Omega$ is a unit vector giving the propagation direction and $\vec{e}_0 \in \mathbb{R}^3$ is a constant vector responsible for the polarization. From the previous consideration one can see that:

$$\begin{aligned}\mathbf{E}_s(x;\vec{\kappa},\vec{e}_0) &= \frac{e^{ik|x|}}{|x|}\left\{\mathbf{E}_\infty(\varpi\vec{\kappa},\vec{e}_0) + O\left(\frac{1}{|x|^2}\right)\right\}, \\ |x| &\to \infty\end{aligned} \qquad [21]$$

Furthermore, the following reciprocity relation for the scattering amplitude holds:

$$\vec{h}_0\cdot\mathbf{E}_\infty(\varpi;\vec{\kappa},\vec{e}_0) = \vec{e}_0\cdot\mathbf{E}_\infty(-\vec{\kappa};-\varpi,\vec{h}_0)$$

for all vectors $\varpi, \vec{\kappa} \in \Omega$ and $\forall \vec{e}_0, \vec{h}_0 \in \mathbb{R}^3$.

Optical Theorem

Far field pattern enables us to evaluate scattering cross section, which is defined as

$$\sigma_s = \int_\Omega |\mathbf{E}_\infty(\varpi)|^2 d\varpi \qquad [22]$$

For the planewave given by eqn [20] and excitation of nonabsorbing obstacle ($\sigma = 0$), the following relation holds:

$$\sigma_s = \frac{4\pi}{k}\mathrm{Im}\{\mathbf{E}_\infty(\vec{\kappa})\cdot\vec{e}_0\} \qquad [23]$$

This is so-called optical theorem.

Huygen's Principle (Green's Theorem). Christian Huygens (1629–1695)

The Huygens' principle shows that a wavefield on the surface determines the wavefield off the surface. Let us introduce the scalar Green's function which satisfies the Helmholtz equation:

$$(\nabla^2 + k^2)g(x,y) = -\delta(x-y) \qquad [24]$$

where $\delta(x-y)$ is the Dirac delta function. For three-dimensional problems, the scalar Green's function $g(x,y)$ for isotropic media accepts the form:

$$g(x,y) = \frac{e^{jk|x-y|}}{4\pi|x-y|} \qquad [25]$$

Let electromagnetic fields $\{\mathbf{E}_s, \mathbf{H}_s\}$ be a radiating solution to the Maxwell eqn [5] in D_e. Then for any $x \in D_e$ we have the Stratton–Chu formulas:

$$\begin{aligned}\mathbf{E}_s(x) &= \nabla\times\int_S g(x,y)\vec{\nu}_y\times\mathbf{E}_s(y)ds_y \\ &\quad + \frac{j}{k}\nabla\times\nabla\times\int_S g(x,y)\vec{\nu}_y\times\mathbf{H}_s(y)ds_y \\ \mathbf{H}_s(x) &= \nabla\times\int_S g(x,y)\vec{\nu}_y\times\mathbf{H}_s(y)ds_y \\ &\quad - \frac{j}{k}\nabla\times\nabla\times\int_S g(x,y)\vec{\nu}_y\times\mathbf{E}_s(y)ds_y\end{aligned} \qquad [26]$$

here S is a smooth closed surface, and $\vec{\nu}$ is the unit normal vector to S directed into D_e. Similar relations are valid for an electromagnetic field inside a bounded domain D_i. The Stratton–Chu formulas are extensively employed to construct boundary integral equations for scattering problems investigating.

Homogeneous Obstacle

Consider a mathematical statement of the scattering problem by a bound homogeneous obstacle. Let external excitation be a planewave $\{\mathbf{E}_0, \mathbf{H}_0\}$: (eqn [20]). In this case the boundary-value scattering problem can be formulated in the form:

$$\begin{gathered}\nabla \times \mathbf{E}_\tau(x) = \mathrm{j}k\mathbf{H}_\tau(x), \quad \nabla \times \mathbf{H}_\tau(x) = -\mathrm{j}k\varepsilon_\tau \mathbf{E}_\tau(x), \\ x \in D_\tau, \tau = \mathrm{i}, \mathrm{e} \\ \vec{\nu} \times (\mathbf{E}_\mathrm{i} + \mathbf{E}_\mathrm{e}) = \vec{\nu} \times \mathbf{E}_0, \\ \vec{\nu} \times (\mathbf{H}_\mathrm{i} + \mathbf{H}_\mathrm{e}) = \vec{\nu} \times \mathbf{H}_0 \ \text{ on } S \\ \frac{x}{|x|} \times \mathbf{H}_\mathrm{e} + \mathbf{E}_\mathrm{e} = o\left(\frac{1}{|x|}\right), \quad |x| \to \infty \end{gathered} \qquad [27]$$

where $\varepsilon_\mathrm{e} = 1$, $\varepsilon_\mathrm{i} = \varepsilon$ and $\{\mathbf{E}_\mathrm{e}, \mathbf{H}_\mathrm{e}\}$ corresponds to scattered field. For a smooth surface and the case $\operatorname{Im} \varepsilon \geq 0$, the problem has a unique solution.

Lippman–Swinger Integral Equation

Assume we deal with the local inhomogeneous obstacle that can be described by dielectric permittivity $\varepsilon(x)$, which is a continuous complex-valued ($\operatorname{Im} \varepsilon(x) \geq 0$) function in D_i. The correspondent scattering problem can be formulated as below in eqn [29]. In this case, the boundary-value scattering problem can be reduced to the solution of the Lippman–Swinger integral equation

$$\begin{aligned}\mathbf{E}_\mathrm{i}(x) = {} & \mathbf{E}_0(x) + \nabla \times \nabla \times \int_{D_\mathrm{i}} g(x, y) k^2 (\varepsilon(y) - 1) \\ & \times \mathbf{E}_\mathrm{i}(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{i} \end{aligned} \qquad [28]$$

Volume integral eqn [28] is equivalent to the corresponding scattering problem. Once the solution is found, the associated scattered field outside the inhomogeneous obstacle can be represented as

$$\begin{aligned}\mathbf{E}_\mathrm{s}(x) = {} & \nabla \times \nabla \times \int_{D_\mathrm{i}} g(x, y) k^2 (\varepsilon(y) - 1) \\ & \times \mathbf{E}_\mathrm{i}(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{e} \end{aligned} \qquad [29]$$

The Lippman–Swinger integral equation plays an important role under analysis of a wave scattering by local inhomogeneous obstacle.

Born Approximation

In the case where $k^2 |\varepsilon(x) - 1| mes D_\mathrm{i} \ll 1$, or where the second term on the right-hand part of the Lippman–Swinger integral eqn [28] is small compared to the first term, one can write $\mathbf{E}_\mathrm{i}(x) \simeq \mathbf{E}_0(x)$. Then the scattered field can be approximately evaluated as

$$\begin{aligned}\mathbf{E}_\mathrm{s}(x) = {} & \nabla \times \nabla \times \int_{D_\mathrm{i}} g(x, y) k^2 (\varepsilon(y) - 1) \\ & \times \mathbf{E}_0(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{e} \end{aligned} \qquad [30]$$

The above representation for the scattered field is known as the first-order Born approximation. It is also the first-order approximation in the Neumann series for the integral eqn [28]. The Born approximation plays an important role in diffraction tomography.

Rayleigh Scattering

Rayleigh scattering characterizes the scattering of electromagnetic waves by particles much smaller then an exciting wavelength. Consider a spherical particle with constant permittivity ε and radius a deposited at the origin of a coordinate system. Let linear polarized planewave be incident upon the particle $\mathbf{E}_0(x) = \vec{e}_x \, e^{\mathrm{j}kz}$, and $\vec{e}_x$ be unit vector Cartesian coordinate system. It is polarized in x-direction and propagates along z-axis. In this case, for the scattered intensity, we get:

$$I_\mathrm{s} = \left(\frac{\varepsilon - 1}{\varepsilon + 2}\right)^2 \frac{k^4 a^6}{r^2} \sin^2 \vartheta \qquad [31]$$

where r is distance from the origin and $\vartheta \in [0, \pi]$.

Corresponding scattering cross-section is calculated as

$$\sigma_\mathrm{s} = \frac{8\pi}{3} \left(\frac{\varepsilon - 1}{\varepsilon + 2}\right)^2 k^4 a^6 \qquad [32]$$

Thus the total scattered power is proportional to the fourth power of the wave number (minus fourth power of the wavelength – Rayleigh's Law). The scattered power is also proportional to the sixth power of the radius.

Mie Scattering

The Mie theory is a complete mathematical–physical theory of the scattering of electromagnetic wave by homogeneous spherical particles, developed by Gustav Mie in 1908. In contrast to Rayleigh scattering, the Mie theory embraces all possible ratios of the particle radius to wavelength. It relays upon the exact solution to Maxwell equation for spherical homogeneous obstacle with planewave excitation. In the frame of Mie theory, the scattered field is represented as an infinite series of fields of spherical multipoles, which are orthogonal at a sphere surface. Mie solution plays an important role in light scattering by clouds of particles.

See also

Scattering: Raman Scattering; Scattering from Surfaces and Thin Films; Scattering Phenomena in Optical Fibers; Stimulated Scattering.

Further Reading

Boheren C and Hoffman D (1983) *Absorption and Scattering of Light by Small Particles*. New York: Wiley.
Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.
Chew WC (1995) *Waves and Fields in Inhomogeneous Media*. New York: IEEE Press.
Colton D and Kress R (1992) *Inverse Acoustic and Electromagnetic Scattering Problems*. New York: Springer-Verlag.
Guenther RD (1990) *Modern Optics*. New York: Wiley.
Kong JA (2000) *Electromagnetic Wave Theory*. Cambridge: EMW Publishing.
Kong JA (2002) *Maxwell Equations*. Cambridge: EMW Publishing.

Stimulated Scattering

M Bashkansky and J Reintjes, Naval Research Laboratory, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Scattering of light involves conversion of an incoming light wave in one mode to an outgoing wave in a different mode. Scattering can involve changes of wavelength, direction, and polarization, and can occur in all types of materials. Without scattering we would not be able to see objects in the world around us. Scattering processes involving particular physical interactions are often given distinctive names. Familiar examples include Rayleigh scattering from the random distribution of air molecules, which makes the sky blue, and Tyndall scattering from small suspended particles, which makes the sunset red. Other types of scattering, e.g., Raman, Brillouin and Rayleigh-wing scattering, are important spectroscopic tools that provide insight into atomic and molecular properties. In their most commonly encountered form, where the intensity of the scattered light is proportional to the intensity of the incident light, the scattering processes are termed spontaneous. With the intense radiation provided by lasers, however, the intensity of the scattered light can grow exponentially, and the processes are termed stimulated scattering.

Stimulated scattering is a subset of the larger discipline of nonlinear optics. The term was first used to describe nonlinear optical interactions (e.g., Raman, Brillouin, Rayleigh) in which the intensity of the scattered light grows exponentially, in a manner similar to the exponential growth of light intensity associated with lasers. These interactions involve exchange of energy with a second mode that is an internal mode of the scattering medium, and the scattered light wave is at a different wavelength from the incoming light. The use of the term stimulated has subsequently been expanded to include a wider class of nonlinear optical interactions, such as parametric fluorescence or certain aspects of self-focusing, in which the second mode is itself a light wave, with either a different frequency or direction of propagation.

In the wider sense, scattering of light occurs through interaction of the incoming light wave with an ‘idler’ mode in a material. Stimulated scattering occurs when the idler mode is itself driven by interference of the incoming and scattered light waves. Under these circumstances, exponential growth of the scattered wave results. Depending on the interaction, the idler mode may be an internal mode of the material or another light wave. Most linear scattering processes have stimulated counterparts. However, there are stimulated scattering interactions that have no spontaneous analogs.

The elements of stimulated light scattering can be summarized mathematically as follows. The scattering process is modeled as an incoming laser or pump, an outgoing scattered wave and an idler mode. The interaction of the laser with the idler mode to produce the scattered light wave is described by the equation

$$\frac{\partial}{\partial z}A_{\mathrm{S}}(z,t) + \frac{n}{c}\frac{\partial}{\partial t}A_{\mathrm{S}}(z,t) = \kappa A_{\mathrm{L}} b^{*} \qquad [1]$$

where A_{S} and A_{L} are the slowly varying amplitudes of the scattered and laser light waves whose electric fields are defined as

$$E_{\mathrm{S(L)}} = \frac{1}{2}\left\{A_{\mathrm{S(L)}}\mathrm{e}^{\mathrm{i}[\omega_{\mathrm{S(L)}}t - k_{\mathrm{S(L)}}z]} + \mathrm{cc}\right\} \qquad [2]$$

$\omega_{S(L)}$ is the frequency of the scattered (laser) light, $k_{S(L)}$ is the corresponding wavevector, κ is a coupling coefficient, cc denotes complex conjugate, and b is the amplitude of the idler mode given by

$$B = \frac{1}{2}\left\{b e^{i[\omega_b t - k_b z]} + cc\right\} \qquad [3]$$

The interaction of the laser and scattered wave to drive the idler mode is described by the equation

$$\mathcal{L}(b^*) = \kappa' A_L^* A_S \qquad [4]$$

where $\mathcal{L}$ is a differential operator that describes the dynamics of the idler mode, e.g., a damped oscillator or a propagating wave, and κ' is a second coupling coefficient. For the class of stimulated processes that involve excitation of an internal mode of the medium, $\mathcal{L}$ has the form

$$\mathcal{L} = \frac{\partial}{\partial t} + \frac{1}{T_2} - v\frac{\partial}{\partial z} \qquad [5]$$

where T_2 is the damping time of the coherence of the internal mode. When interaction times are long compared to internal response and decay times,

$$b^* \sim \kappa' A_L^* A_S$$

and the scattered wave obeys an equation of the form

$$\frac{\partial}{\partial z} A_S(z,t) = \frac{g}{2}|A_L|^2 A_S(z,t)$$

which has an exponential-gain solution for the scattered wave intensity:

$$I_S(z,t) = I_S(0,t)e^{I_L g z}$$

When the idler mode is a light wave, as in stimulated parametric fluorescence, [4] has the same form as [1] and they can be solved as a coupled pair, leading to transcendental or hyperbolic solutions. Within the framework of nonlinear optics, interactions of these types can be expressed as arising from a nonlinear optical polarization, which forms the right-hand side of [1]. The dynamics of the nonlinear polarization are then described by [4].

Stimulated processes are often associated with pumping thresholds. When these processes are configured with cavity resonators, they possess a threshold in pumping intensity just as with lasers or other oscillators. When they are operated in single or multipass scattering cells, they have no true threshold. However, they still have a practical useful threshold when the growth from quantum noise raises the scattered signal to a level comparable to the pump. By convention, this is taken to be the incident light intensity at which the scattered intensity reaches approximately 1% of the pump intensity, typically involving gains of the order of e^{23}–e^{30}. At this level, a small increase in pump intensity raises the scattered intensity to levels at which the pump starts to be depleted.

Attenuation of the scattered light can be accounted for by replacing $I_L g$ in the above expression by $I_L g - \alpha$, where α is the attenuation coefficient. Exponential gain is now experienced when the pump intensity exceeds the level given by

$$I_L > \frac{\alpha}{g}$$

This value can serve as an estimate of threshold pumping in cavity configurations, although in free propagation situations, net gain must still approach e^{23} to exhibit threshold behavior.

A list of stimulated scattering interactions is given in Table 1.

Stimulated Raman Scattering

Stimulated Raman scattering (SRS) involves conversion of an incoming light wave at one frequency, termed the laser or pump wave, into a scattered wave of different frequency, along with excitation or de-excitation of an internal mode of the medium. The frequency difference between the pump and scattered waves is equal to the frequency of the material mode and is characteristic of the material. When the frequency of the scattered wave is lower than that of the incident wave, the scattered wave is termed a Stokes wave, and the internal mode of the medium is excited in the scattering process. When the frequency of the scattered wave is higher than that of the incident wave, the scattered wave is termed an anti-Stokes wave, and the internal mode of the medium is de-excited in the scattering process.

The internal mode of the medium that is involved in Raman scattering is a nonpropagating collective mode. Many different types of internal modes in all types of materials can be involved in Raman scattering. Examples include molecular vibrations and rotations in gases, liquids, or solids, electronic states that are of the same parity as the ground state in gases or solids, optical phonons or polaritons in solids, and Langmuir waves in plasmas. Raman frequency shifts can range from tens of wavenumbers for lattice vibrations to tens of thousands of wavenumbers for electronic transitions.

Spontaneous Raman scattering was first described theoretically by Smekal in 1923 from a quantum mechanical study of light scattering, and was

Table 1 List of stimulated scattering interactions

Stimulated scattering process	*Internal material mode*	*Light mode change*
Stimulated Raman scattering	Nonpropagating internal energy mode: molecular vibrations and rotations, same parity electronic states, optical phonons and Langmuir waves	Change in wavelength, polarization, direction
Stimulated Brillouin scattering	Propagating sound waves	Change in wavelength, polarization, and direction
Stimulated Rayleigh scattering	Temperature and density fluctuations due to light absorption	Change in direction
Stimulated Rayleigh wing scattering	Molecular distribution and orientation fluctuations	Change in wavelength and direction
Stimulated concentration scattering	Fluctuations of concentration in gas mixtures	Change in direction
Photorefractive stimulated scattering	Diffusion of photo-excited carriers	Change in direction
Stimulated Thomson/Compton scattering	Momentum changes in bunched free electrons	Change in wavelength and direction
Stimulated parametric fluorescence	Nonresonant nonlinear polarization	Change in wavelength and direction
Self-focusing	Nonresonant nonlinear polarization	Change in direction

demonstrated experimentally by Raman, and shortly after by Landsberg and Mandelstamm in 1928. Stimulated Raman scattering was first observed in 1962 by Woodbury and Ng, and Eckhardt *et al.* soon after the development of pulsed lasers.

Stimulated Raman scattering can be observed in many different configurations including cavity resonators, single or multiple-pass generator cells or amplifiers. It can be observed as forward scattering, with the scattered light propagating in the same direction as the pump light, or backward scattering, with the scattered light propagating in the opposite direction to the pump light. Its most widespread use is for generation of coherent light at frequencies different from the pump light. Other applications include phase conjugation, optical gating, beam cleanup, and pulse compression.

The equations describing SRS when depletion of the pump light and diffraction of the optical beams are neglected are:

$$2ik_S\left(\frac{\partial A_S}{\partial z} \pm \frac{1}{v_{gS}}\frac{\partial A_S}{\partial t}\right) = -\frac{N}{2}\frac{\omega_S^2}{c^2}\left(\frac{\partial \alpha}{\partial Q}\right)_S Q^* A_L \qquad [6]$$

$$\frac{\partial Q^*}{\partial t} + \Gamma Q^* = -\frac{i}{4\omega_o m}\left(\frac{\partial \alpha}{\partial Q}\right)_S A_L^* A_S \qquad [7]$$

Here A_S and A_L are the slowly varying optical field amplitudes of the Stokes and laser waves as given in [2], $\omega_{S(L)}$ are the optical frequencies, $k_{S(L)}$ are the corresponding k vectors, $v_{S(L)}$ are the group velocities, and Q is the amplitude of the idler mode, in this case the normal mode coordinate of the material excitation, defined by:

$$\mathrm{Q}(x,y,z,t) = \frac{1}{2}\left(Q(x,y,z,t)e^{-i(\omega_o t - k_o z)} + \mathrm{cc}\right) \qquad [8]$$

$(\partial\alpha/\partial Q)_S$ is the Stokes hyperpolarizability, $\omega_0 = \omega_L - \omega_S$ is the Raman transition frequency, and k_0 is the k vector of the material excitation. The phase matching relation $k_S = k_L - k_0$ is automatically satisfied for SRS because the material mode that is involved is nonpropagating. Γ is the half width at half maximum of the Raman linewidth given by $\Gamma = 1/\mathrm{T}_2$, where T_2 is the dephasing time. N is the number of molecules in the ground state and m is the effective reduced mass of the material oscillation.

When the fields are constant or vary slowly compared to the response time of the material, the Stokes intensity is given by

$$I_S(z) = I_S(0)e^{I_L g_{ss} Z} \qquad [9]$$

where the intensity I is given by

$$I_{S(L)} = \frac{1}{2}(cn_{S(L)}\varepsilon_0)|A_{S(L)}|^2$$

and

$$g_{ss} = \frac{N\omega_s}{4\Gamma n_S n_L m \omega_0 c^2 \varepsilon_0}\left(\frac{\partial \alpha}{\partial Q}\right)^2 \qquad [10]$$

where $n_{S(L)}$ is the index of refraction of the Stokes (laser) in the material and ε_0 is the dielectric permittivity of the vacuum.

The Stokes wave grows exponentially with z at a rate that depends on the pump intensity. If the Stokes wave becomes intense enough, its power can become comparable to that of the original pump wave. In that case depletion of the pump wave must be taken into account with an equation for the pump wave amplitude similar to [6]. The Stokes intensity is now given by

$$I_S(z) = \frac{I_S(0)e^{I_L(0)gz}}{1 + \dfrac{\omega_L}{\omega_S}\dfrac{I_S(0)}{I_L(0)}e^{I_L(0)gz}} \qquad [11]$$

which is valid when $I_S(0) \ll I_L(0)$.

In the limit of strong interactions all of the laser energy can be depleted. The maximum Stokes energy is given by $I_S = I_L(\omega_S/\omega_L)$, where ω_S/ω_L is termed the Manly–Rowe ratio. The energy that is removed from the laser but not converted to Stokes radiation is transferred to the material.

For each photon produced in the scattered wave, a corresponding quantum of energy is given to the material mode and a photon is removed from the laser. In principle, 100% of the laser photons can be converted to Stokes photons, but complete conversion is not achieved in practice. Photon conversion efficiencies greater than 90% have been reported in some pulsed laser experiments.

SRS frequency shifts and gains of some materials are shown in Table 2.

When pulses are used with durations comparable to or shorter than the response time of the material, ranging from picoseconds to femtoseconds depending on the material, the interaction is transient and the effects of the time derivatives in [6] and [7] must be accounted for. Under these conditions the Stokes intensity is given as a temporal integral with Bessel function dependence on the laser intensity. When the pulse duration is much shorter than the material dephasing time and the incident Stokes functional form is the same as the pump, the amplified Stokes intensity is given by the expression

$$I_S(z,\tau) = I_S(0,\tau)I_0^2\left(\sqrt{2g_{ss}\Gamma z\int_{-\infty}^{\tau} I_L(\tau')\mathrm{d}\tau'}\right) \qquad [12]$$

where I_0 is a modified Bessel function. In this regime, the Stokes intensity depends only on the integrated pump energy. The Stokes gain is reduced from its steady state value for pulses of constant intensity, but for pulses of constant energy the highest Raman gain is obtained in the limit of transient interactions.

One of the most important applications of Raman scattering is the generation of new coherent waves at frequencies shifted from the pump. In this case, an incident Stokes wave is not supplied as the starting term in [9] or [12]. Rather, the process is initiated by the vacuum fluctuations in the Stokes and material oscillator fields. Quantum field creation and annihilation operators must replace the classical field amplitudes of [6] and [7]. The intensity of the generated Stokes wave is given by the expectation value of the normally ordered number operator and in the extreme transient regime is given by:

$$\begin{aligned} I_S(z,\mathrm{d}\tau) = {} & \frac{1}{2A}\hbar\omega_S\Gamma g_{ss}I_L(\mathrm{d}\tau)z \\ & \times\left\{I_0^2\left(\sqrt{2\Gamma g_{ss}z\int_{-\infty}^{\tau} I_L(\tau')\tau'}\right)\right. \\ & \left. - I_1^2\left(\sqrt{2\Gamma g_{ss}z\int_{-\infty}^{\tau} I_L(\tau')\tau'}\right)\right\} \end{aligned} \qquad [13]$$

where A is the cross-sectional area of the interaction geometry. The Stokes light generated in this fashion has all the characteristics of amplified spontaneous emission of lasers operated below threshold until the Stokes intensity becomes high enough to deplete the laser.

The stochastic nature of the SRS quantum mechanical source is manifest in the macroscopic Stokes light when the Stokes generator is operated well below pump depletion levels. For a single spatial mode the probability density distribution of

Table 2 List of frequency shifts and gains of selected materials for stimulated Raman scattering

Material, λ_L (nm)	*Shift* (cm^{-1})	*Gain (cm/GW)*
H_2 gas (20 atm), 532	4155	2.5
D_2 gas (60 atm), 532	2987	0.45
CH_4 (115 atm), 532	2913	1.26
H_2O, 694	3290	0.14
CS_2, 694	655.6	24
$LiNbO_3$, 694	637	9.4
$Ba(NO_3)_2$, 532	1047	47

Table compiled in parts from: Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman Spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2. pp. 1077–1150. Amsterdam, The Netherlands: North Holland and Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin Scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV, New York: McGraw-Hill.

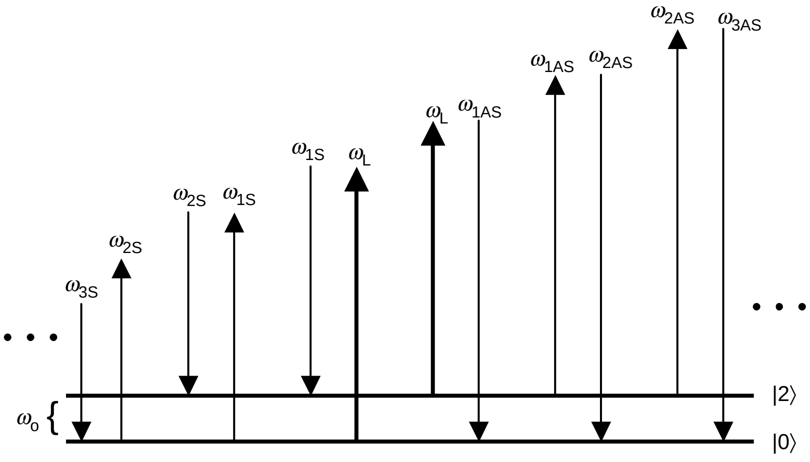

Figure 1 SRS level diagram showing generation of multiple Stokes and anti-Stokes frequencies.

the Stokes energy has the form of a negative exponential:

$$P(W) = \frac{1}{\langle W \rangle} \exp\left[-\frac{W}{\langle W \rangle} \right]$$

Statistical fluctuations in pulse energy, spatial profile, pointing, and in the spectral and temporal structure of the Stokes signal have all been observed.

If the intensity of the generated Stokes light becomes comparable to the pump intensity, it can serve as a pump wave for its own Raman interaction, producing a wave termed the second Stokes wave, shifted from the pump by frequency $2\omega_o$. If the process is driven hard enough third- and higher-order Stokes waves can be produced.

In addition to the interaction between the pump and Stokes beams described above, stimulated Raman scattering can also involve an anti-Stokes wave whose frequency is given by $\omega_{AS} = \omega_L + \omega_o$. The relationship of various Stokes and anti-Stokes waves is depicted in Figure 1. If a population inversion exists between the energy levels of the Raman process, then the anti-Stokes wave experiences exponential gain in a manner similar to that described above for the Stokes gain.

When all the population is in the ground state, which is usually the case, the anti-Stokes wave has exponential loss. The anti-Stokes wave can still grow without population inversion if the phase-matching condition

$$\Delta k = k_{AS} + k_S - 2k_L = 0$$

is approximately satisfied. Because of normal dispersion in the material this condition is usually not satisfied for collinear beams. However, if the Stokes and anti-Stokes beams propagate at small angles as shown in Figure 2, then the phase-matching condition can be met. For angles near but not exactly at phase matching the anti-Stokes wave grows exponentially as part of a mixed Stokes/anti-Stokes mode with a gain given by:

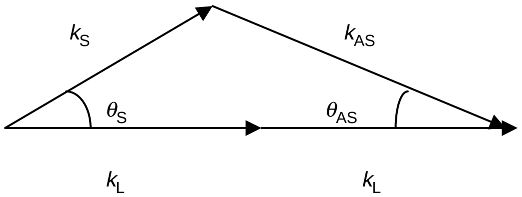

Figure 2 Phase matching diagram for Stokes–anti-Stokes coupling in a medium with positive dispersion.

$$g = \mathrm{Re}\left\{ \tfrac{1}{2}(K_3 - K_1)|A_L|^2 - \tfrac{i}{2}\sqrt{\Delta k^2 + 2i\Delta k(K_3 + K_1)|A_L|^2 - (K_1 - K_3)^2 |A_L|^4} \right\} \quad [14]$$

At exact phase matching the interaction of Stokes and anti-Stokes waves serves to suppress the exponential gain of either wave.

As with Stokes waves, multiple-order anti-Stokes waves can be generated. Stokes and anti-Stokes orders up to at least 13 have been reported.

Stimulated Brillouin and Rayleigh Scattering

Light scattering can also occur through interaction with material density variations. Spontaneous scattering through interaction with this type of internal mode produces a triplet of spectral lines. The up- and down-shifted lines are known as Brillouin anti-Stokes and Stokes lines, respectively, and are due to interaction with acoustic waves. The central peak is termed Rayleigh, or thermal Rayleigh scattering and is due to stationary density fluctuations arising from light absorption. A broad skirt around the central peak, which can be observed in anisotropic media, is due to the fluctuation of molecular distribution and orientation and is termed Rayleigh wing scattering. Each of these processes can give rise to stimulated scattering in the presence of intense radiation.

Density fluctuations are induced directly by light intensity through electrostriction, and indirectly by temperature fluctuations when optical absorption is important. A general classical model of the interaction, including both electrostriction and absorptive heating effects, is described by three equations:

The Navier-Stokes equation:

$$\rho_0\frac{\partial \mathbf{v}}{\partial t}+\frac{v^2}{\delta}\nabla(\Delta\rho)+\frac{v^2\beta\rho_0}{\delta}\nabla(\Delta T)-\eta\nabla^2\mathbf{v} = 2\gamma\varepsilon_0\nabla(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})-2\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})\nabla(\Delta T) \qquad [15]$$

The continuity equation:

$$\frac{\partial}{\partial t}\Delta\rho+\rho_0\nabla\cdot\mathbf{v}=0 \qquad [16]$$

The energy transport equation:

$$\left(\rho_0 C_v\frac{\partial}{\partial t}-\lambda_\mathrm{T}\nabla^2\right)\Delta T-\frac{C_v(\delta-1)}{\beta}\frac{\partial}{\partial t}\Delta\rho = 4nc\alpha\varepsilon_0(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})+4\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho\left(T_0\mathbf{E}_\mathrm{L}\cdot\frac{\partial}{\partial t}\mathbf{E}^*_\mathrm{R/B}\right) \qquad [17]$$

Here $\rho=\rho_0+\Delta\rho$, $T=T_0+\Delta T$, $\mathbf{v}$ is the volume element, v is the acoustic wave velocity, $\delta=C_p/C_v$ is the ratio of heat capacities at constant pressure and at constant volume, $\eta=(2\eta_\mathrm{s}+\eta_\mathrm{d})$ where η_s is the shear viscosity and η_d is the dilational viscosity, $\gamma=\rho_0(\partial\varepsilon/\partial\rho)_T$, λ_T and β are the thermal conductivity and thermal expansion coefficients, ε is the dielectric constant, and α is the linear absorption coefficient. $\mathbf{E}_\mathrm{L}$ is the pump field and $\mathbf{E}_\mathrm{R/B}$ is either the Rayleigh or Brillouin field.

Equations [15] and [16] can be combined to eliminate the $\mathbf{v}$ dependence:

$$\left[-\frac{\partial^2}{\partial t^2}+\left(\frac{v^2}{\delta}+\frac{\eta}{\rho_0}\frac{\partial}{\partial t}\right)\nabla^2\right]\Delta\rho+\frac{v^2\beta\rho_0}{\delta}\nabla^2(\Delta T) = 2\gamma\varepsilon_0\nabla^2(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})-2\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho\nabla\cdot[(\mathbf{E}_\mathrm{L}\cdot\mathbf{E}^*_\mathrm{R/B})\nabla(\Delta T)] \qquad [18]$$

Together with the equations describing the effect of the medium on the optical fields [19a,b], this coupled set of equations describes simultaneously stimulated Brillouin and Rayleigh scattering:

$$\left(\frac{\partial}{\partial z}+\frac{1}{v_\mathrm{g}}\frac{\partial}{\partial t}+\frac{1}{2}\alpha\right)A_\mathrm{L} = \mathrm{i}\frac{\omega_\mathrm{L}}{4cn_\mathrm{L}}\left[\frac{\gamma}{\rho_0}\Delta\rho A_\mathrm{R/B}+\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho A_\mathrm{R/B}\Delta T\right] \qquad [19a]$$

$$\left(\frac{\partial}{\partial z}-\frac{1}{v_\mathrm{g}}\frac{\partial}{\partial t}-\frac{1}{2}\alpha\right)A_\mathrm{R/B} = -\mathrm{i}\frac{\omega_\mathrm{R/B}}{4cn_\mathrm{R/B}}\left[\frac{\gamma}{\rho_0}\Delta\rho^* A_\mathrm{L}+\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho A_\mathrm{L}\Delta T^*\right] \qquad [19b]$$

where the slowly varying envelope approximation [2] is used for the waves described by A_L, $A_\mathrm{R/B}$, $\Delta\rho$, and ΔT. The nonlinear polarization term is related to density and temperature wave amplitudes in the following way:

$$P^\mathrm{NL}=\varepsilon_0 E\left[\left(\frac{\partial\varepsilon}{\partial\rho}\right)_T\rho+\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho T\right] \qquad [20]$$

Without the driving terms involving the optical fields, [17] and [18] describe broadband damped fluctuations in the variables $\Delta\rho$ and ΔT. When the optical driving terms are included, resonant behavior emerges. Steady state solutions can be obtained assuming strongly damped material excitations, allowing neglect of $\Delta\rho$ and ΔT derivatives. Separate contributions of the Rayleigh and Brillouin interactions to $\Delta\rho$ can be identified:

Rayleigh contribution:

$$\Delta\rho_\mathrm{R}=\left[\frac{\gamma(\Omega+\frac{1}{2}\mathrm{i}\delta\Gamma_\mathrm{R})-\mathrm{i}\gamma_\mathrm{a}\Omega_\mathrm{B}}{\Omega+\frac{1}{2}\mathrm{i}\Gamma_\mathrm{R}}\right]\frac{\varepsilon_0}{4v^2}A_\mathrm{L}A^*_\mathrm{R} \qquad [21]$$

where γ_a is the absorptive coupling constant given by

$$\gamma_\mathrm{a}=\frac{2\alpha n v^2\beta}{C_p\Omega_\mathrm{B}}$$

The Brillouin contribution is:

$$\Delta\rho_{\mathrm{B}} = \left[\frac{-(\gamma - \mathrm{i}\gamma_a vq/\Omega)q^2\varepsilon_0}{4(\Omega^2 + \mathrm{i}\Omega\Gamma_{\mathrm{B}} - v^2q^2)}\right]A_{\mathrm{L}}A_{\mathrm{B}}^* \qquad [22]$$

In [21] and [22] the frequencies and wavevectors satisfy the following relations:

$$\omega_{\mathrm{L}} = \omega_{\mathrm{R/B}} + \Omega \qquad [23a]$$

$$k_{\mathrm{L}} = k_{\mathrm{R/B}} + q \qquad [23b]$$

where ω_{L} is the incident laser frequency, $\omega_{\mathrm{R/B}}$ is the frequency of the scattered Rayleigh or Brillouin wave, and k_{L}, $k_{\mathrm{R/B}}$ are the corresponding wavevectors. Ω and q are the frequency and wavevector of the density fluctuation. Equations [23a,b] can also be described as energy and momentum conservation relations among incident and scattered photons and a phonon in the quantum mechanical picture.

The Rayleigh and Brillouin contributions exhibit resonant behavior at different frequency shifts. The Rayleigh contribution has its resonant frequency at $\Omega_{\mathrm{R}} = 0$, resulting in typical diffusion behavior. The Rayleigh linewidth $\Gamma_{\mathrm{R}} = 2\lambda_T q^2/\rho_0 C_p$ is the inverse of the decay time of the isobaric density fluctuations. The Brillouin contribution has its resonance at a frequency shift given by

$$\Omega_{\mathrm{B}} = qv$$

where v is the sound velocity. Thus at resonance the density fluctuations associated with the Brillouin effect correspond to sound waves. The Brillouin linewidth

$$\Gamma_{\mathrm{B}} = \frac{q^2}{\rho_0}\left[\eta + \frac{\kappa}{C_p}\left(\frac{C_p}{C_v} - 1\right)\right]$$

is the inverse of the phonon lifetime τ_{B}.

In general Brillouin and Rayleigh interactions involve noncollinear scattering with the outgoing wave propagating at an arbitrary angle to the incident wave. In this case [23b] is a vector relation and the interaction geometry can be represented as shown in Figure 3.

The resulting relations among the k-vectors and frequencies of the various waves can place restrictions on the values of frequency shifts that are observed under various circumstances.

The coupled equations for the optical intensities in the steady state take the following form:

$$\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}z} = -g_{\mathrm{R/B}}I_{\mathrm{L}}I_{\mathrm{R/B}} - \alpha I_{\mathrm{L}} \qquad [24a]$$

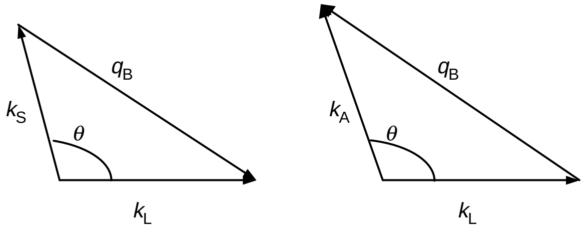

Figure 3 SBS phase matching diagram depicting relations between laser, k_{L}, sound wave, q_{B}, and Stokes and anti-Stokes, k_{S}, k_{A} k-vectors.

$$\frac{\mathrm{d}I_{\mathrm{R/B}}}{\mathrm{d}\zeta} = g_{\mathrm{R/B}}I_{\mathrm{L}}I_{\mathrm{R/B}} - \alpha I_{\mathrm{R/B}} \qquad [24b]$$

where z is the direction of propagation of the incoming wave, ζ is the direction of propagation of the scattered wave, and $g_{\mathrm{R/B}}$ is the gain coefficient of the Rayleigh or Brillouin effect. Equations [24a,b] describe interactions that have exponential gain. Stimulated Brillouin and Rayleigh scattering can be used in both a self-generator and an amplifier mode, but are used most often in a self-generator mode. Because the frequency shifts are small in these interactions, the starting signals for self-generators are thermal excitations of the internal modes rather than vacuum fluctuations.

Stimulated Brillouin Scattering

Stimulated Brillouin scattering (SBS) occurs through interaction with acoustic waves in solids, liquids, and gases and with ion-acoustic waves in plasmas. The Stokes or anti-Stokes frequency shifts for SBS are much smaller than for SRS, with typical values on the order of 0.1–100 GHz, depending on the excitation wavelength and interaction geometry as well as on material properties. Because the internal mode in SBS is a propagating wave, the phase matching condition shown in Figure 3 places restrictions on the frequency of the sound wave that is involved at various scattering angles. SBS is used most often for pulse compression and for correction of phase distortion in propagating light beams through phase conjugation. Although SBS is normally associated with high-power lasers in laboratory experiments, it is also a major factor in limiting the intensity of low-power lasers that can be used in optical fibers because of the long interaction lengths. SBS is also a prominent source of damage in solids and crystals caused by high-power lasers. It can also be used for generation and study of coherent hypersonic waves.

The SBS gain coefficient has both electrostrictive and absorptive contributions, $g_B = g_B^e + g_B^a$ where

$$g_B^e = \frac{\omega_B \gamma^2 \Omega_B}{2n_B n_L c^2 v^2 \rho_0 \Gamma_B} \frac{1}{1 + (2(\Omega - \Omega_B)/\Gamma_B)^2} \qquad [25]$$

and

$$g_B^a = \frac{\omega_B \gamma \gamma_a \Omega_B}{4n_B n_L \rho_0 v^2 c^2 \Gamma_B} \frac{4(\Omega - \Omega_B)/\Gamma_B}{1 + (2(\Omega - \Omega_B)/\Gamma_B)^2} \qquad [26]$$

The SBS gain from electrostriction has a maximum value of

$$g_B^e = \frac{\omega_B \gamma^2 \Omega_B}{2n_B n_L c^2 v^2 \rho_0 \Gamma_B} \qquad [27]$$

at the resonant frequency shift

$$\Omega = \Omega_B = qv \qquad [28]$$

For this condition the interaction is with a sound wave in the material. The frequency of the sound wave, and the gain, depend on the angle between the incident and scattered light. Using the vector relation

$$\mathbf{k}_B = \mathbf{k}_L - \mathbf{q}$$

depicted in Figure 3 and the approximations $\Omega_B << \omega_B, \omega_L, |k_B| \approx |k_L|$ we obtain

$$q = 2|k_L|\sin(\theta/2) \qquad [29]$$

The Brillouin frequency shift is then given by

$$\Omega_B = 2n_L \omega_L (v/c)\sin(\theta/2) \qquad [30]$$

and the Brillouin linewidth is given by

$$\Gamma_B = \frac{4}{\rho_0}\left[\eta + \frac{\kappa}{C_p}\left(\frac{C_p}{C_v} - 1\right)\right]\frac{\omega_L^2 n_L^2}{c^2}\sin^2(\theta/2)$$
$$= \Gamma'' \sin^2(\theta/2) \qquad [31]$$

The peak Brillouin electrostrictive gain is then given by

$$g_B^e = \frac{\omega_B \gamma^2 \omega_L}{n_B c^3 v \rho_0 \Gamma_B}\sin(\theta/2) \qquad [32a]$$

$$g_B^e = \frac{\omega^2 \gamma^2}{n_B c^3 v \rho_0 \Gamma'' \sin(\theta/2)} \qquad [32b]$$

where we have used the approximation $\omega_B \approx \omega_L = \omega$.

The maximum electrostrictive SBS gain is independent of the laser frequency because the Brillouin linewidth is proportional to the square of the laser frequency.

As a function of angle the maximum Brillouin gain is usually observed in the backward direction. This is consistent with the form of the gain coefficient given in [32a], which indicates that the gain is proportional to $\sin(\theta/2)$. However, when the angular variation of the Brillouin linewidth is also accounted for, the gain coefficient is actually minimum in the backward direction [32b]. Typical laboratory experiments involve pencil-like geometries in which the interaction length is significant only in the backward or forward directions. In these situations, the total gain is actually maximum in the backward direction, where the interaction length is longest. Although the formulas appear to indicate that the gain goes to infinity in the forward direction, the damping time, which is proportional to Γ_B^{-1}, also goes to infinity and the interaction becomes transient. Under these conditions the gain parameter is proportional to $\sin(\theta/2)$ and the Brillouin gain goes to zero in the forward direction. Under some geometries Brillouin scattering at 90° to the laser can also be significant. It can be a source of damage to large glass components in high-power lasers.

For thermal Brillouin scattering the gain has a dispersive shape centered on the frequency shift $\Omega = \Omega_B$, with loss occurring on the low-frequency side and gain occurring on the high-frequency side, peaking at a frequency $\Omega = \Omega_B + \Gamma_B/2$. The maximum absorptive SBS gain depends on the laser frequency as ω^{-1}.

Table 3 shows representative values of the Brillouin linewidth and gain coefficients for a selection of substances.

Brillouin-enhanced four-wave mixing is a related effect that involves both a downshifted wave and an upshifted wave in which the coherent interaction is enhanced by the resonance associated with the sound wave. Gains of the order of 10^5 can be realized. Systems based on this effect are often used for phase conjugation.

Stimulated Rayleigh Scattering

Stimulated thermal Rayleigh scattering (R) can be observed in liquids while Stimulated Rayleigh wing scattering (SRWS) is observed in liquids with anisotropic molecules. Additionally, stimulated Rayleigh scattering due to fluctuations of concentration in gas mixtures can also occur. For R the internal mode obeys a diffusion equation. The R gain is given by

$$g_R = g_R^e + g_R^a$$

Table 3 Properties of stimulated Brillouin scattering for selected materials

Substance	*Laser wavelength (nm)*	*Frequency shift (Ghz)*	*$\Delta\nu$ (MHz)*	τ_B (ns)	*g_B^e (cm/GW)*	*g_B^a/α* (cm/GW)*
Liquid						
Acetone	532	5.93	361	0.44	12.9	22
Benzene	532	8.33	515	0.31	12.3	24
CS_2	532	7.7	120	1.9	130	20
CCl_4	532	5.72	890	0.18	8.77	13
Chloroform	532	5.75	635	0.25	11.7	
Ethanol	532	5.91	546	0.29	12*	10
Methanol	532	5.47	325	0.49	10.6	13
Water	532	7.4	607	0.26	2.94	0.8
Gas						
Xenon (7599 torr)	532	0.654 ± 0.024	98.1 ± 8.9	$0.65\lambda_P^2P$	1.38 ± 0.19	
SF_6(20 bar)	1320	0.2		35		
N_2(100 bar)	1320	0.5		30		
Solid						
BK 7	532	34.65 ± 0.039	165.0 ± 8.6		2.15 ± 0.21	
CaF_2	532	37.164 ± 1.185	45.6 ± 8.8		4.11 ± 0.65	
Plexiglas	532	15.687 ± 0.036	253.7 ± 12.6			
SiO_2	488	35.6	156		4.482	

*Absorptive maximum gain SBS values at 694 nm, with permission from Boyd RW (1992) *Nonlinear Optics*. Academic Press: New York. Reproduced from Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV. New York: McGraw-Hill.

where g_R^e and g_R^a are the electrostrictive and absorptive contributions given by

$$g_R^e = \frac{\omega\gamma^2(\delta - 1)}{4\rho_0 n_L n_R v^2 c^2}\,\frac{4\Omega/\Gamma_R}{1 + (2\Omega/\Gamma_R)^2} \qquad [33]$$

$$g_R^a = \frac{-\omega\gamma\gamma_a\Omega_B}{2\rho_0 n^2 v^2 c^2 \Gamma_R}\,\frac{4\Omega/\Gamma_R}{1 + (2\Omega/\Gamma_R)^2} \qquad [34]$$

Both contributions to the STRS gain have dispersive forms. The electrostrictive contribution is maximum for a down-shifted wave with $\Omega = \frac{1}{2}\Gamma_R$. Its maximum value scales linearly with laser frequency. For the absorptive process the gain depends inversely on the laser frequency, and is maximum for an upshifted wave with $\Omega = -\frac{1}{2}\Gamma_R$. Table 4 shows representative values of the Rayleigh linewidth and gain coefficients for a selection of substances.

Table 4 Properties of stimulated Rayleigh scattering for some materials at a wavelength of 694 nm

Substance	*Calculated maximum steady state gain factor*		*Linewidth $\delta\nu_R$ (MHz)*
	g_R^e(max) (cm/MW)	*g_R(max)/α (cm²/MW)*	
CCl_4	2.6×10^{-4}	0.82	17
Methanol	8.4×10^{-4}	0.32	20
CS_2	6.0×10^{-4}	0.62	36
Benzene	2.2×10^{-4}	0.57	24
Aceton	2.0×10^{-4}	0.47	21
H_2O	0.02×10^{-4}	0.019	27.5
Ethanol		0.38	18

Reproduced from Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman Spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2, pp. 1077–1150. Amsterdam, The Netherlands: North Holland.

Parametric Fluorescence

Some types of stimulated interactions do not involve exchange of energy with the material. These include parametric down-conversion and self-focusing. In parametric down-conversion an incident wave at frequency ω_L scatters into two longer-frequency light waves at signal and idler frequencies ω_s and ω_i. They satisfy the frequency and wavevector relations

$$\omega_L = \omega_s + \omega_i$$
$$k_L = k_s + k_i$$

The equations for the waves are:

$$\frac{\partial}{\partial z}A_s = -i\left(\frac{\omega_s}{2n_s c}\right)\chi^{(2)}A_L A_i^*$$

$$\frac{\partial}{\partial z}A_i = -i\frac{\omega_i}{2n_i c}\chi^{(2)}A_L A_s^*$$

where A_s and A_i are called the signal and idler fields. The solution for the signal field takes the following form:

$$A_s(z) = A_s(0)\cosh\left(\sqrt{|\kappa|^2|A_L|^2}z\right) - i\sqrt{\frac{\omega_s n_i}{\omega_i n_s}}A_i^*(0)\sinh\left(\sqrt{|\kappa|^2|A_L|^2}z\right) \quad [35]$$

where

$$|\kappa|^2 = (\omega_s\omega_i/4n_s n_i c^2)|\chi^{(2)}|^2$$

The stimulated parametric interaction provides gain for input signal and idler waves. For large values of $\sqrt{|\kappa|^2 A_L|^2}\ z$ the gain becomes exponential in character, similar to the gain in SRS and SBS. When no incident signals are supplied, the source terms in [35] are the vacuum fluctuations at the signal and idler frequencies. Stimulated parametric fluorescence is the basis for parametric oscillators that provide tunable coherent radiation over a wide wavelength range in the visible and infrared.

Stimulated Photorefractive Scattering

Stimulated light scattering also takes place in photorefractive materials. These materials include crystals such as $LiNbO_3$, $BaTiO_3$, $Bi_{12}SiO_{20}$, and others. They possess a unique property of photoinduced index change with very low light intensities. Some of the potential applications include real-time holography, information recording, phase conjugation, parallel signal processing, and ultrafast optical gating. The nonlinearities leading to stimulated scattering are very large and can produce a significant gain in millimeter thick crystals. In order for amplification of scattered light to take place the nonlinearities have to have a nonlocal character due to diffusion or drift fields. Figure 4 shows a two-beam interaction effect in a photorefractive medium leading to stimulated scattering. Solid lines represent intensity maxima of interfering light beams while dashed lines represent diffused index change maxima. The fact that the two do not overlap allows amplification of the signal beam, with the maximum gain occurring at π/2 phase shift.

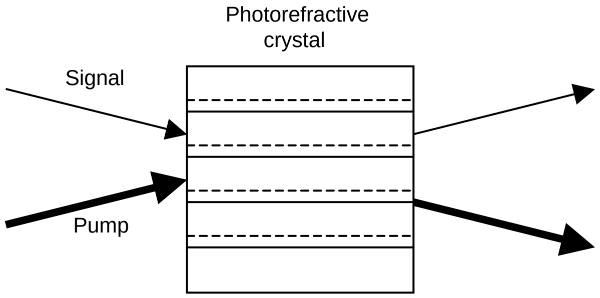

Figure 4 Stimulated photorefractive scattering diagram showing incident pump and signal beams, amplified signal beam, and an index grating formed in the photorefractive crystal.

Stimulated Compton and Thomson Scattering

Light scattering from free electrons is described by Thomson scattering in the low-energy limit, with a cross-section of $(8\pi)/(3)((e^2)/(4\pi\varepsilon_0 mc^2)^2)$, and by Compton scattering in the relativistic limit. Stimulated Thomson or Compton scattering is the source of light emitted from free electron lasers. In these interactions, the electron scatters incoming radiation in its rest frame to a backscattered wave. The ponderomotive force resulting from the interaction with the scattered waves increases the periodic electron bunching, which in turn increases the efficiency of scattering. In some cases, the incoming electromagnetic wave is provided by a microwave source, while in others it arises from a static magnetic wiggler, which appears as an incident electromagnetic wave in the electron rest frame because of a relativistic Lorentz transformation. Free electron lasers are a source of tunable coherent radiation from the microwave region to the ultraviolet.

Self-focusing

In the presence of a strong laser field with intensity I, the refractive index of many materials changes according to the relation

$$n = n_0 + n_2 I$$

where n_2 is termed the nonlinear index. This change in refractive index gives rise to several effects, depending on the temporal and spatial properties of the light. In one type of interaction, a strong forward beam couples to two weaker beams that propagate at small forward angles, providing gain for the weaker beams. One of the weak beams interferes with the strong forward beam to form a phase grating through the nonlinear index, providing gain for the second weak beam. The interference of the second weak beam with the forward beam forms a second phase grating, providing gain for the first weak beam, thereby completing the coupling for the stimulated process. This effect is similar to Stoke/anti-Stokes Raman scattering, except that the coupling is between waves propagating in different directions rather than between waves with different frequencies.

For interactions with an instantaneous relaxation time, the gain for waves propagating at a small angle to the forward direction is given by

$$g = \frac{K_\perp}{2k}\sqrt{K_{\max}^2 - K_\perp^2}$$

where $K_\perp = \sqrt{k_x^2 + k_x^2}$ is the transverse wavevector, corresponding to the angle

$$\theta = \frac{K_\perp}{k}$$

and

$$K_{\max} = 4k^2 n^{\mathrm{L}} n_2 I$$

is the largest transverse k-vector for which gain is observed. The maximum gain is given by

$$g_{\max} = \frac{K_{\max}^2}{4k}$$

and occurs for an angle corresponding to

$$K_\perp = \frac{K_{\max}}{\sqrt{2}}$$

When the physical interaction that causes the refractive index to change with light intensity has a nonzero relaxation time, gain can be observed at angles larger than that given by $K_{\max}$.

This stimulated four-photon effect can be interpreted as the initiating stage of self-focusing. In the self-focusing effect, a laser beam with a normal mode profile (more intense in the center than at the edges) creates a positive lens in the material, focusing the beam. As the beam intensifies, the power of the lens increases, causing the beam to focus more tightly and eventually coming to a catastrophic focus if other nonlinear effects do not limit the intensity. The four-photon stimulated effect described above results in the spread of k-vectors of the forward beam, corresponding to a reduction in its diameter.

Self-focusing is responsible for breakdown and damage in many solids. Self-focusing in high-power lasers is commonly suppressed by restricting transverse k-vectors to values well below those of maximum gain.

See also

Nonlinear Optics, Applications: Self-Focusing and Related Effects (Solitons and Multiphoton Absorption). **Nonlinear Optics, Basics:** Photorefraction. **Scattering:** Scattering Theory.

Further Reading

Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.

Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

Boyd RW (1992) *Nonlinear Optics*. New York: Academic Press.

Fabellinskii IL (1968) *Molecular Scattering of Light*. New York: Plenum.

Gunter P and Huignard J-P (eds) (1989) *Photorefractive Materials and their Applications II*. Springer-Verlag Topics in Applied Physics, vol. 62. Berlin: Springer-Verlag.

Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2, pp. 1077–1150. Amsterdam, The Netherlands: North Holland.

Milanovich F (1986) Stimulated Raman scattering. In: Weber MJ (ed.) *Handbook of Laser Science and Technology. Optical Materials*, vol. III. Boca Raton, FL: CRC Press.

Penzkofer A, Laubereau A and Kaiser W (1979) High intensity Raman interactions. *Progress in Quantum Electronics* 6: 55–140.

Rabin H and Tang CL (eds) (1975) *Quantum Electronics: A Treatise*. Part A, vol. 1. New York: Academic Press.

Raymer MG and Walmsley IA (1990) The quantum coherence properties of stimulated Raman scattering. In: Wolf E (ed.) *Progress in Optics*, vol. XXVIII, p. 183. Amsterdam, The Netherlands: Elsevier.

Reintjes JF (1984) *Nonlinear Optical Parametric Processes in Liquids and Gases*. New York: Academic Press.

Reintjes JF (1995) Stimulated Raman and Brillouin scattering. Handbook of laser science and technology, supplement 2. In: Weber MJ (ed.) *Optical Materials*, pp. 334–364. Boca Raton, FL: CRC Press.

Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV. New York: McGraw-Hill.

Shen YR (1984) *Principles of Nonlinear Optics*. New York: Wiley.

Shen YR and Bloembergen N (1965) Theory of stimulated Brillouin and Raman scattering. *Physical Review* 137: A1787.

SEMICONDUCTOR MATERIALS

Contents

Amorphous Semiconductors
Band Structure Engineering
Dilute Magnetic Semiconductors
GaAs Based Compounds
Group IV Semiconductors, Si/SiGe
III-Nitrides
Large Gap II–VI Semiconductors
Lead Salts
Mercury Cadmium Telluride
Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures
Quantum Dots
Type-II Quantum Wells and Superlattices

Amorphous Semiconductors

J I B Wilson, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Amorphous semiconductors comprise several classes of materials having a range of useful properties that have made many of them commercially important. Their special characteristics are derived from the particular techniques used in their synthesis, and include the ability of some to change their structure under the influence of heat or illumination. Many are photoconductive, making them valuable for photodetectors when large area devices are required (e.g., xerography), or photovoltaic, hydrogenated amorphous silicon being widely used in thin-film solar cells.

'Amorphous' means noncrystalline, thus encompassing glassy materials and disordered covalent-bonded materials. Glasses, such as the amorphous chalcogenides (glassy sulfides, selenides, and tellurides), melt to viscous liquids when heated, whereas disordered semiconductors like amorphous silicon are metastable thin films that vaporize or crystallize when heated. Two other groups of glassy semiconductors, the covalent-bonded II–IV–V_2 compounds and ionic-bonded oxides (e.g., $V_2O_5P_2O_5$) will not be discussed here. Amorphous organic semiconductors are receiving much interest now, because of their potential use in low-cost display devices, but are more logically treated as polymer materials.

The following sections explain the main preparation methods and the structures that result, the importance of defects and the influence these have on optical and electrical behavior, and some of the most important applications.

Amorphous Structure

The common property that distinguishes all amorphous materials from crystalline ones is the absence of long-range order in their atomic structure. These disordered solids have no periodic or translational symmetry. The term is sometimes loosely applied, and it should be noted that elements and compounds that are able to form several structures might contain different phases. An example is 'hard carbon' which in thin-film form may contain tetrahedral amorphous carbon and hydrogenated amorphous carbon, as well as microcrystallites of graphite and diamond. Interestingly, the prediction in 1989 by A.Y. Liu and M.L. Cohen, that crystalline cubic-C_3N_4 would be even harder than diamond, has not been confirmed by experiment. Only a hard amorphous C_xN_y has been produced, even by nitrogen ion beam assisted sputtering of graphite, which it was thought would have provided sufficient energy to form the close C–N bonds.

When the common crystalline semiconductors, Si and Ge, melt they become liquid metallic conductors. This is because the solid is formed by directional covalent bonds that produce tetrahedral packing of the atoms. In the liquid the bonds are weakened and a closer packed structure found in metals results. However a less ordered atomic arrangement results from condensing vaporized Si or Ge on to a cold substrate. In this case the atoms cannot move into their preferred positions for either tetrahedral or close packing arrays and instead are frozen almost wherever they land. Nonetheless, the arrangement of atoms is closer to that of the tetrahedral crystalline semiconductor than that of the metallic liquid. The radial distribution function of an amorphous semiconductor provides the numbers of atoms versus distance from an arbitrary atom. It shows that the nearest-neighbor distance is the same as in the equivalent crystalline material, although with some spread of values, but that the separation of atoms further away has a less well-defined value. Thus pure amorphous Si and Ge retain the directional covalent bonds between nearest neighbors but have distorted bond angles that make next nearest neighbors lie in different positions from the equivalent crystalline solid. Consequently there is only short-range order and the atoms form a continuous random network, a concept introduced by W.H. Zachariasen in 1932 for glasses. The deposition conditions generally lead to further disruptive effects that will incorporate small voids where the distortion is too large to allow a bond between atoms. These voids together with broken bonds give a defective network that is far from the ideal random tetrahedral 'lattice'.

Amorphous Semiconductor Energy Bands

A frequently asked question is 'Why can amorphous materials show semiconducting behavior?' since they do not have the periodic structure that leads to the allowed energy band scheme in crystalline semiconductors. However, note that metals do not become insulators when they melt, demonstrating that electrical conductivity does not require a periodic atomic structure. In fact, electron energy levels in solids may also be derived by considering chemical bonds when discrete atoms are brought closer together. The valence band in a semiconductor arises from the electron bonding orbitals and the conduction band from the antibonding electron orbitals. The forbidden energy gap (bandgap) is the separation between these two bands, which are only partially filled in semiconductors at room temperature. Thus amorphous semiconductors will have bonding and antibonding energy bands, and a more or less well-defined bandgap according to the defect content.

In fact, pure amorphous Si and Ge are extremely poor semiconductors due to an ill-defined bandgap and the large concentration of defects. Distorted bonds produce allowed electron energy levels that lie adjacent to the well-defined bands in crystals, producing tails of energy levels extending into the bandgap. Electrons that take up these 'tail state' energies are not entirely free to move through the solid but can only hop from one defective site to another. Additional defects arise from unsatisfied covalent bonds that 'dangle' in the spaces between the atoms. These produce electron energy levels within the forbidden gap (Figure 1). Indeed, their concentration is so high in amorphous silicon produced from a condensed silicon vapor, that they fix the position of the Fermi level at approximately the middle of the bandgap and the semiconductor has a high resistivity. The usual option of reducing a semiconductor's resistivity by adding small amounts of dopants from adjacent columns of the periodic table has little effect in this case. A donor atom from group V (e.g., phosphorus) may have all five valence electrons satisfied by a locally distorted arrangement of silicon atoms, instead of having only four bonded electrons and a loosely bound electron as in crystalline silicon. A similar picture exists for an acceptor atom from group III (e.g., boron).

Semiconductors containing group VI elements, S, Se, or Te, have two-fold coordination instead of the tetrahedral arrangement of the group IV elements, Si and Ge. The bonding between atoms in the solid is formed from four of the six p-electrons, which give

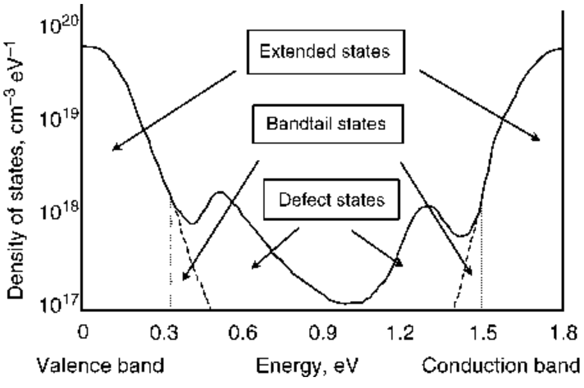

Figure 1 Schematic density of states for hydrogenated amorphous silicon.

rise to the bonding and antibonding energy bands of the solid, and the two remaining p-electrons from each atom form a non-bonding 'lone pair'. The lone pair energy band lies between the bonding and antibonding bands, and is the highest filled band in the solid, thus being the valence band.

Just as in amorphous Si or Ge, dangling bond defects will have energies between the valence and conduction bands, but in lone-pair semiconductors such disturbed energy states will now produce energy levels that fall within the lone pair band instead of within the bandgap. This makes it difficult to dope chalcogenide semiconductors, especially by donors, since the additional energy levels often lie within an existing energy band. Similar considerations apply to ternary chalcogenide alloys: these will have lone-pair bands arising from the group VI element, with additional energy levels from the other components (e.g., from group IV or V). However, amorphous chalcogenides will also contain three-fold coordinated chalcogenide atoms and these defects can associate in pairs or may interact with an adjacent two-fold coordinated atom. Defect centers with two oppositely charged chalcogens may form a 'valence alternation pair' (VAP).

So it may be seen that the semiconducting behavior of solids arises from the nearest-neighbor interactions of the atoms giving rise to allowed electron energy bands, and that long-range order is not necessary to produce the semiconductor's forbidden energy gap. Defects then produce localized energy levels that usually lie within the forbidden gap.

Preparation of Amorphous Silicon

Despite the early use of selenium as a photoconductive semiconductor in xerography, most amorphous semiconductors might have stayed as curious but technologically impractical materials, if an alternative preparation method to physical vapor deposition had not been discovered.

In the 1950s the preparation of high-purity single-crystal silicon was not yet established, and work at Standard Telecommunication Laboratories, UK, was attempting to use radio-frequency heating to produce silicon from low-pressure silane gas (SiH_4). It was noticed that a side effect was a deposit of resistive, noncrystalline material on cool parts of the container, close to unwanted plasma 'glow discharges'. In following up this observation, the technique was developed to produce a variety of other useful compounds from electrical discharges in appropriate gas mixtures, particularly silicon nitrides and oxides. With new scientific interest in amorphous semiconductors, the silicon coatings from silane were examined in more detail, and were shown to be significantly different from amorphous silicon produced by evaporating and condensing silicon. This new form was very photoconductive and could be doped n-type by adding phosphine to the silane. A more detailed study of doping led to the seminal papers by W.E. Spear and P.G. LeComber of Dundee University, UK, and to an explosive growth in the whole subject, once the potential commercial applications became apparent.

Despite many alternative synthesis methods, plasma chemical vapor deposition (CVD) remains the most widely used, and produces material that has not been surpassed in quality. The key features are a vacuum chamber into which a controlled flow of each of the gaseous sources is fed, and a heated plate on which the substrate is placed for coating. The decomposition of the gases is driven by an electrical discharge between two electrodes, which enables the gas temperature to be much lower than in conventional pyrolysis (thermal CVD). An important aspect of the process is the incorporation of hydrogen into the deposited material, which is thus more accurately known as hydrogenated amorphous silicon, a-Si:H. In the best material, hydrogen is mainly bonded singly as Si–H, but if the preparative conditions are not optimum then there may also be large amounts of $Si{-}H_2$ (as may be seen from the infrared absorption spectra). Alternative source gases usually contain hydrogen (e.g., Si_2H_6), and hydrogen may be added as a diluting agent for the doping gases (e.g., B_2H_6 or PH_3). Other precursors include fluorinated gases, but these have failed to give the predicted improvements in performance of the semiconductor in such applications as solar cells, and have greater chemical reactivity.

Since silane is spontaneously flammable in air, and the dopant gases are not only flammable but also very toxic, efforts have been made to produce a-Si:H by other means. Sputtering from undoped or doped silicon targets has been a popular choice, but it is still necessary to incorporate hydrogen into the deposited material, for instance by adding it to the usual argon sputter gas. This is because hydrogen serves to passivate unsatisfied 'dangling' Si bonds and to reduce the stresses that would occur in a pure silicon network. Photo-CVD, using ultraviolet light to drive the reaction without raising the temperature, can produce high-quality material but has the problem of unwanted silicon deposits on the illuminated reaction chamber window. Although conventional silane pyrolysis is not suitable for producing a-Si:H, because the temperatures needed for useful deposition rates also remove much of the hydrogen if the substrate is too hot, it is possible to heat the gas separately from the substrate by means of a hot filament and so produce good-quality films.

Other hydrogenated silicon alloy semiconductors are synthesized by additions to silane of ammonia (giving amorphous silicon nitride, a-Si:N:H), or a hydrocarbon gas such as methane (giving a-Si:C:H), or germane (giving a-Si:Ge:H).

Preparation of Amorphous Chalcogenides

Unlike the tetrahedral group IV elements, useful members of this class of amorphous semiconductors may be produced by vacuum evaporation and condensation on to a relatively cool substrate. Compounds, such as As_2Se_3, do not evaporate congruently and so the deposit will change composition as the evaporation proceeds. More uniform compositions may be obtained by sputtering with argon, and glassy oxides may also be obtained in this way with oxygen added to the inert gas. Chalcogenide glasses may also be prepared by rapid cooling from a melt, or by chemical vapor deposition from heated vapor precursors. Since glasses are thermodynamically unstable, they will tend to transform to a more stable, sometimes crystalline structure. The rate for this process depends on the temperature and may be effectively too slow to observe.

Optical Properties of Amorphous Semiconductors

Optical absorption spectra of amorphous semiconductors clearly show a forbidden energy gap. By plotting a suitable function of the absorption coefficient, α, versus photon energy, $h\nu$, usually that given by J. Tauc, $(\alpha h\nu)^{1/2}$, the 'optical gap' may be determined from the intercept of the linear portion of the plot with the x-axis. Unlike crystalline semiconductors, amorphous ones have no k-selection rules to give indirect and direct transitions across the bandgap, thus a-Si:H behaves as a direct gap material. This 'optical gap' is greater than the bandgap determined from electrical conductivity measurements, since the latter will have contributions from both localized bandtail states (low electrical mobility) and extended states (high electrical mobility) (Figure 2).

The absorption spectra also have a portion at lower photon energies having an exponential dependence on photon energy, known as the Urbach tail. This is related to the structural disorder in the material and the tail states. Transitions involving defect bandgap states lie at even lower photon energy, and the density of defects can be determined from the weak absorption in this region of the spectrum, perhaps by using photothermal deflection spectroscopy to detect the weak optical absorption.

Photoluminescence also provides a probe of the defect content of many materials. Interpretation and identification of the actual defect (e.g., chemical impurity or structural imperfection) is aided by

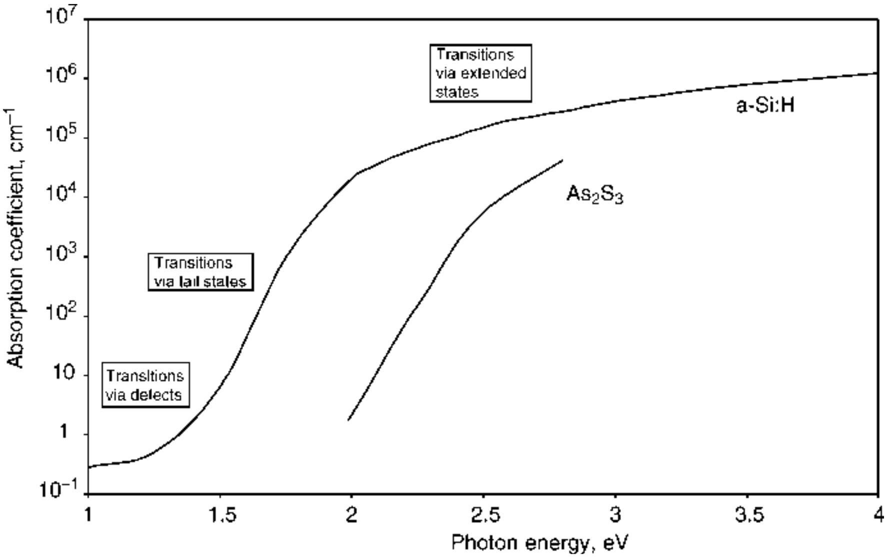

Figure 2 Typical optical absorption spectra of hydrogenated amorphous silicon and arsenic sulfide.

reference to the crystalline form. Both crystalline and amorphous forms of chalcogenide have a wealth of special effects, including reversible photoinduced bleaching and darkening, and laser-induced crystallization and amorphization. In addition the creation and dispersal of defects by appropriate illumination of some amorphous chalcogenides is accompanied by changes in the electron spin resonance (ESR) signal. Some of these effects have been the basis for applications in optical memories. More prosaically, chalcogenide glasses are important infrared transparent materials and amorphous chalcogenide thin films have had historical importance as the photoconductive layer in xerography, although here they are now being supplanted by less toxic amorphous silicon alloys.

Electrical Properties of Amorphous Semiconductors

The electrical conductivity of amorphous semiconductors is often controlled by the defects that are introduced by impurities or by structural flaws. For a-Si:H this provides alternative conduction mechanisms, via extended state electron and hole conduction (as in crystalline silicon), and via hopping of electrons or holes through localized energy levels within the bandtails or within the bandgap. The relative importance of each route depends on the temperature and on the density of states (DOS) in each energy range, as well as on the mobility of the charge carriers.

These different conduction mechanisms are clearly seen in the temperature dependence of the conductivity of a-Si:H and of many amorphous chalcogenides, since a semilog plot of conductivity versus $1000/T$ K will have portions with different slopes as each mechanism dominates a particular temperature range. Near room temperature, good-quality a-Si:H films in the dark exhibit a well-defined Arrhenius relationship with an activation energy that is the energy separation of the Fermi level from the conduction band extended states, electrons dominating the conductivity. The conduction band extended states do not have the high electron mobility found in crystalline silicon, due to the disruption of long-range order and the consequent increased electron scattering. Typical electron mobilities in a-Si:H are $\sim 1\ \mathrm{cm}^2\ \mathrm{v}^{-1}\ \mathrm{s}^{-1}$, and hole mobilities are even lower.

The band tail states have even worse carrier mobilities, since movement is by hopping to vacant sites that might lie at some distance, and hopping is unlikely between sites that have a large energy difference. (The process has an inverse exponential dependence on the energy difference between the initial and final states.) Hopping of carriers between states that lie within the bandgap is a significant contribution to the conductivity only at low temperatures, unless the density of such defects is high. In materials having weakly localized states, nearest-neighbor hopping is supplanted by variable range hopping, as carriers prefer to hop to a more distant site if it has a smaller energy difference from the starting site than do the neighboring sites. As the temperature is lowered, this hopping distance will increase and the conductivity has N.F. Mott's characteristic $\exp(-B/T^{1/4})$ dependence on temperature, T. Lower-quality films deposited by plasma CVD on to unheated substrates tend to be p-type due to hopping conductivity through the high density of defect states. The link between substrate deposition temperature and material quality is caused by the way in which hydrogen is incorporated into the growing material: high temperatures allow only singly bonded hydrogen, which passivates dangling bonds thus removing bandgap defect energy levels. Low temperatures also allow hydrogen to be incorporated as $\mathrm{Si{-}H_2}$ and this disrupts the Si–Si lattice bonding, producing more defect levels. Confirmation of this effect is obtained from sputtered amorphous silicon, with and without hydrogen added to the argon.

When a-Si:H is illuminated it may show strong photoconductivity, depending on the preparative conditions. Free carriers are generated by the illumination but may fall into defect states (energies within the band gap) where they can be released or can recombine with the oppositely charged carriers. The detailed response to light (wavelength, intensity, duration) and temperature thus depends on the distribution and concentration of defects. Luminescence effects are similarly influenced. Most amorphous chalcogenides are only weakly photoconductive due to their high density of gap states, those near the middle of the gap (e.g., VAPs) being highly effective recombination centers.

The electrical properties of other hydrogenated amorphous tetrahedral semiconductors based on silicon are determined by their defect populations, and in general they are less well behaved. a-Si:Ge:H and a-Si:Sn:H have reduced bandgaps and a-Si:C:H and a-Si:N:H have increased bandgaps, which offer many possibilities for multilayer device design, provided that their normally high defect contents can be reduced or controlled.

The study of glassy semiconductors was stimulated in the 1960s by the discovery of electrical switching in sputtered multicomponent thin films.

These simple devices changed from a nonconducting to a conducting state when the applied electric field exceeded a certain threshold, and returned to their initial state when the current fell below a certain value. A variety of other switching types was discovered, especially under pulsed operation, including those having a memory effect based on the reversible crystallization and vitrification of a narrow filament between the contacts. A theoretical model of these devices comprises both electronic and thermal effects, which are responsible for initiating the switching event and for memory effects. A practical problem has been the variable jitter in the switching time.

Applications

Amorphous silicon alloy films are valuable as the active layers in thin-film photovoltaic cells, two-dimensional optical position detectors, linear image sensors (optical scanners), and thin-film transistors used in liquid crystal display panels. They also have uses as antireflection coatings and planar optical waveguides. Amorphous chalcogenide films have been used as electrical switches and memory devices, for optical data storage, and as dry photoresists and electron beam resists. Both types of amorphous material have been used in xerography and medical X-ray imaging detectors.

Amorphous silicon solar cells have power conversion efficiencies of ~12% for the most complicated structures. These are tandem cells that use different alloys (including a-Si:C:H) for the various layers, in order to enhance effective absorption of the solar spectrum. A serious drawback of using amorphous semiconductors in electronic devices is their short carrier diffusion length. In solar cells this is circumvented by ensuring that there is a built-in field to separate photogenerated electron–hole pairs before they can recombine. Because doped amorphous silicon alloys have high defect densities, it is difficult to make the effective pn junctions that provide such built-in fields in conventional crystalline silicon cells. Instead, amorphous silicon cells use pin structures, where the i-layer is effectively undoped and provides an extended electric field between the p-i and i-n junctions. Long periods of illumination increase the dark current in these devices, as additional defects are generated. This 'Staebler–Wronski effect' is believed to be associated with hydrogen movement at a passivated dangling bond (Si–H) and a nearby weak Si–Si bond. It is a cause of slow degradation in unstabilized solar cells although the change may be reversed by thermal annealing.

In contrast to 'minority carrier' photovoltaic devices, thin-film transistors necessitate the control of majority carriers alone. To address each pixel in a liquid crystal display, a switching element is needed and simple thin-film transistors fulfill this role. Although low carrier mobilities in amorphous semiconductors restrict the current passed and switching speeds, the requirement for coating large area glass panels at low temperatures make a-Si:H with a-Si:N:H gate dielectric almost the only options. Improved performance is possible by using pulsed laser irradiation to change the amorphous semiconductor to polycrystalline material.

Further Reading

Brodsky MH (ed.) (1985) *Amorphous Semiconductors*, 2nd edn. Berlin: Springer-Verlag.

Elliot SR (1984) *Physics of Amorphous Materials*. London: Longman.

EMIS Datareviews Series No.1 (1989) *Properties of Amorphous Silicon*, 2nd edn. London: INSPEC.

Joannopoulos JD and Lucovsky G (eds) (1984) *The Physics of Hydrogenated Amorphous Silicon*. Topics in Applied Physics, vols 55 and 56. Berlin: Springer-Verlag.

Journal of Non-Crystalline Solids (2000) *Proceedings of the 18th International Conference on Amorphous and Microcrystalline Semiconductors (1999)*, vol. 266–269. This journal has published the proceedings of most of these well-established annual meetings dealing with all aspects of amorphous semiconductors.

Kramer B and MacKinnon A (1993) Localization – theory and experiment. *Reports on Progress in Physics* 56: 1469–1564.

Luft W and Tsuo YS (1993) *Hydrogenated Amorphous Silicon Alloy Deposition Processes*. New York: Marcel Dekker.

Madan A and Shaw MP (1988) *The Physics and Applications of Amorphous Semiconductors*. San Diego, CA: Academic Press.

Mott NF (1985) Electrons in non-crystalline materials – the last twenty five years. *Contemporary Physics* 26: 203–215.

Pankove JI (volume ed.) (1984) Hydrogenated amorphous silicon. In: Willardson RK and Beer AC (series eds) *Semiconductors and Semimetals*, vol. 21, Parts A–D. Florida: Academic Press.

Zallen R (1983) *The Physics of Amorphous Solids*. New York: Wiley.

Band Structure Engineering

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Direct gap semiconductors have many unique properties which are advantageous for optical and electronic applications, and in particular for detectors and emitters of infrared radiation associated with interband optical transitions. Gallium arsenide, GaAs, based structures have principally been used in the near-infrared, and so-called narrow gap semiconductors in the mid-infrared (MIR) spectral regions. More recently systems based on intersub-band transitions have shown great promise for MIR devices, with the obvious advantage for some applications that they can be constructed from large gap, better controlled materials (mainly InP- and GaAs-based to date) and have a peak operating frequency which is determined simply by the design of the particular quantum structure. Both approaches have been substantially advanced by band structure engineering techniques, reducing Auger losses and phonon losses, respectively, which is the subject of the present work.

The classic examples of narrow gap materials are the III–V semiconductors (InSb, InAs, and their related alloys), II–VI semiconductors (mainly $Hg_{1-x}Cd_xTe$), and the lead salt alloys. As a direct result of the small energy gap, these materials have a very low electron effective mass, very high electron mobility, even at room temperature, and a high saturation velocity. In addition they have a high ratio of stimulated to spontaneous transition probabilities and a large nonlinear susceptibility. However, they have the disadvantage of a high intrinsic carrier density at room temperature and thermal generation rate, so that a fundamental limitation to the performance of devices at elevated temperatures (preferably room temperature) is nonradiative losses resulting from Auger recombination processes. One solution is to cool these systems to 80 K or below which involves expensive and bulky cooling methods such as Stirling-cycle coolers, negating their advantages over other, otherwise inferior, devices. In the past 10 years there has been an intensive effort to circumvent this limitation by quantum structure and so-called 'band structure engineering' techniques made possible by epitaxial growth techniques, whereby the Auger processes are suppressed. Since the emphasis of the band structure engineering work has been mainly towards the optimization of III–V semiconductor lasers rather than detectors, we reflect that emphasis here. Although developed more recently than the bipolar diode lasers, the unipolar quantum cascade (QC) lasers provide perhaps the best example of systems that can be conveniently optimized by band structure engineering techniques. In this case, phonon processes dominate so that the design is to do with phonon suppression. Thus, we are concerned principally with suppression of Auger processes for interband devices, and phonon processes for intraband devices.

Nonradiative Processes and Band Structure Engineering

Suppression of Auger Recombination

Electron recombination processes are dealt with in detail elsewhere in the Encyclopedia (*see* **Semiconductor Physics:** Recombination Processes) but we briefly summarize the main points here. It is well known that there are three principal interband recombination mechanisms by which thermodynamic equilibrium is established in semiconductors: Shockley–Read recombination, radiative recombination, and Auger recombination. These are shown schematically in Figure 1. The Shockley–Read recombination occurs via extrinsic energy levels produced by lattice defects and impurities, Figure 1a, and can in principle be controlled by the procedure used to grow the material; hence it does not provide a fundamental limit to the

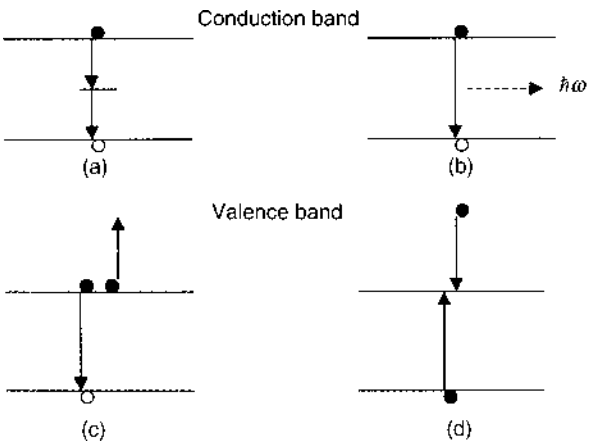

Figure 1 Electronic processes in semiconductors: (a) Shockley–Read recombination; (b) radiative recombination; (c) Auger recombination; (d) impact ionization.

carrier lifetime. The radiative and Auger mechanisms (Figure 1b,c, respectively) are, however, fundamental and are determined by the electronic structure of the material. The former is central to the operation of light-emitting diodes and is approximately inversely proportional to the electron and hole concentration. Optimum detector operation is achieved in materials that are limited by the radiative lifetime mechanism. Thus, in both cases it is desirable to minimize Auger processes.

By contrast with the two-particle radiative recombination process, Auger recombination is a three-particle interaction. For example in one so-called CHCC (or Auger-1) process an electron collides with a second electron; the first electron recombines with a hole in the valence band and the energy produced by the electron–electron collision excites the second electron higher up into the conduction band. This second electron eventually de-excites by phonon emission, and can cause observable heating of carriers by an Auger type of up-conversion. However, for the present article we are only interested in the first part of the process, so we do not consider the heating further here. The Auger process is the inverse of impact ionization where an energetic conduction electron collides with a valence electron, thereby creating an electron–hole pair as shown in Figure 1d. The final state consists of two low-energy electrons and one hole. Analogously, in Auger recombination the initial state has three particles (e.g., two electrons and one hole) and the final state one (e.g., one electron). Due to this three-body nature, the collision probability (or transition rate) per unit volume, R, varies with the cube of the carrier concentration and (if $n = p$) may be written $R = Cn^3$, where n is the electron density and C is referred to as the Auger coefficient. Note that the Auger rate is sometimes defined per carrier as $R' = Cn^2 = 1/\tau_A$, where τ_A is the Auger lifetime.

Narrow gap III–V semiconductors have the familiar three-band (so-called Kane energy band structure – (*see* **Semiconductor Physics:** Band Structure and Optical Properties) consisting of an s-like conduction band separated by a small energy gap from a triply degenerate p-like valence band, part of whose degeneracy is raised by spin–orbit splitting as shown in Figure 2. Thus, a large number of permutations of Auger processes become possible for n- and p-type semiconductors. Those which are initiated by a collision between two electrons are referred to as conduction, or n-n, Auger processes (e.g., the CHCC process), and those by a collision between two holes as valence, or p-p, Auger processes. It can be shown that the n-n process (so-called CHCC or Auger-1) illustrated in Figure 3a is dominant in bulk n-type and the p-p process (so-called CHLH or Auger-7) shown in Figure 3c is dominant in bulk p-type semiconductors, and so these have received the most attention. For larger gap (shorter wavelength) semiconductors there is sometimes a significant resonance between the bandgap and the spin–orbit splitting and the p-p process (CHSH) shown in Figure 3b becomes important. Clearly these simple descriptions require extension when the more complex band structure of strained layer superlattice quantum systems is considered.

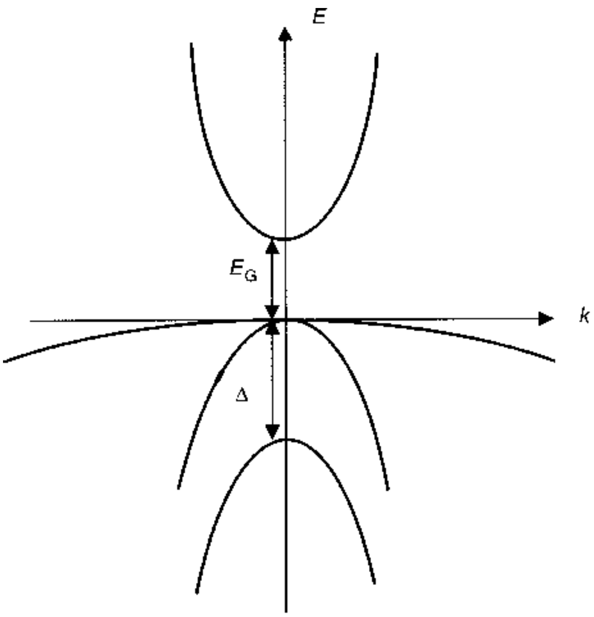

Figure 2 Schematic diagram of the conduction and valence band structure of InSb showing the energy bands around the Γ-point ($k = 0$) that are involved in Auger transitions.

With certain approximations, some analytic solutions for the Auger recombination rates are possible. This model is, while not appropriate for higher levels of band structure engineering, very important for giving an insight into how Auger effects should be studied, and how they themselves might be engineered. The approximations generally used are parabolic bands and Boltzmann statistics, but some further analytic results may also be obtained when these restrictions are relaxed. Bloch functions and Fermi's golden rule are utilized (*see* **Semiconductor Physics:** Recombination Processes) to describe the particle states, so obtaining the general expression for the Auger rate per unit volume. Energy and momentum conservation provide strong constraints on the resulting integral, and give rise to a considerable reduction in the Auger transition rate. However if, for example, the momentum conservation constraint is relaxed due to occupied levels existing with a broad range of momenta, such as at some band extremum with a

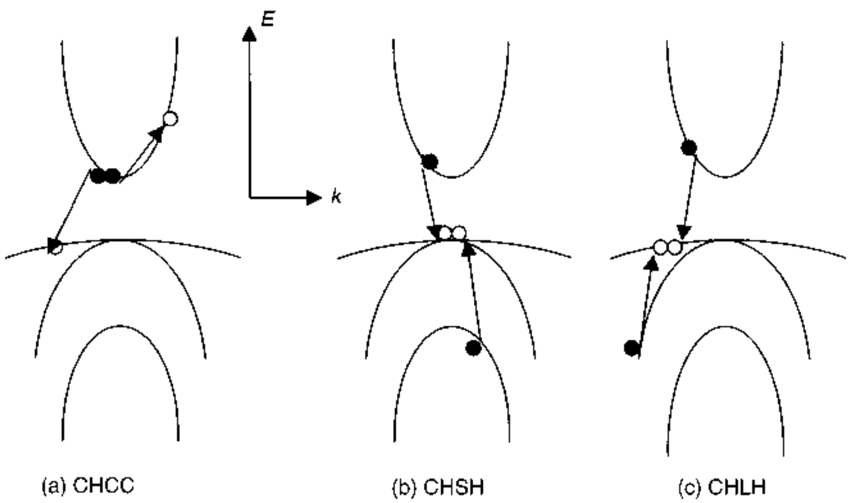

Figure 3 Important Auger recombination mechanisms for mid-infrared (MIR) semiconductor detectors and lasers: (a) CHCC or Auger-1; (b) CHSH; (c) CHLH or Auger-7.

large effective mass, then the Auger rate will be strongly enhanced. Or to put it the other way round, the Auger rate can be reduced by engineering the band structures such that the heavy hole effective mass is reduced (e.g., by stress – see below) to approach that of the conduction band. Furthermore, since we are concerned with nonradiative recombination across the bandgap with energy E_g, any energy levels (again band extrema in particular) separated from the conduction or valence band edges by E_g are particularly susceptible to the excitation part of the Auger process. The Auger rate may therefore be reduced by designing the structure so that there are no band extrema at energies E_g from the conduction and valence band edges.

Interband (Bipolar) Lasers

The technology of band structure engineering, which includes superlattices, quantum confinement, and intentional incorporation of strain, has now become sufficiently advanced that it can be employed in a routine manner. As originally proposed, the modified band structure of strained III–V quantum-well structures can lead to significant benefits for diode laser performance, including reduced threshold current density, improved efficiency, and enhanced dynamic response and speed. Predicted advantages have been demonstrated, although for near-infrared devices these have not been quite as substantial as originally hoped because of the presence of phonon-assisted Auger processes which are not so susceptible to band structure engineering techniques. Strained layer lasers are now commercially available at a variety of MIR wavelengths, where the dominant nonradiative loss is via the direct interband Auger scattering routes.

From the point of view of the present article the main contribution of 'designed' strain splittings and quantum confinement is in the reduction of the Auger coefficient itself, by removing states that easily satisfy the momentum and energy conservation conditions required for the process to take place. However, in addition to this there can be a substantial reduction in the threshold carrier injection level required for population inversion in a laser. In real III–V semiconductors the heavy hole valence band mass, m_h, is so much greater than the electron mass, m_e, that the usual laser picture of degenerate electron and hole distributions does not apply. At normal carrier injection levels the quasi-Fermi level of the holes is above the top of the valence band resulting in a classical heavy hole distribution. Thus, the lower laser levels are almost completely filled with electrons, by contrast with the ideal situation of two mirror bands. This leads to an injection level for the laser threshold of greater than 2:1 required for the actual case compared to the ideal case. Since the Auger recombination rate per unit volume is approximately proportional to the cube of n, this factor of ~ 2 reduction in injection level can produce a factor of almost an order of magnitude reduction in this loss mechanism. Large biaxial strains, either compressive or tensile, can lead to a considerable reduction in the hole mass and hence in the density of states at the top of the valence band, so that the laser characteristics may then approach those of an ideal semiconductor, with $m_e = m_h$.

A further important loss mechanism is free hole absorption of the heavy holes resulting from intervalence band transitions. This can be effectively eliminated, and other improvements made to laser performance, by the introduction of a strained layer structure which raises the degeneracy of the light and heavy hole at $k = 0$ and reduces the mass of the heavy hole. One example of this comes just from the symmetry of the problem, where clearly the cubic symmetry of the p-like valence band is not matched to the uniaxial symmetry of the laser beam. Application of biaxial or uniaxial strain removes the cubic symmetry, changing the relative energies of the p_x, p_y, and p_z valence states, and making it possible for almost all of the injected holes to be in the states of the correct symmetry to take part in the lasing action. The one parasitic phenomenon in semiconductor lasers which is hardly affected by band structure engineering is spontaneous emission, which will always make some contribution to the current density required at threshold.

The band structure of a bulk unstrained direct gap tetrahedral semiconductor is shown again for comparison in Figure 2. The effect of biaxial tension and compression is shown respectively in Figure 4a,b. In each case the lowest conduction band is approximately parabolic near the zone center, but becomes anisotropic with the electron dispersion at small k. The axial strain breaks the cubic symmetry of the semiconductor, introducing a tetragonal distortion which splits the degeneracy of the light and heavy hole states at the valence maximum, Γ, typically by about 60–80 meV for a 1% lattice mismatch. The resulting valence band structure is highly anisotropic (Figure 4a,b), with the band which is heavy along the strain axis, wavevector $k_\perp$, being comparatively light in the plane, along $k_\parallel$, and vice versa. The in-plane mass, $m_\parallel$, determines the density of states, and thus, under compression (Figure 4b), can be reduced according to the principles described above to improve conditions towards those for ideal laser operation: viz, (1) that there is only one band at the valence band maximum; (2) that it has an effective mass as close to that of the electron as possible; and (3) that above the Bernard–Duraffourg transparency condition, the polarization of the gain is anisotropic, with spontaneous emission and gain being suppressed along all directions except along the axis contributing to the laser beam.

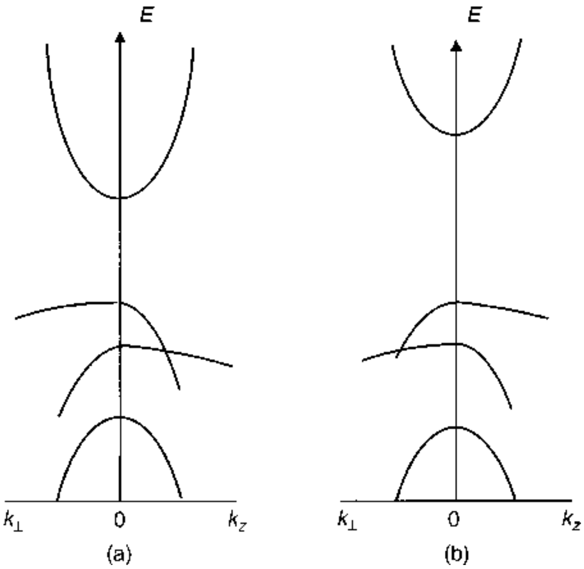

(a)
(b)

Figure 4 Schematic representation of the band structure of a direct gap tetrahedral semiconductor: (a) under biaxial tension the hydrostatic component of the tension reduces the mean bandgap, while the axial component splits the degeneracy of the valence band maximum and introduces an anisotropic valence band structure (the highest band being light along the strain axis, k_z and comparatively heavy perpendicular to that axis, $k_\perp$); (b) under biaxial compression the mean bandgap increases and the valence splitting is reversed.

Since the Auger recombination involves three carriers, the Auger current J_{NR} in a device with an undoped active region varies approximately as $J_{NR}(T) \propto C(T)n_{th}^3$, where $C(T)$ is the temperature-dependent Auger coefficient per unit volume and nth is the threshold carrier density. The influence of strain on Auger recombination has two aspects, as discussed above. First, J_{NR} is very sensitive to any reduction in n_{th} brought about by strain either through a decrease in h (heavy hole) mass or an increase in the optical transition strength. Secondly, strain may change the magnitude of $C(T)$ for the interband Auger processes described earlier (Figure 3). A simple calculation of Auger recombination in a quantum well heterostructure, assuming parabolic bands and Boltzmann statistics, shows that the coefficient increases exponentially with temperature as $C(T) = C_o\exp(-E_a/kT)$, where the activation energy, E_a, is approximately dependent on $(m_h)^{-1}$. Thus, both of these should be reduced by several orders of magnitude due to a reduction in the heavy hole mass, m_h, resulting from the strain effects discussed above.

Phonon Scattering: QWIPs and QC Lasers

Sub-band transitions are useful for MIR detectors and emitters because they make it possible to cover a broad range of MIR wavelengths using the same combination of large gap, InP-based and GaAs-based state of the art semiconductor alloys that are used for high-speed electronics and optical communications. They are the most advanced of the low-dimensional solids from the point of view of demonstrated detector performance. This superior control of materials results in high uniformity and has thus allowed the demonstration of large 128 × 128 arrays with excellent imaging performance. Detectivities $D^* > 10^{10}\,\mathrm{cm\,Hz^{1/2}\,W^{-1}}$, low noise equivalent temperature differences, and excellent infrared video imaging can be achieved at 10 μm for temperatures $T = 10-80$ K.

Although the QWIP (quantum well infrared photodetectors) detectors showed promise for many years as alternatives to HgCdTe systems, the first intersub-band laser was not realized until the demonstration of the quantum cascade (QC) laser. Intersub-band (QC) lasers differ fundamentally from bipolar diode lasers in that they rely on only one type of carrier and on electronic transitions between conduction sub-band states arising from size quantization in semiconductor quantum heterostructures. These in-plane sub-bands have the same curvature (by contrast with the interband devices which have conduction and valence bands of opposite curvature), so that the joint density of states and corresponding gain spectrum (collision limited) is much narrower and approximately symmetric. Thus, the arguments which show that the Kane-type band structure is particularly disadvantageous for the overall Auger loss mechanism in interband diode lasers do not apply; the problems resulting from having $m_e \ll m_h$ are removed and the inversion is much less sensitive to thermal broadening of the electron distribution. In addition, there is no free hole absorption, and, since the component semiconductor materials are of comparatively wide bandgap, there is negligible intrinsic carrier concentration. However, the threshold current density tends to be rather large owing to the rapid nonradiative phonon relaxation of the population inversion.

In the original QC laser structure the electrons streamed down a potential staircase, sequentially emitting photons at the steps, consisting of coupled quantum wells in which the population inversion between discrete conduction band excited states was achieved by control of tunneling. The laser structure was grown with the $Al_{0.48}In_{0.52}As/Ga_{0.47}In_{0.53}As$ heterojunction material system lattice matched to InP. Electrical pumping was achieved by alternating 25 undoped coupled-well active regions with compositionally graded layers. The coupled-well region was essentially a four-level laser system, where a population inversion was achieved between the two excited states, $n = 3$ and $n = 2$. While the nonradiative phonon emission lifetime is short ($\tau_{32} \sim$ 4.3 ps at the electric fields used in the tunneling process), strong coupling to an adjacent GaInAs well gives an even shorter lower state lifetime ($\tau_{21} \sim 0.6$ ps) from optical phonons with nearly zero momentum transfer between the strongly overlapped and closely spaced $n = 2$ and $n = 1$ states. This enables the achievement of a population inversion between these states. (Note that the separation between the $n = 2$ and $n = 1$ states, E_{21}, is designed to be approximately equal to the optical phonon energy.)

The radiative efficiency of these QC devices is unavoidably low ($<10^{-3}$) because most of the electrons in the upper sub-band relax within ~1 ps through nonradiative optical phonon emission before radiative transitions with a lifetime in the nanosecond range have a chance to occur. An alternative configuration has been proposed, the type II interband cascade laser (T2ICL), which retains the advantages of cascaded tunneling injection and wavelength tuneability but eliminates the nonradiative phonon path between valence and conduction bands of opposite curvature. In addition, band structure engineering ideas discussed above are incorporated to suppress the unwanted Auger processes; the design results in a light in-plane hole mass and the removal of resonances between the energy gap and intervalence transitions.

See also

Semiconductor Materials: Group IV Semiconductors, Si/SiGe; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Band Structure and Optical Properties; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Faist J, Capasso F, Sirtori C, Sivco DL, Hutchinson AL and Cho AL (1994) Quantum cascade laser. *Science* 264: 553–556.

O'Reilly EP and Adams A (1994) Band structure engineering in strained semiconductor lasers. *IEEE Journal of Quantum Electronics* 30: 366–379.

Pidgeon CR, Ciesla CM and Murdin B (1997) Suppression of non-radiative processes in semiconductor mid-infrared emitters and detectors. *Progress in Quantum Electronics* 21: 361–419.

Yablonovitch E and Kane EO (1988) Band structure engineering of semiconductor lasers for optical communications. *Journal of Lightwave Technology* 6: 1292–1299.

Youngdale ER, Meyer JR, Hoffman CA, Bartoli FJ, Grein CH, Young PM, Ehrenreich H, Miles RH and Chow DH (1994) Auger lifetime enhancement in InAs-GaInSb superlattices. *Applied Physics Letter* 64: 3160–3162.

Dilute Magnetic Semiconductors

R R Gałązka, Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction, General Remarks

Semimagnetic semiconductors (SMSs) are a group of solids at the interface between semiconductors and magnetic materials. Semimagnetic semiconductors, also referred to as dilute magnetic semiconductors, are semiconductor-based solid solutions where some of the cations are replaced by transition metals or rare earth elements. The crystallographic structure of the semiconductor is conserved; the lattice constant is a function of composition. From a magnetic point of view SMSs are disordered magnetic materials, since magnetic atoms are randomly distributed in the cation sublattice of the semiconductor compound. Table 1 presents the most complete list of bulk ternary SMSs. Quaternary SMSs are also investigated (e.g., PbSnMnTe or CdMnSeTe).

Generally speaking, in SMSs there coexist two interrelated and interacting subsystems: mobile delocalized charge carriers and localized magnetic moments connected with paramagnetic ions. Electronic (mainly optical) properties of SMSs have been the subject of intensive studies since early in 1970. Due to the strong spin exchange interaction between mobile carriers and localized magnetic moments (the exchange constant of the sp-d interaction is of the order of 1 eV for II–VI SMSs) significant changes of the band structure and behavior of the carriers were observed. A number of new physical phenomena were discovered, such as giant Faraday rotation (now important in magnetic field sensors and other applications in, for example, the automobile industry based on this strong magneto-optical effect), the magnetic field induced metal–insulator transition, the bound magnetic polaron and quantum thermomagnetic oscillations. Investigations of the influence of the electron subsystem on the magnetic properties of SMSs are connected with the observation of a ferromagnetic phase in PbSnMnTe induced by carrier concentration.

Molecular beam epitaxy (MBE) grown GaAs with Mn opened a new chapter of III–V SMSs showing ferromagnetic behavior with the highest transition temperature $T_c = 110\,\mathrm{K}$. In 1987 the first paper devoted to layered structure and magnetic properties of low-dimensional (LD) SMSs was published starting intensive studies of superlattices and other LD structures made of SMSs.

Energy Band Structure, Optical and Magneto-optical Properties of SMSs

Replacing cations such as Cd or Hg with a paramagnetic ion such as Mn in the same crystallographic structure does not markedly disturb the semiconductor properties of the material. The energy gap changes (increases) but the conduction and valence bands conserve their symmetry and character as in nonmagnetic semiconductor mixed crystals. The spin momentum of paramagnetic ions is connected with the 3d or 4f shell, for transition metals or rare earth elements, respectively. The energy level of 3d or 4f electrons lies below the top of the valence band and thus has negligible influence on its shape or that of the bottom of the conduction band. Thus all basic semiconductor optical properties connected with the band symmetry and topology are the same as for a typical semiconductor. These are described, for example, in articles on the lead salts (MS641) involving direct optical transitions at the L-point of the Brillouin zone, and on mercury cadmium telluride (MS639) with direct optical transitions at the Γ-point. The only influence of the magnetic subsystems on electrons comes from the spin exchange interaction between mobile carriers and localized magnetic ions. This interaction can be represented by a Heisenberg term

$$H_{\mathrm{ex}} = \bar{\sigma} \sum_i \bar{S}_i J(\mathbf{r} - \mathbf{R}_i) \qquad [1]$$

where $\overline{\sigma}$ and $\overline{S}$ are spin operators of the band electron and magnetic ion, respectively, the summation is over all lattice sites occupied by the magnetic ions, and J is the exchange constant.

Assuming mean field and virtual crystal approximations this term can be rewritten in the form which is periodic with a lattice constant:

$$H_{\mathrm{ex}} = x \langle \bar{S} \rangle_{\mathrm{av}} \bar{\sigma} \sum_R J(\mathbf{r} - \mathbf{R}) \qquad [2]$$

where x is the molar fraction of magnetic ions, $\langle \bar{S} \rangle_{\mathrm{av}}$ is the average over all magnetic ions and is directly

Table 1 Composition and crystal structure of ternary bulk SMSs

Material	*Crystal structure*	*Composition range*
II–VI (Mn)		
$Zn_{1-x}Mn_xS$	ZB	$0 < x \leq 0.10$
	W	$0.10 < x \leq 0.45$
$Zn_{1-x}Mn_xSe$	ZB	$0 < x \leq 0.30$
	W	$0.30 < x \leq 0.57$
$Zn_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.86$
$Cd_{1-x}Mn_xS$	W	$0 < x \leq 0.45$
$Cd_{1-x}Mn_xSe$	W	$0 < x \leq 0.50$
$Cd_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.77$
$Hg_{1-x}Mn_xS$	ZB	$0 < x \leq 0.37$
$Hg_{1-x}Mn_xSe$	ZB	$0 < x \leq 0.38$
$Hg_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.35$
II–V (Mn)		
$(Cd_{1-x}Mn_x)_3As_2$	tetr.	$0 < x \leq 0.18$
$(Zn_{1-x}Mn_x)_3As_2$	tetr.	$0 < x \leq 0.15$
IV–VI (Mn, Eu, Gd)		
$Pb_{1-x}Mn_xS$	RS	$0 < x \leq 0.05$
$Pb_{1-x}Mn_xSe$	RS	$0 < x \leq 0.17$
$Pb_{1-x}Mn_xTe$	RS	$0 < x \leq 0.12$
$Sn_{1-x}Mn_xTe$	RS	$0 < x \leq 0.40$
$Ge_{1-x}Mn_xTe$	RH	$0 < x \leq 0.18$
	RS	$0.18 < x \leq 0.50$
$Pb_{1-x}Eu_xSe$	RS	$0 < x \leq 0.04^*$
$Pb_{1-x}Eu_xTe$	RS	$0 < x \leq 0.32^*$
$Sn_{1-x}Eu_xTe$	RS	$0 < x \leq 0.013$
$Pb_{1-x}Gd_xTe$	RS	$0 < x \leq 0.11^*$
$Sn_{1-x}Gd_xTe$	RS	$0 < x \leq 0.09^*$
II–VI (Fe, Co, Cr)		
$Zn_{1-x}Fe_xS$	ZB	$0 < x \leq 0.26^*$
$Zn_{1-x}Fe_xSe$	ZB	$0 < x \leq 0.22^*$
$Zn_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.01^*$
$Cd_{1-x}Fe_xS$	W	$0 < x < 0.09^*$
$Cd_{1-x}Fe_xSe$	ZB	$0 < x < 0.20$
$Cd_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.06^*$
$Hg_{1-x}Fe_xS$	ZB	$0 < x \leq 0.037^*$
$Hg_{1-x}Fe_xSe$	ZB	$0 < x \leq 0.20^*$
$Hg_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.02^*$
$Zn_{1-x}Co_xS$	ZB	$0 < x \leq 0.15^*$
$Zn_{1-x}Co_xSe$	ZB	$0 < x \leq 0.10^*$
$Zn_{1-x}Co_xTe$	ZB	$0 < x < 0.06^*$
$Cd_{1-x}Co_xS$	W	$0 < x \leq 0.064^*$
$Cd_{1-x}Co_xSe$	W	$0 < x \leq 0.082^*$
$Cd_{1-x}Co_xTe$	ZB	$0 < x \leq 0.04$
$Hg_{1-x}Co_xS$	ZB	$0 < x \leq 0.02$
$Hg_{1-x}Co_xSe$	ZB	$0 < x \leq 0.047^*$
$Zn_{1-x}Cr_xSe$	ZB	$0 < x \leq 0.005^*$
$Zn_{1-x}Cr_xTe$	ZB	$0 < x \leq 0.003^*$
$Cd_{1-x}Cr_xS$	W	$0 < x \leq 0.002^*$

ZB – zinc blende, W – wurtzite, RS – rock salt, RH – rombohedral. Asterisks denote SMSs with the highest reported x value; tetr, tetragonal structure.

Reprinted in part from Gałązka RR (1995) Influence of electron subsystem on magnetic properties of semimagnetic semiconductors. In: Tróc R, Morkowski J and Szymczak M (eds) Proceedings of the International Conference of Magnetism, Warsaw 1994. *Journal of Magnetism and Magnetic Materials*, 113–116. Copyright Elsevier Science.

related to the magnetization of the sample:

$$M = N_0 g \mu \langle \bar{S} \rangle_{\mathrm{av}} \qquad [3]$$

N_0 is the number of unit cells per unit volume, g is the Lande factor for magnetic ions, and μ is the Bohr magneton.

The Heisenberg term must be added to the effective-mass Hamiltonian to solve the energy eigenvalue problem for SMSs. Realizing this was the turning point in understanding their properties and provided a basis for this group of alloys to be distinguished from other semiconductor mixed crystals. Since that time SMSs have become the subject of intensive studies in Europe and elsewhere.

The exchange interaction depends on the ion spin and the exchange constant. The spin of Mn is the highest for transition metals and equals five Bohr magnetons, and whilst a typical value for the exchange coupling between magnetic ions $J_{\mathrm{d-d}}$ is 10^{-3} eV, the exchange constant for electrons and holes with magnetic ions $J_{\mathrm{sp-d}}$ is about 1 eV in SMSs.

The exchange interaction is also strongly temperature and magnetic field dependent. As the macroscopic magnetization of a sample is proportional to the thermodynamic average value of the magnetic ion spin, to a good approximation we can replace the ion spin operators in the Hamiltonian by their average value, calculated or taken from measurements of the magnetization. From typical magnetic functions we can thus obtain information on the electron behavior.

In semiconductors, an external magnetic field acts on both the orbital motion and the spin of electrons producing Landau quantization and spin splitting, to an extent dependent on the effective mass of the carriers. The exchange interaction acts on their spin only and is, in effect, mass independent. Because of this, and also the large value of $J_{\mathrm{sp-d}}$, the band structure in turn changes drastically under the influence of an external magnetic field and depends strongly on temperature. From Figure 1 we can see that the very typical structure of the degeneracy of the light and heavy hole valence bands disappears in SMS when a magnetic field is applied. The change is accompanied by a drastic increase of hole mobility like that observed in low-dimensional structures. In SMS such as $Hg_{1-x}Mn_xTe$ or $Hg_{1-x}Mn_xSe$ (for $x \leq 0.07$) (Figure 2), the magnetic field produces an overlap of the valence and conduction bands, an effect never observed in nonmagnetic semiconductors.

Impurity levels are also influenced by exchange interactions: the ionization energy of an acceptor

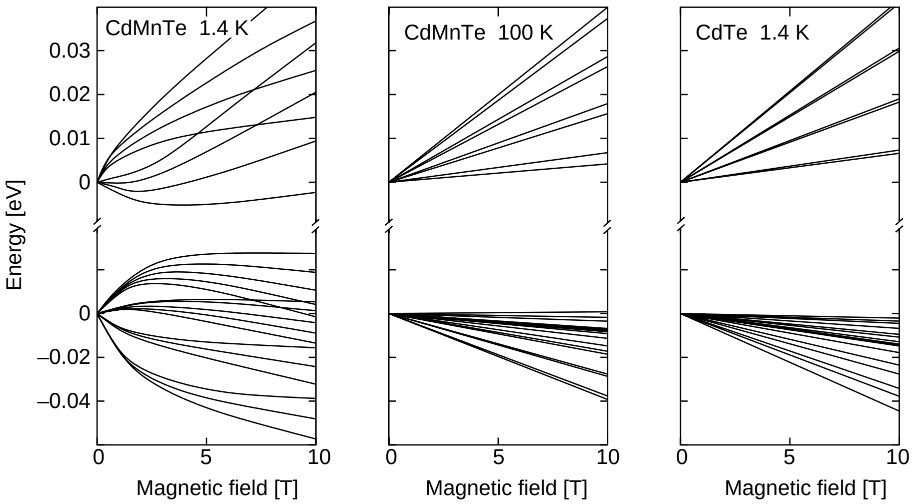

Figure 1 Conduction and valence band quantization in an external magnetic field for $Cd_{0.95}Mn_{0.05}Te$ and CdTe. Spin splitting, which is very weak in CdTe, is the main effect in CdMnTe. Notice the almost equal splitting of the heavy- and light-hole bands in CdMnTe. The influence of exchange interaction is visible even at 100 K. The picture of CdTe over this range of temperature is practically unchanged. Reproduced from Gał̨azka RR (1987) Semimagnetic semiconductors. *Europhysics News* 18(6): 90–92.

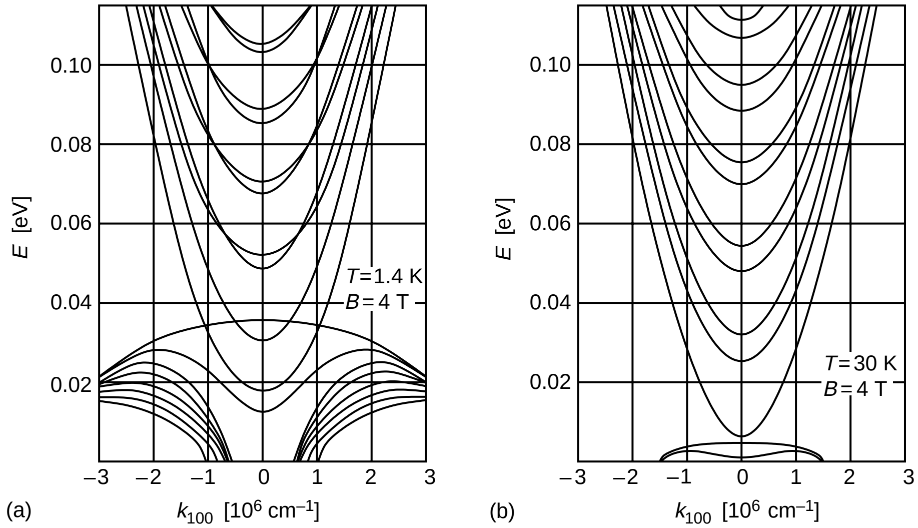

Figure 2 Energy band structure of $Hg_{1-x}Mn_xTe$, $x = 0.02$, calculated at 14 K (a) and 30 K (b) in a magnetic field of 4 T. After Gał̨azka RR (1978) Semimagnetic semiconductors based on HgMnTe and CdMnTe. Reproduced from Wilson BLM (ed) ICPS 14th, Edinburgh Institute of Physics, Conference series no. 43, 1978, p.133–140.

decreases and its wavefunction becomes anisotropic under the influence of a magnetic field. Both effects produce a giant negative magnetoresistance and a field-induced insulator–metal transition. Indeed all (particularly optical) properties susceptible to being changed by a magnetic field are very different in SMSs from those of standard semiconductors: the Faraday effect., Shubnikov–de Haas oscillations, interband magnetoabsorption, luminescence and spin flip Raman scattering (see below) become strongly dependent on temperature and magnetic field.

In the absence of a magnetic field, the average value of the ion spins is zero, the magnetization is zero too, and SMSs should behave as typical non-magnetic semiconductors. Whereas this is true for delocalized band electrons, when an electron is localized around an impurity, a bound magnetic polaron can be observed. This concept was first introduced to explain the behavior of magnetic semiconductors (Eu chalcogenides) and then later was applied also to SMSs. A localized electron can produce a spontaneous magnetization within the range of its wavefunction via the exchange interaction. Local ferromagnetic ordering in this range produces a Stokes shift in spin-flip Raman scattering even in the absence of an external magnetic field and other related phenomena. Local fluctuations of magnetization and a detailed knowledge of electronic states are essential for a theoretical description of this effect. The bound magnetic polaron is a subtle example of the feedback between electronic and magnetic subsystems present in SMS.

Magnetic Properties

Magnetic properties of SMS connected with the exchange interaction between paramagnetic ions are similar to those of other dilute magnetic materials. The broad range of possible magnetic ion concentrations (up to $x = 0.80$ in $Zn_{1-x}Mn_xTe$ and $x = 0.70$ for $Cd_{1-x}Mn_xTe$) allows us to observe the evolution of magnetic interactions in the same crystallographic structure as a function of magnetic ion concentration. A continuous transition from diamagnetic behavior (for $x \leq 0.005$) of the host semiconductor, to paramagnetism, spin-glass (for $x \geq 0.02$) and finally antiferromagnetic order has been observed. Type III antiferromagnetic ordering is observed for $x > 0.6$ indicating that the magnetic elementary cell contains two elementary cells of the cation sublattice. From a crystallographic point of view the magnetic subsystem is disordered; the distribution of magnetic atoms on the cation sublattice is random.

A spin-glass-like state is observed in SMSs in the broadest range of paramagnetic ion concentrations ($0.02 \leq x \leq 0.60$) below and above the percolation threshold (Figure 3). Indeed the temperature dependence of the magnetic susceptibility below and above the transition temperature and their dynamic properties at low and high magnetic fields suggest rather the existence of two spin-glass phases arising from the competition between spin-glass and clustering behavior. Below the percolation threshold ($x \leq 0.17$ for fcc lattice) the mechanism responsible for a spin-glass phase is even less clear, although, for such a dilute system it must be long-range to produce freezing of the spins.

The antiferromagnetic order has been inferred from specific heat and magnetic susceptibility measurements, and additional information concerning this phase has come from neutron diffraction studies: only a certain fraction of the total number of magnetic ions (e.g., about 50% for CdMnTe, $x = 0.65$) is well ordered; the rest remain in a disordered spin-glass phase. Thus a mixed phase (antiferromagnetic and spin-glass coexisting together) rather than a truly antiferromagnetic phase is observed.

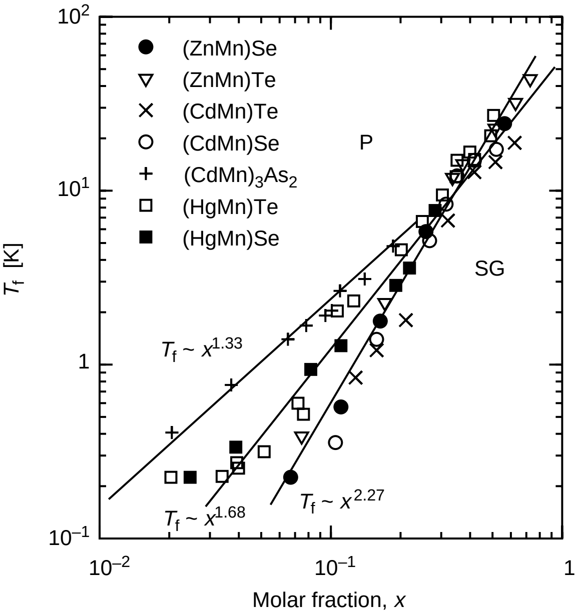

Figure 3 Magnetic phase diagram for the listed SMSs. P indicates the paramagnetic region, SG indicates the spin-glass region. Reprinted in part from Gałązka RR (1995) Influence of electron subsystem on magnetic properties of semimagnetic semiconductors. Reproduced with permission from Tróc R, Morkowski J and Szymczak M (eds) Proceedings of the International Conference of Magnetism, Warsaw 1994. *Journal of Magnetism and Magnetic Materials*, 113–116. Copyright Elsevier Science.

The influence of carrier concentration on magnetic properties has been extensively investigated in metallic alloys. Several IV–VI semiconductors like SnTe can exhibit quite high metallic-like carrier concentrations. In the quarternary alloy PbSnMnTe, the concentration of holes can be varied by annealing in the range 10^{20} to 5×10^{21} cm^{-3}. In $Pb_{0.25}Sn_{0.72}Mn_{0.03}Te$ an abrupt transition from the paramagnetic to ferromagnetic phase has been found for $p > 3 \times 10^{20}\ cm^{-3}$. The Curie temperature is a function of hole concentration, but the Mn ions conserve their magnetic moment of five Bohr magnetons. The magnetic phase diagram for this first ferromagnetic SMS (Figure 4) is thus three dimensional (T, x, n) in contrast to the

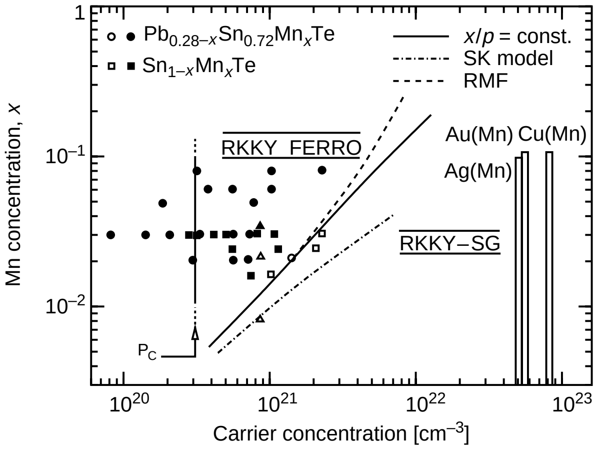

Figure 4 Magnetic phase diagram for PbSnMnTe (•) and SnMnTe (▪). The ferromagnetic phase is observed in the region above the critical carrier concentration p_c and the straight line defined by the equation x/p = const. Away from this region a spin-glass state is observed. Calculated phase boundaries between the ferromagnetic and spin-glass state are also shown for two theoretical models: SK: Sherrington–Kirkpatrick model and RMF: random mean field model. Canonical metallic spin-glasses are also shown. Reproduced from de Jonge WJM, Story T, Swagten HJM and Eggerkamp PJT (1992) Carrier-induced breakdown of ferromagnetism. *Europhysics Letters* 17: 631–636.

two-dimensional diagram (T,x) usually presented for magnetic alloys. Even a low concentration of spin polarized carriers when optically pumped can change the magnetization of HgMnTe owing to the strong exchange interaction of carriers with magnetic ions.

Low-Dimensional (LD) SMSs: Magneto-optical and Spintronic Effects

Modern technology such as MBE makes it possible to obtain quasi-3D SMSs in the whole range of x from 0 to 1. It also gives broad possibilities to create different combinations of semiconductor, SMS, and magnetic materials. In addition the dimensionality of LD structures influences the properties of both electronic and magnetic subsystems. Further there is almost always some strain in the system which also influences the properties of superlattices and other LD structures. Studies of LD SMSs not only extend the material base but open a new area of dimension-dependent phenomena and interactions in SMS systems.

Among many effects observed in LD SMS structures we will mention only a few phenomena, where exchange interaction is essential and which cannot be observed in nonmagnetic LD semiconductors. In all cases the giant (magneto-optical) Zeeman splitting caused by the exchange interaction plays a key role.

In superlattices made of ZnSe/ZnMnSe with a small amount of Mn, the band offset of conduction and valence bands is negligible at zero external magnetic field. In the presence of an external field the conduction and valence bands of ZnMnSe split creating a spin-superlattice because the spin-split states allow only one spin configuration. The possibility of the existence of such a spin superlattice was earlier suggested for SMSs and later experimentally observed.

If the superlattice, quantum well or heterostructure is made of SMS and non-SMS material (e.g. ZnTe/CdMnSe or CdTe/CdMnTe) one can exploit the drastic differences in the Zeeman splitting occurring in different layers to pinpoint the localization in space of specific electronic states. If one (hole or electron) or both states involved in optical transitions originate from the SMS layer, this state will show a very strong Zeeman splitting in the external magnetic field. Thus the SMS component can be used as a ‘marker’ in order to bring out the general properties of the wavefunction distribution.

Excitonic Zeeman splitting in non-magnetic quantum wells with SMS barriers can be applied to studies of the shape of interface profile and the influence of the growth conditions on the shape of the interface. If the interface is not completely flat, even a small fraction of magnetic atoms in the well close to the interface can significantly enhance the Zeeman exciton splitting in the quantum well. The sensitivity of this method allows one to detect the sharpness of an interface with one monolayer accuracy.

The MBE technique opens up possibilities of creating hybrid structures with SMSs, useful for spintronic devices such as spin filters, spin transistors and tunneling magnetoresistance sensors. Very importantly, it has been realized that SMSs are highly suitable to form the ferromagnetic spin injecting electrodes in such systems where simple metallic electrodes fail because of the large electrical resistance mismatch between the electrode and the semiconductor in which the spin is being processed. A further possibility is to fabricate hybrid metallic ferromagnetic–SMS structures. In such a case one can make use of the magnetic field exerted by the metallic ferromagnet positioned very close (on an atomic scale) to the SMS layer or quantum structure, thus affecting the spin splitting in the latter. One could use such a fringe field to shape the band edges in such a fashion that an additional localization of the carriers in a two-dimensional quantum well is induced leading to formation of quantum wires or dots. In other applications a spin-coherence storage effect has been obtained for core electron polarization in Mn-doped ZnSe quantum wells. The optically pumped conduction electron spin polarization generates a nonequilibrium spin polarization of the paramagnetic core electron spins(3d in the case of Mn) through the core–conduction-electron coupling. This core electron spin polarization persists long after the conduction electron spin polarization has decayed away and presents an interesting avenue towards information storage.

The physics of SMSs is currently a well-established part of solid state physics. In the investigations of bulk crystals and LD structures many problems are solved and new problems are waiting further research. Particularly important seems to be investigation of layers and LD structures because of the broad technological possibilities and because of the new important factor (i.e., dimensionality) whereby spin-dependent phenomena can be intentionally regulated.

List of Units and Nomenclature

g	Lande factor for paramagnetic ion
H_{ex}	Heisenberg spin exchange Hamiltonian
$J_{\mathrm{d-d}}$	exchange constant for coupling between paramagnetic ions (transition metals)
$J(\bar{r}-\bar{R})$	exchange constant
$J_{\mathrm{sp-d}}$	exchange constant for electron or hole coupling with paramagnetic ions
LD	low dimensional
M	magnetization
MBE	molecular beam epitaxy
N_0	number of unit cells per unit volume
SMS	semimagnetic (diluted magnetic) semiconductor
$\langle\bar{S}\rangle_{\mathrm{av}}$	average value of spin operator
T_{c}	Curie temperature
V_{h}	hybridization energy
x	molar fraction of paramagnetic ions
$\boldsymbol{\mu}$	Bohr magneton
$\bar{\boldsymbol{\sigma}}, \bar{S}$	spin operators

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Materials:** Lead Salts; Mercury Cadmium Telluride. **Semiconductor Physics:** Spin Transport and Relaxation in Semiconductors; Spintronics.

Further Reading

Bauer G, Pascher H and Zawadzki W (1992) Magneto-optical properties of semimagnetic lead chalcogenides. *Semiconductor Science and Technology* 7: 703–723.

Dietl T (1994) Magnetic semiconductors. Moss TS (ed.) *Handbook of Semiconductors*, vol. 3b, pp. 1251–1342. Amsterdam, The Netherland: North-Holland.

Dietl T (2002) Ferromagnetic semiconductors. *Semiconductor Science and Technology* 17: 377–392. (Special issue: Semiconductor Spintronics, guest ed Hideo Ohno).

Dobrowolski W, Kossut J and Story T (2003) Magnetic semiconductors–new bulk materials and low-dimensional quantum structures. In: Buschow KHJ (ed.) *Handbook of Magnetic Materials, II–VI and IV–VI.* vol. 15, pp. 289–378. Amsterdam: Elsevier North-Holland.

Furdyna JK (1991) In: Averous M and Balkanski M (eds) *Semimagnetic Semiconductors and Diluted Magnetic Semiconductors*. New York: Plenum.

Furdyna JK and Kossut J (eds) (1988) Semiconductors and Semimetals. *Diluted Magnetic Semiconductors*, vol. 25, 470 pp. San Diego, CA: Academic Press.

Gałązka RR and Kossut J (1980) Narrow-gap semimagnetic semiconductors. In: Zawadzki W (ed.) *Lecture Notes in Physics*, vol. 133, pp. 245–265. Berlin: Springer Verlag.

Gałązka RR and Kossut J (1982) Semimagnetic semiconductors. In: Madelung O, Schulz M and Weiss H (eds) *Landolt-Börnstein New Series*, vol. III/17b, pp. 302–318. Berlin: Springer Verlag.

Gałązka RR, Kossut J and Story T (1999) Semimagnetic semiconductors. In: Rössler U (ed.) *Landolt-Börnstein New Series*, vol. III/41B, pp. 650–721. Berlin: Springer Verlag (Update to chapter 5, Vol. III/17b).

Kossut J and Dobrowolski W (1993) Diluted magnetic semiconductors. In: Buschow KHJ (ed.) *Handbook of Magnetic Materials*, vol. 7, pp. 231–305. Amsterdam, The Netherlands: North Holland.

Kossut J and Dobrowolski W (1997) Properties of diluted magnetic semiconductors. In: Capper P (ed.) *Narrow-Gap II–VI Compounds for Optoelectronics and Electromagnetic Applications*, pp. 401–429. London: Chapman and Hall.

Kossut J (2001) Low dimensional structure of diluted magnetic (semimagnetic) semiconductors – a subjective review. *Acta Physica Polonica, A* 100 (supplement): 111–138.

Ohno H, Matsukura F and Ohno Y (2002) Semiconductor spin electronics. *Japan Society of Applied Physics (JSAP International)* 5: 4–13.

Story T (2003) Optoelectronic properties. In: Monasreh MO (series ed.) *Optoelectronic Properties of Semiconductors and Superlattices* vol. 18, Khokholov D. (volume ed.) *Lead Chalcogenides: Physics and Applications, Semimagnetic Semiconductors Based on Lead Chalcogenides*, pp. 385–425. New York, London: Taylor and Francis Books Inc.

GaAs Based Compounds

J J Finley and J P R David, University of Sheffield, Sheffield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Background

Gallium Arsenide (GaAs) is an important semiconductor that has come to dominate the field of optoelectronics by virtue of its favorable electro-optical properties and the ease by which it can be controllably modified by extrinsic means; combining it with its large family of related alloys ($Al_xGa_{1-x}As$, $In_xGa_{1-x}As$, $GaAs_xP_{(1-x)}$, $(In_xGa_{(1-x)})_yAsN_{1-y}$) and via the growth of hetero-structures with reduced dimensionality (Quantum Wells-2D, Wires-1D, and Dots-0D). In such low dimensional semiconductor nanostructures, quantum mechanical effects are directly exploited to customize and dramatically enhance electro-optical properties. Such investigations of the fundamental electro-optical properties of GaAs-based compounds has resulted in the emergence of several new device concepts such as the Quantum Cascade laser, led to the discovery of new physical phenomena like the fractional Quantum Hall effect, and may pave the way for an entirely new class of opto-electronic devices with true quantum functionality.

In the following, we explore the fundamental physical and electro-optical properties of GaAs and its related alloys and illustrate the considerable impact this remarkable family of materials has had on modern semiconductor devices.

Fundamental Physical Properties

The crystal structure of GaAs (Figure 1 – inset) consists of two interpenetrating face centred cubic lattices that accommodate the group III and V atoms respectively. The sub-lattices of Ga and As are shifted from each other by $a/4$ along the [111] body diagonal, where a is the lattice constant, forming the well-known zincblende crystal structure. For stochiometric GaAs, the lattice constant is $a = 0.56536(1)$ nm at room temperature which translates to a Ga–As nearest neighbor separation (bond length) of 0.2308 nm.

Figure 1 summarizes the variation of the lattice constant for a number of GaAs related III–V semiconductor alloys, plotted as a function of their fundamental electronic bandgap. Particularly relevant is AlAs, which also crystallizes in the zincblende form and has a lattice constant ($a(\text{AlAs}) = 0.5660(2)$ nm) almost identical to that of GaAs ($\Delta a < 1\%$). The sole fact that the $Al_xGa_{(1-x)}As$ alloy can be almost perfectly lattice matched to GaAs over its entire compositional range, whilst exhibiting near continuously variable electronic and optical properties, is the principal reason for the key role GaAs-related materials play in semiconductor opto-electronic devices. We continue by investigating the electronic and optical properties of the $Al_xGa_{(1-x)}As$ family of materials and discover why they are particularly suited to opto-electronic device applications.

Summary of Electronic Properties of GaAs and $Al_xGa_{1-x}As$

The optical properties of semiconductors are determined principally by inter-band radiative transitions involving electronic states in the vicinity of the conduction (CB) and valence band (VB) extrema. Consequently, in the following we focus on electronic properties of GaAs close to these points and substantiate our discussion using experimental data obtained from optical spectroscopy.

GaAs Bandstructure

GaAs is a direct-gap semiconductor ($E_g = 1.42$ eV at 300 K), both the conduction band minimum and

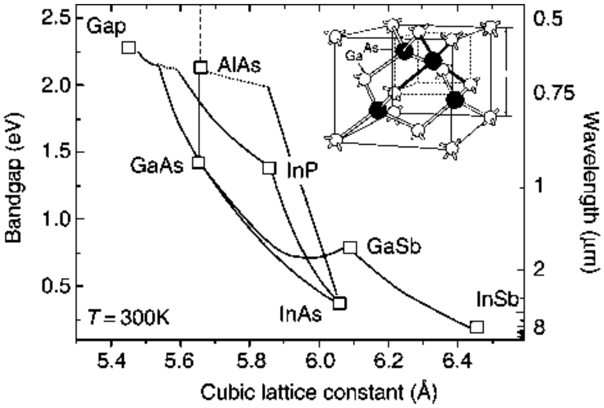

Figure 1 Fundamental bandgap of selected compounds within the III–V semiconductor family plotted as a function of the cubic lattice constant – *a*. The inset depicts the zincblende crystal structure within the GaAs unit cell. Data obtained from various sources, see for example Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16 and Loffe Institute, on NSM semiconductor data repository at http://www.ioffe.rssi.ru

valence band maximum occurring at the center of the Brillouin zone (Γ-point). Figure 2 summarizes our discussion of the electronic structure of GaAs, depicting the bandstructure schematically within the first Brillouin zone. The valence electron configurations of Ga and As atoms result in the band edge CB states being two-fold degenerate due to the electron spin and characterized by an almost isotropic effective mass of $m_\Gamma^* = 0.068\, m_o$. For GaAs heterostructures with reduced dimensionality, the very weak anisotropy of the conduction band states can generally be ignored when calculating the lowest energy confined electron states but has a significant impact on higher quantized levels and must, therefore, be included. In addition to the fundamental CB minimum at the zone center or Γ-point of the Brillouin zone, satellite minima occur at the six equivalent X-points ($k = (1/a, 0, 0)$) and eight equivalent L-points ($k = (1/a, 1/a, 1/a)$). The X-point and L-point minima lie $\Delta E_{\Gamma X} = 0.49$ eV and $\Delta E_{\Gamma L} = 0.32$ eV above the Γ-minimum at low temperatures ($T = 4.2$ K) and are characterized by strongly direction dependent effective masses parallel (||) or perpendicular (⊥) to the principal axes. Electronic states in the vicinity of these minima are termed the X and L-valleys respectively and have a significant impact on high-energy electron transport in devices such as HEMTs and in avalanche carrier multiplication processes, but do not generally play a significant role in the optical properties of the binary – GaAs. However, satellite CB minima can strongly influence the optical properties of GaAs–AlAs superlattices and QWs that can exhibit a type-II band alignment for GaAs layer thicknesses < 6 nm. In this case, the electrons (holes) are localized in the AlAs (GaAs) layer and the nature of the AlAs X-states dominates the optical properties.

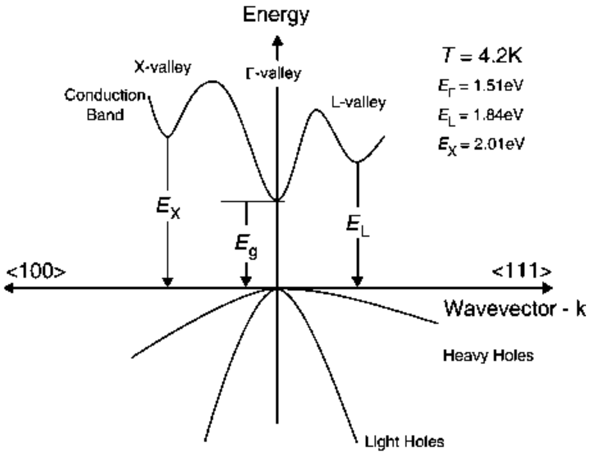

Figure 2 Schematic representation of the bandstructure of GaAs showing just a single conduction band over whole Brillouin zone in Γ-L and Γ-X directions only and neglecting the split off band that does not significantly influence electrical or optical properties.

Table 1 Low temperature GaAs effective masses for electrons (upper panel) and holes (lower panel). Values marked with * are inferred from measurements on GaAs based heterostructures and values marked ** are calculated

	Crystallographic axis	*Effective Mass (m_o)*
CB-Γ minimum	*Isotropic*	0.068(2)
CB-X minimum*	|| [100]	1.9
	⊥ [100]	0.19
CB-L minimum**	|| [111]	1.9
	⊥ [111]	0.075
VB-HH	[111]	0.72(3)
	[100]	0.33(6)
VB-LH	[111]	0.08(3)
	[100]	0.09(4)

The valence band (VB) states of GaAs are four-fold degenerate at the bandedge and, characteristic of most zincblende semiconductors the valence bandstructure, is significantly more strongly anisotropic than the CB states discussed above. Away from the Γ-point, the four-fold degeneracy of the valence band states is lifted into 2 doubly degenerate sets of bands termed the heavy (HH) and light hole (LH) bands by virtue of their zone center effective masses (Table 1). The HH band is considerably more strongly anisotropic than the LH band and is characterized by a much larger effective mass along the [100] direction when compared with [001], $m_h^*[111] > m_h^*[001]$. The anisotropy of the LH states is much weaker than for the HH band, with approximately equivalent effective masses along [100] and [111] crystallographic directions. Typical values for the CB and VB effective masses are summarized in Table 1, obtained using a variety of experimental techniques at low temperature.

Temperature Dependence

Quantitative values given above relate specifically to the bandstructure at liquid helium temperatures ($T = 4.2$ K). Increasing temperature results in a narrowing of many of the conduction–valence band gaps discussed above and lowering of the effective masses. This temperature dependence of the fundamental bandstructure arises from the anharmonic dilation of the lattice and from interactions with phonons, the population of which becomes significant at elevated temperature. In the simplest picture, one can visualize the effect of increasing temperature

Table 2 Varshnii coefficients for the Γ, L, and X CB minima in GaAs relative to the top of the VB. For the Γ minimum a number of experimental works are available and the value cited is the average

	$\alpha(10^{-4}$ eV/K)	β (K)
Γ point	5.45	210
X point	7.20	205
L point	6.05	204

as a shrinking of the first Brillouin zone as the lattice constant increases, and with a subsequent increase of the band curvature leading to ~5–10% lighter effective masses at room temperature than at 4.2 K. In addition, the dramatic increase of the phonon population at elevated temperatures results in a narrowing of the fundamental bandgaps. Phenomenologically, the temperature dependence of the fundamental bandgaps can be well described by the well-known Varshnii relation

$$E_g(T) = E_g^{T=0\text{ K}} - \frac{\alpha T^2}{(T+\beta)} \qquad [1]$$

where α and β are experimentally determined constants which are sensitive to the bandgap of interest (e.g. Γ, X, or L, etc.) as summarized in Table 2.

GaAs/Al$_x$Ga$_{(1-x)}$As – The Most Significant Alloy

Incorporation of aluminum into the GaAs crystal to form Al$_x$Ga$_{(1-x)}$As results in a number of profound and important modifications of the electronic properties compared with pure GaAs. Aluminum is more strongly electronegative than Gallium and consequently the electron charge distribution along the Al–As bonds are more strongly localized towards the group-III site than for Ga–As bonds which qualitatively results in a widening of the fundamental bandgap with increasing aluminum composition (x).

The variation of the Al$_x$Ga$_{(1-x)}$As energy gaps for the Γ, X, and L conduction band minima at low temperature are summarized in Figure 3 as a function of the alloy composition up to $x \sim 0.8$. The data of Figure 3 reflect the results of optical emission and absorption measurements and are most reliable for more dilute compositions since crystalline AlAs and high Al content Al$_x$Ga$_{(1-x)}$As oxidize rapidly under atmospheric conditions.[1] With increasing Al content, the Γ-minimum shifts rapidly to higher energy whilst the shift of the L and X-minima is significantly less pronounced. At $x \sim 0.4$ the Γ and X minima intersect and for $x > 0.4$, the lowest energy conduction band minimum shifts to the X-point and the fundamental bandgap becomes indirect. This effect is responsible for the short wavelength cutoff of Al$_x$Ga$_{(1-x)}$As/GaAs based multi-quantum well LEDs at $\lambda \sim 630$ nm since the transition to an indirect bandgap is accompanied by a drastic reduction of the radiative efficiency. Shorter wavelength LED operation region of the spectrum is possible by moving to other related alloys in the III–V family such as (Al$_x$Ga$_{1-x}$)$_{0.5}$In$_{0.5}$P (520–630 nm) and (GaAs)$_{0.6}$P$_{0.4}$ (~600 nm). We return to discuss the optical properties of Al$_x$Ga$_{(1-x)}$As below in Section 4.

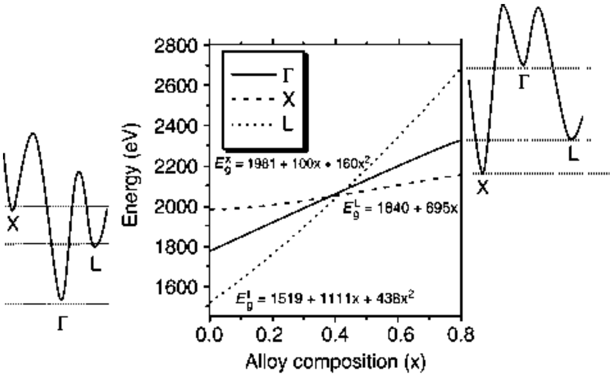

Figure 3 The absolute energy (at 4.2 K) of the Γ, X and L CB minima for Al$_x$Ga$_{(1-x)}$As as a function of aluminum composition up to $x \sim 0.8$. Higher compositions are not given due to the well-known unreliability of experimental data due to instability of high Al content Al$_x$Ga$_{(1-x)}$As due to oxidation. Polynominals describing each curve are also plotted. Data taken from the following sources: Γ-minimum, Bosio C, Staehli JL, Guzzi M, Burri G and Logan RA (1988) Direct-energy-gap dependence on Al concentration in Al$_x$Ga$_{1-x}$As. *Physical Review* 38: 3263–3268, X-minimum, Guzzi M, Grilli E, Oggioni S, Staehli JL, Bosio C and Pavesi L (1992) Indirect-energy-gap dependence on Al concentration in Al$_x$Ga$_{1-x}$As alloys. *Physical Review* B45: 10951, and L-minimum, Henning JCM, Ansens JPM and Roksnoer PJ (1986) Spectroscopic determination of L6 conduction-band minima in Al$_x$Ga$_{1-x}$As. *Journal of Physical Chemistry*, L335.

As expected, the pronounced modifications of the bandgap induced by Al incorporation are accompanied by modifications of the carrier effective masses and temperature dependence of the bandgap. These compositional dependence of the Γ, X, and L conduction band and HH and LH valence band effective masses are summarized in Table 3.

It should be stressed that the relationships given in Table 3 should be considered to be accurate at the

---

[1] The ease by which high Al content Al$_x$Ga$_{(1-x)}$As can be oxidized is of great technological benefit for device fabrication. Examples include the formation of current apertures in injection lasers, Al$_x$Ga$_{(1-x)}$As etch stops etc. For a summary Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16.

Table 3 Compositional variation of the electron and hole effective masses for $Al_xGa_{(1-x)}As$ taken from various sources: Data drawn from the following sources: Lautenschlager Γ-P, Garriga M, Logothetidis S and Cardona M (1997) *Physical Review* B35: 9174 and Grilli E , Guzzi M and Zamboni R (1992) *Physical Review* B45: 1638. Thurmond CD (1975) L-minimum. *Journal Electrochemical Society*, 122: 1133 and Aspines DE (1976) L-Minimum. *Physical Review* B14: 5331

	Crystallographic Axis	*$Al_xGa_{(1-x)}As$ Effective Mass (m^0)*
CB-Γ minimum	*Isotropic*	$0.0682 + 0.0835x$
CB-X minimum	‖ [100]	$1.9 - 1.02x$
	⊥ [100]	$0.19 + 0.06x$
CB-L minimum	‖ [111]	$1.9 - 0.58x$
	⊥ [111]	$0.075 + 0.075x$
VB-HH	DOS	$0.62 + 0.14x$
VB-LH	DOS	$0.087 + 0.09x$

~5–10% level, since reliable experimental data only exists for the GaAs binary endpoint and more dilute alloy compositions ($x < 0.3$).

The Varshnii coefficients of the temperature dependence of the bandgap vary with $x(0 < x < 0.5)$ according to $\alpha(x) = 5.5 + 3.35x(10^{-4}\ \mathrm{eV/K})$ and $\beta(x) = 225 + 88x(\mathrm{K})$ respectively. Recent photoreflectance measurements have indicated that these simple linear relationships do not hold well over the entire compositional range with evidence for some bowing of $\alpha(x)$ and $\beta(x)$ closer to the AlAs binary endpoint.

GaAs/$Al_xGa_{(1-x)}As$ Hetero-interfaces

Modern epitaxial growth techniques now enable the routine fabrication of semiconductor heterostructures consisting of a well-defined sequence of atomically precise layers of different materials, which together manipulate the opto-electronic properties to suit a particular device or application. Of particular technological relevance is the GaAs/$Al_xGa_{(1-x)}As$ hetero-interface since both materials are lattice matched for $0 < x < 1$ (Figure 1), and multiple layers of GaAs and $Al_xGa_{(1-x)}As$ can be readily deposited without introducing defects into the crystal which seriously degrade optical performance. In this section, we summarize the most salient features of the GaAs/$Al_xGa_{(1-x)}As$ hetero-interface, with particular emphasis on the electronic properties before continuing to discuss the optical properties of $Al_xGa_{(1-x)}As$ in Section 4.

The most important parameters that characterize a semiconductor hetero-interface are the magnitude and sign of the conduction (ΔE_c) and valence band (ΔE_v) offsets. These parameters determine how effectively electrons and hole are localized within specific epilayers and are related to the total difference in the bandgap ($\Delta E_g = E_g^{\mathrm{AlGaAs}} - E_g^{\mathrm{GaAs}}$) via $\Delta E_g = |\Delta E_c + \Delta E_v|$. Taking the mean of 21 separate optical and electrical investigations between 1985 and 1995, the currently accepted value for the band-offset ratio is $\Delta E_c/\Delta E_v \approx 64/36$, with an associated variance over the body of the data of approximately ±8%. It should be stressed that the majority of the experimental investigations of the band offsets at the GaAs–$Al_xGa_{(1-x)}As$ hetero-interface have been performed for alloy compositions $x < 0.4$ and much less experimental data exists regarding the variation of $\Delta E_c : \Delta E_v$ with higher Al composition. A rather old review (*c.* 1992), collating data available, at the time has indicated that the absolute energies of the conduction and valence band offsets vary with composition as $\Delta E_c = (800 \pm 30)x$ meV for $(x < 0.4)$ and $\Delta E_c = (510 \pm 40)x\ (0 < x < 1)$. It is widely assumed that the band offsets do not change appreciably with temperature.

Optical Properties of GaAs and Significant Alloys

Optical Absorption

Monochromatic light of energy $E_p = \hbar\omega$ incident on the surface of GaAs may stimulate interband optical transitions in which a photon is absorbed and an electron from the valence band is promoted into the conduction band. In processes involving a single photon, the photon momentum is entirely negligible compared with the electron and optical transitions are vertical in k-space.

Following excitation, the electron–hole (e–h) pairs relax extremely rapidly ($\tau_{LO} \sim 150$ fs) by sequential emission of multiple longitudinal optical (LO) phonons until their energy becomes close to the bandedge and further LO-phonon emission is forbidden by energy conservation requirements. This initial phase of the carrier relaxation can be extremely rapid, the LO-phonon scattering time in GaAs being due to the semi-polar nature of the crystal. After this initial rapid relaxation phase, the electron and hole populations relax principally via acoustic phonon emission ($\Delta E < 2$ meV) over much longer timescales ($\sim 10-300$ ps), which depend upon the lattice temperature, carrier type and density. Finally, the electron and hole populations form thermal distributions via carrier–carrier and Auger scattering over longer timescales up ~ 500 ps. During these final relaxation and thermalization phases, e–h pairs are depleted from the conduction and valence bands by interband radiative recombination.

For energies less than the bandgap, the semiconductor normally appears transparent, although the presence of impurities and defect levels can result in a weak absorption 'tail' below the gap. Optical absorption by impurities results in only electrons in the conduction band or holes in the valence band. The wavelength of light that just causes the creation of an electron–hole pair is given by $\lambda = hc/E_g$, where h is Planck's constant, c the speed of light and E_g the fundamental bandgap. As described in Section 3.1, this corresponds to approximately $\lambda = 870$ nm at room temperature, depending on the type and level of the doping. Light with wavelength much shorter than this, corresponding to energies much larger than E_g, will create electron–hole pairs higher up the band structure with higher energy and be strongly absorbed. These carriers, however, will rapidly lose this excess energy by interacting with the lattice and relax to the minima of the conduction and valence bands.

The optical absorption is usually quantified by the absorption coefficient, α, which is a function of the photon energy or wavelength and is shown for GaAs in Figure 4. The fraction of the light intensity that is absorbed in the semiconductor after passing through a thickness x is given by $f = (1 - \exp(-\alpha x))$ and the fraction of light intensity that is transmitted is then simply $\exp(-\alpha x)$. α is usually expressed in units of cm^{-1} and at energies above the bandgap has values typically between $10^4-10^5\ \text{cm}^{-1}$. Light with wavelength corresponding to energies above the bandgap is usually absorbed very efficiently (>90%) within a few microns. Increasing the bandgap, say by adding aluminum, would have the effect of making the material more transparent at a given wavelength.

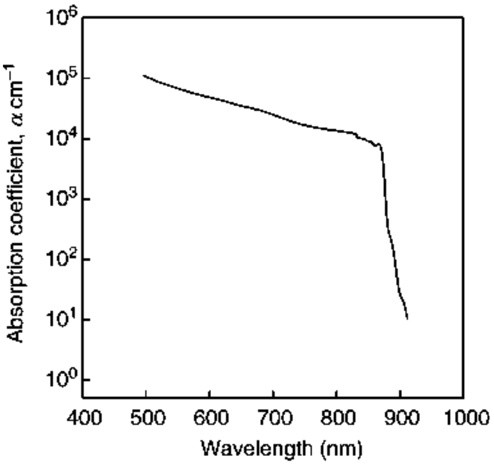

Figure 4 Variation of the absorption coefficient, α, as a function of wavelength in GaAs at room temperature.

Optical absorption is an important process as it allows light to be converted to free carriers, and hence a current in the semiconductor. This forms the basis of optical detection by most types of semiconductor photodetectors.

Refractive Index

The speed (more precisely the phase velocity) of light, of all wavelengths, in a vacuum is constant at $\sim 3.10^8\ \text{m s}^{-1}$. The speed of light in a medium such as a semiconductor is slower due to the microscopic interaction of the electromagnetic wave with the crystal. The refractive index, R_I, is defined as the ratio of the speed in a vacuum to the speed in the semiconductor. Increasing the density of the medium decreases the speed. Changing the semiconductor, for example by adding aluminum, not only increases the bandgap as discussed above, but also decreases R_I for a given wavelength.

At wavelengths corresponding to an energy just below the bandgap of the semiconductor, the material is essentially transparent, i.e., the absorption coefficient is low.

Light can be made to travel in a waveguide comprising this semiconductor, provided that it was surrounded by semiconductor material of a lower R_I. Light is totally internally refracted at the interface between the GaAs and $Al_xGa_{1-x}As$ and so cannot escape. The semiconductors with a smaller R_I usually have a larger bandgap, so are even more transparent at the operating wavelength and help to physically confine the electrons and holes within the GaAs.

The R_I can be changed by injecting charge (electrons or holes) into the semiconductor, by changing the temperature (which changes the bandgap), and by applying an electric field. This forms the basis of operation of devices such as optical modulators. Unfortunately it also can cause 'chirping', i.e., the unwanted change in light intensity or wavelength as the charge or temperature changes.

The R_I which is a function of the wavelength, is sometimes expressed as a complex number:

$$R_I = n + ik \qquad [2]$$

where n is the real part and k is the imaginary part called the extinction coefficient, which is related to the absorption coefficient. Typical room temperature values of n for GaAs vary from about 3.37 at ~1.55 μm to 3.87 at 620 nm.

Common GaAs Based Alloys for Novel Technological Applications

GaInP/AlGaInP – 'Red, Orange and Green Emission'

The alloy $(Al_xGa_{1-x})_{0.52}In_{0.48}P$ is becoming technologically important as it has the same lattice parameter as GaAs for all values of x, so can grow lattice matched on GaAs. This alloy system offers the largest direct bandgap of virtually any III–V (excluding the Nitride family), so has attracted considerable interest for visible lasers and LEDs in the wavelength range 530 nm–670 nm. At room temperature the direct Γ bandgap increases almost linearly from ~1.85 eV for $Ga_{0.52}In_{0.48}P$ to ~2.3 for $(Al_{0.5}Ga_{0.5})_{0.52}In_{0.48}P$. For values of $x > 0.5$, the bandgap becomes indirect as in $Al_xGa_{1-x}As$ ($x > 0.4$) when the X-level indirect bandgap becomes lower than the Γ-level direct bandgap. Earlier work in this alloy suggested that the point at which it became indirect was at $x = 0.66$. The X-level bandgap appears to be relatively insensitive to the aluminum composition and has a value of ~2.3 eV at room temperature.

GaAs/GaInAsN – '1.3 μm and 1.55 μm for telecommunication applications'

One limitation of GaAs is that most of the compounds that are lattice matched to it, such as $Al_xGa_{1-x}As$ or $(Al_xGa_{1-x})_{0.52}In_{0.48}P$, have bandgap energies that are large, making them unsuitable for telecommunication applications which require narrow bandgap materials capable of operating in the range 1.3 μm to 1.55 μm. Adding indium (In) to GaAs can decrease the bandgap but also causes strain as the lattice parameter increases, which limits the amount of indium and total thickness of the structure.

Recently, it has been shown that while adding In to GaAs increases the lattice parameter, adding N to GaAs decreases it. Consequently GaInAsN can be made to be lattice matched to GaAs by controlling the In and N content. Unlike conventional alloy semiconductors, adding In or N both result in a reduction of the bandgap energy. Furthermore, GaInAsN is found to be a direct bandgap material with a type I band lineup when combined with $Al_xGa_{1-x}As$, which is a wide bandgap material also lattice matched to GaAs. These properties make GaInAsN ideally suited for lasers and there have been several reports of devices operating at 1.3 μm in the literature.

See also

Semiconductor Physics: Quantum Wells and GaAs-based Structures.

References

Bosio C, Staehli JL, Guzzi M, Burri G and Logan RA (1988) Direct-energy-gap dependence on Al concentration in $Al_xGa_{1-x}As$. *Physical Review* 38: 3263–3268.

Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16.

Data drawn from the following sources: Lautenschlager Γ-P, Garriga M, Logothetidis S and Cardona M (1997) *Physical Review* B35: 9174 and Grilli E, Guzzi M and Zamboni R (1992) *Physical Review* B45: 1638. Thurmond CD (1975) L-minimum. *Journal Electrochemical Society*, 122: 1133 and Aspines DE (1976) L-Minimum. *Physical Review* B14: 5331.

Guzzi M, Grilli E, Oggioni S, Staehli JL, Bosio C and Pavesi L (1992) Indirect-energy-gap dependence on Al concentration in $Al_xGa_{1-x}As$ alloys. *Physical Review* B45: 10951.

Henning JCM, Ansens JPM and Roksnoer PJ (1986) Spectroscopic determination of L6 conduction-band minima in $Al_xGa_{1-x}As$. *Journal of Physical Chemistry*, L335.

Ioffe Institute, on NSM semiconductor data repository at http://www.ioffe.rssi.ru/SVA/NSM/Semicond/index.html

Landolt-Börnstein (1986) *Numerical data and functional relationships in science and technology.* In: Madelung O (ed.) *Semiconductors*, vol 22. Berlin: Springer Verlag.

Nakwaski W (1995) InAs, AlAs and their ternary compounds. *Physica* B210: 1–25.

Varshni YP (1967) Temperature dependence of the energy gap in semiconductors. *Physica* B34: 149.

Group IV Semiconductors, Si/SiGe

D J Paul, University of Cambridge, Cambridge, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

SiGe is now a mature materials system with electronic circuits on the market place. While no optical devices are available yet, there is a substantial amount of research in the field. In this article, the fundamental materials properties will be reviewed. As SiGe is a strained system, we shall start with the basic material properties as this determines the Ge content and layer thicknesses which are available for use. The band structure and electronic behavior will then be reviewed before the optical properties of the material.

The main drive for SiGe research in optical systems is the ability to use the extensive Si processing knowledge and to integrate a number of components onto a Si chip. Conventional Si transistors may also be used as driver circuits for the optical components, reducing chip counts in systems and therefore reducing costs. As Si processing is substantially cheaper than other semiconductor materials, if SiGe devices can be made on Si processing lines, they have the potential to be substantially cheaper than comparable III–V or II–VI devices. The SiGe material properties that will be reviewed in this article should be considered as an expansion of bulk silicon.

Material Properties and Strain

Ge has a 4.2% larger lattice constant that Si (Table 1). When growing epitaxial films of germanium or the alloy $Si_{1-x}Ge_x$ $(0 < x < 1)$ on a silicon substrate, there exists a maximum thickness called the critical thickness above which it costs too much energy to strain additional layers of material into coherence with the substrate. Instead defects appear, in this case misfit dislocations, which act to relieve the strain in the epitaxial film. Relaxed Si, Ge or SiGe films form diamond, diamond and zincblende lattices respectively while strained heterolayers below the critical thickness form a tetragonal symmetry arrangement. The strain therefore will substantially modify the electronic and optical properties of the material as will be discussed later.

Most SiGe films are produced by either chemical vapor deposition or molecular beam epitaxy. Both these techniques allow the deposition of films at low temperatures (typically 670–970 K) which is essential to prevent strained layers relaxing and forming defects. Chemical vapor deposition uses a number of different gases such as silane, disilane, dichlorosilane, and germane (occasionally with a hydrogen carrier gas) as the main species, while phosphine or arsine are used as n-type dopants and diborane is the typical p-type dopant. With the high melting point of both silicon and germanium, molecular beam epitaxy requires electron beam evaporators for the host materials silicon and germanium while effusion cells of antimony (occasionally phosphorus) and boron are used for n- and p-type dopants, respectively.

There are a number of models used to calculate the equilibrium critical thickness. The most often quoted is the Matthews and Blakeslee model (Figure 1). Experimentally it has been shown that defect-free films with thicknesses above the critical thickness may be grown (Figure 1) although these layers are metastable and will relax if temperatures above the

Table 1 Materials properties of Si, Ge and relaxed $Si_{1-x}Ge_x$

Element		*Si*	*Ge*	*Bulk relaxed* $Si_{1-x}Ge_x$
Lattice		Diamond	Diamond	Zincblende
Lattice constant	a_o (nm)	0.5431	0.5657	$0.5431 + 0.0226x$
Lattice mismatch	η_o (%)	0	4.17	$0.04115x - 0.02500x^2 + 0.05826x^3 - 0.03379x^4$
Density	(kg m^{-3})	2329	5323	$2329 + 2994x$
Sound velocity	m s^{-1} (300 K)	8440	4900	$8440 - 3540x$
Poisson s ratio	ν	0.280	0.273	$0.280 - 0.007x$
Melting point	(K)	1685	1210	
Elastic moduli	c_{11} (GPa)	165.8	128.5	
Elastic moduli	c_{12} (GPa)	63.9	48.3	
Elastic moduli	c_{44} (GPa)	79.6	66.8	
Thermal conductivity	κ (W m^{-1} K^{-1}) (300 K)	140	60	
Dielectric constant	ε	11.9	16.2	$11.9 + 4.1x$

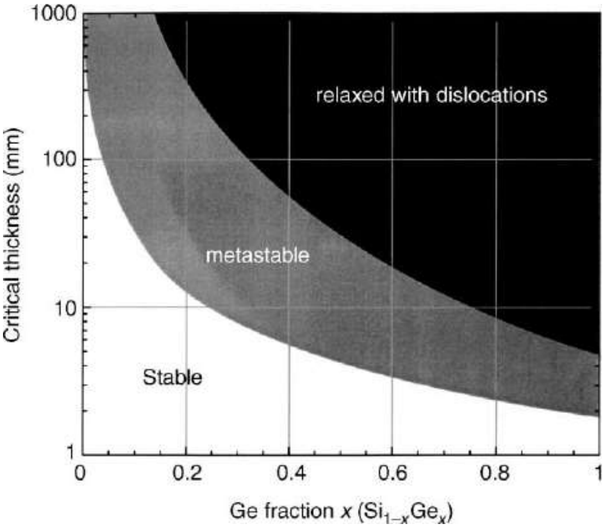

Figure 1 The critical thicknesses for growing a strained $Si_{1-x}Ge_x$ epitaxial layer on a bulk silicon substrate. The white area corresponds to the region below the equilibrium critical thickness as calculated with the Matthews and Blakeslee model while the gray region corresponds to a metastable layer which may relax after thermal treatments. The black area will have dislocations and strain relaxation using any growth technique or temperature.

growth temperature are applied to the heterolayers. For any manufacture, the layers cannot be metastable as high-temperature processes are normally employed in processing. The Matthews and Blakeslee equilibrium critical thickness, h_c for strained-$Si_{1-x}Ge_x$ on bulk Si is given by

$$h_c = \frac{0.55}{x}\ln[10h_c] \qquad [1]$$

while an approximate value can be obtained from

$$h_c \approx 1.7793x^{-1.2371}(\text{nm}) \qquad [2]$$

Figure 1 clearly demonstrates that for thick SiGe films, low Ge contents must be used.

There is also a substantial interest in producing low-defect relaxed $Si_{1-y}Ge_y$ films or strained-Si on a relaxed $Si_{1-y}Ge_y$ substrate which can provide bandgaps and band discontinuities not available to pseudomorphically grown alloy layers. Ideally the relaxed $Si_{1-y}Ge_y$ substrate is grown on a Si substrate as this allows integration with conventional complementary metal oxide semiconductor (CMOS) electronics. To date no high-quality bulk single-crystal $Si_{1-y}Ge_y$ substrates are available due to the strong segregation of Si and Ge phases as liquid SiGe crystallizes. Growing a thick $Si_{1-x}Ge_x$ layer on Si well above the critical thickness so that it relaxes produces a threading dislocation density of about 10^{12} cm^{-2} which is far too large to produce device-quality material for electronics or optics. By growing a linearly graded $Si_{1-y}Ge_y$ with the grading rate sufficiently low (typically 10% per μm), the strain relaxation mechanism changes from a nucleation relaxation mechanism to a glide relaxation mechanism resulting in dislocations with comparatively long misfit segments. The defect densities are typically 10^5–10^6 cm^{-2} (Figure 2) although recent thermal processes and other techniques have now reduced this to 100 cm^{-2}.

Substantial work has appeared in the literature on self-organized growth of islands. Such islands are of interest in many materials systems as the quantum confinement they may offer can potentially enhance luminescence properties in numerous opto-electronic devices. The dots are formed by the Stranski–Kranstanow growth mode in a strained system. Planar epitaxial layers grown under stress conditions can lower their elastic energy by morphological changes. For the growth of strained-Ge or $Si_{1-x}Ge_x$ on bulk Si, this corresponds to the formation of islands when the growth parameters in the system are correctly set. While the islands may have well-defined shapes, the formation of the dots relies on random nucleation followed by ripening and coalescence. These mechanisms have produced large distribution of sizes (up to 30% of the mean size), densities and positions. It is quite common to have distributions of two or more sizes of islands and the densities are typically of order of 10^8–10^9 islands per cm^2. While the heights of these islands may be as low as 3 nm, the

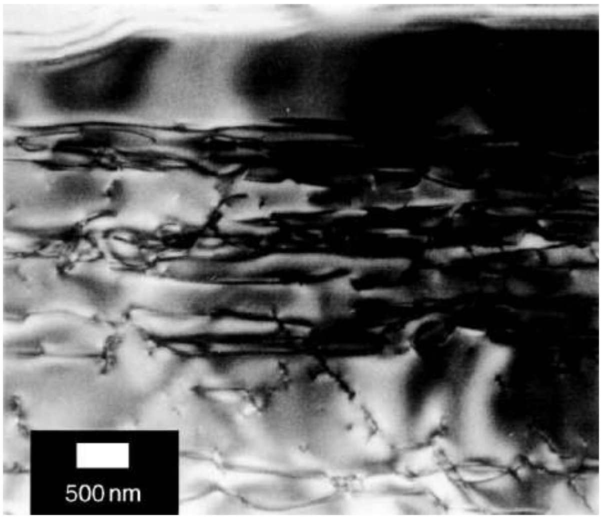

Figure 2 A transmission electron micrograph of a strain relaxation buffer grown by DERA, Malvern. The dense region of dislocations corresponds to an approximately 4 μm thick $Si_{1-y}Ge_y$ layer which is graded from $y = 0$ to $y = 0.23$. Above this is a 1 μm thick $Si_{0.77}Ge_{0.23}$ buffer which has no dislocation segments threading to the surface.

widths of the islands are typically 20 nm or greater, which is not yet small enough to produce the quantum confinement properties useful for significant improvements in opto-electronic devices.

Electronic Properties and Band Structure

As many of the optical properties and responses are derived or result from the band structure of materials, this section will concentrate on the band structure and electronic properties of SiGe layers. The bandgap of SiGe whether strained or unstrained is always below that of bulk Si. This potentially allows applications at longer wavelengths than comparable Si devices. Figure 3 shows the appropriate bandgaps for both the unstrained and strained $Si_{1-x}Ge_x$ grown on relaxed Si. For the unstrained case with a Ge fraction $x < 0.85$, the band structure is Si-like with six Δ-valleys forming the conduction band (Figure 4a). For $x > 0.85$ the band structure is Ge-like with four L-valleys. When strain is applied, both the conduction band and the valence band are split. Figure 4b shows the splitting of the conduction band with the Δ_4 valleys lower in energy and the Δ_2 valleys higher. In addition the valence band is split with the heavy-hole band being higher in energy than the light-hole band (Figure 3). The split-off band is also moved further away from the light and heavy hole bands as the strain in the system is increased. The indirect energy gap for the strained $Si_{1-x}Ge_x$ at 4.2 K as measured by photoluminescence is given by

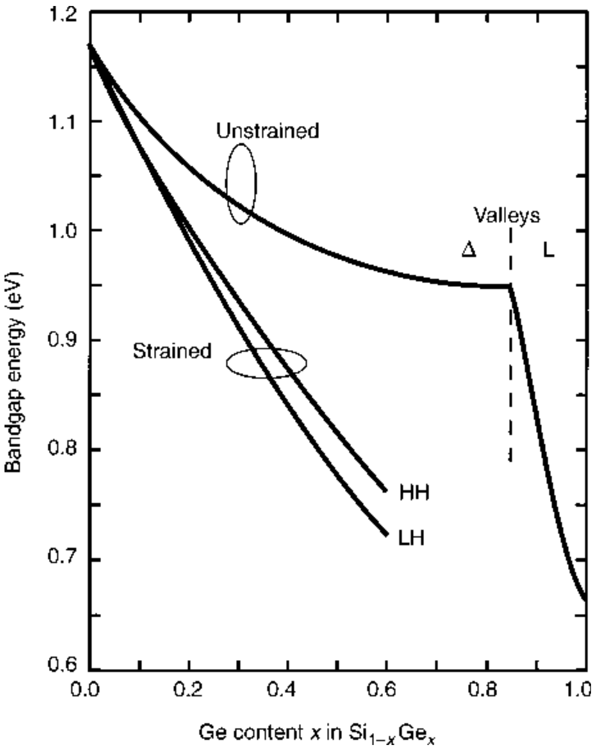

Figure 3 The bandgap energy for strained $Si_{1-x}Ge_x$ on a relaxed Si substrate and unstrained $Si_{1-x}Ge_x$. For the unstrained there is a discontinuity at $x \sim 0.85$ where the band structure changes from Si-like with Δ-valleys to Ge-like with L valleys (see Figure 4).

$$E_g = 1.155 - 0.43x + 0.0206x^2\,[\text{eV}] \text{ for } x < 0.85 \quad [3]$$

$$E_g = 2.010 - 1.27x\,[\text{eV}] \text{ for } x > 0.85 \quad [4]$$

while the direct bandgap is given by

$$E_g = 3.395 - 1.287x + 0.153x(1-x)\,[\text{eV}]\,(300\,\text{K}) \quad [5]$$

Due to the critical thickness, measurements of the bandgap of strained $Si_{1-x}Ge_x$ grown pseudomorphically on bulk Si substrates has been limited to low Ge concentrations. Using photoluminescence measurements at 4.2 K, Robbins has obtained a value for the exciton bandgap for strained $Si_{1-x}Ge_x$ layers for $x < 0.25$ to be

$$E_g = 1.155 - 0.874x + 0.376x^2\,[\text{eV}] \quad [6]$$

The bandgap for any strained $Si_{1-x}Ge_x$ layer grown on a relaxed $Si_{1-y}Ge_y$ substrate is shown in Figure 5. If the active strained $Si_{1-x}Ge_x$ layer is compressively strained then the conduction band is split into Δ_4 valleys being lower in energy and the Δ_2 valleys being higher (Figure 4b). For tensile strain (such as strained Si on a relaxed $Si_{1-y}Ge_y$ substrate), the Δ_2 valleys are lower in energy and the Δ_4 valleys are higher (Figure 4b).

The band alignments for typical heterolayer structures are shown in Figure 6. Pseudomorphically grown strained $Si_{1-x}Ge_x$ layers are only useful for forming quantum wells in the valence band (Figure 6a). Photoluminescence measurements do suggest a type I band alignment for these structures (quantum wells in both the conduction and valence bands) although it was later shown that the material is type II but due to charges producing band bending, the photoluminescence wrongly suggests a type I transition. For a quantum well for electrons, strained Si layers are grown on a relaxed $Si_{1-y}Ge_y$ substrate. In addition, quantum wells in the valence band can be produced for strained $Si_{1-x}Ge_x$ layers with high Ge content grown on relaxed $Si_{1-y}Ge_y$ substrates for $x > y$. In almost all cases the band alignments are type II although a small region exists where theory suggests type I is possible but this has not been demonstrated experimentally at present.

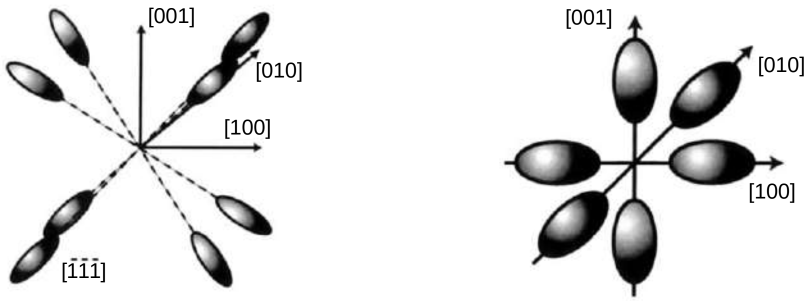

(a) bulk-Ge and Strained $Si_{1-x}Ge_x$ ($x > 0.85$) bulk-Si

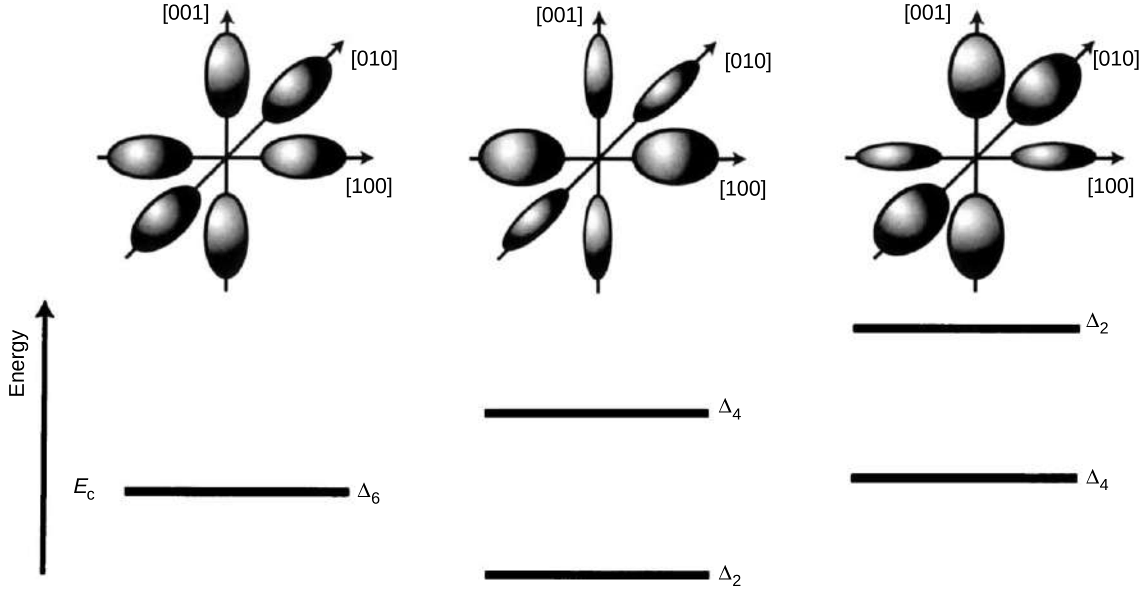

(b) Relaxed $Si_{1-y}Ge_y$ ($y > 0.85$) Strained Si on relaxed $Si_{1-y}Ge_y$ Strained $Si_{1-x}Ge_x$ on relaxed $Si_{1-y}Ge_y$ ($x > y$)

Figure 4 (a) A schematic diagram of the valleys for Ge-like and Si-like material. (b) A schematic diagram of the valley structures and energy levels of the conduction bands for relaxed, tensile strained (strained Si on relaxed $Si_{1-y}Ge_y$) and compressively strained (strained $Si_{1-x}Ge_x$ on relaxed $Si_{1-y}Ge_y$ ($x > y$)) material.

The valence band discontinuity is given by

$$\Delta E_v = (0.74 - 0.06y)(x - y)\,[\text{eV}] \qquad [7]$$

while the conduction band discontinuity for strained Si on a relaxed $Si_{1-y}Ge_y$ substrate is

$$\Delta E_c = 0.6y\,[\text{eV}] \qquad \text{for } x < 0.6 \qquad [8]$$

The complete set of ΔE_c and ΔE_v values for all x and y are shown in Figures 7 and 8 respectively. Table 2 gives some of the electronic data on Si, Ge, and relaxed $Si_{1-x}Ge_x$ including mobilities, bandgaps, and effective masses.

The mobilities for electrons and holes in relaxed $Si_{1-x}Ge_x$ layers are below the bulk Si and Ge layers due to alloy scattering (Table 2). For strained layers, mobility enhancements have been demonstrated in a number of different systems. Electron mobility enhancements result from the splitting of the valleys reducing intervalley scattering. Hole mobility enhancements result from the splitting of the valence band and reduction of effective masses. The hole mobilities for pseudomorphically grown strained $Si_{1-x}Ge_x$ quantum wells have shown values as high as 17 000 $cm^2\,V^{-1}\,s^{-1}$ at a carrier density of $2 \times 10^{11}\,cm^{-2}$ for $Si_{0.87}Ge_{0.13}$ layers. Higher Ge contents reduce the mobility due to alloy scattering. At room temperature, only modest mobility improvements of about 20% have been observed giving 220 $cm^2\,V^{-1}\,s^{-1}$ at the high carrier

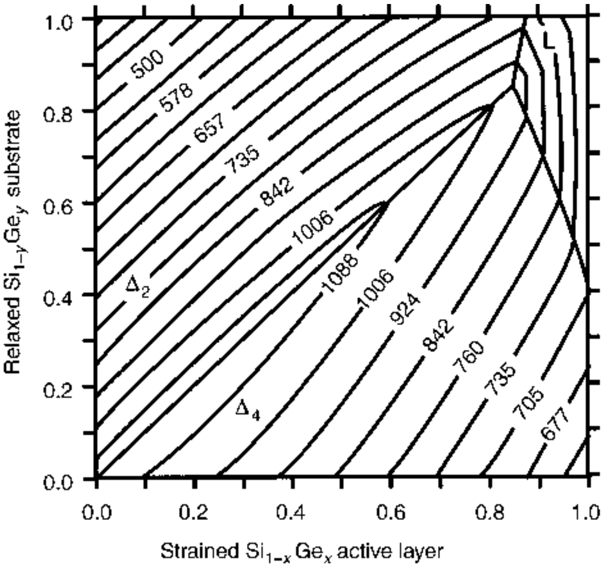

Figure 5 The bandgap energy (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Theoretical curves from Riger and Vogl have been calibrated to the experimental data from pseudomorphic layers.

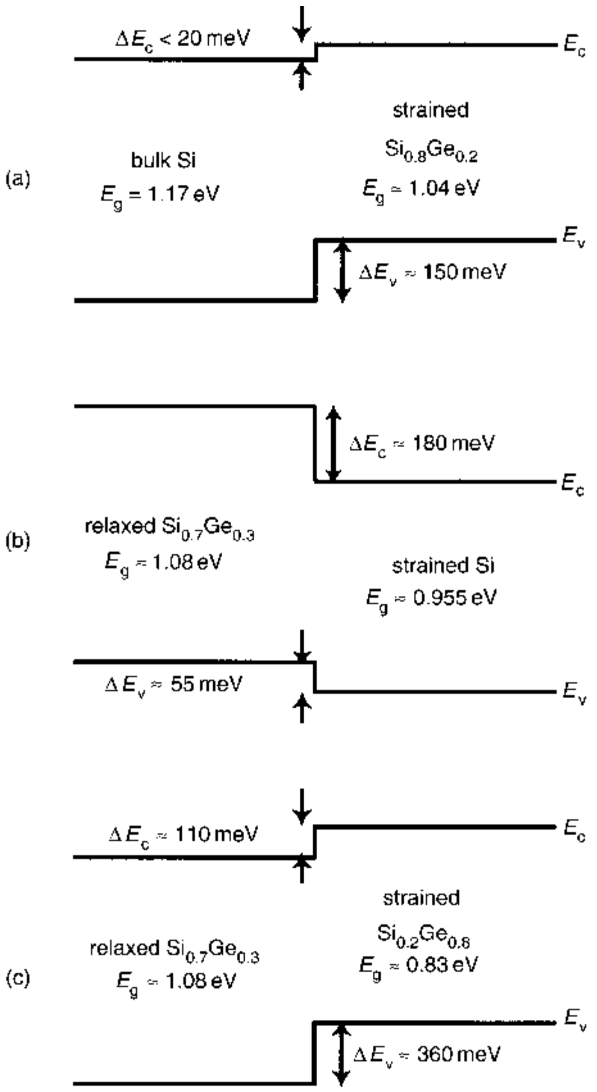
(a)
ΔE_c < 20 meV
E_c
bulk Si
E_g = 1.17 eV
strained
$Si_{0.8}Ge_{0.2}$
$E_g \simeq$ 1.04 eV
E_v
$\Delta E_v \approx$ 150 meV
(b)
$\Delta E_c \approx$ 180 meV
E_c
relaxed $Si_{0.7}Ge_{0.3}$
$E_g \approx$ 1.08 eV
strained Si
$E_g \approx$ 0.955 eV
$\Delta E_v \approx$ 55 meV
E_v
(c)
$\Delta E_c \approx$ 110 meV
E_c
relaxed $Si_{0.7}Ge_{0.3}$
$E_g \approx$ 1.08 eV
strained
$Si_{0.2}Ge_{0.8}$
$E_g \approx$ 0.83 eV
E_v
$\Delta E_v \approx$ 360 meV

Figure 6 A schematic diagram of some typical SiGe heterolayer structures showing the bandgap, conduction band discontinuity, ΔE_c, and the valence band discontinuity, ΔE_v. (a) shows a quantum well in the valence band using a pseudomorphic strained $Si_{0.8}Ge_{0.2}$ layer; (b) shows a quantum well in the conduction band using a strained Si layer on a relaxed $Si_{0.7}Ge_{0.3}$ substrate; and (c) shows a quantum well in the valence band using a high Ge content layer on a relaxed substrate with lower Ge content.

densities required for electronic devices of about 2×10^{12} cm^{-2}. To remove alloy scattering, pure strained Ge channels grown on relaxed $Si_{1-y}Ge_y$ substrates have shown low-temperature mobilities of 55 000 cm^2 V^{-1} s^{-1} for densities of 5×10^{11} cm^{-2} with the room-temperature values reaching 1665 cm^2 V^{-1} s^{-1}. For electrons, a strained Si layer on relaxed $Si_{1-y}Ge_y$ substrates ($0.2 < y < 0.3$) also reduces the alloy scattering in the channel and produces much higher low-temperature mobilities of 390 000 cm^2 V^{-1} s^{-1} for densities of 4×10^{11} cm^{-2} while room-temperature values have peaked at 2830 cm^2 V^{-1} s^{-1}. All these values correspond to electrons or holes at heterointerfaces. Enhanced mobilities have also been observed for strained Si metal oxide semiconductor field effect transistors (MOSFET) where electron mobility enhancements of 75% compared to the bulk Si devices has been reported on relaxed $Si_{0.8}Ge_{0.2}$ although intrinsic transconductances are only increased by 60%. For holes, a strained Si layer on a $Si_{0.8}Ge_{0.2}$ relaxed substrate has demonstrated a 50% mobility enhancement over a bulk Si MOSFET.

The effective mass for electrons in strained Si layers on relaxed $Si_{1-y}Ge_y$ substrates for $0.1 < y < 0.4$ has produced a very small spread of effective masses with values of 0.19 m_0 (where m_0 is the free electron mass) for the longitudinal mass as measured by cyclotron resonance and the temperature dependence of Shubnikov–de Haas oscillations. The electron effective masses are very similar to the bulk values for strained Si quantum wells. Hole effective masses in pseudomorphically strained $Si_{1-x}Ge_x$ layers have produced a large spread of data even for similar alloy and strain values mainly due to the large nonparabolicity of the valence band (Table 3) and many of the measured samples having metastable $Si_{1-x}Ge_x$ quantum wells. The in-plane hole mass is theoretically expected to decrease as the strain in the system is increased but to date the experimental values have been much lower than theoretically calculated values. Experimental data (Table 3)

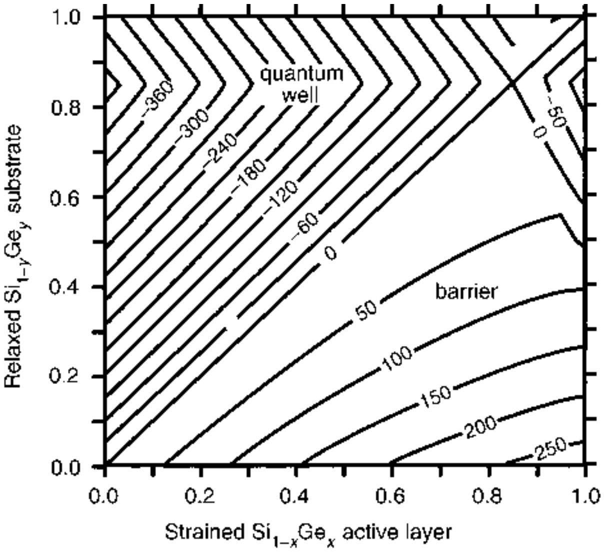

Figure 7 The conduction band discontinuity (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Positive numbers correspond to potential barriers to electrons while negative values represent quantum wells. The general trends and unknown experimental regimes are from the theoretical work of Rieger and Vogl while the quantum well data for $0.1 < y < 0.4$ and $x = 0$ has been recalibrated from experimental work. No experimental data are known to allow calibration of the barrier regime.

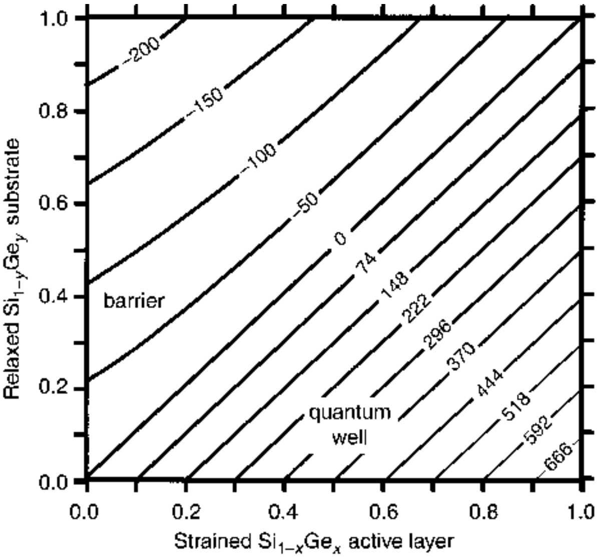

Figure 8 The valence band discontinuity (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Positive numbers correspond to quantum wells for holes while negative values represent potential barriers. The values from theory in the quantum well region agree well with experimental values but little experimental data exist to compare the barrier data from the theorertical values (plot from Rieger and Vogl) with experiment.

suggests that strained $Si_{0.88}Ge_{0.12}$ layers have a heavy effective mass of about $0.3\,m_0$ compared to the bulk value of $0.55\,m_0$.

Optical Properties

Due to the problems of producing high-quality single crystals of $Si_{1-x}Ge_x$, there are relatively little data on the optical properties of relaxed $Si_{1-x}Ge_x$ in the literature. The situation is even worse for strained $Si_{1-x}Ge_x$ as the layer thicknesses are limited to below 30 nm for Ge contents above 0.1, a value much smaller than the penetration depths in the system at most optical frequencies. There is, however, substantial interest in the optical properties, particularly the refractive index, as they suggest that SiGe can be used as a waveguide on Si substrates.

For the relaxed $Si_{1-x}Ge_x$ layers, most of the published work in the field has been from crystals grown by liquid phase epitaxy using melts of indium or tin. The samples had free hole concentrations of $10^{17}\,\mathrm{cm}^{-3}$ along with indium or tin impurities and details can be found in the reviews by Humlícek. Some of the data are also from thick epitaxial layers grown well above the critical thickness where the threading dislocation density is likely to be about $10^{12}\,\mathrm{cm}^{-2}$. The data below 2 eV were obtained by absorption measurements, those between 2 and 5.6 eV have been obtained using ellipsometry and the values between 6 and 12 eV were obtained by reflectance measurements.

Figures 9 and 10 show the real and imaginary parts of the complex dielectric function defined as $\varepsilon = \varepsilon_1 + i\varepsilon_2$ where $i = \sqrt{-1}$ for a number of different Ge contents. Figures 11 and 12 show the real and imaginary parts of the complex refractive index defined by $N = n + ik = \sqrt{\varepsilon}$. The peak starting at 3.395 eV and decreasing as the Ge content is increased corresponds to the E_1 transition across the direct gap and may be used to find the Ge content using the expression

$$x = 4.707 - \sqrt{6.538E_1 - 0.0397} \qquad [9]$$

The refractive index is high owing to the strong electronic interband absorption in the visible and ultraviolet regions of the spectrum.

The other important optical property of the material is the absorption. The absorption coefficient is defined as $K = 4\pi k/\lambda$ where k is the extinction coefficient for a wavelength λ. The intensity of a wave traversing a distance d in a homogeneous medium of extinction coefficient, k, is attenuated by a factor $\exp(-Kd)$. Therefore $1/K$ is the penetration depth. The absorption coefficient is shown in Figure 13.

Table 2 Electronic properties of Si, Ge and relaxed $Si_{1-x}Ge_x$

Element		*Si*	*Ge*	*Bulk unstrained* $Si_{1-x}Ge_x$
Direct bandgap	(eV) at 300 K	3.4	0.80	$3.395 - 1.287x + 0.153x(1-x)$
Indirect bandgap	(eV) at 300 K	1.11 (D)	0.664 (L)	$1.155 - 0.43x + 0.0206x^2$ $(x < 0.85)$ $2.010 - 1.27x$ $(x > 0.85)$
Spin-orbit splitting	Δ_{so} (meV)	44	290	Discontinuity at $x \sim 0.85$
Electron mobility	μ_e ($cm^2\,V^{-1}\,s^{-1}$) at 300 K	1450	3900	$\mu_e = 1466.2 - 6425.9x + 6601.8x^2$ $+7716.5x^3 - 10944x^4 (x < 0.6)$ $\mu_e = 2.9235 \times 10^{-28} e^{71.105x} + 4790.1$ $-14246x + 11117x^2 (x > 0.6)$
Hole mobility	μ_h ($cm^2\,V^{-1}\,s^{1}$) at 300 K	505	1800	$\mu_h = 188.42 - 935.7x + 5455.5x^2 - 8553.7x^3$ $+ 5751.7x^4$(all x) ($N_A = 1 - 50 \times 10^{16}\ cm^{-3}$)
Longitudinal electron mass	m_l^*	0.1905	0.082	Discontinuity at $x \sim 0.85$
Transverse electron mass	m_t^*	0.9163	1.58	Discontinuity at $x \sim 0.85$
Heavy-hole mass	m_{hh}^*	0.537	0.284	Discontinuity at $x \sim 0.85$
Light-hole mass	m_{lh}^*	0.153	0.044	Discontinuity at $x \sim 0.85$
Spin–orbit mass	m_{so}^*	0.234	0.095	Discontinuity at $x \sim 0.85$

Table 3 A comparison of effective masses in modulation-doped samples as measured by the temperature dependence of the Shubnikov–de Haas oscillations (SdH) or by cyclotron resonance (CR) at liquid helium temperatures. All samples have strained $Si_{0.88}Ge_{0.12}$ or $Si_{0.87}Ge_{0.13}$ quantum wells grown on bulk Si substrates. Therefore only the heavy-hole band should be populated at these low temperatures and strain values. ℓ is the mean free path which is the mean scattering distance of a hole in the system and is a quantitative measure of disorder

	Mobility ($cm^2\,V^{-1}\,s^{-1}$)	p ($\times 10^{11}\ cm^{-2}$)	ℓ (nm)	m^* (m_0) SdH	m^* (m_0) CR
DERA	6560	4.4	71	0.304	0.26
NRC	6800	5.9	86	0.31	–
IBM	6000	3.7	60	0.34	0.18
Warwick	9600	2.5	78	0.27	0.23, 0.29
Warwick	10500	2.1	78	0.27	0.24, 0.30

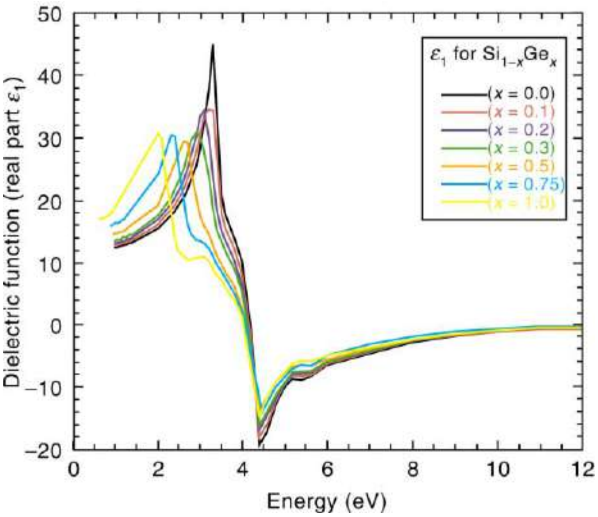

Figure 9 The real part (ε_1) of the complex dielectric function $\varepsilon = \varepsilon_1 + i\varepsilon_2$.

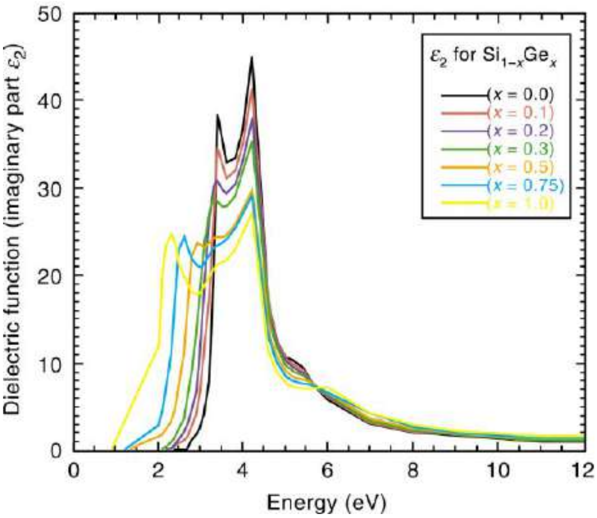

Figure 10 The imaginary part (ε_2) of the complex dielectric function $\varepsilon = \varepsilon_1 + i\varepsilon_2$.

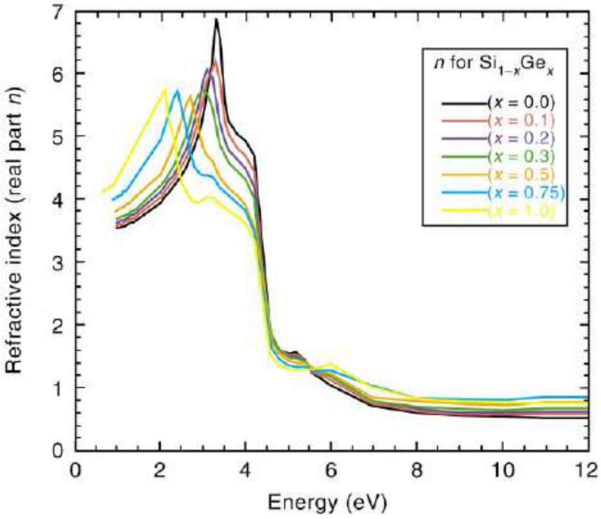

Figure 11 The real part (n) of the complex refractive index $N = n + \mathrm{i}k$.

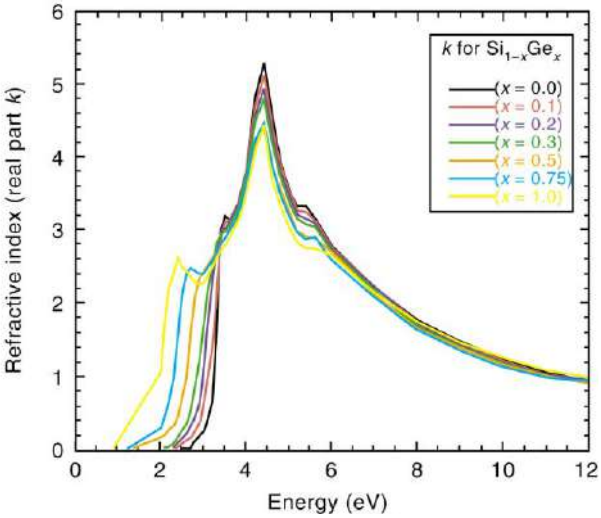

Figure 12 The imaginary part (k) of the complex refractive index $N = n + \mathrm{i}k$.

Determining the optical properties of strained $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ layers has been substantially more difficult. The major problem is that the absorption coefficient is about 10–20 cm^{-1} around the fundamental bandgap (Figure 6) and therefore any shift due to strain effects becomes very difficult to observe using absorption techniques in thin SiGe layers. Additional complexity results as the substrate material has a fundamental bandgap close to the strained layer. Therefore photoluminescence is one of the few methods for extracting experimentally the optical changes due to strain.

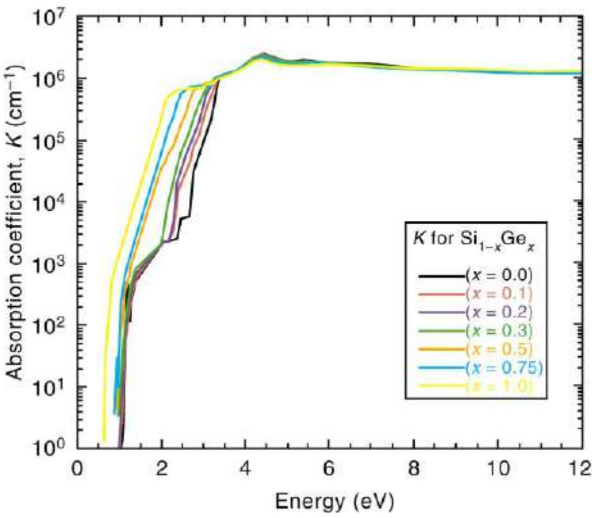

Figure 13 The absorption coefficient K.

The splitting of the conduction (Figure 8) and valence bands (Figure 9) has already been discussed. The photoluminescence signals will also shift with strain for the lowest lying bandgap energy. Almost all the experimental data for strained layers have been obtained for pseudomorphically grown strained $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ alloys on bulk Si substrates and the data agree well with the bandgaps shown in Figures 3 and 5. The quantum well photoluminescence signal is found to shift by −1.4 meV per kbar of pressure. After removal of nonhydrostatic components, a value of −1.52 meV per kbar for the pure hydrostatic shift has been reported.

Conclusions

A review has been presented of the fundamental properties of relaxed and strained $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ alloys. The materials properties were first reviewed and demonstrate the limitations imposed by the critical thickness. The electronic properties including the bandstructure were then discussed and demonstrate the lowered bandgap for $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ alloy compositions compared to bulk Si. This property has been used in $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ p-i-n photodetectors to move the absorption to longer wavelengths in an attempt to compete with III–V photodetectors for the important 1.3 μm and 1.55 μm fiber optic frequencies. In the final section the optical properties of relaxed and strained $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ material were reviewed. The difference in refractive indices has opened research into SiGe on Si waveguides.

List of Units and Nomenclature

Absorption coefficient (K)	[cm^{-1}]
Band gap energy (E_g)	[eV]
Conduction band discontinuity (ΔE_c)	[meV]
Critical thickness	[nm]
Density	[kg m^{-3}]
Dielectric constant (ε)	
Effective mass [m_0 where $m_0 = 9.11 \times 10^{-31}$ *kg*]	
Elastic moduli (c_{ij})	[GPa]
Extinction coefficient (k)	[cm^{-1}]
Lattice constant (a_0)	[nm]
Mean free path (l)	[nm]
Melting point	[K]
Mobility (μ)	[cm^2 V^{-1} s^{-1}]
Pressure	[kbar]
Refractive index (N)	
Sheet carrier densities	[cm^{-2}]
Sound velocity	[m s^{-1}]
Temperature	[K]
Thermal conductivity (κ)	[W m^{-1} K^{-1}]
Valence band discontinuity (ΔE_v)	[meV]
Wavelength (λ)	[nm]

See also

Semiconductor Physics: Band Structure and Optical Properties; Outline of Basic Electronic Properties; Recombination Processes.

Further Reading

Humlícek J, Garriga M, Alonso MI and Cardona M (1989) Optical spectra of $Si_{1-x}Ge_x$ alloys. *Journal of Applied Physics* 65(7): 2827–2832.

Humlícek J, Lukes F and Schmidt E (1991) Silicon-germanium alloys (Si_xGe_{1-x}). In: Palik ED (ed.) *Handbook of Optical Constants of Solids*, vol. II, pp. 607–636. San Diego, CA: Academic Press.

Kasper E and Lyutovich K (eds) *Properties of Silicon Germanium and SiGe:Carbon*. London: INSPEC IEE.

Paul DJ (1999) Silicon-germanium strained layer materials in microelectronics. *Advanced Materials* 11(3): 191–204.

Rieger MM and Vogl P (1993) Electronic-band parameters in strained $Si_{1-x}Ge_x$ alloys on $Si_{1-y}Ge_y$ substrates. *Physical Review B* 48(19): 14276–14287.

Robbins DJ, Canham LT, Barnett SJ, Pitt AD and Calcott P (1992) Near band gap photoluminescence from pseudomorphic $Si_{1-x}Ge_x$ single layers on silicon. *Journal of Applied Physics* 71(3): 1407–1414.

III-Nitrides

K P O'Donnell, University of Strathclyde, Glasgow, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction to InGaN

The optical properties of wurtzite GaN are of interest mainly to band-structure theorists. Luminescence from pure InN has probably never been detected. It is therefore surprising that the combination of these binary III–N compounds produces a solid solution, InGaN, which has revolutionized blue–green optoelectronics in the last 10 years, and may form the basis of a new lighting technology to replace the tungsten lamp in the first half of the twenty-first century.

InGaN is also the material basis of the blue laser diode, which will be used for future generations of optical data storage and for many other applications that require a compact source of short-wavelength coherent light. The market for nitride light emitters has been growing at a rate of ~20% a year since the introduction of the first blue nitride light-emitting diode (LED) by Nichia Chemical in 1993. The first commercial lasers, with lifetimes in excess of 10 000 hours, were released in 1998.

So what is InGaN? The naïve response–that it is a semiconductor alloy–ignores the theoretical demonstration, based on simple equilibrium thermodynamics, that InN is insoluble in GaN at the temperatures usual for growth (using metalorganic vapor phase epitaxy, MOVPE). If we imagine a cation lattice, on which we distribute Ga and In atoms, in a fixed numerical proportion, in order to form a solid solution of given composition, it is thermodynamically probable that the In atoms will associate to form agglomerations in certain regions of the lattice. The size, shape, and properties of the resulting concentration fluctuations have been a topic of intense speculation.

Practically speaking, InGaN layers are grown either as rather thin (<3 nm) quantum wells (QWs) immersed in GaN or AlGaN or as thicker epilayers (10–500 nm) on top of thick GaN templates. Only the QWs are of use in devices, whereas the composition and structure of the thick epilayers are easier to measure in research applications. It is dangerous to suppose that the material retains the same properties when grown in thicker layers. Recently, a third class of sample has emerged, in the form of self-organized 'quantum boxes' (originally French *boîtes*), which we shall call QBs. These samples, grown at low temperatures by molecular beam epitaxy (MBE) comprise pancake-shaped disks of uncertain composition, about 50 nm across and 3–5 nm high.

Optical Spectra of Devices

There have been several 'generations' of nitride LED. We refer here to the Nichia Blue, Extrablue (actually turquoise) and Green diodes which set the scene in 1996, and the Amber diode which first appeared in 1998, although not yet available commercially. The spectral properties of these devices have been stable while their efficiencies have improved in the intervening years. The basic design appears to incorporate a single quantum well. Up to a point, the wavelength tuning depends upon the incorporation of different amounts of indium in the well. (Green and amber devices may have the same amount.)

In what follows, we prefer to specify composition by quoting the indium nitride content, x, according to the formula unit

$$\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N} = x\mathrm{InN} + (1-x)\mathrm{GaN} \qquad [1]$$

Electroluminescence (EL) spectra of blue and green diodes are shown in Figure 1. Spectra are broad (compared to k_BT at room temperature) with exponential tails to high and low photon energies. The linewidth tends to increase as the device peak wavelength shifts further to the red.

Whereas it is difficult to measure directly the optical absorption spectrum of an ultrathin layer of material, an equivalent measurement can be made on an LED by monitoring the photocurrent (PC) generated by incident monochromatic light. PC spectra of the blue and green diode are shown in Figure 1, together with the absorption and photoluminescence (PL) spectra of a thick violet-light-emitting epilayer.

The spectra of Figure 1 transform into each other by 'stretching' the x-axis, using the bandgap energy of GaN as an anchor point. Comparison of many pairs of spectra produces evidence of a *universal scaling behavior*, which relates the details of each spectrum to a single controlling parameter. In the case of epilayers, this is undoubtedly the InN fraction, as we shall see later.

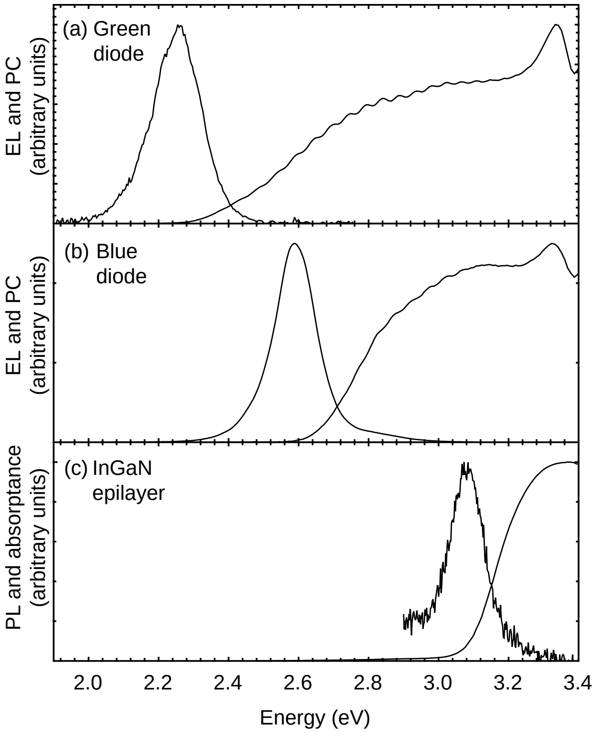

Figure 1 EL and PC (quasi-absorption) spectra of nitride LEDs, compared to PL and absorption spectra of a thick InGaN epilayer.

In order to describe the form of the absorption/excitation curve, we introduce the formal expression

$$\alpha = \frac{\alpha_0}{1 + \exp\left(\dfrac{E_g - E}{\Delta E}\right)} \qquad [2]$$

where α_0 is the plateau absorption, observed experimentally, but not yet theoretically explained, E_g is the (effective) bandgap energy of the sample, E is the photon energy, and ΔE is an energy broadening term, similar to the well-known Urbach parameter.

Optical Energy Relationships

The spectral universality pointed out in the previous section requires the existence of certain relationships between the optical energies. We list the defining energies as the bandgap, E_g, the peak energy of EL or PL, E_p, and the difference between these energies, measured for a particular sample, which we call the Stokes shift:

$$\mathrm{SS} = E_g - E_p \qquad [3]$$

(While the energy broadening tends to increase in a regular fashion with the Stokes shift, there are irregularities in both ΔE and the PL linewidths of samples from different sources. Hence ΔE and PL

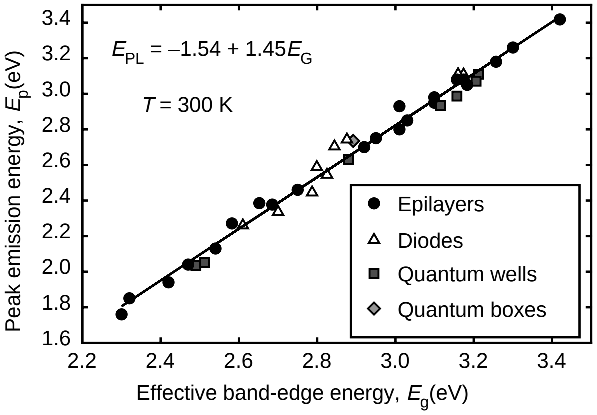

Figure 2 Linear dependence of luminescence peak energy on effective bandgap.

linewidth appear to be local rather than universal parameters.)

The relationship between E_g and E_p for a wide range of emitters is found to be linear (Figure 2):

$$E_p = 1.45E_g - 1.54 \qquad [4]$$

The bandgap energy and luminescence peak energy coincide at $1.54/(1.45 - 1) = 3.42$ eV, which corresponds to the energy bandgap of pure GaN. The simple linear dependence of E_p upon E_g implies a linear increase of the Stokes shift energy with increasing detuning from the GaN band edge.

Dependence of Optical Energies on *x*, the InN Fraction, for Thick Epilayers

As stated above, it is usual to measure the composition of thick layers, rather than quantum wells, for reasons of experimental convenience. Several techniques can be used to this end, including measurement of the lattice constants by X-ray diffraction (XRD) (one asymmetric and one symmetric reflections are required to take into account the strain state of the layer), Rutherford backscattering spectrometry (RBS), wavelength-dispersive (preferable) or energy-dispersive X-ray fluorescence (XRF), sputtered ion mass spectroscopy (SIMS), and extended X-ray absorption fine structure (EXAFS).

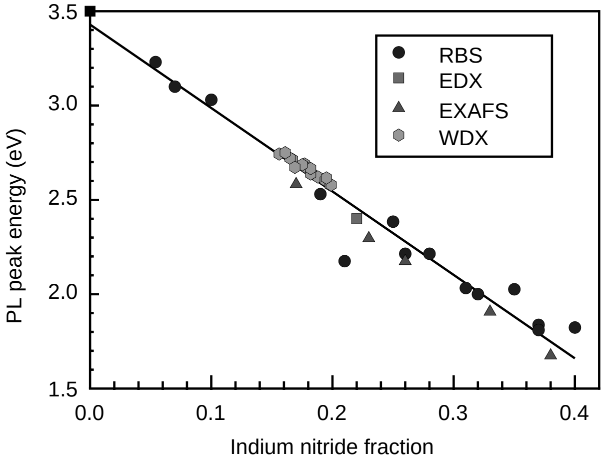

Figure 3 Linear dependence of the PL peak energy on the InN content of InGaN epilayers.

All usable techniques provide comparable results if carried out with sufficient care. We plot the peak energy of PL emission against the measured indium nitride fraction for a range of epilayer samples in Figure 3. Once more, the dependence appears to be linear in a restricted composition range. The best-fit line to a linear function $E_p(x)$ has an x-intercept which is close to the bandgap of GaN. The dependence of the optical energies on the indium nitride fraction can now be summarized by three interdependent linear relations, which are valid in the composition range $0 < x < 0.4$:

$$E_p = 3.4 - 4.4x$$

$$E_g = 3.4 - 3.0x \qquad [5]$$

$$\mathrm{SS} = 1.4x$$

The measurements described above rely on separate determinations of composition and an optical energy for each sample. Since all samples are to some extent inhomogeneous, it would be preferable to make such measurements at one and the same time. This can be done by using an electron probe microanalyzer, modified to allow simultaneous cathodoluminescence (CL) spectroscopy. CL is analogous to PL. A recent study of samples with $0.05 < x < 0.25$ has yielded the result:

$$E_c = 3.40(1) - 3.91(5)x \qquad [6]$$

with error bars (shown) that are about ten times smaller than those of the previous work.

Bandgap of InN

It is dangerous to extrapolate experimentally determined relationships beyond the range of measurements used to derive them.

The bandgap of an alloy is best represented by a quadratic function of x, with the deviation from linearity described by a bowing parameter. The above measurements suggest a small bowing parameter for the InGaN alloy system. If we ignore bowing and equate E_p with E_c, the InN bandgap obtained by setting $x = 1$ is calculated to be 0.71 eV. This is certainly an underestimate, but is similar to recent values obtained experimentally for layers of pure InN grown by Molecular Beam Epitaxy. Before the year 2000, the generally accepted value of the InN bandgap was close to 2 eV.

Origin of the Stokes Shift

Stokes shifts originate in systematic energy loss during the excitation–luminescence cycle. While the bandgap, as defined above, is the mean energy of electron–hole pairs across the bandgap, the PL peak is the modal energy of electron–hole recombination. The large Stokes shift in the InGaN system is probably due to the influence of giant internal electric fields that separate the carriers prior to recombination.

In the absence of a microscopic model, it is difficult to provide a quantitative description of energy loss in the excitation–emission cycle. Experiment shows that the shift can be as large as ~ 0.6 eV in an epilayer with $x = 0.4$. If such a layer were 100 nm thick and uniform in composition, it would be strained by about 4% when pseudomorphic to a GaN template, with an estimated vertical electric field of 6×10^6 V m^{-1}. Unit strain would correspond to a 15-eV shift (equivalent to a piezoelectric deformation potential). While these numbers are large, they seem reasonable in the light of sophisticated theoretical calculations of the magnitude of the piezoelectric effect in InGaN. But it is an unfortunate experimental fact that layers grown with different thicknesses, and presumably under different states of strain, may show the same value of the Stokes shift.

If the material has composition fluctuations, accompanied by strain variations, the recombination energy will be much more sensitive to the local environment. One can imagine a situation in which recombination always takes place at the lowest available energy, as carriers diffuse under the action of local strain/electric fields. This situation would be rather similar to that which pertains in quantum wells with spatial energy disorder, caused by well width fluctuations. A theoretical treatment of localization in three dimensions would be very difficult.

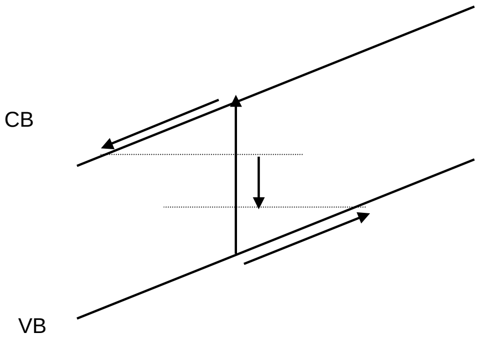

Figure 4 Excitation–emission cycle in the presence of an electric field. CB and VB are conduction band and valence band, respectively.

In any case, the physical situation of excitation and recombination in the presence of an electric field is that sketched in Figure 4.

Quantum Wells and Quantum Boxes

The incorporation of quantum wells has several advantages for light-emitting devices. In nitrides, the expected increase of oscillator strength for radiative transitions, due to the reduced dimensionality, is mitigated to some extent by the opposing effect of the internal electric fields, but the quantum well still acts as an energy sink for mobile excitations and as a site for carrier localization.

Quantum confinement can also act as a wavelength tuning mechanism in QW devices. The effect in blue-emitting nitride devices is rather small until the QW width falls below 3 nm. It should be clear that a reduction in the photon energy of recombination with increasing well width will accompany vertical separation of carriers in a piezoelectric model. Usually, as mentioned above, wavelength tuning in nitride devices is accomplished by changing the composition. What should we make of the fact that amber and green diodes apparently have the same composition? It would appear that the well widths are different. On the other hand, the amber emitters have anomalously narrow linewidths, which may imply that some other mechanism is in force. Here, we must let the matter rest, in the absence of detailed structural information.

Quantum boxes should be distinguished from quantum dots (QDs). The former are mesoscopic objects that form spontaneously under certain growth conditions, which usually involve the influence of strain. The best-characterized example in semiconductor physics is the Stranski–Krastanow growth of InGaAs structures on GaAs. While GaN in AlGaN forms recognizable nanocrystals, InGaN forms less well-defined structures of indeterminate shape, some 30–50 nm in lateral extent and 3–10 nm high. QDs, on the other hand, are truly zero-dimensional objects, in the sense that carrier motion is quantized simultaneously in three spatial dimensions. The utilization of reduced dimensionality to enhance recombination probabilities in active devices is only valid when QB structures meet the criteria of low dimensionality: the carrier motion must be restricted to a distance smaller than the appropriate Bohr radius.

Structure of InGaN Epilayers

The state of strain of InGaN epilayers is usually determined by XRD. Both symmetric and asymmetric reflections are required to calculate the actual values of the out-of-plane and in-plane lattice constants, *c* and *a*, respectively. These can then be compared with those predicted for a relaxed alloy by Vegard's law, for example:

$$c(\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}) = xc_{\mathrm{InGaN}} + (1-x)c_{\mathrm{GaN}} \qquad [7]$$

The composition itself, as well as the state of strain, can be determined from measurements of *c* and *a* with one additional assumption: that the Poisson ratio of the alloy also obeys Vegard's law. If the composition is known from an independent chemical measurement, this assumption can be tested.

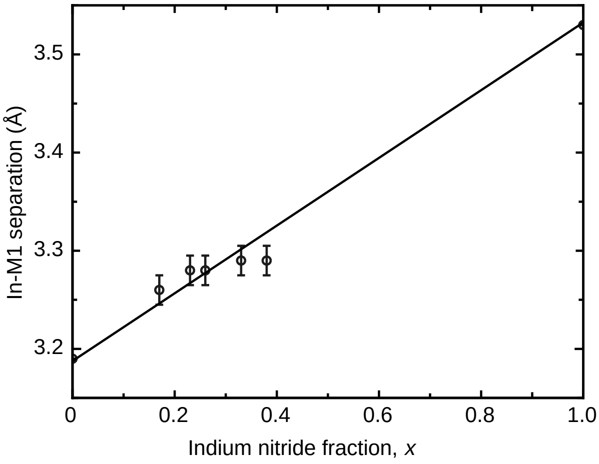

Figure 5 EXAFS-determined local structure parameters of thick InGan films.

InGaN epilayers are usually grown on thick GaN formers. In pseudomorphic growth, the constraint, $a(\text{layer}) = a_{\mathrm{GaN}}$ applies. This is sometimes assumed to be the case when layers are thinner than about 50 nm, but nothing should be taken for granted when dealing with this material. Each layer is different. A similar warning applies to the assumption that thick InGaN layers will be relaxed. It has recently been shown that partially relaxed and pseudomorphic regions may actually coexist within the same InGaN film, which nevertheless has a composition that is uniform throughout. Other films may show a concentration gradient due to compositional pulling during growth.

The local structure of InGaN films has been measured using In-edge EXAFS. Using a tetrahedral two-shell model that does not distinguish the wurtzite and cubic crystal structures, we find that the In–N1 bond length is more or less independent of the alloy composition in the range $0.17 < x < 0.38$ of samples measured, having a value of 2.11(1) Å (compared to 2.16 Å in pure InN). The M1 shell of mixed In and Ga atoms is located on average at about 3.28 Å, as shown in Figure 5.

The straight line on the figure connects the values of the *a* lattice parameters (of GaN at $x = 0$ and InN at $x = 1$) and hence represents the predictions of Vegard's law for a relaxed or bulk alloy. The measured In–M1 separations are less dependent on *x* than Vegard predicts.

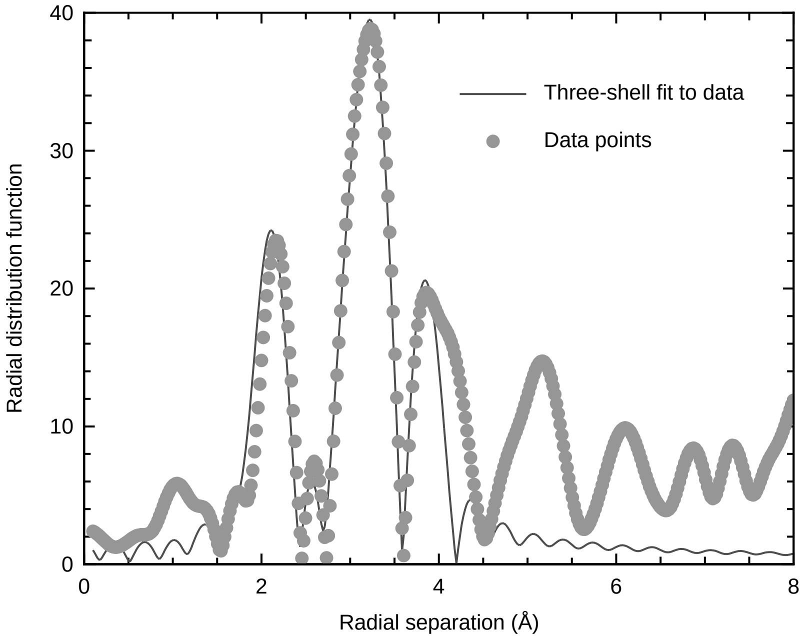

Figure 6 Fourier transform of the EXAFS spectrum yields the radial distribution function of an InGaN QB sample.

Local Structure of InGaN QW and QB

There has been a great deal of speculation in the literature concerning the local structure of InGaN alloys in the active layers of light-emitting devices. Much of this discussion paddles in the shallows of invalid comparison, as researchers extrapolate their findings from the results of measurements and calculations that can be made to those that cannot. One example is the large theoretical effort devoted to the defect physics of GaN, which is supposed somehow to 'explain' the optical properties of InGaN. ('Let us first suppose that all the horses in the race are spherical.')

We will try to avoid invalid comparisons in the brief discussion that follows. Recently, a Strathclyde University/Daresbury Laboratories/CHEA-CNRS Valbonne collaboration has succeeded in measuring the local structure of an InGaN QB sample, grown on a thick GaN template by MBE. The radial distribution function (RDF) around the target In probe atom is shown in Figure 6, together with a three-shell fit to the data.

The first shell of neighbors comprises four N atoms at 2.10(2) Å, the second shell is pure Ga at 3.22(3) Å, and the third shell pure In at 3.56(3) Å. These results suggest that there are two distinct In environments in such samples. One resembles the InGaN alloy, encountered in the epilayer samples, while the other is an environment of pure InN.

On the basis of these results, it is extremely likely that InGaN phase-separates in QB growth to form a two-component mixture comprising an In-poor and an In-rich phase. The manner in which these phases interpenetrate is a matter for speculation.

See also

Semiconductor Materials: Quantum Dots. **Semiconductor Physics:** Infrared Lattice Properties; Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Gil B (ed.) (1998) *Group III Nitride Semiconductor Compounds, Physics and Applications*. Series on Semiconductor Science and Technology 6. Oxford, UK: Oxford Science Publications.

Gil B (ed.) (2002) *Low Dimensional Nitride Semiconductors*. Oxford, UK: Oxford Science Publications.

Nakamura S and Fasol G (2000) *The Blue Laser Diode*, 2nd edn. Berlin: Springer-Verlag.

Large Gap II–VI Semiconductors

J F Donegan, Trinity College Dublin, Dublin, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Background

The chemical reaction of group IIA (Be, Mg), IIB(Zn, Cd) and group VI (O, S, Se, Te) atoms produces solid crystalline materials which are classified as semiconductors. The archetypal compound is ZnSe and this material will be used as the main example in this article. These are termed II–VI binary (involving two elements) compounds. It is also possible to grow alloy materials where more than one atom from each group is used; an example is $Zn_xCd_{1-x}Se$ where x is the proportion of Zn in the material. In general, there is a random arrangement of the Zn and Cd atoms within the crystal structure. Both binary and alloy semiconductor materials are single crystals so that the atoms are arranged in a periodic three-dimensional system called a lattice. The lattice may be represented by a unit cell; simply by repeating the pattern of the unit cell, we can generate the entire crystal lattice.

At the atomic level, the zinc (selenium) atom is bonded to 4 selenium (zinc) atoms which are arranged at the corners of a tetrahedron. The unit cell formed by this tetrahedral arrangement is termed the zincblende structure and is illustrated in Figure 1a. The dimension shown is called the lattice constant a and is characteristic of a particular II–VI semiconductor material. Some II–VI materials crystallize with a different unit cell which is termed the wurtzite structure, though again the atoms are bonded to four others in a tetrahedral arrangement as shown in Figure 1b. In this case, two lattice constants denoted a and c are required to characterize the unit cell. Some of the II–VI materials can occur in either of these two structures depending on how they were grown. The lattice constants of several II–VI compounds are given in Table 1.

II–VI Materials Growth

All of the II–VI family of semiconductors have high melting points due to the strong bonding that exists between the atomic species; values are given in

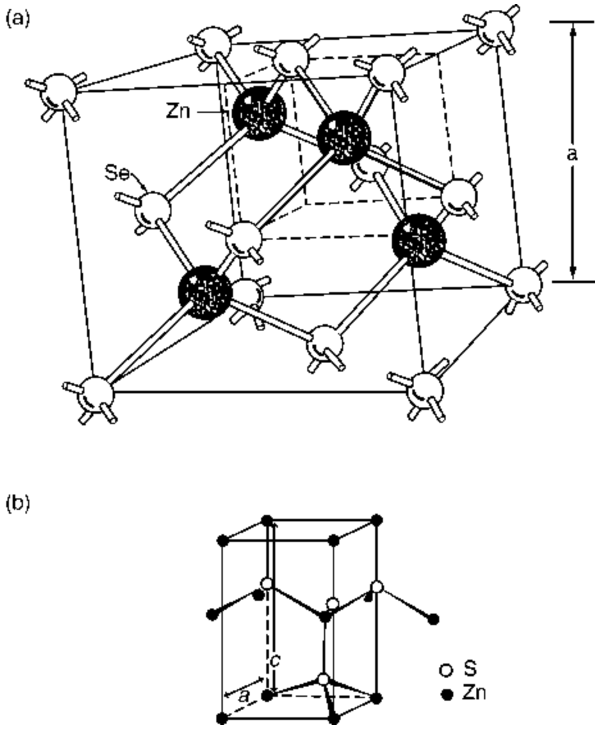

Figure 1 (a) Zincblende lattice; (b) wurtzite lattice. *a* and *c* are the lattice constants as given in Table 1.

Table 1. The crystals can be grown either by cooling a liquid (often called a melt) or by condensation from a gas (vapor). Crystal growth is a very precise and difficult process requiring the control of a large number of variables most notably temperature and pressure. As many of the II–VI materials can exist in either the zincblende or wurtzite structures, producing single crystals with only one structure is often quite difficult. Many of the crystals are 'twinned'; in a zincblende structure, a twin produces a wurtzite structure in one plane. The photograph in Figure 2 shows a ZnSe single crystal. This crystal took several days to grow indicating that very stable growth conditions must be maintained over long periods of time. The vapor phase technique takes advantage of the high vapor pressure of the II–VI compounds that allows for the growth of large single crystals at temperatures much below the melting point of these materials. Vapor growth is helped also by the congruent evaporation of the two constituents into the vapor state in a dissociated form. These two vapor constituents also recombine stoichiometrically into the solid crystal. The melting points and vapor pressures at the melting points of some II–VI semiconductors are given in Table 1.

It is also possible to grow materials in a layer-by-layer fashion onto an existing substrate; this is termed epitaxial growth. Recent developments have allowed the growth to proceed one atomic layer after the other. As a result, materials can be grown that are less than 100 Å in size. At these small sizes, the fundamental electronic and optical behavior of these materials is altered due to quantum mechanical effects. These materials are termed quantum wells. Growth temperatures in these cases are much below the melting point of these materials and allow for uniform single crystals without any 'twin' formation. In molecular beam epitaxy (MBE), atomic beams of the constituents impinge onto a heated substrate where they combine to form the crystalline material. Growth of ZnSe onto III–V substrates such as GaAs is well developed. The quality or purity of the single crystals can often be deduced from their optical properties as we will see later.

The addition of small amounts of other atoms into the crystal during the growth procedure is called doping. Semiconductors are materials whose electrical conductivity can be drastically altered by the addition of small amounts of dopants. This ability to control the electrical conductivity is the key to their use in the high-speed devices used in modern optical communications systems (*see* **Semiconductor Physics:** Impurities and Defects). In ZnSe, doping with Cl produces n-type material while doping with N produces p-type material. The ability to form a pn junction (a thin region between n and p type material) is the precursor for the production of optical devices such as light-emitting diodes (LEDs) or laser diodes (LDs) based on II–VI semiconductors. Doping also changes the optical properties of II–VI materials.

All materials are defective to some degree, and these II–VI materials are no exception. The number and type of defects are determined by thermodynamic and purity considerations during the growth process. There are native defects such as vacancies (missing atoms) and dislocations (extra or missing layers of atoms). There are also impurities present; foreign atoms due to the impurities in the starting growth materials. These impurities have an important effect on the ability of the material to emit light. Dopants and impurities may help to increase the amount of light emitted while dislocations usually reduce the emission efficiency in II–VI materials.

Optical Properties of II–VI Semiconductors: Basic Concepts

Introduction

When light is incident on a solid material, its electric field interacts with the electric field within the atoms comprising the solid. This interaction leads to a

Table 1 Structural, thermodynamic and optical properties of II–VI compound semiconductors

Material	*Lattice constant (Å)*	*Crystal structure*	*Melting point (mp) (°C)*	*Vapor pressure at mp (atm)*	*Refractive index*	*Bandgap (0 K)*	*Bandgap (300 K)*
ZnO	*a* 3.250	W	1975	22	2.0	3.44	3.37
	c 5.207						
ZnS	*a* 3.820	W	1718–1850	3.7	2.8	3.91	3.80
	c 6.260						
	a 5.410	Z			2.4	3.83	3.66
ZnSe	*a* 5.668	Z	1520	1.0	2.8	2.82	2.67
ZnTe	*a* 6.104	Z	1300	0.6	3.1	2.39	2.25
CdS	*a* 4.133	W	1405–1475	3.8	2.5	2.58	2.42
	c 6.710						
	a 5.832	Z			–	–	–
CdSe	*a* 4.299	W	1239–1264	0.4	2.6	1.84	1.74
	c 7.015						
	a 6.057	Z			–	1.77	–
CdTe	*a* 6.477	Z	–	–	2.75	1.60	1.49
MgS	*a* 5.89	Z	–	–	2.6	–	4.5
MgSe	*a* 5.62	Z	–	–	2.6	–	3.6
BeTe	*a* 5.627	Z	–	–	–	–	2.8

Data taken from many sources mentioned in the bibliography but primarily from Bhargava, R. (1997) (ed.) *Wide Bandgap II–VI Semiconductors*, The Institute of Electrical Engineers, London, Table 1 of Chapter 5.4 with permission.

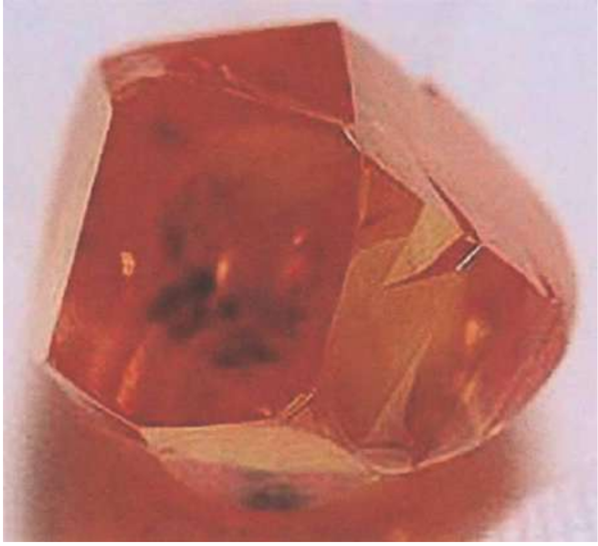

Figure 2 ZnSe single crystal grown by vapor transport. Photograph courtesy of NASA.

number of effects that we can observe. The light will be reflected or scattered from the surface of the solid, depending on whether the surface is smooth or rough on a length-scale similar to the wavelength of light. The materials of interest here are high-quality crystals with a smooth reflective surface. When light enters a material, it is refracted, i.e., it changes direction within the material. The simplest example of this is when a stick is placed in a jug of water, it appears to be bent. It is of course the light that is bent due to its different interaction with the air and with water. Snell's law states that:

$$\frac{\sin \alpha_{\mathrm{i}}}{\sin \alpha_{\mathrm{tr}}} = n \qquad [1]$$

where α_i is the angle of incidence of the light and α_{tr} is the angle of refraction. n is known as the refractive index of the material and is characteristic of it. As well as being refracted, any light that enters the crystal is either absorbed within the material or transmitted through the material to emerge on the other side. If the light is absorbed the material will look colored if the absorption takes place in the visible region where our eyes are sensitive. The ZnSe crystal in Figure 2 is such a material; the orange color is due to the absorption of blue light from the white light shining on the material. Other materials look transparent; ZnO is such an example. Following absorption, some materials produce light of their own; the color of the light is characteristic of the material and this process is termed luminescence. Light may also be scattered within a crystal by imperfections and by vibrations of the atoms in the crystal.

When light enters a solid material, the electric field polarizes the atoms within the structure. The positive and negative charges within the atoms move slightly from the equilibrium positions. All of the optical effects mentioned above are related to the induced polarization. As the constituent atoms in the structure are unique, the induced polarization and the resulting optical properties are specific to particular materials, and this makes optical characterization a very important tool for the investigation of semiconductor materials.

Reflection at the Air-to-Material Interface

Our experience of looking at reflection from glass tells us that these are weakly reflecting materials; the same is true of single crystal II–VI materials such as that in Figure 2. The interface between air and an optical material is a fundamental aspect of optical physics. The amount of light reflected can be determined from Maxwell's equations. The simplest equation is found where light is incident perpendicular to the surface of the material:

$$R = \frac{(n-1)^2}{(n+1)^2} \qquad [2]$$

where R is the reflection coefficient and n is the refractive index of the material. The value of n is characteristic of the particular material. Values of n for several II–VI materials are given in Table 1. The refractive index n for all these II–VI materials is not constant but varies slowly with the wavelength of the light impinging on the material. In Figure 3, we show the wavelength dependence of the refractive index of ZnSe.

Transmission and Absorption of Light within II–VI Materials

Once the light has passed through the interface with the air, it can either be transmitted through the material to emerge on the other side at a second material–air interface or be gradually absorbed within the body of the material. The ZnSe sample in Figure 2 is illuminated with white light, but it appears orange in color to our eyes, i.e., the orange component within the white light is transmitted through the material to our eye. The other components in the light, mostly green and blue, are absorbed within the material. To obtain more quantitative information, we take a spectrum of the absorption within the material. We illuminate the sample with light of different wavelengths and determine how much is transmitted or absorbed at each wavelength. The transmission spectrum of ZnSe is shown in Figure 4.

The visible spectral region extends from 0.4 to 0.7 microns and we see that there is a large change in transmission around 0.5 microns that is responsible for the color of the ZnSe crystal as seen in Figure 2. This spectral region around the onset of a large reduction in transmission at 0.5 microns is termed the bandgap of the material. As in the case of the refractive index, the value is unique to the material. II–VI materials are termed wide bandgap materials as they possess larger values of the bandgap energy than for Si, Ge and the III–V semiconductor materials. If we shine light of wavelength 0.5 microns onto ZnSe material, the transmission is very low. Under these conditions, the material would look black to our eyes. Once the light has been

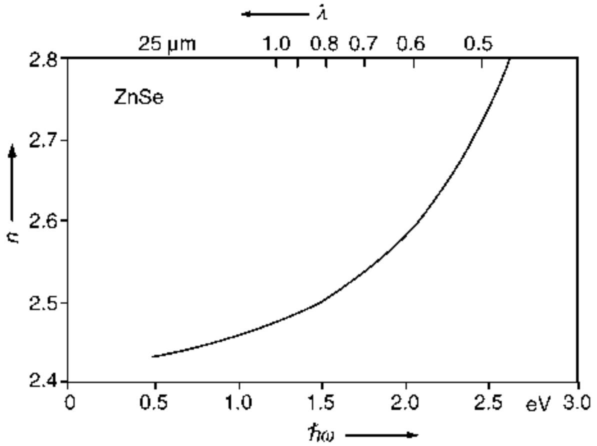

Figure 3 Refractive index n of ZnSe versus wavelength and photon energy at $T = 300\,K$. Adapted from Madelung E (ed.) (1982) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 17b: Physics of II–VI and II–VII Compounds, Semimagnetic Semiconductors*, Berlin: Springer-Verlag, Figure 52, with permission.

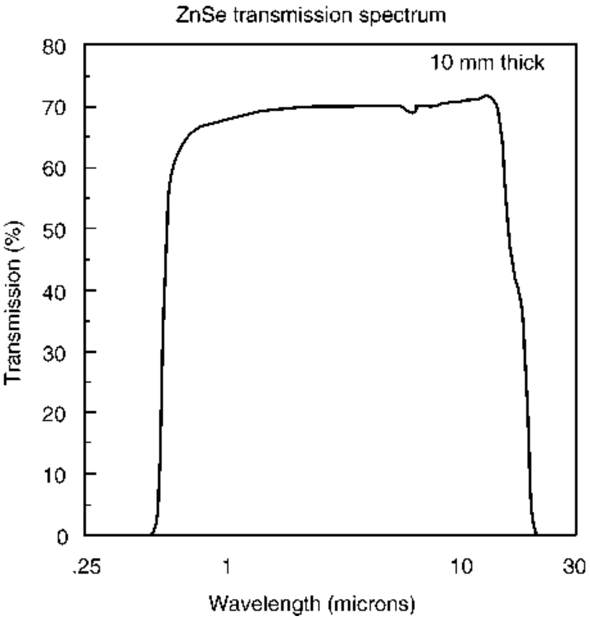

Figure 4 ZnSe transmission spectrum versus wavelength. Courtesy of II–VI Incorporated.

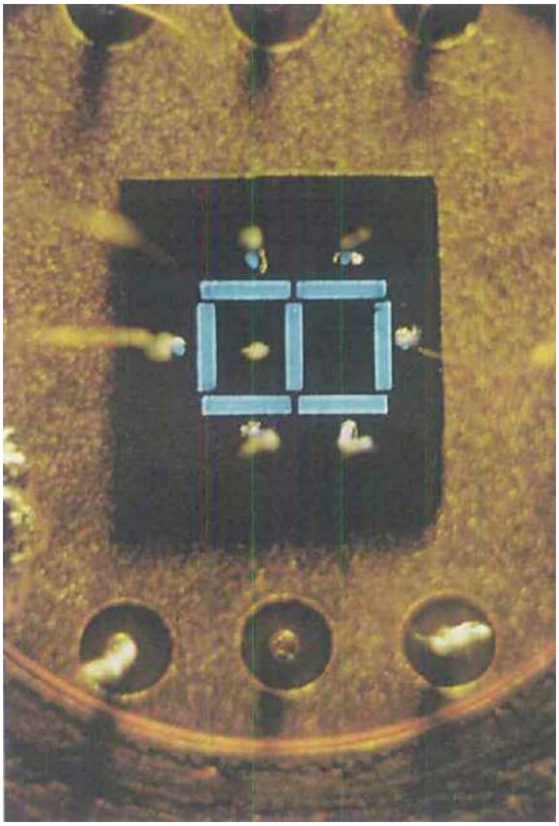

Figure 5 A seven-segment II–VI LED with blue emission taken at room temperature. Reproduced with permission from Hagerott M, Jeon H, Nurmikko AV, *et al.* (1992) Indium tin oxide as transparent electrode material for ZnSe-based blue quantum well light emitters. *Applied Physics Letters* 60: 2825–2827.

absorbed by the II–VI semiconductor, it can in turn produce light of its own. This light, known as luminescence, is again characteristic of the particular material and is the basis for the use of such materials in light-emitting diodes (LEDs) and laser diodes (LDs). In Figure 5 we show the luminescence from a II–VI LED. We see also from Figure 4 that at very long wavelengths, the ZnSe material also exhibits a large reduction in light transmission. This spectral region is termed the far infrared and the light absorbed at this spectral region produces vibrations of the atoms within the material, referred to as lattice absorption.

Light Scattering

Within the bulk of the material, small areas may have a different arrangement of atoms or may contain a substantial number of foreign atoms known as impurities. A large grouping of these impurities is known as an inclusion. In these regions, the optical properties will be different from the bulk of the material. Differences in both absorption coefficients and refractive indices can be found. These different refractive indices cause the light passing through the material to be scattered. Poor-quality material will often contain many inclusions. Even to the eye, these crystals will not look highly transparent. Changes in the refractive indices can also occur in highly pure materials due to vibrations of the atoms in the crystals. The lattice vibrations change the distance between the atoms in the material which alters the electric fields within the material and thereby changes the optical properties. Light can also be scattered due to these vibrations; experimental measurements of light scattering can provide useful information on the vibrations in these materials.

Optical Properties of II–VI Semiconductors: More Advanced Concepts

Light–Semiconductor Interactions

Light is an oscillating electromagnetic field whose interaction with semiconductor materials is in terms of the complex dielectric constant $\varepsilon_c = n_c^2$ where n_c is the complex refractive index defined as $n_c = n - ik$. k is known as the extinction coefficient and the value of k determines the amount of absorption in a material. The optical processes of reflection and absorption in semiconductors are governed by this complex dielectric constant. It turns out that there is a strong relationship between these processes. If the absorption spectrum is determined at all frequencies, then we can determine the refractive index spectrum at all frequencies by what is known as the Kramers–Kronig relations. From this description, we can now refine eqn [2] presented above:

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2} \qquad [3]$$

If the material is transparent at a particular wavelength, then $k = 0$ and we can determine the refractive index n. This is the case for values given in the table. The technique of spectroscopic ellipsometry can be used to determine the optical response of materials over a large wavelength range. This technique relies upon the change from linear polarization to elliptical polarization for light reflected from a material. From these measurements, the dielectric response can be determined.

If we look again at Figure 3, we see that there is a strong change in sample transmission around 0.5 μm. This corresponds to an energy of 2.18 eV and is the bandgap energy of ZnSe. According to the details presented above, transitions are now being made from the full valence to the empty conduction bands of the semiconductor material. Once we have exceeded the bandgap energy, we can make further transitions between the bands. The spectrum shown in the figure is a broad band, reflecting the change from sharp atomic transitions to broad transitions in the solid state. Light with a wavelength greater than 0.5 μm does not have sufficient energy to effect the transitions between the bands. These wavelengths are not absorbed in the material. The material is transparent and the value of the extinction coefficient k is zero. The orange color of the crystal in Figure 2 corresponds to this transparent spectral region.

Cooling samples to low temperatures close to absolute zero has a profound effect on the optical properties of materials, particularly the absorption spectrum. In Figure 6, we show the absorption spectrum of a ZnSe sample at a temperature of 2 K. The absorption spectrum is determined from the measured transmission by the following equation

$$\alpha = \frac{1}{L}\ln\frac{I_0}{I_t} \qquad [4]$$

where α is the absorption coefficient, L the length of the sample, I_0 the light intensity incident on the sample and I_t the light intensity transmitted through the sample. In addition to the broad band spectrum there are now a number of sharp spectral lines near the bandgap energy.

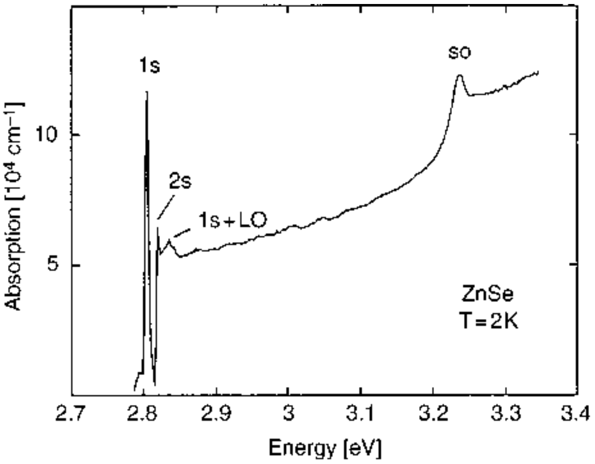

Figure 6 Low-temperature optical absorption spectrum of a ZnSe layer showing the 1S and 2S exciton features, the 1S + 1LO-phonon transition and the split-off (SO) exciton. Reproduced with permission from Bhargava R (ed.) (1997) *Wide Bandgap II–VI Semiconductors*, The Institute of Electrical Engineers: London, Figure 2 of Chapter 4.1.

These lines are due to exciton formation. An exciton is an electron–hole pair bound together by their mutual Coulomb interaction. It is the solid state equivalent of the Hydrogen atom (*see* **Semiconductor Physics:** Excitons). The bandgap energy at 2 K has increased from its value at 300 K. This is due to the thermal changes in the interatomic distance and the vibrational properties of the materials. Values of the bandgap energies determined from the position of the exciton lines in absorption spectra at low and at room temperatures are given in Table 1. At low temperatures, the luminescence is dominated by excitons, and we observe sharp line transitions. The width of the exciton transition is an excellent indicator of the quality of the material, and is often used by crystal growers to determine the quantity of defects in their materials.

It is possible for excitons and for free carriers (those that are not bound together) to recombine without giving out light. This nonradiative process usually involves the creation of local vibrations of the crystal lattice in the vicinity of defects in the material. We can define the radiative efficiency η_r as follows

$$\eta_r = \frac{R_r}{R_r + R_{nr}} \qquad [5]$$

where R_r is the radiative transition rate and R_{nr} is the nonradiative transition rate. In order to develop new optical devices based on II–VI semiconductors, it is essential that the value of R_{nr} be as small as possible. As R_{nr} is directly related to the concentration of defects in these materials, one can see why it is essential to make semiconductors with as few defects as possible. We can study these effects using a pulsed light source. Carriers are created during the pulse of light and recombine after the pulse has stopped. If we measure the luminescence intensity versus time, we can deduce the lifetime, which is given as

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \qquad [6]$$

where τ_r is the radiative and τ_{nr} the nonradiative lifetime, respectively. By varying the carrier density we can separately determine the contribution to the lifetime from radiative and nonradiative effects. In II–VI materials at room temperature, lifetimes are of the order of 1 ns.

Quantum Wells, High Carrier Density Effects, and Lasers

Recent developments in the growth of semiconductor materials has allowed the production of samples with monolayer accuracy (3–5 Å). If we grow a layered structure with materials of high bandgap on either side of a material of lower bandgap, we can observe quantum mechanical effects. These materials are called quantum wells (*see* **Semiconductor Physics:** Quantum Wells and GaAs-Based Structures). As an example, a 75 Å ZnSe quantum well can be grown between layers of ZnSSe. Due to the small size of the active material, quantum mechanical effects come into play, most notably an increase in the bandgap energy as the width of the quantum well decreases. The effect of placing this material between ZnSSe layers that have a larger bandgap is to confine the electrons and holes to the small region of the well. In this small region, they are less likely to encounter defects and so quantum well materials usually have a higher radiative efficiency than bulk II–VI materials which are grown by standard techniques. In Figure 7, we show the exciton absorption and luminescence from a ZnSe quantum well at low temperature.

If we increase the excitation level in these materials, we can eventually cause them to operate as lasers. At high carrier densities, nonradiative transitions become saturated and radiative efficiency increases. As the refractive indices in II–VI materials are quite large, the reflectivity of the two ends of the sample are usually sufficient to act as the mirrors for the laser cavity. All the II–VI semiconductor materials have exhibited laser action due to their direct bandgap. Most of the early studies of lasing were carried out at low temperature. In recent times, lasing action has been observed at room temperature in CdZnSe quantum well materials and there are attempts at present to develop commercial II–VI lasers for use in optical storage devices and flat panel displays. In Figure 8 we show the transition from luminescence through gain and into lasing action for a CdZnSe quantum well material.

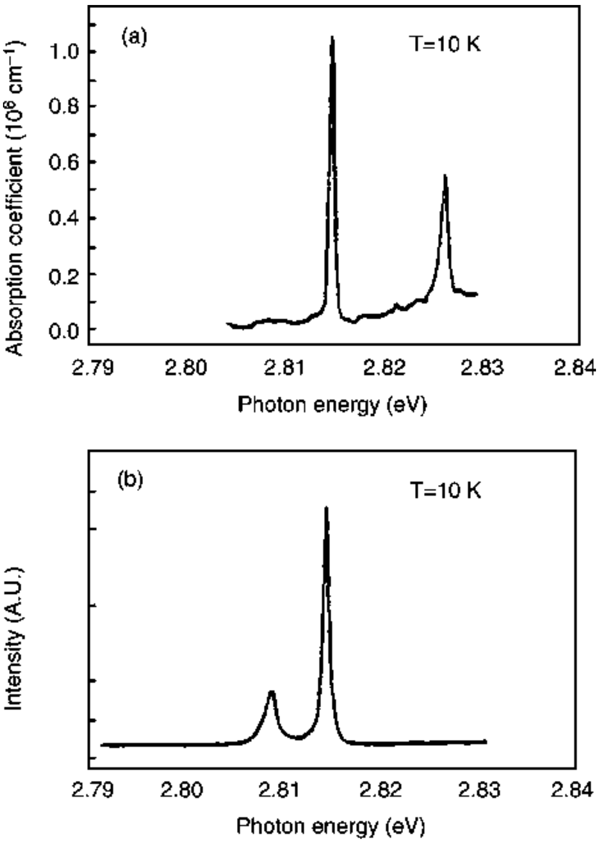

Figure 7 (a) Absorption and (b) luminescence from a single ZnSe quantum well at $T = 10$ K in the vicinity of the $n = 1$ quantum well exciton transition. Reprinted from Nurmikko AV, Jeon H, Gunshor RL and Han J (1996) II–VI lasers – directions. *Journal of Crystal Growth*, 159: 644–651 with permission from Excerpta Media Inc.

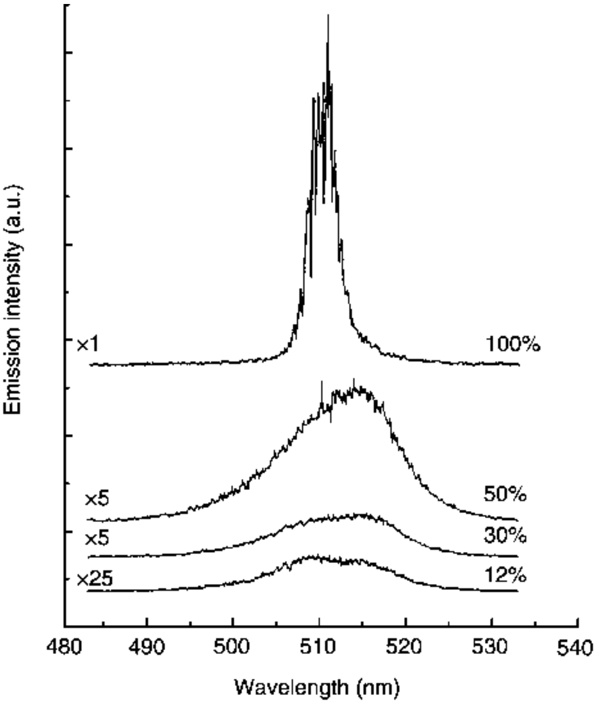

Figure 8 Emission spectra above and below lasing for a 40 Å $Cd_{0.28}/Zn_{0.72}$ Se/ZnSe system at room temperature. 100% corresponds to 1.1 times lasing threshold. Reprinted from Donegan JF, Jordan C, Rees P, Logue F, Heffernan JF and Hegarty J (1996) A study of internal loss in CdZnSe/ZnSe multiple quantum well materials. *Journal of Crystal Growth*, 159: 653–656 with permission from Excerpta Media Inc.

Dopants and Impurities

The key to the large-scale use of semiconductor materials is the large change in the electrical conductivity of the materials with the incorporation of small amounts of other elements. This process is known as doping and the foreign elements are known as dopants. The recent development of II–VI laser diodes in ZnSe materials is based on the p-n diode junction. Electrons injected into the n-type material recombine with holes injected into the p-type material at the junction between these two materials. A particular type of luminescence transition unique to semiconductor materials is the donor–acceptor pair (DAP) recombination. As donor and acceptor atoms are located in specific crystalline sites with respect to one another, we observe that the recombination energy is dependent on separation. Pairs that are close together often exhibit sharp lines while those far apart merge into a broad band. In Figure 9 we show the DAP recombination in ZnSe at 1.6 K.

It is also possible to incorporate foreign atoms in II–VI materials that have a small effect on the electrical properties but significantly change the optical properties. Such foreign atoms are termed impurities to distinguish them from dopants. An example is Mn^{2+} in ZnS. The Mn ions take the place of a small number of Zn atoms in the lattice. The luminescence from such a material is a broad emission in the visible region due to optical transitions on the Mn ions. ZnS:Mn materials are used as phosphors in flat-panel displays.

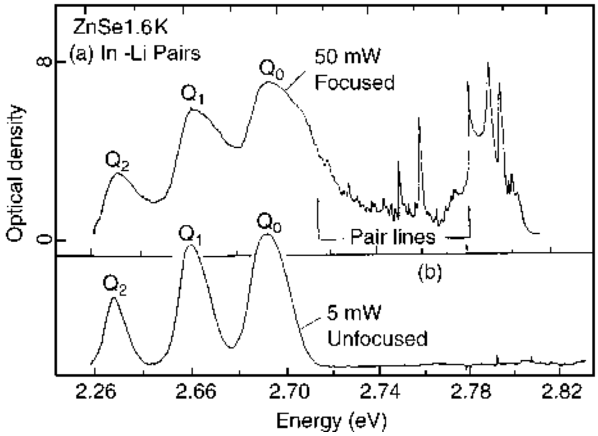

Figure 9 DAP recombination spectra of In-Li pairs in ZnSe showing pair lines, pair band (Q_0) and LO-phonon replicas (Q_1 and Q_2) at two different intensities (a) and (b). Adapted with permission from Merz JL, Nassan K and Shiever JW (1973) Pair spectra and the shallow acceptors in ZnSe. *Physical Review B* 8: 1444–1452.

Effects of Temperature, Pressure, Electric, and Magnetic Fields

The optical properties of II–VI materials are subject to variation due to temperature changes, external pressure and fields. Increasing temperature is found to lead to a decrease in the bandgap of II–VI materials. This decrease is due to the combined effects of an increase in the lattice spacing with temperature and the electron–phonon interaction. As the population of phonons is temperature dependent, this effect increases strongly as temperature increases and is the major contributor to the change in bandgap with temperature. A second effect of increasing the temperature is the change in the luminescence properties. In particular, it is found that the luminescence efficiency decreases and this is reflected in a reduced luminescence lifetime at room temperature compared with low-temperature values. Looking again at eqn [6], it is the nonradiative lifetime that changes rapidly with temperature and results in the decrease in the luminescence efficiency.

Pressure changes may be hydrostatic or uniaxial. In the hydrostatic case, the atoms are all brought closer together, and for the II–VI materials, this results in an increase in the bandgap energy. On the other hand, a uniaxial stress results in a decrease in the bandgap in the direction of the stress and an increase in the transverse directions. In II–VI materials with a wurtzite structure, the situation is more complex with two longitudinal and four transverse deformations possible.

External electric and magnetic fields can also be applied to II–VI semiconductors and thereby the optical properties can be changed. Electric fields polarize the materials and can lead to a reduction in the optical strength of transitions due to this polarization-induced carrier separation. Magnetic fields, on the other hand, lead to a splitting of the conduction and valence bands into Landau levels due to the circular motion of the carriers around the direction of the field.

See also

Semiconductor Physics: Band Structure and Optical Properties; Excitons; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Adachi S and Taguchi T (1991) Optical Properties of ZnSe. *Physical Review B* 43: 9569–9577.

Bhargava R (ed.) (1997) *Wide Bandgap II-VI Semiconductors*. London: Institute of Electrical Engineers.

Madelung E (ed.) (1982) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 17b: Physics of II-VI and I-VII Compounds, Semimagnetic Semiconductors*. Berlin: Springer-Verlag.

Madelung E (ed.) (1987) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 22a: Intrinsic Properties of Group IV Elements and III-V, II-VI and I-VII Compounds*. Berlin: Springer-Verlag.

Mar H (1992) Optical properties of wide bandgap II-VI semiconductors. In: Ruda HE (ed.) *Widegap II-VI Compounds for Opto-electronic Applications*, pp. 167–188. London: Chapman and Hall.

Nurmikko AV and Gunshor RL (1993) Physics and application of II-VI quantum confined structures. In: Henneberger F, Schmitt-Rink S and Göbel EO (eds) *Optics of Semiconductor Nanostructures*, pp. 229–262. Berlin: Akademie Verlag.

Pankove J (1975) *Optical Processes in Semiconductors*, 1st edn. New York: Dover Publications.

Stanley RP and Hegarty J (1993) Physics of excitons in wide bandgap II-VI quantum wells. In: Henneberger F, Schmitt-Rink S and Göbel EO (eds) *Optics of Semiconductor Nanostructures*, pp. 263–290. Berlin: Akademie Verlag.

Yu PY and Cadona M (1996) *Fundamentals of Semiconductors*. Berlin: Springer Verlag.

II–VI Compounds 1999. Ninth International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 214–215, 2000.

II–VI Compounds 1997. Eighth International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 184–185, 1998.

II–VI Compounds 1995. Seventh International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 159, 1996.

Lead Salts

G Bauer and G Springholz, Johannes Kepler Universität, Linz, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The binary IV–VI compounds formed from Pb and group-VI elements like S, Se, and Te are among the oldest known semiconducting materials, and have been used for many years for electronic and optoelectronic device applications. The first ever reported solid state diode was made from single crystalline PbS and its rectifying properties were exploited in early radio receivers. Later on, interest in IV–VI semiconductors shifted to mid-infrared opto-electronic device applications such as photon detectors operating in the 3–14 μm wavelength range, taking advantage of the narrow energy gap between the valence and conduction band of the IV–VI compounds. Soon after the demonstration of lasing in III–V semiconductor diode structures, the first mid-infrared p-n junction laser was reported using $Pb_{1-x}Sn_xTe$, and since then efficient mid- and far-infrared IV–VI compound diode lasers have been fabricated, finding their main applications for remote sensing of gaseous pollutants in trace gas sensing devices, for toxic gas analysis systems, for human breath analysis in medical diagnostics, and for fabrication process control. Recently, IV–VI multiquantum-well structures have also attracted a lot of attention for their potential as efficient thermoelectric devices.

Basic Properties

In lead salt compounds, both the valence band maximum and the conduction band minimum occur at the L-points of the Brillouin zone. Thus these many-valley semiconductors have a direct gap, but not in the center of the Brillouin zone. In addition the small values of the energy gaps ($E_g < 410$ meV at $T = 300$ K) lead to a strong $\mathbf{k} \cdot \mathbf{p}$ interaction which in turn results in a strong nonparabolicity of the $E(k)$ dispersion relations. The small energy gaps also lead to small effective masses of electrons and holes, and the strong spin–orbit interaction (Pb is a very heavy element) leads to large values of the g-factors.

The bandgaps of the lead compounds have a positive temperature coefficient ($dE_g/dT > 0$) and a negative pressure coefficient ($dE_g/dp < 0$), i.e., the gaps increase with temperature and decrease with hydrostatic pressure. These anomalous properties are related to peculiarities of the band structure, among them the fact that in the lead compounds the Pb 6s levels are below the top of the valence band, and thus are occupied, in contrast to the group IV, III–V and II–VI compounds.

Almost all optical device applications of the IV–VI semiconductors are based on the materials subgroup of cubic lead chalcogenides PbTe, PbSe, and PbS, and their quasibinary alloys with other chalcogenide such as compounds, like PbSnTe, PbSnSe, or ternary alloys like PbEuTe, PbSrTe, PbEuSe, PbSrSe, etc. These compounds all crystallize in the rocksalt structure with sixfold coordination and lattice constants in the 5.9 to 6.5 Å range. Mechanically, the lead salt compounds are much softer than the tetrahedrally bonded semiconductors, which has significant implications for device fabrication processes.

A further unusual property of the lead compounds is their huge static dielectric constant ε_0. In particular PbTe is close to a structural phase transition from the NaCl structure to a rhombohedral modification with ferroelectric properties. GeTe and SnTe are indeed ferroelectric with Curie temperatures of about 700 K and 145 K respectively, and also PbGeTe and PbSnTe become ferroelectric when the Ge or Sn content exceeds 1% or 20%, respectively. PbTe exhibits a softening of the transverse optic phonon mode with decreasing temperature. Although PbTe remains cubic, the temperature dependence of the static dielectric constant $\varepsilon_0 \propto C/(T-\theta)$ can be associated with an extrapolated negative Curie temperature θ of −75 K. As a consequence, the static dielectric constant increases strongly with decreasing temperature, reaching a value of $\varepsilon_0 = 1350$ at 4 K. This is the reason for the strong increase of the static dielectric constant at low temperatures as shown in Table 1, in which important properties of PbTe, PbSe, and PbS are summarized.

Also, the high-frequency dielectric constants ε_∞ of the lead salts are extremely high, which leads to refractive indices of about 4.4 for PbS, to 5.0 for PbSe and to about 6.0 for PbTe at $T = 300$ K, for photon energies just below the respective energy gaps.

As a result of the huge static dielectric constants, the hydrogen-like impurity binding energies are in the range of μeV, and the carrier scattering at ionized impurities is effectively screened. Consequently, the low-temperature carrier mobilities are high, exceeding 10^6 cm^2/Vs in PbTe. For doping of the lead salts two alternative methods are used, namely (1) adjustment of the stoichiometry of the compounds, or (2) doping with various kinds of impurities. Because the electronic states of these vacancies are resonant within the conduction or valence bands, each metal (Pb or Sn) vacancy acts as a doubly charged acceptor, and each chalcogen (Te, Se, or S) vacancy acts as a doubly charged donor. As a result, background doping levels below 10^{17} cm^{-3} are difficult. For doping with impurities, atoms from group IIIa (In, Ga, Tl), group Va (As, Sb, Bi), as well as group Ib elements (Cu, Ag) are used. These dopants are amphoteric p- or n-type, depending on whether they are incorporated substitutionally on metal or chalcogen lattice sites.

Table 1 Physical properties of some of binary IV–VI semiconductors

Parameter	*PbTe*	*PbSe*	*PbS*
Lattice constant a_0 (Å)	6.462	6.124	5.936
Melting point (°C)	930	1080	1113
Energy gap E_g			
at 300 K (meV)	319	278	410
at 77 K (meV)	217	176	307
at 4 K (meV)	190	147	286
Static dielectric constant ε_0			
at 300 K	414	210	169
at 77 K	1000	227	181
at 4 K	1350	280	–
High-frequency dielectric constant ε_∞	34	21	17
TO phonon mode frequency	18 (4 K) 33 (300 K)	37	65

Table 2 Expansion coefficients (10^{-6} K^{-1})

Compound	*Expansion coefficient*
PbS	20.3
PbSe	19.4
PbTe	19.8
GaAs	6
Si	2.6
BaF$_2$	18.8

Other important aspects of the lead salts are their rather large linear thermal expansion coefficients of around 20×10^{-6} K^{-1} which are about a factor of three larger than that of GaAs, and six times larger than that of Si as shown in Table 2. The thermal conductivity is much lower than that of other semiconductors, which is favorable for their use in thermoelectric devices but causes problems in the removal of the dissipated heat in IV–VI compound diode lasers.

Band Structure

A main property of the electronic band structure of the lead salt compounds is the small and direct energy gap of less than 410 meV (see Table 1). Since the maxima of the valence band and the minima of the conduction band are located at the same wavevector in reciprocal space, optical photons can be directly absorbed or emitted at the band edges. This makes the lead salts very well suited for opto-electronic applications such as infrared photon detectors or emitters. The band extrema are located at the edges in the <111> direction of the Brillouin zone (L-points). In contrast to most other semiconductors, the conduction and valence bands are nearly mirror symmetric with almost equal effective masses for the electrons and holes. Because of the narrow energy gaps, the energy bands are strongly nonparabolic and the effective masses of electrons as well as holes are rather small (0.02–0.08 m_0 for the transverse masses,

Table 3 Band edge values of effective masses (in units of m_0) and g-factors (absolute values), $T = 2$ K

Parameter	*PbTe*	*PbSe*	*PbS*
m_t^{CB}	0.0207	0.037	0.081
m_l^{CB}	0.0213	0.0687	0.108
m_t^{VB}	0.0241	0.036	0.076
m_l^{VB}	0.263	0.066	0.108
g_t^{CB}	16.7	32.5	–
g_l^{CB}	66.1	44.1	12
g_t^{VB}	13.5	30.6	–
g_l^{VB}	65.6	–	13

Table 4 Pressure and temperature coefficients of the energy gaps

	PbTe	*PbSe*	*PbS*
dE_g/dp(eV Pa^{-1})	-7.5×10^{-11}	-9.1×10^{-11}	-9.1×10^{-11}
dE_g/dT (meV K^{-1})	0.457	0.457	0.462

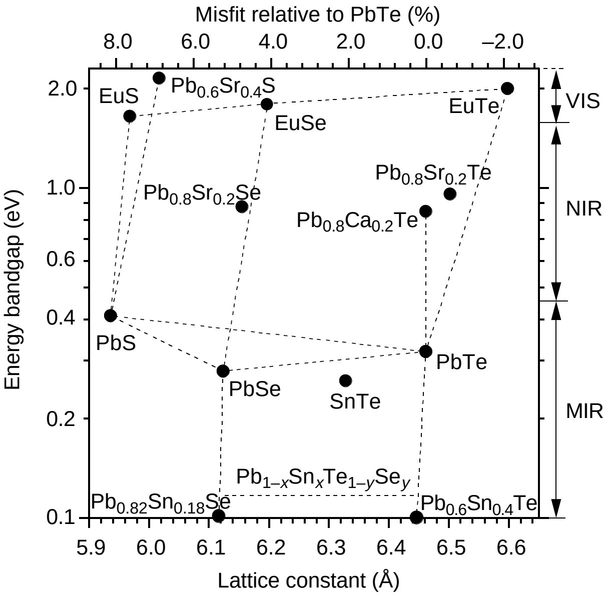

Figure 1 Energy bandgap versus lattice constant for several IV–VI compounds and alloys.

see Table 3). In addition, the bands are anisotropic, i.e., the Fermi surfaces are elongated ellipsoids of revolution around the <111> axes, characterized by a longitudinal and transverse effective mass m_l and m_t parallel and perpendicular to the four-fold <111> directions.

For PbTe the effective mass anisotropy is very large with a ratio of about 10 between longitudinal and transverse masses, whereas for PbSe and PbS this ratio is only 1.8 and 1.2, respectively (Table 2). Furthermore, in the alloy system $PbSe_{1-x}Te_x$ the energy gap shows a bowing, but not with a minimum value but rather a maximum gap for x at about 0.7. This is in contrast to the behavior found in direct-gap zincblende semiconductors: the bandgap $E_g(x)$ of an $A_{1-x}B_x$ alloy can generally be represented by $E_g(x) = (1-x)E_g(A) + xE_g(B) - bx(1-x)$ where b denotes the bowing coefficient. For III–V and II–VI compounds $b > 0$; whereas, for PbSeTe alloys $b < 0$.

Apart from the above-mentioned negative pressure coefficients of the direct gaps and their positive temperature coefficient which are tabulated in Table 4, there is a further intriguing anomaly in the chemical trends of the fundamental gap at the L-points: it decreases from PbS to PbSe, but increases from PbSe to PbTe, i.e., the direct gaps do not decrease monotonically with increasing atomic number of the anions. This decrease of the energy bandgap with decreasing temperature is substantial: i.e. it decreases from room temperature to 4 K by about 130 meV (see Table 1). This represents a large relative change with respect to the absolute value of the bandgap, and thus the emission of lead salt based diode lasers can be tuned over a broad wavelength range just by changing the operation temperature. This effect is actually utilized for spectroscopic applications of these lasers.

For adjustment of the fundamental absorption edge of infrared detectors as well as the adjustment of the wavelength emission of lead salt diode lasers, alloying of the lead salt compounds with other chalcogenide compounds is used. As illustrated in Figure 1, for this purpose many different ternary and quaternary alloys have been used. Important alloy systems for far-infrared applications are the lead–tin chalcogenide alloys. The tin chalcogenides (SnS, SnSe, SnTe) form single-phase pseudobinary alloy systems with the lead salts for all compositions. They have similar band structures as compared to the lead salt compounds; however, the arrangement of valence and conduction bands is exchanged. As a result, with increasing Sn content the energy bandgap of the ternary lead–tin chalcogenide alloys decreases and a zero bandgap semiconductor is reached for a certain ternary composition. For an operation temperature of 77 K, the bandgap is zero for a Sn content of $x_{Sn} \sim 40\%$ in the case of $Pb_{1-x}Sn_xTe$, and of $x_{Sn} \sim 20\%$ in the $Pb_{1-x}Sn_xSe$ case.

Optical Properties

The strong fundamental absorption in the lead salt compounds reflects the particularly high joint density of states at the direct gap at the L-points of the Brillouin zone. The main advantage of the lead compounds with respect to other direct narrow-gap semiconductors like InSb, or HgCdTe alloys, with band extremes at the center, i.e., the Γ-point of the Brillouin zone, is related to the differences in the effective masses. Whereas the transverse effective masses of the IV–VI compounds have comparable values to the effective masses of the narrow-gap III–V and II–VI compounds, the longitudinal effective masses of the lead compounds are much larger. Thus the mass anisotropy together with the fourfold valley degeneracy at the L-point of the Brillouin zone leads to the substantially higher joint density of states and corresponding large absorption constants. Due to the rapid change of $\alpha(\omega)$ in the fundamental absorption range, a peak appears in the refractive index, $n(\omega)$, as shown in Figure 2 for PbTe, PbSe and for PbS. The shift of the peak with increasing temperature to higher photon energies reflects the increasing energy gaps.

In two-dimensional structures the changes of the absorption constant with frequency are particularly steep for energies which correspond to the onset of absorption between quantum confined electric subbands in the valence and conduction band. Such step-like features have indeed been observed in the absorption of PbTe/PbEuTe multiquantum-well (MQW) structures by Yuan *et al.* In Figure 3 the absorption constants and refractive indices are shown for a series of three MQW samples with different PbTe well widths. The steps in the absorption coefficients for the interband transitions associated to the longitudinal valleys with their main axis parallel to the [111] growth direction, denoted by ‘l’ as well as those associated by the oblique valleys, denoted by ‘o’ are accompanied by corresponding cusp-like enhancements of the refractive index. These cusps are of importance for the proper design of MQW laser structures.

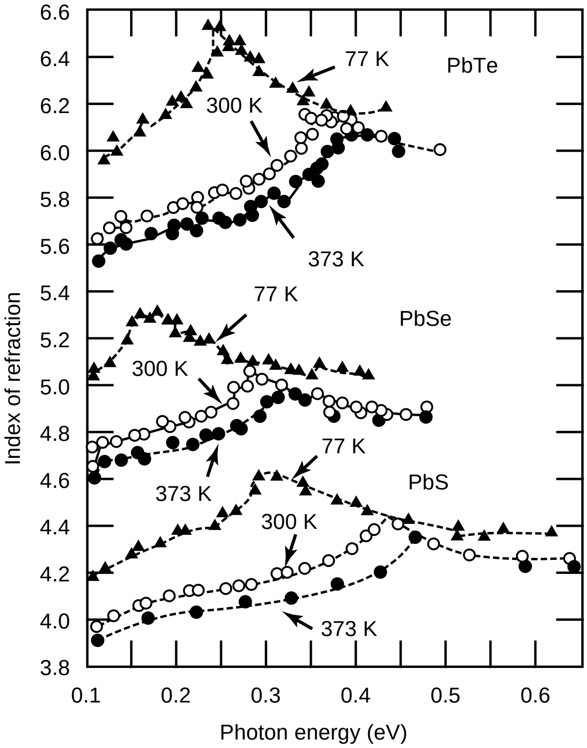

Figure 2 Refractive indices of PbTe, PbSe, and PbS as a function of photon energy at 77, 300, and 373 K. Reproduced with permission from Zemel JN, Jensen JD and Schoolar RB (1965) Electrical and optical properties of epitaxial films of PbS, PbSe, PbTe and SnTe. *Physical Review* 140: A330–A337. Copyright 1965, American Physical Society.

Because of the extremely high values of the refractive indices, the lead salts can easily be combined with materials of much lower refractive indices like EuTe, EuSe, or BaF_2 for efficient multilayer Bragg mirror structures. The high reflectivity of Bragg mirrors is caused by constructive interference of electromagnetic waves reflected at the subsequent interfaces of a multilayer structure. To obtain constructive interference, all interfaces have to be parallel, and, depending on the refractive indices of the media inside the Bragg mirror, the distance between subsequent interfaces should be an even or odd multiple of one quarter of the optical wavelength λ. In practice, stacks of two alternating layers with different refractive indices n_1 and n_2 and with a thickness equal to a quarter optical wavelength are used.

If for these $\lambda/4$ layers $Pb_{0.93}Eu_{0.07}Te$ and EuTe are used, a high index contrast of 68% is obtained and therefore very high reflectivities can be achieved already by a very small number of layer pairs. It was shown that a Bragg mirror with 3.5 periods of EuTe/$Pb_{0.93}Eu_{0.07}Te$ layer pairs for a target wavelength of 3.8 μm has only a total thickness as small as 2.55 μm. For comparison with results obtained for other material combinations it is useful to relate the stop bandwidth $\Delta\lambda$ to the center wavelength λ_T of the Bragg mirror. For EuTe/$Pb_{0.93}Eu_{0.07}Te$ this relative stop bandwidth $\Delta\lambda/\lambda_T$ is as large as 50%. An even wider stop bandwidth of 59% was demonstrated for a mirror with PbTe and EuTe $\lambda/4$ layers. This value represents, to the best of our knowledge, the highest relative stop bandwidth obtained for molecular beam epitaxy (MBE) grown Bragg reflectors.

Infrared Lasers

Among the infrared diode lasers, those based on IV–VI compounds play a particularly important role. They cover a wavelength range from 2.5 μm to about 30 μm. In fact, up to now most commercially

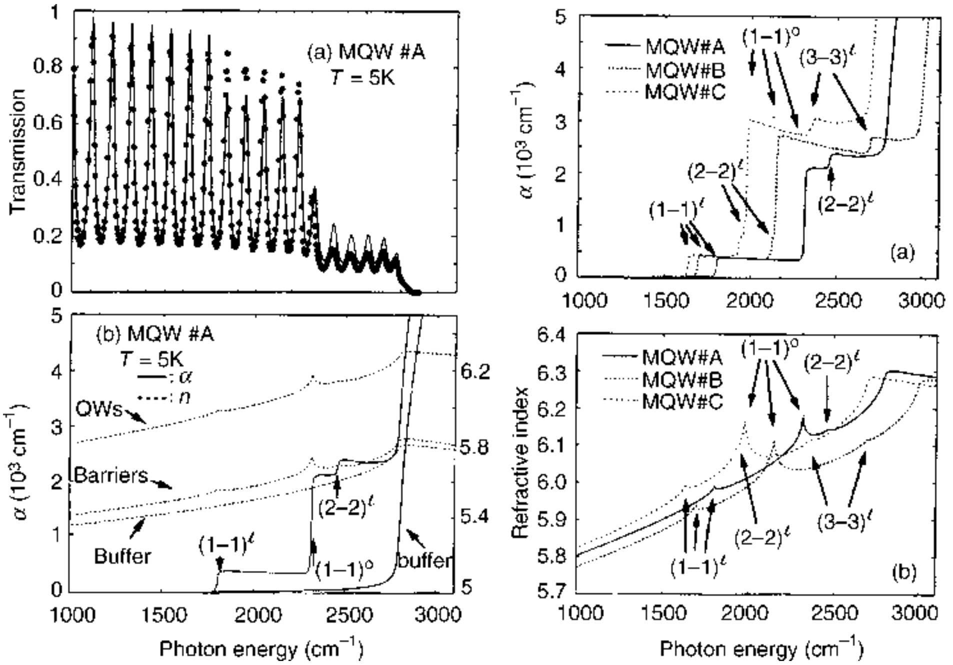

Figure 3 Left panel: (a) transmission versus energy for a PbTe/Pb$_{1-x}$Eu$_x$Te multiquantum-well sample A (6.2 nm/62.1 nm, $x = 0.026$, 46 periods). Dots: experimental data; full line: calculation based on an envelope function model. (b) Frequency dependence of the absorption constant (full line) and of the refractive index (dotted lines). The arrows indicate interband transitions between electric sub-bands associated with valleys with main axis parallel to the [111] growth direction (l) and oblique to it (o). Right panel: (a) Comparison of the absorption constant of three PbTe/PbEuTe MQW samples (B: 9.3 nm/55.4 nm, $x = 0.03$, 30 periods; C: 11.8 nm/48.6 nm, $x = 0.024$, 30 periods). (b) Refractive index versus energy. Steps in the absorption constant due to onset of interband transitions are associated with cusps in the refractive index. Reproduced with permission from Yuan S, Springholz G, Bauer G and Kriechbaum M (1994) Electronic and optical properties of PbTe/Pb$_{1-x}$Eu$_x$Te multiple-quantum-well structures. *Physical Review B* 49: 5476–5489, Copyright 1994, American Physical Society.

available mid- and far-infrared laser diodes are made from IV–VI compounds. The main material properties which are advantageous in this respect are (i) the electronic band structure with mirror-like bands at the L-point of the Brillouin zone and the resulting high joint density of states; (ii) the lower nonradiative Auger recombination rate in IV–VI compounds as compared to III–V and II–VI materials, and (iii) the high dielectric constant and the resulting weak detrimental influence of defects on nonradiative recombination losses. At wavelengths longer than 3 μm, IV–VI interband diode lasers are superior to their III–V counterparts for reaching high operating temperatures. So far IV–VI compound lasers hold extreme high cw operating temperature at wavelengths larger than 3 μm. With PbSe/PbSrSe diode lasers pulsed operation up to 60 °C was reported, whereas cw operation up to 223 K was achieved with separate confinement buried PbTe/PbEuSeTe heterostructure laser. Embedding PbTe quantum wells as the active medium in PbEuTe barriers between two dielectric Bragg mirrors optically pumped pulsed lasing was observed up to 65 °C. The recently developed III–V based quantum cascade lasers are rapidly improving and are becoming important competitors for IV–VI lasers. The commercial IV–VI lasers, however, are still based on conventional designs like homojunction, heterojunction, or single and multiple heterojunction lasers.

A unique property of lead salt laser diodes is the ease of tuning of the emission wavelength due to the strong change of bandgap of the IV–VI materials with increasing temperature. This, together with their narrow linewidths, makes these lasers ideally suited for spectroscopic applications, including research in molecular spectroscopy of ions or radicals in heterodyne and acoustic spectroscopy, or Doppler spectroscopy of molecular beams. So far the main applications are in time resolved combustion analysis in the automotive industry, in the monitoring of emissions of power plants, in exact monitoring of gas compositions in chemical plants, and spatially and time resolved *in situ* analysis.

Apart from the significant advances with IV–VI edge emitting diode lasers, vertical cavity surface

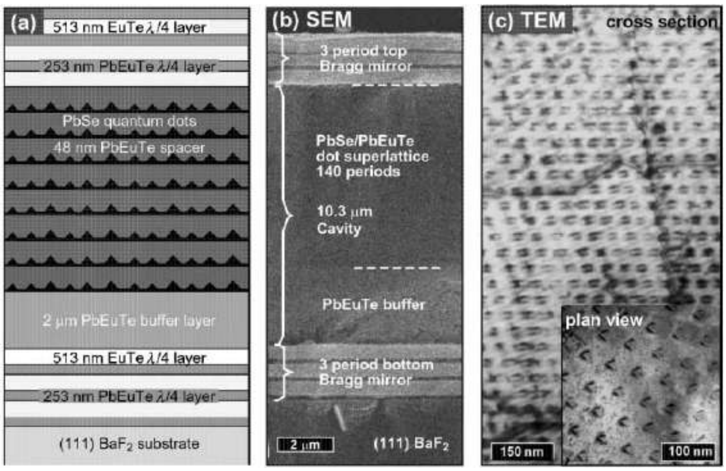
(a)
513 nm EuTe λ/4 layer
253 nm PbEuTe λ/4 layer
PbSe quantum dots
48 nm PbEuTe spacer
2 μm PbEuTe buffer layer
513 nm EuTe λ/4 layer
253 nm PbEuTe λ/4 layer
(111) BaF₂ substrate
(b) SEM
3 period top Bragg mirror
PbSe/PbEuTe dot superlattice 140 periods
10.3 μm Cavity
PbEuTe buffer
3 period bottom Bragg mirror
2 μm
(111) BaF₂
(c) TEM
cross section
plan view
150 nm
100 nm

Figure 4 (a) Schematic cross-section of a PbSe quantum-dot VCSEL structure; (b) corresponding cross-sectional scanning electron micrograph; and (c) cross-sectional and plan-view transmission electron micrographs of a PbSe/PbEuTe dot superlattice sample reference sample. Reproduced with permission from Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Midinfrared surface emitting PbSe/PbEuTe quantum dot lasers. *Applied Physics Letters* 79: 1225–1227. Copyright 2001, American Institute of Physics.

emitting lasers (VCSELs) have been successfully produced with optically pumped pulsed mode operation up to 50 °C. These lasers have attractive features like low-divergence circular output beam, low active volumes, and low thresholds.

High-finesse vertical cavity structures have also made it possible to realize mid-infrared surface-emitting PbSe/PbEuTe quantum dot lasers. As shown in Figure 4 in this case a self-organized PbSe/PbEuTe quantum dot superlattice was introduced into the active cavity region between Bragg mirrors consisting of three λ/4 layer pairs of 513 nm EuTe alternating with 253 nm $\mathrm{Pb}_{0.94}\mathrm{Eu}_{006}\mathrm{Te}$. Optically pumped VCSEL emission at $\lambda = 4.2–3.9$ μm was achieved up to 150 K. The PbSe dots exhibit a face-centered cubic three-dimensional stacking in the superlattice. Lasing occurs simultaneously at the $m = 28^{\text{th}}$ and 29th order cavity modes, which shifts with increasing temperature to the $m = 29^{\text{th}}$ and $m = 30^{\text{th}}$ mode as shown in Figure 5. This simultaneous stimulated two-mode laser operation is a result of the inhomogeneously broadened quantum-dot gain spectrum, due to PbSe dot size fluctuations.

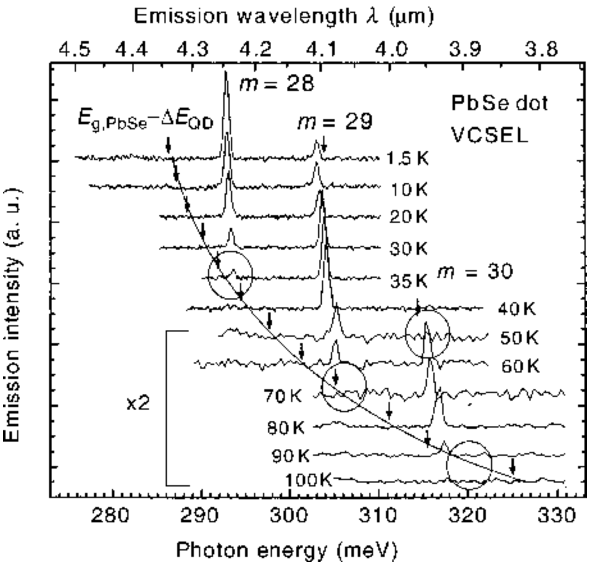

Figure 5 PbSe dot VCSEL emission spectra at various temperatures between 1.5 and 100 K for a pump power twice the threshold, showing the switching of the laser emission to higher cavity modes as the temperature increases. The arrows and the dashed line indicate the low-energy edge of the quantum-dot gain spectrum given by $E_{\mathrm{g,PbSe}}(T) + \Delta E_{\mathrm{QD}}$, where $\Delta E_{\mathrm{QD}} = 141$ meV. Reproduced with permission from Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Midinfrared surface emitting PbSe/PbEuTe quantum dot lasers. *Applied Physics Letters* 79: 1225–1227. Copyright 2001, American Institute of Physics.

Infrared Detectors

The photoconductive properties of PbS and PbSe have been used for the detection of infrared radiation already since about 1930. These detectors are still mass produced either by vacuum deposition or precipitation from aqueous solution in the presence of an oxidizing agent. The polycrystalline films deposited on glass or similar substrates are usually about 1 μm thick with grain sizes of about 1 μm.

Apart from polycrystalline thin-film detectors also single crystalline films have been used for infrared detection. In particular, the photoconductive response of a sequence of n- and p-doped PbTe multilayers (so-called nipi structures) was shown to come close to that for the theoretical limit for the detectivity D^* of an ideal photodetector at temperatures between 77 and 90 K. Recently, lead selenide detectors have been used in advanced focal plane array (FPA) detectors and are commercially available as linear or bilinear FPAs with up to 526 detector elements. These elements are coupled to CMOS multiplexer readout circuits. The PbSe detector elements are thermoelectrically cooled, with entire power requirement for the FPAs and the cooler of less than 5 W. These structures are a low-cost alternative to the more elaborate FPAs based on HgCdTe technology.

The importance of thermal imaging by using focal plane arrays of many detector elements has recently stimulated further work with IV–VI epitaxial layers. In particular, the possibility of using Si substrates, as first shown by Zogg *et al.*, has brought a new impetus to the field of IV–VI infrared devices. With the use of Si wafers it is possible to integrate the detector of an infrared camera which consists of thousands of pixels to the read-out electronics which is usually based on standard silicon CMOS devices. In this case, the active lead salt infrared detector elements are grown by molecular beam epitaxy on CaF_2/BaF_2 buffer layers deposited on the Si (111) substrates after the fabrication of the electronic circuitry for the read-out. Conventional HgCdTe focal plane technology relies on hybrid structures with In bump bonds between the infrared devices and the Si substrate for each of the pixels. Due to the rather high dielectric constants of the IV–VI compounds as compared to HgCdTe, the high-frequency operation of IV–VI detector structures is limited to about 100 MHz. This upper frequency limit is, however, sufficient for infrared focal plane arrays. With IV–VI active epilayer thicknesses of few μm internal quantum efficiencies close to 100% and external efficiencies of about 60% have been achieved. Using materials like PbS, PbTe, PbEuSe, PbEuTe, or PbSnSe, the cutoff wavelength of the detector array can be varied from 4 to 14 μm at $T = 77$ K. The sensitivity of these detector structures is limited by generation-recombination noise for $T < 100$ K, whereas for higher temperatures these sensors are diffusion limited. The variation of quantum efficiencies between different pixels was shown to be less than about 3% for a 256 pixel device. Since all growth steps for these infrared sensors are performed at temperatures below 450 °C, a processing compatibility exists with standard Si-VLSI technology.

Summary

The peculiar structural, electronic, and optical properties of IV–VI semiconductors as compared to other semiconductor materials are a consequence of the ten valence electrons per atomic pair instead of the eight valence electrons typical for the tetrahedrally bonded group IV, III–V, and II–VI semiconductors. The most important group of IV–VI materials is the so-called lead chalcogenide (lead salt) compounds and their ternary and quaternary alloys. Their electronic band structure with the narrow direct minimum gap not at the center but at the edges of the Brillouin zone (L-points) exhibits a number of unique properties. The minimum energy gap increases with increasing temperature and decreases with hydrostatic pressure. The refractive indices are particularly large. As compared to narrow-gap III–V or II–VI semiconductors the Auger recombination rates are by two orders of magnitude smaller. All these properties provide the lead salts with great advantages for achieving infrared p-n junction lasing operation at comparatively high temperatures. Furthermore, efficient microcavity designs with extremely high reflectivities of 99% were realized with only three layer periods of combination like $PbSrTe/BaF_2$ or PbEuTe/EuTe due to their extremely high refractive index contrast. Further applications of the lead compounds are high sensitive infrared detectors, including focal plane array detectors which can be fabricated with comparatively low cost technologies, as well as thermoelectric devices.

List of Units and Nomenclature

Absorption *constant*	α
Absorption index	κ
Composition	x
Damping parameter	Γ_D
Detectivity	D^*
Dielectric function	ε
Energy	E
Energy gap	E_g
Free electron mass	m_0
High frequency dielectric constant	ε_∞
Kramers Kronig relations	KKR
Longitudinal effective mass	m_l
Longitudinal g-factor	g_l
Pressure	p
Refractive index	n
Static dielectric constant	ε_0
Temperature	T
Transverse effective mass	m_t
Transverse g-factor	g_t

See also

Nonlinear Optics, Basics: Kramers–Krönig Relations in Nonlinear Optics.

Further Reading

Bauer G (1980) Magnetooptical properties of IV–VI compounds. In: Zawadzki W (ed.) *Narrow Gap Semiconductors Physics and Applications*, pp. 470–494. Berlin: Springer Verlag.

Bauer G and Krenn H (1985) Lead selenide, lead telluride. In: Palik ED (ed.) *Handbook of Optical Constants*, pp. 517–523 and 535–546. New York: Academic Press.

Felix CL, Bewley WW, Vurgaftman I, *et al.* (2001) Low-threshold optically pumped $\lambda = 4.4$ μm vertical cavity surface emitting laser with a PbS quantum well active region. *Applied Physics Letters* 78: 3770–3772.

Holloway G (1980) Thin film IV–VI semiconductor photodiodes. In: Hass G and Francombe MF (eds) *Physics of Thin Films*, vol. 11, pp. 105–203. New York: Academic Press.

Jantsch W (1983) Dielectric properties and soft modes in semiconducting (PbSnGe)Te. *Springer Tracts in Modern Physics* 99: 1–50.

Katzir A, Rosman R, Shani Y, *et al.* (1989) Tuneable lead salt lasers. In: Cheo PK (ed.) *Handbook of Solid State Lasers*, pp. 227–347. New York: Marcel Dekker.

Khoklov D (2003) Lead chalcogenioles: Physics and applications. In: Manasreh MO (ed.) *Optoelectronic Properties of Semiconductors and Superlattices*, vol. 18. New York: Taylor and Francis.

Lucovsky G, Martin RM and Burstein E (1974) Structure, bonding and lattice dynamical properties of the IV–VI compounds. In: Rabii S (ed.) *Physics of IV–VI Compounds and Alloys*, pp. 93–102. New York: Gordon and Breach.

Nimtz G (1993) IV–VI compounds. In: Hellwege K and Madelung O (eds) *Numerical Data and Functional Relationships in Science and Technology.* Landolt-Börnstein, new series, group III, vol. 17, Part F, pp. 152–180. Berlin: Springer Verlag.

Partin DL and Heremans J (1994) Growth of narrow bandgap semiconductors. In: Moss TS and Mahajan S (eds) *Handbook on Semiconductors*, vol. 3A, pp. 369–450. Amsterdam: North-Holland.

Preier H (1990) Physics and applications of IV–VI compound semiconductor lasers. *Semiconductor Science and Technology* 5: S12–S20.

Ravich YI, Efimova BA and Smirnov IA (1970) Semiconducting Lead Chalcogenides. New York: Plenum.

Schwarzl T, Heiss W and Springholz G (1999) Ultra-high-finesse IV–VI microcavities for the midinfared. *Applied Physics Letters* 75: 1246–1248.

Springholz G, Shi Z and Zogg H (1998) Molecular beam epitaxy of IV–VI semiconductors. In: Liu AWK and Santos MB (eds) *Heteroepitaxial Thin Film Systems*, pp. 621–688. Singapore: World Scientific.

Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Mid-infrared surface-emitting PbSe/PbEuTe quantum dot-lasers. *Applied Physics Letters* 79–81: 1225.

Tacke M (1995) New developments and applications of tunable IR lead salt lasers. *Infrared Physics and Technology* 36: 447–463.

Yuan S, Springholz G, Bauer G and Kriechbaum M (1994) Electronic and optical properties of PBTe/$PB_{1-x}Eu_xTe$ multiple quantum well structures. *Physical Review* B49: 5476–5489.

Zemel JN, Jensen JD and Schoolar RB (1965) Electrical and optical properties of epitaxial films of PbS, PbSe, PbTe and SnTe. *Physical Review* 140: A330–A337.

Zogg H, Fach A, John J, *et al.* (1995) Photovoltaic IV–VI on Si infrared sensor arrrays for thermal imaging. *Optical Engineering* 34: 1964–1969.

Mercury Cadmium Telluride

M B Reine, BAE SYSTEMS, Lexington, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Mercury-cadmium telluride (HgCdTe or $Hg_{1-x}Cd_x$ Te) is a pseudobinary alloy semiconductor that crystallizes in the zincblende structure. The energy bandgap $E_G(x,T)$ varies continuously, and nearly linearly, with alloy composition parameter x, ranging from 1.6 eV for the wide-gap semiconductor CdTe to −0.3 eV for the semimetal HgTe. The bandgap passes through zero at an alloy composition of $x = 0.16$ (for $T = 0$ K). This broad range of energy bandgap provides a continuum of materials whose absorption edge can be tailored to provide high-performance quantum detectors for wavelengths over the 1–30 μm spectral range. Direct allowed valence-to-conduction band transitions are responsible for large values of the optical absorption coefficient, which enable quantum efficiencies approaching 100% to be achieved in practical device designs. Long minority carrier lifetimes result in low thermal noise, and permit high performance to be achieved at operating temperatures that are the highest reported among infrared detectors of comparable cutoff wavelengths. HgCdTe is well-established today as the most broadly applicable high-performance infrared detector material.

This article summarizes the fundamental properties of this versatile alloy semiconductor, and relates the material properties to its success as an infrared photoconductive and photovoltaic detector material.

Historical Background

It is noteworthy that HgCdTe is probably the first example of a semiconductor material that was intentionally created to satisfy well-defined theoretical device requirements. The importance of the infrared atmospheric transmission window at 8–12 μm was well known for thermal imaging systems, which enable night vision by imaging emitted infrared radiation from the scene rather than reflected visible radiation. The only available infrared detectors with the required sensitivity at 8–12 μm were extrinsic photoconductors such as copper-doped germanium, in which the photoinduced transitions were between a donor or acceptor level and the adjacent band extremum. However, such extrinsic detectors required cooling to prohibitively low temperatures, such as 10–15 K, which made thermal imaging systems bulky and impractical.

Theory showed that the reasons for such a low operating temperature were fundamental to the extrinsic detector material itself – essentially a short carrier lifetime that resulted in a large thermal generation rate and a correspondingly large thermal noise. Theory also showed that an intrinsic detector, in which the photoinduced transitions were direct (vertical) transitions between the valence band and the conduction band, could have significantly longer carrier lifetimes. This meant that an intrinsic detector, with a bandgap of 0.1 eV that matched the 8–12 μm band, could achieve the required sensitivity at much higher operating temperatures, as high as 77 K, that of liquid nitrogen. These theoretical and practical considerations motivated the search in the late 1950s for a semiconductor with a direct bandgap of 0.1 eV.

The first report of the synthesis of the semimetal HgTe and the wide-bandgap semiconductor CdTe to form the semiconductor alloy HgCdTe was published in 1959 by the Royal Radar Establishment in Malvern, England. This seminal paper reported both photoconductive and photovoltaic response at wavelengths extending to 12 μm, and made the understated observation that this material showed promise for intrinsic infrared detectors.

Soon thereafter, working under a US Air Force contract with the objective of devising an 8–12 μm background-limited semiconductor infrared detector that would operate at temperatures as high as 77 K, the group led by Paul Kruse at the Honeywell Corporate Research Center in Hopkins, Minnesota developed a modified Bridgman crystal growth technique for HgCdTe. In 1962 they reported both photoconductive and photovoltaic detection in rudimentary HgCdTe devices.

Fundamental Material Properties

Three fundamental properties determine the degree to which a semiconductor will make a successful infrared detector. First, the energy bandgap must be approximately equal to the smallest photon energy to be detected. There can be severe cooling penalties if the bandgap is significantly smaller than the smallest photon energy of interest. Second, the optical absorption coefficient must be large enough so that nearly all incident photons of interest are converted to electron–hole pairs for practical values of device thickness. Third, the inherent recombination mechanisms must allow long carrier lifetimes to be realized so that thermal generation rates, and hence thermal noise, can be minimized for practical devices and operating conditions.

The three key features that make HgCdTe the nearly ideal infrared detector material – tailorable energy band gap over the 1–30 μm range, large optical absorption coefficients that enable high quantum efficiency, and favorable inherent recombination mechanisms that lead to high operating temperature – are direct consequences of the energy band structure of this zincblende semiconductor. They apply no matter whether the device embodiment is a photoconductor, a photodiode or a metal–insulator–semiconductor detector. In addition, HgCdTe has many other favorable material properties, summarized in this article, that are technologically important for realizing practical infrared detectors. Tables 1 and 2 summarize various material properties for $Hg_{1-x}Cd_xTe$. **Table 1** summarizes material properties for $Hg_{1-x}Cd_xTe$ that vary with alloy composition x and temperature T: lattice constant a, energy band gap E_G, cutoff wavelength λ_{CO} (= 1.24/E_G), intrinsic carrier concentration n_i, conduction band effective mass ratio m_C/m_0, conduction band g-factor g_C, static and high frequency dielectric constants $\varepsilon_S/\varepsilon_0$ and $\varepsilon_\infty/\varepsilon_0$, index of refraction $n_R = (\varepsilon_\infty/\varepsilon_0)^{1/2}$, electron and heavy hole mobilities μ_e and μ_{hh}, mobility ratio b, radiative and Auger-1 lifetimes τ_{RAD} and $\tau_{Auger\text{-}1}$, and typical lifetimes $\tau_{typical}$ observed in n-type HgCdTe of low carrier concentration ($<1 \times 10^{15} cm^{-3}$).

Table 1 Material properties for selected compositions of $Hg_{1-x}Cd_xTe$, including the binary components HgTe and CdTe, and several technologically important alloy compositions. The radiative and Auger-1 lifetimes τ_{RAD} and $\tau_{Auger\text{-}1}$ are calculated for n-type HgCdTe with $N_D = 1 \times 10^{15}$ cm^{-3}

Property	*HgTe*	*$Hg_{1-x}Cd_xTe$*						*CdTe*
x	0	0.194	0.205	0.225	0.31	0.44	0.62	1.0
a (Å)	6.461	6.464	6.464	6.464	6.465	6.468	6.472	6.481
T (K)	77	77	77	77	140	200	250	300
E_G (eV)	−0.261	0.073	0.091	0.123	0.272	0.474	0.749	1.490
λ_{CO} (μm)		16.9	13.6	10.1	4.6	2.6	1.7	0.8
n_i (cm^{-3})		1.9×10^{14}	5.8×10^{13}	6.3×10^{12}	3.7×10^{12}	7.1×10^{11}	3.1×10^{10}	4.1×10^{5}
m_C/m_0		0.006	0.007	0.010	0.021	0.035	0.053	0.102
g_C		−150	−118	−84	−33	−15	−7	−1.2
$\varepsilon_S/\varepsilon_0$	20.0	18.2	18.1	17.9	17.1	15.9	14.2	10.6
$\varepsilon_\infty/\varepsilon_0$	14.4	12.8	12.7	12.5	11.9	10.8	9.3	6.2
n_R	3.79	3.58	3.57	3.54	3.44	3.29	3.06	2.50
μ_e (cm^2/V-s)		4.5×10^5	3.0×10^5	1.0×10^5				
μ_{hh} (cm^2/V-s)		450	450	450				
$b = \mu_e/\mu_{hh}$		1000	667	222				
τ_{RAD} (μs)		16.5	13.9	10.4	11.3	11.2	10.6	
$\tau_{Auger\text{-}1}$ (μs)		0.45	0.85	1.8	40	450	4.8×10^3	
$\tau_{typical}$ (μs)		0.4	0.8	1	7			

Table 2 Material properties for $Hg_{1-x}Cd_xTe$ that are independent of or relatively insensitive to alloy composition

Property	*Symbol*	*Value*
Kane matrix element	E_P	19 eV
Split-off band energy	Δ	0.93 eV
Heavy-hole effective mass ratio	m_{hh}/m_0	0.40–0.53
Valence band offset	E_v(HgTe) − E_v(CdTe)	0.35–0.55 eV

Energy Band Gap and Band Structure

The energy band gap $E_G(x, T)$ of $Hg_{1-x}Cd_xTe$ varies continuously, and nearly linearly, with alloy composition parameter x, ranging from 1.6 eV for the wide-gap semiconductor CdTe to −0.3 eV for the semimetal HgTe, as shown in Figure 1a. Also plotted in Figure 1a is the cutoff wavelength $\lambda_{CO}(x, T)$ ($\approx hc/E_G(x, T)$, where h is Planck's constant and c is the speed of light), defined as that wavelength at which the photoresponse falls to 50% of its peak value. The energy bandgap $E_G(x, T)$, defined at the difference between the Γ_6 and Γ_8 band extrema at $\Gamma = 0$, passes through zero at an alloy composition of $x = 0.16$ (for $T = 0$ K), as illustrated in Figure 1b. This broad range of energy bandgap provides a continuum of materials whose absorption edge can be tailored to provide high-performance quantum detectors for wavelengths over the 1–30 μm spectral range.

Optical Absorption Coefficient

Optical absorption coefficient data for several alloy compositions of $Hg_{1-x}Cd_xTe$ are plotted versus wavelength in Figure 2. Several trends are evident. The absorption strength generally decreases as the gap becomes smaller due both to the decrease in the conduction band effective mass and to the $\lambda^{-1/2}$ dependence of the absorption coefficient on wavelength λ. There is an exponential tail at energies just below the bandgap energy.

Carrier Lifetime Mechanisms

There are two fundamental mechanisms that determine the minority carrier lifetime in defect-free HgCdTe: radiative recombination and Auger recombination (Auger-1 in n-type and Auger-7 in p-type). Because they involve only interactions among electrons and holes in the valence and conduction bands, and do not require the intermediary of defect states, these mechanisms are regarded as inherent to the material itself, depending only on the band structure, the donor or acceptor concentration, and the temperature. The important advantage that HgCdTe enjoys as a high-performance detector material is that these fundamental mechanisms provide, for practically achievable

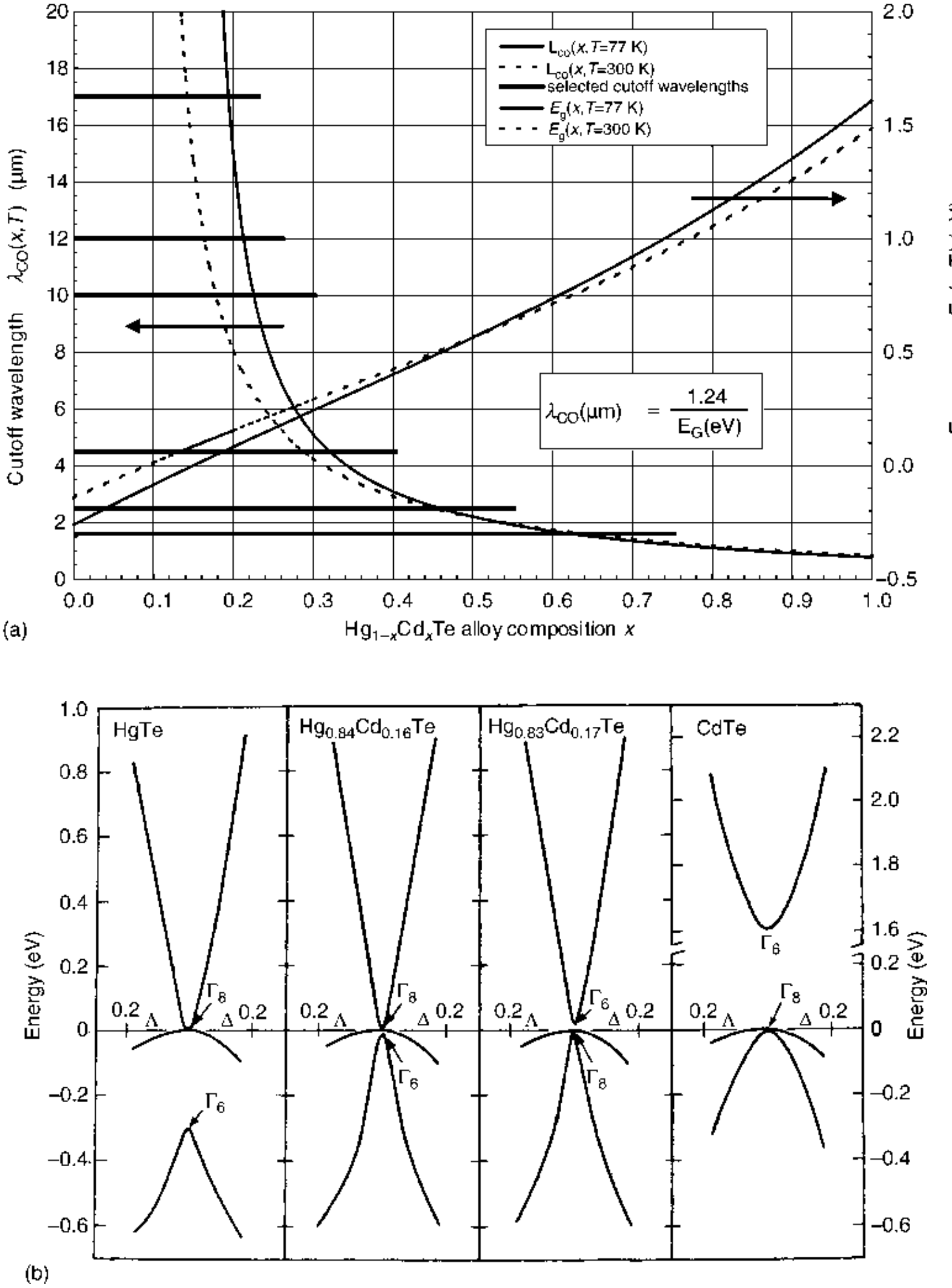

Figure 1 (a) Energy bandgap $E_G(x,T)$ and cutoff wavelength $\lambda_{CO}(x,T)$ for $Hg_{1-x}Cd_xTe$, plotted versus alloy composition parameter x, for temperatures of 77 K (solid curves) and 300 K (dashed curves). Horizontal bars indicate important cutoff wavelengths for HgCdTe detector technology. (b) Band diagrams near the Γ point calculated for HgTe, for two compositions of $Hg_{1-x}Cd_xTe$, and for CdTe, illustrating the transition from negative to positive energy bandgap. The energy bandgap is defined at the difference between the Γ_6 and Γ_8 band extrema at $\Gamma = 0$. Part (b) reproduced with permission from Chadi DJ and Cohen ML (1973) Electronic structure of $Hg_{1-x}Cd_xTe$ alloys and charge-density calculations using representative *K* points. *Physical Review B* 7: 692–699.

carrier concentrations, lifetimes that are long enough (and hence thermal generation rates that are low enough) that background-limited sensitivity is achieved at temperatures substantially higher than for other classes of infrared detectors, such as extrinsic detectors (As-doped Si, Cu-doped Ge) and quantum-well infrared photodetectors (QWIPs).

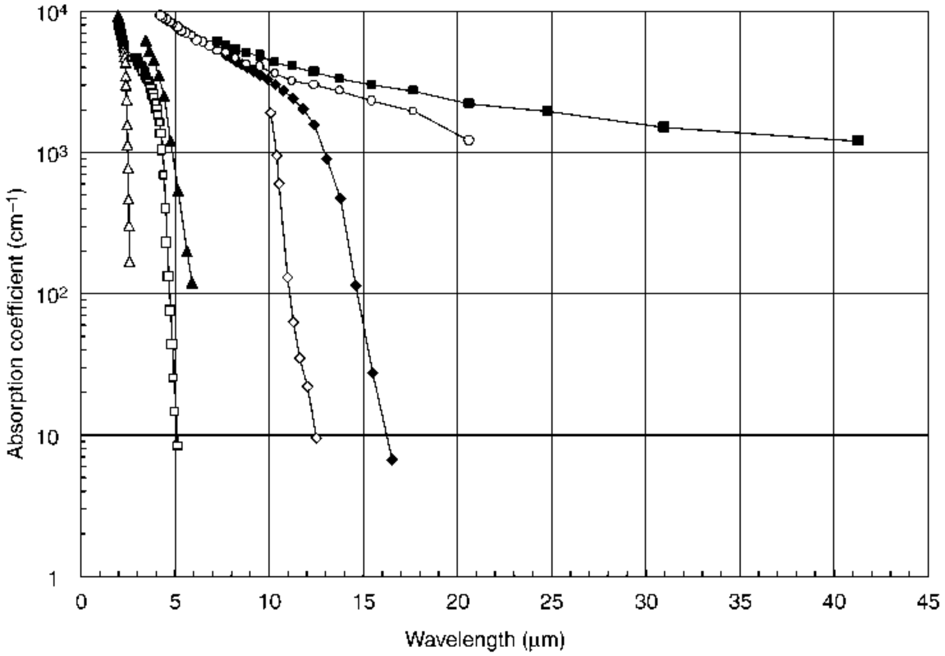

Figure 2 Optical absorption coefficient data for several $Hg_{1-x}Cd_xTe$ alloy compositions, for photon energies near the fundamental absorption edge, plotted versus wavelength.

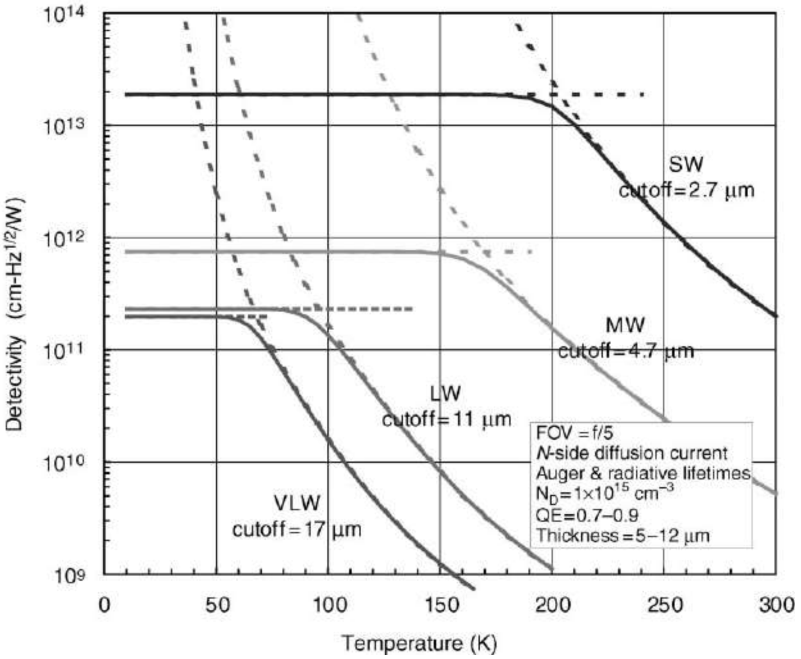

Figure 3 Calculated detectivity (D^*) for $Hg_{1-x}Cd_xTe$ infrared detectors for four technologically important wavelength regions, plotted versus operating temperature.

Figure 3 illustrates the sensitivities and operating temperatures that can be achieved for HgCdTe infrared detectors for four important wavelength regions. Detectivity, D^*, is the normalized signal-to-noise ratio that is the key figure of merit for the sensitivity of an infrared detector. At lower temperatures, detector thermal noise is negligible, and D^* is limited by detector noise due to fluctuations in the arrival rate of photons from the thermal room-background radiation. This limit is referred to as the BLIP (background-limited infrared photodetector) limit, and is indicated in Figure 3 by the horizontal dashed lines. As detector temperature increases, the detector thermal noise increases exponentially, and eventually overcomes the background noise, causing D^* to decrease exponentially for further increases in temperature. Detector thermal noise is proportional to the thermal generation rate, which is inversely proportional to the carrier lifetime.

n-Type and p-Type Doping

HgCdTe can be made n-type or p-type by a number of relatively convenient methods, at carrier concentrations required for high-performance n-type photoconductor and p-n junction photodiode architectures. Common donors are indium and iodine. Arsenic has become the most commonly used acceptor dopant, although copper and gold are used in some cases. Some HgCdTe photodiode designs still employ native metal-vacancy point defects as acceptors, although the clear trend is toward the use of extrinsic acceptor doping to avoid the strong Shockley–Read recombination associated with the Hg vacancy. Residual electrically active impurity concentrations are generally less than $1 \times 10^{14}\ \mathrm{cm}^{-3}$, which allows controllable doping at low-$10^{14}\ \mathrm{cm}^{-3}$ concentrations.

Crystal Growth Methods

The first growth methods for HgCdTe were bulk growth. Liquid phase epitaxy (LPE) began to be developed around 1975, followed by metalorganic chemical vapor deposition (MOCVD) and molecular beam epitaxy (MBE) in the early 1980s. Bulk growth, LPE, MOCVD, and MBE techniques have been refined through the years, and each is still in use today. Each technique has found application for certain types of HgCdTe devices. Bulk growth methods are used for production quantities of n-type wafers, with diameters of 12–20 mm, for LW and VLW photoconductive arrays, which are more tolerant of dislocations. LPE is in use today for production quantities of photovoltaic (PV) HgCdTe arrays, and has also been used for certain photoconductive HgCdTe arrays. Vapor-phase epitaxy (VPE) methods such as MBE and MOVPE are in use for engineering and prototype quantities. Both run-to-run reproducibility and control as well as intra-wafer uniformity of HgCdTe alloy composition are well in hand, meeting or exceeding requirements.

LPE-based one-layer and two-layer photodiode technology is well established today, offering the lowest defect densities and highest operabilities for production quantities of high performance FPAs. The VPE methods, MBE and MOVPE, will probably not entirely replace LPE in the near future, but will continue to improve, particularly with respect to lower defect densities, offering *in situ* growth of advanced bandgap-engineered detector designs that are unwieldy or impossible with LPE, such as dual-band detectors, avalanche photodiodes, and planar buried junctions with *in situ* CdTe passivation. MBE has shown the potential for rapidly switching from growth of one HgCdTe alloy composition to another in successive growth runs, without the necessity of preparatory calibration growth runs. This compositional agility, important for the rapid and cost-effective adjustment from product to product, will be further enhanced by the ongoing efforts to implement and improve *in situ* real-time monitors and feedback controls.

Substrates for Epitaxial Growth

The lattice mismatch between HgTe and CdTe is small, approximately 0.3%. This allows epitaxial growth of high-quality HgCdTe films on IR-transparent CdTe or nearly-lattice-matched IR-transparent $Cd_{1-z}Zn_zTe$ ($z \approx 0.04$) substrates, with dislocation densities in the mid-$10^4\ \mathrm{cm}^{-2}$ range. It also allows the *in situ* growth of various isotype and anisotype heterojunctions with tolerably low or negligible densities of misfit dislocations. This has led to an increasing number of bandgap-engineered HgCdTe photodiode structures, such as dual-band or two-color detector arrays and Auger-suppressed photodiodes with the potential for greatly increased operating temperature.

Sapphire and silicon are IR-transparent substrates that are less costly, that are available in much larger areas, and that are more rugged than CdTe and CdZnTe. Both sapphire and silicon can be used for epitaxial growth of HgCdTe films with dislocation densities that are acceptably low (mid-$10^6\ \mathrm{cm}^{-2}$) for many important photodiode applications, such as for the MW and SW spectral ranges, and for the LW spectral range for high background photon fluxes. LPE growth of HgCdTe($x = 0.4$) is done on 3-inch diameter sapphire substrates with MOCVD-grown CdTe buffer layers. MBE growth of HgCdTe($x = 0.3$) on 4-inch diameter (211) silicon substrates, with ZnTe/CdTe buffer layers, has recently been reported.

Dielectric Constant

The relatively low dielectric constant ($\varepsilon_S = 18\varepsilon_0$) of HgCdTe, in contrast to those of PbTe and PbSnTe, permits low junction capacitance. This is important for fast response in laser pulse detectors where small $R-C$ time constants are needed. It is also important for suppressing preamplifier noise below the detector noise, and thereby achieving high detectivity (D^*), in photodiodes operating at low temperature and low background fluxes, where the junction resistance is sufficiently high that the junction capacitance becomes an important contribution to the junction impedance at frequencies of interest.

Surface Passivation

Passivation of the exposed HgCdTe surfaces is critical for both device performance and long-term stability. The favorable surface properties of HgCdTe have enabled practical surface passivation schemes to be developed for both n-type photoconductors and p-n junction photodiodes of both planar and mesa configurations.

n-type HgCdTe photoconductors are commonly passivated with a native oxide that accumulates the surface, resulting in surface recombination velocities as low as $50\ \mathrm{cm\,s^{-1}}$, along with low $1/f$ noise (knee frequencies less than 50 Hz). The surface accumulation layer also provides a layer of surface electrons with high mobility, though lower than the bulk electron mobility, such that the shunt conductance of this layer is tolerably low compared to the bulk conductance of the detector itself. Several methods have been used for growing this native oxide, including plasma deposition and electrochemical (anodic) oxidation.

HgCdTe photovoltaic detectors require quite different passivation because usually both n-type and p-type surfaces are exposed as well as the depletion region. The CdTe/HgCdTe heterostructure is an important passivation for HgCdTe photovoltaic detectors and arrays. CdTe passivation has enabled arrays of HgCdTe photodiodes with negligible $1/f$ noise (knee frequencies less than 1 Hz), radiation hardness well in excess of the 20–40 kRad(Si) usually encountered in Earth-resource space missions, and excellent stability to thermal bake and to exposure to moisture.

CdTe has a number of important advantages as a passivant for HgCdTe junctions. It is a ‘native’ passivant that is chemically compatible with HgCdTe. Adhesion is excellent. The valence band edge of CdTe is approximately 0.4 eV below that of HgTe, thereby allowing repulsive barriers to form in the underlying HgCdTe for both electrons in the conduction band and holes in the valence band. Both mesa and planar photodiodes have been successfully passivated with CdTe. A wide variety of methods has been reported for deposition of CdTe passivating layers, including MBE, MOVPE, e-beam evaporation, sputtering, and hot wall epitaxy.

Electron-to-Hole Mobility Ratio

The small values for the conduction band effective mass ratio lead to large values for the electron mobility, as large as $4.5 \times 10^5\ \mathrm{cm^2/V\text{-}s}$ for $x = 0.195$ at 77 K. The heavy hole effective mass ratio is quite large, approximately 0.5 for all alloy compositions, leading to low values for the heavy hole mobility, on the order of $500\ \mathrm{cm^2/V\text{-}s}$. This gives large values for the electron-to-heavy hole mobility ratio, b, as high as 1000 for $x = 0.194$ at 77 K.

These large values of b were of great benefit to n-type HgCdTe photoconductors. Most such devices have small active areas, on the order of $50 \times 50\ \mu\mathrm{m}^2$, and the drift length for minority carriers (holes) becomes comparable to the interelectrode spacing at modest bias voltages. In this case, when the contacts have a high recombination velocity, the photoconductive gain saturates to the value $b/2$. High values of b in HgCdTe allowed sufficiently high photoconductive gains to be achieved to raise the detector noise above the preamp noise, an essential condition for achieving BLIP sensitivity.

HgCdTe Infrared Detector Configurations

Three different types of infrared detectors can be realized in the $Hg_{1-x}Cd_xTe$ alloy semiconductor: the photoconductive (PC) detector, the photodiode, also referred to as the photovoltaic (PV) detector, and the metal–insulator–semiconductor (MIS) detector. These are illustrated in Figure 4.

In all three types of HgCdTe detector – PC, PV, and MIS – excess electron–hole pairs are photogenerated by the same mechanism: valence-to-conduction band absorption of infrared radiation with photon energy greater than the bandgap energy E_G. All three types can achieve high quantum efficiencies, approaching 100% for well-designed devices. All three types have the same fundamental recombination mechanisms (Auger-1 and radiative) that determine the highest possible operating temperature to achieve a given sensitivity (signal-to-noise ratio).

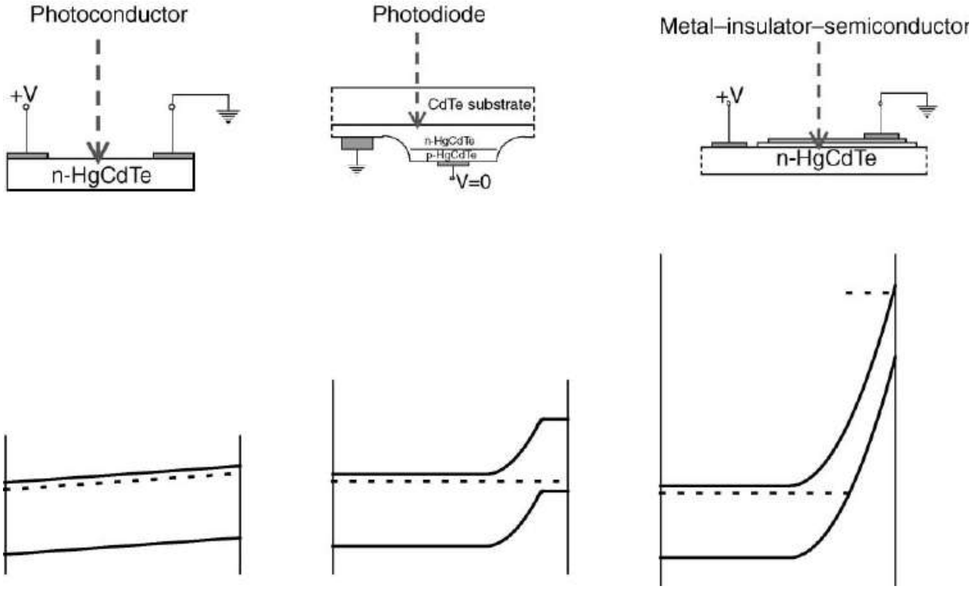

Figure 4 Three different types of infrared detectors have been realized in the $Hg_{1-x}Cd_xTe$ alloy semiconductor: the photoconductive (PC) detector, the photodiode or the photovoltaic (PV) detector, and the metal–insulator–semiconductor (MIS) detector. HgCdTe PC and PV detector arrays are being manufactured for a wide variety of applications. The MIS detector was abandoned in the mid-1980s because it placed unachievable demands on material quality.

The differences among PC, PV, and MIS HgCdTe detectors are due to the way in which the photogenerated electron–hole pairs are manifested electrically in the terminal characteristics of the device. In the PC detector, the excess electron–hole pairs are sensed as a small increase in the conductivity of what is basically a two-terminal resistor. In the PV detector, the excess electron–hole pairs are sensed as a photocurrent in a short-circuited p-n junction photodiode. In the MIS detector, the excess electron–hole pairs are sensed as a small change in the voltage due to minority photocarriers filling a transient charge storage well that has been biased to deep depletion.

PC HgCdTe Detectors

The HgCdTe PC detector is a low-resistance two-terminal device. Resistances are typically 25–100 ohms per square. A dc bias current is applied to convert the conductivity change to an observable voltage change. The device operates under near-equilibrium conditions, with small dc bias voltages typically on the order of 0.1 V. The dc bias electric field within the photoconductor is quite small, generally in the $20-50\,\mathrm{V\,cm^{-1}}$ range. The low-resistance device requires a low-noise bipolar preamplifier, usually external to the dewar. Each PC detector element requires one lead through the dewar wall, thus limiting the practical number of PC elements in an array to about 200–300.

The basic HgCdTe photoconductor is a rectangular area, photolithographically defined in an n-type HgCdTe layer approximately 8–15 μm thick, with two ohmic contacts on opposite edges. Surface passivation is straightforward: a thin native oxide strongly accumulates the surface of n-type HgCdTe, thereby reducing the surface recombination velocity to negligibly small values.

Many useful variations on the simple rectangular geometry are possible because of the favorable material properties of HgCdTe. Several contact geometries have been devised to minimize the recombination of photocarriers at the contacts. Serpentine designs have been used to increase detector resistance and reduce bias power dissipation, especially important issues for large-area and very long-wavelength detectors. The SPRITE (signal processing in the element) design reduces the number of electrical connections and dewar leads for scanned arrays by performing both detection and integration within an elongated HgCdTe bar.

PV HgCdTe Detectors

The HgCdTe photodiode is a more complicated device than the photoconductor, requiring both n-type and p-type layers, and having a surface depletion region that makes stringent demands on surface passivation technology.

The PV detector is a high-resistance p-n junction device. The photodiode is operated very close to equilibrium conditions, either at zero bias voltage or at a small reverse bias voltage whose magnitude is several *kT*, typically 20–30 mV. The absorber layer is usually n-type HgCdTe about 5–15 μm thick, with the thicker values required for longer-wavelength radiation. A p-n junction is formed by a thin p-type layer. A wide variety of junction formation methods are used. Wide-gap-p on narrow-gap-n heterojunctions are formed by two-layer LPE. Planar p-on-n or n-on-p junctions are formed by arsenic or boron implantation into n-type or p-type layers grown by LPE or MBE. Vertical-geometry p-on-n junctions are formed by a damage mechanism during ion beam milling. Both single-junction and multijunction devices are grown *in situ* by vapor phase epitaxial methods such as MBE and MOVPE. Surface passivation is generally accomplished by CdTe. The high-resistance PV device can match well with low-noise silicon CMOS preamplifiers that can be integrated with two-dimensional arrays to form large back-illuminated hybrid focal plane arrays (FPAs). Typical array formats for HgCdTe FPAs are 256 × 256 and 480 × 640, and can be as large as 1024 × 1024 and 2048 × 2048.

MIS HgCdTe Detectors

The metal–insulator–semiconductor (MIS) detector, also referred to as a photocapacitor, is usually formed on an n-type HgCdTe absorber layer. The insulator of choice is a thin native oxide. The gate electrode is a thin semitransparent metal film.

In contrast to the PC and PV detectors, the MIS detector operates in a strongly nonequilibrium mode.

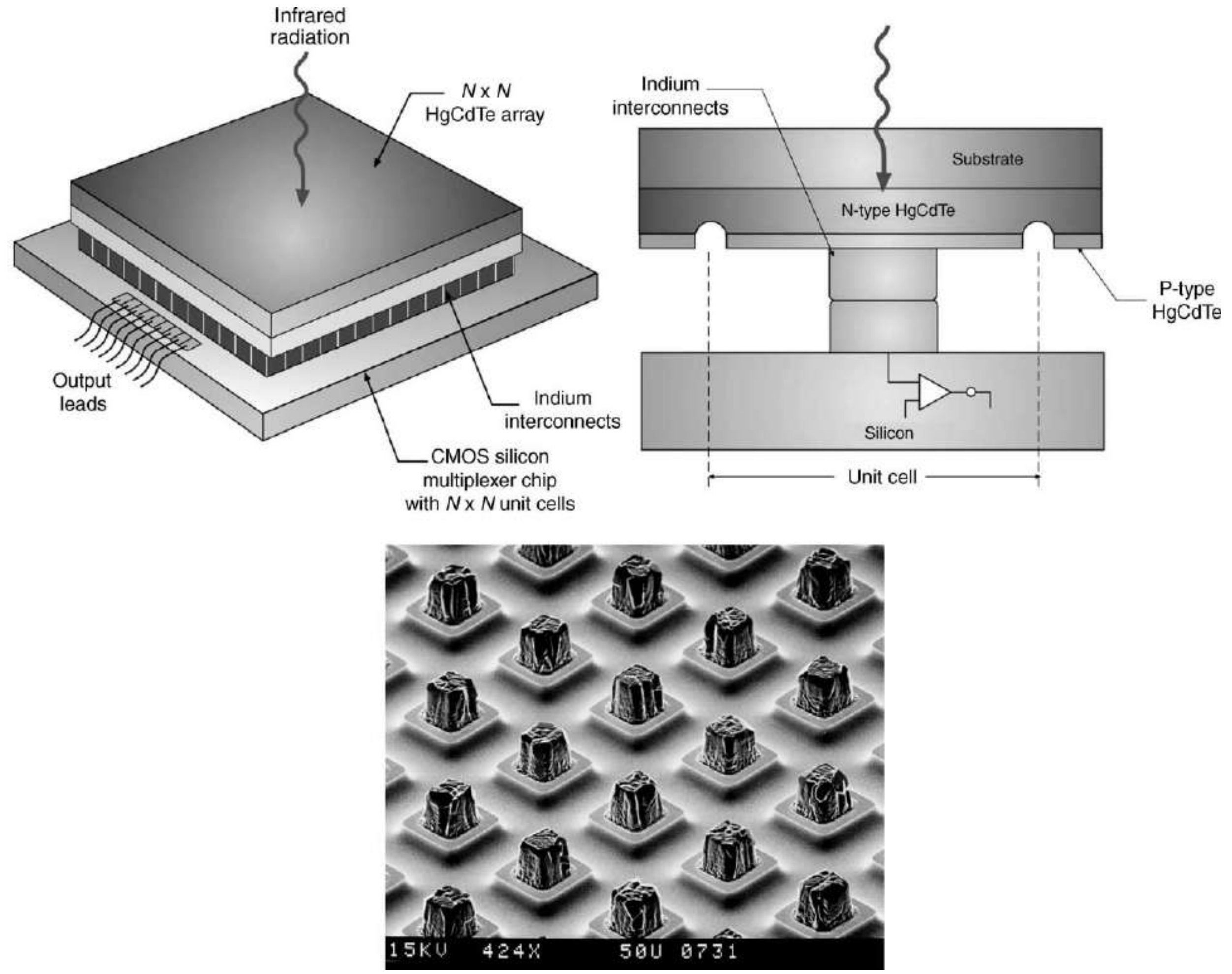
Infrared radiation
N x N HgCdTe array
Output leads
Indium interconnects
CMOS silicon multiplexer chip with N x N unit cells
Indium interconnects
Substrate
N-type HgCdTe
P-type HgCdTe
Silicon
Unit cell
15KV 424X 50U 0731

Figure 5a Back-illuminated bump-interconnected HgCdTe FPA architecture. The SEM photo shows an LPE P-on-n HgCdTe mesa photodiode array with 60×60 μm^2 unit cells and 33×33 μm^2 mesa (junction) areas.

A bias voltage pulse of several volts is applied across the capacitor to drive the HgCdTe surface into deep depletion. Excess electron–hole pairs are photogenerated in the absorber layer and diffuse to the edge of the depletion region, where the holes are collected in the charge well. After a period of time, called the integration time, the voltage across the capacitor is sensed, giving a measure of how much charge was collected, and the cycle is repeated.

Because of the nonequilibrium operation of the MIS detector, much larger electric fields are set up in the depletion region than in the p-n junction, resulting in defect-related tunneling dark current that is orders of magnitude larger than the fundamental dark current. The MIS detector requires much higher material quality than PV or PC detectors, which still has not been achieved. For this reason, all development of the HgCdTe MIS detector was abandoned in the mid-1980s.

Two-Dimensional HgCdTe Infrared Focal Plane Arrays

Among the most important applications of HgCdTe are large two-dimensional electronically scanned hybrid arrays, referred to as focal plane arrays (FPAs). A hybrid HgCdTe FPA consists of a two-dimensional HgCdTe photovoltaic detector array that is interfaced electrically, thermally and mechanically with a matching two-dimensional array of input circuits in a silicon CMOS ROIC chip. Each HgCdTe detector element in the array has one electrical connection to its matching input circuit node in the silicon ROIC.

The thermal expansion coefficient of HgCdTe is sufficiently close to that of silicon to allow several technologically viable hybrid arrangements of HgCdTe detector arrays and silicon multiplexer chips.

There are two main types of hybrid HgCdTe FPAs being developed and manufactured today: the back-illuminated bump-interconnected configuration, shown in Figure 5a, and the front-illuminated or 'loophole' configuration shown in Figure 5b. Each configuration successfully solves the problem of the thermal expansion mismatch between silicon and HgCdTe in different ways, and each requires fundamentally different HgCdTe photodiode designs and processing.

In the back-illuminated configuration, illustrated in Figure 5a, the HgCdTe detector array is bump-mounted onto the silicon ROIC by cold-welded or thermally reflowed indium interconnects. Incident radiation reaches the detectors through an infrared-transparent substrate such as CdTe, CdZnTe, sapphire, or silicon.

In the front-illuminated (or 'loophole') FPA architecture, shown in Figure 5b, the thin HgCdTe layer is epoxied to the silicon ROIC chip, and undergoes elastic deformation when the FPA is cooled. This allows large arrays to be made without need for engineering the thermal expansion of the silicon ROIC chip. Low-temperature processing techniques

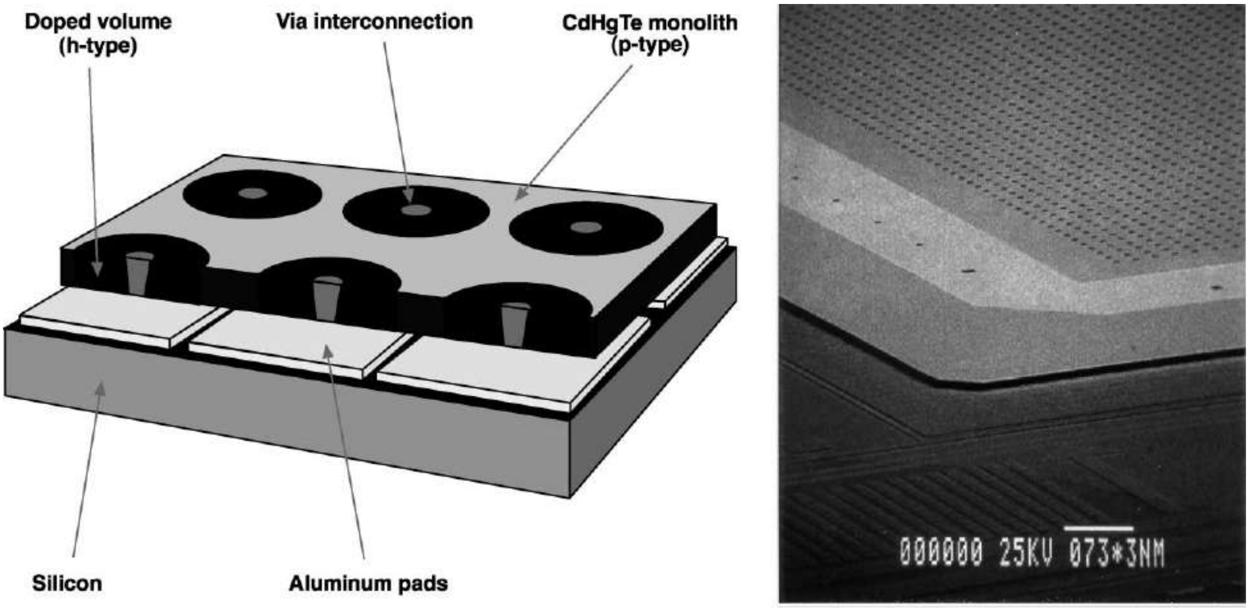

Figure 5b Front-illuminated loophole HgCdTe FPA architecture. The SEM photo shows one corner of a FPA. The metalized via holes are on 20 μm centers.

for junction formation and passivation are required because of the presence of the epoxy.

Conclusions and Trends

There has been enormous progress in HgCdTe material science and device technology over the past 45 years. This progress has been fueled by the many military and space applications for which HgCdTe PC and PV infrared detectors provide nearly ideal solutions. It is important to recognize that the progress so far in HgCdTe detector technology has been possible only through advances in basic materials growth and processing technology.

Further progress can be expected. The fundamental performance limits of HgCdTe photodiodes have not yet been reached. Continued reduction of material defects, both grown-in and process-induced, will reduce thermal generation rates, increase junction impedance, and allow higher operating temperatures for a number of important applications. Continued development of the *in situ* VPE growth methods – MBE and MOVPE – will allow bandgap-engineered heterojunction devices of increasing quality and complexity. Continued development of VPE growth on alternative substrates such as silicon will reduce the cost of two-dimensional arrays, and will make larger arrays practical.

List of Units and Nomenclature

BLIP	Background limited infrared photodetector
CdTe	Cadmium telluride
CdZnTe	Cadmium zinc telluride
CMOS	Complementary metal-oxide-semiconductor
FPA	Focal plane array
HgCdTe	Mercury cadmium telluride
$Hg_{1-x}Cd_xTe$	Mercury cadmium telluride
IR	Infrared
LPE	Liquid phase epitaxy
LW	Long wavelength
LWIR	Long wavelength infrared
MBE	Molecular beam epitaxy
MCT	Mercury cadmium telluride
MIS	Metal-insulator-semiconductor
MOCVD	Metal-organic chemical vapor epitaxy
MOVPE	Metal-organic vapor phase epitaxy
MW	Medium wavelength
MWIR	Medium wavelength infrared
PC	Photoconductive
PV	Photovoltaic
QWIP	Quantum well infrared photodetector
ROIC	Readout integrated circuit
SPRITE	Signal processing in the element
SW	Short wavelength
SWIR	Short wavelength infrared
VLW	Very long wavelength
VLWIR	Very long wavelength infrared
VPE	Vapor phase epitaxy

See also

Semiconductor Physics: Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties.

Further Reading

Capper P (ed.) (1994) *Properties of Narrow Gap Cadmium-based Compounds*. EMIS Datareviews Series No. 10. London: INSPEC, Institution of Electrical Engineers. Reviews and compilations of fundamental and technological properties of HgCdTe and CdTe/CdZnTe/CdTeSe.

Capper P (ed.) (1997) *Narrow-Gap II–VI Compounds for Optoelectronic and Electromagnetic Applications*. London: Chapman and Hall. Twenty review chapters on growth techniques, materials characterization, and device applications of HgCdTe.

Capper P and Elliott CT (eds) (2000) *Infrared Detectors and Emitters: Materials and Devices*. Boston, MA: Kluwer Academic Publishers. Review chapters on HgCdTe materials, photoconductive and nonequilibrium HgCdTe detectors, photovoltaic HgCdTe detectors, and HgTe-CdTe superlattice detectors.

Rogalski A, Adamiec K and Rutkowski J (2000) *Narrow Gap Semiconductor Photodiodes*. Washington, DC: SPIE Press.

Willardson RK and Beer AC (eds) (1981) *Defects, (Hg,Cd)Se, (Hg,Cd)Te*. Semiconductors and Semimetals, vol. 16. New York: Academic Press. Review chapters on magneto-optical properties of HgCdTe and nonlinear optical effects in HgCdTe.

Willardson RK and Beer AC (eds) (1981) *Mercury Cadmium Telluride*. Semiconductors and Semimetals, vol. 18. New York: Academic Press. Seven review chapters: history of HgCdTe, preparation of high-purity Hg, Cd and Te, crystal growth of HgCdTe, Auger recombination in HgCdTe, and photoconductive, photovoltaic and metal-insulator-semiconductor HgCdTe detectors.

Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures

Y-S Huang, National Taiwan University of Science and Engineering, Taipei, Taiwan

F H Pollak, Brooklyn College of the City University of New York, Brooklyn, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modulation spectroscopy is a versatile and powerful optical technique for obtaining valuable information relevant to a large variety of semiconductor systems including bulk/thin films, surfaces/interfaces [semiconductor/air (vacuum), semiconductor/electrolyte, semiconductor/metal, semiconductor/semiconductor (homo- and heterojunctions)], micro- and nanostructures (quantum wells, multiple quantum wells, superlattices, quantum wires, quantum dots), the effects of growth/processing, as well as the characterization of actual device structures [heterojunction bipolar transistors (HBTs), pseudomorphic high electron mobility transistors (PHEMTs), edge-emitting and vertical-cavity surface-emitting (VCSEL) quantum well lasers, etc.].

Modulation spectroscopy deals with the measurement and interpretation of changes in the optical response of a sample which are effected by the modification of the measurement conditions. This can easily be accomplished by periodically modulating either (1) the measurement conditions themselves ('internal' modulation) or (2) some parameters applied to the sample ('external' modulation). The observed normalized changes are usually small so that the difference signals are closely related to a derivative of the absolute spectrum with respect to the modifying parameters. The derivative nature of modulation spectroscopy emphasizes structure localized in the photon energy region of the interband (intersub-band) transitions of semiconductors (semiconductor microstructures) and suppresses uninteresting background effects. As a result, weak features that may not be detected in the absolute spectra are often enhanced. Because of this derivative-like nature a large number of sharp spectral features may be observed, even at room temperature.

The ability to perform a lineshape fit is one of the great advantages of modulation spectroscopy. Since for the modulated signal the features are localized in photon energy it is possible to account for the lineshapes to yield accurate values of energies and broadening parameters of the associated interband transitions. For example, the energies of various interband (intersub-band) transitions may be evaluated to within a few meV, even at room temperature. Thus, the effects of static external perturbations such as electric and magnetic fields, temperature, hydrostatic pressure, uniaxial stress, etc., can be conveniently studied.

A particularly useful form of modulation spectroscopy is electromodulation (EM) since it is sensitive to surface/interface electric fields and can be performed in contactless modes that require no special mounting of the sample. Under appropriate conditions the EM spectrum may exhibit above bandgap oscillatory features, called Franz–Keldysh oscillations (FKOs), which are a direct measure of the relevant electric field (built-in and/or applied).

Two of the most widely used forms of EM are photoreflectance (PR) and contactless electroreflectance (CER). In PR, modulation of the built-in electric field is caused by photo-excited electron–hole pairs created by a pump source of ~3–5 mW power of either (1) an internally modulated laser diode or (2) a mechanically chopped dc laser. The modulating frequency is typically ~100–200 Hz. CER utilizes a condenser-like system consisting of a front wire grid electrode with a second metal electrode separated from the first electrode by insulating spacers, which are ~0.1 mm larger than the sample thickness. The sample is placed between these two capacitor plates. Thus, there is nothing in direct contact with the front surface of the sample. The probe beam is incident through the front wire grid. EM is achieved by apply an ac voltage (~1 kV peak to peak at ~200 Hz) across the electrodes.

Instrumentation

Shown in Figure 1 is a schematic drawing of the experimental arrangement for a general 'external' modulated reflectance experiment (except for PR). Light from an appropriate light source (xenon arc or tungsten halogen lamp) passes through a monochromator. The exit intensity at wavelength λ, $I_o(\lambda)$, is focused onto the sample by means of a lens (or mirror). The modulation (electric field, stress, temperature) is applied to the sample at frequency Ω_m. The reflected light is collected by a second lens (mirror) and is focused onto an appropriate detector

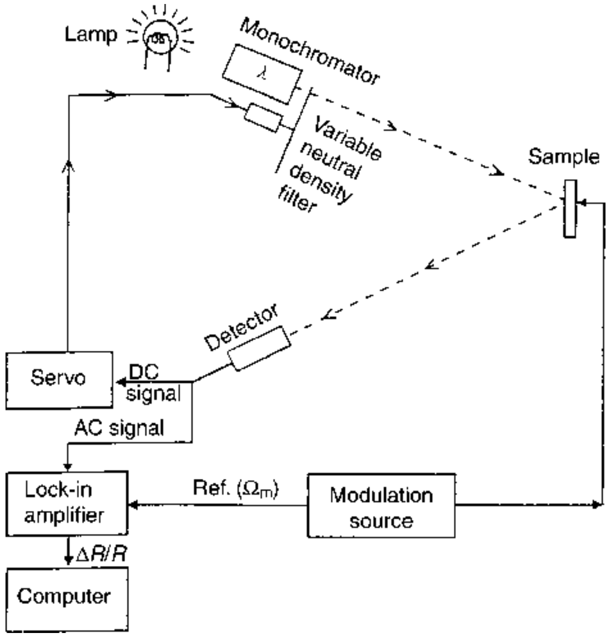

Figure 1 Schematic representation of external modulation experimental apparatus.

(photomultiplier or photodiode). For the sake of simplicity the two lenses (mirrors) are not shown.

The light striking the detector contains two signals: the dc is given by $I_o(\lambda)R(\lambda)$, where $R(\lambda)$ is the dc reflectance of the material while the modulated value (at frequency Ω_m) is $I_o(\lambda)\Delta R(\lambda)$, where $\Delta R(\lambda)$ is the change in reflectance produced by the modulation source. The ac signal from the detector, proportional to $I_o(\lambda)\Delta R(\lambda)$, is measured by a lock-in amplifier. Typically $I_o\Delta R$ is 10^{-4}–10^{-6} of I_oR. In order to evaluate the quantity of interest, i.e., the relative change in reflectance $\Delta R/R$, a normalization procedure must be used to eliminate the uninteresting common feature $I_o(\lambda)$. This can be done in several ways, one of which is shown in Figure 1, where the normalization is performed by a variable neutral density filter (VNDF) connected to a servo mechanism. The dc signal from the detector, which is proportional to $I_o(\lambda)R(\lambda)$, is fed into the servo which moves the VNDF in such a manner as to maintain a constant $I_o(\lambda)R(\lambda)$, i.e., $I_o(\lambda)R(\lambda) = C$. Under these conditions the ac signal $I_o(\lambda)\Delta R(\lambda) = C\Delta R(\lambda)/R(\lambda)$. Subsequently, the signal to the lock-in amplifier is proportional to the quantity of interest, i.e., $\Delta R(\lambda)/R(\lambda)$.

Bulk/Thin-Film Material

Modulation spectroscopy has been used to study the properties of bulk/epitaxial layered semiconductors such as ordering in $GaInP_2$, piezoelectric fields in strained GaN, the influence of isotopes, the properties of low-temperature grown GaAs, the space charge region, carrier concentration, the nature of the band bending (carrier type), impurity effects, deep levels, alloy composition, the influence of static external or internal perturbations such as temperature, external uniaxial stress, lattice-mismatch strain, hydrostatic pressure, the effects of crystal growth/processing/annealing procedures, and amorphous and microcrystalline materials.

The dashed lines in Figure 2 are the piezoreflectance (PZR) spectra in the region of the direct gap of ReS_2 in the temperature range $25\,K < T < 450\,K$. The data at low temperatures exhibit two well-resolved features labeled E_1^{ex} and E_2^{ex}. The solid lines are lineshape fits to the first derivative of a Lorentzian profile which yields the energies indicated by arrows and also the related broadening parameters. The solid/open squares and circles in Figure 3 are the temperature dependence of E_1^{ex}/E_2^{ex} of ReS_2 and $ReSe_2$, respectively. The solid, dot-dashed, and dashed lines in the figure are fits to the Varshni, O'Donnel–Chen, and Bose–Einstein expressions, respectively. The Varshni equation is:

$$E_i^{ex}(T) = \frac{E_i^{ex}(0) - \alpha_i T^2}{\beta_i + T} \qquad [1]$$

where $i = 1$ or 2, $E_i^{ex}(0)$ is the excitonic transition energy at 0 K and α, β are the Varshni coefficients. The O'Donnel–Chen expression is:

$$E_i^{ex}(T) = E_{i0} - S_i\langle\hbar\Omega_i\rangle\left(\frac{\coth\langle\hbar\Omega_i\rangle}{2kT} - 1\right) \qquad [2]$$

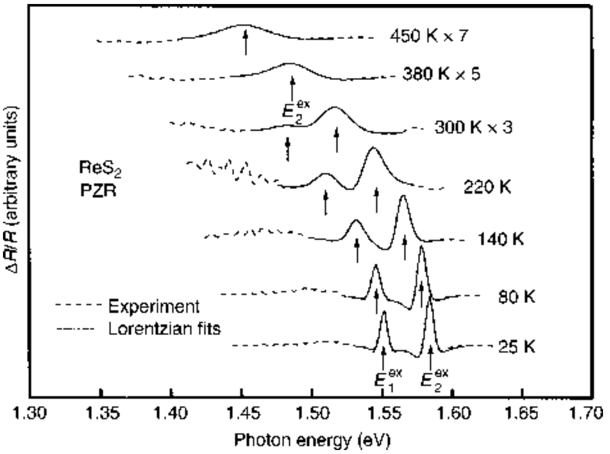

Figure 2 The PZR spectra (dashed lines) in the region of the direct gap of ReS2 in the temperature range $25\,K < T < 450\,K$. The solid lines are lineshape fits yielding the energies indicated by the arrows.

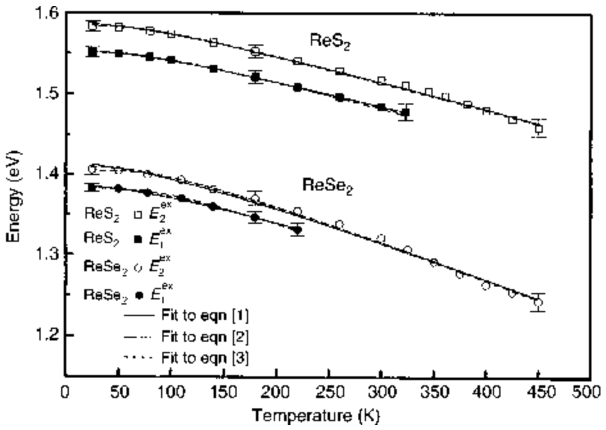

Figure 3 The temperature dependence of the excitonic transition energies of ReS_2 and $ReSe_2$. The solid, dot-dashed, and dashed lines in the figure are fits to the Varshni, O'Donnel–Chen, and Bose–Einstein expressions, respectively.

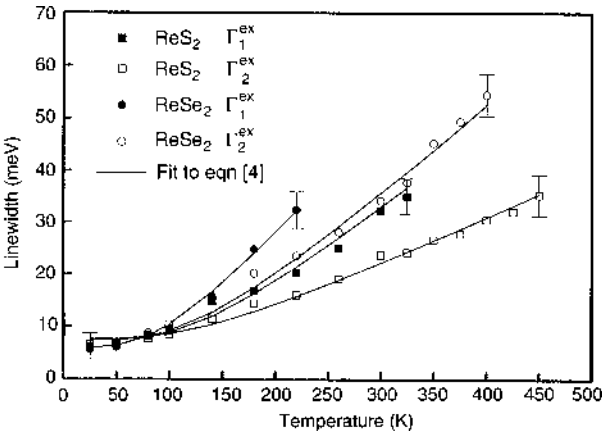

Figure 4 The solid/open squares and circles are the temperature dependence of the broadening parameters of E_1^{ex}/E_2^{ex} of ReS_2 and $ReSe_2$, respectively. The solid lines are fits to the Bose–Einstein-type expression.

where E_{i0} is the excitonic transition energy at 0 K, S_i is a dimensionless coupling constant related to the strength of the electron–phonon interaction, and $\langle \hbar\Omega_i \rangle$ is the average phonon energy. The Bose–Einstein-type expression can be written as

$$E_i^{ex}(T) = E_{iB} - a_{iB}\left[1 + \frac{2}{\exp(E_{ap}/kT) - 1}\right] \quad [3]$$

where a_{iB} represents the strength of the electron (exciton)–phonon interaction, and E_{ap} corresponds to the average phonon energy.

The temperature dependence of the broadening parameters, $\Gamma(T)$, is due only to the electron–phonon interaction and can also be described by a Bose–Einstein-type equation. For the lowest lying direct gap only the longitudinal optic (LO) phonon is involved. In Figure 4 the solid/open squares and circles are the temperature dependence of the broadening parameters of E_1^{ex}/E_2^{ex} of ReS_2 and $ReSe_2$, respectively. The solid lines are fits to the Bose–Einstein-type expression:

$$\Gamma_i(T) = \frac{\Gamma_i(0) + \Gamma_{i,ep}}{\exp(E_{i,LO}/kT) - 1} \quad [4]$$

where $i = 1$ or 2. The first term of eqn [4] corresponds to broadening mechanisms due to intrinsic lifetime, electron–electron interaction, impurity, dislocation, and alloy scattering effects. The parameter Γ_{ep} is an electron (exciton)–phonon coupling coefficient and E_{LO} is the zone-center LO phonon energy. The obtained values of Γ_{ep} for these materials have been compared with those for the direct gap in GaAs and ZnSe.

Surfaces/Interfaces

Almost since the inception of modulation spectroscopy it has been recognized that EM could be used as an effective probe of surfaces and interfaces. For EM this is true not only because of the sharp, derivative-like spectral features but also due to its sensitivity to electric fields, i.e., FKOs. In addition, PR and CER are of considerable interest and usefulness since they are contactless, employ low light levels, require no special mounting of the sample, and can be performed in any transparent ambient including ultrahigh vacuum (UHV).

Of particular usefulness in the study of Fermi level pinning at surfaces are UN(SIN)/UP(SIP) and δ-doping configurations. The UN(SIN)/UP(SIP) structures are made by fabricating an undoped layer of thickness L (~1000 Å) on a buried doped $n^+(p^+)$ buffer on a doped $n^+(p^+)$ substrate. In the $n^+(p^+)$ buffer/substrate the Fermi level occurs near the conduction (valence) band edge. At the surface the Fermi level is pinned at some value, V_F. Therefore, there exists in the undoped region, which has a small broadening parameter, a large, almost constant electric field, F. A similar constant field can be created by placing a δ-doping layer a distance L from the surface.

The relation between F, as observed from the FKOs, and V_F is given by:

$$FL = V_F - (kT/q) - SCC - V_p(T) \quad [5]$$

where the second, third, and fourth terms on the right-hand side are the Debye length, space charge layer, and temperature dependent photovoltaic effect corrections, respectively.

The properties of the metal–semiconductor interface have been reported by a number of investigators using PR and/or ER. Jonker *et al.* have demonstrated that an epitaxial Fe film on GaAs (UN structure) provides a ferromagnetic contact, suppresses midgap state formation, and does not pin the Fermi level. Four samples were investigated: a native oxide reference sample, the same sample after sulfur passivation, Al/GaAs-(2 × 4), and Fe/GaAs-(2 × 4). The width of the undoped GaAs layer was 1070 Å for the first three samples and 1500 Å for the last sample. These results are shown in Figure 5a–d, respectively. The fields obtained from the FKOs are 58 kV/cm, 61 kV/cm, 55 kV/cm and 60 kV/cm. For the first three samples the measured barrier height is about 0.64 eV (not corrected for the photovoltaic effect) while for the Fe/GaAs material it is about 40% higher (~0.9 eV). Another significant feature of the Fe/GaAs spectrum is the large number of FKOs (in relation to the other samples) extending to about 2 eV. The significant reduction in damping indicates a substantial increase in carrier lifetimes.

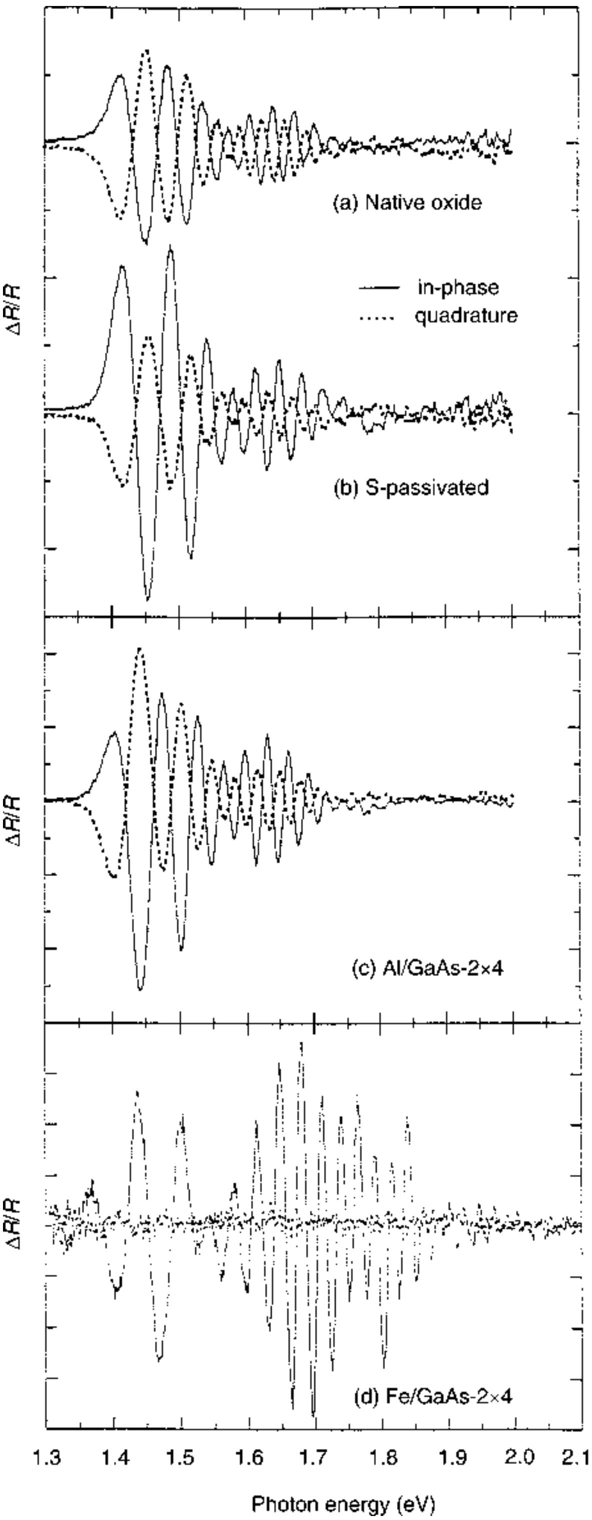

Figure 5 The PR spectra at 300 K for (a) a native oxide reference sample, $F = 58$ kV/cm; (b) the same sample following sulfur passivation, $F = 61$ kV/cm; (c) Al/GaAs-(4 × 2) sample, $F = 55$ kV/cm, and (d) Fe/GaAs-(2 × 4) sample, $F = 60$ kV/cm. The thickness of the undoped spacer region is 1070 Å for (a)–(c) and 1500 Å for (d).

Micro- and Nanostructures

Modulation spectroscopy is a very powerful tool for investigating many of the fundamental aspects of compositional single quantum wells (SQWs)/multiple quantum wells (MQWs), superlattices (SLs), quantum wires, and quantum dots (QDs). Various phenomena can be studied, including band offset, well and barrier widths, excitons, strain, coupling and decoupling between wells, miniband formation, two-dimensional electron gas (2DEG) effects, zone-folding in short-period SLs, built-in electric fields, etc. Work on a large variety of GeSi, III–V and II–VI systems has been reported, including both lattice-matched and strain-layer configurations. In addition the effects of various external perturbations can be evaluated.

The properties of a vertically coupled InAs/GaAs QD-based laser structure fabricated by a self-assembled technique have been investigated by CER (300 K and 20 K). Signals have been observed from all the relevant regions of the sample. The dotted lines are the CER spectra at 300 K and 20 K in Figure 6a,b, respectively. The solid lines are lineshape fits which yield the energies indicated by arrows. The low-energy features, labeled QD_0 (lateral coupling), QD_1 (vertical coupling) and QD_2 (vertical coupling), originate from the QDs while the resonances designated W_1 and W_2, correspond to the first conduction to first heavy- (1C-1H) and light-hole (1C-1L) transitions, respectively, in the QW (approximately one monolayer) formed by the

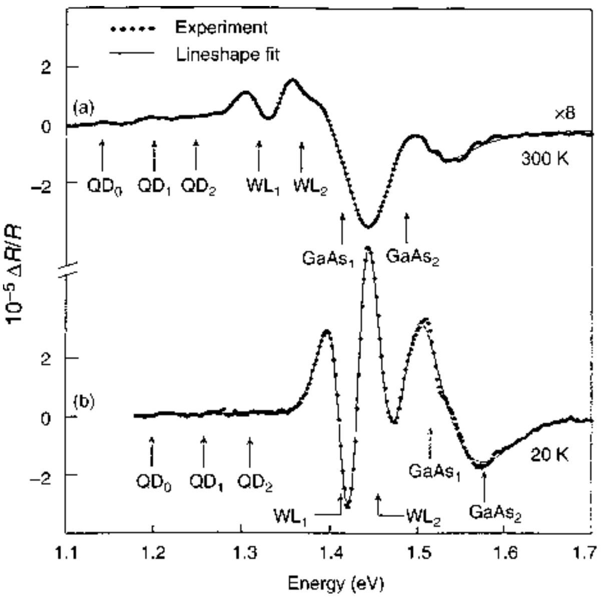

Figure 6 CER spectra (dotted lines) of an InAs/GaAs quantum dot-based laser structure at (a) 300 K and (b) 20 K. The solid lines are a lineshape fit to a first derivative Gaussian.

wetting layer (WL). Although the oscillator strength of the QD features is small compared to the WL resonances, they are clearly visible even at 300 K. The $GaAs_1$ and $GaAs_2$ features are attributed to GaAs-related transitions. The former at both 300 K and 20 K corresponds to the bandgap energy of GaAs and probably comes from the undoped 400 Å-thick GaAs layers which envelop the QD region. The $GaAs_2$ feature is most likely related to the p-doped 6000 Å-thick GaAs cap layer. The heavy doping of this layer induces a Burstein–Moss blue shift of the bandgap.

Device Structures

The methods of PR and CER are valuable tools in the evaluation of important device parameters for structures such as heterojunction bipolar transistors (HBTs), pseudomorphic high electron mobility transistors (PHEMTs), edge-emitting and vertical cavity surface-emitting (VCSEL) quantum well lasers, multiple quantum well infrared detectors, solar cells, superlattice optical mirrors, resonant tunneling structures, metal-oxide-semiconductor (MOS) configurations, among others.

The dashed lines in Figure 7a,b are the 300 K CER spectra of a $GaInP_2$(emitter)/GaAs(collector) HBT (fabricated by metalorganic chemical vapor deposition) for incident light with electric field vector $E \| [110]$ (polarization A) and $E \| [1\bar{1}0]$ (polarization B), respectively. Both collector and emitter signals exhibit well-pronounced FKOs. The solid lines are least-squares fits to a function which contains Lorentzian broadened electro-optic functions, yielding 1.873 ± 0.001 eV and 1.885 ± 0.001 eV for the $GaInP_2$ bandgaps for polarizations A and B, respectively, as indicated by the arrows. The ordering parameter of 0.3 deduced from the differences in the measured bandgaps for the two polarizations was consistent with transmission electron microscope measurements. The electric fields evaluated from the emitter and collector region FKOs were compared with a computer simulation of the field profiles using a comprehensive, self-consistent model, including the photovoltaic effect. The deduced emitter and collector doping densities were in good agreement with the intended growth conditions.

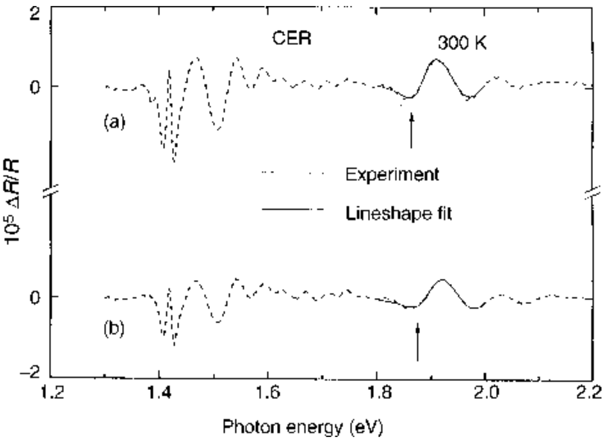

Figure 7 Room-temperature CER spectra (dashed lines) for a GaInP/GaAs HBT sample fabricated by metalorganic chemical vapor deposition for incident light polarized (a) $E \| [110]$ and (b) $E \| [1\bar{1}0]$. The solid lines are a lineshape fit which yields the bandgap energies denoted by arrows.

Summary

Modulation spectroscopy is a very powerful tool for investigating many of the fundamental and applied aspects of a wide variety of semiconductors and semiconductor micro- and nanostructures, including bulk/thin films, surfaces/interfaces, single quantum wells, multiple quantum wells, superlattices, quantum wires, quantum dots, etc. In addition the effects of various perturbations such as alloy composition, temperature, electric and magnetic fields, hydrostatic pressure, internal and/or external stain, processing, etc., can be evaluated. Recently these methods have also been applied for the characterization of actual device structures such as HBTs, PHEMTs, quantum well lasers, etc.

See also

Semiconductor Materials: Quantum Dots. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures; Surface Photovoltage Spectroscopy of Semiconductors.

Further Reading

Aspnes DE (1980) Modulation spectroscopy/electric field effects on the dielectric properties of semiconductors. In: Balkanski M (ed.) *Handbook on Semiconductors*, vol. 2, pp. 109–154. Amsterdam, The Netherlands: North-Holland.

Glembocki OJ and Shanabrook BV (1992) Photoreflectance spectroscopy of microstructures. In: Seiler DG and Littler CL (eds) *Semiconductors and Semimetals*, vol. 36, pp. 222–289. New York: Academic Press.

Pollak FH (1994) Modulation spectroscopy of semiconductors and semiconductor microstructures. In: Balkanski M (ed.) *Handbook on Semiconductors*, vol. 2, pp. 527–635. Amsterdam, The Netherlands: North Holland.

Pollak FH (1995) Electromodulation spectroscopy of semiconductor surfaces and interfaces. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 175–209. Amsterdam, The Netherlands: North-Holland.

Pollak FH (1996) Modulation spectroscopy of quantum wells and superlattices. In: Bhattacharya P (ed.) *Properties of III–V Quantum Wells and Superlattices*: *EMIS Data Reviews Series #15*, pp. 232–240. London: INSPEC.

Pollak FH (1998) Modulation spectroscopy of the group III-nitrides. In: Gil B (ed.) *Group III Nitride Semiconductor Compounds*, pp. 158–181. Oxford, UK: Clarendon.

Pollak FH (2001) Modulation spectroscopy characterization of semiconductors and semiconductor microstructures. In: Buschow KH, Cahn W, Flemmings MC, *et al.* (eds) *Encyclopedia of Materials: Science and Technology*, vol. 6, pp. 5743–5748. Amsterdam, The Netherlands: Elsevier.

Ramdas AK and Rodriguez S (1992) Piezospectroscopy of semiconductors. In: Seiler DG and Littler CL (eds) *Semiconductors and Semimetals*, vol. 36, pp. 137–220. New York: Academic Press.

Quantum Dots

R J Warburton, Heriot-Watt University,
Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Quantum dots are nanometer-sized clusters of semiconductor material. An island of low-bandgap material is embedded in a sea of high-bandgap material so that both electrons and holes are spatially confined in all three directions. In an unstructured semiconductor, the electron energy levels of the individual atoms hybridize to form energy bands. In a quantum dot, however, there are discrete energy levels. This can be understood from the band structure of the unstructured material where there is a so-called dispersion relation connecting the energy, E, and the momentum, k. The confinement potential of a quantum dot quantizes the k-values, giving rise to discrete energy levels. If the energy levels of a quantum dot are separated by more than the typical energy level broadenings and by more than the thermal energy then the quantum dot will behave quite differently to the bulk material. The quantum dot will exhibit atom-like properties – discrete energy levels with small degeneracies – despite the fact that it is embedded in a semiconductor matrix. It is this ability to prepare a variety of atom-like systems in a semiconductor environment which has driven the research into semiconductor quantum dots.

In some sense, nanosized quantum dots have been the subject of research for quite some time. Color filters, for instance, consist of nanosized particles dispersed in a glass matrix, and this is a very mature technology. Also, lithography of layered semiconductor systems already possessing confinement in one dimension has been pursued for at least 15 years. However, developments in the past few years have been rapid, all based on the discovery of self-assembly of quantum dots directly in the growth of semiconductor heterostructures. These quantum dots can be produced with standard materials in standard growth systems, and this allows an easy integration of quantum dots into devices such as diode lasers. Furthermore, the self-assembled quantum dots are of extremely high quality and this is of great value not just for devices but also for fundamental physics experiments, aimed at studying or exploiting the coherence of excitons confined in the quantum dots. For these reasons, this short review concentrates on the properties of self-assembled quantum dots, giving just brief descriptions of other systems. It should be noted that the bibliography is extremely limited and in no way does justice to the large number of research groups which have made valuable contributions to this field. Instead, the bibliography is simply a list of suggestions for finding out more about this fascinating research area.

Quantization in Quantum Dots

Quantum dots are on the one hand large enough that the bulk band structure of the host material can be used to describe the material's electronic properties. In the effective mass approximation, this amounts to giving both electrons and holes a quadratic dispersion on momentum, just as for a free electron, but with an effective mass, typically about 0.1 for the electrons, and 0.3 for the holes. On the other hand, quantum dots are small enough that confinement effects are important. For instance, the energy of an exciton in a nanometer-sized quantum dot is a strong function of the quantum dot size. As the dots become smaller, confinement effects become larger, and the energy separating electron and hole levels increases. This does not increase indefinitely, however. There is a finite-size energy barrier to the matrix surrounding the dot such that in the limit of small dot sizes, the wavefunction spreads out into the barrier region and confinement effects become very weak. Ideally, an electron in a quantum dot has a confinement energy of at least several meV, and an energy barrier to the surrounding matrix of several 100 meV so that confinement effects are visible even at room temperature.

The motivation for the study of quantum dots is to combine the atom-like properties with the flexibility of a solid-state environment. This is perfectly illustrated by considering a quantum dot as the gain medium for a laser. In a conventional semiconductor diode, a band of levels has to be inverted in order to achieve population inversion. This is costly in terms of the threshold current and the temperature stability. In a quantum dot, inversion can be achieved simply by inverting the population of the two-fold degenerate ground state, which is clearly trivial. Furthermore, a strong confinement energy and a large barrier to the surrounding matrix imply that this population inversion is stable against rises in temperature, implying better room-temperature performance. This example also exemplifies the motivation for investigating quantum dots using well-developed inorganic semiconductor technology as in this case the dots can be integrated into established device geometries. Generally speaking, the important concept in quantum dot physics is spatial confinement of electrons in all three directions, giving rise to discrete energy levels.

Lithography for Quantum Dots

Developments in semiconductor growth technology, largely during the 1970s and 1980s, enabled thin layers of different semiconductors to be grown on top of each other. There are two principal growth techniques, molecular beam epitaxy (MBE), in which a heated substrate is bombarded with pure species in a high vacuum, and metal organic chemical vapor deposition (MOCVD), in which chemical precursors flow over a heated substrate and react. In the simplest structure, termed a quantum well, a thin layer of low-bandgap material is sandwiched between high-bandgap material. The discontinuities in the potentials at the interface give rise to confined electronic states for both carrier types, electrons and holes. The quantum well is an example of a two-dimensional system: there is confinement in one direction, the growth direction, but free dispersion in the other two directions. To make a quantum dot, it is necessary to confine the electron and hole in all three directions. An obvious approach is to take a quantum well, with perfect confinement in one direction, and process it laterally with lithographic techniques to complete the quantization. Note that an analogy can be made to the physics of the quantum Hall effect, where a two-dimensional electron gas is subjected to a very high magnetic field. The magnetic field quantizes the lateral motion in a similar way to the electrostatic potential in a quantum dot.

This approach to the formation of optically active quantum dots obviously has some advantages, notably that the shape and position of the quantum dots can be chosen in the lithography. However, there are some significant problems associated with these techniques. As most of these problems do not plague the properties of self-assembled quantum dots, self-assembled quantum dots are more suitable for advanced photonic devices.

A number of different lithographic techniques has been used to process quantum-well heterostructures into quantum dots. For research applications, electron-beam lithography is the method of choice. The sample is coated with a resist, for instance PMMA, and exposed with electrons in an area with nanometer size. A protective layer, either a metal or insulator, is transferred to the exposed area by developing the resist, by depositing the protective layer material, and by removing the unwanted material with a lift-off step. The wafer is then etched, typically in a dry environment, to produce a free-standing column.

The photoluminescence of these quantum dots has been measured by various research groups, and although the results vary quite widely there are some systematic trends. The photoluminescence efficiency decreases rapidly for pillars smaller than a few microns. This is somewhat improved at low temperature, where reasonably strong emission can be observed down to 100-nm diameter pillars. For pillars with micron diameters, the lateral confinement effects are weak, and it is debatable if true quantum dots have been achieved. There are a number of effects which degrade the emission. First, pinning of the Fermi energy by surface states generates large electric fields near the surface, which can ionize electron–hole pairs. This is well known for quantum wells: if a GaAs quantum well for example is placed less than about 25 nm from the surface, the photoluminescence is very weak because of the surface electric field. In an etched pillar, these effects are clearly magnified by the new surface area generated. Secondly, the etching process damages the material to a greater or lesser degree, and the damaged region contains centers for nonradiative recombination.

Other lithography techniques have also been used to generate semiconductor quantum dots for optical experiments. An elegant idea is to deposit small regions of a highly strained film onto the sample surface. In this way, a laterally varying strain film in an underlying quantum well can be set up, generating quantum dots through the effect of strain on the band gap. (This idea has also been married with self-assembly, by using Stranski–Krastanow-grown InP stressors on a GaAs surface, generating lateral confinement in an underlying InGaAs quantum well.) A further idea is to generate a quantum dot by deliberately growing a quantum well with rough interfaces. In such a system, there will inevitably be regions which confine an electron–hole pair through localization. Both of these approaches clearly circumvent the problems generated by the exposed surface in an etched quantum dot. However, they generate small lateral confinement potentials, typically a few meV. This is of value in a research environment where fundamental experiments can be carried out at low temperature, but in a possible device at room temperature, where the thermal energy is 25 meV, quantization effects will not be discernible.

Self-Assembled Quantum Dots

A solution to the difficulties inherent in quantum dots made by lithographic techniques began to emerge in the early 1990s. It was discovered that quantum dots can be produced directly in the semiconductor growth by depositing a material onto a substrate with a substantially smaller lattice constant. The most well-known example is the growth of InAs (lattice constant 6.05 Å) on GaAs (lattice constant 5.65 Å) where In-rich quantum dots form. Because billions of dots form spontaneously in this process, the dots are referred to as self-assembled quantum dots. Since the early experiments, considerable progress has been made, such that the state-of-the-art quantum dots possess many of the hoped-for properties: they are coherent (meaning that they contain no dislocations), highly homogeneous (meaning that fluctuations in shape and composition from dot to dot are small), and have large quantum efficiencies for the emission of light.

Growth of Self-Assembled Quantum Dots

When growing one material on top of another, three principal growth mechanisms have been identified: Frank–van der Merwe, Volmer–Weber, and Stranski–Krastanov. In the Frank–van der Merwe mode, material is deposited layer by layer, as in the growth of the lattice-matched pair GaAs and AlAs. In the Volmer–Weber mode, island formation occurs: the material does not wet the surface because it is energetically unfavorable. Another possibility arises if there is a lattice mismatch. In this case, the epilayer is strained, and the strain energy obviously increases with increasing epilayer thickness. In the Stranski–Krastanow mode, the initial growth is layer by layer, as in Frank–van der Merwe, but beyond a certain thickness, islands form, as shown schematically in Figure 1. On the one hand, the islands limit the strain energy, because some strain relaxation is allowed, which is not possible in a thin film. On the other hand, island formation results in an increase of the

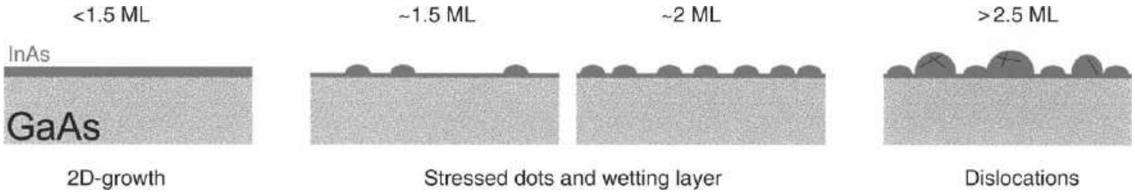

Figure 1 A schematic of the Stranski–Krastanow mode. The material deposited has a substantially larger lattice constant than the substrate, for example InAs on GaAs. Initially, a thin and uniform layer is deposited. At the critical thickness (1.5 monolayers for InAs on GaAs), islands spontaneously form. Further growth results in an increase in the number of islands. At large coverages, large and dislocated islands appear.

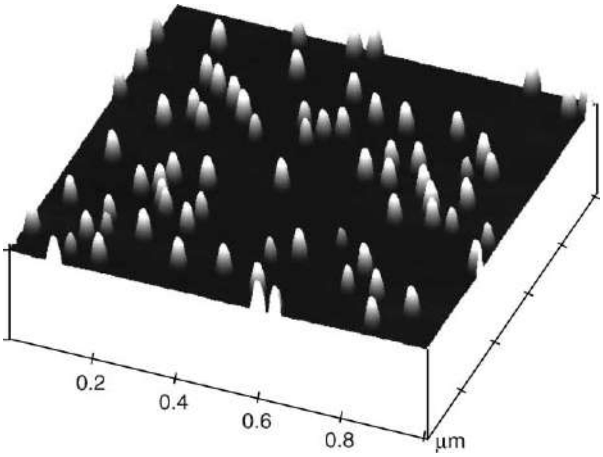

Figure 2 An atomic force micrograph of InAs quantum dots on a GaAs substrate. The InAs dots were grown by MBE at a growth temperature of 530 °C. The image was taken by Axel Lorke.

surface energy. This means that there is a size at which the total energy is minimized.

It is now well established that many combinations of semiconductors grow in the Stranski–Krastanow mode, and this can be utilized for self-assembly of quantum dots. The most studied example is InAs on GaAs where high-quality quantum dots can be grown with both MBE and MOCVD. Dots form at the so-called critical thickness of 1.5 monolayers, corresponding to just 4 Å of material. For thicknesses less than the critical thickness, the covering is not completely uniform; there are island-like structures, elongated along the $[0\bar{1}1]$ direction. At the critical thickness dots form, leaving a thin InAs layer (the wetting layer) between the dots. As more InAs is added, the density of dots increases with only small changes in the dot size. As an example, Figure 2 shows InAs dots on a GaAs substrate.

The dots grown in this Stranski–Krastanow mode are free of dislocations, which is an essential prerequisite in III–V compounds for a high quantum efficiency for optical emission. Furthermore, the dots are remarkably homogeneous. Statistical fluctuations during growth will always give rise to a distribution in dot size and height. Nevertheless, the fluctuations in dot size can be as small as ~10%. The dots are randomly arranged in the lateral plane if the growth proceeds on a flat substrate, and densities are between 10^9 and 10^{11} cm^{-2}. The shape and composition of the dots depend on the growth parameters. For instance, InAs/GaAs dots can have facets along particular crystal directions, or they can be lens-shaped without facets, depending on the growth technique, and in particular the growth temperature. In fact, measuring the shape and composition is a very challenging task, particularly for buried dots.

Stranski–Krastanow growth has turned out to be a robust phenomenon for III–V materials; quantum dots have been produced in this way for a number of material combinations. The most important condition is that the deposited material has a substantially larger lattice constant than the substrate. InAs forms dots not just on GaAs but also on InP. The InAs/InP dots are particularly interesting because they emit at the technologically important 1.5 μm wavelength. InP can also form the dot material by using GaInP substrates where the bandgap is pushed up into the red region of the spectrum. Dots can also be grown with II–VI materials; the growth of CdSe on ZnSe is analogous to the growth of InAs on GaAs. These II–VI dots emit in the green. Finally, dots can also be formed in the nitrides, for instance by depositing InGaN on an AlGaN surface. In fact, the high emission efficiency of GaN, despite the large defect density, has been attributed to the formation of quantum dot-like structures in the active layers.

It is a challenge to arrange the dots in the plane of the substrate. An easy route with self-assembly has not been found to accomplish this. Nevertheless, dots can be encouraged to grow at particular positions by pre-patterning the substrate. This obviously limits the separation of the dots in any such array to the limits set by lithography. However, self-assembly does provide a means to order the dots in the growth direction: if the separation between successive layers is small enough (less than 15 nm for InAs/GaAs) then the dots grow directly above one another, as shown in Figure 3. In a new layer, the residual strain field from the dots in the layer below is sufficient to seed the dots. In this way, vertical stacks of dots can be built up. In fact, it has been argued that the dots tend to become more laterally ordered with each layer. The point is that if two dots are close together in one layer, their strain fields overlap, and provide only one seeding center in the subsequent layer.

Optical Properties

The main interest in quantum dot physics is the quantization in all three spatial directions, giving rise to atomic-like energy levels. For a system of self-assembled quantum dots, it is important to measure the fundamental bandgap, the energetic separation between the confined states both for electrons and for holes, the barriers to the electron continuum and hole continuum, and the oscillator strengths for the various interband transitions. All this information can be gleaned with optical spectroscopy.

Perhaps the simplest experiment to perform, if not to understand, is photoluminescence in which electron–hole pairs are excited with a laser beam.

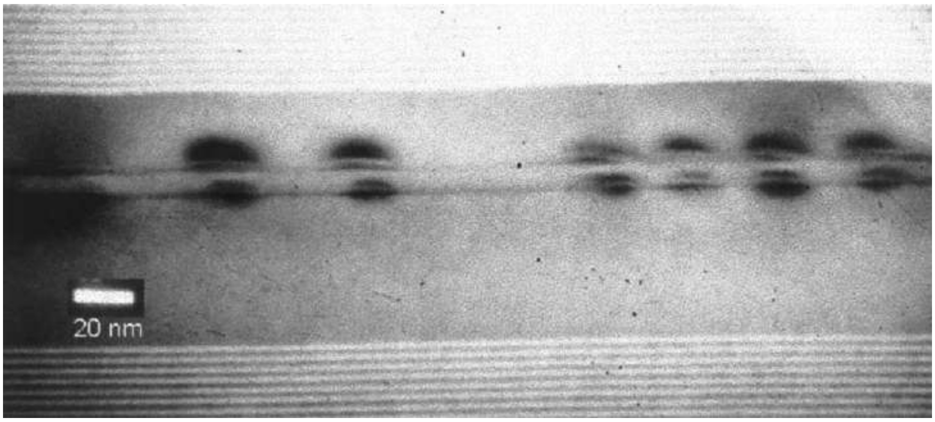

Figure 3 A transmission electron micrograph of two layers of InAs quantum dots in GaAs showing how the dots in the upper layer lie directly above the dots in the first. The image was taken by Gio Medeiros-Ribeiro.

The laser energy can be tuned to lie just above the bandgap of the wetting layer in which case the electron–hole pairs are excited in the wetting layer itself. Alternatively, if the laser energy is larger than the bandgap of the barrier material, electron–hole pairs are excited in the barrier. In both cases, the carriers relax into the dots, and then down the ladder of states in the dot, so that the dot electron and hole ground states are occupied. Emission is detected when an electron and a hole recombine. As such, the experiment measures only the energy of the ground-state exciton. However, if the pump intensity is increased such that electron–hole pairs are generated faster than they can recombine, the excited states of the dots are also occupied and recombination can also occur from excited states. An example of such an experiment is shown in Figure 4. At low pump powers, only the ground state emission is detected; at higher pump powers, the first excited state emerges, and at higher powers still, the second excited state also emerges, and so on. The spectra are broadened by the inhomogeneous broadening, i.e., from fluctuations from dot to dot, but the broadening is small enough that the excited state emission can be easily distinguished from the ground state emission. For these particular quantum dots, the energetic separation between the ground and excited state emission is 15 meV, and this corresponds to the sum of the electron and hole confinement energies. The separation between the other emissions is also about 15 meV, implying that the states are approximately equally spaced, implying in turn that the confining potential is approximately parabolic. Furthermore, in this experiment, emission from the wetting layer can also be made out, and this enables the electron and hole confining potentials to be estimated.

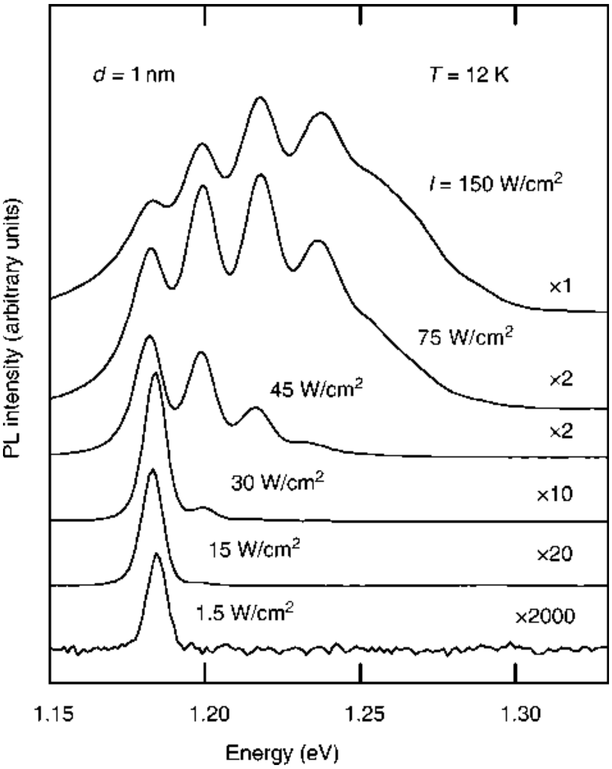

Figure 4 Photoluminescence from an ensemble of quantum dots as a function of laser excitation power. The measurement temperature was 12 K. The quantum dots are induced in an InGaAs/GaAs quantum well by a strain field generated by InP stressors on the sample surface. The stressors are grown in the Stranski–Krastanow mode, and are 1 nm away from the quantum well. At low excitation power, only the ground state emission is observed. As the power increases, emission from excited states is also observed through state filling. Reprinted with permission from Lipsanen H, Sopanen M and Ahopelto J (1995) Luminescence from excited states in strain-induced $In_xGa_{1-x}As$ quantum dots. *Physical Review* 51: 13868. Copyright 1995 American Physical Society.

Photoluminescence invariably yields the sum of electron and hole energies. In order to measure just the electronic properties, it is possible to occupy the dots with, say, two electrons (so that in each dot the two-fold degenerate ground state is fully occupied) and to excite electric dipole transitions between the electron levels. This has been accomplished with so-called charge-tunable heterostructures, with which electrons can be controllably loaded into the dots from an electron reservoir. The interlevel absorption lies in the far-infrared region of the spectrum; for InAs/GaAs quantum dots for instance, the transition is at a wavelength of 25 μm (an energy of 50 meV), as shown in Figure 5. The results of this experiment also point to an approximately parabolic confining potential from the fact that as electrons are added to the dots, the energy of the far-infrared absorption does not change much. (In the limit of a perfect parabolic potential, a famous result known as Kohn's theorem states that the long-wavelength absorption should show no energy dependence on the electron occupation.)

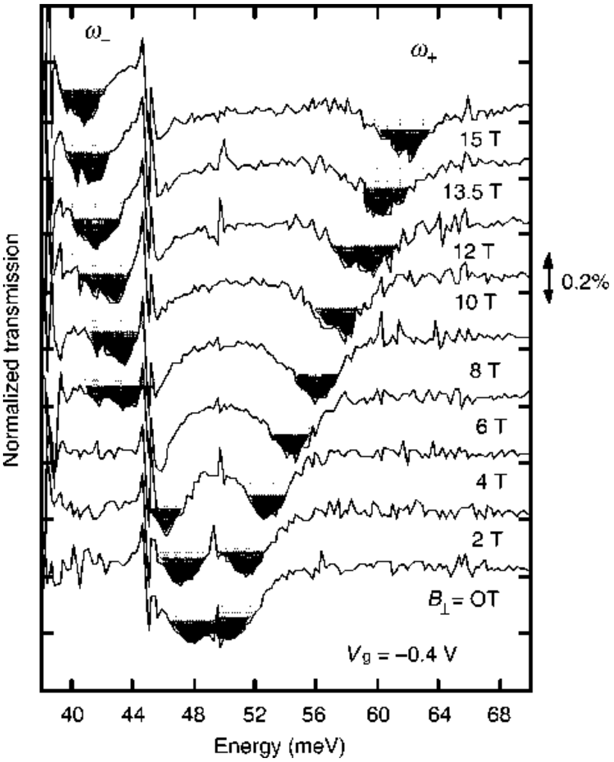

Figure 5 Transmission in the far infrared of an ensemble of InAs/GaAs quantum dots where each quantum dot is occupied with two electrons. The data were taken at a temperature of 4.2 K. The shaded transmission minima correspond to absorption from the quantum dots. The transitions are between the electron ground state and the first excited state. Two peaks are observed even at zero magnetic field, suggesting that the dots are slightly oval in shape. In a magnetic field, the two peaks move apart, which is a consequence of the Zeeman effect. The strong feature at 45 meV arises from an electronic interaction with an interface phonon. Reprinted with permission from Fricke M, Lorke A, Kotthaus JP, Medeiros-Ribeiro G and Petroff PM (1996) Shell structure and electron-electron interaction in self-assembled InAs quantum dots. *Europhysics Letters* 36: 197. Copyright 1996 European Physical Society.

In self-assembled quantum dots, the confining potential can be thought of as very steep in the growth direction, and relatively shallow in the lateral plane. In other words, the dots are essentially two-dimensional disks. The electron ground state is, in analogy with atomic physics, s-like, and the excited state p-like. However, because of the disk-like nature of the dots, the p-state is four-fold degenerate, and not six-fold degenerate as in a conventional atom. One way to demonstrate this is to perform interband absorption experiments on charge-tunable structures, as shown in Figure 6. The first transition corresponds to the transition from the hole s-state to the electron s-state, and disappears when the dots are occupied with two electrons. This is simply because the Pauli principle forbids the transition once the final level is fully occupied. The second transition corresponds to the transition from the hole p-state to the electron p-state, and disappears when the dots are occupied with four electrons, confirming the degeneracies of

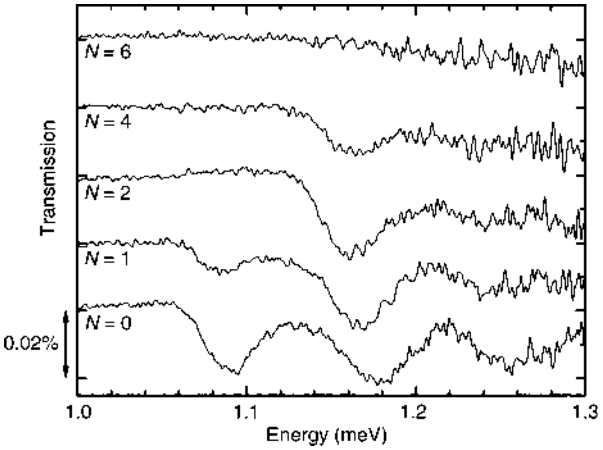

Figure 6 Transmission in the near infrared of an ensemble of InAs/GaAs quantum dots at 4.2 K. Spectra are shown for different values of N, the electron occupation. At $N = 0$, there are three transitions corresponding to transitions from a valence state to an electron state. These are the s–s, p–p and d–d transitions, where the states are labeled in analogy to atomic physics. At $N = 2$, the electron ground state is fully occupied, and the s–s transition disappears because it is blocked, a consequence of the Pauli principle. Similarly, the p–p transition is blocked at $N = 6$ when both the electronic s and p states are fully occupied. Reprinted with permission from Warburton RJ, Dürr CS, Karrai K *et al.* (1997) Charged excitons in self-assembled semiconductor quantum dots. *Physical Review Letters* 79: 5282. Copyright 1997 American Physical Society.

the confined states. This experiment, unlike a photoluminescence experiment, also yields a value for the oscillator strength. The result can be understood simply in terms of the overlap integral between the hole and electron states. In other words, these dots are in the strong confinement regime where the confinement energies are larger than the exciton binding energy. In the other limit, the weak confinement regime, the exciton binding energy dominates over the confinement energies, and it makes sense only in this limit to think of an exciton moving as a coherent entity throughout the dot. In the weak confinement regime, the oscillator strength is actually larger than in the strong confinement regime.

An important issue in interpreting photoluminescence experiments is the rate at which a highly excited electron–hole pair can relax to the ground state of a quantum dot. Relaxation in the continuum is known to occur on a picosecond time-scale: electrons and holes relax by emitting LO-phonons if possible; and if not, acoustic phonons. Capture by a dot also occurs quickly (apart from in low-density samples where capture may take a few hundred picoseconds). It was initially thought that relaxation of carriers once captured by a quantum dot would be hindered by the discrete nature of the density of states. The argument is that the energy separation between the levels is too large for acoustic phonon emission; and there is no reason why the separation between the levels should match the LO-phonon energy (which has a very weak dispersion). Nonetheless, experiments have shown that relaxation from excited dot states to the ground state is rapid, occurring on a picosecond time-scale, pretty much as for quantum wells. Auger processes can be important, whereby an electron in a dot loses energy by promoting an electron in the continuum to a higher energy. But relaxation can also be fast without a large population of mobile carriers in the wetting layer, and this is thought to arise through multiphonon processes. The radiative lifetime is typically 1 ns for self-assembled quantum dots, depending to some extent on the size and material system. As for quantum wells, relaxation occurs much faster than recombination.

Lasers and Self-Assembled Quantum Dots

The motivation for using quantum dots as the gain medium for semiconductor lasers is that they should have advantageous properties over quantum wells. These include a lower threshold current, better temperature stability, and a higher modulation frequency. All of these possible advantages stem from the quantization. In a quantum dot, there are discrete levels each with low degeneracy so that population inversion is easy to achieve. Also, there are large potential barriers to the continuum of states in the surrounding matrix, and an energy separation between levels in the dot at least comparable to the thermal energy at room temperature. This should make a quantum dot device relatively insensitive to temperature. The threshold current of a semiconductor laser usually behaves as $I = I_0 e^{T/T_0}$ where T is the temperature, T_0 is the characteristic temperature, and I_0 the low-temperature threshold current. Generally speaking, lasers are desirable with a small I_0 and a large T_0.

The main obstacle to realizing the benefits of full spatial confinement in self-assembled quantum dots is the inhomogeneity. For a particular lasing wavelength only a fraction of the dots can contribute to the lasing signal. Additionally, the dot densities are typically about 10^{11} cm^{-2} for these applications, implying that in the layer plane there is more unoccupied space than dot material. All in all, the gain at a particular wavelength is limited. Nevertheless, it would appear that even with inhomogeneous dot ensembles the advantages inherent in zero-dimensional states can be exploited.

In addition to the density of states, there is an important difference in the nature of the gain spectrum between quantum well and quantum dot lasers. In a quantum well, gain exists over a range of energies through band filling. If an electron–hole pair is stimulated to emit by the lasing field, then the vacancies will be quickly filled by carrier–carrier scattering in the bands, maintaining a quasi-thermal equilibrium in the conduction and valence bands. In quantum dots at low temperature, capture into a particular dot is a random process, and given the rapid relaxation and the high barriers to the continuum, an electron–hole pair is very likely to emit from the dot where it was created. In other words, the system is not in thermal equilibrium. At higher temperatures, thermal escape out of the dots becomes likely, enabling the system to come into thermal equilibrium. The transition into thermal equilibrium will depend critically on the barrier height, and with appropriate design may take place close to, or even above, room temperature.

Quantum dot lasers have now been realized, and emit continuous wave at room temperature. Progress has been rapid in the past few years. In fact the best quantum dot lasers now have the lowest I_0 and the highest T_0 of any semiconductor laser diode.

Even if the intrinsic advantages of quantum dots are largely offset by the inhomogeneous broadening for laser applications, there are some other advantages. The InAs/GaAs quantum dot emission wavelength is significantly higher than that of InGaAs/GaAs quantum wells, and with subtle changes in the growth can be increased to 1.3 μm, corresponding to one of the low-loss windows in optical fiber. Also, much higher powers have been realized with quantum dot lasers than with quantum well lasers. In quantum well lasers, device failure occurs at high current, often due to degradation of the end mirrors. It would appear that the carrier localization in quantum dots is sufficiently strong to prevent diffusion to the end facets, allowing the devices to withstand higher currents.

Colloidal Quantum Dots

Progress has been made recently in making inorganic quantum dots with techniques from organic chemistry. Organometallic precursors are employed. The growth temperature is typically below 200 °C which has been found to give material with quite high crystalline quality. A crucial development has been the use of various organic groups to passivate the surface so that the quantum dots have bright photoluminescence. The II–VIs, being more polar than the III–Vs, are easier to produce with these techniques, and most work has concentrated on CdS and CdSe quantum dots which emit in the visible. Nevertheless, it is also possible to produce III–V quantum dots with this technology, and although the bulk GaAs bandgap lies in the near-infrared, the quantum confinement effects can be so high as to push the photoluminescence into the visible. Furthermore, HgCdTe has also been developed as a material system for quantum dots, and the emission can be tuned to the all-important 1.5 μm.

The spectral width of emission from these colloidal quantum dots tends to be very large because of a large spread in particle size. However, techniques are being developed to isolate a specific size. With a selective precipitation technique for instance, the spectral width can be reduced considerably to the extent that a fine structure can be observed in the optical response.

The advantage of these quantum dots lies in the possibility of mass-production for applications such as solar cells. The ease with which one material system can be used to cover a large spectral bandwidth is already attracting a lot of interest, for instance in using passivated and water-soluble quantum dots as biological labels. By using different materials, a huge spectral range can be covered with this technology, as illustrated in Figure 7. Of course, there are still challenges to be met in this field. The quantum dots tend to degrade in the course of time, losing their high quantum efficiency for emission, presumably as a result of changes to the surface structure. Improvements have been made recently by capping the dots with a high-bandgap semiconductor, such as ZnS, which hinders interactions of the exciton with the surface states. The emission from organic-based dots also tends to show temporal fluctuations in wavelength, revealed experimentally from the emission of a single dot. This is thought to come about through the charging and discharging of traps in the vicinity of the quantum dot. Fundamentally, there are difficulties injecting carriers across an interface into the colloidal dots. However, progress can be expected in all these areas, and these quantum dots may well find applications in novel photonic devices.

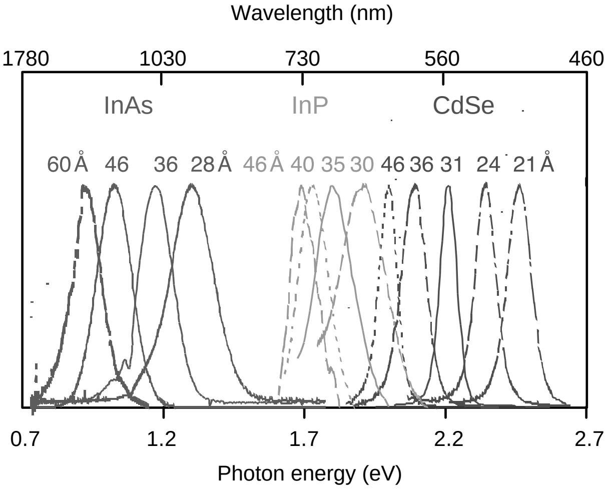

Figure 7 Room temperature emission spectra of surfactant-coated colloidal quantum dots. The quantum dots consist of InAs (left), InP (middle), and CdSe (right). Selection of the size has been used to give a particular wavelength in each band. Reprinted (abstracted/excerpted) with permission from Bruchez M Jr, Moronne M, Gin P, Weiss S and Alivisatos AP (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science* 281: 2013. Copyright 1998 American Association for the Advancement of Science.

Single-Dot Spectroscopy

No matter how quantum dots are prepared, there is always a distribution in emission energy. This constitutes an inhomogeneous broadening mechanism. For self-assembled InAs dots on GaAs, the lowest emission width from an ensemble of dots is about 20 meV. This energy is large compared to several other energies in the system. For example, the binding energy of a charged exciton, or of an exciton–exciton complex (a biexciton), is about 2 meV in this system. Also, the width of single dot emission is two to three orders of magnitude smaller than the ensemble width at low temperature. The width of single-dot emission is inversely related to the time over which the exciton maintains its coherence, and this should be as long as possible for future applications of quantum dots which manipulate the phase of the excitonic wavefunction. In order to investigate the properties of quantum dots in detail it is clearly necessary to measure the optical response of a single quantum dot rather than the response of a large ensemble of dots. The development of this field has been aided by the techniques developed for single-molecule spectroscopy which was first reported in the late 1980s and has subsequently become almost a standard experimental technique.

Typical densities of self-assembled quantum dots correspond to 100 per μm^2. Given that the emission wavelength is around 1 μm, it is clear that an optical technique with subwavelength spatial resolution is required to excite and detect a signal from just one quantum dot. The techniques employed include cathodoluminescence, where a focused electron beam provides the spatial resolution, and near-field optics where a subwavelength-sized aperture in a metal mask determines the spatial resolution. In the latter case, the aperture is either permanently connected to the sample surface, or it is formed on a tapered optical fiber positioned just a few nanometers away from the sample surface. An example is shown in Figure 8: with a confocal microscope the spatial resolution is not high enough to isolate spectroscopically a single dot. With increased spectral resolution, however, just a small number of peaks remain, each corresponding to the emission from a single dot. A further technique is to etch nanometer-sized mesas in the sample such that each mesa contains on average just one quantum dot. In some cases, single quantum dot emission has been detected with a conventional optical microscope (usually a confocal microscope) by adjusting the growth conditions to give particularly low dot densities. This can be achieved with colloidal nanocrystals by spinning a dilute suspension of the dots onto a substrate, a technique which has been frequently used in single-molecule spectroscopy. In each case, the photoluminescence signal from a single quantum dot is not large, and a detector capable of counting individual photons with a high quantum efficiency is required. A cooled silicon CCD camera is very convenient, giving not only a very low dark count and high quantum efficiency, but also a huge multiplexing advantage. However, the response of a Si detector tails off in the visible, and decreases rapidly at about 1050 nm in the near infrared, so it is by no means a universal solution.

The most striking feature of the emission from a single inorganic quantum dot is its spectral purity. Linewidths as small as a few μeV (a few hundred MHz) have been reported at low temperature for

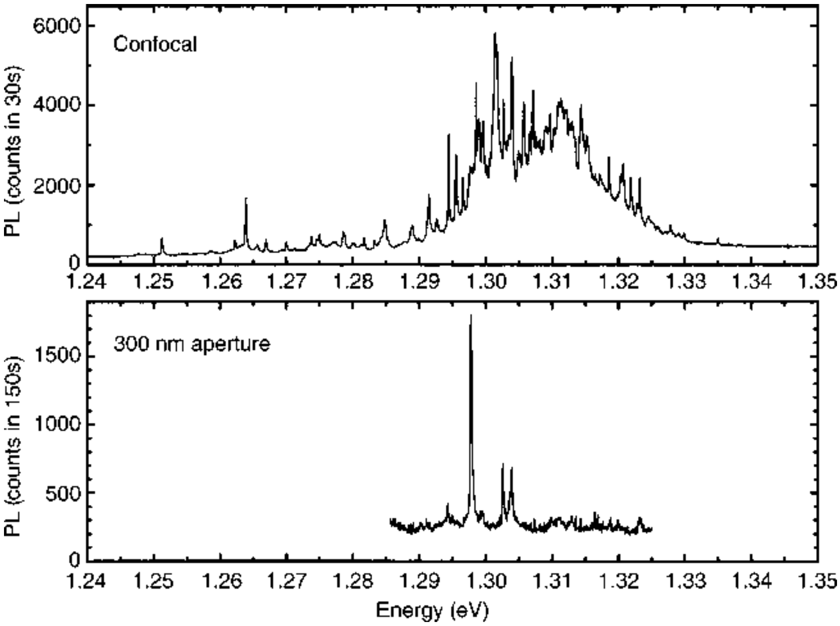

Figure 8 Emission from quantum rings from a 1 μm^2 area of the sample, and from a 0.07 μm^2 area of the sample, both taken at 4.2 K. The quantum rings are generated with a Stranski–Krastanow growth procedure using InAs and GaAs. The first spectrum was taken with a low-temperature confocal microscope; the second with a near-field technique. It can be seen how large statistical fluctuations arise in the first spectrum, because the number of rings probed is not sufficient to give a smooth distribution function. In the second spectrum, just a few peaks can be discerned above the noise level, each arising from the emission from an individual quantum ring.

quantum dots formed in a localizing potential in a quantum well. These linewidths can be interpreted as de-phasing times of the excitons. The maximum possible de-phasing time is obviously the radiative lifetime, in which case the line is homogeneously broadened. This limit has been observed at low temperature on natural quantum dots, but not at elevated temperature: as the temperature increases, scattering of the excitons by phonons decreases the coherence time, observed through a broadening of the emission.

The sharp lines observed in single-dot spectroscopy give access to fine structure on a meV energy scale. Most of this fine structure arises through Coulomb interactions between the electrons and holes. For example, two recent experiments have explored the emission as a function of electron occupation, and exciton population. In both cases, new lines emerge when a single dot is populated by more than one electron–hole pair.

Perspectives

The field of semiconductor quantum dots has made rapid advances in the past decade largely through the discovery of self-assembly in the growth of lattice-mismatched inorganic semiconductors. The most immediate application of these defect-free nanostructures is as the gain medium for semiconductor laser diodes, the aim being to reduce the threshold currents beyond those achievable with quantum wells. This approach is already proving highly successful. However, there are also other possibilities for novel photonic applications of quantum dots. These include memory elements for image processing, detectors and sources for single photons, and perhaps even elements for the coherent manipulation of wavefunctions.

See also

Semiconductor Materials: Band Structure Engineering; GaAs Based Compounds. **Semiconductor Physics:** Band Structure and Optical Properties; Excitons; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Bayer M, Stern O, Hawrylak P, Fafard S and Forchel A (2000) Hidden symmetries in the energy levels of excitonic artificial atoms. *Nature* 405: 923.

Bimberg D, Grundmann M and Ledentsov NN (1998) *Quantum Dot Heterostructures*. Chichester, UK: Wiley.

Bruchez M, Jr, Moronne M, Gin P, Weiss S and Alivisatos AP (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science* 281: 2013.

Empedocles SA and Bawendi MG (1997) Quantum-confined Stark effect in single CdSe nanocrystalline quantum dots. *Science* 278: 2114.

Fricke M, Lorke A, Kotthaus JP, Medeiros-Ribeiro G and Petroff PM (1996) Shell structure and electron–electron interaction in self-assembled InAs quantum dots. *Europhysics Letters* 36: 197.

Gaponenko SV (1998) *Optical Properties of Semiconductor Nanocrystals*. Cambridge, UK: Cambridge University Press.

Leonard D, Pond K and Petroff PM (1994) Critical layer thickness for self-assembled InAs islands on GaAs. *Physical Review B* 50: 11687.

Lipsanen H, Sopanen M and Ahopelto J (1995) Luminescence from excited states in strain-induced $In_xGa_{1-x}As$ quantum dots. *Physical Review B* 51: 13868.

Tersoff J, Teichert C and Lagally MG (1996) Self-organization in growth of quantum dot superlattices. *Physical Review Letters* 76: 1675.

Warburton RJ, Dürr CS, Karrai K, *et al.* (1997) Charged excitons in self-assembled semiconductor quantum dots. *Physical Review Letters* 79: 5282.

Warburton RJ, Schäflein C, Haft D, *et al.* (2000) Optical emission from a charge-tunable quantum ring. *Nature* 405: 926.

Weisbuch C and Vinter B (1991) *Quantum Semiconductor Structures*. San Diego, CA: Academic Press.

Xie Q, Madhukar A, Chen P and Kobayashi NP (1995) Vertically self-organized InAs quantum box islands on GaAs(100). *Physical Review Letters* 75: 2542.

Type-II Quantum Wells and Superlattices

I Vurgaftman and J R Meyer, Naval Research Laboratory, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

When thin layers of two semiconductor materials are combined to form a heterojunction, quantum well (QW), or superlattice, the properties of the resulting 'quantum heterostructure' are to a large extent governed by the alignment of the conduction and valence bands at the interfaces. Usually the overall conduction-band (CB) minimum lies in the same material as the overall valence-band (VB) maximum, as for example in the GaAs/$Al_{0.3}Ga_{0.7}As$ QW illustrated in Figure 1a. In such a type-I structure,

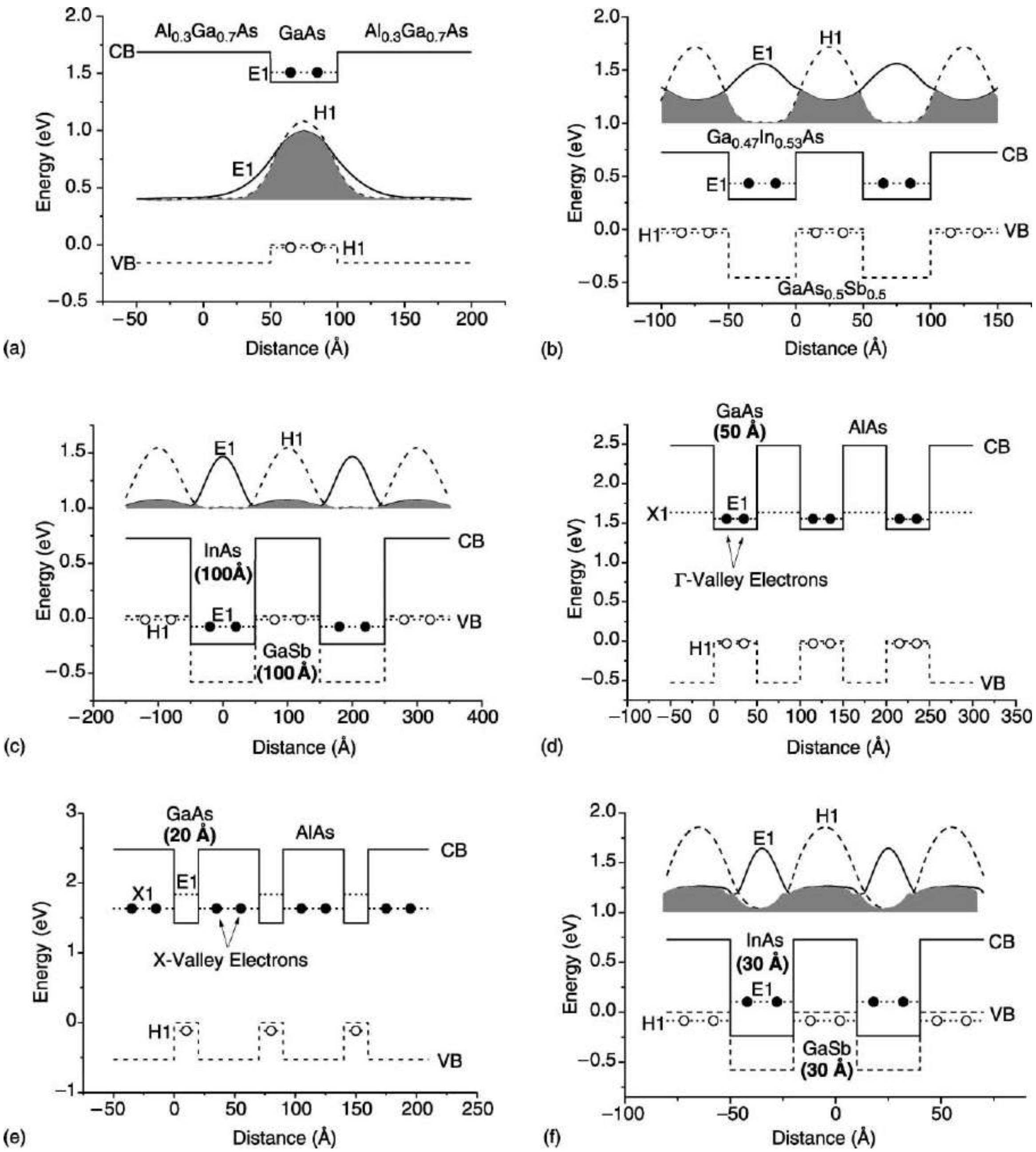
Al0.3Ga0.7As
GaAs
Al0.3Ga0.7As
CB
E1
H1
E1
VB
H1
Energy (eV)
Distance (Å)
(a)
H1
E1
Ga0.47In0.53As
CB
E1
H1
VB
GaAs0.5Sb0.5
Energy (eV)
Distance (Å)
(b)
E1
H1
InAs (100Å)
CB
H1
E1
VB
GaSb (100 Å)
Energy (eV)
Distance (Å)
(c)
GaAs (50 Å)
AlAs
CB
X1
E1
Γ-Valley Electrons
H1
VB
Energy (eV)
Distance (Å)
(d)
GaAs (20 Å)
AlAs
CB
X1
E1
X-Valley Electrons
H1
VB
Energy (eV)
Distance (Å)
(e)
H1
E1
InAs (30 Å)
CB
E1
H1
VB
GaSb (30 Å)
Energy (eV)
Distance (Å)
(f)

Figure 1 Conduction band (CB) and valence band (VB) profiles along with selected sub-band energies and wavefunctions for: (a) 50 Å GaAs/$Al_{0.3}Ga_{0.7}As$ type-I quantum well (GaAs substrate); (b) 50 Å/50 Å $Ga_{0.47}In_{0.53}As/GaAs_{0.5}Sb_{0.5}$ type-II staggered-gap superlattice (InP substrate); (c) 100 Å/100 Å InAs/GaSb type-II broken-gap semimetallic superlattice (GaSb substrate); (d) 50 Å/50 Å GaAs/AlAs type-I superlattice (GaAs substrate); (e) 20 Å/50 Å GaAs/AlAs type-II superlattice with indirect band alignment both in real and momentum spaces; and (f) 30 Å/30 Å InAs/GaSb type-II semiconducting superlattice. E1, H1, and X1 represent the first sub-band in the Γ-valley conduction band, Γ-valley valence band, and X-valley conduction band, respectively. The overlap between the electron and hole wavefunctions is shown in gray.

the electrons and holes both reside mostly in the GaAs and their wavefunctions overlap almost perfectly as shown in the figure.

There are cases, however, when the CB minimum is in one layer while the VB maximum is in the other. For this so-called type-II alignment, which is exemplified by the $Ga_{0.47}In_{0.53}As/GaAs_{0.5}Sb_{0.5}$ superlattice of Figure 1b, the electrons and holes tend to separate spatially. It should not be surprising that this significantly influences many of the basic optical and electronic properties. Just one example is the band-edge absorption, whose strength is proportional to the square of the wavefunction overlap integral.

In contrast to the staggered type-II heterostructure of Figure 1b, another possibility is the broken-gap type-II alignment of the InAs/GaSb superlattice shown in Figure 1c. Since the CB minimum in InAs actually lies below the VB maximum in GaSb, this combination becomes a semimetal even though the bulk constituents are both semiconductors.

The nature of the band alignment can also be strongly influenced by shifts of the electron and hole energy levels induced by quantum confinement. Figure 1d illustrates a type-I GaAs/AlAs superlattice with relatively thick GaAs layers. Notice that the same superlattice becomes type-II if, as in Figure 1e, the GaAs layers are thin enough that the GaAs Γ-valley electron level (center of the Brillouin zone) is pushed to a higher energy than the AlAs X-valley level (boundary of the Brillouin zone). Similarly, quantum confinement can turn the broken-gap (semimetallic) type-II InAs/GaSb superlattice of Figure 1c into a staggered (semiconducting) type-II superlattice as in Figure 1f.

Although many different type-II heterostructures have been identified over the last three decades, only a few are likely to have long-term technological importance for optics or electronics. The system that has received the most attention in terms of both fundamental scientific studies and opto-electronic device development is InAs/GaSb and related alloy combinations such as InAs/GaInSb. Following a brief overview of wider-gap type-II heterostructures, this article will concentrate primarily on that interesting narrow-gap system.

Wide-Gap Type-II Heterostructures

The best-known wide-gap type-II heterostructure is the GaAs/AlAs superlattice of Figure 1e. The band lineup is type-II provided the GaAs thickness is no greater than ≈10 monolayers (28 Å) and the AlAs layer is no thinner than the GaAs layer. Since the CB minimum is in the AlAs X-valley and the VB maximum is in the GaAs Γ-valley, this structure has an indirect energy gap in both real space and momentum space. In type-II GaAs/AlAs, the X-valley states at the CB minimum form one or more superlattice minibands, which display relatively low resistance at small biases, negative differential velocity, etc. A related structure is the GaAs/AlAs/GaAs/AlAs/GaAs double-barrier intervalley resonant tunneling diode, in which resonant or sequential phonon-assisted tunneling proceeds via intervalley transitions whenever the Fermi level in the GaAs emitter lines up with the X-valley minimum in AlAs. Structures containing AlGaAs do not necessarily need quantum confinement to establish a type-II alignment, since it occurs naturally in $Al_xGa_{1-x}As$/AlAs heterojunctions with $x > 0.3$.

The device potential of wide-gap type-II heterostructures is rather limited, owing to their low luminescence efficiencies (due in large part to the doubly indirect band alignment) and low mobilities (due to the heavy X-valley effective mass). A few proposals have focused on modulation of the band alignment between the type-I and type-II regimes. The $Ga_{0.47}In_{0.53}As/GaAs_{0.5}Sb_{0.5}$ superlattices of Figure 1b and related structures have been investigated as alternative active regions for semiconductor lasers emitting in the important 1.3–2 μm wavelength range.

InAs/GaSb/AlSb-Based Type-II Superlattices and Quantum Wells

Molecular beam epitaxy (MBE) was used to fabricate the first InAs/GaSb-based superlattices in 1977, only a few years after the initial GaAs/AlGaAs superlattice growths. It was quickly demonstrated that for relatively thick layers the band alignment is broken-gap type-II, with 100–150 meV overlap between the conduction band minimum in InAs and the valence band maximum in GaSb. In this semimetallic superlattice, GaSb valence-band electrons spill over into the InAs and leave behind holes (see Figure 1c). It was also observed that at equal layer thicknesses of ≈90 Å quantum confinement induces a semimetal-to-semiconductor transition (Figure 1f), and that the gap becomes increasingly positive when the thickness is reduced further. This is no mere scientific curiosity, since being able to tune in a wide range of energy gaps spanning 0 to >0.5 eV has enormous practical consequences for infrared (IR) emitters and detectors. Such a capability cannot be replicated by any other III–V heterostructure system.

Another happy accident is that the lattice constants of InAs (6.058 Å) and GaSb (6.096 Å) are matched closely enough to allow them to be grown together by MBE or metalorganic chemical vapor deposition (MOCVD). A third binary member of the 'antimonide' or '6.1-Å' family of growth-compatible heterostructures is AlSb (6.136 Å), which is valuable in that it serves as a quantum barrier to both InAs conduction states and GaSb valence states. The extended family includes a variety of ternary alloys (GaInSb, GaAlSb, InAsSb, AlAsSb, etc.), whose energy gaps as a function of lattice constant are given by the curves in Figure 2, and also at least three quaternary alloys: GaInAsSb, AlGaAsSb, and InAlAsSb. Taken together, this menagerie of materials

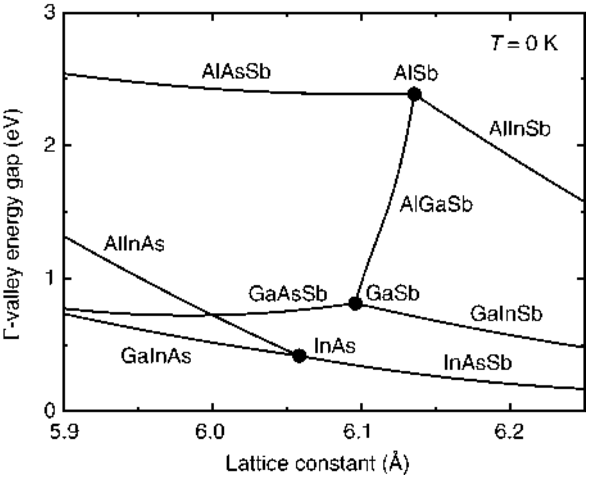

Figure 2 Direct energy gaps at zero temperature for III–V compound semiconductors and their ternary alloys with lattice constants in the vicinity of 6.1 Å.

provides the type-II IR device designer with an exceptional degree of configurational flexibility.

Type-II IR structures are typically grown on either a GaSb or InAs substrate. Since even a small amount of strain build-up can produce dislocation defects that seriously degrade the performance, one usually balances the net tensile strain (associated with layers having a lattice constant slightly larger than that of the substrate) with a compensating amount of net compressive strain (from layers whose lattice constant is slightly smaller than that of the substrate).

Band structures for complicated type-II layering configurations have been calculated by a variety of methods, including $k\cdot p$, tight-binding, and empirical pseudopotentials. While all reproduce the main features of the electron and hole sub-bands and dispersion relations near the center of the Brillouin zone, the relative merits of the different approaches remain controversial. Due to the small energy gaps, band-mixing effects are especially important in this system.

One consequence of the small type-II antimonide energy gap is that the electron and hole bands repel each other and both effective masses can be lighter than in any other III–V semiconductor (e.g., smaller than in the bulk constituents). This leads to high mobilities at room temperature, and promising possibilities for high-electron-mobility transistors (HEMTs). Also of interest are resonant tunneling diodes (RTDs) based on InAs/AlSb/GaSb/AlSb/InAs and related structures, in which interband tunneling via the GaSb valence states results in negative differential resistance with a very high peak-to-valley ratio. Since these devices are primarily electronic rather than optical, we will not discuss them further here.

While interfaces play an important role in any semiconductor heterostructure, they are especially crucial in InAs/GaSb and related type-II systems because the two materials on opposite sides of the interface have neither a common cation nor a common anion (unlike GaAs/AlGaAs, for example, in which both wells and barriers have As anions). By timing the shutter sequences in the growth, one can in principle end the InAs on an In layer and begin the GaSb on Sb, to produce InSb-like interface bonds. Alternatively, one can end the InAs on As and begin the GaSb on Ga to form GaAs-like bonds. Detailed characterizations by a variety of methods have shown that the device properties and even the band structure can be relatively sensitive to whether the interface bonds are InSb-like or GaAs-like.

With regard to opto-electronic devices, a potential limitation of InAs/GaSb superlattices is that in order to reach a long wavelength (small energy gap), the layers must be rather thick (e.g., ≈90 Å for zero gap). In such a structure the overlap between the electron and hole wavefunctions becomes quite small (see Figures 1c and 1f), and consequently the interband optical interactions are weak. However, Mailhiot and Smith pointed out that this issue can be circumvented to a large extent by employing $Ga_{1-x}In_xSb$ (with $x \leq 0.4$) rather than GaSb for the hole QWs. Compressive strain in the GaInSb produces a large splitting of the heavy and light hole sub-bands, along with a substantial increase of the type-II offset energy. The result is that much thinner layers (with large wavefunction overlaps) may be employed in conjunction with energy gaps corresponding to long IR wavelengths. In fact, the overlap for a short-period InAs/GaInSb superlattice such as that illustrated schematically in Figure 3 can be large enough to yield interband optical interaction strengths comparable to those in bulk materials and type-I superlattices.

While practical constraints on the minimum reproducible layer thickness set a lower wavelength limit on the order of 2 μm, the upper limit governed by uniformity considerations is quite high, perhaps in the $\lambda = 25-100$ μm range. This makes type-II antimonide quantum heterostructures a natural choice for mid-IR to far-IR photodetectors and lasers.

Type-II IR Photodetectors

Because IR detectors sense the radiant black-body emission from warm objects rather than relying on reflected visible light, they can be used for 'night vision.' Both photoconductive (PC) and

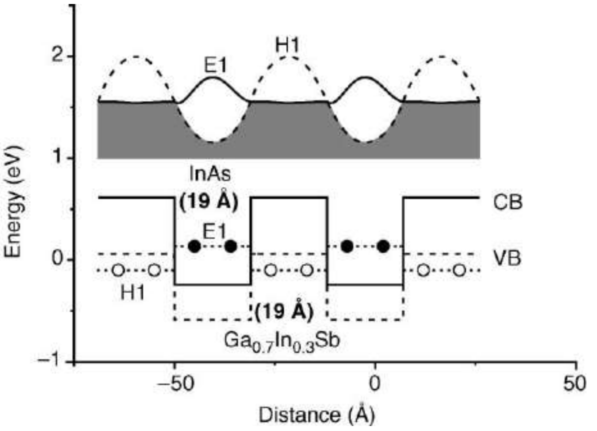

Figure 3 Conduction and valence band profiles along with electron (E1) and hole (H1) sub-band energies and wavefunctions for a 19 Å/19 Å InAs/$Ga_{0.7}In_{0.3}Sb$ type-II superlattice (GaSb substrate) with an energy gap equal to that of the InAs/GaSb superlattice in **Figure 1d**.

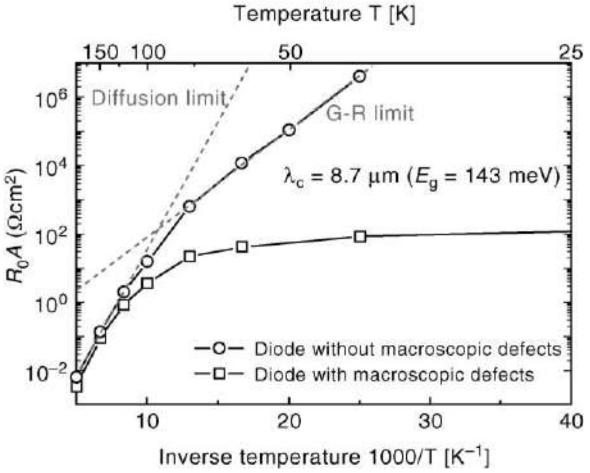

Figure 4 Temperature dependence of the dynamic impedance at zero bias R_0A for two different InAs/GaInSb superlattice diodes with cutoff wavelengths of 8.7 μm. The generation-recombination (G-R) and diffusion-limited R_0A values are indicated by the dashed lines. Reproduced with permission from Bürkle *et al.* (2000) *MRS Symp. Proc.*, 607, pp. 77–82.

photovoltaic (PV) IR detectors are now being developed with type-II active regions, usually consisting of simple InAs/GaInSb superlattices. The PV geometry, in which a reverse bias is applied to a p–n junction, is ultimately preferred since it is easily adaptable to pixelation in a 2D array with many discrete detector elements. That allows an IR camera to construct a detailed thermal picture of any given scene.

An optimized PV photodetector must have high absorption in the active region, unimpeded vertical transport to assure collection of the photogenerated carriers, and a low dark current. The thin layers of a type-II InAs/GaInSb active region help to satisfy the first two criteria, since they assure both a large wavefunction overlap (**Figure 3**) and rapid electron and hole tunneling through the superlattice minibands.

The dark current in a high-quality diode tends to be dominated by generation-recombination currents at low temperatures and diffusion currents at higher *T*. However, trap-assisted tunneling via dislocations and other macroscopic defects can significantly increase the dark current. **Figure 4** plots the resistance–area product (R_0A) for diodes with and without macroscopic defects. Surface leakage resulting from inadequate sidewall passivation can also lead to a substantial contribution that is proportional to the perimeter of the mesa. The diffusion current scales as $1/\tau^{1/2}$, where τ is the recombination lifetime. To date, the defect-mediated Shockley–Read lifetimes of <100 ns have been significantly shorter than in competing HgCdTe IR materials (>1 μs). However, Auger lifetimes are substantially enhanced, as will be discussed further in the context of type-II IR lasers. In an Auger event, the energy and momentum of the recombining electron–hole pair is transferred to a third carrier, which scatters to a higher-lying state in either the same or a different band. Because three different carriers must interact simultaneously, the recombination rate scales with the cube of the carrier density. Usually Shockley–Read recombination dominates at lower temperatures and Auger processes at higher *T*.

Typically the n-side of a type-II diode is heavily doped (10^{18} cm^{-3} range) and relatively thin, while most of the absorption occurs on the lightly doped (10^{16} cm^{-3} range) p-side. A standard figure of merit is the detectivity (D^*), which is essentially the signal-to-noise ratio in a device of standard size. For 77 K operation, type-II PV detectors with cutoff wavelengths of 7.5 μm and 12 μm achieved D^* of 1×10^{12} cm Hz$^{1/2}$/W and 5×10^{10} cm Hz$^{1/2}$/W, respectively. Those values are nearly as high as typical results for the much more mature HgCdTe FPA technology.

It is anticipated that type-II detectors will be particularly attractive at very long wave IR (VLWIR), at which it becomes increasingly difficult to maintain adequate control over the HgCdTe compositional uniformity. Recent PC detectors operated out to 22 μm, and PV devices to 16 μm. Due in part to the strong suppression of Auger recombination, type-II detectors are also expected to be advantageous at noncryogenic temperatures where future systems will increasingly operate.

Type-II IR Lasers

At visible and near-IR wavelengths out to ≈2 μm, the QW diode laser is by now a mature technology that dominates the commercial marketplace due to its low cost (<$1 per laser in some cases), compactness, and reliability. However, the technical challenges become much more severe in the mid-IR spectral region (3–5 μm). Before the advent of the type-II laser, no interband III–V semiconductor device emitting at $\lambda \geq 3$ μm had ever operated at room temperature. Alternatives are the lead salt laser, which is limited to very low-output powers, and the quantum cascade laser (QCL) that is based on optical transitions between two electron QW sub-bands rather than interband electron–hole processes. Mid-IR lasers are needed for high-power military applications such as IR countermeasures against heat-seeking missiles and for the sensitive detection of trace chemicals using the spectroscopic 'fingerprint' that distinguishes each species from any other.

From the 1960s to the late 1990s, conventional III–V double heterostructure (DH) and type-I QW lasers for the mid-IR never achieved the desired levels of performance because they failed to overcome certain fundamental limitations. Those included:

1. inadequate band offsets, which at higher temperatures made it difficult for the QW or DH active regions to retain the injected electrons and holes;
2. rapid Auger nonradiative decay, which led to excessive threshold current densities at higher temperatures; and
3. restricted spectral range, limited to wavelengths shorter than the cutoff for bulk $InAs_{1-x}Sb_x$.

However, it has more recently become apparent that all of these limitations can be either minimized or removed by employing a type-II InAs/GaSb-based active region.

For example, electrical confinement is far more effective because the GaSb conduction band confines InAs electrons and the InAs valence band confines GaSb holes (see Figure 1f). Auger recombination is also significantly suppressed, in part because the holes are nearly as light in the plane as the electrons, making it much more difficult to conserve both momentum and energy in the three-carrier process. Nearly an order-of-magnitude suppression of the room-temperature Auger rate has been reported, which translates directly into a similar reduction of the threshold current density. And as was discussed above in connection with IR photodetectors, very long wavelengths are readily attainable because the energy gap in a type-II active region can be arbitrarily small. Type-II lasers have already operated cw at $\lambda > 7$ μm, and cryogenic operation out to wavelengths as long as 100 μm has been projected theoretically.

In 1986 the first type-II semiconductor laser, which emitted at $\lambda = 1.86$ μm from single liquid-phase epitaxy (LPE)-grown heterojunctions, was demonstrated. Since the mid-1990s, 3–4 μm GaInAsSb/InAs-based single-interface type-II diodes have been developed that operate up to 205 K in pulsed mode.

In 1995, lasing at $\lambda = 3.2$–3.5 μm was reported from type-II $InAs/Ga_{0.75}In_{0.25}Sb$ superlattice diodes grown by MBE. Subsequent work on these devices extended the temperature range to 255 K in pulsed mode and 180 K for cw operation, and the emission wavelength to 4.3 μm.

While the wavefunction overlap and hence the gain for a two-constituent type-II superlattice active region can be quite favorable, penetration of the electron wavefunctions into the thin GaInSb layers is in fact so great that a wide miniband forms along the growth axis. The electrons then have strong energy dispersion in all three dimensions, whereas present-generation near-IR and visible semiconductor lasers nearly always have QW active regions with 2D carriers. The stepped quasi-2D density of states yields a much higher differential gain per injected carrier at threshold, and also narrower spectral emission lines than are attainable with a 3D electron population. The most straightforward approach to blocking the formation of a miniband in the superlattice is to add a high electron barrier between the periods, e.g., InAs/GaInSb/AlSb. However, the drawback of that approach is that the wavefunction overlap becomes much smaller than in the simple superlattice, resulting in significantly lower gain.

On the other hand, if a second electron well is added on the other side of the hole well, the resulting structure with four constituents per period (e.g., InAs/GaInSb/InAs/AlSb) combines the large wavefunction overlap of the two-constituent superlattice with 2D electron and hole confinement. Figure 5 illustrates band profiles, energy levels, and wavefunctions for the 'W' laser, which takes its name from the shape of the conduction-band profile.

Optically pumped type-II W lasers display high-power pulsed operation at temperatures well above ambient. Wavelengths have spanned the range 2.6–7.3 μm, and mid-IR vertical-cavity surface-emitting lasers (VCSELs) were also fabricated. Optical pumping is viewed primarily as a means of achieving high cw or quasi-cw output powers in the mid-IR, and in that regard W lasers have exceeded all other semiconductor approaches with 1.5 W peak output power being obtained at 71 K for 100 μs

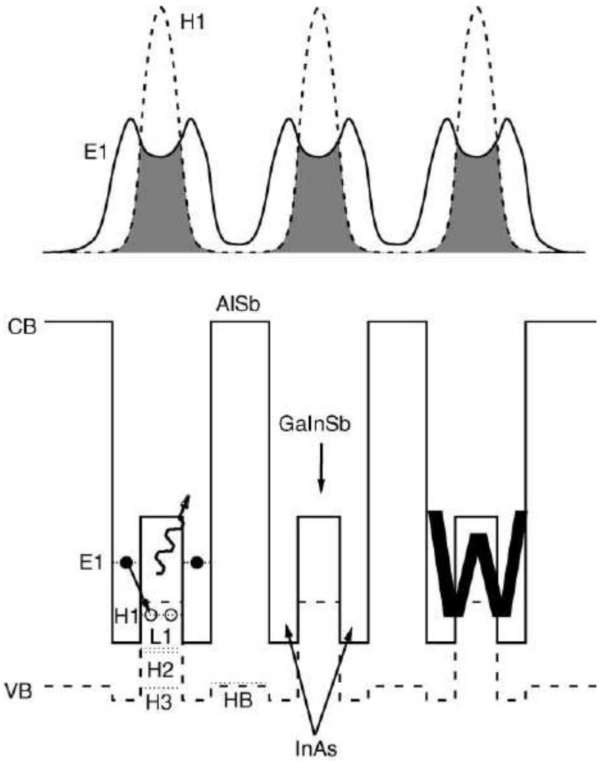

Figure 5 Conduction and valence band profiles along with selected sub-band energies and wavefunctions for InAs/$Ga_{0.7}In_{0.3}Sb$/InAs/AlSb type-II 'W' structure (GaSb substrate). Electron (E1), heavy hole (H1, H2, H3), light hole (L1), and barrier heavy hole (HB) sub-bands are indicated.

pulses with a 10% duty cycle. The average output at 82 K for that device was up to 360 mW for 20 μs pulses and a 50% duty cycle.

Cw operation at $\lambda = 3.0$ μm was observed nearly to room temperature (290 K), and even a 6.1 μm emitter operated cw to 210 K. The cw output power of 540 mW/facet from a device with 80 W QW periods was the highest ever reported for 78 K operation in the mid-IR. Despite the rather high output power, however, the differential power conversion efficiency was only 2.5% at 78 K and fell to 1% at 140 K. Much higher efficiencies were obtained using the optical pumping injection cavity (OPIC) approach, in which the active region was surrounded by semiconductor quarter-wave mirror stacks that formed an etalon cavity resonant at the 2.1 μm pump wavelength. For pulsed operation, OPIC devices with uncoated facets and only 10 active QWs yielded lower thresholds, lower internal losses, and higher power conversion efficiencies than any previous results at $T \geq 200$ K. Figure 6 illustrates that the efficiencies were 7.1% at 220 K for one device and 4% at 275 K for another.

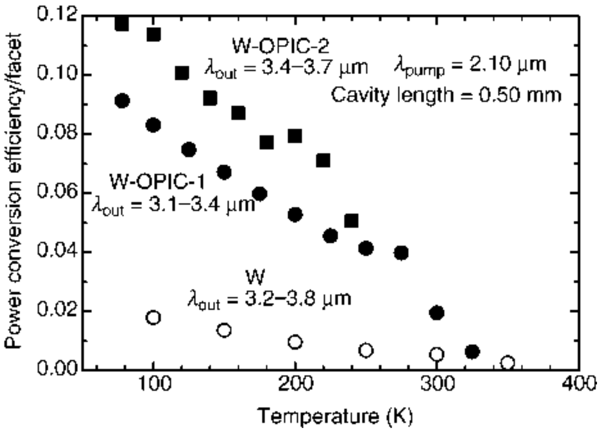

Figure 6 Power conversion efficiency per uncoated facet as a function of temperature for two different W-OPIC lasers emitting in the wavelength ranges of 3.1–3.4 μm (filled circles) and 3.4–3.7 μm (filled squares), respectively. For comparison, conversion efficiencies for a typical 'W' laser without an OPIC emitting in the same spectral range (open circles) are also shown. All of the lasers were optically pumped in pulsed mode by a 2.1 μm Ho:YAG laser, and the cavity length for all three devices was 0.5 mm.

An alternative approach to maximizing the pump absorbance despite a small number of QWs is the 'integrated absorber' (IA) concept. In this approach, the W QWs are separated by intermediate-gap GaInAsSb absorbing layers which donate optically generated electrons and holes to the active wells. At 80 K with 35 μs pulses at 2.5% duty cycle, an IA laser with only five QWs separated by ≈1600-Å-thick absorber layers had a power conversion efficiency of 9.8% and peak output power of 2.1 W. Thus far the IA has not done as well as the OPIC at high temperatures, due mostly to an inadequate valence band offset between the absorber layers and the active QWs.

The beam quality for optically pumped type-II lasers was substantially improved by using the angled-grating distributed-feedback (α-DFB) approach pioneered at shorter wavelengths. Since both the grating and the facets are needed to produce optical feedback, only a narrow angular cone is fed back into the tilted cavity. For a pump-stripe width of 50 μm, the near-diffraction-limited output beam had 15 times narrower angular divergence (1.4° full width at half-maximum (FWHM)) than the double-lobed beam emitted from a conventional Fabry–Perot laser occupying an unpatterned portion of the same bar. Even more promising results have been reported using the photonic-crystal distributed-feedback (PCDFB) laser, in which the grating is defined on a two-dimensional rectangular lattice that is tilted with respect to the facets as in the α-DFB. For a pump-stripe width of 200 μm, the beam quality for the direction along the laser stripe was four times the diffraction limit.

Although optical pumping is the simplest way to inject electrons and holes into the active region, type-II diode lasers will ultimately be much more inexpensive, convenient, compact, and energy efficient. However, generalization of the W configuration to electrical pumping is not entirely straightforward. This is because holes injected from the p-type cladding layer may have poor growth-axis transport, and hence hole populations in the various GaInSb QWs could be highly nonuniform. Whereas the holes in a superlattice are required to tunnel through only one InAs electron well at a time (see Figure 3), the barrier for the structure in Figure 5 consists of two InAs layers in addition to the AlSb barrier. To overcome this problem the AlSb was replaced with a specially designed quaternary $Al_{0.15}Ga_{0.85}As_{0.05}Sb_{0.95}$ barrier, which raised the L1 light-hole sub-band to within ≈100 meV of the H1 heavy-hole valence band maximum. That permitted thermally excited light holes to provide the needed tunneling transport to all of the QWs. Another key design element was the insertion of relatively thick (0.6 μm) undoped $Al_{0.35}Ga_{0.65}As_{0.03}Sb_{0.97}$ 'broadened-waveguide' layers on each side of the active region, in order to minimize free-carrier absorption in the doped $Al_{0.9}Ga_{0.1}As_{0.07}Sb_{0.93}$ cladding layers. Some of the devices also contained InAs/AlSb 'hole-blocking' layers, which prevented hole leakage into the n-type broadened waveguide layer.

A pulsed $\lambda = 3.27$ μm W-diode of this type with 10 active QW periods achieved the first room temperature operation (to 310 K) of an electrically pumped III–V interband laser in the mid-IR. Another device with five QWs operated in cw mode to 200 K, the highest $T_{\max}$ to date for any interband III–V laser beyond 3 μm. The threshold at 78 K was 63 A/cm^2, and Figure 7 shows that the cw output was 140 mW at 78 K and 20 mW at 180 K. Following optimization of the turn-on voltages, series resistances, and internal losses, cw diodes with one–three QWs are projected to operate in the thermo-electric cooler range and ultimately at room temperature.

A more exotic variation on the type-II diode emitter is the interband cascade laser (ICL) that was first proposed by RQ Yang. Like the intersub-band QCL, the ICL generates multiple photons per injected electron by making an optical transition at each step of a staircase-like QW structure. While the threshold current density for the QCL is high because of the inherently short lifetimes associated with intersubband phonon scattering, the interband configuration of the ICL eliminates that nonradiative relaxation path. Early projections that the ICL should display low thresholds, high cw operating temperatures, high output powers, and high differential quantum efficiencies (well above the conventional limit of one photon/electron) have all been confirmed experimentally.

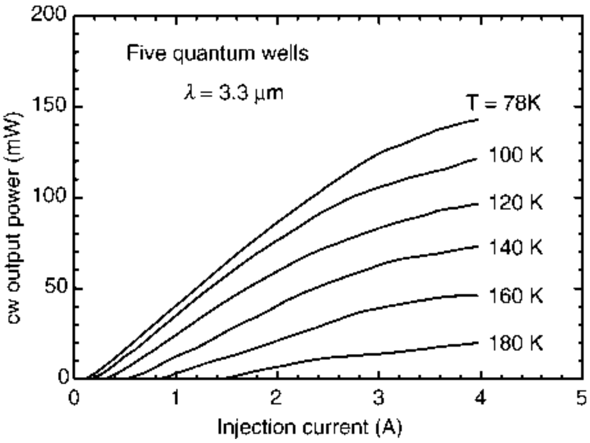

Figure 7 Cw output power as a function of injection current at temperatures between 78 K and 180 K for a 'W' diode laser with a 2-mm-long cavity and 100-μm-wide stripe emitting at a wavelength of 3.3 μm.

Figure 8 illustrates conduction and valence band profiles for a typical single stage of the ICL staircase. While this example employs a W configuration, in which two InAs electron QWs (19 Å and 13 Å) surround the active GaInSb hole QW (32 Å), some designs employ only a single InAs active well. Electrons tunneling into the InAs QWs from the preceding injection region at left emit a photon by making spatially indirect radiative transitions to the GaInSb valence band. They next tunnel into the valence states of the adjacent GaSb and GaInSb hole wells, whose function is to provide a thick barrier to prevent electron leakage from the active InAs QWs directly to the 70 Å InAs well that begins the next injection region. From the final GaInSb hole QW (40 Å), the carriers undergo near-elastic interband scattering into conduction states of the 70 Å InAs QW. Electrons traversing the superlattice miniband of the injection region, which in this case consists of eight digitally graded InAs wells, finally tunnel into the active electron QWs of the next period. This series of events is repeated at each step of the ICL staircase.

Following the first ICL demonstration in 1997, there have been reports of threshold current densities 13 A/cm^2 at 80 K, maximum operating temperatures of up to 300 K for pulsed operation and 150 K for cw (with epitaxial-side-up mounting), pulsed output powers of 6 W/facet and slope quantum efficiencies exceeding 600%. A cw wall plug efficiency of 22%

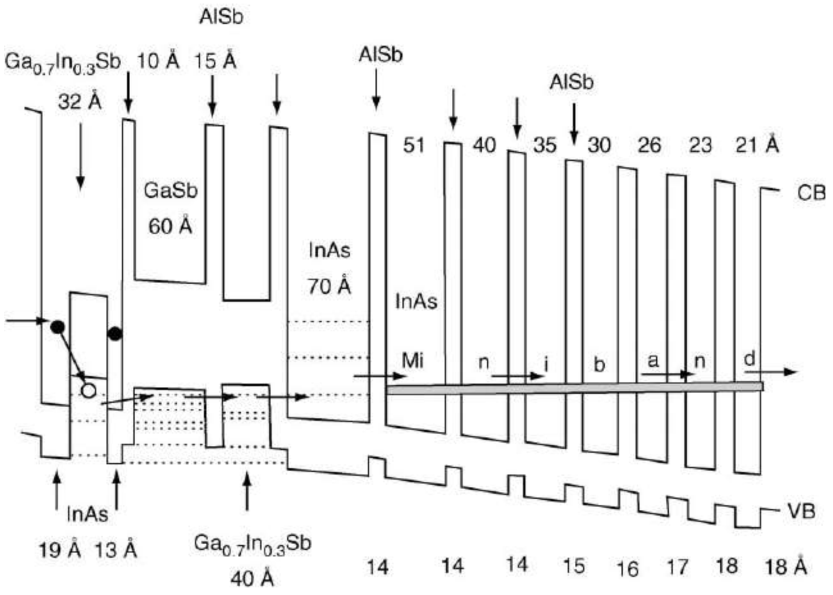

Figure 8 Conduction and valence band profiles for one period of an interband cascade laser with a W active region and three hole wells. Sub-band energy levels are also indicated.

at 80 K was recently reported. While substantial further improvements are expected, the primary disadvantage of the ICL is its complexity.

Despite numerous dramatic advances by type-II mid-IR lasers since their first demonstration in 1995, the performance has not yet progressed to the level that will be required to support a robust commercial technology. One key issue is the cw operating temperature, which must increase to at least the thermo-electric cooler range if the potential market is to extend beyond niche applications. Higher-temperature operation will result automatically once the number of QWs, and hence the threshold current density, can be reduced. Further suppression of the Auger decay rate would also be highly beneficial if it is proven possible.

The laser output power and efficiency are governed by the internal loss, which in type-II QWs tends to be dominated by resonant free-hole intersub-band transitions. The details of the valence sub-band structure are not well known at present, although recent experiments have apparently identified a number of the transition energies. As in the case of the threshold, reducing the number of QWs is highly beneficial since it reduces the number of free carriers that can absorb the lasing photons. The optically pumped OPIC and IA structures discussed above achieved dramatic reductions in the internal losses by following that strategy.

The beam quality attainable from a semiconductor laser is governed by the linewidth enhancement factor (LEF), which is proportional to the variation in the refractive index induced by a small change in the carrier density. A large positive value of the LEF leads to filamentation in broad-area lasers as tiny peaks in the optical intensity cause carrier depletion and a local increase in the refractive index, which in turn stimulates further focusing of the output beam. The output then consists of an incoherent superposition of many rapidly diverging filaments, resulting in a beam quality that is several orders of magnitude worse than the diffraction limit. While theoretical LEFs for optimized type-II W structures can be in the 1.5–2 range even at room temperature, recent Hakki–Paoli measurements have yielded values in the 2.5–5 range. This is possibly due to additional inhomogeneous broadening of the gain spectrum induced by interface roughness.

See also

Semiconductor Materials: Band Structure Engineering; Quantum Dots. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Mailhiot C and Smith DL (1989) Long-wavelength infrared detectors based on strained InAs-GaInSb type-II superlattices. *Journal of Vacuum Science and Technology* A7: 445–449.

Meyer JR, Malin JI, Vurgaftman I, Hoffman CA and Ram-Mohan LR (1997) Antimonide-based quantum heterostructure devices. In: Manasreh MO (ed.) *Strained*

Quantum Wells and Their Applications, pp. 235–272. Amsterdam, The Netherlands: Gordon and Breach.

Olafsen LJ, Vurgaftman I and Meyer JR (in press) Antimonide mid-IR lasers. In: Choi HK (ed.) *Long-Wavelength Semiconductor Lasers*. New York: Wiley.

Ram-Mohan LR, Vurgaftman I, Meyer JR and Dossa D (2000) Wavefunction engineering: a new paradigm in quantum nanostructure modeling. In: Nalwa HS (ed.) *Handbook of Nanostructured Materials and Nanotechnology*, vol. 2, pp. 707–739. New York: Academic Press.

Razeghi M (1998) Current status and future trends of infrared detectors. *Optoelectronics Review* 6: 155–194.

Vurgaftman I, Felix CL, Aifer EH and Meyer JR (2000) Interband III–V mid-IR semiconductor lasers. In: Perera AGU and Liu HU (eds) *Handbook of Thin Film Devices*, vol. 2, pp. 171–193. San Diego, CA: Academic Press.

Vurgaftman I, Felix CL, Bewley WW, *et al.* (2001) Mid-IR "W" lasers. *Philosophical Transactions of the Royal Society A* 359: 489–504.

Vurgaftman I, Meyer JR and Ram-Mohan LR (2001) Band structure parameters for III–V compound semiconductors and their alloys. *Applied Physics Reviews, Journal of Applied Physics* 89: 5815–5875.

Wilson BA (1988) Carrier dynamics and recombination mechanisms in staggered-alignment heterostructures. *IEEE Journal of Quantum Electronics* 24: 1763–1777.

Yang RQ, Bradshaw JL, Bruno JD, Pham JT and Wortman DE (2002) Mid-IR type-II interband cascade lasers. *IEEE Journal of Quantum Electronics* (Feature Section on Mid-IR Quantum Cascade Lasers) 38: 559–568.

SEMICONDUCTOR PHYSICS

Contents

Outline of Basic Electronic Properties
Band Structure and Optical Properties
Excitons
Impurities and Defects
Infrared Lattice Properties
Light Scattering
Polarons
Quantum Wells and GaAs-Based Structures
Recombination Processes
Spin Transport and Relaxation in Semiconductors; Spintronics
Surface Photovoltage Spectroscopy of Semiconductors

Outline of Basic Electronic Properties

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

We briefly introduce the main concepts underlying the detailed treatment of specialist aspects of semiconductor physics in ensuing sections, viz: the free electron model, density of states, Fermi-Dirac statistics, energy bands and energy gaps, effective mass, electrons and holes, intrinsic and extrinsic conductivity, direct and indirect interband optical transitions, excitons and low-dimensional systems.

Introduction

A proper treatment of optoelectronic devices needs to include the physical properties of the materials from which the devices are made, and this subject is covered under the section of the Encyclopedia entitled 'Semiconductor Fundamentals'. This includes sections on the basic physics common to all semiconductors useful in optoelectronic applications, the experimental techniques used to investigate the optical properties of these semiconductors and the more detailed properties of the most widely used semiconductors, and semiconductor systems. The properties of greatest importance here are those of electrons in crystalline solids, but the interaction between the electrons and the atoms of the crystal lattice is also significant. The combination of these microscopic effects is responsible for determining the macroscopic optical and electrical properties of semiconductors, and in this section we give the necessary background to these topics.

The classical treatment is extremely useful in describing the electrical conductivity of solids; the electrons are regarded as charged particles obeying Newton's laws of motion, and electrical resistance is assumed to arise from collisions with the vibrating lattice and/or impurities. However, the classical theory does not give a method of predicting the natural frequencies, or energies, of these electronic oscillators. To do this we need to use quantum mechanics, and the elements of this are reviewed here. We start with the free electron theory of metals, where the electrons are treated as particles constrained to move in a potential box. For convenience we begin by considering the situation in one dimension (1D).

Free Electron Theory of Metals

Free electrons in a 1D potential box of length L in the x-direction are described by a plane wave function that is normalized within the box, $\psi = (1/L)e^{ikx}$, and satisfies the Schrödinger equation in the x-direction:

$$\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + E\psi = 0 \qquad [1]$$

where E is the electron energy. This gives the familiar parabolic $E-k$ dispersion relation (see Figure 1), $E = \hbar^2k^2/2m$ (with electron momentum $p = \hbar k$). In order to evaluate physical problems it is necessary to define the boundary conditions for the wave function. We assume fixed boundary conditions (i.e., $\psi = 0$ at $x = 0, L$) leading to the quantization condition $k = n\pi/L$ where n is a positive integer.

Density of States

Physical properties of solids depend on the number of allowed k-values per unit range of k (or E-values per unit range of E), a quantity referred to as the density of states, D(k) or D(E) – for example, the number of allowed states between k and $k + dk = D(k)dk$. From the fixed boundary condition ($k = n\pi/L$, the number of allowed values of k per unit range of k in 1D is just $D(k) = L/\pi$. In terms of energy, using the $E-k$ dispersion relation, we can write (including the factor 2 for two possible spin states):

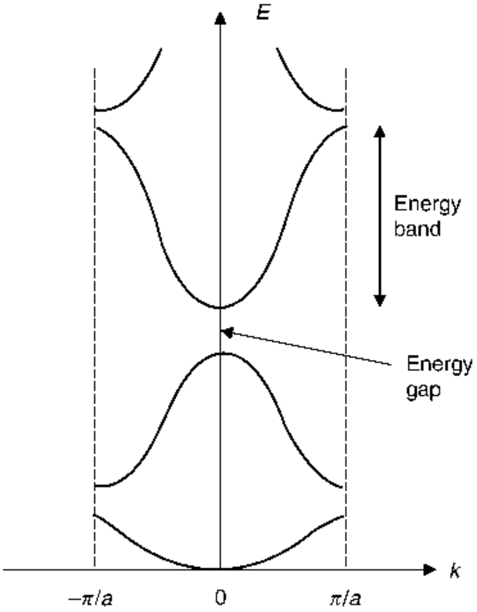

Figure 1 Schematic representation of energy bands in one dimension, characterized by gaps at the high symmetry points of the Brillouin zone.

$$\begin{aligned} D(E)dE &= D(k)(dk/dE)dk \\ &= (L/2\pi)(2m/\hbar^2)^{1/2}E^{-1/2}dE \qquad [2] \end{aligned}$$

i.e., the density of states depends on $E^{-1/2}$ for the 1D situation.

For a cube of side L, in three dimensions we generalize the identity for fixed boundary conditions:

$$\frac{L^2k^2}{\pi^2} = n_x^2 + n_y^2 + n_z^2 = R^2 \qquad [3]$$

where $k_x = n_x\pi/L$ etc., and the states fill up the positive quadrant of a sphere of radius ($R = Lk/\pi$) in k-space. Thus, the number of states in a given volume of k-space is numerically equal to that volume, so that the number of states between R and $R + dR$ in the positive quadrant (hence 1/8th) of a spherical shell of radius R in k-space is

$$D(R)dR = 4\pi R^2 dR(1/8) \qquad [4]$$

i.e., (1/8)th of the surface area of the shell multiplied by the thickness of the element. Changing the variable to k, the number of states between k and $k + dk$ is

$$D(k)dk = D(R)\frac{dR}{dk}dk = \frac{V}{2\pi^2}2k^2dk \qquad [5]$$

where the volume $V = L^3$, and a factor of 2 has again been included to account for the two possible spin states of the electron. This in turn, through the relation between E and k, can be expressed in terms of the density of states per unit energy so that the number of states between E and $E + \mathrm{d}E$ is given by:

$$D(E)\mathrm{d}E = D(k)\frac{\mathrm{d}E}{\mathrm{d}k}\mathrm{d}E = \frac{2V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}\mathrm{d}E \quad [6]$$

In two dimensions, by the same argument:

$$\frac{L^2k^2}{(\pi)^2} = n_x^2 + n_y^2 = R^2 \quad [7]$$

Therefore $D(R)\mathrm{d}R = (2\pi R/4)\,\mathrm{d}R$, where the factor 1/4 implies the positive quarter of the 2D ring between R and $R + \mathrm{d}R$ in k-space, $A = L^2$ and

$$D(k)\mathrm{d}k = D(R)\frac{\mathrm{d}R}{\mathrm{d}k}\mathrm{d}k = \frac{A}{\pi}k\mathrm{d}k \quad [8]$$

From the $E-k$ relation we have the number of states between E and $E + \mathrm{d}E$ as:

$$D(E)\mathrm{d}E = D(k)\frac{\mathrm{d}E}{\mathrm{d}k}\mathrm{d}E = \frac{A}{\pi}\left(\frac{m}{\hbar^2}\right)\mathrm{d}E \quad [9]$$

Thus, in 2D systems we have the important result that $D(E)$ is a constant.

Fermi-Dirac Distribution of Electrons

At zero temperature an atom exists in what is known as the ground state. All the electrons rest in the lowest energy states allowed to them by the Pauli principle, and fill the allowed states up to an energy known as the Fermi energy, E_F. Above absolute zero of temperature this is no longer the case, and thermal energy allows some electrons to exist in higher energy states. The probability $f(E)$ that a state of energy E will be occupied for an ideal electron gas in thermal equilibrium (thermal energy kT) is governed by Fermi-Dirac statistics:

$$f(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_\mathrm{F}}{kT}\right)} \quad [10]$$

For a metal at room temperature the difference between the Fermi energy and the energy of the conduction band minimum is much greater than the thermal energy, $(E_\mathrm{F} - E_\mathrm{co}) \gg kT$; this is also true for a strongly n-type semiconductor (see below) and the distribution is said to be degenerate.

Energy Bands and Energy Gaps: Semiconductors

In the presence of the (perturbing) periodic crystal potential, $V(r)$, the above simplicity is lost: this is the so-called ‘nearly free electron model’, and the perturbed Schrödinger equation is now:

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + V(x)\psi = E\psi \quad [11]$$

for the x-direction. This topic is discussed in all undergraduate texts on solid state physics, but we list the main results of the model as follows:

1. The electrons that we are concerned with are confined within a periodic lattice. Consequently the wave function ψ is modulated with the periodicity of the lattice. By symmetry arguments the correct wave function can be shown to be:

   $$\psi_k(x) = \frac{1}{\sqrt{L}}\mathrm{e}^{ikx}u_k(x) \quad [12]$$

   where $u_k(x)$ is a function that has the symmetry of the lattice. $\psi_k(x)$ is known as the Bloch function and is a satisfactory solution to the Schrödinger equation. Because of the periodic nature of the Bloch equation, it is always possible to select the wavevector index k of any Bloch function to lie within the range $\pi/\mathrm{a} < k < \pi/a$. This region is referred to as the first Brillouin zone, and the representation where the $E-k$ dispersion is mapped back entirely into the first Brillouin zone is called the ‘reduced zone scheme’ (Figure 1).
2. Regions of E appear where solutions $\psi_k(x)$ are not allowed, corresponding to Bragg reflections of the electron waves by the lattice; these are the so-called energy gaps (forbidden energy regions) which separate energy bands (allowed energy regions). In 3D these discrete bands of energy form the conduction and valence bands within materials. The forbidden energy gap separates these bands. An additional quantum number n is required to label the individual energy bands corresponding to successive Brillouin zones in the wavevector, $2n\pi/a$ (where $n = 0, 1\ldots$). The position of the Fermi level determines the electron population of the energy bands, and the character of the solid (see below).
3. The parabolic $E-k$ dispersion relation is retained near $k = 0$, and at other symmetry points of the Brillouin zone, but with the substitution of an effective mass, m^* (see below) for the free electron mass, i.e., $E = \hbar^2k^2/2m^*$ for $k \ll \pi/a$. This is referred to as the effective mass approximation.

The nearly free electron model is one limit for the description of the formation of energy bands and energy gaps in a solid. In real crystals the energy bands derive from, and retain the symmetry of, the parent atomic states of the host atoms. In this limit, the model that starts out from the wave functions of the free atoms is known as the tight binding model or the LCAO (linear combination of atomic orbitals) approximation. As neutral separated atoms are brought together to form a crystal, their wave functions overlap and the Coulomb interaction between the atom cores and the electrons splits the atomic energy levels, spreading them into bands. Each state of given quantum number of the free atom is spread in the crystal into a band of energies; a given s, p, d-state in N separate atoms will produce the corresponding N-orbital s, p, d-energy band in the solid. Including the factor of 2 for spin, a full s-band would contain 2N electrons, a full p-band 6N electrons and so on. The width of the band is proportional to the strength of the overlap interaction between neighboring atoms. Each energy band will not have the same energy as any other band over any substantial range of the wavevector.

Metals, Insulators and Semiconductors

The position of the Fermi level distinguishes whether the solid is a metal, insulator or semiconductor. Generally those bands that derive from full atomic orbital states will be fully occupied with electrons, those that derive from empty atomic states lead to empty energy bands, and those from partially full atomic states lead to partially full energy bands. The highest band of those that are filled with electrons is called the valence band; the next highest band is called the conduction band (the fundamental energy gap, E_G, is the valence/conduction band separation, referred to from now on as simply the energy gap). For a metal, E_F is in the conduction band at all temperatures; for an insulator, E_F is close to the middle of the energy gap. Semiconductors are a class of crystals with small energy gaps; they are insulators at $T = 0\,K$, but are conductors at finite temperatures due to either thermal excitation of electrons across the gap or doping, n- or p-type (see below).

Effective Mass

Consider what happens to an electron when an external electric field F is applied. Suppose the electron is initially in a state k. When the field has acted on the electron for a small time dt, it has gained an energy:

$$\mathrm{d}E = \mathrm{e}Fv\mathrm{d}t = (\mathrm{e}F/\hbar)(\mathrm{d}E/\mathrm{d}k)\mathrm{d}t \qquad [13]$$

since the group velocity of the de Broglie wave (energy, $E = \hbar\omega$) is $v = \hbar^{-1}(\mathrm{d}E/\mathrm{d}k)$. Or simplifying we have: $\mathrm{d}k/\mathrm{d}t = \mathrm{e}F/\hbar$. This gives for the resulting acceleration:

$$\begin{aligned} a &= \mathrm{d}v/\mathrm{d}t = (1/\hbar)(\mathrm{d}^2k/\mathrm{d}t^2)(\mathrm{d}k/\mathrm{d}t) \\ &= (\mathrm{e}F/\hbar^2)(\mathrm{d}^2E/\mathrm{d}k^2) \end{aligned} \qquad [14]$$

Comparing this result for the acceleration of a free electron of mass m under the influence of F, $a = \mathrm{e}F/m$, it follows that the electron behaves as if it had an effective mass, m^*, given by

$$\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{\mathrm{d}^2E}{\mathrm{d}k^2} \qquad [15]$$

Thus, the electron effective mass m^* gives a measure of the conduction band curvature near the minimum at $k = 0$. We note that the same result is achieved by differentiating twice the 'nearly free electron' E–k relation, $E = \hbar^2k^2/2m^*$.

Electrons and Holes

Consider the situation where a single electron is excited either optically or thermally from the top of the filled valence band to the bottom of the conduction band of a 1D lattice, and consider the collective behavior of the valence band when an external electric field is applied. Denoting the charge of an electron by $-$e and the velocities of the electrons by v_i, we may write for the current associated with all electrons in the completely filled valence band in the absence of an external field:

$$\mathrm{I} = -\mathrm{e}\sum_i v_i = -\mathrm{e}[v_j + \sum_{i\neq j} v_i] = 0 \qquad [16]$$

Thus, if the electron j were missing, we should have current

$$I' = -\mathrm{e}\sum_{i\neq j} v_i = \mathrm{e}v_j \qquad [17]$$

Applying an external field F, the rate of change of the current due to the field is:

$$\mathrm{d}I'/\mathrm{d}t = \mathrm{e}(\mathrm{d}v_j/\mathrm{d}t) = -\mathrm{e}^2F/m_j^* \qquad [18]$$

Since the holes are at the top of the valence band, m_j^* is negative and the right hand side of the equation is positive – i.e., a positive external field F gives a positive change of current at the top of the valence band. Thus, a band in which an electron is missing behaves as a 'positive hole' with an effective mass $m_h = -m_e$, and positive charge +e. Since m_e is negative near the top of the valence band, m_h is positive.

Impurities, Intrinsic and Extrinsic Conductivity

The key point about semiconductors from a device point of view is that one can control their conductivity. The basic method is by doping the host crystal with impurities either of valence greater than the host (donors giving excess electrons, called n-type) or of valence less than the host (acceptors giving excess holes, called p-type). The conductivity is referred to as intrinsic conductivity (from intrinsic carrier concentration, N_i) and extrinsic conductivity (from extrinsic carrier concentration, N_e) for undoped and doped material, respectively.

The impurities are chosen to have very low binding energies, so that the electrons (holes) are substantially ionized at room temperature to provide n-type (or p-type) conductivity. These shallow impurities can be treated using the simple Bohr theory of the hydrogen atom, substituting the electron orbiting around the proton in the hydrogen atom under the Coulomb potential e/r, by the electron in the crystal (electron effective mass m^*) orbiting round the ionized (donor/acceptor) impurity atom under the Coulomb potential $e/\varepsilon r$ in the crystal – where ε is the static dielectric constant of the crystal. As with the hydrogen atom, this gives a set of discrete energy levels:

$$E = \frac{m^*/m_o}{\varepsilon^2} \frac{13.6}{n^2} \text{eV} \qquad [19]$$

where 13.6 eV is the binding energy of the H-atom and n is the principal (Bohr) quantum number. The binding energy is then

$$E_B = \frac{m^*/m_o}{\varepsilon^2} \times 13.6 \text{ eV} \qquad [20]$$

For typical semiconductors (Ge, Si...) we have $m^*/m_o \sim 0.05$, $\varepsilon \sim 16$. Thus, E_B is in the range of a few meV (i.e., much less than E_G, and kT at room temperature). The energy level scheme for a doped semiconductor is shown schematically in real space (at room temperature) in Figure 2.

The total conductivity will be $\sigma = \sigma_i + \sigma_e$, where the intrinsic conductivity $\sigma_i = n_i e^2 \tau / m^*$ and the extrinsic conductivity $\sigma_e = n_e e^2 \tau / m^*$. τ is the so-called carrier scattering time. The intrinsic carrier concentration is related to temperature by $n_i = n_0 \times \exp(-E_G/2kT)$, where n_0 is a constant. In the extrinsic regime ($n_e \gg n_i$), at temperatures such that $kT \gg E_d$ (so that the impurities are all ionized), the conduction (valence) carrier concentration $n_e \to n_d$ (or n_a), the donor (or acceptor) concentration.

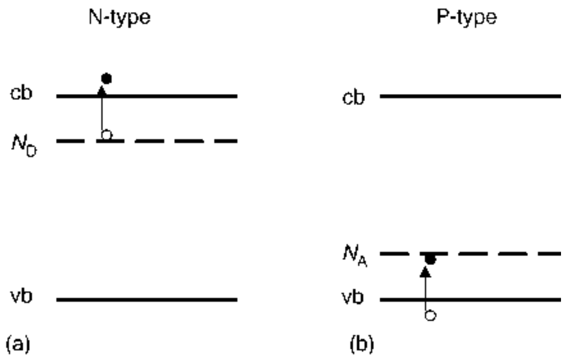

Figure 2 Doping in semiconductors. In n-type material (a), carriers are excited thermally from the donor states to the conduction band where they can contribute to the conductivity. In p-type material (b), electrons are effectively 'captured' by the acceptor states, leaving behind holes in the valence band which can also contribute to conduction. N_D and N_A are the concentration of donor and acceptor impurities, respectively.

Infrared Absorption – Interband Optical Transitions, Excitons

Apart from thermal promotion of electrons and holes via interband (valence to conduction band) transitions responsible for intrinsic conductivity, it is possible to directly excite such transitions optically by subjecting the sample to radiation with $\hbar\omega > E_G$. If the lowest conduction band minimum is situated vertically above the valence band maximum (usually at $k = 0$), then the optical transition process is referred to as a direct transition. If the lowest conduction band minimum is situated at some other point in k-space than that of the valence band maximum, the process is referred to as an indirect transition; this requires the participation of optical phonons (i.e., quantized lattice vibrations) to satisfy momentum conservation, and is therefore a weaker (higher order) process. The semiconductors are respectively referred to as Direct Gap and Indirect Gap Semiconductors. Simplified energy band diagrams for direct and indirect transitions are shown in Figure 3.

The optical spectrum corresponding to the onset of interband electric dipole transitions is referred to as the absorption edge. The most widely used indirect gap materials are the group IV semiconductors Si and Ge. The best known direct gap materials are the III–V's (InSb, GaAs, GaP, etc.), II–VI's (HgTe, CdTe, ZnSe, etc.) and IV–VI's (lead salts PbTe, PnSe, SnTe) and alloy combinations of all these. An important additional effect, which becomes progressively stronger as the energy gap becomes larger in a semiconductor is the Coulomb interaction between the electrons and holes created by interband optical transitions. The electron is raised into an excited state but remains bound in a hydrogen-like orbit around the positive hole creating a two particle excitation called an exciton. The composite particle can then move throughout the crystal. In an ideal crystal, therefore, the absorption spectrum should consist of a series of discrete lines corresponding to the ground and excited exciton states, followed at somewhat high energies by the continuum associated with interband electronic transitions. The energy levels of the exciton spectrum measured from the beginning of the ionization continuum will be given by the Bohr formula:

$$E_{ex} = \frac{\mu}{\varepsilon^2}\frac{13.6}{n^2}\mathrm{eV} \qquad [21]$$

where μ is the reduced mass given by: $1/\mu = 1/m_e + 1/m_h$ (m_e, m_h are electron and hole effective masses). A schematic diagram of the absorption edge is shown in Figure 4 (dashed line – in the absence of excitons; solid line – in the presence of excitons).

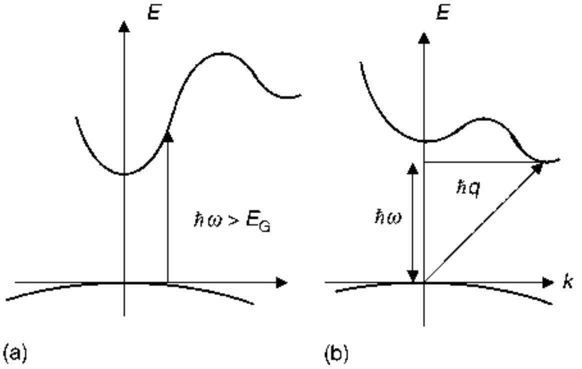

Figure 3 (a) Direct and (b) indirect gap semiconductors. Direct gap transitions involve only the absorption or emission of a single photon and are therefore very nearly 'vertical' transitions in *k*-space. On the other hand, indirect transitions have a 'two-stage' nature, involving the participation of both a photon and an optical phonon that carries sufficient momentum $\hbar q$ to allow the conduction band electron created in the absorption process to have a large non-zero wavevector.

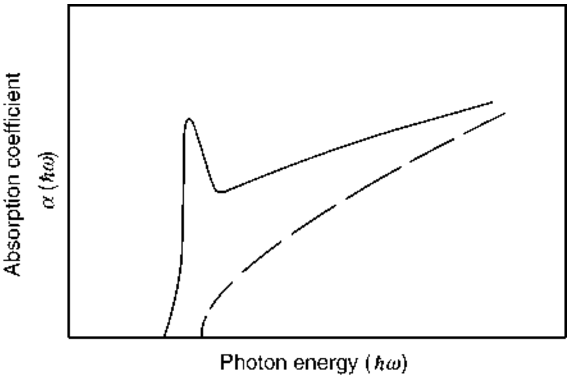

Figure 4 Absorption edge spectra without (dashed curve) and with (solid curve) the excitonic effect.

Conduction and Valence Band Parameters in Typical Semiconductors

As has been described earlier in real semiconductors the conduction and valence bands retain the symmetry of the parent atomic states from which they derive. For example, the conduction and valence bands of silicon can be regarded as deriving from 3s and 3p electrons, and it is generally true that for all group IV, III–V and II–VI semiconductors commonly used in electronic applications the conduction band has s-like symmetry and the valence band has p-like symmetry. This means that the valence band has the additional complication that it is triply degenerate (c.f. the p_x, p_y, p_z components of the hydrogen atom p-orbital); and further, in the presence of spin-orbit coupling (familiar from atomic physics) part of this degeneracy is raised, resulting in doubly degenerate (at $k = 0$) so-called heavy hole (hh) and light hole (lh) bands, and a band split-off by the spin-orbit splitting, Δ:

$$E_{hh} = -\hbar^2k^2/2m_{hh}\ ;\ E_{lh} = -\hbar^2k^2/2m_{lh}\ ;$$
$$E_{so} = -\Delta - \hbar^2k^2/2m_{so} \qquad [22]$$

The resulting conduction and valence bands near $k = 0$ are shown in Figure 5.

Low Dimensional Systems: Quantum Wells

So far we have been concerned almost entirely with the bulk properties of crystalline solids. Such properties can normally be specified in terms of coefficients that are independent of the shape and size of a particular specimen. Thus, for example the electrical

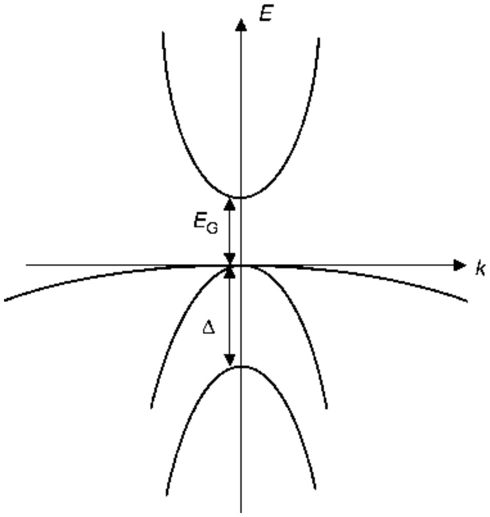

Figure 5 Energy bands (s-like conduction and p-like valence bands) of a direct gap semiconductor near the Brillouin zone center.

conductivity and resistivity, or the optical absorption coefficient and refractive index are properties only of the material and not of the sample dimensions. When one or more dimensions of a solid are reduced sufficiently, the properties are no longer given by these bulk coefficients. The sample is then described as a low-dimensional system (LDS). Such systems form the basis of many optoelectronic devices. LD systems can be classified according to the number of dimensions that are small: thin films (and heterojunctions) are two-dimensional since only the thickness is reduced, and these form so-called quantum wells; fine wires are one-dimensional since two dimensions are reduced, forming quantum wires; dots or specks are zero-dimensional, forming quantum dots.

From an optoelectronic point of view, the most important example is that of electrons confined to a semiconductor film or junction; this system is called the two-dimensional electron gas (2-DEG) and the structure a quantum well. Thus, it is possible to create a narrow electric potential junction region so that the electrons are restricted in 1D (say z) according to the quantum mechanics of a particle in an infinite potential well (described earlier), but 'free' within the xy junction plane.

See also

Semiconductor Materials: Band Structure Engineering; Quantum Dots. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Infrared Lattice Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Hook JR and Hall HE (1991) *Solid State Physics*, 2nd edn. Chichester, UK: John Wiley and Sons, Inc.

Kittel C (1996) *Introduction to Solid State Physics*, 7th edn. Chichester, UK: John Wiley and Sons, Inc.

Seeger K (1989) *Semiconductor Physics: an Introduction*, 4th edn. Berlin: Springer.

Band Structure and Optical Properties

W Zawadzki, Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The purpose of this article is to describe the influence of the electronic energy band structure of semiconductors on their optical properties. By the band structure one understands the energy–wavevector dispersion relation $\varepsilon(\mathbf{k})$ for the electron energy bands and the relative positions of band maxima and minima in the Brillouin zone.

The main problem in calculating the band structure of a semiconductor is that of the periodic potential of the lattice in which the electrons move. It is, in fact, a self-consistent problem since the moving electrons partly screen the potential. Different approximations have been developed to deal with the question, and in all of them the symmetry of the lattice is of fundamental importance. Thus, in the so-called empty-lattice approximation one deals exclusively with the symmetry and periodicity of the lattice without specifying the potential. This can give qualitative ordering and symmetry of the bands but no quantitative results. In the opposite limit one uses the tight binding approximation, in which the bands are constructed from the atomic states of separate atoms. This method gives quite a good description of lower (valence) bands but poor results for higher (conduction) bands. A powerful way to describe real energy bands is obtained by various forms of pseudopotential calculation, where one approximates the potential near actual atoms by simple parametrized forms and then adjusts the parameters to fit experimental (mostly optical) data. The pseudopotential methods give a good overall description of various bands in the entire Brillouin zone, but they do not provide sufficiently precise results near the band extrema. A semi-empirical way to serve the latter purpose is called $\mathbf{k}\cdot\mathbf{p}$ theory which we describe below. Various methods of band structure calculation are reviewed in books on solid state physics as outlined in the suggestions for Further Reading at the end of this article.

In the majority of important semiconductor materials (Si, Ge, many III–V and II–VI compounds) the maxima of the valence bands are at the center of the Brillouin zone (Γ point). The minima of the conduction bands in Si and Ge are not at the Γ point. This means that the fundamental optical interband transitions in these materials are indirect (in the ε–$\mathbf{k}$ diagram), i.e., they require phonon assistance. On the other hand, the minima of conduction bands in important III–V and II–VI compounds are at the Γ point, so that the interband optical transitions are direct and they do not require phonon assistance.

Both systems are utilized in opto-electronic devices, particularly for detectors. However, for emitters (light-emitting diodes and lasers) we are mostly concerned with the second case, on which we concentrate here.

Bloch States

The Bloch theorem states that if the potential $V(\mathbf{r})$ in which the electron moves is periodic with the periodicity of the lattice, then the solutions $\Psi(\mathbf{r})$ of the Schrödinger wave equation

$$\left[\frac{\mathbf{p}^2}{2m_0}+V(\mathbf{r})\right]\Psi(\mathbf{r})=\varepsilon\Psi(\mathbf{r}) \qquad [1]$$

are of the form $\Psi(\mathbf{r})=\exp(i\mathbf{k}\cdot\mathbf{r})u_k(\mathbf{r})$, where $u_\mathrm{k}(\mathbf{r})$ is periodic with the periodicity of the direct lattice, and $\mathbf{k}$ is the wavevector ($\hbar\mathbf{k}$ is the pseudomomentum). The proof of this theorem can be found, for example, in undergraduate textbooks on solid state physics.

k·p Theory

The underlying idea of semi-empirical **k·p** theory is to describe the energy bands in the vicinity of a given point of the Brillouin zone (usually near a band extremum). Symmetry properties are used to minimize the number of unknown band parameters (effective mass, energy gap, etc.) which are then determined by experiment. The initial Schrödinger equation for an electron in the periodic potential $V(\mathbf{r})$ of the crystal lattice reads

$$[\mathbf{p}^2/2m_0+V(r)+H_{\mathrm{so}}]\Psi=\varepsilon\Psi \qquad [2]$$

where m_0 is the free electron mass and H_{so} is the spin–orbit interaction. Since $H_{\mathrm{so}}(\mathbf{r})$ is also periodic with the lattice periodicity, the solutions of eqn [2] are given by the Bloch states above.

We use **k·p** perturbation theory, as follows, to describe the energy bands near $\mathbf{k}=0$ (i.e., the so-called Γ point of the Brillouin zone). Since we are interested in small values of $\mathbf{k}$, we expand our unknown cell periodic functions, $u_{l\mathbf{k}}(\mathbf{r})$, in terms of the set of Γ-point functions, $u_{lo}(\mathbf{r})$, whose symmetry we know. One defines a representation $\Phi_{l\mathbf{k}}=\exp(i\mathbf{k}\cdot\mathbf{r})u_{lo}(r)$, where $u_{lo}(\mathbf{r})$ is the periodic $\mathbf{k}$-independent function satisfying the equation

$$[\mathbf{p}^2/2m_0+V+\mathrm{H}_{\mathrm{so}}]u_{lo}=\varepsilon_{lo}u_{lo} \qquad [3]$$

whose solution we know in terms of the band-edge energies, ε_{lo}. This representation is sometimes referred to as the Luttinger and Kohn (LK) representation. The u_{lo} functions are in general mixtures of spin up and spin down states because of the spin–orbit interaction. One looks for the solutions of eqn [2] in the form

$$\Phi_{n\mathbf{k}}=\exp(i\mathbf{k}\cdot\mathbf{r})\sum_l c_l^{(n)}(\mathbf{k})u_{lo}(\mathbf{r}) \qquad [4]$$

where the sum is over all bands and $c_l^{(n)}(\mathbf{k})$ are the coefficients to be determined. Inserting the above form into eqn [2], performing the operation $\mathbf{p}^2$ (i.e., operating twice on the Bloch product function with the operator $\mathrm{p}=-i\hbar\nabla$) and using the property [3] one eliminates the unknown potential $V(\mathbf{r})$. Multiplying the resulting equation from the left by $u^*_{l'o}$, integrating over the unit cell Ω and using the orthonormality of u_{lo}, one finally obtains

$$\sum_l\left[(\varepsilon_{lo}-\varepsilon')\delta_{l'l}+\frac{\hbar}{m_0}\mathbf{k}\cdot\mathbf{p}_{l'l}\right]c_l^{(n)}(\mathbf{k})=0 \qquad [5]$$

for $l'=1,2,3,\ldots$ Here $\delta_{l'l}$ is the Kronecker delta function, $\varepsilon'=\varepsilon-\hbar^2k^2/2m_0$, and $\mathbf{p}_{l'l}$ is the so-called matrix element of the momentum given by $\mathbf{p}_{l'l}=(1/\Omega)\int_\Omega u^*_{l'o}\mathbf{p}u_{lo}\,\mathrm{d}\mathbf{r}$. The index l' runs over all energy bands.

Equation [5] formulates the famous **k·p** theory. The nondiagonal part $\mathbf{k}\cdot\mathbf{p}_{l'l}$ can be eliminated by the perturbation procedure, and the method is referred to as **k·p** perturbation theory. In the second order of perturbation the bands are separated and in each band the carriers are described by a dispersion relation $\varepsilon_l=(\hbar^2/2)\mathbf{k}(\widehat{1/m^*_{lo}})\mathbf{k}$ where $(\widehat{1/m^*_{lo}})$ is an inverse effective mass tensor at the band edge in question. The inverse mass is generally a 3×3 tensor, but for cubic materials in the center of the Brillouin zone it is a scalar, so that $\varepsilon_l(\mathrm{k})=\hbar^2\mathrm{k}^2/2\mathrm{m}^*_{lo}$ (where $1/\mathrm{m}^*_{lo}\equiv 1/\mathrm{m}_0+(2/\mathrm{m}_0^2)\sum_{l\neq l'}(\mathrm{p}^2_{ll'})/[\varepsilon_{l0}-\varepsilon_{l'0}]$). We then say that the band is spherical and parabolic. The second order of perturbation is a good approximation if the energy ε (as counted from a nondegenerate band edge) is small compared to forbidden energy gaps. For the triply degenerate p-like valence band one has to do degenerate perturbation theory and treat the bands together as a 3×3 matrix equation.

In the same approximation the wavefunction for a given band is: $\Psi_{l\mathbf{k}}(\mathbf{r})=\exp(i\mathbf{k}\cdot\mathbf{r})\,u_{lo}(\mathbf{r})$. In the absence of spin–orbit coupling, H_{so}, the u_{lo} function would just have the symmetry of the parent band (i.e., labeled S for the s-like conduction band, and X, Y or Z for the triply degenerate p-like valence band – each with a single spin up or spin down component). In the presence of H_{so}, the triply degenerate valence band states become mixtures of X, Y and Z with mixed spin character and part of the degeneracy is raised.

In atomic notation the presence of H_{so} has transformed the basis from the (l, s) to the (J, m_J) representation. This results in the well-known light and heavy hole bands (degenerate at $k = 0$) and the spin–orbit split-off band.

Under the influence of a radiation field, of vector potential $\mathbf{A}'$ and frequency ω, the optical transition probability for an electron to be raised from state Ψ_i (initial) to Ψ_f (final) is proportional to the square of $\mathrm{M} = (e/mc)\langle\Psi_f|\mathbf{A}'\cdot\mathbf{p}|\Psi_i\rangle$ i.e., it is determined by the same momentum matrix element, p_{fi}, which governs the effective mass. For interband transitions one immediately gets the selection rule $\Delta k = 0$ (i.e., direct transitions). The polarization selection rules are determined by the $u_{lo}(r)$ components of the initial and final states, through the momentum matrix elements.

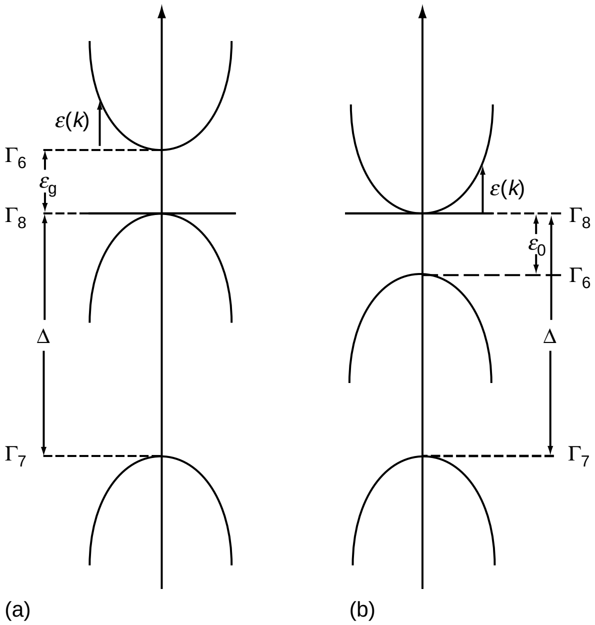

Figure 1 Three-level model of band structure near the Γ point of the Brillouin zone: (a) InSb-type semiconductors; (b) HgTe-type semiconductors.

Narrow-Gap Semiconductors

Semiconductors such as InSb and $\mathrm{Hg}_{1-x}\mathrm{Cd}_x\mathrm{Te}$ (with $x < 0.3$) have small energy gaps and are referred to as narrow gap semiconductors (NGSs). They are commonly used in opto-electronic applications which may be accessing energies in the conduction or valence band which are a significant fraction of the energy gap. Under these circumstances it is not valid to cut off the expansion to order k^2, and one has to deal with so-called nonparabolic bands. Thus, in NGSs (or indeed in any situation where the energy becomes an appreciable fraction of ε_g) the above procedure is not valid and an alternative approximation is in order. Following semidegenerate perturbation theory one includes exactly in eqn [5] a finite number of bands (near each other in energy) and neglects distant bands. This is referred to as the Kane model, and the energy bands near $k = 0$ are shown for InSb-type III–V compounds and for HgTe-type II–VI compounds in Figure 1. In this case we include only the conduction and valence band in our set of states (eqn [5]). Including spin and degeneracy of the Γ_8 symmetry the set [5] then has eight equations. They can be solved analytically and the resulting energies are given by

$$\varepsilon'(\varepsilon'+\varepsilon'_{\mathrm{g}})(\varepsilon'+\varepsilon'_{\mathrm{g}}+\Delta)-\left(\varepsilon'+\varepsilon'_{\mathrm{g}}+\frac{2\Delta}{3}\right)\kappa^2k^2=0 \qquad [6]$$

where $\varepsilon' = \varepsilon - \hbar^2k^2/2m_0$. We see that for this restricted set of states the only unknown parameters are the energy gap, ε_{g}, the spin–orbit splitting energy, Δ, and $\kappa = -\mathrm{i}(\hbar/m_0)\langle S|p_z|Z\rangle$ (note that this is the only non-vanishing matrix element of the momentum). The fourth energy is $\varepsilon' = -\varepsilon_{\mathrm{g}}$. All four energies are spin degenerate. In NGSs one may neglect the free electron term, i.e., set $\varepsilon' \approx \varepsilon$. For $\varepsilon \ll \varepsilon_{\mathrm{g}} + (2/3)\Delta$ the resulting quadratic equation for the conduction and the light-hole bands is

$$\varepsilon(1+\varepsilon/\varepsilon_{\mathrm{g}}) = \hbar^2k^2/2m_0^* \qquad [7]$$

where $1/m_0^* = (4\kappa^2/3\hbar^2\varepsilon_{\mathrm{g}})(\varepsilon_{\mathrm{g}}+2\Delta/3)/(\varepsilon_{\mathrm{g}}+\Delta)$ defines the effective mass at the band edge. The root for the heavy holes is $\varepsilon = -\varepsilon_{\mathrm{g}}$ i.e., this band is not correctly described within the three-level model. The bands given by eqn [6] are spherical but nonparabolic. For $\varepsilon \ll \varepsilon_{\mathrm{g}}$ one recovers the standard dispersion, $\varepsilon = \hbar^2k^2/2m_0^*$.

For HgTe-type materials, setting the zero of energy at the Γ_8 edge and replacing ε_{g} by $-\varepsilon_0$ (cf. Figure 1), one obtains

$$\varepsilon'(\varepsilon'+\varepsilon_0)(\varepsilon'+\Delta)-\left(\varepsilon'+\frac{2\Delta}{3}\right)\kappa^2k^2 = 0 \qquad [8]$$

For $\varepsilon' \ll (2/3)\Delta$ the dispersion [7] is valid with $1/m_0^* = 4\kappa^2/3\hbar^2\varepsilon_0$. In this case the conduction and the heavy hole bands ($\varepsilon \equiv 0$) are degenerate at $k = 0$, i.e., the thermal gap is zero.

The important property of the above model is that it holds also in the limit of $\varepsilon_{\mathrm{g}} \to 0$ (i.e., for mixed $\mathrm{Hg}_{1-x}\mathrm{Cd}_x\mathrm{Te}$ crystals). The effective mass in eqn [7] tends to zero, but the dispersion described by eqn [6] is valid and it gives $\varepsilon = (2/3)\kappa k$, i.e., a linear band.

The electron and hole wavefunctions in NGSs resulting from semidegenerate perturbation theory are given by eqn [4], in which the sum runs over all the included states. Thus they involve truly **k**-dependent amplitudes of the Bloch states, cf. eqn [1]. In addition, these functions are mixtures of spin-up and spin-down states. These features have important consequences for optical and dc transport phenomena.

If the conduction band minimum is not at the Γ point (Ge and Si), there are at least two different matrix elements of momentum and the resulting band is ellipsoidal.

Effective Mass

The nonparabolic dispersions $\varepsilon(k)$ in NGSs require generalizations of important relations. The momentum mass $\widehat{m}^*$ is introduced as a connection between the psuedomomentum and the velocity: $\hbar\mathbf{k} = \widehat{m}^*\mathbf{v}$. Since $v_i = \delta\varepsilon/\delta(\hbar k_i)$, one obtains for a spherical band

$$\frac{1}{m^*} = \frac{1}{\hbar^2 k}\frac{\mathrm{d}\varepsilon}{\mathrm{d}k} \qquad [9]$$

The above mass is a scalar, but it depends on the energy (or momentum). This mass is more useful than the one usually defined in textbooks, relating force to acceleration. The latter is not a scalar even for a spherical band. However, for the standard parabolic band, $\varepsilon = \hbar^2k^2/2m_0^*$, both masses are the same and are equal to m_0^*. The mass [9] is related via velocity to the electric current, so that it enters naturally in transport phenomena (also free carrier optics). In particular, this mass enters the definition of mobility $\mu = q\tau/m^*$, where q is the charge and τ is the relaxation time. It enters the density of states (see below). Finally, it is the momentum mass [9] which is measured in cyclotron resonance.

For the dispersion [7] (the so-called two-level model), the mass [9] in the conduction band is

$$m^*(\varepsilon) = m_0^*\left(1 + \frac{2\varepsilon}{\varepsilon_{\mathrm{g}}}\right) \qquad [10]$$

Again, for $\varepsilon/\varepsilon_{\mathrm{g}} \ll 1$ the energy dependence of the mass is negligible, which represents the standard regime. Figures 2 and 3 illustrate energy dependences of the electron masses in InSb and HgSe, as measured by various optical and transport methods. The increase of the mass with energy is very strong in both cases.

Density of States

The density of states in energy space is calculated beginning with pseudomomentum space. For the spherical band one obtains

$$\rho(\varepsilon) = \frac{k^2}{\pi^2}\frac{\mathrm{d}k}{\mathrm{d}\varepsilon} = \frac{m^*k}{\pi^2\hbar^2} \qquad [11]$$

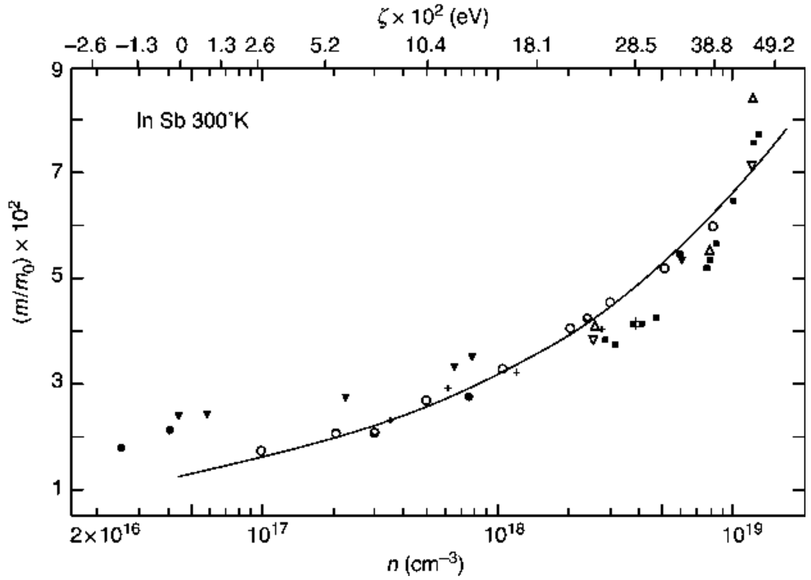

Figure 2 Electron effective mass in InSb at room temperature versus free electron concentration. The solid line, calculated for the two-level model, represents the mass value at the Fermi energy, as indicated on the upper abscissa. The experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

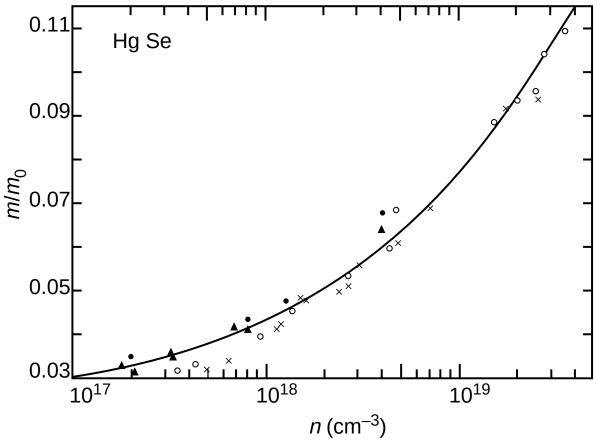

Figure 3 Electron effective mass in zero-gap material HgSe versus free electron concentration. The solid line is calculated for the three-level model. Experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

The momentum mass [9] enters naturally into this important quantity. For the parabolic case $m^* = \text{const.}$ and $k \sim \varepsilon^{1/2}$, so that the standard result is recovered. In the general case one should use eqn [10] for $m^*(\varepsilon)$ and determine k from eqns [6] or [7].

Electron–photon Interaction

The electron–photon interaction can be introduced into the theory on two levels. If one replaces in the initial equation [2] the momentum $\mathbf{p}$ by $\mathbf{p} + (e/c)\mathbf{A}'$, the interaction term is obtained in the scalar form $H_{\text{int}} = (e/m_0c)\mathbf{A}'\cdot\mathbf{p}$, where $\mathbf{A}'$ is the vector potential of radiation. The wavefunctions to be used for the calculation of matrix elements are of the form [4], i.e., they include the periodic LK amplitudes. In fact, the matrix elements of momentum for optical transitions are identical to those of the $\mathbf{k}\cdot\mathbf{p}$ theory, as pointed out earlier. Since (as noted above) the wavefunctions for all bands have the same form [4], differing only by the coefficients $c_l^{(n)}(\mathbf{k})$, the matrix elements for interband and intraband optical transitions have the same form.

The electron–photon interaction can also be introduced directly to the $\mathbf{k}\cdot\mathbf{p}$ theory of eqn [5] replacing $\hbar\mathbf{k}$ by $\hbar\mathbf{k} + (e/c)\mathbf{A}'$. If the free electron term $\hbar^2k^2/2m_0$ is neglected, the interaction matrix involving $\mathbf{A}'$ terms is a number matrix. In this procedure the wavefunctions for initial and final states are simply columns and rows of $c_l^{(n)}$ coefficients and LK amplitudes no longer come into play. Here the $\mathbf{p}_{l'l}$ elements of $\mathbf{k}\cdot\mathbf{p}$ theory determine directly the electron–photon interaction. Both procedures described above give the same results.

Quantum Wells

If charge carriers are placed in a quantum well described by a potential $U(z)$, the motion in the z-direction is quantized while the motion in the x–y plane remains free. One takes the same LK basis (cf. eqn [3]), but the general form of solution is given by

$$\Psi = \sum_l f_l(\boldsymbol{r})u_{l0}(\boldsymbol{r}) \qquad [12]$$

since, in the presence of an external potential, $\mathbf{k}$ is not a good quantum number and the envelope functions $f_l(\mathbf{r})$ are unknown. If $U(z)$ is a slowly varying function within the unit cell the potential appears only on the diagonal of the set [5] and $\hbar\mathbf{k}$ is replaced by $\mathbf{p}$. One obtains

$$\sum_l \left[\left(\varepsilon_{l0} + \frac{\mathbf{p}^2}{2m_0} + U - \varepsilon\right)\delta_{l'l} + \frac{\mathbf{p}_{l'l}\cdot\mathbf{p}}{m_0}\right] f_l(\mathbf{r}) = 0 \qquad [13]$$

One can now eliminate the nondiagonal terms applying the standard Luttinger and Kohn scheme of second-order perturbation theory and arrive at the decoupled band equations with effective masses. This results in the standard quantization due to electric (and magnetic, if present) fields, and wavefunctions are of the form $\Psi_{lm} = f_m u_{l0}$ for the m-th sub-band of the l-th band.

For NGSs this procedure is a poor approximation and one should proceed as before including exactly a finite number of bands (cf. eqn [12]). This results in a set of coupled differential equations for the envelope functions $f_l(\mathbf{r})$. In some important cases one can find the eigenenergies without finding the functions by using the semiclassical approximation (the so-called WKB method).

We shall omit the calculations with coupled differential equations by using a semiclassical procedure also in another sense. Namely, we shall generalize the nonparabolic dispersion relation [7] to include the presence of external fields. To this end we observe that, including the potential $U(\mathbf{r})$ on the diagonal of eqn [13], one replaces $-\varepsilon$ by $-\varepsilon + U$. Further, if a magnetic field is introduced to the $\mathbf{k}\cdot\mathbf{p}$ theory one replaces $\hbar\mathbf{k}$ by $\mathbf{P} = \mathbf{p} + (e/c)\mathbf{A}$, where $\mathbf{A}$ is the vector potential of the magnetic field. Thus the semiclassical equation resulting from eqn [7] is

$$\left[\frac{\mathbf{P}^2}{2m_0^*} - (\varepsilon - U)\left(1 + \frac{\varepsilon - U}{\varepsilon_g}\right)\right]\Psi = 0 \qquad [14]$$

It can be seen that for $\varepsilon - U \ll \varepsilon_g$ one recovers the standard one-band approximation mentioned above. However, below we consider the general situation described by eqn [14].

Let us first consider the case of no magnetic field, i.e., $\mathbf{P} = \mathbf{p}$. For $U(\mathbf{r}) = U(z)$ one can separate the variables by looking for solutions in the form $\Psi = \exp(\mathrm{i}k_x x + \mathrm{i}k_y y)\Phi(z)$. One can now transform eqn [14] into

$$\left[\frac{1}{2m_0^*}p_z^2 - \frac{1}{\varepsilon_\mathrm{g}}(\varepsilon - \varepsilon_\perp - U)(\varepsilon_\mathrm{g} + \varepsilon + \varepsilon_\perp - U)\right]\Phi(z) = 0 \qquad [15]$$

where

$$\varepsilon_\perp = \frac{\varepsilon_\mathrm{g}}{2} + \left[\left(\frac{\varepsilon_\mathrm{g}}{2}\right)^2 + \varepsilon_\mathrm{g}\frac{\hbar^2 k_\perp^2}{2m_0^*}\right]^{1/2} \qquad [16]$$

in which $k_\perp^2 = k_x^2 + k_y^2$. Equation (15) is suitable for the WKB semiclassical quantization with which one determines the eigenenergies ε by one integration.

For a magnetic field applied along the z-direction (parallel electric and magnetic fields) the effects of both fields are separated and the form of eqn [15] is valid, with $\hbar^2 k_\perp^2/2m_0^*$ in eqn [18] replaced by $\hbar\omega_\mathrm{c}(n + 1/2) \pm (1/2)|g_0^*|\mu_\mathrm{B}B$. Here $\omega_\mathrm{c} = eB/m_0^*c$ is the cyclotron frequency, $n = 0, 1, 2, \ldots$ is the Landau quantum number, g_0^* is the spin splitting factor, and μ_B is the Bohr magneton. This results in nonequal spacing of Landau levels for a fixed value of B and in nonlinear B dependence of the Landau energies as functions of B (similarly but not identically to the bulk case). These features are illustrated by the cyclotron resonance experiment and theory shown in Figure 4.

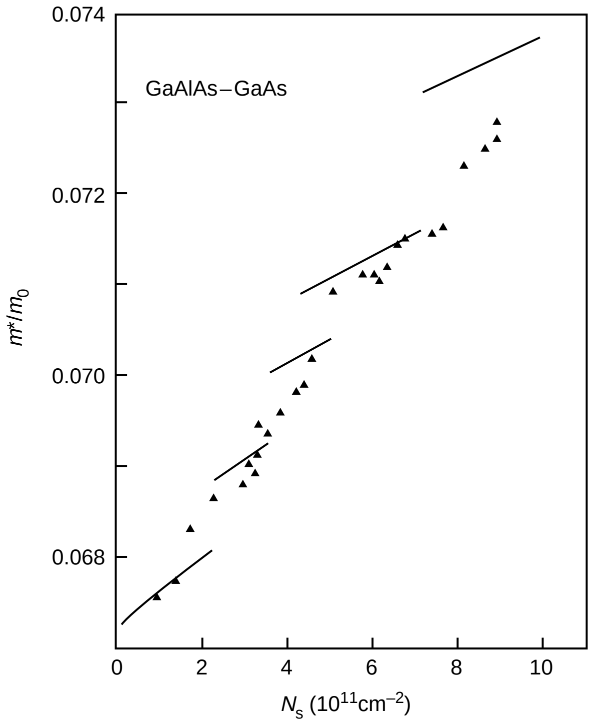

Figure 4 Electron effective mass versus two-dimensional electron density in GaAs/GaAlAs heterostructures, as measured by far infrared cyclotron resonance. The solid lines are calculated using the effective two-level model for GaAs. Reproduced from Zawadzki W, *et al.* (1994) Cyclotron emission study of electron masses in GaAs-GaAlAs heterostructures. *Semiconductor Science and Technology* 9: 320–328.

Selection Rules for Intersub-band Optical Transitions

For the one-band effective mass approximation the electron–photon interaction is $H_\mathrm{int} \sim \mathbf{A}'\cdot\mathbf{p}$ (cf. above), and the wavefunctions for the initial and final sub-bands are $\Psi_m = \exp(\mathrm{i}k_x x + \mathrm{i}k_y y)\Phi_m(z)$, in which the $\Phi_m(z)$ must be orthogonal to each other since they describe different energies. It is then clear that the matrix element $\langle fn|\mathbf{A}'\cdot\mathbf{p}|in\rangle$ is finite only if $\mathbf{A}'$ is polarized along the z-direction (also called the growth direction) since for $\mathbf{A}'$ parallel to x or y the integral over z is $\langle\Phi_{m'}|\Phi_m\rangle = 0$ giving a vanishing matrix element. Experimentally this is a serious problem for spectroscopic applications involving intraband (i.e., intersub-band) transitions in n-type semiconductors, since it is not possible to access these using normal incidence radiation.

This selection rule is somewhat relaxed for p-type semiconductors as a result of the mixed spin states of the complex valence band. In addition the narrow-gap band structure introduces interesting new possibilities into the selection rules. The wavefunctions have the form [12] and for the light polarizations A'_x or A'_y the momentum operators p_x or p_y act also on the periodic amplitudes $u_{l0}(\mathbf{r})$, which leads to a nonvanishing transition probability for intersub-band transitions. This has indeed been observed in narrow-gap materials.

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Materials:** Mercury Cadmium Telluride; Quantum Dots. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Ashcroft NW and Mermin ND (1976) *Solid State Physics*. Holt, Rinehart and Winston.

Kane EO (1980) Band structure of narrow gap semiconductors. In: Zawadzki W (ed.) *Narrow Gap Semiconductors. Physics and Applications*, pp. 13–31. Berlin: Springer.

Luttinger JM and Kohn W (1955) Motion of electrons and holes in perturbed periodic fields. *Physical Review* 97: 869–883.

Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

Zawadzki W (1984) Electric and magnetic quantisation of two-dimensional systems – elementary theory. In: Bauer G, *et al.* (eds) *Two-Dimensional Systems, Heterostructures, and Superlattices*, pp. 2–11. Berlin: Springer.

Zawadzki W (1991) Intraband and interband magneto-optical transitions in semiconductors. In: Landwehr G and Rashba EI (eds) *Landau Level Spectroscopy*, pp. 483–512. Amsterdam: North-Holland.

Excitons

I Galbraith, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

In semiconductors electron–hole pairs can lower their energy by correlating their motion in an exciton state. Such states dominate the bandedge optical properties of direct-gap semiconductors. The resultant absorption peaks are affected by quantum confinement, electric, and magnetic fields and carrier scattering.

Introduction

In the simplest, single-particle picture, the ground state of a semiconductor consists of a full valence band and an empty conduction band. The lowest lying excitation above the ground state in such a scheme is produced by excitation of one electron from the valence band to the conduction band. To achieve this optically requires a photon energy greater than the bandgap, E_g. However in many semiconductor materials this simple picture is unable to explain the observed absorption spectrum in the neighborhood of the fundamental bandgap. The origin of the discrepancy lies in the neglect of the Coulomb interaction between the negatively charged excited electron and the hole which is left in the valence band.

The term exciton refers to a two-particle excitation, consisting of a bound electron and hole. Such excitations dominate the bandedge optical spectra of most direct gap semiconductors, especially at low temperatures. In particular there exists a series of hydrogen-like bound states lying below the bandedge. These states are bound by energies

$$E_b^{3D} = \frac{\mu e^4}{8\mathrm{h}^2 \varepsilon_r^2 \varepsilon_0^2} \frac{1}{n^2} \qquad [1]$$

where $\mu = m_e m_h/(m_e + m_h)$, is the reduced mass, $m_{e(h)}$ the electron (hole) effective mass, e is the electronic charge, h is Planck's constant, ε_0 is the permittivity of free space, ε_r the relative permittivity of the material and $n = 1, 2, 3, \ldots \infty$ is the principal quantum number.

The binding energy of the lowest lying exciton state varies considerably from one semiconductor to another being, e.g., 0.5 meV in InSb and over 60 meV in Zn0. Similarly the spatial extent of the 1 s electron-hole relative wave function or effective Bohr radius is given by

$$a_0^{3D} = \frac{\hbar}{\sqrt{2\mu E_b^{3D}}} \qquad [2]$$

and ranges from 750 μm in InSb to ≈2 nm in Zn0. Typically, in semiconductors the exciton spatial extent is much larger than the lattice constant of around 0.6 nm, and such excitons are classed Wannier excitons. At the opposite limit, which occurs in many molecular materials, the exciton is localized around a particular atomic site and is termed a Frenkel exciton. Wannier excitons are characterized by the hydrogen-like quantum numbers of their relative motion and an overall center of mass momentum which describes the wave-like motion of the bound entity through the crystal.

Another type of exciton is often referred to as a bound exciton. This consists of an electron–hole pair bound to a neutral impurity center. Such excitons are localized around the impurity and cannot move through the crystal in the same way as a free exciton.

Optical Properties

Excitons manifest themselves primarily as strong modifications to the bandedge optical properties of semiconductors. In particular the bound states give rise to discrete absorption lines at lower energy than the bandgap energy. Several peaks corresponding to the different hydrogen-like bound states

(1s, 2s, etc.) can be seen provided the lines are sufficiently narrow. A theoretical calculation of this is shown in Figure 1.

Associated with these absorption changes are concomitant changes in the refractive index. Contributions to the spectral width of the absorption peaks are characterized as either homogeneous or inhomogeneous. Homogeneous broadening refers to broadening associated with the finite lifetime of a particular state, e.g., due to phonon scattering. Inhomogeneous broadening is due to non-uniformity in the system, e.g., in a sample having a spatially non-uniform strain field.

There is also enhancement of the absorption at photon energies above the bandgap due to the residual influence of the Coulomb attraction on the electron-hole scattering states. Although not bound, electrons and holes in these states still have an enhanced probability of being found close together. This is called the Sommerfeld Enhancement Factor (see Figure 1).

It is important to note that although there may be an excitonic peak in the absorption spectrum this does not imply that excitons form a stable population in such a sample. Often, especially at room temperature, the lifetime of an optically generated exciton is determined by the scattering time with optical phonons which may be very short (<1 ps). In this case the excitons are quickly ionized and the quasi-thermal equilibrium which is formed is dominated by essentially plasma-like behavior. In photoluminescence experiments, where carriers are generated high in the band and the spectrum of the resulting emitted luminescence is measured, peaks associated with the free exciton are often observed and taken as a signature of the presence of excitons. This view has been challenged recently by theoretical calculations which show that a photoluminescence peak at the exciton energy can be also be produced from an uncorrelated plasma if proper account of the Coulomb interaction is taken. This issue remains controversial.

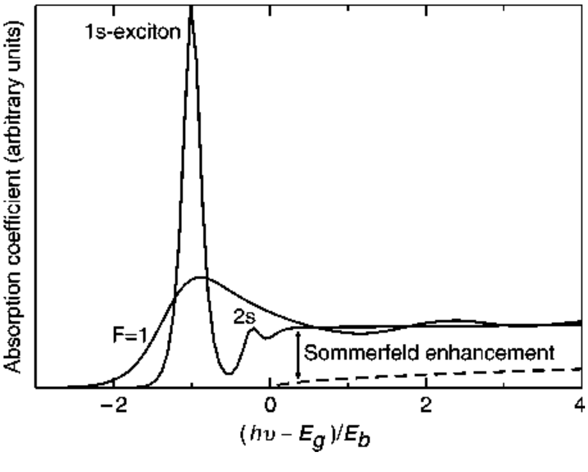

Figure 1 Calculated bulk absorption spectra including (solid-line) and neglecting (dashed line) the Coulomb interaction. The Sommerfeld enhancement is seen as photon energies above the bandgap. Also shown is the spectrum with an applied electric field of one exciton Rydberg per Bohr radius.

In bulk samples, experiments such as absorption and reflectivity are complicated by the existence of polariton effects. An exciton-polariton is a quantum mixture of the propagating photon inside the semiconductor sample and the material excitation. Where the photon dispersion and the exciton energy dispersion meet there is an anti-crossing and this is manifested in, for example, the appearance of two lines in the reflectivity spectrum.

It was hoped during the 1970s that laser action in wide bandgap material such as ZnSe would be possible based on excitonic lasing. In this process a quasi-equilibrium exciton population would form the injected excitation and stimulated recombination would occur with the associated emission of a scattering partner which would take up the necessary momentum to ensure overall momentum conservation. This leads to a light emission wavelength below the absorption edge which is advantageous for minimizing losses. In fact, this process has only been seen at low temperature for scattering with LO-phonons, electrons and other excitons. Each mechanism has its own characteristic lasing threshold and temperature dependence.

Influence of Quantum Confinement

The advent of Molecular Beam Epitaxy enabled the growth of semiconductor layers of thickness similar to the electronic de Broglie wavelength. By sandwiching a low bandgap semiconductor between two high bandgap layers, one can make quantum well structures in which the electronic motion is effectively confined to the plane of the layer. In such structures the excitonic properties are also radically altered by this confinement. An example of the absorption spectrum of a ZnCdSe quantum well embedded in ZnSe is shown in Figure 2. The heavy and light hole excitons correspond to different spin quantum numbers for the electron and hole making up the exciton.

The exciton binding energy is enhanced by a factor of 4 compared to the bulk case and the spatial extent of the 1s wavefunction is also reduced by half:

$$E_b^{2D} = \frac{\mu e^4}{8\mathrm{h}^2\varepsilon_r^2\varepsilon_0^2}\,\frac{1}{(n-\frac{1}{2})^2} \qquad n = 1,2,3\ldots \qquad [3]$$

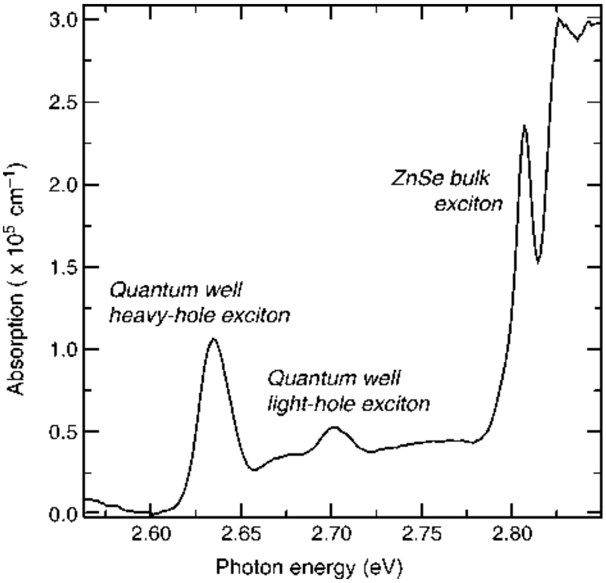

Figure 2 Heavy and light hole exciton absorption spectrum for ZnCdSe quantum wells embedded between ZnSe barriers. Also shown is the bulk exciton of the ZnSe barriers.

This enhancement can be understood since the particles are confined to lie in the plane and are hence closer together on average than in the three-dimensional case. In practice, a four-fold enhancement in the binding energy is never observed since the quantum wells are neither infinitely deep nor infinitely thin. Both these limitations lead to a reduction in the actual enhancement from 4 to about 2, depending on the structure. For example, the binding energy in a realistic quantum well corresponds to the bulk value of the well material for very wide wells and to the bulk value of the barrier material for very narrow wells. In between the value is enhanced by the quantum confinement. The maximum enhancement typically occurs when the well width is around half the bulk Bohr radius for the well material.

Accurate calculation of the exciton binding energy is complicated by the fact that the quantum confinement influences the band structure, making it strongly anisotropic and splitting heavy and light holes. A complete inclusion of these effects is at the limit of current computational power due to the need to calculate the electron-hole interaction matrix elements. Approximate schemes, including variational approaches, have proved successful in fitting a variety of samples, including GaAs and ZnSe based quantum wells. The limitation to such calculations is often the requirement for accurate values of the material parameters such as effective masses and energy band offsets.

The Sommerfeld enhancement of the absorption into the continuum states of quantum wells leads to a doubling of the absorption at the bandedge. This enhancement reduces only slowly as the photon energy is increased.

If multiple quantum wells with very thin barrier layers are used, the electronic states in the neighboring wells couple together to form a superlattice. The superlattice dispersion in the growth direction behaves like a heavy mass and the resulting exciton is three-dimensional but strongly anisotropic.

More extreme quantum confinement is possible making quantum wires, by etching or selective growth of quantum well samples. In these quasi-one dimensional structures, as in quantum wells, the exciton binding energy is again increased further. The Coulomb problem in one-dimension is pathological in that the ground-state binding energy diverges logarithmically, and higher states are pairwise degenerate. However, accounting for the finite cross-section of realistic wires eliminates this problem and recovers a finite binding energy. One consequence of this is that there is no simple enhancement factor which corresponds to the four-fold enhancement in going from bulk to two-dimensional excitons. The density of states in one-dimensional systems diverges at the bandedge, implying the existence of an infinite absorption in a perfect quantum wire. However, the influence of excitonic correlations cancels this divergence, resulting in an enhanced, but not divergent, bandedge absorption feature.

The ultimate limit in quantum confinement can be achieved in quantum dots, where the carriers are confined in all three directions. In this case the term exciton should be used with some care as there are no bandedge states to compare with. The discrete energy levels dictated by the confinement are affected by the Coulomb interaction and the energy is renormalized – in some cases substantially. The absorption spectrum should consist of a series of sharp atomic-like transitions corresponding to s-s, p-p, d-d,... transitions and indeed luminescence spectra obtained from single quantum dots do display such fine structure. Even in the best controlled material systems, such as InAs quantum dots, the spectra are strongly inhomogeneously broadened due to variations in the dot size, shape, and alloy composition.

In large quantum wells, where the width is larger than the exciton Bohr radius the center-of-mass part of the wavefunction can be confined by the barriers. This leads to a quantized motion for the exciton as a whole, rather than for the individual electron and hole separately. One implication of this is that for such excitons the center-of-mass momentum is no longer a good quantum number.

The above effects are a controlled case of disorder induced effects in semiconductors. This is an area which is only now beginning to be understood. The presence of local potential fluctuations will influence the energy spectrum of excitons in that vicinity. Clearly both the depth and spatial extent of these fluctuations will be important – very short range (less than Bohr radius), shallow potential fluctuations will not have any influence on the exciton. While deep potentials can lead to the center of mass confinement above.

Influence of Electric and Magnetic Fields

Just as an externally applied electric field affects the absorption edge via the Franz–Keldysh Effect, broadening the edge itself and inducing oscillations in the continuum part of the spectrum, so, in the case of an excitonic bandedge, there is substantial modification due to the presence of the field. At low fields, when the Coulomb interaction is strong compared to the potential drop across the exciton diameter, the exciton will be little affected by the applied field. At higher fields the exciton will become gradually more polarized by the field until the exciton is field ionized. This ionization process is manifested in the field broadening of the exciton line (see Figure 1).

Of more technological importance is the influence of an electric field applied perpendicular to the plane of a quantum well. The field forces the electron and hole in opposite directions leading to a reduction in the exciton binding energy. This would lead to a blue shift in the exciton absorption peak were it not for an even larger reduction in the single particle electron and hole energy levels within the quantum wells. Overall a strong red-shift of the excitonic absorption peak is observed with increasing field. This is accompanied by a reduction in the oscillator strength due to the reduced electron-hole overlap. This is termed the Quantum Confined Stark Effect and is the basis for the operation of some semiconductor electro-optic modulators. An electric field applied in the plane of the quantum well gives rise to Franz–Keldysh oscillations analogous to the bulk case.

The influence of magnetic fields on excitons is more subtle and richer than the electric field. This is because the magnetic field seeks to induce a circular cyclotron orbit motion on the carriers, but this is influenced by their mutual Coulomb interaction. Two regimes exist where either the cyclotron energy:

$$\hbar\omega_c = \hbar(eB/\mu) \qquad [4]$$

or the exciton binding energy dominates. In the former strong field case, one can treat the Coulomb interaction as a small perturbation on the electron and hole states in the presence of the magnetic field. In exciton states, which have a finite magnetic moment at zero field, we find a linear Zeeman shift in the magneto-exciton energy. In the other case we must use the magnetic field as the perturbation which produces a mixing of the zero-field exciton states. For example, for the ground state the magnetic field induces an admixture of the p-like excited state with the s-like ground state. The total angular momentum of the mixed state is proportional to B and as the energy of a magnetic dipole in a magnetic field is also proportional to B the shift of the magneto-exciton is proportional to B^2. This distinction between linear and quadratic shifts has been used to identify the nature of carrier populations (excitons or unbound free carriers) in a variety of samples. Clearly there exists an intermediate regime where both the magnetic and Coulomb energies are comparable and in this case the precise energy shifts need to be evaluated numerically.

In quantum well samples the orientation of the magnetic field with respect to the confinement direction is crucial in determining the physics. When the field is perpendicular to the confinement plane the cyclotron orbits exist as before and essentially the same phenomenon as in the bulk emerge. For other orientations, the behavior is much more complex and beyond the scope of this article.

Exciton Scattering

An electron and hole in an excitonic bound state execute a correlated motion which can be disturbed by scattering of either partner. This may lead to the ionization of the exciton and the destruction of the correlation or, alternatively, it may change the center of mass momentum of the exciton as a whole. Almost all of the important scattering mechanisms for excitons arise from the charged nature of the electron and hole. Via the Coulomb interaction, excitons can scatter with lattice phonons, other excitons, free carriers, and impurities. Each of these scattering processes has its own regime of dominance, dependent on, for example, temperature or material quality.

When the scattering rate with other particles is much larger than the recombination rate for electrons and holes, a quasi-equilibrium state is reached. The detailed nature of such an interacting electron/hole plasma remains a a long-standing open question. The reasons for this can be traced to interplay between the intrinsic Coulomb interactions within the plasma, which give rise to both bound and scattering states, and Pauli exclusion arising from the fermionic nature of the electrons and holes. This leads to a complex phase diagram which encompasses, e.g., electron-hole droplets, the ionized electron-hole plasma, biexciton phase, and the exciton gas which, some suggest, may undergo Bose–Einstein condensation. The parameters of this phase diagram are carrier density, temperature, and the semiconductor material parameters. Interest in this essentially fundamental question has remained high, stimulated by the stream of technological benefits that even partial answers have brought.

In this context it is worth mentioning that there have been two theoretical approaches to exciton physics which each have their advantages and problems. One approach is to treat excitons as bosonic quasi-particles and derive results from Hamiltonians formulated using bosonic operators. This approach has the appeal of simplicity and produces acceptable results in the low-density regime. It does, however, omit a key feature of the constituent particles making up an exciton namely the fermionic nature of electrons and holes. A more rigorous approach based around electron and hole operators has been followed but this is considerably more complex and numerically involved.

Excitonic Molecules

Just as two hydrogen atoms can lower their total energy by forming a bound hydrogen molecule, so two excitons can make a bound complex which has lower energy than the two isolated excitons. Such complexes are called biexcitons and are mostly important in high quality material and at low temperatures. Their binding energy is less than that of the exciton by a factor of about 0.1–0.3, depending on the ratio of electron to hole effective masses.

Three particle complexes have also been observed consisting of two electrons bound to a hole. These are termed trions or charged excitons. Their binding energy lies intermediate between excitons and biexcitons.

See also

Semiconductor Materials: GaAs Based Compounds; Large Gap II–VI Semiconductors; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Bastard G (1998) Wave Mechanics Applied to Semiconductor Heterostructures. New York: Halstead John Wiley.

Haug H and Koch SW (1993) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. Singapore: World Scientific.

Klingshirn CF (1997) *Semiconductor Optics*. Berlin: Springer.

Rashba EI and Sturge MD (eds) (1982) *Excitons*. Amsterdam: North Holland.

Schmitt-Rink S, Chemla DS and Miller DAB (1989) Linear and nonlinear optical properties of semiconductor quantum wells. *Advances in Physics* 38: 89–188.

Ueta M, Kanzaki H, Kobayashi K, Toyozawa Y and Hanamura E (1986) *Excitonic Processes in Solids*. Berlin: Springer.

Yu PY and Cardona M (1999) *Fundamentals of Semiconductors*. Berlin: Springer.

Impurities and Defects

K A Prior, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A simple definition of a defect is any feature that disturbs the periodicity of the perfect lattice. This definition is sufficiently broad to encompass a whole range of different entities, which are normally divided into three categories based on their dimensionality, or extent in real space.

0D defects are those which have an extent of the order of only a few atomic spacings at most in any direction. They include all cases where one, or at most a few, atoms are removed or displaced, or are replaced by atoms of a different type.

1D defects are linear features within the crystal that extend over many atomic dimensions. The most important type of 1D defect is the dislocation and in a single crystal it is possible for a dislocation to extend from one side of the crystal to the other.

2D defects include a variety of different features that occur primarily in one plane. Any junction between two different types of crystal in an epitaxially grown layer (a heterointerface) fits this description, as of course does the sample surface. There are also crystal defects which exist in two dimensions of which the most important is the stacking fault.

Finally, there are also 3D defects. This category includes voids and inclusions, which are both major perturbations in a crystal structure. In most pure crystals grown by modern techniques these are eliminated and they will not be discussed further.

In the sections below the main types of defect will be discussed according to the scheme described above. Despite the substantial research effort to understand the basic properties of many semiconductors, and of course the huge effort devoted to producing semiconductor devices, it remains true that the defect structure of most semiconductors is still poorly understood, with the exceptions of silicon and gallium arsenide. This is especially true in the case of the nitride semiconductors with the wurtzite structure which have recently been the subject of much commercial interest. In the following sections examples will mainly be taken from gallium arsenide.

Lastly, it is important to note that defects of different dimensionality interact. Several possible interactions are discussed in the last section.

0D Defects

Dopants

There are two basic types of 0D defects to consider: chemical defects, or impurities, and native or point defects caused by missing or misplaced atoms. Table 1 gives the major types of 0D defect found in compound semiconductors together with the symbol used in the common notation for point defects and impurities.

Impurity atoms are those which are introduced either accidentally during the production of the bulk crystal or epitaxial layer, or deliberately by the crystal grower in precisely controlled amounts to modify some property of the material. Usually, this means that the material added is a dopant, and has been added to create a surplus of either free electrons or holes in the semiconductor. Typically, the amounts of dopants added are between 10^{-6} and 10^{-3} mole fraction, giving doping levels between $10^{22}\ \mathrm{m}^{-3}$ and $10^{25}\ \mathrm{m}^{-3}$.

Table 1 Impurities and point defects in a compound *AB* with the zincblende lattice structure

Symbol	*Description*
X_A^n, X_B^n	Impurity atom, *X*, on *A* or *B* site with charge *n*
X_i^n	Interstitial impurity atom with charge *n*
V_A^n, V_B^n	Vacancy on *A* or *B* site with charge *n*
A_i^n, B_i^n	Interstitial *A* or *B* atom with charge *n*
A_B, B_A	*A* atom on *B* site or *B* atom on *A* site (antisite)

In general, any atom added to a material will introduce a localized state into the band structure of the semiconductor. This state may rarely occur in either the conduction or valence band, but is more likely to be located in the bandgap. If the localized energy level is within a few kT of the valance or conduction band edge, then the state is an electron acceptor or donor. If it is much further from the band edge then it is usually known as a deep level. In this case its main effect on the electrical properties will be to act as a trap for electrons or holes, and introduce noise into the current flow. Its effect on the optical properties of the material will be to reduce the radiative recombination rate for band-to-band transitions by acting as an alternate recombination pathway for either radiative or nonradiative transitions.

It is generally assumed that dopant atoms occur only on the site of the appropriate sublattice and are well behaved. While it is true that some elements are well behaved substitutional dopants, such as sulfur in gallium arsenide, which can be denoted $\mathrm{S}_{\mathrm{As}}^{+}$ when acting as a singly charged donor, the behavior of other species is more complex. In many III–V compounds, silicon is an amphoteric dopant, occupying sites on both the metal and nonmetal sublattices which ionize as shown:

$$\mathrm{Si}_{\mathrm{Ga}}^{0} = \mathrm{Si}_{\mathrm{Ga}}^{+} + \mathrm{e}^{-}$$

$$\mathrm{Si}_{\mathrm{As}}^{0} = \mathrm{Si}_{\mathrm{As}}^{-} + \mathrm{h}^{+}$$

In molecular beam epitaxy (MBE) growth on GaAs(100) the ratio between the donor $\mathrm{Si}_{\mathrm{Ga}}^{+}$ and the acceptor $\mathrm{Si}_{\mathrm{As}}^{-}$ varies as the concentration of doping is increased. At low levels silicon is a donor, while at high concentrations the fraction which is an acceptor increases rapidly. The semiconductor is therefore compensated, effectively limiting the maximum concentration of free carriers.

Another type of behavior occurs with beryllium in III–V semiconductors and lithium in II–VI semiconductors. Here, the dopant atom is so small that it can easily be displaced from the lattice site. Thus, for example, in InGaAs, it is energetically favorable for the acceptor $\mathrm{Be}_{\mathrm{In}}^{-}$ atom to move from the lattice site to an interstitial position:

$$\mathrm{Be}_{\mathrm{In}}^{-} = \mathrm{V}_{\mathrm{In}}^{-} + \mathrm{Be}_{\mathrm{i}}^{+} + \mathrm{e}^{-}$$

The interstitial Be ion is a donor and causes some compensation, but even worse, is highly mobile. It can move a substantial fraction of a micron in an epitaxial device and in so doing completely destroys a carefully engineered abrupt p-n junction. It can be seen therefore that dopant interactions with the host lattice are important in determining device performance.

Native Defects

A large variety of native defect species are possible in semiconductors. Interstitial atoms and vacancies are found in both the diamond structure (silicon) and the zincblende structure (most III–V and II–VI semiconductors). In zincblende structures there is also the possibility of placing an atom on the wrong atomic site, or sublattice; for example in GaAs a Ga atom may be found on an As site. This type of defect is known as an antisite.

Bulk semiconductor crystals are grown under conditions close to equilibrium. In the long periods of time needed to grow the crystals, equilibrium concentrations of point defects are established. Epitaxial growth techniques (in particular molecular beam epitaxy, MBE and metalorganic chemical vapor deposition, MOCVD) operate under conditions far from equilibrium and the point defect concentrations may differ markedly from the equilibrium values. Some semiconductor processing operations can also increase point defect concentrations. Etching semiconductors with ion beams or energetic plasmas is a very efficient way of creating in the subsurface region high concentrations of point defects well above the equilibrium values.

It is possible to anneal crystals in the vapor of one of the constituent elements to alter the point defect concentrations. Interstitial and vacancy concentrations can be changed relatively rapidly over the course of several hours (interstitials) or days (vacancies) as these species are much more mobile than dopant atoms. Usually, annealing is done to reduce point defect concentrations, but on occasions, for example with the alloy HgCdTe, it is done to increase the concentration of a type of defect, in this case mercury vacancies which are used to dope the material.

Like dopant atoms, point defects can also introduce defect levels into the band structure of the semiconductor, and some defects, for example the gallium vacancy $\mathrm{V}_{\mathrm{Ga}}^{n-}$, exist in multiple charge states. In some materials, the native defects act either as dopants or as compensating traps, and are present in concentrations that can severely limit the maximum obtainable doping levels. The gallium vacancy, for example, is believed to limit the maximum doping level obtainable in n-type GaAs, and its concentration increases dramatically as the concentration of n-type dopant is increased.

It should be noted that the concentrations of the various types of defect are not independent of one another, and at any given growth (or annealing) temperature and applied partial pressures of the constituent elements, the equilibrium concentrations of the defects are fixed. Examples of the interactions between the types of defect may be seen by considering the displacement of a gallium atom from its lattice site to create a vacancy and an interstitial (the Frenkel reaction):

$$\mathrm{Ga}_{\mathrm{Ga}}^{0} = \mathrm{Ga}_{\mathrm{i}}^{0} + \mathrm{V}_{\mathrm{Ga}}^{0}$$

and the interaction of an arsenic interstitial with a gallium vacancy to create an arsenic antisite:

$$\mathrm{As}_{\mathrm{i}}^{0} + \mathrm{V}_{\mathrm{Ga}}^{0} = \mathrm{As}_{\mathrm{Ga}}^{0}$$

The relative importance of the types of defect varies between semiconductors. In compounds where one of the elements is very volatile, then high concentrations of vacancies on that particular sublattice are to be expected. Antisites occur only at very low concentrations in the ionic II–VI compounds, as the energy involved in placing an atom on the wrong sublattice is very high. However, in the more covalent III–V compounds, there is not such a large energy penalty, and it has been suggested that in GaAs the $\mathrm{As}_{\mathrm{Ga}}$ antisite may be the most common point defect.

In addition to the simple defect types described above, there are other possibilities. One of these is that complexes of defects may form. The best known are the complexes that occur between dopant atoms and vacancies, for example the $[\mathrm{Te}_{\mathrm{As}}\mathrm{V}_{\mathrm{Ga}}]^{-}$ complex found in Te doped GaAs (Figure 1). In this material, partial compensation of the n-type dopant occurs with the formation of gallium vacancies, as described above:

$$\frac{1}{2}\mathrm{As}_2(\nu) = \mathrm{V}_{\mathrm{Ga}}^{0} + \mathrm{As}_{\mathrm{As}}$$

$$\mathrm{V}_{\mathrm{Ga}}^{0} = \mathrm{V}_{\mathrm{Ga}}^{2-} + 2\mathrm{h}^{+}$$

These vacancies are mobile and diffuse in the material. In the vicinity of $\mathrm{Te}_{\mathrm{As}}^{+}$ they experience a strong Coulombic attraction and the complex forms when the vacancy and dopant atom are on nearest neighbor sites:

$$\mathrm{V}_{\mathrm{Ga}}^{2-} + \mathrm{Te}_{\mathrm{As}}^{+} = [\mathrm{V}_{\mathrm{Ga}}\mathrm{Te}_{\mathrm{As}}]^{-}$$

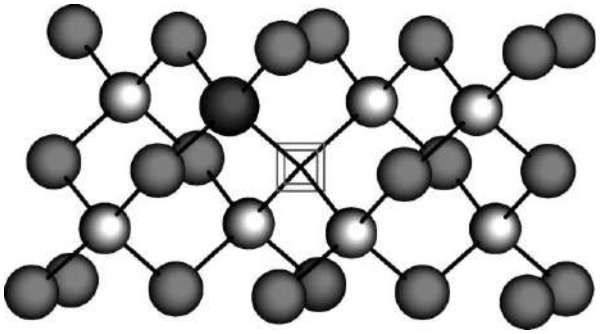

Figure 1 Diagram of the tellurium atom–vacancy complex in GaAs. Gallium and arsenic atoms are shown in mid-gray and white, respectively, while the tellurium atom is shown in dark gray.

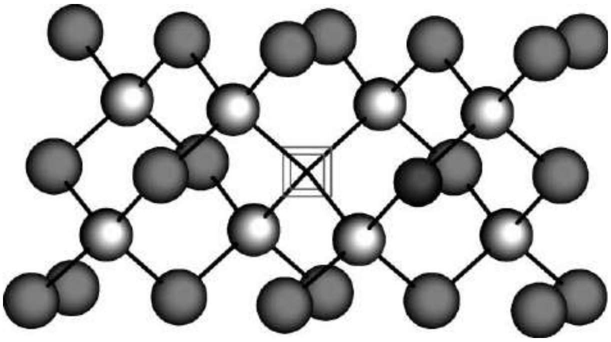

Figure 2 Diagram of the silicon atom–vacancy complex in GaAs. Gallium and arsenic atoms are shown in mid-gray and white, respectively, while the silicon atom is shown in dark gray.

A similar complex also forms in Si doped GaAs: $[\mathrm{Si_{Ga}V_{Ga}}]^-$ (Figure 2), although here both species exist on the same sublattice, and consequently occur as next nearest neighbors in the complex.

In the description above it is assumed that all atoms which are located on a particular site are centered on that site. However, it is possible for a dopant atom to replace an atom of the host semiconductor, but not occupy the lattice site. *Ab initio* calculations of the bonding around the dopant atom suggest that it can be displaced away from one of its nearest neighbors towards the other three atoms (Figure 3). The dopant can be considered as having one bond weaker than the other three.

In the case of n- and p-type doped semiconductors the corresponding defects are known as DX and AX centers respectively. DX centers occur in many semiconductors but are particularly important in AlGaAs, where Si (or other n-type dopant) atoms form DX centers that partially compensate the material and reduce the electron mobility. However, irradiation with white light excites the DX center and converts it back to a normal dopant centered on the lattice site. In AlGaAs the DX center is the stable species and the dopant atom is a higher energy state. Conversion back to the DX center requires surmounting an energy barrier and thus the normal dopant is metastable. The irradiated semiconductor has a lower concentration of DX and a higher doping concentration than before, and consequently has a higher conductivity than the unirradiated material. This results in persistent photoconductivity in the irradiated material, which is a characteristic feature of semiconductors containing DX centers. At low temperatures, this effect can persist for many hours.

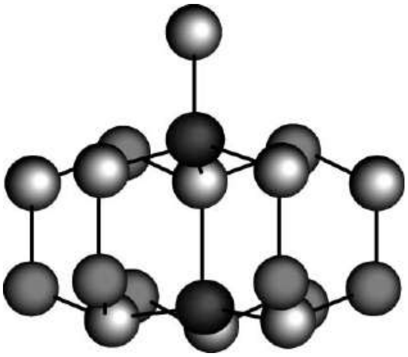

Figure 3 A GaAs crystal shown with [111] oriented vertically. Two Ga atoms have been substituted by silicon atoms which are shown in dark gray. One Si atom is on a substitutional site (top atom) and the other is displaced along [111] away from one of its nearest neighbors (bottom atom). This is one of the structures suggested for the DX center.

1D Defects

Within any crystal, there must exist a set of planes where the interplanar bonding is weakest. In both the diamond and zincblende crystal structures these are the {111} planes. When a crystal is subjected to a shear stress above its elastic limit, then the material accommodates that stress by changing shape, alternately breaking and reforming bonds across these planes where the bonding is weakest. The resulting movement is easily observable under a microscope in many materials and is known as slip or glide; the planes on which this happens are known as slip or glide planes.

Movement on a slip plane does not happen over the whole crystal at once, and there is always a boundary between the material in which slip has occurred and material where it has not. This boundary is the dislocation. A section through a piece of material where slip has occurred is shown in Figure 4. The curved line DD′ marks the dislocation and slip has occurred in the region marked DQD′. Because it marks a boundary, the dislocation may not end within the crystal and the only permitted ways that a dislocation may terminate are to finish at the surface of the crystal (as shown) or to meet another dislocation, forming either a closed loop or a junction.

Across the slip plane, the material has moved by a certain amount, which is, typically, a lattice vector within the crystal. The direction of slip is known as the Burgers vector of the dislocation, and it is a constant for any one dislocation. The Burgers vector at any point may be determined by describing a simple loop in the crystal passing through adjacent lattice points lying in one plane. If a dislocation lies within the loop then the ends will not close, and the displacement needed to close the loop is the Burgers vector. This can be seen in Figure 4 on the left-hand face of the crystal around point D, where the Burgers vector is parallel to the line of the dislocation. It can also be seen on the right hand face of the crystal at point D′, where the Burgers vector, which has not changed direction, now lies perpendicular to the dislocation line. Dislocations where the Burgers vector is parallel or perpendicular to the dislocation line are known as screw and edge dislocations, respectively.

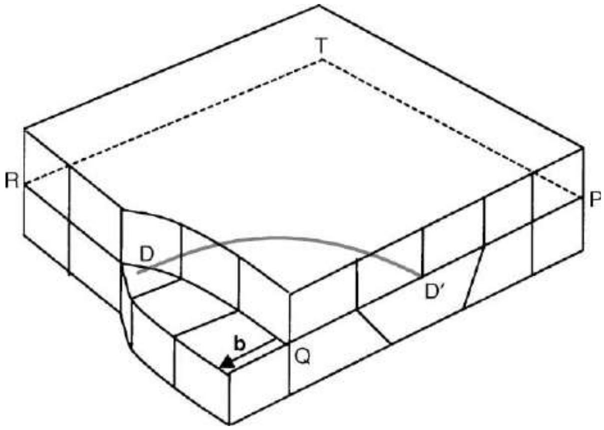

Figure 4 Diagram of a crystal showing a region where slip has occurred. For clarity only the lattice points are shown. The slip plane is PQRT and the slipped region is DQD′. The curve DD′ is the line of the dislocation and has a Burgers vector **b**.

Examples of edge and screw dislocations are shown in Figures 5 and 6, respectively. Edge and screw dislocations are considered to be ‘pure’ types of dislocations. If the Burgers vector makes any other angle to the dislocation line, the dislocation is referred to as a mixed dislocation. In semiconductors mixed dislocations are the most common type. There are many different ways of describing dislocations; however, the most common one is to give the angle between the dislocation line and the Burgers vector. In this system screw and edge dislocations are referred to as 0° and 90° dislocations.

At the core of the dislocation there is a small region, typically only one or two atoms wide, where the atoms have highly distorted bonding. Around this

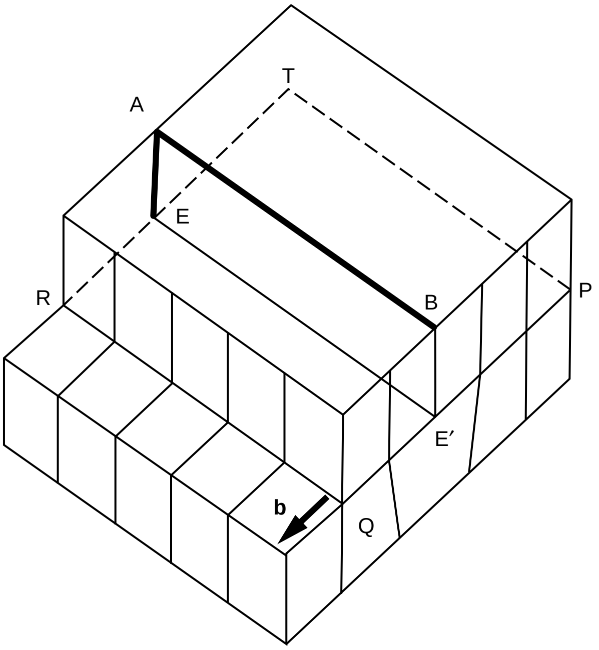

Figure 5 A schematic diagram of an edge dislocation. Only the lattice points are shown for clarity. The slip plane is PQRT and contains the Burgers vector denoted **b**. The line of the edge dislocation is EE′, which can be thought of as terminating a half-plane of atoms denoted ABEE′.

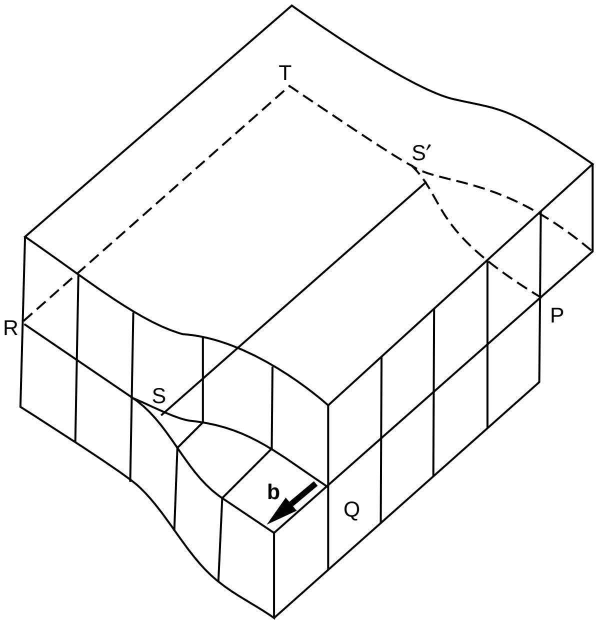

Figure 6 A schematic diagram of a screw dislocation. Only the lattice points are shown for clarity. The slip plane is PQRT and contains the Burgers vector denoted **b**. The line of the screw dislocation is SS′.

core the atoms adopt the normal (tetrahedral) bonding, although they are strained for many nanometers around the dislocation. The strain field around the dislocation means that low dislocation densities are very easy to detect using a technique such as X-ray topography, which images the strain introduced locally in a set of crystal planes by a dislocation. Using this technique, individual dislocations may be determined within crystals. Currently, it is possible to obtain silicon crystals which are essentially dislocation free, while the best GaAs crystals have dislocation densities of $<10^5$–$10^6\,\mathrm{m}^{-2}$. In other materials the densities may be 100–1000 times higher.

The highly distorted bonding at the core of the dislocation inhibits dislocation motion. Under the action of an external stress, dislocations normally move on their glide plane to relieve the stress and deform the material plastically. This is dislocation glide. In covalent semiconductors, such movement means breaking and reforming strong bonds, and does not occur in silicon below the crystal growth temperature. Dislocations are also immobile in GaAs at room temperature, although they can be induced to move in softer, more ionic II–VI semiconductors. Dislocation densities are therefore usually considered 'frozen in' at room temperature.

In addition to macroscopic shear forces, each dislocation experiences the stresses introduced by other dislocations and point defects present in the crystal, and is either attracted or repelled by them. At high densities, therefore, the motion of dislocations is inhibited by the presence of the surrounding defects – the dislocations are said to be pinned.

Generation and movement of dislocations occurs in most semiconductors easily at their growth temperature, and is important in the relief of strain introduced in growing a material on a substrate with a different lattice constant. Edge dislocations always introduce, or remove, extra planes of atoms (Figure 5), and so an array of dislocations can add or remove atoms to accommodate the strain between two semiconductor layers of different lattice constant. These are known as mismatch dislocations and they occur frequently in epitaxial growth when strain relief has occurred.

Typically, growth in MBE or MOCVD occurs on the (001) surface. Mismatch dislocations lie on the {111} planes, which intersect this plane, and at low densities occur in crystals regularly spaced along [110] and $[1\bar{1}0]$ directions. The Burgers vectors are also of $\langle 110 \rangle$ type and are typically at 60° to the dislocation line. The edge component of this vector, projected on to the (001) plane, gives the amount of strain relief to be expected.

Despite the distorted bonding that exists around the dislocation core, there are thought to be remarkably few broken bonds there. An exception to this is thought to be when the dislocation line is not straight, but kinked, the atoms at the kinks having broken bonds. As a result, nonradiative transitions involving dislocations are not strong. They are important, however, during operation of lasers and LEDs. Dislocations can be seen in electroluminescence of light-emitting devices as dark lines. These so-called dark-line defects (DLDs) are the result of nonradiative recombination occurring at dislocations. Not only do they reduce the efficiency of the device, but the energy released by the recombination of the electron and hole is deposited at the dislocation core as thermal energy. Under these conditions, dislocation movement becomes easy and DLDs propagate, eventually destroying the device.

2D Defects

In a simple description of metallic structures, the hexagonal close packed structure is described as a sequence of hexagonal sheets with the stacking pattern *ABABABAB*..., while that of the face centered cubic structure is *ABCABCA*... In a similar manner, we can describe the wurtzite crystal structure by the stacking sequence *ABABABAB*..., although the layers are no longer close packed and the stacking units comprise one metal and one nonmetal atom. The corresponding stacking sequence for the zincblende structure then becomes *ABCABCA*... These structures are shown in Figures 7a and b, respectively. It can be seen that the two atom units are oriented perpendicular to the stacking sequence, and it is possible to rotate the structure about the vertically oriented bonds between any two given planes. A rotation of 180° would place all the atoms back in registry but on the wrong site. For example, in a zincblende crystal, a rotation after the third layer in the sequence *ABCABCA*... would result in *ABCBCAB*... This means that the top part of the crystal is related to the bottom half by a rotation of 180°, and the two parts are said to be twinned. The twin plane extends throughout the crystal, and is a two-dimensional defect. Both above and below the twin plane, the stacking sequence is the usual *ABC*... found in the zincblende structure; however, at the junction there is a region denoted *BCB*. This region is three planes of the wurtzite stacking sequence, and is an example of what is called a stacking fault. In zincblende crystals, stacking faults introduce small regions of the wurtzite structure, while the reverse happens in wurtzite crystals.

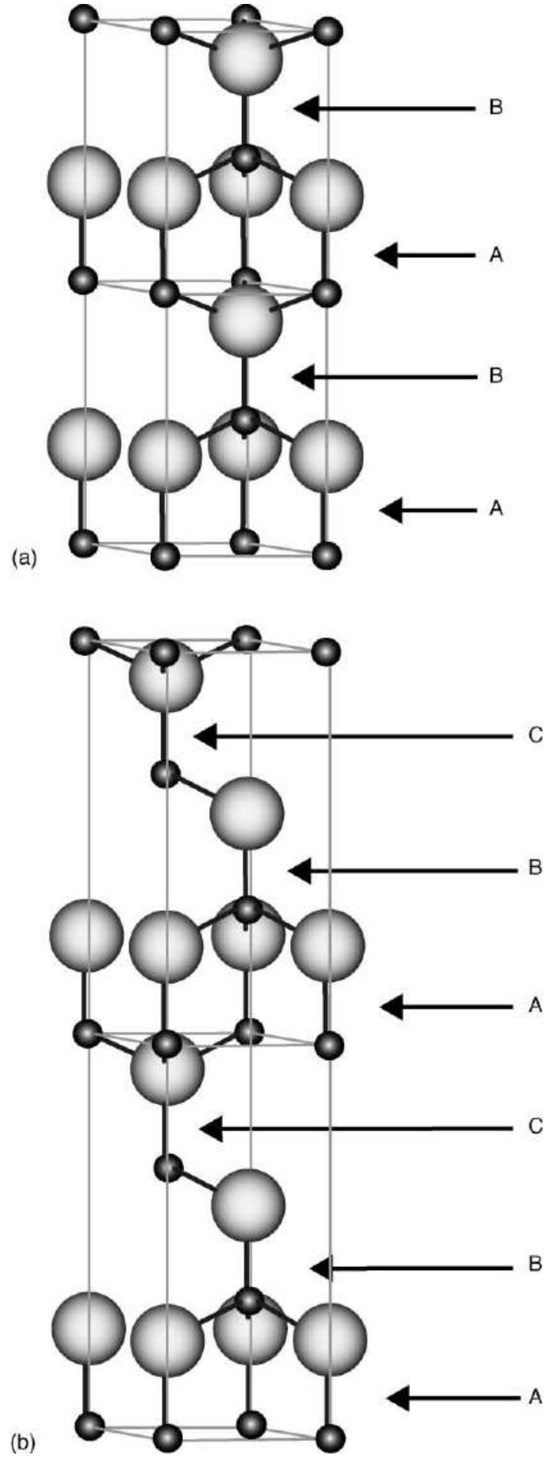

Figure 7 Diagrams of the stacking sequence of layers for two different semiconductor structures. In both cases the metal atoms are small and dark while the larger pale ones are the nonmetal atoms. The structures are: (a). The wurtzite crystal structure. The atoms which occupy the A, B and C sites are marked. Here the vertical direction in which the planes stack is the [0001] direction. (b). The zincblende crystal structure. The atoms which occupy the A and B sites are marked and the vertical direction corresponds to one of the ⟨111⟩ directions in the crystal.

The density of stacking faults in a material is primarily a function of the ease of placing the crystal planes in the wrong order. Under nonequilibrium growth conditions, this can occur easily for many different materials, but for slow growth near equilibrium, the prevalence of stacking faults is governed by the stacking fault energy. This is a measure of the wrong stacking sequence occurring and is also related to the energy of transformation between the zincblende and wurtzite crystal structures. For GaAs, the energy is high and stacking faults are rare, while for II–VI semiconductors the stacking fault energies are low and stacking faults are relatively common.

Typically, stacking faults are observed in TEM planar view micrographs, and on (001) oriented epitaxial samples are clearly visible as triangular regions lying on {111} planes, as described above, nucleated at the interface between the substrate and epitaxial layer.

While stacking faults are accidental introductions in a crystal structure, other 2D defects are introduced deliberately. Heterojunctions between different semiconductor alloys are an obvious example and are responsible for most of the useful properties of advanced semiconductor devices. However, by far the most important 2D crystal defect is the surface of the crystal. This creates a high density of broken bonds, each of which can introduce a defect level in the band structure. Surfaces usually lower their free energy by moving atoms from the bulk equilibrium positions to new sites and forming extra bonds in the plane of the surface. This removes some of the defect levels, but it is still possible to have surface defect densities as large as $10^{18}\ \mathrm{m}^{-2}$ which are associated with enhanced rates of nonradiative recombination of electrons and holes.

Interactions of Defects

Earlier it was assumed that the Burgers vector associated with a dislocation is a lattice vector. The energy of a dislocation per unit length is proportional to the magnitude of the Burgers vector squared:

$$E \propto b^2$$

In some crystals, however, it is possible for the displacement to be composed of two smaller steps, neither of which is a complete lattice vector:

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$$

where $\mathbf{b}$ is a lattice vector and $\mathbf{b}_1$ and $\mathbf{b}_2$ are the smaller displacements. In this case, the total energy

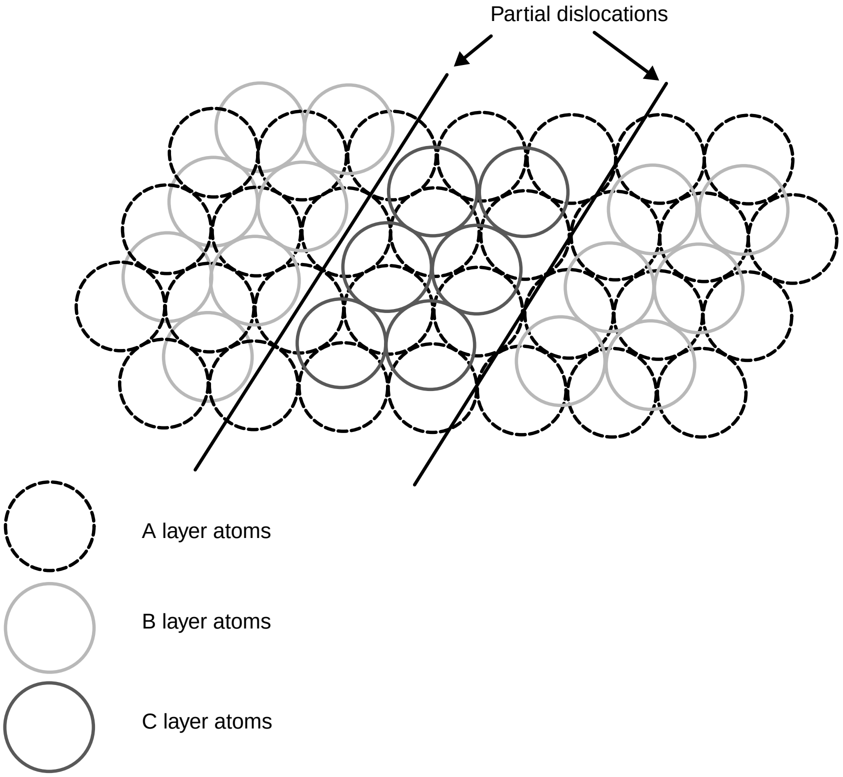

Figure 8 An example of a close packed structure where a region with a stacking fault lies between two partial dislocations.

can be reduced if the dislocation splits into two:

$$E_1 + E_2 \propto b_1^2 + b_2^2 < E$$

These dislocations are known as *partial* dislocations. A normal dislocation splits into two partial dislocations if it is energetically favorable to do so. Figure 8 shows an example in a close packed structure, which illustrates the point without the problem of introducing two different types of atoms. Two partial dislocations run through the top layer. The material lying between them has a different stacking sequence (AC) from the material around it (AB) and must contain a stacking fault. This shows that a 2D defect, such as a stacking fault, is bounded by two 1D defects, in this case partial dislocations.

There also exists an intimate relationship between point defects and dislocations, which may be seen by examining the edge dislocation shown in Figure 5. The edge dislocation may be considered as an extra half-plane of atoms, ABEE′ which terminates along line EE′. The last row of atoms in the plane can be removed from the sheet by one of two processes. Firstly, the atoms at the dislocation core can become interstitials and diffuse away from the dislocation. Alternatively, vacancies can diffuse towards the dislocation. In either case the result is the same; the dislocation line EE′ moves upwards towards AB. The opposite sequence of events (diffusion of interstitials towards the dislocation or vacancies away from it) is also possible and results in a movement of line EE′ downwards. Such processes are known as dislocation climb. Climb is not as common as glide, but it is important as the dislocation acts as a source or sink for point defects. This process contributes to the reduction in radiative efficiency in LEDs and lasers, particularly as the devices age.

See also

Semiconductor Physics: Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures.

Further Reading

Beanland R, Kiely CJ and Pond RC (1992) Dislocations in heteroepitaxial films. In: Mahajan S (ed.) *Handbook on Semiconductors. Materials, Properties and Preparation*, vol. 3a, pp. 1149–1225. Amsterdam: North-Holland.

Fukuda M (1991) Reliability and Degradation of Semiconductor Lasers and LEDs. Boston, MA: Artech House.

Hirth JP and Lothe J (1982) Theory of Dislocations. New York: Wiley.

Hull D (1965) Introduction to Dislocations. Oxford, UK: Pergamon Press.

Hurle DTJ (1999) A comprehensive thermodynamic analysis of native point defect and dopant solubilities in gallium arsenide. *Journal of Applied Physics* 85: 6957–7022.

Kelly A, Groves GW and Kidd P (2000) Crystallography and Crystal Defects, revised edn. Chichester, UK: Wiley.

Kossevich AM (1999) The Crystal Lattice. Berlin: Wiley-VCH.

Malloy KJ and Khachaturyan K (1993) DX and related defects in semiconductors. In: Willardson RK and Beer AC (eds) *Semiconductors and Semimetals*, pp. 235–291. New York: Academic Press.

Schubert EF (1993) Doping in III–V semiconductors. In: Ahmed H (ed.) *Cambridge Studies in Semiconductor Physics and Microelectronic Engineering*, vol. 1. Cambridge, UK: Cambridge University Press.

Skowronski M (1992) Impurity- and defect-centres in compound semiconductors. In: Mahajan S (ed.) *Handbook on Semiconductors. Materials, Properties and Preparation*, vol. 3a, pp. 1343–1398. Amsterdam: North-Holland.

Infrared Lattice Properties

T J Parker and S R P Smith, University of Essex, Colchester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The vibrations of atoms in solids were first studied theoretically at the beginning of the twentieth century to explain the observed temperature dependence of the heat capacity. Following earlier work by Einstein and Debye, Born and von Karman showed in 1912 that the vibrational properties of a linear chain of alternate light and heavy atoms connected by springs account very well for the observed thermal behavior of solids. It was shown many years later that, if the atoms are alternately positively and negatively charged, the model also gives a good description of the optical properties.

In this article the vibrational properties of a linear diatomic chain of atoms are first described. The model gives a good description of the basic vibrational properties of solids with two atoms per unit cell, such as Si, Ge, and the binary III–V (e.g., GaAs) and II–VI (e.g. CdTe) semiconductors, and explains the origin of optic and acoustic branches of the phonon dispersion curves. This leads naturally to an understanding of the basic far-infrared reflectivity of binary semiconductors. The model is then further developed to account for the effects of impurities and the optical properties of alloys.

The far-infrared properties of semiconductors are then reviewed. We begin with the group IV semiconductors Si and Ge in which there is no first-order interaction with infrared radiation, and explain the origins of second-order interactions, and the effects of disorder and impurities. Examples of the first-order reflection spectra of bulk single crystal binary semiconductors and their alloys are then presented. Localized vibrational modes and phonon combination bands are discussed next. We finish with a brief review of more exotic excitations that occur in layered semiconductors: confined optic modes, surface phonon polaritons, guided waves, and interface modes.

Background Theory of Phonons in Semiconductor Materials

Phonons in a Diatomic Linear Chain Lattice

Consider a line of atoms of alternating type A and B, separated by a distance a, as shown in Figure 1. Suppose that the nth A atom is displaced by a small amount $u_{A,n}$ from its equilibrium position. The forces attempting to return it to its equilibrium position will be proportional to the distance it has moved relative to the neighboring atoms. If we assume that the only significant force involves the nearest neighbor B atoms n and $n+1$, the equation of motion for the nth A atom is

$$M_A \frac{d^2 u_{A,n}}{dt^2} = C(u_{B,n} + u_{B,n+1} - 2u_{A,n}) \qquad [1]$$

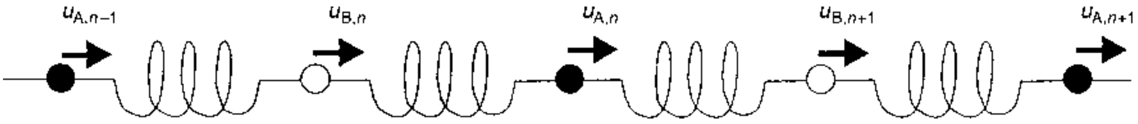

Figure 1 Atomic displacements in a diatomic linear chain.

where M_A is the mass of the atom, and C is the Hooke's law nearest-neighbor force constant. There is a similar equation

$$M_B \frac{d^2 u_{B,n}}{dt^2} = C(u_{A,n-1} + u_{A,n} - 2u_{B,n}) \qquad [2]$$

for the motion of the nth B atom. We assume that the solution of the equations of motion has the wavelike form

$$\begin{aligned} u_{A,n} &= U_A \exp\{i(qan - \omega t\} \\ u_{B,n} &= U_B \exp\{i(qan - \omega t\} \end{aligned} \qquad [3]$$

in which U_A and U_B are amplitude factors, ω is the angular frequency, and q is the wavevector, so that qa is the phase factor by which the phase of the wave changes from one site n to the next site $n + 1$. The equations of motion are satisfied so long as the angular frequency ω is given by

$$\omega^2 = \frac{C}{M_A M_B}\Big[M_A + M_B \pm \sqrt{(M_A + M_B)^2 - 4M_A M_B \sin^2(qa/2)}\Big] \qquad [4]$$

There are two possible solutions here because there are two ions in the unit cell. A plot of ω as a function of q is shown in Figure 2a, with q in the range $-\pi/a$ to $+\pi/a$. In the lower branch, the *acoustic* phonon branch, the vibrations at long wavelength (equivalent to small q) are essentially the same as ordinary sound waves. The A and B atoms vibrate in phase, and at long wavelength there is a linear relation between ω and q:

$$\omega \approx v_s q \qquad [5]$$

where $v_s = \omega_0 a/2$ is the velocity of sound of the acoustic phonons.

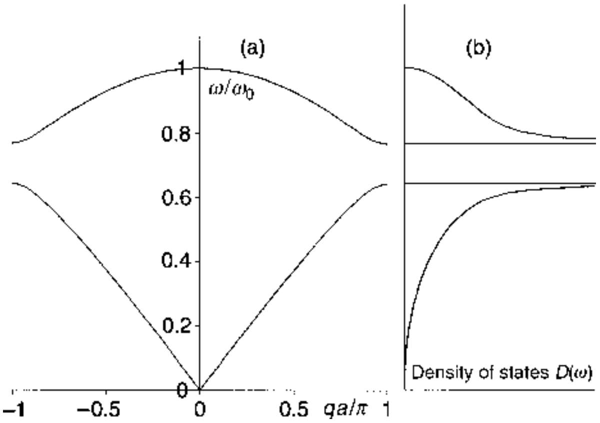

Figure 2 (a) Phonon frequency ω versus wavevector, q, in a one-dimensional diatomic lattice; (b) three-dimensional density of states $D(\omega)$.

The upper branch in Figure 2a, which is of more interest in the present context, is called the *optic* phonon branch. Here the A and B atoms vibrate in antiphase, and because the A and B atoms have equal but opposite charges there is an electric field associated with the vibration. The optic phonon vibrations will consequently affect the propagation of electromagnetic waves through the medium – hence the name – although the vibration frequencies (around 10 THz) are in the far infrared rather than the visible part of the electromagnetic spectrum. At this frequency, the infrared wavelength λ is about 30 μm. The propagation of the infrared will be affected only by phonons whose wavevector q matches the infrared wavevector $k = 2\pi/\lambda$. Since λ is around 10^4 times larger than the typical lattice spacing a in semiconductor materials, it follows that the optical properties of the material may be largely understood merely by considering the long-wavelength phonons with wavevector $q \approx 0$. Thus, in the diatomic case, only the optic phonon of wavevector $q \approx 0$ and frequency

$$\omega_0 = \sqrt{\frac{2C(M_A + M_B)}{M_A M_B}} \qquad [6]$$

is of importance to the propagation of electromagnetic radiation.

Many semiconductor materials are of the diatomic AB type – for example group III–V materials such as GaAs, and II–VI semiconductors such as CdTe. Their phonons are described in detail by three-dimensional versions of the simple linear chain model discussed above. The most basic semiconductor materials, Si and Ge, are clearly monatomic, but their crystal structure is identical to the zincblende structure of GaAs, and they show two phonon bands as in Figure 2a. However, because there is only one type of atom, these materials differ from the diatomic materials in that the optic phonon branch has no electric dipole moment. In Si and Ge, there is no dielectric anomaly around the frequency ω_0 of the type discussed in the following section. In these materials, the only significant optic effects occur when phonons combine in pairs with equal but opposite wavevector q, so that the resultant has very small wavevector that can match the infrared wavevector $k = 2\pi/\lambda$. The significant factor here is the density of states $D(\omega)$, representing the number of phonon states per unit frequency range. In Figure 2b, we show $D(\omega)$ for the diatomic lattice. This is obtained as

$$D(\omega) = D(q)\frac{\mathrm{d}q}{\mathrm{d}\omega} \qquad [7]$$

where $D(q)$ is the density of states with respect to the wavevector, which is proportional to q^2 for a three-dimensional lattice. Note that $D(\omega)$ has peaks at frequencies corresponding to the zone boundary phonon frequencies because the dispersion curves are flat in these regions. These features are known as critical points, and they have well-defined shapes. Their spectroscopic importance is discussed later.

In bulk three-dimensional crystals, both transverse and longitudinal optic and acoustic phonons occur, each represented by different branches of the phonon dispersion curves, and the transverse branches in cubic materials are doubly degenerate. In three dimensions, more elaborate theoretical models are required, such as the rigid ion model and the shell model, and improved results can sometimes be achieved by including next-nearest-neighbor atoms in the model. As an example, the phonon dispersion curves of GaAs are shown in Figure 3.

It should be noted that the dispersion curves, and hence the associated optical properties, differ for propagation in different crystallographic directions, e.g., [100], [110], and [111].

Born–Huang Optic Phonon Model

The infrared lattice properties of materials at long wavelength are usually described by using the relative permittivity ε (often called the 'dielectric constant'), which is defined through the relation

$$D = \varepsilon\varepsilon_0 E \qquad [8]$$

between the electric field E in a medium and the displacement field D (here, ε_0 is the permittivity of free space). D is the sum of the field E and the polarization P of the medium:

$$D = P + \varepsilon_0 E \qquad [9]$$

P has contributions from lattice displacements (phonons), free electrons (plasmons), and the basic atomic polarizabilities. Since these quantities depend on frequency, ε is also a function $\varepsilon(\omega)$ of angular frequency ω. The charges of the A and B sites are taken as $+Q$ and $-Q$ respectively, and their equations of motion [1] and [2] are extended to include the electric forces $\pm EQ$. As a result of these considerations, the dielectric function can be written in the form

$$\varepsilon(\omega) = \varepsilon_\infty\left[1 + \frac{\omega_\mathrm{L}^2 - \omega_\mathrm{T}^2}{\omega_\mathrm{T}^2 - \omega^2}\right] = \varepsilon_\infty\frac{\omega_\mathrm{L}^2 - \omega^2}{\omega_\mathrm{T}^2 - \omega^2} \qquad [10]$$

Here, ε_∞ is the high-frequency dielectric constant due to the atomic polarizabilities, and ω_L and ω_T are frequencies that are close to the natural vibrational frequency ω_0 that occurs in the absence of electric field effects; the difference $(\omega_\mathrm{L}^2 - \omega_\mathrm{T}^2)$ is of order $\omega_0^2 Q^2/(Ca^3\varepsilon_0)$.

Figure 4 shows a plot of $\varepsilon(\omega)$ as a function of ω. Note that $\varepsilon(\omega)$ diverges at $\omega = \omega_\mathrm{T}$, and that $\omega = 0$ at $\omega = \omega_\mathrm{L}$ ($\omega_\mathrm{L}/\omega_\mathrm{T} = 1.25$ in Figure 4). ω_T is called the transverse optic (TO) phonon frequency: it is the resonance frequency of propagation of transverse

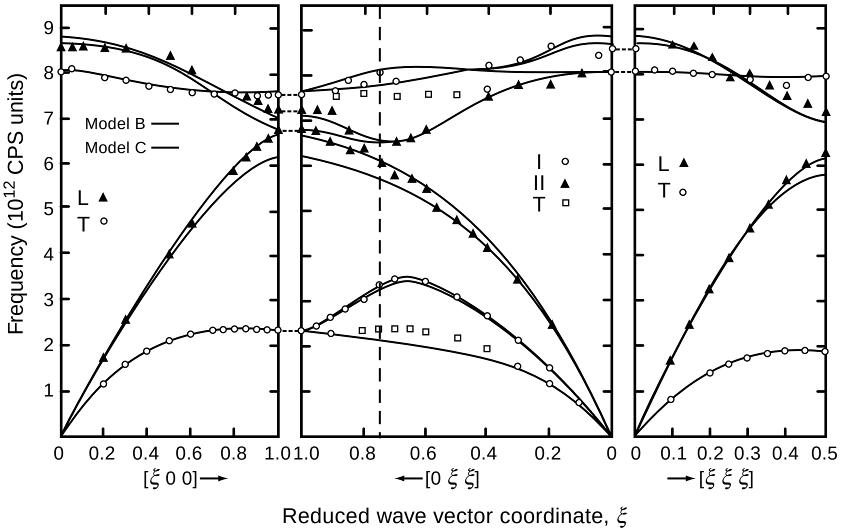

Figure 3 Phonon dispersion curves of GaAs determined by neutron spectrometry (points) compared with calculation (dotted and solid lines). Reproduced with permission from Waugh JLT and Dolling G (1963) Crystal dynamics of gallium arsenide. *Physical Review* 132: 2410.

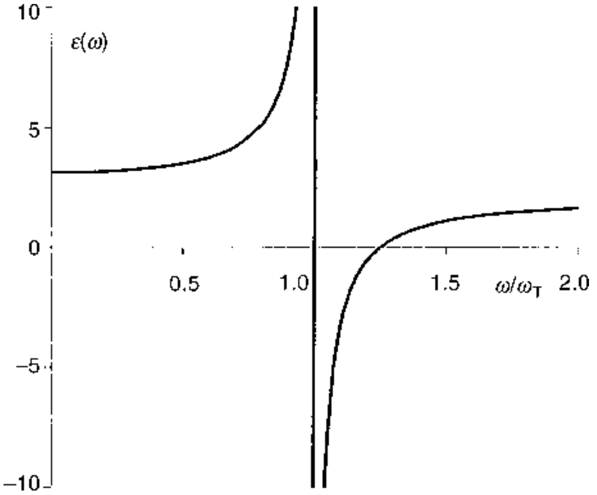

Figure 4 Dielectric function $\varepsilon(\omega)$ as a function of angular frequency, ω.

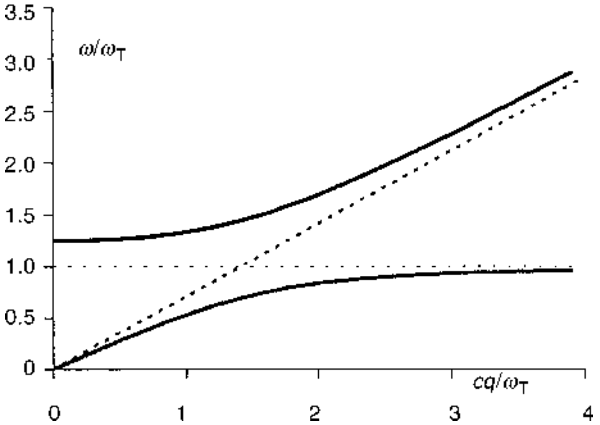

Figure 5 The dispersion curves for phonon polariton modes (solid lines) for a diatomic linear chain of atoms showing the gap in bulk phonon modes between the frequencies ω_T and ω_L.

electromagnetic waves in the material. ω_L is called the longitudinal optic (LO) phonon frequency: it is the frequency of the longitudinal electrostatic wave. Physically, the dielectric constant determines the velocity v of propagation of electromagnetic waves in the material

$$v^2 = \frac{\omega^2}{q^2} = \frac{c^2}{\varepsilon(\omega)} \qquad [11]$$

where c is the velocity of light. Equation [11] has two solutions for ω as a function of wavevector: the two solutions are shown in Figure 5. Note that there is no solution with ω in the range between ω_T and ω_L. The excitations in the region where the graphs are curved are called 'polaritons' – a combination of polar phonon and photon.

The refractive index n of the material is the ratio of v, the velocity of light in the medium, to c, the velocity in free space; thus $n = \sqrt{\varepsilon}$. n is often found from a measurement of the optical reflectivity R. At normal incidence

$$R = \left| \frac{1-n}{1+n} \right|^2 \qquad [12]$$

A graph of R is shown in Figure 6; note that the reflectivity is 100% for ω in the range $\omega_T < \omega < \omega_L$.

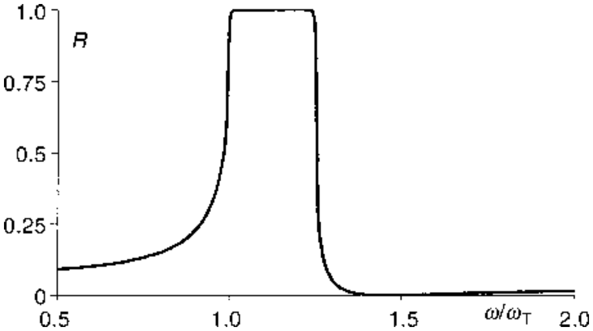

Figure 6 Calculated reflectivity plotted versus angular frequency for a diatomic linear chain of atoms for the case of zero damping.

Impurity Phonon Modes

Impurity atoms in crystal lattices can significantly affect the infrared properties. The simplest case is when an impurity atom has the same binding but a different mass. Consider the monatomic, one-dimensional case, in which the impurity atom has a mass $M + \delta M$, where M is the normal mass. Let the impurity be at the site $n = 0$. The equations of motion are

$$-M\omega^2 u_n = C(u_{n-1} + u_{n+1} - 2u_n) \quad \text{for } n \neq 0$$

$$-(M+\delta M)\omega^2 u_0 = C(u_{-1} + u_1 - 2u_0) \quad \text{for } n = 0 \qquad [13]$$

These equations have a solution provided that ω satisfies the equation

$$N^{-1}\frac{\delta M}{M}\sum_q \frac{\omega^2}{\omega_q^2 - \omega^2} = 1 \qquad [14]$$

where N is the number of atoms in the linear chain, and ω_q the unperturbed normal mode frequencies. This equation has solutions ω that lie close to the bulk phonon frequencies ω_q. However, if $\delta M < 0$ (i.e. the impurity atom is lighter than the bulk atoms), one additional mode is possible at a frequency ω that is higher than the maximum frequency $\omega_0 = \sqrt{C/M}$ of the bulk phonons. This is the so-called impurity or localized vibrational mode (LVM – see Figure 7).

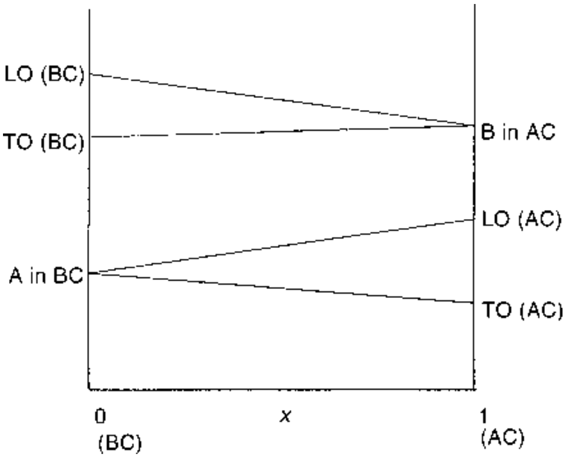

Figure 7 Representation of TO and LO phonon frequencies in a mixed crystal $A_xB_{1-x}C$ plotted versus composition *x*.

The impurity atom, being lighter than the material atoms, vibrates at a higher frequency than the bulk frequencies; the amplitude of the vibration is greatest at the impurity site, and decreases the further away one goes from the impurity site. If the impurity atom is heavier than the bulk atoms ($\delta M > 0$), there is no distinguishable local mode. In the case of diatomic lattices, an analogous situation pertains: if the impurity is lighter than the bulk atoms, an impurity mode is distinguishable above the top of the optic phonon band; if the impurity is heavier, there is the possibility that a distinct impurity mode may appear in the gap between the acoustic and optic phonon branches.

Mixed Crystals

Mixed compound semiconductor crystals are in widespread use because the bandgap can be adjusted by varying the composition. For example, the ternary material $A_xB_{1-x}C$ is an alloy of a fraction x of the material AC with a fraction $1 - x$ of the material BC. A typical example is $Al_xGa_{1-x}As$. The infrared lattice properties of these materials are classified as 'one-mode' or 'two-mode' depending on how many distinct optic phonon frequencies are observed. The distinction can easily be understood in the context of the discussion above of impurity modes. If an A atom in the bulk material BC has a distinct local impurity optic phonon mode frequency, and similarly for a B impurity in AC, then there will usually be two distinct optic phonon modes throughout the whole composition range $0 < x < 1$: one mode is AC-like, and the other BC-like, as shown in Figure 7. On the other hand, if neither A in BC, nor B in AC, provide distinct optic phonons, there is only one distinguishable optic phonon in the alloy $A_xB_{1-x}C$.

Review of Far-Infrared Properties of Semiconductors

Group IV Semiconductors

We begin with Si and Ge, both of which crystallize in the diamond structure. As mentioned earlier, there is no first-order interaction between far infrared radiation and the TO mode. However, there are several other mechanisms that result in absorption. The dipole moment of the zone center TO phonon in an ionic solid is an example of a symmetry property that, in group theory, is described by a set of selection rules for each different crystal lattice. Other mechanisms that can lead to the creation of a dipole moment are:

Disorder

Disorder destroys the translational symmetry, and hence the selection rules. Consequently, phonon modes that, in a perfect lattice, are not infrared active will become active. The most general case occurs if the crystal becomes completely disordered, in which case all phonon modes become dipole-active and the far-infrared absorption spectrum resembles the phonon density of states. This is illustrated in Figure 8, in

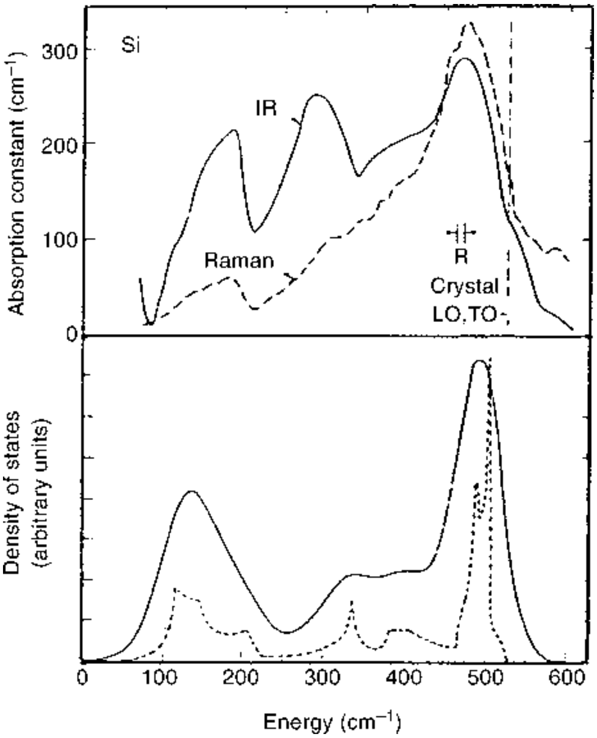

Figure 8 Top: infrared absorption coefficient (solid line) and Raman spectrum (dashed line) of amorphous Si. Bottom: calculated density of states (solid line), and broadened density of states (dashed line). Reproduced with permission from Bradskyn H and Lunio (1974) Infrared vibrational spectra of amorphous Si and Ge. *Physical Review B* 9: 1646, ADS.

which measured infrared and Raman spectra of amorphous Si are compared with the calculated one-phonon density of states.

Phonon combination bands

A combination of the lattice distortions due to two phonons excited simultaneously can produce a resultant displacement of the ions in the lattice with a dipole moment. Such phonon combinations can be excited simultaneously by a far-infrared photon, and they are observed as weak bands in the infrared absorption spectrum. Apart from the dipole moment selection rule, it is also necessary for wavevector conservation to occur. Since $q \sim 0$ for the photon, the two phonons must either be at the zone center, or have equal and opposite wavevectors. An additional consideration is that, if the process is to lead to observable absorption, it should have a high probability as determined by the density of states, so observable absorption features usually arise from critical-point phonon combinations. An example of phonon combination bands in Si is shown in Figure 9. Assignment of the features in the spectrum is usually possible by inspection of the phonon dispersion curves at the zone boundaries, a procedure known as critical-point analysis. Alternatively, the two-phonon density of states can be calculated, taking into account the selection rules, and then compared with the measured spectrum. Similar considerations apply to higher-order phonon combination bands, e.g., three or four phonons, etc., but the strength of the combination bands usually decreases rapidly as the order of interaction increases.

The presence of impurities

As mentioned above, dipole-active localized vibrational modes (LVMs) can occur if the sample is doped with lighter elements as this destroys the local translational symmetry. For instance, an LVM occurs if Si is doped with B, as shown in Figure 10. The LVM frequency is proportional to the ratio $(C/M)^{1/2}$ as seen earlier where C is the force constant and M is the isotopic mass of the substituted atom. As such modes have narrow linewidths, the features in the spectrum due to different isotopes of the same element are easily distinguished, as shown in Figure 10.

Group III–V and II–VI Semiconductors: Bulk Crystals

The group III–V and group II–VI semiconductors crystallize in either the cubic (zincblende) structure, or the hexagonal (wurtzite) structure. Many mixed crystal materials, such as $Al_xGa_{1-x}As$, are grown for

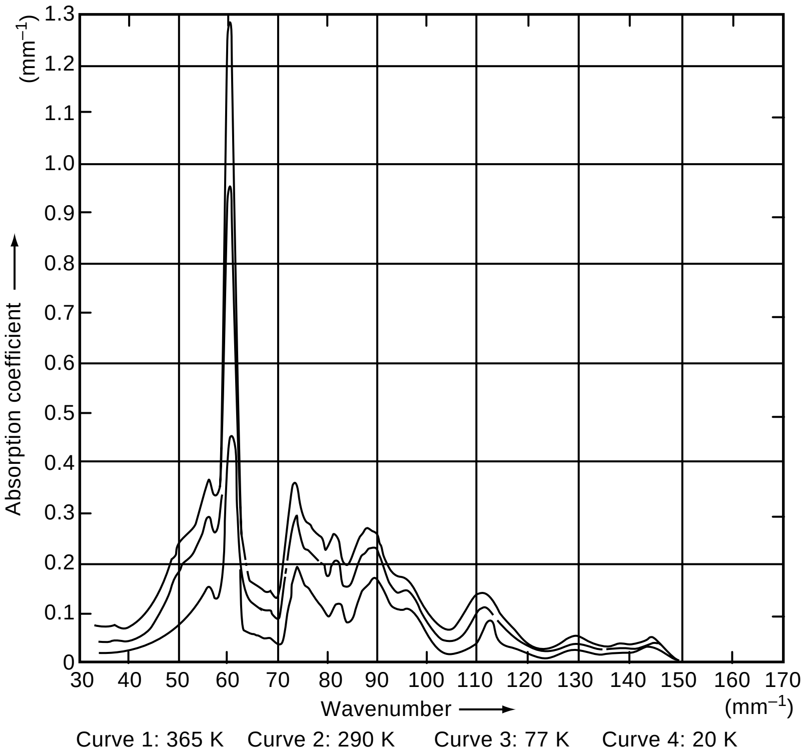

Figure 9 Measured lattice combination bands in a single crystal of silicon at 290 K (top), 77 K (middle), and 20 K (bottom). (Note that 100 mm^{-1} = 1000 cm^{-1}.) Reproduced with permission from Johnson FA (1959) Lattice absorption bands in silicon. *Proceedings of the Physical Society* 73: 265.

electronic applications that require materials with bandgaps not found naturally. Furthermore, a host of low-dimensional structures, such as epilayers, multiple quantum wells (MQWs) and superlattices, are fabricated by molecular beam epitaxy (MBE) or metalorganic vapor phase epitaxy (MOVPE) to create devices that exploit the novel properties which occur when one or more of the physical dimensions are small.

Reflectivity from a bulk single crystal

In the reststrahl frequency range bulk phonon modes cannot propagate and incident radiation is perfectly reflected, i.e., $R = 1$, as shown in the calculated spectrum in Figure 6. Figures 4 and 6 represent the ideal case of a harmonic crystal, in which there is no interaction between the phonon modes, so that all modes have infinite lifetime or, equivalently, spectral features have zero linewidth, corresponding to zero damping. This is close to the behavior of a crystal at absolute zero but, in practice, the phonon modes are coupled, so each mode will decay into combinations of other phonon modes. This leads to finite phonon lifetimes and finite linewidths of all features in the measured spectra. This damping has the effect of

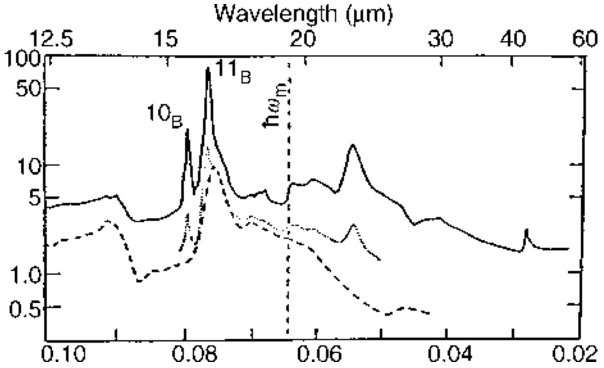

Figure 10 The far-infrared absorption spectrum of Si doped with B (^{10}B + ^{11}B). Reproduced with permission from Angress JF, Goodwin AR and Smith SD (1965) A study of the vibrations of boron and phosphorus in silicon by infrared-red absorption. *Proceedings of the Royal Society A* 287: 64.

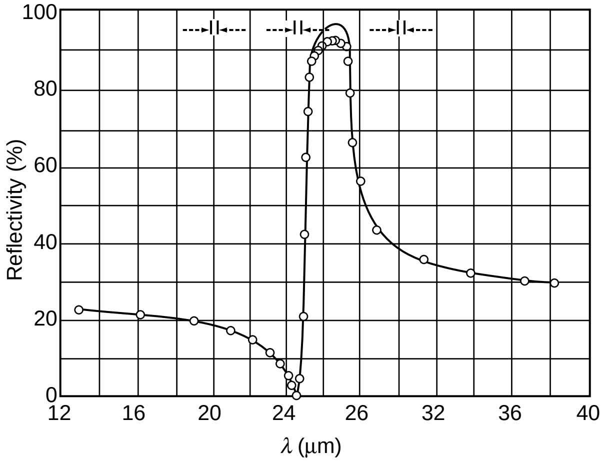

Figure 11 The reststrahlen band reflectivity of GaP, measured (points) and calculated from a damped simple harmonic oscillator (line). Reproduced with permission from Kleinman DA and Spitzer WG (1960) Infrared lattice absorption of GaP. *Physical Review* 118: 110.

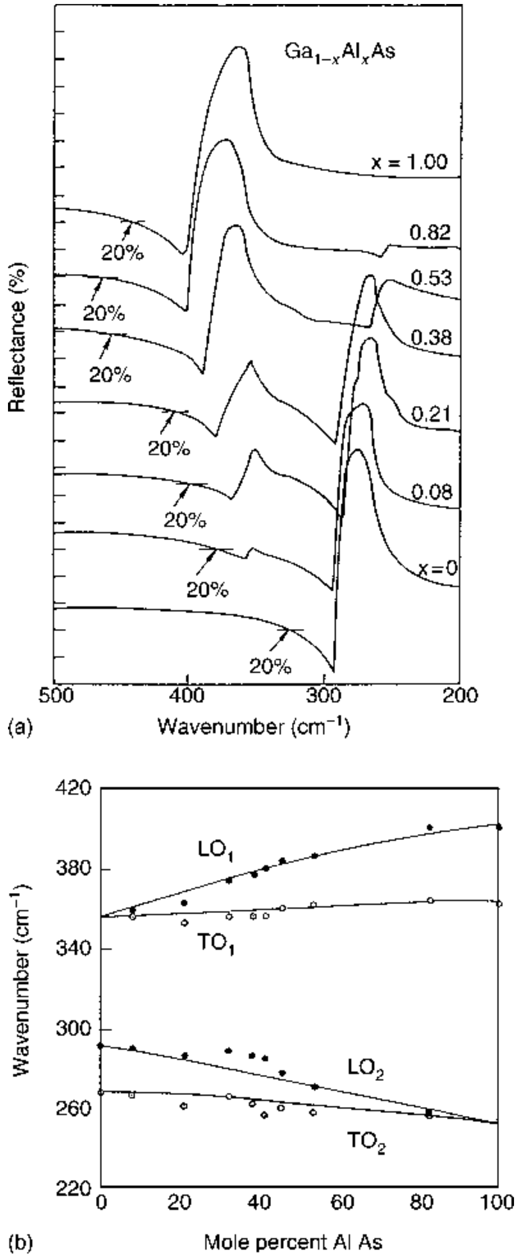

Figure 12 (a) Reflection spectra for GaAs, AlAs, and five $Ga_{1-x}Al_xAs$ mixed crystals. The curves have been displaced for clarity. (b) TO and LO mode frequencies obtained analytically from the spectra, plotted versus alloy composition. Reproduced with permission from Ilegens M and Pearson GL (1970) Infrared reflection spectra of $Ga_{1-x}AL_xAs$ mixed crystals. *Physical Review B* 1: 1576.

rounding off features in the reflectivity compared with Figure 6, and it can be represented analytically by the equation:

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{(\varepsilon(0) - \varepsilon_\infty)\omega_T^2}{\omega_T^2 - \omega^2 - i\gamma\omega} \qquad [15]$$

This differs from eqn [10] only by the addition of the damping coefficient, γ, which takes account in a phenomenological way of the interactions between the phonon modes. Here $\varepsilon(0)$ is the dielectric function at zero frequency, and $\varepsilon(0) - \varepsilon_\infty$ is the oscillator strength (i.e., the dipole strength) of the mode. We shall consider γ again later.

The reflectivity of GaP is shown in Figure 11, and this should be compared with Figure 6. In pure diatomic crystals like the simple ionic solids, e.g., NaCl and KCl, or the binary semiconductors, e.g., GaAs and CdTe, this is the only region of high reflectivity at frequencies below the visible or ultraviolet, and it is often described as the reststrahlen band, from the German term for residual ray.

Reflectivity from mixed crystals of binary semiconductors

The spectra of mixed crystals were discussed above and an example is shown in Figure 12a, which shows the reflection spectra of GaAs, AlAs and five $Ga_{1-x}Al_xAs$ mixed crystals. The TO and LO phonon frequencies were obtained from the spectra using a Kramers–Kronig analysis and classical dispersion theory. Changes in the mode frequencies with composition are interpreted using the so-called random element isodisplacement (REI) model. In Figure 12b, the TO and LO frequencies are plotted versus composition, x, and it can be seen that this system displays typical two-mode behavior, following closely the behavior shown in Figure 7.

Localized vibrational modes due to impurities

LVMs have been widely investigated in binary semiconductors because they offer a useful diagnostic tool for the investigation of impurities. In addition to the identification of isotopic species, high-resolution transmission spectroscopy can be used to determine the sites of the impurity atoms in the lattice. In Figure 13 the LVM absorption band for ^{12}C in GaAs is shown at different resolutions. The ^{12}C atoms are substitutional impurities on the As sites and the different components of the band arise from the different nearest-neighbor isotopic arrangements

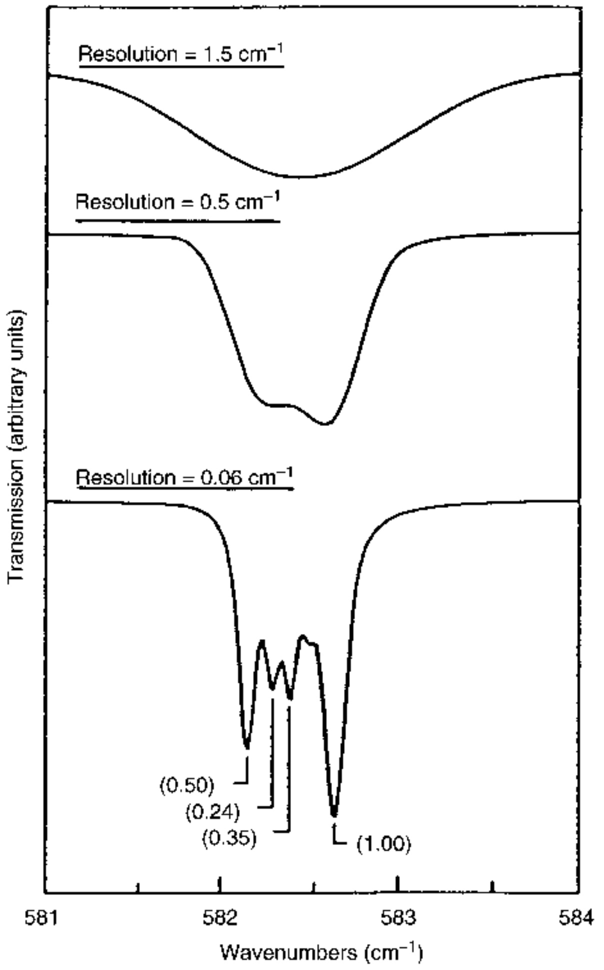

Figure 13 Localized vibrational mode (LVM) absorption band for ^{12}C in GaAs measured at 80 K. The numbers in brackets are estimated band strengths. Reproduced with permission from Theis WM, Bajaj KK, Litton CW and Spitzer WG (1982) Direct evidence for the site of substitutional carbon impurity in GaAs. *Applied Physics Letters* 41: 70.

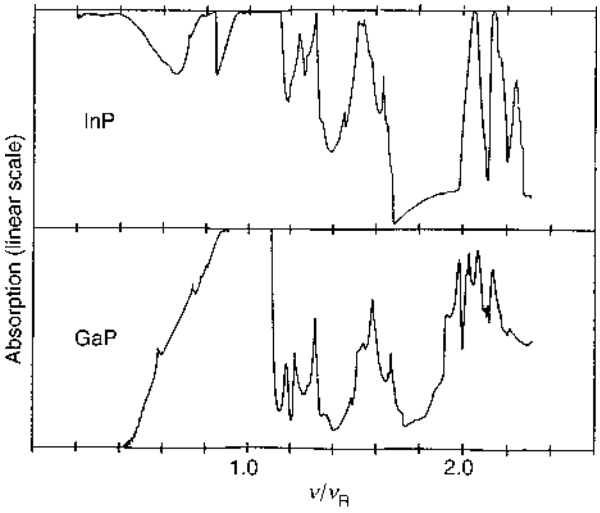

Figure 14 High-resolution far-infrared absorption spectra of GaP and InP with the frequency scale normalized in each case to the reststrahl frequency ($\nu/\nu_R = 1$ corresponds to $\omega = \omega_T$). Reproduced with permission from Koteles ES and Datars WR (1976) Two-phonon absorption in InP and GaP. *Solid State Communications* 19: 221.

surrounding the different ^{12}C atoms in the lattice. Note that ^{12}C is much lighter than either of the host lattice atoms, Ga or As, as required theoretically.

Phonon combination bands

Weak features due to the simultaneous excitation of two or more phonons are observed in the spectra of all semiconductors, and there are two excitation mechanisms. The first is direct excitation of two or more phonons by an incident photon via nonlinear terms in the dipole moment. The second process is indirect: the incident photon first excites a zone center TO phonon and the TO phonon then decays into a combination of phonons elsewhere in the Brillouin zone, a process described as anharmonic decay of the TO phonon. Mathematically, it is described by a complex frequency-dependent damping coefficient, $\gamma(\omega)$, in the oscillator equation [15]. Clearly, the latter mechanism is not available in Si and Ge as the TO phonons are not dipole-active. Anharmonicity is very important in strongly ionic crystals like the alkali halides (i.e., group I–VII compounds like NaCl), leading to prominent phonon combination bands, but it becomes progressively less important in semiconductors as the interatomic bonding becomes less ionic and more covalent, which is the case in the sequence through group II–VI (e.g., semiconductors like CdTe or ZnSe), group III–V (e.g., GaAs or InP), and group IV (Ge and Si), which are entirely covalent. Nearly all two-phonon combinations are dipole-active in zinc-blende structure crystals with the result that the transmission spectrum either side of the reststrahlen band is very rich in weak phonon combination bands, as shown in Figure 14 for GaP and InP.

Layered Semiconductor Structures

Confined optical phonons

Semiconductor superlattices are fabricated for a variety of electronic and optical applications, and the

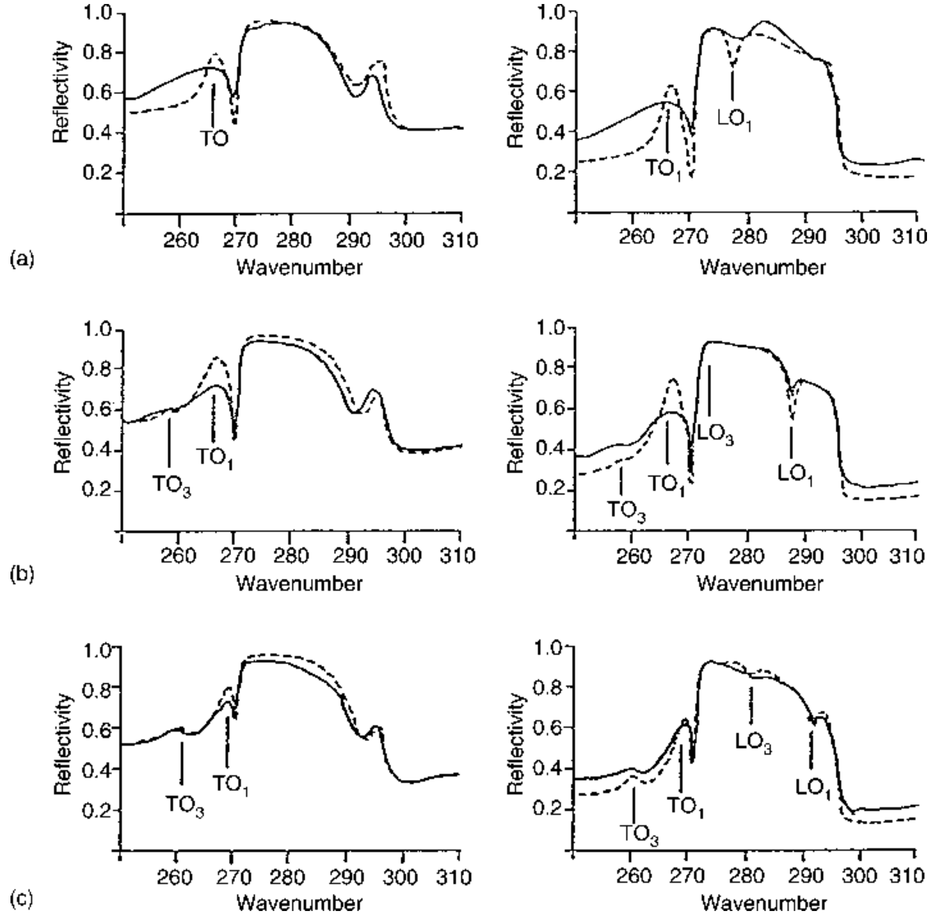

Figure 15 Measured (solid lines) and calculated (dashed lines) oblique incidence (45°) reflectivity spectra at 77 K of three short-period $(GaAs)_n/(AlAs)_n$ superlattices on GaAs substrates: n denotes the number of monolayers. Left curves: s-polarization, right curves: p-polarization. Note that the strong reststrahl reflectivity is from the GaAs substrates and the fine structure is due to confined modes. The subscripts 1, 2, and 3 denote the orders of the confined modes. Reproduced with permission from Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

quality of the interfaces in such structures is important as it determines the amount of electronic scattering at the interfaces. Some structures are fabricated from alternate layers of materials such as GaAs and AlAs, in which there is no overlap between the frequencies of the optical branches of the phonon dispersion curves. This has the effect that optical phonons excited in one type of layer, e.g., in GaAs, cannot propagate into the other type of layer, i.e., AlAs, and vice versa, so that the optical phonon modes are confined in their respective layers. The confined modes behave like standing waves, with frequencies determined by the thicknesses of the layers and the bulk optical phonon dispersion curves. Thus, if the superlattice structure is of poor quality, and the interfaces are broad due to alloying, rather than abrupt, there will be a change in the effective widths of the layers. This leads to small shifts in the frequencies of the confined phonons which can be measured and used to obtain a quantitative estimate of the amount of interface broadening. An example is shown in Figure 15. The degree of interface broadening determined from these spectra was 1.4 atomic spacings, in agreement with estimates using other techniques.

Surface phonon polaritons and guided waves

The reststrahl reflection band (Figure 6) occurs because there is a gap between ω_T and ω_L in the polariton dispersion curves (Figures 4 and 5), with no solutions for bulk phonon modes. Surface phonon polaritons can, however, propagate in this frequency band, but they have wavevectors that are larger than those, ω/c, of far-infrared radiation, so special techniques are required to observe them. The necessary wavevector enhancement can be obtained in two ways. Prism coupling can be used, as shown in Figure 16. When the angle of incidence of the incident beam at the base of the prism exceeds the critical angle for total internal reflection an evanescent wave with advanced wave vector, $q > \omega/c$, is set up at the base of the prism. This wave can couple to a surface polariton, and the energy used to excite the polariton results in a dip in the reflectivity at the frequency of the surface polariton, hence the name of the technique: attenuated total reflection (ATR). An example is shown in Figure 16. Alternatively, wavevector enhancement can be achieved by placing a grating on the surface of the sample. The incident photon wavevector is then enhanced by $+2m\pi/d$, where d is the grating period and m is an integer.

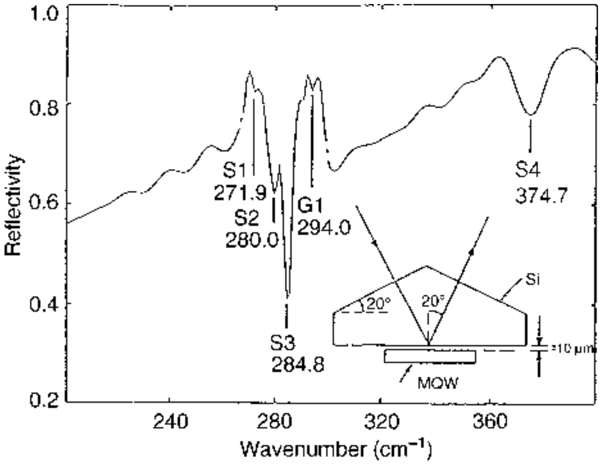

Figure 16 ATR reflection spectrum with dips due to surface phonon polaritons (S) and guided waves (G) in a GaAs/$Al_{0.38}Ga_{0.65}As$ MQW. Inset: schematic diagram of the ATR stage. Reproduced with permission from Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

Interface phonons

Under certain conditions phonons can propagate along the interface between two different media that are in intimate contact, as for instance, in the case of thin epilayers on substrates, or in multilayered

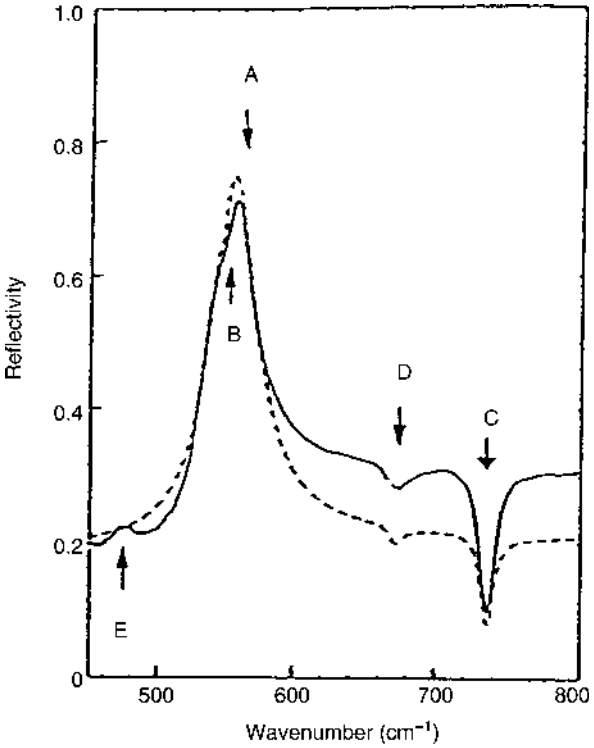

Figure 17 Far-infrared reflectivity of a GaN epilayer on a GaAs substrate at 290 K with dips due to a Brewster mode at C and a Berreman mode at D. Reproduced with permission from Mirjalili G, Parker TJ, Farjami Shayesteh S, Bulbul MM, Smith SRP, Cheng TS and Foxon CT (1998) Far infrared and Raman analysis of phonons and phonon interface modes in GaN epilayers on GaAs and GaP substrates. *Physical Review B* 57: 4656.

structures. There are two common sets of criteria for the existence of such interface modes. The first, leading to so-called Brewster modes, is when the numerator of the Fresnel reflection coefficient passes through zero due to impedance matching in the two media. The second, leading to a Berreman mode, is when the real part of the dielectric constant has a zero crossing in one of the media at a frequency where it is negative in the other medium. Both modes are observed as dips in the reflectivity. A reflection spectrum for a GaN epilayer on a GaAs substrate is shown in Figure 17. The two interface modes occur at frequencies close to the frequencies of GaN LO phonons; the Berreman mode is close to the GaN LO phonon frequency at the base of the GaN epilayer, which is shifted down by about 50 cm^{-1} from the bulk GaN value (735 cm^{-1}) due to disorder arising from the large lattice mismatch between the epilayer and the substrate. Both interface modes couple strongly to free carriers (i.e., plasmons) propagating normal to the interface, and thus act as useful probes of the structural and electronic properties of the epilayer. No other technique is available for probing the electronic properties of the epilayer for propagation normal to the layer.

See also

Semiconductor Materials: GaAs Based Compounds. **Semiconductor Physics:** Band Structure and Optical Properties; Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Angress JF, Goodwin AR and Smith SD (1965) A study of the vibrations of boron and phosphorus in silicon by infrared-red absorption. *Proceedings of the Royal Society A* 287: 64.

Barker AS Jr and Sievers AJ (1975) Optical studies of the vibrational properties of disordered solids. *Reviews of Modern Physics* 47 (Suppl. 2): S1.

Berreman DW (1963) Infrared absorption at longitudinal optic frequency in cubic crystal films. *Physical Review* 130: 2193.

Born M and Huang K (1954) *Dynamical Theory of Crystal Lattices*. Oxford: Clarendon Press.

Dumelow T and Tilley DR (1993) Optical properties of semiconductor superlattices in the far infrared. *Journal of the Optical Society of America* 10: 633.

Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

Johnson FA (1959) Lattice absorption bands in silicon. *Proceedings of the Physical Society* 73: 265.

Kittel C (1996) *Introduction to Solid State Physics*, 7th edn. New York: John Wiley.

Kleinman DA and Spitzer WG (1960) Infrared lattice absorption of GaP. *Physical Review* 118: 110.

Koteles ES and Datars WR (1976) Two-phonon absorption in InP and GaP. *Solid State Communications* 19: 221.

Mirjalili G, Parker TJ, Farjami Shayesteh S, Bulbul MM, Smith SRP, Cheng TS and Foxon CT (1998) Far infrared and Raman analysis of phonons and phonon interface modes in GaN epilayers on GaAs and GaP substrates. *Physical Review B* 57: 4656.

Newman RC (1974) *Studies of Crystal Defects*. London: Taylor and Francis.

Theis WM, Bajaj KK, Litton CW and Spitzer WG (1982) Direct evidence for the site of substitutional carbon impurity in GaAs. *Applied Physics Letters* 41: 70.

Waugh JLT and Dolling G (1963) Crystal dynamics of gallium arsenide. *Physical Review* 132: 2410.

Light Scattering

M Balkanski, Université Pierre et Marie Curie, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

When a monochromatic light beam passes through a transparent medium, a small part of the light is scattered out of the incident direction with a change in frequency. The scattering processes are usually classified into three groups. The light scattering is named:

Brillouin scattering when the light is scattered with a small frequency shift that varies continuously with scattering angle.

Raman scattering when the light is scattered with a relatively large frequency shift that is independent of scattering angle. The possibility of observing a given transition, however, depends on the orientation of the crystal relative to the polarization of the incident light.

Rayleigh scattering when the light is scattered without frequency shift. In this case the scattering is elastic with the incident and scattered frequencies equal: $\omega_\mathrm{i} = \omega_\mathrm{s}$. Rayleigh scattering is particularly useful in the study of critical phenomena or aspects related to the size and polarizability of particles.

Brillouin Scattering

Brillouin scattering is caused by the interaction of light with the acoustic modes of vibration of the crystal. Consider an acoustic wave of frequency ω_a propagating with velocity $\pm v$ as shown in Figure 1. Light with incident frequency ω_i interacts strongly with acoustic waves satisfying the Bragg condition $\omega_a = 2\omega_i(v/c)n\sin(\theta/2)$, where n is the refractive index of the material and θ is the angle of deviation of the scattered beam. Since the acoustic waves are moving with velocity $\pm v$ the scattered light suffers a Doppler shift in frequency and exhibits a frequency doublet at the angle θ given by the Brillouin equation

$$\omega_i \pm \omega_a = \omega_i \pm 2\omega_i(v/c)n\sin(\theta/2)$$

In terms of the quantum picture, Brillouin scattering corresponds to scattering of incident photons of frequency ω_i and wavevector $\mathbf{k}_i$ into scattered photons of frequency ω_s and wavevector $\mathbf{k}_s$ with the emission or absorption of an acoustic phonon of frequency ω_a and wavevector $\mathbf{q}$. The predicted Brillouin spectrum is shown in Figure 2. It consists of Stokes ($\omega_s = \omega_i - \omega_a$) and anti-Stokes ($\omega_s = \omega_i + \omega_a$) components corresponding to phonon emission and absorption, respectively.

The frequency shift $|\omega_i - \omega_s|$ for a well-defined scattering angle θ directly gives the acoustic phonon frequency and velocity. The width Γ of the Brillouin component of the spectrum is a measure of the damping or attenuation of the wave.

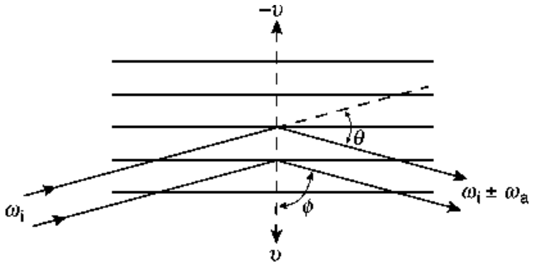

Figure 1 Light scattering at angle ϕ and Bragg reflection at angle θ by moving thermal waves.

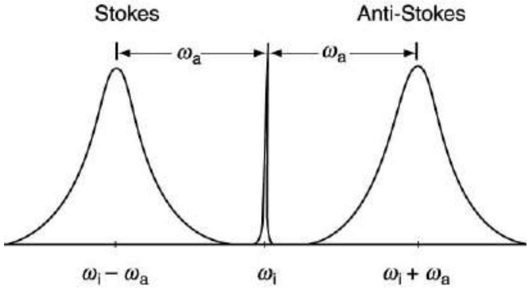

Figure 2 Schematic diagram of a Brillouin spectrum.

Raman Scattering

The essential difference between Brillouin and Raman scattering is that in Raman scattering the incident light beam is scattered with relatively large frequency shift independent of the scattering angle. The same basic considerations apply to Raman scattering that apply to Brillouin scattering. The Raman spectrum has Stokes and anti-Stokes branches corresponding to the emission and absorption, respectively, of an elementary excitation. A variety of elementary excitations are important. They include optical phonons and, in the case of magnetic materials, magnons. Also of interest are electronic excitations such as intraband single-particle excitations, interband excitations, and collective excitations (plasmons).

Raman Scattering by Phonons

An inelastic light scattering event involves the destruction of a photon of frequency ω_i incident from a light source, the creation of a scattered photon of frequency ω_s and the creation or destruction of an optical phonon of frequency ω_o.

Destruction of a photon involves interaction of the radiation field with matter by which a pair of free carriers is created in the semiconductor corresponding to a virtual transition between the valence and conduction bands. The excited state, free electron in the conduction band, interacts with the lattice via the optical deformation potential. This interaction results in the creation or annihilation of an optical phonon of frequency ω_s. When the resulting process corresponds to the creation emission of a phonon, the frequency of the scattered photon is $\omega_s = \omega_i - \omega_a$ and is referred to as the Stokes component of the spectrum. When the process is the annihilation or absorption of a phonon the scattered frequency is $\omega_s = \omega_i + \omega_a$, referred to as the anti-Stokes component.

These two first-order scattering events can be described in terms of Feynman diagrams as shown in Figure 3.

Considering these elementary processes one can calculate, using time-dependent perturbation theory, the Raman amplitude and hence the differential scattering cross-section. The remarkable feature here is that for incident frequencies approaching the frequencies of an electronic excited state of the material the Raman amplitude diverges. This leads to the phenomenon of resonant Raman scattering.

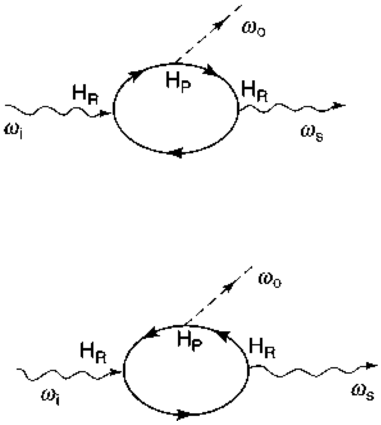

Figure 3 Feynman diagrams for two first-order scattering events.

The Raman scattering efficiency, which is the scattering cross-section per illuminated area can be calculated numerically. Numerical estimates give efficiencies typically of order 10^{-6} or 10^{-7}.

In polar semiconductors the macroscopic electric field leads to an additional contribution to the scattering cross-section for the longitudinal optical (LO) phonon modes. The macroscopic field is a consequence of the Coulomb field of the ionic charges and gives rise to an electron–phonon interaction.

There is no simple relationship between the scattering intensities for transverse optical (TO) and LO phonons in polar semiconductors.

When the probing frequency coincides with the transverse optical mode frequency, i.e., for the regime $cq \cong \omega_{\mathrm{TO}}$ (where c is the light velocity, q the photon wavevector and ω_{TO} the transverse optical phonon frequency) in polar semiconductors it is necessary to take into account the polariton nature of the coupled photon–TO-phonon modes and their associated macroscopic electrical field. The dispersion relation of the coupled photon–TO mode, called a polariton, consists of two branches: an upper branch ω_+, in which the electro-optical contribution dominates, and the lower branch ω_-, which contains essentially the mechanical contribution. This leads to a contribution to the Raman amplitude from both mechanical and electro-optical origins. An exception occurs for the lower polariton branch with $\omega = \omega_{\mathrm{TO}}$; then the electro-optic contribution is very small compared to the mechanical contribution. If $cq \gg \omega_{\mathrm{TO}}$ as in the upper polariton branch, destructive interference between the mechanical and electro-optic contributions can occur.

Selection Rules in Raman Scattering

Not all elementary excitations in semiconductors scatter light. The Raman active modes are determined by selection rules established using group-theoretical methods. The various normal modes in a given crystal correspond to various symmetries of the vibrations of the atoms in the crystal and are characterized by the irreducible representations of the space group of the crystal lattice. One can show that a normal mode can participate in a first-order Raman transition if and only if its irreducible representation is the same as one of the irreducible representations that occur in the reduction of the representation of the Raman tensor.

An important result of group theory is the rule of mutual exclusion which states that, in crystals with a center of inversion, excitations that are active in the first-order infrared spectrum are inactive in the first-order Raman spectrum, and conversely, excitations that are active in the first-order Raman spectrum are inactive in the first-order infrared spectrum. In particular, the first-order spectra due to optical phonons in NaCl are infrared active, but Raman inactive, whereas in Si they are infrared inactive, but Raman active. This difference is related to the fact that each atomic site in NaCl is a center of inversion and the active optical phonons have odd parity whereas each midpoint between two nearest-neighbor atomic sites in Si is a center of inversion and the active optical phonons have even parity.

In a crystal with the zinc blend structure, there is no center of inversion, and the crystal is both infrared and Raman active. The only non-vanishing elements of the Raman tensor $R(\hat{e})$ with $\hat{e} \| z$ are the xy and yx elements. To observe the Raman effect of an LO phonon propagating in the z-direction, one can arrange the polarization of the incident light parallel to the x-axis and observe the scattered light with its polarization parallel to the y-axis or vice versa. An analogous set of constraints applies to TO phonons. A typical spectrum is illustrated in Figure 4.

For higher-order Raman processes in which several phonons participate, it is the product of the irreducible representations of the phonons involved that must be the same as the irreducible representation of the Raman tensor.

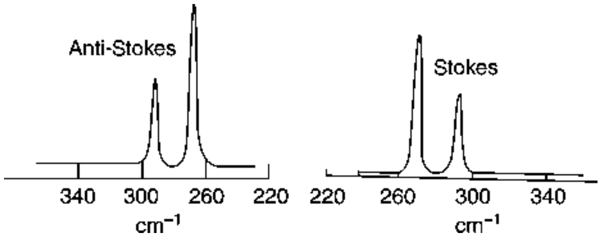

Figure 4 Raman spectrum for optical phonons.

Geometrical Aspects of First-Order Raman Scattering

The observation of light scattering by optical phonons in transparent crystals is usually done in a geometry in which the linearly polarized incident light beam is directed along, say, the x-axis and the scattered beam is observed along the y-axis (see Figure 5). When the crystal is not transparent, as is often the case in semiconductors, the observation is made in the backscattering geometry in which the scattered beam is observed in the opposite direction to the incident beam.

The scattering geometry affects the range of phonon wavevectors that is accessible in first-order Raman scattering. The condition of wavevector conservation is

$$\mathrm{K_i} - \mathrm{K_s} = \pm \mathbf{q}$$

where $\mathbf{q}$ is the phonon wavevector and the plus (minus) sign refers to the Stokes (anti-Stokes) process. The Stokes geometry is shown in Figure 6 and satisfies the following relation:

$$q^2 = k_\mathrm{i}^2 + k_\mathrm{s}^2 - 2k_\mathrm{i}\, k_\mathrm{s} \cos\theta$$

Forward scattering is characterized by $\theta = 0$ and a minimum value of q given, for isotropic media, by

$$q_\mathrm{min} = \frac{n(\omega_\mathrm{i})\ \omega_\mathrm{i} - n(\omega_\mathrm{s})\omega_\mathrm{s}}{c}$$

where $n(\omega_\mathrm{i})$ and $n(\omega_\mathrm{s})$ are the refractive indices of the crystal for the incident and scattered light, respectively, and $k_{\mathrm{i(s)}} = n(\omega_{\mathrm{i(s)}})\omega_{\mathrm{i(s)}}/c$.

Back scattering is characterized by the maximum value of q when $\theta = 180°$ and is given by

$$q_\mathrm{max} = \frac{n(\omega_\mathrm{i})\omega_\mathrm{i} - n(\omega_\mathrm{s})\omega_\mathrm{s}}{c}$$

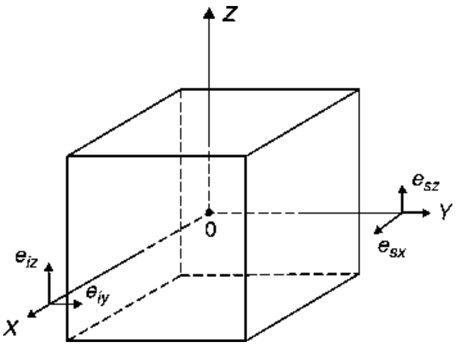

Figure 5 Geometry for light-scattering experiments in transparent crystals.

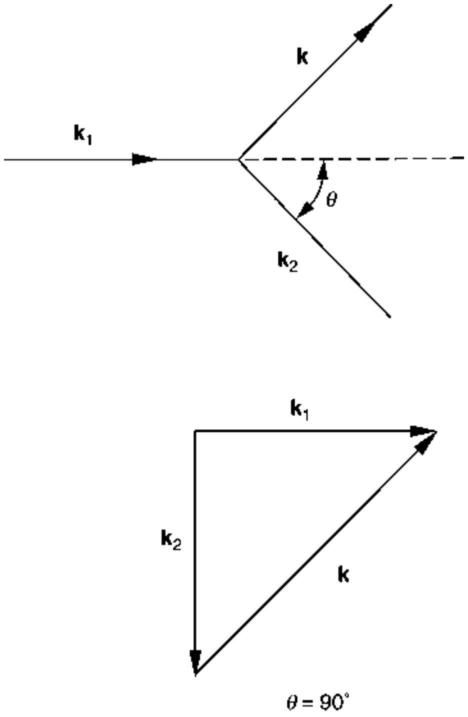

Figure 6 Diagram for Stokes processes.

For typical light-scattering experiments in the visible region the range of the incident wavevector is $0 < k_\mathrm{i} < 10^6\ \mathrm{cm}^{-1}$. This implies that for first-order scattering processes the accessible range of $\mathbf{q}$ under conditions of wavevector conservation is small compared to a nonzero reciprocal lattice wavevector. Light-scattering experiments yield the frequencies of optical modes at essentially the center of the Brillouin zone.

The energy and momentum conservation rules have to be modified when the lifetime of the crystal excitations are strongly limited by their decay into other crystal excitations. Momentum conservation breaks down in imperfect crystals, in solids lacking translational symmetry like amorphous materials, and in crystals which are opaque to incident and scattered light.

In those cases where the incident and scattered waves are damped inside the scattering volume, such as occur in small-gap semiconductors that are opaque at the light frequencies involved, $\mathbf{k}_\mathrm{i}$ and $\mathbf{k}_\mathrm{s}$ are complex. The inelastic scattering is due to excitations having a range of wavevector

$$\Delta q = |\mathrm{Im}\ \mathbf{k}_\mathrm{i}| + |\mathrm{Im}\ \mathbf{k}_\mathrm{s}|$$

about $\mathbf{q} = \mathrm{Re}(\mathbf{k}_\mathrm{i} - \mathbf{k}_\mathrm{s})$. Effects associated with such a wavevector uncertainty have been reported in

Raman scattering spectra of III–V semiconductor compounds.

Second-Order Raman Scattering

Second-order Raman scattering involves two phonons rather than a single phonon as in first-order scattering. The possible second-order processes are the following: two phonons may be created, giving a Stokes component in the scattered light, or one is created and the other destroyed, giving a Stokes or anti-Stokes component, or both may be destroyed giving an anti-Stokes component. For each of these cases there are the possibilities of a pair of first-order electron–phonon interactions or a single second-order electron–phonon interaction.

In the case of second-order Raman scattering the wavevector conservation condition is $\mathbf{q}_1 \pm \mathbf{q}_2 \cong 0$, where $\mathbf{q}_1$ and $\mathbf{q}_2$ are the wavevectors of the phonons involved. This condition places no restriction on the magnitudes of the individual wavevectors, other than $|q_1| \cong |q_2|$, which is in contrast to first-order scattering where $q \cong 0$.

The possible phonon frequency pairings associated with second-order processes are $\omega_1 + \omega_2$ (combination band), $\omega_1 - \omega_2$ (difference band), and $2\omega_1$, (overtone band). The second-order scattering spectrum covers a broad range of frequencies. The overtone spectrum provides a measure of the phonon density of states.

Resonant Light Scattering

General Formulation

Resonant Raman scattering occurs when the incident or scattered photon energy is close to the energy of an intermediate electronic state relative to the ground electronic state. Certain terms in the Raman amplitude then diverge leading to a very large scattering cross-section. Divergences occur in the Stokes spectrum when $\hbar\omega_\mathrm{i} = \hbar\omega_\mathrm{a}$, and $\hbar\omega_\mathrm{s} = \hbar\omega_\mathrm{b}$: the photon energies *resonate* with the excitation energies of the intermediate states *a* or *b*. If *a* and *b* are the same state, resonance occurs nearly simultaneously for both ω_i and ω_s, and gives rise to a particularly strong enhancement of the scattering efficiency. It is thus clear that the resonance phenomenon is specific to the nature of the intermediate state, and its investigation leads to basic information concerning the electronic states of the system. We shall now examine the resonance behavior in several cases where the intermediate states are carrier Bloch states, free excitons, or bound excitons.

Resonance at the Fundamental Bandgap

The divergent behavior of the Raman amplitude is associated with the factor $(\hbar\omega_\mathrm{i} - E_\mathrm{g})^{-1}$ where E_g is the lowest direct bandgap of the material.

Resonance at Free Exciton States

The Raman intensity as a function of the incident photon energy has a Lorentzian line shape in the range of a single intermediate state. It is centered at the resonance frequency and has a width determined by the lifetime of the intermediate state. The distinction between scattering at resonance and scattering off resonance in the wings of the Lorentzian is that at resonance the intensity is determined by the exciton lifetime, whereas off resonance it is determined by the frequency separation from resonance. Since in many instances excitons have very long lifetimes, they can lead to a very large Raman cross-section at resonance.

Resonance at Bound Exciton States

Scattering induced by the presence of impurities involving bound exciton states has the following essential features: (a) observation of sharp resonance at the energies of the bound exciton which form discrete levels below the free exciton resonance; (b) dependence upon impurity concentration; (c) involvement of LO phonons of wavevector of the order of the inverse of the impurity state radius. Resonance at bound exciton energies has been observed in impurity-induced resonant Raman scattering by LO phonons in CdS. The resonance is centered at the absorption peak for the impurity state and has approximately Lorentzian shape. If the impurity gives rise to a localized vibrational mode, bound exciton resonance in Raman scattering due to the local mode can occur.

Anharmonic Effects on Raman Spectra

In a perfect crystal whose vibrations are harmonic and which is transparent to the incident and scattered light, the Raman lineshape is a delta function centered on the optical phonon frequency. Experimentally one finds that even in crystals of very high quality, the Raman line is broadened into a roughly Lorentzian shape with a width that increases with increasing temperature. This width can be attributed to anharmonic terms in the vibrational Hamiltonian.

The temperature dependence of the line width is determined by the phonon occupation factors. As the temperature approaches zero, the phonon occupation factor approaches zero also, and the width approaches the value due to the zero-point motion of the nuclei. At temperatures above the Debye temperature, the width due to cubic anharmonicity becomes proportional to *T*. Quartic anharmonicity imparts a T^2 dependence to the width which becomes significant at very high temperatures.

In addition to line broadening, anharmonicity causes a shift of the frequency of peak intensity to lower values as the temperature increases. Contributions to the shift proportional to T and T^2 arise in the high-temperature regime from cubic and quartic anharmonicity, respectively.

Light Scattering due to Electronic Excitations

Light Scattering by Plasmons

Plasmons are similar to LO phonons in that they have a macroscopic electric field associated with them. The field modulates the electric susceptibility and gives rise to Raman scattering. The difference in frequency of the incident and scattered light is equal to the plasma frequency ω_p specified by $\omega_\mathrm{p} = \mathrm{e}^2 n/\varepsilon_0 \varepsilon_\infty m^*$, where n is the electron concentration, m^* is their effective mass, and ε^∞ is the high-frequency dielectric constant.

When the plasma frequency is close to the LO phonon frequency, there is a plasmon–phonon interaction via the macroscopic electric fields that leads to coupled modes and forces the frequencies apart.

Light Scattering Associated with Interband Transitions

The plasmons dealt with in the preceding section are an intraband type of excitation. Also of interest are interband transitions in which a carrier undergoes a transition from one band to a different band during the scattering process. An example is provided by p-type silicon in which an electron makes a transition from an occupied state in the light-hole band to an unoccupied state in the heavy-hole band via the conduction band as shown in Figure 7.

For the case of isotropic bands the scattering efficiency spectrum at 0 K has sharp edges. If the anisotropy of the valence bands is taken into account, the sharp edges of the spectrum are smoothed out.

The experimental Raman spectrum for p-type Si exhibits a broad continuum due to the electronic transitions but in the same spectral range there are also Raman active optical phonons which superpose on the electronic continuum and distort it. The resulting distortion of the spectrum is associated with resonant interference that occurs between the discrete phonon line and the electronic continuum.

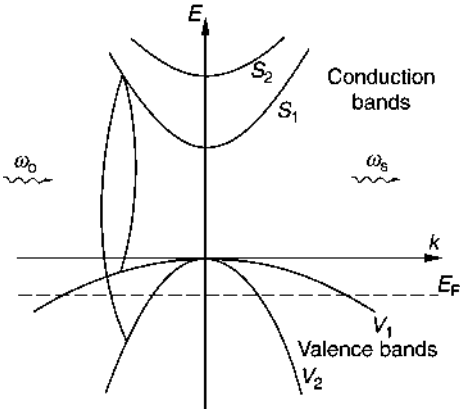

Figure 7 Electronic transitions occurring in intervalence band scattering.

See also

Scattering: Raman Scattering. **Semiconductor Physics:** Outline of Basic Electronic Properties.

Further Reading

Balkanski M (1972) In: Abeles F (ed.) *Optical Properties of Solids*. Amsterdam: North-Holland.

Balkanski M, Jain KP, Beserman R and Jouanne M (1975) *Phys. Rev.* B12: 4328.

Birman JL (1974) *Encyclopedia of Physics*, vol XXV 12b. Berlin: Springer-Verlag.

Born M and Huang K (1954) *Dynamical Theory of Crystal Lattice*. Oxford, UK: Oxford University Press.

Brillouin L (1922) *Ann. Physique* 17: 88.

Buchler S and Burstein E (1974) *Phys. Rev. Letter* 33: 908.

Daman TC and Shah J (1971) *Phys. Rev. Letter* 27: 1506.

Hayes W and Loudon R (1978) Scattering of Light by Crystals. New York: John Wiley.

Haro E, Balkanski M, Wallis RF and Wanser KH (1986) *Phys. Rev.* B34: 5358.

Kanehisa MA, Wallis RF and Balkanski M (1982) *Phys. Rev.* B25: 7619.

Wallis RF and Balkanski M (1986) Many-Body Aspects of Solid Spectroscopy. Amsterdam: North-Holland.

Yu PY, Pilkhun MH and Evangelisti F (1978) *Solid State Communications*, vol. 25, p. 371.

SEMICONDUCTOR PHYSICS

Polarons

J T Devreese, Universiteit Antwerpen, Antwerp, Belgium

© 2005, Elsevier Ltd. All Rights Reserved.

The Polaron Concept

If an electron (or hole) is placed in a polarizable medium (like an ionic crystal or a polar semiconductor), it induces a deformation of polarization field around itself. On the other hand, the electron is attracted to the potential well of this disturbed polarization field. Thus, a feedback loop between the electron and the polarization field arises, which leads to a highly correlated state of the charge carrier and the polarization field. A conduction electron (or hole) together with its self-induced polarization in an ionic crystal or in a polar semiconductor forms a quasiparticle, which is called a polaron. The polaron concept was introduced by Landau in 1933.

The physical properties of the polaron differ from those of the band carrier. In particular, the polaron is characterized by its binding (or self-) energy E_0, effective mass m^* and by its response to external electric and magnetic fields (e.g., dc mobility and optical absorption coefficient).

If the spatial extension of the polaron is large compared to the lattice parameters of the solid, the latter can be treated as a polarizable continuum. This is the case of a 'large polaron', to which most of this article is confined.

The polaron characteristics are determined by the coupling between an electron and the long-wavelength longitudinal optical (LO) phonons. The strength of this coupling is expressed in terms of the constant introduced by Fröhlich:

$$\alpha = \frac{e^2}{\hbar c}\sqrt{\frac{m_b c^2}{2\hbar\omega_{LO}}}\left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_0}\right) \qquad [1]$$

In this definition, ω_{LO} is the frequency of LO phonons, c is the velocity of light, and ε_∞ and ε_0 are, respectively, the electronic and the static dielectric constant of the polar crystal.

In order to find the coupling constant according to eqn [1], it is crucial to know the electron (hole) band mass m_b. This unknown parameter can, in principle, be determined from experiment. However, in ionic crystals and polar semiconductors, experiment measures the polaron effective mass m^* rather than the band mass, because of the polaron effect. The problem of determining the band mass is therefore far from trivial and should be tackled on the basis of the results of polaron theory. In Table 1 the coupling constants are indicated for a number of crystals.

Landau and Pekar were the first to investigate the self-energy and the effective mass of the strong-coupling polaron, when $\alpha \gg 1$. As seen from Table 1, in substances with large values of the electron–phonon coupling constant (like alkali halides, TlBr, CdF_2, $SrTiO_3$) α ranges between 2.5 and 4.5. Though, strictly speaking, polarons in such materials are of intermediate coupling, their properties can be approached within the strong-coupling approximation.

The cases when $\alpha \ll 1$ (like InSb, InAs, GaAs) are referred to as the weak-coupling regime. Polarons in many other substances with $\alpha < 1$ (e.g., GaP, CdTe, ZnSe, CdSe, CdS, CdO) are well described in the weak-coupling approximation.

In the weak-coupling regime, the states of the polaron can be imagined as those of a band carrier perturbed by the electron–phonon interaction. Fröhlich provided in 1954 the first weak-coupling perturbation theory results for the polaron characteristics:

$$E_0 = -\alpha\hbar\omega_{LO} \qquad [2]$$

$$m^* = \frac{m_b}{1 - \alpha/6} \qquad [3]$$

Feynman formulated the polaron problem in the Lagrangian (path integral) form of quantum mechanics and obtained an upper bound for the polaron self-energy at all values of the α, which at weak and strong coupling gave accurate limits. Over the years the Feynman all-coupling model

Table 1 Electron–phonon coupling constants

Material	α	*Material*	α
InSb	0.02	AgCl	1.8
InAs	0.052	KI	2.5
GaAs	0.068	TlBr	2.55
GaP	0.201	KBr	3.05
CdTe	0.286	RbI	3.16
ZnSe	0.43	CdF_2	3.2
CdSe	0.49	KCl	3.5
CdS	0.52	CsI	3.67
CdO	0.74	RbCl	3.81
AgBr	1.6	$SrTiO_3$	4.5

for the polaron has been proven to be in many respects the most successful approach to the polaron problem.

Optical Properties and Internal Structure of Polarons

Optical Absorption of Polarons at Weak Coupling

At zero temperature and in the weak-coupling limit, the optical absorption is due to the following elementary polaron scattering process, schematically shown in Figure 1. An incoming photon is absorbed by a polaron. The polaron emits a phonon during the absorption process and takes recoil energy from the incident light. At zero temperature, the absorption coefficient for absorption of light with frequency Ω can be expressed in terms of elementary functions in two limiting cases: in the region of comparatively low frequencies ($\hbar(\Omega - \omega_{\mathrm{LO}})/\zeta \ll 1$)

$$\Gamma(\omega)=\frac{1}{\varepsilon_0 nc}\frac{2^{1/2}N^{2/3}\alpha}{(3\pi^2)^{1/3}}\frac{e^2}{(\hbar m_{\mathrm{b}}\omega_{\mathrm{LO}})^{1/2}}\frac{\omega-1}{\omega^3}\Theta(\omega-1) \quad [4]$$

and in the high-frequency region ($\hbar(\Omega-\omega_{\mathrm{LO}})/\zeta \gg 1$)

$$\Gamma(\omega)=\frac{1}{\varepsilon_0 nc}\frac{2Ne^2\alpha}{3m_{\mathrm{b}}\omega_{\mathrm{LO}}}\frac{(\omega-1)^{1/2}}{\omega^3}\Theta(\omega-1) \quad [5]$$

Here $\omega=\Omega/\omega_{\mathrm{LO}}$, ε_0 is the dielectric permittivity of the vacuum, n is the refractive index of the medium, N is the concentration of polarons, and ζ is the Fermi level for electrons. The step function

$$\Theta(\omega-1)=\begin{cases}1 & \text{if } \omega>1\\ 0 & \text{if } \omega<1\end{cases} \quad [6]$$

reflects the fact that at zero temperature absorption of light accompanied by emission of a phonon can occur only if the energy of a photon is larger than that of a phonon ($\omega > 1$). In the weak-coupling limit, according to eqn [4], the absorption spectrum consists of a one-phonon line.

Experimentally, the one-phonon line has been clearly seen for free polarons in the infrared absorption spectra of CdO-films (see Figure 2). In CdO, which is a weakly polar material with $\alpha \approx 0.74$, the optical absorption band is observed in the spectral region between 6 and 20 μm (above the LO phonon frequency). As seen from Figure 2, this optical absorption band can be attributed to the weak-coupling polaron absorption as described by eqn [4].

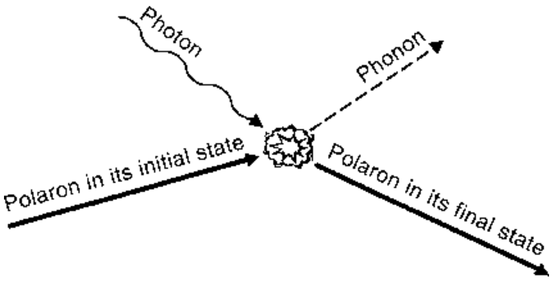

Figure 1 The elementary polaron scattering process leading to absorption of an incoming photon and to generation of an outgoing phonon.

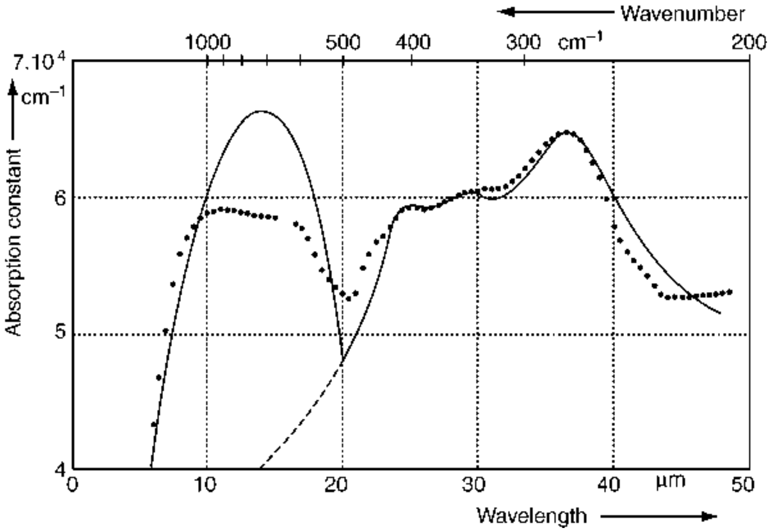

Figure 2 Optical absorption spectrum of a CdO film with the carrier concentration $N = 5.9 \times 10^{19}$ cm^{-3} at $T = 300$ K. The theoretical results are shown with (solid curves) and without (dashed curves) the polaron contribution of eqn [4], and compared to the experimental data (solid dots). (Reproduced with permission from Finkenrath H, Uhle N and Waidelich W (1969) The influence of phonons and polarons on the infrared absorption of cadmium oxide*1. *Solid State Communications* 7: 11–14.)

With strengthening coupling, it becomes possible that absorption of a photon is accompanied by emission of two, three, …, K phonons, thus giving rise to two-, three-, …, K-phonon sub-bands in the optical absorption spectra. At high temperatures, absorption of a photon can be accompanied not only by emission, but also by absorption of one or more phonons.

Optical Absorption of Polarons at Arbitrary Coupling

The theoretical absorption spectrum of a single large polaron (at all electron–phonon coupling strengths) is given by the expression

$$\Gamma(\omega) = \frac{1}{n\varepsilon_0 c} \frac{\mathrm{Im}\Sigma(\omega)}{[\omega - \mathrm{Re}\Sigma(\omega)]^2 + [\mathrm{Im}\Sigma(\omega)]^2} \qquad [7]$$

The so-called memory function $\Sigma(\omega)$ contains the dynamics of the polaron and depends on α and ω. This function is related to the impedance function $Z(\omega)$ of the polaron through $-\mathrm{i}Z(\omega) = \omega - \Sigma(\omega)$.

Optical Absorption of Polarons at Strong Coupling

The absorption of light by free large polarons was treated by Kartheuser, Evrard and Devreese using the adiabatic strong-coupling method by Landau and Pekar. The polaron ground state in this scheme has the energy

$$E_0 = -\frac{\alpha^2}{3\pi}\hbar\omega_{\mathrm{LO}} = -0.106\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [8]$$

If the lattice polarization is allowed to relax or adapt to the electronic distribution of the excited electron (which itself then adapts its wavefunction to the new potential, etc., leading to a self-consistent final state), the so-called relaxed excited state (RES) results. Its energy is

$$E_{\mathrm{RES}} = -0.041\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [9]$$

It was argued by Kartheuser, Evrard and Devreese, that for a sufficiently large α ($\alpha > 3$), the (first) RES of a polaron is a stable state, which can participate in optical absorption transitions. This idea was important in order to understand the optical absorption spectrum in the strong-coupling regime. In particular, the following scenario of a transition, which leads to a zero-phonon peak in the absorption by strong-coupling polaron, can be suggested. If the frequency of a photon is equal to

$$\Omega_{\mathrm{RES}} \equiv \frac{E_{\mathrm{RES}} - E_0}{\hbar} = 0.065\alpha^2\omega_{\mathrm{LO}} \qquad [10]$$

then the electron goes from the ground state (following the analogy with the hydrogen atom, let us call it '1s') to an excited state ('2p'), while the lattice polarization in the final state is adapted to the '2p' state of the polaron.

Each RES of the polaron can be accompanied by 1, 2, 3, … free optical phonons in the lattice. Such states are called scattering states (ScS). They are shown in Figure 3. Starting with the photon energy

$$\Omega_{\mathrm{RES}} + \omega_{\mathrm{LO}} \qquad [11]$$

a transition of the polaron towards the first ScS, belonging to the RES, becomes possible. In the course of this transition, while absorbing a photon, the polaron goes from the ground state to the RES with emission of one extra phonon. This process is called one-phonon sideband absorption.

The one-, two-, …, K-phonon sidebands give rise to the broad structure of the absorption coefficient. These transitions occur without lattice relaxation. The (unstable) polaron state, in which the lattice polarization corresponds to the electron ground state, while the electron is excited ('a frozen lattice'), is referred to as the *Franck–Condon* (FC) state of a polaron. Within the adiabatic approach, the energy of the lowest FC state is

$$E_{\mathrm{FC}} = \frac{\alpha^2}{9\pi}\hbar\omega_{\mathrm{LO}} = 0.035\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [12]$$

The superposition of the one-, two-, …, K-phonon sidebands is centered at the FC transition frequency

$$\Omega_{\mathrm{FC}} \equiv \frac{E_{\mathrm{FC}} - E_0}{\hbar} = 0.141\alpha^2\omega_{\mathrm{LO}} \qquad [13]$$

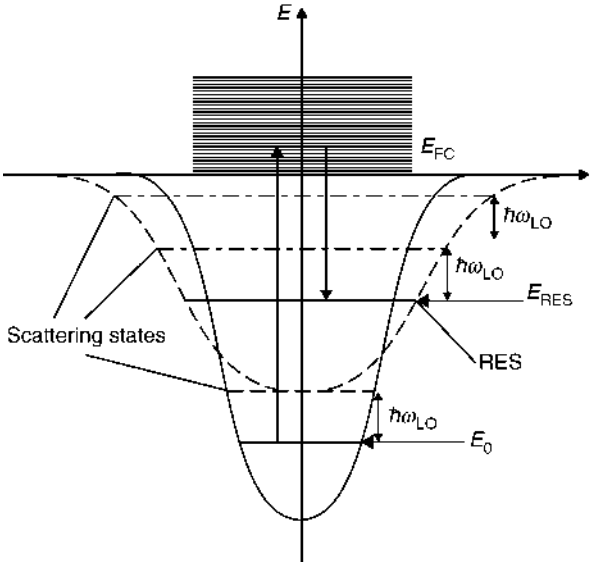
E
E_{FC}
$\hbar\omega_{\mathrm{LO}}$
$\hbar\omega_{\mathrm{LO}}$
E_{RES}
Scattering states
RES
$\hbar\omega_{\mathrm{LO}}$
E_0

Figure 3 Internal excitations of a polaron at strong coupling: E_0 – the ground state, E_{RES} – the (first) relaxed excited state; the Franck–Condon states (E_{FC}) are shown with a set of horizontal lines above the zero energy. In fact, both the Franck–Condon states and the relaxed excited states lie in the continuum and, strictly speaking, are resonances.

The optical absorption spectrum at strong coupling is, thus, characterized by the following features.

(a) There is an intense absorption peak ('zero-phonon line'), which corresponds to a transition from the ground state to the first RES at Ω_{RES}.
(b) At $\Omega_{\text{RES}} + 1$, a phonon sideband structure appears (namely, a K-phonon sideband at $\Omega_{\text{RES}} + K$). The maximum of this structure occurs at the FC transition frequency Ω_{FC}.

The qualitative behavior predicted by Kartheuser, Evrard and Devreese, namely, an intense zero-phonon (RES) line with a broader sideband at the high-frequency side, was revealed using a general all-coupling expression for the optical absorption coefficient [7] at $\alpha = 5, 6, 7$.

Internal Structure of Polarons at Arbitrary Coupling

The optical absorption spectra calculated according to the theory of Devreese *et al.* are displayed in Figure 4 for values of α in the range from 1 to 6.

The absorption spectrum for $\alpha = 1$ consists of a one-phonon line, similar to the weak-coupling result. The absorption peak for $\alpha = 3$ might correspond to a one-phonon peak at this intermediate coupling. It was also argued that part of the oscillator strength in the absorption for $\alpha = 3$ is due to lattice relaxation in the final states, be it to a smaller extent than for $\alpha = 5$.

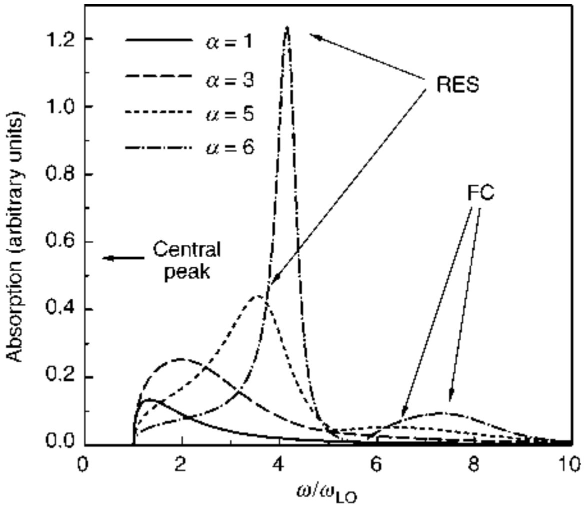

Figure 4 Polaron optical absorption spectra at zero temperature, calculated within the path integral method (according to eqns [11a], [12], [13] of the paper (Devreese JT, De Sitter J and Goovaerts M (1972) Optical absorption of polarons in the Feynman–Hellwarth–Iddings–Platzman approximation. *Physical Review B* 5: 2367–2381)) for different values of the electron–phonon coupling constant α. The peaks are labeled in the picture according to their physical origin as discussed in the text. A δ-like central peak is schematically shown by a vertical line.

It is remarkable that from [7] the three different kinds of polaron excitations appear in the calculated optical absorption spectra for polarons at $\alpha \geq 5$:

- scattering states where, e.g., one real phonon is excited (the structure starting at $\omega = 1$);
- relaxed excited state (RES);
- Franck–Condon (FC) states.

Besides that, at zero temperature, the optical absorption spectrum for one polaron contains a central peak [$\sim\delta(\omega)$] at the origin. For nonzero temperature, this central peak smears out and the optical absorption spectrum consists of both a broad envelope and an anomalous Drude-type low-frequency component.

For example, in Figure 4 the main peak of optical absorption for $\alpha = 5$ at $\omega = 3.51$ is interpreted as due to transitions to RES. The shoulder at the low-frequency side of the main peak is attributed to one-phonon transitions to scattering states. The structure centered at about $\omega = 6.3$ is attributed to a FC band. As seen from Figure 4, when increasing the electron–phonon coupling constant to $\alpha = 6$, the RES peak at $\omega = 4.3$ becomes very intense compared with the FC peak centered at $\omega = 7.5$.

The free-carrier polaron effects were revealed through measurements of optical constants, of cyclotron resonance and of mobility in semiconductors and insulating photoconductors. Experimentally, only the scattering states have been seen for free polarons. However, the full structure of eqn [7] has been revealed through cyclotron resonance measurements.

Magneto-absorption of Polarons

Polaron Cyclotron Resonance

A polaron in a magnetic field is referred to as a *magnetopolaron*. The most powerful technique to study the electron–phonon coupling in semiconductors is cyclotron resonance. Changing the magnetic field B allows one to vary the cyclotron frequency $\omega_c = eB/m_b c$. When it is brought closer to resonance with the phonon frequency ω_{LO}, a strongly renormalized magnetopolaron state is formed. This results in a splitting of the cyclotron resonance, which is a measure of the electron–phonon (polaron) coupling constant α.

A theory of the polaron effects in cyclotron resonance was proposed by Larsen. In particular, the variational approach of Larsen is based on an intermediate-coupling theory to calculate the energy levels (modified Landau levels) of a polaron in a magnetic field. The polaron mass is then defined from the energy differences between the polaron (Landau-) energy levels.

A rigorous way to find the polaron cyclotron mass is to calculate the magneto-optical absorption spectrum of the polaron (the quantity which is actually measured) and to define the polaron mass, in the same way as the experimentalists do, from the peak positions in the spectrum. The magneto-optical absorption of polarons for all α and ω_c at $T = 0$ was calculated by Peeters and Devreese. They extended the memory-function formalism described above to study the response of a polaron in a magnetic field. The magneto-absorption was then obtained from

$$\lim_{\varepsilon \to 0} \mathrm{Re} \frac{1}{z - \omega_c - \Sigma(z)} \qquad z = \omega + \mathrm{i}\varepsilon \qquad [14]$$

where $\omega_c = eB/m_b$ is the cyclotron frequency for a rigid-lattice band mass m_b, and $\Sigma(z)$ is the memory function, which takes into account all the polaron internal states as well as all the Landau levels and depends on the coupling constant α and on the magnetic field. The explicit results for $\Sigma(z)$ depend on the optimal parameters of the anisotropic Feynman polaron model, which are found by a variational calculation of the polaron ground-state energy.

At $T = 0$ and when $\omega < 1$, it follows that $\mathrm{Im}\Sigma(\omega) = 0$. According to eqn [14], the position of the cyclotron resonance line is determined by the equation $\omega - \omega_c - \Sigma(\omega) = 0$. If solutions of this equation are denoted as $\omega = \omega_c^*$ then the polaron cyclotron mass m_p^* is calculated from the relation $\omega_c^* = eB/m_p^*$.

Cyclotron Resonance of Polarons in Silver Halides

In pioneering experimental studies, Brown and coworkers have combined mobility experiments and cyclotron resonance measurements to clearly demonstrate the polaron effect in AgBr. From a theoretical plot of polaron mobility versus band mass compared to experimental Hall mobility data at a given temperature, they estimate the band mass. This allows them to calculate α and the polaron mass m^*. This value of m^* can then be compared to the measured cyclotron mass in order to test the polaron theory. For AgBr, the experimental value $m^*/m_e = 0.27 \pm 0.01$ obtained from the cyclotron resonance at millimeter wave frequencies and at 18 K is shown to agree well with the value $m^*/m_e = 0.27 \pm 0.05$ determined by comparison of mobility experiment and theory.

Precise cyclotron mass measurements in AgBr and AgCl covered the range from zero magnetic field to 16 T (see Figure 5). Several polaron theories were compared in analyzing these experimental data. It should be pointed out that the weak-coupling theories (Rayleigh–Schrödinger perturbation theory, Wigner–Brillouin and its improvements) fail (and are all off by at least 20% at 16 T) to describe the

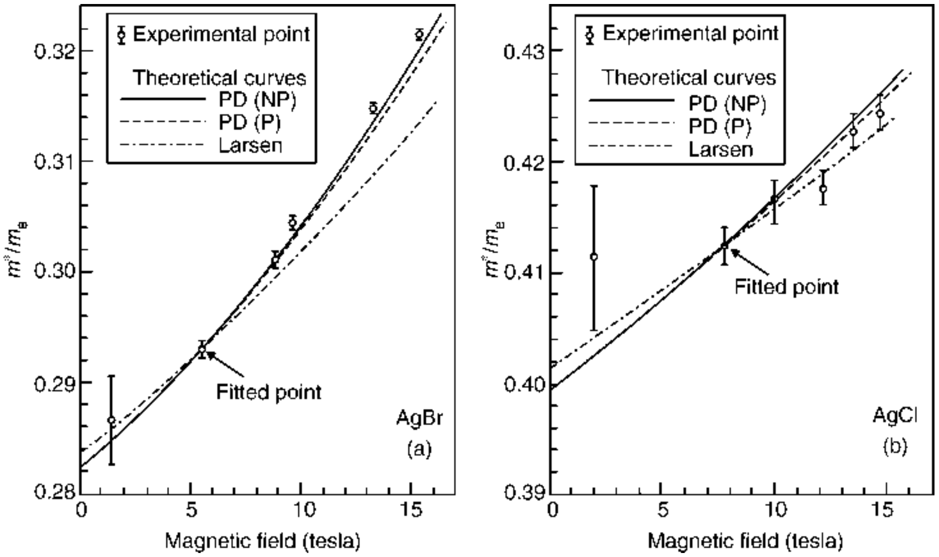

Figure 5 The polaron cyclotron mass in AgBr (a) and in AgCl (b): comparison of experiment and theory. Larsen: (Larsen D (1974) *Journal of Physics C* 7: 2877–2889); PD: (Peeters FM and Devreese JT (1986) *Physical Review B* 34: 7246–7259) (P) – with parabolic band, (NP) – with corrections of the two-band Kane model. In each case the band mass was adjusted to fit the experimental point at 525 GHz. (Reproduced with permission from Hodby JW, Russell G, Peeters F, Devreese JT and Larsen DM (1987) Cyclotron resonance of polarons in the silver halides. *Physical Review Letters* 58: 1471–1474). ©(1987) American Physical Society.

experimental data for the silver halides. The variational approach by Larsen underestimates the polaron cyclotron mass by 2% at 15.3 T. It turns out that the magneto-absorption calculated by Peeters and Devreese leads to the best quantitative agreement between theory and experiment as was analyzed for AgBr and AgCl. This analysis provides a clear confirmation of the Fröhlich description of the polaron in the case where weak-coupling approximations are adequate.

Cyclotron Resonance of Polarons in CdTe

The early infrared transmission study of hydrogen-like shallow donor impurity states in *n*-CdTe was reported by Cohn, Larsen and Lax (see Figure 6). By studying the Zeeman splitting of the $(1s \rightarrow 2p, m = \pm 1)$ transition in the Faraday configuration at magnetic fields up to ~160 kOe, they performed the first quantitative determination of polaron shifts of the energy levels of a bound electron. The experimental data were shown to be in good agreement with the weak-coupling theory of the polaron Zeeman effect. In this comparison, however, the value $\alpha = 0.4$ was used instead of $\alpha = 0.286$, which comes from the definition [1]. Similarly, the value $\alpha \sim 0.4$ was suggested to explain the measured variation of the cyclotron mass with magnetic field in CdTe. This discrepancy gave rise to a discussion where important contributions have been, in particular, due to Harper and Zawadzki.

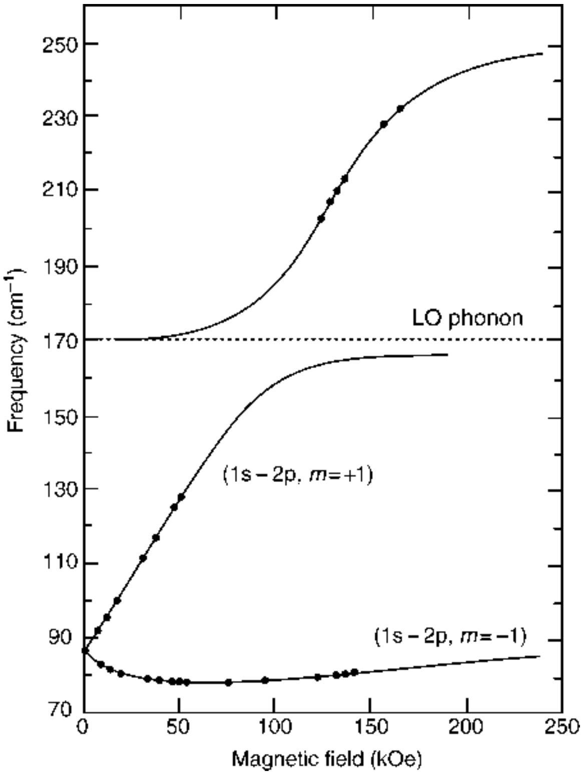

Figure 6 Plot of the experimentally determined magnetic field dependence of the $(1s \rightarrow 2p, m = \pm 1)$ transition frequencies in n-CdTe. The solid lines represent the theoretical results for the effects of the electron–LO-phonon interaction upon the shallow donor impurity levels. α treated as a fitting parameter is taken to be 0.4. (Reproduced with permission from Cohn DR, Larsen DM and Lax B (1972) Polaron Zeeman effect in CdTe. *Physical Review B* 6: 1367–1374). ©(1972) American Physical Society.

Grynberg *et al.* applied the far-infrared photoconductivity technique to study the energy spectrum of shallow In donors in CdTe layers and obtained experimental data over the energy range relevant to study the magnetopolaron effect, as shown in Figure 7. An overall good agreement is found between experiment and the theoretical approach, where the electron–phonon interaction is treated within a second-order improved Wigner–Brillouin perturbation theory and a variational calculation is performed for the lowest-lying donor states ($1s$, $2p^{\pm}$, $2s$, $2p_z$, $3d^{-2}$, $4f^{-3}$). It is to be stressed that this agreement is achieved with the coupling constant $\alpha = 0.286$ as follows from eqn [1] with the known material parameters of CdTe.

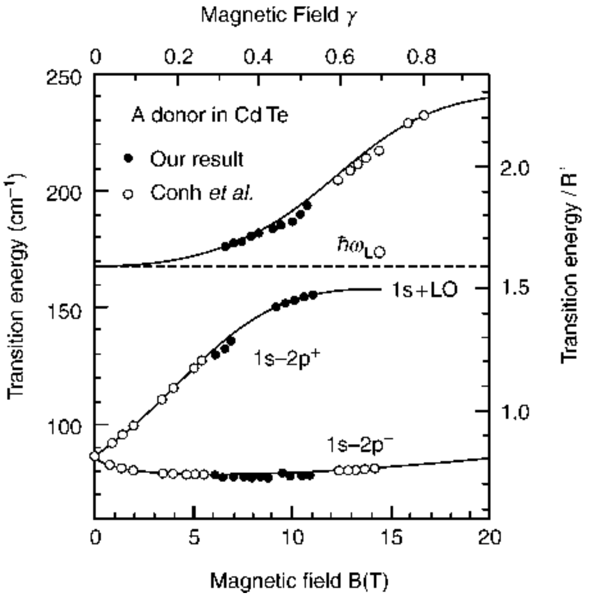

Figure 7 Plot of the experimentally determined magnetic field dependence of the $1s \rightarrow 2p^{\pm 1}$ transition energies in CdTe layers grown by molecular beam epitaxy. The solid lines represent the results of the calculation described in the text without any fitting parameters. The solid dots are the experimental data of the present work and the open circles represent the data of Cohn, Larsen and Lax. (Reproduced with permission from Grynberg M, Huant S, Martinez G, *et al.* (1996) Magnetopolaron effect on shadow indium donors in CdTe. *Physical Review B* 54: 1467–1470). ©(1996) American Physical Society.

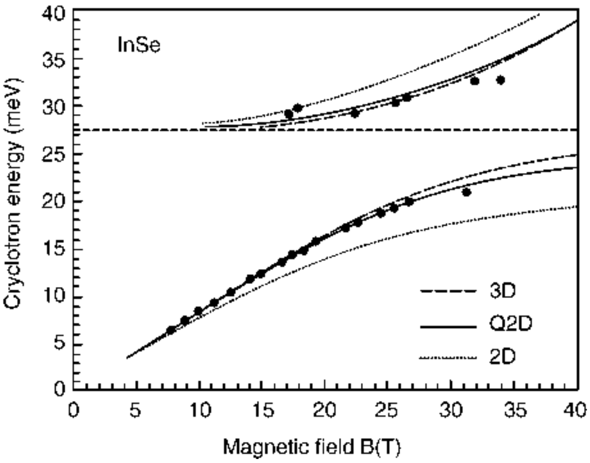

Figure 8 The cyclotron resonance position plotted as a function of magnetic field for InSe. (Reproduced with permission from Nicholas RJ, Watts M, Howell DF, *et al.* (1992) Cyclotron resonance of both magnetopolaron branches for polar and neutral. *Physical Review B* 45: 12144–12147). ©(1992) American Physical Society.

Polarons in the Layered Compound InSe and in Superlattices

Nicholas *et al.* provided a clear demonstration of the polaron coupling by the cyclotron resonance in a two-dimensional electron gas (2DEG), which naturally occurs in the polar semiconductor InSe. One clearly sees, over a wide range of magnetic fields ($B = 18-34$ T), the two distinct magnetopolaron branches separated by as much as 11 meV ($\sim 0.4\omega_{\mathrm{LO}}$) at resonance (Figure 8). The lines show the results of theoretical calculations for coupling to the LO phonon in bulk (3D), sheet (2D) and after correction for the quasi-2D systems at $\alpha = 0.29$ (for motion perpendicular to the c-axis). The agreement is reasonable for the 3D case, and is better for the quasi-2D system, where the finite spatial extent of the 2D electron gas in the symmetric planar layer is taken into account.

For GaAs/Al$_x$Ga$_{1-x}$As quantum wells and superlattices, the polaron effect is found to decrease the energy of the shallow donor states at low magnetic fields and to lead to a resonant splitting of the energies at high magnetic fields. The results are in very good agreement with available experimental far-infrared optical data in the whole magnetic-field range.

Optical Properties of Quantum Dots: Effects of the Polaron Interaction

In order to interpret the phonon-assisted optical transitions in semiconductor quantum dots, a theory was developed by Fomin, Gladilin, Devreese *et al.* comprising the exciton interaction with both adiabatic and Jahn–Teller phonons, and also the external nonadiabaticity (pseudo-Jahn–Teller effect). The effects of nonadiabaticity of the exciton–phonon system are shown to lead to a significant enhancement of phonon-assisted transition probabilities and to multiphonon optical spectra which are considerably different from the Franck–Condon progression (see Figure 9). The calculated relative intensity of the phonon satellites and its temperature dependence compare well with the experimental data on the photoluminescence of CdSe quantum dots, both colloidal and embedded in glass. These results clearly demonstrate a breakdown of the adiabatic approximation when describing the polaron interaction in quantum dots.

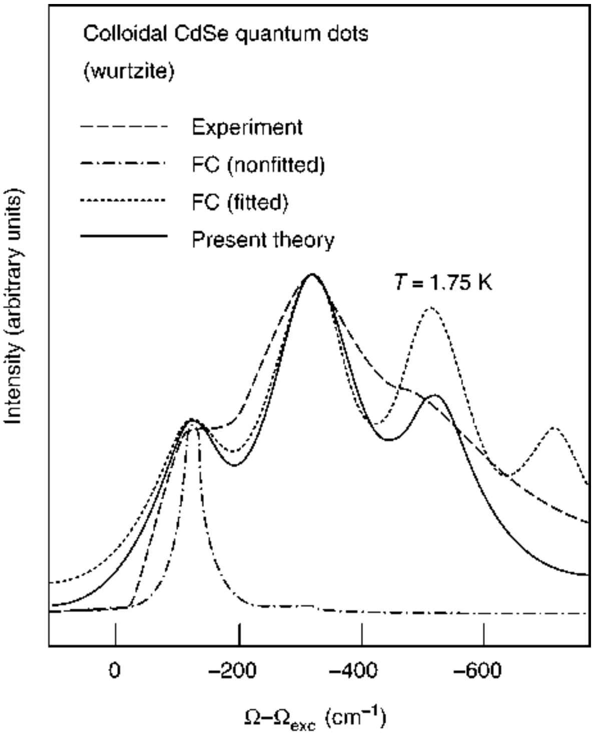

Figure 9 Fluorescence spectra of CdSe quantum dots with wurtzite structure at the average radius $\langle R \rangle = 1.25$ nm. The dashed line represents the experimental data of Nirmal M, Murray CB, Norris DJ and Bawendi MG (1993) *Z. Phys. D* 26: 361–363, the dot-dashed line displays a Franck–Condon progression with the Huang–Rhys parameter $S = 0.06$ calculated using the adiabatic approximation, the dotted line shows another Franck–Condon progression with the Huang–Rhys parameter $S = 1.7$, which is obtained by fitting the ratio of one-phonon and zero-phonon peak heights to the experimental value, and the solid line results from the nonadiabatic theory. (Reproduced with permission from Fomin VM, Gladilin VN, Devreese JT, *et al.* (1988) Photoluminescence of spherical quantum dots. *Physical Reviews B* 57: 2415–2425). ©(1998) American Physical Society.

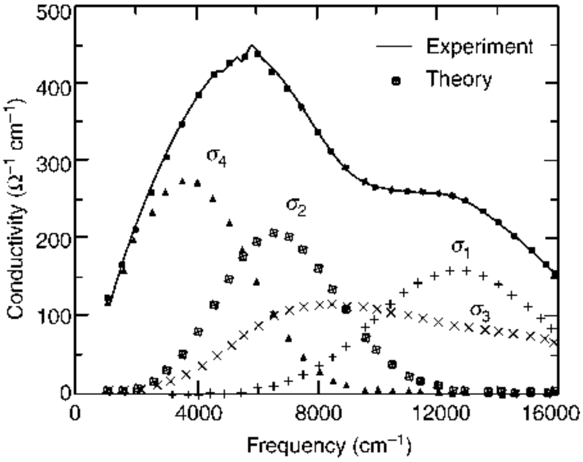

Figure 10 Experimental and theoretical conductivity data at 250 K for $Pr_2NiO_{4.22}$ and four contributions to the theoretical curves resulting from the transitions between the following polaron states: σ_1 – small → small; σ_2 – large → small; σ_3 – small → large; σ_4 – large → large. (Reproduced with permission from Eagles DM, Lobo RPSM and Gervais F (1995) Infrared absorption in oxides in the presence of both large and small polarons. *Physical Review B* 52: 6440–6450). ©(1995) American Physical Society.

Small Polarons

An electron or a hole trapped by its self-induced atomic (ionic) displacement field in a region of linear dimension, which is of the order of the lattice constant, is called a 'small polaron'. As distinct from large polarons, small polarons appear due to short-range forces.

Experimentally small-polaron effects have been analyzed, e.g., in KCl, LiF, NiO, MnO, TiO_2, $BaTiO_3$, $SrTiO_3$, $LaCoO_3$. More recently, Alexandrov and Mott (1996) surveyed both the principles and the main results of the small-polaron theory in the context of the (bi)polaronic approach to describe the physics of high-T_c superconductors.

Spectroscopic manifestations of both large and small polarons have been found by Eagles *et al.* in the infrared optical absorption spectra of $Pr_2NiO_{4.22}$. The infrared reflectivity spectra on a crystal of this material were measured between 30 and 20 000 cm^{-1}. Mid-infrared conductivity data for frequencies between 1000 and 16 000 cm^{-1} were obtained by Kramers–Kronig transformation of those reflectivity spectra. The resulting optical absorption spectrum was quantitatively interpreted (see Figure 10) using the assumption that in $Pr_2NiO_{4.22}$ two types of polarons, small and large, coexist but do not mix with each other. The contribution due to the transitions between the large-polaron states, which dominates the absorption in the low-frequency region, was calculated using the arbitrary coupling theory by Devreese *et al.* for the coupling constant $\alpha = 3$.

In summary, the polaron concept has been proven to be one of the corner-stones of modern solid-state optics.

Acknowledgments

This work has been supported by the BOF NOI (UA-UIA), GOA BOF UA 2000, IUAP, FWO-V projects G.0287.95, 9.0193.97, G.0274.01 and the W.O.G. WO.025.99N (Belgium).

See also

Semiconductor Materials: Quantum Dots.

Further Reading

Alexandrov AS and Mott N (1996) *Polarons and Bipolarons*. Singapore: World Scientific.

Devreese JT (ed.) (1972) *Polarons in Ionic Crystals and Polar Semiconductors*. Amsterdam: The Netherlands North-Holland.

Devreese JT (1996) Polarons. In: Trigg GL (ed.) *Encyclopedia of Applied Physics*, vol. 14, pp. 383–413. Weinheim: VCH.

Devreese JT, De Sitter J and Goovaerts M (1972) Optical absorption of polarons in the Feynman–Hellwarth–Iddings–Platzman approximation. *Physical Review B* 5: 2367–2381.

Feynman RP (1955) Slow electrons in a polar crystal. *Physical Review* 97: 660–665.

Kartheuser E, Evrard R and Devreese J (1969) Mechanism of absorption of light by free continuum polarons. *Physical Review Letters* 22: 94–97.

Kuper CG and Whitfield GD (eds) (1963) *Polarons and Excitons*. Edinburgh, UK: Oliver and Boyd.

Landwehr G and Rashba EI (eds) (1991) *Landau Level Spectroscopy*, 2 vols. Amsterdam, The Netherlands: North-Holland.

Peeters FM and Devreese JT (1986) Magneto-optical absorption of polarons. *Physical Review B* 34: 7246–7259.

Pekar SI (1951) *Issledovanija po Ekektronnoj Teorii Kristallov*. Moscow: Gostekhizdat (in Russian); German translation: (1954) *Untersuchungen über die Elektronentheorie der Kristalle*. Berlin: Akademie Verlag.

Quantum Wells and GaAs-Based Structures

P Blood, Cardiff University, Cardiff, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A quantum well is a potential minimum within a semiconductor structure which is sufficiently thin to localize charge carriers on a length-scale similar to their de Broglie wavelength which, for an electron in GaAs at room temperature, is about 30 nm. When electrons are localized in this way their electronic and optical properties are determined by quantum mechanical aspects of their behavior which are not apparent in larger-scale structures. The potential well is usually formed by a sandwich of a thin layer of narrow-gap semiconductor between two layers of a wider gap material as depicted in Figure 1. Typical quantum wells have widths of about 5 nm and the key to their routine production has been the development of advanced epitaxial semiconductor crystal growth techniques such as metalorganic vapor phase epitaxy (MOVPE) and molecular beam epitaxy (MBE). The most significant application of these structures has been in opto-electronic devices, especially laser diodes, and in this article I describe how electrons behave in quantum wells with particular reference to optical properties.

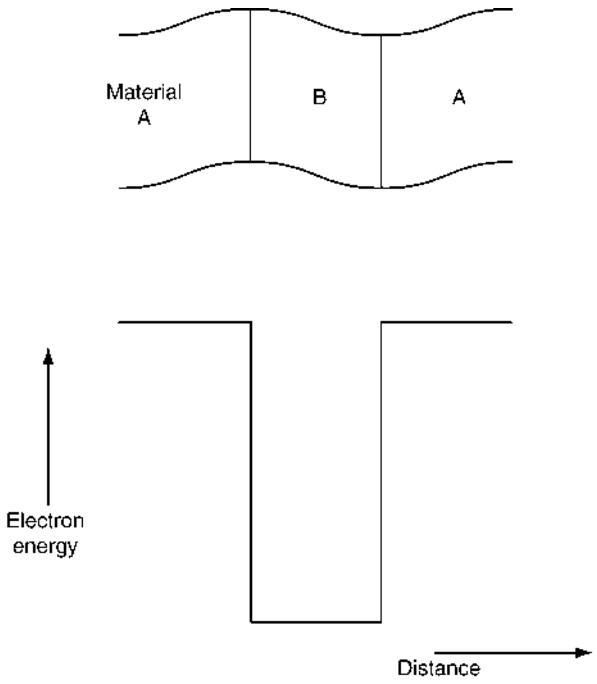

Figure 1 Electron energy diagram illustrating the formation of a potential well by sandwiching a layer of narrow gap semiconductor material (B) between two wider gap layers (A).

The optical properties of semiconductors are usually considered in terms of transitions of electrons between quantum mechanical energy levels, as in the Bohr atom. An alternative description is to consider the spatial displacement of electrons in the oscillating electric field of the electromagnetic radiation. These two descriptions are linked through Schrödinger's equation. For each energy state there is an associated wavefunction which determines the spatial probability distribution of electrons in the state. Consequently, a change in quantum state implies a change in spatial distribution and energy of the electron system. Both approaches are used: here I adopt the viewpoint of transitions between energy states.

We begin with a quantum mechanical description of electrons in a 'bulk' material, as appears in many textbooks. Here electrons are constrained by a three-dimensional potential well of extent given by the sample dimensions, typically millimeters or centimeters and therefore large compared with the de Broglie wavelength. Then we examine what happens when the size of the sample is reduced in one direction: in the limiting case of a quantum well electrons are able to move in only two directions. Quantum confinement is also possible in two or three dimensions to produce a quantum wire or quantum dot, respectively, and it is easy to use the concepts developed in this article to determine the properties of such systems.

Electron States in Bulk Material

We first consider the behavior of an electron in one direction (the x-direction) within a potential well with large dimensions (side length L), then generalize this to three directions (see Figure 2). The wavefunction ψ_n and energy E_n (measured with respect to the bottom of the well) are given by solutions of Schrödinger's equation, which within the well is:

$$\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi_n}{\mathrm{d}x^2} = E_n \qquad [1]$$

It is assumed that the potential is infinitely deep compared with the energy of the electrons.

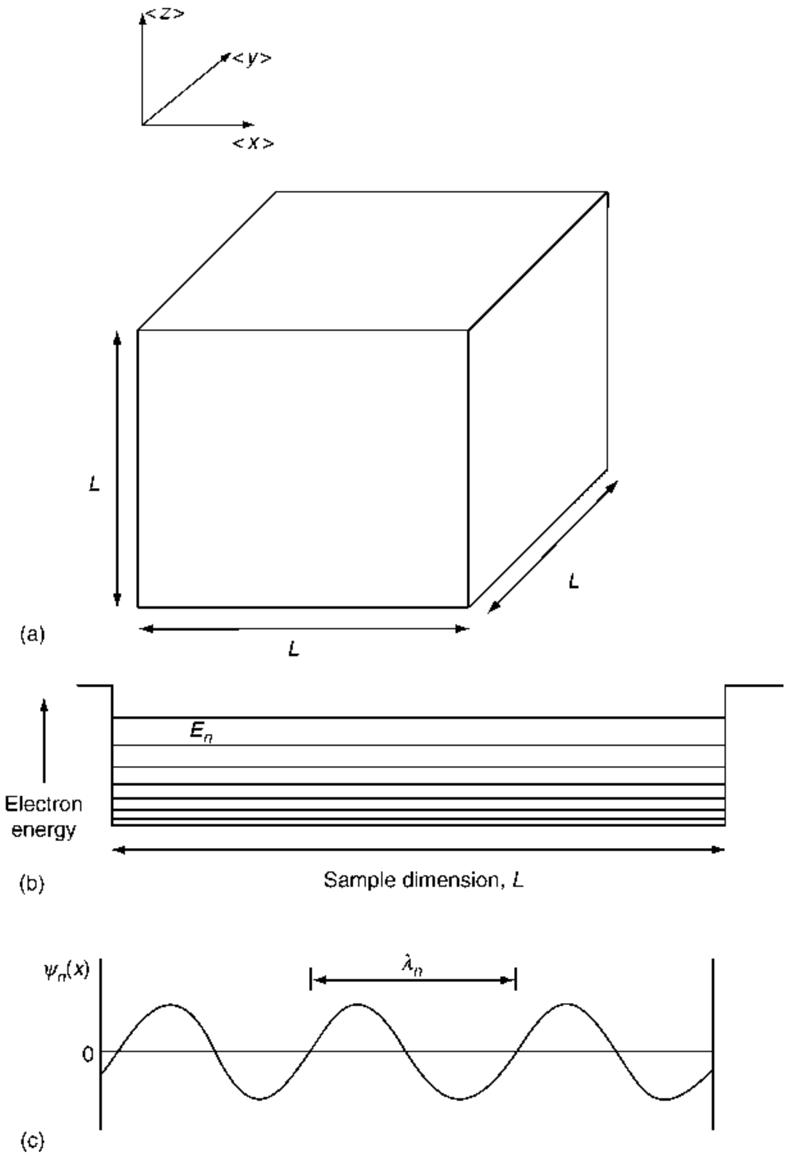

Figure 2 The properties of electrons in a large cubic sample (a) can be obtained by solving Schrödinger's equation for the potential well formed by the sample in each direction (b). This potential well keeps the electrons within the material. The wavefunctions $\psi(x)$ shown in part (c) satisfy cyclic boundary conditions as described in the text, leading to a series of electron energy levels, E_n, given by eqn [3], shown in (b) of the figure.

Within the sample the electrons are described by plane waves of the form $\psi_n = A_n \sin(k_n x)$, where k_n is a wavevector ($= 2\pi/\lambda_n$, where λ_n is the wavelength), and substitution into eqn [1] gives the corresponding energy eigenvalues as

$$E_n = \frac{\hbar^2 k_n^2}{2m} \qquad [2]$$

where m is the mass of the electron. [Substituting $E_n = k_B T$ at room temperature (0.025 eV) and using an effective mass of $0.067m_0$ (for GaAs) in eqn [2] gives $k_n = 2.1 \times 10^8\ \mathrm{m}^{-1}$ and $\lambda = 30$ nm.] Applying cyclic boundary conditions which permit traveling-wave solutions of eqn [1] the wavelength must satisfy the condition $\lambda_n = nL$, i.e., $k_n = 2\pi n/L$, as illustrated in Figure 2, where n is an integer. The dimensions of a typical sample are much greater than the 'size' of an electron so n can be a very large number. The energy eigenvalues are therefore given by

$$E_n = \frac{\hbar^2}{2m}\left(\frac{2\pi n}{L}\right)^2 \qquad [3]$$

This treatment can be extended by solving eqn [1] for motion in three orthogonal directions, x, y, and z. In this case electron motion is described by a wavevector **k** comprising components k_x, k_y, k_z, each of which satisfies the cyclic boundary condition in the respective direction. We take the sample to be of dimension L in each direction with no loss of generality. Thus (representing unit vectors by $\langle \mathbf{x} \rangle$, etc.)

$$\begin{aligned} \mathbf{k} &= k_x\langle \mathbf{x}\rangle + k_y\langle \mathbf{y}\rangle + k_z\langle \mathbf{z}\rangle \\ &= \frac{2\pi n_x}{L}\langle \mathbf{x}\rangle + \frac{2\pi n_y}{L}\langle \mathbf{y}\rangle + \frac{2\pi n_z}{L}\langle \mathbf{z}\rangle \end{aligned} \qquad [4]$$

the energy is

$$E = \frac{\hbar^2}{2m}\left\{k_x^2 + k_y^2 + k_z^2\right\} \qquad [5]$$

and the wavefunction takes the vector form

$$\psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}\cdot\mathbf{r}) \qquad [6]$$

$\psi_n^2(\mathbf{r})$ represents the probability of an electron being at the location **r**. Each allowed electron state obtained by solution of Schrödinger's equation is specified by a unique combination of the numbers $(n_x\, n_y\, n_z)$, which take both positive and negative values corresponding to plane waves traveling in positive and negative directions. The Pauli exclusion principle states that only one electron of each spin can occupy a given quantum state so the amplitude of each wavefunction is normalized such that

$$\int \psi_k^2(\mathbf{r})\,\mathrm{d}\mathbf{r} = \int [A_k \sin(\mathbf{k}\cdot\mathbf{r})]^2\,\mathrm{d}\mathbf{r} = 1 \qquad [7]$$

where the integrals are evaluated over the volume of the sample. Since the amplitude of these solutions is constant throughout the sample electrons may be anywhere in the sample with equal probability. The motion of an *individual* electron of momentum $\hbar\mathbf{k}$ is represented by a wavepacket formed by combination of a number of wavefunctions having similar values of **k**.

Because the sample dimensions are large the energies of the states allowed by Schrödinger's equation take a series of very closely spaced values for increasing integer values of n as shown in Figure 2b. Substituting a typical sample size of $L = 1$ mm into eqn [3], the $n = 1$ and 2 states are separated by only 7×10^{-11} eV in GaAs and these values are very small compared with thermal energies: at room temperature, $k_{\mathrm{B}}T = 0.025$ eV. The allowed values of **k** and E are discrete and there is a finite but very large number of allowed states in the sample. Since these are very closely spaced we can regard **k** and E as continuous variables and the allowed energy states form a continuum. We can calculate the number of allowed energy states dN in a small energy interval, dE, and hence determine the density of states in energy $g = (\mathrm{d}N/\mathrm{d}E)$ (the number per unit energy interval) at any value of energy.

Electron States in Quantum Wells

Now imagine a sample where the length of one side is reduced in the z-direction (L_z), as illustrated in Figure 3. Gradually the allowed values of k_z become more widely spaced as a consequence of the boundary conditions (eqn [4]). Eventually the width of the potential well becomes similar to the wavelength of the plane-wave state corresponding to the lowest energy level in the well. In these circumstances electron motion is not possible in the z-direction and the wavefunctions become standing waves with boundary conditions $n(\lambda_n/2) = L_z$ in an infinitely deep potential (where n_z takes only positive values) and the amplitude of the wavefunction outside the well must be zero. The lowest energy state, given by eqn [5] with $n_x = n_y = n_z = 1$, no longer lies near the bottom of the potential well. Since the energies associated with motion in the x and y directions remain very small (of order 10^{-10} eV derived earlier) because L_x and L_y are large, the energy of the lowest state is effectively determined by the z-dimension because it is very small, so eqn [5] gives

$$E_{n_z} = \frac{\hbar^2}{2m}\left(\frac{n_z\pi}{L_z}\right)^2 \qquad [8]$$

for the lowest energy state in the well. For $L_z = 5$ nm (and $m = 0.067$) E_1 is 0.22 eV above the bottom of the well. Equation [8] shows that this energy can be changed by choice of the well thickness. A simple interpretation of this behavior is as follows. When L_z becomes very small it is no longer possible for the z-component of the wavefunction of the lowest energy state to satisfy the boundary condition of zero amplitude at opposite sides of the sample. The condition can only be satisfied by wavefunctions which have a smaller wavelength, and consequently a higher energy.

Electrons can occupy states defined by all values of $(n_x\, n_y\, n_z)$, thus for $n_z = 1$ there is a continuum of allowed states at increasing energies corresponding to increasing values of $(n_x\, n_y)$ and corresponding to motion in the (x, y) plane (Figure 4a). Since L_x and L_y are both large these states are closely spaced and form a continuum. There is a similar continuum of energy

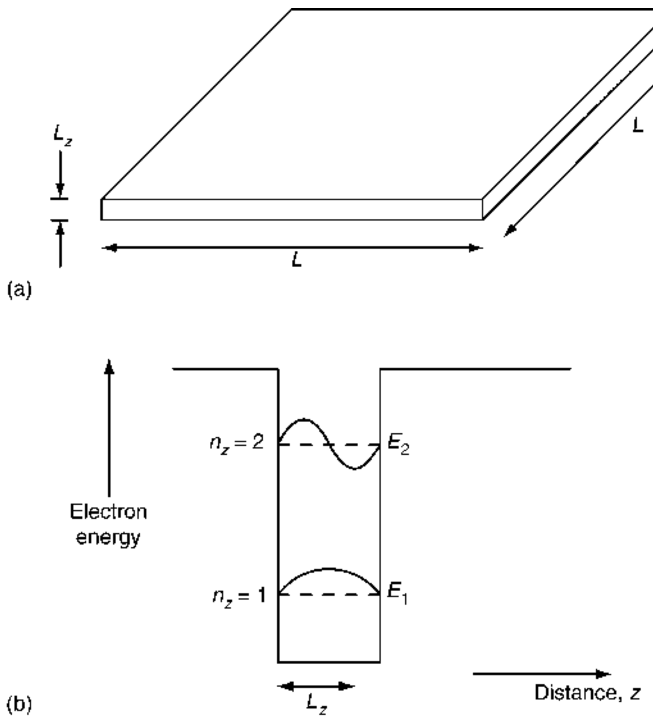
L_z
L
L
(a)
Electron energy
$n_z = 2$
E_2
$n_z = 1$
E_1
L_z
Distance, z
(b)

Figure 3 Electron energy diagram (b) for a sample (a) in which one dimension, L_z, is very small. The electron motion in the z-direction is constrained and the associated energy levels become widely spaced, defined by the integers $n_z = 1, 2, \ldots$ in eqn [8]. The wavefunctions in the z-direction, $\psi(z)$, are also illustrated for the first two electron levels.

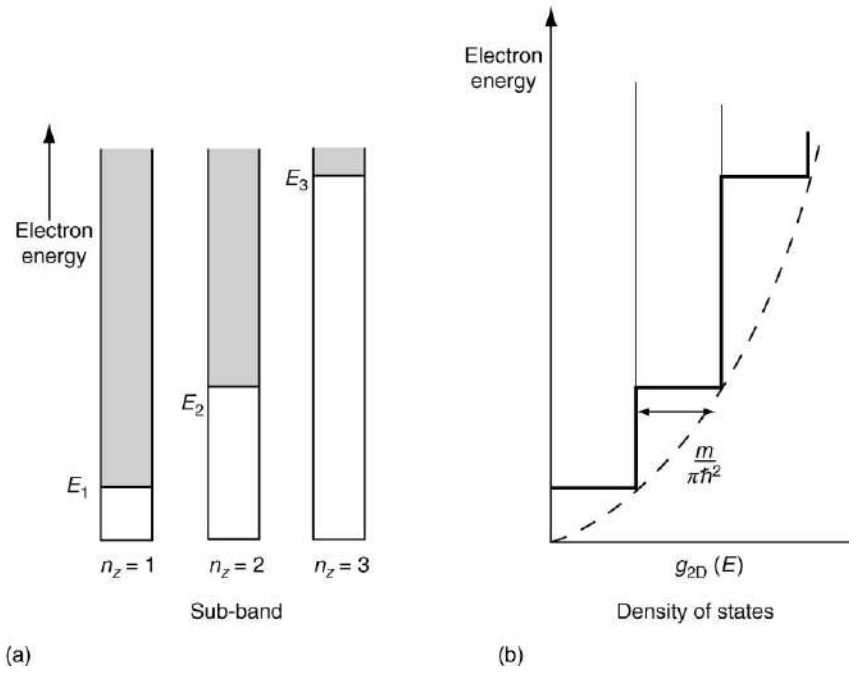
Electron energy
E_1
E_2
E_3
$n_z = 1$
$n_z = 2$
$n_z = 3$
Sub-band
(a)
Electron energy
$\frac{m}{\pi\hbar^2}$
g_{2D} (E)
Density of states
(b)

Figure 4 (a) Electron energy diagram showing the energy states associated with unconstrained motion in the (x, y) plane for each sub-band formed by localizing the electrons in the z-direction, defined by $n_z = 1, 2, 3$, etc. When all these allowed states are summed at any energy and expressed as a number of states per unit energy interval we obtain the density of states function $g_{2D}(E)$ shown in (b). This has a series of steps corresponding to each sub-band edge.

states above the energy levels for $n_z = 2, 3$, etc. Thus the energy level diagram is a series of sub-bands, each defined by n_z and with a continuum of states associated with motion in two dimensions in the (x, y) plane, as shown in Figure 4. The number of continuum energy states in a given sub-band in a small energy interval gives the density of states for both spin directions

$$g_{2\mathrm{D}}(E) = \frac{m}{\pi\hbar^2} \qquad [9]$$

per unit energy interval per unit sample area.

The density of states is independent of n_z and therefore the same for all sub-bands, and is independent of energy. The density of states function is therefore a series of steps of height given by eqn [9] at each sub-band energy as illustrated in Figure 4b. The density of states per unit area within a given sub-band is independent of L_z as a consequence of dealing with a two-dimensional system.

In real samples the quantum well is not infinitely deep: typically the well depth is in the range 100–400 meV, consequently electrons are not totally confined to the well but are able to penetrate into the barrier material by quantum mechanical tunneling. Solution of Schrödinger's equation for a finite well (treated in many quantum mechanics textbooks) again yields a series of sub-bands, finite in number and with an energy spacing smaller than that given by eqn [8]. In a typical GaAs well the lowest energy state may be about 100 meV from the bottom of the well. The density of states in each sub-band, which is due to the (x, y) motion, remains unchanged. The wavefunction in the z-direction $F(z)$ comprises a sine wave-like standing wave within the well and an exponentially decaying wave in the barrier representing tunneling (Figure 5). The wavefunction is therefore made up of (x, y) plane-wave components and the localized z component and is of the form

$$\Psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}_{xy} \cdot \mathbf{r}_{xy}) F(z) \qquad [10]$$

where $\mathbf{k}_{xy}$ and $\mathbf{r}_{xy}$ are the wavevector and position vector in the (x, y) plane. This wavefunction is again normalized according to eqn [7] such that each state is occupied by one electron of a given spin direction. We see from Figure 5 that we cannot regard the electron as being localized in the well: the electron distribution in the z-direction is specified by the probability distribution $F^2(z)$.

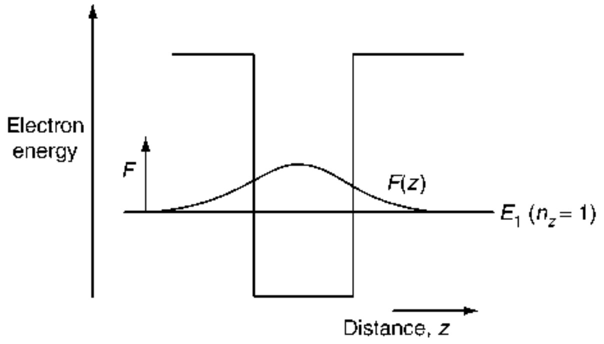

Figure 5 Illustration of the envelope wavefunction $F(z)$ in the z-direction for an electron localized in the $n_z = 1$ sub-band of a well of finite depth. The electron is able to penetrate the barrier by tunneling and this is represented by the decaying part of the wavefunction outside the well.

Occupancy of States in Quantum Wells

The density of states function describes the energy distribution of allowed states in the quantum well. Whether or not particular states are occupied by electrons is determined by the electron concentration and the temperature through the thermodynamic properties of the system. In most cases electron–electron scattering is sufficiently rapid to bring the electron population into an internal equilibrium. Furthermore the interaction between the electrons and the crystal lattice is able to redistribute energy between the two systems so that the lattice and the electron distribution have the same temperature. The electron distribution can therefore be described by Fermi–Dirac statistics for which the probability of a state at energy E being occupied with an electron is

$$f(E) = \frac{1}{1 + \exp\left\{\dfrac{E - E_\mathrm{f}}{k_\mathrm{B}T}\right\}} \qquad [11]$$

where E_f is the chemical potential (which can be equated with the Fermi level for electrons), T is the lattice temperature and all energies are positive quantities measured with respect to the same arbitrary zero. The energy distribution of electrons in the well, $n(E)$, is then the product of the density of states function, $g_{2\mathrm{D}}(E)$, and the occupation probability, $f(E)$, as depicted in Figure 6. The total number of electrons, n, is the integral over energy. Combining eqns [9] and [11], for a single sub-band gives:

$$n_a = \frac{mk_\mathrm{B}T}{\pi\hbar^2} \ell\mathrm{n}\left\{1 + \exp\left(-\frac{E_n - E_\mathrm{f}}{k_\mathrm{B}T}\right)\right\} \qquad [12]$$

per unit area, single sub-band.

We cannot specify the carrier density in the well 'per unit volume' since the extent of the distribution in the z-direction is not defined because electrons

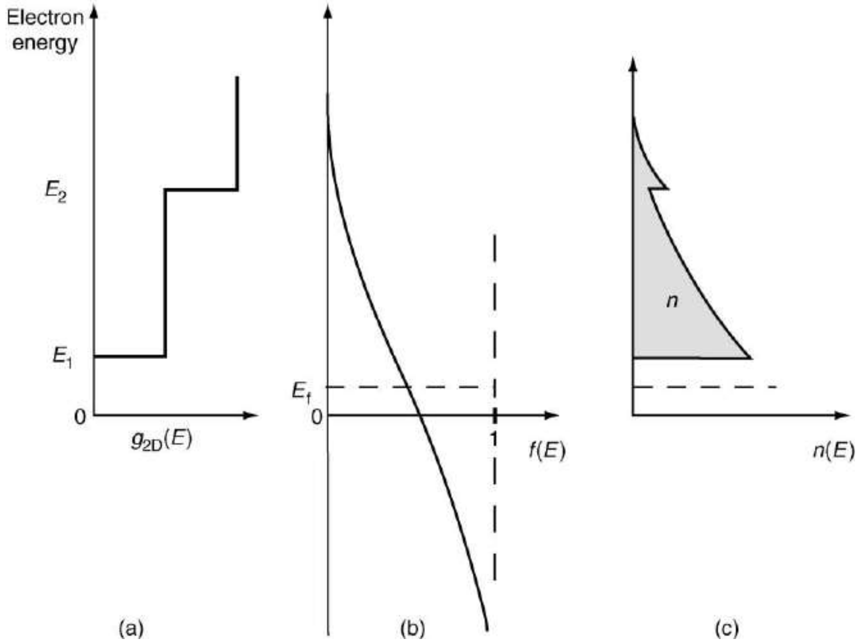

Figure 6 (a) The density of states function for a quantum well, with the value of $g_{2D}(E)$ on the horizontal axis indicating how the density of states varies with increasing electron energy. (b) is an illustration of the probability that a state is occupied by an electron, showing how $f(E)$ varies with increasing electron energy, given by eqn [11]. The actual density of electrons in a given energy interval at any energy $n(E)$ is given by the product of the density of available states and the probability that the state is occupied as shown in (c). The integral of $n(E)$ over energy gives the total density of electrons given for each sub-band by eqn [12].

may tunnel into the barrier material. The electron distribution is properly specified as a number per unit area having a probability distribution in the z-direction given by the function $F(z)$. Where more than one sub-band is populated the total electron density is obtained by adding the contributions from each sub-band each given by eqn [12] with the same Fermi energy and the appropriate sub-band energy. We always use Fermi factors to specify the probability of occupation of a state by an *electron*. The probability that a state is empty is $(1 - f)$.

Formation of Quantum Wells

A quantum well is formed by sandwiching a narrow gap semiconductor layer (E_{g1}) between two layers of wider gap material (E_{g2}). To avoid the formation of defects and dislocations, which have a deleterious effect on many properties of the structure, the constituent materials must have the same lattice parameter. Figure 7a is an energy band diagram for an n-type double heterostructure with layers of micron dimensions. At each interface there are discontinuities in the conduction and valence bands, ΔE_c and ΔE_v, respectively, which account for the difference in the band gaps of the two materials, ΔE_g. While the value of ΔE_g is known for any system, the manner in which this difference is apportioned between the two band edges is not known a priori and the values of ΔE_c and ΔE_v must be obtained from experiment. The discontinuities are often expressed as fractions Q_c, Q_v of the band gap difference:

$$\Delta E_c = Q_c \Delta E_g, \qquad \Delta E_v = Q_v \Delta E_g \qquad [13]$$

In equilibrium the Fermi energy is constant across the diagram and there is a band-bending region on each side of the discontinuity such that the conduction band edges return to their equilibrium energy values at large distances from the interface. The extent of these regions is determined by the discontinuity and the doping level and is typically 200–1000 nm. At each single interface a triangular well is formed which localizes electrons and whose precise shape is determined by the doping density, the carrier distribution and discontinuity through simultaneous solution of Schrödinger's equation and Poisson's equation. Single heterointerfaces only localize the majority carrier. Quantum confinement of both types of carrier can be achieved by a very thin narrow gap semiconductor layer.

When the layer of narrow gap material is made thinner than the band-bending regions, the

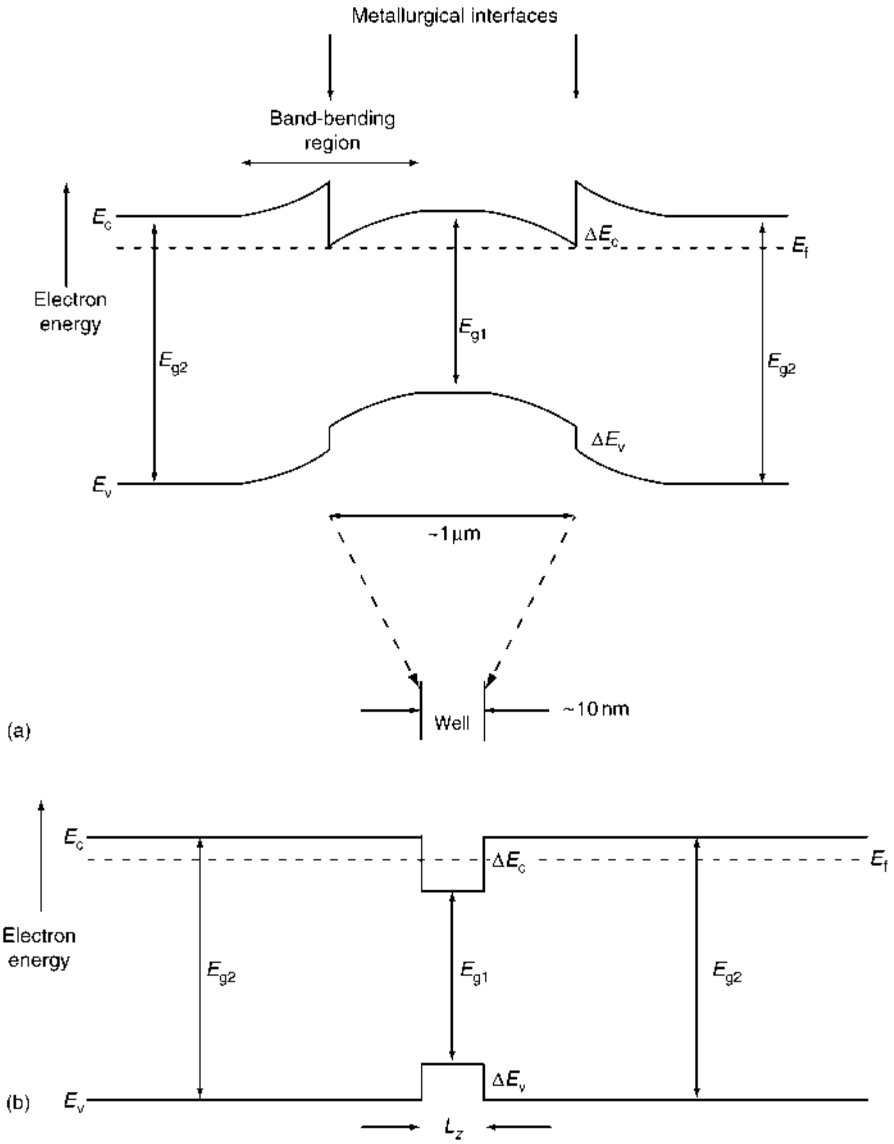

Figure 7 Electron energy diagrams showing the formation of a quantum well as a very thin double heterostructure made up of materials of band gaps E_{g1} and E_{g2}. When the layers are thick the energy diagram is influenced by band bending at the interfaces as shown in (a). When the narrow gap layer is made thinner than the band-bending distances a rectangular well is formed in the conduction and valence bands with depths determined by the respective band offset as shown in (b). On the small distance scale of the well the bands are flat in each region of the structure.

conduction band edge in this layer is not able to regain its bulk equilibrium position and the two triangular wells coalesce as illustrated in Figure 7b. When the well is sufficiently thin, say L_z less than 20 nm, the band-bending across this layer is negligibly small and the well becomes effectively rectangular. This well accumulates electrons from the surrounding barrier material and its potential relative to the Fermi level rises such that there is band bending in the barrier material on each side of the well inhibiting further accumulation. Since the band bending in the barrier occurs on a distance of hundreds of nanometers (being determined by the doping density) it is not apparent over the distance scale of the diagram of the quantum well, which is on the order of 10 nm. Consequently, as shown in Figure 7b, the well can be drawn as a rectangle to a good degree of approximation. This thin, narrow gap layer localizes both electrons and holes. The form of the well is determined by the respective band discontinuities, ΔE_c and ΔE_v, and the thickness of the layer.

Potential wells formed in this way are the basis for most quantum-confined device structures.

Other forms are possible. In certain material combinations ΔE_c or ΔE_v may be negative, meaning that, at the interface, both discontinuities are in the same sense and only one carrier type is confined in the thin narrow gap material. These are known as 'type II' structures. If there is an electric field across the well due to doping or strain (piezoelectric) effects the well becomes triangular. Wells can also be engineered with steps in the potential profile and multiple well systems can be grown in which the states are quantum mechanically coupled. All these variants provide opportunities to engineer the properties of the structure.

One important development of the rectangular quantum well deserves mention. It is possible for the well material to have a different lattice parameter to that of the surrounding barrier material provided that the strain energy can be accommodated elastically within the structure. This means that the strain energy in the layer must be below the energy necessary for formation of dislocations and this translates into an upper limit on the layer thickness (the critical thickness) for a given mismatch. Incorporation of elastic strain is significant because relaxing the lattice match requirement widens the choice of well and barrier materials and because strain modifies the properties of the electronic states in the quantum well and these effects can be used to advantage in the design of devices.

Here we concentrate on the rectangular potential well. Such wells are widely used, particularly in opto-electronic devices, and this structure provides the basic concepts which lie at the heart of all quantum-confined systems. We consider the 'model' lattice-matched quantum well system: a GaAs well bounded by AlGaAs barriers.

GaAs/AlGaAs Quantum Well Structures

As Al is substituted for Ga in the $Al_xGa_{1-x}As$ alloy system the direct band gap increases from 1.424 eV in GaAs ($x = 0$) to 3.018 eV in AlAs ($x = 1$) (Figure 8) while the lattice parameter remains unchanged. A quantum well is formed using a narrow GaAs layer sandwiched between barrier layers of $Al_xGa_{1-x}As$ having a composition chosen to give the desired well depth. The whole structure is grown on a GaAs substrate. The band discontinuity ratio between GaAs and $Al_xGa_{1-x}As$ is independent of alloy composition (ΔE_c and ΔE_v are constant fractions of ΔE_g as x is varied) with $Q_c = 0.66$ and $Q_v = 0.33$. For a typical barrier composition of $x = 0.3$,$\Delta E_g = 0.374$ eV and ΔE_c and ΔE_v are 0.247 eV and 0.127 eV, respectively.

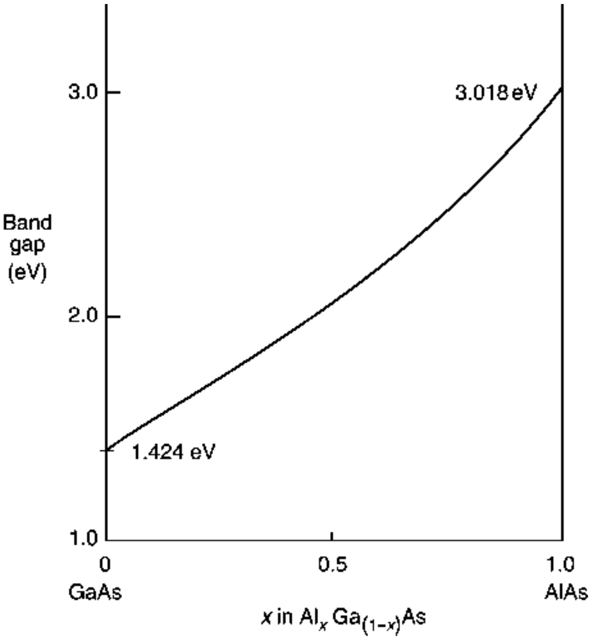

Figure 8 Variation of the direct band gap of $Al_xGa_{(1-x)}As$ with Al content (x).

Optical Properties of AlGaAs/GaAs Quantum Wells

General Principles

Figure 9 illustrates a transition of an electron between an occupied state at E_1 in the $n_z = 1$ sub-band in the conduction band and an empty state (i.e., a hole) at E_2 in the $n_z = 1$ sub-band in the valence band, resulting in emission of a photon of energy $h\nu = E_1 - E_2$ as required by energy conservation. This is the process of luminescence (or spontaneous emission), which requires external excitation such as illumination (photoluminescence) or biasing a p-n junction (electroluminescence, as in a light-emitting diode). Alternatively, incident light may promote an electron from an occupied state in the valence band to an empty state in the conduction band with the consumption of the energy of a photon. This is the process of optical absorption.

At typical temperatures the electrons and holes are at energies near to their respective band edges so the photon energy $h\nu$ is close to the separation of the $n_z = 1$ sub-bands. Since these energies are determined by the well width it is possible to change the photon energy for emission or absorption by changing the well width. As the well width is reduced the energy separation increases and the wavelength of light emitted by the structure is reduced. This ability to engineer the emission wavelength by simply adjusting

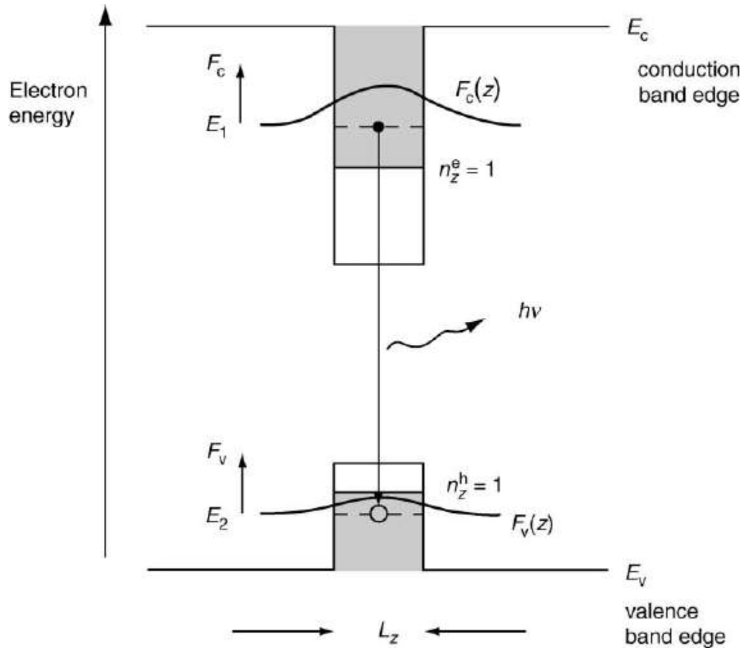

Figure 9 Illustration of a downward optical transition of an electron from a state in the $n_z = 1$ conduction sub-band to an empty state (hole) in the $n_z = 1$ valence sub-band resulting in emission of a photon of energy $h\nu = E_1 - E_2$. The spatial probability distribution of the electron and hole is specified by their respective envelope functions $F_c(z)$ and $F_v(z)$ and the probability of the transition occurring is proportional to their overlap integral, eqn [14].

the well width, without changing the chemical composition of the constituent materials, is one of the major attractions of quantum-confined structures. The shortest wavelength possible for a given material combination is determined by the band gap of the wide gap barrier material alongside the well.

The strength of the luminescence signal or the absorption is determined by the rates at which the electronic transitions occur. In general these rates are determined by the following intrinsic factors.

1. *The quantum mechanical probability of an electron making a transition between a full and an empty state*. This depends upon the 'matrix element' for the transition, which is a fundamental property of the well material and the overlap of the envelope functions. The electron and hole in the initial and final states must be in the same region of space and if the spatial distributions of electrons and holes, determined by F_c and F_v, do not overlap the probability for an electron to make a transition to the hole state is very low. The overlap is almost complete in a rectangular well (see Figure 9), but is only partial in a triangular well. The transition probability is proportional to

$$I_{\text{overlap}} = \left\{ \int F_v(z) F_c(z)\,dz \right\}^2 \qquad [14]$$

2. *The probability of occupation of initial and final states by electrons, specified by the Fermi factors* (eqn [11]). The initial state must be occupied and the Pauli exclusion principle prohibits an electron entering a final state which is already occupied. Thus for a downward transition resulting in emission of light the rate is proportional to $f_c(1 - f_v)$.
3. *The density of initial and final states* (eqn [9]). The more states there are within a given small energy interval, the greater the emission rate per unit energy.
4. *Photon distribution*. For processes such as optical absorption, which are 'induced' by the presence of a photon, the transition rate is also proportional to the photon density in the region of the well. Because the electrons are not wholly localized in the well, and because the spatial extent of the optical field is larger than that of the carriers, the coupling between the photon field (intensity $I_{\text{ph}}(z)$) and the carrier probability distributions is expressed as

$$I_{\text{coupling}} = \frac{\left\{ \int F_v(z)[I_{\text{ph}}(z)]^{1/2} F_c(z)\,dz \right\}^2}{\int I(z)\,dz} \qquad [15]$$

5. *Optical mode density*. The emission rate also depends upon the density of optical modes

available to receive the recombination radiation. In most situations the emission occurs into the cavity of free space which is very large compared with the wavelength of the radiation so the modes are very closely spaced and the cavity can accept radiation at any transition energy (this is the well-known Raleigh–Jeans theory of cavity radiation). However, in some special situations the cavity has dimensions comparable with the wavelength of the radiation (microcavities) and then the modes are widely spaced. In such a microcavity, the optical cavity effectively controls the electronic transitions.

Optical Absorption

Figure 10 is an optical absorption spectrum for a GaAs quantum well measured with light incident normal to the plane of the quantum wells. A series of steps can be seen which correspond to absorption at increasing photon energy between increasing orders of sub-bands, $E_{1e} \rightarrow E_{1h}$, $E_{2e} \rightarrow E_{2h}$, etc. Normally $n_z = 1$ to $n_z = 2$ transitions are forbidden in rectangular wells. At each step there is a pair of peaks (not always resolved) and above each step the absorption is roughly constant, following the step-like density of states function sketched in Figure 4. The peaks are due to the formation of excitons which are weakly bound electron–hole pairs. In bulk materials excitons are only observed at low temperatures whereas in quantum wells the localization increases the binding energy (E_{bX}, typically 6–8 meV) such that they are observed at room temperature. The excitonic peak is at an energy E_{bX} below the sub-band separation. There are in fact two different kinds of electrons in the valence band with different masses so there are pairs of $n_z = 1, 2, 3, \ldots$ sub-bands due to the different confinement energies for the different valence electrons. Spectra of this kind provided the first evidence for the distinctive sub-band structure of quantum well systems produced by carrier confinement. Absorption measurements provide data on the energy spacing of the sub-bands in the well, and it is possible to determine the well width and depths, from which the band offsets can also be determined.

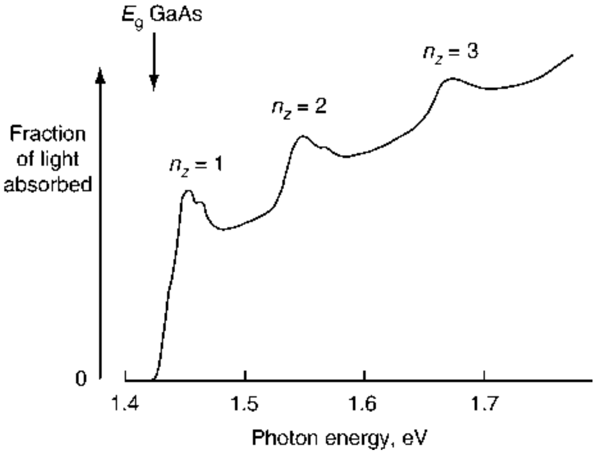

Figure 10 Absorption spectrum of a multiple quantum well structure having 10 nm GaAs wells measured at room temperature. The absorption edges corresponding to transitions between $n_z = 1, 2$, and 3 pairs of sub-bands can be identified. The peaks are due to absorption by formation of excitons.

The strength of the absorption is also of interest because it provides a measure of the optical matrix element if the Fermi factors f_c and f_v are known. If the incident optical beam is very weak such that the absorption process excites very few electrons then the system remains close to thermal equilibrium so the upper states are empty ($f_c = 0$) and the lower states are full ($f_v = 1$).

In *bulk material* the change in photon flux over a small distance Δx is proportional to the flux Φ and the distance traveled

$$\Delta\Phi = -\Phi\alpha\Delta x \qquad [16]$$

where α is the absorption coefficient which has units L^{-1}. This results in an exponential decrease of photon flux with distance traveled through the material

$$\Phi(x) = \Phi_0 \exp(-\alpha x) \qquad [17]$$

For a quantum well structure with light incident in the z-direction perpendicular to the plane of the quantum well, the fraction of light absorbed by a single well by transitions between a single sub-band pair (e.g., transitions between the $n_z = 1$ conduction and valence sub-bands) is independent of the thickness of the well because the density of states for a single sub-band is independent of well thickness. The strength of the absorption cannot be expressed by an absorption coefficient but by the fraction of light absorbed per well:

$$\Delta\Phi = -\gamma_w\Phi \qquad [18]$$

This arises because the quantum well is effectively a sheet in the z-direction: changing its thickness does not change the density of states per unit area (eqn [9]) and we cannot specify the electron concentration along the z-direction. This behavior also occurs because the well is much thinner than the wavelength of the light and it is not possible to specify the variation of photon flux on a distance scale smaller than the wavelength. The barrier material surrounding the well has a wider band gap and does not absorb at the photon energies absorbed by the well. The well thickness does affect the sub-band spacing and at a fixed photon energy the absorption increases as the well width is decreased as light is absorbed by transitions between more sub-bands. Equation [18] applies individually to each sub-band because the density of states is the same for each sub-band.

In a multiple well system, the fraction of light transmitted through N independent wells is

$$T_N = \{1 - \gamma_w\}^N \qquad [19]$$

so the fraction of light absorbed by such a system is

$$\frac{\Delta\Phi}{\Phi} = -\left[1 - \{1 - \gamma_w\}^N\right] = -N\gamma_w \qquad [20]$$

which is proportional to the number of wells when $\gamma_w << 1$.

The fraction of light absorbed by transitions between a single pair of sub-bands for a rectangular GaAs quantum well is predicted to be about 0.01, and a similar value has been obtained from absorption measurements. In a rectangular well the envelope function overlap is complete; however, in a triangular well of the same material the absorption is reduced by the smaller overlap of the envelope functions.

Photoluminescence and Spontaneous Emission

We turn now to discuss the optical emission from quantum wells as a result of external excitation. An external light source with photon energy above the effective band gap produces excess electron–hole pairs by excitation of electrons from the valence band. These rapidly lose their excess energy and thermalize to the sub-band edges to take up Fermi energy distributions. The electrons subsequently recombine to vacant valence band states by spontaneous emission. The low energy edge of this spectrum corresponds to the sub-band separation and the shape of the spectrum at higher energies corresponds to the thermal distribution of carriers in the bands. At very low excitation intensities where the excess electron population is small the luminescence may be due to recombination of electrons within an exciton and occur at an energy E_{bX} below the sub-band separation. Generally the sub-band separation is several $k_B T$ so only the lowest sub-band is populated and photoluminescence is not seen from higher sub-bands. In this respect absorption measurements provide a more complete characterization of the structure.

While the energy position of features in the photoluminescence spectrum provides useful information, it is difficult to use the strength of the photoluminescence signal to provide quantitative information about the rate of optical transitions within the material because the excitation power and volume within the sample must be known and a known fraction of emitted light must be collected and measured with a calibrated detector. The relative intensity of photoluminescence under standard conditions does provide comparative information and the linewidth of the spectrum at low intensity provides a qualitative indication of 'quality' (e.g., well width variations in the structure). A quantitative determination of the internal recombination rate can be obtained by measuring the decay of the luminescence signal following short-pulse excitation. Fuller discussion can be found in books dealing with the characterization of semiconductor structures.

Concluding Remarks

We have shown that an understanding of the electronic structure of quantum wells follows naturally from quantum mechanical treatments of electrons in bulk materials. When one of the dimensions of the structure is reduced to be similar to the de Broglie wavelength, quantization of the electron motion in this direction becomes apparent. The sub-bands which are formed have a profound effect on the electronic and optical properties when their spacing exceeds thermal energies such that only one band is significantly populated with carriers. The density of states function is a series of steps, each corresponding to a sub-band, and this produces characteristic step features in the optical absorption spectrum. The effective band gap of the structure is controlled by the width of the well. Quantum confinement also increases the binding energy of excitons compared with bulk materials and under low excitation conditions absorption and luminescence spectra are dominated by excitonic features even at room temperature.

One of the most successful application areas of quantum well structures has been in laser diodes. Figure 11 shows experimental data for laser emission wavelengths as a function of well width, showing the dramatic reduction which can be achieved simply by changing the width of the quantum well. Furthermore the step density of

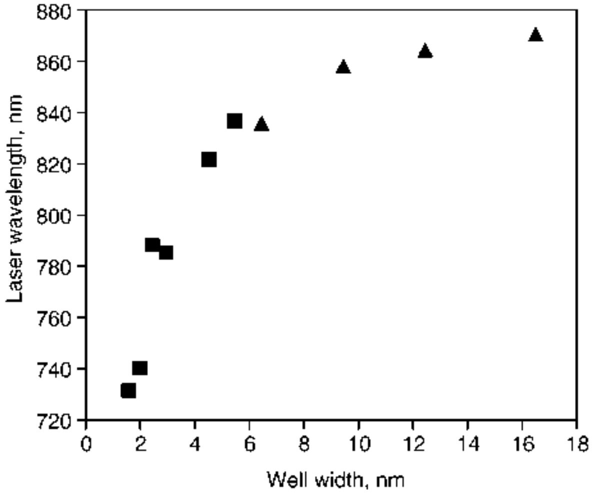

Figure 11 Measured values of the emission wavelength of GaAs/AlGaAs quantum well lasers as a function of well width, showing how the wavelength can be engineered by choice of the well width. Data from: squares: Woodbridge K, Blood P, Fletcher ED and Hulyer PJ (1984) Short wavelength (visible) GaAs quantum well lasers grown by molecular beam epitaxy. *Applied Physics Letters* 45: 16–18; and triangles: Chen HZ, Ghaffari A, Morkoç H and Yariv A (1987) Effect of substrate tilting on molecular beam epitaxial grown AlGaAs/GaAs lasers having very low threshold current densities. *Applied Physics Letters* 51: 2094–2096.

states function, characteristic of quantum wells, is one of the fundamental reasons for the reduction in threshold current of quantum well lasers relative to bulk GaAs devices.

We have considered GaAs because it is a model material system for production of quantum well structures. The concepts developed in this chapter are, however, quite general and can be applied to other material systems, other forms of quantum well and quantum wires and dots.

List of Units and Nomenclature

E	energy
E_f	Fermi energy
E_g	bandgap
F	envelope wavefunction
$f(E)$	Fermi–Dirac distribution function
g_{2D}	two-dimensional density of states $[\text{energy}]^{-1}$ $[\text{area}]^{-1}$
$\hbar$	Planck's constant divided by 2π
I_{ph}	optical field intensity
$\mathbf{k}$	wavevector
k_B	Boltzmann's constant
m	electron mass
n	carrier density, expressed per unit area in quantum well structures
Q_c, Q_v	heterostructure band offset ratios
$\mathbf{r}$	position vector
t	time
T_N	fraction of light transmitted at normal incidence through N quantum wells
$\langle\mathbf{x}\rangle$	unit vector along x-direction, similarly for y, z directions
α	optical absorption coefficient in a bulk material, per unit length
γ	fraction of light absorbed by a quantum well at normal incidence to the plane of the well
λ	wavelength
Φ	photon flux (photons per unit time crossing unit area)
Ψ	electronic wavefunction
ν	light frequency

See also

Semiconductor Materials: Band Structure Engineering; Quantum dots; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Excitons; Outline of Basic Electronic Properties.

Further Reading

Bastard G (1988) *Wave Mechanics Applied to Semiconductor Heterostructures*. New York: Les Editions de Physique, Halstead Press.

Bimberg D, Grundmann M and Ledentsov NN (1999) *Quantum Dot Heterostructures*. Chichester, UK: Academic Press.

Blood P (1991) Heterostructures in lasers. In: Morgan DV and Williams RH (eds) *Physics and Technology of Heterojunction Devices*. Stevenage, UK: Peter Peregrinus.

Blood P (2000) On the dimensionality of optical absorption, gain and recombination in quantum-confined structures. *Journal of Quantum Electronics* 36: 354–362.

Blood P and Orton JW (1992) *The Electrical Characterisation of Semiconductors: Majority Carriers and Electron States*. London: Academic Press.

Casey HC Jr and Panish MB (1978) *Heterostructure Lasers*. San Diego, CA: Academic Press.

Dingle R, Gossard AC and Weigmann W (1975) Direct observation of superlattice formation in semiconductor heterostructures. *Physical Review Letters* 34: 1327–1330.

Eisberg R and Resnick R (1985) *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd edn. New York: John Wiley.

Hook JR and Hall HE (1991) *Solid State Physics*. Chichester, UK: John Wiley.

Kelly MJ (1995) *Low-Dimensional Semiconductors*. Oxford, UK: Clarendon Press.

Orton JW and Blood P (1990) *The Electrical Characterisation of Semiconductors: Measurement of Minority Carrier Properties*. London: Academic Press.

Perkowitz S (1993) *Optical Characterisation of Semiconductors*. London: Academic Press.

Wolfe CM, Holonyak N Jr and Stillman GE (1989) *Physical Properties of Semiconductors*. Englewood Cliffs, NJ: Prentice Hall.

Recombination Processes

P T Landsberg, The University of Southampton, Southampton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In studies of recombination there occur considerable complications; many are associated with interactions between electrons, electrons and phonons, excitons, bi-excitons, impurity centers, etc. Most of these are not required in the present exposition. Modern work normally assumes that the basics of recombination physics are understood and the present exposition offers an appropriate outline.

Basic Assumptions

Electron–Electron Interactions for Electrons in Bands

The recombination problem in semiconductors is greatly complicated by the interaction of the electrons with each other. This allows one to speak only of the quantum states of the semiconductor crystal as a whole. However, as in metals, so also in semiconductors, a simplified picture is successful. In this, the electron interactions, and other interactions, are first neglected, but are later taken into account as a perturbation. Thus, the electrons are first treated as if they can move through each other; the fact that they collide and deflect each other by virtue of their Coulomb interaction is treated as a perturbation. The transitions are still described within the framework of the single-particle states of the unperturbed problem.

The Effect of Electron–Boson Interactions

Our first approximation is to neglect most (but not all) electron interactions. Later we take into account the two-electron transitions that arise. This two-particle recombination process is referred to as Auger recombination and its inverse as the generation of current carriers by impact ionization. In addition, electrons interact with the radiation and lattice fields and emit or absorb photons or phonons. These electron–boson interactions result in transitions of single electrons in an energy band scheme. We shall attach a superfix ‘S’ (for single-electron) to the recombination coefficients for such processes. These are then denoted by B^{S} or T^{S} depending whether only bands are involved or whether traps are also involved. They are illustrated in Figure 1, where the solid horizontal lines represent the conduction and valence band edges and the dashed line refers to traps. Note that the two-electron transitions do not have the superfix ‘S’, while n and p refer to the electron and hole concentrations; N_0 and N_1 are the concentration of unoccupied and occupied trap states, respectively.

Electron–Electron Interaction in Traps

Electron–electron interactions can at least formally be taken into account in connection with electrons trapped in a center: the spectrum is a function of the number, $r\,(= 1, 2, 3, \ldots, M)$, of electrons captured. The ‘irremovable’ electrons can be included with the ion core. Given that the center captured r electrons (say), it can still be in a variety of quantum states, and they will be denoted by the symbol ℓ, yielding a set of quantum states (ℓ, r) for a center. The energy of such a state, divided by kT to make it dimensionless, is denoted by $\eta(\ell, r)$, yielding the canonical partition function

$$Z_r = \sum_{\ell} \exp[-\eta(\ell, r)] \qquad [1]$$

The location of a center in space will here be considered to be of no significance.

Centers can capture several electrons. This brings in a need for the chemical potentials (or Fermi levels) for r-electron centers. They are here denoted by γ (when divided by kT) and the suffix ‘eq’ denotes an equilibrium value. Thus, the equilibrium probability of finding an r-electron center in state ℓ is given by:

$$P(\ell, r)_{\mathrm{eq}} = \frac{\exp[r\gamma_{\mathrm{eq}} - \eta(\ell, r)]}{\sum_{s=0}^{M} (\exp s\gamma_{\mathrm{eq}}) Z_S} \qquad [2]$$

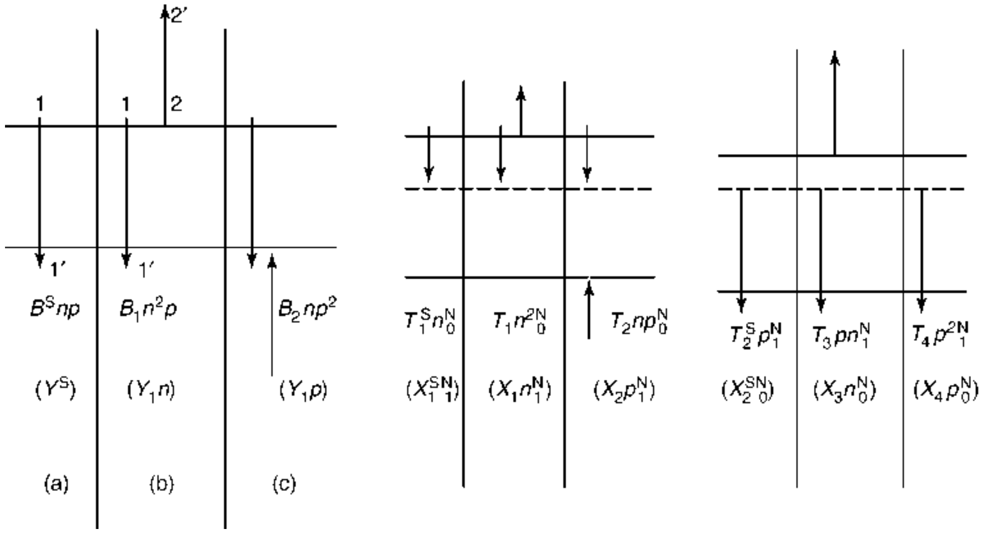

Figure 1 Definition of recombination coefficients. Transition rates per unit volume are stated with each process, and, in brackets, for the reverse process. Thus B_1, B_2, $T_1 - T_4$ refer to Auger processes; B^s, T_1^s, T_2^s refer to single-electron recombination; Y^s, X_1^s, X_2^s refer to carrier generation processes; Y_1, Y_2, $X_1 - X_4$ refer to impact ionization processes. Arrows indicate transitions made by electrons.

One can identify the dominant energies from optical or electrical experiments.

To understand the properties of these centers, suppose we can arrange for the equilibrium Fermi level to rise from the valence band to the conduction band. At first practically no electrons are captured ($r = 0$). As the Fermi level rises the states corresponding to $r = 1$ begin to appear. The states ℓ of the center which correspond to the ground states in equilibrium are always more highly populated than the states ℓ corresponding to excited states, so that we can often confine attention to them. As the Fermi level rises, the ground states for $r = 2$ become more important, and there may now be hardly any centers which have captured fewer electrons. If a larger number of electrons cannot be captured, by the time the Fermi level reaches the conduction band, then states of the center with $r > 2$ can (normally) be neglected as unstable. That this is a satisfactory picture is our second approximation.

Assumptions for Nonequilibrium Statistics

The now much simplified recombination problem is still complex because of the many states available to electrons in bands and centers. A key simplification arises from the fact that it is often possible to talk about a small number of groups of quantum states: let $I\,(= 1, 2, \ldots)$ label quantum states in a group i, let $J\,(= 1, 2, \ldots)$ label quantum states in a group j, etc. Within each group it is supposed that the transitions are much more rapid than they are between groups. In that sense electrons in each group are in equilibrium among themselves. Their quasi-equilibrium is then characterized by what is called a quasi-Fermi level. The dimensionless version, obtained by dividing it by kT, is denoted by γ_i, γ_j, etc., for groups i, j, etc. That this is reasonable is our third assumption. Thus γ_c, γ_v are quasi-Fermi levels which refer to the quasi-equilibrium of the conduction and valence band, respectively, and γ_I to an I-electron center. Groups of states with different quasi-Fermi levels are not in equilibrium with each other.

Recombination problems can now be discussed by neglecting transitions within one (quasi-equilibrium) group, because they proceed at exactly the same rate as their reverse transitions. The number of transition types to be considered is thereby greatly reduced. Away from equilibrium we shall adopt eqn [3] instead of [2] in order to allow for distinct quasi-Fermi levels:

$$P(\ell, r) = \frac{\exp[r\gamma_r - \eta(\ell, r)]}{\sum_{s=0}^{M} (\exp s\gamma_S) Z_S} \qquad [3]$$

Equation [3] is not always correct, but is an approximation arising from the quasi-Fermi level assumption. Among the improved theories are, for example, the ‘cascade’ theories.

The assumption of a quasi-Fermi level for each state of charge of a center implies that all the excited states of an *r*-electron center are populated according to eqn [2].

The Main Recombination Rates

General Results and The Two-Band Case

Recombination and its converse, generation, consist of a transition of an electron from one state to another. The observed rate is the net rate of recombination and is the algebraic sum of the recombination and generation processes. During these processes both total energy and total momentum must be conserved. This is achieved by creation or absorption of photons, or phonons, excitation of secondary electrons, etc.

The transition probability per unit time, S_{IJ}, for a single-electron transition from state I in a band to state J requires that state I should be occupied, with probability p_I say, and state J should be vacant with probability q_J say. The general expression for the average rate of the transition from I to J then takes the form $p_I S_{IJ} q_J$. For the reverse process electron state J has to be occupied and state I vacant. Thus the rate of this reverse process is $p_J S_{JI} q_I$. The net recombination rate per unit volume of the process I to J can then be written as:

$$u_{IJ} = (p_I S_{IJ} q_J - p_J S_{JI} q_I) V^{-1} \qquad [4]$$

By the principle of detailed balance, this expression vanishes at equilibrium. If one puts $X_{JI} \equiv S_{IJ} p_I q_J / S_{JI} p_J q_I$, one has $u_{IJ} = p_J S_{JI} q_I (X_{JI} - 1) V^{-1}$. In equilibrium $X_{JI} \to (X_{JI})_{\mathrm{eq}} = 1$, and the recombination rate is zero.

One can say a little more if one assumes that states I and J are in conduction or valence bands, each with a quasi-Fermi level (divided by kT to make it dimensionless). If these are denoted by γ_{e} and γ_{h}, then $p_I = [1 + \exp(\eta_I - \gamma_{\mathrm{e}})]^{-1}$ for a conduction band. For the total recombination rate per unit volume between the bands i and j one finds:

$$u_{ij} \equiv \left[\sum_{I \in i} \sum_{J \in j} p_J S_{JI} q_I \right] \left[\exp\left(\frac{|e|\phi}{kT} \right) - 1 \right] V^{-1} \qquad [5]$$

where ϕ is essentially the difference between the quasi-Fermi levels: $\gamma_{\mathrm{e}} - \gamma_{\mathrm{h}}$. The first factor depends on the bands involved and it has been assumed that the transition probability ratio S_{IJ}/S_{JI} is independent of the excitation.

A transition rate, when multiplied by the charge of current carriers, is a current, and when divided by the area of the surface involved, is a current density. If i and j denote the states of the conduction and valence bands of a semiconductor, excitation independence may often be assumed, and one then expects a current density proportional to $\exp(|e|\phi/kT) - 1$. This is characteristic of the current through pn junctions, metal semiconductor junctions, etc. In these configurations the Fermi level difference between the ends of the device determines the voltage across it. When radiation is involved, however, excitation dependence of some of the parameters introduced above (e.g. S_{JI}) tends to spoil this simple story.

The Case of Defects

So far we have considered only two bands. When a trap is involved matters are rather different. Because of the interactions among the electrons on a center it is not possible to talk of the same level being occupied or vacant. Consequently, identification of forward and reverse processes in terms of levels becomes impossible. Instead one deals with a center, say an *r*-electron center, as a whole; we then need the probability that a given center is an *r*-electron center. For example, the capture of an electron converts an $(r-1)$-electron center into an *r*-electron center. Thus the p's and q's must be replaced by more complicated expressions. It is convenient to denote u_{ij} by u_{cv} in the simple two-band case and its structure is given (with a sign change) by a recombination coefficient:

$$u_{\mathrm{cv}} = B^{\mathrm{S}} np [1 - \exp(\gamma_{\mathrm{h}} - \gamma_{\mathrm{e}})] \equiv B^{\mathrm{S}} np - Y^{\mathrm{S}} \qquad [6]$$

using Figure 1a. Here, n and p are the electron and hole concentrations. Auger effects in Figures 1b and 1c can also be included, at least formally. Then B^{S} has to be replaced in eqn [6] by $B^{\mathrm{S}} + B_1 n + B_2 p$. Analogous replacements are found for recombination processes involving defects.

In the simplest case of one type of localized defect one finds a steady-state recombination rate per unit volume of the form:

$$u = \frac{np - (np)_0}{\tau_{n0}(p + p_1) + \tau_{p0}(n + n_1)} \qquad [7]$$

Here τ_{n0}, τ_{p0} are parameters with the dimension of time and p_1, n_1 are parameters with the dimension of concentration. The recombination increases with the defect concentration, which is actually in the denominators of τ_{n0} and τ_{p0}. This is an old and much used result associated with the names of W Shockley and W T Read.

Radiative Transitions

Spontaneous and Stimulated Emission

Quantum theory was initiated by Planck's law for black-body radiation at temperature T. This gives the spatial energy density as a function of frequency ν in this system as

$$f(\nu, T) = \frac{8\pi\nu^2}{c^3} \frac{h\nu}{\exp(h\nu/kT) - 1} \qquad [8]$$

For low frequencies one finds the Rayleigh–Jeans law which makes [8] proportional to kT in agreement with the classical equipartition theorem. For high temperatures [8] diverges, as required.

The result [8] may also be obtained by writing

$$N = [A + Bf(\nu, T)]N_{\mathrm{u}} \qquad [9]$$

for the emission rate of photons by atoms, where N_{u} is the number of atoms in the upper of two states which are separated by the energy $h\nu$. The first term on the right-hand side is due to the normal decay of an excited state ('spontaneous emission'). The second term refers to additional emission ('stimulated emission') induced by the radiation of frequency $h\nu$ itself.

The first factor in [8] is due to the density of states and the second factor is due to the fact that the energy is considered. If one considers the number of photons of energy $h\nu$ at temperature T one comes up with

$$N_\nu = \left[\exp\frac{h\nu - \mu}{kT} - 1\right]^{-1} \qquad [10]$$

(Bose–Einstein distribution)

Here μ is a possible chemical potential of the radiation which is non-zero only in nonequilibrium situations such as in a semiconductor laser.

In a pn junction the two bulk materials several diffusion lengths away from the junction are approximately in equilibrium even if a modest current is flowing. On one side one has then just one quasi-Fermi level, say μ_I, and on the other side one has just one quasi-Fermi level, say μ_J. Then the difference

$$\mu \equiv \mu_I - \mu_J = q\varphi \qquad [11]$$

corresponds to the applied voltage φ. This is in agreement with eqn [10], in that $\varphi = 0$ implies no current and hence thermal equilibrium is possible.

Statistical mechanics teaches one the rule that, in equilibrium, occupation probabilities of individual quantum states are always less for states of higher energies. This ensures that the total energy of a quantum system converges, the occupation probability p being a kind of convergence factor. However, if one is away from equilibrium, the above rule can be suspended and one can have population inversion.

For $h\nu = \mu$ there is trouble with [10] because the steady-state photon occupation diverges. This does not correspond to a 'death ray', but is the result of imperfect modeling; for example, the leakage of photons from the cavity may have been neglected. A formula of type [10] is also needed in connection with solar cells.

Donor–Acceptor Pair Recombination

A striking demonstration of radiative donor–acceptor transitions in GaP at low temperature (1.6 K) was revealed by sharp lines (Figure 2) at photon energies $h\nu_i$ given by

$$h\nu_i = E_{\mathrm{G}} - E_{\mathrm{A}} - E_{\mathrm{D}} + e^2/\varepsilon R_i \, [-E] \qquad [12]$$

where R_i is the distance between the ith-order nearest neighbors.

The discrete nature of the peaks is due to the fact that the impurities involved settle in general on lattice sites so that only definite separations R_i are possible. The energy gap is E_{G}; the value of $E_{\mathrm{A}} + E_{\mathrm{D}}$ can be inferred from the experimental lines by extrapolation to $R_i \to \infty$. The energy term E is sometimes neglected.

The ZnS phosphors were the first materials in which donor–acceptor radiation was hypothesized. However, it is hard to control its stoichiometry and its impurity content, and it has relatively large carrier mass and hence relatively large impurity activation energies. In such cases spectra such as those shown in Figure 2 are hard or impossible to obtain. This applies in general to many II–VI compounds.

Quantum Efficiency

The quantum efficiency is the radiative recombination as a fraction of the total recombination. It can also be regarded as the average number of electrons produced per incident photon. Consider an intrinsic semiconductor, i.e., one in which the electron and hole concentrations are equal. Then with the notation of Figure 1 the radiative band–band recombination rate is $B^{\mathrm{s}}n^2$. The nonradiative rate is $(B_1 + B_2)n^3$. We shall add another nonradiative rate, An say, proportional to the injected carrier density $n \sim p$. The quantum efficiency is then

$$\eta = \frac{B^{\mathrm{s}}n^2}{An + (B_1 + B_2)n^3 + B^{\mathrm{s}}n^2} \qquad [13]$$

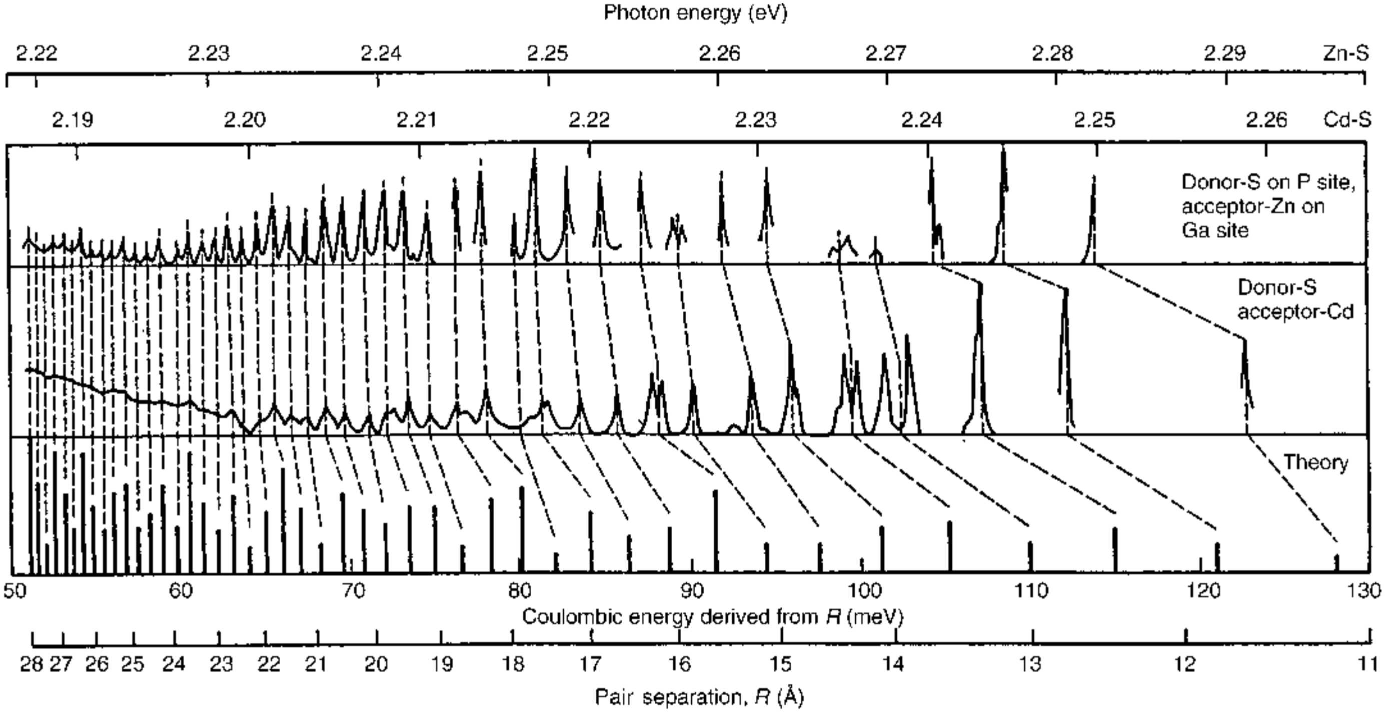

Figure 2 Comparison of the positions and intensities of the sharp line spectra at 1.6 K corresponding to both ZnS and CdS acceptor–donor pairs with the predicted pair distribution. The lower scales show the pair separation (R) and the Coulombic energy derived from R. The emission energy scales for the two measured spectra are shown at the top. Reproduced with Gershenzon M, Logan RA, Nelson DF, *et al.* (1968) *Proceedings of the International Conference on Luminescence*. Budapest, Hungary: Akademia Kiada.

Traps are here neglected, and one finds a maximum

$$\eta_{\max} = \left\{ 1 + \frac{2}{B^{s}} A^{1/2} (B_1 + B_2)^{1/2} \right\}^{-1} \qquad [14]$$

For a high-quality epitaxial AlGaAs/GaAs double heterostructure of great purity we may take

$$A \sim 0.5 \times 10^{6}\ \mathrm{s}^{-1},\ \ B \sim 10^{-10}\ \mathrm{cm}^{3}\ \mathrm{s}^{-1},$$
$$B_1 + B_2 \sim 10^{-29}\ \mathrm{cm}^{6}\ \mathrm{s}^{-1} \qquad [15]$$

whence $n_1 \sim 2 \times 10^{17}\ \mathrm{cm}^{-3}$, $\eta_{\max} \sim 0.96$.

Detailed Balance

It is clear that the radiative recombination rate from a material should be obtainable in terms of its optical 'constants', the absorption coefficient $\alpha(\lambda)$ and the refractive index $\mu(\lambda)$. Both are functions of the wavelength. This connection can be formalized by comparing the optical absorption process with the emission process and this can be done by using the principle of detailed balance. This equates the rate of disappearance by absorption of photons with the rate of production of photons by radiative recombination. The radiative recombination rate can thus be obtained as an integral over the optical functions. This relationship, pioneered by van Roosbroeck and Shockley in 1954, has been used a great deal because the optical quantities are often known with some accuracy.

This result corresponds to balancing the rate $B^{s}np$ and the rate Y^{s} in Figure 1a. Analogous detailed balance results are obtainable for the other pairs of processes in Figure 1. Thus, the Auger recombination rate can be related to an integral over the impact ionization rate. However, this connection is less useful since impact ionization data are generally less well known than the optical absorption information.

Nonradiative Processes

Auger Effects

The Auger effect was discovered in 1925 in gases by Pierre Auger. An atom is ionized in an inner shell. An electron drops into the vacancy from a higher orbit and a second electron takes up the energy which is used to eject it from the atom. In solids the effect is roughly analogous. One of its characteristics is that it is not radiative. Since radiation is what is seen in many experiments, the Auger effect is suspected if and when there is less radiation than expected in the first place. It is rather harder to investigate than radiative transitions.

The effect has proved to be important as it limits the performance of semiconductor lasers, light-emitting diodes and solar cells, and it can be crucial in transistors and similar devices whose performance is governed by lifetimes. When heavy doping is required, as it is in the drive towards microminiaturization, its importance tends to increase, since the Auger recombination rate behaves roughly as n^2p or p^2n (see Figure 1) compared with the radiative rate which behaves more like np. The inverse process is impact ionization, and is important in the photodiode, the impact avalanche transit time (IMPATT) diode, and hot electron devices.

Impact ionization can be regarded as an autocatalytic reaction of order one:

$$\mathrm{e} \rightarrow 2\mathrm{e} + \mathrm{h} \quad \text{or} \quad \mathrm{h} \rightarrow 2\mathrm{h} + \mathrm{e} \qquad [16]$$

i.e., one extra particle is produced of the type present in the first place. This is a key feature for impact-induced nonequilibrium phase transitions in semiconductors.

In the theory one needs wavefunctions for four states of two electrons which are relevant after the many-electron problem has been reduced to a two-electron problem. In a matrix element calculation the four wavevectors imply a 12-fold integration in $\boldsymbol{k}$-space to cover all possible states of the two electrons. However, momentum and energy conservation usually reduce this to an eight-fold integration. Such a calculation is hard, for it requires (1) good wavefunctions and (2) accurate integrations.

Here we shall merely concentrate on the broad principles. In addition, the many-electron nature of the problem implies that another approximation is often inherent in the treatment apart from the use of perturbation theory. This is clear if one considers that the electron interactions are screened twice: once by an exponential screening factor and a second time by the dielectric constant. So there is some double counting, and the treatment of the effect as uncorrelated electronic transitions mediated by screened Coulomb interactions is another approximation. The collective effects that enter require more sophisticated field theoretic methods which take care of electron–hole correlations, plasmon effects and the effect of free excitons, the so-called excitonic Auger effect. These calculations broadly confirm the results obtained by the simpler methods used for energy gaps large compared with the plasmon energies, provided the high-frequency dielectric constant is used and doping is not too heavy. Significant corrections are required for the narrow gap lead compounds, for example, and a considerable computational effort is required.

Let us now consider lifetimes for high carrier concentrations, as limited by band–band processes. They are best studied by looking at the emitted radiation. In this connection we recall the striking rise of the threshold current density J in a semiconductor laser as the material is changed from a direct material like GaAs to an indirect material, by mixing it with a compound such as GaP. Figure 3 shows this spectacular rise for $\mathrm{GaAs}_{1-x}P_x$ at 77 K near $x = 0.38$. It is due to the drop in the radiative transition probabilities, as the substance becomes indirect, thus requiring a higher current density for threshold. So we should start with band–band processes in direct materials as most favorable for the experimentally accessible radiative transitions.

The essential point here is that the injected carrier density behaves as:

$$n_{\mathrm{inj}} = \frac{J\tau}{qd} \sim \frac{(5\times 10^{3}\ \mathrm{A\ cm^{-2}})(10^{-9}\ \mathrm{s})}{(1.6\times 10^{-19}\ \mathrm{C})(10^{-4}\ \mathrm{cm})} \sim 10^{17}\ \mathrm{cm^{-3}} \tag{17}$$

where J/q is the particle current density at threshold, τ is the lifetime of the carriers, and d is the thickness of the active layer. As active layers are made thinner n_{inj} increases to $10^{19}\ \mathrm{cm^{-3}}$ or so, far in excess of the defect concentration in the (undoped) material. This brings in band–band Auger effects. These have also been invoked to explain the undesirable increase in J with temperature. Thus the interest in direct band–band Auger effects in III–V compounds is fuelled by the need for better and smaller optoelectronic devices which work at long wavelengths (1.3–1.5 μm). It matches the interest in indirect band–band Auger and impurity Auger effects in silicon due to the importance of heavy doping in VLSI (very large scale integration) and transistor technology. Figure 4 shows a typical Auger process, called CHHS, as the conduction band (C) and the split-off band (S) are involved. Two relevant states are in the heavy hole band (H). State 2′ is referred to as the Auger carrier, as it has more kinetic energy than the others.

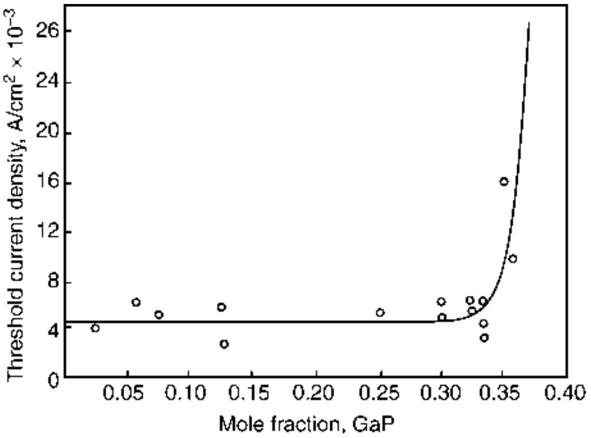

Figure 3 Lowest values of laser threshold current density at 77 K as a function of mole fraction of GaP for $\mathrm{Ga(As}_{1-x}P_x)$ junction lasers. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. *Solid-State Electronics* 9: 735–742, with permission from Elsevier.

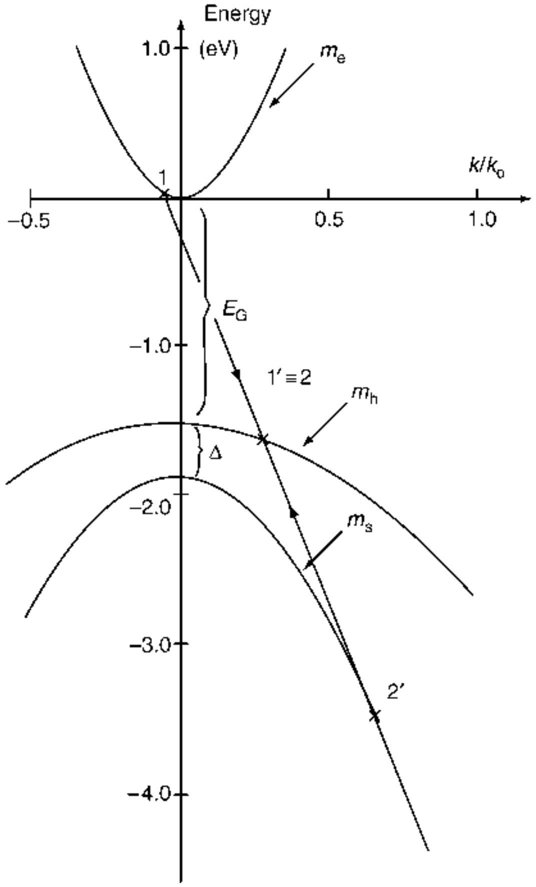

Figure 4 Conduction band, heavy hole and split-off valence bands of GaAs, all treated as parabolic with $\hbar^2 k_0^2/2m_h \equiv E_G - \Delta$. x denotes a quadruplet of states for a most probable transition. The two states in the heavy hole band are not shown separately. The arrows indicate electron transitions. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. Solid-State Electronics 9: 735–742, with permission from Elsevier.

Impact Ionization

The CHHS and CHCC processes and their inverses can be written as

$$\mathrm{e_C + 2h_H \leftrightarrow h_s \quad 2e_C + h_H \leftrightarrow e_C} \qquad [18]$$

where the suffix refers to the band. Viewed from left to right these are Auger processes. Viewed from right to left they are autocatalytic and impact ionizations.

Such processes are important for the impact-induced nonequilibrium phase transition in semiconductors. Also reaction rates with autocatalytic elements imply nonlinear equations in the concentrations and this gives rise to much interesting behavior as regards stability and bifurcation phenomena.

Momentum and energy conservation are very restrictive conditions on the four states involved in [18], and it is easily seen that they cannot normally be satisfied if in a direct band semiconductor the recombining electron drops from the band minimum to the valence band maximum. This effect raises all the kinetic energies of possible processes. The Auger electron (on the right-hand sides of [18]) must also have a minimum energy in order that the impact-ionization process can proceed. This leads to an activation energy for the process. For Figure 4, for example, and in the case of nondegeneracy, the energetic hole has a kinetic energy at threshold of

$$E_{\mathrm{th}} = \frac{m_{\mathrm{e}} + 2m_{\mathrm{h}}}{m_{\mathrm{e}} + 2m_{\mathrm{h}} - m_{\mathrm{s}}}(E_{\mathrm{G}} - \Delta) \quad \text{(CHSS)} \qquad [19]$$

As the band gap increases we see that E_{th} goes up and so the Auger rate for simple parabolic bands decreases. Values of E_{th} for other transitions can be obtained from equation [19].

The total kinetic or threshold energy E_{th} can be converted to an activation energy by subtracting the basic energy which must under all conditions be part of the Auger particle energy. In the case of eqn [19] one finds for $E_G > \Delta$:

$$\begin{aligned} E_{\mathrm{a}} &= E_{\mathrm{th}} - (E_{\mathrm{G}} - \Delta) \\ &= \frac{m_{\mathrm{s}}}{m_{\mathrm{e}} + 2m_{\mathrm{h}} - m_{\mathrm{s}}}(E_{\mathrm{G}} - \Delta) \qquad [20] \end{aligned}$$

This can result in a strong temperature dependence in accordance with an Arrhenius factor $\exp(-E_{\mathrm{th}}/kT)$. However, this is again lost, and the direct band–band Auger effect will certainly be important and only weakly temperature dependent, if $E_{\mathrm{G}} \sim \Delta$. This can occur, for example, for InAs, GaSh and their solid solutions.

The connection between Auger processes and impact ionization has also been formalized. Of all Auger quadruplets of states in a set of nondegenerate bands, the most probable Auger transition involves the quadruplet, which yields the threshold for impact ionization. The three crosses in Figure 4 indicate such a quadruplet (the central cross represents two states).

Identification of Auger Effects

How does one know that an Auger effect has occurred? In pure but highly excited materials there is the original solid-state Auger effect which leads to a lifetime broadening of the electronic states at the band edge. For a semiconductor this effect causes fuzziness in the band edge which is a contributory factor to the overall bandgap shrinkage. The mechanism is illustrated in Figure 5, which shows how the lifetime of a vacant electron state near the band edge is shortened by the Auger processes within this band. Under normal conditions this effect is small in a semiconductor. But under degenerate conditions the effect leads by lifetime broadening to a low-energy tail in emission.

The blanket term 'Auger effect' for all Coulombically excited two-particle transitions has been used here. In semiconductor device work the 'Auger effect' has become associated with transitions in which one electron bridges an energy gap. But there is no need to limit the concept in this way. Its competition with radiative effects means that it is often detrimental and a full discussion of means of suppressing it has recently been given by Pidgeon *et al.*

More spectacular and more convincing is the *detection* of the energetic Auger electron or hole. For this purpose one may look for the weak luminescence emitted when this carrier recombines radiatively. This has been done for the band–band process in Si and for the band–impurity process

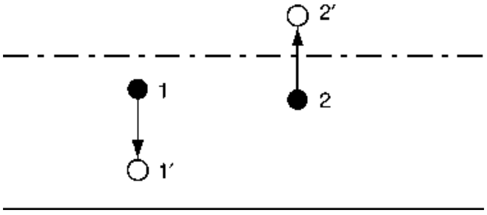

Figure 5 The filling of a state *k* by the Auger effect. The arrows indicate electron transitions.

in GaAs. It leads to so-called $2E_G$-emission. Its rate-limiting step is the rate of the Auger process per unit volume, B_1n^2p, which populates the high-energy level. The radiative process then proceeds at a rate $B'_1n^2p^2$ per unit volume with $B'_1 \sim 10^{-54}$–10^{-51} cm^9 s^{-1}.

More usual is the identification of the band–band Auger recombination mechanism from the minority carrier lifetime τ, which behaves as

$$\frac{1}{\tau_p} = B_1n^2 \quad \text{or} \quad \frac{1}{\tau_n} = B_2P^2 \qquad [21]$$

where B_1 and B_2 are Auger coefficients when the Auger particle is an electron or hole, respectively. A more complete expression allows also for trapping processes.

The departure from parabolic bands plays an important part for the energetic Auger particle (in state $2'$). Theory shows that when this effect is taken into account, it tends to lower the band–band Auger coefficient. In fact in GaAs the CHCC process is ruled out altogether at 0 K for nonparabolic bands, suggesting that it should be an ideal material for radiative transitions. However, the possibility of phonon participation brings the effect back again, though it is still comparatively weak. The difficulty of conserving electron energy and momentum, which gives rise to the activation energies (eqn [20]) is greatly alleviated if phonons can take up some of the momentum in a direct gap semiconductor.

We have seen that the activation energy barrier against the band–band Auger effect can be overcome by a suitable disposition ($E_G = \Delta$) of the three direct bands and by phonon participation. It can also be overcome if there is an indirect minimum near a Brillouin zone edge at about half the direct energy gap.

Acknowledgments

The author wishes to acknowledge support from NATO under their grant PST.CLG 975758.

See also

Semiconductor Physics: Band Structure and Optical Properties; Excitons; Impurities and Defects; Infrared Lattice Properties; Outline of Basic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Fraser DA (1986) *The Physics of Semiconductor Devices*, 4th edn. Oxford, UK: Clarendon Press.

Hangleiter A (1985) Experimental proof of impurity Auger recombination in silicon. *Physical Review Letter* 55: 2976–2978.

Harrison D, Abram RA and Brand S (1999) Characteristics of impact ionization rates in direct and indirect gap semiconductors. *Journal of Applied Physics* 85: 8186.

Landsberg PT (1991) (paperback edn., 2003) *Recombination in Semiconductors*. Cambridge, UK: Cambridge University Press.

Nimtz G (1980) Recombination in narrow gap semiconductors. *Physics Reports* 63: 265.

Pidgeon CR, Ciesla CM and Murdin BN (1997) Suppression of non-radiative processes in semiconductor mid-infrared emitters and detectors. *Progress in Quantum Electronics* 21: 361–419.

Spin Transport and Relaxation in Semiconductors; Spintronics

M E Flatté, University of Iowa, Iowa City, IA, USA

D D Awschalom, University of California, Santa Barbara, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Control of the generation and transport of coherent electron spin in semiconductors suggests new ways to probe the fundamentals of quantum decoherence in solids and to explore device applications that rely on coherence, including quantum computation. The focus here will be on the physical phenomena that govern spin transport and relaxation, with a particular emphasis on those phenomena that are amenable to direct manipulation. Manipulation of material properties in general has been taken to its most advanced level in the design of semiconductor electronic devices, in which operations are performed by perturbing the motion of electrons through their electric charge; hence, ‘spintronics’ is a term introduced to describe both the manipulation of electrons by their spin and control of their spin properties, such as g-factors.

The study of the evolution of spin coherence first requires the generation and detection of spin coherence. Here the ultrafast optical technique for generating and detecting spin-polarized populations will be emphasized. The next logical component to treat will be the persistence of spin coherence, followed by the transport of spin coherence from one region of a material to another. Finally the direct ultrafast manipulation of spin-polarized populations will be described. These components are likely to play a prominent role in any attempt to apply the fundamental physical phenomena of spin coherence to specific device technologies.

Optical Generation of Spin-Polarized Distributions

Optical pulses with definite circular polarization can be used to generate spin-polarized distributions of excited electrons in crystals. In order for this process to work one of the energy bands involved in the optical transition must be characterized by a substantial spin–orbit interaction. The selection rules for circularly polarized photons require a change in the orbital angular momentum of the electron projected along the propagation direction of the photon. For the optical transitions of an atom in free space this azimuthal orbital angular momentum l_z must increase or decrease by $\hbar$, permitting the coupling of the p-state with $l_z = 1$ to an s-state for left circular polarization, and of the p-state with $l_z = -1$ to that s-state for right circular polarization. Thus the optical selection rules pick out a particular orbital angular momentum projection.

These selection rules by themselves do not explain how spin-polarized populations are generated, for both spin directions are possible for each value of l_z. In the presence of spin–orbit interaction, however, the energy of a state with a particular l_z is correlated with the spin orientation. The six p-states, which are degenerate in the absence of spin–orbit interaction, split into two groups: one of four degenerate states and the other of two degenerate states. The higher-energy group of four degenerate states consists of two states with **L** and **S** parallel and two states with a one-third probability for **L** and **S** to be antiparallel and a two-thirds probability for **L** and **S** to be perpendicular. Thus for a transition involving the four degenerate states, a circularly polarized photon (that selects a particular initial l_z) is three times more likely to generate a carrier with spin parallel to the initial l_z than antiparallel to it.

In a solid, although full rotational symmetry is lost, the band symmetries preserve some of these selection rules. In crystals of tetrahedral symmetry, such as zincblende semiconductors, the selection rules for transitions from the valence to the conduction band are identical to the selection rules between p and s states for the atom in free space. For example, the valence heavy-hole states are those for which **L** and **S** are parallel. Thus by setting the optical photon energy to the fundamental gap, the resulting excited conduction electrons and valence holes will be spin polarized. In quantum wells, where the energy of the heavy hole is split from that of the light hole by quantum confinement (and, sometimes, strain), fully spin-polarized carrier distributions are possible. Commonly used group IV, III–V, and II–VI semiconductors have a sufficient spin–orbit interaction to permit the straightforward optical generation of spin-polarized distributions. Shown in Figure 1 are the selection rules

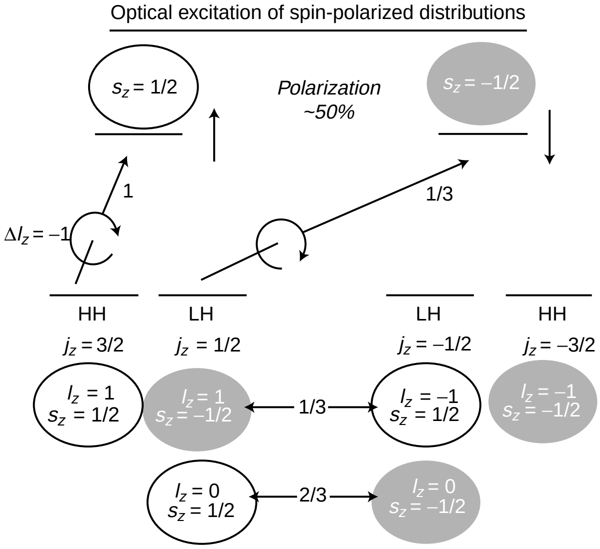

Figure 1 Selection rules for bulk zincblende crystals. Circularly polarized light changes the azimuthal orbital quantum number l_z by $+1$ or -1. Here $\Delta l_z = -1$ is shown. The orbital and spin quantum numbers for the two degenerate conduction band edge states and the four degenerate valence band edge states are shown – unshaded areas correspond to spin up and shaded to spin down. A transition with $\Delta l_z = -1$ couples one heavy-hole state to the conduction band up-spin state, and one light-hole state to the conduction band down-spin state. The orbital angular momentum of the heavy-hole state is entirely parallel to its spin angular momentum. In contrast only one-third of the light-hole state has antiparallel orbital and spin angular momentum. The rest of the light-hole state has perpendicular orbital and spin angular momentum. Thus the oscillator strength of the transition creating the conduction up-spin state is three times larger than that of the transition creating the conduction down-spin state, yielding a 50% spin polarized distribution. In quantum wells for which the heavy–light splitting is larger than the linewidth of the light source driving the transition, only the heavy-hole transition is excited and the spin polarization approaches 100%.

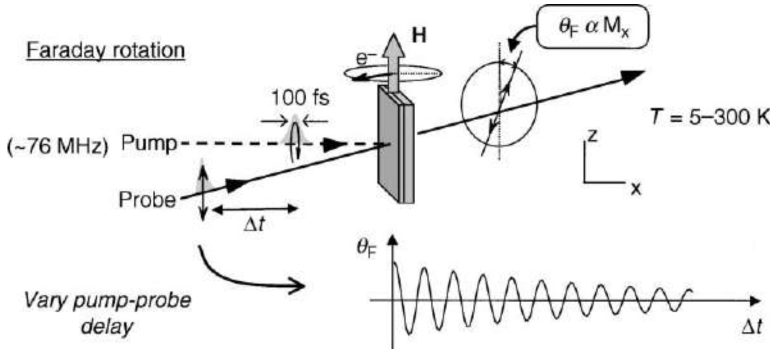

Figure 2 Diagram of the experimental geometry for time-resolved Faraday rotation. A pulse of circularly polarized light illuminates the sample, followed by a probe of linearly polarized light. The linearly polarized light can be considered as a sum of equal parts left and right circularly polarized light. The spin-polarization of the electron population (along the photon propagation direction) induces a difference in index of refraction for left and right circularly polarized light (circular dichroism). This produces a phase change between the left and right polarized components of the probe, which is measured as a rotation (θ_F) of the linear polarization direction. Monitoring the angle of rotation of the probe as a function of time directly indicates the spin polarization along the propagation axis of the photon as a function of time. Reproduced with permission from Awschalom DD, Samarth N and Loss D (eds) (2002) *Semiconductor Spintronics and Quantum Computation*. New York: Springer

which permit circularly polarized optical pulses to generate spin-polarized carriers in the conduction and valence band.

The evolution of a coherent spin-polarized population can be monitored through optical means as well. If the optical transitions used to generate the spin-polarized population are pumped hard enough, the transitions will be bleached. By measuring the polarization dependence of this bleaching the decay of the spin-polarized population can be determined. This can be done either by monitoring the transmission of time-delayed probe pulses of circularly polarized light, or by monitoring the rotation angle (Faraday rotation) of a pulse of linearly polarized light. The first of these depends directly on the imaginary component of the spin-polarized dielectric function, whereas the second depends directly on the real components. Shown in Figure 2 is the geometry for observing Faraday rotation from a spin-polarized distribution in the presence of an applied magnetic field perpendicular to the spin polarization.

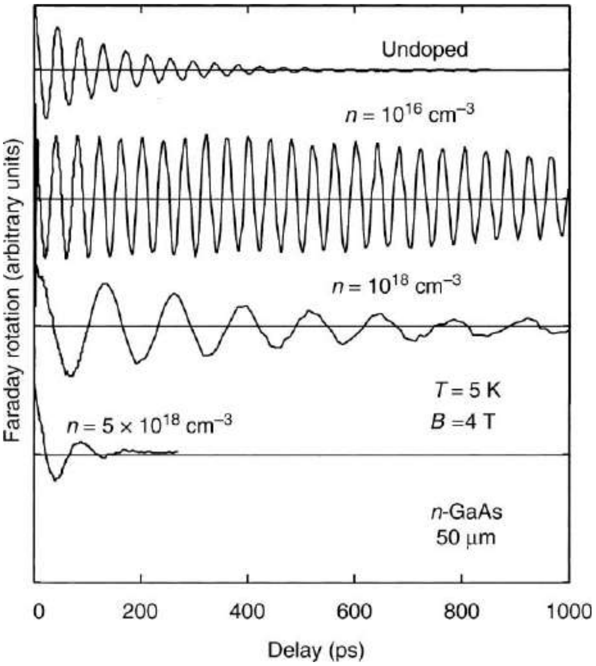

Figure 3 Time-resolved Faraday rotation measurements on bulk GaAs at several doping levels, indicating a sensitive doping dependence of the transverse spin coherence time T_2. T_2 for ideal doping densities can exceed 150 ns at 5 K. The magnetic field is 4 T, and the sample is 50 μm thick. Reproduced with permission from Kikkawa JM and Awschalom DD (1998) Resonant spin amplification in n-type GaAs. *Physical Review Letters* 80: 4313–4316.

Mobile Electron Spin Time Evolution: Decoherence and Precession

Two natural phenomena in the time evolution of a coherent spin-polarized population are decay of that population (decoherence) and rotation of the macroscopic magnetization in the presence of an applied magnetic field (precession). The time-scale of decay of the spin coherence can be very much longer than time-scales for the decay of other coherent phenomena in solids. Typical orbital coherence times, for example, are 100 fs, whereas spin coherence times can exceed microseconds. In Figure 3 are shown time-resolved Faraday rotation data for three bulk GaAs crystals with different doping levels. The oscillations are characteristic of the spin precession, and the oscillation frequency is the Larmor frequency of the electron spin in this crystal. The crystal with the middle doping level, $10^{16}\ \mathrm{cm}^{-3}$, has oscillations that persist for the longest time, and therefore this material has the longest spin coherence time.

The dominant mechanism for the decay of these spin polarized distributions works via the same spin–orbit interaction that permits the generation and detection of these spin-polarized distributions. For mobile electrons in homogeneous environments, the mechanisms of spin coherence relate to the crystal momentum-dependence of the eigenstates of the electronic system. Because of time reversal invariance the electronic states with zero momentum (at the Γ point in the Brillouin zone) are doubly degenerate. The electronic states for momentum $\mathbf{k}$ near the Γ point occur in nearly degenerate pairs. Although it is common usage to refer to these eigenstates as degenerate spin eigenstates (correct in the absence of spin–orbit interaction), with spin–orbit interaction the states are not spin eigenstates and are not necessarily degenerate. As a consequence, when a spin-polarized carrier is scattered by ordinary, non-spin-dependent scattering from one momentum state $\mathbf{k}$ to another $\mathbf{k}'$, and then back again to $\mathbf{k}$, the effective orientation of the carrier spin will have rotated by an angle dependent on the intermediate momentum state $\mathbf{k}'$. For a population of carriers in thermal equilibrium this process will lead to misorientation of the individual spin moments from the mean, and thus a decoherence of the spin population. This process is characterized by an increasing decoherence rate with increasing orbital scattering. Thus the dirtier a material is made, the faster the spin orientations decohere.

In crystals with inversion symmetry the pairs of nearly degenerate states are precisely degenerate. Lack of inversion symmetry, and the consequential energy splitting, can be viewed as an effective (crystal) magnetic field, in which the approximate spin eigenstates precess. The magnitude and direction of this effective field depends on the orbital momentum $\mathbf{k}$, and because of time reversal invariance, the effective fields for $\mathbf{k}$ and $-\mathbf{k}$ are opposite. Thus an initially spin-polarized population will dephase due to experiencing different effective magnetic fields. Ordinary spin-independent scattering can now be seen as randomly changing the precession axis for the spin in this effective field. Once the information about the precession axis orientation has been lost the dephasing is not reversible and the spin orientations have decohered. As the precession angle increases linearly between scattering events, the longer the time between such scattering events the faster the spin population decoheres. Thus for this process the cleaner a material is made, the faster the spin orientations decohere. This precessional decoherence process dominates for mobile conduction electrons in bulk and quantum-well III–V semiconductors near room temperature.

The influence of the spin–orbit interaction is far stronger for the valence electrons than the conduction electrons for two reasons. First, as the effective magnetic fields come from relativistic transformations of internal electric fields, and the electric fields are largest near the nucleus, the valence electrons, which are closer to the nucleus than the conduction electrons, experience a stronger spin–orbit interaction. Second, the valence electrons have a dominant p-like character, whereas the conduction electrons have dominant s-like character (and thus lack orbital angular momentum to be coupled to the spin). Thus the spin coherence times of valence electrons (or holes) are commonly much shorter than the coherence times of the conduction electrons. Spin-flip exchange between conduction electrons and valence holes can therefore be the dominant source of decoherence for conduction electrons. This process can dominate at low temperatures in clean III–V semiconductors, and even up to room temperature in III–V quantum wells grown along the unusual (110) growth direction.

As a consequence one would expect the spin coherence times of electrons in n-doped semiconductors to be longer than those of electrons in p-doped or undoped semiconductors (in which excess holes must also be optically generated). This expected trend is present in Figure 3, for the n-doped materials have longer spin coherence times than the semi-insulating material. However, as the effect of the spin–orbit interaction increases with increasing momentum $\mathbf{k}$, the spin coherence times of the more heavily doped material are shorter than those of the lightly doped material. The doping corresponding to the longest spin coherence times may lie near the metal–insulator transition, although the theory of spin coherence times on the insulating side of that transition is not as well developed as on the metallic side.

Transport and Manipulation of Spin Coherence

An essential component of the motion of spin coherence in semiconductors is the response of carriers to electric fields. Outside of very unusual electronic states in semiconductors, such as the quantum Hall state, there is no spin–charge separation of elementary carriers. The carriers in the semiconductors will move under the influence of an electric field, and if they happen to be spin-polarized, then a spin current will exist. Although there are numerous possible regimes of spin transport (at least as many as there are regimes of charge transport), only two will be presented here, both in the low-field diffusive limit.

Mobilities and diffusion constants are the material properties that describe the low-field charge transport in a homogeneous material. Spin transport can be described similarly, and the spin mobility and diffusion constant can be related to those of charge motion. The relationship between the two emerges from the nature of carrier packets that can be maintained under low-field conditions. Deviations from local charge neutrality cause very large space-charge fields in the semiconductor, and thus under low-field conditions approximate local charge neutrality must be maintained. Shown in Figure 4 are charge and spin packets in an n-doped semiconductor. The charge polarization packet, incorporating extra conduction electrons and valence holes, is shown in Figure 4a. The electron spin polarization packet shown in Figure 4b is qualitatively different from the charge packet as well as the hole spin polarization packet (shown in Figure 4c). Only the electron spin polarization packet does not incorporate both conduction electrons and valence holes.

The difference in packet structure profoundly changes the mobility and diffusion properties of the packet. For packets consisting of both electrons and holes the low-conductivity carrier species dominates the packet motion (in this case, the holes). For the electron spin packet, however, majority carriers dominate the packet motion. Thus motion of this spin packet is very rapid, and also sensitive to the higher Fermi temperature of electrons relative to holes. Experimentally this has been seen by optically creating a local spin-polarized packet of conduction electrons (an electron spin packet) in a thin sample of n-doped GaAs. The application of a lateral electric field drags that spin packet, and an optical probe measures the transport distance (thus the mobility) and the packet spread (yielding the diffusion constant). Such an approach is very similar in concept to the electrical probe of charge packet motion developed by Shockley and Haynes to measure minority hole mobility and diffusion in n-doped Ge.

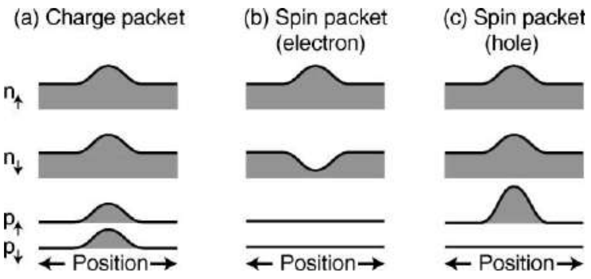

Figure 4 Inhomogeneous carrier packets in n-doped semiconductors in the low-field regime. The charge polarization packet (a) is constructed of approximately equal numbers of electrons and holes. The electron spin polarization packet (b) can be constructed without any excess holes, whereas the hole spin polarization packet (c) requires both excess electrons and excess holes. The presence of minority carriers in a packet significantly alters the mobility and diffusion properties of the packet.

The picture of approximately charge-neutral packets is very helpful for describing spin motion in homogeneous materials, but most semiconductor devices are based on inhomogeneous configurations of semiconducting material. For example, the most ubiquitous charge-based semiconductor device is the p–n diode, whose nonlinear current–voltage characteristics have been used as building blocks for numerous other devices. Such devices rely on the ability of a current to traverse an interface between two dissimilar materials – either different in doping density or different in host structure. Experiments in metallic ferromagnetic systems indicated that, whereas charge current was easily passed from one ferromagnetic metal to another through an interface, the same was not always true for the spin current. Differences in band structure or interface roughness are usually expected in such systems to strongly diminish the efficiency of spin current transfer across the interface.

Semiconductors, however, offer a possible new regime. The bulk band structures of most III–V semiconductors are very compatible – the variations in bandgap may be substantial, but the orbital wavefunctions are very similar. Current growth techniques permit the fabrication of atomically smooth interfaces between two semiconductors, so long as the lattice constants are almost identical. Thus one might expect the spin current to be transferred across such interfaces without tremendous degradation.

The transfer process for spin current across the GaAs/ZnSe interface has been investigated in detail. Here the ZnSe material was n-doped, and GaAs layers that were n-doped, p-doped, and semi-insulating were investigated. Electron spin polarization was generated optically in the GaAs layer. For n-doped GaAs in the absence of an electric field only a small signal for spin-polarized electrons in the ZnSe was detected. An applied electric field, however, which drove the electrons within the GaAs towards the interface, produced a substantial spin-polarized population in the ZnSe. This indicates that spin-polarized packets can be created in one semiconductor, and then controllably moved into a second semiconductor with an applied electric field. This ability may become a key component of manipulating spin in semiconductors.

One reason the p–n diode is a paradigm of charge-based semiconducting electronics is that it exhibits the crucial phenomenon of minority spin injection. Manipulation of minority spin injection permits great control of charge current, for its amplitude is very sensitive to barriers (such as those introduced by gates). Spin-polarized minority carrier injection has not been shown yet experimentally, however the complementary process of minority carrier extraction has. This process corresponds to optically generating spin-polarized electrons in p-doped GaAs, where the electrons are minority carriers. The built-in electric field of the p–n junction drives these electrons out of the GaAs into the n-doped ZnSe with great efficiency. Thus one may imagine that spin-polarized minority carrier injection is plausible in these systems.

A hybrid of these two situations is the spin-polarized light-emitting diode. Here spin-polarized carriers (either electrons or holes) are electrically injected into the insulating region of p–i–n structure. In the insulating region they recombine with unpolarized carriers of the other species and emit light. Due to the selection rules described before, the emitted light is circularly polarized, and the degree of circular polarization is an indication of the spin-polarization of the recombining carriers. This device demonstrates that large numbers of spin-polarized carriers can be accumulated in the insulating region, although the spin-polarized carriers are not minority carriers.

Methods of manipulating electron spin coherence externally without electric fields are also under investigation. One demonstrated technique relies on the shift of energy levels in a semiconductor under illumination by laser light with frequencies below the fundamental bandgap. When the laser light is circularly polarized this ‘ac Stark effect’ shifts the energy levels of spin-up states and spin-down states differently, generating an effective magnetic field. Rotations of electron spin by angles greater than $\pi/2$ have been demonstrated in this ultrafast effective magnetic field in semiconductor quantum wells.

Transfer of Spin Coherence from Mobile Electrons to Other Spin Systems

Another fundamental phenomenon in spin coherence is the transfer of spin coherence from one system to another. If one spin system has an extremely long coherence time, this transfer can be viewed as establishing a reservoir of spin coherence that can be drawn on at later times. Explorations of transfer phenomena have focused on establishing spin coherence in a system that is coupled to the conduction electrons, such as core electron polarization or nuclear polarization. The polarization of these other spin systems can be detected because, when polarized, they exert an effective field on the conduction electrons. For example, when the nuclear system is polarized it exerts an effective field on the electrons through the hyperfine coupling.

It is also possible to polarize these other systems indirectly through their interaction with the conduction electron system. When the conduction electron system is driven far from equilibrium by optically pumping the spin polarization, the coupled spin systems also become spin polarized. In the case of the nuclear spin system this effect is the Overhauser effect. While continually generating conduction electron spin polarization with optical pumping the polarization of the nuclear system begins to rise with a time-scale characterized by the coupling between the nuclear spins and the conduction electron spins (hyperfine coupling). Because of the substantial difference between the nuclear spin coherence times parallel to and perpendicular to the applied magnetic field, the effective nuclear polarization is generated parallel to the applied magnetic field independent of the orientation of the conduction electron spin polarization. In bulk GaAs at 5 K the percentage polarization achieved was a few per cent, yielding an effective field through the hyperfine interaction of ~ 0.4 T. These nuclear polarizations and effective fields are increased over an order of magnitude in certain quantum wells, suggesting that nanostructure design can be used to control the effective fields. The induced nuclear polarization persists for time-scales characteristic of the nuclear longitudinal spin coherence time, on the order of minutes. If the optical pumping is turned off, the conduction electron spin polarization decays away rapidly, whereas the nuclear spin polarization persists for time-scales characteristic of the nuclear longitudinal spin coherence time, on the order of minutes. This nuclear hyper-polarization, and its manipulation through an optically pumped time-dependent conduction electron spin polarization, suggests a possible general technique of ultrafast all-optical nuclear magnetic resonance in solids.

A similar spin-coherence storage effect has been obtained for core electron polarization in Mn-doped ZnSe quantum wells. Again, the optically pumped conduction electron spin polarization generates a nonequilibrium spin polarization of the paramagnetic core electron spins (3d in the case of Mn) through the core-conduction electron coupling. This core electron spin polarization persists long after the conduction electron spin polarization (and, in undoped systems, long after the conduction electron density) has decayed away. The magnetic polarization of the paramagnetic Mn spins dephases over the very long time-scales determined by spin–lattice relaxation, and thus is an interesting avenue towards information storage.

The above systems are not ferromagnetic – the spin systems involved in the process are paramagnetic spin systems coupled to the conduction electrons, which are 'super-polarized' through the conduction electron spin polarization. An alternative approach to manipulating the conduction electrons has recently been demonstrated using a thin layer of a ferromagnetic material placed on top of n-doped GaAs. Here the presence of this ferromagnetic material, even if only a few nanometers thick, leads to a remarkable enhancement of the nuclear spin polarization in the GaAs. For example in the presence of a mere 1000 G applied magnetic field, the ferromagnetic layer generates an effective field on the electrons from the nuclei of up to 9000 G. This effective field also depends on the orientation of the magnetism in the ferromagnetic layer. This intriguing effect has yet to be fully explored, but may yield a new type of magnetic 'gate' on the mobile electrons in a semiconductor without a direct form of electrical spin injection.

The Future of Semiconductor Spintronic Devices

Semiconductor charge-based electronics (e.g., transistors and integrated circuit chips) and magnetoelectronics (e.g., hard disk read heads) are two of the most successful electronic technologies of the twentieth century. A union of elements of these two technologies is anticipated in the area of semiconductor spintronics. Immediate approaches involve placing hybrid magnetic elements within known semiconductor devices. Semiconductor spintronic devices, however, are unlikely to compete with the most successful devices from each of these two areas – the semiconductor transistor is an extremely effective and efficient device, as is the magnetoelectronic hard disk read head. As has occurred with many other new technologies perhaps the most important applications of semiconductor spintronics have not even been envisaged yet.

Despite this, one significant application without a dominant technology suggests itself – that of quantum computation. Recent advances in the theoretical understanding of quantum information technology, including novel algorithms, error-correcting codes, and plausible scalable device architectures, suggest that calculations involving quantum mechanical coherent states are possible. Here the extremely long spin coherence times in semiconductor systems, and the demonstrated ability to manipulate that coherence on ultrafast optical times, indicate that spins in semiconductors (whether electron or nuclear) may constitute a good basis for a physical realization of a quantum computer.

See also

Quantum Optics: Quantum Computing with Atoms. **Semiconductor Physics:** Outline of Basic Electronic Properties.

Further Reading

Abragam A (1961) *Principles of Nuclear Magnetism*. Oxford, UK: Oxford University Press.

Awschalom DD and Kikkawa JM (1999) Electron spin and optical coherence in semiconductors. *Physics Today* 52: 33–38.

Awschalom DD, Samarth N and Loss D (eds) (2002) *Semiconductor Spintronics and Quantum Computation*. New York: Springer.

Ivchenko EL and Pikus GE (1997) *Superlattices and Other Heterostructures*, 2nd edn. New York: Springer.

Jeffries CD (1963) *Dynamic Nuclear Orientation*. New York: Wiley.

Meier F and Zachachrenya BP (eds) (1984) *Optical Orientation*. Amsterdam, The Netherlands: North-Holland.

Slichter CP (1992) *Principles of Magnetic Resonance*, 3rd edn. New York: Springer.

Wolf SA, Awschalom DD, Buhrman RA, *et al.* (2001) Spintronics: a spin-based electronics vision for the future. *Science* 294: 1488–1495.

Yafet Y (1963) g factors and spin-lattice relaxation of conduction electrons. In: *Solid State Physics*, vol. 14. New York: Academic Press.

Surface Photovoltage Spectroscopy of Semiconductors

L Kronik, Weizmann Institute of Science, Rehovot, Israel

Y Shapira, Tel-Aviv University, Tel-Aviv, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Surface photovoltage spectroscopy (SPS) is a well-established contactless technique for semiconductor characterization, which relies on analyzing illumination-induced changes in the surface voltage. SPS traces its origins to the pioneering work of Brattain and Bardeen in the early 1950s, and was extended into a powerful spectroscopic tool by Gatos, Lagowski and Balestra in the early 1970s. It has been used as an extensive source of surface and bulk information on various semiconductors and semiconductor interfaces. In the following, we present the basic theory behind SPS, its experimental setup, and a range of its applications.

Principles of Operation

In general, a surface is defined as a boundary of media with different physical properties. The periodic structure of an ideal crystalline semiconductor results in the appearance of allowed energy bands separated by forbidden energy gaps. In particular, semiconductors are characterized by a nearly full ‘valence band’, separated by a forbidden bandgap from a higher lying, nearly empty ‘conduction band.’ Electron transport takes place within the latter and hole transport takes place within the former. The mere termination of this periodic structure at the surface, as well as the chemical changes in the surface associated with this termination, allow for the formation of surface-localized electronic states within the semiconductor bandgap. Electrons or holes in these local states are referred to as ‘trapped carriers’, as opposed to the ‘free carriers’ in the conduction and valence bands.

The appearance of surface-localized states induces charge transfer between bulk and surface in order to establish thermal equilibrium between the two. The free carrier density in the vicinity of the surface therefore deviates from its equilibrium value in the bulk. This results in a surface space charge region (SCR), which is electrically non-neutral, implying a nonzero electric field in it and therefore a potential drop across it. Thus, even under equilibrium conditions the surface potential, V_s, is different from the electric potential far away in the bulk. As a specific example which illustrates these concepts, consider an n-type semiconductor with a depleted surface. The majority carriers are electrons and some of them have been trapped in surface states, such that their concentration in the vicinity of the surface is smaller than its equilibrium value but larger than that of the holes (see Figure 1a). The potential drop across the SCR is manifested by the bending of the semiconductor bands, which is such that electrons are repelled from the surface and holes are attracted to it, due to the trapped surface electrons. We note that, by definition, the energy band edge is lower the higher the electrical potential, so that a positive V_s corresponds to downward-bent bands.

Generally, the photovoltaic effect comprises an illumination-induced change in the equilibrium potential distribution and is typically the result of some charge transfer and/or redistribution within the device due to the incident illumination. This is the basic principle behind devices ranging from photodiode detectors to photovoltaic solar panels. A specific variant of the photovoltaic effect is the *surface* photovoltaic effect, which is at the center of this text. It is important to note that the formation of a surface photovoltage (SPV) occurs only if carrier photogeneration *per se* is followed by net charge redistribution. Usually, no significant driving force for such redistribution is found beyond the SCR and the underlying bulk remains quasineutral. Thus, we are interested primarily in the surface SCR.

The SPV mechanism depends strongly on whether the incident photon energy is larger or smaller than the semiconductor bandgap (denoted below as the super-bandgap and sub-bandgap, respectively). Dominant SPV mechanisms are demonstrated below for a depleted n-type surface. The dominant mechanism by which super-bandgap illumination results in surface photovoltage generation is shown in Figure 1b. The electric field in the SCR causes excess electrons to be swept away from the surface and excess holes to be swept towards it. This serves to reduce the density of surface-trapped electrons and decreases the band-bending. In a second mechanism, shown in Figure 1c, either electrons or holes are preferentially trapped at surface defects. This effectively charges the surface and thus modifies the

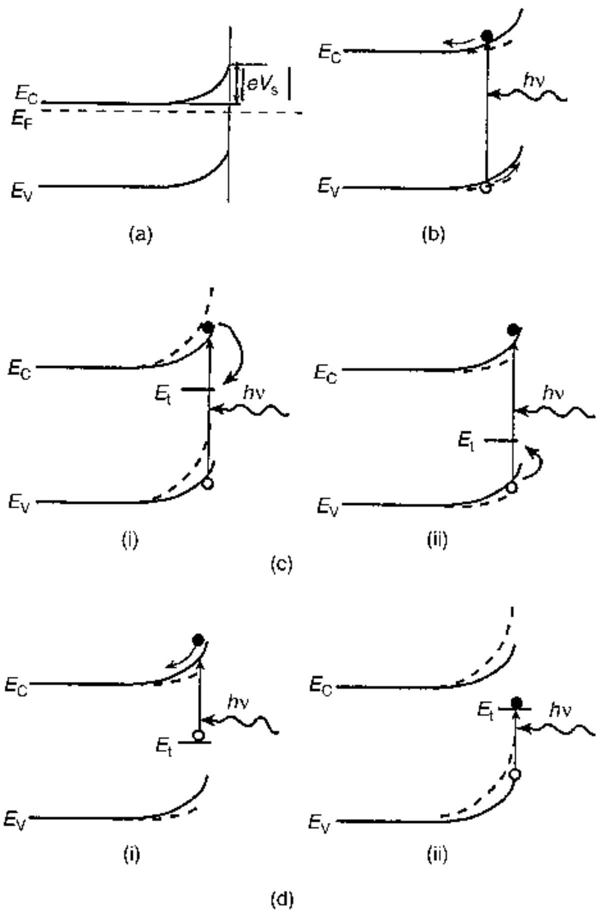

Figure 1 Schematic band diagrams of the surface space charge region at a depleted n-type semiconductor surface under different conditions: (a) thermal equilibrium; (b) super-bandgap illumination with carrier separation under an electric field; (c) super-bandgap illumination with preferential trapping of: (i) electrons, (ii) holes; (d) sub-bandgap illumination with excitation of trapped: (i) electrons, (ii) holes. In all diagrams, solid and dashed lines indicate band positions in the dark and under illumination, respectively. Straight arrows denote carrier generation and curved arrows denote carrier trapping. Adapted from Gal D, Mastai Y, Hodes G and Kronik L (1999) Bandgap determination of semiconductor powders via surface photovoltage spectroscopy. *Journal of Applied Physics* 86: 5573. Reproduced by permission of the American Institute of Physics.

surface potential. This second mechanism is usually apparent in bulk samples only if the equilibrium surface band-bending is fairly small, but increases in importance in polycrystalline materials with decreasing crystallite size, because of the increasing surface to volume ratio.

The most common mechanism for sub-bandgap SPV involves the direct modification of the surface charge, and hence potential, by excitation of trapped carriers, as shown in Figure 1d. Illumination by photons with energy $h\nu > E_c - E_t$ may produce electron transitions from a surface state at an energy E_t into the conduction band, where they are quickly swept to the semiconductor bulk by means of the surface electric field. Hence, the negative surface charge is reduced and the band-bending decreases. Conversely, illumination by photons with energy $h\nu > E_t - E_v$ may produce electron transitions from the valence band into a surface state situated at an energy E_t above the valence band maximum, E_v (which are equivalent to hole transitions from the surface state to the valence band). Such transitions increase the surface negative charge and therefore the surface band-bending. For a SPV to develop in this case, it is necessary to have a significant diffusion of the excess holes into the bulk and/or a significant recombination of electrons and holes inside the SCR, as there is no field-assisted driving force of holes into the bulk.

A typical SPV spectrum shows the SPV as a function of incident photon energy. Its most important features – the ones to watch for when attempting to interpret an SPV spectrum – are the sharp slope changes ('knees') associated with an abrupt onset of an additional carrier excitation mechanism. Consider, as an illustrative example, the top SPV curve in the right-hand part of Figure 2 (the rest of the figure is discussed below). In this curve, two 'knees' are shown. The first 'knee' (marked by an arrow) occurs at sub-bandgap photon energies and corresponds to the onset of carrier excitation from surface states (as in Figure 1d). The second 'knee' occurs in the vicinity of the bandgap energy and corresponds to the onset of band-to-band carrier excitation (as in Figure 1b,c). The proper interpretation of the energy position of these 'knees', their sign (negative or positive SPV change), and the relation of both to the SPV mechanism at work, is elaborated below in the 'Applications' sections.

So far, our discussion has been devoted solely to a study of semi-infinite, uniform bulk samples. However, actual semiconductor samples may be thin, i.e., not much larger than the diffusion length and/or the absorption length, so that excess carriers may be present near their back part as well. Many optically and electronically interesting structures (advanced detectors, transistors, lasers) consist of a multilayer structure, which has several buried interfaces.

Any change in the potential drop across a buried surface will also be measurable at the front surface, because potential drops are additive. For example, the photo-induced potential drop across a photodiode is measured at its contacts, which can be many microns away from its active region. Thus, the sensitivity of SPV measurements to a certain region in the semiconductor is limited only by the

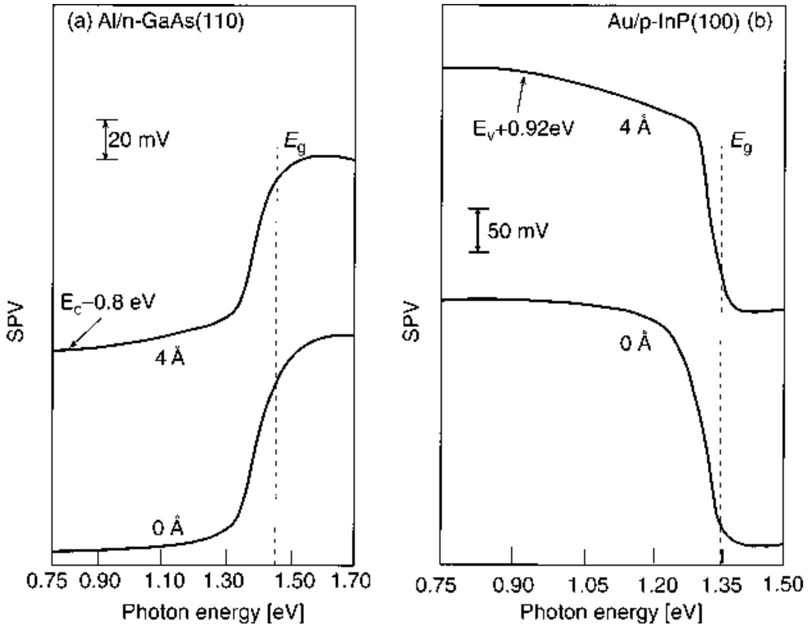

Figure 2 SPV spectra of: (a) an n-type GaAs(110) surface, UHV cleaved (top curve) and with an Al overlayer (bottom curve); (b) a p-type InP(110) surface, UHV cleaved (top curve) and with an Au overlayer (bottom curve). Arrows denote the onset of optically induced transitions and labels denote the corresponding defect level. Adapted from Burstein L, Bregman J and Shapira Y (1991) Characterization of interface states at III–V compound semiconductor–metal interfaces. *Journal of Applied Physics* 69: 2312. Reproduced by permission of the American Institute of Physics.

absorption length of the photons, i.e., by the possibility of introducing an excess of free carriers in the region of interest.

Experimental Details

SPV measurements are nontrivial because the surface potential is a *built-in* potential, rather than an external potential. In other words, the surface potential is not equal to the difference in Fermi levels between the front and back surfaces. Therefore, it cannot be simply measured using some form of voltmeter. Moreover, in the case of a free surface, the application of any contact to indirect electrical measurements of the built-in voltage (e.g., current–voltage or capacitance–voltage analyses) will invariably alter the surface properties and hence the quantity under measurement. Consequently, most of the prevalent techniques for measuring the surface potential (or at least changes of it upon illumination) are based on capacitive coupling schemes that do not require a direct electrical contact.

The two most commonly used methods rely on bringing a metallic probe close to the free surface, but never touching it, thus forming a parallel-plate capacitor. An alternating current, from which surface potential changes can be calculated, is then excited by (a) vibrating the metallic probe, which results in alternating capacitance, or (b) using ‘chopped’ illumination. The two approaches are known as the Kelvin probe and the ac-SPV schemes, respectively. In principle, they provide similar information, but their somewhat different physical principles result in several relative strengths and weaknesses, which make the method of choice application-dependent. In a nutshell, the Kelvin probe is more suitable for following phenomena with large time constants (in particular surface states with long relaxation times) and does not perturb the true band-bending of the free surface. The ac-SPV technique is more suitable for following fast relaxation phenomena, allows for a systematic modification of the surface band-bending in the dark as an extra degree of freedom, and does not require an ohmic back-contact.

Both Kelvin probe and ac-SPV measurements may also be employed in a scanning mode, which allows for recording lateral variations in the SPV, by scanning a small probe or a well-focused light beam, respectively, across the surface. Both approaches are typically limited to a lateral resolution roughly of the order of ~1 μm. SPV measurements can also be performed by a limited modification of two high-resolution techniques for topographic mapping of surfaces that have emerged in the 1980s: scanning tunneling microscopy (STM) and atomic force microscopy (AFM). This opens new horizons for SPV mapping with a resolution on the nanometer scale or even the atomic scale.

A schematic view of a complete, generic SPS setup is given in Figure 3. The sample under study is typically placed inside a metallic box, which serves as both a dark box and a Faraday cage. This box must be such that it effectively includes the probe.

Clearly, for spectroscopic purposes one needs to generate the SPV signal using a broadly tunable light source. The default choice would therefore be the use of a 'white-light' source in conjunction with a monochromator. Because the typical super-bandgap absorption coefficient is typically orders of magnitude larger than the sub-bandgap coefficient, SPS is vulnerable to the spurious contribution of high-order diffraction peaks and stray light. Use of a monochromator where these effects are as small as possible is therefore highly recommended. For some applications, use of an auxiliary light source (e.g., white light, laser light, ultraviolet light, etc.) allows for a 'photobias' of the system and/or for additional quantitative analysis.

The actual recording of an SPV spectrum is coordinated by a real-time computer program. This program controls the monochromator motor and thus wavelength changes, operates the control electronics of the probe, reads the measurement results by interfacing with the probe read-out electronics via an analog-to-digital converter, and finally stores the obtained spectrum for subsequent display and analysis.

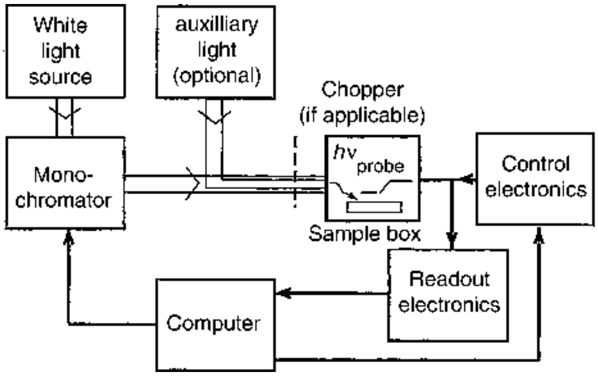

Figure 3 Schematic block diagram of a generic SPS setup. Adapted from Kronik L and Shapira Y (2001) Surface photovoltage spectroscopy of semiconductor structures: at the cross-roads of physics, chemistry and electrical engineering. *Surface and Interface Analysis* 21: 954. New York: John Wiley and Sons, Inc.

Applications I: Uniform Semiconductors

An elementary application of SPS is the approximate determination of the semiconductor bandgap. This determination is based on the large increase in absorption coefficient near the bandgap energy, E_g, found in most semiconductors. This increase brings about a significant change of the SPV signal, which is easily identified as a sharp change in slope of the SPV curve and is often the most significant one in a given spectrum. An example is shown in Figure 2, where the slope changes related to the bandgaps of GaAs and InP are clearly identified. Indeed, the bandgaps of most known semiconductors have been analyzed using this approach.

The use of SPS for extraction of E_g is at heart nothing more than an emulation of an absorption spectrum. However, as opposed to, e.g., transmission spectroscopy, SPS does not require light collection. It can therefore be performed on arbitrarily thick samples (or on layers sandwiched within a heterostructure – see below) and does not require the sample to be removed from the substrate or grown on a transparent one. It is also inherently insensitive to reflection and scattering and is thus eminently useful for micro- and nanocrystallites.

One must bear in mind that the value of E_g obtained from an SPS plot by inspection is only approximate. Indeed, detailed experimental comparisons between SPV and absorption spectra reveal that the two are often similar, but never identical. Thus, the nominal bandgap is nearly always found within the onset of the largest SPV signal, but is usually relatively broad, the exact position of E_g within it is by no means obvious, and the error in E_g is often ~0.1 eV. However, it has been shown that this error can be reduced significantly – often by one to two orders of magnitude – by a careful quantitative analysis of the experimental data.

We now turn our attention to determining the type (p or n) of the semiconductor. This may be achieved with the help of the sign of the knee associated with the onset of the super-bandgap SPV. Most semiconductor surfaces are depleted, which means that the bands of p-type semiconductors are bent downwards towards the surface, whereas the bands of n-type semiconductors are bent upwards. Since super-bandgap illumination typically tends to decrease the surface band-bending, this would result in a positive SPV in n-type semiconductors and a negative SPV in p-type semiconductors. For example, the SPV spectra of n-type GaAs and p-type InP, shown in Figure 2, clearly feature opposite onset signs which obey the above rules.

While in many cases the type of the semiconductor is known a priori, this is not always the case and SPS can become very useful in determining the semiconductor type. For example, SPS was used for fast and nondestructive verification of the semiconductor type in GaN films, where the p-doping is known to be nontrivial and subject to compensation by its native n-type doping.

We note that under conditions where the SPV is dominated by trapping of carriers, as shown in Figure 1c, the sign convention described above clearly does not apply. For example, etched CdSe quantum dot films were found to exhibit a p-type response in a humid ambient and an n-type response in a dry ambient, on account of preferential trapping of holes or electrons in a dry or humid ambient, respectively.

A key strength of SPS is in the determination of the energy position and type of surface states. Because photons of sufficient energy may excite charge carriers from a surface state to a band or vice versa, we expect a knee in the SPV spectrum whenever the photon energy exceeds the threshold energy of a certain transition. If the latter involves excitation of electrons from the surface state to the conduction band (Figure 1d(i)),a positive change in the surface charge is induced, and hence a positive SPV is expected. Conversely, excitation of holes to the valence band (Figure 1d(ii)) makes the surface charge more negative and a negative SPV is expected. Thus, the combination of the SPV threshold energy and slope sign makes the determination of the approximate position of the surface state within the bandgap possible.

As an illustrative example, consider how SPS can be used to monitor metal-induced surface states, as shown in Figure 2. After cleaving in ultrahigh vacuum, neither GaAs nor InP surfaces display any knees associated with deep surface states, as expected for these surfaces. However, upon deposition of a very thin metallic overlayer, distinct sub-bandgap knees emerge. These knees, associated with positive and negative SPV changes at the Al/GaAs and Au/InP interfaces, respectively, indicate the formation of surface states situated ~0.8 eV below the conduction band edge and ~0.92 eV above the valence band edge, respectively.

Because SPS is contactless and nondestructive, it can be applied not only to clean surfaces, but also to real ones, in practically any ambient. This makes SPS highly suitable for a direct correlation of surface electronic structure with chemical treatments. One such area of active research is the study of surface passivation treatments. An illustrative example, in which the passivating effects of various liquids on a specific p-InP surface (denoted as (100)) was studied *in situ*, is shown in Figure 4. The negative SPV change at ~1.3 eV found in all spectra shown is clearly due to the onset of band-to-band transitions in the InP sample, with its sign indicating a p-type sample, as appropriate. The SPV spectrum of the water-exposed sample (which is similar to that obtained in air (not shown)) features two surface states: the negative SPV slope change, at ~1 eV, is due to a photo-induced population of a surface state situated ~1 eV above the valence band edge. The positive SPV slope change at ~1.2 eV is due to a photo-induced depopulation of a surface state situated ~1.2 eV below the conduction band edge. Interestingly, exposure to HF eliminates the latter state, but not the former. The opposite is true for exposure to H_2O_2. Finally, exposure to a Na_2S solution results in removal of both surface states, in agreement with the known passivating action of this solution. This is just one example of many illustrating the strength of SPS for revealing the link between chemical/physical treatments and device-affecting electronic properties, already at the material level.

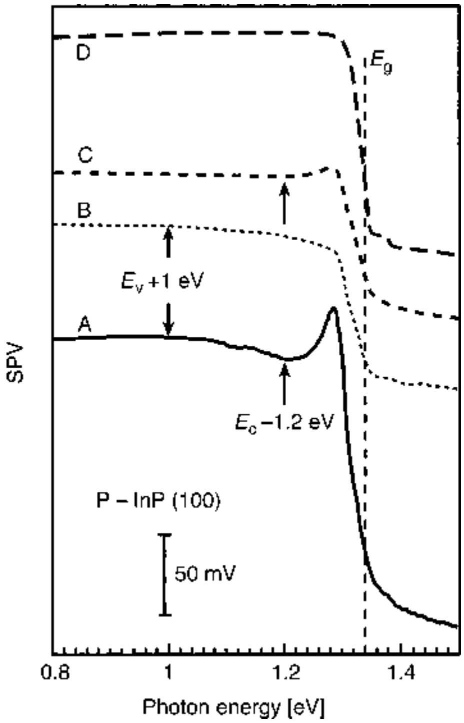

Figure 4 SPV spectra of the p-InP(100) surface in air and after successive immersion in (from top to bottom in the lower energy part): A – 18 MΩ H_2O; B – 0.4% HF; C – 30% H_2O_2; D – a 1.3 M Na_2S solution. Arrows denote the onset of optically induced transitions and labels denote the corresponding defect level. Adapted from Bastide S, Gal D, Cahen D and Kronik L (1999) Surface photovoltage measurements in liquids. *Review of Scientific Instruments* 70: 4032. Reproduced by permission of the American Institute of Physics.

While we have so far considered only surface states, SPS is sensitive to bulk states too, as long as charge excited from or trapped in them can be separated (typically under the electric field of a space charge region). Accordingly, SPS has been used for studying, e.g., Cr-related defects in GaAs or Fe-related defects in InP. The simplest distinction between surface and bulk states in SPS is obtained by examining whether the pertinent spectral feature disappears after surface (e.g., etching) or bulk (e.g., doping) treatment, respectively.

We note that in recent years, SPS has also been applied successfully along the lines outlined here for the study of bandgaps, doping type, and surface and bulk defects in the emerging field of organic semiconductors and organic treatments of semiconductors, as well as for nanostructures, such as fullerenes and nanotubes.

Additional, quantitative analyses of the SPV spectra allow for a determination of densities, thermal and optical transition rates, and the spatial site of gap states. Such analyses are usually based on the SPV response to a systematic modification of one or more external parameters, e.g., time (for transient responses), illumination intensity, temperature, etc.

Applications II: Multilayer Structures

Beyond the study of surface properties, the application of SPS to heterojunctions, multilayer structures, and actual device structures has been most rewarding in recent years. The SPS of multilayer structures is illustrated by Figure 5. This figure features the SPV spectra of a ZnO:Al/ZnO/CdS/Cu(In,Ga)Se$_2$ layered structure, used for the fabrication of thin-film solar cells. Distinct spectral knees at ~1 eV, ~2.4 eV, and ~3.1 eV, corresponding to the bandgaps of Cu(In,Ga)Se$_2$, CdS, and ZnO, respectively, are clearly observed. The unequivocal identification of all three bandgaps confirms the above-mentioned theoretical prediction of SPS sensitivity to buried interfaces. Interestingly, the magnitude of the SPV signal is seen to increase with increasing annealing time (in air, at 200 °C), in agreement with the increase in overall solar cell efficiency with such annealing.

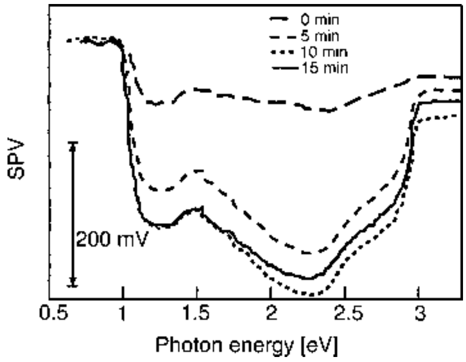

Figure 5 SPV spectra of ZnO:Al/ZnO/CdS/CIGS structures for various annealing times, in air, at 200 °C. Adapted from Gal D, Beier J, Moons E, *et al.* (1995) Band diagram and band line-up of the polycrystalline Cds/Cu(In,Ga)se$_2$ heterojunction and its response to air annealing. *AIP Conference Proceedings* 353: 453. Reproduced by permission of the American Institute of Physics.

Another feature of growing importance in SPS is that it allows for the characterization of quantum wells and other low-dimensional heterostructures. An illustrative example of the SPS of a multiple-quantum-well (MQW) GaAs/AlGaAs structure is shown in Figure 6a. The spectral knees at ~1.4 eV and 1.78 eV correspond to the onset of band-to-band absorption in the GaAs epilayer and the AlGaAs cap layer, respectively. In between these two knees, the shape of the SPV spectrum resembles typical absorption spectra of MQW structures and the observed MQW-related maxima agree well with calculated energy values of the hole level-electron level transitions denoted as 1HH-1E and 2HH-2E transitions. Note that before etching, a nonnegligible sub-bandgap SPV signal was apparent at photon energies below 1.38 eV, indicating the presence of optically active gap states. Since no such signal was observed after etching, these states were located at the external AlGaAs surface and were not associated with the MQW region. Moreover, the SPV spectrum in the MQW absorption energy range has not changed at all (other than undergoing a uniform shift) after etching, showing that the signal SPV in the MQW absorption range is not interfered by surface effects. Thus, all allowed hole–electron transitions have been resolved, at room temperature.

As SPV analyses of multilayer structures offer the possibility of performing contactless and nondestructive electrical and optical characterization of buried interfaces, they have found significant uses in the field of process monitoring and quality control of actual device structures. This is because the possibility of assessing the quality of a layered structure before any actual device processing takes place presents an opportunity for significant savings in time and money. Recent applications include the identification of defective batches and prediction of ultimate performance of solar cells and x-ray detectors, the prediction of energy levels and lasing wavelength in modern laser heterostructures, the assessment of bandgap narrowing and ultimate gain of heterojunction bipolar transistors, and the

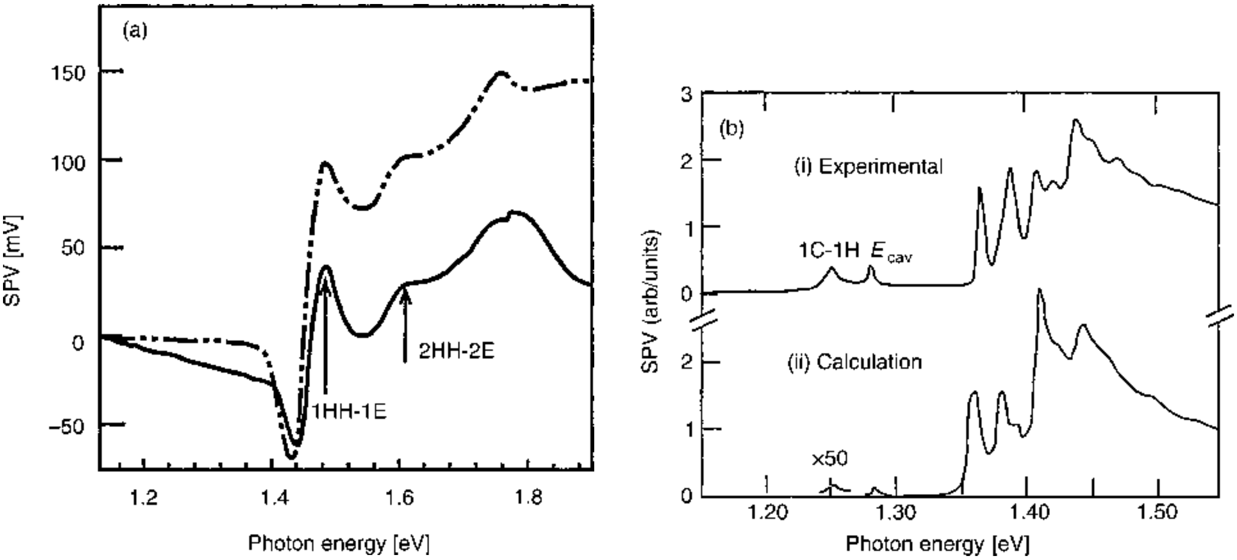

Figure 6 (a) SPV spectra of a GaAs/AlGaAs MQW sample before (solid curve) and after (dashed curve) etching. Adapted from Bachrach-Ashkenasy N, Kronik L, Shapira Y, *et al.* (1996). Surface photovoltage spectroscopy of quantum welts and superlattices. *Applied Physics Letters* 68: 879. Reproduced by permission of the American Institute of Physics. (b) Room temperature experimental SPV spectra (i) and calculated photo-excited carrier spectra (ii) of an InGaAs/GaAs/AlGaAs vertical-cavity surface-emitting laser structure. Adapted from Huang YS, Malikova L, Pollak FH, *et al.* (2000). Surface photovoltage spectroscopy, photoreflectance, and reflectivity characterization of an InGaAs/GaAs/GaAlAs vertical-cavity surface-emitting laser including temperature dependence. *Applied Physics Letters* 77: 37. Reproduced by permission of the American Institute of Physics.

characterization of the electronic structure of high electron mobility transistors and vertical cavity surface emitting lasers.

As an example of such studies, consider the recent characterization of an In-GaAs/GaAs/AlGaAs vertical-cavity surface-emitting laser, shown in Figure 6b. It is readily observed that SPS resolves both the gain providing 1C-1H transition and the lasing-wavelength-related cavity mode, as well as a rich oscillatory structure above ~1.37 eV due to interference effects from the distributed Bragg reflector stacks. The measured SPV spectrum is also in very good agreement with the calculated dependence of the photo-excited carrier density on photon energy, also shown in Figure 6b. In fact, SPS is the first contact-less technique to resolve all of these features in one measurement, and it also allows for a temperature-resolved determination of these quantities.

Summary

SPS features a theory that is relatively straightforward (at least at the qualitative level), and offers typical experimental setups that make the method contactless, nondestructive, ambient insensitive, and capable of any lateral resolution. It can provide a detailed picture of the electronic structure at surfaces, interfaces, heterojunctions, and even complicated multilayer structures. The wide range of SPS applicability also makes it easy to compare and link changes in electronic structure to changes in chemical/physical properties on the one hand, and electrical/optical device properties and performance on the other.

See also

Semiconductor Materials: Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Brattain WH and Bardeen J (1953) Surface properties of germanium. *Bell System Technical Journal* 32: 1–41.

Brillson LJ (1982) The structure and properties of metal–semiconductor interfaces. *Surface Science Reports* 2: 123–326.

Gatos HC and Lagowski J (1973) Surface photovoltage spectroscopy – a new approach to the study of high-gap semiconductor surfaces. *Journal of Vacuum Science and Technology* 10: 130–135.

Kronik L and Shapira Y (1999) Surface photovoltage phenomena: theory, experiment, and applications. *Surface Science Reports* 37: 1–206.

Kronik L and Shapira Y (2001) Surface photovoltage spectroscopy of semiconductor structures: at the crossroads of physics, chemistry and electrical engineering. *Surface and Interface Analysis (special issue on Optical Techniques)* 21: 954–965 (main source for the present text).

Lüth H (1993) *Surfaces and Interfaces of Solids*, 2nd edn. Berlin: Springer-Verlag.

Many A, Goldstein Y and Grover NB (1971) *Semiconductor Surfaces*, 2nd edn. Amsterdam, The Netherlands: North-Holland.

Pollak FH (2000) Modulated electric fields produce semiconductor spectra. *Laser Focus World* 36: S9.

SOLITONS

Contents

Bright Spatial Solitons
Optical Fiber Solitons, Physical Origin and Properties
Soliton Communication Systems
Temporal Solitons

Bright Spatial Solitons

C Conti and G Assanto, University 'Roma Tre', Rome, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical spatial solitons, once considered an exotic and rather abstract concept belonging to the realm of mathematics or mathematical physics, nowadays are pursued both experimentally and theoretically in terms of their rich dynamics and variegated features, with important perspectives in applications to signal processing and switching, routing, filtering and – generally – light-by-light controlling. In an intuitive description of a bright spatial soliton, a self-focusing response balances natural diffraction of a light beam, allowing for its invariant propagation in transverse size and intensity. Spatial solitons can be regarded as the stationary or time-independent self-guided 'eigenwaves' of Maxwell equations in a material system with a nonlinear optical response. Basically, 'bright' spatial solitons (we will not address 'dark' cases here, i.e., a dip on an illuminated background) bear the same relationship to nonlinear optics in space as temporal, or fiber, solitons do in time. While both temporal and spatial solitons are special cases of light localization in the $(3 + 1)$ dimensions (light bullets) encompassing time and two transverse coordinates in space in addition to propagation; the spatial case requires particular attention compared to the fiber inasmuch as, in general, two transverse coordinates (2D) complement the direction of propagation. In other words, the scenario is potentially one-dimension richer than for 'standard' temporal solitons, where light confinement, i.e., the effect of the nonlinearity, takes place along a single coordinate. In this respect, the latter case finds a direct counterpart with spatial solitons in planar waveguides, i.e., $(1 + 1)$D solitons, where light confinement occurs linearly in one spatial dimension, providing group-velocity dispersion in time is replaced by beam diffraction in space, respectively.

When both transverse dimensions are involved, i.e. when a light beam propagates in bulk, in the simplest case of Kerr self-focusing, the nonlinearity can become overcritical and cause filamentation and catastrophic collapse. Additional mechanisms (such as saturation, nonlocality, or higher-order effects) or inherently different nonlinear responses (parametric or molecular), need therefore to be involved in the description of spatial solitons. In several instances, the mathematics differs from the nonlinear Schrödinger equation originally invoked to define a 1D bright optical soliton, leading to nonintegrable systems and to a wider class of solutions, called solitary waves. Since solitary waves constitute the vast majority of the observed spatial light-localization phenomena, we will refer to spatial solitons even in those cases when their model is nonintegrable, defining them as non-diffracting beams in a self-focusing material. While retaining the distinction between 1D and 2D cases, in the following sections we will discuss bright spatial solitons and their features, grouping them on the basis of the underlying nonlinear mechanism and corresponding model, trying to provide a direct link with experimental observations. In doing so, we will leave out discussions on dark solitons and modulationally unstable phenomena, as well as more involved subjects ranging from higher-order to vector, vortex, photonic-crystal, and incoherent spatial solitons.

The General Model of Bright Spatial Solitons

Assuming the paraxial approximation to be valid, the general structure of the evolution equations which lead to spatial solitons propagating along z takes the so-called Foch–Leontovich form:

$$2\mathrm{i}k\frac{\partial A}{\partial z} + \nabla_\perp^2 A = -2k^2\frac{\Delta n}{n_0}A \qquad [1]$$

where A is the complex amplitude of the linearly polarized electric field, $k = \omega/c = 2\pi n_0/\lambda$, with λ the wavelength, $\omega = 2\pi\nu$ the pulsation, n_0 the bulk refractive index, and $\Delta n(\mathbf{r})$ the index change induced by the nonlinear effect. In anisotropic materials, both n_0 and Δn may depend on the polarization of the envelope A. The latter is normalized such that the optical intensity is $I = |A|^2$. Hence the whole optical field is

$$E_{\mathrm{opt}} = \mathrm{Re}\left[\sqrt{\frac{2Z_0}{n_0}}A\mathrm{e}^{-\mathrm{i}\omega t + \mathrm{i}kz}\right] \qquad [2]$$

with Z_0 being the vacuum impedance.

Hereafter we will neglect material losses and assume $\Delta n(\mathbf{r})$ to be real-valued. In the simplest case it is a function of the field intensity I calculated in $\mathbf{r}$. Typical models are the Kerr medium with $\Delta n = n_2 I$, or the saturable Kerr with $\Delta n = n_2 I/(1 + I/I_s)$ and I_s a saturation value; they are particular cases of local nonlinear media defined by a relation of the type $F[\Delta n(\mathbf{r}), I(\mathbf{r})] = 0$, with F a 'well-behaved' function.

There is also the possibility that the refractive index modification depends on the electromagnetic perturbation (e.g. the intensity) in a finite region, as happens with nonlocal solitons. In the latter case, F needs to be substituted by an integro-differential system of equations, i.e., it represents a functional between $\Delta n(\mathbf{r})$ and $I(\mathbf{r})$.

Relevant exceptions to eqn [1] are spatial solitons due to a quadratic nonlinearity, where a system of two coupled equations must be taken into consideration, and a description in terms of a refractive index change only holds in an approximate, and by no means physical, sense. Quadratic solitons will be discussed later. Here, instead of embarking into a general classification, we will describe the relationships to some representative cases. In this simple framework, spatial solitons are defined as solutions of the type $A = a\exp(\mathrm{i}\beta z)$, with a real-valued and satisfying the equation:

$$\frac{1}{2k^2}\nabla_\perp^2 a + \frac{\Delta n(I)}{n_0}a = \frac{\beta}{k}a \qquad [3]$$

Equation [3] is the z-independent counterpart of eqn [1]. It has the form of the stationary Schrödinger equation (with a field-dependent Hamiltonian operator), where the kinetic energy and the potential are given by the diffraction term and by the nonlinear correction to the refractive index, respectively. Its solutions correspond to solutions of eqn [1] due to the balance between diffraction and nonlinearity. β is the nonlinear correction to the dispersion relation $k = n_0\omega/c$, i.e., it yields the resulting longitudinal wavevector $K = k + \beta$. Solitary solutions are often classified in terms of their dimensionality. Solutions of eqn [3], which depend on either 1 or 2 transverse coordinates, define 1D or 2D solitons, respectively. 1D solitons refer to situations in which a bulk nonlinear medium is illuminated with a highly elliptic beam (such that the diffraction along a transversal dimension, denoted by y, is negligible), or a planar waveguide is excited with a linear mode fixing the transverse profile in one of the dimensions (denoted by y). The nonlinear dynamics and the corresponding steady-state solutions are then described by eqns [1] and [3], with $\partial_y = 0$.

Spatial Solitons in Kerr Media

1D Kerr Spatial Solitons and the Nonlinear Schrödinger Equation

The simplest example of a spatial soliton is an optical beam diffracting in one transverse dimension in the presence of the so-called Kerr effect, i.e., a refractive index that increases linearly with the local intensity I:

$$n = n_0 + n_2 I \qquad [4]$$

where $n_2(>0)$ is measured in $\mathrm{m}^2\,\mathrm{W}^{-1}$ and has typical values of 10^{-20} for silica and 10^{-17} for semiconductors at photon energies below half-bandgap, i.e., leaving aside two photon absorption and higher-order effects. For 1D Kerr spatial solitons, eqn [1] becomes

$$2ik\frac{\partial A}{\partial z} + \frac{\partial^2 A}{\partial x^2} = -2k^2\frac{n_2}{n_0}|A|^2 A \qquad [5]$$

Equation [5] is known as the nonlinear Schrödinger equation (NLS), and is the basic model in describing temporal soliton, i.e., nondispersing pulses propagating in optical fibers, provided the variable x is intended as the temporal coordinate. The importance of eqn [5] is due to several remarkable, in some sense unique, mathematical properties, and to the fact that it can be considered a fundamental model for diverse nonlinear phenomena. This circumstance is often referred to as the 'universality of the NLS'.

In the framework of optical spatial solitons, the NLS relevance is moderated by the fact that most experimentally explored environments involve processes significantly more complicated than the Kerr effect. Nevertheless, it settles the basic principles and constitutes a reference point in optical soliton assessment.

Trying to gain some physical insight, the starting point is the beam spreading when $n_2 = 0$ in eqn [5], i.e., the linear regime corresponding to low intensities. The diffraction is inherent to the spectrum of plane waves in a finite beam (along x). To underline the role of the nonlinear response, one can consider the NLS without the diffraction term (the second derivative with respect to x), obtaining:

$$A(x,z) = \sqrt{I(x)}\exp\left[ik\frac{n_2}{n_0}I(x)z\right] \qquad [6]$$

Equation [6] shows how the nonlinearity yields a phase-front curvature related to the sign of n_2: the medium is self-focusing (a converging beam) for $n_2 > 0$ or self-defocusing (a diverging beam) for $n_2 < 0$ (see Figure 1). In the absence of diffraction, i.e., for a plane wave, the Kerr effect modifies the field phase by increasing ($n_2 > 0$) or decreasing ($n_2 < 0$) the local refractive index (and hence the effective optical path) proportional to the intensity. For a beam of finite transverse extent, nonlinearity and diffraction act together. In a self-focusing medium, the concave phase-front due to diffraction can be balanced by the nonlinearity, and the natural tendency of a beam to spread can therefore be compensated. For this to happen, the length-scales associated to the diffraction, i.e., the Rayleigh range $L_R = kw_0^2$, and the Kerr effect must be comparable. A characteristic length $L_N = n_0/(kn_2I_0)$, over which the nonlinear phase shift equals 1, can be associated

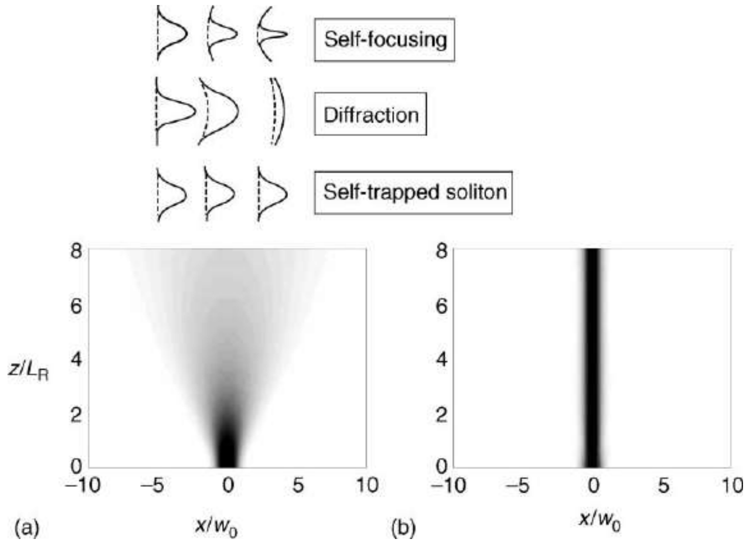

Figure 1 Kerr spatial solitons. Top: sketch of the mechanism leading to Kerr soliton formation. In the absence of diffraction, self-focusing of a Gaussian beam provides a converging phase-front (dashed lines). Conversely, diffraction by itself causes beam divergence. When these two effects act jointly, they can balance each other and soliton generation can be observed, with a flat phase-front. (Adapted with permission from Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6:1419–1426). Bottom: beam propagation method used to simulate the propagation of a one-dimensional Gaussian beam by integrating the nonlinear Schrödinger equation, for (a) low-power input, (b) input high enough to generate a self-trapped solitary wave. The transverse coordinate x is expressed in units of the beam waist, the propagation coordinate z in Rayleigh lengths.

to the nonlinearity, with I_0 the peak intensity across the beam profile. When the two lengths L_R and L_N are nearly equal, i.e., $w_0^2 I_0 \approx n_0/(k^2 n_2)$, the balance between the phase-front curvatures, due to diffraction and nonlinearity, yields a self-localized nondiffracting beam. Such a situation is described by the exact solutions of eqns [1] and [3], which takes the form:

$$A(x,z)=a(x)e^{i\beta z} \quad \text{with} \quad a(x)=\sqrt{I_0}\,\text{sech}\left(x\sqrt{\frac{k^2 n_2}{n_0}I_0}\right) \qquad [7]$$

and $\beta = n_2 I_0/2n_0$ is determined by the peak intensity or, equivalently, by the power P_1 carried by the soliton per unit of transverse wavefront along y:

$$P_1=\int_{-\infty}^{\infty}|A|^2\,dx=\sqrt{\frac{4n_0 I_0}{k^2 n_2}}$$

Note that the overall phase-front is flat, and the spatial extension in width yields $1/(n_2 k^2 I_0)^{1/2}$, as anticipated. The relationship between the soliton width – which determines the strength of diffraction – and the intensity – responsible for the nonlinear phase shift – is known as the existence curve, and is characteristic of the specific nonlinear mechanism.

Equation [7] is the fundamental soliton of the NLS and the paradigmatic 1D spatial soliton. Its sech shape specifically results from the type of nonlinearity; as discussed below, other effects (including parametric generation) can yield different profiles.

Figure 1 shows the simulation of a 1D Kerr soliton generated from a Gaussian input. Self-trapping is attained when the peak intensity or, equivalently, the power, reaches a threshold value depending on experimental details, the size of the nonlinear response, and the beam profile.

Figure 2 shows experimental results on the formation of a 1D Kerr soliton in a semiconductor (AlGaAs, $n_0 \approx 3.5$) waveguide. At low power (center panel) the output beam exhibits diffraction, whereas the soliton (bottom) reproduces the input beam profile (top panel) at the end facet of the sample. These results were obtained at $\lambda = 1.55$ μm, using pulsed laser beams, with temporal pulse durations of about 1 ps and peak powers of about 500 W; the generated 1D solitons had waists of tens of microns.

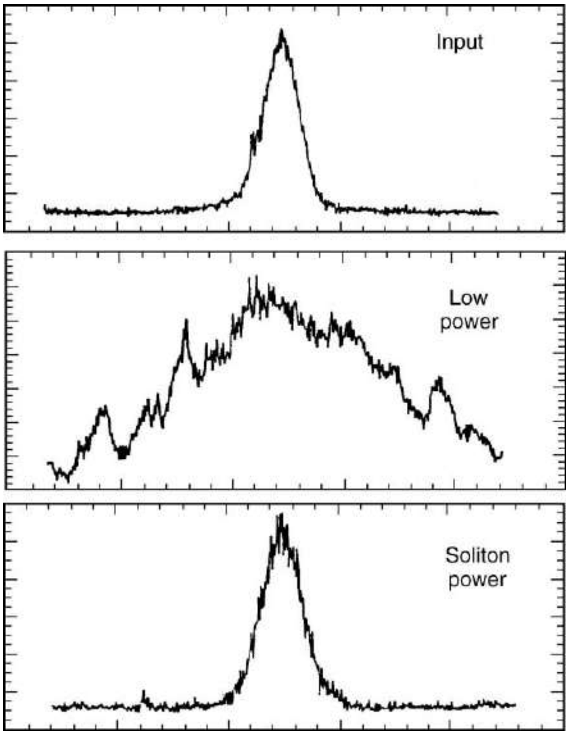

Figure 2 1D Kerr soliton in a planar AlGaAs waveguide. Transverse beam profiles at the input (top) and output facets of a planar waveguide in AlGaAs, at low (center) and high (bottom) power, showing diffraction and soliton formation, respectively. (Adapted with permission from Aitchison JS, Al-hemyari K, Ironside CN, Grant RS and Sibbett W (1992) Observation of spatial solitons in AlGaAs waveguides. *Electronic Letters* 28: 1879.)

In a physically intuitive sense, self-trapping is due to the refractive index increase across the beam profile. Such an increase gives rise to a graded index optical waveguide able to confine all the plane wave components with angles of propagation within the numerical aperture of the input beam. Large spatial-frequency components k_x cannot be trapped and are emitted as radiation during soliton formation. It is possible to experimentally verify that this self-induced waveguide is also able to guide a weak probe uncorrelated to the soliton, provided it is above the cutoff for dielectric confinement. The use of spatial solitons for intensity-controlled or all-optical lightwave circuitry can therefore be envisaged. The first experiments on optical 1D Kerr spatial solitons date back to 1985, and were performed in carbon disulfide. Since then, 1D Kerr spatial solitons have been reported in glasses, semiconductors, and organic materials.

Kerr Spatial Solitons in Two Transverse Dimensions: Saturable and Nonlocal Media

When the dynamics of the optical beam encompasses two transverse dimensions, the idealized discussion above no longer applies. As long as the medium is described by the simple Kerr law (eqn [4]), 2D propagation does not admit stable self-trapped stationary beams. Because of this, two alternatives are available: (i) the beam diffracts; or (ii) the beam progressively self-focuses until it undergoes catastrophic collapse: in the absence of additional limiting phenomena, such as plasma formation or nonparaxial/vectorial coupling, its waist keeps shrinking with a corresponding unbound increase in peak intensity (beam power is conserved). Which regime actually applies, depends on whether the excitation is lower or higher than a self-focusing critical power $P_c = (0.61\lambda)^2/8n_0n_2$. The latter corresponds to a diffraction-driven divergence equal to the critical angle for total internal reflection in the induced waveguide. The self-focusing instability, first predicted by Kelly in 1965, is related to the so-called Derrick's theorem: particularly related to spatial solitons in a positive ($n_2 > 0$) Kerr medium, it states that the equilibrium between diffraction and nonlinearity in 2D is unstable.

The situation is markedly different when saturation is present: self-focusing arrests at the maximum value of the nonlinear refractive index increase, and stable self-trapped beams are allowed. The simplest model for a saturating nonlinearity yields the following stationary-state equation:

$$\frac{1}{2k^2}\nabla_\perp^2 a + \frac{n_2}{n_0}\frac{I}{1 + I/I_s}a = \frac{\beta}{k}a \qquad [8]$$

Its solutions, while retaining a bell-shaped profile, depart from the purely Kerr model corresponding to $I_s \to \infty$, and can only be obtained numerically.

Gases, and specifically sodium vapors, were the first media in which spatial solitons of a saturable Kerr-like nonlinearity were observed in 1974. Historically, these were the first 2D optical spatial solitons.

Most recent developments on optical spatial solitons involve more complicated environments and nonlinear responses. In the next sections we will consider those most investigated experimentally: photorefractive solitons, nonlocal solitons in liquid crystals and parametric solitons.

Spatial Solitons in Photorefractive Media

In recent years, research on spatial optical solitons in photorefractives has been among the most active. Self-trapping in photorefractive materials was first predicted in 1992 and observed in 1993. However, a clear assessment of this result came with the demonstration of the so-called steady-state screening photorefractive solitons in 1994.

The principle underlying the photorefractive response can be described as a feedback mechanism relating to the optical induction of a static electric field **E**. In specific materials, such as semiconductors or ferroelectrics, a light beam generates free charges which diffuse and drift under an externally applied voltage. Owing to free carriers and donor/acceptor sites, the resulting net spatial charge produces a static electric field which, in turn, modifies the refractive index n_0 via the electro-optic effect, phenomenologically described by

$$\Delta n_{ij} = -\frac{1}{2}n_0^3 r_{ijk}E_k - \frac{1}{2}n_0^3\varepsilon_0^2(\varepsilon_r - 1)^2 g_{ijkl}E_kE_l \qquad [9]$$

Equation [9] is a series expansion of the index change in powers of the static electric field: all terms are present if the medium is noncentrosymmetric, otherwise only even-powered terms need be included. In essence, eqn [9] entails the feedback mediated by the static electric field on the optical wave.

The feedback leads to stable solutions under appropriate conditions. These are steady-state solitons if the inhomogeneous light-induced index change makes a suitable waveguide for the beam. The three main types of such self-localized beams are screening, photovoltaic, and centrosymmetric photorefractive solitons. Boundary conditions determine the induced static field to be inserted in eqn [9] and in the propagation eqn [1]. In the simplest cases we consider, the light-wave is linearly polarized and the tensorial nature of the index perturbation in eqn [9] can be neglected.

The equations encompassing most of the physics of the photorefractive effect constitute the Kukhtarev model and can be shown as:

$$\begin{aligned} &s(I + I_b + I_{\text{dark}})(N_d - N_d^i) - \gamma_r\rho N_d^i = 0 \\ &\nabla\cdot\mathbf{J} = \nabla\cdot[q\mu\rho\mathbf{E} + k_B T\mu\nabla\rho + \beta_{ph}(N_d - N_d^i)I\mathbf{c}_{\text{opt}}] = 0 \\ &\nabla\cdot\mathbf{E} + \frac{q}{\varepsilon_S}(\rho + N_A - N_d^i) = 0 \end{aligned} \qquad [10]$$

The relevant quantities in eqns [10] are: the current density **J**; the free charges (electrons) density ρ; the density of ionized (total) donors N_d^i (N_d), the density of negatively charged sites N_A, the photo-ionization cross-section s, the background density I_b (due to background laser light), the dark irradiance I_{dark} (sI_{dark} is the dark generation rate, typically $I_b \gg I_{\text{dark}}$), the recombination rate γ_r, the electron mobility μ, the low frequency dielectric constant $\varepsilon_S = \varepsilon_0\varepsilon_r$, the Boltzmann constant k_B, the versor $\mathbf{c}_{\text{opt}}$ of the optic axis, and the temperature T. In the second of eqn [10], the pertinent element β_{ph} of the photovoltaic tensor needs be taken into account only for the photovoltaic solitons described below. The above must be completed by boundary conditions for the bias V – the line integral of

E – and the geometry of the metal contacts. Typically, $N_d \approx 10^{18}\ \text{cm}^{-3} \gg N_a$, $I_{\text{dark}} \approx 10^{-3}-10^{-6}\ \text{W/cm}^2$, $I_b \approx 10\ \text{W/cm}^2$, $\varepsilon_r \approx 10^3$ and solitons are observed in the visible.

Equations [10] determine the field **E** in terms of the optical intensity I in steady-state, and **E** is inserted in eqn [9] to yield the optically induced index change. A major signature of photorefractive solitons is the cumulative process they rely on: the space-charge field **E** builds up in time along with the distribution of carriers, until equilibrium is reached. Hence the phenomenon does not require high optical intensities but, conversely, a sufficiently long time (of the order of minutes or hours). Photorefractive solitons are typically observed at modest excitations, as low as a few microwatts, with typical waists of tens of microns.

Screening and Photovoltaic Solitons

When an externally applied bias lowers the refractive index of a crystal through its electro-optic response, a light beam can screen this effect, defining a self-waveguide supporting a spatial soliton. In a noncentrosymmetric photorefractive, the beam raises the conductivity in the illuminated region, locally lowering the voltage drop compared to dark regions. The induced space-charge screens the applied static field, and it may bring the refractive index back to its unbiased value. This can be seen as a voltage divider formed by a small optically controlled resistance and two large ones corresponding to the dark regions. The resulting solitary waves are called screening photorefractive solitons. For them to be stable (in time), additional illumination of intensity I_b at the same wavelength is required on the crystal in order to increase the conductivity of the (dark) regions surrounding the soliton. The charge accumulation in the beam-tails would otherwise flatten the induced index distribution and destroy the soliton.

The simplest description of screening solitons is provided by the saturating relationship between the electric field E and the light intensity I:

$$E = -\frac{V}{L}\frac{1}{1 + I/I_b} \qquad [11]$$

to be used in eqn [9] to get the index perturbation via the electro-optic effect. L is the width across which the voltage is applied. Since the quadratic term in eqn [9] (weighed by the tensor g_{ijkl}) can be neglected in noncentrosymmetric crystals:

$$\Delta n \approx -\frac{1}{2} n_0^3 r_{\text{eff}} E \qquad [12]$$

where r_{eff} is the relevant component of the electro-optics tensor; typically $n_0 \approx 2$, $r_{\text{eff}} \approx$ tens or hundreds of pm/V, and the overall index perturbation is of the order of $10^{-4}-10^{-3}$, with $V/L \approx 10^3$ V/cm. This description applies to 1D solitons; for 2D solitons, sometimes referred to as 'needle' solitons, the whole set of constitutive equations must be solved numerically. Nevertheless, in most cases the saturating model, even if not rigorous, describes well solitons and related phenomena both theoretically and experimentally, especially when nonlocal effects can be ignored.

Figure 3 shows the propagation of a 2D spatial screening soliton (top) compared to the linear case (bottom). The diffractive regime is realized by artificially 'turning off' the nonlinearity acting on the applied voltage (when $V = 0$ in eqn [11] then $\Delta n = 0$, corresponding to no induced perturbation).

Another class of photorefractive spatial solitary waves is that of photovoltaic solitons. In this case the electric field in the crystal has an additional component through the photovoltaic effect, the latter being relevant in materials such as LiNbO_3. To sustain bright solitons the index perturbation due to the photovoltaic current, which induces the space-charge field, must be positive. This is the case for Cu:KNSBN, where 2D photovoltaic solitons were first observed.

The space-charge field, due to the photovoltaic effect, can be derived from the Kukhtarev model:

$$E = -E_{\text{ph}}\frac{I/I_b}{1 + I/I_b} \qquad [13]$$

with E_{ph} is a constant depending on the material ($E_{\text{ph}} \approx 10^3$ V/cm) and I_b is the dark-irradiance controlled by the background laser light. Typical

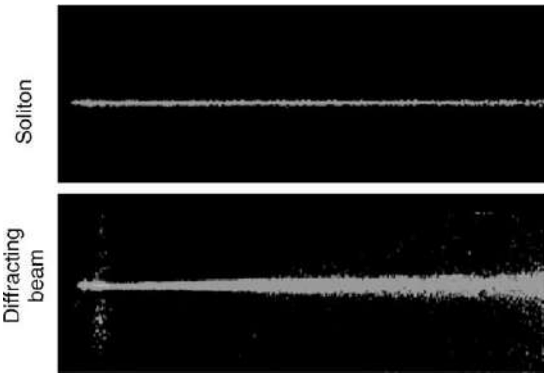

Figure 3 2D photorefractive soliton. Top view photograph of a 10 mm wide spatial soliton propagating in strontium barium niobate (top), and for comparison, the same beam diffracting when the nonlinearity if 'turned off' (bottom). (Adapted with permission from Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1419–1426).

geometries account for the tensorial nature of the effect, and are arranged to make I_b of the same magnitude of the beam intensity I. If $E_{ph} > 0$, as in Cu:KNSBN, eqn [13] in eqn [12] shows that a positive index change of the saturating type can be induced and support bright solitons. This model is successfully used for both 1D and 2D bright photovoltaic spatial solitons.

Photorefractive Solitons in Centrosymmetric Media

Most experiments on screening solitons were performed in ferroelectric noncentrosymmetric crystals. When brought above the Curie temperature T_c, such materials exhibit a phase-transition to a centrosymmetric paraelectric state. Crystals such as Potassium Lithium Tantalate Niobate have room-temperature phase-transitions, conveniently enabling experimental investigations. In paraelectrics, the second-order electro-optic tensor r_{ijk} is zero because of symmetry, but the photorefractive effect is still available through the third-order coefficient g_{ijkl} in eqn [9]. By properly choosing the orientation, eqn [12] can then be replaced by

$$\Delta n = \frac{1}{2} n_0^3 \varepsilon_0^2 (\varepsilon_r - 1)^2 g_{\text{eff}} E^2 \qquad [14]$$

Operation near T_c greatly enhances the relevant electro-optic coefficient $g_{\text{eff}} > 0$ ($g_{\text{eff}} \approx 0.1\ \text{m}^4\ \text{C}^{-2}$), making the observation of photorefractive screening solitons possible. Using eqn [11] in eqn [14] the model is saturating and works for both 1D and 2D bright solitons.

Thanks to the low powers involved and a mature crystal technology, the photorefractive effect and related soliton phenomena have contributed several impressive results in optical morphology of nonlinear light propagation, including incoherent and white light excitations.

Spatial Solitons in Nonlocal Media, Liquid Crystalline Media

A link, such as the Kerr law (eqn [4]), between the index perturbation and the optical intensity is 'local', because Δn at a given point (x, y, z) only depends on the intensity at the same location. Otherwise the medium is said to be 'nonlocal', e.g., one of those described by the general Kukhtarev model (eqn [10]) although, sometimes well approximated by eqns [11] or [13].

There are various mechanisms yielding nonlocality, mainly through spatial diffusion of the index perturbation far from the excitation. Remarkable nonlocal effects are expected when the spatial features of the beam, e.g., its waist, are significantly smaller than those of the induced polarization.

One of the most important effects of nonlocality is the stabilization against the catastrophic self-focusing in 2D, as experimentally investigated by Suter and Blasberg in 1992. Evidence of a highly nonlocal behavior has been very recently reported in liquid crystals, where the nonlinearity can be thermal or re-orientational. In the former case the index perturbation is much wider than the beam waist due to thermal diffusion associated to (absorption mediated) heat transfer. In the latter case the elastic properties of the medium produce a widening of the self-waveguide far from the beam axis. In nematic liquid crystal (NLC), the re-orientational response stems from the optically induced tilt of the elongated molecules towards the linearly polarized field versor. The light forced re-orientation increases the refractive index, thus creating the self-guiding conditions for spatial solitons. However, since in the nematic phase the molecules are linked together in a wide-scale order, the region (kernel) affected by this process is determined by the elastic properties of the medium and can extend to tens of microns.

An effective approach to describe spatial solitons in a highly nonlocal medium (where the kernel is much larger than the beam) postulates that the punctual relationship between Δn and I is replaced by a link between Δn and the beam power P. In such a way the transverse integral of the intensity is fully embedded in the index perturbation.

Assuming $n = n(P, r)$ with the beam mainly subject to the refractive perturbation near its axis, one can write:

$$n^2(r, P) = n_0^2(P) - r^2 \gamma(P) \qquad [15]$$

with $n_0^2(P) \cong n_0^2$ the on-axis value. $\gamma(P)$ depends on the specific geometry and determines the wavefront convexity at the origin. The overall profile is basically determined by the constitutive relationships of the medium. The propagation equation can be shown as

$$2ik \frac{\partial A}{\partial z} + \nabla_\perp^2 A = \frac{k^2}{n_0^2} \gamma^2(P) r^2 A \qquad [16]$$

Equations like [16] also describe the transverse profile of a beam linearly propagating in optical fibers, as well as the quantum harmonic oscillator (interpreting z as time). The soliton is its nondiffracting solution, a particular stationary case of a more general Gaussian beam with waist variable in propagation.

For intensities:

$$I(z) = \frac{P}{\pi w^2(z)} \mathrm{e}^{-r^2/w^2(z)} \qquad [17]$$

the spot-size obeys:

$$\frac{w^2(z)}{w_0^2} = 1 + \left[\frac{1}{k_0^2 w_0^4 \gamma^2(P)} - 1\right] \sin^2\left[\frac{n_0 z}{\gamma(P)}\right] \qquad [18]$$

Therefore, for a Gaussian excitation, the waist oscillates along z between the values w_0 – in general determined by the launch conditions – and $w_m = w_0 \gamma(P_S)/\gamma(P)$. P_s is the 'soliton power' for which $w_m = w_0$ and $k^{02} w_0^4 \gamma^2(P_S) = 1$, hence for $P < P_s$ or $P > P_s$ the beam pulsates cyclically with z (breathes). Such a dynamic balance between diffraction and nonlinearity stems from the spectral broadening associated to an increased beam shrinking: when the spectrum widens sufficiently, diffraction starts to prevail and the beam spreads again. This process repeats itself in propagation.

The stationary solution for $P = P_s$ corresponds to a point in the existence curve of waist versus power, i.e.:

$$k_0 w_0^2 \gamma(P_S) = 1 \qquad [19]$$

Such exact balance is critical, and experimentally observed solitary waves in highly nonlocal media appear as breathers.

Figure 4 displays the intensity profile of a soliton propagating in nematic liquid crystals, as acquired by a digital camera placed on the top of the cell and retrieving the scattered light (see Figure 5). The waist oscillation in the self-trapped beam and the dependence of the period on excitation are apparent.

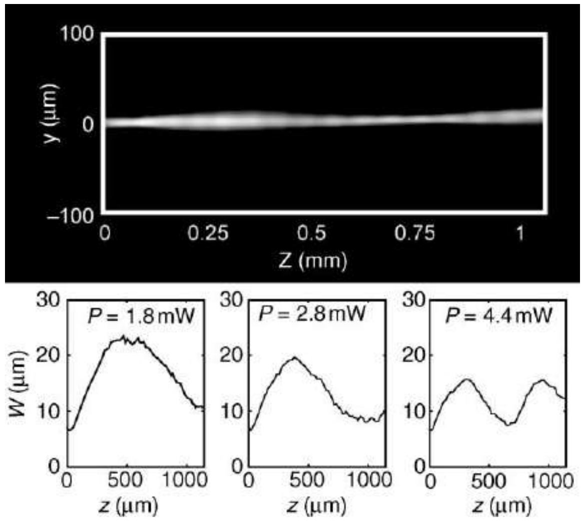

Figure 4 Nonlocal breather. (Top) top view photograph of a spatial soliton in NLC. Beam waist oscillations are apparent in propagation, as graphed (bottom panels) versus z for various excitations.

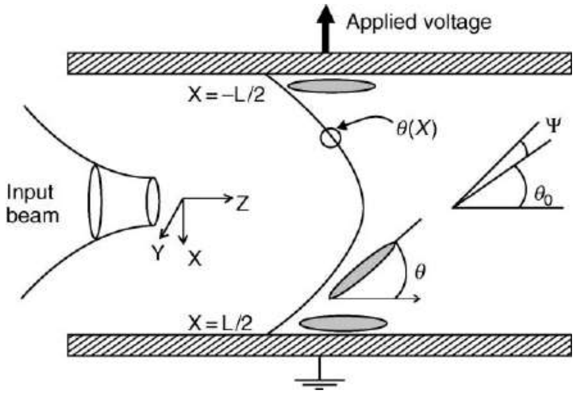

Figure 5 Sketch of a liquid crystal cell for the observation of nonlocal solitons. $\theta(x)$ is the orientation of the main axis of the ellipsoidal molecules, fixed by an external bias. The alignment is planar, with molecules parallel to the cell facets. The optical beam propagates in the center of the cell, where $\theta(x) = \theta_0$, and induces an additional re-orientation ψ. (Adapted from Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901).

A specific highly nonlocal nonlinear system is formed by a linearly polarized beam in NLC with molecules pre-tilted (by an external electric or magnetic field) at $\theta_0 = \pi/4$ with respect to the input polarization. As sketched in Figure 5, the light wave induces an additional tilt Ψ. At the lowest order of approximation, the relevant set of equations can be shown as:

$$\begin{aligned} 2ik\frac{\partial A}{\partial z} + \nabla_\perp^2 A &= -k^2 \frac{n_a^2}{n_0^2} \Psi A \\ K\nabla_\perp^2 \Psi - G\Psi + \frac{n_a^2}{2cn_0}|A|^2 &= 0 \end{aligned} \qquad [20]$$

with K the NLC elastic constant ($K \approx 10^{-11}$ N, for NLC of the E7 type), $n_0 \approx 1.5$, n_a^2 the optical anisotropy ($n_a \approx 1$), i.e., the difference between relative permittivities along the two principal axes, G a constant determined by the specific geometry of the NLC cell and the pre-tilt angle θ_0. Since, for $K = 0$, eqn [20] represents a Kerr medium, it can be considered one of the simplest nonlocal generalizations of local models for spatial solitons. The length $R_c = \sqrt{K/G}$ quantifies the diffusion of the index perturbation away from the beam axis, such that Ψ decays according to the modified Bessel function $K_0(r/R_c)$. Typically, $R_c \approx 20$ µm and solitons are observed in the visible and near-infrared.

Solitary wave profiles are found with the ansatz $A = a \exp(i\beta z)$, obtaining:

$$\begin{aligned} &\frac{\nabla_\perp^2 a}{2k^2} + \frac{n_a^2}{2n_0^2}\Psi a = \frac{\beta}{k} a \\ &R_c^2 \nabla_\perp^2 \Psi + \frac{R_c^2 n_a^2}{2Kcn_0} a^2 = \Psi \end{aligned} \qquad [21]$$

No analytical solutions of eqns [21] are known in 2D, hence the profiles have to be derived numerically. Conversely, it can be shown that 1D sech^2 solutions exist (see next section on parametric solitons). From eqn [21], the whole family can be parameterized by a dimensionless $\alpha = 1/(2k\beta R_c^2)$. Since $(1/2k\beta)^{1/2}$ measures the soliton waist (see eqn [7]), α relates the beam size and the extension of the induced perturbation. Small α solutions access the highly nonlocal regime, as shown in Figure 6, where the ratio R between the index perturbation and the soliton waist is graphed. The insets show sample profiles of a and ψ in two limits, respectively. On the other hand, when $K = 0$, $R_c = 0$, and $\alpha \to \infty$, the local limit yields an index perturbation comparable in size to the beam waist.

It is worth pointing out that a model such as eqn [16], also derived in the framework of plasma physics, predicts unconditionally stable self-trapped beams. Nonlocal solitons realize an energy minimum, thus preventing beam collapse such as in saturable (Kerr) media. The oscillations observed in experiments can be regarded as fluctuations around the lowest energy state.

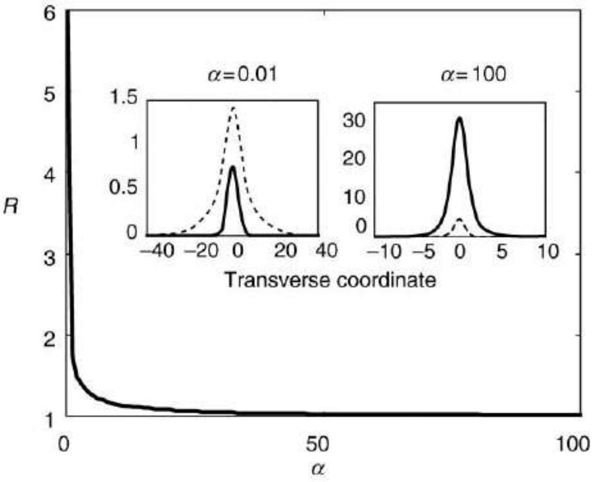

Figure 6 Spatial solitons versus nonlocality. Graph of the ratio *R* between the waist of the induced index perturbation in a nonlocal medium and the soliton waist versus parameter α. The insets show (in normalized units) soliton (solid line) and perturbation profiles (dashed line) for two different α, i.e., in the local ($\alpha = 100$) and highly nonlocal ($\alpha = 0.01$) regimes. (Adapted from Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901).

Parametric Spatial Solitons in Quadratic Media

Solitons due to parametric processes, such as second-harmonic generation, belong to a different class with respect to previous sections. In fact, no refractive index modification can be physically associated to parametric effects. Parametric solitons are characterized by a power spectral density peaked around at least two carrier harmonic frequencies. It is the interaction between these components which provides the diffraction beating mechanism. Their existence was predicted in the early 1970s by Sukhorukov and Karamzin, while the first experimental evidence was reported in 1995.

The simplest process to address is second-harmonic generation (SHG): a fundamental frequency (FF) wave at ω interacts with a component at 2ω. The latter can be spontaneously generated in noncentrosymmetric crystals (e.g., $LiNbO_3$, $KNbO_3$, BBO, KTP, etc.) or externally seeded.

For plane waves, this process is ruled by:

$$\begin{aligned} &i\frac{dA_1}{dz} + \frac{\omega d_{\text{eff}}}{c} A_1^* A_2 \, e^{-i\Delta k z} = 0 \\ &i\frac{dA_2}{dz} + \frac{\omega d_{\text{eff}}}{c} A_1^2 \, e^{i\Delta k z} = 0 \end{aligned} \qquad [22]$$

with $\Delta k = k_1 - 2k_2 = 2\omega/c(n_1 - n_2)$ the phase mismatch, d_{eff} the nonlinear coefficient, n_2 and n_1 the indices at SH and FF, respectively, and

$$E_{\text{opt}} = \text{Re}\left[\sqrt{\frac{2Z_0}{n_1}} A_1 e^{-i\omega t + ik_1 z} + \sqrt{\frac{2Z_0}{n_2}} A_2 e^{-2i\omega t + ik_2 z}\right]$$

the overall electric field. d_{eff} relates to the quadratic susceptibility $d_{\text{eff}}^{(2)}$ (measured in mV^{-1}) through to $d_{\text{eff}} = d_{\text{eff}}^{(2)}\sqrt{2Z_0}/(n_1\sqrt{n_2})$, and typically $d_{\text{eff}} \approx 10^{-10}\,\text{m}^2\,\text{W}^{-1/2}$.

Equations [22] show that, in general, the harmonics with amplitudes A_1 and A_2 affect each other's phase and intensity during propagation. The solution with constant moduli, i.e.:

$$\begin{aligned} A_1 &= \sqrt{I_1} \exp\left[i\frac{\omega}{c}\Delta n_1 z\right] \\ A_2 &= \sqrt{I_2} \exp\left[i\frac{2\omega}{c}\Delta n_2 z\right] \end{aligned} \qquad [23]$$

with I_1 and I_2 the respective intensities, provides a 'pure' phase variation and 'apparent' index changes Δn_1 and Δn_2 on either wave. The latter satisfy:

$$\Delta n_1 = d_{\text{eff}}^2 \frac{I_2 - I_1/2}{n_1 - n_2}$$
$$\Delta n_2 = d_{\text{eff}}^2 \frac{I_1}{2I_2} \frac{I_2 - I_1/2}{n_1 - n_2} \qquad [24]$$

with the constraint (for plane waves to exist):

$$I_1 = 2I_2 - 2\frac{n_1 - n_2}{d_{\text{eff}}}\sqrt{I_2} \qquad [25]$$

Equations [24] encompass a great deal of the physics in self-trapping via SHG. When eqn [25] is satisfied (at least approximately), the effect of SHG is exclusively on the phase. Assuming eqn [24] is valid for (wide) beams, the propagation in the presence of diffraction can be described by two coupled equations of the type [1] with relevant perturbations (eqn [24]). This shows that the sign of the term $(2I_2 - I_1)/(n_1 - n_2)$ is crucial, because focusing at both harmonics ($\Delta n_{1,2} > 0$) requires $(2I_2 - I_1)/(n_1 - n_2) > 0$.

When just a fundamental frequency (ω) is launched, solitons are generated for $\Delta k > 0$: a small 2ω component is provided via SHG and $I_1 > 2I_2$. In fact, since SHG ($\omega + \omega = 2\omega$) is followed by difference frequency generation ($2\omega - \omega = \omega$, DFG), for any finite mismatch $\Delta k > 0$, a portion of the FF travels for a while at 2ω before it is back-converted: this is commonly named 'cascading' and, as a consequence, part of the FF undergoes a positive phase shift. The converse happens in the DFG + SHG process. Finally, if an SH beam with a small FF seed is input, solitons form for $\Delta k < 0$ being $2I_2 > I_1$. From a physical point of view, these mechanisms create the phase-front distortion required to counteract diffraction. Figure 7 is an artist's sketch on the formation of an SHG soliton.

Equations [24] are approximately valid and rely on a plane-wave insight, but they point out the wealth of possibilities offered by three-wave interactions for light transverse localization via parametric processes. The full set of coupled equations describing parametric solitons from SHG differs from eqn [1]:

$$\mathrm{i}\frac{\mathrm{d}A_1}{\mathrm{d}z} + \frac{1}{2k_0 n_1}\nabla_\perp^2 A_1 + \frac{\omega d_{\text{eff}}}{c} A_1^* A_2 \mathrm{e}^{-\mathrm{i}\Delta k z} = 0$$
$$\mathrm{i}\frac{\mathrm{d}A_2}{\mathrm{d}z} + \frac{1}{4k_0 n_2}\nabla_\perp^2 A_2 + \frac{\omega d_{\text{eff}}}{c} A_1^2 \mathrm{e}^{\mathrm{i}\Delta k z} = 0 \qquad [26]$$

as it includes diffraction terms added to eqn [22].

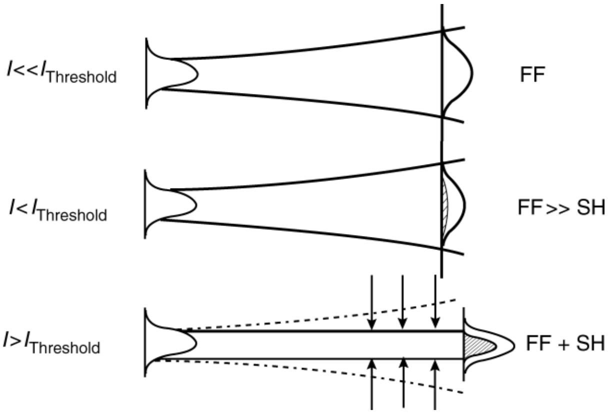
$I \ll I_{\text{Threshold}}$
FF
$I < I_{\text{Threshold}}$
FF>> SH
$I > I_{\text{Threshold}}$
FF + SH

Figure 7 Parametric solitons via SHG. Intuitive description of soliton formation with an FF Gaussian input. At low powers (top) and for a small amount of generated SH (center) linear diffraction dominates. Above threshold parametric gain efficiently counteracts diffraction, originating a spatial soliton via quadratic cascading.

The self-trapped solutions are found as in eqn [23], with

$$A_1 = a_1(x, y)\exp(\mathrm{i}\beta z)$$
$$A_2 = a_2(x, y)\exp(\mathrm{i}\Delta k z)\exp(2\mathrm{i}\beta z), \quad \text{and} \qquad [27]$$
$$a_{1,2}(x, y) = \sqrt{I_{1,2}(x, y)}$$

These show that, even in the presence of diffraction, SHG can give rise to pure phase effect if the profiles satisfy:

$$\frac{1}{2k_0 n_1}\nabla_\perp^2 a_1 + \frac{\omega d_{\text{eff}}}{c} a_1 a_2 = \beta a_1$$
$$\frac{1}{8k_0 n_2}\nabla_\perp^2 a_2 + \frac{\omega d_{\text{eff}}}{2c} a_1^2 = \beta + \frac{\Delta k}{2} a_2 \qquad [28]$$

which are the space-dependent counterpart of eqn [25]. Again, β is a nonlinear contribution to the FF wavevector, and determines the corresponding correction $\beta_2 = 2\beta + \Delta k$ at SH. β is univocally determined by the power of the two harmonics in the solutions of eqn [28].

In the particular case:

$$\Delta k = \beta\frac{n_1 - 4n_2}{2n_2} \qquad [29]$$

1D parametric solitons exhibit profiles

$$a_1 = \sqrt{\frac{n_1}{n_2}} \frac{3c}{4\omega d_{\text{eff}}} \beta \operatorname{sech}\left(\sqrt{\frac{\beta k_0 n_1}{2}} x\right)^2$$
$$a_2 = \frac{3c}{4\omega d_{\text{eff}}} \beta \operatorname{sech}\left(\sqrt{\frac{\beta k_0 n_1}{2}} x\right)^2 \qquad [30]$$

and the powers per unit of transverse wavefront along y at the two components are:

$$\begin{aligned} P_1 &= \int_{-\infty}^{\infty} a_1^2(x)\mathrm{d}x = \frac{n_1}{n_2}\frac{3\sqrt{2}c^2}{4\omega^2 d_{\mathrm{eff}}^2\sqrt{k_0 n_1}}\beta^{3/2} \\ P_2 &= \int_{-\infty}^{\infty} a_2^2(x)\mathrm{d}x = \frac{3\sqrt{2}c^2}{4\omega^2 d_{\mathrm{eff}}^2\sqrt{k_0 n_1}}\beta^{3/2} \end{aligned} \qquad [31]$$

with the total soliton power $P_1 + P_2$, directly related to β.

Equation [29] implies that, for $n_1 > 4n_2$ and therefore $\Delta k > 0$, 1D parametric solitons exhibit the sech^2 profile. Otherwise, no analytical solutions are known even in 1D, and numerical techniques need to be employed. The same applies to 2D soliton profiles in all instances.

Notably, eqns [28] are essentially identical to eqns [21] for solitons in NLC. Therefore, Figure 4 provides the features of 2D parametric solitons as well, with 'a' replaced by 'a_1', and 'Ψ' by 'a_2'. For parametric solitons $\alpha = 4n_2(1 + \Delta k/2\beta)/n_1$, showing that at phase-matching ($n_2 = n_1$, $\Delta k = 0$) $\alpha = 4$, while large α's correspond to large Δk or small β's, e.g., low power and very wide solitons. The sech^2 profile for 1D solitons corresponds to $\alpha = 1$. 1D and 2D soliton solutions exist for any $\alpha > 0$.

The model for SHG solitons is more complicated than the previously discussed Kerr or saturable cases. The stability question can be formulated as the constrained-minimum problem of a specific functional, the Hamiltonian H. In general, for a given soliton power (the sum of the powers at the two harmonics) two branches exist for H. They correspond to two solitary solutions, differing in the ratio between the two powers. The solution with $P_2 > P_1$ corresponds to higher H, and is unstable via parametric amplification. When a harmonic field propagates in the presence of energy at a subharmonic (e.g., a small amount of FF), it tends to break up and transfer energy to it. Looking at Figure 4, this corresponds to small α's, only attainable when $\Delta k < 0$. As a result, quadratic solitons with a relevant SH content are unstable and thus, more difficult to observe.

The generation of a parametric soliton with a Gaussian input at FF, has a threshold (see Figure 7). This is because the input has to generate sufficient SH in order to produce and 'feel' the nonlinear phase-shift. A numerical simulation of eqn [26] shows the formation of a parametric spatial soliton, as displayed in Figure 8. As shown in (a), if only a small second-harmonic is generated by an FF excitation, diffraction dominates. Above threshold, as in (b), a stable self-trapped two-color beam is generated.

Two significant experimental results are reproduced in Figures 9 and 10. The photograph in Figure 10 shows a soliton propagating in periodically poled KTP (PPKTP) over five Rayleigh ranges, obtained by imaging the out-of-plane scattered light. Parametric solitons are typically observed with the FF in the near-infrared and the SH in the visible.

Figure 10 shows the output spots at FF and SH of a beam propagating in PPKTP for various FF excitations, until a soliton is generated (right end).

We close this section by observing that more general models must be considered to describe the whole physics of solitons in media with a quadratic

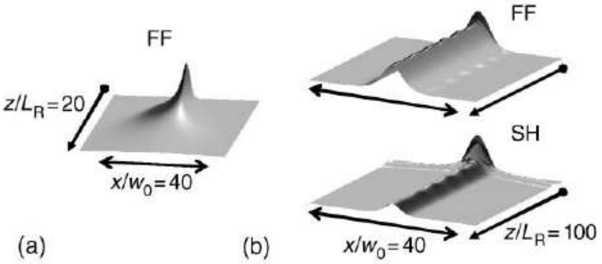

Figure 8 Parametric solitons via SHG. 1D numerical simulations using the beam propagation method. (a) An FF Gaussian excitation diffracts at low powers; (b) at high FF powers, after a transient both harmonics propagate undiffracted as a parametric spatial soliton. (Adapted with permission from Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396).

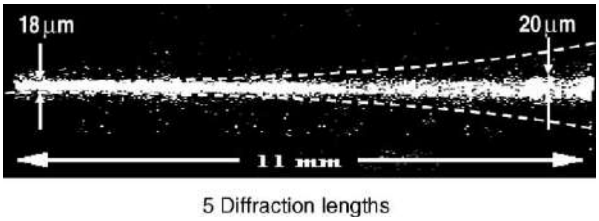

Figure 9 2D parametric solitons via SHG. Top view photograph of a soliton in PPKTP (20 ps pulses at $\lambda \approx 983$ nm and pulse energy of about 0.3 μJ). The dashed lines illustrate the diffractive behavior in the absence of nonlinear effects. (Adapted with permission from Malendevich R, Jankovic L, Polyakov S, Fuerst R, Stegeman G, Bosshard C and Gunter P (2002) Two-dimensional type I quadratic spatial solitons in KNbO3 near non-critical phase-matching. *Optics Letters* 27: 631–673.)

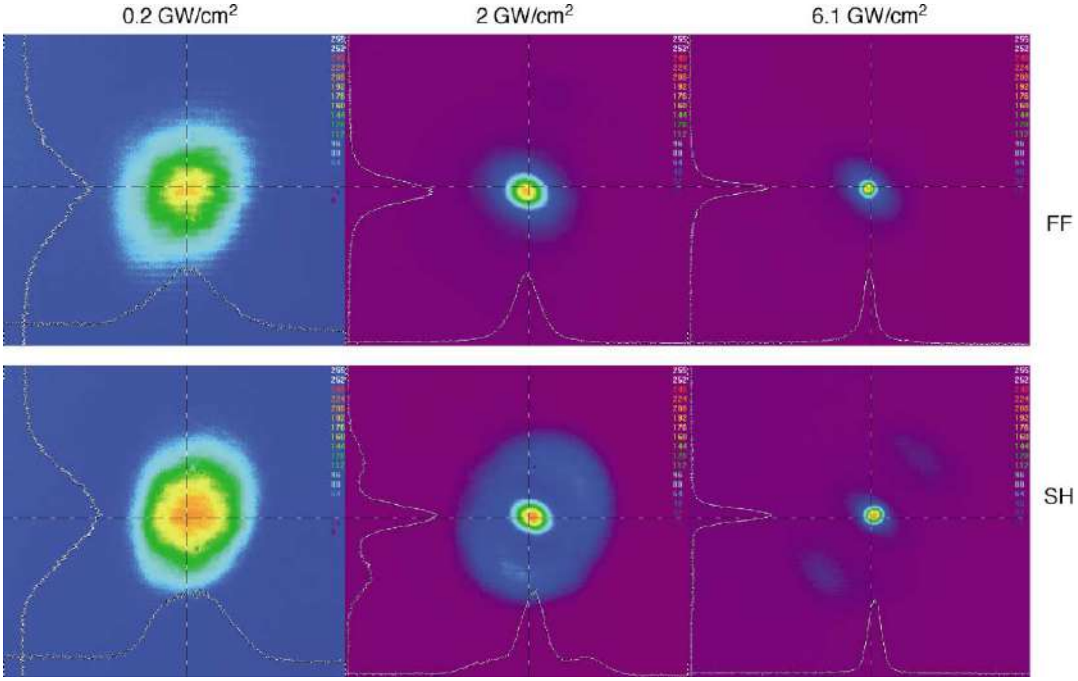

Figure 10 2D parametric solitons via SHG. Beam spots at FF (top row) and SH (bottom) for various input intensities in a PPKTP crystal. (Adapted with permission from Stegeman GI, Jankovic L, Kim H, *et al.* (2003) Generation of, and interactions between quadratic spatial solitons in non-critically-phase-matched crystals. *Journal of Nonlinear Optical Physics and Materials*. 12: 1–20).

response. We mention, for example, the subtleties associated to anisotropy (walk-off), nondegenerate three wave interactions, additional Kerr and higher-order effects.

Conclusions

Bright spatial solitons are widespread, and have been reported in several materials using various wavelengths and geometries. The brief overview presented above was intended to introduce the variety of basic mechanisms underlying their generation and propagation, but could not even touch upon the wealth of additional properties and potential applications of individual and multiple spatial solitons, their interactions and collisional properties, their stability and their control with other linear and nonlinear light waves, as well as with additional nonoptical effects (temperature, voltage, acoustic, and magnetic fields, etc.). Fascinating phenomena involve bright spatial solitons and are the subject of current research. Among them we like to mention all-optical logic and signal readdressing, beam shaping and filtering, limiting, sub-wavelength spatial compression, and digital imaging. To learn more on the subject, we encourage the interested readers to look up the vast literature covering both fundamental and applied aspects of optical spatial solitons.

List of Units and Nomenclature

Angular frequency [s^{-1}]	ω
Beam waist [m]	w
Boltzmann constant [J K^{-1}]	k_{B}
Current density vector [A m^{-2}]	$\mathbf{J}$
Density of ionized donors [m^{-3}]	N_{d}^{i}
Dielectric constant of vacuum [F m^{-1}]	ε_0
Director angle perturbation [rad]	ψ
Director orientation angle [rad]	θ_0
Electric field amplitude, normalized [W$^{1/2}$ m^{-1}]	a, A
Electric field amplitude, optical frequency [Vm^{-1}]	E_{opt}
Electric field vector, low frequency [V m^{-1}]	$\mathbf{E}$
Electron charge [C]	q
Electron density [m^{-3}]	ρ
Electron mobility [m^2 V^{-1} s^{-1}]	μ
Electro-optic second-order tensor [m V^{-1}]	r, r_{ijk}

Electro-optic third-order tensor [m^4 C^{-2}]	$g_{\mathrm{eff}}, g_{ijkl}$
Element of the photovoltaic tensor [C m^3 J^{-1}]	β_{ph}
Kerr nonlinear coefficient [m^{2}W^{-1}]	n_2
Low frequency permittivity [F m^{-1}]	ε_{S}
Medium elastic constant [N]	K
Medium forcing constant [N m^{-2}]	G
Non local decay distance [m]	R_{c}
Nonlinear length [m]	L_N
Nonlinear wavevector correction [m^{-1}]	β, β_2
Optic axis versor	$\mathbf{c}_{\mathrm{opt}}$
Optical background intensity [Wm^{-2}]	I_{b}
Optical dark intensity [Wm^{-2}]	I_{dark}
Optical index anisotropy	n_{a}
Optical intensity [Wm^{-2}]	I
Optical saturation intensity [Wm^{-2}]	I_{S}
Photo-ionization cross-section [m^2 J^{-1}]	s
Photovoltaic index perturbation constant [V m^{-1}]	E_{ph}
Power [W]	P
Power per wavefront unit [W m^{-1}]	P_1, P_2
Propagation coordinate [m]	z
Rayleigh length [m]	L_{R}
Recombination rate [m^3 s^{-1}]	γ_{r}
Refractive index	n
Refractive index perturbation	Δn
Relative dielectric constant	ε_{r}
Sample width [m]	L
Second order effective susceptibility [m V^{-1}]	$d_{\mathrm{eff}}^{(2)}$
Second order nonlinear coefficient [m W$^{-1/2}$]	d_{eff}
Self-focusing power [W]	P_{c}
Soliton power [W]	P_{S}
Speed of light in vacuum [ms^{-1}]	c
Static dielectric constant [F m^{-1}]	ε_{s}
Temperature [K]	T
Total density of negatively charged sites [m^{-3}]	N_{A}
Total donor density [m^{-3}]	N_{d}
Transverse coordinates [m]	x, y, r
Transverse wave-vector [m^{-1}]	k_x
Vacuum impedance [Ω]	Z_0
Wavelength [m]	$\lambda = 2\pi n/k = 2\pi/k_0$
Wavevector in medium [m^{-1}]	$k = n\omega/c$
Wavevector in vacuum [m^{-1}]	$k_0 = \omega/c$

See also

Babinet's Principle. **Fiber and Guided Wave Optics:** Nonlinear Optics. **Nonlinear Optics, Applications:** Self-focusing and Related Effects (Solitons and Multiphoton Absorption). **Nonlinear Optics, Basics:** Cascading; Photorefraction.

Further Reading

Assanto G and Peccianti M (2003) Spatial solitons in nematic liquid crystals. *IEEE Journal of Quantum Electronics* 39: 13–21.

Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396.

Assanto G, Peccianti M and Conti C (2003) Nematicons: optical spatial solitons in nematic liquid crystals. *Optics and Photonics News* 14: 44–48.

Buriak AV, Di Trapani P, Skryabin DV and Trillo S (2002) Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications. *Physics Reports* 370: 63–235.

Chiao RY, Garmire E and Townes CH (1964) Self-trapping of optical beams. *Physics Review Letters* 13: 479–482.

Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901.

Conti C, Peccianti M and Assanto G (2004) Observation of optical spatial solitons in a highly nonlocal medium. *Physics Review Letters* 92: 113902.

Crosignani B and Salamo G (2002) Photorefractive solitons. *Optics and Photonics News* 13: 38–41.

Kivshar Y and Agrawal G (2003) *Optical Solitons: From Fibers to Photonic Crystals*. New York: Academic Press.

Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1419–1426.

Torner L and Sukhorukov AP (2002) Quadratic solitons. *Optics and Photonics News* 13: 42–47.

Various Authors (1998) In: Kajzar F and Reinisch R (eds) *Beam Shaping and Control with Nonlinear Optics*. New York: Plenum Press.

Various Authors (2001) In: Boardman AD and Sukhorukov AP (eds). *Soliton-driven Photonics*. London: Kluwer Academic.

Various Authors (2002) In: Trillo S and Torruellas WE (eds) *Spatial Solitons*. Berlin: Springer-Verlag.

Optical Fiber Solitons, Physical Origin and Properties

G Millot and P Tchofo-Dinda, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the last few decades there has been considerable effort made in understanding energy-localization phenomena in many areas of physics. Optical fibers constitute excellent media where such phenomena are routinely observed by means of relatively simple experimental setups. A well-known example of those energy-localization phenomena is the optical soliton, which corresponds to a light pulse that can propagate in optical fibers over relatively long distance without experiencing significant degradations. From a fundamental point of view, the high stability of the optical soliton results from a delicate balance between the fiber chromatic dispersion and the nonlinear refractive (or Kerr) index. The idea of using optical solitons as information bits in high-speed communication systems was first proposed in 1973, and then demonstrated experimentally in 1980. After these discoveries, interest in optical soliton transmission started to increase. In an ideal fiber, optical solitons can be modeled approximately by the nonlinear Schrödinger equation, whose solutions are well known. In this article we present the main types of soliton behavior and the fundamental physical processes that are involved in soliton stability.

The Nonlinear Schrödinger Equation

In optical fibers, there are many nonlinear phenomena that are difficult to understand by direct analysis of the nonlinear polarization (e.g., the soliton effect) but that can be readily understood by use of the nonlinear Schrödinger equation (NLSE). Let us consider an optical pulse that propagates in the z-direction, in a dielectric Kerr medium such as an optical fiber, with frequency ω_0, wavenumber k_0, and an envelope $\mathcal{A}$:

$$E = \mathcal{A}\exp[\mathrm{i}(\mathbf{k}_0\cdot\mathbf{r} - \omega_0 t)] \qquad [1]$$

We assume that the wave envelope $\mathcal{A}$ varies slowly in time and space compared with the fast variations of the carrier wave ω_0. In this situation, the Fourier spectrum of the wave is restricted to frequencies and wavenumbers close to $\omega = \omega_0$ and $k = k_0$, respectively. Then, the wavenumber k can be expanded in a Taylor series around k_0, as follows:

$$k(\omega) = k_0 + \left(\frac{\partial k}{\partial \omega}\right)_{\omega_0}(\omega-\omega_0) + \frac{1}{2}\left(\frac{\partial^2 k}{\partial \omega^2}\right)_{\omega_0}(\omega-\omega_0)^2 + \left(\frac{\partial k}{\partial |\mathcal{A}|^2}\right)_{\mathcal{A}_0=0}|\mathcal{A}|^2 + \cdots \qquad [2]$$

The nonlinear term $Q = (\partial k/\partial|\mathcal{A}|^2)_{\mathcal{A}_0=0}$ arises from the optical Kerr effect, as $k = n\omega/c = k(\omega, |\mathcal{A}|^2)$.

Setting $K = k - k_0$, $\Omega = \omega - \omega_0$, $\beta_1 = (\partial k/\partial\omega)_{\omega_0}$ and $\beta_2 = (\partial^2 k/\partial\omega^2)_{\omega_0}$, one can rewrite eqn [2] as

$$K = \beta_1\Omega + \frac{1}{2}\beta_2\Omega^2 + Q|\mathcal{A}|^2 \qquad [3]$$

where β_1 is the inverse of the group velocity and β_2 is the group velocity dispersion (GVD) coefficient. Equation [3] represents the nonlinear dispersion relation for the envelope. Using a Fourier transform method in which K is replaced by a spatial operator $-\mathrm{i}\,\partial/\partial z$, and Ω by a temporal operator $\mathrm{i}\,\partial/\partial T$, one can apply eqn [3] on the envelope $\mathcal{A}$ to obtain the following equation:

$$\mathrm{i}\frac{\partial \mathcal{A}}{\partial z} + \mathrm{i}\beta_1\frac{\partial \mathcal{A}}{\partial T} - \frac{1}{2}\beta_2\frac{\partial^2 \mathcal{A}}{\partial T^2} + Q|\mathcal{A}|^2\mathcal{A} = 0 \qquad [4]$$

In practice it is more convenient to make use of the quantity $A = (\alpha A_{\mathrm{eff}})^{1/2}\mathcal{A}$, for which $|A|^2$ represents the wave power. The parameter A_{eff} is the effective core area, which takes into account the variation of the light intensity across the transverse section of the fiber. The value of A_{eff} strongly depends on the mode distribution. Using a system of coordinates that moves at the group velocity v_g, i.e., replacing T by $t = T - z/v_g$, eqn [4] reduces to the basic NLSE:

$$\mathrm{i}\frac{\partial A}{\partial z} - \frac{1}{2}\beta_2\frac{\partial^2 A}{\partial t^2} + \gamma|A|^2 A = 0 \qquad [5]$$

where γ is the nonlinear coefficient defined by $\gamma = Q/(\alpha A_{\mathrm{eff}}) = \omega_0 n_2/(cA_{\mathrm{eff}})$ and expressed in $\mathrm{W}^{-1}\mathrm{m}^{-1}$. This equation is formally analogous to the Schrödinger equation in quantum mechanics but with a nonlinear potential. Furthermore, in this basic NLSE, all higher-order terms that describe effects such as the fiber losses, third-order dispersion, or stimulated Raman scattering, are neglected. The NLSE for pulse propagation in optical fibers was first derived in 1973 by Hasegawa and Tappert. It has also found applications in other areas of physics, such as in plasma physics, fluid dynamics and electrical transmission lines. From a fundamental point of view, the basic NLSE involves the combined effects of the group velocity dispersion (GVD) and self-phase modulation (SPM), which are described by the second and third terms in eqn [5], respectively. Each of these two effects is analyzed below.

Group Velocity Dispersion

The individual effect of the GVD on pulse propagation can be obtained by setting the nonlinear coefficient to zero ($\gamma = 0$). In this case, eqn [5] reduces to

$$\mathrm{i}\frac{\partial A}{\partial z} = \frac{1}{2}\beta_2\frac{\partial^2 A}{\partial t^2} \qquad [6]$$

In the frequency domain, eqn [6] becomes

$$\mathrm{i}\frac{\partial \tilde{A}}{\partial z} = -\frac{1}{2}\beta_2\omega^2\tilde{A} \qquad [7]$$

where $\tilde{A}$ represents the Fourier transform of A. The solution of eqn [7] is given by

$$\tilde{A}(z,\omega) = \tilde{A}(0,\omega)\exp(\mathrm{i}\beta_2\omega^2 z/2) \qquad [8]$$

As eqn [8] shows, the GVD changes the phase of each spectral component of the pulse by an amount that depends on the propagated distance and the frequency of the spectral component. But this effect does not modify the power spectrum: $(|\tilde{A}(z,\omega)|^2 = |\tilde{A}(0,\omega)|^2)$. This phase change can modify the shape of the pulse envelope. For example, a Gaussian pulse with a width of δ_0 at $z = 0$, $A(z=0,t) = \exp\{(-t^2/(2\delta_0^2)\}$, will transform into

$$A(z,t) = [\delta_0^2/(\delta_0^2 - \mathrm{i}\beta_2 z)]^{1/2}\exp[-t^2/2(\delta_0^2 - \mathrm{i}\beta_2 z)] \qquad [9]$$

at a distance z. In other words, the pulse width increases from δ_0 to $\delta_0[1+(z/L_\mathrm{D})^2]^{1/2}$. The parameter $L_\mathrm{D} = \delta_0^2/|\beta_2|$ defines a dispersion length, that is, a length scale over which the dispersive effects become important. At $z = L_\mathrm{D}$ the Gaussian pulse broadens by a factor of $\sqrt{2}$. Thus, the GVD causes the spectral components of the pulse to move at different velocities, which broadens the pulse, as shown by Figure 1.

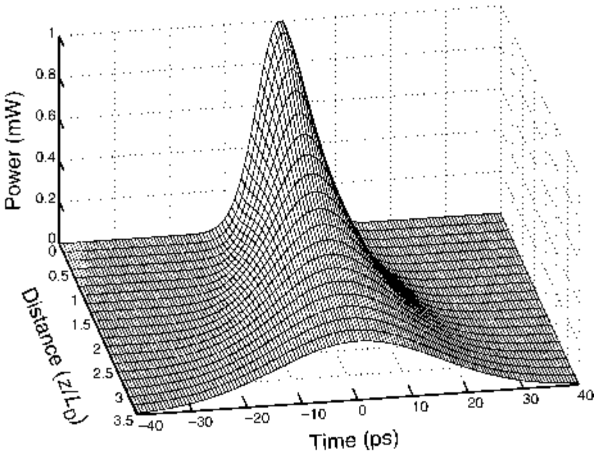

Figure 1 Temporal evolution of the power of a Gaussian pulse, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 1$ mW, $\beta_2 = -20\ \mathrm{ps}^2\ \mathrm{km}^{-1}$, $\gamma = 3\times10^{-3}\ \mathrm{m}^{-1}\ \mathrm{W}^{-1}$ and $\delta_0 = 10$ ps. The dispersion length is $L_\mathrm{D} = 1.8$ km.

Another important effect of the GVD comes from the time dependence of the pulse phase, which one can easily obtain by rewriting eqn [9] as $A(z,t) = |A(z,t)|\exp[\mathrm{i}\phi(z,t)]$. It turns out that the GVD induces an instantaneous variation of the frequency across the pulse, called a frequency chirp. For the initially unchirped Gaussian pulse, the GVD induces a chirp given by

$$\delta\omega = -\frac{\partial\phi}{\partial t} = \frac{t\beta_2(z/L_\mathrm{D})}{|\beta_2|\delta_0^2[1+(z/L_\mathrm{D})^2]} \qquad [10]$$

As eqn [10] illustrates, the GVD-induced frequency chirp depends on the sign of β_2, and varies linearly across the pulse.

Self-Phase Modulation

Similarly to dispersion effects, nonlinear effects in optical fibers require a minimum propagation distance to develop significantly. The distance over which nonlinear effects become important is defined as the nonlinear length L_NL, given by $L_\mathrm{NL} = 1/(\gamma P_0)$, where P_0 is the peak power.

Let us consider the propagation of a light pulse in a fiber of length L, such that $L_\mathrm{NL} < L \ll L_\mathrm{D}$. In this particular case, the effects of dispersion can be neglected, and then the NLS equation becomes

$$\mathrm{i}\frac{\partial A}{\partial z} + \gamma|A|^2 A = 0 \qquad [11]$$

It results from this equation that $|A(z,t)|^2 = |A(z=0,t)|^2$. This means that the pulse shape remains unchanged during propagation. The solution of eqn [11] is given by

$$A(z,t) = A(0,t)\exp[\mathrm{i}\Phi_\mathrm{NL}(z,t)] \qquad [12]$$

where $\Phi_{\mathrm{NL}}(z,t) = \gamma|A(0,t)|^2 z$. Thus, the Kerr effect causes the phase of the wave envelope to vary during the propagation. The nonlinear phase shift Φ_{NL} increases with the propagated distance z. The maximum phase shift $\Phi_{\mathrm{NL}}^{\max}$ occurs at the pulse center and is given by $\Phi_{\mathrm{NL}}^{\max} = \gamma|A(0,0)|^2 z = \gamma P_0 z$. One can see that the nonlinear length represents the propagation distance at which $\Phi_{\mathrm{NL}}^{\max} = 1$. Since the nonlinear phase shift Φ_{NL} depends on the initial pulse shape, it varies nonlinearly with time t. This implies that the instantaneous optical frequency differs across the pulse from its central value ω_0. The frequency chirp is given by

$$\delta\omega(t) = \omega(t) - \omega_0 = -\frac{\partial \Phi_{\mathrm{NL}}}{\partial t} = -\frac{\partial |A(0,t)|^2}{\partial t}\gamma z \qquad [13]$$

The modification of the phase of the wave envelope by itself is called self-phase modulation. The frequency chirp increases with the propagated distance, thus implying that new frequencies are continuously generated as the pulse propagates through the fiber. This spectral broadening depends on the initial pulse shape. Figure 2 shows the variation of Φ_{NL} and the corresponding induced frequency chirp across the pulse at $z = L_{\mathrm{NL}}$ for the case of an initial Gaussian pulse.

The SPM-induced chirp exhibits several interesting features. $\delta\omega$ is negative near the leading edge (red shift) and is positive near the trailing edge (blue shift). The chirp is linear (and increases) in the central region of the pulse. In fact, the chirp $\delta\omega$ at a given propagation distance z depends on the initial chirp. As Figure 3 shows, at a sufficiently long propagated distance, the SPM-induced spectrum exhibits several peaks due to interference effects across the pulse. Indeed as Figure 2 shows, the same frequency chirp occurs at two distinct times associated with two distinct nonlinear phase shifts.

The spectrum width can be estimated by calculating the maximum value $\delta\omega_{\max}$ of the frequency chirp. For an initial unchirped Gaussian pulse one obtains

$$\delta\omega_{\max} \approx \frac{0.86\gamma P_0 z}{\delta_0} \qquad [14]$$

Thus, the spectral width increases linearly with the propagated distance z and the input peak power P_0, and is inversely proportional to the initial pulse width δ_0. The SPM can be detrimental for optical communication systems. The SPM broadens the spectrum of the signal and makes it then more susceptible to GVD effects. Note that in wavelength division multiplexing (WDM) systems, the phase modulation of a given channel can come not only from the power in the considered channel itself (SPM) but also from the power in a different channel of the WDM system (CPM, i.e., cross-phase modulation). By appropriate dispersion compensation, the impact of SPM and CPM can be significantly reduced. On the other hand, the SPM, combined with dispersive effects, leads to very useful effects such as pulse compression, or the generation of optical solitons (described below).

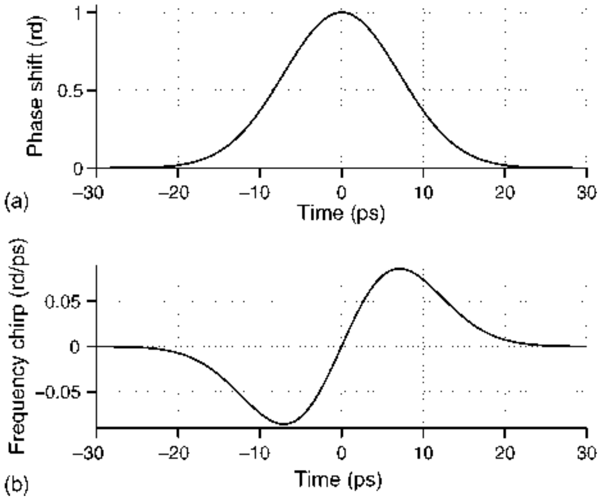

Figure 2 Temporal variation of the nonlinear phase shift (a) and the corresponding frequency chirp induced by SPM (b), for a Gaussian pulse.

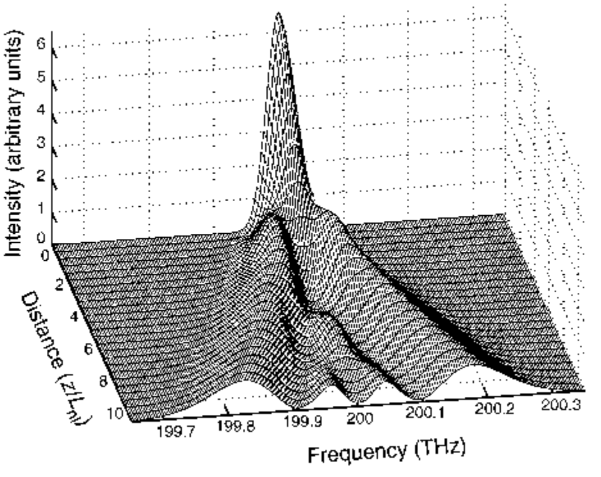

Figure 3 Evolution of the SPM-induced spectral broadening of an initially unchirped Gaussian pulse, as a function of the propagation distance. Here the normalized distance is z/L_{NL}.

Optical Solitons

Bright Solitons

Soliton propagation results fundamentally from a delicate balance between two phenomena: GVD and SPM. As we mentioned above, the GVD causes the spectral components of the pump to move at different velocities, thus leading to a temporal broadening of the pulse as shown in Figure 1 in the case of a Gaussian input pulse. Moreover, eqn [10] shows that the effect of GVD on a propagating Gaussian pulse is associated with a linear frequency chirp $\delta\omega$ and a parabolic temporal phase, proportional to β_2. On the other hand, eqn [12] shows that the SPM induces a frequency chirp across the pulse proportional to the temporal derivative of the pulse intensity. The chirp induced by SPM is then linear only in the central region of the pulse as long as the initial pulse shape is nonparabolic. As a consequence a Gaussian input pulse cannot propagate in a distortion-free manner since the GVD for anomalous dispersion ($\beta_2 < 0$) is able to completely cancel the intensity-dependent phase shift from SPM only over the central region of the pulse. On the other hand we show below that, in the anomalous dispersion regime, the exact cancellation of the nonlinear and dispersive temporal phase components (and corresponding chirps) is obtained in the case of the propagation of a hyperbolic secant pulse, which corresponds to the fundamental bright optical fiber soliton.

We consider the propagation of an unchirped hyperbolic secant pulse at a wavelength of 1.55 μm in a standard telecommunications fiber with $\beta_2 = -21 \times 10^{-3}\ \mathrm{ps}^2\ \mathrm{m}^{-1}$ and $\gamma = 1.2 \times 10^{-3}\ \mathrm{W}^{-1}\ \mathrm{m}^{-1}$. The input pulse at $z = 0$ is given by $A(0,t) = \sqrt{P_0}\ \mathrm{sech}\,(t/\delta_0)$ with $\delta_0 = 10$ ps and $P_0 = 175$ mW. With this choice of parameters the dispersion length is equal to the nonlinear length, i.e. $L_\mathrm{D} = L_\mathrm{NL} = 4.76$ km. In fact, these conditions correspond to those which would support a propagating soliton. However, let us consider first how SPM and GVD modify the incident pulse when acting independently. Figure 4 shows numerically computed results from the solutions of eqns [12] and [8] to illustrate the effects of nonlinear and dispersive evolution over a range of propagation distances. The figure shows the pulse intensity as well as the time-dependent phase and the corresponding frequency chirp. The evolution is clearly very different in the presence of only nonlinear or dispersive effects. Nonlinear SPM alone induces no change in the temporal intensity of the pulse, but the magnitudes of the SPM-induced phase and chirp increase with propagation distance. In contrast, GVD induces changes in both the intensity and the phase of the pulse as it propagates, and the temporal broadening with propagation is associated with a significant change in the functional forms of the phase and frequency chirp. In particular, whilst for large propagation distances it can be seen that GVD induces a parabolic temporal phase and corresponding linear chirp, this is clearly not the case for small propagation distances. This result strongly differs from that obtained for a propagating Gaussian pulse for which the temporal phase is always parabolic. In the context of soliton propagation, the results in Figure 4b are of particular significance, since they clearly show that at $z = L_\mathrm{D}/10$ where neither SPM nor GVD noticeably modify the pulse intensity profile, the SPM and GVD-induced phases and frequency chirps have qualitatively similar forms, yet opposite sign. The split-step Fourier method demonstrates that over a sufficiently small propagation distance (e.g., $\delta z = L_\mathrm{D}/10$) the effects of nonlinearity and dispersion can be considered to act independently on a propagating pulse. This suggests that, over the propagation segment ($0 < z < \Delta z$), the phases are capable of cancelling each other across the entire pulse profile. It is obvious that, if the effects cancel over this distance, then the propagation in the following segment will begin from the same initial condition of an unchirped pulse which was the case at $z = 0$. The nonlinear and dispersive phase components will therefore continue to cancel over successive iterations for an arbitrary length of fiber, and we thus obtain a simple physical interpretation to explain the distortion-free propagation of an optical soliton. Thus, the combined effects of GVD and SPM can prevent the occurrence of a frequency chirp across the pulse. In this situation the pulse propagates with a constant temporal profile. Such light pulses are called optical solitons.

Mathematically, the basic NLSE, eqn [5], belongs to a remarkable class of integrable nonlinear equations, which can be solved exactly for an arbitrary initial condition by means of the inverse scattering method. The solutions depend strongly on the sign of dispersion, and the relative importance of dispersive and nonlinear effects, measured by a dimensionless parameter N:

$$N^2 = \frac{L_\mathrm{D}}{L_\mathrm{NL}} = \frac{\gamma P_0 \delta_0^2}{|\beta_2|} \qquad [15]$$

In the anomalous dispersion regime ($\beta_2 < 0$) the NLS equation admits envelope-soliton solutions. The soliton order is given by the parameter N. Only the fundamental soliton, defined by $N = 1$, can propagate in the fiber with a constant profile, as schematically represented in Figure 5.

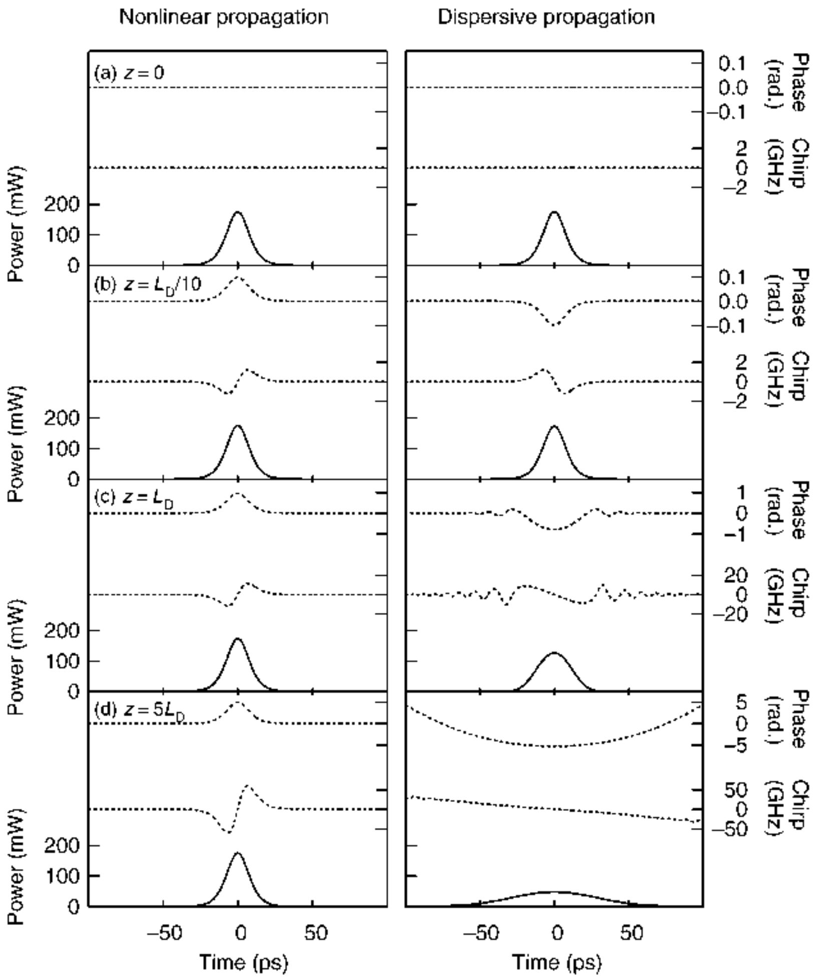

Figure 4 Numerical results showing nonlinear (left column) and dispersive (right column) propagation in an optical fiber, with parameters as described in the text, and for propagation distances as indicated. The solid lines show the pulse intensity (left axis) whilst the dashed lines show the calculated phase (right axis, top) and chirp (right axis, bottom). Constant phase offsets have been removed for clarity.

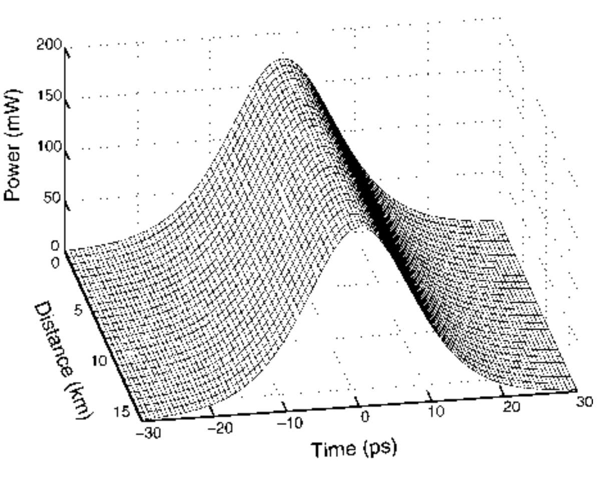

Figure 5 Temporal evolution of the power of a fundamental bright soliton, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 67$ mW, $\beta_2 = -20\ \mathrm{ps}^2\ \mathrm{km}^{-1}$, $\gamma = 3 \times 10^{-3}\ \mathrm{m}^{-1}\ \mathrm{W}^{-1}$ and $\delta_0 = 10$ ps. The nonlinear length is $L_{\mathrm{NL}} = 5$ km.

For the fundamental soliton, dispersion and nonlinearity exactly balance each other in a way such that neither the pulse shape nor the pulse spectrum change during the pulse propagation. The fundamental soliton ($N = 1$) is given by

$$A(z,t) = \sqrt{P_0}\ \mathrm{sech}(t/\delta_0)\ \exp(\mathrm{i}\gamma P_0 z/2) \qquad [16]$$

where the peak power required to generate this soliton is determined by $L_{\mathrm{D}} = L_{\mathrm{NL}}$, and given by

$$P_0 = \frac{|\beta_2|}{\gamma \delta_0^2} \qquad [17]$$

Note that the phase of this bright soliton remains constant across the entire pulse. For a standard single-mode fiber (SMF), at 1.55 μm wavelength, typical values of the dispersion and nonlinear parameters are $\beta_2 = -21\ \mathrm{ps}^2\ \mathrm{km}^{-1}$ and $\gamma = 1.2 \times 10^{-3}\ \mathrm{W}^{-1}\ \mathrm{m}^{-1}$, respectively.

So, the peak power required to propagate a fundamental soliton with pulse width $\delta_0 = 10$ ps, is $P_0 = 175$ mW. This soliton is represented in Figure 5. A further reduction of P_0 can be obtained by using dispersion-shifted fibers (DSF), for which the dispersion can be significantly reduced to values of the order of 1 ps^2 km^{-1}. In this situation, the power required to create a soliton reduces to $P_0 = 8.3$ mW. Thus, fundamental solitons can be generated at power levels that are easily obtainable from conventional laser diodes. The first experimental observation of soliton propagation in optical fibers was reported in 1980 by Mollenauer, Stolen and Gordon.

Note that all the higher-order solitons ($N > 1$) execute a recurrent deformation during their propagation, as illustrated in Figure 6 in the case of the second-order soliton ($N = 2$).

Dark Solitons

In the normal dispersion regime ($\beta_2 > 0$) the chirps induced by GVD and SPM reinforce each other and thus make the hyperbolic secant pulse more susceptible to temporal broadening. Thus, in the normal dispersion regime bright solitons cannot propagate. However, in this case the NLS equation, eqn [5], admits a hole-soliton solution. In other words, the intensity profile of this soliton exhibits a dip in a uniform background, as represented in Figure 7.

Mathematically, this hole-soliton, commonly called a dark soliton, takes the following form:

$$A(z,t) = \pm\sqrt{P_0}\,\tanh(t/\delta_0)\,\exp(i\gamma P_0 z) \qquad [18]$$

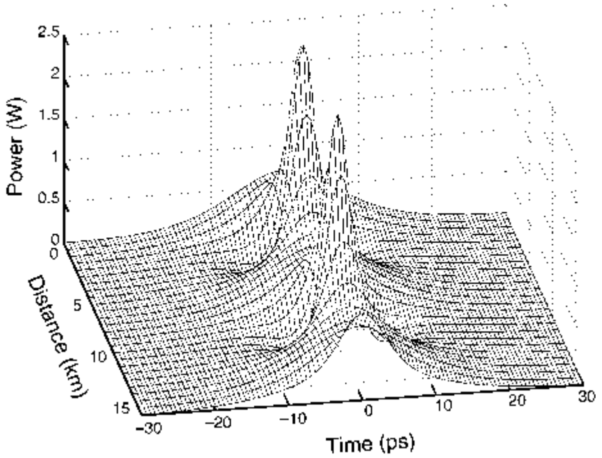

Figure 6 Temporal evolution of the power of a second-order soliton, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 267$ mW. The fiber parameters are the same as in Figure 5. The nonlinear length is $L_{NL} = 1.25$ km.

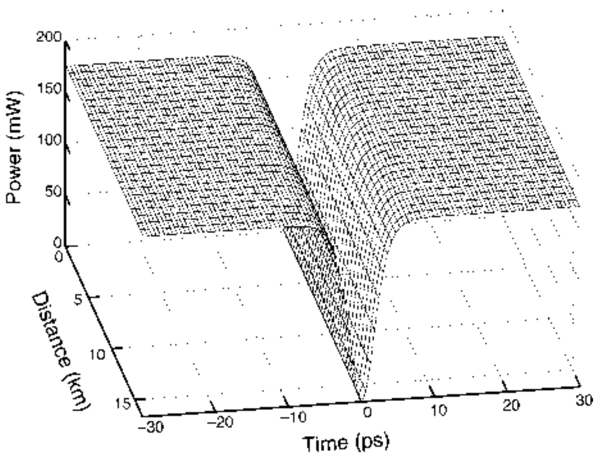

Figure 7 Temporal evolution of the power of a dark soliton, as a function of the propagation distance, for an input peak power $P_0 = 67$ mW.

As eqn [18] shows, the amplitude of the dark soliton is given by an odd function of time. Thus its phase exhibits an abrupt π jump at the dip center, thus implying that dark solitons are chirped. The phase between each jump is constant. Note that there exist other soliton solutions that are similar to the dark soliton, but for which the intensity at the dip center does not fall to zero and the phase variation is more gradual and smaller. These solutions are often called gray solitons.

A common feature of bright and dark solitons is their robustness. This property is very important for ensuring practical applications in optical communications. Moreover, optical solitons emerge unchanged from collision processes. However, dissipative perturbations such as material loss or Raman scattering can destroy such solitons.

Dispersion-Managed Solitons

As we mentioned above, the fundamental soliton is based on a delicate balance between dispersion and nonlinearity, through the relation $L_D = L_{NL}$. In addition, the existence of this soliton requires that only these two effects are present in the fiber. But in practice, additional linear and nonlinear effects such as fiber losses, Raman scattering, or third-order dispersion, are always present in the fiber. In particular, the fiber losses cause the soliton power to continually decrease during the propagation. The use of lumped amplifiers along the transmission line allows one to compensate for the power loss, but does not prevent the occurrence of chirp. As a consequence, the fundamental soliton does not survive over long distances (a few tens of kilometers), without in-line control. Moreover, this soliton is not stable against phenomena such as CPM and FWM processes that occur unavoidably in WDM systems. The CPM between the pulses in a given channel and the pulses in other channels induce a frequency chirp (in the considered channel), which is converted into amplitude variations by the fiber dispersion. The FWM induces energy-exchange processes between different channels, which lead to pulse distortions.

A new transmission scheme, called *dispersion management*, was recently proposed to resolve the above-mentioned limitations. Basically, the dispersion-management technique utilizes a transmission line with a periodic dispersion map, such that each period is built up by two types of fibers, generally with different lengths and opposite group-velocity dispersion. Whereas the fundamental soliton requires total cancellation of the chirp at each propagation distance, the dispersion-managed (DM) soliton induces a small variation of the chirp, with a zero mean value over each period of dispersion management.

The DM soliton exhibits major differences with respect to the fundamental bright soliton. First, whereas the fundamental bright soliton possesses a smooth profile (as Figure 5 shows), the profile of the DM soliton exhibits an internal oscillatory structure, with several side-lobes on each side of the soliton's center. A typical representation of a DM-soliton profile is shown in Figure 8, at the free-chirp point of the dispersion map. Another outstanding difference is the following: whereas the fundamental soliton propagates with a constant profile, the DM soliton executes a periodic deformation, with a period that coincides with that of the dispersion map. The minimum value of the soliton's width occurs at the free-chirp point. From this point, the soliton's width can execute several oscillations within one dispersion map, with a peak value that can be quite large compared with the width at the free-chirp point. Figure 9 illustrates the propagation of a DM soliton within one dispersion map of a lossless DM line.

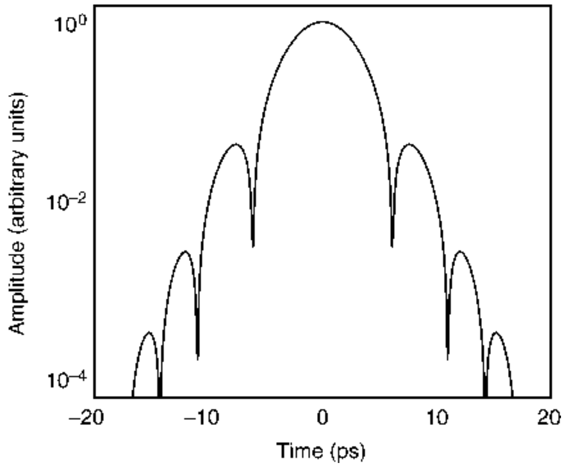

Figure 8 Plot showing the profile of a DM soliton at the free-chirp point of a typical DM line made up of 20.5 km of anomalous-dispersion fiber [$\beta_2 = -1.5 \times 10^{-3}$ ps^2 m^{-1}, $\gamma = 1.4 \times 10^{-3}$ W^{-1} m^{-1}], and 19.5 km of normal-dispersion fiber [$\beta_2 = 1.5 \times 10^{-3}$ ps^2 m^{-1}, $\gamma = 1.4 \times 10^{-3}$ W^{-1} m^{-1}].

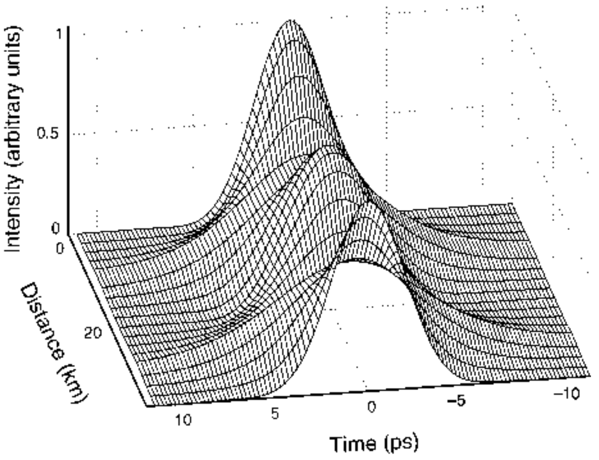

Figure 9 Plot showing the evolution of the DM soliton's profile as a function of the propagation coordinate *z*.

Recent experiments have demonstrated that in this new transmission scheme the pulse propagation becomes much more stable than fundamental solitons against linear and nonlinear effects such that SPM, FWM, CPM, or third-order dispersion. The DM transmission lines are currently subject to intense research with a view both to upgrade the capacity of existing terrestrial networks, and to design submarine fiber systems with ultrahigh capacity.

Modulational Instability

In optical fibers, the interaction between nonlinear and dispersive effects may lead to modulational instability (MI), a phenomenon in which a continuous or quasicontinuous wave undergoes a modulation of its amplitude or phase in the presence of noise or any other small perturbation. The perturbation can originate from quantum noise (spontaneous-MI) or from a frequency shifted signal wave (induced-MI). MI was observed for the first time by Tai, Hasegawa and Tomita in 1986 for a single pump wave propagating in a standard nonbirefringent fiber (scalar MI). Scalar MI occurs only when the group velocity dispersion (GVD) is negative (anomalous dispersion regime). Modulational instability in eqn [5] is examined through a linear stability analysis (LSA) of its steady state solution

$$A = \sqrt{P_0} \exp[\mathrm{i}\gamma P_0 z] \qquad [19]$$

Equation [19] shows that the power remains constant through propagation along the fiber. The linear stability of the steady state solution is examined by looking into the system in the presence of small amplitude and phase perturbations *a;* that is, we consider

$$A = (\sqrt{P_0} + a) \exp[\mathrm{i}\gamma P_0 z] \qquad [20]$$

Then, the linearization of the NLSE with respect to a yields

$$\frac{\partial a}{\partial z} + \frac{\mathrm{i}}{2}\beta_2 \frac{\partial^2 a}{\partial t^2} = \mathrm{i}\gamma P_0 (a + a^*) \qquad [21]$$

Then one can assume for the perturbation a modulational ansatz with wavenumber K and frequency Ω, of the form

$$a(z,t) = u\cos(Kz - \Omega t) + \mathrm{i}v \sin(Kz - \Omega t) \qquad [22]$$

Substitution of eqn [22] into eqn [21] leads to the following dispersion relation for the perturbation

$$K^2 = \frac{\beta_2^2 \Omega^2}{4}\left[\Omega^2 + \frac{4\gamma P_0}{\beta_2}\right] \qquad [23]$$

The MI phenomenon manifests itself by an exponential growth of the amplitude of the perturbation. This behavior occurs when the wavenumber K possesses a nonzero imaginary part. The dispersion relation [23] clearly shows that MI strongly depends on the sign of the dispersion. For normal dispersion ($\beta_2 > 0$), K is always real and the steady state is stable, whereas in the anomalous dispersion regime, MI appears for $\Omega < \Omega_c = \sqrt{4\gamma P_0/|\beta_2|}$. The importance of the phenomenon is measured by a power gain defined by

$$g(\Omega) = 2|\mathrm{Im}(K)| = |\beta_2|\Omega\sqrt{\Omega_c^2 - \Omega^2} \qquad [24]$$

Figure 10 shows the MI-gain spectra at three power levels with fiber parameters appropriate for standard silica fibers at $\lambda_0 = 1.319$ µm.

The peak value of the gain, $g_{\max} = 2\gamma P_0$, occurs at a frequency

$$\Omega_{\mathrm{opt}} = \sqrt{\frac{2\gamma P_0}{|\beta_2|}} \qquad [25]$$

called the optimal modulational frequency. In practice, when the perturbation corresponds to quantum noise, the spectral components that are amplified by the MI-process are close to Ω_{opt}. Here, as shown by Figure 11, the MI phenomenon leads to two sidebands that are situated symmetrically with respect to the pump, at frequency detunings of $\approx \pm\Omega_{\mathrm{opt}}$ (see Figure 10 for a pump peak power $P_0 = 3$ W).

In the time domain, induced-MI leads to the break-up of the quasi-cw pump wave into a train of ultrashort pulses. The repetition rate of the pulses is determined by the frequency detuning between the signal and the pump (i.e. the modulational frequency). The temporal shape of these ultrashort pulses depends not only on the powers

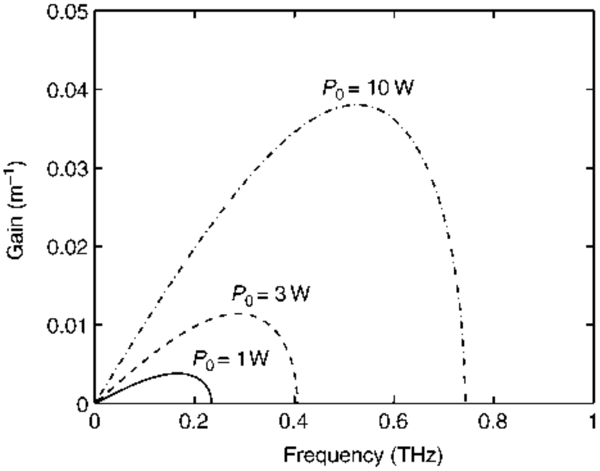

Figure 10 Spectral profiles of MI gain g for a pump peak power $P_0 = 1$ W (solid), 3 W (dashed) and 10 W (dot-dashed). The fiber parameters are $\beta_2 = -3.5$ ps^2 km^{-1} and $\gamma = 1.9 \times 10^{-3}$ m^{-1}W^{-1}.

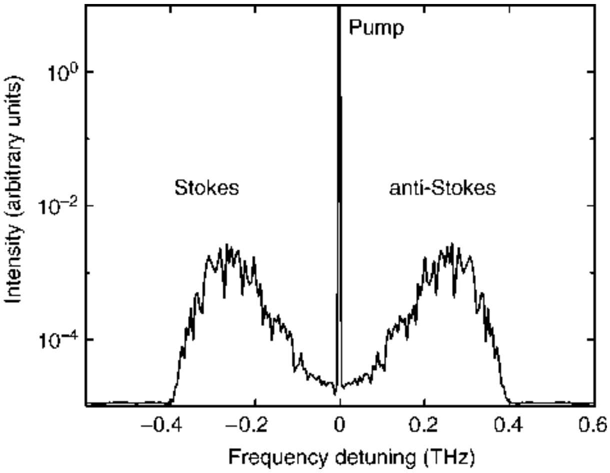

Figure 11 Spontaneous modulational instability spectrum (i.e., parametrically amplified spontaneous noise), as obtained for a peak pump power $P_0 = 3$ W and a fiber length $L = 1$ km. The fiber parameters are the same as in Figure 10.

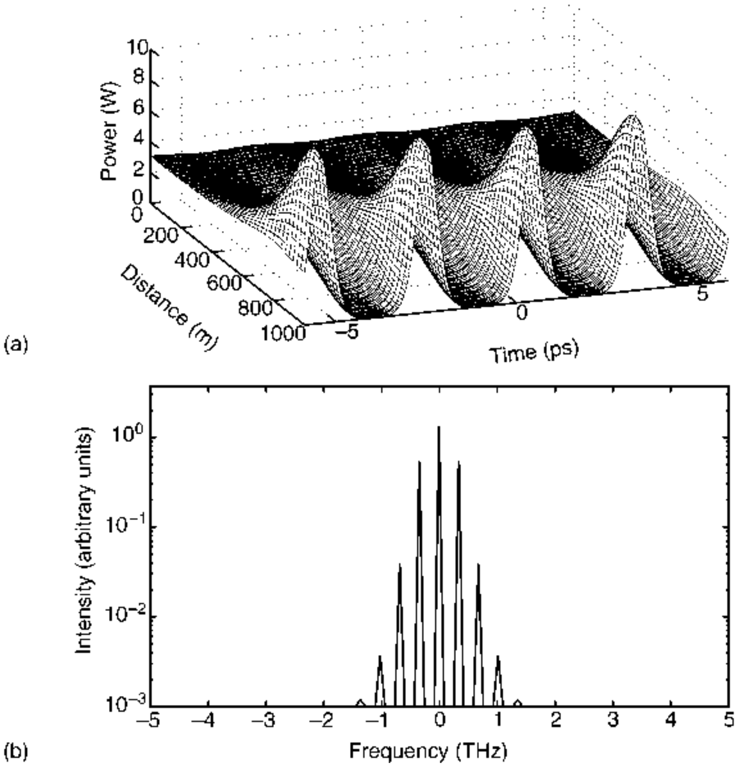

Figure 12 (a) Theoretical temporal evolution of the power for induced-MI, as a function of the propagation distance. (b) Output power spectrum in logarithmic intensity scale. The pump (signal) power is 3 W (0.5 mW), the pump-signal frequency detuning is 0.34 THz, and the fiber length is 1 km. The fiber parameters are the same as in Figure 10.

of the input waves but also on the modulational frequency. In particular, trains of bright and dark solitons with a terahertz repetition rate have been generated by MI in optical fibers. Figure 12a shows theoretically the evolution of a small initial modulation.

As Figure 12a shows, MI causes a small initial perturbation to evolve into a periodic pulse train, whose period is inversely proportional to the frequency spacing Δf between the pump and signal waves. Thus the technique of induced-MI can be exploited for generating high-repetition-rate ultrashort pulse trains. As shown in Figure 12b, such pulse train generation manifests itself in the frequency domain by the growth of a cascade of sidebands. The number of harmonics ($n\Delta f$) is determined by the temporal shape of the generated pulses and the initial conditions at the fiber input.

Cross-phase modulation (CPM) between two modes can be used to extend the instability domain to the normal dispersion regime (positive GVD). This is achieved by using orthogonally polarized pump waves. This CPM-induced MI is called vector modulational instability. The optical field evolutions are described by two coupled NLS equations. Several experiments involving two polarization modes, or two spatial modes, have been done in order to observe MI in the normal GVD regime.

Conclusion

The invention of soliton transmission, predicted in 1973 and experimentally demonstrated in 1980, is one of the major discoveries made in the area of optical telecommunication. Although from the period of invention of optical solitons up to very recent years the main line of research activities was based on the concept of the ‘classical soliton’, which represents an exact balance between the fiber group-velocity dispersion and its intensity-dependent refractive index, most recent experimental and theoretical developments focus on dispersion-managed solitons. It is now widely recognized that DM solitons in fiber-optic links are the most promising way both to upgrade the capacity of existing terrestrial telecommunication networks and to design submarine fiber systems. Most of the current research and development activities in this context aim to develop a non-empirical methodology of designing ultrahigh-speed DM fiber systems.

List of Units and Nomenclature

Dispersion length L_D [m]
Group velocity dispersion coefficient β_2
[$\mathrm{ps}^2\,\mathrm{km}^{-1} = 10^{-27}\,\mathrm{s}^2\,\mathrm{m}^{-1}$]
Nonlinear length L_{NL} [m]
Nonlinear parameter γ [$\mathrm{W}^{-1}\,\mathrm{m}^{-1}$]
Nonlinear refractive index n_2 [$\mathrm{m}^2\,\mathrm{W}^{-1}$]
Optical intensity I [$\mathrm{GW\,cm}^{-2} = 10^9\,\mathrm{W\,cm}^{-2}$]

CPM: cross-phase modulation
DM Soliton: dispersion-managed soliton
FWM: four-wave mixing
GVD: group velocity dispersion
LSA: linear stability analysis
MI: modulational instability
NLS (E): nonlinear Schrödinger (equation)
SPM: self-phase modulation
WDM: wavelength division multiplexing

See also

Fiber and Guided Wave Optics: Dispersion; Light Propagation; Nonlinear Effects (Basics). **Lasers:** Optical Fiber Lasers. **Scattering:** Scattering Theory. **Solitons:** Soliton Communication Systems.

Further Reading

Ablowitz MJ and Clarkson PA (1991) *Solitons, Evolution Equations and Inverse Scattering*. Cambridge, UK: Cambridge University Press.
Agrawal GP (1997) *Fiber-Optic Communication Systems*. New-York: Wiley-Interscience.
Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Hasegawa A (1989) *Solitons in Optical Fibers*. New-York: Springer Verlag.
Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. Oxford, UK: Clarendon Press.
Remoissenet M (1999) *Waves Called Solitons, Concepts and Experiments*, 3rd edn. Berlin: Springer.
Taylor JR (ed.) (1992) *Optical Solitons – Theory and Experiment*. Cambridge, UK: Cambridge University Press.
Zakharov VE and Wabnitz S (eds) (1999) *Optical Solitons: Theoretical Challenges and Industrial Perspectives*. Les Ulis (France) and Berlin (Germany): EDP Sciences and Springer.

Soliton Communication Systems

M Karlsson and P Andrekson, Chalmers University of Technology, Gothenburg, Sweden

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In 1973 Hasegawa and Tappert proposed that optical soliton pulses in fibers could exist. A soliton is a propagating wave packet that is localized in the sense that it does not spread its energy during propagation, and with the additional property that it is so stable that it can collide with other solitons and emerge unaffected with respect to energy, shape, and momentum after the collision. Solitons are based on some kind of nonlinearity in the system, and for optical fibers the weak Kerr nonlinearity (which makes the refractive index increase in proportion to the optical intensity) can counteract the pulse broadening induced by group-velocity dispersion (GVD). The two effects can form a stable balance in the form of a soliton pulse, which then propagates without changing shape.

Due to the lack of short-pulse laser sources at wavelengths above 1.3 μm, and low-loss silica fibers, it took another seven years for optical soliton pulses to be experimentally verified. In an experiment by Mollenauer, Stolen, and Gordon in 1980, soliton pulse transmission over 700 m fiber was demonstrated. During the 1980s soliton research aimed towards the use of solitons as information carriers in optical communications, and in 1990 the first data transmission experiment using solitons (2.8 Gb/s over 23 km of fiber) were reported by Iwatsuki *et al.*

During the 1990s, soliton-based communication systems have matured, an important reason being the development of the erbium-doped fiber amplifier (EDFA), which made high power levels commercially available. The most recent developments have been towards the use of solitons in alternating dispersion maps (so-called dispersion management), and such systems have reached performance levels near commercialization.

It should be stressed that fiber soliton research has been almost exclusively devoted to single-mode fibers at a wavelength of 1550 nm. In principle, solitons can also be used to counteract mode dispersion in multimode fibres, but that generally requires unrealistically high peak powers, and is therefore regarded to be of less interest. Similarly, solitons with wavelengths around 1300 nm in single-mode fibers have been considered, but found to be of less interest due to the lack of low-noise optical amplifiers in this spectral range. Another kind of soliton, the so-called dark soliton, can also be generated in single-mode fibers. However, dark solitons have only been used as information carriers in a very limited number of experiments, and will also be omitted from this discussion.

In the present review we will discuss both theoretical and experimental aspects of soliton transmission. We will distinguish between conventional solitons, which have constant dispersion during transmission, and dispersion-managed solitons (although solitary waves would be the proper mathematical name) for which the dispersion varies periodically during transmission.

Conventional Fiber Solitons

Group-velocity dispersion means that the group velocity of the light changes with the optical frequency, and it is usually defined as the second derivative of the propagation constant β with respect to the angular frequency ω, and denoted β_0'' [ps^2 km^{-1}]. Standard single-mode fibers have a $\beta_0'' \approx -20$ [ps^2 km^{-1}] around a wavelength of 1550 nm, whereas dispersion-shifted fibers (DSFs) have β_0'' close to zero in the 1550 nm region. The dispersion is defined with respect to sign, so that the GVD is said to be normal if the group velocity decreases with frequency (i.e., $\beta_0'' > 0$), and anomalous if the group velocity increases with frequency (i.e., $\beta_0'' < 0$). A pulse propagating in a dispersive fiber will broaden with propagation distance at a rate proportional to the product of β_0'' and the spectral width of the pulse. In a communication system such a broadening is devastating, since neighboring pulses will interfere and disperse power into adjacent bit slots. The problem is increased for high data rates (as this requires short pulses with broad spectra) and/or long transmission distances.

The nonlinearity in optical fibers manifests itself as a dependence of the refractive index with power, and is known as the optical Kerr effect. For pulse propagation the Kerr effect will give rise to self-phase modulation (SPM), which means that the pulse will get a phase modulation proportional to its own intensity. Mathematically speaking, the amplitude of an optical pulse $u(0, t)$ [W$^{1/2}$] launched at $z = 0$ will evolve according to $u(z,t) = u(0,t)\exp(-\mathrm{i}\gamma z|u(0,t)|^2)$, where γ [W^{-1} km^{-1}] is the nonlinear coefficient of the fiber, z [km] is the propagation distance along the fiber, t [ps] is the time coordinate, and $|u(z, t)|^2$ [W] is the power level of the light. The value of γ varies between different fiber types due to the dependence on the core area of the fiber, but for standard single-mode fibers, in the 1550 nm region, it is approximately 2.2 km^{-1} W^{-1}.

Taking both self-phase modulation and dispersion into account, one can derive a propagation equation for light, known as the nonlinear Schrödinger (NLS) equation. Of particular importance for the solution to the NLS equation is solitons, which is a solution that propagates without dispersive broadening. The soliton solution arises as a result of a cancellation of the dispersion and the self-phase modulation effects, and it can only be formed when the dispersion is anomalous. For normal dispersion the nonlinearity and the dispersion will instead act together to increase the spectral broadening of the pulse. The soliton solution of duration t_0 and dispersion β_0'' reads

$$u_{\mathrm{sol}}(z,t) = \sqrt{\frac{1}{\gamma L_{\mathrm{D}}}}\,\mathrm{sech}\left(\frac{t}{t_0}\right)\exp\left(-\mathrm{i}\frac{z}{2L_{\mathrm{D}}}\right) \qquad [1]$$

where the dispersive length is $L_{\mathrm{D}} = t_0^2/|\beta_0''|$ [km]. The soliton peak power P_{s} [W] and energy E_{s} [J] can be expressed as $P_{\mathrm{s}} = |u_{\mathrm{sol}}(z,0)|^2 = |\beta_0''|/\gamma t_0^2$ and $E_{\mathrm{s}} = 2P_{\mathrm{s}}t_0 = 2|\beta_0''|/\gamma t_0$.

The fact that solitons are not broadened by dispersion will make them ideally suited as information carriers in a communication system. However, the solitons are affected by various perturbations that will affect the balance by dispersion and nonlinearity. Unless proper care is taken in the design of the system, those perturbations can destroy the performance of the communication system. Next, we will review the most important perturbations and design issues of soliton-based communication systems.

Solitons in the Presence of Third-Order Dispersion

The third-order dispersion (TOD), or the dispersion slope, is modeled with the derivative of the GVD, β_0'' [ps^3 km^{-1}], which is typically 0.1 ps^3 km^{-1} in standard and dispersion-shifted fibers. For solitons in dispersion-shifted fibers the effect of the dispersion slope is increasingly important the closer the solitons are to the zero dispersion wavelength.

When a soliton is placed close to the zero-dispersion wavelength, it will emit radiation in the normal dispersion regime of the spectrum, lose energy, and recoil further into the anomalous dispersion region so that the radiation stops and a stable soliton is formed. As a design criterion $\beta_0'''/(|\beta_0''|t_0) < 0.24$ have been suggested for this phenomenon to provide negligible energy loss.

Solitons in the Presence of Amplification and Loss

The most important property that has been neglected in the derivation of the NLS equation for optical pulses is the effect of loss. The loss will decrease the power and hence the nonlinearity so that the balance sustaining the soliton is no longer maintained. As a result of the loss (which is of the order of 0.2 dBkm^{-1} in standard transmission fibers) the soliton pulse will be broadened. Partly, this can be circumvented by requiring the launched peak power of the soliton after each amplifier to be such that the path-average power between the amplifiers equals the soliton power, P_{s}. The peak power P_{peak} falls off as $\exp(-\alpha z)$, where $\alpha = 0.042\ \mathrm{km}^{-1}$ is the fiber loss coefficient, and the average peak power over one amplifier span is then $P_{\mathrm{peak}}\ (1 - \exp(-\alpha L_a))/\alpha L_a = P_{\mathrm{peak}}(G - 1)/G\ \ln(G) = P_{\mathrm{s}}$, where $G = \exp(\alpha L_{\mathrm{a}})$ is the gain of each amplifier and L_{a} [km] is the amplifier spacing.

Nevertheless the periodic fluctuation of power and width over each amplifier length may cause the solitons to become unstable and lose power to dispersive wave radiation. To minimize this loss a useful design criterion is that the amplifier length L_{a} should be smaller than the dispersive length, L_{D}.

Sources of Timing Jitter

Another transmission obstacle is the various sources of random movement of the pulse in the bit slot, i.e., timing jitter of the pulses in the data transmission link. There are various sources of timing jitter, e.g., soliton interaction, Gordon–Haus, acoustic, and wavelength division multiplexing collision induced jitter.

Soliton interactions

Solitons, being nonlinear pulses, will interact with adjacent pulses in the pulse train. The interaction between solitons of the same polarization state and wavelength is phase sensitive so that in-phase solitons will attract each other whereas out-of-phase solitons will repel each other. To reduce this effect the solitons must be sufficiently spaced in the data stream. A typical design condition is that the pulse width of the solitons are one fifth of the bit slot, or one fifth of the inverse bit rate. A straightforward way of reducing the interaction is to have alternating amplitudes of the soliton in the pulse train, and the solitons can then be packed more densely.

Orthogonally polarized solitons interact substantially less, since it is the intensity overlap that causes the interaction, rather than the amplitude overlap as for copolarized solitons. In polarization multiplexed systems where adjacent pulse have orthogonal polarization it is usually enough with half the above separation, i.e., a pulse duration of 2.5 less than the bit separation.

Gordon–Haus jitter

Noise from the in-line amplifiers will give rise to a small jitter in the carrier frequency of each soliton, which, via the dispersion, will alter the group velocity and hence affect the arrival time of each pulse. This is known as the Gordon–Haus effect, and it has to be accounted for in long-distance systems. The variance of the timing jitter is proportional to the ratio of the power spectral density of the noise to the energy of the soliton, and also proportional to the transmission distance cubed. The fact that Gordon–Haus jitter grows cubically with distance makes it particularly important at transoceanic lengths, typically exceeding 1 Mm.

Acoustic jitter

The electrostriction nonlinearity in the fiber gives rise to a mechanical pressure proportional to the optical intensity, which in turn modifies the refractive index of the fiber. In particular, an intense optical pulse like a soliton will give rise to a pressure (acoustic) wave moving radially outwards from the fiber center. The pulses in the wake of this wave will experience a randomly changing local refractive index, and hence (just as for Gordon–Haus jitter) a randomly changing carrier frequency that transforms into a timing jitter. The variance of this jitter will grow as the transmission distance to the power of four.

WDM Considerations

Wavelength division multiplexing (WDM) is another way of increasing the bit-rate of soliton systems. That is, several frequency channels are used for solitons transmission. The problem with WDM transmission using solitons stems mainly from collisions between solitons in the different wavelength channels. In a perfectly ideal NLS equation solitons would collide elastically without changing carrier wavelength. However, the presence of losses and amplification may cause the collisions to be asymmetric if they occur around an amplifier, and the result will be a frequency displacement and a concomitant timing jitter.

Soliton Control

Soliton control is the common name for methods to control the soliton parameters such as wavelength and position, mainly in order to reduce the deleterious timing jitter effects. There are two different approaches: passive and active control.

Passive soliton control has been suggested in the form of filters that are inserted along the transmission path. This helps to keep the soliton wavelength fixed. In this way not only Gordon–Haus and acoustic jitter can be remedied, but also interaction jitter and

WDM-collision-induced jitter. A problem with this kind of filtering is that it defines a spectral region with excess gain, in which amplifier noise will grow excessively. A way around that problem is to slightly shift the center wavelength of the filters along the transmission path. In that way the solitons will follow the frequency shift, but the linear noise will not. Such sliding filter experiments have demonstrated 8 × 10 Gb/s WDM soliton transmission over 10 Mm.

Active control usually acts in the time domain by using phase or amplitude modulators to retime and reshape the solitons. Using this technique, 10 Gbit/s over unlimited distances has been demonstrated. However, this kind of active reshaping of the pulses suffers from the same drawbacks as conventional electronic regeneration, i.e., incompatibility with WDM, complexity, and high cost.

Dispersion-Managed Solitons

Introduction

In the last years of the 1990s, solitons have become substantially more attractive through the rapidly emerging strategy of improving the performance of soliton transmission with dispersion management (DM). While the DM strategy, which involves altering the local dispersion between a large positive and a large negative GVD such that the average GVD is small, has long been used in linear systems, it was only relatively recently appreciated that the same technique, if properly implemented, gives rise to several very striking improvements over conventional soliton transmission systems. While DM-solitons are clearly nonlinear pulses they are by no means classical solitons.

From a commercial viewpoint, however, the most important benefit with using DM-solitons is that they, in principle, can use the already installed conventional fibers (with zero dispersion at 1300 nm) allowing a much more cost-effective upgrade together with dispersion-compensating fibers (DCF) or chirped fiber gratings, which can be located together with in-line amplifiers at certain intervals in the link.

In Figure 1, an example of breathing in spectral and temporal widths of a DM-soliton can be seen, together with its dispersion map at the bottom.

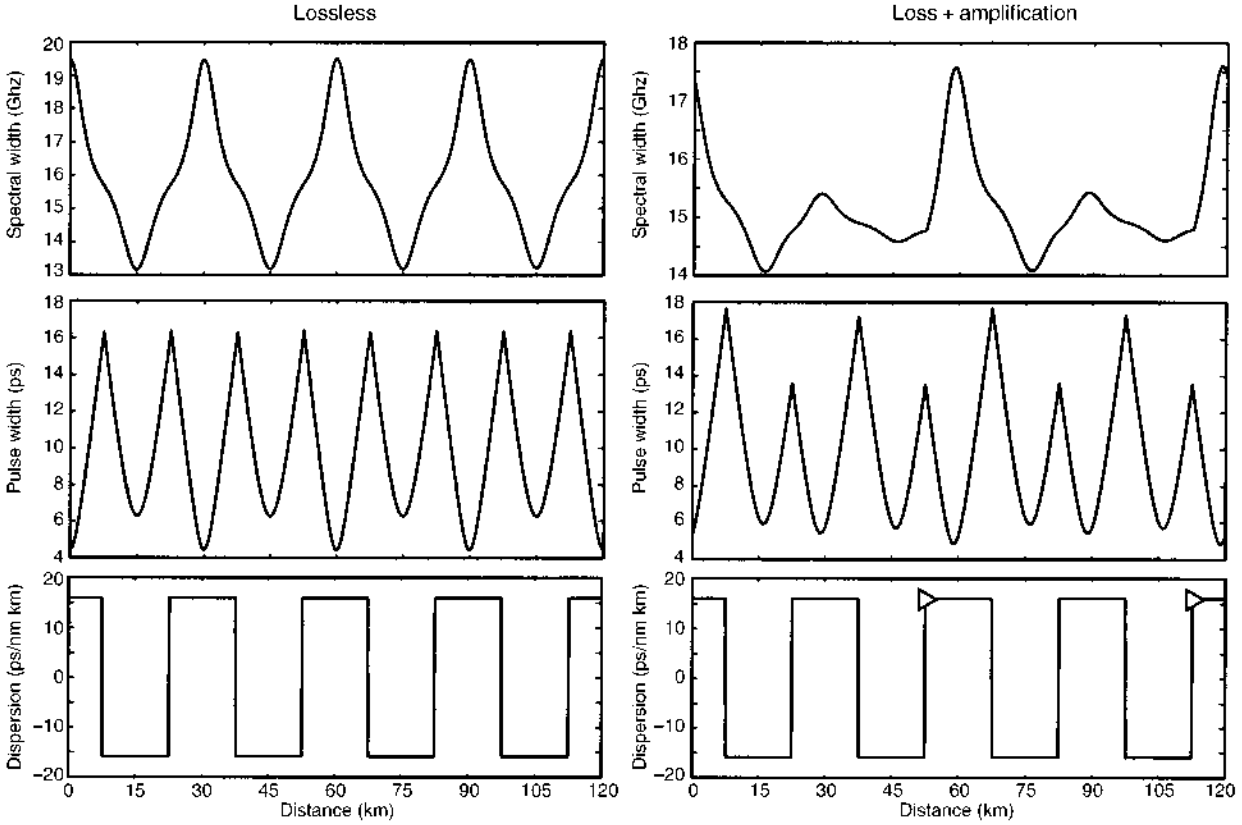

Figure 1 Plot of the spectral width (top), temporal width (middle), dispersion map (bottom) of a dispersion-managed soliton in the lossless case (left column) and the lossy case that is periodically amplified (right column). The positions of the amplifiers are denoted by the small triangles in the plot of the dispersion map.

Two cases are shown; one without fiber attenuation (left column) and one including fiber losses and periodic, lumped amplification every 60 km. One may note the strictly periodic behavior of both the spectral and temporal widths in both cases, and that the evolution of the spectral width is significantly different in the lossy case. This is because a change of the pulse spectrum requires nonlinearities and high power, and the spectral broadening is therefore highest directly after the amplifiers.

Properties

DM-solitons emerged from extensive simulation work, and it was quite surprising to many researchers that the simulations revealed such stable and strictly periodic pulses. The periodicity follows that of the dispersion variation of the system. Usually the system under study has a dispersion map, i.e., a plot of how the GVD varies along the transmission path. The map has a characteristic period (typically 10–100 km) and an average GVD which is significantly lower than the local GVD in the map.

As a result of the massive simulation work done by many groups the following properties of DM-solitons have been found:

- The pulse width, chirp, and spectral width oscillates periodically in the dispersion map. There are two points within the period at which the pulse is chirp-free, and those correspond to local minima of the pulse width. One of those points is the global minimum width, referred to as the 'shortest pulse' below.
- A central parameter that is useful for the characterization of DM-solitons is the map strength, $S = (L_1|\beta_1''| + L_2|\beta_2''|)T_{\mathrm{FWHM}}^2$, where L is the length, β'' is the dispersion, T_{FWHM} is the minimum pulse width in the full width half maximum sense, and the subscripts 1 and 2 refer to the two fibers in the dispersion map. In this example we thus assume the map to consist of two fibers, one with normal and one with anomalous dispersion. Physically the map strength is the number of dispersive lengths the pulse propagates in one period. DM-solitons have been found for map strengths ranging from $S = 0$ (which is the same as conventional solitons) to $S \approx 12$, although this upper limit is a transition regime in which the pulses radiate and perfect periodic evolution never arises.
- DM-solitons have been found for anomalous, normal, and zero average dispersion β'', defined as $\beta_0'' = (L_1\beta_1'' + L_2\beta_2'')/(L_1 + L_2)$. Normal average dispersion is only possible for map strengths above 3.9.
- The shortest pulse shape ranges from hyperbolic secant at $S = 0$ to Gaussian for higher map strengths, and this is also evident from the time-bandwidth product, which increases with S from 0.32 (sech-shape) at $S = 0$ to 0.44 (Gaussian shape) and even higher for large values of S. In addition, the shortest pulse has oscillating tails in the wings.
- The energy of a DM-soliton pulse is enhanced relative to a soliton with the same average dispersion and pulse width.
- The interaction between DM-solitons is less than that of conventional solitons, and an optimum map strength exists that minimizes the interaction.

The fact that DM solitons can work for an average net zero GVD and even for normal dispersion, was a striking and unexpected difference from conventional solitons. This can be understood by the fact that the spectrum is widest in the anomalous dispersion regime (i.e. positive values of the dispersion parameter D, cf. Figure 1), and it has been conjectured that the average dispersion, weighted by the spectral width of the DM-soliton, should always be anomalous. This is then a generalization of the fact that conventional solitons must always have a constant anomalous dispersion.

The technical improvements with DM solitons over conventional ones are numerous. The signal-to-noise ratio is improved since the DM-solitons have a larger peak power than the corresponding conventional solitons. DM-solitons have less Gordon–Haus and acoustic timing jitter, since the system average GVD is much smaller in these systems. A very important added benefit appears in WDM systems. Due to the alteration between large positive and negative GVD along the path, the jitter induced from WDM-soliton collisions is greatly reduced. This, in turn, allows for very dense WDM which will improve the spectral efficiency substantially.

An important practical consequence of using DM-solitons is that they reduce (or eliminate) the need for in-line soliton control such as synchronous modulation or sliding filters. Yet soliton control methods are still applicable and will give improvement in terms of signal-to-noise ratio for DM-solitons as well.

Quite impressive circulating loop experiments including WDM have been reported. For example, 51 densely packed WDM channels each operating at 20 Gb/s were transmitted over 1000 km with 100 km sections of standard fiber, clearly demonstrating the strength of the DM soliton technique.

Quasi-linear Transmission Intrachannel Impairments

As a rule, map strengths in the range 1–8 make best use of the unique features of DM-solitons. This means that the use of installed standard fiber becomes difficult at very high bit-rates (say beyond 30–40 Gb/s) as shorter pulses require a more rapidly (sub-km) varying dispersion map to maintain a proper S-value. If this cannot be maintained, i.e. for S values significantly above 10, one reaches the quasi-linear regime, in which the dispersion significantly dominates the transmission.

In the quasi-linear regime, there are a couple of nonlinear transmission impairments that need to be accounted for and analyzed in more detail. They are the so-called intra-channel effects; intra-channel four-wave mixing (ICFWM), and intra-channel cross-phase modulation (ICXPM), and arise due to the nonlinear interaction between two neighboring pulses. Four-wave mixing (FWM) and cross-phase modulation (XPM) are usually effects associated with WDM transmission. However, the fact that the pulses are chirped and broadened, will cause different frequency components from neighboring pulses within the same wavelength channel to overlap in time, thereby causing FWM and/or XPM within the same channel.

ICFWM arises for two neighboring pulses that, via four-wave mixing (FWM), creates new frequency components that in the time domain will give rise to a new pulse (commonly referred to as a ghost pulse), next to the two. The ghost pulse will then give rise to intersymbol interference and reduction of the eye opening. ICFWM is most prominent for large map strengths and high power.

ICXPM can be viewed as pulse-to-pulse interaction, and physically, it manifests as the frequency shift of one pulse induced by the presence of a neighboring pulse, which, by the dispersion transforms into a timing jitter. The effect can be minimized by selecting proper map strength and pre-chirp of the pulses.

Finally it should be emphasized that these intra-channel impairments will affect also non-soliton systems (e.g., non-return-to-zero (NRZ) modulated systems) if the nonlinearities are significant.

Experiments and Field Trials

Soliton Pulse Sources

When doing soliton experiments, be it conventional or DM-solitons, particular importance is placed on the properties of the pulse source, as it sets the lower limit of the system performance. A high bit-rate soliton pulse source needs to produce low-chirp, low timing-jitter pulses with proper duration (in the ps regime), repetition rate (10–40 GHz), and shape.

One possible choice is gain-switched (GS) laser diodes, possibly with an external cavity for tunability. However, they suffer the drawback of producing pulses that are strongly chirped, asymmetric and often too wide.

For laboratory experiments, fiber ring lasers (FRL) are very attractive, as they provide wavelength and pulse width tunability, besides meeting the above demands. Their drawback is that they are bulky, need active stabilization, and sometimes also temperature control to achieve long-term stability.

Finally it appears quite clear that electro-absorption modulators (EAM), which can be integrated or not with a distributed feedback (DFB) laser, are very useful and simple sources for soliton transmission. While such sources were developed for linear NRZ systems, they have now proven to be near ideal in soliton systems as well. Although EAMs are not commercially available at 40 GHz yet, they are likely to be so soon.

Special considerations need to be taken in DM-soliton systems, however, as the launch condition is different from that for conventional solitons. The pulses should have a linear chirp such that it fits seamlessly in the periodically induced chirp variation along the link. This can be achieved by incorporating a proper length of fiber (or chirped fiber grating) in the transmitter once the overall dispersion map is known.

Loop Experiments

In order to investigate really long distances (megameters) of transmission, loop experiments were developed in the early 1990s. This means that the data pulses are injected in a loop consisting of transmission fiber and amplifier, and then left to propagate a number of laps corresponding to a certain transmission distance. Acousto-optic switches are used to switch the pulse train in and out from the loop at proper time intervals. The drawback of loop experiments is that they may be poor models of reality when it comes to things like dispersion variation along the fiber, polarization mode dispersion (PMD), or various kinds of drift that may arise. In addition, a real system has more options to fine tune, e.g., amplifiers along the transmission line. However, as long as these drawbacks are recognized, loop experiments are very powerful indeed, and invaluable in laboratory evaluations of long-distance transmission.

Field Trials

In the field, there are many transmission link design restrictions and fiber properties which make the systems far from optimal. The actual fiber parameters are nonperiodic with propagation distance as the systems are straight lines rather than relatively short loops. Both loss (in particular when including many contacts and splices along the link) and dispersion might vary significantly along a fiber span. In addition, it may not be possible to tailor the dispersion map and amplifier locations to reach an optimal state. All these examples of nonidealities justify the need for field experiments.

Several soliton field experiments have been conducted over the past few years in Japan by NTT, in the US by MCI/Pirelli, as well as in Europe by ACTS projects, and the results from those experiments are discussed below. It is a good indication that solitons are indeed foreseen as very interesting candidates in commercial systems. Again, it is not very easy to compare the results as the situation in each case differs. All of the systems operated in the 1550 nm range, used optical time division demultiplexing with a 10 Gb/s electronic base rate, and the average loss/km ranged from 0.24 to 0.33 dB/km. Dispersion-shifted fiber was always used, apart from in two cases, where standard fiber was used. One field trial has used DCF for dispersion compensation, which makes this the only DM-soliton field experiment to date.

Polarization multiplexing, for which adjacent pulses have orthogonal polarization, have been used in a few experiments and this serves mainly to allow the use of relatively wide pulses, which in turn allows for larger amplifier spans. Polarization-multiplexing, however, is not as useful if the PMD of the system is high, as then the orthogonally polarized pulses would start to drift statistically in time relative to each other, thereby creating intersymbol interference and increasing the soliton interaction. In the 40 Gb/s cases and above, PMD was found to be the main capacity limiting factor.

Most of the recent experiments used a mode-locked fiber ring laser as a source, mainly because these provide excellent pulse quality as well as tunability in terms of wavelength and pulse width. Other experiments used either gain-switched lasers or electro-absorption modulators.

Future soliton field trials are expected to (1) take advantage of the now well-known strategy of improving soliton transmission performance with dispersion-management, this being very attractive for upgrading existing fiber plants, (2) implement dense WDM (in fiber lines that do not contain dispersion-shifted fibers) to boost aggregate capacity, (3) utilize different forms of in-line control, particularly at high bit rates, and (4) further address the implications of PMD and techniques to combat it. The interesting trade-off between wavelength and time division multiplexing (WDM–TDM) for optimization of overall aggregate capacity will depend on the details of the fiber line parameters.

Future Outlook

To conclude, we note that the motivation for using solitons as information carriers has changed over the years. The property of being resistant to dispersive broadening was originally the main feature, but this was considered less important when dispersion-compensating fibers became commercially available. Instead, this led to the development of the dispersion-managed soliton. The advantages of the DM-soliton over linear transmission are features like the large power which gives a high signal-to-noise ratio. On the other hand, the difference between linear and nonlinear pulses is becoming increasingly fuzzy, and perhaps the distinction should be made between return-to-zero (RZ) and NRZ modulation rather than between linear and nonlinear transmission.

It is nevertheless interesting to note that solitons are now not only considered for oceanic systems but also for shorter terrestrial systems. There are still several challenges and opportunities remaining in order to take full advantage of solitons and to reach a better understanding. Further work is also needed on WDM soliton- and very high-speed TDM soliton-systems. The use of DM-solitons is a very recently established technique and there are thus many issues to consider. These include studies of robustness to deviations of optimum conditions, e.g., improper pulse launch conditions, the impact of nonperiodic dispersion maps and of PMD (both of which are difficult to study in loop experiments), and intrachannel effects. Nevertheless, solitons have now reached a level of maturity such that commercialization seems very near.

List of Units and Nomenclature

Amplifier gain	G	$[-]$
Amplifier separation	L_a	[km]
Average dispersion	β_0''	$[\mathrm{ps}^2\ \mathrm{km}^{-1}]$
Dispersive length	L_D	[km]
Fiber loss coefficient	α	$[\mathrm{km}^{-1}]$
Fiber nonlinear coefficient	γ	$[\mathrm{W}^{-1}\ \mathrm{km} - 1]$
Group velocity dispersion (GVD)	β_0''	$[\mathrm{ps}^2\ \mathrm{km}^{-1}]$

Map strength	S	[−]
Optical field amplitude	$u(z, t)$	[$\mathrm{W}^{1/2}$]
Pulse duration	t_0	[ps]
Soliton energy	E_s	[J]
Soliton peak power	P_s	[W]
Third-order dispersion	β_0'''	[$\mathrm{ps}^3\ \mathrm{km}^{-1}$]
Time	t	[ps]
Transmission distance	z	[km]

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics). **Solitons:** Temporal Solitons.

Further Reading

Agarwal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. San Diego: Academic Press.

Agarwal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego: Academic Press.

Agarwal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.

Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. Oxford: Clarendon Press.

Iannone E, Matera F, Mecozzi A and Settembre M (1998) *Nonlinear Optical Communication Networks*. New York: Wiley.

Mollenauer LF, Gordon JP and Mamyshev PV (1997) Solitons in high bit-rate long-distance transmission. In: Kaminow IP and Koch TL (eds). *Optical Fiber Telecommunications III A*. San Diego: Academic Press.

Smith NJ and Doran NJ (1995) Evaluating the capacity of phase modular-controlled long-haul soliton transmission. *Optical Fiber Technology* 1: 218–235.

Suzuki K and Nakazawa M (1995) Recent progress in optical soliton communication. *Optical Fiber Technology* 1: 289–308.

Taylor JR (ed) (1992) *Optical Solitons–Theory and Experiment*. Cambridge: Cambridge University Press.

The final report of the ACTS-MIDAS project, Jan 1999, can be downloaded at *www.infowin.org/ACTS/RUS/PROJECTS/FINAL-REPORTS/fr-053.pdf*.

Turitsyn SK, Mezentsev VK and Shapiro EG (1998) Dispersion-managed solitons and optimization of the dispersion management. *Optical Fiber Technology* 4: 384–452.

Temporal Solitons

S Trillo, University of Ferrara, Ferrara, Italy

A Tonello, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fact that wave propagation phenomena are intrinsically dispersive, that is localized disturbances tend to spread, is part of our experience. Less well known is the fact that the spreading of wavepackets can be counteracted by the nonlinear response of the host medium. Solitons are wavepackets that exploit this principle, realizing a perfect balance of dispersion and nonlinearity that allows them to self-trap and travel undistorded, or at least to recover their shape periodically, for long (virtually unlimited) distances. More specifically, here we deal with optical temporal solitons constituted by envelopes, or pulses with relatively narrow bandwidths around a carrier optical frequency, whose temporal profile remains unchanged (either strictly or on average) upon propagation. In contrast with spatial solitons predicted earlier and such that nonlinearities are exploited to compensate for diffractive beam spreading even in the two transverse dimensions (i.e., in bulk samples, for details the reader is referred to **Solitons:** Bright Spatial Solitons), temporal soliton trapping is intrinsically a one-dimensional phenomenon typically observed in optical fibers where diffraction is compensated by the standard fiber guidance mechanism. Temporal solitons were first predicted and observed in single-mode fibers (SMFs). Their applications in digital optical communications have increased due to their capability to behave as nearly ideal bits. Meanwhile, however, the concept of temporal soliton has been considerably generalized by investigating new settings and effects such as polarization, parametric amplification (*see* **Nonlinear Optics, Basics:** $\chi^{(2)}$-Harmonic Generation), gratings (*see* **Fiber Gratings**), stimulated scattering (*see* **Scattering:** Scattering Phenomena in Optical Fibers), and fiber lasers (*see* **Lasers**: Optical Fiber Lasers), etc. Here we briefly review the essential features of these different families of temporal solitons. In particular, in the section on Fiber Temporal Solitons below, we introduce the basic model for fiber solitons, whereas we discuss their application to optical communications in the section on Solitons in Optical Communications below. In the sections on Multi-component Temporal Solitons, and Temporal Solitons in Bragg Gratings below, we introduce significant generalizations of the concept involving multicomponent and slow-light grating solitons, respectively. Finally, in the section on Solitons on Dissipative nonlinear optics we discuss issues related to solitons in strongly dissipative systems. Note that, throughout the paper, the word soliton is used in the aforementioned general sense, which does not necessarily require the underlying model to be integrable (*see* the section on Fiber Temporal Solitons below,), as implied by the more restrictive and rigorous mathematical meaning of soliton. We further focus our discussion to bell-shaped solitons (for dark solitons, holes on a continuous-wave background, *see* **Solitons:** Optical Fiber Solitons, Physical Origin and Properties), which can be excited by means of standard laser pulses.

Fiber Temporal Solitons

An optical pulse is described by the complex electric-field envelope $E(z,t)$ that modulates the carrier frequency ω_0. Its evolution along the z axis of a silica SMF is described with good approximation by the following universal model for dispersive and weakly nonlinear systems, known as the nonlinear Schrödinger equation (NLSE),

$$\mathrm{i}\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \gamma |E|^2 E = 0 \qquad [1]$$

In eqn [1] the second term is responsible for dispersive spreading of pulses ruled by the group-velocity dispersion

$$k' = \left.\frac{\mathrm{d}^2 k}{\mathrm{d}\omega}\right|_{\omega_0}$$

(GVD, known also as chromatic or second-order dispersion), while the first-order dispersion

$$k' = \left.\frac{\mathrm{d}k}{\mathrm{d}\omega}\right|_{\omega_0}$$

has been removed in eqn [1] by introducing the retarded time $t = T - k'z$ measured in the pulse frame traveling at the natural group-velocity $V \equiv k'^{-1}$ of light, T being the physical time in the laboratory frame. The cubic term in eqn [1] is responsible for self-phase modulation (SPM) that stems from the instantaneous optical Kerr effect characterized by the effective coefficient $\gamma = k_0 n_{2I}/A_{\mathrm{eff}} = (\omega_0/cA_{\mathrm{eff}})n_{2I}$, where n_{2I} is the nonlinear index of silica, i.e., the nonlinear refractive index change divided by the intensity which has produced it. The real-world total electric field can be recovered as $\mathcal{E}(z,r,t) = \mathrm{Re}[f(r)E(z,t)\exp(\mathrm{i}k_0 z - \mathrm{i}\omega_0 T)]$, where the transverse modal profile $f(r)$ has been averaged out in eqn [1] to yield an effective fiber area A_{eff}.

The mechanism behind soliton formation can be understood as due to the instantaneous frequency changes along the pulse caused by the Kerr SPM, compensating those due to the group velocity dispersion (GVD) in the anomalous case ($k'' < 0$, at carrier wavelengths longer than 1.3 μm in standard SMFs) where the GVD induces a blue-shift on the pulse leading edge and a red-shift on the pulse trailing edge (in the absence of such compensation the pulse broadens because anomalous GVD means that blue-shifted frequencies travel faster than red-shifted frequencies). Such mechanism of compensation becomes ideal in the case of a soliton waveform. This concept finds its mathematical support in the fact that eqn [1] is a conservative (Hamiltonian) model with the remarkable property to possess infinite conserved quantities, a rather exceptional feature which makes it exactly integrable. Among the solutions, the following fundamental (so-called $N = 1$, see below) bright temporal soliton exists whenever $k'' < 0$:

$$\begin{aligned} E(z,t) = {} & \frac{1}{\sqrt{\gamma L_{\mathrm{d}}}}\,\mathrm{sech}[\eta(t - \mathrm{d}k'z - t_0)] \\ & \times \exp\left[-\,\mathrm{i}\delta\omega t + \frac{\mathrm{i}}{2}\left(\frac{1}{L_{\mathrm{d}}} + k''\delta\omega^2\right)z + \mathrm{i}\phi_0\right] \end{aligned} \qquad [2]$$

where t_0 and ϕ_0 are arbitrary initial position and phase, and $L_{\mathrm{d}} = (k''|\eta^2)^{-1}$ is the characteristic distance after which the GVD causes a significant pulse broadening, the so-called dispersion length. Temporal solitons [2] are characterized by two independent parameters, namely the duration $1/\eta$ and the frequency detuning $\delta\omega$, and possess several remarkable features:

(i) they travel with invariant shape by realizing the condition $N \equiv L_{\mathrm{d}}/L_{\mathrm{nl}} = 1$, i.e., exact balance between the two length scales associated with eqn [1], namely L_{d} and the nonlinear length $L_{\mathrm{nl}} = (\gamma P_1)^{-1}$ associated with the peak power $P_1 = (\gamma L_{\mathrm{d}})^{-1} = |k''|\eta^2/\gamma$ needed for the input pulse [2] to be a soliton;
(ii) in any given fiber with fixed GVD k'' and nonlinearity γ, a larger soliton power P_1 is required for a shorter duration, i.e., narrower solitons are taller;
(iii) solitons experience a phase retardation as a whole (particle), which increases with P_1, i.e., taller solitons have larger shifts;

(iv) as a consequence of Galileian invariance, a frequency shift $\delta\omega$ from ω_0 induces the soliton velocity to change from k'^{-1} to $(k' + dk')^{-1}$, with $dk' = dk''\delta\omega$;

(v) unlike other waves in nonlinear media, collisions between solitons with different velocities are strictly elastic, as shown in the example of Figure 1a; and

(vi) solitons are extremely robust: for instance they can be excited with input peak powers differing from P_1, whereas a weak attenuation α $(\alpha L_d \ll 1)$ induces the soliton to reshape adiabatically, i.e., the peak power decreases, yet the pulse remains locally a soliton by adjusting its width, as displayed in Figure 1b.

More general (and cumbersome) solutions of eqn [1] describe periodic soliton evolutions, such as the interaction of in-phase adjacent input pulses at the same frequency (see Figure 1c), or a single input pulse with sech-shape but peak powers corresponding to higher integer values of N, so-called breathers because of their ability to recover periodically the initial shape after 'breathing' (see Figure 1d for $N = 3$) over the distance $\pi L_d/2$.

Solitons [2] describe stable propagation-invariant pulses under ideal conditions (lossless case, relatively long pulses and short SMFs). In practical settings, the importance of solitons stems from their robustness against real-world perturbing phenomena described by terms that in turn break the integrability of the underlying model [1]. Situations of this kind arise, for instance, when one accounts for stimulated Raman scattering whose effect is to down-shift continuously the central frequency of ultra-short solitons, or in optical communication systems treated below.

Solitons in Optical Communications

Solitons [2] represent ideal bits for RZ (return-to-zero) optical data sequences. However, fibers have unavoidable losses of about 0.2 dB/km at the usual operating wavelength $\lambda_0 = 1550$ nm. To ensure long-haul connections, the optical signal must be periodically amplified, commonly using erbium-doped fiber amplifiers (EDFAs). The pulse peak power hence

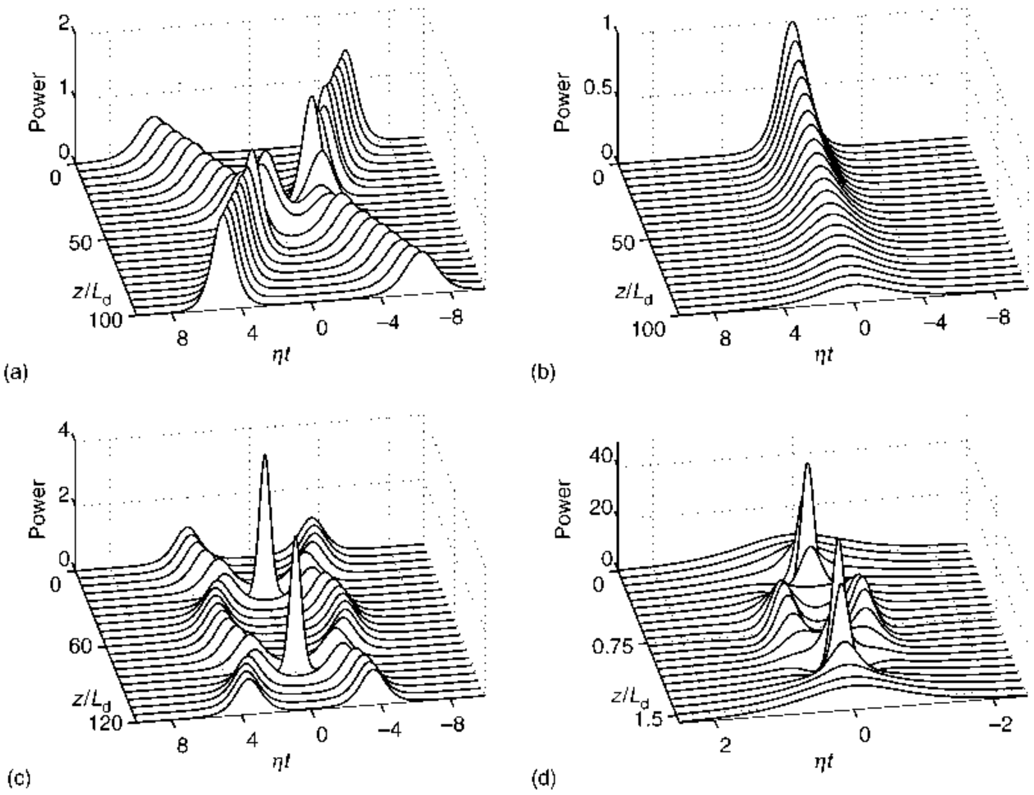

Figure 1 Fundamental features of fiber (NLSE) solitons: (a) elastic collision of two $N = 1$ solitons with different values of parameter pairs (frequency–velocity and amplitude–width); (b) adiabatic soliton damping in the presence of losses; (c) two-soliton bound state (in-phase interacting solitons of equal amplitudes); (d) evolution of a $N = 3$ soliton breather along one period of its evolution. Here the power $|E|^2$ is scaled in units of fundamental soliton power P_1, while distance z and time t are reported in units of L_d and initial pulse-width $1/\eta$, respectively.

undergoes both large exponential decay, because of distributed losses and abrupt lumped amplification under the action of EDFAs. Therefore, the stable local balance between GVD and nonlinearity cannot be rigorously sustained in the fiber. If both losses and amplifiers are accounted for, the mathematical model for pulse envelope propagation along M spans becomes:

$$\mathrm{i}\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \chi|E|^2 E = -\,\mathrm{i}\alpha E + \mathrm{i}[\exp(\alpha Z_\mathrm{A}) - 1] \times \sum_{n=1}^{M} \delta(z - nZ_\mathrm{A})E(nZ_\mathrm{A}, t) \qquad [3]$$

where the RHS contains the perturbing terms arising form distributed losses (α) and lumped amplifications (described by the comb with spacing Z_A).

The analysis of nonintegrable systems, such as eqn [3], can be carried out by appropriate transformations that allow the recovery of a reduced integrable model. The key result is that, under the limitations set by an appropriate length scale, the solitons of the reduced system can be close to those of the original problem. For temporal solitons, such transformation was first called ‘the guiding center soliton’, envisaging a similarity with the guiding center motion of a charged particle in a magnetic field. The main outcome is that eqn [3] can be reduced to the standard NLSE [1], clearly showing that soliton-like propagation still occurs in the presence of losses (even large, $\alpha L_\mathrm{d} \gg 1$) and rapid periodic amplification. The appropriate scale for these systems is the amplifier spacing Z_A, that must be much shorter than L_d. To reproduce an average behavior similar to the lossless case, it is sufficient to multiply the input condition for an enhancement factor that counteracts the effect of fiber losses. Therefore soliton-like solutions do exist even in fiber systems with loss management, though on average (over several fiber spans) rather than locally.

Another issue that makes solitons suitable as optical bits, lies in their intrinsic aptness to be controlled either through passive or through active systems. Solitons tend to follow their relative center of mass in the temporal or spectral domain. For instance, a powerful technique to overcome one of the major problems, i.e., the growth of noise due to amplified spontaneous emission, may be obtained by deploying periodically along the line optical passband filters. Mollenauer and co-workers have demonstrated that, by sliding the filters progressively (i.e., changing slightly the central peak from filter to filter), one can guide a shape-invariant soliton which adapts its carrier frequency while effectively suppressing the noise. Similar control techniques involve amplitude or phase modulators and a consistent amount of work has focused on the problem of all optical soliton regeneration.

To exploit the idea of soliton in optical communications, however, we should consider the fact that ultra-short optical pulses, suitable for high transmission rates, require extremely low fiber GVD to accomplish the stable balance between dispersion and nonlinearity for a fixed pulse-width and available power. For instance, in order to transmit at 40 Gbit/s, we should achieve a fixed GVD coefficient $D = -(2\pi c/\lambda_0^2)k'' \simeq 0.25\ \mathrm{ps/(km\ nm)}$ for the whole link.

On the other hand, in many optical systems the fiber nonlinearity is not exploited at all but rather considered as a perturbing effect and a source of penalty that limits the maximum transmission distance. In such an approach the desired cumulated GVD should be as close as possible to zero to avoid errors coming from inter-symbol interference caused by dispersive broadening. Dispersion-shifted fibers may ensure low GVD only in a narrow spectral region, thus limiting the transmission bandwidth. Conversely GVD can be reduced to zero on average, simply combining fiber spans with opposite GVD. The periodic alternance of fiber types may also carry several benefits, e.g., limiting the impact of resonant nonlinear interactions such as four-wave mixing (*see* **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics) for its definition). The different architectures of dispersion management proposed with the twofold scope of limiting nonlinear effects and reducing the cumulated GVD, have immediately stimulated the investigation of their impact on soliton propagation. In the early schemes, the attention was focused on the so-called adiabatic dispersion management, where the GVD was changed proportionally to the soliton power to improve soliton system performances. In contrast with previous works, nonadiabatic map profiles were introduced in the experiments by Suzuki and co-workers to compensate the cumulated dispersion by locating periodically dispersion compensating fibers (DCFs). In net contrast, Doran and co-workers have suggested the use of a combination of fibers with alternate normal and anomalous GVD (i.e., a periodic stepwise variation of GVD), predicting the existence of solitons with a peak power larger than the equivalent classical soliton power, i.e., dispersion-managed solitons (DMSs) The periodic stepwise GVD map is not only conceptually different from earlier designs (with exponentially decaying GVD), but has opened the field to real application due to its simplicity of implementation. The innovative discovery that special pulse waveforms may propagate in periodically dispersion managed fiber links has triggered the interest of several groups, leading to a complete characterization of the problem, both theoretically and experimentally.

Similar to the case of guiding center solitons, a DMS usually shows a solitonic behavior. However, the DMS has a Gaussian shape surrounded by oscillating tails (see Figure 8 in **Solitons:** Optical Fiber Solitons, Physical Origin and Properties), thus being more suitable, due to its faster decaying tails, to be closely packed in a bit sequence in comparison with a sech-soliton (eqn [2]) of the same FWHM. We show an example of DMS transmission line in the top frame of Figure 2, where we sketched the variation of the GVD of a standard SMF periodically balanced by shorter segments of DCFs with opposite GVD. We report, in the middle frame, the DMS pulsewidth and chirp and in the bottom frame the pulse peak power. As shown, the macroscopic soliton parameters are characterized by large fluctuations: the local minima of pulsewidth correspond to chirp-free points and at such points one observes either a minimum or a maximum of the pulse spectral bandwidth. Nonetheless, all these pulse parameters have periodical evolutions that are fingerprints of an invariant and soliton-like pulse when observed at integer multiples of Z_A. The exact periodicity preserves the pulse waveform and hence the carried information. In DMS transmission we assist at the threefold compensation of energy, GVD and nonlinearity. Similarly to the guiding center solitons, the threefold compensation has an average, rather than pointwise, connotation. However, unlike guiding center solitons, for DMSs the definition of average should take into account the different relevance of pulse peak power in the map. Consequently the nonlinearity is compensated in a weighted manner, and it is not surprising in this context (though astonishing if placed back in the context of the NLSE [1] which requires anomalous GVD), that DMSs have been observed even when the uniform average GVD is weakly normal.

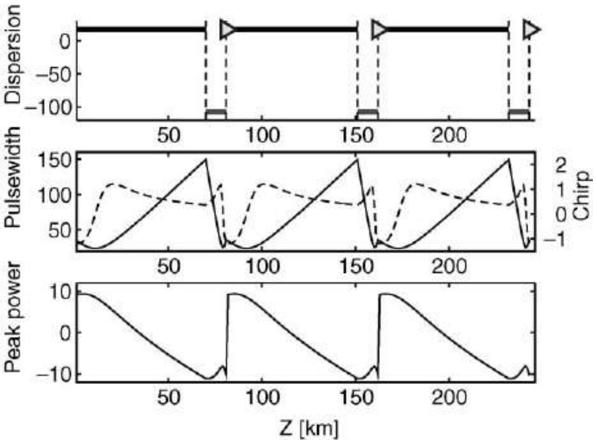

Figure 2 Evolution of soliton parameters along three periods of a fiber link with lumped amplifiers and dispersion management. Top frame: GVD map D(z) [ps/(nm km)] and EDFAs (triangles) location; Middle frame: soliton chirp (in units of 1000 ps^{-2}, dashed line), and pulse-width (in ps, solid line) vs. distance. Bottom frame: DMS peak power (dBm) vs. distance.

The net difference between the local and average GVD allows simultaneous control of different detrimental phenomena: for instance four-wave mixing is reduced by the large local GVD, whereas a small average GVD limits the impact of timing jitter due to amplified spontaneous emission (Gordon–Haus effect). As a result, DMSs perform better than conventional solitons.

Figure 3 shows a typical high-speed (20 Gb/s) single-channel experiment, and the achieved performance in terms of quality factor Q versus propagation distance, with the output after 8100 km. We recall that the Q factor measures the bit-error-rate (an error rate lower than 10^{-9} corresponds to a Q higher than $\sim$ 16 dB). DMSs also allow the control of collisions between solitons that pertain to different channels (we recall that solitons at different carrier frequencies go at different velocities, and hence collide), thus enhancing performances of standard or dense WDM systems.

Multi-Component Temporal Solitons

So far we have discussed temporal solitons described by a single envelope. However, the same mechanism of mutual balance between nonlinearity and dispersion can also be effective for multiple pulses, leading to trap simultaneously multicomponent (or so-called vector) solitons. Solitons of this type are, for instance, those due to the interaction of polarization modes in isotropic or birefringent Kerr media, higher-order modes in large-core (multimode) fibers, or mixing of pulses at different carrier frequencies that interact parametrically (i.e., without net exchange of energy with the medium). The latter case entails a broad class of solitons since virtually any parametric four-photon mixing process in cubic media (e.g., nondegenerate or partially degenerate four-wave mixing, third harmonic generation, ...), or three-photon mixing process in quadratic media (sum- and difference-frequency generation, second-harmonic generation (SHG), optical rectification, etc.) is allowed to sustain them. This opens up the field to an entirely new class of materials, namely noncentrosymmetric (quadratic) materials which usually have stronger nonlinear responses if compared with silica fibers. SHG can be taken as a paradigm to

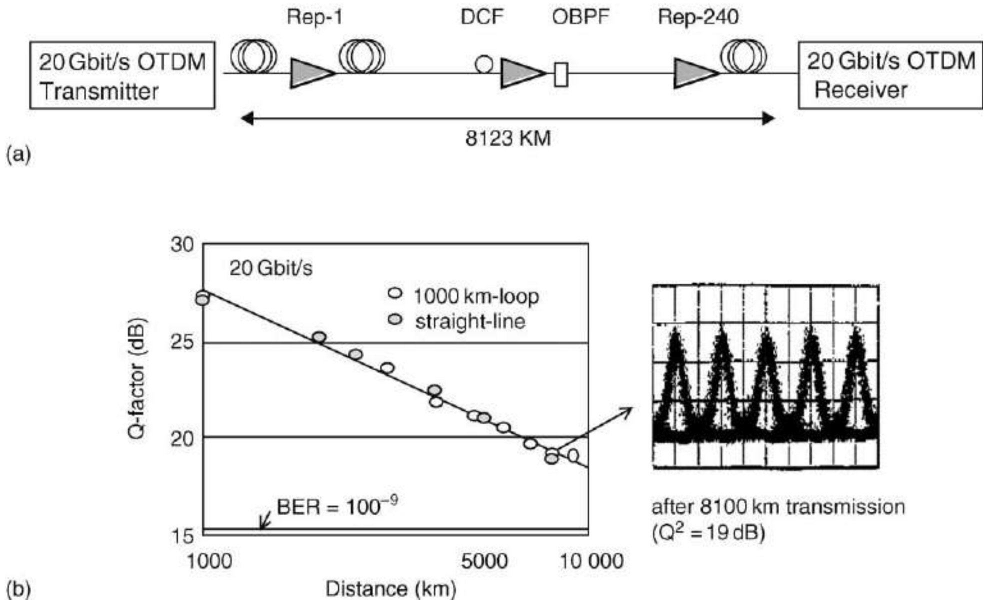

Figure 3 (a) Schematic diagram of the straight-line experiment with optical time-division multiplexing (OTDM) of soliton at 20 Gbit/s; (b) Measured *Q* factor vs. distance comparing experimental results obtained in the straight-line and loop configuration (loop length of 1000 km), respectively. The waveform received after 8100 km is also shown. With permission from Suzuki M and Edagawa N (2003) Dispersion-managed high-capacity ultra-long-haul transmission. *Journal of Lightwave Technology* 21: 916–929. © 2003 IEEE.

understand the main features of these multicomponent solitons. In this case, the evolution of the interacting envelopes E_j at carrier frequencies $j\omega_0$, $j = 1, 2$, is ruled by the coupled equations:

$$\left(\mathrm{i}\frac{\partial}{\partial z} - \frac{k_1''}{2}\frac{\partial^2}{\partial t^2}\right)E_1 + \chi E_2 E_1^* \exp(\mathrm{i}\Delta k z) = 0,$$

$$\left(\mathrm{i}\frac{\partial}{\partial z} + \mathrm{i}\delta k'\frac{\partial}{\partial t} - \frac{k_2''}{2}\frac{\partial^2}{\partial t^2}\right)E_2 + \chi E_1^2 \exp(-\mathrm{i}\Delta k z) = 0, \quad [4]$$

where the subscript $j = 1, 2$ indicates that the relative quantities evaluated at frequency $j\omega_0$, $t = T - k'_1 z$, $k''_{1,2}$ are GVD coefficients, $\delta k' \equiv k'_2 - k'_1$ is the group-delay, χ is the effective nonlinear coefficient, and $\Delta k = k_2 - 2k_1$ is the wave-number (or phase) mismatch. In spite of the fact that eqns [4] are not integrable, their stable soliton solutions can be found (numerically) in a wide range of parameters as symbiotic phase-locked pulses $E_{1,2}(z,t)$ that travel undistorted. For instance, SHG solitons exist either in phase-matching ($\Delta k = 0$) or far-off phase-matching (with some limitations for $\Delta k > 0$), and, unlike fiber solitons, also in normally dispersive media ($k''_{1,2} > 0$). As for fiber solitons, their width is related to their peak power, since the basic mechanism is the mutual balance between the dispersive broadening (that depend on the pulsewidth) and nonlinear phase-shifts (that depend on power), which occur due to repeated up-conversion and down-conversion processes, often referred to as cascading. A SHG soliton is robust enough to permit the observation of trapping, even from an input that differs considerably from the soliton (e.g., from the fundamental only), or in the presence of a relatively strong group-delay which tends to rip apart the two-component soliton, inducing walk-off between the two envelopes.

In contrast with fibers, in this case, experiments are necessarily made in relatively short bulk samples, thus resulting in a limited number of dispersion lengths. Usually broad (quasi-planewave) beams are conveniently employed in order to make diffraction ineffective.

The main experimental problem is that, in crystals commonly employed for SHG, the dynamics of typical laser pulses (duration above 100 fs) is fully dominated by the group-delay $\delta k'$. In other words, the material GVD becomes relevant only for pulses as short as a few fs. Therefore, the observation of solitons with longer pulses has been required to tailor the dispersion, which can be accomplished by means of pulses with tilted phase front. As displayed in Figure 4, successful compression of 200 fs (FWHM) tilted pulses was observed at relatively large negative Δk in a BBO bulk crystal with effective zero group-delay and enhanced (anomalous) GVD.

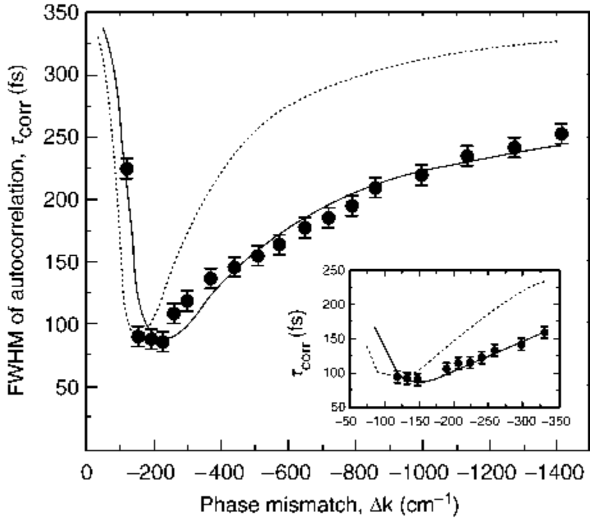

Figure 4 Soliton fomation via SHG: the circles show the output pulse duration (full width at half maximum (FWHM) of autocorrelation τ_{corr}) vs. mismatch Δk, measured after propagation of an input 200 fs tilted pulse (input $\tau_{\text{corr}} = 284$ fs) with 13.4 GW/cm^2 peak intensity, in a 7 mm BBO crystal (inset: same with 10.4 GW/cm^2). Superimposed (dotted line) is the corresponding theoretical value obtained from eqn [4]. The solid curve that fits better the data accounts for the additional effect of the material Kerr effect. With permission from Di Trapani P, Caironi D, Valiulis G, Dubietis A, Danielius R and Piskarskas A (1998) Observation of temporal solitons in second-harmonic generation with tilted pulses. *Physical Review Letters* 81: 570–573. © 1998 APS.

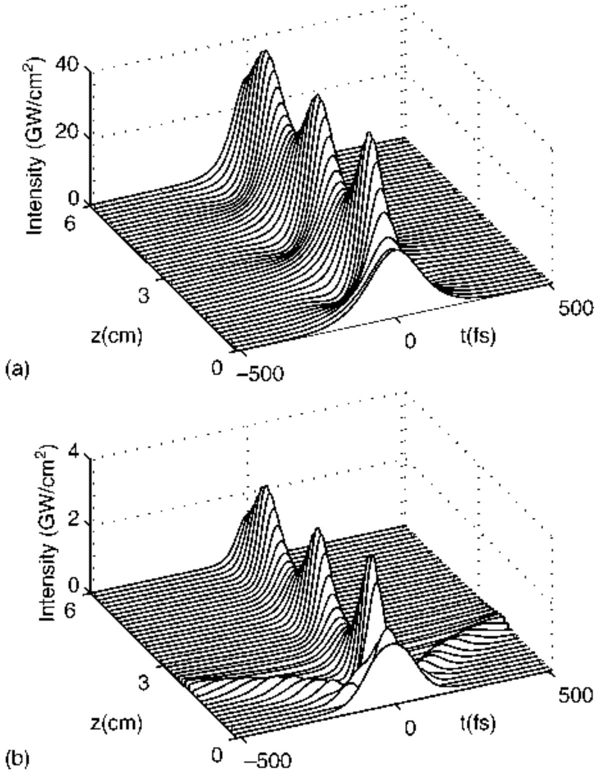

Figure 5 Dynamics of temporal soliton formation ruled by eqn [4]. A compressed soliton is formed via strongly mismatched ($\Delta k = -100\ \text{cm}^{-1}$) SHG from a 200 fs (FWHM) input pulse at fundamental carrier frequency, in a crystal with $\chi = 1.4 \times 10^{-4}\ \text{W}^{-1/2}$, effective dispersion $\delta k' = 0$ (group–velocity matching) and $k_1'' = -1.8\ \text{ps}^2/\text{m}$, $k_2'' = -0.75\ \text{ps}^2/\text{m}$.

The measured compression is a signature of soliton formation, as clearly indicated by the example of simulated pulse dynamics reported in Figure 5: the pulse at second-harmonic is generated and then trapped with the input one as a two-component bound state. Radiation and oscillations occur, in this case, because the launching conditions (single envelope) are far enough from the two-component soliton.

Temporal Solitons in Bragg Gratings

A linear periodic structure, such as a Bragg grating with pitch Λ, couples counter-propagating waves when their frequency lies close to the Bragg frequency ω_{B} (vacuum wavelength $\lambda_{\text{B}} = 2n_0\Lambda$) defined by the resonance condition $k(\omega_{\text{B}}) = \pi/\Lambda$. Such a type of structure supports a different type of temporal solitons, so-called gap solitons (GSs) since they originate from a linear dispersion relationship of the type shown in Figure 6a, characterized by a stop-band (gap) centered around ωB. The gap is characterized by a unitary grating reflectivity, as shown in Figure 6b, and the forward linear propagation is therefore forbidden. However, if the incoming wave is intense enough to induce the refractive index to increase through the optical Kerr effect, the Bragg resonance (and the whole gap) is shifted downwards (in frequency) allowing for self-transparency at the operating frequency. GSs can be understood as bell-shaped envelopes which are allowed to travel along the grating due to a high-intensity core that bleaches the reflectivity, while the high reflectivity seen by the tails takes the pulse together. GSs are expected to be slow because the slope of the dispersion curve tends to zero when approaching the gap, while its curvature results in the grating GVD (huge if compared with material GVD in real-world units), which is exactly balanced by the nonlinearity when the GS is formed. This naive portrait is supported by a more thorough analysis starting from the coupled-mode equations obeyed by the forward $E_{+}(z, T)$ and backward $E_{-}(z, T)$ envelopes at frequency ω_{B}, propagating along a (e.g., fiber) Bragg grating with Kerr effect (inducing SPM and

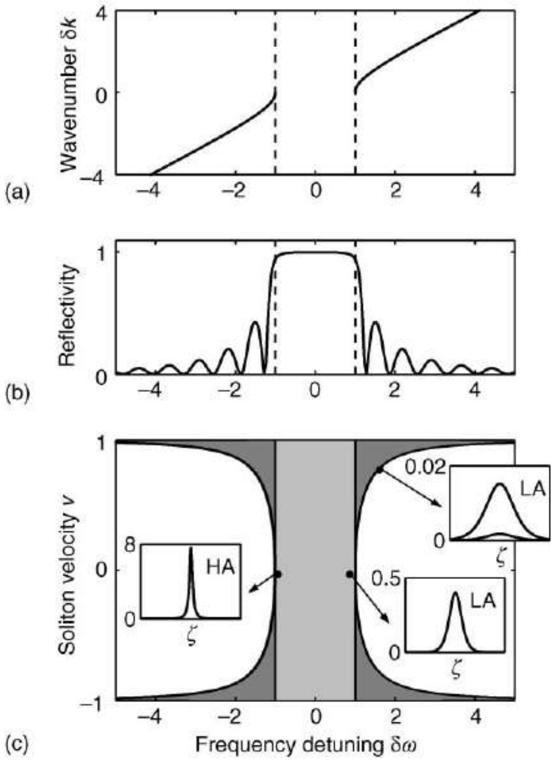

Figure 6 (a) Bragg linear dispersion relationship in normalized (dimensionless) units: wave-number shift $\delta k = (k(\omega) - \pi/\Lambda)/\Gamma$ vs. frequency detuning $\delta\omega = k'(\omega - \omega_B)/\Gamma$; (b) linear grating reflectivity vs. $\delta\omega$ for a grating figure of merit $\Gamma L = 4$; (c) existence domain (whole shaded area) of GSs in the normalized detuning–velocity plane ($\delta\omega, v \equiv k V_{sol}$). The light shaded area corresponds to the stop-band $|\delta\omega| < 1$ (inner region between dashed lines in a–b). The insets show samples of GS intensity profiles vs. the moving frame coordinate $\zeta = (1 - v^2)(z - V_{sol} T)$. Zero-velocity GSs have symmetric envelopes ($|E_+| = |E_-|$) and range from high-amplitude (HA) to low-amplitude (LA) across the gap. The third inset shows a moving GS in the experimentally accessible region, where it retains LA but has a stronger component in the direction of motion.

cross-phase modulation):

$$\mathrm{i}\left(k'\frac{\partial E_\pm}{\partial T} \pm \frac{\partial E_\pm}{\partial z}\right) + \Gamma E_\mp + \gamma\left(|E_\pm|^2 + 2|E_\mp|^2\right)E_\pm = 0 \qquad [5]$$

The linear dispersion associated with eqn [5] is as in Figure 6a, with a stop-band $|\omega - \omega_B| \le \Gamma/k'$ proportional to the index corrugation depth through the Bragg coupling coefficient Γ. In the nonlinear regime, GS solutions of eqn [5] differ qualitatively from conventional fiber solitons [2]:

(i) first, GSs are indeed slow since the two envelopes can travel locked with any soliton velocity V_{sol} such that $|v| \equiv |k' V_{sol}| \le 1$, i.e., lower than the natural group velocity k'^{-1};

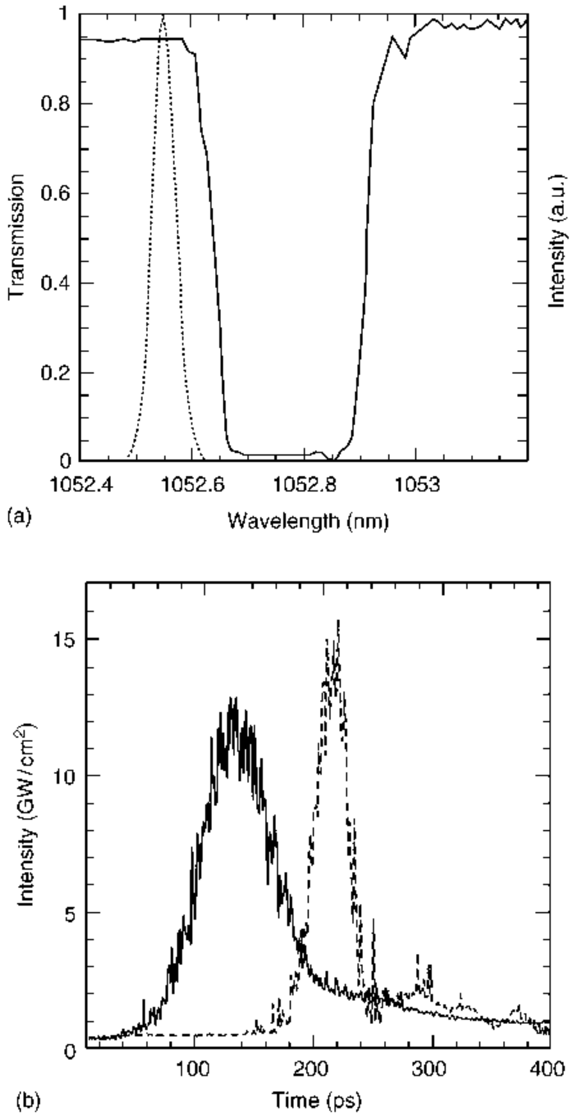

Figure 7 Gap solitons in a fiber Bragg grating: (a) measured pulse spectrum compared with the linear grating transmission, (b) compressed (15 ps) and delayed pulse transmitted at high intensity (dashed line), compared with the pulse transmitted at frequency far from the Bragg resonance (solid line) where it retains its input width of 80 ps. With permission from Eggleton BJ, Slusher RE, de Sterke CM, Krug PA and Sipe JE (1996) Bragg grating solitons. *Physical Review Letters* 76: 1627–1630. ©1996 APS.

(ii) GSs exist only in a narrow bandwidth, i.e., the shaded area in Figure 6c, which corresponds to the gap seen in the soliton frame (it reduces to the conventional stop-band for still solitons and gets wider for faster GSs); and

(iii) the gap is strongly asymmetric since lower intensity is required to induce transparency close to the high-frequency band-edge where, as a consequence, GSs have low amplitudes and are stable (the amplitude grows moving towards the low-frequency bandedge, where GSs become unstable).

Experimentally, GSs have been typically observed with pulses at frequency close to the blue bandedge of fibers (see Figure 7a) and AlGaAs waveguides with built-in Bragg gratings of a few centimeters. The typical signature of GSs, illustrated in Figure 7b, is the enhanced pulse transmission observed under the operating conditions of Figure 7a when the input intensity is sufficiently high. The transmitted pulse experiences pulse compression and a significant group-delay (measured with respect to the transit-time at frequency far off the stop-band), witnessing its nature of slow-light solitons.

Finally, we emphasize that grating solitons can also be supported in the case where the two modal envelopes propagate in the forward direction along a grating whose period Λ is long enough to be resonant with the modal phase slippage, thus inducing efficient linear coupling between them. In this case, the structure has a gap in wave-number, and solitons are slow in the frame traveling at the average modal group-velocity. Soliton formation is expected to lead to grating-induced compensation of the group-delay between the modes. Experiments on such types of structures have been carried out only recently.

Solitons in Dissipative Nonlinear Optics

Temporal solitons can also be observed in nonlinear optical interactions that involve changes of the state (energy level) of matter as in inverted media or scattering phenomena. For instance, early studies in the 1960s have demonstrated that coherent pulses with energy above a critical level can pass through resonant (two-level) atoms without being absorbed, a phenomenon denoted as self-induced transparency. Such pulses are temporal solitons solutions of the Maxwell–Bloch model which describes the atom–light interaction.

Also scattering phenomena occurring in optical fibers, such as Raman and Brillouin scattering, possess three-wave solitons, in which the pump and scattered (Stokes) light waves are coupled to a phonon field of the acoustic or optical branch (for Brillouin or Raman, respectively).

Perhaps more important from the point of view of the modern applications, are fiber lasers composed by spans of active and/or passive fiber closed to form a recirculating loop. These types of structures can often be described by distributed (averaged over several round-trips) models, which take the form of a dissipative perturbed NLSE, known as the Ginzburg–Landau equation (GLE):

$$\mathrm{i}\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \chi|E|^2E = \mathrm{i}gE + \mathrm{i}g_{\mathrm{nl}}|E|^2E + \mathrm{i}q|E|^4E + \frac{\mathrm{i}}{\Omega_{\mathrm{f}}^2}\frac{\partial^2 E}{\partial t^2} \qquad [6]$$

where g is the excess (net) gain ($g > 0$) or loss ($g < 0$), g_{nl} and q describe the nonlinear gain mechanism and its saturation (usually introduced to control the mode-locking), and Ω_{f} is linked to spectral filtering. Several soliton (pulse) solutions of the GLE [6] can be found even when the parameters (g, g_{nl}, q, Ω_{f}) are such as to constitute a strong perturbation of the NLSE. Unlike NLSE solitons [2], pulse solutions of the GLE [6] are usually chirped (i.e., possess a nonlinear phase profile), and have fixed amplitude for a given value of the parameters. The set of soliton waveforms is generally richer and includes, for instance, flat-top stationary pulses, two-pulse bound-states, solitons moving at fixed velocity, etc. Although instabilities of the background (for $g > 0$) or of the pulse itself can affect their observability, large islands of stability in the parameter space can be found. The relevant role of stable solitons in mode-locking and intra-cavity pulse formation mechanisms has been demonstrated in many different fiber laser geometries, for which we refer the reader to the more specialized papers listed as Further Reading at the end of this article. We point out that some schemes, for instance the stretched pulse-fiber laser developed at MIT, involve a basic configuration (alternating GVD within the cavity length) which is quite similar to that seen in the framework of DMSs, thus exhibiting the same basic physical mechanism (strong compression and stretching of the pulse along each round-trip which play the role of the map period for a DMS).

List of Units and Nomenclature

BBO	β-barium borate
DCF	dispersion compensating fiber
DMS	disperion managed soliton
EDFA	Erbium doped fiber amplifier
GLE	Ginzburg–Landau equation
GS	gap soliton
GVD	group-velocity dispersion
NLSE	nonlinear Schrödinger equation
OTDM	optical time-division multiplexing
SHG	second-harmonic generation

SPM	self-phase modulation
WDM	wavelength division multiplexing

Amplifier spacing L_{A} [m]
Bragg frequency ω_{B} [Hz]
Bragg period Λ [m]
Central wavelength λ_0 [nm]
Coupling coefficient Γ [m^{-1}]
Dispersion coefficient D [$\mathrm{ps\ nm^{-1}\ km^{-1}}$]
Dispersion length L_{d} [m]
Effective area A_{eff} [m^2]
First-order dispersion k' [s/m]
Fundamental soliton power P_1 [W]
Gain (linear) g [m^{-1}]
Gain (nonlinear) g_{nl} [m^{-1}]
Group-delay δ [s/m]
Group-velocity $V = k'^{-1}$ [m/s]
Group velocity dispersion k'' [$\mathrm{ps^2\ m^{-1}}$]
Inverse pulse-width η [s^{-1}]
Light velocity in vacuum c [m/s]
Loss coefficient α [m^{-1}]
Nonlinear coefficient (fiber) γ [$\mathrm{m^{-1}\ W^{-1}}$]
Nonlinear coefficient (quadratic bulk materials) χ [$\mathrm{W}^{-1/2}$]
Nonlinear refractive index n_{2I} [$\mathrm{m^2\ W^{-1}}$]
Nonlinear length L_{nl} [m]
Refractive index (linear) n_0
Retarded time (in the moving frame) t [s]
Spectral filtering coefficient Ω_{f} [$\mathrm{Hz\ m^{1/2}}$]
Time in the laboratory frame T [s]
Wave-number mismatch Δk [m^{-1}]

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics); Nonlinear Optics. **Fiber Gratings**. **Lasers:** Optical Fiber Lasers. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation. **Optical Communication Systems:** Wavelength Division Multiplexing. **Scattering:** Scattering Phenomena in Optical Fibers. **Solitons:** Bright Spatial Solitons; Optical Fiber Solitons, Physical Origin and Properties.

Further Reading

Ablowitz MJ, Biondini G and Ostrovsky LA (eds) (2000) Focus issue: optical solitons: perspectives and applications. *Chaos* 10: 471–640.

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. London: Academic Press.

Agrawal GP and Kivshar YS (2003) *Optical Solitons*. London: Academic Press.

Akhmediev NN and Ankiewicz A (1997) *Solitons: Nonlinear Pulses and Beams*. London: Chapman and Hall.

Buryak AV, Di Trapani P, Skryabin D and Trillo S (2002) Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications. *Physics Reports* 370: 63–235.

De Sterke CM, Eggleton BJ and Sipe JE (2001) In: Trillo S and Torruellas W (eds) *Spatial Solitons*. Berlin: Springer-Verlag.

Di Trapani P, Caironi D, Valiulis G, Dubietis A, Danielius R and Piskarskas A (1998) Observation of temporal solitons in second-harmonic generation with tilted pulses. *Physics Review Letters* 81: 570–573.

Eggleton BJ, Slusher RE, de Sterke CM, Krug PA and Sipe JE (1996) Bragg grating solitons. *Physics Review Letters* 76: 1627–1630.

Forysiak W, Knox FM and Doran NJ (1994) Average soliton propagation in periodically amplified systems with stepwise dispersion-profiled fiber. *Optics Letters* 19: 174–176.

Hasegawa A (2000) Soliton-based optical communications: An overview. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1161–1172.

Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. London: Oxford University Press.

Hasegawa A and Matsumoto M (2002) *Optical Solitons in Fibers*. Berlin: Springer-Verlag.

Haus HA (2000) Mode-locking of lasers. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1173–1185.

Mollenauer LF and Mamyshev PV (1998) Massive wavelength-division multiplexing with solitons. *IEEE Journal of Quantum Electronics* 34: 2102.

Nakazawa N, Kubota H, Sahara A, Suzuki K and Yamada E (2000) Ultrahigh-speed long-distance TDM and WDM soliton transmission technologies. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 363.

Porsezian K and Kuriakose VC (2003) *Optical Solitons: Theoretical and Experimental Challenges*. Berlin: Springer-Verlag.

Slusher RE and Eggleton BJ (eds) (2003) *Nonlinear Photonics Crystals*. Berlin: Springer-Verlag.

Suzuki M and Edagawa N (2003) Dispersion-managed high-capacity ultra-long-haul transmission. *Journal of Lightwave Technology* 21: 916–929.

Suzuki M, Morita I, Edagawa N, Yamamoto S, Toga H and Akiba S (1995) Reduction of Gordon-Haus jitter by periodic dispersion compensation in soliton transmission. *Electronic Letters* 31: 2027–2028.

Turitsyn SK, Aceves AB, Jones CKRT, Zharnitsky V and Mezentsev VK (1999) Hamiltonian averaging in soliton-bearing systems with a periodically varying dispersion. *Physics Review E* 59: 3843–3846.

Wabnitz S, Kodama Y and Aceves AB (1995) Control of optical soliton interactions. *Optics and Fiber Technology* 1: 187–217.

Zakharov VE and Wabnitz S (1999) *Optical Solitons: Theoretical Challenges and Industrial Perspectives*. Berlin: Springer-Verlag.

SPECTROSCOPY

Contents

Absolute Optical Frequency Metrology
Fourier Transform Sepctroscopy
Hadamard Spectroscopy and Imaging
Nonlinear Laser Spectroscopy
Raman Spectroscopy
Second-Harmonic Spectroscopy
Single Molecule Spectroscopy

Absolute Optical Frequency Metrology

S T Cundiff, National Institute of Standards and Technology and University of Colorado, Boulder, CO, USA

L Hollberg, National Institute of Standards and Technology, Boulder, CO, USA

Published by Elsevier Ltd.

Introduction

Visible light occupies a portion of the electromagnetic spectrum between frequencies of approximately 400 terahertz (THz) to 750 THz, where 1 THz is 10^{12} oscillations per second. Optical radiation, which includes adjacent infrared and ultraviolet portions of the spectrum, covers a frequency range of approximately 15–1500 THz. The techniques required for measuring frequencies in this range are qualitatively different from those currently employed in the radio frequency (RF) and microwave portions of the spectrum, where high-speed electronic counters make a direct measurement possible.

The frequency of an oscillatory phenomenon is the number of oscillations that occurs in unit time, or alternatively, the inverse of the period of one oscillation. Thus, to understand frequency, we must understand time, which is intrinsically related to phase. Only a time interval, the time elapsed between two events, has fundamental physical meaning. The 'time' that we are most familiar with, the time of day (including the date), is based on an arbitrarily chosen starting point. The starting point chosen in current international agreements can be traced to noon on December 31st, 1899. Furthermore, the time defined by those agreements, Universal Coordinated Time (UTC), is periodically adjusted relative to atomic time (discussed below) to compensate for irregularities in the rotation of the Earth.

Knowledge of the transition frequencies (energies) in simple atoms gives detailed information on the structure of the atoms. This can be used to test our understanding of the fundamental interactions in nature by comparing the measured frequencies with predictions of fundamental physical theories, in this case quantum electrodynamics (QED). Beautiful results along these lines are exemplified by precision measurements of transition frequencies in hydrogen and helium. In addition, the ability to measure and precisely control laser frequencies promises tremendous advances in the performance of the next generation of atomic clocks and frequency standards. These have potential applications to navigation and communication systems.

The measurement of frequency and time intervals also involves a choice, that of the unit of time, i.e., the definition of the second. The original definition of the second was based on the rotation of the Earth. However, not only is it irregular, but it is slowing down. Both of these facts limit its utility as a frequency standard. Once atomic clocks were highly developed, the second was redefined in 1967 to be the time it takes for the $F = 4$, $m_F = 0 \rightarrow F = 3$, $m_F = 0$ transition in the hyperfine structure of the ground state of ^{133}Cs to undergo 9192 631 770 oscillations. Thus an 'absolute' measurement of time interval, and hence frequency, must be directly connected to this defining frequency. The basic structure of an atomic clock is shown in Figure 1.

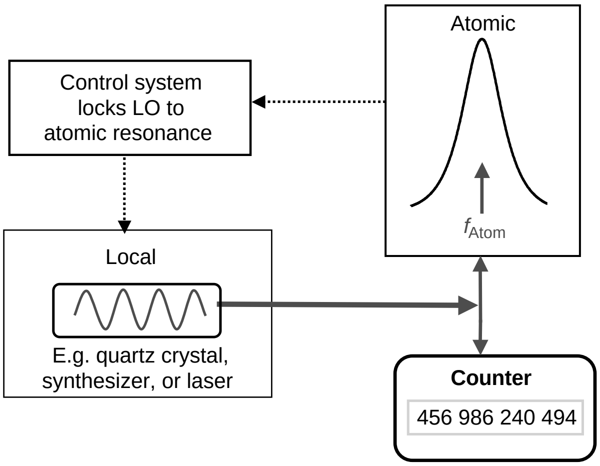
Atomic
Control system locks LO to atomic resonance
f_{Atom}
Local
E.g. quartz crystal, synthesizer, or laser
Counter
456 986 240 494

Figure 1 An atomic clock has three basic components: an atomic resonance, an oscillator to probe and lock to the atomic resonance, and a counter that records the number of oscillations and thus displays the time interval since some chosen starting time.

Frequency standards (or clocks) are characterized by three properties: accuracy, stability, and reproducibility. Accuracy describes how well the natural and fundamental atomic frequency is produced by the atomic frequency standard. Stability describes the frequency fluctuations on short timescales (an accurate but unstable standard produces a frequency that fluctuates around a constant mean that is the atomic frequency). Reproducibility between standards addresses whether or not two implementations of the standard produce exactly the same frequency. Atomic clocks are intrinsically accurate because the frequency of the transitions they are based on is determined by identical atoms in nature. The largest inaccuracies are usually due to environmental perturbations such as tiny variations in the local magnetic field.

A careful analysis of all of these issues, together with technical considerations, yielded the choice of cesium for the current definition of the second. Other common frequency standards include quartz crystal oscillators, hydrogen masers, and rubidium vapor cell standards. Quartz oscillators are inexpensive and have good stability, but in comparison to atomic standards they suffer from accuracy and reproducibility issues. Hydrogen masers have the best stability, but also poor accuracy and reproducibility compared to cesium. Vapor cell rubidium standards can be small and inexpensive but have poorer accuracy than cesium beam standards due to cell and buffer gas effects.

There are several motivations for using optical sources rather than RF or microwaves to measure time intervals and frequencies. Obviously, a higher oscillation frequency can divide time into smaller units and this can provide higher measurement precision. It is thus natural to use the highest oscillation frequency that can be precisely counted to measure time and frequency. Not long after lasers were invented, suggestions were made that it would be possible to make optical frequency standards using lasers locked to narrow atomic resonances.

A basic feature of optical transitions is that they have a very high *Q* factor (the ratio of the center frequency to the linewidth). This allows them to be measured to very high precision (current measurements are good to a few parts in 10^{15}). It may be possible to use this high precision to build improved atomic clocks that use optical transitions instead of the microwave transition currently being used. Candidates for optical standards include both trapped ions (Hg^+, Yb^+, Sr^+, and In^+ are under investigation) and laser-cooled neutral atoms (Ca, Sr, Mg, Ba, and Ag). The high precision is also utilized for fundamental tests of quantum electrodynamics by comparing transition frequencies (usually of hydrogen or helium atoms) to first-principle calculations. For comparison Figure 2 shows the relevant transitions and energy levels in the Cs microwave standard and the Hg^+ optical frequency. It is even possible to use the precision of atomic clocks for high-sensitivity tests of variation in the fundamental constants with time. Finally, because the speed of light is constant length measurements based on the wavelength of light with a known frequency, the precision of length metrology ultimately rests on optical frequency metrology.

An obvious question, which might occur to many in the field of optics, is why a standard spectrometer does not provide absolute frequency measurements. All spectrometers measure wavelength, not frequency, based on interference between light traveling two different path lengths. Thus, knowledge of the path length is needed to determine the wavelength. However, since length is now defined using the speed of light, the definitions become circular unless a source of light with known absolute frequency is available. In addition, frequency measurement is intrinsically more precise and reliable than mechanical length measurement.

Recently, there has been a significant breakthrough in optical-frequency metrology and optical clocks by using mode-locked lasers that generate optical pulses with durations of a few femtoseconds. Before discussing optical frequency metrology based on mode-locked lasers, we will review earlier techniques to provide the necessary background. After describing

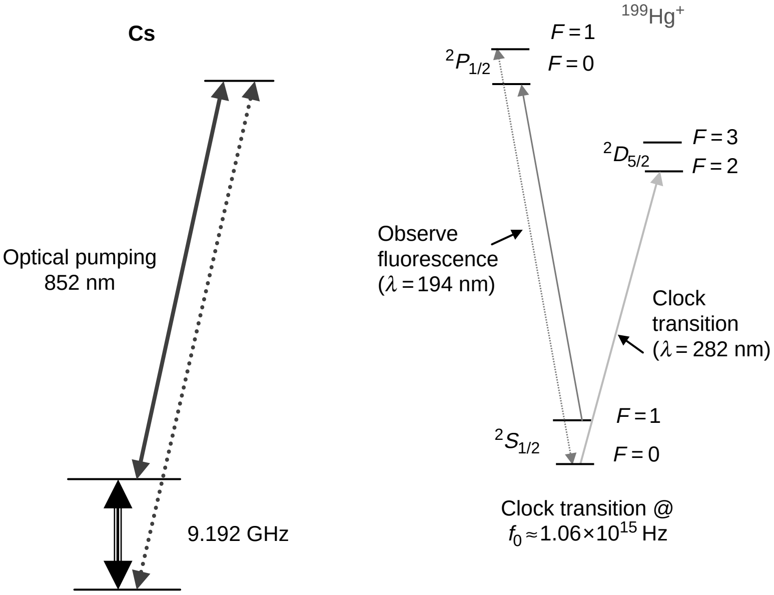
Cs
Optical pumping 852 nm
9.192 GHz
^{199}Hg$^+$
$^2P_{1/2}$
F=1
F=0
$^2D_{5/2}$
F=3
F=2
Observe fluorescence (λ=194 nm)
Clock transition (λ=282 nm)
$^2S_{1/2}$
F=1
F=0
Clock transition @ $f_0 \approx 1.06 \times 10^{15}$ Hz

Figure 2 Simplified energy level diagram of Hg^+ and Cs.

methods that use mode-locked lasers, we will then provide a summary of current measurements and standards. Finally we will briefly describe the outlook for the future of absolute optical frequency metrology.

Background

To imagine being able to measure the frequency of light we need to have laser sources with high coherence, like that of radio waves, where the phase is stable enough to be measured. Even though the first beat-note between two lasers was demonstrated in the 1960s, most lasers were not coherent (stable) enough that the optical phase could be tracked for much longer than about 10 to 100 ns. After many years of research and inspiring technological achievements, Hall, Chebotayev, Bergquist and others were able to stabilize the frequency of tunable laser sources to the point that lasers could indeed be considered coherent sources, with the phase continuously measurable and even predictable for times as long as seconds. Once the lasers were stable enough to be counted, there was still the essential problem that no counter was fast enough to actually count optical frequencies.

With spectrally pure laser sources it was possible to conceive of measuring optical frequencies by extending the nonlinear methods that are used to generate and measure RF and microwave frequencies. By using simple nonlinear components such as diodes, it is easy to generate high harmonics of an input frequency in the RF and microwave regions of the spectrum. At least in principle, and by direct analogy, it seemed that it should be possible to multiply coherent sine waves from radio frequencies up through the microwave region to reach the THz frequencies of far-infrared lasers, and then to use nonlinear optical methods to multiply on up to the infrared (IR) and eventually to the visible region of the spectrum. The concept is simple enough, $f \rightarrow 100 \times f \rightarrow 1000 \times f \rightarrow 2000 \times f \rightarrow 4000 \times f \ldots$ until you reach an optical frequency. Multiplication up to the microwave region works quite well, but unfortunately the nonlinear optical processes in the far-infrared, IR, and optical regions are typically quite inefficient; the power at the second harmonic is typically only 10^{-5} times the square of the fundamental power in watts. Just multiplying a single-frequency laser by a factor of two can be a technological challenge. After reaching a microwave frequency of about 50 GHz we still require a multiplication factor of 10^4 to reach the visible at 500 THz. Since we are usually forced to multiply in the IR and optical by factors of 2, we have $2^N = 10^4$, which means we require $N \approx 13$ multiplication stages. This is approximately the number of stabilized laser oscillators that are required to span the frequency gap from the microwave to the visible. Notwithstanding these technological limitations, Evenson and collaborators did a groundbreaking demonstration of a harmonic frequency chain in 1972 that connected a microwave atomic frequency standard and a stabilized HeNe laser in the IR at 88 THz. With a concerted multiperson effort, more lasers, and 10 years of research, Jennings and co-workers finally succeeded in extending the harmonic optical frequency chain to the visible. That original optical frequency chain from the RF to the visible is diagramed in Figure 3. In the succeeding 18 years, only three other harmonic frequency chains were built world-wide that connected the microwave standards to optical frequencies.

An alternative scheme for measuring optical frequencies is the optical bisection method, a conceptually simple technique proposed by Telle, Meschede and Hänsch, for dividing down from

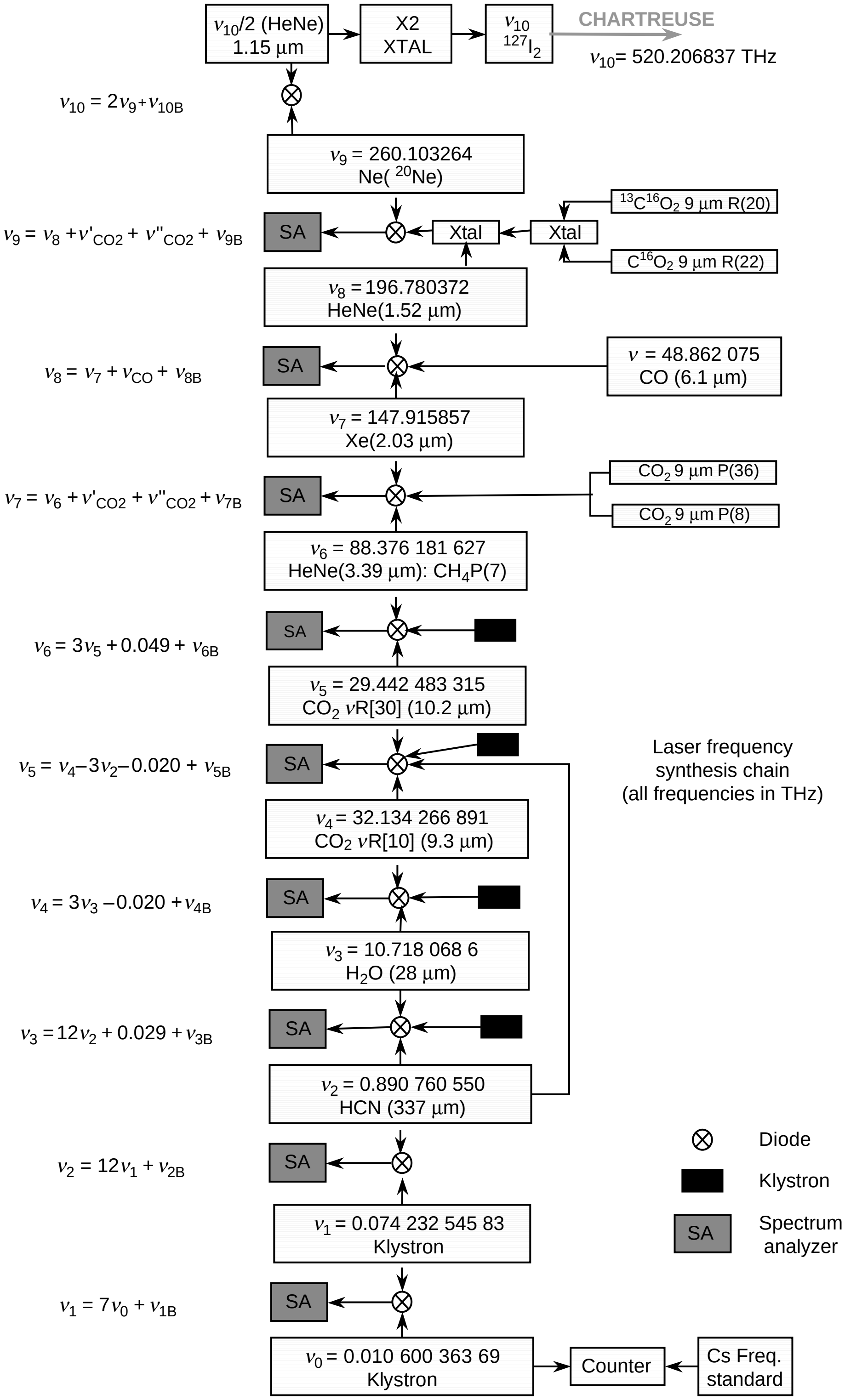
$\nu_{10}/2$ (HeNe) 1.15 µm
X2 XTAL
ν_{10} $^{127}I_2$
CHARTREUSE
ν_{10}= 520.206837 THz
$\nu_{10} = 2\nu_9+\nu_{10B}$
ν_9 = 260.103264 Ne(^{20}Ne)
$^{13}C^{16}O_2$ 9 µm R(20)
Xtal
Xtal
$C^{16}O_2$ 9 µm R(22)
SA
$\nu_9 = \nu_8 + \nu'_{CO2} + \nu''_{CO2} + \nu_{9B}$
ν_8 = 196.780372 HeNe(1.52 µm)
$\nu_8 = \nu_7 + \nu_{CO} + \nu_{8B}$
ν = 48.862 075 CO (6.1 µm)
ν_7 = 147.915857 Xe(2.03 µm)
CO_2 9 µm P(36)
CO_2 9 µm P(8)
$\nu_7 = \nu_6 + \nu'_{CO2} + \nu''_{CO2} + \nu_{7B}$
ν_6 = 88.376 181 627 HeNe(3.39 µm): CH_4P(7)
$\nu_6 = 3\nu_5 + 0.049 + \nu_{6B}$
ν_5 = 29.442 483 315 CO_2 νR[30] (10.2 µm)
$\nu_5 = \nu_4 - 3\nu_2 - 0.020 + \nu_{5B}$
Laser frequency synthesis chain (all frequencies in THz)
ν_4= 32.134 266 891 CO_2 νR[10] (9.3 µm)
$\nu_4 = 3\nu_3 - 0.020 + \nu_{4B}$
ν_3 = 10.718 068 6 H_2O (28 µm)
$\nu_3 = 12\nu_2 + 0.029 + \nu_{3B}$
ν_2 = 0.890 760 550 HCN (337 µm)
$\nu_2 = 12\nu_1 + \nu_{2B}$
Diode
Klystron
Spectrum analyzer
ν_1 = 0.074 232 545 83 Klystron
$\nu_1 = 7\nu_0 + \nu_{1B}$
ν_0 = 0.010 600 363 69 Klystron
Counter
Cs Freq. standard

Figure 3 Schematic of the first frequency chain used to measure optical frequencies relative to the Cs standard.

optical frequencies to microwaves. In fact, the bisection method divides an arbitrary optical frequency interval in half by utilizing second-harmonic generation and sum-frequency-mixing, and forcing the condition that $2f_3 = f_1 + f_2$. Thus, the laser at frequency $f_3 = (f_1 + f_2)/2$ bisects the frequency interval between the lasers at f_1 and f_2. The optical bisection method has the following two significant advantages over the multiplication method: the phase noise is decreased in successive divisions (rather than increased by multiplications), and the system could be constructed out of one convenient and reproducible laser technology, say diode lasers in the near-visible. In principle, any optical frequency could be measured by successive operations of the optical bisector. Optical frequency intervals as large as 8 THz have been measured with this method, but even though it is feasible, no bisection system has been built to date that connects the optical region all the way to countable microwave frequencies. The main limitation is the same as that facing the multiplication scheme: it still requires lots of stabilized laser sources to divide by factors of 2 from 10^{15} to 5×10^{10} Hz. The problem can be reduced somewhat by taking advantage of an optical-frequency comb generator based on an electro-optic modulator inside a resonant optical cavity. These systems, developed by Korougi and others, generate many coherent microwave sidebands on a laser so that optical frequency intervals on the order of 5 THz can be measured.

Even combining all of these ideas and techniques, a convenient system to count optical frequencies remained elusive. This has now changed with the revolutionary new optical frequency combs based on mode-locked femtosecond lasers.

Optical Frequency Metrology with Femtosecond Combs

Recently, the introduction of mode-locked lasers into the field of absolute optical frequency metrology has resulted in an important advance. The extremely complex frequency chains described above can be replaced by a single mode-locked laser if it produces pulses with sufficiently large bandwidth. Although the underlying ideas are not new (they were originally discussed by Hänsch and Chebotayaev in the 1970s) it is only recently that mode-locked lasers with the required characteristics were developed and the concept carefully tested.

The dramatic simplification of a frequency chain that uses a mode-locked laser means that absolute optical-frequency measurements can now be made by a single person as compared to the team of approximately 10 highly trained scientists required to run the previous chains. This has resulted in a large number of measurements being reported in the last 2 years. The simplicity also allows longer averaging times and a greater number of measurements to be used in determining the final reported frequency. In addition, since a simpler system has fewer sources of error, it has improved the quality of measurements made in a given amount of time.

Mode-locked lasers produce ultrashort pulses of light. For a typical modern high-quality mode-locked laser, the pulse duration is around 10 fs (1 fs = 10^{-15} second), corresponding to three or four optical cycles in the near-infrared, where these lasers typically operate. The current best mode-locked lasers can produce pulses that are shorter than two optical cycles in duration (5–6 fs). Since the spectrum of these lasers is complicated, the bandwidth is significantly larger than it would be for a smooth pulse such as one with a Gaussian or hyperbolic-secant temporal intensity profile.

At first sight, using mode-locked lasers for optical-frequency metrology seems counter-intuitive, because optical-frequency metrology clearly requires very well-defined frequencies in contrast to the broad frequency spectrum of ultrashort pulses. This apparent conundrum is resolved by the fact that a mode-locked laser actually produces a very regularly spaced train of ultrashort pulses. The ultrashort pulses emitted by a mode-locked laser are replicas of a steady-state pulse that circulates inside the cavity of the laser. Every time the internal pulse impinges on the output coupler, which is a partially reflective mirror, a portion of it is transmitted, producing the output. The timing between these output pulses is determined purely by the time, τ_c, for the intracavity pulse to make one round trip, which is typically between 1 and 20 ns. The spectrum of a train of pulses is a regularly spaced 'comb' of frequencies spaced by the repetition rate of the train, $f_{rep} = 1/\tau_c$. If all of the pulses are identical, then these comb lines are just integer multiples of f_{rep}. However, due to dispersion, the group and phase velocities are different inside the cavity of a mode-locked laser. This leads to a pulse-to-pulse shift in the phase between the carrier and envelope (which we call the carrier–envelope phase; ϕ_{ce}, the pulse-to-pulse shift of ϕ_{ce} is designated by $\Delta\phi_{ce}$). The presence of $\Delta\phi_{ce}$ results in a rigid shift of all the comb lines by an amount $\delta = 2\pi\Delta\phi_{ce}/f_{rep}$. Thus the frequencies of the comb lines are given by $\nu_n = nf_{rep} + \delta$, where n is a large integer of order 10^6. The essential point of this equation is that it gives optical frequencies in terms of RF frequencies (f_{rep} and δ) that can easily be measured with conventional electronics and

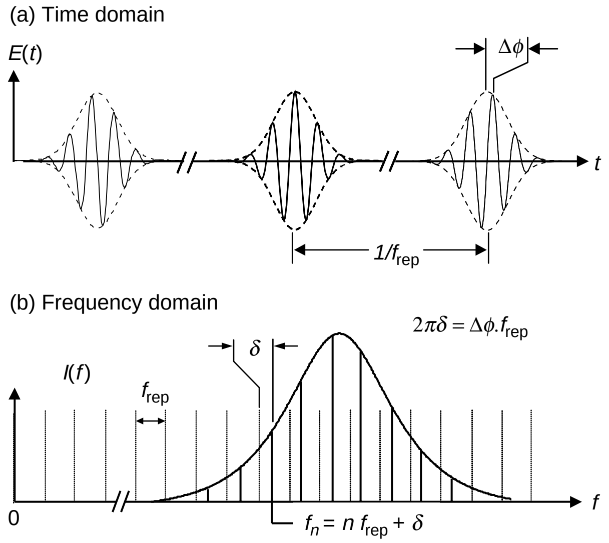

Figure 4 Time–frequency correspondence and relationship between $\Delta\phi$ and δ. (a) In the time domain, the relative phase between the carrier and the envelope evolves from pulse to pulse by the amount $\Delta\phi$. (b) In the frequency domain, the elements of the frequency comb of a mode-locked pulse train are spaced by f_{rep}. The entire comb is offset from integer multiples of f_{rep} by an offset frequency $\delta = \Delta\phi f_{\text{rep}}/2\pi$.

compared to the microwave cesium frequency standard. The relationship between the time and frequency domains is shown in Figure 4.

Given such a comb of frequencies, a heterodyne measurement readily yields the frequency difference between the unknown optical frequency of a narrow-band source (typically a single frequency laser) and the optical frequency of a nearby comb line. Thus, the absolute unknown frequency can be determined if the absolute frequencies of the comb lines are known. This requires measurement of f_{rep} and δ. The repetition rate, f_{rep}, is easily measured with a fast photodiode. Measurement of δ requires somewhat greater effort.

The development of a 'self-referencing' method to easily measure δ is the key enabling breakthrough that has brought about the recent revolution in optical-frequency metrology. It is called self-referencing because it provides a direct measure of δ with no other optical frequencies as input. It works by comparing frequencies that differ by a factor of two (typically these lie in the wings of the spectrum) using second-harmonic generation. Specifically, the frequency difference between the second harmonic of a comb line n and the comb line $2n$ is given by $2f_n - f_{2n} = 2(nf_{\text{rep}} - \delta) - (2nf_{\text{rep}} - \delta) = \delta$. Such a frequency difference is easily measured using a heterodyne beat. This technique is shown schematically in Figure 5. Other variations on this basic idea are possible; for example, rather than frequency doubling the comb itself, it is also possible to extract the same information by comparing the comb lines to a single-frequency laser and its second harmonic.

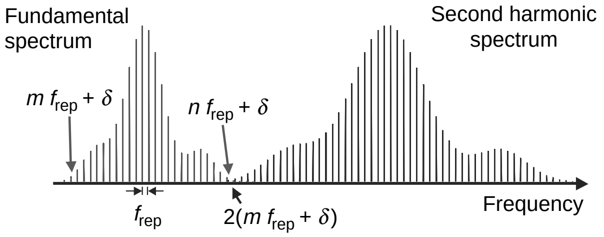

Figure 5 Schematic of the self-referencing technique for determining the offset frequency of the frequency comb produced by a femtosecond mode-locked laser. The frequency difference between the fundamental comb and its second harmonic is the offset frequency. In the region where overlap occurs, this is easily detected as a heterodyne beat.

Implementation of the self-referencing technique requires a pulse spectrum that spans a factor of two in frequency (an optical octave). Although a laser has recently been demonstrated that produces such a broad spectrum, they are not commonly available. However, the development of microstructured fiber has made it easy to broaden the spectrum of ordinary mode-locked lasers so that it spans an octave, as was first demonstrated by Ranka and co-workers at Bell Labs. The broadening occurs because optical nonlinearity in the fiber results in self-phase modulation of the laser pulse. Microstructure fiber achieves guiding by surrounding the core region with microscopic air holes, as compared to regular fiber that uses doping to produce a difference in the index of refraction. The much larger difference in the index of refraction in microstructure fiber allows a smaller core region to be used, which increases the effective nonlinearity. It also makes it possible to modify the dispersion such that the group velocity dispersion goes through zero for light with a wavelength close to 800 nm. This is crucial because group-velocity dispersion in ordinary fiber causes ultrashort pulses to temporally spread, which lowers the peak power and hence nonlinearity. It is possible to implement self-referencing techniques using less bandwidth if higher orders of nonlinearity are used; for example, if the difference between the second harmonic of the high-frequency end and third harmonic of the low-frequency end of the comb are used, only a half-octave bandwidth is required.

Given the ability to measure both f_{rep} and δ, and as long as its frequency lies within the comb spectrum, it is possible to measure the optical frequency of any optical source by measuring the heterodyne beat with a nearby comb line. Thus, an absolute optical frequency measurement comes down to measuring three RF frequencies. It is often advantageous to actively control either one or both of f_{rep} and δ to hold them to a fixed frequency. This requires real-time adjustment of some parameters of the mode-locked laser to affect the necessary feedback. Controlling f_{rep} is straightforward; it is achieved by simply changing the cavity length of the laser, typically by mounting a mirror on a piezoelectric transducer. Controlling δ is more challenging, as it requires a differential change between the cavity group delay and phase delay. This has been achieved by adjusting the pump power, which in turn controls the energy of the intracavity pulse, and/or by tilting the end mirror of the cavity in lasers that employ prism dispersion compensators. Since the optical spectrum is spread out across this mirror, tilting it can be thought of as a frequency-dependent change in the cavity length, with the optical frequency corresponding to the pivot point of the tilt experiencing no length change, while those on opposite sides of the pivot point see changes with opposite signs.

In addition to the three RF frequencies, it is necessary to know the large integer n that describes the comb line that is beat against the unknown source. If a priori knowledge about the unknown source is available with precision somewhat better than $f_{\text{rep}}/2$ or better, n can readily be determined. For typical mode-locked lasers, standard commercial wavemeters can provide this level of precision. In the absence of additional outside information, n can also be determined with some effort and precision measurements by changing f_{rep} and δ in a systematic fashion. In the end, the simple formula $f_n = nf_{\text{rep}} + \delta$ gives the optical frequency of any arbitrary nth mode of the femtosecond optical comb. With this established, the frequency of any single-mode laser within the octave span of the comb can be determined simply by measuring the frequency of the beat-note generated in a photodiode between the cw laser and the known frequencies of the comb.

The diagram in Figure 6 shows how a femtosecond comb acts as an optical frequency synthesizer, which provides frequencies that are known relative to the input frequency (provided by a hydrogen maser in the figure). A typical series of measurements is shown in Figure 7 for the NIST Hg^+ single ion standard. The inset shows the results of numerous measurements for the frequency of Hg^+ relative to the NIST primary cesium standard.

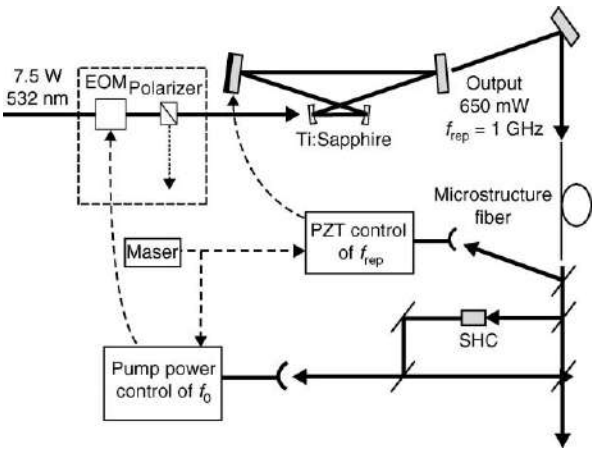

Figure 6 Schematic example of how the femtosecond comb is configured to make optical frequency measurements relative to the reference frequency provided by a hydrogen maser (Maser) that is in turn calibrated to a cesium atomic frequency standard. The repetition rate of the pulses from the Ti:Sapphire laser are detected in a photodiode and are controlled using a piezoelectric transducer (PZT) that changes the length of the laser cavity. The microstructure fiber and second harmonic crystal (SHC), provide the 'self-referencing' beat-note signal that is used to control the offset frequency, δ, by changing the power of the pump laser with an electro-optic modulator (EOM). This 'self-referenced' system then provides a set of modes with known frequencies spanning the visible. An unknown laser frequency can then be measured by making a beat-note with one of the modes of the comb.

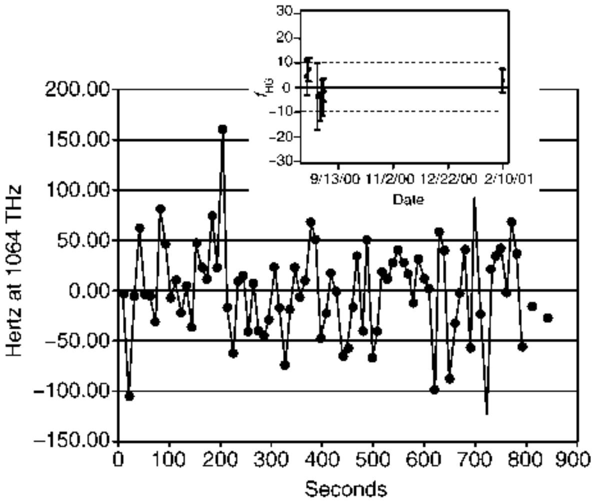

Figure 7 Actual frequency measurements of the beat-note between a hydrogen maser stabilized optical frequency comb (as shown in Figure 6) and the NIST Hg^+ optical frequency standard. The scatter in the data for counter gate time of 10 seconds corresponds to a fractional frequency uncertainty of about 4×10^{-14}, consistent within the instability in the maser. The inset shows the results from numerous datasets of measurement of the frequency of the Hg^+ optical standard relative to the cesium primary standard at NIST. The frequency data are plotted relative to 1064 721 609 899 142.6 Hz. The dotted lines give estimates of systematic uncertainties in lieu of a complete accuracy evaluation.

Table 1 Summary of optical frequency measurements of a selection of molecular, atomic, and ionic transitions. Recent measurements listed for all and additional measurements are given for a few to indicate the level of agreement

Atom/molecule/ion	*Wavelength (nm)*	*Frequency (THz)*	*Uncertainty*		*Year*	*First author and institution*
			± Hz	*Frac.*		
O_sO_4	10318.0	29.054 057 446 660	50	2×10^{-12}	1985	Clarion, LPTF
		29.054 057 446 579	10	3×10^{-13}	1999	Ducos, LPTF
CH_4	3392.2	88.376 181 627 000	50 000	6×10^{-10}	1972	Evenson, NBS
		88.373 149 028 553	200	2×10^{-11}	1998	Ering, PTB
Rb−2 photon	778.11	385.285 142 367 000	8000	2×10^{-11}	1993	Nez, ENS
		385.285 142 374 800	3000	8×10^{-12}	2000	Diddams, JILA
Sr^+	674.03	444.779 904 409 540	200	4×10^{-13}	1999	Bernard, NRC
Ca	657.45	455.986 240 495 150	8	10×10^{-14}	2003	Helmcke, PTB
		455.986 240 494 158	26	6×10^{-14}	2000	Udem, NIST
I_2 a_{15} of $R(127)$ 11-5	632.99	473.612 340 492 000	74 000	2×10^{-10}	1983	Jennings, NBS
I_2 a_{16} of $R(127)$ 11-5	632.99	473.612 353 604 800	1200	3×10^{-12}	2000	Ye, JILA
I_2 a_{10} of $R(56)$ 32-0	532.24	563.260 223 480 000	70 000	1×10^{-10}	1995	Jungner, JILA
		563.260 223 514 000	5000	9×10^{-12}	2000	Diddams, JILA
Yb^+	435.51	688.358 979 230 931	6	9×10^{-15}	2001	Stenger, PTB
H 2*S*-8*S*	389.01	770.649 561 581 100	5900	8×10^{-12}	1997	de Bouroir, ENS
Hg^+	281.57	1064.721 609 899 140	10	9×10^{-15}	2000	Udem, NIST
In^+	236.54	1267.402 452 899 920	230	2×10^{-13}	2000	von Zanthier, MPQ
H 1*S*-2*S*	121.57	2466.061 413 187 100	46	2×10^{-14}	2000	Holzwarth, MPQ

Summary of Measurements/Standards

Starting with the first measurement of the frequency of the methane-stabilized HeNe laser in 1972, a new laser frequency measurement was reported every year or so using harmonic frequency chains. These are summarized in Table 1. An indication of the history of frequency measurements of stable laser references is given in Figure 8. The accuracy, in terms of fractional frequency uncertainty, improved from 10^{-10} in 1972 to 10^{-12} in 1999, at which time the femtosecond optical frequency combs came on line and there was a dramatic improvement in the precision as well as the number of measurements that were completed. It is impossible to predict with certainty but some have projected that optical standards can ultimately reach uncertainties as small as 10^{-18}.

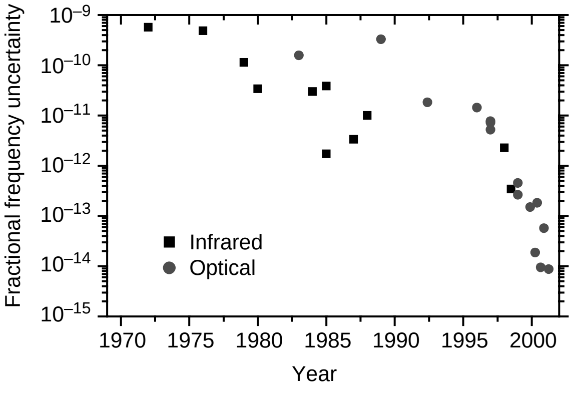

Figure 8 Progress in the accuracy of optical frequency measurements.

Outlook

Three separate technologies have now reached a level of maturity that it is possible to build high-performance optical-frequency standards and clocks. The essential achievements are: laser cooling and trapping of atoms (first proposed by Wineland and Dehmelt, and Hänsch and Schawlow), highly stabilized narrow-linewidth cw lasers, and femtosecond optical frequency combs. Combining these key ingredients, we can construct an optical atomic clock as shown schematically in Figure 9.

Here, for comparison with Figure 1, the cold atoms or single ion provide the narrow atomic resonance, the cw laser serves as the local oscillator to probe the resonance, and the femto-comb serves as the counter. Optical frequency standards of the future are expected to provide orders of magnitude better stability and improved accuracy over the existing atomic frequency standards that now use microwave transitions in atoms. As described in the previous section, frequency combs produced by femtosecond lasers can directly measure the frequency of a stable laser locked to an atomic transition relative to a known microwave frequency standard. This gives fundamental information about the atomic energy levels, and structure, and allows comparisons between different elements. However, to take

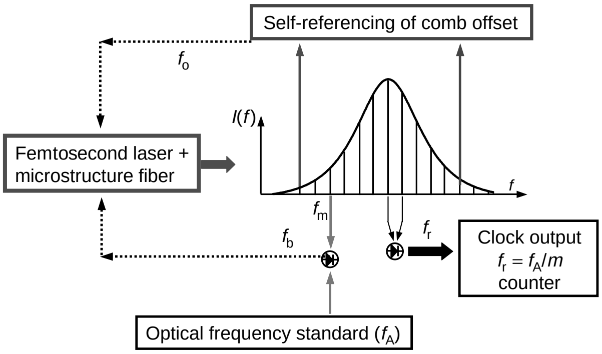
Self-referencing of comb offset
f_o
$I(f)$
Femtosecond laser + microstructure fiber
f
f_m
f_b
f_r
Clock output
$f_r = f_A/m$
counter
Optical frequency standard (f_A)

Figure 9 Simplified schematic of an optical clock.

advantage of the high stability of the optical references we run the system as an optical clock (Figure 9) where the stable laser and the femto-comb are locked to the atomic resonance, and the clock output comes as pulses at the repetition rate (e.g., 1 GHz) of the femtosecond mode-locked laser. With a judicious choice of control parameters it is possible to have the pulse repetition frequency, i.e., the clock output, at an exact subharmonic of the optical transition frequency. It is intriguing to note that a portable optical clock could measure time and length at the same time.

See also

Instrumentation: Spectrometers. **Quantum Electrodynamics:** Quantum Theory of the Electromagnetic Field.

Further Reading

Barnett JE (1998) Time's Pendulum. New York: Plenum.

Bergquist JC, Jefferts SR and Wineland DJ (2001) Time measurement at the millennium. *Physics Today* March: 37–42.

Cundiff ST, Ye J and Hall JL (2001) Optical frequency synthesis based on mode-locked lasers. *Reviews of Scientific Instruments* 72: 3749–3771.

Diddams SA, Jones DJ, Ye J, *et al.* (2000) Towards the ultimate control of light. *Optics and Photonics News* October: 16–22.

Diddams SA, Udem T, Bergquist JC, *et al.* (2001) An optical clock based on a single trapped $^{199}Hg^+$ ion. *Science* 293: 825–828.

Diels JC and Rudolph W (1996) Ultrashort Laser Pulse Phenomena. San Diego, CA: Academic Press.

Hall JL (2000) Optical frequency measurement: 40 years of technology revolutions. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1136–1144.

Hall JL and Ye J (2001) A new era of frequency standards and optical frequency measurements. *Optics and Photonics News* February: 45–50.

Hänsch TW and Walther H (1999) Laser spectroscopy and quantum optics. *Reviews of Modern Physics* 71: S242–S252.

Hollberg L, Diddams SA, Curtis EA, Oates CW and Fox RW (2001) Optical frequency standards for clocks of the future. In: The International Society for Optical Engineering. *Harnessing Light: Optical Science and Metrology at NIST.* Bellingham, DC. *Proceedings of SPIE* 4450: 54–59.

Jennings DA, Evenson KE and Knight DJE (1986) Optical frequency measurements. *Proceedings of the IEEE* 74: 168–179.

Jespersen J and Fitz-Randolph J (1977) From Sundials to Atomic Clocks. New York: Dover.

Jones T (2000) Splitting the Second: The Story of Atomic Time. Bristol, UK: Institute of Physics.

Luiten AN (2001) Frequency Measurement and Control: Advanced Techniques and Future Trends. Berlin: Springer-Verlag.

Menoni CS (ed.) (2001) Feature section on optical frequency synthesis: a new tool for precision optical frequency metrology. *IEEE Journal of Quantum Electronics* 37: 1481–1513.

Quinn TJ (1999) Practical realization of the definition of the metre. *Metrologia* 36: 211–244.

Fourier Transform Spectroscopy

T Fromherz, Johannes Kepler Universität, Linz, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fourier transform spectroscopy (FTS) has emerged as one of the most powerful spectroscopic techniques since the first instruments became commercially available in the late 1960s. With this method, high spectral resolution, high wavenumber accuracy, broad spectral range, high optical throughput, and high signal to noise ratio can be achieved simultaneously. Because of these combined advantages that will be discussed in more detail throughout this article, especially in the far- (10–500 cm^{-1}) and mid-infrared (FIR and MIR) spectral region (500–5000 cm^{-1}), FTS has become the method of choice for the most sensitive spectral investigations. In this article, an overview of the setup, the working principle, and the computation of spectra from the measured interferograms will be given. Finally, by comparing FTS to dispersive spectroscopic methods, the conditions under which FTS can be exploited to advantage, will be discussed.

Spectrometer Setup

Almost all commercially available Fourier transform (FT) spectrometers are based on a Michelson interferometer. A typical layout of this interferometer is sketched in Figure 1. A collimated beam with spectral density $S(\sigma)$ entering the spectrometer is divided into two beams by a beamsplitter. Here, σ denotes the wavenumber that is related to the wavelength (λ) of the radiation by $\sigma = 1/\lambda$. As shown in Figure 1, one part of the radiation is reflected by the beamsplitter and propagates through the compensator plate to a fixed mirror. The mirror reflects the beam through the compensator plate back onto the beamsplitter. The other part of the radiation beam is transmitted through the beamsplitter and its substrate, and hits a moveable mirror from which it is reflected back onto the beamsplitter. At the beamsplitter, both beams are divided again. As shown in Figure 1, two beams, one propagating to the source, the other propagating to the detector, finally result. Both beams contain a contribution that was reflected by the fixed mirror and a second contribution that was reflected by the moveable mirror. In Figure 1, lines indicate the paths of the beams schematically. In the case of a reflection on the beamsplitter, the line type is changed in Figure 1, whereas for a transmission, the line type remains unchanged. The widths of the lines in Figure 1 indicate the size of the field amplitudes and, therefore, are reduced for each reflection/transmission on the beamsplitter. For clarity, the lines indicating the beams reflected into themselves by the fixed and moveable mirrors in Figure 1 are offset laterally.

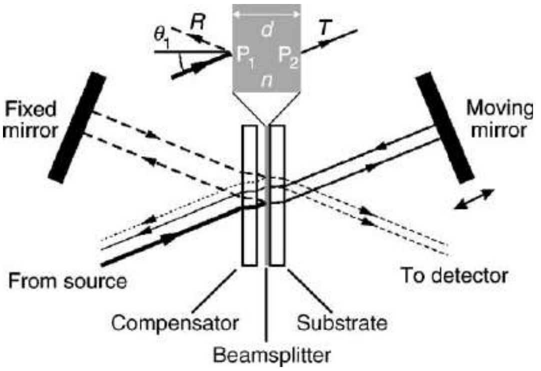

Figure 1 Schematic sketch of the layout of a Michelson interferometer. The paths of the rays are indicated by lines and arrows. At a reflection on the beamsplitter, the line type is changed, whereas for a transmission it remains unchanged. The widths of the lines indicate the size of the field amplitudes. For clarity, the lines indicating the beams reflected into themselves by the fixed and moveable mirrors, are offset laterally. The inset shows a magnified cross-section of the beamsplitter. In the inset, the geometrical details used for calculating the phase difference between *R* and *T* are defined.

By changing the position of the moveable mirror along the beam axis, a difference between the optical paths from the beamsplitter to the fixed and moveable mirror is generated. The compensator consists of the same material and has the same thickness as the beamsplitter substrate and, therefore, minimizes the influence of the dispersion of the beamsplitter substrate on the optical path difference.

Fourier Analysis and Interferometry

To calculate the intensities of the combined beams falling on the detector and on the source, we start by considering the phase difference between the reflected and transmitted beams leaving the beamsplitter. For simplicity, we ignore the compensator and beamsplitter substrate, since ideally, they generate the same phases in both arms of the interferometer, and therefore, do not contribute to the phase difference. Thus the beamsplitter is regarded as a freestanding, dielectric slab with thickness d. The complex reflection (R) and transmission (T) coefficients, including the effect of multiple internal reflections within the beamsplitter layer, are given by:

$$\begin{aligned} R &= r(1 - e^{i2\gamma})/(1 - r^2 e^{i2\gamma}) \\ T &= e^{i\gamma}(1 - r^2)/(1 - r^2 e^{i2\gamma}) \end{aligned} \qquad [1]$$

Here, r is the Fresnel reflection coefficient that describes the ratio of the moduli and the phase difference between incoming and reflected electrical field for a electromagnetic wave incident under a angle Θ_1 on the surface of a dielectric with dielectric constant n, and $\gamma = 2\pi\sigma n d \cos(\Theta_2)$ is the phase difference evaluated at P_1 and P_2 in Figure 1 for a single pass of the electromagnetic wave through the dielectric slab under the angle Θ_2 (by Snellius law, $\sin(\Theta_1) = n \sin(\Theta_2)$). Since n is real, so is r and, therefore, the phase difference $\Delta\varphi$ between R and T is calculated from eqn [1] according to:

$$\begin{aligned} \frac{R}{T} &= \frac{|R|}{|T|} e^{i\Delta\varphi_{RT}} = \frac{r}{1 - r^2}(e^{-i\gamma} - e^{i\gamma}) \\ &= -2i\frac{r}{1 - r^2}\sin(\gamma) \end{aligned} \qquad [2]$$

i.e., $\Delta\varphi_{RT} = -\pi/2$. With this result, we are able to calculate the phase difference between the two rays falling on the detector and the two rays falling back on the source. The geometry considered in the following is sketched in Figure 1. The coefficients R and T connect the phases and magnitudes of reflected and transmitted electric field amplitudes at the points P_1 and P_2. From these points, the radiation beams travel back and forth to the fixed and moveable mirrors, thereby accumulating phases according to: $\varphi_f = 2\pi 2 x_f \sigma$ and $\varphi_m = 2\pi 2(x_f + \Delta x)\sigma$, where we have defined $\Delta x = x_m - x_f$ and x_m, x_f denote the distances of moveable and fixed mirrors from the points P_1 and P_2. The two planewaves falling on the detector interact twice with the beamsplitter. Each planewave is once reflected at and once transmitted through the beamsplitter. Therefore, the beamsplitter adds no additional phase difference to the two planewaves and the total phase difference is given by $\Delta\varphi = \varphi_m - \varphi_f$. The situation is different for the two beams falling back on the source. Here, one beam is reflected at both interactions with the beamsplitter, whereas the other is transmitted at both interactions. According to eqn [2], between these two beams an additional phase shift of π results. Quantitatively, the total field amplitude at the detector, neglecting an arbitrary phase, is given by $E_D = E_0 RT(e^{i\varphi_m} + e^{i\varphi_f})$, and the corresponding intensity is:

$$\begin{aligned} I_D \propto E_D E_D^* &= 2|E_0|^2|R|^2|T|^2\{1 + \mathrm{Re}[e^{i(\varphi_m - \varphi_f)}]\} \\ &= 2|E_0|^2|R|^2|T|^2\{1 + \cos(4\pi\sigma\Delta x)\} \end{aligned} \qquad [3]$$

The sum of the field amplitudes falling back on the source is $E_S = E_0(R^2 e^{i\varphi_m} + T^2 e^{i\varphi_f})$. For the intensity one gets:

$$I_S \propto E_S E_S^* = |E_0|^2\{|R|^4 + |T|^4 + 2\mathrm{Re}[R^2 T^{*2} e^{i(\varphi_m - \varphi_f)}]\} \qquad [4]$$

Denoting the complex coefficients R and T by their modulus and their phase $R = |R|e^{i\varphi_R}$, $T = |T|e^{i\varphi_T}$ and using eqn [2], one gets:

$$I_S \propto |E_0|^2\{|R|^4 + |T|^4 + 2|R|^2|T|^2\cos(4\pi\sigma\Delta x + \pi)\} \qquad [5]$$

Both intensities, I_D and I_S contain a constant component and a component that varies with $\cos(4\pi\sigma\Delta x)$. The varying parts have the same amplitude $2|E_0|^2|R|^2|T^2|$ but are out of phase by π. Consequently, a maximum in I_D corresponds to a minimum in I_S and the sum of both intensities is independent of the mirror displacement Δx.

In most of the commercially available spectrometers, only I_D is used for spectral measurements. Usually, the detector signal is amplified by an AC coupled electronics. Therefore, in a measurement of $I_D(\Delta x)$, the constant part of the signal is suppressed. Assuming a spectral density $S(\sigma)$ of the source and an ideal detector with a responsivity R_D linear in the intensity and independent of the wavenumber σ, the detector signal $D(\Delta x)$ is given by:

$$D(\Delta x) = R_D \int_0^\infty S(\sigma)\cos(4\pi\sigma\Delta x)\,d\sigma \qquad [6]$$

The information about the spectrum is contained in $S(\sigma)$ and can be extracted from the interferogram $D(\Delta x)$ by a cosine Fourier transformation. The transformation has to be done numerically and, therefore, it is more convenient to use a complex, fast Fourier transform algorithm. For that purpose, eqn [6] is modified by using the complex expansion of the cosine $\cos(x) = (e^{ix} + e^{-ix})/2$:

$$\begin{aligned} D(\Delta x) &= \frac{R_D}{2}\left(\int_0^\infty S(\sigma)e^{i4\pi\sigma\Delta x}d\sigma + \int_0^\infty S(\sigma)e^{-i4\pi\sigma\Delta x}d\sigma\right) \\ &= \frac{R_D}{2}\left(\int_0^\infty S(\sigma)e^{i4\pi\sigma\Delta x}d\sigma + \int_{-\infty}^0 S(-\sigma)e^{i4\pi\sigma\Delta x}d\sigma\right) \end{aligned} \qquad [7]$$

Extending $S(\sigma)$ to negative wavenumbers according to:

$$S'(\sigma) = \begin{cases} S(\sigma) & \text{for } \sigma \geq 0 \\ S(-\sigma) & \text{for } \sigma < 0 \end{cases} \qquad [8]$$

leads to the following expression for the interferogram:

$$D(\Delta x) = \frac{R_D}{2}\int_{-\infty}^{\infty} S'(\sigma)e^{i4\pi\sigma\Delta x}d\sigma \qquad [9]$$

and the spectral density $S'(s)$ can be obtained by the complex Fourier transform according to:

$$S'(\sigma) = \frac{4}{R_D}\int_{-\infty}^{\infty} D(\Delta x)e^{-i4\pi\sigma\Delta x}d(\Delta x) \qquad [10]$$

In the following, we will skip the prime and S will refer to the spectral density extended to negative values of σ.

To calculate the spectral density according to eqn [10], one has to use an interferogram extending from $-\infty$ to $+\infty$ measured at infinitely small sampling distances $d(\Delta x)$. In practise, of course, the interferogram is measured only between finite mirror displacements $-\Delta x_{max}$ and Δx_{max}, at sampling points equally spaced by ξ.

The effect of the finite mirror displacement on the resulting spectral density S can be included into eqn [10] by multiplying the interferogram $D(\Delta x)$ with a function that vanishes for $|\Delta x| > \Delta x_{max}$. As an example, we use the boxcar function

$\theta(\Delta x + \Delta x_{\max})\theta(\Delta x_{\max} - \Delta x)$, where $\theta(x)$ is the Heavyside step function with values 0 and 1 for $x < 0$ and $x > 0$, respectively, and calculate the output spectrum for a spectral density $S_1(\sigma)$ consisting of a single line at wavenumber $\sigma_0 : S_1(\sigma) = (\delta(\sigma - \sigma_0) + \delta(\sigma + \sigma_0))/2$ (note that the spectral density is extended to negative wavenumbers according to the definitions given above). According to eqn [9], the interferogram is $D_1(\Delta x) = (R_D/2)\cos(4\pi\sigma_0\Delta x)$. Multiplying the interferogram with the boxcar function and Fourier transforming the product back using eqn [10], the resulting spectral density S_1^{box} is:

$$S_1^{\mathrm{box}} = 2\Delta x_{\max}\left\{\frac{\sin[4\pi(\sigma_0 - \sigma)\Delta x_{\max}]}{4\pi(\sigma_0 - \sigma)\Delta x_{\max}} + \frac{\sin[4\pi(\sigma_0 + \sigma)\Delta x_{\max}]}{4\pi(\sigma_0 + \sigma)\Delta x_{\max}}\right\} \quad [11]$$

Therefore, the approximation to a monochromatic spectrum is a sum of $\mathrm{sinc}(z_\pm)$ ($= \sin(z_\pm)/z_\pm$) functions where $z_\pm = 4\pi(\sigma_0 \pm \sigma)\Delta x_{\max}$. The lineshape function S_1^{box} is plotted around σ_0 in Figure 2 by the full line. Compared to the monochromatic spectrum S_1, the lines of S_1^{box} have a finite linewidth with a fullwidth at half maximum (FWHM) of approximately $0.3/\Delta x_{\max}$. As a general rule in spectroscopy, two spectral lines of equal intensity and linewidth have to be separated at least by their FWHM, in order to be clearly distinguishable. According to this definition, the limit of resolution of a FT spectrometer $\Delta\sigma_{\min}$ is inversely proportional to the maximum mirror displacement and given by $\Delta\sigma_{\min} = 0.3/\Delta x_{\max}$.

A less precise, but more intuitive definition of the resolution refers to the difference of the phases of the interferograms corresponding to two monochromatic lines separated by $\Delta\sigma$. The two lines are defined to be resolvable if the phases of their interferograms differ by more than 2π between the mirror positions $-\Delta x_{\max}$ and $\Delta x_{\max}$, i.e., the beating in the superposition of the interferograms shows at least one period. It is easily seen that, according to this definition $\Delta\sigma_{\min} = 1/(2\Delta x_{\max})$, a result in good agreement but slightly larger than that obtained by the more precise definition given above.

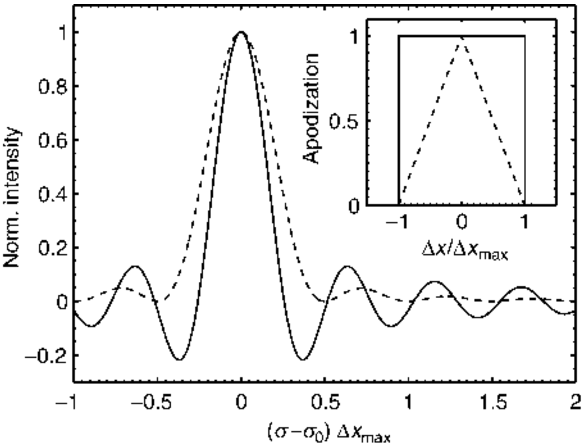

Figure 2 Normalized lineshape functions for boxcar (full line) and triangular apodization (broken line). In the inset, the corresponding apodization functions are shown.

As shown in Figure 2 by the full line, the sinc function has negative side-lobes with magnitudes of up to 22% of the maximum value. For some spectroscopic applications this side lobes are unwanted since they lead to a suppression of smaller peaks separated by approximately the resolution limit of the spectrometer. Thus, in these cases, it appears to be reasonable to multiply the interferogram by a (apodization) function different from the natural boxcar function. Usually, triangular or trapezoidal functions vanishing for $|\Delta x| > \Delta x_{\max}$ are used. By Fourier transformation of these functions, it can be shown that no negative side lobes are generated. However, these functions have a larger (up to ~50%) peak FWHM than the sinc function. For comparison, in Figure 2, also the lineshape of a monochromatic spectrum using a triangular apodization function $A(\Delta x) = \theta(\Delta x + \Delta x_{\max})\theta(\Delta x_{\max} - \Delta x)(1 - |\Delta x/\Delta x_{\max}|)$ is shown by the broken line. In this case, the negative sidelobes are strongly suppressed but the FWHM is increased.

Generally, the output spectrum $S^{\mathrm{A}}(\sigma)$ corresponding to an arbitrary spectral density $S(\sigma)$ and an apodization function $A(\Delta x)$ is given by the convolution $S^{\mathrm{A}}(\sigma) = 2[S\otimes A^{\mathrm{FT}}](\sigma)$, where the lineshape function A^{FT} is the Fourier transform of $A(\Delta x)$:

$$A^{\mathrm{FT}}(\sigma) = \int_{-\infty}^{\infty} A(\Delta x)\mathrm{e}^{-4\pi\sigma\Delta x}\mathrm{d}(\Delta x) \quad [12]$$

and

$$[f\otimes g](t) \equiv \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\mathrm{d}\tau \quad [13]$$

Discrete Sampling

In order to discuss the effect of the discrete, equally spaced sampling points with a mutual distance of ξ on the output spectrum, Δx is replaced by $x_n = n\xi$ in eqn [10], where n is an integer and runs from $-N_{\max}$ to $N_{\max}$ and $N_{\max}\xi = \Delta x_{\max}$. Instead of the continuous interferogram $D(\Delta x)$, a discrete interferogram:

$$D_n = \frac{R_D}{2}\int_{-\infty}^{\infty} S(\sigma)\mathrm{e}^{\mathrm{i}4\pi n\xi\sigma}\mathrm{d}\sigma \quad [14]$$

results. From eqn [14], it is readily seen (by shifting the integration variable) that replacing $S(\sigma)$ with $S[\sigma + m/(2\xi)]$, where m is an integer, does not change the discrete interferogram D_n. Keeping in mind that according to our definition $S(-\sigma) = S(\sigma)$, it is clear that $S[-\sigma + m/(2\xi)]$ also leads to the same interferogram as $S(\sigma)$. Therefore, to be able to transform the interferogram to the spectral density $S(\sigma)$ for a wavenumber band $[0, \sigma_{\max}]$ unambiguously, one has to make sure, by electronically and/or optical filtering, that no wavenumbers outside this range contribute to the interferogram and that the sampling interval ξ is small enough so that $\sigma_{\max} < -\sigma_{\max} + 1/(2\xi)$, i.e., $\sigma_{\max} < 1/(4\xi)$. In Figure 3, a schematic sketch of a spectral density $S(\sigma)$, consisting of two spectral bands shown by the full and dashed lines, is plotted. For the spectral band shown by the full line, the sampling interval ξ is small enough so that in the interval $[0, 1/(4\xi)]$ no ambiguity arises. For the spectral band shown by the dashed line, the interferogram appears to be under-sampled: ξ is too large, and, therefore, the part of the spectral band at negative wavenumbers shifted by $+1/(2\xi)$ falls into the interval $[0, 1/(4\xi)]$ (shown by the dotted line in Figure 3). Thus, for the sampling interval ξ, the interferogram for the sum of spectral bands, shown by the full and the dashed lines in Figure 3, is the same as for the sum of the spectral bands shown by the full and dotted lines. Since the interferograms are the same, the two spectra are indistinguishable at sampling intervals ξ. This ambiguity of spectra for an under-sampled interferogram is called aliasing. Comparing the correct sampling interval $\xi < 1/(4\sigma_{\max})$ with the period of the interferogram D_n corresponding to a spectrum consisting of a single line at $\sigma_{\max}$ ($D_n = R_{\mathrm{D}}\cos(4\pi n\xi\sigma_{\max})$), it follows that the interferogram has to be sampled more than two times per period in order to obtain an unambiguous result (Nyquist Theorem).

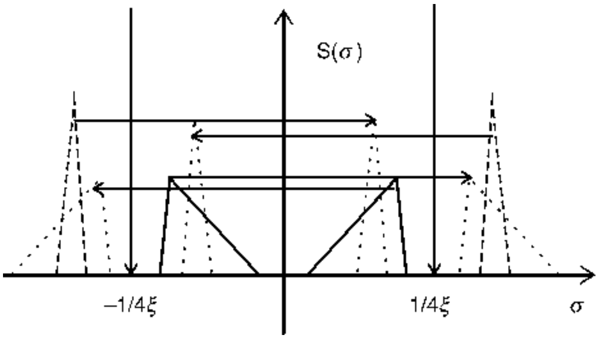

Figure 3 Schematic sketch of aliasing for under-sampled interferograms. For spectral density shown by the full line, the sampling interval ξ is chosen correctly and the $1/(2\xi)$ wavenumber ambiguity indicated by the arrows in the plot does not result in erroneous features in the spectral range $[0, 1/(4\xi)]$. The situation is different for the spectral band sketched by the dashed line: For this band, the sampling interval ξ is too large, and consequently, the $1/(2\xi)$ wavenumber ambiguity maps the line from the negative wavenumber region into the range $[0, 1/(4\xi)]$. The mapped spectral band is indicated by the dotted line.

Mathematically, the spectral density is calculated from the discrete interferogram D_n by a discrete Fourier transformation:

$$S_m = \frac{2}{R_{\mathrm{D}}N}\sum_{n}^{N} D_n \mathrm{e}^{-2\pi\,\mathrm{i}nm/N} \qquad [15]$$

where N is the number of sampling points ($N = 2N_{\max} + 1$) and the wavenumbers corresponding to S_m are given by $\sigma_m = m/2\xi N$.

In order to obtain a high accuracy in the sampling intervals ξ, usually a reference interferogram of a monochromatic laser line (for example from a HeNe laser) is measured simultaneously with the interferogram of the actual experiment. By counting, for example, the zeros with positive slope in the interferogram of the laser line, uniformly spaced sampling points can be achieved with high accuracy. If, in an experiment many interferograms have to be averaged, the absolute position of the sampling points is important, in order to ensure that only points of equal mirror displacement are averaged. Therefore, in some spectrometers, the interferogram of a broadband source is measured simultaneously with the interferograms of the laser and the experiment and the maximum of the interferogram of the broadband source is taken as the origin for counting the zeros of the laser interferogram.

Phase Correction

Ideally, the interferogram is expected to be symmetrically around the position of equal optical path in the two arms of the interferometer. However, in practise this is seldom the case. Many reasons can lead to an asymmetric interferogram, the most obvious being that none of the discrete sampling points of the interferogram coincides with the mirror position for equal optical path in the two interferometer arms. Let Δx_{m} be the measured mirror displacement that differs from the real mirror displacement Δx by x_0, i.e., $\Delta x = \Delta x_{\mathrm{m}} + x_0$. In that case the measured interferogram is:

$$D(\Delta x_{\mathrm{m}}) = \frac{R_{\mathrm{D}}}{2}\int_{-\infty}^{\infty} S(\sigma)\mathrm{e}^{\mathrm{i}[4\pi\sigma(\Delta x_{\mathrm{m}} + x_0)]}\mathrm{d}\sigma \qquad [16]$$

and the Fourier transformation, according to eqn [10] with Δx_{m} as the integration variable, results in

$S(\sigma)e^{i4\pi x_0\sigma}$, that is, in a spectrum with both real and imaginary parts. Also, mirror or compensator misalignment can result in a wavenumber-dependent optical path at a fixed mirror position. Finally, for rapid-scan interferometers, in that the movable mirror is scanned at a constant velocity v, the electronic filtering and amplification of the detector signal will also add different phases, depending on the signal frequency ν that is correlated to the wavenumber σ by $\nu = 2\sigma v$. In general, as long as the phase φ depends only on the wavenumber and not on the mirror displacement, the measured interferogram is:

$$D(\Delta x_m) = \frac{R_D}{2}\int_{-\infty}^{\infty} S(\sigma)e^{i[4\pi\sigma\Delta x_m + \varphi(\sigma)]}d\sigma \qquad [17]$$

and by the Fourier transformation, according to eqn [10], one obtains $S(\sigma)e^{i\varphi(\sigma)}$. There are several methods implemented in the software of commercially available Fourier spectrometers to eliminate the phase from the Fourier transform. In most experimental cases, in that the measured interferograms correspond to spectra with $S(\sigma) \geq 0$ for all $\sigma \in [0, \sigma_{max}]$, these methods usually give the correct results. However, care has to be taken if spectra contain positive and negative components. This can be the case, for example, if the transmission of a sample is periodically modulated by externally applied voltage or a laser light and the detector signal is amplified by a lock-in technique before it is fed into the electronics of the Fourier transform spectrometer. If the modulation increases the sample absorbance in one part of the spectral range and decreases it in a different part, then the measured interferogram corresponds to a spectrum with both negative and positive regions. In the following, first the phase correction methods for spectra with positive values only are discussed before it is argued, why these methods might fail for spectra with positive and negative parts.

For spectra with only positive values, the most simple method to eliminate the unknown phase is to measure the interferogram symmetrically around $\Delta x_m = 0$ and calculate $S(\sigma)e^{i\varphi(\sigma)}$ according to eqn [10]. Including the effect of finite scan length, the output spectrum becomes $S^A = 2[S(\sigma)e^{i\varphi(\sigma)}]\otimes A(\sigma)^{FT}$. Under normal experimental conditions, the phase $\varphi(\sigma)$ does not vary over the peak FWHM of $A(\sigma)^{FT}$ and, therefore, the phase term can be put outside the integral of the convolution and $S^A = 2e^{i\varphi(\sigma)}[S(\sigma)\otimes A(\sigma)^{FT}]$ results. Since $A(\Delta x_m) \geq 0$, and chosen to be symmetric around $\Delta x_m = 0$, the convolution is real and positive and the true spectrum with instrumental resolution $2[S(\sigma)\otimes A(\sigma)^{FT}]$ can be obtained by calculating the modulus of S^A without any knowledge about $\varphi(\sigma)$. However, for several reasons this method for phase correction is not ideal. First of all, since the maximum spectral resolution of a Fourier transform spectrometer is inversely proportional to the maximum mirror displacement, for a given spectrometer layout with a fixed maximal mirror scan length L, the maximal achievable spectral resolution is reduced by a factor of two if scans symmetric around $\Delta x = 0$ have to be performed. Secondly, for signals closer to or smaller than the noise level of the spectrometer, the random noise will be entirely positive by taking the modulus. Hence, for small signals, the noise level will be increased relative to the signal level by taking the modulus.

Therefore, usually the 'Mertz phase correction method' is used. This method is based on the assumption that the phase $\varphi(\sigma)$ is a slowly varying function of σ. In this case, $\varphi(\sigma)$ can be determined with low spectral resolution, i.e., by a short, symmetric scan of length $2L_1$ around $\Delta x = 0$. It can be shown that from this scan the phase can be obtained according to:

$$\varphi(\sigma) = \arctan\left\{\frac{\mathrm{Im}[S(\sigma)e^{i\varphi(\sigma)}\otimes A^{FT}(\sigma)]}{\mathrm{Re}[S(\sigma)e^{i\varphi(\sigma)}\otimes A^{FT}(\sigma)]}\right\} + \pi\,(\text{if denominator is negative}) \qquad [18]$$

Here, $S(\sigma)e^{i\varphi(\sigma)}\otimes A(\sigma)^{FT}$ results from the Fourier transformation of the short interferogram measured between $-L_1$ and L_1 and the apodization function $A(\Delta x_m)$ is adjusted to $-L_1$ and L_1. The addition of π for negative denominators is a consequence of the assumption that $S(\sigma)$ is positive everywhere. Under this assumption, a negative denominator in eqn [18] corresponds to a phase angle $|\varphi(\sigma)| > \pi/2$, that is mapped into $[-\pi/2, \pi/2]$ by the arctan-function. To get the correct phase angle, π has to be added to the phase angle, thereby extending the range of $\varphi(\sigma)$ to $[0, 2\pi]$.

With $\varphi(\sigma)$ known, the interferogram has to be recorded only for mirror displacements $\Delta x_m \geq 0$, since the information about the interferogram for $\Delta x_m < 0$ is contained in the phase spectrum and the interferogram for $\Delta x_m \geq 0$. Therefore, with this method, the totally available mirror displacement length can be used for increasing the spectral resolution.

However, for spectra containing positive and negative spectral bands, the Mertz phase correction usually fails, due to the addition of π if the denominator in eqn [18] becomes negative. For spectra with positive and negative bands, the sign of the denominator is determined by both $S(\sigma)$ and $\varphi(\sigma)$. Applying the Mertz algorithm in a negative region of the spectrum with a vanishing physical phase angle, results in a calculated phase angle of $\varphi = \pi$. As a consequence, the negative spectral band appears as a positive band in the output spectrum calculated according to the Mertz algorithm.

To avoid erroneous phase-flips, the addition of π in eqn [18] for negative denominators can be omitted and the allowed values for $\varphi(\sigma)$ are reduced to $[-\pi/2, \pi/2]$. The corresponding phase correction method is usually referred to as the 'Mertz signed method'. However, it is immediately clear from the discussion above, that in this case any physical phase angle outside the range $[-\pi/2, \pi/2]$ results to an erroneous π-jump of the phase. According to eqn [16], a physical phase shift larger than $\pi/2$ can, for example, be the consequence of a large offset x_0, i.e., of an inaccurate alignment of Δx_m and Δx. At $\Delta x = 0$, all contributions to the interferogram add constructively and the height of the interferogram is proportional to the area under the spectral curve $S(\sigma)$, as can be readily seen by setting $\Delta x = 0$ in eqn [10]. Therefore, for $S(\sigma) > 0$ the position $\Delta x_m = 0$ is usually estimated by the global maximum of the measured interferogram. However, for spectra with both positive and negative bands, in most cases the interferogram does not exhibit a maximum at $\Delta x = 0$ (for positive and negative bands of equal height and width, the interferogram vanishes at $\Delta x = 0$), and estimating $\Delta x_m = 0$ by an extremum of the interferogram results, therefore, in a large offset x_0. Thus, in order to be able to use the 'Mertz signed' algorithm, $\Delta x_m = 0$ has to be determined separately in a calibration experiment in that the condition $S(\sigma) > 0$ is fulfilled. However, even for accurately determined $\Delta x_m = 0$, erroneous phase jumps can appear if lines of different intensity and opposite sign are spaced approximately equal to the resolution used for determining the phase spectrum $\varphi(\sigma)$.

In conclusion, there are no general strategies to completely exclude phase jumps in the Fourier transformation of interferograms corresponding to spectra with positive and negative bands. The experimentalist has to check from case to case if lines with strange shapes appear in the output spectra, and modify the phase correction accordingly. More elaborate strategies for phase correction, that are beyond the scope of this article, are reported in literature.

Example of Experimental Results

To illustrate the features of FTS discussed in the previous paragraphs by a practical example, in Figure 4 the results of a typical measurement in the MIR between 500 cm^{-1} and 5000 cm^{-1} are shown. The experiments were performed with a BRUKER IFS113 spectrometer using a silicon carbide glowbar as a source for the infrared radiation, a potassium bromide beamsplitter, and a liquid-nitrogen cooled mercury cadmium telluride detector. The spectrometer used for the experiments is set up in a vacuum chamber. The two interferograms shown in Figure 4a were measured at two different pressures (broken line: 10 mbar; full line: 300 mbar), i.e., at two different levels of absorption of the atmospheric gas in the interferometer chamber (for clarity, the interferograms are offset vertically). The abscissa axis is broken in Figure 4a and shows portions of the interferograms around zero mirror displacement and around the maximum mirror displacement of 2.5 cm (chosen for 0.2 cm^{-1} spectral resolution). For higher chamber pressures, the increased absorption of the atmospheric gas results in an increased amplitude of the wiggles of the interferogram apparent for large mirror displacements in Figure 4a. The amplitude of these wiggles is approximately three orders of magnitude smaller than the detector signal observed at the center-burst of the interferogram around zero mirror displacement. However, in the corresponding spectra (Figure 4b) calculated from the interferograms after phase correction, the absorption of the atmospheric gas is clearly evident by a strong increase of the dense and narrow absorption lines in the spectral regions 1200 cm^{-1}–2200 cm^{-1}, 2300 cm^{-1}–2400 cm^{-1}, and 3200 cm^{-1}–4000 cm^{-1} (note that in Figure 4b the spectra are offset vertically by 0.1 for clarity). The lines observed between 2300 cm^{-1}–2400 cm^{-1} are due to absorption of the CO_2 molecule and are shown in more detail in Figure 4c from 2335 cm^{-1} to 2345 cm^{-1}. In order to simulate the contribution of the CO_2 lines to the interferogram shown in Figure 4a, two neighboring lines in Figure 4c were fitted to a doublet of Lorentz-absorption lines (indicated by the symbols in Figure 4c). The corresponding interferogram was calculated by Fourier transformation of the Lorentz lines and is shown in Figure 4d. It consists of a rapid oscillating, cosine-like signal (see inset for an expanded plot between 0.6 and 0.602 cm) with an amplitude modulation corresponding to the difference of the resonance wavenumbers of the two Lorentz lines. The ordinate axes of Figures 4a,d have the same arbitrary units and, therefore, the amplitudes of the interferograms shown in these two panels can be directly compared. From Figure 4a, d we conclude that the two CO_2 lines contribute approximately with only one in 10^5 parts to the center-burst of the measured interferogram, i.e., theses lines absorb only a fraction of 10^{-5} of the integrated (over the wavenumbers) intensity of

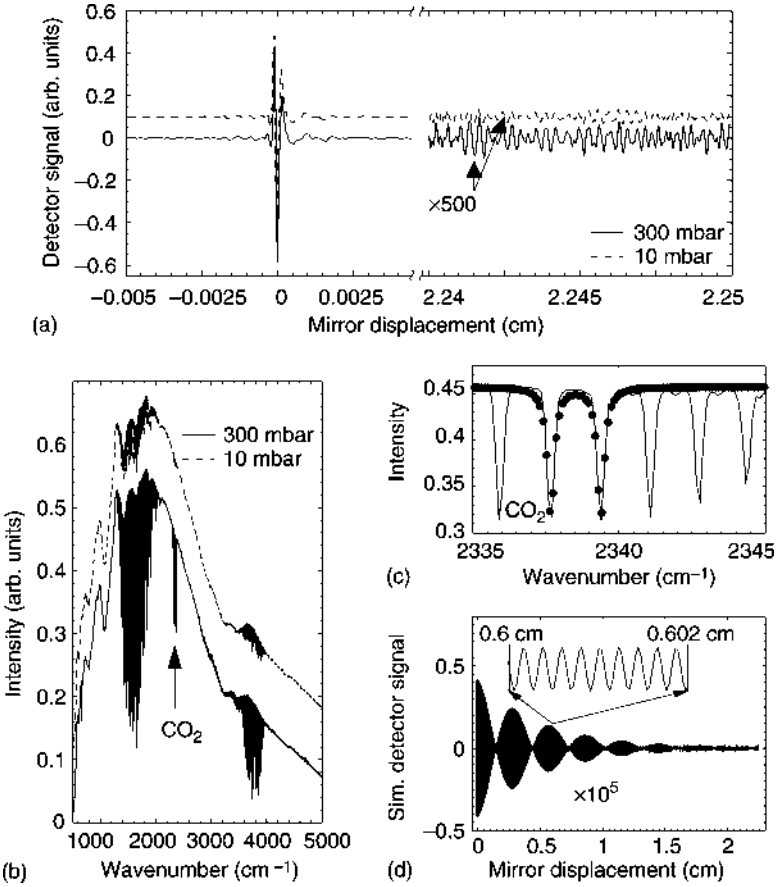

Figure 4 Measurement of the absorption of atmospheric gas for different pressures in the interferometer chamber (broken lines: 10 mbar, full lines: 300 mbar). (a) Interferograms (note the break in the axis of the mirror displacement). For clarity, the interferograms are offset vertically by 0.1 and 0.0002 on the left and right side of the axis break. On the right side of the axis break, the interferograms are amplified by a factor 500. (b) Spectra corresponding to the interferograms shown in (a). The spectra are offset vertically by 0.1 for clarity. (c) Part of the CO_2 absorption lines on an expanded wavenumber axis. The symbols indicate a fit of two Lorentz lines to neighbouring CO_2 absorption lines. The fit is used for calculating the interferogram shown in (d). (d) Simulated interferogram corresponding to the two CO_2 lines shown by the symbols in (c). The arbitrary units of the ordinate axis are the same as those used in (a) but the interferogram is amplified by a factor 10^5. In the inset, the interferogram is shown on an enlarged scale between 0.6 and 0.602 cm mirror displacement.

the incident radiation. Despite this small fraction, they are extremely well resolved with a large signal-to-noise ratio in Figures 4b,c, thus demonstrating the ability of FTS to measure spectra simultaneously with high resolution and over a large spectral range. This ability is one of the advantages (the so-called multiplex advantage) inherent in FTS and will be discussed in more detail in the following paragraph.

Advantage of FTS Over Dispersive Spectrometer

The two main advantages of FTS over dispersive spectroscopic methods, using, for example, a grating for the spectral decomposition of an incoming radiation, are directly connected with the working principle of FT spectrometers; high optical throughput and multiplexed measurement over the whole spectral range of interest. These advantages are discussed in the following, in more detail.

For a fair comparison of the optical throughput, spectrometers with equal spectral resolving power ρ have to be compared. Here, $\rho = \sigma_{\max}/\Delta\sigma$, where $\sigma_{\max}$ is the maximum wavenumber analyzed and $\Delta\sigma$ is the spectral resolution. For an ideal Fourier transform spectrometer, the limit for $\sigma_{\max}$ depends on the divergence angle of the beams propagating in the spectrometer. As sketched in Figure 5, the divergence angle α is determined by the radius r_a of the entrance aperture and the focal length f of the collimating lens by $\alpha \approx r_a/f$. This angle causes a phase shift between the rays propagating parallel to and those propagating under the angle of α inclined to the optical axis

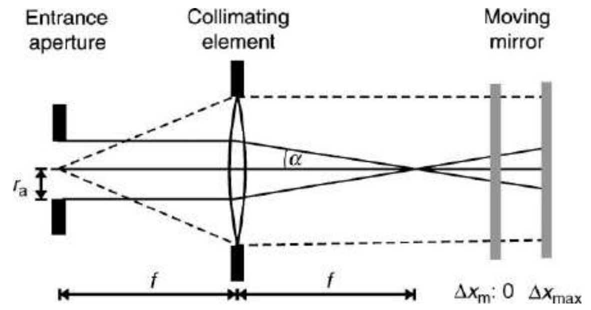

Figure 5 Entrance aperture and collimating element for an interferometer. For simplicity, the beamsplitter and fixed mirror are omitted in the sketch. The extreme rays entering the interferometer are indicated by the full and broken lines. Due to divergence α, a monochromatic light with wavenumber σ_0 produces all periods in the range $[4\pi\sigma_0, 4\pi\sigma_0/\cos(\alpha)]$ in the interferogram.

through the interferometer. The phase difference φ_{div} depends on the mirror displacement according to $\varphi_{\mathrm{div}} = 4\pi\sigma\Delta x((1/\cos(\alpha)) - 1)$ and can be approximated for small α by $\varphi_{\mathrm{div}} \approx 2\pi\sigma\Delta x\alpha^2$. It can be shown, that as long as φ_{div} does not exceed approximately π, the influence of the divergence on the interferogram can be tolerated, i.e., $\sigma_{\max} \approx 1/(2\Delta x_{\max}\alpha^2)$. Using $\Delta\sigma \approx 1/(2\Delta x_{\max})$ for the resolution, one obtains for the resolving power of a FT spectrometer, $\rho = (f/r_a)^2$.

The optical throughput Θ describes the allowable energy per unit time that the system can let through and is related to area πr_a^2 of the aperture and the solid angle Ω of the collimating (or focusing) optics by $\Theta = \pi r_a^2\Omega$. Using $\Omega = A_{\mathrm{FTS}}/f^2$, where A_{FTS} is the clear area of the collimating lens (for a collimating mirror, A_{FTS} is the area of the mirror projected on a plane perpendicular to the beam direction), results in $\Theta_{\mathrm{FTS}} = \pi A_{\mathrm{FTS}} r_a^2/f^2 = \pi A_{\mathrm{FTS}}/\rho$. For a grating spectrometer, $\Theta_{\mathrm{grating}} = hA_G/f\rho$. (see respective chapter of this encyclopedia) where h is the height of the slit, f the focal length of the collimating optics, and A_G the projected area of the grating. Generally, h/f is smaller than 1/20 in up-to-date grating spectrometers. Thus, for the same resolving power ρ and similar instrument size, FT spectrometer can offer a more than $\pi f/h \approx 60$ times larger energy gathering capability. Therefore, for measurements with weak signals in that the detector noise is the dominant noise source, the large throughput of a FTS greatly increases the spectral signal to noise ratio that can be achieved for a fixed measuring time.

The other major principal advantage compared to dispersive methods relies on the simultaneous (multiplexed) gathering of information about all spectral bands $\Delta\sigma$ over a broad spectral range from $\sigma_{\min}$ and $\sigma_{\max}$. For a grating spectrometer using a single detector element, the measurement time T_m available for recording the spectrum in that range has to be distributed between the $M = (\sigma_{\max} - \sigma_{\min})/\Delta\sigma$ spectral bands so that for a single spectral band the available integration time is T_m/M and the integrated signal increases proportional to T_m/M. If the dominant noise source is the detector noise which is independent of the signal level, then the noise will be proportional to $(T_m/M)^{1/2}$ and, therefore, the signal to noise ratio for the grating spectrometer $(S/N)_G$ is:

$$(S/N)_G \propto (T_m/M)^{1/2} \qquad [19]$$

For a FTS the situation is different, since it detects in the band $[\sigma_{\min}, \sigma_{\max}]$ all M small bands over the whole measurement time T_m via their contribution to the interferogram. So the integrated signal in a small band $\Delta\sigma$ is proportional to T_m. If the noise is again assumed to be random and independent of the signal level, for the FTS:

$$(S/N)_{\mathrm{FTS}} \propto T_m^{1/2} \qquad [20]$$

results. For the same detector, assumed in the grating spectrometer and in the FTS, the proportionality constants are the same in eqns [19] and [20] and, therefore

$$\frac{(S/N)_{\mathrm{FTS}}}{(S/N)_G} = M^{1/2} \qquad [21]$$

Equation [21] shows that the multiplex advantage becomes increasingly important for measurements of broad spectral bands with high resolution. In this case, an enhancement of the $(S/N)_{\mathrm{FTS}}$ over $(S/N)_G$ by 2–3 orders of magnitude can be achieved. However, it has to be noted that the multiplex advantage of a FTS can be exploited only in cases where the detector noise is the dominant source of noise. In the visible and near infrared spectral region, low noise detectors, that allow single photon detection, are available. With these detectors, the spectral noise is dominated by the photon noise that is proportional to the square root of the intensity. It is evident that in this case the multiplexed detection is not an advantage, since for M spectral bands of width $\Delta\sigma$ contributing to the interferogram at a given mirror position, the noise is enhanced by $M^{1/2}$ compared to the noise that would be measured for a single band (here, it is assumed that all M bands contribute with the same intensity to the interferogram). Consequently, in this case, the noise enhancement by $M^{1/2}$ cancels the beneficial factor $M^{1/2}$ in eqn [21] and with single-detector grating and FT spectrometers, the same S/N ratio is achieved for a given measurement time T_m. In addition, using linear detector arrays with dispersive spectrometers, the multiplex advantage increases the S/N ratio for the dispersive spectrometer. However, detector arrays and low-noise detectors are mainly available for spectral regions with $\sigma > 6000\ \mathrm{cm}^{-1}$, i.e., in the near-infrared and visible spectral region. In the FIR and MIR region ($\sigma < 6000\ \mathrm{cm}^{-1}$) in that such detectors are absent, the multiplexing inherent in FTS makes it the method of choice if low-noise, broadband, and high-resolution spectroscopic data are required.

List of Units and Nomenclature

Angle of incidence for radiation on beamsplitter	Θ_1
Angle of propagation of radiation in the beamsplitter layer	Θ_2
Beam divergence angle	α
Beam splitter thickness (cm)	d
Complex reflection coefficient	R, r
Complex transmission coefficient	T, t
Detector Signal normalized to detector area (A cm^{-2})	D
Difference between real and measured mirror displacement (cm)	x_0
Dirac delta-function	δ
Discrete interferogram (A cm^{-2})	D_n
Discrete spectral density (W cm^{-1})	S_m
Distance from beamsplitter to fixed mirror (cm)	x_f
Distance from beamsplitter to movable mirror (cm)	x_m
Electric field amplitude of incident radiation (V/m)	E_0
Electric field amplitude of radiation falling back on source (V/m)	E_S
Electric field amplitude of radiation reaching detector (V/m)	E_D
Focal length (cm)	f
Fourier transform of apodization function *A* (cm)	A^FT
General apodization function	A
Heavyside step function	θ
Intensity falling back on source (W cm^{-2})	I_S
Intensity on detector (W cm^{-2})	I_D
Maximum mirror displacement (cm)	Δx_max
Maximum wavenumber (cm^{-1})	σ_max
Measured mirror displacement (cm)	Δx_m
Measured single line spectral density for a finite scan length (W cm^{-1})	S_1^box
Measured single line spectral density for a finite scan length using the apodization function *A* (W cm^{-1})	S^A
Measurement time for complete spectrum (s)	T_m
Minimum wavenumber (cm^{-1})	σ_min
Mirror displacement	Δx
Mirror velocity (cm s^{-1})	v
Modulation frequency (s^{-1})	ν
Number of sampling points of the interferogram	N
Number of spectral bands	M
Optical throughput (cm^2)	Θ
Optical throughput for FTS (cm^2)	Θ_FTS
Optical throughput for grating spectrometer (cm^2)	Θ_grating
Phase accumulation due to optical path to fixed mirror	φ_f
Phase accumulation due to optical path to moveable mirror	φ_m
Phase angle	γ
Phase angle	φ
Phase difference	$\Delta\varphi$
Phase difference between beams reflected at and transmitted through beamsplitter	$\Delta\varphi_\mathrm{RT}$
Phase difference due to beam divergence	φ_div
Phase jump for reflection at beamsplitter	φ_R
Phase jump for transmission through beamsplitter	φ_T
Projected area of collimating element of FTS (cm^2)	A_FTS
Projected area of grating (cm^2)	A_G
Radius of entrance aperture (cm)	r_a
Refractive index	n
Responsivity (A W^{-1})	R_D
Sampling interval (cm)	ξ
Signal to noise ratio for FTS	$(\mathrm{S/N})_\mathrm{FTS}$
Signal to noise ratio for grating spectrometer	$(\mathrm{S/N})_\mathrm{G}$
Single line spectral density (W cm^{-1})	S_1
Solid angle	Ω
Spectral bandwidth (cm^{-1})	$\Delta\sigma$
Spectral density (W cm^{-1})	S
Spectral resolving power	ρ
Wave length (μm)	λ
Wave number (cm^{-1})	σ, σ_0
Wave number resolution (cm^{-1})	$\Delta\sigma_\mathrm{min}$

See also

Interferometry: Overview; Phase Measurement Interferometry.

Further Reading

Bell RJ (1972) *Introductory Fourier Transform Spectroscopy*. New York, London: Academic Press.

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Griffiths PR (1975) *Chemical Infrared Fourier Transform Spectroscopy*. New York, London, Sydney, Toronto: John Wiley & Sons.
Hecht E (1987) *Optics*. Reading, MA: Addison-Wesley.
Mertz L (1965) *Transformations in Optics*. New York, London, Sydney: John Wiley & Sons.
Saptari V (2004) *Fourier-Transform Spectroscopy Instrumentation Engineering*. Bellingham: SPIE Press.

Hadamard Spectroscopy and Imaging

R A DeVerse, Plain Sight Systems, Inc.,
Kailua-Kona, HI, USA

R M Hammaker, Kansas State University,
Manhattan, KS, USA

W G Fateley, Plain Sight Systems, Inc., Manhattan,
KS, USA

F B Geshwind and A C Coppi, Plain Sight Systems,
Inc., Hamden, CT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Consider the task of measuring the spectra of a source, the spectra of an array of contiguous sources, or simply the broadband irradiance of an array of sources. These measures are commonly known as spectrometry, spectral imaging, and imaging, respectively. The sensitivity at which spectrometric, spectral imaging, and imaging systems can be operated is often dominated by noise associated with the detection system and the ability of the optics employed to deliver sufficient photonic flux to the detector subsystem to overcome this noise limitation. The ideal case is where the optical system can be fashioned in such a way as to deliver greater signal to the detector in order to maximize the signal-to-noise ratio (SNR) for a given finite source irradiance and detector noise.

Let us consider the number of discrete spectral resolution elements of spectrometers, spatio-spectral resolution elements of spectral imagers, and spatial resolution elements of imaging systems as being the N quantities of interest we would like to measure. We wish to measure these N quantities in time T with the greatest accuracy and precision as determined by the maximum SNR possible given a finite source flux and a fixed detector noise floor. One option might be to use a multi-element detection system (multi-channel detection system) with N detectors, so each detector is measuring one of the N quantities for the full time T in order to maximize the available integration time. However, in practice, these detectors can be problematic or simply not available for many applications when compared to single-element detectors. When a single-element detector is employed in spectrometry, spectral imaging, or imaging, the measurement methodology required becomes N measurements, each for a time T/N, so that each of the N quantities is measured separately or alternatively, one can record N measurements, each for time T/N, for various groups of the N quantities. The measurement procedure for various groups of the N quantities requires an encoding procedure to choose which of the N quantities to include in each of the N measurements and a decoding procedure to recover the values for each of the N individual quantities. When combinations of N quantities are measured, the signal at the single detector element is increased and the procedure is called multiplexing. Multiplexing may be defined as measuring the sum of groups of quantities rather than measuring each quantity separately with the primary goal of increasing SNR. Multiplexing can and does increase SNR when the detection system of the measuring device is operating in the detector noise limit. The detector noise limit is defined as the noise of the detector output that is independent of the amplitude of the signal. Photon-noise-limited detector systems can benefit from multiplexing strategies, however, one must be careful not to realize a multiplex disadvantage when the source of noise is not detector noise limited.

Fourier Transform Multiplexing

In optical spectrometry a common multiplexing method employs the Michelson interferometer in a Fourier transform (FT) spectrometer where, in the best case, as much as half of the spectral energy of the source entering the aperture is incident upon the single detector element for the duration of the measurement. The output of the detector is recorded as an interferogram, which is the detector response as a function of optical retardation (twice the path difference between the arms of the Michelson interferometer). The recorded interferogram is decoded using a Fourier transform algorithm to yield the desired spectrum. FT spectrometers are routinely employed in near-infrared, mid-infrared, and far-infrared spectrometry. In Raman spectrometry, FT spectrometers are used in the near-infrared spectral region typically using an Nd:YAG laser providing an excitation wavelength of 1064 nm. FT techniques are also employed in nuclear magnetic resonance (NMR) spectrometry and mass spectrometry.

Hadamard Transform Multiplexing

Hadamard transform multiplexing in dispersive optical spectrometry can be accomplished by use of an encoding mask at the focal plane of a dispersive spectrometer, that is operable to select different combinations of spectral resolution elements that are allowed to pass to the single detector element for each of the *N* required measurements. The specification of which ones of the *N* quantities are included in each of the *N* measurements of groups of the *N* quantities is called a weighing design. The term weighing design originally referred to a procedure for accurately determining the mass of a number of objects by weighing them in groups rather than one at a time. The optimal weighing designs and corresponding optimal encoding masks are specified by Hadamard mathematics. The Hadamard matrix or **H** matrix has all elements of +1 and −1 and provides the optimal weighing design for a two-pan balance or an optical instrument having two detector elements for the +1 and −1 measures. The Hadamard matrices specify which quantities go to each balance pan or detector element.

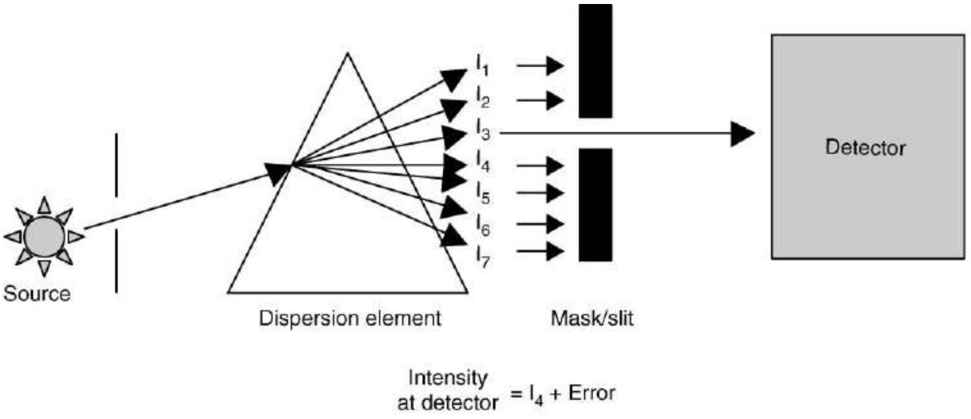

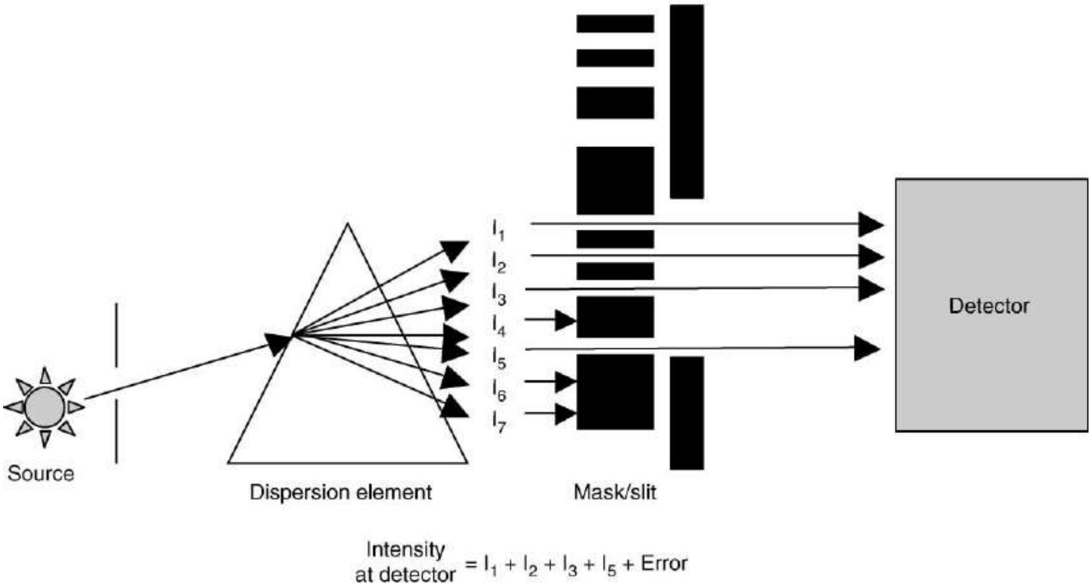

Figure 1 Simple schematic of a conventional and a Hadamard transform spectrometer. The top shows a conventional spectrometer where only one of seven spectral resolution elements ($N = 7$) impinge upon the single element detector for each of *N* measurements. The bottom shows a Hadamard transform spectrometer where $N = 7$. One more than half of the spectral resolution elements impinge upon the single element detector for each of the *N* measurements.

One and Two Detector Hadamard Encodement Matrices

In the case of a dispersive spectrometer with a single detector element, the optimal weighing designs and corresponding optimal encoding masks are provided by the simplex matrix or **S** matrix derived from the **H** matrix. The type of **S** matrix that is most convenient for optical applications is the cyclic or left circulant **S** matrix where only the first row needs to be known to generate the complete cyclic **S** matrix of N columns and N rows. Each row is obtained by shifting the previous row by a left-shift registration in which the element in the far left column moves to the far right column of the next row and each of the other $N-1$ elements moves one column to the left. Each of the N rows of the cyclic **S** matrix specifies one of the N encodements and each of the N columns specifies one of the N spectral resolution elements. The elements of the cyclic **S** matrix are 1 and 0 instead of the +1 and −1 found in the **H** matrix implementation. Spectral resolution elements allowed and not allowed to reach the single detector element are given as 1's and 0's, respectively. Each row and each column of the cyclic **S** matrix is a different sequence of $(N+1)/2$ elements of 1's and $(N-1)/2$ elements 0's so that each encodement in the row of the cyclic **S** matrix has a different combination of $(N+1)/2$ spectral resolution elements that are incident upon the single detector element. Just as the positions of the 1's in a row show which spectral resolution elements are included in a particular encodement, the positions of the 1's in a column show which encodements include that particular spectral resolution element. The result is that provided the detector noise is independent of its output signal, the SNR is increased by a factor of $(N+1)/(2\sqrt{N})$, approximately $(\sqrt{N})/2$, when compared to the method of measuring each spectral resolution element separately as in a raster scan performed by a classical scanning dispersive spectrometer.

Multiplexing in dispersive optical spectrometry as introduced in the previous paragraph is called Hadamard transform spectrometry (HTS) (see Figure 1). It is a unique combination of dispersive and multiplexing spectrometries. The choice of transparent or opaque for the elements of a multislit array or Hadamard encoding mask provides an encoding of spectral information that may be decoded into a conventional spectrum by Hadamard mathematics. In a typical dispersive spectrometer, the source energy is collected and separated into its individual spectral resolution elements by a spectral separator and then collected and focused for spatial presentation on a focal plane. Unlike a scanning dispersive spectrometer, which possesses a single exit slit, the Hadamard transform (HT) spectrometer employs a multislit array as a Hadamard encoding mask. This arrangement allows for the simultaneous measurement of a multitude of spectral resolution elements at a single-element detector and produces a multiplexing spectrometer.

Theory of Hadamard Multiplexing

The operation of the HT spectrometer can be described using matrix algebra with column vectors [**D**] and [**I**] and the N-order square matrices $\mathbf{S}_N$, $\mathbf{S}_N^{-1}$, $\mathbf{S}_N^{\mathrm{T}}$, and $\mathbf{T}_N$. [**D**] is a column vector of N rows containing the detector responses for the N encodements. [**I**] is a column vector of N rows containing the N spectral resolution elements to be determined. $\mathbf{S}_N$ is the cyclic **S** matrix of order N that is the basis for construction and design of the Hadamard encoding mask. $\mathbf{S}_N^{-1}$ is the inverse of $\mathbf{S}_N$ and is given by $(2/(N+1))\mathbf{T}_N$. $\mathbf{T}_N$ is a matrix of +1 and −1

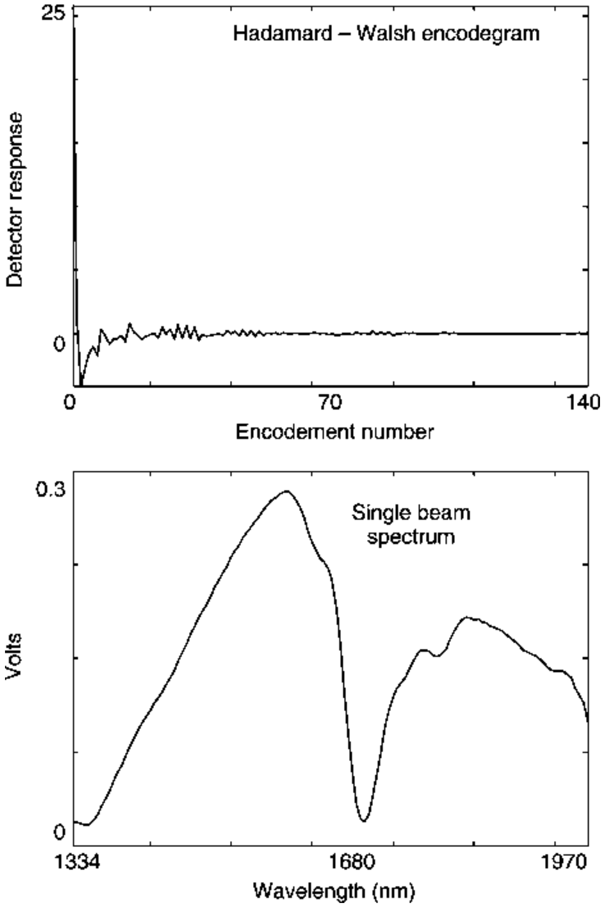

Figure 2 An encodegram (top plot) and the corresponding single-beam spectrum (bottom plot) as a result of decoding the encodegram using Hadamard mathematics.

elements obtained from S_N^T, the transpose of S_N, with T_N having +1's and −1's where S_N^T has 1's and 0's, respectively. A cyclic matrix is required to be symmetric so that an interchange of columns and rows can recover the original matrix and S_N^T then becomes equal to S_N. The encoding of the radiation by the HT spectrometer is given by $[D] = S_N[I]$ and the resulting primary data are a record of detector response versus encodement number. The plot of detector response versus encodement number is called an encodegram (analogous to the interferogram produced by an FT spectrometer). The decoding of the encodegram to recover the conventional spectrum using a Hadamard transform can be described by $[I] = S_N^{-1}[D]$, obtained by left multiplying $[D] = S_N[I]$ by S_N^{-1}. In practice, it is not necessary to invert S_N because there is a fast Hadamard transform (FHT) algorithm available when N is of the form $N = 2^w - 1$, w being an integer. There are three known methods for generating the first row of cyclic S matrices of order $N = 2^w - 1$. Consider the example of $N = 7$ for $w = 3$ where the first row of S_7 is 1 1 1 0 1 0 0. Figure 2 shows an encodegram and the result of decoding using Hadamard mathematics. Figure 3 contains the matrices and matrix equations for $N = 7$.

Hadamard Encoding Masks

The key component of any HT spectrometer is the Hadamard encoding mask (Figure 4). Two choices for the Hadamard encoding mask are the moving mechanical mask and the stationary electro-optical mask. The moving mask for N spectral resolution elements requires $2N - 1$ mask elements that are either completely open with a transmittance of 1 or completely closed with a transmittance of 0. The stationary mask for N spectral resolution elements requires only N mask elements that are selected to be either transparent: transmittance $T_t = 1$ or <1 or opaque: transmittance $T_o = 0$ or >0. The resulting SNR is dependent upon $\Delta T = T_t - T_o$, which has its optimal value of unity only for an ideal mask which is $T_t = 1$, $T_o = 0$. The advantage of the moving mask is

$$[D] = \begin{matrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \\ D_6 \\ D_7 \end{matrix} = \begin{matrix} I_1 + I_2 + I_3 + I_5 \\ I_1 + I_2 + I_4 + I_7 \\ I_1 + I_3 + I_6 + I_7 \\ I_2 + I_5 + I_6 + I_7 \\ I_1 + I_4 + I_5 + I_6 \\ I_3 + I_4 + I_5 + I_7 \\ I_2 + I_3 + I_4 + I_6 \end{matrix} \qquad S_7 = \begin{matrix} 1\,1\,1\,0\,1\,0\,0 \\ 1\,1\,0\,1\,0\,0\,1 \\ 1\,0\,1\,0\,0\,1\,1 \\ 0\,1\,0\,0\,1\,1\,1 \\ 1\,0\,0\,1\,1\,1\,0 \\ 0\,0\,1\,1\,1\,0\,1 \\ 0\,1\,1\,1\,0\,1\,0 \end{matrix} \qquad [I] = \begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \end{matrix}$$

$$T_7 = \begin{matrix} 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & 1 & -1 & 1 & -1 \end{matrix} \qquad S_7^{-1} = (1/4)T_7 = \begin{matrix} (1/4)(1\ 1\ 1 -1\ 1\ -1\ -1) \\ (1/4)(1\ 1 -1\ 1\ -1\ -1\ 1) \\ (1/4)(1 -1\ 1\ -1\ -1\ 1\ 1) \\ (1/4)(-1\ 1\ -1\ -1\ 1\ 1\ 1) \\ (1/4)(1 -1\ -1\ 1\ 1\ 1\ -1) \\ (1/4)(-1\ -1\ 1\ 1\ 1\ -1\ 1) \\ (1/4)(-1\ 1\ 1\ 1\ -1\ 1\ -1) \end{matrix}$$

$$[I] = \begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \end{matrix} = S_7^{-1}[D] = \begin{matrix} (1/4)(D_1 + D_2 + D_3 - D_4 + D_5 - D_6 - D_7) \\ (1/4)(D_1 + D_2 - D_3 + D_4 - D_5 - D_6 + D_7) \\ (1/4)(D_1 - D_2 + D_3 - D_4 - D_5 + D_6 + D_7) \\ (1/4)(-D_1 + D_2 - D_3 - D_4 + D_5 + D_6 + D_7) \\ (1/4)(D_1 - D_2 - D_3 + D_4 + D_5 + D_6 - D_7) \\ (1/4)(-D_1 - D_2 + D_3 + D_4 + D_5 - D_6 + D_7) \\ (1/4)(-D_1 + D_2 + D_3 + D_4 - D_5 + D_6 - D_7) \end{matrix}$$

For example: Note $I_1 = (1/4)\,[\,(D_1 + D_2 + D_3 + D_5) - (D_4 + D_6 + D_7)\,]$ so

$$\begin{aligned} I_1 = (1/4)\,\{\,[\,(I_1 + I_2 + I_3 + I_5) + (I_1 + I_2 + I_4 + I_7) + (I_1 + I_3 + I_6 + I_7) + \\ (I_1 + I_4 + I_5 + I_6)\,] - [\,(I_2 + I_5 + I_6 + I_7) + (I_3 + I_4 + I_5 + I_7) + \\ (I_2 + I_3 + I_4 + I_6)\,]\,\} = (1/4)\,\{\,4I_1\,\} = I_1 \end{aligned}$$

Figure 3 Matrices and matrix equations (note since $\mathbf{S}_7$ is symmetric $\mathbf{S}_7^T = \mathbf{S}_7$).

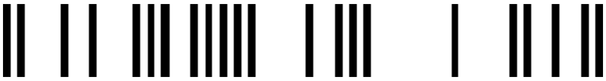

Figure 4 A linear encodement of length $N = 43$ generated by using a quadratic residue construction.

no optical transmission problems. However, it suffers the disadvantage of having moving parts and position repeatability problems. For the case of the stationary mask, it has the advantage that it has no moving parts and no positioning repeatability problems. The disadvantage of the stationary mask is the optical transmission problem where $T_t < 1$ and $T_o > 0$. The moving Hadamard encoding mask of $2N - 1$ mask elements is changed to the next encodement by translation of the Hadamard encoding mask to the left past the aperture by a single mask element. For $N = 7$ and $2N - 1 = 13$, the assignment of 1 and 0 to positions 1 through 13 is [1 1 1 0 1 0 0 1 1 1 0 1 0]. For example, the first encodement corresponding to the first row of S_7 in Figure 3 is from positions 1 through 7 as [1 1 1 0 1 0 0] and the fourth encodement corresponding to the fourth row of S_7 in Figure 3 is from positions 4 through 10 as [0 1 0 0 1 1 1].

2D Encoding Masks for Multiplexing in Spectral Imaging and Imaging

If the conventional one-dimensional (1D) Hadamard encoding mask is folded in some manner, the result is the generation of a two-dimensional (2D) Hadamard encoding mask (Figure 5). Similar to the 1D Hadamard encoding masks that are used for spectral mutiplexing, 2D Hadamard encoding masks can be used for the spatial multiplexing done in Hadamard transform spectral imaging (HTSI) and Hadamard transform imaging (HTI). It is important to note that spectrometry as HTS, spectral imaging as HTSI, and imaging as HTI all realize the same theoretical SNR improvement based on the number of spectral, spatio-spectral, or spatial resolution elements (N), respectively. Consider the folding of a 1D Hadamard encoding mask for $N = 15$ into a 2D Hadamard encoding mask of three rows and five columns (Figure 6). A cyclic S_{15} has a first row of [0 0 0 1 0 0 1 1 0 1 × 0 1 1 1 1] so the 2D Hadamard encoding mask has a first row of [0 0 0 1 0], a second row of [0 1 1 0 1], and a third row of [0 1 1 1 1] from columns 1 through 5, 6 through 10 and 11 through 15, respectively, from S_{15}.

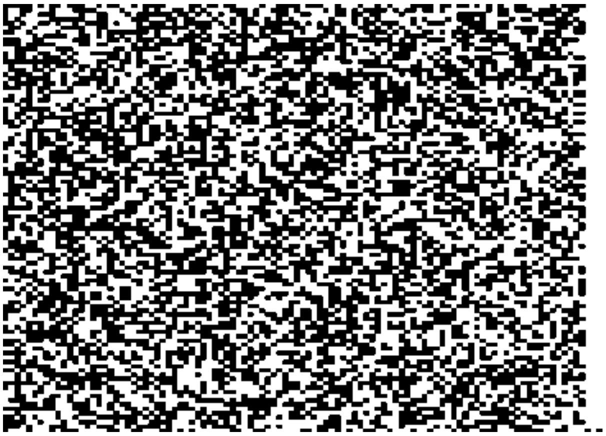

Figure 5 A two dimensional encodement of length $N = 10{,}007$ in a 100 × 100 format generated by quadratic residue construction. The remaining 7 elements are shown as the last 7 elements in the bottom row on the right. In practice these 'extra' data points can be masked off and used as a measure of system noise characteristics.

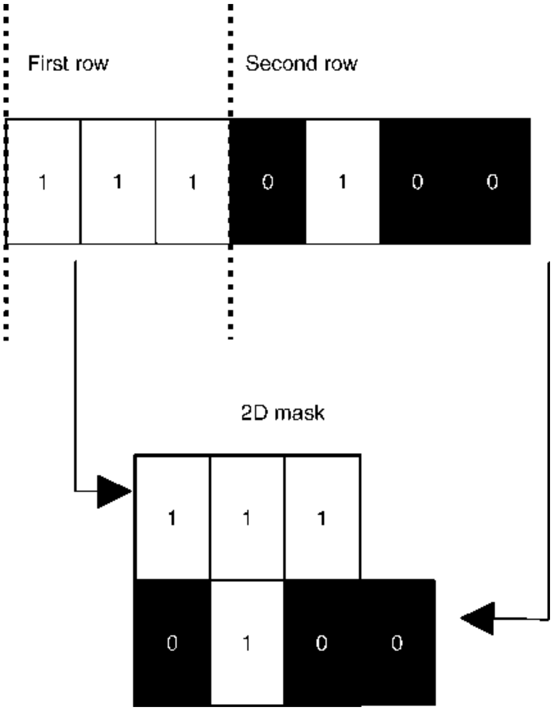

Figure 6 Spectrometry to imagery: The folding of a ID encoding mask that can be used for spectrometry is used to generate a 2D encoding mask that can be used for imaging. Note that using a symmetric aperture window where only 3 × 2 elements are used, the mask element that is not included in the windowed aperture can be used as a measure of the dark noise of the system, however, a more conventional encodement matrix might be one that can be divided into an integer number of rows and columns such as $N = 15$ where one can construct a 5 × 3 2D encodement mask for imagery and use a symmetric aperture that uses the entire length of the encodement matrix.

Hadamard Mathematics

Consider a general multiplexing problem where one seeks to measure N quantities, $X = \{x_0, x_1, \ldots, x_{N-1}\}$, in the presence of noise, and wishes to determine a vector function $F(X)$, such that the measurement of $F(X)$ yields the best estimate of X. Conservation of energy and superposition principles dictate a restriction to linear functions. The principle of rank suggests that we restrict our measurements to N-dimensional linear combinations. Determining the best linear combination is the challenge. In the case of constant error for each measurement, an answer is provided by Hotelling's theorem, which proves Hadamard matrices are optimal for such measurements. Hadamard matrices uniquely achieve the optimal bound for root-mean-square measurement error. A Hadamard matrix is a matrix $\mathbf{H}$ for which $\mathbf{H}_{i,j} = \pm 1$, and $\mathbf{H}^{-1} = \frac{1}{N}\mathbf{H}^{\mathrm{T}}$. The definitions of matrix inverse, transpose, and multiplication dictate that the rows of the matrix $\mathbf{H}$ form an orthogonal basis with each basis element having norm $\sqrt{N}$. The norms of the rows of $\mathbf{H}$ follow from the entries of the matrix that are ± 1, and place no additional restriction on $\mathbf{H}$. The columns of $\mathbf{H}$ are also an orthogonal basis with these same norms. The rows (columns) of the matrix $\frac{1}{\sqrt{N}}\mathbf{H}$ form an orthonormal basis. From these few statements and definitions, one can already derive estimates of the multiplex advantage. For example, one seeks to measure X and in one case measures $Y_1 = X + \varepsilon$, where ε is a random noise or error that has a Gaussian distribution with a mean of 0 and variance σ^2. In another case, suppose that one measures $Y_2 = HX + \varepsilon$. In this second case, one would estimate

$$X_2 = \frac{1}{N}\mathbf{H}^{\mathrm{T}}(\mathbf{H}X + \varepsilon) = X + \left(\frac{1}{\sqrt{N}}\right)^2 \mathbf{H}^{\mathrm{T}}\varepsilon$$
$$= X + \frac{1}{\sqrt{N}}\varepsilon_1$$

where ε_1 is random noise with the same characteristics as ε. It follows that the measurement error in the second case has a standard deviation σ that is less than in the former case, with an improvement by the factor $\frac{1}{\sqrt{N}}$.

Hadamard Computations

The ± 1 Hadamard matrices can be applied to two detector optical systems, and 0,1 matrices can be applied to single detector systems. In a two detector system $\mathbf{H}^{\mathrm{T}}$ is proportional to $\mathbf{H}^{-1}$ and one can invert these transforms. Suppose that one has chosen a particular Hadamard matrix $\mathbf{H}$, with which to conduct a multiplexed measurement. One can show that multiplication of any row or any column by -1 does not affect the efficacy of $\mathbf{H}$ in improving SNR. Hence one may assume that the first row and first column of $\mathbf{H}$ are all $+1$. From $\mathbf{H}$, we can form a new $(N-1) \times (N-1)$ matrix $\mathbf{S}$ by deleting the first row and column of $\mathbf{H}$, and then changing each 1 to a 0, and each -1 to a 1. This new matrix $\mathbf{S}$, an $\mathbf{S}$-matrix, is the one corresponding to $\mathbf{H}$ that is to be used for single-element detector optical systems. It is known that $\mathbf{S}$ matrices are optimal for 0,1 matrices just as Hadamard matrices are optimal for ± 1 matrices, with signal to noise improvement by a factor $\frac{\sqrt{N-1}}{2}$, where here N is the dimension of $\mathbf{H}$.

Letting $\mathbf{I}$ be the identity matrix, and $\mathbf{J}$ be the matrix whose entries are all 1, it follows that $\mathbf{S}^{\mathrm{T}}\mathbf{S} = \frac{N}{4}(\mathbf{I} + \mathbf{J})$, where N is the dimension of $\mathbf{H}$. From this we can ascertain that a multiplexed system using the matrix $\mathbf{S}$, multiplied by the encoded parameters of $\mathbf{S}^{\mathrm{T}}$ recovers the original signal to within the addition of an offset. In practice this is what is done to recover the original X. In certain cases, there is also a fast algorithm for computing $\mathbf{S}^{\mathrm{T}}Y$, without ever explicitly computing $\mathbf{S}^{\mathrm{T}}$.

Constructions for Hadamard Matrices

It is known that if $\mathbf{H}$ is an $N \times N$ Hadamard matrix, then N must be 1, 2, or a multiple of 4 and so nontrivial S-matrices must be of size one less than a multiple of 4. It is conjectured that Hadamard matrices exist for each N which is a multiple of 4. Some constructions of Hadamard matrices are enumerated below.

Walsh–Hadamard Construction

A Hadamard matrix of size $N = 2^K$ can be constructed by sampling the first N Walsh functions at $0, 1/N, 2/N, 3/N, \ldots, (N-1)/N$, where the Walsh functions are defined by $W_0(x) = 1$, for $0 \leq x < 1$, $W_0(x) = 0$ otherwise, and the recurrence relations, for each non-negative integer n:

$$\begin{aligned} W_{2n}(x) &= W_n(2x) + W_n(2x-1), \text{ and} \\ W_{2n+1}(x) &= W_n(2x) - W_n(2x-1) \end{aligned} \qquad [1]$$

Tensor Product Construction

Walsh is a special case of iterating this construction, starting with the matrix $\mathbf{H}_1 = [1]$. In general, if $\mathbf{H}_{\mathrm{N}}$ is a Hadamard matrix of size $N \times N$, then:

$$\mathbf{H}_{2N} = \begin{bmatrix} H_N & H_N \\ H_N & -H_N \end{bmatrix} \qquad [2]$$

is a Hadamard matrix of size $2N \times 2N$.

Quadratic Residue Construction

This construction is an example of a cyclic S-matrix: one for which each row is a cyclic shift of the previous row. Given a prime number $p = 4n + 3$, define, for $i = 0, \ldots, p - 1$, the values $S_{1,i} = 1$ if there is some integer k, such that the remainder of dividing k^2 by p is i, and $S_{1,i} = 0$ otherwise. Taking the cyclic matrix with the first row given by $S_{1,i}$, yields a Hadamard S-matrix.

Other constructions are possible. Maximal length shift register codes, and a construction involving quadratic residues of twin primes are two examples.

Fast Hadamard Transforms

The fast Walsh transform can be deduced from eqn [1]. The basic idea is that, in eqn [1], the function W_n is specified by $\log_2 n$ recurrence relations. For this reason, a Walsh transform (or its transpose) can be computed in $O(n \log_2 n)$ steps. The algorithm goes as follows, for a sequence of length $N = 2^n$:

(i) begin with $Y_{0,0}, Y_{0,1}, \ldots, Y_{0,N-1} = Y_0, Y_1, \ldots, Y_{N-1}$
(ii) for $i = 1, \ldots, n - 1$, let $Y_{i,k} = Y_{i-1,2k} + Y_{i-1,2k+1}$, and $Y_{i,n/2+k} = Y_{i-1,2k} - Y_{i-1,2k+1}$, for $k = 0, \ldots, N/2$
(iii) output $X_i = Y_{n-1,i}$

Then the resulting X_i is the inverse Walsh transform of Y_i.

Note that cyclic S matrices that come from maximal length shift registers are known to be permutations of the Wzalsh–Hadamard S matrices, and hence there is also a fast algorithm for inverting these, albeit involving some tedious permutations.

Optics for Hadamard Encoded Apertures

Hadamard optical systems utilize spatially encoded apertures that can be employed either at the entrance aperture of the optical system, the exit aperture or both. They have the common attribute that they encode the available aperture spatially where the spatial resolution elements that make up the encodement dictate the spectral, spatio-spectral, or spatial resolution elements that propagate through the optical system, including diffractive optical elements and on to the sensor or exit aperture. These masks have some spatial extent that places special requirements on the optics of the system. As the encodement mask grows either by longer length encodements with fixed subapertures or as the subaperture dimension grows for a fixed encodement length, the spatial resolution elements making up the subapertures in the encodement mask depart from the optical axis. When the resolution elements depart from the optical axis or paraxial condition, it becomes important to employ optics that can image the off-axis resolution elements without inducing excessive aberrations that degrade the performance and cripple the advantages gained by HT multiplexing.

Typically the optical path for conventional monochromators begins with a source that is focused onto an aperture plane that has a large aspect ratio aperture know as a slit. This slit is often very small in extent in the dispersion plane compared to the other extent in the spatial plane. However, it is not required that this aspect ratio is large. If the aspect ratio is close to 1 then simple spherical optics can be employed that perform well as long as the departure from the optical axis is kept to a minimum. However, most monochromators have a large aspect ratio in order to increase the opportunity to maximize throughput, and detectors must be able to 'see' the large extent of the slit aperture. The light entering the slit aperture is then dispersed and focused onto an exit slit aperture. Monochromators are only required to perform well on the optical axis and do not typically employ optics that can manage rays that depart from the optical axis in the plane of dispersion as required by HT multiplexing instruments. To employ encoding techniques the optical system is required to utilize optical performance attributes normally found only in imaging and spectral imaging systems. This requirement is driven by the extent of the encoding mask. The extent of the encoding mask is governed by the diffraction limit of the wavelengths within the bandpass, the encodement length N, and the attributes of the optical system.

In a conventional dispersive spectrometer the radiation from a source is collected and separated into its individual spectral resolution elements by a spectral separator such as a diffraction grating or prism and then is collected and focused for spatial

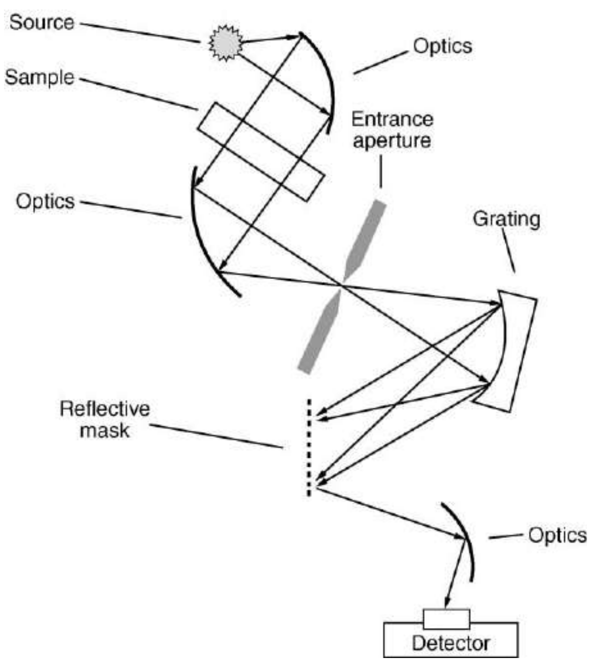

Figure 7 Example implementation of a Hadamard transform spectrometer with reflective optics. Light from a source passes through a sample, is transferred into an entrance aperture, bounced off of a concave grating, and then a reflective mask, and finally transferred to a detector.

presentation on a focal plane. The dispersive spectrometer uses a single exit slit to select one spectral resolution element at any given time, out of N possible spectral resolution elements for measurement by the detector. The Hadamard transform spectrometer (HTS, Figure 7) uses an array of slits (i.e., a mask) at the focal plane to select one more than half, $(N+1)/2$, of the spectral resolution elements at the focal plane for measurement by the detection system. The optical challenge to effect an HT multiplexing spectrometer is to collect all of the spatially distributed individual bandpass images of the entrance slit and transfer them to as small a detector as possible. One desires to keep the area of the detector at a minimum as the noise of many detectors increases with the square of the area. If the optics are able to illuminate a single detector element with all of the available light impinging upon the focal plane containing the spatially distributed images of the slit for each of the N bandpass resolution elements, a multitude of spectral resolution elements can be measured simultaneously using a single detector element. This arrangement results in a multiplexing spectrometer. The recovery of N spectral resolution elements requires measuring the detector response for N different encodements of $(N+1)/2$ open mask elements. The raw data are recorded as the detector response versus encodement number and is called an encodegram. Hadamard transformation of the encodegram yields the spectrum.

The History of Applied Hadamard Multiplexing

The Hadamard transform instruments developed in the 1960s and 1970s employed moving masks. Significant problems such as misalignment and jamming associated with a moving mask led to a reputation of poor reliability and contributed to a dormant period in the development of HTS and HTI. Interest was rekindled in the 1980s using stationary Hadamard encoding masks based on liquid crystal (LC) technology. The first generation 1D stationary Hadamard encoding mask was a cholesteric LC with $N = 127$ mask elements and used polarization as its operating phenomenon. Two parallel polarizers and rotation or lack of rotation of the polarized radiation generated the opaque and transparent states, respectively. The second generation 1D stationary Hadamard encoding mask was fabricated using a polymer dispersed liquid crystal (PDLC) material with $N = 255$ mask elements and used light scattering as its operating phenomenon. The PDLC contained LC droplets dispersed in a polymer matrix whose index of refraction matched the index of refraction in one direction in the birefringent LC droplet. Alignment of the LC droplets optical axis under an applied voltage removed discontinuities in index of refraction at the polymer matrix/LC interface to generate a good transparent state while random orientation of LC droplets in the polymer matrix generated the opaque state from light scattering by the discontinuities in index of refraction at the polymer matrix/LC droplet interface. A 2D stationary Hadamard encoding mask was also based on LC technology. A ferro-electric liquid crystal positioned between a pair of polarizers with perpendicular orientation operated as an electro-optic half-wave plate when a +value of applied voltage rotated the plane of polarization by 90 degrees to produce the transparent state and a −value of applied voltage left the plane of polarization unaltered to produce the opaque state.

Development based on stationary Hadamard encoding masks continued in the 1990s and a 2D moving Hadamard encoding mask was also fabricated and used to perform imaging in the near-infrared and mid-infrared spectral regions. Note that the mid-infrared spectral region is not generally accessible via Hadamard encoding masks based on

LC technology, since any LC material is expected to have strong absorption bands in the mid-infrared spectral region. From the mid-1990s to the present, the stationary Hadamard encoding mask of choice for the visible and near-infrared spectral regions has been the digital micromirror device (DMD), a device based on micro-optoelectromechanical systems (MOEMS) technology and developed by Texas Instruments for projector display applications. One DMD format incorporates 508 800 micromirrors in an 848 column by 600 row array that is 14.4 mm wide by 10.2 mm high. Each individual micromirror is 16 μm square and adjacent micromirrors are separated by a 1 μm gap. The micromirrors are individually addressable and rotate by +10 or −10 degrees about the diagonal axis to produce binary 'on' and 'off' states. The on state has T_t determined by the mirror reflectivity and approaches 1 while the off state approaches $T_o = 0$. The ideal condition of on and off is not realized due to diffraction of the light off of the small and periodic features of the micromirror device. The DMD is an array of spatial resolution elements that may be selected as groups of super-resolution elements or as individual resolution elements consisting of a single micromirror. The DMD resolution elements are realized as spectral resolution elements in the spectrometer with the columns attributed to the frequency or wavelength dimension and the rows attributed to the slit height dimension. The DMD resolution elements are utilized as spatio-spectral resolution elements in the imaging spectrograph with the columns as the frequency or wavelength dimension and the rows as a vertical spatial dimension with the horizontal spatial dimension being accessed, if desired, by translating the sample relative to the imaging spectrograph. The DMD resolution elements are spatial resolution elements in the imager with the columns for the horizontal dimension and the rows for the vertical dimension and the frequency or wavelength dimension provided by other instrumentation. If a photo-acoustic detection system is present then the depth dimension of the sample may also be accessed by changing the modulation frequency used in the photo-acoustic detection system.

Some important features of HTS to keep in mind are:

(i) it is a dispersive technique using a spectral separator;
(ii) it is a multiplexing technique using a single-element detector;
(iii) it uses a Hadamard encoding mask (multislit array) in a focal plane;
(iv) it sends one more than half the resolution elements to the single-element detector in each encodement;
(v) it uses a number of encodements equal to the number of resolution elements desired and the number of mask elements (pixels) in the stationary Hadamard encoding mask (a moving Hadamard encoding mask has $2N - 1$ mask elements);
(vi) it has each encodement containing a different combination of one more than half the resolution elements;
(vii) it has as its primary data the encodegram, a record of detector response versus encodement number; and
(viii) it uses a FHT of the encodegram to decode the encodegram and generate the spectrum or image.

Note that all of these, except for item (i), also apply to HTSI and HTI. The DMD promises to be the best Hadamard encoding mask yet developed for the visible and near-infrared spectral regions. However, its potential applications in spectrometry and imaging are by no means limited to HT techniques since the information corresponding to any micromirror in the DMD may be included in or excluded from any measurement as desired by the investigator. An instrument with no moving parts, other than the micromirrors in the DMD, promises to provide a compact and robust instrument for operation in potentially hostile environments ranging from process control to outer space. It is our belief that the combination of a DMD with a single-element detector may provide an important advance in spectroscopic instrumentation and that instrumentation based on the DMD may lead to a host of environmental, industrial, medical, and military applications.

See also

Fourier Optics. **Imaging:** Multiplex Imaging. **Optical Communication Systems:** Optical Time Division Multiplexing. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Bellamy MK, Mortensen AN, Hammaker RM and Fateley WG (1997) Chemical mapping in the mid- and near-IR spectral regions by Hadamard transform/FT-IR spectrometry. *Applied Spectroscopy* 51: 477–486.

Fateley WG, Hammaker RM, Paukstelis JV, *et al.* (1993) Application of a two-dimensional Hadamard encoding mask for the imaging of thin-layer chromatography plates by laser-induced fluorescence or surface-enhanced Raman scattering and for use with a photoacoustic detector to generate three-dimensional photoacoustic images. *Applied Spectroscopy* 47: 1464–1470.

Fateley WG, Hammaker RM, DeVerse RA, Coifman RR and Geshwind FB (2002) The other spectroscopy: demonstration of a new de-dispersion imaging spectrograph. *Vibrational Spectroscopy* 29: 163–170.

Hammaker RM, Bohlke AP, Freeman RD, *et al.* (1996) Hadamard transform Raman spectrometry. In: Laserna JJ (ed.) *Modern Techniques in Raman Spectroscopy*, pp. 143–226. Chichester: Wiley.

Hammaker RM, DeVerse RA, Asunskis DJ and Fateley WG (2002) Hadamard transform near-infrared spectrometers. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, vol. 1, pp. 453–460. Chichester: Wiley.

Harwit M and Sloane NJA (1979) *Hadamard Transform Optics*. New York: Academic Press.

Nonlinear Laser Spectroscopy

P Ewart, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction: Linear and Nonlinear Spectroscopy

Light propagates in transparent or weakly absorbing media by inducing the atoms or molecules in the material to oscillate in response to the electromagnetic field of the light. The induced dipole oscillations set up a polarization field **P** which radiates the wave in the propagation direction. The induced polarization **P** is proportional to the amplitude of the electric field **E** in the incident light:

$$\mathbf{P} = \varepsilon_0 \chi^{(1)}(\omega)\mathbf{E} \qquad [1]$$

where ε_0 is the permittivity of free space and $\chi^{(1)}(\omega)$ is the first-order, frequency-dependent, susceptibility of the material. This interaction of the light with the medium generally results in some absorption of energy from the field and a shift in the phase of the propagating light. Both the phase delay and the absorption depend on the frequency ω (and wavelength λ) of the light. The phase shifting is responsible for dispersion – the variation, as a function of frequency, of the speed (phase velocity) of light in the medium. The dispersion and absorption effects are described by the real and imaginary parts, respectively, of the complex susceptibility:

$$\chi(\omega) = \chi_{\mathrm{Re}}(\omega) + \mathrm{i}\chi_{\mathrm{Im}}(\omega) \qquad [2]$$

This susceptibility parameter describes the optical properties of the medium. Linear spectroscopy uses the frequency-dependent property of the $\chi(\omega)$ to create spectra by absorption or dispersion of specific frequencies that are resonant with transitions between internal energy states of the molecules. The electric field in the incident light is considered to be weak when it leaves the atomic and molecular properties essentially unaffected. Nonlinear optical spectroscopy, however, is based on the ability of intense laser light to modify the optical properties of the medium through which it propagates. These effects are also frequency dependent and so can be used to generate spectra. When the strength of the electric field becomes comparable with the electric fields within the atoms or molecules (typically $\sim 10^{11}\ \mathrm{V\,m^{-1}}$) then the induced polarization becomes nonlinear and is described by a power series expansion:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega)\mathbf{E}_1 + \chi^{(2)}(\omega)\mathbf{E}_1\mathbf{E}_2 + \chi^{(3)}(\omega)\mathbf{E}_1\mathbf{E}_2\mathbf{E}_3 + \text{higher-order terms}] \qquad [3]$$

$\chi^{(2)}(\omega)$ and $\chi^{(3)}(\omega)$ are the second- and third-order susceptibilities respectively. The real and imaginary parts of these susceptibilities (and of the higher-order terms) describe observable effects arising from absorption and dispersion. The associated interactions involve two and three photon processes, respectively. In general, nonlinear interactions involve multiphoton processes and will be described by the appropriate term in the expansion of the induced polarization.

Broadly speaking, nonlinear spectroscopy is based on detection of these multiphoton interactions by either incoherent or coherent processes. Incoherent processes include fluorescence and ionization where, following the multiphoton interaction, photons or electrons respectively are emitted spontaneously in all directions. Coherent processes, on the other hand, arise from the coupling of the input driving fields via the nonlinear polarization of the medium.

The waveform of the polarization oscillations is a distorted form of the driving field. Fourier analysis of these nonlinear oscillations would reveal extra frequencies not present in the incident field. When a suitably phased array of dipoles is established, the nonlinear optical process is detected by emission of coherent directional light. The efficiency with which this coherent signal beam is generated is crucially dependent upon the correct phasing of the radiating molecules. This process is known as phase matching and is achieved by geometrical arrangements of the input beams or by modifying the refractive index to counteract dispersion between the generating polarization wave and the radiated signal wave.

Apart from the general division of nonlinear spectroscopy according to whether the detection uses incoherent or coherent emission, the techniques are characterized as frequency or time domain methods, according to how the signal is generated and detected. In most cases the signals recorded in either domain may be related by a Fourier transformation to an equivalent signal in the other.

The advantages of nonlinear spectroscopic techniques lie in their ability to provide high spectral, temporal, and spatial resolution. In particular, it is possible, using a variety of nonlinear optical processes, to eliminate inhomogeneous broadening of the spectral lines recorded. Second, when coherent signal beams are generated they can be readily distinguished from incoherent background emissions, thus permitting high signal-to-noise ratios. Third, the use of short laser pulses, down to the order of 10^{-15} s, allows high temporal resolution. High spatial resolution is achieved by generating signals from interaction regions defined by crossed and focused beams. Although the coefficients of successive terms in the power series expansion of the nonlinear polarization decrease by many orders of magnitude with increasing order, signal strengths may be made comparable to those of linear spectroscopy. Strong signals are obtained by the use of resonant enhancement of the nonlinear susceptibility, high field strengths from laser sources, and the use of sensitive photodetectors.

Second-Order Nonlinear Effects

Weak light fields leave unaffected the optical properties of the medium through which they propagate. When, however, the second-order susceptibility $\chi^{(2)}$ contributes significantly to the induced polarization, the complex refractive index includes a field-dependent term:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega) + \chi^{(2)}(\omega)\mathbf{E}_1]\mathbf{E}_2 \qquad [4]$$

Owing to the symmetry properties of $\chi^{(2)}$, second-order nonlinear effects are absent in isotropic media or those possessing a center of inversion symmetry. They may, however, be observed in anisotropic materials such as crystalline solids where they are the basis of frequency mixing schemes leading to second-harmonic generation, sum- and difference-frequency generation, and optical parametric amplification and oscillation. The symmetry of gas phase media, such as atomic vapors, may be broken by application of an external electric or magnetic field. Resonantly enhanced second-harmonic generation has been achieved in alkali metal vapors in the presence of a magnetic field giving potential for spectroscopy. Symmetry breaking is achieved also on solid surfaces where sum-frequency generation has been used to study surface morphology, monatomic layers, and catalytic effects of adsorbed materials.

Third-Order Nonlinear Spectroscopy

In gas phase or isotropic solid and liquid media, the first nonlinear term is of third order and leads to a refractive index that is dependent on the field intensity $|\mathbf{E}|^2$:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega) + \chi^{(3)}(\omega)\mathbf{E}^2]\mathbf{E} \qquad [5]$$

Intensity-dependent refractive index effects lead to a wide range of nonlinear optical phenomena. In general, the polarization contains terms that are the product of three input fields generating Fourier components at sum and difference frequencies. Phase matching constraints will generally determine which of the possible interactions will be observed. A wide range of physical effects contribute to $\chi^{(3)}(\omega)$. The simplest of these nonlinear effects is saturation of the absorption.

Linear absorption is defined by the situation where the intensity decays exponentially with propagation distance x, according to Beer's Law:

$$I(x) = I(0)\exp\{-\alpha_0(\omega)x\} \qquad [6]$$

where $I(0)$ and $I(x)$ are the intensity values at $x = 0$ and x, respectively, and $\alpha_0(\omega)$ is the weak field absorption coefficient. The fractional power absorbed in this case is, therefore:

$$[I(0) - I(x)]/I(0) = \alpha_0(\omega)x \qquad [7]$$

and is independent of the intensity. When the incident intensity increases, until the rate of stimulated absorption from the lower level equals the spontaneous emission rate out of the upper level, the medium is said to be saturated. The intensity is then I_{sat}. A simple rate-equation analysis shows that the absorption coefficient in the presence of a strong light field is:

$$\alpha(\omega) = \frac{\alpha_0(\omega)}{[1 + I/I_{\text{sat}}]} \qquad [8]$$

The absorption is then no longer independent of the light intensity.

Saturation Spectroscopy

In the case of a homogeneously broadened transition the effect of a saturating monochromatic light field is to reduce the absorption across the entire line profile. This is shown in Figure 1a where the frequency of the incident light w_L is detuned from line center ω_0. The degree of saturation and the effect on the absorption coefficient depends on the intensity I and the detuning $\Delta\omega = \omega_L - \omega_0$.

For a transition dominated by inhomogeneous broadening a monochromatic light field interacts strongly only with a sub-group of atoms. This subgroup is determined by those having a resonance frequency within the spectral range covered by the homogeneous width and centered on the incident frequency. Saturation in this case results in 'hole burning' as shown in Figure 1b. A dip in the absorption coefficient $\alpha(\omega)$ is produced centered on ω_L of width given by the homogeneous width.

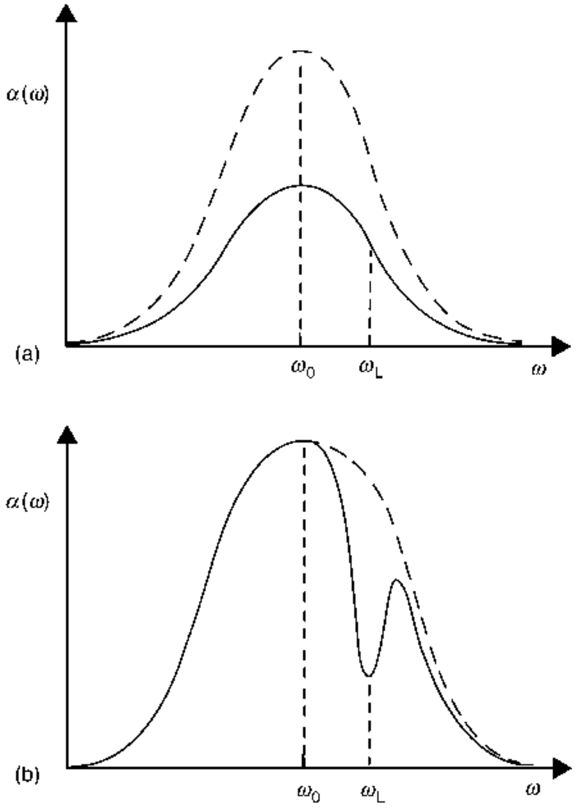

Figure 1 Absorption lineshapes for (a) homogeneously broadened and (b) inhomogeneously broadened lines centered at ω_0. Dashed lines in both figures show lineshape measured by weak probe. Solid lines show lineshape measured by weak probe in presence of a monochromatic, saturating beam at frequency $\omega_L \neq \omega_0$.

Saturation affects the population of both upper and lower levels involved in an optical transition and so affects both absorption and amplification (gain). The phenomenon shows itself as a narrow dip in the spectral profile of lasers operating on an inhomogeneously broadened gain line (e.g., a HeNe gas laser). These 'Lamb dips', as they are called, are the result of saturation by the laser mode which has a very narrow spectral width. Only those atoms in the Maxwell–Boltzmann velocity distribution are saturated, whose velocity results in a Doppler shift of the light that brings it into resonance, i.e., to within the atom's homogeneous linewidth of the resonance frequency. When the laser mode frequency is detuned from atomic line center two velocity subgroups of atoms, within the distribution, are saturated corresponding to positive and negative Doppler shifts for counter-propagating fields of the mode. When the mode frequency is tuned to atomic line center, the counter-propagating fields interact with the same (zero) velocity subgroup. The saturation effect is thereby enhanced, reducing the gain by depletion of the population inversion, leading to a dip in output intensity at line center. The width of the dip is the homogeneous width of the transition (Figure 2).

The variation of output intensity around the Lamb dip is used to provide an error signal in one scheme for laser frequency stabilization. A feedback system is employed to adjust the cavity length, L, in order to maintain the mode frequency, ν, exactly on atomic line-center ($\nu = mc/2L$ where m is an integer). This scheme uses gain saturation of the amplifying medium, whereas an alternative uses saturated absorption. In this scheme, a cell of gas having a narrow absorption line, within the gain profile of the laser, is placed inside the cavity. When the laser mode is tuned to absorption, line-center saturation reduces the absorption loss, resulting in a narrow peak in the laser emission. Drifting off this peak can be used as an error signal, as before, to stabilize the mode frequency.

Saturated absorption spectroscopy uses the same mechanism by placing the absorbing specimen outside the laser cavity. The effect is detected by monitoring the transmission of a weak probe beam in the presence of a strong counter-propagating 'pump' beam. Both probe and pump beams are

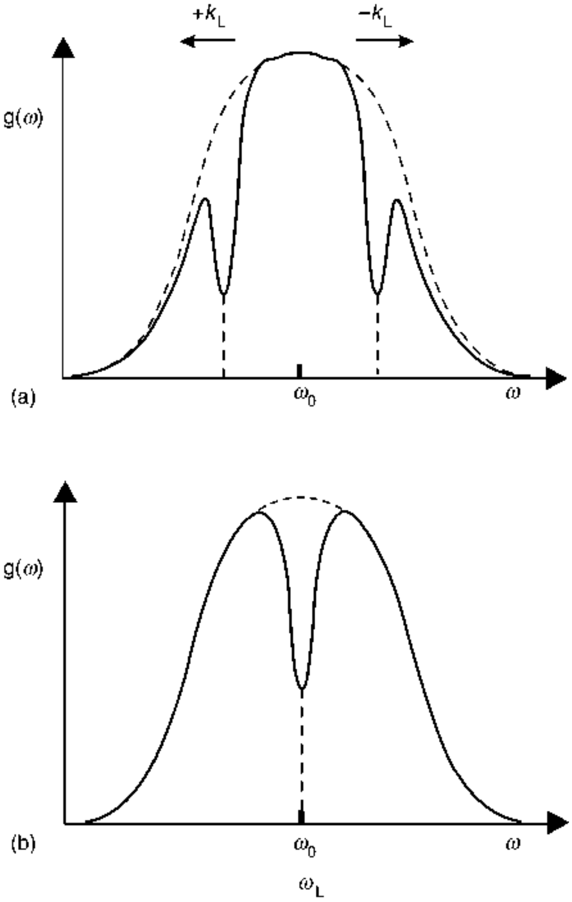

Figure 2 (a) Holes burned in gain profile by two counter-propagating monochromatic waves at $\omega_L \neq \omega_0$. (b) 'Lamb dip' in gain profile $g(\omega)$ (or absorption profile $\alpha(\omega)$) when $\omega_L = \omega_0$.

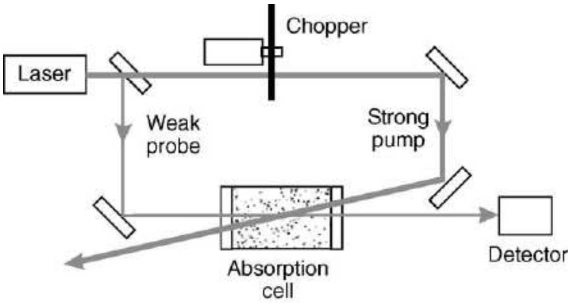

Figure 3 Schematic diagram of saturated absorption spectroscopy experiment. The saturating pump beam is chopped or modulated and the signal is processed using phase-sensitive detection to improve the signal-to-noise ratio.

derived from the same laser using a suitable beamsplitter and, thus, have identical frequency. When tuned to line-center of the absorbing sample, the transmission of the probe increases owing to saturation by the strong pump beam interacting with the same subgroup of the inhomogeneously broadened medium (Figure 3).

The elimination of the first-order Doppler effect, in gaseous media, using this technique results in very high-resolution spectra. The residual linewidth is determined by homogeneous processes such as lifetime broadening, collisional (or pressure) broadening and power (or saturation) broadening. The latter effect is inevitable, as saturation is necessary, but may be controlled so as to contribute very little to the overall broadening. Power broadening becomes significant only when the laser intensity is appreciably in excess of the saturation intensity. Another limiting effect is transit time broadening: moving atoms spend only a limited time interacting with the narrow beams as they pass through. The consequent limit on the interaction time leads to conjugate broadening of the frequency of the transition as a result of the uncertainty principle.

The general problem of the spectral lineshape of signals in saturation spectroscopy is complex. Apart from the effects mentioned above, the lineshape will be affected by relativistic effects (second-order Doppler shifts), velocity-changing collisions, recoil effects from photon absorption and emission, velocity-dependent transit time broadening, optical pumping and laser lineshape effects.

Saturation spectroscopy is widely used in high-resolution spectroscopy for measurement of atomic and molecular transition frequencies, time and frequency standards using stabilized lasers, measurement of fundamental constants such as the Rydberg constant and the fine structure constant. Dynamical effects such as atomic collisions are also studied using saturation, whereby the relaxation of hole burning is measured in the time or frequency domain.

Polarization Spectroscopy

An important variation of saturation spectroscopy uses the change induced by a saturating beam in the real part of the complex refractive index, i.e., dispersion rather than absorption. Specifically, the technique of polarization spectroscopy employs a saturating beam that is circularly polarized to effect optical pumping on the degenerate lower level of a transition. The degenerate states may be identified, for example, with magnetic substates based on orientation of the intrinsic atomic or molecular angular momentum and labeled by the quantum numbers m_J. In equilibrium, the populations of all these degenerate substates will be equal and the optical properties of the system will then be isotropic. In particular, the polarization state of a linearly polarized probe beam will be unaffected by passage through the medium. This probe is totally rejected by an analyzer (a linear polarizer set to transmit only light in an orthogonal plane). The linear polarized beam is a linear combination of two oppositely circular polarization states, σ^+ and σ^-, and both components experience the same complex refractive index. In the presence of a circularly polarized saturating pump beam, the distribution of population amongst the degenerate substates is disturbed from its equilibrium by optical pumping. The result will be an excess of population in higher or lower m_J states (a condition known as orientation) and this leads to optical anisotropy in the medium. The refractive index for σ^+ light is now different from that for σ^- light and a relative phase shift is introduced between these components of the linearly polarized probe resulting in elliptical polarization. A component of this elliptically polarized light will then be transmitted by the analyzer and is detected. When the strong pump beam is closely counter-propagating to the probe, then both beams interact with the same velocity subgroup as in Doppler-free saturation spectroscopy. The inhomogeneous (Doppler) broadening is thus significantly reduced. The technique provides a signal above an essentially 'zero' background and is thus capable of a higher signal-to-noise ratio than saturated absorption spectroscopy. How close to 'zero' the background is depends crucially on the degree of extinction by the analyzer. High-quality polarizer/analyzer devices can achieve extinction ratios of 1:10^{-8}. Stress-induced birefringence in the windows of the cell containing the sample is a common limiting factor (Figure 4).

A probe laser of intensity I_{probe} is transmitted through a sample of length l with an intensity $I_{\text{PS}}(\omega)$ given by the following expression, that takes account of different contributions to the detected signal:

$$I_{\text{PS}}(\omega) = I_{\text{probe}}\left[\xi + \varphi^2 + b^2 + \frac{1}{2}\frac{b\Delta\alpha_{\pm}l}{(1+x^2)} + \frac{1}{2}\varphi\Delta\alpha_{\pm}l\frac{x}{(1+x^2)} + \left(\frac{1}{4}\Delta\alpha_{\pm}l\right)^2\frac{1}{(1+x)^2}\right] \quad [9]$$

where $x = 2(\omega_0 - \omega)/\gamma$ is the detuning from the resonance frequency ω_0 normalized by the halfwidth γ of the collision broadened line. The frequency-dependent terms provide absorptive (Lorentzian) and dispersive contributions to the total lineshape. The first three, frequency independent, terms in eqn [9] are responsible for a constant background signal. They originate from residual transmission of the polarizers (ξ), their accidental uncrossing (φ), and a possible residual birefringence (b) from optical elements between both polarizers in the probe beam path. ($\xi + b^2$), normally is in the range 10^{-6}. The differential absorption coefficient for left and right circularly polarized light, $\Delta\alpha_{\pm}$, at resonance ($x = 0$), is given by:

$$\Delta\alpha_{\pm}(x=0) = \Delta\alpha_{\pm}^0 = \alpha^+ - \alpha^- = \alpha_0 S_0 \Delta C_{JJ_1} \quad [10]$$

where α_0 is the unsaturated absorption coefficient, S_0 is the saturation parameter for the pump wave ($I_{\text{pump}}/I_{\text{sat}}$), and ΔC_{JJ_1} are Clebsch–Gordan coefficients for the respective transition and coupling case.

The dispersion-shaped component may be made to dominate by deliberately uncrossing the polarizer/analyzer combination to a small extent. This technique enhances the resolution of the transition frequency. Linewidths are limited by residual Doppler broadening owing to finite crossing angles of pump and probe beams, collisional broadening, transit time broadening, etc.

Applications of the technique have included Doppler-free spectroscopy of atomic transitions for precision measurements of fundamental constants, molecular spectroscopy and detection of atomic, molecular, and radical species in flames. By employing a different laser to accomplish the induced optical anisotropy from that used to probe the medium, molecular spectra can be simplified by the technique of 'polarization labeling'. Signals are obtained only when the probe and pump excite transitions sharing a common level and the particular transitions observed are sensitive to the selection rules for ΔJ, governing both pump and probe transitions. In this way P- and R-branch transitions can be distinguished.

Advantages of polarization spectroscopy include high sensitivity (10^2–10^3 times that of saturation spectroscopy), high signal-to-noise, simplification of

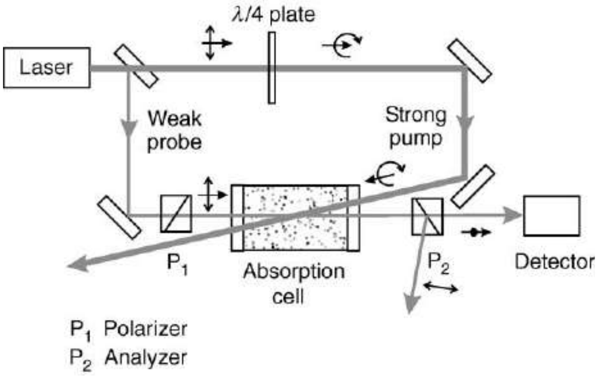

Figure 4 Schematic diagram of polarization spectroscopy experiment. Circularly polarized pump beam may be modulated for phase-sensitive detection of transmitted probe beam.

molecular spectra by polarization labeling, dispersion profiles for enhanced resolution, and for frequency stabilization techniques.

Multiphoton Absorption Spectroscopy

Multiphoton absorption is the term used to describe a process in which an atom or molecule makes a single transition between two of its allowed energy levels by absorbing the energy from more than a single photon. Sequential transitions, by which successive photons are absorbed via single-photon transitions between stationary energy levels, take place provided photons arrive within the lifetime of the intermediate state. Multiphoton transitions take place via 'virtual' energy levels. The lifetime of a virtual level is determined by the uncertainty principle to be the inverse frequency detuning from the nearest stationary energy level. Thus a high photon flux is usually required for multiphoton effects to be observed (Figure 5).

The simplest multiphoton effect, two-photon absorption, may be observed using strong atomic or molecular electronic transitions, even with modest laser power. The absorption is detected by observing the fluorescence from the spontaneous, radiative decay of the excited state. Two-photon absorption spectroscopy may be made Doppler-free by arranging absorption of one photon from each of two counter-propagating beams. The moving atoms experience opposite Doppler shifts $\mathbf{k}\cdot\mathbf{v}$ from each beam. The transition energy is thus:

$$E = (\hbar\omega + \mathbf{k}\cdot\mathbf{v}) + (\hbar\omega - \mathbf{k}\cdot\mathbf{v}) = 2\hbar\omega \qquad [11]$$

Absorption of two photons from a single beam is prevented by choosing a transition between levels of the same parity and angular momentum quantum number, i.e., $\Delta J = 0$ and using circularly polarized light. Absorption of two σ^+ or σ^- photons gives $\Delta J = \pm 2$, which would be forbidden, e.g., on a $n^2S - n'^2S$ transition in an alkali metal. This method of Doppler-free spectroscopy has the advantage that it involves all the atoms and not just the zero-velocity subgroup, as in saturation or polarization spectroscopy. In principle, it is also possible to use three-photon absorption, with the momentum of the wavevectors arranged to sum to zero, to achieve Doppler-free absorption on a transition between opposite parity states.

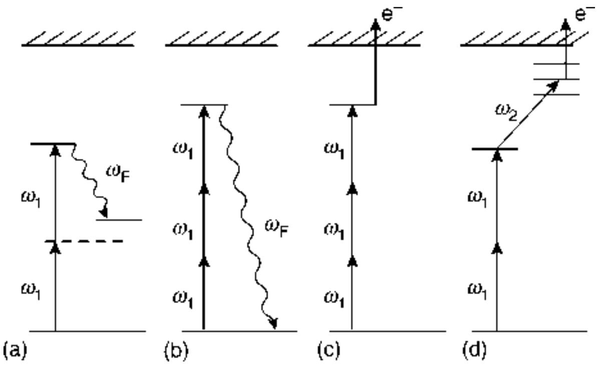

Figure 5 Multiphoton absorption spectroscopy. (a) Two-photon absorption using fluorescence detection of radiative decay at ω_F. (b)Three-photon absorption spectroscopy using fluorescence detection of radiative decay at ω_F. (c) Three-photon absorption spectroscopy using ionization detection following photoionization, collisional or field ionization. (d) Two-photon resonant multiphoton ionization spectroscopy; laser at frequency ω_1 is fixed and laser at ω_2 is tuned over transitions to high-lying levels.

Instead of observing fluorescence following multiphoton absorption, the excited atom or molecule may be ionized by atomic collisions or by applying a d.c. electric field. Excitation of Rydberg levels (i.e., having high value of the principal quantum number n) by two- or three-photon absorption is readily accomplished. The sensitivity of this technique is enhanced by using two lasers. One laser at frequency ω_1 is tuned to a two-photon transition (energy $= 2\hbar\omega_1$) and the second laser ω_2 is tunable to excite transitions from the two-photon intermediate state to higher-lying levels (energy $= 2\hbar\omega_1 + \hbar\omega_2$). Thus, energy levels normally requiring excitation in the vacuum UV may be reached using visible frequency lasers. In such cases ionization detection may be more efficient than fluorescence since the nonradiative decay rate of these highly excited levels may exceed the fluorescence decay rate. The ion signal produced may be significantly amplified by the process of space charge amplification. A weak discharge established in the atomic vapor creates a space charge around the cathode. A single ion moving slowly through the space charge cloud acts to cancel the local field allowing many (up to 10^6) electrons to escape to the anode with a resulting amplification of 10^6 in the 'photoionization current'.

Multiphoton ionization spectroscopy is a sensitive method of recording highly excited energy levels. The sensitivity is enhanced in the technique of resonance ionization spectroscopy (RIS): two lasers are used such that the frequency of one laser is fixed close to a two-photon allowed transition and the second is tuned to excite a range of higher excited states. Absorption of an even number of photons allows excitation of levels having the same parity as the ground state and thus not accessible to normal absorption spectroscopy. In combination with a mass spectrometer to detect the ions produced by the multiphoton absorption resonance ionization mass spectrometry (RIMS), affords additional species selectivity and sensitivity. RIMS finds application in high-resolution spectroscopy of rare isotopic species and is the basis for a scheme for isotope separation.

Coherent Third-Order Processes

An important class of spectroscopic techniques exploits the coherent interaction of three input waves via the nonlinear response of the medium. Specifically, the third-order term has the form:

$$\mathrm{P}_i^{(3)}(\omega_4,\vec{r}) = \chi_{ijkl}^{(3)}(\omega_1,\omega_2,\omega_3,\omega_4)\mathrm{E}_j(\omega_1,\vec{r}) \times \mathrm{E}_k(\omega_2,\vec{r})\mathrm{E}_l(\omega_3,\vec{r}) \quad [12]$$

where $\mathrm{E}_m(\omega_n)$ are complex field amplitudes and the subscript $m = i,j,k,l$ refers to each of the Cartesian coordinates x,y,z of the electric vector. $\chi_{ijkl}^{(3)}$ is a fourth-rank tensor. In condensed phase media, the third-order nonlinearity is readily observed, even with modest laser intensities. In the gas phase, although $\chi_{ijkl}^{(3)}(\omega)$ is small relative to $\chi^{(1)}$, its effect is enhanced when the incident fields are resonant and this nonlinear response may be used for spectroscopy. Equation [12] shows that the induced polarization couples three waves to radiate a fourth and describes a class of interactions known collectively as four-wave mixing.

The physics of four-wave mixing interactions may be visualized as the formation and read-out of a volume hologram. The three input beams are denoted as two pumps and one probe. The process involves the interference of one pump with the probe to form a spatially periodic grating pattern in the electric field distribution. The nonlinear response of the medium via $\chi^{(3)}$ leads to a spatial modulation of the refractive index – a laser-induced grating. The second pump beam generates a signal by scattering off this grating in a manner analogous to Bragg scattering of X-rays from regular crystal planes. The condition for generating a coherent scattered beam is equivalent to matching the momentum (or wavevectors) of the interacting photons, i.e., phase matching.

Degenerate Four-Wave Mixing (DFWM) Spectroscopy

The simplest case of four-wave mixing spectroscopy involves pump and probe beams of identical (i.e., degenerate) frequency to generate a signal at the same frequency. Spectra are generated as the frequency is tuned through atomic or molecular resonances where the interaction is strongly enhanced. In this degenerate case, the pump and probe interference pattern is stationary and the remaining pump beam scatters off this grating with the same frequency. As with all parametric processes, the interaction is phase matched by a suitable geometrical arrangement of the beams. A probe intersecting with counter-propagating pump beams leads to a signal that counter-propagates with the probe. This arrangement provides 'Doppler-free' lineshapes, since only those atoms with essentially zero velocity along the line of intersection of the beams contribute to the signal. Alternatively, pump and probe beams may be arranged to cross at an angle (usually small) in the forward direction – the 'forward geometry', with the resulting signal propagating at the same small angle to the incoming beams. Signals in this case are stronger, since all velocity classes may contribute but the lineshape is Doppler broadened (Figure 6).

An additional feature of the counter-propagating pump geometry is that the phase-matched signal arises from terms that involve products of fields with the complex conjugate of the incident

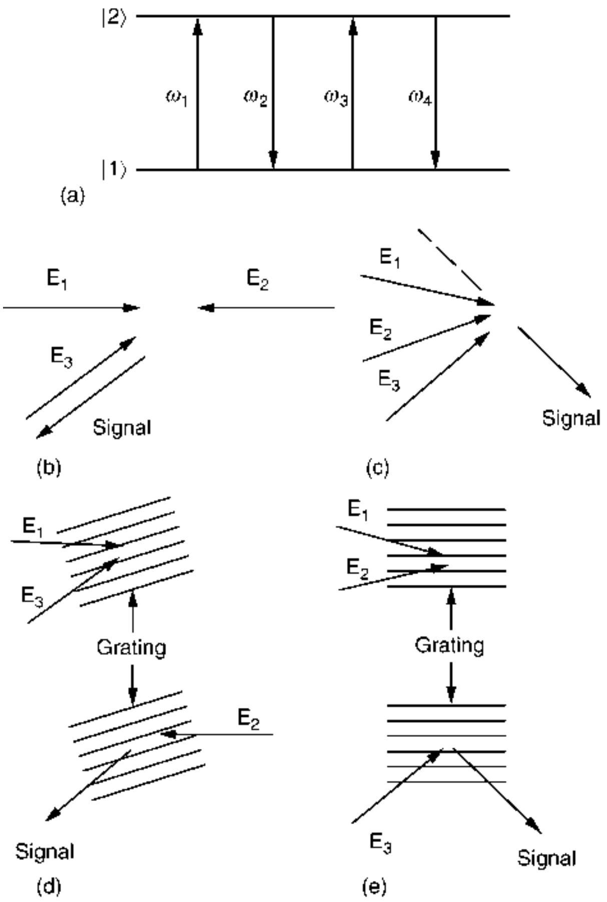

Figure 6 Degenerate four-wave mixing: (a) all input frequencies are degenerate, $\omega_1 = \omega_2 = \omega_3$ and give signal at the same frequency, ω_4. Phase matching geometries are shown in (b) counter-propagating pump (phase conjugate) and (c) forward geometry. (d) and (e) show the grating formation and read-out respectively.

probe field. The induced polarization radiating the signal wave has a component given by:

$$\mathbf{P}^{(3)}_{\text{signal}}(\omega) = \chi^{(3)}(\omega)\mathbf{E}_{\text{pump}}(\omega)\mathbf{E}_{\text{pump}}(\omega)\mathbf{E}^{*}_{\text{probe}}(\omega) \qquad [13]$$

The backward propagating signal appears as a phase conjugated reflection of the incident probe wave. A probe wave that has propagated through a phase-distorting medium thus generates a signal with an initial phase distortion that will be undone as it retraces the path through the distorting medium. This phase conjugating property finds application in adaptive optics but also lends an advantage when DFWM is used as a spectroscopic probe in hostile or distorting environments, such as flames or plasmas. The counter-propagating pump geometry is often referred to as the phase conjugating geometry.

The intensity of the DFWM signal beam may be calculated using perturbation theory to derive $\chi^{(3)}(\omega)$ arising from saturated absorption. In the perturbation approximation, the probe beam is assumed to be weak and interacts with stationary atoms in the presence of counter-propagating, saturating pump fields. The signal is found from the phase-matched terms in the wave radiated from the induced polarization. When integrated over the atomic resonance line, the signal intensity, $I^{\text{int}}_{\text{sig}}$, is given by analytic expressions in the two limiting regimes of weak and strong saturation by the pump:

$$I^{\text{int}}_{\text{sig}} = \frac{|\mu_{12}|^8 N^2 k^2 T_1^2 T_2^3}{[1+4I_{\text{pump}}/I_{\text{sat}}(0)]^{5/2}} I^2_{\text{pump}} I_{\text{probe}} \qquad [14]$$

when $I_{\text{pump}} \ll I_{\text{sat}}(0)$ and

$$I^{\text{int}}_{\text{sig}} \propto |\mu_{12}|^3 N^2 k^2 T_1^{-1/2} T_2^{1/2} I^{-1/2}_{\text{pump}} I_{\text{probe}} \qquad [15]$$

when $I_{\text{pump}} \gg I_{\text{sat}}(0)$.

In these relations μ_{12} is the atomic dipole moment, N the atomic number density, $I_S(0) = \hbar^2/T_1 T_2 |\mu_{12}|^2$ is the line-center saturation intensity, and T_1 and T_2 are the longitudinal and transverse relaxation times, respectively. Extensions of the basic theory accommodate the effects of moving absorbers (Doppler effect) in both the phase-conjugate and forward phase matching geometry.

These results are derived using perturbation theory, by assuming that the probe field is weak. In practice the signal is optimized when pump and probe are of equal intensity and approximately equal to the saturation intensity. This situation violates the perturbation theory assumption but the signal intensity may be derived using nonperturbative analytical methods, although the resulting expression is more complicated. Alternatively, the quantum mechanical equations for the nonlinear susceptibility can be solved numerically to find the signal level for arbitrary intensity of the incident fields. Optimum signal-to-noise ratio in the spectra is achieved using approximately saturating intensities.

The sensitivity of resonant DFWM and the coherent nature of the signal beam suits the process to spectroscopy of nonfluorescing or trace molecular species, such as radicals created in chemical reactions. Equation [15] shows that the DFWM signal becomes relatively insensitive to collisional effects when the incident lasers saturate the transition. This feature of DFWM gives it an additional advantage over laser-induced fluorescence in optical diagnostic applications where quantitative evaluation of fluorescence intensity is hampered by collisional quenching. The state and species selectivity afforded by resonant DFWM is exploited in diagnostics of combustion and plasma processes, where the reduction of Doppler broadening allows high-resolution spectra from high-temperature gases. The relative intensity of spectral lines is used to derive temperature and concentration information with time and space resolution.

Additional selectivity is provided by using two input frequencies in two-color four-wave mixing (TC-FWM), to probe molecular excited state dynamics. The TC-FWM process uses resonant enhancement on two coupled transitions, i.e., the pump beams excite one transition and the probe and signal are generated on another transition having a common level with that excited by the pumps.

Coherent Raman Spectroscopy

The classical theory of spontaneous Raman scattering shows that a nonresonant pump of frequency ω_p wave can excite a resonance ω_m in a molecule with a nonlinear polarizability $\beta(x)$. (x is the spatial coordinate of the internuclear potential.) The wave resulting from this inelastic scattering is a Stokes wave of frequency $\omega_S = \omega_p - \omega_m$. The propagation of the Stokes wave is governed by the complex Raman susceptibility (Figure 7a):

$$\chi_{\text{Raman}}(\omega_s) = \chi'_{\text{Raman}}(\omega_s) + i\chi''_{\text{Raman}}(\omega_s) \qquad [16]$$

This generated Stokes wave may then couple with the pump to drive the molecular resonance coherently leading to amplification of the Stokes wave at the expense of the pump. This process of stimulated Raman scattering (SRS), provides gain if the amplification exceeds the loss due to absorption, scattering, etc. There is, therefore, a definite threshold intensity that the pump must satisfy to observe SRS. The SRS

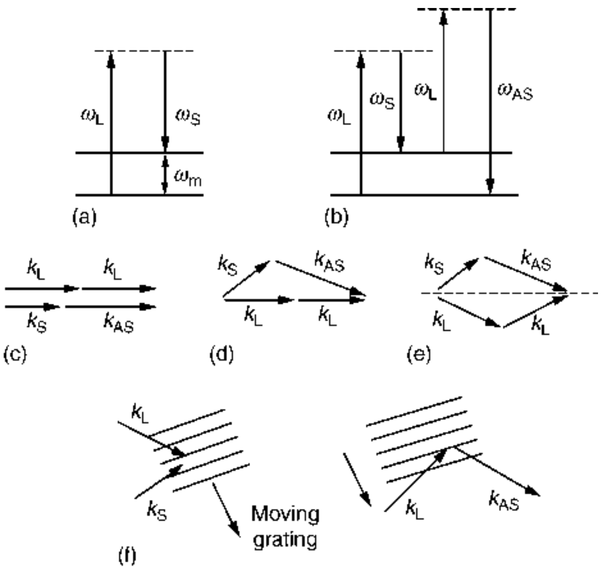

Figure 7 Coherent Raman spectroscopy. (a) Spontaneous Raman scattering of input pump ω_L producing Stokes wave ω_S in inelastic scattering. (b) Coherent anti-Stokes Raman scattering (CARS). Phase matching geometries for CARS is shown for a nondispersive medium in (c) and dispersive media in (d) and (e). The planar boxcars geometry in (e) may be folded along the dotted line (folded boxcars geometry) to separate the signal into an orthogonal plane from the pump beams to reduce scattered pump light reaching the detector.

gain $g(\omega_s)$ at the Stokes frequency ω_s is given by:

$$g(\omega_s) = -\frac{k_s}{2n_s^2}|E(\omega_L)|^2\chi''_{\text{Raman}}(\omega_s) \qquad [17]$$

where k_s and n_s are the wavevector and refractive index for the Stokes wave, $E(\omega_L)$ is the field amplitude at the pump laser frequency ω_L, and $\chi''_{\text{Raman}}(\omega_s)$ is the imaginary part of the complex Raman susceptibility:

$$\chi''_{\text{Raman}}(\omega_s)=\frac{-\varepsilon_0 N\left(\dfrac{\partial\beta}{\partial x}\right)^2\gamma}{32m_e\omega_v\left\{(\omega_m-(\omega_L-\omega_S))^2+\dfrac{\gamma^2}{4}\right\}} \qquad [18]$$

In this expression $\partial\beta/\partial x$ is the variation of polarizability β with internuclear displacement x, N is the molecular population density, and γ is the linewidth due to damping. Spectroscopy based on SRS provides signals that are many orders of magnitude stronger than those from spontaneous Raman scattering, and is widely used to study spectra of complex molecules having no absorption spectra in the visible region.

A four-wave mixing interaction, involving a pump and probe beam of different frequencies, generates a nonstationary grating. The signal generated by scattering of a second pump beam off this moving grating is Doppler-shifted to lower or higher frequency. These signal beams correspond respectively to Stokes or anti-Stokes waves in Raman scattering. A probe beam at the Stokes frequency will thus interact with a pump beam to generate a coherent signal at the anti-Stokes frequency; a process known as coherent anti-Stokes Raman scattering (Figure 7b). A probe at the anti-Stokes frequency correspondingly gives the process of coherent Stokes Raman scattering (CSRS). Phase matching is necessary for both CARS and CSRS and is satisfied by an appropriate arrangement of the incident wavevectors (Figure 7c,d and e).

CARS has proven to be a valuable spectroscopic technique in a wide variety of applications. A CARS spectrum is generated by scanning the frequency of the probe, i.e., the Stokes laser. The signal strength is given by

$$I_4 = I_{\text{CARS}} = \frac{\omega_4^2}{n_1^2 n_2 n_4 c^4 \varepsilon_0^2} I_1^2 I_2 |\chi_{\text{CARS}}|^2 l^2 \times\left(\frac{\sin(\Delta kl/2)}{\Delta kl/2}\right)^2 \qquad [19]$$

where I_1 and I_2 are the intensity of the pump and Stokes beams, respectively, l is the interaction length, and n_i are the refractive indices at frequency ω_i. The maximum signal is generated when the phase mismatch $\Delta k = 0$. The third-order susceptibility, $\chi^{(3)} = \chi_{\text{CARS}}$, in general consists of a nonresonant part $\chi^{(3)}_{\text{nr}}$, which is independent of the frequency of the exciting beams, and a Raman-resonant contribution $\chi^{(3)}_{\text{r}}$:

$$\chi^{(3)} = \chi^{(3)}_{\text{nr}} + \chi^{(3)}_{\text{r}} \qquad [20]$$

The lineshape in coherent Raman spectroscopy is affected by interference between resonant and nonresonant contributions. This effect is exploited to derive the concentration of the resonant (Raman active) species relative to that of the background of known density. A more common application of CARS, however, is for temperature measurement. The dependence of the signal intensity on $|\chi_{\text{CARS}}|^2$, and therefore on the square of the molecular population of the initial state, yields a spectrum from which the temperature may be derived by assuming a Maxwell–Boltzmann distribution over the state populations.

A particularly useful application of CARS employs a narrow linewidth pump laser and a broad bandwidth Stokes laser. CARS signals are generated simultaneously on all frequencies such that the corresponding Stokes frequency is available from the broadband probe. The spectrum recorded with a single laser pulse can thus provide population information on a wide range of molecular states, giving time- and space-resolved measurement of temperature and concentration.

CARS and CSRS spectroscopy is applied to both gas and condensed phase media. Owing to the higher density in the latter, signals may be strong in spite of the two-photon nature of the resonant interaction.

Time Domain Four-Wave Mixing Spectroscopy

Four-wave mixing spectroscopy may also be executed in the time domain. The three incident waves may arrive as pulses separated in time by variable delays. The interaction is, therefore, mediated by the atom or molecule's induced polarization or coherence. DFWM, CARS, and CSRS processes may thus be used to study ultrafast relaxation since the signal intensity gives a measure of the coherence decay in the intervening delay between the pulses. These processes are intimately related to coherent transient spectroscopy using measurements of free induction decay, photon echoes, and stimulated photon echoes.

Laser-Induced Grating Spectroscopy

The four-wave mixing processes discussed above involve a spatial modulation of the molecular population and a consequent modulation of the refractive index for resonantly interacting waves. The signals are radiated by the coherent oscillations of the molecular populations that decay rapidly. Relaxation processes result in transfer of energy from the excited molecular states, leading to a temperature and density grating in the bulk medium with a period Λ, given by:

$$\Lambda = \lambda_{\text{pump}}/(2 \sin \theta_1) \qquad [21]$$

where θ_1 is the angle between the interfering beams of wavelength λ_{pump}. A nonresonant wave of wavelength λ_{probe} may be scattered by this laser-induced grating provided it is incident on the grating at the angle θ_2 satisfying the Bragg condition:

$$\theta_2 = \sin^{-1}(\lambda_{\text{probe}}/2\Lambda) \qquad [22]$$

The periodic temperature/density grating, when induced by a short duration laser pulse, is accompanied by two oppositely propagating acoustic waves. The reflectivity of Bragg scattering is temporally modulated as the acoustic waves traverse the stationary grating at the local sound speed. Measurement of the period of these oscillations yields the sound speed from which the temperature may be derived. The stationary grating and acoustic waves decay by diffusive and damping processes, dependent upon the bulk medium properties. Measurement of the decay time of the grating induced in gas phase media yields the gas pressure if the gas dynamic parameters are known (Figure 8).

Laser-induced grating spectroscopy (LIGS), has been widely employed for studies of condensed phase relaxation processes. Applications have also been developed for gas phase studies using resonant absorption to enhance the thermal grating component. A grating may, however, also be induced in response to the electrostrictive effect of the high electric field in the interference pattern of two intense pump waves. Such electrostrictive gratings do not require resonant absorption and their spectroscopic utility centers on the study of relaxation effects in the molecular medium.

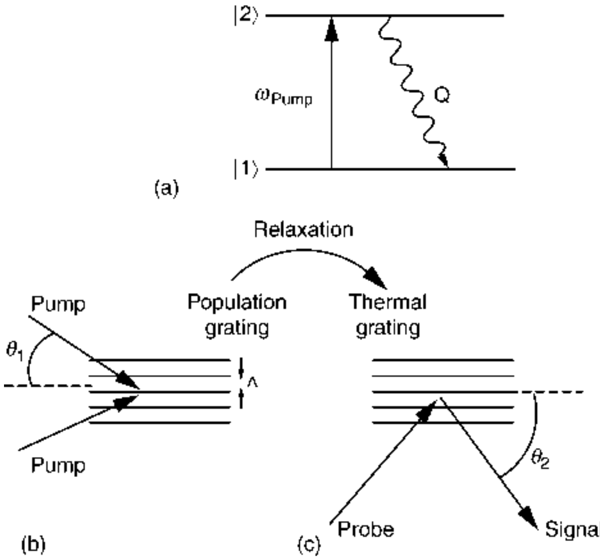

Figure 8 Laser-induced thermal grating spectroscopy. (a) Absorption of energy from pump lasers creates excited molecular population in grating pattern shown in (b). Spacing of laser induced grating Λ is set by angle of intersection of pumps. Quenching, *Q*, of excited population by collisions transfers heat to bulk medium and forms thermal grating. (c) Probe incident at Bragg angle θ_2 is scattered to form signal.

List of Units and Nomenclature

Formulae are expressed in SI units

c	Speed of electromagnetic wave *in vacuo*
$\mathbf{E}$	Electric field amplitude
$g_0(\omega)$, $g(\omega)$	Gain coefficient at line center, at frequency ω
h	Planck's constant
I	Intensity of electromagnetic wave
I_{sat}	Saturation intensity of medium
$\mathbf{k}$, k	Wave vector, scalar value of wave vector ω/c
m_{e}	Mass of electron
n_{m}	Refractive index of medium for mth beam
N	Atomic or molecular population density
$\mathbf{P}$	Polarization field amplitude
T_1	Longitudinal relaxation time
T_2	Transverse relaxation time
α_0, $\alpha(\omega)$	Absorption coefficient at line center, at frequency ω
$\beta(x)$	Molecular polarizability as function of internuclear separation x
ε_0	Permittivity of free space
μ_{12}	Dipole moment for transition between energy levels 1 and 2
ν	Frequency of electromagnetic wave
$\chi^{(n)}$	nth order susceptibility
ω	Angular frequency of electromagnetic wave

See also

Coherent Transients: Coherent Transient Spectroscopy in Atomic and Molecular Vapours; Foundations of Coherent Transients in Semiconductors; Ultrafast Studies of Semiconductors. **Fiber and Guided Wave Optics:** Nonlinear Optics. **Imaging:** Adaptive Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; Four-Wave Mixing; Kramers–Kronig Relations in Nonlinear Optics. **Phase Control:** Phase Conjugation and Image Correction. **Physical Applications of Lasers:** Sum-Frequency Generation at Surfaces. **Scattering:** Raman Scattering; Stimulated Scattering. **Spectroscopy:** Raman Spectroscopy; Second-Harmonic Spectroscopy.

Further Reading

Clark RJH and Hester RE (eds) (1988) Quantitative CARS spectroscopy. In: *Advances in Non-Linear Spectroscopy*, pp. 193–251. New York: John Wiley.

Corney A (1986) *Atomic and Laser Spectroscopy.* Oxford, UK: Oxford University Press.

Demtröder W (1996) *Laser Spectroscopy*, 2nd edn. Berlin: Springer-Verlag.

Dreier T and Ewart P (2002) Coherent techniques for measurements with intermediate concentrations. In: Kohse-Höinghaus K and Jeffries JB (eds) *Applied Combustion Diagnostics*. London: Taylor and Francis.

Eesley GL (1981) *Coherent Raman Spectroscopy.* London: Pergamon.

Eichler HJ, Günter P and Pohl DW (1986) *Laser Induced Dynamic Gratings*. Berlin: Springer-Verlag.

Feld MS and Letokhov VS (eds) (1980) *Coherent Nonlinear Optics, Recent Advances*. Berlin: Springer-Verlag.

Fisher RA (ed.) (1983) *Optical Phase Conjugation*. San Diego, CA: Academic Press.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy.* Revised edn. San Diego, CA: Academic Press.

Raman Spectroscopy

R Withnall, University of Greenwich, Chatham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Inelastic light scattering, the optical analogue of Compton scattering, had been predicted to occur by Smekal in 1923, but it was Chandresekar Venkataraman Raman who provided the first experimental demonstration of the phenomenon in February 1928. Only a few months later in 1928 the Russian scientists, Landsberg and Mandelstam observed inelastic light scattering from a quartz crystal. In spite of initial claims to the contrary, this was the same effect that had been observed by Raman, although the Russian scientists did not acknowledge this, preferring to call the effect 'combinatorial scattering'. It was not until the 1970s that the effect became known as Raman scattering everywhere, including Russia.

As typically practised, Raman spectroscopy involves laser excitation of a sample and measurement of the wavelength and intensity distribution of the scattered Raman light. When a microscope is used for delivering laser excitation and/or collecting the inelastically scattered light, the technique is referred to as Raman microscopy. However, it is important to recognize that, apart from the differences in sampling configuration, there are no fundamental differences between Raman microscopy and Raman spectroscopy; the terms merely identify the different sampling techniques.

Since its invention in 1960, major advances in the technology of the laser have occurred, resulting in the wide variety of laser light sources that are currently available for Raman spectroscopy. Characteristic properties of laser systems for Raman microscopy will be described in this chapter, following an introduction to the technique and its applications in the next section below. Then, the commonly used laser sources for Raman microscopy are categorized according to the wavelength regions of their emissions, along with their merits and drawbacks for each application. The chapter will consider mainly continuous wave laser sources, because pulsed laser beams that are tightly focused with microscope optics give very high irradiance that would destroy most samples.

Raman Microscopy

The first experimental Raman microscopes were developed in the mid-1970s by two independent groups, and it was not long before the first generation of Raman microscopes were subsequently commercialized. Many of these early instruments simply consisted of commercial optical microscopes coupled to Raman spectrometers, as shown in Figure 1. It was recognized that the use of a microscope for sampling in this way can facilitate the Raman examination of tiny particles of micrometer dimensions.

The optical lay-out in Figure 1 shows an infinity corrected microscope. The advantage of using this, rather than a standard tube length microscope, is that its length can be extended in order to incorporate additional optical elements that are required for coupling to the Raman excitation and collection optics. In the optical layout, the incident laser beam is spatially filtered by a pinhole, D_1, in order to remove the diffraction rings and give a point source. The laser light is then partially reflected by beamsplitter, B_1, and focused on to the sample by the microscope objective. The back-scattered Raman radiation is collected by the same microscope objective, partially transmitted by the beamsplitter, and directed into the entrance slit of the Raman spectrometer by means of coupling optics. The accurate location of an aperture, D_2, at the focal point improves the spatial resolution and allows depth profiling of transparent samples. The two pinholes, D_1 and D_2, act as spatial filters and are referred to as confocal diaphragms. The Raman microscope is described as being confocal, because out-of-focus light, collected from outside the focal volume, is rejected.

The use of the beamsplitter, in the optical path of early Raman microscope designs, gives rise to significant losses of the precious Raman

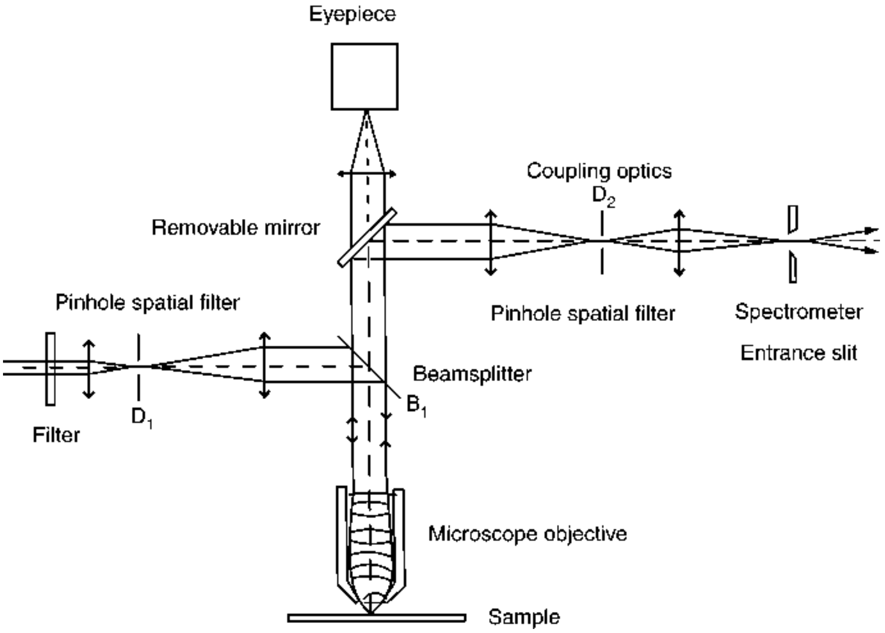

Figure 1 Diagram of the optical layout of a visible Raman microscope (Reproduced from Turrell G, Delhaye M and Dhamelincourt P (1996) Characteristics of Raman microscopy. In: Turrell G and Corset J (eds) *Raman Microscopy. Development and Applications*, 2. London: Academic Press).

scattered light. If a 50:50 beamsplitter is used then half of the incoming laser light is lost by transmission at B_1 and, more significantly, half of the Raman scattered light is lost by reflection at B_1. As long as a higher laser power can be employed, it is better to use a 90:10 beamsplitter that transmits 90% of the light and reflects 10%; then, only 10% of the incoming laser light is directed by B_1 on to the sample, but 90% of the Raman scattered light is transmitted by B_1 towards the detector and only 10% of the Raman light is lost by reflection at B_1. The efficiency of a beamsplitter can be defined as the product of the reflectivity at the laser wavelength and the transmission at the Raman-shifted wavelength. Thus, conventional 50:50 and 90:10 beamsplitters would have efficiencies of 25% and 9%, respectively. Recently, however, holographic notch filters have been developed which can be used as higher efficiency beamsplitters. These typically have a laser rejection contrast ratio of better than 10^4, reflecting ca. 90% of the laser excitation and transmitting ca. 90% of the Raman scattered light, thus giving an efficiency of ca. 81%. Due to these considerations, higher throughput, commercial Raman microscope designs usually incorporate a holographic notch filter.

The importance of Raman microscopy stems from the fact that it is the only microanalytic method available today, by use of which it is possible to identify, or characterize, small particles of micrometer dimensions *in situ*. It is also advantageous that no sample preparation is required when performing Raman microscopy.

The fundamental limit of the lateral spatial resolution (i.e., for a diffraction limited focus) is the separation at which the maximum of one Airy disc function just touches the first minimum of an adjacent Airy disc, given by:

$$\text{Lateral spatial resolution} = \frac{1.22\lambda}{2\text{NA}} \qquad [1]$$

where λ is the wavelength of the light and NA is the numerical aperture of the microscope objective.

The axial Airy disc function determines the resolution in the longitudinal direction. A good estimate of the axial resolution limit for low numerical aperture is given by:

$$\text{Axial spatial resolution} = \frac{2\lambda}{(\text{NA})^2} \qquad [2]$$

When the sample is heterogeneous and exhibits fluorescence that is not evenly distributed within the sample, a region of the sample can be selected with the microscope that shows the minimum amount of fluorescence. If the fluorescence is intrinsic to the sample itself, then it is possible to use the shift subtract technique and/or temporal discrimination between the fluorescence and the Raman scattering. The latter can be achieved by Kerr gate fluorescence rejection, in which a pulse of light is used to close a Kerr gate shutter. The problem with this approach, however, is that it uses a laser pulse, and even pulses of modest energy will have extremely high irradiance when focused to a tiny spot by a microscope objective. Such laser pulses would inevitably destroy the sample under a microscope.

Another approach for reducing fluorescence is to use near infrared excitation, which is low enough in energy so that absorption cannot occur to promote electronic transitions. The technique of Fourier Transform Raman (FTR) spectroscopy often offers the best chance of obtaining Raman spectra from fluorescent samples for this reason. Low energy excitation of wavelength equal to 1064 nm, provided by a Nd:YAG laser, is normally used for FTR. Disadvantages are that the Raman scattering efficiency is low relative to that obtained with visible excitation, due to the ν^4 dependency (ν^4 refers to Rayleigh scattering of the Raman light). This is compensated to some extent by employing a high throughput (Jacquinot advantage) interferometer. However, this works better for macro rather than micro samples. This is because the coupling of a microscope to an FT Raman spectrometer has a fundamental drawback; the microscope is throughput limited, so the high throughput advantage of the FTR spectrometer cannot be realized in FTR microscopy. The image of the Jacquinot stop (the large circular aperture which is the entrance to the interferometer) at the sample plane typically has a diameter of a few hundred μm, which is much larger than particles with diameters of ca. 1 μm having similar dimensions to the waist diameter of the 1064 nm light excitation at its diffraction-limited focus. Consequently the throughput advantage of the interferometer is not fully exploited and there is a trade-off of spatial resolution with signal-to-noise at the detector. For this reason, the typical spatial resolution that is achieved is in the range of 15–100 μm rather than 4 μm, as claimed in the literature. Bruker has commercialized an FTR microscope by coupling an optical microscope to an FTR spectrometer with near infrared optical fibers.

An approach that offers more promise for reducing fluorescence and achieving spatial resolution close to the diffraction limit is to use near infrared excitation in conjunction with dispersive Raman microscopy. For example, semiconductor lasers operating from 785 to 852 nm can be used in conjunction with sensitive multichannel silicon-based CCD arrays. If longer excitation wavelengths are necessary, in order to overcome the problems with fluorescence, then diode lasers can be used that emit at longer wavelengths in conjunction with dispersive Raman spectrometers equipped with multichannel near infrared (NIR) detectors, e.g., Ge- or InGaAs-array detectors.

The applications of Raman microscopy cover many areas, including material science, the earth sciences, environmental science, biology and medicine, forensic science, and even the analysis of artworks. These areas are too wide ranging to be described in detail here, but the interested reader is referred to the Further Reading section at the end of this article.

Characteristics of Laser Sources

The laser is an excellent light source for Raman spectroscopy to such an extent that the terms ‘Raman spectroscopy’ and ‘laser Raman spectroscopy’ are synonymous for all but the most esoteric experiments, for example, with synchotron sources. Indeed, the advent of the laser was the stimulation for the renaissance of Raman spectroscopy in the 1960s, given its special properties, such as monochromaticity, high intensity, beam collimation, and coherence.

The characteristics of laser light and the advantages it offers to Raman microscopy are now considered.

Beam Quality

The light inside a laser tube is formed from a number of standing waves having distinct vibrational modes. There are a limited number of these modes transverse to the beam and these are characterized by the TEM$_{pq}$ number (where p and q can be 0, 1, 2,...) where TEM is an acronym for ‘transverse electromagnetic’. When a laser is operating in its fundamental TEM$_{00}$ mode, light rays are reflected on axes between the end mirrors of the laser cavity. The ‘00’ indicates that there are no nodes in the beam profile (Figure 2a), and the laser beam has a Gaussian intensity profile in the radial direction:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \qquad [3]$$

where $I(r)$ is the irradiance at a radial distance r from the axis of the beam, I_0 is the axial irradiance, and w is the beam radius.

For higher-order modes, a number of nodes are observed in the beam profile, which arise from off-axis light rays being reflected between the end mirrors, for example, the TEM$_{11}$ mode has two nodes which are mutually perpendicular (Figure 2b).

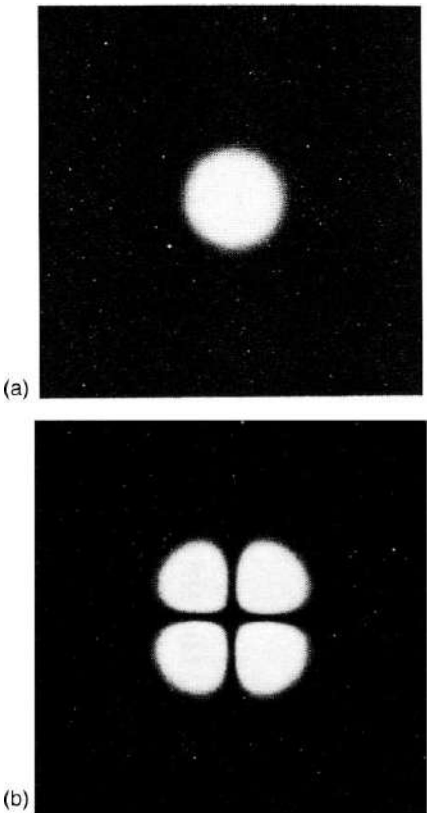

Figure 2 Transverse electromagnetic modes formed with confocal concave mirrors, (a) TEM$_{00}$, and (b) TEM$_{11}$. (Reproduced with permission from Young M (1977) *Optics and Lasers: an Engineering Approach.* Berlin, Heidelberg: Springer-Verlag

The propagation characteristics of a Gaussian beam can be fully defined, either by the diameter of the beam waist or by the far-field divergence. Consequently, it is only necessary to know the diameter of the beam waist ($2w_0$), or the diameter of the beam ($2w_z$) at a longitudinal distance z from the waist, in order to determine the propagation characteristics of the beam:

$$w(z) = w_0\sqrt{1 + \left(\frac{z}{z_\mathrm{R}}\right)^2} \qquad [4]$$

$$R(z) = z\left[1 + \left(\frac{z_\mathrm{R}}{z}\right)^2\right] \qquad [5]$$

where the quantity $z_\mathrm{R} = \pi w_0^2/\lambda$ is known as the Rayleigh range of the beam, λ is the wavelength of the laser radiation, and $R(z)$ is the radius of curvature of the wavefront at a distance z from the beam waist.

The wavefront is planar at the minimum beam waist and the Rayleigh range is the distance from the beam waist to the location at which the wavefront is most curved (Figure 3), the region from the waist to

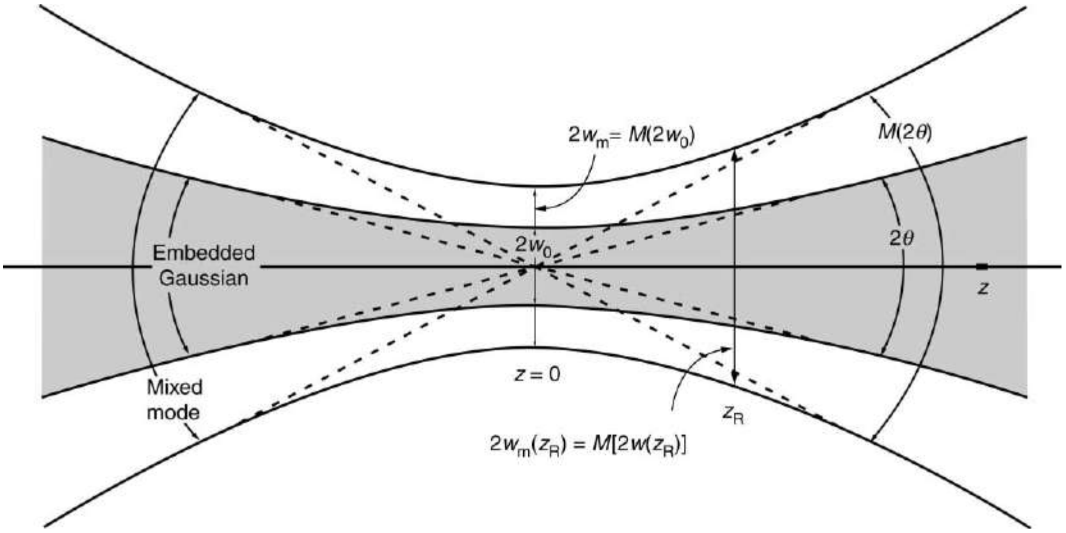

Figure 3 The Rayleigh range and embedded Gaussian (adapted from http://beammeasurement.mellesgriot.com/tut_m2.asp).

the Rayleigh range being the near field. Beyond approximately ten times the Rayleigh range, in the far field, the beam diverges as a cone with approximately straight sides. It can be seen by substituting $z = z_R$ into eqn [4] that the beam diameter at the Rayleigh range is $\sqrt{2}$ times the waist diameter.

Unfortunately laser beams do not conform to pure Gaussian functions and therefore they do not propagate according to the above equations. As a result, a dimensionless beam propagation parameter was devised in the early 1970s. This parameter, known as the M^2 factor or the 'times diffraction limit', is based on the brightness theorem, which states that for any laser beam the product of the beam diameter and the far-field divergence is a constant. Thus, M^2 is defined as the ratio of the laser beam's multimode diameter-divergence product to the ideal diffraction-limited (TEM_{00}) beam diameter-divergence product:

$$M^2 = \left(\frac{2w_m \, \Theta_m}{2w_0 \, \theta_0} \right) \qquad [6]$$

where Θ_m is the laser beam's multimode divergence, θ_0 is the theoretical diffraction-limited divergence, $2w_m$ is the laser beam's multimode waist diameter, and $2w_0$ is the ideal diffraction-limited beam waist diameter.

Alternatively, M^2 can be defined as the square of the ratio of the multimode beam diameter ($2w_m$) to the diffraction-limited beam diameter ($2w_0$):

$$M^2 = \left(\frac{2w_m}{2w_0} \right)^2 \qquad [7]$$

The intensity of the beam has a Gaussian distribution in the radial direction, but the accepted definition of beam diameter is the distance at which the intensity has fallen to $1/e^2$ (i.e., ca. 13.5%) of its axial intensity, as can be seen by setting $r = w$ in eqn [3] above. This definition only applies properly to Gaussian beams; for other beam profiles the diameter can be calculated using a second moment measurement of the entire beam.

The real beam, then, can be treated as Gaussian by substituting an artificial wavelength $M^2\lambda$ into the equations that apply to an ideal diffraction-limited TEM_{00} beam. The ideal TEM_{00} Gaussian can be superimposed as an 'embedded Gaussian' on the optical diagram of the real multimode beam (Figure 3). It can be seen that the multimode beam has a waist diameter and a divergence that are both M times larger than those of the embedded Gaussian.

The M^2 factor has a value of unity for an ideal diffraction-limited TEM_{00} beam, and values of greater than unity for all other beams; it can be as high as several hundred for a distorted beam of poor quality. This is an important parameter where a tightly focused laser spot (e.g., confocal Raman microscopy) or low divergence (e.g., remote sensing) are required.

Polarization

Although polarization is not an inherent property of laser light, lasers provide plane polarized light when the end windows of the laser tube are mounted at Brewster's angle (θ_B):

$$\theta_B = \tan^{-1} n \qquad [8]$$

where n is the refractive index of the window material for the appropriate wavelength of laser light. For this angle, reflected light is completely plane polarized and consequently the laser light that is generated is also plane polarized. Brewster windows are often used inside the laser cavity in order to reduce reflection losses. This is important in low-gain lasers such as HeNe lasers where laser action could be prevented by reflection losses. Air-cooled ionic gas lasers typically deliver highly linearly polarised output with ratios exceeding 1000:1.

The polarization property of the exciting light is used a great deal in Raman spectroscopy. In combination with the polarization properties of the Raman scattered light, it can be used to determine the depolarization ratios, ρ, of Raman bands in the solution state. In solid state studies it can be used to determine the degree of orientation, for example, in polymeric fibers or films, and in single crystal studies it can be used to determine the symmetry classes of the Raman active vibrations. In such studies a linear polarizer is used to improve the degree of plane polarization of the laser light, and a half-wave plate is used to rotate the plane of polarization by 90°.

Plane polarized laser light can be converted to circularly polarized laser light which is required for Raman optical activity (ROA) studies. For example, plane polarized 514.5 nm laser light provided by an argon ion laser can be converted to circularly polarized light by an electro-optic modulator. As CCD detectors are usually employed in present-day ROA experiments and they have relatively long sampling times, slow polarization modulation is required which can be achieved by periodically rotating a quarter-wave plate or by applying a square quarter-wave voltage to a Pockels cell.

Whereas optical isomers or enantiomers give identical Raman spectra, they are distinguished in the ROA experiment by the sign of the ROA signal. Indeed, the sign can be used to determine the absolute configuration of a chiral molecule. Also the enantiomeric excess of mixtures of enantiomers can be determined in such experiments, with obvious applications in the pharmaceutical industry, and the conformations of biological molecules, for example, proteins, nucleic acids, sugars, and viruses, can be determined.

Up until now, the ROA technique has been practised by no more than a handful of research groups world-wide, but it is expected that this situation will change in the near future, since the first commercial ROA spectrometer (the ChiralRAMAN spectrometer) has recently been launched by Biotools Inc. This spectrometer makes use of a solid state laser as the source of the polarized laser light.

Beam Divergence

Referring to Figure 3, the full divergence angle, Θ, for the fundamental TEM_{00} Gaussian beam is given by:

$$\Theta = \lim_{z\to\infty} \frac{2w(z)}{z} = \frac{2\lambda}{\pi w_0} \qquad [9]$$

From eqn [4] it can be seen that the spot size increases linearly with the longitudinal distance z and, from eqn [9], that it diverges at a constant cone angle, Θ. It can also be seen from eqn [9] that the smaller the spot size, w_0, the greater the beam divergence angle, Θ.

Indeed, a highly divergent beam is a problem with edge-emitting laser diodes due to the small dimensions of the light source. Furthermore, it is much smaller in the vertical direction than in the horizontal, giving rise to an elliptical output beam that diverges much more rapidly in the vertical direction than in the horizontal. The highly divergent, elliptical beam can be corrected, to an extent, with a cylindrical lens, but the inherent problem of a small, elliptical source can never be completely corrected.

In contrast, vertical-cavity surface-emitting semiconductor lasers (VCSELs) do not have the same limitations on beam divergence because the cavity is in the vertical direction and the light emission occurs from the surface, which has a much larger area than the light source of an edge emitter. Indeed, frequency-doubled green and blue semiconductor lasers are now available, having beams with M^2 values of less than 1.1 and beam divergences of a few milliradians. In terms of these beam characteristics, it appears that these lasers are beginning to challenge the argon ion laser.

Emission Linewidth

The laser linewidth limit, $\Delta\nu_{\mathrm{L}}$, is given by the Schawlow–Townes expression:

$$\Delta\nu_{\mathrm{L}} = \frac{2\pi h\nu(\Delta\nu_{\mathrm{c}})^2}{P_{\mathrm{out}}} \qquad [10]$$

where $\Delta\nu_{\mathrm{c}}$ is the linewidth of the passive resonator, $h\nu$ is the photon energy, and P_{out} is the laser power output.

The power output, P_{out}, is equal to the number of photons, N_{p}, in the resonator times the energy per photon, $h\nu$, divided by the photon lifetime, τ_{c}:

$$P_{\mathrm{out}} = \frac{N_{\mathrm{p}}h\nu}{\tau_{\mathrm{c}}} \qquad [11]$$

Furthermore, the width $\Delta\nu$ of the resonance curve, at which the intensity has fallen off to half the maximum value, is given by:

$$\Delta\nu_c = (2\pi\tau_c)^{-1} \qquad [12]$$

Substituting for P_{out} and $\Delta\nu_c$ from eqns [8] and [9] into eqn [10] gives the limit for the linewidth, $\Delta\nu_L$:

$$\Delta\nu_L = \frac{\Delta\nu_c}{N_p} \qquad [13]$$

The linewidth of a laser is characterized by its Q factor:

$$Q = \frac{\nu}{\Delta\nu} \qquad [14]$$

where ν is the frequency of the laser line and $\Delta\nu$ is the linewidth.

Other useful relationships involving linewidth parameters are:

$$Q = \frac{\lambda}{\Delta\lambda} \qquad [15]$$

and

$$\Delta\tilde{\nu} = -\frac{\Delta\lambda}{\lambda^2} \qquad [16]$$

where λ and $\Delta\lambda$ are the wavelength and linewidth (in units of metres), respectively, and $\Delta\tilde{\nu}$ is the linewidth expressed as a wavenumber (in units of cm^{-1}).

The Q factor can be as high as 10^8 which is of great use in high resolution spectroscopy.

Visible Lasers

Helium-Neon Laser

The helium-neon laser is continuously pumped electrically using a high dc voltage of up to 1 kV. The gain medium consists of a gas mixture of about 10 parts helium to each part of neon at a pressure of about 3 Torr. The gain of this medium is extremely low, hence the requirement for Brewster-angle windows to eliminate reflection loss of the light polarized in the plane that includes the axis of the laser and the normal to the Brewster window. Due to this requirement, the light output is necessarily plane polarized.

The pumping excites helium atoms by electron impact, and resonant energy transfer to neon atoms then occurs via collisions of the gaseous atoms (Figure 4). This creates a population inversion in the neon atoms, enabling the laser transition to occur at 632.8 nm. Following emission, the neon atoms decay nonradiatively down to their metastable $2p^5 3s^1$ level from which they decay back down to the ground state via collisional de-excitation with the walls of the tube. This mechanism restricts the maximum achievable power output to ca. 50 mW, because of the need to depopulate the metastable neon atoms by wall collisions.

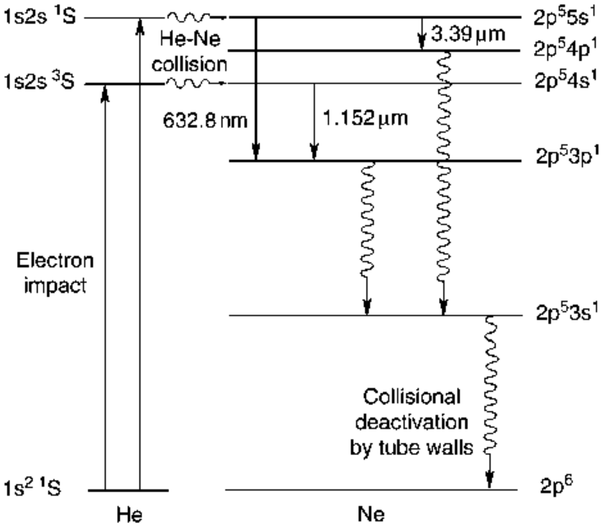

Figure 4 Energy levels of the He-Ne laser.

An advantage of using helium-neon lasers for Raman spectroscopy is that they generally operate in the fundamental TEM_{00} mode, which is critically dependent on the ratio of the length of the tube to its diameter. For this reason they are designed to have small tube diameters of around a few millimetres and lengths of 0.15–0.50 m, the small diameter also aiding collisional de-activation with the walls due to the relatively large surface to volume ratio of the tube. Another advantage of this laser for Raman spectroscopy is that the linewidth of the 632.8 nm emission line is ca. 1.5 GHz.

A disadvantage of the helium-neon laser is that it emits a large number of spontaneous emission lines originating from the excited neon atoms. These plasma emissions are observed in the Raman spectrum as sharp lines unless they are filtered out, for example, by a pre-monochromator or an interference filter. It should be mentioned, in passing, that these plasma lines are not always unwanted, because they can be very useful for wavenumber calibration of the Raman spectrometer.

The low output power of <50 mW is not generally a disadvantage for Raman microscopy since higher laser powers can often cause photolytic or thermal degradation of the sample. This is due to the high irradiance at the sample when the laser light is focused to a tight spot by the microscope objective. The excitation wavelength of 632.8 nm is suitable for combined use of the laser with a silicon-based CCD detector in Raman spectroscopy, because large Stokes shifts in the 3000–4000 cm^{-1} region lie well within the quantum efficiency curve of this detector. It is also possible, though less common, to operate the helium-neon laser on weaker transitions that include the green 543.5 nm line.

Argon Ion Laser

The argon ion laser was one of the first lasers to be discovered following the invention of the laser and up until now it has been used extensively for Raman spectroscopy, among many other applications. Excitation is provided by a continuous electrical discharge, and because of the high energy required to ionize the argon atoms and then promote the ions to an excited state, the efficiency of the laser is very small (ca. 0.1%). In spite of this, once population inversion has been achieved, the gain of the laser is very high and output powers of up to 25 W can be obtained for the strong lines at 488.0 and 514.5 nm and up to 50 W for multiline operation.

An advantage of the argon ion laser is that it can provide emission at more than 35 discrete wavelengths, the strongest of which are listed in Table 1. These lie in the green, blue, and near ultraviolet regions of the spectrum, a number of the ultraviolet lines only being obtained on the larger frame argon ion lasers. Discrete laser emission lines are selected by tuning a prism or grating inside the cavity. The argon ion laser can also be operated in multiline mode, for example, for pumping dye or Ti:sapphire lasers.

Table 1 Continuous wave argon ion laser wavelengths and output powers for the Coherent Inc. Innova 300 argon ion laser system

Wavelength (nm)	*Power (mW)*
228.9	30
238.3	100
244.0	400
248.3	180
257.3	750
275.4	5
300–305.5	20
333.4	40
333.8	30
335.8	20
351.1	200
351.4	60
363.8	240
454.5	140
457.9	420
465.8	180
472.7	240
476.5	720
488.0	1800
496.5	720
501.7	480
514.5	2400
528.7	420

The five lines below 260 nm are frequency doubled.

The linewidths of the argon ion emission lines at 488.0 and 514.5 nm are around 4.0 GHz. The high temperature laser tube has a diameter of approximately 12 mm and a length in the range of 0.1 to 1.8 m. A 240 V three phase power supply and ca. 35 A are required for pumping a medium power 5 W argon ion laser. The tube also requires water cooling at a flow rate of ca. 10 L/min and a pressure of 25 psi because of the large amount of heat dissipation.

For Raman microscopy smaller, air-cooled argon ion lasers, which only require a 240 V single phase power supply, can be used to provide output of a few hundred milliwatts on the 488.0 and 514.5 nm lines. For this application, the excitation is provided by argon ion lasers, which are designed to operate in the TEM_{00} mode.

Until recently, the argon ion laser has been a workhorse as the most common excitation source for Raman spectroscopy, but it is now losing ground to solid state lasers. The principle reasons for this are that the latter are much more efficient and consequently do not in general have three phase power and water-cooling requirements. Nevertheless, argon ion lasers still have a niche when high excitation powers on the order of watts are required in conjunction with a good beam quality ($M^2 < 1.1$), for example, for Raman spectroscopy of gaseous samples. Another advantage of argon ion lasers over solid state lasers is their multiline capability.

Krypton Ion Laser

The krypton ion laser has useful discrete emission lines in the near ultraviolet, blue, yellow, red, and near infrared regions of the electromagnetic spectrum; the strongest lines of a Coherent Inc. Innova 400 krypton ion laser are listed in Table 2.

The argon ion and krypton ion lasers are close relatives, thus the large frame krypton ion laser has similar power and water-cooling requirements to those mentioned above for the argon ion laser. Also, like the argon ion laser, air-cooled models with lower power output are available. The characteristics of the two lasers are also similar; for example, the linewidths of the krypton ion transitions at 530.9, 568.2, and 647.1 nm are about 4.0 GHz, the laser tube has similar dimensions (0.1–2.0 m in length) and the laser can be operated in either TEM_{00} or multimode.

The krypton ion laser is even less efficient than the argon ion laser, consequently lower-power outputs of

Table 2 Continuous wave krypton ion laser wavelengths and output powers for the Coherent Inc. Innova 400 krypton ion laser system

Wavelength (nm)	*Power (mW)*
206.5	4
234	8
337.5–356.4	2000
406.7	900
413.1	1800
415.4	280
468.0	500
476.2	400
482.5	400
520.8	700
530.9	1500
568.2	1100
647.1	3500
676.4	900
720.8	45
752.5	1200
793.1–799.3	300

The 206.5 and 234 nm lines are frequency doubled.

up to about 20 W can be achieved when operated in multiline mode.

Mixed argon and krypton ion laser tubes are also commonly used that provide laser lines originating from both argon ion and krypton ion transitions. As with the individual argon ion and krypton ion lasers, sharp plasma lines, due to spontaneous emission from the rare gas ions, are observed in the Raman spectrum unless they are filtered out. If interference filters are used for this, each laser line requires its own filter that is tailored to the specific emission wavelength.

HeCd Laser

Laser lines at 325 and 442 nm, both capable of providing milliwatts of output power, can be obtained from the helium-cadmium laser. These emission lines result from electronic transitions in free cadmium atoms.

It is obviously advantageous to use blue, rather than longer wavelength, excitation (e.g., provided by the 442 nm line of the HeCd laser) for off-resonance Raman spectroscopy, due to the ν^4 dependence of the Raman light scattering efficiency, provided that the efficiencies of the illumination/collection optics and Raman spectrometer throughput are optimized in the blue region of the spectrum. Unfortunately, far more samples fluoresce when excited with blue light than with red or near infrared radiation, which explains why red or near infrared lasers (e.g., HeNe, semiconductor lasers) are more commonly used for general Raman applications.

Near Infrared Lasers

Although the traditional laser systems for Raman spectroscopy have been the argon ion, krypton ion, and helium-neon lasers for discrete excitation wavelengths, the diode laser has gained popularity over recent years.

The principal advantage of near infrared semiconductor lasers for Raman microscopy is that they generally excite less fluorescence in the Raman spectrum than visible lasers due to their longer wavelength. Commonly available wavelengths are 670, 785, 830, and 852 nm and, of these, the first two can be used in conjunction with silicon-based CCD detectors to give Stokes Raman shifts over the whole range of fundamental vibrations, up to 4000 cm^{-1}. However, for excitation wavelengths of 830 and 852 nm, only Stokes Raman bands in the fingerprint region can be detected with a silicon-based CCD detector because higher wavenumber Raman shifts approach the bandgap of the silicon semiconductor ($\lambda >$ ca. 1050 nm).

The disadvantages of diode lasers are that they cannot supply high power of narrow linewidth in single-mode and they are susceptible to mode hopping which gives rise to a shift in the excitation wavelength. This latter drawback is particularly disadvantageous for Raman spectroscopy because it results in a shift in the wavenumber positions of the Raman bands. When good beam quality is not required, broad stripe, high power laser diodes can be used; these can have output powers that are greater than 1 W but their emission linewidths are equal to ca. 2 nm, which is far too wide for Raman spectroscopic applications. These laser diodes find applications as pumps of solid state lasers, for example, Nd:YAG, Nd:YVO_4 lasers, however. Additionally, amplified stimulated emission (ASE) gives an unwanted background signal that can be very broad (ca. 20–30 nm) and 0.1–1% of the intensity of the laser line. This necessitates the use of bandpass filters in order to reduce this unwanted background.

Although laser diodes having an output of less than 200 mW can operate in TEM_{00} and in single longitudinal modes, mode hopping can occur due to optical feedback or fluctuations in environmental factors and this causes severe broadening of the linewidth. The sensitivity to optical feedback can be greatly reduced, hence the laser linewidth can be narrowed appreciably, by confining the frequency of the photons either in an internal cavity containing a small diffraction grating or in an external cavity. The former types of laser are sometimes called 'distributed feedback' (DFB) lasers because the feedback is distributed over the length of the grating, rather than occurring all at once at a mirror. The wavelength that is fed back is determined by the period of the grating. Usually, a DFB laser has a grating fabricated into the entire length of the laser. A variation referred to as a distributed Bragg reflector (DBR) laser has a distinct grating fabricated into the substrate on each side of the active area. The external cavity semiconductor laser (ECSL) is fabricated by placing the laser diode in a separate resonator like a conventional gas or solid-state laser. The DBR and ECSL lasers are now discussed.

Distributed Bragg Reflector Lasers

The DBR laser consists of a grating on each side of the active region (Figure 5); these gratings act as mirrors having a reflectivity that is optimized at one particular wavelength, in addition to narrowing the laser linewidth. At present, DBR lasers are only commercially available having an excitation wavelength of 852 nm.

The emission linewidth is less than 4 MHz, which is suitable for the vast majority of Raman spectroscopic applications and the laser operates in single TEM_{00} mode. The disadvantages of DBR lasers are that the maximum output power is restricted to around 150 mW, so an optical isolator is usually necessary in order to prevent external facet damage caused by external optical feedback, and only an excitation wavelength of 852 nm is currently available.

External-Cavity Semiconductor Lasers (ECSL)

ECSLs provide higher output power (up to ca. 1 W) and a wider range of excitation wavelengths (630 nm to around 850 nm), as well as being less expensive than DBR lasers. They use the semiconductor chip only as the gain medium and employ an external grating, both as the frequency selector and the reflective mirror. Specific excitation wavelengths are becoming widely adopted in Raman microscopy employing single grating spectrograph in conjunction with CCD detection, such as 670, 785, 830, and 852 nm. The ECSLs have emission linewidths as low as a few MHz, but in general they span the range 2 MHz–30 GHz, and can operate in a nearly diffraction-limited transverse mode.

Three different cavity designs employing a diffraction grating have been discussed in the literature: Littrow type-I, Littrow type-II, and Littman configurations. In the Littrow type-I design, the laser diode light is incident on the grating at an angle of incidence equal to θ_{Littrow} (Figure 6a). The diffracted light in first order is re-directed back to the laser diode to provide feedback and the output is the zeroth order (i.e., specularly reflected) light. A disadvantage of this design is that the bandwidth of the grating is relatively large because a single pass geometry is used. In the Littrow type-II design, light from the rear facet of the laser diode is collimated by a lens and directed on to a grating at an angle of incidence equal to θ_{Littrow} (Figure 6b). The diffracted light in first order is re-directed back to the laser diode, in a similar fashion to the Littrow type-I design, in order to provide optical feedback, and the output is the laser light emitted from the front facet. Disadvantages of the Littrow type-II design are the requirement for

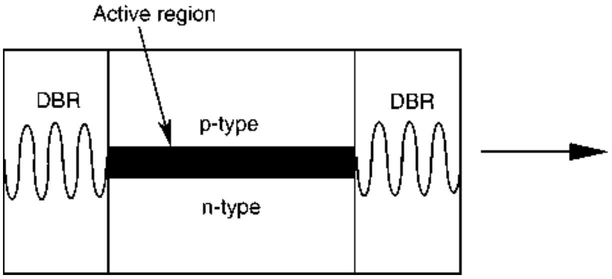

Figure 5 Diagram of the cavity of the DBR laser (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

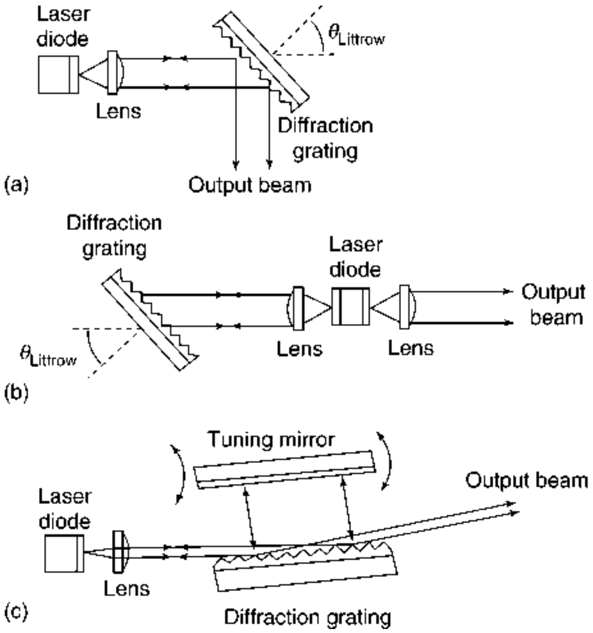

Figure 6 Diagrams of the cavities of (a) Littrow type-I, (b) Littrow type-II, and (c) Littman external cavity diode lasers (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

custom fabrication of antireflection coatings on both facets and the sensitivity to external optical feedback. In the Littman design, the laser diode light is collimated by a lens and directed on to a grating at a grazing angle of incidence. The diffracted light is reflected by a mirror and diffracted back to the laser diode by the grating in order to provide optical feedback (Figure 6c). An advantage of this design is that the grating bandwidth is less than half that of the Littrow designs due to the double pass geometry, but a disadvantage is that the external cavity length is longer, resulting in a narrower spacing of the cavity modes. Another advantage of this design is that the tuning is accomplished by rotating the mirror instead of the grating, thus the alignment of the output beam is not altered when tuning. For Littrow type I and II and Littman ECSL designs, filtering of the ASE is required, as it can have an intensity of ca. 0.1 to 1% of the intensity of the laser line. Thus, these designs necessitate the incorporation of a bandpass filter (having a rejection of better than 10^{-4} at the ASE).

Nd:YAG Laser

Under normal conditions, the Nd:YAG laser oscillates on a transition in the near infrared at 1064 nm, and this is the excitation line that is used in most commercial Fourier Transform Raman spectrometers. The gain medium is a crystal of yttrium aluminium garnet ($Y_3Al_5O_{12}$, YAG) doped with ca. 1.0 mol% Nd^{3+} cations that substitute Y^{3+} ions in the cubic YAG lattice.

The Nd:YAG laser is a four-level system (Figure 7), which has high gain and low threshold due to the narrow fluorescent linewidth of the laser transition. Absorption bands of the Nd^{3+}ions around 808 nm conveniently match the energy of commercially available high-power multimode diode lasers that serve as the pump. The Nd^{3+} ions decay nonradiatively to the upper laser level, the $^4F_{3/2}$ state, thereby creating a population inversion. This is because the lower laser level, the $^4I_{11/2}$ state, has no appreciable thermal population at room temperature, since it is $>2000\ cm^{-1}$ above the $^4I_{9/2}$ ground state, and the $^4F_{3/2}$ excited state has a relatively long lifetime of 230 μs.

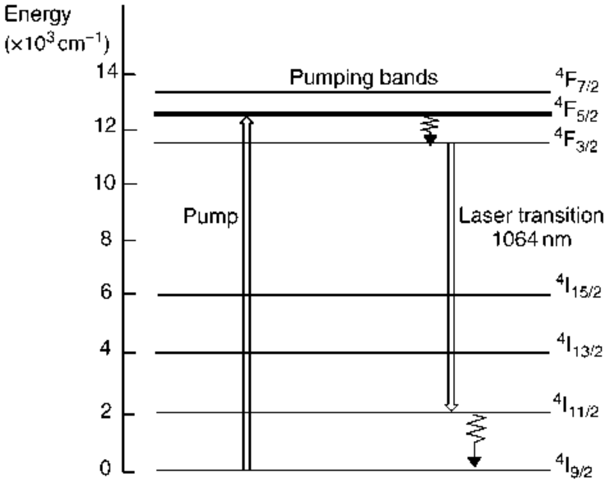

Figure 7 Energy levels of the Nd:YAG laser.

The high thermal conductivity of Nd:YAG enables the laser to operate at high power in either continuous wave (CW) or *Q*-switched modes, and the diode pumped solid state (DPSS) variety of the cw Nd:YAG laser can currently achieve an output power in excess of 20 W on the 1064 nm line. The laser can be designed to operate in the fundamental TEM_{00} mode and the full width at half maximum (FWHM) linewidth of the spontaneous emission of the 1064 nm laser transition is 120 GHz (ca. 0.45 nm). Advantages of the diode pumped Nd:YAG laser are that it is air-cooled and can be operated at 240 V single phase. Also, being all solid state and having a small footprint, it is robust and portable. The lifetime of the laser is dependent on the laser diodes used for pumping, but some Nd:YAG laser designs enable these to be replaced in the field.

The frequency-doubled Nd:YAG laser emitting green light, having a wavelength of 532 nm, is also commonly used nowadays for dispersive Raman spectroscopy, and an output power of 10 W on the 532 nm laser line is provided by commercially available frequency-doubled diode-pumped Nd:YAG lasers.

It is worth mentioning that Nd:YAG lasers have also been fabricated that emit either at 946 or 1330 nm, by suppressing the strong emission at 1064 nm and optimising the optics for the desired wavelength.

Nd:YVO_4 and Nd:YLF Lasers

An intracavity, frequency-doubled DPSS Nd:YVO_4 laser has recently become commercially available, and it has some advantages over the Nd:YAG laser including a larger stimulated emission cross-section and a higher absorption coefficient (along the extraordinary direction of the birefringent crystal). It has the same emission wavelength as the frequency-doubled Nd:YAG laser, i.e., 532 nm, and Nd:YVO_4 is the material of choice for cw end-pumped lasers having around 5 W output power. This is because the diode laser pump beam is tightly focused in the end-pumped system, but a small waist diameter cannot be retained over a distance of more than a few millimeters, consequently a high absorption coefficient and gain are very beneficial.

The gain medium of the Nd:YLF laser consists of a crystal of yttrium lithium fluoride (YLF) that is doped with Nd^{3+} ions on the Y^{3+} cation sites. Unlike the Nd:YAG and $Nd:YVO_4$ lasers, the emission does not occur at 1064 nm; instead the $^4F_{3/2}-^4I_{11/2}$ emission occurs at wavelengths of either 1047 or 1053 nm, depending on the polarization that is selected. The former is due to extraordinary polarized light, whereas the latter is due to ordinary polarization, and either of these emission wavelengths can be selected using an intracavity polarizer. The Nd:YLF laser can offer benefits in *Q*-switched operation when the longer fluorescence lifetime (480 μs) of Nd^{3+} ions in the $^4F_{3/2}$ state enables a higher energy to be stored for the same number of pump laser diodes.

Ti:Sapphire Laser

The Ti:sapphire laser is tunable over the approximate range of 670–1070 nm with the peak of the gain curve at ca. 800 nm. In the gain medium, ca. 0.1% by weight Ti^{3+} is doped into a crystal of sapphire grown by the Czochralski method. The Ti^{3+} ions substitute for Al^{3+} ions in the Al_2O_3 of the sapphire and the laser emission is due to the $^2E-^2T_2$ transition of the Ti^{3+} cation, which has a $3d^1$ valence electronic configuration (Figure 8). The laser has a large stimulated emission cross-section but the fluorescence lifetime of the upper laser level (^2E state) is quite short (3.2 μs), thus the laser is usually laser (e.g., by argon ion or frequency-doubled Nd:YAG or $Nd:YVO_4$ lasers) rather than flashlamp pumped because a very high pump flux is required.

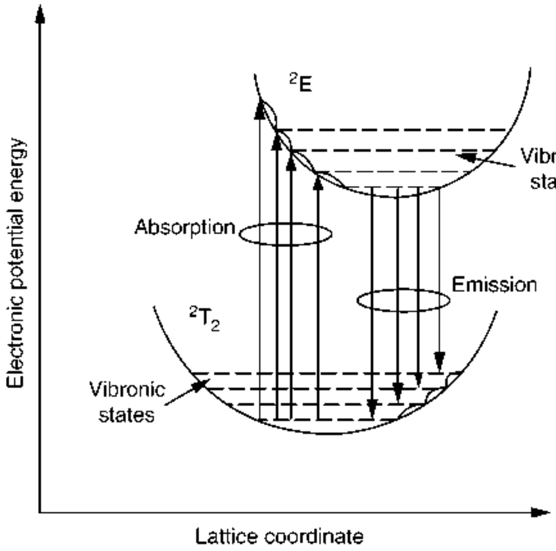

Figure 8 Energy levels of the Ti:sapphire laser (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

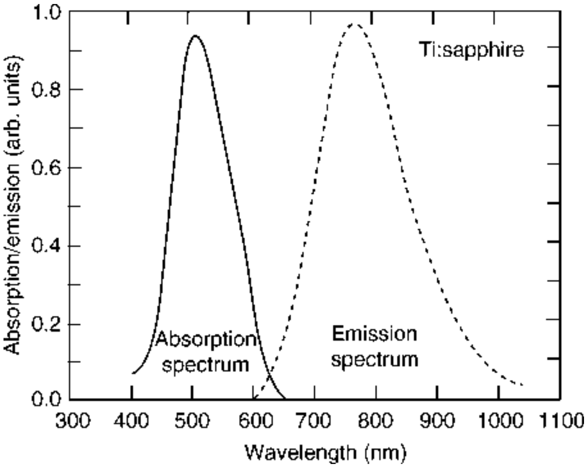

Figure 9 Absorption and emission spectra of the Ti^{3+} ion in sapphire. Al_2O_3 (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

The absorption and emission bands are broad and widely separated, due to the vibronic coupling between the Ti^{3+} host and the Al_2O_3 lattice (Figure 9). The lower laser level is any one of the vibronic levels of the $^2T_2$ state. Following the laser transition, the Ti^{3+} ions decay from the upper vibronic levels of the $^2T_2$ state down to the lower vibronic levels. Hence, this is a four-level laser.

The broad spontaneous emission linewidth of the $^2E-^2T_2$ laser transition is around 100 THz, the ouput power can be as high as 50 W, and the laser can be operated in either TEM_{00} or multimodes.

UV Lasers

UV laser sources offer numerous advantages over visible laser sources for Raman spectroscopy. A major consideration is that many analytes have absorptions in the near UV, making them amenable to resonance Raman spectroscopic studies. The signal enhancement (up to 10^6), that can be achieved in resonance Raman spectra, results in a large increase in detection sensitivity. Furthermore, some vibrational modes, which normally give rise to weak bands in the off resonance Raman spectrum, can show strong enhancement in the UV excited resonance Raman spectrum. A good example of this is the amide II band of peptides and proteins, which is due to a combination of N–H in plane bending and C–N stretching modes of the peptide linkage. This band is normally weak in the Raman spectrum but strong in the infrared spectrum; however, the amide II band can show strong enhancement in the UV excited resonance Raman spectrum and this can be useful for determining secondary structure in proteins. A further benefit of UV excited Raman spectroscopy is that fluorescence can often be avoided as it tends to occur at a lower energy, outside the Stokes Raman spectral window.

It can be advantageous to use cw rather than pulsed excitation for UV resonance Raman spectroscopy, if the analyte has a long excited state lifetime relative to the pulsewidth of the pulsed laser. This is because the concentration of analyte molecules in the ground state is significantly depleted during pulsed excitation, due to Raman saturation giving a lower signal to noise ratio in the Raman spectrum than for the case of cw excitation. It has been found that pulse energy flux densities should be less than 1 mJ/cm^2 and flowing or spinning samples should be used, in order to ensure that nonlinear phenomena and saturation effects do not occur.

For UV Raman microscopy, where the laser beam is focused into a small spot of micrometer dimensions on the sample, it is essential to avoid using pulsed lasers delivering high peak powers. This is because such pulsed lasers can cause dielectric breakdown, and even at lower peak powers they can cause nonlinear effects and Raman saturation phenomena. Consequently, with UV Raman microscopy, one is restricted to the use of cw or quasicontinuous laser excitation.

It is only recently that high thoughput UV Raman microspectrometers have been developed. It has been demonstrated for visible Raman spectroscopy that a significantly higher throughput can be achieved by employing a single stage spectrograph with a notch filter instead of a double or triple monochromator. In the UV region, blocking the Rayleigh scattering is a problem because notch filters are not currently available. However, it has been discovered that this can be overcome by introducing two modifications to the design of the optical layout of a visible Raman microscope. First, two novel dielectric longpass filters were used instead of a notch filter in order to reject the elastically scattered light. Second, an all-reflecting Cassegrain microscope objective was used, instead of a lens, in order to block the specular reflection, and thereby further reduce the Rayleigh scattering background. These design modifications have enabled a single stage spectrograph with a holographic grating to be used to disperse the Stokes Raman radiation in a UV Raman microspectrometer (Figure 10).

In the optical layout of Figure 10 it can be seen that the laser excitation is not focused by the Cassegrain microscope objective because an epi-illumination configuration is not employed. Instead the laser is focused by a separate lens and directed on to the sample with a turning prism located directly underneath the Cassegrain objective. This minimizes loss of laser beam throughput or scattered light throughput to the spectrometer at the expense of spatial resolution, since the laser light is focused by a longer focal length lens than the microscope objective. An additional difference between the optical layouts of the visible and UV Raman microscopes shown in Figures 1 and 10, respectively, is that pinhole spatial filters are not shown in the latter. However, better axial spatial resolution could be achieved if this UV Raman microscope was made confocal by introducing a pinhole and lens.

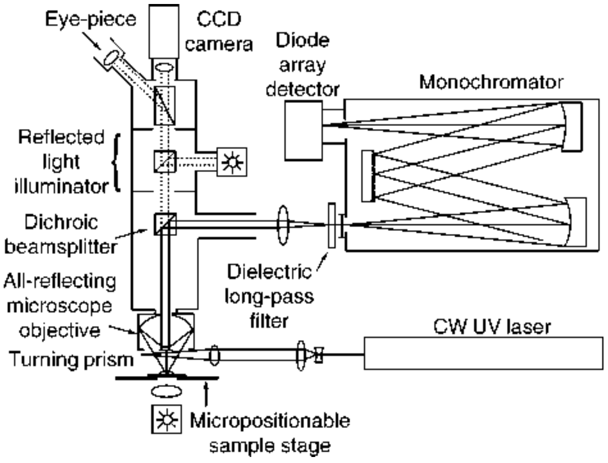
Eye-piece
CCD camera
Diode array detector
Monochromator
Reflected light illuminator
Dichroic beamsplitter
All-reflecting microscope objective
Turning prism
Dielectric long-pass filter
CW UV laser
Micropositionable sample stage

Figure 10 Diagram of the optical layout of a UV Raman microscope (adapted from Pajcini V, Munro CH, Bormett RW, Witkowski RE and Asher SA (1997) UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51: 81–86).

Frequency-Doubled Argon Ion Laser

The frequency doubled argon ion laser is a popular choice among UV laser sources. The cw UV laser contains a nonlinear optical beta-barium borate (BBO) crystal, which is located within the laser cavity; this crystal frequency doubles the strong Ar$^+$ lines to give five UV lines below 260 nm which are listed in Table 1 for the Coherent Inc. Innova 300 Ar$^+$ ion laser system. Of these UV lines, the 228.9 nm line is almost ideal for resonance Raman enhancement of tyrosine and trpytophan residues in proteins and the 244.0 nm line is well suited for studying tyrosinate groups. The 244.0 nm line has also been used to excite selectively UV resonance Raman spectra from spatially resolved areas of biological samples within the nucleus of a single cell, from DNA in particular, using low laser powers and short acquisition times to keep sample damage to a minimum under the microscope.

Frequency-Doubled Krypton Ion Laser

Like the cw frequency-doubled argon ion laser mentioned above, the cw frequency-doubled krypton ion laser contains a nonlinear optical BBO crystal, which is located within the laser cavity; this crystal frequency doubles the strong Kr^+ lines to give 206.5 and 234 nm UV lines which are listed in Table 2 for the Coherent Inc. Innova 400 Kr^+ ion laser system. The 206.5 nm laser line is useful for exciting resonance Raman spectra within the $\pi-\pi^*$ amide transition of the peptide backbone of proteins, and the enhanced protein amide vibrational bands can be used to determine the protein secondary structure.

HeCd Laser

The 325 nm HeCd excitation line, in the near ultraviolet, can be used for combined micro-Raman/photoluminescence studies of, for example, semiconductors. These lasers are suitable for low-power applications, typically having output powers in the 1–100 mW range, and they can be designed to operate in TEM_{00} single-mode or multimode.

Quasi-CW Mode-Locked Ti:Sapphire Laser

The second harmonic of the mode-locked Ti:sapphire laser (ca. 350–500 nm), third harmonic (ca. 233–333 nm), and fourth harmonic (ca. 200–250 nm), are all available with conventional nonlinear crystals.

Typically, the Ti:sapphire crystal is pumped by a cw argon ion laser, for example, as the excitation source for a UV Raman microscope. This quasicontinuous laser system produces 2–3 ps pulses at a 76 MHz repetition rate and is continuously tunable over the 200–300 nm range. The typical power output is ca. 50 mW at 250 nm, ca. 10 mW at 240 nm, and 1 mW at 230 nm. In future, it is anticipated that a frequency-doubled Nd:YAG laser will be increasingly favored as the pump, creating an all-solid-state laser source.

Conclusions

Laser systems for Raman microscopy have been described in this article. Although the differences in the terms ‘Raman spectroscopy’ and ‘Raman microscopy’ only refer to whether a macro or micro sampling configuration is used, this does have a bearing on the types of laser systems that are employed. This is because the latter imposes a restriction on using low duty pulsed lasers with high pulse peak power, as both spatial and temporal confinement of the laser excitation can lead to dielectric breakdown of the sample, nonlinear phenomena, or Raman saturation. Thus, for this reason, one is limited to cw lasers or quasi-cw lasers for Raman microscopy whereas there is no such restriction for Raman spectroscopy.

The quality of the laser beam, measured by the M^2 value or ‘times diffraction limit’ influences the ultimate spatial resolution that can be achieved in Raman microscopy.

In the visible region, argon ion, krypton ion, and frequency-doubled Nd:YAG lasers have good beam quality and high power, with helium-neon and helium-cadmium lasers also having good beam quality but lower power; this makes these lasers good light sources for visible Raman microscopy.

For FT Raman microscopy using Nd:YAG excitation, there is a lack of fluorescence interference from a wide variety of samples, but sensitivity and spatial resolution are compromised in comparison to visible dispersive Raman microscopy.

Red and near infrared semiconductor lasers operating in TEM_{00} single mode can have good beam quality, approaching that of gas lasers, but they are power limited. Obviously, when good beam quality is not necessary, for example, for optically pumping other laser sources, high-power diode lasers can be used in multimode operation. Solid state red and near infrared laser sources are becoming very popular for routine Raman microscopic analysis due to their compactness, robustness, and economical power consumption.

Another reason for the popularity of the red and near infrared diode laser sources for Raman microscopy is that fluorescence is less of a problem due to the lower energy of the light, compared to conventional green 514.5 nm excitation. Although there is a penalty, due to the dependency of the scattering efficiency on the ν^4 term, present day silicon-based CCD arrays can exhibit excellent detection quantum efficiency of Stokes shifted Raman radiation, across the whole vibrational spectrum, when using standard 670 and 785 nm diode lasers and, at least across the fingerprint region, when using 830 and 852 nm diode lasers.

For ultimate spatial resolution in the Raman microscopic experiment, short wavelength excitation is advantageous because it is fundamentally possible to focus to a smaller diffraction-limited laser spot. In this regard, it would be preferable to use UV Raman microscopy where fluorescence is also not as problematic as it is in the visible region. However, in spite of recent improvements in UV laser sources such as the intracavity frequency-doubled argon and krypton ion cw lasers, there are current instrumental limitations on the throughput efficiency of the UV Raman microscope imposed by the lack of availability of suitable notch filters and low stray light and high throughput optics.

It is anticipated that in the future all-solid-state, quasi-cw, diode pumped, mode-locked Ti:sapphire lasers operating on their third or fourth harmonics will become increasingly popular for providing tunable UV excitation affording good beam quality. It is now also possible to fabricate hollow cathode metal ion deep UV lasers, 10–15 cm in length, 2–4 cm in diameter, weighing 50–100 g and requiring only 2–3 W of electrical power. These lasers have low beam quality (M^2 equal to ca. 18), but they make possible the development of portable UV fluorescence imaging and Raman microprobes for geobiological exploration of terrestrial and extraterrestrial environments.

See also

Fourier Optics. **Imaging:** Infrared Imaging. **Introductory Article:** Early Development of the He-Ne Laser (Title TBC). **Lasers:** Noble Gas Ion Lasers; Semiconductor Lasers. **Nonlinear Optics, Applications:** Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Andrews DL (1990) *Lasers in Chemistry* (2nd edition). Berlin, Heidelberg: Springer-Verlag.

Asher SA, Bormett RW, Chen XG, *et al.* (1993) UV resonance Raman spectroscopy using a new cw laser source: Convenience and experimental simplicity. *Applied Spectroscopy* 47: 628–633.

Bell SEJ, Bourguignon ESO, O'Grady A, Villaumie J and Dennis AC (2002) Extracting Raman spectra from highly fluorescent samples with "Scissors" (SSRS, Shifted-Subtracted Raman Spectroscopy). *Spectroscopy Europe* 14: 17.

Best SP, Clark RJH and Withnall R (1992) Non-destructive pigment analysis of artefacts by Raman microscopy. *Endeavour* 16: 66.

Boustany NN, Manoharan R, Dasari RR and Feld MS (2000) Ultraviolet resonance Raman spectroscopy of bulk and microscopic human colon tissue. *Applied Spectroscopy* 54: 24–30.

Delhaye M, Barbillat J, Aubard J, Bridoux M and Da Silva E (1996) Instrumentation. In: Turrell G and Corset J (eds) *Raman Microscopy: Developments and Applications*, chap. 3. London: Academic Press.

Delhaye M and Dhamelincourt P (1975) Raman microprobe and microscope with laser excitation. *Journal of Raman Spectroscopy* 3: 33.

Derbyshire A and Withnall R (1999) Pigment analysis of portrait miniatures using Raman microscopy. *Journal of Raman Sprectroscopy* 30: 185.

Gordeyev SA, Nikolaeva GYu, Prokhorov KA, Withnall R, Dunkin IR and Shilton SJ (2001) Super-selective polysulfone hollow fiber membranes for gas separation: assessment of molecular orientation by Raman spectroscopy. *Laser Physics* 11: 82–85.

Hayward CL, Best SP, Clark RJH, Ross NL and Withnall R (1994) Polarised single crystal Raman spectroscopy of sinhalite, $MgAlBO_4$. *Spectrochimica Acta* 50A: 1287–1294.

Hendra P, Jones C and Warnes G (1991) *Fourier Transform Raman Spectroscopy*. London: Ellis Horwood.

Hitz B (2003) Solid state lasers are gunning for argon ion's place. *Photonics Spectra* 54–58.

Holtz JSW, Bormett RW, Chi Z, *et al.* (1996) Applications of a new 206.5 nm continuous wave laser source: UV Raman determination of protein secondary structure and CVD diamond material properties. *Applied Spectroscopy* 50: 1459–1468.

http://www.btools.com

Koechner W and Bass M (2003) *Solid-State Lasers*. New York: Springer-Verlag.

Landsberg G and Mandelstam L (1928) Eine neue Erschienung bei der Lichtzerstreaung in Krystallen. *Naturwiss* 16: 557.

Matousek P, Towrie M, Stanley A and Parker AW (1999) Efficient rejection of fluorescence from Raman spectra using picosecond Kerr gating. *Applied Spectroscopy* 53: 1485.

Nakashima S, Okumura H, Yamamoto T and Shimidzu R (2004) Deep-ultraviolet Raman microspectroscopy: Characterization of wide-gap semiconductors. *Applied Spectroscopy* 58: 224–229.

Pajcini V, Munro CH, Bormett RW, Witkowski RE and Asher SA (1997) UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51: 81–86.

Pallister DM and Morris MD (1993) In: Morris MD (ed.) *Microscopic and Spectroscopic Imaging of the Chemical State*, chap. 1. New York: Marcel Dekker.

Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy,* 1. Chichester, UK: Wiley.

Pan M-W, Benner RE, Johnson CW and Smith LM (2000) Near-IR lasers rejuvenate Raman spectroscopy. *Optoelectronics World* S5–S10.

Raman CV and Krishnan KS (1928) A new type of secondary radiation. *Nature* 121: 501.

Rosasco GJ (1980) Raman microprobe spectroscopy. In: Clark RJH and Hester RE (eds) *Advances in Infrared and Raman Spectroscopy*, vol. 7. Chichester: Wiley.

Rosasco J, Etz ES and Cassatt WA (1975) The analysis of discrete fine particles by Raman spectroscopy. *Applied Spectroscopy* 29: 396.

Schawlow AL and Townes CH (1958) Infrared and optical masers. *Physics Review* 112: 1940.

Smekal A (1923) Sur Quantentheorie der Dispersion. *Naturwiss* 11: 873.

Storrie-Lombardi MC, Hug WF, McDonald GD, Tsapin AI and Nealson KH (2001) Hollow cathode ion lasers for deep ultraviolet Raman spectroscopy and fluorescence imaging. *Review of Scientific Instruments* 72: 4452–4459.

Turrell G, Delhaye M and Dhamelincourt P (1996) Characteristics of Raman microscopy. In: Turrell G and Corset J (eds) *Raman Microscopy: Development and Applications*, 2. London: Academic Press.

Young M (1977) *Optics and Laser: An Engineering Approach*. Berlin, Heidelberg: Springer-Verlag.

Yu Y-M, Nam S, Byungsung O, *et al.* (2002) Resonant Raman scattering in ZnS epilayers. *Materials Chemistry and Physics* 78: 149–153.

Second-Harmonic Spectroscopy

J I Dadap and T F Heinz, Columbia University, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The demonstration of second-harmonic generation (SHG), one of the early experiments with pulsed lasers, is considered to have marked the birth of the field of nonlinear optics. The importance of second-harmonic generation, and the related phenomena of sum- and difference-frequency generation, as methods for producing new frequencies of coherent radiation was recognized immediately and has in no way diminished over the years. Associated with these applications came an interest in the fundamental nature of the second-order nonlinear response, and some spectroscopic studies were undertaken with this motivation. These investigations were of rather limited scope and were not developed as a general probe of materials. In contrast, the third-order nonlinear optical interactions have given rise to a panoply of significant spectroscopic measurement techniques, including, to name a few, coherent Raman spectroscopy, pump-probe and other four-wave mixing measurements, two-photon absorption and hole burning measurements, photon echoes, and Doppler-free spectroscopy.

A central factor in the lack of spectroscopic applications of the second-order nonlinear response lies in a fundamental symmetry constraint. Unlike the third-order nonlinearity, which is present in all materials, only materials lacking a center of inversion give rise to an (electric-dipole) allowed second-order nonlinear response. As we demonstrate in this article, however, it is this property that makes SHG, as well as the sum-frequency generation process discussed elsewhere in this encyclopedia, such a remarkable probe of surfaces and interfaces. For centrosymmetric materials, the second-order nonlinear response is strong only at interfaces, where only a monolayer or so of material contributes to the response. This yields a purely optical probe (with incident and emitted photons) that gives an inherent surface sensitivity comparable to the best electron spectroscopies, such as electron diffraction, Auger-electron spectroscopy, photoemission spectroscopy, and scanning tunneling microscopy, the workhorses of surface science. Neither linear nor third-order nonlinear optical measurements provide this inherent interface sensitivity. The possibility of probing interfaces by purely optical means permits one to take advantage of several important features of optical radiation. Perhaps most noteworthy is the possibility of probing buried interfaces of every sort, from solid/solid to liquid/liquid, as well as the more conventional cases of solid/vapor, liquid/vapor, and liquid/solid. In addition, laser-based techniques offer unequalled capabilities for spectral or temporal resolution.

In this article, we present a brief overview of SH spectroscopy for the study of surfaces and interfaces. The article is divided into two principal sections: an introductory section that presents some of the fundamentals of SHG and the behavior at interfaces; and a discussion of the different types of information that the method can yield, as illustrated by some representative examples. A brief section on the experimental technique links these two parts.

Fundamentals of Surface SHG

Basic Notions

Within the domain of classical optics, the response of a material to electromagnetic radiation is described by an induced polarization **P** that varies linearly with the electric field strength, **E**. This situation reflects the fact that the strength of electric fields for light encountered under conventional conditions is minute compared to that of the electric fields binding atoms and solids together. The electric field binding an atom may be estimated as $E_a \sim 1\ \text{V/Å} = 10^{10}\ \text{V/m}$. To reproduce this field strength by an electromagnetic wave requires an irradiance of $I_a \sim 10^{17}\ \text{W/m}^2$, so the validity of this linear approximation in classical optics is clear. The advent of the laser, with its capability for producing light beams with very high optical power and irradiance, has permitted us to probe and exploit the nonlinear response of materials, giving rise to the field of nonlinear optics.

The next-order material response beyond the linear approximation consists of an induced nonlinear source polarization $\mathbf{P}^{(2)} = \varepsilon_0 \boldsymbol{\chi}^{(2)} : \mathbf{EE}$ that depends quadratically on the driving electric field $\mathbf{E}$, described by a second-order nonlinear susceptibility $\boldsymbol{\chi}^{(2)}$. For excitation by a single laser at a frequency ω, this response gives rise to SHG at a frequency of 2ω, as well as optical rectification with a dc induced-polarization. For beams at two distinct frequencies of ω_1 and ω_2, the processes of sum- or difference-frequency generation (SFG or DFG) become operative with the production of the new frequencies of $|\omega_1 \pm \omega_2|$. For a relationship of the given form, which assumes only a spatially local quadratic dependence of the polarization on the driving electric field, one can show that $\boldsymbol{\chi}^{(2)}$ must vanish within any centrosymmetric medium. This is a direct consequence of the fact that $\boldsymbol{\chi}^{(2)}$ as a quantity describing a centrosymmetric medium should not change under an inversion operation, while the quantities $\mathbf{P}^{(2)}$ and $\mathbf{E}$ as polar vectors will change sign. The interface specificity of the SHG technique for centrosymmetric media follows from this general observation: only at interfaces is the bulk symmetry broken and the SHG process allowed.

Anharmonic Oscillator Model

Before we embark on our discussion of surface SHG, we first briefly consider a simple classical model of SHG based on an anharmonic oscillator. Although this is not a realistic treatment of the response of materials, it helps to illustrate the characteristic features of the surface SHG process and its spectroscopic attributes. In the anharmonic oscillator model, the medium is treated as a collection of electrons bound to fixed ion cores. While the usual Lorentz model only includes the effect of a linear restoring force, to describe second-order response, we generalize this description to include a leading-order anharmonic restoring force. For simplicity, we consider motion of the anharmonic oscillator in one spatial dimension, in which case the classical equation of motion for the displacement, x, of the electron can then be written as

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + 2\gamma \frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x - bx^2 = -\frac{e}{m} E(t) \qquad [1]$$

Here $-e$ and m denote, respectively, the electron charge and mass, and $E(t)$ is the optical driving field. The (angular) frequency ω_0 defines the resonance of the harmonic component of the response, and γ represents a phenomenological damping rate for the oscillator. The nonlinear restoring force corresponds to the term containing the parameter b, a material-dependent constant. We are interested in the solution of this equation for a monochromatic driving field at frequency ω with amplitude $E(\omega)$ which we write as

$$E(t) = E(\omega) \exp(-\mathrm{i}\omega t) + \mathrm{c.c.} \qquad [2]$$

An exact solution to eqn [1] in the presence of this electric field is not possible. Let us, however, consider a regime in which we do not excite the oscillator too strongly, and the effect of the anharmonic term remains relatively small compared to the harmonic one. In this case, we may solve the problem perturbatively in powers of the driving field. We find a linear response at the fundamental frequency and a nonlinear response at the SH frequency, which we can write in terms of the corresponding induced dipole moments $p = -ex$:

$$p(t) = p^{(1)}(\omega) \exp(-\mathrm{i}\omega t) + p^{(2)}(2\omega) \exp(-\mathrm{i}2\omega t) + \mathrm{c.c.} \qquad [3]$$

We can then write the amplitudes for the harmonically varying dipole moments as

$$p^{(1)}(\omega) = \alpha^{(1)}(\omega) E(\omega) \qquad [4]$$

$$p^{(2)}(2\omega) = \alpha^{(2)}(2\omega = \omega + \omega) E(\omega) E(\omega) \qquad [5]$$

where $\alpha^{(1)}$, the linear polarizability, and $\alpha^{(2)}$, the second-order polarizability (or first hyperpolarizability) for SHG are given, respectively, by

$$\alpha^{(1)}(\omega) = \frac{e^2}{m} \frac{1}{D(\omega)} \qquad [6]$$

$$\alpha^{(2)}(2\omega = \omega + \omega) = \frac{(e^3/m^2) b}{D(2\omega) D^2(\omega)} \qquad [7]$$

Here we have introduced the resonant response at the fundamental frequency by $D(\omega) \equiv \omega_0^2 - \omega^2 - 2\mathrm{i}\gamma\omega$, with a corresponding relation for the response $D(2\omega)$ at the SH frequency. A similar derivation can, of course, be applied to obtain the corresponding material response for the SFG or DFG processes.

The spectral dependence of the polarizabilities $\alpha^{(1)}$ and $\alpha^{(2)}$ is illustrated in Figure 1 for the anharmonic oscillator model with a resonance frequency of ω_0. In the linear case, one sees the expected behavior of the loss, associated with the $\mathrm{Im}[\alpha^{(1)}]$, and dispersion associated with $\mathrm{Re}[\alpha^{(1)}]$. A single resonance in the linear response is seen for the driving frequency ω near the oscillator frequency ω_0. For the case of SHG, on the other hand, we observe both resonances when the fundamental frequency $\omega \approx \omega_0$ and when the SH

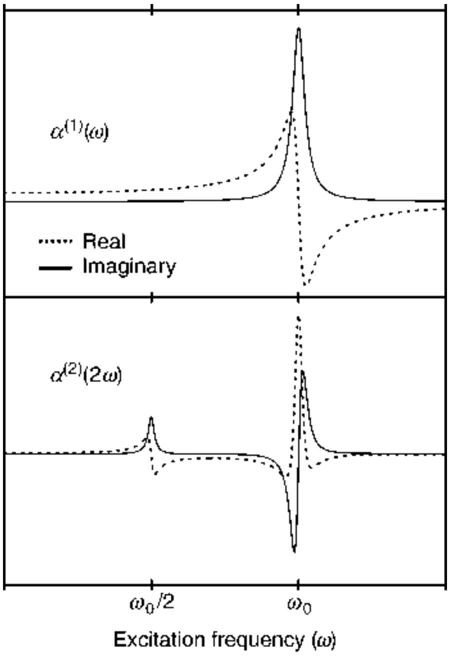

Figure 1 Real and imaginary components of the linear polarizability $\alpha^{(1)}(\omega)$ (top panel) and the second-order nonlinear polarizability $\alpha^{(2)}(2\omega = \omega + \omega)$ (bottom panel) plotted versus frequency according for the anharmonic oscillator model. For the SH response, the resonance features at $\omega = \omega_0$ and $\omega = \omega_0/2$ correspond, respectively, to single-photon and two-photon resonances.

frequency $2\omega \approx \omega_0$. In quantum mechanical language introduced below, these peaks can be considered as arising, respectively, from matching of the fundamental and second-harmonic photon energies with the transition energy in the material.

In SHG spectroscopy, one normally measures the SH power as a function of the laser frequency. The SH field will scale with the nonlinear material response, i.e., with $\alpha^{(2)}(2\omega = \omega + \omega)$, so that the SH power will be proportional to $|\alpha^{(2)}(2\omega = \omega + \omega)|^2$. Figure 2 shows the corresponding spectra in the vicinity of the second-harmonic resonance. An important aspect of this type of spectroscopy, which is also present in other nonlinear spectroscopies, such as coherent Raman measurements, is the role that may be played by a nonresonant background. This may arise either from the off-resonant response of the system under study or from a coherent background from another material. In either case, the effect can be described by adding a spectral flat background response to the nonlinearity. Now when we detect the optical power, we then measure a quantity proportional to $|\alpha^2(2\omega = \omega + \omega) + k|^2$, where k represents the nonresonant background. The presence of k obviously elevates the baseline response near a resonance. In addition, however, depending on the relative phase of k, this extra term can introduce significant changes in the observed lineshapes. Figure 2 illustrates this effect, which must always be borne in mind in the interpretation of experimental spectra.

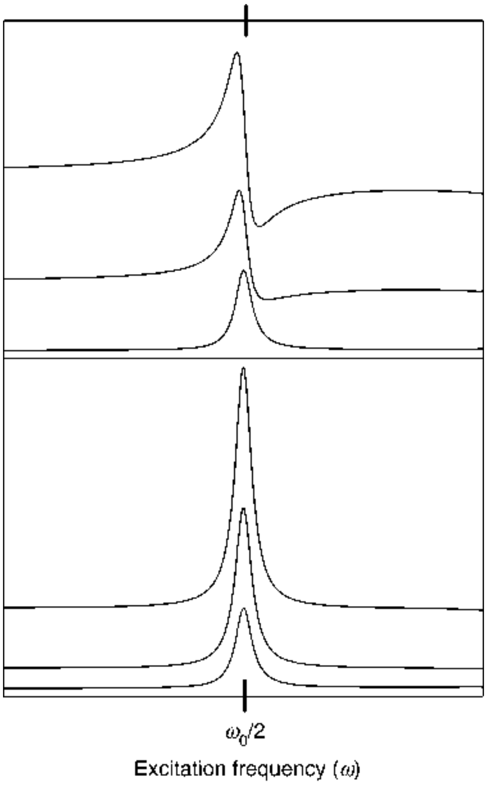

Figure 2 SH spectra $I(2\omega)$ versus pump laser frequency ω corresponding to **Figure 1** in the vicinity of the two-photon resonance, with and without the presence of a nonresonant (frequency-independent) background: $I(2\omega) \sim |\alpha^{(2)}(2\omega = \omega + \omega) + k|^2$, where k represents a constant background and parametrized through $k = m|\alpha^{(2)}(2\omega = \omega_0)|$ and m is constant factor. Top panel: $m = 0, 0.5, 1$ (bottom to top curves). Bottom panel: $m = 0, 0.5i, i$ (bottom to top curves). In the top panel, the upper two lineshapes were displaced vertically to distinguish the three curves.

For the purposes of studies of surfaces and interfaces, the key property of second-order nonlinear processes is the fact that they exhibit, for centrosymmetric bulk media, an inherent interface specificity. This feature can be readily understood in the context of the nonlinear oscillator model in which the SH polarization scales with the material parameter b. Since this term is associated with a potential that varies as x^3, it is clear that the value of the b parameter can be taken as a measure of the material's departure from inversion symmetry. At the interface, such a term can be present, reflecting the inherent asymmetry at a boundary, while it must necessarily be absent in the bulk of a centrosymmetric material.

Quantum-Mechanical Description of SHG

A correct description of SHG naturally requires a quantum mechanical treatment of the material response to the optical field. For our present purposes, we consider a localized entity with a nonlinear response, such as a non-centrosymmetric molecule at a surface. We consider below how these individual units can be added together to yield the surface nonlinear response of the material. Within this picture, we write the induced nonlinear dipole moment $\mathbf{p}^{(2)}$ of each molecule in terms of the driving electric field $\mathbf{E}$ as

$$\mathbf{p}^{(2)}(2\omega) = \boldsymbol{\alpha}^{(2)}(2\omega = \omega + \omega) : \mathbf{E}(\omega)\mathbf{E}(\omega) \qquad [8]$$

This relation is simply the generalization of eqn [5] to include a full 3-dimensional description, where $\mathbf{p}^{(2)}$ and $\mathbf{E}$ are vectors and the second-order nonlinear polarizability $\boldsymbol{\alpha}^{(2)}$ is a third-rank tensor. The quantum mechanical description enters in how we relate the response function $\boldsymbol{\alpha}^{(2)}(2\omega = \omega + \omega)$ to the underlying properties of the material.

We can obtain an expression for the second-order nonlinear polarizability by application of second-order perturbation theory with the light–matter interaction treated as the perturbation. Within the electric-dipole approximation, the interaction Hamiltonian can be taken as $H_{\text{int}} = -\boldsymbol{\mu}\cdot\mathbf{E}(t)$, where $\boldsymbol{\mu} = -e\mathbf{r}$ denotes the electric-dipole operator and $\mathbf{E}(t)$ the electric field of the driving laser beam. Using the standard density-matrix formalism for second-order perturbation theory, one calculates for the Cartesian components $\alpha^{(2)}_{ijk}$ of the tensor $\boldsymbol{\alpha}^{(2)}$ as

$$\alpha^{(2)}_{ijk}(2\omega = \omega + \omega) = -\frac{1}{\hbar^2}\sum_{g,n,n'}\left[\frac{(\mu_i)_{gn}(\mu_j)_{nn'}(\mu_k)_{n'g}}{(2\omega - \omega_{ng} + \mathrm{i}\Gamma_{ng})(\omega - \omega_{n'g} + \mathrm{i}\Gamma_{n'g})} + \cdots\right]\rho^{(0)}_g \qquad [9]$$

In this expression, the letters g, n', and n represent energy eigenstates $|g\rangle$, $|n'\rangle$, and $|n\rangle$ of the system, with $\rho^{(0)}_g$ corresponding to the thermal population for differing available ground states g. As illustrated in Figure 3, the SHG process can be regarded as involving a series of three transitions: two transitions associated with the absorption of two pump photons, each of energy $\hbar\omega$, and a transition associated with emission of a second-harmonic photon of energy $2\hbar\omega$. These transitions occur through the electric-dipole operator $\boldsymbol{\mu}$ and are characterized by the matrix elements $(\mu_i)_{gn}$. In eqn [9], the energy denominators involve the energy differences $\hbar\omega_{ng} \equiv E_n - E_g$ and widths $\hbar\Gamma_{ng}$ for transitions between eigenstates $|n\rangle$ and $|g\rangle$, and similarly for other combinations of states. In addition to the term indicated explicitly in eqn [9], there are seven additional terms with different Cartesian coordinates in the matrix elements and/or frequency denominators.

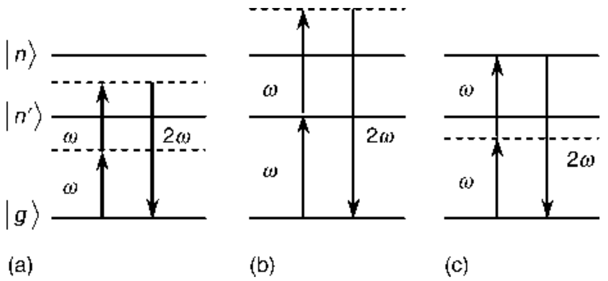

Figure 3 Quantum mechanical description of SHG process for the situations of (a) nonresonant, (b) single-photon resonant, and (c) two-photon resonant interactions. SHG occurs through transition from the initial state electronic state $|g\rangle$ and a first intermediate state $|n'\rangle$, then to a second intermediate state $|n\rangle$, followed by a return to the initial state. The process is just a parametric conversion of two photons at frequency ω into one photon at frequency 2ω.

The frequency denominators in the eight terms of eqn [9] introduce a resonant enhancement in the nonlinearity when either the fundamental frequency ω or the SH frequency 2ω coincides with a transition from a ground state $|g\rangle$ to one of the intermediate states $|n'\rangle$ or $|n\rangle$. Figure 3a shows the situation for a nonresonant nonlinear response, while Figures 3b,c illustrate, respectively, single- and two-photon resonances. The numerator in the perturbation theory expression consists of products of the three dipole matrix elements of the form $(\mu_i)_{gn}(\mu_j)_{nn'}(\mu_k)_{n'g}$. These terms reflect the structure and symmetry of the material that is built into the third-rank tensor $\alpha^{(2)}_{ijk}(2\omega = \omega + \omega)$.

The Surface Nonlinear Response

As a model of the surface nonlinearity, let us consider a monolayer of oriented molecules. The surface nonlinear response accessible to a macroscopic measurement is given, under the neglect of local-field effects, simply by summing the response of the individual molecules. The resulting surface nonlinear susceptibility tensor $\chi^{(2)}_{\mathrm{s},ijk}$ connecting the induced nonlinear sheet polarization to the driving electric field can then be expressed as

$$\chi^{(2)}_{\mathrm{s},ijk} = \frac{N_{\mathrm{s}}}{\varepsilon_0}\langle T_{i\lambda}T_{j\mu}T_{k\nu}\rangle\alpha^{(2)}_{\lambda\mu\nu} \qquad [10]$$

Here N_s denotes the adsorbate surface density, $\alpha^{(2)}_{\lambda\mu\nu}$ the nonlinear polarizability of the molecule expressed in its own coordinate system, $T_{i\lambda}$ the transformation tensor from the molecule's coordinate system to the laboratory frame, $\langle\,\rangle$ an ensemble average over the orientation of the different molecules in the monolayer, and $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m the permittivity of free space. This expression for the surface nonlinear susceptibility has been successful in describing properties of adsorbed molecules, including their coverage, orientation, and spectroscopic features.

For the surfaces of materials such as semiconductors or metals with delocalized electronic states, a somewhat different formulation of the nonlinear response is appropriate. A modified version of eqn [9] can be obtained using the relevant band states. The underlying concepts involved in the surface nonlinear response are, however, similar to those presented for the molecular case.

Regardless of the details of the model describing the surface nonlinear susceptibility, its tensor properties must reflect the symmetry of the interface. In that respect $\chi^{(2)}_{s,ijk}$ is quite analogous to the bulk nonlinear response, $\chi^{(2)}_{ijk}$, in a non-centrosymmetric medium. In the absence of any symmetry constraints, $\boldsymbol{\chi}^{(2)}_s$ will exhibit $3 \times 3 \times 3 = 27$ independent elements for SFG and DFG. For SHG, the order of the last two indices has no significance, and the number of independent elements is reduced to 18. If the surface exhibits a certain in-plane symmetry, then the form of $\boldsymbol{\chi}^{(2)}_s$ will be simplified correspondingly. For the common situation of an isotropic surface, for example, SHG possesses three allowed elements of $\boldsymbol{\chi}^{(2)}_s$ and may be denoted as $\chi^{(2)}_{s,\perp\perp\perp}$, $\chi^{(2)}_{s,\perp\|\|}$, $\chi^{(2)}_{s,\|\perp\|} = \chi^{(2)}_{s,\|\|\perp}$, where $\perp$ corresponds to the direction of the surface normal and $\|$ to an in-plane direction. The nonvanishing elements of $\boldsymbol{\chi}^{(2)}_s$ for other commonly encountered surface symmetries are summarized in several textbooks and reviews listed below.

Radiation Properties for Surface SHG

In order to probe and extract information from interfaces through surface SHG, it is necessary to understand how radiation interacts with the relevant media and gives rise to the experimentally observable signals. Here we present the principal results for the usual case of a spatially homogeneous planar interface. Nonplanar and inhomogeneous geometries, which are also of considerable interest, will be considered briefly under the heading of applications of the SHG technique below.

The SHG process is coherent in nature. Thus, for the planar geometry, a pump beam impinging on the interface will give rise to a well-collimated SH radiation emerging in distinct reflected and transmitted directions. To describe this situation in a more detailed fashion and provide formulas for the radiation efficiency, we introduce a general description of SHG by a planar interface excited by a plane wave, as shown in Figure 4. The nonlinear response of the interface described by the surface nonlinear susceptibility tensor $\boldsymbol{\chi}^{(2)}_s$ is incorporated through a nonlinear source polarization $\mathbf{P}^{(2)}$ of

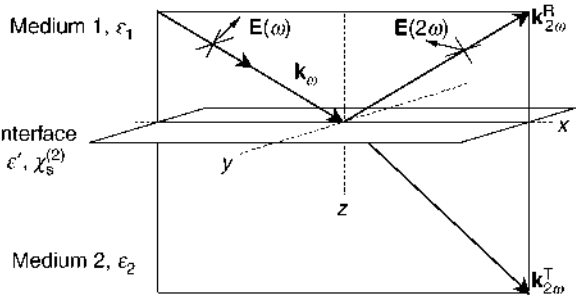

Figure 4 Schematic representation of SHG at the interface between two centrosymmetric media. The pump field $\mathbf{E}(\omega)$ at frequency ω with corresponding wavevector $\mathbf{k}_\omega$ is incident on the interface from medium 1. The SH field $\mathbf{E}(2\omega)$ at frequency 2ω is emitted in directions described by the wavevectors $\mathbf{k}^R_{2\omega}$ (reflected in medium 1) and $\mathbf{k}^T_{2\omega}$ (transmitted in medium 2). The linear dielectric constants of media 1, 2, and the interface are denoted, respectively, by ε_1, ε_2, and ε'. The nonlinear response of the interface ($z = 0$) is given by the surface nonlinear susceptibility tensor $\boldsymbol{\chi}^{(2)}_s$.

$$\begin{aligned}\mathbf{P}^{(2)}(2\omega) &= \mathbf{P}^{(2)}_s(2\omega)\delta(z)\\ &= \varepsilon_0\boldsymbol{\chi}^{(2)}_s(2\omega = \omega + \omega) : \mathrm{E}(\omega)\mathrm{E}(\omega)\delta(z) \qquad [11]\end{aligned}$$

localized at the interface ($z = 0$). In addition to the strong nonlinear response at the interface, one must also generally consider the nonlocal nonlinear response of the bulk media. This response is much weaker than the symmetry-allowed interfacial response, but is permitted even in centrosymmetric materials. Since it is present in a much larger volume, however, its cumulative effect, while generally weaker than the surface response, may be of comparable magnitude. For simplicity, however, we neglect this constant background response in our discussion in this article, but a complete discussion can be found in the referenced works. The overall SH response is naturally also influenced by the linear optical properties of the surrounding media. The two bulk media and the interfacial region are characterized, respectively, by frequency-dependent dielectric functions ε_1, ε_2, and ε', as shown in Figure 4. Again for simplicity, and in accordance with most applications, we take the linear response to be isotropic.

Figure 4 shows the incoming pump radiation and the reflected and transmitted SH beams. The directions of these beams are determined by conservation of the component of the momentum parallel to the interface in the SHG process, i.e., $2\mathbf{k}_{\omega,\|} = \mathbf{k}^{\mathrm{R}}_{2\omega,\|} = \mathbf{k}^{\mathrm{T}}_{2\omega,\|}$. Thus the reflected and transmitted beams both remain in the plane of incidence, with directions governed by the so-called nonlinear Snell's law. For bulk media without dispersion between the frequencies of ω and 2ω, the reflected and transmitted beams simply maintain the same angle of incidence as the pump beam. However, for dispersive bulk media, the directions of the beams are, however, altered, as can be understood immediately from the expression $2k_{\omega}/k_{2\omega} = n_{\omega}/n_{2\omega}$, where n denotes the refractive index of the medium. The surface SHG process does, nonetheless, differ in an important respect. In contrast to the situation for bulk media, the interfacial region is of negligible thickness, so phase shifts can obviously be disregarded. The issue of phase matching, which is of great significance for SHG from bulk media, consequently does not enter into the surface SHG problem.

By application of the Maxwell equations with the indicated nonlinear source polarization, one can derive explicit expressions for the SH radiation in terms of the linear and nonlinear response of the materials and the excitation conditions. For the case of the nonlinear reflection, the irradiance for the SH radiation can be written as

$$I(2\omega) = \frac{\omega^2 \sec^2\theta \left|\mathbf{e}'(2\omega)\cdot\boldsymbol{\chi}_{\mathrm{s}}^{(2)} : \mathbf{e}'(\omega)\mathbf{e}'(\omega)\right|^2 [I(\omega)]^2}{2\varepsilon_0 c^3 [\varepsilon_1(2\omega)]^{1/2} \varepsilon_1(\omega)} \qquad [12]$$

where $I(\omega)$ denotes the irradiance of the pump beam at the fundamental frequency; c is the speed of light in vacuum; θ is the angle of emission of the SH beam with respect to the surface normal. The vectors $\mathbf{e}'(\omega)$ and $\mathbf{e}'(2\omega)$ represent the polarization vectors $\hat{\mathbf{e}}_1(\omega)$, and $\hat{\mathbf{e}}_1(2\omega)$, respectively, after they have been adjusted to account for the linear propagation of the waves to the interface. More specifically, we may write $\mathbf{e}'(\omega) = \mathbf{F}_{1\to 2}\hat{\mathbf{e}}_1(\omega)$. Here the $\mathbf{F}_{1\to 2}$ describes the relationship between the electric field $E\hat{\mathbf{e}}_1$ in medium 1 (propagating towards medium 2), which yields a field $E\mathbf{e}'$ at the interface. For light incident in the x–z plane as shown in Figure 4, $\mathbf{F}_{1\to 2}$ is a diagonal matrix whose elements are $F^{xx}_{1\to 2} = 2\varepsilon_1 k_{2,z}/(\varepsilon_2 k_{1,z} + \varepsilon_1 k_{2,z})$, $F^{yy}_{1\to 2} = 2k_{1,z}/(k_{1,z} + k_{2,z})$, and $F^{zz}_{1\to 2} = 2(\varepsilon_1\varepsilon_2/\varepsilon')k_{1,z}/(\varepsilon_2 k_{1,z} + \varepsilon_1 k_{2,z})$, where the quantity $k_{i,z}$ denotes the magnitude of the z-component of the wavevector in medium i at the relevant wavelength (ω or 2ω) (Figure 4).

A few observations may be made about eqn [12], which describes the relationship between the typical experimental observables of the pump irradiance, $I(\omega)$, and the irradiance of the emitted SH reflection, $I(2\omega)$, from the interface. First, one can immediately recognize the quadratic relationship between the pump irradiance and the SH irradiance that is expected for a second-order nonlinear optical effect. Second, the SH radiation is determined not only by the nonlinear optical response, but by the linear optical properties of the relevant media. Consequently, obtaining the surface nonlinear susceptibility tensor $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ from experiment requires knowledge or measurement of the linear response. (It should be noted, however, that the interfacial linear properties always enter into the response in a fixed manner and can be incorporated into an effective surface nonlinear response.) The linear response can significantly enhance or reduce the efficiency of the SHG process. The use of a total-internal reflection geometry, for example, provides a significant enhancement in the SH radiation through the linear optical response of the media. Third, while full knowledge of the linear and nonlinear response will obviously allow one to predict the strength of the SH radiation for any set of conditions, a simple measurement of the SH response does not permit one to determine $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$, even if the linear optical properties are known. The determination of the tensor $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ requires probing the material with different polarization states and/or angles of incidence. Protocols for making these measurements have been discussed extensively in the literature. For the general case of a material with in-plane anisotropy, such as the surface of a crystal, measurements of the response as the crystal is rotated about its surface normal are also very useful in a complete determination of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$. Fourth, the simple measurement of the SH intensity as assumed here obviously eliminates information about the overall phase of the $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$. This phase information is often of considerable interest and importance. For example, if the response of the interface is dominated by an oriented molecular monolayer, the sign of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ will be inverted if the monolayer is reflected through the plane of the surface, i.e., the phase of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ yields information on the polar ordering of the layer. To obtain such phase information, one must employ a scheme using interference with a known SH reference. This approach, which is sketched briefly in the next section, is implemented quite commonly in experiment.

Experimental Considerations

The merits of the SHG technique for selective probing of interfaces are clear. On the other hand, one needs to be able to detect the optical signal associated with a second-order nonlinear response of a material that arises from an effective thickness of just a monolayer or so. Consequently, it is important to consider various factors that aid in optimizing the strength and detectability of the SH radiation. To discuss how best to achieve this end, let us recast eqn [12] in terms of more convenient experimental parameters. We express the SH signal, S, in units of photons/s and describe the laser excitation by its pulse energy E_p, pulse width τ, and repetition rate R. For irradiation of an area A of the sample, we then obtain

$$S = \frac{\omega \sec^2 \theta E_\mathrm{p}^2 R |\chi_\mathrm{eff}^{(2)}|^2}{4\hbar \varepsilon_0 c^3 \tau A} \qquad [13]$$

where $|\chi_\mathrm{eff}^{(2)}| \equiv |\mathbf{e}'(2\omega)\cdot\boldsymbol{\chi}_\mathrm{s}^{(2)}:\mathbf{e}'(\omega)\mathbf{e}'(\omega)|/[\varepsilon_1^{1/2}(2\omega)\,\varepsilon_1(\omega)]^{\frac{1}{2}}$, whose typical value is of the order of 10^{-21} m^2/V. From this relation, we see that increasing either the pulse energy or repetition rate of the laser will enhance our SH count rate. These parameters are, however, often limited by the characteristics of the available laser, as well as possible damage of the sample. For a given pulse energy and repetition rate, the SH signal is also seen to increase as the laser pulse duration or sample area is decreased. In practice, adequate SH signals can generally be obtained from high-power Q-switched lasers (with nanosecond pulses) with a fairly large focal spot. A preferred solution in most cases is, however, provided by modelocked lasers with high repetition rate and pulses of sub-picosecond duration. Surface SH signals obtained, for example, using a femtosecond mode-locked Ti:sapphire oscillator, are typically in the range of 10^2–10^6 photons/s. Although these signals are relatively weak, they are easily detectable with photomultiplier tubes (PMT) and appropriate electronics.

The basic arrangement for surface SHG measurements is shown in Figure 5. It makes use of the directionality and frequency selectivity of the SHG process not only to detect the desired signal with high efficiency, but to discriminate strongly against background signals, including the reflected laser beam. In the setup, irradiation of the sample involves a pulsed laser source, LS; a halfwave plate, λ, and polarizer, P, for controlling the input polarization. Spectral control is achieved with a filter F1 that passes the fundamental radiation in the pump beam, but rejects other wavelengths, particularly spurious SH radiation arising from other optical components; and a color filter F2 that passes SH, but rejects the reflected fundamental radiation. In addition, a spectrometer is commonly employed to limit the detection bandwidth further. An analyzer A serves to select the desired SH polarization. The SH signal is then measured by a high-quantum efficiency detector and the signal is monitored by processing electronics, typically a photon counter, a gated integrator, or a lock-in amplifier, depending on the strength of the signal and the repetition rate of the laser. The laser source may be either of fixed wavelength or tunable for spectroscopic measurements.

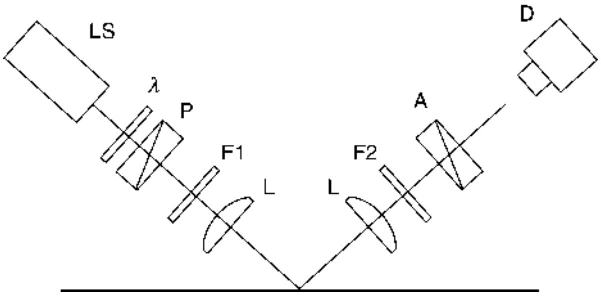

Figure 5 Schematic representation of a basic surface experimental setup for surface SHG. It is comprised of laser source (LS), halfwave plate (λ), polarizer (P), lenses (L), spectral filters (F1, F2), analyzer (A), and detector (D).

There are other important experimental configurations for surface SHG that extend beyond the basic SH reflection geometry, shown schematically in Figure 6a. A common improvement (Figure 6b) is the implementation of a SH reference in which a small fraction of the pump radiation is directed into a crystal with a bulk second-order nonlinear response. This approach provides a reference against which to compare the measured signal, thus compensating appropriately for pulse-to-pulse variation in duration,

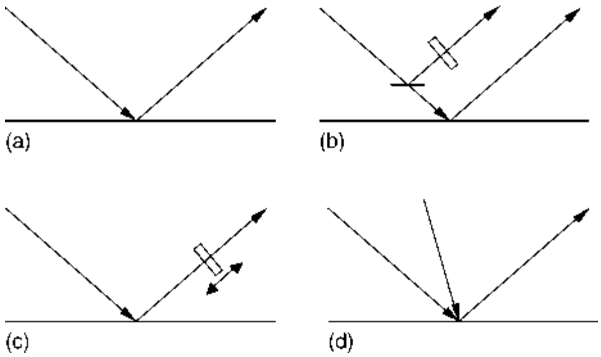

Figure 6 Various schematic representations of different experimental schemes for surface SH measurements. Panel (a) represents the basic scheme as in Figure 5. Panel (b) describes use of a separate SH source as a reference for normalizing the surface SH signals against laser fluctuation and drift. Panel (c) represents the arrangement for a measurement of the phase of the surface SH signal by mixing the fields from the surface and an external source SH radiation. Panel (d) describes the configuration for a time-resolved measurement using the pump-probe scheme.

energy, and spatial profile, as well as for long-term drifts. This type of normalization (against a material with a suitably flat spectral response) is essential for spectroscopy measurements in which the laser frequency is scanned. In addition, use of a calibrated reference is the best way of determining the surface nonlinear response in absolute terms.

It may also be desirable to measure the phase of the signal. For this type of an experiment, a reference field, $\mathbf{E}_{\mathrm{ref}}(2\omega)$, of similar amplitude and known phase response is generated along the path of the surface SH signal $\mathbf{E}(2\omega)$ (Figure 6c). The two collinear SH beams of the same polarization interfere at the detector. The SH power scales as $|\mathbf{E}_{\mathrm{ref}}(2\omega)+\mathbf{E}(2\omega)|^2 = |\mathbf{E}_{\mathrm{ref}}|^2 + 2E_{\mathrm{ref}}E\cos\varphi + |\mathbf{E}|^2$, where φ is the relative phase of the two fields. The relative phase φ may be varied by translating a reference plate along the path of the beam and making use of the dispersion of the ambient air or by inserting another control phase-shifting element. A variant of this phase measurement method is the homodyne scheme in which the reference electric field amplitude $\mathbf{E}_{\mathrm{ref}}(2\omega)$ is significantly stronger than that of the signal amplitude $\mathbf{E}(2\omega)$. The measured SH power then scales as $|\mathbf{E}_{\mathrm{ref}}|^2 + 2E_{\mathrm{ref}}E\cos\varphi$, which yields a signal that varies linearly with the desired SH field strength from the sample.

In the SHG technique, a short-pulse laser is used to interrogate the sample and thus the measurement provides high time-resolution. To make use of this capability, we can add a pump pulse to excite the sample away from equilibrium. This provides the opportunity to examine ultrafast surface dynamics induced by excitation of the sample with a suitable pump laser pulse. For such measurements (Figure 6d), the pump pulse is generally derived from the same laser source as the probe for the SHG measurement, although one or both of these pulses may undergo frequency conversion before being applied to the surface. The critical element is that time synchronism is maintained. The dynamics can then be followed by repeating the SHG measurement at various time delays relative to pump excitation, as controlled by an optical delay line. The ultimate time resolution of this technique is limited only by the duration of the laser pulses and may be as short as a few femtoseconds with state-of-the-art modelocked lasers.

Applications of Surface SHG

The technique of surface SHG is extremely flexible in application, since it relies only on basic symmetry properties for its remarkable surface or interface specificity. SHG measurements can provide quite varied information about the interface, depending on the mode of application of the technique and the material system under study. In this section, we attempt to illustrate the principal types of measurements that can be performed with SHG and the information that can be obtained thereby. For systems with molecular adsorbates, SHG studies can yield information on the density and orientation of molecular species, and can typically do so in real time. For crystalline surfaces, the symmetry and order of the surface can be examined through studies of tensor properties of the surface nonlinear susceptibility. Spectroscopic studies of electronic transitions are also of great value. These can be accomplished for both solids and molecular systems through the same experimental approach of tuning the frequency of the laser source. In all of these investigations, one can achieve very high time resolution, down into the femtosecond range, by the use of pump-probe techniques. In addition, as we discuss below, SHG provides interesting additional capabilities for systems with applied electric and magnetic fields. Finally, the method permits one to probe spatially inhomogeneous systems.

In addition to this diverse range of information that can be provided by surface SHG measurements, the technique is unique in the extremely wide range of material and chemical systems to which it can be – and, indeed, has been – applied. These span the gamut from solids under ultrahigh vacuum to solid/solid, solid/liquid, liquid/vapor, and even liquid/liquid interfaces. Since our examples were drawn to illustrate the type of information obtainable, rather than the range of material systems investigated, the reader is unfortunately obliged to refer to other more detailed treatments of surface SHG measurements to appreciate fully this important facet of the technique.

Adsorbate Density

The remarkably high sensitivity of the SHG technique to the nature of the surface is illustrated by the data in Figure 7, which shows the dramatic change in the SH response from a clean Si(111)-7 × 7 surface upon adsorption of atomic hydrogen. Indeed, it is possible to detect well below 1% of a monolayer (ML) of adsorbed hydrogen. A useful point to note is that the change in $\chi_{\mathrm{s}}^{(2)}$ is linear with adsorbate density for modest values of the coverage, θ. Thus, it is quite easy to obtain a relative gauge of the adsorbate coverage in this regime.

The behavior of the SH response to adsorbates at low coverages can often be described by a simple linearized model of the surface nonlinear susceptibility of $\chi_{\mathrm{s}}^{(2)} = (1-a\theta)\chi_{\mathrm{s},1}^{(2)} + \theta\chi_{\mathrm{s},2}^{(2)}$, where

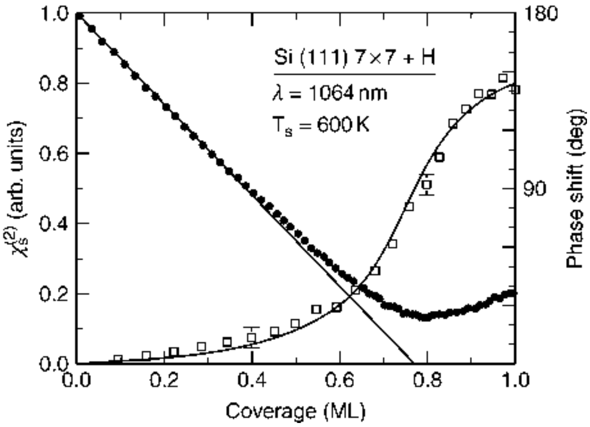

Figure 7 Dependence of the magnitude (full circles) and the phase (open squares) of the nonlinear susceptibility $\chi_s^{(2)}$ for a Si(111)-7 × 7 surface as a function of the coverage of adsorbed atomic hydrogen using a fundamental photon energy 1.17 eV. Reproduced with permission from Höfer U (1996) Nonlinear optical investigations of the dynamics of hydrogen interaction with silicon surfaces. *Applied Physics A* 63: 533.

$\chi_{s,i}^{(2)}$ $(i = 1, 2)$ are complex quantities and a is a constant. This relation describes the surface nonlinear susceptibility as arising from a contribution $\chi_{s,1}^{(2)}$ from the clean surface and a contribution $\chi_{s,2}^{(2)}$ from a saturated monolayer $(\theta = 1)$ of the adsorbate. The term $-a\theta\chi_{s,1}^{(2)}$ then represents the modification of the surface nonlinearity arising from interaction with the adsorbate. For the case of hydrogen adsorption of Figure 7, this is the dominant term, since hydrogen atoms would have no nonlinearity on their own. In other systems, however, the response can be considered as arising essentially from the nonlinear response of aligned adsorbed molecules. The simple linearity of this relation of $\chi_s^{(2)}$ with adsorbate coverage obviously neglects important effects such as adsorbate–adsorbate chemical interactions, as well as local-field effects and the possible influence of differing adsorption sites. All of these factors have been considered in the literature. Nonetheless, the linearity of response with adsorbate coverage is often found to be a good approximation, even for fairly high adsorbate coverages. From the point of view of simply following adsorption dynamics with SHG, it is not, we note, necessary to have any a priori knowledge of the relation between the adsorbate coverage and the SH response. This relation can be established empirically in static measurements and then used advantageously to follow the dynamics of surface processes.

Surface Symmetry and Molecular Orientation

The surface SHG process is described by the third-rank tensor $\boldsymbol{\chi}_s^{(2)}$. The tensorial properties of this response reflect the symmetry of the surface. Other than for a few cases of high symmetry, such a third-rank tensor gives a distinct signature of surfaces with crystalline order compared to surfaces, such as disordered ones, exhibiting effective in-plane isotropy. The form of the $\boldsymbol{\chi}_s^{(2)}$ tensor can be probed both by measurements that vary the polarization of the fundamental and second-harmonic beams and by measurements in which the plane of incidence is changed with respect to the orientation of the surface, generally by rotating the sample about its normal with a fixed experimental geometry.

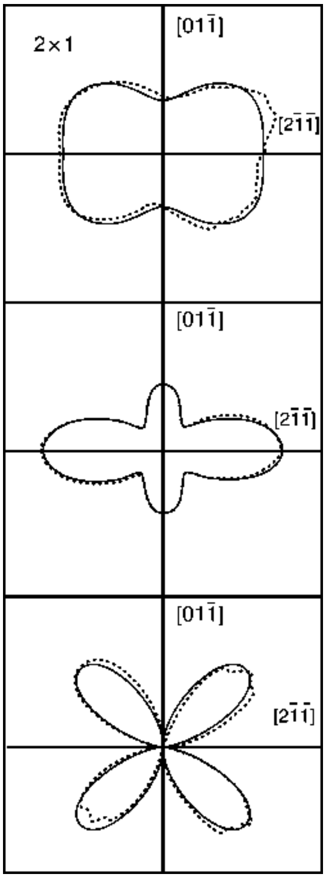

Figure 8 Polarization dependence of SHG for normal incidence excitation of the cleaved Si(111)-2 × 1 surface. The polar plot of the SH signal versus the pump polarization orientation is shown for the case of no polarization analysis (top panel), for the SH analyzer along the [2$\bar{1}\bar{1}$] crystallographic direction of the surface (middle panel), and along the [01$\bar{1}$] direction (bottom panel). The solid curves are fits to theory assuming that the surface has *m* symmetry. Reproduced with permission from Heinz TF, Log MMT and Thomson WA (1985) Study of Si(111) surfaces by optical second-harmonic generation: reconstruction and surface phase-transformation. *Physical Review Letters* 54: 63.

Figure 8 illustrates the capability of surface SHG measurements to probe surface symmetry. The figure shows SHG polarization dependence for a Si(111) surface cleaved in ultrahigh vacuum. While the underlying bulk symmetry suggests that the surface might exhibit 3m symmetry, the analysis of data indicates that only a single mirror *m* plane is present. This lowered symmetry corresponds to the metastable 2×1 reconstruction induced by the crystal cleavage. By heating the sample, one can observe a phase transformation to the equilibrium 7×7 reconstruction and the emergence of the corresponding 3m surface symmetry.

The analysis of surface or interface symmetry by SHG relies only on the form of the nonlinear susceptibility tensor, i.e., which elements are independent and nonvanishing. Within the context of a microscopic model of the surface response, however, the relative values of the tensor elements of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ can also yield valuable information about the nature of the surface. This approach has been developed extensively in the context of the analysis of molecular orientation at surfaces. The basic notion behind this treatment relies on the knowledge of the microscopic molecular nonlinear polarizability $\alpha_{\lambda\mu\nu}^{(2)}$ and the measured surface nonlinear susceptibility $\chi_{\mathrm{s},ijk}^{(2)}$ to infer information about the orientational average that relates these quantities. As discussed above in conjunction with eqn [10], under the neglect of local-field corrections, the relationship between these two quantities involves the ensemble average of the coordinate transformation tensor from the molecular to the laboratory coordinate systems $\langle T_{i\lambda} T_{j\mu} T_{k\nu} \rangle$. Hence a knowledge of the quantities $\alpha_{\lambda\mu\nu}^{(2)}$, N_{s}, and $\chi_{\mathrm{s},ijk}^{(2)}$ allows us to infer information about the orientational distribution of the molecules as described by the quantity $\langle T_{i\lambda} T_{j\mu} T_{k\nu} \rangle$. Even in the absence of knowledge of the magnitudes of the nonlinear polarizabilities and the adsorbate density, one may still obtain useful information for molecules with a simple form of $\alpha_{\lambda\mu\nu}^{(2)}$.

An important example of this analysis of molecular orientation concerns the case of molecules with a dominant nonlinear polarizability along a single axis, i.e., $\boldsymbol{\alpha}^{(2)} = \alpha_{z'z'z'}^{(2)} \hat{\mathbf{z}}'\hat{\mathbf{z}}'\hat{\mathbf{z}}'$. For the usual situation of in-plane isotropy, one can then show that the nonlinear susceptibility elements are given by $\chi_{\mathrm{s},\perp\perp\perp}^{(2)} = N_{\mathrm{s}} \langle \cos^3 \theta \rangle \alpha_{z'z'z'}^{(2)} / \varepsilon_0$ and $\chi_{\mathrm{s},\|\perp\|}^{(2)} = \chi_{\mathrm{s},\|\|\perp}^{(2)} = \frac{1}{2} N_{\mathrm{s}} \langle \cos \theta \times \sin^2 \theta \rangle \; \alpha_{z'z'z'}^{(2)} / \varepsilon_0$, where θ denotes the angle between the surface normal and the z' molecular axis. It should be noted that these orientational moments, reflecting their origin in a third-rank tensor, give a signature of the polar ordering, since a sign change in the elements of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ follows if the molecular orientation is inverted. Also, when only one element of $\boldsymbol{\alpha}^{(2)}$ is present, we can form ratios of the experimentally measurable ratio of tensor elements that depend only on the molecular orientational factors, such as $R \equiv (2\chi_{\mathrm{s},\|\perp\|}^{(2)} + \chi_{\mathrm{s},\perp\perp\perp}^{(2)}) / \chi_{\mathrm{s},\perp\perp\perp}^{(2)} = \langle \cos \theta \rangle / \langle \cos^3 \theta \rangle$. For a narrow distribution of the orientations centered at θ_0, this ratio yields directly the molecular orientation angle through $R \approx \sec^2 \theta_0$.

This type of molecular orientation analysis has been applied to a wide range of material systems. The method, while powerful, relies on several assumptions, including an adequate knowledge of the molecular nonlinear polarizability, the lack of significant local field corrections, the availability of information about the effective linear dielectric constant in the interfacial region, and a suitable method of eliminating any substrate nonlinear response in the measurement. The method is, consequently, best suited for adsorbed layer of molecules with strong and well-defined nonlinear response, present at relatively modest coverages. The use of resonant excitation is a valuable adjunct in these studies, since it generally simplifies the form and strengthens the molecular nonlinear response. An analogous approach to the analysis of molecular orientation can be accomplished with infrared–visible sum-frequency generation. These measurements often provide a superior method of analysis, since the molecular response and its symmetry are better defined by resonant excitation of a molecular vibration.

Surface and Interface Spectroscopy

Surface SHG can be readily adapted to yield spectroscopic information on surface transitions. As indicated above, both resonances at the fundamental and second-harmonic frequencies are manifested in the SHG response. The examination of resonances at the SH frequency is often convenient, since it permits the application of a relatively intense pump radiation at a frequency well separated from that of the resonance. Such SHG spectroscopy measurements have been widely applied to examine electronic transitions at surfaces and interfaces of solids, as well as in adsorbed molecules under a wide variety of conditions.

An example of SHG spectroscopy of electronic transitions at a buried solid/solid interface is presented in Figure 9. For the epitaxial CaF_2/Si(111) interface, distinctive new electronic transitions, associated with interface states, are observed that are absent from the response of either of the bulk materials. These excitations occur in a spectral range where the bulk silicon sample is highly absorbing, which would render the identification of the new interface transitions very difficult using conventional linear optics. The buried nature of the interface naturally also constrains application of conventional surface probing techniques involving electron

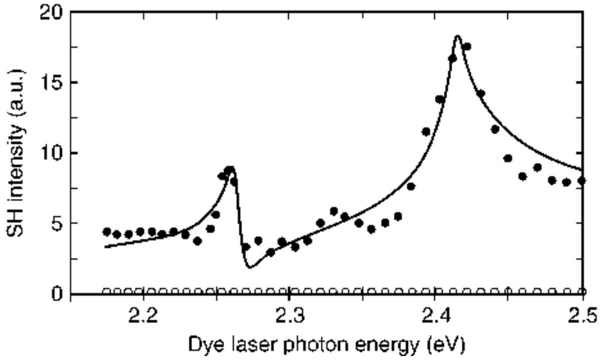

Figure 9 SHG spectrum (solid dots) for the CaF_2/Si(111) interface. For comparison, the open circles represent the corresponding signals for the SiO_2/Si(111) interface, which is seen to give a very weak signal in this frequency range. The solid line is a theoretical fit based on a two-dimensional direct transition of the interface states, together with an excitonic feature slightly below the band gap. Reproduced with permission from Heinz TF, Himpsel FJ, Palange E and Burstein E (1989) Electronic transitions at the CaF_2/Si(111) interface probed by resonant three-wave-mixing spectroscopy. *Physical Review Letters* 63: 644.

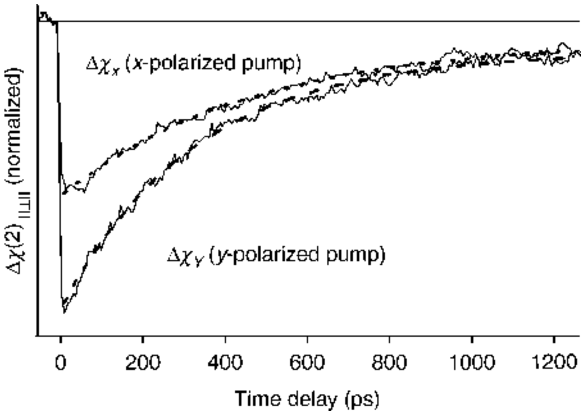

Figure 10 Rotational relaxation of Coumarin 314 molecules at the air–water interface subsequent to photoexcitation. The initial anisotropy in the molecular orientational distribution is created by linearly polarized pump radiation in two orthogonal directions in the surface. Reproduced with permission from Zimdars D, Dadap JI, Eisenthal KB and Heinz TF (1999) Anisotropic orientational motion of molecular adsorbates at the air–water interface. *Journal of Physical Chemistry B* 103: 3425–3433.

irradiation or emission. Two additional general observations are appropriate. First, the lineshape analysis for such measurements must take into account the fact that the SHG response derives from both the real and imaginary parts of $\chi_s^{(2)}$, as we discussed above. Second, in the absence of any other knowledge, SHG spectroscopic data contain an ambiguity about whether the resonances correspond to the fundamental or SH photon energy. This ambiguity, however, can be resolved in a straightforward fashion by complementing the SHG measurement with a sum-frequency measurement in which one frequency is scanned and one is held fixed, as was demonstrated in the study associated with Figure 9.

Time-Resolved Measurements

One of the fascinating frontiers that can be addressed by SHG measurements is probing the ultrafast dynamics of surfaces and interfaces by the pump-probe method. This possibility is exemplified by a variety of studies ranging from the dynamics of laser-driven phase transitions, charge-transfer, solvation, phonon dynamics, and orientational and torsional dynamics.

An example of applying time-resolved SHG to examine ultrafast molecular dynamics at the liquid/vapor interface of water is presented in Figure 10. In these studies, the rotational motion of dye molecules at the interface, which probes the local environment, is followed in real time. This is accomplished by photo-exciting the dye molecules at the interface. Because polarized pump radiation is used, an anisotropic orientational distribution of excited-state molecules is established. Since the excited molecules have a different SHG response, a change in the signal is observed immediately. Subsequently, the anisotropic orientational distribution relaxes to the equilibrium configuration. The time evolution of the rotational anisotropy is followed by detecting the SH of a probe laser pulse as a function of the delay time. Through a comparison of the results for different initial anisotropic distributions (produced by pump beams with differing polarizations), one may deduce rates for both in-plane and out-of-plane orientational relaxation. This example illustrates the possibilities of time-resolved SHG measurements for probing the detailed dynamics of molecules at an interface on the ultrafast timescale.

Spatially Resolved Measurements

Up to this point, our discussions have centered on spatially homogenous surfaces and interfaces. Surface SHG has also been exploited to probe the lateral variation of inhomogeneous surfaces with the micron to sub-micron resolution characteristic of optical microscopy techniques. Spatial resolution may be achieved simply by detecting the nonlinear response with a focused laser beam that is scanned across the surface. A large area of the surface may also be illuminated and the emitted nonlinear radiation imaged. A wide range of applications of this imaging capability has emerged, including the probing of biological materials, such as cells and tissues, to probing spatially varying electric fields and magnetic domains. Within the biological context, SH microscopy has recently become an active research direction. In these applications, while probing with SHG typically senses non-centrosymmetric structures in bulk samples as opposed to interfaces, most of the aspects of SHG presented in this article still apply. To further improve the spatial resolution of SHG microscopy, researchers have also employed near-field techniques for spatial resolutions below the diffraction limit.

Complementary approaches to the real-space imaging just described have also been developed based on SH diffraction. In this scheme, periodically modulated surfaces are established holographically on a surface by utilizing two interfering laser beams. In addition to the reflected and transmitted SH beams, additional beams are then generated at well-defined angles as dictated by the grating period. This approach has been used to monitor the relaxation of a monolayer grating of adsorbates in time, which yields precise information on the rate of adsorbate surface diffusion.

Probing Electric Fields at Interfaces

For centrosymmetric media, the SHG process shows a high degree of sensitivity to the presence of electric fields. This property follows from the fact that an applied electric field can break the inversion symmetry, just as an interface does. The resulting process, termed electric-field induced SHG (or EFISH), has been known for many years. From a phenomenological standpoint, it can usually be described as a nonlinear response of the form $\mathbf{P}^{(3)}(2\omega) = \varepsilon_0 \boldsymbol{\chi}^{(3)} : \mathbf{E}(\omega)\mathbf{E}(\omega)\mathbf{E}$, where $\mathbf{E}$ is the electric field being probed and $\mathbf{E}(\omega)$ is the laser pump field. Within the context of surfaces and interfaces, there are many circumstances where probing such electric fields is of interest and importance. The SHG technique provides high sensitivity, good spatial resolution, accessibility to buried interfaces, and, in the pump-probe implementation, extremely high time resolution. In addition, with appropriate measurements of the polarization dependence and homodyne detection, the full vector character of the electric field can be determined.

The behavior of electric-field induced SHG is similar to that arising from the asymmetry of an interface. Indeed, it is formally equivalent to the response of an interface provided that the electric field is present in (or radiates from) a region of less than a wavelength's thickness. The applications of this measurement capability are quite diverse, ranging from studies of charging and acid/base reactions at liquid/solid interfaces to probing fields in metal-oxide-semiconductor structures and Schottky barriers to the measurement of freely-propagating terahertz radiation. In Figure 11, SHG with appropriate polarization control and homodyne detection scheme has been applied to the measurement of the vector character of electric fields present in a dc-biased dipole structure on silicon. By spatially scanning the laser beam, a two-dimensional electric field map can be produced. In a pump/probe scheme, the same measurements can be made with femtosecond time resolution.

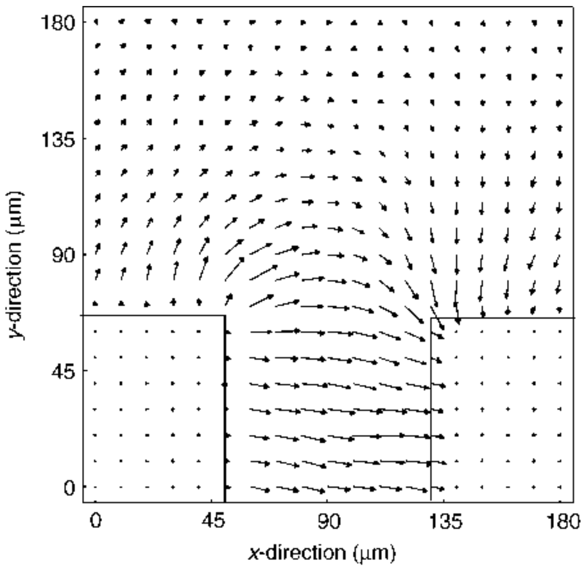

Figure 11 Vector map of the electric field in Si-on-sapphire substrate as deduced by SHG measurements. A voltage is imposed on the rectangular metallic electrodes, which are separated by 80 μm. Reproduced with permission from Dadap JI, Shan J, Weling AS, Misewich JA, Nahata A and Heinz TF (1999) Measurement of the vector character of electric fields by optical second-harmonic generation. *Optics Letters* 24: 1059–1061.

Probing Magnetization at Interfaces

In contrast to the application of an external electric field (a polar vector), the presence of an external magnetic field (an axial vector) does not break the inversion symmetry of a material. Thus, SHG retains its interface selectivity for magnetic effects. SHG measurements, as an interface-specific probe, thus provide an excellent complement to the linear magneto-optical Kerr effect (MOKE), which is sensitive to magnetic effects within the penetration depth of the light.

The ability of the SHG technique to sense the magnetization at a surface is illustrated in Figure 12 for a film of epitaxially-grown magnetic Co on nonmagnetic substrate Cu. A hysteresis loop of the magnetization versus the strength of an applied magnetic field is shown, where the magnetization is gauged from the SHG response. Through measurements of the response as a function of the thickness of

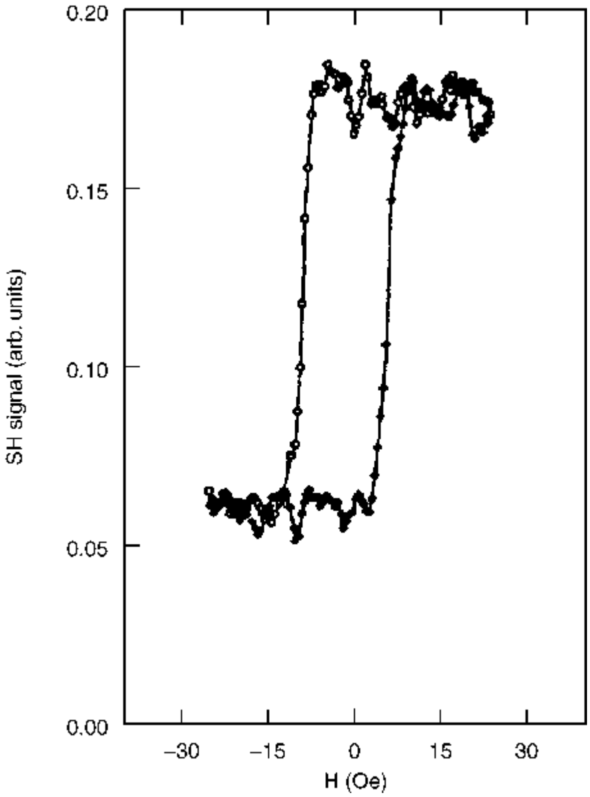

Figure 12 SH response of a Co film on a Cu(001) substrate as a function of the applied magnetic field. A hysteresis loop corresponding to the surface response is observed. Reproduced with permission from Wierenga HA, de Jong W, Prins MWJ *et al.* (1995) Interface magnetism and possible quantum-well oscillations in ultrathin Co/Cu films observed by magnetization induced second harmonic generation. *Physical Review Letters* 74: 1462–1465.

the magnetic film, the researchers demonstrated that the SH response reflected the interface properties, rather than those of the bulk film.

Probing Micro- and Nanoscale Structures

In addition to the microscopy studies described above that probe spatially inhomogeneous surfaces, there has been considerable interest in SHG from a variety of strongly textured surfaces and small particles. From the standpoint of the optical response of materials, particular interest has been attached to studies of metallic surfaces with surface roughness. For appropriate roughness of low-loss materials, such as silver, very significant local-field enhancements can be achieved. Since these field enhancements enter into the strength of the SH emission nonlinearly, overall enhancements in the SH power by several orders of magnitude have been demonstrated. This provides both a useful method of study of such surface enhancement effects and a scheme for amplifying the relatively weak surface nonlinear response.

Beyond these problems associated with roughened surfaces, considerable attention has recently been directed towards the application of the SHG technique for probing micro- and nano-scale particles. The fundamental principle of the interface selectivity of the SHG process for such systems comprised of centrosymmetric materials remains unaltered. The interaction of these particles with the radiation fields is, however, significantly modified. For the case of particles with dimensions of several optical wavelengths, the overall SHG signal strength may be comparable to that from a corresponding area of a planar surface. The radiation pattern will, of course, be strongly modified and will generally be diffuse in character. While this makes detection (and background discrimination) somewhat more difficult than for the planar geometry, many SHG measurements of surface and interface properties have been carried out in this regime. As the particle size becomes progressively smaller, particles with overall inversion symmetry begin to exhibit appreciable cancellation of the SH emission. Indeed, in the limit of vanishing size, full cancellation is expected. This tendency is, however, offset by the increasing surface-to-bulk ratio and the possibility of increasing the number of particles probed. It has been demonstrated experimentally that the surfaces of particles in the nanometer length scale can indeed be probed by SHG.

From the point of view of applications, this extension of the SHG technique to probe micro- and nano-structured materials opens up a rich variety of new possibilities. Results have been reported on metal and semiconductor nanoparticles in suspension, as well as on colloids of minerals. In addition, liquid droplets have been investigated. Self-assembling structures such as liposomes and vesicles have been examined. For these structures, many of the same capabilities of SHG as demonstrated for the planar geometry, such as probing of adsorption and charging effects, have been demonstrated. More uniquely, molecular transport *through* the liposome membrane could also be observed and characterized. The possibility of extension of these measurements to biological systems is of obvious interest and promise.

Concluding Remarks

In this article, we provided a brief review of surface SHG as a technique for interface-specific probing of materials. In addition to discussing some of the fundamental ideas underlying the method, we outlined the wide variety of information that may be obtained from such measurements, ranging from surface symmetry and molecular ordering to interface spectroscopy and ultrafast dynamics. While several aspects of the method have now become quite well established, developments in widely available modelocked lasers with excellent performance characteristics have greatly enhanced the ease of making surface SHG measurements in a wide class of material systems. In addition, new developments and applications of the method have continued to emerge apace. Of particular note in this regard are applications to probe electric fields and magnetization at interfaces, as well as advances in the use of the method for interface-specific probing of micro- and nanoscale structures. The sharp increase in applications of SHG for probing systems of biological importance constitutes another important development, and one with much promise for the future.

See also

Materials for Nonlinear Optics: Organic Nonlinear Materials. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation. **Physical Applications of Lasers:** Sum-Frequency Generation at Surfaces.

Further Reading

Bennemann KH (ed.) (1998) *Nonlinear Optics in Metals*. Oxford, UK: Clarendon.

Bloembergen N (1965) *Nonlinear Optics*. New York: WA Benjamin.

Boyd RW (1992) *Nonlinear Optics*. New York: Academic.

Brevet PF (1997) *Surface Second Harmonic Generation*. Lausanne, Switzerland: Presses Polytechniques et Universitaires Romandes.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge.

Corn RM and Higgins DA (1994) Optical second-harmonic generation as a probe of surface-chemistry. *Chemical Review* 94: 107–125.

Dadap JI and Heinz TF (2001) Nonlinear optical spectroscopy of surfaces and interfaces. In: Moore JH and Spencer ND (eds) *Encyclopedia of Chemical Physics and Physical Chemistry*, pp. 1089–1125. Bristol, UK: IOP.

Eisenthal KB (1996) Liquid interfaces probed by second-harmonic and sum-frequency spectroscopy. *Chemical Review* 96: 1343–1360.

Heinz TF (1991) Second-order nonlinear optical effects at surfaces and interfaces. In: Ponath H-E and Stegeman GI (eds) *Nonlinear Surface Electromagnetic Phenomena*, pp. 353–416. Amsterdam: North Holland.

Liebsch A (1995) Electronic excitations at metal surfaces. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 479–532. Amsterdam: Elsevier.

Lüpke G (1999) Characterization of semiconductor interfaces by second-harmonic generation. *Surface Science Report* 35: 75–161.

McGilp JF (1995) Optical characterisation of semiconductor surfaces and interfaces. *Progress in Surface Science* 49: 1–106.

McGilp JF (1999) Second-harmonic generation at semiconductor and metal surfaces. *Surface Review Letters* 6: 529–558.

McGilp JF, Patterson CH and Weaire DL (eds) (1995) *Epioptics: Linear and Nonlinear Optical Spectroscopy of Surfaces and Interfaces*. Berlin, New York: Springer.

Prasad PN and Williams DJ (1991) *Introduction to Nonlinear Optical Effects in Molecules and Polymers*. New York: Wiley.

Reider GA and Heinz TF (1995) Second-order nonlinear optical effects at surfaces and interfaces: recent advances. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 413–478. Amsterdam: Elsevier.

Richmond GL (1997) Vibrational spectroscopy of molecules at liquid/liquid interfaces. *Analytical Chemistry* 69: A536–A543.

Richmond GL, Robinson JM and Shannon VL (1988) Second harmonic generation studies of interfacial structure and dynamics. *Progress in Surface Science* 28: 1–70.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Shen YR (1989) Optical second harmonic-generation at interfaces. *Annual Review of Physical Chemistry* 40: 327–350.

Single Molecule Spectroscopy

X Michalet and S Weiss, University of California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Textbooks in physics, chemistry, and biology abound in diagrams depicting single-particle, single-molecule or single-enzyme properties, reactions or interactions with their environment. Such pictures and the numbers associated with them, however, were, until recently, extrapolated from measurements performed on large ensembles of molecules, which only yield quantities averaged out over time, space, conformations and/or the local environment. Advances in detector sensitivity and improvements in instrument design do now provide scientists with tools to probe single molecules with light, and monitor their photophysical properties with exquisite sensitivity and time resolution. This spectroscopic information is used to explore the molecular structure, conformational dynamics, local environment, and intermolecular interactions. The collected data generate a wealth of new information on single-molecule physical chemistry or biomechanics, but also in polymer gel physics or glass dynamics and other domains of solid state physics.

This article reviews the field of single-molecule spectroscopy. First, the principles underlying single-molecule spectroscopy approaches are introduced, and this is followed by a discussion of the different spectroscopic techniques available to the experimentalist and what kind of information can be extracted with each of them. It also discusses important photophysical properties of molecules that influence the outcome of single-molecule spectroscopic measurements. The next section presents some applications of these techniques, focusing on physical properties of materials and biological applications of single-molecule spectroscopy. The last part briefly reviews recent developments and outlines the future prospects of single-molecule spectroscopy.

Principles of Single-Molecule Spectroscopy

Motivation

Measurements performed on ensembles of molecules only yield averaged-out information. This averaging process is of several kinds: average over molecular heterogeneity (e.g., chemical composition, oxidation state, or spatial conformation) and average over different or changing local environments. Moreover, dynamic processes studied by ensemble measurements cannot sort out molecules exhibiting fast or slow reaction times, although such differences are expected from the stochastic nature of any chemical kinetics. The interest in single-molecule analysis is for these reasons manifold. At low temperature, being able to study a single molecule and its evolution under the influence of various external parameters allows for precise tests of molecular quantum electrodynamics. At higher temperature, it gives access to the temporal evolution of different nanodomains of a bulk material in which single molecules are embedded as reporters of their environment. In biology and biochemistry, the building blocks of any process are individual molecules (genes, enzymes, motors, filaments, etc.); the ability to fully understand their kinetic behavior at the single-molecule level is thus of utmost importance. Finally, reliable observation of single molecules is a prerequisite of the full development of nanotechnology and quantum computing.

Historical Outlook

Single-molecule experiments were first performed on isolated ion channels by the patch clamp technique, which allows the recording of single-ion translocation through a pore-like protein embedded in a fluid membrane. Scanning tunneling and atomic force microscope observations later opened the way to investigations of nanometer-scale objects on surfaces. Only recently have optical methods reached the sensitivity required to detect, image, manipulate, and follow the spectroscopic evolution of single molecules on surfaces, in solids, liquids, and biological membranes as well as inside live cells.

Single-molecule spectroscopy (SMS) was initially demonstrated in a system of pentacene molecules embedded in a *p*-terphenyl host crystal at liquid helium temperature. Absorption (leading to fluorescence excitation) spectroscopy in this system takes advantage of the very narrow zero-phonon absorption line (ZPL) of the rigid pentacene molecule (Figure 1). The position in laser-frequency space of each molecule's ZPL indeed depends on the local fields, allowing for precise molecular selectivity. However this approach is limited to a few molecule–host pairs and requires cryogenic techniques, due to broadening of the ZPL above 10 K.

Room temperature studies started with total internal reflection (TIR) microscopy and standard flow-cytometry methods in the mid-1980s. In the latter case, single molecules randomly passing through the excitation volume gave rise to sudden bursts of photons emitted during their brief transit. Initially developed for rapid DNA sequencing, this simple approach is now widely used for more sophisticated studies in its confocal version (Figure 2), which will be detailed in the following.

The first image of a single molecule at room temperature was later obtained using scanning near-field optical microscopy (Figure 3), rapidly followed by similar achievements using far-field confocal microscopy, at the expense of a lower spatial resolution, but with much more ease of use. TIR and wide-field fluorescence imaging are now used to image and study the two-dimensional and three-dimensional diffusion trajectories of single fluorescent molecules (Figure 4), making a come-back after the pioneering studies of the early 1980s.

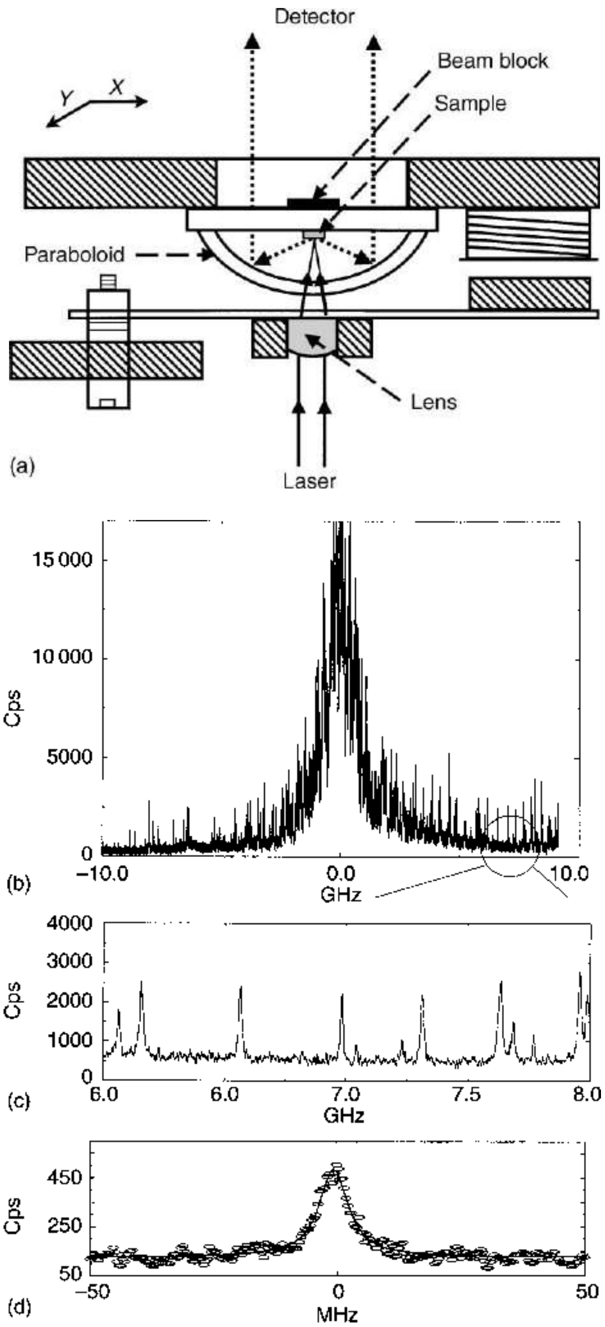

Figure 1 Fluorescence excitation spectra of pentacene in *p*-terphenyl at 1.5 K. (a) Schematic of the cryogenic setup used. A laser beam is focused onto the sample through a lens into the cryogenic chamber. The sample can be moved in the *XY* plane, with a magnet and coil pair. A beam block rejects the incident laser light (black arrow). A paraboloidal mirror collects the emitted fluorescence (dashed arrows) and redirects it towards the detector (photomultiplier tube). The laser frequency (linewidth 3 MHz) is tuned around the absorption line center at 592.321 nm. (b) Broad scan showing the inhomogeneously broadened line shape of pentacene molecules embedded in a *p*-terphenyl crystal at 1.5 K. The sharp lines correspond to spectra of individual molecules (see panel c) (cps: counts per second). (c) Detail of the previous spectrum showing spectra of several individual molecules. The variability of intensities is due to the different positions of the molecule in the excitation beam. (d) Individual molecule spectrum obtained at lower excitation power exhibiting a Lorentzian shape (solid line) centered at 592.407 nm with a lifetime-limited width of 7.8 MHz (intensity is measured in counts-per-second for all three spectra). (a): adapted from Ambrose WP, Basché T and Moerner WE (1991) Detection and spectroscopy of single pentacene molecules in a *p*-terphenyl crystal by means of fluorescence excitation. *Journal of Chemical Physics* 95: 7150–7163. (b)–(d): after Moerner WE (1994) Examined nanoenvironments in solids on the scale of a single, isolated impurity molecule. *Science* 265: 46–53. Copyright 1994, The American Association for the Advancement of Science.

Requirements for Single-Molecule Sensitivity

Observing single molecules may be as easy as finding a needle in a haystack if background, signal and noise are not carefully optimized. Simple considerations illustrated in Figure 5a help measure the technical challenge bearing on single-molecule optical detection. Whatever the type of detected signal (fluorescence, Raman scattering, or others), it is characterized by an absorption cross-section σ (probability of photon absorption) and a quantum yield Q (number of emitted photons per absorbed photon). The optical setup transmits an excitation power P at a frequency ν, resulting in a signal rate $s = EQ\sigma P/(Ah\nu)$ expressed in counts per second (or Hz). This rate is proportional to P/A where A is the cross-section of the excitation beam in the plane of the molecule. E is the collection efficiency of the setup. The environment adds a background rate b per unit volume and unit power, and the detector contributes a dark count rate D. Neglecting noise from the readout electronics, a measurement of duration τ will yield the following signal-to-noise ratio (SNR):

$$\mathrm{SNR} = \frac{s\tau}{\sqrt{(s + bVP + D)\tau}} \qquad [1]$$

where V is the excited volume. The denominator is the root mean square of all contributions to the measured count rate, considered shot-noise limited. The SNR obviously increases at larger integration time τ. Another important quantity is the signal-to-background ratio (SBR):

$$\mathrm{SBR} = \frac{EQ\sigma}{bVAh\nu} \qquad [2]$$

These ratios can be increased by improving the collection efficiency E or decreasing the excitation volume V. A larger excitation power or a longer integration time will improve the SNR without affecting the SBR. Typical values for the above

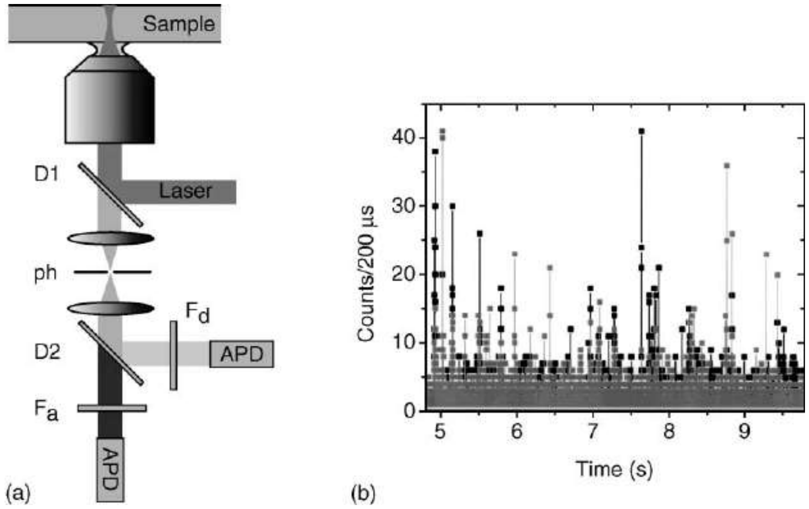

Figure 2 Single-molecule fluorescence burst detection. (a) Schematic of the confocal setup used for single-molecule fluorescence detection. A collimated laser beam is focused through a high numerical aperture (NA) objective in a liquid cell. Molecules diffusing through the diffraction-limited excitation volume (~1 fl) emit photons during their passage, which are collected by the same objective. A dichroic mirror (D1) separates excitation and emission wavelengths. The emitted light is focused on a pinhole (ph) in order to reject out-of-focus background light, and finally recollimated onto two different avalanche photodiodes (APDs) after spectral separation by a dichroic mirror (D2). Filters specific for each channel (F_a: acceptor channel, F_d: donor channel) may be used to improve spectral separation. (b) 5 s time trace of recorded bursts for a 30 pM solution of doubly labeled DNA in 20 mM sodium phosphate buffer. The DNA constructs are used for distance measurement (detailed in **Figure 9**) and have a TMR fluorophore attached to one end of the DNA molecule and a CY5 fluorophore attached to the *n*th base from the end. The time traces correspond to $n = 12$. Gray: TMR fluorescence (donor); black: CY5 fluorescence (acceptor). The laser excitation (0.6 mW, 514 nm) is focused 10 μm within the solution through an oil immersion objective (numerical aperture: 1.4). For each detected photon, the APD generates a pulse, which is recorded by a computer-embedded acquisition board. Counting these events into 200 μs bins, the resulting time trace exhibits distinct bursts distributed randomly in time, with duration and amplitude corresponding to a Brownian diffusion of individual molecules through the excitation volume. (b): is adapted from Deniz AA, Dahan M, Grunwell JR, *et al.* (1999) Single-pair fluorescence resonance energy transfer on freely diffusing molecules: Observation of Förster distance dependence and subpopulations. *Proceedings of the National Academy Sciences USA* 96: 3670–3675. Copyright 1999, The National Academy of Sciences.

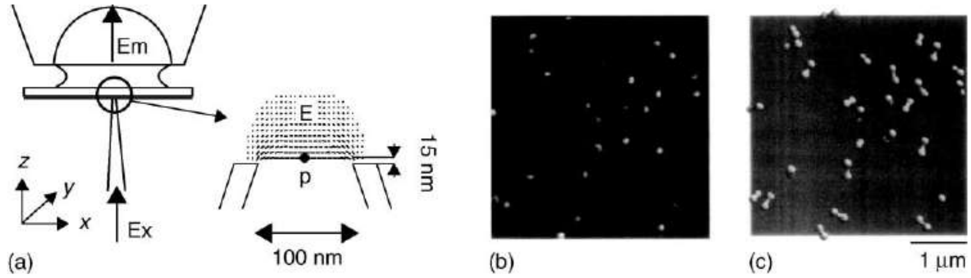

Figure 3 Scanning near-field imaging of single molecules at room temperature. (a) Schematic of a scanning near-field optical microscope setup used for single-molecule imaging. An aluminum-coated tapered optical fiber (raster-scanned at nanometer distance from the sample) serves as a waveguide for laser excitation. Shear-force feedback keeps the tip at a constant distance from the sample, resulting in a signal used for nanometer-resolution topographic reconstruction of the scanned area. The excitation volume and the corresponding local evanescent electric field are detailed in the expanded view. Fluorescence light emitted by individual molecules is collected by an oil immersion, high NA objective and recorded by an avalanche photodiode. (b) $4 \times 4\ \mu m^2$ area of a polymethylmethacrylate (PMMA) coated coverslip on which a diluted methanol solution of a lipophilic carbocyanine dye, $diIC_{12}$, has been deposited. Scanning was performed with randomly polarized light. (c) Orientation of the corresponding individual emitting dipole determined by modeling the field distribution (expanded view in a), simulation of the absorption intensity with polarized excitation and comparison with data obtained at two different excitation polarizations. (a): Adapted and (b), (c) reprinted with permission from Betzig E and Chichester RJ (1993) Single molecules observed by near-field scanning optical microscopy. *Science* 262: 1422–1425. Copyright the American Association for the Advancement of Science.

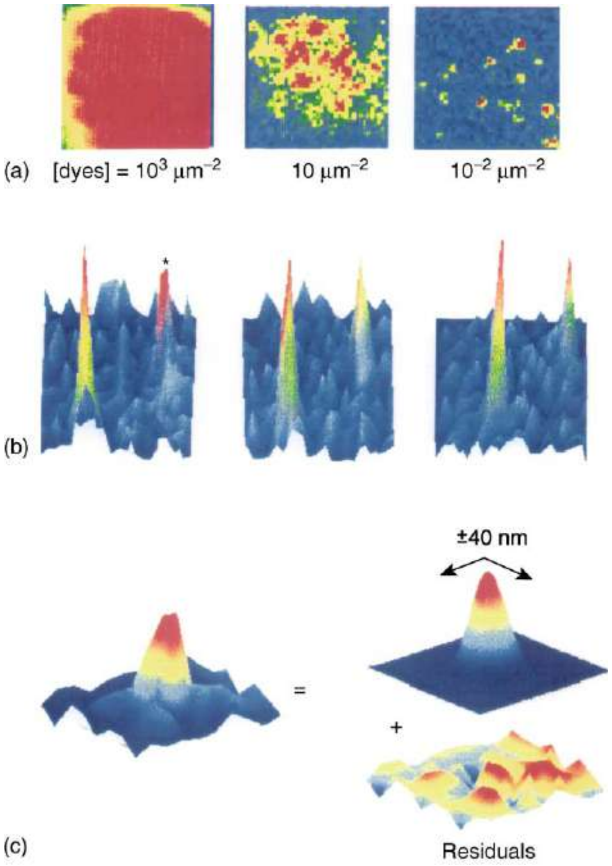

Figure 4 Single molecule diffusion in two dimensions. (a) 6.8 × 6.8 μm^2 images of fluorescently labeled lipids (tetramethylrhodamineinthiocarbamoyl, linked to 1,2-dihexadecanoyl-*sn*-glycero-3-phospho-ethanolamine, DHPE) in a lipid double layer membrane (palmitoyloleoylphosphocholine, POPC) at different concentration. From left to right: 10^3 dyes/μm^2, 10 dyes/μm^2, 10^{-2} dyes/μm^2. Each image was obtained with an epifluorescence microscope, illuminated by a defocused 514 nm Ar laser line (~60 kW/cm^2), and 5 ms integration on a nitrogen cooled CCD camera. Color scale: blue = 0, $d = 60$ counts). Image intensity at 10^3 dyes/μm^2 was divided by 60. (b) Sequence of three 5.4 × 5.4 μm^2 images of two labeled lipids observed at 105 ms intervals (actual acquisition rate was one frame every 35 ms, integration time: 5 ms). (c) Subregion of a 2.4 × 2.4 μm^2 membrane area showing the peak marked by a star in B. The nonlinear least-square fit of a two-dimensional Gaussian profile to the peak yields values of the fluorescence intensity $I = 168 \pm 25$ counts, width $r = 480 \pm 80$ nm and two-dimensional position with 40 nm resolution. Adapted from Schmidt T, Schutz GJ, Baumgartner W, *et al.* (1996) Imaging of single molecule diffusion. *Proceedings of the National Academy of Sciences USA* 93: 2926–2929. Copyright 1996, The Academy of Sciences.

parameters in the case of a confocal microscopy study of freely diffusing fluorescein isothiocyanate (FITC) in water, using avalanche photodiode detectors (APD) are: $E = 10^{-2}$, $Q = 0.85$, $\sigma = 2.8 \times 10^{-16}\,\text{cm}^2$, $A = 0.1\,\mu\text{m}^2$, $\lambda = c/\nu = 525\,\text{nm}$, $bV = 10\,\text{Hz}/\mu\text{W}$, $D = 100\,\text{Hz}$. With an excitation power of 100 μW and a 1 ms integration time, a count rate of 629 counts/ms can be obtained, of which 1 comes from the background sources, leading to a SNR ~25 and SBR ~629. Although these are acceptable figures, they have to be balanced by the fact that single molecules have a finite lifespan. Single molecules in an oxygen-rich environment typically emit on the order of 10^6 photons before irreversible degradation (photobleaching). With the above parameters, this amounts to a total emission of ~3.2 ms. This is enough to observe freely diffusing molecules during their transit time of a few hundred microseconds. For a fixed molecule, however, it sets a stringent limit on the total duration of a single-molecule observation, and all efforts are put on reducing the excitation volume and intensity, as well as background sources. Similar considerations hold on the detector side, where a trade-off has to be found between time-resolution and quantum efficiency.

In addition to SNR and SBR issues, care has to be taken to ensure that the collected signal originates from a single molecule. Two different strategies can be envisioned:

(i) work at low concentration, such that at most one molecule is present in the excitation volume (Figure 5b), or equivalently, minimize the excitation volume;
(ii) use a selective excitation or emission detection protocol, such that only one molecule is excited or detected within the sampled volume (Figure 5c).

The first case has already been illustrated in Figure 2, in which a solution of fluorophores at low concentration yields separate bursts corresponding most of the time to transits of individual molecules. The second strategy can be adopted for the example illustrated in Figure 1, in which the ZPL of individual molecules lying in the wings of the ensemble spectrum are well separated. Tuning the laser to the peak absorption frequency of a molecule will excite this molecule only, allowing for a separate study of this molecule.

In summary, a set of six criteria can be defined for SMS:

1. The observed density of emitter is compatible with the known concentration of individual molecules and scales with the original concentration.
2. The observed intensity level is consistent with that of a single emitting molecule.
3. Each immobilized emitter has a well-defined absorption or emission dipole (for organic dyes usually linear, but for other emitters such as a fluorescent semiconductor nanocrystal, possibly circular).

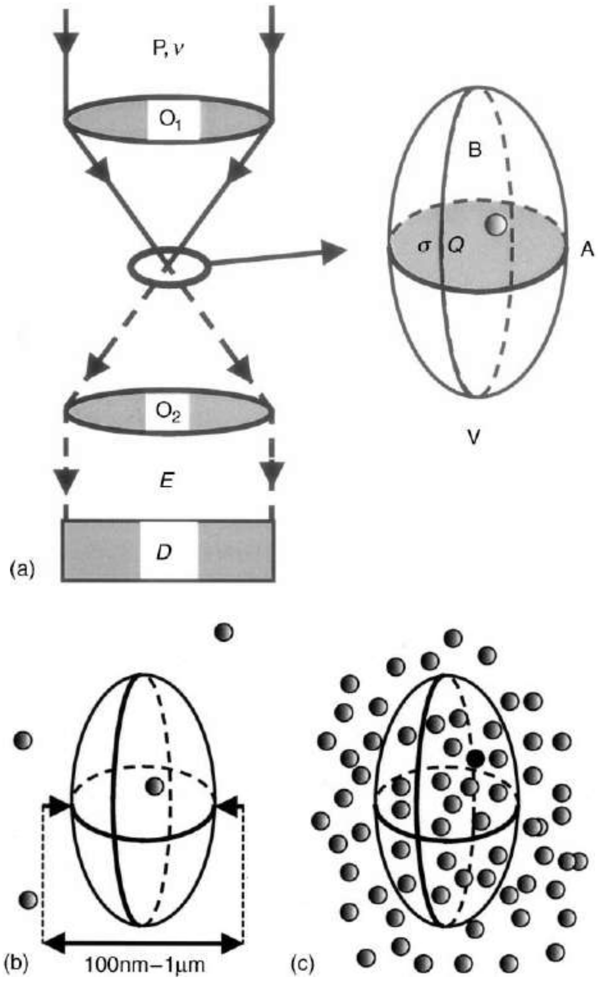

Figure 5 Signal, background and noise issues in single-molecule detection. (a) In a single-molecule spectroscopic experiment, excitation light (power *P*, frequency ν) is focused on the sample through an optical element O_1 (objective lens, fiber tip), exciting a diffraction-limited volume of section *A*, represented by an ellipsoid in the expanded view. The absorption process is characterized by an absorption cross-section σ and the emission by a quantum yield *Q*. The background contribution per watt of incident power is *B*, whereas the detector will have a dark count rate *D*. The signal is collected by an optical element O_2 (possibly confounded with O_1), and further guided to the detector, with an overall efficiency *E*. (b) The signal of a single molecule can be separated from that of other molecules if the probability of finding two or more molecules in the probed volume is negligible. This can be obtained either by a low concentration, or by reducing the excitation volume. (c) If each molecule in the excited volume can be characterized by a different spectral property (which may or may not depend on its exact environment), it is possible to use very selective excitation or detection techniques to probe just one molecule in the sampled volume (the black sphere in the ellipsoid).

4. Fluorescence emission exhibits only two levels (on/off) exemplified in blinking or photobleaching.
5. If there are more emission levels due to spectral jumps (see the section on fluorescence), blinking and spectral jumps are correlated.
6. Antibunching of the emitted photons permits us to test whether the detected signal comes from a single quantum emitter or more (see the section on biological studies below).

Types of Spectroscopy

Fluorophores

A variety of fluorescence emitters (or fluorophores) have been investigated with single-molecule spectroscopy techniques. Fluorescent organic molecules, green fluorescent proteins (GFP) and other biological fluorescent proteins, conjugated polymers (J-aggregates), light-harvesting complexes (e.g., B-phycoerythrin) comprising several fluorophores acting effectively as a single quantum emitter, or semiconductor nanocrystals, are but a few examples of systems which have been extensively studied. Each of these systems exhibits fluorescence based on specific processes, which cannot be described in detail here. A simple picture valid for single fluorescent dyes will be presented instead. It captures, however, the essential ingredients needed to describe the different parameters accessible with single-molecule spectroscopy.

Fluorescence Emission

Most single-molecule experiments use one-photon fluorescence excitation because of its relatively large cross-section ($\sim 10^{-16}\,\text{cm}^2$). It also allows for a simple rejection of the excitation light: the emitted photons have a lower energy because of losses in intramolecular relaxation processes (the so-called Stokes shift). Excitation and emission processes are conveniently represented by a Jablonski diagram (Figure 6a) depicting the initial, final, and intermediate electronic and vibrational states of the molecule. Upon excitation to the first excited state, the molecule has a certain probability to be shelved in a nonemitting triplet state for up to milliseconds, resulting in 'dark' or off-states (Figure 6b,c). This results in a saturation of the emitted count rate *R* (Figure 6d) according to:

$$R = R_s \frac{I/I_s}{1 + I/I_s}, \qquad [3]$$

where the saturation intensity $I_s \sim 10\text{–}100\,\text{kW/cm}^2$ and $R_s \sim 1\text{–}20\,\text{MHz}$ for typical fluorophores. Incidentally, this puts a limit in SNR enhancement due to saturation of the emitted signal (triplet bottleneck). Other fluorescence emitting 'single molecules' like semiconductor nanocrystals exhibit similar dark state intervals, although for different reasons (Auger ionization or surface trapping of carriers). Finally, spectral jumps are also observed at room temperature (Figure 6e). This phenomenon results in a shift of absorption and emission maxima due to the changing environment of the molecule, or a sudden conformational change. The corresponding changes in excitation and emission spectra are noticed by sudden fluctuations in the emitted intensity

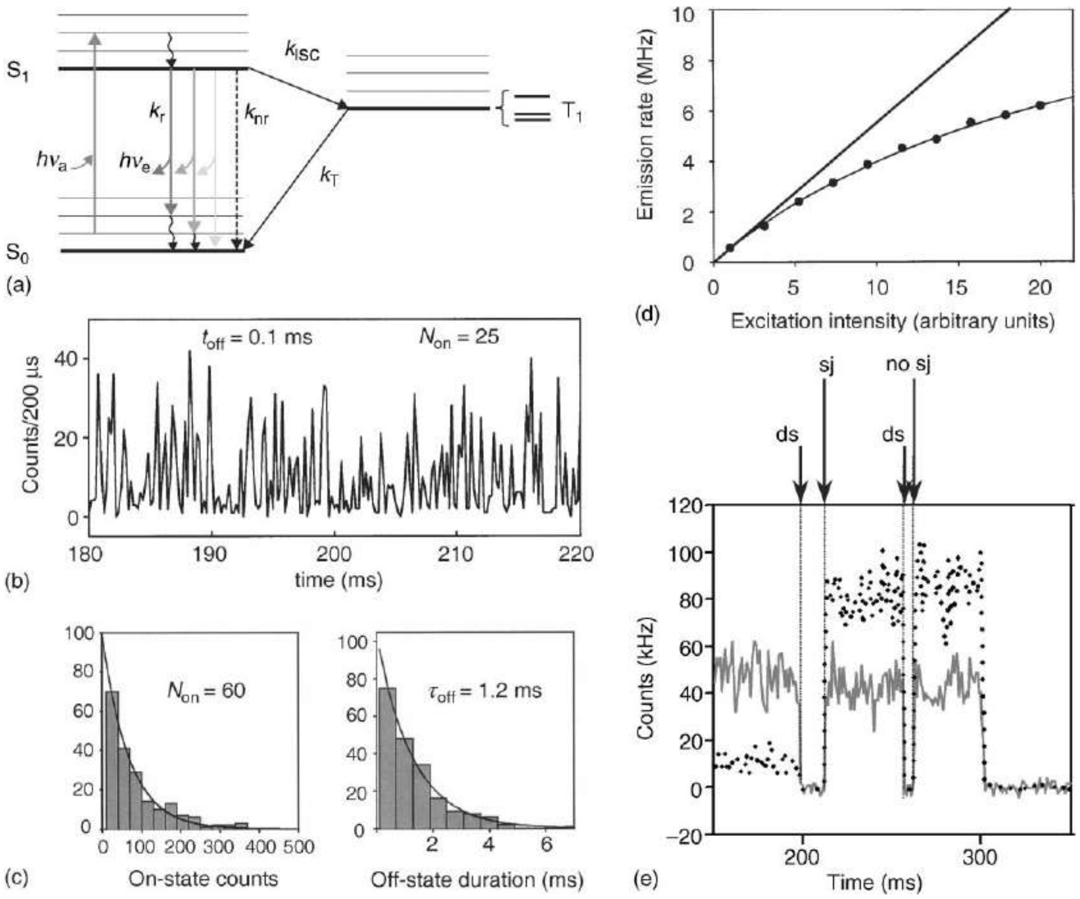

Figure 6 Single-molecule fluorescence. (a) Jablonski diagram for fluorescence (at room temperature). Upon absorption of a photon of energy $h\nu_a$ close to the resonance energy $E_{S_1} - E_{S_0}$, a molecule in a vibronic sublevel of the ground singlet state S_0 is promoted to a vibronic sublevel of the lowest excited singlet state S_1. Nonradiative, fast relaxation brings the molecule down to the lowest sublevel in picoseconds. Emission of a photon of energy $h\nu_e < h\nu_a$ (radiative rate k_r) can take place within nanoseconds and bring the molecule back to one of the vibronic sublevels of the ground state. Alternatively, collisional quenching may bring the molecule back to its ground state without photon emission (nonradiative rate k_{nr}). A third type of process present in organic dye molecules is intersystem crossing to the first excited triplet state T_1 (rate k_{ISC}). Relaxation from this excited state back to the ground state is spin-forbidden and thus the lifetime of this state is in the order of microseconds to milliseconds. Relaxation to the ground state takes place either by photon emission (phosphorescence) or nonradiative relaxation. The fluorescence lifetime is defined by $\tau = 1/\Gamma = (k_r + k_{nr})^{-1}$. (b) Fluorescence time trace (40 ms excerpt) of a Texas Red (TR) molecule attached to a 20-nucleotides-long single-stranded DNA molecule immobilized on a silanized glass coverslip. The confocal setup is similar to that depicted in Figure 2. Excitation: 514 nm, circularly polarized, 6 kW cm^{-2}, binning: 200 µs. On- and off-states (the latter corresponding to the molecule being in the triplet state) alternate stochastically. (c) Histograms of on-state count (left) and off-state duration (right) for the time trace partly shown in (b). The decay parameters of the exponential fits (solid line) are indicated. $t_{off} = 1.2$ ms is the triplet state lifetime. $N_{on}^{-1} = k_{ISC} = 1.7\%$. (d) Intensity dependence of the emission rate of a single Texas Red molecule. The data departs from a linear dependence because of saturation due to intersystem crossing. The fitted curve yields a saturation rate of $R_S = 6.5$ MHz and a triplet lifetime $t_{off} = (k_{ISC}\ R_S)^{-1} = 10$ µs. The difference between this value and the value obtained in (c) illustrates the heterogeneity uncovered by single-molecule measurements. (e) Spectral jumps of a single Texas Red molecule attached to a single-stranded DNA molecule. ds: dark state, sj: spectral jump, no sj: dark state not followed by a spectral jump. Cases where spectral jumps are observed without noticeable dark states may still correspond to dark states with durations shorter than the time resolution of the experiment. (b)–(d) Adapted from Ha T, Enderle T, Chemla DS, *et al.* (1997) Quantum jumps of single molecules at room temperature. *Chemical Physics Letters* 271: 1–5. Copyright 1997, Elsevier Science B. V. (e) Adapted from Ha T, Enderle T, Chemla DS and Weiss S (1996) Dual-molecule spectroscopy: molecular rules for the study of biological macromolecules. *IEEE Journal of Selected Topics in Quantum Electronics* 2: 1115–1127. Copyright 1996 IEEE.

(due to wavelength dependence of the absorption cross-section).

Fluorescence Lifetime

A fluorescence photon is emitted within nanoseconds after absorption of the excitation photon. The decay law is usually mono-exponential, characterized by a lifetime $\tau = \Gamma^{-1}$ (Figure 6a). However, the lifetime depends on the emission spectrum (which can change due to spectral shifts) and the local environment. For instance, the proximity to a dielectric or metallic surface markedly changes it via perturbations of the intramolecular transition matrix elements. Thus, fluorescence lifetime is a sensitive probe of the environment of the molecule (Figure 7). Lifetime measurements require a pulsed laser source and time-correlated single-photon detection and analysis techniques.

Fluorescence Polarization

The efficiency of photon absorption by a fluorescent molecule is proportional to $(\boldsymbol{E}\cdot\boldsymbol{\mu})^2$ where $\boldsymbol{E}$ represents the local electric field and μ the molecule's absorption dipole moment. For a fixed fluorophore, modulation of the excitation polarization thus allows the determination of the absorption dipole orientation (the emission dipole being generally nearly parallel to it). This information in turn allows the determination of the spatial orientation of the molecule. For a mobile molecule, however, more information is needed, since the emission dipole has time to tumble significantly during the few nanoseconds between absorption and emission. The emission polarization is needed to fully recover the relevant information, as illustrated in Figure 8. Different experimental schemes can be used, besides the near-field optical one described in Figure 3 and the confocal one described in Figure 8. In particular, it is important to recover the projection of the polarization on more than two orthogonal axes. Approaches taking advantage of aberrations induced by index mismatch in high numerical aperture (NA) objectives, or using TIR excitation with alternating polarizations, or using amplitude or phase masks, allow a complete determination of this orientation.

Fluorescence Resonance Energy Transfer

A special case of the local influence of the environment on fluorescence is encountered in fluorescence resonance energy transfer (FRET, see Figure 9a), first theoretically explained by Förster in the late 1940s. In case of the presence of a nearby molecule (acceptor) having an absorption spectrum overlapping that of the donor emission, there is a possibility that the energy absorbed by the donor is transferred nonradiatively to the acceptor. The efficiency E of this transfer is given by:

$$E = \left(1 + \left(\frac{r}{R_0}\right)^6\right)^{-1} \qquad [4]$$

where r is the distance between the two emitting centers and R_0 is the Förster radius (in Å):

$$R_0 = a(\kappa^2 n^{-4} Q_D J(\lambda))^{1/6} \qquad [5]$$

In this complex expression, $a = 9.78 \times 10^3$, n is the medium refractive index, Q_D the donor fluorescence quantum yield in the absence of the acceptor and $J(\lambda)$ is a spectral overlap integral (in M^{-1} cm^3). κ^2 is a geometric factor, which averages out to 2/3 in the case of freely rotating dyes. A typical order of magnitude for R_0 is a few nanometers, which gives the range of the energy transfer. FRET is thus very sensitive to the donor–acceptor distance, which makes it a useful molecular ruler. Experimentally, the efficiency is measured as a function of the recorded donor (F_D) and acceptor (F_A) emissions (dual-channel measurement) via:

$$E = \left(1 + \alpha \frac{F_D}{F_A} \frac{Q_A}{Q_D}\right)^{-1} \qquad [6]$$

where Q_A and Q_D are the fluorescence quantum efficiencies of the donor and acceptor, respectively, and α is the ratio of the acceptor and donor channel detection efficiencies. An illustration of this utilization is given in Figure 9b–d, where different predefined distances between donors and acceptors are set by rigid double-stranded DNA molecules. The corresponding donor–acceptor distances are readily measured on diffusing molecules, with subnanometer precision.

Alternatively, the transfer efficiency can be measured via the fluorescence lifetimes as:

$$E = 1 - \frac{\tau_{D(A)}}{\tau_{D(0)}} \qquad [7]$$

where $\tau_{D(A)}$ and $\tau_{D(0)}$ are the donor lifetimes in the presence or absence of an acceptor, respectively.

Other Spectroscopies

Other types of spectroscopic techniques can be used at the single-molecule level. Fluorescence can for instance be excited via a two-photon absorption process, whereby a laser excitation with half the photon energy needed to attain the excited state

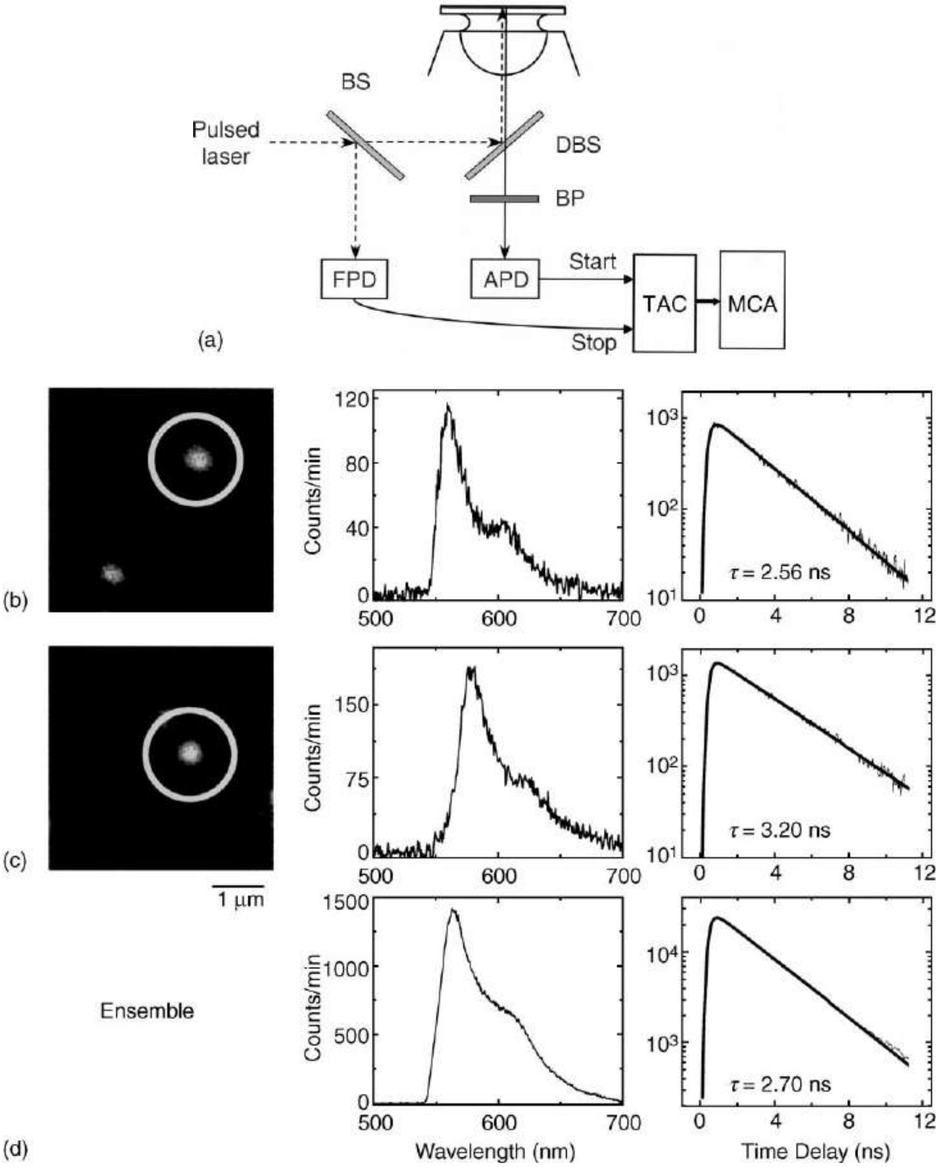

Figure 7 Lifetime measurements. (a) Schematic of a lifetime measurement setup. BS: beamsplitter used to deflect part of the incident pulse towards a fast photodiode (FPD); DBS: dichroic beam splitter used to spectrally separate incident from emitted light; BP: bandpass filter further rejecting stray light; APD: avalanche photodiode. For each detected photon, the APD emits a pulse, which is fed to the start input of a time-to-amplitude converter (TAC). A stop pulse from the laser immediately follows it (regularly emitted at a rate of τ_L^{-1}), allowing the measurement of the separation between both pulses. A multichannel analyzer (MCA) histograms the arrival times, resulting in the curves shown in the right part of Figure 2b–d. (b) and (c) Fluorescence images (left panels), corresponding spectra (middle), and fluorescence decays (right) for two molecules located at a PMMA–air interface. The peak emission wavelength λ_p was 560 nm in (b) and 578 nm in (c). Lifetimes were fitted to a single exponential (dotted curves), with decay times of $\tau = 2.56$ ns (b) and 3.20 ns (c). A ensemble measurement of spectrum and fluorescence decay averaged over several hundred of molecules is shown in (d). The ensemble values are $\lambda_p = 565$ nm and $\tau = 2.70$ ns. (b)–(d): Adapted from Macklin JJ, Trautman JK, Harris TD and Brus LE (1996) Imaging and time-resolved spectroscopy of single molecules at an interface. *Science* 272: 255–258. Copyright 1996, The American Association for the Advancement of Science.

is used. However, due to the very low cross-section and quadratic dependence in the incident energy (two simultaneous photons are needed for excitation), an excitation power several orders of magnitude larger than that for one-photon excitation is needed. In compensation, excitation takes place in a much more reduced volume of the sample, reducing the out-of-focus background contribution. The higher power required increases the probability of photobleaching due to the increased probability of

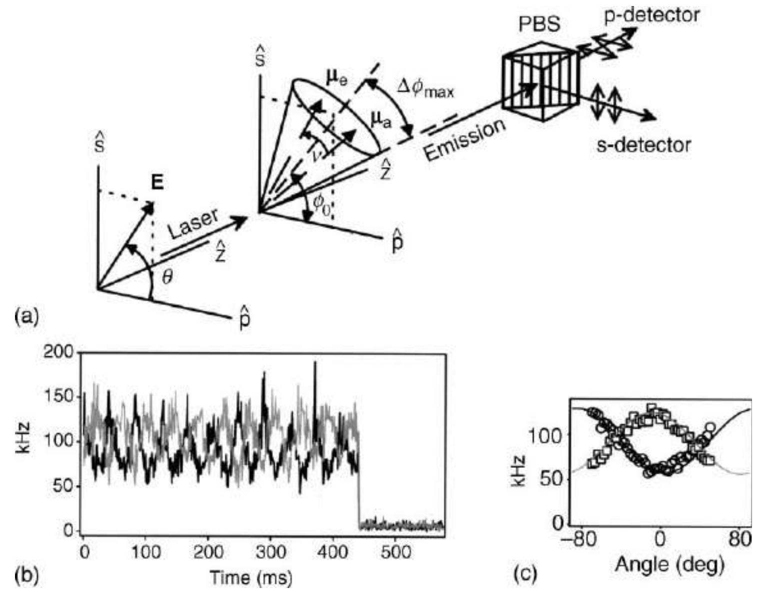

Figure 8 Polarization Measurements. (a) Experiment schematic. **E** is the electric field, making an angle θ with the p polarization axis. The excitation propagates along the z axis, which is also the collection axis. $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_e$ are the absorption and emission dipole moments, initially aligned. ν represents the rotational diffusion of the emission dipole during the excited lifetime. The dipole is supposed to be confined in a cone positioned at an angle ϕ_0 projected on the (s, p) plane, and having a half-angle $\Delta\phi_{max}$. A polarizing beam-splitter (PBS) splits the collected emission into two signals I_s and I_p, which are simultaneously recorded by APDs. (b) Simultaneously recorded I_s (black) and I_p (gray) of a molecule rapidly rotating in a liquid. (c) Same data as in (b), but averaged over 11 on-periods. The fit corresponds to a $\Delta\phi_{max}$ close to 90° (freely rotating molecule) and permits the determination of the constrained rotational diffusion parameters. Adapted from Ha T, Glass J, Enderle T, *et al.* (1998) Hindered rotational diffusion and rotational jumps of single molecules. *Physical Review Letters* 80: 2093–2096. Copyright 1998, The American Physical Society.

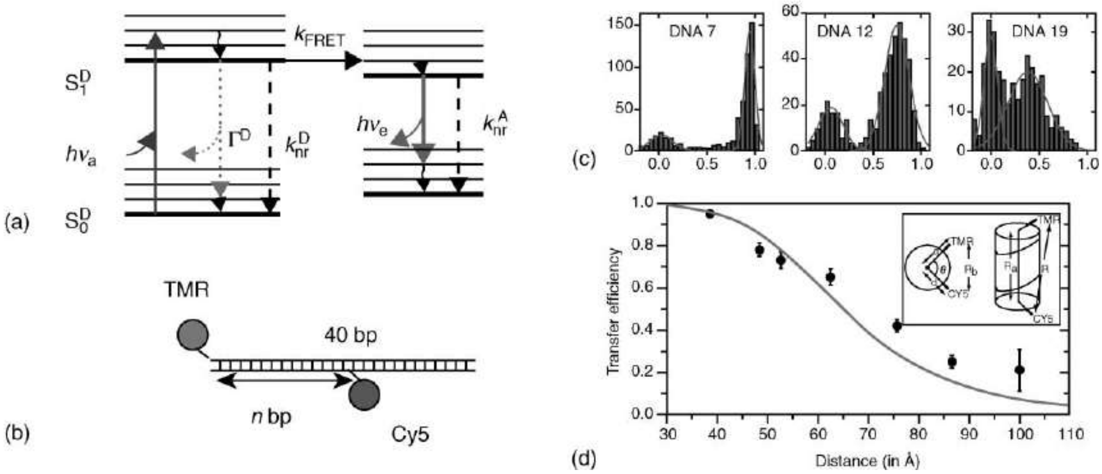

Figure 9 Single pair fluorescence resonance energy transfer. (a) Jablonski diagram for FRET. Fluorescence energy transfer involves two molecules: a donor D and an acceptor A whose absorption spectrum overlaps the emission spectrum of the donor. Excitation of the acceptor to the lowest singlet excited state is a process identical to that previously described for single-molecule fluorescence (**Figure 6a**). In the presence (within a few nm) of a nearby acceptor molecule, donor fluorescence emission is largely prevented by energy transfer to the acceptor by dipole–dipole interaction. The donor exhibits fluorescent emission following the rules outlined in **Figure 6a**. (b) DNA n constructs used for the FRET distance study. Tetramethylrhodamine (TMR) is attached to one end of the DNA and CY5 is attached to the nth base from the end ($n = 7$, 12, 14, 16, 19, 24 and 27). (c) FRET histograms extracted from time traces for DNAs 7, 12, and 19, as shown in **Figure 2b**. Double Gaussian fits extract numbers for the mean (width) of the higher efficiency peak of 0.95 ± 0.05, 0.75 ± 0.13, and 0.38 ± 0.21, respectively. The peak around 0 efficiency corresponds to bleached or inactive acceptor molecules. (d) Mean FRET efficiencies extracted from FRET histograms plotted as a function of distance for DNA 7, 12, 14, 16, 19, 24, and 27. Distances are calculated using the known B-DNA double helix structure. The solid line corresponds to the expected Förster transfer curve with $R_0 = 65$ Å. (b)–(c) Adapted from Deniz AA, Dahan M, Grunwell JR, *et al.* (1999) Single-pair fluorescence resonance energy transfer on freely diffusing, molecules: Observation of Förster distance dependence and subpopulations. *Proceedings of the National Academy of Sciences USA* 96: 3670–3675. Copyright 1999, The National Academy of Sciences.

photochemical degradation in the long-lived triplet state and the interplay of multiphoton ionization processes.

Although Raman scattering cross-sections ($\sim 10^{-30}\,\mathrm{cm}^2$) are orders of magnitude lower than the fluorescence cross-section, they can be enormously amplified if the single molecule is trapped in a 'hot spot' of a metallic nanoparticle, giving rise to surface-enhanced Raman scattering (SERS). Highly resolved vibrational information can be recovered, exhibiting much more stable emission than fluorescence. Moreover, this technique can be applied to nonfluorescent molecules. Similarly to fluorescence, the technique can be extended to nonlinear (two-photon) regimes, as in surface-enhanced hyper-Raman scattering and surface-enhanced anti-Stokes Raman scattering. However, since they require the single molecule to be located on rare sites of a metallic particle, these methods have only been applied to fundamental chemical studies of molecular species.

Finally, extending the range of the FRET 'molecular ruler', electron transfer allows measuring distances between an electron donor and an acceptor within a few ångströms (Å), via opening of a nonradiative channel for the relaxation of the electron in the excited state. Lifetime measurements ($\tau = 1/\Gamma = (k_r + k_{nr})^{-1}$ as defined in Figure 6) yield distance information via the electron transfer rate E_{ET}:

$$E_{ET} = k_0 e^{-\beta r}, \qquad [8]$$

where β^{-1} is of the order of one Å and k_0 of the order of $10^{15}\ \mathrm{s}^{-1}$ are constants characterizing the donor–acceptor pair.

Applications of Single-Molecule Spectroscopy

Single-molecule spectroscopy has already proven to be a useful tool in various domains. The following examples are but a few of an expanding number of applications of SMS.

Molecular Physics

The realization that the fluorescence of a single molecule could be detected with a good SNR over extended periods of time, especially embedded in a crystalline matrix at cryogenic temperature, has led to high-resolution experiments aimed at determining molecular energy levels and transitions. In particular, studies of transitions between sublevels of the triplet state by electron spin resonance (ESR) using superimposed radio-frequency pulses have permitted a precise test of the quantum description of such systems. The dc and ac Stark effect (the shift of a transition energy in the presence of a fixed or alternating electric field) has also been studied in several systems, allowing the symmetry and deformations of the molecule to be determined. The method has been extended to study the photophysics of luminescent polymers (J-aggregates), light-harvesting complexes, green fluorescent proteins (GFP) or semiconductor nanocrystals, and other fundamental systems.

Material Science Studies

Single-molecule photophysical properties are extremely sensitive to their local (nanometer-sized) environment (Figure 1). A precise knowledge of these properties allows using single molecules as sensitive probes of their environment, as illustrated by experiments designed to study the local dynamics of a polymer matrix at the onset of the glass transition (Figure 10).

Molecules of the organic fluorophore Rhodamine 6G (R6G) dispersed within a thin (250 nm) poly(methacrylate) film were observed at temperatures slightly above the melting temperature of the polymer, using a confocal microscope detecting the *s* and *p* polarization of the emitted fluorescence (Figure 8a). Observations are performed under nitrogen atmosphere and at low excitation power (20 nW) in order to prevent oxidation, thus allowing long (several hours) observation time. Each observed R6G molecule exhibits a slow rotational diffusion, sometimes accompanied by lateral or transverse diffusion on a much longer time-scale (Figure 10a). Computation of the autocorrelation function of the emission polarization shows a multi-exponential decay, pointing to a distribution of time-scales of rotational diffusion. A closer look at subwindows of the time trace reveals, however, that the decay comes closer to mono-exponential behavior at shorter time-scales (Figure 10b,c). This study demonstrates that an individual molecule probes an increasing number of different environments over time. At long time-scales, the observable characteristics (such as the autocorrelation function) are similar to those measured on an ensemble of molecules, as expected from the ergodicity hypothesis. At short time-scale, the molecule reveals the local homogeneous and stable characteristic of its nanoenvironment.

Biophysical and Biological Studies

SMS has allowed a reassessment of long-standing questions in biophysics, biochemistry, and biology, by giving scientists the possibility to look at individual building blocks of biological processes: DNA, RNA, proteins and other biomolecules or biomolecular

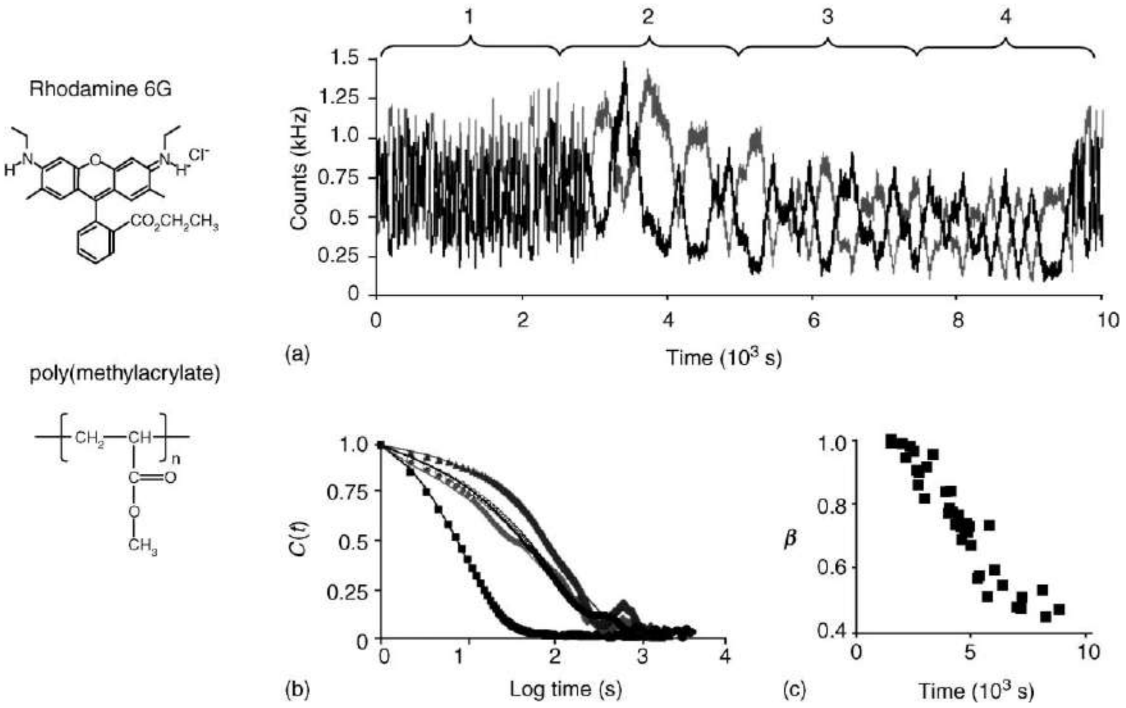

Figure 10 Polymer dynamics near the glass transition. (a) Fluorescence intensity of a single long-lived Rhodamine 6G molecule in poly(methylacrylate) (PMA) at 291 K detected at orthogonal polarizations (gray: *s* and black: *p* polarizations). The melting temperature of PMA is $T_g = 281$ K. The time trace integration is 1 s. This time trace shows the anticorrelation of both polarizations, characteristic of rotational diffusion of the molecule. Different regimes are observed along the time trace (1–4), which could be due to either a modification of the local 'cage' in which the molecule rotates, or escape of the molecule to another nearby nano-environment. The autocorrelation function of polarization computed over the whole trace (not shown) is well fitted by a model adapted to the nonexponential relaxation of polymers, the Kohlrausch–Williams–Watt (KWW) model:

$$\mathrm{KWW}(t) \propto e^{-(t/\tau)^{\beta}}.$$

with $\tau = 88$ s, $\beta = 0.46$. (b) Correlation functions of subsets 1–4 of the full trace. 1 (squares): $\tau = 15$ s, $\beta = 1.0$. 2 (triangles): $\tau = 112$ s, $\beta = 0.77$. 3 (open circles): $\tau = 90$ s, $\beta = 0.59$. 4 (diamonds): $\tau = 64$ s, $\beta = 0.59$. The β exponent is closer to 1, showing that the rotational diffusion is an unhindered one at short time-scales, converging to the bulk exponent at larger time-scale due to ergodicity. This experiment illustrates changes of environment as the observed molecule diffuses and rotates through the polymer mesh, changes that are hidden in ensemble measurements. (c) Values of β as a function of photochemical survival time in PMA at $T_g + 5$ K. Adapted from Deschenes L and Vanden Bout DA (2001) Single-molecule studies of heterogeneous dynamics in polymer melts near the glass transition. *Science* 292: 255–258. Copyright 2001, The American Association for the Advancement of Science.

complexes formed by association of a few of these molecules.

To simplify the task, some biological molecules are intrinsically fluorescent in their active state, as the cholesterol oxidase (COx) enzyme of Brevibacterium sp., an enzyme that catalyzes the oxidation of cholesterol by oxygen. As a result, it is relatively easy to study the dynamics of this enzyme at the single-molecule level using a confocal setup with enzymes immobilized in an agarose gel (Figure 11). The intervals during which the enzyme is active ('on' state) are followed by intervals during which it is inactive ('off' states). The fluorescence recording of a single enzyme thus appears as a 'blinking' time trace, as shown in Figure 11b. Relevant kinetic parameters can be extracted from the distribution of 'on' and 'off' times. Apart from the usual heterogeneity between enzymes of the same batch characterized by different kinetic parameters for the catalysis reaction, this experiment reveals a dynamic heterogeneity, which shows up for each individually studied enzyme as a tendency for short 'on' times to follow short 'on' times, and for long 'on' times to follow long 'on' times (Figure 11c). This variable behavior of a single enzyme is reminiscent of the variable rotational dynamics of Rhodamine 6G molecules in a polymer matrix (see the previous section), and can be studied only thanks to single-molecule techniques.

Not all biological molecules can be studied in such a convenient way using their intrinsic fluorescence. Several labeling schemes are currently available, which allow quantitative, site-specific labeling of proteins, nucleic acid or lipid molecules with few biochemical side effects.

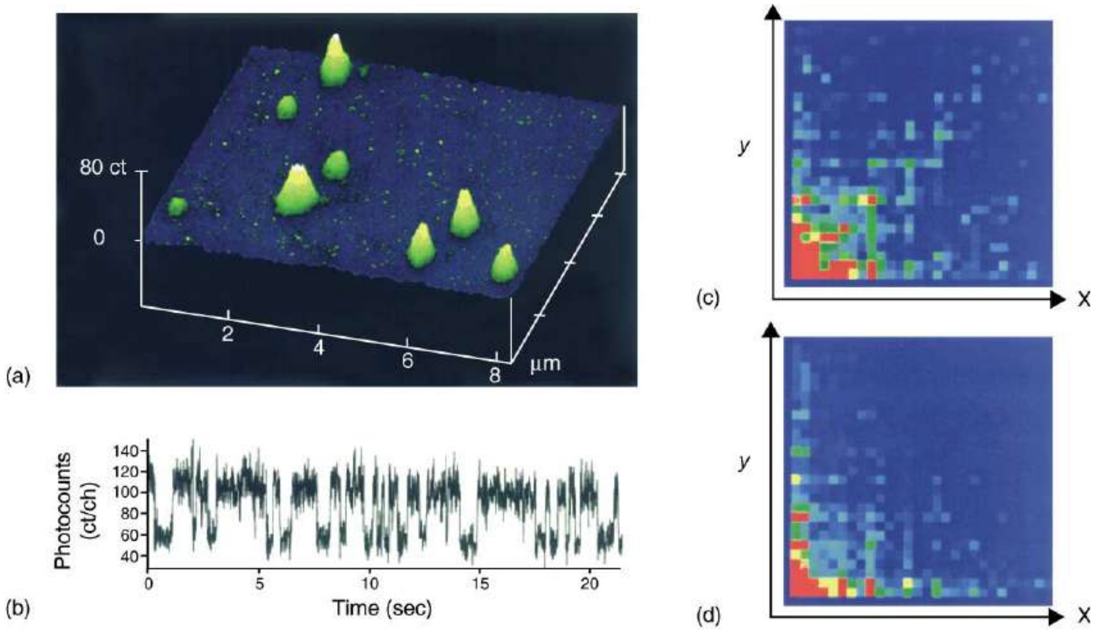

Figure 11 Single-enzyme dynamics. (a) Fluorescence image of single CO_x molecules immobilized in a 10-µm-thick film of agarose gel of 99% pH 7.4 buffer solution (8 µm × 8 µm, 4 min scan, 500 nW excitation at 442 nm). The emission is from the fluorescent active site, FAD, which is tightly bound to the center of CO_x. Each individual peak is attributed to a single CO_x molecule. The intensity variation between the molecules is due to different vertical positions in the excitation volume. (b) Real-time observation of enzymatic turnovers of a single CO_x molecule catalyzing oxidation of cholesterol molecules. This panel shows a portion of an emission intensity trajectory recorded in 13.1 ms per channel. The trajectory was recorded with a cholesterol concentration of 0.2 mM and saturated oxygen concentration of 0.25 mM. The emission exhibits stochastic blinking behavior as the flavin adenine dinucleotide (FAD), the CO_x fluorescent active site toggles between oxidized (fluorescent) and reduced (nonfluorescent) states, each on–off cycle corresponding to an enzymatic turnover. More than 500 on–off cycles are recorded in this trajectory. (c) The 2D conditional probability distribution for on-times (x and y) of two adjacent turnovers as derived from the trajectories of 33 CO_x molecules with 2 mM 5-pregene-3β-20α diol substrate, a derivative of cholesterol. The scales of the x- and y-axes are from 0 to 1 s. A subtle diagonal feature is present, indicative of a memory effect. (d) The 2D conditional histogram for two on-times separated by 10 turnovers for the CO_x molecules in (c). The diagonal feature vanishes because the two on-times become independent of each other at the 10-turnover separation. The color code in (c) and (d) represents the occurrence (z-axis from 350 (red) to 0 (purple). Adapted from Lu HP, Xun L and Xie XS (1998) Single-molecule enzymatic dynamics. *Science* 282: 1877–1882. Copyright 1998, The American Association for the Advancement of Science.

An example of this versatility has been provided by the labeling of the central part of a rotary protein, F_1-ATPase. In this experiment (not shown), the molecular rotor was labeled with a single fluorophore whose orientation, detected by emission polarization measurements, directly reported the angle of the rotor with respect to the shaft. The small size of the fluorophore guaranteed that the protein motion would not be hindered, contrary to previous experiments, which used much larger reporters (micron-size latex beads or fluorescent actin filaments). This experiment reproduced the previous results, showing that the rotor performed 120° steps (recent work hints at substeps of 90° and 30°).

One of the most promising domains where SMS can be of great importance is the study of protein folding. Doubly labeled proteins can be monitored as they undergo conformational changes, taking advantage of the distance dependence of fluorescence energy transfer (demonstrated for a model DNA system in the section on FRET, Figure 9).

Figure 12 shows one of many examples of such a study, performed on diffusing molecules in solution, a configuration that suppresses uncontrolled effects of nearby surfaces. The enzyme chymotrypsin inhibitor 2 (CI2) is a model system for protein folding, believed to have two clearly distinct folding states, which can be controlled by the concentration of denaturant guanidinium chloride. Using single-pair FRET (spFRET) techniques, it is possible to sort out the proportion of folded and unfolded molecules present at different denaturant concentration (Figure 12b). These measurements yield the same titration curve as obtained by ensemble measurements (Figure 12b) and can give access, under some simplifying assumptions, to the energy landscape of the folding reaction. The additional information provided by spFRET is

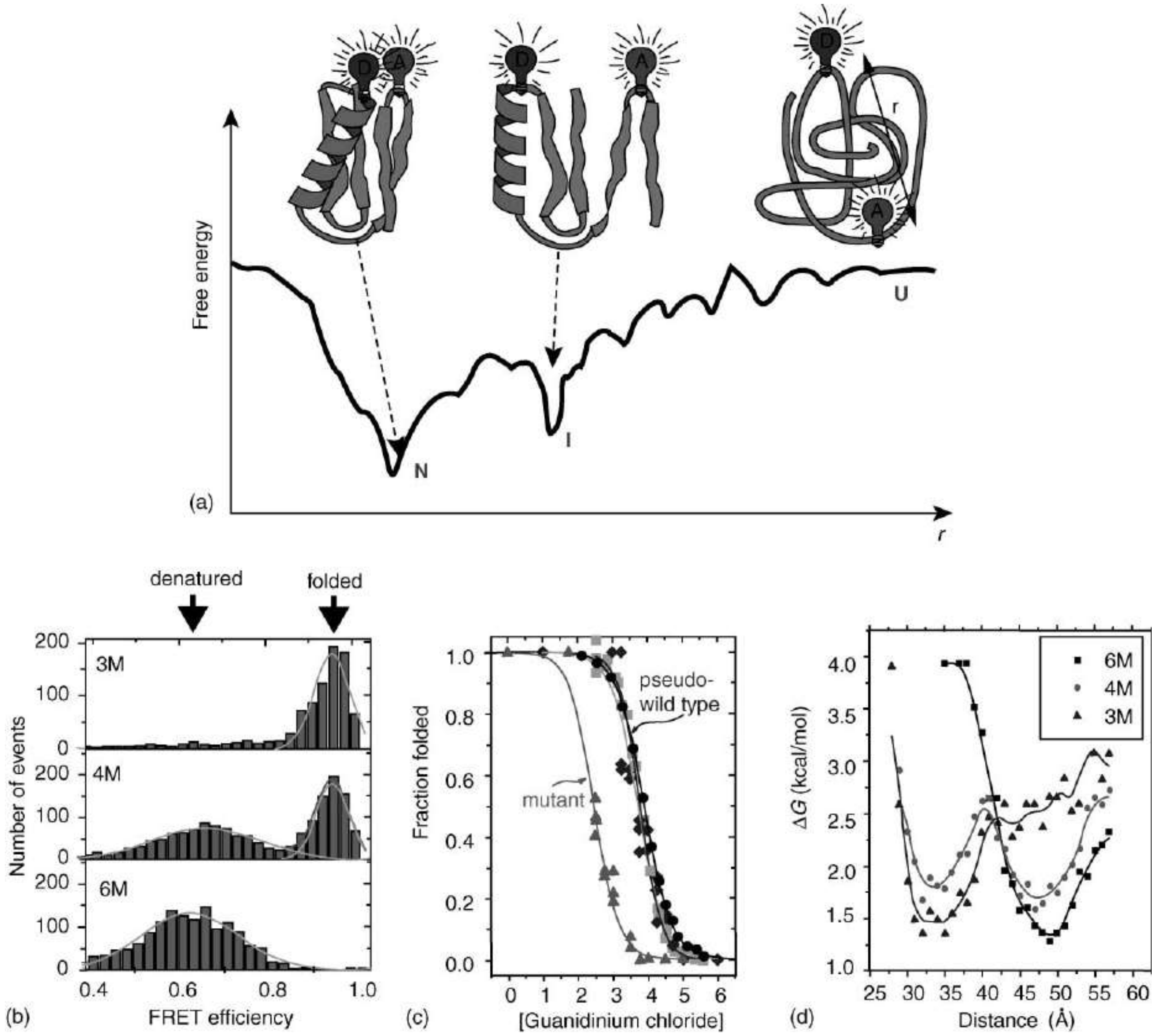

Figure 12 Protein folding. (a) Energy landscape for a protein folding reaction represented as a cartoon of the free energy as a function of distance *r* between donor and acceptor. The rugged energy landscape has a funnel shape with three extrema: U (unfolded state), I (intermediate state) and N (folded or native state). In this labeling scheme, donor and acceptor are close to one another in the native, folded state (N), resulting in a high FRET efficiency. There is reduced energy transfer (larger distance) in the denatured, unfolded state (U). (b) Single-molecule protein folding data. Single-pair FRET (spFRET) histograms of the pseudo-wild type enzyme chymotrypsin inhibitor 2 (PWT CI2) at 3, 4, and 6M denaturant (guanidinium chloride). The FRET efficiency exhibits two peaks, one for folded molecules (high FRET) and one for unfolded molecules (low FRET). Population is transferred from folded to unfolded at higher denaturant concentration, as expected from the protein labeling schematically depicted in (a). (c) Ensemble and single molecule denaturation curves for PWT and K17G mutant CI2. Symbols correspond to data; lines showing sigmoidal fits. PWT CI2: ensemble tryptophan fluorescence experiment (circles), ensemble FRET (squares), spFRET (diamonds); K17G mutant CI2 spFRET (triangles). (d) Free energy functions for PWT CI2 at 3, 4, and 6M denaturant. The solid lines represent a smoothing of the data and are only meant to guide the eye. (a) Adapted from Weiss S (2000) Measuring conformational dynamics of biomolecules by single-molecule fluorescence spectroscopy. *Nature Structural Biology* 7: 724–729. Copyright 2000, Nature Structural Biology. (b)–(d) Adapted from Deniz AA, Laurence TA, Beligere GS, *et al.* (2000) Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proceedings of the National Academy of Sciences USA* 97: 5179–5184. Copyright 2000, The National Academy of Sciences.

the number of molecules in the two different states, providing direct evidence of the two-states model hypothesized from ensemble measurements.

Other biological systems have been addressed using the spFRET approach. Intramolecular spFRET (used in the CI2 study) allows monitoring conformational changes, whereas by labeling two different molecules, intermolecular spFRET permits the detection of association and dissociation events. For instance, spFRET has revealed transient intermediate states in the Tetrahymena ribozyme, which had remained unnoticed in ensemble studies.

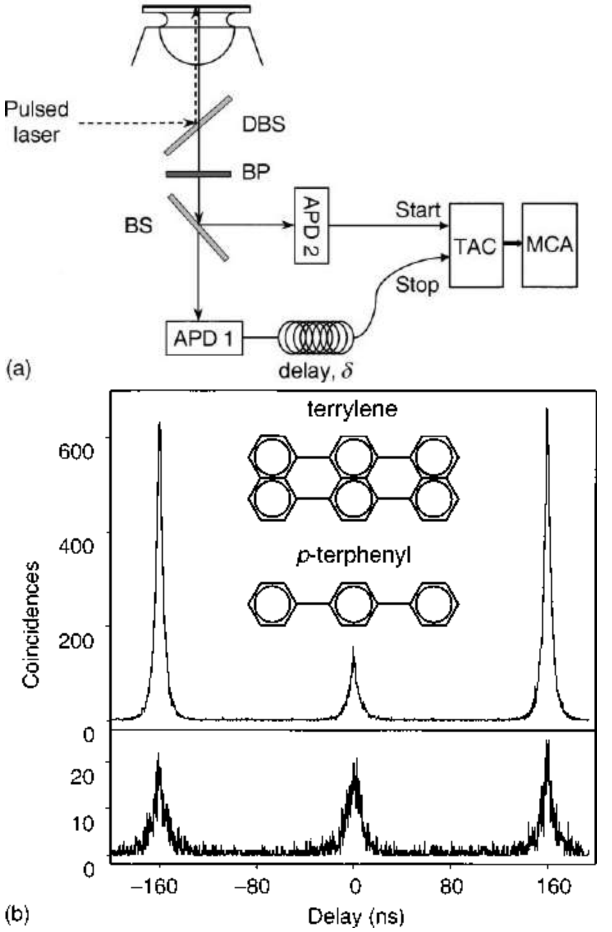

Figure 13 Single-photon source at room temperature. (a) Antibunching measurement setup. The basic elements are similar to those of a lifetime measurement setup (**Figure 7a**) except that the emitted filtered signal (BP: bandpass filter) is split in two with a 50/50 beamsplitter (BS). Two avalanche photodiodes (APD) detect the arrival of photons. One APD pulse is used as the start signal for a time-to-amplitude converter (TAC), which is stopped by the pulse coming from the second APD. Since the TAC response time is finite, it cannot detect simultaneously arriving pulses. A delay time δ is thus added to the stop signal, such that two photons impinging simultaneously on the two APDs will generate pulses with a separation of δ. If the start APD detects a photon from the previous excitation, the measured delay will be $160\,\mathrm{ns} + \delta$, whereas if the opposite happens, the delay will be $\delta - 160\,\mathrm{ns}$. (b) Results of measurements on a terrylene molecule (upper chemical structure) embedded in a p-terphenyl (lower chemical structure) at room temperature. Upper panel: histogram of delay measurements shifted by an amount δ. The histogram corresponds to 120 s of recording on an isolated emitting spot. The residual central peak corresponding to coincidence of photon arrival at APD 1 and 2 is the result of background contamination. Lower panel: histogram acquired over 300 s on a background spot. In both measurements, the laser pulse width was 35 ps, each pulse being separated by 160 ns. The lifetime of terrylene is measured to be 3.8 ns, which sets the width of the observed peaks. Excitation is performed at 532 nm. (b): Adapted with permission from Lounis B and Moerner WE (2000) Single photons on demand from a single molecule at room temperature. *Nature* 407: 491–493. Copyright 2000, Nature Publishing Group.

Quantum Optics

Low-temperature spectroscopic studies of single molecules have turned into a benchmark for testing predictions of quantum mechanics. In particular, one effect expected for a single quantum emitter is the absence of simultaneous emission of two photons: a fluorescent system which is excited at time t_0 has to emit a photon before it can return to the ground state, and be re-excited at a later time t_1. In conditions where the excitation takes place immediately as the electron returns to the ground state, the time difference $t_1 - t_0$ is of the order of the excited state lifetime, of the order of a few nanoseconds. This antibunching phenomenon was first observed in atoms and ions, and rapidly confirmed by low-temperature SMS. This observation is possible either using continuous wave or using pulsed excitation. In fact, nothing prevents this effect from occurring at room temperature, as demonstrated in Figure 13 for pulsed excitation in which excitation photons are provided as well-separated bunches of a few hundred femtoseconds duration.

In this experiment, a system originally used at low temperature (terrylene in a crystal of p-terphenyl) is excited with a repetition rate of 6.25 MHz. The fluorescence emitted by a single molecule is detected in a confocal setup, split in two and directed to two different detectors (Figure 13a). In this way, the time interval between each pair of photons detected by the two detectors can be precisely measured and histogrammed. The result of this time-correlated single-photon counting approach is the absence of coincidence of two photons (Figure 13b), besides the residual counts due to background contamination.

Since the quantum efficiency of this system is high, the experimental setup can be seen as a periodic single-photon source, with an extremely low probability for the simultaneous emission of two photons. This is precisely the kind of device that is needed for quantum cryptography, in which information encoded in the polarization of a beam should not be sensitive to a 'beamsplitter' attack. Future efforts will undoubtedly build on such 'turnstile' photon sources to move forward into optical quantum computing applications.

Perspectives

SMS is a mature discipline, which will certainly prove to be most useful in domains beyond those previously mentioned. Its relative ease of implementation will certainly attract more scientists of other disciplines to this methodology, each of them having specific goals in mind that are difficult to foresee. The principal advantages of SMS outlined at the beginning of this article will remain, but improvements, especially towards simultaneous acquisition of all the fluorescence parameters described earlier are promising. New detectors will permit the combination of the high time resolution of single-photon counting devices with the large field of view and spectral resolution allowed by two-dimensional detectors. SMS will be associated with single molecule manipulation techniques, to correlate applied forces and molecular conformations. New ways of controlling local fields (electric, magnetic, or others) would take advantage of the non-invasiveness, high-temporal and spatial resolution of SMS to get a direct feedback of events at the nanometer scale. Undoubtedly, SMS will be a major tool in the future of nanotechnology.

List of Units and Nomenclature

Å	Ångström
cm^2	square centimeter
fl	femtoliter
K	kelvin
ms	millisecond
MHz	megahertz
μW	microwatt
nm	nanometer

APD	Avalanche photodiode
DNA	Deoxyribonucleic acid
FRET	Fluorescence resonance energy transfer
NA	Numerical aperture
PMMA	Polymethylmethacrylate
SERS	Surface-enhanced Raman scattering
SMS	Single-molecule spectroscopy
SNR	Signal-to-noise ratio
SNE	Signal-to-background ratio
TIR	Total internal reflection
ZPL	Zero phonon line

See also

Scattering: Raman Scattering.

Further Reading

Basché T, Moerner WE, Orrit M and Wild UP (1997) Single-Molecule Optical Detection, Imaging and Spectroscopy. Weinheim: VCH.

Craig DP and Thirunamachandran T (1998) Molecular Quantum Electrondynamics. An Introduction to Radiation Molecule Interactions. Mineola: Dovers.

Forkey JN, Quinlan ME and Goldman YE (2000) Protein structural dynamics by single-molecule fluorescence polarization. *Progress in Biophysics and Molecular Biology* 74: 1–35.

Kneipp K, Katrun H, Kneipp I, *et al.* (1999) Surface-enhanced non-linear Raman scattering at the single-molecule level. *Chemical Physics* 247: 155–162.

Marcus RA and Sutin N (1985) Electron transfers in chemistry and biology. *Biochimica et Biophysica Acta* 811: 265–322.

Mehta AD, Rief M, Spudich JA, Smith DA and Simmons RM (1999) Single-molecule biomechanics with optical method. *Science* 283: 1689–1695.

Moerner WE and Orrit M (1999) Illuminating single molecules in condensed matter. *Science* 283: 1670–1676.

Nie S and Zare RN (1997) Optical detection of single molecules. *Annual Review of Biophysics and Biomolecular Structure* 26: 567–596.

Plakhotnik T, Donley EA and Wild UP (1997) Single-molecule spectroscopy. *Annual Review of Physical Chemistry* 48: 181–212.

Tamarat P, Maali A, Lounis B and Orrit M (2000) Ten years of single-molecule spectroscopy. *Journal of Physical Chemistry A* 104: 1–16.

Weiss S (1999) Fluorescence spectroscopy of single biomolecules. *Science* 283: 1676–1683.

Xie XS and Trautman JK (1988) Optical studies of single molecules at room temperature. *Annual Review of Physical Chemistry* 49: 441–480.

T

TERAHERTZ TECHNOLOGY

Contents

Coherent Terahertz Sources
Terahertz Physics of Semiconductor Heterostructures

Coherent Terahertz Sources

L Wang, Chinese Academy of Sciences, Beijing, China

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Terahertz (THz) radiation is usually referred to as an electromagnetic wave with frequencies ranging from 1 to a few terahertz ($1\,\text{THz} = 10^{12}\,\text{Hz}$). In the electromagnetic spectrum, terahertz radiation lies between the infrared and microwave, as shown in Figure 1. The techniques for generation and detection of THz radiation bridge photonics and electronics in the sense of the concepts and the techniques discussed below.

Physical quantities corresponding to 1 THz are listed as follows;

- Frequency, 1 THz
- Wavelength, 300 μm
- Wavenumber, 33 cm^{-1}
- Energy, 4.1 meV
- Temperature, 48 K

In this spectral region, there exist many rich physical, chemical, and biological phenomena. For example, many phonon resonance and other elementary excitations in condensed matter fall in this region; many molecular vibrations and rotations occur at THz frequency; conformation-related collective vibrational modes in macromolecules, especially in bio-molecules such as proteins and DNA, also lie in THz frequency. However, the development of THz technique lags far behind that of photonics and electronics, due to lack of feasible, reliable, and economic coherent THz sources. The situation has been rapidly changing since the 1990s. Benefiting from advanced material science and the ultrafast laser techniques, various coherent THz sources are being developed and commercialization is soon to follow. We will introduce these coherent THz sources and their main features, and discuss their generation mechanisms. More details can be found in the listed books and articles in the Further Reading section at the end of this article.

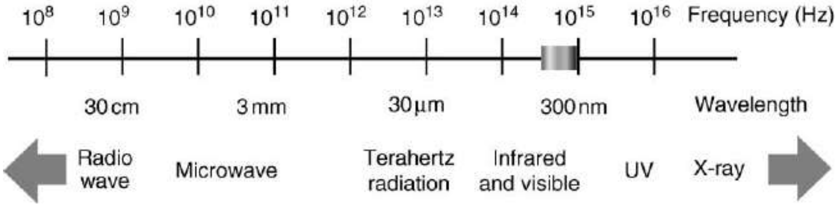

Figure 1 The electromagnetic spectrum with THz radiation sitting between the microwave and the infrared.

Coherent THz Radiation

When a time varying polarization, $\mathbf{P}(\mathbf{r},t)$, is generated in a medium and oscillates at THz frequency, it will radiate a THz wave, $\mathbf{E}(\mathbf{r},t)$, according to Maxwell's equation:

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}, t) + \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \mathbf{E}(\mathbf{r}, t) = -\mu_0 \frac{\partial^2}{\partial t^2} \mathbf{P}(\mathbf{r}, t) \qquad [1]$$

or

$$\nabla \times \nabla \times \mathbf{E}(\omega) - \frac{\omega^2}{c^2} \mathbf{E}(\omega) = \omega^2 \mu_0 \mathbf{P}(\omega) \qquad [2]$$

in frequency domain, for a monochromatic wave or a Fourier component of a wave with a finite bandwidth. The properties of THz radiation are determined by the polarization source $\mathbf{P}(\mathbf{r},t)$. Incoherent THz sources exist in blackbody radiation, and are also detected in astronomy observations. If the phase evolution of the THz wave keeps in step, both spatially and temporally, it is called coherent radiation. A coherent THz source is more powerful and useful in spectroscopy, material characterization, imaging, and many other applications. It is clear that a coherent polarization source is first required for generation of a coherent THz wave. In practice, the working media for THz emitting can be free electrons, quasi-free electrons in solids, or bound electrons and other charge oscillations, such as plasmon oscillation, lattice vibration, etc. Classified by the excitation properties, the emitting process can be of resonance or nonresonance. Coherent polarization sources with different features can be created using various excitation techniques under different principles and device configurations, which lead to different coherent THz sources. Here we focus on coherent THz sources, primarily based on photonic techniques, and working in the spectral range from 1 THz to a few THz.

Gas THz Laser

Molecules have many rotational and vibrational energy levels spaced at THz frequency. For polar molecules, the radiative transition between these levels can radiate THz wave. Following the standard laser technology, molecules pumped either optically or electronically can be used as laser gain media for coherent THz wave generation. For THz gas lasers, the very low THz photon energy poses a problem for lasing, because the coherent radiative transition can be disrupted by thermal depopulation and dephasing. An effective population inversion is difficult to set up and maintain. Therefore, it is important to carefully choose working medium and energy levels, as well as the way to excite the molecules. A number of media has been used for THz laser, such as methanol and HCN vapor, working in both continuous wave (CW) and pulsed modes. For example, methanol lasers operate under excitation of the rocking and asymmetric deformation modes of methanol excited by a CO_2 infrared laser. HCN laser is pumped electrically, which needs a large voltage and current to excite the molecules, a long cavity to obtain sufficient gain, and dedicated control of the temperature in the cavity for stable operation.

In general, THz gas lasers can be tuned discretely in thousands of lasing lines, ranging from a few tens to a few hundreds of micrometers, with power output from μW to mW. THz gas lasers have existed since the early 1970s, and are the only commercialized THz lasers to date. They are still subject to development, for less bulky volume, enhanced tunability and efficiency, and reduced costs.

Free Electron THz Laser

A free electron laser (FEL) uses free electrons as working medium, rather than bound atomic or molecular states in a conventional laser. A typical FEL consists of three parts: (i) an electron source that generates an electron beam with high current; (ii) an accelerator to raise the electron energy; and (iii) a lasing cavity. After the electron beam is generated, electrons are first accelerated to a relativistic velocity, typically with energy of hundreds of MeV, and then enter the cavity consisting of end mirrors and a spatially periodic magnets array called a wiggler (Figure 2). Due to the Lorentzean force imposed by the periodic magnetic field, the high-energy electrons move in the cavity along a sinusoidal path, and emit coherent radiation with a wavelength determined by the spacing between magnets and the electron velocity, as well as the magnetic induction. With the feedback provided by the cavity mirrors, electrons are accelerated or decelerated continuously by the optical field, and are bunched via the resonant interaction. The collective motion of the electron bunches radiates powerful coherent synchrotron radiation. In a standard configuration, the wavelength of the radiation can be expressed as:

$$\lambda = \frac{\Lambda}{2\gamma^2}[1 + (k\Lambda B_0)^2/2] \qquad [3]$$

where Λ is the spatial period of the wiggler magnets, γ the Lorentz factor of the electron beam, B_0 the peak magnetic induction, and k a constant. With a suitable

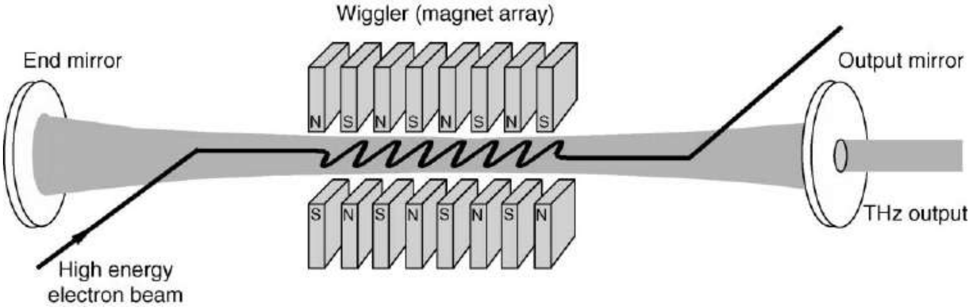

Figure 2 A schematic of THz generation in a free electron laser.

arrangement of the magnet array, as well as other related factors, the facility can work in THz spectral region.

An FEL generates tunable, coherent, high-power THz radiation. It is the most powerful coherent source of THz radiation available. Up to 50 W of average THz power has been generated in the Jefferson Lab recently. The free electron THz laser is becoming an important tools for studying THz radiation and its interactions with condensed matter and biological materials. On the other hand, the FEL is a complicated and expensive facility, and needs dedicated maintenance.

Semiconductor THz Laser

In the twentieth century, semiconductor devices have achieved tremendous success, both in electronic and photonic regime, such as transistors and diode lasers. The semiconductor industry has resulted in revolutionary changes to our world and everyday life. To fill the THz gap between electronics and photonics, great efforts have been made to meet the demand in compact, economic, tunable, and highly efficient THz sources. Encouraging advances have been made since we stepped into the new millennium.

In a semiconductor quantum well, the energy difference between a pair of sublevels can be artificially designed at THz frequency. As early as 1971, the concept of far infrared lasers based on intersub-band transition in semiconductor superlattices was proposed. Although the sublevels can be populated by electrical or optical pumping, the low photon energy at THz frequency causes serious difficulties for lasing, because the two levels sit so close to each other that many thermal processes could destroy the population inversion. Various designs and configurations have been proposed to tackle the problem. An attractive idea is so-called quantum cascade heterostructure. In this kind of device, the electrons are injected into a series of coupled quantum wells electrically. When the electrons are driven by the biased voltage, they can make radiative transition between a pair of sub-bands in a well. Subsequently, electrons enter the next well through resonant tunneling, and so on. However, this kind of device needs the materials of very high quality and suffers severely from fast depopulation of the excited states, even at very low temperatures.

As the advance in material science and unremitting efforts in pursuing the dedicated design of the device structures continued, the real breakthrough came in 2001, when Köhler *et al.* at Pisa demonstrated THz lasing in a quantum cascade heterostructure. Many improvements have been made since then; for example, the idea of using phonon resonance to selectively deplete the population of the lower sub-band has been successful, so that much stabler and robuster population inversion can be set up. The operation temperature has been raised to 136 K, well above the liquid nitrogen temperature, which opens the way for semiconductor THz lasers stepping into many more daily applications. These semiconductor devices are pumped with low voltages and currents at mA levels, and generate narrow bandwidth radiation with mW output at a frequency of a few THz. It is also probable that a four-level lasing system could even working at room temperature.

Another promising device is the p-germanium (Ge) laser developed recently. The Ge laser operates through the electrical excitation of hot holes in p-doped Ge. The laser cavity is formed by polishing the surfaces of the Ge crystal. These lasers could run at 5% duty cycles, with several mW output power. The output wavelength can be continuously tuned from 1 to 4 THz. At the moment, the devices have to operate at 20–30 K and consume about 10–20 W for refrigeration.

It is safe to predict that the feasibility, low cost, and compact semiconductor THz laser will emerge in the near future.

Beside the THz lasers, coherent THz radiation can also be generated by coherently exciting suitable media in other ways. Since the mid-1990s, the development of ultrafast laser techniques has led to two important pulsed coherent THz sources, i.e., photoconductive antenna and optical rectification THz sources. A unique feature of the pulsed coherent sources is broad bandwidth, which is useful in spectroscopic applications. Figure 3 is a typical setup for pulsed THz generation and spectroscopic study, using photoconductive antenna or optical rectification as THz sources.

THz Photoconductive Antenna (PCA)

Different from a conventional antenna working in radio and microwave frequency that are driven directly by electrical current, a photoconductive antenna for THz wave generation works with ultrafast pulsed lasers. When the photoconductive gap is irradiated by an ultrafast laser pulse, photocarriers are generated in the semiconductor. These photocarriers are driven by a DC voltage bias and form a current transient. The antenna couples the electromagnetic field associated with the time varying current into free space, and produces a THz pulse. The THz pulse usually has a duration of a few picoseconds, with a broad frequency bandwidth centered at THz frequency. The output THz power scales with both the optical pulse power and the DC bias field:

$$E_{\mathrm{THz}} \propto \frac{\partial J}{\partial t} \propto \frac{\partial^2 n_{\mathrm{e}}}{\partial t^2} E_{\mathrm{bias}} \qquad [4]$$

The output power, frequency bandwidth, and polarization of the THz radiation depend on the photoconductive features of the semiconductor layer and the geometric structure of the antenna. For an optimal performance, the substrate should have a suitable bandgap for optical interband excitation, a high electron mobility and short carrier lifetime to support an ultrafast current transient, and high dielectric breakdown threshold and to sustain high bias voltage. High-power PCAs have been demonstrated with semi-insulating GaAs, ion-implanted GaAs, and In-GaAs. In the aspects of antenna structures, coplanar strip lines and large aperture emitters are the most effective. Average output power from a good system can reach 30–40 μW.

A broad bandwidth is a special characteristic of a THz source, which is particularly desirable in the spectroscopic study as a coherent probe source. From the property of Fourier transform, we know that the shorter the THz pulse duration, the broader the spectral band. PCAs made of GaAs typically have a useful bandwidth extending from ~100 GHz to 3 THz, which can be extended to 4 THz by tuning the pumping laser wavelength close to the bandedge. Up to 6 THz bandwidths have been reported for PCAs. The pulse duration and, therefore the achievable bandwidth, are ultimately limited by carrier mobility and TO phonon resonant absorption which is around 8.3 THz in GaAs.

In many applications, THz radiation is coupled to free space from the antenna using a closely attached silicon hemispherical lens. This practice increases the system's output and also provides control of the radiation pattern. The radiation pattern for the common dipole antenna is essentially dipolar, with a weak quadrupole component perpendicular to the bias field. In the far-field, the THz beam has a Gaussian cross-section with high-frequency components concentrated in the center.

As the first practical pulsed coherent THz source, PCA has been widely used for many applications in

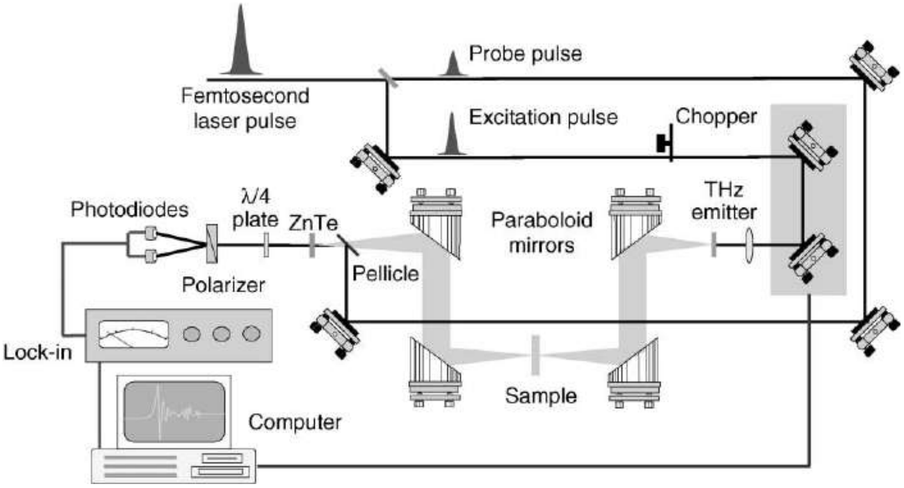

Figure 3 A typical setup for broadband pulsed THz generation and spectroscopic study with a femtosecond laser as excitation source.

scientific research and material characterization. Combined with commercial ultrafast optical fiber lasers, full functional and self-contained THz spectroscopy systems with PCAs as emitters, have been demonstrated and are under commercialization.

Optical Rectification

Optical rectification is a second-order nonlinear optical effect originating from susceptibility $\chi^{(2)}$, and was first observed in 1962 using two intense monochromatic nanosecond laser pulses. In this process, the interaction between laser and nonlinear optical crystal produces a frequency difference polarization:

$$P(0) = \chi^{(2)} E(\omega_1) E(-\omega_1) \qquad [5]$$

When a nonlinear optical crystal is irradiated by an ultrafast laser pulse, the different frequency components in the laser pulse will be coupled with each other by $\chi^{(2)}$, to induce frequency difference polarizations as:

$$P(\Omega) = \chi^{(2)}(\omega_1 - \omega_2 = \Omega) E(\omega_1) E(-\omega_2) \qquad [6]$$

A femtosecond laser pulse usually has a spectral linewidth of at least a few nm, the induced polarization $P(\Omega)$ has a finite frequency distribution centered at THz frequency in the far field, and results in THz radiation $E_{\mathrm{THz}} \propto \partial^2 P(t)/\partial t^2$.

THz radiation in free space by optical rectification was obtained in the early 1990s. One of the differences from a PCA device is that the THz output by optical rectification scales with both the optical pulse power and the DC bias field, while the THz radiation generated by optical rectification solely results from the incident laser. The optimal performance depends on several factors. The generation efficiency first depends on the magnitude of the $\chi^{(2)}$ and phase-matching condition between the optical and THz pulses. The effective magnitude of the $\chi^{(2)}$ coefficient varies with the crystal cut and orientation. Up to now, the most popular material is ZnTe, for its physical durability and excellent phase matching, when a Ti:sapphire femtosecond laser is used as the excitation source. With a moderately focused optical pump beam and a ZnTe crystal, nW T-ray average power can be generated by optical rectification. In general, the generation efficiency may be increased by increasing the laser power, but there are two factors limiting the effect. First, the incident laser power cannot exceed the damage threshold. Second, other second-order nonlinear processes may compete with the optical rectification, such as second-harmonic generation of the pump beam at high optical flux.

For THz radiation generated by optical rectification, the ultimate bandwidth is primarily determined only by the linewidth of the pump laser pulse. Because it is a nonresonance process in most cases, optical rectification can reach broader bandwidths than PCA. Using ultrashort optical pulses and thin nonlinear crystal, THz pulses have been generated with spectra extending to the mid-infrared, well beyond the phonon band of the materials. Like the photoconductive antenna, optical rectification has become a popular technique to generate coherent broadband THz radiation in various applications. Beside the conventional bulk inorganic crystals, many different media have been used for THz generation, such as organic crystal DAST, biased quantum wells, periodically poled LiNbO3 (PPLN), poled polymers, super-conducting thin films, etc.

Other Coherent THz Sources

When two light sources with slightly different wavelengths interact together with a nonlinear optical crystal, a beating polarization will be generated and will radiate a THz wave. It is called difference frequency generation, or three-wave mixing. Optical parametric processing is another way to generate a THz wave. In this process, a near-infrared pump beam generates a second NIR idler beam in a nonlinear crystal, and THz radiation can be generated from the beating of the pump and the idler. The merit of this technique is the continuous frequency tunability of the THz output, which is valuable in spectroscopic applications, and relatively cheaper nanosecond lasers can be used. If the nonlinear crystal is further placed in a cavity and the idler beam is amplified by feedback from end mirrors, a THz optical parametric oscillator is formed. A number of improvements have been made since the mid-1990s with this kind of device. More recently, THz parametric generation with an injection seeded idler beam has shown a reduced linewidth ($\Delta v/v \approx 10^{-4}$) and a peak THz power of over 100 mW for 3.4 ns pulses. A narrow linewidth combined with reasonable tunability make this kind of THz source attractive in spectroscopic studies.

Besides using photonic techniques, coherent THz radiation can also be generated by electronic techniques through increasing the output frequency of the microwave devices. For example, an electrically driven microwave generator, backward wave oscillator, can generate CW THz radiation up to 2 THz. A backward wave oscillator, running under optimal conditions, can provide up to 300 mW of polarized, narrowband radiation with a limited tunability about 30% of its central frequency.

See also

Imaging: Infrared Imaging. **Lasers:** Free Electron Lasers. **Optical Parametric Devices:** Overview. **Semiconductor Physics:** Infrared Lattice Properties. **Terahertz Technology:** Terahertz Physics of Semiconductor Heterostructures.

Further Reading

Grüner G (1998) *Millimeter and Submillimeter Wave Spectroscopy of Solids*, *Vol. 74 of Topics in Applied Physics*. Berlin: Springer-Verlag.

Kaiser W (1988) *Ultrashort Laser Pulses and Applications*, *Vol. 60 of Topics in Applied Physics*. Berlin: Springer-Verlag.

Khurgin JB (1994) Optical rectification and terahertz emission in semiconductors excited above the bandgap. *Journal of the Optical Society of America B* 11: 2492–2501.

Rullière C (1998) *Femtosecond Laser Pulses*. Berlin: Springer-Verlag.

Saldin EL, Schneidmiller EA and Yurkov MV (2000) *The Physics of Free Electron Lasers*. *Advanced Texts in Physics*. Berlin: Springer-Verlag.

Siegel PH (2002) Terahertz technology. *IEEE Transactions on Microwave Theory and Techniques* 50: 910–928.

Tredicucci A, Gmachl C, Capasso F, *et al.* (2001) Novel quantum cascade devices for long wavelength IR emission. *Optics Materials* 17: 211–217.

Tsen KT (2001) *Ultrafast Physical Processes in Semiconductors*, *Vol. 67 of Semiconductors and Semimetals*. New York: Academic Press.

Terahertz Physics of Semiconductor Heterostructures

R Bratschitsch and K Unterrainer, Technische Universität Wien, Vienna, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The terahertz (THz) regime of the electromagnetic spectrum holds promise for a large number of important applications in areas like environmental sensing, quality control, medicine, military applications, astronomy, etc. The development of THz devices has been mainly hindered by the lack of compact sources. In this article we overview the types of tabletop semiconductor sources that are available right now. We focus on the use of pulsed THz sources that are optically excited by femtosecond laser pulses. The generated bursts of THz radiation are extremely useful for time-resolved THz spectroscopy of semiconductor heterostructures.

THz Generation from Semiconductors

The first freely propagating THz pulses using femtosecond lasers were generated in 1984 when Auston *et al.* fabricated a Hertzian dipole antenna. It mainly consists of an electrically biased transmission line structure mounted on a semiconductor, i.e. a photoconductive switch that is triggered by a femtosecond laser pulse (Figure 1). The generated transient photocurrent $J(t)$ emits radiation according to $E(t) \propto \partial J(t)/\partial t$. Since 1989, sapphire half-spheres or refractive silicon lenses have been attached to these dipole antennas, which made it possible to produce collimated THz beams for use in everyday spectroscopic experiments. Today generation and detection of THz radiation by semiconductor photoconductive antennas is widely used. However, most of the antennas are limited to the low-frequency regime of the THz spectrum since they usually have a high-frequency cutoff of about 1–2 THz.

THz generation from semiconductor surfaces was introduced by Zhang in 1990. There are three different physical mechanisms that are responsible for the surface emission of THz radiation. One is the transient current, i.e., the acceleration of photogenerated carriers in the semiconductor surface electric field. This process is accompanied by field induced optical rectification which is often called instantaneous polarization, where the electron–hole pairs are polarized in the dc electric field. Finally, bulk optical rectification as a source of THz radiation can be observed in noncentrosymmetric crystals.

The surface emission intensity can be enhanced by using an externally applied electric field instead of the intrinsic electric field. Until recently, the use of laser amplifiers that optically pump these so-called large-aperture antennas has led to the brightest ultrafast tabletop THz emitters.

Recently, Sarukura and co-workers found that the application of a magnetic field dramatically increases the THz emission from bulk InAs. This led to a re-examination of Zhang's work, in which he reported the magnetic field enhancement in bulk GaAs,

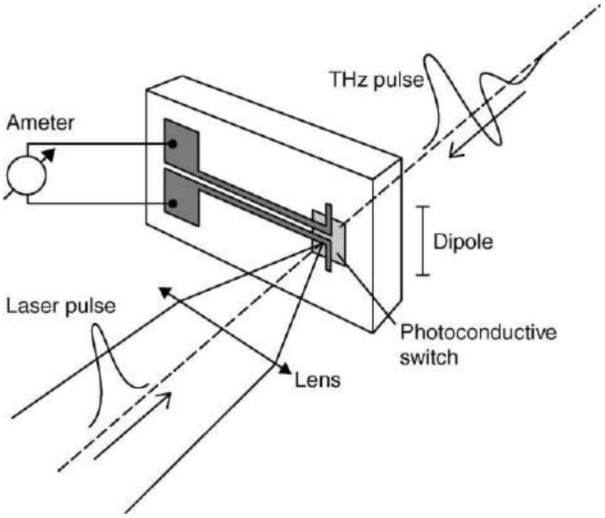

Figure 1 Photoconducting dipole antenna. Reproduced with permission from Gornik E and Kersting R (2001) Coherent THz emission in semiconductors. In: Tsen KT (ed.) *Ultrafast Physical Processes in Semiconductors*, vol. 25, pp. 396. New York: Academic Press.

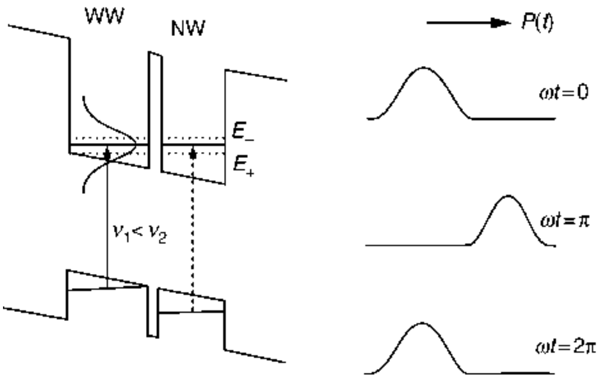

Figure 2 Schematic band diagram of an asymmetric double quantum well structure. A short laser pulse with a spectral width larger than the intersub-band splitting can excite an electronic wavepacket which tunnels back and forth between both wells. Reproduced with permission from Roskos H, Nuss M, Shah J, Leo K, Miller D, Fox A, Schmitt-Rink S and Köhler K (1992) Coherent submillimeter-wave emission from charge oscillations in a double-well potential. *Physical Review Letters* 68: 2216. Copyright (1992) by the American Physical Society.

and to a large number of studies of different types of other semiconductors in various magnetic field geometries. Currently, there is no unified model of the enhancement mechanism(s).

The use of a semiconductor heterostructure by Roskos *et al.* in 1992 (a coupled double quantum well) marked the advent of a new class of semiconductor THz emitters that can be deliberately designed using bandgap engineering. In these devices the quantum nature of electrons in coupled potential wells is used to generate electron wavepackets that tunnel back and forth between different potential wells (Figure 2). The coherent charge oscillation is triggered by a femtosecond laser pulse that has a spectral width that is larger than the separation between the quantized states in the well. In this way a coherent superposition of these levels is created. THz emission from these quantum beats has also been observed from charge oscillations due to light and heavy hole excitons in a quantum well by Planken and co-workers. The concept of quantum beats can even be transferred to the high-THz regime (tens of THz) which was successfully demonstrated by Bonvalet and co-workers. By using an InAs step quantum well they generated THz radiation at a frequency of 30 THz.

Bloch oscillations, one of the fundamental concepts of solid state physics, were predicted in 1934 by Zener. He stated that electrons in a periodic potential under an applied electric field should perform oscillations. Using a quantum mechanical picture, Bloch oscillations can be described as a quantum beat of Wannier–Stark states and in 1993 Waschke *et al.* detected THz emission from these oscillations.

These heterostructure THz sources are inherently frequency tunable since the energy level spacing can be altered by an applied electric field that tilts the quantum wells. The Bloch frequency is given by:

$$\omega_{\mathrm{Bloch}} = \frac{eFd}{\hbar} \qquad [1]$$

where F is the applied electric field and d is the superlattice period. Unfortunately, these devices suffer from their low emission power.

An alternative concept is the use of collective excitations in semiconductor bulk material or heterostructures. Dekorsy and co-workers showed in 1995 that coherent phonons emit THz radiation when excited by femtosecond laser pulses. In the same year Sha and co-workers studied the absorption changes within an electrically biased GaAs p-i-n structure and found distinct oscillations which they ascribed to another collective phenomenon, namely coherent plasma oscillations of the photogenerated carriers (Figure 3). This explanation was supported by the square-root dependence of the observed frequencies on the photogenerated carrier density which is predicted by a simple oscillator model:

$$\omega_{\mathrm{p}} = \sqrt{\frac{ne^2}{\varepsilon\varepsilon_0 m^*}} \qquad [2]$$

where n is the carrier density, ε the dielectric constant, and m^* the effective mass of the carriers. Fischler and co-workers confirmed this finding and in particular investigated the damping behavior of the plasmons. In 1997, Kersting *et al.* detected THz radiation from

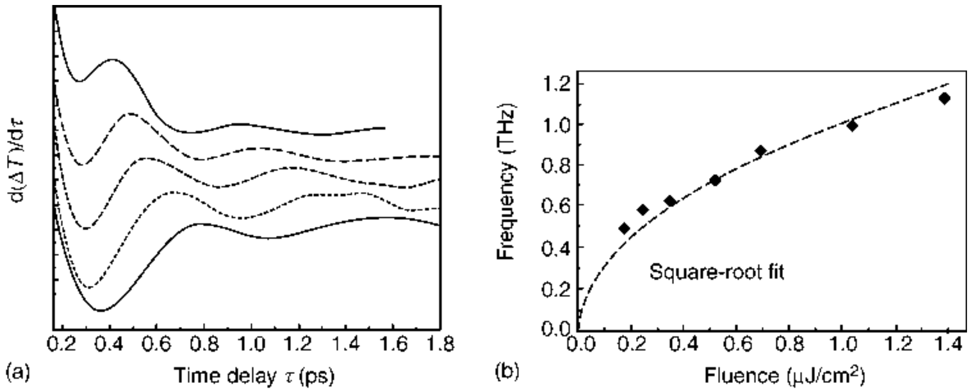

Figure 3 (a) Time-derivative of the charge in electro-absorption versus pump-probe time delay for decreasing fluences (top to bottom respectively); (b) the frequency of the oscillations as a function of fluence. The dashed curve is a square-root function fit. Reproduced with permission from Sha W, Smirl A and Tseng W (1995) Coherent plasma oscillations in bulk semiconductors. *Physical Review Letters* 74: 4273. Copyright (1995) by the American Physical Society.

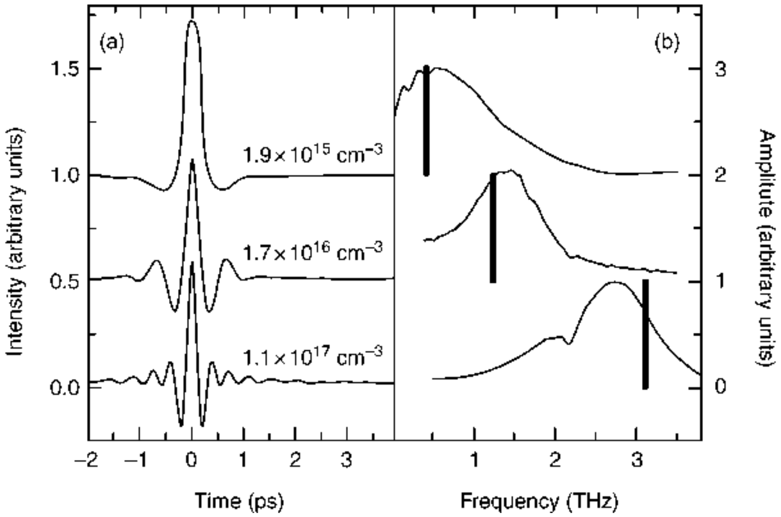

Figure 4 (a) THz autocorrelation traces for three different doping densities; (b) THz spectra; lines denote the frequencies calculated by formula [2]. Reproduced with permission from Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmas in n-doped GaAs. *Physical Review B* 58: 4553. Copyright (1998) by the American Physical Society.

these p-i-n structures and identified heating of the photogenerated carriers as the origin for the strong damping of the plasma oscillations resulting in a low emission intensity.

This problem can be circumvented by using coherent plasma oscillations of extrinsic charge carriers. These electrons are cold compared to the photo-excited carriers and should undergo less scattering. Kersting *et al.* showed that THz pulses are indeed emitted from coherent three-dimensional (3D) plasma oscillations in n-doped GaAs epilayers. The THz pulses are temporally and spatially coherent with intensities up to 100 nW.

Figure 4 shows the THz emission from three GaAs epilayer samples which have been doped with densities 1.9×10^{15}, 1.7×10^{16}, and $1.1 \times 10^{17}\,\mathrm{cm}^{-3}$. The curves are recorded using an autocorrelation technique: 13 nJ, 100 fs laser pulses at a repetition rate of 80 MHz and a central wavelength of 800 nm are split into two pulses with a Michelson interferometer. Every visible pulse generates a THz pulse in the transmission or pseudo-reflection direction. After collimating and focusing of the THz radiation by parabolic mirrors, the interference of the two THz pulses can be recorded by changing the time delay between the visible pulses (Figure 5). Assuming negligible nonlinear interactions between the two excitation processes, the linear autocorrelation function can be recorded. The spectrum is calculated by simply Fourier transforming (Figure 4b). The experiments are performed at room temperature, and the whole setup is purged with dry nitrogen gas to prevent

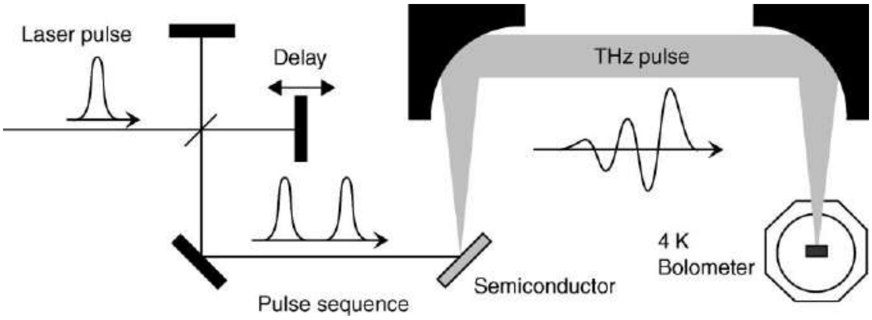

Figure 5 THz autocorrelation setup. Reproduced with permission from Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmons in n-doped GaAs. *Physical Reviews B* 58: 4553. Copyright (1998) by the American Physical Society.

water vapor absorption. Although the autocorrelation traces are recorded at the same optical excitation density of $n_{ex} = 2 \times 10^{16}\,\text{cm}^{-3}$, the emission frequency changes significantly. In addition, the autocorrelation trace of the highly doped structure ($1.1 \times 10^{17}\,\text{cm}^{-3}$) shows multiple oscillations while in the case of the sample with the lowest doping ($1.9 \times 10^{15}\,\text{cm}^{-3}$) the slow oscillation is damped out rather quickly. The frequency behavior of the three different samples can be analyzed from the spectra shown in Figure 4b. The bars indicate the frequency calculated with [2] and setting n equal to the doping density determined by standard Hall measurements. Obviously, the calculated frequencies match very well the experimentally determined ones. This indicates that the THz emission results from a collective oscillation of the extrinsic electrons within the epilayer. This is also supported by the observation that a variation of the excitation density of the photogenerated carriers over two orders of magnitude does not change the THz emission frequency. This is in clear contrast to the experiments with undoped semiconductors where the emission frequency shows a square-root dependence on the optically excited carrier density.

Excitation of the coherent plasma oscillation results from the ultrafast screening of the surface electric field by the femtosecond laser pulse. Before excitation, the doping electrons are confined between the surface electric field and the undoped GaAs substrate (Figure 6). A short laser pulse with an energy above the bandgap of the semiconductor generates electron–hole pairs that screen this surface electric field for a short time. The doping electrons inside the material respond to the change in potential at the surface of the sample leading to the oscillation behavior. A detailed model calculation of the excitation mechanism can be found in Kersting *et al.* (1998). The oscillation frequency – the eigenfrequency of the electron system – is simply the plasma frequency given by the simple formula [2].

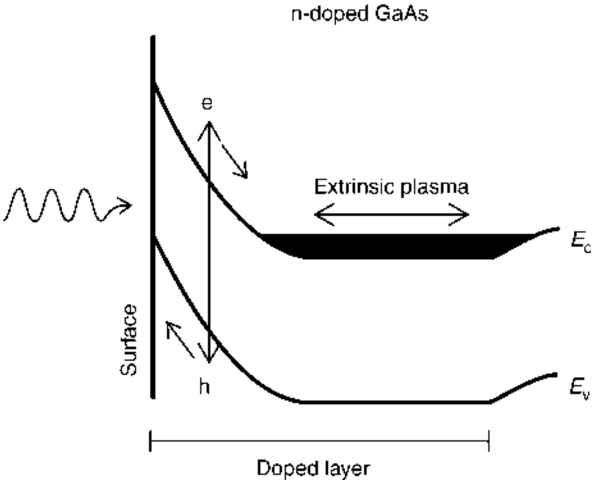

Figure 6 Schematic drawing of the 3D plasmon excitation. After Kersting R, Bratschitsch R, Heyman JN and Unterrainer K (1999) Unpublished.

There are two mechanisms that are responsible for the damping of the plasma oscillations. If the doping density is lower than $10^{16}\,\text{cm}^{-3}$ the coherence time of the 3D plasmon is mainly limited by optical phonon scattering. This can be verified by measuring the plasmon damping times at different temperatures and comparing the results with mobility data which show the same tendency. Since optical phonon scattering is suppressed at low temperatures, the THz emission intensity can be substantially increased by cooling the sample. In the high-doping-density regime $>10^{16}\,\text{cm}^{-3}$ the observed damping times are shorter than those expected due to the optical phonon scattering mechanism alone. This is because in this regime impurity scattering is the dominant scattering mechanism.

One way around this problem is the use of a remotely doped structure. In this case the scatterers are not in the same region as the electrons that perform the plasma oscillation (Figure 7). The higher mobility should lead to a longer decoherence time.

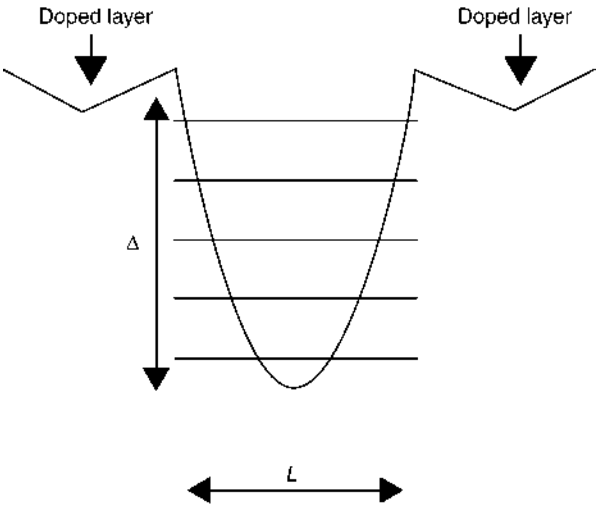

Figure 7 Schematic drawing of a modulation-doped PQW. The doped layers are outside the quantum well.

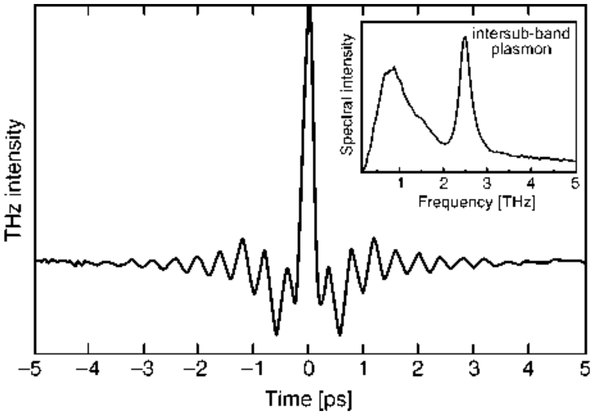

Figure 8 THz autocorrelation signal of a 1400 Å PQW excited by 780 nm laser pulses ($T = 5$ K). Inset: Fourier transform of the recorded autocorrelation. Reproduced with permission from Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent THz emission from optically pumped intersubband plasmons in parabolic quantum wells. In: *Ultrafast phenomena XII*. Springer series in Chemical Physics, vol. 66, pp. 204. New York: Springer-Verlag.

For this purpose Bratschitsch *et al.* used modulation doped GaAs/AlGaAs parabolic quantum wells (PQWs), with widths L in the range of 1200–2000 Å and carrier sheet densities n_s of 1.7×10^{11}–5×10^{11} cm^{-2}. The electrons confined within the parabolic potential form a so-called quasi-3D electron system. Figure 8 shows the THz autocorrelation trace of a modulation doped PQW ($L = 1400$ Å, $n_s = 5 \times 10^{11}$ cm^{-2}) excited by 780 nm laser pulses. The density of the optically generated carriers is kept well below the carrier density due to the modulation doping inside the PQW.

The spectrum of the emitted THz radiation (inset of Figure 8) consists of two components: a broad background and a narrow line (FWHM: 0.3 THz) with a center frequency of 2.55 THz. The low-frequency broadband component is found for all the different PQW samples and is independent of the PQW sample structure. It can also be observed at room temperature while the narrowband component completely vanishes. The origin of the broadband component is due to THz generation at the surface of the sample.

The THz emission at 2.55 THz can be observed within a wide range of excitation wavelengths (815–760 nm) above the 819 nm GaAs bandgap ($T = 5$ K). When varying the excitation density over one order of magnitude ($5.6 \times 10^{10} - 6.2 \times 10^{11}$ cm^{-2}), the emission frequency does not change. The narrowband emission results from the oscillation of the carriers inside the PQW. The observed frequency corresponds to the intersub-band plasmon of the PQW. The expected frequency of the intersub-band plasmon is given by

$$\omega_0 = \sqrt{\frac{8\Delta}{L^2 m^*}} \qquad [3]$$

where Δ is the energetic depth of the well, L is the well width, and m^* the effective mass. For the 1400 Å wide PQW the formula yields 2.2 THz for the intersubband plasmon frequency.

The excitation mechanism for the intersub-band plasmon is due to screening of the surface depletion field by the electron–hole pairs injected by the ultrafast laser pulse. In this way the electrons inside the quantum well experience a kick and begin to oscillate with their eigenfrequency (Figure 9). This is supported by the fact that the oscillation can be excited over a large wavelength range of the femtosecond pulses. The onset of intersub-band plasmon emission appears at 815 nm near the 819 nm GaAs bandgap ($T = 5$ K) and can be observed to up to 760 nm. This large range implies that the excitation mechanism is clearly a nonresonant phenomenon, in contrast to the THz quantum beat experiments. Due to ultrafast field screening, a symmetric nanostructure can emit optically driven THz radiation.

The width of the intersub-band plasmon emission line is 0.3 THz (full width at half maximum; FWHM) for the 1400 Å PQW and even 0.2 THz for the 2000 Å PQW sample. The linewidth is a factor of 5 smaller than the smallest linewidth measured for a 3D plasmon emitter. The width of the intersub-band emission is nearly unaffected when increasing the excitation wavelength from near bandgap to approximately 60 nm above bandgap. However, it shows a linear dependence ($5.6 \times 10^{10} - 6.2 \times 10^{11}$ cm^{-2}) on

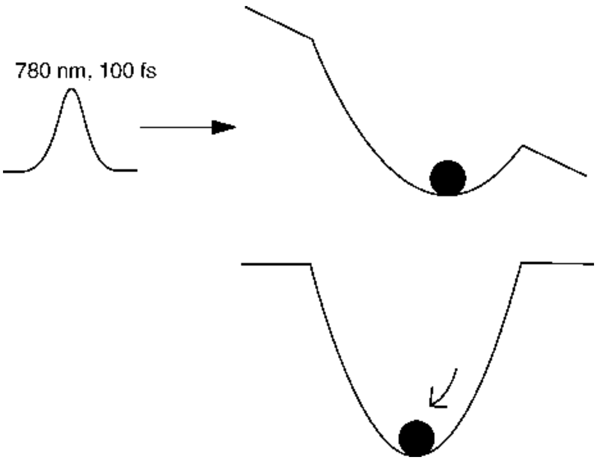

Figure 9 Symbolic drawing of the excitation mechanism: an oscillation of the carriers in the parabolic potential is initiated by ultrafast field screening. Reproduced from Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent terahertz emission from optically pumped intersubband plasmons in parabolic quantum wells. *Applied Physics Letters* 76: 3501, with permission from the American Institute of Physics.

the number of optically excited carriers. These optically injected carriers are added to those from the modulation doping and cause a faster dephasing of the intersubband plasmon oscillation. This linear dependence due to carrier–carrier scattering has also been observed in Wolter's experiments dealing with the dephasing of Bloch oscillations in semiconductor superlattices. The linewidth stays constant up to about 50 K (0.33 THz = 1.36 meV) and then starts to broaden until the sample temperature reaches 150 K where the PQW emission is buried by the surface-generated THz emission. The broadening of the line between $T = 50$ K and $T = 150$ K can be explained by the onset of optical phonon scattering in addition to interface roughness scattering which is the dominant scattering mechanism at low temperatures. The onset and linear increase of the line broadening has also been observed with Fourier Transform Infrared (FTIR) absorption and THz time-domain cross-correlation spectroscopy measurements by Kersting and co-workers.

The combination of the designability of the transition frequency, the narrowband emission, and the absence of any processing of the sample make modulation-doped PQWs attractive and easy-to-use THz emitters.

Other THz emission experiments of heterostructures include the optically pumped emission of THz radiation from grating-coupled intrasub-band (2D) plasmons in a doped single quantum well observed by Sekine and co-workers or optically excited THz emission from coherent plasmons in doped superlattices detected by Bratschitsch and co-workers.

Few-Cycle THz Spectroscopy of Semiconductor Quantum Structures

There are a large number of studies that use few-cycle THz radiation for the investigation of semiconductor bulk materials. The complex index of refraction in the THz regime gives insight into the conductivity (i.e., carrier mobility) of the material. The free carriers that change the index of refraction by Drude absorption can be provided either by doping (n-doped silicon or GaAs) or by photo-excitation. An optical pump-THz probe experiment with undoped GaAs reveals the different intervalley scattering times.

There are compelling reasons to use THz radiation to excite and probe carriers in semiconductor quantum structures. The photon energies are comparable to the sub-band spacings, carrier kinetic energies, and phonon energies. In addition, THz radiation does not generate minority carriers so that the experiments directly probe free carriers rather than excitons.

Heyman *et al.* investigated a modulation-doped GaAs/AlGaAs multiple quantum well ($d = 510$ Å, 10×) with a transition energy between the two lowest sub-bands of 1.5 THz and a carrier density of $n_s = 2.75 \times 10^{10}\ \mathrm{cm}^{-2}$. On top of the sample an aluminum Schottky gate is evaporated and the well is contacted with AuGe alloyed ohmic contacts. The wells can be depleted of carriers by applying a negative gate voltage to the Schottky gate (−10 V gate bias). By modulating the voltage between 0 V and −10 V the response of the 2D electron system can be measured. Otherwise, the recorded signal is dominated by the contribution of the bulk GaAs substrate that dramatically chirps the THz pulse due to its frequency-dependent index of refraction. The THz pulses are coupled into the cleaved edge of the quantum well sample so that the electric field is parallel to the growth direction in order to excite the intersub-band transition. The electrons' response is measured with a cross-correlation technique. Here the THz radiation transmitted through the quantum well is mixed with a short THz pulse emitted from low-temperature grown GaAs to get the desired time resolution in the experiment. This allows the detection of the transmitted signal in amplitude and phase. Figure 10 shows the cross-correlation signal of the QW. The signal from the carriers rises during the first 2 ps in response to the THz field and then continues to oscillate at a constant frequency of 1.5 THz. Due to the femtosecond time resolution in this experiment it is possible to observe the phase of the electrons even during the process of excitation. Initially, the incident THz pulse generates a coherent superposition of states in the first and second sub-band. After the THz driving pulse is over

(>2 ps) the carriers continue to radiate at the intersub-band transition frequency, and the signal is damped out by the free induction decay. Since the amplitude and phase of the radiated electric field is recorded, not only the absorption due to the QW electrons but also the phase shift associated with the absorption can be extracted (Figure 11).

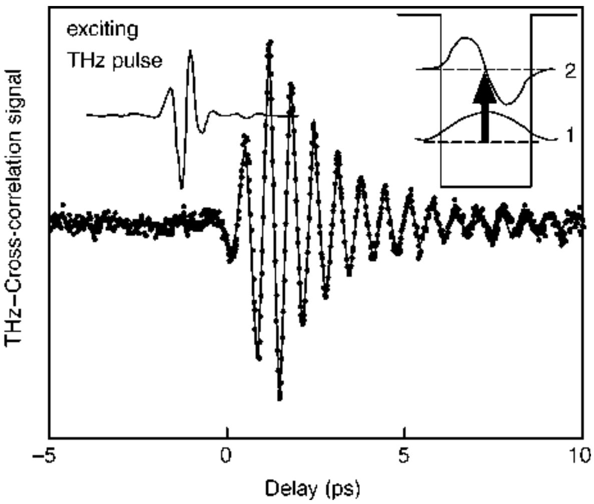

Figure 10 Cross-correlation signal obtained by electrically modulating the charge density in a 510 Å wide quantum well. The THz pulse generates a coherent superposition of the $n = 1$ and $n = 2$ sub-band states. The resulting oscillating polarization radiates and is damped out by the free induction decay. The upper trace shows the incident THz pulse. Reprinted from Unterrainer K, Kersting R, Bratschitsch R, *et al.* (2001) Few-cycle THz spectroscopy of semiconductor quantum structures. *Physica E* 9: 76, with permission from Elsevier.

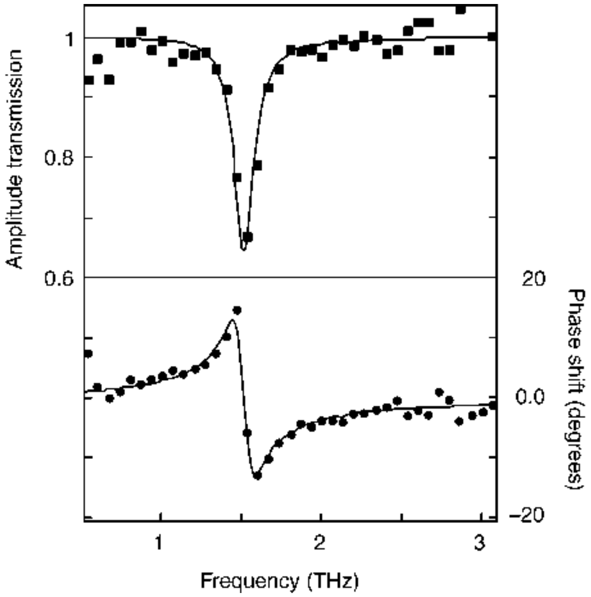

Figure 11 Change in the absorption coefficient and index of refraction in the quantum well due to the quantum well electrons. Solid points are calculated from the time-domain data. Reproduced from Heyman J, Kersting R and Unterrainer K (1998) Time-domain measurement of intersubband oscillations in a quantum well. *Applied Physics Letters* 72: 644, with permission from the American Institute of Physics.

This experiment showed the response of the QW electrons to radiation which is resonant to the intersub-band transition. Using the same experimental technique Kersting *et al.* investigated the response of electrons to a non-resonant excitation of the intersub-band transition. Instead of a square QW, a modulation-doped parabolic quantum well is used in this experiment which has the advantage that the transition frequency is independent of the carrier concentration (Kohn's theorem) and the applied electric field. Furthermore the THz radiation is coupled to the intersub-band transition via a metallic Schottky grating. Figure 12 shows the response of the QW electrons to resonant THz radiation. The electrons follow the driving pulse and continue to radiate at the intersub-band transition frequency. The response to a nonresonant THz pulse is shown in Figure 13. Initially, the electrons inside the QW follow the driving pulse but when the driving pulse has ended a phase jump occurs and the electrons lock to their eigenfrequency and slowly lose their coherence. This behavior can be understood by using time-dependent perturbation theory to solve Schrödinger's

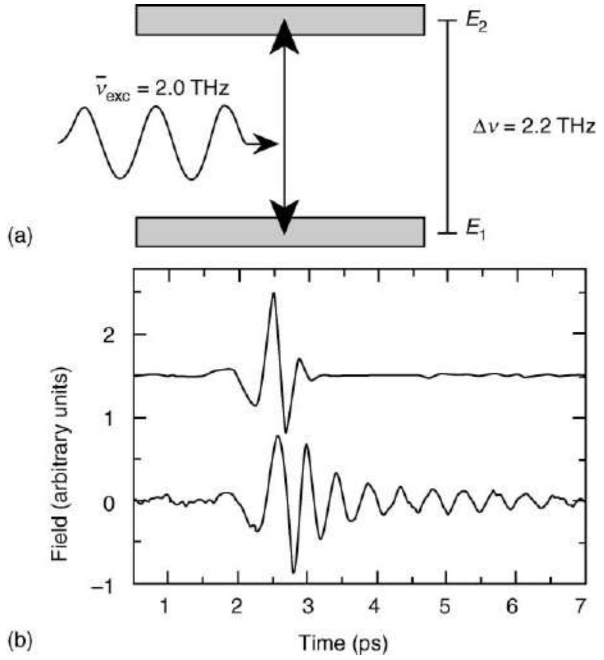

Figure 12 Resonant excitation: (a) band diagram of an intersub-band transition of 2.2 THz and driving THz field at 2.0 THz; (b) exciting THz pulse (upper curve) and cross-correlation signal (lower curve). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

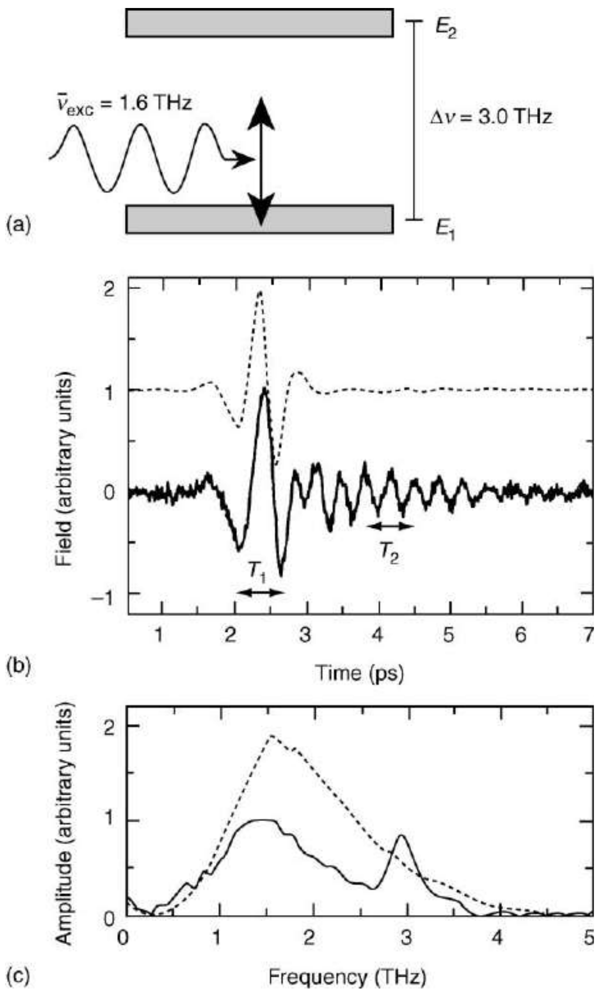

Figure 13 Nonresonant excitation: (a) band diagram of an intersub-band transition of 3.0 THz and driving THz field at 1.6 THz; the oscillation frequency changes from T_1 to T_2; (b) exciting THz pulse (upper curve) and cross-correlation signal (lower curve); (c) spectra of the exciting THz pulse (dotted curve) and cross-correlation signal (solid curve). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

equation for a two-level system with an electromagnetic driving field. While the two-level system is treated quantummechanically, the dipole Hamiltonian is used for the incoming THz radiation: $H = H^0 + H'(t)$ and $H'(t) = \mu E(t)$. Making the ansatz $\Psi(t) = a_1(t)\Psi_1 + a_2(t)\Psi_2$ one gets a system of two coupled equations for $a_1(t)$ and $a_2(t)$. They can be decoupled by the condition that the population in the lower level is close to one ($\left|\left|a_1^2\right|\right| \approx 1$) since the THz radiation field is not very strong. Figure 14 shows the result of the model calculation that reproduces the change in frequency as well as the observed phase jump after the driving pulse is over.

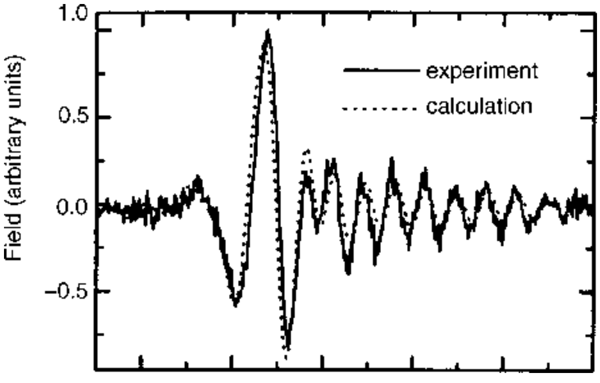

Figure 14 Model calculation of the polarization (dotted curve) and experimental data (solid line). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

These experiments are all performed in the linear regime but with nonadiabatic excitation conditions (half-cycle THz pulses). The excellent time resolution together with sufficiently high THz intensities can open up a previously unexplored field of quantum optics beyond the slowly varying envelope approximation.

Conclusions

The recent progress in the development of semiconductor sources together with room temperature THz detection techniques (free-space electro-optic sampling) paves the way for the commercialization of THz technology. In particular, spectrally sensitive THz imaging seems to have great potential in medicine and quality control of products.

See also

Semiconductor Materials: GaAs Based Compounds. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures.

Further Reading

Auston DH, Cheung KP and Smith PR (1984) Picosecond photoconducting Hertzian dipoles. *Applied Physics Letters* 45: 284.

Bonvalet A, Nagle J, Berger V, Migus A, Martin J-L and Joffre M (1996) Femtosecond infrared emission resulting from coherent charge oscillations in quantum wells. *Physical Review Letters* 76: 4392.

Bratschitsch R, Kersting R, Müller T, Strasser G, Unterrainer K, Fischler W and Höpfel RA (1999) Coherent THz plasmons in GaAs/AlGaAs superlattices. *Physica B* 272/1-4: 375.

Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent terahertz emission from optically pumped intersubband plasmons in parabolic quantum wells. *Applied Physics Letters* 76: 3501.

Chamberlain M (ed.) (1999) *Terahertz Spectroscopy and Applications*, SPIE Proceedings (Europto Series), vol. 3828, pp. 1–454.

Dekorsy T, Auer H, Waschke C, Bakker H, Roskos H, Kurz H, Wagner V and Grosse P (1995) Emission of submillimeter electromagnetic waves by coherent phonons. *Physical Review Letters* 74: 738.

Gornik E and Kersting R (2001) Coherent THz emission in semiconductors. In: Tsen KT (ed.) *Ultrafast Physical Processes in Semiconductors*, vol. 25, pp. 396. New York: Academic Press.

Heyman J, Kersting R and Unterrainer K (1998) Time-domain measurement of intersubband oscillations in a quantum well. *Applied Physics Letters* 72: 644.

Kersting R, Unterrainer K, Strasser G, Kauffmann H and Gornik E (1997) Few-cycle thz emission from cold plasma oscillations. *Physical Review Letters* 79: 3038.

Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmons in n-doped GaAs. *Physical Reviews B* 58: 4553.

Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272.

Nun MC and Orenstein J (1997) Terahertz time-domain spectroscopy. In: Gruener G (ed.) *Millimeter Wave Spectroscopy of Solids*. Berlin: Springer.

Roskos H, Nuss M, Shah J, Leo K, Miller D, Fox A, Schmitt-Rink S and Köhler K (1992) Coherent submillimeter-wave emission from charge oscillations in a double-well potential. *Physical Review Letters* 68: 2216.

Sarukura N, Ohtake H, Izumida S and Liu Z (1998) High average power radiation from femtosecond laser-irradiated InAs in a magnetic field and its elliptical polarization characteristics. *Journal of Applied Physics* 84: 654.

Sekine N, Yamanaka K, Hirakawa K, Voßebürger M, Haring-Bolivar P and Kurz H (1999) Observation of terahertz radiation from higher-order two-dimensional plasmon models in GaAs/AlGaAs single quantum wells. *Applied Physics Letters* 74: 1006.

Sha W, Smirl A and Tseng W (1995) Coherent plasma oscillations in bulk semiconductors. *Physical Review Letters* 74: 4273.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. Springer Series in Solid-State Sciences, vol. 115. Berlin: Springer.

Waschke C, Roskos H, Schwedler R, Leo K, Kurz H and Köhler K (1993) Coherent submillimeter-wave emission from bloch oscillations in a semiconductor superlattice. *Physical Review Letters* 70: 3319.

Wolter F, Martini R, Tolk S, *et al.* (1999) Influence of carrier–carrier scattering on intraband dephasing. *Superlattice and Microstructure* 26: 93.

Zhang X-C, Hu B, Darrow J and Auston D (1990) Generation of femtosecond electromagnetic pulses from semiconductor surfaces. *Applied Physics Letters* 56: 1011.

TIME-RESOLVED FLUORESCENCE

Contents

Laser Applications
Measurements in Polymer Science

Laser Applications

B Valeur, Conservatoire National des Arts et Métiers, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fluorescence is extensively used in physical, chemical, material, biological, and medical sciences, for investigating the structure and dynamics of matter or living systems. After excitation of a fluorescent sample by a short pulse of light, the fluorescence decay is, in the simplest case, a single exponential whose time constant is the lifetime of the excited state of the fluorescent species (i.e., the average time during which these species stay in the excited state). However, interactions of a fluorescent species with its micro-environment often perturb the fluorescence decay so that information on this micro-environment can be obtained. More specifically, when intermolecular photophysical processes (such as electron transfer, proton transfer, energy transfer, excimer, or exciplex formation) occur during the lifetime of the excited state, competition with normal de-excitation results in a fluorescence decay that contains the rate constants for those processes. Consequently, the fluorescence decay is generally a sum of discrete exponentials, or a more complicated function; for example, sometimes the system is characterized by a distribution of decay times. Thus it is of major interest to recover the parameters characterizing the δ-pulse response of a fluorescent sample (i.e., the response to an iznfinitely short pulse of light expressed as the Dirac function δ). With this aim, two time-resolved techniques, pulse fluorometry and phase-modulation fluorometry, are usually employed.

In pulse fluorometry, the sample is excited by a short pulse of light, and the fluorescence response is the δ-pulse response of the sample convoluted by the instrument response. In phase-modulation fluorometry, the sample is excited by modulated light at variable frequency, and the harmonic response of the sample obtained is the Fourier transform of the δ-pulse response. The first technique works in the time domain, and the second technique in the frequency domain. Pulse fluorometry and phase-modulation fluorometry are theoretically equivalent, but the principles of instrumentation are different. In both techniques, lasers are currently used as sources of pulsed or modulated light.

It is the aim of this article to explain the principles of pulse and phase fluorometries using lasers as light sources, and to compare these techniques.

General Principles of Time-Resolved Fluorometry

The principles of time-resolved fluorometry are illustrated in Figure 1. For any excitation function $E(t)$, the fluorescence response $R(t)$ of the sample is the convolution product of this function by the δ-pulse response $I(t)$:

$$R(t) = E(t) \otimes I(t) = \int_{-\infty}^{t} E(t')I(t - t')\,\mathrm{d}t' \qquad [1]$$

In pulse fluorometry, the sample is excited by a short pulse of light and the fluorescence response is recorded as a function of time. If the duration of the pulse is not short with respect to the time constants of the fluorescence decay, the fluorescence response is the convolution product given by eqn [1]: the fluorescence intensity increases, goes through a maximum, and becomes identical to the true δ-pulse response $i(t)$ as soon as the intensity of the light pulse becomes negligible. In this case, data analysis for the determination of the parameters characterizing the δ-pulse response requires deconvolution of the fluorescence response.

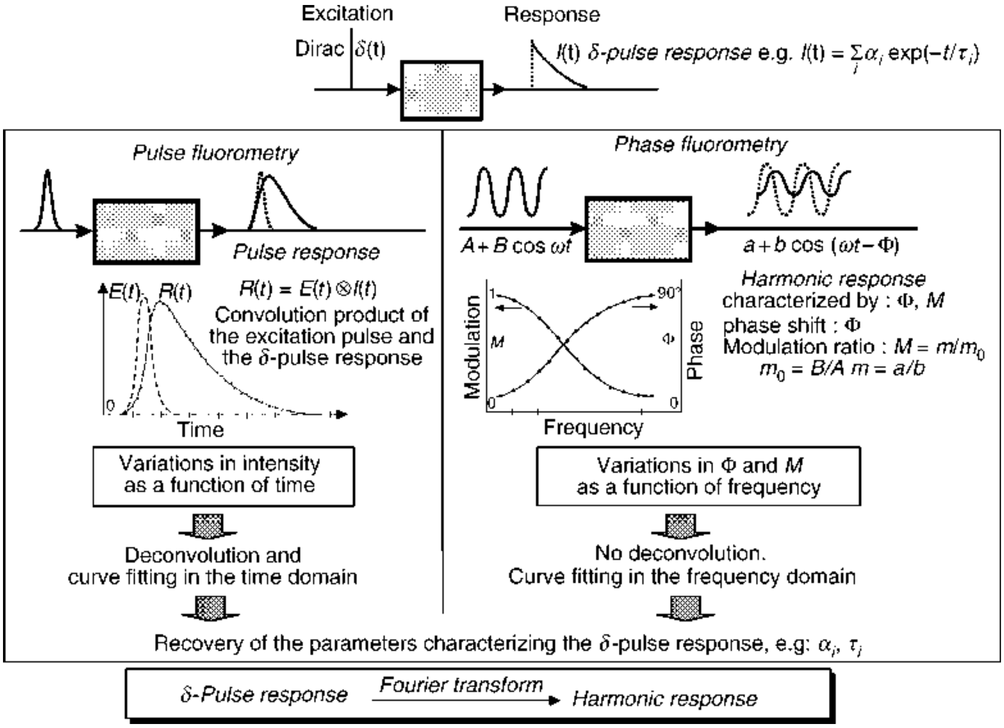
Excitation
Dirac $\delta(t)$
Response
$I(t)$ δ-pulse response e.g. $I(t) = \sum_i \alpha_i \exp(-t/\tau_i)$
Pulse fluorometry
Pulse response
$E(t)$, $R(t)$
$R(t) = E(t) \otimes I(t)$
Convolution product of the excitation pulse and the δ-pulse response
0
Time
Variations in intensity as a function of time
Deconvolution and curve fitting in the time domain
Phase fluorometry
$A + B\cos\omega t$
$a + b\cos(\omega t - \Phi)$
Harmonic response characterized by : Φ, M
phase shift : Φ
Modulation ratio : $M = m/m_0$
$m_0 = B/A$ $m = a/b$
Modulation
Phase
1
90°
M
Φ
0
0
Frequency
Variations in Φ and M as a function of frequency
No deconvolution. Curve fitting in the frequency domain
Recovery of the parameters characterizing the δ-pulse response, e.g: α_i, τ_i
δ-Pulse response
Fourier transform
Harmonic response

Figure 1 Principles of time-resolved fluorometry.

In phase fluorometry, the sample is excited by a sinusoidally modulated light at high frequency. The fluorescence response, which is the convolution product (eqn [1]) of the δ-pulse response by the excitation function, is sinusoidally modulated at the same frequency, but delayed in phase and partially demodulated with respect to the excitation. The modulation ratio M is defined as the ratio of the modulation depth m (AC/DC ratio) of the fluorescence and the modulation depth of the excitation m_0 (see Figure 1). The phase shift Φ and the modulation ratio M characterize the harmonic response of the system. These parameters are measured as a function of the modulation frequency. No deconvolution is necessary because the data are directly analyzed in the frequency domain.

The relationship between the harmonic response and the δ-pulse response can be obtained as follows. When the excitation function is sinusoidal, i.e., of the following form:

$$E(t) = E_0[1 + m_0 \exp(\mathrm{i}\omega t)] \qquad [2]$$

where ω is the angular frequency ($= 2\pi f$), the response of the system can be calculated using eqn [1]. It can then be shown that

$$M \exp(-\mathrm{j}\Phi) = \int_0^\infty i(t) \exp(-\mathrm{i}\omega t)\, \mathrm{d}t \qquad [3]$$

where $i(t)$ is the normalized δ-pulse response according to

$$\int_0^\infty i(t)\, \mathrm{d}t = 1 \qquad [4]$$

This important expression shows that the harmonic response expressed as $M \exp(-\mathrm{j}\Phi)$ is the Fourier transform of the δ-pulse response.

It is convenient to introduce the sine and cosine transforms P and Q of the δ-pulse response:

$$P = \int_0^\infty i(t) \sin(\omega t)\, \mathrm{d}t \qquad [5]$$

$$Q = \int_0^\infty i(t) \cos(\omega t)\, \mathrm{d}t \qquad [6]$$

If the δ-pulse response is not normalized according to eqn [4], then eqns [5] and [6] should be replaced by

$$P = \frac{\int_0^\infty I(t) \sin(\omega t)\, \mathrm{d}t}{\int_0^\infty I(t)\, \mathrm{d}t} \qquad [7]$$

$$Q = \frac{\int_0^\infty I(t) \cos(\omega t)\, \mathrm{d}t}{\int_0^\infty I(t)\, \mathrm{d}t} \qquad [8]$$

Since eqn [3] can be rewritten as $M \cos \Phi - iM \times \sin \Phi = Q - iP$, it is easy to show that the phase shift and the modulation ratio are given by

$$\Phi = \tan^{-1}\left(\frac{P}{Q}\right) \qquad [9]$$

$$M = [P^2 + Q^2]^{1/2} \qquad [10]$$

For a single exponential decay, the δ-pulse response is

$$I(t) = \alpha \exp(-t/\tau) \qquad [11]$$

where τ is the decay time and α is the pre-exponential factor or amplitude. The phase shift and relative modulation are related to the decay time by

$$\tan \Phi = \omega\tau \qquad [12]$$

$$M = \frac{1}{(1 + \omega^2\tau^2)^{1/2}} \qquad [13]$$

For a multi-exponential decay with n components, the δ-pulse response is

$$I(t) = \sum_{i=1}^{n} \alpha_i \exp(-t/\tau_i) \qquad [14]$$

Note that the fractional intensity of component i, i.e., the fractional contribution of component i to the total steady-state intensity, is

$$f_i = \frac{\int_0^\infty I_i(t)\, \mathrm{d}t}{\int_0^\infty I(t)\, \mathrm{d}t} = \frac{\alpha_i\tau_i}{\sum_{i=1}^{n} \alpha_i\tau_i} \qquad [15]$$

with, of course, $\sum_{i=1}^{n} f_i = 1$.

Using eqns [7] and [8], the sine and cosine Fourier transforms, P and Q, are given by

$$P = \frac{\omega\sum_{i=1}^{n} \frac{\alpha_i\tau_i^2}{1 + \omega^2\tau_i^2}}{\sum_{i=1}^{n} \alpha_i\tau_i} = \omega\sum_{i=1}^{n} \frac{f_i\tau_i}{1 + \omega^2\tau_i^2} \qquad [16]$$

$$Q = \frac{\sum_{i=1}^{n} \frac{\alpha_i\tau_i}{1 + \omega^2\tau_i^2}}{\sum_{i=1}^{n} \alpha_i\tau_i} = \sum_{i=1}^{n} \frac{f_i}{1 + \omega^2\tau_i^2} \qquad [17]$$

It should be noted that the optimum frequency for decay time measurements using either the phase shift or the modulation ratio is, according to eqns [12] and [13], such that $\omega\tau$ is close to 1, i.e., $f \approx 1/(2\pi\tau)$. Therefore, for decay times of 10 ps, 1 ns and 100 ns, the optimum frequencies are about 16 GHz, 160 MHz and 1.6 MHz, respectively.

In the case of a single exponential decay, eqns [12] and [13] provide two independent ways of measuring the decay time:

– by phase measurements:

$$\tau_\Phi = \frac{1}{\omega}\tan\Phi \qquad [18]$$

– by modulation measurements:

$$\tau_M = \frac{1}{\omega}\left(\frac{1}{M^2} - 1\right)^{1/2} \qquad [19]$$

The values measured in these two ways should, of course, be identical and independent of the modulation frequency. This provides two criteria to check whether an instrument is correctly calibrated, by using a lifetime standard whose fluorescence decay is known to be a single exponential.

It should be noted that the measurement of a decay time is fast (a fraction of a second) for a single exponential decay since a single frequency suffices. It is also of interest that a significant difference between values obtained by means of eqns [18] and [19] is compelling evidence for nonexponentiality of the fluorescence decay.

Pulse Fluorometers

The most widely used technique in time domain is the time-correlated single-photon counting technique, preferably called single-photon timing technique (SPT). Time-gated systems are less popular. Streak cameras offer a very good time resolution (a few picoseconds or less) but the dynamic range is smaller than that of the single photon-timing technique. The instruments that provide the best time resolution (a few tens of femtoseconds) are based on fluorescence up-conversion, but they are very expensive and not commercially available. Because of space limitations, only the single-photon timing technique will be presented below.

Principle of the Single-Photon Timing Technique

The basic principle of this technique relies on the fact that the probability to detect a single photon at time t, after an exciting pulse, is proportional to the fluorescence intensity at that time. After timing and recording the single photons following a large number of exciting pulses, the decay of the fluorescence intensity is reconstructed.

Figure 2 shows a schematic diagram of a conventional single-photon counting instrument using a pulsed laser as an excitation source. An electrical pulse associated to the optical pulse is generated (e.g. by a photodiode or the electronics associated to the excitation source) and routed – through a constant-fraction discriminator – to the start input of the time-to-amplitude converter (TAC). Meanwhile, the sample is excited by the optical pulse and emits fluorescence. The optics is tuned (e.g., by means of a neutral density filter) so that the photomultiplier detects no more than one fluorescence photon for each exciting pulse. The corresponding electrical pulse is routed – through a constant-fraction discriminator – to the stop input of the TAC. The latter generates an output pulse whose amplitude is directly proportional to the delay time between the start and the stop pulses. The height analysis of this pulse is achieved by an analog-to-digital converter and a multichannel

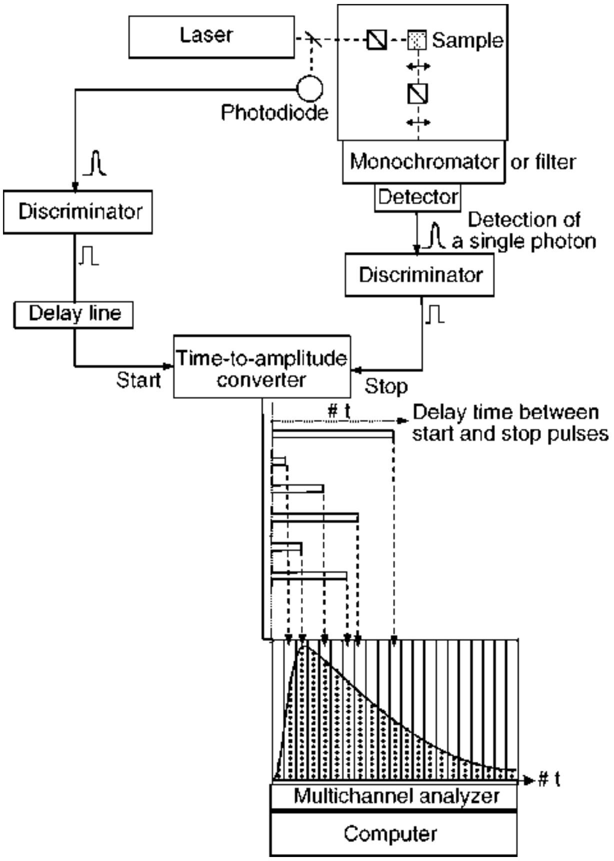

Figure 2 Schematic diagram of a single-photon timing fluorometer.

analyzer (MCA), which increases by one the contents of the memory channel corresponding to the digital value of the pulse. After a large number of excitation and detection events, the histogram of pulse heights represents the fluorescence decay curve.

Obviously, the larger the number of events, the better the accuracy of the decay curve. The required accuracy depends on the complexity of the δ-pulse response of the system; for instance, a high accuracy is, of course, necessary for recovering a distribution of decay times.

When deconvolution is required, the time profile of the exciting pulse is recorded under the same conditions by replacing the sample with a scattering solution (e.g., a suspension of colloidal silica (Ludox)).

It is important to note that the number of detected fluorescence photons must be kept much smaller than the number of exciting pulses ($<0.01-0.05$ stops per pulse), so that the probability of detecting two fluorescence photons per exciting photon is negligible. Otherwise, the TAC will take into account only the first fluorescence pulse and the counting statistics will be distorted: the decay will appear shorter than it is in reality. This effect is called the 'pile-up effect'.

Laser Sources for the Single-Photon Timing Technique

Mode-locked dye lasers and titanium-doped sapphire lasers are well suited to the single-photon timing technique: in fact, they can generate pulses over broad wavelength ranges, and the pulse widths are in the femtosecond and picosecond ranges with a high repetition rate. Flash lamps running in air, or filled with gas, can also be used, but the range of excitation wavelengths is limited; they deliver nanosecond pulses, and the repetition rate is not high (10^4–10^5 Hz). Dye lasers require the handling of large volumes of dye solutions which must be replaced when the dye is bleached. In contrast, the Ti:sapphire laser is a solid-state laser that does not have this drawback and delivers shorter light pulses. Moreover, long operational lifetime, operational simplicity, reliability, and broad tunability are distinct advantages. For these reasons, Ti:sapphire lasers have become more popular, and are described below.

Ti:sapphire lasers

Titanium-doped sapphire crystals are obtained by replacing a small fraction of the Al atoms of the host (Al_2O_3) with Ti atoms. When the concentration of Ti is low, the optical quality and the mechanical properties of Ti:sapphire crystals are comparable to those of ruby crystals in ruby lasers. The absorption spectrum of Ti:sapphire crystals exhibits a maximum at 490 nm and a shoulder at 540 nm. Therefore, pumping can be achieved by an Argon ion laser or a green solid-state laser. The emission spectrum ranges from 600 to 1,100 nm. Such a broad emission spectrum results from electronic coupling between the electronic levels of the 3d electrons with the lattice vibrations. Since excited-state absorption is not possible in the spectral region of the pump and the fluorescence spectrum, the tuning range is broad: it extends from 680 to 1,100 nm, i.e., almost over the entire fluorescence spectrum.

The width of the pulses that can be generated by Ti:sapphire lasers ranges from a few tens of femtoseconds to about 100 picoseconds. Commercially available Ti:sapphire lasers can operate either in the femtosecond configuration or in the picosecond configuration. In time-resolved fluorescence experiments using the single-photon timing technique, the instrument response is limited by the time response of the photomultiplier; the shortest width of this response that can be measured with the faster detectors (i.e., microchannel plate photomultipliers) is about 30–40 ps (see below). Therefore, there is no reason to operate the Ti:sapphire laser in the femtosecond configuration. In the picosecond configuration, the pulse width is <2–100 ps.

Figure 3 shows the six-mirror folded cavity design for the picosecond configuration in a Tsunami

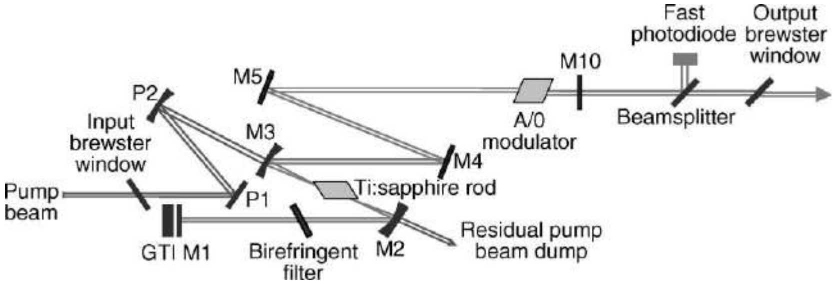

Figure 3 Beam path of the laser cavity of the Tsunami Ti:sapphire laser (Spectra Physics) operating in the picosecond configuration (Adapted from the Tsunami brochure with permission of Spectra Physics).

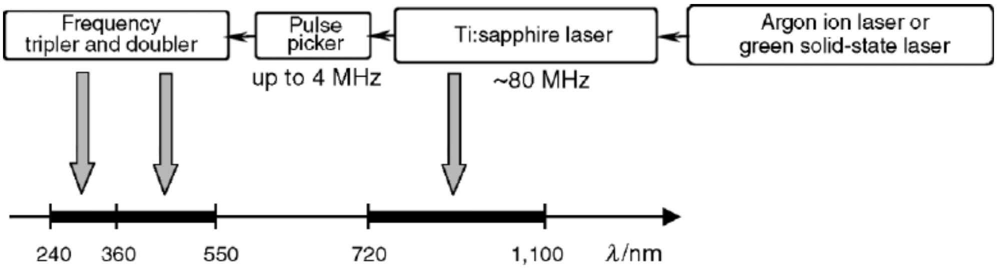
Frequency tripler and doubler
Pulse picker
up to 4 MHz
Ti:sapphire laser
~80 MHz
Argon ion laser or green solid-state laser
240
360
550
720
1,100
λ/nm

Figure 4 Schematics of the excitation system in time-resolved fluorescence experiments.

Ti:sapphire laser (Spectra Physics). A Brewster's angle acousto-optic modulator, driven by a low-power rf source, insures an 80 MHz repetition rate. Consequently, in contrast to standard passive mode-locking systems, the laser can operate for long periods without dropouts or shut-downs. Gires–Tournois Interferometer (GTI) mirror spacing is used for tuning the laser cavity. Wavelength tuning is achieved by means of a two-plate birefringent filter placed within the laser cavity at Brewster's angle. The tuning range depends on the mirror sets that are used. With broadband cavity mirrors, the entire tuning range of Ti:sapphire is covered, i.e., from below 700 nm to greater than 1,000 nm, and changes in wavelength can be made in a matter of minutes.

When pumping with a 10 W argon ion laser, the average output power is about 1.5 W at 800 nm and the pulse energy is about 15 nJ. Argon ion lasers can now be advantageously replaced by cw green solid-state lasers (532 nm) in which the gain medium ($Nd:YVO_4$) is pumped by diodes, and a LBO crystal placed inside the cavity ensures frequency doubling. These lasers offer several advantages: the head is compact and does not require external cooling water; it is insensitive to temperature, and thermal equilibration is very rapid.

Since the excited-state lifetime of fluorescent compounds can be as long as several hundreds of nanoseconds, the repetition rate of 80 MHz is too high to let the fluorescence of long lifetime samples vanish before a new exciting pulse is generated. Therefore, it must be limited to a few MHz, thanks to a pulse picker using a Bragg-angle acousto-optic modulator synchronized to the mode-locked 80 MHz pulse train. Selection of pulses from the train can thus be achieved at a maximum pulse selection rate of 8 MHz.

For excitation of fluorescent samples, wavelengths in the visible and UV regions are desirable. Frequency doublers and triplers using LBO and BBO crystals, respectively, have been designed with this aim. The possible tuning ranges are shown in Figure 4. When pumping with a 10 W argon ion laser, the average output powers are about 300 mW and 120 mW for frequency doubling and tripling, respectively.

With Ti-sapphire lasers and microchannel plate photomultipliers, decay times as short as 10 ps can be measured.

The results of an experiment are shown in Figure 5.

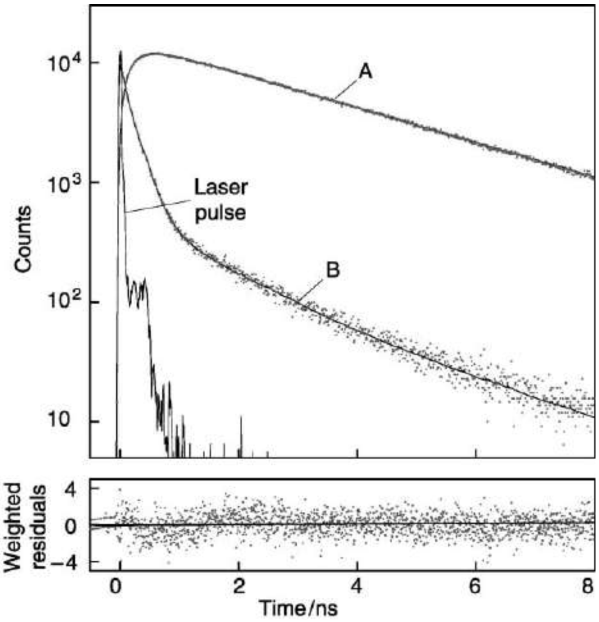

Figure 5 Example of data obtained by the single-photon timing technique using a Spectra-Physics setup composed of a titanium-sapphire Tsunami laser pumped by an argon ion laser, a pulse selector, and doubling (LBO) and tripling (BBO) crystals. The detector is a Hamamatsu MCP photomultiplier R3809U. The sample is a solution of DANS1 (a derivative of dansylamide) in a mixture CH_3CN/H_2O (60:40 v/v) at pH 0.8. Excitation wavelength: 286 nm. Channel width: 9.8 ps. A: Observation through a bandpass filter (300–370 nm) of the protonated form that undergoes photoinduced deprotonation. B: Observation through a cut-off filter ($\lambda > 400$ nm) of the neutral form; the observed risetime corresponds to the formation of the neutral form from the protonated form. Global analysis of the two curves leads to two time constants: 183 ± 9 ps and 3.01 ± 0.03 ns (Global $\chi^2_R = 1.26$) (Métivier R, Leray I and Valeur B, unpublished results. Métivier R, doctoral thesis, Ecole Normale Supérieure de Cachan, 2003).

Pulsed diode lasers

Picosecond diode laser heads offer an interesting alternative to mode-locked lasers and are much less expensive. They can produce light pulses as short as 50–90 ps with repetition rates from single shot to 40–80 MHz. Peak powers up to 500 mW can be obtained. However, the main disadvantage is the absence of tunability, and the number of wavelengths is limited: 375, 400, 440, 635 to 1,550 nm. No wavelength below 375 nm is presently available.

Phase-Modulation Fluorometers

In the frequency domain, the phase and modulation measurements can be taken by using either a cw laser and an optical modulator, or the harmonic content of a pulsed laser.

Phase Fluorometers Using a cw Laser and an Optical Modulator

The optical modulator is generally a Pockels cell. Low-power cw lasers are sufficient and not as expensive as pulsed lasers, so that one can afford several lasers of this type in order to get various excitation wavelengths. Available wavelengths from cw lasers are given in Table 1. A Xenon lamp can also be used; however, the optical modulator works better with a cw laser.

Figure 6 shows the schematic diagram of a multifrequency phase-modulation fluorometer using a cw laser and a Pockels cell. A beamsplitter reflects a few percent of the incident light towards a reference photomultiplier (via, or not, a cuvette containing a reference scattering solution). The fluorescent sample and a reference solution (containing either a scatter or a reference fluorescent compound) are placed in a rotating turret. The emitted fluorescence or scattered light is detected by a photomultiplier through a monochromator or an optical filter. The Pockels cell is driven by a frequency synthesizer and the photomultiplier response is modulated by varying the voltage at the second dynode by means of another frequency synthesizer locked in phase with the first one. The two synthesizers provide modulated signals that differ in frequency by a few tens of Hz in order to achieve cross-correlation (heterodyne detection). This procedure offers excellent accuracy because the phase and modulation information contained in the signal is transposed to the low-frequency domain, where phase shifts and modulation depths can be measured with a much better accuracy than in the high-frequency domain.

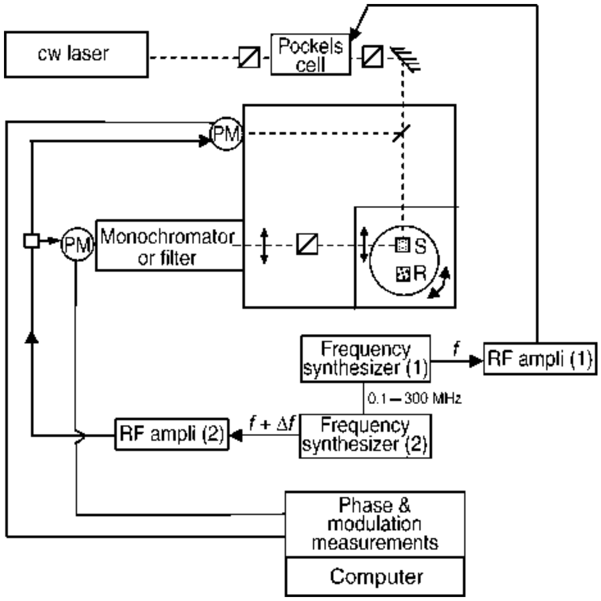

Figure 6 Schematic diagram of a multifrequency phase-modulation fluorometer using a cw laser and a Pockels cell (S: sample; R: reference; PM: photomultiplier).

Table 1 Examples of available wavelengths from low-power cw lasers

Laser	*Wavelength (nm)*
He–Cd	325; 442.5
He–Ne	543.3; 594; 611.9; 632.8
Ar^{+a}	457.9; 476.5; 488.0; 514.5

[a] High-power argon ion lasers can also provide wavelengths of 351.1 and 363.8 nm.

Practically, the phase delay ϕ_R and the modulation ratio m_R of the light emitted by the scattering solution are measured with respect to the signal detected by the reference photomultiplier. Then, after rotation of the turret, the phase delay ϕ_F and the modulation ratio m_F for the sample fluorescence are measured with respect to the signal detected by the reference photomultiplier. The absolute phase shift and modulation ratio of the sample are then $\Phi = \phi_F - \phi_R$ and $M = m_F/m_R$, respectively.

Figure 7 shows an example of an experiment.

Phase Fluorometers Using the Harmonic Content of a Pulsed Laser

The type of laser source that can be used is exactly the same as that for single-photon timing pulse fluorometry (see above). Such a laser system that delivers pulses in the picosecond range, with a repetition rate of a few MHz, can be considered as an intrinsically modulated source. The harmonic content of the pulse train – which depends on the

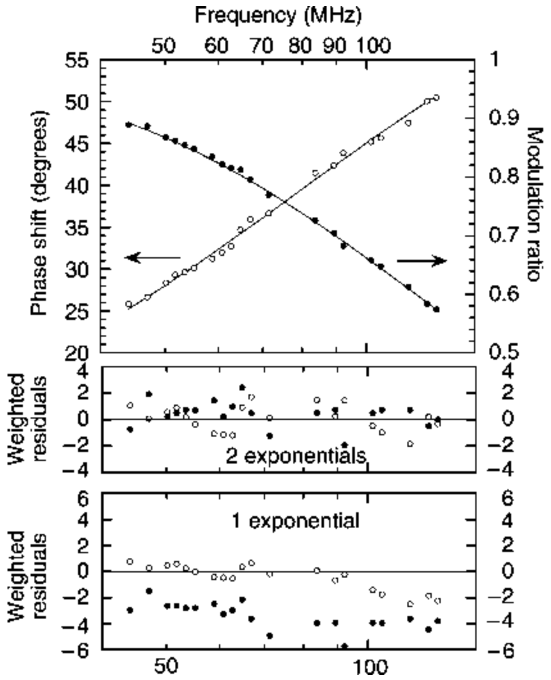

Figure 7 Example of data obtained by the phase and modulation technique using a fluorometer is equipped with a cw Omnichrome He–Cd laser (442 nm) and a Pockels cell operating at frequencies ranging from 0.1 to 200 MHz. The sample is a solution of a calcium complex of a crowned merocyanine (DCM-crown) in acetonitrile (observation through a Balzers bandpass filter centered at 581 nm). The solid line corresponds to the best fit with a biexponential decay. Weighted residuals are shown for the best fit with one and two exponentials. The time constants are 1.87 ± 0.04 and 0.35 ± 0.09 ns with fractional intensities of 0.85 and 0.15, respectively. The value of χ_r^2 is 1.01. (Redrawn from Martin MM, Plaza P, Meyer YH, *et al.* (1996) Steady-state and picosecond spectroscopy of Li^+ or Ca^{2+} complexes with a crowned merocyanine. Reversible photorelease of cations. *Journal of Physical Chemistry* 100: 6879–6888, with permission from the American Chemical Society).

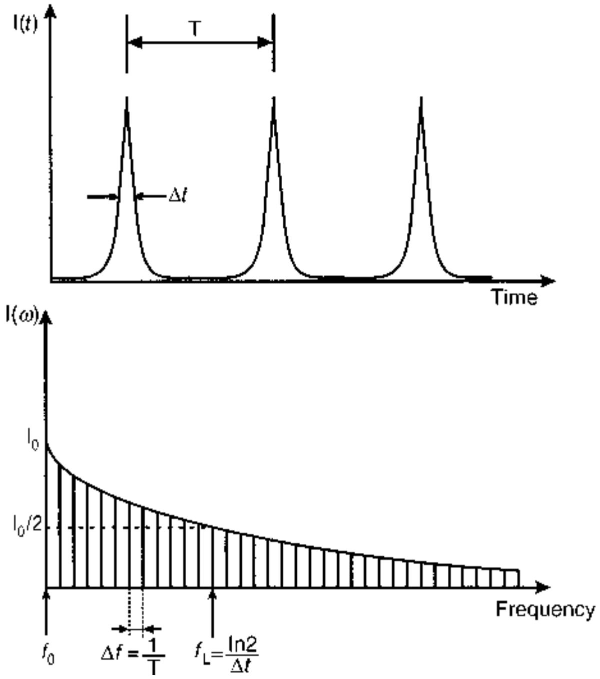

Figure 8 Harmonic content of a pulse train. Example of numerical values: Δ*f* = 4 MHz; Δ*t* = 4 Ps; f_L = 44 GHz.

pulse width (as illustrated in Figure 8) – extends to several gigahertz.

For high-frequency measurements, normal photomultipliers are too slow, and microchannel plate photomultipliers are required. However, internal cross-correlation is not possible with the latter and an external mixing circuit must be used.

The time resolution of a phase fluorometer, using the harmonic content of a pulsed laser and a microchannel plate photomultiplier, is comparable to that of a single-photon timing instrument using the same kind of laser and detector.

Data Analysis

In both pulse and phase fluorometries, the most widely used method of data analysis is based on a nonlinear least-squares method. The basic principle of this method is to minimize a quantity which expresses the mismatch between data and fitted function. This quantity is the reduced chi-square χ_r^2, defined as the weighted sum of the squares of the deviations of the experimental response $R(t_i)$ from the calculated responses $R_c(t_i)$:

$$\chi_r^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left[\frac{R(t_i) - R_c(t_i)}{\sigma(i)}\right]^2 \qquad [20]$$

where N is the total number of data points and $\sigma(i)$ is the standard deviation of the ith data point, i.e., the uncertainty expected from statistical considerations (noise). ν is the number of degrees of freedom ($\nu = N - p$, where p is the number of fitted parameters). The value of χ_r^2 should be close to 1 for a good fit.

In the single-photon timing technique, the statistics obey the Poisson distribution and the expected deviation $\sigma(i)$ is approximated to $[R(t_i)]^{1/2}$ so that eqn [20] becomes

$$\chi_r^2 = \frac{1}{\nu}\sum_{i=1}^{N}\frac{[R(t_i) - R_c(t_i)]^2}{R(t_i)} \qquad [21]$$

In phase fluorometry, no deconvolution is required: curve fitting is performed in the frequency domain, directly using the variations of the phase shift Φ and the modulation ratio M as functions of the modulation frequency. Phase data and modulation data can be analyzed separately, or simultaneously. In the latter case the reduced chi-squared is given by

$$\chi_r^2 = \frac{1}{\nu}\left[\sum_{i=1}^{N}\left[\frac{\Phi(\omega_i) - \Phi_c(\omega_i)}{\sigma_\Phi(\omega_i)}\right]^2 + \sum_{i=1}^{N}\left[\frac{M(\omega_i) - M_c(\omega_i)}{\sigma_M(\omega_i)}\right]^2\right] \qquad [22]$$

where N is the total number of frequencies. In this case, the number of data points is twice the number of frequencies, so that the number of degrees of freedom is $\nu = 2N - p$.

In addition to the value of χ_r^2, it is useful to display graphical tests. The most important of them is the plot of the weighted residuals defined as

$$W(t_i) = \frac{R(t_i) - R_c(t_i)}{\sigma(i)} \qquad [23]$$

where $\sigma(i) = [R(t_i)]^{1/2}$ for single-photon counting data. The fit is satisfactory when the weighted residuals are randomly distributed around zero.

Concluding Remarks

From the instrumental point of view, the latest generations of pulse and phase fluorometers both use pulsed lasers (usually Ti:sapphire lasers) and microchannel plate detectors; only the electronics is different. Since the instrument response is mainly limited by the time response of the detector, the time resolution is the same for both techniques. Moreover, the optical module is identical so that the total cost of instrumentation is similar.

There are, however, significant differences between pulse and phase fluorometers from the methodological point of view. In pulse fluorometry, the single-photon timing technique offers various advantages: direct visualization of the fluorescence decay; high sensitivity; outstanding dynamic range and linearity; and well-defined statistics (Poisson distribution) allowing the accurate weighing of each point in data analysis. In phase-modulation fluorometry, the main advantage is that no deconvolution is necessary, while this operation is often necessary in pulse fluorometry and requires great care in recording the instrument response, especially for very short decay times. Moreover, the short acquisition time for phase shift and modulation ratio measurements at a given frequency is a distinct advantage.

See also

Ultrafast Laser Techniques: Pulse Characterization Techniques.

Further Reading

Alcala JR, Gratton E and Jameson DM (1985) A multifrequency phase fluorometer using the harmonic content of a mode-locked laser. *Analytical Instrumentation* 14: 225–250.

Demas JN (1983) *Excited-State Lifetime Measurements*. New York: Academic Press.

Jameson DM, Gratton E and Hall RD (1984) The measurement and analysis of heterogeneous emissions by multifrequency phase and modulation fluorometry. *Applied Spectroscopy Reviews* 20: 55–106.

Lakowicz JR (ed.) (1991) *Topics in Fluorescence Spectroscopy, Vol. 1: Techniques*. New York: Plenum Press.

Lakowicz JR, Laczko G and Gryczynski I (1986) 2 GHz frequency-domain fluorometer. *Review of Scientific Instruments* 57: 2499–2506.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*. London: Academic Press.

Valeur B (2002) *Molecular Fluorescence. Principles and Applications*, chap. 6. Weinheim, Germany: Wiley-VCH.

Measurements in Polymer Science

T A Smith, University of Melbourne, Melbourne, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In this article we explore the use of ultrashort pulsed lasers in time-resolved fluorescence measurements for the study of the dynamics of excimer formation and energy transfer/migration in synthetic aromatic polymers. The temporal profile of monomer and excimer emission, and the use of time-resolved fluorescence polarization approaches are discussed.

The use of lasers in chemistry has most often exploited the high degree of spectral purity achievable with these devices, in both photochemical and high-resolution spectroscopic applications. Through the use of lasers, we also have the ability to generate pulses of light of ultrashort duration (a few femtoseconds) and of incredibly high peak powers (petawatts), characteristics significantly beyond those achievable with conventional (incoherent) light sources. This enables the kinetics of the many photophysical and photochemical processes that occur on ultrafast time-scales with high efficiency within polymers, to be determined through various time-resolved measurements.

These processes are dealt with in detail in specialist photochemistry textbooks and so knowledge of the terms associated with photochemistry and photophysics is assumed (e.g., electronic configurations, orbitals, spin, multiplicity, etc.).

Following the absorption of light, an isolated chromophore will be in one of its excited electronic and vibrational states. Once formed, the excited state of a molecule will only survive for a limited time, with the molecule losing its excess energy to return to a preferred lower energy state (electronic configuration). It can do this through a number of spontaneous, radiative, and nonradiative deactivation processes, each with its own characteristic rate constant k_{ij}. Nonradiative pathways include internal conversion (IC), such as intramolecular vibrational redistribution, and vibrational coupling to the bath, and inter-system crossing (ISC) between singlet and triplet manifolds. Radiative pathways involve the emission of a photon of light, in the form of fluorescence or phosphorescence.

The time the molecule spends in the excited state is determined by the rates of all the pathways that contribute to the overall deactivation of that state. The lifetime, τ, of a singlet state is therefore defined as:

$$\tau = \frac{1}{k_{\mathrm{ISC}} + k_{\mathrm{IC}} + k_{\mathrm{Rad}} + \cdots} = \frac{1}{\sum k_{ij}} \qquad [1]$$

In systems of higher chromophore concentration, such as aromatic-containing polymers, additional, bimolecular pathways are available through which the excess excited state energy may be dissipated, making it rare that a single fluorescence ‘lifetime’ is observed from polymer systems. The observation of complex (multi- or nonexponential) fluorescence decay profiles sometimes invokes the use of the terms ‘average lifetimes’ or ‘lifetime distributions’.

The complex nonexponential fluorescence decay behavior observed, particularly from polymer systems, has been variously attributed to one or more of the following factors: molecular aggregation (e.g., excimer/exciplex equilibria); acid–base equilibria (e.g., proton transfer); one-, two-, or three-dimensional energy transfer and energy migration (and variations in the singlet energy migration length); electron transfer; molecular motion (translational diffusion, rotational relaxation); vibrational or solvent relaxation; concentration quenching; fluorescent impurities, etc. The kinetics of excimer formation are further complicated by heterogeneity of kinetically distinct monomer and/or excimer species, configurational and conformational influences upon excimer formation, and the time dependence of energy trapping (excimer formation) through nonequilibrium diffusion-controlled processes. Here, we will concentrate on excimer formation and energy migration in synthetic aromatic polymers, as shown schematically in Figure 1.

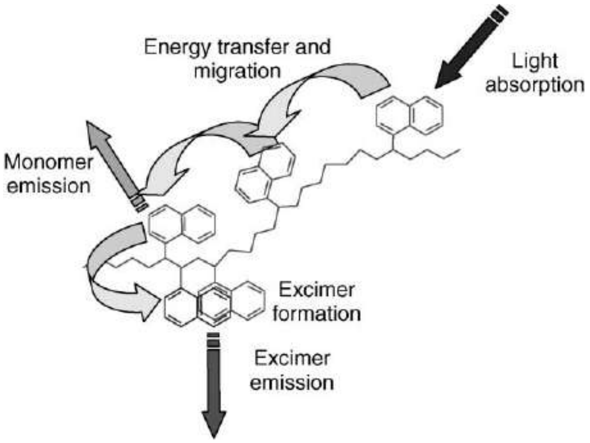

Figure 1 Schematic representation of some of the photophysical processes in polymers.

Ultrafast Laser Spectroscopic Techniques

A wide range of experimental techniques and instrumentation is available for the study of time-resolved processes in polymers. The most widely used methods rely on the detection of emission from the polymer. Fluorescence-based techniques are popular for investigating time-dependent processes in polymers, since the time-scales involved in many of the photochemical and photophysical processes discussed correspond with the time-scales on which fluorescence occurs (normally on the subpicosecond to nanosecond time-scales). Phosphorescence usually occurs on significantly longer time-scales and therefore is less likely to be competitive with fluorescence or other radiative or nonradiative deactivation pathways under normal conditions. Fluorescence-based techniques are attractive for reporting on a number of time-dependent processes since fluorescence is inherently multidimensional, it can be readily resolved temporally, and fluorescence detection can be extremely sensitive, achieving single photon and single molecule detection levels.

Amongst the most commonly used fluorescence-based techniques are: time-correlated single photon counting (TCSPC, or single photon timing), fluorescence transient digitization, and ultrafast streak cameras. Methods such as transient absorption (flash photolysis) can also provide some details regarding the identity of short-lived species formed following photo-excitation. Other ultrafast laser techniques have been applied to the study of time-dependent processes in polymers in order to achieve higher temporal resolution. So-called pump–probe methods provide the ultimate in time-resolution and can be combined with fluorescence detection through the use of Kerr shutters or ultrafast fluorescence upconversion techniques. Other, more complex pump–probe methods, such as transient grating and multiple pulse photon echo measurements, are potentially very powerful tools in the study of time-resolved processes in polymers.

The generation of short pulses of light can be achieved through numerous mechanisms, including active or passive mode-locking and Q-switching techniques, and superradiance processes (e.g., in nitrogen lasers). Lasers capable of producing short (~ns) pulses of light by Q-switching are usually characterized by high energies per pulse (from μJ up to many hundreds of mJ/pulse) at low (~Hz–kHz) pulse repetition rates. Lasers producing pulses of ultrashort duration (often in the picosecond or femtosecond region) through mode-locking techniques, characteristically operate at high pulse repetition rates (~MHz) and with lower energies per pulse (~nJ/pulse). It should be noted that even pulses of a few nJ energy delivered in ~100 fs, still have peak powers of tens of kW. These high peak powers enable the generation of a range of excitation wavelengths suitable for the study of polymer systems through nonlinear optical wavelength conversion processes, such as second-harmonic generation (SHG) and optical parametric oscillation (OPO) devices.

Excimer Formation

Excimers (or *exci*ted-state di*mers*) are formed when an aromatic molecule, while in its excited electronic state, $^1M^*$, interacts with a second molecule of the same chemical type in its ground electronic state, ^1M. Excimer formation is often observed in concentrated molecular systems such as concentrated solutions and molecular crystals; however, it can also occur in dilute solutions of polymers containing aromatic chromophores, or in the confined environment of solid polymer systems such as homopolymer films.

The structure of excimers in solution is generally thought to be a face-to-face, coplanar, sandwiched, or in some cases, partially eclipsed arrangement of the two constituent aromatic moieties, which attempt to adopt the lowest possible energy configuration, with a separation of 3–4 Å, as depicted in Figure 2. A range of possible excimer configurations may exist rather than just the two extremes.

A feature of excimers is that they have a dissociative ground state (Figure 3), i.e., they do not exist prior to the absorption of light by one monomer chromophore and thus cannot be formed directly without the coming together of the two constituent chromophores within the lifetime of the $^1M^*$. Excimers, being of lower energy than the corresponding excited monomer species and with this dissociative ground state, usually exhibit characteristically

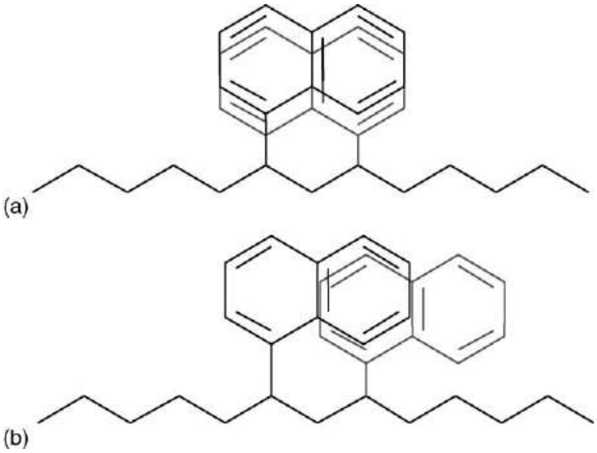

Figure 2 (a) Sandwiched and (b) partially eclipsed excimer configurations.

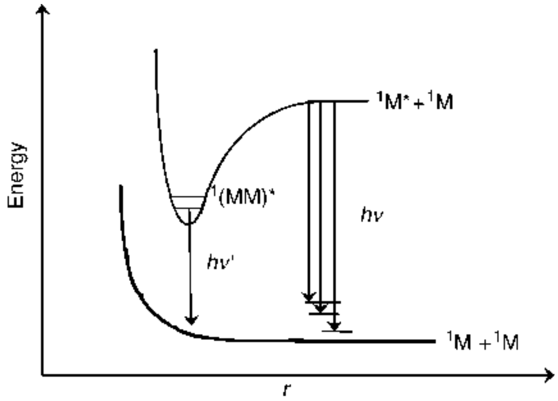

Figure 3 Schematic potential energy diagram for excimer formation and decay. *r* represents the monomer–monomer separation.

broad, structureless emission to the red of the monomer emission (Figure 4). Upon the emission of fluorescence from an excimer, the complex loses its excess energy and dissociates back to two ground state monomer units (i.e., 2 ^{1}M). Excimers may also dissociate back to a noninteracting pair of $^1M^*$ and ^{1}M.

Excimer formation of simple aromatic molecules in solution is often modeled on the kinetic scheme proposed by Birks and co-workers (Figure 5). From Figure 5, if a δ-pulse excitation, represented by $h\nu_{\mathrm{abs}}$, is used to produce an initial concentration of excited singlet monomer molecules, $[^1M^*]_o$, then such a scheme results in the following rate equations:

$$\frac{d[^1M^*]}{dt} = k_{MD}[^1(MM)^*] - X[^1M^*] \qquad [2]$$

$$\frac{d[^1(MM)^*]}{dt} = k_{DM}[^1M][^1M^*] - Y[^1(MM)^*] \qquad [3]$$

where $X = k_{FM} + k_{NM} + k_{DM}[^1M^*] = k_M + k_{DM}[^1M]$, $Y = k_{FD} + k_{ND} + k_{MD} = k_D + k_{MD}$.

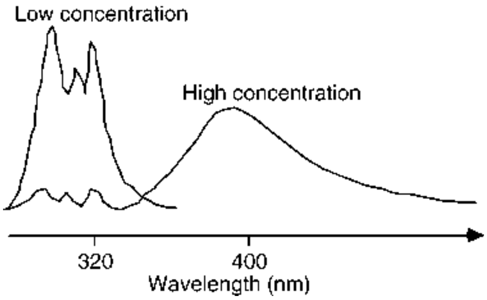

Figure 4 Schematic representation of monomer and excimer emission.

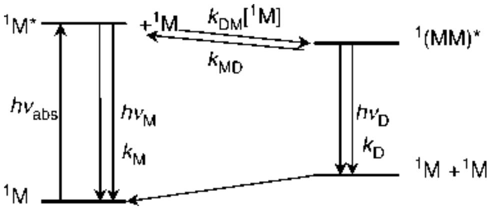

Figure 5 Kinetic scheme for excimer formation where $h\nu_M$ and $h\nu_D$ represent the emission of photons by the monomer and excimer species respectively, k_M and k_D comprise the sum of the various rate constants for all deactivation pathways for the monomer and excimer respectively (excluding k_{MD} and k_{DM}) (e.g., $k_M = k_{FM} + k_{NM}$ where k_{FM} and k_{NM} are the radiative (fluorescence) and nonradiative deactivation rate constants respectively of the monomer).

Solving these rate equations and applying the initial conditions $[^1M^*]_t = [^1M^*]_o$ and $[^1(MM)^*]_o = 0$ at $t = 0$, yields expressions given below for the decay of monomer and excimer fluorescence intensity with time. The emission intensity (monitored in a spectral region in which only the monomer $[^1M^*]$ species emits), will have the temporal form given by eqn [4]:

$$\begin{aligned} i_M(t) &= k_{FM}\frac{[^1M^*]}{[^1M^*]_o} \\ &= \left[k_{FM}\frac{(\lambda_2 - X)}{(\lambda_2 - \lambda_1)}\right]\left[\exp(-\lambda_1 t) + A\exp(-\lambda_2 t)\right] \end{aligned} \qquad [4]$$

where

$$A = \frac{(X - \lambda_1)}{(\lambda_2 - X)},$$

$$\lambda_{1,2} = \frac{1}{2}\left[X + Y \mp \left\{(Y - X)^2 + 4k_{MD}k_{DM}[^1M^*]\right\}^{1/2}\right]$$

The emission intensity collected from a spectral region in which the excimer ($^1[MM]^*$) species only emits, will have the temporal form given by eqn [5]:

$$\begin{aligned} i_D(t) &= k_{FD}k_{DM}\frac{[^1D^*]}{[^1M^*]_o} \\ &= \left[\frac{k_{FD}k_{DM}[^1M]}{(\lambda_2 - \lambda_1)}\right]\left[\exp(-\lambda_1 t) - \exp(-\lambda_2 t)\right] \end{aligned} \qquad [5]$$

If excimer formation follows Birks' scheme, then the decay profile of the monomer fluorescence should therefore be described by the sum of two exponential decay terms, and the excimer emission by an exponential growth term and a decay term (Figure 6). Only two decay constants, λ_1 and λ_2, should be required to fit both the monomer and excimer fluorescence profiles successfully. This is often not observed, which has led to the introduction of more complex interpretations.

In synthetic aromatic polymers, the effective local concentration of the aromatic molecules can be high while the concentration of polymer chains in solution may be very low. This means that excimer formation can be significant in such polymers, even in dilute solution. In fact, in some polymers the emission is dominated by excimers over monomers. The fact that the excimers are of lower energy than the corresponding excited monomer species also means that they can act as effective traps of excitation energy that might be migrating/relocating within a polymer chain as discussed below.

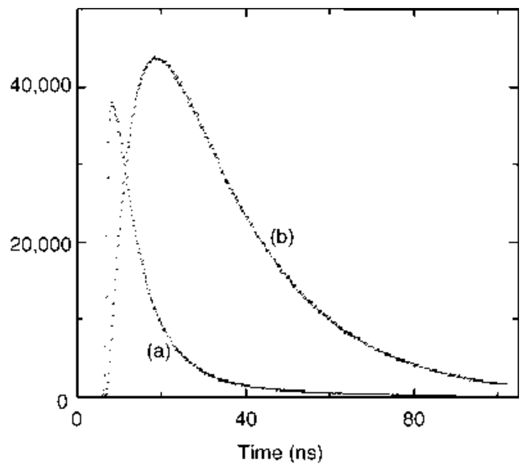

Figure 6 Examples of (a) monomer and (b) excimer fluorescence decays.

Excimer formation in solutions of aromatic polymers is generally thought to proceed via a complicated route that may involve a combination of electronic energy transfer and migration, segmental rotation, and chromophore diffusion. It is believed that energy migrates along a polymer chain until sites suitable for excimer formation are encountered, at which point the energy is trapped. A number of mechanisms for the formation of suitable excimer sites in polymers are then conceivable. These include:

(i) interactions between chromophores belonging to different chains (this *inter*molecular excimer formation is highly concentration dependent);

(ii) interactions between two chromophores far removed from each other, but on the same chain. This necessitates mainchain conformational changes and side group motions during the excited state lifetime, so as to bring both molecules together;

(iii) near neighbor interactions (c.f. Hirayama's '$n = 3$' rule). The configuration required for excimer formation need not exist prior to excitation provided that the two constituent groups can attain the required geometry within the lifetime of the excited chromophore; or

(iv) preformed sites, suitable for excimer formation, present in the polymer before the initial absorption of light occurs. These may be formed by interactions between nearest neighbor chromophores or between two chromophores on loops of the same, or different chains in cases where chain motion is restricted.

The first two of these mechanisms depend upon segmental motion of the polymer chain and the third mechanism may depend upon rotation of the chromophore/s to adopt the close proximity and optimum orientational configuration required for excimer formation. In solution, this may be, at least partially, diffusion-controlled, causing a dominant complicating factor in the kinetics of excimer formation in polymeric systems. One approach is therefore to substitute a time-dependent expression for the rate coefficient of excimer formation, $k_{\mathrm{DM}}(t) = A + B/\sqrt{t}$, (based on the Einstein–Smoluchowski diffusion theory), for k_{DM}, into the rate equations. Such an approach produces the following expression for the time dependence of the monomer fluorescence, $I_{\mathrm{M}}(t)$ in the presence of diffusion-controlled excimer formation and the absence of excimer back dissociation:

$$I_{\mathrm{M}}(t) = A \exp\left\{-at - b\sqrt{t}\right\} \qquad [6]$$

In some cases, such as when excimer back dissociation cannot be neglected, more complex functional forms have been derived, such as:

$$I_{\mathrm{M}}(t) = A \exp\left\{-at - b\sqrt{t}\right\} + B \exp\left\{-t/\tau\right\} \qquad [7]$$

Similar expressions can be derived for the time dependence of the excimer fluorescence, $I_{\mathrm{E}}(t)$. More detailed expressions for the term $k(t)_{\mathrm{DM}}$, involving the complete form of the diffusion rate coefficient, have also been used in such approaches.

A related approach is the concept of 'convolution kinetics', in which the intimate dependence of excimer formation upon the monomer decay is emphasized. Once formed, the excimer is assumed to decay exponentially. The excimer fluorescence temporal profile is predicted to take the form:

$$I_{\mathrm{E}}(t) = k_1(t) I_{\mathrm{M}}(t) \otimes \exp\left(\frac{-t}{\tau}\right) \qquad [8]$$

where $\otimes$ represents the convolution integral and $I_{\mathrm{E}}(t)$ and $I_{\mathrm{M}}(t)$ are the measured decay profiles in the excimer and monomer emission regions respectively.

Electronic Energy Transfer, Energy Migration and Trapping

In polymers the high local concentration of chromophores makes additional deactivation pathways competitive with the other photophysical processes.The excess energy may transfer from the excited fluorophore of interest to a lower energy chromophore. Three of the more common possible mechanisms, through which energy transfer can occur, are:

(i) *Radiative* ('*trivial*') – involves the reabsorption of donor emission by an acceptor chromophore.

(ii) *Resonance* (*Dipole–Dipole*, *Förster*) – long-range (15–100 Å), nonradiative, single-step energy transfer process arising from Coulombic (dipole–dipole) interactions between an excited-state donor chromophore, D*, and an initially unexcited acceptor, *A*. Förster derived a statistical expression for the (first-order) rate constant, for such a process, considering three-dimensional, dipole–dipole interactions between stationary molecules:

$$k(R)_{\mathrm{D}^*\to\mathrm{A}} = \frac{9000\kappa^2\phi_{\mathrm{o}}\ln 10}{128\pi^5\eta^4 N\tau_{\mathrm{D}}R^6}\int_0^\infty \frac{f_{\mathrm{D}}(\bar{\nu})\varepsilon_{\mathrm{A}}(\bar{\nu})}{\bar{\nu}}\mathrm{d}\bar{\nu} = \frac{1}{\tau_{\mathrm{D}}}\left(\frac{R_{\mathrm{o}}}{r_{\mathrm{D}}}\right)^6 \quad [9]$$

where R is the donor–acceptor distance, ϕ_{o} is the quantum yield for emission from the donor in the absence of acceptor molecules, η is the refractive index of the solvent, τ_{D} is the lifetime of the donor excited state (reduced by the onset of electronic energy transfer), N is Avogadro's number, κ is a constant dependent on the mutual orientation of the donor and acceptor molecules, and the integral term is a measure of the spectral overlap between the donor emission profile, $f_{\mathrm{D}}(\bar{\nu})$ (in quanta cm^{-1} and normalized to unit area) and the acceptor absorption profile, $\varepsilon_{\mathrm{A}}(\bar{\nu})$ (converted to molar extinction coefficients and measured on a wavenumber scale). R_{o} is the critical transfer distance, at which the probability of resonance transfer, $k(R)_{\mathrm{D}^*\to\mathrm{A}}$, is equal to the probability of spontaneous decay, $1/\tau_{\mathrm{D}}$.

(iii) *Exchange mechanism* – also a nonradiative, single-step process but involves the exchange of electronic excitation when the donor and acceptor are close enough so that their electron clouds overlap. Consequently it only operates over small distances. Dexter has derived a rate expression for exchange energy transfer, again assuming that the donor and acceptor molecules are stationary:

$$k_{\mathrm{D}^*\to\mathrm{A}} = \left(\frac{2\pi}{\hbar}\right)Z^2\int f_{\mathrm{D}^*}(\bar{\nu})\varepsilon_{\mathrm{A}}(\bar{\nu})\mathrm{d}\bar{\nu} \quad [10]$$

where $Z^2 = K^2\exp(-2R/L)$ and K and L are simply constants.

It has been clearly demonstrated that energy transfer, from pendant aromatic monomer units to chemically bound acceptor chromophores, can occur with high efficiency and is influenced significantly by polymer structure and conformation.

In addition to energy transfer between D* and A (being different chemical species), more than one exchange of excitation energy between chemically and spectroscopically identical chromophores, attached by covalent bonds to a polymer chain, can occur. A single step between identical units is still termed energy transfer, while more than one such step in a sequence would constitute intramolecular energy migration.

Energy migration is a long-range process, thought to occur via a series of energy transfer steps occurring between chromophores, either: (i) along the polymer chain (often the adjacent groups); (ii) across loops in a single polymer chain; or (iii) across different polymer chains. The migration continues until either a trap site is encountered or one of the chromophores excited in the migration process deactivates spontaneously. Each step is believed to occur with a probability defined by the exchange or Förster mechanisms, but since the energy transfer occurs between donor and acceptor groups, which are of the same type, the spectral overlap integral involved in both mechanisms will usually be small. The distances over which such energy transfer can occur therefore cannot be large and typically takes values of <10–20 Å. It is worth noting that many authors consider the energy to migrate as an incoherent exciton (where the term exciton is used in a much looser sense than is customary among solid-state physicists).

Long-range energy migration can only occur by energy transfer from the initially excited chromophore to ground state chromophores of the same, or lower, energy. If energy transfer is favorable between one chromophore and another of the same energy, then back transfer to the initially excited chromophore is also highly probable. When there is a change in the emission and absorption dipole orientations, caused by rotation of the chromophores relative to one another on the time-scale of the energy transfer step, or the transfer is to a chromophore of significantly lower energy, then the probability of energy transfer back to the original chromophore might be reduced. The lower energy chromophores on the polymer chain may therefore act as 'traps' of the excitation originally absorbed by a higher energy chromophore, even in the absence of excimer energy trap sites.

Various approaches have been proposed for modeling the kinetics of energy migration. One can consider the time-dependent nature of each energy transfer step of the migration process, and introduce a time-dependent rate coefficient for each step of the energy migration process, $k_{\mathrm{DD}}(t)$. The form for $k_{\mathrm{DD}}(t)$ by Sienicki *et al.*, is given by eqn [11], which leads to nonexponential functions predicted for the ensemble averaged probability that an originally excited donor molecule is still excited at time t:

$$k_{\mathrm{DD}}(t) = \left(\frac{\Delta}{n}\right)\gamma_\Delta c_{\mathrm{D}}\Gamma(1-\Delta/n)\left(\frac{t}{\tau_{\mathrm{OD}}}\right)^{\Delta/n-1} \qquad [11]$$

where $\Delta(=1,\ 2,\ 3)$ is the dimensionality of the medium, $n(=6,\ 8,\ 10,\ldots)$ is the order of the interaction, τ_{OD} is the lifetime of the donor fluorescence, $\Gamma(x)$ is the gamma function, γ_Δ is a constant dependent upon Δ and c_{D} is the donor concentration.

Energy migration has also been considered as a diffusion-like process, and the Smoluchowski diffusion theory adopted. The rate coefficient for energy migration, $k(t)_{\mathrm{EM}}$, is then expressed in a form similar to the time-dependent rate coefficient for partially diffusion-controlled reactions (eqn [12]). In this type of treatment, the rate of change of the concentration of excited monomer chromophores, $^1\mathrm{M}^*$, is assumed to be given by eqn [13]:

$$k(t)_{\mathrm{EM}} = 4\pi N'_{\mathrm{A}} R D\rho\left[1 + R\rho(\pi D t)^{-1/2}\right] \qquad [12]$$

$$\frac{\mathrm{d}[^1\mathrm{M}^*]}{\mathrm{d}t} = I_{\mathrm{o}} - \left(\frac{1}{\tau_{\mathrm{o}}} + k(t)_{\mathrm{EM}}\right)[^1\mathrm{M}^*] \qquad [13]$$

In these expressions, τ_{o} is the reciprocal of the sum of the rate constants for radiative and nonradiative deactivation, ρ is the reaction probability per collision, R is the interaction distance, and N'_{A} is the number density of the energy acceptors or traps ($N'_{\mathrm{A}} = NC_{\mathrm{A}}$, with the concentration of acceptors, C_{A}, in mol cm^{-3}). D is the mutual diffusion coefficient which can be replaced by a migration coefficient, Λ, R can be replaced by the Förster critical transfer distance, R_{o}, and the reaction probability, ρ, can be set to 0.5 (since R_{o} is the distance at which the probability of energy transfer is 0.5). This leads to a simplified expression for $k(t)_{\mathrm{EM}}$, which can be substituted into eqn [13], to produce an expression (eqn [14]) for the monomer decay:

$$I(t) = A\exp\left(-\left(\frac{1}{\tau_{\mathrm{o}}} + 2\pi N'_{\mathrm{A}} R_{\mathrm{o}}\Lambda[^1\mathrm{M}^*]\right)t - 2N'_{\mathrm{A}} R_{\mathrm{o}}^2(\pi\Lambda)^{1/2}[^1\mathrm{M}^*]\sqrt{t}\right) \qquad [14]$$

Yokota and Tanimoto have also developed a statistical representation of Förster's theory and derived an expression for dipole–dipole transfer in a fluid medium in which the distribution of excited-state donor and acceptor molecules is determined, both by diffusion and by the decay and transfer of the donor excited state. According to the Yokota and Tanimoto approach, the decay law is approximated by the expression:

$$I(t) = A\exp\left\{\frac{-t}{\tau_{\mathrm{o}}} - 2B\delta\left(\frac{-t}{\tau_{\mathrm{o}}}\right)^{1/2}\right\} \qquad [15]$$

where $\delta = [\mathrm{C}]/[\mathrm{C}]_{\mathrm{o}}$ (C_{o} is the critical molar concentration of acceptor molecules given by: $[\mathrm{C}_{\mathrm{o}}] = 3000/2\pi^{3/2}NR_{\mathrm{o}}^3$, and B is a complex, time-dependent function given by: $B = (1 + 10.87x + 15.5x^2 / 1 + 8.743x)^{3/4}$ with $x = D\alpha^{-1/3}t^{2/3}$, $\alpha = R_{\mathrm{o}}^6/\tau_{\mathrm{o}}$ and D is the 'diffusion coefficient' representing energy migration.

Emission depolarization is often exploited in the study of energy migration in polymers. Only molecules with some vector component of their absorption transition dipole lying parallel to the plane of polarization of the incident radiation, will absorb incoming radiation. In the absence of angular motion of the absorbing chromophore within the lifetime of the excited state, emission from the same chromophore will also be highly polarized; however, upon each step in the energy migration process, some loss of polarization will be incurred (Figure 7). The extent of depolarization of emission is usually referred to by the term fluorescence anisotropy, $r(t)$ (eqn [16]):

$$r(t) = \frac{I_{\parallel}(t) - I_{\perp}(t)}{I_{\parallel}(t) + 2I_{\perp}(t)} \qquad [16]$$

The rate of energy migration can be estimated directly through observation of the time-dependent decay of $r(t)$ in the absence of other fluorescence depolarizing processes, such as molecule rotation (Figure 8).

One of the important properties which characterizes excitation energy transport in a system of n identical molecules, which are distributed randomly in space, is the Green function, $G(r,t)$. The Green function can be thought of as the time-dependent probability of finding an excitation at position r on the polymer chain at time t. It has been shown that the overwhelming contribution to fluorescence polarization is due to fluorescence from the sites that were initially excited. Thus, $G^{\mathrm{s}}(r,t)$ is directly related to the fluorescence anisotropy if the polarization memory of secondary excited molecules can be neglected. The $r(t)$ for donor–donor energy transfer is then given by:

$$r(t) = \alpha r_{\mathrm{rot}}(t)G^{\mathrm{s}}(t) \qquad [17]$$

in which r_{rot} describes the anisotropy in the absence of energy transfer, i.e., in the presence of chromophore

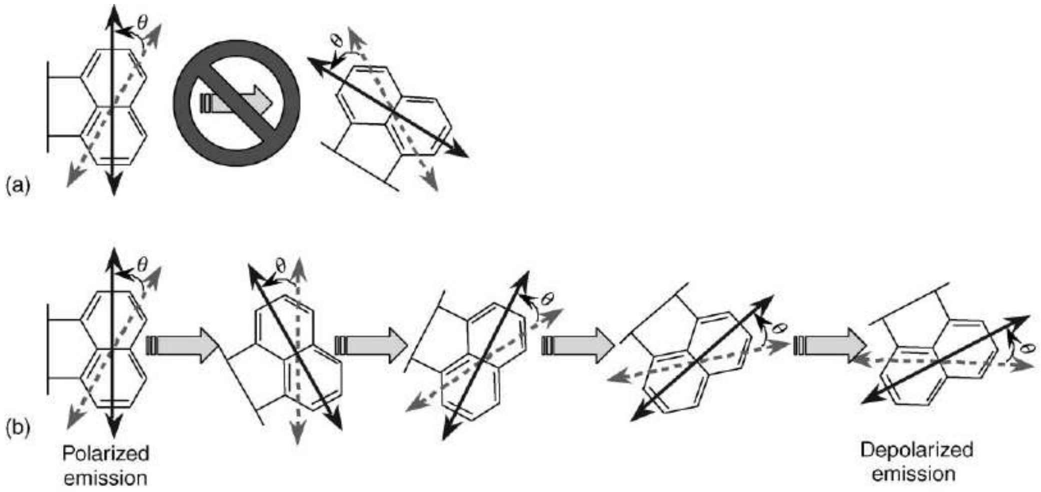

Figure 7 Schematic representation of energy migration. θ is the inherent angle between the donor emission (dotted arrow) and acceptor absorption (solid arrow) transition dipoles of the chromophore (a) orthogonal donor emission and acceptor absorption dipoles – no energy transfer/migration, emission depolarized only due to θ. (b) some vector component of the absorption dipole of the acceptor is parallel with that of the donor emission dipole – this and subsequent energy transfer steps can occur resulting in loss of emission polarization.

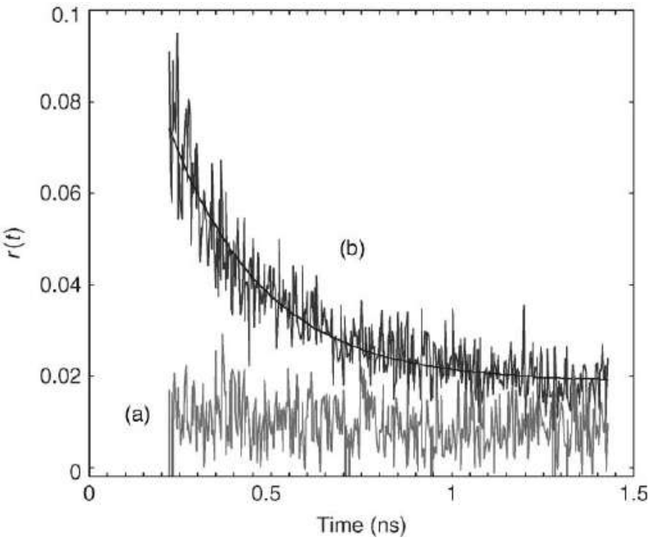

Figure 8 Temporal decays of $r(t)$ from (a) a polymer in which emission depolarization due to energy migration is complete on ultrafast time-scales, and (b) a polymer in which the rate of energy migration has been inhibited. Emission collected from polymers in frozen glass to exclude polymer motion contributions to emission depolarization.

rotation only, and α is the anisotropy from time-independent sources of depolarization.

The determination of the functional form of $G^s(r,t)$, resulting from excitation energy migration in polymer systems, is a challenging theoretical problem to which a number of approaches have been applied. The Gochanour, Anderson and Fayer approach has had some success in arriving at an expression for $G^s(r,t)$, represented as the time decay of emission polarization curves.

Polymeric systems that undergo energy migration, yet contain no excimer forming or other nonmonomeric, energy-trapping sites, are rare. In the presence of singlet energy traps, the problem can be treated as being analogous to energy transfer (via the Förster mechanism), following a diffusive process in a similar manner to that described above.

Fredrickson and Frank have derived expressions for the temporal decays of the monomer and excimer fluorescence when one-dimensional electronic singlet energy migration is the major excimer-forming path, and segmental rotation and chromophore diffusion are considered. The Fredrickson and Frank model assumes inherently that excimers are formed predominantly via a nearest-neighbor mechanism, where rapid energy migration is at least an order of magnitude faster than any competing rotationally sampled pathways. The expressions for the fluorescence decay of both the monomer, $I_M(t)$, and excimer species, $I_E(t)$, are highly nonexponential, with the expression for $I_M(t)$ being given by eqn [18]:

$$I_M(t) = Q_{fm}k_m(1-q)^2 \times \exp[(4q^2W - k_m - k_{rot})t]\mathrm{erfc}[2q(Wt)^{1/2}] \qquad [18]$$

In eqn [18], Q_{fm} is the quantum efficiency for monomer fluorescence and k_m is a rate constant, which encompasses the total rate of monomer decay by radiative and nonradiative pathways in the absence of energy migration and rotational sampling. Similarly, k_{rot} represents the rate at which an excited monomer is transformed into an excimer by segmental rotation. The parameter W represents the rate constant for nearest-neighbor electronic energy transport between the donor (monomer) chromophores, averaged over both the transition moment orientation and the interchromophore separation associated with each dyad.

Energy transfer from pendant aromatic monomer units to a specific chemically bound acceptor chromophore can occur in some polymers following energy migration. This phenomenon is known as the 'antenna effect', which has commercial applications such as in the sensitization of photo-oxidation–reduction, i.e. reactions, in the stabilization of polymers against photodegradation or in the modeling of photosynthesis by light harvesting. The kinetic behavior of such processes is yet to be modeled definitively, but the approaches discussed above (and others) have been used with varying degrees of success to date.

List of Units and Nomenclature

Ångstrom (Å)	10^{-10} m
chromophore	An aromatic molecule that absorbs (and may emit) light
exciton	An excited state localized on a particular chromophore at any given time
femtoseconds	10^{-15} s
intermolecular	Occurring between molecules or polymer chains
intramolecular	Occurring within a single molecule or polymer chain
petawatts	10^{15} W
δ-pulse excitation	Excitation by an infinitely short laser pulse

See also

Chemical applications of lasers: Non-linear spectroscopies. **Lasers:** Excimer Laser. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed). **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Andrews DL and Demidov AA (1999) *Resonance Energy Transfer*. New York: Wiley.

Birks JB (1970) *Photophysics of Aromatic Molecules*. London: Wiley-Interscience.

Fleming GR (1986). *Chemical Applications of Ultrafast Spectroscopy*, vol. 13. New York: Oxford University Press.

Fredrickson GH and Frank CW (1983) Nonexponential transient behavior in fluorescent polymeric systems. *Macromolecules* 16: 572–577.

Ghiggino KP and Smith TA (1993) Dynamics of energy migration and trapping in photoirradiated polymers. *Progress in Reaction Kinetics* 18: 375–436.

Gochanour CR, Andersen HC and Fayer MD (1979) Electronic excited state transport in solution. *Journal of Chemical Physics* 70: 4254–4271.

Guillet JE (1985) *Polymer Photophysics and Photochemistry: An Introduction to the Study of Photoprocesses in Macromolecules*. New York: Cambridge University Press.

Lakowicz JR (1999) *Principles of Fluorescence Spectroscopy*, 2nd edn. New York: Plenum.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*. London: Academic Press.

Sienicki K, Itagaki H and Mattice WL (1989) On the theory of concentration depolarization of fluorescence in one- and two-component systems for multipole interactions in one, two, and three dimensional medium. *Journal of Chemical Physics* 91: 4515–4521.

Van der Meer BW, Coker G and Chen S-YS (1994) *Resonance Energy Transfer: Theory and Data*. New York: VCH.

Van Hecke GR and Karukstis KK (1998) *A Guide to Lasers in Chemistry*. Boston, MA: Jones and Bartlett.

Yokota M and Tanimoto O (1967) Effects of diffusion on energy transfer by resonance. *Journal of the Physics Society of Japan* 22: 779–784.

TOMOGRAPHY

Contents

Optical Coherence Tomography
Tomography and Optical Imaging

Optical Coherence Tomography

S A Boppart, University of Illinois, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical coherence tomography (OCT) is an emerging imaging technique for a wide range of biological, medical, and material investigations. OCT was initially developed for imaging biological tissue because it permits the imaging of tissue microstructure *in situ*, yielding micron-scale image resolution without the need for excision of a specimen and tissue processing. OCT is analogous to ultrasound B-mode imaging except that it uses low-coherence light rather than sound and performs cross-sectional imaging by measuring the backscattered intensity of light from structures in tissue. The OCT image is a gray-scale or false-color two-dimensional representation of backscattered light intensity in a cross-sectional plane. The OCT image represents the differential backscattering contrast between different tissue types on a micron scale. Because OCT performs imaging using light, it has a one to two order-of-magnitude higher spatial resolution than ultrasound and does not require contact with the specimen or sample.

OCT was originally developed and demonstrated in ophthalmology for high-resolution tomographic imaging of the retina and anterior eye. Because the eye is transparent and is optically accessible, it is well suited for diagnostic OCT imaging. OCT is promising for the diagnosis of retinal disease because it can provide images of retinal pathology with 10 micron resolution, almost one order-of-magnitude higher than previously possible using ultrasound. Recently, OCT has been applied for imaging in a wide range of nontransparent tissues. In tissues other than the eye, the imaging depth is limited by optical attenuation due to scattering and absorption. Ophthalmic imaging is typically performed at 800 nm wavelengths. However, because optical scattering decreases with increasing wavelength, OCT imaging in nontransparent tissues is possible using longer near-infrared wavelengths. In most tissues, imaging depths of 2–3 mm can be achieved using a system detection sensitivity of 100 to 110 dB. Imaging studies have been performed in a wide range of biological, medical, and surgical specialties including developmental biology, cardiology, gastroenterology, urology, and neurosurgery. High-resolution OCT using short coherence length, short pulsed light sources has also been demonstrated and axial resolutions less than 2 μm have been achieved. High-speed real-time OCT at image acquisition rates of 4 to 8 frames per second have also been demonstrated. OCT has been extended to perform Doppler imaging of blood flow and birefringence imaging to investigate tissue injury. Different imaging delivery systems including transverse imaging catheters and forward imaging devices have been developed to enable internal body OCT imaging. Most recently, an OCT catheter has been combined with endoscope-based delivery to perform *in vivo* imaging in animal models and human patients.

This chapter will provide an overview of the OCT technology, beginning with the basic principles of operation. Technological advancements over the last decade have enabled OCT to transition from laboratory- to clinical-based imaging. These enabling advancements will be discussed, followed by representative applications in biology, medicine, surgery, and materials.

Principles of Operation

OCT is based on optical ranging, the high-resolution, high dynamic range detection of backscattered light. In contrast to ultrasound, the velocity of light is extremely high. Therefore, the echo time delay of reflected light cannot be measured directly and interferometric detection techniques must be used. One method for measuring echo time delay is to use low coherence interferometry or optical coherence domain reflectometry. Low-coherence interferometry was first developed for measuring reflections in fiber optics and optoelectronic devices.

The echo time delay of reflected light is measured by using a Michelson-type interferometer (Figure 1). The light reflected from the specimen or sample is interfered with light that is reflected from a reference path of known path length. Interference of the light reflected from the sample arm and reference arm of the interferometer can occur only when the optical path lengths of the two arms match to within the coherence length of the optical source. As the reference arm optical path length is scanned, different echo delays of backscattered light from within the sample are measured. The interference signal is detected at the output port of the interferometer, electronically band-pass filtered, demodulated, digitized, and stored on a computer. The position of the incident beam on the specimen is typically scanned in the transverse direction and multiple axial measurements are performed. This generates a two-dimensional data array that represents the optical backscattering through a cross-sectional plane in the specimen (Figure 2). The logarithm of the backscatter intensity is then mapped to false-color or gray-scale and displayed as an OCT image. The interferometer in an OCT instrument can be implemented using a fiber optic coupler and beam-scanning can be performed with small mechanical galvanometers in order to yield a compact and robust system (Figure 3).

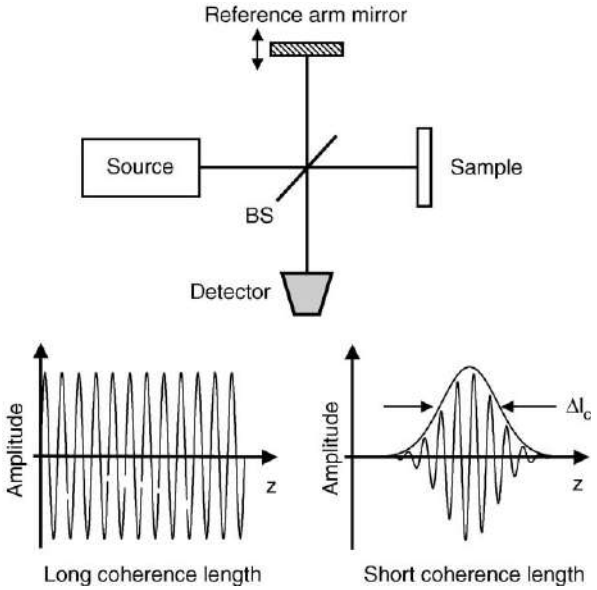

Figure 1 Schematic illustrating the concept of low-coherence interferometry. Using a short coherence-length light source and a Michelson-type interferometer, interference fringes are observed only when the path lengths of the two interferometer arms are matched to within the coherence length of the optical source. Abbreviations: BS, beam splitter; Δl_c, coherence length.

In contrast to conventional microscopy, the axial resolution in OCT images is determined by the coherence length of the light source. The axial point spread function of the OCT measurement as defined by the signal detected at the output of the

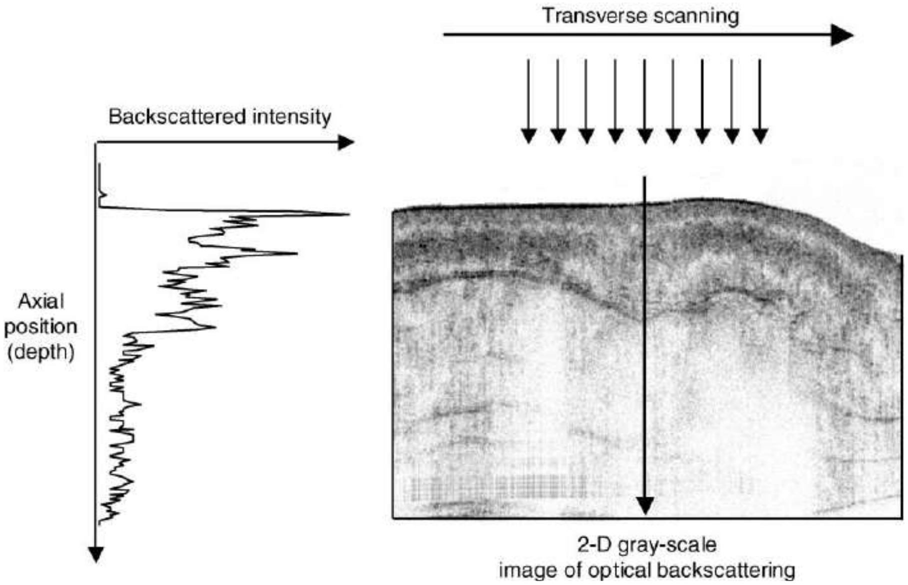

Figure 2 OCT image generation. The OCT image is typically acquired by performing axial measurements of optical backscatter at different transverse positions on the specimen and displaying the resulting two-dimensional data set as a gray-scale (or false-color) image.

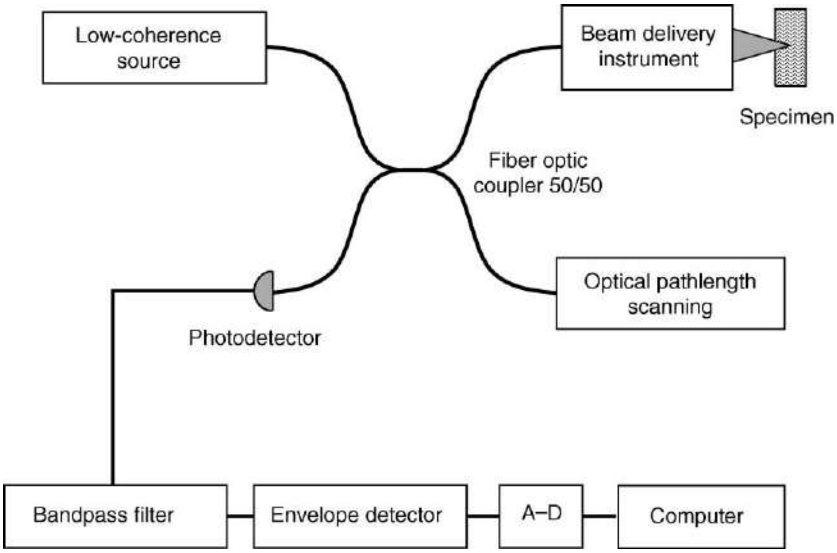

Figure 3 Schematic representation of an OCT system implemented using fiber optics. The Michelson interferometer is implemented using a fiber-optic coupler (beam splitter). The sample arm contains a beam-delivery instrument and the reference arm contains a mechanism for rapidly scanning the optical pathlength. The output of the interferometer is detected with a photodiode, filtered, demodulated, analog-to-digital (A–D) converted, and processed by a computer.

interferometer is the electric-field autocorrelation of the source. The coherence length of the light is the spatial width of the field autocorrelation and the envelope of the field autocorrelation is equivalent to the Fourier transform of its power spectrum. Thus, the width of the autocorrelation function, or the axial resolution, is inversely proportional to the width of the power spectrum. For a source with a Gaussian spectral distribution, the axial resolution Δz is given:

$$\Delta z = \frac{2\ln 2}{\pi} \cdot \frac{\lambda^2}{\Delta\lambda}$$

where Δz and $\Delta\lambda$ are the full-widths-at-half-maximum of the autocorrelation function and power spectrum respectively and λ is the source central wavelength. Figure 4 illustrates the dependence of the coherence length (axial resolution) on the bandwidth of the optical source. To achieve high axial resolution requires broad bandwidth optical sources. Resolution is also improved by using shorter wavelengths, however, shorter wavelengths are scattered and absorbed more in biological tissue.

The transverse resolution in an OCT imaging system is determined by the focused spot size in analogy with conventional microscopy and is given by:

$$\Delta x = \frac{4\lambda}{\pi} \cdot \frac{f}{d}$$

Where d is the spot size on the objective lens and f is its focal length. High transverse resolution can be obtained by using a large numerical aperture and focusing the beam to a small spot size. The transverse resolution is also related to the depth of focus or the confocal parameter $2z_R$ (two times the Raleigh range).

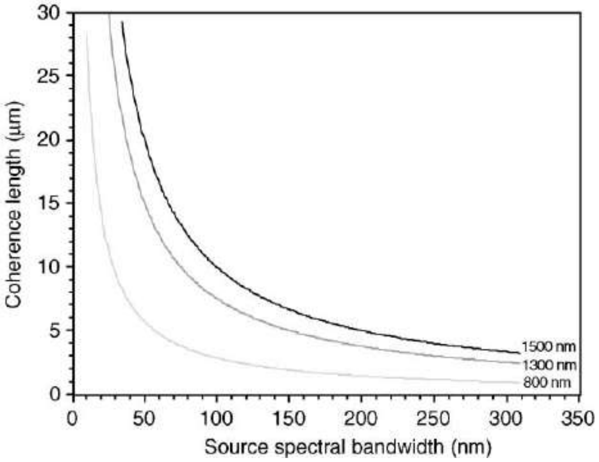

Figure 4 Dependence of coherence length (axial resolution) on the bandwidth of the optical source. Curves are shown for imaging wavelengths of 800 nm, 1300 nm, and 1500 nm. To achieve micron-scale resolution imaging, broad spectral bandwidths are necessary.

$$2z_R = \frac{\pi \Delta x^2}{2\lambda}$$

Thus, increasing the transverse resolution results in a reduced depth of field. Typically, the confocal parameter or depth of focus is chosen to match the desired depth of imaging. Increased resolution may also be obtained by using a higher numerical aperture lens or objective and spatially tracking the focus through the specimen.

Finally, the detection signal-to-noise is given by the optical power backscattered from the sample divided by the noise equivalent bandwidth:

$$\mathrm{SNR} = 10\log\left(\frac{\eta}{\hbar\omega}\frac{P_{\mathrm{SAM}}}{\mathrm{NEB}}\right)$$

Depending upon the desired signal to noise performance, incident powers of 5–10 mW are typically required for OCT imaging of 250–500 square pixel images at several frames per second. If lower data acquisition speeds or signal-to-noise can be tolerated, power requirements can be reduced accordingly.

Technological Developments

Since the inception of OCT in the early 1990s, there has been rapid technological developments aimed at improving the imaging resolution, acquisition rate, and methods for beam delivery to the tissue or sample. Investigators have also explored other imaging methods using the principles of OCT to extract information from the tissue or sample. Some of these methods have included acquiring optical Doppler signals from moving scatterers or structures, obtaining images based on the polarization state of the returned light, and extracting spectroscopic information based on the local absorption or scattering properties of the tissue.

Optical Sources / High-Resolution Imaging

The majority of OCT imaging systems to date have used superluminescent diodes (SLDs) as low coherence light sources. SLDs are manufactured with a similar structure to that of laser diodes, but commonly end facets are angle-cleaved to suppress lasing. These device structures do not support oscillation modes and generate output based on amplified spontaneous emission, resulting in emission spectra that are broader than laser diodes. SLDs are commercially available at a range of wavelengths including 800 nm, 1.3 μm, and 1.5 μm and are attractive because they are compact, have high efficiency, and low noise. However, output powers are typically limited to only a few milliwatts which limits fast real-time acquisition rates, and the available bandwidths are relatively narrow, permitting imaging with 10–15 micron resolution. Recent advances in short-pulse solid-state laser technology make these sources attractive for OCT imaging in research applications. Femtosecond solid-state lasers can generate tunable, low-coherence light at powers sufficient to permit high-speed OCT imaging. Short pulse generation has been achieved across the full wavelength range in titanium:sapphire ($Ti:Al_2O_3$) from 0.7 μm to 1.1 μm and over more limited tuning ranges near 1.3 μm and 1.5 μm in chromium: forsterite ($Cr^{4+}:Mg_2SiO_4$) and chromium:yttrium-aluminum-garnet (Cr^{4+}:YAG) lasers, respectively. OCT imaging with resolutions of 1 μm and 5 μm has been demonstrated at 800 nm and 1.3 μm respectively using $Ti:Al_2O_3$ and $Cr^{4+}:Mg_2SiO_4$ sources. More compact and convenient sources such as superluminescent fiber sources, are currently under investigation. The titanium:sapphire laser technology is routinely used in multi-photon microscopy applications for its high peak intensities to enable multi-photon absorption and subsequent emission of fluorescence from exogenous fluorescent contrast agents. Combined OCT and multi-photon microscopy has been used to provide complementary image data using a single optical source.

Fast Scanning

The short-pulse solid-state laser technology not only provides broad spectral bandwidths for high-resolution OCT imaging, but also higher output powers to enable fast real-time OCT imaging. Higher incident powers are required to maintain equivalent signal-to-noise ratios when scanning at a faster rate. Linearly translating a reference arm mirror is problematic at high rates and provides axial scan frequencies of approximately 100 Hz, depending on the mirror size and the translating galvanometer. Several investigators have utilized rotating glass cubes, piezoelectric modulators, and multi-pass optical cavities to increase axial scan rates while maintaining scan ranges of 1–2 mm. An optical delay based on the principles used in femtosecond pulse shaping has been demonstrated for OCT. This delay line spectrally disperses the reference arm beam with a grating. The dispersed beam is then focused by a lens on to a rotating mirror mounted on a galvanometer. The mirror, located in the Fourier-transform plane of the lens, imparts a wavelength-dependent phase shift on the light. Subsequently, when re-coupled back into the interferometer, this phase-shift is equivalent to a time-delay in the time domain. The use of high-speed resonant galvanometers has permitted axial scan rates as high as 8 kHz over scan ranges of several millimeters. Depending on the image pixel size (number of axial scans within each image), this scan rate can provide video-rate OCT imaging (30 frames per second). An example of high-speed *functional* OCT imaging is shown in Figure 5. Images of a

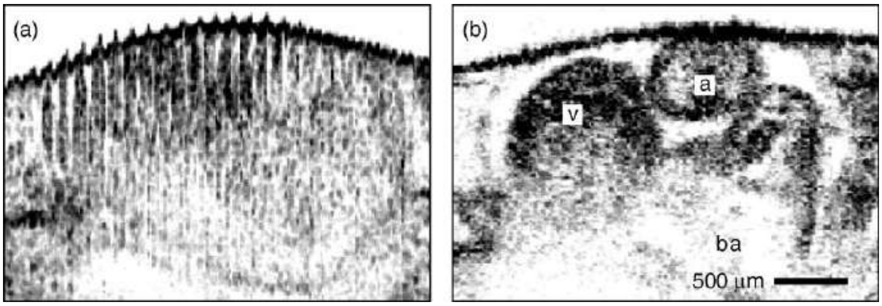

Figure 5 Comparison of slow and fast OCT imaging of a beating *Xenopus* (African frog) tadpole heart. The OCT image in (a) was acquired in 30 seconds and shows oscillations and blurring from the beating heart. The image on the right (b) was acquired in 250 ms (4 frames per second), fast enough to capture the anatomy of the heart with minimal motion artifacts. Abbreviations: a, atrium; ba, bulbous arteriosus; v, ventricle. (Images reprinted with permission from Boppart SA, Tearney GJ, Bouma BE, *et al.* (1997) Noninvasive assessment of the developing *Xenopus* cardiovascular system using optical coherence tomography. *Proceedings of the National Academy of Sciences* 94: 4256–4261.)

beating *Xenopus laevis* (African frog) tadpole were compared at slow (30 s per image) and at fast (250 ms per image) acquisition rates. Artifacts due to cardiac motion were present at slow acquisition rates, but were minimized with acquisition rates of 4 frames per second. Functional cardiac parameters could be measured and high-speed processes such as chamber and valve function could be visualized in real-time.

Depth-priority scanning, as described above, is performed by rapidly varying the optical delay in the reference arm and collecting a single axial scan before translating the beam laterally and repeating this depth-scanning. An alternative method for generating OCT images is with transverse-priority scanning. This is equivalent to optical sectioning in confocal and multi-photon microscopy. Three-dimensional OCT imaging using transverse-priority scanning can be obtained by stepping the position of the reference arm mirror after each *en face* image is acquired. The OCT images produced from transverse-priority scanning can be correlated with confocal or multi-photon microscopy images, which are acquired with the same scanning method. This method can also utilize higher numerical aperture objective lenses to provide high transverse resolutions since a large depth of focus is not needed as in depth-priority OCT scanning. The combination of OCT with high-numerical aperture objectives has been termed optical coherence microscopy.

Doppler Optical Coherence Tomography

Conventional OCT detects the amplitude of the optical backscatter versus position within the sample. Doppler OCT is a technique which has been applied to measuring fluid flow within small capillary tubes and *in vivo* vessels in the skin, retina, and heart of small animal models and humans. This technique is based on digitally sampling the interference fringes that are produced from a moving scatterer within the sample or specimen and determining the Doppler frequency shift in the signal due to the moving scatterer. Figure 6 illustrates two- and three-dimensional optical Doppler data for fluid flow through a cylindrical tube. A two-dimensional cross-sectional profile of fluid flow through a silicon tube with a 600 µm inner-diameter is plotted in Figure 6. The acquired data (points) closely match the predicted profile (line) for these tube dimensions. Below the plot, one method of obtaining a three-dimensional fluid flow profile is illustrated. While imaging the tube in cross-section, the fluid flow velocity was altered with a perfusion pump. The detection filter bandwidth of the OCT electronics served a windowing function, detecting only flow velocities with frequencies within the detection bandwidth. For an unknown flow profile, the center frequency of the detection filter can be shifted to map the flow velocities present within the sample.

Beam Delivery

The OCT imaging technology is modular in design and a variety of optical instruments can be used to deliver the OCT beam to the tissue or sample. Because OCT is fiber-optic based, single optical fibers can be used to deliver the OCT beam and collect the reflected light. The OCT technology can readily be integrated into existing optical instruments such as research and surgical microscopes, ophthalmic slit-lamp biomicroscopes, and hand-held imaging probes.

Imaging penetration is determined by the optical absorption and scattering properties of the tissue or specimen. The imaging penetration for OCT ranges from tens of millimeters for transparent tissues such as the eye to less than 3 mm in highly-scattering tissues such as skin. To image highly-scattering tissues deep within the body, novel beam-delivery

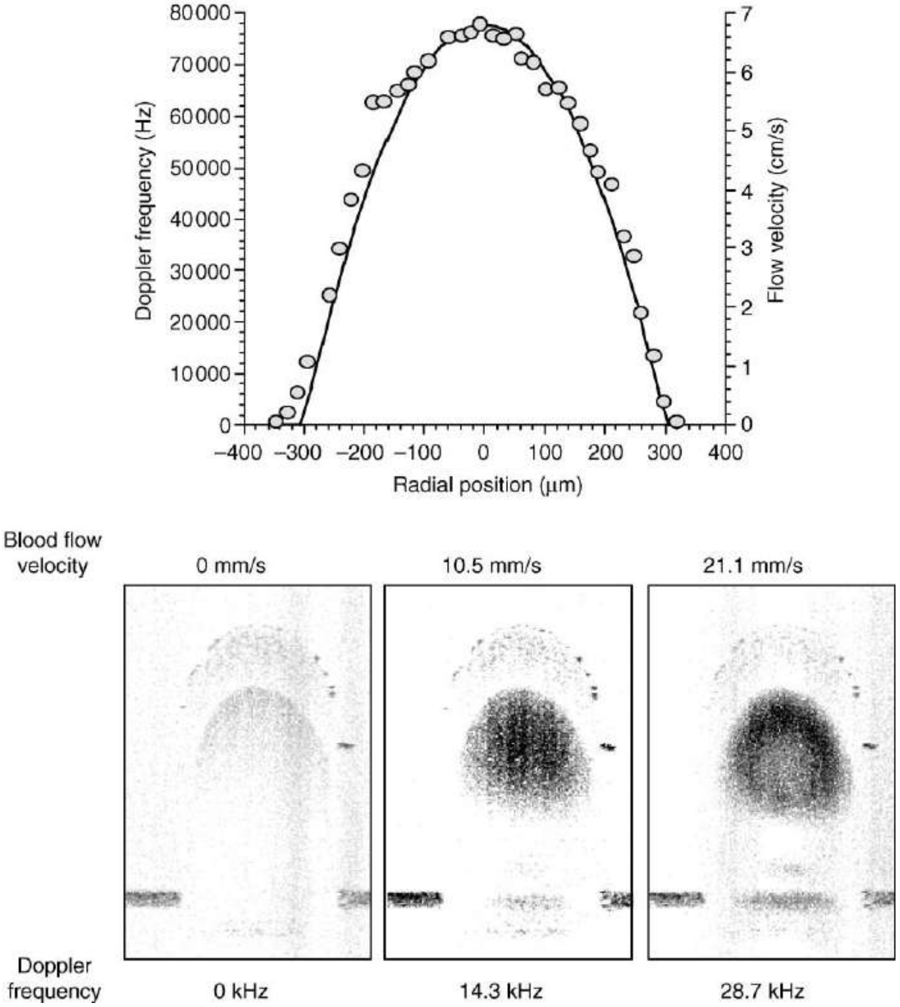

Figure 6 Doppler OCT. A two-dimensional plot of the fluid-flow velocity profile of flowing blood through a 600 μm inner-diameter tube is shown. Data points acquired using OCT closely follow the theoretical laminar flow profile (line). Two-dimensional Doppler OCT images of blood flow through the tube are shown below the plot. As fluid flow increases, only fluid-flow velocities within the bandwidth of the detection electronics are detected. This represents one method by which 3-D fluid flow can be imaged.

instruments have been developed to relay the OCT beam to the site of the tissue to be imaged. An OCT catheter has been developed for insertion into biological lumens such as the gastrointestinal tract. Used in conjunction with endoscopy, the 1 mm diameter catheter can be inserted through the working channel of the endoscope for simultaneous OCT and video imaging. Minimally-invasive surgical procedures utilize laparoscopes, which are long, thin, rigid optical instruments that permit video-based imaging within the abdominal cavity. Laparoscopic OCT imaging has been demonstrated by passing the OCT beam through the optical elements of a laparoscope. Deep solid-tissue imaging is possible with the use of fiber-needle probes. Small (400 μm diameter) needles housing a single optical fiber and micro-optic elements can be inserted into solid tissues and rotated to acquire OCT images. Recently, microfabricated micro-electro-optical-mechanical systems (MEOMS) technology has been used to miniaturize the OCT beam scan mechanism.

Applications

Developmental Biology

OCT has been demonstrated in the field of developmental biology as a method to perform high-resolution, high-speed imaging of developing morphology and function. Cellular-level imaging is possible, providing a non-invasive technique for visualizing cellular processes such as mitosis and migration. Imaging studies have been performed on several standard biological animal models commonly employed in developmental biology investigations including *Rana pipiens* (Leopard frog), *Xenopus laevis*

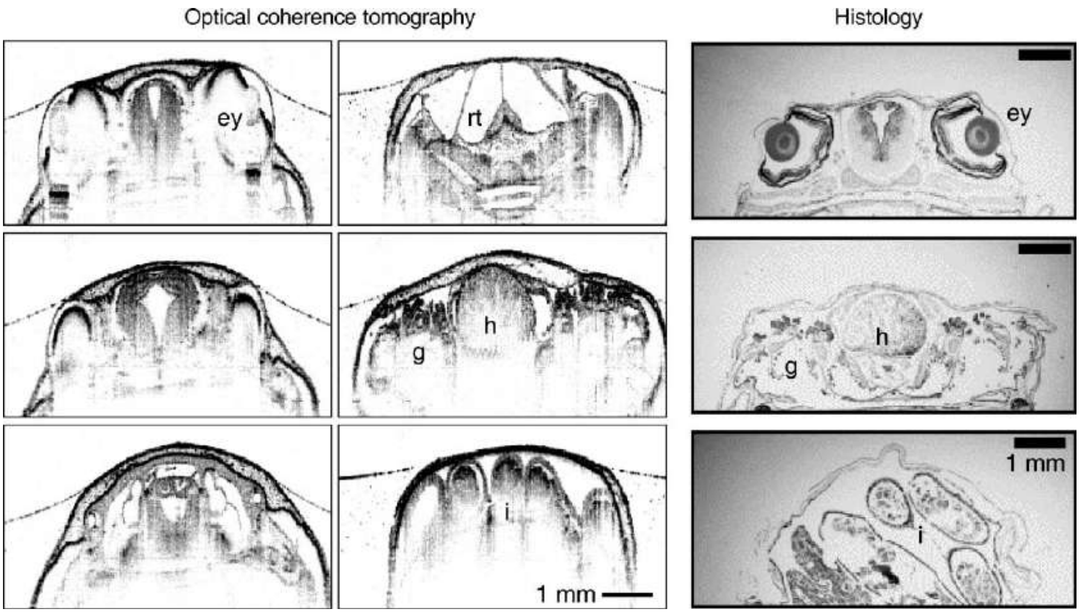

Figure 7 OCT images and corresponding histology of a Stage 49 (12 day) *Rana pipiens* (Leopard frog) tadpole. OCT images in the left and middle columns were acquired with the OCT beam incident from the dorsal and ventral sides of the specimen, respectively. The corresponding histology in the right column illustrates strong correlations between OCT images and the actual tissue morphology. Abbreviations: ey, eye; g, gills; h, heart; i, intestines; rt, respiratory tract. (Images reprinted with permission from Boppart SA, Brezinski ME, Bouma BE, *et al.* (1996) Investigation of developing embryonic morphology using optical coherence tomography. *Dev. Biol.* 177: 54–64.)

(African frog), and *Brachydanio rerio* (zebrafish) embryos and eggs, and the murine (mouse) model.

A series of cross-sectional images acquired *in vitro* from the dorsal and ventral sides of a Stage 49 (12 day) *Rana pipiens* (Leopard frog) tadpole is shown in Figure 7. Features of internal architectural morphology are clearly visible in the images. The image of the eye differentiates structures corresponding to the cornea, lens, and iris. Internal morphology not accessible in one orientation due to the specimen size or shadowing effects can be imaged by reorienting the specimen and scanning in the same cross sectional image plane. With the OCT beam incident on the ventral side, images of the respiratory tract, ventricle of the heart, internal gills, and gastrointestinal tract were acquired.

These images can be compared with corresponding histology (Figure 7). Histological images are acquired by euthanizing the specimen, immersing the specimen in a chemical fixative, and physically sectioning thin (2–5 micron-thick) slices using a microtome. The slices are placed on a microscope slide, selectively stained to highlight particular features, and viewed with light microscopy. The correlations between OCT and histology images are strong, suggesting that OCT images can accurately represent the *in vivo* specimen morphology. The potential exists to repeatedly image specimens to quantify organo- and morphogenesis throughout development. Technologies such as OCT are likely to become increasingly important in functional genomics, relating genetic features to the morphology and function in living specimens.

OCT images represent the optical backscatter intensity from regions within the tissue or sample. Because OCT relies on the inherent optical scattering changes to produce imaging contrast, no exogenous contrast agents or fluorophores are necessary. This permits long-term sequential imaging of development *in vivo* without loss of specimen viability. Repeated images of a developing zebrafish embryo within its egg beginning immediately after fertilization and up until hatching have been demonstrated without loss of specimen viability or without developmental abnormalities. In this example, the zebrafish egg and embryo were semi-transparent and the use of OCT significantly complemented observations made using light microscopy. By imaging subtle differences in backscattering intensity, interfacial structural layers millimeters deep within specimens can be clearly delineated.

Previous OCT images have characterized morphological features within biological specimens. These structures are static even though they may have been acquired from *in vivo* specimens. *In vivo* imaging in living specimens, particularly in larger organisms and for medical diagnostic applications, must be performed at high speeds to eliminate motion artifacts within the images. Functional OCT imaging is the quantification of *in vivo* images which yield information characterizing the functional properties of the organ system or organism. High speed OCT permits both the positioning and manipulation of specimens as well as imaging in real time and is a powerful technology for functional imaging in developmental biology animal models.

Studies investigating normal and abnormal cardiac development have been frequently limited by an inability to access cardiovascular function within the intact organism. OCT has been demonstrated for the high-resolution assessment of structure and function in the developing *Xenopus laevis* (African frog) cardiovascular system (Figure 5). The morphology of the *in vivo* cardiac chambers is clearly delineated. Image acquisition rates are fast enough to capture the cardiac chambers in mid-cycle. With this capability, images can be acquired at various times during the cardiac cycle. These frames can be displayed in real-time to produce a movie illustrating the dynamic, functional behavior of the developing heart. OCT, unlike technologies such as computed tomography and magnetic resonance imaging, provides high-speed *in vivo* imaging, allowing quantitative dynamic activity, such as ventricular ejection fraction, to be assessed.

Cellular Imaging

Although previous studies have demonstrated *in vivo* OCT imaging of tissue morphology, most have imaged tissue at ~10–15 μm resolutions, which does not allow differentiation of cellular structure. The ability of OCT to identify the mitotic activity, the nuclear-to-cytoplasmic ratio, and the migration of cells has the potential to not only impact the fields of cell and developmental biology, but also impact medical and surgical disciplines for the early diagnostics of disease such as cancer.

The *Xenopus laevis* (African frog) tadpole has been used to demonstrate the feasibility of OCT for high-resolution *in vivo* cellular and subcellular imaging. Many of the cells in this common developmental biology animal model are rapidly dividing and migrating during the early growth stages of the tadpole, providing an opportunity to image dynamic cellular processes. Three-dimensional volumes of high-resolution OCT data have been acquired from these specimens throughout development. From this 3-D data set, cells undergoing division were identified and tracked in three dimensions. In a similar manner, 3-D data sets were acquired to track single melanocytes (neural crest cells) as they migrated through the living specimens. The ability of OCT to characterize cellular processes such as mitosis and migration not only are of interest in cell and developmental biology, but also have relevance for cancer diagnostics and tumor metastasis.

An example of cellular-level OCT imaging in these specimens is shown in Figure 8. This composite image (0.83 × 1 mm, 1800 × 1000 pixels) was acquired using a titanium:sapphire laser with a broad bandwidth (~260 nm). The axial and transverse resolution in this image is 1 μm and 5 μm, respectively (see Figure 4). Because the high transverse resolution reduced the depth-of-focus to 49 μm, separate OCT images were first acquired with the focus at different depths within the specimen. These images were then assembled to produce the composite image shown in Figure 8. This type of image construction is similar to C-mode ultrasound. Cellular features including cell membranes, nuclei, and nuclear morphology are clearly observed.

Medicine – Imaging Barrett's Esophagus

OCT performs *in situ* imaging and has the potential to be used as a screening technique of early pathological changes in patients. OCT imaging has the potential to be a general diagnostic in many organ systems,

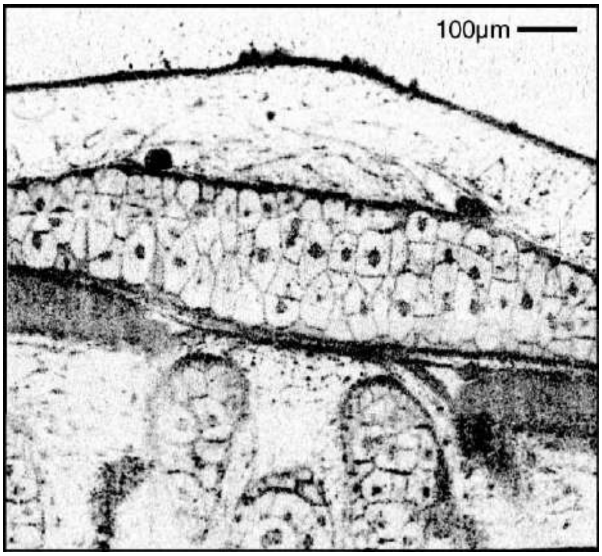

Figure 8 Cellular-level OCT imaging using a broad bandwidth titanium:sapphire laser. Axial and transverse image resolutions are 1 μm and 5 μm, respectively, enabling the visualization of cell membranes, nuclear morphology, and sub-cellular organelles. Cellular processes such as mitosis and migration can be visualized in real-time in living specimens. (Image reprinted with permission from Drexler W, Morgner U, Kartner FX, *et al.* (1999) *In vivo* ultrahigh resolution optical coherence tomography. *Optics Letters* 24: 1221–1223.)

particularly where current methods of screening by excisional biopsy are limited. Barrett's esophagus is a condition where the cells of the distal esophagus undergo a metaplastic change, resembling cells of the lower gastrointestinal tract. This condition is believed to be caused by chronic gastroesophageal reflux. Several studies have demonstrated that Barrett's esophagus is associated with a 30–125 times increased risk of developing adenocarcinoma. For this reason, endoscopic surveillance of Barrett's epithelium every 12–18 months is recommended. Endoscopic screening currently involves random four-quadrant biopsies every 1–2 cm along the length of suspect mucosa. However, excisional biopsy is prone to sampling errors and small foci of carcinoma or dysplasia may be missed. Because of the imprecision and high cost associated with screening, new methods are being developed to assess patients at increased risk. Endoscopic ultrasound catheters have been used for imaging the gastrointestinal tract with 50–100 μm resolution. However, these resolutions are insufficient to resolve early epithelial changes that occur in Barrett's esophagus and the pre-malignant changes that lead to adenocarcinoma of the esophagus.

The ability of OCT to differentiate normal and pathologic tissue is a central question being addressed by many research groups. Comparisons with histology show good correlations between OCT images and histological findings. The image resolution of conventional OCT (10–15 μm) is sufficient to differentiate architectural but not cellular morphology. Still, endoscopic OCT resolution can differentiate normal from Barrett's epithelium in real-time based on differences in epithelial architecture. Crypt- and gland-like structures that disrupt the relatively uniform layers of squamous epithelium can be readily identified, enabling differentiation between normal and Barrett's epithelium (Figure 9). The ability to differentiate normal from Barrett's epithelium suggests that the OCT technology could be used for screening applications.

Oncology – Identifying Tumors and Tumor Margins

OCT has been used to differentiate between the morphological structure of normal and cancerous tissue for a wide-range of tumors. The use of OCT to identify tumors and tumor margins *in situ* will represent a significant advancement for medical or image-guided surgical applications. OCT has been demonstrated for the detection of brain tumors and their margins with normal brain parenchyma, suggesting a role for guiding surgical resection. A hand-held surgical imaging probe was constructed for this application. The compact and portable probe permits OCT imaging within the surgical field while

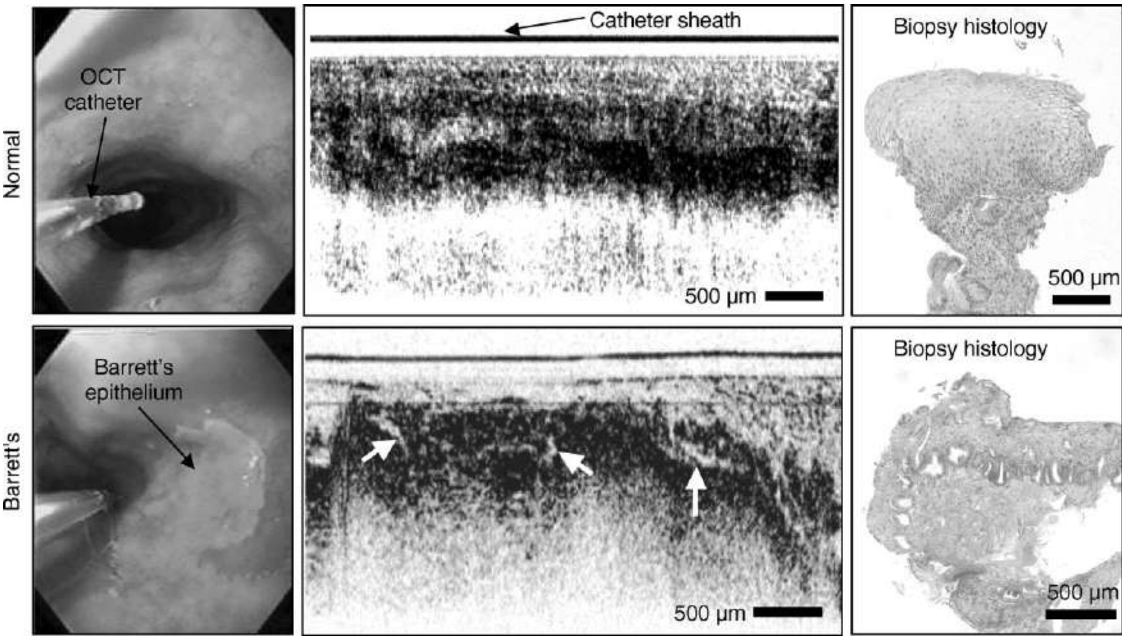

Figure 9 Endoscopic OCT imaging in humans. An OCT catheter was inserted down the working channel of an endoscope. Guided by video imaging, regions of normal and Barrett's esophagus were imaged. Corresponding biopsy histology is shown. OCT can identify Barrett's esophagus by the presence of gland- and crypt-like structures (arrows). (Images reprinted with permission from Li XD, Boppart SA, Van Dam J, *et al.* (2000) Optical coherence tomography: advanced technology for the endoscopic imaging of Barrett's esophagus. *Endoscopy* 32: 921–930.)

the OCT instrument can be remotely located in the surgical suite.

Figure 10 shows a specimen of outer human cerebral cortex with metastatic melanoma. The OCT images in Figure 10a,b were acquired through the tumor. These original images were threshold segmented to identify regions of high backscatter within the tumor. The original images were then overlaid with the segmented data and are shown in Figure 10c,d. The OCT images show increased optical backscattering in the region of the larger tumor (arrows). Smaller tumor lesions also appear within the image (arrows). A shadowing effect is observed below each tumor site due to the increased optical backscatter and the subsequent loss of optical power penetrating beneath the tumor. In Figure 10a,c, the boundary of the tumor can be identified. In Figure 10b,d, the tumor is identified below the surface of normal cortex. The histology in Figure 10e,f confirms the presence and relative size of the tumor.

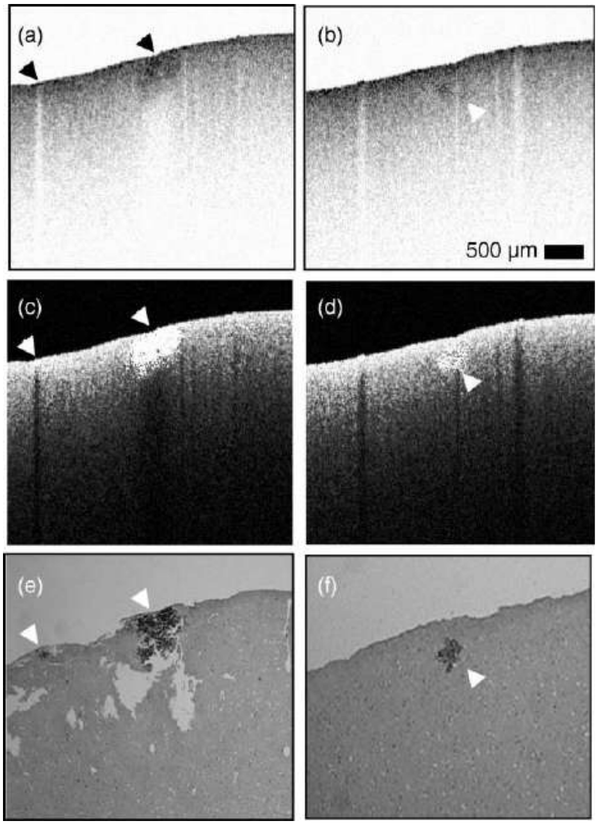

Figure 10 Cancer diagnostics. OCT imaging was used to image metastatic melanoma in an *ex vivo* specimen of human cortex (brain). Raw OCT images (a,b) were threshold-segmented (c,d) to highlight regions of increased optical backscatter and to identify tumor margins. OCT images correlate well with histological findings (e,f). (Images reprinted from Boppart SA, Brezinski ME, Pitris C and Fujimoto JG (1998) Optical coherence tomography for neurosurgical imaging of human intracortical melanoma. *Neurosurgery* 43: 834–841, with permission from Lippincott Williams and Wilkins.)

The image resolutions used to acquire the images in Figure 10 were as high as 16 μm, higher than any current ultrasound, CT, or MRI intraoperative imaging technique. This allowed the tumor-cortex interface and the extent of tumor below the surface to be defined with high resolution. At higher imaging resolutions, it may be possible to image individual tumor cells which have migrated away from the central tumor. OCT represents a new high-resolution optical imaging technology that has the potential for identifying tumors and tumor margins on the micron scale and in real-time. OCT offers imaging performance not achievable with current imaging modalities and may contribute significantly toward the surgical resection of neoplasms.

Image-Guided Surgery

The repair of vessels and nerves is necessary to restore function following traumatic injury. Although the repair of these sensitive structures is performed with the aid of surgical microscopes and loupes to magnify the surgical field, surgeons are limited to the *en face* view that they provide. A technique capable of subsurface, three-dimensional, micron-scale imaging in real-time would permit the intraoperative monitoring of microsurgical procedures. The capabilities of OCT for the intraoperative assessment of microsurgical procedures have been demonstrated. High-speed OCT imaging was integrated with a surgical microscope to performed micron-scale three-dimensional imaging on microsurgical specimens.

OCT has been used to image *in vitro* peripheral nerves and identify individual fascicles. Longitudinal tracking of the spatial orientation of rabbit peripheral nerve fascicles is demonstrated in Figure 11. Representative cross-sectional images of the peripheral nerve are shown in Figure 11a–d. For each slice, one fascicle was manually segmented, colored white, and tracked through the acquired volume of data. Forty images at 100 μm spacing were assembled for the 3-D projections shown in Figure 11e,f. The horizontally- and vertically-rotated projections of the peripheral nerve dramatically reveal the twisted path of the segmented fascicle along the longitudinal axis of the nerve. In addition, a branch in an unsegmented fascicle is observed in Figure 11f. The use of OCT to acquire multiple cross-sectional images and three-dimensionally reconstruct the peripheral nerve offers the opportunity to determine the relative diameters of individual fascicles and to longitudinally track their spatial orientation. OCT may also be useful at identifying and grading the degree of injury in nerves during surgical repair. These results have shown how 2-D OCT images and 3-D OCT projections can

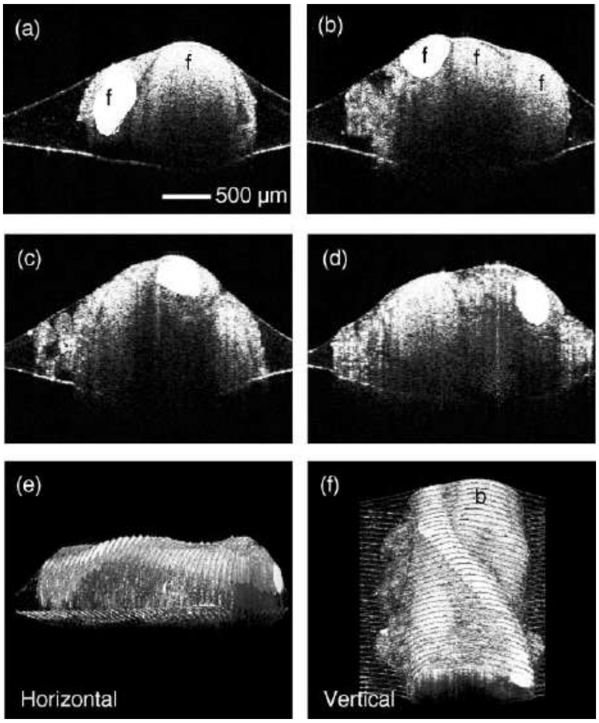

Figure 11 OCT image-guided microsurgery. Multiple cross-sectional OCT images were acquired through a resected rabbit peripheral nerve (a–d). A single nerve fascicle (f) was identified and segmented longitudinally along the length of the nerve. Three-dimensional reconstruction projections of the nerve in (e,f) reveal the tortuous path of the segmented fascicle as well as reveal a bifurcation (b). Intraoperative OCT can guide surgical repair and potentially identify regions of damaged tissue. (Images reprinted from Boppart SA, Bouma BE, Pitris C, *et al.* (1998) Intraoperative assessment of microsurgery with three-dimensional optical coherence tomography. *Radiology* 208: 81–86, with permission from the Radiology Society of North America.)

provide diagnostic feedback to assess microsurgical anastomoses. This previously unavailable diagnostic ability offers the potential to directly impact and improve patient outcome by incorporating high-speed, high-resolution intraoperative image-guidance during microsurgical procedures.

Surgical intervention requires visualization to identify tissue morphology, precision to avoid sensitive tissue structures, and continuous feedback to monitor the extent of the intervention. OCT may provide the technological advancements to improve the operative procedure. The feasibility of OCT to perform image-guided surgical intervention has been investigated. OCT has been used to monitor laser ablation therapy in real-time and may enable more precise control of laser delivery. An argon-laser ablation sequence of fresh *ex vivo* rat rectus abdominis muscle is shown in Figure 12. A pair of blood-filled vessels were located 1.5 mm below the tissue surface and centered within the OCT image. Because blood has a higher absorption coefficient than muscle at the 514 nm argon laser wavelength, coagulation and thermal heating of the blood occurred prior to thermal damage to the overlying tissue. The accompanying histology reveals both the thermally-damaged overlying tissue and coagulated blood present within the vessel pair.

These examples demonstrate the use of OCT for guiding and monitoring surgical intervention. The laser is only a representative interventional surgical technique for a wide range of instruments and techniques including scalpels, electrosurgery, radio-frequency, microwaves, and ultrasound ablation. OCT imaging was performed at 8 frames per second, fast enough to capture dynamic changes in the optical properties of the tissue during thermal ablation. These image sequences provided interesting insight into ablation mechanisms for a variety of tissue types. OCT can monitor the extent of thermal injury below the surface of the tissue by imaging the changes in optical backscatter. OCT imaging can therefore provide empiric information for dosimetry to minimize the extent of collateral injury. The use of OCT for guiding surgical interventions has the potential to improve intraoperative monitoring and more effectively control interventional procedures.

Materials

While the majority of OCT applications have been in the fields of biology and medicine. OCT has also been demonstrated in the non-biological areas of materials investigation, optical data storage, and microfluidic devices. The highly-scattering or reflecting optical properties of many materials prohibits deep imaging penetration using OCT. Many material defects, however, originate near the surface or at interfacial boundaries, making the use of OCT a possibility for inspection and quality control. OCT has been used to identify subsurface defects in ceramics and polymer composites. The optical ranging capabilities of OCT through scattering materials has been utilized for increasing the data storage capacity by assembling multiple layers of optically-accessible data.

The advancement of microfabrication techniques has led to increasingly complex microfluidic and bioMEM (biological micro-electro-mechanical) systems. Microstructures within microfluidic systems range from 10–1000 μm, within both the imaging depth and resolution of OCT. In addition, microfluidic systems are typically fabricated from transparent or semi-transparent substrates, facilitating imaging penetration to deeper three-dimensional features.

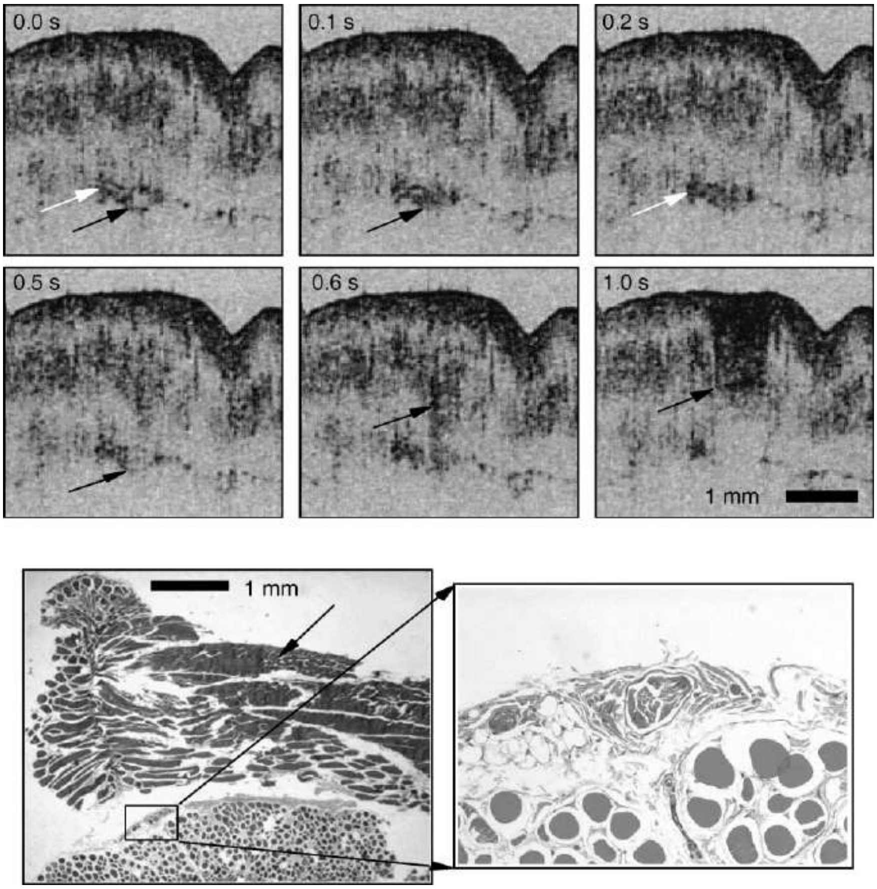

Figure 12 OCT image-guided laser ablation. High-speed OCT image sequence of argon-laser ablation of blood vessels in *ex vivo* rat muscle tissue. Thermal heating within the blood vessels (arrows, 0.0–0.5 s) induces coagulation prior to injury to superficial muscle tissue (0.6–1.0 s). Accompanying histology shows coagulated blood within the vessels.

OCT is not only capable of imaging three-dimensional microstructures at micron-scale resolutions, but is also capable of obtaining dynamic functional data from microfluidic systems. Microfluidic devices made from polymeric materials are attractive since they provide flexibility in the design of the channels, can be manufactured in large volumes, and are mostly bio-compatible. A three-dimensional imaging technique, such as OCT, will be instrumental in the analysis of these devices and the monitoring of device performance.

Conclusions

The capabilities of OCT offer a unique and informative means of imaging biological specimens and non-biological samples. The non-contact nature of OCT and the use of low-power near-infrared radiation for imaging causes few harmful effects on living cells or damage to materials. OCT imaging does not require the addition of fluorophores, dyes, or stains in order to improve contrast in images. Instead, OCT relies on the inherent optical contrast generated from variations in optical scattering and index of refraction. These factors permit the use of OCT for extended imaging over the course of hours, days, or weeks. OCT permits the cross-sectional imaging of tissue and samples and enables *in vivo* structure to be visualized in opaque specimens or in specimens too large for high-resolution confocal or light microscopy.

Imaging at cellular and subcellular resolutions with OCT is an important area of ongoing research. The *Xenopus* (African frog) developmental animal model has been commonly used because its care and handling are relatively simple while allowing cells with a high mitotic index to be assessed. Many of the cells observed were as large as 100 μm in diameter, but ranged in size down to a few microns, below the resolution of most OCT systems. Imaging human cells *in vivo* remains a challenge since differentiated human cells are 10 to 20 microns in size, too small for most OCT systems. With further advances in OCT technology, improved discrimination and imaging of more detailed structures should be possible. New laser sources at other wavelengths in the near-infrared can enhance tissue contrast as well as potentially provide functional information since tissue scattering and absorbance properties in specimens are wavelength dependent. Short coherence length short pulse laser sources have been used to achieve higher axial resolutions on the order of 1 to 3 microns. Unfortunately, unlike superluminescent diode source, these high-speed and high-resolution systems utilize femtosecond lasers that are relatively complex and costly. Developing compact and portable optical sources at near-infrared wavelengths, broad spectral bandwidths, and high output powers is an area of active research.

Optical coherence tomography provides high-resolution morphological, functional, and cellular information of biological, medical, and material specimens and samples. OCT represents a multi-functional investigative tool that not only complements many of the existing imaging technologies available today, but also is poised to become a major optical imaging modality.

List of Units and Nomenclature

Cr^{4+}:Mg_2SiO_4	chromium:forsterite
Cr^{4+}:YAG	chromium:yttrium-aluminum-garnet
CT	computed tomography
d	beam diameter
f	focal length
MRI	magnetic resonance imaging
NEB	noise equivalent bandwidth
OCT	optical coherence tomography
P_{SAM}	power from sample
SNR	signal-to-noise ratio
SLD	superluminescent diode
Ti:Al_2O_3	titanium:sapphire
z_R	Rayleigh range
3-D	three-dimensional
Δx	transverse resolution
$\Delta\lambda$	spectral bandwidth
Δz	axial resolution
η	quantum efficiency
λ	center wavelength

See also

Coherence: Coherence and Imaging; Overview. **Coherent Lightwave Systems**. **Interferometry:** Overview. **Fiber and Guided Wave Optics:** Optical Fiber Cables. **Imaging:** Imaging Through Scattering Media; Interferometric Imaging. **Terahertz Technology:** Coherent Terahertz Sources. **Tomography:** Tomography and Optical Imaging. **Microscopy:** Confocal Microscopy; Imaging multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Boppart SA, Tearney GJ, Bouma BE, *et al.* (1997) Noninvasive assessment of the developing *Xenopus* cardiovascular system using optical coherence tomography. *Proceedings of the National Academy of Sciences* 94: 4256–4261.

Boppart SA, Bouma BE, Pitris C, *et al.* (1998) *In vivo* cellular optical coherence tomography imaging. *Nature Medicine* 4: 861–864.

Boppart SA, Deutsch TF and Rattner DW (1999) Optical imaging technology in minimally invasive surgery. *Surgical Endoscopy* 13: 718–722.

Boppart SA, Brezinski ME and Fujimoto JG (2000) Optical Coherence Tomography Imaging. In: Tuan RS and Lo CW (eds) *Methods in Molecular Biology, Vol. 135: Developmental Biology Protocols*, vol. 1, pp. 217–233. Totowa, NJ: Humana Press, Inc.

Bouma BE and Tearney GJ (eds) (2001) *The Optical Coherence Tomography Handbook*. New York: Marcel Dekker Inc.

Chen Z, Zhao Y, Srinivas SM, *et al.* (1999) Optical Doppler tomography. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1134–1142.

Fujimoto JG, Brezinski ME, Tearney GJ, *et al.* (1995) Biomedical imaging and optical biopsy using optical coherence tomography. *Nature Medicine* 1: 970–972.

Fujimoto JG, Pitris C, Boppart SA and Brezinski ME (2000) Optical coherence tomography: An emerging technology for biomedical imaging and optical biopsy. *Neoplasia* 2(1–2): 9–25.

Huang D, Swanson EA, Lin CP, *et al.* (1991) Optical coherence tomography. *Science* 254: 1178–1181.

Masters BR (ed.) (2001) *Selected Papers on Optical Low-Coherence Reflectometry & Tomography*. SPIE Milestone Series, Vol. MS 165, Thompson BJ (general ed.) Bellingham, WA: SPIE Optical Engineering Press.

Puliafito CA, Hee MR, Schuman JS and Fujimoto JG (1995) Optical Coherence Tomography of Ocular Diseases. Thorofare, NJ: Slack, Inc.

Schmitt M (1999) Optical coherence tomography (OCT): A review. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1205–1215.

Tadrous PJ (2000) Methods for imaging the structure and function of living tissues and cells: 1. Optical coherence tomography. *Journal of Pathology* 191(2): 115–119.

Tearney GJ, Brezinski ME, Bouma BE, *et al.* (1997) *In vivo* endoscopic optical biopsy with optical coherence tomography. *Science* 276: 2037–2039.

Welzel J (2001) Optical coherence tomography in dermatology: a review. *Skin Research and Technology* 7(1): 1–9.

Tomography and Optical Imaging

Z Chen, University of California, Irvine, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

'Tomos' means section, in Greek. A tomogram is a sectional image of the internal structure of a three-dimensional object. One of the challenges for tomographic imaging of biological tissue using visible or near-infrared light is tissue optical scattering. When photons propagate inside a tissue, they undergo multiple scattering. The mean free path of photons in the skin, for example, is 5 μm. If one focuses the light 1 mm below the skin surface, most photons undergo multiple scattering and never reach the target point. Less than one out of a million photons will be able to reach the target, undergo a single backscattering, and return to the detector (Figure 1). Most photons that are backscattered and reach the detector are multiple scattered photons that carry little information on the target point. The challenge for tomographic imaging of tissues is to separate the single backscattered photons, that carry information about the target point, from multiple scattered photons.

Several optical tomography techniques have been developed. Each uses different techniques to discriminate the multiple scattering photons from the minimally scattered photons for imaging constructions. The confocal microscope, for example, uses a spatial filtering technique to reject multiple scattering photons because multiple scattered photons are more spread out in their angular distribution than the minimally scattered photons. Time-of-flight imaging uses a time gating technique to discriminate diffuse photons because single backscattered photons arrive earlier than diffuse photons. Optical coherence tomography (OCT) uses coherent gating to discriminate single scattered photons from multiple scattered photons. These gating techniques provide good discrimination of minimally scattered photons from multiple scattered photons and can form high-resolution sectional images, albeit with a limited imaging depth. Diffuse photons can propagate deep into the tissue and these photons can also be used for optical tomography. If a modulated photon density wave is propagated inside the tissue, the scattering and absorption properties of the tissue will modulate the phase and amplitude of the photon density wave. However, the resolution of optical tomography, based on diffuse photons, is usually much lower than that of the tomography constructed from single or minimally scattered photons. Three of the most important parameters for optical tomography are contrast, spatial resolution, and image depth.

Most optical tomography uses absorption, scattering, and fluorescent properties of materials as contrast. OCT, for example, uses variations in tissue scattering as contrast for tomographic imaging. However, in many instances, and especially during the early stages of disease, the change in tissue scattering properties between normal and diseased tissues is small and difficult to measure. One of the great challenges for extending applications of OCT is to find more contrast mechanisms that can provide physiological information in addition to morphological structure. This article will focus on optical tomography technologies that use nonconventional contrast mechanisms to obtain tomographic images. These include optical Doppler tomography (ODT), optical polarization tomography, and optical second-harmonic tomography. These optical tomography technologies are extensions of OCT with contrast mechanisms that can provide physiological information, in addition to morphological structure. ODT, also known as Doppler OCT, combines the Doppler principle with OCT to obtain high-resolution tomographic images of tissue structure and blood flow simultaneously. Optical polarization tomography, also known as polarization sensitive OCT (PS-OCT), combines polarization sensitive

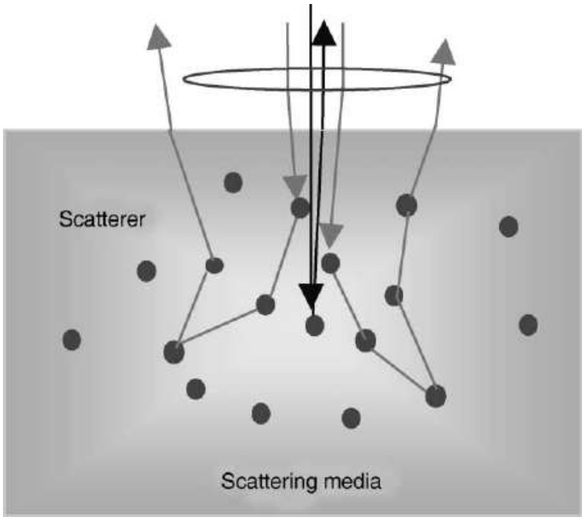

Figure 1 Optical imaging in scattering media.

detection with OCT, to determine tissue birefringence. Optical second-harmonic tomography, also known as second-harmonic optical coherence tomography (SH-OCT), combines second-harmonic generation (SHG) with coherence gating to obtain images with molecular contrast. These functional extensions of OCT provide clinically important information on tissue physiology, such as tissue blood perfusion, oxygen saturation, hemodynamics, and structural remodeling. Each provides several potential clinical applications, such as vasoactive drug screening, tissue viability and burn depth determination, tumor angiogenesis studies, tumor diagnosis, bleeding ulcer management, and ocular pathology evaluation.

Optical Doppler Tomography

ODT combines the Doppler principle with OCT to obtain high-resolution tomographic images of static and moving constituents in high scattering media. When light backscattered from a moving particle interferes with the reference beam, a Doppler frequency shift (f_D) occurs in the interference fringe:

$$f_D = \frac{1}{2\pi}(\mathbf{k}_s - \mathbf{k}_i)\cdot\mathbf{v} \qquad [1]$$

where $\mathbf{k}_i$ and $\mathbf{k}_s$ are wavevectors of incoming and scattered light, respectively, and $\mathbf{v}$ is the velocity vector of the moving particle (Figure 2). Since ODT measures the backscattered light, assuming the angle between flow and sampling beam is θ, the Doppler shift equation is simplified to:

$$f_D = \frac{2V\cos\theta}{\lambda_0} \qquad [2]$$

where λ_0 is the vacuum center wavelength of the light source.

The optical system of ODT is similar to that of OCT. The primary difference is in signal processing. Figure 3 illustrates an ODT instrument that uses a fiber optic Michelson interferometer with a broadband light source. Light from a broadband partial coherence source is coupled into a fiber interferometer by a 2 × 2 fiber coupler and then split equally into reference and target arms of the interferometer. Light backscattered from the turbid sample is coupled back into the fiber and forms interference fringes with the light reflected from the reference arm. A rapid-scanning optical delay line is used for group delay and axial scanning. Because this delay line can decouple the group delay from the phase delay, an electro-optical phase modulator is introduced to produce a stable carrier frequency. The interference fringe intensity signal is amplified, bandpass filtered, and digitized with a high-speed analog-to-digital converter. The signal processing is carried out at the same time as data are transferred to the computer, and real-time display can be accomplished with the use of a digital signal processing board.

To understand the signal processing of ODT, let us look at the fringe signal due to the moving particles.

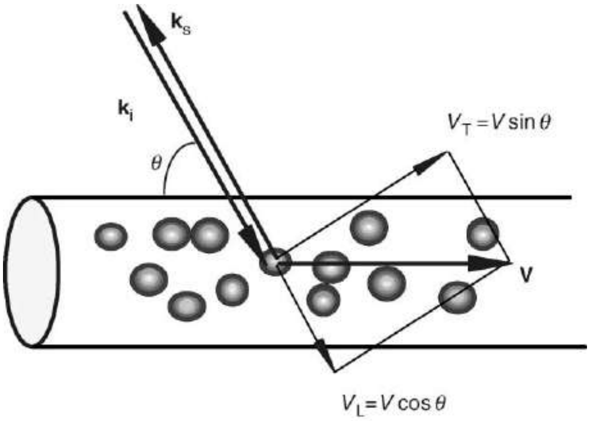

Figure 2 Schematic of flow direction and probe beam angle.

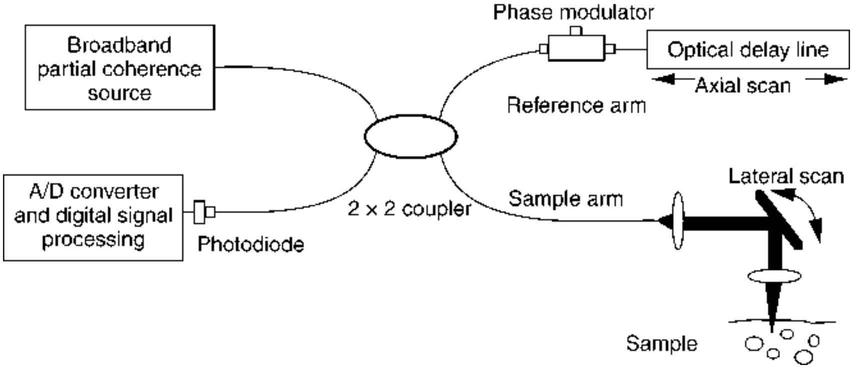

Figure 3 Schematic of ODT system consists of a fiber-based Michelson interferometer with a partially coherent light source.

If we denote $U(t)$ as a complex-valued analytic signal of a stochastic process representing the field amplitude emitted by a low coherent light source and $\bar{U}(\nu)$ as the corresponding spectral amplitude at optical frequency ν, the amplitude of a partially coherent source light coupled into the interferometer at time t, is written as a harmonic superposition:

$$U(t) = \int_0^\infty \bar{U}(\nu) e^{2\pi i \nu t} \, d\nu \qquad [3]$$

Because the stochastic process of a partially coherent light source is stationary, the cross spectral density of $\bar{U}(\nu)$ satisfies:

$$\langle \bar{U}^*(\nu)\bar{U}(\nu')\rangle = S_0(\nu)\delta(\nu - \nu') \qquad [4]$$

where $S_0(\nu)$ is the source power spectral density, and $\delta(\nu - \nu')$ is the Dirac delta function. Assuming that light couples equally into the reference arm and sample arm with spectral amplitude of $\bar{U}_0(\nu)$, the light coupled back to the detector from the reference, $\bar{U}_r(\nu)$, and sample, $\bar{U}_s(\nu)$, are:

$$\bar{U}_r(\nu) = e^{2\pi i\nu(2L_r + L_d)/c} K_r(\nu) e^{i\alpha_r} \bar{U}_0(\nu) \qquad [5]$$

$$\bar{U}_s(\nu) = e^{2\pi i\nu(2L_s + L_d)/c} K_s(\nu) e^{i\alpha_s} \bar{U}_0(\nu) \qquad [6]$$

where L_r and L_s are the optical pathlengths from the beamsplitter to the reference mirror and sample, respectively; L_d is the optical pathlength from the beamsplitter to the detector; and $K_r(\nu)e^{i\alpha_r(\nu)}$ and $K_s(\nu)e^{i\alpha_s(\nu)}$ are the amplitude reflection coefficients of light backscattered from the reference mirror and turbid sample, respectively.

The total power detected at the interferometer output is given by a time-average of the squared light amplitude:

$$P_d(\tau) = \langle |U_s(t) + U_r(t + \tau)|^2 \rangle \qquad [7]$$

where τ is the depth scan time delay. Combining harmonic expansions for $U_s(t)$ and $U_r(t)$ and applying eqn [4] when calculating the time-average, the total power detected is a sum of three terms representing reference (I_r), sample (I_s), and the interference fringe intensity ($\Gamma_{\mathrm{ODT}}(\tau)$):

$$\begin{aligned} P_d(\tau) &= \int_0^\infty (P_r(\nu) + P_s(\nu) + P_{\mathrm{ODT}}(\nu))\,d\nu \\ &= I_r + I_s + \Gamma_{\mathrm{ODT}}(\tau) \end{aligned} \qquad [8]$$

with

$$P_r(\nu) = S_0(\nu)|K_r(\nu)|^2 \qquad [9]$$

$$P_s(\nu) = S_0(\nu)|K_s(\nu)|^2 \qquad [10]$$

$$\begin{aligned} P_{\mathrm{ODT}}(\nu) = {} & 2S_0(\nu)K_r(\nu)K_s(\nu) \\ & \times \cos[2\pi\nu(\Delta_d/c + \tau) + \alpha_s(\nu) - \alpha_r(\nu)] \end{aligned} \qquad [11]$$

and

$$I_r = \int_0^\infty P_r(\nu)\,d\nu \qquad [12]$$

$$I_s = \int_0^\infty P_s(\nu)\,d\nu \qquad [13]$$

$$\Gamma_{\mathrm{ODT}}(\tau) = \int_0^\infty P_{\mathrm{ODT}}(\nu)\,d\nu \qquad [14]$$

When there is a moving particle, light scattered from this particle is equivalent to a moving phase front; therefore, Δ_d in eqn [11] can be written as:

$$\Delta_d = \Delta + 2\bar{n}V_z t \qquad [15]$$

where Δ is the optical pathlength difference between light in the sampling and reference arms; V_z is the velocity of a moving particle parallel to the probe beam; and $\bar{n}$ is the refractive index of flow media.

To simplify the computation, we assume α_s and α_r are constants over the source spectrum and can be neglected. The spectral domain fringe signal, $P_{\mathrm{ODT}}(\nu)$, is simplified to:

$$\begin{aligned} P_{\mathrm{ODT}}(\nu) = {} & 2S_0(\nu)K_r(\nu)K_s(\nu) \\ & \times \cos[2\pi\nu((\Delta + 2\bar{n}V_z t)/c + \tau)] \end{aligned} \qquad [16]$$

The corresponding time domain signal, $\Gamma_{\mathrm{ODT}}(\tau)$, is given by:

$$\begin{aligned} \Gamma_{\mathrm{ODT}}(\tau) = {} & 2\int_0^\infty S_0(\nu)K_r(\nu)K_s(\nu) \\ & \times \cos[2\pi\nu((\Delta + 2\bar{n}V_z t)/c + \tau)]\,d\nu \end{aligned} \qquad [17]$$

A comparison of eqns [16] and [17] shows that there is a Fourier transformation relation between spectral domain and time domain signals. Consequently, there are two methods to acquire the ODT signal: the time domain method and the spectral domain method.

In the time domain method, a delay line is incorporated in the reference arm to generate a delay. A spectrogram analysis or phase-resolved algorithm is then used to determine the Doppler frequency shift. In the spectral domain method, the reference mirror is fixed, and there is no depth scan ($\tau = 0$). The spectral domain fringe signal, $P_{\mathrm{ODT}}(\nu)$, is obtained either by a spectrometer at the detection arm, or by a frequency sweeping light source. The time domain signal, $\Gamma_{\mathrm{ODT}}(\tau)$, is determined from the spectral domain signal by a Fourier transformation.

Spectrogram Method

The spectrogram method uses time–frequency analysis to calculate the Doppler shift. Signal processing algorithms to obtain structural and velocity images from recorded temporal interference fringe intensity using the spectrogram method, are illustrated in Figure 4.

The spectrogram is an estimate of the power spectrum of the temporal interference fringe intensity ($\Gamma_{\mathrm{ODT}}(\tau_i, f_j)$) in the ith time delay window. The power spectrum of the temporal interference fringe at the ith pixel, corresponding to time delay τ_{i} in the structural and velocity images, is calculated by a short-time fast Fourier transformation (STFFT) or a wavelet transformation:

$$\hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_m) = |\mathrm{STFFT}(f_m; \Gamma_{\mathrm{ODT}}(\tau))|^2 \qquad [18]$$

where f_m is the discrete frequency value. A tomographic structural image is obtained by calculating the value of the power spectrum at the phase modulation frequency (f_0). Fluid flow velocity is determined from the Doppler frequency shift (f_{D}), which is the difference between the carrier frequency established by the optical phase modulation (f_0) and the centroid (f_{c}) of the measured power spectrum at the ith pixel:

$$V_{\mathrm{ODT}}(i) = \frac{\lambda_0 f_{\mathrm{D}}}{2\bar{n}\cos\theta} = \frac{\lambda_0(f_{\mathrm{c}} - f_0)}{2\bar{n}\cos\theta} \qquad [19]$$

where we have assumed, $\mathbf{k}_{\mathrm{s}} = -\mathbf{k}_{\mathrm{i}}$ and θ is the angle between $\mathbf{k}_{\mathrm{i}}$ and $\mathbf{v}$. The centroid of the measured power spectrum is determined by:

$$f_{\mathrm{c}} = \sum_m f_m \hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_m) \Big/ \sum_m \hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_m) \qquad [20]$$

Figure 5 shows the *in vivo* structural and blood flow images from a chick chorioallantoic membrane. The lumen wall, chorion membrane, and yolk sac membrane are observed in structural image (a). In the velocity image (b), static regions ($V = 0$) appear dark, while blood flowing at different velocities appears as different brightnesses on the gray scale. The velocity profile taken from a horizontal cross-section passing through the center of the vessel is shown in (c).

Phase-Resolved ODT Method

Although spectrogram methods allow simultaneous imaging of *in vivo* tissue structure and flow velocity, the velocity sensitivity is limited for high-speed imaging. When STFFT or wavelet transformation is used to calculate flow velocity, the resolution is determined by the window size of the Fourier transformation for each pixel. The minimum detectable Doppler frequency shift, f_{D}, varies inversely with the STFFT window size. Because pixel acquisition time is proportional to the STFFT window size, the image frame rate is limited by velocity resolution. Furthermore, spatial resolution is also proportional to the STFFT window size. Therefore, a large STFFT window size increases velocity resolution while decreasing imaging speed and spatial resolution. This coupling between velocity sensitivity, spatial resolution, and imaging speed prevents the spectrogram method from achieving simultaneously both high imaging speed and high velocity sensitivity, which are essential for measuring flow in small blood vessels where flow velocity is low.

Phase-resolved ODT overcomes the compromise between velocity sensitivity and imaging speed by using the phase change between sequential scans to construct flow velocity images. The phase information of the fringe signal can be determined from the complex analytical signal $\tilde{\Gamma}(t)$, which is determined through analytic continuation of the measured interference fringe function, $\Gamma(t)$, using a Hilbert transformation:

$$\tilde{\Gamma}(t) = \Gamma(t) + \frac{i}{\pi} P \int_{-\infty}^{\infty} \frac{\Gamma(\tau)}{\tau - t}\, \mathrm{d}\tau = A(t)\mathrm{e}^{i\phi(t)} \qquad [21]$$

where P denotes the Cauchy principle value, i is the complex number, and $A(t)$ and $\phi(t)$ are amplitude and phase of $\tilde{\Gamma}(t)$, respectively. Because the interference signal $\Gamma(t)$ is quasi-monochromatic, the complex

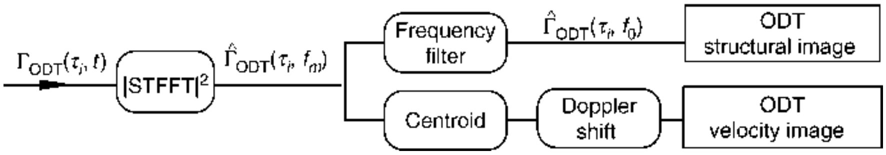
$\Gamma_{\mathrm{ODT}}(\tau_i, t)$
$|\mathrm{STFFT}|^2$
$\hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_n)$
Frequency filter
$\hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_0)$
ODT structural image
Centroid
Doppler shift
ODT velocity image

Figure 4 Signal processing algorithms for ODT structural and velocity images.

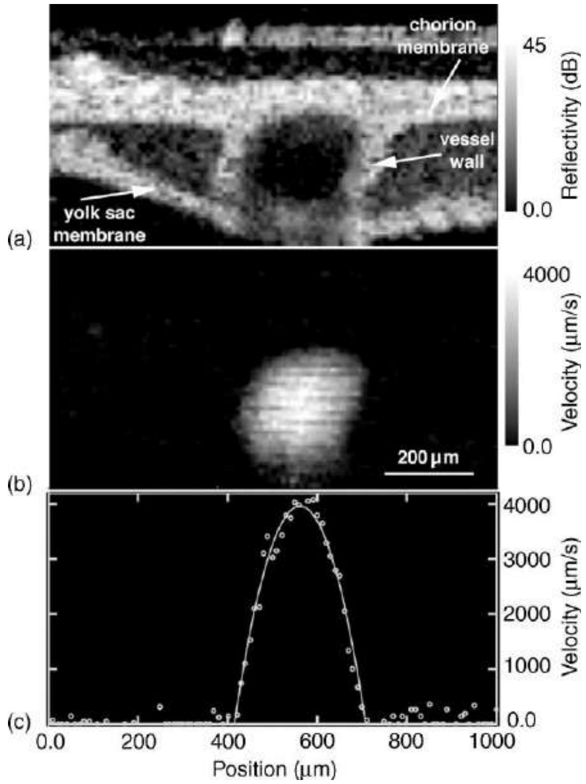

Figure 5 ODT images of *in vivo* blood flow in a CAM vein. (a) Structural image, (b) velocity image, and (c) velocity profile taken from a horizontal cross-section passing through the center of the vein, where the open circles are experimental data and the solid line is a parabolic fit. Reproduced with permission from Chen Z, MilnerTE, Srinivas S, Wang XJ, Malekafzali A, van Gemert MJC and Nelson JS (1997) Noninvasive imaging of *in vivo* blood flow velocity using optical doppler tomography. *Optics Letters* 22: 1119–1121.

analytical signal can be determined by:

$$\tilde{\Gamma}(t) = 2\int_0^\infty \int_0^\tau \Gamma(t')\exp(-2\pi i v t')\,\mathrm{d}t' \times \exp(2\pi i v t)\,\mathrm{d}v \qquad [22]$$

where τ is the time duration of the fringe signal in each axial scan.

The Doppler frequency shift (f_n) at nth pixel in the axial direction is determined from the average phase shift between sequential A-scans. This can be accomplished by calculating the phase change of sequential scans from the individual analytical fringe signal:

$$f_n = \frac{\Delta\phi}{2\pi T} = \frac{1}{2\pi T}\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\left[\tan^{-1}\left(\frac{\mathrm{Im}\tilde{\Gamma}_{j+1}(t_m)}{\mathrm{Re}\tilde{\Gamma}_{j+1}(t_m)}\right) - \tan^{-1}\left(\frac{\mathrm{Im}\tilde{\Gamma}_{j}(t_m)}{\mathrm{Re}\tilde{\Gamma}_{j}(t_m)}\right)\right] \qquad [23]$$

Alternatively, the phase change can also be calculated by the cross-correlation method:

$$f_n = \frac{1}{2\pi T}\tan^{-1}\left(\frac{\mathrm{Im}\left[\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\tilde{\Gamma}_j(t_m)\cdot\tilde{\Gamma}^*_{j+1}(t_m)\right]}{\mathrm{Re}\left[\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\tilde{\Gamma}_j(t_m)\cdot\tilde{\Gamma}^*_{j+1}(t_m)\right]}\right) \qquad [24]$$

where $\tilde{\Gamma}_j(t_m)$ and $\tilde{\Gamma}^*_j(t_m)$ are the complex signals at axial time t_m corresponding to the jth A-scan and its respective conjugate; $\tilde{\Gamma}_{j+1}(t_m)$ and $\tilde{\Gamma}^*_{j+1}(t_m)$ are the complex signals at axial time t_m corresponding to the next A-scan and its respective conjugate; M is an even number that denotes the window size in the axial direction for each pixel; N is the number of sequential scans used to calculate the cross correlation; and T is the time duration between A-scans. Because T is much longer than the pixel time window within each scan used in the spectrogram method, high-velocity sensitivity can be achieved.

Phase-resolved ODT decouples spatial resolution and velocity sensitivity in flow images and increases imaging speed by more than two orders of magnitude, without compromising spatial resolution and velocity sensitivity. In addition, because two sequential A-line scans are compared at the same location, speckle modulations in the fringe signals cancel each other and, therefore, will not affect the phase difference calculation. Consequently, the phase-resolved method reduces the speckle noise in the velocity image. Furthermore, if the phase difference between sequential frames is used, the velocity sensitivity can be further increased. Real-time imaging with velocity sensitivity on the order of 10 μm/s has been demonstrated.

In addition to digital processing of the fringe signal using Hilbert transformation, the complex analytical signal can also be achieved through hardware implementation. Optical Hilbert transformation using polarization optics has been implemented for real-time phase-resolved ODT imaging. Real-time ODT imaging using hardware demodulation of the ODT signal has also been demonstrated.

Spectral Domain Phase-Resolved ODT Method

A schematic diagram of a spectral domain ODT system is shown in Figure 6. The signal from the Michelson interferometer is directly coupled to a

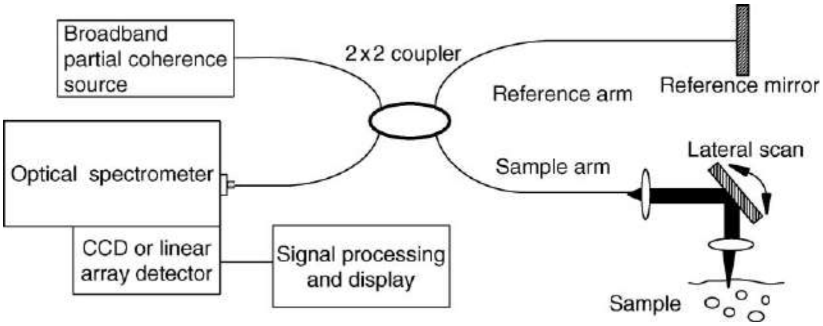
Broadband partial coherence source
2x2 coupler
Reference arm
Reference mirror
Optical spectrometer
Sample arm
Lateral scan
CCD or linear array detector
Signal processing and display
Sample

Figure 6 Schematic diagram of a spectral domain ODT instrument.

spectrometer that records the spectral fringe pattern, $P_{\mathrm{ODT}}(\nu)$. The temporal interference fringe can be calculated by a Fourier transformation of the spectral fringe pattern (eqn [17]). The Doppler shift can then be determined from the phase shift between sequential scans using the phase-resolved ODT algorithm (eqns [23] and [24]).

Spectral domain phase-resolved ODT has the advantage that no optical delay line is required. In addition, the spectral domain method can achieve high signal to noise ratio over time domain ODT. Since parallel detection can be implemented with a high-speed spectrometer, high-speed imaging acquisition is possible. Finally, because the dynamic range of the phase-resolved ODT depends on the speed of the line scans, spectral domain ODT has the advantage in terms of imaging speed and velocity dynamic range.

Transverse Flow Velocity and Doppler Angle Determination

One of the limitations of using the Doppler shift to determine the flow is that the technique is only sensitive to longitudinal flow velocity (Figure 2). If one knows the flow direction, Doppler shift measurement can fully quantify the flow. However, in many biological cases, where flow direction is not known, Doppler shift measurement alone is not enough to fully quantify the flow. Furthermore, there are many clinical cases, such as ocular bloodflow, where vessels are in the plane perpendicular to the probing beam. When flow direction is perpendicular to the probing beam, the Doppler shift is not sensitive to transverse bloodflow (V_{T}). Therefore, a method to measure transverse flow velocity is essential.

Standard deviation of the Doppler spectra can be used to determine the transverse flow. The technique is based on the fact that ODT imaging uses a relatively large numeric aperture lens in the sampling arm. The beam from different sides of the edges will produce different Doppler shifts, f_1 and f_2, as indicated in Figure 7. Consequently, the Doppler spectra will be broadened by the transverse flow. If the incident beam has a Gaussian spectral profile and contributions from Brownian motion and other sources that are independent of the macroscopic flow velocity are included, a linear relation between standard deviation of the Doppler spectra and transverse flow velocity, $V_{\mathrm{T}} = V \sin\theta$, can be derived:

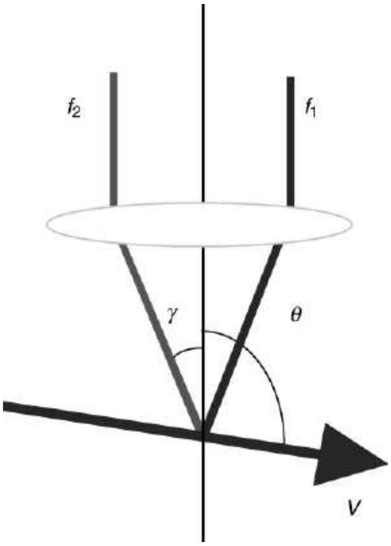

Figure 7 Effect of numerical aperture and transverse flow velocity on Doppler bandwidth.

$$\sigma = \frac{\pi V \sin\theta NA_{\mathrm{eff}}}{8\lambda} + b \qquad [25]$$

where b is a constant, and NA_{eff} is the effective numeric aperture. The standard deviation can be determined from the measured analytical fringe signal:

$$\sigma^2 = \frac{1}{2\pi^2 T^2} \times \left(1 - \frac{\left| \sum_{m=(n-1)M}^{nM} \sum_{j=1}^{N} \tilde{\Gamma}_j(t_m) \cdot \tilde{\Gamma}^*_{j+1}(t_m) \right|}{\frac{1}{2} \sum_{m=(n-1)M}^{nM} \sum_{j=1}^{N} [\tilde{\Gamma}_j(t_m) \cdot \tilde{\Gamma}^*_j(t_m) + \tilde{\Gamma}_{j+1}(t_m) \cdot \tilde{\Gamma}^*_{j+1}(t_m)]} \right) \qquad [26]$$

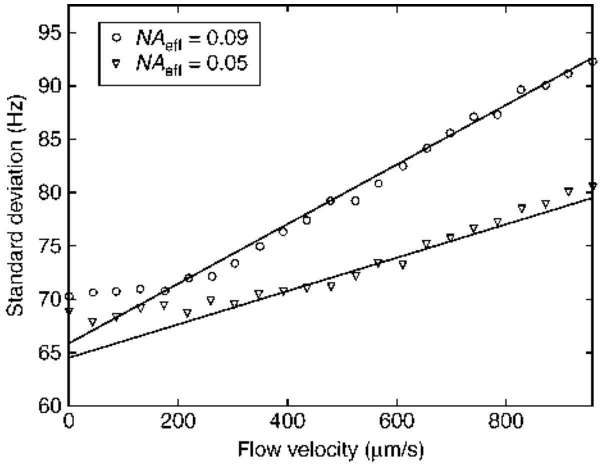

Figure 8 Standard deviation as a function of flow velocity for two different numeric apertures. Reproduced with permission from Ren *et al.* (2002) *Optics Letters* 27: 409–411.

The measured standard deviation as a function of transverse flow velocity is shown in Figure 8. As predicted, the Doppler bandwidth is a linear function of transverse flow velocity above a certain threshold level. The effective numerical aperture of the optical objective in the sample arm determines the slope of this dependence. This result indicates that standard deviation can be used to determine the transverse flow velocity. Since both longitudinal and transverse flow velocity (V_L and V_T) can be measured by the Doppler shift and standard deviation, respectively, flow direction can be determined from a single measurement of the Doppler fringe signal.

Applications

Due to its exceptionally high spatial resolution and velocity sensitivity, several clinical applications of ODT have been demonstrated, including screening vasoactive drugs, monitoring changes in image tissue morphology and hemodynamics following pharmacological intervention and photodynamic therapy, evaluating the efficacy of laser treatment in port wine stain (PWS) patients, assessing the depth of burn wounds, mapping cortical hemodynamics for brain research, imaging ocular bloodflow, and

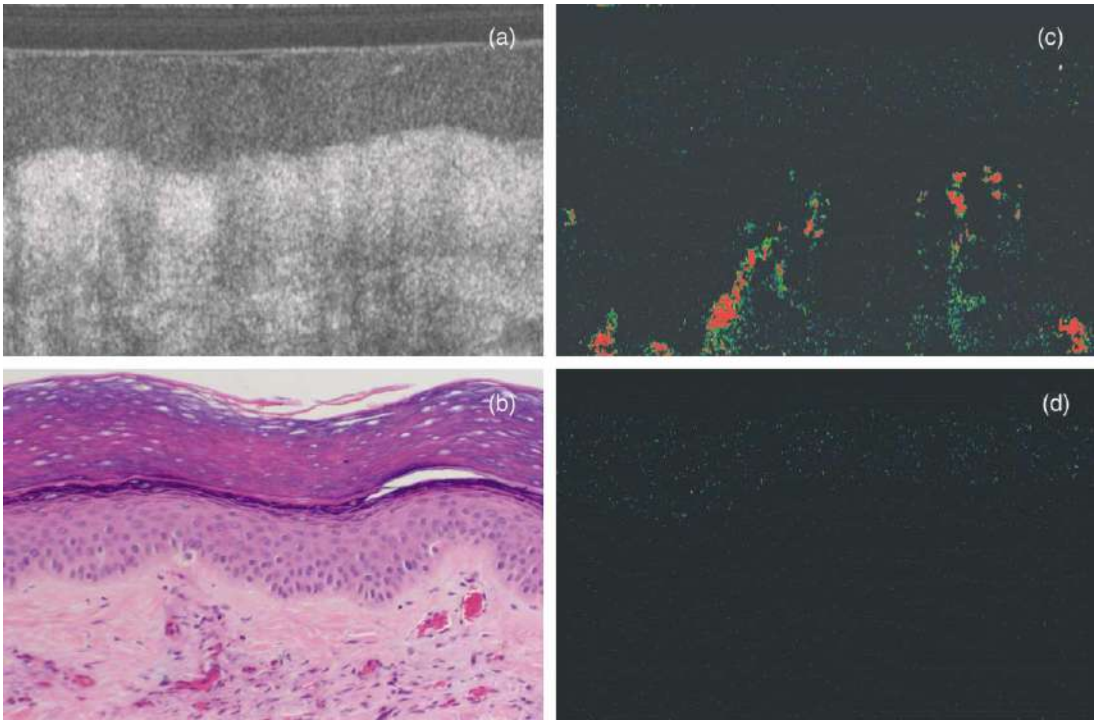

Figure 9 ODT images taken *in situ* from PWS human skin. (a) Structural image, (b) histological section, (c) image before laser treatment, and (d) image after laser treatment (scale bar 100 μm). Reproduced with permission from Nelson *et al.* (2001) *Archives of Dermatology* 137: 741–744. Copyright © 2001, American Medical Association. All rights reserved.

mapping bloodflow in gastrointestinal tracts. Figure 9 shows ODT structural and flow velocity images of a patient with PWS before and after laser treatment, respectively. For comparison, a histology picture taken at the same site is included. The vessel location from the ODT measurement and histology agree very well. Furthermore, the destruction of the vessel by laser can be identified, since no flow appears on the Doppler flow image after laser treatment.

Three-dimensional images of a microvascular network can be obtained by stacking 2D scans together. It is known that the microvasculature of mammary tumors has several distinct differences from normal tissues. Mapping of a 3D cancer microvascular network could provide additional information for cancer diagnosis. Figure 10 shows multiple blood vessels imaged from a patient with an abnormal microvascular network. Different colors represent different signs of the Doppler shift, which depends on the angle between the direction of flow and probing beam.

ODT is also a powerful tool to study flow dynamics in microfluidic channels. In addition to imaging flow velocity profile, ODT can also be used to measure osmotic mobility, quantify size of the scattering particle, and study flow dynamics of microfluids in microchannels of different materials, geometry, and surface treatment.

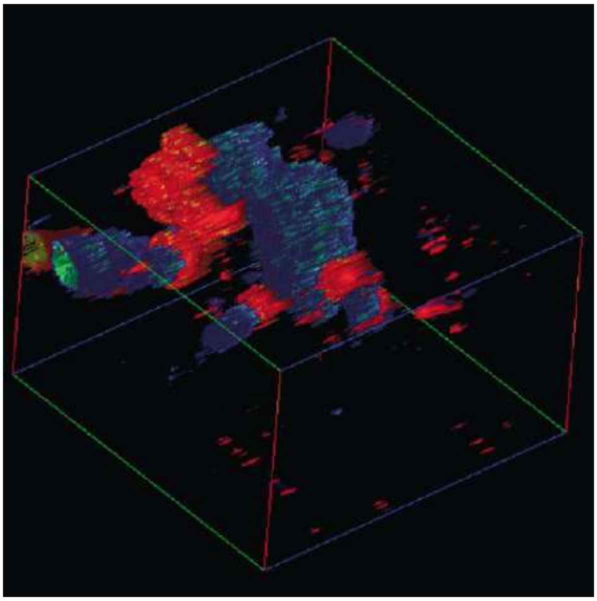

Figure 10 Three-dimensional ODT images of multiple blood vessels in human skin from a patient with a PWS birthmark. Reproduced with permission from Zhao *et al.* (2001) *IEEE J. Select Topics Quantum Electro.* 7: 931–935. © 2001, IEEE.

Polarization Sensitive OCT

PS-OCT combines polarization sensitive detection with OCT to determine tissue birefringence. PS-OCT can obtain enhanced image contrast and additional physiological information by studying polarization properties of biological tissues. Many biological tissues, such as tendon, muscle, nerve, bone, cartilage, and skin, exhibit birefringence properties. Consequently, PS-OCT can provide contrast mechanisms that are directly related to the physiological conditions of these tissues.

The polarization state of a light beam can be described by a Stokes vector **S** as:

$$\mathbf{S}=\begin{bmatrix} I \\ Q \\ U \\ V \end{bmatrix}=\begin{bmatrix} a_x^2+a_y^2 \\ a_x^2-a_y^2 \\ 2a_xa_y\cos\varphi \\ 2a_xa_y\sin\varphi \end{bmatrix} \qquad [27]$$

where I, Q, U, V are Stokes parameters, a_x and a_y are amplitudes of two orthogonal components of the electric vector, and φ represents the phase difference between the two components.

The effect of an optical device on the polarization of light can be characterized by a 4×4 Mueller matrix. The matrix acts on the input state, $\mathbf{S}_1$, to give the output state, $\mathbf{S}_2$:

$$\mathbf{S}_2=\begin{bmatrix} I_2 \\ Q_2 \\ U_2 \\ V_2 \end{bmatrix}=\mathbf{MS}_1$$

$$=\begin{bmatrix} M_{00} & M_{01} & M_{02} & M_{03} \\ M_{10} & M_{11} & M_{12} & M_{13} \\ M_{20} & M_{21} & M_{22} & M_{23} \\ M_{30} & M_{31} & M_{32} & M_{33} \end{bmatrix}\begin{bmatrix} I_1 \\ Q_1 \\ U_1 \\ V_1 \end{bmatrix} \qquad [28]$$

In the Poincaré sphere representation, a polarization state, **S**, can be represented by a point (Q, U, V). The Poincaré sphere representation provides a convenient method to evaluate changes of Stokes vectors. Determination of either the Stokes vector or the Mueller matrix allows full quantification of birefringence properties of the biological sample.

A fiber-based, high-speed PS-OCT system is shown in Figure 11. The system is similar to the ODT system, except a polarization modulator and two orthogonal detection channels are included. Due to the coherence detection intrinsic to OCT, backscattered light that

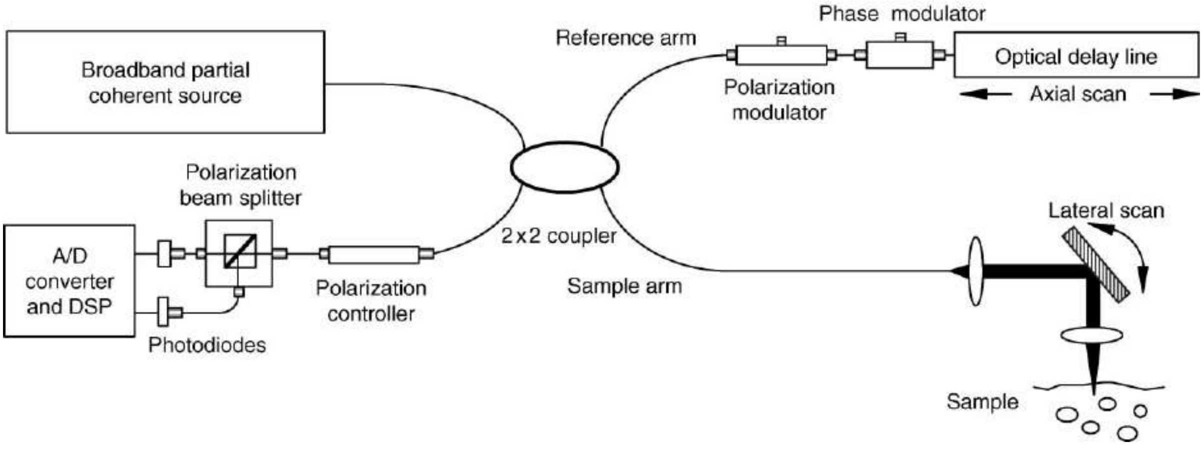

Figure 11 Fiber-based PS-OCT system. DSP: digital signal processing.

originates from incident light in different polarization states can be determined by controlling polarization states in the reference arm. Modulating the reference light in four orthogonal polarization states allows coherent detection of backscattered light from the sample under equivalent illumination in four different polarization states. The choice of orthogonal polarization states in the Poincaré sphere is important because it ensures that the birefringence measurements will be independent of the orientation of the optical axis in the sample.

For every polarization state controlled by the polarization modulator, the A-scan signals corresponding to the two orthogonal polarization diversity channels are digitized. The coherence matrix can be calculated from the complex electrical field vector from these two channels:

$$J = \begin{bmatrix} \langle E_{\mathrm{H}}^{*}(t)E_{\mathrm{H}}(t)\rangle & \langle E_{\mathrm{H}}^{*}(t)E_{\mathrm{V}}(t)\rangle \\ \langle E_{\mathrm{V}}^{*}(t)E_{\mathrm{H}}(t)\rangle & \langle E_{\mathrm{V}}^{*}(t)E_{\mathrm{V}}(t)\rangle \end{bmatrix} = \begin{bmatrix} J_{\mathrm{HH}} & J_{\mathrm{HV}} \\ J_{\mathrm{VH}} & J_{\mathrm{VV}} \end{bmatrix} \qquad [29]$$

where E_{H}, E_{V} are the components of the complex electric field vector corresponding to horizontal and vertical polarization channels, respectively, and E_{H}^{*}, E_{V}^{*} are their conjugates, respectively. The Stokes vector can be derived from the coherence matrix:

$$\begin{aligned} S_0 &= J_{\mathrm{HH}} + J_{\mathrm{VV}} & S_2 &= S_0^{-1}(J_{\mathrm{HV}} + J_{\mathrm{VH}}) \\ S_1 &= S_0^{-1}(J_{\mathrm{HH}} - J_{\mathrm{VV}}) & S_3 &= iS_0^{-1}(J_{\mathrm{VH}} - J_{\mathrm{HV}}) \end{aligned} \qquad [30]$$

where S_1, S_2, and S_3 are the normalized coordinates of the Stokes vector in the Poincaré sphere, characterizing the polarization state of the backscattered light, and S_0 is the module of the Stokes vector characterizing light intensity. The Stokes vector, corresponding to each of the input polarization states for the *n*th pixel in the *j*th A-scan, can be calculated as:

$$S_{0,n} = \sum_{m=(n-1)M}^{nM} [\tilde{\Gamma}_j^{\mathrm{H}}(t_m)\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m) + \tilde{\Gamma}_j^{\mathrm{V}}(t_m)\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)]$$

$$S_{1,n} = S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} [\tilde{\Gamma}_j^{\mathrm{H}}(t_m)\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m) - \tilde{\Gamma}_j^{\mathrm{V}}(t_m)\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)]$$

$$S_{2,n} = S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} 2\,\mathrm{Re}(\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)\tilde{\Gamma}_j^{\mathrm{V}}(t_m))$$

$$S_{3,n} = S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} 2\,\mathrm{Im}(\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)\tilde{\Gamma}_j^{V}(t_m)) \qquad [31]$$

where M is an integer number that determines the size of the average in the axial direction for each pixel; $\tilde{\Gamma}_j^{\mathrm{H}}(t_m)$ and $\tilde{\Gamma}_j^{\mathrm{V}}(t_m)$ are complex signals detected from the two orthogonal polarization channels at axial time t_m, for the *j*th A-scan; and $\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)$ and $\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)$ are their conjugates, respectively.

In order to measure the birefringence properties of the sample accurately, four states of light polarization are generated for each lateral pixel. For each polarization state, one A-line scan is performed. Therefore, a total of four A-line scans are used to calculate the Stokes vector and phase retardation images simultaneously.

For a sample with an assumed linear birefringence, there exists two eigenwaves that are polarized along the projected fast and slow axes of the sample normal to the propagation direction of incident light. The Stokes vectors of these eigenwaves determine a rotation axis in the equator plane of a Poincaré sphere.

The effect of birefringence is to rotate the Stokes vector about this axis through an angle that is equal to the phase retardation of the sample. Conversely, the rotation axis can be determined from the known polarization states of incident and backscattered light at the sample location. The phase retardation image, which characterizes the accumulated birefringence distribution in the sample, is calculated by the rotation of the Stokes vectors in the Poincaré sphere.

The simultaneous Stokes vectors and phase retardation images of a rat muscle are shown in Figure 12. Stokes vectors, corresponding to four different polarization states, are shown in the top figure. From Stokes vector images, the phase retardation image can be calculated as shown in the bottom figure. The banded structure in the phase retardation image indicates the accumulated tissue birefringence distribution in the rat muscle. In addition, differential optical axis and birefringence images can also be calculated.

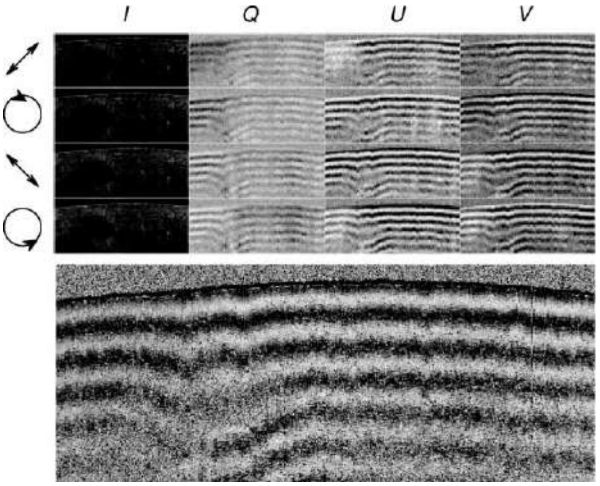

Figure 12 Stokes vectors images (top) and phase retardation image (bottom) in fresh rat muscle. The image area is 2 mm × 5 mm. Reproduced with permission from Ren *et al.* (2003) *SPIE Proceedings* 4956: 320–327.

There are a number of clinical applications for PS-OCT, including burn depth determination and retinal fiber nerve layer thickness evaluation. Figure 13 shows *in vivo* structural and phase retardation images from rat skin burned at 100°C for 10 seconds. Although the structure image shows very little contrast, the phase retardation image clearly shows the contrast between the normal and burned regions. As light propagates into the tissue, the retardation increases due to tissue birefringence. For normal tissue, it only takes less than 20 μm to cause a phase retardation of 90°. However, for the burned tissue, it takes much thicker tissue to cause the same phase retardation. The decrease of birefringence in thermally damaged tissue can clearly be seen in PS-OCT scans in comparison to normal rat skin.

Second Harmonic OCT

SH-OCT combines SHG with coherence gating for high-resolution tomographic imaging of tissue structures with molecular contrast. SHG is the lowest-order nonlinear optical process where the second-order nonlinear optical susceptibility is

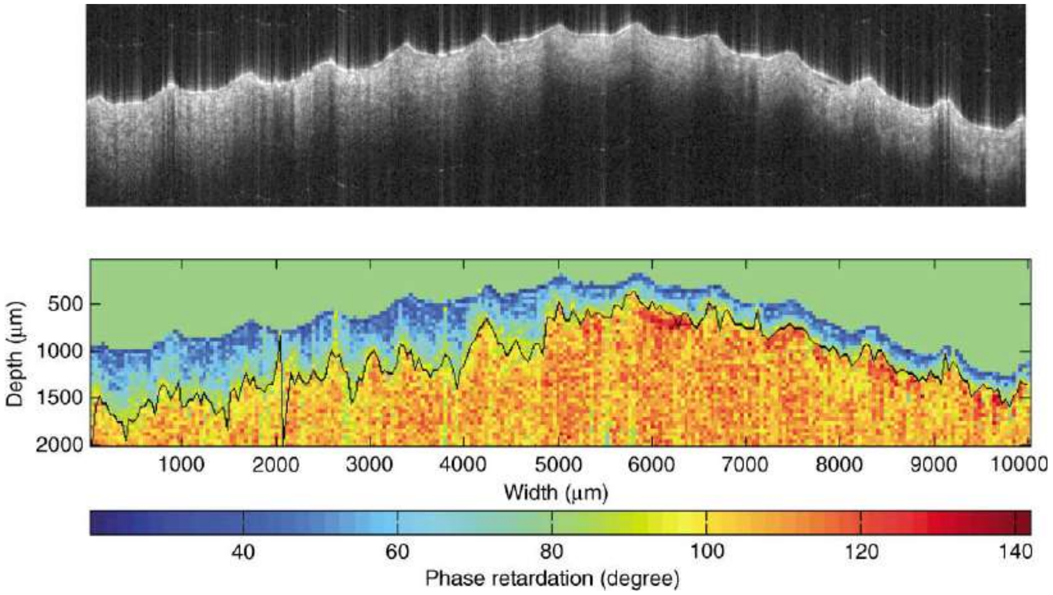

Figure 13 *In vivo* structure (top) and phase retardation image (bottom) from rat skin burned at 100°C for 10 seconds. The black contour line in the phase retardation image demarcates the depth at which 90° phase retardation has been reached with respect to the incident polarization (image size 2 mm × 10 mm).

responsible for the generation of light at second-harmonic frequency. Because second-order nonlinear optical susceptibility is very sensitive to electronic configurations, molecular structure and symmetry, local morphology, and ultrastructures, SHG provides molecular contrast for the coherence image.

The schematic diagram for SH-OCT is shown in Figure 14. A broadband short pulse laser is used as a light source in a Michelson interferometer. The light is split into the reference arm and the sample arm. In the reference arm, a thin nonlinear crystal is used to convert the input radiation to SH photons. Both SH and fundamental waves are then reflected by a metal mirror (M1) mounted on a motorized translation stage, which acts as the delay line in this SH-OCT system. A backscattered SH wave from the sample recombines with the SH wave from the reference arms to form interference fringe. The SH interference fringe signal is detected by a photomultiplier tube (PMT) after passing through a short-pass filter (F1). The fundamental interference fringe signal is detected by a photodiode (PD) after passing through a long-pass filter (F2).

A SH-OCT image of a rat tendon is shown in Figure 15. Compared with conventional OCT performed at fundamental wavelengths, SH-OCT offers enhanced molecular contrast and spatial resolution. It is also an improvement over existing SHG scanning microscopy technology as the intrinsic coherence gating mechanism enables the detection and discrimination of SH signals generated at deeper locations. The enhanced molecular contrast of SH-OCT extends conventional OCT's capability for detecting small changes in molecular structure. SH-OCT is promising for the diagnosis of cancers and other diseases at an early stage when changes in tissue and molecular structure are small.

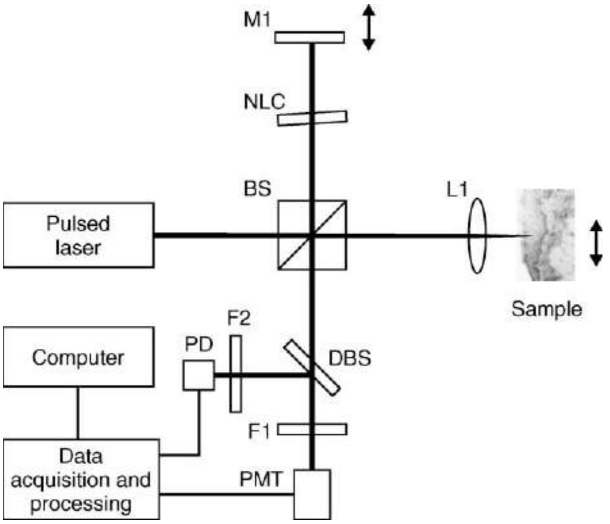

Figure 14 Schematic of experiment set-up for SH-OCT. M1, mirror; NLC, nonlinear crystal; BS, broadband nonpolarization beamsplitter; DBS, dichroic beamsplitter; L1, lens; F1–F2, filters; PD, photo diode; PMT, photomultiplier tube.

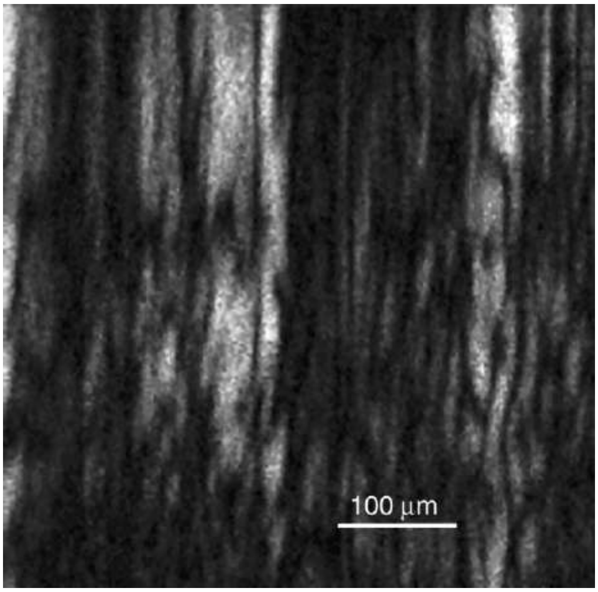

Figure 15 SH-OCT image of a rat tendon; image size 250 μm × 250 μm. Scale bar 100 μm. Image reprinted with permission from Jiang Y *et al.* (2004) Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Conclusions

Optical tomography is a rapidly developing imaging technology with many potential applications. Through different contrast enhancement mechanisms, ODT, PS-OCT, and SH-OCT provide clinically important physiological information that is not available in the structure image. Given the noninvasive nature and exceptionally high spatial resolution, optical tomography can simultaneously provide tissue structure, blood perfusion, birefringence, and other physiological information and have great potential for basic biomedical research and clinical medicine.

Acknowledgments

I would like to thank many of my colleagues who have contributed to the functional OCT project at the Beckman Laser Institute and Center for Biomedical Engineering at UCI, particularly the students and postdoctoral fellows. I also want to acknowledge the research grants awarded from the National Institutes of Health (EB-00293, NCI-91717, RR-01192 and EB-00255), the National Science Foundation (BES-86924), the Whitaker

Foundation (WF-23281) and the Defense Advanced Research Program Agency (Bioflip program). Institutional support from the Air Force Office of Scientific Research (F49620-00-1-0371), and the Beckman Laser Institute Endowment are also gratefully acknowledged. Dr Chen's e-mail address is zchen@bli.uci.edu.

List of Units and Nomenclature

OCT	Optical coherence tomography
ODT	Optical Doppler tomography
PS-OCT	Polarization sensitive OCT
SH-OCT	Second harmonic OCT
STFFT	Short time fast Fourier transformation

See also

Coherence: Coherence and Imaging; Overview. **Fiber and Guided Wave Optics:** Nonlinear Optics; Optical Fiber Cables. **Imaging:** Imaging Through Scattering Media; Interferometric Imaging; Photon Density Wave Imaging. **Interferometry:** Overview. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy. **Polarization:** Introduction; Matrix Analysis. **Tomography:** Optical Coherence Tomography; Tomography and Optical Imaging.

Further Reading

Bouma BE and Tearney GJ (eds) (2001) *The Optical Coherence Tomography Handbook*. New York: Marcel Dekker, Inc.

Chen Z (2003) Functional optical coherence tomograpy. In: Hwang NHC and Woo LY (eds) *Frontiers in Biomedical Engineering*, pp. 345–364. New York: Kluwer Academic Publishers.

Chen Z, Milner TE, Srinivas S, Wang XJ, Malekafzali A, van Gemert MJC and Nelson JS (1997) Noninvasive imaging of *in vivo* blood flow velocity using optical doppler tomography. *Optics Letters* 22: 1119–1121.

Chen Z, Zhao Y, Srinivas SM, Nelson S, Prakash N and Frostig RD (1999) Optical Doppler tomography. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1134–1142.

de Boer JF and Milner TE (2002) Review of polarization sensitive optical coherence tomography and Stokes vector determination. *Journal of Biomedical Optics* 7: 359–371.

Fercher AF and Hizenberger CK (2002) Optical coherence tomography. *Progress in Optics* 44: 215–296.

Huang D, Swanson EA, Lin CP, *et al.* (1991) Optical coherence tomography. *Science* 254: 1178–1181.

Izatt JA, Kulkarni MD, Yazdanfar S, Barton JK and Welch AJ (1997) *In vivo* bidirectional color Doppler flow imaging of picoliter blood volumes using optical coherence tomography. *Optics Letters* 22: 1439–1441.

Jiang Y, Tomov I, Wang Y and Chen Z (2004) Second harmonic optical coherence tomography. *Optics Letters* 29: 1090–1092.

Jiao S and Wang LV (2002) Jones-matrix imaging of biological tissues with quadruple-channel optical coherence tomography. *Journal of Biomedical Optics* 7: 350–358.

Masters BR and Thompson BJ (eds) (2001) *Selected Papers on Optical Low-Coherence Reflectometry & Tomography*. SPIE Milestone Series, vol. MS 165. Bellingham, WA: SPIE Optical Engineering Press.

Tuchin VV (ed.) (2004) *Coherent-Domain Optical Methods for Biomedical Diagnostics, Environmental and Material Science*, vol. 2. Boston, MA: Kluwer Academic Publishers.

Vo-Dinh T (ed.) (2003) *Biomedical Photonics Handbook*. New York: CRC Press.

U

ULTRAFAST LASER TECHNIQUES

Contents

Generation of Femtosecond Pulses
Pulse Characterization Techniques

Generation of Femtosecond Pulses

D T Reid, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical pulses from femtosecond lasers represent some of the shortest physical events ever observed and their high instantaneous powers allow them to be focused to intensities that can heat and compress matter to resemble the extreme conditions found in a nuclear explosion. The extreme units of time and power needed to describe this area of physics can easily distract from the simple and elegant techniques by which femtosecond pulses can be generated. The review presented here is divided into two main sections. The first explains the theory of pulse generation and shaping and the second concentrates on actual femtosecond laser sources in common use today.

Theory of Femtosecond Pulse Generation

Active Mode-locking

In an inhomogeneously broadened laser system, optical frequencies within the emission bandwidth of the material can experience gain simultaneously at discrete frequencies corresponding to the longitudinal modes of the laser resonator. In the absence of any mechanism to coherently couple energy between these modes, the laser will operate with a continuous-wave (cw) output containing a narrow band of frequencies whose roundtrip cavity gain is highest. An alternative operating regime exists in which each mode is related to its neighbor by a fixed phase relationship and it is this mode-locked condition that is used to generate a periodic sequence of ultrashort optical pulses.

To understand how mode-locking works, consider the frequency separation between adjacent longitudinal resonator modes which can be expressed as

$$\Delta\omega = 2\pi c/l \qquad [1]$$

where l is the optical roundtrip length of the resonator and c is the speed of light. If the nth longitudinal mode has an amplitude E_n then the total optical field can be denoted in complex notation as

$$E(t) = \sum_n E_n \exp \mathrm{i}[(\omega_0 + n\Delta\omega)t + n\Delta\varphi] \qquad [2]$$

with ω_0 being the center frequency of the output and $\Delta\varphi$ being the phase difference between adjacent modes. Figure 1 illustrates how a simple pulse can be produced by coherently adding together adjacent modes. In Figure 1a, three modes with equal amplitudes and the same initial phase ($\Delta\varphi = 0$) are interfered to give a pulse. As further modes are added (Figure 1b) the pulse duration decreases.

Coupling of energy between adjacent cavity modes can be achieved in practice by modulating either the

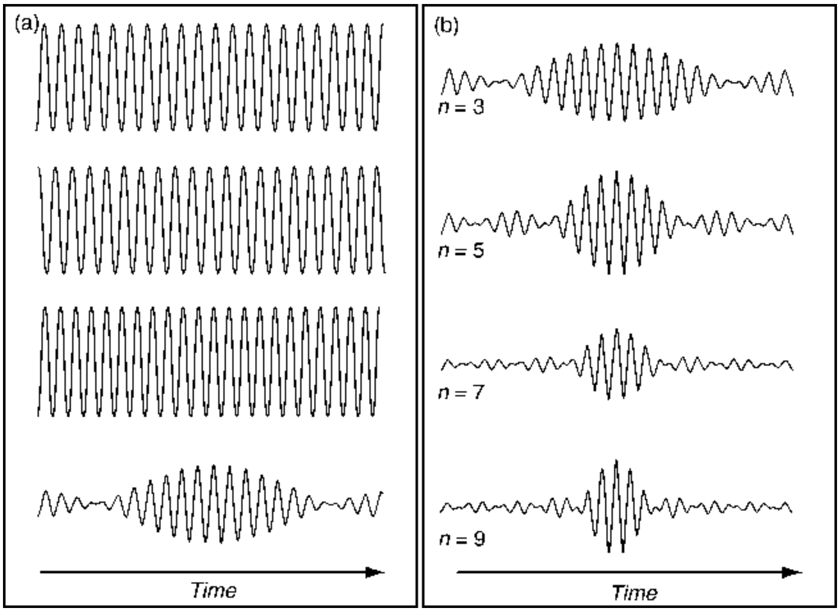

Figure 1 The principles of mode-locking: (a) three adjacent modes add coherently to produce a pulse; (b) shorter pulses require a larger number of *n* coupled modes.

intracavity loss (amplitude modulation (AM) mode-locking) or the intracavity phase (frequency modulation (FM) mode-locking). A common practical amplitude modulator is an acousto-optic modulator and for operation in the visible and near-infrared, fused silica is a common acousto-optic modulator material. Lithium-niobate devices based on the Pockels effect are the most commonly used phase modulators and are attractive for mode-locked fiber-laser applications because fiber-pigtailed waveguide devices are readily available.

Active mode-locking has been employed in a wide variety of laser systems including argon ion, dye, erbium-doped fiber, semiconductor, and solid-state lasers. Unlike passive mode-locking, no optical nonlinearity is required which means that actively mode-locked lasers can be operated with low intracavity powers. The technique is particularly suitable for mode-locked sources which require active synchronization to an external clock source.

A form of 'optical' active mode-locking exists in which gain modulation resulting from optical pumping using another mode-locked laser is the mode-locking mechanism. Known as synchronously pumped mode-locking, this technique has been used to obtain ultrashort pulses from oscillators incorporating gain media of laser or nonlinear optical materials. The technique is most appropriate to laser gain materials which have a short-lived fluorescence lifetime and whose gain therefore exhibits a significant change when the pump light is modulated at high frequencies. Synchronous pumping has been used to generate mode-locked pulses from color-center lasers, dye lasers, and optical parametric oscillators.

Passive Mode-locking

To produce the shortest optical pulses, passive mode-locking, a technique based on exploiting intracavity optical nonlinearities, must be employed. Unlike the active technique, the strength of the mode-locking action increases as the propagating pulses shorten and their peak intensities rise. One consequence of this nonlinear behavior is that the passive technique commonly generates and mode-locks entirely new cavity frequencies which were not originally present during cw operation.

Passive mode-locking has been implemented using a wide range of different nonlinear effects, but in every technique the mode-locking process can be described in terms of dynamic loss saturation combined, in certain systems, with dynamic gain saturation. The principal mode-locking element is a saturable absorber whose loss decreases as the incident pulse intensity increases. A saturable absorber can take the form either of a resonant component such as a solid-state semiconductor device or a dye jet with a fixed absorption wavelength, or a nonresonant component that replicates a saturable-absorber-like action using self-focusing or self-phase-modulation effects. Saturable absorbers can be categorized as either slow saturable absorbers or fast saturable absorbers, depending on whether their recovery time is longer or shorter than the pulse duration, respectively. The pulse-shaping mechanisms are different in each case and will be described separately.

Slow saturable absorbers are physical devices, commonly dye molecules in solution or semiconductor-doped glass, which rely on resonant excitation to produce an excited electronic state whose transmission is greater than that of the ground state. When an optical pulse is incident on a slow absorber, its leading edge is absorbed and creates an excited state which is relatively transparent to the trailing edge. In a laser medium whose gain is readily saturated, a pulse propagating first through the absorber and then through the gain medium experiences a 'gain window' with a finite duration which strongly shapes the pulse (see Figure 2a,b).

A fast saturable absorber is required to realize mode-locking in laser media which are not easily saturated. The time-dependent gain and loss profiles of the absorber and the gain medium are again depicted in Figure 2c,d and refer to an absorber with both a rapid response and recovery time. In this system, the dynamics of the absorber and gain medium no longer represent the limiting factor influencing the pulse duration, and other effects such as linear and nonlinear dispersion become the dominant pulse-shaping mechanisms.

Pulse Shaping by Material Dispersion

Intense pulses propagating in a mode-locked laser experience strong linear and nonlinear shaping effects associated, respectively, with group-velocity dispersion and self-phase-modulation. Contributions to the net intracavity dispersion arise from the gain medium, the cavity mirrors, other elements such as intracavity prisms, even the air in the beam path. Dispersive effects in materials originate as a result of a frequency-dependent dielectric susceptibility:

$$P(\omega) = \varepsilon_0 \chi^{(1)}(\omega) E(\omega) + \varepsilon_0 \chi^{(2)}(\omega) E(\omega)^2 + \varepsilon_0 \chi^{(3)}(\omega) E(\omega)^3 + \cdots \quad [3]$$

which is due to physical resonances (poles) in the electronic response to an applied E-field.

Linear dispersion

Linear dispersion refers to the variation of the refractive index associated with the real part of the $\chi^{(1)}$ term, namely,

$$n_0 = \sqrt{1 + \mathrm{Re}\{\chi^{(1)}\}} \quad [4]$$

and leads to a wavelength-dependent group velocity which is responsible for pulse broadening and break-up in optical systems. Dispersive effects are best analyzed in terms of optical phase which is related to the refractive index by

$$\varphi(\omega) = \frac{n_0(\omega)\omega L}{c} \quad [5]$$

where L is the medium length and c is the vacuum speed of light. The local variation of the spectral

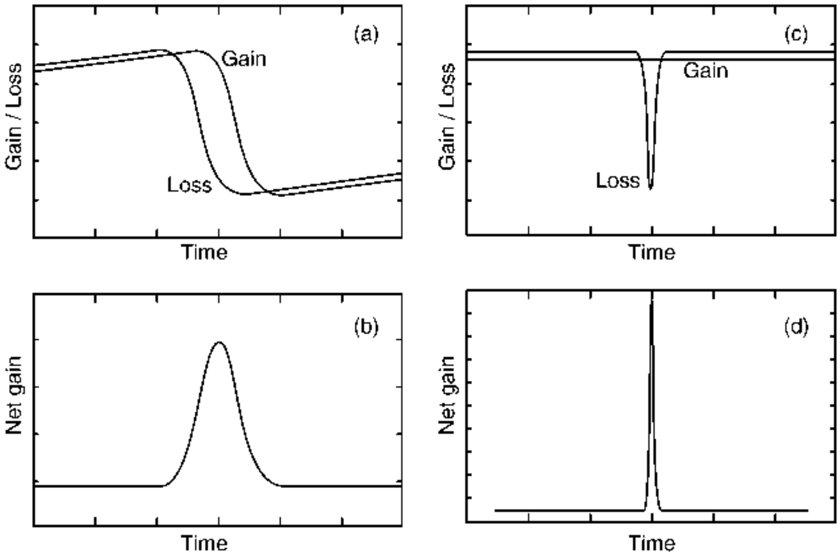

Figure 2 Saturable absorber action: (a) intracavity gain and loss profiles for a laser comprising a slow saturable absorber and a rapid relaxation time gain medium, and (b) the net gain (gain − loss) for this system; (c) gain and loss profiles for a laser comprising a near-instantaneous saturable absorber and a slow relaxation time gain medium, and (d) the net gain for this system.

phase due to an optical medium can be represented as the Taylor series

$$\varphi(\omega) = \varphi(\omega_0) + (\omega - \omega_0)\left.\frac{\mathrm{d}\varphi}{\mathrm{d}\omega}\right|_{\omega_0} + \frac{1}{2}(\omega - \omega_0)^2 \left.\frac{\mathrm{d}^2\varphi}{\mathrm{d}\omega^2}\right|_{\omega_0} + \left|\frac{1}{6}(\omega - \omega_0)^3 \frac{\mathrm{d}^3\varphi}{\mathrm{d}\omega^3}\right\|_{\omega_0} + \cdots \qquad [6]$$

where ω_0 is the center frequency of the incident light. When an ultrashort pulse propagates through the medium, only the quadratic and higher-order terms in [6] have a direct influence on the shape of the pulse in time. The $(\omega - \omega_0)^2$ term is responsible for adding linear chirp which refers to a time-varying linear increase or decrease in the instantaneous optical frequency across the pulse. Obtaining the shortest pulses from a mode-locked laser requires the quadratic and higher-order terms in the roundtrip intracavity spectral phase to be zero. In practice this means minimizing the material group-velocity dispersion of the gain medium by using a short but highly doped laser crystal and then identifying additional optical elements whose group-velocity dispersion is equal in magnitude but opposite in sign to that of the crystal. Optical components for compensating group-velocity dispersion fall into two categories: those which use bulk optics and geometry to achieve a wavelength-dependent path length, and others based on interference effects within a dielectric optical coating.

Geometrical methods of dispersion compensation are commonly based on pairs of diffraction gratings or prisms and in both cases the geometry of the systems means that the group delay is wavelength dependent and can be made greater for longer wavelengths than for shorter ones. Analytic expressions describing the quadratic and cubic spectral phase of prism and diffraction grating pairs are summarized in Table 1. Dispersion compensation using prisms has been extremely successful and 11-fs pulses were reported using this approach. The simultaneous compensation of arbitrary second- and third-order spectral phase is, however, difficult and involves identifying a suitable prism material, apex spacing, and tip insertion. The expressions listed in Table 1 indicate that the form of the spectral phase depends on the refractive index of the prism material and its wavelength derivatives, while the apex separation, l, also contributes to a large extent. A further contribution to the spectral phase comes from the glass encountered by the beam passing through the prism tips which contributes a positive dispersion described by the equations for a bulk material given in Table 1. In general, an iterative solution of the equations given in Table 1 is needed to determine the geometry of a prism pair suitable for producing a desired spectral phase profile.

Laser cavity mirrors can be designed with chirped multilayer dielectric coatings which possess a chosen group-velocity dispersion profile. A Bragg mirror with smaller layer periods at the surface and larger periods deeper in the coating provides negative group-velocity dispersion because the longer wavelengths experience larger group delays within the coating than the shorter ones. These chirped mirrors provide constant reflectivity across their stopband and therefore only modify the spectral phase of a pulse, leaving the intensity unchanged. The design of chirped mirrors has been refined in recent years to enhance their operating bandwidth and to reduce undesirable modulations in the group-delay response. The technique has been used to obtain 6.5-fs pulses directly from a Ti:sapphire oscillator and to achieve compression of amplified pulses to below 5 fs.

Nonlinear dispersion

Nonlinear effects arise in all materials due to the $\chi^{(3)}$ term in [3]. In a transparent medium the refractive index due to this net susceptibility is approximately

$$n \approx n_0 + \frac{\chi^{(3)}E^2}{2n_0} = n_0 + n_2 I \qquad [7]$$

where I is the optical intensity. The importance of the $\chi^{(3)}$ susceptibility is therefore that it causes the refractive index to become intensity-dependent, and significant changes in the refractive index can be induced by the high peak intensities of ultrashort pulses.

Self-phase-modulation (SPM) resulting from the nonlinear refractive index is essential in generating the shortest pulses from modern ultrafast lasers and is caused by the time-varying intensity profile of the pulse. The higher-intensity components induce a larger nonlinear refractive index change and therefore experience a greater phase shift than the weaker components. The resulting frequency shift across the pulse can be shown to be

$$\Omega(t) = -\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = -\frac{\omega n_2 L}{c}\frac{\mathrm{d}I(t)}{\mathrm{d}t} \qquad [8]$$

which describes a redshift of the pulse leading edge and a blueshift of the trailing edge. Only the pulse center frequency remains unchanged and SPM therefore broadens the pulse spectrum by redistributing energy from the center to the wings. Across the center of the pulse, SPM leads to an approximately linear positive chirp that, with appropriate optics, can be

Table 1 Expressions for the second and third-order dispersion of common optical systems

Bulk material, thickness *t* 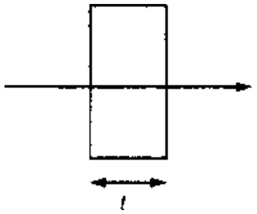 	$\frac{d^2\varphi}{d\omega^2} = t\frac{\lambda^3}{2\pi c^2}\frac{d^2 n}{d\lambda^2}$ $\frac{d^3\varphi}{d\omega^3} = -t\frac{\lambda^4}{4\pi^2 c^3}\left(3\frac{d^2 n}{d\lambda^2} + \lambda\frac{d^3 n}{d\lambda^3}\right)$
Double-pass prism-pair, apex separation *l*, angular deviation within ray of *β* 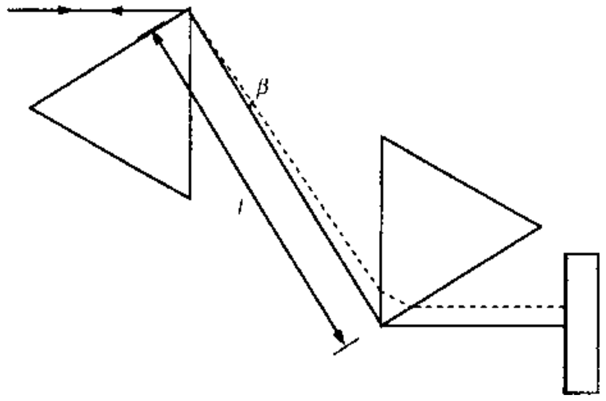 	$\frac{d^2\varphi}{d\omega^2} = \frac{\lambda^3}{2\pi c^2}\frac{d^2 P}{d\lambda^2}$ $\frac{d^3\varphi}{d\omega^3} = -\frac{\lambda^4}{4\pi^2 c^3}\left(3\frac{d^2 P}{d\lambda^2} + \lambda\frac{d^3 P}{d\lambda^3}\right)$ $\frac{d^2 P}{d\lambda^2} = 4l\sin\beta\left[\frac{d^2 n}{d\lambda^2} + (2n - 1/n^3)\left(\frac{dn}{d\lambda}\right)^2\right] - 8l\cos\beta\left(\frac{dn}{d\lambda}\right)^2$ $\frac{d^2 P}{d\lambda^3} = l\cos\beta\left[\left(24/n^3 - 48n\right)\left(\frac{dn}{d\lambda}\right)^3 - 24\frac{dn}{d\lambda}\frac{d^2 n}{d\lambda^2}\right] + l\sin\beta\left[\begin{aligned}&\left(\frac{dn}{d\lambda}\right)^3\left(12/n^6 + 12/n^4 + 8/n^3 - 16/n^2 + 32n\right)\\ &+\left(24n - 12/n^3\right)\frac{dn}{d\lambda}\frac{d^2 n}{d\lambda^2} + 4\frac{d^3 n}{d\lambda^3}\end{aligned}\right]$
Double-pass grating-pair, separation *l*, angle of incidence *γ*, line spacing *d* 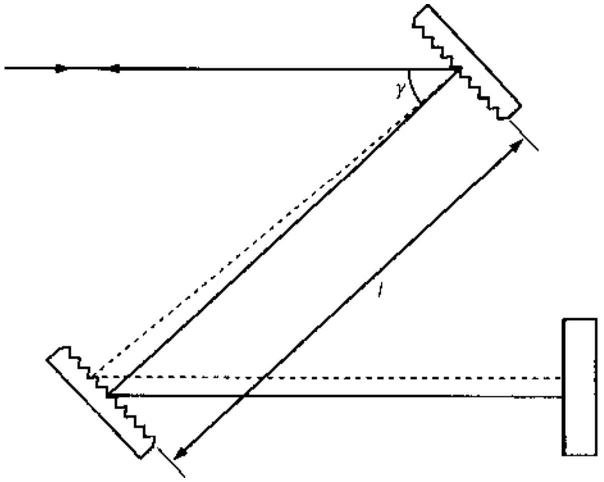 	$\frac{d^2\varphi}{d\omega^2} = l\frac{\lambda^3}{\pi c^2 d^2}\left[1 - \left(\frac{\lambda}{d} - \sin\gamma\right)^2\right]^{-3/2}$ $\frac{d^3\varphi}{d\omega^3} = -6l\frac{\lambda^4}{c^3 d^2}\left(1 + \frac{\lambda}{d}\sin\gamma - \sin^2\gamma\right)\left[1 - \left(\frac{\lambda}{d} - \sin\gamma\right)^2\right]^{-5/2}$

corrected using an equal but opposite amount of negative second-order dispersion. The combination of spectral broadening using SPM and subsequent chirp-removal using linear dispersion is the basis for optical pulse compression techniques that have been applied effectively to create pulses which exist for only a few carrier-wave periods.

When the pulse propagates in an environment where SPM and linear dispersion are well balanced and evenly distributed, a soliton will evolve which, in the absence of loss, can propagate indefinitely without changing shape. The nonlinear Schrödinger equation predicts that lasers in which the dominant pulse-shaping mechanisms are negative group-velocity dispersion and self-phase-modulation will produce soliton-like pulses with a $\mathrm{sech}^2(t)$ intensity profile. In experimental measurements of ultrafast lasers the pulse shape is commonly assumed to be $\mathrm{sech}^2(t)$ and the chirp of the pulses is then estimated by measuring their duration–bandwidth product, $\Delta\nu\Delta\tau$, which is a dimensionless quantity that can be treated as a figure-of-merit for ultrashort pulses. For unchirped pulses the duration–bandwidth product is independent of their duration or spectral bandwidth but is sensitive to their shape: chirp-free Gaussian pulses have $\Delta\nu\Delta\tau = 0.441$ while $\mathrm{sech}^2(t)$ pulses have $\Delta\nu\Delta\tau = 0.315$.

Sources of Femtosecond Pulses

Dye Lasers

The first practical passively mode-locked ultrafast lasers were those based on a gain medium of a thin (~10 μm) organic dye jet and incorporating a slow saturable absorber jet at a separate cavity focus. The most common combination was Rhodamine 6G in ethylene glycol (gain jet) and 3,3-diethyloxadicarbocyanine iodide in ethylene glycol (absorber jet) which produced pulses at a wavelength of 620 nm. The most successful approach was colliding-pulse mode-locking in which two counter-propagating pulses within a ring cavity are synchronized to arrive simultaneously at the absorber jet, so achieving greater saturation of the absorber and allowing the pulses to be transmitted with a loss significantly smaller than in the single-pulse case. The first sub-100-fs optical pulses were reported in 1981 from a colliding-pulse mode-locking dye laser and later work reported pulse durations as low as 27 fs but the configuration was inefficient and produced low average output powers. Other approaches such as synchronous pumping and hybrid mode-locking produced higher output powers but the maintenance-intensive nature of dye lasers and the hazardous nature of many of the chemicals used led to their gradual replacement by more convenient solid-state alternatives, based principally on Ti:sapphire.

Color-Center Lasers

A new source of femtosecond pulses in the near-infrared was reported in 1984. Coined the soliton laser the system was a synchronously pumped KCl:Tl color-center laser tunable from 1.4 to 1.6 μm and modified to allow direct generation of pulses as short as 50 fs, substantially shorter than the 8 ps achieved in synchronously pumped operation alone. Femtosecond operation was achieved by coupling a length of polarization-preserving anomalously dispersive fiber to the main laser cavity using a beamsplitter and retroreflector arrangement. This coupled-cavity arrangement later became known as additive pulse mode-locking (APM) and produced short pulses when the optical length of the fiber was adjusted to be an integral multiple of the main cavity length. The long upper-state lifetime of the laser-active centers in the KCl:Tl gain medium allowed particularly stable APM operation, but mode-locking of this kind was also successfully employed in an actively stabilized NaCl:OH$^-$ color-center laser to produce 110-fs pulses. The need to maintain color-center crystals at liquid-nitrogen temperatures limited the convenience of these lasers but until the advent of Ti:sapphire they remained the only solid-state sources capable of femtosecond operation.

Ti:sapphire Lasers

The creation in 1986 of a new broadband laser material – titanium-doped sapphire or Ti:sapphire – began a revolution in ultrafast laser sources that still continues today. Unlike other contemporary solid-state materials for mode-locked lasers such as Nd:YAG, Ti:sapphire is a vibronic gain medium which exhibits a strong electron-phonon coupling between the titanium ion transition and the host lattice. This vibronic coupling leads to a broad continuum of possible transition energies and results in a fluorescence bandwidth covering 670–1050 nm, making Ti:sapphire capable of supporting extremely short femtosecond pulses. The long fluorescence lifetime of Ti:sapphire means that, unlike dye lasers, femtosecond mode-locking requires a passive method using a fast saturable absorber and two methods have prevailed and are discussed below.

The first passively mode-locked Ti:sapphire lasers were based on the APM technique in which mode-locking is achieved by a fast-saturable absorber action due to interference between the intracavity pulse and a self-phase-modulated replica propagating in an auxiliary cavity. An important breakthrough was made in 1990 when it was reported that an APM Ti:sapphire laser continued to operate even when the auxiliary cavity was blocked. This discovery was attributed to an entirely new mode-locking effect known as Kerr-lens mode-locking (KLM) or self-mode-locking in which the presence of a Kerr lens within the gain medium during mode-locked operation changes the mode focusing within the cavity so that, compared to cw operation, mode-locked pulses experience higher gain. Two configurations of KLM exist and are known as soft aperture mode-locking, where the presence of the Kerr lens increases the gain by improving the overlap between the pump and laser modes, and hard-aperture mode-locking, in which a physical aperture such as a slit or the edge of a prism is adjusted to introduce greater loss for cw operation. Because KLM is based on a nonresonant nonlinearity, Ti:sapphire lasers can be mode-locked at any wavelength in their gain bandwidth using this technique. Recent advances in controlling the intracavity group-velocity dispersion characteristics have led to the generation of sub-5-fs pulses directly from a Ti:sapphire oscillator. The principal drawback of the KLM method is that it is not self-starting because mode-locked operation must be seeded by an intense noise spike or other short fluctuation. Different starting methods have been applied successfully including mirror tapping, acousto-optic modulation, and mode-dragging, and once mode-locking has been initiated it is generally stable until some external perturbation disturbs the intracavity beam.

An alternative method for obtaining femtosecond pulses has been pioneered largely in parallel with the KLM efforts and is based on using resonant nonlinearities in semiconductors as fast saturable absorbers. The approach is to fabricate a semiconductor multiple-quantum-well (MQW) device whose bandgap has been chosen to match the laser wavelength and whose absorption can be saturated at high fluences. Initially these devices were utilized in an APM geometry in which the MQW device was situated in an auxiliary cavity but it was later realized that an equivalent monolithic configuration could be obtained by sandwiching the MQW layer between two high-reflectivity mirrors (see Figure 3, inset) to form a Fabry–Perot structure and the resulting element was known as an antiresonant Fabry–Perot saturable absorber. The first application of a such a semiconductor saturable absorber mirror (SESAM) for mode-locking Ti:sapphire was reported in 1995 where self-starting operation was demonstrated. A typical modern Ti:sapphire laser design is shown in Figure 3 and illustrates how a saturable absorber mirror can be integrated into the cavity. Further design enhancements such as reducing the finesse of the Fabry–Perot cavity resulted in SESAMs with a broader wavelength response which enabled sub-10-fs self-starting operation.

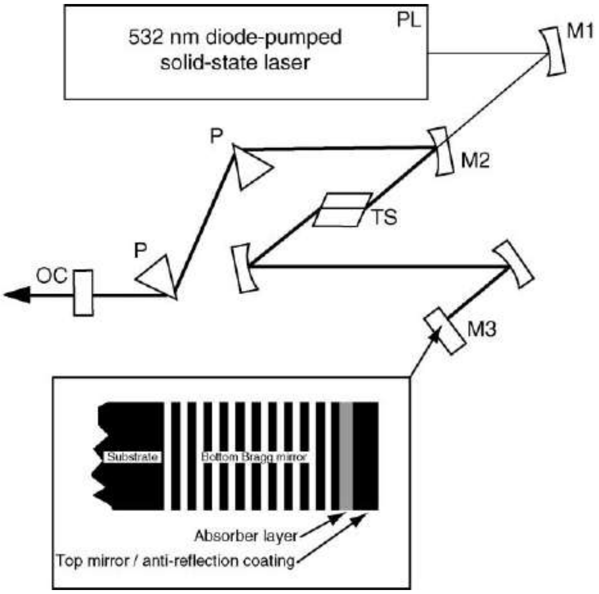
532 nm diode-pumped solid-state laser
PL
M1
P
M2
TS
P
OC
M3
Substrate
Bottom Bragg mirror
Absorber layer
Top mirror / anti-reflection coating

Figure 3 Typical configuration of a modern femtosecond Ti:sapphire laser. PL, pump laser; M1, pump focusing mirror; M2, high-reflectivity cavity mirror (800 nm); TS, Ti:sapphire crystal, typically, 2–10 mm long; M3, SESAM mirror; P, dispersion-compensating prisms; OC, output coupler. Inset: a simplified schematic of a saturable-absorber Bragg mirror.

Fiber Lasers

Ultrafast fiber lasers, particularly those based on silica doped with rare earth Er^{3+}, Nd^{3+}, or Yb^{3+}ions, have enjoyed renewed attention in recent years as efforts to scale their average output powers and pulse energies to practical levels have proved successful. Four main mode-locking strategies have been applied to these systems: active mode-locking which is generally restricted to picosecond generation; intensity-dependent feedback using a nonlinear amplifying loop mirror; polarization APM; and semiconductor saturable absorber mode-locking. The upper-state lifetime of rare-earth-doped fibers is long and therefore a fast saturable absorber mechanism is necessary for passive mode-locking, limiting the choice of mode-locking elements to those based on Kerr or semiconductor nonlinearities.

The development of mode-locked fiber sources has concentrated on Er-doped lasers because of the compatibility of their output wavelengths with the 1.5 μm optical communication window in standard silica fiber. Passively mode-locked designs based on the nonlinear loop mirror work by exploiting the Kerr effect in an intracavity fiber Sagnac interferometer containing a gain section. With suitable control of the intracavity polarization, such a nonlinear amplifying loop mirror (NALM) can be set up to transmit light of high intensities but reflect low intensities because of the differential phase shift induced between the two interferometer arms. By adding the loop mirror to a unidirectional ring cavity a figure-of-eight laser is formed (Figure 4) which is capable of producing femtosecond soliton pulses with average powers of around 1 mW. Drawbacks of the figure-of-eight laser include the presence of spectral sidebands at the shortest pulse durations and a tendency for multiple pulse operation. Mode-locking can also be achieved

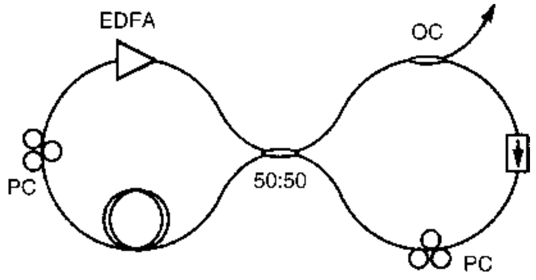

Figure 4 Schematic of a figure-of-eight femtosecond fiber laser. EDFA, erbium-doped fiber amplifier gain section; OC, output coupler; PC, polarization controller.

using a variation on the APM technique and significant average output powers have been demonstrated from stretched-pulse-APM lasers, with up to 90 mW being reported at pulse energies of 2.25 nJ.

Semiconductor saturable absorber mode-locking has been demonstrated in fiber lasers using a variety of configurations including bulk InGaAsP on InP and saturable Bragg reflectors based on InGaAs/InP MQWs on a AlAs/GaAs mirror structure. Fiber lasers mode-locked using semiconductor saturable absorbers are self-starting and exhibit stable mode-locking, normally at the fundamental cavity frequency. The narrow bandwidth of saturable absorber devices at near-infrared wavelengths makes the generation of sub-100-fs pulses difficult.

Semiconductor Sources

Diode lasers and external-cavity lasers based on semiconductor optical amplifiers (SOAs) are of importance because of their ability to produce ultrashort pulses from a compact system at wavelengths of relevance to optical fiber communications. The two principal methods used to generate ultrashort pulses from such devices are mode-locking with an external cavity or gain-switching (equivalently known as *Q*-switching) in a bow-tie waveguide structure. Subpicosecond pulse durations are now routinely produced using dispersion compensation and/or nonlinear pulse compression techniques and, by combining these strategies with Er:fiber amplification, 20-fs pulses have been produced. The mode-locking and gain-switching techniques each have their individual advantages. Gain-switching provides a simple and compact approach offering high average output power, but significant timing jitter and poor spectral quality can accompany pulses produced in this way. Mode-locked systems are physically larger than gain-switched devices because they require an external cavity and additional optical components, but shorter and higher quality pulses are achievable.

Other Common Solid-State Laser Sources

Certain solid-state gain media offer the potential for direct diode pumping, miniaturization and operation at wavelengths or average powers not accessible with other sources. Systems based on Nd or Cr ion transitions have been shown to be capable of ultrashort picosecond or femtosecond operation. Femtosecond mode-locking (175 fs) has been reported in Nd:glass which has one of the broadest fluorescence spectra of the Nd materials. Crystals with active ions of Cr^{2+}, Cr^{3+}, and Cr^{4+} have broad transitions covering wavelengths in the 2.0 μm, 800 nm and 1.4 μm regions and femtosecond operation has been demonstrated in Cr^{4+}:YAG at 1.52 μm, Cr^{4+}:Mg_2SiO_4 (Cr:forsterite) at 1.25 μm, Cr^{3+}:$LiSrAlF_4$ (Cr:LiSAF) at 850 nm, and Cr^{3+}:$LiSrGaF_6$ (Cr:LiSGAF). Direct diode pumping of Cr-ion-based lasers is attractive because of the opportunities for producing low-noise, compact, and efficient femtosecond sources, but the limited power available from high-quality pump diodes results in low intracavity powers which make KLM operation difficult. As a result, stable operation is better achieved using a physical saturable absorber, and SESAM mode-locking has been applied successfully to many of these lasers.

Sources Based on Nonlinear Frequency Conversion

Laser sources alone are unable to provide ultrashort pulses continuously tunable from the ultraviolet to the mid-infrared, and large gaps exist at wavelengths where no broadband laser materials are available. Harmonic generation using Ti:sapphire can be used to access some parts of the visible and ultraviolet regions, but infrared and certain visible wavelengths cannot be generated in this way and coverage is limited by the ability to tune the source itself. The nonlinear production of infrared wavelengths requires parametric down-conversion using a synchronously pumped optical parametric oscillator. Ti:sapphire pumped synchronously pumped optical parametric oscillators have produced, either directly or using harmonic generation, outputs from the green to beyond 7 μm.

Sources of Amplified Ultrashort Pulses

Modern ultrafast oscillators can produce pulses with peak powers of >1 MW, average powers of several watts, and cavity frequencies from 4 MHz to 2 GHz, but many applications require pulses with higher intensities or lower pulse repetition frequencies. Additional amplification stages can be used to satisfy these requirements and these are commonly implemented in Nd:glass for picosecond pulses and Ti:sapphire for femtosecond pulses. In order to avoid damage and unwanted nonlinear effects caused by high pulse intensities within the amplifier system, most practical sources now apply the chirped-pulse amplification (CPA) approach in which a low-energy femtosecond seed pulse is stretched to subnanosecond durations to reduce its peak power then amplified and finally compressed to its original duration. Using the chirped-pulse amplification technique, pulses with peak powers as high as 1.5 petawatt have been generated.

Two possible amplifier geometries are used which are categorized either as multipass or regenerative amplifiers. In the multipass configuration the injected pulse is refocused many times through the same gain crystal using a system of mirrors. The gain crystal is pumped above the saturation fluence in order to extract the maximum stored energy from the medium per pass. The multipass geometry is more commonly used for high-average-power operation and critical alignment is required to maximize the overlap between the pump mode profile and the multiple beams intersecting the gain medium. The alternative regenerative amplifier is essentially a stable laser resonator which includes a Pockel cell and a polarizing beamsplitter to enable pulses to be switched into and out of the cavity. The gain per pass is considerably lower than in the multipass configuration but regenerative systems produce superior beam quality and their performance can be optimized without affecting the beam pointing of the output. Commercially available Ti:sapphire amplifiers commonly use the regenerative method and are available with pulse energies of up to 1 mJ at repetition frequencies in the 1–300 kHz range.

Further Reading

Agrawal GP (1995) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Duling IN (1995) *Compact Sources of Ultrashort Pulses*. Cambridge University Press.

Fork RL, Martinez OE and Gordon JP (1984) Negative dispersion using pairs of prisms. *Optics Letters* 9: 150–152.

Keller U, Weingarten KJ, Kartner FX, *et al.* (1996) Semiconductor saturable absorber mirrors for femtosecond to nanosecond pulse generation. *Solid-State Lasers, IEEE Selected Topics in Quantum Electronics* 2: 435–453.

Siegman AE (1986) *Lasers*. University Science Books.

Spence DE, Kean PN and Sibbett W (1991) 60-fsec pulse generation from a self-mode-locked Ti-sapphire laser. *Optics Letters* 16: 42–44.

Szipocs R, Ferencz K, Spielmann C and Krausz F (1994) Chirped multilayer coatings for broad-band dispersion control in femtosecond lasers. *Optics Letters* 9: 201–203.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Pulse Characterization Techniques

D J Kane, Southwest Sciences, Inc., Santa Fe, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ultrashort optical pulses have very short time durations, typically less than a few tens of picoseconds. As a result, these pulses are spectrally broad. Because the index of refraction of materials is a function of wavelength, different wavelengths of light travel at different speeds in optical materials, causing the properties of ultrashort optical pulses to change as they propagate. However, the shape of the pulse can influence how the pulse itself interacts with materials. Thus, the goal of ultrafast laser pulse measurement is to obtain not only the intensity profile of the pulse, but also the actual variation of the frequencies that make up the pulse. This is called measuring the intensity and phase of the pulse, respectively.

Even though the development of techniques to characterize ultrashort optical pulses has not been easy, a myriad of pulse measurement techniques have been developed. In this chapter, we will confine our discussion to the most well-known pulse and generally accepted pulse characterization methods in an effort to provide a basic working knowledge of pulse measurement techniques and a fundamental understanding of the principles behind pulse measurement. After a brief discussion of the mathematical representation of ultrashort optical pulses, we will discuss pulse measurement in (roughly) chronological order, starting with autocorrelation methods. The next section introduces a useful method for the measurement of the relative phase between two pulses called spectral interferometry (SI). Following SI, we introduce the notion of the time-frequency representation of ultrashort optical pulses where techniques such as frequency-resolved optical gating (FROG) are discussed. Following the time-frequency section, a relatively new technique known as spectral phase interferometry for direct electric field reconstruction (SPIDER) is discussed.

Mathematical Representation of an Optical Pulse

The time-dependent variations of the pulse are embodied in the pulse electric field, $A(t)$, which can be written:

$$A(t) = \mathrm{Re}[E(t)\mathrm{e}^{\mathrm{i}\omega_0 t}] \qquad [1]$$

where ω_0 is the carrier frequency and Re refers to the real part. While we can use $A(t)$ as it stands for calculations, it is much easier to remove the rapidly varying ω_0 part, $\mathrm{e}^{\mathrm{i}\omega_0 t}$, and use a slowly varying envelope together with a phase term that contains only the frequency variations, not the rapidly varying carrier frequency:

$$E(t) = [I(t)]^{1/2}\mathrm{e}^{\mathrm{i}\varphi(t)} \qquad [2]$$

where $I(t)$ and $\varphi(t)$ are the time-dependent intensity and phase of the pulse. (Note that $E(t)$ is complex.) The frequency variation, $\Omega(t)$, is the derivative of $\varphi(t)$ with respect to time:

$$\Omega(t) = -\mathrm{d}\varphi(t)/\mathrm{d}t \qquad [3]$$

The pulse field can be written equally well in the frequency domain by taking the Fourier transform of eqn [2]:

$$\tilde{E}(\omega) = [\tilde{I}(\omega)]^{1/2}\mathrm{e}^{-\mathrm{i}\phi(\omega)} \qquad [4]$$

where $\tilde{I}(\omega)$ is the spectrum of the pulse, and $\phi(\omega)$ is its phase in the frequency domain. The spectral phase contains time versus frequency information; that is, the derivative of the spectral phase with respect to frequency yields the time arrival of the frequency – the group delay.

Obtaining the intensity and phase, $I(t)$ and $\varphi(t)$ (or $\tilde{I}(\omega)$ and $\phi(\omega)$) is called full characterization of the pulse. Common phase distortions include linear chirp, where the phase (either the time domain or frequency domain) is parabolic. When the frequency is increasing in time, the pulse is said to have positive linear chirp; negative linear chirp is when the high frequencies lead the lower frequencies. Higher-order chirps are common, but for these, differentiation between spectral and temporal chirp is required because spectral phase and temporal phase are not interchangeable.

Autocorrelation

Traditionally, we measure events using shorter events. Unfortunately, for the ultrafast researcher, shorter events do not exist and modern electronics are not fast enough. Therefore, because the shortest event we have is the event we wish to measure, traditional pulse measurement methods use the pulse itself to determine the approximate duration of the pulse in question giving birth to the ubiquitous intensity autocorrelation (see Figure 1). While autocorrelations cannot be used to fully characterize ultrashort optical pulses, the methods used in autocorrelations are fundamental to all pulse measurement schemes.

The intensity autocorrelation is measured by combining a pulse and a delayed replica of a pulse in a nonlinear medium such as a second harmonic generation (SHG) crystal. A pulse is sent onto a beamsplitter to produce the two replicas. One pulse is delayed relative to the other and both are focused together into a SHG crystal. As the delay of one pulse relative to the other is varied, the intensity of the second-harmonic signal is recorded. The intensity autocorrelation is not limited to using SHG; any nonlinearity may be utilized. Indeed, two-photon absorption in a semiconductor LED or photodiode often acts as a convenient nonlinearity. Sometimes a third-order nonlinearity is used to provide some direction of time information about the pulse. When the generation of the signal involves phase matching, such as second harmonic generation, care must be taken to use a thin crystal. Typically, 10 μm to 1 mm thick SHG crystals are used, depending on the bandwidth of the pulse to be measured. The exact thickness depends on the SHG crystal, the pulse width, and the requirements of the measurement.

Regardless of the nonlinearity used, an autocorrelation yields only the approximate duration and

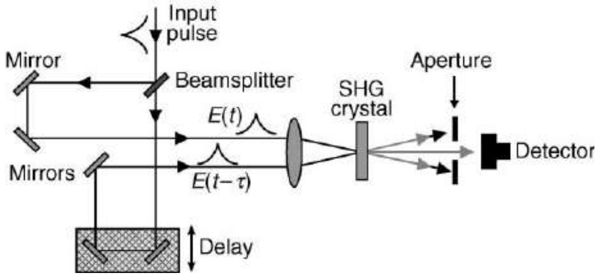

Figure 1 An intensity autocorrelator used to determine the approximate duration of a pulse. The pulse is split into two replicas that are sent into delay lines. The outputs from both delay lines are focused into a doubling crystal. The second harmonic output is monitored as a function of delay between the two pulses. A plot of the signal versus time is the intensity autocorrelation of the pulse.

shape of the pulse. The structure of the pulse is smeared, producing a smooth, featureless profile for even a complex pulse. Thus, the intensity autocorrelation alone does not determine the intensity profile of the pulse, $I(t)$. Only the interaction of the intensity of the pulse is recorded, producing no phase information for the pulse. Some qualitative information can be gleaned from the spectrum and the autocorrelation, but usually only that the pulse is chirped, and higher-order chirps and complex pulse structures elude such an analysis.

Another useful form of intensity autocorrelation is the single-shot autocorrelator. In this case, fairly large beams are used, and these beams are set to intersect at an angle of 2Φ, tilting the pulse fronts, so that delay is mapped onto a spatial coordinate (see Figure 2). The beams are then focused into a nonlinear medium using a cylindrical lens. The interaction region in the SHG crystal is imaged onto a linear array or CCD camera. Typically, when second-harmonic generation is used, the SHG crystal is oriented for Type I phase matching. The time window is proportional to $\Delta w(\sin\Phi)/\Delta v_g$, where Δw is the beam waist at the SHG crystal, and v_g is the group velocity of the pulse in the crystal. Like all SHG interactions, care must be exercised to insure that the phase matching is sufficient to mix the entire bandwidth of the pulse.

Jean-Claude Diels improved the intensity autocorrelation by the development of the interferometric autocorrelation. In this configuration, a Michelson interferometer is typically used so that the beams are collinear when arriving at the SHG crystal, which allows the beams to interfere. As a result, fringes appear on the autocorrelation. By examining the shape and extent of the fringes, some information about the pulse chirp can be obtained. While this does add to the information provided by an intensity autocorrelation, the added information is only the spectrum of the second harmonic. Unfortunately the addition of the second harmonic spectrum does not provide enough information for full retrieval of the intensity and phase.

Even though autocorrelation alone does not completely determine pulse intensity and phase, it is a very simple and useful technique to determine approximate pulse duration. Sometimes, cross-correlation can be used to determine the duration of a long pulse, if a shorter one is available, or a pulse too weak (or at an inappropriate wavelength) to measure with autocorrelation. The addition of the spectrum and/or the second harmonic spectrum can provide more,

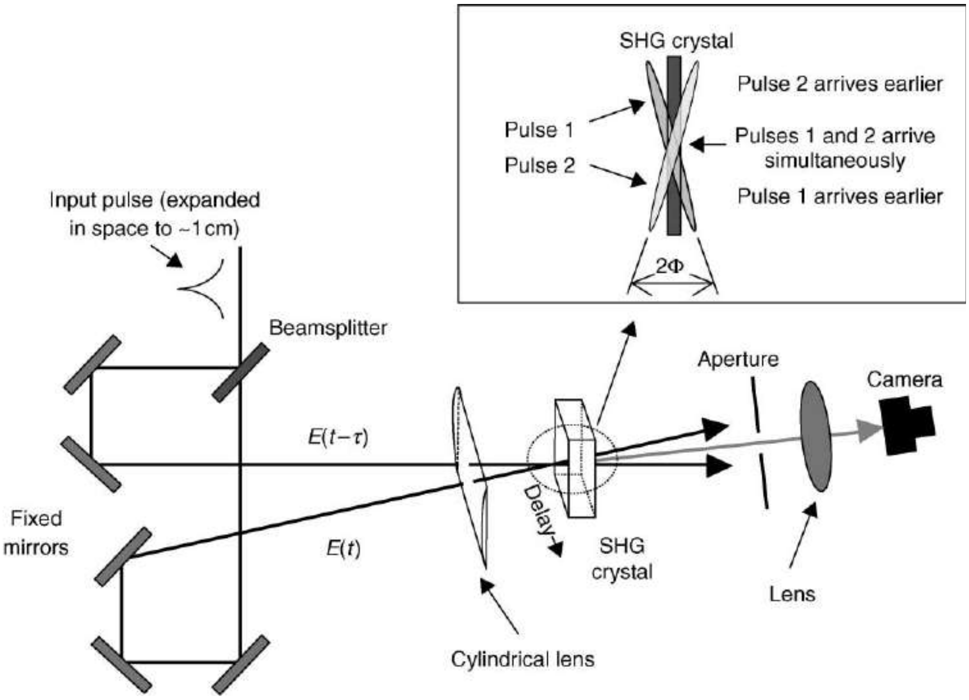

Figure 2 A single shot intensity autocorrelator. For this type of autocorrelator, the beam size is large, on the order of 1 cm in diameter, and the pulse fronts are tilted with respect to each other in order to map relative delay to position. The two beams are focused using a cylindrical lens and the output is recorded using an array detector. The inset shows how delay is mapped into position. Inset: the tilted pulse fronts cause time delay to be mapped spatially.

albeit, incomplete information about pulse chirp. More importantly, the experimental techniques used by correlation methods are fundamental to all pulse measurement methods.

Spectral Interferometry – Relative Phase Measurements

While determination of the absolute intensity and phase of an ultrashort laser pulse is difficult, determination of the relative phase is not. Invented in the late 1800s, spectral interferometry (SI), measures the relative phase between two pulses. SI, a linear technique, can also be modified (albeit, a some-what complex modification) to be self-referencing for full characterization of the ultrashort optical pulse (See the section on SPIDER below). In addition, when combined with a full pulse characterization technique, SI, because it is a linear technique, provides a method by which very weak optical pulses can be characterized (see below).

Typically, an SI apparatus uses a Mach–Zender interferometer (see Figure 3). A pulse is split into two replicas; one is sent through a region or medium to measure and the other through a known path. The two pulses remain separated by some known time and are sent, collinearly, into a spectrometer; no time scanning is required. An analysis of the fringe pattern yields the relative phase between the two pulses. In other words, SI provides $\phi_{\text{unk}}(\omega) - \phi_{\text{ref}}(\omega)$, where $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase versus frequency. The spectrum of the two pulses in the frequency domain is

$$\tilde{I}_{\text{SI}}(\omega) = |\tilde{E}_0(\omega) + \tilde{E}_{\text{unk}}(\omega)|^2$$

$$\tilde{I}_{\text{SI}}(\omega) = \tilde{I}_0(\omega) + \tilde{I}_{\text{unk}}(\omega) + 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)} \times \cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \qquad [5]$$

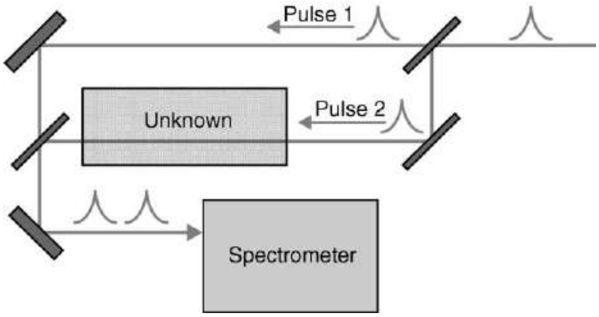

Figure 3 A Mach–Zender interferometer is the typical arrangement for spectral interferometry. One path acts as the reference arm and the sample to be measured is placed in the other arm. Unlike standard interferometry, the delay is fixed in a spectral interferometry experiment.

where $\tilde{I}_{\text{SI}}(\omega)$ is the output spectrum, $\tilde{E}_0(\omega)$ is the reference pulse electric field, $\tilde{E}_{\text{unk}}(\omega)$ is the unknown pulse electric field, $\tilde{I}_0(\omega)$ is the reference pulse spectrum, $\tilde{I}_{\text{unk}}(\omega)$ is the unknown pulse spectrum, $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase, $\phi_0(\omega)$ is the reference pulse phase, and τ is the delay between the two pulses. To facilitate extracting the phase difference from the SI output, τ is chosen to yield fringes in the sum spectrum. The spectral fringes, which have a period inversely proportional to the optical path difference between the two beams (see Figure 4), contain all of the phase difference information.

Figure 4 shows the steps required to obtain the phase difference between the two pulses. The first step is to subtract out the spectra of the individual pulses in order to isolate the spectral interferogram, $S(\omega)$ (Figure 4a), where

$$S(\omega) = 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)}\cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \qquad [6]$$

By Fourier transforming $S(\omega)$, we obtain

$$\Im^{-1}[S(\omega)] = f(t-\tau) + f(-t-\tau) \qquad [7]$$

where $f(t)$ is the correlation product between the reference and the unknown electric field (see Figure 4b). The Fourier transform of eqn [7] multiplied by a Heaviside function, $\Theta(t)$ (to remove $f(-t-\tau)$), recovers the amplitude and phase of $f(\omega)$, the Fourier transform of $f(t-\tau)$, which contains the relative spectral phase between the unknown pulse and the unmodified pulse. Some care, however, needs to be taken to correctly remove the linear phase due to the time delay between the two pulses, $\omega\tau$. Also, for best results, the spectrometer should be well calibrated and care should be taken to properly window the spectrum before taking the Fourier transform.

Spectral interferometry is a straightforward and simple technique that always provides a relative phase. For this very reason, care must be taken to make sure the beams are mode-matched as well as collinear; also, interferometric stability must be maintained over the course of the measurement. Furthermore, SI is nongating; that is, CW background in the laser can add to the interferogram, masking transient effects. Nevertheless, spectral interferometry is a useful technique that has been applied to the measurement of the linear and nonlinear spectral phase introduced by optical fibers. More recently, SI has been applied to phase-locking and to phase-resolved pump probe experiments.

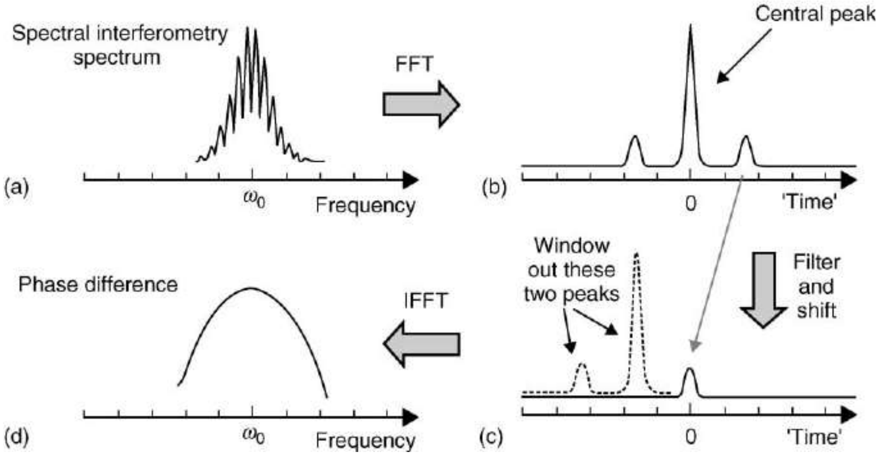

Figure 4 The steps required for the analysis of spectral interferograms. Part (a) shows a sample SI interferogram. The dominant frequency of the fringes is the delay multiplied by the speed of light. The actual phase differences of interest are the perturbations on this frequency. Part (b) is the Fourier transform of the SI interferogram in Part (a). (Note that this is not the true time domain.) The central peak contains only spectral information. All of the phase information is contained in the satellite peaks. The next step is to mask out the central peak and one of the satellite peaks (Part (c)). The phase of the inverse Fourier transform of Part (c) yields the relative phase between the two pulses (Part (d)).

Time-Frequency Representation

In 1971, E.B. Treacy laid the foundations for a revolution in pulse measurement by introducing the idea of measuring the intensity versus time for different spectral splices of a pulse. Because his measurements provided both time and frequency information simultaneously, these measurements could be thought of as applying in a hybrid time-frequency domain. At first, this may seem confusing, but similar methods have been used to visualize sounds patterns (a musical score, for example) and speech. A time-frequency plot of this type, where spectral slices of a pulse are plotted versus time, is called a sonogram. The mathematical formalism of a sonogram is:

$$S_{\mathrm{E}}(\Omega, T) = \left| \int_{-\infty}^{\infty} \tilde{E}(\omega) h(\omega - \Omega) \mathrm{e}^{-\mathrm{i}\omega T} \mathrm{d}\omega \right|^2 \qquad [8]$$

where $\tilde{E}(\omega)$ is the electric field of the pulse to be measured in the frequency domain and $h(\omega - \Omega)$ is a frequency gate that varies with frequency. The magnitude squared of the inverse Fourier transform of spectral slices yields the sonogram.

In 1991, Chilla and Martinez showed that the sonogram could be used to reconstruct the full intensity and phase of the pulse. That is, under certain conditions, the approximate group delay could be determined as a function of frequency from the sonogram by finding the peak time arrival of each spectral slice; integration of the group delay yields the spectral phase, which together, with the pulse spectrum, provides the intensity and phase of the pulse in the frequency domain. They labelled this technique as frequency domain phase measurement (FDPM).

An experimental diagram of an FDPM apparatus is shown in Figure 5. The pulse is split into two replicas, and one of the pulse replicas is spectrally filtered. The spectrally filtered pulse is cross-correlated with the original pulse to find the 'time arrival' of the spectrally filtered pulse – defined as the peak of the cross-correlated pulse. Because the spectrally filtered pulse has a much longer time duration than the original pulse, the small perturbation caused by the finite length of the original pulse is neglected.

The main difficulty of the Chilla–Martinez method is that the frequency gate must be very narrow, reducing the filtered pulse energy significantly, which greatly reduces the measured signal strength. If the spectral phase is not well behaved, the filtered pulse may not be much longer than the original pulse. In addition, for a given frequency, the sonogram may have two or more peaks in time, that is, the group delay may not be a function. However, a method called 2D phase retrieval, discussed in the next section, can alleviate many of these issues.

A spectrogram is a relative of the sonogram. Rather than determining the time arrival of spectral slices of a pulse, a spectrogram is obtained when the spectrum of time slices of the pulse are measured. The spectrogram is experimentally much easier to obtain than the sonogram; however, the measured quantity is not

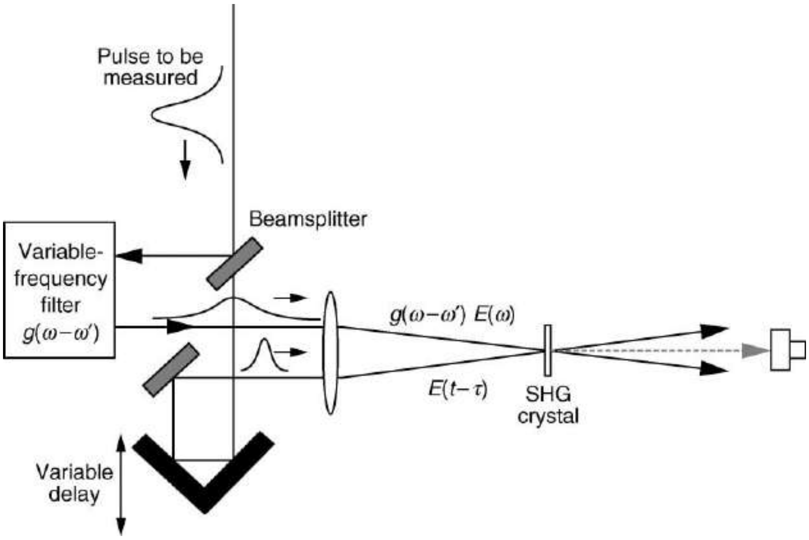
Pulse to be measured
Beamsplitter
Variable-frequency filter $g(\omega-\omega')$
$g(\omega-\omega')\ E(\omega)$
$E(t-\tau)$
SHG crystal
Variable delay

Figure 5 Measuring a sonogram requires determining the time arrival of spectral slices of the pulse. The pulse is split into two replicas using a beamsplitter. One replica is spectrally filtered using a tunable filter. The other pulse is cross-correlated with the spectrally filtered pulse to determine the time arrival.

quite as mathematically useful as the sonogram because the time slicing of the pulse occurs in the time domain. Thus, if a narrow time gate is used, the time domain phase is obtained. The time domain phase together with the intensity of the pulse, $I(t)$, provides the full intensity and phase of the pulse. Unfortunately, the intensity profile is not as readily measurable as the pulse spectrum. Consequently, to obtain the full intensity and phase of a pulse from its spectrogram requires an iterative 2D phase retrieval algorithm. This is the basis of a pulse measurement technique called frequency-resolved optical gating (FROG).

Frequency-Resolved Optical Gating

Frequency-resolved optical gating (FROG), developed by Kane and Trebino, measures the spectrum of a particular temporal component of the pulse (see Figures 6 and 7) by spectrally resolving the signal pulse in an autocorrelation-type experiment using an instantaneously responding nonlinear medium. As shown in Figure 6, FROG involves splitting a pulse and then overlapping the two resulting pulses in an instantaneously responding $\chi^{(3)}$ or $\chi^{(2)}$ medium. Even though any instantaneous nonlinear interaction may be used to implement FROG, perhaps the most intuitive is the polarization-gating configuration. In this case, induced birefringence, due to the electronic Kerr effect, is used as the nonlinear-optical process. In other words, the 'gate' pulse causes the $\chi^{(3)}$ medium, which is placed between two crossed polarizers, to become slightly birefringent. The polarization of the 'gated' probe pulse (which is cleaned up by the first polarizer) is rotated slightly by the induced birefringence allowing some of the 'gated' pulse to leak through the second polarizer. This is referred to as the signal. Because most of the signal emanates from the region of temporal overlap between the gate pulse and the probe pulse, the signal pulse contains the frequencies of the 'gated' probe pulse within this overlap region. The signal is then spectrally resolved, and the signal intensity is measured as a function of wavelength and delay time τ. The resulting trace of intensity versus delay and frequency is a spectrogram, a time- and frequency-resolved transform that intuitively displays time-dependent spectral information of a waveform.

The spectrogram can be expressed as:

$$S_{\mathrm{E}}(\omega,\tau)=\left|\int_{-\infty}^{\infty}E(t)g(t-\tau)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t\right|^2 \qquad [9]$$

where $E(t)$ is the measured pulse's electric field, $g(t-\tau)$ is the variable-delay gate pulse, and the subscript E on S_{E} indicates the spectrogram's dependence on $E(t)$. The gate pulse $g(t)$ is usually somewhat shorter in length than the pulse to be measured, but not infinitely short. This is an important point: an infinitely short gate pulse yields only the intensity $I(t)$ and conversely, a CW gate yields only the spectrum $I(\omega)$. On the other hand,

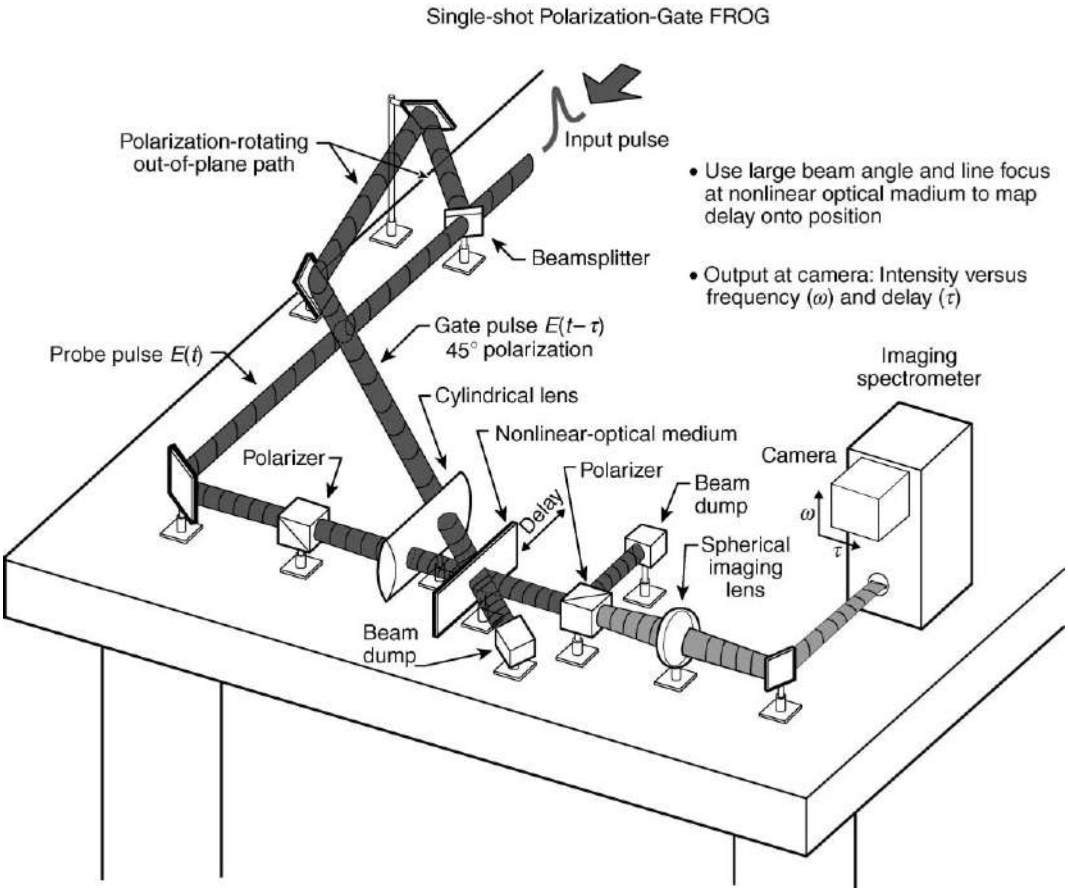
Single-shot Polarization-Gate FROG
Polarization-rotating out-of-plane path
Input pulse
• Use large beam angle and line focus at nonlinear optical medium to map delay onto position
Beamsplitter
• Output at camera: Intensity versus frequency (ω) and delay (τ)
Gate pulse $E(t-\tau)$ 45° polarization
Probe pulse $E(t)$
Imaging spectrometer
Cylindrical lens
Nonlinear-optical medium
Camera
Polarizer
Polarizer
Beam dump
Delay
ω
τ
Spherical imaging lens
Beam dump

Figure 6 Measuring the spectrogram of a pulse is easier than measuring its sonogram – a spectrogram is a spectrally resolved autocorrelation. In this figure, the optical Kerr–effect is used (polarization-gate) as the nonlinearity. (Adapted with permission from Kane DJ and Trebino R (1993) Single shot measurement of the intensity and phase of an arbitrary ultrashort pulse by using frequency-resolved optical gating. *Optics Letters* 18(10): 823–825.)

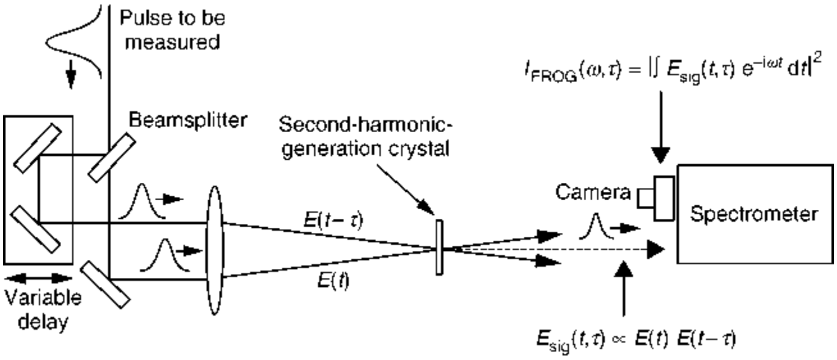
Pulse to be measured
$I_{FROG}(\omega,\tau) = |\int E_{sig}(t,\tau)\, e^{-i\omega t}\, dt|^2$
Beamsplitter
Second-harmonic-generation crystal
Camera
Spectrometer
$E(t-\tau)$
$E(t)$
Variable delay
$E_{sig}(t,\tau) \propto E(t)\, E(t-\tau)$

Figure 7 This figure shows a schematic of an SHG FROG device. The SHG signal from the autocorrelation is spectrally resolved. SHG FROG is very simple and sensitive, but it has a direction-of-time ambiguity. For example, if the pulse has chirp, only the magnitude of the chirp is determined – the sign of the chirp remains unknown.

a finite-length gate pulse yields the spectrum of all of the finite pulse segments with duration equal to that of the gate. While the phase information remains lacking in each of these short-time spectra, having spectra of an infinitely large set of pulse segments compensates for this loss. The spectrogram has been shown to nearly uniquely determine both the intensity $I(t)$ and phase $\varphi(t)$ of the pulse, even if the gate pulse is longer than the pulse to be measured (although if the gate is too long, sensitivity to noise and other practical problems arise).

In FROG, when using optically induced birefringence as the nonlinear effect, the signal pulse is given by:

$$E_{\mathrm{sig}}(t,\tau) \propto E(t)|E(t-\tau)|^2 \qquad [10]$$

So the measured signal intensity $I_{\mathrm{FROG}}(\omega,\tau)$, after the spectrometer is:

$$I_{\mathrm{FROG}}(\omega,\tau) = \left| \int_{-\infty}^{\infty} E(t)|E(t-\tau)|^2 \mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t \right|^2 \qquad [11]$$

We see that the FROG trace is a spectrogram of the pulse $E(t)$ although the gate, $|E(t-\tau)|^2$, is a function of the pulse itself.

To see that the FROG trace essentially uniquely determines $E(t)$ for an arbitrary pulse, it is first necessary to observe that $E(t)$ is easily obtained from $E_{\mathrm{sig}}(t,\tau)$. Then it is simply necessary to write eqn [11] in terms of $E_{\mathrm{sig}}(t,\Omega)$, the Fourier transform of the signal field $E_{\mathrm{sig}}(t,\tau)$, with respect to delay variable τ. We then have what appears to be a more complex expression, but one that will give us better insight into the problem:

$$I_{\mathrm{FROG}}(\omega,\tau) = \left| \int_{-\infty}^{\infty} E_{\mathrm{sig}}(t,\Omega)\mathrm{e}^{-\mathrm{i}\omega t-\mathrm{i}\Omega\tau}\mathrm{d}t\,\mathrm{d}\Omega \right|^2 \qquad [12]$$

Equation [12] indicates that the problem of inverting the FROG trace $I_{\mathrm{FROG}}(\omega,\tau)$, to find the desired quantity $E_{\mathrm{sig}}(t,\Omega)$, is that of inverting the squared magnitude of the two-dimensional (2D) Fourier transform of $E_{\mathrm{sig}}(t,\Omega)$. This problem, which is called the 2D phase-retrieval problem, is well known in many fields, especially in astronomy, where the squared magnitude of the Fourier transform of a 2D image is often measured. At first glance, this problem appears unsolvable; after all, much information is lost when the magnitude is taken. It is well known that the one-dimensional (1D) phase retrieval problem is unsolvable (for example, infinitely many pulse fields give rise to the same spectrum). Intuition fails in this case, however; two- and higher-dimensional phase retrieval essentially always yields unique results.

The Simplified FROG – GRENOUILLE

A simplified version of the FROG device, known as grating eliminated no-nonsense observation of ultrafast incident laser light e-fields (GRENOUILLE), can be constructed using the doubling crystal as both the gating medium and the spectrometer (see Figure 8). In this case, the phase matching condition in certain thick doubling crystals causes a wavelength dependent output from the doubling crystal. A cylindrical lens is used in a Fourier transform configuration to image the wavelength variation onto a CCD camera. A cylindrical lens set at 90 degrees to the Fourier transform lens images the time axis onto the CCD camera. A Fresnel biprism replaces the traditionally used Mach–Zender interferometer to produce two pulses propagating at an angle with respect to each other. Presently, the GRENOUILLE technique works with BBO, in the wavelength region around 500 nm–1200 nm, and with proustite in the wavelength region around 1200 nm–2000 nm.

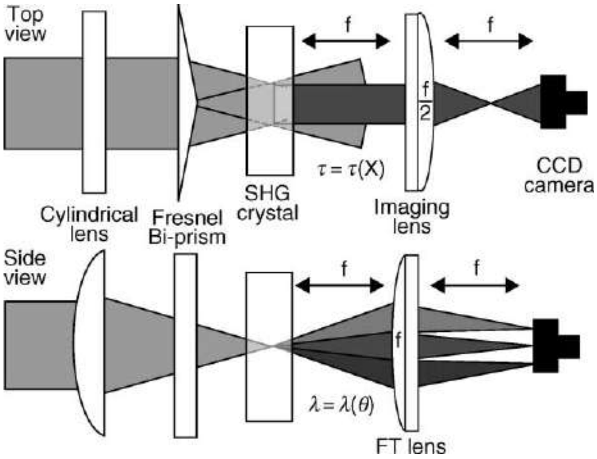

Figure 8 Figure showing a schematic of a GRENOUILLE FROG device which uses a thick doubling crystal as both the gating nonlinearity and the spectrometer. Like single-shot autocorrelators, a cylindrical lens focuses a spatially large input beam into an SHG crystal in only one dimension. Unlike most autocorrelators, the Mach–Zender interferometer is replaced with a Fresnel biprism that forces each half of the input beam to propagate, at an angle, toward the SHG crystal. An imaging cylindrical lens images the spatial dependence of the delay onto a CCD camera. A Fourier transform cylindrical lens maps the angular dependence of wavelength from the SHG crystal to a spatial coordinate on the CCD camera. (Adapted with permission from O'Shea P, Kimmel M, Xun G and Trebino R (2001) Highly simplified device for ultrashort-pulse measurement. *Optics Letters* 26(12): 932–934.)

FROG Inversion Algorithms

An iterative 2D phase retrieval algorithm is required to extract the pulse information from the measured FROG trace (see Figures 9 and 10). This algorithm converges to a pulse that minimizes the difference

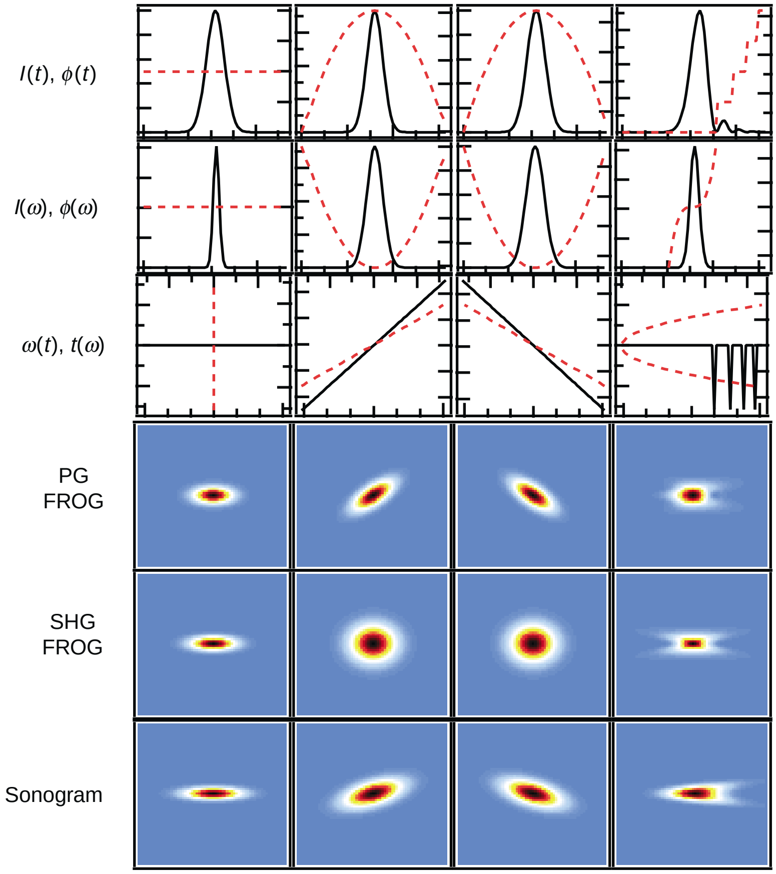

Figure 9 Example FROG traces and sonograms are shown for four different pulses: a transform limited pulse, a positively chirped pulse, and a negatively chirped pulse. The top series of plots are the time domain representation of the pulse intensity (solid line) and phase (dashed line). The second row of four plots is the frequency domain representation of the sample pulse's intensity and phase. The next four plots show the instantaneous frequency and the group delay. The vertical axis is frequency and the horizontal axis is time. The remaining rows show the PG FROG traces, the SHG FROG traces and sonograms of the pulse shown in the first two rows.

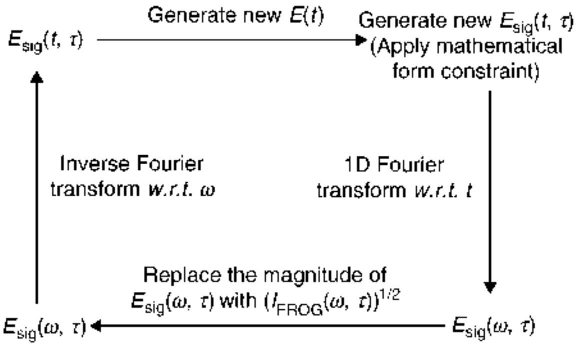
$E_{sig}(t, \tau)$
Generate new $E(t)$
Generate new $E_{sig}(t, \tau)$ (Apply mathematical form constraint)
Inverse Fourier transform *w.r.t.* ω
1D Fourier transform *w.r.t.* t
Replace the magnitude of $E_{sig}(\omega, \tau)$ with $(I_{FROG}(\omega, \tau))^{1/2}$
$E_{sig}(\omega, \tau)$
$E_{sig}(\omega, \tau)$

Figure 10 Phase retrieval algorithm for the inversion of FROG spectrograms. Start with an initial guess for the pulse to generate an initial $E_{sig}(t,\tau)$. A 1D Fourier transform generates the FROG trace. The next step is to replace the magnitude of the calculated FROG trace with the square root of the measured FROG trace. Inverse Fourier transform with respect to ω to produce the new signal field and generate a new guess for $E(t)$. Interestingly, it is only the step that produces $E(t)$ from $E_{sig}(t,\tau)$ that differentiates all the FROG algorithms.

between the measured and the calculated FROG trace. While this aspect of FROG has been its Achilles heel in the past, in reality, new generalized projections algorithms (together with faster computers) converge quickly and can track pulse changes at rates up to 30 Hz and beyond, making FROG a true, real-time pulse measurement technique. Indeed, excellent algorithms for the analysis of FROG traces in real-time are commercially available.

The original FROG inversion algorithm, commonly referred to as the vanilla algorithm, is simple and iterates quickly, but tends to stagnate, giving erroneous results, especially for geometries that use a complex gate function such as SHG or self-diffraction. An improved algorithm, incorporating several different algorithms including brute force minimization, was developed to alleviate stagnation programs at the expense of both speed and convergence time. By the mid-1990s, a significant advance in both speed and stability was made by the addition of a numerical method called generalized projections. This algorithm determines the next guess by constructing a projection – minimizing the error (distance) between the FROG electric field, $E_{\mathrm{sig}}(t, \tau)$, obtained immediately after the application of the intensity constraint, and the FROG electric field calculated from the mathematical form constraint.

The first generalized projections algorithm used a standard minimization procedure to find the electric field for the next iteration, which can still be slow. For the most common FROG geometries, PG and SHG, a different algorithm that determines the next iteration directly has been developed. This algorithm, called the principal components generalized projections (PCGP) algorithm, converts the generalized projections algorithm into an eigenvector problem. Using this algorithm, pulse measurement rates of at least 30 Hz have been achieved.

The goal of the phase retrieval algorithm is to find the $E(t)$ that satisfies two constraints. The first is the FROG trace itself which is the magnitude squared of the 1D Fourier transform of $E_{\mathrm{sig}}(t, \tau)$:

$$I_{\mathrm{FROG}}(\omega, \tau) = \left| \int_{-\infty}^{\infty} E_{\mathrm{sig}}(t, \tau) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t \right|^2 \qquad [13]$$

The other constraint is the mathematical form of the signal field, $E_{\mathrm{sig}}(t, \tau)$, for the nonlinear interaction used. The signal forms for a variety of FROG beam geometries are:

$$E_{\mathrm{sig}}(t, \tau) \propto \begin{Bmatrix} E(t)|E(t-\tau)|^2 & \text{PG FROG} \\ E(t)^2 E^*(t-\tau) & \text{SD FROG} \\ E(t)E(t-\tau) & \text{SHG FROG} \\ E(t)^2 E(t-\tau) & \text{THG FROG} \end{Bmatrix} \qquad [14]$$

where PG is polarization gate, SD is self-diffraction, SHG is second harmonic generation, and THG is third harmonic generation FROG.

All FROG algorithms work by iterating between two different data sets: the set of all signal fields that satisfy the data constraint, $I_{\mathrm{FROG}}(\omega, \tau)$, and the set of all signal fields that satisfy eqn [14]. The difference between the FROG algorithms is how the iteration between the two sets is completed. In the case of generalized projections, the $E(t)'$s are chosen such that the distance between the $E(t)$ on the magnitude set and the $E(t)$ on the mathematical form set is minimized. This is accomplished by minimizing the following equation:

$$Z = \sum_{i,j=1}^{N} \left| E^{(k)}_{\mathrm{sig(DC)}}(t_i, \tau_j) - E^{(k+1)}_{\mathrm{sig(MF)}}(t_i, \tau_j) \right|^2 \qquad [15]$$

where $E^{(k)}_{\mathrm{sig(DC)}}(t_i, \tau_j)$ is the signal field generated by the data constraint, and $E^{(k+1)}_{\mathrm{sig(MF)}}(t_i, \tau_j)$ is the signal field produced from one of the beam geometry equations in eqn [14]. For the normal generalized projection, the minimization of Z is completed using a standard steepest descent algorithm; the derivative of Z with respect to the signal field is computed to determine the direction of the minimum. The computation of the derivatives are tedious; they are tabulated elsewhere.

An alternative to an algorithm that minimizes eqn [15] is the PCGP algorithm. This algorithm converts the computation of the next guess into an eigenvector problem, reducing the computation of the next guess to simple matrix–vector multiplications. This algorithm works for both the PG and SHG beam geometries, it is robust and the fastest FROG algorithm. Indeed, the PCGP algorithm was used in the original real-time FROG work by D.J. Kane.

Self Checks in FROG Measurements

Unlike any other pulse measurement techniques, FROG can provide a great deal of feedback about both the quality of the measurement (systematic errors) and the quality of the algorithm's performance. The most common check for convergence is the FROG trace error together with a visual comparison between the retrieved FROG trace and the measured FROG trace. The FROG trace error is given by:

$$G = \sqrt{\frac{1}{N^2} \sum_{i,j=1}^{N} \left| I_{\mathrm{FROG}}(\omega_i, \tau_j) - \alpha I^{(k)}_{\mathrm{FROG}}(\omega_i, \tau_j) \right|^2} \qquad [16]$$

where α is a renormalization constant, I_{FROG} is the measured trace, and $I^{(k)}_{\mathrm{FROG}}(\omega, \tau)$ is computed from the retrieved electric field. Typically, the FROG trace error of a PG measurement should be less than 2% for a 64 × 64 pixel trace, while the FROG trace error of a 64 × 64 pixel SHG FROG trace should be about 1% or less. Acceptable FROG trace errors decrease as FROG trace size increase for smaller FROG traces. These values are only rules of thumb only;

for example, acceptable retrievals of large and very complicated FROG traces can produce larger FROG trace errors.

In a good FROG measurement, the spectrum of the retrieved pulse should faithfully reproduce the salient features of the pulse's measured spectrum. SHG FROG even provides an additional check called the frequency marginal. The sum of an SHG FROG trace along the time axis yields the autoconvolution of the pulse's spectrum, providing two ways the FROG measurement can be checked. First, the autoconvolution of an independently measured spectrum can be compared to the sum of the FROG trace along the time axis, providing an indication of how well the measurement was made. For example, if the doubling crystal was too thick in the pulse measurement, the FROG trace's frequency marginal will be narrower than the autoconvolution of the measured spectrum. Second, comparing the autoconvolution of the retrieval pulse spectrum with the FROG trace marginal can provide a test of algorithm convergence in addition to a test of the measurement. Phase matching problems appear as a mismatch between the FROG trace frequency marginal and the autoconvolution of the retrieved spectrum.

Because FROG is a spectrally resolved autocorrelation, summing any FROG trace along the frequency axis yields the autocorrelation of the measured pulse. This autocorrelation can be compared to an independently measured autocorrelation, or a comparison can be made between the frequency sum of the FROG trace and the autocorrelation calculated from the retrieved pulse to determine algorithm convergence and the quality of the measurement.

Measuring Pulses Directly – SPIDER

SPIDER is a novel technique that measures the pulse directly – no iterative phase retrieval method is required. To accomplish this feat, spectral interferometry is conducted on two pulse replicas that are shifted in frequency with respect to one another. The original pulse is split into two replicas. The two pulse replicas are sent into a phase modulator that shifts the center frequency of each pulse slightly. The frequency-shifted pulses are sent into a spectrometer and a standard spectral interferometry analysis yields the derivative of the phase. Integration of the phase derivative, together with the spectrum of pulse, provides the full intensity and phase.

Thus, analysis of the spectral interferogram gives:

$$\Delta\phi = (\phi(\omega_1) - \phi(\omega_2))\Delta\omega \qquad [17]$$

where $\Delta\phi$ is the measured phase difference from the spectral interferogram, $\Delta\omega$ is the difference in the center frequency between the two pulses, and $\phi(\omega)$ is the pulse phase as a function of frequency. Consequently, the phase difference must be divided by the frequency difference between the two pulses to determine the true derivative. It is also important to subtract out the linear phase difference resulting from the delay between the two pulses (the delay produces the fringes in the resulting interferogram) as this integrates to a linear chirp.

Because phase modulators are too slow to be used in the femtosecond region, a different method is used to shift the frequency of the two pulses – they are mixed with a highly chirped pulse (see Figure 11).

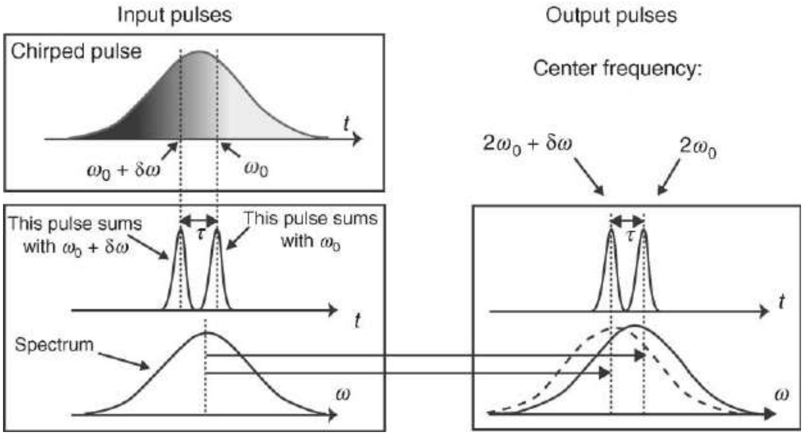
Input pulses
Chirped pulse
t
$\omega_0 + \delta\omega$
ω_0
This pulse sums with $\omega_0 + \delta\omega$
This pulse sums with ω_0
τ
t
Spectrum
ω
Output pulses
Center frequency:
$2\omega_0 + \delta\omega$
$2\omega_0$
τ
t
ω

Figure 11 Figure showing how a SPIDER device works. The highly chirped pulse has a frequency that varies with time. Mixing the chirped pulse with two time delayed replicas produces two pulses, separated in time, that have slightly different frequencies, but are otherwise identical. By interfering the two pulses in a spectrometer, the relative phase between $\omega_0 + \delta\omega$ portions of the pulse and ω_0 portions of the pulse are determined.

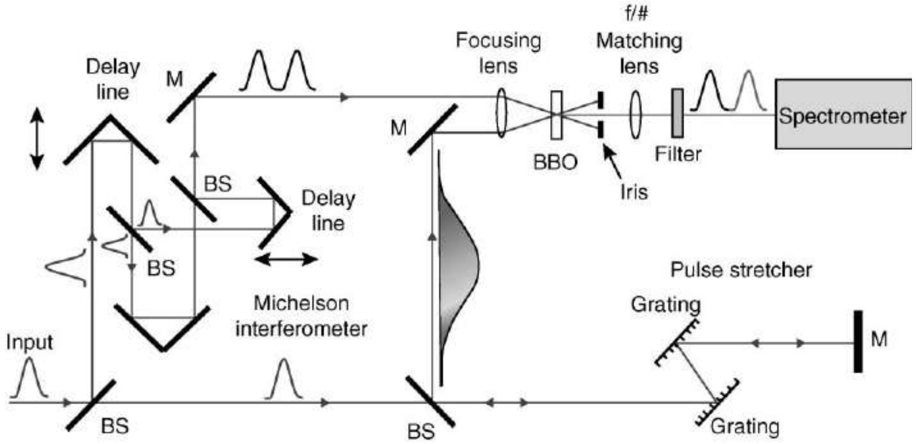

Figure 12 SPIDER apparatus. The input pulse is first split into two pulses. One pulse is stretched in a pulse stretcher. The other pulse is sent into a Michelson interferometer to produce two time-delayed pulses. The two time-delayed pulses are mixed with the stretched pulse to produce two pulses at a slightly different center frequency. The two pulses are sent into a spectrometer to produce a spectral interferogram. Analysis of the spectral interferogram yields the derivative of the pulse phase.

Because the pulses are slightly time delayed, each one aligns with a slightly different frequency in the chirped pulse. The chirped pulse must be so highly chirped that the pulse frequency does not change significantly over the duration of the pulse to be measured.

A schematic of a SPIDER device for the measurement of femtosecond laser pulses is shown in Figure 12. The input pulse is first split into two replicas. The first replica is sent to a pulse stretcher to produce to highly chirped pulse. The other replica is sent into a Michelson interferometer (or a simple étalon may sometimes suffice) to create two time-delayed replicas. The time-delayed replicas are mixed with the chirped pulse. Because each replica is mixed with a slightly different frequency in the chirped pulse, each replica produces a pulse with a slightly different center frequency. When the time-delayed replicas are spectrally resolved, each frequency of the pulse interferes with a frequency-shifted portion of the same pulse. Thus, the measured phase difference is proportional to the derivative of the phase of the original pulse.

Measuring Weak Pulses

Because virtually all the pulse measurement techniques, that do not require a reference pulse, use a nonlinearity, measuring weak ultrafast pulses directly is difficult. Fortunately, because most weak pulses are generated from a stronger pulse, spectral interferometry can be used to determine the relative phase between the weaker pulse and the stronger pulse (for regions where there is spectral overlap between the two pulses). If the stronger pulse is characterized using another pulse measurement technique, then the weaker pulse phase can be determined from the relative phase measurement and the known phase of the stronger pulse. For example, the Trebino group termed the combination of FROG and SI as TADPOLE (Temporal Analysis by Dispersing a Pair Of Light E-fields).

Which Pulse Measurement Method Should I Use?

Choosing a pulse measurement method depends on the required measurement as well as the precision and accuracy required. All the pulse measurement techniques have advantages and disadvantages, but excellent measurements have been made with all of them. Autocorrelation is adequate when all that is required is the approximate pulse duration without any phase information. Spectral interferometry is quite useful for measuring transient changes in a sample in pump-probe experiments. FROG is perhaps the most commonly used and general-purpose pulse measurement technique, providing a great deal of feedback about the quality of the pulse measurement. Inexpensive and convenient FROG devices, together with software, are available commercially, adding to their acceptance. While SPIDER is experimentally complex and does not have the cross checks that FROG has, it provides the quickest answer, requiring no iterative phase retrieval algorithm to obtain pulse intensity and phase.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant numbers DMI-9801116 and DMI-0091454. The author would also like to acknowledge Rick Trebino for the use of figures from his lectures.

List of Units and Nomenclature

$A(t)$	Full electric field
$E(t)$	Complex electric field without carrier frequency
$E_{\mathrm{sig}}(t, \tau)$	Complex electric field of the FROG trace versus time and time delay
$E_{\mathrm{sig}}(t, \Omega)$	Fourier transform of $E_{\mathrm{sig}}(t, \tau)$ with respect to τ
$\tilde{E}(\omega)$	Complex electric field in the frequency domain
$f(t)$	Correlation product
$f(\omega)$	Fourier transform of the correlation product
$g(t - \tau)$	time gate or filter
G	FROG trace error
$h(\omega - \Omega)$	frequency gate or filter
$I(t)$	Intensity profile of the electric field in the time domain
$\tilde{I}(\omega)$	spectral intensity
$I_{\mathrm{FROG}}(\omega, \tau)$	Intensity of the FROG trace
$\Theta(t)$	Heaviside function $= 0 \;\; t < 0, \;\; = 1 \;\; t \geqq 0$
$\phi(\omega)$	Frequency domain phase
$S_{\mathrm{E}}(\omega, \tau)$	Spectrogram
$S_{\mathrm{E}}(\Omega, T)$	Sonogram
$S(\omega)$	Spectral interferogram
t, T	variables for time
τ	Fourier transform
v_{g}	group velocity
$\varphi(t)$	Time domain phase
Δw	Beam waist
Z	Distance metric between the FROG signal field generated by the data constraint and the FROG signal field generated by the mathematical form constraint
Φ	Geometric angle between two beams
$\Delta\phi$	Phase difference
$\Pi^{(n)}$	nth order nonlinear susceptibility
ω	Primary variable for frequency
$\Delta\omega$	Difference in center frequency between two pulses
Ω	Secondary variable for frequency
$\Omega(t)$	Instantaneous frequency

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses. **Ultrafast Technology:** Femtosecond Chemical Dynamics: Gas-Phase; Femtosecond Condensed Phase Spectroscopy: Structural Dynamics; Ultrafast Illumination and Processing.

Further Reading

Diels JC and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena*. San Diego, CA: Academic Press Inc.

Dorrer C, Belabas N, Likforman J-P and Joffre M (2000) Spectral resolution and sampling issues in Fourier-transform spectral interferometry. *Journal of the Optical Society B* 17: 1795.

Kane DJ, Weston J and Chu K-CJ (2003) Real-time inversion of polarization gate frequency-resolved optical gating spectrograms. *Applied Optics* 42: 1140–1144.

Iaconis C and Walmsley IA (1999) Self-referencing spectral interferometry for measuring ultrashort optical pulses. *IEEE Journal of Quantum Electronics* 35: 501–509.

Lepetit L, Cheriaux G and Joffre M (1995) Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy. *Journal of the Optical Society of America B* 12: 2467–2474.

Rulliere C (ed.) (2000) *Femtosecond Laser Pulses*. New York: Springer Verlag.

Shuman TM, Anderson ME, Bromage J, Iaconis C, Waxer L and Walmsley IA (1999) Real-time SPIDER: ultrashort pulse characterization at 20 Hz. *Optics Express* 5: 134–143.

Treacy EB (1971) Measurement and interpretation of dynamic spectrograms of picosecond light pulses. *Journal of Applied Physics* 42: 3848–3858.

Trebino R (2000) *Frequency Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses*. Boston, MA: Kluwer Academic Publisher.

Trebino R, DeLong KW, Fittingholf DN, Sweetser JN, Krunbugel MA, Richman BA and Kane DJ (1997) Measuring ultrashort laser pulses in the time-frequency domain using frequency-resolved optical gating. *Review of Scientific Instruments* 68: 3277–3295.

Various authors (1999) Feature issue on ultrashort-laser-pulse measurement and applications. *IEEE Journal of Quantum Electronics* 35: 4.

ULTRAFAST TECHNOLOGY

Contents

Femtosecond Chemical Dynamics: Gas-Phase
Femtosecond Condensed Phase Spectroscopy: Structural Dynamics
Ultrafast Illumination and Processing

Femtosecond Chemical Dynamics: Gas-Phase

M Dantus, Michigan State University, East Lansing, MI, USA

E J Brown, McDaniel College, Westminster, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Chemical reactions are at the heart of chemistry; the quest to understand them has sparked research and study along many different avenues. The fundamental dynamics involved in bond breakage, bond formation, and transition states take place on a very fast time-scale, usually in the tens to hundreds of femtoseconds, where a femtosecond (fs) is equal to 10^{-15} seconds. Therefore, ultrashort laser pulses are required to measure these fundamental processes as well as the rotational and vibrational motion before and after the reaction. The whole set of measurements can then be used for mapping a potential energy surface for the reaction. As a result of these studies, one learns the key details that are required to develop a picture of how chemical reactions occur. The measurement of these fundamental dynamics, using ultrashort laser pulses, is the subject of this article.

To illustrate how long it takes to form or break a chemical bond, one needs to consider the following analogy. The time it takes to break a chemical bond is approximately a tenth of a trillionth of a second (10^{-13} s). During such a short time, light can travel only 0.03 mm in vacuum. This time is so short that it cannot be measured electronically. Its measurement requires the use of ultrashort laser pulses, with pulse durations of 50 femtoseconds or shorter. The development of methods aimed at measuring the fundamental steps involved in chemical reactions culminated in the development of femtosecond transition state spectroscopy (FTS), a method developed in the 1980s by Zewail and co-workers at Caltech. This work allows researchers to observe bond breakage and formation, vibrational and rotational dynamics, and transition state configurations, as the reaction occurs. The Nobel Prize in Chemistry was awarded in 1999 to Ahmed Zewail for this development. Since the 1980s, femtosecond lasers have been used to study molecular dynamics and chemical dynamics on very fast time-scales. The field, known as femtochemistry, has grown quickly and diversified to involve hundreds of scientists around the world. This article discusses the measurement of ultrafast dynamics measured in gas-phase samples. The presentation encompasses some of the measurement techniques and the fundamental values to be measured and concludes with an outlook towards the future.

From Nanosecond to Picosecond to Femtosecond Reaction Kinetics

The quest for shorter light pulses to study chemical reactions predates the invention of the laser. In the early days, electric discharges were used to measure fast dynamics in the millisecond (10^{-3} s) and then the microsecond (10^{-6} s) time-scales. With the invention of the laser, measurements improved drastically with temporal resolution going from the nanosecond (10^{-9} s) to the picosecond (10^{-12} s) time-scales. Since the early days, it was recognized that the best time resolution was obtained when pairs of light pulses were used – one to initiate the reaction and the second to probe it. These pump-probe measurements have progressed with the laser technology almost to the single femtosecond level.

The measurements being considered here go beyond the measurement of fast reaction kinetics. When a chemical reaction occurs in a beaker, one usually measures a statistical rate of reaction that involves diffusion of reagents in a solvent. These kinetic measurements, which involve an ensemble of molecules, are much slower and do not reveal the fundamental steps of the reaction such as bond formation. The introduction of ultrafast pulses allowed, for the first time, the measurement of the motion of the atoms in real time, as if the motion was frozen by ultrafast flash photography. Capturing this motion helps to reveal the internal forces, described by a potential energy surface, that act on the atoms in the molecule during the chemical reaction. This concept is illustrated in Figure 1, where we see the reagents and the products for the reaction. Typically, the starting and ending point of a chemical reaction are very well known and the compounds are very well characterized. However, how the reagents become the products is not always known. Chemists typically deduce a reaction mechanism based on a large body of experiments where different parameters including the structure of the reagents are modified to evaluate their effect on the outcome of the reaction. Unfortunately in most cases, reaction mechanisms, even when consistent with all available experimental evidence, may not accurately reflect how a particular chemical reaction takes place.

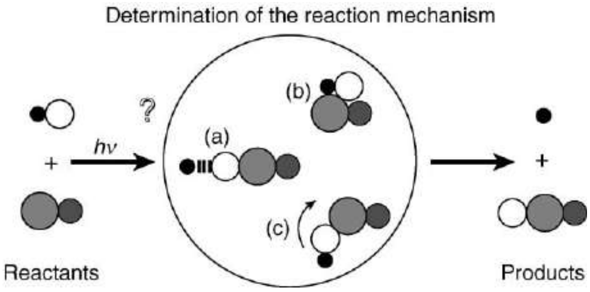
Determination of the reaction mechanism
?
hν
+
(a)
(b)
(c)
+
Reactants
Products

Figure 1 A sketch of a chemical reaction, AB + CD → A + BCD. The reactants and products are well known, but it is not known how they change from reactants to products. This mechanism may involve a transition state which is (a) a linear complex with vibrational motion, (b) a nonlinear complex, (c) a bent complex with rotational motion, or many other possibilities and combinations. The elucidation of the mechanism and transition state dynamics are the goals of femtosecond time-resolved studies.

The ideal method to make the determination involves direct observation. As mentioned earlier, direct observation requires femtosecond laser pulses, just like a fast camera shutter is required to take pictures of fast-moving objects. The concept of a femtosecond pump-probe measurement is illustrated in Figure 2. The chemical reaction is initiated at time $t = 0$ fs by the pump laser. The pump laser provides the energy required for initiating the chemical transformation. The probe laser, delayed in time, probes the formation of the product. Figure 2 illustrates a number of pump-probe measurements

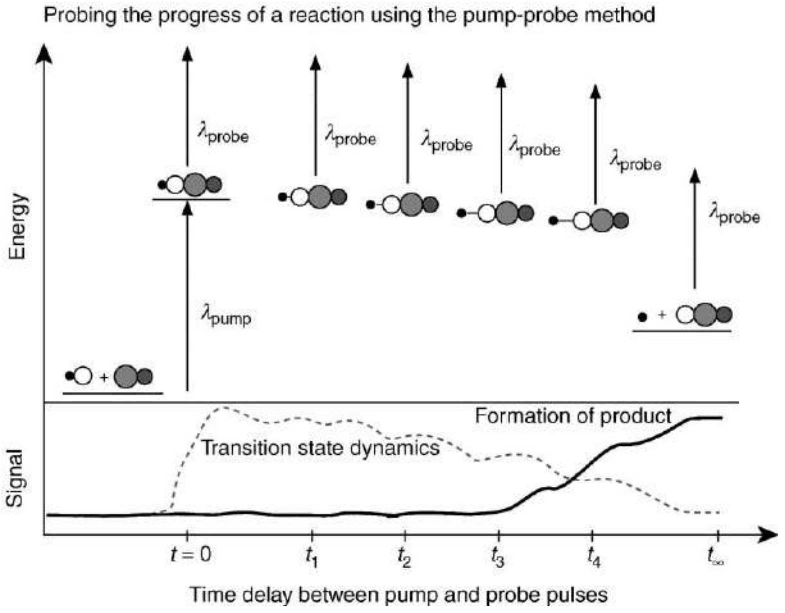
Probing the progress of a reaction using the pump-probe method
Energy
λ_{probe}
λ_{probe}
λ_{probe}
λ_{probe}
λ_{probe}
λ_{probe}
λ_{pump}
+
+
Signal
Formation of product
Transition state dynamics
$t = 0$
t_1
t_2
t_3
t_4
t_∞
Time delay between pump and probe pulses

Figure 2 Energy schematic of the reaction of AB + CD → A + BCD assuming vibrational motion in a linear transition state. The pump pulse provides enough energy to excite the reactants to the transition state at $t = 0$. This transition state is probed by the second pulse. This pulse arrives at different time delays, from no delay ($t = 0$), to increasing delays (t_1, t_2,…), to infinite delay (t_∞). The signal varies depending on the time delay between the two pulses and these changes in signal intensity correspond to vibrational and rotational information about the transition state. To obtain a transient as shown at the bottom, the time delay between the pump and probe pulses is scanned. The dashed line shows the transient that would be obtained when the transition state is monitored; the solid line shows the transient that would be obtained when the product is monitored.

each obtained at a different time delay between the pump and probe laser pulses. When a series of pump-probe measurements are gathered as a function of time delay, one obtains a transient that contains a record for the evolution of the reagents to product as a function of time.

The Formation of Coherent Wave Packets and their Motion

The illustration shown in Figure 2 shows atoms in a molecule moving as classical particles. Because of their small size, atoms behave quantum mechanically and this has some implications on the observed motion. A prototypical pump-probe measurement is illustrated for this purpose in Figure 3. For resonant excitation, a photon is absorbed by a molecular system, originally in the ground electronic state (V_0). This absorption causes a vertical transition to an excited electronic state (V_1). Consistent with Heisenberg's uncertainty principle, a pulse that has a short temporal duration has a large spectral bandwidth; for most molecular systems and ultrashort pulses, the frequency spread of the pulse is wider than the vibrational energy spacing in the excited state. Thus, a number of quantum vibrational states are populated on V_1 by the absorption of a short laser pulse. When a short pulse excites a number of states, a coherent superposition of states is formed. One can understand the formation of a wave packet by simply realizing that in the short time the laser pulse excites the system there is no opportunity for the atoms in the molecule to move. For this reason the initial wave packet is very similar to the ground state atomic arrangement.

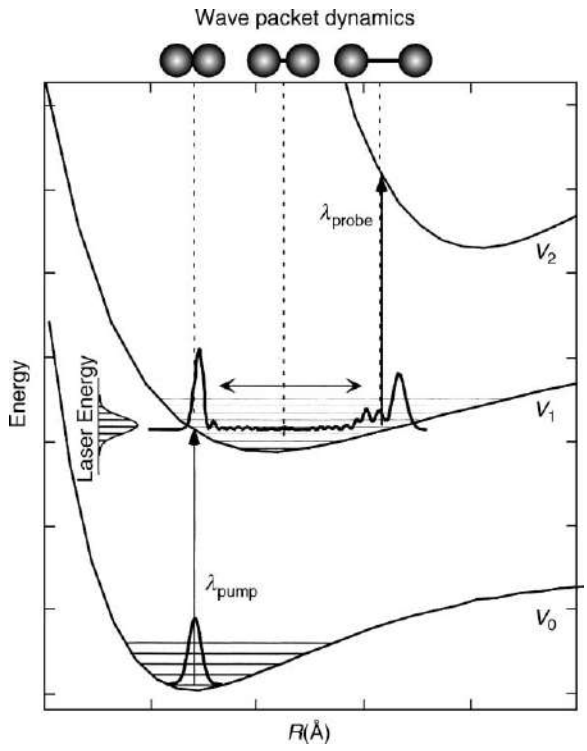

Figure 3 Pump-probe experiment. The pump pulses excite the molecular system from the ground state V_0 to excited state V_1. The large spectral bandwidth of femtosecond pulses allows for a number of vibrational levels in V_1 to be populated. A wave packet is formed that vibrates on V_1; the classical motion of the particle is shown as well at the top. In this case the energy of the probe pulse will excite the molecules to V_2 when the bond is stretched (right side of oscillation); however, when the bond is compressed, the energy is not resonant with any excitation and the probe pulse will not be absorbed.

When a short laser pulse excites a molecule to a bound electronic state, the resulting wave packet, $\Psi(t)$, is a sum of all the possible states φ_n in V_1 with amplitudes, a_n, $\Psi = \sum_n a_n \varphi_n$. The amplitudes, a_n, depend on the overlap between the initial (ground state) and final (excited state) spatial overlap. This overlap indicates that the excitation process is faster than the motion of atoms in a molecule. Based on this approximation for the overlap integral, the amplitudes are replaced by the Franck–Condon overlaps to obtain $\Psi = C\sum_n \langle \varphi_n | \varphi_0^{(0)} \rangle \varphi_n$. The wave packet on V_1 evolves in time according to $\Psi(t) = \sum_n a_n \exp[-\mathrm{i}E_n t/\hbar]\varphi_n$, where E_n is the energy that corresponds to the state φ_n and $\hbar$ is Planck's constant divided by 2π. When a pulse is very short in time, the coherent superposition of states closely resembles the initial distribution of atoms in the molecule. The motion from inner to outer turning point of the quantum mechanical wave packet is illustrated in Figure 3. One can imagine this wave packet as a classical particle oscillating on the excited potential energy curve with the periodic nature of the molecular vibrations. When the pulse is very long and only one state is populated, the dynamics can no longer be observed, as is the case with conventional frequency-resolved spectroscopy.

Molecular dynamics can be inferred from high-resolution frequency-resolved spectroscopy; however, the spectra can become congested and the data analysis becomes more complicated for large molecules, for high temperatures, or in solutions. Molecular dynamics are obtained from frequency-resolved spectra by taking the Fourier transform of the data. It seems more intuitive to consider these dynamics in the time domain where the dynamics can be observed directly as 'snapshot' taken by the femtosecond pulses. Ultrafast time-resolved spectroscopy has other advantages – high peak intensities allow for multiphoton excitation and other nonlinear processes, the selection of the detection wavelength discriminates between different species and results in an increased signal-to-noise ratio for intermediate short-lived species, collisions take place on a longer time-scale and do not affect these measurements, and the characteristics of the ultrafast laser pulse (such as phase, frequency components, and chirp) can be manipulated to control the dynamics and the yield of chemical reactions.

Methods for Measuring Femtosecond Dynamics

Pump-Probe Method

Pump-probe techniques use a pair of pulses separated in time, one to initiate the reaction and the second to determine its progress. The first femtosecond pulse with wavelength λ_{pump} excites a molecule from its (bound) ground electronic state (V_0) to an excited (V_1) state. After a variable time delay, a second femtosecond pulse with wavelength λ_{probe} causes a transition to another excited state (V_2) as shown in Figures 3 and 4.

The progress of the reaction dynamics is followed through the measurement of signal resulting from the absorption of the probe pulse by the system. There are different detection modes depending on the potential energy surfaces reached by the pump and probe pulses. In one case (Figure 4a), the first excited state (V_1) is nonfluorescent but the second one (V_2) fluoresces and this fluorescence is measured. In another case (Figure 4b), state V_1 is fluorescent but state V_2 is not; in this case the depletion of the fluorescence by the probe is measured. Finally (Figure 4c), excited states V_1 and V_2 may be dissociative; therefore, the fluorescence detected is emitted by the products. In any of these cases, as the wave packet on the intermediate state propagates (e.g., vibrates, dissociates), the transition probability varies (i.e., the ability of the probe pulse to be absorbed by the molecule changes). Therefore, the amount of laser-induced fluorescence (LIF) emitted varies with the pump-probe time delay. In some cases, the probe pulse produces ions and there are a number of methods to detect the electrons or the ions with great sensitivity.

Measurement of rotational dynamics

Measurement of rotational dynamics requires a vectorial frame of reference and this is achieved with linearly polarized laser beams. Excitation by the pump pulse selects an initial population distribution described by $\cos^2\theta$ where θ is the angle between the transition dipole and the pump laser polarization vector. As the rotational wave packet dephases due to the distribution of the population in different rotational levels, the probe pulse probes the transient alignment of the molecules in space. As the molecules rotate away from the initial alignment, the signal decreases. In Figure 5 rotational dynamics are depicted schematically. First, a single (classical) molecule is considered (top panel). As the molecule rotates, its dipole changes from being aligned with the polarization vector of the laser field

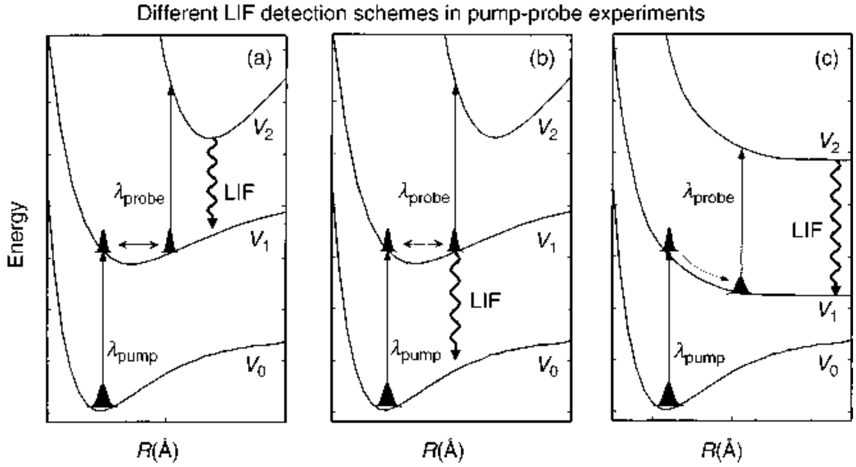
Different LIF detection schemes in pump-probe experiments
(a)
(b)
(c)
Energy
V_2
V_1
V_0
λ_{probe}
λ_{pump}
LIF
R(Å)

Figure 4 Different pump-probe experiments.(a) Bound-to-bound transitions from $V_0 \rightarrow V_1$ and from $V_1 \rightarrow V_2$. The V_2 state fluoresces and this increase in LIF from V_2 is measured as signal. (b) Bound-to-bound transition from $V_0 \rightarrow V_1$ where V_1 fluoresces. When the probe pulse causes an excitation from $V_1 \rightarrow V_2$, a loss of LIF signal is observed. The depletion of fluorescence is measured in this case. (c) Bound-to-repulsive surface transition from $V_0 \rightarrow V_1$. The energy of the probe causes an excitation to V_2 at a particular molecular distance. Fluorescence of a product fragment is measured.

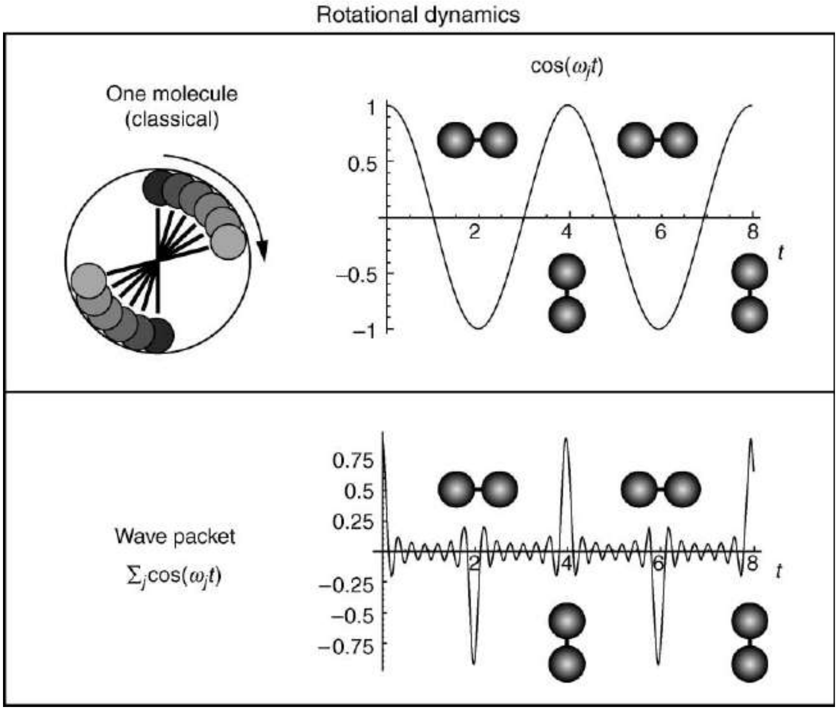

Figure 5 Rotational motion of molecules. Excitation by a linearly polarized laser pulse initially selects molecules that are well aligned ($\cos^2\theta$) with its polarization. Considering only one molecule evolving in time (top), the rotational motion in the molecule causes it to move away from the initial alignments and then eventually return to it. The signal will decrease when the molecule is not aligned with the polarization of the probe laser and will increase when aligned well. The signal looks like a cosine function that depends on the rotational frequency of the molecule (ω_j). For a rotational wave packet, the rotational dynamics are similar except that a number of rotational levels are populated; thus, the signal depends on the summation of many cosine functions and the features in the data become sharper.

(assumed vertical) to being perpendicular. The rotational motion can be described by a cosine function. The rotational frequency ω_j is proportional to the rotational level and the rotational constant of the molecule. Because the rotational constant is inversely proportional to the angular momentum, larger molecules rotate slower than smaller molecules. When several rotational levels are contained in a rotational wave packet the motion produces sharper features (bottom panel). The sharp features are called rotational recurrences or revivals. Note that the opposite trend is observed if the probe pulse is polarized perpendicular to the pump pulse; the signal increases as the molecules rotate away from the initial alignment and decreases as the alignment recurs. The time-resolved data for each polarization arrangement (parallel and perpendicular) contain both an isotropic component (vibrational motion) and an anisotropic component (rotational motion). In rotational anisotropy measurements, data from these two polarization arrangements are manipulated mathematically to yield just the rotational motion; rotational dynamics and rotational information such as the spectroscopic rotational constants and the rotational energy partitioning in the reactants and products can be extracted from these measurements.

Three-Pulse Four-Wave Mixing Method

Some methods are more complicated than pump-probe technique but they are more powerful. Three-pulse four-wave mixing (FWM) is one such method and is briefly presented here. Time-resolved four-wave mixing techniques can be understood based on the formation of transient gratings by two of the incident lasers. When two laser beams cross, the spatial modulation of their electric fields varies due to constructive and destructive interference as shown in Figure 6. The molecules in the interaction region experience varying electric field intensities according to their position and this leads to the formation of a transient grating of polarized molecules in space. The transient grating formation can be probed easily by the detection of Bragg scattering of a third laser beam, also known as homodyne detection; here, the signal beam is scattered in the phase-matching direction $\mathbf{k}_{\text{signal}}$ as indicated in Figure 6a. Alternatively, probing can be achieved by detection of changes in the intensity of a fourth beam that propagates in the

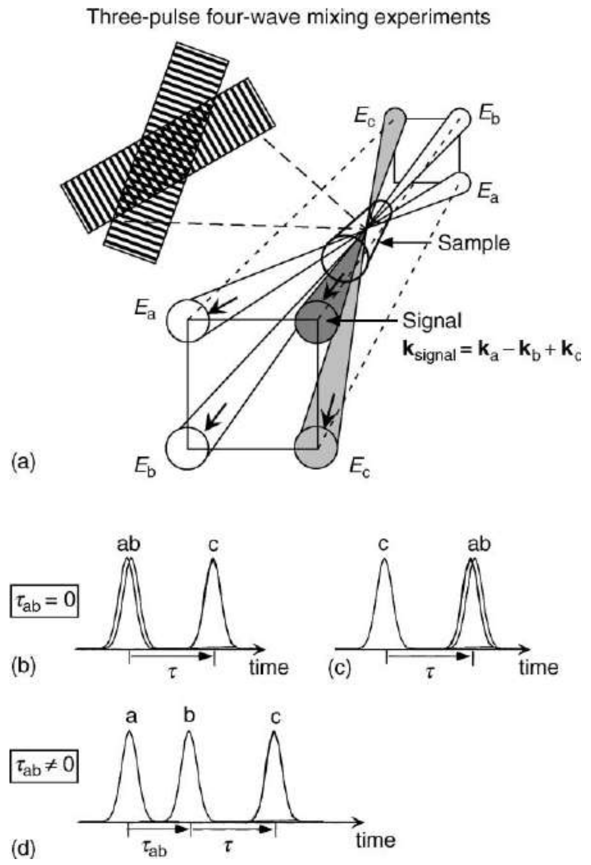

Figure 6 (a) FWM experimental configuration. Two of the beams cross which forms a grating and the third beam scatters off the grating to generate a fourth beam (the signal). The three laser pulses are noted as electric fields E_a, E_b, and E_c. These three laser pulses can be delayed or advanced in time relative to each other giving rise to a number of different pulse sequences. (b) Pulse sequence with fields E_a and E_b overlapped in time ($\tau_{ab} = 0$). Field E_c follows at a time delay τ. (c) Pulse sequence in which E_c precedes E_a and E_b by time delay τ. For both of these cases, τ is scanned in time to obtain the transients. (d) Pulse sequence in which no pulses are overlapped in time and are applied in the order E_a, E_b, and E_c. Time delay τ_{ab} can be set to specific values while τ is scanned in time.

same direction as the signal; this is called heterodyne detection. The medium that constitutes the gratings is composed of the sample molecules, which in the gas phase move freely. Therefore as time evolves, the molecular dynamics (rotation and vibration) modulate the transient grating and hence the signal.

Three-pulse four-wave mixing is a nonlinear spectroscopic method that combines the interaction of three laser pulses in a phase-matched geometry with a well-defined time sequence of the pulses. The signal from these measurements arises from a third-order polarization resulting from the interaction of the three electric fields. Three-pulse FWM is similar to the pump-probe technique in that a preparation step is followed, after some variable time delay, by a probing step. However, three-pulse FWM allows for a greater degree of control over the preparation and probing processes.

When the lasers are in resonance with an electronic state, FWM measurements can provide dynamic information from the ground and excited electronic states. When two pulses are incident on the sample at $t = 0$ and a third pulse probes the dynamics at a later time τ (as in Figure 6b), the measurements provide information that is very similar to that from a pump-probe measurement. However, having three pulses allows for other different setups. For example, when one pulse arrives at the sample first and the other two pulses follow after a time delay (as in Figure 6c), the resulting signal provides dynamic information from the excited state. This type of measurement also provides a measurement of the electronic coherence of the sample, i.e., for how long the molecules remember the phase of the incident laser pulses. This information is important if one wants to take advantage of the phase of laser pulses to cause quantum mechanical interference in the molecular sample. These measurements are known as photon echo and are much more common in condensed-phase samples. There are numerous pulse sequences possible; only three examples are shown in Figure 6.

FWM measurements do not depend on resonance excitation to an electronic excited state. In the absence of such resonance, the grating is formed in the ground state and ground state vibrational and rotational dynamics are observed. These measurements, also known as transient grating (TG), depend on the impulsive Raman scattering process. By changing the time delay (in Figure 6b) between the electric fields that generate the grating and the third (probe) beam, the signal is modulated by the molecular dynamics. Analysis of the time-dependent signal yields the rotational and vibrational spectroscopic values from the sample being studied.

Studies of Reaction Dynamics

Dissociation on a Repulsive Surface

In 1987, Zewail and co-workers studied the photodissociation of cyanogen iodide (ICN) using femtosecond lasers; the first direct observation of a chemical reaction from reactants to products was conducted. The excitation and decay of the reactants (ICN), the formation and decay of the intermediates (I–CN), and the formation of the products (I + CN) were observed in this experiment.

The pump pulse (306 nm) excited ground state cyanogen iodide to a repulsive excited state of ICN which dissociated into I and CN fragments (both in the ground state). The probe pulse when set to 388.5 nm was resonant with an excitation of the CN fragment; therefore, the reaction was monitored by detecting the laser-induced fluorescence (LIF) of this CN^* fragment using a probe pulse at different time delays (as shown in Figure 4c). The wavelength of the probe pulse could be detuned to different wavelengths for making observations at different lengths of the I–C bond. The results of these experimental studies showed that ICN dissociated in about 200 fs and that the transition state lived for only about 50 fs. The data from the wavelength detuning studies were used to map the excited state potential of ICN both in time and in space; this data analysis is called inversion to the potential.

A schematic representation of the ICN experiment is shown in Figure 7. The potential energy curves are shown in Figure 7a. The pump laser initiates the chemical reaction. Three different probe wavelengths are depicted, the first probing early in the reaction, the second at an intermediate distance, and the final one probing the free products. Notice that the wavelength of the probe determines the location in the potential energy curve that is probed. This concept allows one to study reagents as they are transformed into products. The signals obtained from the three different probe wavelengths are shown in Figure 7b.

Reactions Involving Crossing of Potential Energy Surfaces

For many chemical reactions, the progress of the reaction requires changes of the electronic state. This can take place where potential energy surfaces cross. The best-known example of reaction dynamics involves the experiment on the photodissociation of NaI by Zewail and co-workers. This experiment is depicted in Figure 8. The pump pulse excites the wave packet from ground state (ionic) to an excited state (covalent) where it begins to oscillate. The ground and excited states overlap with an avoided crossing in this molecule (ionic versus covalent nature, as shown in Figure 8a); when the wave packet approaches this avoided crossing, it splits into two partial waves. One wave stays on the excited state while the other 'escapes' onto the dissociative surface and separates into Na and I. The partial wave on the excited state again oscillates and then splits at the avoided crossing and the pattern continues. The time delay of the probe pulse shows the oscillations of the wave packet on the excited surface when detecting the activated complex $[NaI]^{*\ddagger}$ (thin line), while the data of the product fragments (free Na atoms detected by using the D-line absorption) (thick line) show a step-wise accumulation in the signal corresponding to the sequential formation of products during each oscillation near the avoided crossing region of the surfaces (see Figure 8b).

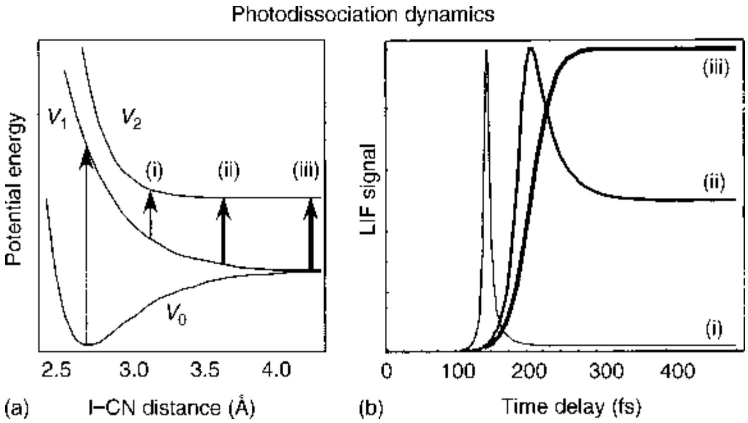

Figure 7 Photodissociation of ICN. (a) Potential energy surface schematic showing the initial excitation by the pump laser to a dissociative surface. The reaction could be monitored at different I–CN distances by using different probe wavelengths as shown by three examples here, (i), (ii) and (iii), from shorter to longer distances. (b) At short distances (long probe wavelengths), the signal is observed at shorter time delays only. The signal decreases at later time delays because the internuclear distance continues to lengthen as the dissociation occurs over time. When the distance increases as monitored by shorter probe wavelengths, the signal rises to a maximum at longer time delays reflecting the dissociation of the reaction in progress. At 'infinite' I–CN distance (i.e., the molecule is dissociated and only free I atoms and CN fragments remain), the signal reaches a maximum and remains there because the reaction is complete.

Chemical Dynamics with Multiple Pathways

The pump-probe concept can be used to probe more complex reactions. To date, a great number of different chemical reactions have been studied (see Further Reading). Among the most important reactions we list ring opening, isomerization, proton transfer, and concerted chemical reactions. We illustrate these measurements with experiments on the photodissociation of dihalogenated alkanes. Upon high-energy excitation (12 eV), the methylene iodide (CH_2I_2) molecule dissociates and forms molecular iodine in excited states D′ and f as well as producing I atoms (main channel). The detection wavelength used in the experiment can distinguish between these products; D′ fluorescence can be collected at 342 nm and f fluorescence is collected at 272 nm. The Dantus group has studied the dynamics of this type of concerted molecular elimination reaction extensively. The photodissociation reaction occurs by a multiphoton excitation (three-photons from the pump beam at 310 nm) reaching a dissociative state where the CH_2 fragment and excited molecular iodine are formed. The 620 nm probe pulse depletes the fluorescence from the excited I_2 molecule (either D′ or f state) as shown in Figure 4b; this fluorescence depletion is measured and analyzed to determine the dynamics of the reaction. The object of this experimental work was to determine if the photodissociation mechanism for forming molecular iodine followed a step-wise, synchronous concerted, or asynchronous concerted mechanism as shown in Figure 9.

The time transients reveal coherent vibrational motion of the I_2 product. Analysis of the transition state dynamics shows that the breaking of the carbon–halogen bond and the formation of the halogen–halogen bond occurs within 50 fs. Further, the iodine molecule is formed with a very hot rotational distribution indicating that the symmetry of the molecule is broken during the process; the torque imparts a large amount of rotational motion. Overall, the results of these studies show that the photodissociation process of CH_2I_2 follows an asynchronous concerted reaction mechanism and corresponds to a symmetry barrier that exists in C_{2v} molecules preventing the formation of the interhalogen bond. The existence of this symmetry barrier was explored by using CH_2ICl, a molecule that does not have C_{2v} symmetry; as expected, the halogen molecule (ICl) was formed with a cold rotational distribution indicating that it was able to follow a synchronous concerted mechanism. The dissociation times were also studied on the family of compounds CX_2Y_2 and $R{-}CHI_2$ to investigate the influence of thermodynamics and reduced mass changes on the reaction.

The Dynamics of Bond Formation

Most chemical reactions that are studied with femtosecond spectroscopy are unimolecular reactions (photodissociation); however, bimolecular reactions (photo-association) are interesting and chemically relevant in trying to understand how chemical reactions occur. In the photo-association process, a pair of atoms or molecules cooperatively absorbs the photon undergoing a free-to-bound transition and forming a bond. When the ultrashort pulse initiates the reaction in femtosecond photo-association spectroscopy, a clearly defined initial time for the reaction is given

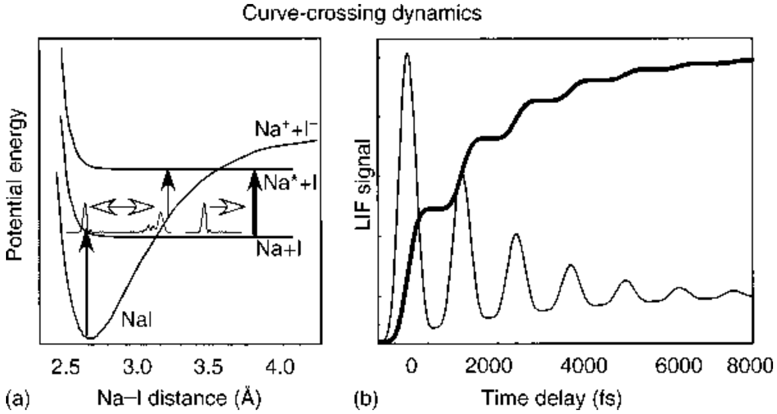

Figure 8 Photodissociation of sodium iodide. (a) Potential energy curve schematic showing the curve-crossing between the ground (ionic) and excited (covalent) states. As the wave packet oscillates on the excited curve, it splits into two wave packets at the curve crossing. (b) The thin line shows the vibrations of the excited complex as it oscillates on the excited surface. The signal intensity decreases because some of the molecules have dissociated to form Na and I atoms when the wave packet splits. The thick line shows the stepwise increase in free Na atom signal each time the wave packet oscillates and splits.

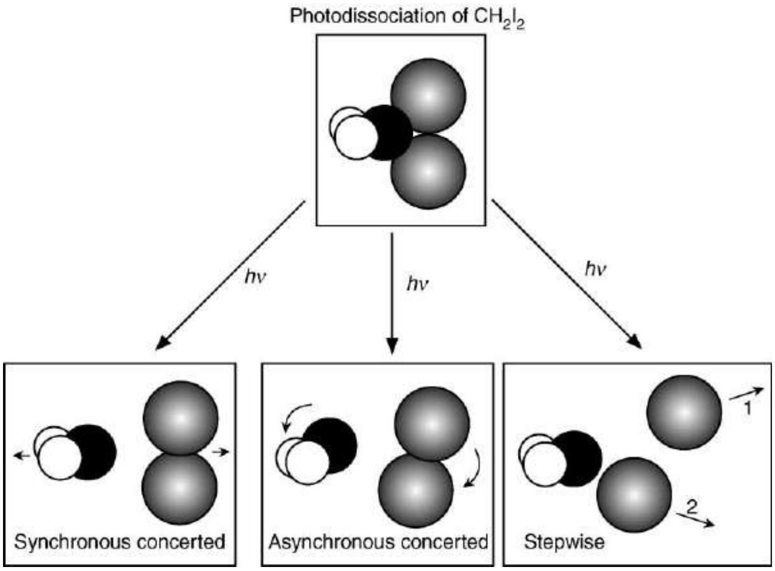

Figure 9 Possible mechanisms for the photodissociation of CH_2I_2 following multiphoton excitation. The molecule could break both C–I bonds and form an I–I bond all at the same time (synchronous concerted mechanism); this would result in very little rotational motion in the products. The molecule could begin to break one C–I bond slightly before breaking the second one and forming the I–I bond, but all three events still occur within the duration of the laser pulse (asynchronous concerted mechanism). This would result in significant rotational motion in the products. The molecule could also lose one I atom and at some later time lose the other I atom; these I atoms could collide later to form I_2 (stepwise mechanism).

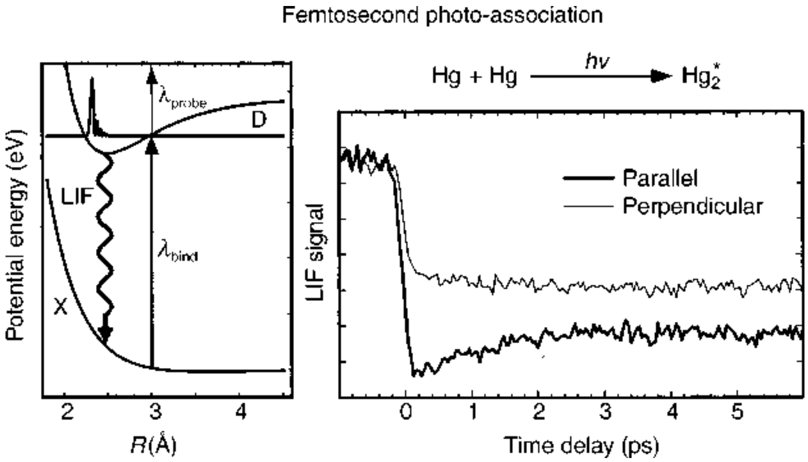

Figure 10 Photo-association of mercury dimers. The pump (bind) pulse is absorbed by two ground-state Hg atoms to form Hg_2^* on the excited state which is fluorescent. The probe pulse depletes this fluorescence causing the LIF signal to decrease at $t = 0$. Analysis of the transient reveals that the time of association is <60 fs. Parallel and perpendicular pump-probe data were examined to obtain the rotational distribution of this reaction.

and the alignment conditions of the reaction can be examined by taking advantage of the polarization of the lasers. The femtosecond photo-association spectroscopy technique was used by the Dantus group to study the formation of an excited mercury dimer. Their experiment is depicted in Figure 10. Two ground-state mercury atoms absorbed the pump (bind) pulse (at 312 nm) forming a wave packet in the excited D state. The probe pulse depleted the fluorescence from the newly formed molecules. The time of association was determined to be within the laser pulse duration (<60 fs). The dimers were found to have a narrow rotational distribution which together with the wavelength of the binding pulse reflects the impact parameter (geometry) of the collision. Calculations reveal that the 'optimal'

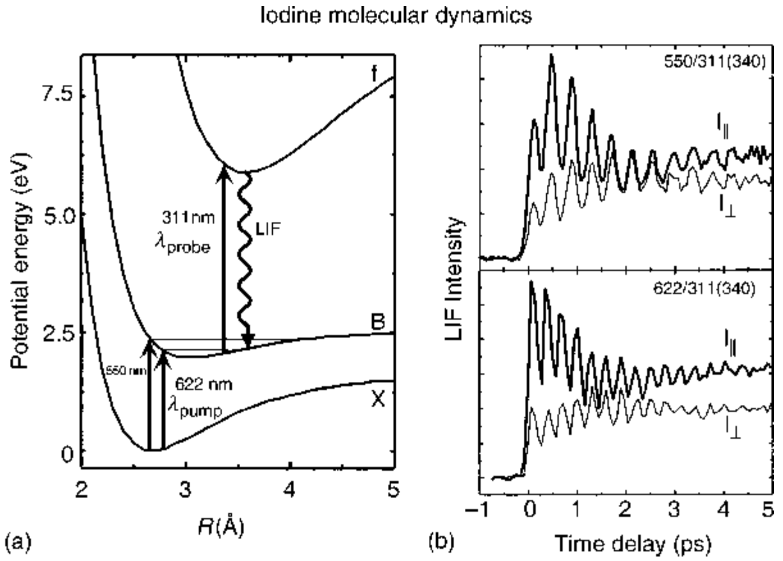

Figure 11 Molecular dynamics of iodine using two different pump laser wavelengths. (a) The pump laser, either at 550 or 622 nm, causes an excitation from the ground state to the B excited state. The choice of wavelength determines the vibrational levels which are excited on the B state. In either case, the probe at 311 nm causes an excitation to the f excited state which fluoresces around 340 nm. (b) Parallel and perpendicular pump-probe transients were obtained using each pump wavelength. At higher pump energies (550 nm), the molecule vibrates more slowly (higher vibrational level) and is reflected by the 410 fs vibrational period observed in the data. At lower pump energies (622 nm), the molecule vibrates faster and corresponds to the 300 fs vibrational period observed in the data.

binding pulse for this association reaction is 10 fs; such short pulses would overlap the entire absorption band from the repulsive ground state to the bound D state.

Bimolecular reactions can also be studied by preparing a low-temperature cluster that contains the reagents of interest. For example, a cluster containing one HI and one CO_2 molecule can be used to study the reaction $H + CO_2 \rightarrow OH + CO$. Initiation of the reaction takes place when the pump laser cleaves the H–I bond and the H atom attacks the CO_2 molecule present in the cluster. The reaction dynamics are probed as a function of time delay between pump and probe pulses as for other pump-probe experiments. The femtosecond dynamics of this reaction have been studied independently by Wittig's and Zewail's research groups.

Studies of Molecular Dynamics

I_2-Pump-Probe, Rotation and Vibration (Bound Surface)

Aside from observing chemical reactions, using ultrafast spectroscopy allows for studying the motion of isolated molecules to elucidate the internal motions of molecules themselves and to extract spectroscopic parameters. Iodine is one of the most frequently used molecules in femtosecond spectroscopy. Here we highlight only a small subset of the femtosecond studies on I_2 that have been conducted by Zewail's and other research groups. Using a pump beam at 550 nm, ground state I_2 is excited to the B state; probing with 310 nm excites the molecule to an even higher bound potential (f state) from which fluorescence at 340 nm can be obtained (as shown in Figure 11). By changing the pump-probe time delay, oscillations on the B excited state are clearly visible and show a vibrational period of approximately 410 fs. Excitation at 622 nm produces a wave packet lower in the potential well, with a vibrational period of approximately 300 fs. The differences observed for parallel and perpendicular probing correspond to the early rotational motion of the molecules. Fourier transfer analysis of the data yields the energy of each vibrational state that forms part of the vibrational wave packet. Different pump wavelengths can be used to access different vibrational levels such that vibrational frequencies and rotational constants can be obtained for a wide range of vibrational levels of the potential. This information can then be used to refine the B excited state potential energy curve of iodine.

Different Measurements Possible with Three-Pulse FWM

Ground and excited state molecular dynamics

Molecular dynamics can also be examined with four-wave mixing techniques. Once again we use iodine as a model system to illustrate that unlike the previous experimental transients of iodine obtained with pump-probe measurements, both excited and ground state dynamics of I_2 are revealed with FWM. Materny, Kiefer, and co-workers have carried out coherent anti-Stokes Raman spectroscopy and degenerate four-wave mixing measurements on this molecule using two pulses overlapped in time followed by a third pulse after a time delay; both ground and excited state dynamic are observed in the Fourier transform of data. Dantus and co-workers explored all the pulse sequences possible with three-pulse FWM using molecular iodine as a model system (see below).

Double-sided Feynman diagrams and third-order response functions can be used to describe the interaction of the three pulses with the molecular system in three-pulse four-wave mixing. These mathematical tools are described in several books on nonlinear optical spectroscopy (see Further Reading). Figure 12 illustrates how different pulse sequences reveal different molecular dynamics. The transient grating method (Figure 12a) shows both ground (160 fs oscillations) and excited state (310 fs oscillations) wave packet dynamics. When data collection occurs over long time delays, the transient grating signal persists with a similar time constant to that of the lifetime of the excited state. The photon echo pulse sequence (shown in Figure 12b) reveals only excited state dynamics (310 fs oscillations). Note that at longer time delays, the signal decays exponentially with a time constant that reflects the rate of electronic coherence dephasing. This fundamental decay parameter indicates how long molecules remember the phase of the electric field that acted on them.

Off-resonance transient grating molecular dynamics

Off-resonance transient grating can be used to study the molecular motion of molecules in the ground state. The pulse sequence used is shown in Figure 6b. With this method it is possible to examine the rotational and vibrational dynamics of diatomic and polyatomic molecules. The vibrational dynamics belong to the Raman-active modes. From the rotational recurrences, one can obtain rotational constants with great accuracy. Dantus and co-workers have studied carbon dioxide at room temperature and in flames. At high temperatures, rotational recurrences narrow because more rotational levels are populated and the rotational recurrence moves towards longer times because of centrifugal distortion. Further studies on carbon dioxide and carbon disulfide have shown that strong fields can impart an increase in initial alignment (due to torque along the polarization vector of the laser) as well as molecular deformation such as bending (due to electronic state mixing). Heterodyne detected off-resonance transient grating data for CO_2 at room temperature are shown in Figure 13. The data closely resemble the schematic illustration in Figure 5 where rotational dynamics were introduced. The data in Figure 13 have been analyzed to yield the B and D rotational constants which give the structural parameters of the molecules and the centrifugal distortions, respectively. The type of analysis given here for CO_2 can be done on molecules

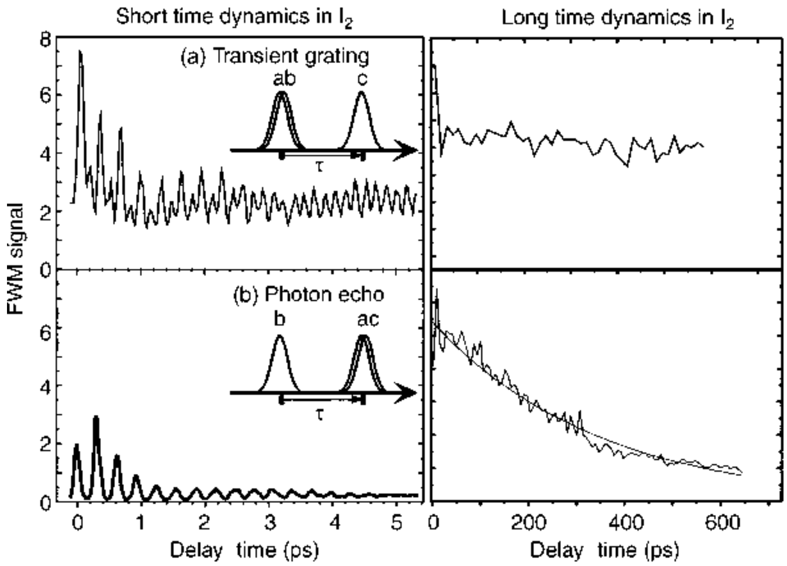

Figure 12 FWM signal in molecular iodine using two pulse sequences at short and long delay times. (a) Transient grating – E_a and E_b overlapped in time followed by E_c. Similar amounts of ground (208 cm^{-1}) and excited (105 cm^{-1}) state dynamics are observed with short delay times. In the transients, these correspond to oscillations of 160 fs and 307 fs respectively. (b) Photon echo – E_a and E_c overlapped in time but preceded by E_b. Excited state dynamics are observed in the transient with short delay times. At long time delays, the decrease in the signal reflects (a) the lifetime of the excited state or (b) the rate of the electronic coherence dephasing.

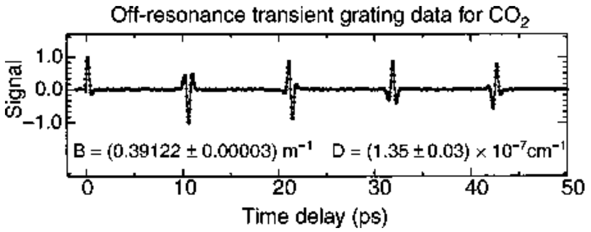

Figure 13 Heterodyne detected off-resonance transient grating on carbon dioxide at room temperature. The experimental data (dots) reflect the rotational motion of the CO_2 molecule. The data have been analyzed (the fit is shown as a line that goes through the experimental data points) to obtain the rotational constant (B) and the centrifugal distortion constant (D).

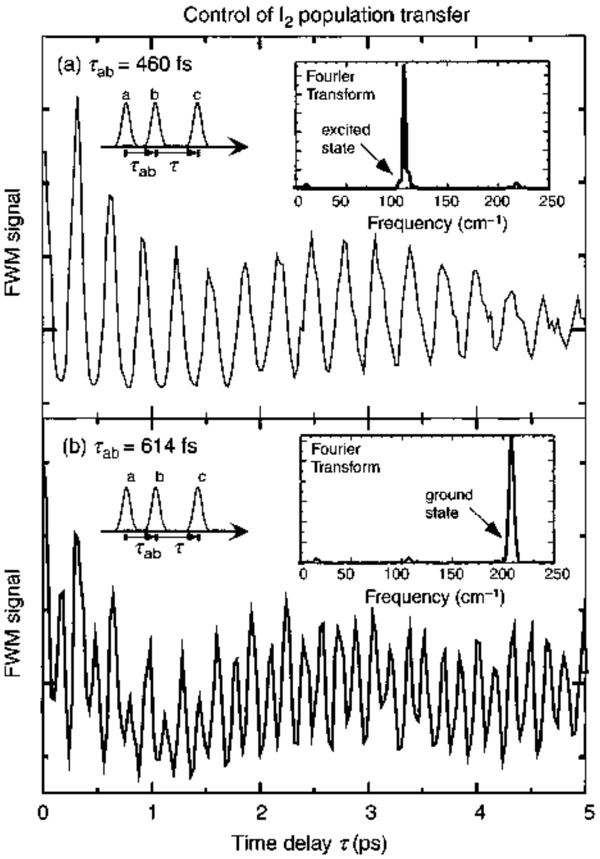

Figure 14 FWM signal in molecular iodine demonstrating control over observation of molecular dynamics. By using the pulse sequence shown, τ_{ab} was defined to be either (a) 460 fs or (b) 614 fs while τ was scanned in time. As is clearly evident in the data and the FFT, excited-state dynamics alone (307 fs oscillations) are observed with $\tau_{ab} = 460$ fs and ground-state dynamics (160 fs oscillations) are observed with $\tau_{ab} = 614$ fs.

and molecular fragments resulting from chemical reactions. This method can also be used for molecular identification.

Perspectives for the Future

Most of this article has focused on the observation of ultrafast dynamics involved in chemical reactions. Presently, it is possible to generate pulses with durations as short as 5 fs. Such short pulses are capable of observing motion involving frequencies near 3000 cm^{-1}. This implies that one can observe the dynamics of most chemical bonds. The availability of shorter pulses will allow one to observe electronic dynamics. Advances in laser technology have also produced ultrashort pulses in the infrared region as well as in the vacuum ultraviolet and X-ray region. The wide spectrum available permits experiments that follow specific chemical bonds or specific transitions in elements. This in turn gives a wide range of possibilities for probing ultrafast chemical reaction dynamics.

Having pulses that are shorter than the duration of the dynamics they probe allows one to consider influencing the evolution of a chemical reaction as it is taking place. Controlling chemical reactions with lasers is not a new theme; however, the availability of ultrashort laser pulses has given new power to these efforts. Here we present a brief sample of some of the current efforts.

Control Using Three-Pulse Four-Wave Mixing

Coherent control of chemical reactions depends on the relative phase of two different laser pulses that interact with the sample in order to optimize the transfer of population between electronic states. The three-pulse four-wave mixi\ng technique can be utilized to attain coherent control; the specific timing between the pulses can be used to achieve near-unity or near-zero values of the transfer of population for molecules that interacted with the two electric fields. The Dantus group has also used three-pulse FWM to demonstrate control over optimizing the population transfer between two electronic states in iodine. The populations of the ground and excited states are controlled based on the time delay between the first two pulses τ_{ab} in the pulse sequence shown in Figure 6d. The transient is obtained as a function of the time delay between the second and third pulses, which probes the vibrational dynamics of the population; this experiment is illustrated in Figure 14. For Figure 14a using $\tau_{ab} = 460$ fs, only excited-state dynamics are observed; while for Figure 14b using $\tau_{ab} = 614$ fs, primarily ground-state dynamics are observed. This observation is confirmed by the Fourier transform of the time-resolved transient. The top transient has a vibrational frequency of 108 cm^{-1}, corresponding to the vibrational frequency of the excited B state. The bottom transient has a vibrational frequency of 208 cm^{-1}, corresponding to the vibrational frequency of the ground state of iodine molecules.

Control Using Feedback and Evolutionary Algorithm

For complex molecules, one seldom knows the precise sequence of pulses or phase modulation of a given pulse that can optimize the yield of a chemical reaction. In these cases, a pulse-shaping device which manipulates the phase and amplitude of frequencies within an ultrashort pulse can be used. The device is part of a learning loop that includes a computer and an experimental setup. The learning algorithm running in the computer generates different pulses to be evaluated experimentally. Signal from the experiments is sent back to the computer for further refinement. The learning algorithm iteratively arrives at the optimum pulse. This method, first introduced by Rabitz, was first tested by Wilson. Gerber and co-workers have demonstrated automated control over the photodissociation reaction of organometallic compounds using an evolutionary algorithm with feedback. The organometallic compounds chosen were iron pentacarbonyl ($Fe(CO)_5$) and dicarbonylchloro(η^5-cyclopentadienyl)iron ($CpFe(CO)_2Cl$). For the first molecule, multiphoton absorption of femtosecond pulses at 800 nm caused ionization and fragmentation resulting in numerous ionic species – Fe^+, $Fe(CO)^+$, $Fe(CO)_2^+$, $Fe(CO)_3^+$, $Fe(CO)_4^+$, and $Fe(CO)_5^+$. These species were detected by mass spectrometry; the measured signal was used as feedback for the evolutionary algorithm that adjusts the phase of the femtosecond laser pulse. The $Fe^+$and $Fe(CO)_5^+$ exit channels were used to demonstrate the control available over branching ratios with this automated control method. Maximization (5.0) of the $Fe(CO)_5^+/Fe^+$ ratio was achieved with very short laser pulses while minimization (0.057) of this ratio was achieved with long (ps) pulses. These results were found to correspond well to changing the pulse duration of bandwidth-limited pulses. The optimization procedure did not attempt to control the other ionic species. This same automated optimization procedure was used to control the branching ratio of $CpFeCOCl^+/FeCl^+$ for the second organometallic compound. The maximum branching ratio was 4.9; the minimum was 1.2; and bandwidth-limited pulses yielded 2.4. Unlike the simpler carbonyl compound, simple changes to the pulse duration of a laser pulse could not optimize this branching ratio. The evolutionary algorithm with the feedback from the mass spectrometer allowed for these tailored pulses to be obtained and optimized the desired exit channel without any knowledge of the specific molecular system. The full potential of shaped femtosecond pulses has yet to be realized. It is possible that shaped femtosecond pulses will play an important role in future time-resolved experiments where they will be used to improve time resolution or to access potential energies that could not be reached with simpler pulses. Pulse shapers will also play a role in the preparation of ultrashort pulses with single femtosecond or attosecond duration.

List of Units and Nomenclature

Å	angstrom (10^{10} Å = 1 meter)
Attosecond	10^{18} as = 1 second
B	rotational constant
cm^{-1}	wavenumbers
D	centrifugal distortion constant
electron volts	1 eV = 8065 cm^{-1}
fs	femtosecond (10^{15} fs = 1 second)
FTS	femtosecond transition state spectroscopy
FWM	four-wave mixing
k	wavevector
LIF	laser-induced fluorescence
nm	nanometer (10^9 nm = 1 meter)
ps	picosecond (10^{12} ps = 1 second)
R(Å)	internuclear distance (given in angstroms)
$t = 0$	initiation of reaction (pump and probe are overlapped in time)
τ	time delay between pump and probe pulses being applied
TG	transient grating
ω_j	rotational frequency

See also

Chemical Applications of Lasers: Non-Linear Spectroscopies; Pump and Probe Studies of Femtosecond Kinetics. **Coherent Control:** Experimental; Theory. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** Four-Wave Mixing; Ultrafast and Intense-Field Nonlinear Optics; $\chi^{(3)}$ – Third-Harmonic Generation. **Spectroscopy:** Nonlinear Laser Spectroscopy; Raman Spectroscopy. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses. **Ultrafast Technology**: Femtosecond Condensed Phase Spectroscopy: Structural Dynamics.

Further Reading

Bandrauk AD, Fujimura Y and Gordon RJ (eds) (2002) *Laser Control and Manipulation of Molecules*, vol. 821 Washington, DC: ACS Press.

Chergui M (ed.) (1996) *Femtochemistry: Ultrafast Chemical and Physical Processes in Molecular Systems*. Singapore: World Scientific.

Dantus M (2001) Coherent nonlinear spectroscopy: from femtosecond dynamics to control. *Annual Reviews of Physical Chemistry* 52: 639–679.

De Schriber FC, De Feyter S and Schweitzer G (eds) (2001) *Femtochemistry: With the Nobel Lecture of A. Zewail.* Weinheim, Germany: Wiley-VCH.

Douhal A and Santamaria J (eds) (2002) *Femtochemistry and Femtobiology.* Singapore: World Scientific.

Goswami D (2003) Optical pulse shaping approaches to coherent control. *Physics Reports* 374: 385–481.

Grimberg I, Lozovoy VV, Dantus M and Mukamel S (2002) Femtosecond three pulse spectrocopies in the gas phase: density matrix representation. *Journal of Physical Chemistry* 106: 697–718.

Manz J and Woste L (eds) (1995) *Femtosecond Chemistry*, vols. I and II. Heidelberg, Germany: Verlag Chemie.

Mukamel S (1995) Principles of Nonlinear Optical Spectroscopy. New York: Oxford University Press.

Sundstrom V (ed.) (1997) *Femtochemistry and Femtobiology.* Singapore: World Scientific.

Zewail AH (ed.) (1994) *Femtochemistry – Ultrafast Dynamics of the Chemical Bond*, vols. I and II. Singapore: World Scientific.

Femtosecond Condensed Phase Spectroscopy: Structural Dynamics

E T J Nibbering, Max Born Institut, Berlin, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Ultrafast Chemistry

Chemistry means rearrangement of atoms in/between molecules. Chemical bonds are broken and/or formed, leading to a change in chemical composition, from reactants to products, during a reaction. Following the temporal behavior of the reaction all the way from reactant – via possible intermediates – towards the product state(s), amounts to deciphering the molecular processes that lead to these chemical reaction dynamics. Molecular rearrangements from reactant to product may encompass many orders of magnitude in time (see Figure 1). Typically the larger and more involved the rearrangements are, the more time it takes for a molecular species to reach the final state. Rearrangements of biomolecular structure occur in micro- to milliseconds (spontaneous folding of small proteins consisting of several hundreds of amino acid units), or longer (DNA multiplication, protein synthesis in ribosomes). In these conformational dynamics, hundreds to thousands of chemical bonds change their nature. Bimolecular reaction dynamics in liquid solution occur on time-scales of nanoseconds or longer, being controlled by the relatively slow diffusional motions of the reaction partners to each other. Many elementary processes in chemistry, however, occur on much faster time-scales. In fact, when one considers the dynamical event of a single-bond rearrangement, the relocation of the (relatively few) atoms appears to take place as an ultrafast event. Bond fission, hydrogen and proton transfer, electron transfer, and cis/trans-isomerizations have been found to occur on femto- to picosecond time-scales ('femtochemistry'). The dynamics of these elementary processes are not only determined by the energy landscapes of the reacting partners. The energy landscapes (potential energy surfaces) are determined by molecular parameters such as relative orientation (distances, angles) of the reaction partners, the energy levels of the reactant, intermediate and product species, and – very important – the energy barriers represented by transition states. When considering condensed phase reaction dynamics, the important role of the surrounding solvent has been recognized. The solvent shells modulate energy levels of the reactant, intermediate and product states, through electrostatic interactions. In the liquid phase the solvent fluctuations lead through these interactions to fluctuating energy levels of the reaction partners. In addition, the solvent may facilitate chemical reactions through energy exchange with the reaction partners, leading to efficient dissipation of excess energy, making chemical reactions often irreversible. The processes that are extremely dominated by these solvent interactions (such as molecular collisions, electronic and vibrational dephasing, vibrational relaxation) are also often found to occur on femto- to picosecond time-scales. The outcome of a bond fission may be a null effect, when the dissociating fragments are forced by the surrounding solvent to recombine and relax ('cage effect').

Spectroscopy has been a prime tool to obtain insight into these key molecular processes. Reactant, intermediate, and products contribute to the spectra with their respective molecular resonances. With spectral domain spectroscopy one may, in principle, obtain a precise determination of structural information. Fluctuations of energy levels of these species by the solvent interactions, and finite lifetimes, contribute to a broadening of these molecular resonances. Using steady-state spectroscopy one

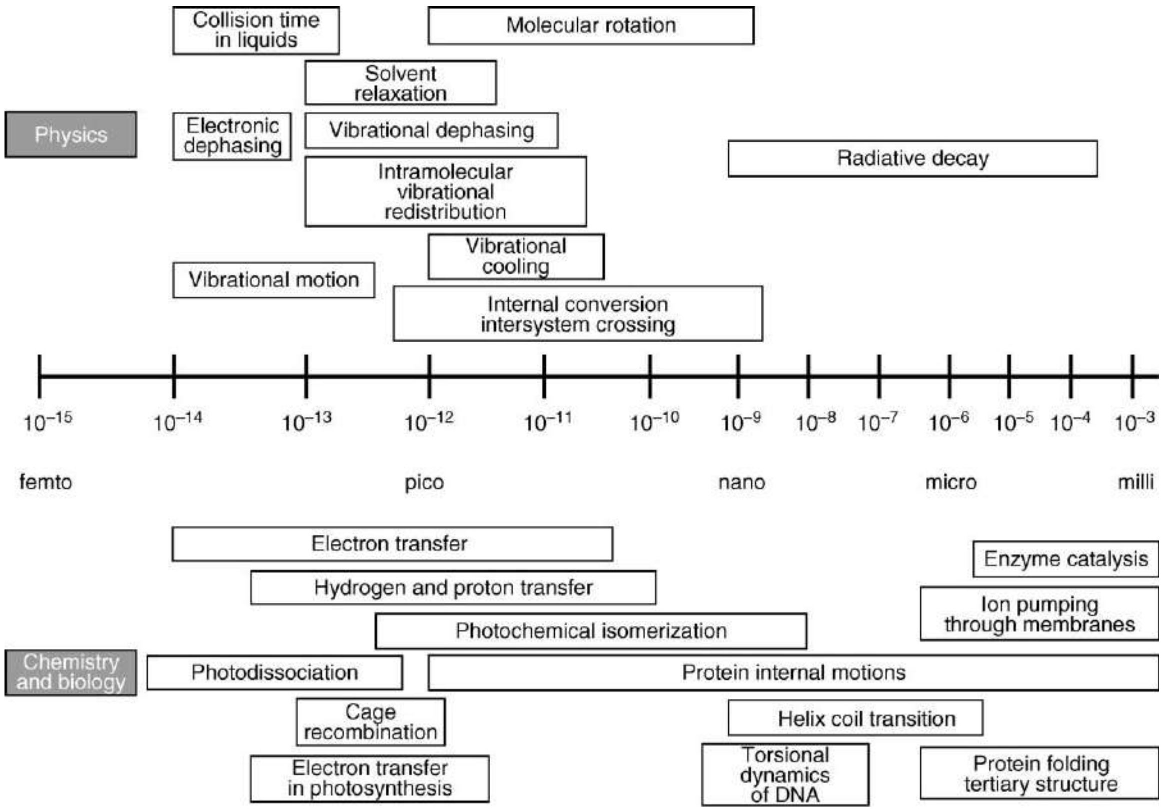

Figure 1 Chart showing the ultrafast time-scales of fundamental physical and chemical processes (inspired by charts of Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Clarendon Press and Zewail AH (2000) Femtochemistry: atomic-scale dynamics of the chemical bond. Nobel lecture as appeared in *Journal of Physical Chemistry A* 104(24): 5660–5694).

could try to determine these effects through a lineshape analysis. Using this approach, the results have to be analyzed with an a priori defined model representing the dynamics. The disadvantage lies in the averaging over all time-scales inherent in spectral domain spectroscopy, where details of transient states may be lost in this averaging procedure. In contrast, in experiments designed to grasp the molecular dynamics in real time, the temporal resolution is the prerequisite quantity in observation of these processes. In time-resolved experiments one can trigger a chemical reaction at a well-defined point in time, after which one can follow the conformational changes from reactant – via possible intermediates – to the product, and identify any possible transition states along the reaction pathways.

During the last century the technological advances in time-resolved spectroscopy have prompted significant breakthroughs in the study of chemical reaction dynamics. Initially the temporal resolution was given by the duration of flashes from light bulbs ('flash' spectroscopy) or by rapid mixing of reaction partners (stopped-flow technique), as developed by Nobel Laureates Eigen, Norrish, and Porter, and many others. These techniques allow a temporal resolution of at best milliseconds. Already, up to the early 1960s, it was understood that elementary reaction dynamics occur on much faster time-scales. Fourier spectroscopy may improve time resolution down to microseconds or slightly less. With the advent of pulsed laser sources with ever-increasing temporal resolution, a wide arena of time-resolved spectroscopic techniques has emerged. Nanosecond laser systems are known since the early 1960s, picosecond pulses can be generated since the 1970s, the first femtosecond laser system was reported in 1981, and currently laser engineering groups are breaking into the attosecond domain. The majority of pulsed laser systems operate at visible and near-infrared wavelengths. This means that usually chemical events in ultrafast time-resolved spectroscopy are initiated by electronic excitation ('photochemistry'). Subsequent molecular rearrangements have then to be followed by probing the electronic states through electronic resonances (UV/VIS pump–UV/VIS probe or UV/VIS four-wave mixing

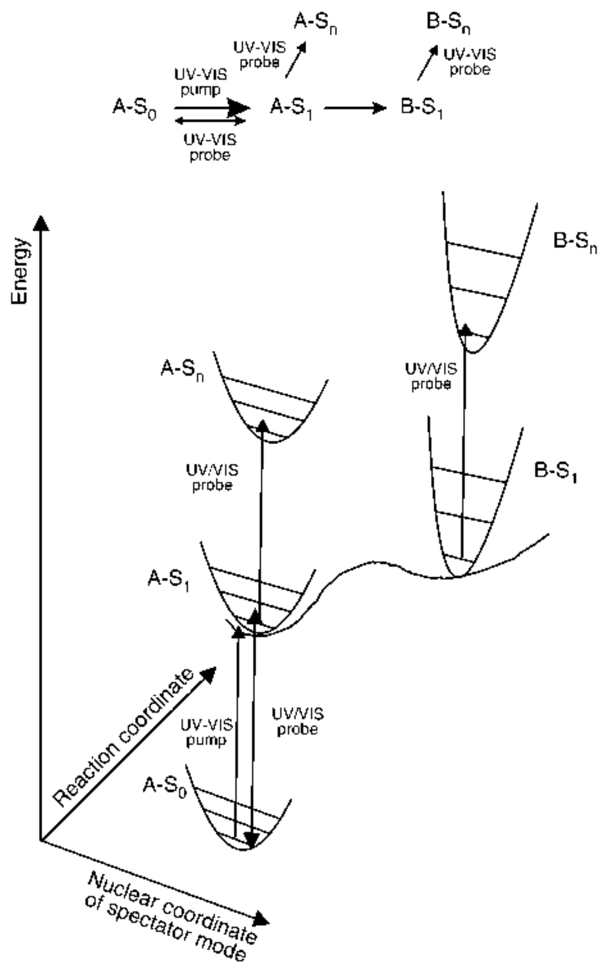
A-S$_n$
B-S$_n$
UV-VIS probe
UV-VIS probe
UV-VIS pump
A-S$_0$
A-S$_1$
B-S$_1$
UV-VIS probe
Energy
B-S$_n$
A-S$_n$
UV/VIS probe
UV/VIS probe
B-S$_1$
A-S$_1$
Reaction coordinate
UV-VIS pump
UV/VIS probe
A-S$_0$
Nuclear coordinate of spectator mode

Figure 2 Ultrafast time-resolved electronic spectroscopy. A UV/VIS pump pulse excited an electronic resonance, and a probe pulse determines, depending on the tuning in the UV/VIS population changes in the reactant ground or excited states or in the product state.

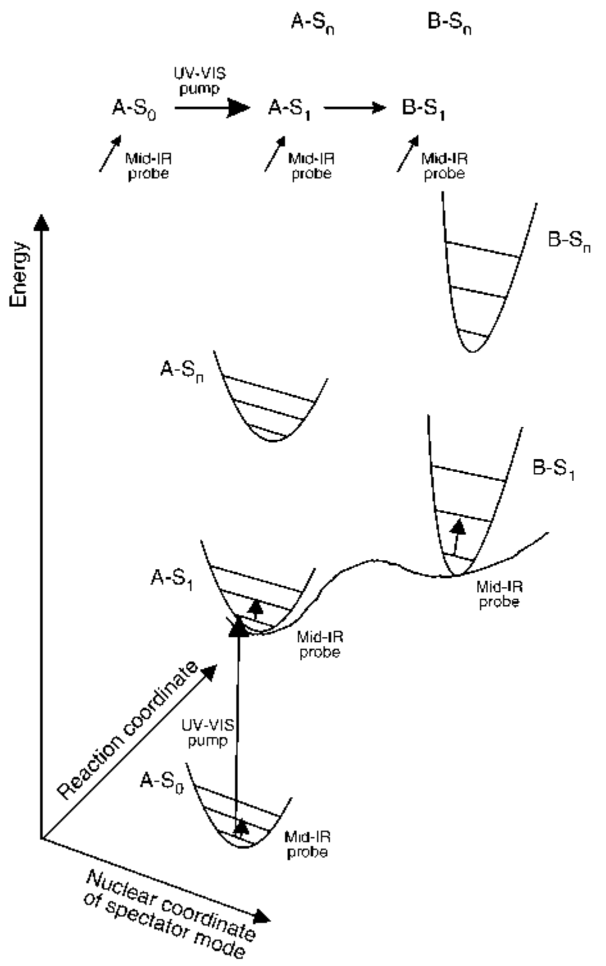
A-S$_n$
B-S$_n$
UV-VIS pump
A-S$_0$
A-S$_1$
B-S$_1$
Mid-IR probe
Mid-IR probe
Mid-IR probe
Energy
B-S$_n$
A-S$_n$
B-S$_1$
A-S$_1$
Mid-IR probe
Mid-IR probe
Reaction coordinate
UV-VIS pump
A-S$_0$
Mid-IR probe
Nuclear coordinate of spectator mode

Figure 3 Ultrafast time-resolved infrared spectroscopy. A UV/VIS pump pulse excites an electronic resonance, after which the populations are followed with an infrared probe pulse by inspection of marker modes in the different electronic states.

spectroscopy) (Figure 2), or via vibrational resonances (UV/VIS pump–IR/Raman probe spectroscopy) (see Figures 3 and 4).

Laser Technology

Since the beginning of the 1990s, a major advancement in femtosecond laser technology has emerged by the discovery of Kerr lens mode locking in laser oscillators with Ti:sapphire as lasing material. Ultrastable and ultrashort laser pulses tunable between 700 and 1,000 nm can be generated with durations as short as 6 fs. The second advancement is due to the ability of Ti:sapphire as laser amplifying material, using the method of chirped pulse amplification. Pulses with output powers up to several W, with repetition rates in the kHz-regime, are now routinely generated. These developments facilitate the efficient generation of ultrashort pulses tunable from the UV to the mid-infrared, and even beyond to the THz-regime using schemes based on nonlinear optics. Frequency conversion of the fundamental output of an amplified Ti:sapphire laser system is then achieved with nonlinear processes such as self-phase modulation and supercontinuum generation, n^{th}-order harmonic generation ($n = 2, 3, 4, \ldots$), sum and difference frequency generation, or parametric generation and amplification. As a result, one can now almost arbitrarily tune the excitation and probe wavelengths of the applied laser pulses. Additional parameters, that one may alter, are pulse duration (from less than 10 fs up to several ps) and pulse energy (in the mJ range or less; typically one uses for

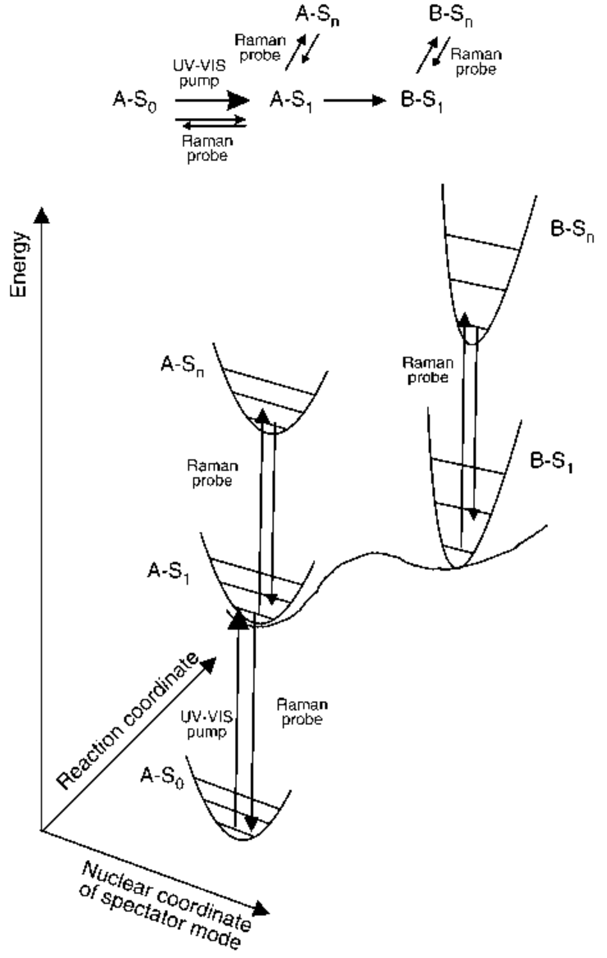

Figure 4 Ultrafast time-resolved Raman spectroscopy. A UV/VIS pump pulse excites an electronic resonance, after which the populations are followed with (resonance) Raman probe pulse by inspection of marker modes in the different electronic states.

condensed phase spectroscopy pulse energies in the range of nJ to μJ). More advanced pulse manipulation schemes use amplitude and phase masks. The latter aspect holds the promise of amplitude and phase control of molecular excitation, feeding the ultrafast chemist's dream of laser control of molecular reaction dynamics.

Ultrafast Electronic Spectroscopic Techniques

Pulsed laser sources enable a variety of experimental approaches to determine transient states. Already, since the development of nanosecond laser sources, time-resolved pump–probe and four-wave mixing spectroscopy have been widely used. In a first approach, often one uses these forms of spectroscopy to probe in real-time populations of reactant, transient, and product states as function of pulse delay. With the advent of pico- and femtosecond laser sources, coherence properties of the material response have been investigated as well.

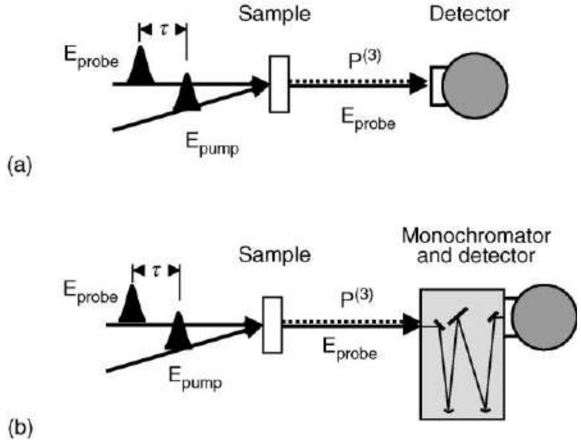

Figure 5 Pump–probe technique: A pump pulse induces an absorbance change in a sample, that is subsequently measured by a time delayed probe pulse, either spectrally integrated (a) or spectrally resolved (b).

The following techniques have been widely used in ultrafast electronic condensed phase spectroscopy.

UV/VIS Pump–UV/VIS Probe Spectroscopy

In this technique a 'pump' pulse resonant to a transition promotes the molecules to an electronic excited state, and the time-dependent populations are determined by measurement of the transmission of a time-delayed 'probe' pulse (Figure 5). The absorbance change $\Delta A(\tau) = -{}^{10}\log[T(\tau)/T(\tau = 0)]$ (with T for transmission change and τ the pulse delay) is a direct indication of populations of states. Here one can tune the probe pulse to the same transition as the pump pulse, and one measures an absorbance decrease due to a bleach of ground state population (i.e., fraction of molecules being excited), and due to molecules in the excited state stimulated back to the ground state. One could also tune the probe pulse to transitions between the transient state and higher lying states. For electronic spectroscopy, often transitions between the ground and first excited state ($S_0 \rightarrow S_1$) overlap with excited state absorptions ($S_1 \rightarrow S_n$). One has to rely then on numerical analysis procedures such as singular value decomposition and decay associated spectra, that correlate spectral features with temporal behavior.

The time resolution of the experiment is given by the cross-correlation between pump and probe pulses. In the case of extremely short pulses one has to consider the effects of group velocity dispersion between different frequency components due to different travel times through the samples, leading to pulse temporal broadening. For the case of pump and probe tuned to different wavelengths, the time-resolution is usually dominated by group velocity mismatch between pump and probe pulses.

When ultrashort laser pulses are used, often the applied spectral bandwidth enables simultaneous excitation of several vibrational sublevels in vibronic transitions. Due to the phase-relationship of the applied laser pulses, these vibrational sublevels are prepared in states that are well-defined in temporal phase with respect to each other. Such a coherent superposition of vibrational eigenstates is nothing less than a vibrational wavepacket, that will evolve in time. The occurrence of vibrational wavepacket motions leads to oscillatory modulations of the pump–probe signals.

Time-Resolved Fluorescence Spectroscopy

Here again a 'pump' pulse excites an electronic resonance. The transient excited states are now probed through their spontaneous fluorescence emission, either by use of time-correlated single photon counting or by fluorescence upconversion with a gating pulse in a nonlinear medium (Figure 6). Single-photon counting has a time resolution of at best a few picoseconds. Fluorescence upconversion can be performed with a time resolution of 100–200 fs. The advantage lies in the fact that the measured signals are more likely due to the specific excited states only, and not corrupted by other contributions.

Electronic Four-Wave Mixing Spectroscopy

In time-resolved degenerate electronic four-wave mixing spectroscopy, where all the three applied laser pulses (with wavevectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$) are tuned to the same electronic resonance, one can detect a nonlinear signal in the phase-matched direction $\mathbf{k}_S = \mathbf{k}_3 + \mathbf{k}_2 - \mathbf{k}_1$ (Figure 7). The first laser pulse generates a coherent superposition between the electronic ground and excited states. This electronic coherence evolves in time during the period (denoted as coherence time τ), until the second pulse converts this coherence in a population frequency grating in the electronic ground and excited states. After a second time period of free evolution (denoted as population or waiting time T), this frequency grating is again transformed into an electronic coherence, which after some time evolution generates a macroscopic polarization in the phase-matched directions. This macroscopic polarization is either directly measured by a time-integrating detector (in which case one measures the intensity signal proportional to the absolute square of the nonlinear polarization), or one uses another light pulse that mixes as local oscillator with the nonlinear polarization (and the signal is proportional to the amplitude of the polarization). The time resolution is determined by the third-order autocorrelation of the applied pulses.

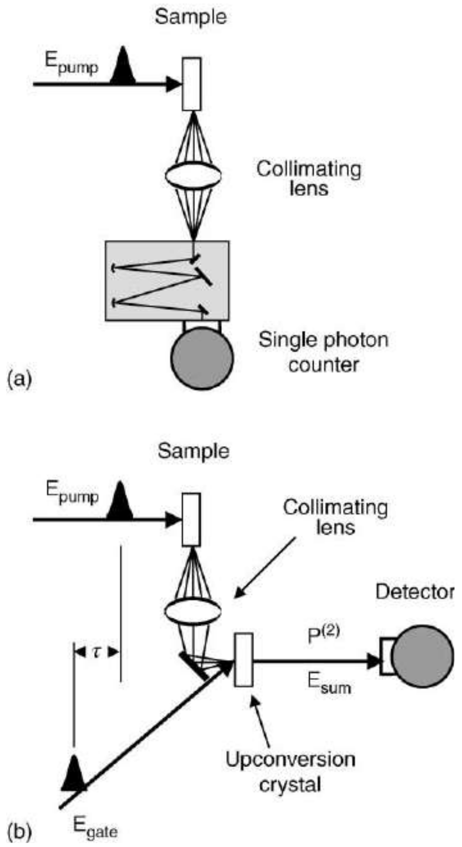

Figure 6 Time-resolved emission detection: The fluorescence emission is either temporally resolved by time-correlated single photon counting (a) or by up-conversion with a gating pulse (b).

Depending on the delay between the three input pulses, and which delay one varies, different dynamics can be measured. When one performs an experiment with τ scanned and T set at zero, one measures the electronic coherence decay (due to electronic dephasing) time in a two-pulse photon echo experiment. When, on the other hand, τ is set to zero, and T is scanned, one measures the evolution of a population grating, and electronic lifetimes can be deduced. In a three-pulse photon echo experiment both pulse delays are scanned, and the signals give insight into both electronic coherence and frequency grating decay (due to electronic dephasing and spectral diffusion, respectively). A novel extension of four-wave mixing technique is the three-pulse echo peak shift (3PEPS). In a 3PEPS measurement, the delay time $\tau_{\max}$ is

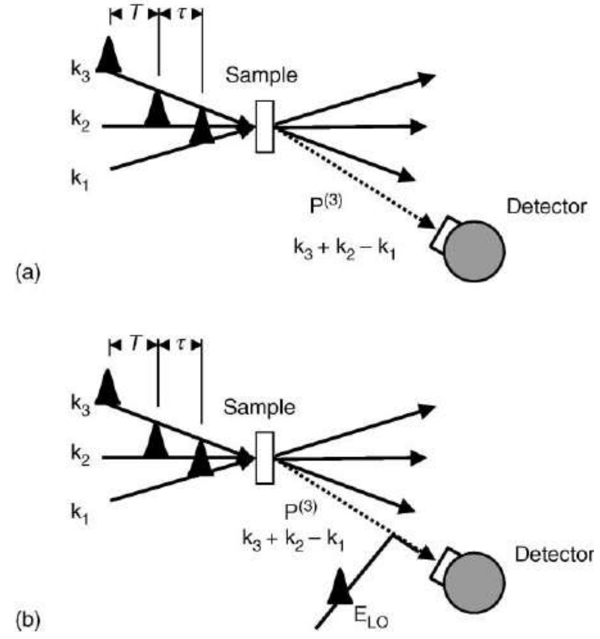

Figure 7 Beam geometry for time-resolved four-wave mixing (grating/photon echo) spectroscopy: The nonlinear signal is transmitted by the sample in a phase-matched direction and its intensity signal (a) or its amplitude by interference with a local oscillator field (b) is detected by a time-integrating detector.

determined, where the echo signal has a maximum along the coherence delay τ, as a function of the population delay T. This method provides direct insight into the frequency fluctuation correlation function that governs the linear and nonlinear signals. Extension of the method to multi-color four-wave mixing spectroscopy has been demonstrated.

Electronic Condensed Phase Spectroscopy: Femtochemistry and Solvation Dynamics

In the narrow sense of meaning, 'femtochemistry' deals with the real-time observation of nuclear motions during chemical reactions. Quantum mechanics teaches us that this is only possible when molecular systems are prepared in vibrational superpositions. Only in the case of quantum states describing vibrational wavepackets evolving in time the narrow-defined positions of nuclei allow for a real-time description of molecular rearrangements, or in other words: dynamics. This has been demonstrated in the gas-phase landmark experiments by Nobel Laureate Zewail (Figure 8).

In the condensed phase at room temperature, however, often such well-defined quantum states are hard to prepare, since the ensembles of molecular systems exist in mixed states. Coherent laser excitation on an ensemble will then prepare mixed states as well, and then information that one obtains from the experiments is less explicit. In addition, room temperature liquid solutions induce extremely fast fluctuations in the states of the molecules under study and as a consequence the electronic and vibrational coherences show ultrafast decaying behavior. This limits the temporal window with which one can follow the evolution of vibrational wavepackets.

Most of the ultrafast spectroscopic experimental work performed on photochemistry in the condensed phase, being described as 'femtochemistry', only deals with the time-scales of changes in electronic states (usually indicated with reaction rates typified with exponentially decaying or rising functions). In this regard it is better to speak of 'kinetics' rather than 'dynamics', although both terms are used in condensed phase spectroscopy. Due to the relative broad electronic spectra of condensed phase molecules masking any structural detail, it is hard to derive nuclear motions in real time, and as such it is difficult to make any statements about dynamics of molecular structures during chemical reactions.

Electronic resonances have also been used to probe the dynamical interactions between a nonreactive solute and the surrounding fluctuating solvent. The idea behind these experiments is to achieve information about the time-scales of solvent fluctuations on (and solvent motions due to a change in) the electronic charge distribution in a molecule (as induced by an electronic excitation). Usually the solvent motions are characterized with a frequency fluctuation correlation function (Figure 9). This information is highly relevant for the case of reacting molecules, where electronic motions accompany the relocations of nuclei, even in electron transfer reactions, where the nucleic rearrangements are modest. Solvent motions have been shown to often control reaction rates. Solvent rearrangement, as a response to a change in electronic charge distribution, is known as solvation dynamics. Experimental probes for solvent motional fluctuations are electronic coherence decay (electronic dephasing) and population frequency grating decay (spectral diffusion) that can be measured in two- and three-pulse photon echoes. Solvent adaptation to new charge distributions can be followed with the time-dependent Stokes-shift in fluorescence emission, as well as in three-pulse photon echoes (in so-called three-pulse echo peak shift measurements described above). Ultimately these experiments resolve the temporal characteristics of the frequency fluctuation correlation function.

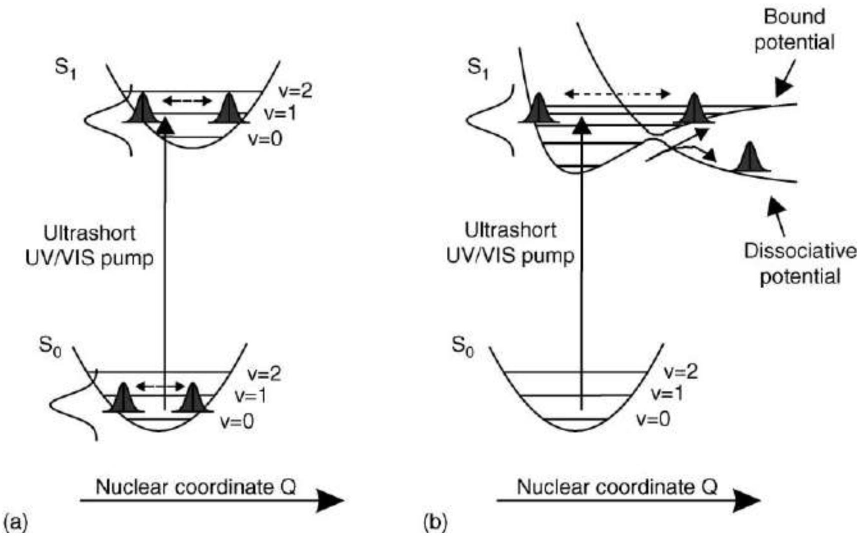

Figure 8 Ultrafast excitation with broadband pulses creates coherent superpositions of vibrational eigenstates, i.e., vibrational wavepackets. These wavepackets are generated in electronic excited and ground states (a), and their time-dependence can be determined in pump–probe and four-wave mixing experiments. When level crossing occurs to another electronic state, the vibrational wavepacket may be observed to evolve along the reaction coordinate (b).

Ultrafast Vibrational Spectroscopy in Photochemistry: Structural Dynamics

Vibrational spectroscopy has, despite a smaller signal sensitivity due to smaller transition cross-sections, several advantages over electronic spectroscopy. Vibrational transitions have typically smaller bandwidths than electronic transitions, due to longer dephasing times (the exception to the rule: O–H/O–D stretching bands of hydrogen bonded hydroxyl groups). Specific signal contributions are thus easier tractable in vibrational spectroscopy than in electronic spectroscopy. Vibrational bands can often be correlated to specific vibrational motions by inspection of the transition frequencies, (e.g., O–H, N–H, and C–H stretching bands can be found in the 3000 cm^{-1} range, C=O and C=N stretching modes are located near 1,600–1,750 cm^{-1}, O–H, N–H, and C–H bending vibrations between 1,500–1,650 cm^{-1}, C–O stretching mode around 1,200–1,300 cm^{-1} etc.), and, in particular, the frequency range between 1,000 and 1,800 cm^{-1} is called the fingerprint region because of this reason. Measurement of vibrational bands thus leads to identification of particular vibrational motions, and conclusions can be drawn on specific structural motifs in the molecules. Vibrational bands can be infrared (IR) or Raman active, sometimes both, and both IR and Raman techniques have been widely applied in structural determination studies ranging from small molecules to larger biomolecular systems. In ultrafast photochemistry one thus excites the molecules with a UV/VIS pump pulse, and an IR/Raman probe pulse follows the outcome of the chemical reaction by inspection of vibrational bands of reactant, transients, and products.

Vibrational spectroscopy has the potential of revealing site-specific information if the marker modes are due to nuclear motions of specific molecular side-groups. For instance, hydrogen bonding induces marked shifts of O–H, N–H, C=O, and C=N bands. Observation of changes in spectral shifts reveals important information on hydrogen bond interactions (weakening/strengthening or even hydrogen bond cleavage). Infrared spectroscopy is able to probe small molecular species in solution that typically have their electronic resonances in the far-UV (making electronic spectroscopy impossible, since normally the solvent would absorb this radiation) (Figure 10). For instance, in acid–base neutralization reactions, where the acidity of a so-called photoacid is switched on by use of a UV pulse, a VIS probe pulse would only be able to probe the photoacid (or its conjugate photobase), thus revealing only when a proton leaves the photoacid. An infrared pulse can, besides probing vibrational resonances of the photoacid, also probe bands of a small base (or its conjugated acid) indicating when a proton arrives at the base.

In the case where the vibrational normal modes do not allow such a structural insight into site-specific groups, one can make a comparison of the

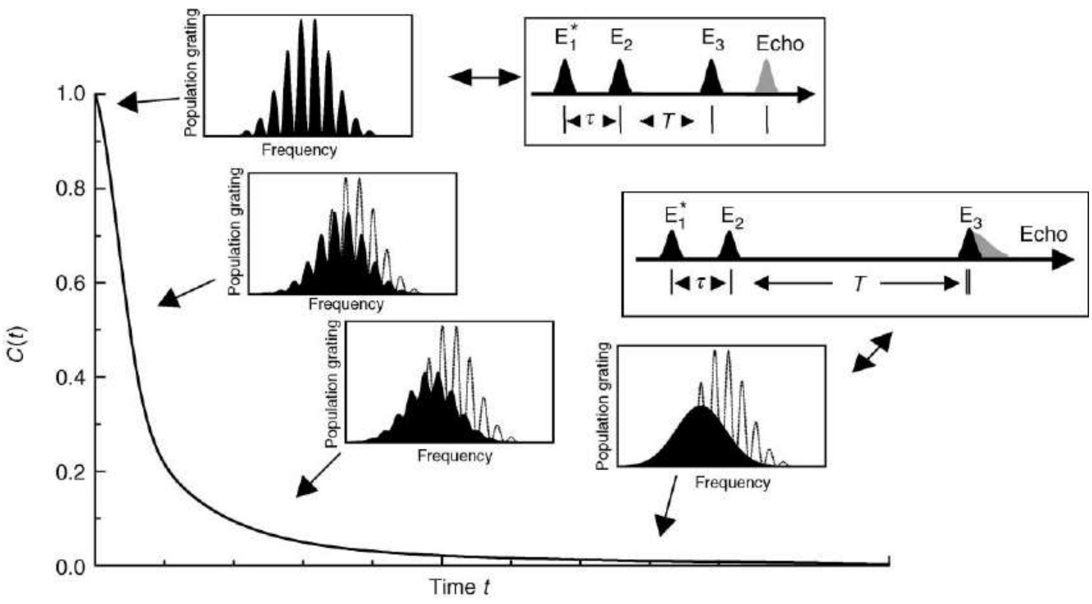

Figure 9 Solute–solvent interactions as characterized by the transition frequency fluctuation correlation function $C(t)$. This quantity can be measured in photon echo and time-resolved fluorescence experiments. In a three-pulse photon echo experiment phase information as given by the excited state frequency grating created with the first two interactions, is washed out due to solvent fluctuations and Stokes shifts to lower frequencies due to solvent rearrangement. A similar plot can be drawn for the ground state frequency grating.

experimentally observed vibrational bands with predictions made by quantum chemical calculations. When a full correspondence between experiment and theory is possible, one can make statements about the three-dimensional structure. With current quantum chemical calculational routines, such as density functional theory, medium-sized molecules are routinely calculated for electronic ground-state conformations. Reliable results of transient and product states in electronic excited states can be obtained with the routine *ab initio* complete active space self-consistent field (CASSCF), albeit for mid-size molecules of at most on the order of about 20 atoms. New developments in numerical procedures, such as time-dependent density functional theory (TD-DFT) may prove fruitful in the calculation of larger molecular systems in electronic excited states.

During recent years, femtosecond IR spectroscopy has been used in photo-induced chemical reactions ranging from excited state hydrogen and proton transfer, transformations of photochromic switches, excited state charge transfer, and cis–trans isomerizations. Here an ultrashort UV/VIS 'pump' pulse promotes the molecule to an electronic excited state, and the reaction is followed by a measurement of the absorbance change of a 'probe' pulse tuned in the mid-IR region where the vibrational marker modes sensitive to structural changes have their resonances. Experimentally, one performs spectrally resolved transient IR spectroscopy. Femtosecond IR parametric devices deliver pulses with bandwidths of $150\,\mathrm{cm}^{-1}$ or more. In order to be able to observe shifts as small as the linewidths of IR-active vibrations, one usually measures the IR absorbance change with a detector after spectral dispersion with a monochromator. As a side-effect of this spectral dispersion ground-state bleach signals often appear to grow in at negative pulse delay with the dephasing time of the transition. This effect, known as 'perturbed free induction decay' is a common feature of spectrally resolved nonlinear pump–probe spectroscopy of bleached transitions with dephasing times much longer than the time resolution of the experiment. The time resolution of the experiment is given by the cross-correlation between the UV/VIS-pump and IR-probe pulse (about 100–200 fs), and is typically dominated by group velocity mismatch in samples with thicknesses of about 100 μm.

In principle, the same approach can be followed by probing Raman-active vibrations. In this case the spectral resolution is not only determined by the monochromator through which the spontaneous Raman emission dispersed, but also by the bandwidth of the gating pulse by which the Raman effect is induced. As a result, UV/VIS pump–Raman probe

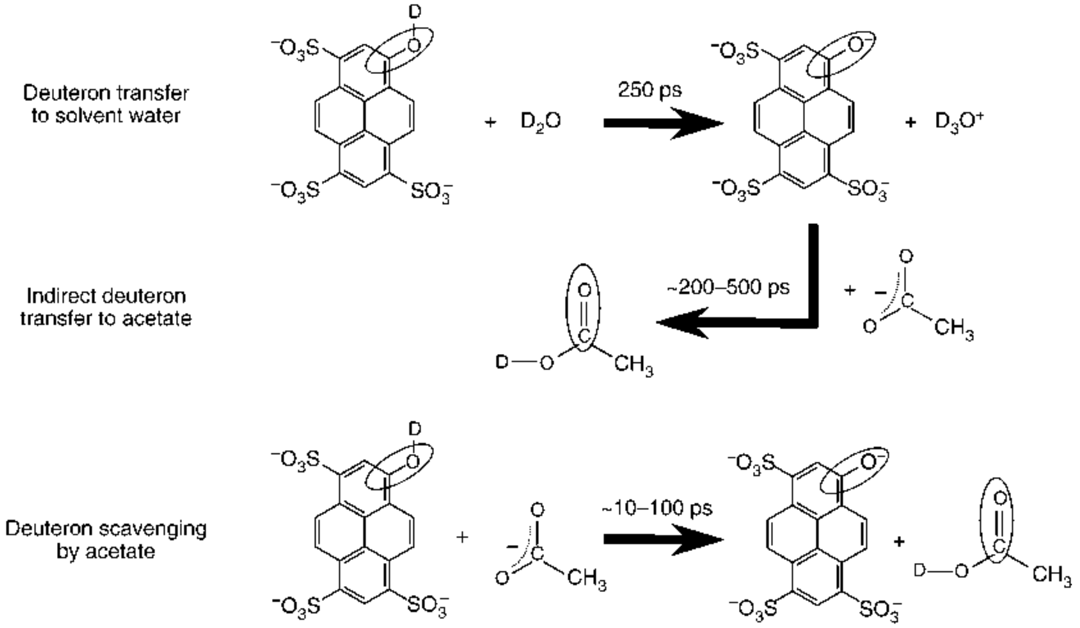
Deuteron transfer to solvent water
+ D2O
250 ps
+ D3O+
Indirect deuteron transfer to acetate
~200–500 ps
Deuteron scavenging by acetate
~10–100 ps
D
O
-O3S
SO3-
CH3

Figure 10 In excited state acid–base neutralization experiments, the proton dissociation can be followed on two sides with ultrafast infrared spectroscopy. Vibrational marker modes of photoacid and conjugated photobase indicate when the proton leaves the acid. The marker mode of acid formed by proton pick-up by the accepting base is a direct measure of the arrival time of the proton at the base. Depending on relative concentrations, either the indirect proton transfer via the solvent or direct scavenging of the proton by the base dominates the dynamics.

spectroscopy has a temporal resolution of around 1 ps (a compromise between spectral and time resolution). Due to the even weaker cross-sections of Raman transitions, one often uses resonance enhancement by tuning the gating pulse close to or resonant with an electronic transition of the state that is probed. The advantage is that one usually only observes Raman bands of the state under inspection. The drawback is that fluorescence resulting from the resonant electronic excitation often inhibits detection of Raman bands in extended spectral ranges.

Raman spectroscopy offers an additional insight into how chemical reactions evolve. By comparison of the intensities of anti-Stokes and Stokes lines of a particular vibration, it is possible to derive time-dependent changes in the excitation level of this vibration. One can draw conclusions on whether particular modes initially drive a chemical reaction ('state transition promoting modes'), or only get excited after the transition is made, by taking up the excess energy released by the reaction ('accepting modes'). In such a way, insight is obtained on the energy flow inside a molecular system (intramolecular vibrational redistribution, abbreviated as IVR) and vibrational energy dissipation to the surrounding solvent (vibrational cooling). Infrared spectroscopy is less powerful in revealing this, since red-shifted transient absorption of vibrational bands may either mean that the particular mode is highly excited (the red-shift is a consequence of the diagonal anharmonicity of the vibration), or it may mean that other modes are highly excited and these cause a red-shift of the marker mode under inspection (the red-shift is then due to off-diagonal anharmonic coupling with other modes).

Equilibrium Structural Dynamics in the Electronic Ground State

Existence of anharmonic coupling between vibrational modes means that the vibrational motions are not decoupled from each other. These couplings always exist, since otherwise no IVR and vibrational cooling (no relaxation) would occur. When the couplings are significantly large, excitation of one specific mode will induce a significant instantaneous shift of other vibrations anharmonically coupled to this particular mode. Estimation of the magnitude of these couplings should lead to a determination of the curvatures of potential energy surfaces along the respective coordinates. With current state-of-the-art femtosecond infrared technology (with time resolution of about 150 fs or less) it is possible to excite

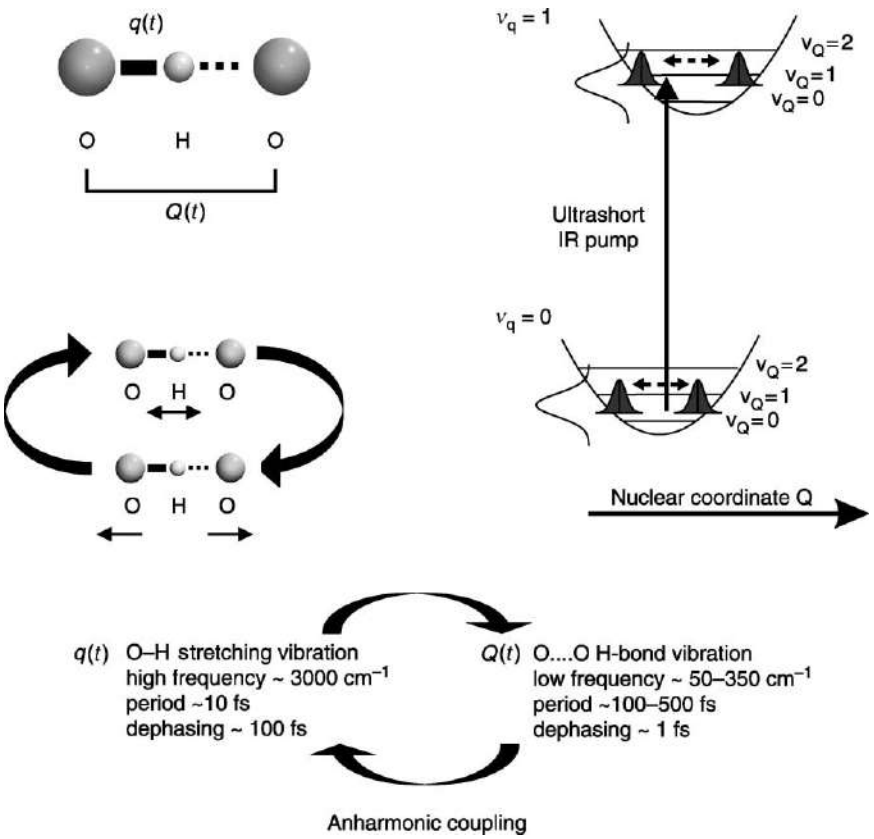

Figure 11 In hydrogen-bonded hydroxyl groups a strong anharmonic coupling exists between the hydroxyl stretching and modes modulating the hydrogen-bond distance. A Born–Oppenheimer-like separation of time-scales of vibrational motion between the high-frequency O–H/O–D stretching and the low-frequency hydrogen bond modes enables a description with potential energy surfaces for the low-frequency mode as a function of the quantum state of the hydroxyl stretching vibration. The formation of low-frequency vibrational wavepackets in ultrafast IR spectroscopy of O–H/O–D stretching modes is then fully analogous to wavepacket formation in ultrafast electronic spectroscopy.

infrared-active vibrations and to probe the same or other vibrations (either in the IR or by the Raman process), enabling insight into the anharmonic couplings between these vibrational modes.

Direct anharmonic couplings are probed in IR-pump–IR-probe and IR photon echo spectroscopy (Figure 11). In the case of peptides, the excitonic (Davydov) couplings between amide I vibrations of the amino acid units have been the subject of extensive study since these lead to spatial information of the relative orientations of the different amide I modes, and thus, of the spatial orientation of the amino acid units inside the peptide. Exploring the effects of the delay between the IR pulses on the observed signals, give insight into the fluctuations of these orientational features, or in other words, into structural dynamics of peptides. In the case of hydrogen bonded O–H/O–D stretching vibrations, the anharmonic couplings with underdamped low-frequency modes modulating the hydrogen bond distance, lead to marked modulations of observed pump–probe and echo signals as function of pulse delay. These experiments reveal which modes couple strongly to the hydrogen stretching oscillator.

Fluctuations in couplings between vibrational modes (in particular with those of the solvent) induce dephasing of the vibrational coherences. The dephasing and spectral diffusion dynamics of vibrational transitions, e.g., the O–H stretch vibration of HOD dissolved in D_2O, have been determined by IR photon echo spectroscopy. From these studies the time-scales of structural memory decay of hydrogen bond networks can be derived. Couplings between vibrational modes ultimately lead to population relaxation through intramolecular vibrational redistribution effects and vibrational cooling. Ultrafast two-color IR-pump–IR-probe reveals whether excitation of one IR-active vibration is followed by population transfer into another IR-active mode.

With IR-pump–Raman-probe, one can estimate whether Raman-active vibrations are transiently excited. In the case of hydrogen-bonded O–H/O–D stretching bands, an extremely short population relaxation time T_1 (below 1 ps) is observed, as well as an ultrarapid dephasing time T_2 (200 fs or less). These values are definitely longer for other vibrations, from which it has been concluded that the strong anharmonicities in hydrogen-bonded vibrational systems enable dephasing and relaxation processes to have such a rapid impact.

Future Developments

Future activities in the determination of dynamically evolving molecular structures in the condensed phase, may involve the aforementioned application of amplitude and phase controlled excitation pulses in optimal control of chemical reactions. Vibrational spectroscopy may be used as a tool in determination of evolving structures after electronic excitation by a shaped pulse tuned in the UV/VIS. Infrared pulses could also be shaped in amplitude and phase, and may be used in the exploration of steering of chemical reactions in the electronic ground state. Multidimensional vibrational spectroscopy is an object of extensive research. Extension of the method to determination of structure of transient states will lead to new results on molecular rearrangements, e.g., polypeptides reaching a new equilibrium after inducing a UV-excitation induced geometric change of a photochromic switch incorporated in the peptide structure. Much activity is currently also being put in the development of ultrafast structure resolving techniques such as electron diffraction, X-ray diffraction, and X-ray spectroscopy. While still extremely demanding at a technological level, final implementation of these latter techniques will open up new areas of ultrafast chemical dynamics.

See also

Chemical Applications of Lasers: Pump and Probe Studies of Femtosecond Kinetics.

Further Reading

Cerullo G and De Silvestri S (2003) Ultrafast optical parametric amplifiers. *Review of Scientific Instruments* 74(1): 1–18.

de Boeij WP, Pshenichnikov MS and Wiersma DA (1998) Ultrafast solvation dynamics explored by femtosecond photon echo spectroscopies. *Annual Review of Physical Chemistry* 49: 99–123.

Elsaesser T and Bakker HJ (eds) (2002) *Ultrafast Hydrogen Bonding Dynamics and Proton Transfer in the Condensed Phase*. Dordrecht, Germany: Kluwer.

Elsaesser T and Kaiser W (1991) Vibrational and vibronic relaxation of large polyatomic molecules. *Annual Review of Physical Chemistry* 42: 83–107.

Fayer MD (ed.) (2001) *Ultrafast Infrared and Raman Spectroscopy*. New York: Marcel Dekker.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Clarendon Press.

Fleming GR and Cho M (1996) Chromophore-solvent dynamics. *Annual Review of Physical Chemistry* 47: 109–134.

Fleming GR and Hänggi P (eds) (1993) *Activated Barrier Crossing: Applications in Physics, Chemistry and Biology*. Singapore: World Scientific.

Kaiser W (ed.) (1988) *Ultrashort Laser Pulses and Applications*. Berlin: Springer.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. Oxford, UK: Oxford University Press.

Nibbering ETJ and Elsaesser T (2004) Ultrafast vibrational dynamics of hydrogen bonds in the condensed phase. *Chemical Reviews* 104: 1887–1914.

Owrutsky JC, Raftery D and Hochstrasser RM (1994) Vibrational relaxation dynamics in solutions. *Annual Review of Physical Chemistry* 45: 519–555.

Sundström V (ed.) (1996) *Nobel Symposium: Femtochemistry and Femtobiology: Ultrafast Reaction Dynamics at Atomic-Scale Resolution*. Singapore: World Scientific.

Wiersma DA (ed.) (1994) *Femtosecond Reaction Dynamics*. Amsterdam: North-Holland.

Zewail AH (2000) Femtochemistry: atomic-scale dynamics of the chemical bond. Nobel lecture as appeared in *Journal of Physical Chemistry A* 104(24): 5660–5694.

Web-information: http://www.nobel.se
Nobel web-site showing information on Nobel Laureates M Eigen, RGW Norrish and G Porter (The Nobel Prize in Chemistry 1967: ‘for their studies of extremely fast chemical reactions, effected by disturbing the equilibrium by means of very short pulses of energy’) and AH Zewail (The Nobel Prize in Chemistry 1999: ‘for his studies of the transition states of chemical reactions using femtosecond spectroscopy’).

Ultrafast Illumination and Processing

Y Fainman, University of California, San Diego, CA, USA

D M Marom, Bell Laboratories, Holmdel, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Mode-locked lasers emitting ultrashort pulses have expanded the range of available instruments and techniques to the scientific community. The unique properties of these ultrashort pulses – their high peak powers and short duration – enable such diverse applications as multiphoton microscopy, pump-probe experiments, and coherent control of molecular and atomic wavepackets, to name a few. Ultrashort pulses can also be used as a basis for ultrahigh data rate optical communications using time division multiplexing (TDM) in a fiber communication application. These ultrafast applications often require complementary real-time optical processors based on instantaneous nonlinear phenomena, since conventional electronic means are limited by the response time of electronic devices.

During the last few years, we have developed real-time optical processors utilizing nonlinear three- and four-wave mixing of signals originating from spatial or temporal channels for synthesizing, processing, and detecting ultrafast waveforms. We perform real-time optical signal processing that can be applied to data conversion between slow parallel channels in space and an ultrafast serial signal. A time-to-space mapping processor has been built for detection of ultrafast waveforms. It mixes two spatially inverted temporal frequency information waves in a three-wave mixing arrangement, generating a quasi-monochromatic spatial signal carrying the temporal image for detection by slower electronic means. A space-to-time mapping processor was constructed for generation of arbitrary ultrafast waveforms from an input ultrashort pulse. The waveform is a time-scaled replica of the employed input spatial image, converted by a four-wave mixing arrangement. By introducing only temporal information channels to our four-wave mixing processors, real-time optical processing of temporal waveforms was enabled. We have demonstrated time reversal experiments based on performing spectral phase conjugation and spectral inversion operations, achieving time reversal of the electrical field and of the complex amplitude waveform, respectively. The three- and four-wave mixing arrangements are performed within a spectral processing device – a free-space optical setup consisting of diffraction gratings and lenses – utilizing a $\chi^{(2)}$ nonlinear crystal. A cascaded second-order nonlinearity technique is used for four-wave mixing within the $\chi^{(2)}$ crystal (a frequency-sum generation process followed by a frequency-difference generation process satisfying the type-II noncollinear phase matching condition). However, a $\chi^{(3)}$ nonlinear crystal may be readily substituted without loss of generality (albeit with a likely smaller conversion efficiency).

This article reviews the spatial/temporal processing techniques for synthesizing, processing, and detecting ultrafast waveforms. The generation of spectrally decomposed waves within a spectral processing device is described in the following section. The subsequent section articulates on the employed nonlinear wave mixing processes, and the ensuing sections express the synthesis, processing, and detection of ultrafast waveforms.

Generation and Characteristics of Spectrally Decomposed Waves

Our wave mixing experiments are performed with spectrally decomposed waves, in which the spectral frequency content is spatially dispersed. When the pulse is decomposed with high resolution, the signal at every spatial location can be treated as quasi-monochromatic. This allows for increased interaction length within a nonlinear crystal, allowing us to use long crystals, whereas typically the interaction length is limited by temporal walkoff. In this section we analyze the creation of spectrally decomposed waves.

We introduce an ultrashort optical waveform, with a temporal envelope of $s(t/\tau)$, into the spectral processing device, where τ parameterized the waveform's duration. The waveform is propagating in free space towards a diffraction grating, at an incidence angle θ (see Figure 1). The input ultrashort waveform is characterized in its coordinate system (x_1, z_1) as:

$$E_{\text{pulse}}(x_1, z_1; t) = w\left(\frac{x_1}{L_1}\right) s\left(\frac{t - t_0 - z_1/c}{\tau}\right) \times \exp\left(\mathrm{i}\left(\omega_0 t - \frac{\omega_0}{c} z_1\right)\right) \qquad [1]$$

where $w(\cdot)$ defines the transversal field distribution with a scale size parameter L_1, t_0 is a time delay parameter, c is the speed of light in vacuum, and ω_0 is the center optical frequency. The waveform is

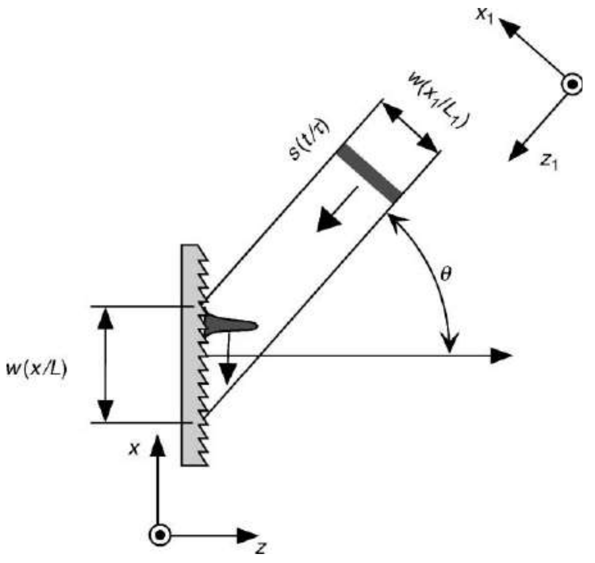

Figure 1 An ultrashort optical waveform, $s(t/\tau)$, propagating in free-space is incident upon a diffraction grating at angle θ. The transverse spatial mode of the pulse is stretched due to the projection onto the grating.

propagating in the z_1 direction at a group velocity and phase velocity of the speed of light in vacuum. The spatial confinement in the y_1 direction is omitted for brevity. We perform a time-domain Fourier transform (TFT) on eqn [1] to perform a frequency decomposition of the ultrafast waveform, yielding:

$$\begin{aligned}\tilde{E}_{\mathrm{pulse}}(x_1, z_1; \omega) = {} & \tau w\left(\frac{x_1}{L_1}\right) \\ & \times \exp\left(-\mathrm{i}\frac{\omega}{c} z_1\right)\tilde{s}(\tau(\omega-\omega_0)) \\ & \times \exp(-\mathrm{i}(\omega-\omega_0)t_0)\end{aligned} \qquad [2]$$

where the tilde over the script denotes a FT relationship, defined for the TFT as $\tilde{s}(\omega)A\int s(t)\exp(-\mathrm{i}\omega t)\mathrm{d}t$. Next we perform a rotation of the coordinate system from (x_1, z_1) to (x, z), for compatibility with the coordinate system of the spectral processing device, yielding:

$$\begin{aligned}\tilde{E}_{\mathrm{pulse}}(x, z; \omega) = {} & \tau w\left(\frac{x\cos\theta - z\sin\theta}{L_1}\right) \\ & \times \exp\left(-\mathrm{i}\frac{\omega}{c}(x\sin\theta + z\cos\theta)\right) \\ & \times \tilde{s}(\tau(\omega-\omega_0))\exp(-\mathrm{i}(\omega-\omega_0)t_0)\end{aligned} \qquad [3]$$

To find the incident field on the input grating of the spectral processing device, the field of eqn [3] is evaluated at $z = 0$. The effect of the grating diffraction can be modeled by adding the grating momentum, k_{g}, in the x direction to the k-vector that characterizes the propagation direction of the wave, resulting in:

$$\begin{aligned}\tilde{E}_{\mathrm{input}}(x, 0; \omega) = {} & \tau w\left(\frac{x\cos\theta}{L_1}\right) \\ & \times \exp\left(-\mathrm{i}\left(\frac{\omega}{c}\sin\theta - k_{\mathrm{g}}\right)x\right) \\ & \times \tilde{s}(\tau(\omega-\omega_0))\exp(-\mathrm{i}(\omega-\omega_0)t_0)\end{aligned} \qquad [4]$$

The grating momentum, k_{g}, and the incidence angle, θ, are chosen such that the center frequency ω_0 will diffract in the direction of the optical axis of the system. Setting $\omega_0/c\cdot\sin\theta = k_{\mathrm{g}}$ and substituting the grating's k-vector, $k_{\mathrm{g}} = 2\pi/\Lambda$, where Λ is the grating period, yields $\sin\theta = \lambda_0/\Lambda \equiv \alpha$, where λ_0 is the optical wavelength corresponding to the center carrier frequency. To further simplify the notation, we define a new spatial width scaling parameter $L = L_1/\cos(\theta)$, accounting for the aperture size increase due to the projection onto the diffraction grating. With these substitutions, the expression takes on the more familiar form, given by

$$\begin{aligned}\tilde{E}_{\mathrm{input}}(x; \omega) = {} & \tau w\left(\frac{x}{L}\right)\exp\left(-\mathrm{i}\frac{(\omega-\omega_0)\alpha x}{c}\right) \\ & \times \tilde{s}(\tau(\omega-\omega_0))\exp(-\mathrm{i}(\omega-\omega_0)t_0)\end{aligned} \qquad [5]$$

We may perform an inverse TFT on eqn [5] to characterize the input signal in the time domain, yielding:

$$E_{\mathrm{input}}(x; t) = w\left(\frac{x}{L}\right)s\left(\frac{t - t_0 - \alpha x/c}{\tau}\right)\exp(\mathrm{i}\omega_0 t) \qquad [6]$$

Equation [6] describes the input waveform scanning across the fixed aperture at velocity of c/α in the x-direction.

The spectrally decomposed wave is generated by performing a spatial Fourier transform (SFT) employing a lens of focal length f on the input wave (either frequency domain representation, eqn [4], or time domain representation, eqn [6]). Using the frequency domain representation, the spectrally decomposed wave U_{SDW} of the input waveform is given by

$$\begin{aligned}U_{\mathrm{SDW}}(x'; \omega) = {} & \tau\tilde{s}(\tau(\omega-\omega_0)) \\ & \times \exp(-\mathrm{i}(\omega-\omega_0)t_0)\int_{-\infty}^{\infty} w\left(\frac{x}{L}\right) \\ & \times \exp\left(-\,\mathrm{i}(\omega-\omega_0)\frac{\alpha x}{c}\right) \\ & \times \exp\left(-\mathrm{i}2\pi\frac{xx'}{\lambda_0 f}\right)\mathrm{d}x\end{aligned} \qquad [7]$$

where we use the central wavelength in the Fourier transform kernel. Evaluating the integration of eqn [7], yields:

$$U_{\mathrm{SDW}}(x';\omega) = L\tau\tilde{s}(\tau(\omega-\omega_0))\exp(-\mathrm{i}(\omega-\omega_0)t_0) \times \tilde{w}\left(\frac{\alpha L}{2\pi c}\left[\frac{\omega_0 x'}{\alpha f}+(\omega-\omega_0)\right]\right) \quad [8]$$

where the SFT of the input pupil function, $\tilde{w}(f_x)A\int w(x)\exp(-\mathrm{i}2\pi x f_x)\mathrm{d}x$, has mixed space-domain and time-domain frequency variables in its argument, characterizing the dispersion. A spectral component shifted away from the center frequency, $\Delta\omega = \omega - \omega_0$, will be centered at a displaced location $\Delta x' = -\Delta\omega\cdot\alpha f/\omega_0$, defining the spatial dispersion term $\partial x'/\partial\omega = -\alpha f/\omega_0$. One can see that in order to get fine localization of each spectral component, the spatial extent of the function $\tilde{w}(\cdot)$ must be minimized. This can be readily achieved by increasing the size of the input aperture, or the parameter L, as well as increasing α (grating's angular dispersion). Let us make the following observation on system resolution: The ultrafast pulse's spectral width is proportional to $1/\tau$, and its spatial extent is therefore proportional to $\alpha f/\omega_0\tau$ (multiplying by the spatial dispersion term). The spatial width of each spectral component is proportional to $cf/\omega_0\tau$. The dimensionless ratio of these two spatial widths, which is indicative of the number of resolvable spectral components, is $N \sim \alpha L/c\tau$, where N is defined as the system spectral resolution. When the spectral processing device is designed to satisfy the requirement of $N \gg 1$, then the spectrally decomposed wave can be considered quasi-monochromatic at every spatial position. In our experiments, we use $\lambda_0 = 800$ nm, gratings of 600 lines/mm frequency ($\Lambda = 1666$ nm), $L \approx 1$ cm, and $\tau \approx 100$ femtoseconds, yielding $\alpha \cong 0.5$ and $N \sim 160$, easily satisfying the high spectral resolution requirement. Let us make a second observation on system resolution using the time domain representation of eqn [6]: The spatial extent of the input aperture is proportional to L, whereas the instantaneous spatial extent of the scanning waveform is proportional to $\tau c/\alpha$ (see Figure 2). The dimensionless ratio of these two spatial widths is again $N \sim \alpha L/c\tau$, exemplifying the concept of resolvable positions and demonstrating the conservation of the space-bandwidth product.

To further facilitate the ensuing analysis, let us assume that we operate in the high-resolution regime (satisfied with the values given above), and we seek to simplify the expressions defining the spectrally decomposed wave. The spectrally decomposed wave can be defined in the time domain by performing a SFT on eqn [6], yielding:

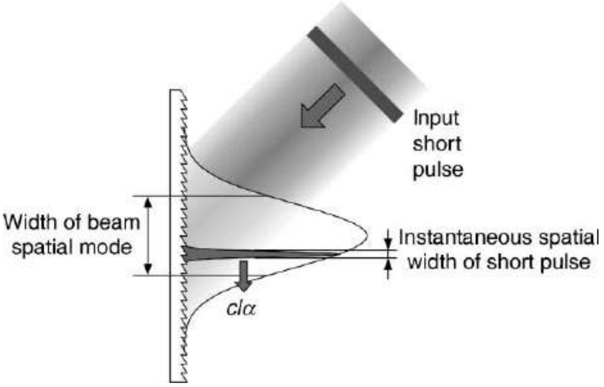

Figure 2 The input short pulse scans across the diffraction grating at velocity of *c/α*. The ratio of the width of the spatial mode of the beam to the instantaneous spatial width of the scanning waveform determines the system resolution. © 2001 Optical Society of America.

$$\begin{aligned} U_{\mathrm{SDW}}(x';t) &= \int w\left(\frac{x}{L}\right)s\left(\frac{t-t_0-\alpha x/c}{\tau}\right)\exp(\mathrm{i}\omega_0 t) \times \exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f}xx'\right)\mathrm{d}x \\ &\approx w\left(\frac{c(t-t_0)}{\alpha L}\right)\exp\left(-\mathrm{i}\frac{\omega_0(t-t_0)x'}{\alpha f}\right) \times \tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \end{aligned} \quad [9]$$

where by account of the high-resolution approximation, the instantaneous location of the traveling signal waveform samples the value of the aperture function and removes the aperture dependence from the Fourier integral. This approximation decouples the aperture function from the temporal waveform, and illustrates that the duration of the spectrally decomposed wave is defined by the aperture function (time of flight through aperture of size L), and that the spatial dependence is a mapping of the incident waveform spectrum. A rotating wavevector in time is also present, on account of the traveling (moving) waveform on the input plane. Again, we may perform a TFT on eqn [9], defining the high-resolution limit of the spectrally decomposed wave, yielding:

$$U_{\mathrm{SDW}}(x';\omega) = \tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(-\mathrm{i}(\omega-\omega_0)t_0) \times \tilde{w}\left(\frac{\alpha L}{2\pi c}\left[\frac{\omega_0 x'}{\alpha f}+(\omega-\omega_0)\right]\right) \quad [10]$$

One can compare eqns [8] and [10], as both characterize the spectrally decomposed wave in the

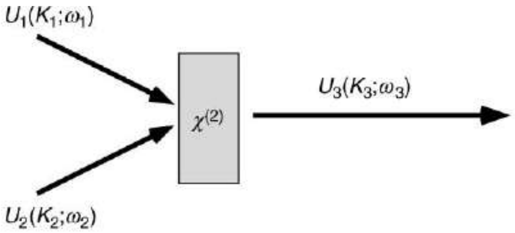

(a) Three-wave mixing in a $\chi^{(2)}$ crystal

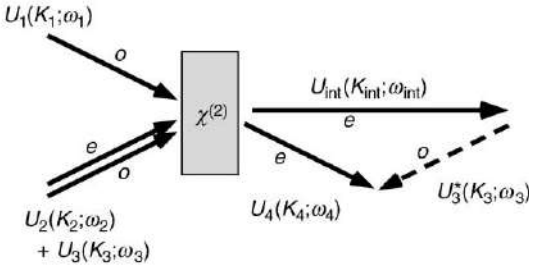

(b) Four-wave mixing by cascaded $\chi^{(2)}$ processes

Figure 3 (a) Three-wave mixing interaction: two input waves U_1 and U_2, generate a third wave U_3. (b) Four-wave mixing by cascaded second order nonlinearities: three input waves, U_1, U_2, and U_3, generate a fourth wave U_4. The first two waves generate an intermediate wave $U_{\rm int}$, by a frequency-sum process, followed by a frequency-difference process between the intermediate wave and the third input wave.

temporal frequency domain, and observe that the difference lies with the argument of the spectral term. If we apply the high-resolution approximation to eqn [8], then each spectral component would occupy a small region in space, as defined by the extent of the aperture function's SFT. Substituting for $\omega - \omega_0$ by $-\omega_0 x'/\alpha f$ for the spectral argument of eqn [8], corresponding to the high-resolution limit, equates the two expressions (essentially the same approximation as was performed in deriving eqn [9]).

Nonlinear Wave Mixing With Nondepleting Pumps

A nonlinear crystal exhibiting a large nonlinear susceptibility $\chi^{(2)}$ is placed at the Fourier plane of the spectral processing device. We employ noncollinear phase matching in our experiments. Either type-I or type-II phase matching conditions can be used in three-wave mixing experiments, with the resultant frequency-sum wave propagating along the bi-sector direction (see Figure 3a). Since both input waveforms are at ω_0 frequency, the generated field will be at a doubled carrier frequency, $2\omega_0$, due to the frequency-sum process. Four-wave mixing experiments use type-II phase matching in the $\chi^{(2)}$ crystal, involving a frequency-sum process generating an intermediate wave followed by a frequency-difference process between the intermediate wave and the third input wave. The resultant wave will co-propagate with one of the input waves and at the same temporal carrier frequency (propagation direction and frequency ω_0), albeit at an orthogonal polarization state (see Figure 3b). The output signal can be extracted using a polarizer or a polarizing beam splitter. The interacting input waves propagate noncollinearly at the Fourier transform plane by offsetting the signals from each other at the input plane of the spectral processing device.

While it is possible, though often difficult, to obtain a precise analytic solution to the coupled mode equations governing the wave-mixing processes for the generated waves, it is beyond the scope of this article whose focus is on the signal processing functionality. We therefore limit the analysis to weak interactions with nondepleting input waves. Using the nondepleted input waves approximation, the nonlinear polarization term, which drives the generation of new waves, is constant and an ordinary differential equation characterizes the generation process.

The nonlinear polarization for the frequency-sum process is proportional to the product of the input waves, when expressed in the time domain: $P_{\rm NL}^{\rm sum}(x';t) = \chi_{\rm eff}^{(2)} U_1(x';t)U_2(x';t)$. The nonlinear polarization for the frequency-difference process is proportional to the product of an input waves and a conjugated input wave: $P_{\rm NL}^{\rm diff}(x';t) = \chi_{\rm eff}^{(2)} U_1(x';t)U_2^*(x';t)$.

The solution of the ordinary differential equation driving the generation of the new wave, takes into account the possible phase mismatch between the propagation of the nonlinear polarization and the generated wave within the nonlinear crystal medium. Since we are interested in the signal processing functionality only, we approximate the generated waves as being directly proportional to the driving nonlinear polarization. This approach is valid for obtaining the generated signal forms in the case of short nonlinear crystals (limited interaction length) satisfying the phase matching condition. In practice, the temporal and spatial bandwidths, as well as the crystal length, will introduce filtering effects which will slightly modify the output signals. However, it can be shown that when wave mixing with spectrally decomposed waves, especially mutually inverted waves, the phase matching requirement is well satisfied and the filtering effects are negligible for the employed crystal lengths. In our experiments, we typically used 2 mm long β-barium borate (BBO) crystals. Using the simplified method for obtaining the form of the generated waves, arrangements for synthesizing, processing, and detecting ultrafast waveforms can be explored.

Detection of Ultrafast Waveforms

Ultrafast optical waveforms are too rapid for acquisition by opto-electronic means alone, due to the latter's finite response time. An optical processor is usually required to convert a finite time window (typically in the tens to hundreds of ps range) to a stationary signal in space for detection by slower means. We mix two mutually inverted spectrally decomposed waves in the Fourier plane of a spectral processing device, and spatially Fourier transform the resultant wave to the output plane. As will be shown below, when one of the spectrally decomposed waves is generated by a transform limited short pulse, the output spatial signal carries the temporal information of the input waveform. We call this process a time-to-space converter.

Let two ultrafast waveforms, $s((t-t_0)/\tau)$ and $r(t/\tau)$, enter the spectral processing device from opposite directions (one strikes the grating at angle θ and the other at $-\theta$, see Figure 4). Each waveform will generate a spectrally decomposed wave in the spatial Fourier transform plane. The spectrally decomposed wave of the signal $s(t)$, as defined by eqn [9], yields the first input field:

$$U_1(x';t) \approx w\left(\frac{c(t-t_0)}{\alpha L}\right)\exp\left(-\mathrm{i}\frac{\omega_0(t-t_0)x'}{\alpha f}\right) \times \tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \qquad [11a]$$

whereas the inverted spectrally decomposed wave for $r(t/\tau)$ is given by

$$U_2(x';t) \approx w\left(-\frac{ct}{\alpha L}\right)\exp\left(\mathrm{i}\frac{\omega_0 t x'}{\alpha f}\right)\tilde{r}\left(\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \qquad [11b]$$

Without loss of generality, we set the time delay parameter in eqn [11b] to zero, such that the parameter t_0 in the spectrally decomposed wave of the signal waveform is a measure of the relative time delay of the two waveforms. With full knowledge of the two input waves, U_1 and U_2, we can express the resultant waveform of the three-wave mixing process as:

$$U_3(x';t) \propto P_{\mathrm{NL}}^{\mathrm{sum}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right)w\left(-\frac{ct}{\alpha L}\right) \times \exp\left(\mathrm{i}\frac{\omega_0 t_0 x'}{\alpha f}\right)\tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\tilde{r}\left(\frac{\omega_0\tau x'}{\alpha f}\right) \times \exp(\mathrm{i}2\omega_0 t) \qquad [12]$$

Next we apply a SFT to propagate the wave U_3 to the output plane, yielding:

$$U_{\mathrm{out}}(x'';t) = \int U_3(x';t)\exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f/2}x'x''\right)\mathrm{d}x' = y\left(\frac{2\alpha x''}{c}-t_0\right)w\left(\frac{c(t-t_0)}{\alpha L}\right)w\left(-\frac{ct}{\alpha L}\right) \times \exp(\mathrm{i}2\omega_0 t_0) \qquad [13]$$

where $y(t) = s(t/\tau)\otimes r(-t/\tau)$, as the Fourier transform of a product is the convolution of their Fourier transforms (the $\otimes$-symbol denotes the convolution operation), and we use the spatial Fourier transform kernel at the doubled carrier frequency. The stationary spatial signal, $y(\cdot)$, is proportional to the convolution of the two input waveforms, where

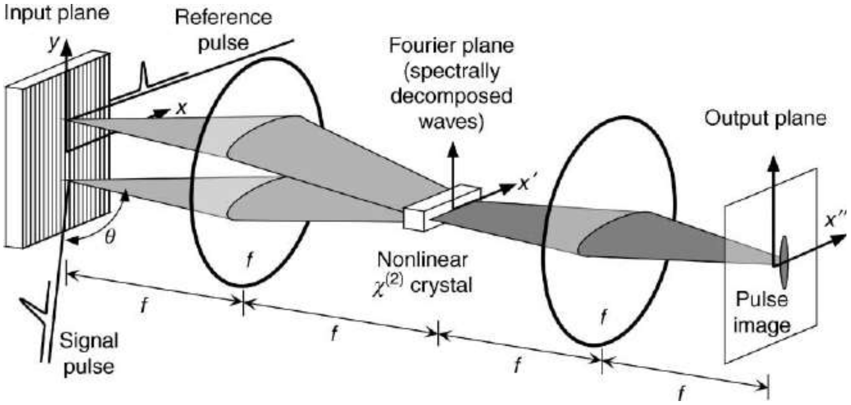

Figure 4 Arrangement for imaging ultrafast waveforms using a time-to-space conversion. The mutually inverted, spectrally decomposed waves of a signal waveform and that of a reference pulse are mixed in a nonlinear crystal placed at the Fourier plane. The generated wave is spatially Fourier transformed to the output plane, yielding a stationary image that carries the signal waveform information. © 2001 Optical Society of America.

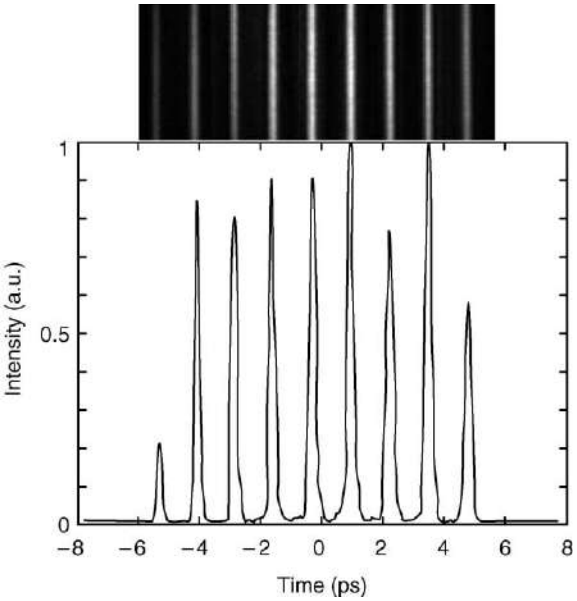

Figure 5 Output of a time-to-space converter. Top: image acquired by a CCD camera of the stationary image of a pulse sequence. Botttom: extracted time domain data from the image, showing inter-pulse time separation of 1.25 ps.

one of them is reversed in time. This mixing process was developed for time-to-space conversion by using a reference transform limited pulse for the reversed waveform. The resulting spatial image will be that of the signal waveform, albeit smoothed out due to the convolution operation with the reference pulse. The spatial signal is centered at $x'' = ct_0/2\alpha$, illustrating that the relative time delay between the two waveforms shifts the location of the output signal. The temporal characteristic is determined by the product of the time-scaled input apertures. The relative time delay, t_0, serves as an attenuation mechanism when the time windows are not perfectly aligned. When the time delay between the signal pulse and the reference pulse is too large, no output signal will emerge as the two signals do not overlap for the wave mixing process. Therefore, the time window of the spectral processing device is determined by the input aperture size, not the physical dimension of the nonlinear crystal. Consequently, we have been able to demonstrate high-resolution temporal imaging with time windows in the range of tens of picoseconds (see Figure 5). Additionally, as mentioned earlier, one of the significant advantages of this technique is the favorable phase matching condition with the spectrally decomposed wave and its inverted counterpart.

Processing of Ultrafast Waveforms

The simplest form of ultrafast waveform processing uses linear filtering, or placement of an element with transmittance characteristics that are frequency selective. However, a linear filter is a time invariant process and cannot perform more advanced processing operations of ultrafast waveforms such as interaction between ultrafast waveforms. For such processing applications, we mix three spectrally decomposed waves in the Fourier plane of a spectral processing device to generate a resultant output ultrafast waveform in real time. We illustrate advanced processing by performing a time reversal experiment using the spectral phase conjugation and spectral inversion techniques in four-wave mixing experiments.

Consider three ultrafast waveforms, $s_1(t - t_0/\tau)$, $s_2((t - t_1)/\tau)$ and $s_3(t/\tau)$, which enter a spectral processing device for generation of their respective spectrally decomposed waves. Let all three waves enter the spectral processing device from the same direction, such that the spatial dispersion is in the same direction for each signal (as defined by eqn [11a]), and generate the spectrally decomposed waves $U_1(x';t)$, $U_2(x';t)$, and $U_3(x';t)$, respectively (see Figure 6a). Since we utilize a cascaded second-order nonlinearity process, we let the first two waves interact first in a frequency-sum process, generating the intermediate field:

$$U_{\mathrm{int}}(x';t) \propto P_{\mathrm{NL}}^{\mathrm{sum}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right)w\left(\frac{c(t-t_1)}{\alpha L}\right) \times \exp\left(-\mathrm{i}\frac{\omega_0(2t-t_0-t_1)x'}{\alpha f}\right)\tilde{s}_1\left(-\frac{\omega_0\tau x'}{\alpha f}\right) \times \tilde{s}_2\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}2\omega_0 t) \qquad [14]$$

It should be noted that upon a more rigorous analysis of the generated frequency-sum wave, phase matching is not well met across the temporal bandwidth of the waveform, as is the case when wave mixing mutually inverted spectrally decomposed waves. This will introduce time domain spectra filtering effects, but for simplicity these effects are ignored below. The generated frequency-sum wave interacts with the third input wave in a frequency-difference process, giving rise to the fourth wave:

$$U_4(x';t) \propto P_{\mathrm{NL}}^{\mathrm{diff}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right)w\left(\frac{c(t-t_1)}{\alpha L}\right) \times w\left(\frac{ct}{\alpha L}\right)\exp\left(-\mathrm{i}\frac{\omega_0(t-t_0-t_1)x'}{\alpha f}\right) \times \tilde{s}_1\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\tilde{s}_2\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\tilde{s}_3^*\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \qquad [15]$$

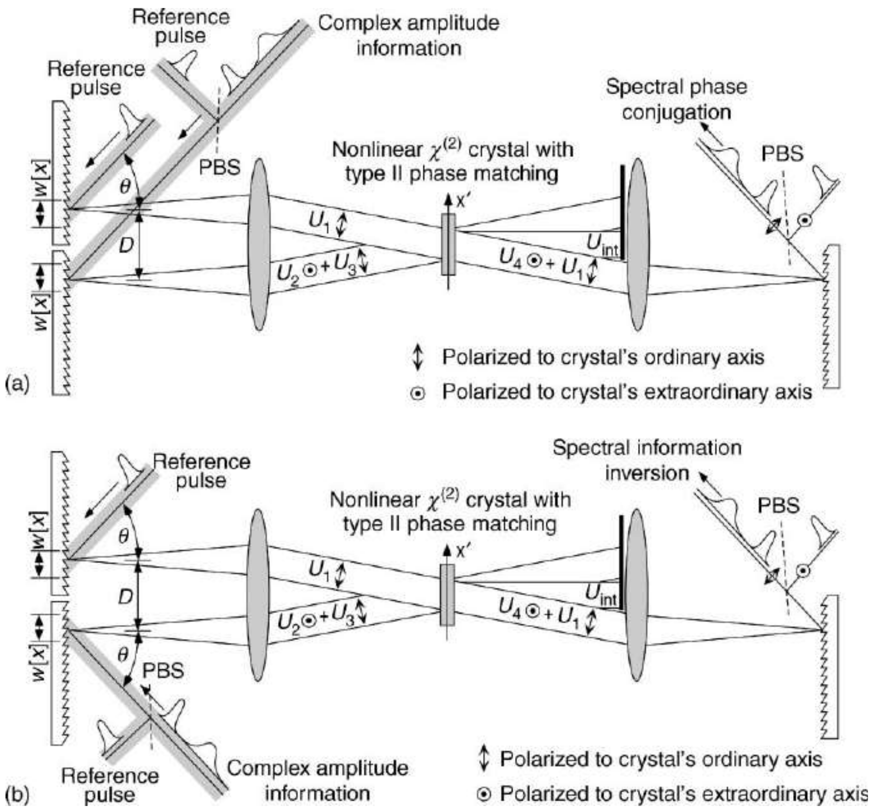
Reference pulse
Complex amplitude information
Reference pulse
Spectral phase conjugation
PBS
Nonlinear χ(2) crystal with type II phase matching
x′
w[x]
θ
D
U1
U2 ⊙ + U3
U4 ⊙ + U1
Uint
PBS
w[x]
Polarized to crystal's ordinary axis
⊙ Polarized to crystal's extraordinary axis
(a)
Reference pulse
Spectral information inversion
PBS
Nonlinear χ(2) crystal with type II phase matching
x′
w[x]
θ
D
U1
U2 ⊙ + U3
U4 ⊙ + U1
Uint
w[x]
θ
PBS
Reference pulse
Complex amplitude information
Polarized to crystal's ordinary axis
⊙ Polarized to crystal's extraordinary axis
(b)

Figure 6 Arrangements of processing ultrafast waveforms within a spectral processing device. (a) Three input waveforms are introduced from the same direction, and their spectrally decomposed waves interact via four-wave mixing. (b) The three input waveforms are introduced from opposite directions, giving rise to mutually inverted spectrally decomposed waves. These two arrangements can perform time reversal, the first by spectral phase conjugation and the second by spectral information inversion. © 2001 IEEE.

where we assume that the window function $w(\cdot)$ is real. Next we apply a SFT to propagate the wave U_4 to the output plane, yielding:

$$\begin{aligned} U_{\text{out}}(x'';t) &= \int U_4(x';t)\exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f}x'x''\right)\mathrm{d}x' \\ &= y\left(\frac{\alpha x''}{c}+t-t_0-t_1\right)w\left(\frac{c(t-t_0)}{\alpha L}\right) \\ &\quad \times w\left(\frac{c(t-t_1)}{\alpha L}\right)w\left(\frac{ct}{\alpha L}\right)\exp(\mathrm{i}\omega_0 t_0) \qquad [16] \end{aligned}$$

where $y(t)=s_1(t/\tau)\otimes s_2(t/\tau)\otimes s_3^*(-t/\tau)$. The new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction (compare to the input signals scanning in the opposite direction along the input plane, eqn [6]). By placing another diffraction grating at this plane, the new ultrafast waveform can be recomposed and made to propagate once more in free space. The waveform information consists of a convolution between the first and second signals, followed by a convolution with the third signal, which is both conjugated and reversed. If we set the first and second signals to be transform limited pulses, then the resultant waveform will carry the reversed and conjugated information of the third signal. Such capability can be very important for dispersion compensation in optical fiber communication links.

Let us now consider an alternate arrangement, where the second and third input signals are introduced to the spectral processing device from the opposite direction, and consequently have reversed spatial dispersion direction (see Figure 6b). Therefore, the first signal will generate a spectral decomposed wave $U_1(x';t)$, as defined by eqn [11a], whereas the second and third signals will give rise to the spectrally decomposed waves $U_2(x';t)$ and $U_3(x';t)$, with inverted spectra relative to that of U_1, as defined by eqn [11b]. The interaction of the first two waves gives rise to:

$$\begin{aligned} U_{\text{int}}(x';t) \propto P_{\text{NL}}^{\text{sum}}(x';t) &= w\left(\frac{c(t-t_0)}{\alpha L}\right) \\ &\quad \times w\left(-\frac{c(t-t_1)}{\alpha L}\right)\exp\left(\mathrm{i}\frac{\omega_0(t_0+t_1)x'}{\alpha f}\right) \\ &\quad \times \tilde{s}_1\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\tilde{s}_2\left(\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}2\omega_0 t) \end{aligned} \qquad [17]$$

which is identical to the wave mixing process of the time-to-space converter (see eqn [12]). However, in this case, we follow the first wave mixing process by a frequency-difference process, generating:

$$
\begin{aligned}
U_4(x';t) \propto P_{\mathrm{NL}}^{\mathrm{diff}}(x';t) = {} & w\left(\frac{c(t-t_0)}{\alpha L}\right) w\left(-\frac{c(t-t_1)}{\alpha L}\right) \\
& \times w\left(-\frac{ct}{\alpha L}\right) \exp\left(-\mathrm{i}\frac{\omega_0(t-t_0-t_1)x'}{\alpha f}\right) \tilde{s}_1\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \\
& \times \tilde{s}_2\left(\frac{\omega_0 \tau x'}{\alpha f}\right) \tilde{s}_3^*\left(\frac{\omega_0 \tau x'}{\alpha f}\right) \exp(\mathrm{i}\omega_0 t) \qquad [18]
\end{aligned}
$$

Next we apply a SFT to propagate the wave U_4 to the output plane, yielding the output waveform:

$$
\begin{aligned}
U_{\mathrm{out}}(x'';t) &= \int U_4(x';t) \exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f} x' x''\right) \mathrm{d}x' \\
&= y\left(\frac{\alpha x''}{c} + t - t_0 - t_1\right) w\left(\frac{c(t-t_0)}{\alpha L}\right) \\
&\quad \times w\left(-\frac{c(t-t_1)}{\alpha L}\right) w\left(-\frac{ct}{\alpha L}\right) \exp(\mathrm{i}\omega_0 t_0) \qquad [19]
\end{aligned}
$$

where $y(t) = s_1(t/\tau) \otimes s_2(-t/\tau) \otimes s_3^*(t/\tau)$. As before, the new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction, and will propagate in free-space upon another grating diffraction. This time, however, the waveform information consists of a convolution between the first signal and a time-reversed second signal, followed by a convolution with the conjugated third signal. If we set the first and second signals to be transform limited pulses, then the resultant waveform will carry the conjugated information of the third signal. If we set the first and third signals to be transform limited pulses, then the resultant waveform will carry the reversed information of the second signal.

We have experimentally demonstrated the two time-reversal techniques: the first technique whereby the resultant signal is both reversed and conjugated is labeled 'spectral phase conjugation', as the signal's spectrum is conjugated by the frequency difference process, and the second technique, where the signal is purely reversed in time, is named 'spectral information inversion', as the spectrum is flipped about the carrier frequency (see Figure 7). By proper selection of the input waveforms, one can also obtain higher-level operations and functionality such as convolutions and correlations among the different signals. All these operations are time-variant, and cannot be performed by passive filtering alone.

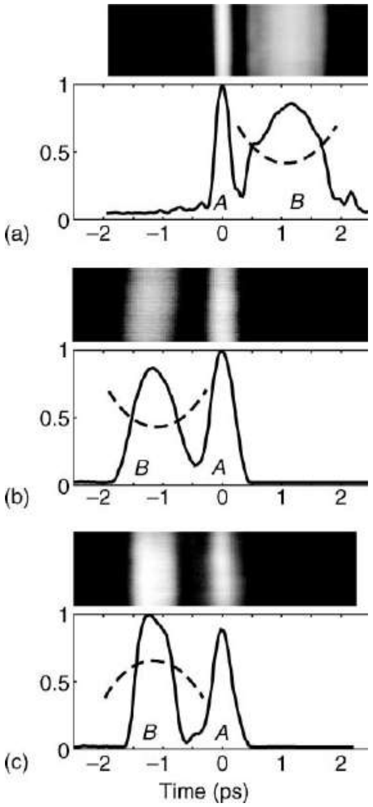

Figure 7 Time reversal experimental results. (a) Input pulse pair consisting of a transform limited pulse followed by a chirped pulse. (b) Time reversal by spectral information inversion technique. Chirped pulse leads transform limited pulse and maintains the sign of the chirp. (c) Time reversal by spectral phase conjugation technique. Chirped pulse leads transform limited pulse and flips the sign of the chirp. Data acquired with time-to-space imaging technique and sign of quadratic term (dashed line) determined by translating output plane.

Synthesis of Ultrafast Waveforms

Shaping of ultrafast waveforms is usually performed by filtering an input transform limited pulse by a fixed filter. Adaptive filtering can be performed by various means incorporating optoelectronic modulators such as liquid-crystal-based or acousto-optic-based spatial light moduloators, however the response time required for changing the waveform is typically limited to the ms and μs regime, respectively. For ultrafast reconfigurability, we utilize our spectral processing device and wave mixing, to achieve pulse shaping controlled by light, or a second optical signal.

Let the three input signals entering the spectral processing device be an ultrashort waveform, $s(t/\tau)$, and two additional spatial domain signals; one carrying spatial information $m(x/\chi)$, where χ is a scaling factor, and the second a point source (see Figure 8). Both spatial domain signals are illuminated

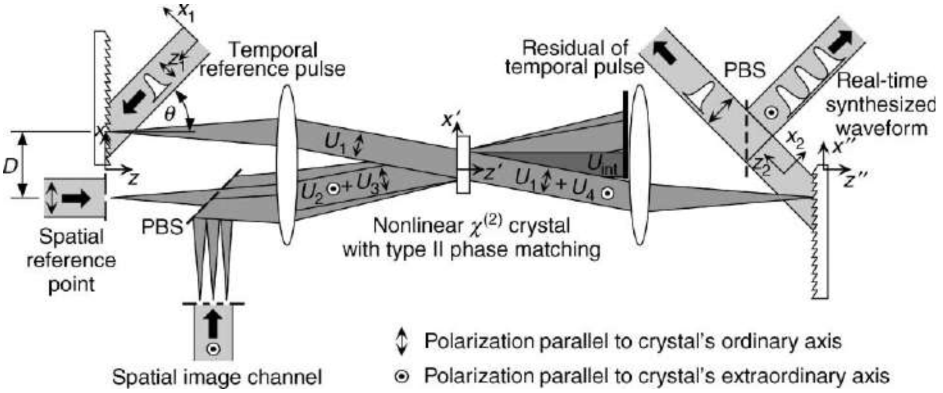
x_1
Temporal reference pulse
z_1
θ
x
z
D
Spatial reference point
PBS
Spatial image channel
U_1
U_2+U_3
x'
z'
U_1+U_4
U_{int}
Residual of temporal pulse
Nonlinear $\chi^{(2)}$ crystal with type II phase matching
PBS
Real-time synthesized waveform
x_2
z_2
x''
z''
Polarization parallel to crystal's ordinary axis
Polarization parallel to crystal's extraordinary axis

Figure 8 Arrangement for ultrafast waveform synthesis by wave mixing temporal and spatial information channels. Three input channels consist of a temporal channel, a spatial channel carrying information, and a second spatial channel consisting of a point source. The information from the spatial channel is transferred to the temporal channel via the wave mixing process. © 2001 IEEE.

by a monochromatic light source at frequency ω_0. The input ultrashort waveform gives rise to a spectrally decomposed wave $U_1(x';t)$, as described by eqn [11a]. The spatial information channel is SFT by a lens, generating the wave:

$$U_2(x';t) \propto \tilde{m}\left(\frac{\omega_0 \chi x'}{2\pi cf}\right)\exp(\mathrm{i}\omega_0 t) \qquad [20]$$

The signal from the point source (or $\delta(x)$) gives rise to a wave $U_3(x';t)$ with no spatial information (plane wave). As before, we utilize cascaded second-order nonlinearities for the wave mixing process. The interaction of the first two waves gives rise to the intermediate frequency-sum wave:

$$\begin{aligned} U_{\mathrm{int}}(x';t) &\propto P_{\mathrm{NL}}^{\mathrm{sum}}(x';t) \\ &= w\left(\frac{ct}{\alpha L}\right)\tilde{s}\left(-\frac{\omega_0 \tau x'}{\alpha f}\right)\tilde{m}\left(\frac{\omega_0 \chi x'}{cf}\right)\exp(\mathrm{i}2\omega_0 t) \end{aligned} \qquad [21]$$

The frequency-difference process with an informationless waveform serves to reduce the carrier frequency of resultant waveform back to ω_0. The down-converted wave U_4 is SFT to the output plane, yielding:

$$U_{\mathrm{out}}(x'';t) = y\left(\frac{\alpha x''}{c} + t\right) w\left(\frac{ct}{\alpha L}\right)\exp(\mathrm{i}\omega_0 t_0) \qquad [22]$$

where $y(t) = s(t/\tau) \otimes m(-ct/\alpha\chi)$. As in the processing of ultrafast waveforms section above, the new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction. The synthesized ultrafast waveform can propagate in freespace after an additional grating diffraction. The waveform information, in this case, consists of a convolution between the input temporal signal and a time-scaled and reversed version of the spatial information channel. Since the input temporal channel is typically a transform limited pulse, the output waveform carries the spatial information image, mapped to the time domain. In analogy to the imaging of ultrafast waveforms section, we call this technique space-to-time conversion. We have generated experimentally numerous waveforms that are controlled by a space domain image, or mask (flat-top waveforms, Figure 9, and pulse sequences of Figure 5 were generated by a cylindrical lens array in the spatial information channel).

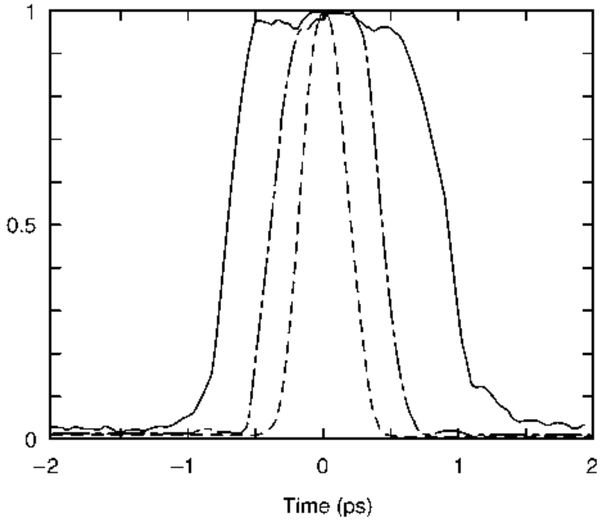

Figure 9 Synthesized flat-top waveforms via space-to-time conversion utilizing a variable slit in the spatial information channel. Increasing the width of the slit results in broader square-shaped pulses in real-time. Solid line: 2 ps pulse, dot-dash line: ~1 ps pulse, dashed line: ~300 fs pulse.

Conclusions

To properly utilize ultrafast waveforms in scientific and advanced technology applications requires instantaneous techniques from generation to processing and imaging of the waveforms. We have studied the nonlinear wave-mixing approach within a spectral processing device, and demonstrated the ability to perform all of these tasks. The wave mixing was performed in a second-order nonlinear crystal, using three-wave mixing for detection and four-wave mixing for synthesis and processing of ultrafast waveforms. The three-wave mixing process results in a second-harmonic output image, suitable for acquisition by silicon-based CCD imagers. The four-wave mixing processes result in an output waveform of an identical carrier frequency, as typically required in processing and synthesis applications. The high nonlinear coefficient of $\chi^{(2)}$ crystals enabled us to demonstrate these techniques with ultrashort pulse energies ranging from nJ to mJ levels. Further advantages afforded by the techniques are expanded time windows and favorable phase matching.

This review of the instantaneous optical signal processing capabilities with optical nonlinearities and interaction with ultrafast waveforms introduced and demonstrated the basic signal processing capabilities to manipulate optical signals in amplitude and phase. The same techniques can be used and expanded upon to achieve different signal processing abilities, which are beyond the scope of this article.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Ultrafast Laser Techniques:** Pulse Characterization Techniques. **Ultrafast Technology:** Femtosecond Chemical Dynamics: Gas-Phase; Femtosecond Condensed Phase Spectroscopy: Structural Dynamics; Ultrafast Illumination and Processing.

Further Reading

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena: Fundamentals, Techniques, and Applications on a Femtosecond Time Scale*. San Diego: Academic Press.

Marom DM, Panasenko D, Sun P-C and Fainman Y (2000) Femtosecond-rate space-to-time conversion. *Journal of the Optical Society of America B* 17: 1759–1773.

Marom DM, Panasenko D, Sun P-C, Mazurenko YY and Fainman Y (2001) Real-time spatial-temporal signal processing with optical nonlinearities. *IEEE Journal of Selected Topics in Quantum Electronics* 7: 683–693.

Miller DR, Murnane MM, Scherer NF, Weiner AM (eds) (2003) *Ultrafast Phenomena XIII, Proceeding of the 13th International Conference*. Berlin: Springer-Verlag.

Weiner AM (1995) Femtosecond optical pulse shaping and processing. *Progress in Quantum Electronics* 19: 161–238.

Weiner AM and Kanan M (1998) Femtosecond pulse shaping for synthesis, processing, and time-to-space conversion of ultrafast optical waveforms. *IEEE Journal of Selected Topics in Quantum Electronics* 4: 317–331.

EARLY HISTORY OF QUANTUM ELECTRONICS

C H Townes, University of California, Berkeley, CA, USA

© 2005, C H Townes. Published by Elsevier Ltd. All rights reserved.

All the ideas essential to making a laser were known before 1930, but there was no operating laser before 1960. So why didn't the laser come sooner? There are several reasons. One important impediment to the laser invention was that a combination of ideas from quantum mechanics and from electrical engineering was needed, and these two fields were not well mixed in the early days. Another is that, while some physicists recognized that amplification could occur if there was a population inversion of states, they did not consider the coherence of such amplifications and did not recognize its importance or usefulness. It was just something that in principle could happen, but was not very interesting. What was required was a combination of physics and electrical engineering thinking and recognition of the coherence of such amplification. In addition the importance of such a development had to be visualized and recognized in a way that it led to very devoted work towards its achievement.

It is striking that lasers grew out of the study of microwave spectra of molecules, for which engineering and quantum mechanics were both important and which helped orient thinking in the appropriate direction. That this origin was not accidental is convincingly demonstrated by the fact that three independent ideas for amplification by stimulated emission were generated about 1950. They were from Joe Weber, at the University of Maryland, from Nikolai Basov and Alexander Prokhovov at the Soviet Academy, and myself at Columbia University. Weber primarily wanted to point out the possibility, but didn't try to do it. In addition, his numbers were a bit off and no practical system was suggested. Basov and Prokhovov actually worked towards microwave amplification using a beam of molecules, as did I.

That stimulated amplification was recognized early, but not thought through, is illustrated by the fact that Prof. Richard Tolman, a theoretical physicist, wrote a discussion in 1924 of the net absorption of light by molecules, pointing out in particular that induced emission counteracted absorption and noting that if there were more molecules in the upper than in the lower state there could be 'negative absorption'. But, he wrote, 'This would usually be very small'. The Russian physicist Vitaly Ginsburg wrote me, after the maser and laser had appeared, that his professor, S.M. Levi, had been well aware of such effects back in the 1930s and had told him 'create an overpopulation at higher atomic levels and you will obtain an amplifier; the whole trouble is that it is difficult to create a substantial overpopulation of levels'.

The German physicist F.G. Houtermans said to me that in 1932, when told by a colleague of an unusual light intensity in a gaseous discharge, he had thought it might be a 'photon avalanche', i.e. multiplication of photons by stimulated emission.

Another Russian physicist, V.A. Fabricant, wrote a thesis in 1939 in which he discussed absorption and emission of light radiation in a gas and looked for 'negative absorption', or amplification. He did not discuss coherence or a resonant cavity, and was not able to achieve any amplification so his work was quickly forgotten. None of these early mentions of stimulated emission proposed how to actually get amplification, that it would be useful, nor clearly noted its coherence. Tolman did, however, write in 1927 that, 'We should expect radiation induced by an external field to be coherent with the radiation associated with that field'. I know of no proposal to actually make use of such amplification, until those made by microwave spectroscopists in the early 1950s.

Another clear indicator that even in the 1950s physicists and engineers did not think amplification by stimulated emission was particularly interesting nor useful is that during the $2\frac{1}{2}$ years when Jim Gordon, Herb Zeiger, and I were working on trying to obtain microwave amplification (the maser), a large number of scientists visited my laboratory, saw what we were doing, but no one bothered to also try to obtain such amplification. After the maser worked, it hit the newspapers and then very quickly became a popular and intense field of interest for a number of physicists.

My own drive to produce oscillators by stimulated emission came from my strong interest in obtaining sources of waves shorter than the few millimeters wavelength which could be produced by electronic devices, in order to extend the high resolution study of molecular spectra to wavelengths shorter than microwaves, down into the far infrared. My students and I worked on several possible schemes for producing waves shorter than those produced by klystrons or magnetrons – frequency multiplication by nonlinearities, electronic beams passing over surfaces of solid materials with resonances, and anything else I could think of. None worked very well.

In early 1950, I was asked by the Office of Naval Research to form and chair a committee which would examine possible research towards obtaining wavelengths down to one millimeter and shorter. I chose outstanding scientists and engineers in a variety of fields which might touch on this problem. We met together, and visited many pertinent laboratories and individuals interested in such research. Nothing very promising seemed to turn up. But we wanted, of course, to at least provide a report summarizing the situation as we saw it. Our last meeting was April 26, 1951, in Washington, D.C. Worrying over our lack of success, I woke up early before the meeting. Breakfast was not ready, so I walked over to nearby Franklin Park, sat on a bench in front of beautiful blooming azaleas, and puzzled over why neither I nor the Committee had found any promising solutions.

I went over all the ideas I had had previously. Molecules, of course, can produce high frequencies. But I had previously concluded, I thought wisely and rigorously, that one could not obtain intense radiation from them because radiation intensity was a function of temperature, and the temperature could not be very high without destroying the molecules. Suddenly I realized that they did not need to have a temperature in the usual sense – they need not be in temperature equilibrium. There could be more molecules in an upper than a lower excitation state, which could in principle produce indefinitely intense radiation. I pulled a piece of paper out of my pocket and wrote down the equations and numbers for such a case, using a molecule beam sent into a resonant cavity and ammonia molecules with which I was very familiar. My equations said one could get enough excited molecules, low enough loss in a resonant cavity, and it would work! Why hadn't I thought of it before?!

Back at Columbia, where I worked, and about 4 months later, the graduate student Jim Gordon agreed to work on trying to obtain such an oscillator using a beam of ammonia. I assured him that if it didn't work he could modify the experiment to do interesting spectroscopy and thus complete a thesis. But we both thought he had a good chance of making it work. And a young post doc working with me, Herb Zeiger, joined the effort.

Actually, I had first thought of obtaining stimulated emission from a molecular beam back in 1948, but simply as a demonstration of physical principles rather than as a useful amplification. Also, about a year later a young post doc, J.W. Trischka, working on molecular beams with Professors Rabi and Kusch, had also thought of demonstrating stimulated emission. We talked about it together, and he decided it wasn't worthwhile just demonstrating the effect because it wouldn't really prove any new physics. Neither of us at that time had recognized the real point and the possibility of useful amplification, which is one of the reasons mentioned above that the idea was delayed as long as it was.

It is perhaps important to emphasize again and to illustrate how out-of-the-way the use of stimulated emission was at that time for physicists, and how distant stimulated emission was from engineers. While we were working on the ammonia beam maser, Prof. L.H. Thomas, an outstanding physicist known for the 'Thomas Effect', frequently would run into me in the hallway at Columbia University and say that I didn't understand, and that my proposed ammonia oscillator could not work. However, I never got a clear explanation from him of what it was I didn't understand. And after we had been working on the ammonia maser system for about 2 years (a more-or-less normal time for a student thesis project), the Physics Department Chairman, Prof. Polycarp Kusch, and the previous chairman, Prof. I.I. Rabi, came into my office to object. They were excellent physicists, both received Nobel Prizes, and both were experts on molecular beams. They sat down in my office and said 'Look, Charlie, that is not going to work. We know it won't work and you know it won't work. You are wasting departmental money, and must stop!' Other people had also questioned what I was doing, frequently in particular whether the stimulated radiation would be coherent. So I had already thought over the situation many times. I had full notes from the quantum mechanics course I took as a student back in 1938, and could derive from them a proof of coherence. I felt also that I knew the quantitative numbers, such as the molecular beam intensity and the possible 'Q', or loss, in the cavity resonator, very well. I was also by then an Associate Professor, and the department chairman could not fire me simply because of stupidity. I replied to Rabi and Kusch 'No, I believe it has a reasonable chance of working and I'm going to continue!' Annoyed, they stomped out of the room.

About 2 months after the Rabi/Kusch incident, Jim Gordon dashed into the classroom where I was lecturing and said loudly 'It's working!' Most of the class then went up to the lab to see the new device. Rabi and Kusch were not against me; even though they were outstanding physicists, they probably just didn't quite comprehend the device. A couple of months after its successful operation, Kusch more or less apologized by saying 'Well, I should have realized that you probably know more about what you are doing than I do.'

After the maser successfully operated, there were other incidences showing the lack of appropriate focus of physicists on stimulated emission. I was a friend of Aage Bohr, the son of Niels Bohr, and in that connection was visiting him in Denmark. Niels Bohr asked me what research I was presently doing, so I told him about our new oscillator, the maser, and its remarkably pure frequency. He looked at me and said 'Oh no, that's not possible. You must be misunderstanding something.' I emphasized again what it was really doing, but he still seemed not to believe it could function that way. I presumed he was thinking in terms of the uncertainty principle and the finite time of passage of a molecule through the cavity, though I never was quite sure just why he felt it impossible. A similar thing happened shortly after that at a cocktail party in Princeton, where John Von Neumann asked the same question – what was my research at the moment? After telling him about the maser oscillator and the purity of frequency he reacted very similarly 'Oh no,' he said, 'That can't be right. You must be misunderstanding something.' After arguing a little more, he left to get another drink. Fifteen minutes later he came back, saying 'Hey, you are right!' He had understood, and wanted to talk much more about the maser and of possibly using excited semiconductors. Only after his death did I learn from his notebooks that he had considered exciting electrons in semiconductors with neutrons from a reactor, and had written Edward Teller about whether some experiments might be done with this to obtain intense light. However, he had not considered coherence, and Teller was apparently not interested enough to respond, so the matter was dropped.

The above account reminds me a bit of the amusing comments of Arthur Clark on 'Change'. He writes:

'People go through four states before any revolutionary development:

1. It's nonsense, don't waste my time
2. It's interesting, but not important
3. I always said it was a good idea
4. I thought of it first'.

It's clear that the world of physics was thinking very little in the direction of masers or lasers, and that many preconceptions as well as lack of interest stood in the way. My own experience with engineering at Bell Telephone Labs during World War II, in designing radar and electronic systems, plus my intense interest in obtaining short-wave oscillators, were clearly important in bringing me to the right ideas.

As masers became very interesting to the physics community and the field grew rapidly, my engineering experiences continued to be of help. I was well acquainted with the theoretical examination of noise in vacuum tube amplifiers by various individuals at Bell Labs, and recognized that the maser could provide much more sensitive amplification than could common electronic amplifiers, where discrete electronic charges produce the basic noise. On sabbatical in France, I worked on electron spin masers with Jean Combrisson and Arnold Honig who had the appropriate equipment. And then in Japan, Koichi Shimoda, Hidetoshi Takahashi, and I wrote a theoretical paper on the basic quantum noise of maser (or laser) amplification. The theory showed that maser amplifiers of microwaves should be about 100 times more sensitive than the existing electronic ones.

After 2–3 years of maser experiments and development I felt I wanted to move on to the shorter wavelengths for which I had generated the maser idea. Although I had first tried the idea at microwave frequencies because that seemed the easiest way to test out the general idea and the result had been exciting, I still wanted those shorter wavelengths. I had not come up with any great ideas of just how to get to much shorter wavelengths, which is why I waited several years after the maser worked before moving on. However, in 1957, 3 years after the first successful operation of the maser, I decided it was high time to simply figure out what was the best way I could imagine to move on into the infrared region and do it. A number of physicists had concluded that of course one couldn't make masers work at much shorter wavelengths, certainly not in the visible region, because spontaneous emission becomes so much faster as the wavelength is shortened and adequate inversion of population was not practical. But that was intuition, not quantitative science. As I wrote down equations for what might be done to move towards shorter wavelengths using a model of atomic or molecular excitation by radiation and a reasonably high Q resonator, it became clear that it was quite practical to move on even into the visible region. That was exciting! Why hadn't I, or someone else, looked at it carefully and quantitatively before that time?

I was at the time consulting at Bell Labs, with the assignment to spend a day there every 2 weeks and just talk with the Lab's scientific personnel. Since my former post doc and now brother-in-law Arthur Schawlow was then at Bell Labs, I of course talked with him. On telling him of my ideas for an 'optical maser', using a resonant cavity and excitation of atoms with optical radiation, he said he too was interested in that, and we decided to work together to optimize a system. It was Art who then suggested use of a Fabry–Perot resonator rather than the cavity with large holes I had used for a model, and that was an excellent addition. Why I did not think of that is a mystery, but Art had done his thesis at the University of Toronto in Fabry–Perot spectroscopy, and that might have been why the thought came to him.

Since Art Schawlow was participating, I decided the patent for the new 'optical maser' should belong to the Bell Labs (I already claimed ownership of the basic maser patent, which covered all wavelengths). Hence we recognized that the new idea must be kept confidential until Bell Labs lawyers had worked out and applied for an appropriate patent. This delay in public information sheds some additional light on how the scientific and technical world was thinking at the time. I had written down my original ideas for an 'optical maser' in my notebook and had it witnessed by my student Joe Giordmaine at Columbia University. I had also talked with a Columbia student Gordon Gould because he had been doing research with an intense light source and I wanted to know how much intensity he had in order to be sure I could get enough excited atoms and provide an oscillator at these short wavelengths. Except for these two persons, neither Art Schawlow nor I told anyone outside of Bell Labs about the 'optical maser', or laser idea until after Bell Labs had properly prepared its patent, which was about 11 months after my first notebook entry. For that entire time, no one has produced a record of any thoughts about extending the maser to optical wavelengths except Gordon Gould, who entered some ideas in his notebook about 1 month after I talked with him about the possibility of an optical maser, and 2 months after my original notebook entry. His notes were later to be the source of a long patent case.

The striking observation is that no one outside of Bell Labs except Gordon Gould, to whom I had explained my ideas, seems to have written or noted down anything about extending masers to these shorter wavelengths during the 11 months from my recognition that it could well be done until after the Schawlow and Townes paper on how to do it became available. After that there was considerable excitement and a number of ideas.

It is also noteworthy that, because of low general interest and competition, I did not publish a paper on how a maser might be made, but waited for publication until we demonstrated its operation, about 3 years after the idea had arrived. But after the maser worked the field became exciting and competitive. Hence, Schawlow and I thought we surely should publish a theoretical paper establishing the idea before taking the time to make a laser work. And indeed, there was much competition to build the first laser.

I myself helped a couple of my graduate students start work on trying to build an 'optical maser' or laser. But at about that time (1958), I was urged to undertake a job in Washington as Vice President and Director of Research of the Institute for Defense Analysis, an organization put together largely by the presidents of several universities to try to help advise the government on matters of science and technology. Sputnik had gone up the year before, and everyone was worried about the position of the U.S. with respect to Communist Russia, which seemed to be ahead particularly in some areas important to defense. I decided I should try to help, and accepted a two-year appointment in Washington. I recognized that this seriously distracted my attention from developing a laser quickly, but knew there were many others working towards lasers and so such devices would certainly be developed and our doing it just at Columbia University was neither critically important nor highly probable.

The field of masers and 'optical masers' or lasers was becoming so exciting and a bit scrambled that The Office of Naval Research asked me if I would organize a meeting on the subject. I did, with the help of a committee of many distinguished people in the field or closely related science and technology. And it was in a meeting of the Committee that we christened the field with the name 'Quantum Electronics'. This first international meeting on the subject was at Shawanga Lodge in New York State in September 1959. It made an occasion for the Russians Basov and Prokhorov to visit the United States (and my lab and home), and was about 6 months before Ted Maiman made the first working laser. There were many interesting discussions of masers and their coming operation at optical wavelengths.

It is significant to note that, while the maser grew out of basic research in universities (with Russian work at the Russian Academy) and industry had little to do with early masers, all the first lasers were made in industrial laboratories. This illustrates the sociology and the strengths and weaknesses of industrial and of academic laboratories. While masers (and from them lasers) originated from research on microwave spectroscopy of molecules, industrial laboratories believed the field of microwave spectroscopy had little to offer in the way of commercial results. Because of equipment available and the interest of physicists in industry, the field was initiated and pursued immediately after World War II in commercial laboratories – by myself at Bell Labs, my friends at the RCA Labs, at Westinghouse, and at General Electric. For lack of interest in industry, such work was soon shut down, except at Bell Labs, and it moved to universities. Bell Labs generously allowed me to continue such work, although they wanted me to do some 'more useful' engineering. Once the field had obvious commercial possibilities, commercial laboratories began to support it well and the clear importance of masers and lasers made industry very interested. Really interested industry can more easily put strong new resources and push harder on a field than can academic laboratories, where money has to be granted and professors have a variety of other assignments. The first laser was made to work, of course, by Ted Maiman at Hughes Research Laboratories. Ted had been a student of Willis Lamb at Stanford, and there worked on radio spectroscopy. The second type of laser, rather similar to Maiman's but using a different material, was made to work at the General Electric Laboratories by Peter Sorokin and Mirek Stevenson. Sorokin had been a student of Bloembergen at Harvard in microwave or radio spectroscopy and Stevenson was one of my students in microwave spectroscopy at Columbia University. The next type of laser, and one I particularly admire, was made by Ali Javan, William Bennett, and Don Herriott. Javan had been a student with me in microwave spectroscopy at Columbia University, Bill Bennett a student in the molecular beam group at Columbia working on radio spectroscopy, and Don Harriott was an optics specialist. All of these originators, with the exception of Herriott, had been working at universities in the field which originated the idea and were recently hired by industry. The next important laser, involving semiconductors, was invented by Robert Hall and collaborators at General Electric Labs. Note that every one of these early lasers was created in industry.

After the first laser was operated, my own students at Columbia quickly turned from trying to make a laser to using lasers to explore more physics, the normal university function. And I am delighted that masers and lasers have provided such excellent tools for research, as well as for commercial and medical applications.

After a few years of exploring new physics with lasers, I decided that since there were many excellent scientists doing such work, I should move into fields which it seemed to me were being relatively neglected. I moved to the University of California at Berkeley to look for molecules in interstellar space by microwave spectroscopy, and to do infrared astronomy. Very soon after initiating work at Berkeley, one of my students, Albert Cheung, not only discovered the first polyatomic molecules in space, he found powerful water masers. A while before our discovery of water and identification of its radiation as due to maser action, it had been deduced that some microwave radiation of OH must be due to maser action. And since then, many, many masers due to a wide variety of molecules have been found in astronomical sources as well as a few powerful lasers. Since deviations from thermal equilibrium are common in the very thin gases excited by powerful sources in space, we should have expected this, but didn't. And these masers in space could have been easily detected with radio technology available back in the 1930s if anyone had searched the microwave spectrum carefully.

Clearly, masers and lasers could have been discovered and used much sooner than they actually were. We simply were neither thinking nor looking in the right directions. And this raises the natural question – what important and more-or-less obvious ideas are we missing now because of our lack of imaginative exploration?

Further Reading

Basov N and Prokhorov A (1954) Application of molecular beams to radio spectroscopic studies of rotation spectra of molecules. *J. Experimental. Theoretical Physics U.S.S.R.* 27: 431–438.

Bloembergen N (1956) Proposal for a new type of solid state maser. *Physical Review* 104: 324–327.

Cheung A, Rank D, Townes C, Thornton D and Welch J (1969) Detection of water in interstellar regions by its microwave radiation. *Nature* 121: 626–628.

Combrisson J, Honig A and Townes C (1956) Utilization de la resonance de spins electroniques pour realiser un oscillateur ou un amplificateur en hyperfrequencies. *Comptes Rendues* 242: 2451–2453.

Gordon J, Zeiger H and Townes C (1954) Molecular microwave oscillator and new hyperfine spectrum of NH_3. *The Physical Review* 95: 282–284.

Hall R, Fenner G, Kingsley G, Soltys J and Carlson R (1962) Coherent light emission from GaAs junction. *Physical Review Letters* 9: 366–368.

Javan A, Bennett W and Herriot D (1961) Population inversion and continuous optical maser oscillation in a gas discharge containing a He–Ne mixture. *Physical Review Letters* 6: 106–110.

Knowles S, Mayer C, Cheung A, Rank D and Townes C (1969) Spectra, variability, size, and polarization of H_2O microwave emission sources in the galaxy. *Science* 163: 1055–1057.

Maiman T (1960) Stimulated optical radiation in ruby. *Nature* 187: 493–494.

Perkins F, Gold T and Salpeter E (1966) Maser action in interstellar OH. *Ap. J.* 145: 361–366.

Schawlow A and Townes C (1958) Infrared and optical masers. *Physical Review* 112: 1940–1949.

Shimoda K, Takahashi H and Townes C (1957) Fluctuations in amplification of quanta with application to maser amplifiers. *Journal of the Physical Society of Japan* 12: 686–700.

Sorokin P and Stevenson M (1960) Stimulated infrared emission from trivalent uranium. *Physical Review Letters* 5: 557–559.

Tolman R (1924) Weak quantization. *The Physical Review* 24: 287–295.

Tolman R (1927) Statistical mechanics with applications to physics and chemistry. The Chemical Catalog Co., New York, pp. 169.

Weber J (1953) Amplification of microwave radiation by substances not in thermal equilibrium. *Trans. Inst. Radio Engineers Professional Group on Electron Devices* 3: 1–4.

LET THERE BE COHERENT LIGHT: THE DISCOVERY OF THE LASER

T H Maiman, Vancouver, BC, Canada

© 2005, T H Maiman. Published by Elsevier Ltd. All rights reserved.

Editor's Note

The discovery of the first laser was a remarkable culmination of human creativity and ingenuity. However the achievement was finally made by one individual young scientist, Theodore Maiman, working independently of the mainstream and without the benefit of strong government or private financial support. With persuasive arguments, he convinced his management at the Hughes Research Laboratories to allow him to continue his work, albeit with limited resources. And then, working alone with his assistant he saw the light that now illuminates the world of laser science and engineering. This is the story of how thinking outside the box created a new field of science, whose developing technology has advanced the state of medicine and engineering and the quality of life in general.

Introduction

I am frequently asked how it was that I discovered the laser. Many assume that the concept evolved from some sudden, inspirational thought. It didn't happen that way.

It is dramatic and exciting to have a scientific discovery emerge from a dream or a vision that comes out of nowhere, but it seldom does. In reality, almost all scientific discoveries come from building on other, prior, scientific developments.

So it was with the laser. In 1916, Albert Einstein laid the foundation and conceived the basic underlying principles on which lasers are based. He formulated and explained the relations that govern the way that atoms and molecules absorb and emit radiation. He introduced the key concept of stimulated emission. Then, in the 1920s, physicists C.H. Füchtbauer and Rudolph Ladenburg added formulations that used the Einstein theory to go further, and tie the absorption of light in a material medium to the fundamental properties of its constituent atoms. But it was the Russian physicist A.V. Fabricant who first had the vision to propose the concept of a laser in 1940.

Unlike the Sun and other incoherent light sources, the laser produces light of one precise color. Its waves are emitted in phase with each other, and its radiation can be efficiently focused to a precise spot. In short, the laser produces coherent light, with the properties we were already used to with radio, television and microwave sources.

Fabricant specified the conditions needed for amplification of light via stimulated emission. He appreciated the concept of an inverted population and the concept of coupling such an inverted population medium to a resonant structure (resonator). He proposed using a gaseous electrical discharge to achieve laser action as one possibility. Later, he proposed the use of a helium discharge lamp to optically pump the gaseous form of cesium.

Scientists Purcell and Pound were the first to report evidence of net stimulated emission in 1951. Their observation, in the radio part of the spectrum, was a by-product of their pioneering developments in the techniques of nuclear induction, the foundation for magnetic resonance imaging.

The first achievement of a working device that utilized stimulated emission as its operating principle was a microwave oscillator/amplifier. That device used a beam of ammonia molecules as the working medium. The design details were worked out independently by groups at the Lebedev Institute in the Soviet Union and Columbia University in the USA. The device was dubbed a *maser*, which is an acronym for *microwave amplification by stimulated emission of radiation*.

In 1955, the year after I was awarded my Ph.D., I joined the Hughes Research Laboratory in Culver City, California. One of my major assignments was to head a project developing a miniature, liquid-nitrogen ruby maser for the US Army Signal Corps.

Race to the Light

Before starting work on the maser I had begun to germinate some ideas about the possibility of a laser. My concept so far was to try to use a solid material for the lasing medium, and to fabricate the potential laser material into a rod shape. In microwave thinking this configuration would be called a *dielectric wave-guide*. In optics it is known as a *light pipe*.

Abridged from "The Laser Odyssey" by Theodore H. Maiman. Copyright © 2000 Theodore H. Maiman. Reproduced with the permission of the author and Laser Press, Blaine, WA, USA.

I planned to put mirrors on each end of the rod to form a resonator.

Despite the advances in physics it represents, the maser does not in any way extend the coherent electromagnetic spectrum. Furthermore, its use as an amplifier turned out to be impractical because cryogenic temperatures were required for proper operation. There was strong motivation to push the frontiers of coherent electromagnetic radiation to higher frequencies, principally the promise of higher energies and greater concentrations. The reduction in wavelength from microwave to visible is a factor of about 10^4. This would lead to a consequent 10^4-fold reduction in the minimum achievable focal size, and a 10^8-fold increase in energy concentration delivered by the laser.

Work on the maser project therefore proved something of a distraction, and it was not until August 1959 that I was able to devote full effort to the consideration and analysis of my concepts for a laser.

Why would I be willing to enter such a race? There was rather formidable global competition already in play, well-funded and very competently manned. The answer lies in my knowledge of the proposals that had been floated about. Generally, the scientists who pursued these notions were not particularly secretive about what they were working on. They published and gave talks at conferences. On the whole, however, I found that the authors and conference presenters were only offering vague proposals. To be sure, this served as an exchange of information and stimulated ideas. But it didn't appear to me that anyone was close to the answer. In short, no viable laser concept was yet in existence.

As I look back, I was a little brash. I would be thrusting myself, in a sense, into a technological Olympics. The competition was of the best quality and of international scope. But my competitive spirit won out. The challenge of working in the top league of such an exciting project, that had so many questions and problems to resolve, was very compelling to me.

We know *now* that many kinds of lasers can be made. But back then, in 1959, we didn't know. We didn't even know with any confidence if it was really possible to make a laser at all. My strategy was to limit myself to potential solutions to the making of a laser that did not have appreciable distractions in the design. That way I could focus strictly on just the laser problem itself.

For various reasons I was reluctant to work on alkali-vapor or gas discharge systems. I chose to work with solid-state crystals. The main appeal that solids held for me was simplicity. By that, I mean simple in analysis and understanding, and simple in device conception.

In contrast to a gas discharge, the energy level diagram in an appropriate crystal is very limited. There are relatively few possibilities for the energy states, and by and large, the pertinent parameters for a potential laser candidate are amenable to a combination of calculations and relatively straightforward direct measurements.

Another advantage, in principle, to a solid crystal is its relatively high gain coefficient. By that, I mean the amplification in a given length of material is of reasonable proportions. This meant that the laser medium could be relatively small in size and short in extent, and I would not have the problem of developing or depending on the use of special mirrors. Indeed, my first laser used a crystal that was only 2 cm long.

I was also intrigued with the concept of a solid medium since I would not have to deal with vacuum pumps, impurity problems and gas handling apparatus, or complex mirror mechanisms. I could put simple silver mirror coatings directly on the crystal as I had done with my small ruby maser. In principle, a solid crystal laser could be designed to be very simple, compact and rugged.

My first choice to study and contemplate was that of a ruby crystal. Ruby is the result when a water-clear cast of aluminum oxide is doped with a small amount of chromium oxide impurity. It is the chromium that is responsible for the red color. The rubies used for devices are usually not natural gemstones, which have a chromium impurity level of about 0.5%. The chromium concentration in device rubies is around ten times less than this, and hence they display a lighter red color than gemstone ruby and are referred to as *pink ruby*.

Why choose ruby as a potential laser candidate? Most importantly, I was quite familiar with and fascinated by the interesting optical properties of the crystal. Ruby is a fluorescent mineral; if ultraviolet light is shone on a ruby, it will glow with deep red fluorescence. Furthermore, ruby also gives off a red glow when either blue or green light is shone upon it. It is these blue and green absorption bands that give ruby its red color. When a green photon impinges on and is absorbed by the ruby, a chromium impurity ion is raised from its ground state into a broad, excited band. Although the chromium ion has the possibility of radiating by spontaneous emission from that excited level, another process comes into play. The competing process uses the thermal vibrations of the crystal lattice to interact with the excited ion and deposit most of the excitation energy to another slightly lower excited level of the ruby chromium where it stays for a while. This latter process is much more probable and dominates.

The energy from this third level is radiated as spontaneous incoherent emission. This spontaneous emission is made up of red photons and is the observed fluorescence. The level where the fluorescence emanates is sometimes called a metastable level, since the chromium ions linger in that energy state for a comparatively long time before they radiate red photons.

I developed a model that could be mathematically analyzed and I set up kinetic equations to take into account the various mechanisms taking place in the fluorescent process. I also set up simple intuitive criteria for establishing the condition for laser action. This model and these equations have subsequently become a standard way for others to analyze crystal lasers.

I was able to determine which material parameters were important and relevant to a laser by getting the solution to the equations describing the model. Using known and estimated values for the pertinent parameters in ruby, I found that ruby would require a very bright pump lamp to excite the crystal sufficiently to allow laser action to take place. The brightness of a lamp is an important parameter for lasers. It is a measure not of the total power radiated by the lamp, but rather the power per unit area.

Obstacles and Solutions

Ruby has many desirable qualities. It is a very stable, very rugged crystal to work with. It has some broad pumping bands that make it helpful to get a reasonable energy transfer from an incoherent pumping lamp, and compared to some other crystals I considered (e.g., gadolinium-doped crystals), a working ruby laser would produce visible light.

However, there was some uncertainty over the crystal's measured quantum efficiency. The quantum efficiency for a fluorescent crystal is the number of fluorescent photons emitted, divided by the number of pumping photons. In the case of ruby, this is the number of red photons radiated out compared to the number of green photons absorbed. A report in November 1959 had suggested that ruby's quantum efficiency was only 1%. This would effectively mean that it would be impossible for ruby to achieve laser action. However, I was able to make my own measurements of ruby's fluorescent efficiency, and these showed that it was nearer 75%.

I now had enough information that I could proceed to work out an actual laser design. But how would I pump the ruby? What would be the ruby's shape and size?

I knew that I needed a very bright lamp. Among the brightest around was a high-pressure mercury arc lamp manufactured by General Electric. In addition to being one of the brightest available laboratory lamps, it had the advantage of radiating most of its energy in the green and blue-violet parts of the spectrum. This was a good match to what I needed for the ruby.

I drew up a paper design with the mercury arc lamp at the focus of an elliptic cylinder. A small ruby rod was to be mounted on the other focus of the elliptic cylinder. It is a property of an elliptically shaped reflector that a point of light placed at one focus of the ellipse will be imaged at the other focus. The elliptic cylinder that I had in mind would be highly polished on its inner surface to a mirror finish.

But I had a problem. My paper analysis of the design showed that although it should work, it would only do so by a slim margin. I studied the design in more detail, looking for ways to optimize and improve it, but I couldn't convince myself that it was anything but marginal.

I stewed over the problem at hand. I started to think about other ways to look at the problem. I was frustrated: I felt that, on the one hand I was very close to an answer, but on the other hand it was eluding me.

I went back to my analytical model. I pondered my options and decided to put the pumping lamp requirement for the ruby in a different form. I calculated the equivalent black-body temperature of a suitable pumping lamp. The temperature I calculated to have enough brightness capable of driving a ruby into laser action was close to 5000 K. By contrast an ordinary tungsten light bulb has a brightness temperature of about 2800 K.

Once I put the pumping requirement in terms of brightness temperature, I began to think in a different way. I remembered reading an article about photographic strobe lamps, a camera's flash mechanism. The article said that strobe lamps could reach brightness temperatures of 8000 K or more.

I now had my 'aha'!

Most scientists had been thinking in terms of a continuous laser, and that was certainly my thinking to start with. But why should I place such restrictions on myself? At this juncture I was only trying to demonstrate that coherent light could be made at all. Besides, a pulsed source is not only acceptable in many applications, it may even be preferable.

I went back to my analytical model, modified the equations to account for a pulsed light source, and then analyzed the results to guide me in the actual laser design. I had already determined that the most important parameter of the light source would be its brightness, that is, a lamp that maximized the power per unit area into the useful pump bands of the ruby.

My calculations also instructed me that this lamp brightness requirement is largely independent of the ruby crystal dimensions and its chromium concentration over a reasonably wide range. The model assumption was that of a three-level system (appropriate to ruby) where the lower laser-energy level is the ground state.

Next, I scoured every flashlamp catalogue that I could find. I calculated the luminous intensity for promising candidates and found that the highest intensity came from three General Electric spiral-shaped flashlamps. My calculations showed that I would have an adequate margin of safety, in terms of excess brightness intensity for the laser, of some two to three times, even with the smallest lamp.

Let there be 'Coherent' Light

Based on my most current calculations and measurements, I was becoming optimistic about the possibility of creating a laser. The tension was building and I started to have dreams that I actually did it. It was a wildly exciting thought!

Still, I was worried that I might be missing something. Instead of proceeding directly to the actual design, I decided to do an experiment that would check the validity of my model.

I made up a cube of pink ruby with each dimension equal to one centimeter in length, about the size and shape of a sugar cube. The crystal axis was perpendicular to one set of faces of the cube. I then placed the cube between two parallel plates, thus forming a microwave cavity. The cavity resonance was designed to be equal to the natural ground-state splitting of the ruby. These are the ground-state sublevels used in a ruby maser.

The purpose of this arrangement and the subsequent experiment was to monitor the ground state population of the ruby. My calculations had shown that it would be possible to substantially reduce the ruby ground state ion population necessary to make a laser, but I wanted tangible evidence to confirm and justify my optimism.

I was still harassed by the thought that no one had yet ever made a laser, although by now several teams of scientists had been diligently working away for nearly two years. Was there a fundamental insurmountable problem that I had missed?

I proceeded with the experiment using the ruby cube. As described above, one set of parallel faces of the cube served as the resonant microwave cavity. I connected one of the open (second set of) faces of the cube to one end of a polished quartz rod called a light pipe. The other end of the light pipe was connected to the flashlamp. The third set of faces of the cube was used to probe with selected wavelengths of light. I used the 'loading' on the parallel-plate microwave cavity to monitor the ground state population of the ruby, making use of the microwave properties of ruby that I had learned from my work on the ruby maser.

When I flashed the lamp, the cavity loading did indeed change. It decreased because, as expected, there were fewer chromium ions in the ground state. More important, the *magnitude* of the population change, 3%, was very close to the value I had predicted from my analysis of the experimental parameters. This was an extremely gratifying result since it gave further confirmation to my model and its analysis.

Now I was getting more excited. Since I couldn't think of anything else to check, it was time to proceed with the definitive design of a laser.

The Laser Design

My first thought was to use some modification of the elliptic-cylinder configuration that I considered for the mercury arc lamp as discussed in the previous section. The problem was, I would need a straight, rod-shaped flashlamp, but the straight lamps in the catalogues that I checked didn't have high enough intensity. Therefore, I resorted to a different design and stuck with what was available – the spiral-shaped lamps. I didn't want to get sidetracked at this time into the development of a special lamp.

It dawned on me that I would not have to devise a focusing arrangement, since the brightness of the radiation that is attainable at the focus of a mirror or lens can only approach, but not exceed, the brightness of the source. So, why not just place the crystal in close proximity to the lamp? That is, why not place the crystal *inside* the lamp helix?

To help gather the light, I placed a polished aluminum cylinder around the outside of the spiral. To fit the lamp, I used a pink ruby crystal in the shape of a right circular cylinder rod, about 1 × 2 cm. The ends of the ruby cylinder were polished flat, parallel to each other and perpendicular to the axis of the cylinder. For mirrors, I used evaporated silver.

I chose silver for the mirrors because it has the highest reflectivity of any metal at the deep red ruby laser wavelength. However, silver tarnishes quickly, and so with time, the thin output layer will change its transmission characteristics; it is not stable. To solve this problem, I had a thick layer of silver evaporated on both ends of the crystal and I scraped a tiny hole in the coating at one end. The laser beam would emanate from that coupling hole.

The Crucial Test

As I neared completion of my laser design, Bob Hellwarth, one of my colleagues in the Atomic Physics Department, asked, 'How will you know if it's working?'

At first, I was concerned that if I were stuck with a ruby crystal that deviated too much from optical perfection, these imperfections could prevent measurable laser behavior.

To get a better understanding of the processes, I elaborated on my existing analysis to account for an imperfect crystal. I found that if I were able to drive the ruby crystal reasonably far above the point of inverted population, or laser threshold, that I would see very substantial evidence of an inverted population and impressive laser behavior even with a poor crystal.

Just exactly what would I see?

I planned to monitor the red light emitted from the ruby through the hole in the output silver coating and expected to observe three kinds of laser behavior evidence: a reduction in the decay time of the fluorescent level; the bunching of the radiation into a beam; and a significant reduction of the spectral width of the red light.

The Laser Process

Laser action can be understood by following through the operational details in the above-mentioned ruby laser design. The process starts when the ruby crystal is excited by the flashlamp and chromium ions are excited into the metastable fluorescent level. The ions lose energy by radiating randomly red photons. This is the familiar red fluorescence (spontaneous radiation).

When the excitation level is great enough, that revered 'inverted population' condition is reached. In that case, more chromium ions are in the ruby's metastable upper fluorescent level than are in the ground state. Therefore, chromium ions can radiate by *stimulated* emission as well as by the normal spontaneous emission process.

Because the metastable level now empties much faster than it normally would, the fluorescent lifetime is reduced. That decreased lifetime can be observed by monitoring the red ruby glow with a photoelectric cell connected to an oscilloscope when the ruby is flashlamp-excited.

When the ruby is in the inverted population condition, as explained previously, it becomes an amplifier. The red photons are amplified as they progress through the crystal.

An important selection process starts to take place. The red photons are initially emitted in random directions. But the fluorescent photons that happen to be radiated at large angles to the mechanical axis of the ruby cylinder are lost through the sides of the crystal. On the other hand, photons radiated along the crystal axis, or at small angles with respect to the crystal axis, are in effect, trapped. They are reflected when they strike either one of the end mirrors and move back and forth through the crystal. As these axial photons move through the ruby they are amplified and consequently generate more photons in the same direction. They quickly become the dominant stimulated radiation from the excited metastable level.

Keep in mind that photons moving through the crystal are responsible for the stimulated emission. The axial photons pick up a following and march down that crystal axis. Consequently, the photons that emerge from the coupling hole in the output silver mirror are concentrated in a direction along, or nearly along, the crystal axis.

The red ruby fluorescent (spontaneous) emission extends over a distribution of frequencies in a curve that resembles a bell shape. The fluorescent photons are most concentrated at the center of that curve. When the inverted population condition exists, the top of the curve has the highest amplification. The consequence is (as with the discussion of beam angle) that photons near the center of the distribution are favored, since it is these photons that get amplified most in the back and forth transits, through the crystal from the multiple mirror reflections. It is this last process that explains why the frequency distribution, the linewidth of a laser, is so small.

I planned to vary the excitation to the flashlamp. In doing so, as explained above, I would expect to see a reduction of the fluorescent lifetime giving evidence of stimulated emission. As I varied the ruby excitation level, the fluorescent intensity should increase proportionately, as long as I was below threshold. But, when above threshold, small increases in excitation should make much bigger changes in the output since the detector and monochromator combination are more responsive to the smaller beam and narrower linewidth.

Do it!

It was the afternoon of May 16, 1960; it was time to confirm or deny all the fears of why 'ruby can't work' or why 'lasers can't be made to work.' No more new calculations, no more diversionary experiments. This was the moment of truth!

The laser head was mounted on a workbench. The flashlamp was connected to the power supply. The trigger electrode was connected to the spark coil (the mechanism that initiates the flash from the strobe lamp). The light output from the coupling hole in the end of the ruby was directed through the monochromator to a photomultiplier tube. The electrical signal from the photomultiplier was connected to an oscilloscope.

Irnee D'Haenens, my laboratory assistant, and I were the only ones performing and observing the experiment. We first took a test shot so that we could adjust the monitoring equipment. We turned the power supply up to about 500 volts. We fired the flashtube. Indeed, we observed a trace on the oscilloscope!

The trace was a recording of the red ruby fluorescence. The decay in the trace was about three milliseconds, the lifetime of the upper possible laser level. We made the appropriate adjustments to optimize the monitor display.

We continued. We progressively increased the supply voltage, each time monitoring and recording the light output trace. As we did so, the peak output increased proportionately to the energy input, the decay time remained the same. So far, so good.

But, when we got past 950 volts on the power supply, everything changed! The output trace started to shoot up in peak intensity and the initial decay time rapidly decreased.

Voilà. This was it! The laser was born!

The Light Fantastic

When Irnee and I observed the first laser go into action, Irnee was smiling and jumping up and down with glee. I was numb and emotionally drained from all the tension and excitement.

Imagine: in the 10 years prior to the laser, the coherent electromagnetic spectrum had been extended by perhaps a factor of five. Now, with the advent of the laser, there was a quantum jump in that spectrum of ten thousand!

The significance of my historic accomplishment didn't sink in right away. I'm not sure that it has yet.

THE HELIUM–NEON LASER

A Javan, Massachusetts Institute of Technology,
Cambridge, MA, USA

© 2005, A Javan. Published by Elsevier Ltd. All rights reserved.

> The date, twelve of December nineteen sixty, and the time, 4:20 PM, are the date and the time of an event that triggered such an imortant segment of what has now evolved in modern optics. Here is how:

It had snowed heavily on that day. The laboratory at Murray Hill, New Jersey, the Bell Telephone Laboratory, BTL, had been closed at 4. What occurred was a triumph of our depth of understanding of the atomic spectra, the physics of the inter-atom collisions and electron impacts and exchange effects, the fascinating sequence of what takes place in a *gas discharge plasma* when the discharge is switched on. The triumph was also that of the art in spectroscopy, this time that of the colorful glow of a gas discharge plasma in a controlled mixture of helium and neon. Performing the spectroscopy, one could *effectively* watch and follow a set of fascinating physics effects in the discharge plasma, that would lead to an *inversion* to set-in in a specific transition between a high-lying pair of energy levels in neon, by the now well-known transition at 1.15 μm wavelength in neon. Helium atoms in a metastable state in the discharge plasma played the key role.

On that day: The laser had already been de-gassed for 5 days and nights at 600 °C under the high vacuum, to drive off the impurities prior to introducing the gas mixture in the laser-plasma quartz tubing and sealing it off. We had a way to optimize the mixture, at the exact mixture ratio for the optimum inversion and the laser gain at the 1.15 μm neon transition wavelength. The laser design had a *daring* interferometer length *one meter long*. Why *daring* at *one meter long*, I will have a chance to explain why below. The laser-mirrors at each end of the one-meter-long interferometer were situated internally in the laser, exposed to the gas mixture. Two fine micrometers at each end of the laser provided the fine control for the laser–mirror alignment control externally.

To convey the scene, we had known already, that the inversion and the laser gain at the 1.15 μm neon transition will be *higher* in the discharge '*afterglow*', the afterglow following *rapidly switching-off the discharge*. The inversion and the laser gain at the 1.15 μm wavelength, we knew, would last for as long as a millisecond in the afterglow. Certain that the inversion and the gain will be higher in the afterglow, we had set-up the electronics to switch on the discharge for about 10 milliseconds, and switching it off fast at a one microsecond switch-off time, and leaving it off for the same 10 milliseconds time interval, and repeating it at a low rep-rate, about 20 times a second on that day. Expecting that while searching for the laser signal, the laser will first break into oscillation in the afterglow where the laser-gain will be higher, knowing that pulling it into the glow afterward will be easy.

With me on that day I had two of the three good people who worked with me on the experiment. One of the two, Donald Herriott, was about to miss his car pool. He had his hand at the time on the laser–mirror tuning-alignment, the fine micrometer at one end of the laser interferometer on his side. I had my back towards the oscilloscope that would display the laser signal, aligning the detection system. My journey had begun over two years previously. I was *certain* on that day, and at that time, that possibly right then, or no later than hours later or by midnight, the moment would come for me and others that I had with me, to witness the onset of the event sought after so hard. And then … .

I had my back towards Don Herriott, urging him to stay a bit longer. I heard Don's voice, saying '*What is that*?'. I turned and looked. The laser signals in the discharge afterglow on the oscilloscope screen, synchronized to the gas discharge rep rate. I reached the laser interferometer fine-tuning micrometer on my end … a gentle tuning, in a mere few seconds time, optimized the afterglow signal, and then pulled the laser signal right into the glow. I loudly called on 'Ed' to switch on to CW. He, Edward Ballick, jumped over the equipment and the electronics on the floor in the lab, reached the switching electronics, switched it on to CW:

> The laser signal on the oscilloscope … switched to CW … a coherent light beam … CW … continuous … and at a degree of coherence … that as it proved later and I know then … to the limits that the laws of nature will permit. I looked at my watch, 4:20 PM.

The year 1960 was the beginning. Some months earlier, in the summer of that year, Theodore Maiman at Hughes Laboratories, in Malibu, CA, succeeded to operate an optically pumped laser, operating in short-duration laser pulses, the Ruby laser at 0.694 μm. That approach to the laser differed fundamentally from the discharge plasma approach. The two were different inventions and served different purposes. One, the *optically pumped*, converts the *optical energy* in the pump, to the laser light output. The other, *electrically pumped,* converts the *electrical energy* in the *electrical current* that excites the laser plasma, to the laser light output. The rest on how things evolved is history.

The Scene Before Lasers

Before lasers, optics had long passed its peaks. The last words had been written by Born and Wolf in 1957, in an impressive volume on Physical Optics. When I went through my graduate studies at Columbia University in the early fifties, there were no graduate courses offered in optics. The subject was dealt with incidentally in courses in electricity and magnetism. Action was all on the microwaves at the time. Optical spectroscopy had reached its peaks much earlier yet, before World War II. Condon and Shortly had written the last words on Theory of Atomic Spectra in 1935, amazingly, as it seems, in *hardly a decade* after the discoveries in quantum mechanics. As events had it, optics gave way to the microwaves after the war, and optical spectroscopy to the RF and microwave spectroscopy – and all this, a result of the discoveries in Radar at the microwave frequencies in World War II.

We owed everything we knew of the atomic spectra to the *gas discharges* used as spectroscopic light sources at the time, from the high-resolution spectra of the hydrogen atoms at one end of the periodic table, to that of the trans-uranium elements on the other. We need not be reminded, everyone knows, how much we owed then and do now, to the high-resolution spectra of the hydrogen atoms, all because of the early hydrogen discharge light sources and then at the beginning, what led to the discovery of the quantum mechanics, and the wave nature of the atoms. The excitement that followed has remained unparalleled in physics since. Within hardly a decade after as a result, the cleverly designed spectroscopic gas-discharge sources, of different makes and designs, exhaustively revealed every feature predicted by the quantum mechanics in the high-resolution spectra of every element we know in nature. Gas discharge light sources set the course at the time, as did the lasers after the events in 1960.

Optics had given way to the microwaves in the fifties after the war. The physics of the gas discharges, however, continued as an active research field after the war, although of short duration. Bell Telephone Laboratory had a most distinguished research department in gas discharges in the early fifties. Towards the late fifties, as it seemed, the gas discharges as a field in research, also seemed to have been exhausted. The payoff at the time had evolved in the colorful neon light sources and fluorescence lamps, a thriving industry all over the country.

In the summer of 1958, Bell Telephone Laboratory had dissolved their research department in gas discharges. They had transferred every one of their distinguished research people to the other research areas in the laboratory. Some good people had left earlier, already.

As the fate in gas discharge physics had it, I joined the Laboratory in September 1958 – with ZERO backgrounds in gas discharge physics. I knew physics.

How Things Evolved

It has been said that when time is ripe for a new idea, if one person misses it, the next one will not. Time was ripe in the late fifties for the optically pumped idea for the laser, but far from it for the *electrically pumped*. Optical pumping was an active field in research at the time. The French had introduced the ‘pompage optic’, the optical pumping. With the French pompage optic, one would by optical pumping, resonantly excite an atomic species from its ground into its excited electronic state, and perform a double-resonance at an RF frequency on a fine or hyperfine structure of the excited electronic state. There were other versions of the French pompage optic. Each version used the optical pumping to induce resonantly, an atomic species into an excited electronic state.

Important, and more so, in the fifties, in the second half, the discoveries of the three-level masers at the microwave frequencies had attracted much attention. The three-level masers are the counterparts at the microwaves of our now optically pumped lasers. In three-level masers, one drives to saturation a pair of energy levels at microwave energy-spacing in a paramagnetic solid, and allows the inversion and the gain to set-in in a transition to a low-lying third energy level at the microwave frequencies. Nicholas Blumbergen at Harvard had proposed the three-level maser idea in an important publication in 1956. I had my own original three-level maser ideas, and had them published in 1957. In my work I had discovered a fascinating new effect, and for the first time, that in the transition to a third level in a three-level maser, it will be possible to drive the maser into its self-oscillation, *without* requiring the inversion. The effect had led me to the idea of the Raman masers, requiring a ‘*pumping*’ microwave field, followed by a Raman transition across a pair of energy levels at microwave energy-spacing. In my publication of the effect, I had also considered the pumping field at an optical frequency, instead of at the microwave.

The ideas were 'in the air' at the time, to extend the three-level maser approach into the optical frequencies. Yet, in the summer of 1958, on an occasion in June of that year, I think on the first Monday in June, *flashed* through my mind a far different approach, all because of the *helium metastables*. Ideas always come to mind in a *flash*, the rest follows.

Summer of 1958: A Bit of History

I think it was on the first Monday in June 1958. I had been on an occasion visiting the Bell Telephone Laboratory (BTL) on that day. They had approached me that I join their research staff at Bell. My work on masers – I had done as a part of my post-doctoral research position in physics at Columbia University in New York City – had in part overlapped the then ongoing work at Bell. They knew me, and I knew Bell.

At the visit meeting with Arthur Schawlow, whom I knew well, he gave me his news on that day, that he and Charles Townes, had submitted to the *Physical Review* a paper for publication proposing an optically pumped counterpart of a three-level maser at the optical frequencies – by-now, the well-known Schawlow–Townes optically pumped laser publication. Specially, I was attracted to Schawlow's idea on a two-mirror Perot–Fabry-type interferometer as a resonator at the optical frequencies for the laser. The use of it in that publication was Schawlow's idea.

My thoughts, it may seem paradoxical, right there and then moved far away from an optically pumped laser as a possibility. The helium atoms in a *metastable state* flashed through my mind. I will avoid explaining why or how, other than, of a three-level atom, two of the levels were now to belong to a *different* atomic species, the helium atom, and that there will be no optical pumping … and so much '*electrical energy*' could be converted to and stored in the energetic metastable helium. A new possibility presented itself, with no resemblance to the optical pumping – an energy-exchange effect taking place resonantly, between the highly energetic helium metastables, and – as it evolved – another inert gas, the neon atoms in a helium–neon gas mixture. The resonant nature of the effect embodied everything we know of the wave nature of the atoms. I will have a chance to explain the effect shortly below.

Driving home at the end of the day on that day, a two hour drive from Murray Hill, New Jersey to uptown New York City, the thought of the metastable helium had crystallized before me the task to examine the effect in a helium-plasma in a gas discharge, in a gas mixture with the other *four* well-behaved *inert gases*, from neon through xenon. There are four.

Within weeks in June, searching through the literature I learned, that over the years, and mostly in the early fifties, the helium plasmas in a gas discharge in mixtures with *every other* inert gases had been the subject of most extensive research, *except in a mixture with neon*. The mixture with neon, for the reason that the reader will find below, had been thought of as *uninteresting*. The matter was left for me.

I had accepted in June already, to join the Bell research staff at Murray-Hill, and I did on September. From my work that entire summer, unfolded the entire theory for the helium–neon gas laser. There was also another far-reaching offshoot.

The Physics Effect

Next to the hydrogen atom, the helium is the most fundamental of all atoms in physics. Hydrogen has one electron. Its spectrum, detached from the complexities of the multi-electron atoms, carries the entire signature of the wave nature of the atom, in ways that are subject to accurate predictions and the measurements. The helium atom has two electrons. Its spectrum carries the entire signature of the physics effect of the entire multi-electron atoms, also in ways that are subject to accurate predictions and the measurements. In its metastable states, it carries special *dynamic features*, deeply fundamental, and special *only* to the helium metastables.

In a gas discharge in pure helium at a moderate gas pressure, the discharge color takes on a grayish-white with a tint of blue. Over 500 excited electronic states in helium have been identified – their excitation by the energetic electrons in the discharge, followed by their spontaneous radiative decays, the emissions of photons, accounts for the high-resolution spectra of the discharge grayish-white with a tint of blue in color. Of these 500, all of them, high lying in energy, are short-lived radiatively, with the exception, however, of two amongst them. The two have the lowest excitation energy. Their radiative decays to the ground electronic state are deeply forbidden; they cannot decay radiatively. The two form the well-known helium metastables, identified by their spectroscopic designations: $^3S_{1/2}$, the triplet S one half, and $^1S_0$, the singlet S zero – generally referred to in the literature as, the triplet S and the singlet S metastables. Of the two, the triplet S metastable has the lowest excitation energy. It lies at 19.8 electron volts (eV), above the helium ground state. The singlet S metastable lies higher, at 20.2 eV.

Helium atoms in their ground electronic state are chemically inert. A helium atom's two electrons form a closed shell. It cannot enter in a chemical reaction, neither with its own species, nor with any other, as we all know. The situation is, however, far different with every one of the 500 excited electronic states in a helium gas discharge. They are energetically high-reactive, except, and with the exception of the two metastables, they are all *short-lived*, and cannot play an important role.

The effect manifests itself differently in the long-lived metastables, and accounts for their *special dynamic features,* special only to the helium metastables. What follows outlines the effect.

Dynamic Features: Why so Special

By far the majority of the atoms in the helium gas discharge are in their inert, ground electronic states. Once in its energetic metastable state, a helium metastable in the gas discharge encounters in an impact – in a collision – with a helium atom in its ground electronic state. The encounter in the impact is of a short duration in our scale of time, but long in the atomic time-scale, of about 2×10^{-13} second, a fifth of a picosecond. What occurs is strictly wave-mechanical, dictated by the wave nature of the electrons in the atoms. During the encounter, the electron wave functions in the encountering atoms overlap momentarily. A momentary binding occurs. In no ways can one tell during the encounter which electron is which: Electrons are the properties of their wave functions – once overlapped, individual electrons lose their identity, in no way can one tell which is which. As the atoms fly apart, caused by the kinematics of the impact, an *electron exchange* occurs, *wave-mechanically.* The helium atom initially in its ground electronic state, will fly away from the encounter with one of its two electrons now in the *highly excited* energetic *metastable* state. The other, the one initially in the excited metastable state, emerges from the encounter in the *unexcited* ground electronic state.

At a gas pressure in the discharge of about one torr, a millimeter, the mean free-time between two impacts is about one microsecond, a very long time in the scale of the atoms. The helium atom that flew away from the encounter in the metastable state remains in its now metastable state, during the entire mean-free collision time. Once it encounters the next helium atom in its ground electronic state, the act repeats again, a momentary binding occurs, an electron exchange takes place, and a new metastable flying away. The effect repeats over and over again, at about every microsecond. A *diffusion* of the metastability through the helium in the discharge sets in. Depending on the diameter of the discharge tubing, it takes about a millisecond until the helium that emerges from the encounter in the metastable state reaches the walls of the discharge tubing. The act will then *terminate*, in a *most violent* collision effect with the wall. The energetic metastable helium *knocks out* an electron off the wall, as itself emerges from the collision in its ground electronic state – all intact.

With one exception, the encounter between a helium metastable with any species other than of its own kind, a helium, is energetically *highly violent.* The exception is the neon atoms. In encounters with all atomic species other than neon, the metastable helium *knocks out* an electron off the species, ionizes the species, and itself emerges from the encounter in its ground electronic state, intact. The ionizing impacts had attracted so much attentions in the early fifties, the effect in the helium gas discharge in mixtures specially with every other inert gases, had been exhaustively investigated, except with neon. Neon has its ionization potential higher than the two metastable excitation energies. It can not undergo an ionizing impact with either of the two metastables. The mixture with neon had thus been considered uninteresting, and was discarded.

What emerged by the end of the summer of 1958 was the entire theory for the helium neon laser – and the recognition of a deeply wave-mechanical effect, the resonant energy exchange taking place between the energetic helium metastable and an inert neon atom in the discharge gas mixture. In the mixture, a momentary binding occurs between the helium metastable encountering an inert neon atom in the gas discharge. Wave-function overlaps occur for the electrons in the encountering atoms. A *resonant* energy exchange takes place during the encounter. Helium metastable will fly away from the impact in its now ground electronic state, while neon emerges in one of its high-lying excited energy states, at the near-resonant with the helium metastable.

A special effect sets in in the energy exchange, as well. The kinetics of the impact in the encounter enters in satisfying the exact resonance in the energy exchange. The singlet S metastable flying away from the encounter after the energy exchange, will be slightly 'cooler' kinetically, and the triplet S flying away from the encounter after the energy-exchange, will be slightly 'hotter'. The kinetic energy of the impact enters in the energy exchange.

Optical Gain and Amplification Effect

Within weeks, early in the summer of 1958, investigating the energy transfer effect led to the prediction that there can be inversion and optical gain in two sets of transitions in neon, one as a result of the excitation transfer from the triplet S metastables, and the other a result of the transfer from the singlet S. By far the more challenging task that summer, however, was to estimate theoretically the *values* for the *gain and amplification factors* obtainable for the predicted inverted transitions, and more specifically, for the inverted transitions that will have the highest gain and amplification factors. There followed the theoretical estimates on which relied the entire experimental phase of the work, the planning of it commenced within weeks after my joining the Bell Telephone Laboratory on September of that year.

The set of transitions originating from the triplet S energy transfer to neon, occur in the near-red region of the spectrum, the 1 μm wavelength region. The strongest inverted transition at the highest predicted gain value, lies at 1.15 μm transition wavelength, the spectral laser line that first broke into self-oscillation (see the opening page in this article). The set of transitions originating from the singlet S energy transfer to neon occur in the red-yellow region of the spectrum. The strongest, although at less gain-value than 1.15 μm, occurs at the 0.633 μm in the red.

In the early phase of the investigation, it came to light also that the achievement of the inverted transitions with optical gains, was, interestingly, a *commune occurrence* in gaseous discharges in a variety of the atomic species, including in pure neon. However, the complexities in the excitations by the hot electrons of the high-lying energy levels, and in particular in the presence of the secondary excitation mechanism, and a list of other similar effects, in no way would allow a quantitative estimate of the gain and amplification factors in the inverted transitions obtainable in the gas discharge.

The reliance on the excitations of the high-lying energy levels in neon via the resonant energy transfer from the helium metastables, as per the physics effect I have outlined above, placed at my disposal the one gas discharge system that lent itself to the theoretical estimates of the predicted optical gains and the amplification factor for the inverted transitions.

Without a knowledge of the predicted values for the optical gain and the amplification factor, it would have been in no way possible to perform the experiment.

Experimental Phase

The Bell Telephone Laboratory, considered as a foremost research institution in the country at the time, operated at a $250 million yearly budget – in 1960s' dollars. Transistors had been invented at Bell in the forties. With the awareness at the time at Bell, that a light beam at the degree of frequency-purity and the coherence expected of the laser, will have far-reaching consequences on the telecommunication technology, without hesitation I received from Bell an open ticket to do what it takes – one key person in the administration understood in-depth my approach to the laser – via the gas discharge and the physics effects I had invoked. Within weeks following joining Bell in September, I had my planning underway for what came to be known in a short while later, as a most aggressive research in gas discharges – and at a time when Bell had already, in the summer of 1958, dismantled the entire research activities in gas discharges at Bell. That speaks a great deal for Bell at the time … at the time before the AT&T break up.

In lasers, the knowledge of the full length of the amplifying medium across which the optical gain obtainable at the laser transition will suffice to drive the laser into its self-oscillation, will dictate the full length of the laser interferometer resonator. The spacing between the two high-reflectance interferometer end-mirrors will have to accommodate the full length of the amplifying medium, thus dictating the interferometer full length. With respect to the mirror reflectivity obtainable at the time, one had to be content with the values at the highest close to 98%. The highest gain transition from my theoretical estimates, that of the neon transition at the 1.15 μm wavelength in the near-red, indicated an interferometer length of no less than 50 centimeters, or possibly as long as one meter – and there lay the challenge, achieving the required degree of parallelism needed for the interferometer end-mirrors at such spacing. One could show, from simple inspection, that the parallelism needed for the mirrors had to be at values better than one part in several-hundredths of a wavelength. This value applied to the use of the flat interferometer mirrors known to me at that time.

An alignment to such a degree of accuracy could well be obtainable at the time for Perot–Fabry-type interferometer mirrors at the mirror-spacing not exceeding about eight or ten centimeters at the most – where one would rely on the observation of the interference fringes using the available spectroscopic light sources to observe the fringes, a principle well known in physical optics. At longer mirror spacing, the spectroscopic light sources lacked the degree of coherence, the color purity, needed to observe fringes for the mirror alignment. The fringes will get wiped out.

These days, using a helium–neon laser light source available at every laboratory in optics, one can align a two-mirror interferometer at tens of meters mirror-spacing, or hundreds, or even as far apart as several miles in some experiments. Ideas are floating around now of the possibility, using gas lasers, to align mirrors placed on satellites in orbit, with the mirror on two different satellites at controlled spacing as far apart as thousands of miles – the experiment thought of is that of the detection of the gravity waves. Without a laser operating and available at the time, in no ways would it have been possible to align an interferometer at a *daring* mirror-spacing as far apart as one meter, or even half as long. One had to rely on an auto-collimator, an optical instrument that at the best would provide an alignment to within one or two wavelengths, nowhere sufficient to drive the laser into its self oscillation.

The challenge mapped the entire course for the experimental phase, foremost aimed at showing every feature of the predicted effects experimentally, from the energy transfer effect, to observing and measuring the gain and amplification effect in the predicted inverted transitions in neon. The feat was that of the art in spectroscopy. The observation of gain and amplification effect was a highlight in the course of the experiment. Up to that point, in optical spectroscopy, one always measured the spectral lines in either emission or in absorption. The feat was the first in spectroscopy, observing the spectral line of an atomic species not in absorption, but in amplification – the amplification of an external light source transmitted through the gas discharge in the experiment. The feat, yet, lay in the art in spectroscopy, throughout the experiments.

The event at 4:20 PM in December two years later, speaks for itself.

> It may seem now, as if one needed an operating helium–neon laser at the time, to align the laser mirrors for the first operating helium–neon laser. It seemed evident from the beginning, that the presence in the very first helium–neon laser, of the optical-gain at the predicted value in the predicted inverted transition, will guide achieving the alignment of its own mirrors, to the exact optical degree of the alignment accuracy needed for the self-oscillation. Driving the laser into its self-oscillation signaled the feat. The feat, however, was that of the power in spectroscopy, making it possible to predetermine the presence in the first helium–neon laser, of the optical gain, at the predicted optimum value needed for self oscillation.
>
> There were doubters and doubters, in the technical staff and the administration at Bell. Donald Herriott, William Bennett and Edward Ballick collaborated with me in the experiment.

The Impact

Within months of the publication of the work in January 1961 in the *Physical Review Letters*, walking through the long hallways at BTL where the laboratories were situated, one could see through the open doors in many laboratories down the long halways, the pink colors of long helium–neon gas discharges plasma tubing, with people in the labs searching for new laser transitions. BTL had revived the gas discharge research activities they had dismantled in the summer of 1958. Not surprisingly, the second set of new laser transitions driven into self-oscillation took place at Bell in 1962, that of the set of transitions in neon originating from the energy transfer from the helium singlet S metastable – with the most intense in the series, that of the by now well-known 0.6328 μm neon transition in the red. There followed the discoveries of the argon–ion gas laser and shortly thereafter the CO_2 gas laser, the two amongst the foremost in the many new gas lasers driven to self-oscillation at Bell.

Every gas laser discovered since the December 12 event, has enjoyed the use of a '*no-longer*' a '*daring*' '*long*' interferometer resonator, merely because one would use a He–Ne laser to *pre-align* the laser mirrors to within a small fraction of the wavelength needed, prior to searching for a new laser line in the gas discharge chosen for the search. The already aligned resonator removed the need to *predetermine* the presence of the gain and the amplification effect at the values needed for self-oscillation. Within a span of about a decade, virtually hundreds of different gases were driven to self-oscillation with this method, operating literally at thousands of different wavelengths, ranging from the far infrared, the 100 100ηn region, to the near-UV.

The helium–neon laser triggered so many of the key discoveries that followed. Within months, in the spring of 1961, one of the distinguished engineers at BTL, Rudolf Komphner, the discoverer of the backward tube oscillators at the microwave frequencies, began experimenting with the transmission through the improvised thin fiberoptic, of the output beam of a helium–neon laser at the 1.15 μm transition in the near-red, the near-IR, having his eyes on the potential someday in communication on the use of a continuously operating laser light beam at the near-IR frequencies. I recollect the very first ad that appeared on lasers nationally was a BTL ad in the spring of 1961, showing a cross-section of a beam of light, underneath it saying: '*Someday a Beam of Light This Size Will Carry Millions of Telephone Conversations Simultaneously*'. The ad speaks for itself.

In every gas laser, the active laser medium – being a gas – is highly homogeneous. As a result, with a reasonably good optics, one can obtain a nearly perfect Gaussian output beam. For this reason, the helium–neon lasers available commercially since the mid-sixties, have served as the indispensable tools for the alignment of the complex optics in every research laboratory in lasers, or in every manufacturer of the lasers in the country, or the manufacturers of the precise optical components, or the medical uses and so on.

In the applied areas, the helium–neon laser triggered so many of the key experiments in the early days. Philips at Eindhoven can be considered as the first discoverer of a video disc recorder. The model used a helium–neon red laser, at a most impressive fidelity and color resolution, in recording and in reading for the display. I recollect, in an exhibit shown to me at a Philips Lab, one groove on the recorder was that of a horse in a prairie, and the next groove, one '*micrometer away*' on the recorder, was that of a crowd in a park – so impressive at the time, with no cross talks. They had it packaged in a compact unit already. In other areas, all bar-code readings were done using a helium–neon laser, and some still do.

In the scientific area, from high-resolution spectroscopy at unprecedented accuracy, to the accurate clocks operating at optical frequencies, or in the measurements of the absolute frequencies of the accurate laser clocks, or the early experiments in relativity and so on, all are owed to the precision gas lasers. Helium–neon laser has played the key role, in the early days throughout, and continues today.

A note to the reader

In the early publications on laser, the device was generally known and referred to at the time as the 'Optical Maser'. The word laser became an accepted designation for the device in the late 1960s and the early 1970s.

Suggested Readings

Javan A (1961) In: Singer J (ed.) *Advances in Quantum Electronics*, pp. 18. Columbia University Press.

Javan A, Bennett WR, Jr. and Herriott DR (1961) Publications reporting operation of the first helium–neon laser. *Physical Review Letters* 6: 106.

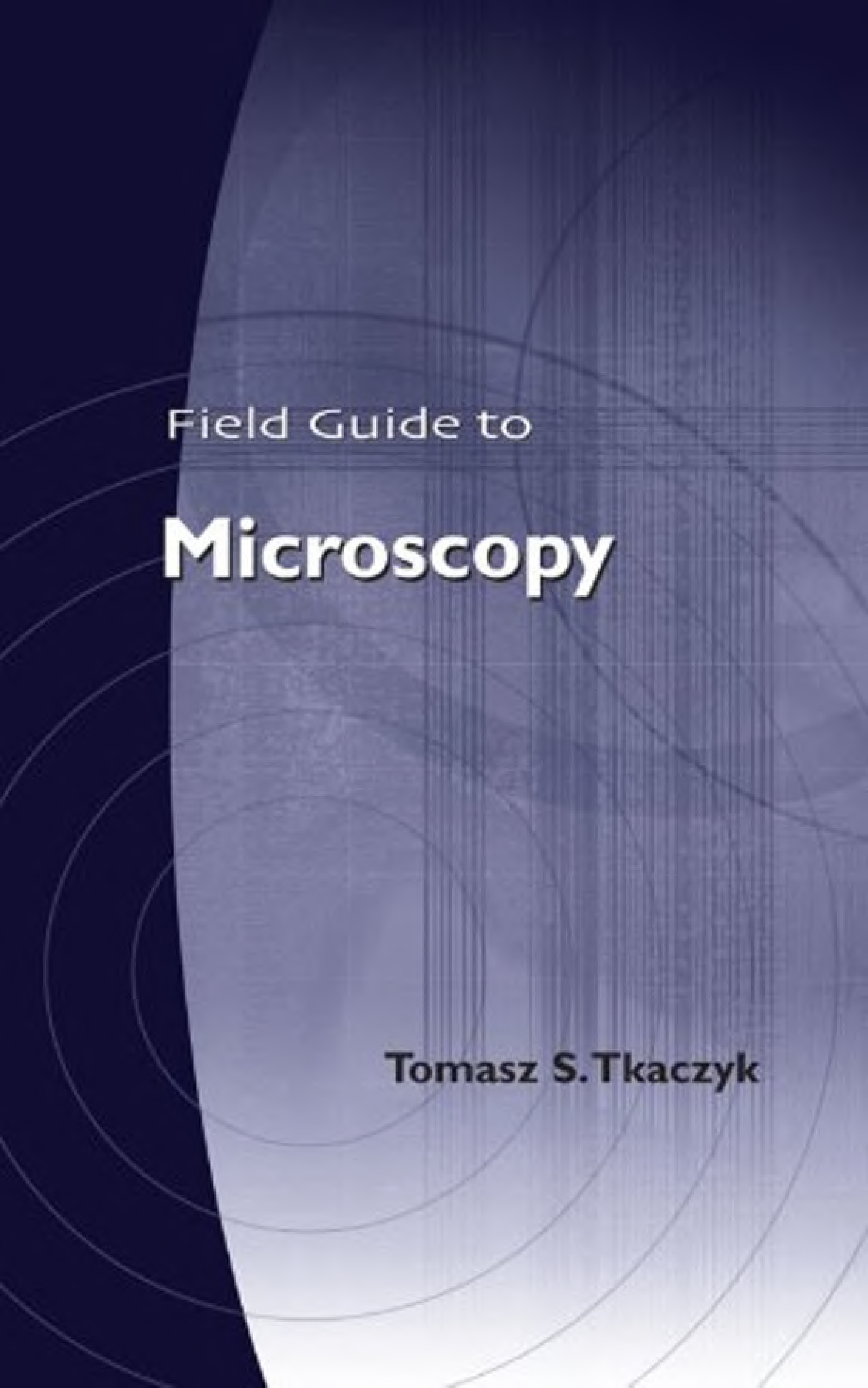

Field Guide to

Microscopy

Tomasz S. Tkaczyk

Field Guide to
Microscopy

Tomasz S. Tkaczyk

SPIE Field Guides
Volume FG13

John E. Greivenkamp, Series Editor

SPIE
PRESS
Bellingham, Washington USA

Library of Congress Cataloging-in-Publication Data

Tkaczyk, Tomasz S.
Field guide to microscopy / Tomasz S. Tkaczyk.
p. cm. -- (The field guide series)
Includes bibliographical references and index.
ISBN 978-0-8194-7246-5
1. Microscopy--Handbooks, manuals, etc. I. Title.
QH205.2.T53 2009
502.8'2--dc22
2009049648

Published by

SPIE
P.O. Box 10
Bellingham, Washington 98227-0010 USA
Phone: +1 360.676.3290
Fax: +1 360.647.1445
Email: spie@spie.org
Web: http://spie.org

Copyright © 2010 The Society of Photo-Optical Instrumentation Engineers (SPIE)

All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means without written permission of the publisher.

The content of this book reflects the work and thought of the author. Every effort has been made to publish reliable and accurate information herein, but the publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Printed in the United States of America.

Introduction to the Series

Welcome to the *SPIE Field Guides*—a series of publications written directly for the practicing engineer or scientist. Many textbooks and professional reference books cover optical principles and techniques in depth. The aim of the *SPIE Field Guides* is to distill this information, providing readers with a handy desk or briefcase reference that provides basic, essential information about optical principles, techniques, or phenomena, including definitions and descriptions, key equations, illustrations, application examples, design considerations, and additional resources. A significant effort will be made to provide a consistent notation and style between volumes in the series.

Each *SPIE Field Guide* addresses a major field of optical science and technology. The concept of these *Field Guides* is a format-intensive presentation based on figures and equations supplemented by concise explanations. In most cases, this modular approach places a single topic on a page, and provides full coverage of that topic on that page. Highlights, insights, and rules of thumb are displayed in sidebars to the main text. The appendices at the end of each *Field Guide* provide additional information such as related material outside the main scope of the volume, key mathematical relationships, and alternative methods. While complete in their coverage, the concise presentation may not be appropriate for those new to the field.

The *SPIE Field Guides* are intended to be living documents. The modular page-based presentation format allows them to be easily updated and expanded. We are interested in your suggestions for new *Field Guide* topics as well as what material should be added to an individual volume to make these *Field Guides* more useful to you. Please contact us at **fieldguides@SPIE.org.**

John E. Greivenkamp, *Series Editor*
College of Optical Sciences
The University of Arizona

The Field Guide Series

Keep information at your fingertips with all of the titles in the *Field Guide* series:

Field Guide to Geometrical Optics, John E. Greivenkamp (FG01)

Field Guide to Atmospheric Optics, Larry C. Andrews (FG02)

Field Guide to Adaptive Optics, Robert K. Tyson & Benjamin W. Frazier (FG03)

Field Guide to Visual and Ophthalmic Optics, Jim Schwiegerling (FG04)

Field Guide to Polarization, Edward Collett (FG05)

Field Guide to Optical Lithography, Chris A. Mack (FG06)

Field Guide to Optical Thin Films, Ronald R. Willey (FG07)

Field Guide to Spectroscopy, David W. Ball (FG08)

Field Guide to Infrared Systems, Arnold Daniels (FG09)

Field Guide to Interferometric Optical Testing, Eric P. Goodwin & James C. Wyant (FG10)

Field Guide to Illumination, Angelo V. Arecchi; Tahar Messadi; R. John Koshel (FG11)

Field Guide to Lasers, Rüdiger Paschotta (FG12)

Field Guide to Microscopy, Tomasz Tkaczyk (FG13)

Field Guide to Laser Pulse Generation, Rüdiger Paschotta (FG14)

Field Guide to Infrared Systems, Detectors, and FPAs, Second Edition, Arnold Daniels (FG15)

Field Guide to Optical Fiber Technology, Rüdiger Paschotta (FG16)

Preface to the Field Guide to Microscopy

In the 17th century Robert Hooke developed a compound microscope, launching a wonderful journey. The impact of his invention was immediate; in the same century microscopy gave name to "cells" and imaged living bacteria. Since then microscopy has been the witness and subject of numerous scientific discoveries, serving as a constant companion in humans' quest to understand life and the world at the small end of the universe's scale.

Microscopy is one of the most exciting fields in optics, as its variety applies principles of interference, diffraction, and polarization. It persists in pushing the boundaries of imaging limits. For example, life sciences in need of nanometer resolution recently broke the diffraction limit. These new super-resolution techniques helped name microscopy the method of the year by *Nature Methods* in 2008.

Microscopy will critically change over the next few decades. Historically, microscopy was designed for visual imaging; however, enormous recent progress (in detectors, light sources, actuators, etc.) allows the easing of visual constrains, providing new opportunities. I am excited to witness microscopy's path toward both integrated, digital systems and nanoscopy.

This Field Guide has three major aims: (1) to give a brief overview of concepts used in microscopy; (2) to present major microscopy principles and implementations; and (3) to point to some recent microscopy trends. While many presented topics deserve a much broader description, the hope is that this Field Guide will be a useful reference in everyday microscopy work and a starting point for further study.

I would like to express my special thanks to my colleague here at Rice University, Mark Pierce, for his crucial advice throughout the writing process and his tremendous help in acquiring microscopy images.

This Field Guide is dedicated to my family: my wife, Dorota, and my daughters, Antonina and Karolina.

Tomasz Tkaczyk
Rice University

Table of Contents

Glossary of Symbols **xi**

Basics Concepts **1**
Nature of Light 1
The Spectrum of Microscopy 2
Wave Equations 3
Wavefront Propagation 4
Optical Path Length (OPL) 5
Laws of Reflection and Refraction 6
Total Internal Reflection 7
Evanescent Wave in Total Internal Reflection 8
Propagation of Light in Anisotropic Media 9
Polarization of Light and Polarization States 10
Coherence and Monochromatic Light 11
Interference 12
Contrast vs Spatial and Temporal Coherence 13
Contrast of Fringes (Polarization and Amplitude Ratio) 15
Multiple Wave Interference 16
Interferometers 17
Diffraction 18
Diffraction Grating 19
Useful Definitions from Geometrical Optics 21
Image Formation 22
Magnification 23
Stops and Rays in an Optical System 24
Aberrations 25
Chromatic Aberrations 26
Spherical Aberration and Coma 27
Astigmatism, Field Curvature, and Distortion 28
Performance Metrics 29

Microscope Construction **31**
The Compound Microscope 31
The Eye 32
Upright and Inverted Microscopes 33
The Finite Tube Length Microscope 34

Infinity-Corrected Systems 35
Telecentricity of a Microscope 36
Magnification of a Microscope 37
Numerical Aperture 38
Resolution Limit 39
Useful Magnification 40
Depth of Field and Depth of Focus 41
Magnification and Frequency vs Depth of Field 42
Köhler Illumination 43
Alignment of Köhler Illumination 45
Critical Illumination 46
Stereo Microscopes 47
Eyepieces 48
Nomenclature and Marking of Objectives 50
Objective Designs 51
Special Objectives and Features 53
Special Lens Components 55
Cover Glass and Immersion 56
Common Light Sources for Microscopy 58
LED Light Sources 59
Filters 60
Polarizers and Polarization Prisms 61

Specialized Techniques 63
Amplitude and Phase Objects 63
The Selection of a Microscopy Technique 64
Image Comparison 65
Phase Contrast 66
Visibility in Phase Contrast 69
The Phase Contrast Microscope 70
Characteristic Features of Phase Contrast 71
Amplitude Contrast 72
Oblique Illumination 73
Modulation Contrast 74
Hoffman Contrast 75
Dark Field Microscopy 76
Optical Staining: Rheinberg Illumination 77

Optical Staining: Dispersion Staining 78
Shearing Interferometry: The Basis for DIC 79
DIC Microscope Design 80
Appearance of DIC Images 81
Reflectance DIC 82
Polarization Microscopy 83
Images Obtained with Polarization Microscopes 84
Compensators 85
Confocal Microscopy 86
Scanning Approaches 87
Images from a Confocal Microscope 89
Fluorescence 90
Configuration of a Fluorescence Microscope 91
Images from Fluorescence Microscopy 93
Properties of Fluorophores 94
Single vs Multi-Photon Excitation 95
Light Sources for Scanning Microscopy 96
Practical Considerations in LSM 97
Interference Microscopy 98
Optical Coherence Tomography/Microscopy 99
Optical Profiling Techniques 100
Optical Profilometry: System Design 101
Phase-Shifting Algorithms 102

Resolution Enhancement Techniques 103
Structured Illumination: Axial Sectioning 103
Structured Illumination: Resolution Enhancement 104
TIRF Microscopy 105
Solid Immersion 106
Stimulated Emission Depletion 107
STORM 108
4Pi Microscopy 109
The Limits of Light Microscopy 110

Other Special Techniques 111
Raman and CARS Microscopy 111
SPIM 112

Array Microscopy 113

Digital Microscopy and CCD Detectors 114

Digital Microscopy 114

Principles of CCD Operation 115

CCD Architectures 116

CCD Noise 118

Signal-to-Noise Ratio and the Digitization of CCD 119

CCD Sampling 120

Equation Summary 122

Bibliography 128

Index 133

Glossary of Symbols

$a(x, y)$, $b(x, y)$	Background and fringe amplitude
AO	Vector of light passing through amplitude Object
A_R, A_S	Amplitudes of reference and sample beams
b	Fringe period
c	Velocity of light
C	Contrast, visibility
C_ac	Visibility of amplitude contrast
C_ph	Visibility of phase contrast
$C_\mathrm{ph\text{-}min}$	Minimum detectable visibility of phase contrast
d	Airy disk dimension
d	Decay distance of evanescent wave
d	Diameter of the diffractive object
d	Grating constant
d	Resolution limit
D	Diopters
D	Number of pixels in the x and y directions (N_x and N_y, respectively)
DA	Vector of light diffracted at the amplitude object
DOF	Depth of focus
DP	Vector of light diffracted at phase object
$\mathrm{D}_\mathrm{pinhole}$	Pinhole diameter
d_{xy}, d_z	Spatial and axial resolution of confocal microscope
E	Electric field
E	Energy gap
E_x and E_y	Components of electric field
F	Coefficient of finesse
F	Fluorescent emission
f	Focal length
f_C, f_F	Focal length for lines C and F
f_e	Effective focal length
FOV	Field of view
FWHM	Full width at half maximum
h	Planck's constant
h, h'	Object and image height
H	Magnetic field
I	Intensity of light
i, j	Pixel coordinates
I_o	Intensity of incident light

Glossary of Symbols (cont.)

I_{max}, I_{min}	Maximum and minimum intensity in the image
I, I_{t}, I_{r}	Irradiances of incident, transmitted, and reflected light
I_1, I_2, I_3 …	Intensity of successive images
I_0, $I_{2\pi/3}$, $I_{4\pi/3}$	Intensities for the image point and three consecutive grid positions
k	Number of events
k	Wave number
L	Distance
l_{c}	Coherence length
m	Diffraction order
M	Magnification
M_{min}	Minimum microscope magnification
MP	Magnifying power
MTF	Modulation transfer function
M_{u}	Angular magnification
n	Refractive index of the dielectric media
n	Step number
N	Expected value
N	Intensity decrease coefficient
N	Total number of grating lines
n_{a}	Probability of two-photon excitation
NA	Numerical aperture
n_{e}	Refractive index for propagation velocity of extraordinary wave
n_{m}, n_{r}	Refractive indices of the media surrounding the phase ring and the ring itself
n_{o}	Refractive index for propagation velocity of ordinary wave
n_1	Refractive index of media 1
n_2	Refractive index of media 2
o, e	Ordinary and extraordinary beams
OD	Optical density
OPD	Optical path difference
OPL	Optical path length
OTF	Optical transfer function
OTL	Optical tube length
P	Probability
P_{avg}	Average power
PO	Vector of light passing through phase object

Glossary of Symbols (cont.)

PSF	Point spread function
Q	Fluorophore quantum yield
r	Radius
r	Reflection coefficients
r, m, and o	Relative, media, or vacuum
r_{AS}	Radius of aperture stop
r_{PR}	Radius of phase ring
r_{λ}	Radius of filter for wavelength λ
s, s'	Shear between wavefront object and image space
S	Pinhole/slit separation
$s_{\perp}$	Lateral shift in TIR perpendicular component of electromagnetic vector
$s_{\parallel}$	Lateral shift in TIR for parallel component of electromagnetic vector
S_{FD}	Factor depending on specific Fraunhofer approximation
SM	Vector of light passing through surrounding media
SNR	Signal-to-noise ratio
SR	Strehl ratio
t	Lens separation
t	Thickness
t	Time
t	Transmission coefficient
T	Time required to pass the distance between wave oscillations
T	Throughput of a confocal microscope
t_{c}	Coherence time
u, u'	Object, image aperture angle
V_{d}, V_{e}	Abbe number as defined for lines d, F, C or e, F′, C′
V_{m}	Velocity of light in media
w	Width of the slit
x, y	Coordinates
z	Distance along the direction of propagation
z	Fraunhofer diffraction distance
z, z'	Object and image distances
z_{m}, z_{o}	Imaged sample depth, length of reference path

Glossary of Symbols (cont.)

α	Angle between vectors of interfering waves
α	Birefringent prism angle
α	Grating incidence angle in plane perpendicular to grating plane
α_s	Visual stereo resolving power
β	Grating diffraction angle in plane perpendicular to grating plane
γ	Angle of fringe localization plane
γ	Convergence angle of a stereo microscope
γ	Incidence / diffraction angle from the plane perpendicular to grating plane
Γ	Retardation
δ	Birefringence
δ	Excitation cross section of dye
δz	Depth perception
Δ_b	Axial phase delay
Δf	Variation in focal length
$\Delta\mathbf{k}$	Phase mismatch
Δz, $\Delta z'$	Object and image separations
$\Delta\lambda$	Wavelength bandwidth
$\Delta\nu$	Frequency bandwidth
$\Delta\varphi$	Phase delay
$\Delta\varphi$	Phase difference
$\Delta\varphi$	Phase shift
$\Delta\varphi_{min}$	Minimum perceived phase difference
ε	Angle between interfering beams
ε	Dielectric constant, i.e., medium permittivity
η	Quantum efficiency
θ'	Refraction angle
θ_{cr}	Critical angle
θ_i	Incidence angle
θ_r	Reflection angle
λ	Wavelength of light
λ_p	Peak wavelength for the mth interference order
μ	Magnetic permeability
ν	Frequency of light
ν	Repetition frequency
ξ	Propagation direction
ρ	Spatial frequency
σ	Molecular absorption cross section

Glossary of Symbols (cont.)

σ_{dark}	Dark noise
σ_{photon}	Photon noise
σ_{read}	Read noise
τ	Integration time
τ	Length of a pulse
τ	Transmittance
φ	Incident photon flux
φ	Optical power
φ	Phase difference generated by a thin film
φ_o	Initial phase
φ_o	Phase delay through the object
φ_p	Phase delay in a phase plate
φ_{TF}	Phase difference generated by a thin film
ω	Angular frequency
ω	Bandgap frequency
$\perp$ and $\parallel$	Perpendicular and parallel components of the light vector

Nature of Light

Optics uses two approaches, termed the **particle** and **wave models,** to describe the nature of light. The former arises from atomic structure, where the transitions between energy levels are quantized. Electrons can be excited into higher energy levels via external processes with the release of a discrete quantum of energy (a photon) upon their decay to a lower level.

The wave model complements this corpuscular theory and explains optical effects involving diffraction and interference. The wave and particle models can be related through the frequency of oscillations, with the relationship between quanta of energy and frequency given by

$$E = h\nu \text{ [in eV or J]},$$

where $h = 4.135667\times10^{-15}$ [eV·s] $= 6.626068\times10^{-34}$ [J·s] is **Planck's constant**, and ν is the **frequency of light** [Hz]. The frequency ν of a wave is the reciprocal of its period T [s]:

$$\nu = \frac{1}{T}\,.$$

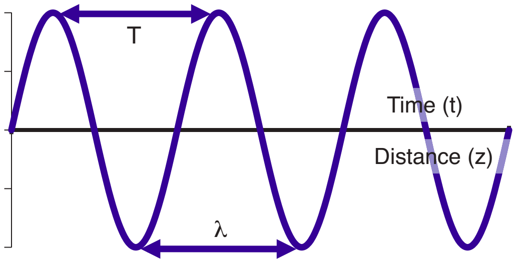

The period T [s] corresponds to one cycle of a wave or, if defined in terms of the distance required to perform one full oscillation, describes the **wavelength of light** λ [m]. The **velocity of light** in free space is 2.99792×10^8 m/s and is defined as the distance traveled in one period (λ) divided by the time taken (T):

$$c = \frac{\lambda}{T}\,.$$

Note that wavelength is often measured indirectly as time T required to pass the distance between wave oscillations.

The relationship for the velocity of light c can be rewritten in terms of wavelength and frequency as

$$c = \lambda\nu\,.$$

The Spectrum of Microscopy

The range of electromagnetic radiation is called the **electromagnetic spectrum**. Various microscopy techniques employ wavelengths spanning from **x-ray radiation** and **ultraviolet radiation** (UV) through the **visible spectrum** (VIS) to **infrared radiation** (IR). The wavelength in combination with the microscope parameters determines the resolution limit of the microscope (0.61λ/NA). The smallest feature resolved using light microscopy and determined by diffraction is approximately 200 nm for UV light with a high numerical aperture (for more details, see *Resolution Limit* on page 39). However, recently emerging super-resolution techniques overcome this barrier, and features are observable in the 20–50 nm range.

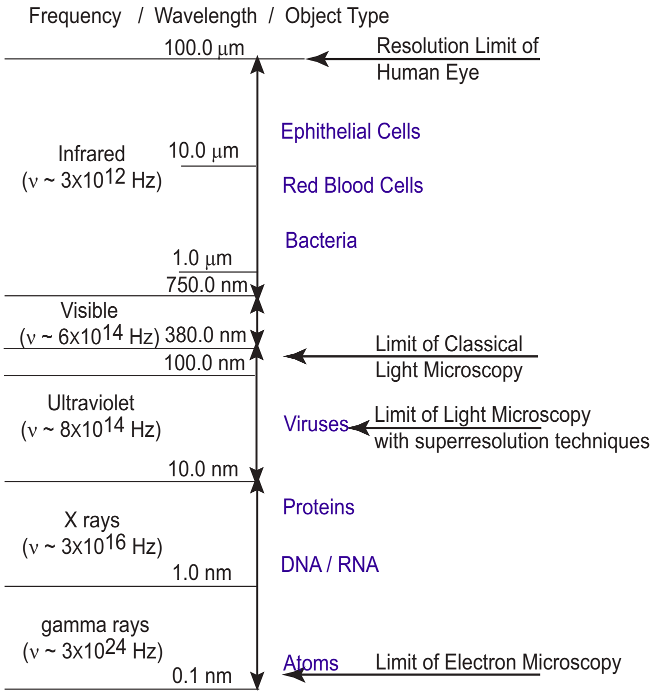
Frequency / Wavelength / Object Type
100.0 μm
Resolution Limit of Human Eye
Ephithelial Cells
Infrared
(ν ~ 3x10^12 Hz)
10.0 μm
Red Blood Cells
Bacteria
1.0 μm
750.0 nm
Visible
(ν ~ 6x10^14 Hz)
380.0 nm
Limit of Classical Light Microscopy
100.0 nm
Ultraviolet
(ν ~ 8x10^14 Hz)
Viruses
Limit of Light Microscopy with superresolution techniques
10.0 nm
Proteins
X rays
(ν ~ 3x10^16 Hz)
1.0 nm
DNA / RNA
gamma rays
(ν ~ 3x10^24 Hz)
0.1 nm
Atoms
Limit of Electron Microscopy

Wave Equations

Maxwell's equations describe the propagation of an electromagnetic wave. For homogenous and isotropic media, magnetic (*H*) and electric (*E*) components of an electromagnetic field can be described by the wave equations derived from Maxwell's equations:

$$\nabla^2 \vec{E} - \varepsilon_{\mathrm{m}} \mu_{\mathrm{m}} \frac{\partial^2 \vec{E}}{\partial t^2} = 0$$

$$\nabla^2 \vec{H} - \varepsilon_{\mathrm{m}} \mu_{\mathrm{m}} \frac{\partial^2 \vec{H}}{\partial t^2} = 0 ,$$

where ε is a **dielectric constant,** i.e., **medium permittivity** while μ is a **magnetic permeability**:

$$\varepsilon_{\mathrm{m}} = \varepsilon_{\mathrm{o}} \varepsilon_{\mathrm{r}} \qquad \mu_{\mathrm{m}} = \mu_{\mathrm{o}} \mu_{\mathrm{r}}$$

Indices r, m, and o stand for relative, media, or vacuum, respectively.

The above equations indicate that the time variation of the electric field and the current density creates a time-varying magnetic field. Variations in the magnetic field induce a time-varying electric field. Both fields depend on each other and compose an electromagnetic wave. Magnetic and electric components can be separated.

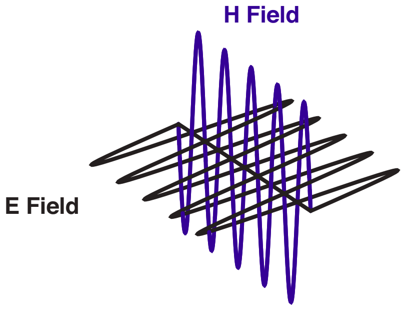

The electric component of an electromagnetic wave has primary significance for the phenomenon of light. This is because the magnetic component cannot be measured with optical detectors. Therefore, the magnetic component is most often neglected, and the electric vector $\vec{E}$ is called a light vector.

Wavefront Propagation

Light propagating through isotropic media conforms to the equation

$$E = A\sin(\omega t - kz + \varphi_o),$$

where t is time, z is distance along the direction of propagation, and ω is an **angular frequency** given by

$$\omega = \frac{2\pi}{T} = \frac{2\pi V_m}{\lambda}.$$

The term kz is called the **phase of light**, while φ_o is an initial phase. In addition, k represents the **wave number** (equal to $2\pi/\lambda$) and V_m is the velocity of light in media:

$$kz = \frac{2\pi}{\lambda} nz$$

The above propagation equation pertains to the special case when the electric field vector varies only in one plane.

The **refractive index** of the dielectric media n describes the relationship between the speed of light in a vacuum and in media. It is

$$n = \frac{c}{V_m},$$

where c is the speed of light in vacuum.

An alternative way to describe the propagation of an electric field is with an exponential (complex) representation:

$$E = A\exp\left[-i(\omega t - kz + \varphi_o)\right].$$

This form allows the easy separation of phase components of an electromagnetic wave.

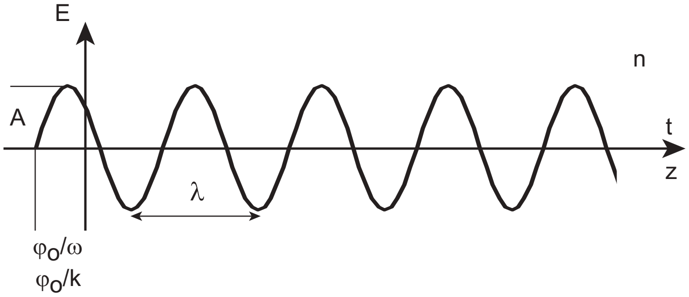

Optical Path Length (OPL)

Fermat's principle states that "The path traveled by a light wave from a point to another point is stationary with respect to variations of this path." In practical terms, this means that light travels along the minimum time path.

Optical path length (OPL) is related to the time required for light to travel from point P1 to point P2. It accounts for the media density through the refractive index:

$$t = \frac{1}{c}\int_{P1}^{P2} n\,ds$$

or

$$OPL = \int_{P1}^{P2} n\,ds\ ,$$

where

$$ds^2 = dx^2 + dy^2 + dz^2\ .$$

Optical path length can also be discussed in the context of the number of periods required to propagate a certain distance *L*. In a medium with refractive index *n*, light slows down and more wave cycles are needed. Therefore, OPL is an equivalent path for light traveling with the same number of periods in a vacuum:

$$OPL = nL\ .$$

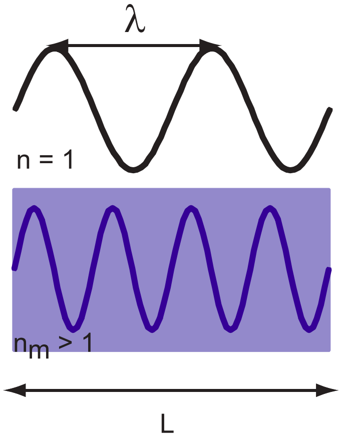

Optical path difference (OPD) is the difference between optical path lengths traversed by two light waves:

$$OPD = n_1L_1 - n_2L_2\ .$$

OPD can also be expressed as a phase difference:

$$\Delta\varphi = 2\pi\frac{OPD}{\lambda}\ .$$

Laws of Reflection and Refraction

Light rays incident on an interface between different dielectric media experience reflection and refraction as shown.

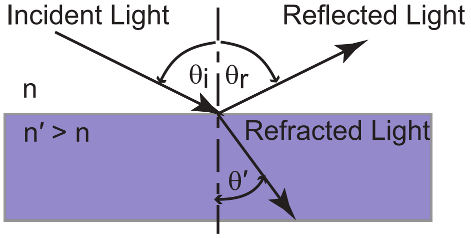

Reflection Law: Angles of **incidence** and **reflection** are related by

$$\theta_i = -\theta_r \, .$$

Refraction Law (**Snell's Law**): Incident and refracted angles are related to each other and to the refractive indices of the two media by Snell's Law:

$$n \sin\theta_i = n' \sin\theta' \, .$$

Fresnel reflection: The division of light at a dielectric boundary into transmitted and reflected rays is described for nonlossy, nonmagnetic media by the Fresnel equations:

Reflectance coefficients	**Transmission Coefficients**
$r_\perp = \dfrac{I_{r\perp}}{I_\perp} = \dfrac{\sin^2(\theta_i - \theta')}{\sin^2(\theta_i + \theta')}$	$t_\perp = \dfrac{I_{t\perp}}{I_\perp} = \dfrac{4\sin^2\theta'\cos^2\theta_i}{\sin^2(\theta_i + \theta')}$
$r_\parallel = \dfrac{I_{r\parallel}}{I_\parallel} = \dfrac{\tan^2(\theta_i - \theta')}{\tan^2(\theta_i + \theta')}$	$t_\parallel = \dfrac{I_{t\parallel}}{I_\parallel} = \dfrac{4\sin^2\theta'\cos^2\theta_i}{\sin^2(\theta_i + \theta')\cos^2(\theta_i - \theta')}$

t and r are transmission and reflection coefficients, respectively. I, I_t, and I_r are the irradiances of incident, transmitted, and reflected light, respectively. $\perp$ and $\parallel$ denote perpendicular and parallel components of the light vector with respect to the plane of incidence. θ_i and θ' in the table are the angles of incidence and reflection/refraction, respectively. At a normal incidence ($\theta' = \theta_i = 0$ deg), the Fresnel equations reduce to

$$r = r_\perp = r_\parallel = \frac{(n' - n)^2}{(n' + n)^2} \quad \text{and} \quad t = t_\perp = t_\parallel = \frac{4nn'}{(n' + n)^2} \, .$$

Total Internal Reflection

When light passes from one medium to a medium with a higher refractive index, the angle of the light in the medium bends toward the normal, according to Snell's Law. Conversely, when light passes from one medium to a medium with a lower refractive index, the angle of the light in the second medium bends away from the normal. If the angle of refraction is greater than 90 deg, the light cannot escape the denser medium, and it reflects instead of refracting. This effect is known as **total internal reflection** (TIR). TIR occurs only for illumination with an angle larger than the **critical angle**, defined as

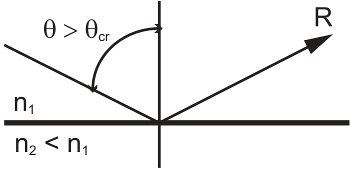

$$\theta_{cr} = \arcsin\left(\frac{n_1}{n_2}\right).$$

It appears, however, that light can propagate through (at a limited range) to the lower-density material as an **evanescent wave** (a nearly standing wave occurring at the material boundary), even if illuminated under an angle larger than θ_{cr}. Such a phenomenon is called **frustrated total internal reflection**. Frustrated TIR is used, for example, with thin films (TF) to build beam splitters. The proper selection of the film thickness provides the desired beam ratio (see the figure below for various split ratios). Note that the maximum film thickness must be approximately a single wavelength or less, otherwise light decays entirely. The effect of an evanescent wave propagating in the lower-density material under TIR conditions is also used in total internal reflection fluorescence (**TIRF**) microscopy.

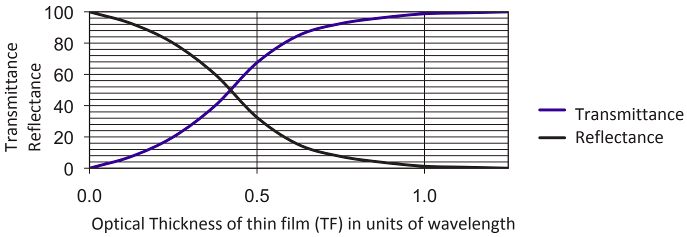

Evanescent Wave in Total Internal Reflection

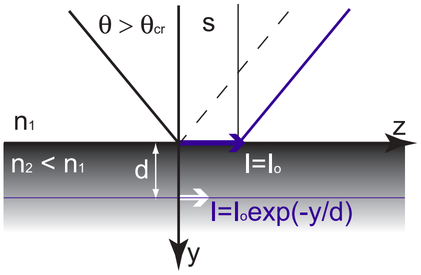

A beam reflected at the dielectric interface during **total internal reflection** is subject to a small lateral shift, also known as the **Goos-Hänchen effect**. The shift is different for the parallel and perpendicular components of the electromagnetic vector.

	Lateral shift in TIR
Parallel* component of e/m vector	$s_{\parallel} = \frac{\lambda n_1}{\pi n_2^2} \frac{\tan\theta}{\sqrt{\sin^2\theta - \sin^2\theta_c}}$
Perpendicular* component of e/m vector	$s_{\perp} = \frac{\lambda}{\pi n_1} \frac{\tan\theta}{\sqrt{\sin^2\theta - \sin^2\theta_c}}$
* to the plane of incidence	

The intensity of the illuminating beam decays exponentially with distance y in the direction perpendicular to the material boundary:

$$I = I_o \exp(-y/d).$$

Note that d denotes the distance at which the intensity of the illuminating light I_o drops by e. The decay distance is smaller than a wavelength. It is also inversely proportional to the illumination angle.

Decay distance (d) of evanescent wave	
Decay distance as a function of incidence and critical angles	$d = \frac{\lambda}{4\pi n_1} \frac{1}{\sqrt{\sin^2\theta - \sin^2\theta_c}}$
Decay distance as a function of incidence angle and refractive indices of media	$d = \frac{\lambda}{4\pi} \frac{1}{\sqrt{n_1^2 \sin^2\theta - n_2^2}}$

Propagation of Light in Anisotropic Media

In anisotropic media the velocity of light depends on the direction of propagation. Common anisotropic and optically transparent materials include **uniaxial crystals**. Such crystals exhibit one direction of travel with a single propagation velocity. The single velocity direction is called the **optic axis** of the crystal. For any other direction, there are two velocities of propagation.

The wave in birefringent crystals can be divided into two components: ordinary and extraordinary. An **ordinary wave** has a uniform velocity in all directions, while the velocity of an **extraordinary wave** varies with orientation. The extreme value of an extraordinary wave's velocity is perpendicular to the optic axis. Both waves are linearly polarized, which means that the electric vector oscillates in one plane. The vibration planes of extraordinary and ordinary vectors are perpendicular. Refractive indices corresponding to the direction of the optic axis and perpendicular to the optic axis are $n_o = c/V_o$ and $n_e = c/V_e$ respectively.

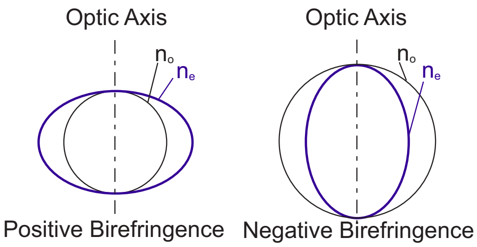

Uniaxial crystals can be positive or negative (see table). The **refractive index** $n(\xi)$ for velocities between the two extreme (n_o and n_e) values is

$$\frac{1}{n^2(\xi)} = \frac{\cos^2\xi}{n_o^2} + \frac{\sin^2\xi}{n_e^2}.$$

Positive Birefringence	$V_e \leq V_o$
Negative Birefringence	$V_e \geq V_o$

Note that the propagation direction ξ (with regard to the optic axis) is slightly off from the ray direction, which is defined by the Poynting (energy) vector.

Uniaxial Crystal	**Refractive index**	**Abbe Number**	**Wavelength Range** [μm]
Quartz	$n_o = 1.54424$	70	0.18–4.0
	$n_e = 1.55335$	69	
Calcite	$n_o = 1.65835$	50	0.2–2.0
	$n_e = 1.48640$	68	

The refractive index is given for the *D* spectral line (589.2 nm).

(Adapted from Pluta, 1988)

Polarization of Light and Polarization States

The orientation characteristic of a light vector vibrating in time and space is called the **polarization of light**. For example, if the vector of an electric wave vibrates in one plane, the state of polarization is linear. A vector vibrating with a random orientation represents unpolarized light.

The wave vector E consists of two components E_x and E_y.

$$E(z,t) = E_x + E_y$$

$$E_x = A_x \exp\left[-i\left(\omega t - kz + \varphi_x\right)\right]$$

$$E_y = A_y \exp\left[-i\left(\omega t - kz + \varphi_y\right)\right].$$

The **electric vector** rotates periodically (the periodicity corresponds to wavelength λ) in the plane perpendicular to the propagation axis (z) and generally forms an elliptical shape. The specific shape depends on the A_x and A_y ratios and the phase delay between the E_x and E_y components, defined as $\Delta\varphi = \varphi_x - \varphi_y$.

Linearly polarized light is obtained when one of the components E_x or E_y is zero, or when $\Delta\varphi$ is zero or π. **Circularly polarized light** is obtained when $E_x = E_y$ and $\Delta\varphi = \pm\pi/2$. The light is called right circularly polarized (RCP) if it rotates in a clockwise direction or left circularly polarized (LCP) if it rotates counterclockwise when looking at the oncoming beam.

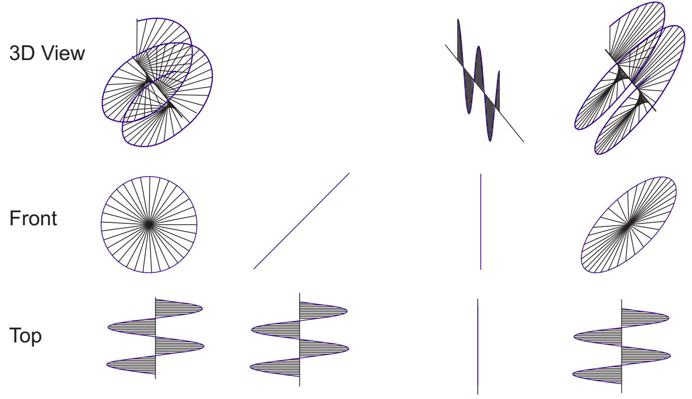

Polarization	Circular	Linear	Linear	Elliptical
Amplitudes	1 1	1 1	0 1	1 1
Δφ	π/2	0	0	π/4

Coherence and Monochromatic Light

An ideal light wave that extends in space at any instance to $\pm\infty$ and has only one wavelength λ (or frequency ν) is said to be **monochromatic**. If the range of wavelengths λ or frequencies ν is very small (around λ_o or ν_o, respectively), the wave is **quasi-monochromatic**. Such a wave packet is usually called a **wave group**.

Note that for monochromatic and quasi-monochromatic waves, no phase relationship is required, and the intensity of light can be calculated as a simple summation of intensities from different waves; phase changes are very fast and random, so only the average intensity can be recorded.

If multiple waves have a common phase relation dependence, they are **coherent** or **partially coherent**. These cases correspond to full- and partial-phase correlation respectively. A common source of a coherent wave is a laser where waves must be in resonance and therefore in phase. The average length of the wave packet (group) is called the **coherence length**, while the time required to pass this length is called **coherence time**. Both values are linked by the equations

$$\Delta\nu = \frac{1}{t_c} = \frac{V}{l_c}$$

where coherence length is

$$l_c = \frac{\lambda^2}{\Delta\lambda}.$$

The coherence length l_c and **temporal coherence** t_c are inversely proportional to the bandwidth $\Delta\lambda$ of the light source.

Spatial coherence is a term related to the coherence of light with regard to the extent of the light source. The fringe contrast varies for interference of any two spatially different source points. Light is **partially coherent** if its coherence is limited by the source bandwidth, dimension, temperature, or other effects.

Interference

Interference is a process of superposition of two coherent (correlated) or partially coherent waves. Waves interact with each other, and the resulting intensity is described by summing the **complex amplitudes** (**electric fields**) E_1 and E_2 of both wavefronts:

$$E_1 = A_1 \exp\left[-i\left(\omega t - kz + \varphi_1\right)\right]$$

and

$$E_2 = A_2 \exp\left[-i\left(\omega t - kz + \varphi_2\right)\right].$$

The resultant field is

$$E = E_1 + E_2 .$$

Therefore, the **interference** of the two beams can be written as

$$I = EE^*,$$

$$I = A_1^2 + A_2^2 + 2A_1A_2\cos\left(\Delta\varphi\right),$$

$$I = I_1 + I_2 + 2\sqrt{I_1I_2}\cos\left(\Delta\varphi\right),$$

$$I_1 = E_1E_1^*,\ \ I_2 = E_2E_2^*,\ \text{and}$$

$$\Delta\varphi = \varphi_2 - \varphi_1,$$

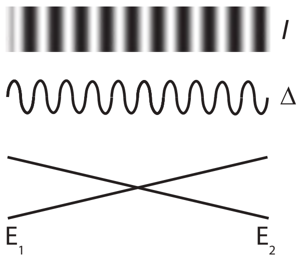

Propagating Wavefronts

where * denotes a conjugate function, I is the **intensity of light**, A is an amplitude of an electric field, and Δφ is the phase difference between the two interfering beams. **Contrast** C (called also **visibility**) of the interference fringes can be expressed as

$$C = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}.$$

The fringe existence and visibility depends on several conditions. To obtain the interference effect,

- Interfering beams must originate from the same light source and be temporally and spatially coherent;
- The polarization of interfering beams must be aligned;
- To maximize the contrast, interfering beams should have equal or similar amplitudes.

Conversely, if two noncorrelated, random waves are in the same region of space, the sum of intensities (irradiances) of these waves gives a total intensity in that region: $I = I_1 + I_2$.

Contrast vs Spatial and Temporal Coherence

The result of interference is a periodic intensity change in space, which creates fringes when incident on a screen or detector. **Spatial coherence** relates to the contrast of these fringes depending on the extent of the source and is not a function of the phase difference (or OPD) between beams. The intensity of interfering fringes is given by

$$I = I_1 + I_2 + 2C\left(\text{Source Extent}\right)\sqrt{I_1 I_2}\cos\left(\Delta\varphi\right),$$

where C is a constant, depending on the extent of the source.

$C = 1$

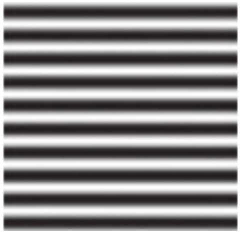

$C = 0.5$

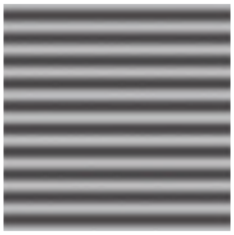

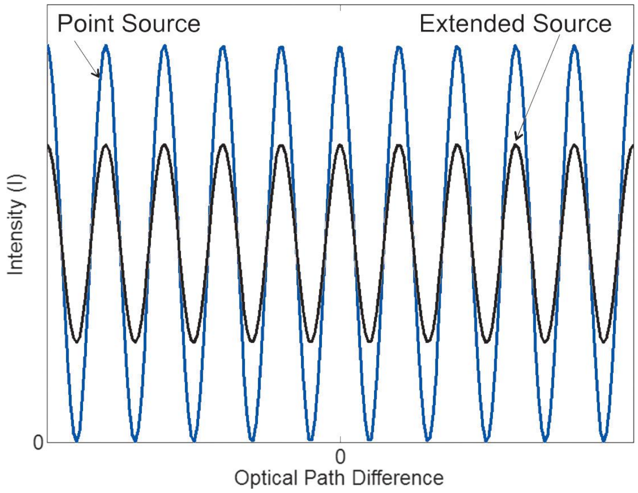

The spatial coherence can be improved through spatial filtering. For example, light can be focused on the pinhole (or coupled into the fiber) by using a microscope objective. In microscopy, spatial coherence can be adjusted by changing the diaphragm size in the conjugate plane of the light source.

Contrast vs Spatial and Temporal Coherence (cont.)

The intensity of the fringes depends on the OPD and the **temporal coherence** of the source. The fringe contrast trends toward zero as the OPD increases beyond the coherence length:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\,C(OPD)\cos(\Delta\varphi).$$

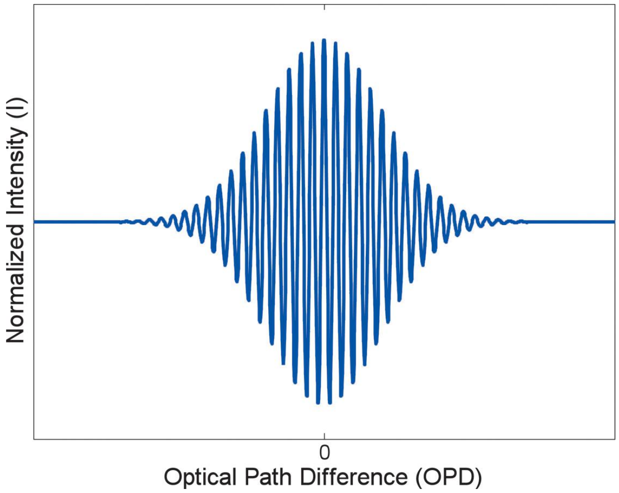

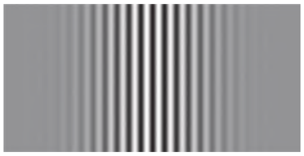

The spatial width of a fringe pattern envelope decreases with shorter temporal coherence. The location of the envelope's peak (with respect to the reference) and narrow width can be efficiently used as a gating mechanism in both imaging and metrology. For example, it is used in interference microscopy for optical profiling (white light interferometry) and for imaging using optical coherence tomography. Examples of short coherence sources include white light sources, luminescent diodes, and broadband lasers.

Long coherence length is beneficial in applications that do not naturally provide near-zero OPD, e.g., in surface measurements with a Fizeau interferometer.

Contrast of Fringes (Polarization and Amplitude Ratio)

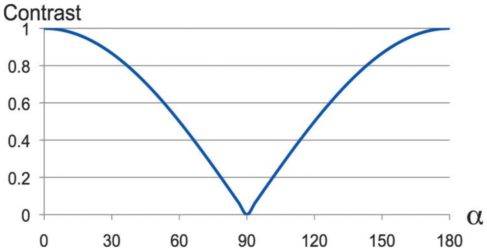

The contrast of fringes depends on the polarization orientation of the electric vectors. For linearly polarized beams, it can be described as $C = |\cos\alpha|$, where α represents the angle between polarization states.

Angle between Electric Vectors of Interfering Beams			
0	π/6	π/3	π/2
Interference Fringes			
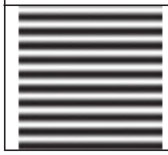	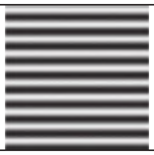	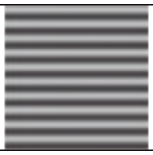	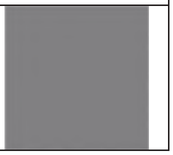

Fringe contrast also depends on the interfering beams' amplitude ratio. The intensity of the fringes is simply calculated by using the main interference equation. The contrast is maximum for equal beam intensities and for the interference pattern, defined as

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\Delta\varphi),$$

it is

$$C = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}.$$

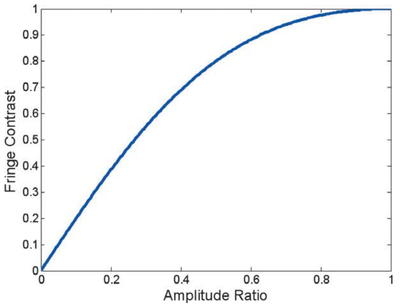

Ratio between Amplitudes of Interfering Beams			
1	0.5	0.25	0.1
Interference Fringes			
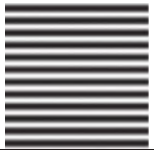	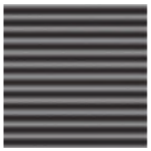	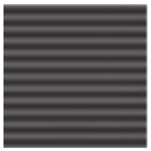	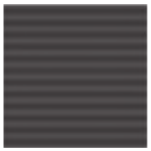

Multiple Wave Interference

If light reflects inside a thin film, its intensity gradually decreases and **multiple beam interference** occurs.

The intensity of reflected light is

$$I_r = I_i \frac{F \sin^2(\varphi/2)}{1 + F \sin^2(\varphi/2)} ,$$

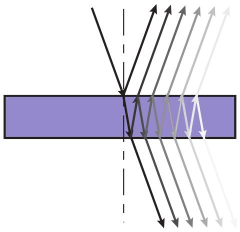

and for transmitted light is

$$I_t = I_i \frac{1}{1 + F \sin^2(\varphi/2)} .$$

The coefficient of finesse *F* of such a resonator is

$$F = \frac{4r}{(1-r)^2} .$$

Because phase depends on the wavelength, it is possible to design selective **interference filters**. In this case, a dielectric thin film is coated with metallic layers. The peak wavelength for the mth interference order of the filter can be defined as

$$\lambda_p = \frac{2nt \cos\theta'}{m - \left(\frac{\varphi_{TF}}{2\pi}\right)} ,$$

where φ_{TF} is the phase difference generated by a thin film of thickness *t* and a specific incidence angle. The **interference order** relates to the phase difference in multiples of 2π.

Interference filters usually operate for illumination with flat wavefronts at a normal incidence angle. Therefore, the equation simplifies as

$$\lambda_p = \frac{2nt}{m} .$$

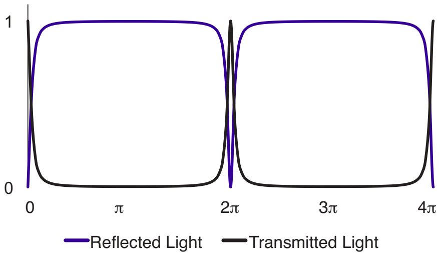

The **half bandwidth** (HBW) of the filter for normal incidence is

$$HBW = \frac{1-r}{m\pi} \lambda_p \text{ [m]} .$$

The peak intensity transmission is usually at 20% to 50%, or up to 90% of the incident light for metal-dielectric or multi-dielectric filters respectively.

Interferometers

Due to the high frequency of light, it is not possible to detect the phase of the light wave directly. To acquire the phase, interferometry techniques may be used. There are two major classes of interferometers, based on **amplitude splitting** and **wavefront splitting**.

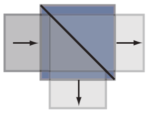
Amplitude Split

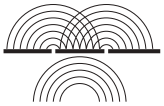
Wavefront Split

From the point of view of fringe pattern interpretation, interferometers can also be divided into those that deliver fringes, which directly correspond to the phase distribution, and those providing a derivative of the phase map (also called shearing interferometers). The first type is realized by obtaining the interference between the test beam and a reference beam. An example of such a system is the **Michelson interferometer**. **Shearing interferometers** provide the intereference of two shifted object wavefronts.

Michelson Interferometer

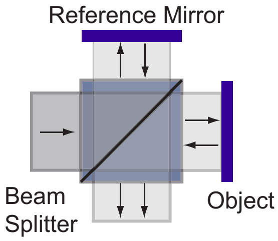
Reference Mirror
Beam
Splitter
Object

There are numerous ways of introducing shear (linear, radial, etc.). Examples of shearing interferometers include the parallel or wedge plate, the grating interferometer, the **Sagnac**, and polarization interferometers (commonly used in microscopy). The **Mach-Zehnder interferometer** can be configured for direct and differential fringes, depending on the position of the sample.

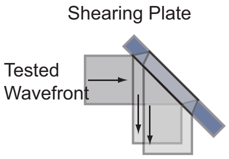
Shearing Plate
Tested
Wavefront

Mach Zehnder Interferometer
(Direct Fringes)

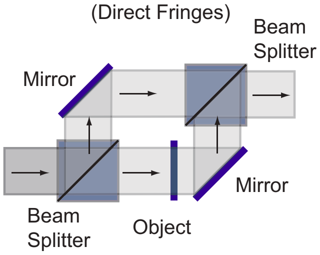
Beam
Splitter
Mirror
Mirror
Beam
Splitter
Object

Mach Zehnder Interferometer
(Differential Fringes)

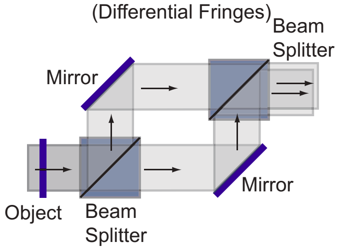
Beam
Splitter
Mirror
Mirror
Object
Beam
Splitter

Diffraction

The bending of waves by apertures and objects is called **diffraction of light**. Waves diffracted inside the optical system interfere (for path differences within the coherence length of the source) with each other and create diffraction patterns in the image of an object.

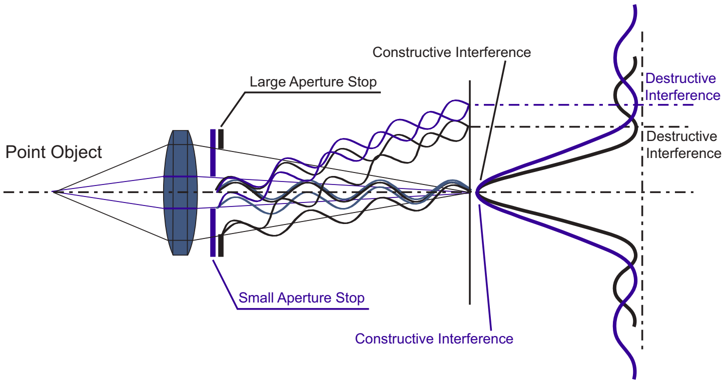
Constructive Interference
Large Aperture Stop
Destructive Interference
Destructive Interference
Point Object
Small Aperture Stop
Constructive Interference

There are two common approximations of diffraction phenomena: **Fresnel diffraction** (near-field) and **Fraunhofer diffraction** (far-field). Both diffraction types complement each other but are not sharply divided due to various definitions of their regions. Fraunhofer diffraction occurs when one can assume that propagating wavefronts are flat (collimated), while the Fresnel diffraction is the near-field case. Thus, Fraunhofer diffraction (distance z) for a free-space case is infinity, but in practice it can be defined for a region

$$z > S_{\mathrm{FD}} \frac{d^2}{\lambda},$$

where d is the diameter of the diffractive object, and S_{FD} is a factor depending on approximation. A conservative definition of the Fraunhofer region is $S_{\mathrm{FD}} = 1$, while for most practical cases it can be assumed to be 10 times smaller ($S_{\mathrm{FD}} = 0.1$).

Diffraction effects influence the resolution of an imaging system and are a reason for fringes (ring patterns) in the image of a point. Specifically, Fraunhofer fringes appear in the conjugate image plane. This is due to the fact that the image is located in the Fraunhofer diffracting region of an optical system's aperture stop.

Diffraction Grating

Diffraction gratings are optical components consisting of periodic linear structures that provide “organized” diffraction of light. In practical terms, this means that for specific angles (depending on illumination and grating parameters), it is possible to obtain constructive (2π phase difference) and destructive (π phase difference) interference at specific directions. The angles of constructive interference are defined by the diffraction grating equation and called **diffraction orders** (m).

Gratings can be classified as transmission or reflection (mode of operation), amplitude (periodic amplitude changes) or phase (periodic phase changes), or ruled or holographic (method of fabrication). Ruled gratings are mechanically cut and usually have a triangular profile (each facet can thereby promote select diffraction angles). Holographic gratings are made using interference (sinusoidal profiles).

Diffraction angles depend on the ratio between the grating constant and wavelength so various wavelengths can be separated. This makes them applicable for spectroscopic detection or spectral imaging. The **grating equation** is

$$m\lambda = d \cos \gamma \, (\sin \beta \pm \sin \alpha)$$

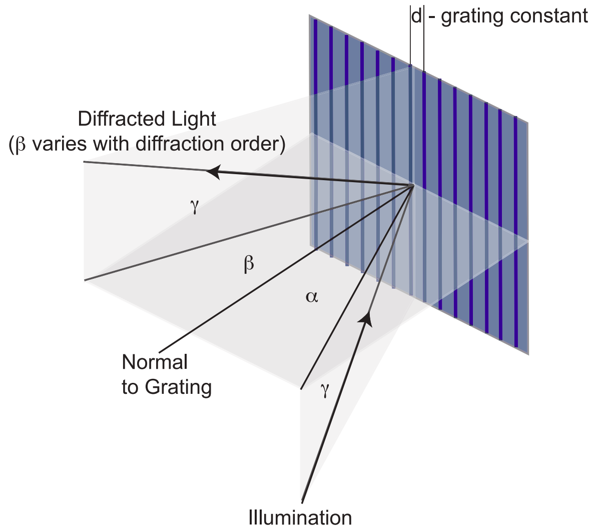

Diffraction Grating (cont.)

The sign in the diffraction grating equation defines the type of grating: A **transmission grating** is identified with a minus sign (–), and a **reflective grating** is identified with a plus sign (+).

For a grating illuminated in the plane perpendicular to the grating, its equation simplifies and is

$$m\lambda = d\left(\sin\beta \pm \sin\alpha\right).$$

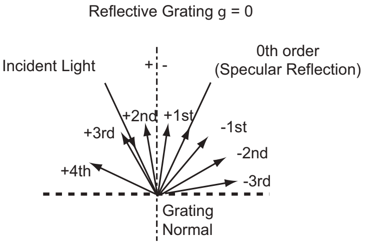
Reflective Grating g = 0
Incident Light
+
-
0th order
(Specular Reflection)
+2nd
+1st
+3rd
-1st
-2nd
+4th
-3rd
Grating
Normal

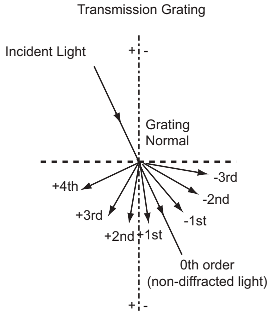
Transmission Grating
Incident Light
+
-
Grating
Normal
-3rd
+4th
-2nd
+3rd
-1st
+2nd
+1st
0th order
(non-diffracted light)
+
-

For normal illumination, the grating equation becomes

$$m\lambda = d\sin\beta.$$

The **chromatic resolving power** of a grating depends on its size (total number of lines *N*). If the grating has more lines, the diffraction orders are narrower, and the resolving power increases:

$$\frac{\lambda}{\Delta\lambda} = mN.$$

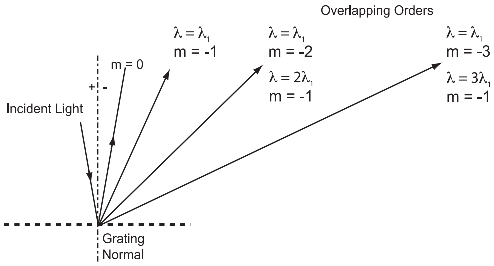
Overlapping Orders
$\lambda = \lambda_1$
m = -1
$\lambda = \lambda_1$
m = -2
$\lambda = 2\lambda_1$
m = -1
$\lambda = \lambda_1$
m = -3
$\lambda = 3\lambda_1$
m = -1
m = 0
+
-
Incident Light
Grating
Normal

The **free spectral range** $\Delta\lambda$ of the grating is the bandwidth in the mth order without overlapping other orders. It defines useful bandwidth for spectroscopic detection as

$$\Delta\lambda = \frac{m+1}{m}\lambda_1 - \lambda_1 = \frac{\lambda_1}{m}.$$

Useful Definitions from Geometrical Optics

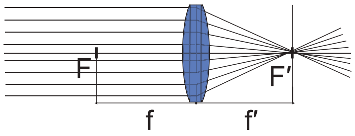

All optical systems consist of refractive or reflective interfaces (surfaces) changing the propagation angles of **optical rays**. Rays define the propagation trajectory and always travel perpendicular to the wavefronts. They are used to describe imaging in the regime of geometrical optics and to perform optical design.

There are two **optical spaces** corresponding to each optical system: **object space** and **image space**. Both spaces extend from –∞ to +∞ and are divided into real and virtual parts. Ability to access the image or object defines the real part of the optical space.

Paraxial optics is an approximation assuming small ray angles and is used to determine first-order parameters of the optical system (image location, magnification etc.).

Thin lenses are optical components reduced to zero thickness and are used to perform first-order optical designs (see next page).

The **focal point** of an optical system is a location that collimated beams converge to or diverge from. Planes perpendicular to the optical axis at the focal points are called **focal planes**. **Focal length** is the distance between the lens (specifically, its principal plane) and the focal plane. For thin lenses, principal planes overlap with the lens.

Sign Convention: The common sign convention assigns positive values to distances traced from left to right (the direction of propagation) and to the top. Angles are positive if they are measured counterclockwise from normal to the surface or optical axis. If light travels from right to left, the refractive index is negative. The surface radius is measured from its vertex to its center of curvature.

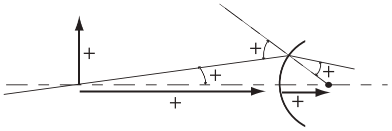

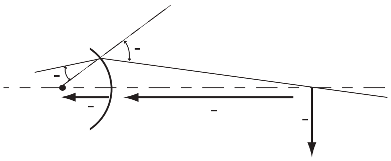

Image Formation

A simple model describing imaging through an optical system is based on thin lens relationships. A **real image** is formed at the point where rays converge.

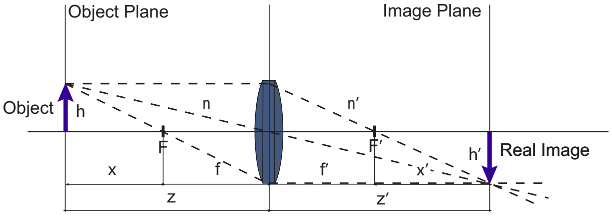
Object Plane
Image Plane
Object
h
n
n′
F
F′
h′
Real Image
x
f
f′
x′
z
z′

A **virtual image** is formed at the point from which rays appear to diverge.

For a lens made of glass, surrounded on both sides by air ($n = n' = 1$), the imaging relationship is described by the **Newtonian equation**

$$xx' = ff' \qquad \text{or} \qquad xx' = -f'^2 .$$

Note that Newtonian equations refer to the distance from the focal planes so, in practice, they are used with thin lenses.

Imaging can also be described by the **Gaussian imaging equation**

$$\frac{f}{z} + \frac{f'}{z'} = 1 .$$

The **effective focal length** of the system is

$$f_e = \frac{1}{\varphi} = -\frac{f}{n} = \frac{f'}{n'} ,$$

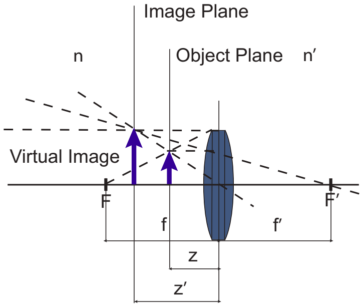
Image Plane
n
Object Plane
n′
Virtual Image
F
F′
f
f′
z
z′

where φ is the **optical power** expressed in diopters D [m^{-1}].

Therefore,

$$\frac{n'}{z'} = \frac{n}{z} + \frac{1}{f_e} .$$

For air, when n and n' are equal to 1, the imaging relation is

$$\frac{1}{z'} = \frac{1}{z} + \frac{1}{f_e} .$$

Magnification

Transverse magnification (also called **lateral magnification**) of an optical system is defined as the ratio of the image and object size measured in the direction perpendicular to the optical axis:

$$M = \frac{h'}{h} = -\frac{x'}{f'} = -\frac{f}{x} = \frac{z'}{z} = \frac{-f}{z-f} = \frac{-(z'-f')}{f'}\,.$$

Longitudinal magnification defines the ratio of distances for a pair of conjugate planes:

$$\frac{\Delta z'}{\Delta z} = \left(-\frac{f'}{f}\right) M_1 M_2\,,$$

where

$$\Delta z = z_2 - z_1$$
$$\Delta z' = z_2' - z_1'$$

and

$$M_1 = \frac{h_1'}{h_1}$$
$$M_2 = \frac{h_2'}{h_2}\,.$$

Angular magnification is the ratio of angular image size to angular object size and can be calculated with

$$M_u = \frac{u'}{u} = \frac{z}{z'}\,.$$

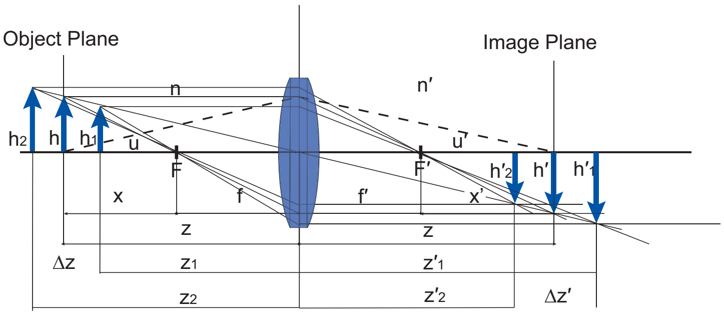

Stops and Rays in an Optical System

The primary stops in any optical system are the **aperture stop** (which limits light) and **field stop** (which limits the extent of the imaged object or the field of view). The aperture stop also defines the resolution of the optical system. To determine the aperture stop, all system diaphragms including the lens mounts should be imaged to either the image or the object space of the system. The aperture stop is determined as the smallest diaphragm or diaphragm image (in terms of angle) as seen from the on-axis object/image point in the same optical space.

Note that there are two important conjugates of the aperture stop in object and image space. They are called the **entrance pupil** and **exit pupil**, respectively.

The physical stop limiting the extent of the field is called the **field stop**. To find a field stop, all of the diaphragms should be imaged to the object or image space, with the smallest diaphragm defining the actual field stop, as seen from the entrance/exit pupil. Conjugates of the field stop in the object and image space are called the **entrance window** and **exit window**, respectively.

Two major rays that pass through the system are the **marginal ray** and the **chief ray**. The marginal ray crosses the optical axis at the conjugate of the field stop (and object) and the edge of the aperture stop. The chief ray goes through the edge of the field and crosses the optical axis at the aperture stop plane.

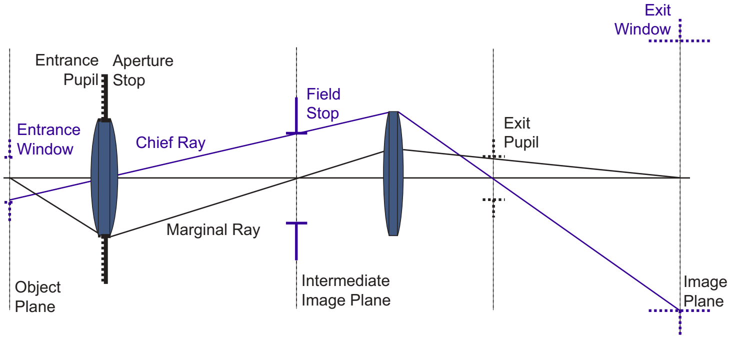

Aberrations

Individual spherical lenses cannot deliver perfect imaging because they exhibit errors called aberrations. All aberrations can be considered either **chromatic** or **monochromatic**. To correct for aberrations, optical systems use multiple elements, aspherical surfaces, and a variety of optical materials.

Chromatic aberrations are a consequence of the **dispersion** of optical materials, which is a change in the refractive index as a function of wavelength. The parameter-characterizing dispersion of any material is called the **Abbe number** and is defined as

$$V_{\mathrm{d}} = \frac{n_{\mathrm{d}} - 1}{n_{\mathrm{F}} - n_{\mathrm{C}}}.$$

Alternatively, the following equation might be used for other wavelengths:

$$V_{\mathrm{e}} = \frac{n_{\mathrm{e}} - 1}{n_{\mathrm{F'}} - n_{\mathrm{C'}}}.$$

λ [nm]	Symbol	Spectral Line
656	C	red hydrogen
644	C′	red cadmium
588	d	yellow helium
546	e	green mercury
486	F	blue hydrogen
480	F′	blue cadmium

In general, V can be defined by using refractive indices at any three wavelengths, which should be specified for material characteristics. Indices in the equations denote spectral lines. If V does not have an index, V_{d} is assumed.

Geometrical aberrations occur when optical rays do not meet at a single point. There are longitudinal and transverse ray aberrations describing the axial and lateral deviations from the paraxial image of a point (along the axis and perpendicular to the axis in the image plane), respectively.

Wave aberrations describe a deviation of the wavefront from a perfect sphere. They are defined as a distance (the optical path difference) between the wavefront and the reference sphere along the optical ray.

Chromatic Aberrations

Chromatic aberrations occur due to the dispersion of optical materials used for lens fabrication. This means that the refractive index is different for different wavelengths; consequently, various wavelengths are refracted differently.

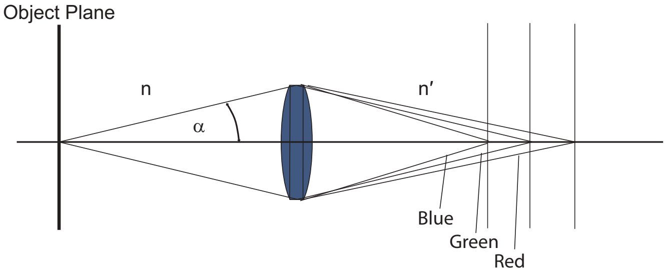

Chromatic aberrations include axial (longitudinal) or transverse (lateral) aberrations. **Axial chromatic aberration** arises from the fact that various wavelengths are focused at different distances behind the optical system. It is described as a variation in focal length:

$$\frac{\Delta f}{f} = \frac{f_F - f_C}{f} = \frac{1}{V}\,.$$

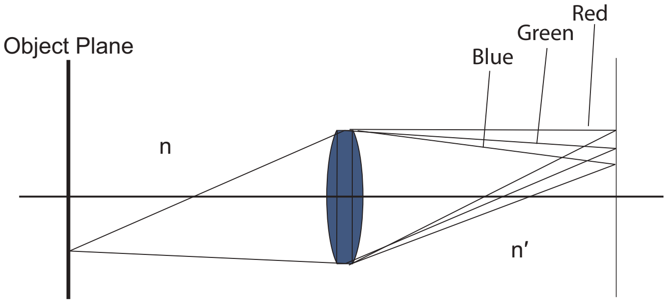

Transverse chromatic aberration is an off-axis imaging of colors at different locations on the image plane.

To compensate for chromatic aberrations, materials with low and high Abbe numbers are used (such as flint and crown glass). Correcting chromatic aberrations is crucial for most microscopy applications, but it is especially important for multi-photon microscopy. Obtaining multi-photon excitation requires high laser power and is most effective using short pulse lasers. Such a light source has a broad spectrum, and chromatic aberrations may cause pulse broadening.

Spherical Aberration and Coma

The most important wave aberrations are spherical, coma, astigmatism, field curvature, and distortion.

Spherical aberration (on-axis) is a consequence of building an optical system with components with spherical surfaces. It occurs when rays from different heights in the pupil are focused at different planes along the optical axis. This results in an axial blur. The most common approach for correcting spherical aberration uses a combination of negative and positive lenses. Systems that correct spherical aberration heavily depend on imaging conditions. For example, in microscopy a cover glass must be of an appropriate thickness and refractive index in order to work with an objective. Also, the media between the objective and the sample (such as air, oil, or water) must be taken into account.

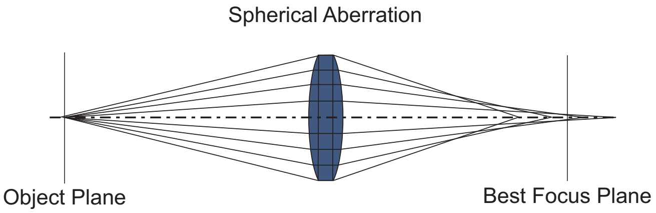

Coma (off-axis) can be defined as a variation of magnification with aperture location. This means that rays passing through a different azimuth of the lens are magnified differently. The name “coma” was inspired by the aberration’s appearance, because it resembles a comet’s tail as it emanates from the focus spot. It is usually stronger for lenses with a larger field, and its correction requires accommodation of the field diameter.

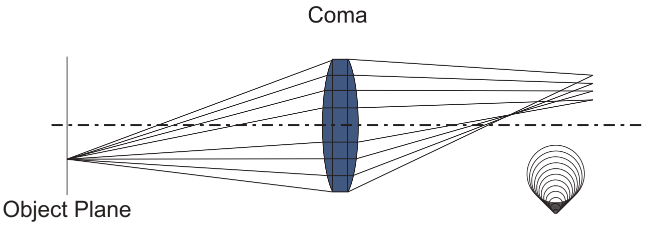

Astigmatism, Field Curvature, and Distortion

Astigmatism (off-axis) is responsible for different magnifications along orthogonal meridians in an optical system. It manifests as elliptical, elongated spots for the horizontal and vertical directions on opposite sides of the best focal plane. It is more pronounced for an object farther from the axis and is a direct consequence of improper lens mounting or an asymmetric fabrication process.

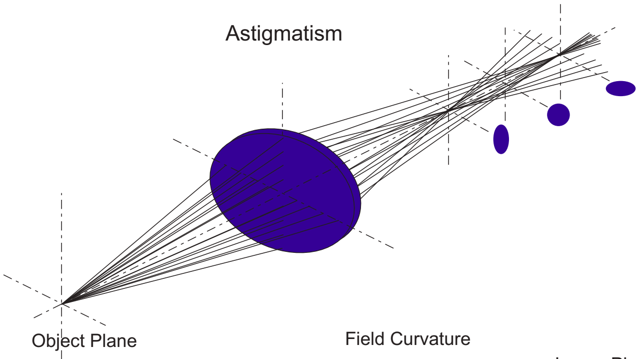

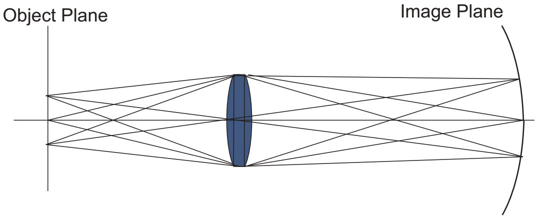

Field curvature (off-axis) results in a non-flat image plane. The image plane created is a concave surface as seen from the objective; therefore, various zones of the image can be seen in focus after moving the object along the optical axis. This aberration is corrected by an objective design combined with a tube lens or eyepiece.

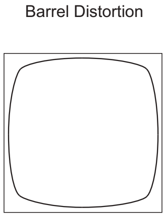

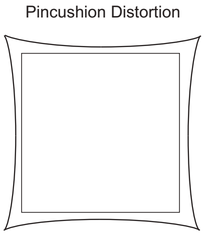

Distortion is a radial variation of magnification that will image a square as a pincushion or barrel. It is corrected in the same manner as field curvature. If preceded with system calibration, it can also be corrected numerically after image acquisition.

Performance Metrics

The major metrics describing the performance of an optical system are the **modulation transfer function** (MTF), the **point spread function** (PSF), and the **Strehl ratio** (SR).

The MTF is the modulus of the **optical transfer function** described by

$$OTF(\rho) = MTF(\rho)\exp\left[i\varphi(\rho)\right],$$

where the complex term in the equation relates to the phase transfer function. The MTF is a contrast distribution in the image in relation to contrast in the object as a function of spatial frequency ρ (for sinusoidal object harmonics) and can be defined as

$$MTF(\rho) = \frac{C(\rho)_{\text{image}}}{C(\rho)_{\text{object}}}.$$

The PSF is the intensity distribution at the image of a point object. This means that the PSF is a metric directly related to the image, while the MTF corresponds to spatial frequency distributions in the pupil. The MTF and PSF are closely related and comprehensively describe the quality of the optical system. In fact, the amplitude of the Fourier transform of the PSF results in the MTF.

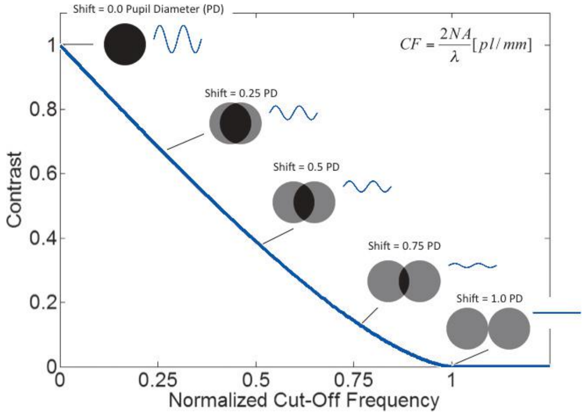

The MTF can be calculated as the autocorrelation of the **pupil function**. The pupil function describes the field distribution of an optical wave in the pupil plane of the optical system. In the case of a uniform pupil's transmission, it directly relates to the field overlap of two mutually shifted pupils where the shift corresponds to spatial frequency.

Performance Metrics (cont.)

The **modulation transfer function** has different results for coherent and incoherent illumination. For incoherent illumination, the phase component of the field is neglected since it is an average of random fields propagating under random angles.

For coherent illumination, the contrast of transferring harmonics of the field is constant and equal to 1 until the location of the spatial frequency in the pupil reaches its edge. For higher frequencies, the contrast sharply drops to zero since they cannot pass the optical system. Note that contrast for the coherent case is equal to 1 for the entire MTF range.

The **cutoff frequency** for an incoherent system is two times the cutoff frequency of the equivalent aperture coherent system and defines the **Sparrow resolution limit**.

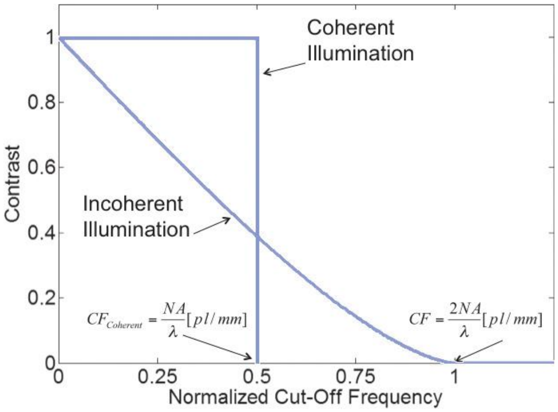

The **Strehl ratio** is a parametric measurement that defines the quality of the optical system in a single number. It is defined as the ratio of irradiance within the theoretical dimension of a diffraction-limited spot to the entire irradiance in the image of the point. One simple method to estimate the Strehl ratio is to divide the field below the MTF curve of a tested system by the field of the diffraction-limited system of the same numerical aperture. For practical optical design consideration it is usually assumed that the system is diffraction limited if the Strehl ratio is equal to or larger than 0.8.

The Compound Microscope

The primary goal of microscopy is to provide the ability to resolve the small details of an object. Historically, microscopy was developed for visual observation and, therefore, was set up to work in conjunction with the human eye. An effective high-resolution optical system must have resolving ability and be able to deliver proper magnification for the detector. In the case of visual observations, the detectors are the cones and rods of the retina.

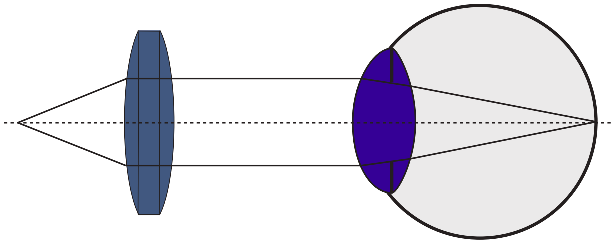

A basic microscope can be built with a single-element, short-focal-length lens (**magnifier**). The object is located in the focal plane of the lens and is imaged to infinity. The eye creates a final image of the object on the retina. The system stop is the eye's pupil.

To obtain higher resolution for visual observation, the **compound microscope** was first built in the 17th century by Robert Hooke. It consists of two major components: the **objective** and the **eyepiece**. The short-focal-length lens (objective) is placed close to the object under examination and creates a real image of an object at the focal plane of the second lens. The eyepiece (similar to the magnifier) throws an image to infinity, and the human eye creates the final image. An important function of the eyepiece is to match the eye's pupil with the system stop, which is located in the back focal plane of the microscope objective.

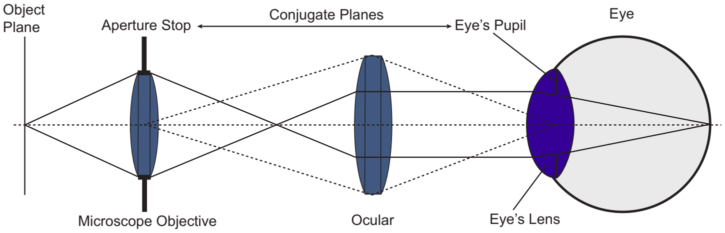

The Eye

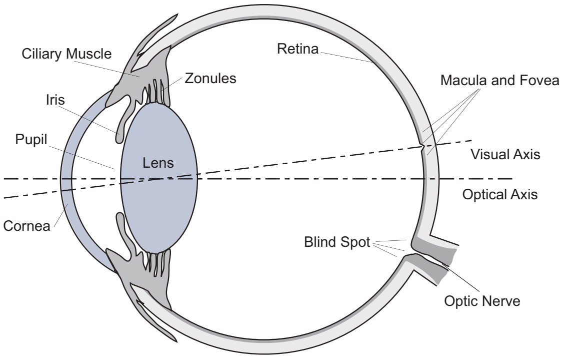

The **eye** was the first—and for a long time, the only—real-time detector used in microscopy. Therefore, the design of the microscope had to incorporate parameters responding to the needs of visual observation.

Cornea—the transparent portion of the sclera (the white portion of the human eye), which is a rigid tissue that gives the eyeball its shape. The cornea is responsible for two thirds of the eye's refractive power.

Lens—the lens is responsible for one third of the eye's power. Ciliary muscles can change the lens's power (accommodation) within the range of 15–30 diopters.

Iris—controls the diameter of the pupil (1.5–8 mm).

Retina—a layer with two types of photoreceptors: rods and cones. **Cones** (about 7 million) are in the area of the **macula** (~3 mm in diameter) and **fovea** (~1.5 mm in diameter, with the highest cone density), and they are designed for bright vision and color detection. There are three types of cones (red, green, and blue sensitive), and the spectral range of the eye is approximately 400–750 nm. **Rods** are responsible for night/low-light vision, and there are about 130 million located outside the fovea region.

It is arbitrarily assumed that the eye can provide sharp images for objects between 250 mm and infinity. A 250-mm distance is called the **minimum focus distance** or **near point**. The maximum eye resolution for bright illumination is 1 arc minute.

Upright and Inverted Microscopes

The two major microscope geometries are **upright** and **inverted.** Both systems can operate in reflectance and transmittance modes.

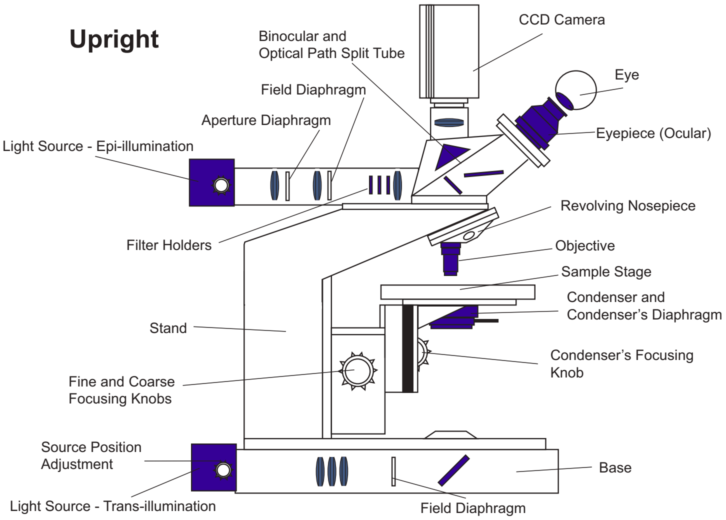
Upright
CCD Camera
Binocular and Optical Path Split Tube
Eye
Field Diaphragm
Aperture Diaphragm
Eyepiece (Ocular)
Light Source - Epi-illumination
Revolving Nosepiece
Filter Holders
Objective
Sample Stage
Condenser and Condenser's Diaphragm
Stand
Condenser's Focusing Knob
Fine and Coarse Focusing Knobs
Source Position Adjustment
Base
Light Source - Trans-illumination
Field Diaphragm

The inverted microscope is primarily designed to work with samples in cell culture dishes and to provide space (with a long-working distance condenser) for sample manipulation (for example, with patch pipettes in electrophysiology).

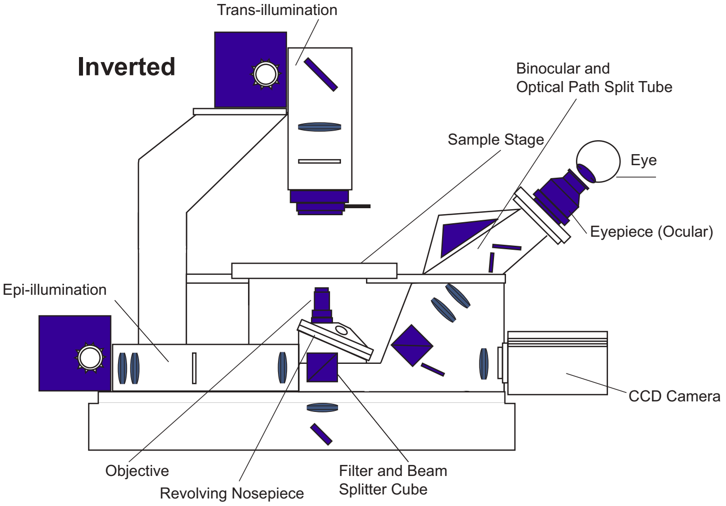
Trans-illumination
Inverted
Binocular and Optical Path Split Tube
Sample Stage
Eye
Eyepiece (Ocular)
Epi-illumination
CCD Camera
Objective
Revolving Nosepiece
Filter and Beam Splitter Cube

The Finite Tube Length Microscope

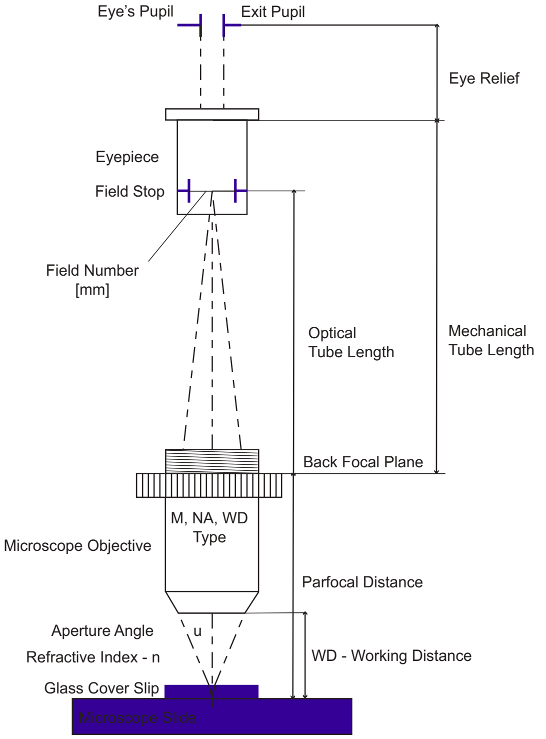
Eye's Pupil
Exit Pupil
Eye Relief
Eyepiece
Field Stop
Field Number [mm]
Optical Tube Length
Mechanical Tube Length
Back Focal Plane
M, NA, WD Type
Microscope Objective
Parfocal Distance
Aperture Angle
u
Refractive Index - n
WD - Working Distance
Glass Cover Slip
Microscope Slide

Historically, microscopes were built with a **finite tube length**. With this geometry, the microscope objective images the object into the tube end. This intermediate image is then relayed to the observer by an eyepiece. Depending on the manufacturer, different **optical tube lengths** are possible (for example, the standard tube length for Zeiss is 160 mm). The use of a standard tube length by each manufacturer unifies the optomechanical design of the microscope.

A constant **parfocal distance** for microscope objectives enables switching between different magnifications without defocusing. The **field number**, corresponding to the physical dimension (in millimeters) of the **field stop** inside the eyepiece, allows the microscope's field of view (FOV) to be determined according to

$$FOV = \frac{FieldNumber}{M_{\text{objective}}} \; [\text{mm}] .$$

Infinity-Corrected Systems

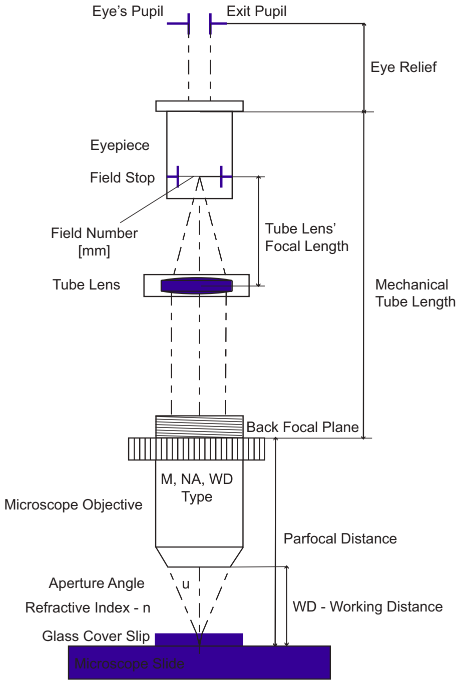
Eye's Pupil
Exit Pupil
Eye Relief
Eyepiece
Field Stop
Field Number [mm]
Tube Lens' Focal Length
Tube Lens
Mechanical Tube Length
Back Focal Plane
M, NA, WD Type
Microscope Objective
Parfocal Distance
Aperture Angle
u
Refractive Index - n
WD - Working Distance
Glass Cover Slip
Microscope Slide

Microscope objectives can be corrected for a conjugate located at infinity, which means that the microscope objective images an object into infinity. An infinite conjugate is useful for introducing beam splitters or filters that should work with small incidence angles. Also, an infinity-corrected system accommodates additional components like DIC prisms, polarizers. etc. The collimated beam is focused to create an intermediate image with additional optics, called a **tube lens**. The tube lens either creates an image directly onto a CCD chip or an intermediate image, which is further reimaged with an eyepiece. Additionally, the tube lens might be used for system correction. For example, Zeiss corrects aberrations in its microscopes with a combination objective-tube lens.

In the case of an **infinity-corrected objective**, the transverse magnification can only be defined in the presence of a tube lens that will form a real image. It is given by the ratio between the tube lens's focal length and the focal length of the microscope objective.

Manufacturer	Focal Length of Tube Lens
Zeiss	164.5 mm
Olympus	180.0 mm
Nikon	200.0 mm
Leica	200.0 mm

Telecentricity of a Microscope

Telecentricity is a feature of an optical system where the principal ray in object, image, or both spaces is parallel to the optical axis. This means that the object or image does not shift laterally, even with defocus; the distance between two object or image points is constant along the optical axis.

An optical system can be **telecentric** in

- Object space, where the entrance pupil is at infinity, and the aperture stop is in the back focal plane;
- Image space, where the exit pupil is at infinity, and the aperture stop is in the front focal plane; or
- Both (doubly telecentric), where the entrance and exit pupils are at infinity, and the aperture stop is at the center of the system, in the back focal plane of the element before the stop and front focal length of the element after the stop (afocal system).

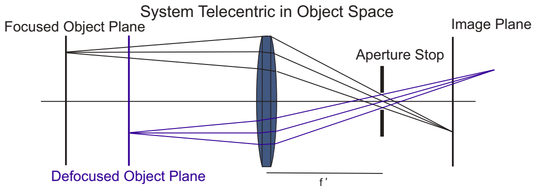
System Telecentric in Object Space
Focused Object Plane
Image Plane
Aperture Stop
Defocused Object Plane
f ′

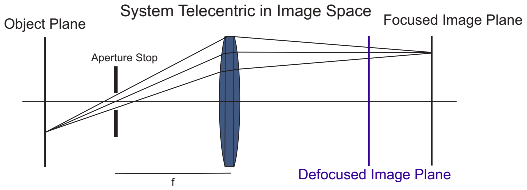
System Telecentric in Image Space
Object Plane
Focused Image Plane
Aperture Stop
Defocused Image Plane
f

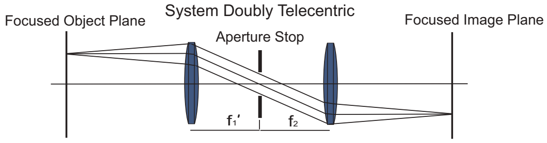
System Doubly Telecentric
Focused Object Plane
Focused Image Plane
Aperture Stop
f₁′
f₂

The aperture stop in a microscope is located at the back focal plane of the microscope objective. This makes the microscope objective telecentric in object space. Therefore, in microscopy, the object is observed with constant magnification, even for defocused object planes. This feature of microscopy systems significantly simplifies their operation and increases the reliability of image analysis.

Magnification of a Microscope

Magnifying power (MP) is defined as the ratio between the angles subtended by an object with and without magnification. The magnifying power (as defined for a single lens) creates an enlarged virtual image of an object. The angle of an object observed with magnification is

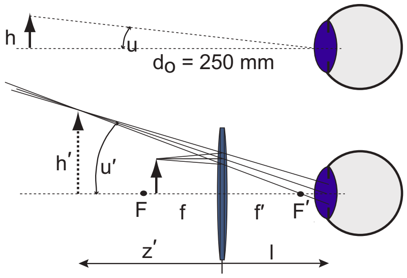
h
u
d_o = 250 mm
h′
u′
F
f
f′
F′
z′
l

$$u' = \frac{h'}{z'-l} = \frac{h(f-z')}{f(z'-l)}.$$

Therefore,

$$MP = \frac{u'}{u} = \frac{d_o(f-z')}{f(z'-l)}.$$

The angle for an unaided eye is defined for the **minimum focus distance** (d_o) of 10 inches or 250 mm, which is the distance that the object (real or virtual) may be examined without discomfort for the average population. A distance l between the lens and the eye is often small and can be assumed to equal zero:

$$MP \approx \frac{250\text{mm}}{f} - \frac{250\text{mm}}{z'}.$$

If the virtual image is at infinity (observed with a relaxed eye), $z' = -\infty$, and

$$MP \approx \frac{250\text{mm}}{f}.$$

The total magnifying power of the microscope results from the magnification of the microscope objective and the magnifying power of the eyepiece (usually 10×):

$$M_{\text{objective}} = \frac{OTL}{f_{\text{objective}}}$$

$$MP_{\text{microscope}} = M_{\text{objective}} MP_{\text{eyepiece}} = -\frac{OTL}{f_{\text{objective}}} \frac{250\text{mm}}{f_{\text{eyepiece}}}.$$

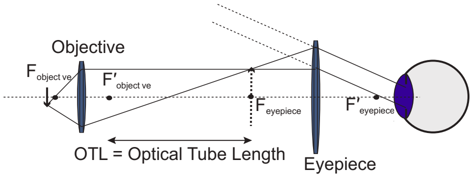
Objective
$F_{\text{objective}}$
$F'_{\text{objective}}$
F_{eyepiece}
F'_{eyepiece}
OTL = Optical Tube Length
Eyepiece

Numerical Aperture

The aperture diaphragm of the optical system determines the angle at which rays emerge from the axial object point and, after refraction, pass through the optical system. This acceptance angle is called the object space aperture angle. The parameter describing system throughput and this acceptance angle is called the **numerical aperture** (NA):

$$NA = n \sin u \,.$$

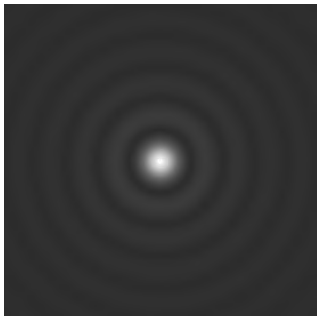

As seen from the equation, throughput of the optical system may be increased by using media with a high refractive index *n*, e.g., oil or water. This effectively decreases the refraction angles at the interfaces.

The dependence between the numerical aperture in the object space NA and the numerical aperture in the image space between the objective and the eyepiece NA′ is calculated using the objective magnification:

$$NA = NA' M_{\text{objective}} \,.$$

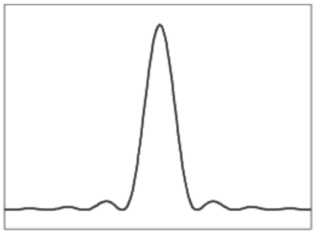

As a result of diffraction at the aperture of the optical system, self-luminous points of the object are not imaged as points but as so-called Airy disks. An **Airy disk** is a bright disk surrounded by concentric rings with gradually decreasing intensities. The disk diameter (where the intensity reaches the first zero) is

$$d = \frac{1.22\lambda}{n \sin u} = \frac{1.22\lambda}{NA} \,.$$

Media	Refractive Index
Air	1
Water	1.33
Oil	1.45–1.6 (1.515 is typical)

Note that the refractive index in the equation is for media between the object and the optical system.

Resolution Limit

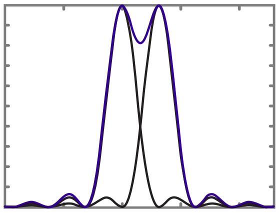

The lateral resolution of an optical system can be defined in terms of its ability to resolve images of two adjacent, self-luminous points. When two **Airy disks** are too close, they form a continuous intensity distribution and cannot be distinguished. The **Rayleigh resolution limit** is defined as occurring when the peak of the Airy pattern arising from one point coincides with the first minimum of the Airy pattern arising from a second point object. Such a distribution gives an intensity dip of 26%. The distance between the two points in this case is

$$d = \frac{0.61\lambda}{n \sin u} = \frac{0.61\lambda}{NA}.$$

The equation indicates that the resolution of an optical system improves with an increase in *NA* and decreases with increasing wavelength λ. For example, for $\lambda = 450$ nm (blue) and oil immersion $NA = 1.4$, the microscope objective can optically resolve points separated by less than 200 nm.

The situation in which the intensity dip between two adjacent self-luminous points becomes zero defines the **Sparrow resolution limit**. In this case $d = 0.5\lambda/NA$.

The **Abbe resolution limit** considers both the diffraction caused by the object and the *NA* of the optical system. It assumes that if at least two adjacent diffraction orders for points with spacing *d* are accepted by the objective, these two points can be resolved. Therefore, the resolution depends on both imaging and illumination apertures and is

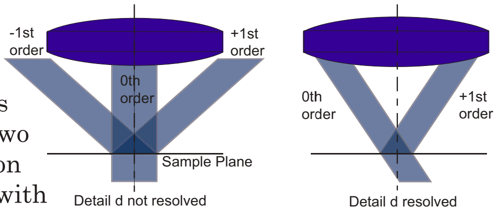

$$d = \frac{\lambda}{NA_{\text{objective}} + NA_{\text{condenser}}}.$$

Useful Magnification

For visual observation, the **angular resolving power** can be defined as 1.5 arc minutes. For unmagnified objects at a distance of 250 mm (the eye's minimum focus distance), 1.5 arc minutes converts to d_{eye} = 0.1 mm. Since microscope objectives by themselves do not usually provide sufficient magnification, they are combined with oculars or eyepieces. The resolving power of the microscope is then

$$d_{\text{mic}} = \frac{d_{\text{eye}}}{M_{\text{min}}} = \frac{d_{\text{eye}}}{M_{\text{obj}} M_{\text{eyepiece}}} .$$

In the Sparrow resolution limit, the minimum microscope magnification is

$$M_{\text{min}} = 2 d_{\text{eye}} NA / \lambda .$$

Therefore, a total minimum magnification M_{min} can be defined as approximately 250–500 NA (depending on wavelength). For lower magnification the image will appear brighter, but imaging is performed below the overall resolution limit of the microscope. For larger magnification, the contrast decreases and resolution does not improve. While a theoretically higher magnification should not provide additional information, it is useful to increase it to approximately 1000 NA to provide comfortable sampling of the object. Therefore, it is assumed that the **useful magnification** of a microscope is between 500 NA and 1000 NA. Usually, any magnification above 1000 NA is called **empty magnification**. The image size in such cases is enlarged, but no additional useful information is provided. The highest useful magnification of a microscope is approximately 1500 for an oil-immersion microscope objective with $NA = 1.5$.

Similar analysis can be performed for digital microscopy, which uses CCD or CMOS cameras as image sensors. Camera pixels are usually small (between 2 and 30 microns), and useful magnification must be estimated for a particular image sensor rather than the eye. Therefore, digital microscopy can work at lower magnification, and magnification of the microscope objective alone is usually sufficient.

Depth of Field and Depth of Focus

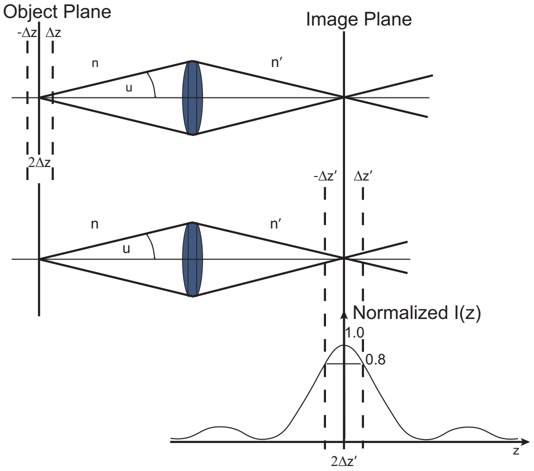

Two important terms are used to define axial resolution: **depth of field** and **depth of focus**. Depth of field refers to the object thickness for which an image is in focus, while depth of focus is the corresponding thickness of the image plane. In the case of a diffraction-limited optical system, the depth of focus is determined for an intensity drop along the optical axis to 80%, defined as

$$DOF = 2\Delta z' = \frac{n'\lambda}{NA'^{2}}.$$

The relation between depth of field ($2\Delta z$) and depth of focus ($2\Delta z'$) incorporates the objective magnification:

$$2\Delta z' = M^{2}_{\text{objective}}\, 2\Delta z \frac{n'}{n},$$

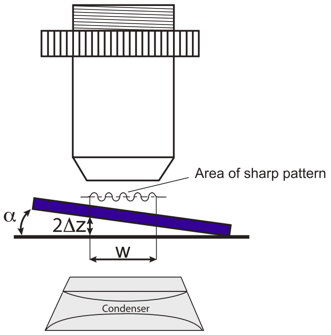

where n' and n are medium refractive indexes in the object and image space, respectively.

To quickly measure the depth of field for a particular microscope objective, the grid amplitude structure can be placed on the microscope stage and tilted with the known angle α. The depth of field is determined after measuring the grid zone w while in focus:

$$2\Delta z = nw\tan\alpha\,.$$

Magnification and Frequency vs Depth of Field

Depth of field for visual observation is a function of the total microscope magnification and the NA of the objective, approximated as

$$2\Delta z = n\frac{0.5\lambda}{NA^2} + n\frac{340}{M_{\text{microscope}}NA}\ .$$

Note that estimated values do not include eye accommodation. The graph below presents depth of field for visual observation. Refractive index n of the object space was assumed to equal 1. For other media, values from the graph must be multiplied by an appropriate n.

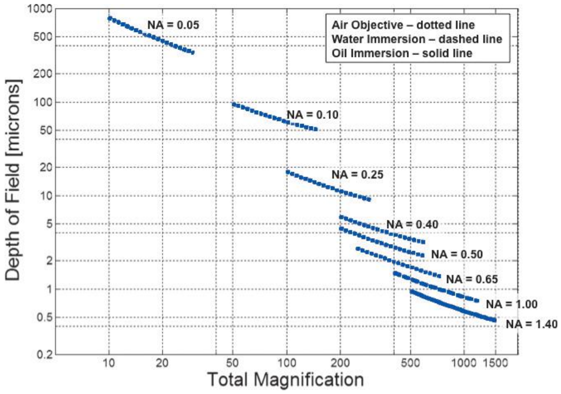

Depth of field can also be defined for a specific frequency present in the object because imaging contrast changes for a particular object frequency, as described by the modulation transfer function. The approximated equation is

$$2\Delta z = \frac{0.4}{\rho NA},$$

where ρ is the frequency in cycles per millimeter.

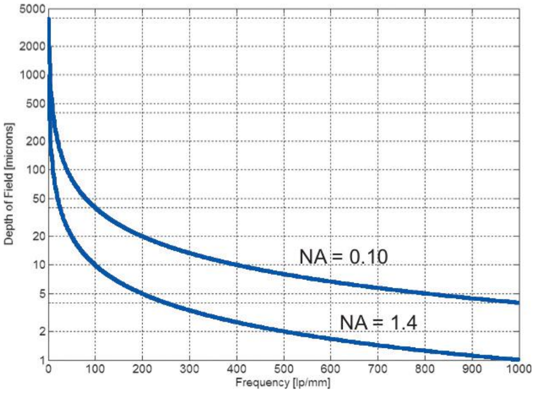

Köhler Illumination

One of the most critical elements in efficient microscope operation is proper set up of the illumination system. To assist with this, August Köhler introduced a solution that provides uniform and bright illumination over the field of view, even for light sources that are not uniform (e.g., a lamp filament). This system consists of a light source, a field-stop diaphragm, a condenser aperture, and collective and condenser lenses. This solution is now called **Köhler illumination** and is commonly used for a variety of imaging modes.

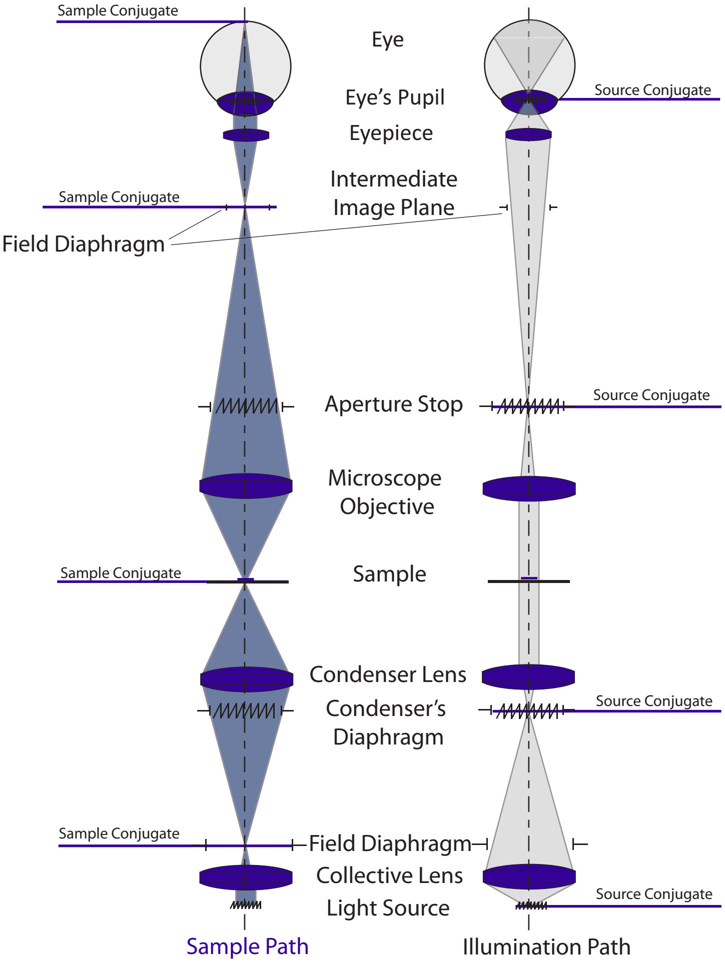
Sample Conjugate
Eye
Eye's Pupil
Source Conjugate
Eyepiece
Intermediate Image Plane
Sample Conjugate
Field Diaphragm
Aperture Stop
Source Conjugate
Microscope Objective
Sample Conjugate
Sample
Condenser Lens
Condenser's Diaphragm
Source Conjugate
Sample Conjugate
Field Diaphragm
Collective Lens
Light Source
Source Conjugate
Sample Path
Illumination Path

Köhler Illumination (cont.)

The illumination system is configured such that an image of the light source completely fills the condenser aperture.

Köhler illumination requires setting up the microscope system so that the field diaphragm, object plane, and intermediate image in the eyepiece's field stop, retina, or CCD are in conjugate planes. Similarly, the lamp filament, front aperture of the condenser, microscope objective's back focal plane (aperture stop), exit pupil of the eyepiece, and the eye's pupil are also at conjugate planes. Köhler illumination can be set up for both the transmission mode (also called **DIA**) and the reflectance mode (also called **EPI**).

The uniformity of the illumination is the result of having the light source at infinity with respect to the illuminated sample.

EPI illumination is especially useful for the biological and metallurgical imaging of thick or opaque samples.

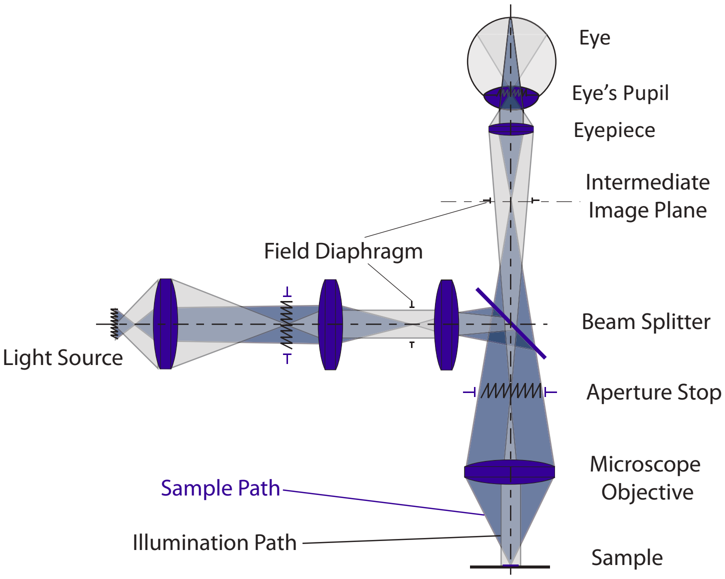
Eye
Eye's Pupil
Eyepiece
Intermediate Image Plane
Field Diaphragm
Beam Splitter
Light Source
Aperture Stop
Microscope Objective
Sample Path
Illumination Path
Sample

Alignment of Köhler Illumination

The procedure for **Köhler illumination** alignment consists of the following steps:

1. Place the sample on the microscope stage and move it so that the front surface of the condenser lens is 1–2 mm from the microscope slide. The condenser and collective diaphragms should be wide open.
2. Adjust the location of the source so illumination fills the condenser's diaphragm. The sharp image of the lamp filament should be visible at the condenser's diaphragm plane and at the back focal plane of the microscope objective. To see the filament at the back focal plane of the objective, a **Bertrand lens** can be applied (also see *Special Lens Components* on page 55). The filament should be centered in the aperture stop using the bulb's adjustment knobs on the illuminator.
3. For a low-power objective, bring the sample into focus. Since a microscope objective works at a parfocal distance, switching to a higher magnification later is easy and requires few adjustments.
4. Focus and center the condenser lens. Close down the field diaphragm, focus down the outline of the diaphragm, and adjust the condenser's position. After adjusting the x-, y-, and z-axis, open the field diaphragm so it accommodates the entire field of view.
5. Adjust the position of the condenser's diaphragm by observing the back focal objective plane through the Bertrand lens. When the edges of the aperture are sharply seen, the condenser's diaphragm should be closed to approximately three quarters of the objective's aperture.
6. Tune the brightness of the lamp. This adjustment should be performed by regulating the voltage of the lamp's power supply or, preferably, through the neutral density filters in the beam path. Do not adjust the brightness by closing the condenser's diaphragm because it affects the illumination setup and the overall microscope resolution.

Note that while three quarters of the aperture stop is recommended for initial illumination, adjusting the aperture of the illumination system affects the resolution of the microscope. Therefore, the final setting should be adjusted after examining the images.

Critical Illumination

An alternative to Köhler illumination is **critical illumination**, which is based on imaging the light source directly onto the sample. This type of illumination requires a highly uniform light source. Any source non-uniformities will result in intensity variations across the image. Its major advantage is high efficiency, since it can collect a larger solid angle than Köhler illumination and therefore provides a higher energy density at the sample. For parabolic or elliptical reflective collectors, critical illumination can utilize up to 85% of the light emitted by the source.

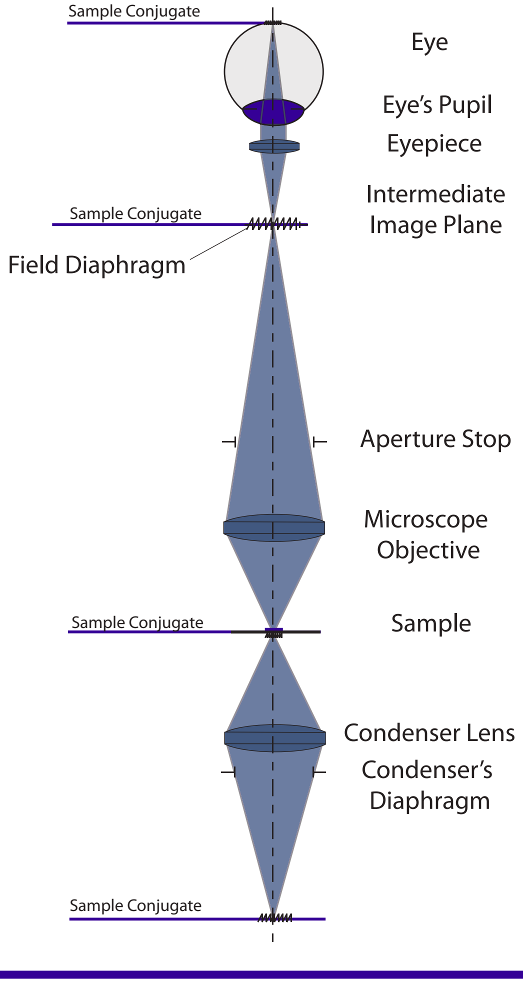
Sample Conjugate
Eye
Eye's Pupil
Eyepiece
Intermediate Image Plane
Sample Conjugate
Field Diaphragm
Aperture Stop
Microscope Objective
Sample Conjugate
Sample
Condenser Lens
Condenser's Diaphragm
Sample Conjugate

Stereo Microscopes

Stereo microscopes are built to provide depth perception, which is important for applications like micro-assembly and biological and surgical imaging. Two primary stereo-microscope approaches involve building two separate tilted systems, or using a common objective combined with a binocular system.

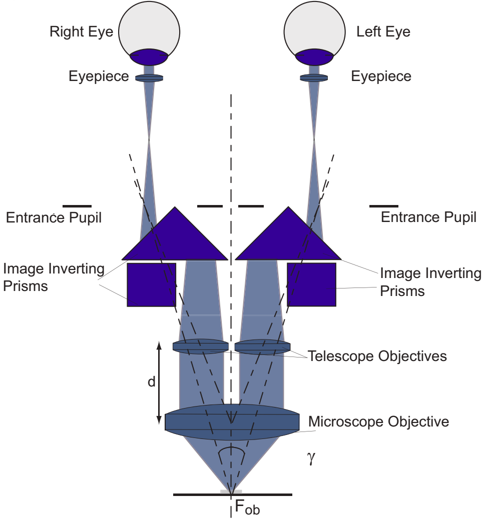
Right Eye
Left Eye
Eyepiece
Eyepiece
Entrance Pupil
Entrance Pupil
Image Inverting Prisms
Image Inverting Prisms
Telescope Objectives
d
Microscope Objective
γ
F_{ob}

In the latter, the angle of convergence γ of a stereo microscope depends on the focal length of a microscope objective and a distance d between the microscope objective and telescope objectives. The entrance pupils of the microscope are images of the stops (located at the plane of the telescope objectives) through the microscope objective.

Depth perception δz can be defined as

$$\delta z = \frac{250\alpha_s}{M_{\text{microscope}} \tan\gamma} [\text{mm}],$$

where α_s is a **visual stereo resolving power**, which for daylight vision is approximately 5–10 arc seconds.

Stereo microscopes have a convergence angle γ in the range of 10–15 deg. Note that γ is 15 deg for visual observation and 0 for a standard microscope. For example, depth perception for the human eye is approximately 0.05 mm (at 250 mm), while for a stereo microscope of $M_{\text{microscope}} = 100$ and $\gamma = 15$ deg, it is $\delta z = 0.5$ μm.

Eyepieces

The eyepiece relays an intermediate image into the eye, corrects for some remaining objective aberrations, and provides measurement capabilities. It also needs to relay an exit pupil of a microscope objective onto the entrance pupil of the eye. The location of the relayed exit pupil is called the **eye point** and needs to be in a comfortable observation distance behind the eyepiece. The clearance between the mechanical mounting of the eyepiece and the eye point is called **eye relief**. Typical eye relief is 7–12 mm.

An eyepiece usually consists of two lenses: one (closer to the eye) which magnifies the image, and a second working as a collective lens and also responsible for the location of the exit pupil of the microscope. An eyepiece contains a field stop that provides a sharp image edge.

Parameters like **magnifying power** of an eyepiece and **field number** (FN), i.e., field of view of an eyepiece, are engraved on an eyepiece's barrel: M × / FN. Field number and magnification of a microscope objective allow quick calculation of the imaged area of a sample (see also *The Finite Tube Length Microscope* on p. 34). Field number varies with microscopy vendors and eyepiece magnification. For 10× or lower magnification eyepieces, it is usually 20–28 mm while for higher magnification, wide-angle oculars can get down to approximately 5 mm.

The majority of eyepieces are **Huygens**, **Ramsden**, or derivations of them. The Huygens eyepiece consists of two plano-convex lenses with convex surfaces facing the microscope objective:

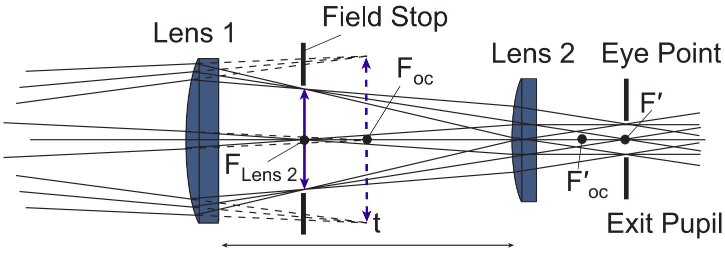

Eyepieces (cont.)

Both lenses are usually made with crown glass and allow for some correction of lateral color, coma, and astigmatism. To correct for color, the following conditions should be met:

$$f_1 \approx f_2 \quad \text{and} \quad t \approx 1.5 f_2 .$$

For higher magnifications, the exit pupil moves toward the ocular, making observation less convenient, and so this eyepiece is used only for low magnifications (≤10×). In addition, Huygens oculars effectively correct for chromatic aberrations, and therefore can be more effectively used with lower-end microscope objectives (e.g., achromatic).

The Ramsden eyepiece consists of two plano-convex lenses with convex surfaces facing each other. Both focal lengths are very similar, and the distance between lenses is smaller than f_2:

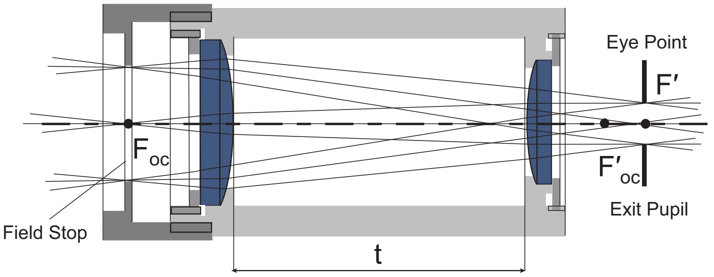

The field stop is placed in the front focal plane of the eyepiece. A collective lens does not participate in creating an intermediate image, so the Ramsden eyepiece works as a simple magnifier:

$$f_1 = f_2 \quad \text{and} \quad t < f_2 .$$

Compensating eyepieces work in conjunction with microscope objectives to correct for lateral color (apochromatic).

High-eye point oculars provide an extended distance between the last mechanical surface and the exit pupil of the eyepiece. They allow users with glasses to comfortably use a microscope. The convenient high eye-point location should be 20–25 mm behind the eyepiece.

Nomenclature and Marking of Objectives

Objective parameters include:

Objective correction—such as Achromat, Plan Achromat, Fluor, Plan Fluor, Apochromat (Apo), and Plan Apochromat.

Magnification—the lens magnification for a finite tube length, or for a microscope objective in combination with a tube lens. In infinity-corrected systems, magnification depends on the ratio between the focal lengths of a tube lens and a microscope objective. A particular objective type should be used only in combination with the proper tube lens.

Magnification	Zeiss Code
1×, 1.25×	Black
2.5×	Khaki
4×, 5×	Red
6.3×	Orange
10×	Yellow
16×, 20×, 25×, 32×	Green
25×, 32×	Green
40×, 50×	Light Blue
63×	Dark Blue
> 100×	White

Application—specialized use or design of an objective, e.g., H (bright field), D (dark field), DIC (differential interference contrast), RL (reflected light), PH (phase contrast), or P (polarization).

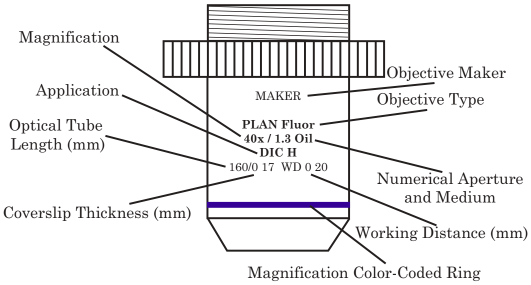
Magnification
Application
Optical Tube Length (mm)
Coverslip Thickness (mm)
MAKER
PLAN Fluor
40x / 1.3 Oil
DIC H
160/0 17 WD 0 20
Objective Maker
Objective Type
Numerical Aperture and Medium
Working Distance (mm)
Magnification Color-Coded Ring

Tube length—an infinity corrected (∞) or finite tube length in mm.

Cover slip—the thickness of the cover slip used (in mm). "0" or "–" means no cover glass or the cover slip is optional, respectively.

Numerical aperture (NA) and medium—defines system throughput and resolution. It depends on the media between the sample and objective. The most common media are air (no marking), oil (Oil), water (W), or Glycerine.

Working distance (WD)—the distance in millimeters between the surface of the front objective lens and the object.

Objective Designs

Achromatic objectives (also called **achromats)** are corrected at two wavelengths: 656 nm and 486 nm. Also, spherical aberration and coma are corrected for green (546 nm), while astigmatism is very low. Their major problems are a secondary spectrum and field curvature.

Achromatic objectives are usually built with a lower NA (below 0.5) and magnification (below 40×). They work well for white light illumination or single wavelength use. When corrected for field curvature, they are called plan-achromats.

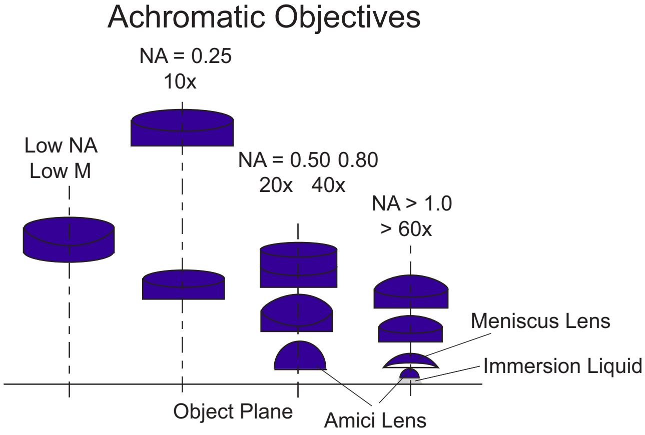

Fluorites or **semi-apochromats** have similar color correction as **achromats**; however, they correct for spherical aberration for two or three colors. The name “fluorites” was assigned to this type of objective due to the materials originally used to build this type of objective. They can be applied for higher NA (e.g., 1.3) and magnifications, and used for applications like immunofluorescence, polarization, or differential interference contrast microscopy.

The most advanced microscope objectives are **apochromats**, which are usually chromatically corrected for three colors, with spherical aberration correction for at least two colors. They are similar in construction to fluorites but with different thicknesses and surface figures. With the correction of field curvature, they are called plan-apochromats. They are used in fluorescence microscopy with multiple dyes and can provide very high NA (1.4). Therefore, they are suitable for low-light applications.

Objective Designs (cont.)

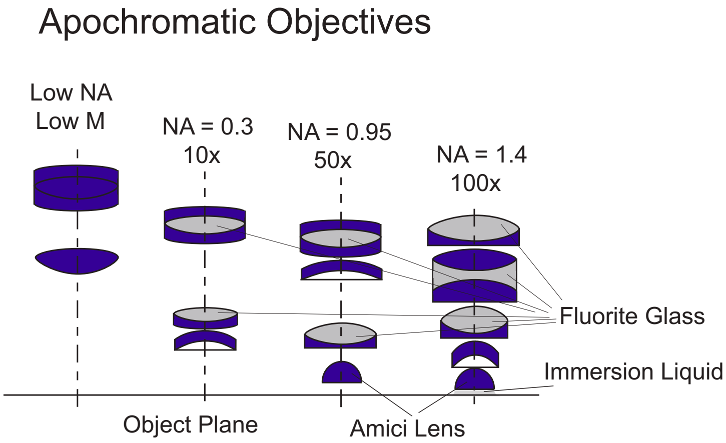
Apochromatic Objectives
Low NA
Low M
NA = 0.3
10x
NA = 0.95
50x
NA = 1.4
100x
Fluorite Glass
Immersion Liquid
Object Plane
Amici Lens

Type	Number of wavelengths for spherical correction	Number of colors for chromatic correction
Achromat	1	2
Fluorite	2–3	2–3
Plan-Fluorite	2–4	2–4
Plan-Apochromat	2–4	3–5

(Adapted from Murphy, 2001 and http://www.microscopyu.com/)

Examples of typical objective parameters are shown in the table below:

M	Type	Medium	*WD*	NA	*d*	DOF
10	Achromat	Air	4.4	0.25	1.34	8.80
20	Achromat	Air	0.53	0.45	0.75	2.72
40	Fluorite	Air	0.50	0.75	0.45	0.98
40	Fluorite	Oil	0.20	1.30	0.26	0.49
60	Apochromat	Air	0.15	0.95	0.35	0.61
60	Apochromat	Oil	0.09	1.40	0.24	0.43
100	Apochromat	Oil	0.09	1.40	0.24	0.43
Refractive index of oil is $n = 1.515$						

(Adapted from Murphy, 2001)

Special Objectives and Features

Special types of objectives include long working distance objectives, ultra-low-magnification objectives, water-immersion objectives, and UV lenses.

Long working distance objectives allow imaging through thick substrates like culture dishes. They are also developed for interferometric applications in Michelson and Mirau configurations. Alternatively, they can allow for the placement of instrumentation (e.g., micropipettes) between a sample and an objective. To provide a long working distance and high NA, a common solution is to use reflective objectives, or extend the working distance of a standard microscope objective by using a reflective attachment. While reflective objectives have the advantage of being free of chromatic aberrations, their serious drawback is a smaller field of view relative to refractive objectives. The central obscuration also decreases throughput by 15%–20%.

LWD Reflective Objective

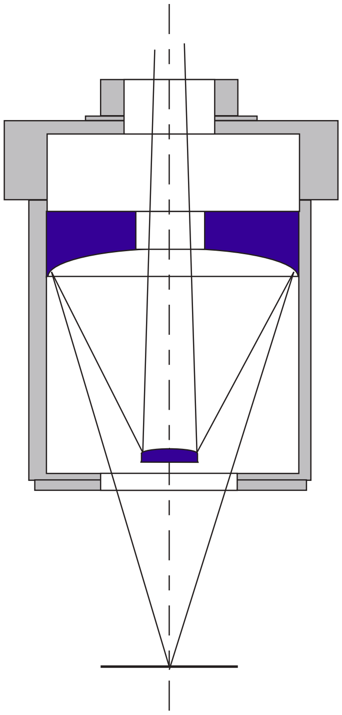

Special Objectives and Features (cont.)

Low magnification objectives can achieve magnifications as low as 0.5. However, in some cases they are not fully compatible with microscopy systems, and may not be telecentric in the image space. They also often require special tube lenses and special condensers to provide Köhler illumination.

Water immersion objectives are increasingly common, especially for biological imaging, because they provide a high NA and avoid toxic immersion oils. They usually work without a cover slip.

UV objectives are made using UV transparent, low-dispersion materials such as quartz. These objectives enable imaging at wavelengths from the near-UV through the visible spectrum, e.g., 240–700 nm. Reflective objectives can also be used for the IR bands.

Reflective Adapter to extend working distance WD

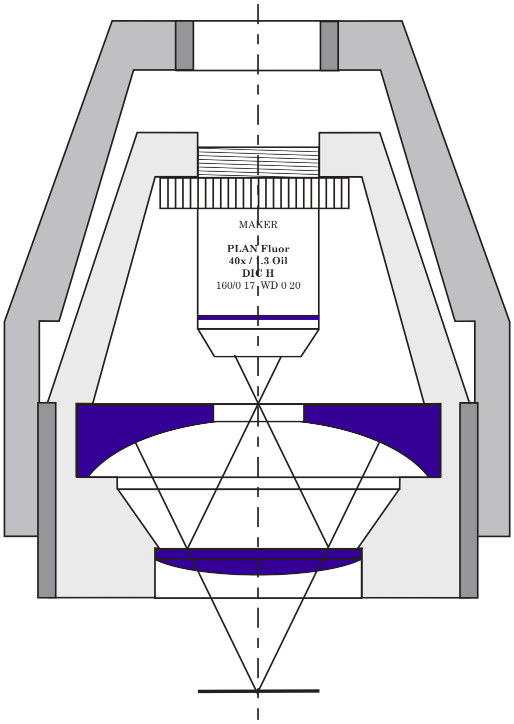

Special Lens Components

The **Bertrand lens** is a focusable eyepiece telescope that can be easily placed in the light path of the microscope. This lens is used to view the back aperture of the objective, which simplifies microscope alignment, specifically when setting up Köhler illumination.

To construct a high-numerical-aperture microscope objective with good correction, numerous optical designs implement the **Amici lens** as a first component of the objective. It is a thick lens placed in close proximity to the sample. An Amici lens usually has a plane (or nearly plane) first surface and a large-curvature spherical second surface. To ensure good chromatic aberration correction, an achromatic lens is located closely behind the Amici lens. In such configurations, microscope objectives usually reach an NA of 0.5–0.7.

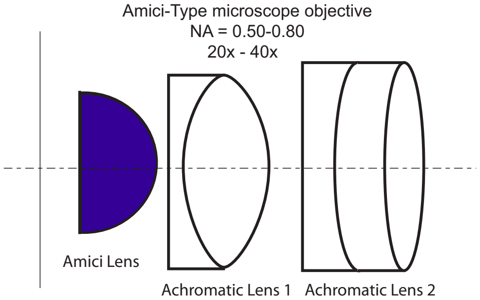
Amici-Type microscope objective
NA = 0.50-0.80
20x - 40x
Amici Lens
Achromatic Lens 1
Achromatic Lens 2

To further increase the NA, an Amici lens is improved by either cementing a high-refractive-index meniscus lens to it or placing a meniscus lens closely behind the Amici lens. This makes it possible to construct well-corrected high magnification (100×) and high numerical aperture (NA > 1.0) oil-immersion objective lenses.

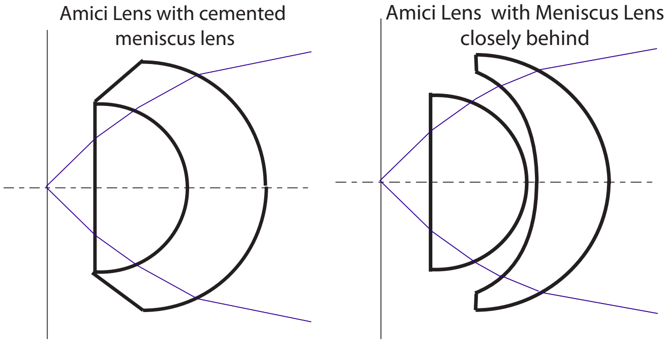
Amici Lens with cemented meniscus lens
Amici Lens with Meniscus Lens closely behind

Cover Glass and Immersion

The **cover slip** is located between the object and the microscope objective. It protects the imaged sample and is an important element in the optical path of the microscope. Cover glass can reduce imaging performance and cause spherical aberration, while different imaging angles experience movement of the object point along the optical axis toward the microscope objective. The object point moves closer to the objective with an increase in angle.

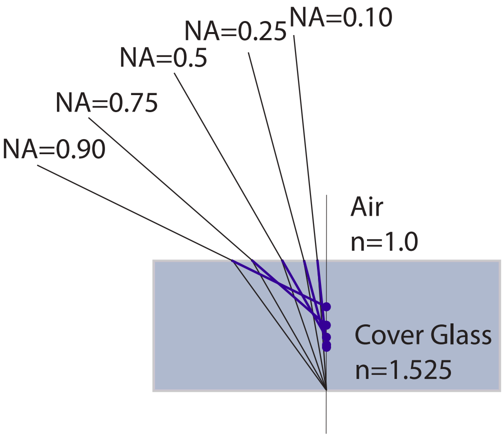

The importance of the cover glass increases in proportion to the objective's NA, especially in high-NA dry lenses. For lenses with an NA smaller than 0.5, the type of cover glass may not be a critical parameter, but should always be properly used to optimize imaging quality. Cover glasses are likely to be encountered when imaging biological samples. Note that the presence of a standard-thickness cover glass is accounted for in the design of high-NA objectives. The important parameters that define a cover glass are its thickness, index of refraction, and Abbe number. The ISO standards for refractive index and Abbe number are $n = 1.5255 \pm 0.0015$ and $V = 56 \pm 2$, respectively.

Microscope objectives are available that can accommodate a range of cover-slip thicknesses. An adjustable collar allows the user to adjust for cover slip thickness in the range from 100 microns to over 200 microns.

Cover Glass Grade Number	Thickness (in microns)
0	83–130
1	130–160
1.5	160–190
2	190–250

Cover Glass and Immersion (cont.)

The table below presents a summary of acceptable cover glass thickness deviations from 0.17 mm and the allowed thicknesses for Zeiss air objectives with different NAs.

NA of the Objective	Allowed Thickness Deviations (from 0.17 mm)	Allowed Thickness Range
<0.3	–	0.000–0.300
0.30–0.45	0.07	0.100–0.240
0.45–0.55	0.05	0.120–0.220
0.55–0.65	0.03	0.140–0.200
0.65–0.75	0.02	0.150–0.190
0.75–0.85	0.01	0.160–0.180
0.85–0.95	0.005	0.165–0.175

(Adapted from Pluta, 1988)

To increase both the microscope's resolution and system throughput, immersion liquids are used between the cover-slip glass and the microscope objective, or directly between the sample and objective. An **immersion liquid** effectively increases the NA and decreases the angle of refraction at the interface between the medium and the microscope objective. Common immersion liquids include oil (n = 1.515), water (n = 1.34), and glycerin (n = 1.48).

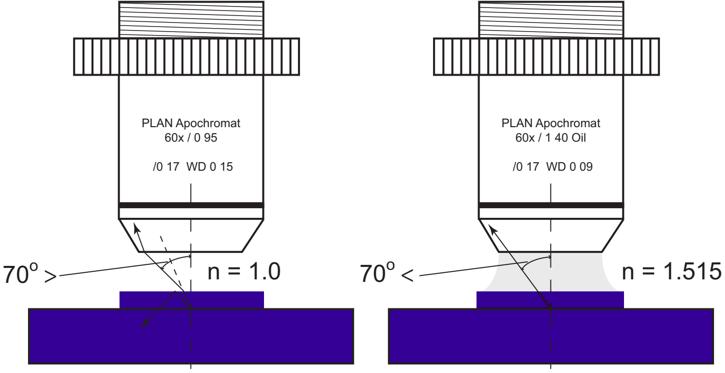

Water, which is more common, or glycerin immersion objectives are mainly used for biological samples, such as living cells or a tissue culture. They provide a slightly lower NA than oil objectives but are free of the toxicity associated with immersion oils. Water immersion is usually applied without a cover slip. For fluorescent applications it is also critical to use non-fluorescent oil.

Common Light Sources for Microscopy

Common light sources for microscopy include **incandescent lamps**, such as tungsten-argon and tungsten (e.g., quartz halogen) lamps. A **tungsten-argon lamp** is primarily used for bright-field, phase-contrast, and some polarization imaging. **Halogen lamps** are inexpensive and a convenient choice for a variety of applications that require a continuous and bright spectrum.

Arc lamps are usually brighter than incandescent lamps. The most common examples include xenon (XBO) and mercury (HBO) arc lamps. Arc lamps provide high-quality monochromatic illumination if combined with the appropriate filter. Arc lamps are more difficult to align, are more expensive, and have a shorter lifetime. Their spectral range starts in the UV range and continuously extends through visible to the infrared. About 20% of the output is in the visible spectrum, while the majority is in the UV and IR. The usual lifetime is 100–200 hours.

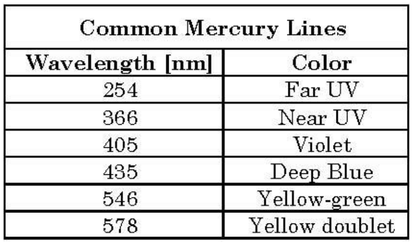

Common Mercury Lines	
Wavelength [nm]	**Color**
254	Far UV
366	Near UV
405	Violet
435	Deep Blue
546	Yellow-green
578	Yellow doublet

Another light source is the **gas-arc discharge lamp**, which includes mercury, xenon, and halide lamps. The **mercury lamp** has several strong lines, which might be 100 times brighter than the average output. About 50% is located in the UV range, and should be used with protective glasses. For imaging biological samples, the proper selection of filters is necessary to protect living cell samples and micro-organisms (e.g., UV-blocking filters/cold mirrors). The xenon arc lamp has a uniform spectrum, can provide an output power greater than 100 W, and is often used for fluorescence imaging. However, over 50% of its power falls into the IR; therefore, IR-blocking filters (hot mirrors) are necessary to prevent the overheating of samples.

Metal halide lamps were recently introduced as high-power sources (over 150 W) with lifetimes several times longer than arc lamps. While in general the metal-halide lamp has a spectral output similar to that of a mercury arc lamp, it extends further into longer wavelengths.

LED Light Sources

Light-emitting diodes (LEDs) are a new, alternative light source for microscopy applications. The LED is a semiconductor diode that emits photons when in forward-biased mode. Electrons pass through the depletion region of a p-n junction and lose an amount of energy equivalent to the bandgap of the semiconductor. The characteristic features of LEDs include a long lifetime, a compact design, and high efficiency. They also emit narrowband light with relatively high energy.

Wavelengths [nm] of High-power LEDs Commonly Used in Microscopy		**Total Beam Power [mW] (approximate)**
455	Royal Blue	225–450
470	Blue	200–400
505	Cyan	150–250
530	Green	100–175
590	Amber	15–25
633	Red	25–50
435–675	White Light	200–300

An important feature of LEDs is the ability to combine them into arrays and custom geometries. Also, LEDs operate at lower temperatures than arc lamps, and due to their compact design they can be cooled easily with simple heat sinks and fans.

The radiance of currently available LEDs is still significantly lower than that possible with arc lamps; however, LEDs can produce an acceptable fluorescent signal in bright microscopy applications. Also, the pulsed mode can be used to increase the radiance by 20 times or more.

LED Spectral Range [nm]	**Semiconductor**
350–400	GaN
400–550	$In_{1-x}Ga_xN$
550–650	$Al_{1-x-y}In_yGa_xP$
650–750	$Al_{1-x}Ga_xAs$
750–1000	$GaAs_{1-x}P_x$

Filters

Neutral density (ND) filters are neutral (wavelength wise) gray filters scaled in units of **optical density** (OD):

$$OD = \log_{10}\left(\frac{1}{\tau}\right),$$

where τ is the transmittance. ND filters can be combined to provide OD as a sum of the ODs of the individual filters. ND filters are used to change light intensity without tuning the light source, which can result in a spectral shift.

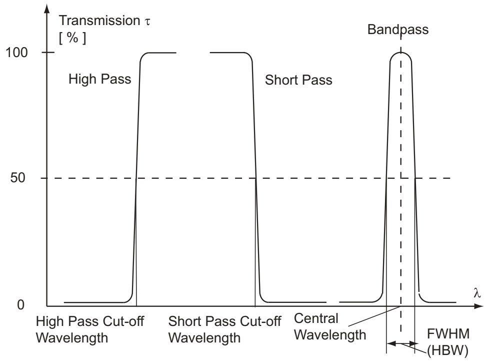

Color **absorption filters** and **interference filters** (also see *Multiple Wave Interference* on page 16) isolate the desired range of wavelengths: e.g. bandpass and edge filters.

Edge filters include short-pass and long-pass filters. Short-pass filters allow short wavelengths to pass and stop long wavelengths, and long-pass filters allow long wavelengths to pass while stopping short wavelengths. Edge filters are defined for wavelengths with a 50% drop in transmission.

Bandpass filters allow only a selected spectral bandwidth to pass and are characterized with a central wavelength and full-width-half-maximum (FWHM) defining the spectral range for transmission of at least 50%.

Color absorption glass filters usually serve as broad bandpass filters. They are less costly and less susceptible to damage than interference filters.

Interference filters are based on multiple beam interference in thin films. They combine between three to over 20 dielectric layers of $\lambda/2$ and $\lambda/4$, separated by metallic coatings. They can provide a sharp bandpass transmission down to the sub-nm range and full-width-half-maximum of 10–20 nm.

Polarizers and Polarization Prisms

Polarizers are built with birefringent crystals using polarization at multiple reflections, selective absorption (dichroism), form birefringance, or scattering.

For example, polarizers built with birefringent crystals (Glan-Thompson prisms) use the principle of total internal reflection to eliminate ordinary or extraordinary components (for positive or negative crystals).

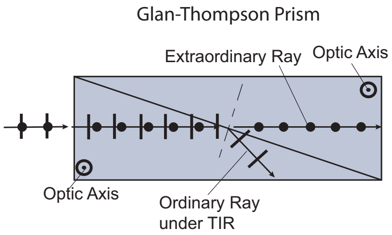
Glan-Thompson Prism
Extraordinary Ray
Optic Axis
Optic Axis
Ordinary Ray
under TIR

Birefringent prisms are crucial components for numerous microscopy techniques, e.g., differential interference contrast. The most common birefringent prism is a **Wollaston prism**. It splits light into two beams with orthogonal polarization and propagates under different angles.

The angle between propagating beams is

$$\varepsilon = 2|n_e - n_o|\tan\alpha .$$

Both beams produce interference with fringe period b:

$$b = \frac{\lambda}{\varepsilon} = \frac{\lambda}{2|n_e - n_o|\tan\alpha} .$$

The localization plane of fringes is

$$\gamma = \frac{1}{2}\left(\frac{1}{n_e} + \frac{1}{n_o}\right)\alpha .$$

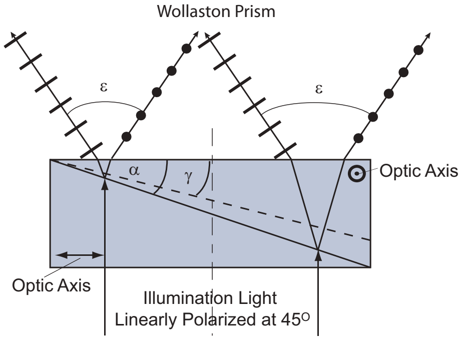
Wollaston Prism
ε
ε
α
γ
Optic Axis
Optic Axis
Illumination Light
Linearly Polarized at 45°

Polarizers and Polarization Prisms (cont.)

The fringe localization plane tilt can be compensated by using two symmetrical Wollaston prisms.

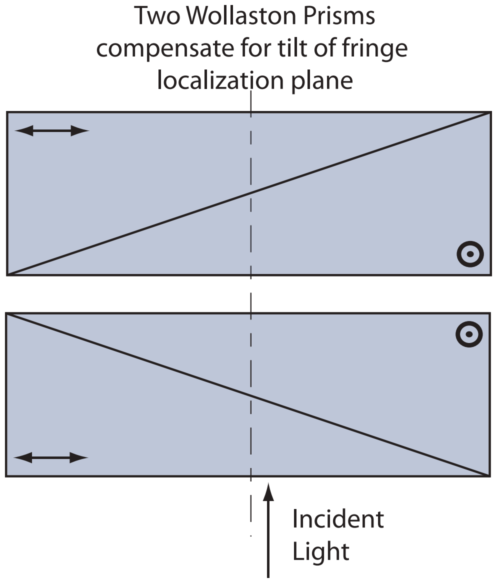
Two Wollaston Prisms
compensate for tilt of fringe
localization plane
Incident
Light

Wollaston prisms have a fringe localization plane inside the prism. One example of a modified Wollaston prism is the **Nomarski prism**, which simplifies the DIC microscope setup by shifting the plane outside the prism, i.e., the prism does not need to be physically located at the condenser front focal plane or the objective's back focal plane.

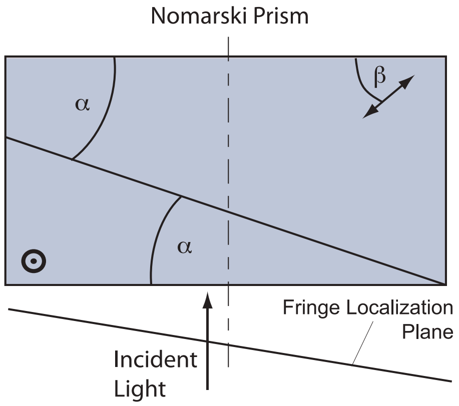
Nomarski Prism
α
β
α
Fringe Localization
Plane
Incident
Light

Amplitude and Phase Objects

The major object types encountered on the microscope are **amplitude** and **phase** objects. The type of object often determines the microscopy technique selected for imaging.

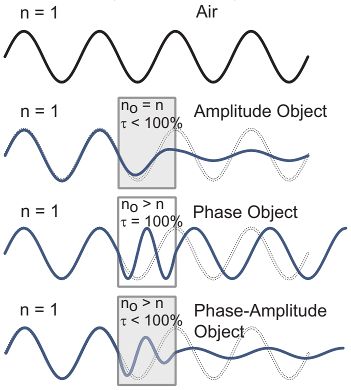

The amplitude object is defined as one that changes the amplitude and therefore the intensity of transmitted or reflected light. Such objects are usually imaged with bright-field microscopes. A stained tissue slice is a common amplitude object.

Phase objects do not affect the optical intensity; instead, they generate a phase shift in the transmitted or reflected light. This phase shift usually arises from an inhomogeneous refractive index distribution throughout the sample, causing differences in optical path length (OPL). Mixed amplitude-phase objects are also possible (e.g., biological media), which affect the amplitude and phase of illumination in different proportions.

Classifying objects as **self-luminous** and **non-self-luminous** is another way to categorize samples. Self-luminous objects generally do not directly relate their amplitude and phase to the illuminating light. This means that while the observed intensity may be proportional to the illumination intensity, its amplitude and phase are described by a statistical distribution (e.g., fluorescent samples). In such cases, one can treat discrete object points as secondary light sources, each with their own amplitude, phase, coherence, and wavelength properties.

Non-self-luminous objects are those that affect the illuminated beam in such a manner that discrete object points cannot be considered as entirely independent. In such cases, the wavelength and temporal coherence of the illuminating source needs to be considered in imaging. A diffusive or absorptive sample is an example of such an object.

The Selection of a Microscopy Technique

Microscopy provides several imaging principles. Below is a list of the most common techniques and object types:

Technique	Type of sample
Bright-field	Amplitude specimens, reflecting specimens, diffuse objects
Dark-field	Light-scattering objects
Phase contrast	Phase objects, light-scattering objects, light-refracting objects, reflective specimens
Differential interference contrast (DIC)	Phase objects, light-scattering objects, light-refracting objects, reflective specimens
Polarization microscopy	Birefringent specimens
Fluorescence microscopy	Fluorescent specimens
Laser scanning, Confocal microscopy, and Multi-photon microscopy	3D samples requiring optical sectioning, fluorescent and scattering samples
Super-resolution microscopy (**RESOLFT**, **4Pi**, **I^{5}M**, **SI**, **STORM**, **PALM**, and others)	Imaging at the molecular level; imaging primarily focuses on fluorescent samples where the sample is a part of an imaging system
Raman microscopy, **CARS**	Contrast-free chemical imaging
Array microscopy	Imaging of large FOVs
SPIM	Imaging of large 3D samples
Interference Microscopy	Topography, refractive index measurements, 3D coherence imaging

All of these techniques can be considered for transmitted or reflected light. Below are examples of different sample types:

Sample type	Sample Example
Amplitude specimens	Naturally colored specimens, stained tissue
Specular specimens	Mirrors, thin films, metallurgical samples, integrated circuits
Diffuse objects	Diatoms, fibers, hairs, micro-organisms, minerals, insects
Phase objects	Bacteria, cells, fibers, mites, protozoa
Light-refracting samples	Colloidal suspensions, minerals, powders
Birefringent specimens	Mineral sections, crystallized chemicals, liquid crystals, fibers, single crystals
Fluorescent specimens	Cells in tissue culture, fluorochrome-stained sections, smears and spreads

(Adapted from www.microscopyu.com)

Image Comparison

The images below display four important microscope modalities: **bright field**, **dark field**, **phase contrast**, and **differential interference contrast**. They also demonstrate the characteristic features of these methods. The images are of a blood specimen viewed with the Zeiss upright Axiovert Observer Z1 microscope. The microscope was set with an objective at 40×, NA = 0.6, Ph2, LD Plan Neofluor, and a cover slip glass of 0.17 mm. The pictures were taken with a monochromatic CCD camera.

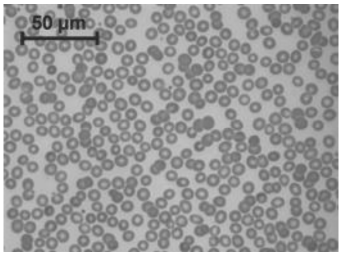

Bright Field

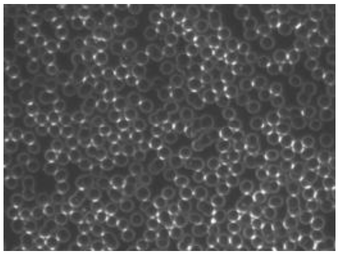

Dark Field

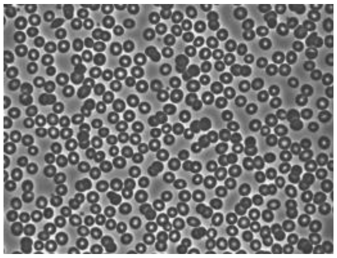

Phase Contrast

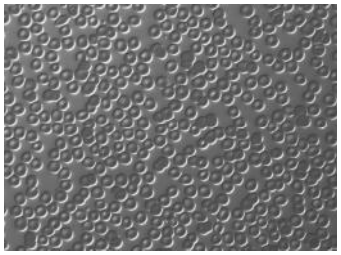

Differential Interference Contrast (DIC)

The bright-field image relies on absorption and shows the sample features with decreasing amounts of passing light. The dark-field image only shows the scattering sample components. Both the phase contrast and the differential interference contrast demonstrate the optical thickness of the sample. The characteristic halo effect is visible in the phase-contrast image. The 3D effect of the DIC image arises from the differential character of images; they are formed as a derivative of the phase changes in the beam as it passes through the sample.

Phase Contrast

Phase contrast is a technique used to visualize phase objects by phase and amplitude modifications between the direct beam propagating through the microscope and the beam diffracted at the phase object. An object is illuminated with monochromatic light, and a phase-delaying (or advancing) element in the aperture stop of the objective introduces a phase shift, which provides interference contrast. Changing the amplitude ratio of the diffracted and non-diffracted light can also increase the contrast of the object features.

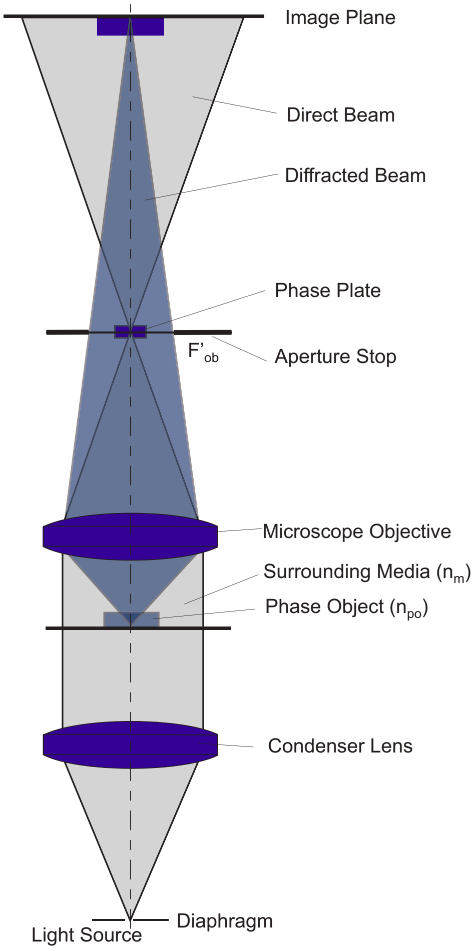
Image Plane
Direct Beam
Diffracted Beam
Phase Plate
F'$_{ob}$
Aperture Stop
Microscope Objective
Surrounding Media (n_m)
Phase Object (n_{po})
Condenser Lens
Diaphragm
Light Source

Phase Contrast (cont.)

Phase contrast is primarily used for imaging living biological samples (immersed in a solution), unstained samples (e.g., cells), thin film samples, and mild phase changes from mineral objects. In that regard, it provides qualitative information about the optical thickness of the sample. Phase contrast can also be used for some quantitative measurements, like an estimation of the refractive index or the thickness of thin layers. Its phase detection limit is similar to the **differential interference contrast** and is approximately $\lambda/100$. On the other hand, phase contrast (contrary to DIC) is limited to thin specimens, and phase delay should not exceed the depth of field of an objective. Other drawbacks compared to DIC involve its limited optical sectioning ability and undesired effects like halos or shading-off. (See also *Characteristic Features of Phase Contrast Microscopy* on page 71).

The most important advantages of phase contrast (over DIC) include its ability to image birefringent samples and its simple and inexpensive implementation into standard microscopy systems.

Presented below is a mathematical description of the phase contrast technique based on a vector approach. The phase shift in the figure (See page 68) is represented by the orientation of the vector. The length of the vector is proportional to amplitude of the beam. When using standard imaging on a transparent sample, the length of the light vectors passing through sample PO and surrounding media SM is the same, which makes the sample invisible. Additionally, vector PO can be considered as a sum of the vectors passing through surrounding media SM and diffracted at the object DP.

$$PO = SM + DP$$

If the wavefront propagating through the surrounding media can be the subject of an exclusive phase change (diffracted light DP is not affected), the vector SM is rotated by an angle corresponding to the phase change. This exclusive phase shift is obtained with a small circular or ring phase plate located in the plane of the aperture stop of a microscope.

Phase Contrast (cont.)

Consequently, vector PO, which represents light passing through a phase sample, will change its value to PO' and provide contrast in the image:

$$PO' = SM' + DP,$$

where SM' represents the rotated vector SM.

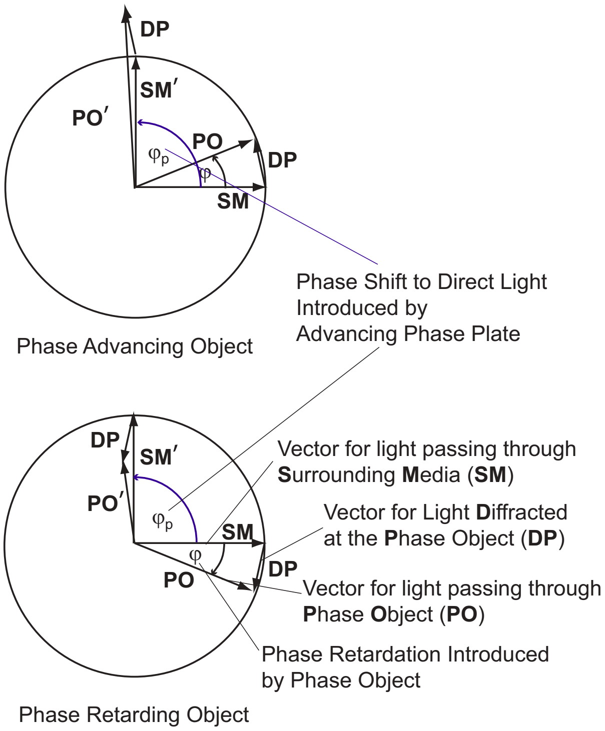
DP
SM′
PO′
PO
φp
φ
DP
SM
Phase Shift to Direct Light
Introduced by
Advancing Phase Plate
Phase Advancing Object
DP
SM′
PO′
φp
SM
φ
DP
PO
Vector for light passing through
Surrounding Media (SM)
Vector for Light Diffracted
at the Phase Object (DP)
Vector for light passing through
Phase Object (PO)
Phase Retardation Introduced
by Phase Object
Phase Retarding Object

Phase samples are considered to be **phase retarding objects** or **phase advancing objects** when their refractive index is greater or less than the refractive index of the surrounding media, respectively.

Phase plate	Object Type	Object Appearance
$\varphi_p = +\pi/2$ (+90°)	phase-retarding	brighter
$\varphi_p = +\pi/2$ (+90°)	phase-advancing	darker
$\varphi_p = -\pi/2$ (–90°)	phase-retarding	darker
$\varphi_p = -\pi/2$ (–90°)	phase-advancing	brighter

Visibility in Phase Contrast

Visibility of features in **phase contrast** can be expressed as

$$C_{\text{ph}} = \frac{I_{\text{media}} - I_{\text{object}}}{I_{\text{media}}} = \frac{|SM'|^2 - |PO'|^2}{|SM'|^2}.$$

This equation defines the phase object's visibility as a ratio between the intensity change due to phase features and the intensity of surrounding media $|SM'|^2$. It defines the negative and positive contrast for an object appearing brighter or darker than the media, respectively. Depending on the introduced phase shift (positive or negative), the same object feature may appear with a **negative** or **positive contrast**. Note that C_{ph} relates to classical contrast C as

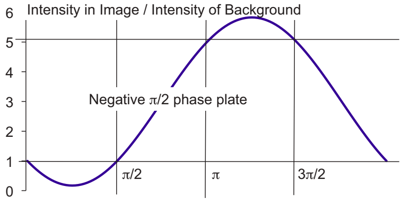

$$C = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} = \frac{|I_1 - I_2|}{I_1 + I_2} = |C_{\text{ph}}| \frac{|SM'|^2}{|SM'|^2 + |PO'|^2}.$$

Contrast	Phase Plate
-2φ	$\varphi_p = +\pi/2$ (+90°)
$+2\varphi$	$\varphi_p = -\pi/2$ (−90°)

For phase changes in the 0–2π range, intensity in the image can be found using vector relations; small phase changes in the object $\varphi << 90$ deg can be approximated as $\pm 2\varphi$.

To increase the contrast of images, the intensity in the direct beam is additionally changed with beam attenuation τ in the phase ring. τ is defined as a transmittance, $\tau = 1/N$, where N is a dividing coefficient of intensity in a direct beam (intensity is decreased N times). The contrast in this case is

$$C_{\text{ph}} = -2\varphi / \sqrt{\tau} = -2\varphi\sqrt{N}$$

for a +π/2 phase plate, and

$$C_{\text{ph}} = +2\varphi / \sqrt{\tau} = +2\varphi\sqrt{N}$$

for −π/2 phase plate. The minimum perceived phase difference with phase contrast is

$$\Delta\varphi_{\text{min}} = \frac{C_{\text{ph-min}}\lambda}{4\pi\sqrt{N}}.$$

$C_{\text{ph-min}}$ is usually accepted at the contrast value of 0.02.

The Phase Contrast Microscope

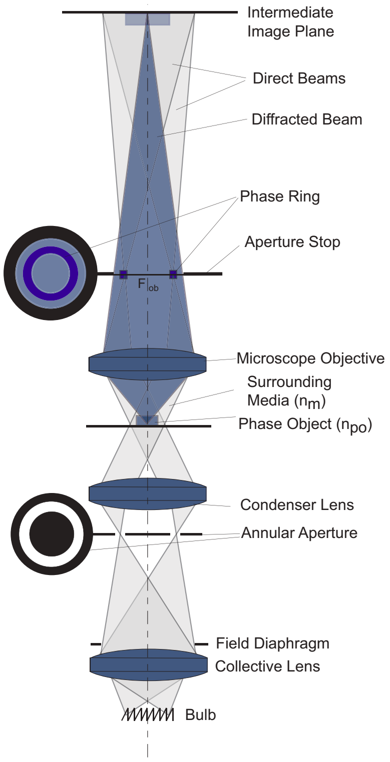
Intermediate Image Plane
Direct Beams
Diffracted Beam
Phase Ring
Aperture Stop
F_ob
Microscope Objective
Surrounding Media (n_m)
Phase Object (n_po)
Condenser Lens
Annular Aperture
Field Diaphragm
Collective Lens
Bulb

The common phase-contrast system is similar to the bright-field microscope but with two modifications:

1. The condenser diaphragm is replaced with an annular aperture diaphragm.
2. A ring-shaped phase plate is introduced at the back focal plane of the microscope objective. The ring may include metallic coatings, typically providing 75% to 90% attenuattion of the direct beam.

Both components are in conjugate planes and are aligned such that the image of the annular condenser diaphragm overlaps with the objective phase ring. While there is no direct access to the objective phase ring, the anular aperture in front of the condenser is easily accessible.

The phase shift introduced by the ring is given by

$$\varphi_{\mathrm{p}} = 2\pi \frac{OPD}{\lambda} = \frac{2\pi}{\lambda}\left(n_{\mathrm{m}} - n_{\mathrm{r}}\right)t\,,$$

where n_m and n_r are refractive indices of the media surrounding the phase ring and the ring itself, respectively. t is the physical thickness of the phase ring.

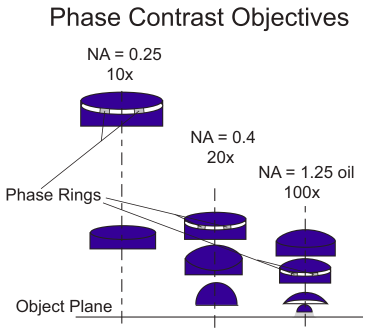
Phase Contrast Objectives
NA = 0.25
10x
NA = 0.4
20x
NA = 1.25 oil
100x
Phase Rings
Object Plane

Characteristic Features of Phase Contrast

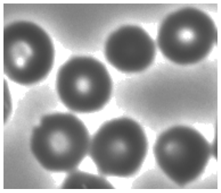

Images in phase contrast are dark or bright features on a background (positive and negative contrast, respectively). They contain undesired image effects called halo and shading-off, which are a result of the incomplete separation of direct and diffracted light. The **halo effect** is a phase contrast feature that increases light intensity around sharp changes in the phase gradient.

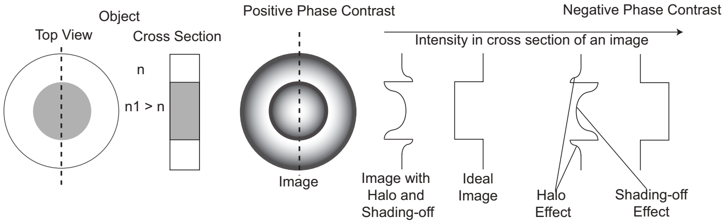
Object
Top View
Cross Section
n
n1 > n
Positive Phase Contrast
Image
Negative Phase Contrast
Intensity in cross section of an image
Image with Halo and Shading-off
Ideal Image
Halo Effect
Shading-off Effect

The **shading-off effect** is an increase or decrease (for dark or bright images, respectively) of the intensity of the phase sample feature.

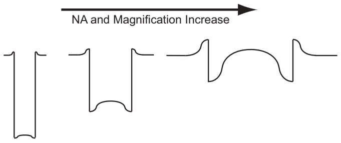
NA and Magnification Increase

Both effects strongly increase with an increase in numerical aperture and magnification. They can be reduced by surrounding the sides of a phase ring with ND filters.

Lateral resolution of the phase contrast technique is affected by the annular aperture (the radius of phase ring r_{PR}) and the aperture stop of the objective (the radius of aperture stop is r_{AS}). It is

$$d = f'_{\text{objective}} \frac{\lambda}{(r_{\text{AS}} + r_{\text{PR}})},$$

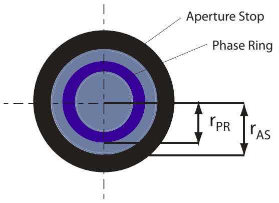
Aperture Stop
Phase Ring
r_{PR}
r_{AS}

compared to the resolution limit for a standard microscope:

$$d = f'_{\text{objective}} \frac{\lambda}{r_{\text{AS}}}.$$

Amplitude Contrast

Amplitude contrast changes the contrast in the images of absorbing samples. It has a layout similar to phase contrast; however, there is no phase change introduced by the object. In fact, in many cases a phase-contrast microscope can increase contrast based on the principle of amplitude contrast. The visibility of images is modified by changing the intensity ratio between the direct and diffracted beams.

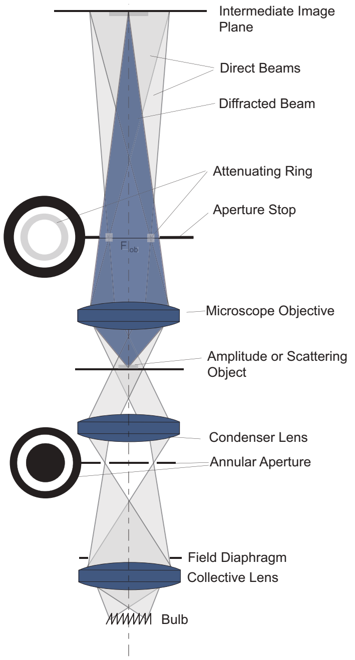
Intermediate Image Plane
Direct Beams
Diffracted Beam
Attenuating Ring
Aperture Stop
Fob
Microscope Objective
Amplitude or Scattering Object
Condenser Lens
Annular Aperture
Field Diaphragm
Collective Lens
Bulb

A vector schematic of the technique is as follows:

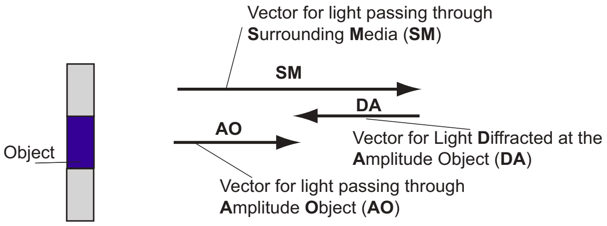
Vector for light passing through Surrounding Media (SM)
SM
DA
AO
Object
Vector for Light Diffracted at the Amplitude Object (DA)
Vector for light passing through Amplitude Object (AO)

Similar to visibility in phase contrast, image contrast can be described as a ratio of the intensity change due to amplitude features surrounding the media's intensity:

$$C_{\text{ac}} = \frac{I_{\text{media}} - I_{\text{object}}}{I_{\text{media}}} = \frac{2|SM||DA| - |DA|^2}{|SM|^2}.$$

Since the intensity of diffraction for amplitude objects is usually small, the contrast equation can be approximated with $C_{\text{ac}} = 2|DA|/|SM|$. If a direct beam is further attenuated, contrast C_{ac} will increase by factor of $N^{-1} = 1/\tau^{-1}$.

Oblique Illumination

Oblique illumination (also called **anaxial illumination**) is used to visualize phase objects. It creates pseudo-profile images of transparent samples. These reliefs, however, do not directly correspond to the actual surface profile.

The principle of oblique illumination can be explained by using either refraction or diffraction and is based on the fact that sample features of various spatial frequencies will have different intensities in the image due to a nonsymmetrical system layout and the filtration of spatial frequencies.

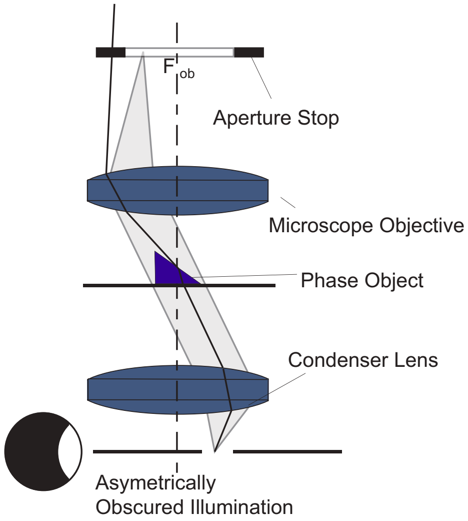

In practice, oblique illumination can be achieved by obscuring light exiting a condenser; translating the sub-stage diaphragm of the condenser lens does this easily.

The advantage of oblique illumination is that it improves spatial resolution over Köhler illumination (up to two times). This is based on the fact that there is a larger angle possible between the 0th and 1st orders diffracted at the sample for oblique illumination. In Köhler illumination, due to symmetry, the 0th and –1st orders at one edge of the stop will overlap with the 0th and +1st order on the other side. However, the gain in resolution is only for one sample direction; to examine the object features in two directions, a sample stage should be rotated.

Modulation Contrast

Modulation contrast microscopy (MCM) is based on the fact that phase changes in the object can be visualized with the application of multilevel gray filters.

The intensities of light refracted at the object are displayed with different values since they pass through different zones of the filter located in the stop of the microscope. MCM is often configured for oblique illumination since it already provides some intensity variations for phase objects. Therefore, the resolution of the MCM changes between normal and oblique illumination.

MCM can be obtained through a simple modification of a bright-field microscope by adding a slit diaphragm in front of the condenser and amplitude filter (also called the modulator) in the aperture stop of the microscope objective. The modulator filter consists of three areas: 100% (bright), 15% (gray), and 1% (dark). This filter acts in a similar way to the knife-edge in the Schlieren imaging techniques.

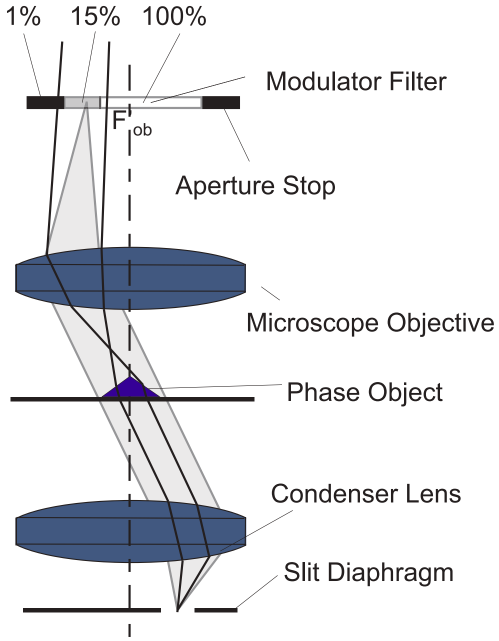
1%
15%
100%
Modulator Filter
F_{ob}
Aperture Stop
Microscope Objective
Phase Object
Condenser Lens
Slit Diaphragm

Hoffman Contrast

A specific implementation of MCM is **Hoffman modulation contrast**, which uses two additional polarizers located in front of the slit aperture. One polarizer is attached to the slit and obscures 50% of its width. A second can be rotated, which provides two intensity zones in the slit (for crossed polarizers, half of the slit is dark and half is bright; the entire slit is bright for the parallel position).

The major advantages of this technique include imaging at a full spatial resolution and a minimum depth of field, which allows optical sectioning. The method is inexpensive and easy to implement. The main drawback is that the images deliver a pseudo-relief image, which cannot be directly correlated with the object's form.

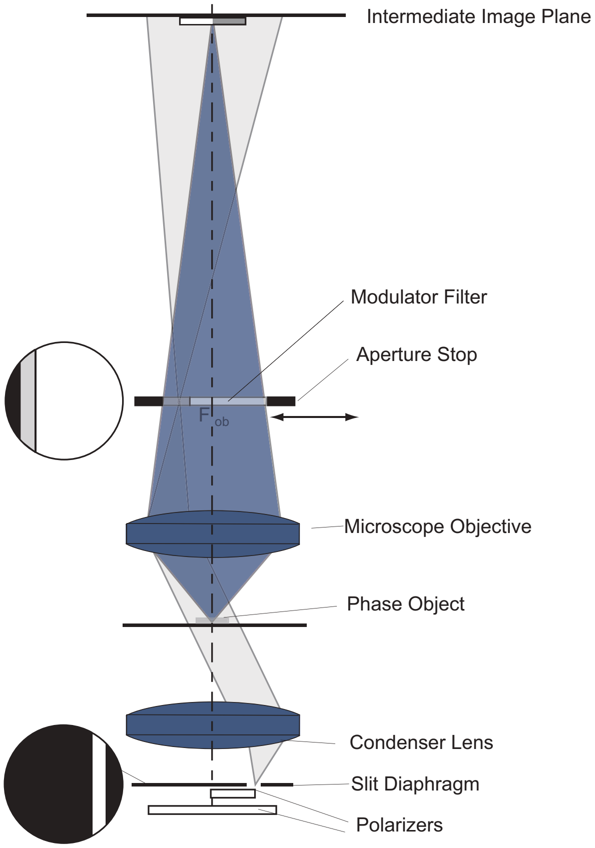
Intermediate Image Plane
Modulator Filter
Aperture Stop
F ob
Microscope Objective
Phase Object
Condenser Lens
Slit Diaphragm
Polarizers

Dark Field Microscopy

In **dark-field microscopy**, the specimen is illuminated at such angles that direct light is not collected by the microscope objective. Only diffracted and scattered light are used to form the image.

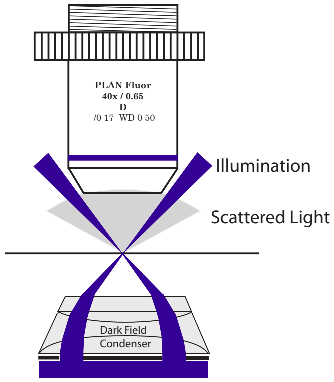
PLAN Fluor
40x / 0.65
D
/0 17 WD 0 50
Illumination
Scattered Light
Dark Field Condenser

The key component of the dark-field system is the condenser. The simplest approach uses an annular condenser stop and a microscope objective with an NA below that of the illumination cone. This approach can be used for objectives with NA < 0.65. Higher-NA objectives may incorporate an adjustable iris to reduce their NA below 0.65 for dark-field imaging. To image at a higher NA, specialized reflective dark-field condensers are necessary. Examples of such condensers include the **cardioid** and **paraboloid** designs, which enable dark-field imaging at an NA up to 0.9–1.0 (with oil immersion).

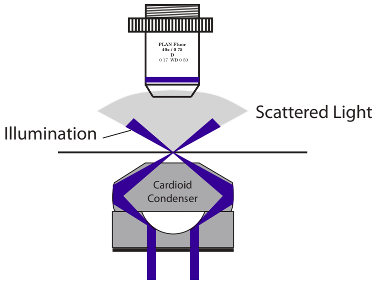
Scattered Light
Illumination
Cardioid Condenser

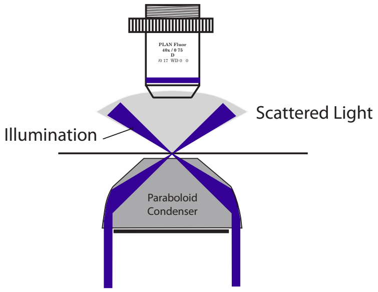
Scattered Light
Illumination
Paraboloid Condenser

For dark-field imaging in epi-illumination, objective designs consist of external illumination, and internal imaging paths are used.

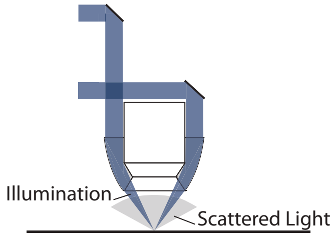
Illumination
Scattered Light

Optical Staining: Rheinberg Illumination

Optical staining is a class of techniques that uses optical methods to provide color differentiation between different areas of a sample**.**

Common Color Combinations	
Inner Filter	Annular Filter
Violet	Red
Blue	Yellow-Orange
Green	Red

Rheinberg illumination is a modification of dark-field illumination. Instead of an annular aperture, it uses two-zone color filters located in the front focal plane of the condenser. The overall diameter of the outer annular color filter is slightly greater than that corresponding to the NA of the system:

$$2r_{\lambda 2} > 2NAf'_{\text{condenser}} .$$

To provide good contrast between scattered and direct light, the inner filter is darker. Rheinberg illumination provides images that are a combination of two colors. Scattering features in one color are visible on the background of the other color.

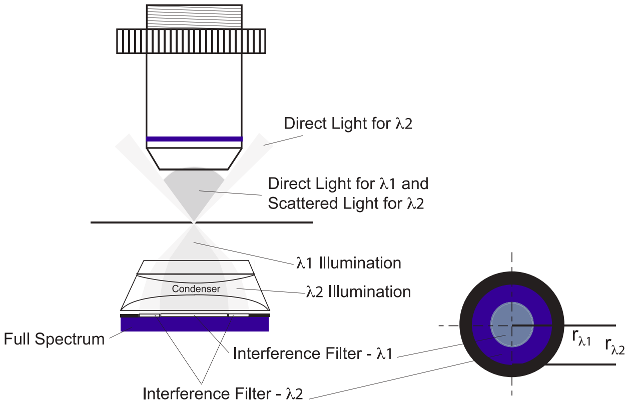
Direct Light for λ2
Direct Light for λ1 and Scattered Light for λ2
λ1 Illumination
Condenser
λ2 Illumination
Full Spectrum
Interference Filter - λ1
Interference Filter - λ2
$r_{\lambda 1}$
$r_{\lambda 2}$

Color-stained images can also be obtained in a simplified setup where only one (inner) filter is used. For such an arrangement, images will present white-light scattering features on the colored background. Other deviations from the above technique include double illumination where transmittance and reflectance are separated into two different colors.

Optical Staining: Dispersion Staining

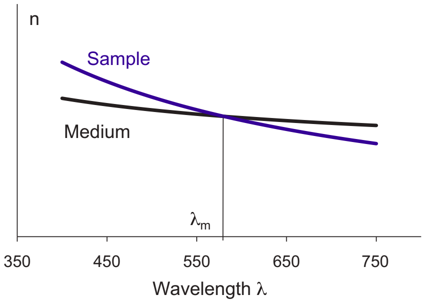

Dispersion staining is based on using highly dispersive media that matches the refractive index of the phase sample for a single wavelength λ_m (the matching wavelength). The system is configured in a setup similar to dark-field or oblique illumination. Wavelengths close to the matching wavelength will miss the microscope objective, while the other will refract at the object and create a colored dark-field image. Various configurations include illumination with a dark-field condenser and an annular aperture or central stop at the back focal plane of the microscope objective.

Color	λ Range [nm]	λ Mean [nm]
Violet	400–450	440
Blue	450–480	470
Blue-Green	480–510	495
Green	510–550	530
Green-Yellow	550–570	560
Yellow	570–590	580
Orange	590–630	605
Red	630–700	650

(Adapted from Pluta, 1989)

The annular-stop technique uses low-magnification (10×) objectives with a stop built as an opaque screen with a central opening, and is used for bright-field dispersion staining. The annular stop absorbs red and blue wavelengths and allows yellow through the system. A sample appears as white with yellow borders. The image background is also white.

The central-stop system absorbs unrefracted yellow light and direct white light. Sample features show up as purple (a mixture of red and blue) borders on a dark background. Different media and sample types will modify the color's appearance.

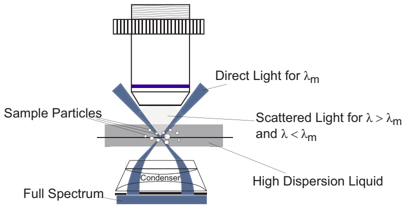
Direct Light for λ_m
Sample Particles
Scattered Light for $\lambda > \lambda_m$ and $\lambda < \lambda_m$
Condenser
High Dispersion Liquid
Full Spectrum

Shearing Interferometry: The Basis for DIC

Differential interference contrast (DIC) microscopy is based on the principles of shearing interferometry. Two wavefronts propagating through an optical system containing identical phase distributions (introduced by the sample) are slightly shifted to create an interference pattern. The phase of the resulting fringes is directly proportional to the derivative of the phase distribution in the object, hence the name "differential." Specifically, the intensity of the fringes relates to the derivative of the phase delay through the object (φ_o) and the local delay between wavefronts (Δ_b):

$$I \propto I_{\max} \cos^2\left(\Delta_b \pm \Delta\varphi_o\right)$$

or

$$I \propto I_{\max} \cos^2\left(\Delta_b \pm s\frac{d\varphi_o}{dx}\right),$$

where *s* denotes the shear between wavefronts, and Δ_b is the axial delay.

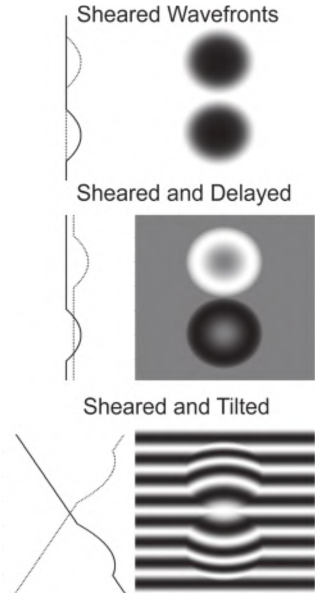

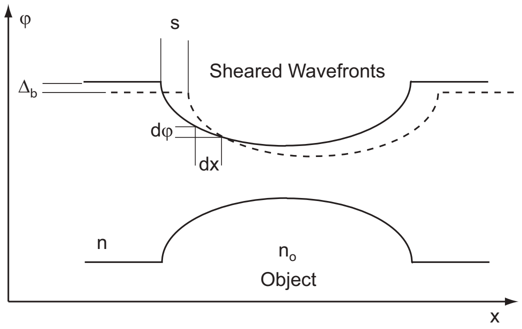

Wavefront shear is commonly introduced in DIC microscopy by incorporating a Mach-Zehnder interferometer into the system (e.g., Zeiss) or by the use of birefringent prisms. The appearance of DIC images depends on the sample orientation with respect to the shear direction.

DIC Microscope Design

The most common **differential interference contrast** (**Nomarski DIC**) design uses two **Wollaston** or **Nomarski birefringent prisms**. The first prism splits the wavefront into ordinary and extraordinary components to produce interference between these beams. Fringe localization planes for both prisms are in conjugate planes. Additionally, the system uses a crossed polarizer and analyzer. The polarizer is located in front of prism I, and the analyzer is behind prism II. The polarizer is rotated by 45 deg with regard to the shear axes of the prisms.

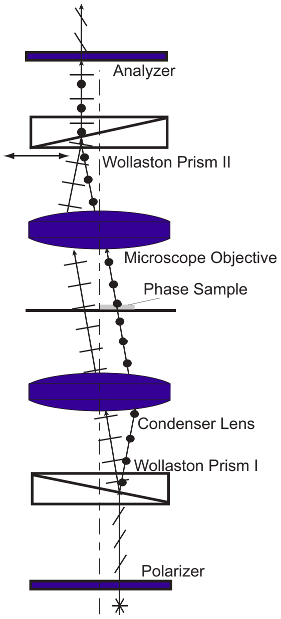

If prism II is centrally located, the intensity in the image is

$$I \propto \sin^2\left(\frac{\pi s \dfrac{d\varphi_o}{dx}}{\lambda}\right).$$

For a translated prism, phase **bias** Δ_b is introduced, and the intensity is proportional to

$$I \propto \sin^2 \frac{\pi\left(\Delta_b \pm s\dfrac{d\varphi_o}{dx}\right)}{\lambda}.$$

The sign in the equation depends on the direction of the shift. Shear s is

$$s = \frac{s'}{M_{\text{objective}}} = \varepsilon \frac{OTL}{M_{\text{objective}}},$$

where ε is the angular shear provided by the birefringent prisms, s' is the shear in the image space, and OTL is the optical tube length. The high spatial coherence of the source is obtained by the slit located in front of the condenser and oriented perpendicularly to the shear direction. To assure good interference contrast, the width of the slit w should be

$$w = \frac{\lambda f'_{\text{condenser}}}{4s}.$$

Low-strain (low-birefringence) objectives are crucial for high-quality DIC.

Appearance of DIC Images

In practice, shear between wavefronts in **differential interference contrast** is small and accounts for a fraction of a fringe period. This provides the differential character of the phase difference between interfering beams introduced to the interference equation.

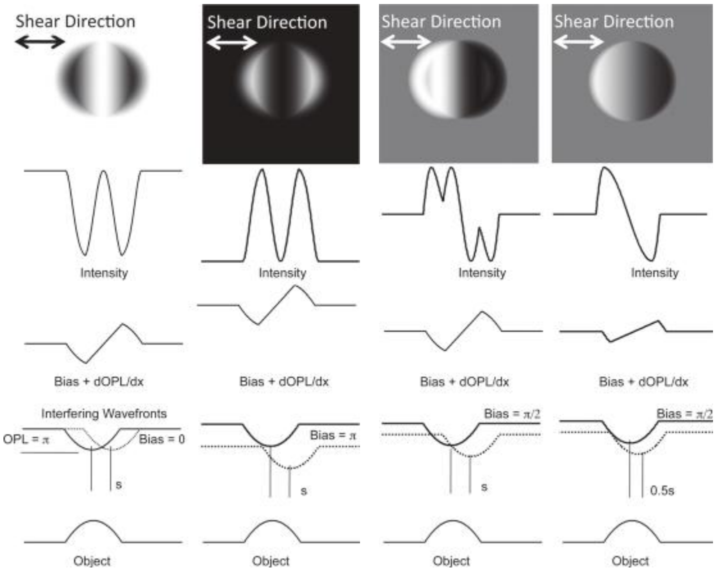
Shear Direction
Shear Direction
Shear Direction
Shear Direction
Intensity
Intensity
Intensity
Intensity
Bias + dOPL/dx
Bias + dOPL/dx
Bias + dOPL/dx
Bias + dOPL/dx
Interfering Wavefronts
OPL = π
Bias = 0
s
Bias = π
s
Bias = π/2
s
Bias = π/2
0.5s
Object
Object
Object
Object

Increased shear makes the edges of the object less sharp, while increased bias introduces some background intensity. Shear direction determines the appearance of the object.

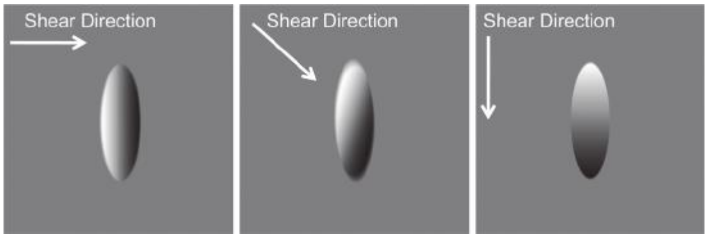
Shear Direction
Shear Direction
Shear Direction

Compared to phase contrast, DIC allows for larger phase differences throughout the object and operates in full resolution of the microscope (i.e., it uses the entire aperture). The depth of field is minimized, so DIC allows optical sectioning.

Reflectance DIC

A **Nomarski interference microscope** (also called a polarization interference contrast microscope) is a reflectance DIC system developed to evaluate roughness and the surface profile of specular samples. Its applications include metallography, microelectronics, biology, and medical imaging. It uses one Wollaston or Nomarski prism, a polarizer, and an analyzer. The information about the sample is obtained for one direction parallel to the gradient in the object. To acquire information for all directions, the sample should be rotated.

If white light is used, different colors in the image correspond to different sample slopes. It is possible to adjust the color by rotating the polarizer.

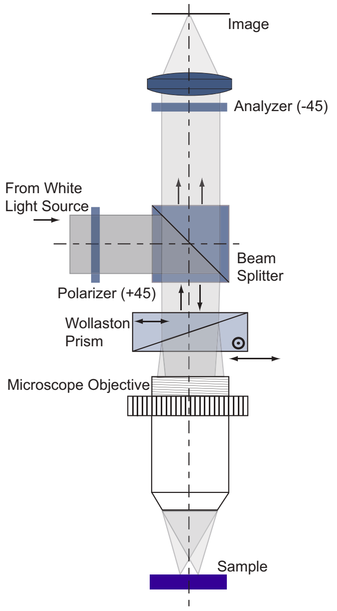
Image
Analyzer (-45)
From White
Light Source
Beam
Splitter
Polarizer (+45)
Wollaston
Prism
Microscope Objective
Sample

Phase delay (**bias**) can be introduced by translating the birefringent prism. For imaging under monochromatic illumination, the brightness in the image corresponds to the slope in the sample. Images with no bias will appear as bright features on a dark background; with bias they will have an average background level, and slopes will vary between dark and bright. This way, it is possible to interpret the direction of the slope. Similar analysis can be performed for the colors from white-light illumination.

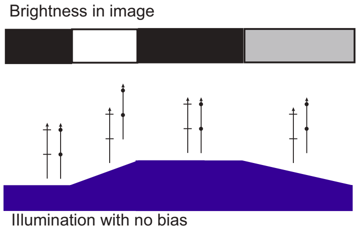
Brightness in image
Illumination with no bias

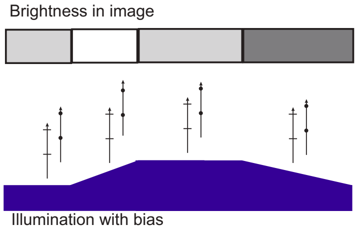
Brightness in image
Illumination with bias

Polarization Microscopy

Polarization microscopy provides images containing information about the properties of anisotropic samples. A polarization microscope is a modified compound microscope with three unique components: the **polarizer**, **analyzer**, and **compensator**. A linear polarizer is located between the light source and the specimen, close to the condenser's diaphragm. The analyzer is a linear polarizer with its transmission axis perpendicular to the polarizer. It is placed between the sample and the eye/camera, at or near the microscope objective's aperture stop. If no sample is present, the image should appear uniformly dark. The compensator (a type of retarder) is a birefringent plate of known parameters used for quantitative sample analysis and contrast adjustment.

Quantitative information can be obtained by introducing known retardation. Rotation of the analyzer correlates the analyzer's angular position with the retardation required to minimize the light intensity for object features at selected wavelengths. **Retardation** introduced by the compensator plate can be described as

$$\Gamma = \left(n_\mathrm{e} - n_\mathrm{o}\right)t\,,$$

where t is the sample thickness, and subscripts e and o denote the extraordinary and ordinary beams. Retardation in concept is similar to the optical path difference for beams propagating through two different materials:

$$OPD = \left(n_1 - n_2\right)t = \left(n_\mathrm{e} - n_\mathrm{o}\right)t\,.$$

The phase delay caused by sample birefringence is therefore

$$\delta = \frac{2\pi OPD}{\lambda} = \frac{2\pi\Gamma}{\lambda}\,.$$

A polarization microscope can also be used to determine the orientation of the optic axis.

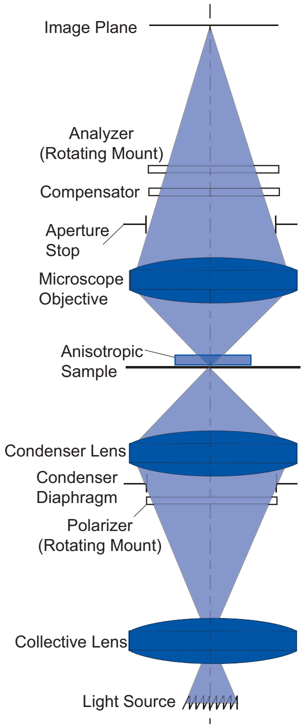
Image Plane
Analyzer (Rotating Mount)
Compensator
Aperture Stop
Microscope Objective
Anisotropic Sample
Condenser Lens
Condenser Diaphragm
Polarizer (Rotating Mount)
Collective Lens
Light Source

Images Obtained with Polarization Microscopes

Anisotropic samples observed under a **polarization microscope** contain dark and bright features and will strongly depend on the geometry of the sample. Objects can have characteristic elongated, linear, or circular structures:

- linear structures appear as dark or bright, depending on the orientation of the analyzer.
- circular structures have a “Maltese Cross” pattern with four quadrants of different intensities.

While polarization microscopy uses monochromatic light, it can also use white-light illumination. For broadband light, different wavelengths will be subject to different retardation; as a result, samples can provide different output colors. In practical terms, this means that color encodes information about retardation introduced by the object. The specific retardation is related to the image color. Therefore, the color allows for determination of sample thickness (for known retardation) or its retardation (for known sample thickness). If the polarizer and the analyzer are crossed, the color visible through the microscope is a complementary color to that having full wavelength retardation (a multiple of 2π, and the intensity is minimized for this color). Note that the two complementary colors combine into white light. If the analyzer could be rotated to be parallel to the analyzer, the two complementary colors will be switched (the one previously displayed will be minimized, while the other color will be maximized).

Polarization microscopy can provide quantitative information about the sample with the application of compensators (with known retardation). Estimating retardation might be performed visually or by integrating with an image acquisition system and CCD. This latter method requires one to obtain a series of images for different retarder orientations and calculate sample parameters based on saved images.

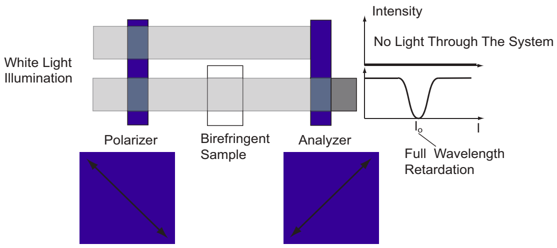

Compensators

Compensators are components that can be used to provide quantitative data about a sample's retardation. They introduce known retardation for a specific wavelength and can act as nulling devices to eliminate any phase delay introduced by the specimen. Compensators can also be used to control the background intensity level.

A **wave plate compensator** (also called a first-order red or red-I plate) is oriented at a 45-deg angle with respect to the analyzer and polarizer. It provides retardation equal to an even number of waves for 551 nm (green), which is the wavelength eliminated after passing the analyzer. All other wavelengths are retarded with a fraction of the wavelength and can partially pass the analyzer and appear as a bright red magenta. The sample provides additional retardation and shifts the colors towards blue and yellow. Color tables determine the retardation introduced by the object.

The **de Sénarmont compensator** is a quarter-wave plate (for 546 nm or 589 nm). It is used with monochromatic light for samples with retardation in the range of $\lambda/20$ to λ. A quarter-wave plate is placed between the analyzer and the polarizer with its slow ellipsoid axis parallel to the transmission axis of the analyzer. Retardation of the sample is measured in a two-step process. First, the sample is rotated until the maximum brightness is obtained. Next, the analyzer is rotated until the intensity maximally drops (the extinction position). The analyzer's rotation angle θ finds the phase delay related to retardation: $\Gamma_{\text{sample}} = 2\theta$.

A **Brace Köhler** rotating **compensator** is used for small retardations and monochromatic illumination. It is a thin birefringent plate with an optic axis in its plane. The analyzer and polarizer remain fixed while the compensator is rotated between the maximum and minimum intensities in the sample's image. Zero position (the maximum intensity) is determined for a slow axis being parallel to the transmission axis of the analyzer, and retardation is

$$\Gamma_{\text{sample}} = \Gamma_{\text{compensator}} \sin 2\theta \,.$$

Confocal Microscopy

Confocal microscopy is a scanning technique that employs pinholes at the illumination and detection planes so that only in-focus light reaches the detector. This means that the intensity for each sample point must be obtained in sequence through scanning in the x and y directions. An illumination pinhole is used in conjunction with the imaged sample plane and only illuminates the point of interest. The detection pinhole rejects the majority of out-of-focus light.

Confocal microscopy has the advantage of lowering the background light from out-of-focus layers, increasing spatial resolution, and providing the capability of imaging thick 3D samples, if combined with z scanning. Due to detection of only the in-focus light, confocal microscopy can provide images of thin sample sections. The system usually employs a photo-multiplier tube (PMT), avalanche photodiodes (APD), or a charge-coupled device (CCD) camera as a detector. For point detectors, recorded data is processed to assemble x-y images. This makes it capable of quantitative studies of an imaged sample's properties. Systems can be built for both reflectance and fluorescence imaging.

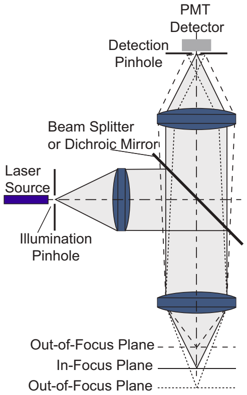

Spatial resolution of a confocal microscope can be defined as

$$d_{xy} \approx \frac{0.4\lambda}{NA},$$

and is slightly better than wide-field (bright-field) microscopy resolution. For pinholes larger than an Airy disk, the spatial resolution is the same as in a wide-field microscope. The **axial resolution** of a confocal system is

$$d_z \approx \frac{1.4n\lambda}{NA^2}.$$

The optimum pinhole value is at the full width at half maximum (FWHM) of the Airy disk intensity. This corresponds to ~75% of its intensity passing through the system, minimally smaller than the Airy disk's first ring:

$$D_{\text{pinhole}} = \frac{0.5\lambda M}{NA}.$$

Scanning Approaches

A scanning approach is directly connected with the temporal resolution of confocal microscopes. The number of points in the image, scanning technique, and the frame rate are related to the signal-to-noise ratio SNR (through time dedicated to detection of a single point). To balance these parameters, three major approaches were developed: point scanning, line scanning, and disk scanning.

A **point-scanning confocal microscope** is based on point-by-point scanning using, for example, two mirror galvanometers or resonant scanners. Scanning mirrors should be located in pupil conjugates (or close to them) to avoid light fluctuations. Maximum spatial resolution and maximum background rejection are achieved with this technique.

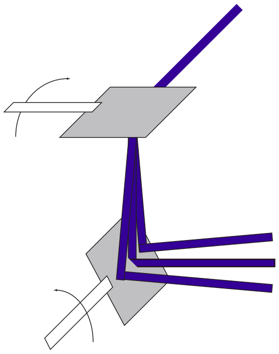

A **line-scanning confocal microscope** uses a slit aperture that scans in a direction perpendicular to the slit. It uses a cylindrical lens to focus light onto the slit to maximize throughput. Scanning in one direction makes this technique significantly faster than a point approach. However, the drawbacks are a loss of resolution and sectioning performance for the direction parallel to the slit.

Feature	**Point Scanning**	**Slit Slit Scanning**	**Disk Spinning**
z resolution	High	Depends on slit spacing	Depends on pinhole distribution
x,y resolution	High	Lower for one direction	Depends on pinhole spacing
Speed	Low to moderate	High	High
Light sources	Lasers	Lasers	Laser and other
Photobleaching	High	High	Low
QE of detectors	Low (PMT) Good (APD)	Good (CCD)	Good (CCD)
Cost	High	High	Moderate

Scanning Approaches (cont.)

Spinning-disk confocal imaging is a parallel-imaging method that maximizes the scanning rate and can achieve a 100–1000 speed gain over point scanning. It uses an array of pinholes/slits (e.g., Nipkow disk, Yokogawa, Olympus DSU approach). To minimize light loss, it can be combined (e.g., with the Yokogawa approach) with an array of lenses, so each pinhole has a dedicated focusing component. Pinhole disks contain several thousand pinholes, but only a portion is illuminated at one time.

	Throughput of a confocal microscope T [%]
Array of pinholes	$T_{\text{pinhole}} = 100 \times (D/S)^2$
Multiple slits	$T_{\text{slit}} = 100 \times (D/S)$
Equations are for a uniformly illuminated mask of pinholes/slits (array of microlenses not considered). D is the pinhole's diameter or slit's width respectively, while S is the pinhole/slit separation.	

Crosstalk between pinholes and slits will reduce the background rejection. Therefore a common pinhole/slit separation S is 5−10 times larger than the pinhole's diameter or the slit's width.

Spinning-disk and slit-scanning confocal microscopes require high-sensitivity array image sensors (CCD cameras) instead of point detectors. They can use lower excitation intensity for fluorescence confocal systems; therefore, they are less susceptible to photobleaching.

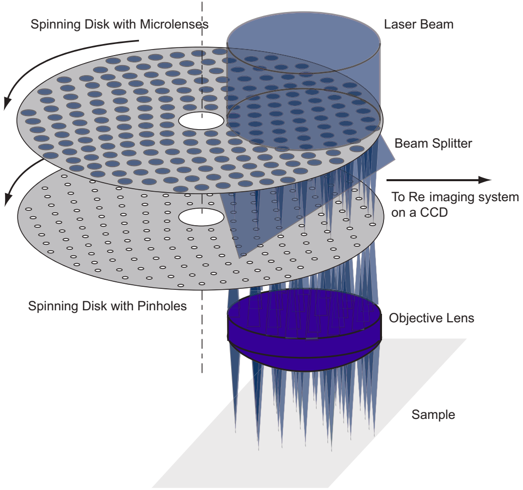
Spinning Disk with Microlenses
Laser Beam
Beam Splitter
To Re imaging system on a CCD
Spinning Disk with Pinholes
Objective Lens
Sample

Images from a Confocal Microscope

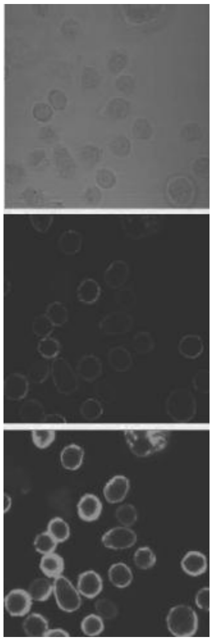

Laser-scanning confocal microscopy rejects out-of-focus light and enables optical sectioning. It can be assembled in reflectance and fluorescent modes. A 1483 cell line, stained with membrane anti-EGFR antibodies labeled with fluorecent Alexa488 dye, is presented here. The top image shows the bright-field illumination mode. The middle image is taken by a confocal microscope with a pinhole corresponding to approximately 2.5 Airy disks. In the bottom image, the pinhole was entirely open. Clear cell boundaries are visible in the confocal image, while all out-of-focus features were removed.

Images were taken with a Zeiss LSM 510 microscope using a 63× Zeiss oil-immersion objective. Samples were excited with a 488-nm laser source.

Another example of 1483 cells labeled with EGF-Alexa647 and proflavine obtained on a Zeiss LSM 510 confocal at 63× / 1.4 oil is shown below. The proflavine (staining nuclei) was excited at 488 nm, with emission collected after a 505-nm long-pass filter. The EGF-Alexa647 (cell membrane) was excited at 633 nm, with emission collected after a 650−710 nm bandpass filter. The sectioning ability of a confocal system is nicely demonstrated by the channel with the membrane labeling.

EGF-Alexa647 Channel (Red)

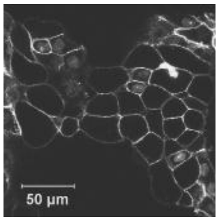

Proflavine Channel (Green)

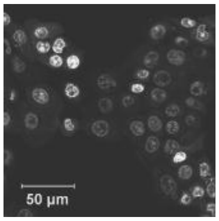

Combined Channels

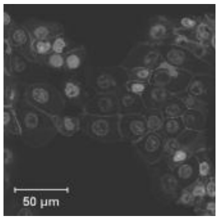

Fluorescence

Specimens can absorb and re-emit light through **fluorescence**. The specific wavelength of light absorbed or emitted depends on the energy level structure of each molecule. When a molecule absorbs the energy of light, it briefly enters an excited state before releasing part of the energy as fluorescence. Since the emitted energy must be lower than the absorbed energy, fluorescent light is always at longer wavelengths than the excitation light. Absorption and emission take place between multiple sub-levels within the ground and excited states, resulting in absorption and emission spectra covering a range of wavelengths. The loss of energy during the fluorescence process causes the **Stokes shift** (a shift of wavelength peak from that of excitation to that of emission). Larger Stokes shifts make it easier to separate excitation and fluorescent light in the microscope. Note that energy quanta and wavelength are related by $E = hc/\lambda$.

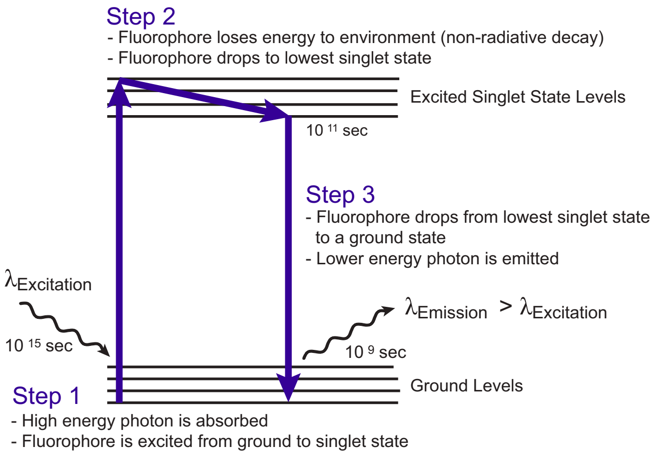
Step 2
- Fluorophore loses energy to environment (non-radiative decay)
- Fluorophore drops to lowest singlet state
Excited Singlet State Levels
10^{11} sec
Step 3
- Fluorophore drops from lowest singlet state to a ground state
- Lower energy photon is emitted
$\lambda_{\text{Excitation}}$
10^{15} sec
$\lambda_{\text{Emission}} > \lambda_{\text{Excitation}}$
10^{9} sec
Ground Levels
Step 1
- High energy photon is absorbed
- Fluorophore is excited from ground to singlet state

The fluorescence principle is widely used in fluorescence microscopy for both organic and inorganic substances. While it is most common for biological imaging, it is also possible to examine samples like drugs and vitamins.

Due to the application of filter sets, the fluorescence technique has a characteristically low background and provides high-quality images. It is used in combination with techniques like wide-field microscopy and scanning confocal-imaging approaches.

Configuration of a Fluorescence Microscope

A fluorescence microscope includes a set of three **filters**: an **excitation filter**, **emission filter**, and a **dichroic mirror** (also called a **dichroic beam splitter**). These filters separate weak emission signals from strong excitation illumination. The most common fluorescence microscopes are configured in epi-illumination mode. The dichroic mirror reflects incoming light from the lamp (at short wavelengths) onto the specimen. Fluorescent light (at longer wavelengths) collected by the objective lens is transmitted through the dichroic mirror to the eyepieces or camera. The transmission and reflection properties of the dichroic mirror must be matched to the excitation and emission spectra of the fluorophore being used.

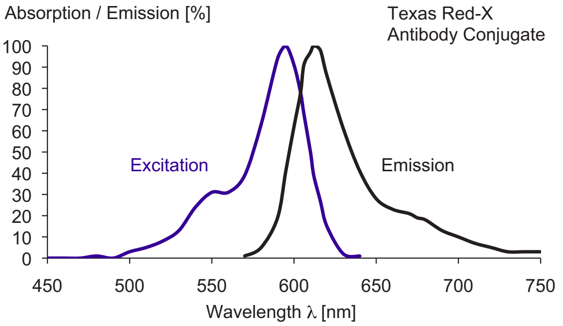

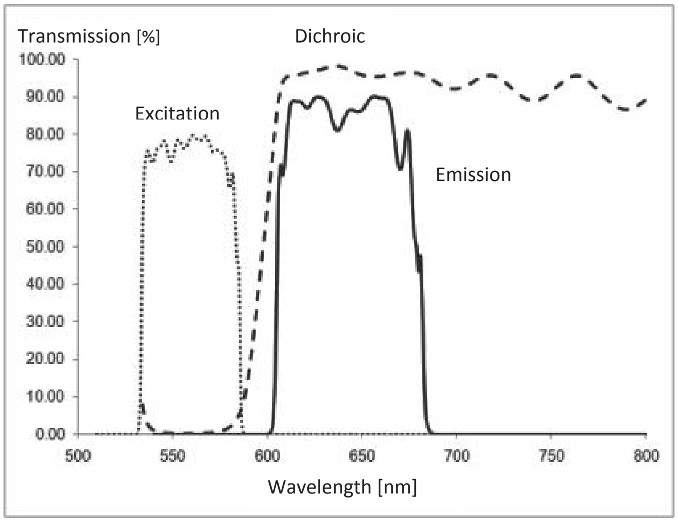

Configuration of a Fluorescence Microscope (cont.)

A sample must be illuminated with light at wavelengths within the excitation spectrum of the fluorescent dye. Fluorescence-emitted light must be collected at the longer wavelengths within the emission spectrum. Multiple fluorescent dyes can be used simultaneously, with each designed to localize or target a particular component in the specimen.

The **filter turret** of the microscope can hold several cubes that contain filters and dichroic mirrors for various dye types. Images are then reconstructed from CCD frames captured with each filter cube in place. Simultanous acquisition of emission for multiple dyes requires application of multiband dichroic filters.

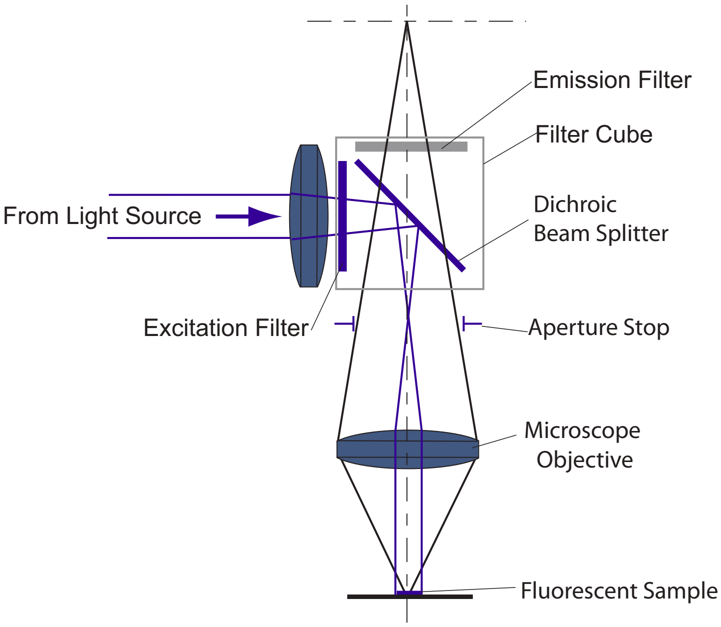
Emission Filter
Filter Cube
From Light Source
Dichroic Beam Splitter
Excitation Filter
Aperture Stop
Microscope Objective
Fluorescent Sample

Images from Fluorescence Microscopy

Fluorescent images of cells labeled with three different fluorescent dyes, each targeting a different cellular component, demonstrate fluorescence microscopy's ability to target sample components and specific functions of biological systems. Such images can be obtained with individual filter sets for each fluorescent dye, or with a multiband (a triple-band in this example) filter set, which matches all dyes.

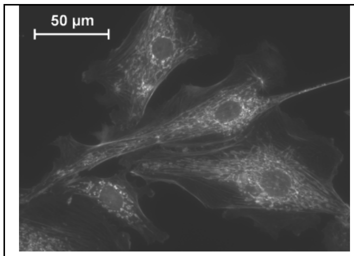 Triple-band Filter	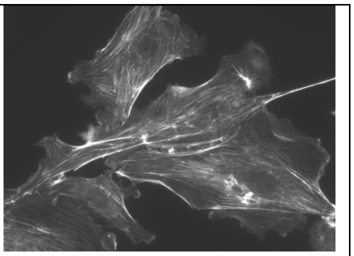 BODIPY FL phallacidin (F-actin)
MitoTracker Red CMXRos (mitochondria)	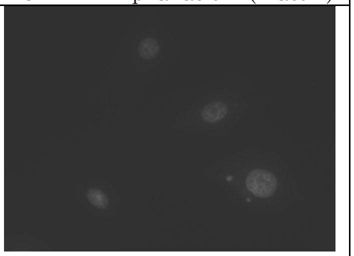 DAPI (nuclei)

Sample: Bovine pulmonary artery epithelial cells (Invitrogen FluoCells No. 1)			
Components: Plan-apochromat 40×/0.95, MRm Zeiss CCD (1388×1040 mono)			
Fluorescent labels:	Peak Excitation		Peak Emission
DAPI	358 nm		461 nm
BODIPY FL	505 nm		512 nm
MitoTracker Red CMXRos	579 nm		599 nm
Filter:	Excitation	Dichroic	Emission
Triple-band	395–415 480–510 560–590	435 510 600	448–472 510–550 600–650
DAPI	325–375	395	420–470
GFP	450–490	495	500–550
Texas Red	530–585	600	615LP

Properties of Fluorophores

Fluorescent emission F [photons/s] depends on the intensity of excitation light I [photons/s] and fluorophore parameters like the **fluorophore quantum yield** Q and **molecular absorption cross section** σ:

$$F = \sigma Q I \ ,$$

where the molecular absorption cross section is the probability that illumination photons will be absorbed by fluorescent dye. The intensity of excitation light depends on the power of the light source and throughput of the optical system. The detected fluorescent emission is further reduced by the quantum efficiency of the detector.

Quenching is a partially reversible process that decreases fluorescent emission. It is a non-emissive process of electrons moving from an excited state to a ground state.

Fluorescence emission and excitation parameters are bound through a photobleaching effect that reduces the ability of fluorescent dye to fluoresce. **Photobleaching** is an irreversible phenomenon that is a result of damage caused by illuminating photons. It is most likely caused by the generation of long-living (in triplet states) molecules of singlet oxygen and oxygen free radicals generated during its decay process. There is usually a finite number of photons that can be generated for a fluorescent molecule. This number can be defined as the ratio of emission quantum efficiency to bleaching quantum efficiency, and it limits the time a sample can be imaged before entirely bleaching. Photobleaching causes problems in many imaging techniques, but it can be especially critical in time-lapse imaging. To slow down this effect, optimizing collection efficiency (together with working at lower power at the steady-state condition) is crucial.

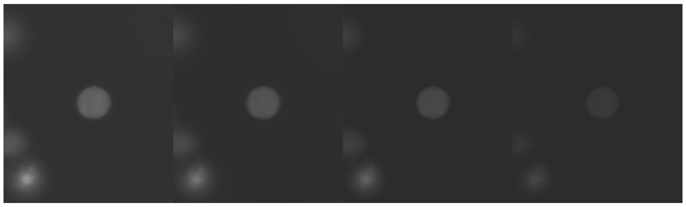

Photobleaching effect as seen in consecutive images

Single vs Multi-Photon Excitation

Multi-photon fluorescence is based on the fact that two or more low-energy photons match the energy gap of the fluorophore to excite and generate a single high-energy photon. For example, the energy of two 700-nm photons can excite an energy transition approximately equal to that of a 350-nm photon (it is not an exact value due to thermal losses). This is contrary to traditional fluorescence, where a high-energy photon (e.g., 400 nm) generates a slightly lower-energy (longer wavelength) photon. Therefore, one big advantage of multi-photon fluorescence is that it can obtain a significant difference between excitation and emission.

Fluorescence is based on stochastic behavior, and for single-photon excitation fluorescence, it is obtained with a high probability. However, multi-photon excitation requires at least two photons delivered in a very short time, and the probability is quite low:

$$n_a \propto \delta\left(\frac{P_{\text{avg}}^2}{\tau\nu^2}\right)\left(\pi\frac{NA^2}{hc\lambda}\right)^2.$$

δ is the excitation cross section of dye, P_{avg} is the average power, τ is the length of a pulse, and ν is the repetition frequency. Therefore, multi-photon fluorescence must be used with high-power laser sources to obtain favorable probability. Short-pulse operation will have a similar effect, as τ will be minimized.

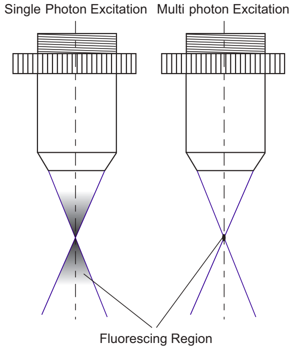

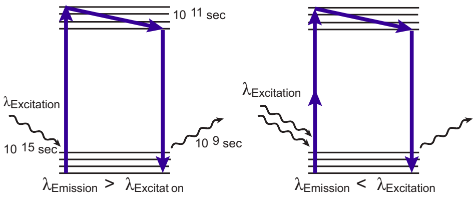

Due to low probability, fluorescence occurs only in the focal point. Multi-photon imaging does not require scanning through a pinhole and therefore has a high signal-to-noise ratio and can provide deeper imaging. This is also because near-IR excitation light penetrates more deeply due to lower scattering, and near-IR light avoids the excitation of several autofluorescent tissue components.

Light Sources for Scanning Microscopy

Lasers are an important light source for scanning microscopy systems due to their high energy density, which can increase the detection of both reflectance and fluorescence light. For laser-scanning confocal systems, a general requirement is a single-mode TEM_{00} laser with a short coherence length. Lasers are used primarily for point-and-slit scanning modalities. There are a great variety of laser sources, but certain features are useful depending on their specific application:

- Short-excitation wavelengths are preferred for fluorescence confocal systems because many fluorescent probes must be excited in the blue or ultraviolet range.
- Near-IR wavelengths are useful for reflectance confocal microscopy and multi-photon excitation. Longer wavelengths are less suseptible to scattering in diffused objects (like tissue) and can provide a deeper penetration depth.
- Broadband short-pulse lasers (like a Ti-Sapphire mode-locked laser) are particularily important for multi-photon generation, while shorter pulses increase the probability of excitation.

Laser Type	**Wavelength [nm]**
Argon-Ion	351, 364, 458, 488, 514
HeCd	325, 442
HeNe	543, 594, 633, 1152
Diode lasers	405, 408, 440, 488, 630, 635, 750, 810
DPSS (diode pumped, solid state)	430, 532, 561
Krypton-Argon	488, 568, 647
Dye	630
Ti-Sapphire*	710–920, 720–930, 750–850, 690–100, 680–1050

*High-power (1000mW or less), pulses between 1 ps and 100 fs

Non-laser sources are particularly useful for disk-scanning systems. They include arc lamps, metal-halide lamps, and high-power photo-luminescent diodes (See also pages 58 and 59).

Practical Considerations in LSM

A light source in laser-scanning microscopy must provide enough energy to obtain a sufficient number of photons to perform successful detection and a statistically significant signal. This is especially critical for non-laser sources (e.g., arc lamps) used in disk-scanning systems. While laser sources can provide enough power, they can cause fast photobleaching or photo-damage to the biological sample.

Detection conditions change with the type of sample. For example, fluorescent objects are subject to photobleaching and saturation. On the other hand, back-scattered light can be easily rejected with filter sets. In reflectance mode, out-of-focus light can cause background comparable or stronger than the signal itself. This background depends on the size of the pinhole, scattering in the sample, and overall system reflections.

Light losses in the optical system are critical for the signal level. Collection efficiency is limited by the NA of the objective; for example, a NA of 1.4 gathers approximately 30% of the fluorescent/scattered light. Other system losses will occur at all optical interfaces, mirrors, filters, and at the pinhole. For example, a pinhole matching the Airy disk allows approximately 75% of the light from the focused image point.

Quantum efficiency (QE) of a detector is another important system parameter. The most common photodetector used in scanning microscopy is the photomultiplier tube (PMT). Its quantum efficiency is often limited to the range of 10–15% (550–580 nm). In practical terms, this means that only 15% of the photons are used to produce a signal. The advantages of PMTs for laser-scanning microscopy are high speed and high signal gain. Better QE can be obtained for CCD cameras, which is 40–60% for standard designs and 90–95% for thinned, back-illuminated cameras. However, CCD image sensors operate at lower rates.

Spatial and temporal sampling of laser-scanning microscopy images imposes very important limitations on image quality. The image size and frame rate are often determined by the number of photons sufficient to form high-quality images.

Interference Microscopy

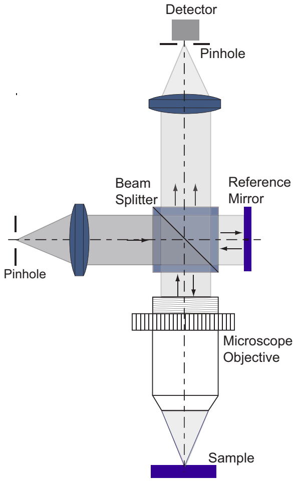

Interference microscopy includes a broad family of techniques for quantitative sample characterization (thickness, refractive index, etc.). Systems are based on microscopic implementation of interferometers (like the Michelson and Mach-Zehnder geometries), or polarization interferometers.

In interference microscopy, short coherence systems are particularly interesting and can be divided into two groups: **optical profilometry** and **optical coherence tomography** (OCT) [including **optical coherence microscopy** (OCM)]. The primary goal of these techniques is to add a third (*z*) dimension to the acquired data.
Optical profilers use interference fringes as a primary source of object height. Two important measurement approaches include **vertical scanning interferometry** (VSI) (also called **scanning white light interferometry** or **coherence scanning interferometry**) and **phase-shifting interferometry**. Profilometry techniques are capable of achieving nanometer-level resolution in the *z* direction while *x* and *y* are defined by standard microscope limitations.

Profilometry is used for characterizing the sample surface (roughness, structure, microtopography, and in some cases film thickness) while OCT/OCM is dedicated to 3D imaging, primarily of biological samples.
Both types of systems are usually configured using the geometry of the Michelson or its derivatives (e.g., Mirau).

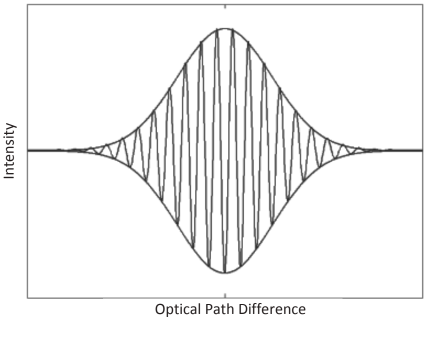

Optical Coherence Tomography/Microscopy

In early 3D coherence imaging, information about the sample was gated by the coherence length of the light source (**time-domain OCT**). This means that the maximum fringe contrast is obtained at a zero optical path difference, while the entire fringe envelope has a width related to the coherence length. In fact, this width defines the axial resolution (usually a few microns) and images are created from the magnitude of the fringe pattern envelope. **Optical coherence microscopy** is a combination of OCT and confocal microscopy. It merges the advantages of out-of-focus background rejection provided by the confocal principle and applies the coherence gating of OCT to improve optical sectioning over confocal reflectance systems.

In the mid-2000s, time-domain OCT transitioned to **Fourier-domain OCT** (FDOCT), which can acquire an entire depth profile in a single capture event. Similar to Fourier Transform Spectroscopy, the image is acquired by calculating the Fourier transform of the obtained interference pattern. The measured signal can be described as

$$I(k, z_\mathrm{o}) = \int A_\mathrm{R} A_\mathrm{S}(z) \cos\left[k\left(z_\mathrm{m} - z_\mathrm{o}\right)\right] dz \,,$$

where A_R and A_S are amplitudes of the reference and sample light, respectively, and z_m is the imaged sample depth (z_o is the reference length delay, and k is the wave number).

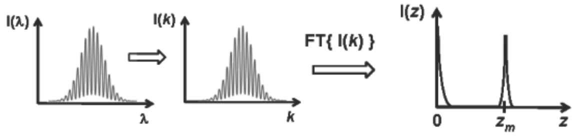

The fringe frequency is a function of wave number k and relates to the OPD in the sample. Within the Fourier-domain techniques, two primary methods exist: **spectral-domain OCT** (SDOCT) and **swept-source OCT** (SSOCT). Both can reach detection sensitivity and signal-to-noise ratios over 150 times higher than time-domain OCT approaches. SDOCT achieves this through the use of a broadband source and spectrometer for detection. SSOCT uses a single photodetector but rapidly tunes a narrow linewidth over a broad spectral profile. Fourier-domain systems can image at hundreds of frames per second with sensitivities over 100 dB.

Optical Profiling Techniques

There are two primary techniques used to obtain a surface profile using white-light profilometry: **vertical scanning interferomtry** (VSI) and **phase-shifting interferometry** (PSI).

VSI is based on the fact that a reference mirror is moved by a piezoelectric or other actuator, and a sequence of several images (in some cases hundreds or thousands) is acquired along the scan. During post-processing, algorithms look for a maximum fringe modulation and correlate it with the actuator's position. This position is usually determined with capacitance gauges; however, more advanced systems use optical encoders or an additional Michelson interferometer. The important advantage of VSI is its ability to ambiguously measure heights that are greater than a fringe.

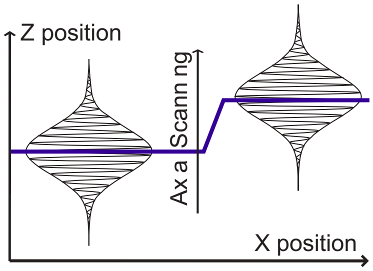

The PSI technique requires the object of interest to be within the coherence length and provides one order of magnitude or greater z resolution compared to VSI. To extend coherence length, an optical filter is introduced to limit the light source's bandwidth. With PSI, a phase-shifting component is located in one arm of the interferometer. The reference mirror can also provide the role of the phase shifter. A series of images (usually three to five) is acquired with appropriate phase differences to be used later in phase reconstruction. Introduced phase shifts change the location of the fringes. An important PSI drawback is the fact that phase maps are subject to 2π discontinuities. Phase maps are first obtained in the form of so-called phase fringes and must be unwrapped (a procedure of removing discontinuities) to obtain a continuous phase map.

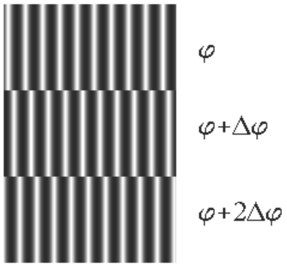

Note that modern short-coherence interference microscopes combine phase information from PSI with a long range of VSI methods.

Optical Profilometry: System Design

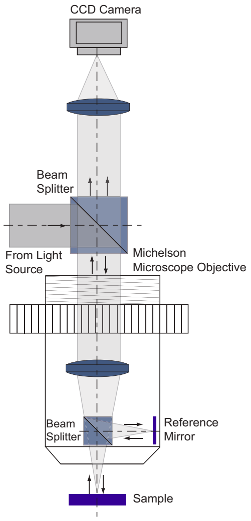
CCD Camera
Beam Splitter
From Light Source
Michelson Microscope Objective
Beam Splitter
Reference Mirror
Sample

Optical profilometry systems are based on implementing a Michelson interferometer or its derivatives into their design. There are three primary implementations: classical **Michelson** geometry, **Mirau**, and **Linnik**. The first approach incorporates a Michelson inside a microscope objective using a classical interferometer layout. Consequently, it can work only for low magnifications and short working distances. The Mireau objective uses two plates perpendicular to the optical axis inside the objective. One is a beam-splitting plate, while the other acts as a reference mirror. Such a solution extends the working distance and increases the possible magnifications.

Design	**Magnification**
Michelson	1× to 5×
Mireau	10× to 100×
Linnik	50× to 100×

The Linnik design utilizes two matching objectives. It does not suffer from NA limitation, but it is quite expensive and susceptible to vibrations.

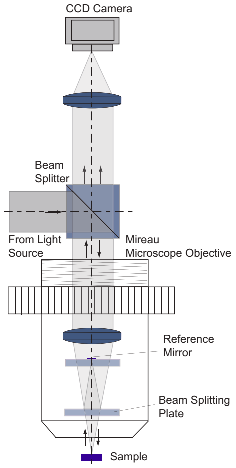
CCD Camera
Beam Splitter
From Light Source
Mireau Microscope Objective
Reference Mirror
Beam Splitting Plate
Sample

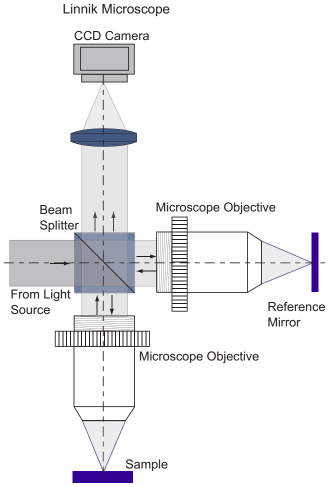
Linnik Microscope
CCD Camera
Beam Splitter
Microscope Objective
From Light Source
Reference Mirror
Microscope Objective
Sample

Phase-Shifting Algorithms

Phase-shifting techniques find the phase distribution from interference images given by

$$I(x,y) = a(x,y) + b(x,y)\cos(\varphi + n\Delta\varphi),$$

where $a(x, y)$ and $b(x, y)$ correspond to background and fringe amplitude, respectively. Since this equation has three unknowns, at least three measurements (images) are required. For image acquisition with a discrete CCD camera, spatial (x, y) coordinates can be replaced by (i, j) pixel coordinates.

The most basic PSF algorithms include the **three-**, **four-**, and **five-image** techniques.

I denotes the intensity for a specific image (1st, 2nd, 3rd, etc.) in the selected i, j pixel of a CCD camera. The phase shift for a three-image algorithm is $\pi/2$. Four- and five-image algorithms also acquire images with $\pi/2$ phase shifts.

$$\varphi = \arctan\left(\frac{I_3 - I_2}{I_1 - I_2}\right)$$

$$\varphi = \arctan\left(\frac{I_4 - I_2}{I_1 - I_3}\right)$$

$$\varphi = \arctan\left(\frac{2[I_2 - I_4]}{I_1 - I_3 + I_5}\right)$$

A reconstructed phase depends on the accuracy of the phase shifts. $\pi/2$ steps were selected to minimize the systematic errors of the above techniques. Accuracy progressively improves between the three- and five-point techniques. Note that modern PSI algorithms often use seven images or more, in some cases reaching thirteen.

After the calculations, the wrapped phase maps (modulo 2π) are obtained (arctan function). Therefore, unwrapping procedures have to be applied to provide continuous phase maps.

Common unwrapping methods include the line by line algorithm, least square integration, regularized phase tracking, minimum spanning tree, temporal unwrapping, frequency multiplexing, and others.

Structured Illumination: Axial Sectioning

A simple and inexpensive alternative to confocal microscopy is the **structured-illumination technique**. This method uses the principle that all but the zero (DC) spatial frequencies are attenuated with defocus. This observation provides the basis for obtaining the **optical sectioning** of images from a conventional wide-field microscope. A modified illumination system of the microscope projects a single spatial-frequency grid pattern onto the object. The microscope then faithfully images only that portion of the object where the grid pattern is in focus. The structured-illumination technique requires the acquisition of at least three images to remove the illumination structure and reconstruct an image of the layer. The reconstruction relation is described by

$$I = \sqrt{\left(I_0 - I_{2\pi/3}\right)^2 + \left(I_0 - I_{4\pi/3}\right)^2 + \left(I_{2\pi/3} - I_{4\pi/3}\right)^2}\ ,$$

where I denotes the intensity in the reconstructed image point, while I_0, $I_{2\pi/3}$, and $I_{4\pi/3}$ are intensities for the image point and the three consecutive positions of the sinusoidal illuminating grid (zero, 1/3 and 2/3 of the structure period).

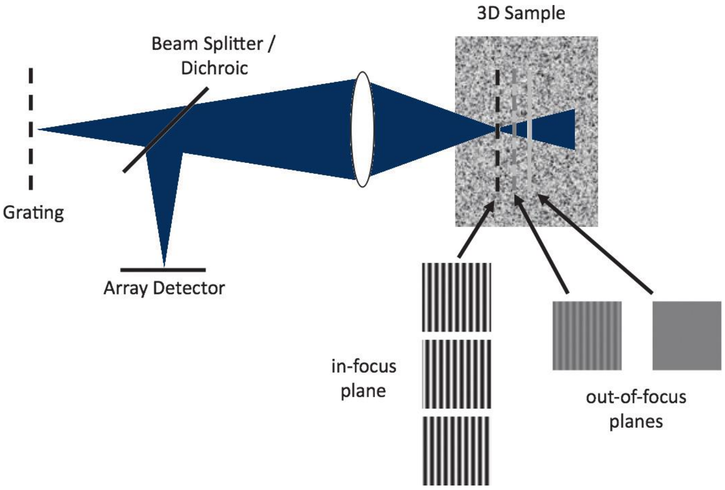

Note that the maximum sectioning is obtained for 0.5 of the objective's cut-off frequency. This sectioning ability is comparable to that obtained in confocal microscopy.

Structured Illumination: Resolution Enhancement

Structured illumination can be used to enhance the spatial resolution of a microscope. It is based on the fact that if a sample is illuminated with a periodic structure, it will create a low-frequency beating effect (Moiré fringes) with the features of the object. This new low-frequency structure will be capable of aliasing high spatial frequencies through the optical system. Therefore, the sample can be illuminated with a pattern of several orientations to accommodate the different directions of the object features. In practical terms, this means that system apertures will be doubled (in diameter), creating a new synthetic aperture.

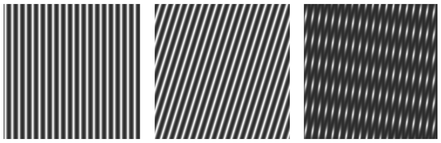

To produce a high-frequency structure, the interference of two beams with large illumination angles can be used. Note that the blue dots in the figure represent aliased spatial frequencies.

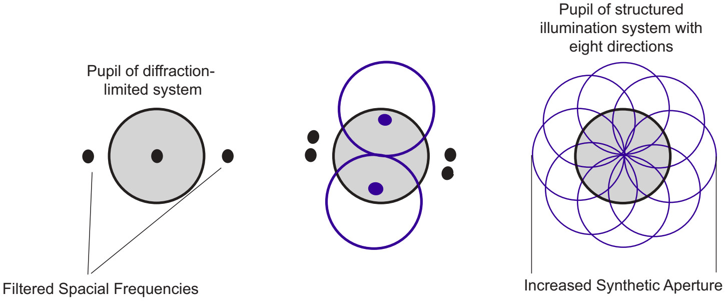
Pupil of diffraction-limited system
Pupil of structured illumination system with eight directions
Filtered Spacial Frequencies
Increased Synthetic Aperture

To reconstruct distribution, at least three images are required for one grid direction. In practice, obtaining a high-resolution image is possible with seven to nine images and a grid rotation, while a larger number can provide a more efficient reconstruction.

The linear-structured illumination approach is capable of obtaining two-fold resolution improvement over the diffraction limit. The application of nonlinear gain in fluorescence imaging improves resolution several times while working with higher harmonics. Sample features of 50 nm and smaller can be successfully resolved.

TIRF Microscopy

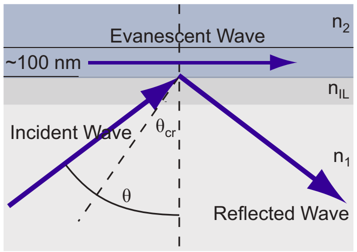

Total internal reflection fluorescence (TIRF) **microscopy** (also called evanescent wave microscopy) excites a limited width of the sample close to a solid interface. In TIRF, a thin layer of the sample is excited by a wavefront undergoing total internal reflection in the dielectric substrate, producing an **evanescent wave** propagating along the interface between the substrate and object.

The thickness of the excited section is limited to less than 100–200 nm. Very low background fluorescence is an important feature of this technique, since only a very limited volume of the sample in contact with the evanescent wave is excited.

This technique can be used for imaging cells, cytoplasmic filament structures, single molecules, proteins at cell membranes, micro-morphological structures in living cells, the adsorption of liquids at interfaces, or Brownian motion at the surfaces. It is also a suitable technique for recording long-term fluorescence movies.

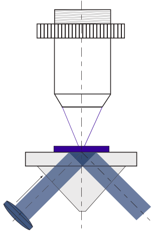

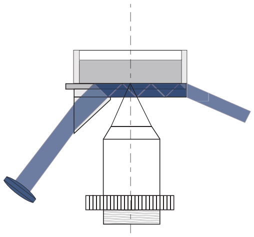

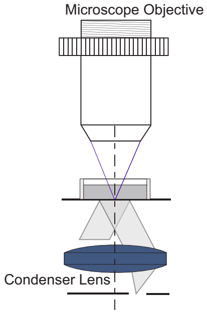

While an evanescent wave can be created without any layers between the dielectric substrate and sample, it appears that a thin layer (e.g., metal) improves image quality. For example, it can quench fluorescence in the first 10 nm and therefore provide selective fluorescence for the 10–200 nm region. TIRF can be easily combined with other modalities like optical trapping, multi-photon excitation, or interference. The most common configurations include oblique illumination through a high-NA objective or condenser and TIRF with a prism.

Solid Immersion

Optical resolution can be enhanced using **solid immersion** imaging. In the classical definition, the diffraction limit depends on the NA of the optical system and the wavelength of light. It can be improved by increasing the refractive index of the media between the sample and optical system or by decreasing the light's wavelength. The solid immersion principle is based on improving the first parameter by applying **solid immersion lenses** (SILs). To maximize performance, SILs are made with a high refractive index glass (usually greater than 2). The most basic setup for a SIL application involves a hemispherical lens. The sample is placed at the center of the hemisphere so the rays propagating through the system are not refracted (they intersect the SIL surface direction) and enter the microscope objective. The SIL is practically in contact with the object, but has a small sub-wavelength gap (<100 nm) between the sample and optical system; therefore, an object is always in an evanescent field and can be imaged with high resolution. Therefore, the technique is confined to a very thin sample volume (up to 100 nm) and can provide optical sectioning. Systems based on solid immersion can reach sub-100-nm lateral resolutions.

The most common solid-immersion applications are in microscopy, including fluorescence, optical data storage, and lithography. Compared to classical oil-immersion techniques, this technique is dedicated to imaging thin samples only (sub-100 nm), but provides much better resolution depending on the configuration and refractive index of the SIL.

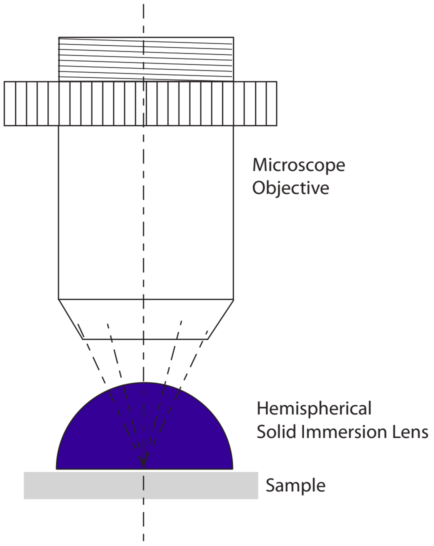

Stimulated Emission Depletion

Stimulated emission depletion (STED) microscopy is a fluorescence super-resolution technique that allows imaging with a spatial resolution at the molecular level. The technique has shown an improvement 5–10 times beyond the possible diffraction-resolution limit.

The STED technique is based on the fact that the area around the excited object point can be treated with a pulse (depletion pulse or STED pulse) that depletes high-energy states and brings fluorescent dye to the ground state. Consequently, an actual excitation pulse excites only the small sub-diffraction resolved area. The depletion pulse must have high intensity and be significantly shorter than the lifetime of the excited states. The pulse must be shaped, for example, with a half-wave phase plate and imaging optics to create a 3D doughnut-like structure around the point of interest. The STED pulse is shifted toward red with regard to the fluorescence excitation pulse and follows it by a few picoseconds. To obtain the complete image, the system scans in the x, y, and z directions.

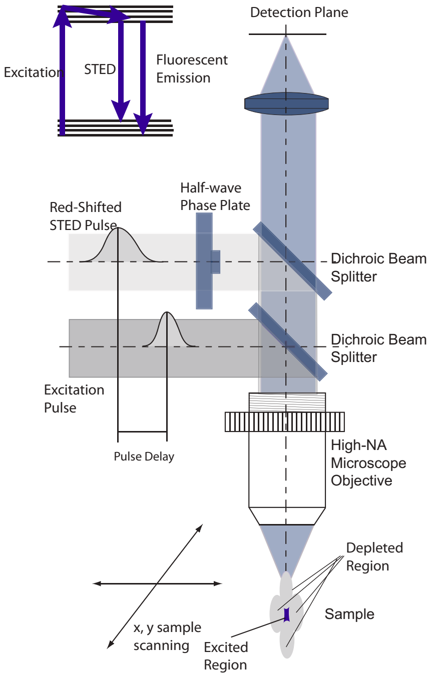
Excitation
STED
Fluorescent Emission
Detection Plane
Red-Shifted STED Pulse
Half-wave Phase Plate
Dichroic Beam Splitter
Dichroic Beam Splitter
Excitation Pulse
Pulse Delay
High-NA Microscope Objective
Depleted Region
Sample
x, y sample scanning
Excited Region

STORM

Stochastic optical reconstruction microscopy (STORM) is based on the application of multiple photo-switchable fluorescent probes that can be easily distinguished in the spectral domain. The system is configured in **TIRF** geometry. Multiple fluorophores label a sample, and each is built as a dye pair where one component acts as a fluorescence activator. The activator can be switched on or deactivated using a different illumination laser (a longer wavelength is used for deactivation—usually red). In the case of applying several dual-pair dyes, different lasers can be used for activation and only one for deactivation.

A sample can be sequentially illuminated with activation lasers, which generate fluorescent light at different wavelengths for slightly different locations on the object (each dye is sparsely distributed over the sample and activation is stochastic in nature). Activation for a different color can also be shifted in time. After performing several sequences of activation/deactivation, it is possible to acquire data for the centroid localization of various diffraction-limited spots. This localization can be performed with precision to about 1 nm. The spectral separation of dyes detects closely located object points encoded with a different color.

The STORM process requires 1000+ activation/deactivation cycles to produce high-resolution images. Therefore, final images combine the spots corresponding to individual molecules of DNA or an antibody used for labeling. STORM images can reach a lateral resolution of about 25 nm and an axial resolution of 50 nm.

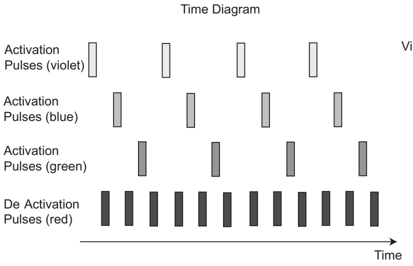
Time Diagram
Activation Pulses (violet)
Activation Pulses (blue)
Activation Pulses (green)
De Activation Pulses (red)
Time

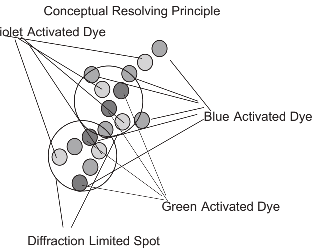
Conceptual Resolving Principle
Violet Activated Dye
Blue Activated Dye
Green Activated Dye
Diffraction Limited Spot

4Pi Microscopy

4Pi microscopy is a fluorescence technique that significantly reduces the thickness of an imaged section. The 4Pi principle is based on coherent illumination of a sample from two directions. Two constructively interfering beams improve the axial resolution by three to seven times. The possible values of z resolution using 4Pi microscopy are 50–100 nm and 30–50 nm when combined with STED.

The undesired effect of the technique is the appearance of side lobes located above and below the plane of the imaged section. They are located about λ/2 from the object. To eliminate side lobes, three primary techniques can be used:

- Combine 4Pi with confocal detection. A pinhole rejects some of the out-of-focus light.
- Detect in multi-photon mode, which quickly diminishes the excitation of fluorescence.
- Apply a modified 4Pi system, which creates interference at both the object and detection planes (two imaging systems are required).

It is also possible to remove side lobes through numerical processing, if they are smaller than 50% of the maximum intensity. However, side lobes increase with the NA of the objective; for an NA of 1.4, they are about 60–70% of the maximum intensity. Numerical correction is not effective in this case.

4Pi microscopy improves resolution only in the z direction; however, the perception of details in the thin layers is a useful benefit of the technique.

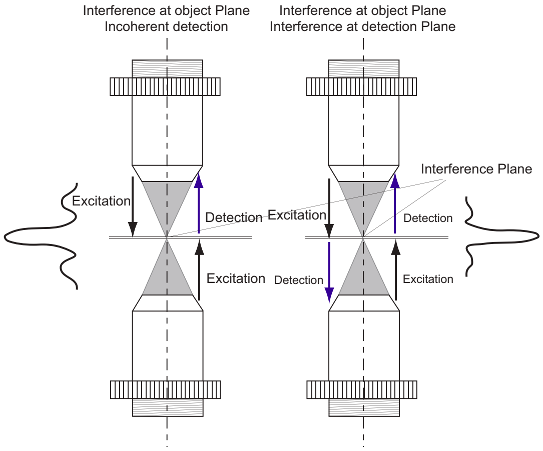
Interference at object Plane
Incoherent detection
Interference at object Plane
Interference at detection Plane
Interference Plane
Excitation
Detection
Excitation
Detection
Excitation
Detection
Detection
Excitation

The Limits of Light Microscopy

Classical light microscopy is limited by diffraction and depends on the wavelength of light and the parameters of an optical system (NA). However, recent developments in super-resolution methods and enhancement techniques overcome these barriers and provide resolution at a molecular level. They often use a sample as a component of the optical train [resolvable saturable optical fluorescence transitions (**RESOLFT**), saturated structured-illumination microscopy (**SSIM**), photoactivated localization microscopy (**PALM**) and **STORM**] or combine near-field effects (**4Pi, TIRF, solid immersion**) with far-field imaging. The table below summarizes these methods.

		Demonstrated Values	
Method	**Principles**	**Lateral**	**Axial**
Bright field	Diffraction	200 nm	500 nm
Confocal	Diffraction (slightly better than bright field)	200 nm	500 nm
Solid immersion	Diffraction / evanescent field decay	< 100 nm	< 100 nm
TIRF	Diffraction / evanescent field decay	200 nm	< 100 nm
4Pi / I5	Diffraction/ interference	200 nm	50 nm
RESOLFT (e.g.,STED)	Depletion, molecular structure of sample–fluorescent probes	20 nm	20 nm
Structured illumination (SSIM)	Aliasing, nonlinear gain in fluorescent probes (molecular structure)	25–50 nm	50–100 nm
Stochastic techniques (PALM, STORM)	Fluorescent probes (molecular structure), centroid localization, time-dependant fluorophore activation	25 nm	50 nm

Raman and CARS Microscopy

Raman microscopy (also called chemical imaging) is a technique based on inelastic **Raman scattering** and evaluates the vibrational properties of samples (minerals, polymers, and biological objects).

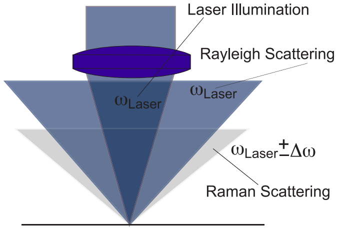

Raman microscopy uses a laser beam focused on a solid object. Most of the illuminating light is scattered, reflected, and transmitted, and preserves the parameters of the illumination beam (the frequency is the same as the illumination). However, a small portion of the light is subject to a shift in frequency: $\omega_{\text{Raman}} = \omega_{\text{laser}} \pm \Delta\omega$. This frequency shift between illumination and scattering bears information about the molecular composition of the sample. Raman scattering is weak and requires high-power sources and sensitive detectors. It is examined with spectroscopy techniques.

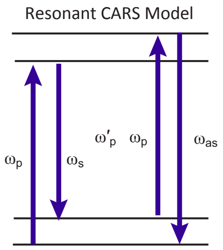

Coherent Anti-Stokes Raman Scattering (**CARS**) overcomes the problem of weak signals associated with Raman imaging. It uses a pump laser and a tunable Stokes laser to stimulate a sample with a four-wave mixing process. The fields of both lasers [pump field $E(\omega_p)$, Stokes field $E(\omega_s)$, and probe field $E'(\omega_p')$] interact with the sample and produce an anti-Stokes field [$E(\omega_{as})$] with frequency ω_{as}, so $\omega_{as} = 2\omega_p - \omega_s$.

CARS works as a resonant process, only providing a signal if the vibrational structure of the sample specifically matches $\omega_p - \omega_s$. It also must assure phase matching so that l_c (the coherence length) is greater than $\pi/\Delta\mathbf{k} = \pi/\left|\mathbf{k}_{as} - (2\mathbf{k}_p - \mathbf{k}_s)\right|$, where $\Delta\mathbf{k}$ is a phase mismatch. CARS is orders of magnitude stronger than Raman scattering, does not require any exogenous contrast, provides good sectioning ability, and can be set up in both transmission and reflectance modes.

SPIM

Selective plane illumination microscopy (SPIM) is dedicated to the high-resolution imaging of large 3D samples. It is based on three principles:

- The sample is illuminated with a **light sheet**, which is obtained with cylindrical optics. The light sheet is a beam focused in one direction and collimated in another. This way, the thin and wide light sheet can pass through the object of interest (see figure).
- The sample is imaged in the direction perpendicular to the illumination.
- The sample is rotated around its axis of gravity and linearly translated into axial and lateral directions.

Data is recorded by a CCD camera for multiple object orientations and can be reconstructed into a single 3D distribution. Both scattered and fluorescent light can be used for imaging.

The maximum resolution of the technique is limited by the longitudinal resolution of the optical system (for a high numerical aperture, objectives can obtain micron-level values). The maximum volume imaged is limited by the working distance of a microscope and can be as small as tens of microns or exceed several millimeters. The object is placed in an air, oil, or water chamber.

The technique is primarily used for bio-imaging, ranging from small organisms to individual cells.

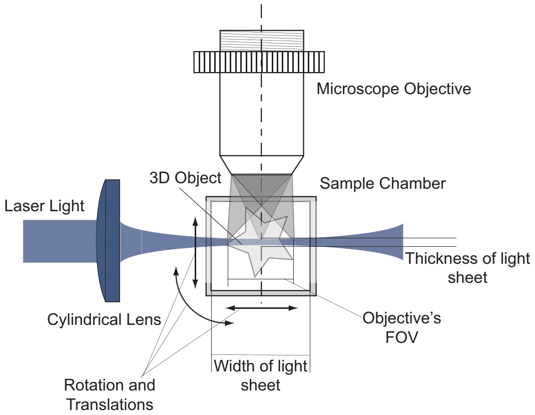
Microscope Objective
3D Object
Sample Chamber
Laser Light
Thickness of light sheet
Cylindrical Lens
Objective's FOV
Width of light sheet
Rotation and Translations

Array Microscopy

An **array microscope** is a solution to the trade-off between field of view and lateral resolution. In the array microscope, a miniature microscope objective is replicated tens of times. The result is an imaging system with a field of view that can be increased in steps of an individual objective's field of view, independent of numerical aperture and resolution. An array microscope is useful for applications that require fast imaging of large areas at a high level of detail. Compared to conventional microscope optics for the same purpose, an array microscope can complete the same task several times faster due to its parallel imaging format.

An example implementation of the array-microscope optics is shown. This array microscope is used for imaging glass slides containing tissue (histology) or cells (cytology). For this case, there are a total of 80 identical miniature 7×/0.6 microscope objectives. The summed field of view measures 18 mm. Each objective consists of three lens elements. The lens surfaces are patterned on three plates that are stacked to form the final array of microscopes. The plates measure 25 mm in diameter. Plate 1 is near the object at a working distance of 400 μm. Between plates 2 and 3 is a baffle plate. A second baffle is located between the third lens and the image plane. The purpose of the baffles is to eliminate cross talk and image overlap between objectives in the array.

The small size of each microscope objective and the need to avoid image overlap jointly dictate a low magnification. Therefore, the array microscope works best in combination with an image sensor divided into small pixels (e.g., 3.3 μm).

Focusing the array microscope is achieved by an up/down translation and two rotations: a pitch and a roll.

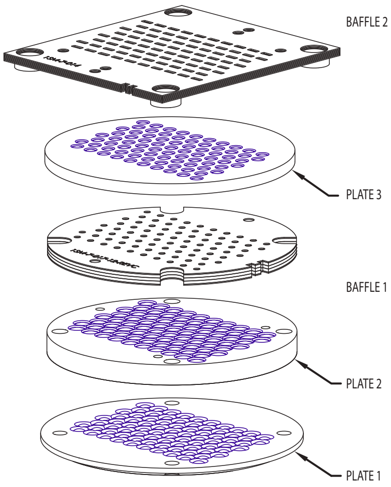

Digital Microscopy

Digital microscopy emerged with the development of electronic array image sensors and replaced the microphotography techniques previously used in microscopy. It can also work with point detectors when images are recombined in post or real-time processing. Digital microscopy is based on acquiring, storing, and processing images taken with various microscopy techniques. It supports applications that require:

- Combining data sets acquired for one object with different microscopy modalities or different imaging conditions. Data can be used for further analysis and processing in techniques like hyperspectral and polarization imaging.
- Image correction (e.g., distortion, white balance correction, or background subtraction). Digital microscopy is used in deconvolution methods that perform numerical rejection of out-of-focus light and iteratively reconstruct an object's estimate.
- Image acquisition with a high temporal resolution. This includes short integration times or high frame rates.
- Long time experiments and remote image recording.
- Scanning techniques where the image is reconstructed after point-by-point scanning and displayed on the monitor in real time.
- Low light, especially fluorescence. Using high-sensitivity detectors reduces both the excitation intensity and excitation time, which mitigates photobleaching effects.
- Contrast enhancement techniques and an improvement in spatial resolution. Digital microscopy can detect signal changes smaller than possible with visual observation.
- Super-resolution techniques that may require the acquisition of many images under different conditions.
- High throughput scanning techniques (e.g., imaging large sample areas).
- UV and IR applications not possible with visual observation.

The primary detector used for digital microscopy is a CCD camera. For scanning techniques, a photomultiplier or photodiodes are used.

Principles of CCD Operation

Charge-coupled device (CCD) sensors are a collection of individual photodiodes arranged in a 2D grid pattern. Incident photons produce electron-hole pairs in each pixel in proportion to the amount of light on each pixel. By collecting the signal from each pixel, an image corresponding to the incident light intensity can be reconstructed.

Here are the step-by-step processes in a CCD:

1. The CCD array is illuminated for integration time τ.
2. Incident photons produce mobile charge carriers.
3. Charge carriers are collected within the p-n structure.
4. The illumination is blocked after time τ (full-frame only).
5. The collected charge is transferred from each pixel to an amplifier at the edge of the array.
6. The amplifier converts the charge signal into voltage.
7. The analog voltage is converted to a digital level for computer processing and image display.

Each pixel (**photodiode**) is reverse biased, causing photoelectrons to move towards the positively charged electrode. Voltages applied to the electrodes produce a potential well within the semiconductor structure. During the integration time, electrons accumulate in the potential well, up to the full-well capacity. The full-well capacity is reached when the repulsive force between electrons exceeds the attractive force of the well. At the end of the exposure time, each pixel has stored a number of electrons in proportion to the amount of light received. These charge packets must be transferred from the sensor, from each pixel to a single amplifier without loss. This is accomplished by a series of parallel and serial shift registers. The charge stored by each pixel is transferred across the array by sequences of gating voltages applied to the electrodes. The packet of electrons follows the positive clocking waveform voltage from pixel-to-pixel or row-to-row. A potential barrier is always maintained between adjacent pixel charge packets.

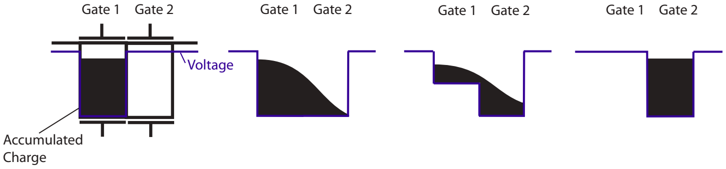

CCD Architectures

In a **full-frame architecture**, individual pixel charge packets are transferred by a parallel row shift to the serial register, then one-by-one to the amplifier. The advantage of this approach is a 100% photosensitive area, while the disadvantage is that a shutter must block the sensor area during readout, and consequently limits the frame rate.

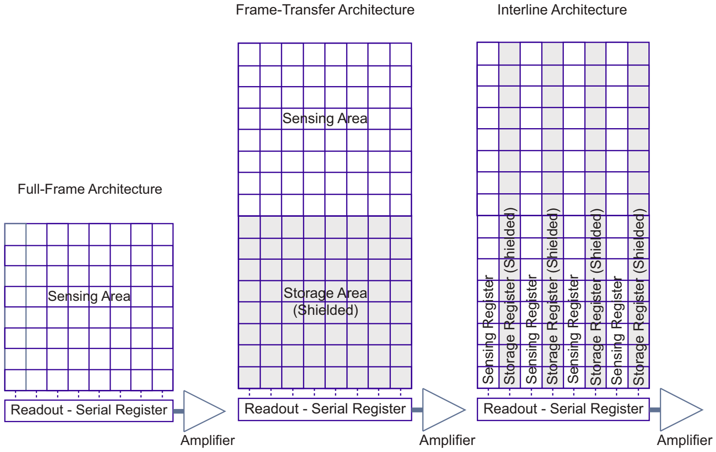
Full-Frame Architecture
Sensing Area
Readout - Serial Register
Amplifier
Frame-Transfer Architecture
Sensing Area
Storage Area (Shielded)
Readout - Serial Register
Amplifier
Interline Architecture
Sensing Register
Storage Register (Shielded)
Sensing Register
Storage Register (Shielded)
Sensing Register
Storage Register (Shielded)
Sensing Register
Storage Register (Shielded)
Readout - Serial Register
Amplifier

The **frame-transfer architecture** has the advantage of not needing to block an imaging area during readout. It is faster than full-frame, since readout can take place during the integration time. The significant drawback is that only 50% of the sensor area is used for imaging.

The **interline transfer architecture** uses columns with exposed imaging pixels interleaved with columns of masked storage pixels. A charge is transferred into an adjacent storage column for readout. The advantage of such an approach is a high frame rate due to a rapid transfer time (< 1 ms), while again the disadvantage is that 50% of the sensor area is used for light collection. However, recent interleaved CCDs have an increased fill factor due to the application of microlens arrays.

CCD Architectures (cont.)

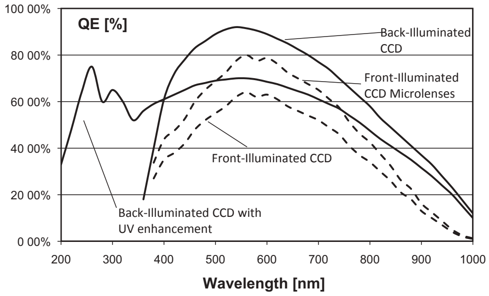

Quantum efficiency (QE) is the percentage of incident photons at each wavelength that produce an electron in a photodiode. CCDs have a lower QE than individual silicon photodiodes due to the charge-transfer channels on the sensor surface, which reduces the effective light-collection area. A typical CCD QE is 40–60%. This can be improved for back-thinned CCDs, which are illuminated from behind; this avoids the surface electrode channels and improves QE to approximately 90–95%.

Recently, **electron-multiplying CCDs** (EM-CCD), primarily used for biological imaging, have become more common. They are equipped with a gain register placed between the shift register and output amplifier. A multi-stage gain register allows high gains. As a result, single electrons can generate thousands of output electrons. While the read noise is usually low, it is therefore negligible in the context of thousands of signal electrons.

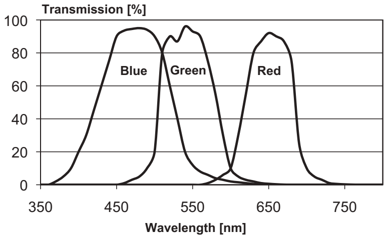

CCD pixels are not color sensitive, but use color filters to separately measure red, green, and blue light intensity. The most common method is to cover the sensor with an array of **RGB** (red, green, blue) filters, which combine four individual pixels to generate a single color pixel. The **Bayer mask** uses two green pixels for every red and blue pixel, which simulates human visual sensitivity.

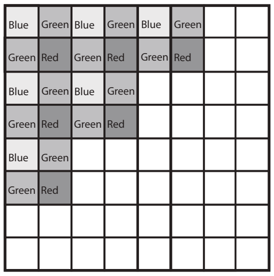

Blue	Green	Blue	Green	Blue	Green		
Green	Red	Green	Red	Green	Red		
Blue	Green	Blue	Green				
Green	Red	Green	Red				
Blue	Green						
Green	Red						

CCD Noise

The three main types of noise that affect CCD imaging are **dark noise**, **read noise**, and **photon noise**.

Dark noise arises from random variations in the number of thermally generated electrons in each photodiode. More electrons can reach the conduction band as the temperature increases, but this number follows a statistical distribution. These random fluctuations in the number of conduction band electrons results in a **dark current** that is not due to light. To reduce the influence of dark current for long time exposures, CCD cameras are cooled (which is crucial for the exposure of weak signals, like fluorescence, with times of a few seconds and more).

Read noise describes the random fluctuation in electrons contributing to measurement, due to electronic processes on the CCD sensor. This noise arises during the charge transfer, the charge-to-voltage conversion, and the analog-to-digital conversion. Every pixel on the sensor is subject to the same level of read noise, most of which is added by the amplifier.

Dark noise and read noise are due to the properties of the CCD sensor itself.

Photon noise (or **shot noise**) is inherent in any measurement of light, due to the fact that photons arrive at the detector randomly in time. This process is described by **Poisson statistics**. The probability P of measuring k events, given an expected value N is

$$P(k,N)=\frac{N^k e^{-N}}{k!}.$$

The standard deviation of the Poisson distribution is $N^{1/2}$. Therefore, if the average number of photons arriving at the detector is N, then the noise is $N^{1/2}$. Since the average number of photons is proportional to incident power, shot noise increases with $P^{1/2}$.

Signal-to-Noise Ratio and the Digitization of CCD

Total **noise** as a function of the number of electrons from all three contributing noises is given by

$$Noise\left[N_{\text{electrons}}\right] = \sqrt{\sigma_{\text{Photon}}^2 + \sigma_{\text{Dark}}^2 + \sigma_{\text{Read}}^2}\ ,$$

where

$\sigma_{\text{Photon}} = \sqrt{\varphi\eta\tau}$, $\sigma_{\text{Dark}} = \sqrt{I_{\text{Dark}}\tau}$, and $\sigma_{\text{Read}} = N_{\text{R}}$, I_{Dark} is dark current (electrons per second), and N_{R} is read noise (electrons, rms).

The quality of an image is determined by the **signal-to-noise ratio** (SNR). The signal can be defined as

$$Signal\left[N_{\text{electrons}}\right] = \varphi\eta\tau\ ,$$

where φ is the incident photon flux at the CCD (photons per second), η is the quantum efficiency of the CCD (electrons per photon), and τ is the integration time (in seconds). Therefore, SNR can be defined as:

$$SNR = \frac{\varphi\eta\tau}{\sqrt{\varphi\eta\tau + I_{\text{dark}}\tau + N_{\text{R}}^2}}\ .$$

It is best to use a CCD under photon-noise-limited conditions. If possible, it would be optimal to increase the integration time to increase the SNR and work in photon-noise-limited conditions. However, an increase in integration time is possible until reaching full-well capacity (saturation level):

$$SNR = \frac{\varphi\eta\tau}{\sqrt{\varphi\eta\tau}} = \sqrt{\varphi\eta\tau}.$$

The **dynamic range** can be derived as a ratio of full-well capacity and read noise. Digitization of the CCD output should be performed to maintain the dynamic range of the camera. Therefore, an analog-to-digital converter should support (at least) the same number of gray levels calculated by the CCD's dynamic range. Note that a high bit depth extends readout time, which is especially critical for large-format cameras.

CCD Sampling

The maximum spatial frequency passed by the CCD is one half of the sampling frequency—the **Nyquist frequency**. Any frequency higher than Nyquist will be aliased at lower frequencies.

Undersampling refers to a frequency where the sampling rate is not sufficient for the application. To assure maximum resolution of the microscope, the system should be optics limited, and the CCD should provide sampling that reaches the diffraction resolution limit. In practice, this means that at least two pixels should be dedicated to a distance of the resolution. Therefore, the maximum pixel spacing that provides the diffraction limit can be estimated as

$$d_{\text{pix}} = M\frac{0.61\lambda}{2NA},$$

where M is the magnification between the object and the CCD's plane. A graph representing how sampling influences the MTF of the system, including the CCD, is presented below.

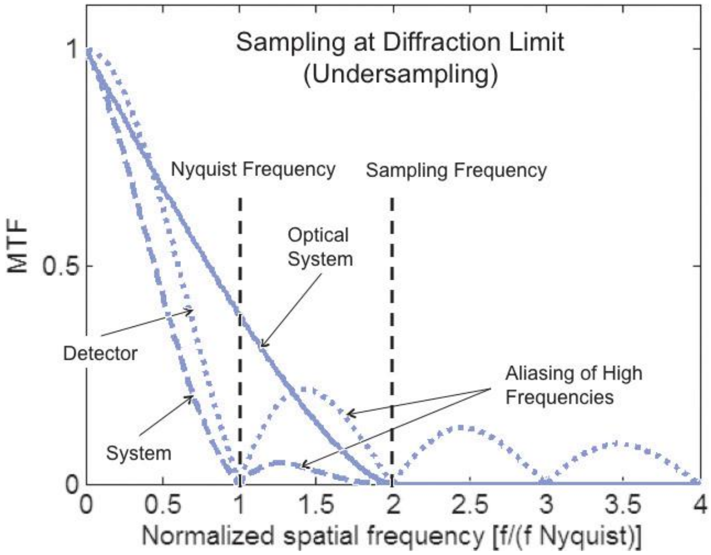

CCD Sampling (cont.)

Oversampling means that more than the minimum number of pixels, according to the Nyquist criteria, are available for detection, which does not imply excessive sampling of the image. However, excessive sampling decreases the available field of view of a microscope. The relation between the extent of field D, the number of pixels in the x and y directions (N_x and N_y, respectively), and the pixel spacing d_{pix} can be calculated from

$$D_x = \frac{N_x d_{\text{pix}-x}}{M}$$

and

$$D_y = \frac{N_y d_{\text{pix}-y}}{M}.$$

This assumes that the object area imaged by the microscope is equal to or larger than D. Therefore, the size of the microscope's field of view and the size of the CCD chip should be matched.

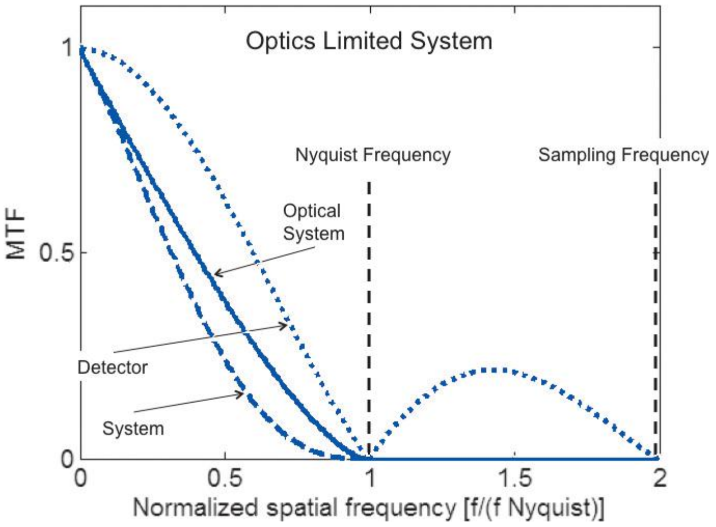

Equation Summary

Quantized Energy:

$$E = h\nu = hc/\lambda$$

Propagation of an electric field and wave vector:

$$E = A\sin\left(\omega t - kz + \varphi_o\right) = A\exp\left[-i\left(\omega t - kz + \varphi_o\right)\right]$$

$$\omega = \frac{2\pi}{T} = \frac{2\pi V_m}{\lambda} \qquad E(z,t) = E_x + E_y$$

$$E_x = A_x\exp\left[-i\left(\omega t - kz + \varphi_x\right)\right] \qquad E_y = A_y\exp\left[-i\left(\omega t - kz + \varphi_y\right)\right]$$

Refractive index:

$$n = \frac{c}{V_m}$$

Optical path length:

$$\mathrm{OPL} = nL$$

$$\mathrm{OPL} = \int_{P1}^{P2} n\,ds$$

$$ds^2 = dx^2 + dy^2 + dz^2$$

Optical path difference and phase difference:

$$\mathrm{OPD} = n_1L_1 - n_2L_2$$

$$\Delta\varphi = 2\pi\frac{\mathrm{OPD}}{\lambda}$$

TIR:

$$\theta_{cr} = \arcsin\left(\frac{n_1}{n_2}\right)$$

$$I = I_o\exp(-y/d)$$

$$d = \frac{\lambda}{4\pi n_1}\frac{1}{\sqrt{\sin^2\theta - \sin^2\theta_c}}$$

Coherence length:

$$l_c = \frac{\lambda^2}{\Delta\lambda}$$

Equation Summary (cont'd)

Two-beam interference:

$$I = EE^*$$

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left(\Delta\varphi\right)$$

$$\Delta\varphi = \varphi_2 - \varphi_1.$$

Contrast:

$$C = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$$

Diffraction grating equation:

$$m\lambda = d\left(\sin\beta \pm \sin\alpha\right)$$

Resolving power of diffraction grating:

$$\frac{\lambda}{\Delta\lambda} = mN$$

Free spectral range:

$$\Delta\lambda = \frac{m+1}{m}\lambda_1 - \lambda_1 = \frac{\lambda_1}{m}.$$

Newtonian equation:

$$xx' = ff'$$

$$xx' = -f'^2$$

Gaussian imaging equation:

$$\frac{n'}{z'} = \frac{n}{z} + \frac{1}{f_e}$$

$$\frac{1}{z'} = \frac{1}{z} + \frac{1}{f_e}$$

$$f_e = \frac{1}{\varphi} = -\frac{f}{n} = \frac{f'}{n'}$$

Transverse magnification:

$$M = \frac{h'}{h} = -\frac{x'}{f'} = -\frac{f}{x} = \frac{z'}{z} = \frac{-f}{z-f} = \frac{-\left(z' - f'\right)}{f'}$$

Longitudinal magnification:

$$\frac{\Delta z'}{\Delta z} = \left(-\frac{f'}{f}\right)M_1 M_2$$

Equation Summary (cont'd)

Optical transfer function:

$$\mathrm{OTF}(\rho) = \mathrm{MTF}(\rho)\exp\left[i\varphi(\rho)\right]$$

Modulation transfer function:

$$\mathrm{MTF}(\rho) = \frac{C(\rho)_{\mathrm{image}}}{C(\rho)_{\mathrm{object}}}$$

Field of view of microscope:

$$\mathrm{FOV} = \frac{FieldNumber}{M_{\mathrm{objective}}}[\mathrm{mm}]$$

Magnifying power:

$$\mathrm{MP} = \frac{u'}{u} = \frac{d_o\left(f - z'\right)}{f\left(z' - l\right)}$$

$$MP \approx \frac{250\ \mathrm{mm}}{f}$$

Magnification of the microscope objective:

$$M_{\mathrm{objective}} = \frac{OTL}{f_{\mathrm{objective}}}$$

Magnifying power of the microscope:

$$\mathrm{MP}_{\mathrm{microscope}} = M_{\mathrm{objective}}\mathrm{MP}_{\mathrm{eyepiece}} = -\frac{\mathrm{OTL}}{f_{\mathrm{objective}}}\frac{250\,\mathrm{mm}}{f_{\mathrm{eyepiece}}}$$

Numerical aperture:

$$NA = n\sin u$$

$$NA = NA'M_{\mathrm{objective}}$$

Airy disk:

$$d = \frac{1.22\lambda}{n\sin u} = \frac{1.22\lambda}{NA}$$

Rayleigh resolution limit:

$$d = \frac{0.61\lambda}{n\sin u} = \frac{0.61\lambda}{NA}$$

Sparrow resolution limit:

$$d = 0.5\lambda/NA$$

Equation Summary (cont'd)

Abbe resolution limit:

$$d = \frac{\lambda}{NA_{\text{objective}} + NA_{\text{condenser}}}$$

Resolving power of the microscope:

$$d_{\text{mic}} = \frac{d_{\text{eye}}}{M_{\text{min}}} = \frac{d_{\text{eye}}}{M_{\text{obj}} M_{\text{eyepiece}}}$$

Depth of focus and depth of field:

$$DOF = 2\Delta z' = \frac{n'\lambda}{\text{NA}'^2}$$

$$2\Delta z' = M^2_{\text{objective}} 2\Delta z \frac{n'}{n}$$

Depth perception of stereoscopic microscope:

$$\delta z = \frac{250\alpha_s}{M_{\text{microscope}} \tan\gamma}[\text{mm}]$$

Minimum perceived phase in phase contrast:

$$\Delta\varphi_{\text{min}} = \frac{C_{\text{ph-min}}\lambda}{4\pi\sqrt{N}}$$

Lateral resolution of the phase contrast:

$$d = f'_{\text{objective}} \frac{\lambda}{(r_{\text{AS}} + r_{\text{PR}})}$$

Intensity in DIC:

$$I \propto \sin^2\left(\frac{\pi s \dfrac{d\varphi_o}{dx}}{\lambda}\right)$$

$$I \propto \sin^2 \frac{\pi\left(\Delta_b \pm s\dfrac{d\varphi_o}{dx}\right)}{\lambda}$$

Retardation:

$$\Gamma = (n_e - n_o)t$$

Equation Summary (cont'd)

Birefringence:

$$\delta = \frac{2\pi OPD}{\lambda} = \frac{2\pi\Gamma}{\lambda}$$

Resolution of a confocal microscope:

$$d_{xy} \approx \frac{0.4\lambda}{NA},$$

$$d_z \approx \frac{1.4n\lambda}{NA^2}$$

Confocal pinhole width:

$$D_{\text{pinhole}} = \frac{0.5\lambda M}{NA}$$

Fluorescent emission:

$$F = \sigma QI$$

Probability of two-photon excitation:

$$n_{\text{a}} \propto \delta\left(\frac{P_{\text{avg}}^2}{\tau\nu^2}\right)\left(\pi\frac{NA^2}{hc\lambda}\right)^2$$

Intensity in FD-OCT:

$$I(k, z_{\text{o}}) = \int A_{\text{R}} A_{\text{S}}(z)\cos\left[k(z_{\text{m}} - z_{\text{o}})\right]dz$$

Phase-shifting equation:

$$I(x, y) = a(x, y) + b(x, y)\cos(\varphi + n\Delta\varphi)$$

Three-image algorithm (π/2 shifts):

$$\varphi = \arctan\left(\frac{I_3 - I_2}{I_1 - I_2}\right)$$

Four-image algorithm:

$$\varphi = \arctan\left(\frac{I_4 - I_2}{I_1 - I_3}\right)$$

Five-image algorithm:

$$\varphi = \arctan\left(\frac{2[I_2 - I_4]}{I_1 - I_3 + I_5}\right)$$

Equation Summary (cont'd)

Image reconstruction structure illumination (sectioning):

$$I = \sqrt{\left(I_0 - I_{2\pi/3}\right)^2 + \left(I_0 - I_{4\pi/3}\right)^2 + \left(I_{2\pi/3} - I_{4\pi/3}\right)^2}$$

Poisson statistics:

$$P\left(k, N\right) = \frac{N^k e^{-N}}{k!}$$

Noise:

$$Noise\left[N_{\text{electrons}}\right] = \sqrt{\sigma_{\text{Photon}}^2 + \sigma_{\text{Dark}}^2 + \sigma_{\text{Read}}^2}$$

$$\sigma_{\text{Photon}} = \sqrt{\varphi\eta\tau}\ ,\ \sigma_{\text{Dark}} = \sqrt{I_{\text{Dark}}\tau}\ ,\ \sigma_{\text{Read}} = N_{\text{R}}\ ,\ I_{\text{Dark}}$$

Signal-to-noise ratio (SNR):

$$\text{Signal}\left[N_{\text{electrons}}\right] = \varphi\eta\tau$$

$$\text{SNR} = \frac{\varphi\eta\tau}{\sqrt{\varphi\eta\tau + I_{\text{dark}}\tau + N_{\text{R}}^2}}$$

$$\text{SNR} = \frac{\varphi\eta\tau}{\sqrt{\varphi\eta\tau}} = \sqrt{\varphi\eta\tau}$$

Bibliography

M. Bates, B. Huang, G.T. Dempsey, and X. Zhuang, "Multicolor Super-Resolution Imaging with Photo-Switchable Fluorescent Probes." *Science* **317**, 1749 (2007).

J. R. Benford, Microscope objectives, Chapter 4 (p. 178), in *Applied Optics and Optical Engineering, Vol. III,* R. Kingslake, ed., Academic Press, New York, NY (1965).

M. Born and E. Wolf, *Principles of Optics*, Sixth Edition, Cambridge University Press, Cambridge, UK (1997).

S. Bradbury and P. J. Evennett, *Contrast Techniques in Light Microscopy*, BIOS Scientific Publishers, Oxford, UK (1996).

T. Chen, T. Milster, S. K. Park, B. McCarthy, D. Sarid, C. Poweleit, and J. Menendez, "Near-field solid immersion lens microscope with advanced compact mechanical design." *Optical Engineering* **45**(10), 103002 (2006).

T. Chen, T. D. Milster, S. H. Yang, and D. Hansen, "Evanescent imaging with induced polarization by using a solid immersion lens." *Optics Letters* **32**(2), 124–126 (2007).

J.-X. Cheng and X. S. Xie, "Coherent anti-stokes raman scattering microscopy: instrumentation, theory, and applications." *J. Phys. Chem. B* **108***,* 827-840 (2004).

M. A. Choma, M. V. Sarunic, C. Yang, and J. A. Izatt, "Sensitivity advantage of swept source and Fourier domain optical coherence tomography." *Opt. Express* **11**, 2183–2189 (2003).

J. F. de Boer, B. Cense, B. H. Park, M.C. Pierce, G. J. Tearney, and B. E. Bouma, "Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography." *Opt Lett* **28**, 2067–2069 (2003).

E. Dereniak, materials for SPIE Short Course on Imaging Spectrometers, SPIE, Bellingham, WA (2005).

E. Dereniak, *Geometrical Optics*, Cambridge University Press, Cambridge, UK (2008).

M. Descour, materials for OPTI 412, "Optical Instrumentation," University of Arizona (2000).

Bibliography

D. Goldstein, *Polarized Light*, Second Edition, Marcel Dekker, New York, NY (1993).

D. S. Goodman, Basic optical instruments, Chapter 4, in Geometrical and Instrumental Optics, D. Malacara, ed., Academic Press, New York, NY (1988).

J. Goodman, *Introduction to Fourier Optics*, 3rd Edition, Roberts and Company Publishers, Greenwood Village, CO (2004).

E. P. Goodwin and J. C. Wyant, *Field Guide to Interferometric Optical Testing*, SPIE Press, Bellingham, WA (2006).

J. E. Greivenkamp, *Field Guide to Geometrical Optics*, SPIE Press, Bellingham, WA (2004).

H. Gross, F. Blechinger, and B. Achtner, *Handbook of Optical Systems*, *Vol. 4*: *Survey of Optical Instruments*, Wiley-VCH, Germany (2008).

M. G. L. Gustafsson, "Nonlinear structured-illumination microscopy: Wide-field fluorescence imaging with theoretically unlimited resolution." *PNAS*, **102**(37) 13081–13086 (2005).

M. G. L. Gustafsson, "Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy." *Journal of Microscopy*, **198**(2), 82-87 (2000).

Gerd Häusler and Michael Walter Lindner, ""Coherence Radar" and "Spectral Radar"—New tools for dermatological diagnosis." *Journal of Biomedical Optics* **3**(1), 21–31 (1998).

E. Hecht, *Optics*, Fourth Edition, Addison-Wesley, Upper Saddle River, New Jersey (2002).

S. W. Hell, "Far-field optical nanoscopy." *Science* **316**, 1153 (2007).

B. Herman and J. Lemasters, *Optical Microscopy: Emerging Methods and Applications*, Academic Press, New York, NY (1993).

P. Hobbs, *Building Electro-Optical Systems: Making It All Work*, Wiley and Sons, New York, NY (2000).

Bibliography

G. Holst and T. Lomheim, *CMOS/CCD Sensors and Camera Systems*, JCD Publishing, Winter Park, FL (2007).

B. Huang, W. Wang, M. Bates, and X. Zhuang, "Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy." *Science* **319**, 810 (2008).

R. Huber, M. Wojtkowski, and J. G. Fujimoto, "Fourier domain mode locking (FDML): A new laser operating regime and applications for optical coherence tomography." *Opt. Express* **14**, 3225–3237 (2006).

Invitrogen, http://www.invitrogen.com

R. Jozwicki, *Teoria Odwzorowania Optycznego* (in Polish), PWN (1988).

R. Jozwicki, *Optyka Instrumentalna* (in Polish), WNT (1970).

R. Leitgeb, C. K. Hitzenberger, and A. F. Fercher, "Performance of Fourier-domain versus time-domain optical coherence tomography." *Opt. Express* **11**, 889–894 (2003).

D. Malacara and B. Thompson, Eds., *Handbook of Optical Engineering*, Marcel Dekker, New York, NY (2001).

D. Malacara and Z. Malacara, *Handbook of Optical Design*, Marcel Dekker, New York, NY (1994).

D. Malacara, M. Servin, and Z. Malacara, *Interferogram Analysis for Optical Testing*, Marcel Dekker, New York, NY (1998).

D. Murphy, *Fundamentals of Light Microscopy and Electronic Imaging*, Wiley-Liss, Wilmington, DE (2001).

P. Mouroulis and J. Macdonald, *Geometrical Optics and Optical Design*, Oxford University Press, New York, NY (1997).

M. A. A. Neil, R. Juškaitis, and T. Wilson, "Method of obtaining optical sectioning by using structured light in a conventional microscope." *Optics Letters*, **22**(24), 1905–1907 (1997).

Nikon Microscopy U, http://www.microscopyu.com/

Bibliography

C. Palmer (Erwin Loewen, First Edition), *Diffraction Grating Handbook*, Newport Corp. (2005).

K. Patorski, *Handbook of the Moiré Fringe Technique*, Elsevier, Oxford, UK(1993).

J. Pawley, Ed., *Biological Confocal Microscopy*, Third Edition, Springer, New York, NY (2006).

M. C. Pierce, D. J. Javier and R. Richards-Kortum, "Optical contrast agents and imaging systems for detection and diagnosis of cancer." *Int. J. Cancer* **123**, 1979–1990 (2008).

M. Pluta, *Advanced Light Microscopy, Volume One: Principle and Basic Properties*, PWN and Elsevier, New York, NY (1988).

M. Pluta, *Advanced Light Microscopy Volume Two: Specialized Methods*, PWN and Elsevier, New York, NY (1989).

M. Pluta, *Advanced Light Microscopy, Volume Three: Measuring Techniques*, PWN, Warsaw, Poland; and North Holland, Amsterdam, Holland (1993).

E. O. Potma, C. L. Evans, and X. S. Xie, "Heterodyne coherent anti-Stokes Raman scattering (CARS) imaging." *Optics Letters*, **31**(2), 241–243 (2006).

D. W. Robinson, G. T.Reed,, Eds., *Interferogram Analysis*, IOP Publishing, Bristol, UK (1993).

M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)." *Nature Methods*, **3**, 793–796 (2006).

B. Saleh and M. C. Teich, *Fundamentals of Photonics*, SecondEdition, Wiley, New York, NY (2007).

J. Schwiegerling, *Field Guide to Visual and Ophthalmic Optics*, SPIE Press, Bellingham, WA (2004).

W. Smith, *Modern Optical Engineering,* Third Edition, McGraw-Hill, New York, NY (2000).

D. Spector and R. Goldman, Eds., *Basic Methods in Microscopy*, Cold Spring Harbor Laboratory Press, Woodbury, NY (2006).

Bibliography

Thorlabs website resources: http://www.thorlabs.com/

P. Török and F. J. Kao, Eds., *Optical Imaging and Microscopy*, Springer, New York, NY (2007).

Veeco Optical Library entry, http://www.veeco.com/pdf/Optical_Library/Challenges_in_White_Light.pdf

R. Wayne, *Light and Video Microscopy*, Elsevier (reprinted by Academic Press), New York, NY (2009).

H. Yu, P. C. Cheng, P. C. Li, F. J. Kao, Eds., *Multi Modality Microscopy*, World Scientific, Hackensack, NJ (2006).

S. H. Yun, G. J. Tearney, B. J. Vakoc, M. Shishkov, W. Y. Oh, A. E. Desjardins, M. J. Suter, R. C. Chan, J. A. Evans, I. K. Jang, N. S. Nishioka, J. F. de Boer, and B. E. Bouma, "Comprehensive volumetric optical microscopy *in vivo*." *Nature Med* **12**, 1429–1433 (2006).

Zeiss Corporation, http://www.zeiss.com

Index

4Pi, 64, 109, 110
4Pi microscopy, 109

Abbe number, 25
Abbe resolution limit, 39
absorption filter, 60
achromatic objectives, 51
achromats, 51
Airy disk, 38
Amici lens, 55
amplitude contrast, 72
amplitude object, 63
amplitude ratio, 15
amplitude splitting, 17
analyzer, 83
anaxial illumination, 73
angular frequency, 4
angular magnification, 23
angular resolving power, 40
anisotropic media, propagation of light in, 9
aperture stop, 24
apochromats, 51
arc lamp, 58
array microscope, 113
array microscopy, 64, 113
astigmatism, 28
axial chromatic aberration, 26
axial resolution, 86

bandpass filter, 60
Bayer mask, 117
Bertrand lens, 45, 55
bias, 80
Brace Köhler compensator, 85
bright field, 64, 65, 76

charge-coupled device (CCD), 115
chief ray, 24
chromatic aberration, 26
chromatic resolving power, 20
circularly polarized light, 10
coherence length, 11
coherence scanning interferometry, 98
coherence time, 11
coherent, 11
coherent anti-Stokes Raman scattering (CARS), 64, 111
coma, 27
compensators, 83, 85
complex amplitudes, 12
compound microscope, 31
condenser's diaphragm, 46
cones, 32
confocal microscopy, 64, 86
contrast, 12
contrast of fringes, 15
cornea, 32
cover glass, 56
cover slip, 56
critical angle, 7
critical illumination, 46
cutoff frequency, 30

dark current, 118
dark field, 64, 65
dark noise, 118
de Sénarmont compensator, 85
depth of field, 41
depth of focus, 41
depth perception, 47
destructive interference, 19

DIA, 44
DIC microscope design, 80
dichroic beam splitter, 91
dichroic mirror, 91
dielectric constant, 3
differential interference contrast (DIC), 64–65, 67, 79
diffraction, 18
diffraction grating, 19
diffraction orders, 19
digital microscopy, 114
digitiziation of CCD, 119
dispersion, 25
dispersion staining, 78
distortion, 28
dynamic range, 119

edge filter, 60
effective focal length, 22
electric fields, 12
electric vector, 10
electromagnetic spectrum, 2
electron-multiplying CCDs, 117
emission filter, 91
empty magnification, 40
entrance pupil, 24
entrance window, 24
epi-illumination, 44
epithelial cells, 2
evanescent wave microscopy, 105
excitation filter, 91
exit pupil, 24
exit window, 24
extraordinary wave, 9
eye, 32, 46
eye point, 48
eye relief, 48
eye's pupil, 46
eyepiece, 48

Fermat's principle, 5
field curvature, 28
field diaphragm, 46
field number (FN), 34, 48
field stop, 24, 34
filter turret, 92
filters, 91
finite tube length, 34
five-image technique, 102
fluorescence, 90
fluorescence microscopy, 64
fluorescent emission, 94
fluorites, 51
fluorophore quantum yield, 94
focal length, 21
focal plane, 21
focal point, 21
Fourier-domain OCT (FDOCT), 99
four-image technique, 102
fovea, 32
frame-transfer architecture, 116
Fraunhofer diffraction, 18
frequency of light, 1
Fresnel diffraction, 18
Fresnel reflection, 6
frustrated total internal reflection, 7
full-frame architecture, 116

gas-arc discharge lamp, 58
Gaussian imaging equation, 22
geometrical aberrations, 25

Index

geometrical optics, 21
Goos-Hänchen effect, 8
grating equation, 19–20

half bandwidth (HBW), 16
halo effect, 65, 71
halogen lamp, 58
high-eye point, 49
Hoffman modulation contrast, 75
Huygens, 48–49

I^5M, 64
image comparison, 65
image formation, 22
image space, 21
immersion liquid, 57
incandescent lamp, 58
incidence, 6
infinity-corrected objective, 35
infinity-corrected systems, 35
infrared radiation (IR), 2
intensity of light, 12
interference, 12
interference filter, 16, 60
interference microscopy, 64, 98
interference order, 16
interfering beams, 15
interferometers, 17
interline transfer architecture, 116
intermediate image plane, 46
inverted microscope, 33
iris, 32

Köhler illumination, 43–45

laser scanning, 64
lasers, 96
laser-scanning microscopy (LSM), 97
lateral magnification, 23
lateral resolution, 71
laws of reflection and refraction, 6
lens, 32
light sheet, 112
light sources for microscopy, common, 58
light-emitting diodes (LEDs), 59
limits of light microscopy, 110
linearly polarized light, 10
line-scanning confocal microscope, 87
Linnik, 101
long working distance objectives, 53
longitudinal magnification, 23
low magnification objectives, 54

Mach-Zehnder interferometer, 17
macula, 32
magnetic permeability, 3
magnification, 21, 23
magnifier, 31
magnifying power (MP), 37
marginal ray, 24
Maxwell's equations, 3
medium permittivity, 3
mercury lamp, 58
metal halide lamp, 58
Michelson, 17, 53, 98, 100, 101

Index

minimum focus distance, 32
Mirau, 53, 98, 101
modulation contrast microscopy (MCM), 74
modulation transfer function (MTF), 29
molecular absorption cross section, 94
monochromatic light, 11
multi-photon fluorescence, 95
multi-photon microscopy, 64
multiple beam interference, 16

nature of light, 1
near point, 32
negative birefringence, 9
negative contrast, 69
neutral density (ND), 60
Newtonian equation, 22
Nomarski interference microscope, 82
Nomarski prism, 62, 80
non-self-luminous, 63
numerical aperture (NA), 38, 50
Nyquist frequency, 120

object space, 21
objective, 31
objective correction, 50
oblique illumination, 73
optic axis, 9
optical coherence microscopy (OCM), 98
optical coherence tomography (OCT), 98
optical density (OD), 60
optical path difference (OPD), 5
optical path length (OPL), 5
optical power, 22
optical profilometry, 98, 101
optical rays, 21
optical sectioning, 103
optical spaces, 21
optical staining, 77
optical transfer function, 29
optical tube length, 34
ordinary wave, 9
oversampling, 121

PALM, 64, 110
paraxial optics, 21
parfocal distance, 34
partially coherent, 11
particle model, 1
performance metrics, 29
phase-advancing objects, 68
phase contrast microscope, 70
phase contrast, 64, 65, 66
phase object, 63
phase of light, 4
phase retarding objects, 68
phase-shifting algorithms, 102
phase-shifting interferometry (PSI), 98, 100
photobleaching, 94
photodiode, 115
photon noise, 118
Planck's constant, 1

Index

point spread function (PSF), 29
point-scanning confocal microscope, 87
Poisson statistics, 118
polarization microscope, 84
polarization microscopy, 64, 83
polarization of light, 10
polarization prisms, 61
polarization states, 10
polarizers, 61–62
positive birefringence, 9
positive contrast, 69
pupil function, 29

quantum efficiency (QE), 97, 117
quasi-monochromatic, 11
quenching, 94

Raman microscopy, 64, 111
Raman scattering, 111
Ramsden eyepiece, 48
Rayleigh resolution limit, 39
read noise, 118
real image, 22
red blood cells, 2
reflectance coefficients, 6
reflected light, 6, 16
reflection law, 6
reflective grating, 20
refracted light, 6
refraction law, 6
refractive index, 4
RESOLFT, 64, 110

resolution limit, 39
retardation, 83
retina, 32
RGB filters, 117
Rheinberg illumination, 77
rods, 32

Sagnac interferometer, 17
sample conjugate, 46
sample path, 44
scanning approaches, 87
scanning white light interferometry, 98
Selective plane illumination microscopy (SPIM), 64, 112
self-luminous, 63
semi-apochromats, 51
shading-off effect, 71
shearing interferometer, 17
shearing interferometry, 79
shot noise, 118
SI, 64
sign convention, 21
signal-to-noise ratio (SNR), 119
Snell's law, 6
solid immersion, 106
solid immersion lenses (SILs), 106
Sparrow resolution limit, 30
spatial coherence, 11
spatial resolution, 86
spectral-domain OCT (SDOCT), 99
spectrum of microscopy, 2
spherical aberration, 27
spinning-disc confocal imaging, 88
stereo microscopes, 47

stimulated emission depletion (STED), 107
stochastic optical reconstruction microscopy (STORM), 64, 108, 110
Stokes shift, 90
Strehl ratio, 29, 30
structured illumination, 103
super-resolution microscopy, 64
swept-source OCT (SSOCT), 99

telecentricity, 36
temporal coherence, 11, 13–14
thin lenses, 21
three-image technique, 102
total internal reflection (TIR), 7
total internal reflection fluorescence (TIRF) microscopy, 105
transmission coefficients, 6
transmission grating, 20
transmitted light, 16
transverse chromatic aberration, 26
transverse magnification, 23
tube length, 34
tube lens, 35
tungsten-argon lamp, 58

ultraviolet radiation (UV), 2
undersampling, 120
uniaxial crystals, 9
upright microscope, 33
useful magnification, 40
UV objectives, 54

velocity of light, 1
vertical scanning interferometry (VSI), 98, 100
virtual image, 22
visibility, 12
visibility in phase contrast, 69
visible spectrum (VIS), 2
visual stereo resolving power, 47

water immersion objectives, 54
wave aberrations, 25
wave equations, 3
wave group, 11
wave model, 1
wave number, 4
wave plate compensator, 85
wavefront propagation, 4
wavefront splitting, 17
wavelength of light, 1
Wollaston prism, 61, 80
working distance (WD), 50

x-ray radiation, 2

Tomasz S. Tkaczyk is an Assistant Professor of Bioengineering and Electrical and Computer Engineering at Rice University, Houston, Texas, where he develops modern optical instrumentation for biological and medical applications. His primary research is in microscopy, including endo-microscopy, cost-effective high-performance optics for diagnostics, and multi-dimensional imaging (snapshot hyperspectral microscopy and spectro-polarimetry).

Professor Tkaczyk received his M.S. and Ph.D. from the Institute of Micromechanics and Photonics, Department of Mechatronics, Warsaw University of Technology, Poland. Beginning in 2003, after his postdoctoral training, he worked as a research professor at the College of Optical Sciences, University of Arizona. He joined Rice University in the summer of 2007.

Fundamental Optics

1

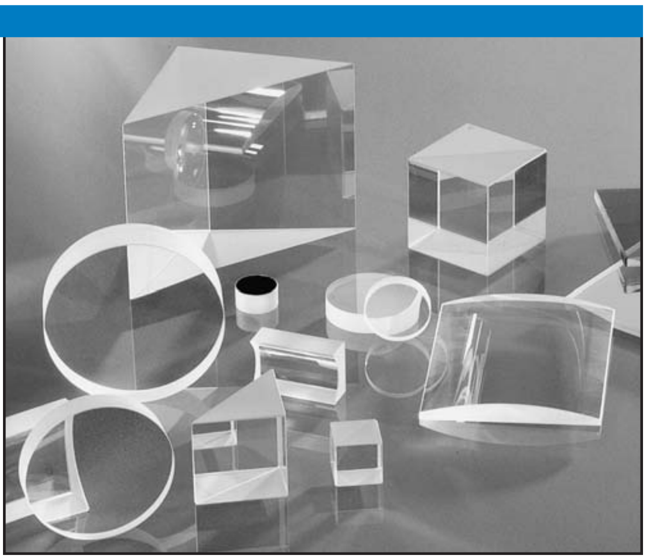

Introduction	**1.2**
Paraxial Formulas	**1.3**
Imaging Properties of Lens Systems	**1.6**
Lens Combination Formulas	**1.8**
Performance Factors	**1.11**
Lens Shape	**1.17**
Lens Combinations	**1.18**
Diffraction Effects	**1.20**
Lens Selection	**1.23**
Spot Size	**1.26**
Aberration Balancing	**1.27**
Definition of Terms	**1.29**
Paraxial Lens Formulas	**1.32**
Principal-Point Locations	**1.36**
Prisms	**1.37**
Polarization	**1.41**
Waveplates	**1.46**
Etalons	**1.49**
Ultrafast Theory	**1.52**

1ch_FundamentalOptics_Final_a.qxd 6/15/2009 2:28 PM Page 1.2

Introduction

The process of solving virtually any optical engineering problem can be broken down into two main steps. First, paraxial calculations (first order) are made to determine critical parameters such as magnification, focal length(s), clear aperture (diameter), and object and image position. These paraxial calculations are covered in the next section of this chapter.

Second, actual components are chosen based on these paraxial values, and their actual performance is evaluated with special attention paid to the effects of aberrations. A truly rigorous performance analysis for all but the simplest optical systems generally requires computer ray tracing, but simple generalizations can be used, especially when the lens selection process is confined to a limited range of component shapes.

In practice, the second step may reveal conflicts with design constraints, such as component size, cost, or product availability. System parameters may therefore require modification.

Because some of the terms used in this chapter may not be familiar to all readers, a glossary of terms is provided in *Definition of Terms*.

Finally, it should be noted that the discussion in this chapter relates only to systems with uniform illumination; optical systems for Gaussian beams are covered in *Gaussian Beam Optics*.

Engineering Support

CVI Melles Griot maintains a staff of knowledgeable, experienced applications engineers at each of our facilities worldwide. The information given in this chapter is sufficient to enable the user to select the most appropriate catalog lenses for the most commonly encountered applications. However, when additional optical engineering support is required, our applications engineers are available to provide assistance. Do not hesitate to contact us for help in product selection or to obtain more detailed specifications on CVI Melles Griot products.

THE OPTICAL ENGINEERING PROCESS

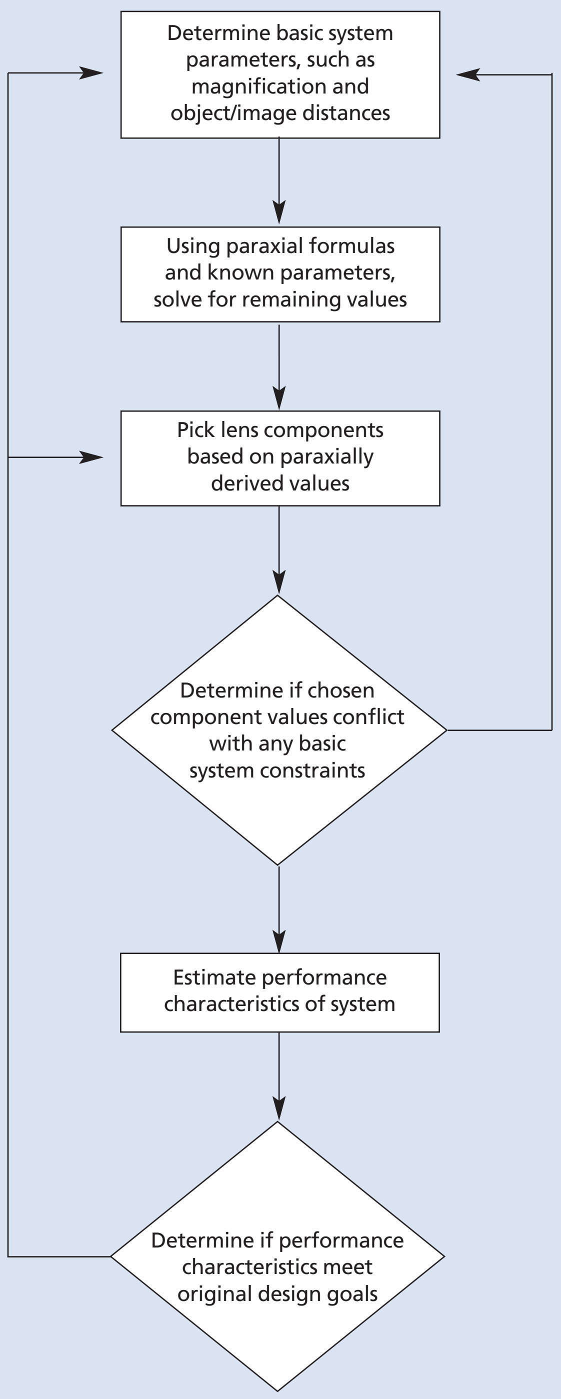
Determine basic system parameters, such as magnification and object/image distances
Using paraxial formulas and known parameters, solve for remaining values
Pick lens components based on paraxially derived values
Determine if chosen component values conflict with any basic system constraints
Estimate performance characteristics of system
Determine if performance characteristics meet original design goals

1ch_FundamentalOptics_Final_a.qxd 6/15/2009 2:28 PM Page 1.3

Paraxial Formulas

Sign Conventions

The validity of the paraxial lens formulas is dependent on adherence to the following sign conventions:

For lenses: (refer to figure 1.1)

s is + for object to left of H (the first principal point)

s is − for object to right of H

s'' is + for image to right of H″ (the second principal point)

s'' is − for image to left of H″

m is + for an inverted image

m is − for an upright image

For mirrors:

f is + for convex (diverging) mirrors

f is − for concave (converging) mirrors

s is + for object to left of H

s is − for object to right of H

s'' is − for image to right of H″

s'' is + for image to left of H″

m is + for an inverted image

m is − for an upright image

When using the thin-lens approximation, simply refer to the left and right of the lens.

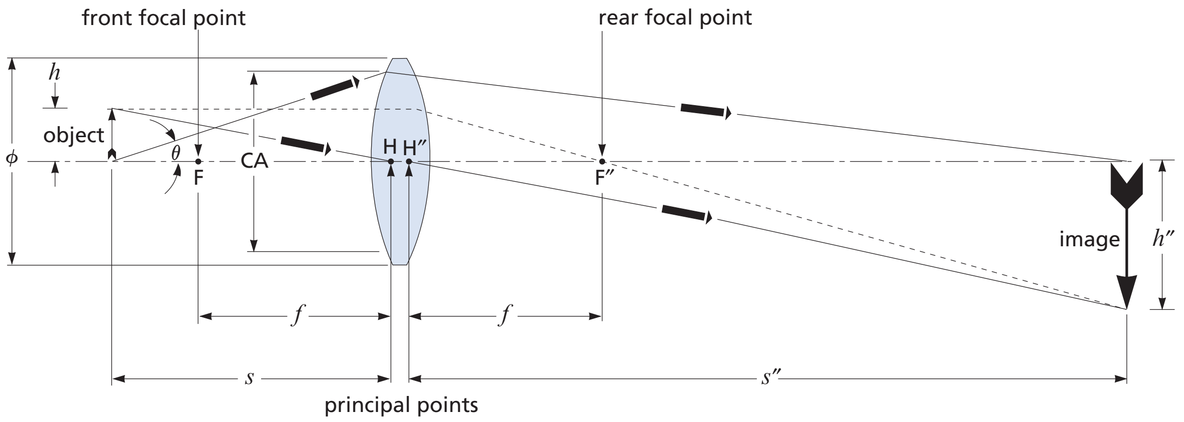

Note location of object and image relative to front and rear focal points.

ϕ = lens diameter

CA = clear aperture (typically 90% of ϕ)

f = effective focal length (EFL) which may be positive (as shown) or negative. f represents both FH and H″F″, assuming lens is surrounded by medium of index 1.0

m = $s''/s = h''/h$ = magnification or conjugate ratio, said to be infinite if either s'' or s is infinite

θ = arcsin (CA/2s)

s = object distance, positive for object (whether real or virtual) to the left of principal point H

s'' = image distance (s and s'' are collectively called conjugate distances, with object and image in conjugate planes), positive for image (whether real or virtual) to the right of principal point H″

h = object height

h'' = image height

Figure 1.1 **Sign conventions**

Typically, the first step in optical problem solving is to select a system focal length based on constraints such as magnification or conjugate distances (object and image distance). The relationship among focal length, object position, and image position is given by

$$\frac{1}{f}=\frac{1}{s}+\frac{1}{s''} \tag{1.1}$$

This formula is referenced to figure 1.1 and the sign conventions given in *Sign Conventions*.

By definition, magnification is the ratio of image size to object size or

$$m=\frac{s''}{s}=\frac{h''}{h}. \tag{1.2}$$

This relationship can be used to recast the first formula into the following forms:

$$f=m\frac{(s+s'')}{(m+1)^2} \tag{1.3}$$

$$f=\frac{sm}{m+1} \tag{1.4}$$

$$f=\frac{s+s''}{m+2+\frac{1}{m}} \tag{1.5}$$

$$s(m+1)=s+s'' \tag{1.6}$$

where $(s+s'')$ is the approximate object-to-image distance.

With a real lens of finite thickness, the image distance, object distance, and focal length are all referenced to the principal points, not to the physical center of the lens. By neglecting the distance between the lens' principal points, known as the hiatus, $s+s''$ becomes the object-to-image distance. This simplification, called the thin-lens approximation, can speed up calculation when dealing with simple optical systems.

Example 1: Object outside Focal Point

A 1-mm-high object is placed on the optical axis, 200 mm left of the left principal point of a LDX-25.0-51.0-C ($f=50$ mm). Where is the image formed, and what is the magnification? (See figure 1.2.)

$$\frac{1}{s''}=\frac{1}{f}-\frac{1}{s}$$

$$\frac{1}{s''}=\frac{1}{50}-\frac{1}{200}$$

$$s''=66.7\ \text{mm}$$

$$m=\frac{s''}{s}=\frac{66.7}{200}=0.33$$

or real image is 0.33 mm high and inverted.

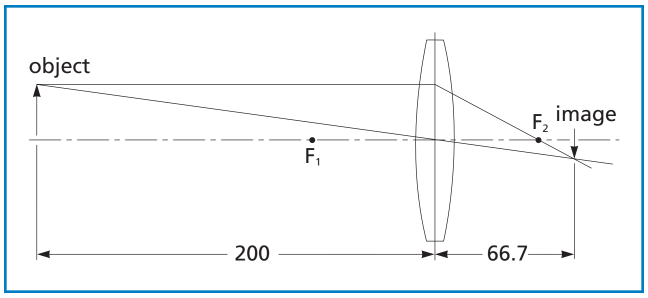

Figure 1.2 **Example 1** ($f=50$ mm, $s=200$ mm, $s''=66.7$ mm)

Example 2: Object inside Focal Point

The same object is placed 30 mm left of the left principal point of the same lens. Where is the image formed, and what is the magnification? (See figure 1.3.)

$$\frac{1}{s''}=\frac{1}{50}-\frac{1}{30}$$

$$s''=-75\ \text{mm}$$

$$m=\frac{s''}{s}=\frac{-75}{30}=-2.5$$

or virtual image is 2.5 mm high and upright.

In this case, the lens is being used as a magnifier, and the image can be viewed only back through the lens.

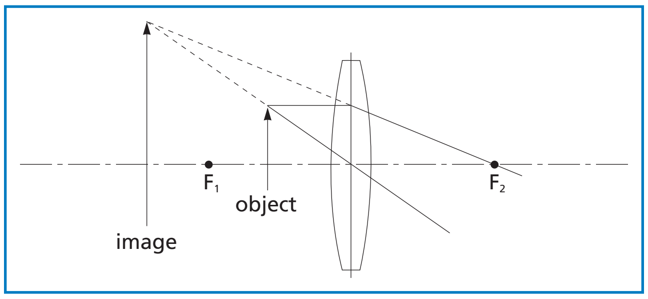

Figure 1.3 **Example 2** ($f=50$ mm, $s=30$ mm, $s''=-75$ mm)

Example 3: Object at Focal Point

A 1-mm-high object is placed on the optical axis, 50 mm left of the first principal point of an LDK-50.0-52.2-C ($f=-50$ mm). Where is the image formed, and what is the magnification? (See figure 1.4.)

$$\frac{1}{s''}=\frac{1}{-50}-\frac{1}{50}$$

$$s''=-25\ \text{mm}$$

$$m=\frac{s''}{s}=\frac{-25}{50}=-0.5$$

or virtual image is 0.5 mm high and upright.

Fundamental Optics
Gaussian Beam Optics
Optical Specifications
Material Properties
Optical Coatings

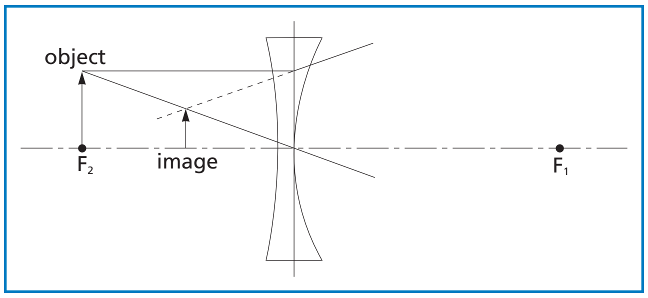

Figure 1.4 **Example 3** ($f = -50$ mm, $s = 50$ mm, $s'' = -25$ mm)

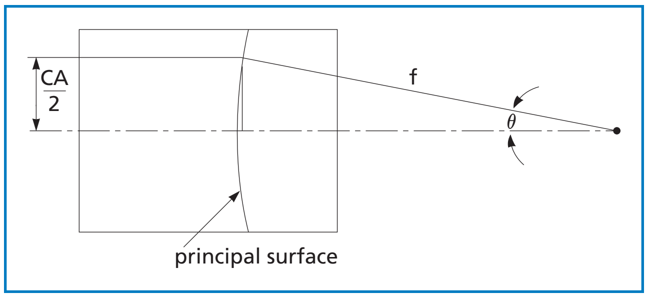

Figure 1.5 **F-number and numerical aperture**

A simple graphical method can also be used to determine paraxial image location and magnification. This graphical approach relies on two simple properties of an optical system. First, a ray that enters the system parallel to the optical axis crosses the optical axis at the focal point. Second, a ray that enters the first principal point of the system exits the system from the second principal point parallel to its original direction (i.e., its exit angle with the optical axis is the same as its entrance angle). This method has been applied to the three previous examples illustrated in figures 1.2 through 1.4. Note that by using the thin-lens approximation, this second property reduces to the statement that a ray passing through the center of the lens is undeviated.

F-NUMBER AND NUMERICAL APERTURE

The paraxial calculations used to determine the necessary element diameter are based on the concepts of focal ratio (f-number or f/#) and numerical aperture (NA). The f-number is the ratio of the focal length of the lens to its "effective" diameter, the clear aperture (CA).

$$\text{f-number} = \frac{f}{\text{CA}} . \tag{1.7}$$

To visualize the f-number, consider a lens with a positive focal length illuminated uniformly with collimated light. The f-number defines the angle of the cone of light leaving the lens which ultimately forms the image. This is an important concept when the throughput or light-gathering power of an optical system is critical, such as when focusing light into a monochromator or projecting a high-power image.

The other term used commonly in defining this cone angle is numerical aperture. The NA is the sine of the angle made by the marginal ray with the optical axis. By referring to figure 1.5 and using simple trigonometry, it can be seen that

$$\text{NA} = \sin\theta = \frac{\text{CA}}{2f} \tag{1.8}$$

and

$$\text{NA} = \frac{1}{2(\text{f-number})} . \tag{1.9}$$

Ray f-numbers can also be defined for any arbitrary ray if its conjugate distance and the diameter at which it intersects the principal surface of the optical system are known.

NOTE

Because the sign convention given previously is not used universally in all optics texts, the reader may notice differences in the paraxial formulas. However, results will be correct as long as a consistent set of formulas and sign conventions is used.

Imaging Properties of Lens Systems

THE OPTICAL INVARIANT

To understand the importance of the NA, consider its relation to magnification. Referring to figure 1.6,

$$\text{NA (object side)} = \sin\theta = \frac{\text{CA}}{2s} \tag{1.10}$$

$$\text{NA}'' \text{ (image side)} = \sin\theta'' = \frac{\text{CA}}{2s''} \tag{1.11}$$

which can be rearranged to show

$$\text{CA} = 2s\ \sin\theta \tag{1.12}$$

and

$$\text{CA} = 2s''\sin\theta'' \tag{1.13}$$

leading to

$$\frac{s''}{s} = \frac{\sin\theta}{\sin\theta''} = \frac{\text{NA}}{\text{NA}''}. \tag{1.14}$$

Since $\frac{s''}{s}$ is simply the magnification of the system,

we arrive at

$$m = \frac{\text{NA}}{\text{NA}''}. \tag{1.15}$$

The magnification of the system is therefore equal to the ratio of the NAs on the object and image sides of the system. This powerful and useful result is completely independent of the specifics of the optical system, and it can often be used to determine the optimum lens diameter in situations involving aperture constraints.

When a lens or optical system is used to create an image of a source, it is natural to assume that, by increasing the diameter (ϕ) of the lens, thereby increasing its CA, we will be able to collect more light and thereby produce a brighter image. However, because of the relationship between magnification and NA, there can be a theoretical limit beyond which increasing the diameter has no effect on light-collection efficiency or image brightness.

Since the NA of a ray is given by CA/2s, once a focal length and magnification have been selected, the value of NA sets the value of CA. Thus, if one is dealing with a system in which the NA is constrained on either the object or image side, increasing the lens diameter beyond this value will increase system size and cost but will not improve performance (i.e., throughput or image brightness). This concept is sometimes referred to as the optical invariant.

SAMPLE CALCULATION

To understand how to use this relationship between magnification and NA, consider the following example.

Example: System with Fixed Input NA

Two very common applications of simple optics involve coupling light into an optical fiber or into the entrance slit of a monochromator. Although these problems appear to be quite different, they both have the same limitation — they have a fixed NA. For monochromators, this limit is usually expressed in terms of the f-number. In addition to the fixed NA, they both have a fixed entrance pupil (image) size.

Suppose it is necessary, using a singlet lens, to couple the output of an incandescent bulb with a filament 1 mm in diameter into an optical fiber as shown in figure 1.7. Assume that the fiber has a core diameter of 100 μm and an NA of 0.25, and that the design requires that the total distance from the source to the fiber be 110 mm. Which lenses are appropriate?

By definition, the magnification must be 0.1. Letting $s+s''$ total 110 mm (using the thin-lens approximation), we can use equation 1.3,

$$f = m\frac{(s+s'')}{(m+1)^2}, \qquad \text{(see eq. 1.3)}$$

to determine that the focal length is 9.1 mm. To determine the conjugate distances, s and s'', we utilize equation 1.6,

$$s(m+1) = s+s'', \qquad \text{(see eq. 1.6)}$$

and find that $s = 100$ mm and $s'' = 10$ mm.

We can now use the relationship NA = CA/2s or NA″ = CA/2s'' to derive CA, the optimum clear aperture (effective diameter) of the lens.

With an image NA of 0.25 and an image distance (s'') of 10 mm,

$$0.25 = \frac{\text{CA}}{20}$$

and

$$\text{CA} = 5\text{ mm}.$$

Accomplishing this imaging task with a single lens therefore requires an optic with a 9.1-mm focal length and a 5-mm diameter. Using a larger diameter lens will not result in any greater system throughput because of the limited input NA of the optical fiber. The singlet lenses in this catalog that meet these criteria are LPX-5.0-5.2-C, which is plano-convex, and LDX-6.0-7.7-C and LDX-5.0-9.9-C, which are biconvex.

Making some simple calculations has reduced our choice of lenses to just three. The following chapter, *Gaussian Beam Optics*, discusses how to make a final choice of lenses based on various performance criteria.

1ch_FundamentalOptics_Final_a.qxd 6/15/2009 2:28 PM Page 1.7

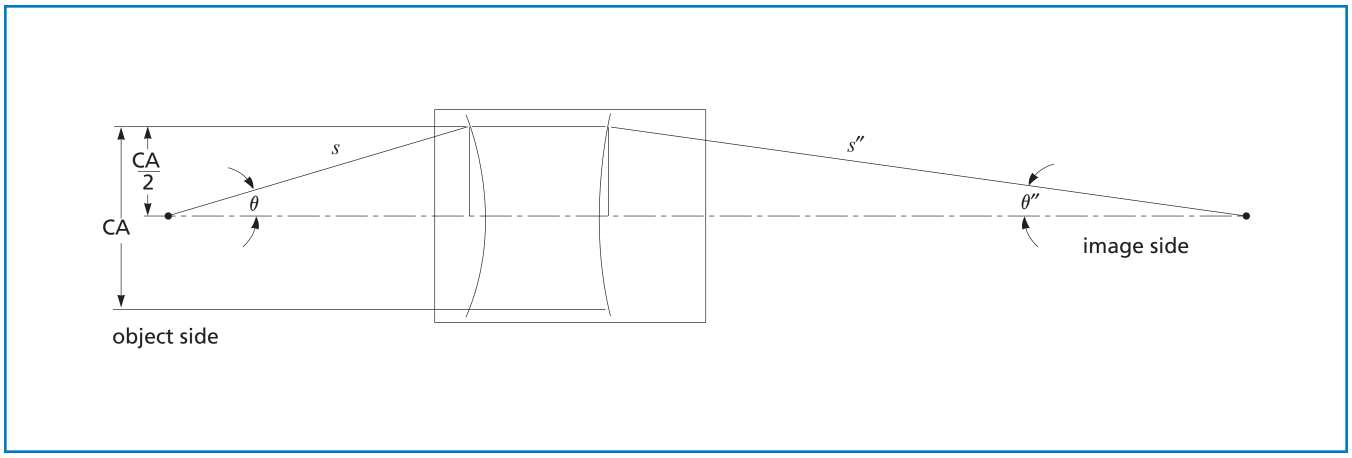
CA
2
CA
s
θ
s″
θ″
image side
object side

Figure 1.6 **Numerical aperture and magnification**

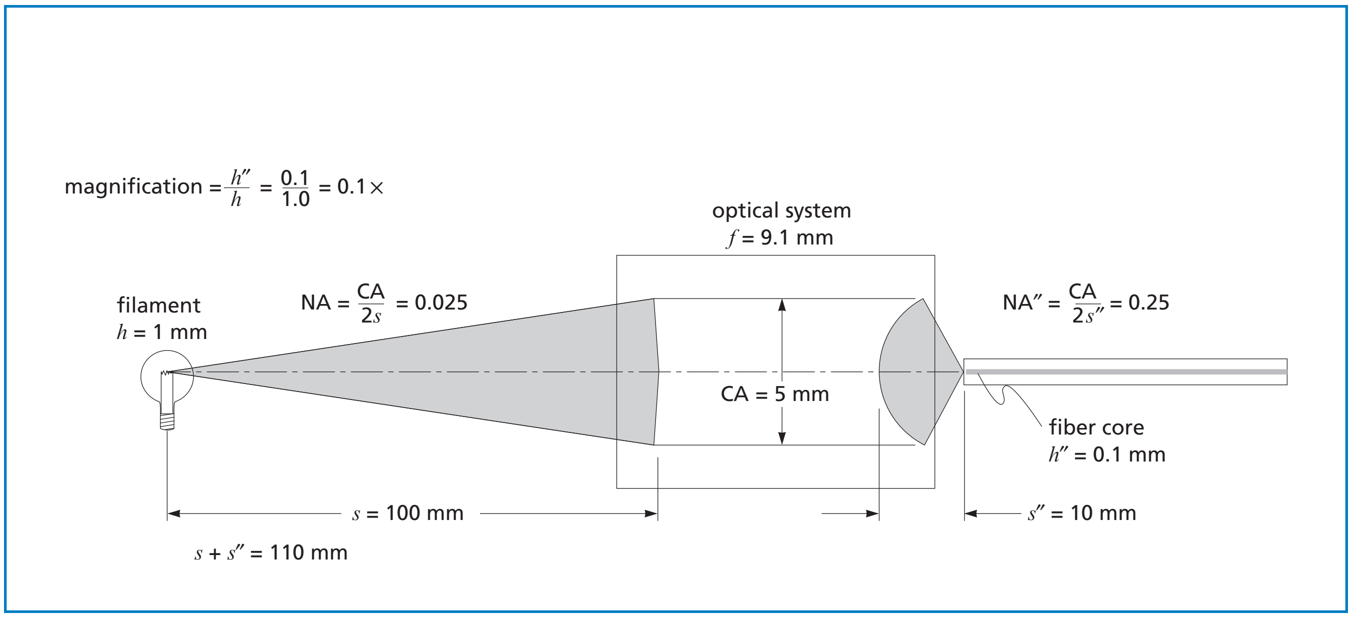
magnification = $\frac{h''}{h}$ = $\frac{0.1}{1.0}$ = 0.1×
optical system
f = 9.1 mm
filament
h = 1 mm
NA = $\frac{CA}{2s}$ = 0.025
NA″ = $\frac{CA}{2s''}$ = 0.25
CA = 5 mm
fiber core
h'' = 0.1 mm
s = 100 mm
s'' = 10 mm
$s + s''$ = 110 mm

Figure 1.7 **Optical system geometry for focusing the output of an incandescent bulb into an optical fiber**

Lens Combination Formulas

Many optical tasks require several lenses in order to achieve an acceptable level of performance. One possible approach to lens combinations is to consider each image formed by each lens as the object for the next lens and so on. This is a valid approach, but it is time consuming and unnecessary.

It is much simpler to calculate the effective (combined) focal length and principal-point locations and then use these results in any subsequent paraxial calculations (see figure 1.8). They can even be used in the optical invariant calculations described in the preceding section.

EFFECTIVE FOCAL LENGTH

The following formulas show how to calculate the effective focal length and principal-point locations for a combination of any two arbitrary components. The approach for more than two lenses is very simple: Calculate the values for the first two elements, then perform the same calculation for this combination with the next lens. This is continued until all lenses in the system are accounted for.

The expression for the combination focal length is the same whether lens separation distances are large or small and whether f_1 and f_2 are positive or negative:

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}. \qquad (1.16)$$

This may be more familiar in the form

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2}. \qquad (1.17)$$

Notice that the formula is symmetric with respect to the interchange of the lenses (end-for-end rotation of the combination) at constant *d*. The next two formulas are not.

COMBINATION FOCAL-POINT LOCATION

For all values of f_1, f_2, and d, the location of the focal point of the combined system (s_2''), measured from the secondary principal point of the second lens (H_2''), is given by

$$s_2'' = \frac{f_2(f_1 - d)}{f_1 + f_2 - d}. \qquad (1.18)$$

This can be shown by setting $s_1 = d - f_1$ (see figure 1.8a), and solving

$$\frac{1}{f_2} = \frac{1}{s_1} + \frac{1}{s_2''}$$

for s_2''.

Symbols

f_c = combination focal length (EFL), positive if combination final focal point falls to the right of the combination secondary principal point, negative otherwise (see figure 1.8c).

f_1 = focal length of the first element (see figure 1.8a).

f_2 = focal length of the second element.

d = distance from the secondary principal point of the first element to the primary principal point of the second element, positive if the primary principal point is to the right of the secondary principal point, negative otherwise (see figure 1.8b).

s_1'' = distance from the primary principal point of the first element to the final combination focal point (location of the final image for an object at infinity to the right of both lenses), positive if the focal point is to left of the first element's primary principal point (see figure 1.8d).

s_2'' = distance from the secondary principal point of the second element to the final combination focal point (location of the final image for an object at infinity to the left of both lenses), positive if the focal point is to the right of the second element's secondary principal point (see figure 1.8b).

z_H = distance to the combination primary principal point measured from the primary principal point of the first element, positive if the combination secondary principal point is to the right of secondary principal point of second element (see figure 1.8d).

z_H'' = distance to the combination secondary principal point measured from the secondary principal point of the second element, positive if the combination secondary principal point is to the right of the secondary principal point of the second element (see figure 1.8c).

Note: *These paraxial formulas apply to* ***coaxial*** *combinations of both thick and thin lenses immersed in air or any other fluid with refractive index independent of position. They assume that light propagates from left to right through an optical system.*

1ch_FundamentalOptics_Final_a.qxd 6/15/2009 2:28 PM Page 1.9

COMBINATION SECONDARY PRINCIPAL-POINT LOCATION

Because the thin-lens approximation is obviously highly invalid for most combinations, the ability to determine the location of the secondary principal point is vital for accurate determination of d when another element is added. The simplest formula for this calculates the distance from the secondary principal point of the final (second) element to the secondary principal point of the combination (see figure 1.8b):

$$z = s_2'' - f \,. \tag{1.19}$$

COMBINATION EXAMPLES

It is possible for a lens combination or system to exhibit principal planes that are far removed from the system. When such systems are themselves combined, negative values of d may occur. Probably the simplest example of a negative d-value situation is shown in figure 1.9. Meniscus lenses with steep surfaces have external principal planes. When two of these lenses are brought into contact, a negative value of d can occur. Other combined-lens examples are shown in figures 1.10 through 1.13.

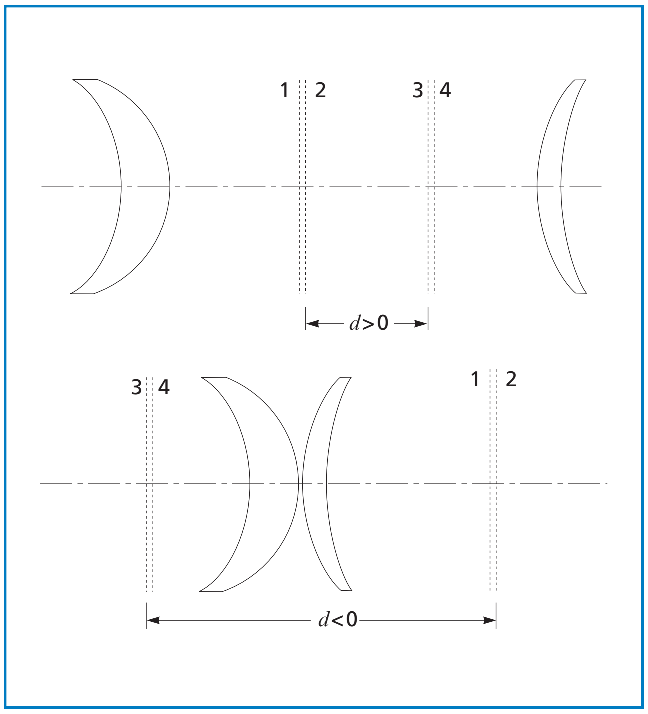

Figure 1.9 **"Extreme" meniscus-form lenses with external principal planes** (drawing not to scale)

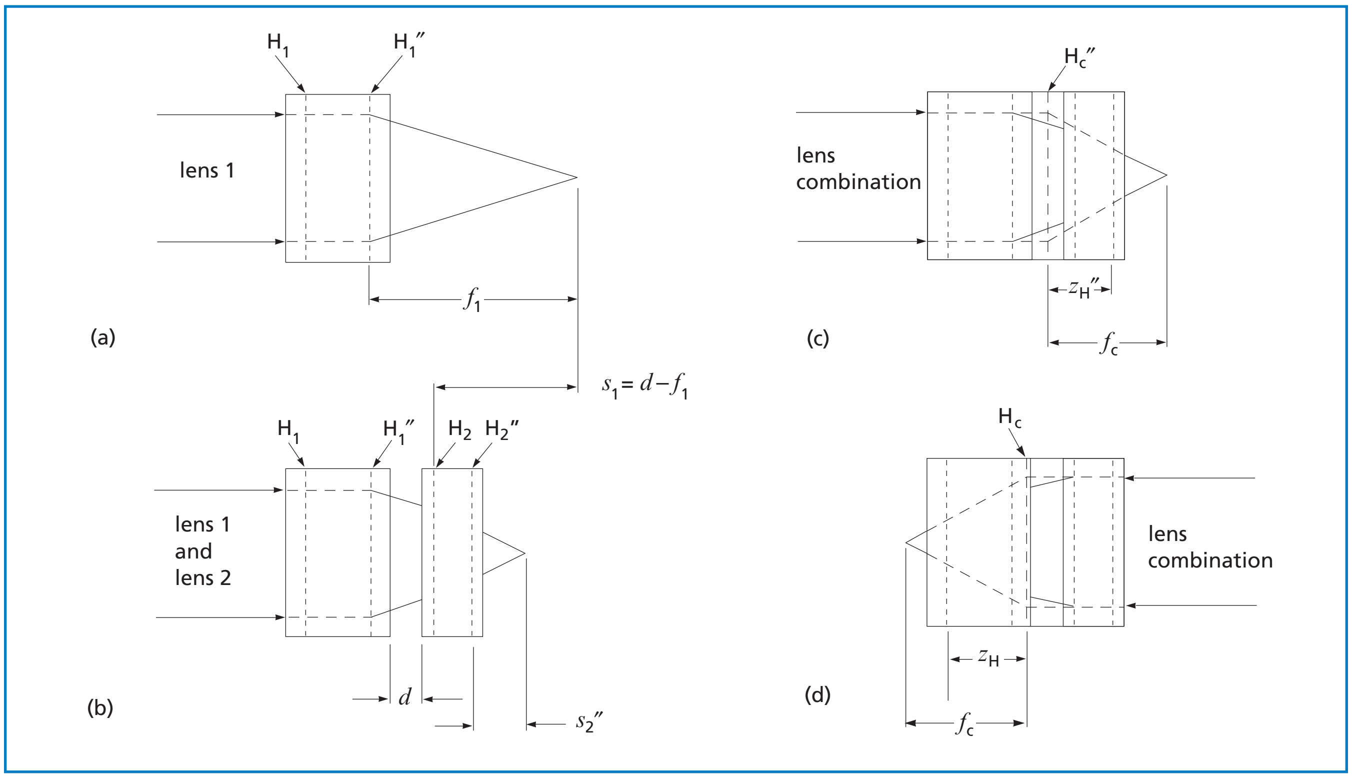

Figure 1.8 **Lens combination focal length and principal planes**

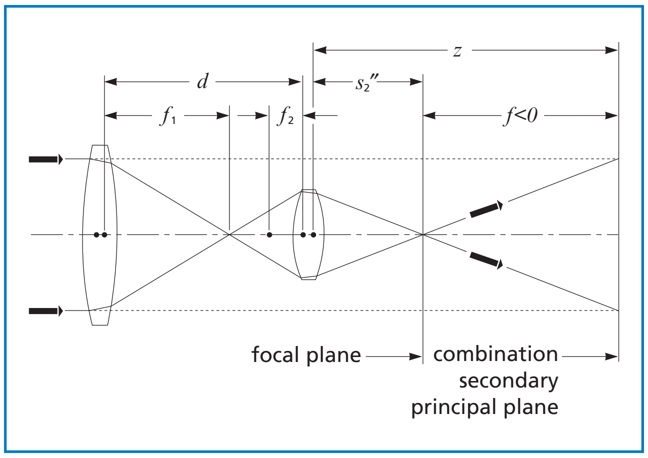

Figure 1.10 **Positive lenses separated by distance greater than $f_1 + f_2$:** f is negative and both s_2'' and z are positive. Lens symmetry is not required.

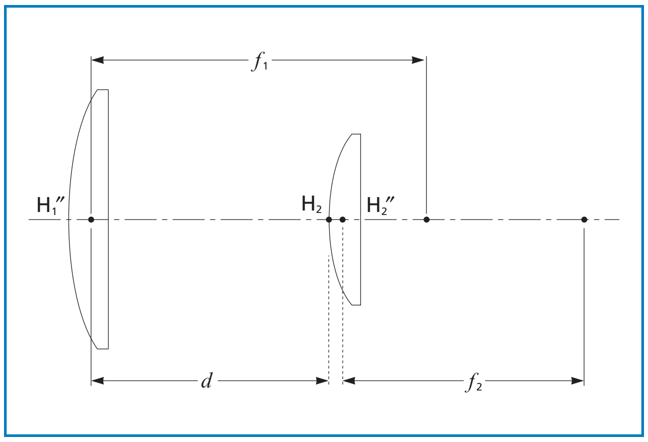

Figure 1.11 **Achromatic combinations:** Air-spaced lens combinations can be made nearly achromatic, even though both elements are made from the same material. Achieving achromatism requires that, in the thin-lens approximation,

$$d = \frac{(f_1 + f_2)}{2}.$$

This is the basis for Huygens and Ramsden eyepieces.

This approximation is adequate for most thick-lens situations. The signs of f_1, f_2, and d are unrestricted, but d must have a value that guarantees the existence of an air space. Element shapes are unrestricted and can be chosen to compensate for other aberrations.

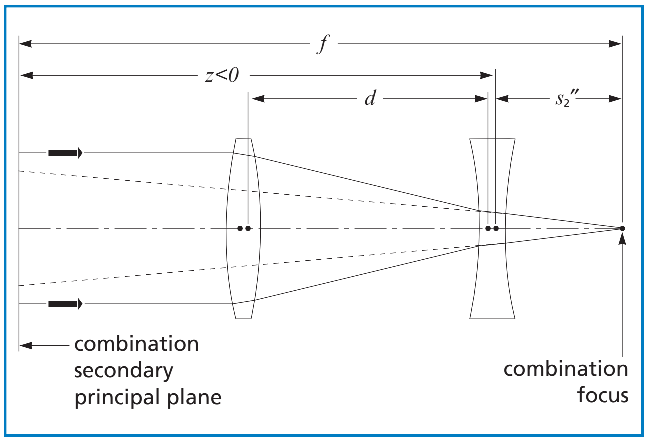

Figure 1.12 **Telephoto combination:** The most important characteristic of the telephoto lens is that the EFL, and hence the image size, can be made much larger than the distance from the first lens surface to the image would suggest by using a positive lens followed by a negative lens (but not necessarily the lens shapes shown in the figure). For example, f_1 is positive and $f_2 = -f_1/2$. Then f is negative for d less than $f_1/2$, infinite for $d = f_1/2$ (Galilean telescope or beam expander), and positive for d larger than $f_1/2$. To make the example even more specific, catalog lenses LDX-50.8-130.4-C and LDK-42.0-52.2-C, with $d = 78.2$ mm, will yield $s_2'' = 2.0$ m, $f = 5.2$ m, and $z = 43.2$ m.

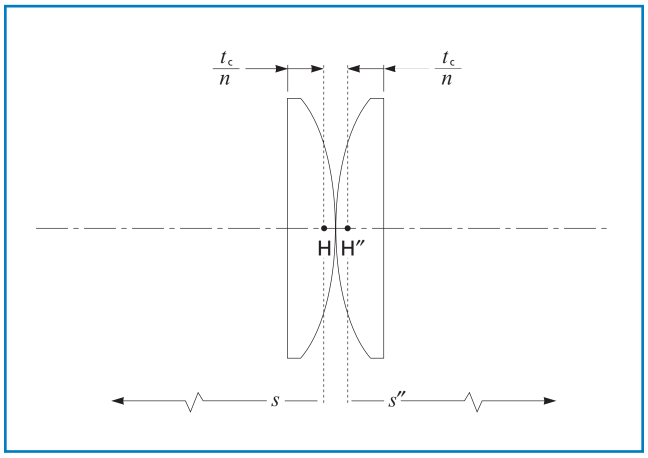

Figure 1.13 **Condenser configuration:** The convex vertices of a pair of identical plano-convex lenses are on contact. (The lenses could also be plano aspheres.) Because $d = 0$, $f = f_1/2 = f_2/2$, $f_1/2 = s_2''$, and $z = 0$. The secondary principal point of the second element and the secondary principal point of the combination coincide at H″, at depth t_c/n beneath the vertex of the plano surface of the second element, where t_c is the element center thickness and n is the refractive index of the element. By symmetry, the primary principal point of the combination is similarly located in the first element. Combination conjugate distances must be measured from these points.

Performance Factors

After paraxial formulas have been used to select values for component focal length(s) and diameter(s), the final step is to select actual lenses. As in any engineering problem, this selection process involves a number of tradeoffs, including performance, cost, weight, and environmental factors.

The performance of real optical systems is limited by several factors, including lens aberrations and light diffraction. The magnitude of these effects can be calculated with relative ease.

Numerous other factors, such as lens manufacturing tolerances and component alignment, impact the performance of an optical system. Although these are not considered explicitly in the following discussion, it should be kept in mind that if calculations indicate that a lens system only just meets the desired performance criteria, in practice it may fall short of this performance as a result of other factors. In critical applications, it is generally better to select a lens whose calculated performance is significantly better than needed.

DIFFRACTION

Diffraction, a natural property of light arising from its wave nature, poses a fundamental limitation on any optical system. Diffraction is always present, although its effects may be masked if the system has significant aberrations. When an optical system is essentially free from aberrations, its performance is limited solely by diffraction, and it is referred to as diffraction limited.

In calculating diffraction, we simply need to know the focal length(s) and aperture diameter(s); we do not consider other lens-related factors such as shape or index of refraction.

Since diffraction increases with increasing f-number, and aberrations decrease with increasing f-number, determining optimum system performance often involves finding a point where the combination of these factors has a minimum effect.

ABERRATIONS

To determine the precise performance of a lens system, we can trace the path of light rays through it, using Snell's law at each optical interface to determine the subsequent ray direction. This process, called ray tracing, is usually accomplished on a computer. When this process is completed, it is typically found that not all the rays pass through the points or positions predicted by paraxial theory. These deviations from ideal imaging are called lens aberrations.

The direction of a light ray after refraction at the interface between two homogeneous, isotropic media of differing index of refraction is given by Snell's law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{1.20}$$

where θ_1 is the angle of incidence, θ_2 is the angle of refraction, and both angles are measured from the surface normal as shown in figure 1.14.

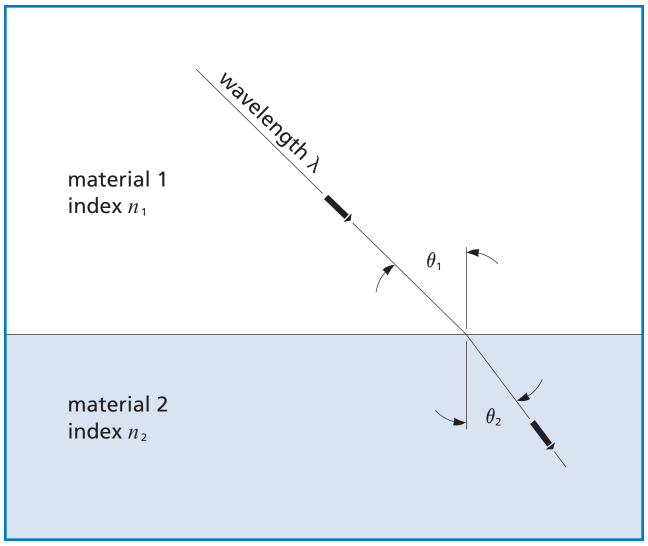

Figure 1.14 **Refraction of light at a dielectric boundary**

APPLICATION NOTE

Technical Assistance

Detailed performance analysis of an optical system is accomplished by using computerized ray-tracing software. CVI Melles Griot applications engineers are able to provide a ray-tracing analysis of simple catalog-component systems. If you need assistance in determining the performance of your optical system, or in selecting optimum components for your particular application, please contact your nearest CVI Melles Griot office.

Even though tools for the precise analysis of an optical system are becoming easier to use and are readily available, it is still quite useful to have a method for quickly estimating lens performance. This not only saves time in the initial stages of system specification, but can also help achieve a better starting point for any further computer optimization.

The first step in developing these rough guidelines is to realize that the sine functions in Snell's law can be expanded in an infinite Taylor series:

$$\sin\theta_1 = \theta_1 - \theta_1^3/3! + \theta_1^5/5! - \theta_1^7/7! + \theta_1^9/9! - \ldots \qquad (1.21)$$

The first approximation we can make is to replace all the sine functions with their arguments (i.e., replace $\sin\theta_1$ with θ_1 itself and so on). This is called first-order or paraxial theory because only the first terms of the sine expansions are used. Design of any optical system generally starts with this approximation using the paraxial formulas.

The assumption that $\sin\theta = \theta$ is reasonably valid for θ close to zero (i.e., high f-number lenses). With more highly curved surfaces (and particularly marginal rays), paraxial theory yields increasingly large deviations from real performance because $\sin\theta \neq \theta$. These deviations are known as aberrations. Because a perfect optical system (one without any aberrations) would form its image at the point and to the size indicated by paraxial theory, aberrations are really a measure of how the image differs from the paraxial prediction.

As already stated, exact ray tracing is the only rigorous way to analyze real lens surfaces. Before the advent of electronic computers, this was excessively tedious and time consuming. Seidel* addressed this issue by developing a method of calculating aberrations resulting from the $\theta_1{}^3/3!$ term. The resultant third-order lens aberrations are therefore called Seidel aberrations.

To simplify these calculations, Seidel put the aberrations of an optical system into several different classifications. In monochromatic light they are spherical aberration, astigmatism, field curvature, coma, and distortion. In polychromatic light there are also chromatic aberration and lateral color. Seidel developed methods to approximate each of these aberrations without actually tracing large numbers of rays using all the terms in the sine expansions.

In actual practice, aberrations occur in combinations rather than alone. This system of classifying them, which makes analysis much simpler, gives a good description of optical system image quality. In fact, even in the era of powerful ray-tracing software, Seidel's formula for spherical aberration is still widely used.

SPHERICAL ABERRATION

Figure 1.15 illustrates how an aberration-free lens focuses incoming collimated light. All rays pass through the focal point F″. The lower figure shows the situation more typically encountered in single lenses. The farther from the optical axis the ray enters the lens, the nearer to the lens it focuses (crosses the optical axis). The distance along the optical axis between the intercept of the rays that are nearly on the optical axis (paraxial rays) and the rays that go through the edge of the lens (marginal rays) is called longitudinal spherical aberration (LSA). The height at which these rays intercept the paraxial focal plane is called transverse spherical aberration (TSA). These quantities are related by

$$\text{TSA} = \text{LSA} \times \tan(u''). \qquad (1.22)$$

Spherical aberration is dependent on lens shape, orientation, and conjugate ratio, as well as on the index of refraction of the materials present. Parameters for choosing the best lens shape and orientation for a given task are presented later in this chapter. However, the third-

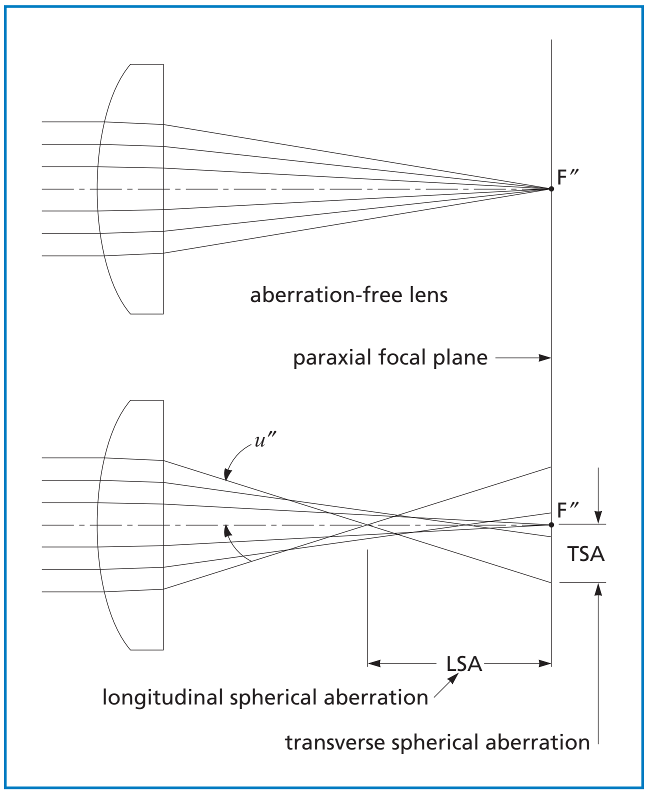
F″
aberration-free lens
paraxial focal plane
u″
F″
TSA
LSA
longitudinal spherical aberration
transverse spherical aberration

Figure 1.15 **Spherical aberration of a plano-convex lens**

order, monochromatic, spherical aberration of a plano-convex lens used at infinite conjugate ratio can be estimated by

---

* *Ludwig von Seidel, 1857.*

$$\text{spot size due to spherical aberration} = \frac{0.067\,f}{f/\#^3}. \qquad (1.23)$$

Theoretically, the simplest way to eliminate or reduce spherical aberration is to make the lens surface(s) with a varying radius of curvature (i.e., an aspheric surface) designed to exactly compensate for the fact that $\sin\theta \neq \theta$ at larger angles. In practice, however, most lenses with high surface accuracy are manufactured by grinding and polishing techniques that naturally produce spherical or cylindrical surfaces. The manufacture of aspheric surfaces is more complex, and it is difficult to produce a lens of sufficient surface accuracy to eliminate spherical aberration completely. Fortunately, these aberrations can be virtually eliminated, for a chosen set of conditions, by combining the effects of two or more spherical (or cylindrical) surfaces.

In general, simple positive lenses have undercorrected spherical aberration, and negative lenses usually have overcorrected spherical aberration. By combining a positive lens made from low-index glass with a negative lens made from high-index glass, it is possible to produce a combination in which the spherical aberrations cancel but the focusing powers do not. The simplest examples of this are cemented doublets, such as the LAO series which produce minimal spherical aberration when properly used.

ASTIGMATISM

When an off-axis object is focused by a spherical lens, the natural asymmetry leads to astigmatism. The system appears to have two different focal lengths.

As shown in figure 1.16, the plane containing both optical axis and object point is called the tangential plane. Rays that lie in this plane are called tangential, or meridional, rays. Rays not in this plane are referred to as skew rays. The chief, or principal, ray goes from the object point through the center of the aperture of the lens system. The plane perpendicular to the tangential plane that contains the principal ray is called the sagittal or radial plane.

The figure illustrates that tangential rays from the object come to a focus closer to the lens than do rays in the sagittal plane. When the image is evaluated at the tangential conjugate, we see a line in the sagittal direction. A line in the tangential direction is formed at the sagittal conjugate. Between these conjugates, the image is either an elliptical or a circular blur. Astigmatism is defined as the separation of these conjugates.

The amount of astigmatism in a lens depends on lens shape only when there is an aperture in the system that is not in contact with the lens itself. (In all optical systems there is an aperture or stop, although in many cases it is simply the clear aperture of the lens element itself.) Astigmatism strongly depends on the conjugate ratio.

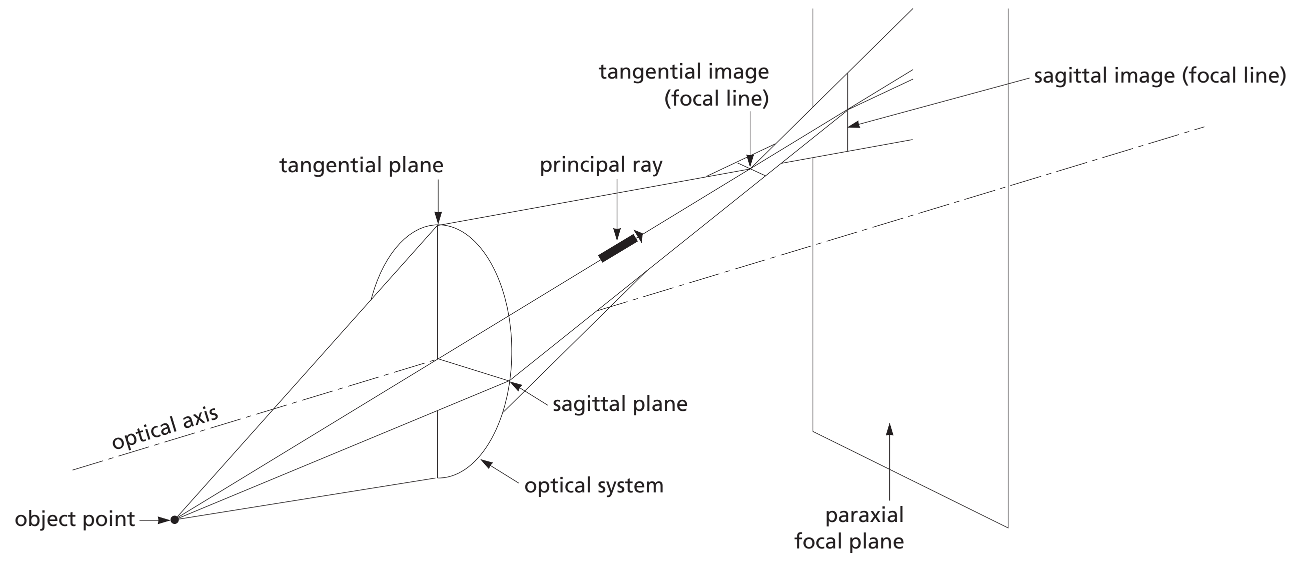

Figure 1.16 **Astigmatism represented by sectional views**

COMA

In spherical lenses, different parts of the lens surface exhibit different degrees of magnification. This gives rise to an aberration known as coma. As shown in figure 1.17, each concentric zone of a lens forms a ring-shaped image called a comatic circle. This causes blurring in the image plane (surface) of off-axis object points. An off-axis object point is not a sharp image point, but it appears as a characteristic comet-like flare. Even if spherical aberration is corrected and the lens brings all rays to a sharp focus on axis, a lens may still exhibit coma off axis. See figure 1.18.

As with spherical aberration, correction can be achieved by using multiple surfaces. Alternatively, a sharper image may be produced by judiciously placing an aperture, or stop, in an optical system to eliminate the more marginal rays.

FIELD CURVATURE

Even in the absence of astigmatism, there is a tendency of optical systems to image better on curved surfaces than on flat planes. This effect is called field curvature (see figure 1.19). In the presence of astigmatism, this problem is compounded because two separate astigmatic focal surfaces correspond to the tangential and sagittal conjugates.

Field curvature varies with the square of field angle or the square of image height. Therefore, by reducing the field angle by one-half, it is possible to reduce the blur from field curvature to a value of 0.25 of its original size.

Positive lens elements usually have inward curving fields, and negative lenses have outward curving fields. Field curvature can thus be corrected to some extent by combining positive and negative lens elements.

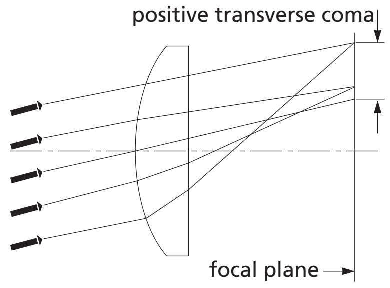

Figure 1.18 **Positive transverse coma**

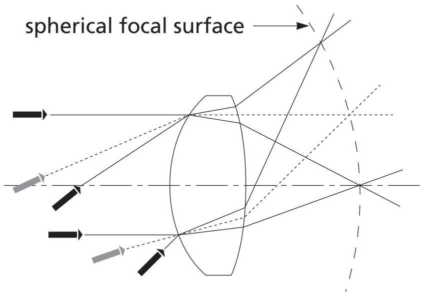

Figure 1.19 **Field curvature**

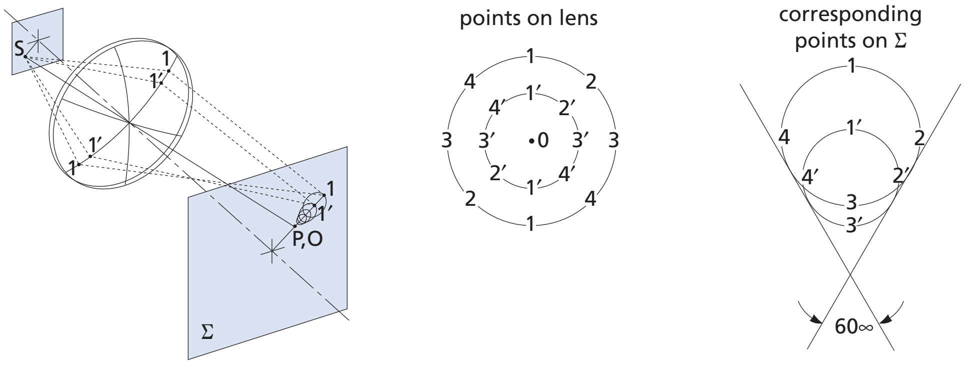

Figure 1.17 **Imaging an off-axis point source by a lens with positive transverse coma**

DISTORTION

The image field not only may have curvature but may also be distorted. The image of an off-axis point may be formed at a location on this surface other than that predicted by the simple paraxial equations. This distortion is different from coma (where rays from an off-axis point fail to meet perfectly in the image plane). Distortion means that even if a perfect off-axis point image is formed, its location on the image plane is not correct. Furthermore, the amount of distortion usually increases with increasing image height. The effect of this can be seen as two different kinds of distortion: pincushion and barrel (see figure 1.20). Distortion does not lower system resolution; it simply means that the image shape does not correspond exactly to the shape of the object. Distortion is a separation of the actual image point from the paraxially predicted location on the image plane and can be expressed either as an absolute value or as a percentage of the paraxial image height.

It should be apparent that a lens or lens system has opposite types of distortion depending on whether it is used forward or backward. This means that if a lens were used to make a photograph, and then used in reverse to project it, there would be no distortion in the final screen image. Also, perfectly symmetrical optical systems at 1:1 magnification have no distortion or coma.

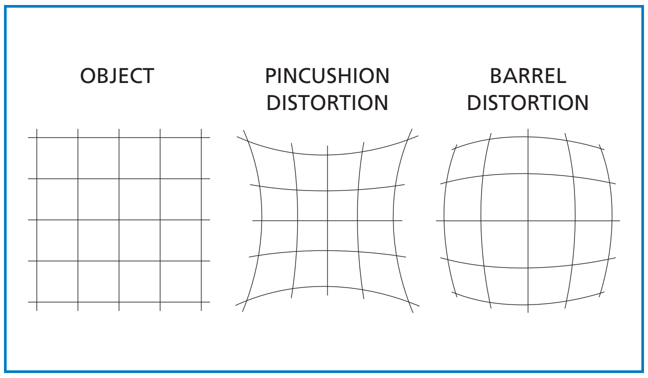

Figure 1.20 **Pincushion and barrel distortion**

CHROMATIC ABERRATION

The aberrations previously described are purely a function of the shape of the lens surfaces, and they can be observed with monochromatic light. Other aberrations, however, arise when these optics are used to transform light containing multiple wavelengths. The index of refraction of a material is a function of wavelength. Known as dispersion, this is discussed in *Material Properties*. From Snell's law (see equation 1.20), it can be seen that light rays of different wavelengths or colors will be refracted at different angles since the index is not a constant. Figure 1.21 shows the result when polychromatic collimated light is incident on a positive lens element. Because the index of refraction is higher for shorter wavelengths, these are focused closer to the lens than the longer wavelengths. Longitudinal chromatic aberration is defined as the axial distance from the nearest to the farthest focal point. As in the case of spherical aberration, positive and negative elements have opposite signs of chromatic aberration. Once again, by combining elements of nearly opposite aberration to form a doublet, chromatic aberration can be partially corrected. It is necessary to use two glasses with different dispersion characteristics, so that the weaker negative element can balance the aberration of the stronger, positive element.

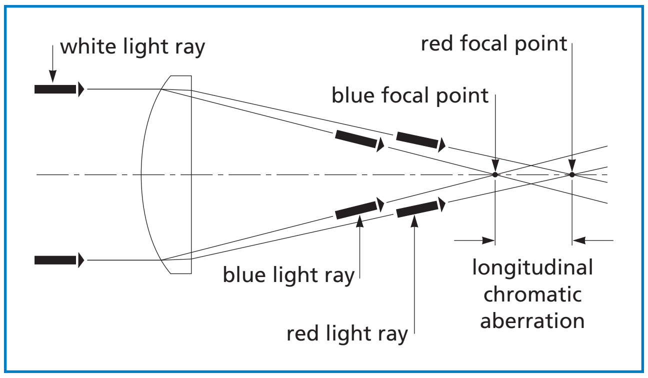

Figure 1.21 **Longitudinal chromatic aberration**

Variations of Aberrations with Aperture, Field Angle, and Image Height

Aberration	Aperture (ϕ)	Field Angle (θ)	Image Height (y)
Lateral Spherical	ϕ^3	—	—
Longitudinal Spherical	ϕ^2	—	—
Coma	ϕ^2	θ	y
Astigmatism	ϕ	θ^2	y^2
Field Curvature	ϕ	θ^2	y^2
Distortion	—	θ^3	y^3
Chromatic	—	—	—

LATERAL COLOR

Lateral color is the difference in image height between blue and red rays. Figure 1.22 shows the chief ray of an optical system consisting of a simple positive lens and a separate aperture. Because of the change in index with wavelength, blue light is refracted more strongly than red light, which is why rays intercept the image plane at different heights. Stated simply, magnification depends on color. Lateral color is very dependent on system stop location.

For many optical systems, the third-order term is all that may be needed to quantify aberrations. However, in highly corrected systems or in those having large apertures or a large angular field of view, third-order theory is inadequate. In these cases, exact ray tracing is absolutely essential.

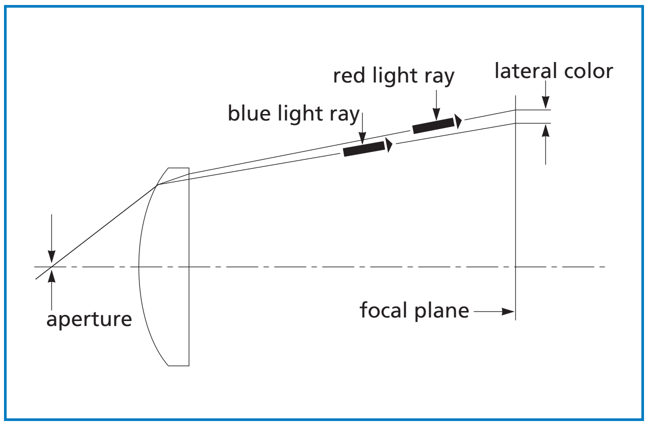

Figure 1.22 **Lateral Color**

APPLICATION NOTE

Achromatic Doublets Are Superior to Simple Lenses

Because achromatic doublets correct for spherical as well as chromatic aberration, they are often superior to simple lenses for focusing collimated light or collimating point sources, even in purely monochromatic light.

Although there is no simple formula that can be used to estimate the spot size of a doublet, the tables in *Spot Size* give sample values that can be used to estimate the performance of catalog achromatic doublets.

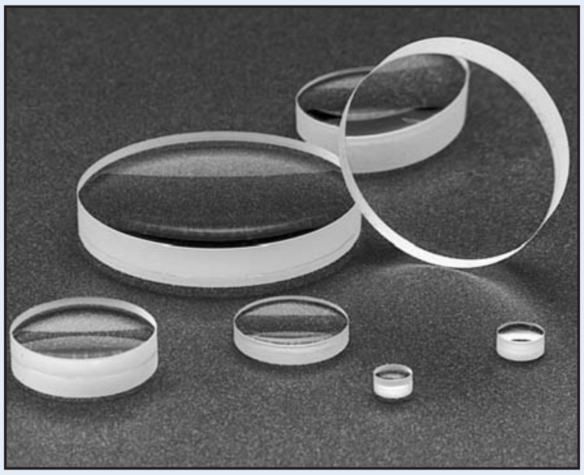

Lens Shape

Aberrations described in the preceding section are highly dependent on application, lens shape, and material of the lens (or, more exactly, its index of refraction). The singlet shape that minimizes spherical aberration at a given conjugate ratio is called best-form. The criterion for best-form at any conjugate ratio is that the marginal rays are equally refracted at each of the lens/air interfaces. This minimizes the effect of $\sin\theta \neq \theta$. It is also the criterion for minimum surface-reflectance loss. Another benefit is that absolute coma is nearly minimized for best-form shape, at both infinite and unit conjugate ratios.

To further explore the dependence of aberrations on lens shape, it is helpful to make use of the Coddington shape factor, q, defined as

$$q = \frac{(r_2 + r_1)}{(r_2 - r_1)}. \qquad (1.24)$$

Figure 1.23 shows the transverse and longitudinal spherical aberrations of a singlet lens as a function of the shape factor, q. In this particular instance, the lens has a focal length of 100 mm, operates at f/5, has an index of refraction of 1.518722 (BK7 at the mercury green line, 546.1 nm), and is being operated at the infinite conjugate ratio. It is also assumed that the lens itself is the aperture stop. An asymmetric shape that corresponds to a q-value of about 0.7426 for this material and wavelength is the best singlet shape for on-axis imaging. It is important to note that the best-form shape is dependent on refractive index. For example, with a high-index material, such as silicon, the best-form lens for the infinite conjugate ratio is a meniscus shape.

At infinite conjugate with a typical glass singlet, the plano-convex shape ($q = 1$), with convex side toward the infinite conjugate, performs nearly as well as the best-form lens. Because a plano-convex lens costs much less to manufacture than an asymmetric biconvex singlet, these lenses are quite popular. Furthermore, this lens shape exhibits near-minimum total transverse aberration and near-zero coma when used off axis, thus enhancing its utility.

For imaging at unit magnification ($s = s'' = 2f$), a similar analysis would show that a symmetric biconvex lens is the best shape. Not only is spherical aberration minimized, but coma, distortion, and lateral chromatic aberration exactly cancel each other out. These results are true regardless of material index or wavelength, which explains the utility of symmetric convex lenses, as well as symmetrical optical systems in general. However, if a remote stop is present, these aberrations may not cancel each other quite as well.

For wide-field applications, the best-form shape is definitely not the optimum singlet shape, especially at the infinite conjugate ratio, since it yields maximum field curvature. The ideal shape is determined by the situation and may require rigorous ray-tracing analysis. It is possible to achieve much better correction in an optical system by using more than one element. The cases of an infinite conjugate ratio system and a unit conjugate ratio system are discussed in the following section.

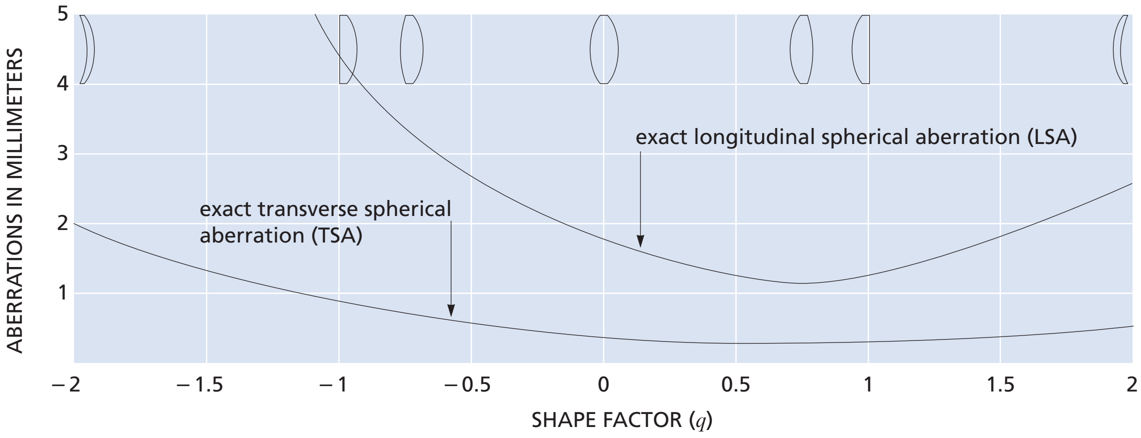

Figure 1.23 **Aberrations of positive singlets at infinite conjugate ratio as a function of shape**

Lens Combinations

INFINITE CONJUGATE RATIO

As shown in the previous discussion, the best-form singlet lens for use at infinite conjugate ratios is generally nearly plano-convex. Figure 1.24 shows a plano-convex lens (LPX-15.0-10.9-C) with incoming collimated light at a wavelength of 546.1 nm. This drawing, including the rays traced through it, is shown to exact scale. The marginal ray (ray f-number 1.5) strikes the paraxial focal plane significantly off the optical axis.

This situation can be improved by using a two-element system. The second part of the figure shows a precision achromat (LAO-21.0-14.0), which consists of a positive low-index (crown glass) element cemented to a negative meniscus high-index (flint glass) element. This is drawn to the same scale as the plano-convex lens. No spherical aberration can be discerned in the lens. Of course, not all of the rays pass exactly through the paraxial focal point; however, in this case, the departure is measured in micrometers, rather than in millimeters, as in the case of the plano-convex lens. Additionally, chromatic aberration (not shown) is much better corrected in the doublet. Even though these lenses are known as achromatic doublets, it is important to remember that even with monochromatic light the doublet's performance is superior.

Figure 1.24 also shows the f-number at which singlet performance becomes unacceptable. The ray with f-number 7.5 practically intercepts the paraxial focal point, and the f/3.8 ray is fairly close. This useful drawing, which can be scaled to fit a plano-convex lens of any focal length, can be used to estimate the magnitude of its spherical aberration, although lens thickness affects results slightly.

UNIT CONJUGATE RATIO

Figure 1.25 shows three possible systems for use at the unit conjugate ratio. All are shown to the same scale and using the same ray f-numbers with a light wavelength of 546.1 nm. The first system is a symmetric biconvex lens (LDX-21.0-19.2-C), the best-form singlet in this application. Clearly, significant spherical aberration is present in this lens at f/2.7. Not until f/13.3 does the ray closely approach the paraxial focus.

A dramatic improvement in performance is gained by using two identical plano-convex lenses with convex surfaces facing and nearly in contact. Those shown in figure 1.25 are both LPX-20.0-20.7-C. The combination of these two lenses yields almost exactly the same focal length as the biconvex lens. To understand why this configuration improves performance so dramatically, consider that if the biconvex lens were split down the middle, we would have two identical plano-convex lenses, each working at an infinite conjugate ratio, but with the convex surface toward the focus. This orientation is opposite to that shown to be optimum for this shape lens. On the other hand, if these lenses are reversed, we have the system just described but with a better correction of the spherical aberration.

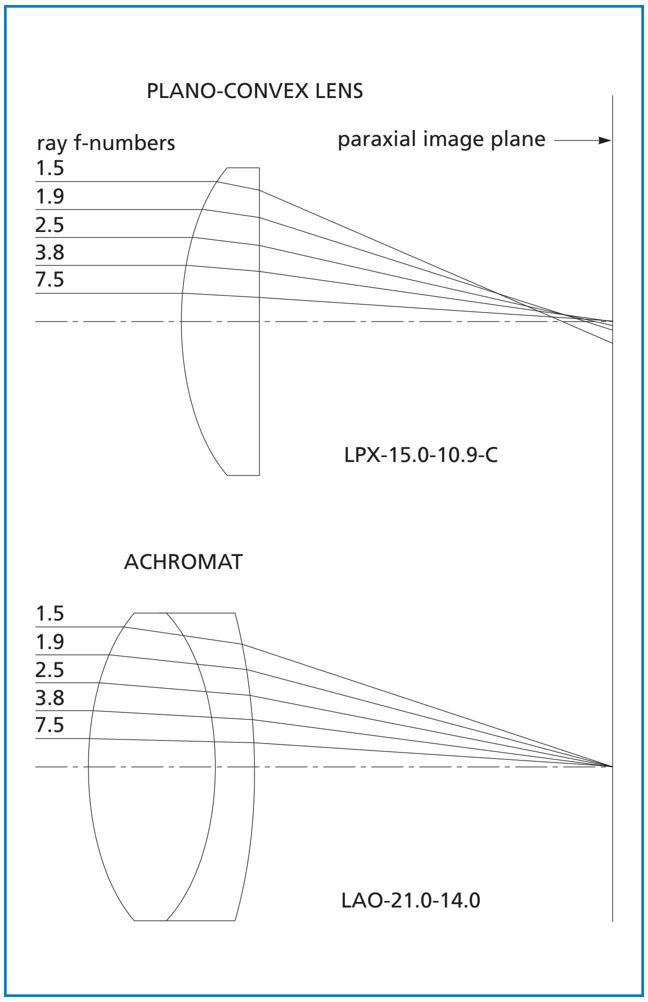

Figure 1.24 **Single-element plano-convex lens compared with a two-element achromat**

Previous examples indicate that an achromat is superior in performance to a singlet when used at the infinite conjugate ratio and at low f-numbers. Since the unit conjugate case can be thought of as two lenses, each working at the infinite conjugate ratio, the next step is to replace the plano-convex singlets with achromats, yielding a four-element system. The third part of figure 1.25 shows a system composed of two LAO-40.0-18.0 lenses. Once again, spherical aberration is not evident, even in the f/2.7 ray.

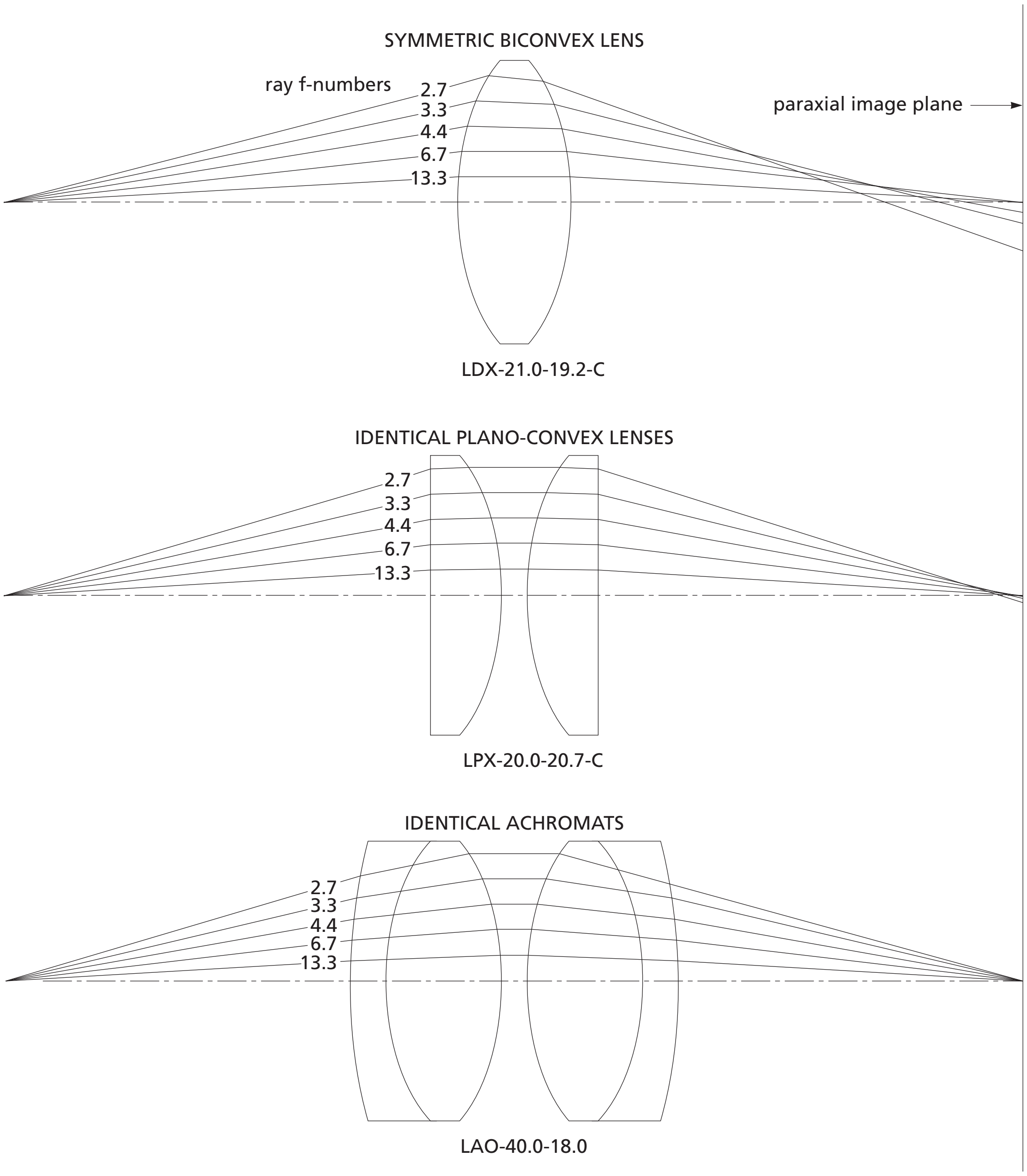
SYMMETRIC BICONVEX LENS
ray f-numbers
2.7
3.3
4.4
6.7
13.3
paraxial image plane
LDX-21.0-19.2-C
IDENTICAL PLANO-CONVEX LENSES
2.7
3.3
4.4
6.7
13.3
LPX-20.0-20.7-C
IDENTICAL ACHROMATS
2.7
3.3
4.4
6.7
13.3
LAO-40.0-18.0

Figure 1.25 **Three possible systems for use at the unit conjugate ratio**

Diffraction Effects

In all light beams, some energy is spread outside the region predicted by geometric propagation. This effect, known as diffraction, is a fundamental and inescapable physical phenomenon. Diffraction can be understood by considering the wave nature of light. Huygens' principle (figure 1.26) states that each point on a propagating wavefront is an emitter of secondary wavelets. The propagating wave is then the envelope of these expanding wavelets. Interference between the secondary wavelets gives rise to a fringe pattern that rapidly decreases in intensity with increasing angle from the initial direction of propagation. Huygens' principle nicely describes diffraction, but rigorous explanation demands a detailed study of wave theory.

Diffraction effects are traditionally classified into either Fresnel or Fraunhofer types. Fresnel diffraction is primarily concerned with what happens to light in the immediate neighborhood of a diffracting object or aperture. It is thus only of concern when the illumination source is close to this aperture or object, known as the near field. Consequently, Fresnel diffraction is rarely important in most classical optical setups, but it becomes very important in such applications as digital optics, fiber optics, and near-field microscopy.

Fraunhofer diffraction, however, is often important even in simple optical systems. This is the light-spreading effect of an aperture when the aperture (or object) is illuminated with an infinite source (plane-wave illumination) and the light is sensed at an infinite distance (far-field) from this aperture.

From these overly simple definitions, one might assume that Fraunhofer diffraction is important only in optical systems with infinite conjugate, whereas Fresnel diffraction equations should be considered at finite conjugate ratios. Not so. A lens or lens system of finite positive focal length with plane-wave input maps the far-field diffraction pattern of its aperture onto the focal plane; therefore, it is Fraunhofer diffraction that determines the limiting performance of optical systems. More generally, at any conjugate ratio, far-field angles are transformed into spatial displacements in the image plane.

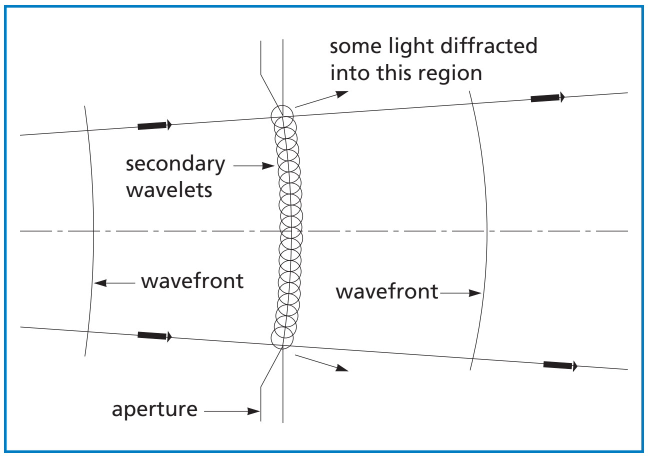

Figure 1.26 **Huygens' principle**

CIRCULAR APERTURE

Fraunhofer diffraction at a circular aperture dictates the fundamental limits of performance for circular lenses. It is important to remember that the spot size, caused by diffraction, of a circular lens is

$$d = 2.44\lambda(f/\#) \qquad (1.25)$$

where d is the diameter of the focused spot produced from plane-wave illumination and λ is the wavelength of light being focused. Notice that it is the f-number of the lens, not its absolute diameter, that determines this limiting spot size.

The diffraction pattern resulting from a uniformly illuminated circular aperture actually consists of a central bright region, known as the Airy disc (see figure 1.27), which is surrounded by a number of much fainter rings. Each ring is separated by a circle of zero intensity. The irradiance distribution in this pattern can be described by

$$I_x = I_0\left[\frac{2J_1(x)}{x}\right]^2 \qquad (1.26)$$

where

I_0 = peak irradiance in the image

$$J_1(x) = \text{Bessel function of the first kind of order unity}$$
$$= x\sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^{2n-2}}{(n-1)!n!2^{2n-1}}$$

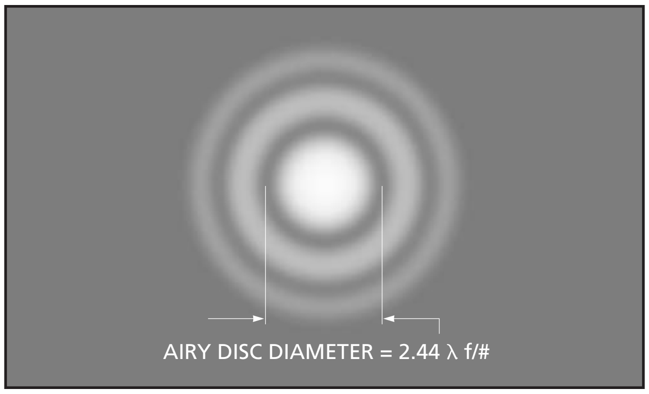

Figure 1.27 **Center of a typical diffraction pattern for a circular aperture**

and

$$x=\frac{\pi D}{\lambda}\sin\theta$$

where

λ = wavelength

D = aperture diameter

θ = angular radius from the pattern maximum.

This useful formula shows the far-field irradiance distribution from a uniformly illuminated circular aperture of diameter D.

SLIT APERTURE

A slit aperture, which is mathematically simpler, is useful in relation to cylindrical optical elements. The irradiance distribution in the diffraction pattern of a uniformly illuminated slit aperture is described by

$$I_x=I_0\left[\frac{\sin x}{x}\right]^2 \qquad (1.27)$$

where I_0 = peak irradiance in image

$$x=\frac{\pi w\sin\theta}{\lambda}$$

where λ = wavelength

w = slit width

θ = angular deviation from pattern maximum.

APPLICATION NOTE

Rayleigh Criterion

In imaging applications, spatial resolution is ultimately limited by diffraction. Calculating the maximum possible spatial resolution of an optical system requires an arbitrary definition of what is meant by resolving two features. In the Rayleigh criterion, it is assumed that two separate point sources can be resolved when the center of the Airy disc from one overlaps the first dark ring in the diffraction pattern of the second. In this case, the smallest resolvable distance, d, is

$$d=\frac{0.61\lambda}{\text{NA}}=1.22\lambda(\text{f}/\#)$$

ENERGY DISTRIBUTION TABLE

The accompanying table shows the major features of pure (unaberrated) Fraunhofer diffraction patterns of circular and slit apertures. The table shows the position, relative intensity, and percentage of total pattern energy corresponding to each ring or band. It is especially convenient to characterize positions in either pattern with the same variable x. This variable is related to field angle in the circular aperture case by

$$\sin\theta=\frac{\lambda x}{\pi D} \qquad (1.28)$$

where D is the aperture diameter. For a slit aperture, this relationship is given by

$$\sin\theta=\frac{\lambda x}{\pi w} \qquad (1.29)$$

where w is the slit width, π has its usual meaning, and D, w, and λ are all in the same units (preferably millimeters). Linear instead of angular field positions are simply found from $r=s''\tan\theta$ where s'' is the secondary conjugate distance. This last result is often seen in a different form, namely the diffraction-limited spot-size equation, which, for a circular lens is

$$d=2.44\,\lambda(\text{f}/\#) \qquad \text{(see eq. 1.25)}$$

This value represents the smallest spot size that can be achieved by an optical system with a circular aperture of a given f-number, and it is the diameter of the first dark ring, where the intensity has dropped to zero.

The graph in figure 1.28 shows the form of both circular and slit aperture diffraction patterns when plotted on the same normalized scale. Aperture diameter is equal to slit width so that patterns between x-values and angular deviations in the far-field are the same.

GAUSSIAN BEAMS

Apodization, or nonuniformity of aperture irradiance, alters diffraction patterns. If pupil irradiance is nonuniform, the formulas and results given previously do not apply. This is important to remember because most laser-based optical systems do not have uniform pupil irradiance. The output beam of a laser operating in the TEM_{00} mode has a smooth Gaussian irradiance profile. Formulas used to determine the focused spot size from such a beam are discussed in *Gaussian Beam Optics*. Furthermore, when dealing with Gaussian beams, the location of the focused spot also departs from that predicted by the paraxial equations given in this chapter. This is also detailed in *Gaussian Beam Optics*.

Energy Distribution in the Diffraction Pattern of a Circular or Slit Aperture

	Circular Aperture			Slit Aperture		
Ring or Band	Position (x)	Relative Intensity (I_x/I_0)	Energy in Ring (%)	Position (x)	Relative Intensity (I_x/I_0)	Energy in Band (%)
Central Maximum	0.0	1.0	83.8	0.0	1.0	90.3
First Dark	1.22π	0.0		1.00π	0.0	
First Bright	1.64π	0.0175	7.2	1.43π	0.0472	4.7
Second Dark	2.23π	0.0		2.00π	0.0	
Second Bright	2.68π	0.0042	2.8	2.46π	0.0165	1.7
Third Dark	3.24π	0.0		3.00π	0.0	
Third Bright	3.70π	0.0016	1.5	3.47π	0.0083	0.8
Fourth Dark	4.24π	0.0		4.00π	0.0	
Fourth Bright	4.71π	0.0008	1.0	4.48π	0.0050	0.5
Fifth Dark	5.24π	0.0		5.00π	0.0	

Note: Position variable (x) is defined in the text.

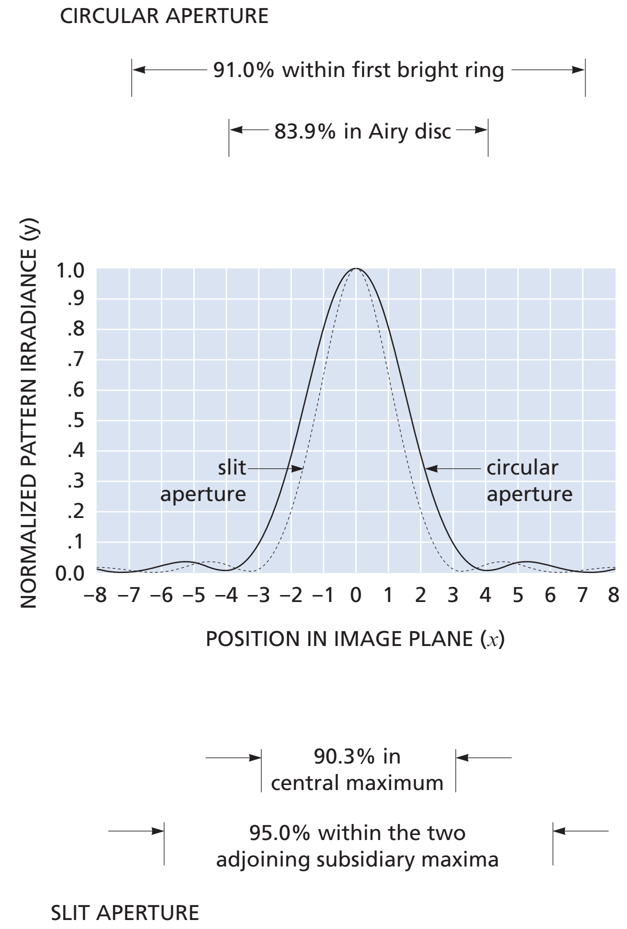

Figure 1.28 **Fraunhofer diffraction pattern of a singlet slit superimposed on the Fraunhofer diffraction pattern of a circular aperture**

Lens Selection

Having discussed the most important factors that affect the performance of a lens or a lens system, we will now address the practical matter of selecting the optimum catalog components for a particular task. The following useful relationships are important to keep in mind throughout the selection process:

- Diffraction-limited spot size = 2.44 λ f/#
- Approximate on-axis spot size of a plano-convex lens at the infinite conjugate resulting from spherical aberration = $\frac{0.067\,f}{\mathrm{f}/\#^3}$
- Optical invariant = $m = \frac{\mathrm{NA}}{\mathrm{NA}''}$.

Example 1: Collimating an Incandescent Source

Produce a collimated beam from a quartz halogen bulb having a 1-mm-square filament. Collect the maximum amount of light possible and produce a beam with the lowest possible divergence angle.

This problem, illustrated in figure 1.29, involves the typical tradeoff between light-collection efficiency and resolution (where a beam is being collimated rather than focused, resolution is defined by beam divergence). To collect more light, it is necessary to work at a low f-number, but because of aberrations, higher resolution (lower divergence angle) will be achieved by working at a higher f-number.

In terms of resolution, the first thing to realize is that the minimum divergence angle (in radians) that can be achieved using any lens system is the source size divided by system focal length. An off-axis ray (from the edge of the source) entering the first principal point of the system exits the second principal point at the same angle. Therefore, increasing the system focal length improves this limiting divergence because the source appears smaller.

An optic that can produce a spot size of 1 mm when focusing a perfectly collimated beam is therefore required. Since source size is inherently limited, it is pointless to strive for better resolution. This level of resolution can be achieved easily with a plano-convex lens.

While angular divergence decreases with increasing focal length, spherical aberration of a plano-convex lens increases with increasing focal length. To determine the appropriate focal length, set the spherical aberration formula for a plano-convex lens equal to the source (spot) size:

$$\frac{0.067\,f}{\mathrm{f}/\#^3} = 1\ \mathrm{mm}.$$

This ensures a lens that meets the minimum performance needed. To select a focal length, make an arbitrary f-number choice. As can be seen from the relationship, as we lower the f-number (increase collection efficiency), we decrease the focal length, which will worsen the resultant divergence angle (minimum divergence = 1 mm/f).

In this example, we will accept f/2 collection efficiency, which gives us a focal length of about 120 mm. For f/2 operation we would need a minimum diameter of 60 mm. The LPX-60.0-62.2-C fits this specification exactly. Beam divergence would be about 8 mrad.

Finally, we need to verify that we are not operating below the theoretical diffraction limit. In this example, the numbers (1-mm spot size) indicate that we are not, since

$$\text{diffraction-limited spot size} = 2.44 \times 0.5\ \mu\mathrm{m} \times 2 = 2.44\ \mu\mathrm{m}.$$

Example 2: Coupling an Incandescent Source into a Fiber

In *Imaging Properties of Lens Systems* we considered a system in which the output of an incandescent bulb with a filament of 1 mm in diameter was to be coupled into an optical fiber with a core diameter of 100 µm and a numerical aperture of 0.25. From the optical invariant and other constraints given in the problem, we determined that $f = 9.1$ mm, CA = 5 mm, $s = 100$ mm, $s'' = 10$ mm, NA″ = 0.25, and NA = 0.025 (or f/2 and f/20). The singlet lenses that match these specifications are the plano-convex LPX-5.0-5.2-C or biconvex lenses LDX-6.0-7.7-C and LDX-5.0-9.9-C. The closest achromat would be the LAO-10.0-6.0.

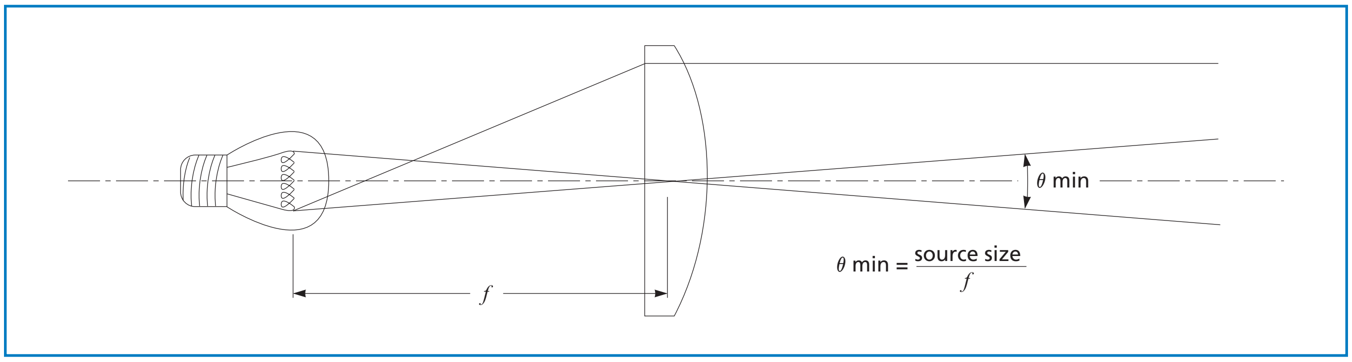

Figure 1.29 **Collimating an incandescent source**

We can immediately reject the biconvex lenses because of spherical aberration. We can estimate the performance of the LPX-5.0-5.2-C on the focusing side by using our spherical aberration formula:

$$\text{spot size} = \frac{0.067\ (10)}{2^3} = 84\ \mu\text{m}.$$

We will ignore, for the moment, that we are not working at the infinite conjugate.

This is slightly smaller than the 100-μm spot size we are trying to achieve. However, since we are not working at infinite conjugate, the spot size will be larger than that given by our simple calculation. This lens is therefore likely to be marginal in this situation, especially if we consider chromatic aberration. A better choice is the achromat. Although a computer ray trace would be required to determine its exact performance, it is virtually certain to provide adequate performance.

Example 3: Symmetric Fiber-to-Fiber Coupling

Couple an optical fiber with an 8-μm core and a 0.15 numerical aperture into another fiber with the same characteristics. Assume a wavelength of 0.5 μm.

This problem, illustrated in figure 1.30, is essentially a 1:1 imaging situation. We want to collect and focus at a numerical aperture of 0.15 or f/3.3, and we need a lens with an 8-μm spot size at this f-number. Based on the lens combination discussion in *Lens Combination Formulas*, our most likely setup is either a pair of identical plano-convex lenses or achromats, faced front to front. To determine the necessary focal length for a plano-convex lens, we again use the spherical aberration estimate formula:

$$\frac{0.067\ f}{3.3^3} = 0.008\ \text{mm}.$$

This formula yields a focal length of 4.3 mm and a minimum diameter of 1.3 mm. The LPX-4.2-2.3-BAK1 meets these criteria. The biggest problem with utilizing these tiny, short focal length lenses is the practical considerations of handling, mounting, and positioning them. Because using a pair of longer focal length singlets would result in unacceptable performance, the next step might be to use a pair of the slightly longer focal length, larger achromats, such as the LAO-10.0-6.0. The performance data, given in *Spot Size*, show that this combination does provide the required 8-μm spot diameter.

Because fairly small spot sizes are being considered here, it is important to make sure that the system is not being asked to work below the diffraction limit:

$$2.44 \times 0.5\ \mu\text{m} \times 3.3 = 4\ \ \mu\text{m}\,.$$

Since this is half the spot size caused by aberrations, it can be safely assumed that diffraction will not play a significant role here.

An entirely different approach to a fiber-coupling task such as this would be to use a pair of spherical ball lenses (LMS-LSFN series) or one of the gradient-index lenses (LGT series).

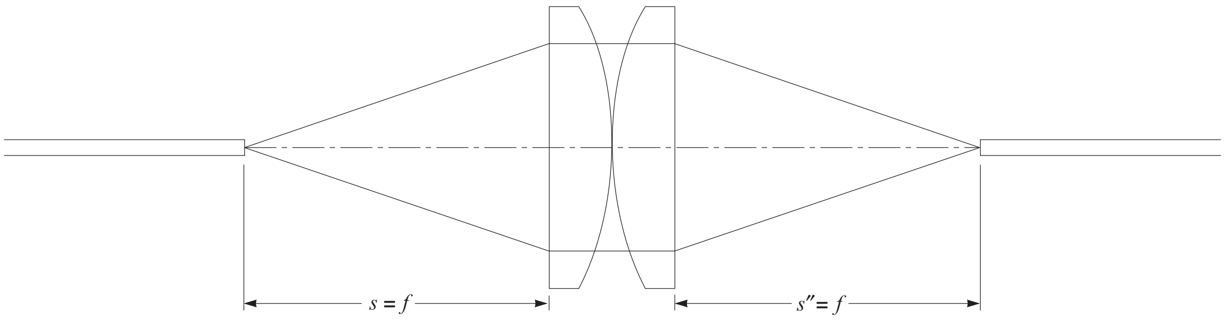

Figure 1.30 **Symmetric fiber-to-fiber coupling**

Example 4: Diffraction-Limited Performance

Determine at what f-number a plano-convex lens being used at an infinite conjugate ratio with 0.5-μm wavelength light becomes diffraction limited (i.e., the effects of diffraction exceed those caused by aberration).

To solve this problem, set the equations for diffraction-limited spot size and third-order spherical aberration equal to each other. The result depends upon focal length, since aberrations scale with focal length, while diffraction is solely dependent upon f-number. By substituting some common focal lengths into this formula, we get f/8.6 at $f = 100$ mm, f/7.2 at $f = 50$ mm, and f/4.8 at $f = 10$ mm.

$$2.44 \times 0.5\ \mu\text{m} \times \text{f}/\# = \frac{0.067 \times f}{\text{f}/\#^3}$$

or

$$\text{f}/\# = (54.9 \times f)^{1/4}.$$

When working with these focal lengths (and under the conditions previously stated), we can assume essentially diffraction-limited performance above these f-numbers. Keep in mind, however, that this treatment does not take into account manufacturing tolerances or chromatic aberration, which will be present in polychromatic applications.

APPLICATION NOTE

Spherical Ball Lenses for Fiber Coupling

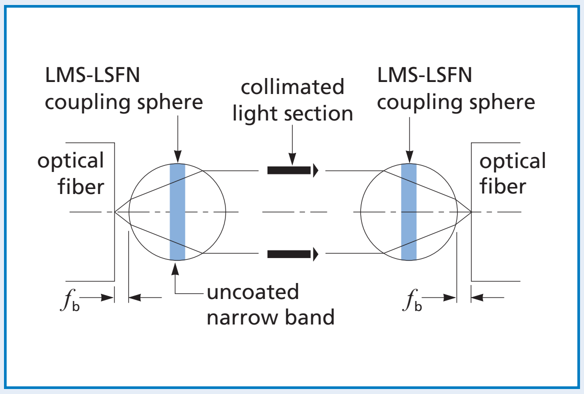

Spheres are arranged so that the fiber end is located at the focal point. The output from the first sphere is then collimated. If two spheres are aligned axially to each other, the beam will be transferred from one focal point to the other. Translational alignment sensitivity can be reduced by enlarging the beam. Slight negative defocusing of the ball can reduce the spherical aberration third-order contribution common to all coupling systems. Additional information can be found in "Lens Coupling in Fiber Optic Devices: Efficiency Limits," by A. Nicia, *Applied Optics*, vol. 20, no. 18, pp 3136–45, 1981. Off-axis aberrations are absent since the fiber diameters are so much smaller than the coupler focal length.

Spot Size

In general, the performance of a lens or lens system in a specific circumstance should be determined by an exact trigonometric ray trace. CVI Melles Griot applications engineers can supply ray-tracing data for particular lenses and systems of catalog components on request. In certain situations, however, some simple guidelines can be used for lens selection. The optimum working conditions for some of the lenses in this catalog have already been presented. The following tables give some quantitative results for a variety of simple and compound lens systems, which can be constructed from standard catalog optics.

In interpreting these tables, remember that these theoretical values obtained from computer ray tracing consider only the effects of ideal geometric optics. Effects of manufacturing tolerances have not been considered. Furthermore, remember that using more than one element provides a higher degree of correction but makes alignment more difficult. When actually choosing a lens or a lens system, it is important to note the tolerances and specifications clearly described for each CVI Melles Griot lens in the product listings.

The tables give the diameter of the spot for a variety of lenses used at several different f-numbers. All the tables are for on-axis, uniformly illuminated, collimated input light at 546.1 nm. They assume that the lens is facing in the direction that produces a minimum spot size. When the spot size caused by aberrations is smaller or equal to the diffraction-limited spot size, the notation "DL" appears next to the entry. The shorter focal length lenses produce smaller spot sizes because aberrations increase linearly as a lens is scaled up.

The effect on spot size caused by spherical aberration is strongly dependent on f-number. For a plano-convex singlet, spherical aberration is inversely dependent on the cube of the f-number. For doublets, this relationship can be even higher. On the other hand, the spot size caused by diffraction increases linearly with f-number. Thus, for some lens types, spot size at first decreases and then increases with f-number, meaning that there is some optimum performance point at which both aberrations and diffraction combine to form a minimum.

Unfortunately, these results cannot be generalized to situations in which the lenses are used off axis. This is particularly true of the achromat/aplanatic meniscus lens combinations because their performance degrades rapidly off axis.

Focal Length = 10 mm

	Spot Size (μm)*		
f/#	LDX-5.0-9.9-C	LPX-8.0-5.2-C	LAO-10.0-6.0
f/2	—	94	11
f/3	36	25	7
f/5	8	6.7 (DL)	6.7 (DL)
f/10	13.3 (DL)	13.3 (DL)	13.3 (DL)

**Diffraction-limited performance is indicated by DL.*

Focal Length = 30 mm

	Spot Size (μm)*		
f/#	LPX-18.5-15.6-C	LAO-30.0-12.5	LAO-50.0-18.0 & MENP-18.0-4.0-73.5-NSF8
f/2	295	—	3
f/3	79	7	4 (DL)
f/5	17	6.7 (DL)	6.9 (DL)
f/10	13.3 (DL)	13.3 (DL)	13.8 (DL)

**Diffraction-limited performance is indicated by DL.*

Focal Length = 60 mm

	Spot Size (μm)*			
f/#	LDX-50.0-60.0-C	LPX-30.0-31.1-C	LAO-60.0-30.0	LAO-100.0-31.5 & MENP-31.5-6.0-146.4-NSF8
f/2	816	600	—	—
f/3	217	160	34	4 (DL)
f/5	45	33	10	6.7 (DL)
f/10	13.3 (DL)	13.3 (DL)	13.3 (DL)	13.3 (DL)

**Diffraction-limited performance is indicated by DL.*

Fundamental Optics
Gaussian Beam Optics
Optical Specifications
Material Properties
Optical Coatings

Aberration Balancing

To improve system performance, optical designers make sure that the total aberration contribution from all surfaces taken together sums to nearly zero. Normally, such a process requires computerized analysis and optimization. However, there are some simple guidelines that can be used to achieve this with lenses available in this catalog. This approach can yield systems that operate at a much lower f-number than can usually be achieved with simple lenses.

Specifically, we will examine how to null the spherical aberration from two or more lenses in collimated, monochromatic light. This technique will thus be most useful for laser beam focusing and expanding.

Figure 1.31 shows the third-order longitudinal spherical aberration coefficients for six of the most common positive and negative lens shapes when used with parallel, monochromatic incident light. The plano-convex and plano-concave lenses both show minimum spherical aberration when oriented with their curved surface facing the incident parallel beam. All other configurations exhibit larger amounts of spherical aberration. With these lens types, it is now possible to show how various systems can be corrected for spherical aberration.

A two-element laser beam expander is a good starting example. In this case, two lenses are separated by a distance that is the sum of their focal lengths, so that the overall system focal length is infinite. This system will not focus incoming collimated light, but it will change the beam diameter. By definition, each of the lenses is operating at the same f-number.

The equation for longitudinal spherical aberration shows that, for two lenses with the same f-number, aberration varies directly with the focal lengths of the lenses. The sign of the aberration is the same as focal length. Thus, it should be possible to correct the spherical aberration of this Galilean-type beam expander, which consists of a positive focal length objective and a negative diverging lens.

If a plano-convex lens of focal length f_1 oriented in the normal direction is combined with a plano-concave lens of focal length f_2 oriented in its reverse direction, the total spherical aberration of the system is

$$\mathrm{LSA} = \frac{0.272\,f_1}{\mathrm{f}/\#^2} + \frac{1.069\,f_2}{\mathrm{f}/\#^2}.$$

After setting this equation to zero, we obtain

$$\frac{f_1}{f_2} = -\frac{1.069}{0.272} = -3.93.$$

To make the magnitude of aberration contributions of the two elements equal so they will cancel out, and thus correct the system, select the focal length of the positive element to be 3.93 times that of the negative element.

Figure 1.32(a) shows a beam-expander system made up of catalog elements, in which the focal length ratio is 4:1. This simple system is corrected to about 1/6 wavelength at 632.8 nm, even though the objective is operating at f/4 with a 20-mm aperture diameter. This is remarkably good wavefront correction for such a simple system; one would normally assume that a doublet objective would be needed and a complex diverging lens as well. This analysis does not take into account manufacturing tolerances.

A beam expander of lower magnification can also be derived from this information. If a symmetric-convex objective is used together] with a reversed plano-concave diverging lens, the aberration coefficients are in the ratio of 1.069/0.403 = 2.65. Figure 1.32(b) shows a system of catalog lenses that provides a magnification of 2.7 (the closest possible given the available focal lengths). The maximum wavefront error in this case is only a quarter-wave, even though the objective is working at f/3.3.

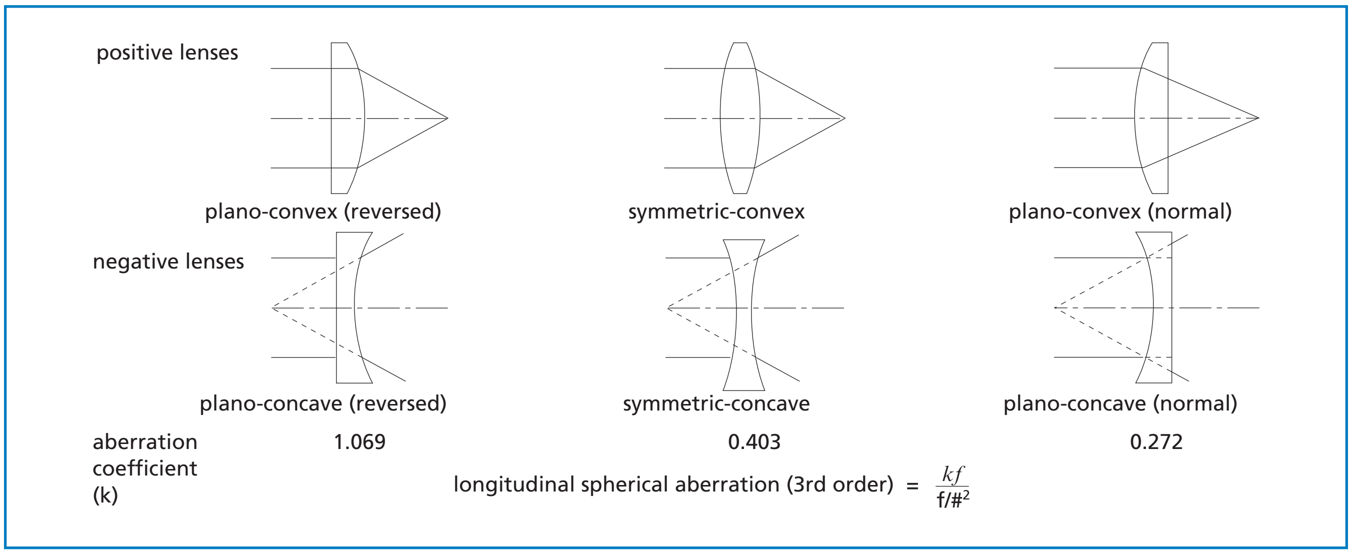

Figure 1.31 **Third-order longitudinal spherical aberration of typical lens shapes**

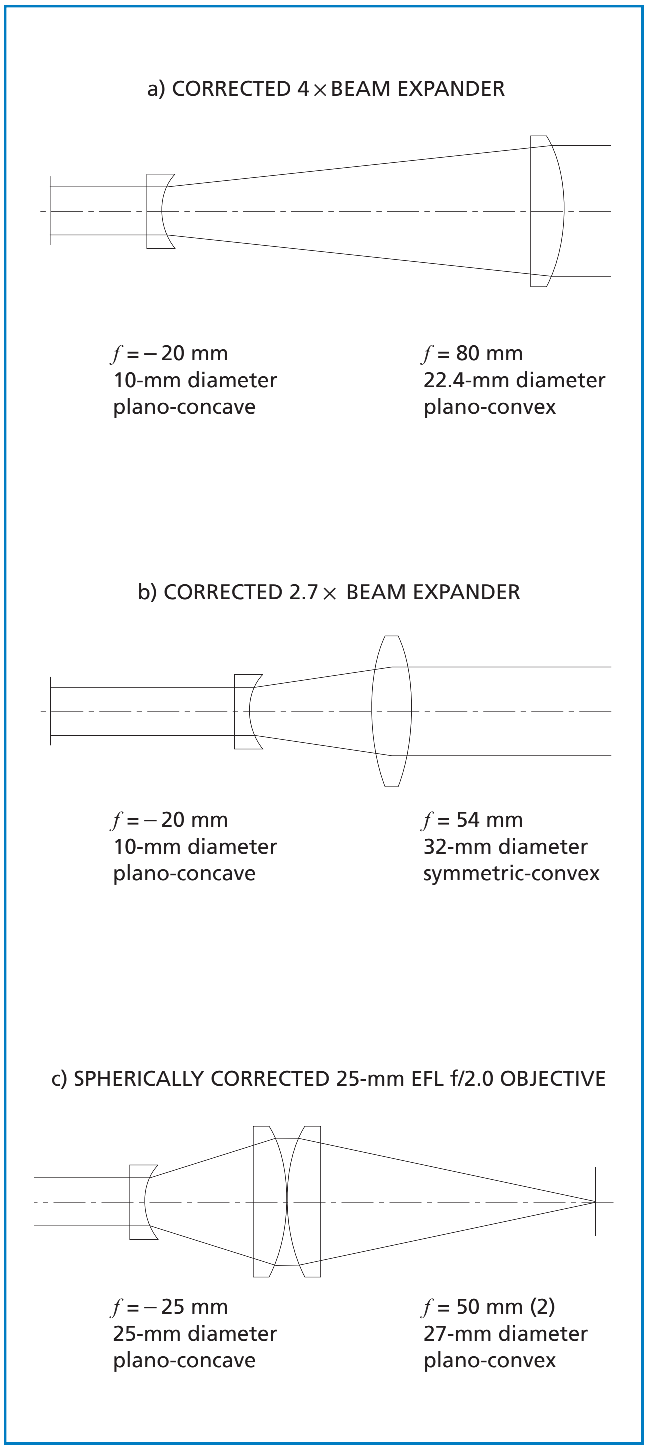
a) CORRECTED 4 × BEAM EXPANDER
f = – 20 mm
10-mm diameter
plano-concave
f = 80 mm
22.4-mm diameter
plano-convex
b) CORRECTED 2.7 × BEAM EXPANDER
f = – 20 mm
10-mm diameter
plano-concave
f = 54 mm
32-mm diameter
symmetric-convex
c) SPHERICALLY CORRECTED 25-mm EFL f/2.0 OBJECTIVE
f = – 25 mm
25-mm diameter
plano-concave
f = 50 mm (2)
27-mm diameter
plano-convex

Figure 1.32 **Combining catalog lenses for aberration balancing**

The relatively low *f* numbers of these objectives is a great advantage in minimizing the length of these beam expanders. They would be particularly useful with Nd:YAG and argon-ion lasers, which tend to have large output beam diameters.

These same principles can be utilized to create high numerical aperture objectives that might be used as laser focusing lenses. Figure 1.32(c) shows an objective consisting of an initial negative element, followed by two identical plano-convex positive elements. Again, all of the elements operate at the same f-number, so that their aberration contributions are proportional to their focal lengths. To obtain zero total spherical aberration from this configuration, we must satisfy

$$1.069\, f_1 + 0.272\, f_2 + 0.272\, f_2 = 0$$

or

$$\frac{f_1}{f_2} = -0.51.$$

Therefore, a corrected system should result if the focal length of the negative element is just about half that of each of the positive lenses. In this case, $f_1 = 425$ mm and $f_2 = 50$ mm yield a total system focal length of about 25 mm and an f-number of approximately f/2. This objective, corrected to 1/6 wave, has the additional advantage of a very long working distance.

Definition of Terms

FOCAL LENGTH (f)

Two distinct terms describe the focal lengths associated with every lens or lens system. The effective focal length (EFL) or equivalent focal length (denoted f in figure 1.33) determines magnification and hence the image size. The term f appears frequently in lens formulas and in the tables of standard lenses. Unfortunately, because f is measured with reference to principal points which are usually inside the lens, the meaning of f is not immediately apparent when a lens is visually inspected.

The second type of focal length relates the focal plane positions directly to landmarks on the lens surfaces (namely the vertices) which are immediately recognizable. It is not simply related to image size but is especially convenient for use when one is concerned about correct lens positioning or mechanical clearances. Examples of this second type of focal length are the front focal length (FFL, denoted f_f in figure 1.33) and the back focal length (BFL, denoted f_b).

The convention in all of the figures (with the exception of a single deliberately reversed ray) is that light travels from left to right.

FOCAL POINT (F OR F″)

Rays that pass through or originate at either focal point must be, on the opposite side of the lens, parallel to the optical axis. This fact is the basis for locating both focal points.

PRIMARY PRINCIPAL SURFACE

Let us imagine that rays originating at the front focal point F (and therefore parallel to the optical axis after emergence from the opposite side of the lens) are singly refracted at some imaginary surface, instead of twice refracted (once at each lens surface) as actually happens. There is a unique imaginary surface, called the principal surface, at which this can happen.

To locate this unique surface, consider a single ray traced from the air on one side of the lens, through the lens and into the air on the other side. The ray is broken into three segments by the lens. Two of these are external (in the air), and the third is internal (in the glass). The external segments can be extended to a common point of intersection (certainly near, and usually within, the lens). The principal surface is the locus of all such

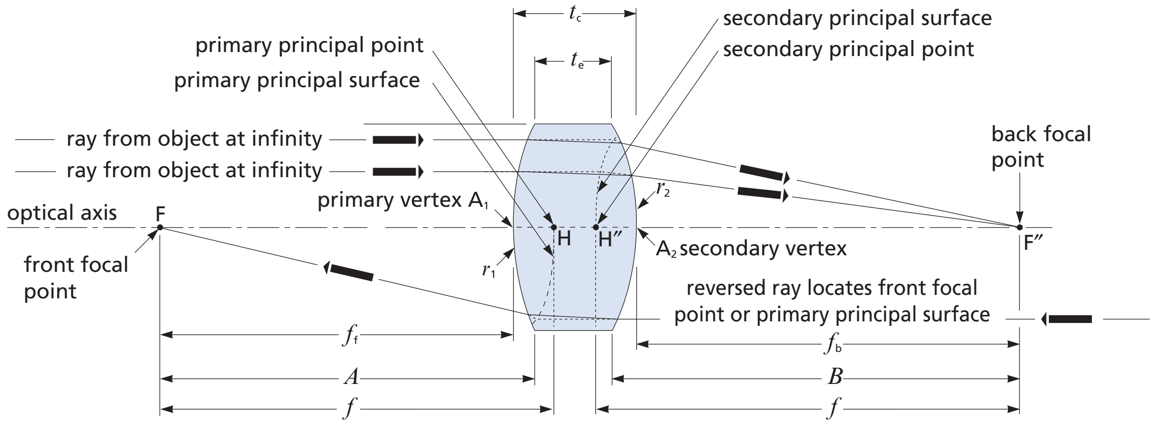
t_c
t_e
secondary principal surface
secondary principal point
primary principal point
primary principal surface
ray from object at infinity
ray from object at infinity
back focal point
primary vertex A_1
r_2
optical axis
F
H
H″
A_2 secondary vertex
F″
front focal point
r_1
reversed ray locates front focal point or primary principal surface
f_f
f_b
A
B
f
f

A = front focus to front edge distance

B = rear edge to rear focus distance

f = effective focal length; may be positive (as shown) or negative

f_f = front focal length

f_b = back focal length

t_e = edge thickness

t_c = center thickness

r_1 = radius of curvature of first surface (positive if center of curvature is to right)

r_2 = radius of curvature of second surface (negative if center of curvature is to left)

Figure 1.33 **Focal length and focal points**

points of intersection of extended external ray segments. The principal surface of a perfectly corrected optical system is a sphere centered on the focal point.

Near the optical axis, the principal surface is nearly flat, and for this reason, it is sometimes referred to as the principal plane.

SECONDARY PRINCIPAL SURFACE

This term is defined analogously to the primary principal surface, but it is used for a collimated beam incident from the left and focused to the back focal point F″ on the right. Rays in that part of the beam nearest the axis can be thought of as once refracted at the secondary principal surface, instead of being refracted by both lens surfaces.

PRIMARY PRINCIPAL POINT (H) OR FIRST NODAL POINT

This point is the intersection of the primary principal surface with the optical axis.

SECONDARY PRINCIPAL POINT (H″) OR SECONDARY NODAL POINT

This point is the intersection of the secondary principal surface with the optical axis.

CONJUGATE DISTANCES (s AND s'')

The conjugate distances are the object distance, s, and image distance, s''. Specifically, s is the distance from the object to H, and s'' is the distance from H″ to the image location. The term infinite conjugate ratio refers to the situation in which a lens is either focusing incoming collimated light or is being used to collimate a source (therefore, either s or s'' is infinity).

PRIMARY VERTEX (A_1)

The primary vertex is the intersection of the primary lens surface with the optical axis.

SECONDARY VERTEX (A_2)

The secondary vertex is the intersection of the secondary lens surface with the optical axis.

EFFECTIVE FOCAL LENGTH (EFL, f)

Assuming that the lens is surrounded by air or vacuum (refractive index 1.0), this is both the distance from the front focal point (F) to the primary principal point (H) and the distance from the secondary principal point (H″) to the rear focal point (F″). Later we use f to designate the paraxial EFL for the design wavelength (λ_0).

FRONT FOCAL LENGTH (f_f)

This length is the distance from the front focal point (F) to the primary vertex (A_1).

BACK FOCAL LENGTH (f_b)

This length is the distance from the secondary vertex (A_2) to the rear focal point (F″).

EDGE-TO-FOCUS DISTANCES (A AND B)

A is the distance from the front focal point to the primary vertex of the lens. B is the distance from the secondary vertex of the lens to the rear focal point. Both distances are presumed always to be positive.

REAL IMAGE

A real image is one in which the light rays actually converge; if a screen were placed at the point of focus, an image would be formed on it.

VIRTUAL IMAGE

A virtual image does not represent an actual convergence of light rays. A virtual image can be viewed only by looking back through the optical system, such as in the case of a magnifying glass.

F-NUMBER (f/#)

The f-number (also known as the focal ratio, relative aperture, or speed) of a lens system is defined to be the effective focal length divided by system clear aperture. Ray f-number is the conjugate distance for that ray divided by the height at which it intercepts the principal surface.

$$\mathrm{f/\#} = \frac{f}{\mathrm{CA}}. \qquad \text{(see eq. 1.7)}$$

NUMERICAL APERTURE (NA)

The NA of a lens system is defined to be the sine of the angle, θ_1, that the marginal ray (the ray that exits the lens system at its outer edge) makes with the optical axis multiplied by the index of refraction (n) of the medium. The NA can be defined for any ray as the sine of the angle made by that ray with the optical axis multiplied by the index of refraction:

$$\mathrm{NA} = n\sin\theta. \qquad (1.30)$$

MAGNIFICATION POWER

Often, positive lenses intended for use as simple magnifiers are rated with a single magnification, such as 4 × . To create a virtual image for viewing with the human eye, in principle, any positive lens can be used at an infinite number of possible magnifications. However, there is usually a narrow range of magnifications that will be comfortable for the viewer. Typically, when the viewer adjusts the object distance so that the image appears to be essentially at infinity (which is a comfortable viewing distance for most individuals), magnification is given by the relationship

$$\text{magnification} = \frac{254\text{ mm}}{f} \qquad (f \text{ in mm}). \tag{1.31}$$

Thus, a 25.4-mm focal length positive lens would be a 10 × magnifier.

DIOPTERS

The term diopter is used to define the reciprocal of the focal length, which is commonly used for ophthalmic lenses. The inverse focal length of a lens expressed in diopters is

$$\text{diopters} = \frac{1000}{f} \qquad (f \text{ in mm}). \tag{1.32}$$

Thus, the smaller the focal length is, the larger the power in diopters will be.

DEPTH OF FIELD AND DEPTH OF FOCUS

In an imaging system, depth of field refers to the distance in object space over which the system delivers an acceptably sharp image. The criteria for what is acceptably sharp is arbitrarily chosen by the user; depth of field increases with increasing f-number.

For an imaging system, depth of focus is the range in image space over which the system delivers an acceptably sharp image. In other words, this is the amount that the image surface (such as a screen or piece of photographic film) could be moved while maintaining acceptable focus. Again, criteria for acceptability are defined arbitrarily.

In nonimaging applications, such as laser focusing, depth of focus refers to the range in image space over which the focused spot diameter remains below an arbitrary limit.

APPLICATION NOTE

Technical Reference

For further reading about the definitions and formulas presented here, refer to the following publications:

Rudolph Kingslake, *Lens Design Fundamentals*
(Academic Press)

Rudolph Kingslake, System Design
(Academic Press)

Warren Smith, *Modern Optical Engineering*
(McGraw Hill).

Donald C. O'Shea, *Elements of Modern Optical Design*
(John Wiley & Sons)

Eugene Hecht, *Optics* (Addison Wesley)

Max Born, Emil Wolf, *Principles of Optics*
(Cambridge University Press)

If you need help with the use of definitions and formulas presented in this guide, our applications engineers will be pleased to assist you.

Paraxial Lens Formulas

PARAXIAL FORMULAS FOR LENSES IN AIR

The following formulas are based on the behavior of paraxial rays, which are always very close and nearly parallel to the optical axis. In this region, lens surfaces are always very nearly normal to the optical axis, and hence all angles of incidence and refraction are small. As a result, the sines of the angles of incidence and refraction are small (as used in Snell's law) and can be approximated by the angles themselves (measured in radians).

The paraxial formulas do not include effects of spherical aberration experienced by a marginal ray — a ray passing through the lens near its edge or margin. All EFL values (f) tabulated in this catalog are paraxial values which correspond to the paraxial formulas. The following paraxial formulas are valid for both thick and thin lenses unless otherwise noted. The refractive index of the lens glass, n, is the ratio of the speed of light in vacuum to the speed of light in the lens glass. All other variables are defined in figure 1.33.

Focal Length

$$\frac{1}{f}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right)+\frac{(n-1)^2}{n}\,\frac{t_c}{r_1 r_2} \qquad (1.33)$$

where n is the refractive index, t_c is the center thickness, and the sign convention previously given for the radii r_1 and r_2 applies. For thin lenses, $t_c \cong 0$, and for plano lenses either r_1 or r_2 is infinite. In either case the second term of the above equation vanishes, and we are left with the familiar lens maker's formula:

$$\frac{1}{f}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right). \qquad (1.34)$$

Surface Sagitta and Radius of Curvature

(refer to figure 1.34)

$$r^2=(r-s)^2+\left(\frac{d}{2}\right)^2 \qquad (1.35)$$

$$s=r-\sqrt{r^2-\left(\frac{d}{2}\right)^2}>0 \qquad (1.36)$$

$$r=\frac{s}{2}+\frac{d^2}{8s}. \qquad (1.37)$$

An often useful approximation is to neglect $s/2$.

Symmetric Lens Radii ($r_2 = -r_1$)

With center thickness constrained,

$$r_1=(n-1)\left[f\pm\sqrt{f^2-\left(\frac{f\,t_c}{n}\right)}\right] \qquad (1.38)$$

$$=(n-1)\,f\left[1+\sqrt{1-\left(\frac{t_c}{nf}\right)}\right]$$

where, in the first form, the + sign is chosen for the square root if f is positive, but the − sign must be used if f is negative. In the second form, the + sign must be used regardless of the sign of f.

Plano Lens Radius

Since r_2 is infinite,

$$r_1=(n-1)\,f. \qquad (1.39)$$

Principal-Point Locations (signed distances from vertices)

$$\mathrm{A_2H''}=\frac{-r_2\,t_c}{n(r_2-r_1)+t_c(n-1)} \qquad (1.40)$$

$$\mathrm{A_1H}=\frac{-r_1\,t_c}{n(r_2-r_1)+t_c(n-1)} \qquad (1.41)$$

where the above sign convention applies.

For symmetric lenses ($r_2 = -r_1$),

$$\mathrm{A_1H}=-\mathrm{A_2H''}$$

$$=\frac{r_1\,t_c}{2nr_1-t_c(n-1)}. \qquad (1.42)$$

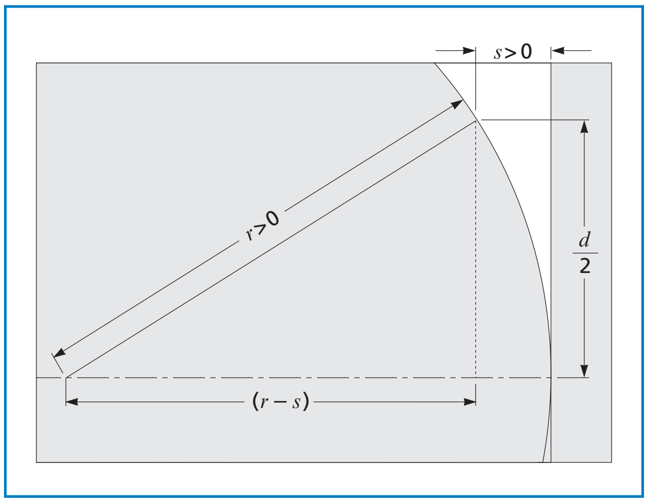

Figure 1.34 **Surface sagitta and radius of curvature**

If either r_1 or r_2 is infinite, l'Hôpital's rule from calculus must be used. Thus, referring to *Aberration Balancing*, for plano-convex lenses in the correct orientation,

$$A_1H = 0$$

and

$$A_2H'' = -\frac{t_c}{n}. \tag{1.43}$$

For flat plates, by letting $r_1 \rightarrow \infty$ in a symmetric lens, we obtain $A_1H = A_2H'' = t_c/2_n$. These results are useful in connection with the following paraxial lens combination formulas.

Hiatus or Interstitium (principal-point separation)

$$HH'' = t_c\left\{1 - \frac{f}{n}\left[\frac{1}{f} - \frac{(n-1)^2}{n}\frac{t_c}{r_1 r_2}\right]\right\} \tag{1.44}$$

which, in the thin-lens approximation (exact for plano lenses), becomes

$$HH'' = t_c\left(1 - \frac{1}{n}\right). \tag{1.45}$$

Solid Angle

The solid angle subtended by a lens, for an observer situated at an on-axis image point, is

$$\begin{aligned}\Omega &= 2\pi(1-\cos\theta)\\ &= 4\pi\sin^2\left(\frac{\theta}{2}\right)\end{aligned} \tag{1.46}$$

where this result is in steradians, and where

$$\theta = \arctan\left(\frac{CA}{2s}\right) \tag{1.47}$$

is the apparent angular radius of the lens clear aperture. For an observer at an on-axis object point, use s instead of s''. To convert from steradians to the more intuitive sphere units, simply divide Ω by 4π. If the Abbé sine condition is known to apply, θ may be calculated using the arc sine function instead of the arctangent.

Back Focal Length

$$\begin{aligned}f_b &= f'' + A_2H''\\ &= f'' - \frac{r_2 t_c}{n(r_2 - r_1) + t_c(n-1)}\end{aligned} \tag{1.48}$$

where the sign convention presented above applies to A_2H'' and to the radii. If r_2 is infinite, l'Hôpital's rule from calculus must be used, whereby

$$f_b = f'' - \frac{t_c}{n}. \tag{1.49}$$

Front Focal Length

$$\begin{aligned}f_f &= f - A_1H\\ &= f + \frac{r_1 t_c}{n(r_2 - r_1) + t_c(n-1)}\end{aligned} \tag{1.50}$$

where the sign convention presented above applies to A_1H and to the radii. If r_1 is infinite, l'Hôpital's rule from calculus must be used, whereby

$$f_f = f - \frac{t_c}{n}. \tag{1.51}$$

Edge-to-Focus Distances

For positive lenses,

$$A = f_f + s_1 \tag{1.52}$$

and

$$B = f_b + s_2 \tag{1.53}$$

where s_1 and s_2 are the sagittas of the first and second surfaces. Bevel is neglected.

Magnification or Conjugate Ratio

$$\begin{aligned}m &= \frac{s''}{s}\\ &= \frac{f}{s-f}\\ &= \frac{s''-f}{f}.\end{aligned} \tag{1.54}$$

PARAXIAL FORMULAS FOR LENSES IN ARBITRARY MEDIA

These formulas allow for the possibility of distinct and completely arbitrary refractive indices for the object space medium (refractive index n'), lens (refractive index n''), and image space medium (refractive index n). In this situation, the EFL assumes two distinct values, namely f in object space and f'' in image space. It is also necessary to distinguish the principal points from the nodal points. The lens serves both as a lens and as a window separating the object space and image space media.

The situation of a lens immersed in a homogenous fluid (figure 1.35) is included as a special case ($n = n''$). This case is of considerable practical importance. The two values f and f'' are again equal, so that the lens combination formulas are applicable to systems immersed in a common fluid. The general case (two different fluids) is more difficult, and it must be approached by ray tracing on a surface-by-surface basis.

Lens Constant (k)

This number appears frequently in the following formulas. It is an explicit function of the complete lens prescription (both radii, t_c and n') and both media indices (n and n''). This dependence is implicit anywhere that k appears.

$$k = \frac{n'-n}{r_1} + \frac{n''-n'}{r_2} - \frac{t_c(n'-n)(n''-n')}{n' r_1 r_2}. \qquad (1.55)$$

Effective Focal Lengths

$$f = \frac{n}{k} \qquad f'' = \frac{n''}{k}. \qquad (1.56)$$

Lens Formula (Gaussian form)

$$\frac{n}{s} + \frac{n''}{s''} = k. \qquad (1.57)$$

Lens Formula (Newtonian form)

$$xx'' = f''f = \frac{nn''}{k^2}. \qquad (1.58)$$

Principal-Point Locations

$$\mathrm{A_1H} = \frac{nt_c}{k}\left(\frac{n''-n'}{n' r_2}\right) \qquad (1.59)$$

$$\mathrm{A_2H''} = \frac{-n''t_c}{k}\left(\frac{n'-n}{n' r_1}\right). \qquad (1.60)$$

Object-to-First-Principal-Point Distance

$$s = \frac{ns''}{ks''-n''}. \qquad (1.61)$$

Second Principal-Point-to-Image Distance

$$s'' = \frac{n''s}{ks-n}. \qquad (1.62)$$

Magnification

$$m = \frac{ns''}{n''s}. \qquad (1.63)$$

Lens Maker's Formula

$$\frac{n}{f} = \frac{n''}{f''} = k. \qquad (1.64)$$

Nodal-Point Locations

$$\mathrm{A_1N} = \mathrm{A_1H} + \mathrm{HN} \qquad (1.65)$$

$$\mathrm{A_2N''} = \mathrm{A_2H''} + \mathrm{H''N''}. \qquad (1.66)$$

Separation of Nodal Point from Corresponding Principal Point

HN = H″N″ = $(n''-n)/k$, positive for N to right of H and N″ to right of H″.

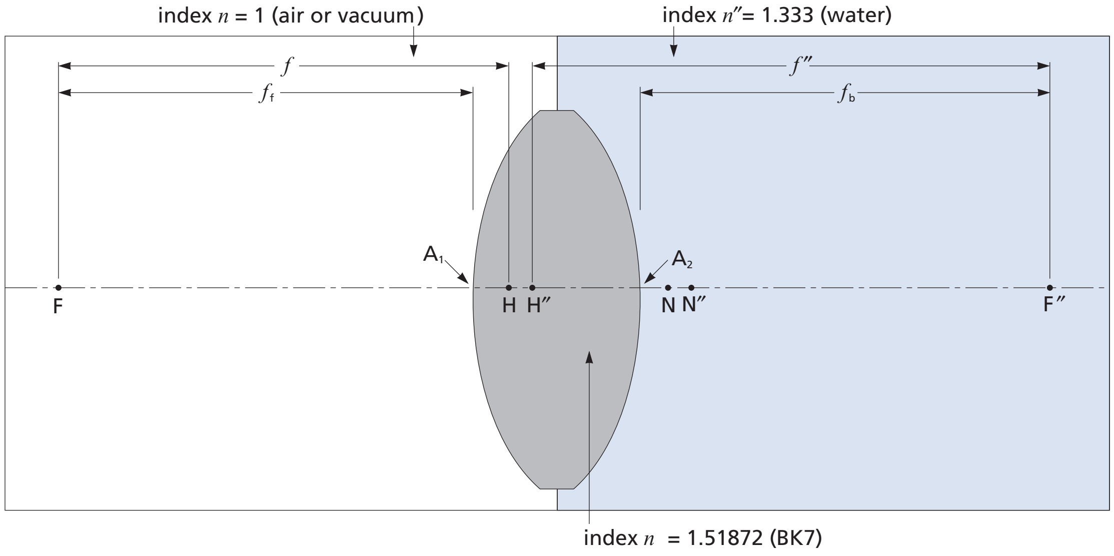
index n = 1 (air or vacuum)
index n''= 1.333 (water)
f
f_f
f''
f_b
A₁
A₂
F
H H″
N N″
F″
index n = 1.51872 (BK7)

Figure 1.35 **Symmetric lens with disparate object and image space indexes**

Back Focal Length

$$f_{\mathrm{b}} = f'' + \mathrm{A_2H''}. \qquad \textit{(see eq. 1.48)}$$

Front Focal Length

$$f_{\mathrm{f}} = f - \mathrm{A_1H}. \qquad \textit{(see eq. 1.50)}$$

Focal Ratios

The focal ratios are f/CA and f''/CA, where CA is the diameter of the clear aperture of the lens.

Numerical Apertures

$$n\sin\theta$$

$$\text{where } \theta = \arcsin\left(\frac{\mathrm{CA}}{2s}\right)$$

and

$$n''\sin\theta''$$

$$\text{where } \theta'' = \arcsin\left(\frac{\mathrm{CA}}{2s''}\right).$$

Solid Angles (in steradians)

$$\Omega = 2\pi\,(1-\cos\theta) \qquad \textit{(see eq. 1.46)}$$

$$= 4\pi\sin^2\left(\frac{\theta}{2}\right)$$

$$\text{where } \theta = \arctan\left(\frac{\mathrm{CA}}{2\mathrm{s}}\right)$$

$$\Omega = 2\pi(1 - \cos\theta'')$$

$$= 4\pi\sin^2\left(\frac{\theta}{2}\right)$$

$$\text{where } \theta'' = \arctan\left(\frac{\mathrm{CA}}{2\mathrm{s}''}\right).$$

To convert from steradians to spheres, simply divide by 4π.

APPLICATION NOTE

For Quick Approximations

Much time and effort can be saved by ignoring the differences among f, f_{b}, and f_{f} in these formulas by assuming that $f = f_{\mathrm{b}} = f_{\mathrm{f}}$. Then s becomes the lens-to-object distance; s'' becomes the lens-to-image distance; and the sum of conjugate distances $s + s''$ becomes the object-to-image distance. This is known as the thin-lens approximation.

APPLICATION NOTE

Physical Significance of the Nodal Points

A ray directed at the primary nodal point N of a lens appears to emerge from the secondary nodal point N″ without change of direction. Conversely, a ray directed at N″ appears to emerge from N without change of direction. At the infinite conjugate ratio, if a lens is rotated about a rotational axis orthogonal to the optical axis at the secondary nodal point (i.e., if N″ is the center of rotation), the image remains stationary during the rotation. This fact is the basis for the nodal slide method for measuring nodal-point location and the EFL of a lens. The nodal points coincide with their corresponding principal points when the image space and object space refractive indices are equal ($n = n''$). This makes the nodal slide method the most precise method of principal-point location.

Principal-Point Locations

Figure 1.36 indicates approximately where the principal points fall in relation to the lens surfaces for various standard lens shapes. The exact positions depend on the index of refraction of the lens material, and on the lens radii, and can be found by formula. in extreme meniscus lens shapes (short radii or steep curves), it is possible that both principal points will fall outside the lens boundaries. For symmetric lenses, the principal points divide that part of the optical axis between the vertices into three approximately equal segments. For plano lenses, one principal point is at the curved vertex, and the other is approximately one-third of the way to the plane vertex.

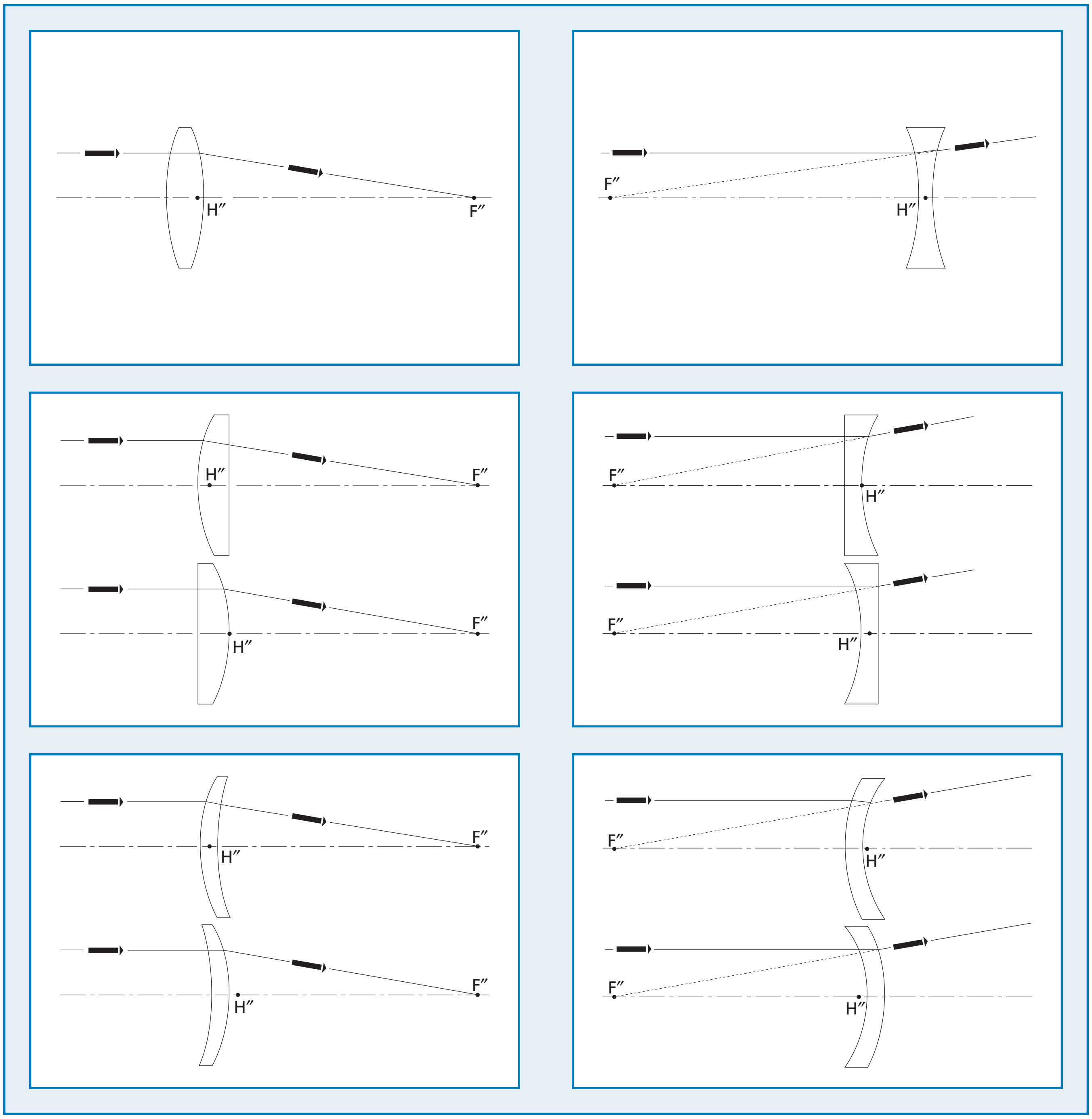

Figure 1.36 **Principal points of common lenses**

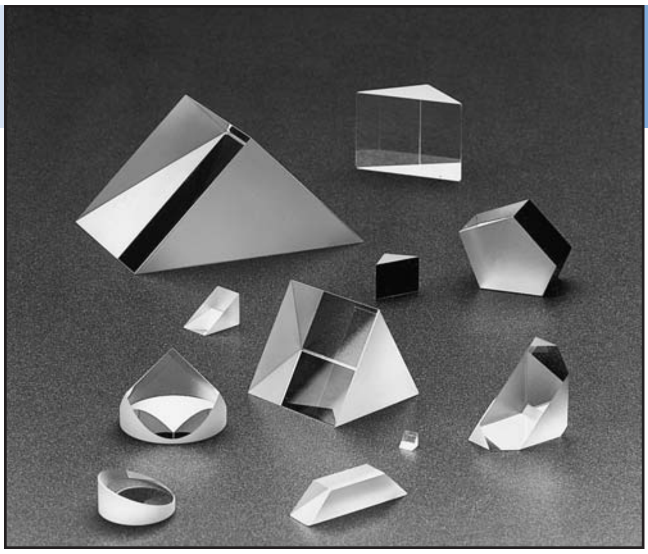

Prisms

Prisms are blocks of optical material whose flat, polished sides are arranged at precisely controlled angles to each other. Prisms may be used in an optical system to deflect or deviate a beam of light. They can invert or rotate an image, disperse light into its component wavelengths, and be used to separate states of polarization.

PRISM ORIENTATION

The orientation of a prism determines its effect on a beam of light or an image.

A viewer looks through a prism at an object and sees a virtual image (see Figure 1.37). This image may be displaced from the original object, or, if a dove prism is used, it may coincide with the object. Furthermore, image orientation may differ from the object; in the case of a right-angle prism, the image is reversed.

A real image (see Figure 1.38) can be formed only if imaging optics are present in the system. Without imaging optics, the image is virtual. A virtual image has the same orientation as the real image shown, but it can be viewed by the observer only by looking back through the prism system.

TOTAL INTERNAL REFLECTION

Rays incident upon a glass/air boundary (i.e., an internal reflection) at angles that exceed the critical angle are reflected with 100-percent efficiency regardless of their initial polarization state. The critical angle is given by

$$\theta_c(\lambda) = \arcsin\frac{1}{n_\lambda} \qquad (1.67)$$

and depends on the refractive index, which is a function of wavelength. If, at some wavelength, the refractive index should fall to less than $\sqrt{2} = 1.414$, the critical angle will exceed 45 degrees, and total internal reflection (TIR) will fail for a collimated beam internally incident at 45 degrees on the hypotenuse face of a right-angle prism. Reflectance decreases rapidly at angles of incidence smaller than the critical angle.

The index of BK7 is sufficiently high to guarantee the TIR of a collimated beam at 45 degrees internal incidence over the visible and near-infrared region. The possibility of significant TIR failure with convergent or divergent beams should be kept in mind if polarization is important. TIR can also fail if the hypotenuse face is not kept extremely clean. Even an almost invisible fingerprint can lead to TIR failure. An aluminum- or silver-coated hypotenuse is recommended for applications where the right-angle prism is frequently handled, or where convergent or divergent beams are used. There is a slight loss of reflectance at all internal angles with the coating, and no critical angle exists.

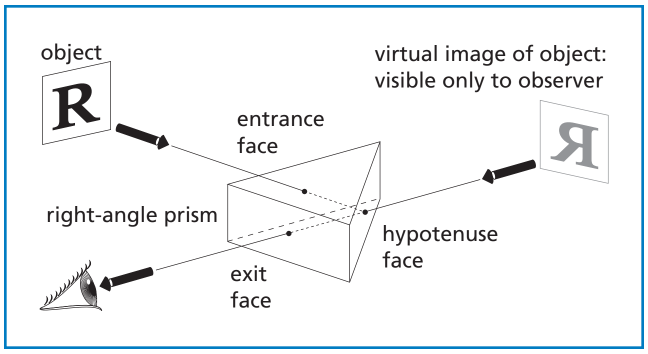

Figure 1.37 **Virtual imaging using a prism**

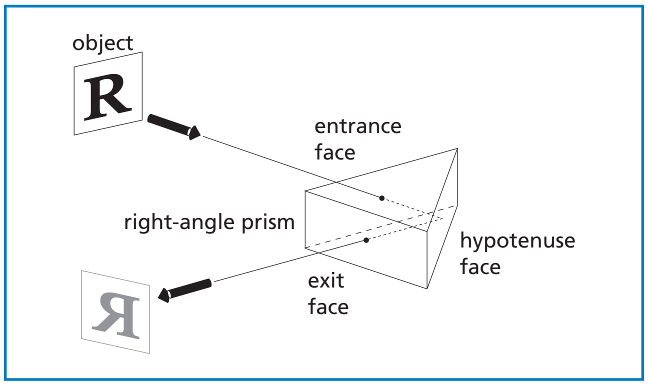

Figure 1.38 **Real imaging using a prism**

ABERRATIONS FOR PRISMS

Prisms will introduce aberrations when they are used with convergent or divergent beams of light. Using prisms with collimated or nearly collimated light will help minimize aberrations. Conjugate distances that include prisms should be long.

DISPERSING PRISMS

Dispersing prisms are used to separate a beam of white light into its component colors. Generally, the light is first collimated and then dispersed by the prism. A spectrum is then formed at the focal plane of a lens or curved mirror. In laser work, dispersing prisms are used to separate two wavelengths following the same beam path. Typically, the dispersed beams are permitted to travel far enough so the beams separate spatially.

A prism exhibits magnification in the plane of dispersion if the entrance and exit angles for a beam differ. This is useful in anamorphic (one-dimensional) beam expansion or compression, and may be used to correct or create asymmetric beam profiles.

As shown in Figure 1.39, a beam of width W_1 is incident at an angle α on the surface of a dispersing prism of apex angle A. The angle of refraction at the first surface, β, the angle of incidence at the second surface, γ, and the angle of refraction exiting the prism, δ, are easily calculated:

$$\beta = \sin^{-1}((\sin\alpha)\ /\ \eta)$$
$$\gamma = A - \beta$$
$$\delta = \sin^{-1}(\eta\sin\gamma)$$

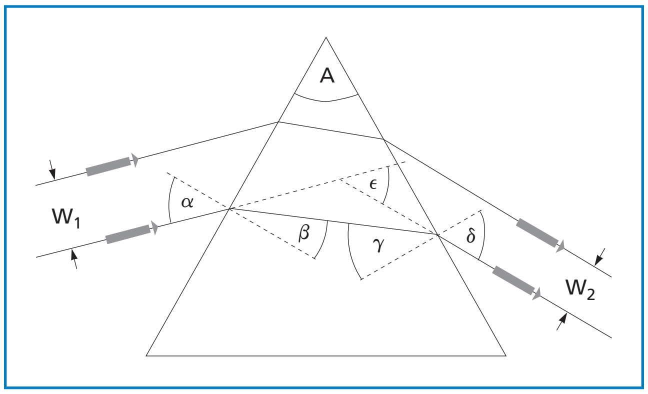

Figure 1.39 **Diagram of dispersing prism**

The beam deviation, ϵ, is of greatest importance. It is the angle the exit beam makes with its original direction.

$$\epsilon = \alpha + \delta - A \quad (1.68)$$

The magnification W_2/W_1 is given by:

$$M = \frac{\cos\delta\cos\beta}{\cos\alpha\cos\gamma} \quad (1.69)$$

The resolving power of a prism spectrometer angle α, the angular dispersion of the prism is given by:

$$\frac{d\delta}{d\lambda} = \left(\frac{\sin A}{\cos\delta\cos\beta}\right)\left(\frac{d\eta}{d\lambda}\right) \quad (1.70)$$

If the spectrum is formed by a diffraction limited focal system of focal length f, the minimum spot size is $dx \sim f\lambda/W$. This corresponds to a minimum angular resolution $d\delta \sim \lambda/w$ for a beam of diameter w. The diffraction limited angular resolution at a given beam diameter sets the limit on the spectral resolving power of a prism. Setting the expression for $d\delta$ equal to the minimum angular resolution, we obtain:

$$RP = \frac{\lambda}{d\lambda} = \left(\frac{w_2\sin A}{\cos\delta\cos\beta}\right)\left(\frac{d\eta}{d\lambda}\right) \quad (1.71)$$

where RP is the resolving power of the prism.

At a given wavelength, the beam deviation ϵ is a minimum at an angle of incidence:

$$\alpha_{\text{min dev}} = \sin^{-1}[\eta\sin(A/2)] \quad (1.72)$$

where η is the prism index of refraction at that wavelength. At this angle, the incident and exit angles are equal, the prism magnification is one, and the internal rays are perpendicular to the bisector of the apex angle.

By measuring the angle of incidence for minimum deviation, the index of refraction of a prism can be determined. Also, by proper choice of apex angle, the equal incident and exit angles may be made Brewster's angle, eliminating losses for p-polarized beams. The apex angle to choose is:

$$A = \pi - 2\theta_B \quad (1.73)$$

If, in addition, the base angles of the prism are chosen as Brewster's angle, an isosceles Brewster prism results.

Another use is illustrated next.

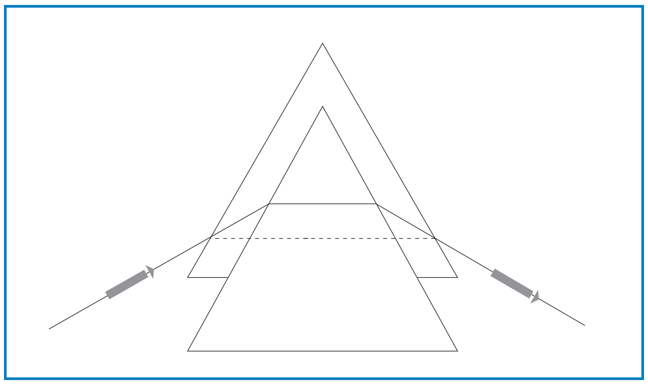

Figure 1.40 **Translation of a prism at minimum deviation**

At minimum deviation, translating a prism along the bisector of the apex angle does not disturb the direction of the output rays. See Figure 1.40. This is important in femtosecond laser design where intracavity prisms are used to compensate for group velocity dispersion. By aligning a prism for minimum deviation and translating it along its apex bisector, the optical path length in material may be varied with no misalignment, thus varying the contribution of the material to overall group velocity dispersion. Finally, it is possible to show that at minimum deviation

$$\mathrm{RP} = \left(\mathrm{b}_2 - \mathrm{b}_1\right)\frac{d\eta}{d\lambda} \qquad (1.74)$$

where the relevant quantities are defined in Figure 1.41.

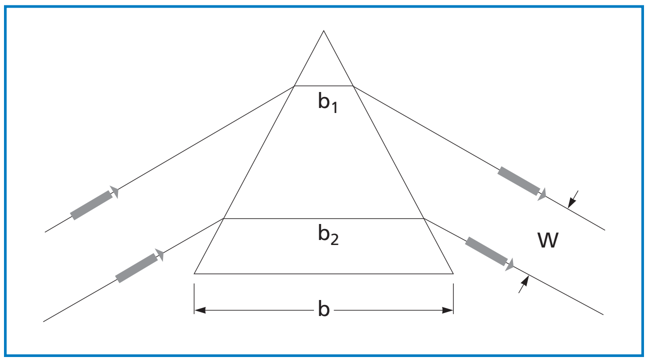

Figure 1.41 **Ray path lengths of a prism at minimum deviation**

If the beam is made to fill the prism completely, $b_1 = 0$, and $b_2 = b$, the base of the prism. So, we have the classical result that the resolving power of a prism spectrometer is equal to the base of the prism times the dispersion of the prism material.

As an example, consider CVI Melles Griot EDP-25-F2 prism, operating in minimum deviation at 590 nm. The angle of incidence and emergence are both then 54.09° and $d\eta/d\lambda$ is $-0.0854\ \mu\mathrm{m}^{-1}$ for F2 glass at 590 nm. If the 25-mm prism is completely filled, the resolving power, $\lambda/d\lambda$, is 2135. This is sufficient to resolve the Sodium D lines.

PELLIN BROCA PRISMS

In a Pellin Broca prism, an ordinary dispersing prism is split in half along the bisector of the apex angle. Using a right angle prism, the two halves are joined to create a dispersing prism with an internal right angle bend obtained by total internal reflection, as shown in Figure 1.42.

In principle, one can split any type of dispersing prism to create a Pellin Broca prism. Typically the Pellin Broca prism is based on an Isosceles Brewster prism. Provided the light is p-polarized, the prism will be essentially lossless.

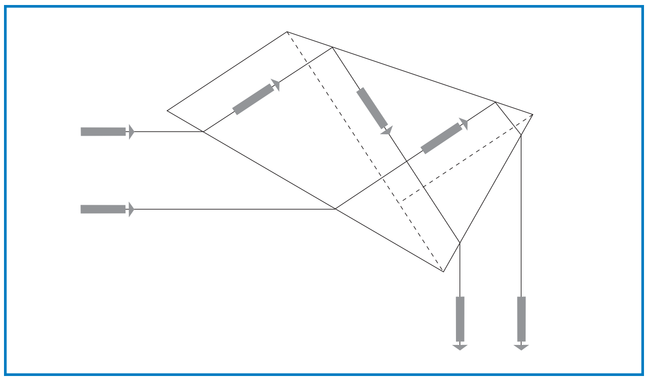

Figure 1.42 **One of the wavelengths deviates at exactly 90° to its initial direction**

Suppose wavelengths λ_1 and λ_2 are superimposed in a collimated beam, as at the output of a harmonic generating crystal, the diagram in Figure 1.42 suggests that it is always possible to find a rotation of the prism in its plane that ensures that one of the two wavelengths will operate at minimum deviation when refracting at the input face of the first of the half-dispersing prisms. This means that it will enter the right angle prism normal to one of its faces, be turned exactly 90°, be presented to the second half-dispersing prism in minimum deviation, and hence exit the Pellin Broca prism deviated at exactly 90° to its initial direction.

A simple dispersing prism always deviates the longer wavelength less than the shorter wavelength. In a Pellin Broca prism, whether the longer wavelength is deviated more or less depends on the orientation of the prism. This is an important consideration when designing a high power Pellin Broca beam separator, as shown in Figures 1.43 and 1.44.

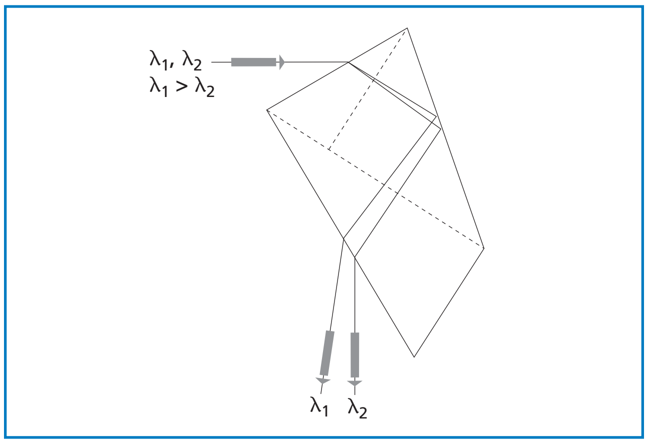

Figure 1.43 **Longer wavelength is deviated more than the shorter wavelength**

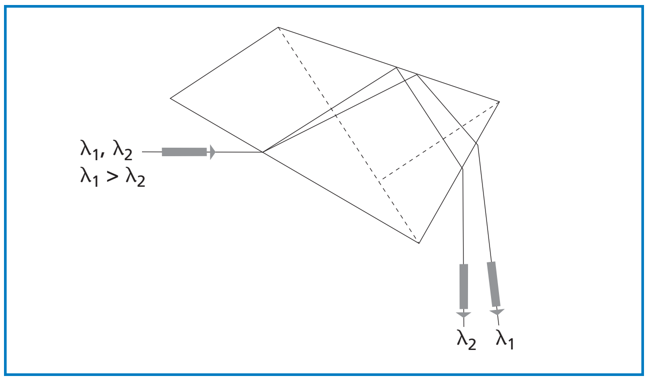

Figure 1.44 **Longer wavelength is deviated less than the shorter wavelength**

CVI Melles Griot offers Brewster angle Pellin Brocca prisms in a number of sizes and materials. BK7 prisms are used in the visible and near IR, and is the least expensive. UV-grade fused silica Pellin Broca prisms are used from 240 nm to 2000 nm. Excimer-grade prisms are used in the 180-nm to 240-nm region. Crystal-quartz Pellin Broca prisms are specifically designed for high-power Q-switched 266-nm laser pulses at fluence levels of 50 mJ/cm^2. Fused silica prisms track (i.e., suffer internal catastrophic damage) above this fluence, probably due to self-focusing.

PORRO PRISMS

A Porro prism, named for its inventor Ignazio Porro, is a type of reflection prism used to alter the orientation of an image. In operation, light enters the large face of the prism, undergoes total internal reflection twice from the 45° sloped faces, and exits again through the large face. An image traveling through a Porro prism is rotated by 180° and exits in the opposite direction offset from its entrance point, as shown in Figure 1.45. Since the image is reflected twice, the handedness of the image is unchanged. Porro prisms have rounded edges to minimize breakage and facilitate assembly.

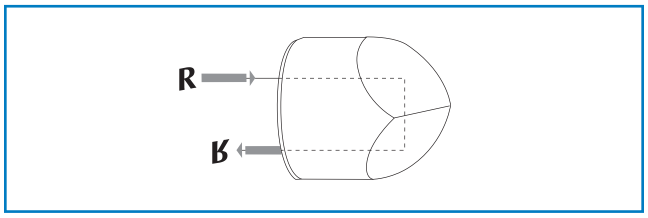

Figure 1.45 **Porro prisms retroreflect and invert the image**

Porro prisms are most often used in pairs, forming a double Porro prism, as shown in Figure 1.46. A second prism, rotated 90° with respect to the first, is placed such that the beam will traverse both prisms. The net effect of the prism system is a beam parallel to but displaced from its original direction, with the image rotated 180°. As before, the handedness of the image is unchanged.

Double Porro prism systems are used in small optical telescopes to reorient an inverted image and in many binoculars to both re-orient the image and provide a longer, folded distance between the objective lenses and the eyepieces.

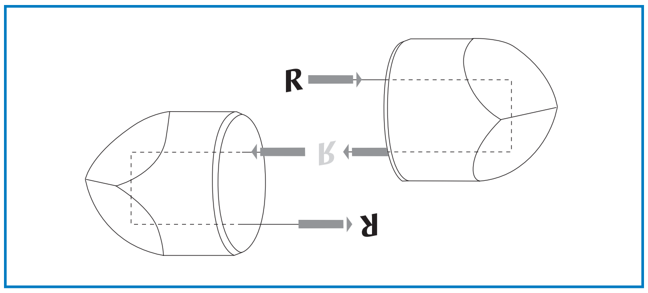

Figure 1.46 **Double porro prism results in a beam parallel to but displaced from its original direction, with the image rotated 180°**

MELLES GRIOT GLAN THOMPSON 03 PTO
MELLES GRIOT GLAN TAYLOR 03 PTY
MELLES GRIOT GLAN LASER PRISM 03 PLS 301

Polarization

Polarization States

Four numbers are required to describe a single plane wave Fourier component traveling in the $+z$ direction. These can be thought of as the amplitude and phase shift of the field along two orthogonal directions.

1. CARTESIAN REPRESENTATION

In Cartesian coordinates, the propagation equation for an electric field is given by the formula

$$E(x,y,z) = (xE_x e^{i\phi_x} + yE_y e^{i\phi_y})e^{i(kz-\omega t)} \tag{1.75}$$

where E_x, E_y, ϕ_x, and ϕ_y are real numbers defining the magnitude and the phase of the field components in the orthogonal unit vectors x and y. If the origin of time is irrelevant, only the relative phase shift

$$\phi = \phi_x - \phi_y \tag{1.76}$$

need be specified.

2. CIRCULAR REPRESENTATION

In the circular representation, we resolve the field into circularly polarized components. The basic states are represented by the complex unit vectors

$$e_+ = \left(1/\sqrt{2}\right)\left(x+iy\right) \text{ and} \tag{1.77}$$

$$e_- = \left(1/\sqrt{2}\right)\left(x-iy\right) \tag{1.78}$$

where e_+ is the unit vector for left circularly polarized light; for positive helicity light; for light that rotates counterclockwise in a fixed plane as viewed facing into the light wave; and for light whose electric field rotation obeys the right hand rule with thumb pointing in the direction of propagation.

APPLICATION NOTE

Polarization Convention

Historically, the orientation of a polarized electromagnetic wave has been defined in the optical regime by the orientation of the electric vector. This is the convention used by CVI Melles Griot.

e_- is the unit vector for right circularly polarized light; for negative helicity light; for light that rotates clockwise in a fixed plane as viewed facing into the light wave; and for light whose electric field rotation disobeys the right hand rule with thumb pointing in the direction of propagation.

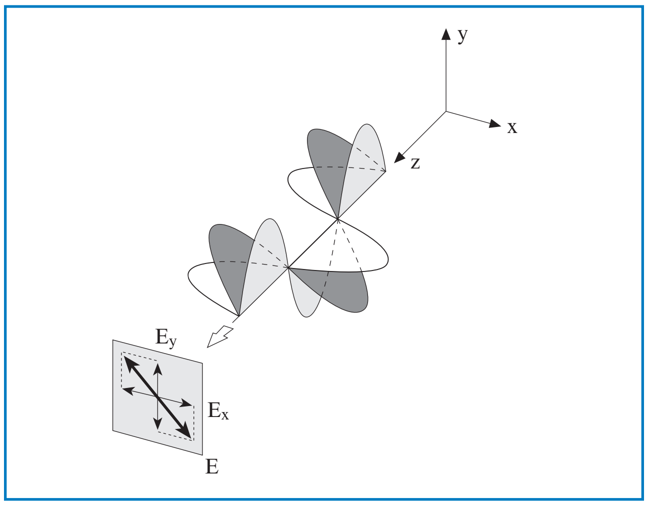

Figure 1.47 **Linearly polarized light. E_x and E_y are in phase**

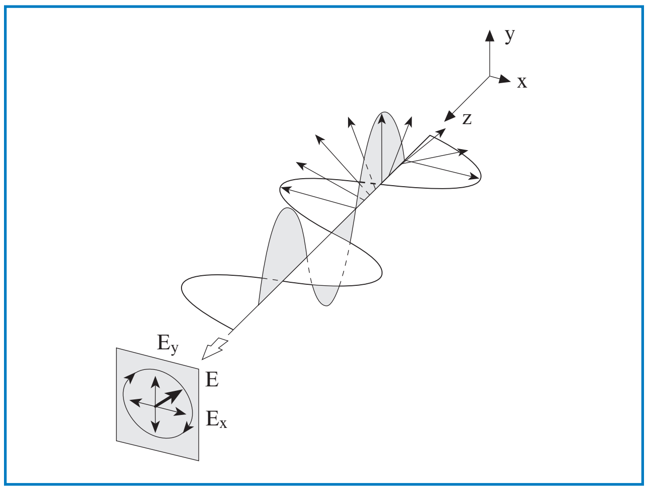

Figure 1.48 **Circularly polarized light. E_x and E_y are out of phase by angular frequency ω**

1ch_FundamentalOptics_Final_a.qxd 6/15/2009 2:31 PM Page 1.42

As in the case of the Cartesian representation, we write

$$E=(e_{+}E_{+}\mathrm{e}^{i\phi+}+e_{-}E_{-}\mathrm{e}^{i\phi-})\,e^{i(kz-\omega t)} \qquad (1.79)$$

where E_{+}, E_{-}, ϕ_{+}, and ϕ_{-} are four real numbers describing the magnitudes and phases of the field components of the left and right circularly polarized components. Note that

$$E_{+}=e_{-}\bullet E \qquad (1.80)$$

$$E_{-}=e_{+}\bullet E \qquad (1.81)$$

3. ELLIPTICAL REPRESENTATION

An arbitrary polarization state is generally elliptically polarized. This means that the tip of the electric field vector will describe an ellipse, rotating once per optical cycle.

Let a be the semimajor and b be the semiminor axis of the polarization ellipse. Let ψ be the angle that the semimajor axis makes with the x axis.

Let ξ and $\boldsymbol{\eta}$ be the axes of a right-handed coordinate system rotated by an angle $+\psi$ with respect to the x axis and aligned with the polarization ellipse as shown in Figure 1.49.

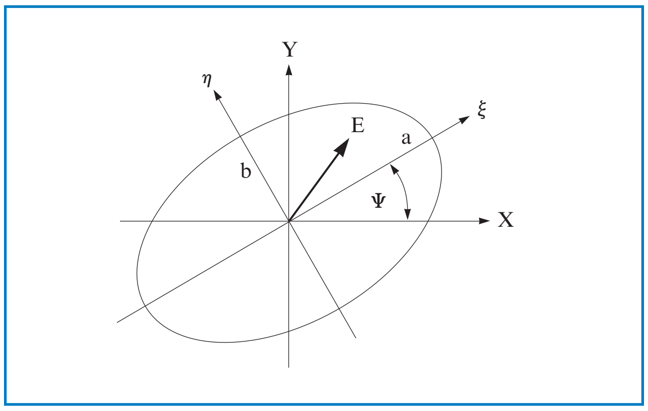

Figure 1.49 **The polarization ellipse**

The elliptical representation is:

$$E=\left(a\hat{\xi}+b\hat{\eta}\right)e^{i\delta_0}e^{i(kz-\omega t)} \qquad (1.82)$$

Note that the phase shift δ_o above is required to adjust the time origin, and the parameter ψ is implicit in the rotation of the ξ and $\boldsymbol{\eta}$ axes with respect to the x and y axes.

CONVERSIONS BETWEEN REPRESENTATIONS

For brevity, we will provide only the Cartesian to circular and Cartesian to elliptical transformations. The inverse transformations are straightforward. We define the following quantities:

$$g_1=E_x\cos\phi_x-E_y\sin\phi_y \qquad (1.83)$$

$$g_2=E_x\sin\phi_x+E_y\cos\phi_y \qquad (1.84)$$

$$g_3=E_x\cos\phi_x+E_y\sin\phi_y \qquad (1.85)$$

$$g_4=E_x\sin\phi_x-E_y\cos\phi_y \qquad (1.86)$$

$$u=\left(g_1^2+g_2^2\right)^{1/2} \qquad (1.87)$$

$$v=\left(g_3^2+g_4^2\right)^{1/2} \qquad (1.88)$$

$$\phi_{12}=\mathrm{atan}\,(g_1,g_2) \qquad (1.89)$$

$$\phi_{34}=\mathrm{atan}\,(g_3,g_4) \qquad (1.90)$$

In the above, atan(x,y) is the four quadrant arc tangent function. This means that atan$(x,y)=$ atan(y/x) with the provision that the quadrant of the angle returned by the function is controlled by the signs of both x and y, not just the sign of their quotient; for example, if $g_2=g_1=-1$, then ϕ_{12} above is $5\pi/4$ or $-3\pi/4$, not $\pi/4$.

A. CARTESIAN TO CIRCULAR TRANSFORMATION

$$E_{+}=v/\sqrt{2} \qquad (1.91)$$

$$E_{-}=u/\sqrt{2} \qquad (1.92)$$

$$\phi_{+}=\phi_{34} \qquad (1.93)$$

$$\phi_{-}=\phi_{12} \qquad (1.94)$$

B. CARTESIAN TO ELLIPTICAL TRANSFORMATION

$$a=(v+u)/2 \qquad (1.95)$$

$$b=(v-u)/2 \qquad (1.96)$$

$$\psi=(\phi_{12}-\phi_{34})/2 \qquad (1.97)$$

$$\delta_0=(\phi_{12}+\phi_{34})/2 \qquad (1.98)$$

LINEAR POLARIZERS

A linear polarizer is a device that creates a linear polarization state from an arbitrary input. It does this by removing the component orthogonal to the selected state. Unlike plastic sheet polarizers which absorb the rejected beam (which turns into heat), cube polarizers and thin-film plate polarizers reflect the rejected beam, creating two usable beams. Still others may refract the two polarized beams at different angles, thereby separating them. Examples are Wollaston and Rochon prism polarizers.

Suppose the pass direction of the polarizer is determined by unit vector p. Then the transmitted field E_2, in terms of the incident field E_1, is given by

$$E_2 = p\left(p \bullet E_1\right) \tag{1.99}$$

where the phase shift of the transmitted field has been ignored.

A real polarizer has a pass transmission, $T_\parallel$, less than 1. The transmission of the rejected beam, $T_\perp$, may not be 0. If r is a unit vector along the rejected direction, then

$$E_2 = \left(T_\parallel\right)^{1/2} p\left(p \bullet E_1\right)e^{i\phi_\parallel} + \left(T_\perp\right)^{1/2} r\left(r \bullet E_1\right)e^{i\phi_\perp} \tag{1.100}$$

In the above, the phase shifts along the two directions must be retained. Similar expressions could be arrived at for the rejected beam. If θ is the angle between the field E_1 and the polarizer pass direction p, the above equation predicts that

$$T = T_\parallel \cos^2\theta + T_\perp \sin^2\theta \tag{1.101}$$

The above equation shows that, when the polarizer is aligned so that $\theta = 0$, $T = T_\parallel$. When it is "crossed", $\theta = \pi/2$, and $T = T_\perp$. The extinction ratio is $\epsilon = T_\parallel / T_\perp$. A polarizer with perfect extinction has $T_\perp = 0$, and thus $T = T_\parallel \cos^2\theta$ is a familiar result. Because $\cos^2\theta$ has a broad maximum as a function of orientation angle, setting a polarizer at a maximum of transmission is generally not very accurate. One has to either map the $\cos^2\theta$ with sufficient accuracy to find the $\theta = 0$ point, or do a null measurement at $\theta = \pm\pi/2$.

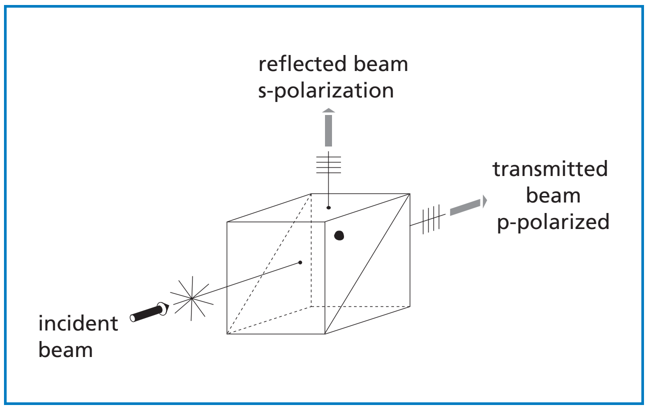

Figure 1.50 **At CVI Melles Griot, the DOT marks preferred input face. This is the tested direction for transmitted wavefront. Damage threshold is also higher for this orientation as well.**

Polarization Definitions

BIREFRINGENCE

A birefringent crystal, such as calcite, will divide an entering beam of monochromatic light into two beams having opposite polarization. The beams usually propagate in different directions and will have different speeds. There will be only one or two optical axis directions within the crystal in which the beam will remain collinear and continue at the same speed, depending on whether the birefringent crystal is uniaxial or biaxial.

If the crystal is a plane-parallel plate, and the optical axis directions are not collinear with the beam, radiation will emerge as two separate, orthogonally polarized beams (see Figure 1.51). The beam will be unpolarized where the beams overlap upon emergence. The two new beams within the material are distinguished from each other by more than just polarization and velocity. The rays are referred to as extraordinary (E) and ordinary (O). These rays need not be confined to the plane of incidence. Furthermore, the velocity of these rays changes with direction. Thus, the index of refraction for extraordinary rays is also a continuous function of direction. The index of refraction for the ordinary ray is constant and is independent of direction.

The two indexes of refraction are equal only in the direction of an optical axis within the crystal. The dispersion curve for ordinary rays is a single, unique curve when the index of refraction is plotted against wavelength. The dispersion curve for the extraordinary ray is a family of curves with different curves for different directions. Unless it is in a particular polarization state, or the crystalline surface is perpendicular to an optical axis, a ray normally incident on a birefringent surface will be divided in two at the boundary. The extraordinary ray will be deviated; the ordinary ray will not. The ordinary ray index n, and the most extreme (whether greater or smaller) extraordinary ray index n_e, are together known as the principal indices of refraction of the material.

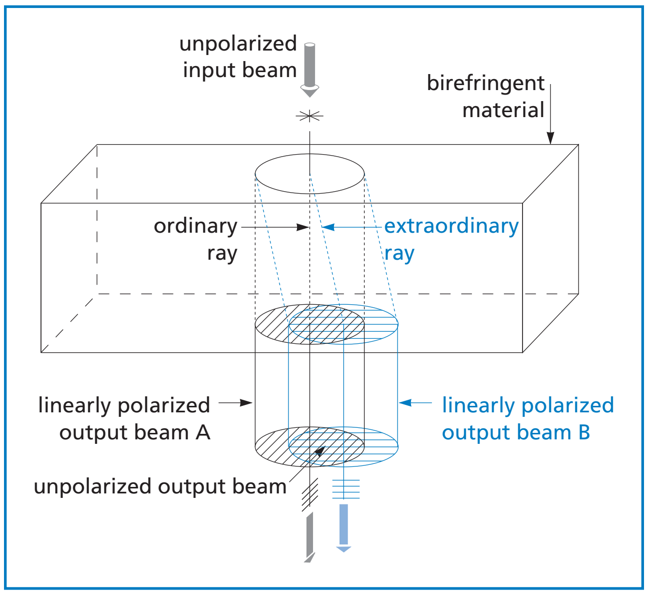

Figure 1.51 **Double refraction in a birefringent crystal**

If a beam of linearly polarized monochromatic light enters a birefringent crystal along a direction not parallel to the optical axis of the crystal, the beam will be divided into two separate beams. Each will be polarized at right angles to the other and will travel in different directions. The original beam energy, which will be divided between the new beams, depends on the original orientation of the vector to the crystal.

The energy ratio between the two orthogonally polarized beams can be any value. It is also possible that all energy will go into one of the new beams. If the crystal is cut as a plane-parallel plate, these beams will recombine upon emergence to form an elliptically polarized beam.

The difference between the ordinary and extraordinary ray may be used to create birefringent crystal polarization devices. In some cases, the difference in refractive index is used primarily to separate rays and eliminate one of the polarization planes, for example, in Glan-type polarizers. In other cases, such as Wollaston and Thompson beamsplitting prisms, changes in propagation direction are optimized to separate an incoming beam into two orthogonally polarized beams.

DICHROISM

Dichroism is selective absorption of one polarization plane over the other during transmission through a material. Sheet-type polarizers are manufactured with organic materials imbedded into a plastic sheet. The sheet is stretched, aligning molecules and causing them to be birefringent, and then dyed. The dye molecules selectively attach themselves to aligned polymer molecules, so that absorption is high in one plane and weak in the other. The transmitted beam is linearly polarized. Polarizers made of such material are very useful for low-power and visual applications. The usable field of view is large (up to grazing incidence), and diameters in excess of 100 mm are available.

POLARIZATION BY REFLECTION

When a beam of ordinary light is incident at the polarizing angle on a transmissive dielectric such as glass, the emerging refracted ray is partially linearly polarized. For a single surface (with n=1.50) at Brewster's angle, 100 percent of the light whose electric vector oscillates parallel to the plane of incidence is transmitted. Only 85 percent of the perpendicular light is transmitted (the other 15 percent is reflected). The degree of polarization from a single-surface reflection is small.

If a number of plates are stacked parallel and oriented at the polarizing angle, some vibrations perpendicular to the plane of incidence will be reflected at each surface, and all those parallel to it will be refracted. By making the number of plates within the stack large (more than 25), high degrees of linear polarization may be achieved. This polarization method is utilized in CVI Melles Griot polarizing beamsplitter cubes which are coated with many layers of quarter-wave dielectric thin films on the interior prism angle. This beamsplitter separates an incident laser beam into two perpendicular and orthogonally polarized beams.

THIN METAL FILM POLARIZERS

Optical radiation incident on small, elongated metal particles will be preferentially absorbed when the polarization vector is aligned with the long axis of the particle. CVI Melles Griot infrared polarizers utilize this effect to make polarizers for the near-infrared. These polarizers are considerably more effective than dichroic polarizers.

Polarizing thin films are formed by using the patented Slocum process to deposit multiple layers of microscopic silver prolate spheroids onto a polished glass substrate. The exact dimensions of these spheroids determine the optical properties of the film. Peak absorption can be selected for any wavelength from 400 to 3000 nm by controlling the deposition process. Contrast ratios up to 10,000:1 can be achieved with this method. Other CVI Melles Griot high-contrast polarizers exhibit contrasts as high as 100,000:1.

CALCITE

Calcite, a rhombohedral crystalline form of calcium carbonate, is found in various forms such as limestone and marble. Since calcite is a naturally occurring material, imperfections are not unusual. The highest quality materials, those that exhibit no optical defects, are difficult to find and are more expensive than those with some defects. Applications for calcite components typically fall into laser applications or optical research. CVI Melles Griot offers calcite components in two quality grades to meet those various needs.

There are three main areas of importance in defining calcite quality.

Spectral Properties

Trace amounts of chemical impurities, as well as lattice defects, can cause calcite to be colored, which changes absorption. For visible light applications, it is essential to use colorless calcite. For near-infrared applications, material with a trace of yellow is acceptable. This yellow coloration results in a 15-percent to 20-percent decrease in transmission below 420 nm.

Wavefront Distortion (Striae)

Striae, or streaked fluctuations in the refractive index of calcite, are caused by dislocations in the crystal lattice. They can cause distortion of a light wavefront passing through the crystal. This is particularly troublesome for interferometric applications.

Scatter

Small inclusions within the calcite crystal account for the main source of scatter. They may appear as small cracks or bubbles. In general, scatter presents a significant problem only when the polarizer is being used with a laser. The amount of scatter centers that can be tolerated is partially determined by beam size and power.

CVI MELLES GRIOT CALCITE GRADES

CVI Melles Griot has selected the most applicable calcite qualities, grouped into two grades:

Laser Grade

Calcite with a wavefront deformation of $\lambda/4$ at 633 nm or better due to striae only.

Optical Grade

Calcite with a wavefront deformation of 1λ to $\lambda/4$ at 633 nm due to striae only.

1ch_FundamentalOptics_Final_a.qxd 6/15/2009 2:31 PM Page 1.46

Waveplates

Waveplates use birefringence to impart unequal phase shifts to the orthogonally polarized field components of an incident wave, causing the conversion of one polarization state into another.

There are two types of birefringence. With linear birefringence, the index of refraction (and hence the phase shift) differs for two orthogonally polarized linear polarization states. This is the operation mode of standard waveplates. With circular birefringence, the index of refraction and hence phase shift differs for left and right circularly polarized components. This is the operation mode of polarization rotators.

STANDARD WAVEPLATES: LINEAR BIREFRINGENCE

Suppose a waveplate made from a uniaxial material has light propagating perpendicular to the optic axis. This makes the field component parallel to the optic axis an extraordinary wave and the component perpendicular to the optic axis an ordinary wave. If the crystal is positive uniaxial, $\eta_e > \eta_o$, then the optic axis is called the slow axis, which is the case for crystal quartz. For negative uniaxial crystals $\eta_e < \eta_o$, the optic axis is called the fast axis.

The equation for the transmitted field E_2, in terms of the incident field E_1 is:

$$E_2 = \mathrm{s}(\mathrm{s} \bullet E_1)e^{i\phi_s} + \mathrm{f}(\mathrm{f} \bullet E_1)e^{i\phi_f} \qquad (1.102)$$

where s and f are unit vectors along the slow and fast axes. This equation shows explicity how the waveplate acts on the field. Reading from left to right, the waveplate takes the component of the input field along its slow axis and appends the slow axis phase shift to it. It does a similar operation to the fast component.

The slow and fast axis phase shifts are given by:

$$\phi_s = \eta_s(\omega)\omega t / c = 2\pi\eta_s(\lambda)t / \lambda \qquad (1.103)$$

$$\phi_f = \eta_f(\omega)\omega t / c = 2\pi\eta_f(\lambda)t / \lambda \qquad (1.104)$$

where η_s and η_f are, respectively, the indices of refraction along the slow and fast axes, and t is the thickness of the waveplate.

To further analyze the effect of a waveplate, we throw away a phase factor lost in measuring intensity, and assign the entire phase delay to the slow axis:

$$E_2 = \mathrm{s}(\mathrm{s} \bullet E_1)e^{i\phi} + \mathrm{f}(\mathrm{f} \bullet E1) \qquad (1.105)$$

$$\phi = \phi_s - \phi_f = 2\pi\left[\eta_s(\lambda) - \eta_f(\lambda)\right]t / \lambda \qquad (1.106)$$

$$\phi = 2\,\pi\,\Delta\eta(\lambda)t / \lambda \qquad (1.107)$$

In the above, $\Delta\eta(\lambda)$ is the birefringence $\eta_s(\lambda)$ - $\eta_f(\lambda)$. The dispersion of the birefringence is very important in waveplate design; a quarter waveplate at a given wavelength is never exactly a half waveplate at half that wavelength.

Let E_1 be initially polarized along x, and let the waveplate slow axis make an angle θ with the x axis. This orientation is shown in Figure 1.52.

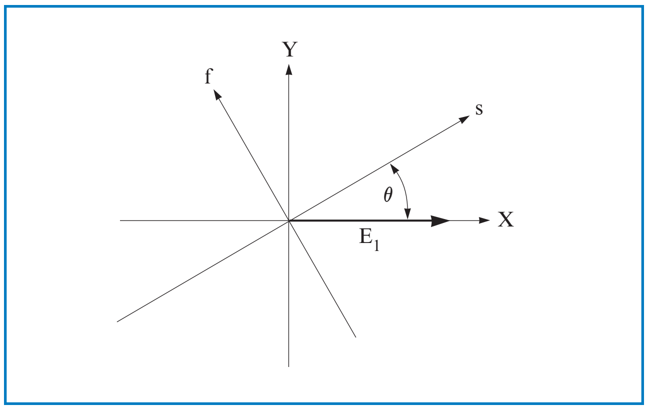

Figure 1.52. Orientation of the slow and fast axes of a waveplate with respect to an x-polarized input field

When the waveplate is placed between parallel and perpendicular polarizers the transmissions are given by:

$$T_{\parallel} \propto |E_2 x|^2 = 1 - \sin^2\ 2\theta \sin^2\phi/2 \qquad (1.108)$$

$$T_{\perp} \propto |E_2 y|^2 = \sin^2\ 2\theta \sin^2\phi/2 \qquad (1.109)$$

Note that θ is only a function of the waveplate orientation, and ϕ is only a function of the wavelength, the birefringence is a function of wavelength and the plate thickness.

For a full-wave waveplate:

$\phi = 2m\pi$, $T_{\parallel} = 1$, and $T_{\perp} = 0$, regardless of waveplate orientation.

For a half-wave waveplate:

$$\phi = (2m + 1)\pi,\ T_{\parallel} = \cos^2 2\theta, \text{ and } T_{\perp} = \sin^2 2\theta. \qquad (1.110)$$

This transmission result is the same as if an initial linearly polarized wave were rotated through an angle of 2θ. Thus, a half-wave waveplate finds use as a polarization rotator.

For a quarter waveplate,

$$\phi = (2m + 1)\pi/2 \text{ (i.e., an odd multiple of } \pi/2). \qquad (1.111)$$

To analyze this, we have to go back to the field equation. Assume that the slow and fast axis unit vectors s and f form a right handed coordinate system such that s x f = +z, the direction of propagation. To obtain circularly polarized light, linearly polarized light must be aligned midway between the slow and fast axes. There are four possibilities listed in the table below.

Phase Shift	Input Field Along $(\mathbf{s} + \mathbf{f})/\sqrt{2}$	Input Field Along $(\mathbf{s} - \mathbf{f})/\sqrt{2}$
$\phi = \pi/2 + 2m\pi$	RCP	LCP
$\phi = 3\pi/2 + 2m\pi$	LCP	RCP

Sometimes, waveplates described by the second line above are called 3/4 waveplates. For multiple order waveplates, CVI Melles Griot permits the use of either of the above classes of waveplates to satisfy the requirements of a quarter-wave waveplate.

MULTIPLE-ORDER WAVEPLATES

For the full-, half-, and quarter-wave waveplate examples given in standard waveplates, the order of the waveplate is given by the integer m. For $m > 0$, the waveplate is termed a multiple-order waveplate. For $m = 0$, we have a zero order waveplate.

The birefringence of crystal quartz near 500 nm is approximately 0.00925. Consider a 0.5-mm-thick crystal quartz waveplate. A simple calculation shows that this is useful as a quarter waveplate for 500 nm; in fact, it is a 37 $\lambda/4$ waveplate at 500 nm with $m = 18$. Multiple-order waveplates are inexpensive, high-damage-threshold retarders. Further analysis shows that this same 0.5mm plate is a 19 $\lambda/2$ half waveplate at 488.2 nm and a 10λ full-wave waveplate at 466.5 nm. The transmission of this plate between parallel polarizers is shown in Figure 1.53 as a function of wavelength. The retardance of the plate at various key points is also shown. Note how quickly the retardance changes with wavelength. Because of this, multiple-order waveplates are generally useful only at their design wavelength.

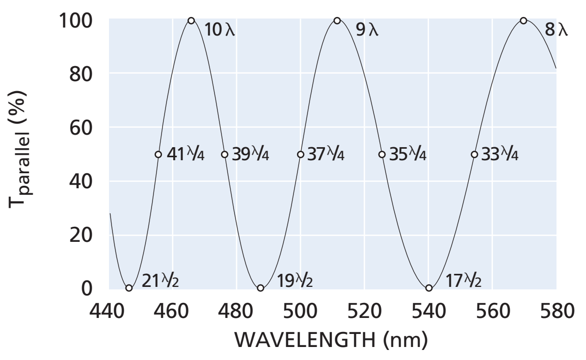

Figure 1.53 **Transmission of a 0.5-mm-thick crystal quartz waveplate between parallel polarizers**

ZERO-ORDER WAVEPLATES

As discussed above, multiple-order waveplates are not useful with tunable or broad bandwidth sources (e.g., femtosecond lasers). A zero-order waveplate can greatly improve the useful bandwidth in a compact, high-damage-threshold device.

As an example, consider the design of a broadband half-wave waveplate centered at 800 nm. Maximum tuning range is obtained if the plate has a single π phase shift at 800 nm. If made from a single plate of crystal quartz, the waveplate would be about 45 μm thick, which is too thin for easy fabrication and handling. The solution is to take two crystal quartz plates differing in thickness by 45 μm and align them with the slow axis of one against the fast axis of the other. The net phase shift of this zero-order waveplate is π. The two plates may be either air-spaced or optically contacted. The transmission of an 800-nm zero-order half-wave waveplate between parallel polarizers is shown in Figure 1.54 using a 0-10% scale. Its extinction is better than 100:1 over a bandwidth of about 95 nm centered at 800 nm.

CVI Melles Griot produces multiple order and zero order crystal quartz waveplates at any wavelength between 193 nm and 2100 nm. Virtually all popular laser wavelengths are kept in stock, and custom wavelength parts are available with short delivery time.

Mica waveplates are an inexpensive zero-order waveplate solution for low-power applications and in detection schemes.

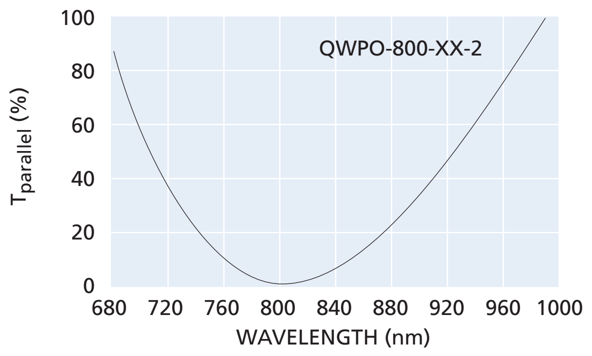

Figure 1.54 **Zero-order crystal quartz half-wave waveplate for 800 nm**

POLYMER WAVEPLATES

Polymer waveplates offer excellent angular field of view since they are true zero-order waveplates. Figure 1.55 compares the change in retardance as function of incidence angle for polymer and quartz waveplates. A polymer waveplate changes by less than 1% over a ±10° incidence angle.

Retardance accuracy with wavelength change is often of key concern. For example, an off-the-shelf diode laser has a center wavelength tolerance of ±10 nm. Changes with temperature and drive conditions cause wavelength shifts which may alter performance. These polymer waveplates maintain excellent waveplate performance even with minor shifts in the source wavelength. The temperature sensitivity of laminated polymer waveplates is about 0.15 nm/°C, allowing operation over moderate temperature ranges without significantly degrading retardance accuracy. A comparison of different waveplate types and their dependence on wavelength is shown in figure 1.56.

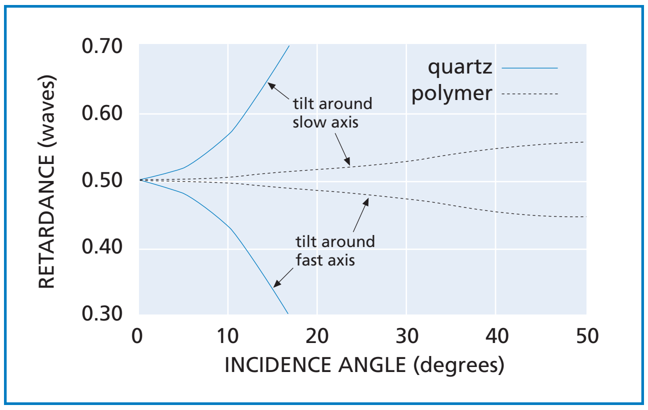

Figure 1.55 **Retardance vs incidence angle for quartz and polymer waveplates**

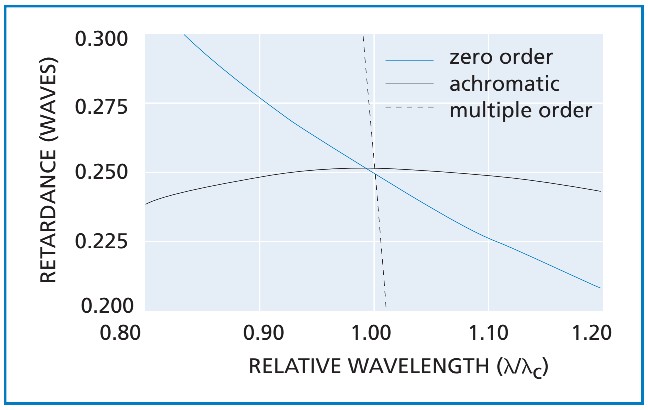

Figure 1.56 **Wavelength performance of common quarter wave retarders.**

ACHROMATIC WAVEPLATES

At 500 nm, a crystal quartz zero-order half-wave waveplate has a retardation tolerance of λ/50 over a bandwidth of about 50 nm. This increases to about 100 nm at a center wavelength of 800 nm. A different design which corrects for dispersion differences over the wavelength range is required for bandwidths up to 300 nm.

If two different materials are used to create a zero-order or low-order waveplate, cancellation can occur between the dispersions of the two materials. Thus, the net birefringent phase shift can be held constant over a much wider range than in waveplates made from one material. ACWP-series achromatic waveplates from CVI Melles Griot (see Figure 1.57) are comprised of crystal quartz and magnesium fluoride to achieve achromatic performance.

Three wavelength ranges are available in both quarter and half wave retardances. Retardation tolerance is better than λ/100 over the entire wavelength range.

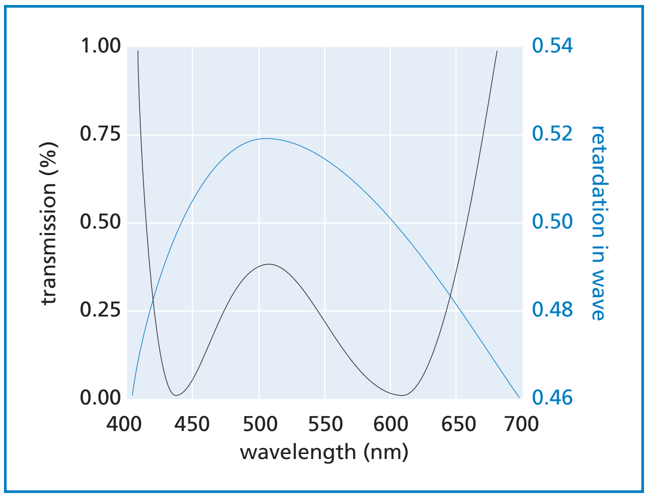

Figure 1.57 **ACWP-400-700-10-2**

For quarter-wave waveplates, perfect retardance is a multiple of 0.25 waves, and transmission through a linear polarizer must be between 33% and 67%. (In all but the shortest wavelength design, quarter-wave retardation tolerance is better than λ/100.) For half-wave waveplates, perfect retardance is 0.5 waves, while perfect transmission through a linear polarizer parallel to the initial polarization state should be zero. A high degree of achromatization is achievable by the dual material design. In addition, we manage low group velocity dispersion for ultrashort pulse applications through the use of thin plates.

DUAL-WAVELENGTH WAVEPLATES

Dual-wavelength waveplates are used in a number of applications. One common application is separation of different wavelengths with a polarizing beamsplitter by rotating the polarization of one wavelength by 90°, and leaving the other unchanged. This frequently occurs in nonlinear doubling or tripling laser sources such as Nd:YAG (1064/532/355/266).

One way to achieve the multiple retardation specifications is through careful selection of multiple-order waveplates which meet both wavelength and retardation conditions. This often results in the selection of a relatively high order waveplate. Therefore, these dual-wavelength waveplates operate best over a narrow bandwidth and temperature range.

Another approach is to combine two quartz waveplates with their optical axes orthogonal to one another, effectively creating a zero-order waveplate. In this configuration, the temperature dependence is a function of the thickness difference between the waveplates, resulting in excellent temperature stability. The retardation of the compound waveplate is also a function of the thickness difference enabling wide bandwidth performance.

Etalons

Etalons are most commonly used as line-narrowing elements in narrow-band laser cavities or as bandwidth-limiting and coarse-tuning elements in broadband and picosecond lasers. Further applications are laser line profile monitoring, diagnosis.

The etalons described in this section are all of the planar Fabry-Perot type. Typical transmission characteristics for this type etalon are shown in Figure 1.58.

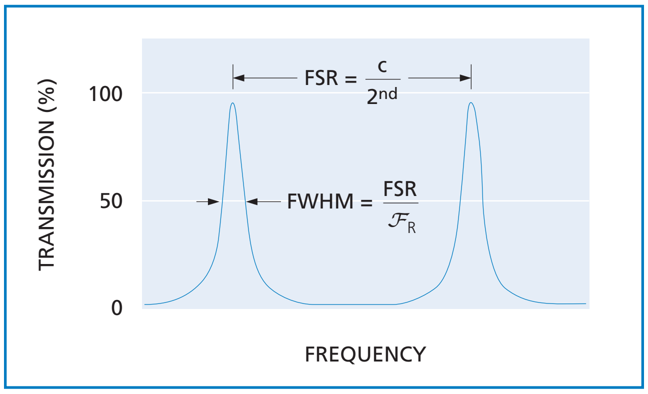

Figure 1.58 **Transmission characteristics of a Fabry-Perot etalon**

Air-Spaced Etalons consist of pairs of very flat plano-plano plates separated by optically contacted spacers. The inner surfaces of the plates are coated with partially reflecting coatings, the outer surfaces are coated with antireflection coatings.

Solid Etalons are made from a single plate with parallel sides. Partially reflecting coatings are then deposited on both sides. The cavity is formed by the plate thickness between the coatings.

Deposited Solid Etalons are a special type of solid etalon in which the cavity is formed by a deposited layer of coating material. The thickness of this deposited layer depends on the free spectral range required and can range from a few nanometers up to 15 micrometers. The cavity is sandwiched between the etalon reflector coatings and the whole assembly is supported on a fused-silica base plate.

Etalon plates need excellent surface flatness and plate parallelism. To avoid peak transmission losses due to scatter or absorption, the optical coatings also have to meet the highest standards.

For a plane wave incident on the etalon, the transmission of the etalon is given by:

$$T = \frac{I_{\text{trans}}}{I_{\text{inc}}} = \frac{1}{1 + \frac{4R}{(1-R)^2}\sin^2(\delta/2)} \tag{1.112}$$

Here, R is the reflectance of each surface; δ is the phase shift

$$\delta = \frac{2\pi}{\lambda}\eta d\cos\theta \tag{1.113}$$

where,

η is the refractive index (e.g., 1 for air-spaced etalons)

d is the etalon spacing or thickness

θ is the angle of incidence

The free spectral range (FSR) of the etalon is given by

$$\begin{aligned} \text{FSR} &= \frac{c}{2\eta d} \text{ in Hz} \\ &= \frac{1}{2\eta d} \text{ in cm}^{-1} \\ &= \frac{\lambda^2}{2\eta d} \text{ in nm} \end{aligned} \tag{1.114}$$

The reflectivity finesse, $\mathcal{F}_R$ is given by

$$\mathcal{F}_R = \frac{\pi\sqrt{R}}{1-R} \tag{1.115}$$

Figure 1.59 shows the reflectivity finesse as a function of the coating reflectivity.

The bandwidth (FWHM) is given by

$$\text{FWHM} = \frac{\text{FSR}}{\mathcal{F}_R} \tag{1.116}$$

However, the above applies to theoretical etalons which are assumed to be perfect. In reality, even the best etalon will show defects that limit theoretically expected performance. Therefore, in a real etalon, the actual finesse will usually be lower than the reflectivity finesse.

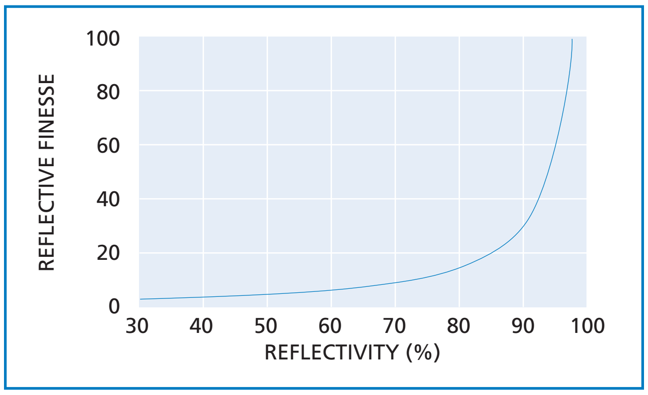

Figure 1.59 **Reflectivity finesse vs. coating reflectance of each surface**

$$\mathcal{F}_{\mathrm{R}} = \frac{\pi\sqrt{R}}{1-R} \tag{1.117}$$

$$\mathcal{F}_{\mathrm{S}} = \frac{M}{2}\,\frac{\lambda}{633\text{ nm}} \tag{1.118}$$

$$\mathcal{F}_{\theta} = \frac{\lambda}{d\tan^2\theta} \tag{1.119}$$

$$\mathcal{F}_{\mathrm{D}} = \frac{\mathrm{CA}^2}{2d\lambda} \tag{1.120}$$

where $\mathcal{F}_{\mathrm{R}}$ is the reflectivity finesse, $\mathcal{F}_{\mathrm{s}}$ is the plate spherical deviation finesse coefficient, $\mathcal{F}_{\theta}$ is the incident beam divergence finesse coefficient, and $\mathcal{F}_{\mathrm{d}}$ is the diffraction-limited finesse coefficient.

The defects that contribute to this reduction are as shown in Figure 1.60 (graphical representations are exagerated for clarification).

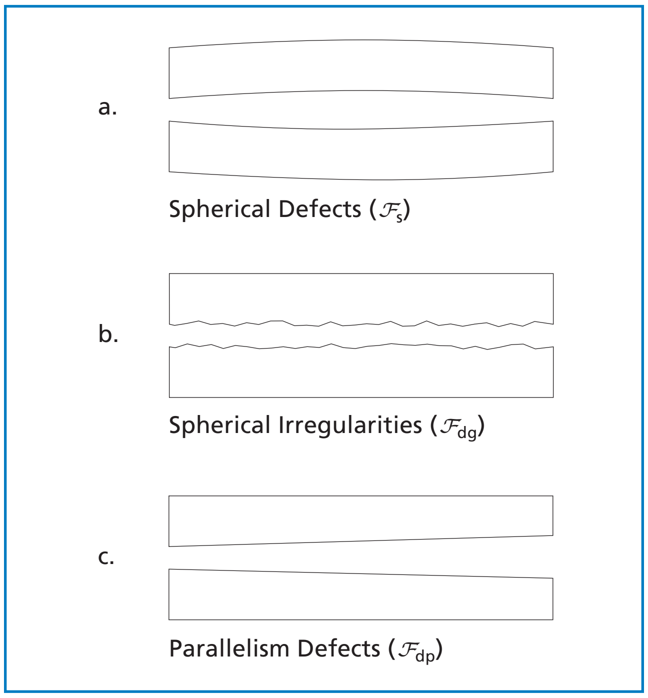
a.
Spherical Defects ($\mathcal{F}_{\mathrm{s}}$)
b.
Spherical Irregularities ($\mathcal{F}_{\mathrm{dg}}$)
c.
Parallelism Defects ($\mathcal{F}_{\mathrm{dp}}$)

Figure 1.60 **Three types of defects contributing to the total defect finesse**

All three types of defects contribute to the total defect finesse $\mathcal{F}_{\mathrm{d}}$:

$$\frac{1}{\mathcal{F}_{\mathrm{d}}{}^2} = \frac{1}{\mathcal{F}_{\mathrm{S}}{}^2} + \frac{1}{\mathcal{F}_{\mathrm{dg}}{}^2} + \frac{1}{\mathcal{F}_{\mathrm{dp}}{}^2} \tag{1.121}$$

The beam divergence also influences the actual finesse of an etalon. Taking into account all these contributions, the effective finesse ($\mathcal{F}_{\mathrm{e}}$) of an etalon (with $\mathcal{F}_{\mathrm{R}}$ being the reflectivity finesse and $\mathcal{F}_{\theta}$ the divergence finesse) is:

$$\frac{1}{\mathcal{F}_{\mathrm{e}}} = \sqrt{\frac{1}{\mathcal{F}_{\mathrm{R}}{}^2} + \frac{1}{\mathcal{F}_{\mathrm{D}}{}^2} + \frac{1}{\mathcal{F}_{\theta}{}^2} + \frac{1}{\mathcal{F}_{\mathrm{S}}{}^2}} \tag{1.122}$$

The effective finesse a user sees when using the etalon depends not only on the absolute clear aperture, but also on the used aperture of the etalon, especially when a high finesse is required.

The examples below show how the effective finesse varies with plate flatness and clear aperture.

Example 1: Air-spaced etalon,

$R = 95\%$ ($\pm 1\%$) at 633 nm

CA = 25 mm

Used aperture = 20 mm

Spacer (air gap) = 1 mm

Spherical / parallelism defects = <λ/20

Plate rms = 0.80 nm

Beam divergence = 0.1 mRad

$\mathcal{F}_{\mathrm{R}} = 61$, $\mathcal{F}_{\mathrm{e}} = 10$

Example 2: Air-spaced etalon,

Same parameters as example 1 except:

Used aperture = 5 mm

$\mathcal{F}_{\mathrm{R}} = 61$, $\mathcal{F}_{\mathrm{e}} = 40\ (\pm 4)$

Example 3: Air-spaced etalon,

Same parameters as example 1 except:

Spherical / parallelism defects = λ/100

Plate rms = 0.40 nm

Beam divergence: 0.1mRad

$\mathcal{F}_{\mathrm{R}} = 61$, $\mathcal{F}_{\mathrm{e}} = 40\ (\pm 8)$

These examples illustrate that, for large-aperture applications, it is important to use very high-quality plates to ensure a high finesse and good transmission values.

TUNING AN ETALON

Etalons can be tuned over a limited range to alter their peak transmission wavelengths. These techniques are:

Angle tuning or tilting the etalon: As the angle of incidence is increased, the center wavelength of the etalon can be tuned down the spectrum.

Temperature tuning: Primarily used for solid etalons, temperature-tuning changes both the actual spacing of the reflective surfaces via expansion and the index of refraction of the material, which changes the optical spacing. The tuning result can be given by

$$\frac{\partial(\mathrm{FSR})}{\partial T} = -(\mathrm{FSR})\left[\frac{1}{\mathrm{n}}\frac{\partial \eta}{\partial T} + \frac{1}{d}\frac{\partial d}{\partial T}\right] \qquad (1.123)$$

Pressure tuning: Air-spaced etalons can be tuned by increasing the pressure in the cavity between the optics, thereby increasing the effective index of refraction, and thus the effective spacing.

The above examples illustrate how critical the optical surface flatness, plate parallelism and surface quality are to the overall performance of an etalon. At CVI Melles Griot we have developed sophisticated software that allows us to simulate all effects that influence the performance of an etalon. To order an etalon, FSR, finesse and used aperture are required.

Ultrafast Theory

The distinguishing aspect of femtosecond laser optics design is the need to control the phase characteristic of the optical system over the requisite wide pulse bandwidth. CVI Melles Griot has made an intensive theoretical study of these effects. Certain coating designs have been modified with control of the phase characteristics in mind. New proprietary designs have been created with desirable characteristics for femtosecond researchers. All optics in this section have been tested by researchers in the field and we are constantly fielding new requests.

Assume that the power, reflectivity, and polarization characteristics of a laser mirror are acceptable over the bandwidth of a femtosecond pulse. This means that, over the entire pulse bandwidth, a cavity mirror may have a reflectivity greater than 99.8%; a 50% beamsplitter may have a fairly constant reflection; a polarizer may maintain its rejection of one polarization with an acceptable transmission of the other. It is not enough, however, to simply preserve the power spectrum $S(\omega) = |E(\omega)|^2$ when dealing with femtosecond pulses. The phase relationship among the Fourier components of the pulse must also be preserved in order that the pulse not be broadened or distorted. What constraint on the performance of a mirror or transmissive optic does this imply?

Consider a general initial pulse shape $E_0(t)$. As a function of its Fourier components, it may be expressed as:

$$E_0(t) = \int E(\omega)^{-i\omega t}\, d\omega \qquad (1.124)$$

Suppose this pulse reflects off of a mirror. For this example, we assume the mirror is "ideal", and use the Fourier transform of its complex amplitude reflectance:

$$r(\omega) = re^{+i\Phi(\omega)} = re^{+i\omega t_d} \qquad (1.125)$$

For this "ideal" mirror, r is a real constant equal to the amplitude reflectivity that is assumed constant over the pulse bandwidth. All phase effects have been assumed to be describable by a single phase shift $\phi(\omega)$ that is linearly proportional to frequency with proportionality constant t_d. The reflected pulse is then:

$$\begin{aligned} E_r(t) &= r\int E(\omega)^{-i\omega(t-t_d)}\, d\omega \\ &= rE(t-t_d) \end{aligned} \qquad (1.126)$$

Thus, provided the phase shift is linear in frequency over the pulse bandwidth, the reflected pulse is scaled by the amplitude reflectance r, and delayed in time by the constant group delay t_d. It is, otherwise, an undistorted replica of the original pulse.

Examined over a large enough bandwidth, no optical system will exhibit the constant group delay over frequency needed for perfect fidelity. In general, the phase shift near some center frequency ω_0 may be expanded in a Taylor series for frequencies near ω_0:

$$\begin{aligned} \Phi(\omega) = \Phi(\omega_0) + \Phi'(\omega_0)(\omega-\omega_0) + \\ \Phi''(\omega_0)(\omega-\omega_0)^2/2! + \\ \Phi'''(\omega_0)(\omega-\omega_0)^3/3!\ \ldots \end{aligned} \qquad (1.127)$$

These derivatives are, respectively, the group delay $\Phi'(\omega_0)$, the group velocity dispersion $\Phi''(\omega_0)$, and the cubic term $\Phi'''(\omega_0)$, evaluated at a center frequency ω_0. This expansion is heuristically useful, in an exactly soluble model, for the propagation of a transform-limited Gaussian pulse. Note, however, that for extremely short pulses the expansion above may be insufficient. A full numerical calculation may have to be performed using the actual phase shift function $\Phi(\omega)$. CVI Melles Griot will be happy to assist those interested in the modeling of real optical elements.

To illustrate pulse distortion due to the dependence of the group delay on frequency, consider what happens when an unchirped, transform-limited Gaussian pulse passes through a medium, or is incident on a mirror whose dominating contribution to phase distortion is non-zero group velocity dispersion. The field envelope of the pulse is assumed to be of the form:

$$E(t) = \exp\left[-\left(2\ \ln 2t^2/\tau_0^2\right)\right] \qquad (1.128)$$

where τ_0 is the initial pulse duration (FWHM of the pulse intensity). Let the pulse enter a medium or reflect off of a mirror with non-zero $\Phi''(\omega)$, measured in fsec^2 radians. (For a continuous medium-like glass, $\Phi''(\omega) = \beta''(\omega) \times z$ where $\beta''(\omega)$ is the group velocity dispersion (GVD) per centimeter of material, and z is the physical path length, in centimeters, traveled through the material.) The Gaussian pulse will be both chirped and temporally broadened by its encounter with group velocity dispersion. The power envelope will remain Gaussian; the result for the broadened FWHM is:

$$\tau_1 = \tau_0\left[1+\left(4\ln 2\Phi''(\omega)/\tau_0^2\right)^2\right]^{1/2} \qquad (1.129)$$

This result, valid only for initially unchirped, transform-limited Gaussian pulses, is nevertheless an excellent model to study the effects of dispersion on pulse propagation. The graphs shown in Figure 1.61 represent the theoretical broadening from dispersion for initial pulse widths ranging from 10 to 100 femtoseconds.

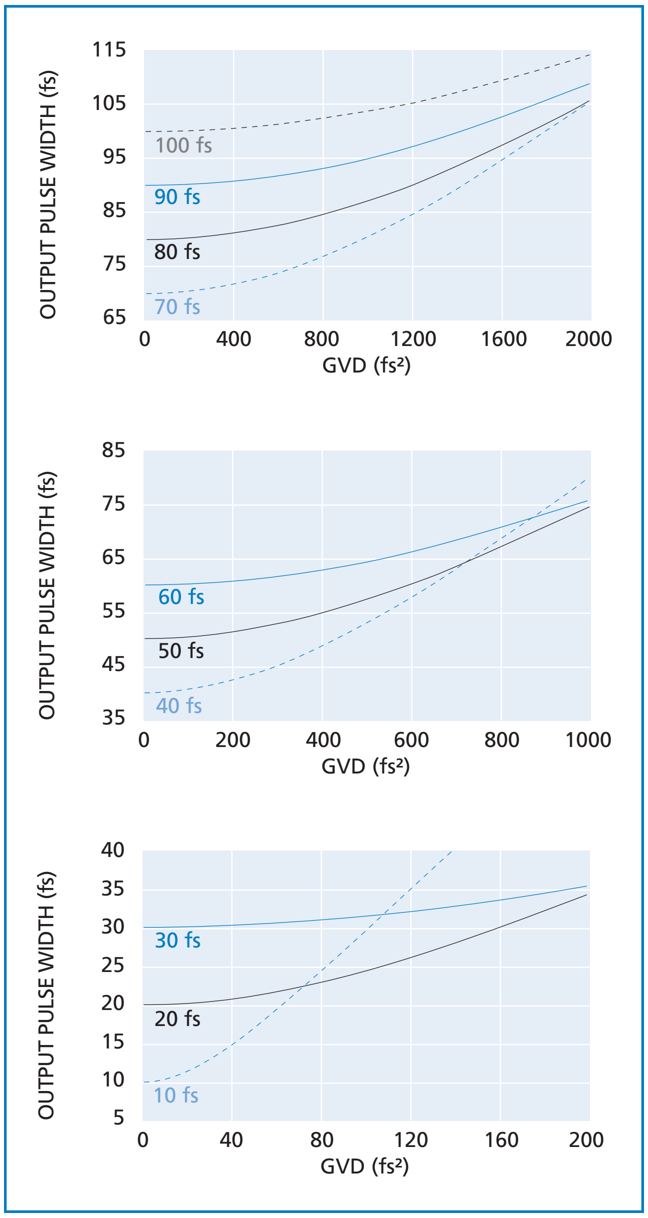

Figure 1.61 **Output pulse width vs. GVD**

GROUP-VELOCITY & CUBIC DISPERSION FOR VARIOUS OPTICAL MATERIALS

Figures 1.62 and 1.63 show the GVD and cubic dispersion respectively for some common used glasses. Some of the glasses can be used in the UV region. They should be useful in estimating material dispersion and pulse distortion effects. Please check these calculations independently before using them in a final design.

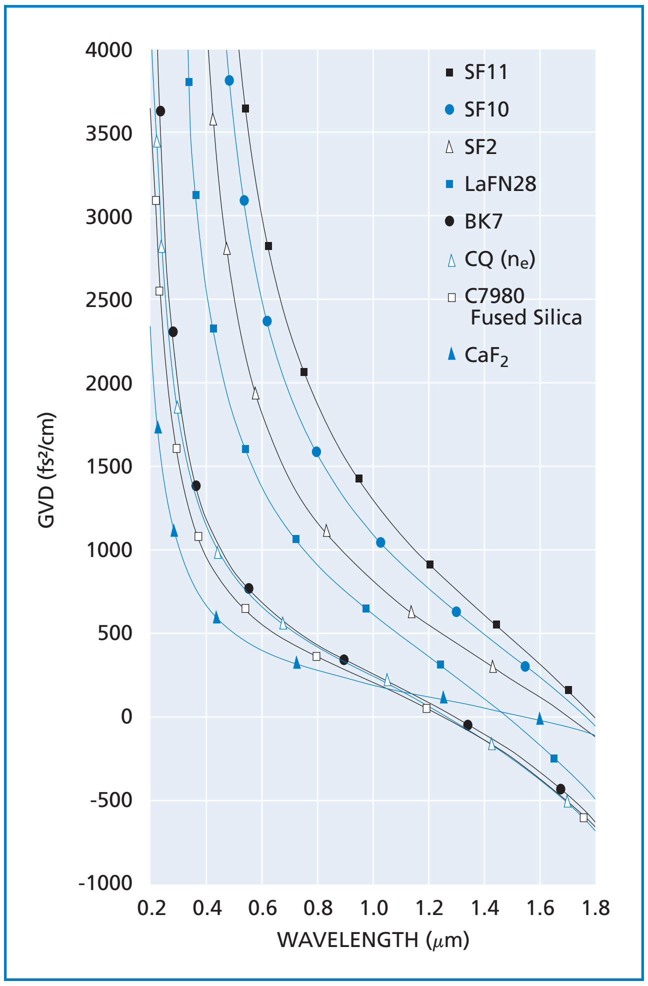

Figure 1.62 **GVD for common glasses**

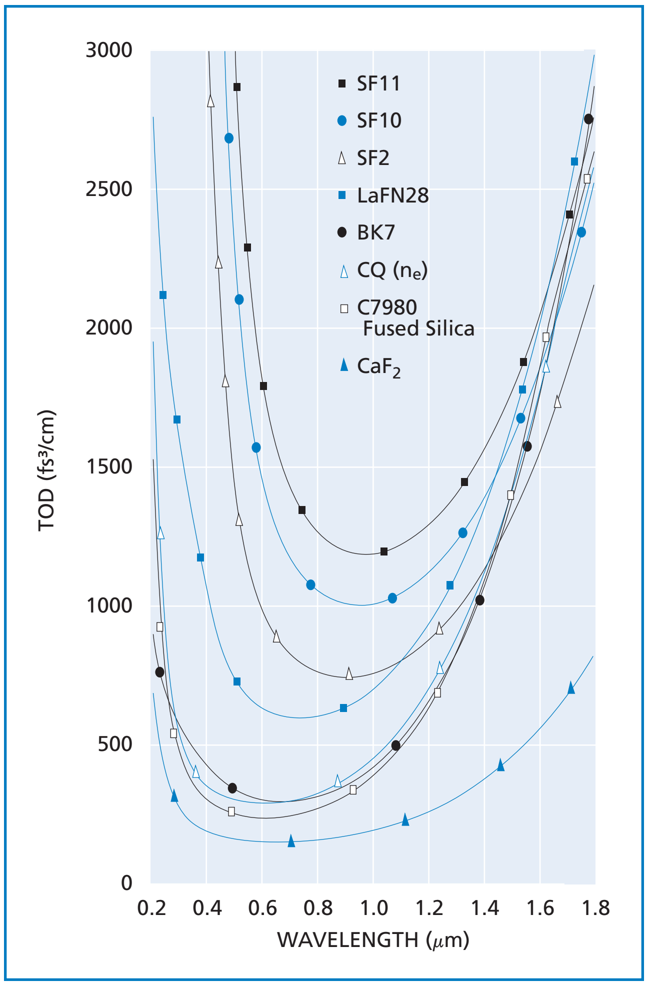

Figure 1.63 **Cubic dispersion for common glasses**

DISPERSIVE PROPERTIES OF MIRRORS

CVI Melles Griot uses three basic designs; TLM1 mirrors for energy fluence greater than 100 mJ/cm^2, TLM2 mirrors for cw oscillators and low-fluence pulses, and TLMB mirrors which are a hybrid of the two. The reflectivity, GVD parameter, and cubic dispersion parameter for TLM2 high reflectors are shown in Figure 1.64. In these examples, the mirrors are centered at 800 nm and are designed for use at normal incidence and at 45 degrees. Note that, at the design wavelength, (a) GVD is zero, (b) the cubic term is minimized, and (c) at 45° incidence, the GVD of the p-polarization component is very sensitive to wavelength, while the GVD for s-polarization component is nearly zero over a broad wavelength range. Thus one should avoid using mirrors at 45° incidence with the p-polarization. On the other hand, at 45° incidence, s-polarization provides very broad bandwidth and minimizes pulse distortion problems and should be used whenever possible.

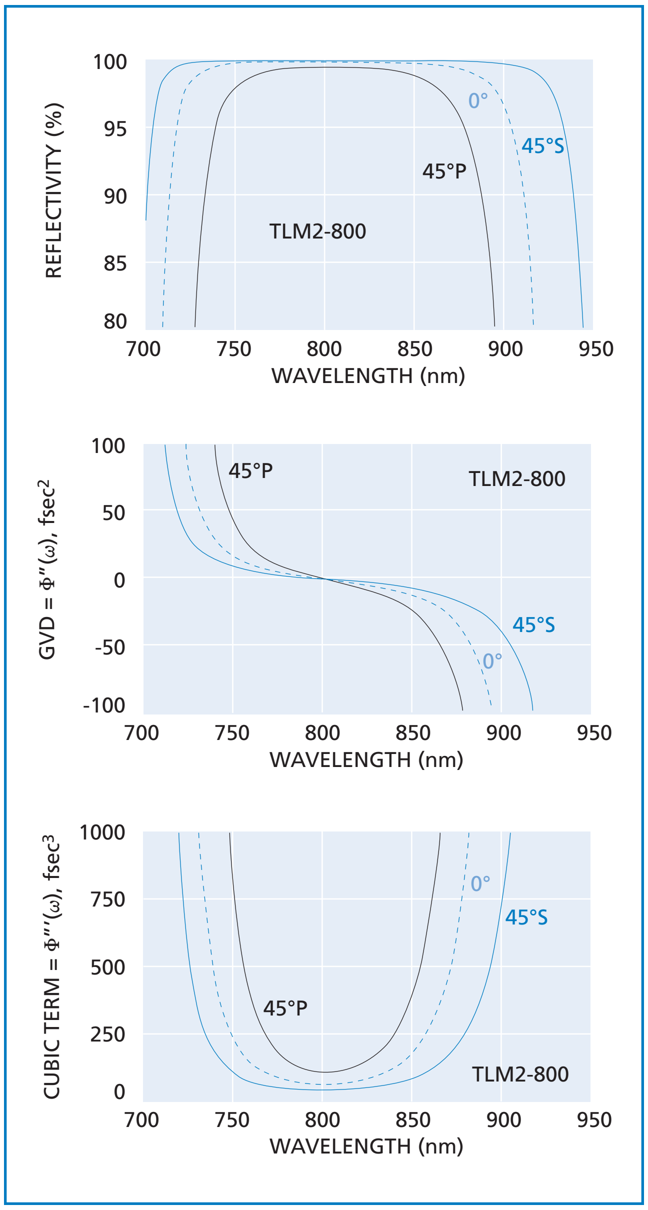

Figure 1.64 **Dispersion and reflectivity for mirrors TLM2-800-0 and TLM2-800-45**

Ti:Sapphire and other femtosecond laser systems need prismless compensation of the built-in positive chirp encountered in the laser optical circuit. This becomes mandatory in industrial and biomedical applications where the laser must provide a compact, stable, and reliable solution.

Negative Group Velocity Dispersion Mirrors (TNM2) meet these needs with off-the-shelf availability, and can be employed both intracavity and extracavity to satisfy chirp control requirements.

In experiments using CVI Melles Griot TNM2 negative group velocity dispersion mirrors, 200-mW, 80-fsec pulses centered at 785 nm were achieved in a simple, prismless, Ti:Sapphire oscillator. The configuration is shown in Figure 1.65.

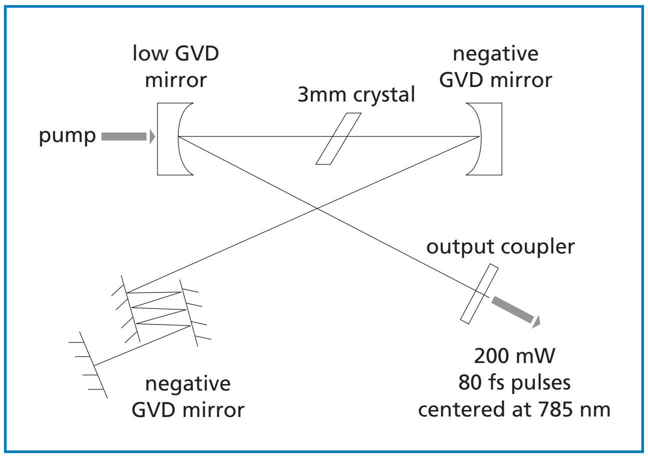

Figure 1.65 **Typical optical setup incorporating low GVD and Negative GVD mirrors in an ultrafast application**

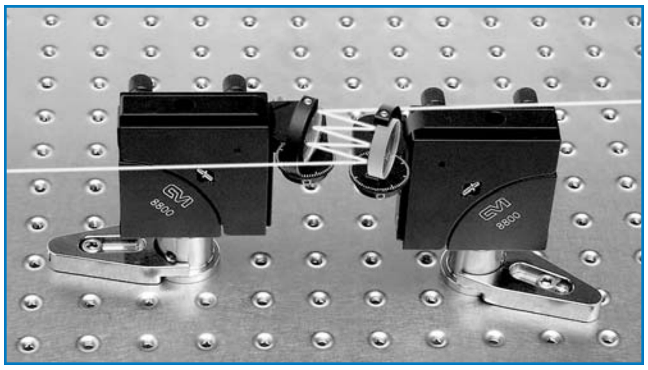

Figure 1.66 **Typical optical set-up of negative GVD mirrors**

OUTPUT COUPLERS AND BEAMSPLITTERS

Output-coupler partial reflectors and beamsplitters behave similarly; however, here is an additional consideration in their analysis. The behavior of the transmitted phase of the coating and the effect of material dispersion within the substrate on the transmitted beam have to be taken into account in a detailed analysis. In general, the coating transmitted phase has similar properties and magnitudes of GVD and cubic to the reflected phase. As usual, centering is important. As a beamsplitter, we recommend the 1.5 mm thick fused silica substrate PW-1006-UV. As an output coupler substrate, we recommend the 3.0 mm thick, 30 minute wedge fused silica substrate IF-1012-UV.

CVI Melles Griot has developed the TFPK Series Broadband Low Dispersion Polarizing Beamsplitters to satisfy requirements for very-high-power, short-pulse lasers. These optics are ideal for intracavity use in femtosecond regenerative amplifiers. The main emphasis is on linear phase characteristics. See Chapter 9 of Lasers, A. E. Siegman (University Science Books, Mill Valley, California, 1986), for a good discussion of linear pulse propagation.

In chirped pulse regenerative amplification, the pulse may have to pass through one or two polarizers twice per round trip. There can be 10 to 20 round trips before the gain is saturated and the pulse is ejected. At this stage the pulse is long (100–1000 psec); however the phase shift at each frequency must still be maintained to minimize the recompressed pulse width. The many round trips of the pulse in the regenerative amplifier put stringent requirements on the phase characteristics of the coatings.

Figure 1.67 shows the power transmission curves for both s- and p-polarization and the transmitted phase characteristics of the p component for a TFPK optimized at 800 nm. (Users may specify any wavelength

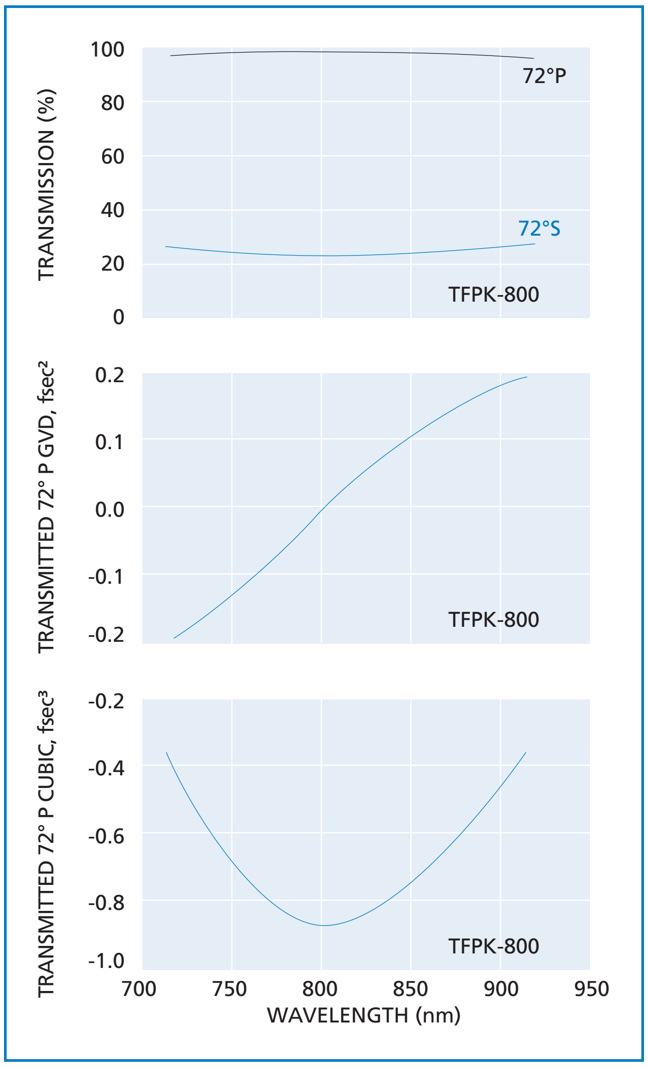

Figure 1.67 **Properties for one coated side of a TFPK polarizing beamsplitter optimized for 800 nm. Both sides are coated for these properties.**

from 250 nm to 1550 nm.) The phase characteristics shown are the GVD and the cubic phase term. Not shown are the reflected phase characteristics for *s*-polarization; they are similar to the *p*-polarization transmission curves, and have the same low nonlinearity and broad bandwidth. Note that both sides of the optic have the coating whose properties are described in Figure 1.67. Therefore, the *s*- and *p*-polarization transmissions per surface should be squared in determining the specifications. The phase characteristics show that in all modes of operation, the TFPK polarizer performance is dominated by the substrate.

There are some subtleties associated with the TFPK. The near 72° angle has to be set properly and optimized. Some thought has to be given to mechanical clearances of the laser beam at such a steep incidence angle. The reflectivity for *s*-polarization is limited to 75%. Variant designs can increase this at a slight loss in bandwidth, increase in incidence angle, and increase in insertion loss for the transmitted *p*-polarized component.

The FABS autocorrelator beamsplitters from CVI Melles Griot are broadband, 50% all-dielectric beamsplitters. They are useful in many types of pump-probe experiments and in the construction of antiresonant ring configurations. They are essentially lossless and extremely durable. Both have advantages over partially reflecting metal coatings.

Power transmission curves for the *s*- and *p*-polarized versions of the FABS, along with the corresponding reflected phase characteristics for beamsplitters optimized at 800 nm, are shown in Figures 1.68 and 1.69. The linear pulse propagation properties of these beamsplitters are dominated by the substrate material dispersion. As with virtually all dielectric coated optics, the *s*-polarized version is broader than *p*-polarized version. CVI Melles Griot can produce FABS in other than 50:50 with excellent phase characteristics.

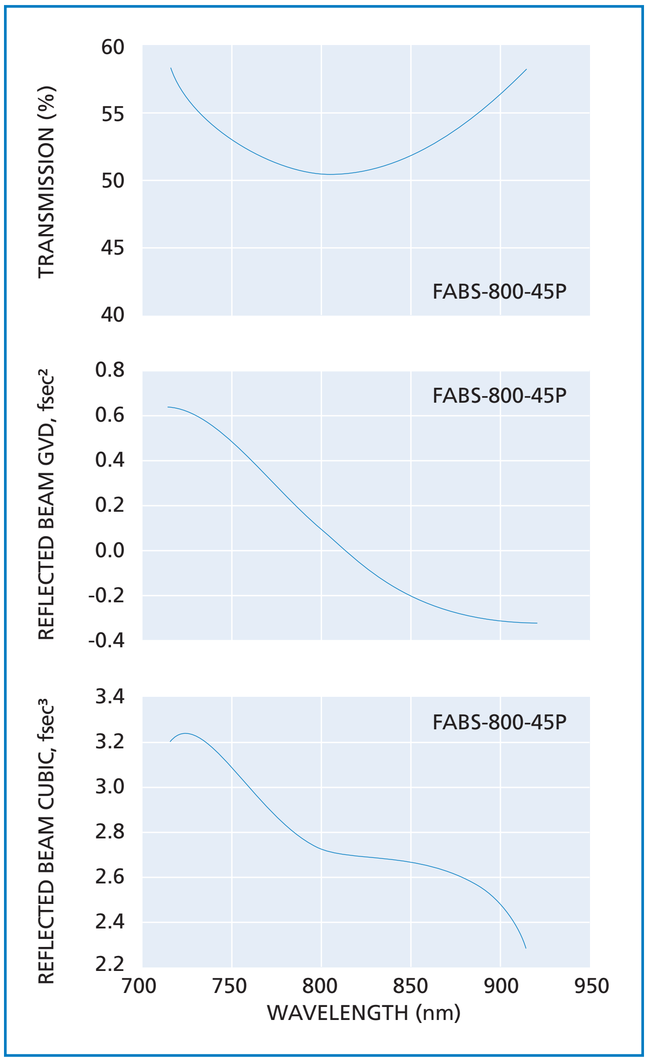

Figure 1.68 **Transmission characteristics for FABS series polarizers with *p*-polarized light**

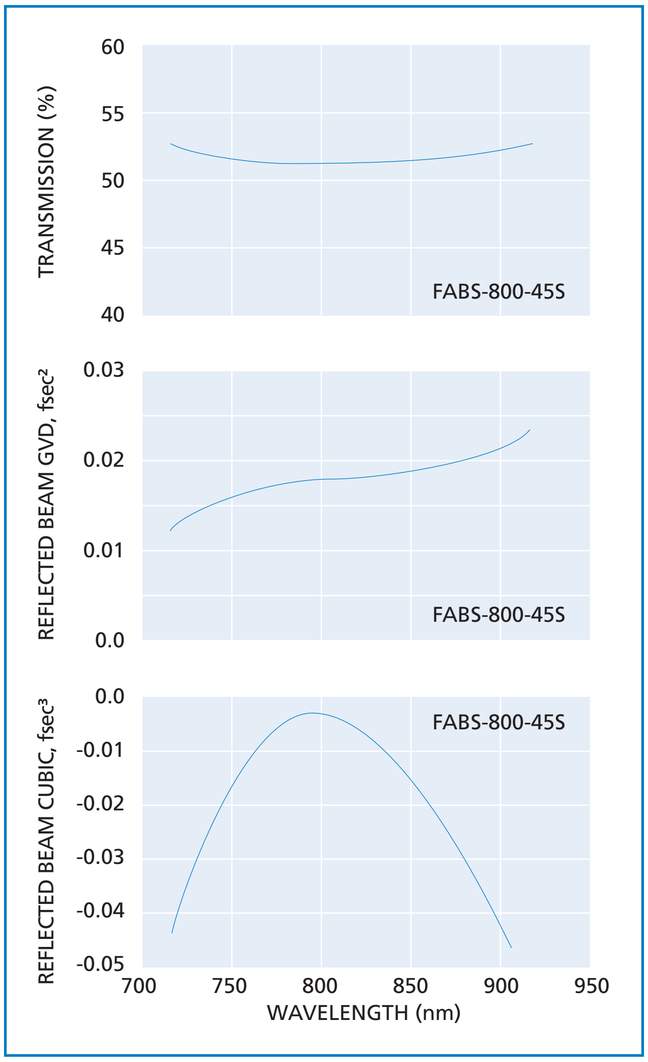

Figure 1.69 **Transmission characteristics for FABS series polarizers with *s*-polarized light**

ANTIREFLECTION COATINGS

All CVI Melles Griot antireflection coating designs work well in femtosecond operation as the forward-going phasor is the dominant contribution to the phase shift; the AR coating is very thin and simply "fixes" the small Fresnel reflection of the substrate.

PRISMS

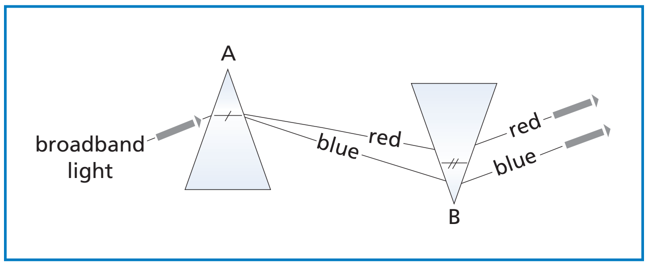

Figure 1.70 **Brewster prism**

Very-high-quality isosceles Brewster's angle prisms for intra and extra-cavity use are available from CVI Melles Griot. The design of these prisms satisfies the condition of minimum loss due to entrance and exit at Brewster's angle. To calculate GVD at Brewsters angle, refer to Figure 1.70 and use the following equation:

$$\mathrm{GVD} = \frac{d^2}{d\omega^2}\psi\Big|_{\omega-\omega_l} \approx \frac{\lambda_l^3}{2\pi c^2}\left[L\,\frac{d^2\eta}{d\lambda^2}\Big|_{\lambda_l} - 4l\left(\frac{d\eta}{d\lambda}\Big|_{\lambda_l}\right)^2\right] \qquad (1.130)$$

where

η = refractive index of the prisms (assuming the same material)
l = tip to tip distance ($\overline{AB}$)
L = total avg. glass path
ψ = spectral phase of the electric field
$\omega_l \lambda_l = 2\pi_c$ (assumes Brewster prism at minimum deviation).

For more on the Ultrafast phenomena, see J.C. Diels and W. Rudolph, Ultrashort Laser Pulse Phenomena, Academic Press, 1996.

CVI MELLES GRIOT. YOUR SOURCE FOR

BUILD-YOUR-OWN, RAPID DELIVERY SOLUTIONS.

CVI Melles Griot offers two-week delivery of optics at catalog prices.

- 4000 uncoated substrates
- 2500 coating options
- A wide variety of in-stock substrates available (curves, flats, prisms, waveplates)
- Coatings from 193 nm to 2300 nm
- Your order ships in two weeks or less

How can we help you make your project a success?

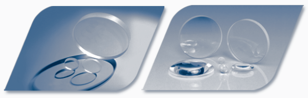

Lasers | Lenses | Mirrors | Assemblies | Windows | Shutters | Waveplates | Mounts

WELDING SECRETS

SECOND EDITION

by Hal Wilson

WELDING SECRETS

Second Edition

A welding guide for the self taught welder,
as well as the more experienced welder.

by

HAL WILSON

© Copyright, 1990 by the Flyco Machine Company

All rights reserved. No part of this book may be reproduced by any means without the written permission of the Flyco Machine Company.

Published by

FLYCO MACHINE COMPANY
3712 Ringgold Road #155
Chattanooga, TN 37412-1638

Web Site:
http://www.entermall.com/weldingsecrets

Printed in the United States of America

Table of Contents

How To Remove A Bad Bearing Race Or Cup From A Hole 1
How To Build Up A Worn Shaft 1
How To Burn A Nut Off Of A Bolt And Save The Threads On The Bolt 3
How To Expand A Nut 3
How To Shrink A Nut 3
Camouflage 4
How To Remove A Broken-Off Bolt 4
Case Hardening 5
How To Square A Frame By Welding Or Peening 5
How To Make A Warped Steel Frame Lie Flat 6
How To Weld A Screen Wire To A Steel Frame 8
How To Make A Smooth Cut With A Cutting Torch 9
How To Do Overhead Welding 9
How To Weld Thick To Thin 9
How To Make A Smooth Bead 10
Vertical Weld 10
How To Flux Lead 10
Bulge 10
How To Burn A Weld And Save Both Pieces 11
Big John 12
Notch Effect 13
How To Remove A Broken-Off Tap 13
Built In Stresses 14
Crystals 14
How To Burn Through A Large Steel Shaft 15
My First Commercial Weld 15
Cast Iron Welded With Nickel Electrodes 16
Straighten 6 Inch Strip 16
Clevis On Channel Iron 17
How To Temper A Chisel 18
How To Make A Circle Lie Flat 18
How To Remove A Stuck Sleeve From A Hole 19
How To Shrink A Steel Pulley 19
Reinforce Inside Of Tubing Or Pipe 19
How To Repair A Crack In A Piece Of Tubing 20
Weld Inserted Pipe 21
Another Method For Welding Pipe 22
Build Up Inside Of Tubing 23

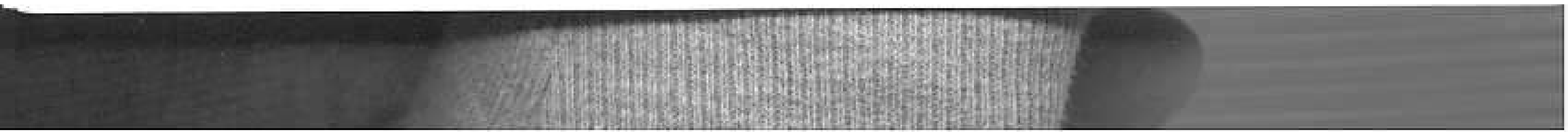

Test Welds 24
Flying Coaster 25
Roller Coaster 27
Cracked Center 28
Peening Test 28
Cross Brace 29
Test Welds On Low Carbon Steel 30
"Direction Of Travel" and "Back-Stepping" 33
Why A Weld Bead Shrinks More Where It Stops Than It Does Where It Starts 34
Confined Expansion 34
The Exception To The Rule 35
Magnaflux Testing For Cracks 36
Water Tank 36
Back Stepping 37
Broken Trailer Frame 37
Air Lifts 37
Subjects Of Non-Welding 39
Cover A Padded Bar With Upholstery 40
Hooks Or Eyes On Spring 41
Make A Coil Spring 42
Cooling Fin 42
Shaft Will Break At Shoulder 43
How To Temper Small Parts 43
Rod Ends 44
Find Diameter Of Numbered Machine Screws 44
Which Way To Turn A Turnbuckle 45
Dish In Wheel 45
Blowing Grip On Bar 45
How To Find A Tap Drill Size Without A Chart 46
Finding Half Of A Mixed Common Fraction 47
Simple Hardness Tester 49
Fusibility: The Ease Of Melting 50
Heat Of Fusion Of Metals 51
Thermal Conductivity Of Metals 52
Thermal Expansion 53
Typical Ferrous Materials 54
Uses For Steel By Carbon Content 55
AWS Electrode Classification System 57
Temper Colors 58
Melting Points Of Some Elements 58

Foreword

I have seen a lot of welds in my life. I've seen them on new products and on repair jobs. The variety of welds I've seen lately is enough to convince me that some people **still** don't know where **not** to weld.

In this book I've tried to help those of you who will go on welding a long time after I am gone. I've tried to illustrate as clearly as I know how — in words and in pictures — some of the places that should **not** be welded.

I've also included some tips that will help you do a better, safer job in places that should be welded.

I hope this book will help you be a better welder. The information in it — gathered the hard way over 49 years — has certainly helped me.

Hal Wilson

NOTE

This publication is designed to provide the author's opinion in regard to the subject matter covered.

It is sold with the understanding that the Publisher or Author does not assume responsibility or liability for any applications or installations produced from the design, products, processes, techniques or data set forth in this booklet.

How To Remove A Bad Bearing Race Or Cup From A Hole

If the race or cup is bad, it will be discarded as scrap. Take your arc welder and carefully run a bead around the center of the race, keeping it in the center to prevent drifting off to the other surface. After cooling, it will have loosened enough to fall out.

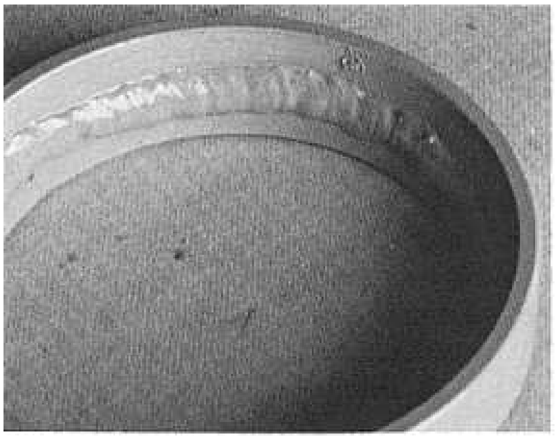

Fig. 1

The above race was 4 7/16 inches in diameter before it was welded. After it cooled it was 7 thousandths of an inch smaller in diameter.

Replace New Race

An easy way to replace the bad race with a new one is to place the new race in the freezing compartment of a refrigerator and freeze it before inserting it. Handle it with gloves.

Occasionally, a race will turn in the hole and wear the hole out too big. If an internal knurling tool is unavailable, take a center punch and a hammer and make center punch marks all the way around the entire surface that is loose. The amount of wear or looseness determines how hard you must hit the center punch. When a piece of mild steel is hit hard enough to make an indentation, the surrounding metal swells outwardly.

How To Build Up A Worn Shaft

First, turn the worn part off in a lathe. Then, preheat it with a torch to about 250°. Have the shaft in a set of rollers so you can turn it from side to side. Weld a bead on the top side. Turn the shaft a half-turn so the new bead is on the bottom. Then weld a bead on the new top side exactly opposite the first bead. Turn the top of the shaft slightly toward you to make the side of your second bead level. (See Figure 2)

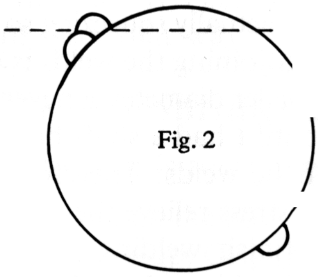

Fig. 2

Then run another bead overlapping to the center of the second bead. Turn the shaft back over to the first bead and run another bead overlapping the center of the first bead.

You don't need to knock the flux off. After the second and third beads, turn the shaft after each bead to the other side. On the last two beads, weave the rod from side to side to overlap the first and last beads on that side.

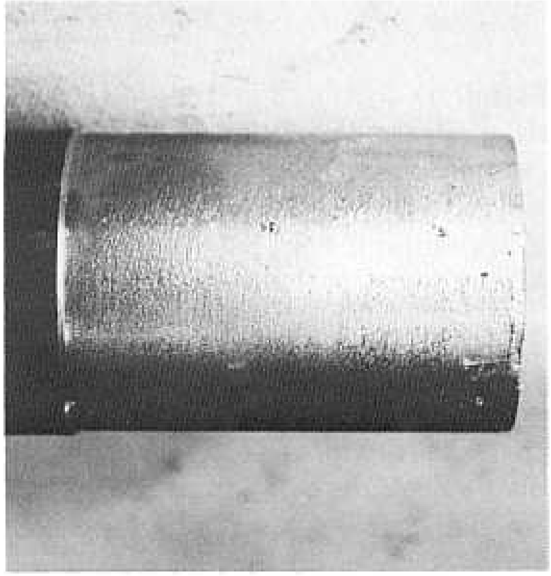

Fig. 3
Shaft turned down ready to be built up.

Fig. 4 shows a built-up shaft. It is hard to believe how much tension is on these welds. They are actually squeezing so hard that the shaft adjoining the welds is compressed to a smaller diameter. I never gave it a thought until I had a shaft break off at the end of the welds. Therefore, you should always stress relieve the welds. As soon as you finish welding, take your acetylene torch and heat the welds to a red heat. Turn the shaft and heat evenly. Hold the heat on it and give it time to get red to the center of the shaft.

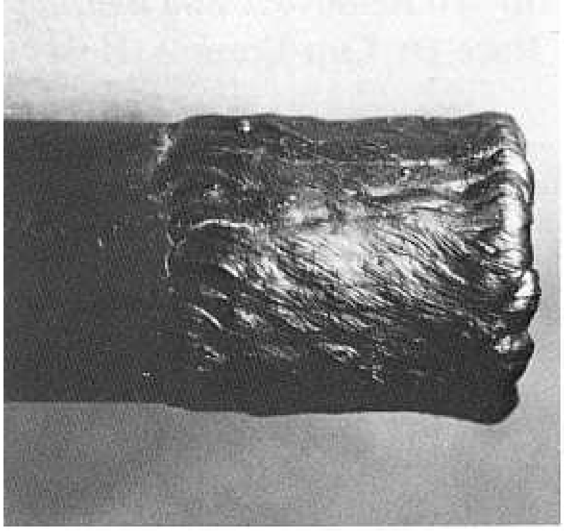

Fig. 4

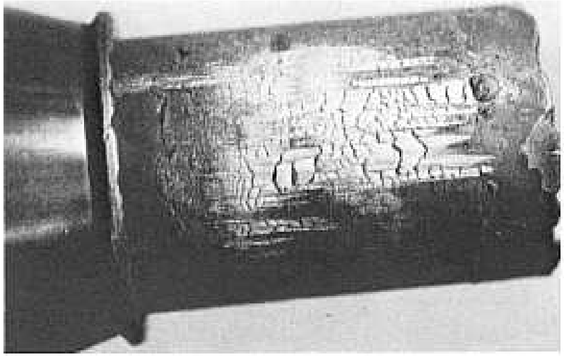

Fig. 5

To prove a point, I turned the welds down on a lathe, to the original size, freely slipped the bearing race on to the other end of the shaft, stress relieved the end that was welded, and let it cool. When it was cold, it was 3 thousandths larger in diameter than it was before I stress relieved it. The bearing race would not come off. I pressed it part of the way off. You can see where I stopped. Then pressed it back on.

How To Burn A Nut Off Of A Bolt And Save The Threads On The Bolt

It's not as hard as it sounds.

You can't burn the threads on the bolt unless you get them red hot. Place the bolt and nut in a horizontal position. Using a small cutting tip, hold the tip parallel to the bolt. (See Below) Start on one of the points and burn through the nut, cutting parallel to the bolt until you get near the bottom. Then, tilt the end of the cutting tip downward and move rapidly through the slot with your stream of oxygen, washing the threads of the nut out of the threads of the bolt.

You can use the same procedure to remove a piece of broken off pipe from a fitting (steel pipe only). Concentrate the heat on the broken off pipe.

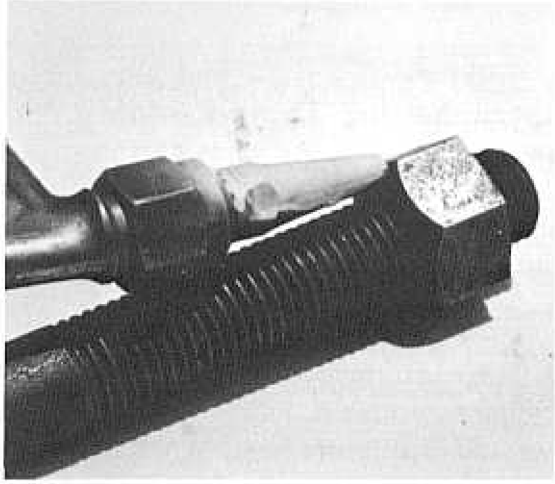

Fig. 6

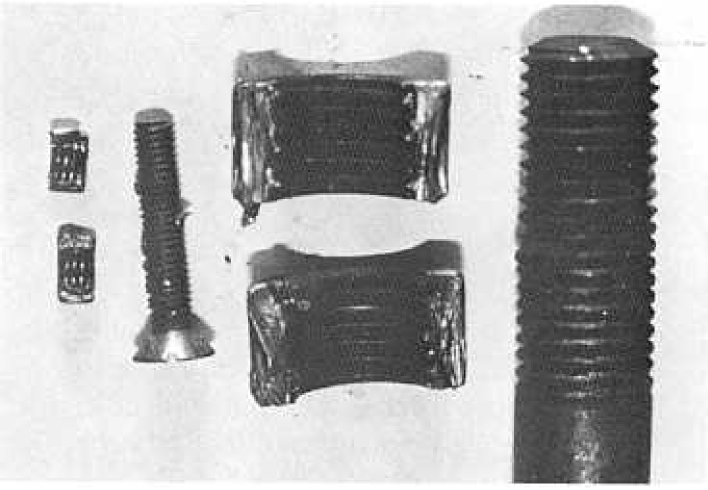

Fig. 7

How To Expand A Nut

To make a nut that is too tight, loose enough to turn with your fingers, do this:

Screw the nut on a bolt. Lay the nut on an anvil and, with a very small ball-peen hammer, peen the flat sides very, very lightly with the ball end. Be extremely careful not to hit it hard enough to make a dent in the flat side of the nut. You can hit it fast, but not hard. Raise the hammer not more than two inches above the nut. If necessary, you may peen all sides that way.

How To Shrink A Nut

When the nut is too loose, do this:

With the nut on the bolt, place the nut in a vise flat side up. Grip the two points of the nut in the jaws of the vise very lightly. Then heat the flat side on top red hot. When it turns black, chill it in water and check it for fit. If necessary, you can heat all six sides this way, one at a time.

Camouflage

The nut below appears to be welded onto the bolt. As you can see, it is not. The nut was welded onto another bolt, cut off, and then drilled to the original "tap drill" size. Then the tap was screwed thru from the back side to tap the weld. This is for larger bolts, the smaller bolts twist off too easily. You can also drill a hole through the nut and bolt, drive a rolled pin through it, putty the ends and paint over it. Occasionally, this procedure may deter a thief.

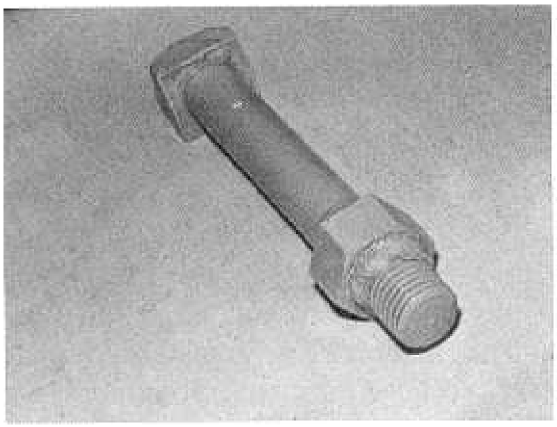

Fig. 8

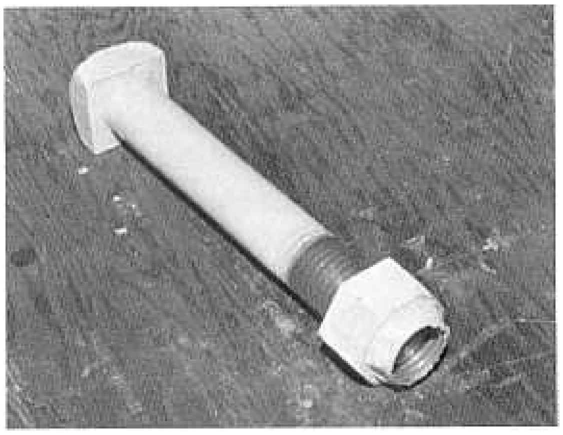

Fig. 9

How To Remove A Broken-Off Bolt

There are several ways to remove a broken-off bolt. "Ridgid" has a very good screw extractor set #10.

Another way is to grind and break off the bent end of an Allen wrench, then grind across the end until it is square. The points should be sharp. Drill a hole in the broken-off bolt the size of the diameter of the flats on the Allen wrench. Then drive the Allen wrench into the hole. The points will cut the metal out and make a hexagonal hole.

These two methods work well — but if the bolt is steel, you can't beat welding a nut on it (See "How To Remove A Broken-Off Tap," page 13).

For example, our shop was once faced with 16 flat head screws, 1/4 by 3/8 — all stuck. The head had a hole for an Allen wrench. We bent an Allen wrench and stripped the hole in another screw. Then, we tried penetrating oil. We hit them with a punch.

Nothing worked — and only three threads holding. Then we welded a nut on it, using the vise grips as shown in "How To Weld A Screen Wire To A Steel Frame." See page 8) It's the only thing that worked.

The reason this method works so well is that when the bolt is welded, the heat radiates through it. Filling the hole in the nut adds still more heat. When the bolt expands and meets resistance from the hole, something has to give. Either the bolt is compressed or the hole is stretched. Either way, the bolt comes out when it cools.

Shown is one of the 1/4 by 3/8 inch flat head screws with a nut welded to it.

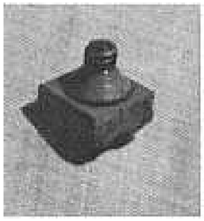

Fig. 10

Case Hardening

Caution: When cutting with a cutting torch, be sure that you do not run out of oxygen. If your oxygen pressure gets too low, a feather will appear on your flame, showing an excess of acetylene. The melted metal on each side of your cut will absorb carbon from the excess acetylene. As you know, when cutting steel with a cutting torch, a very small area gets red hot. The surrounding cold steel absorbs the heat so fast, it is almost like dunking it in water. The result is, the steel that was melted, and absorbed the carbon is case hardened.

I remember 20 years ago, I was cutting some flanges for hydraulic cylinders out of 1/2" plate with a circle burner. My teen-aged son said, "Let me do that dad." I handed him the torch. Later he came to me and said there was something wrong. He was running out of oxygen. We changed oxygen tanks. When I started to machine the 3 pieces that were case hardened, instead of the lathe tool cutting the flange, the hard place cut a groove in the lathe tool. I heated them red hot and covered them with a bucket to retard the cooling. When they were cold, they were soft enough to machine.

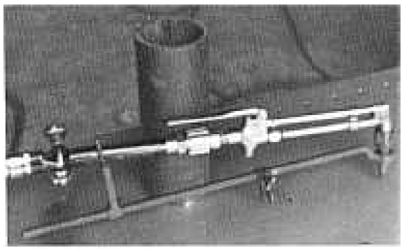

Fig. 11

How To Square A Frame By Welding Or Peening

Every weld in a frame pulls the frame toward being square or out of square. The direction of travel causes it. Every weld pulls the metal closer together at the point where you stop welding than where you started. The first welds on the four corners of a frame should be made in the same direction — from inside to outside or vice versa. After you weld the corners, check the frame for square by measuring diagonally across the corners.

If two corners are farther apart, weld a bead on the other side of the frame from the inside to the outside on both corners. If the frame does not square with that, weld another bead on the other two corners, traveling from outside to inside. In other words, **open** an angle by traveling from inside to outside. **Close** an angle by traveling outside to inside.

You can also open an angle by peening on the inside edge. Close it by peening on the outside edge.

How To Flatten A Warped Steel Frame

Flattening a warped steel frame is simpler than it sounds.

When an angle iron frame is warped — if two diagonal corners are raised when the frame is placed on a flat surface, causing the frame to rock — here's how to solve the problem:

Lay the frame on a flat surface with one leg of the angles pointing down and the other ones pointing toward the inside of the frame. (See Figure 12)

Place an adjustable wrench on the end piece of angle on the leg pointing to the inside of the frame. Place it about one inch from the corner that is higher.

At this point, you have to do just the opposite of what it looks like you should do. You push **down**, not up, on the wrench handle until you can feel it give. (See Figure 13)

Then, do the same thing on the opposite high corner. If the frame still refuses to lie flat, place the tool on the angles on each side as shown in Fig. 15 and 16, and push down. The frame is turned over in Fig. 17. If necessary you may pull inward on the high corners and outward on the low corners on the legs that point upward.

Some angle irons are too large to be bent with an adjustable wrench. For these, cut a slot in a piece of heavy steel and weld a long handle on it. Make the handle strong enough to bend the angle iron leg. This procedure works on many different shapes.

When you bend a piece of cold mild steel, it always springs back a little. If you bend it more, it doesn't spring back any more than it did the first time.

When you bend the leg on an angle to flatten a frame, the leg springs back a little. That puts a torsion stress on the angle iron causing it to push down on the high corner. It isn't necessary to bend the leg too much. You have 16 points on an angle iron frame, (four on each corner) to flatten it with.

The iron-worker who taught me how to do this also told me a story:

A new man came to work at the shop where my friend worked one day. The new man built a gate out of very heavy, closely woven wire. When he finished it, it was warped. He worked all morning, hammering and twisting. He even blocked up the low corners and pressed down on the high corners. It flopped — and warped the other way. What had been the high corners were now the low corners.

My friend watched out of the corner of his eye until the new man took off for lunch.

"Then I went over there and straightened it out with a pair of pliers," said my friend.

When the new man came back from lunch, he checked all four corners. They were all flat on the floor. He turned the gate over and checked it. He stared at it, scratched his head, looked all around the shop and demanded, "Who straightened this for me?"

He never did find out.

My friend's secret? He put stress on the wires near the frame with his pliers. Just the opposite from the way it looks.

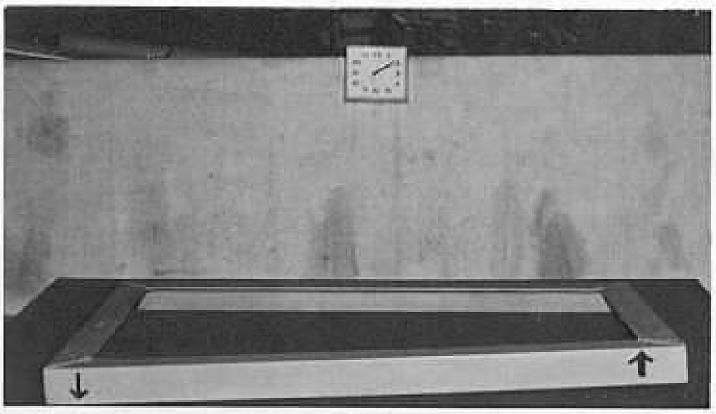

Fig. 12

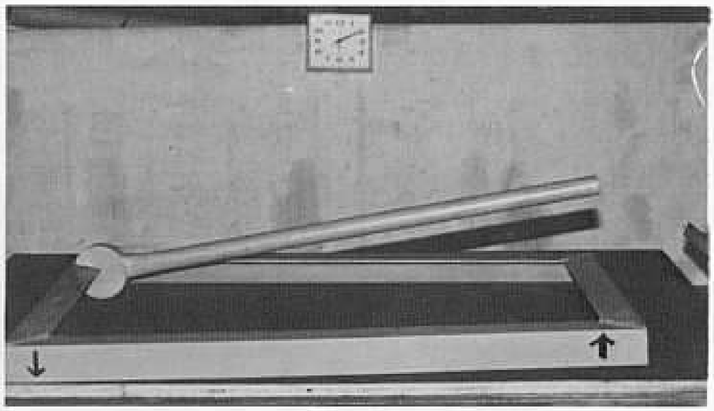

Fig. 13

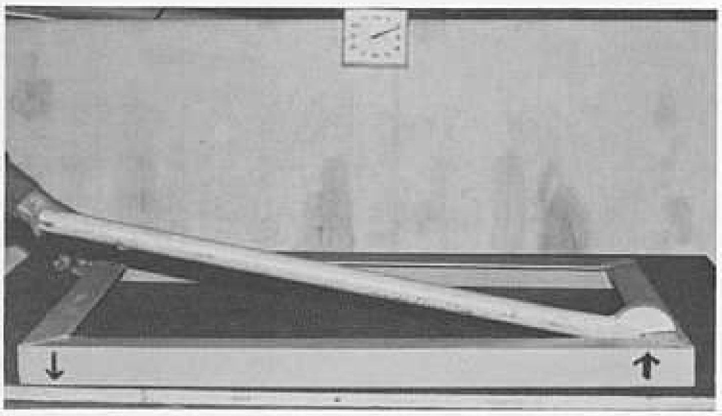

Fig. 14

Notice the clock. I placed the tool, pushed down on the handle, left the tool in place and then made the picture, on each one of these. I straightened one without making pictures in less than one minute.

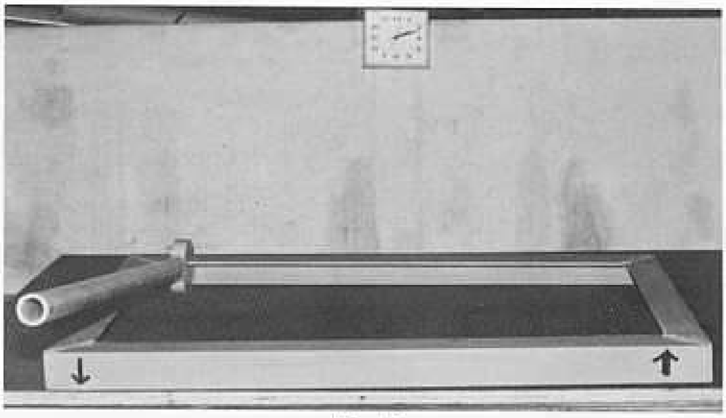

Fig. 15

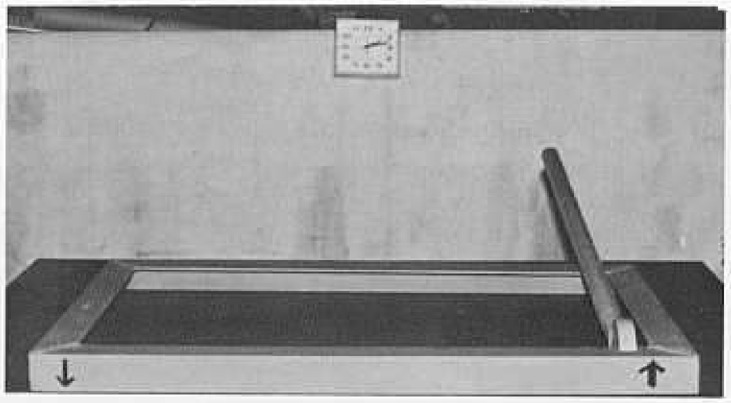

Fig. 16

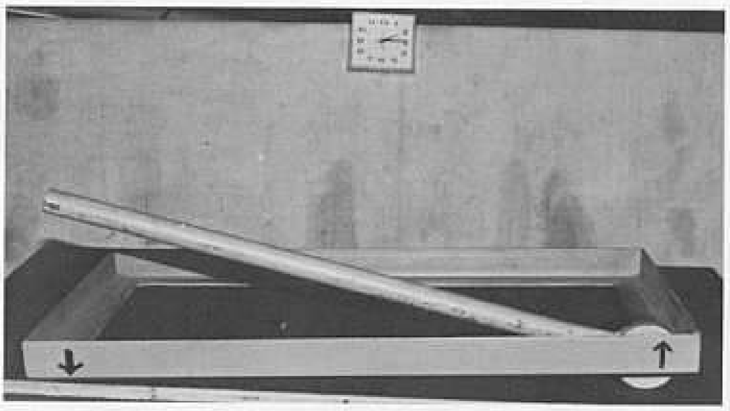

Fig. 17

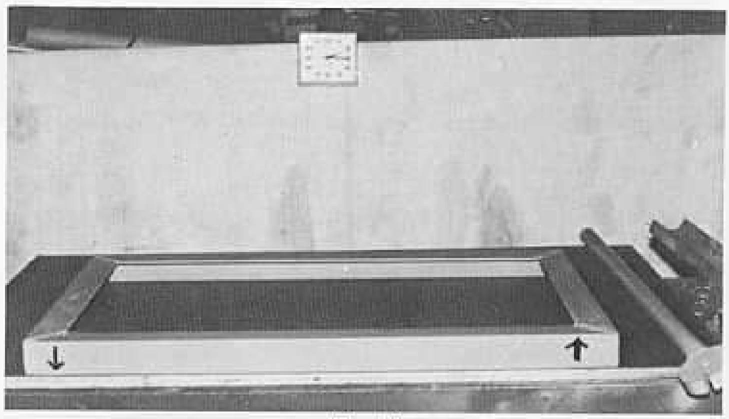

Fig. 18

How To Weld A Screen Wire To A Steel Frame

If you're having trouble welding a screen wire or light expanded metal to a frame, try using washers.

Place a flat washer on the wire where you want to weld it to the frame. Clamp it down with a pair of vise-grips. (See Figure 19) The big, partly cut-off washer is added on, lightly tacked.

Weld through the hole in the washer and into the frame. Fill the holes in the washers with weld metal. (See Figure 20) If you want a really neat job, cut a strip of 1/8 x 3/4 inch to the correct length and drill 5/16 inch holes in the center of it. Clamp it in place and weld through the holes. (See Figures 21 and 22)

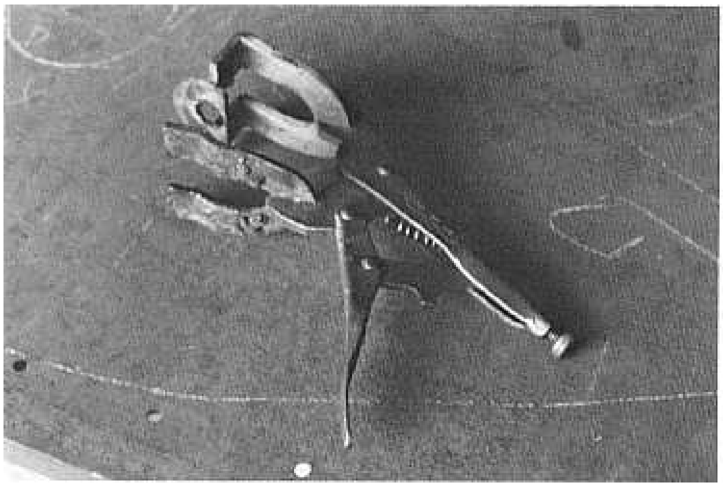

Fig. 19

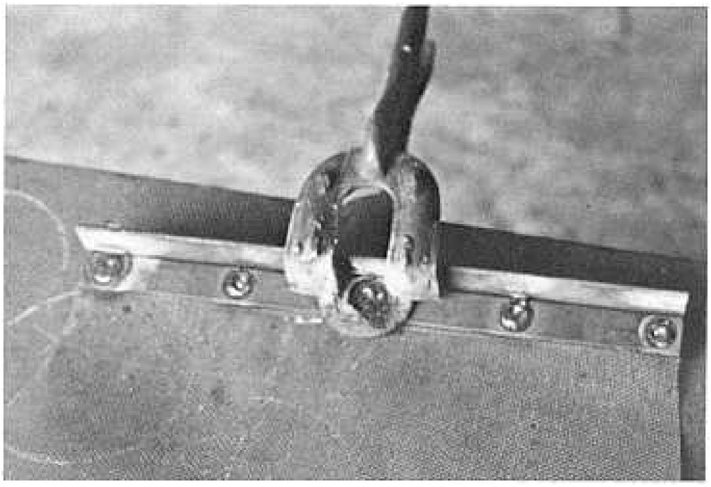

Fig. 21

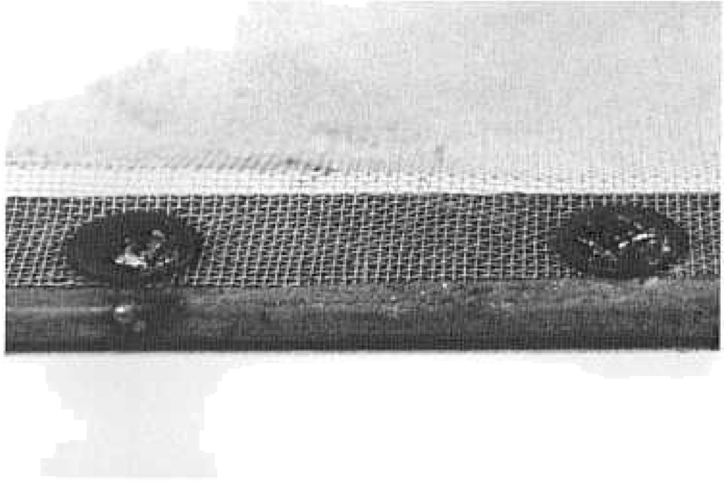

Fig. 20

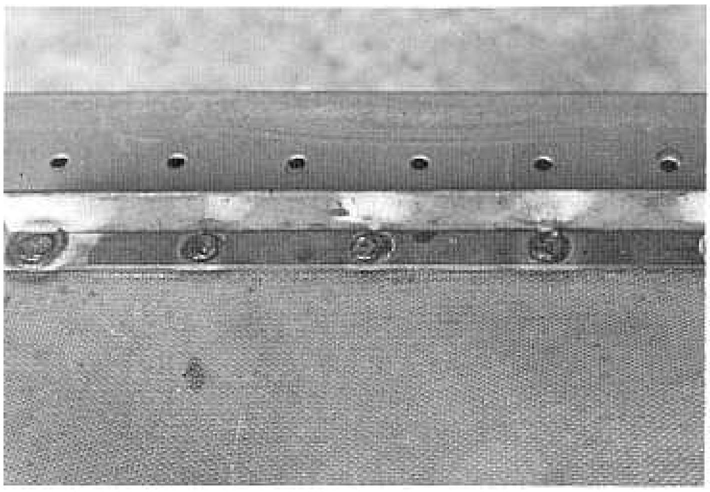

Fig. 22

How To Make A Smooth Cut With A Cutting Torch

For a machine-like cut, use a straight edge. Take a 1/2 by 2-inch piece about 12 to 14 inches long and a 1/8 by 3-inch piece the same length. Weld or tack the 1/8-inch piece on top of the 1/2-inch piece, letting the smaller one overhang a half-inch on each side. File the 1/8-inch edges smooth. Place the file flat on the 1/8-inch edge lengthwise to keep it straight. Keep the edge filed and sanded to keep it smooth.

To cut, place straight edge far enough away from the cutting line to make the center of the cutting tip directly over the cutting line when the tip is rested vertically and lightly against the straight edge. Drag the torch toward you. You will not be able to see the flow of metal. If all the sparks and fire are landing on your feet, you're traveling at the right speed.

If the slag on the bottom of the cut is too hard to knock off, you are using too much heat. Cut down on the flame, use a smaller tip, or travel a little faster.

Incidentally, the acetylene gas is not what does the cutting. When you pull the trigger on red hot steel, the oxygen attacks the carbon in the steel. The carbon burns and supplies most of the heat. You can prove it by burning the end off a piece of 1/4 by 2-inches. Quickly start another cut on the end. Have your hand on the acetylene valve. Get the cut started and turn off the valve. You can finish the cut without a flame.

That's why you can't cut other metals or even stainless steel with a cutting torch.

How To Do Overhead Welding

Never try an overhead weld with a cold machine. Have the machine hot enough to melt the base metal almost instantly. If your machine isn't hot enough, you'll have to stay in one place too long to melt the base metal and your weld will start bulging downward — it may even drop off. Lay in a little rod and move it away for an instant, before it starts to bulge downward. This lets the crater chill. Then go back, but not as far as you were before.

You can also run light stringer beads. Hold a short arc and move along fast enough to keep it from bulging, but slow enough to fuse.

How To Weld Thick To Thin

To weld a thick piece of steel to a thin one without burning a hole in the thin one, try concentrating the arc entirely on the thick piece, moving slowly toward the thin piece with the crater edge until it fuses to the thin piece. Then back away before it burns a hole.

How To Make A Smooth Bead

Every welder is judged by the looks of his work.

To run a smooth bead, position is very important. Get comfortable. Use a rest for your left hand. With the electrode holder in your right hand, rest your right hand on your left hand so that you can burn a complete rod in a smooth, steady movement. Practice the movement without the rod.

Vertical Weld

If a vertical weld bulges in the center or drops off, you are probably coming back down too low. When you raise your rod to give the crater a chance to chill (freeze or solidify), you are coming back down too far. Try this method:

Bring the end of your rod back down and across the top edge of the crater that just chilled. You can also run a vertical bead downward. Use a fairly hot machine, tilt your rod upward, and run a light bead with a close (or short) arc. Travel fast enough to keep the flux and melted weld metal from trying to get ahead of or below the end of your welding rod, but slowly enough to get fusion.

How To Flux Lead

If you have a pot of melted lead and it has a lot of skimmings on top of the lead, you can skim it back and it will be a different color. The color changes back rapidly to the original color. This is because it oxidizes when it comes in contact with the oxygen in the air. Don't skim the skimmings off and throw them away. Sprinkle a little pine rosin on the top of the skimmings and stir it up. The pine rosin is a flux for the lead. It breaks the oxide and it all melts. This also works on babbitt and solder - you have heard of rosin core solder.

Bulge

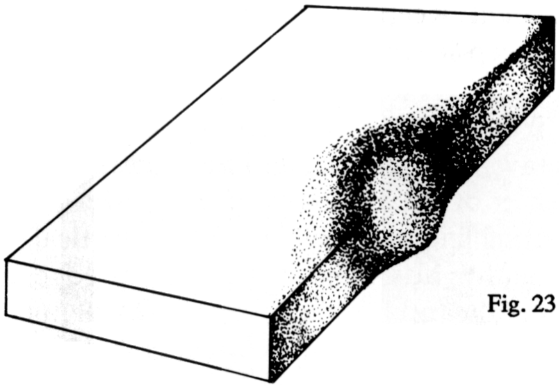

Fig. 23

The bulge in this sketch is highly exaggerated.

If you can imagine that this really happens when a piece of steel is heated red hot in one spot, it will help you figure out a lot of problems that arise in expansion and contraction.

When you start heating one spot, the steel begins to expand on that side. The center of the strip will be under tension and the other side will be under compression. But when the spot gets red hot it has very little strength. The red hot part is pushed together causing it to swell. When it starts to cool and begins to turn black it has more strength. As it contracts, one side of the steel becomes shorter than the other. And the strip bends edgewise.

How To Burn A Weld And Save Both Pieces

If, for instance, you have two quarter-inch pieces lap welded to each other, here's how to handle it:

Light your cutting torch, then cut the oxygen pressure down at the regulator — to the point where you can barely hear it when you pull the trigger.

Position is crucial. Position yourself slightly to the left of the work. Hold the torch with your left hand near the tip. (Have some sort of support handy to rest your left hand on.) Hold the torch with your right hand ready to pull the trigger. Start at the left and travel right. Place your head where you can see the flow of metal. The tip of the torch should be at about a 35-degree angle. Heat the corner red hot and pull the trigger.

When the flow of metal goes through the weld, you will see a black spot at the bottom of the flow. It's okay. Keep on going. If the flow of metal flies back in your face, the oxygen pressure is too high. If you don't see that black spot at the bottom of the flow, you are cutting through the weld and into the other piece of quarter-inch plate. Tilt the tip down more and get below the weld with the flow of metal.

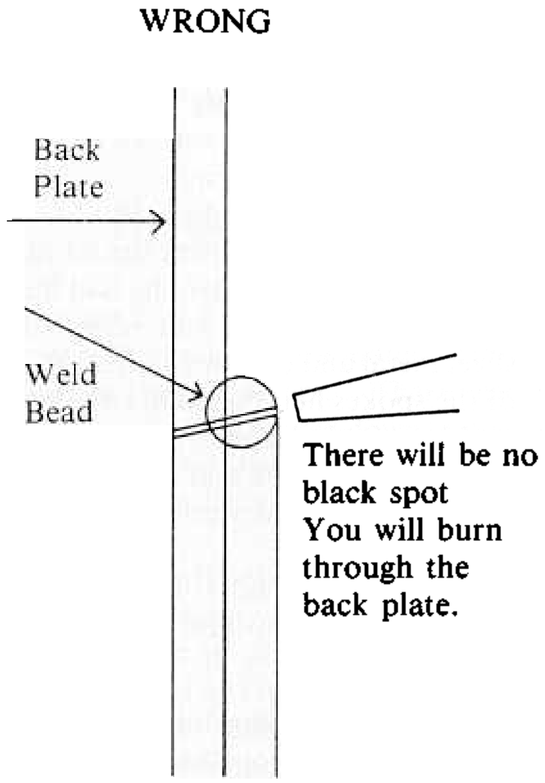
WRONG
Back Plate
Weld Bead
There will be no black spot
You will burn through the back plate.

Fig. 24

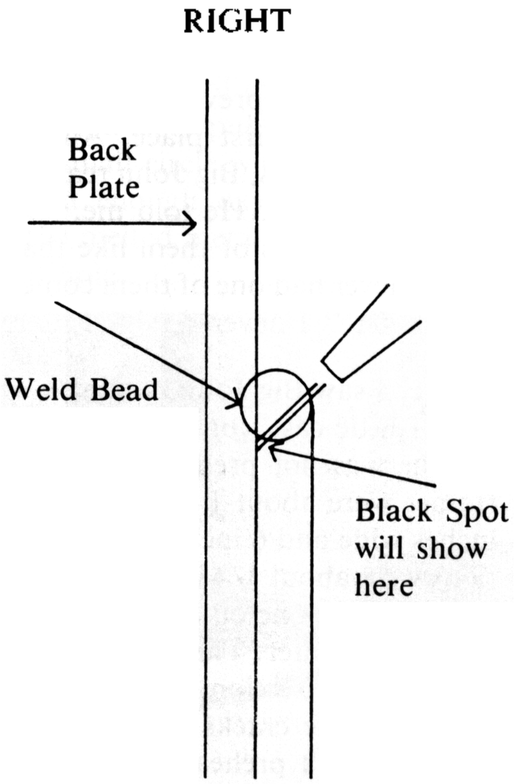
RIGHT
Back Plate
Weld Bead
Black Spot will show here

Fig. 25

Big John

Big John had been working at the shop for 16 years when I started my apprenticeship. He was a top notch welder. He welded all of the broken truck frames — and there were a lot of them back then! Big John had his own way of reinforcing them. He didn't put a fishplate on the side of the channel as you might expect. After he welded the crack up, he would weld a flat bar (about 1/4 x 2-inches or thereabouts) on the top or bottom leg of the channel, extending six or eight inches on each side of the weld. Some of the frames started cracking at the top if the load was too far to the rear of the rear axle.

Big John argued that the weld would hold as much weight as any other part of the channel, and if the reinforcement was put on the bottom (or top) leg of the channel it would prevent the crack from starting in the first place. And if the crack never starts, Big John reasoned, it will never break. He told me, "I have welded hundreds of them like that and I have never had one of them come back on me yet." I never doubted him.

Once I saw Big John brazing a whole pile of little cast iron frames. I noticed that he was not preheating them. The frames were about 1/8 inch thick by 4 inches wide and 6 inches long. The sections were about 3/4 inches wide — one on each side, one on each end, and one across the center. They were all broken on the center section. He had previously ground the cracks out. I asked him why he didn't preheat them.

He told me, but only after twisted his arm. Here's his trick:

Big John held the torch at an angle so that, while he as heating the crack on the center section, the other part of the flame extended over to the end section and heated it. That let the two sections expand together and then shrink together. If I had been doing it, I would have heated all three cross sections.

learned a lot from Big John

Another time, I was helping him with a big steel flywheel, about six or eight feet across with a heavy rim and spokes. We blocked it up on bricks to keep it off the dirt floor. Then, we placed several natural gas torches around it and started heating it with soft flames (no air pressure), on the rim only and covered it with some sheet metal. The heat had not been on it for more than five or ten minutes when there was an explosion that shook the building.

The boss came running out of his office to see what catastrophe had hit. By that time, Big John had removed the sheet metal and cut the torches off. One of the spokes had pulled in two, leaving a wide gap (the spokes were oval, three or four inches thick and ten or twelve inches wide — solid steel).

"John," the boss growled, "you heated the rim too hot."

"Look boss," Big John said, "if that crack goes back together when it cools, you can have my job. But if it doesn't, don't open your mouth."

Here's what happened.

The guys who removed the flywheel had trouble getting it off the shaft. They heated the hub of the flywheel red hot in order to remove it. In doing so, the hub tried to expand outwardly, but it couldn't. The rim and spokes were cold. They removed the flywheel and the hub began to shrink. The shrinking of the hub put tons and tons of tension stress on all of the spokes. When just a little heat was added to the rim, which was under compression, it expanded enough to pull the weakest spoke in two. The gap did not close up when it cooled.

Notch Effect

Cut a notch in a shaft, and you know where it will break - in the notch. This is called a "notch effect." There are many others. A shaft turned down in one place to a smaller diameter will break right next to the larger diameter.

A built-up shaft will break at the end of the welds. A poor weld is a notch effect. A shoulder of almost any kind is a notch effect. The end of a reinforcement is a notch effect which can be minimized by tapering the ends.

How To Remove A Broken Tap

Place a nut over the tap and weld the tap and the nut together. Fill the hole in the nut with weld metal. Use a reverse rod with a light flux. Let it cool.

Before you start, the nut needs to be held in place. One way to do this is to use a washer with a hole bigger than the hole in the nut. Weld it to the end of a flat strip of metal. Place the washer over the nut. Block up the other end of the flat strip to make it level and place a weight on the flat strip. (See Below)

With a little skill, a broken off tap can be removed with a cutting torch. Concentrate the heat on the tap only. Hold the tip in a straight line with the tap. Let about half of the flames go down the side of one of the flutes, and the other half on the center part of the tap. When the tap gets red hot, pull the trigger and make a small circle with your tip. You will need enough oxygen pressure to blow the melted metal back out if it is a blind hole.

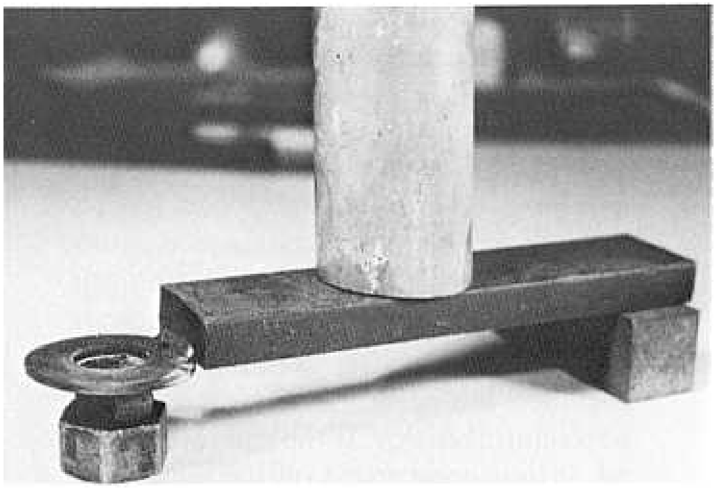

Fig. 26

Built In Stresses

Every weld contracts or shrinks in all directions. Steel is made up of round crystals and expands when it is hot. When a piece of steel is heated uniformly to a red heat, it is free to expand. But if it is heated in one spot, it is not free to expand. When it tries to expand outwardly the cold metal around it prevents it. The crystals are rearranged to make the metal thicker at that point. Then, when the metal cools, it contracts. This puts the crystals under tension. This will cause a thin sheet of metal to buckle. All welds are under stress unless the stress has been relieved. A light bead on a very heavy piece of steel usually cracks the weld wide open. If the hub of a cast iron wheel is heated red hot, it usually pulls one of the spokes in two.

However steel sometimes has stresses already built up in it. When you weld, you may relieve more stresses than you create. Anytime a piece doesn't move the way you expect it to, blame it on the built-in stresses. When a machinist takes a cut on a straight shaft, the shaft sometimes bends a little due to the stresses being cut off. That is probably why the ball-peen hammer was invented. You can peen the inside of a curve on a shaft to straighten it.

If you repair a broken piece of steel you should reinforce it with another piece of steel, after welding it back together. But **never, never** weld the ends of a reinforcement. If the ends are welded, it puts great stress on the metal adjoining the weld. A little more stress and vibration will cause a crack right next to the weld. Cut the ends of the reinforcement on a 45° angle instead of a 90° angle. Always skip weld the reinforcement — weld an inch and skip an inch. Never weld all the way to the ends of a reinforcement. Leave about a quarter-inch, or the thickness of the reinforcement.

Crystals

When a machine part is subjected to severe stresses over a period of years, sometimes the crystals become elongated. If a part breaks, they say that the steel crystalized. I was told that if you heat a piece of steel to a red heat, the crystals would return to their original shape.

To prove this to myself, I cut some 1/2'' round, mild steel bars. I hammered one of them cold, on the end, to bring it to a point. It didn't taper much until it cracked lengthways. The crystals in the steel elongate when the steel is stretched, but they will only elongate so far and then they separate, and that is a crack. Then I hammered another one, but I stopped before it cracked, then heated it to a red heat and waited until it was cold before hammering again.

I continued this process a number of times. The center of the bar did not stretch much, if any. A hole was left in the center. It seems to be quite evident that the crystals return to their original shape when heated red hot.

How To Burn Through A Large Steel Shaft

You will need the biggest cutting tip you have and all the oxygen pressure your regulator will give you. If your tip is big enough and your oxygen strong enough, you can cut right through. But if the tip is just a little too small for the job, try this:

You must be able to see the flow of metal all the way through the shaft at all times. With the shaft in a horizontal position, place yourself on one side of it. The cutting tip should be on about a 45° angle from vertical. You should be able to see over the torch and watch the flow. Before you light the torch, practice moving it in the path that it will travel. Then light it and heat a starting spot. Don't be in a hurry to start cutting; get it red hot. When you do start cutting, weave right and left and make a gap about three-quarters of an inch wide (for a four-inch shaft, wider for a larger shaft). If the gap is not wide enough and you can't burn all the way through, go back and start over again on one side. Try burning off another half-inch. If that doesn't work, your tip may be too small.

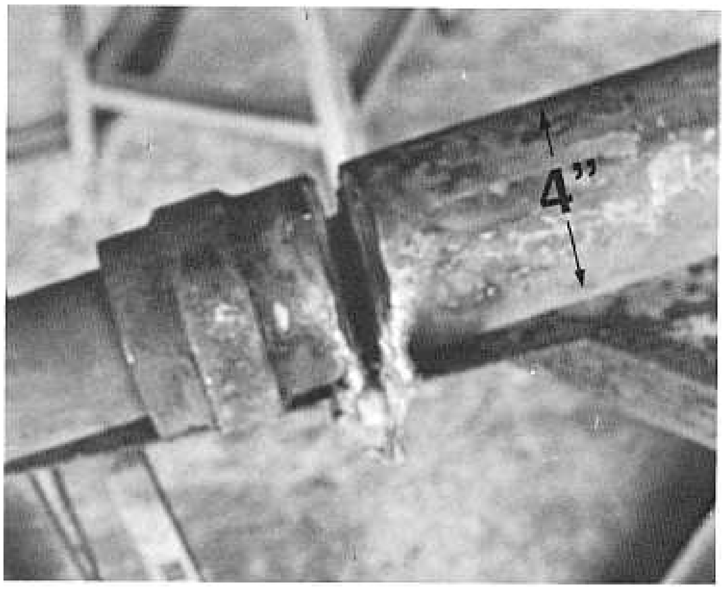

Fig. 27

My First Commercial Weld

In 1936 I was a welder's helper. I had been practicing running a bead for some time. One day someone brought in a broken bumper support from a Model T Ford into the shop. The welder welded it back together and tacked a reinforcement over the weld and looked at me and said: "You can weld this, son. Go ahead."

I welded the top and bottom and both ends. Later, the vice president of the company came through and stopped to talk with the welder. In a few minutes he noticed my weld. He jumped straight up and yelled, "Who did that?"

The welder looked at it and realized what I had done. He said, "It's all my fault, boss. I didn't tell him not to weld the ends. Don't worry about it. I will cut it off and put a longer piece on it."

The boss cooled down. And that was the only time I **ever** welded the ends of a reinforcement.

Cast Iron Welded W ı Nickel Electrodes

If you have a cast iron ornament that was broken accidentally, you can successfully weld it with a nickel electrode. The cast iron should be preheated. However if a cast iron part is subjected to stress and you weld it with a nickel electrode, it usually breaks right next to the weld. I am not a metallurgist and I don't know why, but I think it has something to do with getting the cast iron too hot. Cast iron can be successfully brazed with a bronze rod and an acetylene torch. Chamfer the break and grind 3/4 inches clean on all sides. The bronze will have to be built up a little bigger to equal the strength of the cast iron. Heat the cast iron to a dull red, heat the rod and dip it into "Cast Iron Brazing Flux" and melt it into the Vee. It has to tin the surface of the cast iron (spread out) if it balls up, it will not stick. Fill the Vee and then spread out over the edges that you ground clean. Dip your rod in the flux whenever it is necessary to keep the bronze tinning and flowing into the pores of the cast iron.

Now this procedure is for a single piece of cast iron that is free to expand and contract without putting a stress on another place. On a wheel, pulley, or gear, they can be brazed if you heat them slowly and uniformly to a dull red heat. Some welders can heat the rim of a wheel at each spoke (that allows the heat to radiate each way on the rim and inward on the spokes). As the rim expands and gets larger in diameter, the spokes expand to keep out stresses.

At some places, they weld cast iron blocks and heads for Diesel engines. They grind the cracks out then place the block in a portable furnace. Heat it slowly to a red heat, open a door on top leaving the heaters on, they reach down in the furnace and acetylene weld it with a cast iron rod. Then they close the door and leave the heat on for awhile. When they cut the heat off they leave the block in the furnace. The block is surrounded by red hot fire bricks. It takes some time for it to cool. Then they machine it. It costs about half as much as a new block.

A foundry made me a cast iron sheave with an extra large hub in order to bolt a brake drum to it. When I picked it up, it had a crack in one of the spokes. I showed it to the foreman. He said the mold should have been stripped. He explained that when it took the hub longer to cool than the rim and spokes, they were supposed to remove the sand off of the hub to make it cool as fast as the metal on the rim and spokes. The rim cooled first and shrank, forcing the spokes to squeeze the hub together, making it thicker. Then when the hub cooled and shrank, the rim would not "give" and the hub pulled one of the spokes in two.

Straighten 6 Inch Strip

If you cut a 6" strip off of a 1/2" plate. 8 feet long, with a cutting torch, it will be curved. The side that you cut contracted and now it is shorter than the other side. It can be straightened by stretching the side that you cut, by peening or by shrinking the other side. Stand on edge and with an acetylene torch heat the edge red hot. As it gets red move along. Have someone come behind you cooling it with water.

Clevis On Channel Iron

Figure 28 shows how **NOT** to weld a clevis on to a stress member.

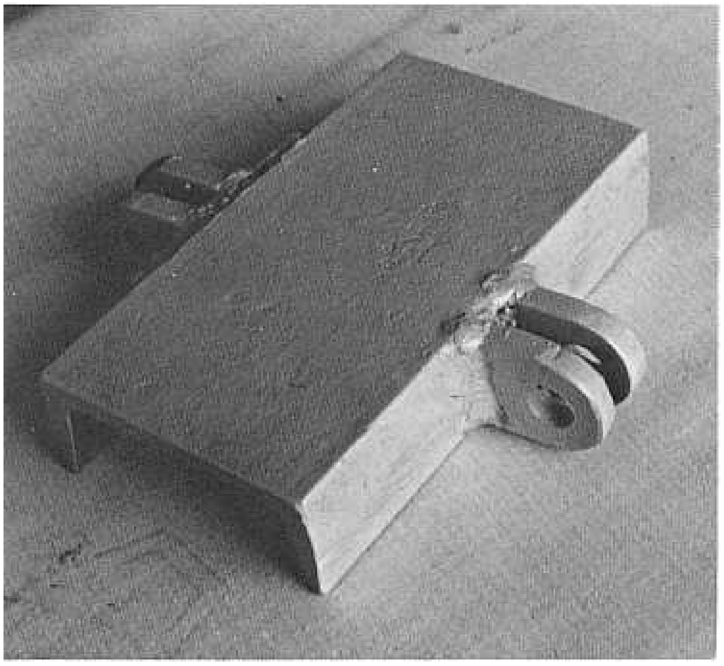

Fig. 28

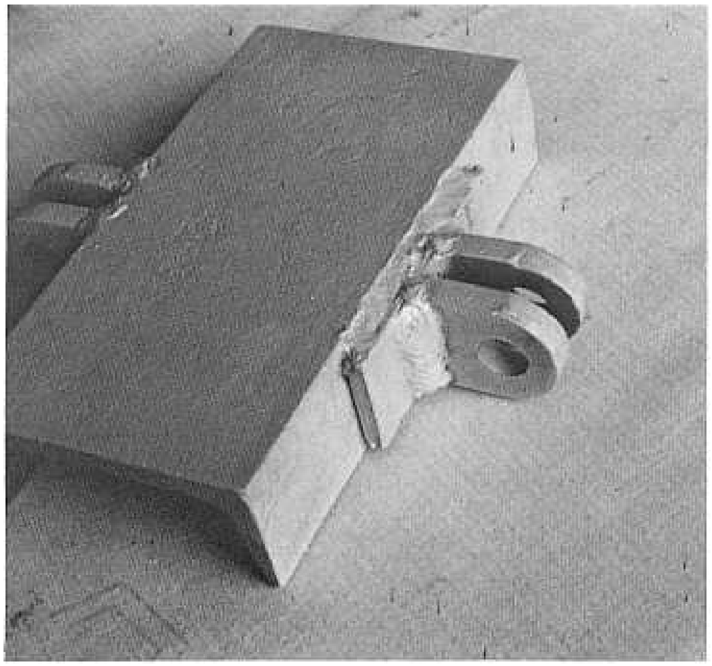

Fig. 29

Figure 29 shows a better way to weld it. The two pieces with the holes are welded on to the patch, then the patch is welded on to the channel iron lengthways with the channel. No doubt you are tired of reading "Do not weld the ends."

How To Temper A Chisel

The cutting edge of a chisel should be soft enough to be filed with a file, but no softer. If it is too hard it will break.

You will need a tub of water, a file, a pair of tongs or vice-grips, and a torch.

Slowly heat the chisel from the cutting edge back about three or four inches to a red heat. With the tongs holding the chisel on the back end in a straight line with the chisel, place the tip of the chisel into the water about a half-inch or more. Hold it there until it cools. Then rapidly plunge the entire chisel into the water and back out. Rest the chisel on the edge of the tub and, with the file, rub the end of the chisel. It should be too hard for the file to cut. Keep stroking the tip until the file starts cutting. Then, instantly shove the whole chisel into the water and back out again. The tip should again be too hard for the file to cut. Continue filing and dunking until it is soft enough to file when you take it out of the water. Then place the chisel in the water and move it around until it is cold.

When I was serving my apprenticeship, a contractor would send in an armful of drills every few weeks to be drawn out and tempered. The drills were about one and a quarter inches in diameter and about 18 inches long. They were used in an air hammer to break rock and concrete.

A friend had told me how to temper a chisel. I decided to try it out on these pointed drills. Without permission.

A few weeks later they brought in another batch with special instructions that the same man who tempered the last ones to temper these.

It worked.

How To Make A Circle Lie Flat

An 18-inch circle that has been burned out of a 1/8-inch plate will bulge in the center. It does so because the outside edge that was heated to a red heat is under tension. To flatten it, remove the slag, place the edge of the circle on an anvil, and hammer it with the flat face of the hammer centered on the edge of the circle. Hold the other side of the circle up so the edge will be in flat contact with the anvil.

Turn the circle and hit the edge all the way around, hard enough to stretch the edge, but not hard enough to distort the metal too much. Try leaving a little space between each stroke and, if that isn't enough, go around again, hitting between the last strokes. If the edge is stretched too much it will buckle. The buckle may be corrected by heating red hot on the edge in very small spots. Cool each spot with water before heating another one. Or you can skip from one side to the other, letting someone else cool the spot you just heated.

On the deck of a ship which has many buckles, one man heats spots and another one cools them with air and water.

How To Remove A Stuck Sleeve From A Hole

Screw a tap that is slightly larger than the hole in the sleeve into the sleeve — it needs only to scratch the inside of the sleeve. Drive it out from the other end with a mild steel rod (never hit a hard surface with another hard surface).

Another way to remove a stuck sleeve is to heat the sleeve red hot and let it cool. You can heat a large sleeve red hot in a straight line from one end to the other — it does not have to be red hot all at the same time. Heat red hot in one spot and, as it reaches red hot stage, move forward. The black part behind you has already shrunk.

This works on steel. Don't heat cast iron unless you are sure you know what you're doing.

How To Shrink A Steel Pulley

We once had a steel fabricated pulley that had a tapered hole in the hub to fit a split Q.D. hub. The Q.D. hub had a flange which would pull up against the hub on the pulley without being tight on the shaft.

To fix it we heated the rim of the pulley red hot with a torch and slowly moved toward the center, heating it red hot as we went — heating the hub red hot in one place also. It helped, but it wasn't enough.

When we heated it in four places on quarters it worked fine. There was a quarter-inch of space between the flange and the pulley and the hub was tight on the shaft.

CAUTION: do not try this method on a cast iron pulley.

Reinforce Inside Of Tubing Or Pipe

If you reinforce a broken-off piece of tubing with a smaller piece of tubing or solid bar that fits on the inside of the broken off tubing, do not weld through the broken piece and into the smaller reinforcement on the inside. If you do, it puts it under stress and may cause a break. You can see that if you run a light bead around a solid shaft, the light bead has to shrink, but it will have a hard time compressing all of that cold steel on the inside of the small piece.

Cut slots lengthways on the broken tubing to weld through and into the reinforcement before you put the reinforcement inside. The size of the slots depend on the size of the tubing. There should be enough welds on each side of your weld (that welds the outside piece back together) to equal the strength of the tubing. Stagger the slots. With the reinforcement in place and the broken tubing lined up, weld the broken tubing back together. Try not to weld through to the inside piece of tubing. Let it cool, then weld up the slots.

How To Repair A Crack In A Piece Of Tubing

Cut out piece as shown in Fig. 30 - shaded area.

Then cut out patch to fit hole, out of same size tubing. Fit patch in hole smoothly and tack at each end, A and B. Then weld one side only — from D to B — C to D, then A to C. Let cool, then weld other side — same way.

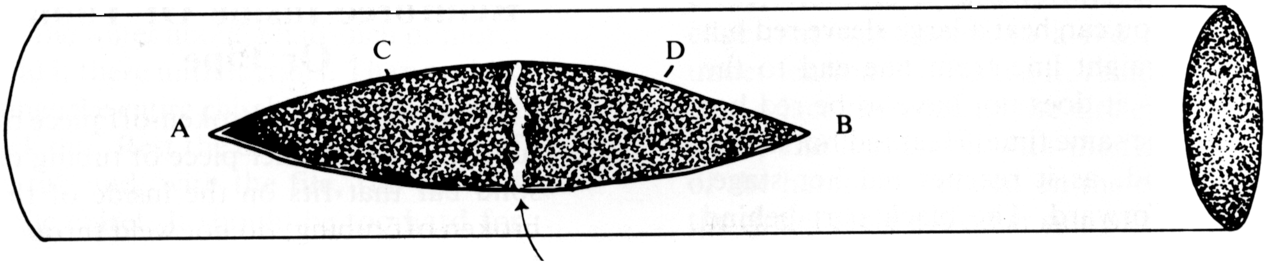

Fig. 30

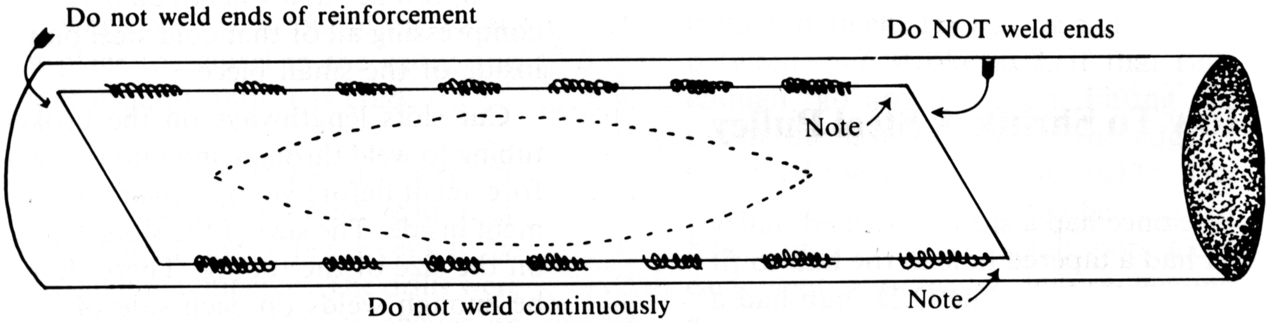
Do not weld ends of reinforcement
Do NOT weld ends
Note
Note
Do not weld continuously
Weld 1 inch and skip 1 inch

Fig. 31

Grind welds on patch and place reinforcement over patch as shown in Fig. 31. It is best to always grind long ways with a stress member. Cross scratches can sometimes cause a crack. Do not weld all the way to the ends of the reinforcement — leave 1/4 inch or the thickness of the reinforcement.

Weld Inserted Pipe

If you weld around the pipe as shown in Fig. 32, you will create a "notch effect" which usually causes a break adjoining the weld.

Slots may be cut in the larger pipe to weld through as shown in Fig. 33 or holes may be drilled to weld through. Slots or holes should be staggered, with enough weld metal to equal the strength of the pipe.

If you don't have room enough for the slots or holes, you may cut slots on the end of the larger pipe as shown in Fig. 34. The direction of travel should be away from the end of the larger pipe. Start your weld slightly to the inside of the slot.

If the joint must be air-tight, the larger pipe can be cut as shown in Fig. 35. When two beads meet at a point or crotch, extend one bead a little further and try to taper it out to nothing. This prevents stopping the shrinkage abruptly.

Some welders may not agree with me on the "notch effect." You think about it. If every weld shrinks, the last part of your bead will shrink more than the first part. The first part of your bead is already black. Even if the weld didn't shrink, there would still be a shoulder.

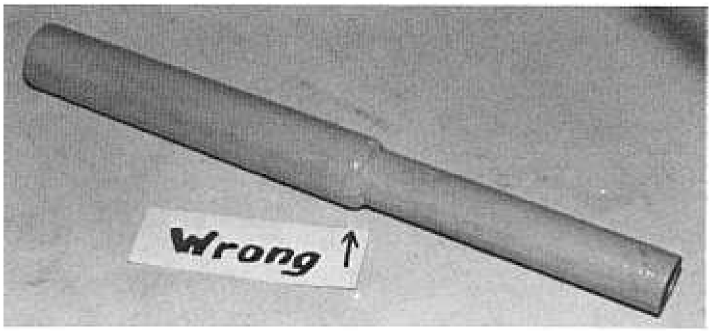

Fig. 32

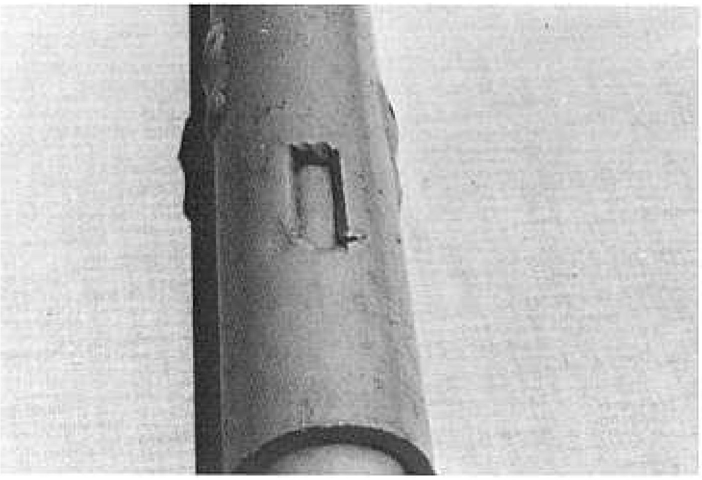

Fig. 33

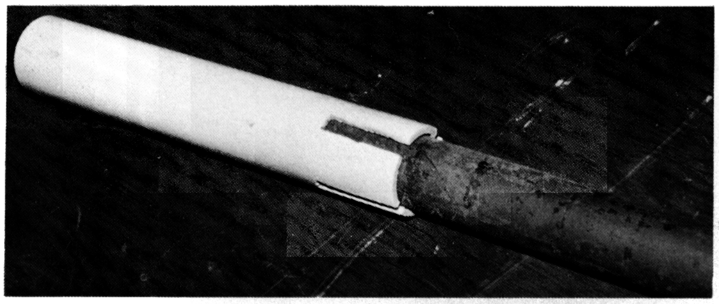

Fig. 34

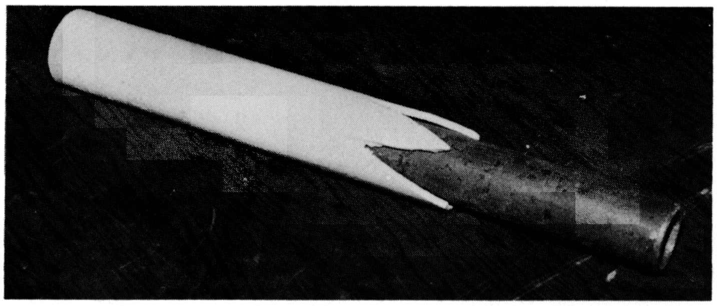

Fig. 35

Another Method For Welding Pipe

It is general practice to bevel two pieces of pipe and butt weld them together. However, if you are using pipe or tubing for structural purposes, and the welds are subjected to tension, torsion, impact or shear stresses, and you have trouble with cracks, you may want to consider the method shown below. The method shown below has no cross welds at all. The welds are back-stepped, and at the ends where the two welds meet, they are reduced to one weld which tapers out to nothing to avoid stopping the shrinkage abruptly. This method is more expensive than a butt weld, but you have a lot more weld metal securing the joint.

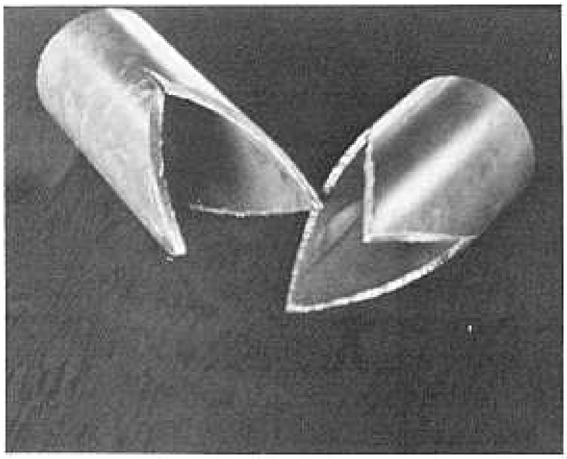

Fig. 36

Make a symmetrical pattern to cut the pipe out with, as shown above.

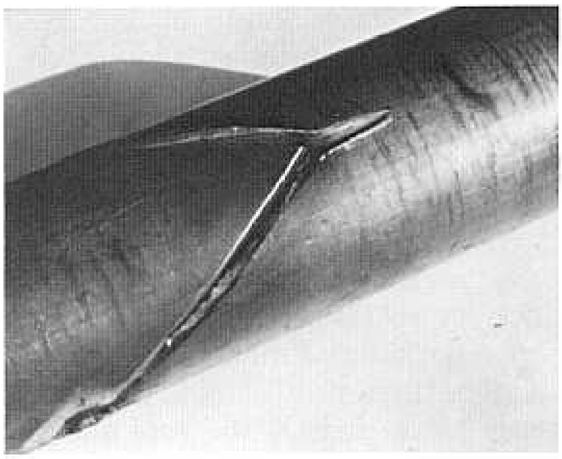

Fig. 37

Bevel edges and grind slot at end.

Another Method For Welding Pipe

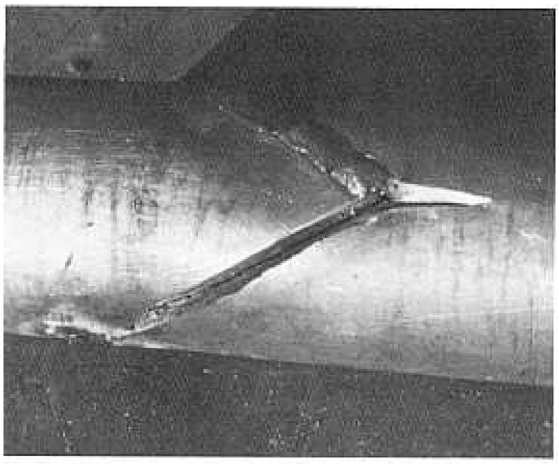

Fig. 38

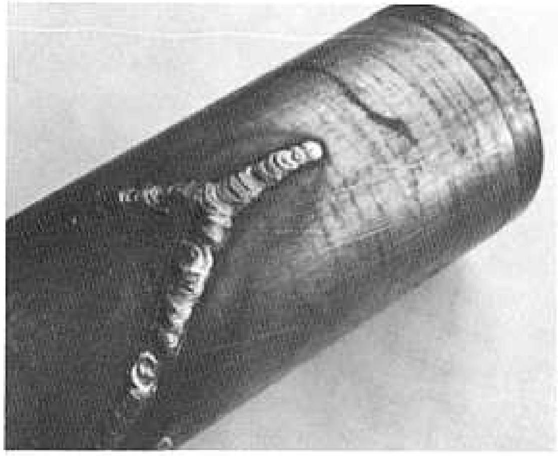

Fig. 39

Weld up two grooves that are exactly opposite of each other and let them cool. The shrinkage will open the other two grooves slightly. Then weld the other grooves. This way, you will have less shrinkage on the joint.

Build Up Inside Of Tubing

The pictures below show that if a tubing is built up on the inside, it will cause the tubing to shrink. This 4 1/4" O.D. piece shrunk .045" in diameter.

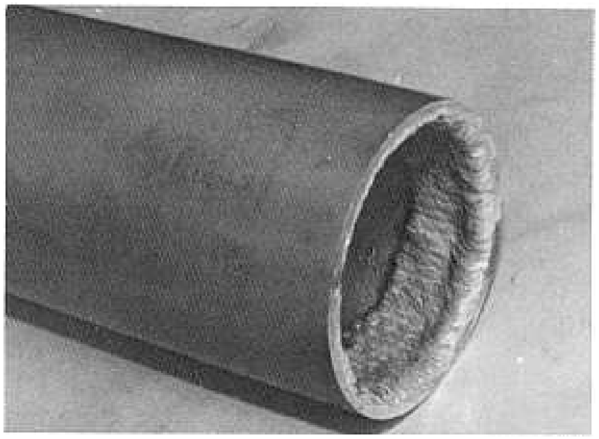

Fig. 40

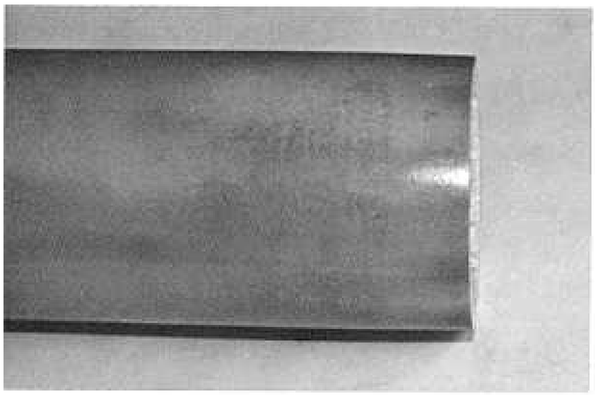

Fig. 41

Test Welds

Figures 42 and 43 are test welds that show how much the tubing shrank from end to end. Figure 42, welded lengthwise, shrank **four** thousandths of an inch in length. Figure 42 probably shrank more crosswise, but that does not cause it to crack.

Figure 43, welded crosswise, shrank **twelve** thousandths of an inch in length. The cold metal surrounding the welds resisted. The cold metal is now under compression stress and the welds and the adjoining metal that got red hot is under tension stress almost to the yield point.

Think about this and **please** don't weld crosswise on a stress member.

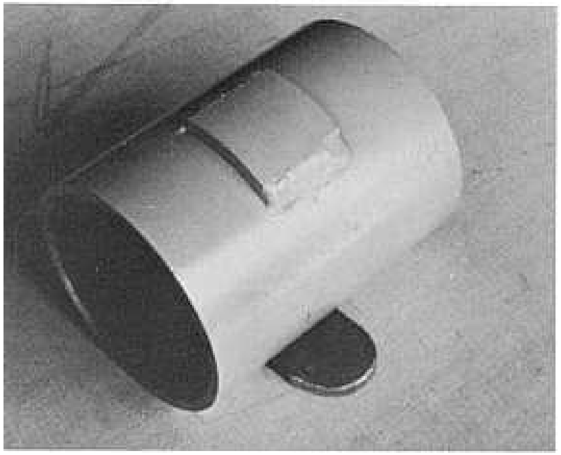

Fig. 42

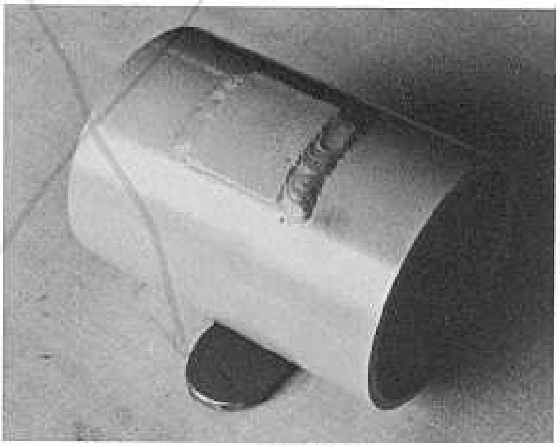

Fig. 43

Flying Coaster

No. 1 shows large gap to prevent someone from welding crossways on the clevis. On the old ride they were welded crossways and they broke off.

No. 2 shows end of patch is not welded.

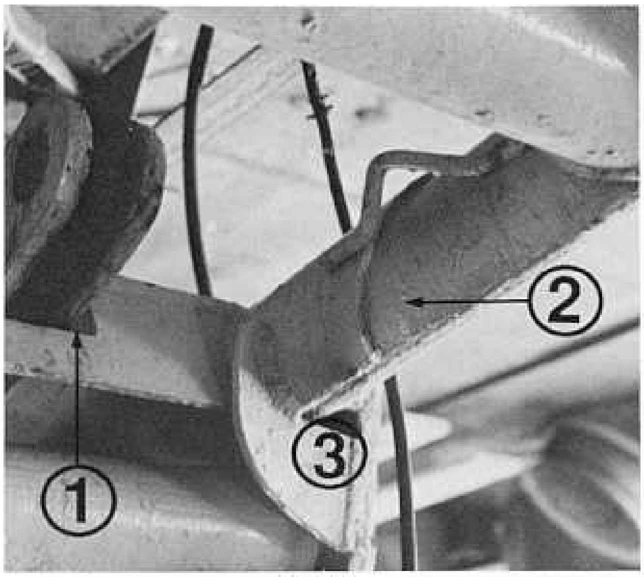

Fig. 44

No. 4 welded crossways, but it is on the inside. I have never seen a failure here.

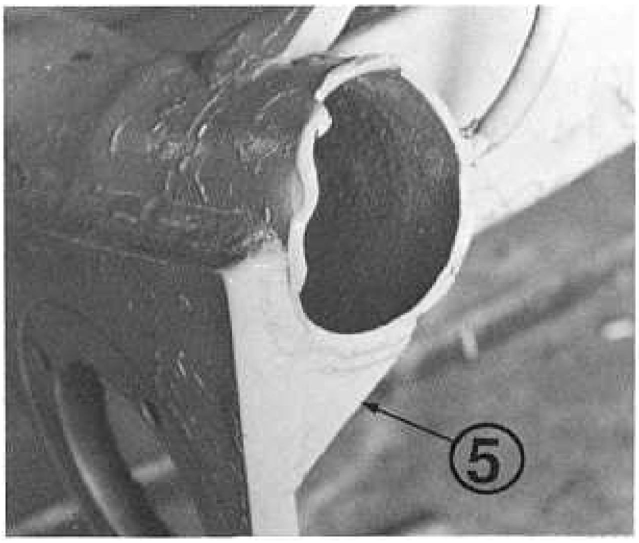

Fig. 45

No. 3 shows that we did not weld a gusset on to the tubing as shown in No. 5.

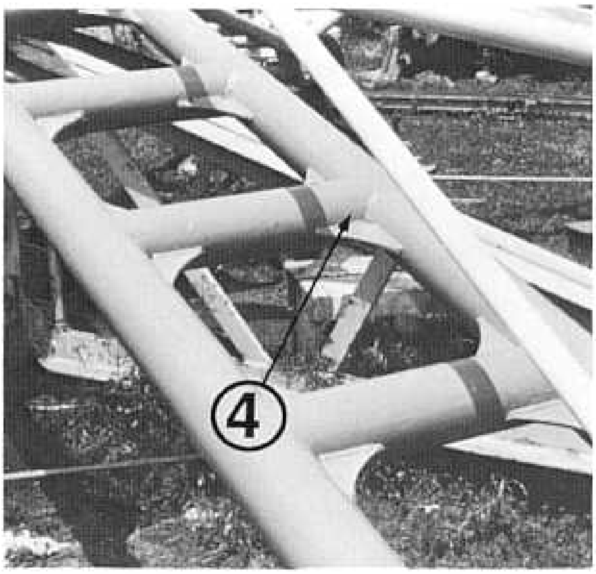

Fig.46

No. 5 weld on gusset and tubing caused the break on both pieces of tubing.

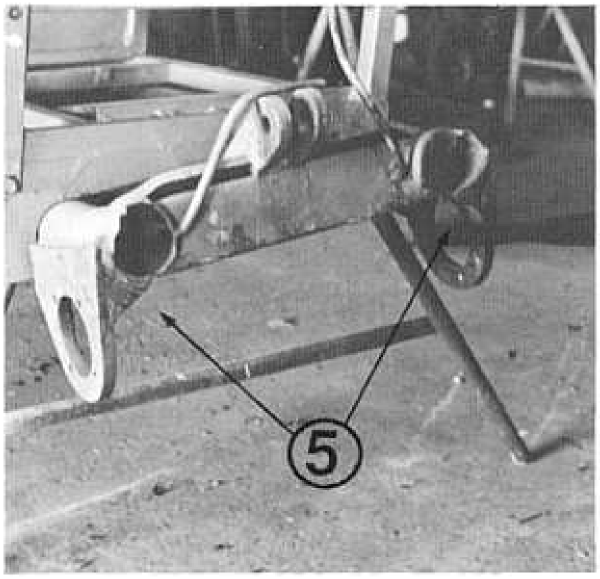

Fig. 47

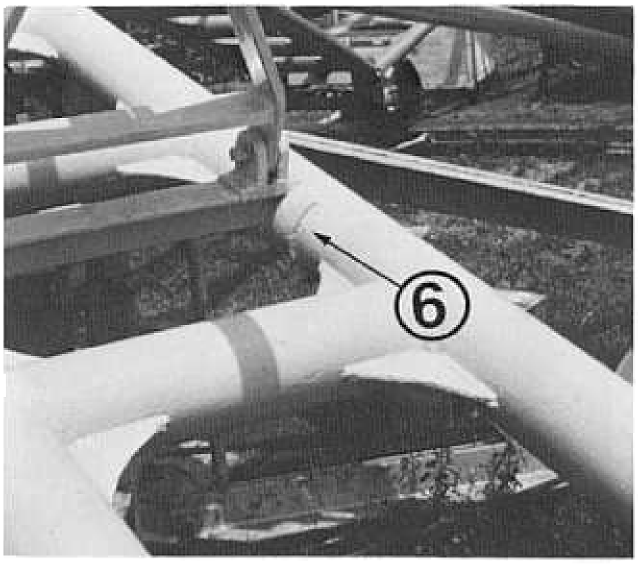

Fig. 48

No. 6 shows T bar welded to patch that is welded lengthways on tubing not welded on ends. On old ride, T bar was welded to tubing and it cracked many times.

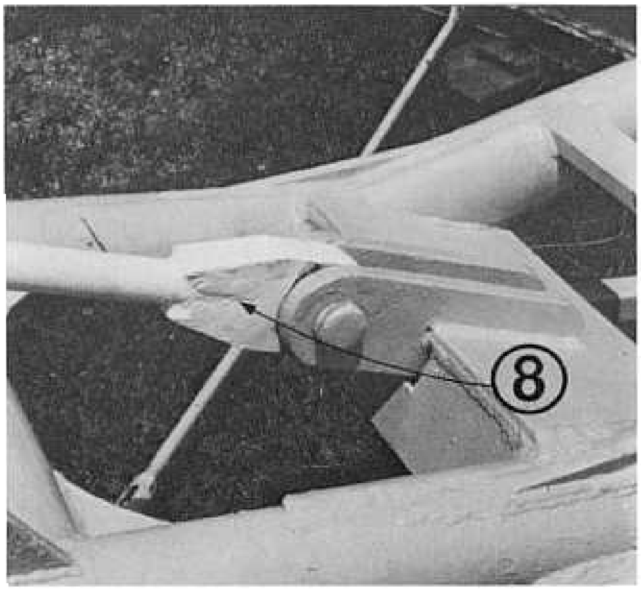

Fig. 49

No. 8 shows rod end is tapered on the end, not welded across the end and the welds are tapered. Even a square shoulder here could cause a break.

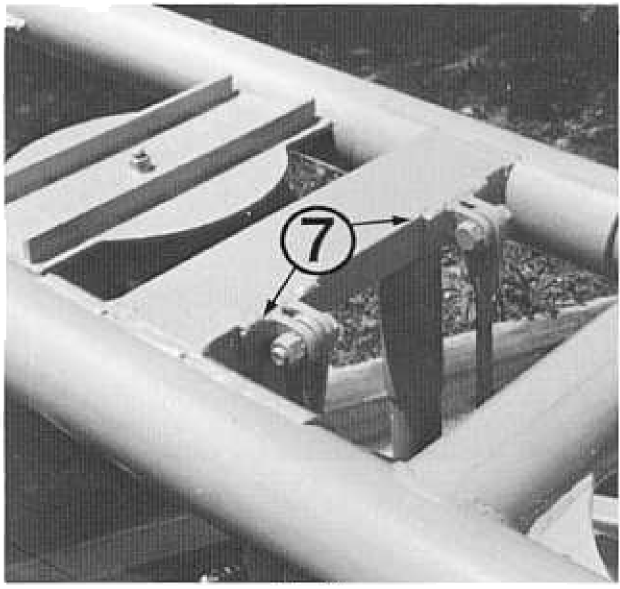

Fig. 50

No. 7 shows short pieces of angle irons welded lengthways with channel iron.

Roller Coaster

Almost every time I saw a certain type of portable steel roller coaster there was a welder on top of it welding something. With a little investigation, I found out why.

The track pieces were angle irons. The cross ties were channel irons. They were fastened together with very short pieces of angle irons. One leg of the angle iron welded flat on top of the channel iron cross tie and the other welded to the side of the angle iron track. The angle irons were not welded lengthways with the angle iron track, but **crosswise.** (See Figure 52) This causes a notch effect. The angle iron track always cracked right next to the weld. They'd weld it. And it would crack again, right next to the new weld. I actually saw one place with 20 or more welds, side by side, overlapping each other.

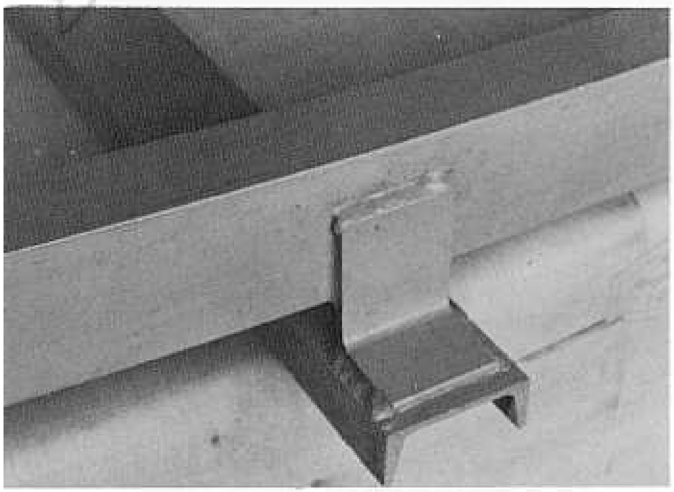

Fig. 51

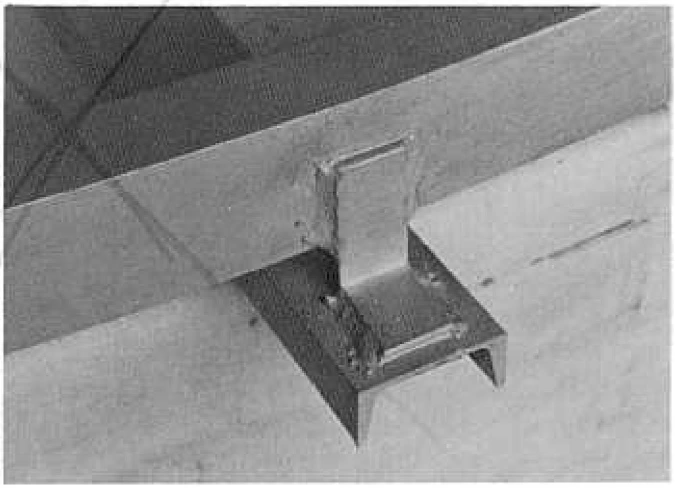

Fig. 52

Cracked Center

This old center for an amusement ride shown below has been abandoned since 1977.

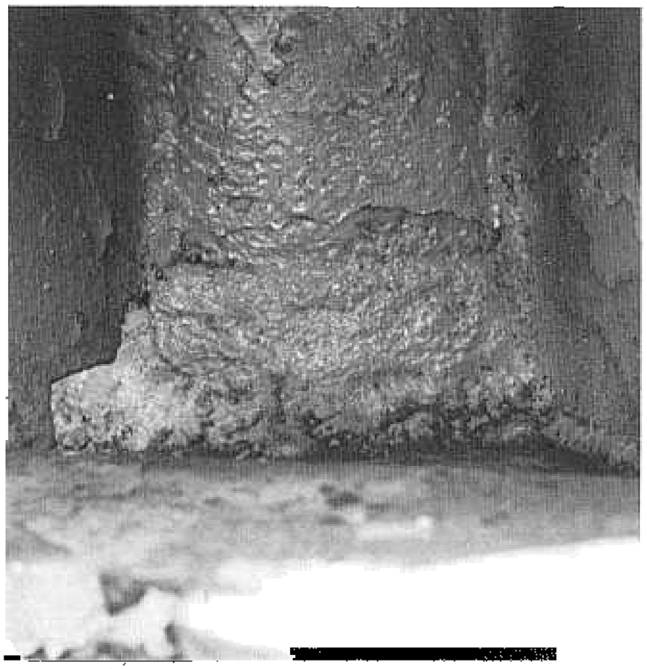

Fig. 53

I remember back in the 1960's, my foreman called me from 600 miles away. He told me the 12'' tubing on the center was cracked. I knew that, with all of the gussets and stiffeners that were welded to it, there could be only one reason for it to crack and that is the shrinkage of the welds. I told him to go find a welding shop that could weld it and call me and let me talk to the welder. I asked the welder if he had an air-hammer. He said "I think I can get one." I told him to vee out the crack and then run light stringer beads and peen each bead with the air hammer before running another bead over it. He said "OK." When the ride was brought back to the shop a few weeks later, it had a big beautiful bead, but it was cracked wide open. The welder couldn't get an air hammer. I had to weld it myself. I peened each bead and also peened the top and bottom beads that were also shrinking. It didn't crack again.

Peening Test

If you don't know how much peening should be done to relieve the stress, there is a simple test you can make. Stand a plate straight up on another plate one inch from the edge. Tack it on and square it to a 90 degree. Weld a bead 2 or 3 inches long on the 1 inch side. Place your square on it to see how much the weld pulled it. Then use an air hammer with a blunt chisel and peen the weld and the adjoining metal which got red hot until the vertical plate is back to a 90 degree.

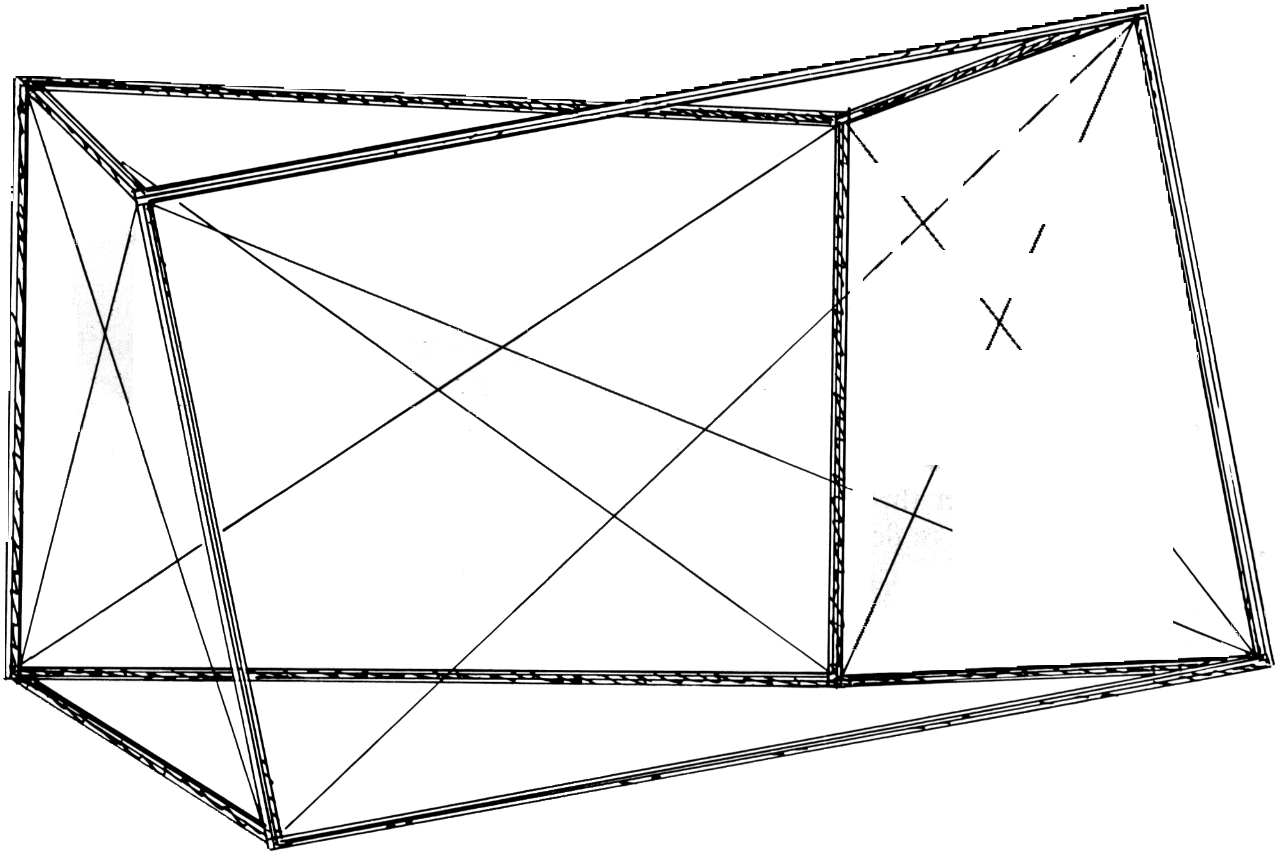

Fig. 54

Cross Brace

A box built of steel, and cross-braced on all four sides will still be flexible. It can be twisted as shown above. To make it rigid, cross-brace the top and bottom also.

For example, twist an open top cardboard box. Two diagonal corners will be farther apart and the other two will be closer together.

Therefore, never cross-brace the top of a truck body. If you do, the cross-brace will tend to lift one wheel off of the ground when it comes to a low place in the road. It could stretch a brace, break something loose or hold the wheel up off of the low place.

Test Welds On Low Carbon Steel

The following subject has been a controversial topic for many years. Whether or not to weld the ends of a reinforcement (or patch).

One writer stated: "Low carbon steel has such a high ductility that it is *not harmed* by being stressed above its yield point. Hence, it is *not weakened* by any stresses that are left in it after welding.

The distortion and the stresses that are caused by welding have been overemphasized. Perhaps this is due to our zeal to make our welds stronger than the parent metal rather than to make them just strong enough to do the job. So frequently after a weld fails, we ask, "Where did it break?" when the question should be, "At what load did it fail?"

We admit that low carbon steel is very ductile, but almost everything has a limit. The above quotation, in my opinion, has helped cause many failures. I have seen hundreds of reinforcements welded on the ends, with a crack adjoining the weld. I would guess that only about 10% of the welders do *not* weld the ends of a reinforcement.

You can see that if you heat a piece of ½" round steel in the center, to a red heat, it will expand in length. If it is placed across the jaws of a vice while heating, the ends of the specimen will be free to move outwardly as it is heated and free to move inwardly as it cools. That is "free expansion". After it cools there will be little if any difference in the length of the specimen. Now open the jaws of the vice and place the ends of the specimen between the jaws and tighten them very lightly, to restrict the expansion.

As the specimen is heated, it expands and puts compressive stress on the specimen. The specimen will compress a little. But when the spot that you are heating gets red hot, it has very little strength. It will push together and cause a bulge. It will be larger in diameter at the place where it was red hot. As it cools, it will fall out of the vice. It is shorter now than it was before.

This is "confined expansion". There is a great difference between "free expansion" and "confined expansion". Some tests were made to show what happens. A 3/16 inch plate was sawed to about a 6"x 6" square. Three overlaping beads, 2 inches long were made in the center of the plate. After they cooled, the welds were ground down flush with the plate, to avoid making a "shoulder effect" which will cause a crack. There were 2 inches on each end of the welds that were not welded.

The cold metal on each end of the welds resisted the expansion and caused the red hot place to be pushed together (up-set). Now, the places which were red hot are thicker than they would have been if the expansion had been "free". As the welds cool, they shrink from all directions, something is forced to stretch. If a weld is not ground down flush, it is thicker than the adjoining metal, a crack usually starts adjoining the weld. If the weld is ground down flush, it usually cracks in the weld.

The 3/16"x 6"x 6" specimen was bent double on about a 3/4" radius. It did *not* crack. See Fig. 55.

Then we sawed a specimen from a 3/8" plate (twice as thick) same size, same welds, same type electrodes and same setting on the machine. We ground it, bent it to an 81 degree and the welds cracked, as shown in Fig. 56. Of course the metal had to stretch much further on the thicker specimen.

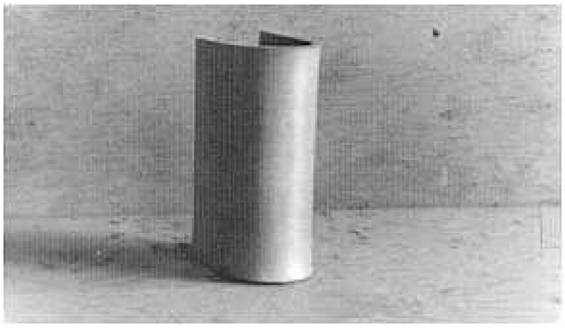

Fig. 55

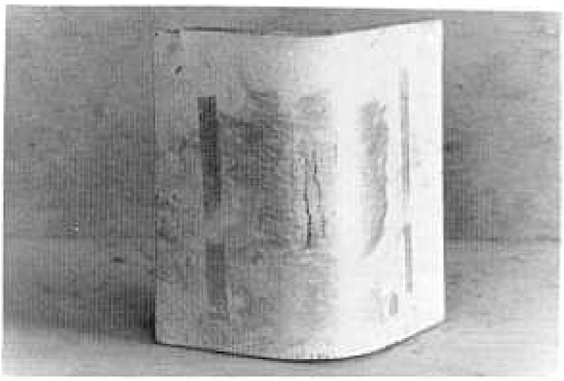

Fig. 56

Then we sawed a specimen 2 inches wide from the 3/8 inch plate. The welds traveled from edge to edge. There was no cold metal on each end of the welds to resist the expansion. Of course, there would be some stress because the welds were not all red hot at the same time. The specimen was ground down and bent double. It did not crack. See Fig. 57.

The specimens in Fig. 56, Fig. 57 and Fig. 58 were all sawed from the same 3/8 inch plate. In Fig. 58 one was sawed across the grain and the other one was sawed with the grain, 1/2 inch wide, no welding was done. They were bent double and then pressed flat. Neither of them cracked. That is ductile.

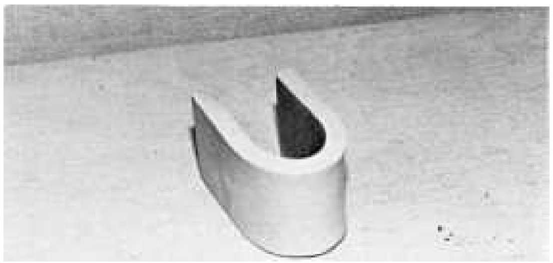

Fig. 57

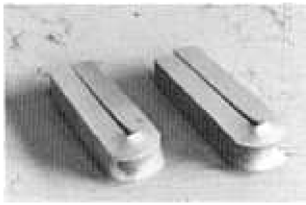

Fig. 58

The specimen in Fig. 59 shows 2 Grade #2 bolts. The one on the left was welded and it cracked. The one on the right was not welded and it did not crack.

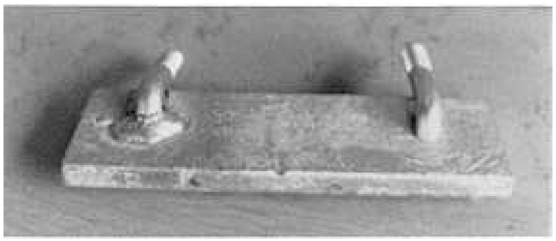

Fig. 59

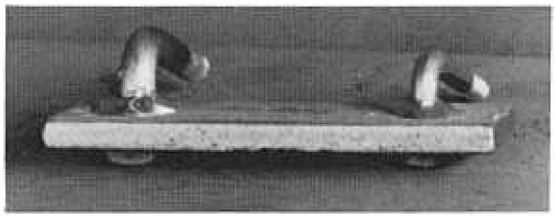

Fig. 60

The bolts in Fig. 60 are ungraded (the square head type). They were both welded. One was a 3/8", the other a 1/2" diameter. They did not crack.

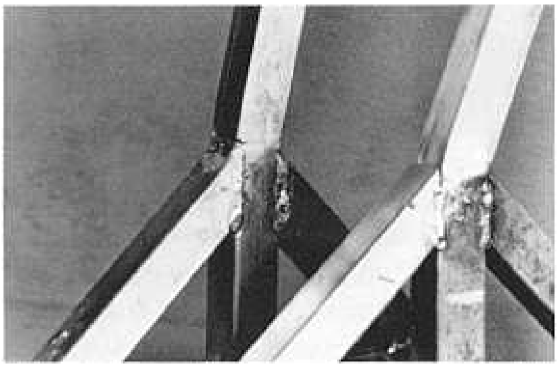

Fig. 61

The 1 inch square tubing in Fig. 61 shows that if you weld across the tubing, it will crack. If you don't weld it crossways, it doesn't crack.

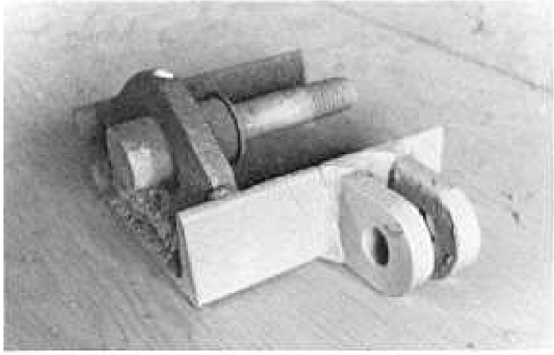

Fig. 62

Fig. 62 shows a crack adjoining a crossweld on a stress member. See a better way on page 26- Fig. 50.

Fig. 63 shows that if you weld around a bolt that comes through a hole, it may break off. Better grades of bolts have markings on top of the head. Grades above one are more likely to break. If you weld the head on the backside, it probably will not affect the bend.

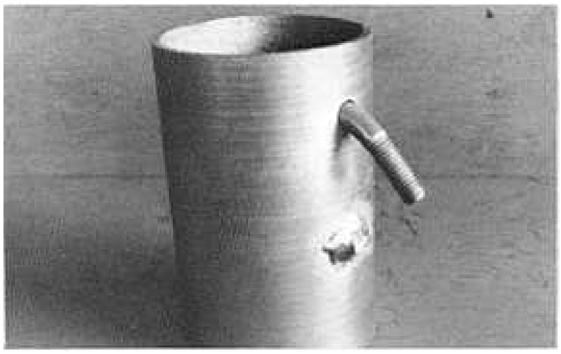

Fig. 63

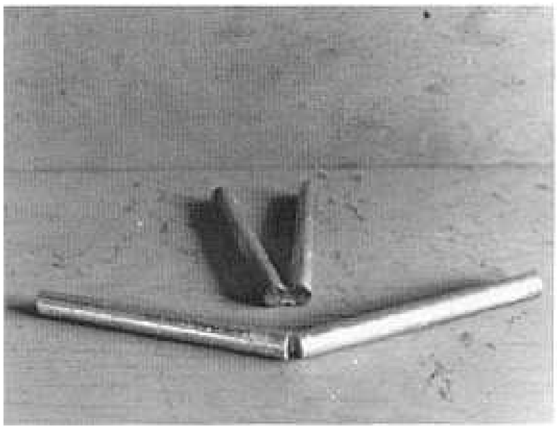

Fig. 64

The above specimens in Fig. 64 were cut from the same 1/2" bar. A stop on the saw was used to cut them the same depth.

The one on the bottom was bent cold to a 38 degree. It cracked because the only place that the metal could stretch was the width of the saw cut at the bottom of the cut.

The specimen on top was bent *cold* also. It was bent cold less than 38 degrees. It had not cracked. Then it was heated red hot and left to cool. When it was cooled to room temperature, it was slightly bent a little more. This procedure was continued until the specimen was bent on a 160 degree. It did not crack. When steel is bent cold, the crystals on the outside stretch. When it is heated to a red heat (non-magnetic), the crystals return to their original shape.

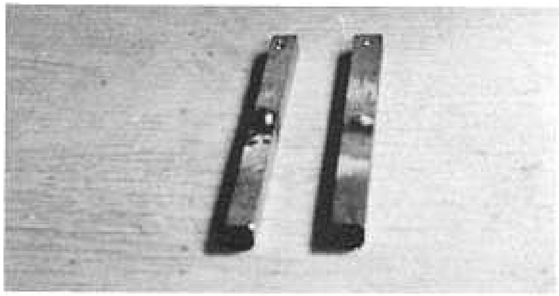

Fig. 65

A very small spot was welded on each of the 3/8" x about 1/2" specimens. The weld area did not get red hot all the way through. The one on the right was heated red hot in the weld area. The one on the left was not. See Fig. 65.

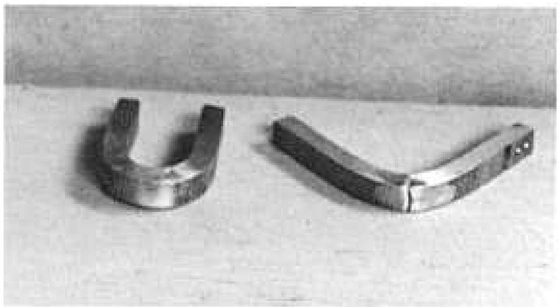

Fig. 66

They were ground and bent. The one that was reheated did *not* crack. The one that was *not reheated* cracked. See Fig. 66.

"Direction of Travel" and "Back Stepping"

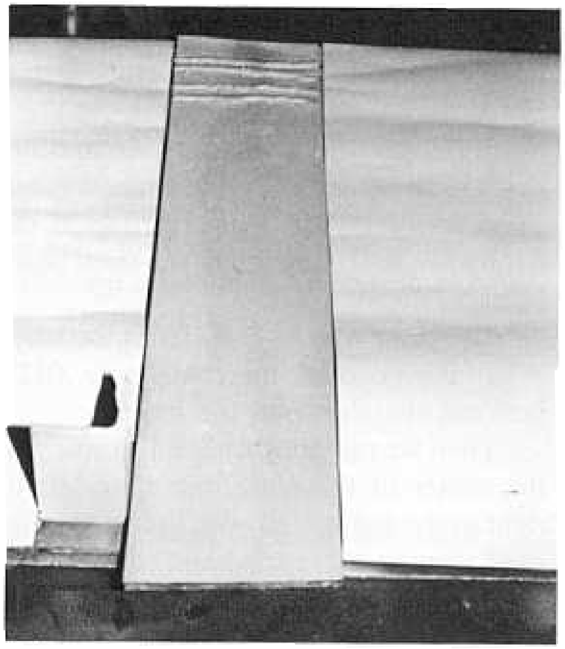

Fig. 67

Fig. 67 shows a piece of scrap steel 1/4"x 4 5/8"x 23" long was tacked (welded) on one end to a steel table. A stop was tacked against one side on the other end to show how much the "Direction of Travel" while welding, effects the other end.

One continuous bead was welded left to right across the end that was welded to the table, 1 inch from the welds that held it to the table. The stop was 19 5/16" from the cross weld. When the weld cooled, there was a space .100" or 1/10 of an inch wide between the stop and the 1/4" plate. The 1/4" plate was not cut in two. The welds were made on a solid 1/4" plate.

The stop was knocked off and again placed against the side of the 1/4" plate and tacked.

To show the results of "Back-Stepping", 4 short beads were "back-stepped" across the plate. That is, start about 1 5/32" from the right hand edge of the plate. Travel left to right to the right hand edge. Then start at about the center of the plate. Travel left to right to the first bead, fusing them together. Then another short bead running into the second bead. Then another short bead starting at the left hand edge running into the third bead. (4 back steps).

The bead was 18 1/4" from the stop. When it all cooled, there was only .012" between the stop and the plate.

Then we ran another bead, starting in the center of the plate, traveling left to right to the right hand edge. Then another bead starting at the left hand edge, traveling left to right to the center, meeting the first bead. (2 back-steps) This bead was 17 1/2" from the stop. It moved away from the stop .021". This is only 9 thousandths of an inch more than the 4 back-steps.

On the next bead, we traveled in both directions, starting at the right hand edge traveling right to left, stopping a little short of the center. (The first bead usually pulls more than the second bead.) Then starting at the left hand edge, traveling left to right, meeting the first bead in the center. This bead was 19 1/2" from the stop. It moved away from the stop .002" (2 thousandths of an inch).

I would not recommend starting in the center and traveling to the outside edges. The center would be under compressive stress, and the edges would be under tensional stress. I think it would be more likely that a crack would start on an edge that was under tension stress.

SUMMARY

	Moved	Distance from Stop
Continuous bead L to R	.100"	19 5/16"
Back-stepped 4 times	.012"	18 1/4"
Back-stepped 2 times	.021"	17 1/2"
2 beads R to L to center and L to R to center	.002"	19 1/2"

Why a Weld Bead Shrinks More Where You Stop Welding Than It Does Where You Start

The weld bead and surrounding metal behind the weld crater are hot. The metal in front of the weld crater is cold. This creates a bending force on the workpiece. The cold metal in front of the weld bead is under a compressive stress, (pushing together). As the weld crater moves forward and melts the compressed metal, the weld crater and the red hot metal around it are relaxed. The crystals are rearranged which makes the weld thicker at that point. This is "confined expansion". As soon as the crater solidifies, it starts shrinking which puts **more** compression on the cold metal in front.

Confined Expansion

A condition which resists the free expansion of a weld bead while welding. Now think back to the time that you made your test welds, were the ends of the specimen free to move in and out? Did you weld from edge to edge on the specimen? Did the speicmen get red hot all the way through? If the answers are "yes" that is "free expansion". The specimen can be bent without cracking.

If you weld a crack on an I beam, angle iron, channel iron, etc. the cold metal partly surrounding the weld will resist the expansion. This is "confined expansion".

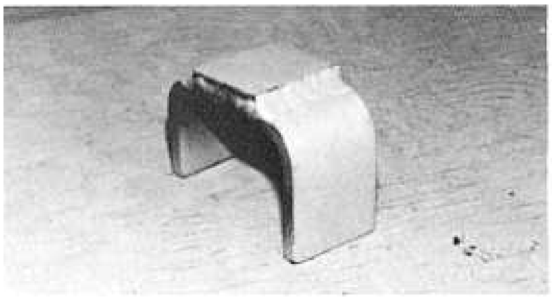

Fig. 68

The specimen in Fig. 68 shows that if it is not too thick, and welded from edge to edge, it can be bent without cracking. It is free to expand in all directions.

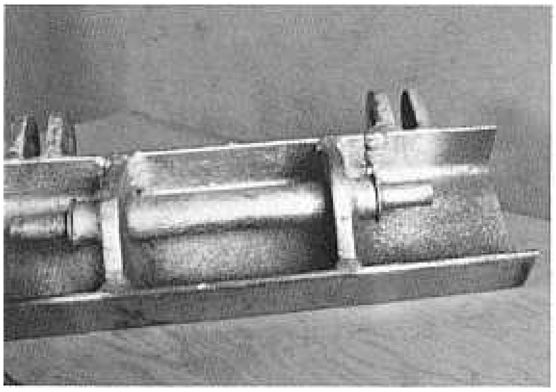

Fig. 69

The specimen in Fig. 69 shows 2 cracks on rewelds, that were welded crossways on a stress member.

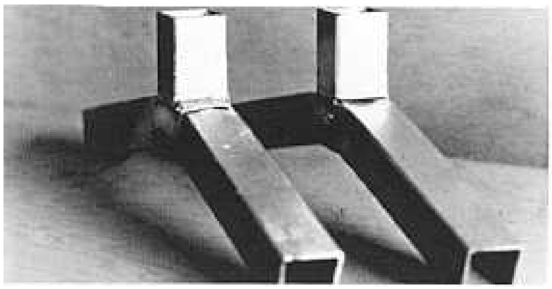

Fig. 70

Fig. 70 shows the same thing. If a crossweld is made on a 1 inch square tubing, it will crack.

The Exception to the Rule

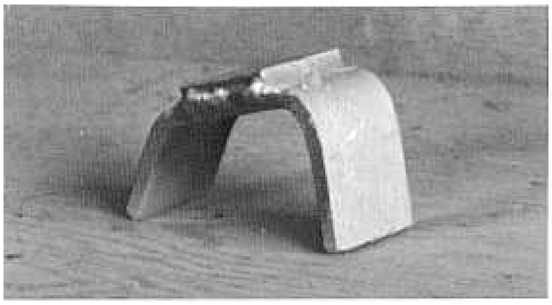

Fig. 71

The test weld in Fig. 71 above was made to show the exception to the rule. This test weld was made with no restraints. The ends were free to move outwardly as it was heated, and free to move inwardly as it cooled. The welds are not under as much tensional stress as they would have been if the ends had been welded to a heavy piece or if the patch had been welded in the center of a much larger plate. This piece can be bent without causing a crack adjoining the weld. However, I can't say what would happen under a work-load.

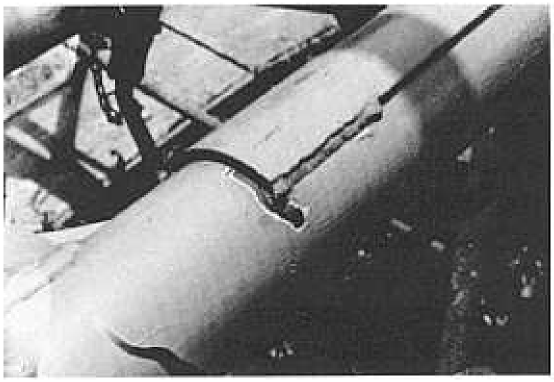

Fig. 72

The above photograph shows a crack (with magnetic particles, see Fig. 72) at the end of a 3 inch weld. The reinforcement is cut on a 90 degree angle. The reinforcement is too wide. It should have been tapered on both sides to about 1 1/2 or 2 inches wide. The 1 1/2 inch width should have been about 3 or 4 inches long and then cut on a 45 degree angle. When a reinforcement wraps half way around a piece of pipe, it makes it rigid. It should not stop abruptly. It causes a shoulder effect which can cause a crack. The shoulder effect and the shrinkage of the weld caused this crack. Both sides of the reinforcement should have been "skip-welded". This means to weld one inch and skip one inch, weld one inch and skip one inch. Repeat all the way on both sides, not the ends.

Magnaflux Testing For Cracks

Magnaflux is a brand name for an Electro magnetic device that assists in finding a crack that cannot be seen with the naked eye. It works like this — place the magnetic poles on each side of the suspected crack. Turn the electric on and spray powdered iron over the surface. The iron dust will stick to the crack. Gently blow the excess away. The magnetic particles will show the crack. **CAUTION:** Do not breathe the iron particles! If you don't have the magnetic device, a large horseshoe magnet will do. If you don't have iron powder, the sawdust from under a metal (steel) saw will do. You may have to sift it to get the smaller pieces. This method also works on a weld that does not go all the way through the work. That is, if you weld over the top of a crack and leave part of the crack below the weld, the dust will stick to the weld bead.

Another way to test for a crack is to clean the surface, pour kerosene over the surface, wipe dry with a clean rag, and then cover with blackboard chalk. The crack will fill with kerosene and when the chalk is applied, the kerosene will wet the chalk to show the crack. You can also buy a kit that works the same way. The kit contains a red liquid dye, and the chalk is brushed on in liquid form.

Water Tank

I remember, back in about 1948, a friend of mine insisted that I help him straighten out a tank that had been screwed up by two beginners. The tank was about 18 or 20 feet in diameter and about 10 or 12 feet high. The 1/4 inch plates had already been rolled to the correct radius. (As you know, the ends of a rolled plate are straight for the distance between the centers on the rolls. They should have been sheared off.) But the beginners had already welded together the bottom of the tank and the first layer of the vertical plates. When they started with the second layer of the plates, (the seams were staggered) the flat places on the ends of the bottom plates took a short cut across the radius of the top plate and left a wide gap horizontally between the top and bottom plates. With a sledge hammer, we drove the seams out to meet the radius on the top plate.

The beginners had already welded some of the second layer plates solidly, except the gap. I remember one place where they had tacked a plate in place and then welded the end vertically from bottom to top, to the end of the other plate. The vertical weld pulled together so much at the top that it broke the tack on the other end and lifted the plate up on the other end, making a gap too large to be welded. (See "Direction Of Travel", pg. 33, fig. 67.) I started welding horizontally from the bottom of their vertical weld toward the loose end. I would weld a couple of inches, then wait for it to cool. (The plate was 10 or 12 feet long.) Weld a couple more inches and wait for it to cool. This method was continued until the short welds pulled the other end back down to the proper place. The end was tacked again and the rest of the weld was "back-stepped" (See pg. 33, fig. 67.) to reduce some of the stresses.

If you have a long bead to run, it is best to back-step it. Back-stepping will not relieve all of the stress, but it will help.

Back-Stepping

Start a bead 6" from the edge of a plate. Travel to the edge of the plate. Start another bead 12" from the edge and weld toward your first bead and into it. Then back up 6" from where you started the last time, each time.

Broken Trailer Frame

I had a truck driver who had a trailer frame welded in Canfield, Ohio. He drove it to Knoxville, Tennessee, and it was broken when he arrived there. When he arrived in Chattanooga, it was broken again. Both welders had welded the ends of the reinforcements. Both breaks were adjoining the end welds. The reinforcements were removed and replaced with a longer one. The ends were *not* welded. It did *not* break again.

Most people agree that: As it cools, every weld shrinks from all directions. When a weld shrinks, the adjoining metal is forced to stretch or crack. If the metal is stretched, it is thinner. Is thin metal as strong as thicker metal?

Some welding instructors teach the beginners to weld crossways on a stress member. They bend a test weld and it doesn't break. This is explained in "The Exception To The Rule" on page 35, fig. 71.

Air Lifts

Years ago, I had some air-lifts installed on my car. They were rubber bags that fit on the inside of the coil springs on the rear of the car. The rear end could be overloaded and then add air pressure to the bags to raise the car up. They worked really well. There was a rubber hose from the bags to the trunk of the car. The air pressure could be adjusted without going under the car. Then one day, one of the bags went flat. It was taken off and cut open to see what caused the crack. There was a reinforcement molded to fit the inside and probably vulcanized to it. The reinforcement was about 1/8 of an inch thick and where it stopped on the sidewall, it made a sharp shoulder. That is where it cracked. If a shoulder will cause a crack on rubber, what will it do to steel?

Subjects of Non-Welding

The following subjects do not pertain to welding. However, a good welder is expected to perform many tasks other than welding. Welding is not a trade, but a sideline of many trades, such as: machinist, iron worker, carpenter, electrician, pipe fitter, etc.

We hope some of you will benefit from this effort. At least, some of them are interesting.

Cover A Padded Bar With Upholstery

This method is for problem applications, where you have a thousand wrestlers a day trying to twist them off. Use a **sponge** rubber pad that is tough (foam will disintegrate). The holes in foam rubber are connected. You can blow smoke through it, but not in sponge.

Glue the pad on to the bar solidly. Leave a 1/4" gap where the sides of pad come together. Try to place the gap out of sight. The gap will contain the loops and wires, as you will see later.

The padding should be wrapped with strips of cloth to hold it in place until the glue dries. When the glue dries, remove the cloth and determine how wide the cover should be to fit around the padding correctly. Cut a strip of your material 1 inch wide and 2 inches or more longer than the circumference around the padding. Fold 1 inch back and staple it with a paper stapler. Wrap it around the padding with the other end folded back loosely. Bring them together at the gap in the padding. Slip the loose end until the loops or folds touch. This measurement is important. If you like, you may try a piece of scrap material first.

With the material cut to exact size, and sewed lapped over 1 inch on the sides and 1/2 inch on the ends, as shown below fold together, place one end of both sides in the vice and cut both pieces 1/4 inch deep every 1/2 inch, as shown above.

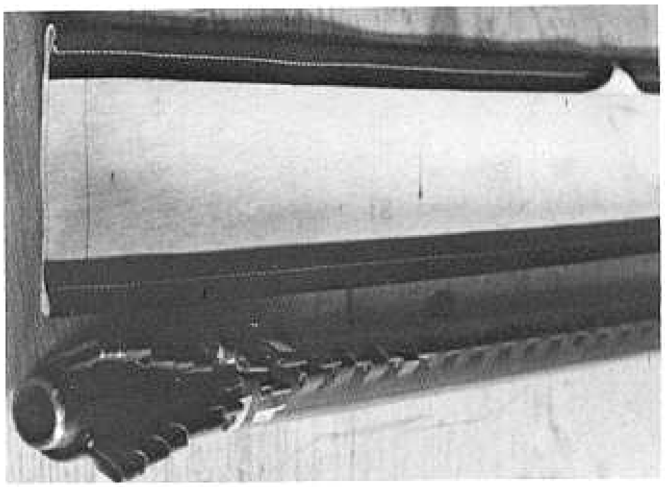

Fig. 73

Apply glue to padded bar, all except 1/2 inch on each side of the gap. If you get glue on the loops, you may not be able to push the wires through. In applying the glue to the cover, raise the flap on the seam up and brush the glue all the way back to the threads, then when your threads rot out, the glue will still be holding.

In placing the cover on the padded bar, line the loops up so that the wire which will be on the inside of the loops will be centered over the gap in the padding.

With the cover glued on, take a piece of 3/32 acetylene welding rod and file one end smooth and bend it 1/2 inch from the end about 1/8 inch off center. Then bend about 2 inches on the other end on a 90 degree for a handle. Push the welding rod through the first loop on the left hand side, then through the second loop on the right hand side, third loop on left, etc.

When you push the wire through one loop, turn the wire about a half turn so that the bent end points to the next loop on the other side. It will start through the loop much easier, and it will pull the cover tight. Push the wire all the way through and out the other end. Then push another welding rod through the remaining loops the same way.

On the ends, punch a hole in the cover so you can run a wire around the ends on the inside of the hem. Now go back to the wires that go through the loops. Cut the wires off and file the sharp edges off and with a pair of needle nose pliers, bend the wires and tuck them under the padding. Then twist the wires together that go around the ends, cut them off, file the sharp edges off and tuck them under the padding. Don't twist them too tight. Leave room to tuck them under.

It is important that you cut the material lengthways with the roll instead of crossways — it will stretch more crossways.

When you cut the material, cut it 1 inch longer and 2 inches wider to allow for the hems. Make a straight line 2 inches in from the sides and 1 inch in from the ends. All lines and cuts should be straight. This gives the one who does the sewing something to lap the edge of the material to, which makes the width correct.

Hooks Or Eyes On Spring

To make the hooks or eyes on an expansion spring, place spring in vise vertically, with one round of the wire above the jaws. Use a pair of duck bill pliers to twist the top round straight up. You may have to hold the second round down with a screwdriver. Now the hook is standing straight up and you have not bent the wire. You have twisted it a quarter turn as shown on one end below and your hook is on one side of the spring. To center the hook, twist it one quarter turn to the center of the spring as shown above. This method does not bend the spring steel wire, it only twists, it. You can also use this method to replace a broken off hook or eye.

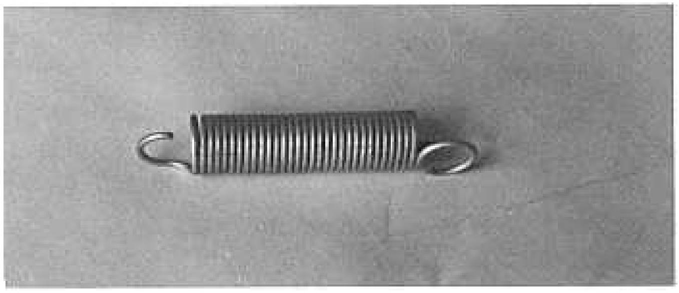

Fig. 74

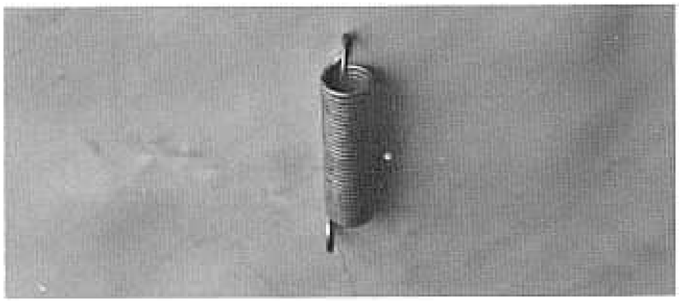

Fig. 75

Make A Coil Spring

In an emergency, you can make a coil spring. You will need some spring steel wire of the correct size. You will need a mandrel the right size.

Drill a hole slightly larger than your wire, 1/4 inch from the end. Then saw a slot with a hacksaw, slightly larger than your wire, a few degrees clockwise of your hole as shown below.

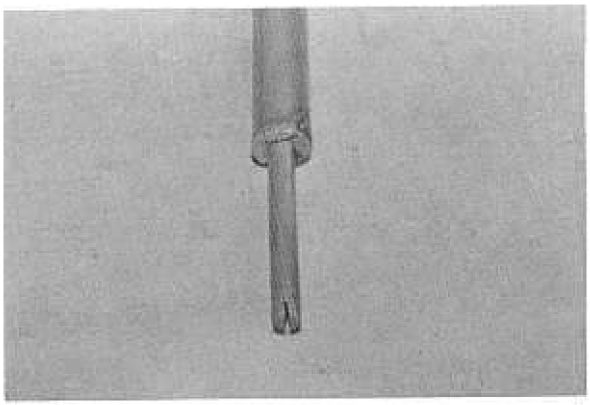

Fig. 76

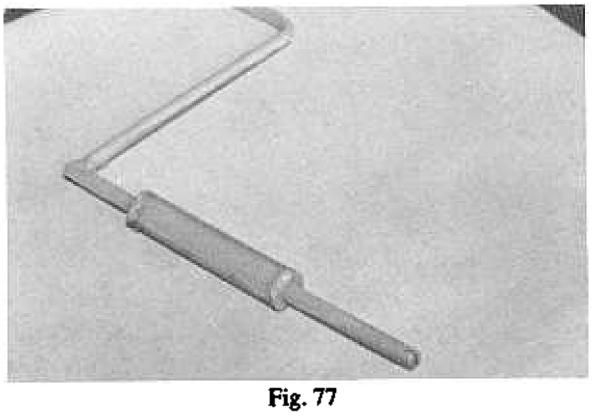

Fig. 77

If you don't have a lathe, cut a piece of pipe 2 inches longer than the jaws of your vise. Weld a washer on each end of the pipe with holes drilled in the washers to fit the mandrel. Weld a handle on the other end of your mandrel to turn it with as shown above. You will need someone else to turn the crank with the pipe held in the vise. Place the end of your spring wire in the slot and bend it. With the mandrel turning, pull tightly on the wire. For an expansion spring, wind the wire close together. For a compression spring, lead the wire away from the last round so the spring will have room to compress. When you finish, turn the crank the other way to relieve the tension. The spring will be much larger in diameter than the mandrel.

Cooling Fin

I was once told by an engineer that a 7" aluminum cooling fin placed on a 4" steel hydraulic cylinder, would contract when heated, which would make it tighter on the cylinder. I didn't believe it. Aluminum expands more than steel. Then I asked another engineer. He agreed with the first one. He explained that it would expand both ways from a point slightly to the inside of the center between the two circumferences, and it would tighten on the cylinder. I still didn't believe it, so I pressed one fin onto a cylinder. With the cylinder standing vertically, heat was added to the inside of the cylinder. When the fin got hot, it dropped. I think the aluminum expanded in **all** directions, causing the hole to be larger as well as the outside circumference. After all, the sides of the fin are only 1 1/2 inches and the circumference is over 21 inches.

Shaft Will Break At Shoulder

Figure 78 shows a broken shaft at the end of a rod end. It broke for two reasons: partly because the threads are the weakest place and partly because the rod end was not tapered enough to give. It broke off right at the end of the lock nut. That is a notch effect.

The same thing happens when you saw a half-inch rod with a hacksaw, leaving only 1/8 inch holding. Bend it and it breaks off. That is because the metal can only stretch in one place, the width of your saw blade. Take a 1/8 inch welding rod and try to break it. It bends in a much larger radius.

Years ago electric irons came with a drop cord the same size all the way. With all the flex required for ironing clothes, the cords inevitably broke off right at the plug end that attaches to the iron. A heavy rubber sheath, tapered out six or eight inches caused the cord to bend in a much larger radius and the breaking stopped.

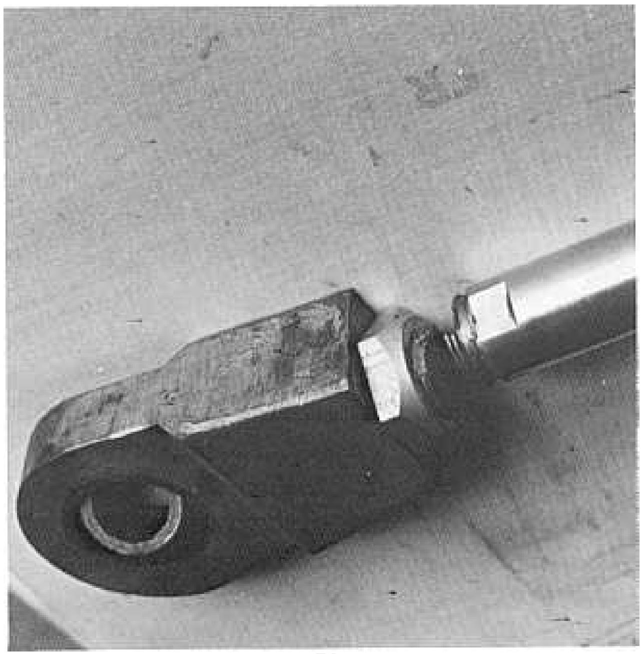

Fig. 78

The same principle applies in both cases.

In 20 years, I had 3 rods to break as shown in Fig. 78 (Shaft will break at shoulder).

Then I made new ones, 1/8 of an inch in diameter larger. To avoid cutting as deep into the shaft, I used 18 threads per inch instead of 12. At the end of the threads, the shaft was turned down about 1/2 inch in length to the size of the bottom of the threads to let it bend in a larger radius. Then it was filed, sanded and polished to remove any cross scratches that might start to crack. They have been in service about 6 years with no failures.

How To Temper Small Parts

Using water hardening drill rod, cut to desired lengths. Heat them to a red heat and plunge them into water. This makes them hard and brittle. Take them out of the water and sand the oxide off so you may be able to see the heat colors. Heat a piece of heavier steel to a red heat, and place the parts on top of it. Watch the heat colors. If you quench them when they first turn dark blue, they will be hard enough to break. If you want them to be soft enough to bend, wait a few seconds until they turn light blue then quench them.

Rod Ends

If you want to pin a threaded rod end to a rod to prevent it from turning, and the rod-end and pin still be interchangeable with the other rods, here's how. Weld a piece of steel onto a nut to hold a hardened drill bushing. The hole for the pin should be near the end of the rod. Screw the nut onto the rod until it is flush with the end of the rod. Then place the rod-end against the rod and start turning the rod-end, nut and drill bushing all together. There may be a space between the nut and the rod-end. If the space is wider than the distance between two threads, take it off. You held the rod-end crooked and didn't catch the first thread. Screw the rod-end on to the desired depth and drill hole through drill bushing into and out the other side of the rod-end and rod.

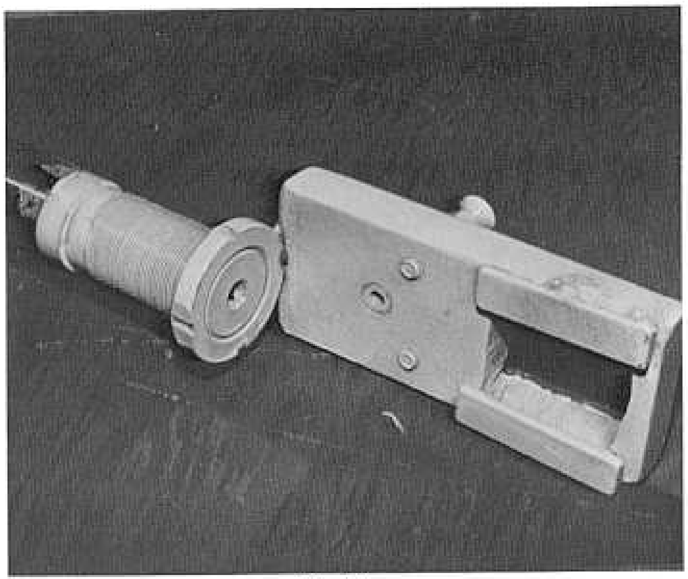

Fig. 79

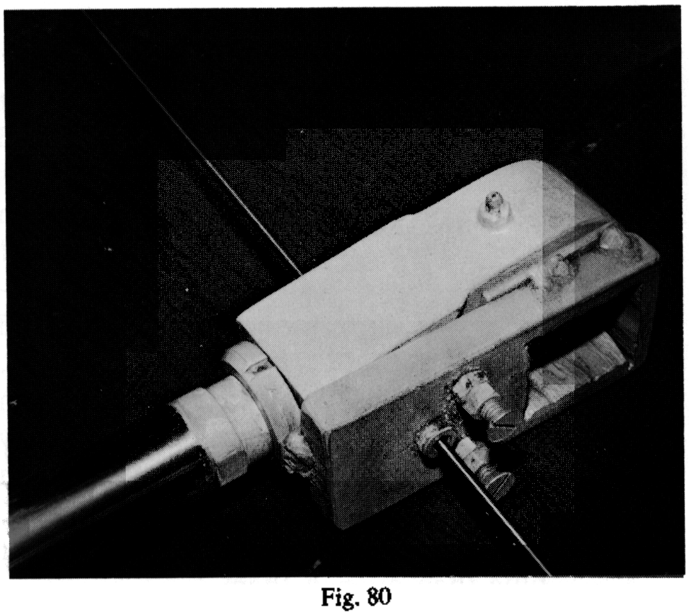

Fig. 80

It is important to have a lock-nut **tightly** against the rod end. I had a nut get loose and wore the threads on the rod-end too much. I shrank the rod-end at the threads (as explained in "How To Shrink A Steel Pulley") a little too much. I had to tap it out to get it on.

Find Diameter of Numbered Machine Screws

Multiply number of screw by .013 then add .060.

Example: The screws that hold the cover plates on your electric wall switches at home are 6-32.

$$\begin{array}{r} .013 \\ \times 6 \\ \hline .078 \\ +.060 \\ \hline .138 \end{array}$$

Diameter of screw is 138 thousandths of an inch

Which Way To Turn A Turnbuckle

Look at the two threads on this turnbuckle. If the threads on one of them look like they are slanted more than the other ones, they are the left-handed threads. Facing either end of the bolt, turn a left-handed nut counter clockwise to move it away from you. Turn a right-handed nut clockwise to move it away from you.

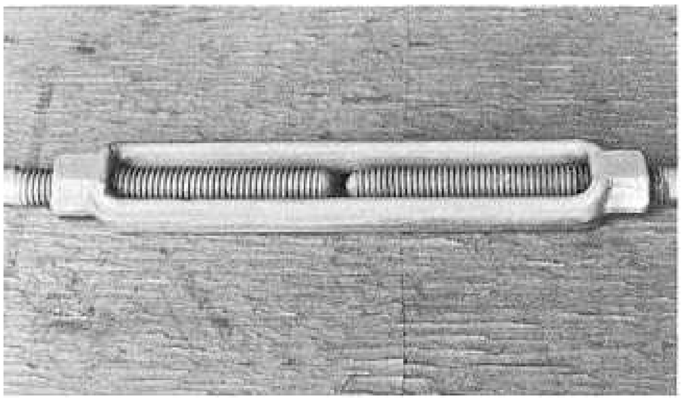

Fig. 81

Dish In Wheel

Almost every wheel has a dish in it. If the hub was not off-center to the rim, it would be like a square tank to hold pressure. A flat surface bends more than a curved surface when pressure is applied. A welder once asked me if I could build a square hot water tank. I said, "yes, but it would take a lot of stay-bolts to prevent the sides from bulging." He said he built one and the next morning it was almost round. Same thing goes for a retaining wall. A straight wall sometimes falls over or breaks. A curved or zig-zag wall is much stronger.

Try to stand a flat card on edge. Then bend it and see the difference.

To make a flat frame rigid, cross-brace from corner to corner. Where the braces cross in the center, they must be bent to make that point higher than the frame. The cross must be bolted or welded to prevent sliding.

Blowing Grip On Bar

The picture below shows a steel bar 3/4" in diameter. The plastic grip has a 5/8" hole in it. The grip was forced on with air pressure. Notice that the grip is not on all the way. The end of the air nozzle hit the end of the bar and stopped. This method may be used to cover a pipe with rubber or plastic tubing. The end of the pipe must be sealed and one end of the rubber tubing fastened to the air nozzle air tight. The rubber tubing will slide over the pipe on a layer of escaping air.

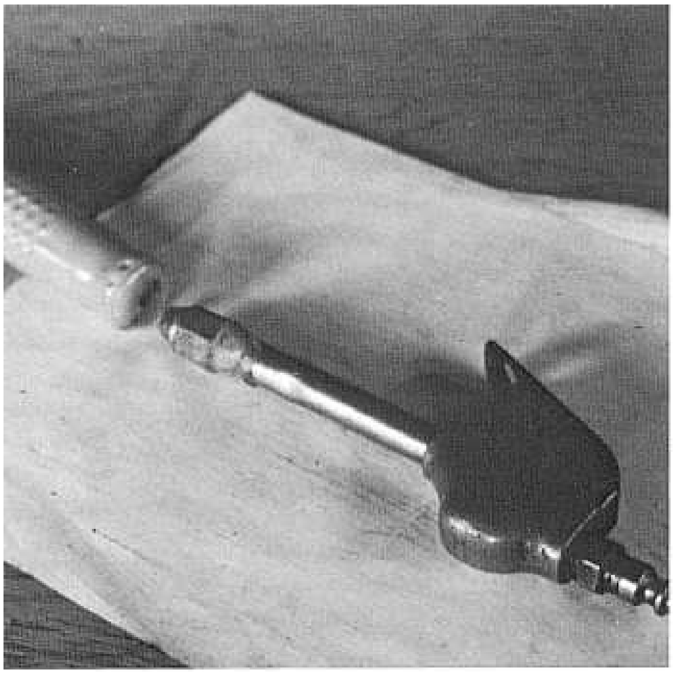

Fig. 82

How To Find A Tap Drill Size Without A Chart

Example: Find a drill size for 1/2 inch 13 thread tap. Divide number of threads per inch into one inch.

$$\begin{array}{r} .0769 \\ 13\,\overline{)1.000} \\ \underline{91} \\ 90 \\ \underline{78} \\ 120 \\ 117 \end{array}$$

Then subtract .0769 from tap size — $\frac{1}{2''}$ = .500

$$\begin{array}{l} .500 \\ \underline{.0769} \\ .4231 \ = \text{drill size} \end{array}$$

$\frac{1}{64''}$ = .0156 — then divide .0156 into .4231

$$\begin{array}{r} 27 \\ .0156\,\overline{).4231} \\ \underline{312} \\ 1111 \\ \underline{1092} \\ 19 \end{array} \quad \frac{27}{64} \quad \text{OR} \quad \begin{array}{r} .4231 \\ \underline{64} \\ 16924 \\ \underline{25386} \\ 27.0784 \end{array} \quad \frac{27}{64''}$$

In other words: divide the number of threads per inch into one inch and subtract the answer from the tap size. Then find the nearest standard drill size. A few thousandths does not matter.

Finding Half Of A Mixed Common Fraction

Sometimes it is necessary to find the center of something. That is — find half of the total length.

Example: Whole number **ODD**

$$5\frac{1}{2} \div 2 = 2\frac{3}{4}$$

DISCARD the odd whole number making the 5 a 4. Divide 4 by 2 and place here.

Multiply the denominator by 2 and place here.

Add the numerator and the denominator together and place here.

Example: Whole number **EVEN**

Divide the whole number by 2 and place here

$$4\frac{1}{2} \div 2 = 2\frac{1}{4}$$

Multiply denominator by 2 and place here.

MOVE numerator to here.

In other words: Take the odd whole number and throw it away — forget about it. That changes the 5 into a 4. Take half of the 4 and write it down, 2 - with a dash to the right of it. Then add the top number and the bottom number of the fraction together and place it above the dash 2 $\underline{3}$, then double the bottom number of the fraction and place it below the dash this way: $2\frac{3}{4}$.

$$37\frac{21}{32} \div 2 = 18\frac{53}{64}$$

If someone tells you something that you don't already know, don't even think about making a defensive statement. You are not supposed to know everything. Nobody else does. Check it out and see Who Is Right.

However, you can't believe everthing you see, hear, or read.

Facts

Nickel is magnetic.

U.S. Paper Currency is slightly magnetic in places. When you turn the steering wheel on your car to the left, the left front wheel turns more than the right one does, and vice versa.

The file and scratch test is a quick procedure for checking hardness. Shopmen frequently use this method to predict the machinability of a metal. It consists of simply trying to scratch or cut the surface of a metal with a file or pointed object of known hardness. All scratch tests reveal only a superficial or outer skin hardness. They tell us nothing about the hardness ⅛ in. below the surface.

Table 3-2 gives a relation between Brinell hardness and the hardness as estimated with a machinist's new hand file.

Table 3-2. File Hardness Test Data

BRINELL HARDNESS	FILE ACTION
100	File bites into surface very easily
200	File removes metal with slightly more pressure
300	Metal exhibits its first real resistance to the file
400	File removes metal with difficulty
500	File just barely removes metal
600	File slides over surface without removing metal. File teeth are dulled.

Simple Hardness Tester

Welders sometimes make their own set of hardness testers, by heat treating ¼-in. high carbon steel electrodes to give a graduated series of hardnesses.The rods are labeled with their respective Rockwell C hardnesses. A sharp point is then ground on the end of each rod, taking care not to soften the steel by overheating.

By drawing these scratch testers, one at a time, over a smooth metal surface, the welder can determine which rod has a hardness about the same as the metal being scratched. If the sharp point of the scratch tester digs in and leaves a scratch, its hardness is greater than that of the metal being tested. If the point only slides over the surface without scratching, the scratch tester is softer than the metal tested.

The hardness set is inexpensive and is convenient to use in hard-to-reach places, where a conventional test would be impossible.

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Fusibility: The Ease of Melting

Fusibility is a measure of the ease of melting. Mercury (the metal with the lowest melting point) melts at -38 degrees Fahrenheit, while tungsten, which has the highest melting point, melts at 6,100 F.

A pure metal has a definite melting point, which is the same temperature as its freezing point. Alloys and mixtures of metals, however, have a temperature at which melting starts and a higher temperature at which the melting is complete.

Figure 3-9 gives the melting point of a few metals and other temperatures of interest.

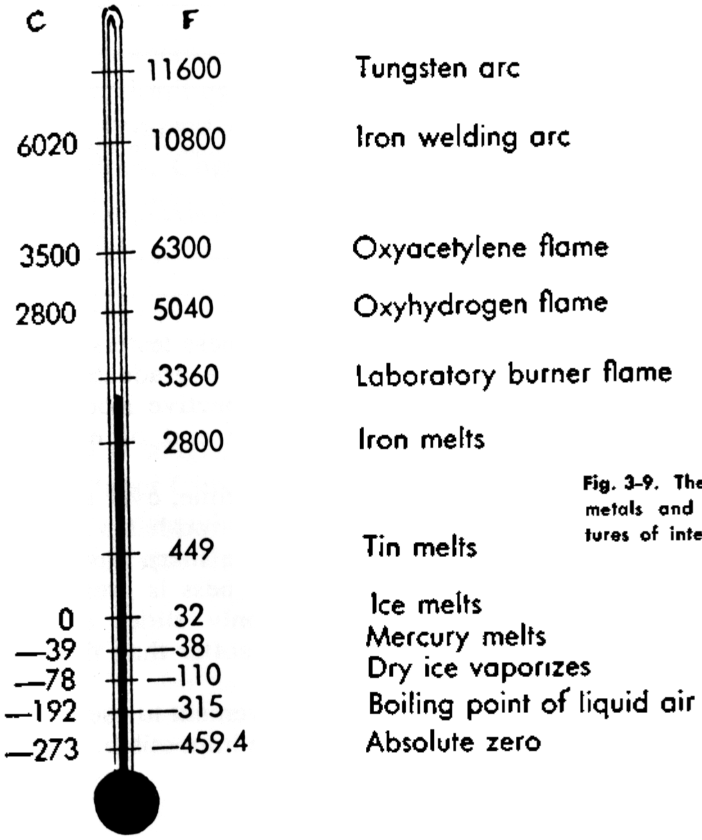
C
F
11600 Tungsten arc
6020 10800 Iron welding arc
3500 6300 Oxyacetylene flame
2800 5040 Oxyhydrogen flame
3360 Laboratory burner flame
2800 Iron melts
449 Tin melts
0 32 Ice melts
—39 —38 Mercury melts
—78 —110 Dry ice vaporizes
—192 —315 Boiling point of liquid air
—273 —459.4 Absolute zero

Fig. 3-9. The melting points of some metals and a few other temperatures of interest for comparison.

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Heat of Fusion

The *heat of fusion* is the quantity of heat necessary to change one pound of a solid material to a liquid without temperature change.

The British thermal unit (Btu) is used to measure the quantity of heat; for all practical purposes, it is the amount of heat required to raise the temperature of 1 pound of water 1 degree Fahrenheit.

The heat of fusion of ice is 144 Btu per lb. In comparison, here are the heats of fusion of a few metals:

Table 3-3. Heat of Fusion of Metals

Aluminum	170 Btu per lb	Tungsten	79 Btu per lb
Magnesium	160 " " "	Silver	45 " " "
Chromium	136 " " "	Zinc	43 " " "
Nickel	133 " " "	Gold	29 " " "
Molybdenum	126 " " "	Tin	26 " " "
Iron	117 " " "	Lead	11 " " "
Manganese	115 " " "	Mercury	5 " " "
Copper	91 " " "		

More heat is required to melt a pound of ice than is needed to melt a pound of iron (provided of course that both materials are at their melting point). The figures are 144 Btu for water and 117 Btu for iron. It takes more heat, however, to melt a pound of aluminum than a pound of ice—170 Btu as against 144 Btu.

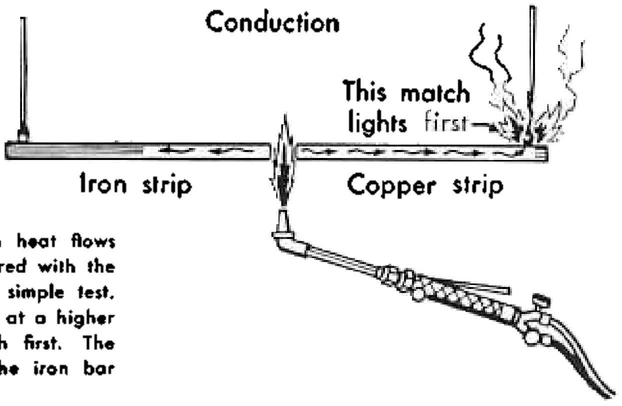

Fig. 3-10. The rate at which heat flows through copper can be compared with the same property in iron by this simple test. The metal which conducts heat at a higher rate (copper) lights the match first. The match at the outer end of the iron bar will burst into flame later.

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Thermal Conductivity

Thermal conductivity is a measure of the rate at which heat will flow through a material. The difference in thermal conductivity between iron and copper is easily demonstrated, Fig. 3-10. The copper conducts heat much faster than does the iron.

If one end of a copper bar is kept in boiling water (212 F) and the other end in chipped ice, heat will flow into the bar from the water, then through the bar and to the ice, causing it to melt. The rate at which the ice melts indicates the rate heat is flowing through the bar.

The amount of ice that melts depends upon:

1. Time	The longer the time, the more ice will melt.
2. Size of the bar	The larger the cross-sectional area of the bar, the more heat will flow.
3. Length of the bar	The shorter the bar, the faster the ice will melt.
4. The temperature to which the bar is heated	The higher the temperature of the hot end of the bar, the faster the ice will melt.
5. Thermal conductivity	The higher the thermal conductivity of the bar, the more heat will flow.

The amount of heat flow, therefore, depends upon time, area, length, temperature difference and thermal conductivity.

Table 3-4. Thermal Conductivity of Metals

Metal	Chemical Symbol	Btu/sq ft/in. hr/°F	Relative Conductivity Based on Silver as 100%
Aluminum	Al	1428	49.7%
Copper	Cu	2664	92.7
Gold	Au	2037	70.9
Iron-Pure	Fe	467	16
Iron-Steel		313	10.9
Iron-Cast		316	11
Lead	Pb	241	8
Mercury	Hg	476	17
Molybdenum	Mo	1004	34.9
Nickel	Ni	412	14.3
Platinum	Pt	483	16.8
Silver	Ag	2873	100
Tin	Sn	450	15.6
Tungsten	W	1381	48
Zinc	Zn	770	27

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Thermal Conductivity

Thermal conductivity is a measure of the rate at which heat will flow through a material. The difference in thermal conductivity between iron and copper is easily demonstrated, Fig. 3-10. The copper conducts heat much faster than does the iron.

If one end of a copper bar is kept in boiling water (212 F) and the other end in chipped ice, heat will flow into the bar from the water, then through the bar and to the ice, causing it to melt. The rate at which the ice melts indicates the rate heat is flowing through the bar.

The amount of ice that melts depends upon:

1. Time	The longer the time, the more ice will melt.
2. Size of the bar	The larger the cross-sectional area of the bar, the more heat will flow.
3. Length of the bar	The shorter the bar, the faster the ice will melt.
4. The temperature to which the bar is heated	The higher the temperature of the hot end of the bar, the faster the ice will melt.
5. Thermal conductivity	The higher the thermal conductivity of the bar, the more heat will flow.

The amount of heat flow, therefore, depends upon time, area, length, temperature difference and thermal conductivity.

Table 3-4. Thermal Conductivity of Metals

Metal	Chemical Symbol	Btu/sq ft/in. hr/°F	Relative Conductivity Based on Silver as 100%
Aluminum	Al	1428	49.7%
Copper	Cu	2664	92.7
Gold	Au	2037	70.9
Iron-Pure	Fe	467	16
Iron-Steel		313	10.9
Iron-Cast		316	11
Lead	Pb	241	8
Mercury	Hg	476	17
Molybdenum	Mo	1004	34.9
Nickel	Ni	412	14.3
Platinum	Pt	483	16.8
Silver	Ag	2873	100
Tin	Sn	450	15.6
Tungsten	W	1381	48
Zinc	Zn	770	27

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Thermal Expansion

Thermal expansion is the increase in dimensions of a body due to a change in its temperature. See Fig. 3-11 for thermal expansion of selected materials.

The coefficient of linear expansion is the ratio of the change in length of a material, caused by heating it one degree, divided by the original length.

The coefficient of linear expansion of iron at room temperature is 0.0000065 per degree F. (6.5 x 10^{-6}/°F).

The total increase in length of an iron bar 100 feet long which is heated from 10 F to 110 F will be:

0.0000065 x (110-10) x 100 = 0.065 ft or 0.78 in.

The coefficient of cubical expansion equals approximately three times the coefficient of linear expansion.

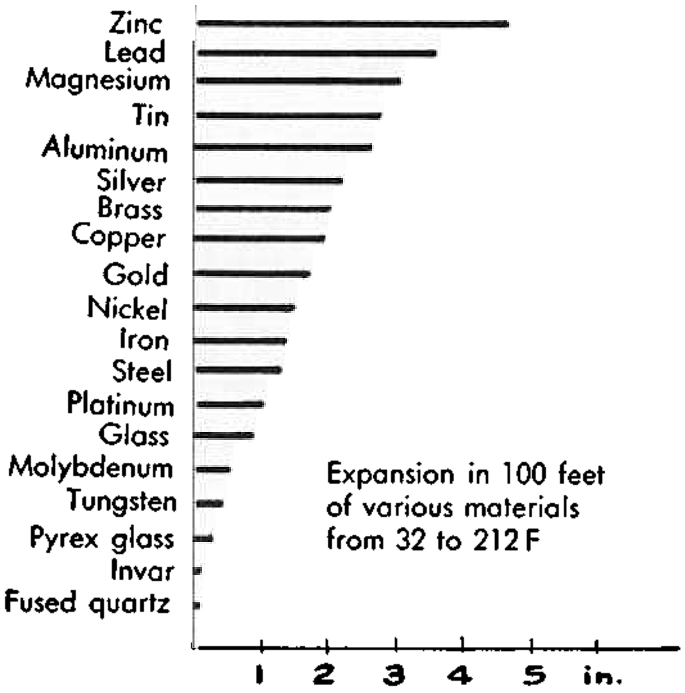

Fig. 3-11. Linear thermal expansion (in in.) of 19 materials when heated from 32 F to 212 F.

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Table 4-1. Typical Ferrous Materials

MATERIAL	SPEC.	CHEMICAL ANALYSIS			MECHANICAL PROPERTIES			
		CARBON	OTHERS		YIELD STRENGTH	TENSILE STRENGTH		
Cast Iron, Gray, Grade 20	ASTM A48-56	3.00-4.00				20,000		163
Gray Grade 30	ASTM A48-56	3.00-4.00				30,000		180
Nickel		2.00-3.50	Ni 0.25-0.50			40,000		310
Chrome-Nickel		2.00-3.50	Ni 1.00-3.00	Cr 0.50-1.00		53,000		510
White		2.00-4.00	Si 0.80-1.50			46,000		420
Malleable	ASTM A47-52	1.75-2.30	Si 0.85-1.20		35,000	53,000	18	140
Iron, Wrought, Plates	ASTM A42-55	0.08	Si 0.15	Slag 1.20	26,000	46,000	35	105
" " Forgings	ASTM A75-55	0.01-0.05	Fe 99.45-99.80		25,000	44,000	30	100
Steel, Cast, Low Carbon		0.11	Mn 0.60	Si 0.40	35,000	60,000	22	120
" " Medium Carbon		0.25	Mn 0.68	Si 0.32	44,000	72,000	18	140
" " High Carbon		0.50			40,000	80,000	17	182
Steel, Rolled, Carbon	SAE 1010	0.05-0.15			28,000	56,000	35	110
" " Carbon	SAE 1015	0.15-0.25			30,000	60,000	26	120
" " Carbon	SAE 1025	0.20-0.30			33,000	67,000	25	135
" " Carbon	SAE 1035	0.30-0.40			52,000	87,000	24	175
" " Carbon	SAE 1045	0.40-0.50			58,000	97,000	22	200
" " Carbon	SAE 1050	0.45-0.55			60,000	102,000	20	207
" " Carbon	SAE 1095	0.90-1.05			100,000	150,000	15	300
" " Nickel	SAE 2315	0.10-0.20	Ni 3.25-3.75		90,000	125,000	21	230
" " Ni.-Cr.	SAE 3240	0.35-0.45	Ni 1.50-2.00	Cr 0.90-1.25	113,000	136,000	21	280
" " Moly.	SAE 4130	0.25-0.35	Cr 0.50-0.80	Mo 0.15-0.25	115,000	139,000	18	280
" " Cr.	SAE 5140	0.35-0.45	Cr 0.80-1.10		128,000	150,000	19	300
" " Cr.-Va.	SAE 6130	0.25-0.35	Cr 0.80-1.10	Va 0.15-0.18	125,000	150,000	18	310
" " Si.-Mn.	SAE 9260	0.55-0.65	Mn 0.60-0.90	Si 1.80-2.20	180,000	200,000	12	390

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Steels in the low-carbon group are generally tough, ductile and easily formed, machined and welded. Although low-carbon steels are not especially hard, some grades respond to heat treatment and are readily case-hardened by carburizing, cyaniding, flame-hardening, etc.

Table 4-2. Uses for Steel by Carbon Content

CARBON CLASS	CARBON RANGE %	TYPICAL USES
Low		Chain, nails, pipe, rivets, screws, sheets for pressing and stamping, wire.
		Bars, plates, structural shapes.
Medium		Axles, connecting rods, shafting.
High	0.45 - 0.60	
	0.60 - 0.75	
Very High	0.75 - 0.90	Chisels, punches, sand tools.
	0.90 - 1.00	Knives, shear blades, springs.
	1.00 - 1.10	Milling cutters, dies, taps.
	1.10 - 1.20	Lathe tools, woodworking tools.
	1.20 - 1.30	Files, reamers.
	1.30 - 1.40	Dies for wire drawing.
	1.40 - 1.50	Metal cutting saws.

Medium-carbon steels—have a carbon range of 0.30 to 0.45%. They are strong, hard and not so easily forged or welded as low-carbon steels. If extensive welding is to be done on medium-carbon steels, it is desirable to use those steels at the lower end of the carbon range—between 0.30 and 0.35% carbon. As the carbon content increases above 0.35%, the steel becomes increasingly difficult to weld since there is a greater tendency toward brittleness of the weld. Special electrodes and procedures are often necessary to prevent weld cracking.

High-carbon (0.45 to 0.75%) and **very-high-carbon** (0.75 to 1.5%) **steels**—are very strong and hard. Both properties increase with an increased carbon content.

While carbon has an important influence on the characteristics of

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

steel, the degree to which impurities are not removed by refining is also important. A slight increase in the percentage of phosphorus or sulphur will materially lower the ductility, malleability, fatigue and shock resistance and welding qualities of a steel.

High- and very-high-carbon steels respond well to heat treating. Nearly any degree of hardness, temper or strength may be obtained. In the annealed state, most of these materials may be readily machined. They also may be hot-worked for forming.

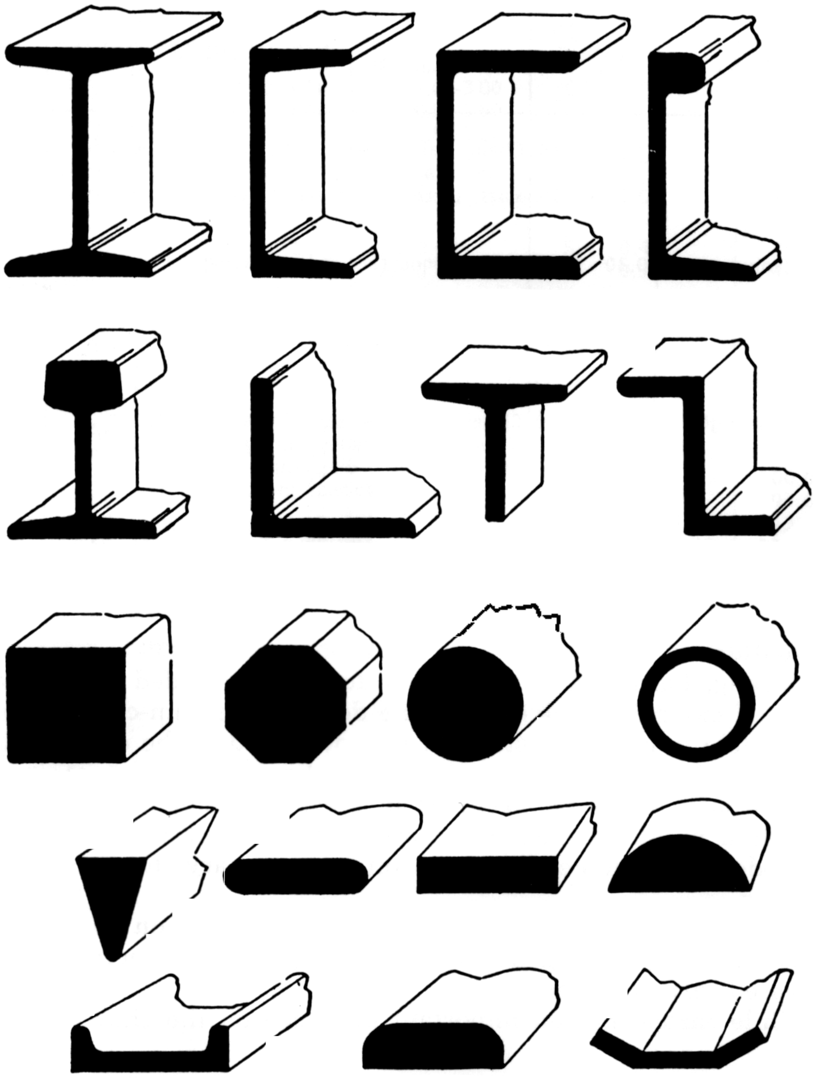

Fig. 4-1. A great many different shapes of rolled carbon and low-alloy steels are available from steel warehouses.

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Table 6-1. AWS Electrode Classification System

DIGIT	SIGNIFICANCE	EXAMPLE
1st two or 1st three	Min. tensile strength (stress relieved)	E-60xx = 60,000 psi (min) E-110xx = 110,000 psi (min)
2nd last	Welding position	E-xx1x = all positions E-xx2x = horizontal and flat E-xx3x = flat
Last	Power supply, type of slag, type of arc, amount of penetration, presence of iron powder in coating	See Table 6-2

Note: Prefix "E" (to left of a 4 or 5-digit number) signifies arc welding electrode

Measuring Preheat and Interpass Temperatures

Instruments are available for measuring surface temperatures. One such instrument has a wire coil, which is placed on the work; within a few seconds the temperature may be read on the dial at the handle end of the pyrometer. Preheat temperatures, however, can be satisfactorily estimated without expensive instruments.

Several suppliers* offer materials that melt over comparatively narrow ranges of temperatures. These indicators are furnished in sticks, liquids or pellets. If a preheat temperature of, say, 400 F, is wanted, the 400 F stick is scratched over the surface of the metal; if it leaves a white line, the metal is not at 400 F. If a transparent liquid line is formed, the metal is at or above the temperature printed on the stick.

Some common items used as rough temperature indicators are listed here:

1. Blue chalk: A mark made with carpenter's blue chalk on hot metal will turn whitish gray when the metal is above 625 F.
2. Red lumber-marking crayon: Its mark melts at about 615 F. The red mark turns pink at about 635 F and turns a dirty gray at about 650 F.
3. Solder: 50-50 solder (50% lead and 50% tin) starts melting at 360 F and is completely molten at 420 F.
4. Pine stick: White pine stick chars at about 635 F (used to determine the proper pouring temperature for some babbitts).

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

Another method for checking preheat temperatures is by temper colors: The temperature and accompanying color that appears on a freshly filed surface of carbon steel are given in Table 13-1.

Table 13-1. Temper Colors

TEMPERATURE, F	COLOR
400	Faint straw
440	Straw
475	Deep straw
520	Bronze
540	Peacock
590	Full blue
640	Light blue

It should be noted that the presence of alloying elements in the steel affects the colors formed, and some of the higher alloy steels have no temper colors.

---

*The Tempil Corporation; Markal Company; and Curtiss-Wright Corporation.

Appendix 2—MELTING POINTS OF SOME ELEMENTS

ELEMENT	CENT.	FAHR.	ELEMENT	CENT.	FAHR.
Hydrogen	—259	—434	Uranium		
Oxygen	—218	—360	Manganese		
Nitrogen	—210	—346	Beryllium		
Phosphorus	44	111	Silicon		
Sodium	97	207	Nickel		
Sulphur	113	236	Cobalt		
Selenium	218	424	Chromium		
Tin	232	450	Iron		
Bismuth	271	520	Titanium		
Cadmium	321	610	Thorium		
Lead	327	621	Zirconium		
Zinc	419	737	Vanadium		
Antimony	630	1,166	Hafnium		
Magnesium	651	1,204	Boron		
Aluminum	659	1,218	Columbium (Niobium)		
Calcium	810	1,490	Molybdenum		
Barium	850	1,562	Tantalum		
Silver	960	1,761	Tungsten		
Copper	1,083	1,981	Carbon	[illegible]	[illegible]

Reprinted With Permission From: The James F. Lincoln Arc Welding Foundation

We Support AWS
SENSE

WELDING

Skills, Processes and Practices for Entry-Level Welders

BOOK 2

- Gas Metal Arc Welding
- Flux Cored Arc Welding

LARRY JEFFUS • LAWRENCE BOWER

Welding

Skills, Processes and Practices for Entry-Level Welders

Book 2

- Gas Metal Arc Welding
- Flux Cored Arc Welding

First Edition

Larry Jeffus
Lawrence Bower

DELMAR
CENGAGE Learning™

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

DELMAR
CENGAGE Learning™

Welding: Skills, Processes and Practices for Entry-Level Welders: Book Two
Larry Jeffus/Lawrence Bower

Vice President, Editorial: Dave Garza
Director of Learning Solutions: Sandy Clark
Executive Editor: David Boelio
Managing Editor: Larry Main
Senior Product Manager: Sharon Chambliss
Editorial Assistant: Lauren Stone
Vice President, Marketing: Jennifer McAvey
Executive Marketing Manger: Deborah S. Yarnell
Senior Marketing Manager: Jimmy Stephens
Marketing Specialist: Mark Pierro
Production Director: Wendy Troeger
Production Manager: Mark Bernard
Content Project Manager: Cheri Plasse
Art Director: Benj Gleeksman
Technology Project Manager: Christopher Catalina
Production Technology Analyst: Thomas Stover

© 2010 Delmar, Cengage Learning

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, Web distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the publisher.

For product information and technology assistance, contact us at
Professional & Career Group Customer Support,
1-800-648-7450

For permission to use material from this text or product,
submit a request online at **cengage.com/permissions.**
Further permissions questions can be e-mailed to
permissionrequest@cengage.com.

Library of Congress Control Number: 2008911231

ISBN-13: 978-1-4354-2790-7
ISBN-10: 1-4354-2790-4

Delmar
5 Maxwell Drive
Clifton Park, NY, 12065-2919
USA

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your lifelong learning solutions, visit **delmar.cengage.com**
Visit our corporate website at **cengage.com.**

Notice to the Reader
Publisher does not warrant or guarantee any of the products described herein or perform any independent analysis in connection with any of the product information contained herein. Publisher does not assume, and expressly disclaims, any obligation to obtain and include information other than that provided to it by the manufacturer. The reader is expressly warned to consider and adopt all safety precautions that might be indicated by the activities described herein and to avoid all potential hazards. By following the instructions contained herein, the reader willingly assumes all risks in connection with such instructions. The publisher makes no representations or warranties of any kind, including but not limited to, the warranties of fitness for particular purpose or merchantability, nor are any such representations implied with respect to the material set forth herein, and the publisher takes no responsibility with respect to such material. The publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or part, from the readers' use of, or reliance upon, this material.

Printed in the United States of America
1 2 3 4 5 XX 11 10 09

Brief Contents

Preface viii

AWS Acknowledgment xi

Acknowledgments xii

About the Authors xiii

Chapter 1 Gas Metal Arc Welding Equipment, Setup, and Operation 1

Chapter 2 Gas Metal Arc Welding 55

Chapter 3 Flux Cored Arc Welding Equipment, Setup, and Operation 113

Chapter 4 Flux Cored Arc Welding 143

Appendix 197

Glossary 199

Index 203

This page intentionally left blank

Contents

Preface viii

AWS Acknowledgment xi

Acknowledgments xii

About the Authors xiii

Chapter 1 Gas Metal Arc Welding Equipment, Setup, and Operation 1

Introduction 2
Metal Transfer 4
  Axial Spray Metal Transfer 5
  Globular Transfer 7
  Pulsed-Spray Transfer: GMAW-P 7
  Shielding Gases for Spray or
    Pulsed-spray Transfer 9
  Buried-arc Transfer 9
  Short-circuiting Transfer: GMAW-S 11
Filler Metal Specifications 13
  Filler Metal Selection 14
  Stainless Steel Electrodes 15
Aluminum and Aluminum Alloys 15
  Aluminum Bare Welding Rods and Electrodes 20
Wire Melting and Deposition Rates 20
Welding Power Supplies 21
  Speed of the Wire Electrode 23
  Power Supplies for Short-circuiting Transfer 23
  Inductance 25
Molten Weld Pool Control 25
  Shielding Gas 26
  Power Settings 29
  Gun Manipulation 30
  Travel Speed 30
  Electrode Extension 30
  Gun Angle 31
Equipment 33
  Power Source 34
  Electrode (Wire) Feed Unit 37
  Push-type Feed System 37

Pull-type Feed System 38
Push-pull-type Feed System 39
Linear Electrode Feed System 39
Spool Gun 40
Electrode Conduit 41
Welding Gun 41
Spot Welding 42
Face, Eye, and Ear Protection 44
Face and Eye Protection. 44
Ear Protection. 48
General Work Clothing. 49
Special Protective Clothing 51
Hand Protection 51
Body Protection 52
Waist and Lap Protection 52
Arm Protection 52
Leg and Foot Protection. 52

Chapter 2 Gas Metal Arc Welding 55

Introduction 56
Setup. 57
Gas Density and Flow Rates 65
Arc-voltage and Amperage Characteristics 68
Electrode Extension 70
Welding Gun Angle 73
Effect of Shielding Gas on Welding 74
Practices. 77
Metal Preparation 78
Flat Position, 1G and 1F Positions 79
Vertical Up 3G and 3F Positions 84
Vertical Down 3G and 3F Positions 86
Horizontal 2G and 2F Positions 89
Overhead 4G and 4F Positions. 91
Axial Spray Transfer 94

Chapter 3 Flux Cored Arc Welding Equipment, Setup, and Operation 113

Introduction 114
Principles of Operation. 114
Equipment 117
Power Supply 117
Guns 117
Smoke Extraction Nozzles. 119
Electrode Feed 119

Advantages 119
  High Deposition Rate. 119
  Minimum Electrode Waste 119
  Narrow Groove Angle 120
  Minimum Precleaning 120
  All-position Welding 120
  Flexibility 120
  High Quality. 120
  Excellent Control. 121
Limitations 121
Electrodes 121
  Methods of Manufacturing 121
  Electrode Cast and Helix 124
Flux. 125
  Flux Cored Steel Electrode Identification 128
  Metal Cored Steel Electrode Identification 130
  Care of Flux Core Electrodes. 130
Shielding Gas 131
Welding Techniques 132
  Gun Angle 132
  Forehand/Perpendicular/Backhand Techniques. . 134
  Travel Speed. 136
  Mode of Metal Transfer 137
  Electrode Extension 138
  Porosity 138

Chapter 4 Flux Cored Arc Welding. 143

Introduction 144
Practices. 145
Flat-position Welds 151
Square-groove Welds 152
V-Groove and Bevel-Groove Welds 154
  Root Pass 155
  Filler Pass 156
  Cover Pass 157
Fillet Welds. 160
Vertical Welds 164
Horizontal Welds. 170
Overhead-position Welds 175
Thin-gauge Welding. 180

Appendix 197

Glossary 199

Index. 203

Preface

ABOUT THE SERIES

Welding: Skills, Processes and Practices for Entry-Level Welders is an exciting new series that has been designed specifically to support the American Welding Society's (AWS) SENSE EG2.0 training guidelines. Offered in three volumes, these books are carefully crafted learning tools consisting of theory-based texts that are accompanied by companion lab manuals, and extensive instructor support materials. With a logical organization that closely follows the modular structure of the AWS guidelines, the series will guide readers through the process of acquiring and practicing welding knowledge and skills. For schools already in the SENSE program, for those planning to join, or for schools interested in obtaining certifiable outcomes based on nationally recognized industry standards order to comply with the latest Carl D. Perkins Career and Technical Education in requirements, *Welding: Skills, Processes and Practices for Entry-Level Welders* offers a turnkey solution of high quality teaching and learning aids.

Career and technical education instructors at the high school level are often called upon to be multi-disciplinary educators, teaching welding as only one of as many as five technical disciplines in any given semester. The *Welding: Skills, Processes and Practices for Entry-Level Welders* package provides these educators with a process-based, structured approach and the tools they need to be prepared to deliver high level training on processes and materials with which they may have limited familiarity or experience. Student learning, satisfaction and retention are the target of the logically planned practices, supplements and full color textbook illustrations. While the AWS standards for entry level welders are covered, students are also introduced to the latest in high technology welding equipment such as pulsed gas metal arc welding (GMAW-P). Career pathways and career clusters may be enhanced by the relevant mathematics applied to real world activities as well as oral and written communication skills linked to student interaction and reporting.

Book 1, the core volume, introduces students to the welding concepts covered in AWS SENSE Modules 1, 2, 3, 8 and 9 (Occupational Orientation, Safety and Health of Welders, Drawing and Welding Symbol Interpretation, Thermal Cutting, and Weld Inspection Testing and Codes). Book 1 contains all the material needed for a SENSE program that prepares students for qualification in Thermal Cutting processes. The optional Books 2 and 3 cover other important welding processes and are grouped in logical combinations. Book 2 corresponds to AWS SENSE Modules 5 and 6 (GMAW, FCAW), and Book 3 corresponds to AWS SENSE Modules 4 and 7 (SMAW, GTAW).

The texts feature hundreds of four-color figures, diagrams and tight shots of actual welds to speed beginners to an understanding of the most widely used welding processes.

FEATURES

- Produced in close collaboration with experienced instructors from established SENSE programs to maximize the alignment of the content with SENSE guidelines and to ensure 100% coverage of Level I-Entry Welder Key Indicators.
- Chapter introductions contain general performance objectives, key terms used, and the AWS SENSE EG2.0 Key Indicators addressed in the chapter.
- Coverage of Key Indicators is indicated in the margin by a torch symbol and a numerical reference.
- Contains scores of fully illustrated Practices, which are guided exercises designed to help students master processes and materials. Where applicable, the Practices reproduce and reference actual AWS technical drawings in order to help students create acceptable workmanship samples.
- Each section introduces students to the materials, equipment, setup procedures and critical safety information they need in order to weld successfully.
- Hundreds of four-color figures, diagrams and tight shots of actual welds to speed beginners to an understanding of the most widely used welding processes.
- End of chapter review questions develop critical thinking skills and help students to understand “why” as well as “how.”

SUPPLEMENTS

Each book in the Welding Skills series is accompanied by a ***Lab Manual*** that has been designed to provide hands-on practice and reinforce the student’s understanding of the concepts presented in the text. Each chapter contains practice exercises to reinforce the primary objectives of the lesson, including creation of workmanship samples (where applicable), and a quiz to test knowledge of the material. Artwork and safety precautions are included throughout the manuals.

Instructor Resources (on CD-ROM), designed to support Books 1–3 and the accompanying Lab Manuals, provide a wealth of time-saving tools, including:

- An Instructor’s Guide with answers to end-of-chapter Review Questions in the texts and Lab Manual quizzes.
- Modifiable model Lesson Plans that aid in the design of a course of study that meets local or state standards and also maps to the SENSE guidelines.
- An extensive ExamView Computerized Test Bank that offers assessments in true/false, multiple choice, sentence completion and short answer formats. Test questions have been designed to expose students to the types of questions they’ll encounter on the SENSE Level 1 Exams.
- PowerPoint Presentations with selected illustrations that provide a springboard for lectures and reinforce skills and processes covered in the texts. The PowerPoint Presentations can be modified or expanded as instructors desire, and can be augmented with additional illustrations from the Image Library.
- The Image Library contains nearly all (well over 1000!) photographs and line art from the texts, most in four-color.
- A SENSE Correlation Chart that shows the close alignment of the *Welding* series to the SENSE Entry Level 1 training guidelines. Each Key Indicator within each SENSE Module is mapped to the relevant text and lab manual page or pages.

TITLES IN THE SERIES

Welding: Skills, Processes and Practices for Entry-Level Welders: Book 1, Occupational Orientation, Safety and Health of Welders, Drawing and Welding Symbol Interpretation, Thermal Cutting, Weld Inspection Testing and Codes
(Order #: 1-4354-2788-2)
Lab Manual, Book One (Order #: 1-4354-2789-0)

Welding: Skills, Processes and Practices for Entry-Level Welders: Book 2, Gas Metal Arc Welding, Flux Cored Arc Welding (Order #:1-4354-2790-4)
Lab Manual, Book Two (Order #: 1-4354-2795-5)

Welding: Skills, Processes and Practices for Entry-Level Welders: Book 3, Shielded Metal Arc Welding, Gas Tungsten Arc Welding (Order #:1-4354-2796-3)
Lab Manual, Book Three (Order #: 1-4354-2797-1)

AWS Acknowledgment

The Authors and Publisher gratefully acknowledge the support provided by the American Welding Society in the development and publication of this textbook series. "American Welding Society," the AWS logo and the SENSE logo are the trade and service marks of the American Welding Society and are used with permission.

For more information on the American Welding Society and the SENSE program, visit **http://www.aws.org/education/sense/** or contact AWS at (800) 443-9353 ext. 455 or by email: **education@aws.org**.

Acknowledgments

The authors and publisher would like to thank the following individuals for their contributions to this series:

Garey Bish, *Gwinnett Technical College, Lawrenceville, GA*
Julius Blair, *Greenup County Area Technology Center, Greenup, KY*
Rick Brandon, *Pemiscot County Career & Technical Center, Hayti, MO*
Stephen Brandow, *University of Alaska Southeast, Ketchikan, Ketchikan, AK*
Francis X Brieden, *Career Technology Center of Lackawanna County, Scranton, PA*
John Cavenaugh, *Community College of Southern Nevada, Las Vegas, NV*
Clay Corey, *Washington-Saratoga BOCES, Fort Edward, NY*
Keith Cusey, *Institute for Construction Education, Decatur, IL*
Craig Donnell, *Whitmer Career Technology Center, Toledo, OH*
Steve Farnsworth, *Iowa Lakes Community College, Emmetsburg, IA*
Ed Harrell, *Traviss Career Center, Lakeland, FL*
Robert Hoting, *Northeast Iowa Community College, Sheldon, IA*
Steve Kistler, *Moberly Area Technical Center, Moberly, MO*
David Lynn, *Lebanon Technology & Career Center, Lebanon, MO*
Frank Miller, *Gadsden State Community College, Gadsden, AL*
Chris Overfelt, *Arnold R Burton Tech Center, Salem, VA*
Kenric Sorenson, *Western Technical College, LaCrosse, WI*
Pete Stracener, *South Plains College, Levelland, TX*
Bill Troutman, *Akron Public Schools, Akron, OH*
Norman Verbeck, *Columbia/Montour AVTS, Bloomsburg, PA*

About The Authors

Larry Jeffus is a dedicated teacher and author with over twenty years experience in the classroom and several Delmar Cengage Learning welding publications to his credit. He has been nominated by several colleges for the Innovator of the Year award for setting up nontraditional technical training programs. He was also selected as the Outstanding Post-Secondary Technical Educator in the State of Texas by the Texas Technical Society. Now retired from teaching, he remains very active in the welding community, especially in the field of education.

Lawrence Bower is a welding instructor at Blackhawk Technical College, an AWS SENSE School, in Janesville, Wisconsin. Mr. Bower is an AWS-certified Welding Inspector and Welding Educator. In helping to create *Welding: Skills, Processes and Practices for Entry-Level Welders*, he has brought to bear an excellent mix of training experience and manufacturing know-how from his work in industry, including fourteen years at United Airlines, and six years in the US Navy as an aerospace welder.

This page intentionally left blank

CHAPTER

1

Gas Metal Arc Welding Equipment, Setup, and Operation

OBJECTIVES

After completing this chapter, the student should be able to

- demonstrate the proper use of personal protective equipment (PPE) for gas metal arc welding (GMAW)
- list the various terms used to describe gas metal arc welding
- describe methods of metal transfer including the axial spray metal transfer process, globular transfer, pulsed-arc metal transfer (GMAW-P), and short-circuiting transfer (GMAW-S)
- list four common shielding gases or gas mixtures used for short-circuiting, spray, and pulsed-spray transfer on plain carbon steel
- locate the GMA welding filler metal on a welding procedure specification (WPS)
- define *deposition efficiency,* and tell how a welder can control the deposition rate
- define *voltage, electrical potential, amperage,* and *electrical current* as related to GMA welding
- tell how wire-feed speed is determined and demonstrate its relationship to welding current
- list five ways the GMAW molten weld pool can be controlled by varying the shielding gas, power settings, travel speed, electrode extension, and gun angle
- describe and demonstrate the backhand and forehand welding techniques and their relationship to weld bead profile and penetration in the short-circuiting transfer mode
- list and describe the basic GMAW equipment
- use a chart to select the correct eye and face protective devices for working and welding in a shop
- describe what type of general work clothing should be worn in a welding shop
- describe special protective clothing worn by welders to protect hands, arms, body, waist, legs, and feet

KEY TERMS

axial spray metal transfer

electrode extension (stickout)

globular transfer

pinch effect

pulsed-arc metal transfer

short-circuiting transfer

slope

synergic systems

transition current

welding helmet

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter:

Module 2: Safety and Health of Welders

Key Indicator 1: Demonstrates proper use and inspection of personal protection equipment (PPE)

Module 5: Gas Metal Arc Welding (GMAW-S, GMAW spray transfer)

Key Indicator 1: Performs safety inspections of GMAW equipment and accessories

Key Indicator 3: Sets up for GMAW-S operations on carbon steel

Key Indicator 8: Sets up for GMAW (spray) operations on carbon steel

INTRODUCTION

In the 1920s, a metal arc welding process using an unshielded wire was being used to assemble the rear axle housings for automobiles. The introduction of the shielded metal arc welding electrode rapidly replaced the bare wire. The shielded metal arc welding electrode made a much higher-quality weld. In 1948, the first inert gas metal arc welding (GMAW) process, as it is known today, was developed and became commercially available, Figure 1.1. In the beginning, the GMAW process was used to weld aluminum using argon (Ar) gas for shielding. As a result, the process was known as MIG, which stands for metal inert gas welding. The later introduction of CO_2 and O_2 to the shielding gas has resulted in the American Welding Society's preferred term *gas metal*

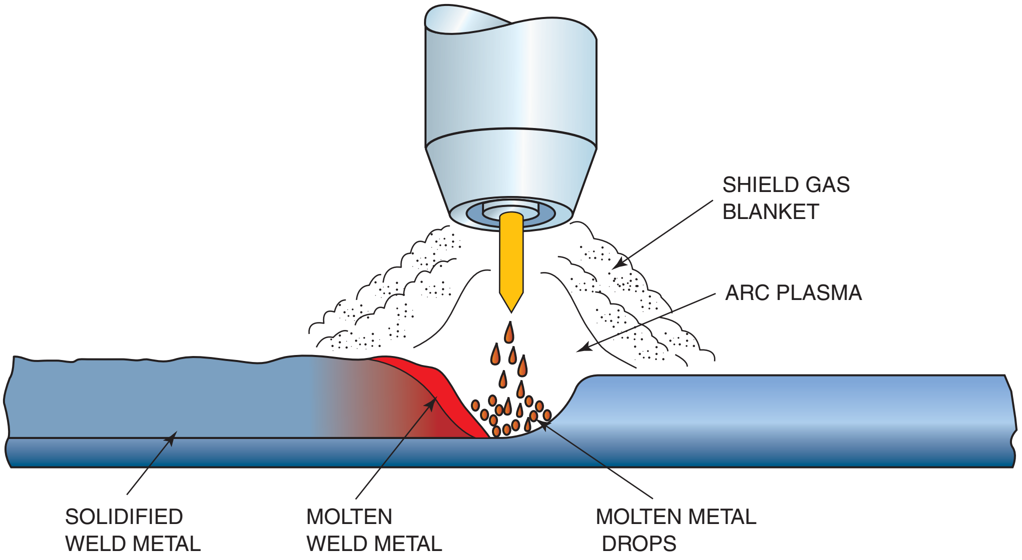

Figure 1.1 Gas-shielded metal arc welding (GMAW)

Table 1.1 **Methods of Performing Welding Processes**

Function	Manual (MA) (Example: SMAW)	Semiautomatic (SA) (Example: GMAW)	Machine (ME) (Example: GMAW)	Automatic (AU) (Example: GMAW)
Maintain the arc	Welder	Machine	Machine	Machine
Feed the filler metal	Welder	Machine	Machine	Machine
Provide the joint travel	Welder	Welder	Machine	Machine
Provide the joint guidance	Welder	Welder	Welder	Machine

arc welding (GMAW). Although the American Welding Society uses the term *gas metal arc welding* to describe this process, it is known in the field by several other terms, such as

- MIG, which is short for metal inert gas welding
- MAG, which is short for metal active gas welding
- *wire welding*, which describes the electrode used

The GMAW process may be performed as semiautomatic (SA), machine (ME), or automatic (AU) welding, Table 1.1. The GMA welding process is commonly performed as a semiautomatic process and is often mistakenly referred to as "semiautomatic welding." Equipment is available to perform most of the wire-feed processes semiautomatically, and the GMAW process can be fully automated. Robotic arc welding often uses GMAW because of the adaptability of the process in any position, Figure 1.2.

The rising use of all the various types of consumable wire welding processes has resulted in the increased sales of wire. At one time, wire made up less than 1% of the total market of filler metal. The total tonnage of filler metals used has grown and so has the percentage of wire. Today, wire exceeds 50% of the total tonnage of filler metals produced and used.

Much of the increase in the use of the wire welding processes is due to the increases in the quality of the welds produced. This improvement is due to an increased reliability of the wire-feed systems, improvements in the filler metal, smaller wire sizes, faster welding speed, higher weld deposition rates, less expensive shielding gases, and improved welding techniques. Table 1.2 shows the typical weld deposition rates using the GMA welding process. The increased usage has led to a reduction in the cost of equipment. GMA welding equipment is now found even in small shops.

In this chapter, the semiautomatic GMA welding process will be covered. The skill required to set up and operate this process is basic to the understanding and operation of other wire-feed processes. The reaction of the weld to changes in voltage, amperage, feed speed, stickout, and gas is similar to that of most wire-feed processes.

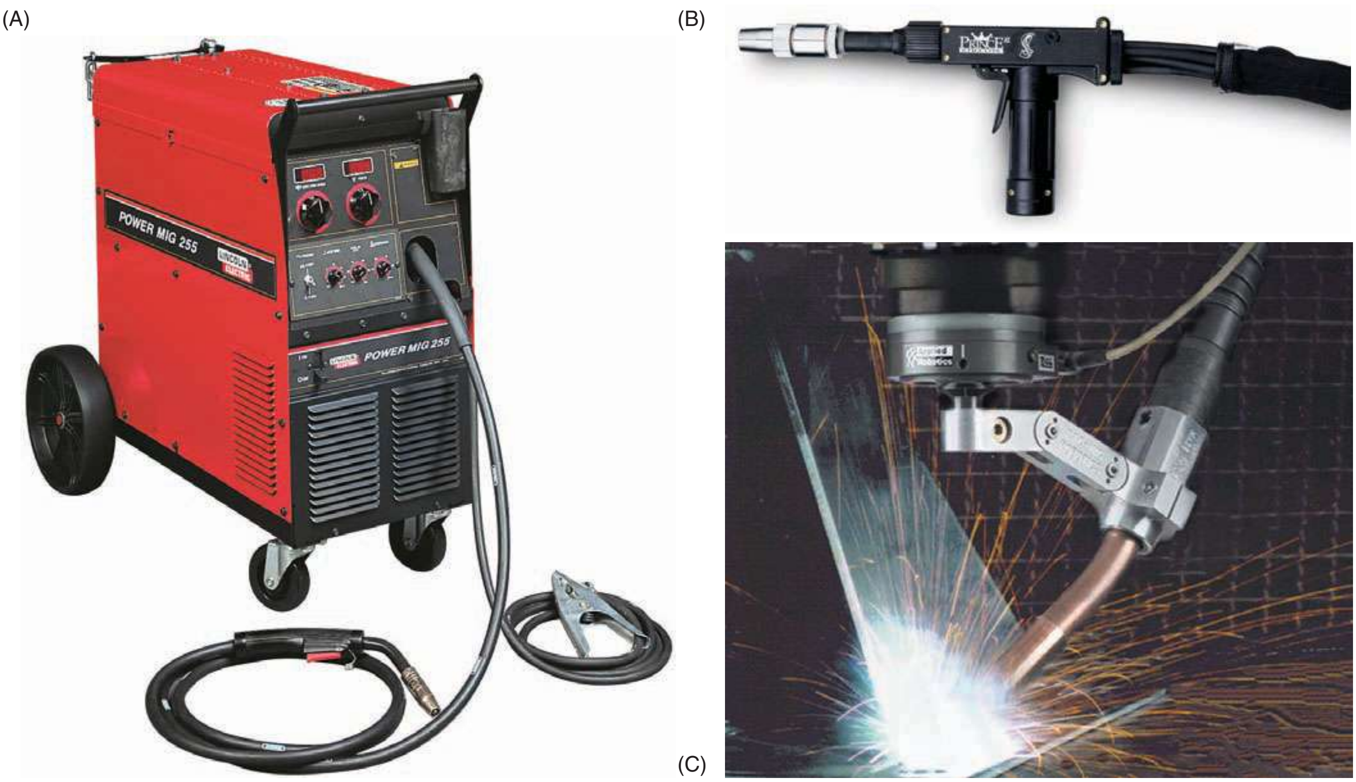

Figure 1.2 **GMAW equipment**
(A) Semiautomatic GMA welding setup. (B) Machine GMA welding gun with a friction drive, which provides both uniform nozzle-to-work distance and travel speed. (C) Automatic GMA welding.
Source: Courtesy of Lincoln Electric Company

Table 1.2 **GMA Weld Deposition Rates**

Electrode Diameter	Pounds per Hour		
Amperage	**0.35**	**0.45**	**0.63**
50	2.0	–	–
100	4.8	4.2	–
150	7.5	6.7	5.1
200	–	8.7	7.8
250	–	12.7	11.1
300	–	–	14.4

METAL TRANSFER

When first introduced, the GMA welding process was used with argon as a shielding gas to weld aluminum. Even though argon (Ar) was then expensive, the process was accepted immediately because it was much more productive than the gas tungsten arc (GTA) process and because it produced higher-quality welds than the shielded metal arc (SMA) process. This new arc welding process required very little postweld cleanup because it was slag and spatter free.

Axial Spray Metal Transfer

The freedom from spatter associated with the argon-shielded GMAW process results from a unique mode of metal transfer called **axial spray metal transfer**, Figure 1.3. This process is identified by the pointing of the wire tip from which droplets smaller than the diameter of the electrode wire are projected axially across the arc gap to the molten weld pool. There are hundreds of drops per second crossing from the wire to the base metal. These drops are propelled by arc forces at high velocity in the direction the wire is pointing. In the case of plain carbon steel and stainless steels, the molten weld pool may be too large and too hot to be controlled in vertical or overhead positions. Because the drops are very small and directed at the molten weld pool, the process is spatter free.

The spray transfer mode for carbon steels requires three conditions: argon-rich shielding gas mixtures, direct current electrode positive (DCEP) polarity (also called direct current reverse polarity, DCRP), and a current level above a critical amount called the **transition current**. The shielding gas is usually a mixture of 95% to 98% argon and 5% to 2% oxygen, or 80% to 90% argon and 20% to 10% carbon dioxide. The added percentage of active gases allows greater weld penetration. Figure 1.4 illustrates how the rate of drops transferred changes in relationship to the welding current. At low currents, the drops are large and are transferred at rates below 10 per second. These drops move slowly, falling from the electrode tip as gravity pulls them down. They tend to bridge the gap between the electrode

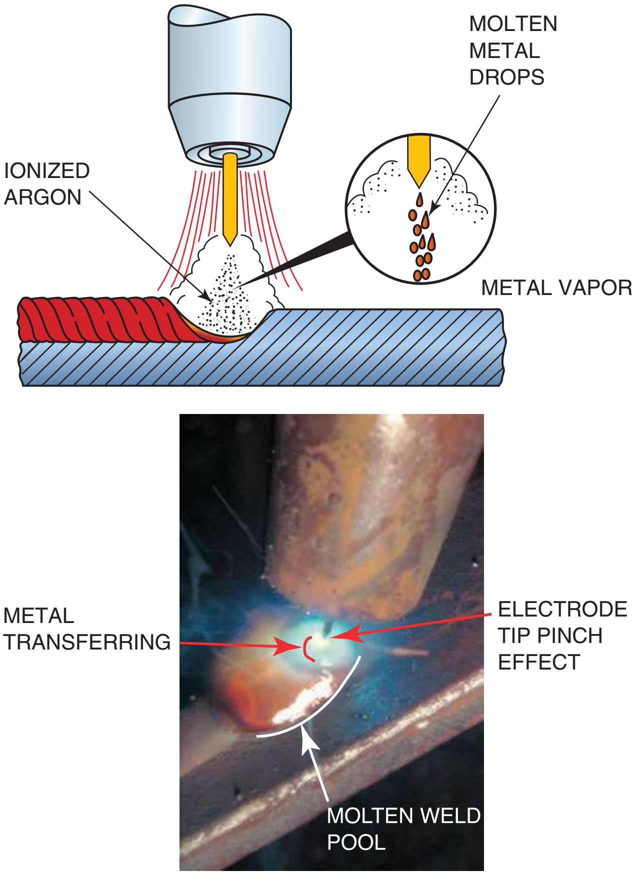

Figure 1.3 Axial spray metal transfer
Note the pinch effect of filler wire and the symmetrical metal transfer column.
Source: Courtesy of Larry Jeffus

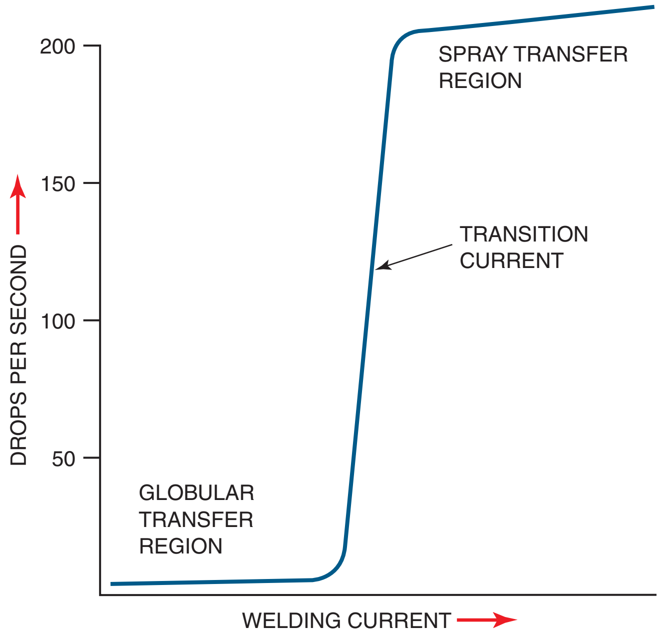

Figure 1.4 Desirable spray transfer shown schematically

tip end and the molten weld pool. This produces a momentary short circuit that throws off spatter. However, the mode of transfer changes abruptly above the critical current, producing the desirable spray.

The transition current depends on the alloy being welded. It is also proportional to the wire diameter, meaning that higher currents are needed with larger diameter wires. The need for high current density imposes some restrictions on the process. The high current hinders welding of sheet metal because the high heat cuts through sheet metal. High current also limits its use to the flat and horizontal welding positions. Weld control in the vertical or overhead position is very difficult to impossible to achieve. Table 1.3 lists the welding parameters for a variety of gases, wire sizes, and metal thicknesses for GMA welding of mild steel.

Table 1.3 GMA Welding Parameters for Mild Steel

Mild Steel	Wire-feed Speed, in./min		Voltage, V				
Base-material Thickness, in.	0.035-in.	0.045-in.	CO$_2$	75 Ar-25 CO$_2$	Ar	98 Ar-2O$_2$	Current A
0.036	105–115	–	18	16	–	–	50–60
0.048	140–160	70	19	17	–	–	70–80
0.060	180–220	90–110	20	17.7	–	–	90–110
0.075	240–260	120–130	20.7	18	20	–	120–130
1/8	280–300	140–150	21.5	18.5	20.5	–	140–150
3/16	320–340	160–175	22	19	21.5	23.5	160–170
1/4	360–380	185–195	22.7	19.5	22.5	24.5	180–190
5/16	400–420	210–220	23.5	20.5	23.5	25	200–210
3/8	420–520	220–270	25	22	25	26.5	220–250
1/2 and up	–	375	28	26	29	31	300

Globular Transfer

The **globular transfer process** is rarely used by itself because it transfers the molten metal across the arc in much larger droplets. It is used in combination with pulsed-spray transfer.

Globular transfer can be used on thin materials and at a very low current range. It can be used with higher current but is not as effective as other welding modes of metal transfer.

Pulsed-Spray Transfer: GMAW-P

Because the change from spray arc to globular transfer occurs within a very narrow current range and the globular transfer occurs at the rate of only a few drops per second, a controlled spray transfer at significantly lower average currents is achievable. Gas metal arc welding with pulsed-spray transfer (GMAW-P), or **pulsed-arc metal transfer** involves pulsing the current from levels below the transition current to those above it. The time interval below the transition current is short enough to prevent a drop from developing. About 0.1 second is needed to form a globule, so no globule can form at the electrode tip if the time interval at the low base current is about 0.01 second. Actually, the energy produced during this time is very low—just enough to keep the arc alive.

The arc's real work occurs during those intervals when the current pulses to levels above the transition current. The time of that pulse is controlled to allow a single drop of metal to transfer. This is typical of the drops normally associated with spray transfer. In fact, with many power supplies, a few small drops could transfer during the pulse interval. As with conventional spray arc, the drops are propelled across the arc gap, allowing metal transfer in all positions.

The average current can be reduced sufficiently to reduce penetration enough to weld sheet metal or reduce deposition rates enough to control the molten weld pool in all positions. This level controlling the weld heat input and rate of weld metal deposit is achieved by changing the following variables, graphed in Figure 1.5:

- Frequency—The number of times the current is raised and lowered to form a single pulse; frequency is measured in pulses per second.
- Amplitude—The amperage or current level of the power at the peak or maximum, expressed in amperage.
- Width of the pulses—The amount of time the peak amperage is allowed to stay on.

Figure 1.6 shows a typical pulsed-arc welding system. Although developed in the mid-1960s, this technology did not receive much attention until solid state electronics were developed to handle the high power required of welding power supplies. Solid state electronics provided a better, simpler, and more economical way to control the pulsing process. The newest generation of pulsed-arc systems interlocks the power supply and wire feeder so that the proper settings of the wire-feed and power supply are obtained for any given job by adjusting a single knob. Such systems have been termed **synergic systems**. The greatest benefit to synergic GMAW-P is that the power supply reacts to changes in the

Caution

The heat produced during pulsed-spray transfer welding using large-diameter wire or high current may be intense enough to cause the filter lens in a welding helmet to shatter. Be sure the helmet is equipped with a clear plastic lens on the inside of the filter lens. Avoid getting your face too close to the intense heat.

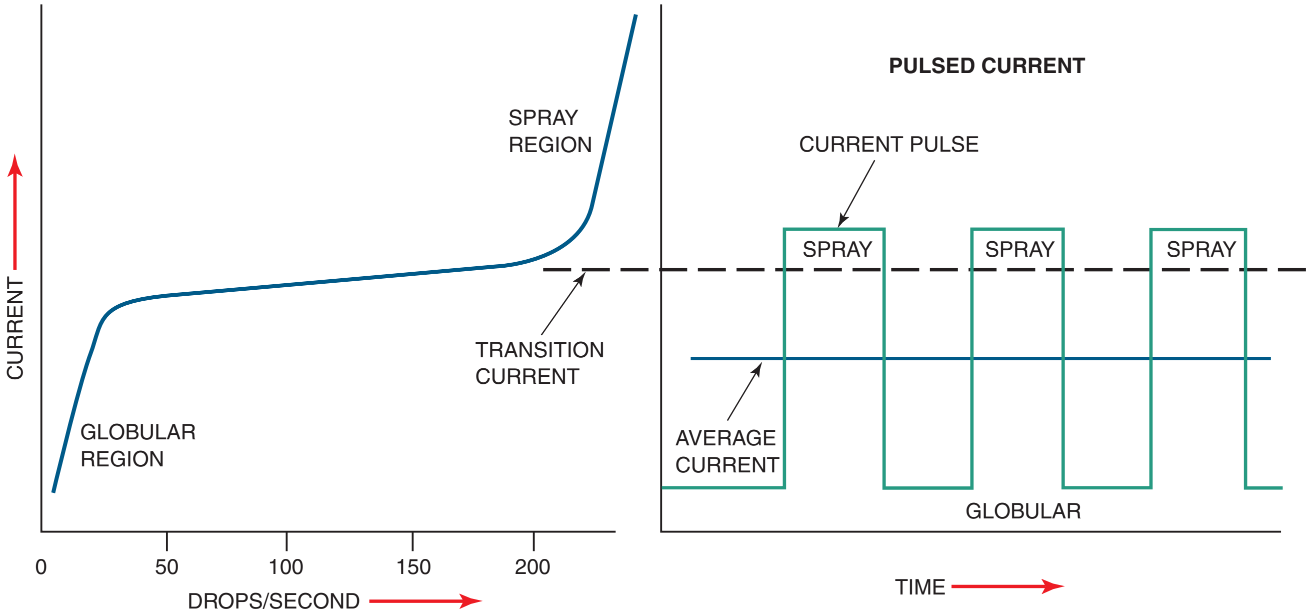

Figure 1.5 **Mechanism of pulsed-arc spray transfer at a low average current**

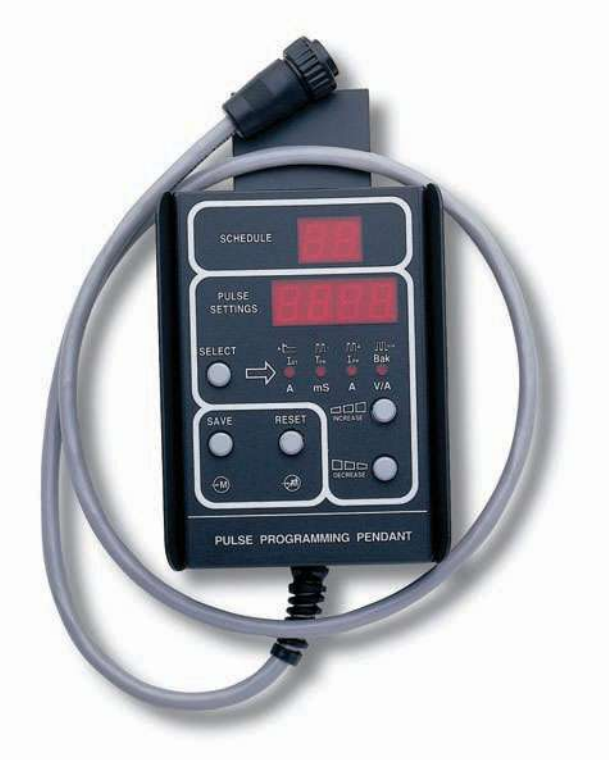
SCHEDULE
PULSE SETTINGS
SELECT
Bak
A
mS
A
V/A
SAVE
RESET
INCREASE
DECREASE
PULSE PROGRAMMING PENDANT

Figure 1.6 **GMA pulsed-arc welding system controller**
Source: Courtesy of Thermal Arc, a Thermadyne Company

work conditions such as inconsistent fitups or operator changes in stickout. For example, with a traditional GMA weld, amperage is reduced as the operator increases the stickout; but with synergic GMAW-P, the power supply senses the change in stickout and adjusts the current to maintain a more consistent weld.

In some respects, synergic GMAW-P systems are more complex to use because the correct relationships between the wire-feed speeds and power supply settings must be programmed into the equipment, and each wire composition, wire size, and shielding gas requires a special program. The manufacturer generally programs the most common combinations, allowing space in the computer for additional user input. In addition, welding power supplies that produce synergic pulsation must be capable of both constant-current (CC) and constant-voltage (CV) power output.

Shielding Gases for Spray or Pulsed-spray Transfer

Axial spray transfer is impossible without shielding gases rich in argon. Pure argon is used with all metals except steels. As much as 80% helium can be added to the argon to increase the heat in the arc without affecting the desirable qualities of the spray mode. With more helium, the transfer becomes progressively more globular, forcing the use of a different welding mode, to be described later. Since these gases are inert, they do not react chemically with any metals. This factor makes the GMAW process the only productive manual or semiautomatic method for welding metals sensitive to oxygen (essentially all metals except iron or nickel). The cathodic cleaning action which helps to remove the thin layer oxides that form on metals is associated with argon at DCEP (DCRP) and is also very important for fabricating metals such as aluminum, which quickly develop these undesirable surface oxides when exposed to air.

This same cleaning action causes problems with steels. Iron oxide in and on the steel surface is a good emitter of electrons that attracts the arc. But these oxides are not uniformly distributed, resulting in very irregular arc movement and in turn irregular weld deposits. This problem was solved by adding small amounts of an active gas such as oxygen or carbon dioxide to the argon. The reaction produces a uniform film of iron oxide on the weld pool and provides a stable site for the arc. This discovery enabled uniform welds in ferrous alloys and expanded the use of GMAW to welding those materials.

The amount of oxygen needed to stabilize arcs in steel varies with the alloy. Generally, 2% is sufficient for carbon and low-alloy steels. In the case of stainless steels, about 0.5% should prevent a refractory scale of chromium oxide, which can be a starting point for corrosion in stainless steels. Carbon dioxide can substitute for oxygen. Between 8% and 10% is optimum for low-alloy steels. In many applications, carbon dioxide is the preferred addition because the weld bead has a better contour and the arc appears to be more stable. Gas mixes of 98% argon–2% oxygen as well as 98% argon–2% carbon dioxide are commonly used for spray transfer GMA welding of stainless steels.

Buried-arc Transfer

Carbon dioxide was one of the first gases studied during the development of the GMAW process. It was abandoned temporarily because of excessive spatter and porosity in the weld. After argon was accepted for shielding, further work with carbon dioxide demonstrated that the spatter was associated with globular metal transfer. The large drops are partially supported by arc forces, Figure 1.7. As they become heavy enough to overcome those forces and drop into the pool, they bridge the gap between the wire and the weld pool, producing explosive short circuits and spatter.

Additional work showed that the arc in carbon dioxide was very forceful. Because of this, the wire tip could be driven below the surface of the molten weld pool. With the shorter arcs, the drop size is reduced, and any spatter produced as the result of short circuits is trapped in the cavity produced by the arc—hence the name **buried-arc transfer**, Figure 1.8. The resultant welds tend to be more highly crowned than those produced with open arcs, but they are relatively free of spatter and offer a decided advantage of welding speed. These characteristics make the buried-arc process useful for high-speed mechanized welding of thin sections, such as that found in compressor domes for hermetic air-conditioning and refrigeration equipment or for automotive components.

Because carbon dioxide is an oxidizing gas, its applications to welding carbon steels are restricted. It cannot be used to fabricate most nonferrous materials. Neither should it be used to weld stainless steels, because carbon corrodes the weld metal.

Carbon dioxide and helium are similar in that metal transfer in both gases is globular. Helium, too, can be used with the buried-arc technique. It has the advantage of being inert, potentially making it useful for the same types of applications as carbon dioxide but in nonferrous alloys.

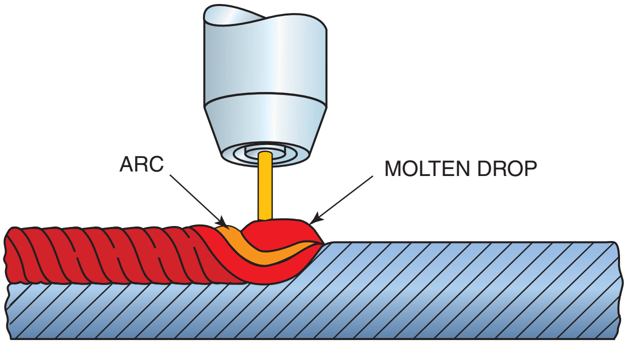

Figure 1.7 Globular metal transfer
Large drop is supported by arc forces.

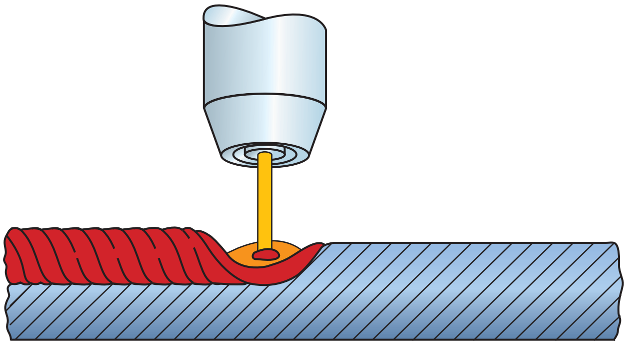

Figure 1.8 Buried-arc transfer
Wire tip is within the weld crater, so spatter is trapped.

Short-circuiting Transfer: GMAW-S

Low currents allow the liquid metal at the electrode tip to be transferred by direct contact with the molten weld pool. This process requires close interaction between the wire feeder and the power supply. This technique is called **short-circuiting transfer**, or gas metal arc welding with short-circuiting transfer (GMAW-S).

The short-circuiting mode of transfer is the most common process used with GMA welding

- on thin or properly prepared thick sections of material
- on a combination of thick to thin materials
- with a wide range of electrode diameters
- with a wide range of shielding gases
- in all positions

The 0.023, 0.030, 0.035, and 0.045 wire electrodes are the most common diameters for the short-circuiting mode. The most popular shielding gases used on carbon steel are 100% carbon dioxide (CO_2) or a combination of 75% argon (Ar) and 25% CO_2. The amperage range may be as low as 35 for materials of 24 gauge or as high as 225 for materials up to 1/8 inch in thickness on square groove weld joints. Thicker base metals can be welded if the edges are beveled to accept complete joint weld penetration.

The transfer mechanisms in this process are quite simple and straightforward, as shown schematically in Figure 1.9. To start, the wire is in direct contact with the molten weld pool, Figure 1.9A. Once the electrode touches the molten weld pool, the arc and its resistance are removed. Without the arc resistance, the welding amperage quickly rises as it begins to flow freely through the tip of the wire into the molten weld pool. The resistance to current flow is highest at the point where the electrode touches the molten weld pool. The resistance is high because both the electrode tip and weld pool are very hot. The higher the temperature,

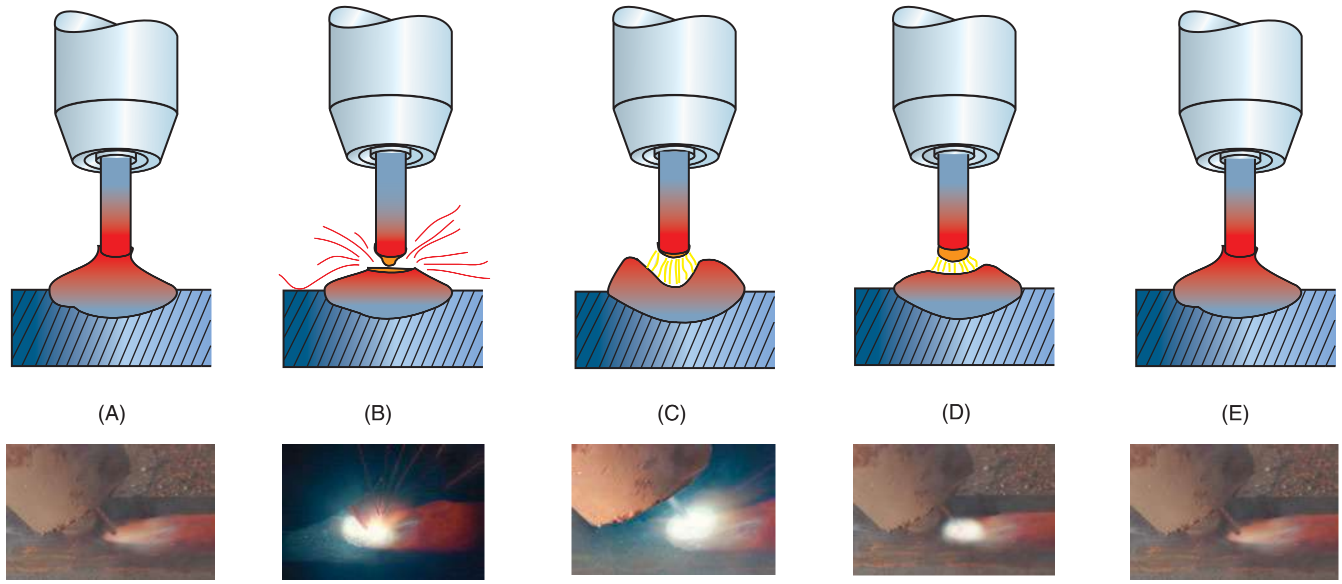

Figure 1.9 Schematic of short-circuiting transfer
Source: Courtesy of Larry Jeffus

the higher the resistance to current flow. A combination of high current flow and high resistance causes a rapid rise in the temperature of the electrode tip.

As the current flow increases, the interface between the wire and molten weld pool is heated until it explodes into a vapor (Figure 1.9B), establishing an arc. This small explosion produces sufficient force to depress the molten weld pool. A gap between the electrode tip and the molten weld pool (Figure 1.9C) immediately opens. With the resistance of the arc reestablished, the voltage increases as the current decreases.

The low current flow is insufficient to continue melting the electrode tip off as fast as it is being fed into the arc. As a result, the arc length rapidly decreases (Figure 1.9D) until the electrode tip contacts the molten weld pool (Figure 1.9E). The liquid formed at the wire tip during the arc-on interval is transferred by surface tension to the molten weld pool, and the cycle begins again with another short circuit.

If the system is properly tuned, the rate of short circuiting can be repeated from approximately 20 to 200 times per second, causing a characteristic buzzing sound. The spatter is low and the process easy to use. The low heat produced by GMAW-S makes the system easy to use in all positions on sheet metal, low-carbon steel, low-alloy steel, and stainless steel ranging in thickness from 25 gauge (0.02 in.; 0.5 mm) to 12 gauge (0.1 in.; 2.6 mm). The short-circuiting process does not produce enough heat to make quality welds in sections much thicker than 1/4 in. (6 mm) unless it is used for the root pass on a grooved weld or to fill gaps in joints. Although this technique is highly effective, lack-of-fusion defects can occur unless the process is perfectly tuned and the welder is highly skilled, especially on thicker metal. For this reason the American Welding Society D1.1 code for structural steel does not list short-circuiting GMAW as a prequalified process. In this code all short-circuiting procedures must be qualified by extensive testing.

Carbon dioxide works well with this short-circuiting process because it produces the forceful, high-energy arc needed during the arc-on interval to displace the weld pool. Helium can be used as well. Pure argon is not as effective because its arc tends to be sluggish and not very fluid. However, a mixture of 25% carbon dioxide and 75% argon produces a less harsh arc and a flatter, more fluid, and desirable weld profile. Although more costly, this gas mixture is preferred. A gas mixture of 98% argon and 2% oxygen may also be used on thinner carbon steels and sheet metal. This mixture produces lower-energy short-circuiting transfer and can be an advantage on thin-gauge metals, producing minimal burn-through at voltages as low as 13 volts.

New technology in wire manufacturing has allowed smaller wire diameters to be produced. These smaller diameters have become the preferred size even though they are more expensive due to the cost of drawing wires down to these desirable sizes. The short-circuiting process works better with a short electrode stickout.

The power supply is most critical. It must have a constant voltage, otherwise known as constant-potential output, and sufficient inductance to slow the time rate of current increase during the short-circuit interval. Too little inductance causes spatter due to high-current surges. Too much inductance causes the system to become sluggish. The short-circuiting rate decreases enough to make the process difficult to use. Also, the power supply must sustain an arc long enough to premelt the electrode tip in anticipation of the transfer at recontact with the weld pool.

FILLER METAL SPECIFICATIONS

GMA welding filler metals are available for a variety of base metals, Table 1.4. The most frequently used filler metals are AWS specification A5.18 for carbon steel and AWS specification A5.9 for stainless steel. Wire electrodes are produced in diameters of 0.023, 0.030, 0.035, 0.045, and 0.062 inch. Other, larger diameters are available for production work and can include wire diameter sizes such as 1/16, 5/64, and 7/64 inch. Table 1.5 lists the most common sizes and the amperage ranges for these electrodes. The amperage will vary depending on the method of metal transfer, type of shielding gas, and base metal thickness. Some steel wire electrodes have a thin copper coating. This coating provides some protection to the electrode from rusting and improves the electrical contact between the wire electrode and the contact tube. It also acts as a lubricant to help the wire move more smoothly through the liner and contact tube. These electrodes may look like copper wire because of the very thin copper cladding. The amount of copper is so small that it

Table 1.4 **AWS Filler Metal Specifications for Different Base Metals**

Base Metal Type	AWS Filler Metal Specification
Aluminum and aluminum alloys	A5.10
Copper and copper alloys	A5.6
Magnesium alloys	A5.19
Nickel and nickel alloys	A5.14
Stainless steel (austenitic)	A5.9
Steel (carbon)	A5.18
Titanium and titanium alloys	A5.16

Table 1.5 **Filler Metal Diameters and Amperage Ranges**

	Electrode Diameter		Amperage
Base Metal	Inch	Millimeter	Range
Carbon Steel	0.023	0.6	35–190
	0.030	0.8	40–220
	0.035	0.9	60–280
	0.045	1.2	125–380
	1/16	1.6	275–450
Stainless Steel	0.023	0.6	40–150
	0.030	0.8	60–160
	0.035	0.9	70–120
	0.045	1.2	140–310
	1/16	1.6	280–450

either burns off or is diluted into the weld pool with no significant effect on the weld deposit.

Filler Metal Selection

Solid Wire

The AWS specification for carbon steel filler metals for gas-shielded welding wire is A5.18. Filler metal classified within this specification can be used for GMAW, gas tungsten arc welding (GTAW), and plasma arc welding (PAW) processes. Because in GTAW and PAW the wire does not carry the welding current, the letters *ER* are used as a prefix. The *ER* is followed by two numbers to indicate the minimum tensile strength of a good weld. The actual strength is obtained by adding three zeroes to the right of the number given. For example, ER70S-x is 70,000 psi.

The *S* located to the right of the tensile strength indicates that this is a solid wire. The last number—2, 3, 4, 5, 6, or 7—or the letter *G* is used to indicate the filler metal composition and the weld's mechanical properties, Figure 1.10.

ER70S-2

This is a deoxidized mild steel filler wire. The deoxidizers allow this wire to be used on metal that has light coverings of rust or oxides. There may be a slight reduction in the weld's physical properties if the weld is made on rust or oxides, but this reduction is only slight, and the weld will usually still pass the classification test standards. This is a general-purpose filler that can be used on killed, semikilled, and rimmed steels. Argon-oxygen, argon-CO_2, and CO_2 can be used as shielding gases. Welds can be made in all positions.

ER70S-3

This is a popular filler wire. It can be used in single or multiple-pass welds in all positions. ER70S-3 does not have the deoxidizers required to weld over rust, over oxides, or on rimmed steels. It produces high-quality welds on killed and semikilled steels. Argon-oxygen, argon-CO_2, and CO_2 can be used as shielding gases. This is the low-carbon steel filler most commonly used to weld galvanized steel.

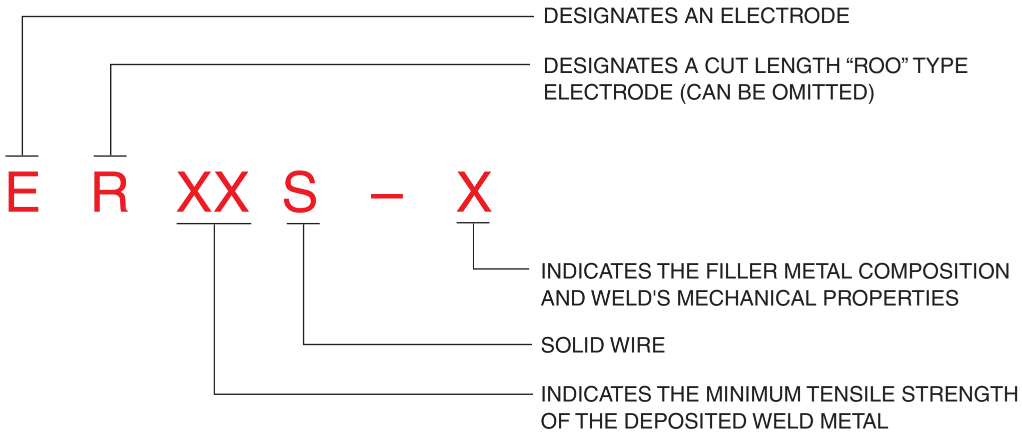
DESIGNATES AN ELECTRODE
DESIGNATES A CUT LENGTH "ROO" TYPE ELECTRODE (CAN BE OMITTED)
E R XX S – X
INDICATES THE FILLER METAL COMPOSITION AND WELD'S MECHANICAL PROPERTIES
SOLID WIRE
INDICATES THE MINIMUM TENSILE STRENGTH OF THE DEPOSITED WELD METAL

Figure 1.10 AWS numbering system for carbon steel filler metal for GMAW

ER70S-6

This is a good general-purpose filler wire. It has the highest levels of manganese and silicon. The wire can be used to make smooth welds on sheet metal or thicker sections. Welds over rust, oxides, and other surface impurities will lower the mechanical properties, but not normally below the specifications of this classification. Argon-oxygen, argon-CO_2, and CO_2 can be used as shielding gases. Welds can be made in all positions. This filler wire is not suggested for galvanized steel as its higher silicon content may induce cracking.

Stainless Steel Electrodes

The AWS specification for stainless steel–covered arc electrodes is A5.4 and for stainless steel bare, cored, and stranded electrodes and welding rods is A5.9. Filler metal classified within A5.4 uses the letter E as its prefix, and the filler metal within A5.9 uses the letters ER as its prefix.

Following the prefix, the three-digit stainless steel number from the American Iron and Steel Institute (AISI) is used. This number indicates the type of stainless steel in the filler metal, Table 1.6.

ALUMINUM AND ALUMINUM ALLOYS

The International Alloy Designation System is the most widely accepted naming scheme for aluminum alloys. Each alloy has a four-digit number, in which the first digit indicates the major alloying elements. Heat-treatable alloys can be strengthened by furnace operations after they are produced, while non-heat-treatable alloys are work hardened by plastic deformation such as bending and rolling operations. Heat treatment of aluminum alloys is sometimes referred to as precipitation hardening, and work hardening is sometimes referred to as strain hardening.

- 1000 series are essentially pure aluminum with a minimum 99% aluminum content by weight, and can be work hardened.
- 2000 series are alloyed with copper, and are heat treatable.
- 3000 series are alloyed with manganese, and can be work hardened.
- 4000 series are alloyed with silicon and can be work hardened.
- 5000 series are alloyed with magnesium, derive most of their strength from heat treatment, but can also be work hardened.
- 6000 series are alloyed with magnesium and silicon, are easy to machine, and are heat treatable, but not to the high strengths that 2000, 5000, and 7000 can reach.
- 7000 series are alloyed with zinc, and can be heat treated to the highest strengths of any aluminum alloy.
- 8000 series are a miscellaneous category where other elements not listed above may be used.

The AWS specifications for aluminum and aluminum alloy filler metals are A5.3 for covered arc welding electrodes and A5.10 for bare welding rods and electrodes. Filler metal classified within A5.3 uses the atomic symbol Al, and in A5.10 the prefix ER is used with the Aluminum Association number for the alloy, Table 1.7.

Aluminum Bare Welding Rods and Electrodes

ER1100

1100 aluminum has the lowest percentage of alloy agents of all the aluminum alloys, and it melts at 1215°F. The filler wire is also relatively pure. ER1100 produces welds that have good corrosion resistance and high ductility, with tensile strengths ranging from 11,000 to 17,000 psi. The weld deposit has a high resistance to cracking during welding. This wire can be used with oxyfuel gas welding (OFW), GTAW, and GMAW. Preheating to 300°F to 350°F is required for GTA welding on plate or pipe 3/8 in. and thicker to ensure good fusion. Flux is required for OFW. 1100 aluminum is commonly used for items such as food containers, food-processing equipment, storage tanks, and heat exchangers. ER1100 can be used to weld 1100 and 3003 grade aluminum.

ER4043

ER4043 is a general-purpose welding filler metal. It has 4.5% to 6.0% silicon added, which lowers its melting temperature to 1155°F. The lower melting temperature helps promote a free-flowing molten weld pool. The welds have high ductility and a high resistance to cracking during welding. This wire can be used with OFW, GTAW, and GMAW. Preheating to 300°F to 350°F is required for GTA welding on plate or pipe 3/8 in. and thicker to ensure good fusion. Flux is required for OFW. ER4043 can be used to weld on 2014, 3003, 3004, 4043, 5052, 6061, 6062, and 6063 and cast alloys 43, 355, 356, and 214.

ER5356

ER5356 has 4.5% to 5.5% magnesium added to improve the tensile strength. The weld has high ductility but only an average resistance to cracking during welding. This wire can be used for GTAW and GMAW. Preheating to 300°F to 350°F is required for GTA welding on plate or pipe 3/8 in. and thicker to ensure good fusion. ER5356 can be used to weld on 5050, 5052, 5056, 5083, 5086, 5154, 5356, 5454, and 5456.

ER5556

ER5556 has 4.7% to 5.5% magnesium and 0.5% to 1.0% manganese added to produce a weld with high strength. The weld has high ductility and only average resistance to cracking during welding. This wire can be used for GTAW and GMAW. Preheating to 300°F to 350°F is required for GTA welding on plate or pipe 3/8 in. and thicker to ensure good fusion. ER5556 can be used to weld on 5052, 5083, 5356, 5454, and 5456.

WIRE MELTING AND DEPOSITION RATES

The wire melting rates, deposition rates, and wire-feed speeds of the consumable wire welding processes are affected by the same variables. Before discussing them, however, these terms need to be defined. The wire melting rate, measured in inches per minute (in/min) or pounds per hour (lb/hr), is the rate at which the arc consumes the wire. The deposition rate, the measure of weld metal deposited, is nearly always less than the melting rate because not all of the wire is converted to weld metal. Some is lost as slag, spatter, or fume. The amount of weld metal deposited in ratio to the wire used is called the deposition efficiency.

Deposition efficiencies depend on the process, on the gas used, and even on how the welder sets welding conditions. With efficiencies of approximately 98%, solid wires with argon-rich shield gas mixes are best. Some of the self-shielded cored wires are poorest, with efficiencies as low as 80%.

Welders can control the deposition rate by changing the current, electrode extension, and diameter of the wire. To obtain higher melting rates, they can increase the current or wire extension or decrease the wire diameter. Knowing the precise constants is unimportant. However, it is important to know that current greatly affects melting rate and that the electrode extension must be controlled if results are to be reproducible.

WELDING POWER SUPPLIES

To better understand the terms used to describe the different welding power supplies, you need to know the following electrical terms:

- Voltage, or volts (V), is a measurement of electrical pressure and is the force that causes the current (amperage) to flow, in the same way that pounds per square inch is a measurement of water pressure.
- Electrical potential means the same thing as voltage and is usually expressed by using the term *potential (P)*. The terms *voltage*, *volts*, and *potential* can all be interchanged when referring to electrical pressure.
- Amperage, or amps (A), is the measurement of the total number of electrons flowing, in the same way that gallons are a measurement of the amount of water flowing.
- Electrical current means the same thing as amperage and is usually expressed by using the term *current (C)*. The terms *amperage*, *amps*, and *current* can all be interchanged when referring to electrical flow.

GMAW power supplies are constant-voltage, constant-potential (CV, CP) machines, unlike shielded metal arc welding (SMAW) power supplies, which are constant-current (CC) machines and are sometimes called drooping arc voltage (DAV). It is impossible to make acceptable welds using the wrong type of power supply. GMAW power supplies are available as transformer rectifiers, motor generators, or inverters, Figure 1.11. Some newer machines use electronics, enabling them to supply both types of power at the flip of a switch; they may be referred to as CC/CV power supplies.

The relationships between current and voltage with different combinations of arc length or wire-feed speeds are called volt-ampere characteristics. The volt-ampere characteristics of arcs in argon with constant arc lengths or constant wire-feed speeds are shown in Figure 1.12. To maintain a constant arc length while increasing current, it is necessary to increase voltage. For example, with a 1/8-in. (3-mm) arc length, increasing current from 150 to 300 amperes requires increasing the

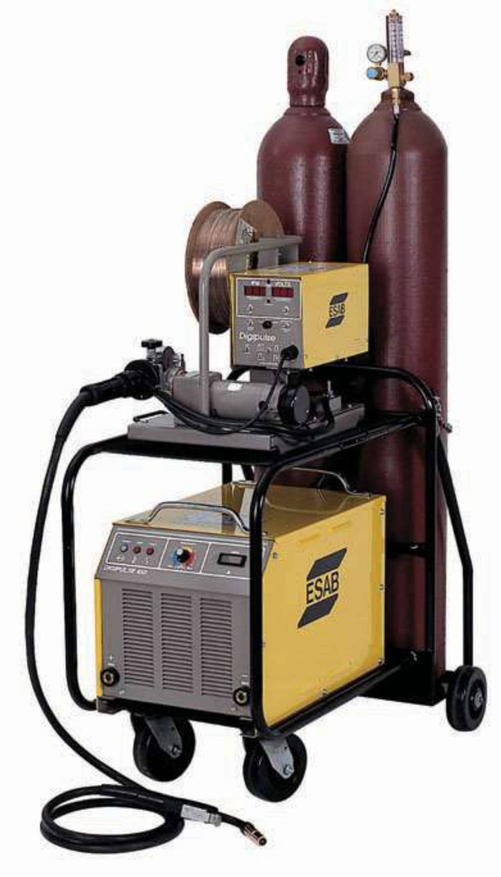

Figure 1.11 **Transformer rectifier welding power supply**
Source: Courtesy of ESAB Welding & Cutting Products

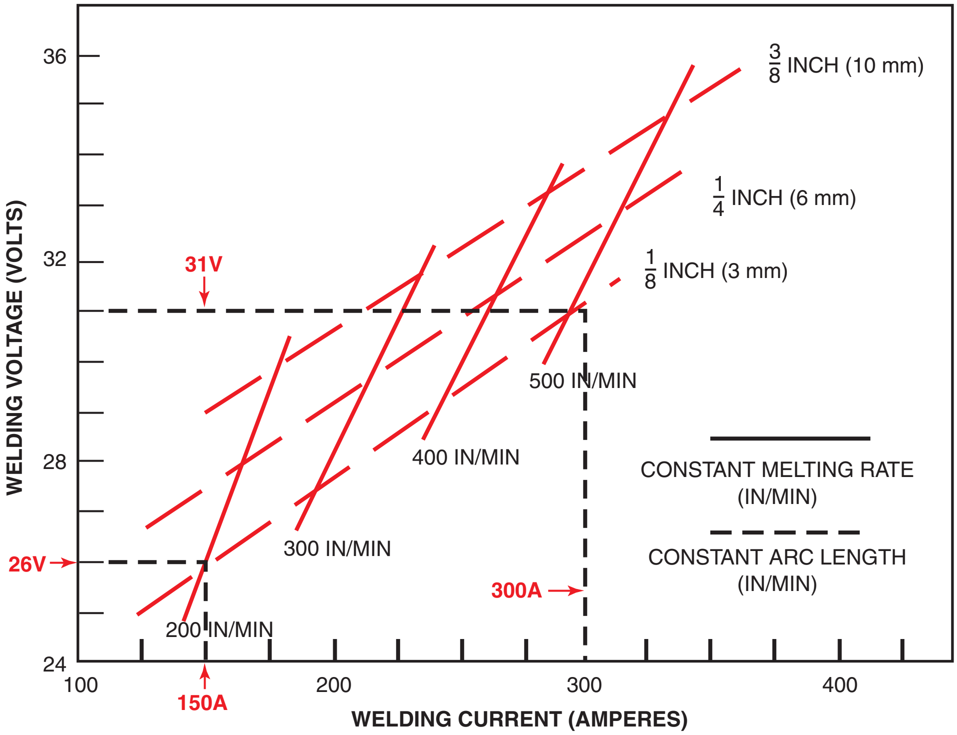

Figure 1.12 **Volt-ampere characteristics of arcs with argon**
The arc length and arc voltage are affected by the welding current and wire-feed speed (0.045-in. [1.43-mm] wire; 1-in. [25-mm] electrode extension).

voltage from about 26 to 31 volts. The current increase illustrated here results from increasing the wire-feed speed from 200 to 500 inches per minute.

Speed of the Wire Electrode

The wire-feed speed is generally recommended by the electrode manufacturer and is selected in inches per minute (ipm), or how fast the wire exits the contact tube. The welder uses a wire-speed control dial on the wire-feed unit to control ipm. It can be advanced or slowed to control the burn-off rate, or how fast the electrode transfers into the weld pool, to meet the welder's skill in controlling the weld pool. Note the direct relationship between current (amps) and wire-feed speed (wfs); as wire-feed speed increases, the amperage increases. If wire-feed speed is reduced, amperage will decrease, Table 1.8.

To accurately measure wire-feed ipm, snip off the wire at the contact tube. Wearing safety glasses and pointing the contact tube away from your face, squeeze the trigger for ten seconds; release and snip off the wire electrode. Measure the number of inches of wire that was fed out in the ten seconds. Now using basic shop math, multiply its total length in inches by six. The result is how many inches of wire were fed per minute.

Power Supplies for Short-circuiting Transfer

Although the GMA power source is said to have a constant voltage (CV) or constant potential (CP), it is not perfectly constant. The graph in Figure 1.13 shows that there is a slight decrease in voltage as the amperage increases within the working range. The rate of decrease is known as **slope**. It is expressed as the voltage decrease per 100-ampere increase—for example, 10 V/100 A. For short-circuiting welding, some welding power supplies are equipped to allow changes in the slope by steps or continuous adjustment.

The slope, which is called the *volt-ampere curve*, is often drawn as a straight line because it is fairly straight within the working range of the machine. Whether it is drawn as a curve or a straight line, the slope can be found by finding two points. The first point is the set voltage as read from the voltmeter when the gun switch is activated but no welding is

Table 1.8 Typical Amperages for Carbon Steel

	Wire Diameter Amperages			
Wire-feed Speed* in./min (m/min)	.030 in. (0.8 mm)	.035 in. (0.9 mm)	.045 in. (1.2 mm)	.062 in. (1.6 mm)
100 (2.5)	40	65	120	190
200 (5.0)	80	120	200	330
300 (7.6)	130	170	260	425
400 (10.2)	160	210	320	490
500 (12.7)	180	245	365	–
600 (15.2)	200	265	400	–
700 (17.8)	215	280	430	–

*To check feed speed, run out wire for one minute and then measure its length.

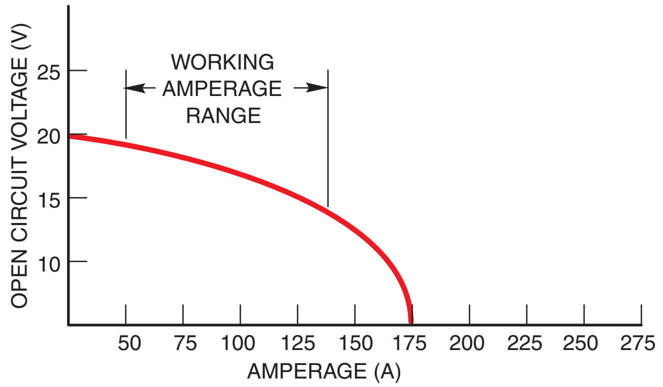

Figure 1.13 Constant potential welder slope

being done. This is referred to as the open circuit voltage. The second point is the voltage and amperage as read during a weld. The voltage control is not adjusted during the test but the amperage can vary. The slope is the voltage difference between the first and second readings. The difference can be found by subtracting the second voltage from the first voltage. Therefore, for settings over 100 amperes, it is easier to calculate the slope by adjusting the wire feed so that you are welding with 100 amperes, 200 amperes, 300 amperes, and so on. In other words, the voltage difference can be simply divided by 1 for 100 amperes, 2 for 200 amperes, and so forth.

The machine slope is affected by circuit resistance. Circuit resistance may result from a number of factors, including poor connections, long leads, or a dirty contact tube. A higher resistance means a steeper slope. In short-circuiting machines, increasing the inductance increases the slope. This increase slows the current's rate of change during short circuiting and the arcing intervals, Figure 1.14. Therefore, slope and inductance become synonymous in this discussion. As the slope increases, both the short-circuit current and the **pinch effect** are reduced. A flat slope has both an increased short-circuit current and a greater pinch effect.

The machine slope affects the short-circuiting metal transfer mode more than it does the other modes. Too much current and pinch effect from a flat slope cause a violent short and arc restart cycle, which results in increased spatter. Too little current and pinch effect from a steep slope result in the short circuit not being cleared as the wire freezes in the molten pool and piles up on the work, Figure 1.15.

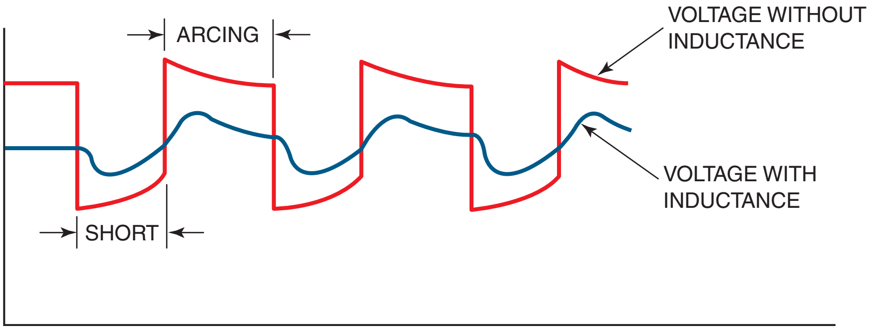

Figure 1.14 Voltage pattern with and without inductance

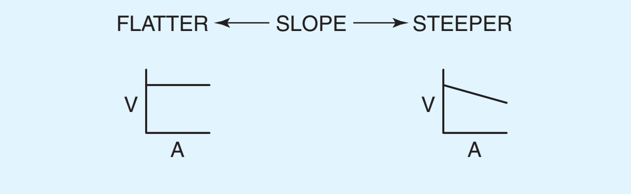

Buildup	Decreases	Increases
Depth of fusion	Increases	Decreases
Spatter	Increases	Decreases
Shorting	Violently cleared	May not clear

Figure 1.15 Effect of slope

The slope should be adjusted so that a proper spatter-free metal transfer occurs. On machines that have adjustable slopes, this is easily set. Experiment 2-2 describes a method of adjusting the circuit resistance to change the slope on machines that have a fixed slope. This is done by varying the contact tube-to-work distance. The GMA filler wire is much too small to carry the welding current and heats up due to its resistance to the current flow. The greater the tube-to-work distance, the greater the circuit resistance and the steeper the slope. By increasing or decreasing this distance, a proper slope can be obtained so that the short circuiting is smoother with less spatter.

Inductance

Many newer GMA welding power supplies are supplied with an inductance control. Inductance controls are used primarily in the GMA short-circuiting transfer mode, especially when open root joints are involved. Inductance controls the rise up to peak current during GMAW short-circuiting transfer. A low inductance control setting will provide a greater short-circuiting frequency, which may be beneficial in welding thin material where burn-through is an issue. A higher setting will reduce the frequency of short circuits, creating slightly longer arc periods, which allows greater current to go to the work. The greater current can be used to increase penetration in thicker sections or where complete penetration is required on open root joints welded from one side.

In the case of the GMA welding arc, the rate of change in the amperage relative to the arc voltage determines how the metal droplet detaches from the end of the electrode. If the rate of change is too rapid (inductance too low), the droplet detaches violently and produces excessive spatter. If the rate of change is too slow (inductance too high) the metal droplet doesn't detach cleanly and the arc is unstable. A midrange setting on the inductance control can be used for general-purpose short-circuit GMA welding and adjustments made for thick, thin, or open root conditions.

MOLTEN WELD POOL CONTROL

The GMAW molten weld pool can be controlled by varying the following factors: shielding gas, power settings, gun manipulation, travel speed, electrode extension, and gun angle.

Shielding Gas

The shielding gas selected for a weld has a definite effect on the weld produced. The properties that can be affected include the method of metal transfer, welding speed, weld contour, arc cleaning effect, and fluidity of the molten weld pool.

In addition to the effects on the weld itself, the metal to be welded must be considered in selecting a shielding gas. Some metals must be welded with an inert gas such as argon or helium or mixtures of argon and helium. Other metals weld more favorably with reactive gases such as carbon dioxide or with mixtures of inert gases and reactive gases such as argon and oxygen or argon and carbon dioxide, Table 1.9. The most commonly used shielding gases are 75% argon + 25% CO_2, argon + 1% to 5% oxygen, and carbon dioxide, Figure 1.16.

- **Argon:** The atomic symbol for argon is *Ar*, and it is an inert gas. *Inert gases* do not react with any other substance and are insoluble in molten metal. One hundred percent argon is used on nonferrous metals such as aluminum, copper, magnesium, nickel, and their alloys; but 100% argon is not normally used for making welds on ferrous metals.

  Because argon is denser than air, it effectively shields welds by pushing the lighter air away. Argon is relatively easy to ionize. Easily ionized gases can carry long arcs at lower voltages. This makes it less sensitive to changes in arc length.

  Argon gas is naturally found in all air and is collected in air separation plants. There are two methods of separating air to extract nitrogen, oxygen, and argon. In the cryogenic process, air is super-cooled to temperatures that cause it to liquefy and the gases are separated. In the noncryogenic process, molecular sieves (strainers with very small holes) separate the various gases, much like using a screen to separate sand from gravel.
- **Argon gas blends:** Oxygen, carbon dioxide, helium, and nitrogen can be blended with argon to change argon's welding characteristics. Adding reactive gases (oxidizing), such as oxygen or carbon dioxide, to argon tends to stabilize the arc, promote favorable metal transfer, and minimize spatter. As a result, the penetration pattern is improved, and undercutting is reduced or eliminated. Adding helium or nitrogen gases (nonreactive or inert) increases the arc heat for deeper penetration.

  The amount of the reactive gases, oxygen or carbon dioxide, required to produce the desired effects is quite small. As little as a half percent change in the amount of oxygen will produce a noticeable effect on the weld. Most of the time, blends containing 1% to 5% of oxygen are used. Carbon dioxide may be added to argon in the range of 10% to 30%. Blends of argon with less than 10% carbon dioxide may not have enough arc voltage to give the desired results. The most commonly used blend for short-circuiting transfer is 75% argon and 25% CO_2.

  When using oxidizing shielding gases with oxygen or carbon dioxide added, a suitable filler wire containing deoxidizers should be used to prevent porosity in the weld. The presence of oxygen in the shielding gas can also cause some loss of certain alloying

Table 1.9 **(A) Metals Matched with GMAW Shielding Gases and Gas Blends; (B) GMAW Shielding Gases and Gas Blends Matched with Metals**

GMAW Shielding Gas, Gas Blends, Metals, and Welding Process							
Gases/ Blend	**Gas Reaction**	**Application**	**Remarks**	**Gas/ Blend**	**Gas Reaction**	**Application**	**Remarks**
Ar + 5% CO_2	Oxidizing	Low-alloy steel	Pulse spray and short-circuit transfer in out-of-position welds	Ar + CO_2 + O_2	Oxidizing	Low-alloy steel and some stainless steels	All metal transfer for automatic and robotic applications
Ar + 10% CO_2	Oxidizing	Low-alloy steel	Same as above with a wider, more fluid weld pool	Ar + CO_2 + N	Almost inert	Stainless steel	All metal transfer, excellent for thin gauge material
Ar + 25% CO_2	Oxidizing	Mild, low-alloy steels and stainless steel	Smooth weld surface, reduces penetration with short-circuiting transfer	He + 7.5% Ar + 2.5% CO_2	Almost inert	Stainless steel and some low-alloy steels	Excellent toughness, arc stability, wetting characteristics, and bead contour, little spatter with short-circuiting transfer

(B)

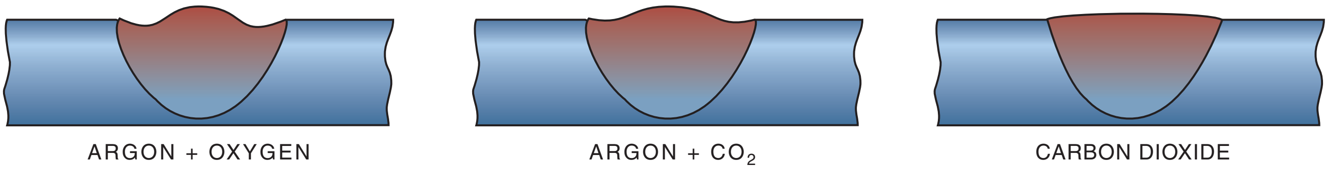

Figure 1.16 Effect of shielding gas on weld bead shape

elements, such as chromium, vanadium, aluminum, titanium, manganese, and silicon.

- **Helium:** The atomic symbol for helium is *He*, and it is an inert gas that is a product of the natural gas industry. It is removed from natural gas as the gas undergoes separation (fractionation) for purification or refinement.

  Helium is lighter than air; thus its flow rates must be about twice as high as argon's for acceptable stiffness in the gas stream to be able to push air away from the weld. Proper protection is difficult in drafts unless high flow rates are used. It requires a higher voltage to ionize, which produces a much hotter arc. There is a noticeable increase in both the heat and temperature of a helium arc. This hotter arc makes it easier to make welds on thick sections of aluminum and magnesium.

  Small quantities of helium are blended with heavier gases. These blends take advantage of the heat produced by the lightweight helium and weld coverage by the other heavier gas. Thus, each gas is contributing its primary advantage to the blended gas.
- **Carbon dioxide:** Carbon dioxide is a compound made up of one carbon atom (C) and two oxygen atoms (O_2), and its molecular formula is CO_2. One hundred percent carbon dioxide is widely used as a shielding gas for GMA welding of steels. In the short-circuiting transfer mode it allows higher welding speed, better penetration, good mechanical properties, and costs less than the inert gas mixes. The chief drawback in the use of carbon dioxide is the less-steady arc characteristics and a considerable increase in weld spatter. The spatter can be kept to a minimum by maintaining a very short, uniform arc length and strict attention to amperage and voltage parameters. CO_2 can produce sound, spatter-free welds of the highest quality, provided established procedures are followed and a filler wire having the proper deoxidizing additives is selected.
- **Nitrogen:** The atomic symbol for nitrogen is *N*. It is not an inert gas but is relatively nonreactive to the molten weld pool. It is often used in blended gases to increase the arc's heat and temperature. One hundred percent nitrogen can be used to weld copper and copper alloys and is an economical choice for gas purging of some austenitic stainless steel pipe welds.

Power Settings

As the power settings, voltage, and amperage are adjusted, the weld bead is affected. Making an acceptable weld requires a balancing of the voltage and amperage. If either or both are set too high or too low, the

Module 5
Key Indicator 3, 8

weld penetration can decrease. A GMA welding machine has no direct amperage settings. Instead, the amperage at the arc is adjusted by changing the wire-feed speed. As a result of the welding machine's maintaining a constant voltage when the wire-feed speed increases, more amperage flows across the arc. This higher amperage is required to melt the wire so that the same arc voltage can be maintained. The higher amperage is used to melt the filler wire and does not increase the penetration. In fact, the weld penetration may decrease significantly.

Increasing and decreasing the voltage changes the arc length; however, it may not put more heat into the weld. Like changes in the amperage, these voltage changes may decrease weld penetration.

Gun Manipulation

The GMA welding process is greatly affected by the location of the electrode tip and molten weld pool. During the short-circuiting process if the arc is directed to the base metal and outside the molten weld pool, the welding process may stop. Without the resistance of the hot molten metal, high-amperage surges occur each time the electrode tip touches the base metal, resulting in a loud pop and a shower of sparks. It is something that occurs each time a new weld is started. So when making a weave pattern, which is a gun manipulation technique of moving side to side in order to produce a wider weld bead, you must keep the arc and electrode tip directed into the molten weld pool. Other than the sensitivity to arc location, most of the SMAW weave patterns that keep the electrode wire at or near the leading edge of the weld pool can be used for short-circuiting GMA welds.

Travel Speed

Because the location of the arc inside the molten weld pool is important, the welding travel speed cannot exceed the ability of the arc to melt the base metal. Too high a travel speed can result in overrunning of the weld pool and an uncontrollable arc. Fusion between the base metal and filler metal can completely stop if the travel rate is too fast. If the travel rate is too slow and the weld pool size increases excessively, it can also restrict fusion to the base plate.

Electrode Extension

Module 5
Key Indicator 3, 8

The **electrode extension (stickout)** is the distance from the contact tube to the arc measured along the wire. Adjustments in this distance cause a change in the wire resistance and the resulting weld bead, Figure 1.17.

GMA welding currents are relatively high for the wire sizes, even for the low current values used in short-circuiting arc metal transfer, Figure 1.18. As the length of wire extending from the contact tube to the work increases, the voltage, too, should increase. Since this change is impossible with a constant-voltage power supply, the system compensates by reducing the current. In other words, by increasing the electrode extension and maintaining the same wire-feed speed, the current has to change to provide the same resistance drop. This situation leads to a reduction in weld heat, penetration, and fusion, and an increase in buildup. On the

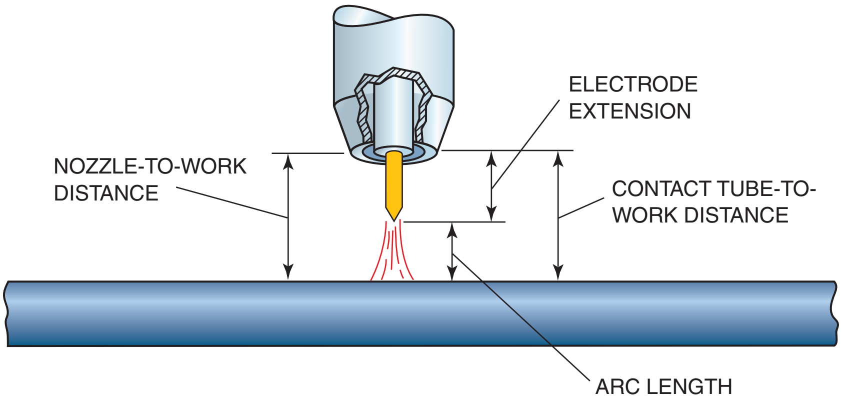

Figure 1.17 **Electrode-to-work distances**

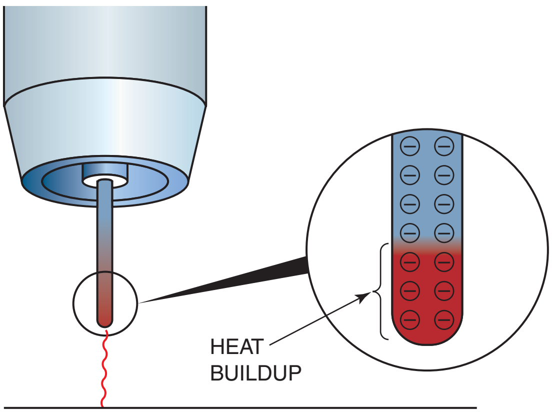

Figure 1.18 **Heat buildup due to the extremely high current for the small conductor (electrode)**

other hand, as the electrode extension distance is shortened, the weld heats up, penetrates more, and builds up less, Figure 1.19.

Experiment 2-3 explains the technique of using varying extension lengths to change the weld characteristics. Using this technique, a welder can make acceptable welds on metal ranging in thickness from 16 gauge to 1/4 in. (6 mm) or more without changing the machine settings. When using this technique, the nozzle-to-work distance should be kept the same so that enough shielding gas coverage is provided. Some nozzles can be extended to provide coverage. Others must be exchanged with the correct-length nozzle, Figure 1.20.

Gun Angle

Module 5
Key Indicator 3, 8

The GMA welding gun may be held so that the relative angle between the gun, work, and welding bead being made is either vertical or has a drag angle or a push angle. Changes in this angle will affect the weld bead. The effect is most noticeable during the short-circuiting arc and globular transfer modes.

Backhand welding is the welding technique that uses a drag angle, Figure 1.21. The welding technique that uses a push angle is known as forehand welding, Figure 1.22.

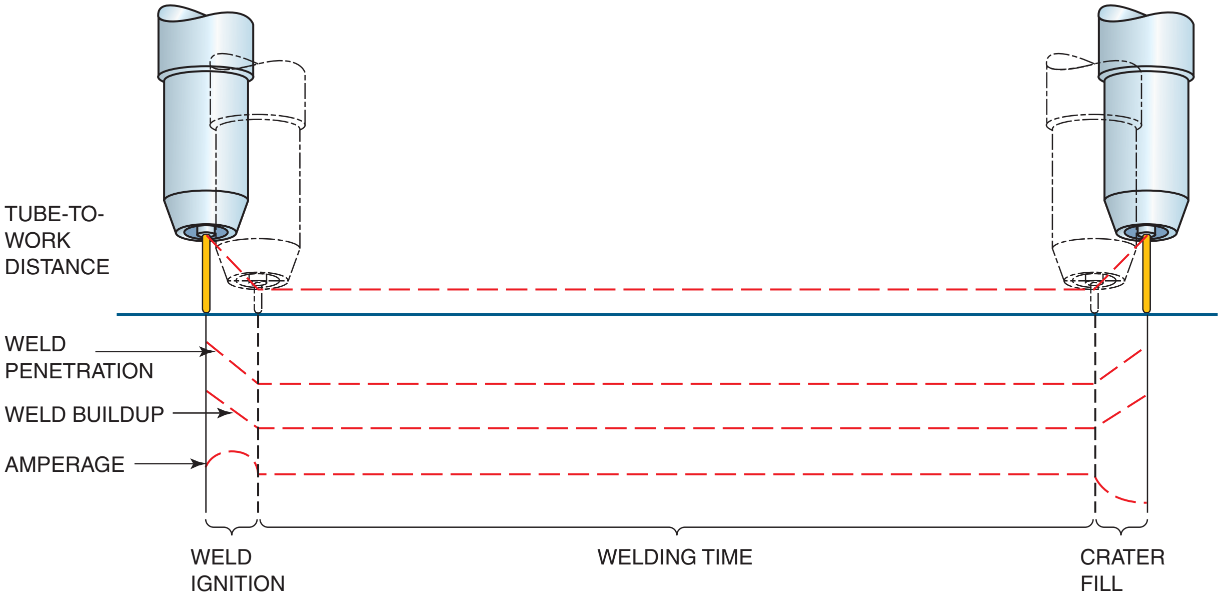
TUBE-TO-
WORK
DISTANCE
WELD
PENETRATION
WELD BUILDUP
AMPERAGE
WELD
IGNITION
WELDING TIME
CRATER
FILL

Figure 1.19 Using the changing tube-to-work distance to improve both the starting and stopping points of a weld

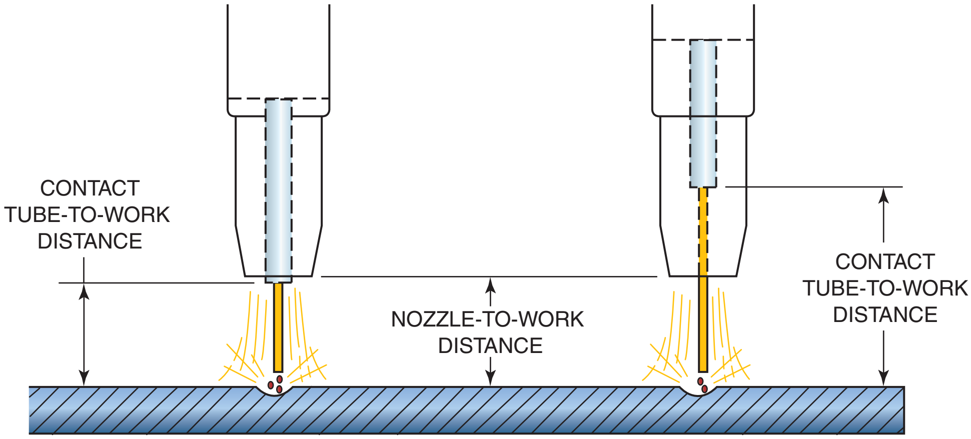
CONTACT
TUBE-TO-WORK
DISTANCE
NOZZLE-TO-WORK
DISTANCE
CONTACT
TUBE-TO-WORK
DISTANCE

Figure 1.20 Nozzle-to-work distance can differ from the contact tube-to-work distance

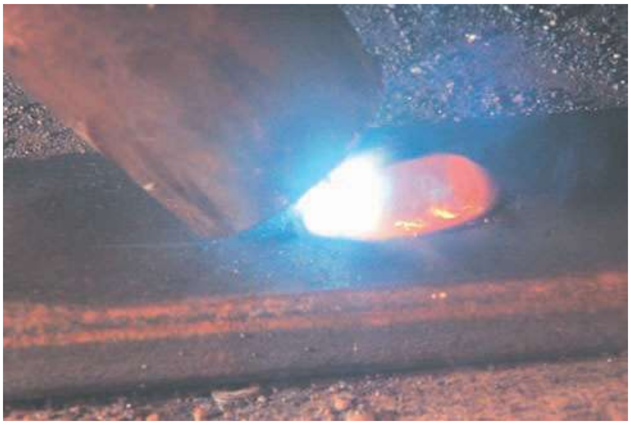

Figure 1.21 Backhand welding, or drag angle
Source: Courtesy of Larry Jeffus

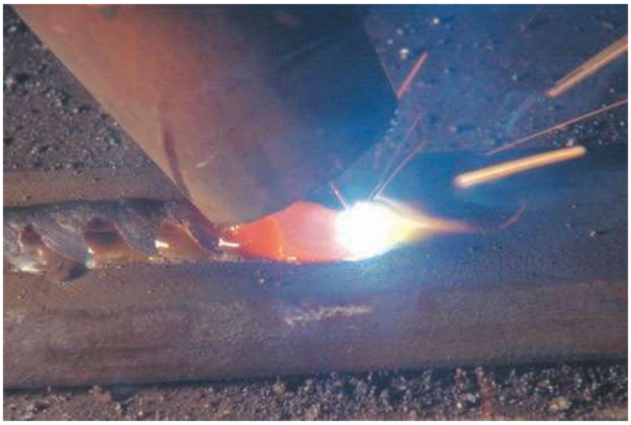

Figure 1.22 Forehand welding, or push angle
Source: Courtesy of Larry Jeffus

Backhand Welding

Module 5
Key Indicator 3, 8

A dragging angle, or backhand, welding technique directs the arc force into the molten weld pool of metal. This action, in turn, forces the molten metal back onto the trailing edge of the molten weld pool and exposes more of the unmelted base metal, Figure 1.23. The digging action pushes the penetration deeper into the base metal while building up the weld head. If the weld is sectioned, the profile of the bead is narrow and deeply penetrated, with high buildup.

Forehand Welding

Module 5
Key Indicator 3, 8

In a push angle, or forehand, welding technique, the arc force pushes the weld metal forward and out of the molten weld pool onto the cooler metal ahead of the weld, Figure 1.24. The heat and metal are spread out over a wider area. The sectional profile of the bead is wide, showing shallow penetration with little buildup.

The greater the angle, the more defined is the effect on the weld. As the angle approaches vertical, the effect is reduced. This allows the welder to change the weld bead as effectively as the changes resulting from adjusting the machine current settings.

EQUIPMENT

Module 5
Key Indicator 3, 8

The basic GMAW equipment consists of the gun, electrode (wire) feed unit, electrode (wire) supply, power source, shielding gas supply with flowmeter/regulator, control circuit, and related hoses, liners, and cables, Figure 1.25 and Figure 1.26. Larger, more complex systems may have water for cooling, solenoids for controlling gas flow, and carriages for moving the work or the gun or both, Figure 1.27. The system may be

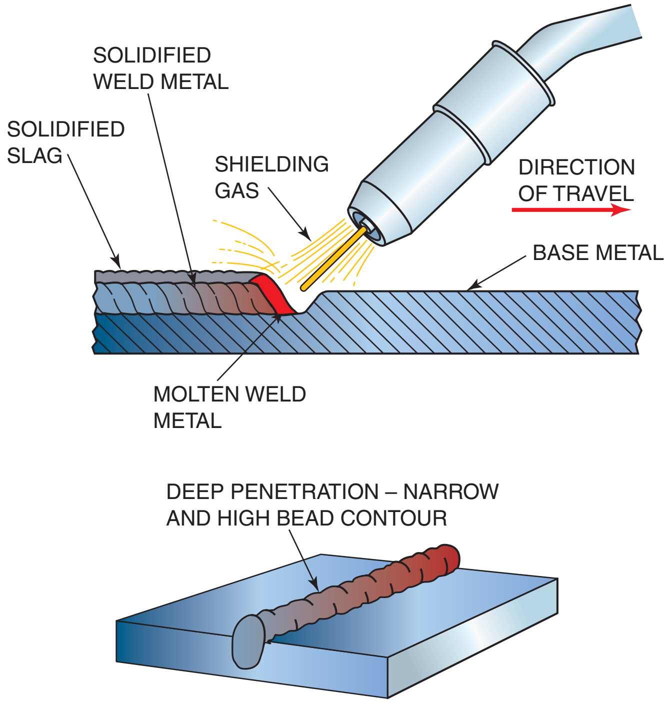

Figure 1.23 **Backhand welding, or dragging angle**

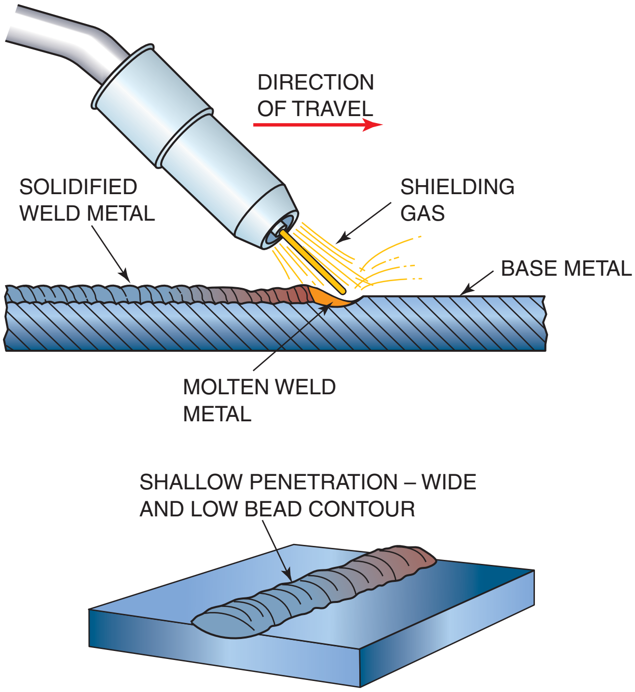
DIRECTION
OF TRAVEL
SOLIDIFIED
WELD METAL
SHIELDING
GAS
BASE METAL
MOLTEN WELD
METAL
SHALLOW PENETRATION – WIDE
AND LOW BEAD CONTOUR

Figure 1.24 Forehand welding, or pushing angle

stationary or portable, Figure 1.28. In most cases, the system is meant to be used for only one process. Some manufacturers, however, do make power sources that can be switched over for other uses.

Power Source

The power source may be either a transformer rectifier, inverter, or generator type. The transformers are stationary and commonly require a three-phase power source. The inverter power sources are smaller, lighter, and may be designed to accept a variety of different electrical inputs, from 208 volts to 440 volts, single or three-phase. Engine generators are ideal for portable use or where sufficient power is not available.

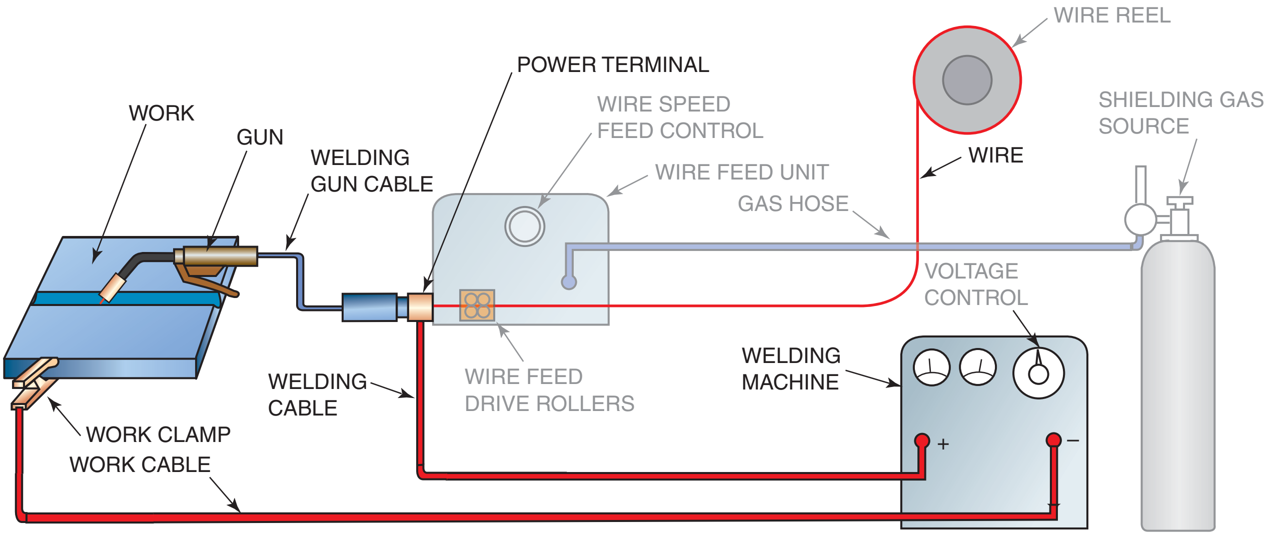
WIRE REEL
POWER TERMINAL
WORK
WIRE SPEED
FEED CONTROL
SHIELDING GAS
SOURCE
GUN
WELDING
GUN CABLE
WIRE
WIRE FEED UNIT
GAS HOSE
VOLTAGE
CONTROL
WELDING
MACHINE
WELDING
CABLE
WIRE FEED
DRIVE ROLLERS
WORK CLAMP
WORK CABLE
+
–

Figure 1.25 Schematic of equipment setup for GMA welding
Source: Courtesy of Hobart Brothers Company

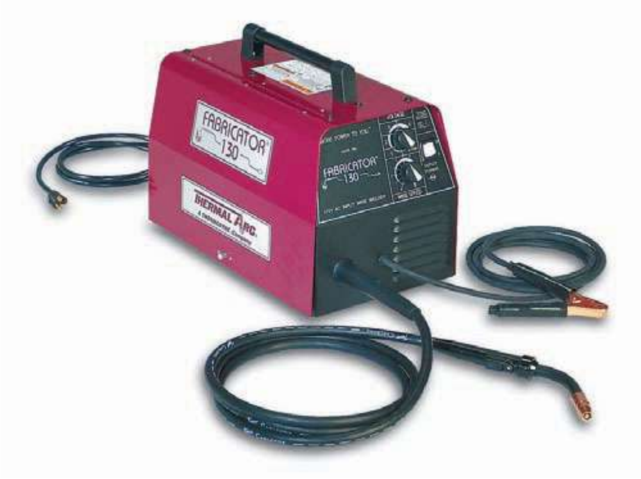

Figure 1.26 **Small 110-V GMA welder**
Source: Courtesy of Thermal Arc, a Thermadyne Company

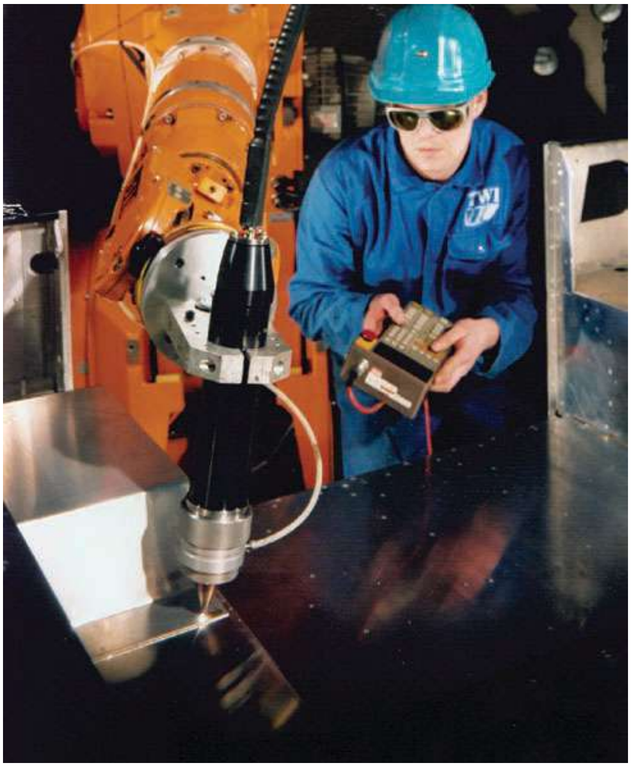

Figure 1.27 **Robot welding**
Source: Courtesy of ESAB Welding & Cutting Products

Figure 1.28 **Portable water cooler for GMA welding equipment**
Source: Courtesy of Lincoln Electric Company

Typical GMA welding machines produce a DC welding current ranging from 40 amperes to 600 amperes with 10 volts to 40 volts, depending upon the machine. In the past, some GMA processes used AC welding current, but DCRP is used almost exclusively now. Typical power supplies are shown in Figure 1.29.

Because many GMAW power supplies are used in automation and require long periods of continuous use, it is not unusual for GMA welding machines to have a 100% duty cycle. This allows the machine to be run continuously at its highest-rated output without damage.

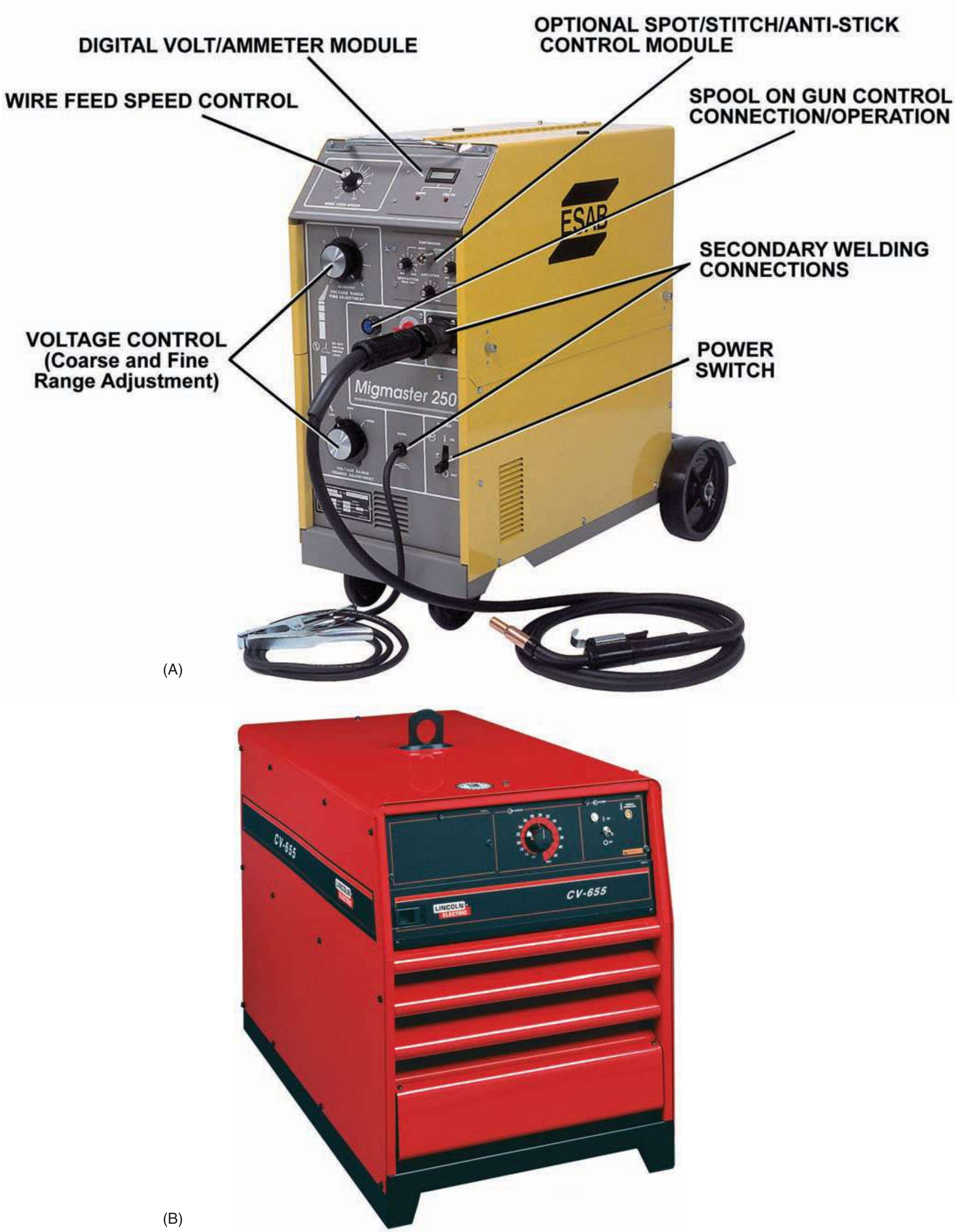
DIGITAL VOLT/AMMETER MODULE
OPTIONAL SPOT/STITCH/ANTI-STICK CONTROL MODULE
WIRE FEED SPEED CONTROL
SPOOL ON GUN CONTROL CONNECTION/OPERATION
SECONDARY WELDING CONNECTIONS
VOLTAGE CONTROL (Coarse and Fine Range Adjustment)
POWER SWITCH
ESAB
Migmaster 250
(A)
CV-655
LINCOLN ELECTRIC
CV-655
(B)

Figure 1.29 **Two power supplies for multipurpose GMAW applications**
(A) An expensive 200-ampere constant-voltage power supply, and (B) a 650-ampere constant-voltage and constant-current power supply.
Source: (A) Courtesy of ESAB Welding & Cutting Products (B) Courtesy of Lincoln Electric Company

Electrode (Wire) Feed Unit

The purpose of the electrode feeder is to provide a steady and reliable supply of wire to the weld. Slight changes in the rate at which the wire is fed have distinct effects on the weld.

The motor used in a feed unit is usually a DC type that can be continuously adjusted over the desired range. Figure 1.30 and Figure 1.31 show typical wire-feed units and accessories.

Push-type Feed System

The wire rollers are clamped securely against the wire to provide the necessary friction to push the wire through the conduit to the gun. The pressure applied on the wire can be adjusted. A groove is provided in the roller to aid in alignment and to lessen the chance of slippage. Most manufacturers provide rollers with smooth or knurled U-shaped or V-shaped grooves, Figure 1.32. Knurling (a series of ridges cut into the groove) helps grip larger-diameter wires so that they can be pushed along more easily. Soft wires, such as aluminum, are easy to damage if knurled rollers are used. Soft wires are best used with U-grooved rollers. Even V-grooved rollers can distort the surface of the wire, causing problems. V-grooved rollers are best suited for hard wires, such as mild steel and stainless steel. It is also important to use the correct-size grooves in the rollers.

Variations of the push-type electrode wire feeder include the pull type and push-pull type. The difference is in the size and location of the drive rollers. In the push-type system, the electrode must have enough strength to be pushed through the conduit without kinking. Mild steel and stainless steel can be readily pushed 15 to 20 ft (4 to 6 m), but aluminum is much harder to push more than 10 ft (3 m).

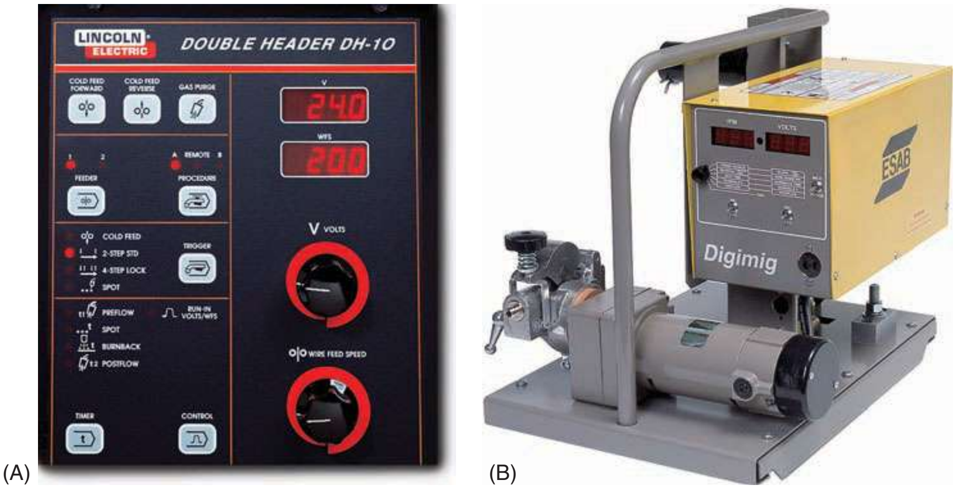
LINCOLN ELECTRIC
DOUBLE HEADER DH-10
COLD FEED FORWARD
COLD FEED REVERSE
GAS PURGE
V
240
WFS
200
FEEDER
REMOTE
PROCEDURE
COLD FEED
2-STEP STD
4-STEP LOCK
SPOT
TRIGGER
V VOLTS
PREFLOW
SPOT
BURNBACK
POSTFLOW
RUN-IN VOLTS/WFS
WIRE FEED SPEED
TIMER
CONTROL
ESAB
Digimig
(A)
(B)

Figure 1.30 Examples of wire feeders
(A) A 90-ampere power supply and wire feeder for welding sheet steel with carbon dioxide shielding. (B) Modern wire feeder with digital preset and readout of wire-feed speed and closed-loop control.
Source: (A) Courtesy of Lincoln Electric Company (B) Courtesy of ESAB Welding & Cutting Products

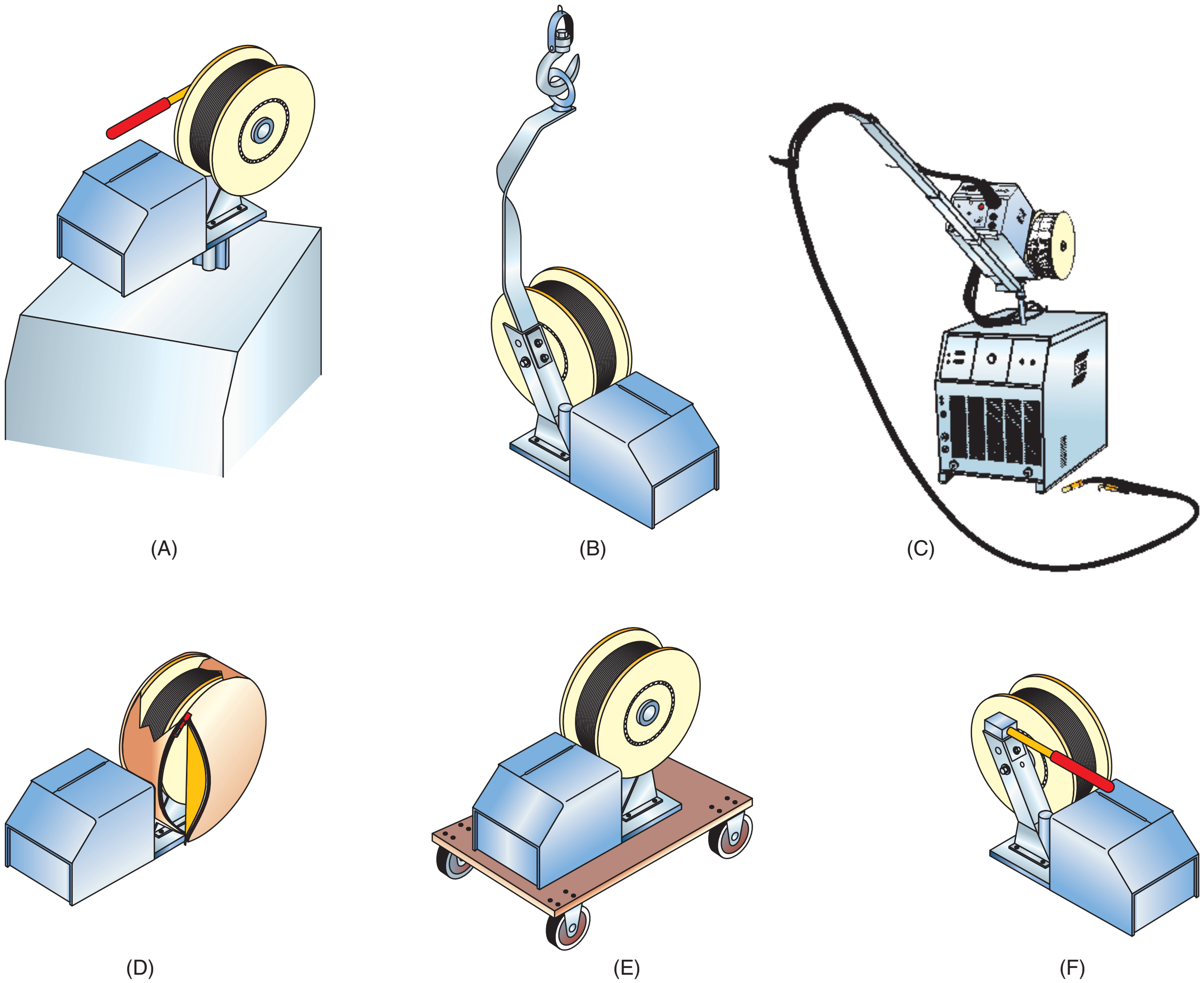
(A)
(B)
(C)
(D)
(E)
(F)

Figure 1.31 **A variety of accessories are available for most electrode feed systems.**
(A) Swivel post, (B) boom hanging bracket, (C) counterbalance mini-boom, (D) spool cover, (E) wire feeder wheel cart, and (F) carrying handle.

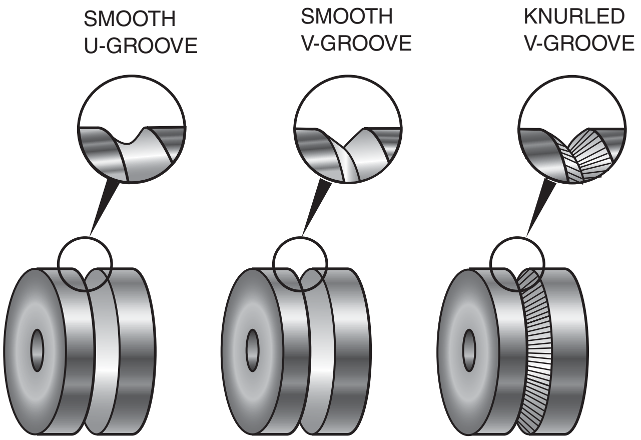

Figure 1.32 **Feed rollers**

Pull-type Feed System

In pull-type systems, a smaller but higher-speed motor is located in the gun to pull the wire through the conduit. Using this system, it is possible to move even soft wire over great distances. The disadvantages are that the gun is heavier and more difficult to use, rethreading the wire takes more time, and the operating life of the motor is shorter.

Push-pull-type Feed System

Push-pull-type feed systems use a synchronized system with feed motors located at both ends of the electrode conduit, Figure 1.33. This system can be used to move any type of wire over long distances by periodically installing a feed roller into the electrode conduit. Compared to the pull-type system, the advantages of this system include the ability to move wire over longer distances, faster rethreading, and increased motor life due to the reduced load. A disadvantage is that the system is more expensive.

Linear Electrode Feed System

Linear electrode feed systems use a different method to move the wire and change the feed speed. Standard systems use rollers that pinch the wire between the rollers. A system of gears is used between the motor and rollers to provide roller speed within the desired range. The linear feed system does not have gears or conventional-type rollers.

The linear feed system uses a small motor with a hollow armature shaft through which the wire is fed. The rollers are attached so that they move around the wire. Changing the roller pitch (angle) changes the speed at which the wire is moved without changing the motor speed.

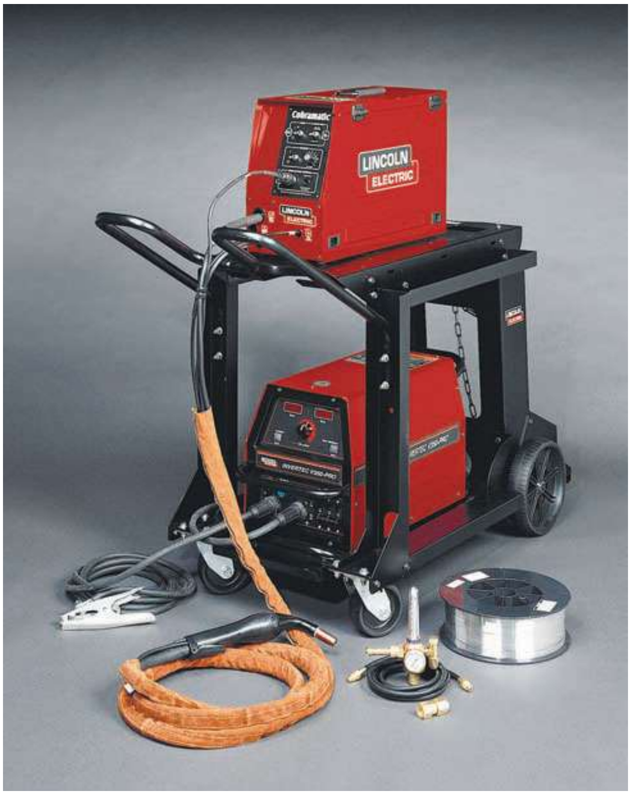

Figure 1.33 **Wire-feed system that enables the wire to be moved through a longer cable**
Source: Courtesy of Lincoln Electric Company

This system works in the same way that changing the pitch on a screw, either coarse threads or fine threads, affects the rate that the screw will move through a spinning nut.

The advantage of a linear system is that the bulky system of gears is eliminated, thus reducing weight, size, and wasted power. The motor operates at a constant high speed where it is more efficient. The reduced size allows the system to be housed in the gun or within an enclosure in the cable. Several linear wire feeders can be synchronized to provide an extended operating range. The disadvantage of a linear system is that the wire may become twisted as it is moved through the feeder.

Spool Gun

A spool gun is a compact, self-contained system consisting of a small drive system and a wire supply, Figure 1.34A. This system allows the welder to move freely around a job with only a power lead and shielding gas hose to manage. The major control system is usually mounted on the welder. The feed rollers and motor are found in the gun just behind the

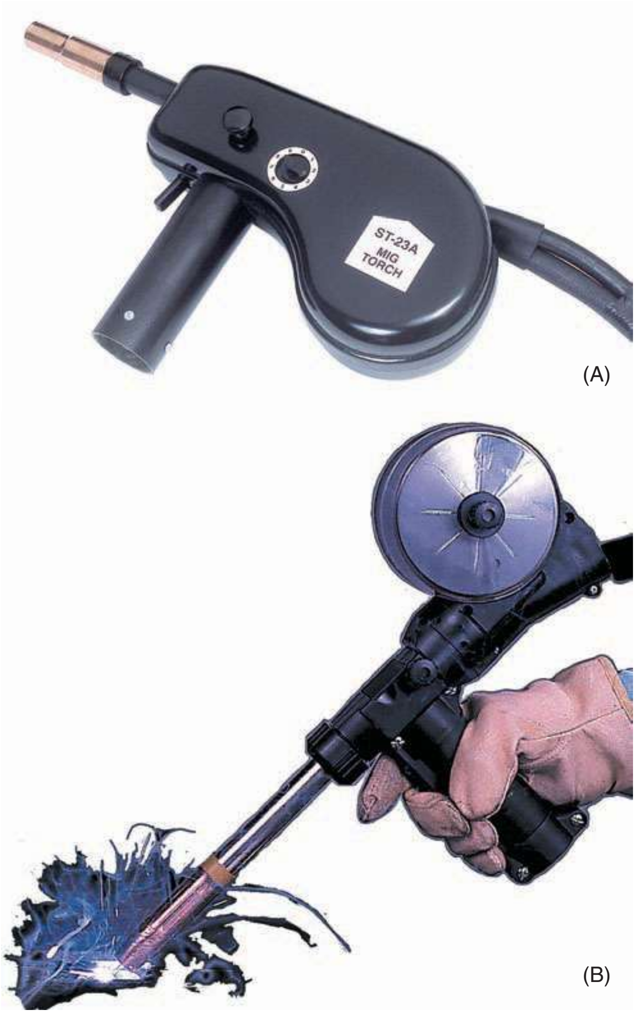

Figure 1.34 **Feeder/guns for GMA welding**
Source: Courtesy of ESAB Welding & Cutting Products

nozzle and contact tube, Figure 1.34B. Because of the short distance the wire must be moved, very soft wires (aluminum) can be used. A small spool of welding wire is located just behind the feed rollers. The small spools of wire required in these guns are often very expensive. Although the guns are small, they feel heavy when being used.

Electrode Conduit

The electrode conduit or liner guides the welding wire from the feed rollers to the gun. It may be encased in a lead that contains the shielding gas.

Power cable and gun switch circuit wires are contained in a conduit that is made of a tightly wound coil having the needed flexibility and strength. The steel conduit may have a nylon or Teflon liner to protect soft, easily scratched metals, such as aluminum, as they are fed.

If the conduit is not an integral part of the lead, it must be firmly attached to both ends of the lead. Failure to attach the conduit can result in misalignment, which causes additional drag or makes the wire jam completely. If the conduit does not extend through the lead casing to make a connection, it can be drawn out by tightly coiling the lead, Figure 1.35. Coiling will force the conduit out so that it can be connected. If the conduit is too long for the lead, it should be cut off and filed smooth. Too long a lead will bend and twist inside the conduit, which may cause feed problems.

Welding Gun

The welding gun attaches to the end of the power cable, electrode conduit, and shielding gas hose, Figure 1.36. It is used by the welder to produce the weld. A trigger switch is used to start and stop the weld cycle. The gun also has a contact tube, which is used to transfer welding current to the electrode moving through the gun, and a gas nozzle, which directs the shielding gas onto the weld, Figure 1.37.

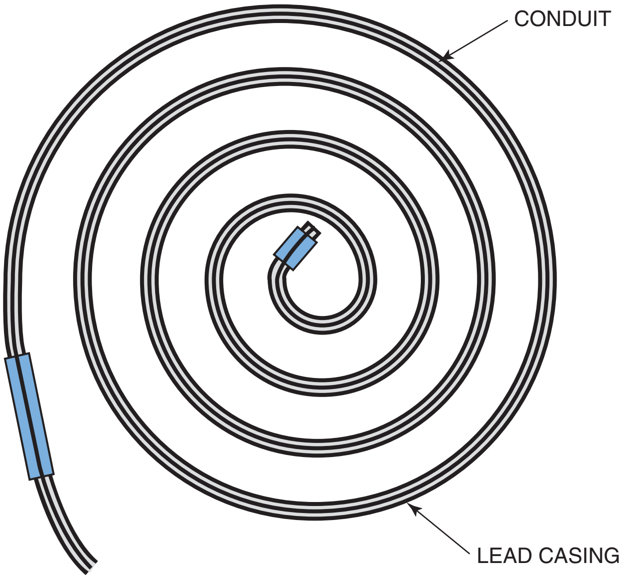

Figure 1.35 **Tightly coiled lead casing will force the liner out of the gun**

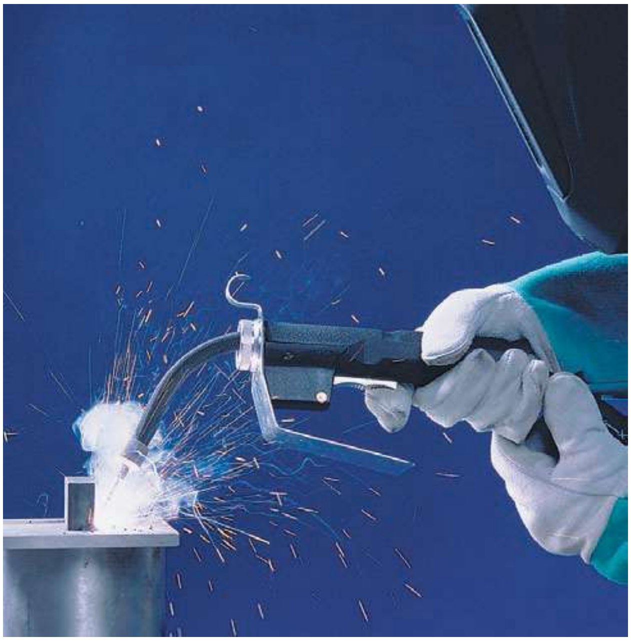

Figure 1.36 **A typical GMA welding gun**
Guns like this are used for most welding processes with a heat shield attached to protect the welder's gloved hand from intense heat generated when welding with high amperages.
Source: Courtesy of Tweco, a Thermadyne Company

SPOT WELDING

GMA can be used to make high-quality arc spot welds. Welds can be made using standard or specialized equipment, Figure 1.38. The arc spot weld produced by GMAW differs from electric resistance spot welding. The GMAW spot weld starts on one surface of one member and burns through to the other member, Figure 1.39. Fusion between the members occurs, and a small nugget is left on the metal surface.

GMA spot welding has some advantages such as the following: (1) welds can be made in thin-to-thick materials; (2) the weld can be made when only one side of the materials to be welded is accessible; and (3) the weld can be made when there is paint on the interfacing surfaces. The arc spot weld can also be used to assemble parts for welding to be done at a later time.

Thin metal can be attached to thicker sections using an arc spot weld. If a thin-to-thick butt, lap, or tee joint is to be welded with complete joint penetration, often the thin material will burn back, leaving a hole, or there will not be enough heat to melt the thick section. With an arc spot weld, the burning back of the thin material allows the thicker metal to be melted. As more metal is added to the weld, the burn-through is filled, Figure 1.39.

The GMA spot weld is produced from only one side. Therefore, it can be used on awkward shapes and in cases where the other side of the surface being welded should not be damaged. This makes it an excellent process for auto body repair. In addition, because the metals are melted

Caution

Safety glasses and/or flash glasses must be worn to protect the eyes from flying sparks.

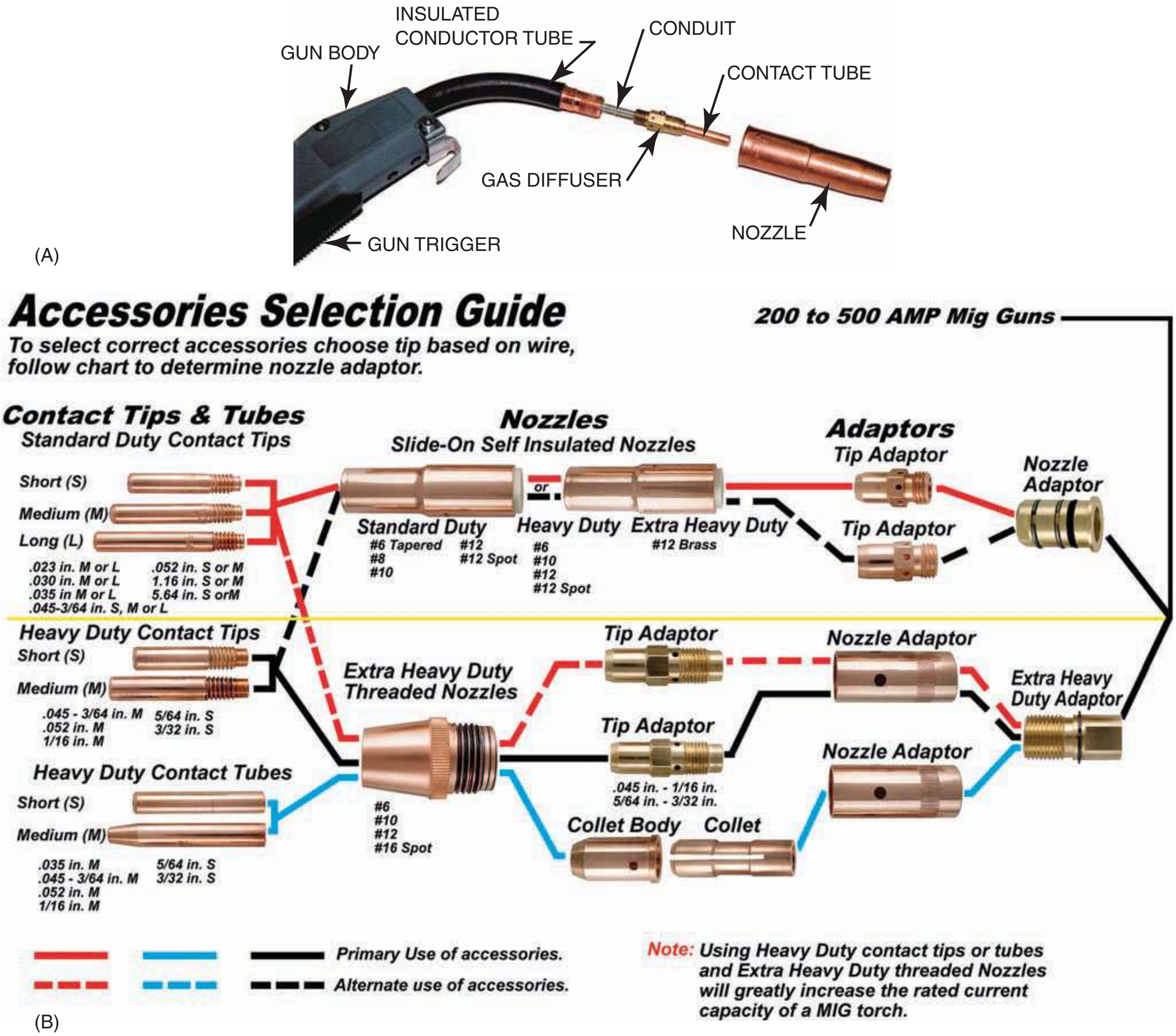
(A)
GUN BODY
INSULATED CONDUCTOR TUBE
CONDUIT
CONTACT TUBE
GAS DIFFUSER
NOZZLE
GUN TRIGGER
(B)
Accessories Selection Guide
To select correct accessories choose tip based on wire, follow chart to determine nozzle adaptor.
200 to 500 AMP Mig Guns
Contact Tips & Tubes
Standard Duty Contact Tips
Short (S)
Medium (M)
Long (L)
.023 in. M or L
.030 in. M or L
.035 in M or L
.045-3/64 in. S, M or L
.052 in. S or M
1.16 in. S or M
5.64 in. S orM
Nozzles
Slide-On Self Insulated Nozzles
or
Standard Duty
#6 Tapered
#8
#10
#12
#12 Spot
Heavy Duty
#6
#10
#12
#12 Spot
Extra Heavy Duty
#12 Brass
Adaptors
Tip Adaptor
Tip Adaptor
Nozzle Adaptor
Heavy Duty Contact Tips
Short (S)
Medium (M)
.045 - 3/64 in. M
.052 in. M
1/16 in. M
5/64 in. S
3/32 in. S
Extra Heavy Duty Threaded Nozzles
#6
#10
#12
#16 Spot
Tip Adaptor
Nozzle Adaptor
Tip Adaptor
.045 in. - 1/16 in.
5/64 in. - 3/32 in.
Nozzle Adaptor
Extra Heavy Duty Adaptor
Collet Body
Collet
Heavy Duty Contact Tubes
Short (S)
Medium (M)
.035 in. M
.045 - 3/64 in. M
.052 in. M
1/16 in. M
5/64 in. S
3/32 in. S
Primary Use of accessories.
Alternate use of accessories.
Note: Using Heavy Duty contact tips or tubes and Extra Heavy Duty threaded Nozzles will greatly increase the rated current capacity of a MIG torch.

Figure 1.37 **GMA welding gun parts**
(A) Typical replaceable parts of a GMA welding gun. (B) Accessories and parts selection guide for a GMA welding gun.
Source: Courtesy of ESAB Welding & Cutting Products

and the molten weld pool is agitated, thin films of paint between the members being joined need not be removed. This is an added benefit for auto body repair work.

Specially designed nozzles provide flash protection, part alignment, and arc alignment, Figure 1.40. As a result, for some small jobs it is possible to perform the weld with only safety glasses. The optional control timer provides weld time and burn-back time. To make a weld, the amperage, voltage, and length of welding time must be set correctly. The burn-back time is a short period at the end of the weld when the wire feed stops but the current does not. This allows the wire to be burned back so it does not stick in the weld, Figure 1.39D.

Caution

It is not advisable for any spot welding work requiring more than just a few spot welds to be done without full welder's safety gear. Prolonged exposure to the reflected ultraviolet light will cause skin burns.

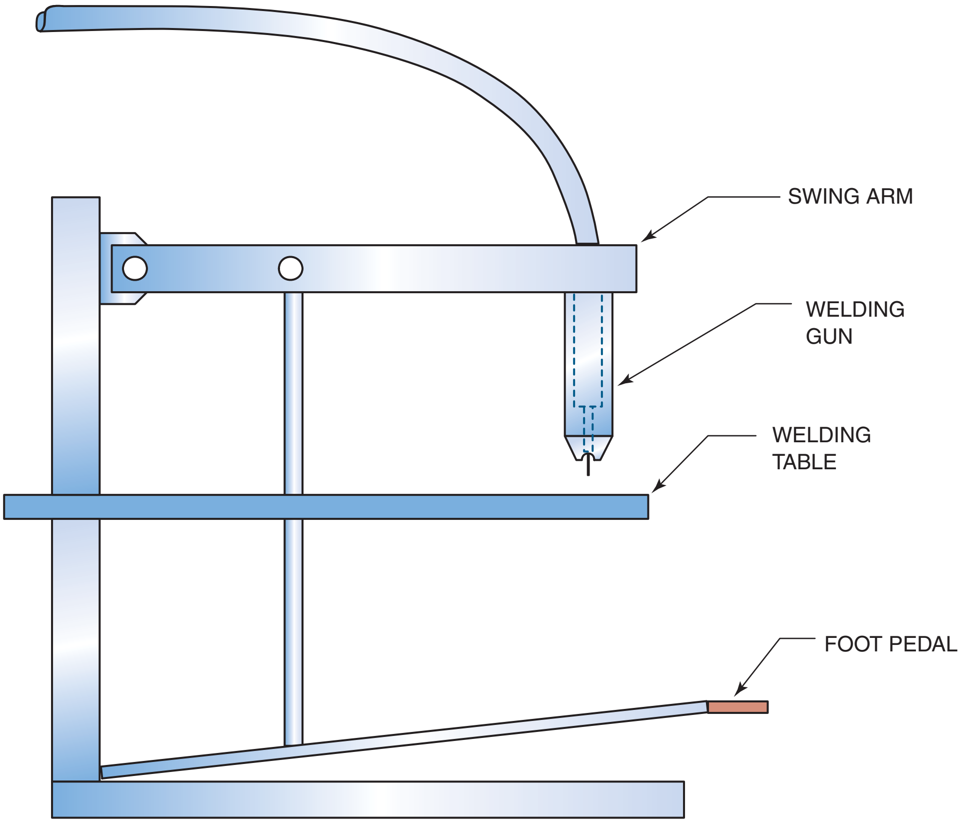

Figure 1.38 GMA spot welding machine

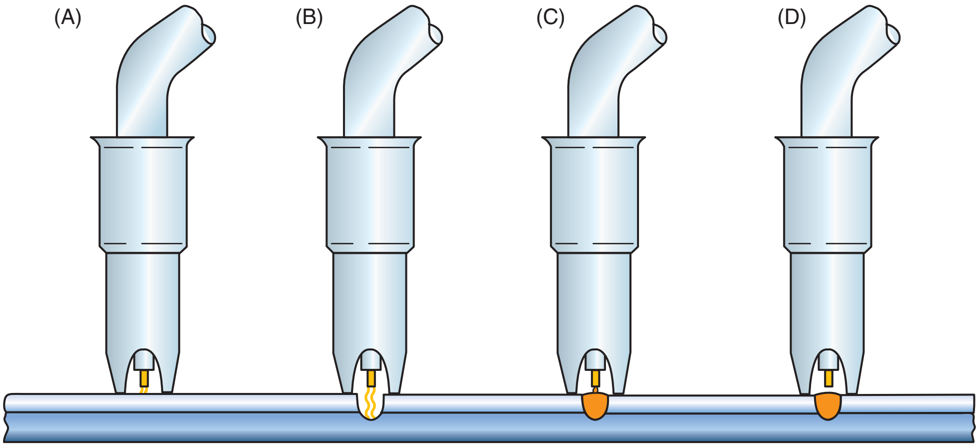

Figure 1.39 A GMA spot weld
(A) The arc starts, (B) a hole is burned through the first plate, (C) the hole is filled with weld metal, and (D) the wire feed stops and the arc burns the electrode back.

FACE, EYE, AND EAR PROTECTION

Module 2
Key Indicator 1

Module 5
Key Indicator 1

Face and Eye Protection

Eye protection must be worn in the shop at all times. Eye protection can be safety glasses with side shields, Figure 1.41, goggles, or a full face shield. To give better protection when working in brightly lit areas or outdoors, some welders wear flash glasses, which are special, lightly tinted, safety glasses. These safety glasses provide protection from both flying debris and reflected light.

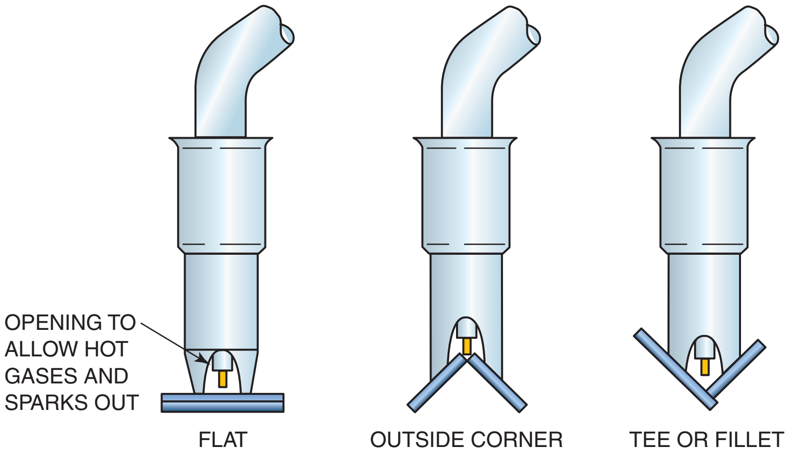

Figure 1.40 **Specialized nozzles for GMA spot welding**

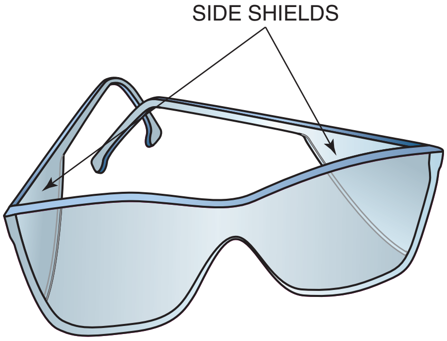

Figure 1.41 **Safety glasses with side shields**

Suitable eye protection is important because eye damage caused by excessive exposure to arc light is not noticed. Welding light damage often occurs without warning, like a sunburn's effect that is felt the following day. Therefore, welders must take appropriate precautions in selecting filters or goggles that are suitable for the process being used, Figure 1.42. Selecting the correct shade lens is also important, because both extremes of too light or too dark can cause eye strain. New welders often select too dark a lens, assuming it will give them better protection, but this results in eye strain in the same manner as if they were trying to read in a poorly lit room. In reality, any approved arc welding lenses will filter out the harmful ultraviolet light. Select a lens that lets you see comfortably. At the very least, the welder's eyes must not be strained by excessive glare from the arc.

Ultraviolet light can burn the eye in two ways. This light can injure either the white of the eye or the retina, which is the back of the eye, Figure 1.43. Burns on the retina are not painful but may cause some loss of eyesight. The whites of the eyes are very sensitive, and burns are very painful. The eyes are easily infected because, as with any burn, many cells are killed. These dead cells in the moist environment of the eyes will promote the growth of bacteria that cause infection. When the eye is burned, it feels as though there is something in the eye. Without a professional examination, however, it is impossible to tell if there is something in the eye. Because there may actually be something in the eye and because of the high risk of infection, home remedies or other medicines should never be used for eye burns. Anytime you receive an eye injury you should see a doctor.

Welding Helmets

Even with quality **welding helmets**, like those shown in Figure 1.44, the welder must check for potential problems that may occur from accidents or daily use. Small, undetectable leaks of ultraviolet light in an arc welding helmet can cause a welder's eyes to itch or feel sore after a day of welding. To prevent these leaks, make sure the lens gasket is installed correctly, Figure 1.45. The outer and inner clear lenses must be plastic. As shown in Figure 1.46, a lens can be checked for cracks by twisting it between your fingers. Worn or cracked spots on a helmet must be repaired.

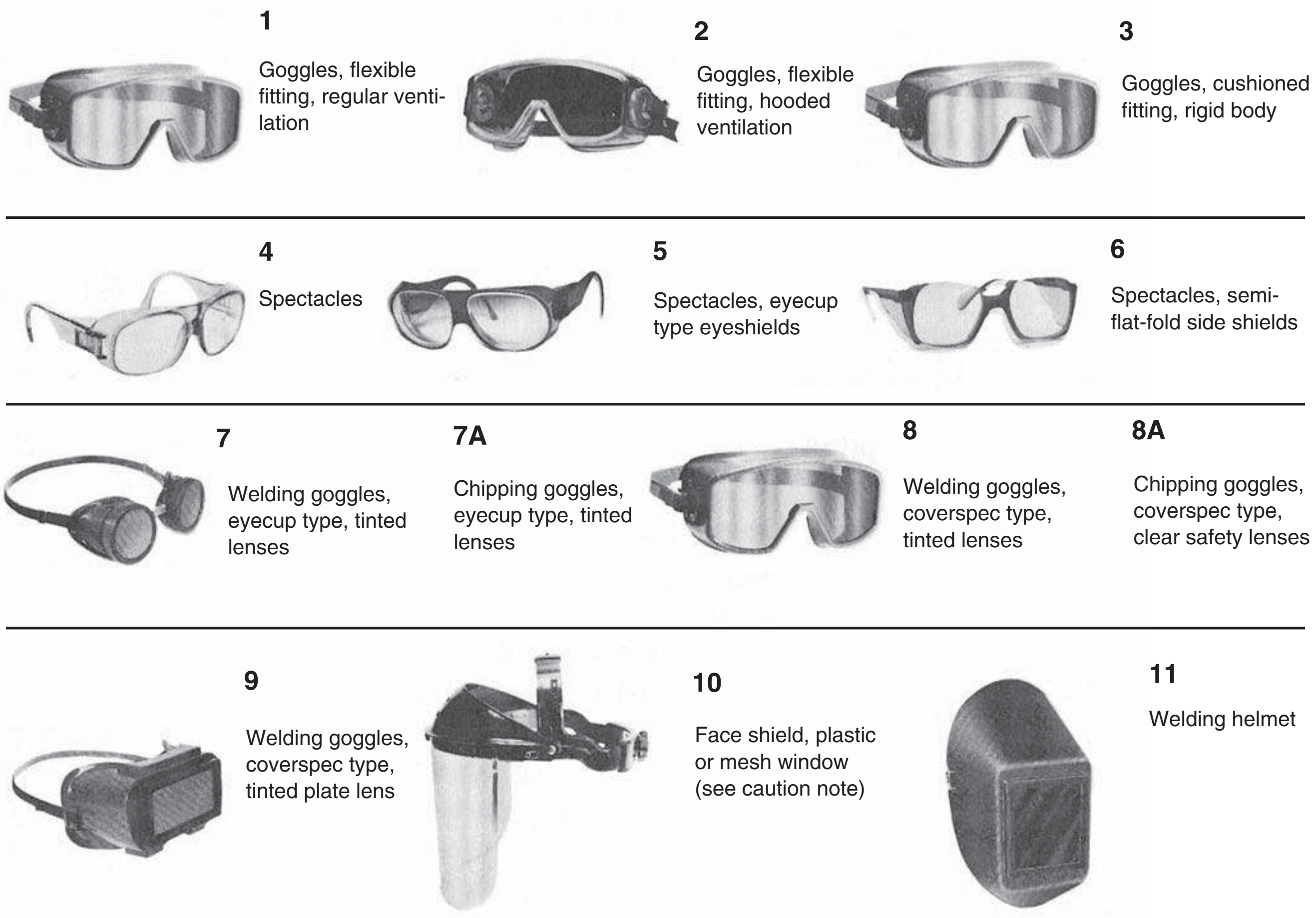
1
Goggles, flexible fitting, regular ventilation
2
Goggles, flexible fitting, hooded ventilation
3
Goggles, cushioned fitting, rigid body
4
Spectacles
5
Spectacles, eyecup type eyeshields
6
Spectacles, semi-flat-fold side shields
7
Welding goggles, eyecup type, tinted lenses
7A
Chipping goggles, eyecup type, tinted lenses
8
Welding goggles, coverspec type, tinted lenses
8A
Chipping goggles, coverspec type, clear safety lenses
9
Welding goggles, coverspec type, tinted plate lens
10
Face shield, plastic or mesh window (see caution note)
11
Welding helmet

**Non-side-shield spectacles are available for limited hazard use requiring only frontal protection.*

Applications

Operation	Hazards	Protectors
Acetylene-burning Acetylene-cutting Acetylene-welding	Sparks, harmful rays, molten metal, flying particles	7,8,9
Chemical handling	Splash, acid burns, fumes	2 (for severe exposure add 10)
Chipping	Flying particles	1,2,4,5,6,7A,8A
Electric (arc) welding	Sparks, intense rays, molten metal	11 (in combination with 4,5,6 in tinted lenses advisable)
Furnace operations	Glare, heat, molten metal	7,8,9 (for severe exposure add 10)
Grinding—light	Flying particles	1,3,5,6 (for severe exposure add 10)
Grinding—heavy	Flying particles	1,3,7A,8A (for severe exposure add 10)
Laboratory	Chemical splash, glass breakage	2 (10 when in combination with 5,6)
Machining	Flying particles	1,3,5,6 (for severe exposure add 10)
Molten metals	Heat, glare, sparks, splash	7,8 (10 in combination with 5,6 in tinted lenses)
Spot welding	Flying particles, sparks	1,3,4,5,6 (tinted lenses advisable, for severe exposure add 10)

CAUTION:
Face shields alone do not provide adequate protection. Plastic lenses are advised for protection against molten metal splash. Contact lenses, of themselves, do not provide eye protection in the industrial sense and shall not be worn in a hazardous environment without appropriate covering safety eyewear.

Figure 1.42 **Huntsman selector chart**
Source: Courtesy of Kedman Co., Huntsman Product Division

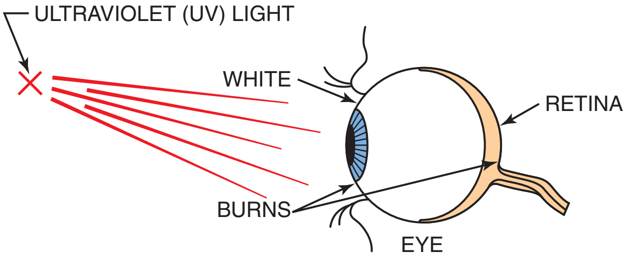
ULTRAVIOLET (UV) LIGHT
WHITE
RETINA
BURNS
EYE

Figure 1.43 **The eye can be burned on the white or on the retina by ultraviolet light**

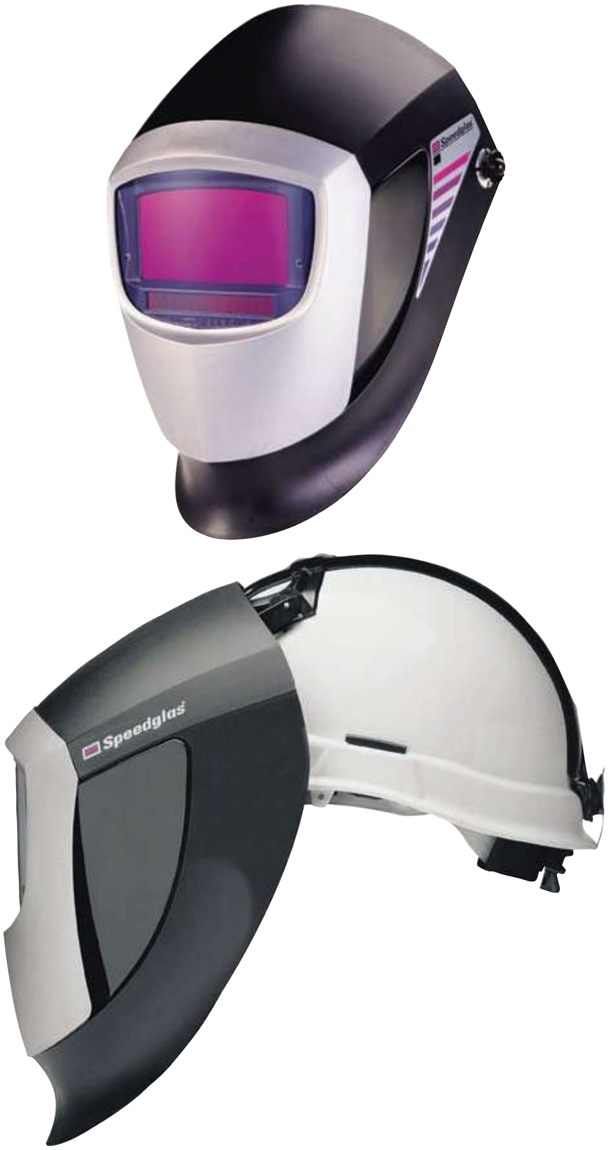
Speedglas
Speedglas

Figure 1.44 **Typical arc welding helmets for eye and face protection during welding**
Source: Courtesy of Hornell, Inc.

Safety Glasses

Safety glasses with side shields are adequate for general use, but if heavy grinding, chipping, or overhead work is being done, goggles or a full face shield should be worn in addition to safety glasses, Figure 1.47. Safety glasses are best for general protection. They must be worn under an arc welding helmet at all times.

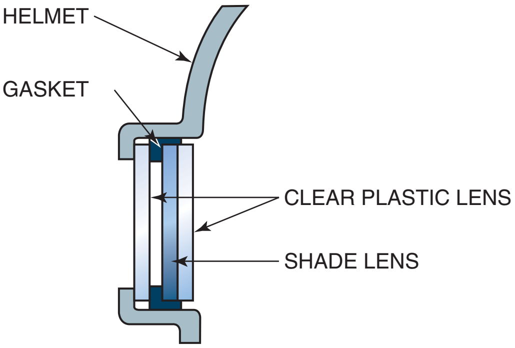

Figure 1.45 **Correct placement of the gasket around the shade lens**
Correct gasket placement around the shade lens of a welding helmet is important because it can stop ultraviolet light from bouncing around the lens assembly.

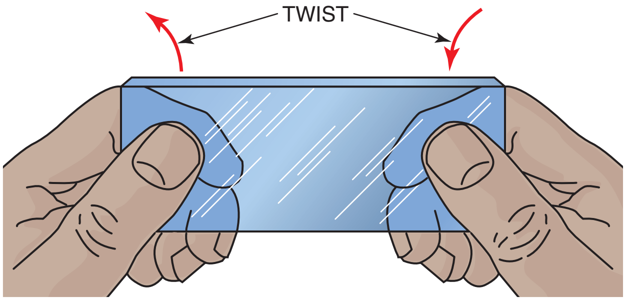

Figure 1.46 **To check the shade lens for possible cracks, gently twist it**

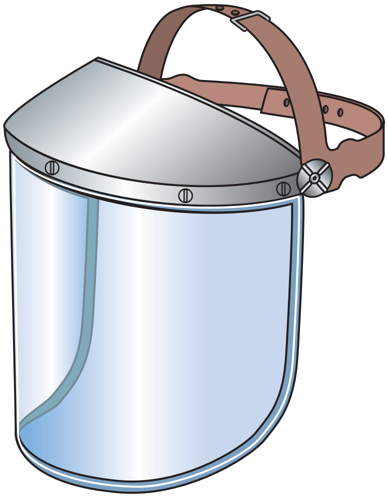

Figure 1.47 **Full face shield**

Ear Protection

The welding environment can be very noisy. The sound level is at times high enough to cause pain and some loss of hearing if the welder's ears are unprotected. Hot sparks can also drop into an open ear, causing severe burns.

Ear protection is available in several forms. One form of protection is earmuffs that cover the outer ear completely, Figure 1.48. Another form

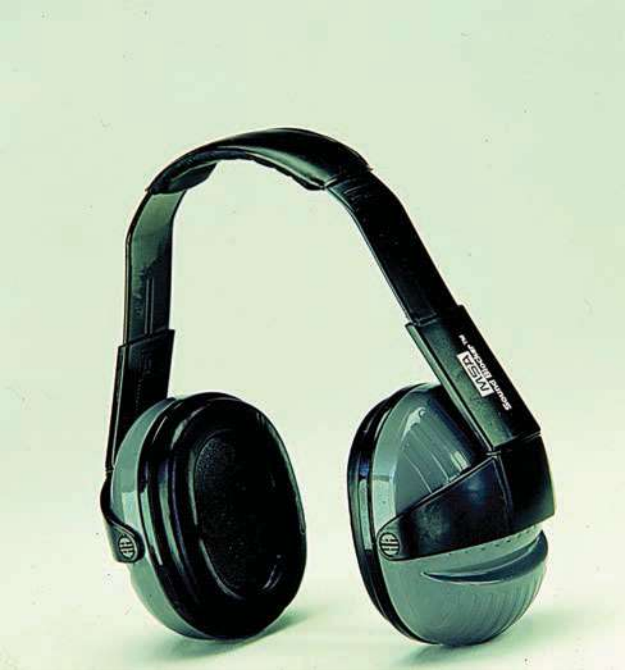

Figure 1.48 Earmuffs
Earmuffs provide complete ear protection and can be worn under a welding helmet.
Source: Courtesy of Mine Safety Appliances Company

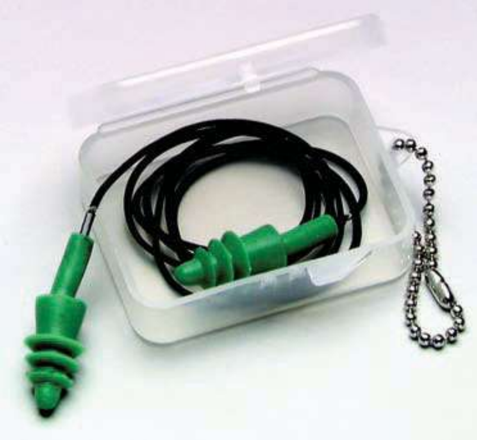

Figure 1.49 Earplugs
Earplugs are used as protection from noise only.
Source: Courtesy of Mine Safety Appliances Company

of protection is earplugs that fit into the ear canal, Figure 1.49. Both of these protect a person's hearing, but only the earmuffs protect the outer ear from burns.

GENERAL WORK CLOTHING

Special protective clothing cannot be worn at all times. It is, therefore, important to choose general work clothing that will minimize the possibility of getting burned because of the high temperature and amount of hot sparks, metal, and slag produced during welding, cutting, or brazing.

Caution

Damage to your hearing caused by high sound levels may not be detected until later in life, and the resulting loss in hearing is permanent. Your hearing will not improve with time, and each exposure to high levels of sound will further damage your hearing.

Work clothing must also stop ultraviolet light from passing through it. This is accomplished if the material chosen is a dark color, thick, and tightly woven. The best choice is 100% wool but it is difficult to find. Another good choice is 100% cotton clothing, the most popular fabric used.

You must avoid wearing synthetic materials including nylon, rayon, and polyester. They can easily melt or catch fire. Some synthetics produce a hot, sticky residue that can make burns more severe. Others may produce poisonous gases.

The following are some guidelines for selecting work clothing:

- Shirts must be long-sleeved to protect the arms, have a high-buttoned collar to protect the neck, Figure 1.50, be long enough to tuck into the pants to protect the waist, and have flaps on the pockets to keep sparks out (or have no pockets).
- Pants must have legs long enough to cover the tops of the boots and must be without cuffs that would catch sparks.
- Boots must have high tops to keep out sparks, have steel toes to prevent crushed toes, Figure 1.51, and have smooth tops to prevent sparks from being trapped in seams.
- Caps should be thick enough to prevent sparks from burning the top of a welder's head.

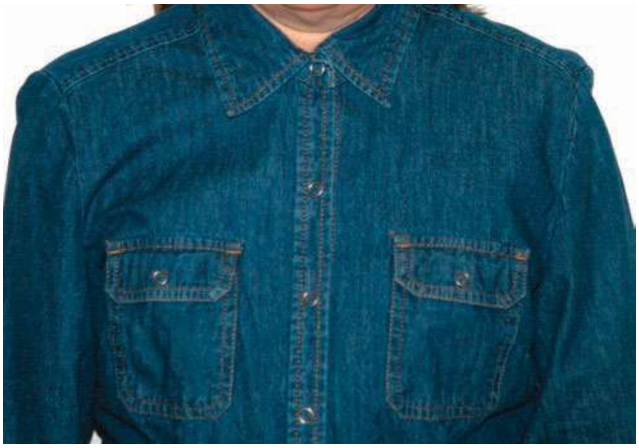

Figure 1.50 Neck protection
The top button of the shirt worn by the welder should always be buttoned in order to avoid severe neck burns.
Source: Courtesy of Larry Jeffus

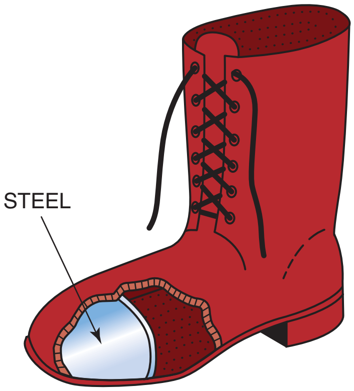

Figure 1.51 Safety boots with steel toes are required by many welding shops

All clothing must be free of frayed edges and holes. The clothing must be relatively tight-fitting in order to prevent excessive folds or wrinkles that might trap sparks.

Some welding clothes have pockets on the inside to prevent the pockets from collecting sparks. However, it is not safe to carry butane lighters or matches in these or any pockets while welding. Lighters and matches can easily catch on fire or explode if they are subjected to the heat and sparks of welding.

Caution

There is no safe place to carry butane lighters or matches while welding or cutting. They can catch fire or explode if subjected to welding heat or sparks. Butane lighters may explode with the force of a quarter of a stick of dynamite. Matches can erupt into a ball of fire. Both butane lighters and matches must always be removed from the welder's pockets and placed a safe distance away before any work is started.

SPECIAL PROTECTIVE CLOTHING

General work clothing is worn by each person in the shop. In addition to this clothing, extra protection is needed for each person who is in direct contact with hot materials. Leather is often the best material to use, as it is lightweight, is flexible, resists burning, and is readily available. Synthetic insulating materials are also available. Ready-to-wear leather protection includes capes, jackets, aprons, sleeves, gloves, caps, pants, knee pads, and spats, among other items.

Hand Protection

All-leather, gauntlet-type gloves should be worn when doing any welding, Figure 1.52. Gauntlet gloves that have a cloth liner for insulation are best for hot work. Noninsulated gloves will give greater flexibility for fine work. Some leather gloves are available with a canvas gauntlet top, which should be used for light work only.

When a great deal of manual dexterity is required for gas tungsten arc welding, brazing, soldering, oxyfuel gas welding, and other delicate processes, soft leather gloves may be used, Figure 1.53. All-cotton gloves are sometimes used when doing very light welding.

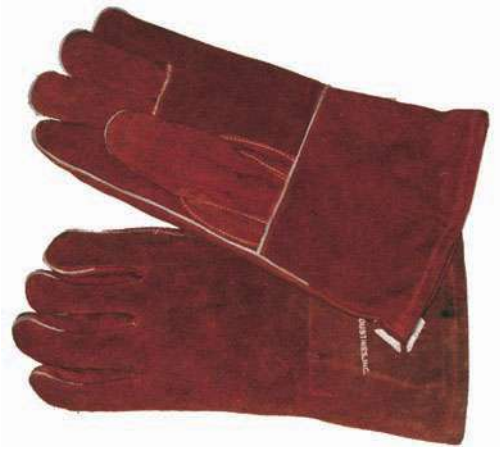

Figure 1.52 **All-leather, gauntlet-type welding gloves**
Source: Courtesy of Larry Jeffus

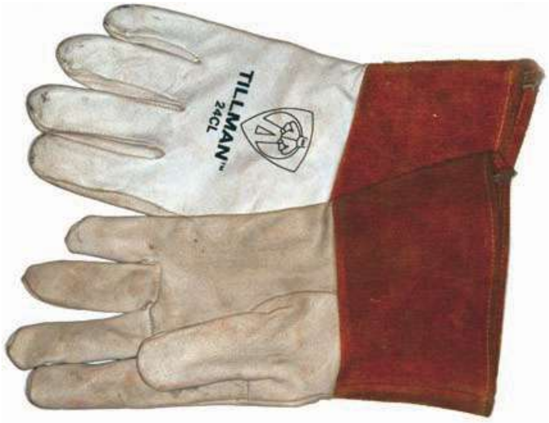

Figure 1.53 **Soft leather gloves**
For welding that requires a great deal of manual dexterity, soft leather gloves can be worn.
Source: Courtesy of Larry Jeffus

Body Protection

Full leather jackets and capes will protect a welder's shoulders, arms, and chest, Figure 1.54. A jacket, unlike the cape, protects a welder's back and complete chest. A cape is open and much cooler but offers less protection. The cape can be used with a bib apron to provide some additional protection while leaving the back cooler. Either the full jacket or the cape with a bib apron should be worn for any out-of-position work.

Waist and Lap Protection

Bib aprons or full aprons will protect a welder's lap. Welders will especially need to protect their laps if they squat or sit while working and when they bend over or lean against a table.

Arm Protection

For some vertical welding, a full or half sleeve can protect a person's arm, Figure 1.55. The sleeves work best if the work level is not above the welder's chest. Work levels higher than this usually require a jacket or cape to keep sparks off the welder's shoulders.

Leg and Foot Protection

When heavy cutting or welding is being done and a large number of sparks are falling, leather pants and spats should be used to protect the welder's legs and feet. If the weather is hot and full leather pants are uncomfortable, leather aprons with leggings are available. Leggings can be strapped to the legs, leaving the back open. Spats will prevent sparks from burning through the front of lace-up boots.

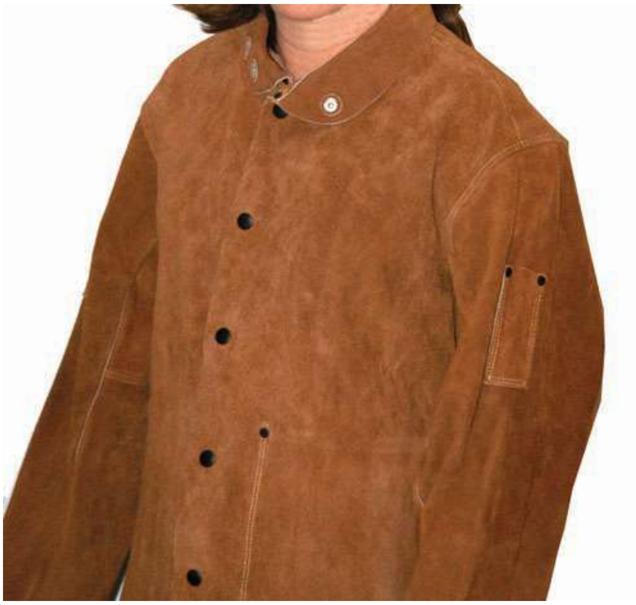

Figure 1.54 Full leather jacket
Source: Courtesy of Larry Jeffus

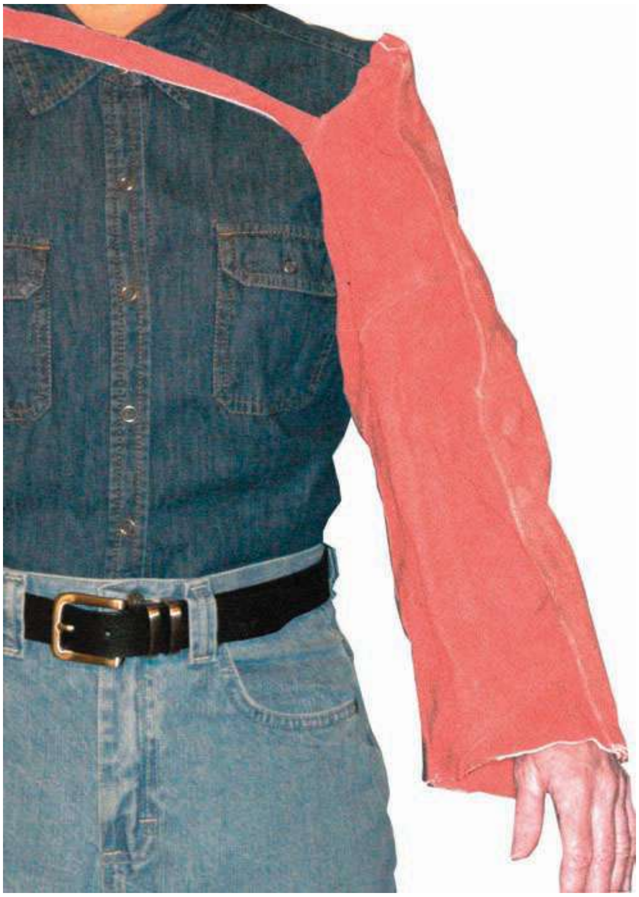

Figure 1.55 **Full leather sleeve**
Source: Courtesy of Larry Jeffus

SUMMARY

The keys to producing quality GMA welds are equipment, setup, and adjustments. Once you have mastered these skills the only remaining obstacle to your producing consistent, uniform, high-quality welds is your ability to follow, or track, the joint consistently. Some welders find that lightly dragging their glove along the metal surface or edge of the fabrication can aid them in controlling the weld consistency. One of the advantages of the GMA welding process is its ability to produce long, uninterrupted welds. However, this often leads to welder fatigue. Finding a comfortable welding position that you can maintain for several minutes at a time will both improve your weld quality and reduce your fatigue.

Selecting the proper method of metal transfer—short arc, globular, or spray transfer—is normally done by the welding shop foreman or supervisor. He or she makes these selections based on the material being welded, the welding position, and other factors, including welding procedure specifications and applicable codes. A welder must be proficient with each of the various methods of metal transfer. It is therefore important that you spend time practicing and developing your skills with each of these processes.

REVIEW

1. Why is usage of the term *GMAW* preferable to *MIG* for gas metal arc welding?
2. Using Table 1.1, answer the following:
   a. What maintains the arc in machine welding?
   b. What feeds the filler metal in manual welding?
   c. What provides the joint travel in automatic welding?
   d. What provides the joint guidance in semiautomatic welding?
3. What factors have led to increased usage of the GMAW process?
4. In what form is metal transferred across the arc in the axial spray metal transfer method of GMA welding?
5. What three conditions are required for the spray transfer process to occur?
6. Using Table 1.3, answer the following:
   a. What should the wire-feed speed and voltage ranges be to weld 1/8-in. metal with 0.035-in. wire using argon shielding gas?
   b. What should the amperage and voltage ranges be when using 98% Ar + 2% O_2 to weld 1/4-in. metal with 0.045-in. wire?
7. What ranges does the pulsed-arc metal transfer shift between?
8. How do frequency, amplitude, and width of the pulses affect the GMA pulse welding process?
9. How have electronics helped the pulsed-arc process?
10. Why is helium added to argon when making some spray or pulsed-spray transfer welds?
11. Why does DCEP help with welds on metals such as aluminum?
12. Why is CO_2 added to argon when making GMA spray transfer welds?
13. Why should CO_2 not be used to weld stainless steel?
14. How is the metal transferred from the electrode to the plate during the GMAW-S process?
15. Using Figure 1.12, what should be the approximate voltage at 175 amps at 200 in./min when using 0.035-in. ER70S-6 electrode wire?
16. Using Table 1.8, what would the amperage be for 0.035-in. (0.9-mm) wire at 200 in./min (5 m/min)?
17. Using Table 1.9, what shielding gas should be used for welding on copper?
18. What may happen if the GMA welding electrode is allowed to strike the base metal outside the molten weld pool?
19. What effect does shortening the electrode extension have on weld penetration?
20. Describe the weld produced by a backhand welding angle.
21. Describe the weld produced by a forehand welding angle.
22. What components make up a GMA welding system?
23. Why must GMA welders have a 100% duty cycle?
24. What can happen if rollers of the wrong shape are used on aluminum wire?
25. Where is the drive motor located in a pull-type wire-feed system?
26. How is the wire-feed speed changed with a linear feed system?
27. What type of liner should be used for aluminum wire?
28. What parts of a typical GMA welding gun can be replaced?
29. Describe the spot welding process using a GMA welder.

CHAPTER

2

Gas Metal Arc Welding

OBJECTIVES

After completing this chapter, the student should be able to

- set up a GMA welding work station
- thread the electrode wire on a GMAW machine
- set the shielding gas flow rate on a GMAW machine
- use various settings on a GMA welding machine, and compare the effects on a weld
- demonstrate the effect of changing the electrode extension on a weld
- describe the effects of changing the welding gun angle on the weld bead
- list six variables to consider when selecting the shielding gas for a particular application
- evaluate weld beads made with various shielding gas mixtures
- explain why hot-rolled steel should be cleaned to bright metal before welding
- make GMA welds in butt joints, lap joints, and tee joints in all positions that will pass a specified standard's visual or destructive examination criteria

KEY TERMS

bird-nesting
cast
conduit liner
contact tube
feed rollers
flow rate
spool drag
wire-feed speed

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter:

Module 1: Occupational Orientation

Key Indicator 1: Prepares time or job cards, reports or records
Key Indicator 2: Performs housekeeping duties
Key Indicator 3: Follows verbal instructions to complete work assignments
Key Indicator 4: Follows written details to complete work assignments

Module 2: Safety and Health of Welders

Key Indicator 1: Demonstrates proper use and inspection of personal protection equipment (PPE)
Key Indicator 2: Demonstrates proper work area operation

Key Indicator 3: Demonstrates proper use and inspection of ventilation equipment
Key Indicator 4: Demonstrates proper hot zone operation
Key Indicator 7: Demonstrates proper inspection and operation of equipment for each welding or thermal cutting process used (This is best done as part of the process module/unit for each of the required welding and thermal cutting processes.)

Module 3: Drawing and Welding Symbol Interpretation

Key Indicator 1: Interprets basic elements of a drawing or sketch
Key Indicator 2: Interprets welding symbol information
Key Indicator 3: Fabricates parts from a drawing or sketch

Module 5: Gas Metal Arc Welding [GMAW-S, GMAW (spray)]

Key Indicator 1: Performs safety inspections of GMAW equipment and accessories
Key Indicator 2: Makes minor external repairs to GMAW equipment and accessories
Short-circuit Transfer
Key Indicator 3: Sets up for GMAW-S operations on carbon steel
Key Indicator 4: Operates GMAW-S equipment on carbon steel
Key Indicator 5: Makes GMAW-S fillet welds, in all positions, on carbon steel
Key Indicator 6: Makes GMAW-S groove welds, in all positions, on carbon steel
Key Indicator 7: Passes GMAW-S welder performance qualification test (workmanship sample) on carbon steel
Spray Transfer
Key Indicator 8: Sets up for GMAW (spray) operations on carbon steel
Key Indicator 9: Operates GMAW (spray) equipment on carbon steel
Key Indicator 10: Makes GMAW (spray) fillet welds, in the 1F and 2F positions, on carbon steel
Key Indicator 11: Makes GMAW (spray) groove welds, in the 1G position, on carbon steel
Key Indicator 12: Passes GMAW (spray) welder performance qualification test (workmanship sample) on carbon steel

Module 9: Welding Inspection and Testing Principles

Key Indicator 1: Examines cut surfaces and edges of prepared base metal parts.
Key Indicator 2: Examines tacks, root passes, intermediate layers, and completed welds

INTRODUCTION

Performing a satisfactory GMA weld requires more than just manipulative skill. The setup, voltage, amperage, electrode extension, and welding angle, as well as other factors, can dramatically affect the weld produced. The very best welding conditions are those that will allow a welder to produce the largest quantity of successful welds in the shortest period of time with the highest productivity. Because these are semiautomatic or automatic processes, increased productivity may require only that the welder increase the travel speed and current. This does not mean that the welder will work harder but, rather, that the welder will work more productively, resulting in a greater cost efficiency.

The more cost efficient welders can be, the more competitive they and their companies become. This can make the difference between being awarded a job or losing work.

SETUP

The same equipment may be used for semiautomatic GMAW, flux cored arc welding (FCAW), and submerged arc welding (SAW). Often, FCAW and SAW equipment have a higher amperage range. In addition, equipment for FCAW and SAW is more likely to be automated than that for GMAW. However, GMA welding equipment can easily be automated.

The basic GMAW installation consists of the following: welding gun, gun switch circuit, electrode conduit or liner, welding contractor control, electrode feed unit, electrode supply, power source, shielding gas supply, shielding gas flowmeter regulator, shielding gas hoses, and both power and work cables. Typical water-cooled and air-cooled guns are shown in Figure 2.1. The equipment setup in this chapter is similar to equipment built by other manufacturers, which means that any skills developed can be transferred easily to other equipment.

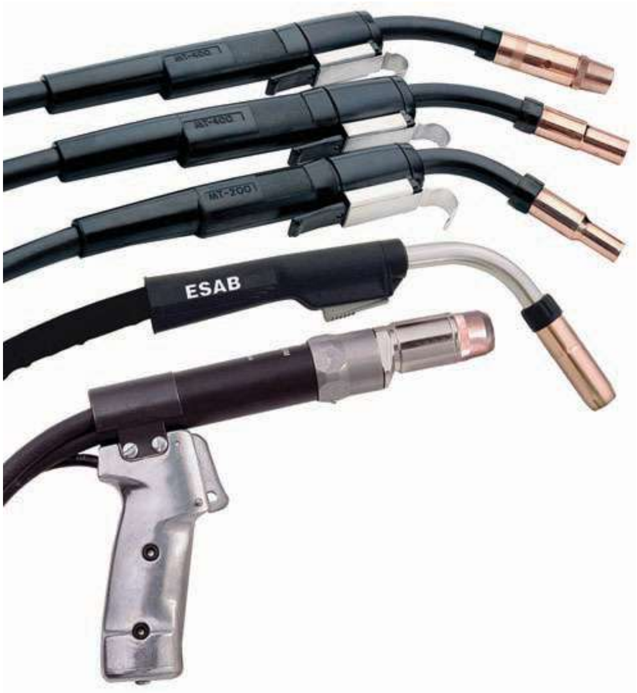

Figure 2.1 **GMA welding guns are available in a variety of sizes and shapes**
Source: Courtesy of ESAB Welding & Cutting Products

PRACTICE 2-1

GMAW Equipment Setup

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Short-circuit Transfer
Key Indicator 3
Spray Transfer
Key Indicator 8

For this practice, you will need a GMAW power source, a welding gun, an electrode feed unit, an electrode supply, a shielding gas supply, a shielding gas flowmeter regulator, electrode conduit, power and work leads, shielding gas hoses, assorted hand tools, spare parts, and any other required materials. In this practice, you will properly set up a GMA welding installation.

If the shielding gas supply is a cylinder, it must be chained securely in place before the valve protection cap is removed, Figure 2.2. Standing to one side of the cylinder and making sure no bystanders are in line with the valve, quickly crack the valve to blow out any dirt in the valve before the flowmeter regulator is attached, Figure 2.3. With the flowmeter regu-

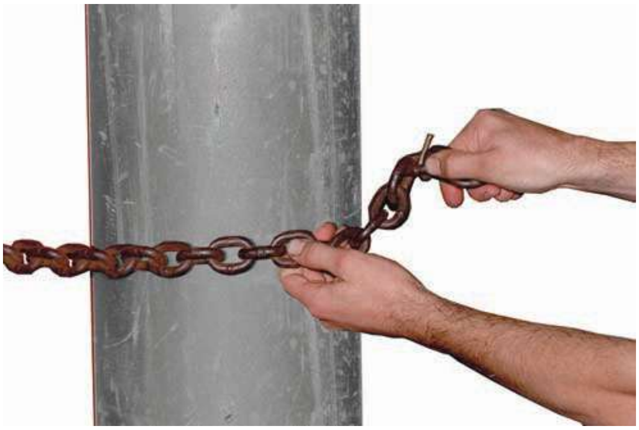

Figure 2.2 **Make sure the gas cylinder is chained securely in place before removing the safety cap**
Source: Courtesy of Larry Jeffus

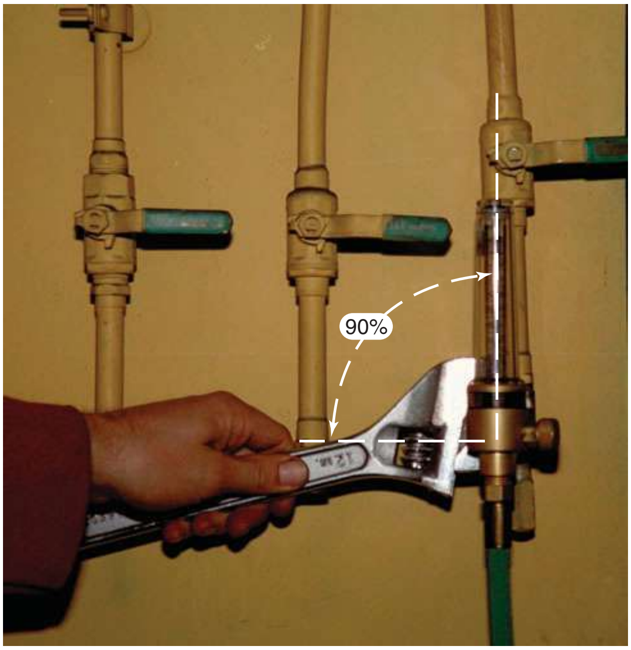

Figure 2.3 **Attaching the flowmeter regulator**
Be sure the tube is vertical.
Source: Courtesy of Larry Jeffus

lator attached securely to the cylinder valve, attach the correct hose from the flowmeter to the "gas-in" connection on the electrode feed unit or machine.

Install the reel of electrode (welding wire) on the holder and secure it, Figure 2.4. Check the feed roller size to ensure that it matches the wire size, Figure 2.5. The conduit liner size should be checked to be sure that it is compatible with the wire size. Connect the conduit to the feed unit. The conduit or an extension should be aligned with the groove in the roller and set as close to the roller as possible without touching, Figure 2.6. Misalignment at this point can contribute to a bird's nest, Figure 2.7. **Bird-nesting** of the electrode wire results when the feed roller pushes the wire into a tangled ball, like a bird's nest, because the wire would not go through the outfeed side conduit.

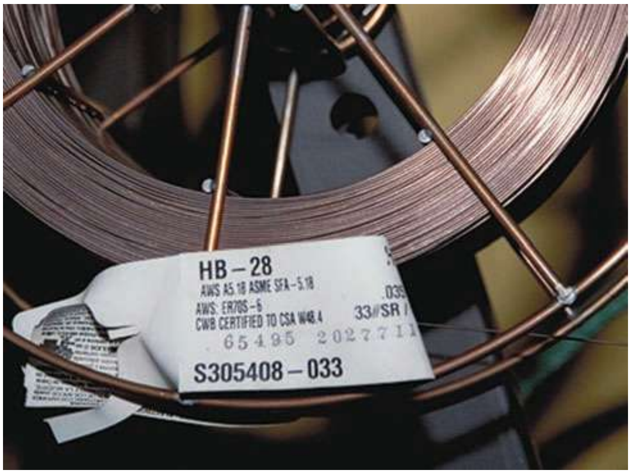
HB-28
AWS A5.18 ASME SFA-5.18
.035
AWS: ER70S-6
33#SR
CWB CERTIFIED TO CSA W48.4
65495 2027711
S305408-033

Figure 2.4 Wire label
When installing the spool of wire, check the label to be sure that the wire is the correct type and size.
Source: Courtesy of Larry Jeffus

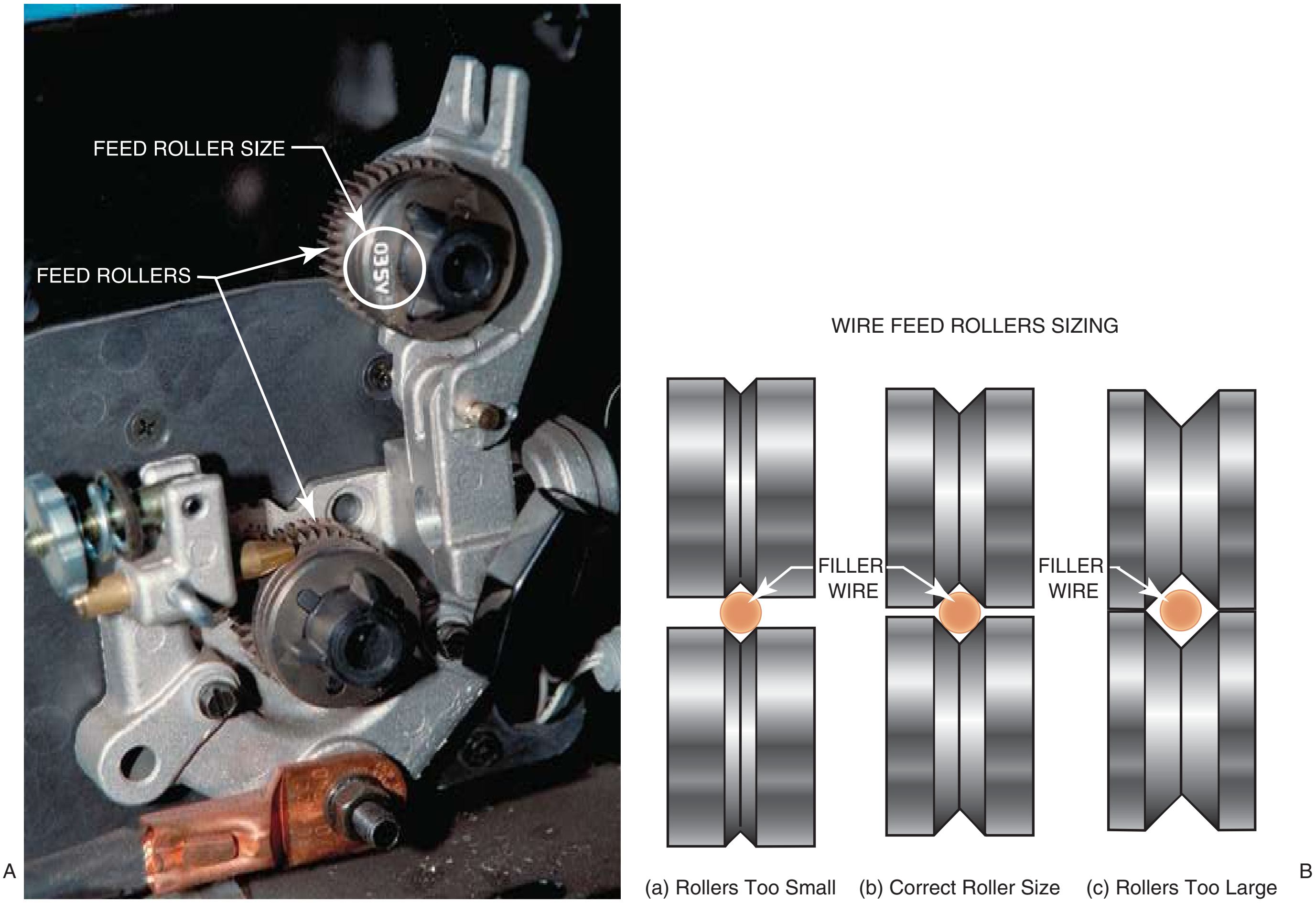
FEED ROLLER SIZE
FEED ROLLERS
.035V
A
WIRE FEED ROLLERS SIZING
FILLER WIRE
FILLER WIRE
(a) Rollers Too Small
(b) Correct Roller Size
(c) Rollers Too Large
B

Figure 2.5 Checking feed roller size
(A) Check to be certain that the feed rollers are the correct size for the wire being used. (B) If the wirefeed rollers are too small, the welding wire could be damaged. If the wirefeed rollers are too large, the rollers will not grip the wire.
Source: (A) Courtesy of Larry Jeffus

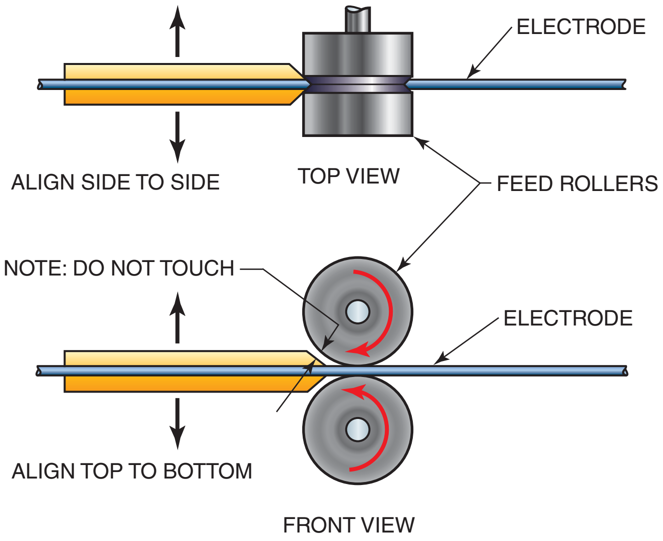
ELECTRODE
ALIGN SIDE TO SIDE
TOP VIEW
FEED ROLLERS
NOTE: DO NOT TOUCH
ELECTRODE
ALIGN TOP TO BOTTOM
FRONT VIEW

Figure 2.6 **Feed**

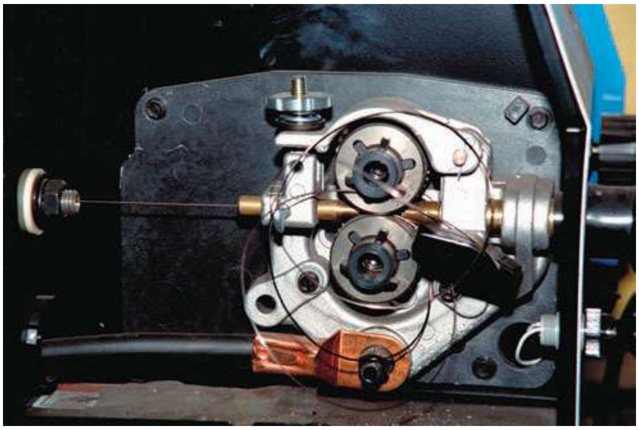

Figure 2.7 **"Bird's nest" in the filler wire at the feed rollers**
Source: Courtesy of Larry Jeffus

Be sure the power is off before attaching the welding cables. The electrode and work leads should be attached to the proper terminals. The electrode lead should be attached to the terminal marked electrode or positive (+). If necessary, it is also attached to the power cable part of the gun lead. The work lead should be attached to the terminal marked work or negative (–).

The shielding "gas-out" side of the solenoid is then also attached to the gun lead. If a separate splice is required from the gun switch circuit to the feed unit, it should be connected at this time, Figure 2.8. Check to see that the welding contractor circuit is connected from the feed unit to the power source.

The welding gun should be securely attached to the main lead cable and conduit, Figure 2.9. There should be a gas diffuser attached to the end of the conduit liner to ensure proper alignment. A **contact tube** (tip) of the correct size to match the electrode wire size being used should be installed, Figure 2.10. A shielding gas nozzle is attached to complete the assembly.

Recheck all fittings and connections for tightness. Loose fittings can leak; loose connections can cause added resistance, reducing the welding

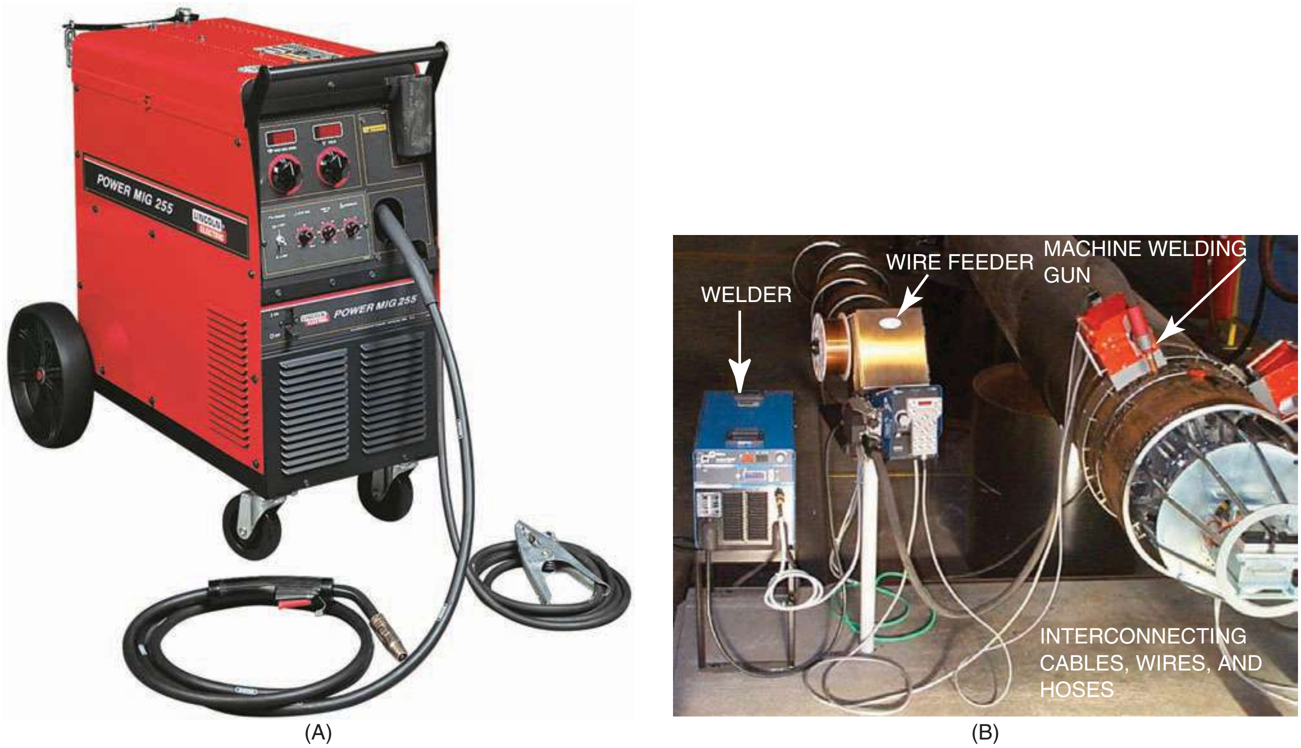
POWER MIG 255
WELDER
WIRE FEEDER
MACHINE WELDING GUN
INTERCONNECTING CABLES, WIRES, AND HOSES
(A)
(B)

Figure 2.8 **GMAW station setup**
(A) Typical GMA welding machine. (B) Typical interconnecting cables and wires for a semiautomatic GMA welding station.
Source: (A) Courtesy of Lincoln Electric Company (B) Courtesy of Dynatorque

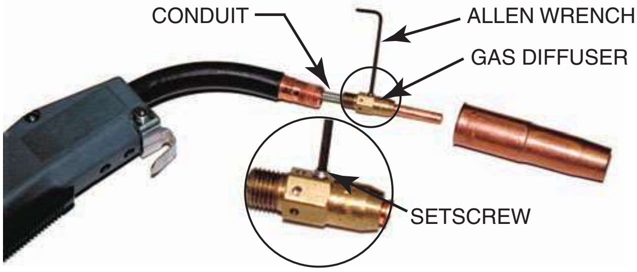
CONDUIT
ALLEN WRENCH
GAS DIFFUSER
SETSCREW

Figure 2.9 **GMA welding gun assembly**
Source: Courtesy of Larry Jeffus

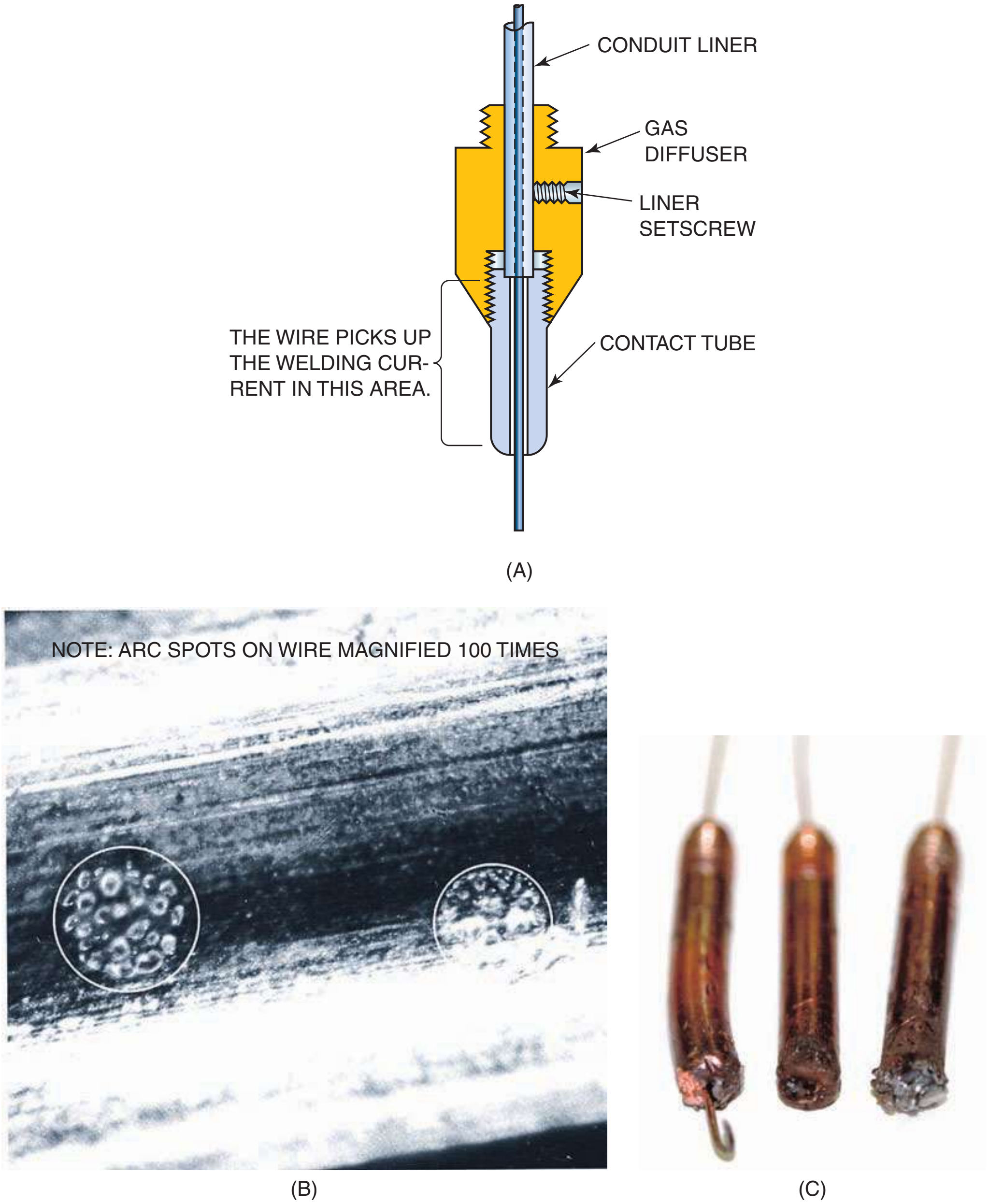
CONDUIT LINER
GAS DIFFUSER
LINER SETSCREW
THE WIRE PICKS UP THE WELDING CURRENT IN THIS AREA.
CONTACT TUBE
(A)
NOTE: ARC SPOTS ON WIRE MAGNIFIED 100 TIMES
(B)
(C)

Figure 2.10 The contact tube must be the correct size
(A) Too small a contact tube will cause the wire to stick. (B) Too large a contact tube can cause arcing to occur between the wire and tube. (C) Heat from the arcing can damage the tube.
Source: (B) Courtesy of Brett V. Hahn (C) Courtesy of Larry Jeffus

efficiency. Some manufacturers include detailed setup instructions with their equipment, Figure 2.11.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Short-circuit Transfer
Key Indicator 3
Spray Transfer
Key Indicator 8

PRACTICE 2-2

Threading GMAW Wire

Using the GMAW machine that was properly assembled in Practice 2-1, you will turn the machine on and thread the electrode wire through the system.

Open the side cover.

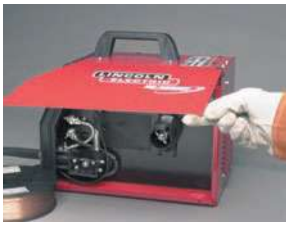

Remove the empty wire spool.

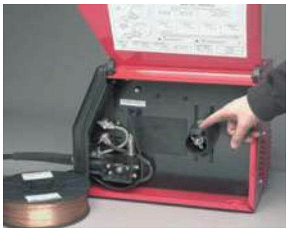

Release upper feed roller.

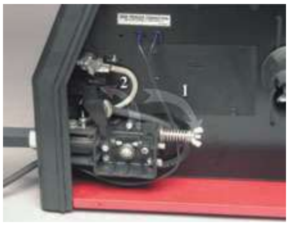

Reload the wire spool with the free end unreeling from the bottom.

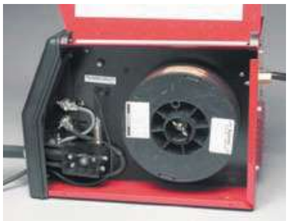

Thread wire through guide between rollers and into wire cable.

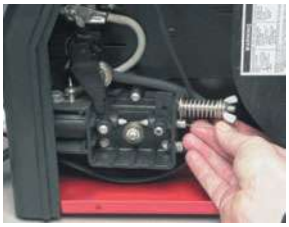

Set the polarity as DCEP from GMA welding.

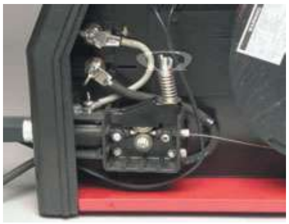

Turn the input switch on.

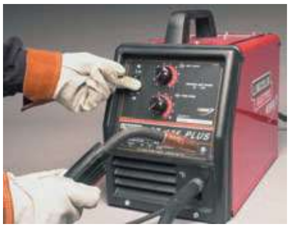

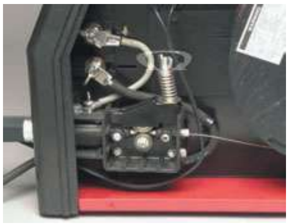

With the gun trigger pressed, adjust the feed roller tension.

Check the setting guide inside the machine door.

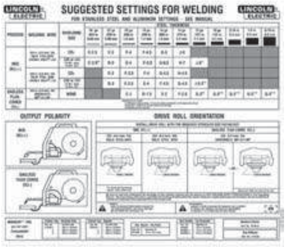

SUGGESTED SETTINGS FOR WELDING

OUTPUT POLARITY

DRIVE ROLL ORIENTATION

Set the voltage and wire feed for the metal you are going to be welding.

Attach work cable clamp to work to be welded.

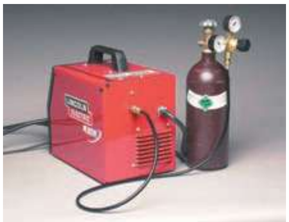

Connect gas to coupling at rear of case and turn on shielding gas.

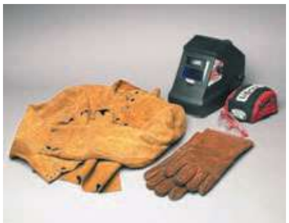

ALWAYS WEAR PROPER SAFETY EQUIPMENT. Pull trigger and weld.

Figure 2.11 **Example of manufacturer's setup instructions**
Source: Courtesy of Lincoln Electric Company

Check to see that the unit is assembled correctly according to the manufacturer's specifications. Switch on the power and check the gun switch circuit by depressing the switch. The power source relays, feed relays, gas solenoid, and feed motor should all activate.

Cut the end of the electrode wire free. Hold it tightly so that it does not unwind. The wire has a natural curve that is known as its **cast**. The cast is measured by the diameter of the circle that the wire would make if it were loosely laid on a flat surface, Figure 2.12. The cast helps the wire make a good electrical contact as it passes through the contact tube. However, the cast can be a problem when threading the system. To make threading easier, straighten about 12 in. (305 mm) of the end of the wire and cut any kinks off.

Separate the wire **feed rollers** and push the wire first through the guides, Figure 2.13, then between the rollers, and finally into the **conduit liner**. Reset the rollers so there is a slight amount of compression on the wire, Figure 2.14. Set the **wire-feed speed** control to a slow speed. Hold the welding gun so that the electrode conduit and cable are as straight as possible.

With safety glasses on and the gun pointed away from the welder's face, press the gun switch, or the cold feed switch if your wire feeder is equipped with one. The cold feed switch feeds wire without delivering current to the gun. The wire should start feeding into the liner. Watch to make certain that the wire feeds smoothly and release the gun switch as soon as the end comes through the contact tube.

With the wire feed running, adjust the feed roller compression so that the wire reel can be stopped easily by a slight pressure. Too light a roller pressure will cause the wire to feed erratically. Too high a pressure can turn a minor problem into a major disaster. If the wire jams at a high roller pressure, the feed rollers keep feeding the wire, causing it to bird-nest and possibly short out. With a light pressure, the wire can stop, preventing bird-nesting. This is very important with soft wires. The other

Caution

If the wire stops feeding before it reaches the end of the contact tube, stop and check the system. If no obvious problem can be found, mark the wire with tape and remove it from the gun. It then can be held next to the system to determine the location of the problem.

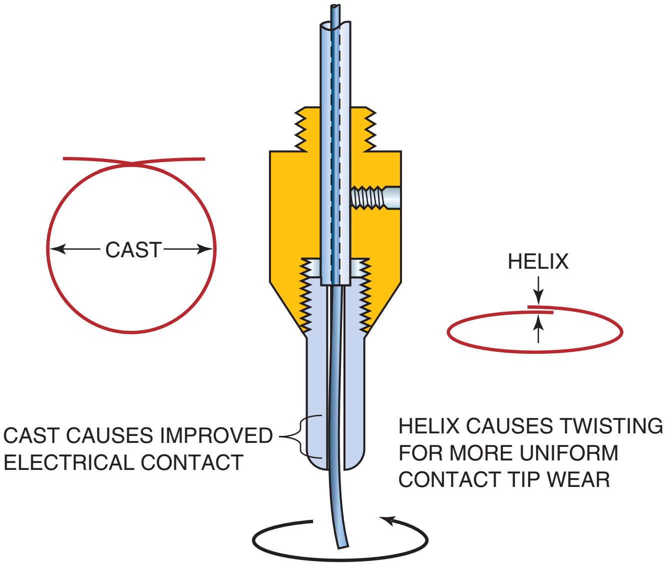
CAST
HELIX
CAST CAUSES IMPROVED ELECTRICAL CONTACT
HELIX CAUSES TWISTING FOR MORE UNIFORM CONTACT TIP WEAR

Figure 2.12 Cast of a welding wire
The cast of the welding wire causes it to rub firmly inside the contact tube for good electrical contact. The helix causes the electrode to twist inside the contact tube so that the tube is worn uniformly.

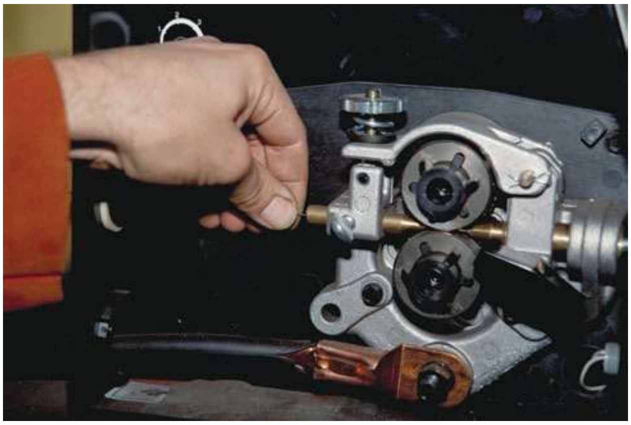

Figure 2.13 **Push the wire through the guides by hand**
Source: Courtesy of Larry Jeffus

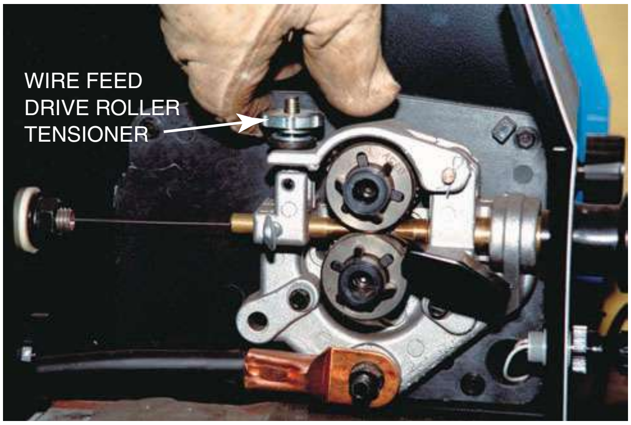

Figure 2.14 **Adjust the wire-feed tensioner**
Source: Courtesy of Larry Jeffus

advantage of a light pressure is that the feed will stop if something like clothing or a gas hose is caught in the reel.

With the feed running, adjust the **spool drag** so that the reel stops when the feed stops. The reel should not coast to a stop because the wire can be snagged easily. Also, when the feed restarts, a jolt occurs when the slack in the wire is taken up. This jolt can be enough to momentarily stop the wire, possibly causing a discontinuity in the weld.

When the test runs are completed, the wire can either be rewound or cut off. Some wire-feed units have a retract button. This allows the feed driver to reverse and retract the wire automatically. To rewind the wire on units without this retraction feature, release the rollers and turn them backward by hand. If the machine will not allow the feed rollers to be released without upsetting the tension, you must cut the wire.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Caution

Do not discard pieces of wire on the floor. They present a hazard to safe movement around the machine. In addition, a small piece of wire can work its way into a filter screen on the welding power source. If the piece of wire shorts out inside the machine, it could become charged with high voltage, which could cause injury or death. Always wind the wire tightly into a ball or cut it into short lengths before discarding it in the proper waste container.

GAS DENSITY AND FLOW RATES

Density is the chief determinant of how effective a gas is for arc shielding. The lower the density of a gas, the higher will be the flow rate required for equal arc protection. Flow rates, however, are not in proportion to the densities. Helium, with about one-tenth the density of argon, requires about twice the flow for equal protection.

The correct flow rate can be set by checking welding guides that are available from the welding equipment and filler metal manufacturers. These welding guides list the gas flow required for various nozzle sizes and welding amperage settings. Some welders feel that a higher gas flow will provide better weld coverage, but that is not always the case. High gas flow rates waste shielding gases and may lead to contamination. The contamination comes from turbulence in the gas at high flow rates. Air is drawn into the gas envelope by the venturi effect around the edge of the nozzle. Also, the air can be drawn in under the nozzle if the torch is held at too sharp an angle to the metal, Figure 2.15.

NOTE

If you need more shielding gas coverage in a windy or drafty area, use both a larger-diameter gas nozzle and a higher gas flow rate. The larger the nozzle size, the higher the permissible flow rate without causing turbulence. Larger nozzle sizes may restrict your view of the weld. You might also consider setting up a wind barrier to protect your welding from the wind, Figure 2.16.

EXPERIMENT 2-1

Setting Gas Flow Rate

Module 1
Key Indicator 1, 2, 4

Module 5
Short-circuit Transfer
Key Indicator 3
Spray Transfer
Key Indicator 8

Using the equipment setup as described in Practice 2-1, and the threaded machine as described in Practice 2-2, you will set the shielding gas flow rate.

The exact **flow rate** required for a certain job will vary depending upon welding conditions. This experiment will help you determine how those conditions affect the flow rate. You will start by setting the shielding gas flow rate at 35 cubic feet per hour (cfh) (16 L/min).

Standing to one side, turn on the shielding gas supply valve. If the supply is a cylinder, the valve is opened all the way. With the machine power on and the welding gun switch depressed, you are ready to set

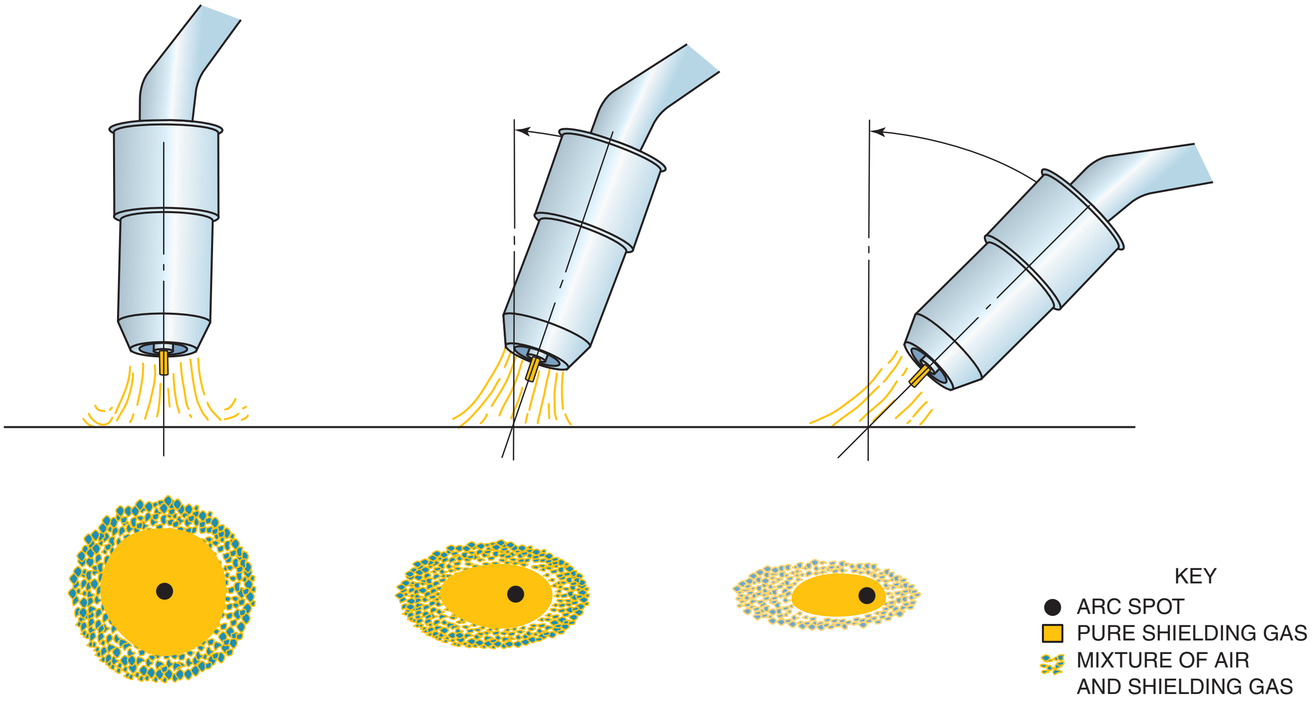

Figure 2.15 The welding gun angle affects the shielding gas coverage for the molten weld pool.

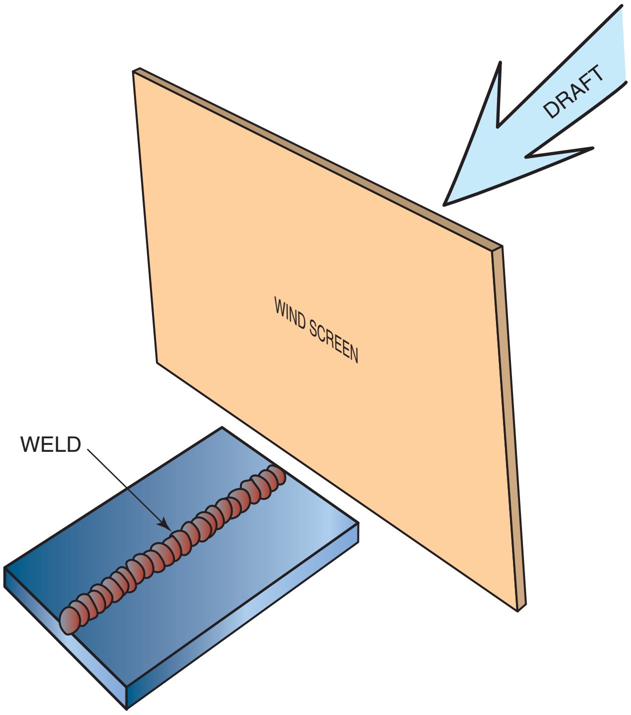

Figure 2.16 **A wind screen can help prevent the shielding gas from being blown away.**

the flow rate. Slowly turn in the adjusting screw and watch the float ball as it rises in a tube on a column of gas. The faster the gas flows, the higher the ball will float. A scale on the tube allows you to read the flow rate. Different scales are used with each type of gas being used. Since various gases have different densities (weights), the ball will float at varying levels even though the flow rates are the same, Figure 2.17. The line corresponding to the flow rate may be read as it compares to the top, center, or bottom of the ball, depending upon the manufacturer's instructions. There should be some marking or instruction on the tube or regulator to tell how it should be read, Figure 2.18.

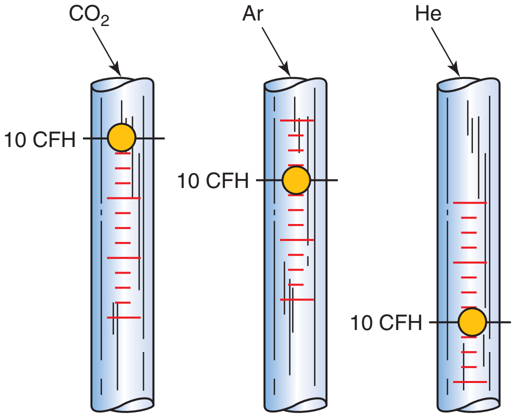

Figure 2.17 **Reading gas flow rates**
Each of these gases is flowing at the same cfh (L/min) rate. Because helium (He) is less dense, its indicator ball is the lowest. Be sure that you are reading the correct scale for the gas being used.

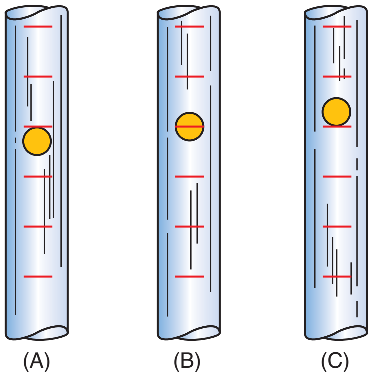

Figure 2.18 **Three methods of reading a flowmeter**
(A) Top of ball, (B) center of ball, and (C) bottom of ball.

Release the welding gun switch, and the gas flow should stop. Turn off the power and spray the hose fittings with a leak-detecting solution.

When you stop for more than a short period, close the shielding gas supply valve and release the hose pressure.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

ARC-VOLTAGE AND AMPERAGE CHARACTERISTICS

The arc-voltage and amperage characteristics of GMA welding are different from those for most other welding processes. The voltage is set on the welder, and the amperage is set by changing the wire-feed speed. At any one voltage setting the amperage required to melt the wire must change as it is fed into the weld. More amperage is required to melt the wire the faster it is fed, and less the slower it is fed.

Because changes in the wire-feed speed directly change the amperage, it is possible to set the amperage by using a chart and measuring the length of wire fed per minute, Table 2.1. The voltage and amperage required for a specific metal transfer method differ for various wire sizes, shielding gases, and metals.

The voltage and amperage setting will be specified for all welding done according to a welding procedure specification (WPS) or other codes and standards. However, most welding—like that done in small production shops, as maintenance welding, for repair work, in farm shops, and the like—is not done to a specific code or standard and therefore no specific setting exists. For that reason, it is important to learn to make the adjustments necessary to allow you to produce quality welds.

EXPERIMENT 2-2

Setting the Current

Module 1
Key Indicator 1, 2, 4

Module 5
Short-circuit Transfer
Key Indicator 4
Spray Transfer
Key Indicator 9

Using a properly assembled GMA welding machine, proper safety protection, and one piece of mild steel plate approximately 12 in. (305 mm) long × 1/4 in. (6 mm) thick, you will change the current settings and observe the effect on GMAW. On a scale of 0 to 10, set the wire-feed speed control dial at 5, or halfway between the low and high settings of

Table 2.1 Typical Amperages for Carbon Steel

	Wire Diameter			
Wire-feed Speed* (in./min)	.030 in. (0.8 mm)	.035 in. (0.9 mm)	.045 in. (1.2 mm)	.062 in. (1.6 mm)
100 (2.5)	40	65	120	190
200 (5.0)	80	120	200	330
300 (7.6)	130	170	260	425
400 (10.2)	160	210	320	490
500 (12.7)	180	245	365	–
600 (15.2)	200	265	400	–
700 (17.8)	215	280	430	–

*To check feed speed, run out wire for one minute and then measure its length.

the unit. The voltage is also set at a point halfway between the low and high settings. The shielding gas can be CO_2, argon, or a mixture. The gas flow should be adjusted to a rate of 35 cfh (16 L/min).

Hold the welding gun at a comfortable angle, lower your welding hood, and pull the trigger. As the wire feeds and contacts the plate, the weld will begin. Move the gun slowly along the plate. Note the following welding conditions as the weld progresses: voltage, amperage, weld direction, metal transfer, spatter, molten weld pool size, and penetration. Stop and record your observations in Table 2.2. Evaluate the quality of the weld as acceptable or unacceptable.

Reduce the voltage somewhat and make another weld, keeping all other weld variables (travel speed, stickout, direction, amperage) the same. Observe the weld and upon stopping record the results. Repeat this procedure until the voltage has been lowered to the minimum value indicated on the machine. Near the lower end the wire may stick, jump, or simply no longer weld.

Return the voltage indicator to the original starting position and make a short test weld. Stop and compare the results to those first observed. Then slightly increase the voltage setting and make another weld. Repeat the procedure of observing and recording the results as the voltage is increased in steps until the maximum machine capability is obtained. Near the maximum setting the spatter may become excessive if CO_2 shielding gas is used. Care must be taken to prevent the wire from fusing to the contact tube.

Return the voltage indicator to the original starting position and make a short test weld. Compare the results observed with those previously obtained.

Lower the wire-feed speed setting slightly and use the same procedure as before. First lower and then raise the voltage through a complete range and record your observations. After a complete set of test results are obtained from this amperage setting, again lower the wire-feed speed for a new series of tests. Repeat this procedure until the amperage is at the minimum setting shown on the machine. At low amperages and high voltage settings, the wire may tend to pop violently as a result of the uncontrolled arc.

Return the wire-feed speed and voltages to the original settings. Make a test weld and compare the results with the original tests. Slightly raise the wire speed and again run a set of tests as the voltage is changed in small steps. After each series, return the voltage setting to the starting point and increase the wire-feed speed. Make a new set of tests.

All of the test data can be gathered into an operational graph for the machine, wire type, size, and shielding gas. Set up a graph like that in Figure 2.19 to plot the data. The acceptable welds should be marked on

Table 2.2 **Setting the Current**

Weld Acceptability	Voltage	Amperage	Spatter	Molten Pool Size	Penetration
Good	20	75	Light	Small	Little

Electrode diameter	0.035 in. (0.9 mm)
Shielding gas CO_2	
Welding direction	Backhand

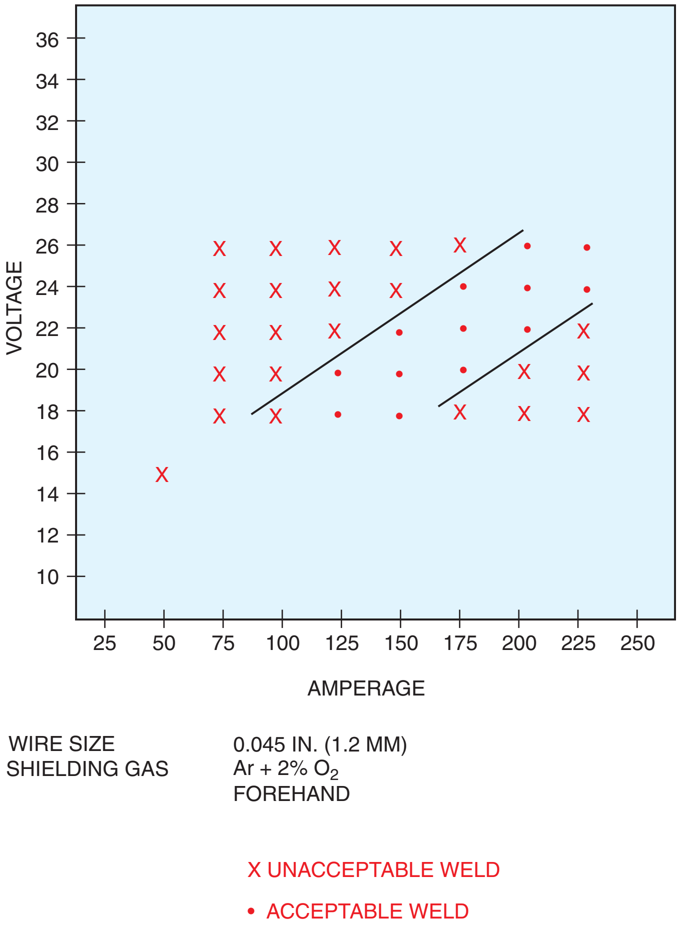

Figure 2.19 Graph for GMAW machine settings

the lines that extend from the appropriate voltages and amperages. Upon completion, the graph will give you the optimum settings for the operation of this particular GMAW setup. The optimum settings are along a line in the center of the acceptable welds.

Experienced welders will follow a much shorter version of this type of procedure anytime they are starting to work on a new machine or testing for a new job. This experiment can be repeated using different types of wire, wire sizes, shielding gases, and weld directions. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

ELECTRODE EXTENSION

Because of the constant-potential (CP) power supply, the welding current will change as the distance between the contact tube and the work changes. Although this change is slight, it is enough to affect the weld being produced. The longer the electrode extension, the greater the resistance will be to the welding current flowing through the small welding wire. This results in some of the welding current being changed to heat at the tip of the electrode, Figure 2.20. With a standard constant-current (CC) power supply for SMA welding, the heat buildup would also reduce the arc voltage, but with a CP power supply the voltage remains constant and the amperage increases. If the electrode extension is shortened, the welding current decreases.

The increase in current does not result in an increase in penetration, because the current is being used to heat the electrode tip and not being transferred to the weld metal. Penetration is reduced and buildup is increased as the electrode extension is lengthened. Penetration is increased and buildup decreased as the electrode extension is shortened. Controlling the weld penetration and buildup by changing the electrode will help maintain weld bead shape during welding. It will also help you better understand what may be happening if a weld starts out correctly but begins to change as it progresses along the joint. You may be changing the electrode extension without noticing the change. Short electrode stickout gives a hotter weld, and long stickout results in a cooler weld.

EXPERIMENT 2-3

Electrode Extension

Module 1
Key Indicator 1, 2, 4

Module 5
Short-circuit Transfer
Key Indicator 4
Spray Transfer
Key Indicator 9

Using a properly assembled GMA welding machine, proper safety protection, and a few pieces of mild steel, each about 12 in. (305 mm) long and ranging in thickness from 16 gauge to 1/2 in. (13 mm), you will observe the effect of changing electrode extension on the weld.

Start at a low current setting. Using the graph developed in Experiment 2.2, set both the voltage and amperage. The settings should be equal to those on the optimum line established for the wire type and size being used with the same shielding gas.

Holding the welding gun at a comfortable angle and height, lower your helmet and start to weld. Make a weld approximately 2 in. (51 mm) long. Then reduce the distance from the gun to the work while continuing to weld. After a few inches, again shorten the electrode extension even

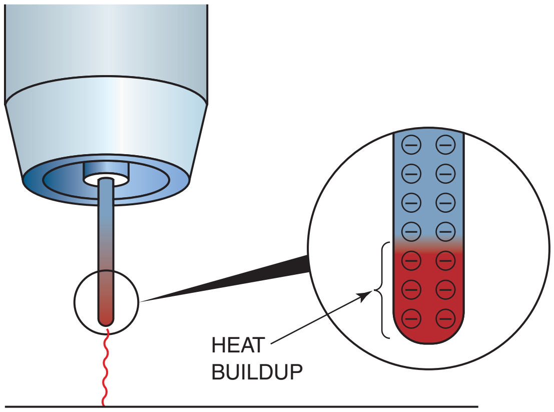

Figure 2.20 **Heat buildup due to the extremely high current for the small conductor (electrode)**

more. Keep doing this in steps until the nozzle is as close as possible to the work. Stop and return the gun to the original starting distance.

Repeat the process just described but now increase the electrode extension and make welds of a few inches at each extension. Keep increasing the electrode extension until the weld will no longer fuse or the wire becomes impossible to control.

Change the plate thickness and repeat the procedure. When the series has been completed with each plate thickness, raise the voltage and amperage to a medium setting and repeat the process. Upon completing this series of tests, adjust the voltage and amperage upward to a high setting. Make a full series of tests using the same procedures as before.

Record the results in Table 2.3 after each series of tests. The final results can be plotted on a graph, as was done in Figure 2.21, to establish the optimum electrode extension for each thickness, voltage, and amperage. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Table 2.3 **Electrode Extension**

Weld Acceptability	Voltage	Amperage	Electrode Extension	Contact Tube to Work Distance	Bead Shape
Poor	20	100	1 in. (25 mm)	1-1/4 in. (31 mm)	Narrow, high, with little penetration

Electrode diameter: 0.035 in. (0.9 mm)
Shielding gas: CO_2
Welding direction: Forehand

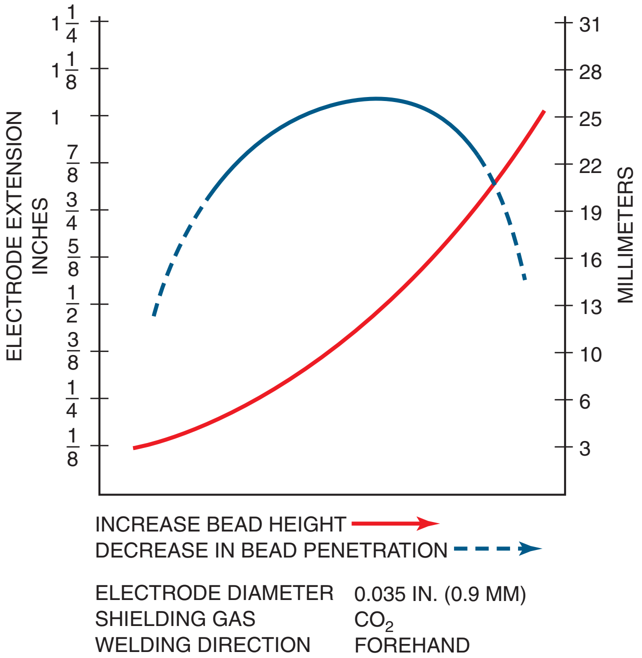

Figure 2.21 **Plot of Experiment 2-3 results**

WELDING GUN ANGLE

The term *welding gun angle* refers to the angle between the GMA welding gun and the work as it relates to the direction of travel. Backhand welding, or dragging angle, Figure 2.22, produces a weld with deep penetration and higher buildup. Forehand welding, or pushing angle, Figure 2.23, produces a weld with shallow penetration and little buildup.

Slight changes in the welding gun angle can be used to control the weld as the groove spacing changes. A narrow gap may require more penetration, but as the gap spacing increases a weld with less penetration may be required. Changing the electrode extension and welding gun angle at the same time can result in a quality weld being made under less than ideal conditions.

EXPERIMENT 2-4

Welding Gun Angle

Module 1
Key Indicator 1, 2, 4

Module 5
Short-circuit Transfer
Key Indicator 4
Spray Transfer
Key Indicator 9

Using a properly assembled GMA welding machine, proper safety protection, and some pieces of mild steel, each approximately 12 in. (305 mm) long and ranging in thickness from 16 gauge to 1/2 in. (13 mm), you will observe the effect of changing the welding gun angle on the weld bead.

Starting with a medium current setting and a plate that is 1/4 in. (6 mm) thick, hold the welding gun at a 30° angle to the plate in the direction of the weld, Figure 2.24. Lower your welding hood and depress the trigger. When the weld starts, move in a straight line and slowly pivot the gun angle as the weld progresses. Keep the travel speed, electrode

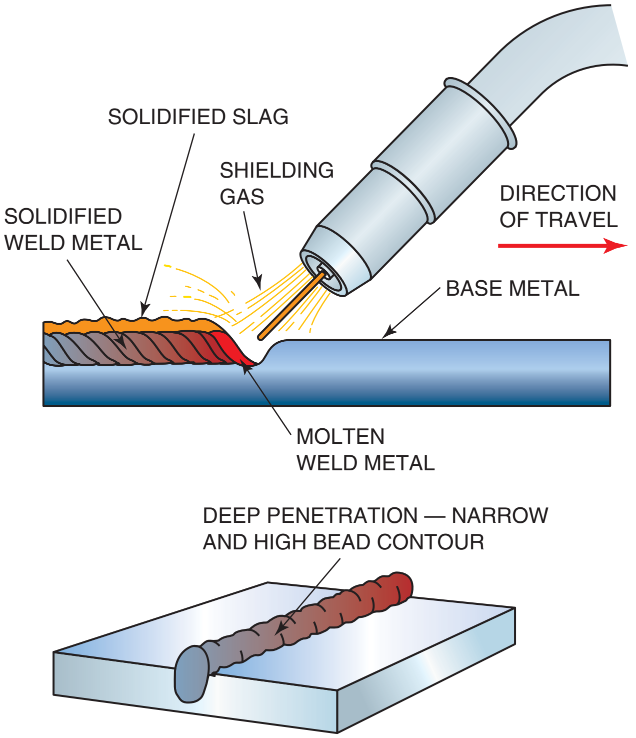
SOLIDIFIED SLAG
SHIELDING GAS
SOLIDIFIED WELD METAL
DIRECTION OF TRAVEL
BASE METAL
MOLTEN WELD METAL
DEEP PENETRATION — NARROW AND HIGH BEAD CONTOUR

Figure 2.22 **Backhand welding, or dragging angle**

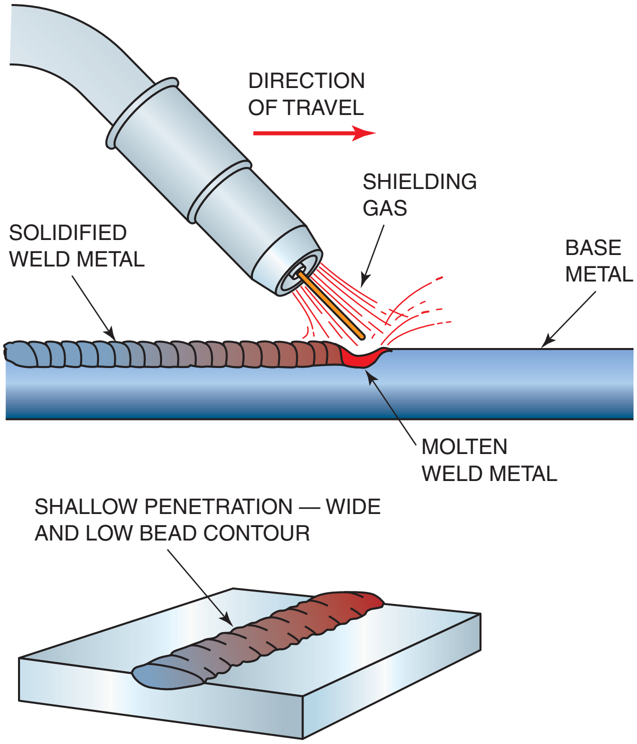
DIRECTION OF TRAVEL
SHIELDING GAS
SOLIDIFIED WELD METAL
BASE METAL
MOLTEN WELD METAL
SHALLOW PENETRATION — WIDE AND LOW BEAD CONTOUR

Figure 2.23 **Forehand welding, or pushing angle**

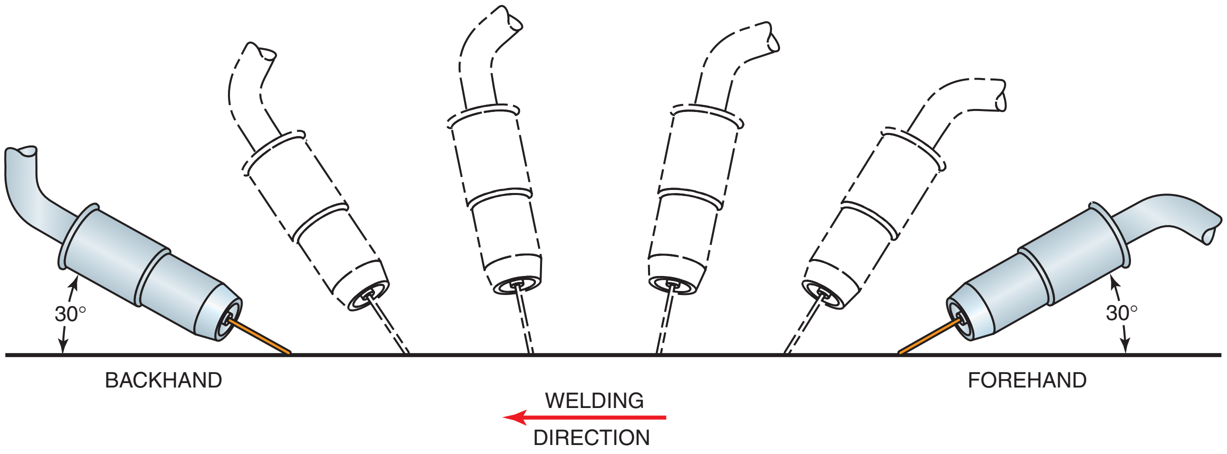

Figure 2.24 Welding gun angle

extension, and weave pattern (the side-to-side motion, if used) constant so that any change in the weld bead is caused by the angle change.

The pivot should be completed in the 12 in. (305 mm) of the weld. You will proceed from a 30° pushing angle to a 30° dragging angle. Repeat this procedure using different welding currents and plate thicknesses.

After the welds are complete, note the differences in width and reinforcement along the welds. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

EFFECT OF SHIELDING GAS ON WELDING

Shielding gases in the gas metal arc process are used primarily to protect the molten metal from oxidation and contamination. Other factors must be considered, however, in selecting the right gas for a particular application. Shielding gas can influence arc and metal transfer characteristics, weld penetration, width of fusion zone, surface shape patterns, welding speed, and undercut tendency. Inert gases such as argon and helium provide the necessary shielding because they do not form compounds with any other substance and are insoluble in molten metal. When used as pure gases for welding ferrous metals, argon and helium may produce an erratic arc action, promote undercutting, and result in other flaws.

It is therefore usually necessary to add controlled quantities of reactive gases to achieve good arc action and metal transfer with these materials. Adding oxygen or carbon dioxide to the inert gas tends to stabilize the arc, promote favorable metal transfer, and minimize spatter. As a result, the penetration pattern is improved and undercutting is reduced or eliminated.

Oxygen or carbon dioxide is often added to argon. The amount of reactive gas required to produce the desired effects is quite small. As little as 0.5% of oxygen will produce a noticeable change; 1% to 5% of oxygen is more common. Carbon dioxide may be added to argon in the 10% to 30% range. Mixtures of argon with less than 10% carbon dioxide may not have enough arc voltage to give the desired results.

Adding oxygen or carbon dioxide to an inert gas causes the shielding gas to become oxidizing. This in turn may cause porosity in some ferrous metals. In this case, a filler wire containing suitable deoxidizers should be used. The presence of oxygen in the shielding gas can also cause some loss of certain alloying elements, such as chromium, vanadium, aluminum, titanium, manganese, and silicon. Again, the addition of a deoxidizer to the filler wire is necessary.

Pure carbon dioxide has become widely used as a shielding gas for GMA welding of steels. It allows higher welding speed, better penetration in the short-circuiting transfer mode, and good mechanical properties, and it costs less than the inert gases. The chief drawback in the use of carbon dioxide is the less-steady-arc characteristics and considerable weld-metal-spatter losses. The spatter can be kept to a minimum by maintaining a very short, uniform arc length. Consistently sound welds can be produced using carbon dioxide shielding, provided that a filler wire having the proper deoxidizing additives is used.

EXPERIMENT 2-5

Effect of Shielding Gas Changes

Module 1
Key Indicator 1, 2, 4

Module 5
Short-circuit Transfer
Key Indicator 4
Spray Transfer
Key Indicator 9

Using a properly assembled GMA welding machine; proper safety protection; a source of CO_2, argon, and oxygen gases or a variety of premixed shielding gases; two flowmeters (or one two-gas mixing regulator); and some pieces of mild steel plate, each about 12 in. (305 mm) long and ranging in thickness from 16 gauge to 1/2 in. (13 mm), you will observe the effect of various shielding gas mixtures on the weld.

Using a mixing flowmeter regulator will allow the gases to be mixed in any desired mixture. A mixing ratio chart to be used to arrive at the approximate gas percentages appears in Figure 2.25. The exact ratios are not so important to you, as a student, as they are on code work.

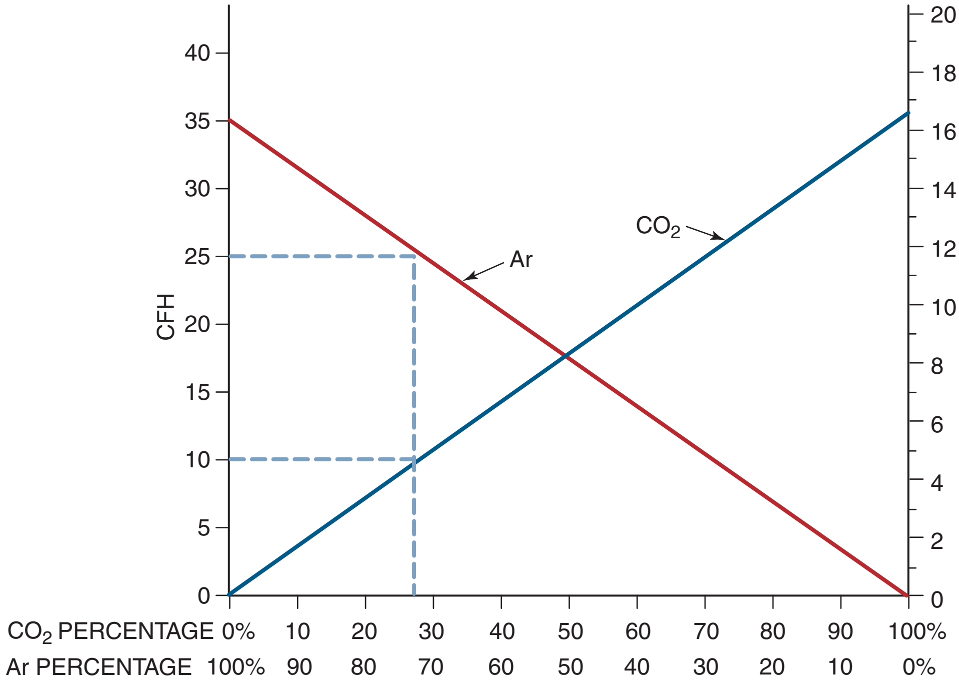

EXAMPLE:
10 CFH CO_2 = 28%
25 CFH Ar = 72%
35 CFH TOTAL = 100%

4.75 L/MIN CO_2 = 28%
11.75 L/MIN Ar = 72%
16.5 L/MIN TOTAL = 100%

Figure 2.25 Gas mixing percentages

With a medium voltage and amperage setting and using a 100% carbon dioxide (CO_2) shielding gas, start making a weld. Either change the mixture after each weld or have another person change the shielding gas during the weld. Keep the total flow rate the same by adding argon (Ar) while reducing the CO_2 to preserve the same flow rate. During the experiment, change over the shielding gas to 100% argon (Ar).

After the weld is complete, evaluate it for spatter, penetration, undercut, buildup, width, or other noticeable changes along its length. Using Table 2.4, record the results of your evaluation.

Repeat the procedure just explained two times more with both low and high power settings. Again, record your observations.

Starting with 100% argon (Ar), add oxygen (O_2) to the shielding gas. The oxygen percentage will range from 0% to 10%, Figure 2.26. Very slight changes in the percentage will have dramatic effects on the weld. You will make three welds using low, medium, and high power settings. For each weld, you will record your observations.

Table 2.4 Shielding Gas Mixtures

Weld Acceptability	Voltage	Spatter	Penetration	Puddle Size	Bead Appearance
Good	75 Ar 25 CO_2	Very little	Deep	Large	Wide with little buildup

Electrode diameter: 0.035 in. (0.9 mm)
Welding direction: Backhand
Voltage: 25
Amperage: 150

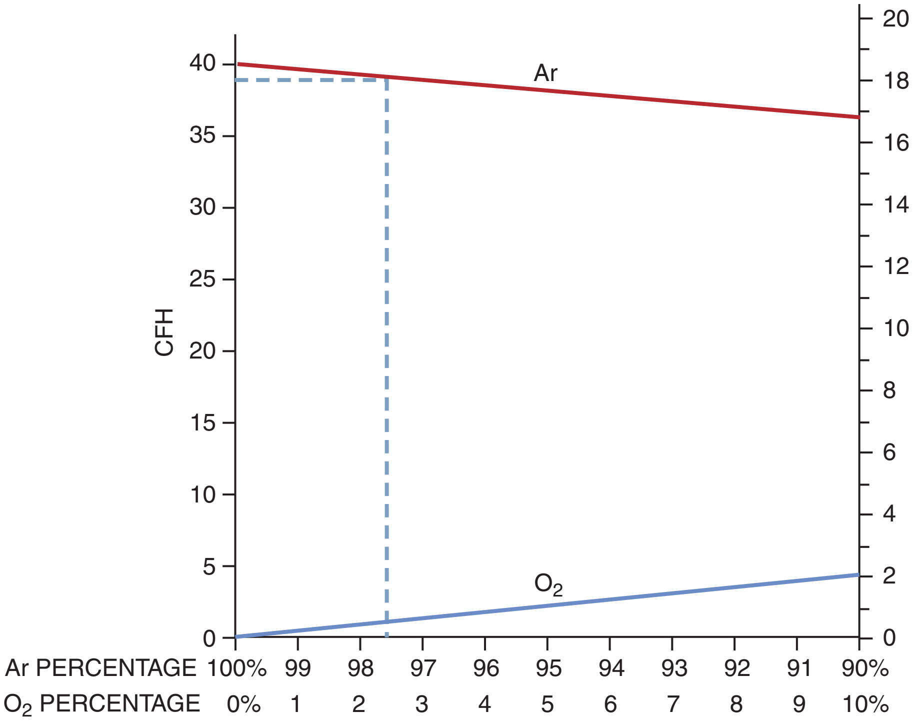

EXAMPLE:

39 CFH	Ar	= 97.5%
1 CFH	O_2	= 2.5%
40 CFH	TOTAL	= 100 %
18.4 L/MIN	Ar	= 97.5%
0.5 L/MIN	O_2	= 2.5%
18.9 L/MIN	TOTAL	= 100 %

Figure 2.26 Ar and O_2 Mixture Percentages

During some of the welding tests, you will notice a change in the method of metal transfer, weld heat, and general weld performance without a change in the current settings. The shielding gas mixture can have major effects on the rate of metal transfer and the welding speed, as well as other welding variables. Higher speeds and greater production can be obtained by using some gas mixtures. However, the savings can be completely offset by the higher gas cost. Before making a final decision about the gas to be used, all the variables must be compared. Table 2.5 lists premixed shielding gases and their uses.

Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICES

The practices in this chapter are grouped according to those requiring similar techniques and setups. To make acceptable GMA welds consistently, the major skill required is the ability to set up the equipment and weldment. Changes such as variations in material thickness, position, and type of joint require changes in both technique and setup. A correctly set up GMA welding station can, in many cases, be operated with minimum skill. Often the only difference between a welder earning a minimum wage and one earning the maximum wage is the ability to make correct machine setups.

Ideally, only a few tests would be needed for the welder to make the necessary adjustments in setup and manipulation techniques to achieve

Table 2.5 Shielding Gases and Gas Mixtures Used for Gas Metal Arc Welding

Shielding Gas	Chemical Behavior	Uses and Usage Notes
1. Argon	Inert	Welding virtually all metals except steel
2. Helium	Inert	Al and Cu alloys for greater heat and to minimize porosity
3. Ar and He (20% to 80% to 50% to 50%)	Inert	Al and Cu alloys for greater heat and to minimize porosity but with quieter, more readily controlled arc action
4. N_2	Reducing	On Cu, very powerful arc
5. Ar + 25% to 30% N_2	Reducing	On Cu, powerful but smoother operating, more readily controlled arc than with N_2
6. Ar + 1% to 2% O_2	Oxidizing	Stainless and alloy steels, also for some deoxidized copper alloys
7. Ar + 3% to 5% O_2	Oxidizing	Plain carbon, alloy, and stainless steels (generally requires highly deoxidized wire)
8. Ar + 3% to 5% O_2	Oxidizing	Various steels using deoxidized wire
9. Ar + 20% to 30% O_2	Oxidizing	Various steels, chiefly with short-circuiting arc
10. Ar + 5% O_2 + 15% CO_2	Oxidizing	Various steels using deoxidized wire
11. CO_2	Oxidizing	Plain-carbon and low-alloy steels, deoxidized wire essential
12. CO_2 + 3% to 10% O_2	Oxidizing	Various steels using deoxidized wire
13. CO_2 + 20% O_2	Oxidizing	Steels

a good weld. The previous welding experiments should have given the welder a graphic set of comparisons to help that welder make the correct changes. In addition to keeping the test data, you may want to keep the test plates for a more accurate comparison.

The grouping of practices in this chapter will keep the number of variables in the setup to a minimum. Often, the only change required before going on to the next weld is to adjust the power settings.

Figures that are given in some of the practices will give the welder general operating conditions, such as voltage, amperage, and shielding gas or gas mixture. These are general values, so the welder will have to make some fine adjustments. Differences in the type of machine being used and the material surface condition will affect the settings. For this reason, it is preferable to use the settings developed during the experiments.

METAL PREPARATION

All hot-rolled steel has an oxide layer, which is formed during the rolling process, called mill scale. *Mill scale* is a thin layer of dark gray or black iron oxide. Some hot-rolled steels that have had this layer removed either mechanically or chemically can be purchased. However, almost all of the hot-rolled steel used today still has this layer because it offers some protection from rusting.

Mill scale is not removed for noncode welding, because it does not prevent most welds from being suitable for service. For practice welds that will be visually inspected, mill scale can usually be left on the plate. Filler metals and fluxes usually have deoxidizers added to them so that the adverse effects of the mill scale are reduced or eliminated, Table 2.6. But with GMA welding wire it is difficult to add enough deoxidizers to remove all effects of mill scale. The porosity that mill scale causes is most often confined to the interior of the weld and is not visible on the surface, Figure 2.27. Because it is not visible on the surface, it usually goes unnoticed and the weld passes visual inspection.

If the practice results are going to be destructively tested or if the work is of a critical nature, then all welding surfaces within the weld groove and the surrounding surfaces within 1 in. (25 mm) must be cleaned to bright metal, Figure 2.28. Cleaning may be either grinding, filing, sanding, or blasting.

Table 2.6 **Deoxidizing Elements in Filler Wire**

Deoxidizing Element	Strength
Aluminum (Al)	Very strong
Manganese (Mn)	Weak
Silicon (Si)	Weak
Titanium (Ti)	Very strong
Zirconium (Zr)	Very strong

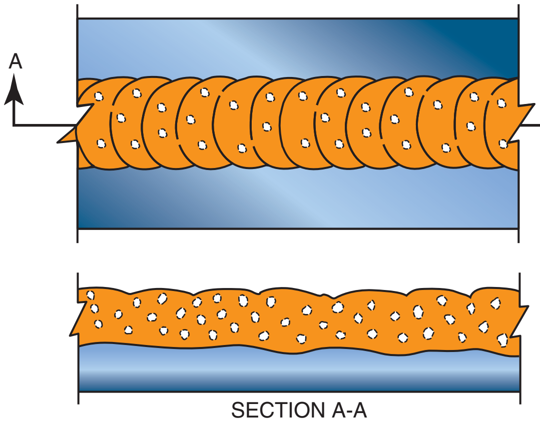

Figure 2.27 Uniformly scattered porosities

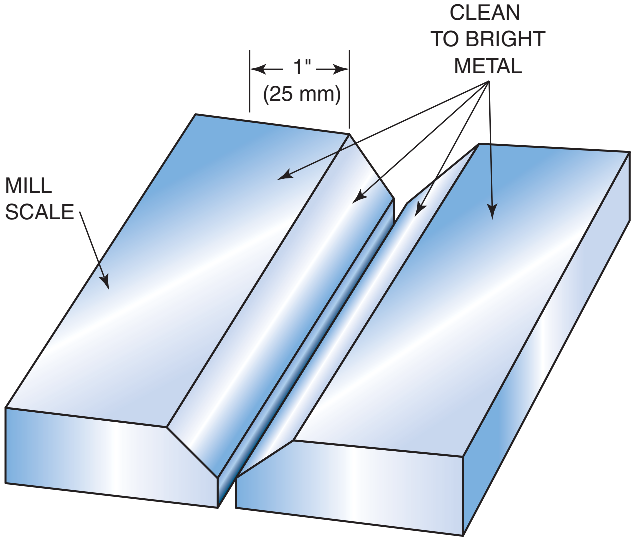

Figure 2.28 Clean all surfaces to bright metal before welding

FLAT POSITION, 1G AND 1F POSITIONS

PRACTICE 2-3

Stringer Beads Using the Short-circuiting Metal Transfer Method in the Flat Position

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Short-circuit Transfer
Key Indicator 4

Using a properly set up and adjusted GMA welding machine, Table 2.7, proper safety protection, 0.035-in. and/or 0.045-in. (0.9-mm and/or 1.2-mm) diameter wire, and two or more pieces of mild steel sheet 12 in. (305 mm) long and 16 gauge and 1/8 in. (3 mm) thick, you will make a stringer bead weld in the flat position, Figure 2.29.

Starting at one end of the plate and using either a pushing or dragging technique, make a weld bead along the entire 12-in. (305-mm) length of the metal. After the weld is complete, check its appearance. Make any needed changes in voltage, wire feed speed, or electrode extension to correct the weld (refer to Figure 2.19 and Figure 2.21). Repeat the weld and make additional adjustments. After the machine is set, start to work on improving the straightness and uniformity of the weld.

Keeping the bead straight and uniform can be hard because of the limited visibility due to the small amount of light and the size of the molten weld pool. The welder's view is further restricted by the shielding gas nozzle, Figure 2.30. Even with limited visibility, it is possible to make a

Table 2.7 Typical Welding Current Settings for Short-circuiting Metal Transfer for Mild Steel

Process	Wire Diameter	Amperage Range (Optimum)	Voltage Range (Optimum)	Shielding Gas
Short-circuiting	0.030	60 (100) 140	14 (15) 16	100% CO_2 75% Ar + 25% CO_2 98% Ar + 2% O
	0.035	90 (130) 150	16 (17) 20	

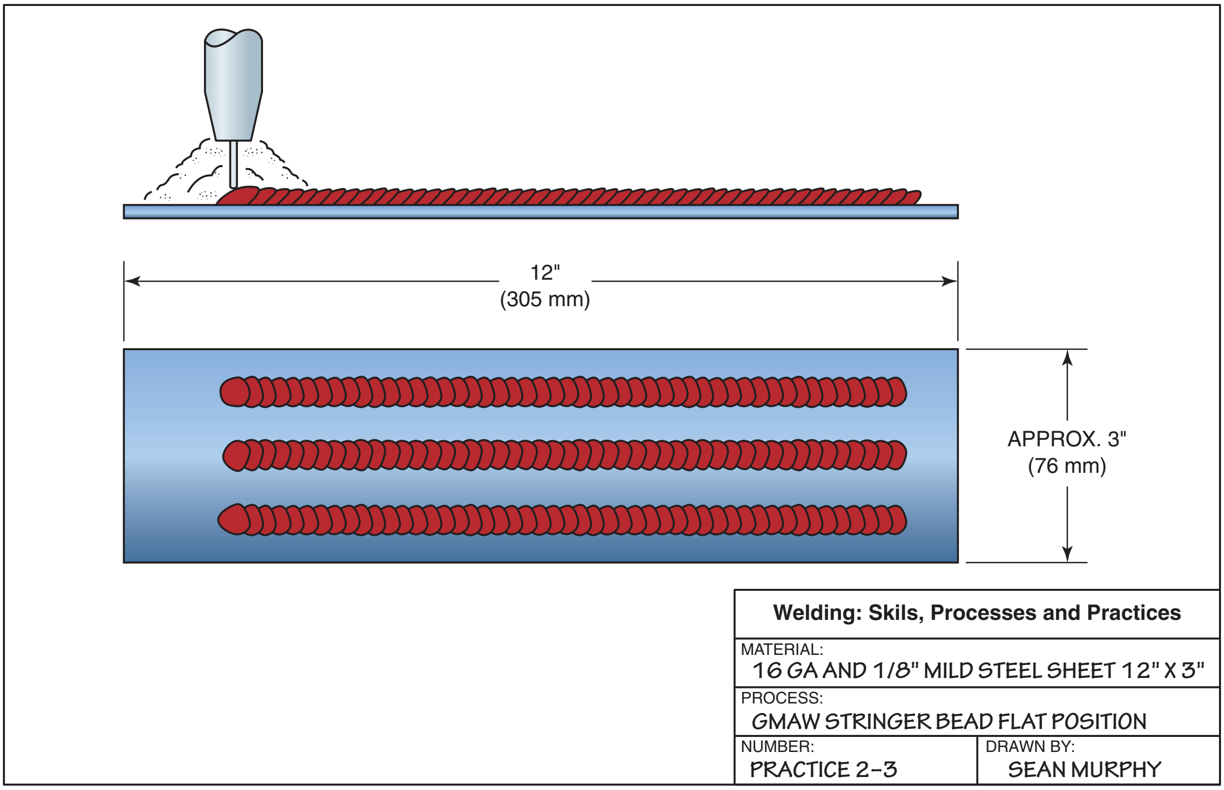
12"
(305 mm)
APPROX. 3"
(76 mm)
Welding: Skils, Processes and Practices
MATERIAL:
16 GA AND 1/8" MILD STEEL SHEET 12" X 3"
PROCESS:
GMAW STRINGER BEAD FLAT POSITION
NUMBER:
PRACTICE 2-3
DRAWN BY:
SEAN MURPHY

Figure 2.29 **Stringer beads in the flat position**

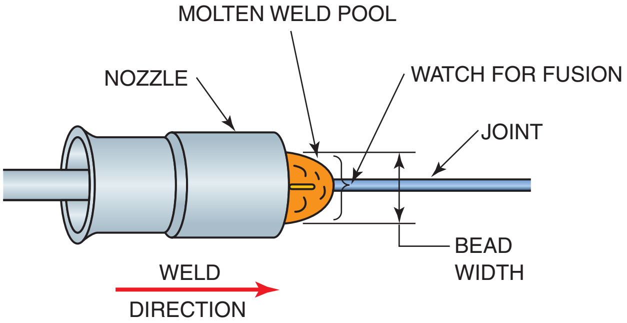
MOLTEN WELD POOL
NOZZLE
WATCH FOR FUSION
JOINT
BEAD WIDTH
WELD DIRECTION

Figure 2.30 **Watching the weld in forehand welding**
The shielding gas nozzle restricts the welder's view of the weld bead when pushing.

satisfactory weld by watching the edge of the molten weld pool, the sparks, and the weld bead produced. Watching the leading edge of the molten weld pool (forehand welding, pushing technique) will show you the molten weld pool fusion and width. Watching the trailing edge of the molten weld pool (backhand welding, dragging technique) will show you the amount of buildup and the relative heat input, Figure 2.31. The quantity and size of sparks produced can indicate the relative location of the filler wire in the molten weld pool. The number of sparks will increase as the wire strikes the solid metal ahead of the molten weld pool. The gun itself will begin to vibrate or bump as the wire momentarily pushes against the cooler, unmelted base metal before it melts. Changes in weld width, buildup, and proper joint tracking can be seen by watching the bead as it appears from behind the shielding gas nozzle.

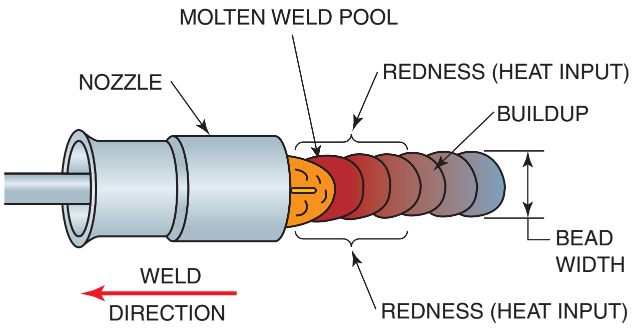
MOLTEN WELD POOL
NOZZLE
REDNESS (HEAT INPUT)
BUILDUP
BEAD WIDTH
WELD DIRECTION
REDNESS (HEAT INPUT)

Figure 2.31 **Watching the weld in backhand welding**
Watch the trailing edge of the molten weld pool.

Repeat each type of bead as needed until consistently good beads are obtained. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-4

Flat Position Butt Joint, Lap Joint, and Tee Joint

Using the same equipment, materials, and procedures listed in Practice 2-3, make welded butt joints, lap joints, and tee joints in the flat position, Figure 2.32A, 2.32B and 2.32C.

- Tack weld the sheets together and place them flat on the welding table, Figure 2.33.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Flat" position portion of the all-position requirement.

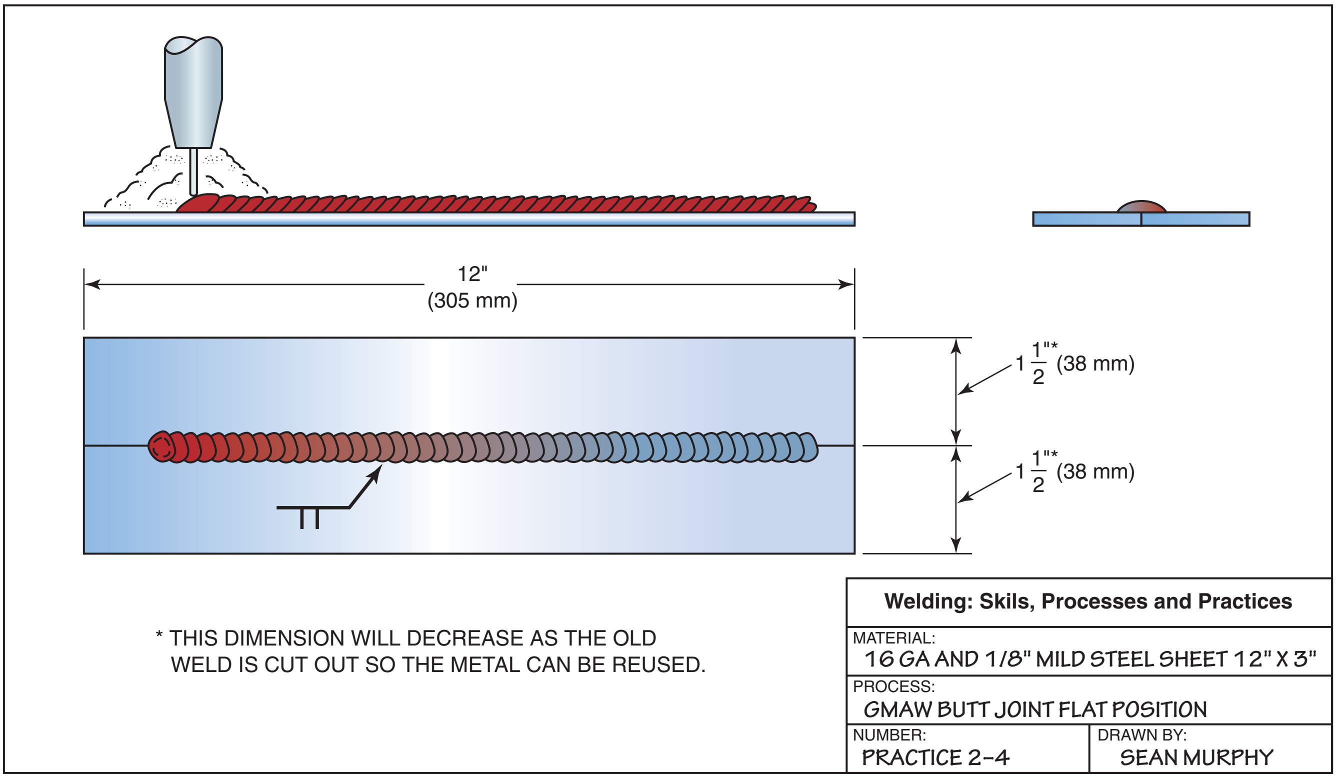
12"
(305 mm)
1 1/2"* (38 mm)
1 1/2"* (38 mm)
* THIS DIMENSION WILL DECREASE AS THE OLD WELD IS CUT OUT SO THE METAL CAN BE REUSED.
Welding: Skils, Processes and Practices
MATERIAL:
16 GA AND 1/8" MILD STEEL SHEET 12" X 3"
PROCESS:
GMAW BUTT JOINT FLAT POSITION
NUMBER:
PRACTICE 2-4
DRAWN BY:
SEAN MURPHY

Figure 2.32 **(A) Butt joint in the flat position**

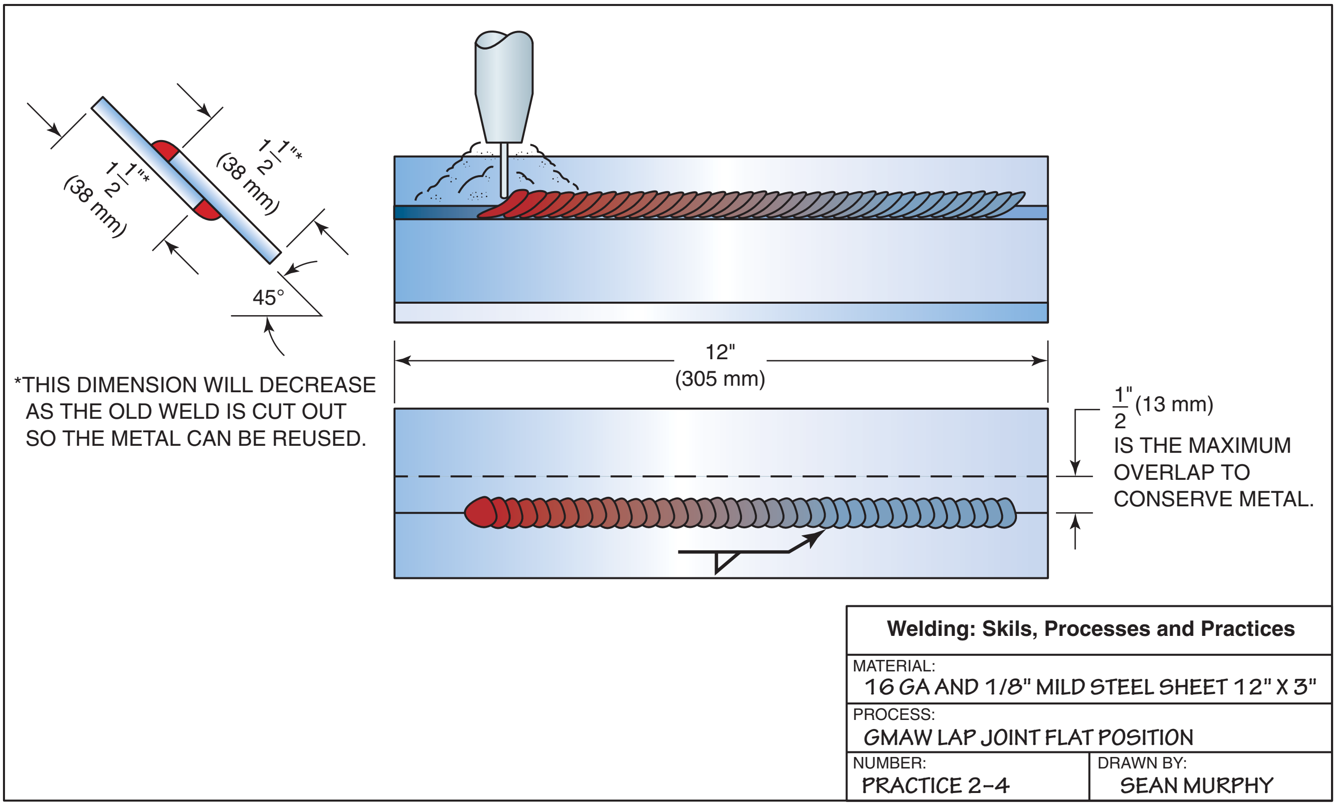
1 1/2"*
(38 mm)
1 1/2"*
(38 mm)
45°
12"
(305 mm)
Welding: Skils, Processes and Practices
MATERIAL:
16 GA AND 1/8" MILD STEEL SHEET 12" X 3"
PROCESS:
GMAW LAP JOINT FLAT POSITION
NUMBER:
PRACTICE 2-4
DRAWN BY:
SEAN MURPHY

*THIS DIMENSION WILL DECREASE AS THE OLD WELD IS CUT OUT SO THE METAL CAN BE REUSED.

1/2" (13 mm) IS THE MAXIMUM OVERLAP TO CONSERVE METAL.

Figure 2.32 (B) Lap joint in the flat position

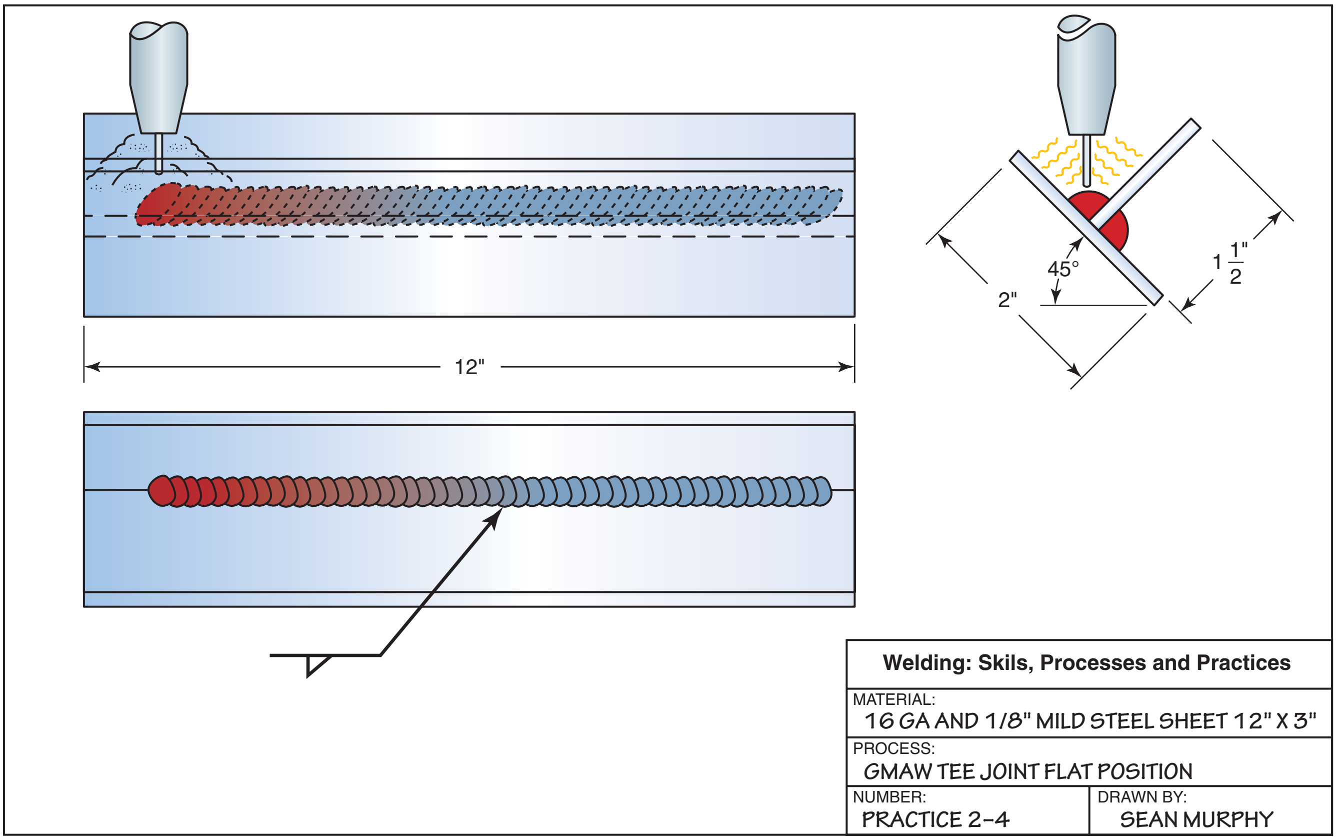
12"
45°
2"
1 1/2"
Welding: Skils, Processes and Practices
MATERIAL:
16 GA AND 1/8" MILD STEEL SHEET 12" X 3"
PROCESS:
GMAW TEE JOINT FLAT POSITION
NUMBER:
PRACTICE 2-4
DRAWN BY:
SEAN MURPHY

Figure 2.32 (C) Tee joint in the flat position

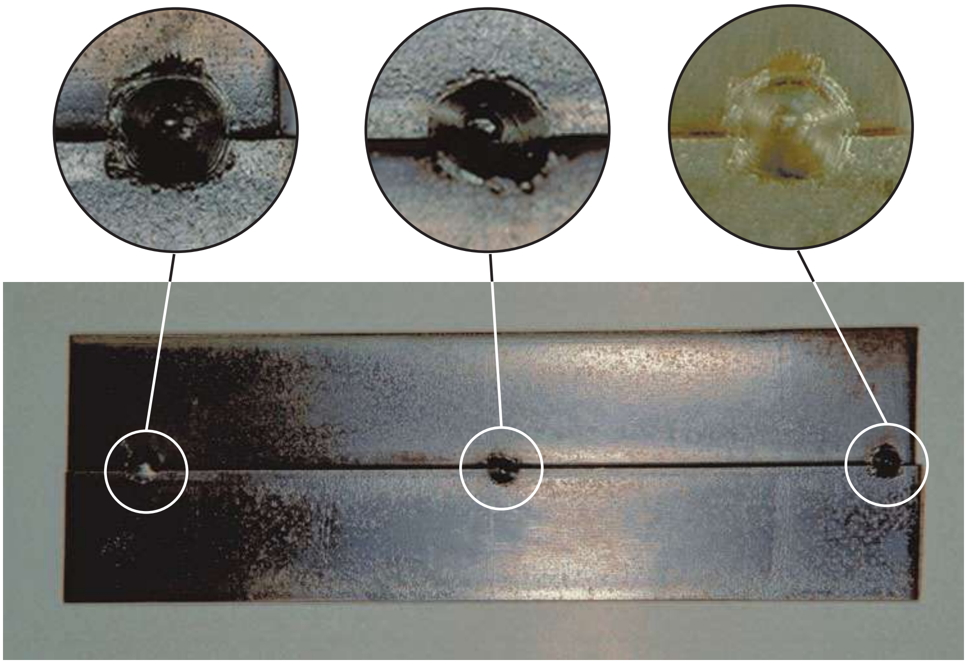

Figure 2.33 **Tack welds**
Use enough tack welds to keep the joint in alignment during welding. Small tack welds are easier to weld over without adversely affecting the weld.
Source: Courtesy of Larry Jeffus

- Starting at one end, run a bead along the joint. Watch the molten weld pool and bead for signs that a change in technique may be required.
- Make any needed changes as the weld progresses. By the time the weld is complete, you should be making the weld nearly perfectly.
- Using the same technique that was established in the last weld, make another weld. This time, the entire 12 in. (305 mm) of weld should be flawless.

Repeat each type of joint with both thicknesses of metal until consistently good beads are obtained. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-5

Flat Position Butt Joint, Lap Joint, and Tee Joint, All with 100% Penetration

Using the same equipment, materials, and setup listed in Practice 2-3, make a welded joint in the flat position with 100% penetration along the entire 12-in. (305-mm) length of the welded joint. Repeat each type of joint until consistently good beads are obtained. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Flat" position portion of the all-position requirement.

VERTICAL UP 3G AND 3F POSITIONS

PRACTICE 2-6

Stringer Bead at a 45° Vertical Up Angle

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Vertical" position portion of the all-position requirement.

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make a vertical up stringer bead on a plate at a 45° inclined angle.

Start at the bottom of the plate and hold the welding gun at a slight pushing or upward angle to the plate, Figure 2.34. Brace yourself, lower your hood, and begin to weld. Depending upon the machine settings and type of shielding gas used, you will make a weave pattern.

If the molten weld pool is large and fluid (hot), use a "C" or "J" weave pattern to allow a longer time for the molten weld pool to cool, Figure 2.35. Do not make the weave so long or fast that the wire is allowed to strike the metal ahead of the molten weld pool. If this happens, spatter increases and a spot or zone of incomplete fusion may occur, Figure 2.36.

If the molten weld pool is small and controllable, use a small "C," zigzag, or "J" weave pattern to control the width and buildup of the weld. A slower speed can also be used. Watch for complete fusion along the leading edge of the molten weld pool. Figure 2.37 shows a weld that did not fuse with the plate.

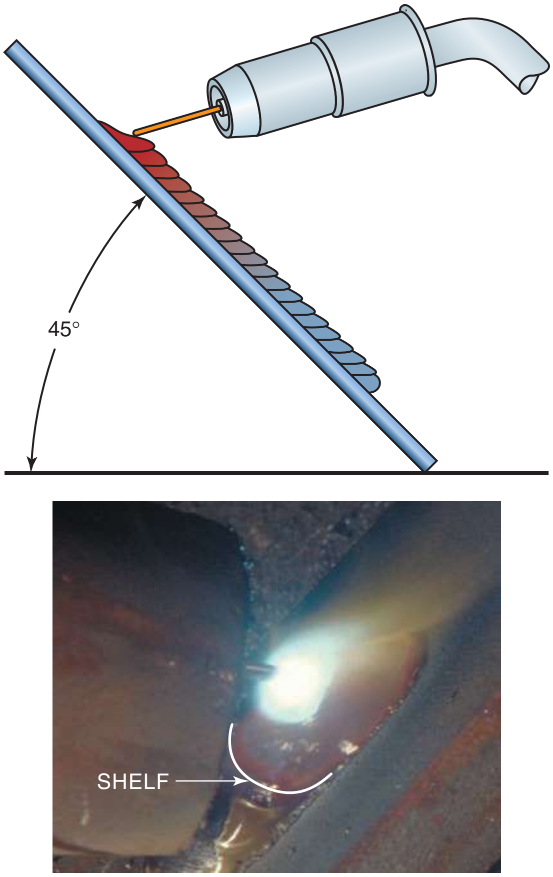

Figure 2.34 **Vertical up position**
Source: Photo courtesy of Larry Jeffus

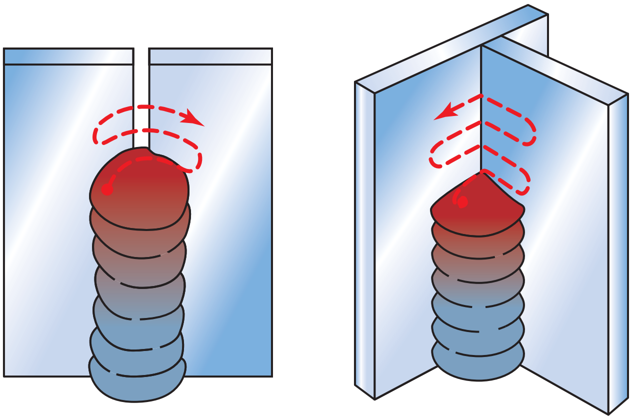

Figure 2.35 **Vertical up weld patterns**
Left, "C" pattern; right, "J" pattern.

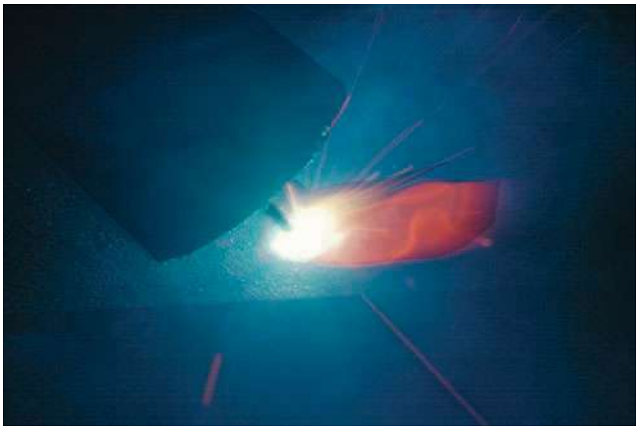

Figure 2.36 **Burst of spatter caused by incorrect electrode contact with base metal**
Source: Courtesy of Larry Jeffus

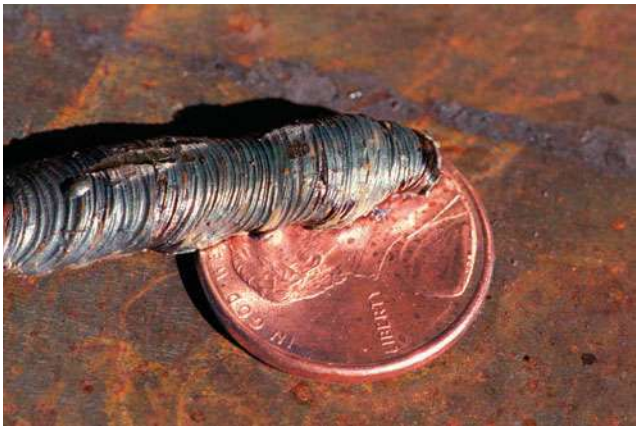

Figure 2.37 **Weld separated from the plate**
There is no fusion between the weld and plate.
Source: Courtesy of Larry Jeffus

A weld that is high and has little or no fusion is too "cold." Changing the welding technique will not correct this problem. The welder must stop welding and make the needed adjustments.

As the weld progresses up the plate, the back or trailing edge of the molten weld pool will cool, forming a shelf to support the molten metal. Watch the shelf to be sure that molten metal does not run over, forming a drip. When it appears that the metal may flow over the shelf, either increase the weave lengths, lengthen the electrode extension, or stop and start the current for brief moments to allow the weld to cool. Stopping for brief moments will not allow the shielding gas to be lost.

Continue to weld along the entire 12-in. (305-mm) length of the plate. Repeat this weld until a straight and uniform weld bead is produced. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-7

Stringer Bead in the Vertical Up Position

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Short-circuit Transfer
Key Indicator 4

Repeat Practice 2-6 and increase the angle of the plate until you have mastered a straight and uniform weld bead in the vertical up position. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-8

Butt Joint, Lap Joint, and Tee Joint in the Vertical Up Position at a 45° Angle

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Vertical" position portion of the all-position requirement.

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make vertical up welded joints on a plate at a 45° inclined angle.

Tack weld the metal pieces together and brace them in position. Check to see that you have free movement along the entire joint to prevent stopping and restarting during the weld. Avoiding stops and starts both speeds up the welding time and eliminates discontinuities.

The weave pattern should allow for adequate fusion on both edges of the joint. Watch the edges to be sure that they are being melted so that adequate fusion and penetration occur.

Repeat each type of joint as needed until consistently good beads are obtained. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-9

Butt Joint, Lap Joint, and Tee Joint in the Vertical Up Position with 100% Penetration

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Vertical" position portion of the all-position requirement.

Using the same equipment, materials, and setup as listed in Practice 2-3, you will increase the plate angle gradually as you develop skill until you are making satisfactory welds in the vertical up position.

Repeat each type of joint as needed until consistently good beads are obtained. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

VERTICAL DOWN 3G AND 3F POSITIONS

The vertical down welding technique can be useful when making some types of welds. The major advantages of this technique are the following:

- Speed—Very high rates of travel are possible.
- Shallow penetration—Thin sections or root openings can be welded with little burn-through.
- Good bead appearance—The weld has a nice width-to-height ratio and is uniform.

Vertical down welds are often used on thin sheet metals or in the root pass in grooved joints. The combination of controlled penetration and higher welding speeds makes vertical down the best choice for such welds. The ease with which welds having a good appearance can be made is deceiving. Generally, more skill is required to make sound welds with this technique than in the vertical up position. The most common problem with these welds is lack of fusion or overlap. To prevent these problems, the arc must be kept at or near the leading edge of the molten weld pool.

PRACTICE 2-10

Stringer Bead at a 45° Vertical Down Angle

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Short-Circuit Transfer
Key Indicator 4

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make a vertical down stringer bead on a plate at a 45° inclined angle.

Holding the welding gun at the top of the plate with a slight dragging angle, Figure 2.38, will help to increase penetration, hold back the molten weld pool, and improve visibility of the weld. Be sure that your movements along the 12-in. (305-mm) length of plate are unrestricted.

Lower your hood and start the weld. Watch both the leading edge and the sides of the molten weld pool for fusion. The leading edge should flow into the base metal, not curl over it. The sides of the molten weld pool should also show fusion into the base metal and not be flashed (ragged) along the edges.

The weld may be made with or without a weave pattern. If a weave pattern is used, it should be a "C" pattern. The "C" should follow the leading edge of the weld. Some changes on the gun angle may help to increase penetration. Experiment with the gun angle as the weld progresses.

Repeat these welds until you have established a rhythm and technique that work well for you. The welds must be straight and uniform and have complete fusion. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

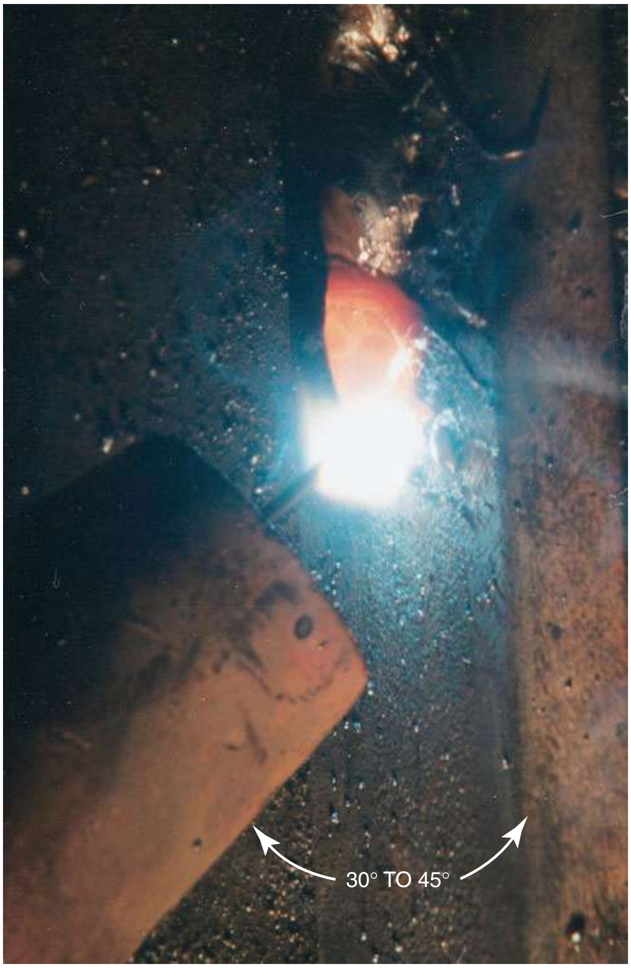

Figure 2.38 Vertical down position
Source: Courtesy of Larry Jeffus

PRACTICE 2-11

Stringer Bead in the Vertical Down Position

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Short-Circuit Transfer
Key Indicator 4

Repeat Practice 2-10 and increase the angle of the plate until you have developed the skill to repeatedly make good welds in the vertical down position. The weld bead must be straight and uniform and have complete fusion. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-12

Butt Joint, Lap Joint, and Tee Joint in the Vertical Down Position

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Vertical" position portion of the all-position requirement.

Using a properly set up and adjusted GMA welding machine, Table 2.7, proper safety protection, 0.035-in. and/or 0.045-in. (0.9-mm and/or 1.2-mm) diameter wire, and two or more pieces of mild steel sheet 12 in. (305 mm) long and 16 gauge and 1/8 in. (3 mm) thick, you will make vertical down welded joints.

Holding the welding gun at the top of the plate with a slight dragging angle, Figure 2.38, will help to increase penetration, hold back the molten weld pool, and improve visibility of the weld. Be sure that your movements along the 12-in. (305-mm) length of plate are unrestricted.

Lower your hood and start the weld. Watch both the leading edge and sides of the molten weld pool for fusion. The leading edge should flow into the base metal, not curl over it. The sides of the molten weld pool should also show fusion into the base metal and not be flashed (ragged) along the edges.

The weld may be made with or without a weave pattern. If a weave pattern is used, it should be a "C" pattern. The "C" should follow the leading edge of the weld. Some changes on the gun angle may help to increase penetration. Experiment with the gun angle as the weld progresses.

Repeat these welds until you have established a rhythm and technique that work well for you. The welds must be straight and uniform and have complete fusion. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-13

Butt Joint and Tee Joint in the Vertical Down Position with 100% Penetration

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Vertical" position portion of the all-position requirement.

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make welded joints with 100% weld penetration.

It may be necessary to adjust the root opening to meet the penetration requirements. The lap joint was omitted from this practice because little additional skill can be developed with it that is not already acquired with the tee joint. Repeat each type of joint until consistently good welds are obtained. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

HORIZONTAL 2G AND 2F POSITIONS

PRACTICE 2-14

Horizontal Stringer Bead at a 45° Angle

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Short-circuit Transfer
Key Indicator 4

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make a horizontal stringer bead on a plate at a 45° reclined angle.

Start at one end with the gun pointed in a slightly upward direction, Figure 2.39. You may use a pushing or a dragging (a leading or a trailing) gun angle, depending upon the current setting and penetration desired. Undercutting along the top edge and overlap along the bottom edge are problems with both gun angles. Careful attention must be paid to the manipulation "weave" technique used to overcome these problems.

The most successful weave patterns are the "C" and "J" patterns. The "J" pattern is the most frequently used. It allows weld metal to be deposited along a shelf created by the previous weave, Figure 2.40. The length of the "J" can be changed to control the weld bead size. Smaller weld beads are easier to control than large ones.

Repeat these welds until you have established the rhythm and technique that work well for you. The weld must be straight and uniform and have complete fusion. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

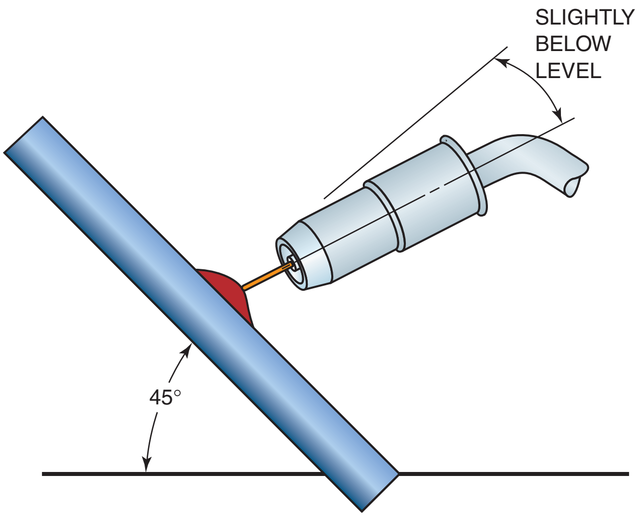

Figure 2.39 45° horizontal position

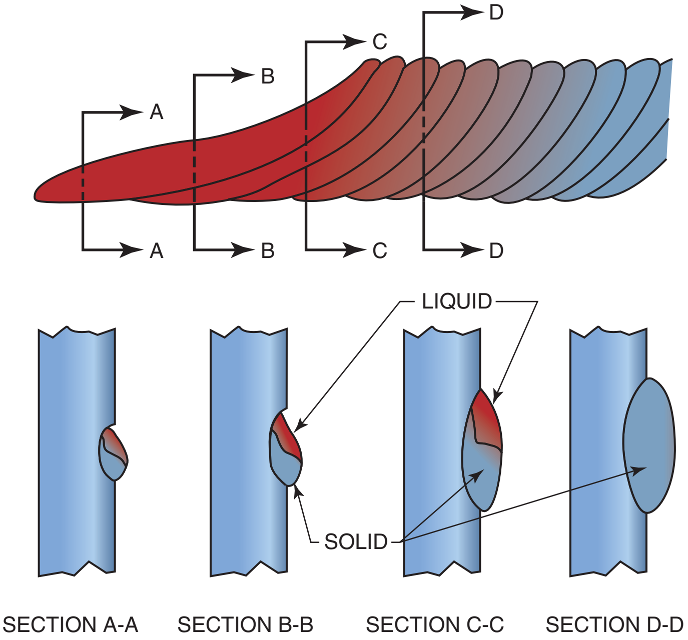

Figure 2.40 **Sections of a "J" weave pattern**
The actual size of the molten weld pool remains small along the weld.

PRACTICE 2-15

Stringer Bead in the Horizontal Position

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Repeat Practice 2-14 and increase the angle of the plate until you have developed the skill to repeatedly make good horizontal welds on a vertical surface. The weld bead must be straight and uniform and have complete fusion. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-16

Butt Joint, Lap Joint, and Tee Joint in the Horizontal Position

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Horizontal" position portion of the all-position requirement.

Using the same equipment, materials, and setup listed in Practice 2-3, you will make horizontal welded joints.

Tack weld the pieces of metal together and brace them in position using the same skills developed in Practice 2-14. Starting at one end, make a weld along the entire length of the joint. When making the butt or lap joints, it may help to recline the plates at a 45° angle until you have developed the technique required. Repeat each type of joint as needed until consistently good welds are obtained. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-17

Butt Joint and Tee Joint in the Horizontal Position with 100% Penetration

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make overhead joints having 100% penetration in the horizontal position.

It may be necessary to adjust the root opening to meet the penetration requirements. Repeat each type of joint as needed until consistently good welds are obtained. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Horizontal" position portion of the all-position requirement.

OVERHEAD 4G AND 4F POSITIONS

There are several advantages to the use of short-circuiting arc metal transfer in the overhead position, including the following:

- Small molten weld pool size—The smaller size of the molten weld pool allows surface tension to hold it in place. Less molten weld pool sag results in improved bead contour with less undercut and fewer icicles, Figure 2.41.
- Direct metal transfer—The direct metal transfer method does not rely on other forces to get the filler metal into the molten weld pool. This results in efficient metal transfer and less spatter and loss of filler metal.

PRACTICE 2-18

Stringer Bead in the Overhead Position

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make a welded stringer bead in the overhead position.

The molten weld pool should be kept as small as possible for easier control. A small molten weld pool can be achieved by using lower current settings, by using a longer wire stickout, by traveling faster, or by pushing the molten weld pool. The technique used is the welder's choice. Often a combination of techniques can be used with excellent results.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Short-circuit Transfer
Key Indicator 4

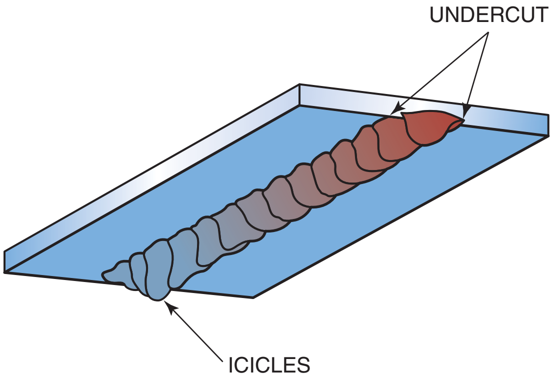

Figure 2.41 **Overhead weld**

Lower current settings require closer control of gun manipulation to ensure that the wire is fed into the molten weld pool just behind the leading edge. The low power will cause overlap and more spatter if this contact position of wire to molten weld pool is not closely maintained.

Faster travel speeds allow the welder to maintain a high production rate even if multiple passes are required to complete the weld. Weld penetration into the base metal at the start of the bead can be obtained by using a slow start or quickly reversing the weld direction. Both the slow start and reversal of weld direction put more heat into the start to increase penetration, Figure 2.42. The higher speed also reduces the amount of weld distortion by reducing the amount of time that heat is applied to a joint.

The pushing, or trailing, gun angle forces the bead to be flatter by spreading it out over a wider area as compared to the bead resulting from a dragging, or backhand, gun angle. The wider, shallow molten weld pool cools faster, resulting in less time for sagging and the formation of icicles.

When welding overhead, extra personal protection is required to reduce the danger of burns. Wear leather sleeves or a leather jacket, and a cap.

Much of the spatter created during overhead welding falls into the shielding gas nozzle. The effectiveness of the shielding gas is reduced, Figure 2.43, and the contact tube may short out to the gas nozzle, Figure 2.44. Turbulence caused by the spatter obstructing the gas may lead to weld contamination. The shorted gas nozzle may arc to the work, causing damage both to the nozzle and to the plate. To control the amount of spatter, a longer stickout and/or a sharper gun-to-plate angle is required to allow most of the spatter to fall clear of the gas nozzle. The nozzle can be dipped, sprayed, or injected automatically, Figure 2.45, with antispatter to help prevent the spatter from sticking. Applying antispatter will not stop the spatter from building up, but it does make its removal much easier. Ensure that the antispatter gel does not restrict the flow of shield gas from the nozzle.

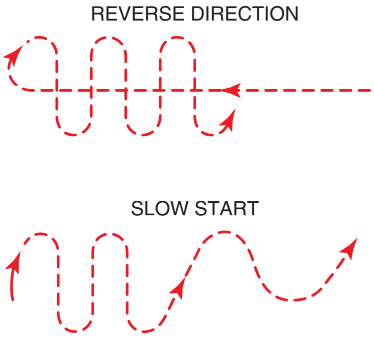

Figure 2.42 Two methods of concentrating heat at the beginning of a weld bead to aid in penetration depth

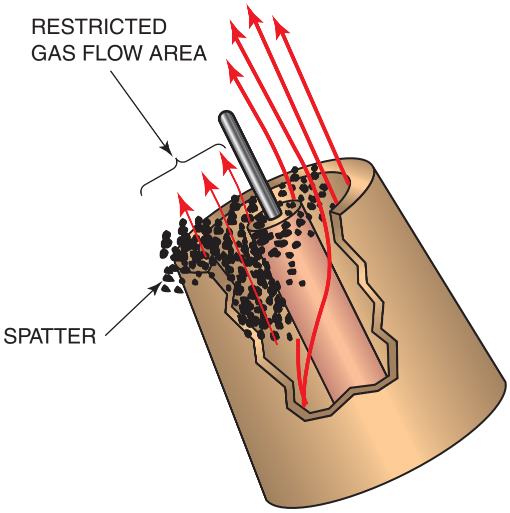

Figure 2.43 Shielding gas flow affected by excessive weld spatter in nozzle

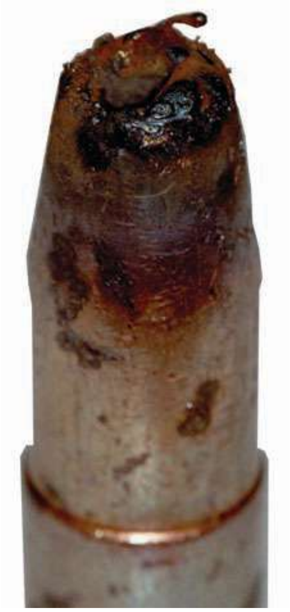

Figure 2.44 Gas nozzle damaged after shorting out against the work

Source: Courtesy of Larry Jeffus

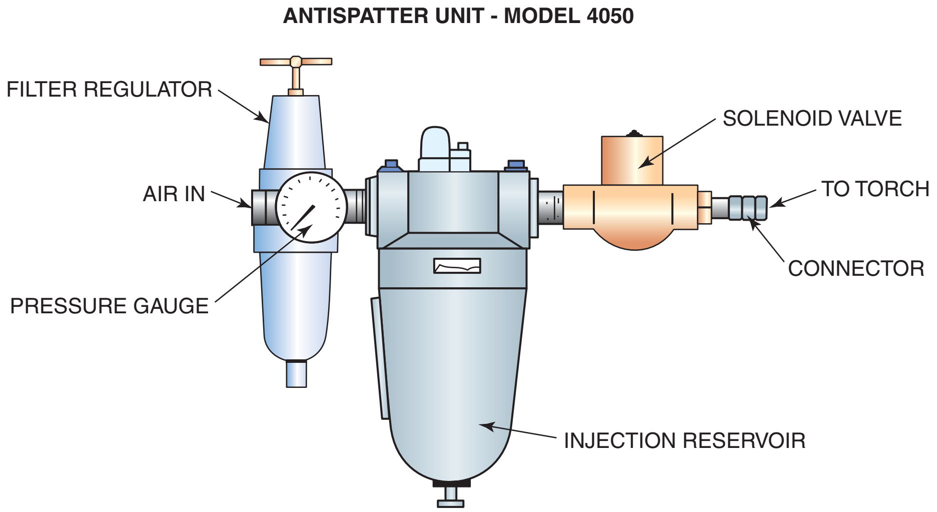

Figure 2.45 Automatic antispatter system that can be added to a GMA welding gun

Make several short weld beads using various techniques to establish the method that is most successful and most comfortable for you. After each weld, stop and evaluate it before making a change. When you have decided on the technique to be used, make a welded stringer bead that is 12 in. (305 mm) long.

Repeat the weld until it can be made straight, uniform, and free from any visual defects. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-19

Butt Joint, Lap Joint, and Tee Joint in the Overhead Position

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make an overhead welded joint.

Tack weld the pieces of metal together and secure them in the overhead position. Be sure you have an unrestricted view and freedom of movement along the joint. Start at one end and make a weld along the joint. Use the same technique developed in Practice 2-18.

Repeat the weld until it can be made straight, uniform, and free from any visual defects. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Overhead" position portion of the all-position requirement.

PRACTICE 2-20

Butt Joint and Tee Joint in the Overhead Position with 100% Penetration

Using the same equipment, materials, and setup as listed in Practice 2-3, you will make overhead welded joints having 100% penetration.

Tack weld the metal together. It may be necessary to adjust the root opening to allow 100% weld metal penetration. During these welds, it may be necessary to use a dragging, or backhand, torch angle. When

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Key Indicator 5, 6
This exercise addresses the "Overhead" position portion of the all-position requirement.

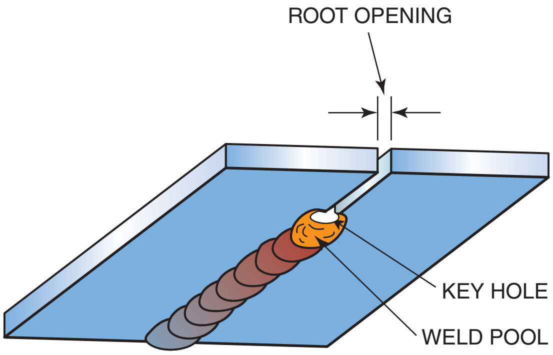

Figure 2.46 Overhead welding

used with a "C" or "J" weave pattern, this torch angle helps to achieve the desired depth of penetration. A key hole just ahead of the molten weld pool is a good sign that the metal is being penetrated, Figure 2.46.

Repeat the weld until it can be made straight, uniform, and free from any visual defects. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

AXIAL SPRAY TRANSFER

PRACTICE 2-21

Stringer Bead, 1G Position

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Spray Transfer
Key Indicator 9

Using a properly set up and adjusted GMA welding machine (see Table 2.8), proper safety protection, 0.035-in. and/or 0.045-in. (0.9-mm and/or 1.2-mm)-diameter wire, and two or more pieces of mild steel plate 12 in. (305 mm) long × 1/4 in. (6 mm) thick, you will make a welded stringer bead in the flat position.

Start at one end of the plate and use either a push or drag technique to make a weld bead along the entire 12-in. (305-mm) length of the metal using spray or pulsed-arc metal transfer. After the weld is complete, check its appearance and make any changes needed to correct the weld, Figure 2.47. Repeat the weld and make any additional adjustments required. After the machine is set, start working on improving the straightness and uniformity of the weld. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Table 2.8 Typical Welding Current Settings for Axial Spray Metal Transfer for Mild Steel

Process	Wire Diameter	Amperage Range (Optimum)	Voltage Range (Optimum)	Shielding Gas
Axial spray	0.030	115–200	15–27	98% Ar + 2% O_2
	0.035	165–300	18–32	
	0.045	200–450	20–34	

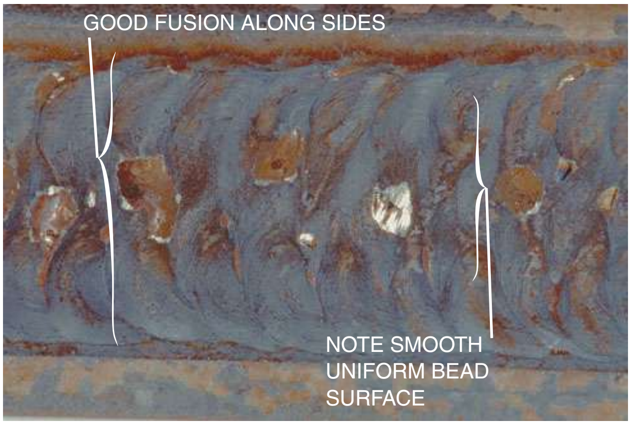

Figure 2.47 **Weld bead made with GMAW axial spray metal transfer**
Source: Courtesy of Larry Jeffus

PRACTICE 2-22

Butt Joint, Lap Joint, and Tee Joint Using the Spray Transfer Method

Using the same equipment, materials, and setup as listed in Practice 2-21, you will make a flat and horizontal weld using spray transfer or pulsed-spray metal transfer, Figure 2.48.

Tack weld the metal together and place the assembly in the flat position on the welding table. Start at one end and make a uniform weld along the entire 12-in. (305-mm) length of the joint. Watch the sides of the fillet weld for signs of undercutting.

Repeat the weld until it can be made straight, uniform, and free from any visual defects. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Spray Transfer
Key Indicator 10, 11

PRACTICE 2-23

Butt Joint and Tee Joint

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 5
Spray Transfer
Key Indicator 10, 11

Using the same equipment, materials, and setup as listed in Practice 2-21, you will make a flat weld using spray transfer. Each weld must pass the

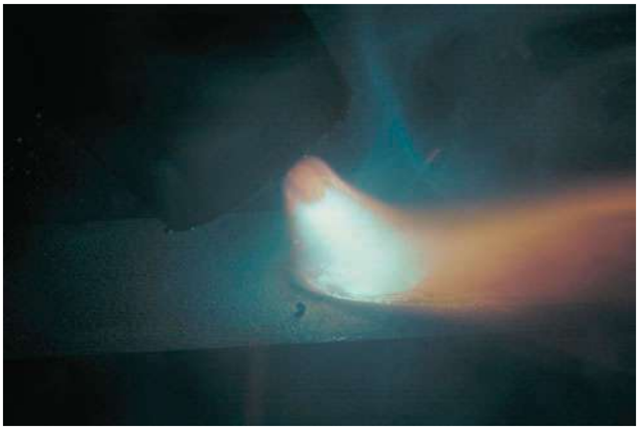

Figure 2.48 **GMAW axial spray metal transfer**
Source: Courtesy of Larry Jeffus

guided bend test. Repeat each type of weld joint as needed until the bend test can be passed. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-24

Gas Metal Arc Welding—Short-Circuit Metal Transfer (GMAW-S) Workmanship Sample

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 3
Key Indicator 1, 2, 3

Module 5
Short-circuit Transfer
Key Indicator 3, 4, 5, 6, 7

Module 9
Key Indicator 1, 2

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 2-24 Date:

Title:

Welding GMAW-S of plate to plate.

Scope:

This procedure is applicable for square groove and fillet welds within the range of 10 ga. (3.4 mm) through 14 ga. (1.9 mm).

Welding may be performed in the following positions: all.

Base Metal:

The base metal shall conform to carbon steel M-1, P-1, and S-1 Group 1 or 2.

Backing material specification: none.

Filler Metal:

The filler metal shall conform to AWS specification no. E70S-X from AWS specification A5.18. This filler metal falls into F-number F-6 and A-number A-1.

Shielding Gas:

The shielding gas, or gases, shall conform to the following compositions and purity:

CO_2 at 30 to 50 cfh or 75% Ar/25% CO_2 at 30 to 50 cfh.

Joint Design and Tolerances:

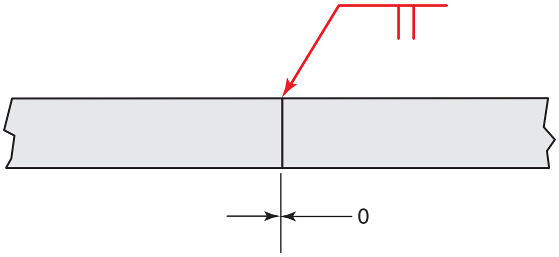

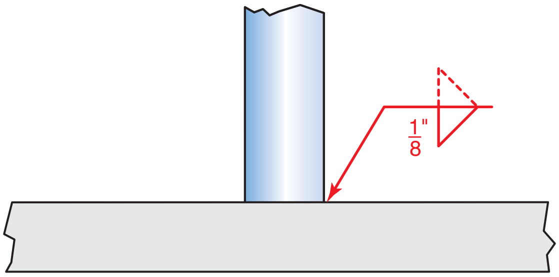

Preparation of Base Metal:

All parts may be mechanically cut or machine PAC unless specified as manual PAC.

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The groove face and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically cleaned of slag, rust, and mill scale. Cleaning must be done with a wire brush or grinder down to bright metal.

Electrical Characteristics:

The current shall be direct-current electrode positive (DCEP). The base metal shall be on the negative side of the line.

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	Size	Amps	Wire-feed Speed, ipm (cm/min)	Volts	Type	Flow	Type	Thickness
E70S-X	0.035 in. (0.9 mm)	90 to 120	180 to 300 (457 to 762)	15 to 19	CO_2 or 75% Ar/ CO_2 25%	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E70S-X	0.045 in. (1.2 mm)	130 to 200	125 to 200 (318 to 508)	17 to 20	CO_2 or 75% Ar/ CO_2 25%	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)

Preheat:

The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:

N/A

Safety:

Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:

Using a 1/2-in. (13-mm) or larger gas nozzle for all welding, first tack weld the plates together according to the drawing. Use the E70S-X filler metal to fuse the plates together. Clean any silicon slag, being sure to remove any trapped silicon slag along the sides of the weld.

Using the E70S-X arc welding electrodes, make a series of stringer beads, no thicker than 3/16 in. (4.7 mm). The 1/8-in. (3.1-mm) fillet welds are to be made with one pass. All welds must be placed in the orientation shown in the drawing.

Interpass Temperature:

The plate should not be heated to a temperature higher than 350°F (175°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:
Any slag must be cleaned off between passes. The weld beads may be cleaned by a hand wire brush, a hand chipping, a punch and hammer, or a needle-scaler. All weld cleaning must be performed with the test plate in the welding position.

Visual Inspection:
Visually inspect the weld for uniformity and discontinuities.

1. There shall be no cracks, no incomplete fusion.
2. There shall be no incomplete joint penetration in groove welds except as permitted for partial joint penetration welds.
3. The Test Supervisor shall examine the weld for acceptable appearance, and shall be satisfied that the welder is skilled in using the process and procedure specified for the text.
4. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm).
5. Where visual examination is the only criterion for acceptance, all weld passes are subject to visual examination, at the discretion of the Test Supervisor.
6. The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length and the maximum diameter shall not exceed 3/32 in. (2.4 mm)
7. Welds shall be free from overlap.

Sketches:
See Figure 2.49.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-25

This practice parallels an AWS D1.1 structural steel limited thickness welder performance qualification test.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 3
Key Indicator 1, 2, 3

Module 5
Short-circuit Transfer
Key Indicator 3, 4, 5, 6

Module 9
Key Indicator 1, 2

Gas Metal Arc Welding—Short-Circuit Metal Transfer (GMAW-S) Limited Thickness Welder Performance Qualification Test Plate for 2G, 3G, and 4G Positions without Backing

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 2-25 Date:

Title:
Welding GMAW-S of plate to plate.

Scope:
This procedure is applicable for V-groove, bevel, or single-bevel welds within the range of 1/8 in. (3.2 mm) through 3/4 in. (19 mm).

Welding may be performed in the following positions: all.

Base Metal:
The base metal shall conform to carbon steel M-1, P-1, and S-1 Group 1 or 2.

Backing material specification: none.

Filler Metal:
The filler metal shall conform to AWS specification no. E70S-X from AWS specification A5.18. This filler metal falls into F-number F-6 and A-number A-1.

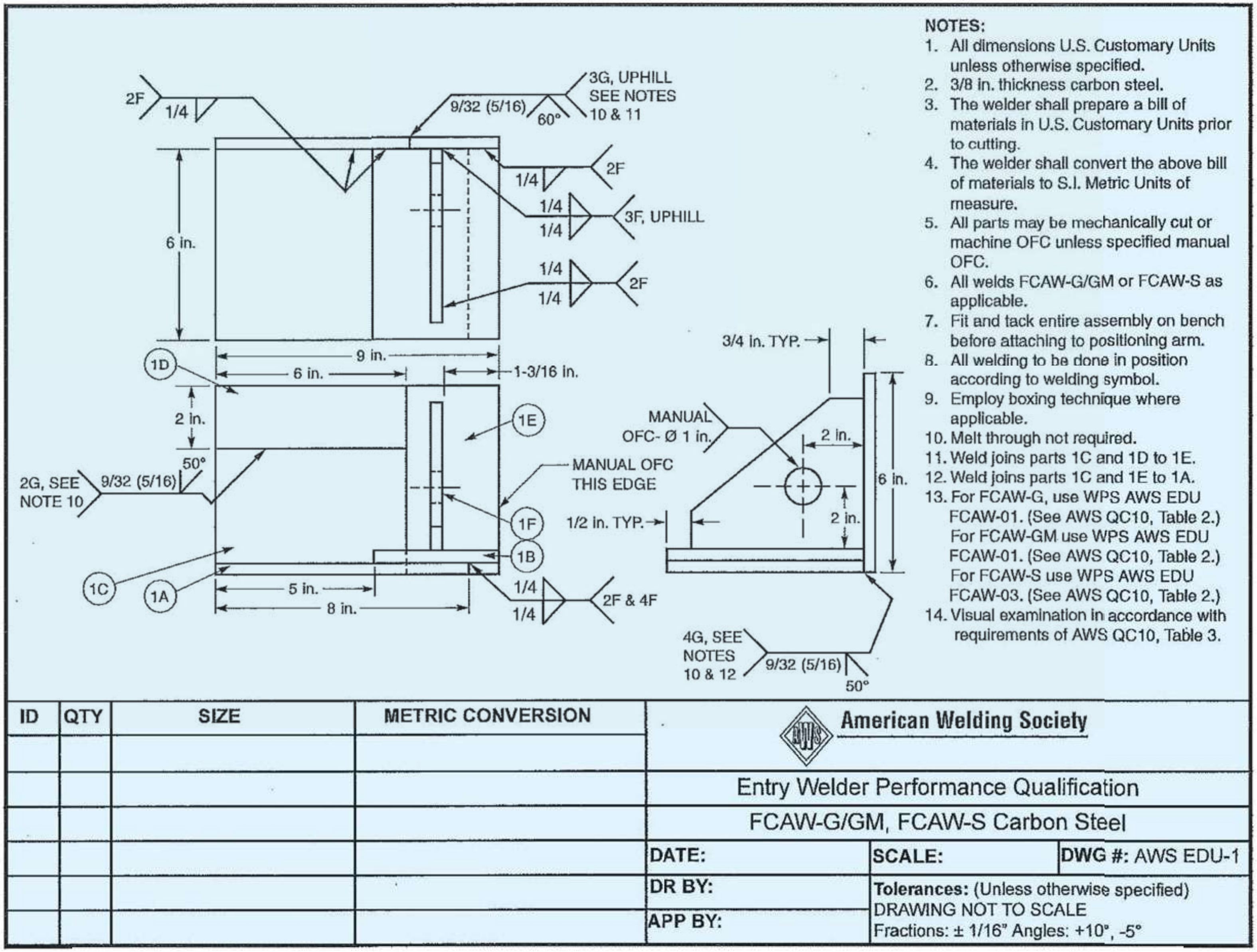
2F
1/4
9/32 (5/16)
60°
3G, UPHILL
SEE NOTES
10 & 11
1/4
2F
1/4
1/4
3F, UPHILL
1/4
1/4
2F
6 in.
9 in.
6 in.
1-3/16 in.
1D
2 in.
50°
2G, SEE
NOTE 10
9/32 (5/16)
1E
MANUAL OFC
THIS EDGE
1F
1B
1C
1A
5 in.
8 in.
1/4
1/4
2F & 4F
3/4 in. TYP.
MANUAL
OFC- Ø 1 in.
2 in.
2 in.
6 in.
1/2 in. TYP.
4G, SEE
NOTES
10 & 12
9/32 (5/16)
50°

NOTES:

1. All dimensions U.S. Customary Units unless otherwise specified.
2. 3/8 in. thickness carbon steel.
3. The welder shall prepare a bill of materials in U.S. Customary Units prior to cutting.
4. The welder shall convert the above bill of materials to S.I. Metric Units of measure.
5. All parts may be mechanically cut or machine OFC unless specified manual OFC.
6. All welds FCAW-G/GM or FCAW-S as applicable.
7. Fit and tack entire assembly on bench before attaching to positioning arm.
8. All welding to be done in position according to welding symbol.
9. Employ boxing technique where applicable.
10. Melt through not required.
11. Weld joins parts 1C and 1D to 1E.
12. Weld joins parts 1C and 1E to 1A.
13. For FCAW-G, use WPS AWS EDU FCAW-01. (See AWS QC10, Table 2.) For FCAW-GM use WPS AWS EDU FCAW-01. (See AWS QC10, Table 2.) For FCAW-S use WPS AWS EDU FCAW-03. (See AWS QC10, Table 2.)
14. Visual examination in accordance with requirements of AWS QC10, Table 3.

ID	QTY	SIZE	METRIC CONVERSION

American Welding Society

Entry Welder Performance Qualification

FCAW-G/GM, FCAW-S Carbon Steel

DATE:
SCALE:
DWG #: AWS EDU-1
DR BY:
APP BY:
Tolerances: (Unless otherwise specified)
DRAWING NOT TO SCALE
Fractions: ± 1/16" Angles: +10°, -5°

Figure 2.49 **GMAW-S workmanship qualification test**

Source: Courtesy of the American Welding Society

Shielding Gas:
The shielding gas, or gases, shall conform to the following compositions and purity: CO_2 at 30 to 50 cfh or 75% Ar/25% CO_2 at 30 to 50 cfh.

Joint Design and Tolerances:

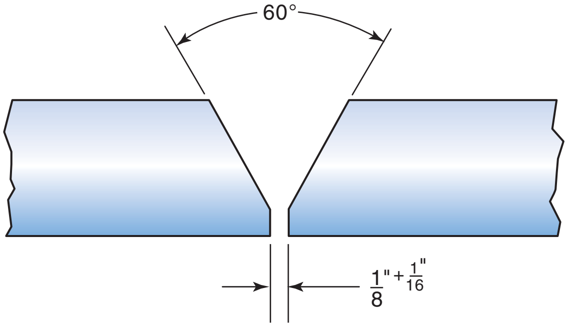

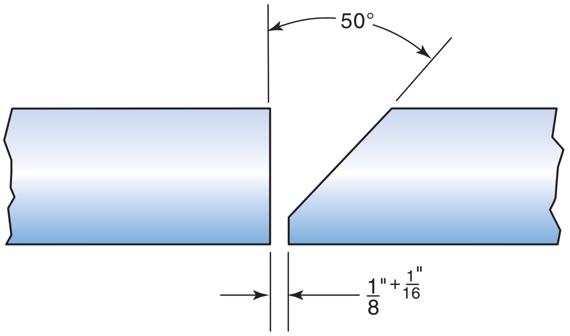

Preparation of Base Metal:
The bevels are to be flame cut on the edges of the plate before the parts are assembled. The beveled surface must be smooth and free of notches. Any roughness or notches deeper than 1/64 in. (0.4 mm) must be ground smooth.

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The groove face and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically cleaned of slag, rust, and mill scale. Cleaning must be done with a wire brush or grinder down to bright metal.

Electrical Characteristics:
The current shall be direct-current electrode positive (DCEP). The base metal shall be on the negative side of the line.

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	Size	Amps	Wire-feed Speed, ipm (cm/min)	Volts	Type	Flow	Type	Thickness
E70S-X	0.035 in. (0.9 mm)	90 to 120	180 to 300 (457 to 762)	15 to 19	CO_2 or 75% Ar/ CO_2 25%	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E70S-X	0.045 in. (1.2 mm)	130 to 200	125 to 200 (318 to 508)	17 to 20	CO_2 or 75% Ar/ CO_2 25%	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)

Preheat:
The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:
N/A

Safety:
Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:
Using a 1/2-in. (13-mm) or larger gas nozzle for all welding, first tack weld the plates together according to the drawing. There should be about a 1/8-in. (1.6-mm) root gap between the plates with V-grooved or beveled edges and 1/16-in. root faces. Use the E70S-X filler wire to make a root pass to fuse the plates together. Clean any silicon slag from the root pass, being sure to remove any trapped silicon slag along the sides of the weld.

Using the E70S-X filler wire, make a series of stringer or weave filler welds, no thicker than 1/4 in. (6.4 mm), in the groove until the joint is filled. Note: The horizontal (2G) weldment should be made with stringer beads only.

Interpass Temperature:
The plate should not be heated to a temperature higher than 350°F (175°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:
Any slag must be cleaned off between passes. The weld beads may be cleaned by a hand wire brush, a hand chipping, a punch and hammer, or a needle-scaler. All weld cleaning must be performed with the test plate in the welding position.

Visual Inspection:*
Visually inspect the weld for uniformity and discontinuities.

1. There shall be no cracks, no incomplete fusion.
2. There shall be no incomplete joint penetration in groove welds except as permitted for partial joint penetration welds.
3. The Test Supervisor shall examine the weld for acceptable appearance, and shall be satisfied that the welder is skilled in using the process and procedure specified for the text.
4. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm)
5. Where visual examination is the only criterion for acceptance, all weld passes are subject to visual examination, at the discretion of the Test Supervisor.

*From Table 3, AWS SENSE QC-10 2004 (Courtesy of the American Welding Society)

6. The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length and the maximum diameter shall not exceed 3/32 in. (2.4 mm).
7. Welds shall be free from overlap.

Bend Test:

The weld is to be mechanically tested only after it has passed the visual inspection. Be sure that the test specimens are properly marked to identify the welder, the position, and the process.

Specimen Preparation:

For 3/8-in. test plates, two specimens are to be located in accordance with the requirements of the figure below left. One is to be prepared for a "transverse face bend," and the other is to be prepared for a "transverse root bend."

Transverse face bend. The weld is perpendicular to the longitudinal axis of the specimen and is bent so that the weld face becomes the tension surface of the specimen. Transverse face-bend specimens shall comply with the requirements of the figure below bottom.

Transverse root bend. The weld is perpendicular to the longitudinal axis of the specimen and is bent so that the weld root becomes the tension surface of the specimen. Transverse face-bend specimens shall comply with the requirements of the figure below bottom.

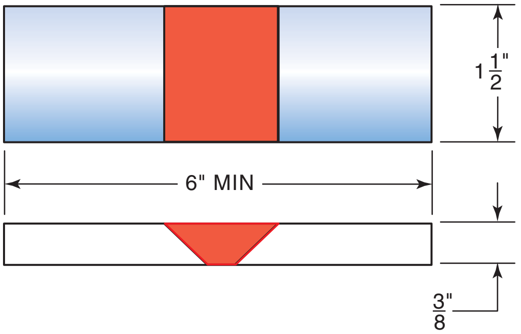

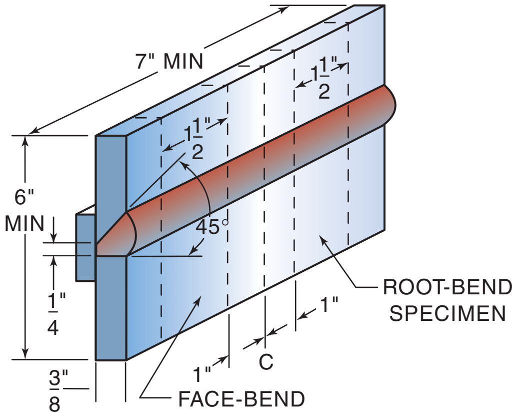

Acceptance Criteria for Face and Root Bends:*

For acceptance, the convex surface of the face- and root-bend specimens shall meet both of the following requirements:

---

*From Table 4, AWS-SENSE QC-10:2004 (Courtesy of the American Welding Society)

1. No single indication shall exceed 1/8 in. (3.2 mm), measured in any direction on the surface.
2. The sum of the greatest dimensions of all indications on the surface, which exceed 1/32 in. (0.8 mm) but are less than or equal to 1/8 in. (3.2 mm), shall not exceed 3/8 in. (9.6 mm).

Cracks occurring at the corner of the specimens shall not be considered unless there is definite evidence that they result from slag inclusion or other internal discontinuities.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 2-26

Gas Metal Arc Welding Spray Transfer (GMAW) Workmanship Sample

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 2-26. Date:

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 3
Key Indicator 1, 2, 3

Module 5
Spray Transfer
Key Indicator 8, 9, 10, 11, 12

Module 9
Key Indicator 1, 2

Title:
Welding GMAW of plate to plate.

Scope:
This procedure is applicable for V-groove and fillet welds within the range of 1/8 in. (3.2 mm) through 1-1/2 in. (38 mm).

Welding may be performed in the following positions: 1G, 1F, 2F.

Base Metal:
The base metal shall conform to carbon steel M-1, P-1, and S-1, Group 1 or 2.

Backing material specification: none.

Filler Metal:
The 0.035 to 0.045 dia. filler metal shall conform to AWS specification no. E70S-X from AWS specification A5.18. This filler metal falls into F-number F-6 and A-number A-1.

Shielding Gas:
The shielding gas, or gases, shall conform to the following compositions: 98% Ar/2% O_2 or 90% Ar/10% CO_2.

Preparation of Base Metal:
The bevels are to be flame cut on the edges of the plate before the parts are assembled. The beveled surface must be smooth and free of notches. Any roughness or notches deeper than 1/64 in. (0.4 mm) must be ground smooth.

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The groove face and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically cleaned of slag, rust, and mill scale. Cleaning must be done with a wire brush or grinder down to bright metal.

Electrical Characteristics:
The current shall be direct-current electrode positive (DCEP). The base metal shall be on the negative side of the line.

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	Size	Amps	Wire-feed Speed, ipm (cm/min)	Volts	Type	Flow	Type	Thickness
E70S-X	0.035 in. (0.9 mm)	180 to 230	400 to 550 (1016 to 1397)	25 to 27	Ar plus 2% O_2 or 90% Ar/ 10% CO_2	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E70S-X	0.045 in. (1.2 mm)	260 to 340	300 to 500 (762 to 1270)	25 to 30	Ar plus 2 O_2 or 90% Ar/ 10% CO_2	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)

Preheat:

The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:

N/A

Safety:

Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:

Using a 3/4-in. (19-mm) or larger gas nozzle for all welding, first tack weld the plates together according to the drawing. There should be about a 1/16-in. (1.6-mm) root gap between the plates with V-grooved or beveled edges. Use the E70S-X arc welding electrodes to make a root pass to fuse the plates together. Clean any silicon slag from the root pass, being sure to remove any trapped silicon slag along the sides of the weld.

Using the E70S-X arc welding electrodes, make a series of stringer or weave filler welds, no thicker than 1/4 in. (6.4 mm), in the groove until the joint is filled. The 1/4-in. (6.4-mm) fillet welds are to be made with one pass.

Interpass Temperature:

The plate should not be heated to a temperature higher than 350°F (175°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:

Any slag must be cleaned off between passes. The weld beads may be cleaned by a hand wire brush, a hand chipping, a punch and hammer, or a needle-scaler. All weld cleaning must be performed with the test plate in the welding position.

Inspection:*

Visually inspect the weld for uniformity and discontinuities.

1. There shall be no cracks, no incomplete fusion.
2. There shall be no incomplete joint penetration in groove welds except as permitted for partial joint penetration welds.
3. The Test Supervisor shall examine the weld for acceptable appearance, and shall be satisfied that the welder is skilled in using the process and procedure specified for the text.

*From Table 3, AWS SENSE QC-10:2004 (Courtesy of the American Welding Society)

4. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm)
5. Where visual examination is the only criterion for acceptance, all weld passes are subject to visual examination, at the discretion of the Test Supervisor.
6. The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length and the maximum diameter shall not exceed 3/32 in. (2.4 mm).
7. Welds shall be free from overlap.

Sketches:

See Figure 2.50.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-27

Gas Metal Arc Welding (GMAW) Spray Transfer Limited Thickness Welder Performance Qualification Test Plate for 1-G, Position, with Backing

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice (2-27). Date:

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 3
Key Indicator 1, 2, 3

Module 5
Spray Transfer
Key Indicator 8, 9, 11

Module 9
Key Indicator 1, 2

Title:

Welding GMAW of plate to plate.

Scope:

This procedure is applicable for V-groove and fillet welds within the range of 1/8 in. (3.2 mm) through 1-1/2 in. (38 mm).

Welding may be performed in the following positions: 1G, 2F.

Base Metal:

The base metal shall conform to carbon steel M-1, P-1, and S-1, Group 1 or 2.

Backing material specification: none.

Filler Metal:

The filler metal shall conform to AWS specification no. 0.035 to 0.045 dia. E70S-X from AWS specification A5.18. This filler metal falls into F-number F-6 and A-number A-1.

Shielding Gas:

The shielding gas, or gases, shall conform to the following compositions and purity: 98% Ar/2% O_2 or 90% Ar/10% CO_2

Joint Design and Tolerances:

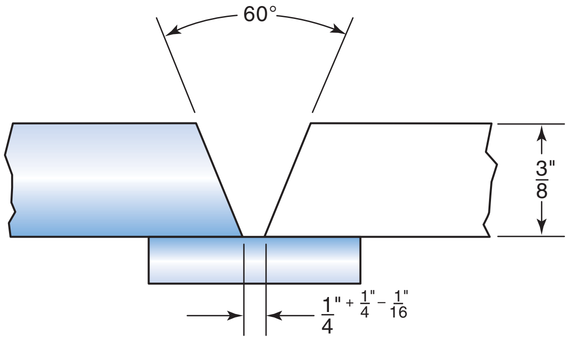

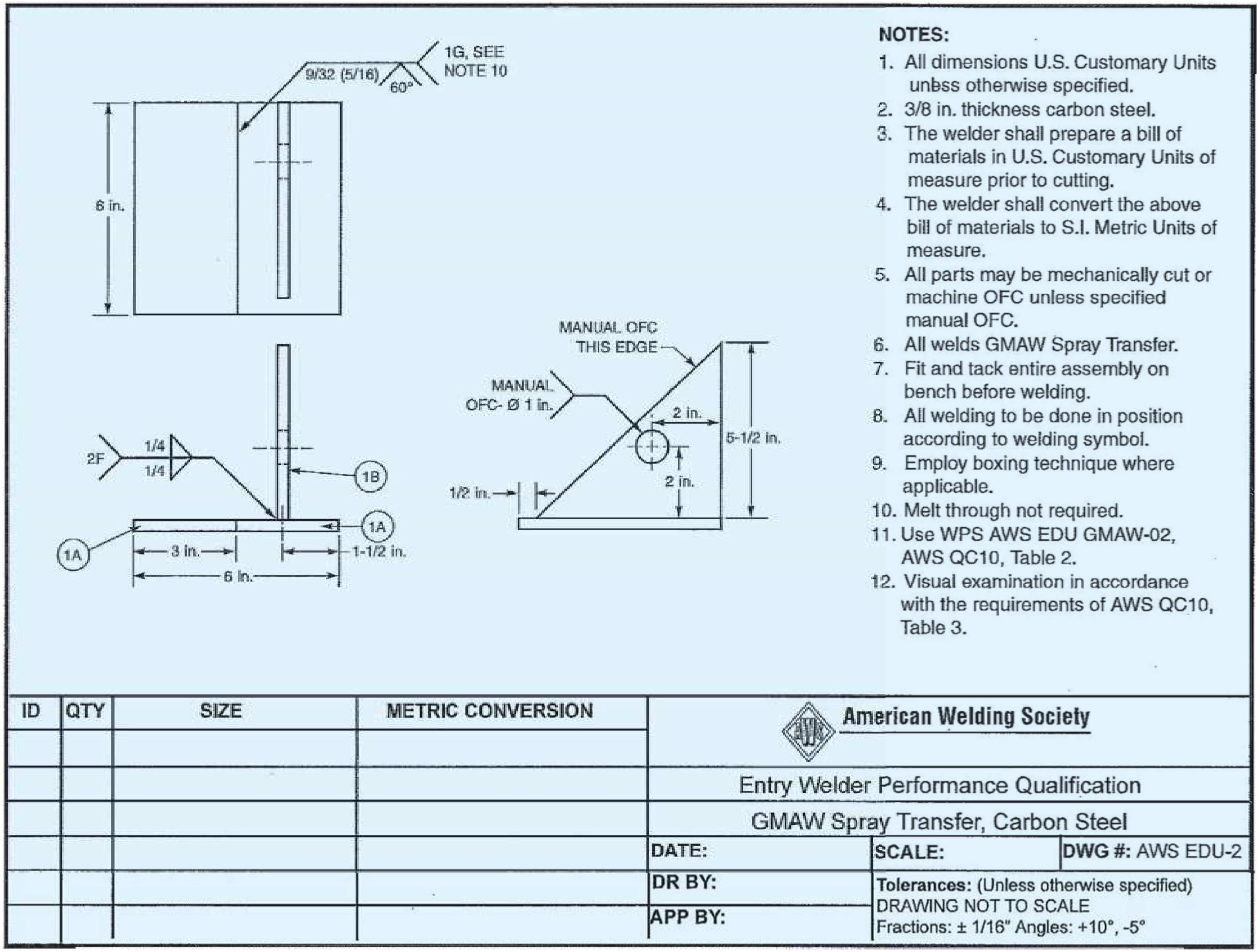

NOTES:

1. All dimensions U.S. Customary Units unbss otherwise specified.
2. 3/8 in. thickness carbon steel.
3. The welder shall prepare a bill of materials in U.S. Customary Units of measure prior to cutting.
4. The welder shall convert the above bill of materials to S.I. Metric Units of measure.
5. All parts may be mechanically cut or machine OFC unless specified manual OFC.
6. All welds GMAW Spray Transfer.
7. Fit and tack entire assembly on bench before welding.
8. All welding to be done in position according to welding symbol.
9. Employ boxing technique where applicable.
10. Melt through not required.
11. Use WPS AWS EDU GMAW-02, AWS QC10, Table 2.
12. Visual examination in accordance with the requirements of AWS QC10, Table 3.

ID	QTY	SIZE	METRIC CONVERSION

American Welding Society

Entry Welder Performance Qualification

GMAW Spray Transfer, Carbon Steel

DATE:
SCALE:
DWG #: AWS EDU-2
DR BY:
APP BY:
Tolerances: (Unless otherwise specified)
DRAWING NOT TO SCALE
Fractions: ± 1/16" Angles: +10°, -5°

Figure 2.50 GMAW (spray transfer) workmanship qualification test

Source: Courtesy of the American Welding Society

Preparation of Base Metal:

The bevels are to be flame cut on the edges of the plate before the parts are assembled. The beveled surface must be smooth and free of notches. Any roughness or notches deeper than 1/64 in. (0.4 mm) must be ground smooth.

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The groove face and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically cleaned of slag, rust, and mill scale. Cleaning must be done with a wire brush or grinder down to bright metal.

Electrical Characteristics:

The current shall be direct-current electrode positive (DCEP). The base metal shall be on the negative side of the line.

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	**Size**	**Amps**	**Wire-feed Speed ipm (cm/min)**	**Volts**	**Type**	**Flow**	**Type**	**Thickness**
E70S-X	0.035 in. (0.9 mm)	180 to 230	400 to 550 (1016 to 1397)	25 to 28	Ar plus 2% O_2 or 90% Ar/10% CO_2	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E70S-X	0.045 in. (1.2 mm)	260 to 340	300 to 500 (762 to 1270)	25 to 30	Ar plus 2 O_2 or 90% Ar/10% CO_2	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)

Preheat:

The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:

N/A

Safety:

Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:

Using a 3/4-in. (19-mm) or larger gas nozzle for all welding, first tack weld the plates and backing strip together according to the drawing below. There should be about a 1/4-in. (1.6-mm) root gap between the plates with V-grooved or beveled edges. Use the E70S-X arc welding electrodes to make a root pass to fuse the plates together. Clean any silicon slag from the root pass, being sure to remove any trapped silicon slag along the sides of the weld.

Using the E70S-X arc welding electrodes, make a series of stringer filler welds, no thicker than 1/4 in. (6.4 mm), in the groove until the joint is filled.

Interpass Temperature:

The plate should not be heated to a temperature higher than 350°F (175°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:

Any slag must be cleaned off between passes. The weld beads may be cleaned by a hand wire brush, a hand chipping, a punch and hammer, or a needle-scaler. All weld cleaning must be performed with the test plate in the welding position.

Visual Inspection:*

Visually inspect the weld for uniformity and discontinuities.

1. There shall be no cracks, no incomplete fusion.
2. There shall be no incomplete joint penetration in groove welds except as permitted for partial joint penetration welds.
3. The Test Supervisor shall examine the weld for acceptable appearance, and shall be satisfied that the welder is skilled in using the process and procedure specified for the text.
4. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm).
5. Where visual examination is the only criterion for acceptance, all weld passes are subject to visual examination, at the discretion of the Test Supervisor.
6. The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length and the maximum diameter shall not exceed 3/32 in. (2.4 mm).
7. Welds shall be free from overlap.

Sketches:

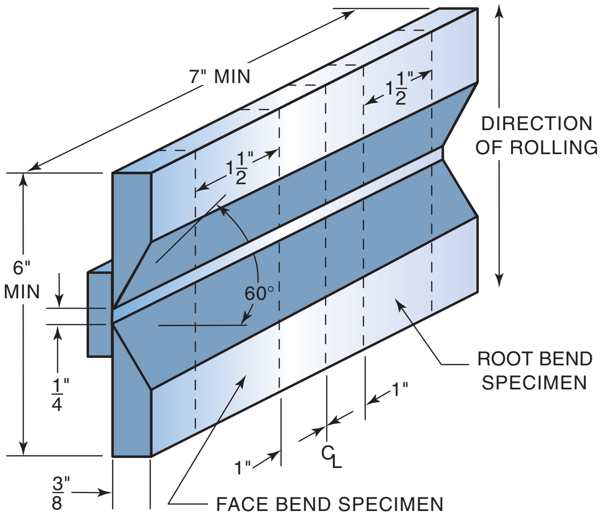

Bend-Test:

The weld is to be mechanically tested only after it has passed the visual inspection. Be sure that the test specimens are properly marked to identify the welder, the position, and the process.

Specimen Preparation

For 3/8-in. test plates, two specimens are to be located in accordance with the requirements below. One is to be prepared for a transverse face bend, and the other is to be prepared for a transverse root bend.

*From Table 3, AWS SENSE QC-10:2004 (Courtesy of the American Welding Society)

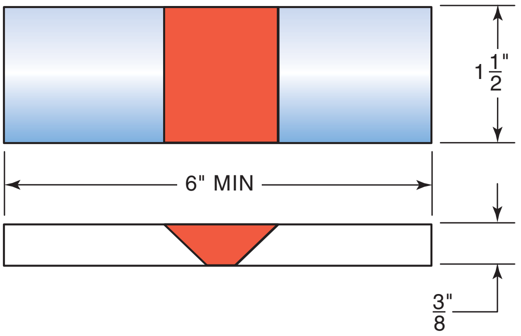

Transverse face bend. The weld is perpendicular to the longitudinal axis of the specimen and is bent so that the weld face becomes the tension surface of the specimen. Transverse face bend specimens shall comply with the requirements below.

Transverse root bend. The weld is perpendicular to the longitudinal axis of the specimen and is bent so that the weld root becomes the tension surface of the specimen. Transverse root bend specimens shall comply with the requirements below.

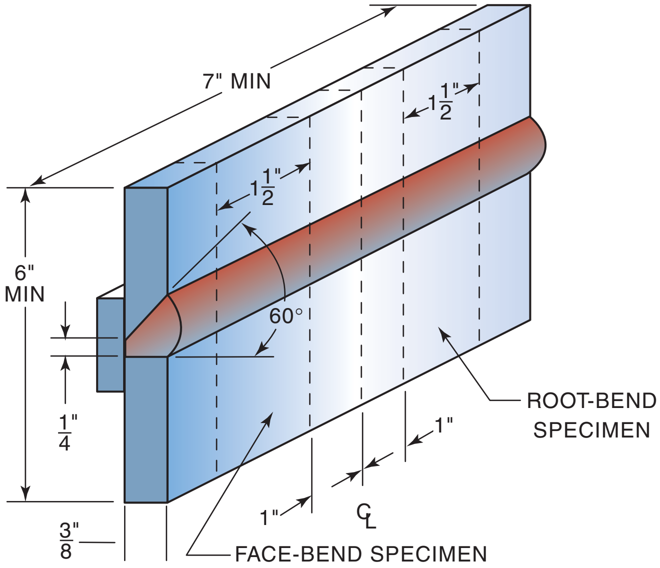

Acceptance Criteria for Face and Root Bends*

For acceptance, the convex surface of the face and root bend specimens shall meet both of the following requirements:

1. No single indication shall exceed 1/8 in. (3.2 mm), measured in any direction on the surface.
2. The sum of the greatest dimensions of all indications on the surface, which exceed 1/32 in. (0.8 mm), but are less than or equal to 1/8 in. (3.2 mm), shall not exceed 3/8 in. (9.6 mm).

Cracks occurring at the corner of the specimens shall not be considered unless there is definite evidence that they result from slag or inclusions or other internal discontinuities.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

*From Table 4, AWS SENSE QC-10:2004 (Courtesy of the American Welding Society)

SUMMARY

Slight changes in welding gun angle and electrode extension can make significant differences in the quality of the weld produced. As a new welder you might find it difficult to tell the effect of these changes if they are slight. Therefore, as you start to learn this process it is a good idea to make more radical changes so it is easier for you to see their effects on the weld. Later as you develop your skills you can use these slight changes to aid in controlling the weld's quality and appearance as it progresses along the joint. Small adjustments in your welding technique are required to compensate for slight changes that occur along a welding joint, such as joint gap and the increasing temperature of the base metal.

Variations in conditions can significantly affect welding setup for the GMA process. Before starting an actual weld in the field you should practice to test your setup. Practice on scrap metal of a similar thickness and type of metal to be welded. A practice weld before you begin welding can significantly increase the chances that your weld will meet standards or specifications. Making these sample or test welds is more important when you are welding in the field, since welds outside the shop are more difficult to control and anticipate. Think of this much as an athlete warms up before competing.

You will find it beneficial when you are initially setting up your welder to have someone assist you, so that he or she can make changes in the welding machine's settings as you are welding. This teamwork can significantly increase your setup accuracy and reduce setup time. Later on in the field, having developed a keen eye for watching the weld, you can then make these adjustments for yourself more rapidly and accurately. Working with another student in a group effort like this will also give you a better understanding of how other individuals' setup preferences affect their welds. Welding is an art, and therefore each welder may have slight differences in preference for voltage, amperage, gas flow, and other setup variables. This gives you an opportunity to learn more from others.

REVIEW

1. What items make up a basic semiautomatic welding system?
2. What must be done to the shielding gas cylinder before the valve protection cap is removed?
3. Why is the shielding gas valve "cracked" before the flowmeter regulator is attached?
4. What causes the electrode to bird-nest?
5. Why must all fittings and connections be tight?
6. What parts should be activated by depressing the gun switch?
7. What benefit does a welding wire's cast provide?
8. What can be done to determine the location of a problem that stops the wire from being successfully fed through the conduit?
9. What are the advantages of using a feed roller pressure that is as light as possible?

10. Why should the feed roller drag prevent the spool from coasting to a stop when the feed stops?
11. Why must you always wind the wire tightly into a ball or cut it into short lengths before discarding it in the proper waste container?
12. Why would the flowmeter ball float at different heights with different shielding gases if the shielding gases are flowing at the same rate?
13. Using Table 2.1, determine the amperage if 400 in. (10.2 m) of 0.45-in. (1.2-mm) steel wire is fed in one minute.
14. How is the amperage adjusted on a GMA welder?
15. What happens to the weld as the electrode extension is lengthened?
16. What is the effect on the weld of changing the welding angle from a dragging to a pushing angle?
17. What are the advantages of adding oxygen or CO_2 to argon for welds on steel?
18. What are the advantages of using CO_2 for making GMA welds on steel?
19. What is mill scale?
20. What type of porosity is most often caused by mill scale?
21. What should the welder watch if the view of the weld is obstructed by the shielding gas nozzle?
22. When you are making a vertical weld and it appears that the weld metal is going to drip over the shelf, what should you do?
23. What are the advantages of making vertical down welds?
24. How can small weld beads be maintained during overhead welds?
25. How can spatter be controlled on the nozzle when making overhead welds?
26. How should the electrode be manipulated for the deepest penetration when using the pulsed-arc metal transfer process?

This page intentionally left blank

CHAPTER 3

Flux Cored Arc Welding Equipment, Setup, and Operation

OBJECTIVES

After completing this chapter, the student should be able to

- describe the flux cored arc (FCA) welding process
- list the equipment required for an FCA welding workstation
- list five advantages of FCA welding, and explain four of its limitations
- tell how electrodes are manufactured and explain the purpose of the electrode cast and helix
- list four things flux can provide to the weld and how fluxes are classified
- explain what each of the digits in a standard FCAW electrode identification number mean
- describe the proper care and handling of FCAW electrodes
- list two common shielding gases used in FCAW, and contrast their benefits related to cost, productivity, and quality
- list three differences in an FCA weld when the gun angle is changed
- identify the two modes of metal transfer and contrast them in regard to application and quality
- list four effects that electrode extension has on FCA welding
- list three things that can cause weld porosity and how it can be prevented

KEY TERMS

air-cooled
coils
deoxidizers
dual shield
flux cored arc welding (FCAW)
lime-based flux
rutile-based flux
self-shielding
slag
smoke extraction nozzles
spools
water-cooled

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter:

Module 6: Flux Cored Arc Welding (FCAW-G, FCAW-S)

Key Indicator 1: Performs safety inspections of FCAW-G/GM, FCAW-S equipment and accessories

Key Indicator 2: Makes minor external repairs to FCAW-G/GM, FCAW-S equipment and accessories

Key Indicator 3: Sets up for FCAW-G/GM, FCAW-S operations on carbon steel

INTRODUCTION

Flux cored arc welding (FCAW) is a fusion welding process in which weld heating is produced from an arc between the work and a continuously fed filler metal electrode. Atmospheric shielding is provided completely or in part by the flux sealed within the tubular electrode, Figure 3.1. Extra shielding may or may not be supplied through a nozzle in the same way as in GMAW.

Although the process was introduced in the early 1950s, it represented less than 5% of the total amount of welding done in 1965. In 2005, it passed the 50% mark and is still rising. The rapid rise in the use of FCAW has been due to a number of factors. Improvements in the fluxes, smaller electrode diameters, increased reliability of the equipment, better electrode feed systems, and improved guns have all led to the increased usage. Guns equipped with **smoke extraction nozzles** and electronic controls are the latest in a long line of improvements to this process, Figure 3.2.

PRINCIPLES OF OPERATION

FCA welding is similar in a number of ways to the operation of GMA welding, Figure 3.3. Both processes use a constant-potential (CP) or constant-voltage (CV) power supply. Constant potential and voltage are terms that have the same meaning. CP power supplies provide a controlled voltage (potential) to the welding electrode. The amperage (current) varies with the speed that the electrode is being fed into the molten weld pool. Just as in GMA welding, higher electrode feed speeds produce higher currents and slower feed speeds result in lower currents, assuming all other conditions remain constant.

The effects on the weld of electrode extension, gun angle, welding direction, travel speed, and other welder manipulations are similar to those experienced in GMA welding. As in GMA welding, having a correctly set welder does not ensure a good weld. The skill of the welder is an important factor in producing high-quality welds.

The flux inside the electrode protects the molten weld pool from the atmosphere, improves strength through chemical reactions and alloys, and improves the weld shape.

Atmospheric contamination of molten weld metal occurs as it travels across the arc gap and within the pool before it solidifies. The major atmospheric contaminations come from oxygen and nitrogen, the major elements in air. The addition of fluxing and gas-forming elements to the core electrode reduces or eliminates their effects.

Improved strength and other physical or corrosion-resistant properties of the finished weld are improved by the flux. Small additions of alloying elements, deoxidizers, and gas-forming and slag agents all can improve the desired weld properties. Carbon, chromium, and vanadium can be added to improve hardness, strength, creep resistance, and corrosion resistance. Aluminum, silicon, and titanium all help remove oxides and/or nitrides in the weld. Potassium, sodium, and zirconium are added to the flux and form a slag.

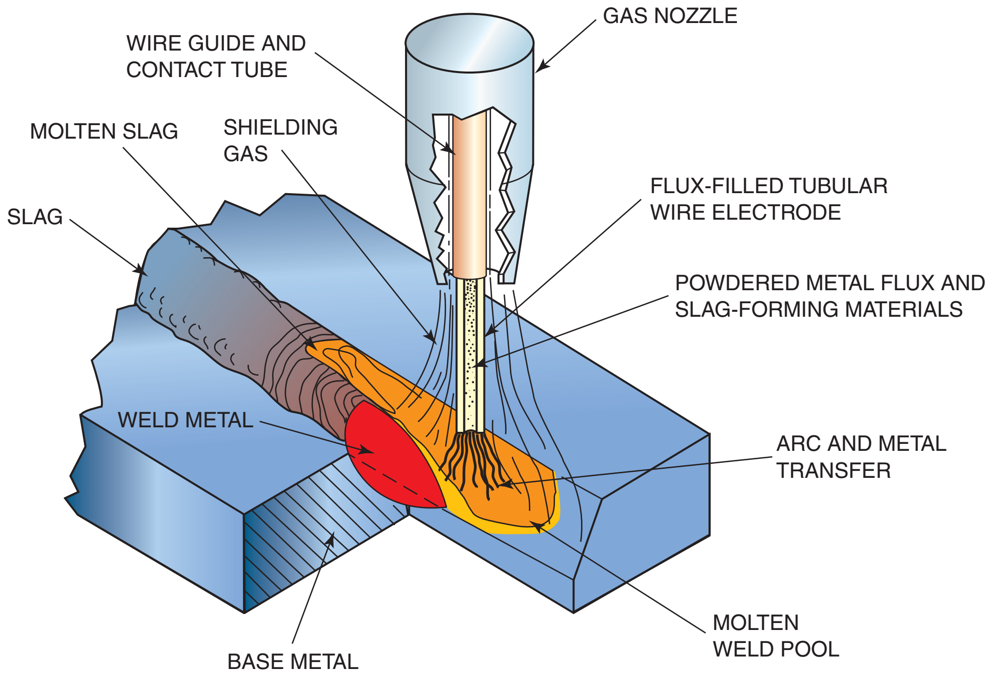
GAS NOZZLE
WIRE GUIDE AND CONTACT TUBE
MOLTEN SLAG
SHIELDING GAS
SLAG
FLUX-FILLED TUBULAR WIRE ELECTRODE
POWDERED METAL FLUX AND SLAG-FORMING MATERIALS
WELD METAL
ARC AND METAL TRANSFER
MOLTEN WELD POOL
BASE METAL

(A) GAS-SHIELDED FLUX CORED ARC WELDING (FCAW-G)

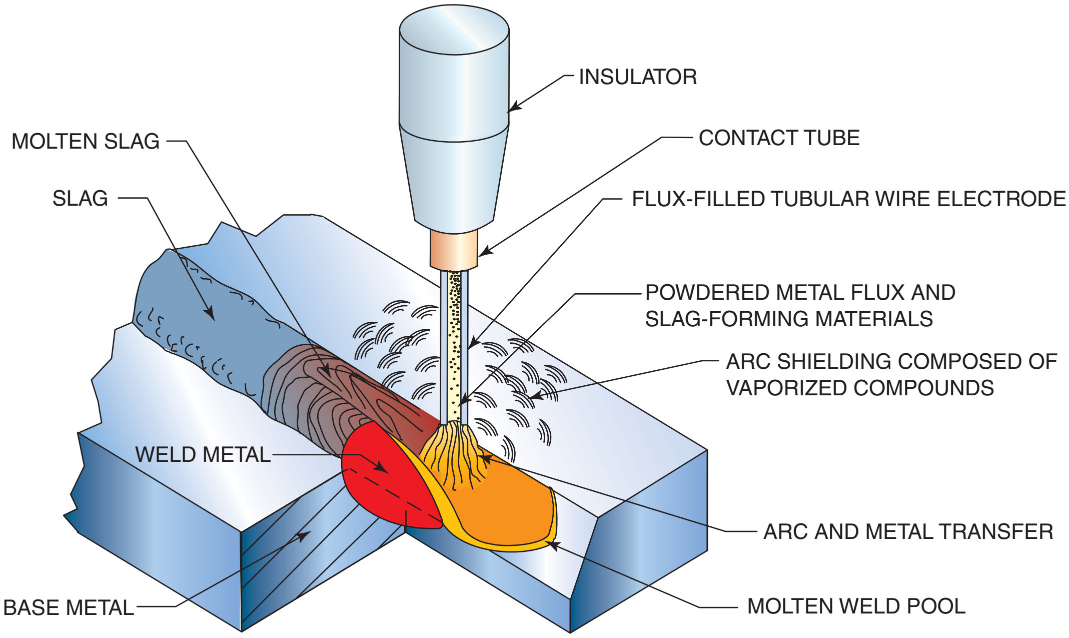
INSULATOR
CONTACT TUBE
MOLTEN SLAG
SLAG
FLUX-FILLED TUBULAR WIRE ELECTRODE
POWDERED METAL FLUX AND SLAG-FORMING MATERIALS
ARC SHIELDING COMPOSED OF VAPORIZED COMPOUNDS
WELD METAL
ARC AND METAL TRANSFER
BASE METAL
MOLTEN WELD POOL

(B) SELF-SHIELDED FLUX CORED ARC WELDING (FCAW-S)

Figure 3.1 Two types of flux cored arc welding
FCA welding may have extra shielding provided by a gas nozzle (A), or be self-shielding only (B).
Source: Courtesy of the American Welding Society

A discussion of weld metal additives and flux elements and their effects on the weld can be found later in this chapter.

The flux core additives that serve as deoxidizers, gas formers, and slag formers either protect the molten weld pool or help to remove impurities from the base metal. Deoxidizers may convert small amounts of surface

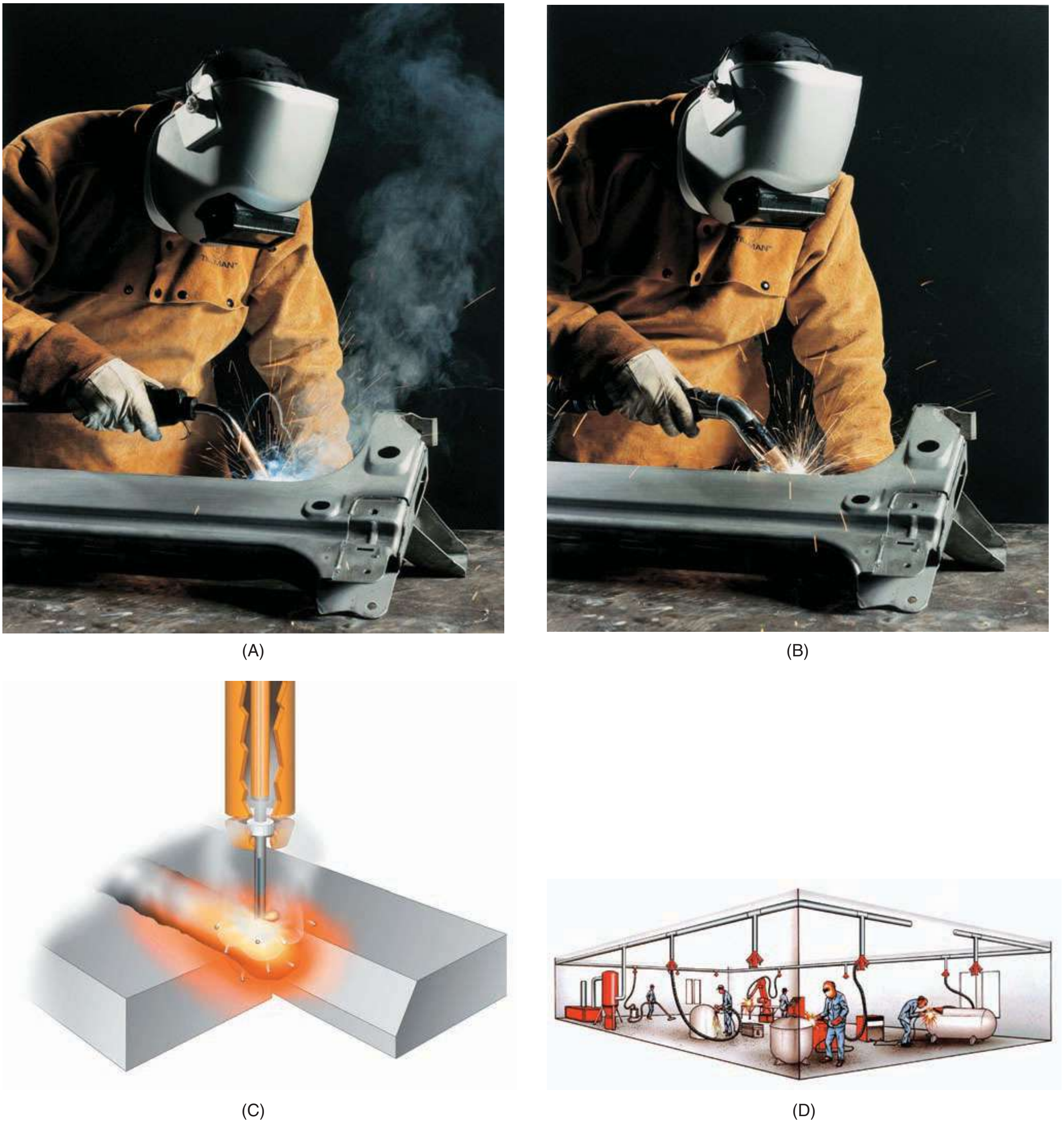

Figure 3.2 **Smoke extraction**
(A) FCA welding without smoke extraction and (B) with smoke extraction. (C) Typical FCAW smoke extraction gun. (D) Typical smoke exhaust system.
Source: Courtesy of Lincoln Electric Company

oxides like mill scale back into pure metal. They work much like the elements used to refine iron ore into steel.

Gas formers rapidly expand and push the surrounding air away from the molten weld pool. If oxygen in the air were to come in contact with the molten weld metal, the weld metal would quickly oxidize. Sometimes this can be seen at the end of a weld when the molten weld metal erupts in a shower of tiny sparks.

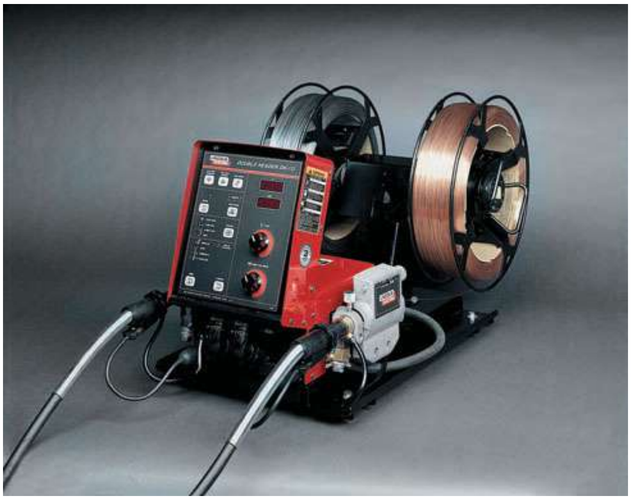

Figure 3.3 **Large-capacity wire-feed unit used with FCAW or GMAW**
Source: Courtesy of Lincoln Electric Company

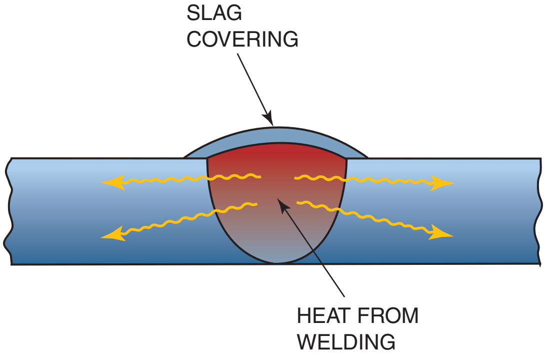

Figure 3.4 **Slag blanketing the weld**
The slag covering keeps the welding heat from escaping quickly, thus slowing the cooling rate.

The slag covering of the weld is useful for several reasons. **Slag** is a nonmetallic product resulting from the mutual dissolution of the flux and nonmetallic impurities in the base metal. Slag helps the weld by protecting the hot metal from the effects of the atmosphere, controlling the bead shape by serving as a dam or mold, and serving as a blanket to slow the weld's cooling rate, which improves its physical properties, Figure 3.4.

EQUIPMENT

Module 6
Key Indicator 1, 2, 3

Power Supply

The FCA welding power supply is the same type that is required for GMAW, called constant-potential, constant-voltage (CP, CV). The words *potential* and *voltage* have the same electrical meaning and are used interchangeably. FCAW machines can be much more powerful than GMAW machines and are available with up to 1500 amperes of welding power.

Guns

FCA welding guns are available as **water-cooled** or **air-cooled**, Figure 3.5. Although most of the FCA welding guns that you will find in schools are air-cooled, our industry often needs water-cooled guns because of the

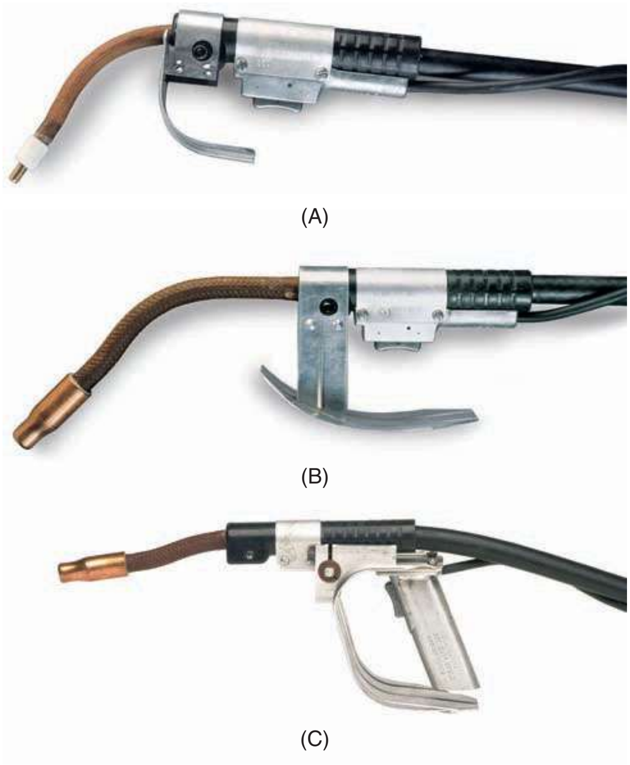
(A)
(B)
(C)

Figure 3.5 **Typical FCA welding guns**
(A) 350 ampere rating self-shielding, (B) 450 ampere rating gas-shielding, and (C) 600 ampere rating gas-shielding.
Source: Courtesy of Lincoln Electric Company

higher heat caused by longer welds made at higher currents. The water-cooled FCA welding gun is more efficient than an air-cooled gun at removing waste heat. The air-cooled gun is more portable because it has fewer hoses, and it may be made lighter so it is easier to manipulate than the water-cooled gun.

Also, the water-cooled gun requires a water reservoir or another system to give the needed cooling. There are two major ways that water can be supplied to the gun for cooling. Cooling water can be supplied directly from the building's water system, or it can be supplied from a recirculation system.

Cooling water supplied directly from the building's water system is usually dumped into a wastewater drain once it has passed through the gun. When this type of system is used, a pressure regulator must be installed to prevent pressures that are too high from damaging the hoses. Water pressures higher than 35 psi (241 kg/mm^2) may cause the water hoses to burst. Check valves must also be installed in the supply line to prevent contaminated water from being drawn back into the water supply. Some cities and states have laws that restrict the use of open systems because of the need for water conservation. Check with your city or state for any restrictions before installing an open water-cooling system.

Recirculating cooling water systems eliminate any of the problems associated with open systems. Chemicals may be added to the water in recirculating systems to prevent freezing, to aid in pump lubrication, and to prevent algae growth. Only manufacturer-approved additives should be used in a recirculation system. Read all of the manufacturer's safety and data sheets before using these chemicals.

Smoke Extraction Nozzles

Because of the large quantity of smoke that can be generated during FCA welding, systems for smoke extraction that fit on the gun have been designed, Figure 3.2B. These systems use a vacuum to pull the smoke back into a specially designed smoke extraction nozzle on the welding gun. The disadvantage of this slightly heavier gun is offset by the system's advantages. The advantages of the system are as follows:

- Cleaner air for the welder to breathe because the smoke is removed before it rises to the welder's face.
- Reduced heating and cooling cost because the smoke is concentrated, so less shop air must be removed with the smoke.

Electrode Feed

Electrode feed systems are similar to those used for GMAW; in fact many feed systems are designed with dual feeders so that solid wire and flux core may be run in sequence. The major difference is that the more robust FCAW feeders are designed to use large-diameter wire and most often have two sets of feed rollers. The two sets of rollers help reduce the drive pressure on the electrode. Excessive pressure can distort the electrode wire diameter, which can allow some flux to be dropped inside the electrode guide tube.

ADVANTAGES

FCA welding offers the welding industry a number of important advantages.

High Deposition Rate

High rates of depositing weld metal are possible. FCA welding deposition rates of more than 25 lb/hr (11 kg/hr) of weld metal are possible. This compares to about 10 lb/hr (5 kg/hr) for shielded metal arc (SMA) welding using a very large-diameter electrode of 1/4 in. (6 mm).

Minimum Electrode Waste

The FCA method makes efficient use of filler metal; from 75% to 90% of the weight of the FCAW electrode is metal, the remainder being flux. SMAW electrodes have a maximum of 75% filler metal; some SMAW electrodes have much less. Also, a stub must be left at the end of each SMA welding electrode. The stub will average 2 in. (51 mm) in length, resulting in a loss of 11% or more of the SMAW filler electrode purchased. FCA welding has no stub loss, so nearly 100% of the FCAW electrode purchased is used.

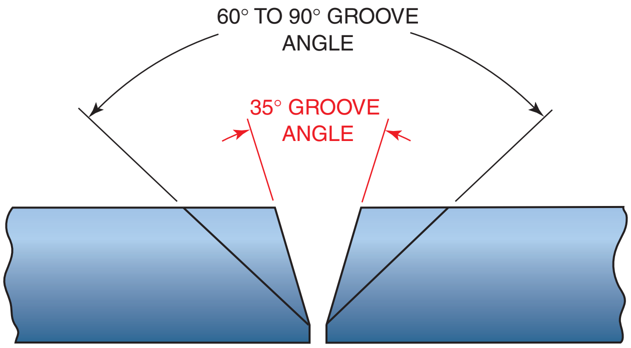

Figure 3.6 **Narrower groove angle for FCAW**
The narrower groove angle for FCAW compared to other welding processes saves on filler metal, welding time, and heat input into the part.

Narrow Groove Angle

Because of the deep penetration characteristic of FCAW, no edge-beveling preparation is required on some joints in metal up to 1/2 in. (13 mm) in thickness. When bevels are cut, the joint-included angle can be reduced to as small as 35°, Figure 3.6. The reduced groove angle results in a smaller-sized weld. This can save 50% of filler metal with about the same savings in time and weld power used.

Minimum Precleaning

The addition of **deoxidizers**, which combine with and remove harmful oxides on the base metal or its surface, and other fluxing agents permits high-quality welds to be made on plates with light surface oxides and mill scale. This eliminates most of the precleaning required before GMA welding can be performed. Often it is possible to make excellent welds on plates in the “as cut” condition; no cleanup is needed.

All-position Welding

Small-diameter electrode sizes in combination with special fluxes allow excellent welds in all positions. The slags produced assist in supporting the weld metal. This process is easy to use, and, when properly adjusted, it is much easier to use than other all-position arc welding processes.

Flexibility

Changes in power settings can permit welding to be done on thin-gauge sheet metals or thicker plates using the same electrode size. Multipass welds allow joining metals with no limit on thickness. This, too, is attainable with one size of electrode.

High Quality

Many codes permit welds to be made using FCAW. The addition of the flux gives the process the high level of reliability needed for welding on boilers, pressure vessels, and structural steel.

Excellent Control

The molten weld pool is more easily controlled with FCAW than with GMAW. The surface appearance is smooth and uniform even with less operator skill. Visibility is improved by removing the nozzle when using self-shielded electrodes.

LIMITATIONS

The main limitation of flux cored arc welding is that it is confined to ferrous metals and nickel-based alloys. Generally, all low- and medium-carbon steels and some low-alloy steels, cast irons, and a limited number of stainless steels are presently weldable using FCAW.

The equipment and electrodes used for the FCAW process are more expensive. However, the cost is quickly recoverable through higher productivity.

The removal of postweld slag requires another production step. The flux must be removed before the weldment is finished (painted) to prevent crevice corrosion.

With the increased welding output comes an increase in smoke and fume generation. The existing ventilation system in a shop might need to be increased to handle the added volume.

ELECTRODES

Methods of Manufacturing

The electrodes have flux tightly packed inside. One method used to make them is to first form a thin sheet of metal into a U-shape, Figure 3.7. A measured quantity of flux is poured into the U-shape before it is squeezed shut. The wire is then passed through a series of dies to size it and further compact the flux.

A second method of manufacturing the electrode is to start with a seamless tube. The tube is usually about 1 in. in diameter. One end of the tube is sealed, and the flux powder is poured into the open end. The tube is vibrated during the filling process to ensure that it fills completely. Once the tube is full, the open end is sealed. The tube is now sized using a series of dies, Figure 3.8.

In both these methods of manufacturing the electrode, the sheet and tube are made up of the desired alloy. Also in both cases, the flux is compacted inside the metal skin. This compacting helps make the electrode operate more smoothly and consistently.

Electrodes are available in sizes from 0.030 in. to 5/32 in. (0.8 mm to 3.9 mm) in diameter. Smaller-diameter electrodes are much more expensive per pound than the same type in a larger diameter due to the high cost of drawing and filling cored wires to small sizes. Larger-diameter electrodes produce such large welds they cannot be controlled in all positions. The most popular diameters range from 0.035 in. to 3/32 in. (0.9 mm to 2.3 mm).

The finished FCA filler metal is packaged in a number of forms for purchase by the end user, Figure 3.9. The AWS has a standard for the size of

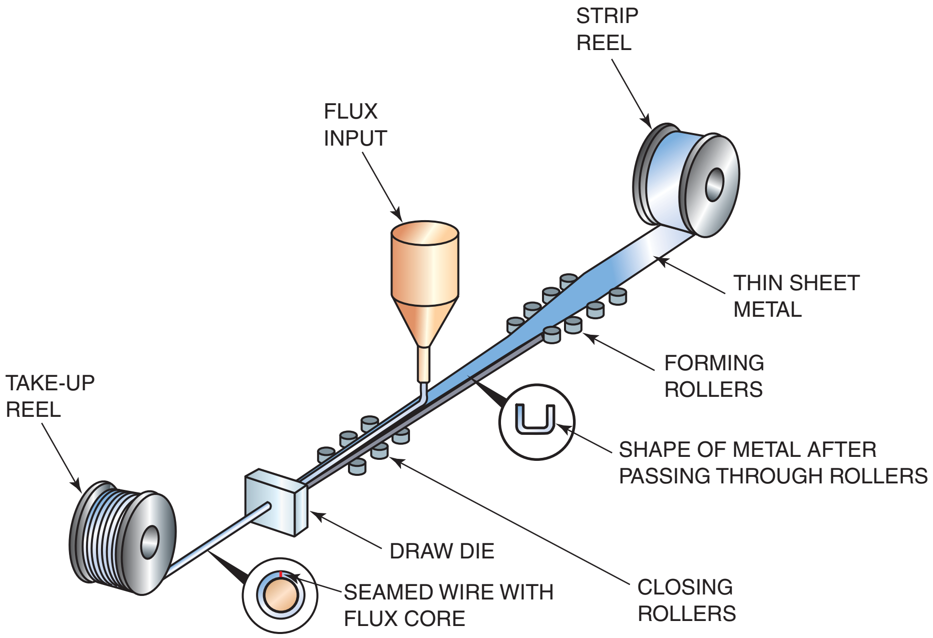
STRIP REEL
FLUX INPUT
THIN SHEET METAL
FORMING ROLLERS
TAKE-UP REEL
SHAPE OF METAL AFTER PASSING THROUGH ROLLERS
DRAW DIE
SEAMED WIRE WITH FLUX CORE
CLOSING ROLLERS

Figure 3.7 **Putting the flux in the flux cored wire**

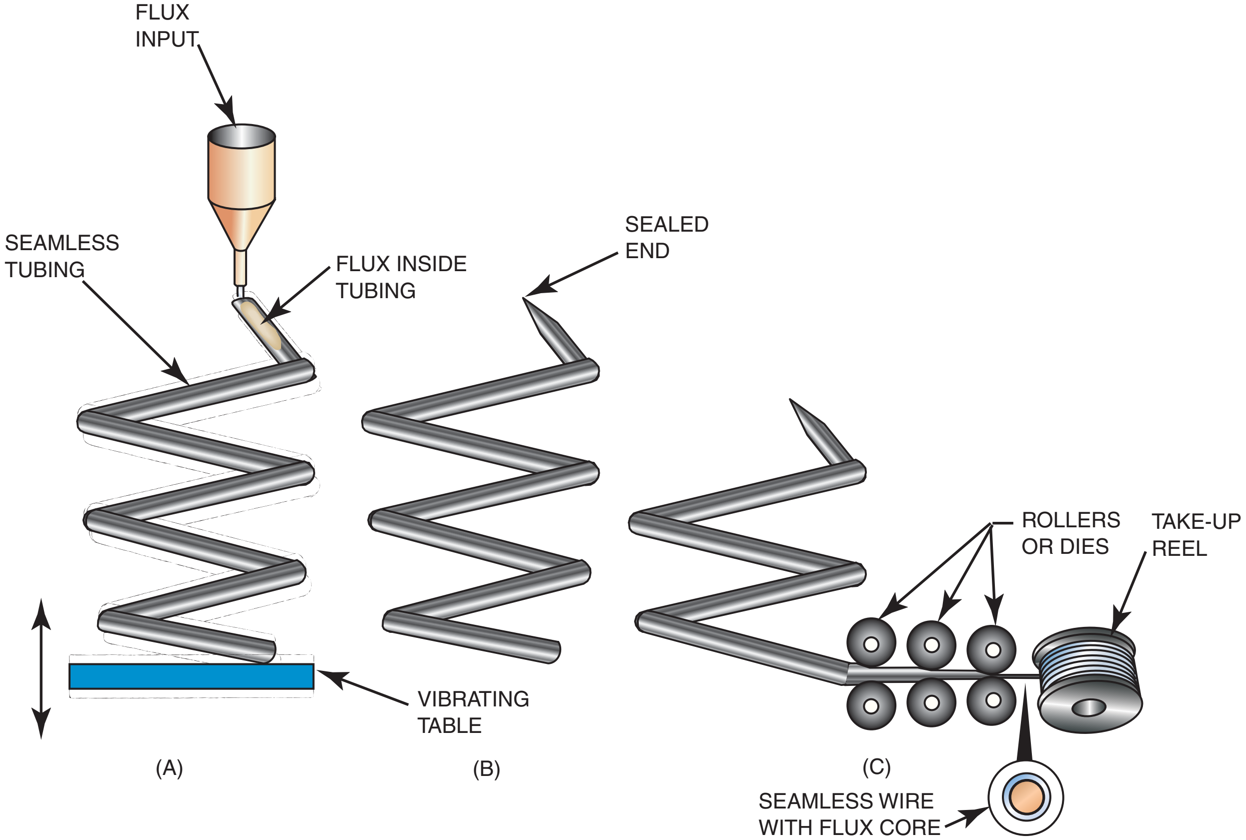
FLUX INPUT
SEAMLESS TUBING
FLUX INSIDE TUBING
SEALED END
VIBRATING TABLE
ROLLERS OR DIES
TAKE-UP REEL
SEAMLESS WIRE WITH FLUX CORE
(A)
(B)
(C)

Figure 3.8 **One method of filling seamless FCA welding filler metal with flux**
The vibration helps compact the granular flux inside the tube.

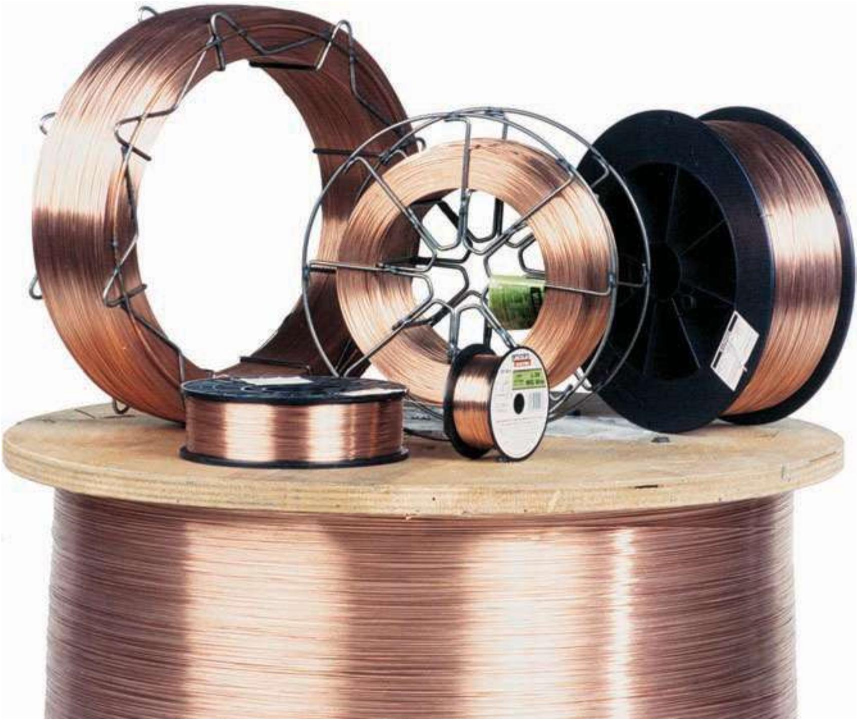

Figure 3.9 **A few packaged forms of FCA filler metal**
FCAW filler metal weights are approximate. They will vary by alloy and manufacturer.
Source: Courtesy of Lincoln Electric Company

each of the package units. Although the dimensions of the packages are standard, the weight of filler wire is not standard. More of the smaller-diameter wire can fit into the same space compared with a larger-diameter wire, so a package of 0.030-in. (0.8-mm) wire weighs more than the same-sized package of 3/32-in. (2.3-mm) wire. The standard packing units for FCAW wires are **spools**, **coils**, reels, and drums, Table 3.1.

Spools are made of plastic or fiberboard and are disposable. They are completely self-contained and are available in approximate weights from

Table 3.1 **Packaging Size Specification for Commonly Used FCA Filler Wire**

Packaging	Outside Diameter	Width	Arbor (Hole) Diameter
Spools	4 in. (102 mm)	1-3/4 in. (44.5 mm)	5/8 in. (16 mm)
	8 in. (203 mm)	2-1/4 in. (57 mm)	2-1/16 in. (52.3 mm)
	12 in. (305 mm)	4 in. (102 mm)	2-1/16 in. (52.3 mm)
	14 in. (356 mm)	4 in. (102 mm)	2-1/16 in. (52.3 mm)
Reels	22 in. (559 mm)	12-1/2 in. (318 mm)	1-5/16 in. (33.3 mm)
	30 in. (762 mm)	16 in. (406 mm)	1-5/16 in. (33.3 mm)
Coils	16-1/4 in. (413 mm)	4 in. (102 mm)	12 in. (305 mm)

	Outside Diameter	Inside Diameter	Height
Drums	23 in. (584 mm)	16 in. (406 mm)	34 in. (864 mm)

1 lb up to around 50 lb (0.5 kg to 25 kg). The smaller spools, 4 in. and 8 in. (102-mm and 203 mm), weighing from 1 lb to 7 lb, are most often used for smaller production runs or for home/hobby use; 12-in. and 14-in. (305-mm and 356-mm) spools are often used in schools and welding fabrication shops.

Coils come wrapped and/or wire tied together. They are unmounted, so they must be supported on a frame on the wire feeder in order to be used. Coils are available in weights around 60 lb (27 kg). Because FCAW wires on coils do not have the expense of a disposable core, these wires cost a little less per pound, so they are more desirable for higher-production shops.

Reels are large wooden spools, and drums are shaped like barrels. Both reels and drums are used for high-production jobs. Both can contain approximately 300 lb to 1000 lb (136 kg to 454 kg) of FCAW wire. Because of their size, they are used primarily at fixed welding stations. Such stations are often associated with some form of automation, such as turntables or robotics.

Electrode Cast and Helix

To see the cast and helix of a wire, feed out 10 ft of wire electrode and cut it off. Lay it on the floor and observe that it forms a circle. The diameter of the circle is known as the cast of the wire, Figure 3.10.

Note that the wire electrode does not lay flat. One end is slightly higher than the other. This height is the helix of the wire.

The AWS has specifications for both cast and helix for all FCA welding wires.

The cast and helix cause the wire to rub on the inside of the contact tube, Figure 3.11. The slight bend in the electrode wire ensures a positive electrical contact between the contact tube and filler wire.

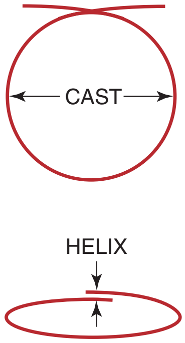

Figure 3.10 Method of measuring cast and helix of FCAW filler wire

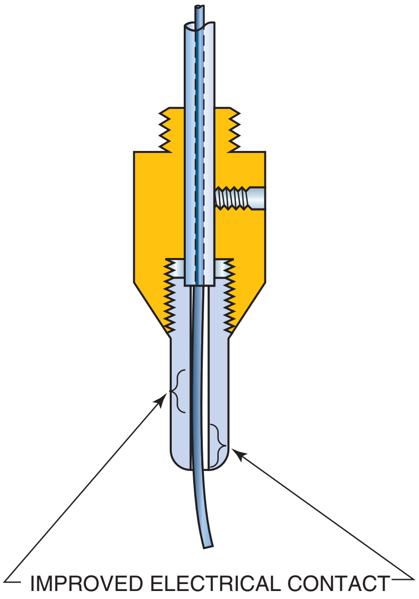

Figure 3.11 Cast forces the wire to make better electrical contact with the tube

FLUX

The fluxes used are mainly based on lime or rutile (a mineral compound consisting of titanium dioxide, usually with a little iron). The purpose of the fluxes is the same as in the shielded metal arc welding (SMAW) process. That is, they can provide all or part of the following to the weld:

- *Deoxidizers:* Oxygen that is present in the welding zone has two forms. It can exist as free oxygen from the atmosphere surrounding the weld. Oxygen can also exist as part of a compound such as an iron oxide or carbon dioxide (CO_2). In either case it can cause porosity in the weld if it is not removed or controlled. Chemicals are added that react to the presence of oxygen in either form and combine to form a harmless compound, Table 3.2. The new compound can become part of the slag that solidifies on top of the weld, or some of it may stay in the weld as very small inclusions. Both methods result in a weld with better mechanical properties because of lower porosity.
- *Slag formers:* Slag serves several vital functions for the weld. It can react with the molten weld metal chemically, and it can affect the weld bead physically. In the molten state it moves through the molten weld pool and acts as a magnet or sponge to chemically combine with impurities in the metal and remove them, Figure 3.12. Slags can be refractory, so that they become solid at a high temperature. As they solidify over the weld, they help it hold its shape and they slow its cooling rate.
- *Fluxing agents:* Molten weld metal tends to have a high surface tension, which prevents it from flowing outward toward the edges of the weld. This causes undercutting along the junction of the

Table 3.2 **Deoxidizing Elements Added to Filler Wire (to Minimize Porosity in the Molten Weld Pool)**

Deoxidizing Element	Strength
Aluminum (Al)	Very strong
Manganese (Mn)	Weak
Silicon (Si)	Weak
Titanium (Ti)	Very strong
Zirconium (Zr)	Very strong

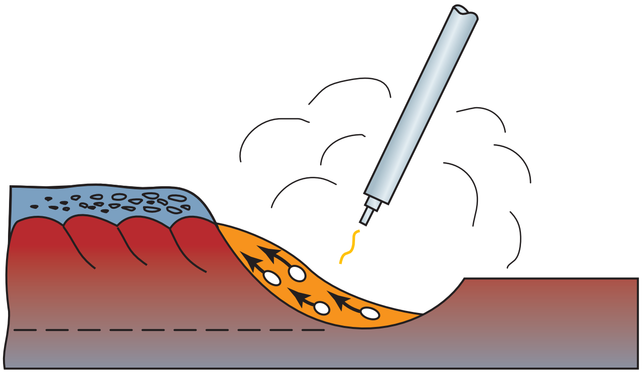

Figure 3.12 **Impurities being floated to the surface by slag**

weld and the base metal. Fluxing agents make the weld more fluid and allow it to flow outward, filling the undercut.

- *Arc stabilizers:* Chemicals in the flux affect the arc resistance. As the resistance is lowered, the arc voltage drops and penetration is reduced. When the arc resistance is increased, the arc voltage increases and weld penetration is increased. Although the resistance within the ionized arc stream may change, the arc is more stable and easier to control. It also improves the metal transfer by reducing spatter caused by an erratic arc.
- *Alloying elements:* Because of the difference in the mechanical properties of metal that is formed by rolling or forging and metal that is melted to form a weld bead, the metallurgical requirements of the two also differ. Some elements change the weld's strength, ductility, hardness, brittleness, toughness, and corrosion resistance. Other alloying elements in the form of powder metal can be added to increase deposition rates.
- *Shielding gas:* As elements in the flux are heated by the arc, some of them vaporize and form voluminous gaseous clouds hundreds of times larger than their original volume. This rapidly expanding cloud forces the air around the weld zone away from the molten weld metal, Figure 3.13. Without the protection this process affords the molten metal, it would rapidly oxidize. Such oxidization would severely affect the weld's mechanical properties, rendering it unfit for service.

All FCAW fluxes are divided into two groups based on the acid or basic chemical reactivity of the slag. The AWS classifies T-1 as acid and T-5 as basic.

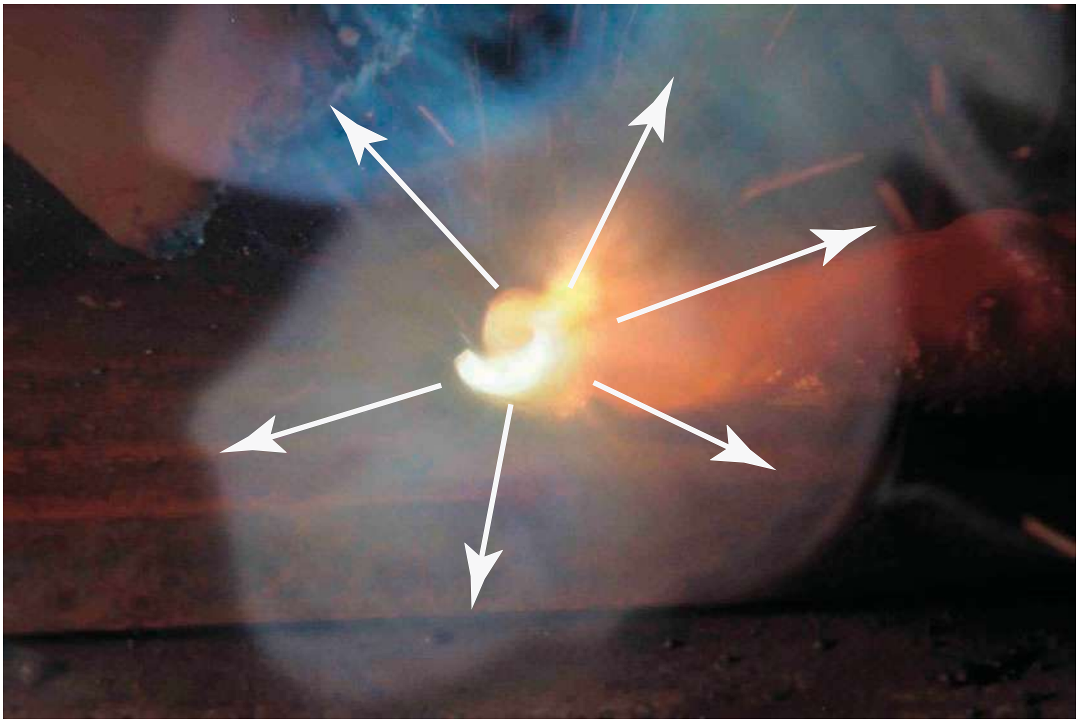

Figure 3.13 **Rapidly expanding gas cloud**
Source: Courtesy of Larry Jeffus

The **rutile-based flux** is acidic, T-1. It produces a smooth, stable arc and a refractory high-temperature slag for out-of-position welding. These electrodes produce a fine drop transfer, a relatively low fume, and an easily removed slag. The main limitation of the rutile fluxes is that their fluxing elements do not produce as high a quality deposit as do the T-5 systems.

The **lime-based flux** is basic, T-5. It is very good at removing certain impurities from the weld metal, but its low-melting-temperature slag is fluid, which makes it generally unsuitable for out-of-position welding. These electrodes produce a more globular transfer, more spatter, more fume, and a more adherent slag than do the T-1 systems. These characteristics are tolerated when it is necessary to deposit very tough weld metal and for welding materials having a low tolerance for hydrogen.

Some rutile-based electrodes allow the addition of a shielding gas. With the weld partially protected by the shielding gas, more elements can be added to the flux, which produces welds with the best of both flux systems, high-quality welds in all positions.

Some fluxes can be used on both single- and multiple-pass welds, and others are limited to single-pass welds only. Using a single-pass welding electrode for multipass welds may result in an excessive amount of manganese. The manganese is necessary to retain strength when making large, single-pass welds. However, with the lower dilution associated with multipass techniques, it can strengthen the weld metal too much and reduce its ductility. In some cases, small welds that deeply penetrate the base metal can help control this problem.

Table 3.3 lists the shielding and polarity for the flux classifications of mild steel FCAW electrodes. The letter *G* is used to indicate an unspecified classification. The *G* means that the electrode has not been classified by the American Welding Society. Often the exact composition of fluxes is kept as a manufacturer's trade secret. Therefore, only limited information about the electrode's composition will be given. The only information often supplied is current, type of shielding required, and some strength characteristics.

Table 3.3 Welding Characteristics of Seven Flux Classifications

Classification	Comments	Shielding Gas
T-1	Requires clean surfaces and produces little spatter. It can be used for single- and multiple-pass welds in all positions.	Carbon dioxide (CO_2) or argon/carbon dioxide mixes
T-2	Requires clean surfaces and produces little spatter. It can be used for single-pass welds in the flat (1G and 1F) and horizontal (2F) positions only.	Carbon dioxide (CO_2)
T-3	Used on thin-gauge steel for single-pass welds in the flat (1G and 1F) and horizontal (2F) positions only.	None
T-4	Low penetration and moderate tendency to crack for single- and multiple-pass welds in the flat (1G and 1F) and horizontal (2F) positions.	None
T-5	Low penetration and a thin, easily removed slag, used for single- and multiple-pass welds in the flat (1G and 1F) position only.	With or without carbon dioxide (CO_2)
T-6	Similar to T-5 without externally applied shielding gas.	None
T-G	The composition and classification of this electrode are not given in the preceding classes. It may be used for single- or multiple-pass welds.	With or without shielding

Table 3.4 Ferrite-forming Elements Used in FCA Welding Fluxes

Element	Reaction in Weld
Silicon (Si)	Ferrite former and deoxidizer
Chromium (Cr)	Ferrite and carbide former
Molybdenum (Mo)	Ferrite and carbide former
Columbium (Cb)	Strong ferrite former
Aluminum (Al)	Ferrite former and deoxidizer

As a result of the relatively rapid cooling of the weld metal, the weld may tend to become hard and brittle. This factor can be controlled by adding elements to the flux that affect the content of both the weld and the slag, Table 3.4. Ferrite is the softer, more ductile form of iron. The addition of ferrite-forming elements can control the hardness and brittleness of a weld. Refractory fluxes are sometimes called “fast-freeze” because they solidify at a higher temperature than the weld metal. By becoming solid first, this slag can cradle the molten weld pool and control its shape. This property is very important for out-of-position welds.

The impurities in the weld pool can be metallic or nonmetallic compounds. Metallic elements that are added to the metal during the manufacturing process in small quantities may be concentrated in the weld. These elements improve the grain structure, strength, hardness, resistance to corrosion, or other mechanical properties in the metal’s as-rolled or formed state. But the deposited weld metal, or weld nugget, is like a small casting because the liquid weld metal freezes in a controlled shape, and some alloys adversely affect the properties of this casting (weld metal). Nonmetallic compounds are primarily slag inclusions left in the metal from the fluxes used during manufacturing. The welding fluxes form slags that are less dense than the weld metal so that they will float to the surface before the weld solidifies.

Flux Cored Steel Electrode Identification

Module 6
Key Indicator 3

The American Welding Society revised its A5.20 *Specification for Carbon Steel Electrodes for Flux Cored Arc Welding* in 1995 to reflect changes in the composition of the FCA filler metals. Table 3.5 lists the AWS specifications for flux cored filler metals.

Mild Steel

The electrode number *E70T-10* is used as an example to explain the classification system for mild steel FCAW electrodes (Figure 3.14):

- *E*—Electrode.
- *7*—Tensile strength in units of 10,000 psi for a good weld. This value is usually either *6* for 60,000 or *7* for 70,000 psi minimum

Table 3.5 Filler Metal Classification Numbers

Metal	AWS Filler Metal Classification
Mild steel	A5.20
Stainless steel	A5.22
Chromium–molybdenum	A5.29

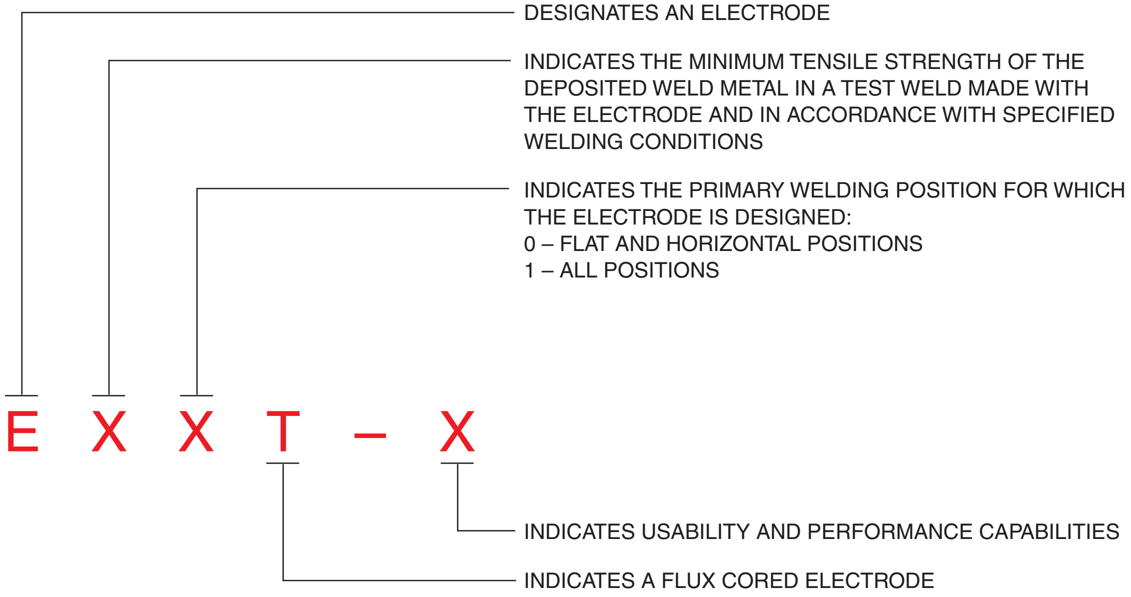
DESIGNATES AN ELECTRODE
INDICATES THE MINIMUM TENSILE STRENGTH OF THE DEPOSITED WELD METAL IN A TEST WELD MADE WITH THE ELECTRODE AND IN ACCORDANCE WITH SPECIFIED WELDING CONDITIONS
INDICATES THE PRIMARY WELDING POSITION FOR WHICH THE ELECTRODE IS DESIGNED:
0 – FLAT AND HORIZONTAL POSITIONS
1 – ALL POSITIONS
E X X T – X
INDICATES USABILITY AND PERFORMANCE CAPABILITIES
INDICATES A FLUX CORED ELECTRODE

Figure 3.14 **Identification system for mild steel FCAW electrodes**
Source: Courtesy of the American Welding Society

weld strength. An exception is for the number *12,* which is used to denote filler metals having a range from 70,000 to 90,000 psi.

- *0—0* is used for flat and horizontal fillets only, and *1* is used for all-position electrodes.
- *T*—Tubular (flux cored) electrode.
- *10*—The number in this position can range from *1* to *14* and is used to indicate the electrode's shielding gas if any, number of passes that may be applied one on top of the other, and other welding characteristics of the electrode. The letter *G* is used to indicate that the shielding gas, polarity, and impact properties are not specified. The letter *G* may or may not be followed by the letter *S. S* indicates an electrode suitable only for single-pass welding.

The electrode classification *E70T-10* can have some optional identifiers added to the end of the number, as in *E70T-10MJH8.* These additions are used to add qualifiers to the general classification so that specific codes or standards can be met. These additions have the following meanings:

- *M*—Mixed gas of 75% to 80% Ar and CO_2 for the balance. If there is no *M,* either the shielding gas is CO_2 or the electrode is self-shielded.
- *J*—Describes the Charpy V-notch impact test value of 20 ft-lb at 40°F.
- *H8*—Describes the residual hydrogen levels in the weld: *H4* equals less than 4 ml/100 g; *H8,* less than 8 ml/100 g; *H16,* less than 16 ml/100 g.

Stainless Steel Electrodes

The AWS classification for stainless steel for FCAW electrodes starts with the letter *E* as its prefix. Following the *E* prefix, the American Iron and Steel Institute's (AISI) three-digit stainless steel number is used. This number indicates the type of stainless steel in the filler metal.

To the right of the AISI number, the AWS adds a dash followed by a suffix number. The number 1 is used to indicate an all-position filler metal, and the number 3 is used to indicate an electrode to be used in the flat and horizontal positions only.

Metal Cored Steel Electrode Identification

The addition of metal powders to the flux core of FCA welding electrodes has produced a new classification of filler metals. The new filler metals evolved over time, and a new identification system was established by the AWS to identify these filler metals. Some of the earlier flux cored filler metals that already had powder metals in their core had their numbers changed to reflect the new designation. The designation was changed from the letter *T* for *tubular* to the letter *C* for *core.* For example, E70T-1 became E70C-3C. The complete explanation of the cored electrode *E70C-3C* follows:

- *E*—Electrode.
- 7—Tensile strength in units of 10,000 psi for a good weld. This value is usually either *6* for 60,000 or *7* for 70,000 psi minimum weld strength. An exception is for the number *12,* which is used to denote filler metals having a range from 70,000 to 90,000 psi.
- *0—0* is used for flat and horizontal fillets only, and *1* is used for all-position electrodes.
- *C*—Metal-cored (tubular) electrode.
- *3—3* is used for a Charpy impact of 20 ft-lb at 0°F, and *6* represents a Charpy impact of 20 ft-lb at 20°F.
- *C*—The second letter *C* indicates CO_2. The letter *M* in this position would indicate a mixed gas, 75% to 80% Ar, with the balance being CO_2. If there is no *M* or *C,* then the shielding gas is CO_2. The letter *G* is used to indicate that the shielding gas, polarity, and impact properties are not specified. The letter *G* may or may not be followed by the letter *S. S* indicates an electrode suitable only for single-pass welding.

NOTE

The powdered metal added to the core flux can provide additional filler metal and/or alloys. This is one way the micro-alloys can be added in very small and controlled amounts, as low as 0.0005% to 0.005%. These are very powerful alloys that dramatically improve the metal's mechanical properties.

Care of Flux Core Electrodes

Wire electrodes may be wrapped in sealed plastic bags for protection from the elements. Others may be wrapped in a special paper, and some are shipped in cans or cardboard boxes.

A small paper bag of a moisture-absorbing material, crystal desiccant, is sometimes placed in the shipping containers to protect wire electrodes from moisture. Some wire electrodes require storage in an electric rod oven to prevent contamination from excessive moisture. Read the manufacturer's recommendations located in or on the electrode shipping container for information on use and storage.

Weather conditions affect your ability to make high-quality welds. Humidity increases the chance of moisture entering the weld zone. Water (H_2O), which consists of two parts hydrogen and one part oxygen, separates in the weld pool. When only one part of hydrogen is expelled, hydrogen entrapment occurs. Hydrogen entrapment can cause weld beads to crack or become brittle. The evaporating moisture will also cause porosity.

To prevent hydrogen entrapment, porosity, and atmospheric contamination, it may be necessary to preheat the base metal to drive out moisture. Storing the wire electrode in a dry location is recommended. The electrode may develop restrictions due to the tangling of the wire or become oxidized with excessive rusting if the wire electrode package is mishandled, thrown, dropped, or stored in a damp location.

Caution

Always keep the wire electrode dry and handle it as you would any important tool or piece of equipment.

SHIELDING GAS

FCA welding wire can be manufactured so that all of the required shielding of the molten weld pool is provided by the vaporization of some of the flux within the tubular electrode. When the electrode provides all of the shielding, it is called **self-shielding** and the welding process is abbreviated FCAW-S (S for self-shielding). Other FCA welding wire must use an externally supplied shielding gas to provide the needed protection of the molten weld pool. When a shielding gas is added, the combined shielding is called **dual shield** and the process is abbreviated FCAW-G (G for gas).

Note: Sometimes the shielding gas(es) are referred to as the shielding *medium*. For example, the shielding gas, or medium, for E71T-5 is either 75% argon with 25% CO_2 or 100% CO_2.

Care must be taken to use the cored electrodes with the recommended gases, and not to use gas at all with the self-shielded electrodes. Using a self-shielding flux cored electrode with a shielding gas may produce a defective weld. The shielding gas will prevent the proper disintegration of much of the deoxidizers. This results in the transfer of these materials across the arc to the weld. In high concentrations, the deoxidizers can produce slags that become trapped in the welds, causing undesirable defects. Lower concentrations may cause brittleness only. In either case, the chance of weld failure is increased. If these electrodes are used correctly, there is no problem.

The selection of a shielding gas will affect the arc and weld properties. The weld bead width, buildup, penetration, spatter, chemical composition, and mechanical properties are all affected as a result of the shielding gas selection.

Shielding gas comes in high-pressure cylinders. These cylinders are supplied with 2000 psi of pressure. Because of this high pressure, it is important that the cylinders be handled and stored safely. For specific cylinder safety instructions see Chapter 2 in Introduction to Welding, the first book in this series.

Gases used for FCA welding include CO_2 and mixtures of argon and CO_2. Argon gas is easily ionized by the arc. Ionization results in a highly concentrated path from the electrode to the weld. This concentration results in a smaller droplet size that is associated with the axial spray mode of metal transfer, Figure 3.15. A smooth, stable arc results and there is a minimum of spatter. This transfer mode continues as CO_2 is added to the argon until the mixture contains more than 25% of CO_2.

As the percentage of CO_2 increases in the argon mixture, weld penetration increases. This increase in penetration continues until a 100% CO_2 shielding gas is reached. But as the percentage of CO_2 is increased the arc stability decreases. The less stable arc causes an increase in spatter.

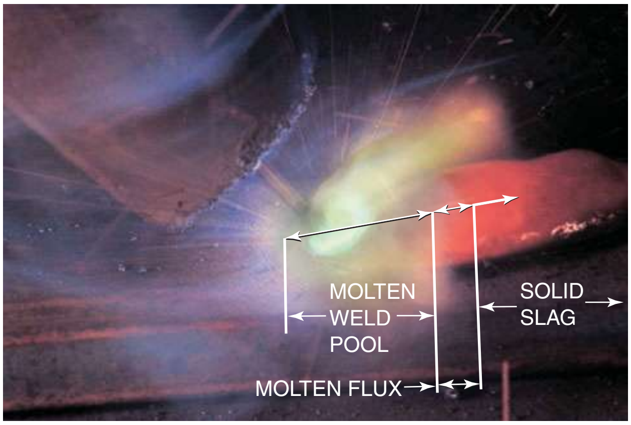

Figure 3.15 **Axial spray transfer mode**
Source: Courtesy of Larry Jeffus

A mixture of 75% argon and 25% CO_2 works best for jobs requiring a mixed gas. This mixture is sometimes called C-25.

Straight CO_2 is used for some welding. But the CO_2 gas molecule is easily broken down in the welding arc. It forms carbon monoxide (CO) and free oxygen (O). Both gases are reactive to some alloys in the electrode. As these alloys travel from the electrode to the molten weld pool, some of them form oxides. Silicon and manganese are the primary alloys that become oxidized and lost from the weld metal.

Most FCA welding electrodes are specifically designed to be used with or without shielding gas and for a specific shielding gas or percentage mixture. For example, an electrode designed specifically for use with 100% CO_2 will have higher levels of silicon and manganese to compensate for the losses to oxidization. But if 100% argon or a mixture of argon and CO_2 is used, the weld will have an excessive amount of silicon and manganese. The weld will not have the desired mechanical or metallurgical properties. Although the weld may look satisfactory, it will probably fail prematurely.

Caution

Never use an FCA welding electrode with a shielding gas it is not designated to be used with. The weld it produces may be unsafe.

WELDING TECHNIQUES

A welder can control weld beads made by FCA welding by making changes in the techniques used. The following explains how changing specific welding techniques will affect the weld produced.

Gun Angle

Gun angle, work angle, and *travel angle* are terms used to refer to the relation of the gun to the work surface, Figure 3.16. The gun angle can be used to control the weld pool. The electric arc produces an electrical force known as the arc force. The arc force can be used to counteract the gravitational pull that tends to make the liquid weld pool sag or run ahead of the arc. By manipulating the electrode travel angle for the flat and horizontal position of welding to a 20° to 45° angle from the vertical, the weld pool can be controlled. A 40° to 50° angle from the vertical plate is recommended for fillet welds.

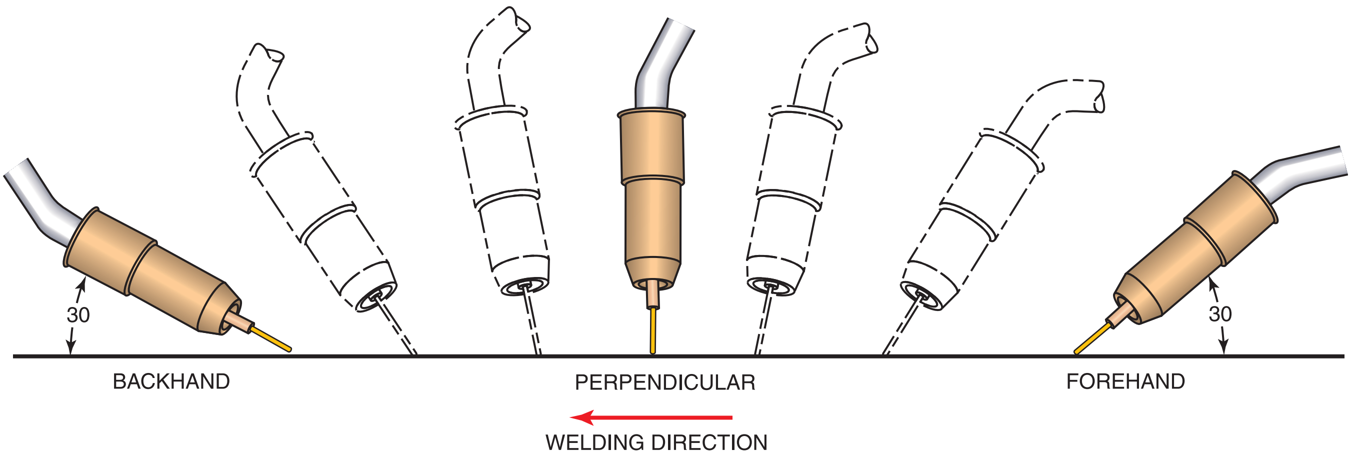

Figure 3.16 **Welding gun angles**

Changes in this angle will affect the weld bead shape and penetration. Shallower angles are needed when welding thinner materials to prevent burn-through. Steeper, perpendicular angles are used for thicker materials.

FCAW electrodes have a flux that is mineral based, often called low-hydrogen. These fluxes are refractory and become solid at a high temperature. If too steep a forehand, or pushing, angle is used, slag from the electrode can be pushed ahead of the weld bead and solidify quickly on the cooler plate, Figure 3.17. Because the slag remains solid at higher temperatures than the temperature of the molten weld pool, it can be trapped under the edges of the weld by the molten weld metal. To avoid this problem, most flat and horizontal welds should be performed with a backhand angle.

Vertical up welds require a forehand gun angle. The forehand angle is needed to direct the arc deep into the groove or joint for better control of the weld pool and deeper penetration, Figure 3.18. Slag entrapment associated with most forehand welding is not a problem for vertical welds.

A gun angle around 90° to the metal surface either slightly forehand or backhand works best for overhead welds, Figure 3.19. The slight angle aids with visibility of the weld, and it helps control spatter buildup in the gas nozzle.

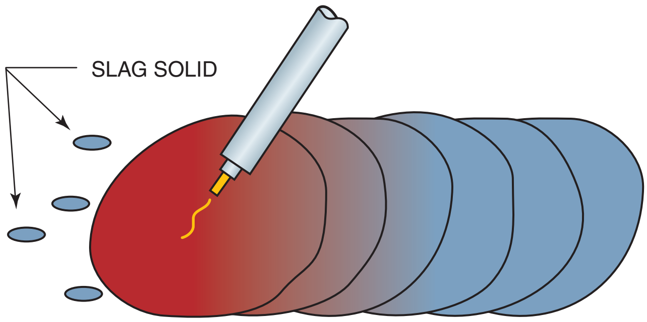

Figure 3.17 **Problem of trapped slag**
Large quantities of solid slag in front of a weld can cause slag to be trapped under the weld bead.

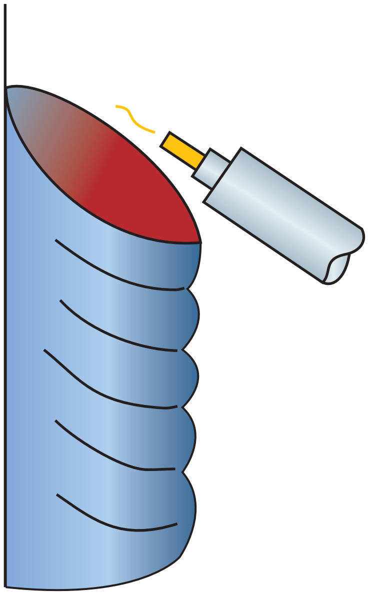

Figure 3.18 **Vertical up gun angle**

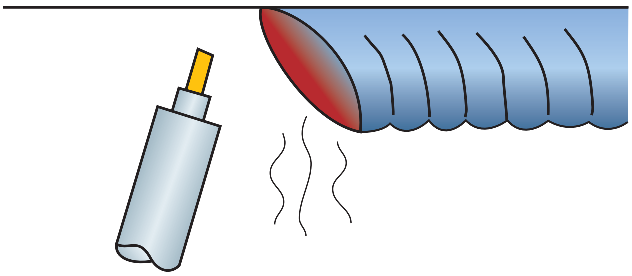

Figure 3.19 **Weld gun position to control spatter buildup on an overhead weld**

Forehand/Perpendicular/Backhand Techniques

Forehand, perpendicular, and *backhand* are the terms most often used to describe the gun angle as it relates to the work and the direction of travel. The forehand technique is sometimes referred to as *pushing* the weld bead, and backhand may be referred to as *pulling* or *dragging* the weld bead. The term *perpendicular* is used when the gun angle is at approximately 90° to the work surface, Figure 3.20.

Advantages of the Forehand Technique

The forehand welding technique has several advantages:

- Joint visibility—You can easily see the joint where the bead will be deposited, Figure 3.21.
- Electrode extension—The contact tube tip is easier to see, making it easier to maintain a constant extension length.
- Less weld penetration—It is easier to weld on thin sheet metal without melting through.
- Out-of-position welds—This technique works well on vertical up and overhead joints for better control of the weld pool.

Disadvantages of the Forehand Technique

The disadvantages of using the forehand welding technique are the following:

- Weld thickness—Thinner welds may occur because less weld reinforcement is applied to the weld joint.

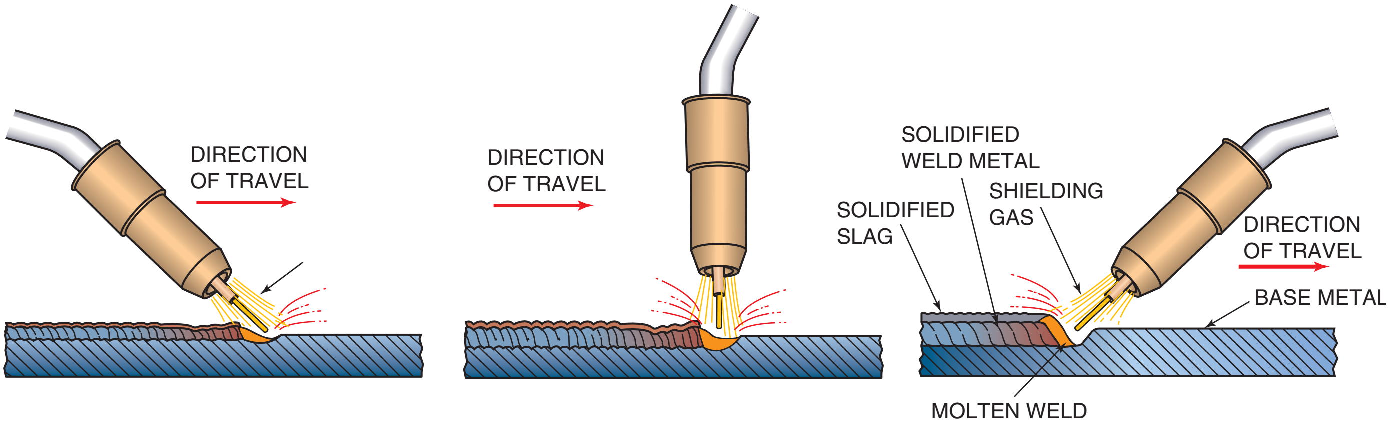

Figure 3.20 **Gun angles**

Changing the welding gun angle between forehand, perpendicular, and backhand angles will change the shape of the weld bead produced.

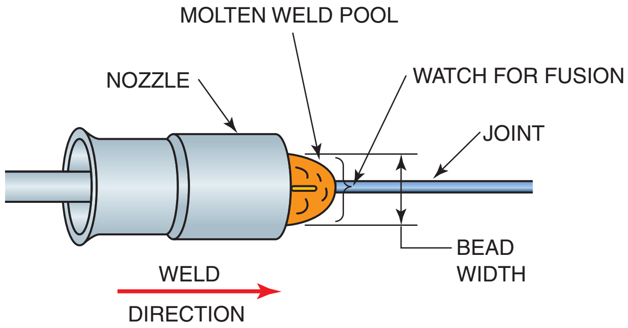

Figure 3.21 Welder's view with a forehand angle
This angle keeps the shielding gas nozzle from restricting the welder's view.

- Welding speed—Because less weld metal is being applied, the rate of travel along the joint can be faster, which may make it harder to create a uniform weld.
- Slag inclusions—Some spattered slag can be thrown in front of the weld bead and be trapped or included in the weld, resulting in a weld defect.
- Spatter—Depending on the electrode, the amount of spatter may be slightly increased with the forehand technique.

Advantages of the Perpendicular Technique

The perpendicular welding technique has the following advantages:

- Machine and robotic welding—The perpendicular gun angle is used on automated welding because there is no need to change the gun angle when the weld changes direction.
- Uniform bead shape—The weld's penetration and reinforcement are balanced between those of forehand and backhand techniques.

Disadvantages of the Perpendicular Technique

The disadvantages of using the perpendicular welding technique are the following:

- Limited visibility—Because the welding gun is directly over the weld, there is limited visibility of the weld unless you lean your head way over to the side.
- Weld spatter—Because the weld nozzle is directly under the weld in the overhead position, more weld spatter can collect in the nozzle, causing gas flow problems or even shorting the tip to the nozzle.

Advantages of the Backhand Technique

The backhand welding technique has the following advantages:

- Weld bead visibility—It is easy to see the back of the molten weld pool as you are welding, which makes it easier to control the bead shape, Figure 3.22.
- Travel speed—Because of the larger amount of weld metal being applied, the rate of travel may be slower, making it easier to create a uniform weld.
- Depth of fusion—The arc force and the greater heat from the slower travel rate both increase the depth of weld joint penetration.

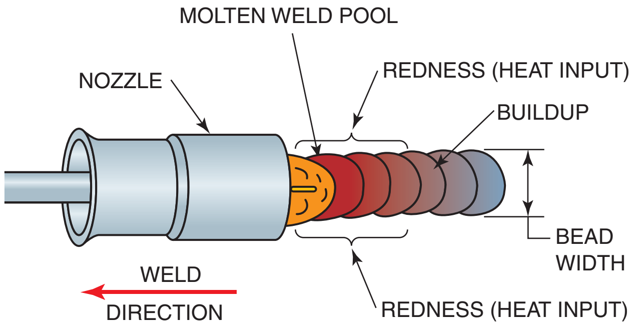

Figure 3.22 **Welder's view with a backhand angle**
Watch the trailing edge of the molten weld pool.

Disadvantages of the Backhand Technique

The disadvantages of the backhand welding technique are the following:

- Weld buildup—The weld bead may have a convex (raised or rounded) weld face when you use the backhand technique.
- Postweld finishing—Because of the weld bead shape, more work may be required if the product has to be finished by grinding smooth.
- Joint following—It is harder to follow the joint because your hand and the FCAW gun are positioned over the joint, and you may wander from the seam.
- Loss of penetration—An inexperienced welder sometimes directs the wire too far back into the weld pool causing the wire to build up in the face of the weld pool reducing joint penetration.

Travel Speed

The American Welding Society defines travel speed as the linear rate at which the arc is moved along the weld joint. Fast travel speeds deposit less filler metal. If the rate of travel increases, the filler metal cannot be deposited fast enough to adequately fill the path melted by the arc. This causes the weld bead to have a groove melted into the base metal next to the weld and left unfilled by the weld. This condition is known as undercut.

Undercut occurs along the edges or toes of the weld bead. Slower travel speeds will, at first, increase penetration and increase the filler weld metal deposited. As the filler metal increases, the weld bead will build up in the weld pool. Because of the deep penetration of flux cored wire, the angle at which you hold the gun is very important for a successful weld.

If all welding conditions are correct and remain constant, the preferred rate of travel for maximum weld penetration is a travel speed that allows you to stay within the selected welding variables and still control the fluidity of the weld pool. This is an intermediate travel speed, or progression, which is not too fast or too slow.

Another way to figure out correct travel speed is to consult the manufacturer's recommendations chart for the inches per minute (ipm) burn-off rate for the selected electrode.

Mode of Metal Transfer

The mode of metal transfer is used to describe how the molten weld metal is transferred across the arc to the base metal. The mode of metal transfer that is selected, the shape of the completed weld bead, and the depth of weld penetration depend on the welding power source, wire electrode size, type and thickness of material, type of shielding gas used, and best welding position for the task.

Spray Transfer with FCAW-G

The spray transfer mode is the most common process used with gas-shielded FCAW (FCAW-G), Figure 3.15.

As the gun trigger is depressed, the shielding gas automatically flows and the electrode bridges the distance from the contact tube to the base metal, making contact with the base metal to complete a circuit. The electrode shorts and becomes so hot that the base metal melts and forms a weld pool. The electrode melts into the weld pool and burns back toward the contact tube. A combination of high amperage and the shielding gas along with the electrode size produces a pinching effect on the molten electrode wire, causing the end of the electrode wire to spray across the arc.

The characteristic of spray-type transfer is a smooth arc, through which hundreds of small droplets per second are transferred through the arc from the electrode to the weld pool. At that moment a transfer of metal is taking place. Spray transfer can produce a high quantity of metal droplets, up to approximately 250 per second above the transition current, or critical current. This means the current required for a spray transfer to take place is dependent on the electrode size, composition of the electrode, and shielding gas. Below the transition current (critical current), globular transfer takes place.

In order to achieve a spray transfer, high current and larger-diameter electrode wire are needed. A shielding gas of carbon dioxide (CO_2), a mixture of carbon dioxide (CO_2) and argon (Ar), or an argon (Ar) and oxygen (O_2) mixture is needed. FCAW-G is a welding process that, with the correct variables, can be used

- on thin or properly prepared thick sections of material
- on a combination of thick to thin materials
- with small or large electrode diameters
- with a combination of shielding gases

Globular Transfer with FCAW-G

Globular transfer occurs when the welding current is below the transition current, Figure 3.23. The electrode forms a molten ball at its end that grows in size to approximately two to three times the original electrode diameter. These large molten balls are then transferred across the arc at the rate of several drops per second.

The arc becomes unstable because of the gravitational pull from the weight of these large drops. A spinning effect caused by a natural phenomenon takes place when argon gas is introduced to a large ball of molten metal on the electrode. The molten ball spins as it transfers across the arc to the base metal. This unstable globular transfer can produce excessive spatter.

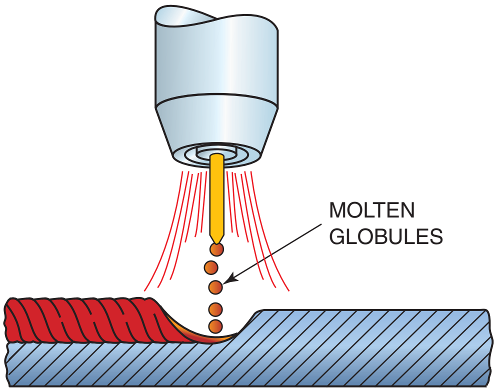

Figure 3.23 Globular transfer method

Both FCAW-S and FCAW-G use direct current electrode negative (DCEN) when welding on thin-gauge materials to keep the heat in the base metal and the small-diameter electrode at a controllable burn-off rate. The electrode can then be stabilized, and it is easier to manipulate and control the weld pool in all weld positions. Larger-diameter electrodes are welded with direct current electrode positive (DCEP) because the larger diameters can keep up with the burn-off rates.

The recommended weld position means the position in which the workpiece is placed for welding. All welding positions use either spray or globular transfer, but for now we will concentrate on the flat and horizontal welding positions.

In the flat welding position the workpiece is placed flat on the work surface. In the horizontal welding position the workpiece is positioned perpendicular to the workbench surface.

The amperage range may be from 30 to 400 amperes or more for welding materials from gauge thickness up to 1-1/2 inches. On square groove weld joints, thicker base metals can be welded with little or no edge preparation. This is one of the great advantages of FCAW. If edges are prepared and cut at an angle (beveled) to accept a complete joint weld penetration, the depth of penetration will be greatly increased. FCAW is commonly used for general repairs to mild steel in the horizontal, vertical, and overhead welding positions, sometimes referred to as out-of-position welding.

Electrode Extension

The electrode extension is measured from the end of the electrode contact tube to the point the arc begins at the end of the electrode, Figure 3.24. Compared to GMA welding, the electrode extension required for FCAW is much greater. The longer extension is required for several reasons. The electrical resistance of the wire causes the wire to heat up, which can drive out moisture from the flux. This preheating of the wire also results in a smoother arc with less spatter.

Porosity

FCA welding can produce high-quality welds in all positions, although porosity in the weld can be a persistent problem. Porosity can be caused by moisture in the flux, improper gun manipulation, or surface contamination.

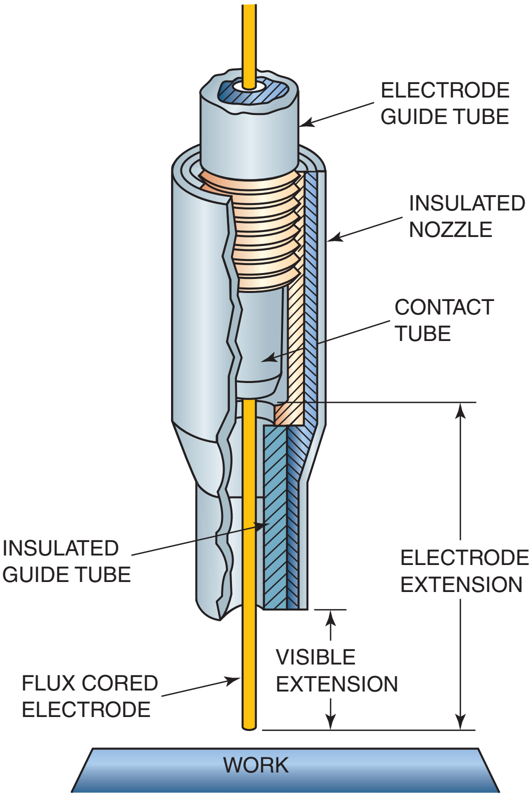

Figure 3.24 **Self-shielded electrode nozzle**
Source: Courtesy of the American Welding Society

The flux used in the FCA welding electrode is subject to picking up moisture from the surrounding atmosphere, so the electrodes must be stored in a dry area. Once the flux becomes contaminated with moisture, it is very difficult to remove. Water (H_2O) breaks down into free hydrogen and oxygen in the presence of an arc, Figure 3.25. The hydrogen can be absorbed into the molten weld metal, where it can cause postweld cracking. The oxygen is absorbed into the weld metal also, but it forms oxides in the metal.

If a shielding gas is used, the FCA welding gun gas nozzle must be close enough to the weld to provide adequate shielding gas coverage. If there is a wind or if the nozzle-to-work distance is excessive, the shielding will be inadequate and allow weld porosity. If welding is to be done outside or in an area subject to drafts, the gas flow rate must be increased or a wind shield must be placed to protect the weld, Figure 3.26.

A common misconception is that the flux within the electrode will either remove or control weld quality problems caused by surface contaminations. That is not true. The addition of flux makes FCA welding more tolerant to surface conditions than GMA welding, although it still is adversely affected by such contaminations.

New hot-rolled steel has a layer of dark gray or black iron oxide called mill scale. Although this layer is very thin, it may provide a source of enough oxygen to cause porosity in the weld. If mill scale causes porosity, it is usually uniformly scattered through the weld, Figure 3.27. Unless it is severe, uniformly scattered porosity is usually not visible in the finished weld. It is trapped under the surface as the weld cools.

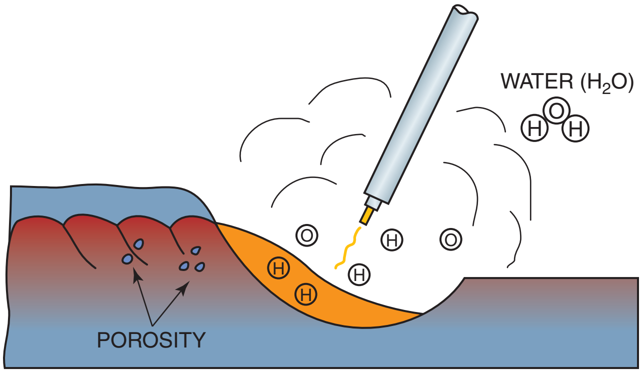
WATER (H_2O)
O
H
H
O
H
O
H
H
H
POROSITY

Figure 3.25 **Water and porosity**
Water (H_2O) breaks down in the presence of the arc and the hydrogen (H) is dissolved in the molten weld metal.

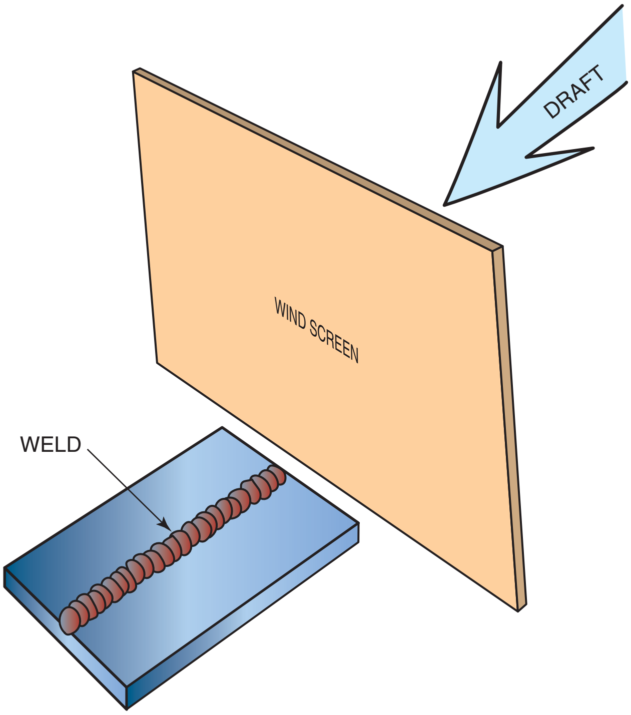
DRAFT
WIND SCREEN
WELD

Figure 3.26 **Wind and draft protection**
A wind screen can keep the welding shielding from being blown away.

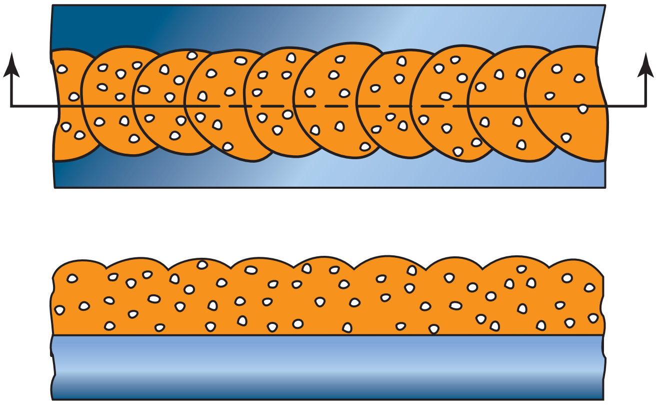

Figure 3.27 **Uniformly scattered porosity**

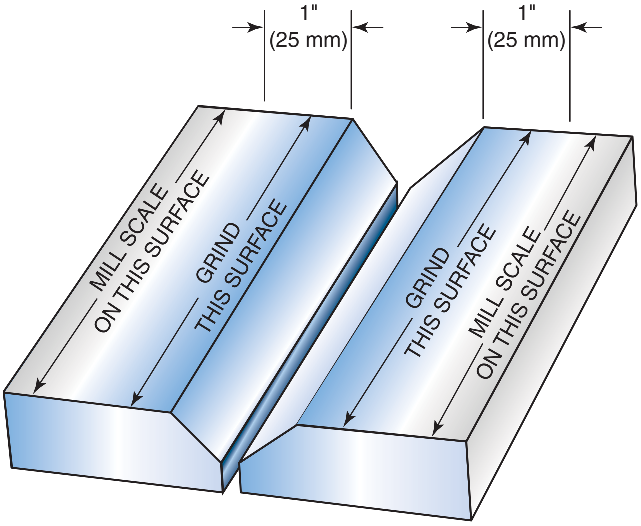

Figure 3.28 Grinding requirements
Grind mill scale off plates within 1 in. (25 mm) of the groove.

Because porosity is under the weld surface, nondestructive testing methods, including X ray, magnetic particle, and ultrasound, must be used to locate it in a weld. It can be detected by mechanical testing such as guided bend, free bend, and nick-break testing for establishing weld parameters. Often it is better to remove the mill scale before welding rather than risking the production of porosity.

All welding surfaces within the weld groove and the surrounding surfaces within 1 in. (25 mm) must be cleaned to bright metal, Figure 3.28. Cleaning may be either grinding, filing, sanding, or blasting.

Any time FCA welds are to be made on metals that are dirty, oily, rusty, or wet or that have been painted, the surface must be precleaned. Cleaning can be done chemically or mechanically.

One advantage of chemically cleaning oil and paint is that it is easier to clean larger areas. Both oil and paint smoke easily when heated, and such smoke can cause weld defects. They must be removed far enough from the weld so that weld heat does not cause them to smoke. In the case of small parts the entire part may need to be cleaned.

Caution

Chemically cleaning oil and paint off metal must be done according to the cleaner manufacturer's directions. The work must be done in an appropriate, approved area. The metal must be dry, and all residues of the cleaner must be removed before welding begins.

SUMMARY

Flux cored arc welding is used to produce more tons of welded fabrications than any other process. The ability to produce high-quality welds on a wide variety of material thicknesses and joint configurations has led to its popularity. As you learn and develop these skills, you will therefore be significantly increasing your employability and productivity in the welding industry.

A wide variety of filler metals and shielding gas combinations for flux cored arc welding are available to you in industry. These various materials aid in producing welds of high quality under various welding conditions. Although the selection of the proper filler metal and gas coverage, if used, will significantly affect the finished weld's quality in the field, there are very few differences in manipulation and setup among these filler metals. Therefore, as you practice welding in a school or training program and learn to use a specific wire and shielding gas mixture, these skills are easily transferable to the next group of materials you will encounter on the job.

REVIEW

1. List some factors that have led to the increased use of FCA welding.
2. How is FCAW similar to GMAW?
3. What does the FCA flux provide to the weld?
4. What are the major atmospheric contaminations of the molten weld metal?
5. How does slag help an FCA weld?
6. What is the electrical difference between a constant-potential and a constant-current power supply?
7. How can FCA welding guns be cooled?
8. What problems does excessive drive roller pressure cause?
9. List the advantages that FCA welding offers the welding industry.
10. Describe the two methods of manufacturing FCA electrode wire.
11. Why are large-diameter electrodes not used for all-position welding?
12. How do deoxidizers remove oxygen from the weld zone?
13. What do fluxing agents do for a weld?
14. Why are alloying elements added to the flux?
15. How does the flux form a shielding gas to protect the weld?
16. What are the main limitations of the rutile fluxes?
17. Why is it more difficult to use lime-based fluxed electrodes on out-of-position welds?
18. What benefit does adding an externally supplied shielding gas have on some rutile-based electrodes?
19. How do excessive amounts of manganese affect a weld?
20. Why are elements added that cause ferrite to form in the weld?
21. Why are some slags called refractory?
22. Why must a flux form a less dense slag?
23. Referring to Table 3.5, what is the AWS classification for FCA welding electrodes for stainless steel?
24. Describe the meaning of each part of the following FCA welding electrode identification: E81T-5.
25. What does the number 316 in E316T-1 mean?
26. What is the advantage of using an argon-CO_2 mixed shielding gas?
27. What are the primary alloying elements lost if 100% CO_2 shielding gas is used?
28. What can cause porosity in an FCA weld?
29. What happens to water in the welding arc?
30. What is the thin, dark gray or black layer on new hot-rolled steel? How can it affect the weld?
31. Why is uniformly scattered porosity hard to detect in a weld?
32. What cautions must be taken when chemically cleaning oil or paint from a piece of metal?
33. What can happen to slag that solidifies on the plate ahead of the weld?
34. How is the electrode extension measured?

CHAPTER 4

Flux Cored Arc Welding

OBJECTIVES

After completing this chapter, the student should be able to

- set up the FCA weld station
- thread the electrode wire through the system
- list three disadvantages of having to bevel a plate before welding
- make root, filler, and cover passes with the FCAW process
- make butt welds in all positions that can pass a specified standard's visual or destructive examination criteria
- make fillet welds in tee joints and lap joints in all positions that can pass a specified standard's visual or destructive examination criteria

KEY TERMS

amperage range
conduit liner
contact tube
critical weld
feed rollers
lap joint
root face
stringer bead
tee joint
voltage range
weave bead
wire-feed speed

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter:

Module 1: Occupational Orientation

Key Indicator 1: Prepares time or job cards, reports, or records
Key Indicator 2: Performs housekeeping duties
Key Indicator 3: Follows verbal instructions to complete work assignments
Key Indicator 4: Follows written instructions to complete work assignments

Module 6: Flux Cored Arc Welding (FCAW-G/GM, FCAW-S)

Key Indicator 1: Performs safety inspections of FCAW equipment and accessories
Key Indicator 2: Makes minor external repairs to FCAW equipment and accessories
Gas Shielded
Key Indicator 3: Sets up for FCAW-G/GM operations on carbon steel
Key Indicator 4: Operates FCAW-G/GM equipment on carbon steel
Key Indicator 5: Makes FCAW-G/GM fillet welds, in all positions, on carbon steel

Key Indicator 6: Makes FCAW-G/GM groove welds, in all positions, on carbon steel

Key Indicator 7: Passes FCAW-G/GM welder performance qualification testing (workmanship sample) on carbon steel

Self Shielded

Key Indicator 8: Sets up for FCAW-S operations on carbon steel

Key Indicator 9: Operates FCAW-S equipment on carbon steel

Key Indicator 10: Makes FCAW-S fillet welds, in all positions, on carbon steel

Key Indicator 11: Makes FCAW-S groove welds, in all positions, on carbon steel

Key Indicator 12: Passes FCAW-S welder performance qualification test (workmanship sample) on carbon steel

Module 9: Welding Inspection and Testing Principles

Key Indicator 1: Examines cut surfaces and edges of prepared base metal parts

Key Indicator 2: Examines tacks, root passes, intermediate layers, and completed welds

INTRODUCTION

Setup of the flux cored arc welding (FCAW) work station is the key to making quality welds. It may be possible, using a poorly set up FCA welder, to make an acceptable weld in the flat position. The FCA welding process is often forgiving; thus welds can often be made even when the welder is not set correctly. However, such welds will have major defects such as excessive spatter, undercut, overlap, porosity, slag inclusions, and poor weld bead contours. Setup becomes even more important for out-of-position welds. Making vertical and overhead welds can be difficult for a student welder with a properly set up system, but it becomes impossible with a system that is out of adjustment.

Learning to set up and properly adjust the FCA welding system will allow you to produce high-quality welds at a high level of productivity.

FCAW is set up and manipulated in a manner similar to that of GMAW. The results of changes in electrode extension, voltage, amperage, and torch angle are essentially the same.

Although every manufacturer's FCA welding equipment is designed differently, all equipment is set up in a similar manner. It is always best to follow the specific manufacturer's recommendations regarding setup as provided in its equipment literature. You will find, however, that, in the field, manufacturers' literature is not always available for the equipment you are asked to use. It is therefore important to have a good general knowledge and understanding of the setup procedure for FCA welding equipment. Figure 4.1 shows all of the various components that make up an FCA welding station.

Caution

FCA welding produces a lot of ultraviolet light, heat, sparks, slag, and welding fumes. Proper personal, protective clothing and special protective clothing must be worn to prevent burns from the ultraviolet light and hot weld metal. Eye protection must be worn to prevent injury from flying sparks and slag. Forced ventilation and possibly a respirator must be used to prevent fume-related injuries. Refer to the safety precautions provided by the equipment and electrode manufacturers and to Chapter 2 in *Welding Skills, Processes and Practices for Entry-Level Welders: Book One* for additional safety help.

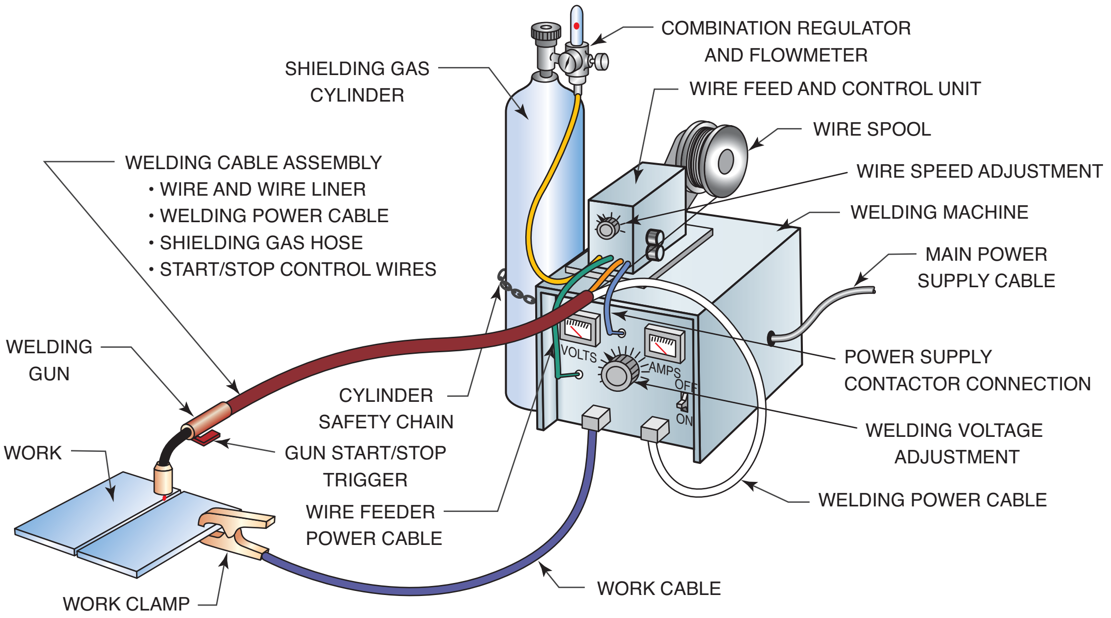
COMBINATION REGULATOR AND FLOWMETER
SHIELDING GAS CYLINDER
WIRE FEED AND CONTROL UNIT
WIRE SPOOL
WELDING CABLE ASSEMBLY
• WIRE AND WIRE LINER
• WELDING POWER CABLE
• SHIELDING GAS HOSE
• START/STOP CONTROL WIRES
WIRE SPEED ADJUSTMENT
WELDING MACHINE
MAIN POWER SUPPLY CABLE
WELDING GUN
VOLTS
AMPS
OFF
ON
POWER SUPPLY CONTACTOR CONNECTION
CYLINDER SAFETY CHAIN
WORK
GUN START/STOP TRIGGER
WELDING VOLTAGE ADJUSTMENT
WIRE FEEDER POWER CABLE
WELDING POWER CABLE
WORK CABLE
WORK CLAMP

Figure 4.1 **Basic FCA welding equipment identification**

PRACTICES

The practices in this chapter are grouped according to those requiring similar techniques and setups. Plate welds are covered first, then sheet metal. The practices start with 1/4-in. (6-mm) mild steel plates; they are used because they require the least preparation times. The thicker 3/8-in. (9.5-mm) plates provide the basics of practicing groove welding. The 3/4-in. (19-mm) and thicker plates are used to develop the skills required to pass the unlimited thickness test often given to FCA welders. Sheet metal is grouped together because it presents a unique set of learning skills.

The major skill required for making consistently acceptable FCA welds is the ability to set up the welding system. Changes such as variations in material thickness, position, and type of joint require changes both in technique and setup. A correctly set up FCA welding station can, in many cases, be operated by a less-skilled welder. Often the only difference between a welder earning a minimum wage and one earning the maximum wage is the ability to correct machine setups.

For several reasons the FCA welding practice plates will be larger than most other practice plates. Welding heat and welding speed are the major factors that necessitate this increased size. FCA welding is both high energy and fast, and the welding energy (heat) input is so great that small practice plates may glow red by the end of a single weld pass. This would seriously affect the weld quality. To prevent this from happening, wider plates are used. Because of the higher welding speeds, longer plates are usually used.

Plates less than 1/2 in. (13 mm) will be 12 in. (305 mm) long for most practices. In addition to controlling the heat buildup, the longer plates are needed to give the welder enough time to practice welding. Learning to make longer welds is a skill that must also be practiced, because the FCA welding process is used in industry to make long production welds.

Plates thicker than 1/2 in. (13 mm) can be shorter than 12 in. (305 mm). Most codes allow test plates of "unlimited thickness" to be as short as 7 in. (178 mm).

PRACTICE 4-1

FCAW Equipment Setup

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Key Indicator 2
Gas Shielded
Key Indicator 3
Self Shielded
Key Indicator 8

For this practice, you will need a semiautomatic welding power source approved for FCA welding, welding gun, electrode feed unit, electrode supply, shielding gas supply, shielding gas flowmeter regulator, electrode conduit, power and work leads, shielding gas hoses (if required), assorted hand tools, spare parts, and any other required materials. In this practice, you will demonstrate to a group of students and your instructor how to properly set up an FCA welding station. Some manufacturers include detailed setup instructions with their equipment. If such instructions are available for your equipment, follow them. Otherwise, use the following instructions.

If the shielding gas is to be used and it comes from a cylinder, the cylinder must be chained securely in place before the valve protection cap is removed, Figure 4.2. Standing to one side of the cylinder, make sure nobody is in line with the valve and quickly crack the valve to blow out any dirt in the valve before the flowmeter regulator is attached, Figure 4.3. Attach the correct hose from the regulator to the "gas-in" connection on the electrode feed unit or machine.

Install the reel of electrode (welding wire) on the holder and secure it, Figure 4.4. Check the feed roller size to ensure that it matches the wire size, Figure 4.5. Also check the **conduit liner** size for compatibility with the wire size. Connect the conduit to the feed unit. The conduit or an extension should be aligned with the groove in the roller and set as close to the roller as possible without touching, Figure 4.6. Misalignment at this point can contribute to a bird's nest, Figure 4.7. Bird-nesting of the electrode wire, so called because it looks like a bird's nest, results when the feed roller pushes the wire into a tangled ball because the wire would not go through the outfeed side conduit.

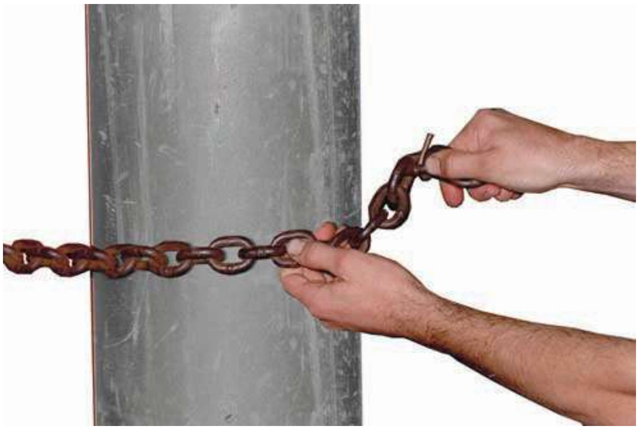

Figure 4.2 Make sure the gas cylinder is chained securely in place before removing the safety cap
Source: Courtesy of Larry Jeffus

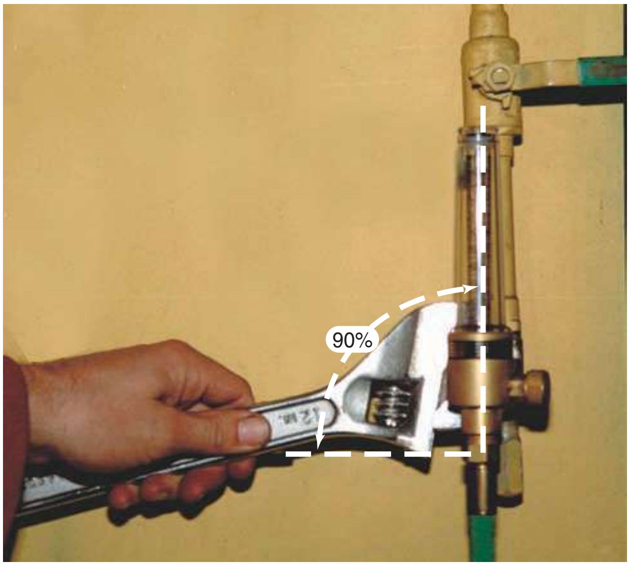

Figure 4.3 Attaching the flowmeter regulator
Be sure the tube is vertical.
Source: Courtesy of Larry Jeffus

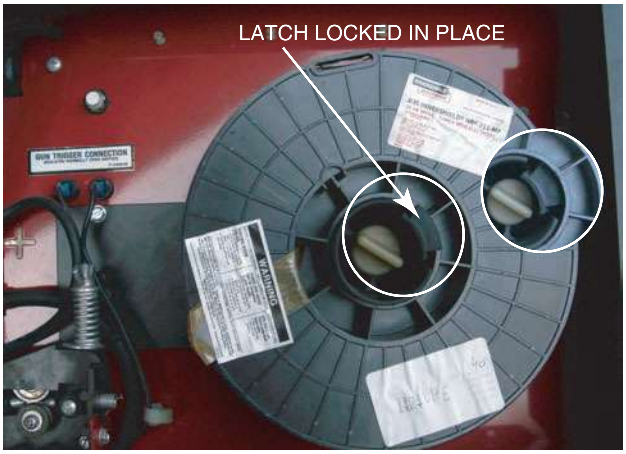
LATCH LOCKED IN PLACE

Figure 4.4 Wire reel may be secured by a center nut or locking lever
Source: Courtesy of Larry Jeffus

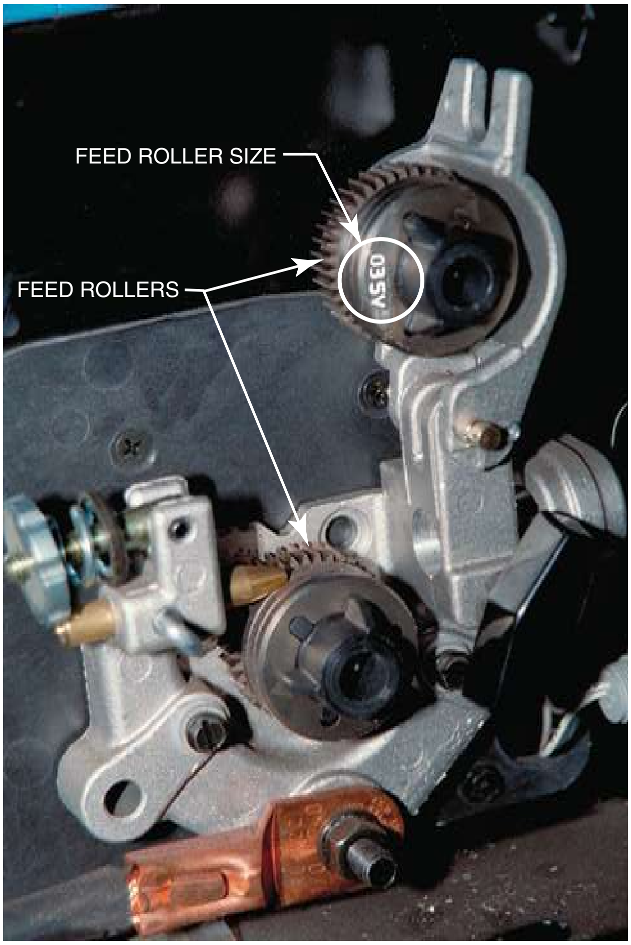
FEED ROLLER SIZE
FEED ROLLERS

Figure 4.5 Checking feed roller size
Check to be certain that the feed rollers are the correct size for the wire being used.
Source: Courtesy of Larry Jeffus

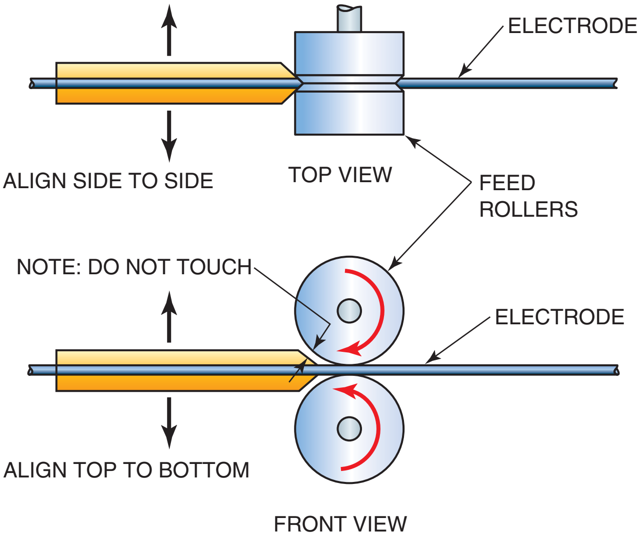

Figure 4.6 Feed roller and conduit alignment

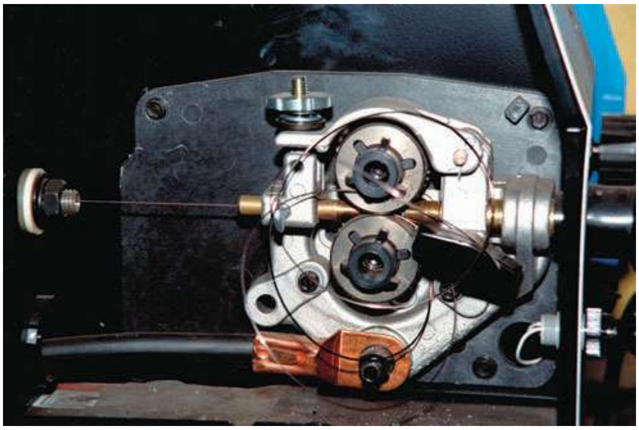

Figure 4.7 "Bird's nest" in the filler wire at the feed rollers
Source: Courtesy of Larry Jeffus

Be sure the power is off before attaching the welding cables. The electrode and work leads should be attached to the proper terminals. The electrode lead should be attached to the electrode or positive (+) terminal. If necessary, it is also attached to the power cable part of the gun lead. The work lead should be attached to the work or negative (–) terminal.

The shielding "gas-out" side of the solenoid is then also attached to the gun lead. If a separate splice is required from the gun switch circuit to the feed unit, it should be connected at this time. Check that the welding contactor circuit is connected from the feed unit to the power source.

The welding cable liner or wire conduit must be securely attached to the gas diffuser and contact tube, Figure 4.8. The **contact tube** (tip), Figure 3.1, must be the correct size to match the electrode wire size being used. If a shielding gas is to be used, a gas nozzle would be attached to complete the assembly. If a gas nozzle is not needed for a shielding gas, it may still be installed. Because it is easy for a student to touch the work with the contact tube during welding, an electrical short may occur. This short-out of the contact tube will immediately destroy the tube. Although the gas nozzle may interfere with some visibility, it may be worth the trouble for a new welder. FCA welding is more sensitive to changes in arc voltage than is SMA (stick) welding. Such variations in FCA welding voltage can dramatically and adversely affect your ability to maintain weld bead control.

A loose or poor connection will result in increased circuit resistance and a loss of welding voltage. To be sure that you have a good work

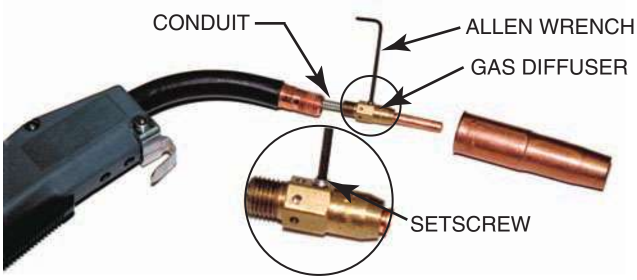

Figure 4.8 Securely attach conduit to gas diffuser and contact tube to prevent wire jams caused by misalignment
Source: Courtesy of Larry Jeffus

connection, remove any dirt, rust, oil, or other surface contamination at the point where the work clamp is connected to the weldment.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 4-2

Threading FCAW Wire

Using the FCAW machine that was properly assembled in Practice 4-1, you will turn the machine on and thread the electrode wire through the system.

Check that the unit is assembled correctly according to the manufacturer’s specifications. Switch on the power and check the gun switch circuit by depressing the switch. The power source relays, feed relays, gas solenoid, and feed motor should all activate.

Cut off the end of the electrode wire if it is bent. When working with the wire, be sure to hold it tightly. The wire will become tangled if it is released. The wire has a natural curl known as *cast.* Straighten out about 12 in. (300 mm) of the curl to make threading easier. Separate the wire **feed rollers** and push the wire first through the guides, then between the rollers, and finally into the conduit liner, Figure 4.9. Reset the rollers so there is a slight amount of compression on the wire, Figure 4.10. Set the **wire-feed speed** control to a slow speed. Hold the welding gun so that the electrode conduit and cable are as straight as possible.

Wearing safety glasses and pointing the gun away from the welder’s face, press the gun switch or the cold feed switch on the feeder. Pressing the gun switch to start the wire feeder is called triggering the gun. The cold feed switch on the feeder is a safety option built into some equipment that advances the wire without current being sent to the gun. The wire should start feeding into the liner. Watch to make certain that the wire feeds smoothly and release the gun switch as soon as the end comes through the gun.

If the wire stops feeding before it reaches the end of the gun, stop and check the system. If no obvious problem can be found, mark the wire with tape and remove it from the gun. It can then be held next to the system to determine the location of the problem.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Key Indicator 2
Gas Shielded
Key Indicator 3
Self Shielded
Key Indicator 8

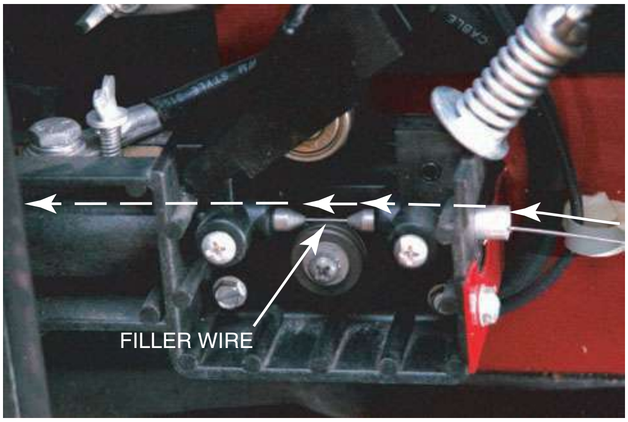

Figure 4.9 Push the wire through the guides by hand
Source: Courtesy of Larry Jeffus

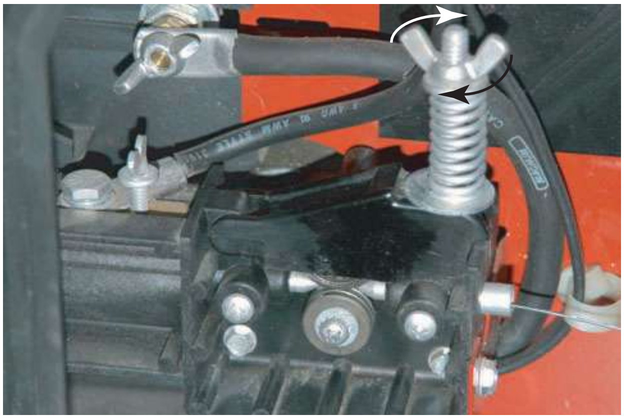

Figure 4.10 Adjust the wire-feed tensioner
Source: Courtesy of Larry Jeffus

With the wire feed running, adjust the feed roller compression so that the wire reel can be stopped easily by a slight pressure. Too light a roller pressure will cause the wire to feed erratically. Too high a pressure can crush some wires, causing some flux to be dropped inside the wire liner. If this happens, you will have a continual problem with the wire not feeding smoothly or jamming.

With the feed running, adjust the spool drag so that the reel stops when the feed stops. The reel should not coast to a stop, because it allows slack in the wire that can easily be snagged. Also, when the feed restarts, a jolt occurs when the slack in the wire is taken up. This jolt can be enough to momentarily stop the wire, possibly causing a discontinuity in the weld.

When the test runs are completed, you can either rewind or cut off the wire. Some wire-feed units have a retract button. This allows the feed driver to reverse and retract the wire automatically. To rewind the wire on units without this retraction feature, release the rollers and turn them backward by hand. If the machine will not allow the feed rollers to be released without upsetting the tension, you must cut the wire. Some wire reels have covers to prevent the collection of dust, dirt, and metal filings on the wire, Figure 4.11.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

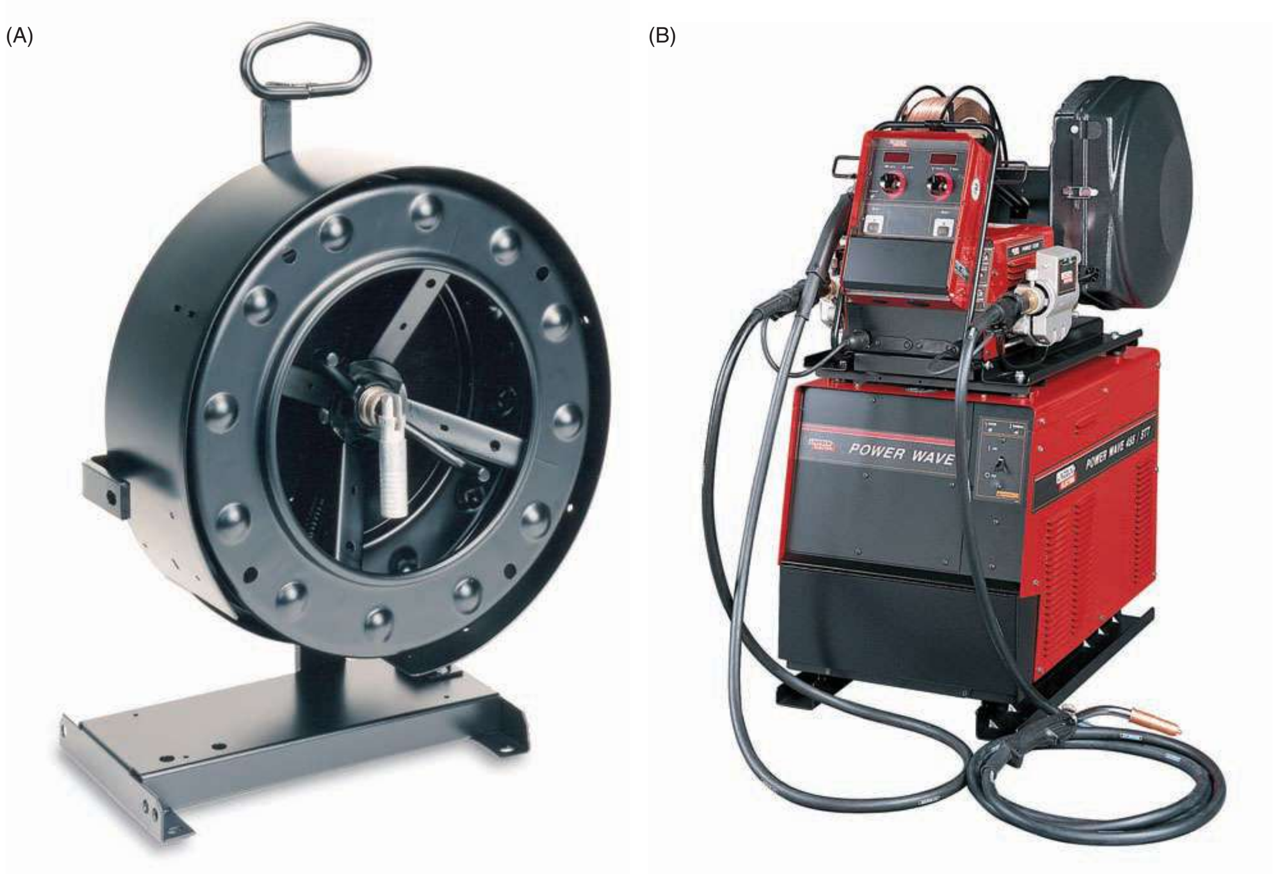

Figure 4.11 Wire protection
(A) Covered wire reel. (B) Wire cover on a dual wire-feed system.
Source: Courtesy of Lincoln Electric Company

FLAT-POSITION WELDS

PRACTICE 4-3

Stringer Beads, Flat Position

Using a properly set up and adjusted FCA welding machine, Table 4.1, proper safety protection, E70T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm), and one or more pieces of mild steel plate, 12 in. (305 mm) long and 1/4 in. (6 mm) or thicker, you will make a stringer bead weld in the flat position, Figure 4.12.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Key Indicator 1
Gas Shielded
Key Indicator 4
Self Shielded
Key Indicator 9

Table 4.1 **FCA Welding Parameters for Use if Specific Settings Are Unavailable from Electrode Manufacturer (base metal thickness 1/4 to 1/2 inch)**

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	Size	Amps	Wire-feed Speed, ipm (cm/min.)	Volts	Type	Flow	Type	Thickness
E70T-1 E71T-1	0.035 in. (0.9 mm)	130 to 150	288 to 380 (732 to 975)	22 to 25	None	n/a	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E70T-1 E71T-1	0.045 in. (1.2 mm)	150 to 210	200 to 300 (508 to 762)	28 to 29	None	n/a	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E70T-5 E71T-11	0.035 in. (0.9 mm)	130 to 200	288 to 576 (732 to 1463)	20 to 28	75% argon 25% CO_2	30 cfh	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E70T-5 E71T-11	0.045 in. (1.2 mm)	150 to 250	200 to 400 (508 to 1016)	23 to 29	75% argon 25% CO_2	35 cfh	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)

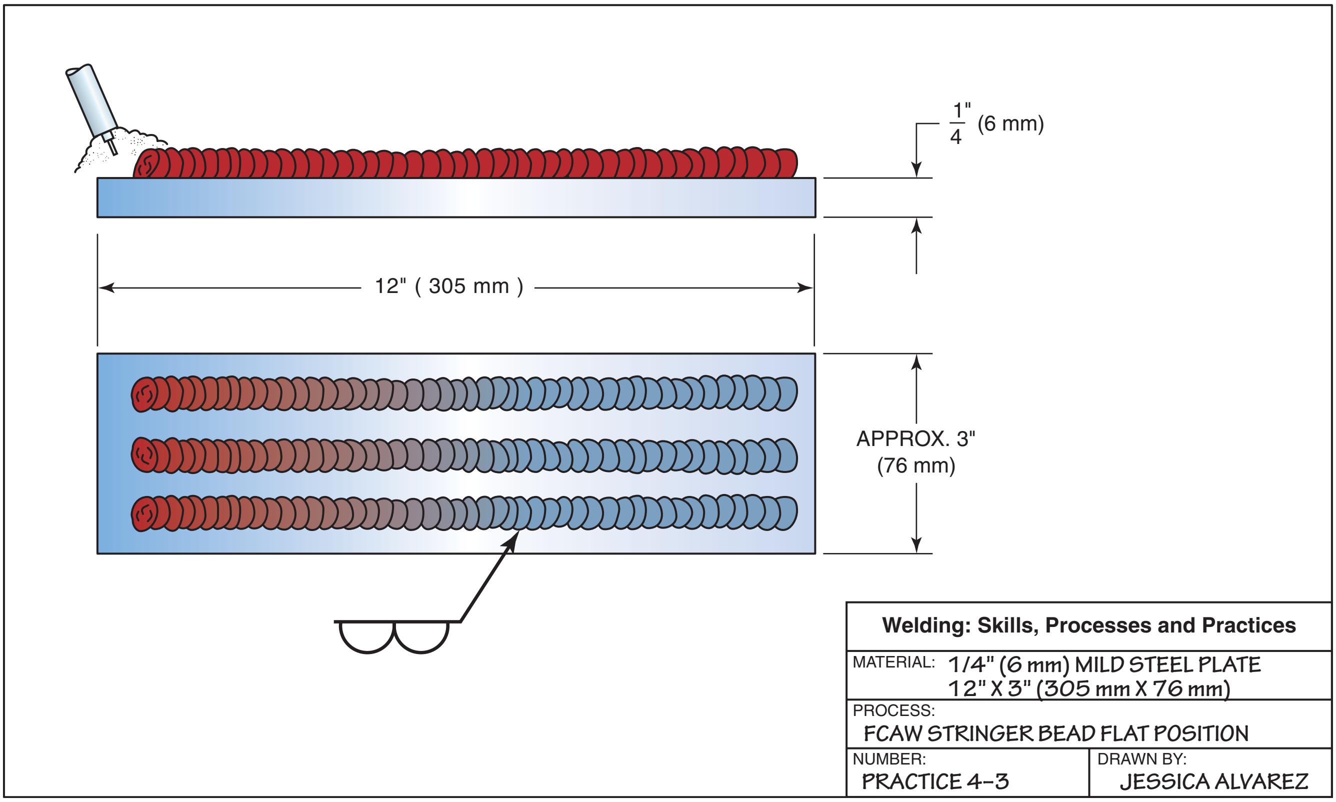

Welding: Skills, Processes and Practices
MATERIAL: 1/4" (6 mm) MILD STEEL PLATE 12" X 3" (305 mm X 76 mm)
PROCESS: FCAW STRINGER BEAD FLAT POSITION
NUMBER: PRACTICE 4-3
DRAWN BY: JESSICA ALVAREZ

Figure 4.12 **FCAW stringer bead, 1/4 in. mild steel, flat position**

Starting at one end of the plate and using a dragging technique, make a weld bead along the entire 12-in. (305-mm) length of the metal. After the weld is complete, check its appearance. Make any needed changes to correct the weld. Repeat the weld and make additional adjustments. After the machine is set, start to work on improving the straightness and uniformity of the weld. Use weave patterns of different widths and straight stringers without weaving.

Repeat with both classifications of electrodes until beads can be made straight, uniform, and free from any visual defects. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

SQUARE-GROOVE WELDS

One advantage of FCA welding is the ability to make 100%-joint-penetrating welds without beveling the edges of the plates. These full-joint-penetrating welds can be made in plates that are 1/4 in. (6 mm) or less in thickness. Welding on thicker plates risks the possibility of a lack of fusion on both sides of the **root face**, Figure 4.13.

There are several disadvantages of having to bevel a plate before welding:

- Beveling the edge of a plate adds an operation to the fabrication process.
- Both more filler metal and welding time are required to fill a beveled joint than are required to make a square jointed weld.
- Beveled joints have more heat from the thermal beveling and additional welding required to fill the groove. The lower heat input to the square joint means less distortion.

The major disadvantage of making square jointed welds is that as the plate thickness approaches 1/4 in. (6 mm) or if the weld is out of

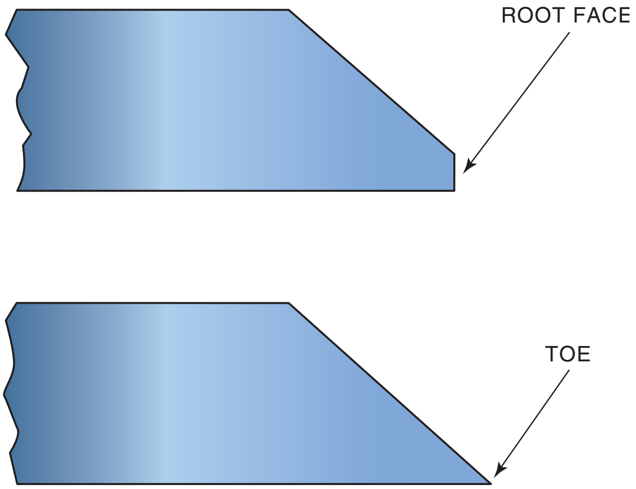

Figure 4.13 A beveled joint may or may not have a flat surface, called a root face

position, a much higher level of skill is required. The skill required to make quality square welds can be acquired by practicing on thinner metal. It is much easier to make this type of weld in metal 1/8 in. (3 mm) thick and then move up in thickness as your skills improve.

PRACTICE 4-4

Butt Joint 1G

Using a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm), and one or more pieces of mild steel plate, 12 in. (305 mm) long and 1/4 in. (6 mm) or less in thickness, you will make a groove weld in the flat position, Figure 4.14.

- Tack weld the plates together and place them in position to be welded.
- Starting at one end, run a bead along the joint. Watch the molten weld pool and bead for signs that a change in technique may be required.
- Make any needed changes as the weld progresses in order to produce a uniform weld.

Repeat with both classifications of electrodes until defect-free welds can consistently be made in the 1/4-in.-thick (6-mm-thick) plate. Turn

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Flat" position portion of the all-position requirement of 6 and 11.

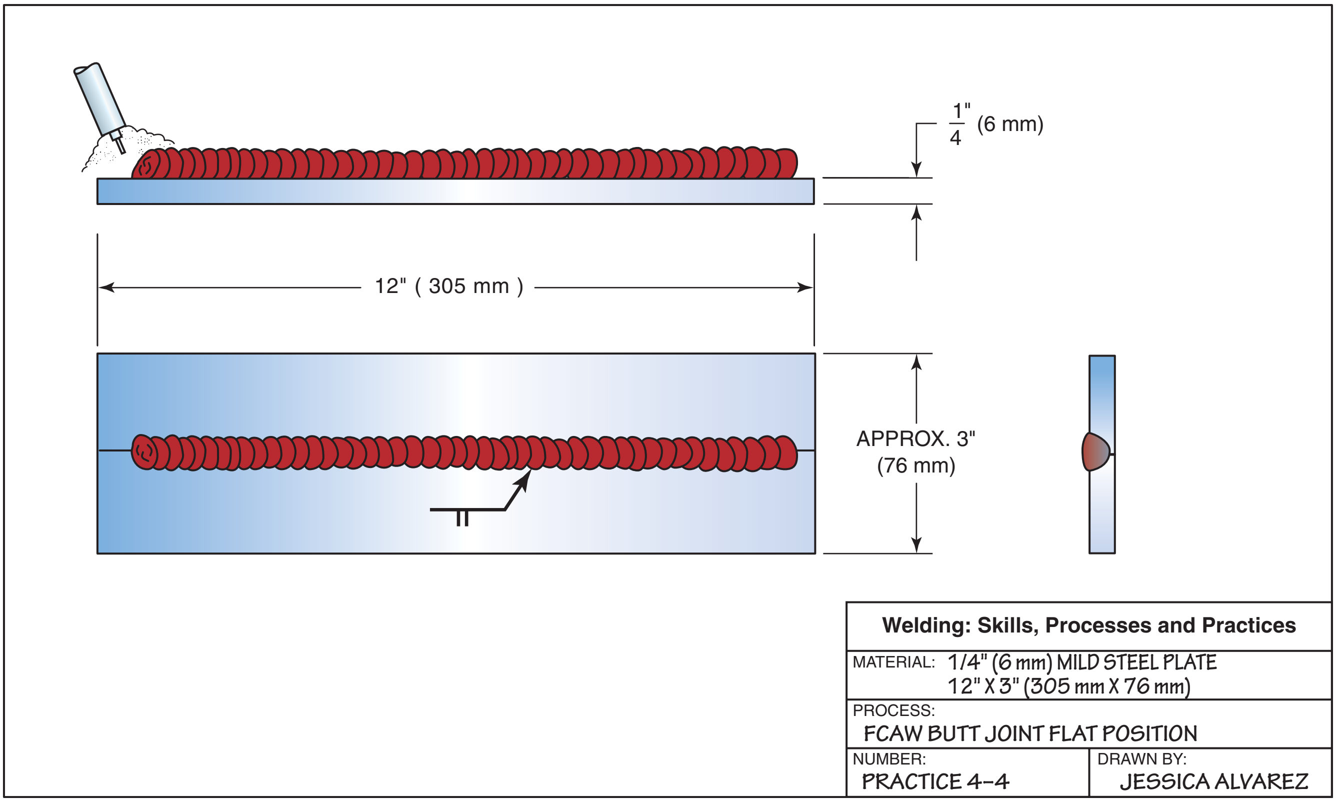

Welding: Skills, Processes and Practices
MATERIAL: 1/4" (6 mm) MILD STEEL PLATE
12" X 3" (305 mm X 76 mm)
PROCESS:
FCAW BUTT JOINT FLAT POSITION
NUMBER: PRACTICE 4–4
DRAWN BY: JESSICA ALVAREZ

Figure 4.14 **FCAW butt joint, 1/4 in. mild steel, flat position**

off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

V-GROOVE AND BEVEL-GROOVE WELDS

Although for speed and economy engineers try to avoid specifying welds that require beveling the edges of plates, it is not always possible. Anytime the metal being welded is thicker than 1/4 in. (6 mm) and a 100% joint penetration weld is required, the edges of the plate must be prepared with a bevel. Fortunately, FCA welding allows a narrower groove to be made and still achieve a thorough thickness weld, due to the deeper penetration characteristics of the FCAW process, Figure 4.15.

All FCA groove welds are made using three different types of weld passes, Figure 4.16.

- *Root pass:* The first weld bead of a multiple-pass weld. The root pass fuses the two parts together and establishes the depth of weld metal penetration.
- *Filler pass:* Made after the root pass is completed and used to fill the groove with weld metal. More than one pass is often required.
- *Cover pass:* The last weld pass on a multipass weld. The cover pass may be made with one or more welds. It must be uniform in width, reinforcement, and appearance.

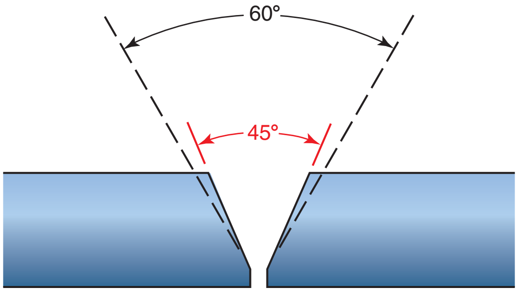

Figure 4.15 Reduced groove angle for FCAW
A smaller groove angle reduces both weld time and filler metal required to make the weld.

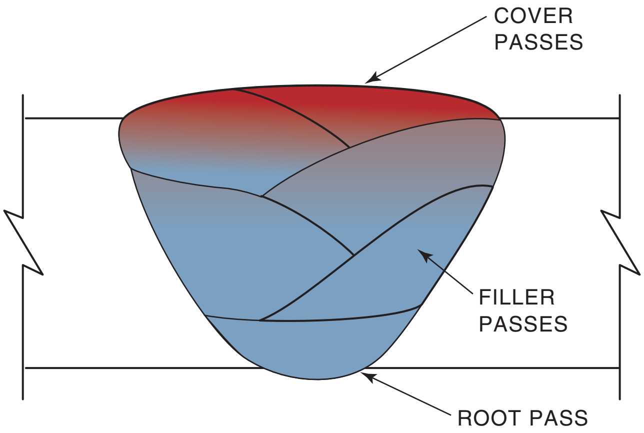

Figure 4.16 Three types of weld passes make up a weld

Root Pass

A good root pass is needed in order to obtain a sound weld. The root may be either open or closed using a backing strip, Figure 4.17.

The backing strips are usually made from a piece of metal 1/4 in. (6 mm) thick and 1 in. (25 mm) wide, and should be 2 in. (50 mm) longer than the base plates. The strip is attached to the plate by tack welds made on the sides of the strip, Figure 4.18.

Most production welds do not use backing strips, so they are made as open root welds. Because of the difficulty in controlling root weld face contours in FCAW, however, open-root joints are often avoided on **critical welds**. If an open-root weld is needed because of weldment design, the root pass may be put in with an SMAW electrode or the root face of the FCA weld can be retouched by grinding and/or back welding.

Care must be taken with any root pass not to have the weld face too convex, Figure 4.19. Convex weld faces tend to trap slag along the toe of

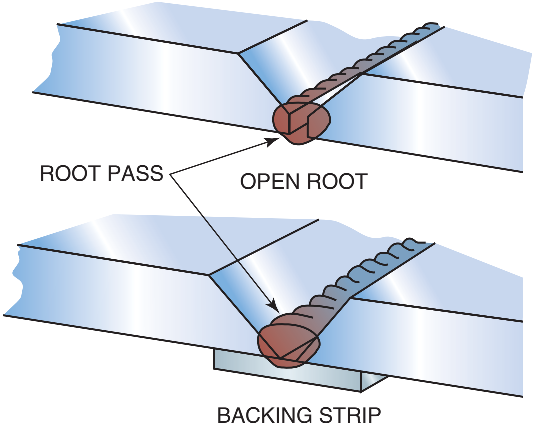

Figure 4.17 Root pass
The maximum deposit for a root pass is 1/4 in. (6 mm) thick.

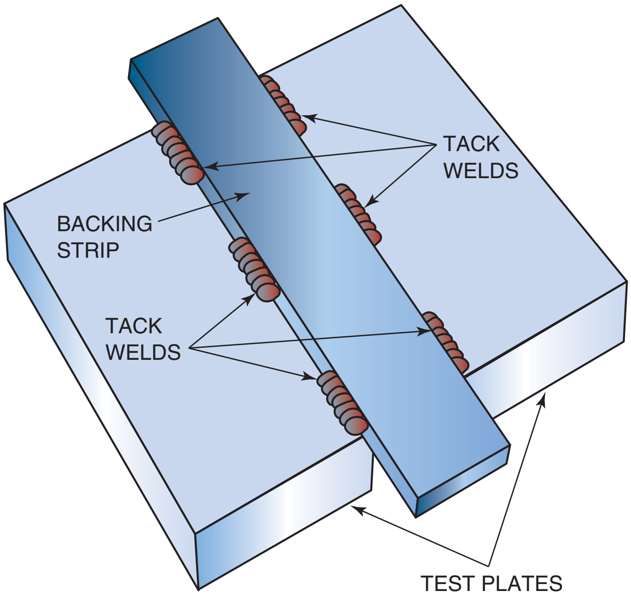

Figure 4.18 Backing strip
Securely tack weld the backing strip to the test plates.

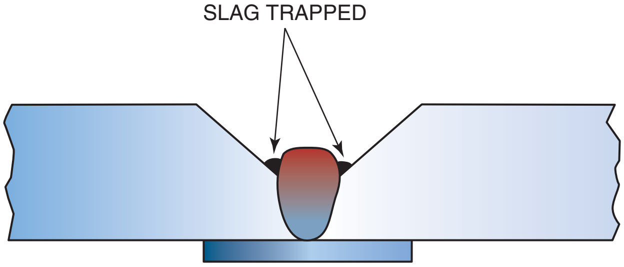

Figure 4.19 Slag trapped beside weld bead is hard to remove

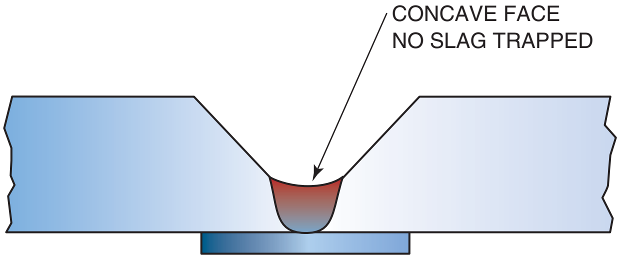

Figure 4.20 Flat or concave weld faces are easier to clean off

the weld. FCA weld slag can be extremely difficult to remove in this area, especially if there is any undercutting. To avoid this, adjust the welding power settings, speed, and weave pattern so that a flat or slightly concave weld face is produced, Figure 4.20.

Filler Pass

Filler passes are made with either **stringer beads**, which are made with a straight progression and very little gun manipulation, or **weave beads**, in which the operator oscillates the gun from side to side in order to widen the weld profile. Either bead type works well for flat or vertically positioned welds, but stringer beads work best for horizontal and overhead-positioned welds. When multiple-pass filler welds are required, each weld bead must overlap the others along the edges. Edges should overlap smoothly enough so that the finished bead is uniform, Figure 4.21. Stringer beads usually overlap about 25% to 50%, and weave beads overlap approximately 10% to 25%.

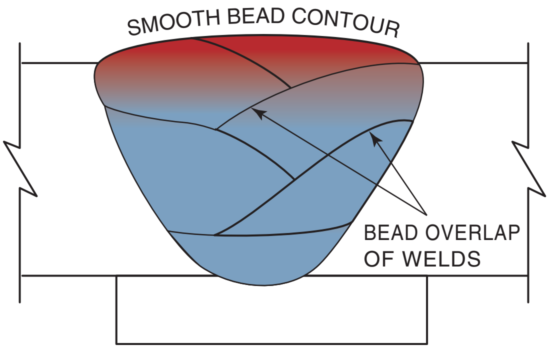

Figure 4.21 The surface of a multipass weld should be as smooth as if it were made by one weld

Each weld bead must be cleaned before the next bead is started. The filler pass ends when the groove has been filled to a level just below the plate surface.

Cover Pass

The cover pass may or may not simply be a continuation of the weld beads used to make the filler pass(es). The major difference between the filler pass and the cover pass is the weld face importance. Keeping the face and toe of the cover pass uniform in width, reinforcement, and appearance and free of defects is essential. Most welds are not tested beyond a visual inspection. For that reason the appearance might be the only factor used for accepting or rejecting welds.

The cover pass must meet a strict visual inspection standard. The visual inspection looks to see that the weld is uniform in width and reinforcement. There should be no arc strikes or hammer marks from chipping or slag removal operations on the plate other than those on the weld itself. The weld must be free of both incomplete fusion and cracks. The weld must be free of overlap, and undercut must not exceed either 10% of the base metal or 1/32 in. (0.8 mm), whichever is less. Reinforcement must have a smooth transition with the base plate and be no higher than 1/8 in. (3 mm), Figure 4.22.

PRACTICE 4-5

Butt Joint 1G

Use a properly set up and adjusted FCA welding machine, Table 4.1; proper safety protection; E70T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick; and a backing strip 14 in. (355 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the flat position, Figure 4.23.

Tack weld the backing strip to the plates. There should be a root gap of approximately 1/8 in. (3 mm) between the plates. The beveled surface can be made with or without a root face, Figure 4.24.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Gas Shielded
Key Indicator 11
This practice addresses the "Flat" position portion of the all-position requirement of 6 and 11.

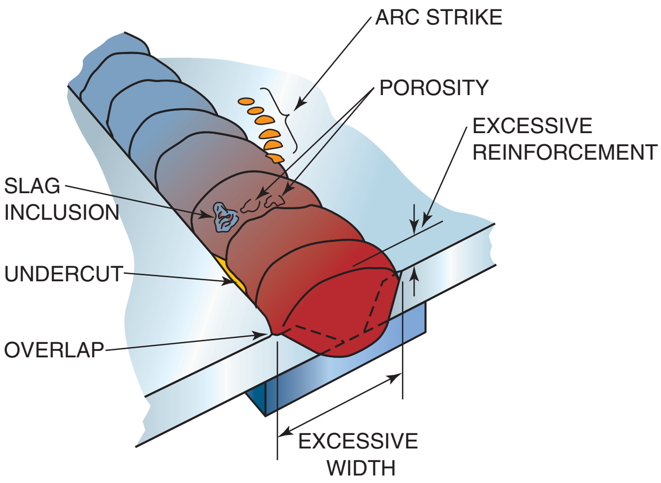

Figure 4.22 Common discontinuities found during a visual examination

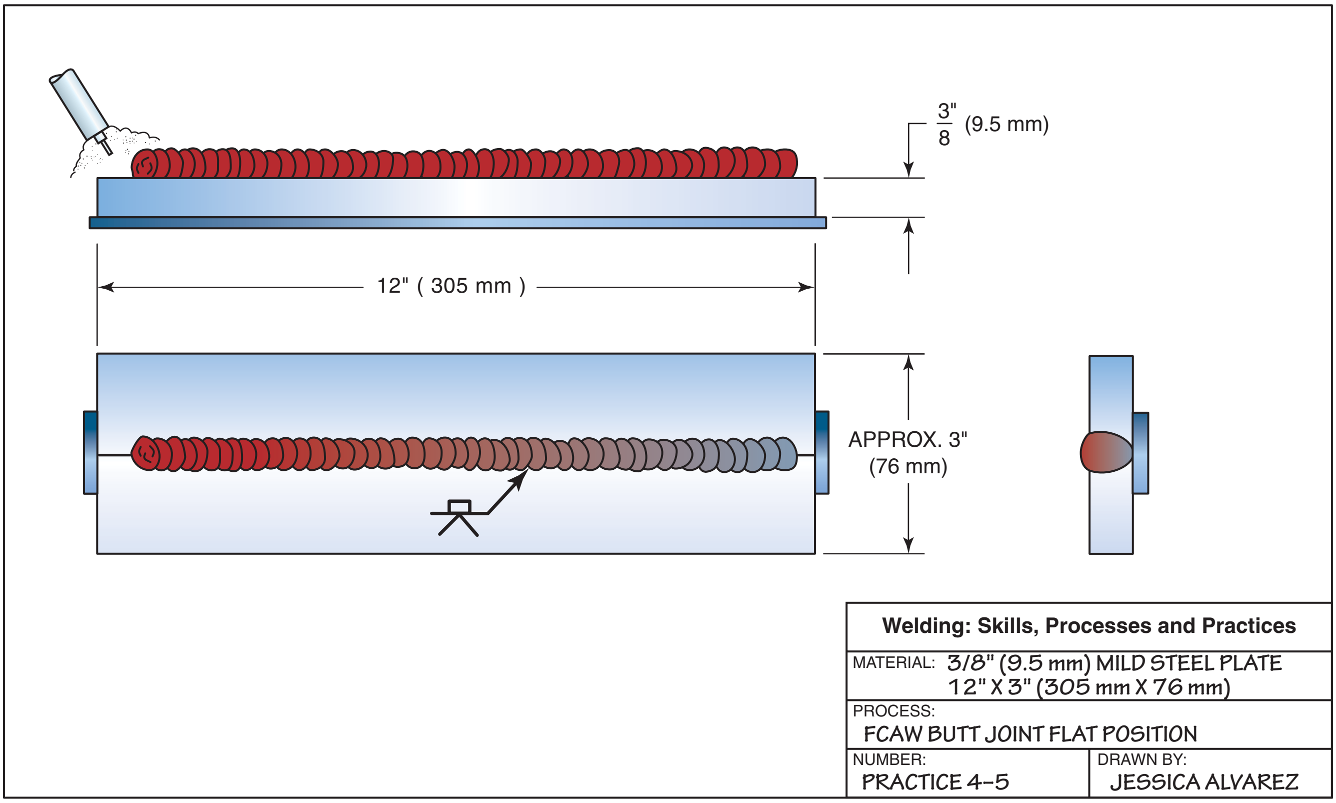
3" / 8 (9.5 mm)
12" (305 mm)
APPROX. 3" (76 mm)
Welding: Skills, Processes and Practices
MATERIAL: 3/8" (9.5 mm) MILD STEEL PLATE 12" X 3" (305 mm X 76 mm)
PROCESS: FCAW BUTT JOINT FLAT POSITION
NUMBER: PRACTICE 4-5
DRAWN BY: JESSICA ALVAREZ

Figure 4.23 FCAW butt joint, 3/8 in. mild steel, flat position

Place the test plates in position at a comfortable height and location. Be sure that you have complete and free movement along the full length of the weld joint. It is often a good idea to make a practice pass along the joint with the welding gun without power to make sure nothing will interfere with your making the weld. Be sure the welding cable is free and will not get caught on anything during the weld.

Start the weld outside the groove on the backing strip tab, Figure 4.25. This is done so that the arc is smooth and the molten weld pool size is

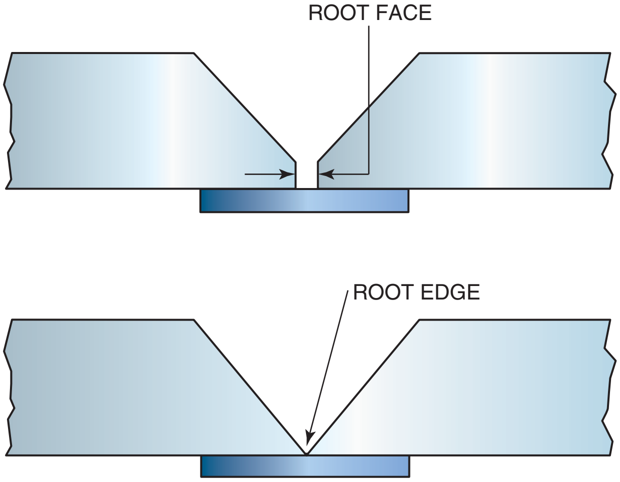
ROOT FACE
ROOT EDGE

Figure 4.24 Groove layout with and without a root face

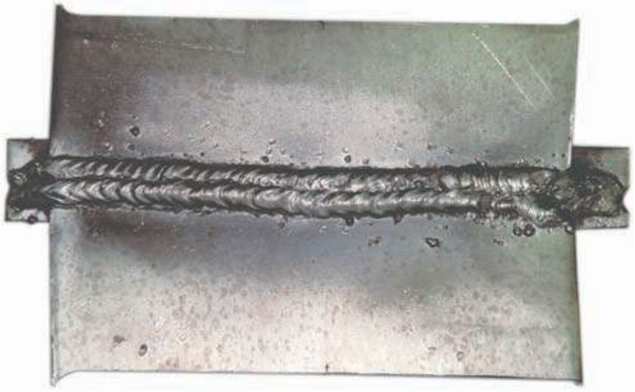

Figure 4.25 Using the ends of a backing strip
Run-off tabs, in which the weld starts and stops on the ends of the backing strip, help control possible underfill or burn-back at the starting and stopping points of a groove weld.
Source: Courtesy of Larry Jeffus

established at the beginning of the groove. Continue the weld out onto the tab at the outer end of the groove. This process ensures that the end of the groove is completely filled with weld.

Repeat with both classifications of electrodes until consistently defect-free welds can be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-6

Butt Joint 1G 100% to Be Tested

Use a properly set up and adjusted FCA welding machine; proper safety protection; E70T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick; and a backing strip 14 in. (355 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the flat position, Figure 4.26.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-5, repeat the weld until you can use each electrode type to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Flat" position portion of the all-position requirement of 6 and 11.

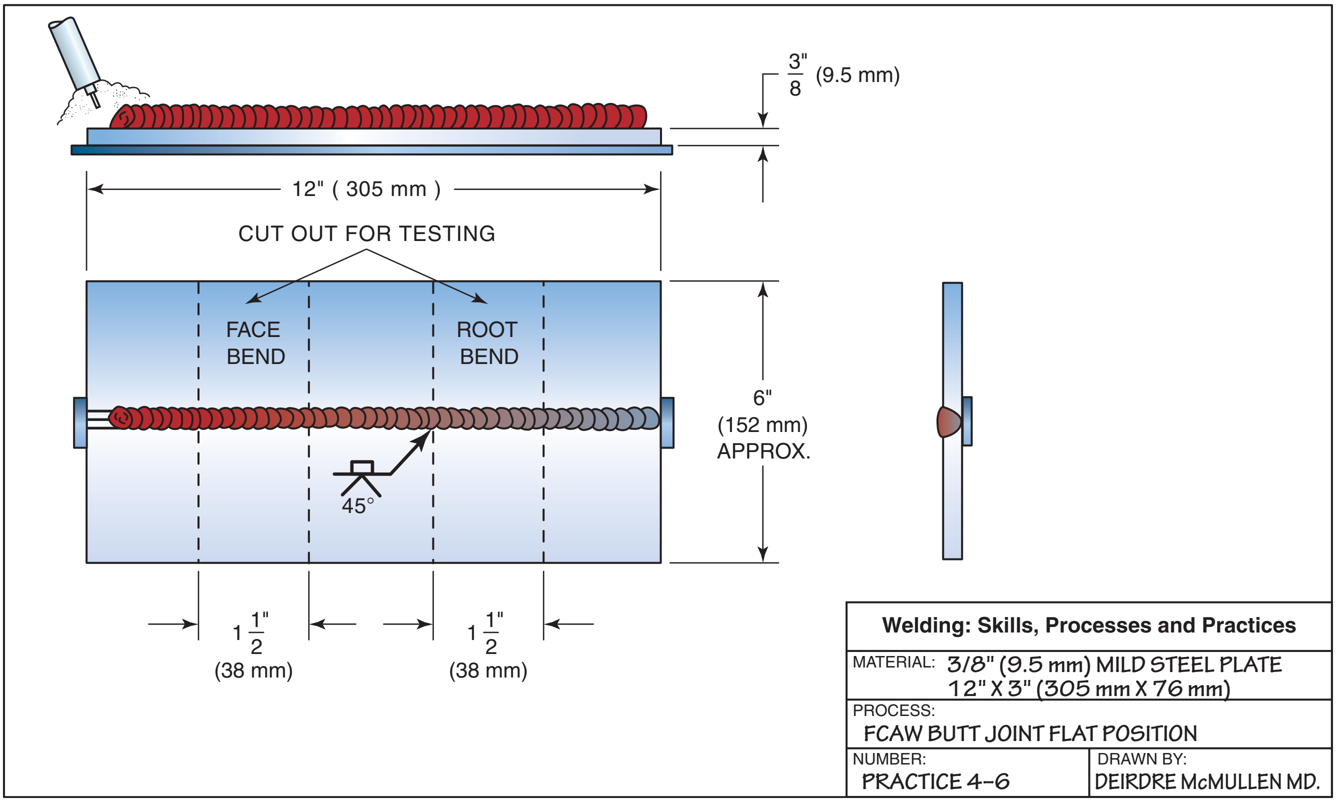

Welding: Skills, Processes and Practices	
MATERIAL: 3/8" (9.5 mm) MILD STEEL PLATE 12" X 3" (305 mm X 76 mm)	
PROCESS: FCAW BUTT JOINT FLAT POSITION	
NUMBER: PRACTICE 4-6	DRAWN BY: DEIRDRE McMULLEN MD.

Figure 4.26 FCAW butt joint, 3/8 in. mild steel, flat position

FILLET WELDS

A fillet weld is the type of weld made on a **lap joint** and a **tee joint**. It should be built up equal to the thickness of the plate, Figure 4.27. On thick plates the fillet must be made up of several passes as with a groove weld. The difference with a fillet weld is that a smooth transition from the plate surface to the weld is required. If this transition is abrupt, it can cause stresses that will weaken the joint.

The lap joint is made by overlapping the edges of the plates. They should be held together tightly before tack welding them together. A small tack weld may be added in the center to prevent distortion during welding, Figure 4.28. Chip and wire brush the tacks before you start to weld.

The tee joint is made by tack welding one piece of metal on another piece of metal at a right angle, Figure 4.29. After the joint is tack welded together, the slag is chipped from the tack welds. If the slag is not removed, it will cause a slag inclusion in the final weld.

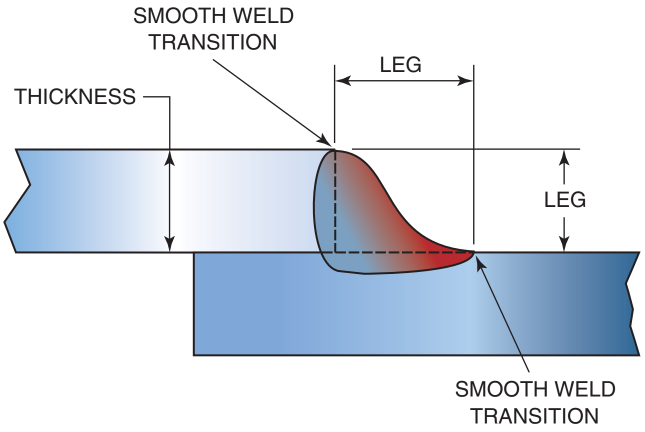
SMOOTH WELD TRANSITION
LEG
THICKNESS
LEG
SMOOTH WELD TRANSITION

Figure 4.27 **The legs of a fillet weld should generally be equal to the thickness of the base metal**

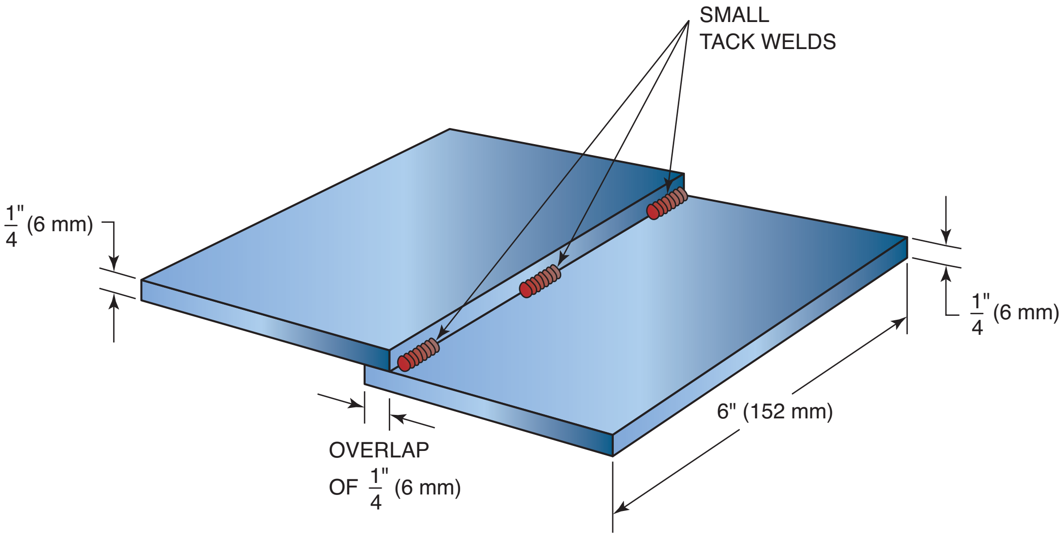
SMALL TACK WELDS
$\frac{1}{4}$" (6 mm)
$\frac{1}{4}$" (6 mm)
6" (152 mm)
OVERLAP OF $\frac{1}{4}$" (6 mm)

Figure 4.28 **Tack welding the plates of a lap joint together**

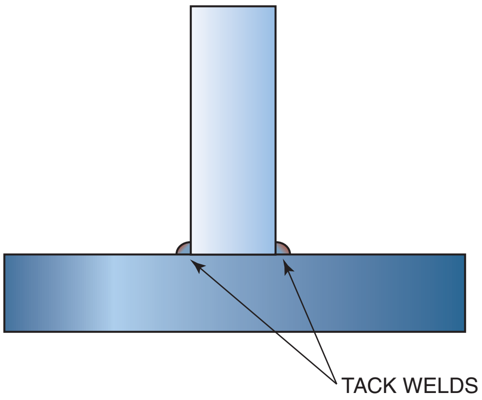

Figure 4.29 Tack welding both sides of a tee joint
Tack welding will help keep the tee square for welding.

Holding thick plates tightly together on tee joints may cause underbead cracking, or lamellar tearing, Figure 4.30. On thick plates the weld shrinkage can be great enough to pull the metal apart well below the bead or its heat-affected zone. In production welds, cracking can be controlled by not assembling the plates tightly together. The space between the two plates can be set by placing a small wire spacer between them, Figure 4.31.

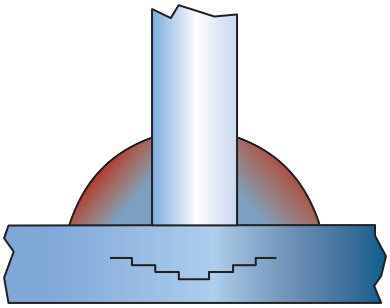

Figure 4.30 Underbead cracking, or lamellar tearing, of the base plate

A fillet welded lap or tee joint can be strong if it is welded on both sides, even without having deep penetration, Figure 4.32. Some tee joints may be prepared for welding by cutting either a bevel or a J-groove in the vertical plate. This cut is not required for strength but may be necessary because of design limitations. Unless otherwise specified, most fillet welds will be equal in size to the plates welded. A fillet weld will be as strong as the base plate if the size of the two welds equals the total thickness of the base plate. The weld bead should have a flat or slightly concave appearance to ensure the greatest strength and efficiency, Figure 4.33.

The root of fillet welds must be melted to ensure a completely fused joint. A notch along the root of the weld pool is an indication that the root is not being fused together, Figure 4.34. To achieve complete root fusion, move the arc to a point as close as possible to the leading edge of the weld pool, Figure 4.35. If the arc strikes the unmelted plate ahead of the molten weld pool, it may become erratic, which will increase weld spatter.

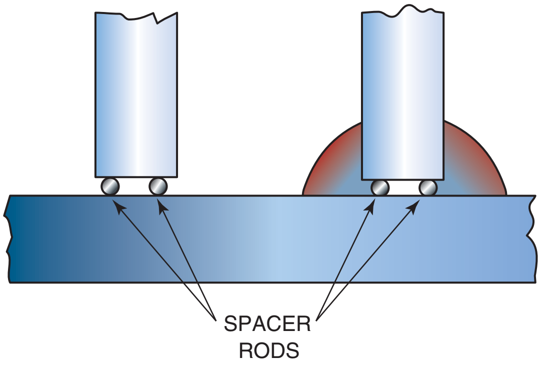

Figure 4.31 Spacers in tee joints
Base plate cracking can be controlled by placing spacers in the joint before welding.

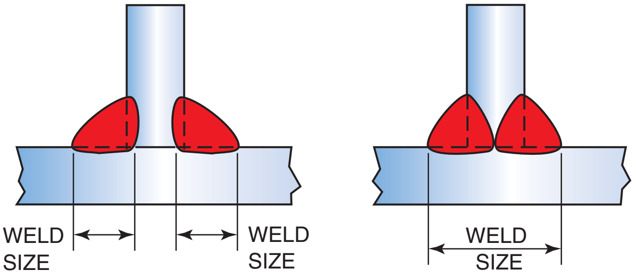

Figure 4.32 Fillet weld size
If the total weld sizes are equal, then both tee joints would have equal strength.

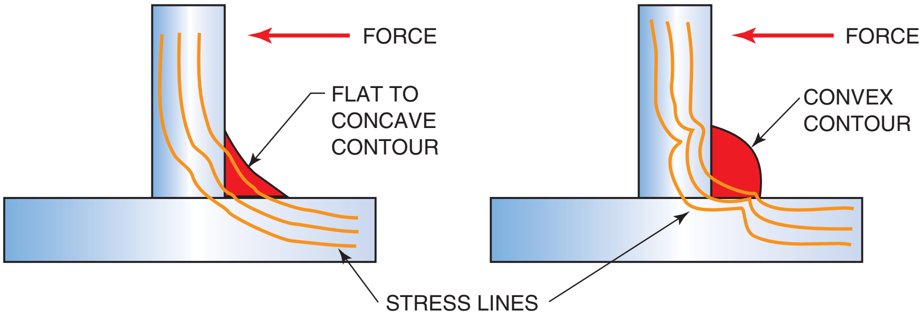

Figure 4.33 **Fillet weld shape**
The stresses are distributed more uniformly through a flat or concave fillet weld.

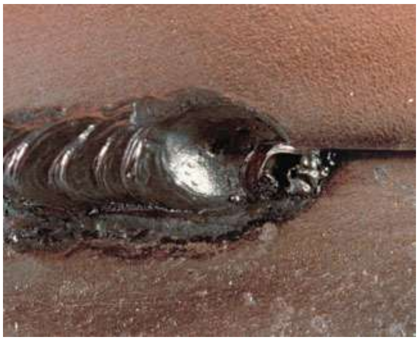

Figure 4.34 **Watch the root of the weld bead to be sure there is complete fusion**
Source: Courtesy of Larry Jeffus

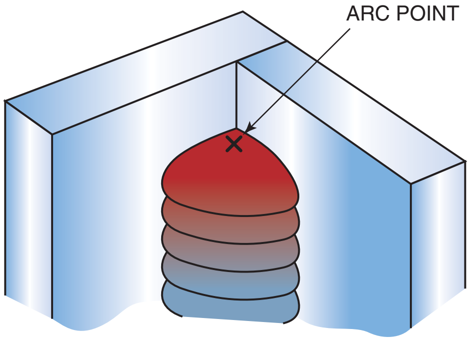

Figure 4.35 **Moving the arc as close as possible to the leading edge of the weld will provide good root fusion.**

PRACTICE 4-7

Lap Joint and Tee Joint 1F

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10
This practice addresses the "Flat" position portion of the all-position requirement for 5 and 10.

Use a properly set up and adjusted FCA welding machine; proper safety protection; E70T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); and one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick. You will make a fillet weld in the flat position.

Tack weld the pieces of metal together and brace them in position. When making the lap or tee joints in the flat position, the plates must be at a 45° angle so that the surface of the weld will be flat, Figure 4.36A and Figure 4.36B. Starting at one end, make a weld along the entire length of the joint.

Repeat each type of joint with both classifications of electrodes until consistently defect-free welds can be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-8

Tee Joint 1F

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10

Use a properly set up and adjusted FCA welding machine, Table 4.2; proper safety protection; E70T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or through 1/16 in. (0.9 mm and/or through 1.6 mm); one

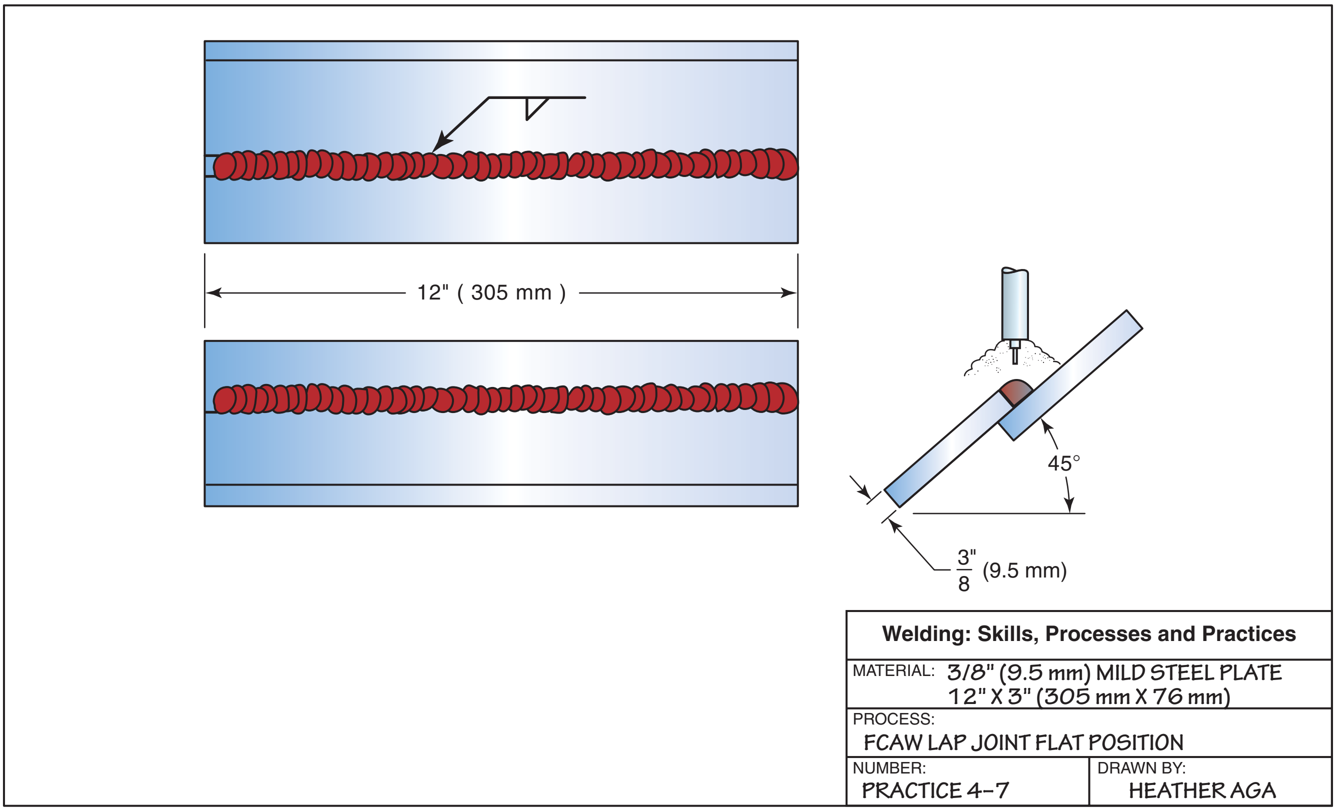
12" (305 mm)
45°
3"/8 (9.5 mm)
Welding: Skills, Processes and Practices
MATERIAL: 3/8" (9.5 mm) MILD STEEL PLATE
12" X 3" (305 mm X 76 mm)
PROCESS:
FCAW LAP JOINT FLAT POSITION
NUMBER:
PRACTICE 4-7
DRAWN BY:
HEATHER AGA

Figure 4.36 (A) FCAW lap joint, 3/8 in. mild steel, flat position

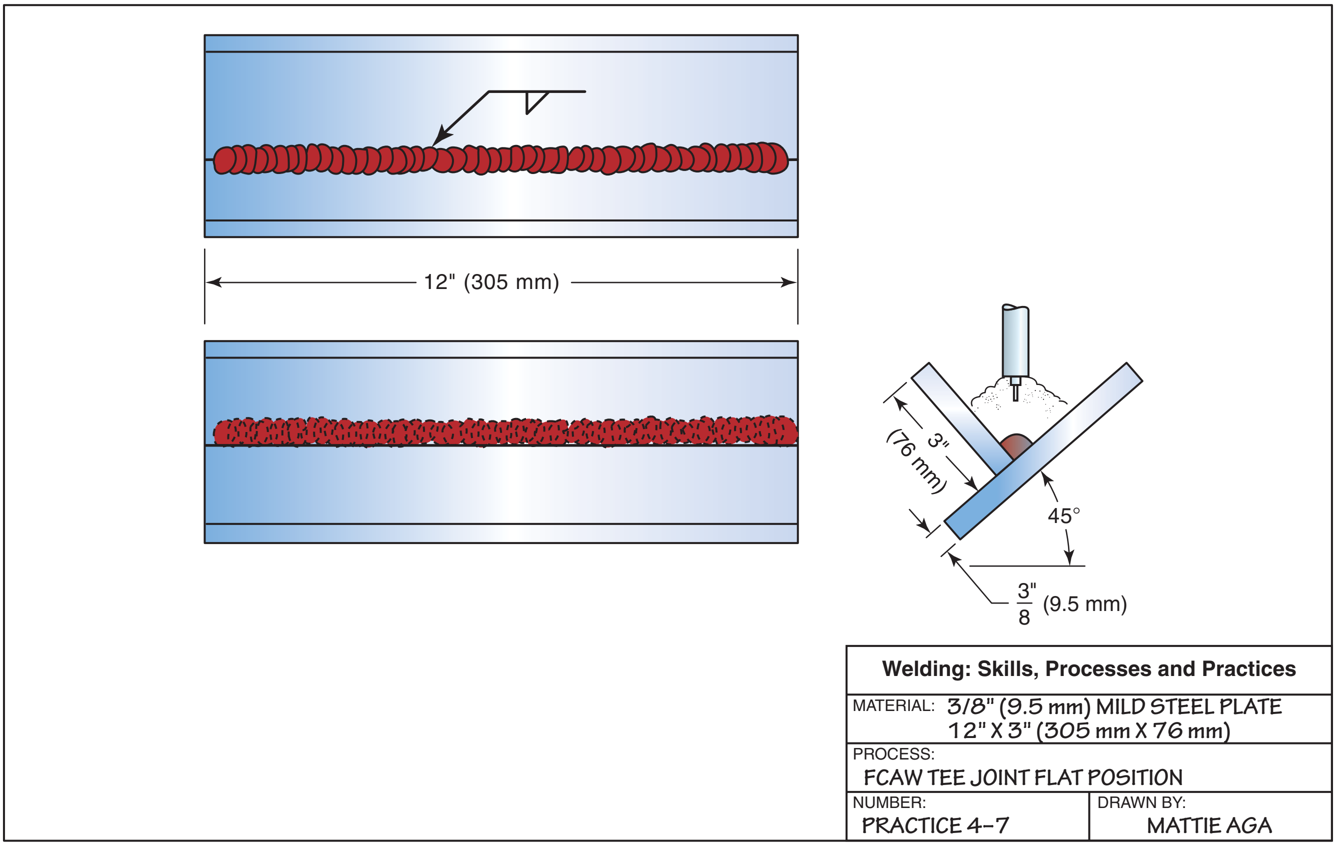
12" (305 mm)
3" (76 mm)
45°
3"/8 (9.5 mm)
Welding: Skills, Processes and Practices
MATERIAL: 3/8" (9.5 mm) MILD STEEL PLATE
12" X 3" (305 mm X 76 mm)
PROCESS:
FCAW TEE JOINT FLAT POSITION
NUMBER:
PRACTICE 4-7
DRAWN BY:
MATTIE AGA

Figure 4.36 (B) FCAW tee joint, 3/8 in. mild steel, flat position

Table 4.2 FCA Welding Parameters for Use if Specific Settings Are Unavailable from Electrode Manufacturer (base metal thickness 1/2 to 3/4 inch)

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	Size	Amps	Wire-feed Speed, ipm (cm/min)	Volts	Type	Flow	Type	Thickness
E70T-1 E71T-1	0.035 in. (0.9 mm)	130 to 150	288 to 380 (732 to 975)	22 to 25	None	n/a	Low-carbon steel	1/2 in. to 3/4 in. (13 mm to 19 mm)
E70T-1 E71T-1	0.045 in. (1.2 mm)	150 to 210	200 to 300 (508 to 762)	28 to 29	None	n/a	Low-carbon steel	1/2 in. to 3/4 in. (13 mm to 19 mm)
E70T-1 E71T-1	.052 in. (1.4 mm)	150 to 300	150 to 350 (381 to 889)	25 to 33	None	n/a	Low-carbon steel	1/2 in. to 3/4 in. (13 mm to 19 mm)
E70T-1 E71T-1	1/16 in. (1.6 mm)	200 to 400	150 to 300 (381 to 762)	27 to 33	None	n/a	Low-carbon steel	1/2 in. to 3/4 in. (13 mm to 19 mm)
E70T-5 E71T-11	0.035 in. (0.9 mm)	130 to 200	288 to 576 (732 to 1463)	20 to 28	75% argon 25% CO_2	30 cfh	Low-carbon steel	1/2 in. to 3/4 in. (13 mm to 19 mm)
E70T-5 E71T-11	0.045 in. (1.2 mm)	150 to 250	200 to 400 (508 to 1016)	23 to 29	75% argon 25% CO_2	35 cfh	Low-carbon steel	1/2 in. to 3/4 in. (13 mm to 19 mm)
E70T-5 E71T-11	0.052 in. (1.4 mm)	150 to 300	150 to 350 (381 to 889)	21 to 32	75% argon 25% CO_2	35 cfh	Low-carbon steel	1/2 in. to 3/4 in. (13 mm to 19 mm)
E70T-5 E71T-11	1/16 in. (1.6 mm)	180 to 400	145 to 350 (368 to 889)	21 to 34	75% argon 25% CO_2	40 cfh	Low-carbon steel	1/2 in. to 3/4 in. (13 mm to 19 mm)

or more pieces of mild steel plate, beveled, 7 in. (178 mm) long and 3/4 in. (19 mm) thick or thicker. You will make a fillet weld in the flat position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-7, repeat each type of joint with both classifications of electrodes until consistently defect-free welds can be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

VERTICAL WELDS

PRACTICE 4-9

Butt Joint at a 45° Vertical Up Angle

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Key Indicator 1, 2
Gas Shielded
Key Indicator 3
Gas Shielded
Key Indicator 4
Self Shielded
Key Indicator 8
Self Shielded
Key Indicator 9

Using a properly set up and adjusted FCA welding machine, proper safety protection, 0.035-in. and/or 0.045-in. (0.9-mm and/or 1.2-mm)-diameter E71T-1 and/or E71T-11 electrodes, and one or more pieces of mild steel plate, 12 in. (305 mm) long and 1/4 in. (6 mm) thick or thinner, you will increase the plate angle gradually as you develop skill until you are making satisfactory welds in the vertical up position, Figure 4.37.

- Start practicing this weld with the plate at a 45° angle.
- Gradually increase the angle of the plate to vertical as skill is gained in welding this joint. A straight stringer bead or slight zigzag will work well on this joint.
- Establish a molten weld pool in the root of the joint.
- Cool, chip, and inspect the weld for uniformity and defects.

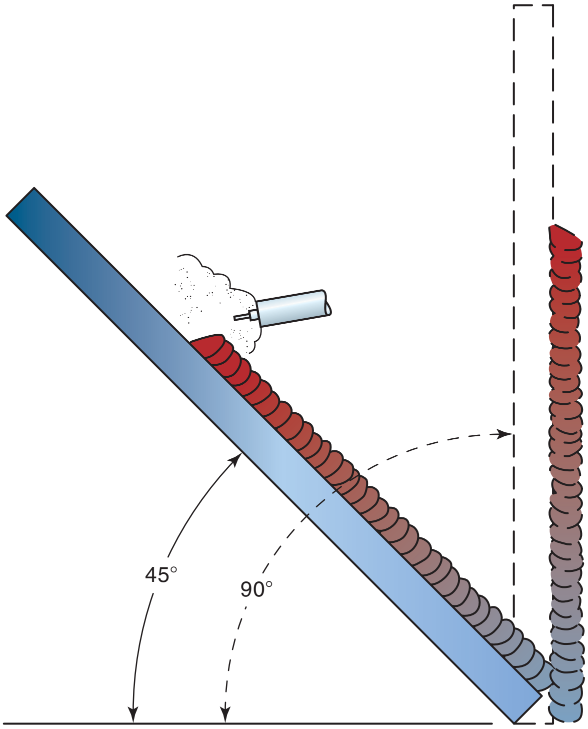

Figure 4.37 Start making welds with the plate at a 45° angle
As your skill develops, increase the angle until the plate is vertical.

It is easier to make a quality weld in the vertical up position if lower settings are used in both the **amperage range** and **voltage range**. This will make the molten weld pool smaller, less fluid, and easier to control. A problem with lower power settings is that the weld bead often can be very convex, Figure 4.38. Faster travel speed and/or slightly wider weave patterns can be used to control the bead shape.

Start at the bottom of the plate and hold the welding gun at a slight upward angle to the plate, Figure 4.39. Brace yourself, lower your hood, and begin to weld. Depending on the machine settings and type of electrode used, you will make a weave pattern.

If the molten weld pool is large and fluid (hot), use a C or J weave pattern to allow a longer time for the molten weld pool to cool, Figure 4.40. Do not make the weave so long or fast that the electrode is allowed to strike the metal ahead of the molten weld pool. If this happens, spatter increases and a spot or zone of incomplete fusion may occur.

A weld that is high and has little or no fusion is too "cold." Changing the welding technique will not correct this problem. The welder must stop welding and make the needed adjustments to the power supply or electrode feeder. Continue to weld along the entire 12-in. (305-mm) length of plate.

Repeat welds with both electrodes until defect-free welds can be consistently made vertically in the 1/4-in. (6-mm)-thick plate. Turn off the

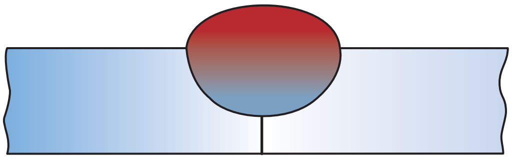

Figure 4.38 Low amperage causes too much buildup and not enough penetration

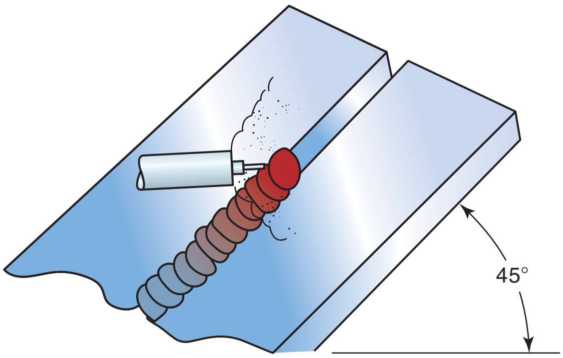

Figure 4.39 45° vertical up

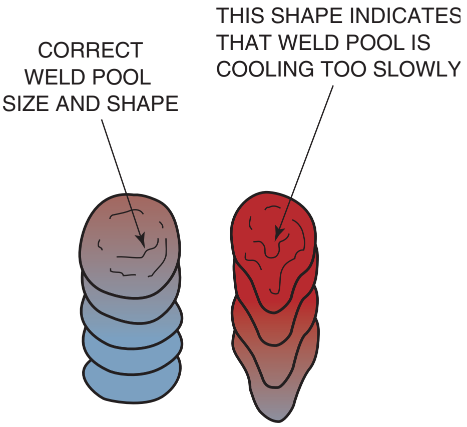
CORRECT WELD POOL SIZE AND SHAPE
THIS SHAPE INDICATES THAT WELD POOL IS COOLING TOO SLOWLY

Figure 4.40 The shape of the weld pool can indicate the temperature of the surrounding base metal

welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Vertical" portion of the all-position requirement.

PRACTICE 4-10

Butt Joint 3G

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm), and one or more pieces of mild

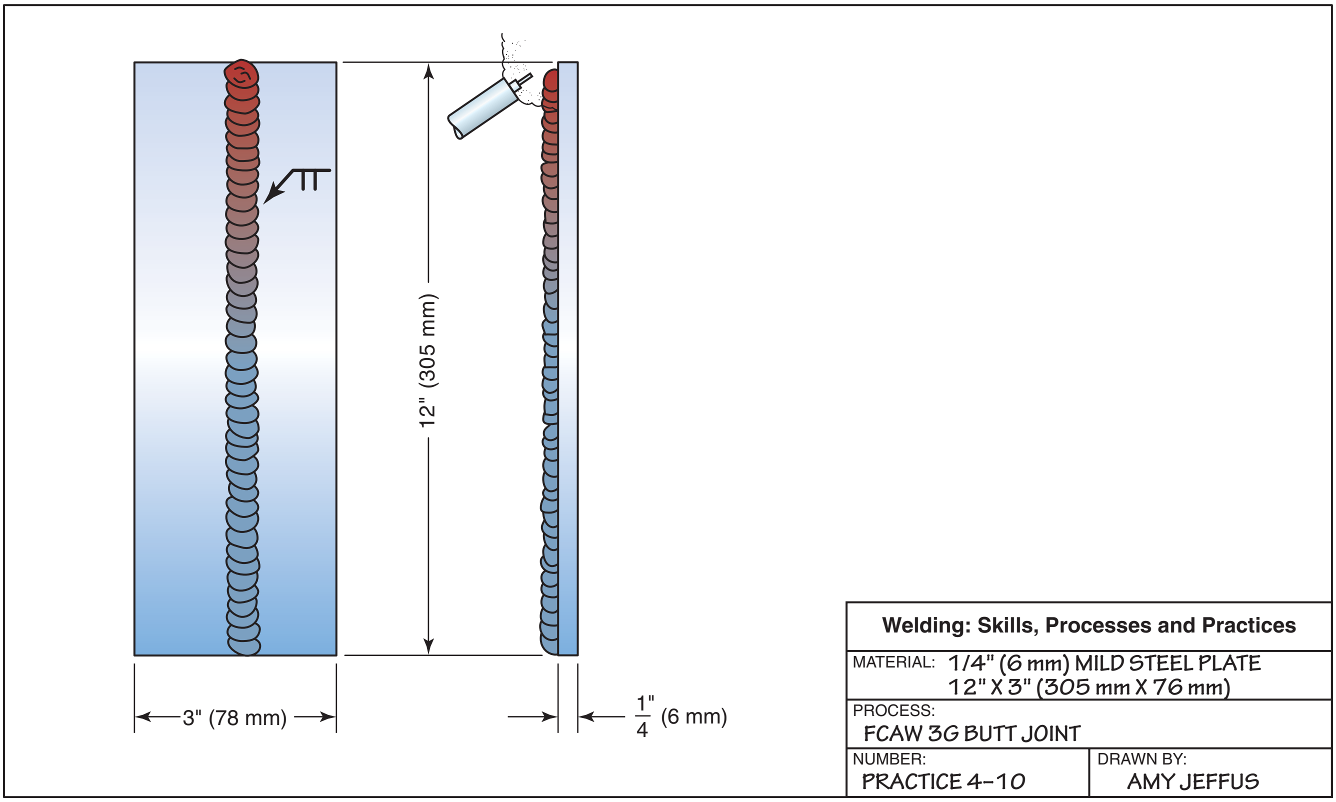
12" (305 mm)
3" (78 mm)
$\frac{1}{4}$" (6 mm)

Welding: Skills, Processes and Practices	
MATERIAL: 1/4" (6 mm) MILD STEEL PLATE 12" X 3" (305 mm X 76 mm)	
PROCESS: FCAW 3G BUTT JOINT	
NUMBER: PRACTICE 4-10	DRAWN BY: AMY JEFFUS

Figure 4.41 FCAW 3G butt joint, 1/4 in. mild steel

steel plate, 12 in. (305 mm) long and 1/4 in. (6 mm) thick or thinner, you will make a groove weld in the vertical position, Figure 4.41.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-9, repeat with both classifications of electrodes until defect-free welds can be consistently made in the 1/4-in. (6-mm)-thick plate. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-11

Butt Joint 3G

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick; and a backing strip 14 in. (355 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the vertical position.

Following the same instructions for assembly and welding procedure as outlined in Practice 4-9, repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-12

Butt Joint 3G 100% to Be Tested

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the "Vertical" component of the all-position requirement of 6 and 11.

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick; and a backing strip 14 in. (355 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the vertical position, Figure 4.42.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-9, repeat the weld until you can use each electrode type to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-13

Butt Joint 3G 100% to Be Tested

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the "Vertical" portion of the all-position requirement of 6 and 11.

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.045 in. and/or through 1/16 in. (0.9 mm and/or through 1.6 mm); one or more pieces of mild steel plate, beveled, 7 in. (178 mm) long and

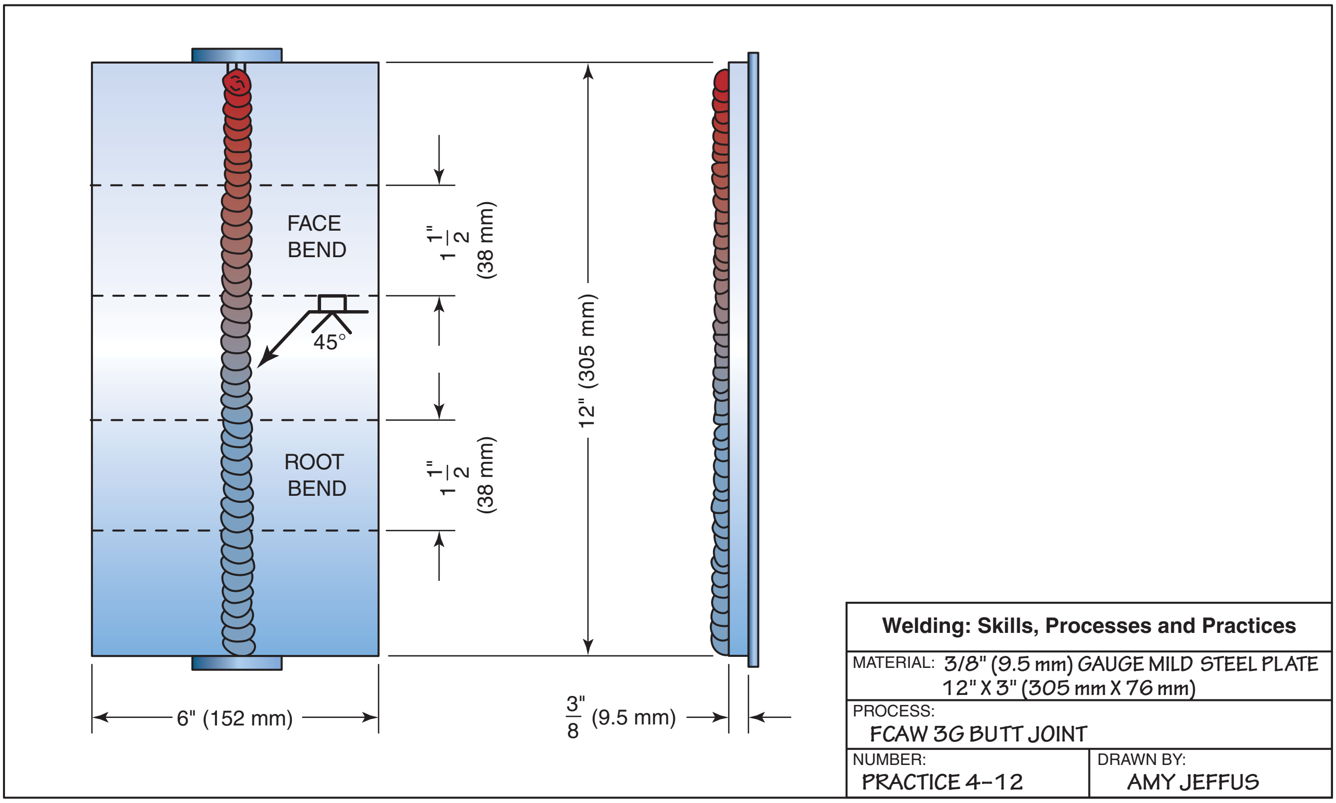
FACE
BEND
45°
ROOT
BEND
1 1/2" (38 mm)
12" (305 mm)
1 1/2" (38 mm)
6" (152 mm)
3/8" (9.5 mm)
Welding: Skills, Processes and Practices
MATERIAL: 3/8" (9.5 mm) GAUGE MILD STEEL PLATE
12" X 3" (305 mm X 76 mm)
PROCESS:
FCAW 3G BUTT JOINT
NUMBER:
PRACTICE 4-12
DRAWN BY:
AMY JEFFUS

Figure 4.42 FCAW 3G butt joint, 3/8 in. mild steel

3/4 in. (19 mm) thick or thicker; and a backing strip 9 in. (230 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the vertical position, Figure 4.43.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-9, repeat the weld until you can use each electrode type to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Key Indicator 1, 2
Gas Shielded
Key Indicator 3
Gas Shielded
Key Indicator 4
Self Shielded
Key Indicator 8
Self Shielded
Key Indicator 9
Self Shielded
Key Indicator 10
Self Shielded
Key Indicator 11
Self Shielded
Key Indicator 12

Module 9
Key Indicator 1, 2

PRACTICE 4-14

Fillet Weld Joint at a 45° Vertical Up Angle

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm), and one or more pieces of mild steel plate, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick, you will increase the plate angle gradually as you develop skill until you are making satisfactory welds in the vertical up position, Figure 4.44.

Tack weld the metal pieces together and brace them in position. Check to see that you have free movement for your gun along the entire joint to prevent stopping and restarting during the weld. Avoiding stops and starts both speeds up the welding time and eliminates discontinuities.

It is easier to make a quality weld in the vertical up position if both the amperage and voltage are set at the lower end of their ranges. This

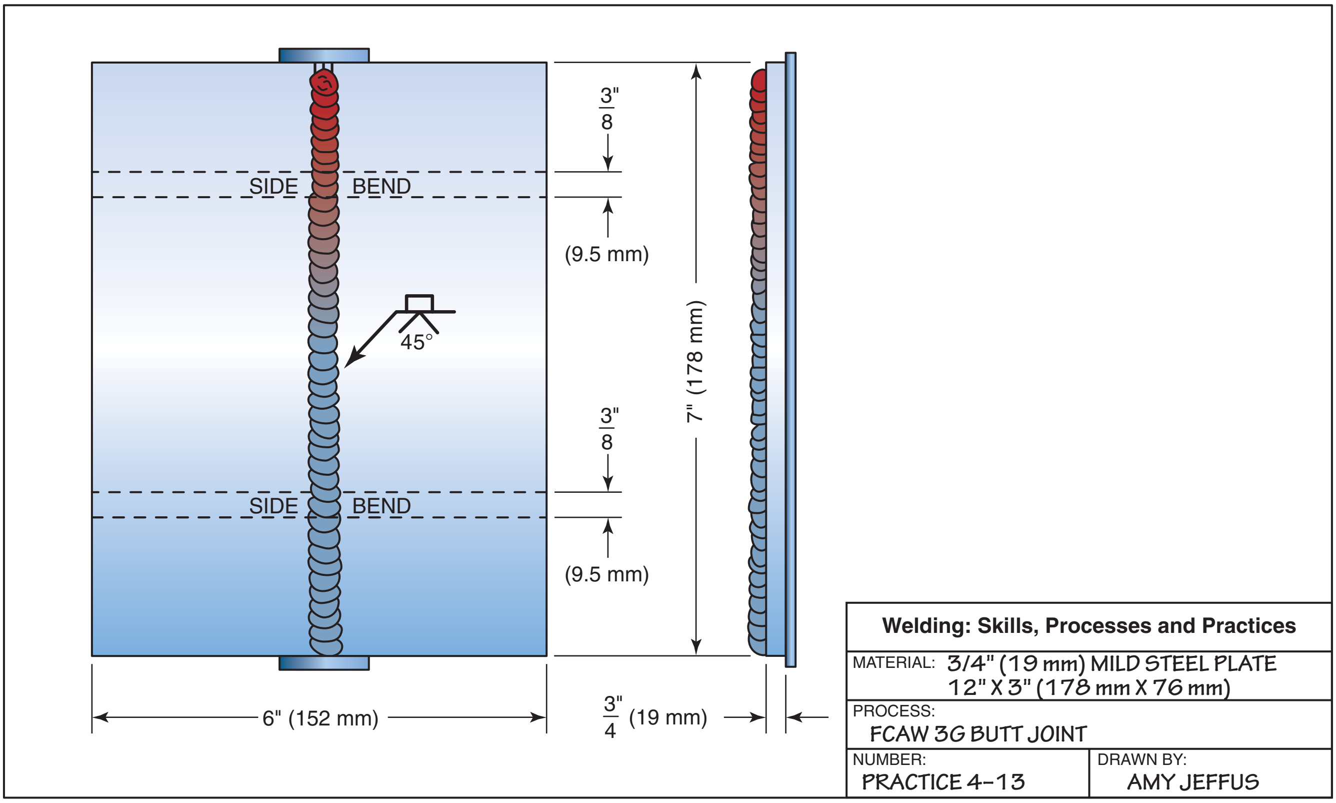
SIDE
BEND
SIDE
BEND
45°
3"/8
(9.5 mm)
3"/8
(9.5 mm)
7" (178 mm)
6" (152 mm)
3"/4 (19 mm)
Welding: Skills, Processes and Practices
MATERIAL: 3/4" (19 mm) MILD STEEL PLATE
12" X 3" (178 mm X 76 mm)
PROCESS:
FCAW 3G BUTT JOINT
NUMBER:
PRACTICE 4-13
DRAWN BY:
AMY JEFFUS

Figure 4.43 FCAW 3G butt joint, 3/4 in. mild steel

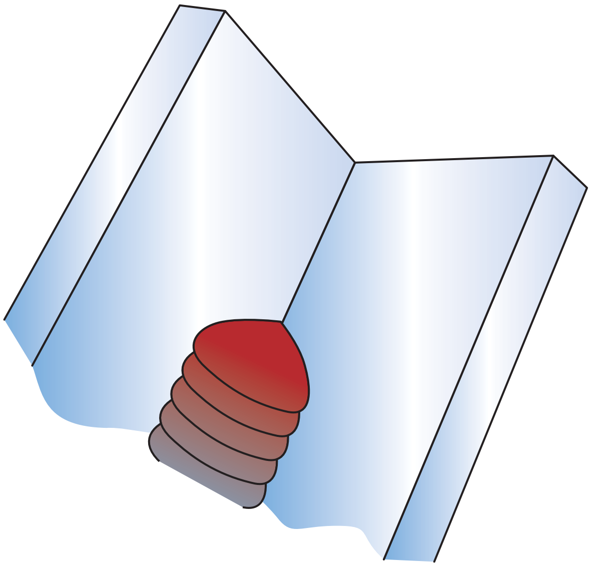

Figure 4.44 45° vertical up fillet weld

will make the molten weld pool smaller, less fluid, and easier to control. A problem with the lower power settings is that the weld bead often is very convex. A convex face on a weld bead often makes it more difficult to remove the slag along the toe of the weld.

The weave pattern should allow for adequate fusion on both edges of the joint. Watch the edges to be sure that they are being melted so that adequate fusion and penetration occur.

Repeat the weld with each electrode type until defect-free welds can consistently be made vertically. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4
Module 2
Key Indicator 1, 2, 3, 4, 7
Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10
Module 9
Key Indicator 1, 2
This practice addresses the “Vertical” position component of the all-position requirement.

PRACTICE 4-15

Lap Joint and Tee Joint 3F 100% to Be Tested

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); and one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick. You will make a fillet weld in the vertical position.

Following the same instructions for assembly and welding procedure as outlined in Practice 4-14, repeat each type of joint with both classifications of electrodes until you can make welds with 100% penetration that will pass the test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4
Module 2
Key Indicator 1, 2, 3, 4, 7
Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10
This practice addresses the “Vertical“ component of the all-position requirement for 5 and 10.

PRACTICE 4-16

Tee Joint 3F

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-11 electrodes of diameter 0.045 in. and/or through 1/16 in. (0.9 mm and/or through 1.6 mm), and one or more pieces of mild steel plate, 7 in. (178 mm) long and 3/4 in. (19 mm) thick, you will make a fillet weld in the vertical position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-14, repeat each type of joint with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4
Module 2
Key Indicator 1, 2, 3, 4, 7
Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10
This practice addresses the “Horizontal” component of the all-position requirement for 5 and 10.

HORIZONTAL WELDS

PRACTICE 4-17

Lap Joint and Tee Joint 2F

Use a properly set up and adjusted FCA welding machine; proper safety protection; E70T-1 and/or E71T-11 (or E71T-1 and/or E70T-5) electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); and one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick. You will make a fillet weld in the horizontal position, Figure 4.45A and Figure 4.45B.

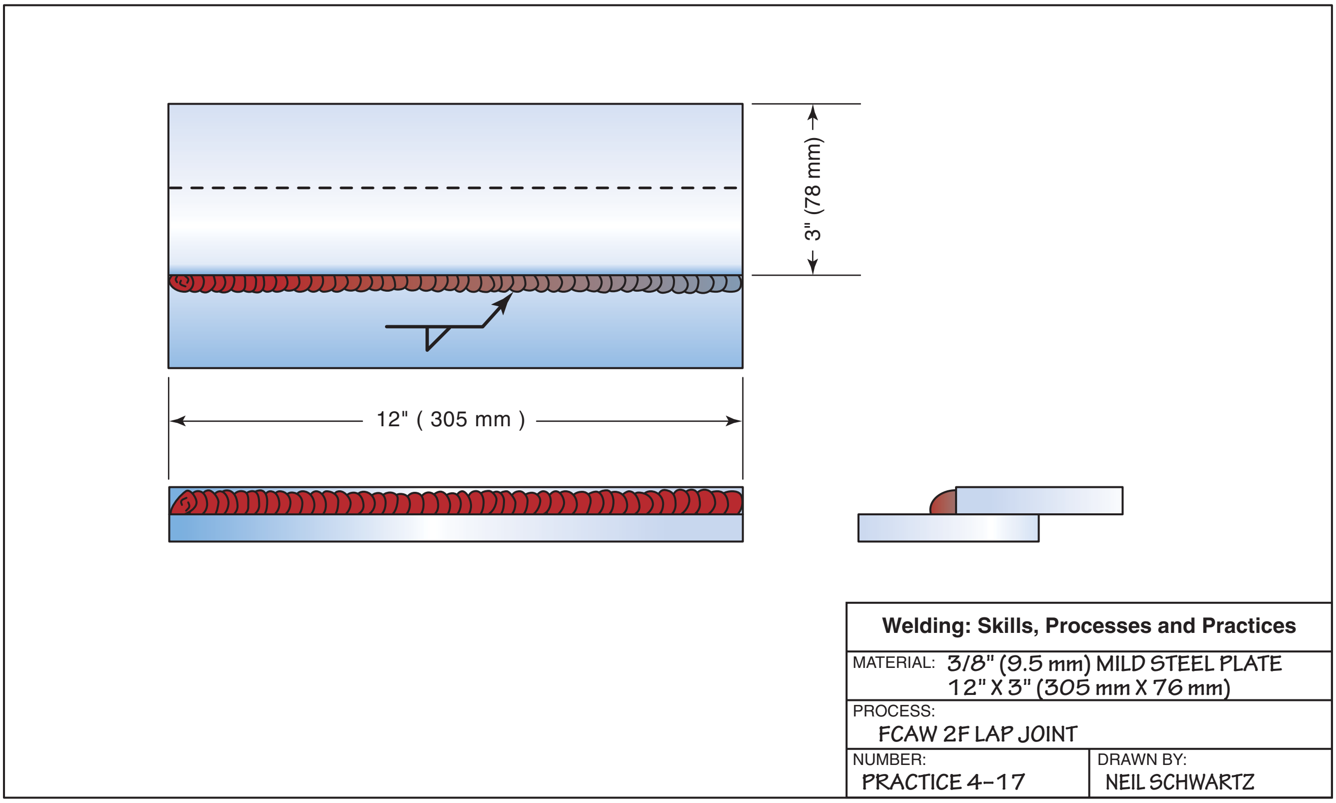

Welding: Skills, Processes and Practices
MATERIAL: 3/8" (9.5 mm) MILD STEEL PLATE
12" X 3" (305 mm X 76 mm)
PROCESS: FCAW 2F LAP JOINT
NUMBER: PRACTICE 4-17
DRAWN BY: NEIL SCHWARTZ

Figure 4.45 (A) FCAW 2F lap joint, 3/8 in. mild steel

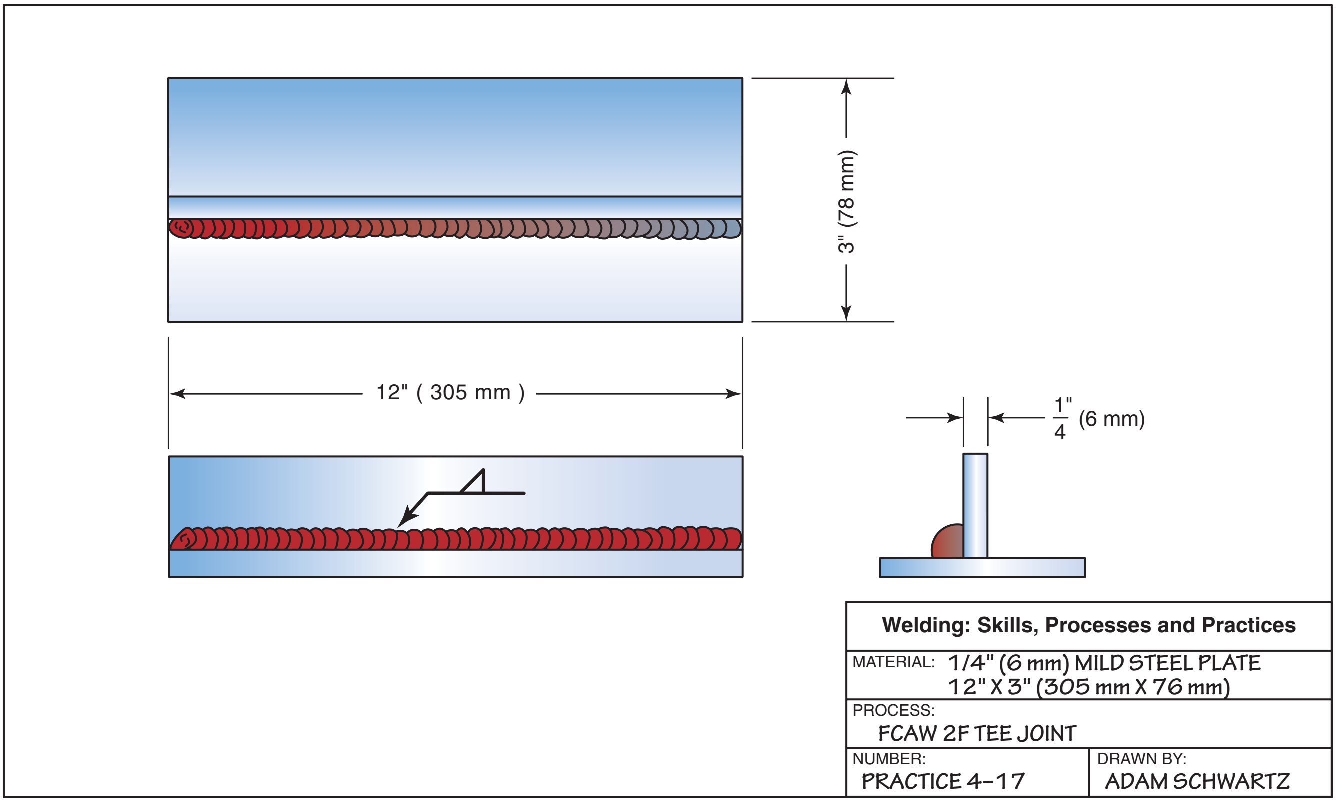

Welding: Skills, Processes and Practices
MATERIAL: 1/4" (6 mm) MILD STEEL PLATE
12" X 3" (305 mm X 76 mm)
PROCESS: FCAW 2F TEE JOINT
NUMBER: PRACTICE 4-17
DRAWN BY: ADAM SCHWARTZ

Figure 4.45 (B) FCAW 2F tee joint, 1/4 in. mild steel

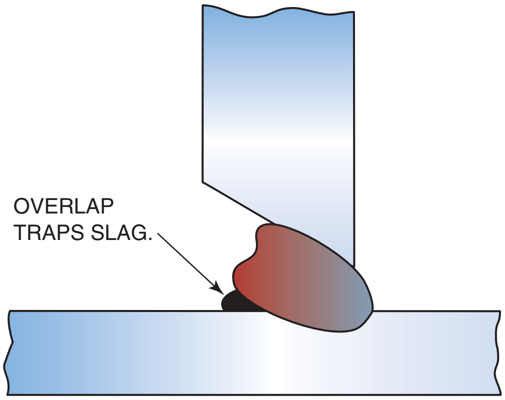

Figure 4.46 Slag can be trapped along the side of the root pass

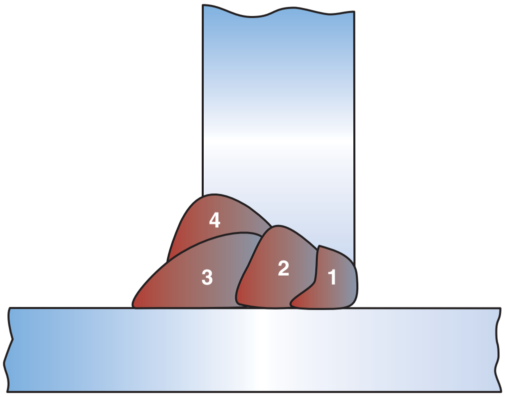

Figure 4.47 FCAW weld bead positions for a 100% penetration grooved tee joint

The root weld must be kept small so that its contour can be controlled. Too large a root pass can trap slag under overlap along the lower edge of the weld, Figure 4.46. Clean each pass thoroughly before the weld bead is started. Follow the weld bead sequence shown in Figure 4.47. Use stringer beads rather than weave beads in order to eliminate overlap. Keeping all of the weld beads small will help control their contour.

Repeat each type of joint with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10
This practice addresses the "Horizontal" component of the all-position requirement for 5 and 10.

PRACTICE 4-18

Tee Joint 2F

Use a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 (or E71T-1 and/or E70T-5) electrodes of diameter 0.045 in. and/or through 1/16 in. (0.9 mm and/or through 1.6 mm), and one or more pieces of mild steel, beveled, 7 in. (178 mm) long and 3/4 in. (19 mm) thick or thicker. You will make a fillet weld in the horizontal position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-17, repeat each type of joint with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Horizontal" position portion of the all-position requirement of 6 and 11.

PRACTICE 4-19

Butt Joint 2G

Use a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 (or E71T-1 and/or E70T-5 electrodes) of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm),

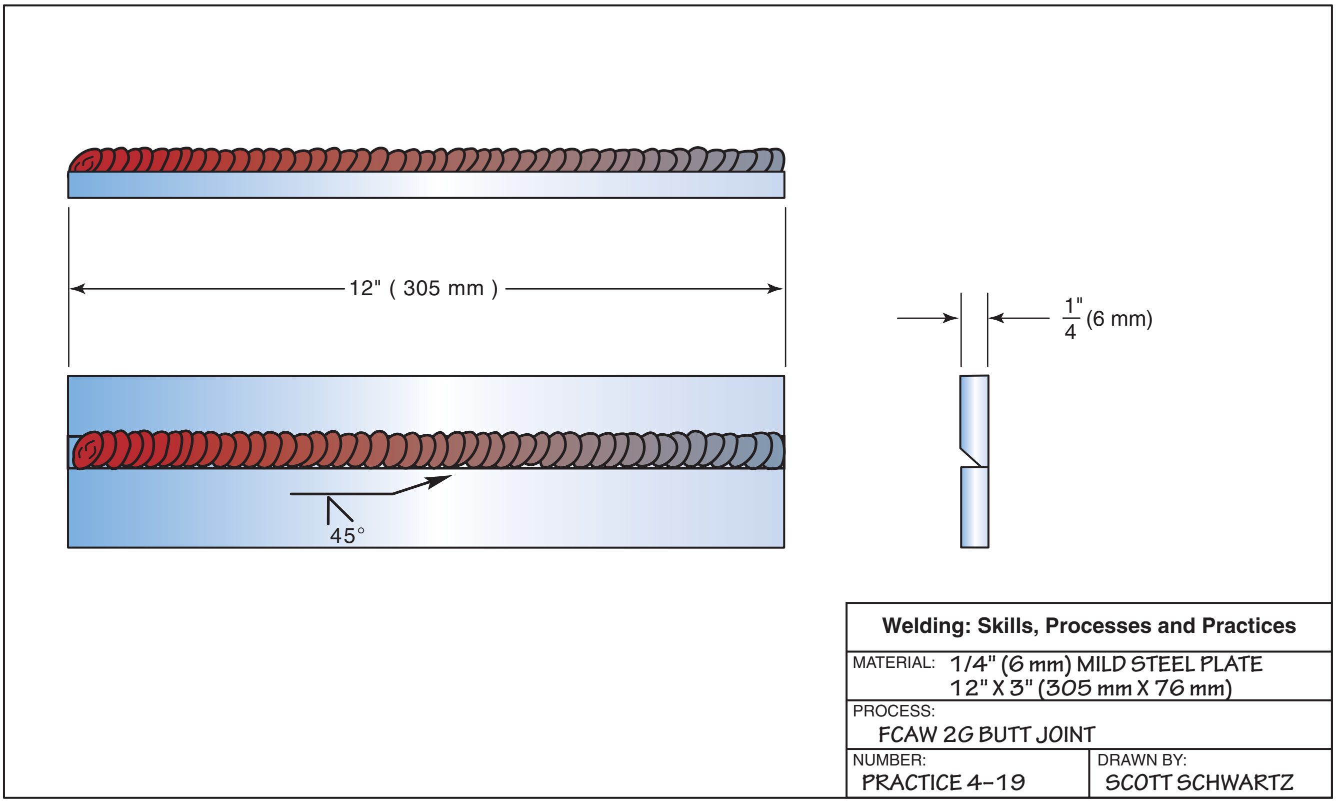
12" (305 mm)
45°
1"/4 (6 mm)
Welding: Skills, Processes and Practices
MATERIAL: 1/4" (6 mm) MILD STEEL PLATE
12" X 3" (305 mm X 76 mm)
PROCESS:
FCAW 2G BUTT JOINT
NUMBER:
PRACTICE 4-19
DRAWN BY:
SCOTT SCHWARTZ

Figure 4.48 **FCAW 2G butt joint, 1/4 in. mild steel**

and one or more pieces of mild steel plate, 12 in. (305 mm) long and 1/4 in. (6 mm) thick or thinner. You will make a groove weld in the horizontal position, Figure 4.48.

Repeat with both classifications of electrodes until defect-free welds can consistently be made in the 1/4-in. (6-mm)-thick plate. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-20

Butt Joint 2G

Use a properly set up and adjusted FCA welding machine; proper safety protection; E70T-1 and/or E71T-11 or 5 (or E71T-1 and/or E70T-5) electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick; and a backing strip 14 in. (355 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the horizontal position.

Repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Horizontal" position portion of the all-position requirement of 6 and 11.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the “Horizontal” position portion of the all-position requirement of 6 and 11.

PRACTICE 4-21

Butt Joint 2G 100% to Be Tested

Use a properly set up and adjusted FCA welding machine; proper safety protection; E70T-1 and/or E71T-11 or 5 (or E71T-1 and/or E70T-5) electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick; and a backing strip 14 in. (355 mm) long, 1 in. (25 mm) wide and 1/4 in. (6 mm) thick. You will make a groove weld in the horizontal position.

Repeat the weld using each electrode classification until you can make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Self Shielded
Key Indicator 6
Gas Shielded
Key Indicator 11
This practice addresses the “Horizontal” position portion of the all-position requirement of 6 and 11.

PRACTICE 4-22

Butt Joint 2G

Use a properly set up and adjusted FCA welding machine; proper safety protection; E70T-1 and/or E71T-11 electrodes (or E71T-1 and/or E70T-5) of diameter 0.035 in. and/or through 1/16 in. (0.9 mm and/or through 1.6 mm); one or more pieces of mild steel plate, beveled, 7 in. (178 mm) long and 3/4 in. (19 mm) thick or thicker; and a backing strip 9 in. (230 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the horizontal position, Figure 4.49.

Repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the “Horizontal” position portion of the all-position requirement of 6 and 11.

PRACTICE 4-23

Butt Joint 2G 100% to Be Tested

Use a properly set up and adjusted FCA welding machine; proper safety protection; E70T-1 and/or E71T-11 electrodes (or E71T-1 and/or E70T-5) of diameter 0.045 in. and/or through 1/16 in. (0.9 mm and/or through 1.6 mm); one or more pieces of mild steel plate, beveled, 7 in. (178 mm) long and 3/4 in. (19 mm) thick or thicker; and a backing strip 9 in. (230 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the horizontal position.

Repeat the weld until you can use each electrode type to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

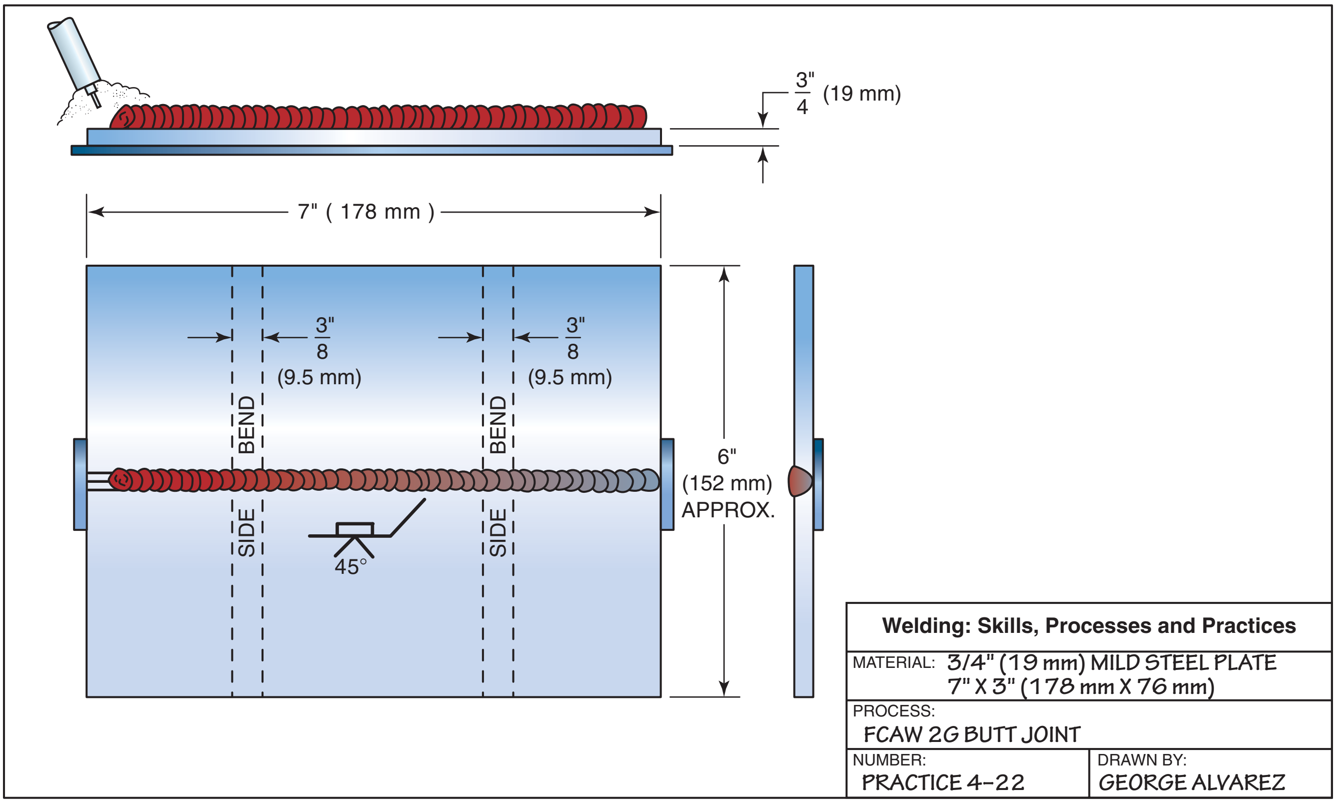

Welding: Skills, Processes and Practices
MATERIAL: 3/4" (19 mm) MILD STEEL PLATE
7" X 3" (178 mm X 76 mm)
PROCESS:
FCAW 2G BUTT JOINT
NUMBER:
PRACTICE 4-22
DRAWN BY:
GEORGE ALVAREZ

Figure 4.49 **FCAW 2G butt joint, 3/4 in. mild steel**

OVERHEAD-POSITION WELDS

PRACTICE 4-24

Butt Joint 4G

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm), and one or more pieces of mild steel plate, 12 in. (305 mm) long and 1/4 in. (6 mm) thick or thinner, you will make a groove weld in the overhead position.

The molten weld pool should be kept as small as possible for easier control. A small molten weld pool can be achieved by using lower current, faster traveling speeds, and settings.

Lower current settings require closer control of gun manipulation to ensure that the electrode is fed into the molten weld pool just behind the leading edge. The low power will cause overlap and more spatter if this electrode-to-molten weld pool contact position is not closely maintained.

Faster travel speeds allow the welder to maintain a high production rate even if multiple passes are required to complete the weld. Weld penetration into the base metal at the start of the bead can be obtained by using a slow start or quickly reversing the weld direction. Both the slow start and reversal of weld direction put more heat into the weld start to increase penetration. The higher speed also reduces the amount of weld distortion by reducing the amount of time that heat is applied to a joint.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Overhead" position portion of the all-position requirement of 6 and 11.

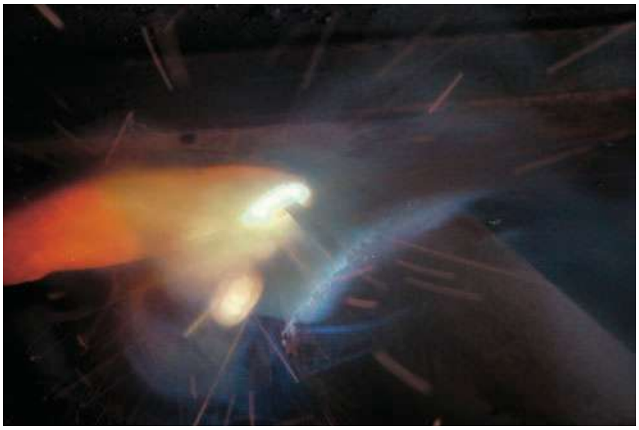

Figure 4.50 Hold the gun so that weld spatter will not fall onto the gun
Source: Courtesy of Larry Jeffus

For overhead welding, extra personal protection is required to reduce the danger of burns. Leather sleeves or leather jackets should be worn.

Much of the spatter created during overhead welding falls into or on the nozzle and contact tube. The contact tube may short out to the gas nozzle. The shorted gas nozzle may arc to the work, causing damage both to the nozzle and to the plate. To control the amount of spatter, a longer stickout and/or a sharper gun-to-plate angle is required to allow most of the spatter to fall clear of the gun or nozzle, Figure 4.50.

Make several short weld beads using various techniques to establish the method that is most successful and most comfortable for you. After each weld, stop and evaluate it before making a change. When you have decided on the technique to be used, make a welded stringer bead that is 12 in. (305 mm) long.

Repeat with both classifications of electrodes until defect-free welds can consistently be made in the 1/4-in. (6-mm)-thick plate. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the "Overhead" position portion of the all-position requirement of 6 and 11.

PRACTICE 4-25

Butt Joint 4G 100% to Be Tested

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm), and one or more pieces of mild steel plate, 12 in. (305 mm) long and 1/4 in. (6 mm) thick, you will make a groove weld in the overhead position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-24, repeat the weld until you can use each electrode type to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-26

Butt Joint 4G

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick; and a backing strip 14 in. (355 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the overhead position, Figure 4.51.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-24, repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Overhead" position portion of the all-position requirement of 6 and 11.

PRACTICE 4-27

Butt Joint 4G 100% to Be Tested

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/8 in. (9.5 mm) thick; and a backing strip 14 in. (355 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the overhead position.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the "Overhead" position portion of the all-position requirement of 6 and 11.

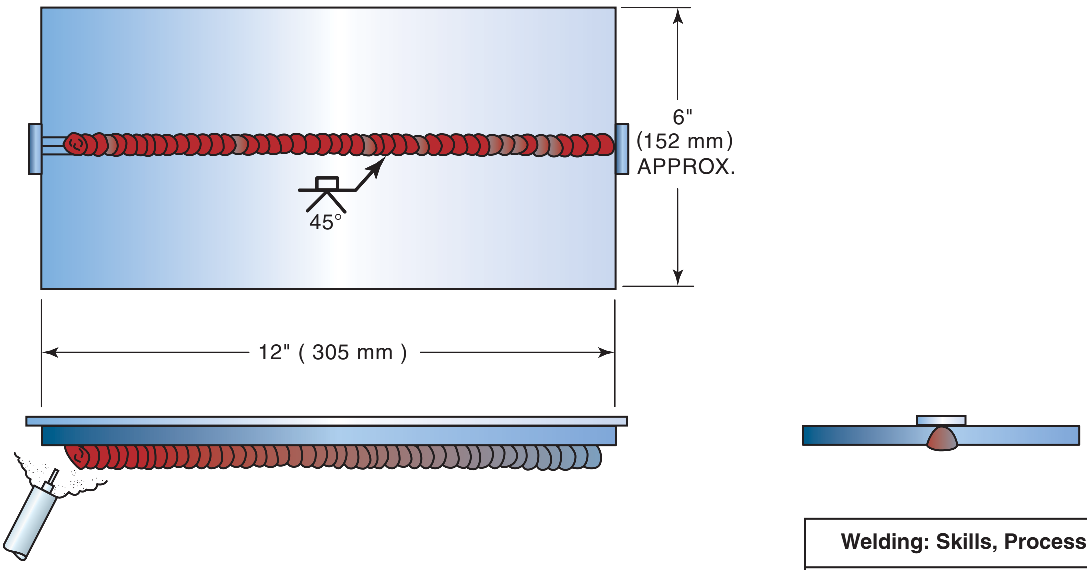

Welding: Skills, Processes and Practices	
MATERIAL: 3/8" (9.5 mm) MILD STEEL PLATE 12" X 3" (305 mm X 76 mm)	
PROCESS: FCAW 4G BUTT JOINT	
NUMBER: PRACTICE 4–26	DRAWN BY: AMY JEFFUS

Figure 4.51 **FCAW 4G butt joint, 3/8 in. mild steel**

Following the same instructions for the assembly and welding procedure outlined in Practice 4-24, repeat the weld until you can use each electrode type to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Overhead" position portion of the all-position requirement of 6 and 11.

PRACTICE 4-28

Butt Joint 4G

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.045 in. and/or through 1/16 in. (0.9 mm and/or through 1.6 mm); one or more pieces of mild steel plate, beveled, 7 in. (178 mm) long and 3/4 in. (19 mm) thick or thicker; and a backing strip 9 in. (230 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the overhead position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-24, repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the "Overhead" position portion of the all-position requirement of 6 and 11.

PRACTICE 4-29

Butt Joint 4G 100% to Be Tested

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.045 in. and/or through 1/16 in. (0.9 mm and/or through 1.6 mm); one or more pieces of mild steel plate, beveled, 7 in. (178 mm) long and 3/4 in. (19 mm) thick or thicker; and a backing strip 9 in. (230 mm) long, 1 in. (25 mm) wide, and 1/4 in. (6 mm) thick. You will make a groove weld in the overhead position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-24, repeat the weld until you can use each electrode classification to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10
This practice addresses the "Overhead" component of the all-position requirement for 5 and 10.

PRACTICE 4-30

Lap Joint and Tee Joint 4F

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.035 in. and/or 0.045 in. (0.9 mm and/or 1.2 mm); and one or more pieces of mild steel plate, beveled, 12 in. (305 mm) long and 3/4 in. (19 mm) thick. You will make a fillet weld in the overhead position, Figure 4.52A and Figure 4.52B.

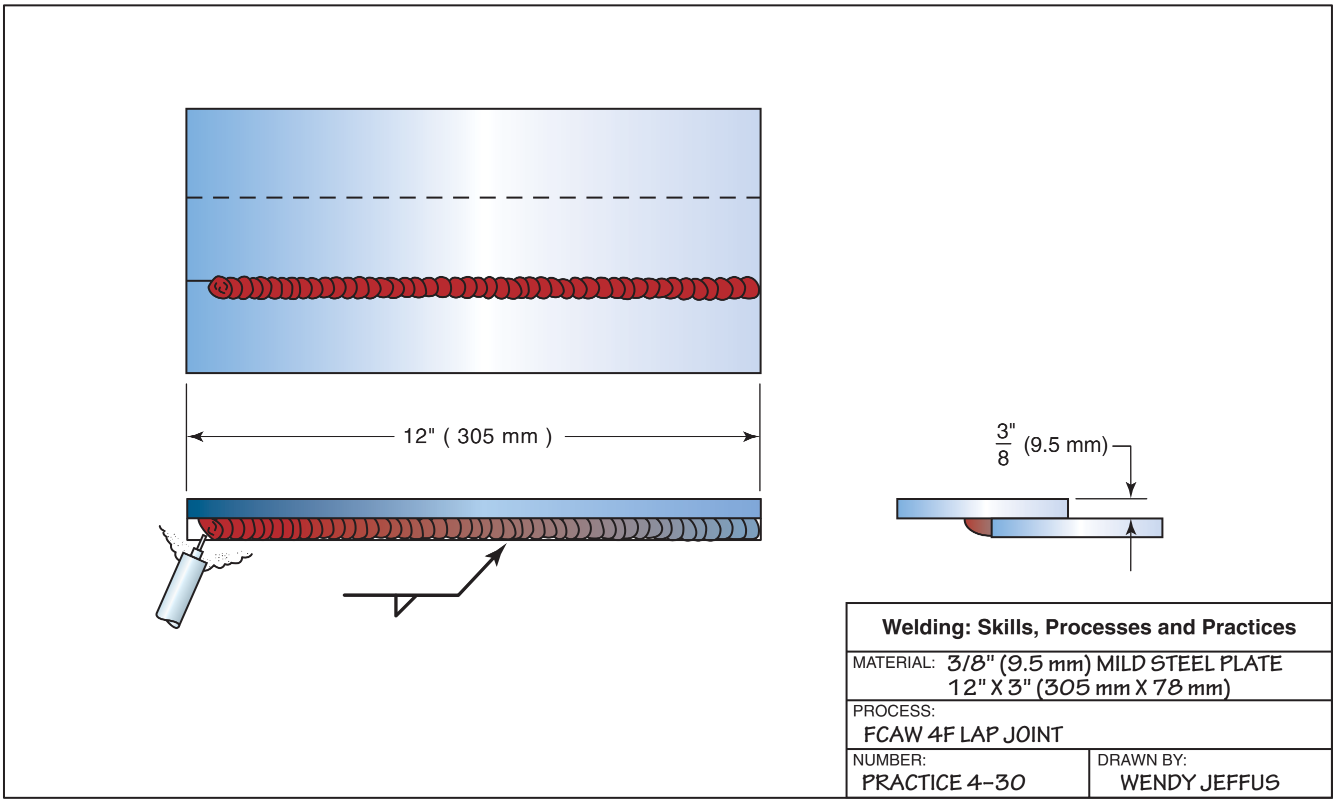
12" (305 mm)
3"/8 (9.5 mm)
Welding: Skills, Processes and Practices
MATERIAL: 3/8" (9.5 mm) MILD STEEL PLATE
12" X 3" (305 mm X 78 mm)
PROCESS:
FCAW 4F LAP JOINT
NUMBER:
PRACTICE 4-30
DRAWN BY:
WENDY JEFFUS

Figure 4.52 (A) FCAW 4F lap joint, 3/8 in. mild steel

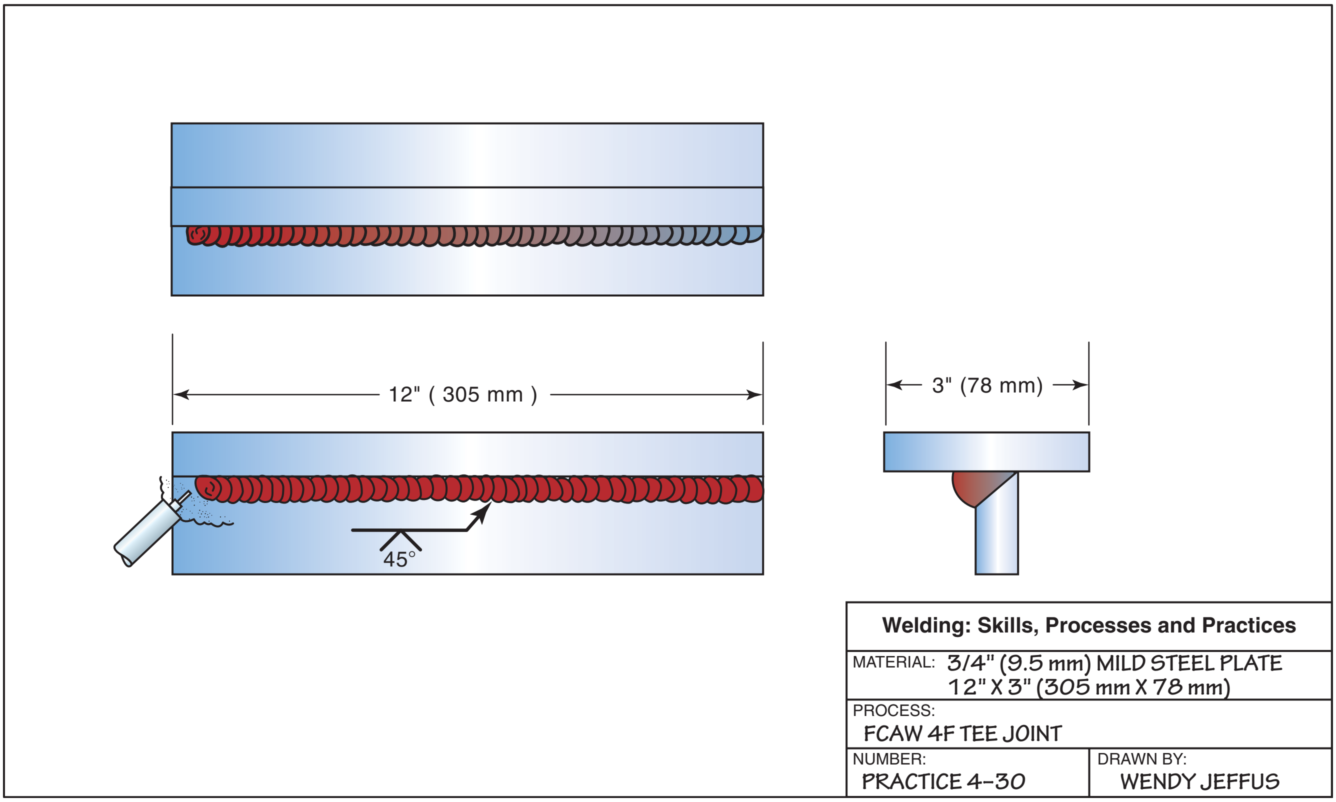
12" (305 mm)
3" (78 mm)
45°
Welding: Skills, Processes and Practices
MATERIAL: 3/4" (9.5 mm) MILD STEEL PLATE
12" X 3" (305 mm X 78 mm)
PROCESS:
FCAW 4F TEE JOINT
NUMBER:
PRACTICE 4-30
DRAWN BY:
WENDY JEFFUS

Figure 4.52 (B) FCAW 4F tee joint, 3/4 in. mild steel

Following the same instructions for the assembly and welding procedure outlined in Practice 4-24, repeat each type of joint with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-31

Tee Joint 4F 100% to Be Tested

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10

Module 9
Key Indicator 1, 2
This practice addresses the "Overhead" component of the all-position requirement for 5 and 10.

Use a properly set up and adjusted FCA welding machine; proper safety protection; E71T-1 and/or E71T-11 electrodes of diameter 0.045 in. and/or through 1/16 in. (0.9-mm and/or through 1.6-mm); and one or more pieces of mild steel plate, beveled, 7 in. (178 mm) long and 3/4 in. (19 mm) thick or thicker. You will make a fillet weld in the overhead position.

Following the same instructions for the assembly and welding procedure as outlined in Practice 4-24, repeat the weld until you can use each electrode classification to make welds with 100% penetration that will pass the bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

THIN-GAUGE WELDING

The introduction of small electrode diameters has allowed FCA welding to be used on thin sheet metal. Usually these welds will be a fillet type, the easiest weld to make on thin stock. An effort should be made when possible to design the weld so it is not a butt-type joint. A common use for FCA welding on thin stock is to join it to a thicker member, Figure 4.53. This type of weld is used to put panels in frames.

The following practices include some butt-type joints. You will find that the vertical down welds are the easiest ones to make. If it is possible to position the weldment for a vertical down weld, production speeds for butt joints can be increased.

PRACTICE 4-32

Butt Joint 1G

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Use a properly set up and adjusted FCA welding machine, Table 4.3; proper safety protection; E71T-1 and/or E70T-5 or E70T-1 and/or E71T-11 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm); and one or more pieces of mild steel sheet, 12 in. (305 mm) long

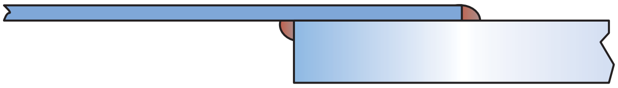

Figure 4.53 FCA welding of thin to thick metal

Table 4.3 **FCA Welding Parameters for Use if Specific Settings Are Unavailable from Electrode Manufacturer**

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	**Size**	**Amps**	**Wire-feed Speed, ipm (cm/min)**	**Volts**	**Type**	**Flow**	**Type**	**Thick**
E70T-1 E71T-1	0.030 in. (0.8 mm)	40 to 145	90 to 340 (228 to 864)	20 to 27	None	n/a	Low-carbon steel	16 gauge to 18 gauge
E70T-1 E71T-1	0.035 in. (0.9 mm)	130 to 200	288 to 576 (732 to 1463)	20 to 28	None	n/a	Low-carbon steel	16 gauge to 18 gauge
E70T-5 E71T-11	0.035 in. (0.9 mm)	90 to 200	190 to 576 (483 to 1463)	16 to 29	57% argon 25% CO_2	35 cfh	Low-carbon steel	16 gauge 18 gauge

and 16-gauge to 18-gauge thick. You will make a butt weld in the flat position, Figure 4.54.

Do not leave a root opening for these welds. Even the slightest opening will result in a burn-through. If a burn-through occurs, the welder can be pulsed off and on so that the hole can be filled. This process will leave a larger than usual buildup. Excessive buildup could be ground off if necessary as part of the postweld cleanup.

Repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I provided by your instructor.

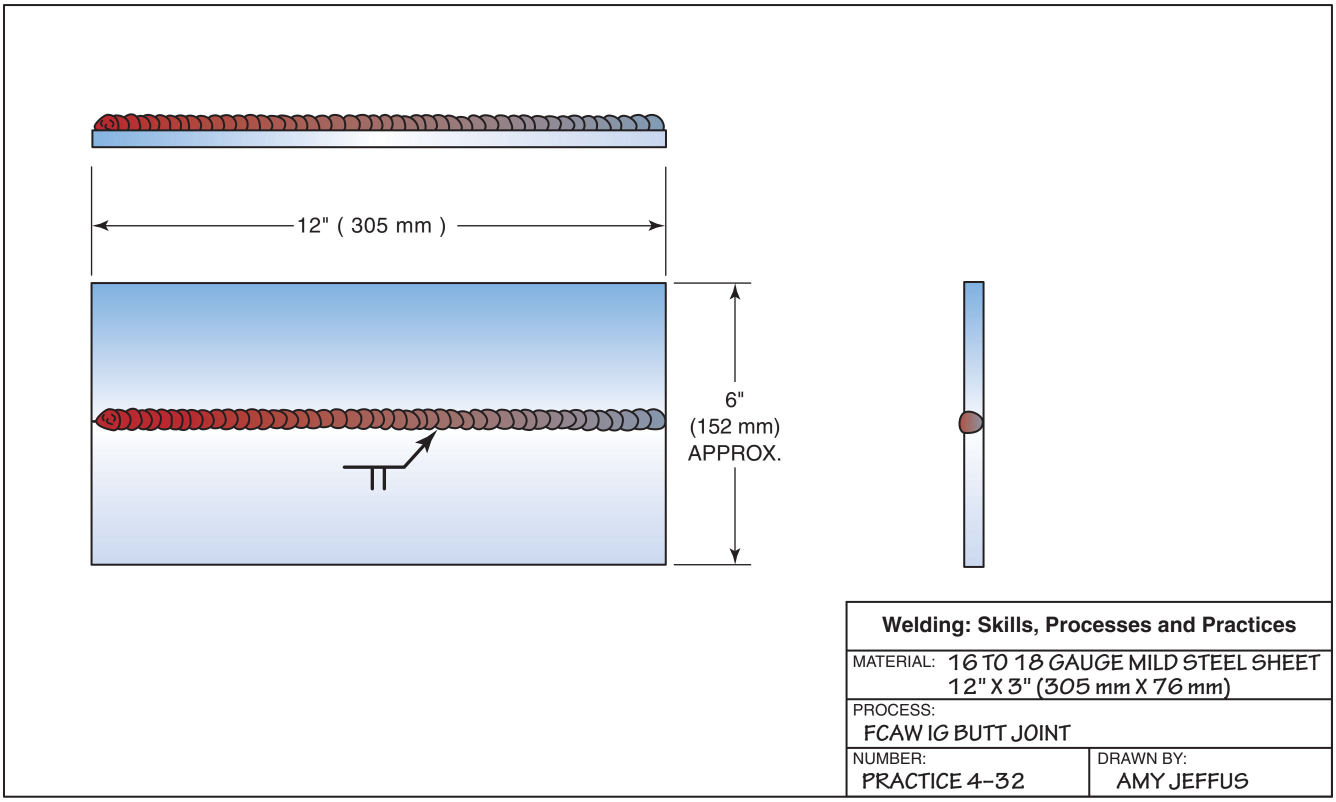

Welding: Skills, Processes and Practices

MATERIAL: 16 TO 18 GAUGE MILD STEEL SHEET 12" X 3" (305 mm X 76 mm)

PROCESS: FCAW IG BUTT JOINT

NUMBER: PRACTICE 4-32

DRAWN BY: AMY JEFFUS

Figure 4.54 **FCAW 1G butt joint, 16- to 18-gauge mild steel**

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2

PRACTICE 4-33

Butt Joint 1G 100% to Be Tested

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E70T-5 or E70T-1 and/or E70T-11 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a butt weld in the flat position, Figure 4.55.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat the weld until you can use each electrode classification to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Gas Shielded
Key Indicator 10
This practice addresses the “Flat” component of the all-position requirement for 5 and 10.

PRACTICE 4-34

Lap Joint and Tee Joint 1F

Using a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 or E71T-1 and/or E70T-5 electrodes of diameter 0.030-in. and/or 0.035-in. (0.8-mm and/or 0.9-mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a fillet weld in the flat position.

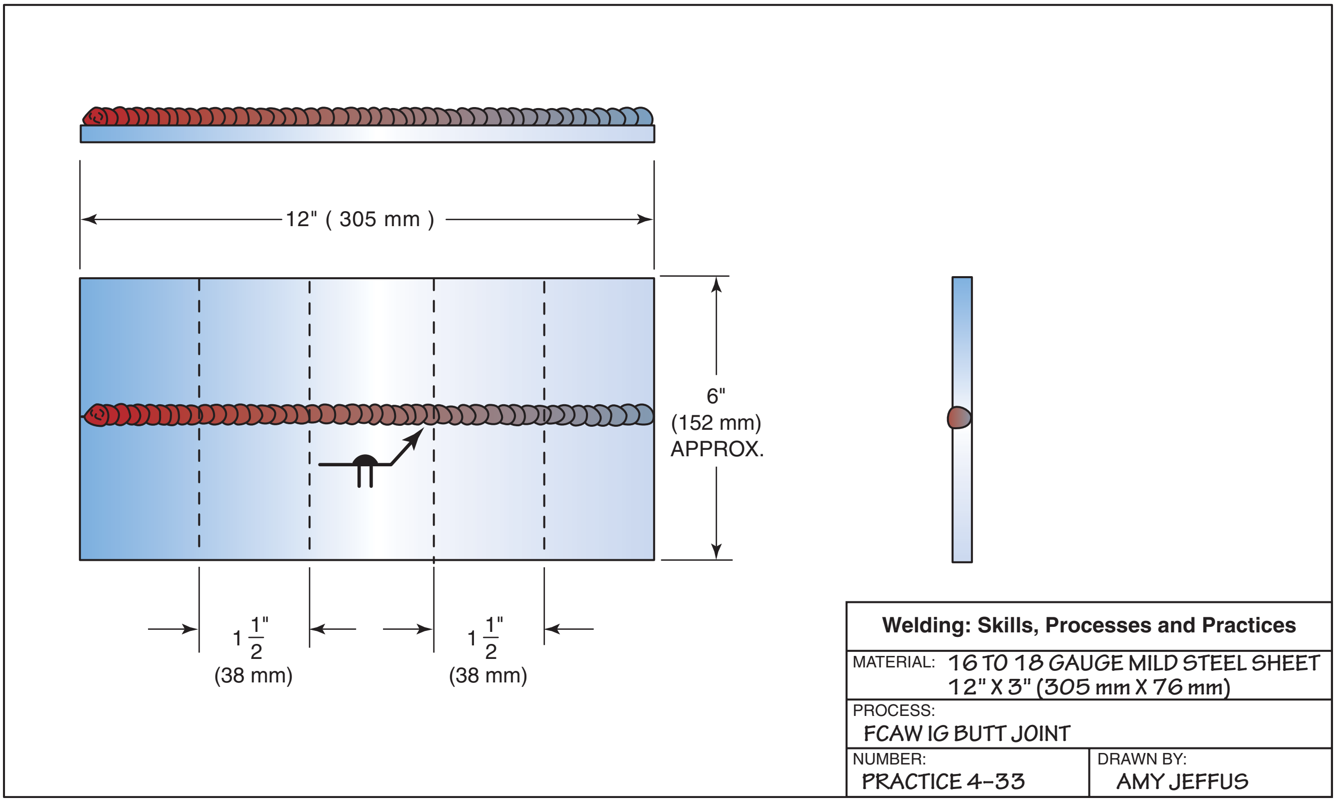
12" (305 mm)
6" (152 mm) APPROX.
1 1/2" (38 mm)
1 1/2" (38 mm)
Welding: Skills, Processes and Practices
MATERIAL: 16 TO 18 GAUGE MILD STEEL SHEET 12" X 3" (305 mm X 76 mm)
PROCESS: FCAW IG BUTT JOINT
NUMBER: PRACTICE 4-33
DRAWN BY: AMY JEFFUS

Figure 4.55 FCAW 1G butt joint, 16- to 18-gauge mild steel

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat each type of joint with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10

Module 9
Key Indicator 1, 2

PRACTICE 4-35

Lap Joint and Tee Joint 1F 100% to Be Tested

Using a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 or E71T-1 and/or E70T-5 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a fillet weld in the flat position, Figure 4.56A and Figure 4.56B.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat each type of joint with both classifications of electrodes until you can make welds with 100% penetration that will pass the bend test, Figure 4.57A and Figure 4.57B. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the "Vertical" position portion of the all-position requirement of 6 and 11.

PRACTICE 4-36

Butt Joint 3G

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-11 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild

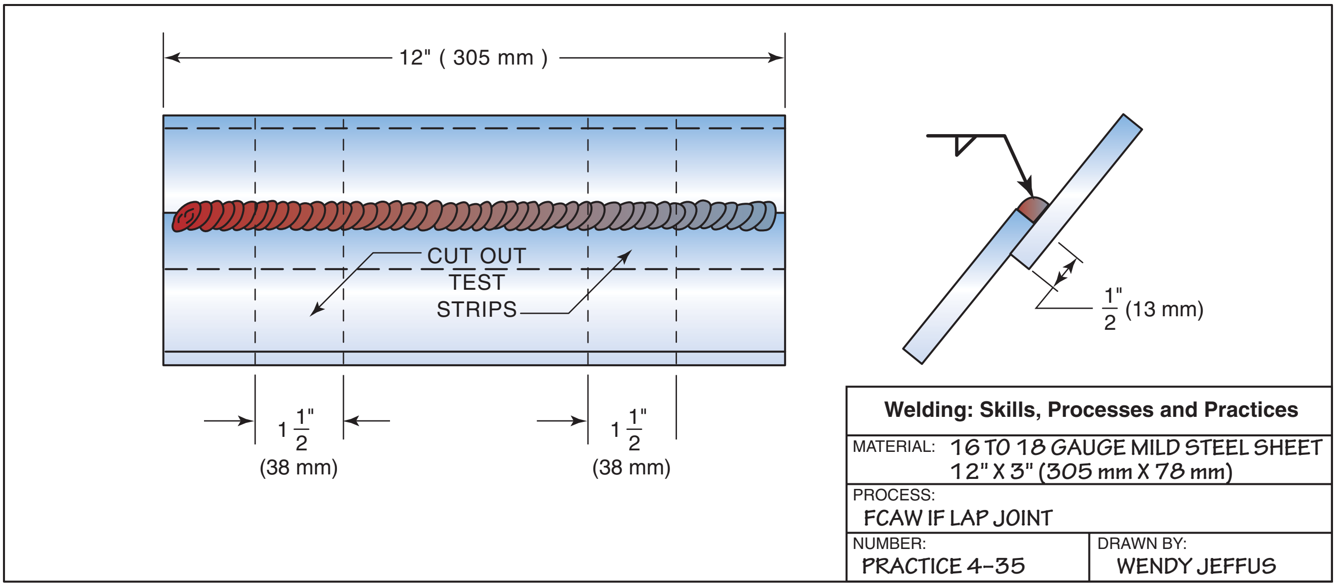

Figure 4.56 **(A) FCAW 1F lap joint, 16- to 18-gauge mild steel**

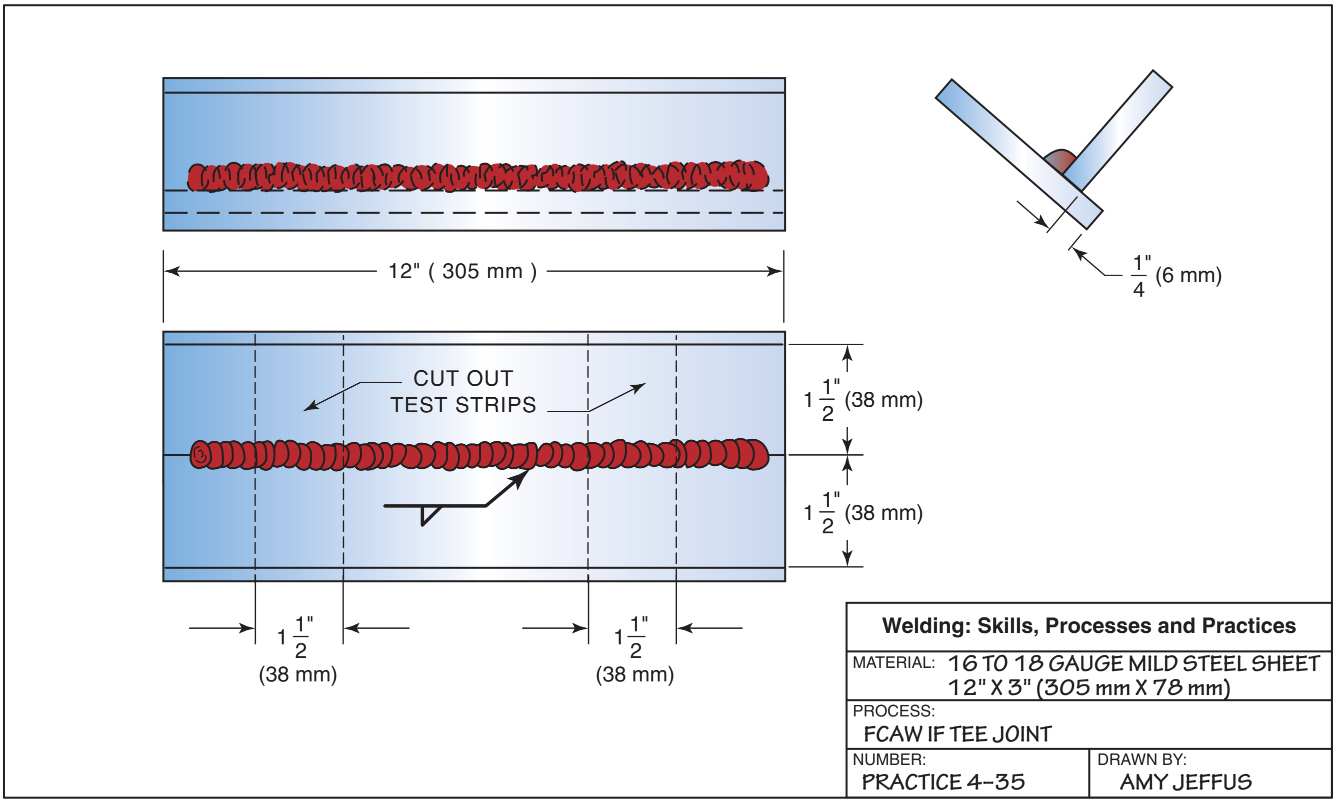
12" (305 mm)
$\frac{1}{4}$" (6 mm)
CUT OUT
TEST STRIPS
1 $\frac{1}{2}$" (38 mm)
1 $\frac{1}{2}$" (38 mm)
1 $\frac{1}{2}$" (38 mm)
1 $\frac{1}{2}$" (38 mm)
Welding: Skills, Processes and Practices
MATERIAL: 16 TO 18 GAUGE MILD STEEL SHEET 12" X 3" (305 mm X 78 mm)
PROCESS: FCAW IF TEE JOINT
NUMBER: PRACTICE 4-35
DRAWN BY: AMY JEFFUS

Figure 4.56 (B) FCAW 1F tee joint, 16- to 18-gauge mild steel

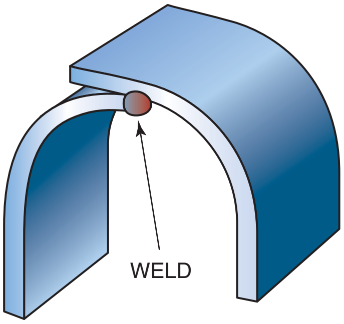

Figure 4.57 (A) 180° bend to test lap weld quality

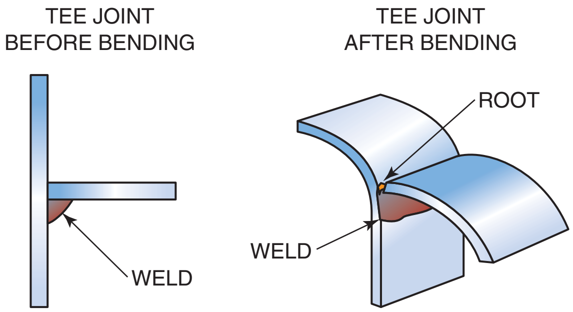
TEE JOINT
BEFORE BENDING
WELD
TEE JOINT
AFTER BENDING
ROOT
WELD

Figure 4.57 (B) Bend the test strip to be sure the weld had good root fusion

steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a butt weld in the vertical up or down position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10
This practice addresses the “Vertical” component of the all-position requirement for 5 and 10.

PRACTICE 4-37

Lap Joint and Tee Joint 3F

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-11 electrodes of diameter 0.030 in.

and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18- gauge thick, you will make a fillet weld in the vertical up or down position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat each type of joint with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10

Module 9
Key Indicator 1, 2
This practice addresses the "Vertical" component of the all-position requirement for 5 and 10.

PRACTICE 4-38

Lap Joint and Tee Joint 3F 100% to Be Tested

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-11 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a fillet weld in the vertical up or down position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat each type of joint with both classifications of electrodes until you can make welds with 100% penetration that will pass the test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10
This practice addresses the "Horizontal" component of the all-position requirement for 5 and 10.

PRACTICE 4-39

Lap Joint and Tee Joint 2F

Using a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a fillet weld in the horizontal position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat each type of joint with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 5
Self Shielded
Key Indicator 10

Module 9
Key Indicator 1, 2
This practice addresses the "Horizontal" component of the all-position requirement for 5 and 10.

PRACTICE 4-40

Lap Joint and Tee Joint 2F 100% to Be Tested

Using a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 or E71T-1 and/or E70T-5 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and

16-gauge to 18-gauge thick, you will make a fillet weld in the horizontal position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat each type of joint with both classifications of electrodes until you can make welds with 100% penetration that will pass the bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the “Horizontal” position portion of the all-position requirement of 6 and 11.

PRACTICE 4-41

Butt Joint 2G

Using a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 or E71T-1 and/or E70T-5 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a butt weld in the horizontal position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the “Horizontal” position portion of the all-position requirement of 6 and 11.

PRACTICE 4-42

Butt Joint 2G 100% to Be Tested

Using a properly set up and adjusted FCA welding machine, proper safety protection, E70T-1 and/or E71T-11 or E71T-1 and/or E70T-5 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a butt weld in the horizontal position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat the weld until you can use each electrode classification to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11
This practice addresses the “Overhead” position portion of the all-position requirement of 6 and 11.

PRACTICE 4-43

Butt Joint 4G

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-5 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a butt weld in the overhead position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat with both classifications of electrodes until defect-free welds can consistently be made. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 4-44

Butt Joint 4G 100% to Be Tested

Using a properly set up and adjusted FCA welding machine, proper safety protection, E71T-1 and/or E71T-5 electrodes of diameter 0.030 in. and/or 0.035 in. (0.8 mm and/or 0.9 mm), and one or more pieces of mild steel sheet, 12 in. (305 mm) long and 16-gauge to 18-gauge thick, you will make a butt weld in the overhead position.

Following the same instructions for the assembly and welding procedure outlined in Practice 4-32, repeat the weld until you can use each electrode classification to make welds with 100% penetration that will pass a bend test. Turn off the welding machine and shielding gas and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Gas Shielded
Key Indicator 6
Self Shielded
Key Indicator 11

Module 9
Key Indicator 1, 2
This practice addresses the "Overhead" position portion of the all-position requirement of 6 and 11.

PRACTICE 4-45

AWS SENSE Entry-Level Welder Workmanship Sample for Flux Cored Arc Welding, Gas-Shielded (FCAW)

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Key Indicator 7

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 4-45. Date:

Title:
Welding FCAW of plate to plate.

Scope:
This procedure is applicable for V-groove, bevel, and fillet welds within the range of 1/8 in. (3.2 mm) through 1-1/2 in. (38 mm).

Welding may be performed in the following positions: all.

Base Metal:
The base metal shall conform to carbon steel M-1, P-1, and S-1, Group 1 or 2.

Backing material specification: none.

Filler Metal:
The filler metal shall conform to AWS specification no. E71T-1 from AWS specification A5.20. This filler metal falls into F-number F-6 and A-number A-1.

Shielding Gas:
The shielding gas, or gases, shall conform to the following compositions and purity:

CO_2 at 30 to 50 cfh or 75% Ar/25% CO_2 at 30 to 50 cfh.

Joint Design and Tolerances:

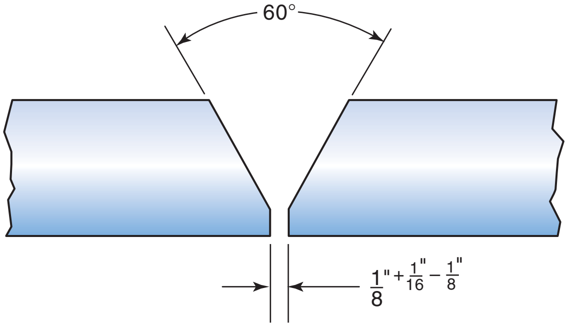

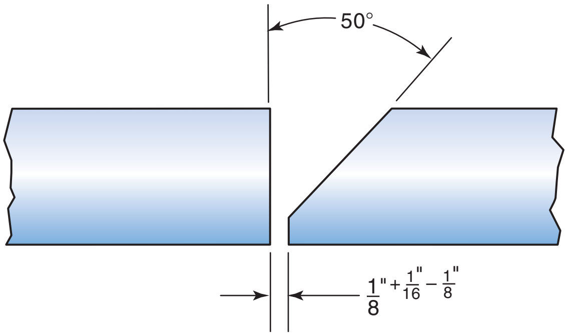

Preparation of Base Metal:

The bevels are to be flame cut on the edges of the plate before the parts are assembled. The beveled surface must be smooth and free of notches. Any roughness or notches deeper than 1/64 in. (0.4 mm) must be ground smooth.

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The groove face and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically cleaned of slag, rust, and mill scale. Cleaning must be done with a wire brush or grinder down to bright metal.

Electrical Characteristics:

The current shall be direct current electrode positive (DCEP). The base metal shall be on the negative side of the line.

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	Size	Amps	Wire-feed Speed, ipm (cm/min)	Volts	Type	Flow	Type	Thickness
E71T-1	0.035 in. (0.9 mm)	130 to 150	288 to 380 (732 to 975)	22 to 25	CO_2 or 75% Ar/ CO_2 25%	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E71T-1	0.045 in. (1.2 mm)	150 to 210	200 to 300 (508 to 762)	28 to 29	CO_2 or 75% Ar/ CO_2 25%	30 to 50	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)

Preheat:

The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:

N/A

Safety:
Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:
Using a 1/2-in. (13-mm) or larger gas nozzle and a distance from contact tube to work of approximately 3/4 in. (19 mm) for all welding, first tack weld the plates together according to Figure 4.58. There should be a root gap of about 1/8 in. (3.2 mm) between the plates with V-grooved or beveled edges. Use an E71T-1 arc welding electrode to make a root pass to fuse the plates together. Clean the slag from the root pass, being sure to remove any trapped slag along the sides of the weld.

Using an E71T-1 arc welding electrode, make a series of stringer or weave filler welds, no thicker than 1/4 in. (6.4 mm), in the groove until the joint is filled. The 1/4-in. (6.4-mm) fillet welds are to be made with one pass.

Interpass Temperature:
The plate should not be heated to a temperature higher than 350°F (175° C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:
The slag must be cleaned off between passes. The weld beads may be cleaned by a hand wire brush, a hand chipping, a punch and hammer, or a needle-scaler. All weld cleaning must be performed with the test plate in the welding position. A grinder may not be used to remove weld control problems such as undercut, overlap, or trapped slag.

Inspection:
Visually inspect the weld for uniformity and discontinuities. There shall be no cracks, no incomplete fusion, and no overlap. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm). The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length, and the maximum diameter shall not exceed 3/32 in. (2.4 mm).

Sketches:
See Figure 4.58.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 4-46

AWS SENSE Entry-Level Welder Workmanship Sample for Flux Cored Arc Welding Self-Shielded (FCAW)

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 7

Module 6
Key Indicator 12

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 4-46 Date:

Title:
Welding FCAW of plate to plate.

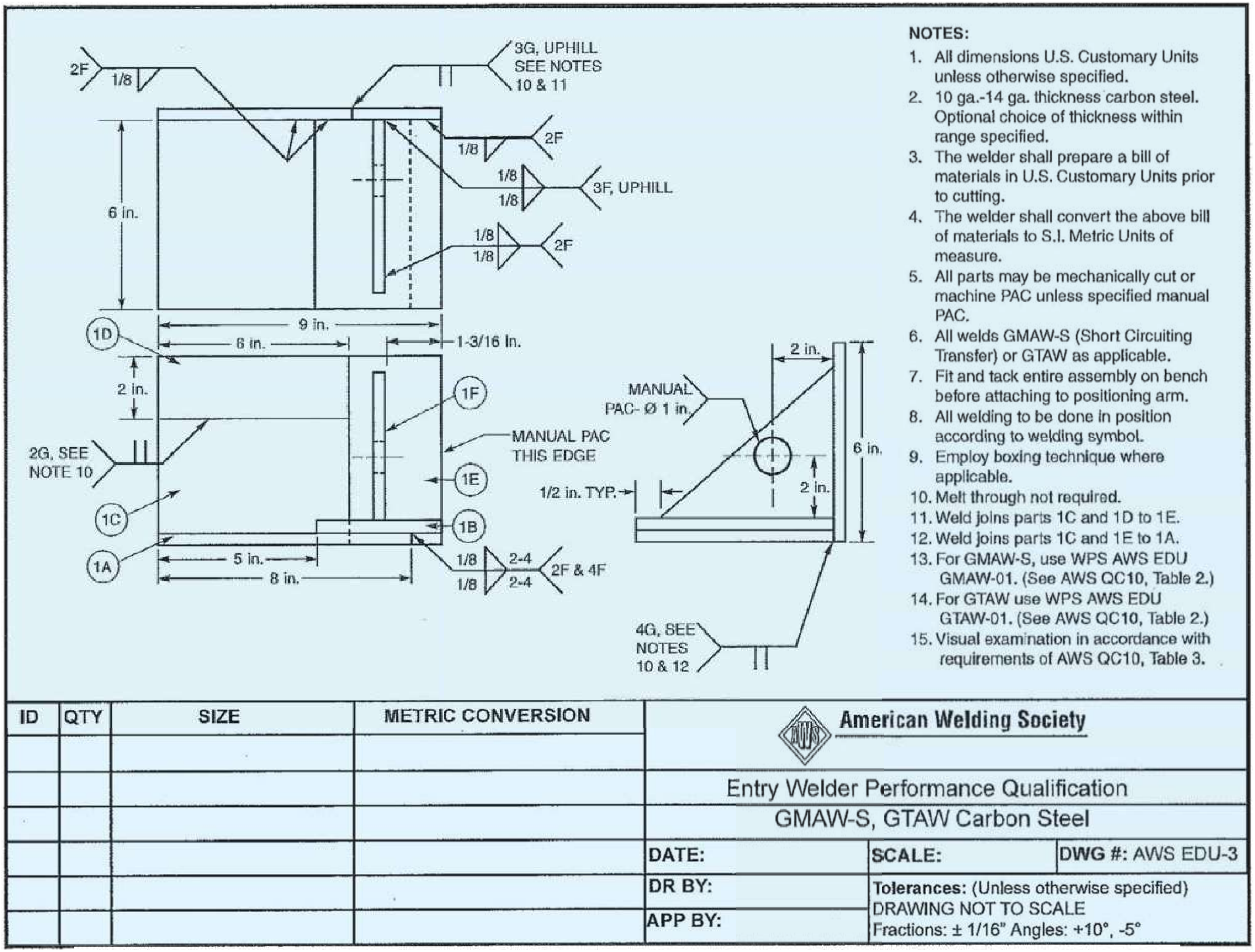
2F
1/8
3G, UPHILL
SEE NOTES
10 & 11
1/8
2F
1/8
1/8
3F, UPHILL
1/8
1/8
2F
6 in.
9 in.
6 in.
1-3/16 in.
1D
2 in.
1F
2G, SEE
NOTE 10
MANUAL PAC
THIS EDGE
1E
1C
1B
1A
5 in.
8 in.
1/8
2-4
1/8
2-4
2F & 4F
MANUAL
PAC- Ø 1 in.
2 in.
6 in.
1/2 in. TYP.
2 in.
4G, SEE
NOTES
10 & 12

NOTES:

1. All dimensions U.S. Customary Units unless otherwise specified.
2. 10 ga.-14 ga. thickness carbon steel. Optional choice of thickness within range specified.
3. The welder shall prepare a bill of materials in U.S. Customary Units prior to cutting.
4. The welder shall convert the above bill of materials to S.I. Metric Units of measure.
5. All parts may be mechanically cut or machine PAC unless specified manual PAC.
6. All welds GMAW-S (Short Circuiting Transfer) or GTAW as applicable.
7. Fit and tack entire assembly on bench before attaching to positioning arm.
8. All welding to be done in position according to welding symbol.
9. Employ boxing technique where applicable.
10. Melt through not required.
11. Weld joins parts 1C and 1D to 1E.
12. Weld joins parts 1C and 1E to 1A.
13. For GMAW-S, use WPS AWS EDU GMAW-01. (See AWS QC10, Table 2.)
14. For GTAW use WPS AWS EDU GTAW-01. (See AWS QC10, Table 2.)
15. Visual examination in accordance with requirements of AWS QC10, Table 3.

ID	QTY	SIZE	METRIC CONVERSION

American Welding Society

Entry Welder Performance Qualification
GMAW-S, GTAW Carbon Steel

DATE:	**SCALE:**	**DWG #:** AWS EDU-3
DR BY:	**Tolerances:** (Unless otherwise specified) DRAWING NOT TO SCALE Fractions: ± 1/16" Angles: +10°, -5°	
APP BY:		

Figure 4.58 **FCAW-G workmanship qualification test**
Source: Courtesy of the American Welding Society

Scope:
This procedure is applicable for V-groove, bevel, and fillet welds within the range of 1/8 in. (3.2 mm) through 1 1/2 in. (38 mm).

Welding may be performed in the following positions: all.

Base Metal:
The base metal shall conform to carbon steel M-1, P-1, and S-1, Group 1 or 2.

Backing material specification: none.

Filler Metal:
The filler metal shall conform to AWS specification no. 0.035 to 0.0415 dia. E71T-11 from AWS specification A5.20. This filler metal falls into F-number F-6 and A-number A-1.

Shielding Gas:
The shielding gas, or gases, shall conform to the following compositions and purity: none.

Joint Design and Tolerances:

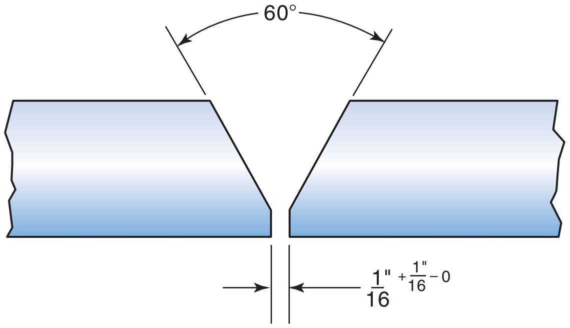

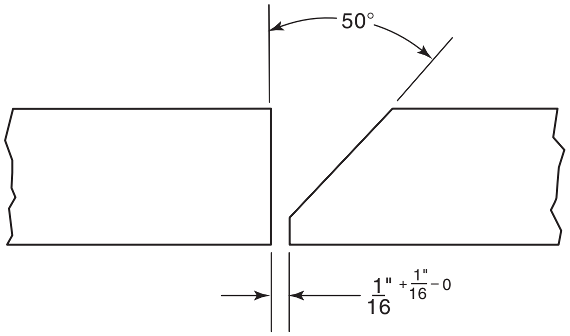

Preparation of Base Metal:
The bevels are to be flame cut on the edges of the plate before the parts are assembled. The beveled surface must be smooth and free of notches. Any roughness or notches deeper than 1/64 in. (0.4 mm) must be ground smooth.

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The groove face and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically cleaned of slag, rust, and mill scale. Cleaning must be done with a wire brush or grinder down to bright metal.

Electrical Characteristics:
The current shall be direct current electrode negative (DCEN). The base metal shall be on the positive side of the line.

Electrode		Welding Power			Shielding Gas		Base Metal	
Type	Size	Amps	Wire-feed Speed, ipm (cm/min)	Volts	Type	Flow	Type	Thickness
E71T-11	0.035 in. (0.9 mm)	130 to 150	150 to 225 (381 to 571)	22 to 25	None	—	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)
E71T-11	0.045 in. (1.2 mm)	150 to 210	105 to 195 (266 to 495)	15 to 18	None	—	Low-carbon steel	1/4 in. to 1/2 in. (6 mm to 13 mm)

Preheat:
The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:
N/A

Safety:
Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:
Using a 1/2-in. (13-mm) or larger gas nozzle and a distance from contact tube to work of approximately 3/4 in. (19 mm) for all welding, first tack weld the plates together according to Figure 4.59. There should be a root gap of about 1/8 in. (3.2 mm) between the plates with V-grooved or beveled edges. Use an E71T-11 arc welding electrode to make a root pass to fuse the plates together. Clean the slag from the root pass, being sure to remove any trapped slag along the sides of the weld.

Using an E71T-11 arc welding electrode, make a series of stringer or weave filler welds, no thicker than 1/4 in. (6.4 mm), in the groove until the joint is filled. The 1/4-in. (6.4-mm) fillet welds are to be made with one pass.

Interpass Temperature:
The plate should not be heated to a temperature higher than 350°F (175°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:
The slag must cleaned off between passes. The weld beads may be cleaned by a hand wire brush, a hand chipping, a punch and hammer, or a needle-scaler. All weld cleaning must be performed with the test plate in the welding position. A grinder may not be used to remove weld control problems such as undercut, overlap, or trapped slag.

Inspection:
Visually inspect the weld for uniformity and discontinuities. There shall be no cracks, no incomplete fusion, and no overlap. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm). The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length, and the maximum diameter shall not exceed 3/32 in. (2.4 mm).

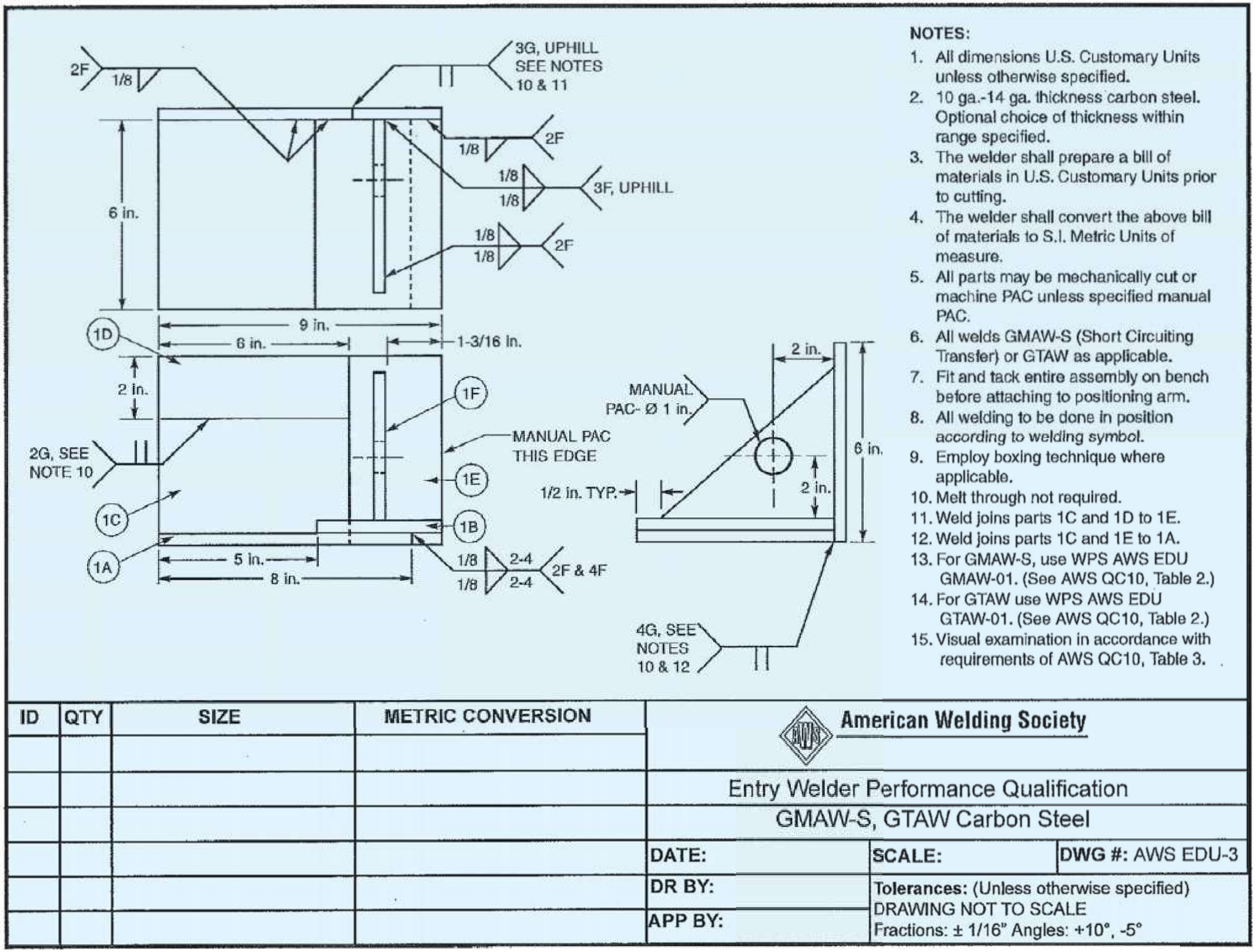
2F
1/8
3G, UPHILL
SEE NOTES
10 & 11
1/8
2F
1/8
1/8
3F, UPHILL
1/8
1/8
2F
6 in.
9 in.
6 in.
1-3/16 in.
1D
2 in.
1F
MANUAL PAC
THIS EDGE
2G, SEE
NOTE 10
1E
1C
1B
1A
5 in.
8 in.
1/8
1/8
2-4
2-4
2F & 4F
MANUAL
PAC- Ø 1 in.
2 in.
6 in.
2 in.
1/2 in. TYP.
4G, SEE
NOTES
10 & 12

NOTES:
1. All dimensions U.S. Customary Units unless otherwise specified.
2. 10 ga.-14 ga. thickness carbon steel. Optional choice of thickness within range specified.
3. The welder shall prepare a bill of materials in U.S. Customary Units prior to cutting.
4. The welder shall convert the above bill of materials to S.I. Metric Units of measure.
5. All parts may be mechanically cut or machine PAC unless specified manual PAC.
6. All welds GMAW-S (Short Circuiting Transfer) or GTAW as applicable.
7. Fit and tack entire assembly on bench before attaching to positioning arm.
8. All welding to be done in position *according to welding symbol.*
9. Employ boxing technique where applicable.
10. Melt through not required.
11. Weld joins parts 1C and 1D to 1E.
12. Weld joins parts 1C and 1E to 1A.
13. For GMAW-S, use WPS AWS EDU GMAW-01. (See AWS QC10, Table 2.)
14. For GTAW use WPS AWS EDU GTAW-01. (See AWS QC10, Table 2.)
15. Visual examination in accordance with requirements of AWS QC10, Table 3.

ID	QTY	SIZE	METRIC CONVERSION

American Welding Society

Entry Welder Performance Qualification

GMAW-S, GTAW Carbon Steel

DATE:

SCALE:

DWG #: AWS EDU-3

DR BY:

APP BY:

Tolerances: (Unless otherwise specified)
DRAWING NOT TO SCALE
Fractions: ± 1/16" Angles: +10°, -5°

Figure 4.59 FCAW-G workmanship qualification test

Source: Courtesy of the American Welding Society

We Support AWS

SENSE

WELDING

Skills, Processes and Practices for Entry-Level Welders

BOOK 3

- Shielded Metal Arc Welding
- Gas Tungsten Arc Welding

LARRY JEFFUS • LAWRENCE BOWER

Welding

Skills, Processes and Practices for Entry-Level Welders

Book 3

- Shielded Metal Arc Welding
- Gas Tungsten Arc Welding

Larry Jeffus
Lawrence Bower

DELMAR
CENGAGE Learning™

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

DELMAR
CENGAGE Learning™

Welding Skills, Processes and Practices for Entry-Level Welders: Book Three
Larry Jeffus/Larry Bower

Vice President, Editorial: Dave Garza

Director of Learning Solutions: Sandy Clark

Executive Editor: David Boelio

Managing Editor: Larry Main

Senior Product Manager: Sharon Chambliss

Editorial Assistant: Lauren Stone

Vice President, Marketing: Jennifer McAvey

Executive Marketing Manager: Deborah S. Yarnell

Senior Marketing Manager: Jimmy Stephens

Marketing Specialist: Mark Pierro

Production Director: Wendy Troeger

Production Manager: Mark Bernard

Content Project Manager: Cheri Plasse

Art Director: Benj Gleeksman

Technology Project Manager: Christopher Catalina

Production Technology Analyst: Thomas Stover

© 2010 Delmar, Cengage Learning

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, Web distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the publisher.

For product information and technology assistance, contact us at
Professional & Career Group Customer Support, 1-800-648-7450
For permission to use material from this text or product,
submit all requests online at **cengage.com/permissions.**
Further permissions questions can be e-mailed to
permissionrequest@cengage.com.

Library of Congress Control Number: 2009920893

ISBN-13: 978-1-4354-2796-9
ISBN-10: 1-4354-2796-3

Delmar
5 Maxwell Drive
Clifton Park, NY 12065-2919
USA

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your lifelong learning solutions, visit **delmar.cengage.com**
Visit our corporate website at **cengage.com.**

Notice to the Reader

Publisher does not warrant or guarantee any of the products described herein or perform any independent analysis in connection with any of the product information contained herein. Publisher does not assume, and expressly disclaims, any obligation to obtain and include information other than that provided to it by the manufacturer. The reader is expressly warned to consider and adopt all safety precautions that might be indicated by the activities described herein and to avoid all potential hazards. By following the instructions contained herein, the reader willingly assumes all risks in connection with such instructions. The publisher makes no representations or warranties of any kind, including but not limited to, the warranties of fitness for particular purpose or merchantability, nor are any such representations implied with respect to the material set forth herein, and the publisher takes no responsibility with respect to such material. The publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or part, from the readers' use of, or reliance upon, this material.

Printed in Canada
1 2 3 4 5 XX 11 10 09

Brief Contents

Preface. ix

AWS Acknowledgment xii

Acknowledgments xiii

About the Authors xiv

Chapter 1 Shielded Metal Arc Equipment, Setup, and Operation 1

Chapter 2 Shielded Metal Arc Welding of Plate. 25

Chapter 3 Advanced Shielded Metal Arc Welding 69

Chapter 4 Gas Tungsten Arc Welding Equipment, Setup, Operation, and Filler Metals. 105

Chapter 5 Gas Tungsten Arc Welding of Plate 141

Appendix 193

Glossary 195

Index 201

This page intentionally left blank

Contents

Preface . . . ix
AWS Acknowledgment . . . xii
Acknowledgments . . . xiii
About the Authors . . . xiv

Chapter 1 Shielded Metal Arc Equipment, Setup, and Operation . . . 1

- *Introduction* . . . 2
- *Welding Current* . . . 3
  - Measurement . . . 4
  - Temperature . . . 4
  - Currents . . . 5
- *Types of Welding Power* . . . 6
- *Open Circuit Voltage* . . . 7
- *Operating Voltage* . . . 7
- *Arc Blow* . . . 7
- *Types of Power Sources* . . . 9
  - Transformers . . . 9
  - Multiple-coil Machines . . . 10
  - Movable-coil or Movable-core Machines . . . 12
  - Inverter Machines . . . 13
- *Generators and Alternators* . . . 13
- *Rectifiers* . . . 15
- *Duty Cycle* . . . 17
- *Welding Cables and Leads* . . . 18
- *Electrode Holders* . . . 20
- *Work Clamps* . . . 21
- *Setup* . . . 22
- *Summary* . . . 23
- *Review* . . . 23

Chapter 2 Shielded Metal Arc Welding of Plate . . . 25

- *Introduction* . . . 26
- *Effect of Too-High or Too-Low Current Settings* . . . 29
- *Electrode Size and Heat* . . . 31
- *Arc Length* . . . 32
- *Electrode Angle* . . . 34

Leading Angle 34
Trailing Angle 35
Electrode Manipulation 37
Positioning of the Welder and the Plate 40
Practice Welds 40
Electrodes 40
Stringer Beads 41
Square Butt Joint 44
Edge Weld 49
Outside Corner Joint 54
Lap Joint 58
Tee Joint 63
Summary 66
Review 67

Chapter 3 Advanced Shielded Metal Arc Welding 69

Introduction 70
Root Pass 71
Hot Pass 79
Filler Pass 81
Cover Pass 84
Plate Preparation 85
Preparing Specimens for Testing 88
Acceptance Criteria for Face Bends and Root Bends 88
Restarting a Weld Bead 89
Preheating and Postheating 92
Poor Fitup 101
Summary 103
Review 103

Chapter 4 Gas Tungsten Arc Welding Equipment, Setup, Operation, and Filler Metals 105

Introduction 106
Tungsten 107
Types of Tungsten Electrodes 110
Pure Tungsten Electrodes, EWP 111
Thoriated Tungsten Electrodes, EWTh-1 and EWTh-2 111
Zirconium Tungsten Electrodes, EWZr 112
Cerium Tungsten Electrodes, EWCe-2 112
Lanthanum Tungsten Electrodes, EWLa-1, EWLa-1.5, EWLa-2 112
Alloy Not Specified, EWG 113
Shaping the Tungsten Electrode 113

Grinding 113
Breaking and Remelting 114
Chemical Cleaning and Pointing 116
Pointing and Remelting 116
GTA Welding Equipment 117
Torches 117
Hoses 118
Nozzles 121
Flowmeter 122
Types of Welding Current 123
AC Balance Control 126
Shielding Gases 129
Argon 129
Helium 130
Hydrogen 130
Nitrogen 130
Hot Start 131
Preflow Time and Postflow Time 131
Gas Flow Rate 132
Remote Controls 133
Summary 138
Review 138

Chapter 5 Gas Tungsten Arc Welding of Plate 141

Introduction 143
Torch Angle 143
Filler Rod Manipulation 144
Tungsten Contamination 146
Current Setting 146
Experiments 147
Gas Flow 149
Practice Welds 151
Low-Carbon and Mild Steels 151
Stainless Steel 152
Aluminum 153
Metal Preparation 154
Summary 191
Review 191

Appendix 193

Glossary 195

Index 201

This page intentionally left blank

Preface

ABOUT THE SERIES

Welding: Skills, Processes and Practices for Entry-Level Welders is an exciting new series that has been designed specifically to support the American Welding Society's (AWS) SENSE EG2.0 training guidelines. Offered in three volumes, these books are carefully crafted learning tools consisting of theory-based texts that are accompanied by companion lab manuals, and extensive instructor support materials. With a logical organization that closely follows the modular structure of the AWS guidelines, the series will guide readers through the process of acquiring and practicing welding knowledge and skills. For schools already in the SENSE program, for those planning to join, or for schools interested in obtaining certifiable outcomes based on nationally recognized industry standards in order to comply with the latest Carl D. Perkins Career and Technical Education Requirements. *Welding: Skills, Processes and Practices for Entry-Level Welders* offers a turnkey solution of high quality teaching and learning aids.

Career and technical education instructors at the high school level are often called upon to be multi-disciplinary educators, teaching welding as only one of as many as five technical disciplines in any given semester. The *Welding: Skills, Processes and Practices for Entry-Level Welders* package provides these educators with a process-based, structured approach and the tools they need to be prepared to deliver high level training on processes and materials with which they may have limited familiarity or experience. Student learning, satisfaction and retention are the target of the logically planned practices, supplements and full color textbook illustrations. While the AWS standards for entry level welders are covered, students are also introduced to the latest in high technology welding equipment such as pulsed gas metal arc welding (GMAW-P). Career pathways and career clusters may be enhanced by the relevant mathematics applied to real world activities as well as oral and written communication skills linked to student interaction and reporting.

Book 1, the core volume, introduces students to the welding concepts covered in AWS SENSE Modules 1, 2, 3, 8 and 9 (Occupational Orientation, Safety and Health of Welders, Drawing and Welding Symbol Interpretation, Thermal Cutting, and Weld Inspection Testing and Codes). Book 1 contains all the material needed for a SENSE program that prepares students for qualification in Thermal Cutting processes. The optional Books 2 and 3 cover other important welding processes and are grouped in logical combinations. Book 2 corresponds to AWS SENSE Modules 5 and 6 (GMAW, FCAW), and Book 3 corresponds to AWS SENSE Modules 4 and 7 (SMAW, GTAW).

The texts feature hundreds of four-color figures, diagrams and tight shots of actual welds to speed beginners to an understanding of the most widely used welding processes.

FEATURES

- Produced in close collaboration with experienced instructors from established SENSE programs to maximize the alignment of the content with SENSE guidelines and to ensure 100% coverage of Level I-Entry Welder Key Indicators.
- Chapter introductions contain general performance objectives, key terms used, and the AWS SENSE EG2.0 Key Indicators addressed in the chapter.
- Coverage of Key Indicators is indicated in the margin by a torch symbol and a numerical reference.
- Contains scores of fully illustrated Practices, which are guided exercises designed to help students master processes and materials. Where applicable, the Practices reproduce and reference actual AWS technical drawings in order to help students create acceptable workmanship samples.
- Each section introduces students to the materials, equipment, setup procedures and critical safety information they need in order to weld successfully.
- Hundreds of four-color figures, diagrams and tight shots of actual welds to speed beginners to an understanding of the most widely used welding processes.
- End of chapter review questions develop critical thinking skills and help students to understand "why" as well as "how."

SUPPLEMENTS

Each book in the Welding Skills series is accompanied by a ***Lab Manual*** that has been designed to provide hands-on practice and reinforce the student's understanding of the concepts presented in the text. Each chapter contains practice exercises to reinforce the primary objectives of the lesson, including creation of workmanship samples (where applicable), and a quiz to test knowledge of the material. Artwork and safety precautions are included throughout the manuals.

Instructor Resources (on CD-ROM), designed to support Books 1–3 and the accompanying Lab Manuals, provide a wealth of time-saving tools, including:

- An Instructor's Guide with answers to end-of-chapter Review Questions in the texts and Lab Manual quizzes.
- Modifiable model Lesson Plans that aid in the design of a course of study that meets local or state standards and also maps to the SENSE guidelines.
- An extensive ExamView Computerized Test Bank that offers assessments in true/false, multiple choice, sentence completion and short answer formats. Test questions have been designed to expose students to the types of questions they'll encounter on the SENSE Level 1 Exams.
- PowerPoint Presentations with selected illustrations that provide a springboard for lectures and reinforce skills and processes covered in the texts. The PowerPoint Presentations can be modified or expanded as instructors desire, and can be augmented with additional illustrations from the Image Library.
- The Image Library contains nearly all (well over 1000!) photographs and line art from the texts, most in four-color.
- A SENSE Correlation Chart that shows the close alignment of the *Welding* series to the SENSE Entry Level 1 training guidelines. Each Key Indicator within each SENSE Module is mapped to the relevant text and lab manual page or pages.

TITLES IN THE SERIES

Welding: Skills, Processes and Practices for Entry-Level Welders: Book 1, Occupational Orientation, Safety and Health of Welders, Drawing and Welding Symbol Interpretation, Thermal Cutting, Weld Inspection Testing and Codes
(Order #: 1-4354-2788-2)
Lab Manual, Book One (Order #: 1-4354-2789-0)

Welding: Skills, Processes and Practices for Entry-Level Welders: Book 2, Gas Metal Arc Welding, Flux Cored Arc Welding (Order #: 1-4354-2790-4)
Lab Manual, Book Two (Order #: 1-4354-2795-5)

Welding: Skills, Processes and Practices for Entry-Level Welders: Book 3, Shielded Metal Arc Welding, Gas Tungsten Arc Welding (Order #: 1-4354-2796-3)
Lab Manual, Book Three (Order #: 1-4354-2797-1)

AWS Acknowledgment

The Authors and Publisher gratefully acknowledge the support provided by the American Welding Society in the development and publication of this textbook series. "American Welding Society," the AWS logo and the SENSE logo are the trade and service marks of the American Welding Society and are used with permission.

For more information on the American Welding Society and the SENSE program, visit **http://www.aws.org/education/sense/** or contact AWS at (800) 443-9353 ext. 455 or by email: **education@aws.org**.

Acknowledgments

The authors and publisher would like to thank the following individuals for their contributions:

Garey Bish, *Gwinnett Technical College, Lawrenceville, GA*
Julius Blair, *Greenup County Area Technology Center, Greenup, KY*
Rick Brandon, *Pemiscot County Career & Technical Center, Hayti, MO*
Stephen Brandow, *University of Alaska Southeast, Ketchikan, Ketchikan, AK*
Francis X Brieden, *Career Technology Center of Lackawanna County, Scranton, PA*
John Cavenaugh, *Community College of Southern Nevada, Las Vegas, NV*
Clay Corey, *Washington-Saratoga BOCES, Fort Edward, NY*
Keith Cusey, *Institute for Construction Education, Decatur, IL*
Craig Donnell, *Whitmer Career Technology Center, Toledo, OH*
Steve Farnsworth, *Iowa Lakes Community College, Emmetsburg, IA*
Ed Harrell, *Traviss Career Center, Lakeland, FL*
Robert Hoting, *Northeast Iowa Community College , Sheldon, IA*
Steve Kistler, *Moberly Area Technical Center, Moberly, MO*
David Lynn, *Lebanon Technology & Career Center, Lebanon, MO*
Frank Miller, *Gadsden State Community College , Gadsden, AL*
Chris Overfelt, *Arnold R Burton Tech Center, Salem, VA*
Kenric Sorenson, *Western Technical College, LaCrosse, WI*
Pete Stracener, *South Plains College, Levelland, TX*
Bill Troutman, *Akron Public Schools, Akron, OH*
Norman Verbeck, *Columbia/Montour AVTS, Bloomsburg, PA*

About The Authors

Larry Jeffus is a dedicated teacher and author with over twenty years experience in the classroom and several Delmar Cengage Learning welding publications to his credit. He has been nominated by several colleges for the Innovator of the Year award for setting up nontraditional technical training programs. He was also selected as the Outstanding Post-Secondary Technical Educator in the State of Texas by the Texas Technical Society. Now retired from teaching, he remains very active in the welding community, especially in the field of education.

Lawrence Bower is a welding instructor at Blackhawk Technical College, an AWS SENSE School, in Janesville, Wisconsin. Mr. Bower is an AWS-certified Welding Inspector and Welding Educator. In helping to create *Welding: Skills, Processes and Practices for Entry-Level Welders,* he has brought to bear an excellent mix of training experience and manufacturing know-how from his work in industry, including fourteen years at United Airlines, and six years in the US Navy as an aerospace welder.

CHAPTER

1

Shielded Metal Arc Equipment, Setup, and Operation

OBJECTIVES

After completing this chapter, the student should be able to

- describe the process of shielded metal arc welding (SMAW)
- list and define the three units used to measure a welding current
- contrast how adding various chemicals to the coverings of the electrodes affects the welding arc
- contrast the three different types of current used for welding and their effects on welds
- contrast constant current (CC) and constant voltage (CV) welding power supplies and which type the shielded metal arc welding process requires
- define open circuit voltage and operating voltage
- identify arc blow and apply three different techniques to control it
- tell what the purpose of a welding transformer is and what kind of change occurs to the voltage and amperage with a step-down transformer
- tell the purpose of a rectifier
- read a welding machine duty cycle chart and explain its significance
- determine the proper welding cable size
- service and repair SMAW electrode holders
- identify three problems that can occur as a result of poor work lead clamping
- describe the factors that should be considered when placing an arc welding machine in a welding area

KEY TERMS

amperage
anode
cathode
duty cycle
electrons
inverter
magnetic flux lines
open circuit voltage
operating voltage
output
rectifier
step-down transformer
voltage
wattage
welding cables
welding leads

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter

Module 1: Occupational Orientation

Key Indicator 1: Prepares time or job cards, reports or records
Key Indicator 3: Follows verbal instructions to complete work assignments
Key Indicator 4: Follows written instructions to complete work assignments

Module 4: Shielded Metal Arc Welding (SMAW)

Key indicator 1: Performs safety inspections of SMAW equipment and accessories
Key Indicator 2: Makes minor external repairs to SMAW equipment and accessories
Key Indicator 3: Sets up SMAW operations on carbon steel

INTRODUCTION

Shielded metal arc welding (SMAW) is a welding process that uses a flux-covered metal electrode to carry an electrical current, **Figure 1.1.** The current forms an arc across the gap between the end of the electrode and the work. The electric arc creates sufficient heat to melt both the electrode and the work. Molten metal from the electrode travels across the arc to the molten pool on the base metal, where they mix together. The end of the electrode and molten pool of metal is surrounded, purified, and protected by a gaseous cloud and a covering of molten flux is produced as the flux coating of the electrode burns or vaporizes. As the arc moves away, the mixture of molten electrode and base metal solidifies and becomes one piece. At the same time, the molten flux solidifies forming a solid slag. Some electrode types produce heavier slag coverings than others.

SMAW is a widely used welding process because of its low cost, flexibility, portability, and versatility. The machine and the electrodes are low cost. The machine itself can be as simple as a 110-V step-down transformer. The electrodes are available from a large number of manufacturers in packages from 1 lb (0.5 kg) to 50 lb (22 kg).

The SMAW process is very flexible in terms of the metal thicknesses that can be welded and the variety of positions it can be used in. Metal as thin as 1/16 in. (2 mm) thick, or approximately 16 gauge, to several feet thick can be welded using the same machine with different settings and sizes of electrodes. The flexibility of the process also allows metal in this thickness range to be welded in any position.

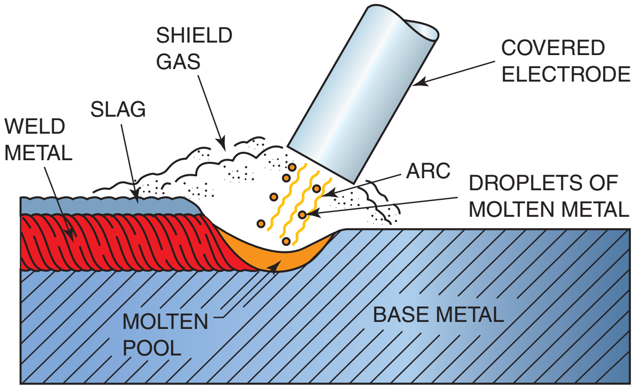
SHIELD GAS
COVERED ELECTRODE
SLAG
WELD METAL
ARC
DROPLETS OF MOLTEN METAL
MOLTEN POOL
BASE METAL

Figure 1.1
Shielded metal arc welding

SMAW is a very portable process because it is easy to move the equipment, and engine-driven generator-type welders are available. Also, the limited amount of equipment required for the process makes moving easy.

The process is versatile, and it is used to weld almost any metal or alloy, including cast iron, aluminum, stainless steel, and nickel.

WELDING CURRENT

The source of heat for arc welding is an electric current. An electric current is the flow of **electrons.** Electrons flow through a conductor from negative (−) to positive (+), **Figure 1.2.** Resistance to the flow of electrons (electricity) produces heat. The greater the resistance, the greater the heat. Air has a high resistance to current flow. As the electrons jump the air gap between the end of the electrode and the work, a great deal of heat is produced. Electrons flowing across an air gap produce an arc.

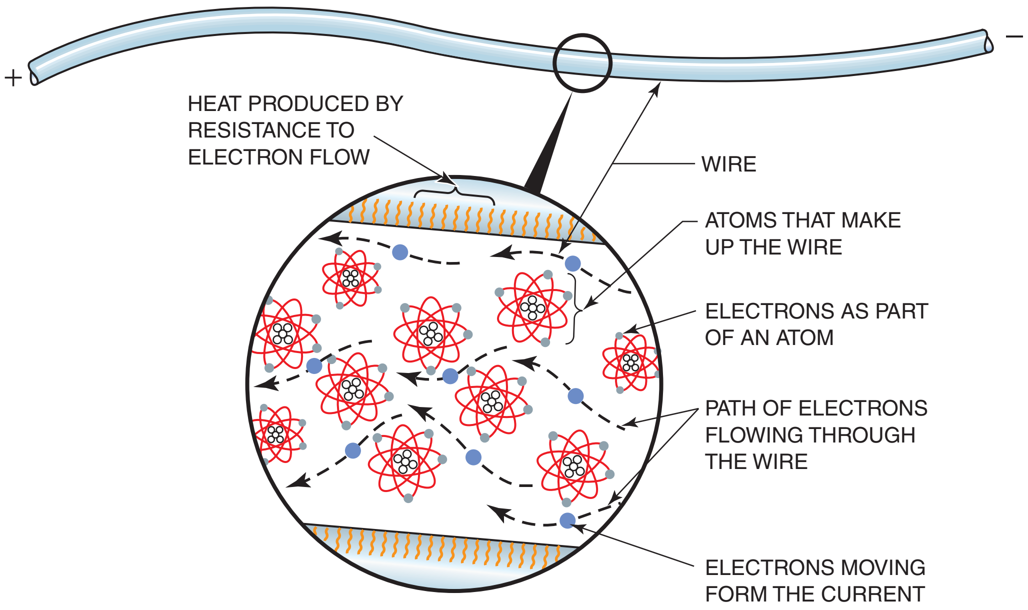
+
−
HEAT PRODUCED BY RESISTANCE TO ELECTRON FLOW
WIRE
ATOMS THAT MAKE UP THE WIRE
ELECTRONS AS PART OF AN ATOM
PATH OF ELECTRONS FLOWING THROUGH THE WIRE
ELECTRONS MOVING FORM THE CURRENT

Figure 1.2
Electrons traveling along a conductor

Measurement

Three units are used to measure a welding current. The three units are voltage (V), amperage (A), and wattage (W). **Voltage,** or volts (V), is the measurement of electrical pressure, in the same way that pounds per square inch is a measurement of water pressure. Voltage controls the maximum gap the electrons can jump to form the arc. A higher voltage can jump a larger gap. **Amperage,** or amperes (A), is the measurement of the total number of electrons flowing, in the same way that gallons is a measurement of the amount of water flowing. Amperage controls the size of the arc. **Wattage,** or watts (W), is calculated by multiplying volts (V) times amperes (A), **Figure 1.3.** Wattage is a measurement of the amount of electrical energy or power in the arc. The amount of watts being put into a weld per inch (cm) controls the width and depth of the weld bead, **Figure 1.4.**

Temperature

The temperature of a welding arc exceeds 11,000°F (6000°C). The exact temperature depends on the resistance to the current flow. The resistance is affected by the arc length and the chemical composition of the gases formed as the electrode covering burns and vaporizes. As the arc lengthens, the resistance increases, thus causing a rise in the arc voltage and temperature. The shorter the arc, the lower the arc temperature produced.

Most shielded metal arc welding electrodes have chemicals added to their coverings to stabilize the arc. These arc stabilizers reduce the arc resistance, making it easier to maintain an arc. By lowering the resistance, the arc stabilizers also lower the arc temperature. Other chemicals within the gaseous cloud around the arc may raise or lower the resistance.

The amount of heat produced is determined by the size of the electrode and the amperage setting. Not all of the heat produced by an arc reaches the weld. Some of the heat is radiated away in the form of light and heat waves, **Figure 1.5.** Additional heat is carried away with the hot gases formed by the electrode covering. Heat also is lost through conduction in the work. In total, about 50% of all heat produced by an arc is missing from the weld.

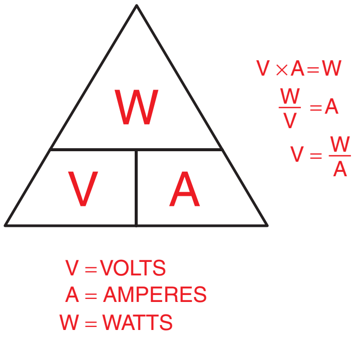

Figure 1.3
Ohm's law

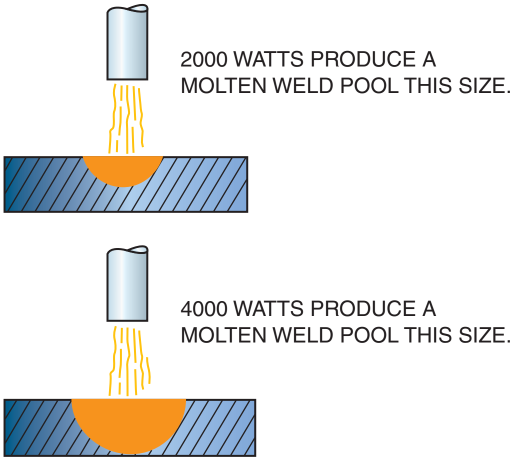
2000 WATTS PRODUCE A MOLTEN WELD POOL THIS SIZE.
4000 WATTS PRODUCE A MOLTEN WELD POOL THIS SIZE.

Figure 1.4
The molten weld pool size depends upon the energy (watts), the metal mass, and thermal conductivity

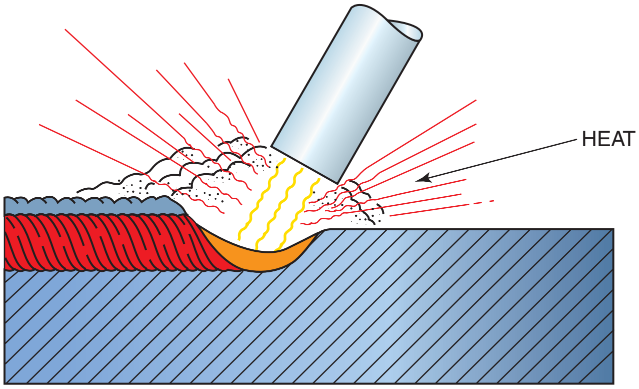

Figure 1.5
Energy is lost from the weld in the forms of radiation and convection

The 50% of the remaining heat the arc produced is not distributed evenly between both ends of the arc. This distribution depends on the composition of the electrode's coating, the type of welding current, and the polarity of the electrode's coating.

Currents

The three different types of current used for welding are alternating current (AC), direct current electrode negative (DCEN), and direct current electrode positive (DCEP). The terms DCEN and DCEP have replaced the former terms *direct current straight polarity (DCSP)* and *direct current reverse polarity (DCRP).* DCEN and DCSP are the same currents, and DCEP and DCRP are the same currents. Some electrodes can be used with only one type of current. Others can be used with two or more types of current. Each welding current has a different effect on the weld.

DCEN

In direct current electrode negative, the electrode is negative, and the work is positive, **Figure 1.6.** DCEN welding current produces a high electrode melting rate.

DCEP

In direct current electrode positive, the electrode is positive, and the work is negative, **Figure 1.7.** DCEP current produces the deepest penetrating welding arc characteristics.

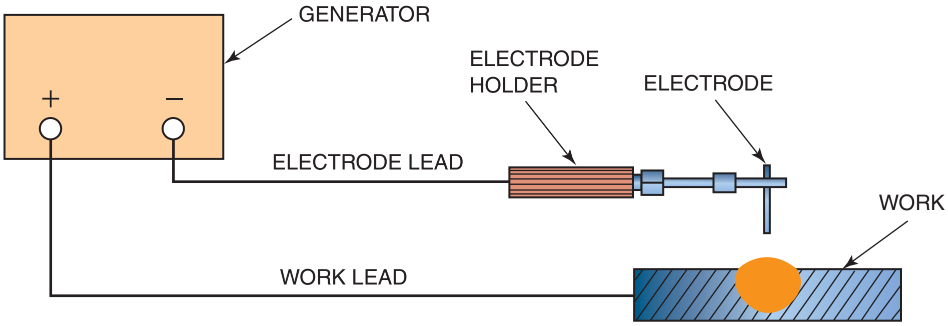

Figure 1.6
Electrode negative (DCEN), straight polarity (DCSP)

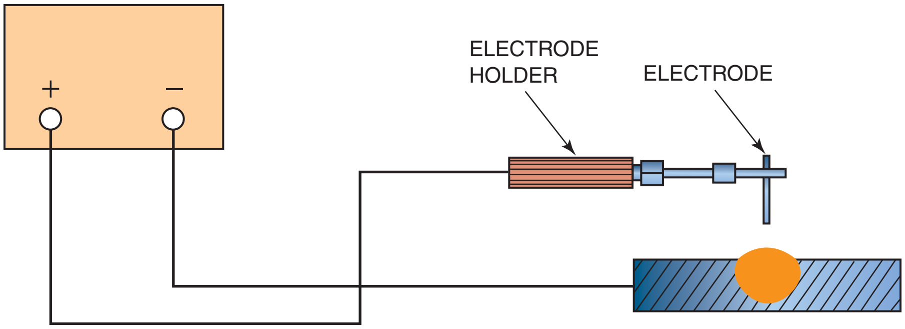

Figure 1.7
Electrode positive (DCEP), reverse polarity (DCRP)

AC

In alternating current, the electrons change direction every 1/120 of a second so that the electrode and work alternate from **anode** to **cathode, Figure 1.8.** The positive side of an electrode arc is called the anode, and the negative side is called the cathode. The rapid reversal of the current flow causes the welding heat to be evenly distributed on both the work and the electrode—that is, half on the work and half on the electrode. The even heating gives the weld bead a balance between penetration and buildup.

TYPES OF WELDING POWER

Welding power can be supplied as

- Constant voltage (CV)—The arc voltage remains constant at the selected setting even if the arc length and amperage increase or decrease.
- Rising-arc voltage (RAV)—The arc voltage increases as the amperage increases.
- Constant current (CC)—The total welding current (watts) remains the same. This type of power is also called drooping-arc voltage (DAV), because the arc voltage decreases as the amperage increases.

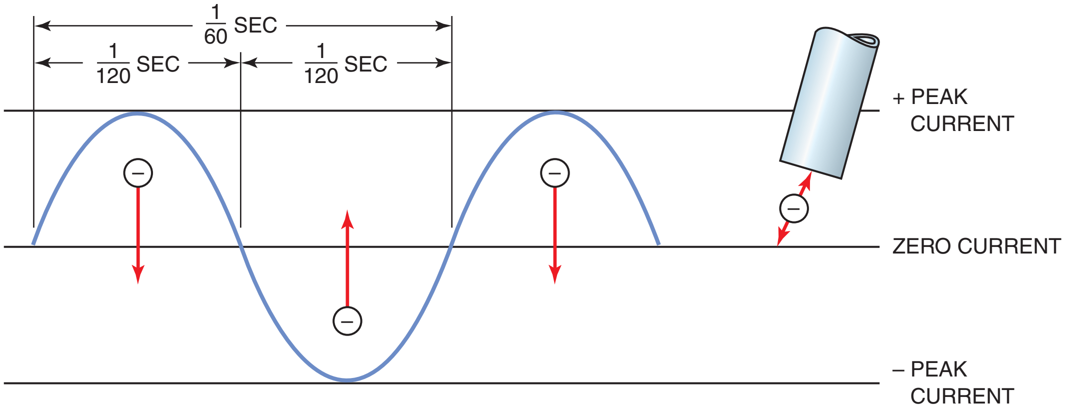

Figure 1.8
Alternating current sine wave (AC)

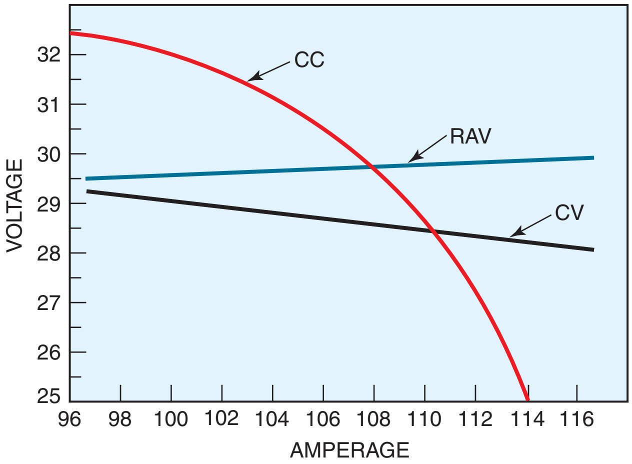

Figure 1.9
Constant voltage (CV), rising arc voltage (RAV), and constant current (CC)

The shielded metal arc welding (SMAW) process requires a constant current arc voltage characteristic, illustrated by the CC line in **Figure 1.9.** The shielded metal arc welding machine's voltage **output** decreases as current increases. This output power supply provides a reasonably high open circuit voltage before the arc is struck. The high open circuit voltage quickly stabilizes the arc. The arc voltage rapidly drops to the lower closed circuit level after the arc is struck. Following this short starting surge, the power (watts) remains almost constant despite the changes in arc length. With a constant voltage output, small changes in arc length would cause the power (watts) to make large swings. The welder would lose control of the weld.

OPEN CIRCUIT VOLTAGE

Open circuit voltage is the voltage at the electrode before striking an arc (with no current being drawn). This voltage is usually between 50 V and 80 V. The higher the open circuit voltage, the easier it is to strike an arc. The higher voltage also increases the chance of electrical shock.

CAUTION

The maximum safe open circuit voltage for welders is 80 V.

OPERATING VOLTAGE

Operating voltage, or closed circuit voltage, is the voltage at the arc during welding. This voltage will vary with arc length, type of electrode being used, type of current, and polarity. The operating voltage will be between 17 V and 40 V.

ARC BLOW

When electrons flow, they create lines of magnetic force that circle around the line of flow, **Figure 1.10.** Lines of magnetic force are referred to as **magnetic flux lines.** These lines space themselves evenly along a current-carrying wire. If the wire is bent, the flux lines on one side are

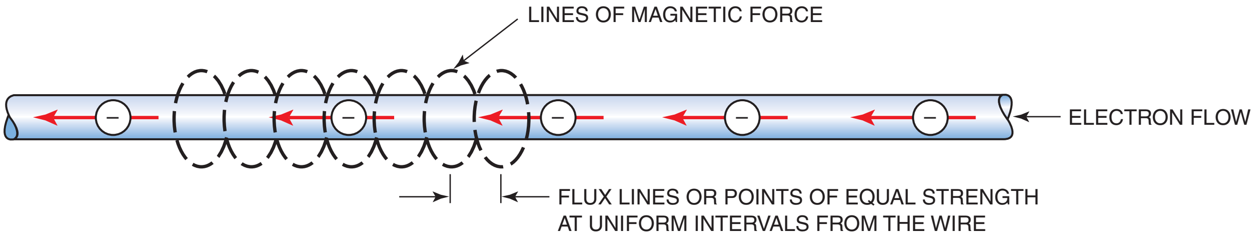
LINES OF MAGNETIC FORCE
ELECTRON FLOW
FLUX LINES OR POINTS OF EQUAL STRENGTH AT UNIFORM INTERVALS FROM THE WIRE

Figure 1.10
Magnetic force around a wire

compressed together, and those on the other side are stretched out, **Figure 1.11.** The unevenly spaced flux lines try to straighten the wire so that the lines can be evenly spaced once again. The force that they place on the wire is usually small. However, when welding with very high amperages, 600 A or more, the force may cause the wire to move.

The welding current flowing through a plate or any residual magnetic fields in the plate will result in uneven flux lines. These uneven flux lines can, in turn, cause an arc to move during a weld. This movement of the arc is called *arc blow.* Arc blow makes the arc drift as a string would drift in the wind. Arc blow is more noticeable in corners, at the ends of plates, and when the work lead is connected to only one side of a plate, **Figure 1.12.** If arc blow is a problem, it can be controlled by connecting the work lead to the end of the weld joint and making the weld in the direction toward the work lead, **Figure 1.13.** Another way of controlling arc blow is to use two work leads, one on each side of the weld. The best way to eliminate arc blow is to use alternating current. AC usually does not allow the flux lines to build long enough to bend the arc before the current changes direction. If it is impossible to move the work connection or to change to AC, a very short arc length can help control arc blow. A small tack weld or a change in the electrode angle can also help control arc blow.

Arc blow may not be a problem as you are learning to weld in the shop, because most welding tables are all steel. If you are using a pipe stand to hold your welding practice plates, arc blow can become a problem. In this case, try reclamping your practice plates or switch to alternating current.

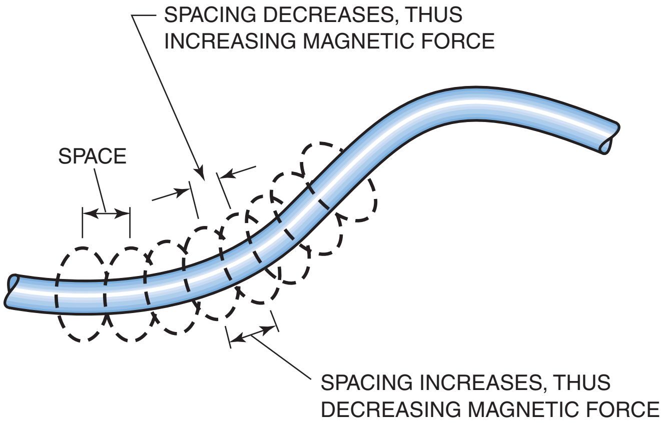
SPACING DECREASES, THUS INCREASING MAGNETIC FORCE
SPACE
SPACING INCREASES, THUS DECREASING MAGNETIC FORCE

Figure 1.11
Magnetic forces concentrate around bends in wires

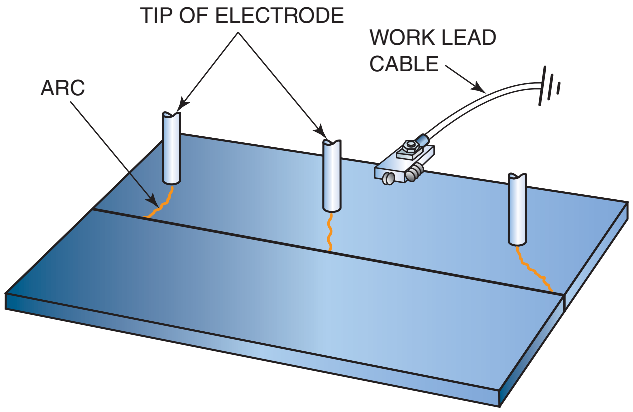
TIP OF ELECTRODE
WORK LEAD CABLE
ARC

Figure 1.12
Arc blow

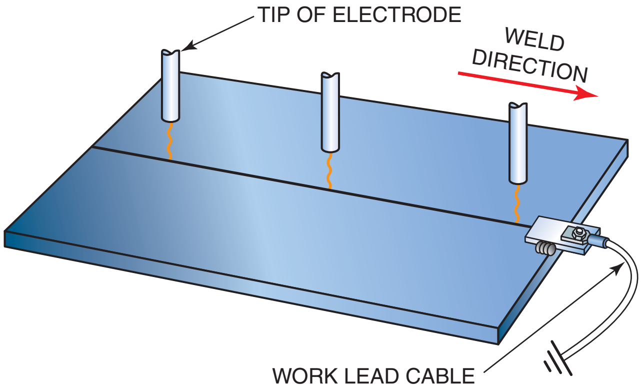

Figure 1.13
Correct current connections to control arc blow

TYPES OF POWER SOURCES

Two types of electrical devices can be used to produce the low-voltage, high-amperage current combination that arc welding requires. One type uses electric motors or internal combustion engines to drive alternators or generators. The other type uses **step-down transformers.** Because transformer-type welding machines are quieter, are more energy efficient, require less maintenance, and are less expensive, they are now the industry standard. However, engine-powered generators are still widely used for portable welding.

Transformers

A welding transformer uses the alternating current (AC) supplied to the welding shop at a high voltage to produce the low-voltage welding power. As electrons flow through a wire they produce a magnetic field around the wire. If the wire is wound into a coil the weak magnetic field of each wire is concentrated to produce a much stronger central magnetic force. Because the current being used is alternating or reversing each 1/120 of a second, the magnetic field is constantly being built and allowed to collapse. By placing a second, or secondary, winding of wire in the magnetic field produced by the first, or primary, winding, a current will be induced in the secondary winding. The placing of an iron core in the center of these coils will increase the concentration of the magnetic field, **Figure 1.14.**

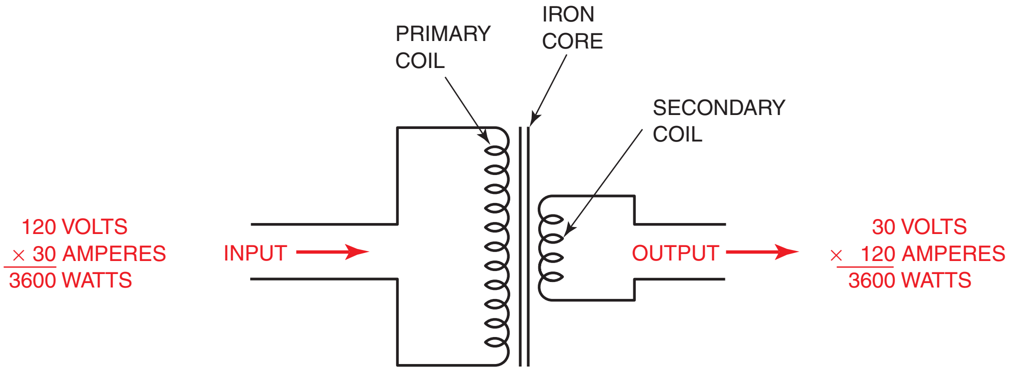

Figure 1.14
Diagram of a step-down transformer

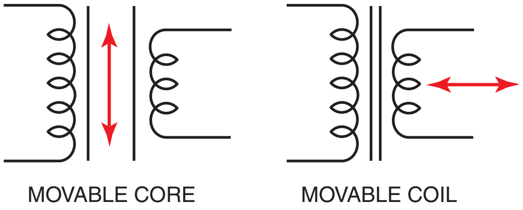

Figure 1.15
Major types of adjustable welding transformers

A transformer with more turns of wire in the primary winding than in the secondary winding is known as a step-down transformer. A step-down transformer takes a high-voltage, low-amperage current and changes it into a low-voltage, high-amperage current. Except for some power lost by heat within a transformer, the power (watts) into a transformer equals the power (watts) out because the volts and amperes are mutually increased and decreased.

A transformer welder is a step-down transformer. It takes the high line voltage (110 V, 220 V, 440 V, etc.) and low-amperage current (30 A, 50 A, 60 A, etc.) and changes it into 17 V to 45 V at 190 A to 590 A.

Welding machines can be classified by the method through which they control or adjust the welding current. The major classifications are multiple-coil, called tap-type, movable-coil or movable-core, **Figure 1.15,** and inverter-type machines.

Multiple-coil Machines

The multiple-coil machine, or tap-type machine, allows the selection of different current settings by tapping into the secondary coil at a different turn value. The greater the number of turns, the higher the amperage is

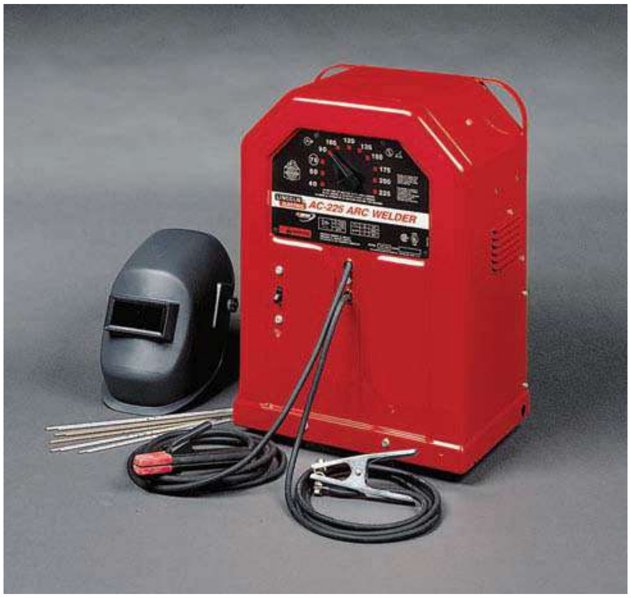

Figure 1.16
Tap-type transformer welding machine
Courtesy of Lincoln Electric Company

induced in the turns. These machines may have a large number of fixed amperes, **Figure 1.16,** or they may have two or more amperages that can be adjusted further with a fine adjusting knob. The fine adjusting knob may be marked in amperes, or it may be marked in tenths, hundredths, or in any other unit.

EXPERIMENT 1-1

Estimating Amperages

Module 1
Key Indicator 1, 3, 4

Module 4
Key Indicator 3

Using a pencil and paper, you will prepare a rough estimate of the amperage setting of a welding machine. **Figure 1.17** shows a welding machine with low, medium, and high tap amperage ranges. A fine adjusting knob is marked with ten equal divisions, and each division is again divided by ten smaller lines.

The machine is set on the medium range, 50 A to 250 A, and the fine adjusting knob is turned until it points to the line marked 5 (halfway between 0 and 10). This means that the amperage is halfway from 50 to 250, or 150 A. If the fine adjusting knob points between 2 and 3, the resulting amperage is one-quarter of the way from 50 to 250, or about 100 A. If the knob points between 7 and 8, the amperage is three-quarters of the way from 50 to 250, or about 200 A. If the knob points at 4, the amperage is more than 100 but a little less than 150, or about 130 A to 140 A. What is the amperage if the knob points at 6?

Since this is a method of estimating only, the amperage value obtained is close enough to allow an arc to be struck. The welder can then finish the fine adjusting to obtain a good weld.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

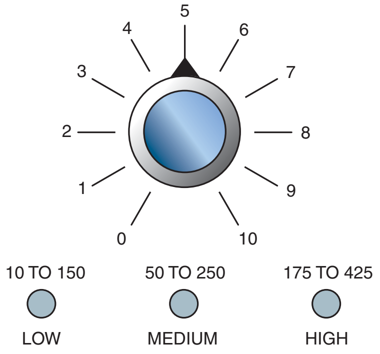

Figure 1.17
Tap-type welder knobs

EXPERIMENT 1-2

Calculating the Amperage Setting

Module 1
Key Indicator 1, 3, 4

Module 4
Key Indicator 3

Using a pencil and paper or calculator, you will calculate the exact value for each space on the fine adjusting knob of a welding machine.

With the machine set on the medium range, from 50 A to 250 A, first subtract the low amperage from the high amperage to get the amperage spread ($250 - 50 = 200$). Now divide the amperage spread by the number of units shown on the fine adjusting knob ($200 \div 10 = 20$). Each unit is equal to a 20-A increase, **Table 1.1.** When the knob points to 0, the amperage is 50; when the knob points to 1, the amperage is 70; and at 2, the amperage is 90, **Figure 1.18.** There are 100 small units on the fine adjusting knob. Dividing the amperage spread by the number of small units gives the amperage value for each unit ($200 \div 100 = 2$). Therefore, if the knob points to 6.1, the amperage is set at a value of $50 + 120 + 2 = 172$ A. This method provides a good starting place for the current setting, but if the welding is to be made in accordance with a welding procedure's specific amperage setting it will be necessary to use a calibrated meter to make the correct setting.

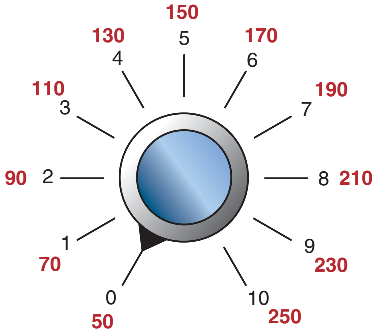

Figure 1.18
Fine adjusting knob

Table 1.1 Example of a Table Used to Calculate the Amperage Setting

Setting	Value in Amperes
0 = 50 + 0,	or 50 A
1 = 50 + 20,	or 70 A
2 = 50 + 40,	or 90 A
3 = 50 + 60,	or 110 A
4 = 50 + 80,	or 130 A
5 = 50 + 100,	or 150 A
6 = 50 + 120,	or 170 A
7 = 50 + 140,	or 190 A
8 = 50 + 160,	or 210 A
9 = 50 + 180,	or 230 A
10 = 50 + 200,	or 250 A

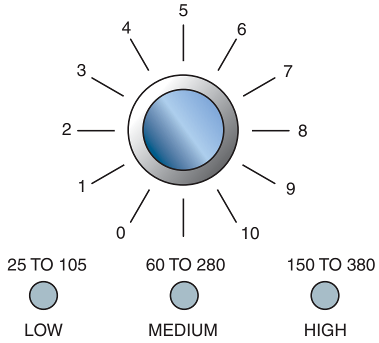

Figure 1.19 Practice 1.1

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 1-1

Estimating Amperages

Module 1
Key Indicator 1, 3, 4

Module 4
Key Indicator 3

Using a pencil and paper and the amperage ranges given in this practice (or from machines in the shop), you will estimate the amperage when the knob is at the 1/4, 1/2, and 3/4 settings, **Figure 1.19.**

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 1-2

Calculating Amperages

Module 1
Key Indicator 1, 3, 4

Module 4
Key Indicator 3

Using a pencil and paper or a calculator, and the amperage ranges given in this practice (or from machines in the shop), you will calculate the amperages for each of the following knob settings: 1, 4, 7, 9, 2.3, 5.7, and 8.5.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Movable-coil or Movable-core Machines

Movable-coil or movable-core machines are adjusted by turning a handwheel that moves the internal parts closer together or farther apart. The adjustment may also be made by moving a lever, **Figure 1.20.** These machines may have a high and low range, but they do not have a fine adjusting knob. The closer the primary and secondary coils are, the greater is the induced current; the greater the distance between the coils, the smaller the induced current, **Figure 1.21.** Moving the core in concentrates more of the magnetic force on the secondary coil, thus increasing the current. Moving the core out allows the field to disperse, and the current is reduced, **Figure 1.22.**

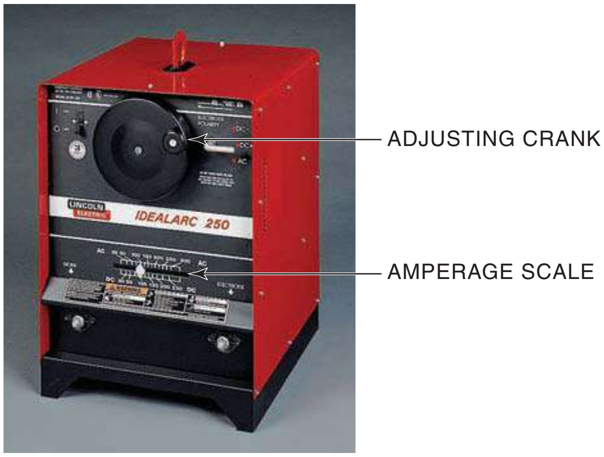

Figure 1.20
A movable core-type welding machine
Courtesy of Lincoln Electric Company

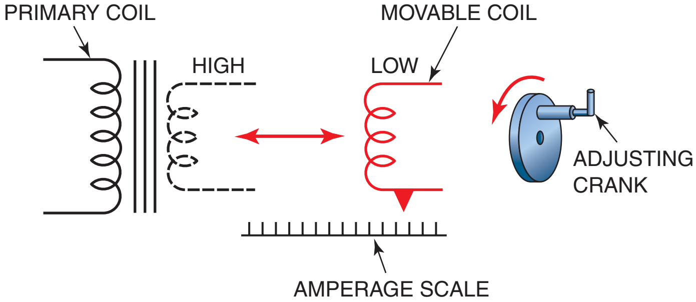

Figure 1.21
Movable coil

Inverter Machines

Inverter welding machines are much smaller than other types of machines of the same amperage range. This smaller size makes the welder much more portable and increases the energy efficiency. In a standard welding transformer, the iron core used to concentrate the magnetic field in the coils must be a specific size. The size of the iron core is determined by the length of time it takes for the magnetic field to build and collapse. By using solid-state electronic parts, the incoming power in an inverter welder is changed from 60 cycles a second to several thousand cycles a second. This higher frequency allows the use of a transformer that may be as light as 7 lb and still do the work of a standard transformer weighing 100 lb. Additional electronic parts remove the high frequency for the output welding power.

The use of electronics in the inverter-type welder allows it to produce any desired type of welding power. Before the invention of this machine, each type of welding required a separate machine. Now a single welding machine can produce the specific type of current needed for shielded metal arc welding, gas tungsten arc welding, gas metal arc welding, and plasma arc cutting. Because the machine can be light enough to be carried closer to work, shorter welding cables can be used. The welder does not have to walk as far to adjust the machine. Welding machine power wire is cheaper than welding cables. Some manufacturers produce machines that can be stacked so that when you need a larger machine all you have to do is add another unit to your existing welder, **Figure 1.23.**

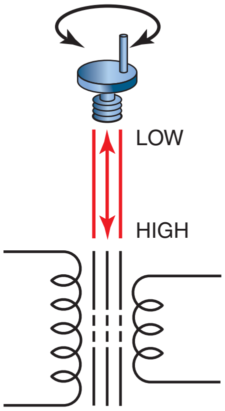

Figure 1.22
Movable core

GENERATORS AND ALTERNATORS

Generators and alternators both produce welding electricity from a mechanical power source. Both devices have an armature that rotates and a stator that is stationary. As a wire moves through a magnetic force field, electrons in the wire are made to move, producing electricity.

STICKWELD 300S
ARCON WELDING, LLC

Figure 1.23
Typical 300-A inverter-type power supply weighing only 70 lb
Courtesy of Arcon Welding, L.L.C.

In an alternator, magnetic lines of force rotate inside a coil of wire, **Figure 1.24.** An alternator can produce AC only. In a generator, a coil of wire rotates inside a magnetic field. A generator produces DC. It is possible for alternators to use diodes to change the AC to DC for welding. In generators, the welding current is produced on the armature and is picked up with brushes, **Figure 1.25.** In alternators, the welding current is produced on the stator, and only the small current for the electromagnetic force field goes across the brushes. Therefore, the brushes in an alter-

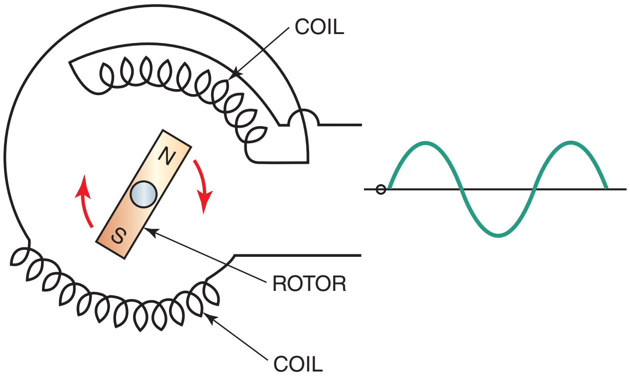
COIL
N
S
ROTOR
COIL

Figure 1.24
Schematic diagram of an alternator

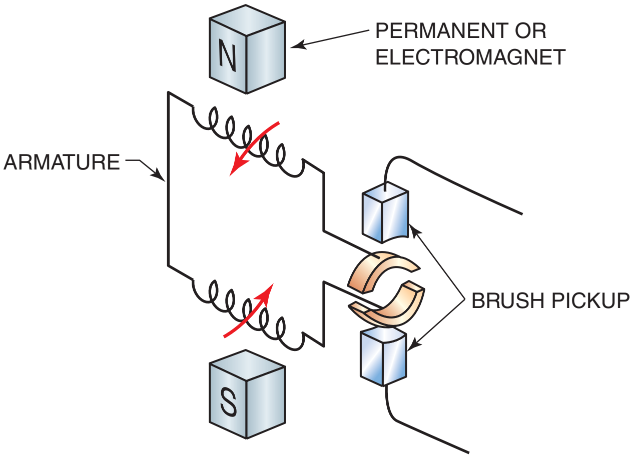
N
PERMANENT OR
ELECTROMAGNET
ARMATURE
BRUSH PICKUP
S

Figure 1.25
Diagram of a generator

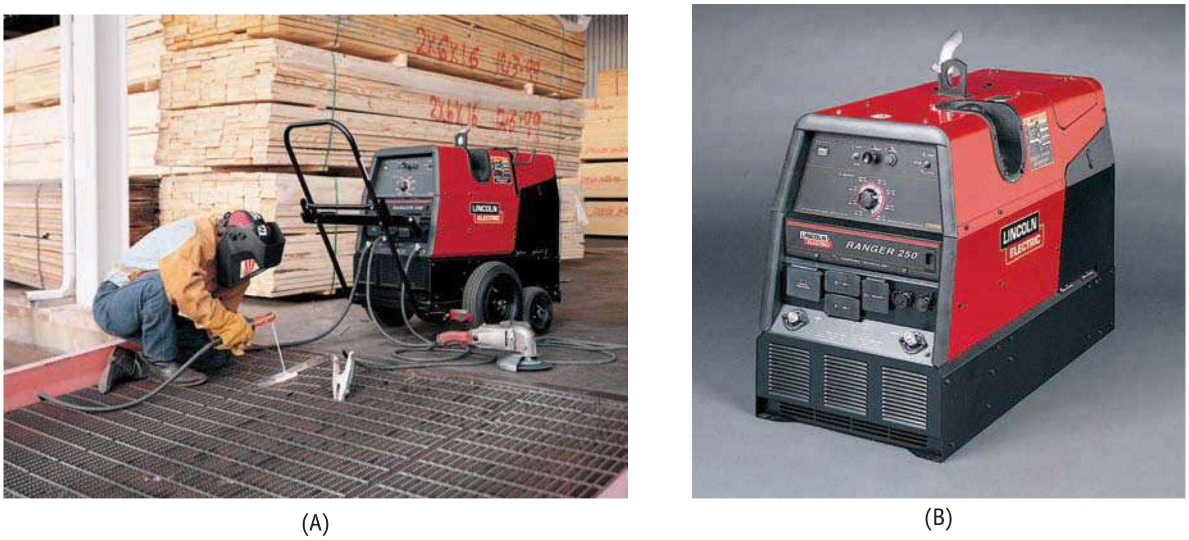
(A) (B)

Figure 1.26
Portable engine generator welder
Courtesy of Lincoln Electric Company

nator are smaller and last longer. Alternators can be smaller in size and lighter in weight than generators and still produce the same amount of power.

Engine-driven generators and alternators may run at the welding speed all the time, or they may have an option that reduces their speed to an idle when welding stops. This option saves fuel and reduces wear on the welding machine. To strike an arc when using this type of welder, stick the electrode to the work for a second. When you hear the welding machine (welder) pick up speed, remove the electrode from the work and strike an arc. In general, the voltage and amperage are too low to start a weld, so shorting the electrode to the work should not cause the electrode to stick. A timer can be set to control the length of time that the welder maintains speed after the arc is broken. The time should be set long enough to change electrodes without losing speed.

Portable welders often have 110-V or 220-V plug outlets, which can be used to run grinders, drills, lights, and other equipment. The power provided may be AC or DC. If DC is provided, only equipment with brush-type motors or tungsten light bulbs can be used. If the plug is not specifically labeled 110 V AC, check the owner's manual before using it for such devices as radios or other electronic equipment. A typical portable welder is shown in **Figure 1.26.**

It is recommended that a routine maintenance schedule for portable welders be set up and followed. By checking the oil, coolant, battery, filters, fuel, and other parts, the life of the equipment can be extended. A checklist can be posted on the welder, **Table 1.2.**

RECTIFIERS

Alternating-welding current can be converted to direct current by using a series of rectifiers. A **rectifier** allows current to flow in one direction only, **Figure 1.27.**

Table 1.2 Portable Welder Checklist. The owner's manual should be checked for any additional items that might need attention.

Check Each Day before Starting
Oil level
Water level
Fuel level
Check Each Monday
Battery level
Cables
Fuel line filter
Check at Beginning of Month
Air filter
Belts and hoses
Change oil and filter
Check Each Fall
Antifreeze
Test battery
Pack wheel bearings
Change gas filter

If one rectifier is added, the welding power appears as shown in **Figure 1.28.** It would be difficult to weld with pulsating power such as this. A series of rectifiers, known as a bridge rectifier, can modify the alternating current so that it appears as shown in **Figure 1.29.**

Rectifiers become hot as they change AC to DC. They must be attached to a heat sink and cooled by having air blown over them. The heat produced by a rectifier reduces the power efficiency of the welding machine. **Figure 1.30** shows the amperage dial of a typical machine. Notice that at the same dial settings for AC and DC, the DC is at a lower amperage. The difference in amperage (power) is due to heat lost in the rectifiers. The loss in power makes operation with AC more efficient and less expensive compared to DC.

A DC adapter for small AC machines is available from manufacturers. For some types of welding, AC does not work properly.

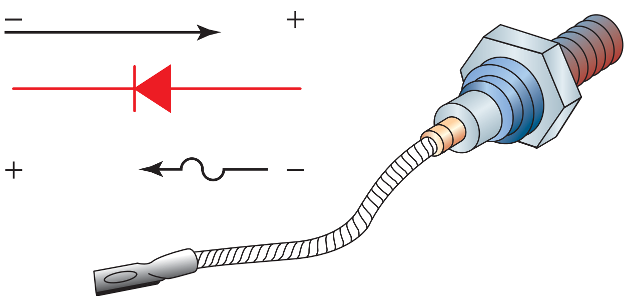

Figure 1.27
Rectifier

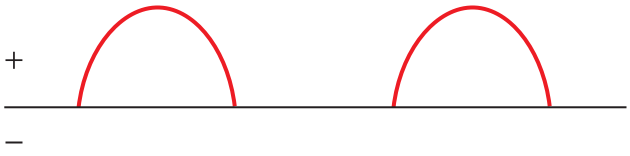

Figure 1.28
One rectifier in a welding power supply results in pulsating power

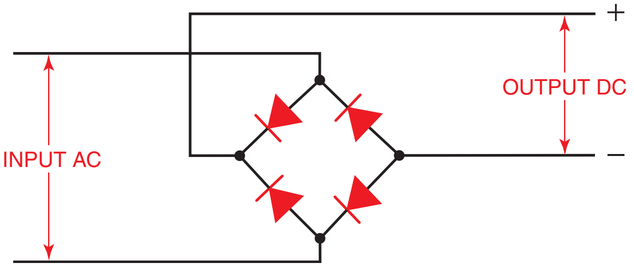

Figure 1.29
Bridge rectifier

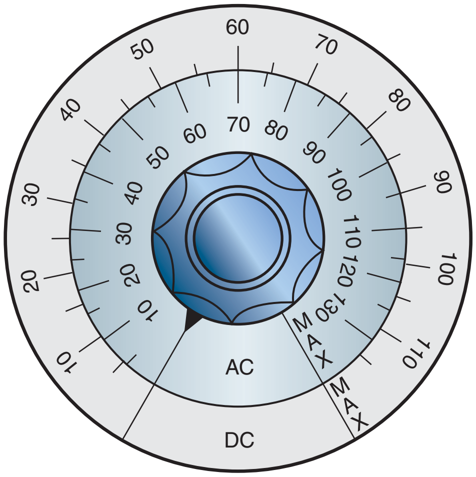

Figure 1.30
Typical dial on an AC-DC transformer rectifier welder

DUTY CYCLE

Welding machines produce internal heat at the same time that they produce the welding current. Except for automatic welding machines, welders are rarely used every minute for long periods of time. The welder must take time to change electrodes, change positions, or change parts. Shielded metal arc welding never continues for long periods of time.

The **duty cycle** is the percentage of time a welding machine can be used continuously. A 60% duty cycle means that out of any 10 minutes, the machine can be used for a total of six minutes at the maximum rated current. When providing power at this level, it must be cooled off for four minutes out of every 10 minutes. The duty cycle increases as the amperage is lowered and decreases for higher amperages, **Figure 1.31.** Most welding machines weld at a 60% rate or less. Therefore, most manufacturers list

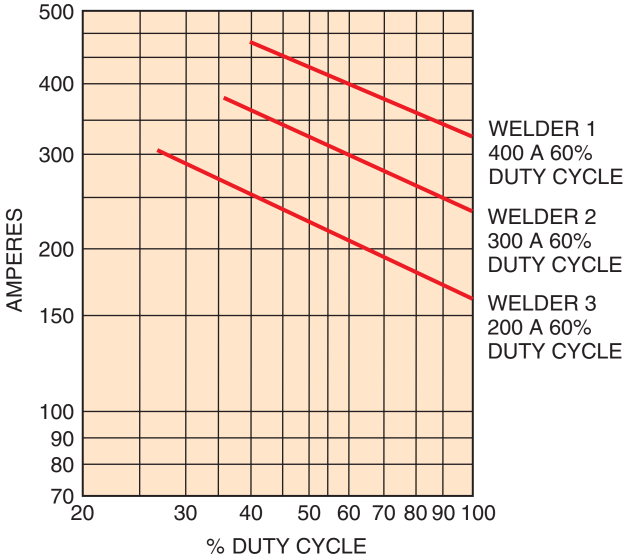

Figure 1.31
Duty cycle of a typical shielded metal arc welding machine

the amperage rating for a 60% duty cycle on the nameplate that is attached to the machine. Other duty cycles are given on a graph in the owner's manual.

The manufacturing cost of power supplies increases in proportion to their rated output and duty cycle. To reduce their price, it is necessary to reduce either their rating or their duty cycle. For this reason, some home-hobby welding machines may have duty cycles as low as 20% even at a low welding setting of 90 A to 100 A. The duty cycle on these machines should never be exceeded because a buildup of the internal temperature can cause the transformer insulation to break down, damaging the power supply.

PRACTICE 1-3

Reading Duty Cycle Chart

Module 1
Key Indicator 1, 3, 4

Module 4
Key Indicator 3

Using a pencil and paper and the duty cycle chart in Figure 1.31 (or one from machines in the shop), you will determine the following:

Welder 1: maximum welding amperage percent duty cycle at maximum amperage
Welder 2: maximum welding amperage percent duty cycle at maximum amperage
Welder 3: maximum welding amperage percent duty cycle at maximum amperage
Welder 1: maximum welding amperage at 100% duty cycle
Welder 2: maximum welding amperage at 100% duty cycle
Welder 3: maximum welding amperage at 100% duty cycle

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

WELDING CABLES AND LEADS

The terms **welding cables** and **welding leads** mean the same thing. Cables to be used for welding must be flexible, well insulated, and the correct size for the job. Most welding cables are made from stranded copper wire. Some manufacturers sell a newer type of cable made from aluminum wires. The aluminum wires are lighter and less expensive than copper. Because aluminum as a conductor is not as good as copper for a given wire size, the aluminum wire should be one size larger than would be required for copper.

The insulation on welding cables is exposed to hot sparks, flames, grease, oils, sharp edges, impact, and other types of wear. To withstand such wear, only specially manufactured insulation should be used for welding cable. Several new types of insulation are available that give longer service against these adverse conditions.

As electricity flows through a cable, the resistance to the flow causes the cable to heat up and increase the voltage drop. To minimize the loss of power and prevent overheating, the electrode cable and work cable must be the correct size. **Table 1.3** lists the minimum size cable that is required for each amperage and length. Large welding lead sizes make electrode

Table 1.3 Copper and Aluminum Welding Lead Sizes

	Amperes		Copper Welding Lead Sizes								
	ft	m	100	150	200	250	300	350	400	450	500
Length of Cable	50	15	2	2	2	2	1	1/0	1/0	2/0	2/0
	75	23	2	2	1	1/0	2/0	2/0	3/0	3/0	4/0
	100	30	2	1	1/0	2/0	3/0	4/0	4/0		
	125	38	2	1/0	2/0	3/0	4/0				
	150	46	1	2/0	3/0	4/0					
	175	53	1/0	3/0	4/0						
	200	61	1/0	3/0	4/0						
	250	76	2/0	4/0							
	300	91	3/0								
	350	107	3/0								
	400	122	4/0								

	Amperes		Aluminum Welding Lead Sizes								
	ft	m	100	150	200	250	300	350	400	450	500
Length of Cable	50	15	2	2	1/0	2/0	2/0	3/0	4/0		
	75	23	2	1/0	2/0	3/0	4/0				
	100	30	1/0	2/0	4/0						
	125	38	2/0	3/0							
	150	46	2/0	3/0							
	175	53	3/0								
	200	61	4/0								
	225	69	4/0								

manipulation difficult. Smaller cable can be spliced to the electrode end of a large cable to make it more flexible. This whip-end cable must not be more than 10 ft (3 m) long.

PRACTICE 1-4

Determining Welding Lead Sizes

Using a pencil and paper and Table 1.3, Copper and Aluminum Welding Lead Sizes, you will determine the following:

1. the minimum copper welding lead size for a 200-A welder with 100-ft (30-m) leads
2. the minimum copper welding lead size for a 125-A welder with 225-ft (69-m) leads
3. the maximum length aluminum welding lead that can carry 300 A

Splices and end lugs are available from suppliers. Be sure that a good electrical connection is made whenever splices or lugs are used. A poor electrical connection will result in heat buildup, voltage drop, and poor service from the cable. Splices and end lugs must be well insulated against possible electrical shorting, **Figure 1.32.**

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

CAUTION

A splice in a cable should not be within 10 ft (3 m) of the electrode because of the possibility of electrical shock.

Module 1
Key Indicator 1, 3, 4

Module 4
Key Indicator 3

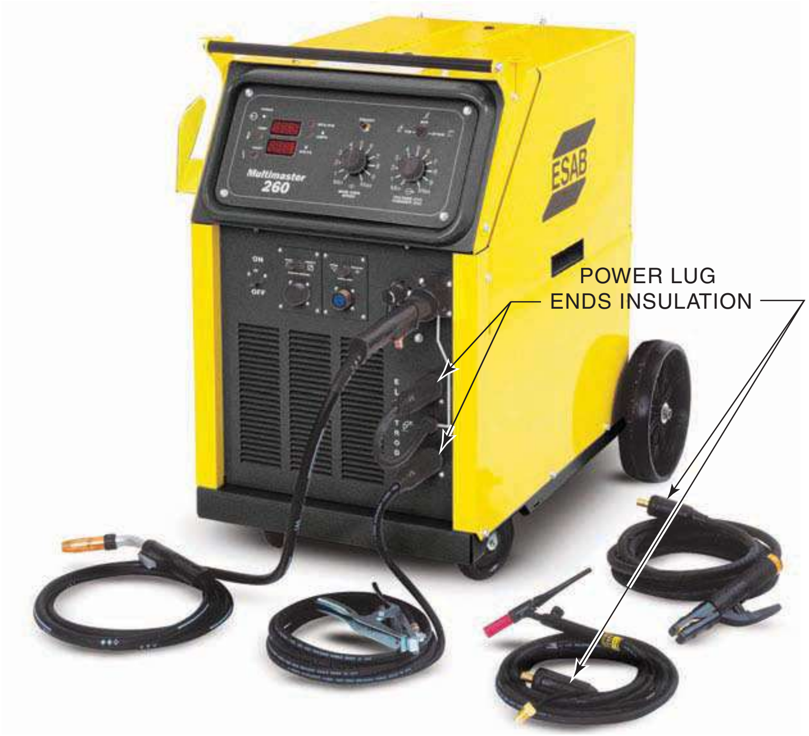

Figure 1.32
Power lug protection is provided by insulators
Courtesy of ESAB Welding & Cutting Products

CAUTION

Never dip a hot electrode holder in water to cool it off. The problem causing the holder to overheat should be repaired.

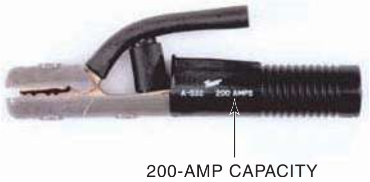

Figure 1.33
The amperage capacity of an electrode holder is often marked on its side
Courtesy of Tweco®, a Thermadyne® Company

Module 1
Key Indicator 1, 3, 4

Module 4
Key Indicator 1, 2

ELECTRODE HOLDERS

The electrode holder should be of the proper amperage rating and in good repair for safe welding. Electrode holders are designed to be used at their maximum amperage rating or less. Higher amperage values will cause the holder to overheat and burn up. If the holder is too large for the amperage range being used, manipulation is hard, and operator fatigue increases. Make sure that the correct amperage holder is chosen, **Figure 1.33.**

A properly sized electrode holder can overheat if the jaws are dirty or loose, or if the cable is loose. If the holder heats up, welding power is being lost. In addition, a hot electrode holder is uncomfortable to work with.

Replacement springs, jaws, insulators, handles, screws, and other parts are available to keep the holder in good working order, **Figure 1.34.** To prevent excessive damage to the holder, welding electrodes should not be burned too short. A 2-in. (51-mm) electrode stub is short enough to minimize electrode waste and save the holder.

PRACTICE 1-5

Repairing Electrode Holders

Using the manufacturer's instructions for your type of electrode holder, required hand tools, and replacement parts, you will do the following:

1. Remove the electrode holder from the welding cable.
2. Remove the jaw insulating covers.
3. Replace the jaw insulating covers.

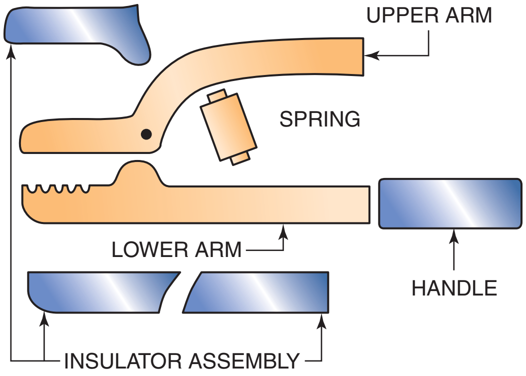

Figure 1.34
Replaceable parts of an electrode holder

4. Reconnect the electrode holder to the welding cable.
5. Turn on the welding power or reconnect the welding cable to the welder.
6. Make a weld to ensure that the repair was made correctly.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

CAUTION

Before starting any work, make sure that the power to the welder is off and locked off or the welding lead has been removed from the machine.

WORK CLAMPS

The work clamp must be the correct size for the current being used, and it must clamp tightly to the material. Heat can build up in the work clamp, reducing welding efficiency, just as was previously described for the electrode holder. Power losses in the work clamp are often overlooked. The clamp should be carefully touched occasionally to find out if it is getting hot.

In addition to power losses due to poor work lead clamping, a loose clamp may cause arcing that can damage a part. Improper work clamp placement may also contribute to arc blow. To avoid arc blow, keep the work clamp situated as close to the weld as practical, and weld away from it. If the part is to be moved during welding, a swivel-type work clamp may be needed, **Figure 1.35.** It may be necessary to weld a tab to thick parts so that the work lead can be clamped to the tab, **Figure 1.36.**

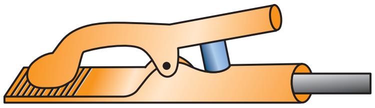

Figure 1.35
A work clamp may be attached to the workpiece

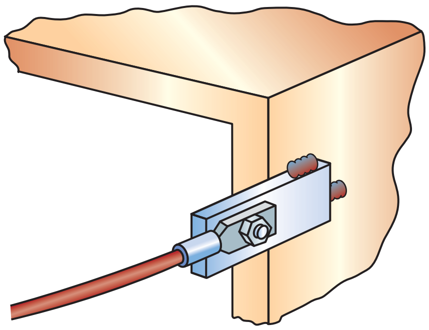

Figure 1.36
Tack welded ground to part

SETUP

Module 1
Key Indicator 1

Module 4
Key Indicator 3

Arc welding machines should be located near the welding site, but far enough away so that they are not covered with spark showers. The machines may be stacked to save space, but there must be enough room between the machines to ensure the air can circulate so as to keep the machines from overheating. The air that is circulated through the machine should be as free as possible of dust, oil, and metal filings. Even in a good location, the power should be turned off periodically and the machine blown out with compressed air, **Figure 1.37.**

The welding machine should be located away from cleaning tanks and any other sources of corrosive fumes that could be blown through it. Water leaks must be fixed and puddles cleaned up before a machine is used.

Power to the machine must be fused, and a power shut-off switch provided. The switch must be located so that it can be reached in an emergency without touching either the machine or the welding station. The machine case or frame must be grounded.

The welding cables should be sufficiently long to reach the work station but not so long that they must always be coiled. Cables should not be placed on the floor in aisles or walkways. If cables must cross a walkway, the cable must be installed overhead, or it must be protected by a ramp, **Figure 1.38.** The welding machine and its main power switch should be off while a person is installing or working on the cables.

The work station must be free of combustible materials. Screens or curtains should be provided to protect other workers from the arc light.

The welding cable should never be wrapped around arms, shoulders, waist, or any other part of the body. If the cable was caught by any moving equipment, such as a forklift, crane, or dolly, a welder could be pulled off balance or more seriously injured. If it is necessary to hold the weight off the cable so that the welding can more easily be done, a free hand can be used. The cable should be held so that if it is pulled it can be easily released.

Check the surroundings before starting to weld. If heavy materials are being moved in the area around you, there should be a safety watch. A safety watch can warn a person of danger while that person is welding.

CAUTION

The cable should never be tied to scaffolding or ladders. If the cable is caught by moving equipment, the scaffolding or ladder may be upset, causing serious personal injury.

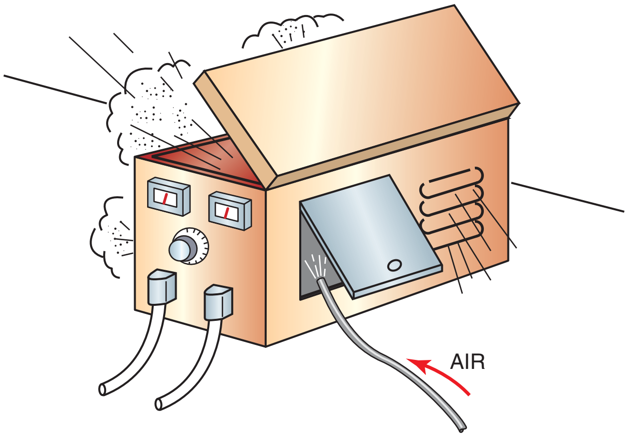

Figure 1.37
Slag, chips from grinding, and dust must be blown out occasionally so that they will not start a fire or cause a short-out or other types of machine failure

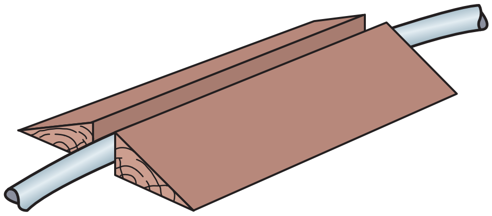

Figure 1.38
To prevent people from tripping, when cables must be placed in walkways, lay two blocks of wood beside the cables

SUMMARY

Understanding the scientific theory of electricity and magnetism will aid you in understanding how the welding currents are produced and their reactions to changes in their physical surroundings. Understanding electromagnetic phenomena will aid you in controlling arc blow. Failure to control arc blow can result in weld failures. In addition, understanding electricity will help you interpret information given on manufacturers' tables, charts, and equipment specifications.

Before starting any new job or welding operation, be sure to check with the equipment manufacturer's safety guidelines for proper operation and maintenance. Follow all recommended guidelines.

Keeping your work area clean and orderly will help prevent accidents.

REVIEW

1. Describe the welding current.
2. What produces the heat during a shielded metal arc weld?
3. Voltage can be described as ______.
4. Amperage can be described as ______.
5. Wattage can be described as ______.
6. What determines the exact temperature of the shielded metal welding arc?
7. Does all of the heat produced by an SMA weld stay in the weld? Why or why not?
8. What do the following abbreviations mean: AC, DCEN, DCEP, DCSP, and DCRP?
9. Sketch a welding machine, an electrode lead, an electrode holder, an electrode, a work lead, and work connected for DCEN welding.
10. Sketch a welding machine, an electrode lead, an electrode holder, an electrode, a work lead, and work connected for DCEP welding.
11. Why is SMA welding current referred to as *constant current?*
12. What is the higher voltage at the electrode before the arc is struck called? What is its advantage to welding?
13. Referring to the graph in Figure 1.9, what would the voltage be for the CC power supply at 110 A? What would the watts be?
14. How does arc blow affect welding?
15. How can arc blow be controlled?
16. How does a welding transformer work?
17. What are *taps* on a welding transformer?
18. What would the approximate amperage setting be if a welder were set to the high range (150 A to 350 A) and the fine adjustment knob were pointing at 5 on a 10-point scale?
19. What are the advantages of the inverter-type welding power supply?
20. What is the difference between the welding current produced by alternators and the one produced by generators?
21. What are the advantages of alternators over generators?
22. What must be checked before using the 110-V power plug on a portable welder?
23. What is meant by a *welder's duty cycle?*

24. Why must a welding machine's duty cycle never be exceeded?
25. Why is copper better than aluminum for welding cables?
26. A splice in a welding cable should never be any closer than __________ to the electrode holder.
27. Why must the electrode holder be correctly sized?
28. What can cause a properly sized electrode holder to overheat?
29. What problem can occur if welding machines are stacked or placed too closely together?
30. Why must welding cables never be tied to scaffolding or ladders?

CHAPTER 2

Shielded Metal Arc Welding of Plate

OBJECTIVES

After completing this chapter, the student should be able to

- demonstrate safe SMAW work practices
- strike an arc at a specific point
- contrast the problems that can result if the welding current is set too low or too high
- select the correct diameter of welding electrode for a weld
- compare a weld bead's shape for width, reinforcement, and appearance with differing heat inputs
- define arc length, and describe the effects of using too short or too long an arc length
- compare a leading electrode angle to a trailing electrode angle
- list three characteristics of the weld bead that can be controlled by the movement or weaving of the welding electrode
- demonstrate ten weave patterns for weld beads
- match various SMAW electrodes to the four filler metal groups
- define stringer beads and tell how they are used
- make a vertical up stringer bead and a horizontal stringer bead
- make a welded square butt joint in the flat, vertical up, and the horizontal positions
- on an edge joint, make a flat weld, a vertical down weld, a vertical up weld, a horizontal weld, and an overhead weld
- on an outside corner joint, make a flat weld, a vertical down weld, a vertical up weld, a horizontal weld, and an overhead weld
- make fillet welds in lap joints in all positions
- make fillet welds in tee joints in all positions
- make groove welds in plate in all positions without backing

KEY TERMS

amperage range
arc length
cellulose-based fluxes
electrode angle
lap joint
mineral-based fluxes
rutile-based fluxes
square butt joint
stringer bead
tee joint
weave pattern

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter

Module 1: Occupational Orientation

Key Indicator 1: Prepares time or job cards, reports or records
Key Indicator 2: Performs housekeeping duties
Key Indicator 3: Follows verbal instructions to complete work assignments
Key Indicator 4: Follows written instructions to complete work assignments

Module 2: Safety and Health of Welders

Key indicator 1: Demonstrates proper use and inspection of personal protective equipment (PPE)
Key Indicator 2: Demonstrates safe operation practices in the work area
Key Indicator 3: Demonstrates proper use and inspection of ventilation equipment
Key Indicator 4: Demonstrates proper hot zone operation
Key Indicator 5: Demonstrates proper work actions for work in confined spaces
Key Indicator 6: Understands proper use of precautionary labeling and MSDS information

Module 4: Shielded Metal Arc Welding (SMAW)

Key Indicator 1: Performs safety inspections of SMAW equipment and accessories
Key Indicator 3: Sets up SMAW operations on carbon steel
Key Indicator 4: Operates SMAW equipment on carbon steel
Key Indicator 5: Makes SMAW fillet welds, in all positions, on carbon steel
Key Indicator 6: Makes SMAW groove welds, in all positions, on carbon steel

Module 9: Welding Inspection and Testing Principles

Key Indicator 2: Examines tacks, root passes, intermediate layers, and completed welds

INTRODUCTION

Shielded metal arc welding (SMAW), or stick welding, is a common method used to join plate. This method provides a high temperature and concentration of heat, which allows a small molten weld pool to be built up quickly. The addition of filler metal from the electrode adds reinforcement and increases the strength of the weld. SMAW can be performed on almost any type of metal 1/8 in. (3 mm) thick or thicker. A minimum amount of equipment is required, and it can be portable.

High-quality welds can be consistently produced on almost any type of metal and in any position. The quality of the welds produced depends largely upon the skill of the welder. Developing the necessary skill level requires practice. However, practicing the welds repeatedly without changing techniques will not aid in developing the required skills. Each time a weld is completed it should be evaluated, and then a change should be made in the technique to improve the next weld.

PRACTICE 2-1

Shielded Metal Arc Welding Safety

Module 1
Key Indicator 1, 2, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 5, 6

Module 4
Key Indicator 1

Module 9
Key Indicator 2

Using a welding workstation, welding machine, welding electrodes, welding helmet, eye and ear protection, welding gloves, proper work clothing, and any special protective clothing that may be required, demonstrate, to your instructor and other students, the safe way to prepare yourself and the welding workstation for welding. Include in your demonstration appropriate references to burn protection, eye and ear protection, material safety data sheets, ventilation, electrical safety, general work clothing, special protective clothing, and area cleanup. More information on welding safety can be found in Chapter 2 in *Welding Skills, Processes and Practices for Entry-Level Welders: Book One* and ANSI Z49.1, *Safety in Welding, Cutting, and Allied Processes.*

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

EXPERIMENT 2-1

Striking the Arc

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 4

Module 9
Key Indicator 2

Using a properly set up and adjusted arc welding machine, the proper safety protection as demonstrated in Practice 2-1, E6011 or E6103 welding electrodes having a 1/8-in. (3-mm) core wire diameter measured at the bare end of the electrode, and one piece of mild steel plate, 1/4 in. (6 mm) thick, you will practice striking an arc, **Figure 2.1.**

With the electrode held over the plate, lower your helmet. Scratch the electrode across the plate (like striking a large match), **Figure 2.2.** As the arc is established, slightly raise the electrode to the desired arc length. Hold the arc in one place until the molten weld pool builds to the desired size. Slowly lower the electrode as it burns off and move it forward to start the bead.

If the electrode sticks to the plate, quickly squeeze the electrode holder lever to release the electrode. Break the electrode free by bending it back and forth a few times. Do not touch the electrode without gloves because it will still be hot. If the flux breaks away very far from the end of the electrode, throw out the electrode because restarting the arc will be very difficult, **Figure 2.3.**

Once you are able to easily strike an arc and make a weld, try to strike the arc where it will be re-melted by the weld you are making. Arc strikes on the metal's surface that are not covered up by the weld are considered to be weld defects by most codes.

Break the arc by rapidly raising the electrode after completing a 1-in. (25-mm) weld bead. Restart the arc as you did before, and make another short weld. Repeat this process until you can easily start the arc each time.

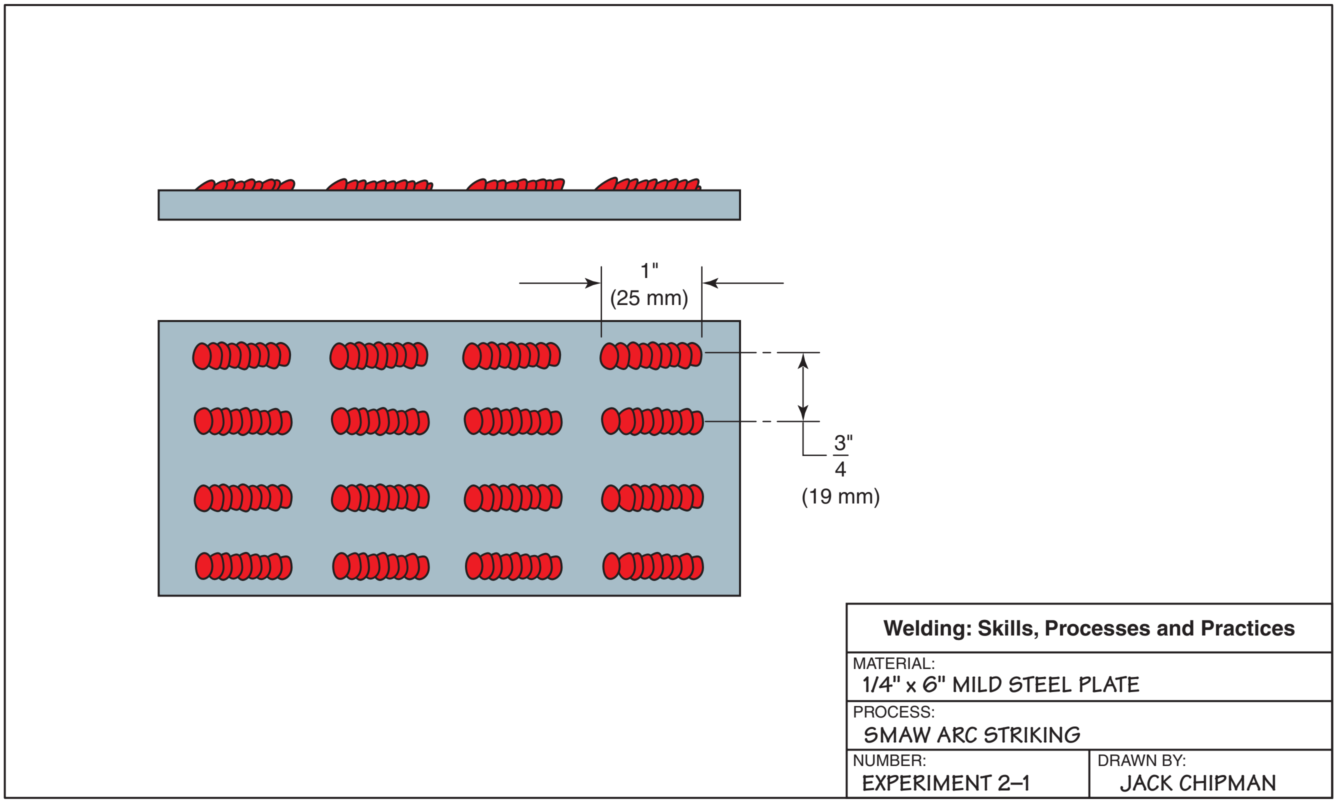
1"
(25 mm)
3"
4
(19 mm)
Welding: Skills, Processes and Practices
MATERIAL:
1/4" x 6" MILD STEEL PLATE
PROCESS:
SMAW ARC STRIKING
NUMBER:
EXPERIMENT 2–1
DRAWN BY:
JACK CHIPMAN

Figure 2.1
Striking an arc and running short beads

Figure 2.2
Striking the arc
Courtesy of Larry Jeffus

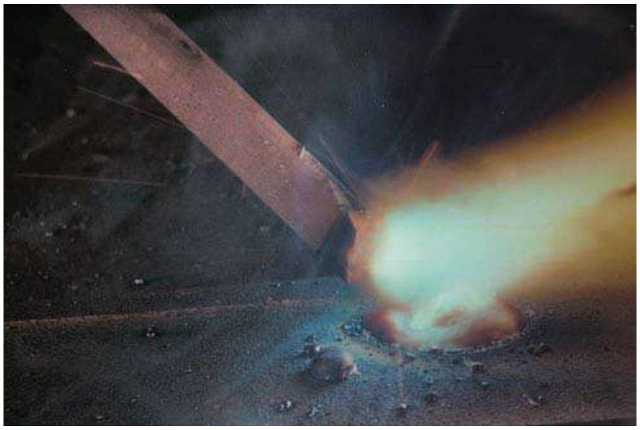

Figure 2.3
If the flux is broken off the end completely or on one side, the arc can be erratic or forced to the side
Courtesy of Larry Jeffus

Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

EXPERIMENT 2-2

Striking the Arc Accurately

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 4

Module 9
Key Indicator 2

Using the same materials and setup as described in Experiment 2-1, you will start the arc at a specific spot in order to prevent damage to the surrounding plate.

Hold the electrode over the desired starting point. After lowering your helmet, swiftly bounce the electrode against the plate, **Figure 2.4.** A lot of practice is required to develop the speed and skill needed to prevent the electrode from sticking to the plate.

A more accurate method of starting the arc involves holding the electrode steady by resting it on your free hand like a pool cue. The electrode is rapidly pushed forward so that it strikes the metal exactly where it should. This is an excellent method of striking an arc. Striking an arc in an incorrect spot may cause damage to the base metal.

Practice starting the arc until you can start it within 1/4 in. (6 mm) of the desired location. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

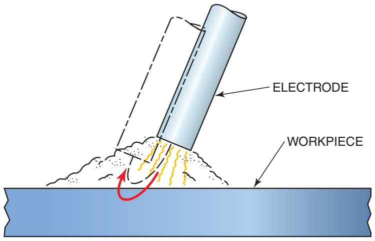

Figure 2.4
Striking the arc on a spot

EFFECT OF TOO-HIGH OR TOO-LOW CURRENT SETTINGS

Each welding electrode must be operated in a specified current (amperage) range, **Table 2.1.** Welding with the current set too low results in poor fusion and poor arc stability, **Figure 2.5.** The weld may have slag or gas inclusions because the molten weld pool was not fluid long enough for the flux to react. Little or no penetration of the weld into the base plate may also be evident. With the current set too low, the arc length is very short. A very short arc length results in frequent shorting and sticking of the electrode.

The core wire of the welding electrode is limited in the amount of current it can carry. As the current is increased, the wire heats up because of electrical resistance. This preheating of the wire causes some of the chemicals in the covering to be burned out too early, **Figure 2.6.** The loss of the proper balance of elements causes poor arc stability. This condition leads to spatter, porosity, and slag inclusions.

Table 2.1 Welding Amperage Range for Common Electrode Types and Sizes

Electrode	Classification					
Size	**E6010**	**E6011**	**E6012**	**E6013**	**E7016**	**E7018**
3/32 in. (2.4 mm)	40–80	50–70	40–90	40–85	75–105	70–110
1/8 in. (3.2 mm)	70–130	85–125	75–130	70–120	100–150	90–165
5/32 in. (4 mm)	110–165	130–160	120–200	130–160	140–190	125–220

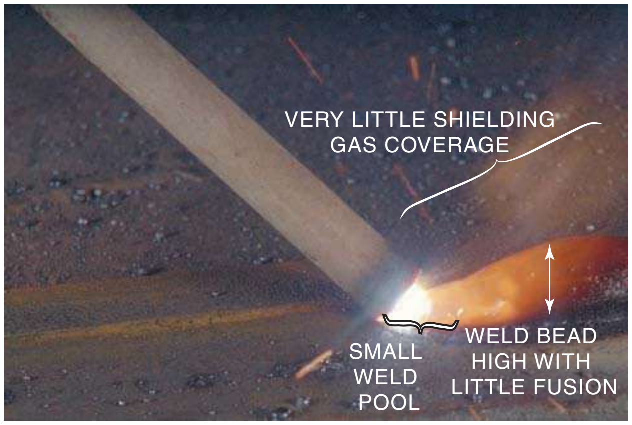
VERY LITTLE SHIELDING GAS COVERAGE
SMALL WELD POOL
WELD BEAD HIGH WITH LITTLE FUSION

Figure 2.5
Welding with the amperage set too low
Courtesy of Larry Jeffus

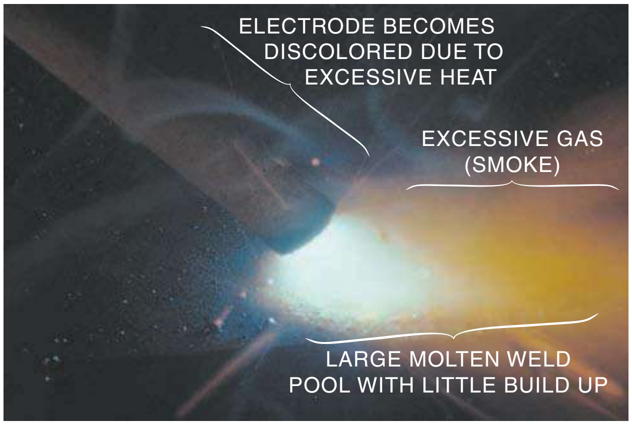
ELECTRODE BECOMES DISCOLORED DUE TO EXCESSIVE HEAT
EXCESSIVE GAS (SMOKE)
LARGE MOLTEN WELD POOL WITH LITTLE BUILD UP

Figure 2.6
Welding with too high an amperage
Courtesy of Larry Jeffus

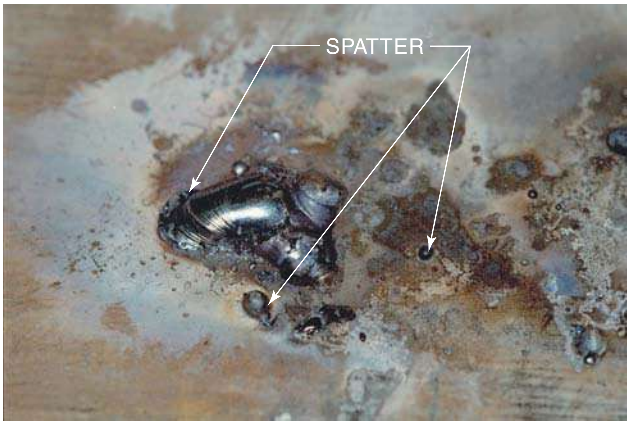

Figure 2.7
Hard weld spatter is fused to base metal and is difficult to remove
Courtesy of Larry Jeffus

An increase in the amount of spatter is also caused by longer arc lengths. The weld bead made at a high amperage setting is wide and flat with deep penetration. The spatter is excessive and is mostly hard. The spatter is called hard because it fuses to the base plate and is difficult to remove, **Figure 2.7.** The electrode covering is discolored more than 1/8 in. (3 mm) to 1/4 in. (6 mm) from the end of the electrode. Extremely high settings may also cause the electrode to discolor, crack, glow red, or burn.

EXPERIMENT 2-3

Effect of Amperage Changes on a Weld Bead

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 3, 4

Module 9
Key Indicator 2

For this experiment, you will need an arc welding machine, welding gloves, safety glasses, welding helmet, appropriate clothing, E6011 or E6013 welding electrodes having a 1/8-in. (3-mm) diameter, and one piece of mild steel plate, 1/4 in. (6 mm) to 1/2 in. (13 mm) thick. You will observe what happens to the weld bead when the amperage settings are raised and lowered.

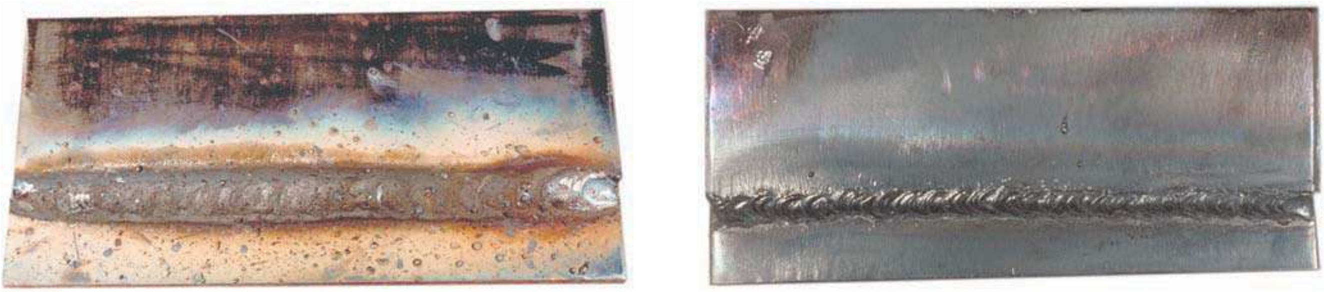

Figure 2.8
Weld before cleaning and after cleaning
Courtesy of Larry Jeffus

Starting with the machine set at approximately 90 A AC or DCRP, strike an arc and make a weld 1 in. (25 mm) long. Break the arc. Raise the current setting by 10 A, strike an arc, and make another weld 1 in. (25 mm) long. Repeat this procedure until the machine amperage is set at the maximum value.

Replace the electrode and reset the machine to 90 A. Make a weld 1 in. (25 mm) long. Stop and lower the current setting by 10 A. Repeat this procedure until the machine amperage is set at a minimum value.

Cool and chip the plate, comparing the different welds for width, buildup, molten weld pool size, spatter, slag removal, and penetration, **Figure 2.8.** In addition, compare the electrode stubs. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

CAUTION

Do not change the current settings during welding. A change in the setting may cause arcing inside the machine, resulting in damage to the machine.

ELECTRODE SIZE AND HEAT

The selection of the correct size of welding electrode for a weld is determined by some or all of the following: the skill of the welder, the thickness of the metal to be welded, the size of the metal, and welding codes or standards. Using small diameter electrodes requires less skill than using large diameter electrodes. The deposition rate, or the rate at which the weld metal is added to the weld, is slower when small diameter electrodes are used. Small diameter electrodes will make acceptable welds on thick plate, but more time is required to make the weld.

Large diameter electrodes may overheat the metal if they are used with thin or small pieces of metal. To determine if a weld is too hot, watch the shape of the trailing edge of the molten weld pool, **Figure 2.9.** Rounded ripples indicate the weld is cooling uniformly and that the heat is not excessive. If the ripples are pointed, the weld is cooling too slowly because of excessive heat. Extreme overheating can cause a burn-through. Once a burn-through occurs, it is hard to repair.

To correct an overheating problem, a welder can turn down the amperage, use a shorter arc, travel at a faster rate, use a chill plate (a large piece of metal used to absorb excessive heat), or use a smaller electrode at a lower current setting.

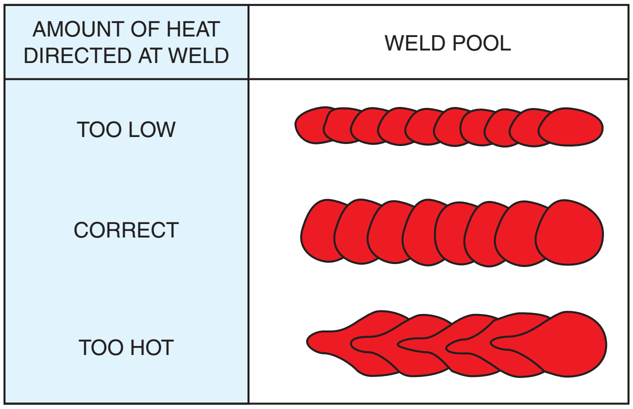

Figure 2.9
The effect on the shape of the molten weld pool caused by the heat input

EXPERIMENT 2-4

Excessive Heat

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 3, 4

Module 9
Key Indicator 2

Using a properly set up and adjusted arc welding machine, the proper safety protection, E6011 or E6013 welding electrodes having a 1/8-in. (3-mm) diameter, and three pieces of mild steel plate, 1/8 in. (3 mm), 3/16 in. (5 mm), and 1/4 in. (6 mm) thick, you will observe the effects of overheating on the weld. Make a stringer weld on each of the three plates using the same amperage setting, travel rate, and arc length for each weld. Cool and chip the welds. Then compare the weld beads for width, reinforcement, and appearance.

Using the same amperage settings, make additional welds on the 1/8-in. (3-mm) and 3/16-in. (5-mm) plates. Vary the arc lengths and travel speeds for these welds. Cool and chip each weld and compare the beads for width, reinforcement, and appearance. Make additional welds on the 1/8-in. (3-mm) and 3/16-in. (5-mm) plates, using the same arc length and travel speed as in the earlier part of this experiment, but at a lower amperage setting. Cool and chip the welds and compare the beads for width, reinforcement, and appearance.

Welders often use the terms *heat* and *amperage* as interchangeable when they are speaking about making changes to the welding current. For example, a welder may say, “Turn up the heat a little,” or “Turn up the amperage a little.” In both cases what’s being asked is that the welding amperage be increased a little.

The plates should be cooled between each weld so that the heat from the previous weld does not affect the test results. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

ARC LENGTH

The **arc length** is the distance the arc must jump from the end of the electrode to the plate or weld pool surface. As the weld progresses, the electrode becomes shorter as it is consumed. To maintain a constant arc length, the electrode must be lowered continuously. Maintaining a con-

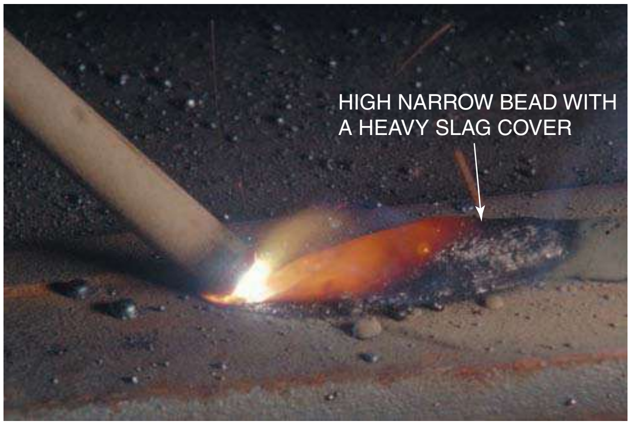

Figure 2.10
Welding with too short an arc length
Courtesy of Larry Jeffus

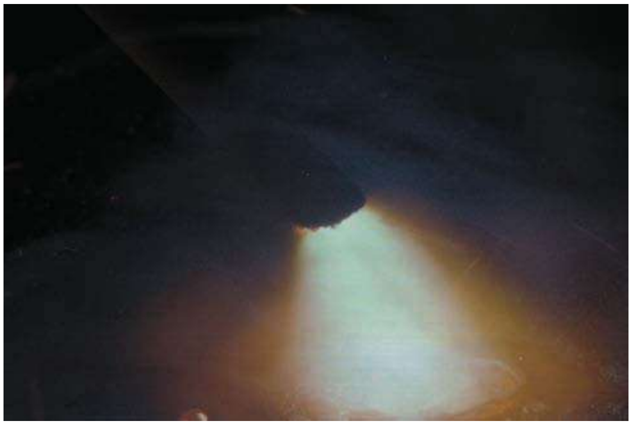

Figure 2.11
Welding with too long an arc length
Courtesy of Larry Jeffus

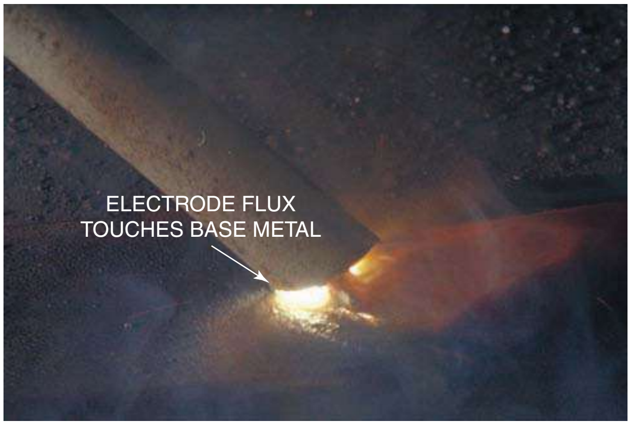

Figure 2.12
Welding with a drag technique
Courtesy of Larry Jeffus

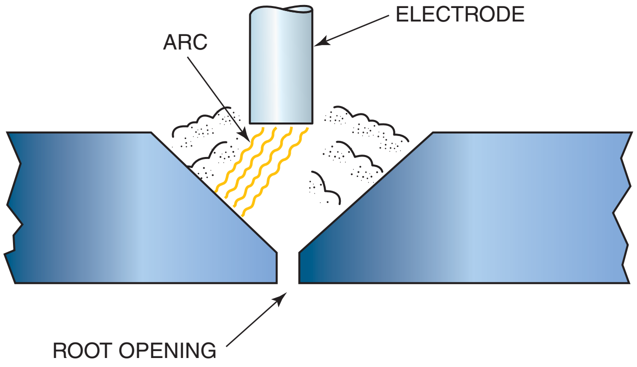

Figure 2.13
The arc may jump to the closest metal, reducing root penetration

stant arc length is important, as too great a change in the arc length will adversely affect the weld.

As the arc length is shortened, metal transferring across the gap may short out the electrode, causing it to stick to the plate. The weld that results from a short arc is narrow and has a high buildup, **Figure 2.10.**

Long arc lengths produce more spatter because the metal being transferred may drop outside of the molten weld pool. The weld is wider and has little buildup, **Figure 2.11.**

There is a narrow range for the arc length in which stability is maintained, metal transfer is smooth, spatter is minimized, and the bead shape is controlled. Factors affecting the length are the type of electrode, joint design, metal thickness, and current setting.

Some welding electrodes, such as E7024, have a thick flux covering. The rate at which the covering melts is slow enough to permit the electrode coating to be rested against the plate. The arc burns back inside the covering as the electrode is dragged along touching the joint, **Figure 2.12.** For that reason, electrodes like E7024 are sometimes referred to as drag electrodes, or drag rods. For this type of welding electrode, the arc length is maintained by the electrode covering.

An arc will jump to the closest metal conductor. On joints that are deep or narrow, the arc is pulled to one side and not to the root, **Figure 2.13.** As a result, the root fusion is reduced or may be nonexistent, thus causing a poor weld. If a very short arc is used, the arc is forced into the root for better fusion.

Because shorter arcs produce less heat and penetration, they are best suited for use on thin metal or thin-to-thick metal joints. Using this technique, metal as thin as 16 gauge can be arc welded easily. Higher amperage settings are required to maintain a short arc that gives good fusion with a minimum of slag inclusions. The higher settings, however, must be within the **amperage range** for the specific electrode.

Finding the correct arc length often requires some trial and adjustment. Most welding jobs require an arc length of 1/8 in. (3 mm) to 3/8 in. (10 mm) when using a 1/8-in. electrode, but this distance varies. It may be necessary to change the arc length when welding to adjust for varying welding conditions.

EXPERIMENT 2-5

Effect of Changing the Arc Length on a Weld

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 3, 4

Module 9
Key Indicator 2

Using an arc welding machine, welding gloves, safety glasses, welding helmet, appropriate clothing, E6011 or E6013 welding electrodes having a 1/8-in. (3-mm) diameter, and one piece of mild steel plate, 1/4 in. (6 mm) to 1/2 in. (13 mm) thick, you will observe the effect of changing the arc length on a weld.

Starting with the welding machine set at approximately 90 A AC or DCRP, strike an arc and make a weld 1 in. (25 mm) long. Continue welding while slowly increasing the arc length until the arc is broken. Restart the arc and make another weld 1 in. (25 mm) long. Welding should again be continued while slowly shortening the arc length until the arc stops. Quickly break the electrode free from the plate, or release the electrode by squeezing the lever on the electrode holder.

Cool and chip both welds. Compare both welding beads for width, reinforcement, uniformity, spatter, and appearance. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

ELECTRODE ANGLE

The **electrode angle** is measured from the electrode to the surface of the metal. The term used to identify the electrode angle is affected by the direction of travel, generally leading or trailing, **Figure 2.14.** The relative angle is important because there is a jetting force blowing the metal and flux from the end of the electrode to the plate.

Leading Angle

A leading electrode angle pushes molten metal and slag ahead of the weld, **Figure 2.15.** When welding in the flat position, caution must be taken to prevent overlap and slag inclusions. The solid metal ahead of the weld cools and solidifies the molten filler metal and slag before they can melt

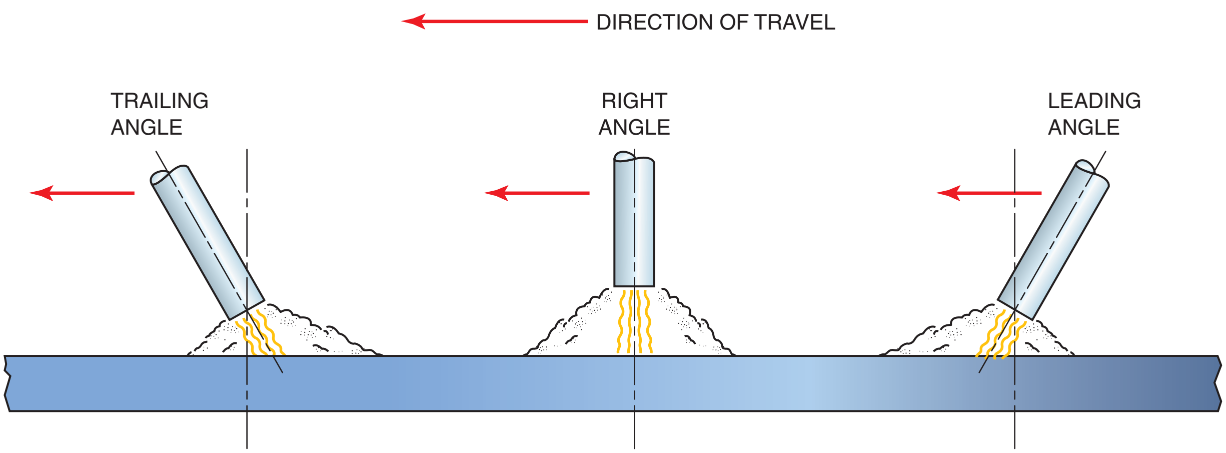
DIRECTION OF TRAVEL
TRAILING ANGLE
RIGHT ANGLE
LEADING ANGLE

Figure 2.14
Direction of travel and electrode angle

the solid metal. This rapid cooling prevents the metals from fusing together, **Figure 2.16.** As the weld passes over this area, heat from the arc may not melt it. As a result, some overlap and slag inclusions are left.

The following are suggestions for preventing cold lap and slag inclusions:

- Use as little leading angle as possible.
- Ensure that the arc melts the base metal completely, **Figure 2.17.**
- Use a penetrating-type electrode that causes little buildup.
- Move the arc back and forth across the molten weld pool to fuse both edges.

A leading angle can be used to minimize penetration or to help hold molten metal in place for vertical welds, **Figure 2.18.**

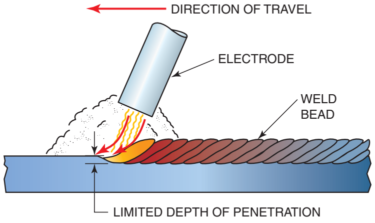
DIRECTION OF TRAVEL
ELECTRODE
WELD BEAD
LIMITED DEPTH OF PENETRATION

Figure 2.15
Leading, lag, or pushing electrode angle

Trailing Angle

A trailing electrode angle pushes the molten metal away from the leading edge of the molten weld pool toward the back where it solidifies, **Figure 2.19.** As the molten metal is forced away from the bottom of the

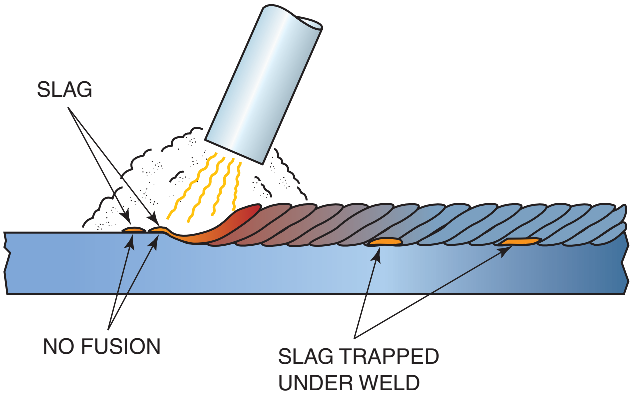
SLAG
NO FUSION
SLAG TRAPPED UNDER WELD

Figure 2.16
Some electrodes, such as E7018, may not remove the deposits ahead of the molten weld pool, resulting in discontinuities within the weld

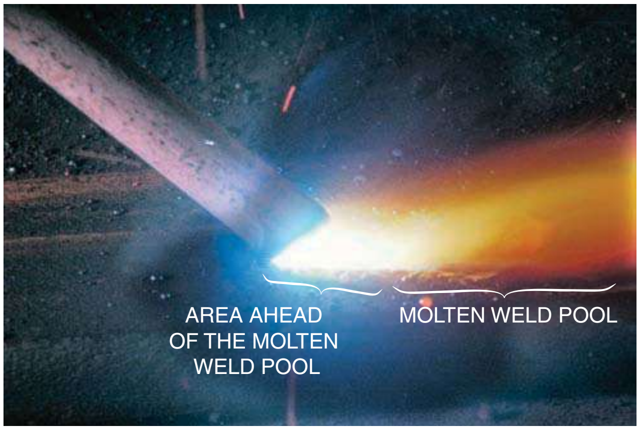
AREA AHEAD OF THE MOLTEN WELD POOL
MOLTEN WELD POOL

Figure 2.17
Metal being melted ahead of the molten weld pool helps to ensure good weld fusion
Courtesy of Larry Jeffus

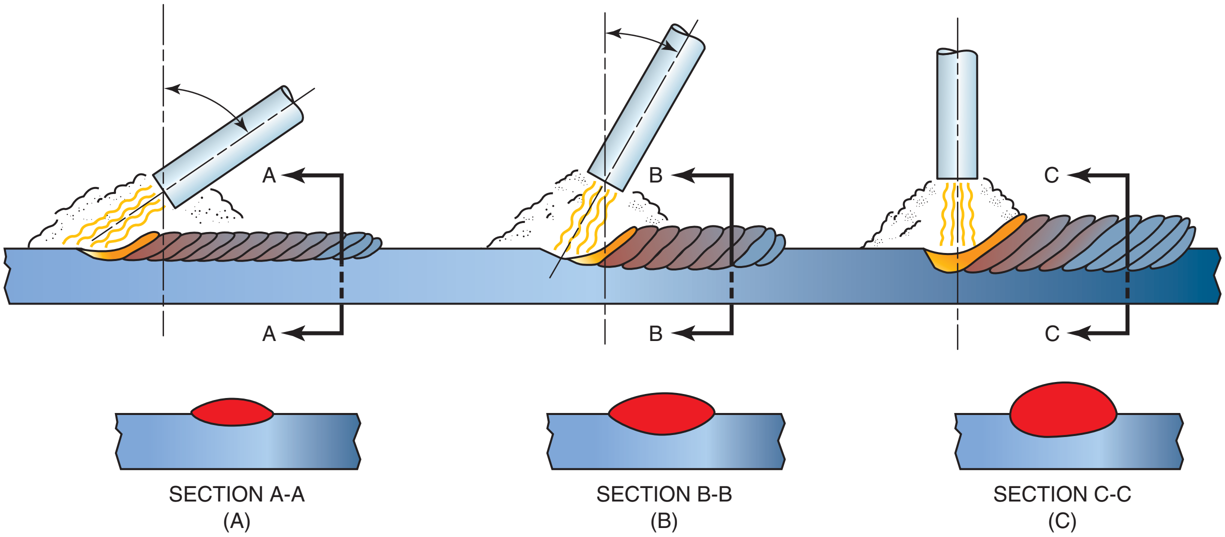
A
A
B
B
C
C
SECTION A-A
(A)
SECTION B-B
(B)
SECTION C-C
(C)

Figure 2.18
Effect of a leading angle on weld bead buildup, width, and penetration. As the angle increases toward the vertical position (C), penetration increases

weld, the arc melts more of the base metal, which results in deeper penetration. The molten metal pushed to the back of the weld solidifies and forms reinforcement for the weld, **Figure 2.20.**

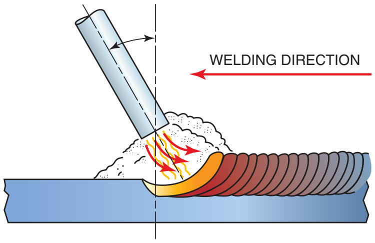
WELDING DIRECTION

Figure 2.19
Trailing electrode angle

EXPERIMENT 2-6

Effect of Changing the Electrode Angle on a Weld

Using a properly set up and adjusted arc welding machine, the proper safety protection, E6011 welding electrodes having a 1/8-in. (3-mm) diameter, and one piece of mild steel plate, 1/4 in. (6 mm) to 1/2 in. (13 mm) thick, you will observe the effect of changes in the electrode angle on a weld.

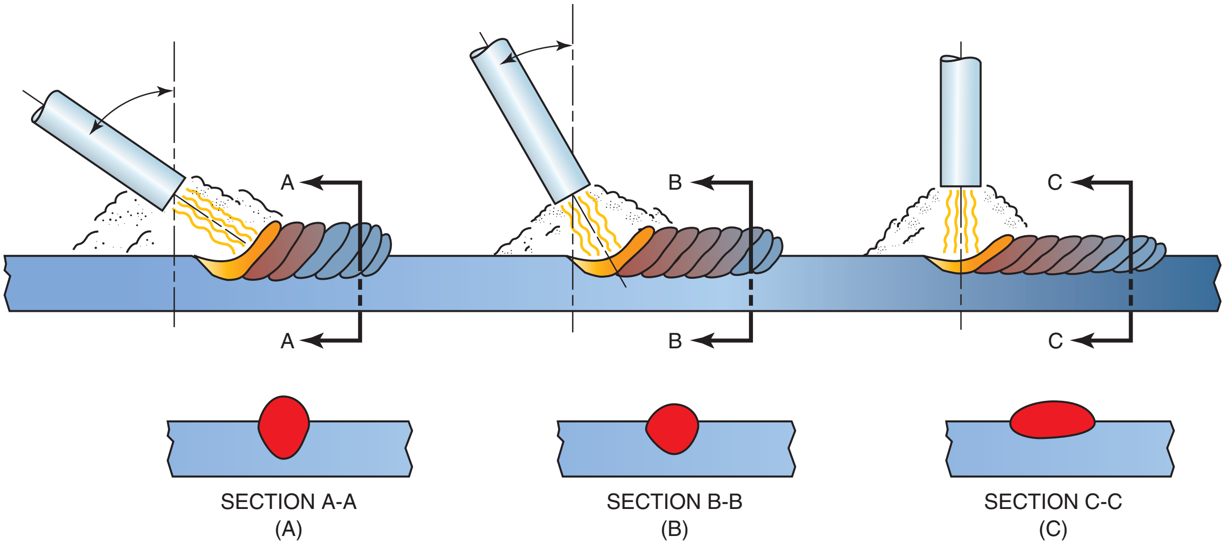
A
A
B
B
C
C
SECTION A-A
(A)
SECTION B-B
(B)
SECTION C-C
(C)

Figure 2.20
Effect of a trailing angle on weld bead buildup, width, and penetration. Section A-A shows more weld buildup due to a greater angle of the electrode

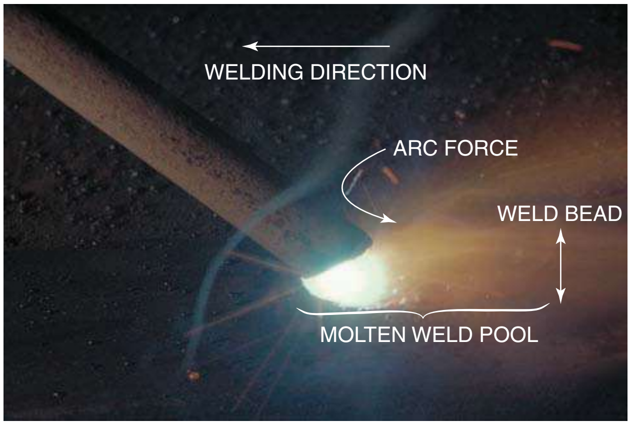

Figure 2.21
Welding with a trailing angle
Courtesy of Larry Jeffus

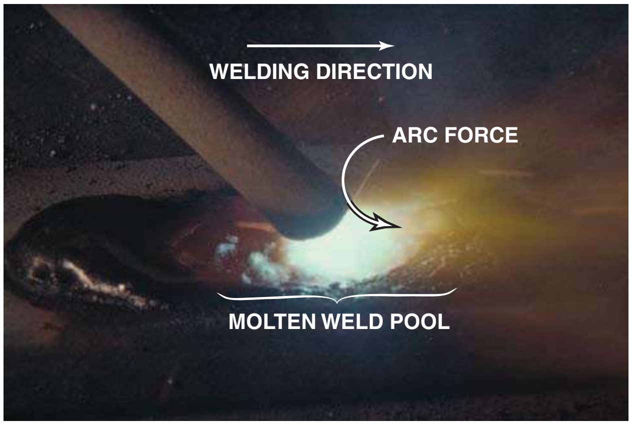

Figure 2.22
Welding with a leading angle
Courtesy of Larry Jeffus

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 3, 4

Module 9
Key Indicator 2

Start welding with a sharp trailing angle. Make a weld about 1 in. (25 mm) long. Closely observe the molten weld pool at the points shown in **Figure 2.21.** Slowly increase the electrode angle and continue to observe the weld.

When you reach a 90° electrode angle, make a weld about 1 in. (25 mm) long. Observe the parts of the molten weld pool as shown in Figure 2.21.

Continue welding and change the electrode angle to a sharp leading angle. Observe the molten weld pool at the points shown in **Figure 2.22.**

During this experiment, you must maintain a constant arc length, travel speed, and weave pattern if the observations and results are to be accurate.

Cool and chip the weld. Compare the weld bead for uniformity in width, reinforcement, and appearance. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

ELECTRODE MANIPULATION

The movement or weaving of the welding electrode can control the following characteristics of the weld bead: penetration, buildup, width, porosity, undercut, overlap, and slag inclusions. The exact **weave pattern** for each weld is often the personal choice of the welder. However, some patterns are especially helpful for specific welding situations. The pattern selected for a flat (1G) butt joint is not as critical as is the pattern selection for other joints and other positions.

Many weave patterns are available for the welder to use. **Figure 2.23** shows ten different patterns that can be used for most welding conditions.

The circular pattern is often used for flat position welds on butt, tee, outside corner joints, and for buildup or surfacing applications. The circle can be made wider or longer to change the bead width or penetration, **Figure 2.24.**

The “C” and square patterns are both good for most 1G (flat) welds, but can also be used for vertical (3G) positions. These patterns can also be

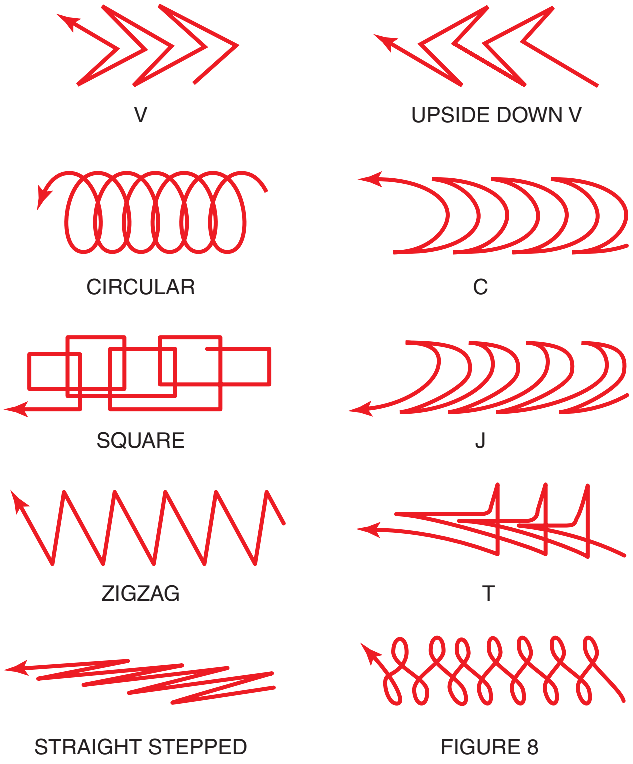
V
UPSIDE DOWN V
CIRCULAR
C
SQUARE
J
ZIGZAG
T
STRAIGHT STEPPED
FIGURE 8

Figure 2.23
Weave patterns

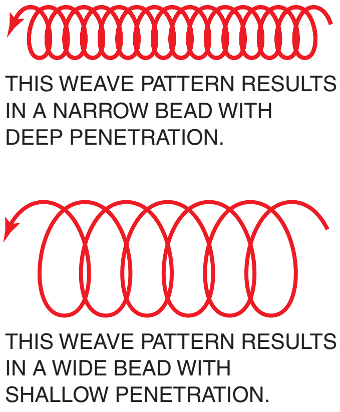
THIS WEAVE PATTERN RESULTS IN A NARROW BEAD WITH DEEP PENETRATION.
THIS WEAVE PATTERN RESULTS IN A WIDE BEAD WITH SHALLOW PENETRATION.

Figure 2.24
Changing the weave pattern width to change the weld bead characteristics

used if there is a large gap to be filled when both pieces of metal are nearly the same size and thickness.

The “J” pattern works well on flat (1F) lap joints, all vertical (3G) joints, and horizontal (2G) butt and lap (2F) welds. This pattern allows the heat to be concentrated on the thicker plate, **Figure 2.25.** It also allows the reinforcement to be built up on the metal deposited during the first part of the pattern. As a result, a uniform bead contour is maintained during out-of-position welds.

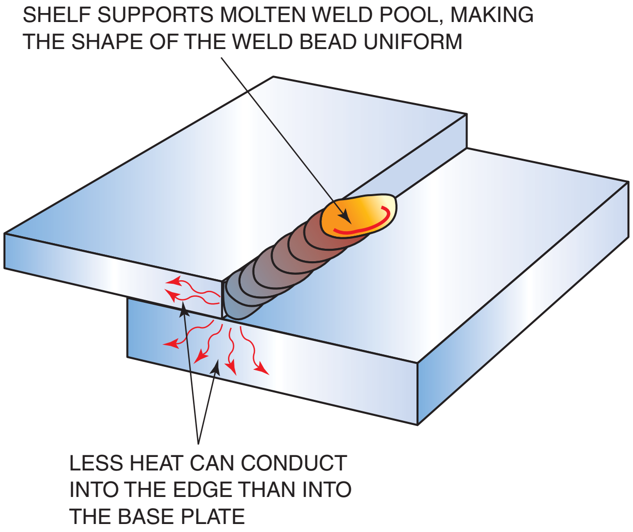
SHELF SUPPORTS MOLTEN WELD POOL, MAKING THE SHAPE OF THE WELD BEAD UNIFORM
LESS HEAT CAN CONDUCT INTO THE EDGE THAN INTO THE BASE PLATE

Figure 2.25
The “J” pattern allows the heat to be concentrated on the thicker plate

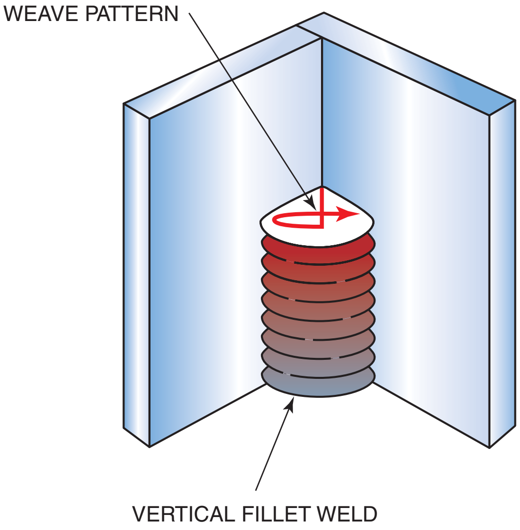
WEAVE PATTERN
VERTICAL FILLET WELD

Figure 2.26
“Inverted T” pattern

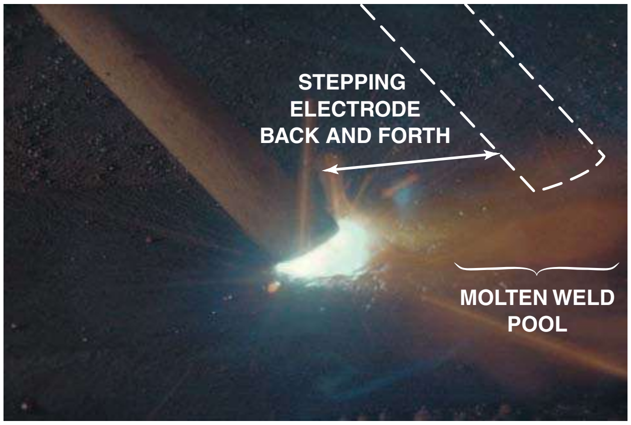

Figure 2.27
The electrode is moved slightly forward and then returned to the weld pool
Courtesy of Larry Jeffus

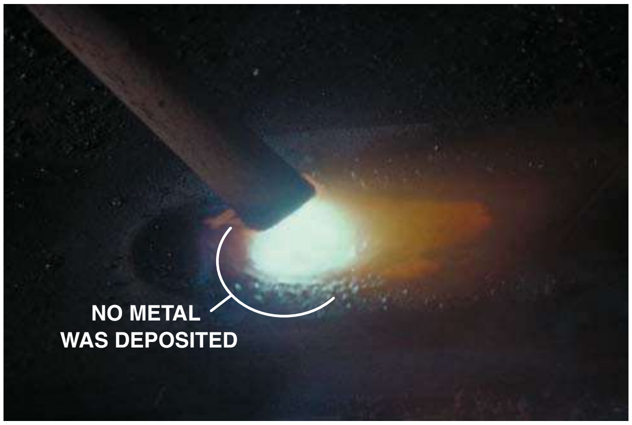

Figure 2.28
The electrode does not deposit metal or melt the base metal
Courtesy of Larry Jeffus

The inverted "T" pattern works well with fillet welds in the vertical (3F) and overhead (4F) positions, **Figure 2.26.** It also can be used for deep groove welds for the hot pass. The top of the "T" can be used to fill in the toe of the weld to prevent undercutting.

The straight step pattern can be used for stringer beads, root pass welds, and multiple pass welds in all positions. For this pattern, the smallest quantity of metal is molten at one time as compared to other patterns. Therefore, the weld is more easily controlled. At the same time that the electrode is stepped forward, the arc length is increased so that no metal is deposited ahead of the molten weld pool, **Figure 2.27** and **Figure 2.28.** This action allows the molten weld pool to cool to a controllable size. In addition, the arc burns off any paint, oil, or dirt from the metal before it can contaminate the weld.

The figure-eight pattern and the zigzag pattern are used as cover passes in the flat and vertical positions. Do not weave more than 2-1/2 times the width of the electrode. These patterns deposit a large quantity of metal at one time. A shelf can be used to support the molten weld pool when making vertical welds using either of these patterns, **Figure 2.29.**

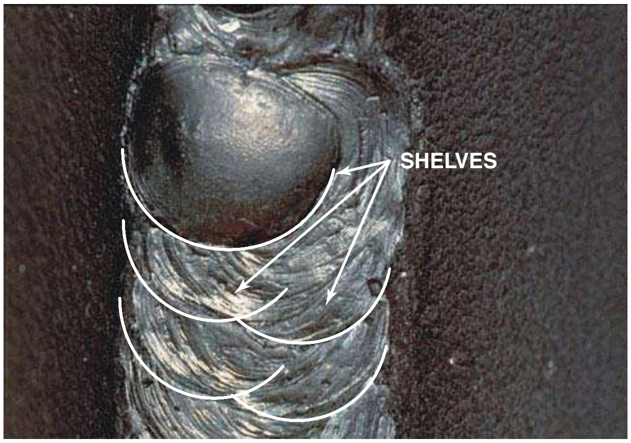

Figure 2.29
Using the shelf to support the molten pool for vertical welds
Courtesy of Larry Jeffus

POSITIONING OF THE WELDER AND THE PLATE

The welder should be in a relaxed, comfortable position before starting to weld. A good position is important for both the comfort of the welder and the quality of the welds. Welding in an awkward position can cause welder fatigue, which leads to poor welder coordination and poor-quality welds. Welders must have enough freedom of movement so that they do not need to change position during a weld. Body position changes should be made only during electrode changes.

When the welding helmet is down, the welder is blind to the surroundings. Due to the arc, the field of vision of the welder is also very limited. These factors often cause the welder to sway. To stop this swaying, the welder should lean against or hold on to a stable object. When welding, even if a welder is seated, touching a stable object will make that welder more stable and will make welding more relaxing.

Welding is easier if the welder can find the most comfortable angle. The welder should be in either a seated or a standing position in front of the welding table. The welding machine should be turned off. With an electrode in place in the electrode holder, the welder can draw a straight line along the plate to be welded. By turning the plate to several different angles, the welder should be able to determine which angle is most comfortable for welding, **Figure 2.30.**

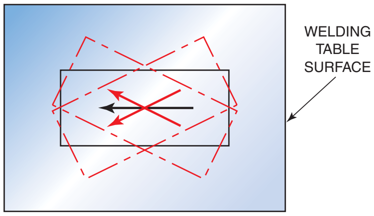

Figure 2.30
Change the plate angle to find the most comfortable welding position

PRACTICE WELDS

Practice welds are grouped according to the type of joint and the type of welding electrode. The welder or instructor should select the order in which the welds are made. The stringer beads should be practiced first in each position before the welder tries the different joints in each position. Some time can be saved by starting with the stringer beads. If this is done, it is not necessary to cut or tack the plate together, and a number of beads can be made on the same plate.

Students will find it easier to start with butt joints. The lap, tee, and outside corner joints are all about the same level of difficulty. Starting with the flat position allows the welder to build skills slowly so that out-of-position welds become easier to do. The horizontal tee and lap welds are almost as easy to make as the flat welds. Overhead welds are as simple to make as vertical welds, but they are harder to position. Horizontal butt welds are more difficult to perform than most other welds.

Electrodes

Arc welding electrodes used for practice welds in the following chapters are grouped into three filler metal (F-number) classes according to their major welding characteristics. E6012 and E6013 fall into the F2 group, E6010 and E6011 fall into the F3 group, and E7016 and E7018 fall into the F4 group. The F1 electrode group (not used in this chapter) is limited flat and horizontal fillet welds, E7024 and E7028 are examples of F1 electrodes.

F1 E7024 and E7028 Electrodes

These electrodes are primarily produced with iron-powder-based fluxes that are substantially thicker than the other groups. As a result, they have very high deposition rates and produce welds with medium to low penetration, but are restricted to groove welds in the flat position and fillet welds in the flat or horizontal positions.

F2 E6012 and E6013 Electrodes

These electrodes have rutile-based fluxes, giving a smooth, easy arc with a thick slag left on the weld bead. They may be used in all positions and do not require special rod ovens. These electrodes are popular because slag is easily removed, and they may be used with all polarities to produce welds with medium to low penetration.

F3 E6010 and E6011 Electrodes

Both of these electrodes have cellulose-based fluxes. As a result, these electrodes have a forceful deep penetrating arc with little slag left on the weld bead. They do not require a rod oven, and they are often the electrode of choice when surface conditions on the base metal are less than optimal. F3 electrodes are also commonly used for open root welds on plate and pipe.

F4 E7016 and E7018 Electrodes

Both of these electrodes have a mineral-based flux. The resulting medium penetration arc is smooth and easy, with a very heavy slag left on the weld bead. F4 electrodes are also referred to as low hydrogen or low-hi electrodes. They require special handling and storage in a rod oven after being removed from their factory packaging. Refer to manufacturer's requirements or the applicable welding code for specific handling directions.

The cellulose- and rutile-based groups of electrodes have characteristics that make them the best electrodes for starting specific welds. The electrodes with the **cellulose-based fluxes** do not have heavy slags that may interfere with the welder's view of the weld. This feature is an advantage for flat tee and lap joints. Electrodes with the **rutile-based fluxes** (giving an easy arc with low spatter) are easier to control and are used for flat stringer beads and butt joints.

Unless a specific electrode has been required by a welding procedure specification (WPS), welders can select what they consider to be the best electrode for a specific weld. Without a WPS, the welder has the final choice. An accomplished welder can make defect-free welds on all types of joints using all types of electrodes in any weld position.

Electrodes with **mineral-based fluxes** should be the last choice. Welds with a good appearance are more easily made with these electrodes, but strong welds are hard to obtain. Without special care being taken during the start of the weld, porosity will be formed in the weld. **Figure 2.31** shows a starting tab used to prevent this porosity from becoming part of the finished weld.

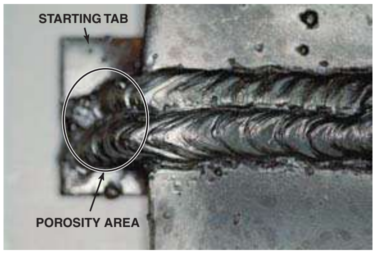

Figure 2.31
Porosity is found on the starting tab where it will not affect the weld
Courtesy of Larry Jeffus

STRINGER BEADS

A straight weld bead on the surface of a plate, with little or no side-to-side electrode movement, is known as a **stringer bead.** Stringer beads are used by students to practice maintaining arc length and electrode angle so that their welds will be straight, uniform, and free from defects. Stringer beads,

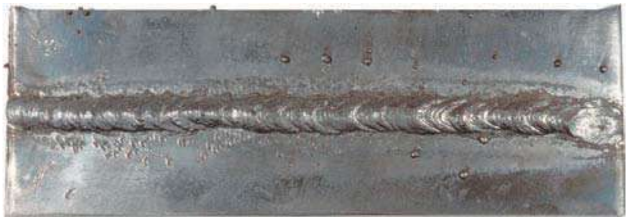

Figure 2.32
Stringer bead
Courtesy of Larry Jeffus

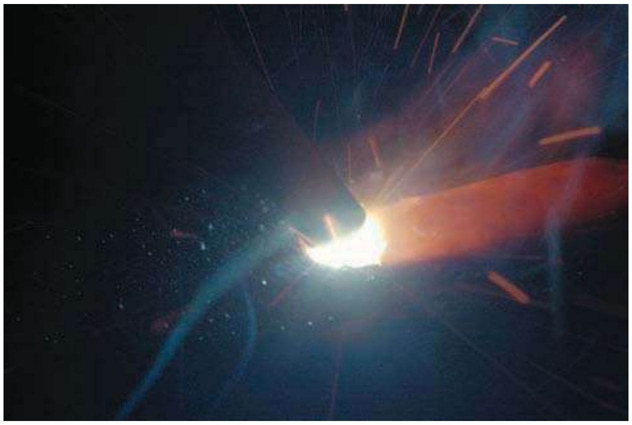

Figure 2.33
New welders frequently see only the arc and sparks from the electrode
Courtesy of Larry Jeffus

Figure 2.32, are also used to set the machine amperage and for buildup or surfacing applications. Stringer beads are the most commonly used type of bead for vertical, horizontal, and overhead welds.

The stringer bead should be straight. A beginning welder needs time to develop the skill of viewing the entire welding area. At first, the welder sees only the arc, **Figure 2.33.** With practice, the welder begins to see parts of the molten weld pool. After much practice, the welder will see the molten weld pool (front, back, and both sides), slag, buildup, and the surrounding plate, **Figure 2.34.** Often, at this skill level, the welder may not even notice the arc.

A straight weld is easily made once the welder develops the ability to view the entire welding zone. The welder will occasionally glance around to ensure that the weld is straight. In addition, it can be noted if the weld is uniform and free from defects. The ability of the welder to view the entire weld area is demonstrated by making consistently straight and uniform stringer beads.

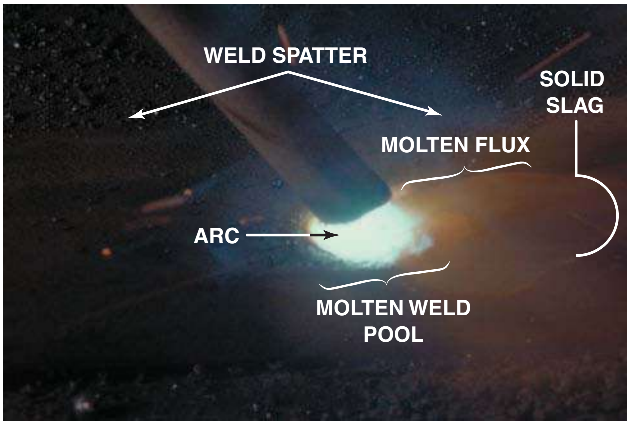

Figure 2.34
More experienced welders can see the molten pool, metal being transferred across the arc, and penetration into the base metal
Courtesy of Larry Jeffus

After making practice stringer beads, a variety of weave bead patterns should be practiced to gain the ability to control the molten weld pool when welding out of position.

PRACTICE 2-2

Straight Stringer Beads in the Flat Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using a properly set up and adjusted arc welding machine, proper safety protection as demonstrated in Practice 2-1, arc welding electrodes with a 1/8-in. (3-mm) diameter, and one piece of mild steel plate, 6 in. (152 mm) long × 1/4 in. (6 mm) thick, you will make straight stringer beads.

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

- Starting at one end of the plate, make a straight weld the full length of the plate.
- Watch the molten weld pool at this point, not the end of the electrode. As you become more skillful, it is easier to watch the molten weld pool.
- Repeat the beads with all three (F) groups of electrodes until you have consistently good beads.
- Cool, chip, and inspect the bead for defects after completing it. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-3

Stringer Beads in the Vertical Up Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using the same setup, materials, and electrodes as listed in Practice 2-2, you will make vertical up stringer beads. Start with the plate at a 45° angle.

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

This technique is the same as that used to make a vertical weld. However, a lower level of skill is required at 45°, and it is easier to develop your skill. After the welder masters the 45° angle, the angle is increased successively until a vertical position is reached, **Figure 2.35.**

Before the molten metal drips down the bead, the back of the molten weld pool will start to bulge, **Figure 2.36.** When this happens, increase the speed of travel and the weave pattern.

Cool, chip, and inspect each completed weld for defects. Repeat the beads as necessary with all three (F) groups of electrodes until consistently good beads are obtained in this position. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

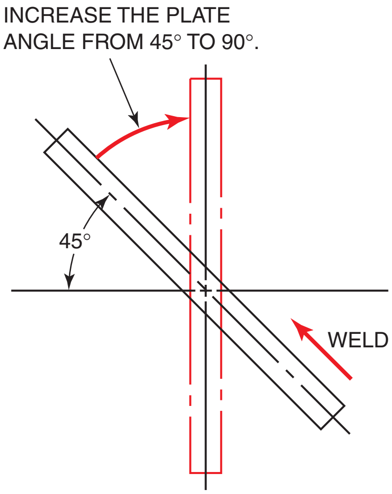

Figure 2.35
Once the 45° angle is mastered, the plate angle is increased successively until a vertical position (90°) is reached

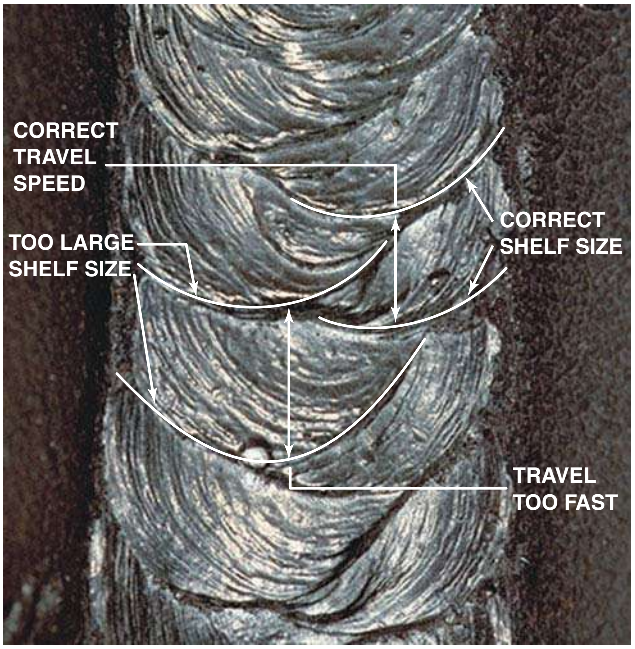

Figure 2.36
E7018 vertical up weld
Courtesy of Larry Jeffus

PRACTICE 2-4

Horizontal Stringer Beads Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-2, you will make horizontal stringer beads on a plate.

When the welder begins to practice the horizontal stringer bead, the plate may be reclined slightly, **Figure 2.37.** This placement allows the welder to build the required skill by practicing the correct techniques successfully. The "J" weave pattern is suggested for this practice. As the electrode is drawn along the straight back of the "J," metal is deposited. This metal supports the molten weld pool, resulting in a bead with a uniform contour, **Figure 2.38.**

Angling the electrode up and back toward the weld causes more metal to be deposited along the top edge of the weld. Keeping the bead small allows the surface tension to hold the molten weld pool in place.

Gradually increase the angle of the plate until it is vertical and the stringer bead is horizontal. Repeat the beads as needed with all three (F) groups of electrodes until consistently good beads are obtained in this position. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

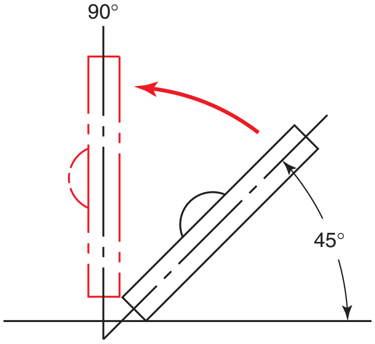

Figure 2.37
Change the plate angle as welding skill improves

SQUARE BUTT JOINT

The **square butt joint** is made by tack welding two flat pieces of plate together, **Figure 2.39.** The space between the plates is called the root opening or root gap. Changes in the root opening will affect penetration. As the

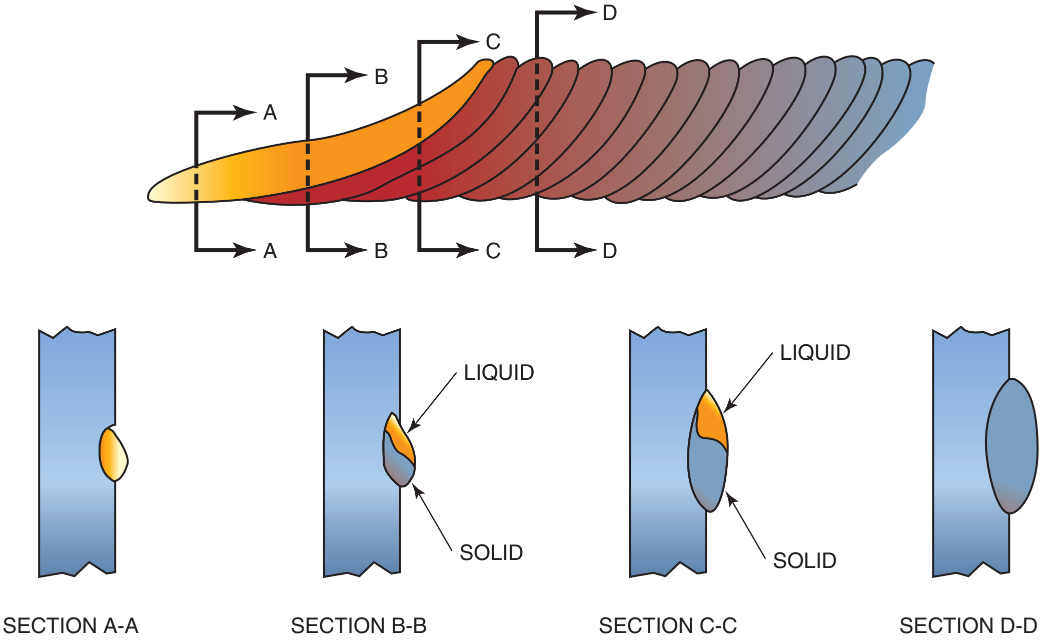
A
B
C
D
A
B
C
D
LIQUID
SOLID
LIQUID
SOLID
SECTION A-A
SECTION B-B
SECTION C-C
SECTION D-D

Figure 2.38
The progression of a horizontal bead

space increases, the weld penetration also increases. The root opening for most butt welds will vary from 0 in. (0 mm) to 1/8 in. (3 mm). Excessively large openings can cause burn-through or a cold lap at the weld root, **Figure 2.40.**

After a butt weld is completed, the plate can be cut apart so it can be used for rewelding. The strips for butt welding should be no smaller than 1 in. (25 mm) wide. If they are too narrow, there will be a problem with heat buildup.

If the plate strips are no longer flat after the weld has been cut out, they can be tack welded together and flattened with a hammer, **Figure 2.41.**

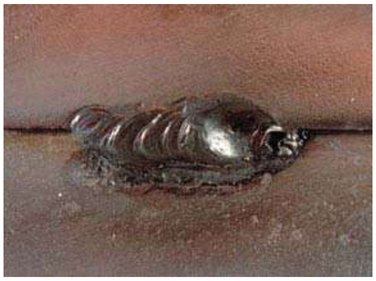

Figure 2.39
The tack weld should be small and uniform to minimize its effect on the final weld
Courtesy of Larry Jeffus

PRACTICE 2-5

Welded Square Butt Joint in the Flat Position (1G) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using a properly set up and adjusted arc welding machine, proper safety protection, arc welding electrodes having a 1/8-in. (3-mm) diameter, and two or more pieces of mild steel plate, 6 in. (152 mm) long × 1/4 in.

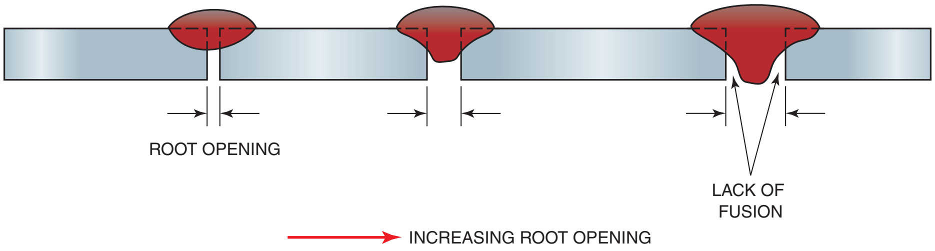
ROOT OPENING
LACK OF FUSION
INCREASING ROOT OPENING

Figure 2.40
Effect of root opening on weld penetration

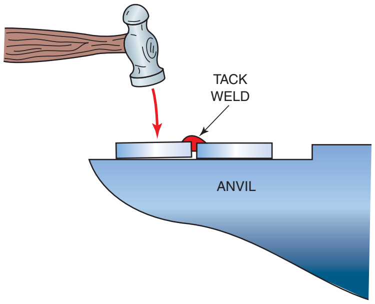
TACK WELD
ANVIL

Figure 2.41
After the plates are tack welded together, they can be forced into alignment by striking them with a hammer

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 6
This practice addresses the "Flat" position portion of the all-position requirement of 6.

Module 9
Key Indicator 2

(6 mm) thick, you will make a welded square butt joint in the flat position, **Figure 2.42**.

Tack weld the plates together and place them flat on the welding table. Starting at one end, establish a molten weld pool on both plates. Hold the electrode in the molten weld pool until it flows together, **Figure 2.43.** After the gap is bridged by the molten weld pool, start weaving the electrode slowly back and forth across the joint. Moving the electrode too quickly from side to side may result in slag being trapped in the joint, **Figure 2.44.**

Continue the weld along the 6-in. (152-mm) length of the joint. Normally, deep penetration is not required for this type of weld. If full plate penetration is required, the edges of the butt joint should be beveled or a larger than normal root gap should be used. Cool, chip, and inspect the weld for uniformity and soundness. Repeat the welds as needed to master all three (F) groups of electrodes in this position. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

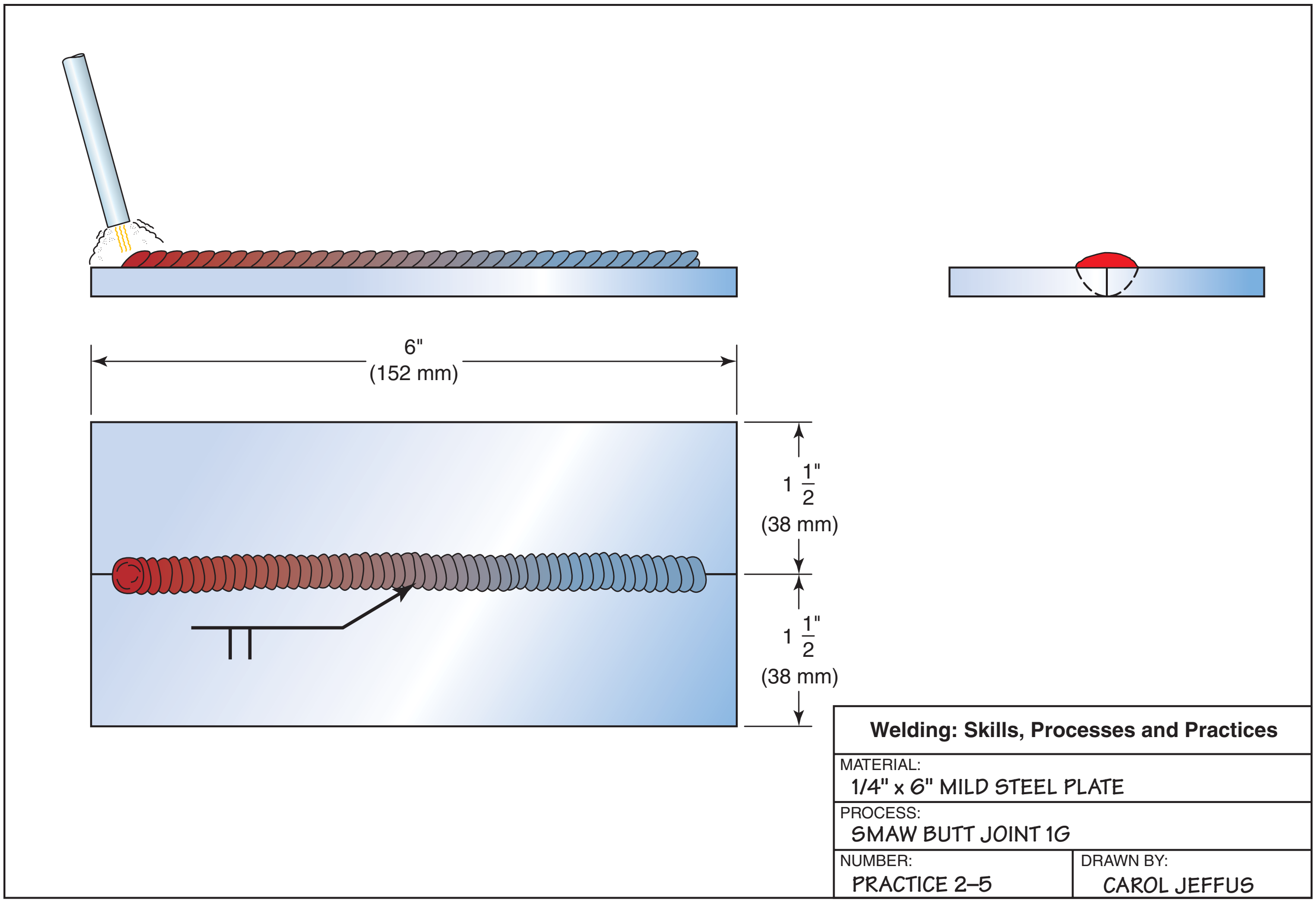
6"
(152 mm)
1 1/2"
(38 mm)
1 1/2"
(38 mm)
Welding: Skills, Processes and Practices
MATERIAL:
1/4" x 6" MILD STEEL PLATE
PROCESS:
SMAW BUTT JOINT 1G
NUMBER:
PRACTICE 2–5
DRAWN BY:
CAROL JEFFUS

Figure 2.42
Square butt joint in the flat position

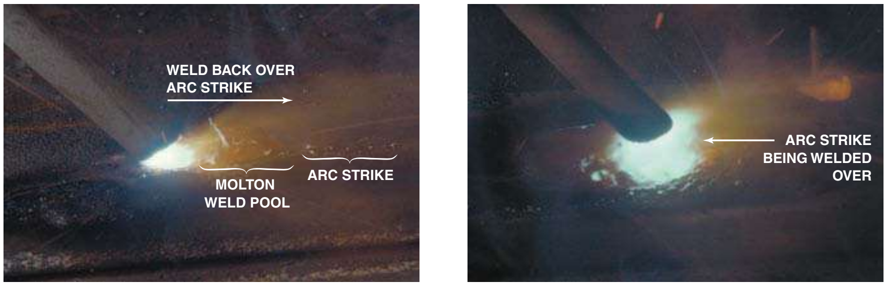
WELD BACK OVER
ARC STRIKE
MOLTON
WELD POOL
ARC STRIKE
ARC STRIKE
BEING WELDED
OVER

Figure 2.43
(A) After the arc is established, hold it in one area long enough to establish the size of molten weld pool desired. (B) Weld back over the arc strike to melt into the weld.
Courtesy of Larry Jeffus

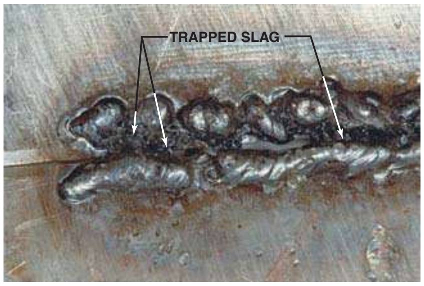
TRAPPED SLAG

Figure 2.44
Moving the electrode from side to side too quickly can result in slag being trapped between the plates
Courtesy of Larry Jeffus

PRACTICE 2-6

Welded Vertical Up (3G) Square Butt Joint Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using the same setup, materials, and electrodes as listed in Practice 2-5, you will make welded vertical up square butt joints.

With the plates at a 45° angle, start at the bottom and make the molten weld pool bridge the gap between the plates, **Figure 2.45.** Build the bead size slowly so that the molten weld pool has a shelf for support. The “C,” “J,” or square weave pattern works well for this joint.

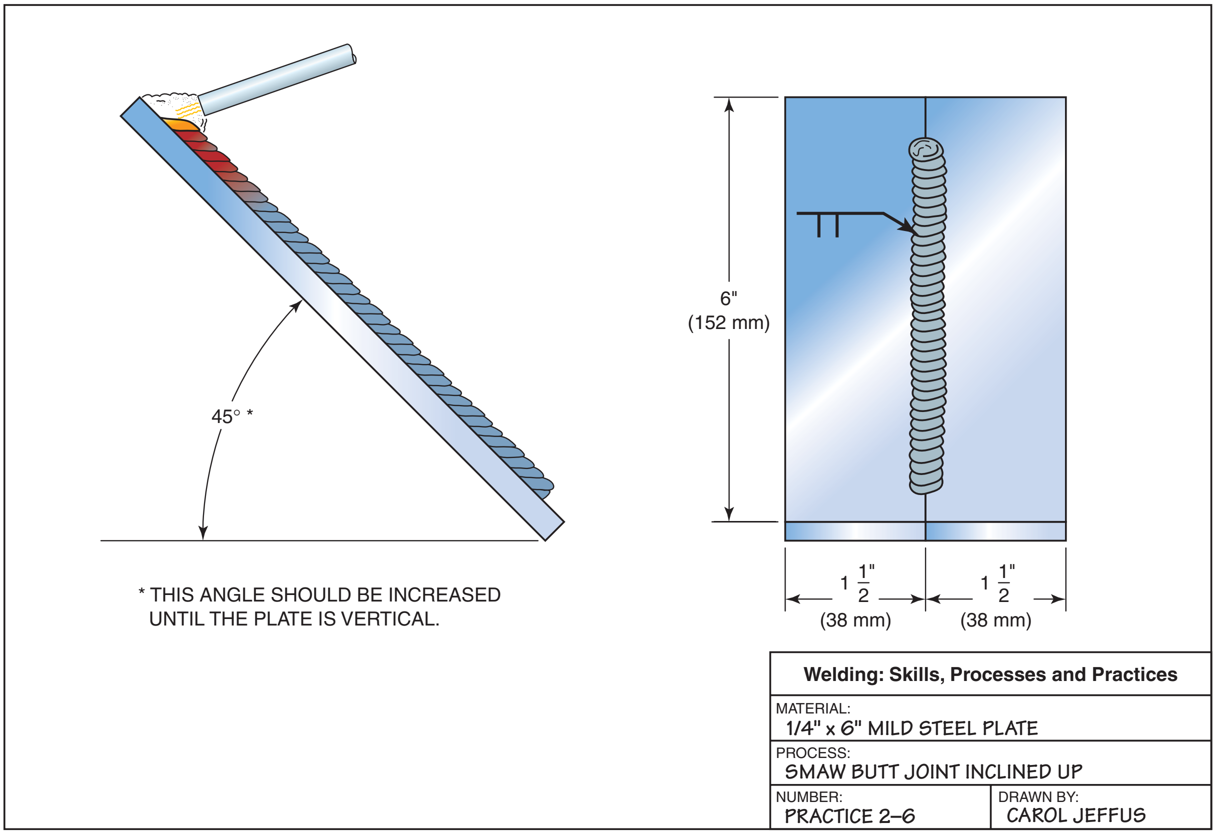
45° *
* THIS ANGLE SHOULD BE INCREASED
UNTIL THE PLATE IS VERTICAL.
6"
(152 mm)
1 1/2"
(38 mm)
1 1/2"
(38 mm)

Welding: Skills, Processes and Practices

MATERIAL: 1/4" x 6" MILD STEEL PLATE

PROCESS: SMAW BUTT JOINT INCLINED UP

NUMBER: PRACTICE 2–6

DRAWN BY: CAROL JEFFUS

Figure 2.45
Square butt joint in the vertical up position

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 6
This practice addresses the "Vertical" position portion of the all-position requirement of 6.

Module 9
Key Indicator 2

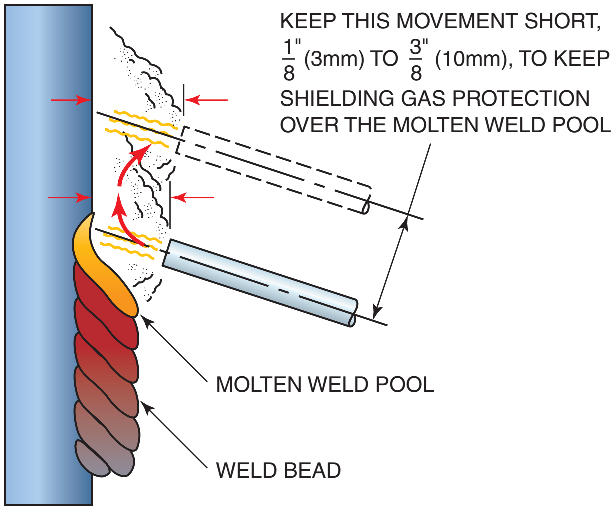
KEEP THIS MOVEMENT SHORT, $\frac{1}{8}$" (3mm) TO $\frac{3}{8}$" (10mm), TO KEEP SHIELDING GAS PROTECTION OVER THE MOLTEN WELD POOL
MOLTEN WELD POOL
WELD BEAD

Figure 2.46
Electrode movement for vertical up welds when using F3 electrodes

As the electrode is moved up the weld, the arc is lengthened slightly so that little or no metal is deposited ahead of the molten weld pool. When the electrode is brought back into the molten weld pool, it should be lowered to deposit metal, **Figure 2.46.**

As skill is developed, increase the plate angle until it is vertical. Cool, chip, and inspect the weld for uniformity and defects. Repeat the welds with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-7

Welded Horizontal (2G) Square Butt Joint Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 6
This practice addresses the "Horizontal" position portion of the all-position requirement of 6.

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as described in Practice 2-5, you will make a welded horizontal square butt joint.

- Start practicing these welds with the plate at a slight angle.
- Strike the arc on the bottom plate and build the molten weld pool until it bridges the gap.

If the weld is started on the top plate, slag will be trapped in the root at the beginning of the weld because of poor initial penetration. The slag may cause the weld to crack when it is placed in service.

The "J" weave pattern is recommended in order to deposit metal on the lower plate so that it can support the bead. By pushing the electrode inward as you cross the gap between the plates, deeper penetration is achieved.

As you acquire more skill, gradually increase the plate angle until it is vertical and the weld is horizontal.

- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding

machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

EDGE WELD

An edge weld joint is made by placing the edges of the plate evenly, **Figure 2.47.** When assembling the edge joint the plates should be clamped tightly together; there should not be any gap between the plates. Both edges of the plate assembly can be welded. Make the tack welds to hold the plates together along the ends of the joint, **Figure 2.48.**

The size of the weld should equal the thickness of the plate being joined. A good indication the weld is being made large enough is when the weld bead width is equal to the width of the joint, **Figure 2.49.** The weld bead should also have a slight buildup.

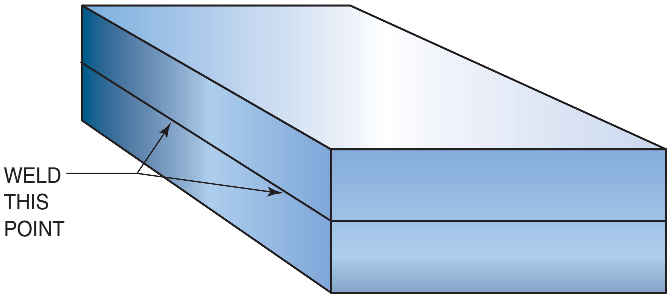

Figure 2.47
Edge joint

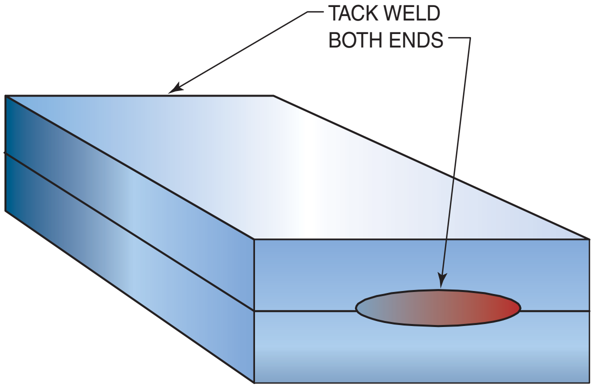

Figure 2.48
Make tack welds at the ends of the joint

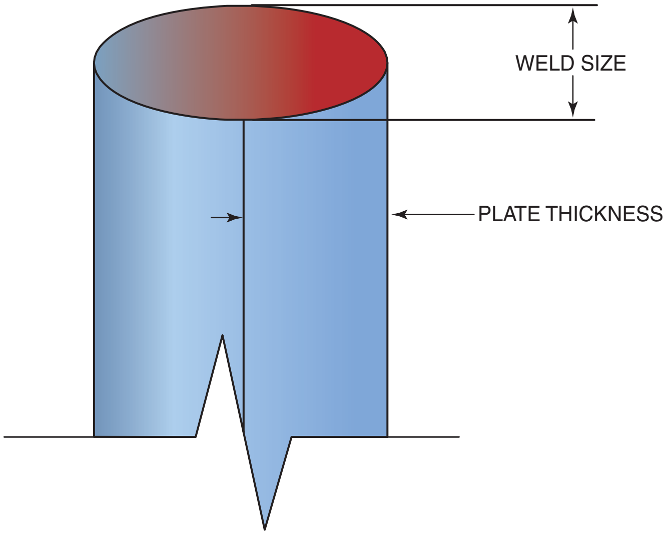

Figure 2.49
Edge weld size

PRACTICE 2-8

Edge Weld in the Flat Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

Using a properly set up and adjusted arc welding machine, proper safety protection as demonstrated in Practice 2-1, arc welding electrodes with a 1/8-in. (3-mm) diameter, and two pieces of mild steel plate, 6 in. (152 mm) long × 1/4 in. (6 mm) thick, you will make a weld on an edge joint, **Figure 2.50.**

- Clamp the plates flat together and make a tack weld along each end of the plates.
- Starting at one end of the plate, make a straight weld the full length of the plate. Make the weld bead as wide as the width of the edge joint.
- Watch the molten weld pool, not the end of the electrode.
- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects.
- Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

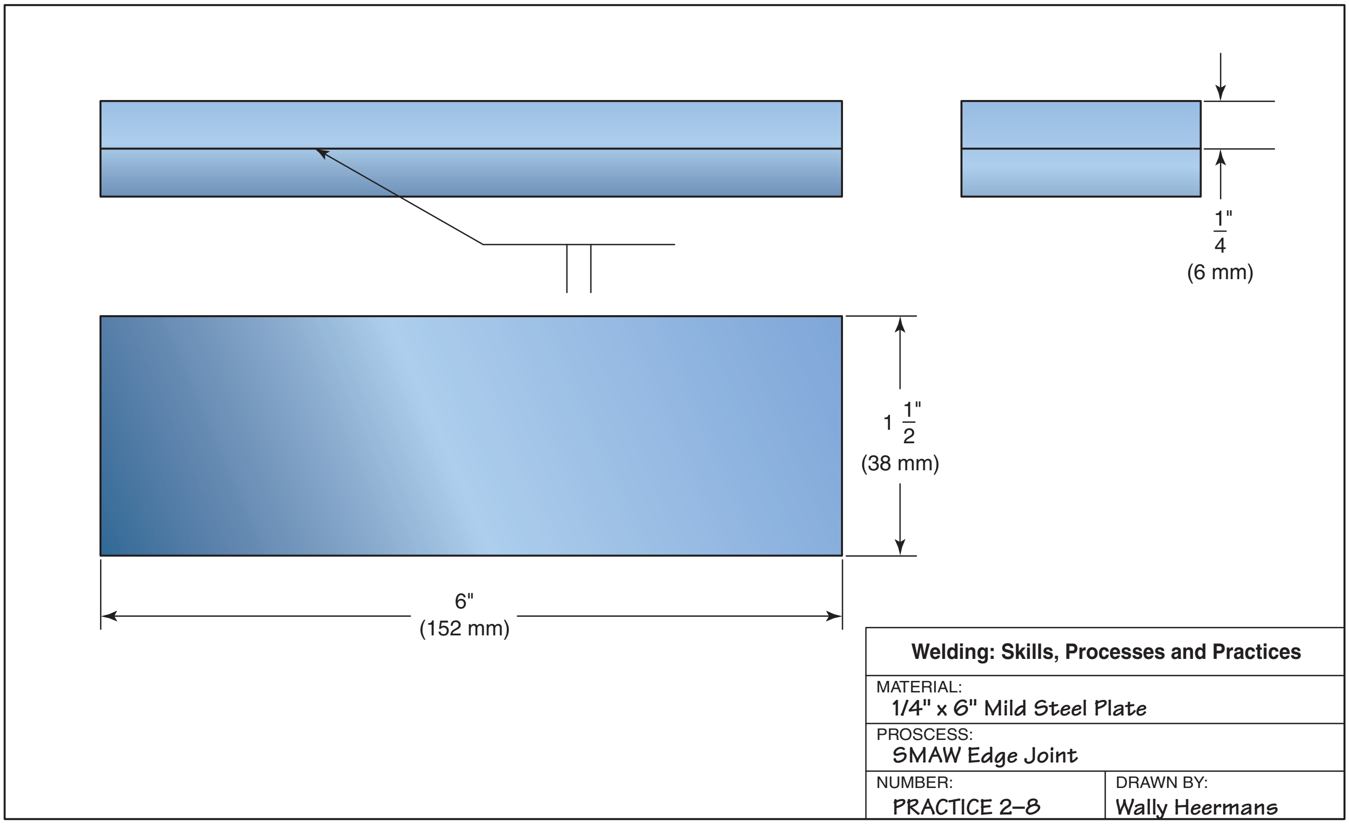
1"/4
(6 mm)
1 1"/2
(38 mm)
6"
(152 mm)
Welding: Skills, Processes and Practices
MATERIAL:
1/4" x 6" Mild Steel Plate
PROSCESS:
SMAW Edge Joint
NUMBER:
PRACTICE 2–8
DRAWN BY:
Wally Heermans

Figure 2.50
Practice 2.8 edge joint

PRACTICE 2-9

Edge Joint in the Vertical Down Position Using E6010 or E6011 Electrodes and E6012 or E6013 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-8, you will make a vertical down weld on an edge joint. Start with the plates at a 45° angle.

This technique is the same as that used to make vertical down welds. However, a lower level of skill is required at 45°, and it is easier to develop your skill. After you master the 45° angle, the angle is increased successively until a vertical position is reached, **Figure 2.51.**

- Make the weld bead as wide as the joint. Controlling a weld bead this size is more difficult, but you must develop the skill required to control this larger molten weld pool.
- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

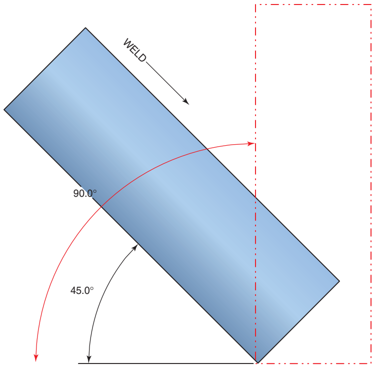

Figure 2.51
Vertical down

PRACTICE 2-10

Edge Joint in the Vertical Up Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-8, you will make a vertical up weld on an edge joint. Start with the plates at a 45° angle.

This technique is the same as that used to make vertical up welds. However, a lower level of skill is required at 45°, and it is easier to develop your skill. After you master the 45° angle, the angle is increased successively until a vertical position is reached, **Figure 2.52.**

Before the molten metal drips down the bead, the back of the molten weld pool will start to bulge, **Figure 2.53.** When this happens, increase the speed of travel and the weave pattern.

- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-11

Edge Joint in the Horizontal Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-8, you will make a horizontal weld on an edge joint. When you begin

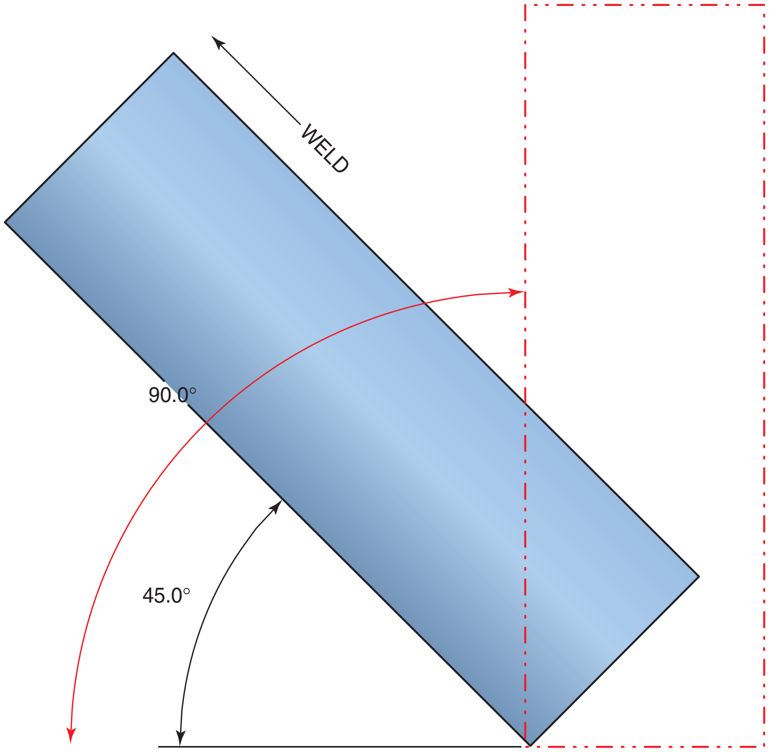

Figure 2.52
Vertical up

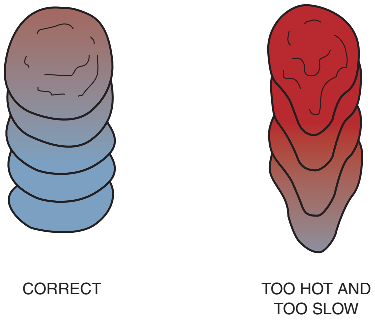

Figure 2.53
Watch the trailing edge of the weld pool to judge the correct travel speed

to practice the horizontal weld, the plate may be reclined slightly, **Figure 2.54.** This placement allows the welder to build the required skill by practicing the correct techniques successfully. The "J" weave or stepped pattern is suggested for this practice. As the electrode is drawn back to the back edge of the weld pool, metal is deposited. Use the metal that's being deposited to support the molten weld pool.

Angling the electrode up and back toward the weld causes more metal to be deposited along the top edge of the weld. Keeping the bead small allows the surface tension to hold the molten weld pool in place.

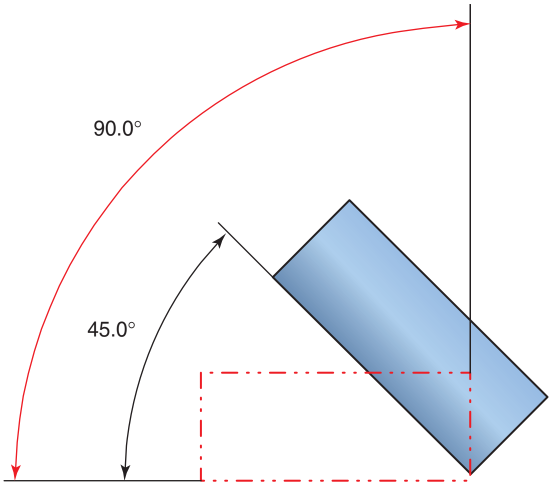

Figure 2.54
Incline angle

Gradually increase the angle of the plate until it and the weld bead are horizontal.

- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-12

Edge Joint in the Overhead Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 1, 2

Using the same setup, materials, and electrodes as listed in Practice 2-8, you will make an overhead weld on an edge joint.

- With the electrode pointed in a slightly trailing angle, **Figure 2.55,** strike the arc in the joint.
- Keep a very short arc length.
- Use the stepped pattern and move the electrode forward slightly when the molten weld pool grows to the correct size, **Figure 2.56.**

As the molten weld pool gets larger, it has a tendency to quickly become convex. If you keep the arc in the molten weld pool once the joint is filled and the weld face is flat, it will quickly overfill and become convex. This can result in the weld face forming drips of metal that hang from the weld like icicles, **Figure 2.57.**

- When the molten weld pool cools and begins to shrink, move the arc back near the center of the weld.

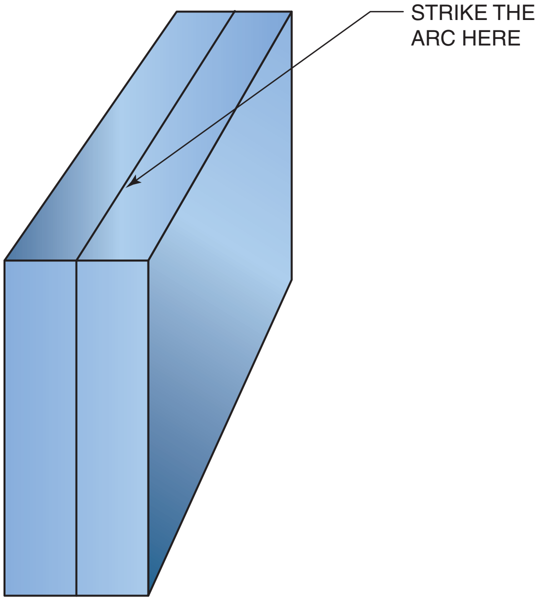

Figure 2.55
Strike the arc in the joint

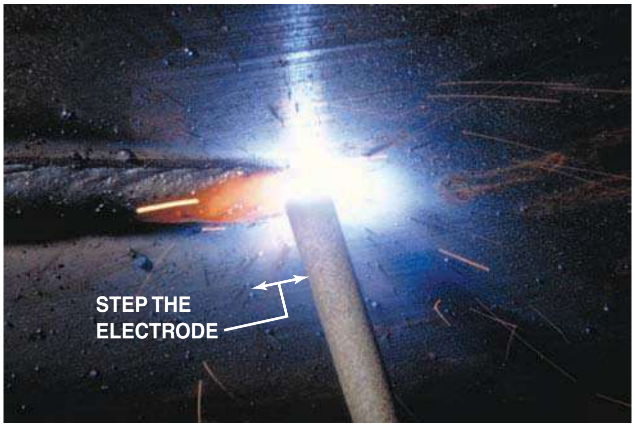

Figure 2.56
Step the electrode
Courtesy of Larry Jeffus

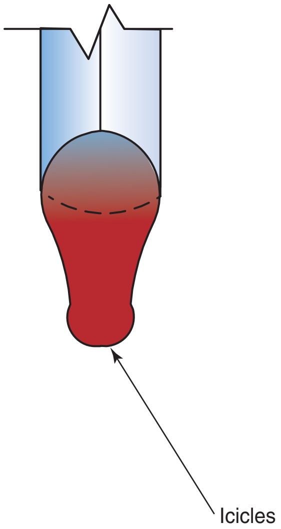

Figure 2.57
Welding too slow or with too high of an amperage setting will result in the weld metal dripping down like icicles

- Hold the arc in this new location until the molten weld pool again grows to the correct size.
- Step the electrode forward again and keep repeating this pattern until the weld progresses along the entire weld joint length.
- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

OUTSIDE CORNER JOINT

An outside corner joint is made by placing the plates at a 90° angle to each other, with the edges forming a V-groove, **Figure 2.58.** There may or may not be a slight root opening left between the plate edges. Small tack welds should be made approximately 1/2 in. (13 mm) from both ends of the joint.

The weld bead should completely fill the V-groove formed by the plates and may have a slightly convex surface buildup. The back side of an outside corner joint can be used to practice fillet welds, or four plates can be made into a box tube shape, **Figure 2.59.**

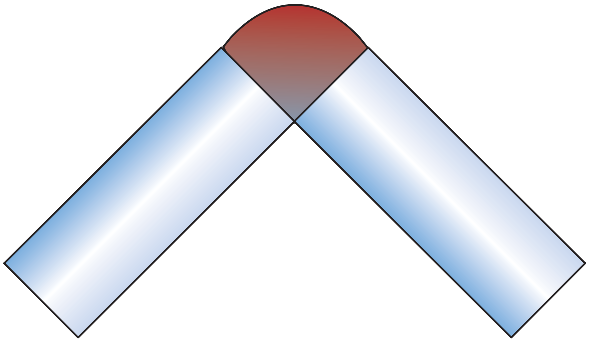

Figure 2.58
V formed by an outside corner joint

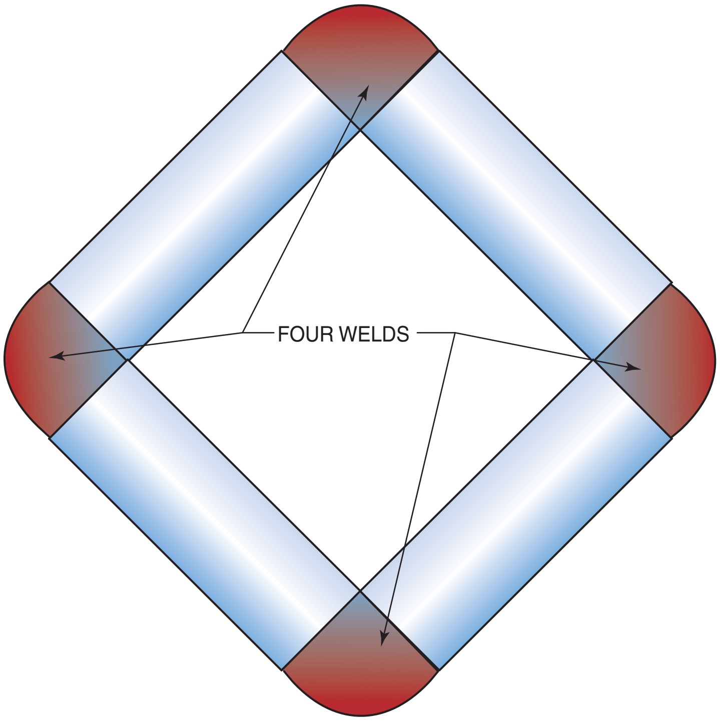

Figure 2.59
Box tube made from four outside corner joint welds

PRACTICE 2-13

Outside Corner Joint in the Flat Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using a properly set up and adjusted arc welding machine, proper safety protection as demonstrated in Practice 2-1, arc welding electrodes with a 1/8-in. (3-mm) diameter, and two pieces of mild steel plate, 6 in. (152 mm) long × 1/4 in. (6 mm) thick, you will make a weld on an outside corner joint.

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 1, 2

- Starting at one end of the plate, make a straight weld the full length of the plate.
- Watch the molten weld pool at this point, not the end of the electrode. As you become more skillful, it is easier to watch the molten weld pool.
- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-14

Outside Corner Joint in the Vertical Down Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes.

Using the same setup, materials, and electrodes as listed in Practice 2-13, you will make a vertical down weld on an outside corner joint. Start with the plate at a 45° angle.

This technique is the same as that used to make vertical down welds. However, a lower level of skill is required at 45°, and it is easier to develop your skill. After you master the 45° angle, the angle is increased successively until a vertical position is reached, **Figure 2.60.**

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 1, 2

- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

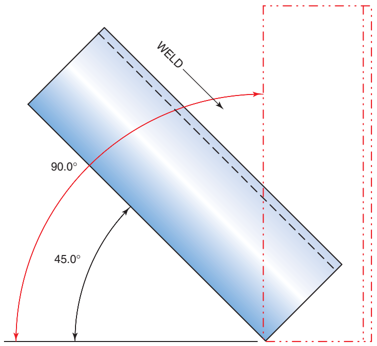

Figure 2.60
Start a 45° angle with and increase it to 90°

PRACTICE 2-15

Outside Corner Joint in the Vertical Up Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using the same setup, materials, and electrodes as listed in Practice 2-13, you will make a vertical up weld on an outside corner joint. Start with the plate at a 45° angle.

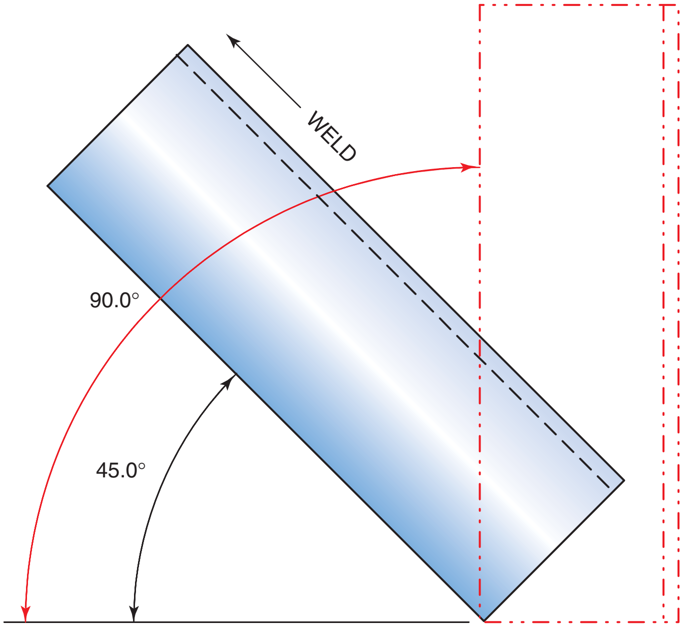

Figure 2.61
Vertical up

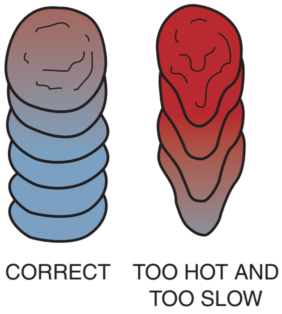
CORRECT
TOO HOT AND TOO SLOW

Figure 2.62
Watch the trailing edge of the weld pool to judge the correct travel speed

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

This technique is the same as that used to make vertical up welds. However, a lower level of skill is required at 45°, and it is easier to develop your skill. After the welder masters the 45° angle, the angle is increased successively until a vertical position is reached, **Figure 2.61.**

Before the molten metal drips down the bead, the back of the molten weld pool will start to bulge, **Figure 2.62.** When this happens, increase the speed of travel and the weave pattern.

- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 2-16

Outside Corner Joint in the Horizontal Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-13, you will make a horizontal weld on an outside corner joint. When the welder begins to practice the horizontal weld, the joint may be reclined slightly, **Figure 2.63.** This placement allows the welder to build the required skill by practicing the correct techniques successfully. The “J” weave or stepped pattern is suggested for this practice. As the electrode is drawn back into the weld pool, metal is deposited. This metal supports the molten weld pool, resulting in a bead with a uniform contour, **Figure 2.64.**

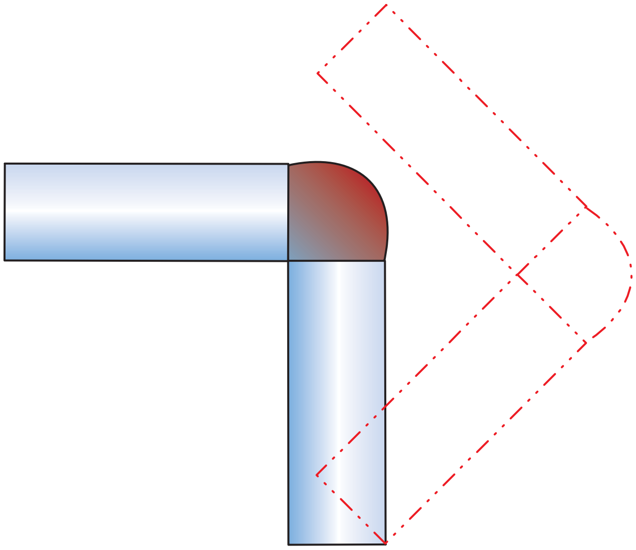

Figure 2.63
Incline angle

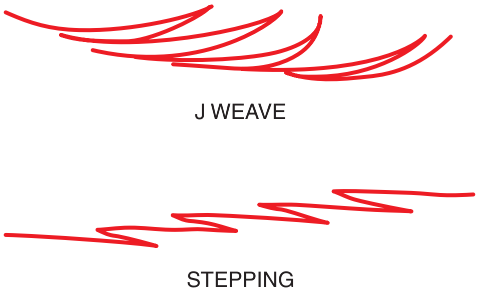

Figure 2.64
"J" weave or stepping

Angling the electrode up and back toward the weld causes more metal to be deposited along the top edge of the weld. Keeping the bead small allows the surface tension to hold the molten weld pool in place.

Gradually increase the angle of the plate until it is vertical and the weld bead is horizontal.

- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-17

Outside Corner Joint in the Overhead Position Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using the same setup, materials, and electrodes as listed in Practice 2-13, you will make an overhead weld on an outside corner joint.

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 3, 4

Module 9
Key Indicator 2

- With the electrode pointed slightly into the joint, **Figure 2.65,** strike the arc in the joint.
- Keep a very short arc length.
- Use the stepped pattern and move the electrode forward slightly when the molten weld pool grows to the correct size, **Figure 2.66.**

As the molten weld pool gets larger, it has a tendency to quickly become convex. If you keep the arc in the molten weld pool once the joint is filled and the weld face is flat, it will quickly overfill and become convex. This can result in the weld face forming drips of metal that hang from the weld like icicles, **Figure 2.67.**

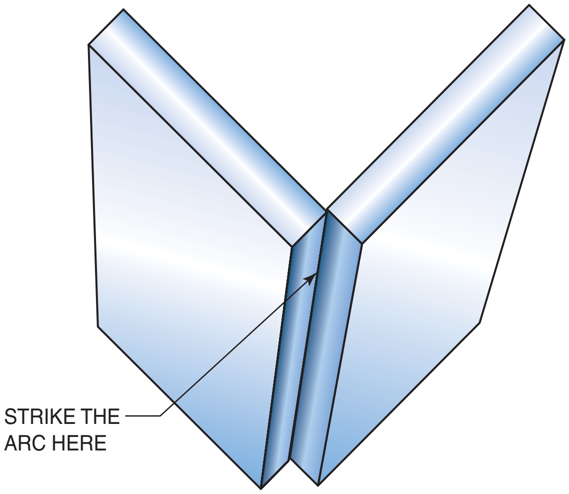

Figure 2.65
Strike arc in the joint

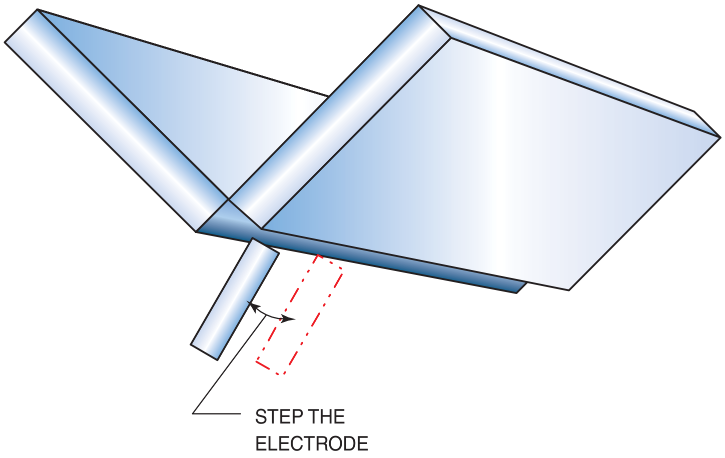

Figure 2.66
Stepping the electrode to control weld size

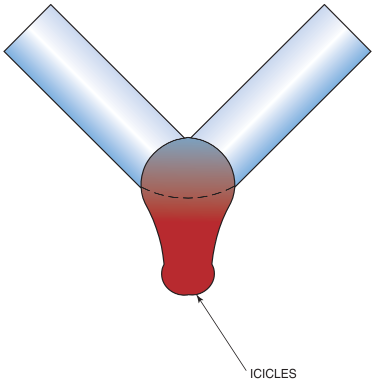

Figure 2.67
Welding too slowly or with too high of an amperage setting will result in the weld metal dripping down like icicles

- When the molten weld pool cools and begins to shrink, move the arc back near the center of the weld.
- Hold the arc in this new location until the molten weld pool again grows to the correct size.
- Step the electrode forward again and keep repeating this pattern until the weld progresses along the entire weld joint length.
- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

LAP JOINT

A **lap joint** is made by overlapping the edges of the two plates, **Figure 2.68.** The joint can be welded on one side or both sides with a fillet weld. In Practice 2-7, both sides should be welded unless otherwise noted.

As the fillet weld is made on the lap joint, the buildup should equal the thickness of the plate, **Figure 2.69.** A good weld will have a smooth transition from the plate surface to the weld. If this transition is abrupt, it can cause stresses that will weaken the joint.

Penetration for lap joints does not improve their strength; complete fusion is required. The root of fillet welds must be melted to ensure a completely fused joint. If the molten weld pool shows a notch during the weld, **Figure 2.70,** this is an indication that the root is not being fused together. The weave pattern will help prevent this problem, **Figure 2.71.**

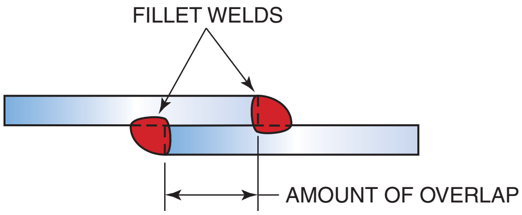

Figure 2.68
Lap joint

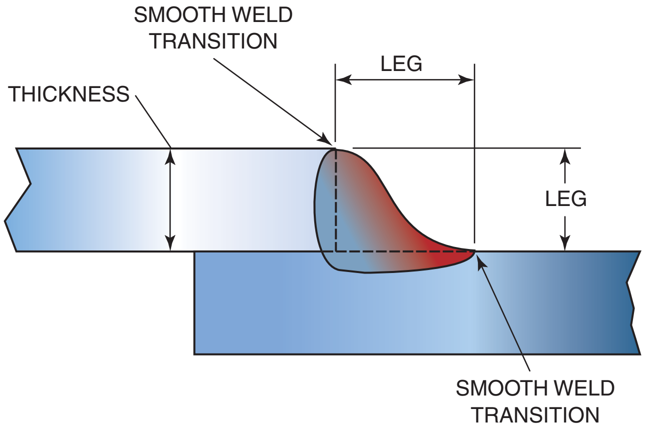

Figure 2.69
The legs of a fillet weld generally should be equal to the thickness of the base metal

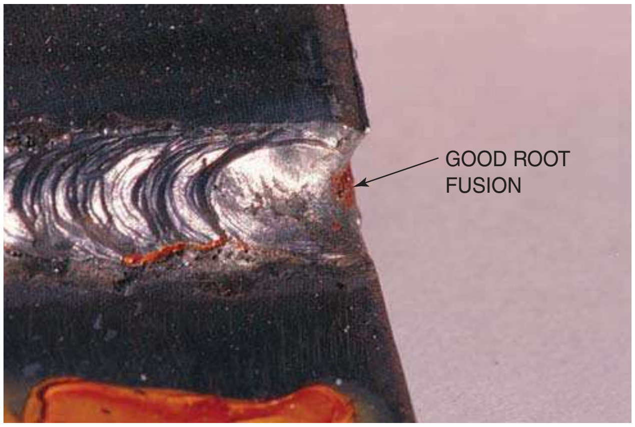

Figure 2.70
Watch the root of the weld bead to be sure there is a complete fusion
Courtesy of Larry Jeffus

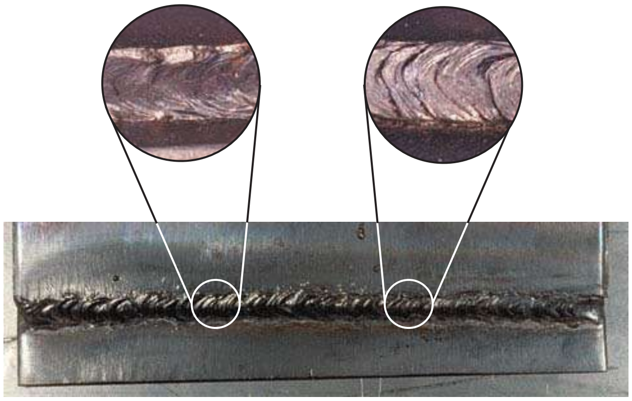

Figure 2.71
Lap joint
Courtesy of Larry Jeffus

PRACTICE 2-18

Welded Lap Joint in the Flat Position (1F) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using a properly set up and adjusted arc welding machine, proper safety protection, arc welding electrodes having a 1/8-in. (3-mm) diameter, and two or more pieces of mild steel plate, 6 in. (152 mm) long × 1/4 in. (6 mm) thick, you will make a welded lap joint in the flat position, **Figure 2.72.**

Hold the plates together tightly with an overlap of no more than 1/4 in. (6 mm). Tack weld the plates together. A small tack weld may be added in the center to prevent distortion during welding, **Figure 2.73.** Chip the tacks before you start to weld.

The "J," "C," or zigzag weave pattern works well on this joint. Strike the arc and establish a molten pool directly in the joint. Move the electrode out on the bottom plate and then onto the weld to the top edge of the top plate, **Figure 2.74.** Follow the surface of the plates with the arc. Do not follow the trailing edge of the weld bead. Following the molten weld pool will not allow for good root fusion and will cause slag to collect in the root. If

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 5

Module 9
Key Indicator 1, 2
This practice addresses the "Flat" position portion of the all-position requirement of 5.

Module 9
Key Indicator 2

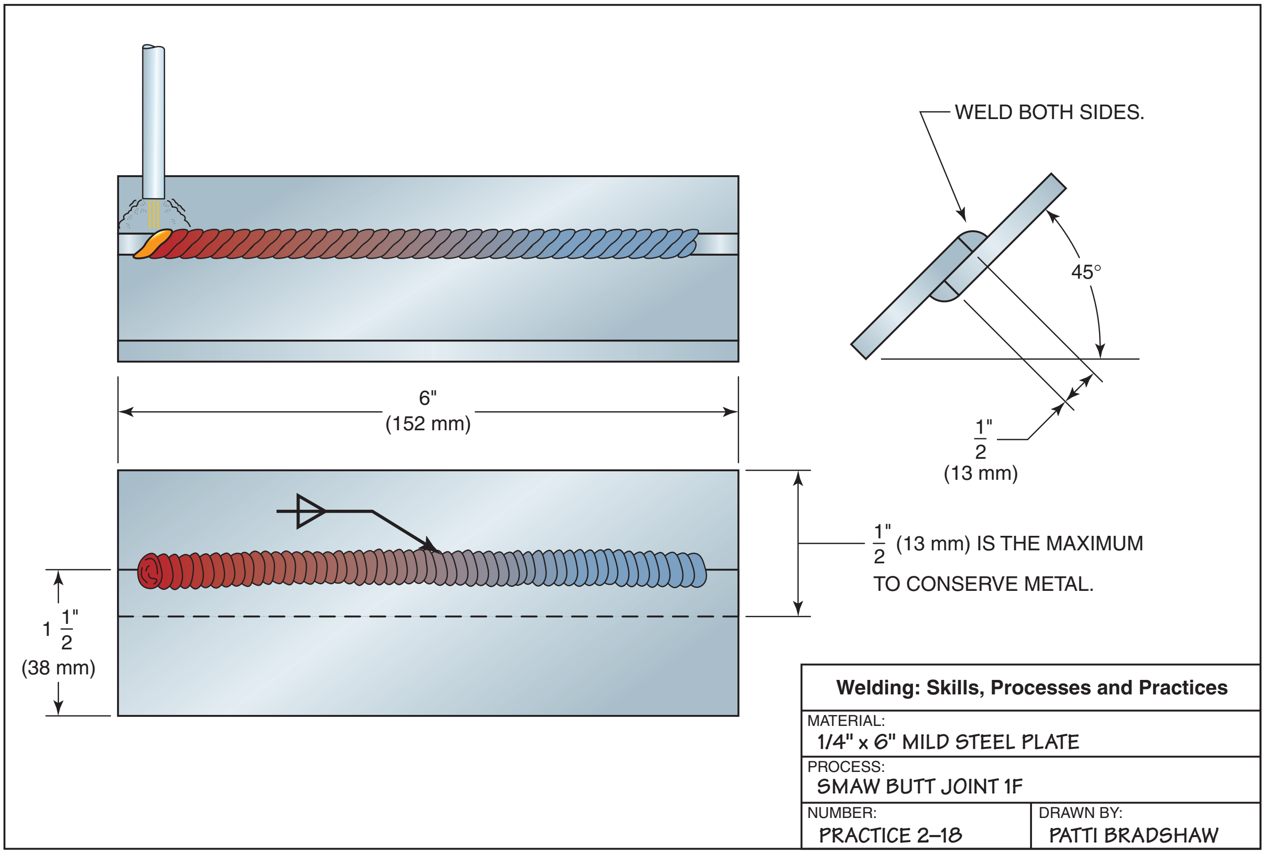
WELD BOTH SIDES.
45°
6" (152 mm)
1/2" (13 mm)
1/2" (13 mm) IS THE MAXIMUM TO CONSERVE METAL.
1 1/2" (38 mm)
Welding: Skills, Processes and Practices
MATERIAL:
1/4" x 6" MILD STEEL PLATE
PROCESS:
SMAW BUTT JOINT 1F
NUMBER:
PRACTICE 2–18
DRAWN BY:
PATTI BRADSHAW

Figure 2.72
Lap joint in the flat position

slag does collect, a good weld is not possible. Stop the weld and chip the slag to remove it before the weld is completed. Cool, chip, and inspect the weld for uniformity and defects. Repeat the welds with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

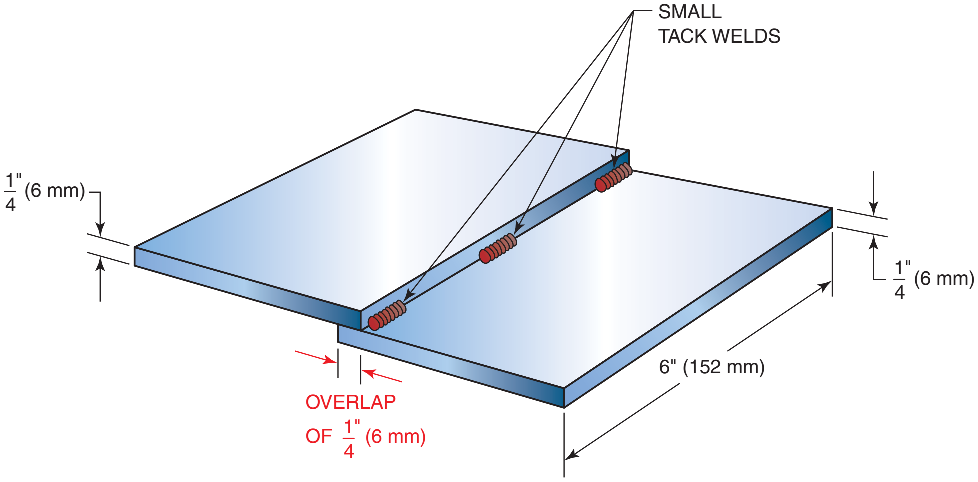
SMALL TACK WELDS
1/4" (6 mm)
1/4" (6 mm)
6" (152 mm)
OVERLAP OF 1/4" (6 mm)

Figure 2.73
Tack welding the plates together

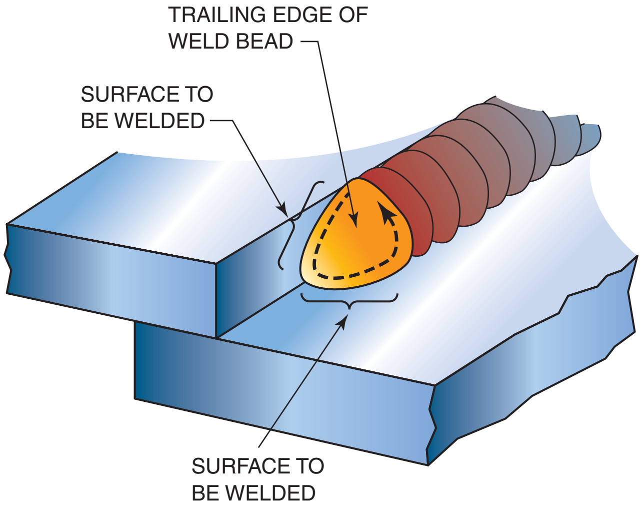

Figure 2.74
Follow the surface of the plate to ensure good fusion

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-19

Welded Lap Joint in the Horizontal Position (2F) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using the same setup, materials, and electrodes as listed in Practice 2-18, you will make a welded horizontal lap joint.

The horizontal lap joint and the flat lap joint require nearly the same technique and skill to achieve a proper weld, **Figure 2.75.** Use the "J," "C," or zigzag weave pattern to make the weld. Do not allow slag to collect in the root. The fillet must be equally divided between both plates for good strength. After completing the weld, cool, chip, and inspect the weld for uniformity and defects. Repeat the welds using all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 5

Module 9
Key Indicator 1, 2
This practice addresses the "Horizontal" position portion of the all-position requirement of 5.

Module 9
Key Indicator 2

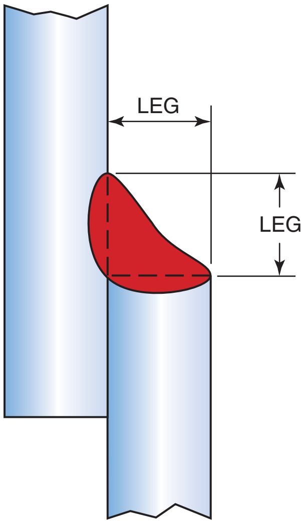

Figure 2.75
The horizontal lap joint should have a fillet weld that is equal on both plates

PROACTICE 2-20

Lap Joint in the Vertical Up Position (3F) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using the same setup, materials, and electrodes as listed in Practice 2-18, you will make a vertical up weld in a lap joint.

- Start practicing this weld with the plate at a 45° angle.
- Gradually increase the angle of the plate to vertical as skill is gained in welding this joint. The "J" or "T" weave pattern works well on this joint.
- Establish a molten weld pool in the root of the joint.

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 5
This practice addresses the "Vertical" position portion of the all-position requirement of 5.

Module 9
Key Indicator 2

- Use the "T" pattern to step ahead of the molten weld pool, allowing it to cool slightly. Do not deposit metal ahead of the molten weld pool.
- As the molten weld pool size starts to decrease, move the electrode back down into the molten weld pool.
- Quickly move the electrode from side to side in the molten weld pool, filling up the joint.
- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as necessary with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-21

Lap Joint in the Overhead Position (4F) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 5
This practice addresses the "Overhead" position portion of the all-position requirement of 5.

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-18, you will make an overhead weld in a lap joint.

- With the electrode pointed slightly into the joint, **Figure 2.76,** strike the arc in the inside corner of the lap joint.
- Keep a very short arc length.
- Use the stepped pattern and move the electrode forward slightly when the molten weld pool grows to the correct size, **Figure 2.77.**

As the molten weld pool gets larger, it has a tendency to quickly become convex. If you keep the arc in the molten weld pool once the joint is filled and the weld face is flat, it will quickly overfill and become convex. This can result in the weld face forming drips of metal that hang from the weld like icicles, **Figure 2.78.**

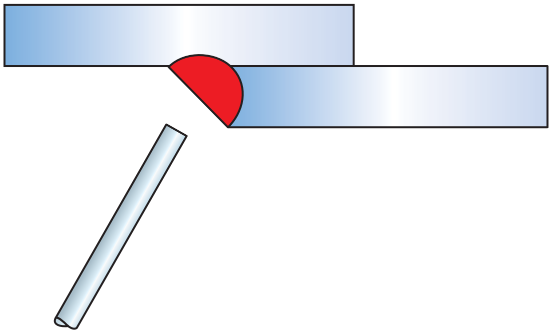

Figure 2.76
Point the electrode slightly toward the root of the joint

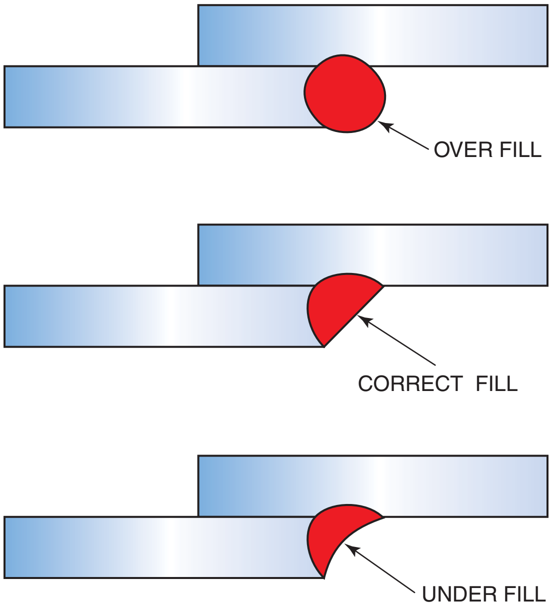

Figure 2.77
Correct fillet weld size for overhead welds

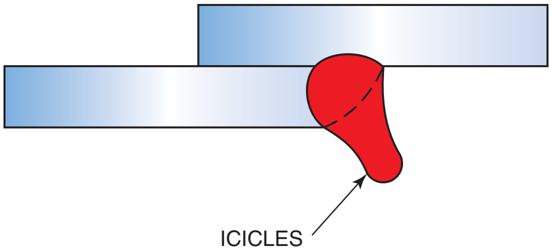

Figure 2.78
Overfilling the molten weld pool will result in drips of metal called icicles

- When the molten weld pool cools and begins to shrink, move the arc back near the center of the weld.
- Hold the arc in this new location until the molten weld pool again grows to the correct size.
- Step the electrode forward again and keep repeating this pattern until the weld progresses along the entire weld joint length.
- Cool, chip, and inspect the weld for uniformity and defects.
- Repeat the welds as needed with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

TEE JOINT

The **tee joint** is made by tack welding one piece of metal on another piece of metal at a right angle, **Figure 2.79.** After the joint is tack welded together, the slag is chipped from the tack welds. If the slag is not removed, it will cause a slag inclusion in the final weld.

The heat is not distributed uniformly between both plates during a tee weld. Because the plate that forms the stem of the tee can conduct heat away from the arc in only one direction, it will heat up faster than the base plate. Heat escapes into the base plate in two directions. When using a weave pattern, most of the heat should be directed to the base plate to keep the weld size more uniform and to help prevent undercut.

A welded tee joint can be strong if it is welded on both sides, even without having deep penetration, **Figure 2.80.** The weld will be as strong as the

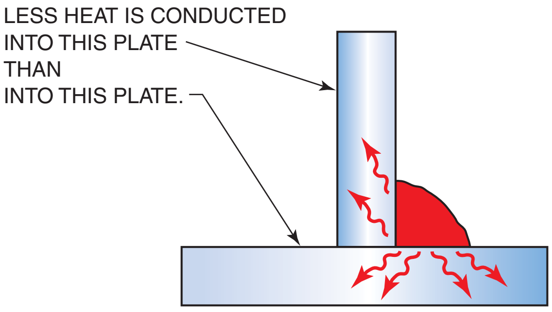

Figure 2.79
Tee joint

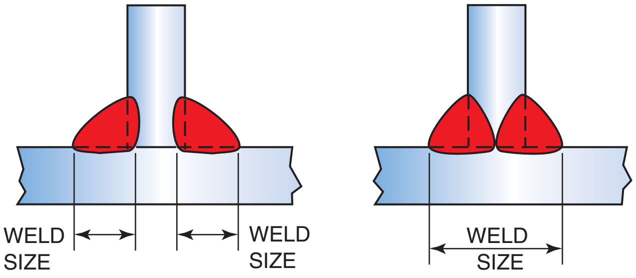

Figure 2.80
If the total weld sizes are equal, then both tee joints would have equal strength

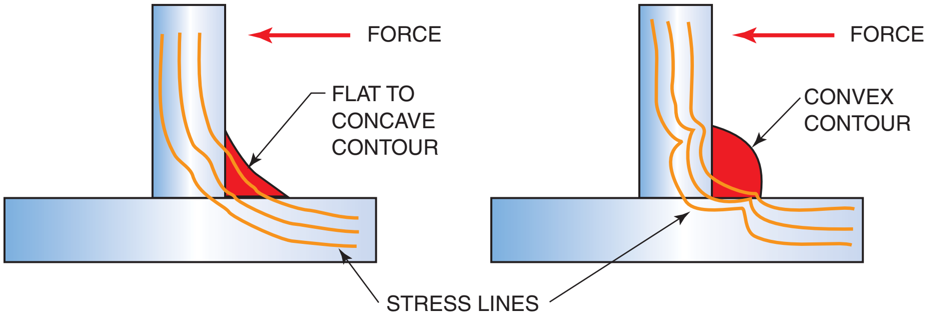

Figure 2.81
The stresses are distributed more uniformly through a flat or concave fillet weld

base plate if the size of the two welds equals the total thickness of the base plate. The weld bead should have a flat or slightly concave appearance to ensure the greatest strength and efficiency, **Figure 2.81.**

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 5
This practice addresses the "Flat" position portion of the all-position requirement of 5.

Module 9
Key Indicator 2

PRACTICE 2-22

Tee Joint in the Flat Position (1F) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Using a properly set up and adjusted arc welding machine, proper safety protection, arc welding electrodes having a 1/8-in. (3-mm) diameter, and two or more pieces of mild steel plate, 6 in. (152 mm) long × 1/4 in. (6 mm) thick, you will make a welded tee joint in the flat position, **Figure 2.82.**

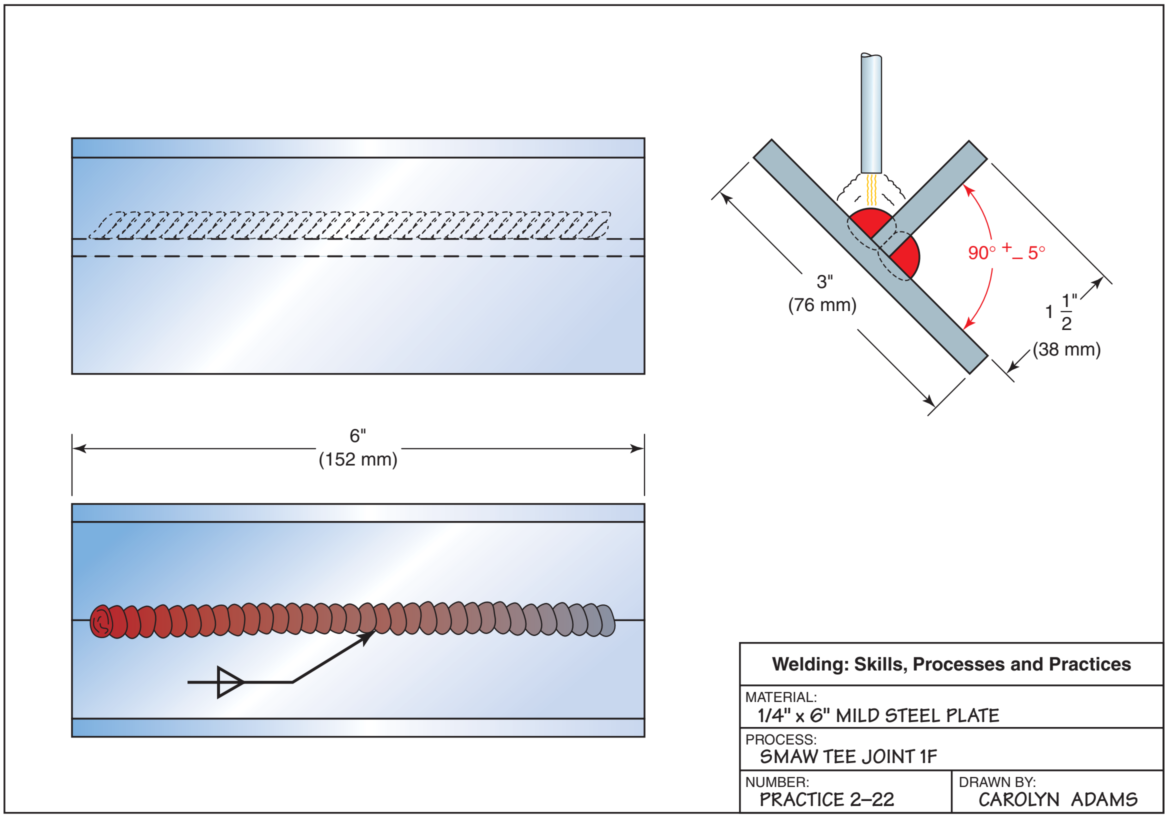

Welding: Skills, Processes and Practices
MATERIAL: 1/4" x 6" MILD STEEL PLATE
PROCESS: SMAW TEE JOINT 1F
NUMBER: PRACTICE 2–22
DRAWN BY: CAROLYN ADAMS

Figure 2.82
Tee joint in the flat position

After the plates are tack welded together, place them on the welding table so the weld will be flat. Start at one end and establish a molten weld pool on both plates. Allow the molten weld pool to flow together before starting the bead. Any of the weave patterns will work well on this joint. To prevent slag inclusions, use a slightly higher than normal amperage setting.

When the 6-in. (152-mm)-long weld is completed, cool, chip, and inspect it for uniformity and soundness. Repeat the welds as needed for all these groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-23

Tee Joint in the Horizontal Position (2F) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 5
This practice addresses the "Horizontal" position portion of the all-position requirement of 5.

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-22, you will make a welded tee joint in the horizontal position.

Place the tack welded tee plates flat on the welding table so that the weld is horizontal and the plates are flat and vertical, **Figure 2.83.** Start the arc on the flat plate and establish a molten weld pool in the root on both plates. Using the "J" or "C" weave pattern, push the arc into the root and slightly up the vertical plate. You must keep the root of the joint fusing together with the weld metal. If the metal does not fuse, a notch will appear on the leading edge of the weld bead. Poor or incomplete root fusion will cause the weld to be weak and easily cracked under a load.

When the weld is completed, cool, chip, and inspect it for uniformity and defects. Undercut on the vertical plate is the most common defect. Repeat the welds with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

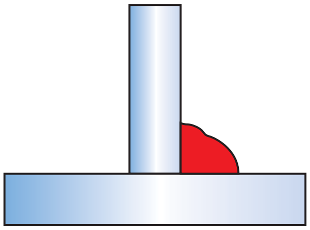

Figure 2.83
Horizontal tee

PRACTICE 2-24

Tee Joint in the Vertical Up Position (3F) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 5
This practice addresses the "Vertical" position portion of the all-position requirement of 5.

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-22, you will make a welded tee joint in the vertical up position.

Practice this weld with the plate at a 45° angle. This position will allow you to develop your skill for the vertical up position. Start the arc and molten weld pool deep in the root of the joint. Build a shelf large enough to support the bead as it progresses up the joint. The square, "J," or "C" pattern can be used, but the "T" or stepped pattern will allow deeper root penetration.

For this weld, undercut is a problem on both sides of the weld. It can be controlled by holding the arc on the side long enough for filler metal

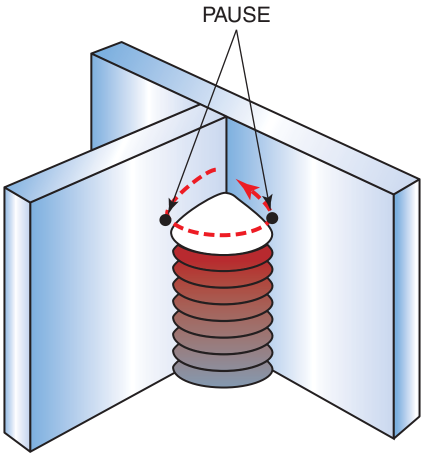

Figure 2.84
Pausing just above the undercut will fill it. This action also causes undercut, but that will be filled on the next cycle.

to flow down and fill it, **Figure 2.84.** Cool, chip, and inspect the weld for uniformity and defects. Repeat the welds as necessary with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 2-25

Tee Joint in the Overhead Position (4F) Using E6010 or E6011 Electrodes, E6012 or E6013 Electrodes, and E7016 or E7018 Electrodes

Module 1
Key Indicator 1, 2, 3, 4

Module 4
Key Indicator 1, 5
This practice addresses the "Overhead" position portion of the all-position requirement of 5.

Module 9
Key Indicator 2

Using the same setup, materials, and electrodes as listed in Practice 2-12, you will make a welded tee joint in the overhead position.

Start the arc and molten weld pool deep in the root of the joint. Keep a very short arc length. The stepped pattern will allow deeper root penetration.

For this weld, undercut is a problem on both sides of the weld with a high buildup in the center. It can be controlled by holding the arc on the side long enough for filler metal to flow in and fill it. Cool, chip, and inspect the weld for uniformity and defects. Repeat the welds as necessary with all three (F) groups of electrodes until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

SUMMARY

The shielded metal arc welding process is most often referred to in welding shops as stick welding. Some people say that it gets this name for one of two reasons. The first is most obviously a result of the stick shape of the electrode. The second reason is experienced by all new welders; it is the tendency for the electrode to stick to the workpiece. All new welders experience this, and your ability to control the sticking of the electrode can be improved as you develop the proper arc-striking techniques.

For a new welder, it is often difficult to concentrate on anything other than the bright sparks and glow at the end of the electrode. But, with time, as you develop your skills, your visual field will increase, allowing you to see a much larger welding zone. This skill comes with time and practice. Developing this skill is essential for you to become a highly proficient welder. Nothing enhances your welding skills more than time under the hood, actually welding, cleaning off the weld, inspecting it, determining the necessary corrections to be made, and immediately trying to produce the next weld with a higher level of quality.

REVIEW

1. Describe two methods of striking an arc with an electrode.
2. Why is it important to strike the arc only in the weld joint?
3. What problems may result by using an electrode at too low a current setting?
4. What problems may result by using an electrode at too high a current setting?
5. According to **Table 2.1,** what would the amperage range be for the following electrodes?
   a. 1/8 in. (3 mm), E6010
   b. 5/32 in. (4 mm), E7018
   c. 3/32 in. (2.4 mm), E7016
   d. 1/8 in. (3 mm), E6011
6. What makes some spatter “hard”?
7. Why should you never change the current setting during a weld?
8. What factors should be considered when selecting an electrode size?
9. What can a welder do to control overheating of the metal pieces being welded?
10. What effect does changing the arc length have on the weld?
11. What arc problems can occur in deep or narrow weld joints?
12. Describe the difference between using a leading and a trailing electrode angle.
13. Can all electrodes be used with a leading angle? Why or why not?
14. What characteristics of the weld bead does the weaving of the electrode cause?
15. What are some of the applications for the circular pattern in the flat position?
16. Using a pencil and paper, draw two complete lines of the weave patterns you are most comfortable making.
17. Why is it important to find a good welding position?
18. Which electrodes would be grouped in the following F-numbers: F2, F3, F4?
19. Give one advantage of using electrodes with cellulose-based fluxes.
20. What are stringer beads?
21. Describe an ideal tack weld.

22. What effect does the root opening or root gap have on a butt joint?
23. What can happen if the fillet weld on a lap joint does not have a smooth transition?
24. Which plate heats up faster on a tee joint? Why?
25. Can a tee weld be strong if the welds on both sides do not have deep penetration? Why or why not?

CHAPTER

3

Advanced Shielded Metal Arc Welding

OBJECTIVES

After completing this chapter, the student should be able to

- prepare metal before welding
- make the root pass, filler weld, and cover pass in all positions and techniques
- explain the purpose of a hot pass
- make a visual inspection, and describe the appearance of an acceptable weld
- make a root pass on plate in all positions
- make a root pass on plate with an open root in all positions
- make an open root weld on plate using the step technique in all positions
- make a multiple pass filler weld on a V-joint in all positions using E7018 electrodes
- make groove welds on a limited thickness plate with a backing strip using E6010 root and hot passes and E7018 fill passes in all positions
- make an SMAW workmanship sample and welder qualification test plate for limited thickness horizontal 2G and 3G positions with E7018 electrodes
- make a cover bead in all positions
- make limited thickness welder performance qualification test plate without backing in all positions
- make a single V-groove open root butt joint with an increasing root opening
- make a single V-groove open root butt joint with a decreasing root opening
- make SMAW welds of plate to plate

KEY TERMS

back gouging
burn-through
cover pass
filler pass
guided bend specimen
hot pass
interpass temperature
key hole
molten weld pool
multiple pass weld
postheating
preheating
root pass
wagon tracks
weld groove
weld specimen

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter

Module 1: Occupational Orientation

Key Indicator 1: Prepares time or job cards, reports or records
Key Indicator 2: Performs housekeeping duties
Key Indicator 4: Follows written instructions to complete work assignments

Module 4: Shielded Metal Arc Welding (SMAW)

Key Indicator 1: Performs safety inspections of SMAW equipment and accessories
Key Indicator 2: Makes minor external repairs to SMAW equipment and accessories
Key Indicator 3: Sets up SMAW operations on carbon steel
Key Indicator 4: Operates SMAW equipment on carbon steel
Key Indicator 5: Makes SMAW fillet welds, in all positions, on carbon steel
Key Indicator 6: Makes SMAW groove welds, in all positions, on carbon steel
Key Indicator 7: Passes SMAW welder performance qualification test (2G and 3G, uphill, limited thickness test plates) on carbon steel

Module 9: Welding Inspection and Testing

Key Indicator 1: Examines cut surfaces and edges of prepared base metal parts
Key Indicator 2: Examines tacks, root passes, intermediate layers, and completed welds

INTRODUCTION

The SMAW process can be used to consistently produce high-quality welds. Sometimes it is necessary to make welds in less than ideal conditions. Knowing how to produce a weld of high strength in an out-of-position, difficult situation or on an unusual metal takes both practice and knowledge. A welder is frequently required to make these types of welds to a code or standard. This chapter covers the high-quality welding of plate. The practices are designed to give you the experience of taking code-type tests in a variety of materials and positions as well as to develop good workmanship.

Any time a code-quality weld requiring 100% joint penetration is to be made on metal thicker than 1/4 in. (6 mm), the metal edges must be prepared before welding. A joint is prepared for welding by cutting a groove in the metal along the edge. The preparation is done to allow deeper penetration into the joint of the weld for improved strength. Prepared joints often require more than one weld pass to complete them. By preparing the joint, metal

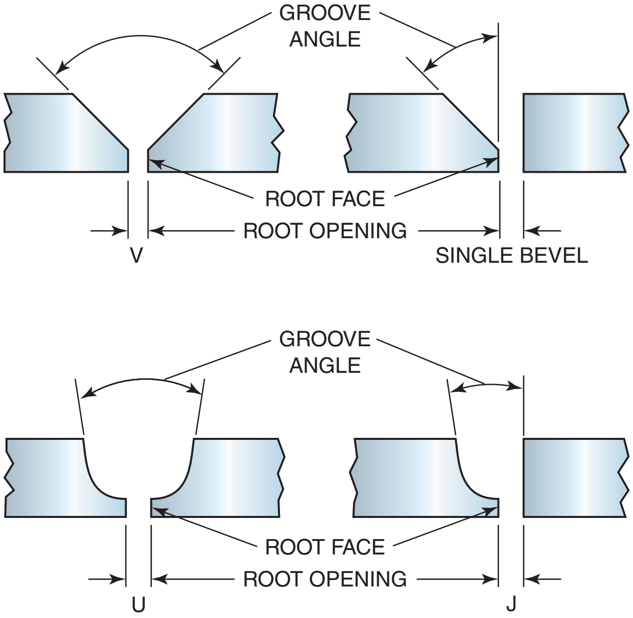

Figure 3.1
Standard grooves used to ensure satisfactory joint penetration

several feet thick can be welded with great success, **Figure 3.1.** The same welding techniques are used when making prepared welds of any thickness.

The root pass is used to fuse the parts together and seal off possible atmospheric contamination from the filler weld. Once the root pass is completed and cleaned, a hot pass may be used to improve the weld contour and burn out small spots of trapped slag. For high-quality welds, a grinder should be used on the root pass to clean it.

Filler welds and cover welds are often made with low hydrogen electrodes such as E7018. These passes are used to fill and cap the weld groove.

Welders are often qualified by passing a required qualification test using a groove weld. The type of joint, thickness of metal, type and size of electrode, and position are all specified by agencies issuing codes and standards. Except for the American Welding Society's Certified Welder program, taking a test according to one company or agency's specifications may not qualify a welder for another company or agency's testing procedures. But being able to pass one type of test will usually help the welder to pass other tests for the same type of joint, thickness of metal, type and size of electrode, and position. Information about the AWS Certified Welder program is available from the AWS's main office in Miami, Florida.

ROOT PASS

The **root pass** is the first weld bead of a **multiple pass weld.** The root pass fuses the two parts together and establishes the depth of weld metal penetration. A good root pass is needed in order to obtain a sound weld. The

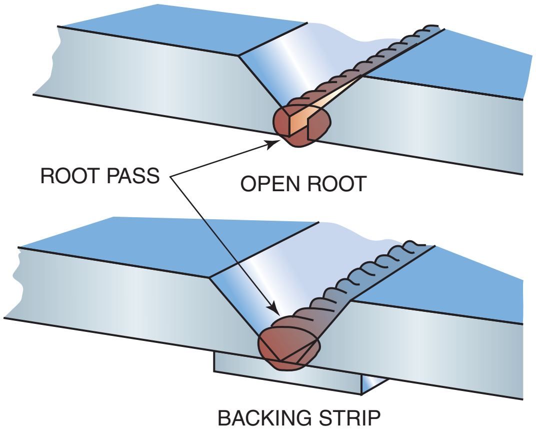

Figure 3.2
Root pass maximum deposit 1/4 in. (6 mm) thick

root may be either open or closed, using a backing strip or backing ring, **Figure 3.2.**

The backing strip used in a closed root may remain as part of the weld, or it may be removed. Because leaving the backing strip on a weld may cause it to fail due to concentrations of stresses along the backing strip, removable backup tapes have been developed. Backup tapes are made of high-temperature ceramics, **Figure 3.3,** that can be used to increase penetration and prevent **burn-through.** The tape can be peeled off after the weld is completed. Most welds do not use backing strips.

On plates that have the joints prepared on both sides, the root face may be ground or gouged clean before another pass is applied to both sides, **Figure 3.4.** This practice has been applied to some large diameter pipes. Welds, however, that can be reached from only one side must be produced adequately, without the benefit of being able to clean and repair the back side.

The open root weld is widely used in plate and pipe designs. The face side of an open root weld is not so important as the root surface on the back or inside, **Figure 3.5.** The face of a root weld may have some areas

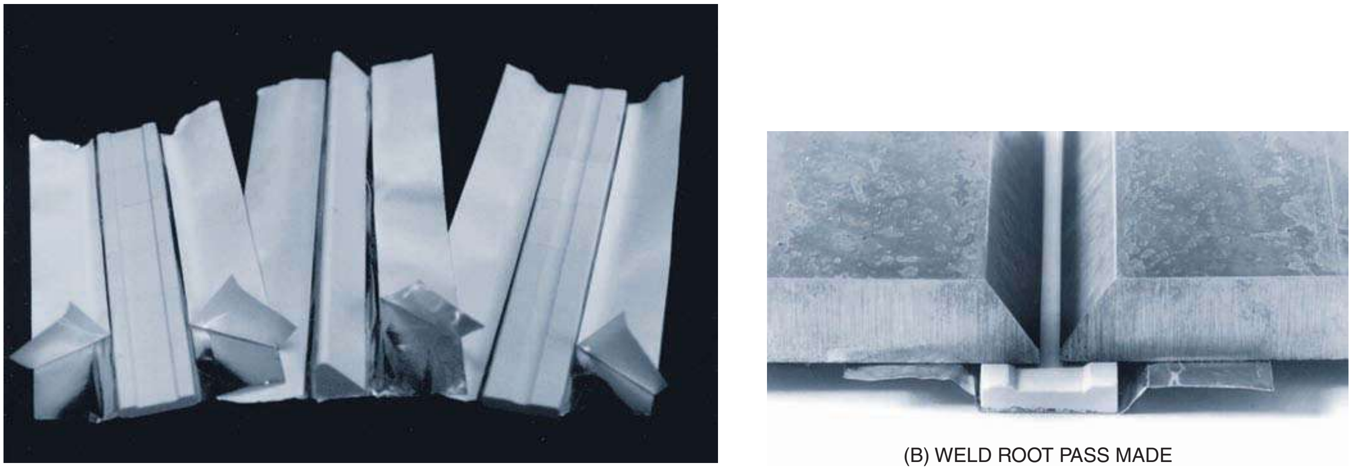
(A) FIBERGLASS

(B) WELD ROOT PASS MADE
USING CERAMIC BACKING TAPE

Figure 3.3
Welding backing tapes are available in different materials and shapes
Courtesy of ESAB Welding & Cutting Products

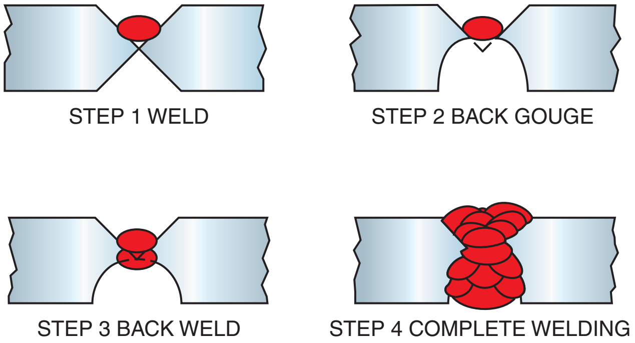
STEP 1 WELD
STEP 2 BACK GOUGE
STEP 3 BACK WELD
STEP 4 COMPLETE WELDING

Figure 3.4
Using back gouging to ensure a sound weld root

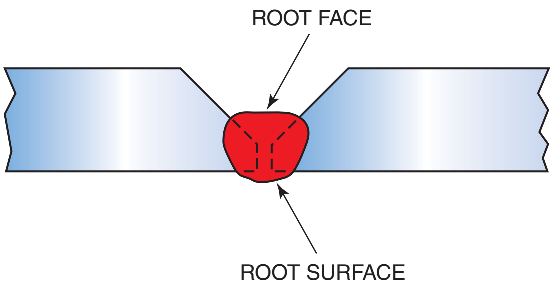
ROOT FACE
ROOT SURFACE

Figure 3.5
Ideal bead shape for the root pass

of poor uniformity in width, reinforcement, and buildup, or it may have other defects, such as undercut or overlap. As long as the root surface is correct, the front side can be ground, gouged, or burned out to produce a sound weld, **Figure 3.6.** For this reason, during the root pass practices, the weld will be evaluated from the root side only, as long as there are not too many defects on the face. To practice the open root welds, the welder will be using mild steel plate that is 1/8 in. (3 mm) thick. The root face for most grooved joints will be about the same size. This thin plate will help the welder build skill without taking too much time beveling the plate just to practice the root pass. Two different methods are used to make a root pass. One method is used only on joints with little or no root gap. This method requires a high amperage and short arc length. The arc length is so short that the electrode flux may drag along on the edges of the joint. The setup for this method must be correct in order for it to work.

The other method can be used on joints with wide, narrow, or varying root gaps. A stepping electrode manipulation and key hole control the penetration. The electrode is moved in and out of the **molten weld pool** as the weld progresses along the joint. The edge of the metal is burned back slightly by the electrode just ahead of the molten weld pool, **Figure 3.7.**

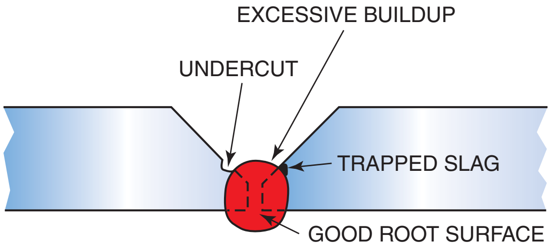
EXCESSIVE BUILDUP
UNDERCUT
TRAPPED SLAG
GOOD ROOT SURFACE

AS WELDED

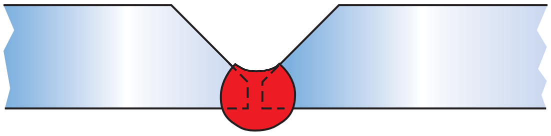

ROOT PASS–CLEANED UP;
READY FOR NEXT WELD PASS

Figure 3.6
Grinding back the root pass to ensure a sound second pass

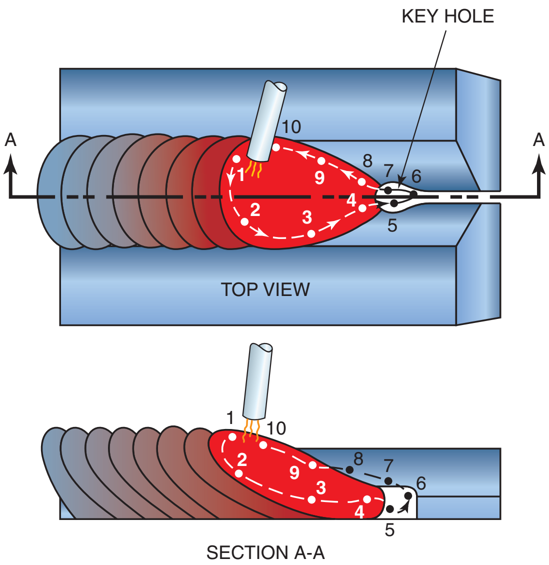
KEY HOLE
A
A
10
1
9
8
7
6
2
3
4
5
TOP VIEW
1
10
2
9
3
8
7
6
4
5
SECTION A-A

Figure 3.7
Electrode movement to open and use a keyhole

This is referred to as a **key hole,** and metal flows through the key hole to the root surface. The key hole must be maintained to ensure 100% penetration. This method requires more welder skill and can be used on a wide variety of joint conditions. The face of the bead resulting from this technique often is defect free.

PRACTICE 3-1

Root Pass on Plate with a Backing Strip in All Positions

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 6

Module 9
Key Indicator 1, 2

Using a properly set up and adjusted welding machine, proper safety protection, E6010 or E6011 arc welding electrodes having a 1/8-in. (3-mm) diameter, and one or more pieces of mild steel plate, 1/8 in. (3 mm) thick × 6 in. (152 mm) long, and one strip of mild steel, 1/8 in. (3 mm) thick × 1 in. (25 mm) wide × 6 in. (152 mm) long, you will make a root weld in all positions, **Figure 3.8.** Tack weld the plates together with a 1/16-in. (2-mm) to 1/8-in. (3-mm) root opening. Be sure there are no gaps between the backing strip and plates when the pieces are tacked together, **Figure 3.9.** If there is a small gap between the backing strip and the plates, it can be removed by placing the assembled test plates on an anvil and striking the tack weld with a hammer. This will close up the gap by compressing the tack welds, **Figure 3.10.**

Use a straight step or "T" pattern for this root weld. Push the electrode into the root opening so that there is good fusion with the backing strip and bottom edge of the plates. Failure to push the penetration deep into the joint will result in a cold lap at the root, **Figure 3.11.**

Watch the molten weld pool and keep its size as uniform as possible. As the molten weld pool increases in size, move the electrode out of the weld pool. When the weld pool begins to cool, bring the electrode back into the

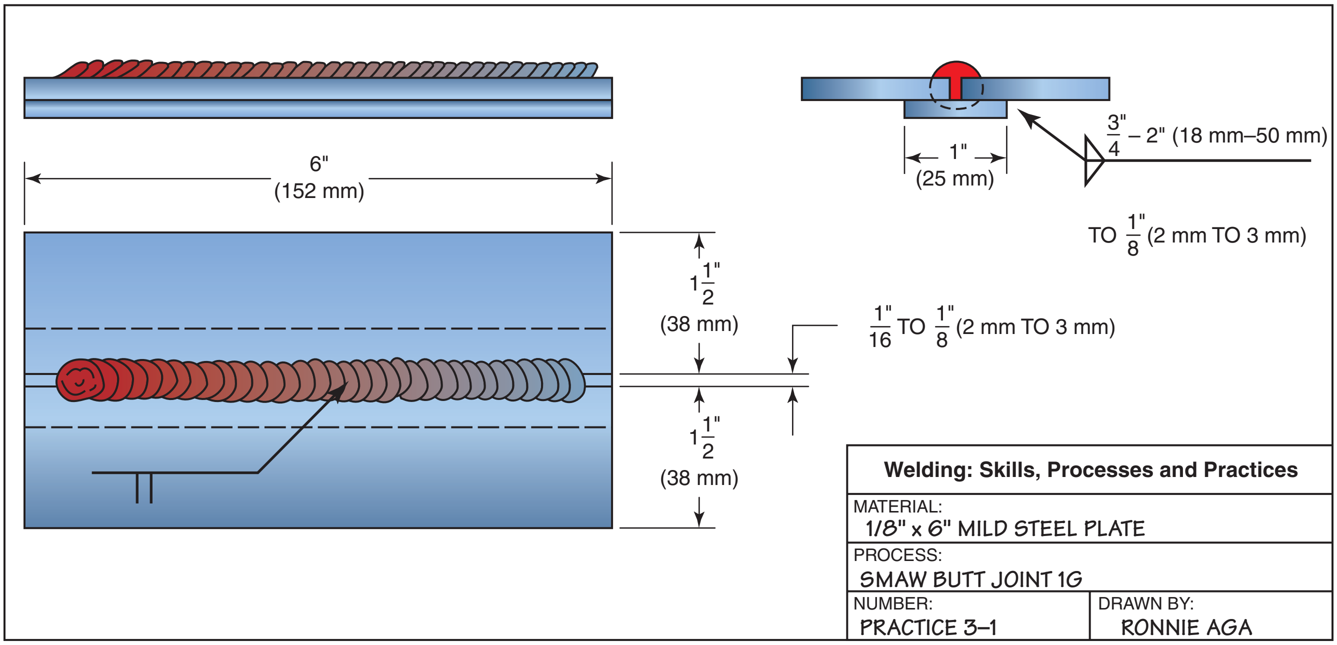
6" (152 mm)
1 1/2" (38 mm)
1/16" TO 1/8" (2 mm TO 3 mm)
1 1/2" (38 mm)
1" (25 mm)
3/4" – 2" (18 mm–50 mm)
TO 1/8" (2 mm TO 3 mm)
Welding: Skills, Processes and Practices
MATERIAL:
1/8" x 6" MILD STEEL PLATE
PROCESS:
SMAW BUTT JOINT 1G
NUMBER:
PRACTICE 3–1
DRAWN BY:
RONNIE AGA

Figure 3.8
Square butt joint with a backing strip

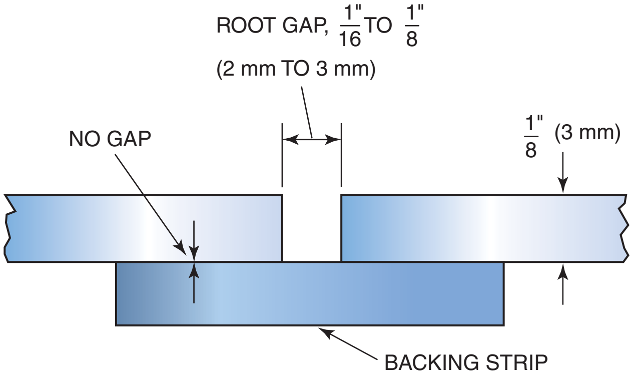

Figure 3.9
Backing strip

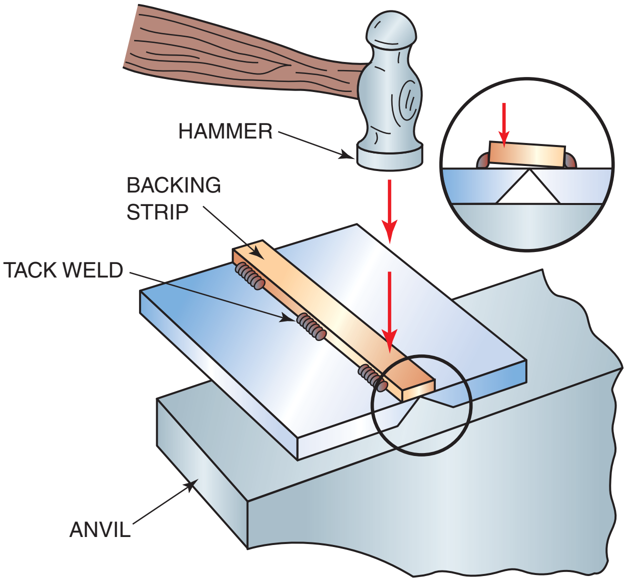

Figure 3.10
Using a hammer to align the backing strip and weld plates

molten weld pool. Use these weld pool indications to determine how far to move the electrode and when to return to the molten weld pool. After completing the weld, cut the plate and inspect the cross section of the weld for good fusion at the edges. Repeat the welds as necessary until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

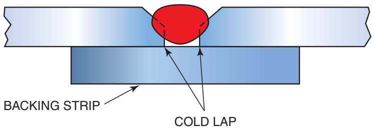

Figure 3.11
Incomplete root fusion

PRACTICE 3-2

Root Pass on Plate with an Open Root in All Positions

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 6

Module 9
Key Indicator 1, 2

Using a properly set up and adjusted arc welding machine, proper safety protection, E6010 or E6011 arc welding electrodes with a 1/8-in. (3-mm) diameter, and two or more pieces of mild steel plate, 6 in. (152 mm) long × 1/8 in (3 mm) or 3/16 in. (4.7 mm.) thick, you will make a welded butt joint in all positions with 100% root penetration.

- Tack weld the plates together with a root opening of 0 in. (0 mm) to 1/16 in. (2 mm).
- Using a short arc length and high amperage setting, make a weld along the joint.

You can change the electrode angle to control penetration and burn-through. As the trailing angle is decreased, making the electrode flatter to the plate, penetration, depth, and burn-through decrease, **Figure 3.12,** because both the arc force and heat are directed away from the bottom of the joint back toward the weld. Surface tension holds the metal in place, and the mass of the bead quickly cools the molten weld pool holding it in place. Increasing the electrode angle toward the perpendicular will increase penetration depth and possibly cause more burn-through. The arc

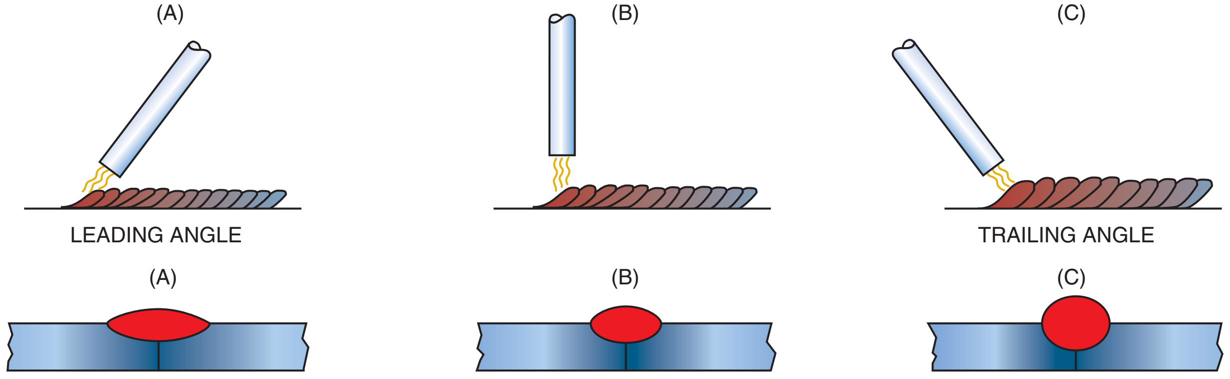

Figure 3.12
Effect of rod angle on weld bead shape

force and heat focused on the gap between the plates will push the molten metal through the joint.

The electrode holder can be slowly rocked from side to side while keeping the end of the electrode in the same spot on the joint, **Figure 3.13.** This will allow the arc force to better tie in the sides of the root to the base metal.

- When a burn-through occurs, rapidly move the electrode back to a point on the weld surface just before the burn-through.
- Lower the electrode angle and continue welding. If the burn-through does not close, stop the weld, chip, and wire brush the weld.
- Check the size of the burn-through. If it is larger than the diameter of the electrode, the root pass must be continued with the step method described in Practice 3-3. If the burn-through is not too large, lower the amperage slightly and continue welding.
- Watch the color of the slag behind the weld. If the weld metal is not fusing to one side, the slag will be brighter in color on one side. The

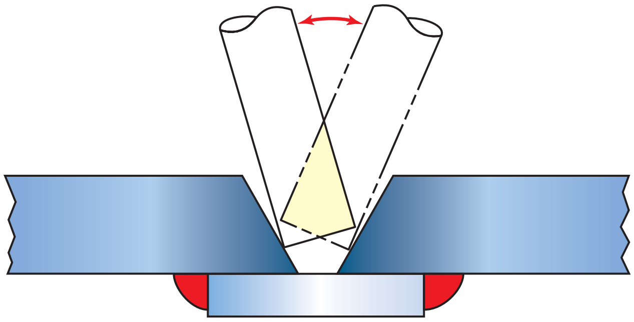

Figure 3.13
Rocking the top of the electrode while keeping the end in the same place helps control the bead shape

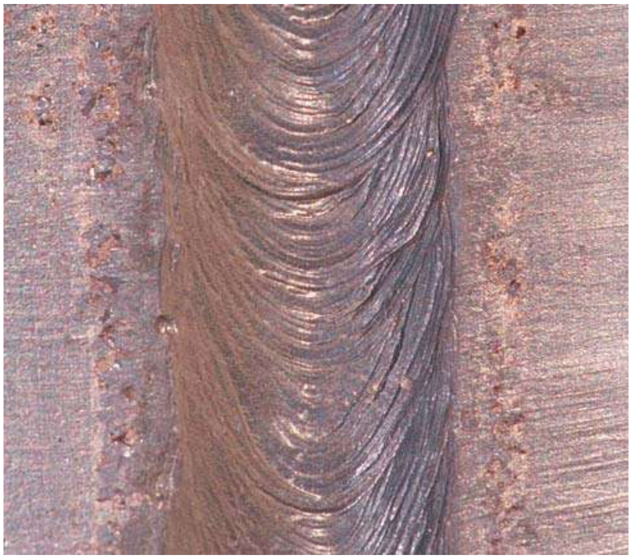

Figure 3.14
The weld toes appear uniform
Courtesy of Larry Jeffus

brighter color is caused by the slower cooling of the slag because there is less fused metal to conduct the heat away quickly.

- After the weld is completed, cooled, and chipped, check the back side of the plate for good root penetration. The root should have a small bead that will look as though it was welded from the back side, **Figure 3.14.** The penetration must be completely free of any drips of metal from the root face, called icicles.
- Repeat the welds as necessary until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 3-3

Open Root Weld on Plate Using the Step Technique in All Positions

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 6

Module 9
Key Indicator 1, 2

Using the same setup, materials, and electrodes as described in Practice 3-2, you will make a welded butt joint in all positions with 100% root penetration.

Tack weld the plates together with a root opening from 0 in. (0 mm) to 1/8 in. (3 mm). Using a medium amperage setting and a short stepping electrode motion, make a weld along the joint.

The electrode should be pushed deeply into the root to establish a key hole that will be used to ensure 100% root penetration. Once the key hole is established, the electrode is moved out and back in the molten weld pool at a steady, rhythmic rate. Watch the molten weld pool and key hole size to determine the rhythm and distance of electrode movement.

If the molten weld pool size decreases, the key hole will become smaller and may close completely. To increase the molten weld pool size and maintain the key hole, slow the rate of electrode movement and shorten the distance the electrode is moved away from the molten weld pool. This will increase the molten weld pool size and penetration because of increased localized heating.

If the molten weld pool becomes too large, metal may drip through the key hole, forming an icicle on the back side of the plate. Extremely large molten weld pool sizes can cause a large hole to be formed or cause burn-through. Repairing large holes can require much time and skill. To keep the molten weld pool from becoming too large, increase the travel speed, decrease the angle, shorten the arc length, or lower the amperage, **Table 3.1.**

The distance the electrode is moved from the molten weld pool and the length of time in the molten weld pool are found by watching the molten

Table 3.1 Changes Affecting Molten Weld Bead Size

	Amperage	Travel Speed	Electrode Size	Electrode Angle
To decrease puddle size	Decrease	Increase	Decrease	Leading
To increase puddle size	Increase	Decrease	Increase	Trailing

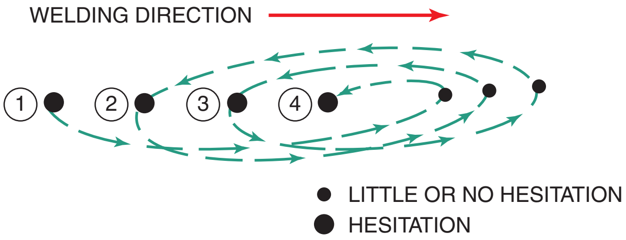

Figure 3.15
Weave pattern used to control the molten weld bead size

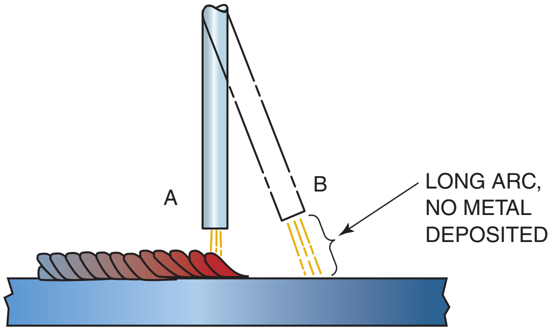

Figure 3.16
A long arc prevents metal or slag from being deposited ahead of the weld bead. A relatively quick movement from position "A" to position "B" and back to "A" is required.

weld pool. The molten weld pool size increases as you hold the arc in the molten weld pool until it reaches the desired size, about twice the electrode diameter, **Figure 3.15.** Move the electrode ahead of the molten weld pool, keeping the arc in the joint but being careful not to deposit any slag or metal ahead of the weld. To prevent metal and/or slag from transferring, raise the electrode to increase the arc length, **Figure 3.16.** Keep moving the electrode slowly forward as you watch the molten weld pool. The molten weld pool will suddenly start to solidify. At that time, move the electrode quickly back to the molten weld pool before it totally solidifies. Moving the electrode in a slight arc will raise the electrode ahead of the molten weld pool and automatically lower the electrode when it returns to the molten weld pool. Metal or slag deposited ahead of the molten weld pool may close the key hole, reduce penetration, and cause slag inclusions. Raising the end of the electrode too high or moving it too far ahead of the molten weld pool can cause all of the shielding gas to be blown away from the molten weld pool. If this happens, oxides can cause porosity. Keeping the electrode movement in balance takes concentration and practice.

Changing from one welding position to another requires an adjustment in timing, amperage, and electrode angle. The flat, horizontal, and overhead positions use about the same rhythm, but the vertical position may require a shorter time cycle for electrode movement. The amperage for the vertical position can be lower than that for the flat or horizontal, but the overhead position uses nearly the same amperage as flat and horizontal. The electrode angle for the flat and horizontal positions is about the same. For the vertical position, the electrode uses a sharper leading angle than does overhead, which is nearly perpendicular and may even be somewhat trailing.

Cool, chip, wire brush, and inspect both sides of the weld. The root surface should be slightly built up and look as though it was welded from that side (refer to Figure 3.14). Repeat the welds as necessary until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

HOT PASS

The surface of a root pass may be irregular, have undercut, overlap, slag inclusions, or other defects, depending upon the type of weld, the code or standards, and the condition of the root pass. The surface of a root pass can be cleaned by grinding or by using a **hot pass.**

On critical, high-strength code welds, it is usually required that the root pass as well as each filler pass be ground (refer to Figure 3.6). This grinding eliminates weld discontinuities caused by slag entrapments. It also can be used to remove most of the E60 series weld metal so that the stronger weld metal can make up most of the weld. When high-strength, low-alloy welding electrodes are used, this grinding is important to remove most of the low-strength weld deposit. This will leave the weld made up of nearly 100% of the high-strength weld metal, **Figure 3.17.**

The fastest way to clean out trapped slag and make the root pass more uniform is to use a hot pass. The hot pass uses a higher than normal amperage setting and a fast travel rate to reshape the bead and burn out trapped slag. After chipping and wire brushing the root pass to remove all the slag possible, a welder is ready to make the hot pass. The ideal way to apply a hot pass is to rapidly melt a large surface area, **Figure 3.18,** so that the trapped slag can float to the surface. The slag, mostly silicon dioxide (SiO_2), may not melt itself, so the surrounding steel must be melted to enable it to float free. The silicon dioxide may not melt because it melts at about 3100°F (1705°C), which is more than 500°F (270°C) hotter than the temperature at which the surrounding steel melts, around 2600°F (1440°C).

A very small amount of metal should be deposited during the hot pass so that the resulting weld is concave. A concave weld, compared to a

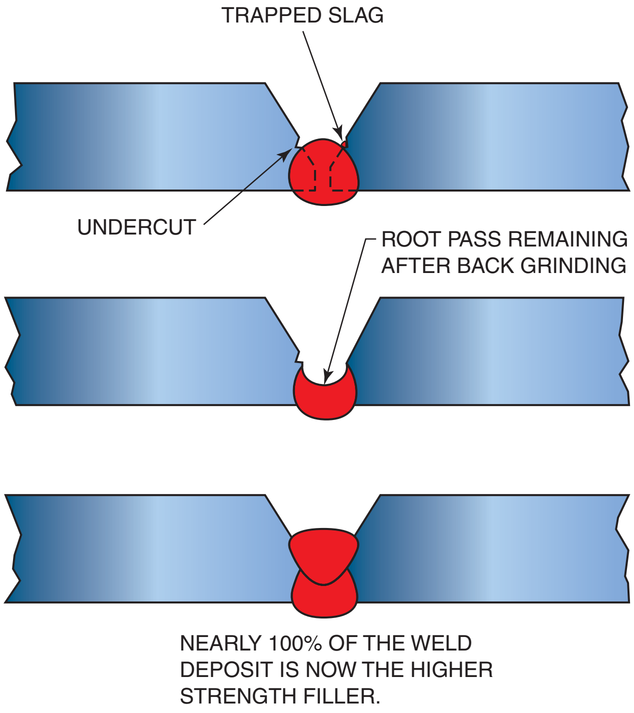
TRAPPED SLAG
UNDERCUT
ROOT PASS REMAINING AFTER BACK GRINDING
NEARLY 100% OF THE WELD DEPOSIT IS NOW THE HIGHER STRENGTH FILLER.

Figure 3.17
Back grinding to remove both discontinuities; filler metal used for the root pass

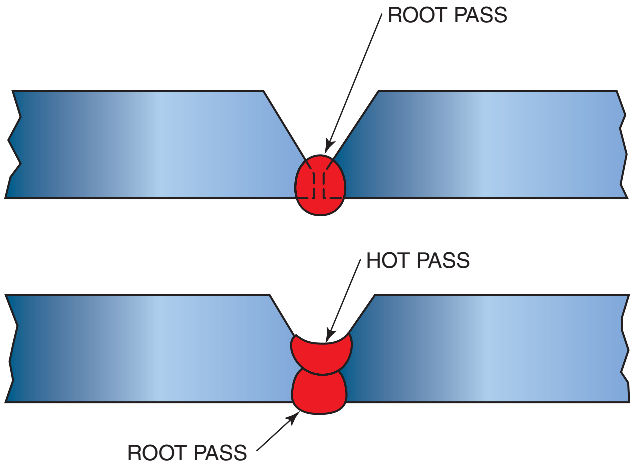

Figure 3.18
Using the hot pass to clean up the face of the root pass

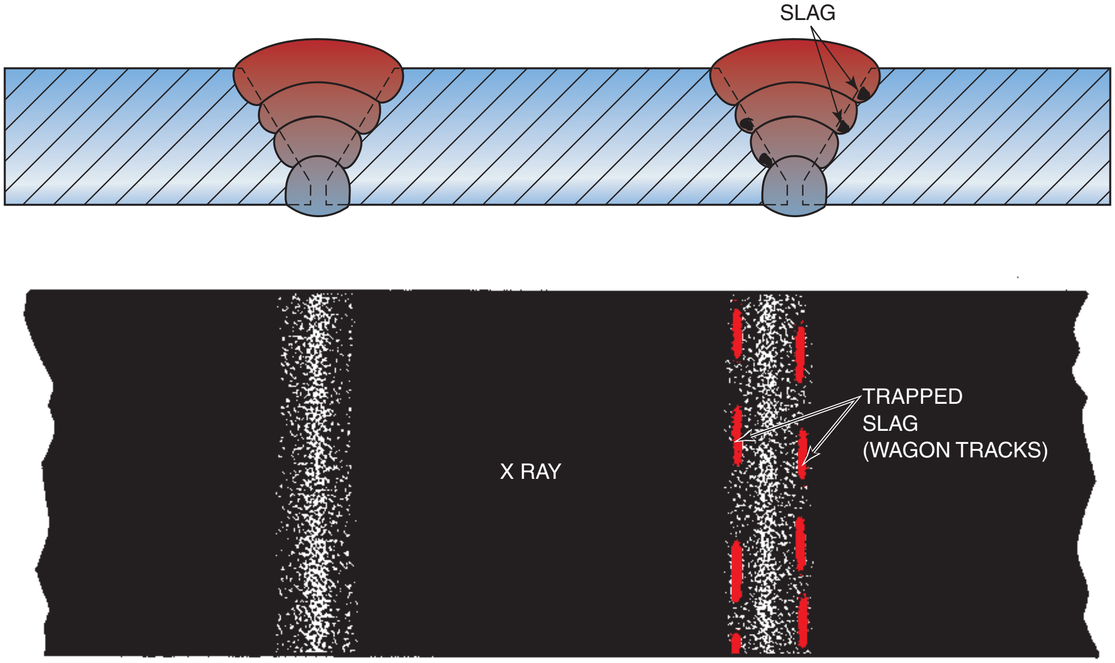

Figure 3.19
Slag trapped between passes will show on an X ray

convex weld, is more easily cleaned by chipping, wire brushing, or grinding. Failure to clean the convex root weld will result in a discontinuity showing up on an X ray. Such discontinuities are called **wagon tracks, Figure 3.19.**

The hot pass can also be used to repair or fill small spots of incomplete fusion or pinholes left in the root pass.

The normal weave pattern for a hot pass is the straight step or "T" pattern. The "T" can be used to wash out stubborn trapped slag better than the straight step pattern. The frequency of electrode movement is dependent upon the time required for the molten weld pool to start cooling. As with the root pass, metal or slag should not be deposited ahead of the bead. Do not allow the molten weld pool to cool completely or let the shielding gas covering to be blown away from the molten weld pool.

The hot pass technique can also be used to clean some welds that may first require grinding or gouging for a repair. The penetration of the molten weld pool must be deep enough to free all trapped slag and burn out all porosity.

EXPERIMENT 3-1

Hot Pass to Repair a Poor Weld Bead

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 4

Module 9
Key Indicator 1, 2

Using a properly set up and adjusted arc welding machine, proper safety protection, E6010 or E6011 arc welding electrodes having a 1/8-in. (3-mm) diameter, and two or more plates that have welds containing slag inclusions, lack of fusion, porosity, or other defects, you will make a hot pass to burn out the defects.

Chip and wire brush the weld bead. If necessary, use a punch to break apart large trapped slag deposits. The poorer the condition of the weld,

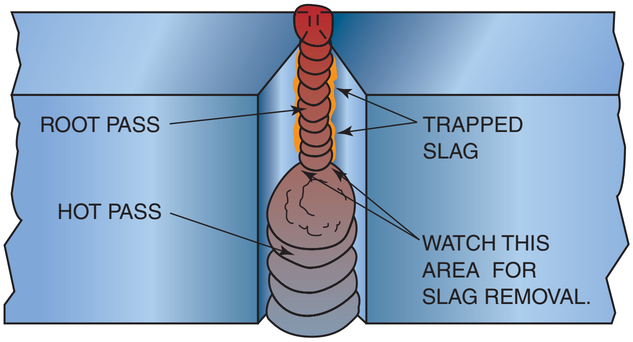

Figure 3.20
Burning out trapped slag by using a hot pass

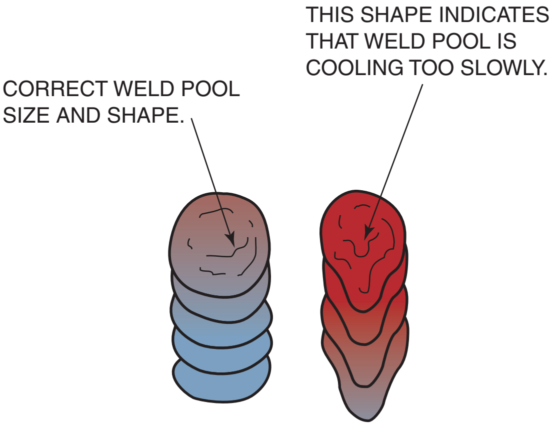
THIS SHAPE INDICATES THAT WELD POOL IS COOLING TOO SLOWLY.
CORRECT WELD POOL SIZE AND SHAPE.

Figure 3.21
The shape of the weld pool can indicate the temperature of the surrounding base metal

the more vertical the joint should be for the hot pass. Large slag deposits tend to float around the molten weld pool and stay trapped in deep pockets in the flat position. With the weld in the vertical position, the slag can run out of the joint and down the face of the weld. Set the amperage as high as possible without overheating and burning up the electrode. Start at the bottom and weld upward using a combination of straight-step and "T" patterns to keep the weld deposit uniform. Watch the back edge of the molten weld pool for size and the weld crater for the complete burning out of impurities, **Figure 3.20.**

The plate may start to become overheated because of the high heat input. If you notice that the weld bead is starting to cool too slowly and is growing in length, you should stop welding, **Figure 3.21.** Allow the plate to cool before continuing the weld.

After the weld is completed, cool, chip, and inspect it for uniformity. The plate can be cut at places where you know large discontinuities existed before to see if they were repaired or only covered up. If you wish, this experiment can be repeated on other defects and joints.

Welds that have large defects in addition to excessive buildup may require some grinding to remove the buildup. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

CAUTION

This hot pass technique is designed to be used on noncritical, noncode welds only. It should not be used to cover bad welds or as a means of repairing the work of a welder who is less skilled.

FILLER PASS

After the root pass is completed and it has been cleaned, the groove is filled with weld metal. These weld beads make up the **filler passes.** More than one pass is often required.

Filler passes are made with stringer beads or weave beads. For multiple pass welds, the weld beads must overlap along the edges. They should overlap enough so that the finished bead is smooth, **Figure 3.22.** Stringer beads usually overlap about 50%, and weave beads overlap approximately 25%.

Each weld bead must be cleaned before the next bead is started. Slag left on the plate between welds cannot be completely burned out because

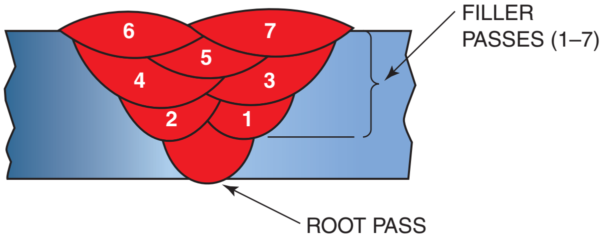

Figure 3.22
Filler passes—maximum thickness 1/8 in. (3 mm) each pass

filler welds should be made with a low amperage setting. Deep penetration will slow the rate of buildup in the joint. Deeply remelting the previous weld metal may weaken the joint. All that is required of a filler weld is that it be completely fused to the base metal.

Chipping, wire brushing, and grinding are the best ways to remove slag between filler weld passes. After the weld is completed, it can be checked by ultrasonic or radiographic nondestructive testing. Most schools are not equipped to do this testing. Therefore, a quick check for soundness can be made by destructive testing. One method of testing the deposited weld metal is by cutting and cross-sectioning the weld with an abrasive wheel and inspecting the weld. Another fast way to inspect filler passes is to cut a groove through the weld with a gouging tip. Watch the hot metal as it is washed away. The black spots that appear in the cut are slag inclusions. If only a few small spots appear, the weld probably will pass most tests. But, if a long string or large pieces of inclusions appear, the weld will most likely fail.

PRACTICE 3-4

Multiple Pass Filler Weld on a V-joint in All Positions

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 5

Module 9
Key Indicator 1, 2

Using a properly set up and adjusted arc welding machine, proper safety protection, E6010 or E6011 arc welding electrodes having a 1/8-in. (3-mm) diameter, and two or more pieces of mild steel plate, 6 in. (152 mm) long × 3 in. (76 mm) wide × 3/8 in. (10 mm) thick, you will make a multiple pass filler weld on a V-joint.

Tack weld the plates together at the corner so that they form a V, **Figure 3.23.** Starting at one end, make a stringer bead along the entire length using the straight step or "T" weave pattern. Thoroughly clean off the slag from the weld before making the next bead. **Figure 3.24** shows the suggested sequence for locating the beads. Continue making welds and cleaning them until the weld is 1 in. (25 mm) or more thick. Both ends of the weld may taper down. If it is important that the ends be square, metal tabs are welded on the ends of the plate for starting and stopping, **Figure 3.25.** The tabs are removed after the weld is completed.

After the weld is completed, it can be visually inspected for uniformity. If nondestructive testing is available, it may be checked for discontinuities. The weld also may be inspected by sectioning it with an abrasive wheel or gouging out with a torch. Repeat these welds until they are mastered. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

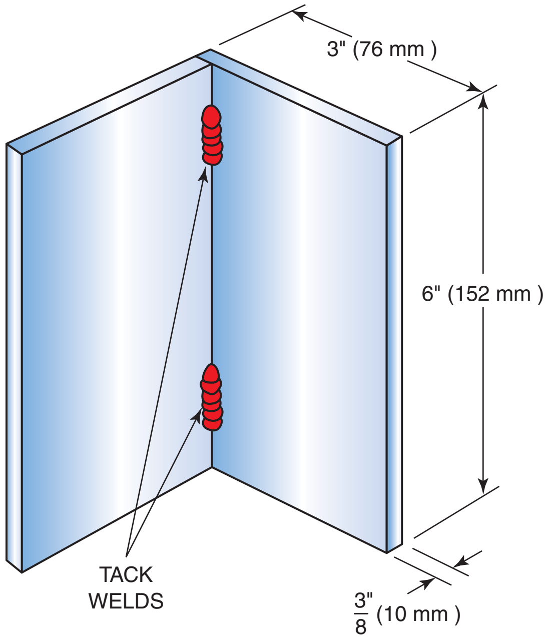

Figure 3.23
Tack weld the plates for the filler weld practice

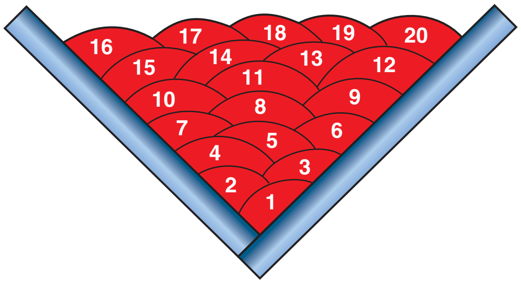

Figure 3.24
Filler pass buildup sequence

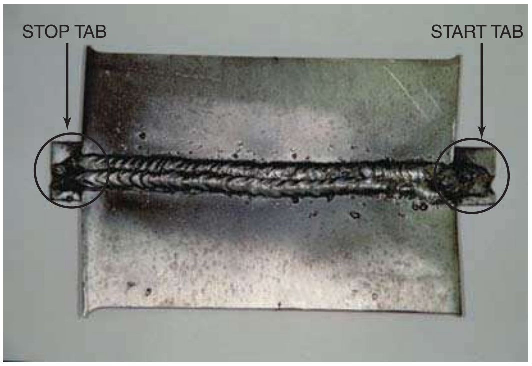

Figure 3.25
Run-off tabs help control possible under-fill or burn back at the starting and stopping points of a groove weld. The tabs can be removed.
Courtesy of Larry Jeffus

PRACTICE 3-5

Multiple Pass Filler Weld on a V-joint in All Positions Using E7018 Electrodes

Using the same setup and materials as in Practice 3-4, and E7018 arc welding electrodes in place of E6010 or E6011 electrodes, you will make a multiple pass filler weld on a V-joint.

- Tack weld the two plates together at the corners so that they form a V.
- Using a slow, straightforward motion, with little or no stepping, "T" or inverted "V" motion, make a stringer bead along the root of the joint.
- Chip the slag and repeat the weld until there is a buildup of 1 in. (25 mm) or more.

If during the weld the buildup should become uneven or large slag entrapments occur, they should be ground out. In industry, groove welds in plate 1 in. (25 mm) or more are normally repaired. Making these repairs now is good experience for the welder. All welders will at some time make a weld that may need repairing.

- After the weld is completed, visually test and nondestructive test the weld for external and internal discontinuities.
- Repeat the welds as necessary until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 5

Module 9
Key Indicator 1, 2

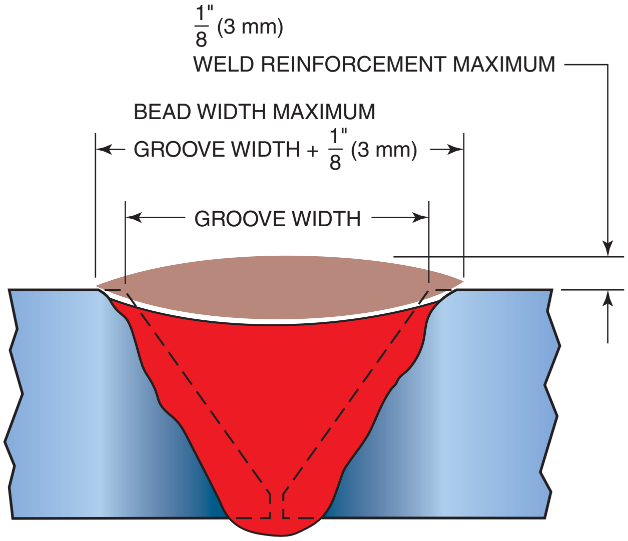

Figure 3.26
The cover pass should not be excessively large

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

COVER PASS

The last weld bead on a multiple pass weld is known as the **cover pass.** The cover pass may use a different electrode weave, or it may be the same as the filler beads. Keeping the cover pass uniform and neat looking is important. Most welds are not tested, and often the inspection program is only visual. Thus, the appearance might be the only factor used for accepting or rejecting welds.

The cover pass should be free of any visual defects such as undercut, overlap, porosity, or slag inclusions. It should be uniform in width and reinforcement. A cover pass should not be more than 1/8 in. (3 mm) wider than the groove opening, **Figure 3.26.** Cover passes that are too wide do not add to the weld strength.

PRACTICE 3-6

Cover Bead in All Positions

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 4

Module 9
Key Indicator 1, 2

Using a properly set up and adjusted arc welding machine, proper safety protection, E7018 arc welding electrodes having a 1/8-in. (3-mm) diameter, and one or more pieces of mild steel plate, 6 in. (152 mm) long × 1/4 in. (6 mm) thick, you will make a cover bead in each position.

Remember, any time an E7018 low-hydrogen-type electrode is to be used, the weave pattern, if used, must not be any larger than 2-1/2 times the diameter of the electrode. This weave cannot be any larger than 5/16 in. (7 mm) wide. Start welding at one end of the plate and weld to the other end. The weld bead should be about 5/16 in. (7 mm) wide having no more than 1/8 in. (3 mm) of uniform buildup, **Figure 3.27.** The weld buildup should have a smooth transition at the toe, with the plate and the face somewhat

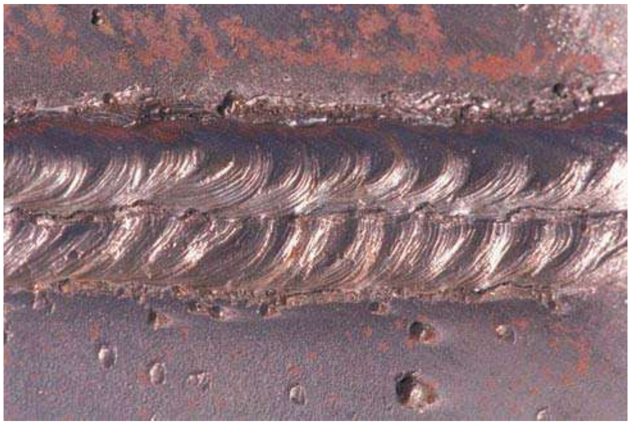

Figure 3.27
Practice cover pass
Courtesy of Larry Jeffus

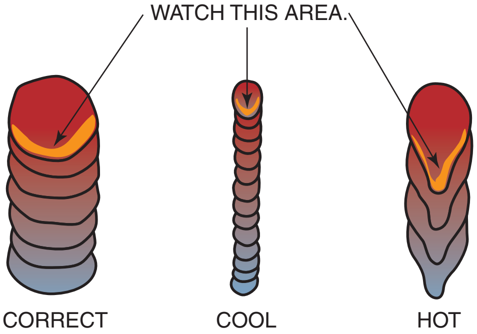

Figure 3.28
Watch the back edge of the weld pool to determine the correct current

convex. Undercut at the toe and a concave or excessively built-up face are the most common problems. Watch the sides of the bead for undercut. When undercut occurs, keep the electrode just ahead of the spot until it is filled in. There should be a smooth transition between the weld and the plate (refer to Figure 3.26). The shape of the bead face can be controlled by watching the trailing edge of the molten weld pool. That trailing edge is the same as the finished bead, **Figure 3.28.**

Deep penetration is not required with this weld and may even result in some weakening. After the weld is completely cooled, chip and inspect it for uniformity and defects. Repeat the welds as necessary until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PLATE PREPARATION

When welding on thick plate, it is impossible or impractical for the welder to try to get 100% penetration without preparing the plate for welding. The preparation of the plate is usually in the form of a **weld groove.** The groove can be cut into one side or both sides of the plate, and it may be cut into either just one plate or both plates of the joint, **Figure 3.29.** The type, depth, angle, and location of the groove are usually determined by a code standard that has been qualified for the specific job.

For SMA welds on plate 1/4 in. (6 mm) or thicker that need to have a weld with 100% joint penetration, the plate must be grooved. The groove may be ground, flame-cut, gouged, or machined on the edge of the plate before or after the assembly. Bevels and V-grooves are best if they are cut before the parts are assembled. J-grooves and U-grooves can be cut either before or after assembly, **Figure 3.30.** The lap joint is seldom prepared with a groove because little or no strength can be gained by grooving this joint. The only advantage to grooving the lap joint design is to give additional clearance.

Plates that are thicker than 3/8 in. (10 mm) can be grooved on both sides but may be prepared on only one side. The choice to groove one

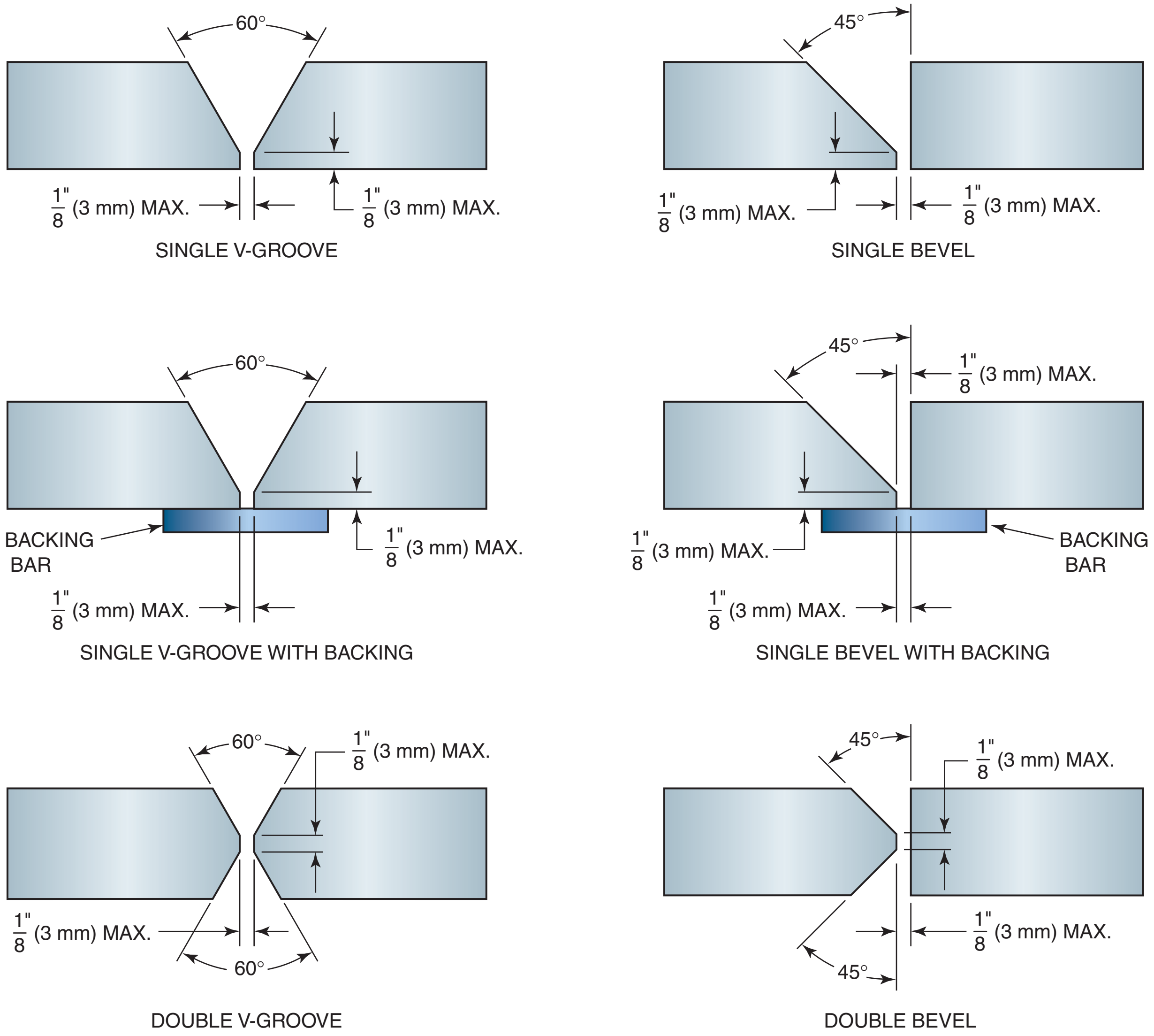
60°
1/8" (3 mm) MAX.
1/8" (3 mm) MAX.
SINGLE V-GROOVE
45°
1/8" (3 mm) MAX.
1/8" (3 mm) MAX.
SINGLE BEVEL
60°
BACKING BAR
1/8" (3 mm) MAX.
1/8" (3 mm) MAX.
SINGLE V-GROOVE WITH BACKING
45°
1/8" (3 mm) MAX.
1/8" (3 mm) MAX.
BACKING BAR
1/8" (3 mm) MAX.
SINGLE BEVEL WITH BACKING
60°
1/8" (3 mm) MAX.
1/8" (3 mm) MAX.
60°
DOUBLE V-GROOVE
45°
1/8" (3 mm) MAX.
1/8" (3 mm) MAX.
45°
DOUBLE BEVEL

Figure 3.29
Typical butt joint preparations

or both sides is most times determined by joint design, position, and application. A tee joint in thick plate is easier to weld and will have less distortion if it is grooved on both sides. Plate in the flat position is usually grooved on only one side unless it can be repositioned. Welds that must have little distortion or that are going to be loaded equally from both sides are usually grooved on both sides. Sometimes plates are either grooved and welded or just welded on one side, and then back gouged and welded, **Figure 3.31. Back gouging** is a process of cutting a groove in the back side of a joint that has been welded. Back gouging can ensure 100% fusion at the root and remove discontinuities of the root pass. This process can also remove the root pass metal if the properties of the metal are not desirable to the finished weld, **Figure 3.32.** After back gouging, the groove is then welded. See Chapters 6, 7, and 8 in *Welding: Skills, Processes and Practices for Entry-Level Welders: Book One* for more information on the various methods of gouging.

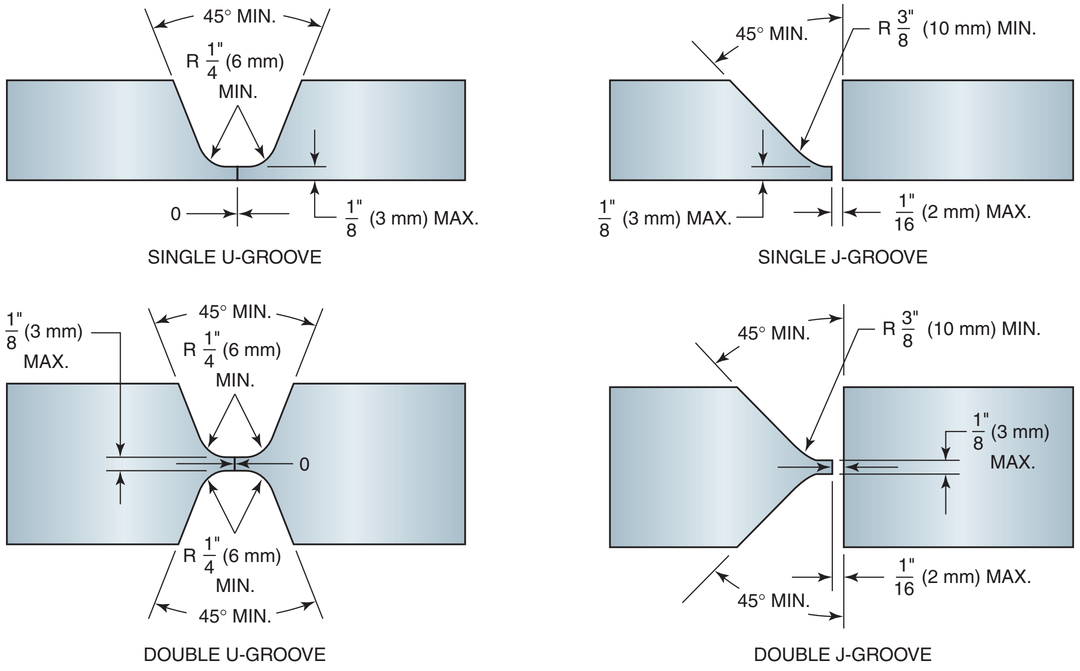
45° MIN.
R 1/4" (6 mm) MIN.
0
1/8" (3 mm) MAX.
SINGLE U-GROOVE
45° MIN.
R 3/8" (10 mm) MIN.
1/8" (3 mm) MAX.
1/16" (2 mm) MAX.
SINGLE J-GROOVE
1/8" (3 mm) MAX.
45° MIN.
R 1/4" (6 mm) MIN.
0
R 1/4" (6 mm) MIN.
45° MIN.
DOUBLE U-GROOVE
45° MIN.
R 3/8" (10 mm) MIN.
1/8" (3 mm) MAX.
45° MIN.
1/16" (2 mm) MAX.
DOUBLE J-GROOVE

Figure 3.30
Typical butt joint preparations

Heavy plate and pipe sections requiring preparations are often used in products manufactured under a code or standard. The American Welding Society, American Society of Mechanical Engineers, and American Bureau of Ships are a few of the agencies that issue codes and specifications. The AWS D1.1 and the ASME Boiler and Pressure Vessel (BPV) Section IX standards will be used in this chapter as the standards for multiple pass groove welds that will be tested. The groove depth and angle are determined by the plate or pipe thickness and process.

After the plate is beveled, a grinder can be used to clean off oxides and improve the fitup.

Note

Root faces on bevels and grooves with backing bars show a 1/8-in. root face. The 1/8-in. root face is a maximum allowable dimension and many welders prefer root faces prepared thinner, or all the way to a "knife edge" (see Figure 3.2).

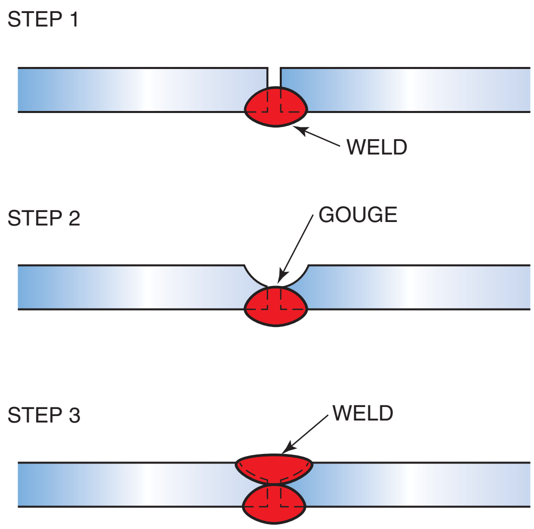
STEP 1
WELD
STEP 2
GOUGE
STEP 3
WELD

Figure 3.31
Back gouging sequence for a weld to ensure 100% joint penetration

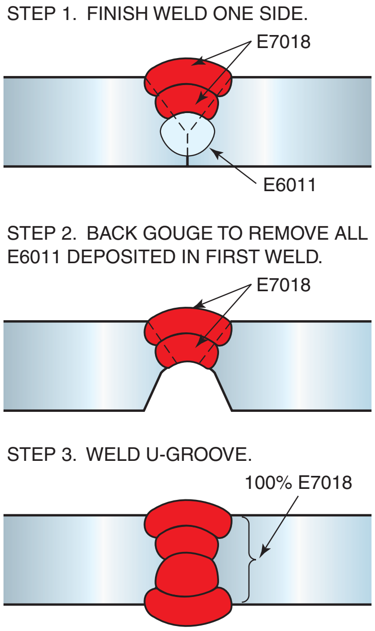
STEP 1. FINISH WELD ONE SIDE.
E7018
E6011
STEP 2. BACK GOUGE TO REMOVE ALL E6011 DEPOSITED IN FIRST WELD.
E7018
STEP 3. WELD U-GROOVE.
100% E7018

Figure 3.32
Back gouging to remove all weld metal used for the root pass or tacking

PREPARING SPECIMENS FOR TESTING

The detailed preparation of specimens for testing in this chapter is based on the structural welding code AWS D1.1 and the ASME BPV Code, Section IX. The maximum allowable size of fissures (crack or opening) in a **guided bend specimen** is given in codes for specific applications. Some of the standards are listed in ASTM E190 or AWS B4.0, AWS QC10, AWS QC11, and others. Copies of these publications are available from the appropriate organizations.

Acceptance Criteria for Face Bends and Root Bends

The weld specimen must first pass visual inspection before it can be prepared for bend testing. Visual inspection looks to see that the weld is uniform in width and reinforcement. There should be no arc strikes on the plate other than those on the weld itself. The weld must be free of both incomplete fusion and cracks. The joint penetration must be either 100% or as specified by the specifications. The weld must be free of overlap and undercut must not exceed either 10% of the base metal or 1/32 in. (0.8 mm), whichever is less.

Correct **weld specimen** preparation is essential for reliable results. The weld must be uniform in width and reinforcement and have no undercut or overlap. The weld reinforcement and backing strip, if used, must be removed flush to the surface, **Figure 3.33.** They can be machined or ground off. The plate thickness after removal must be a minimum of 3/8 in. (10 mm), and the pipe thickness must be equal to the pipe's original wall thickness. The specimens may be cut out of the test weldment by using an abrasive disc, by sawing, or by cutting with a torch. Flame-cut specimens must have the edges ground or machined smooth after cutting.

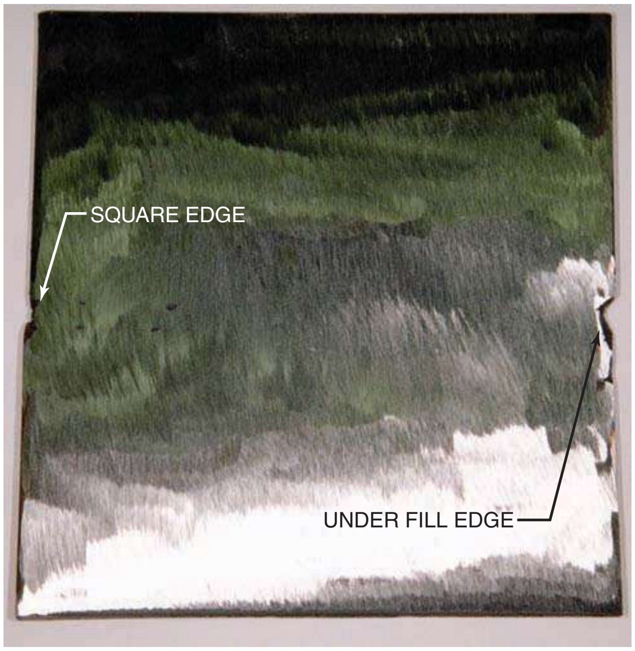

Figure 3.33
Plate ground in preparation for removing test specimens
Courtesy of Larry Jeffus

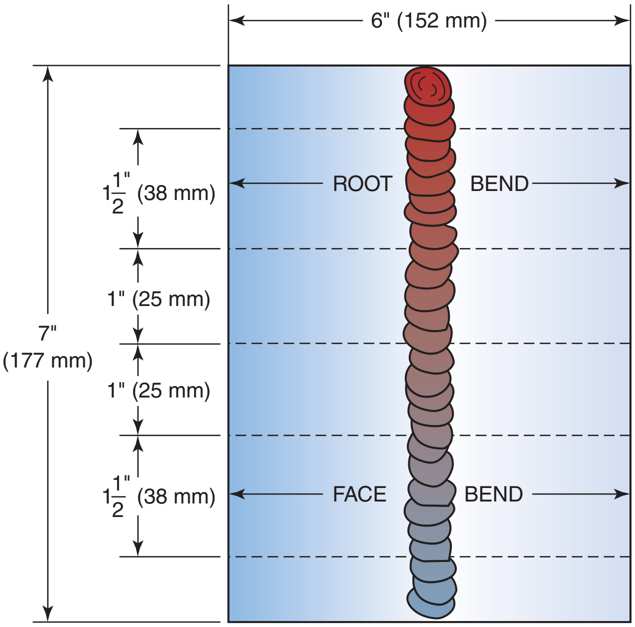

Figure 3.34
Sequence for removing guided bend specimens from the plate once welding is complete

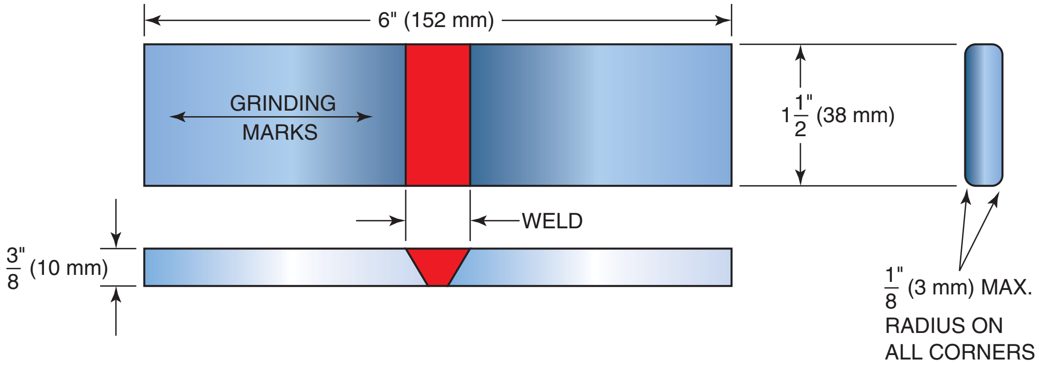

Figure 3.35
Guided bend specimen

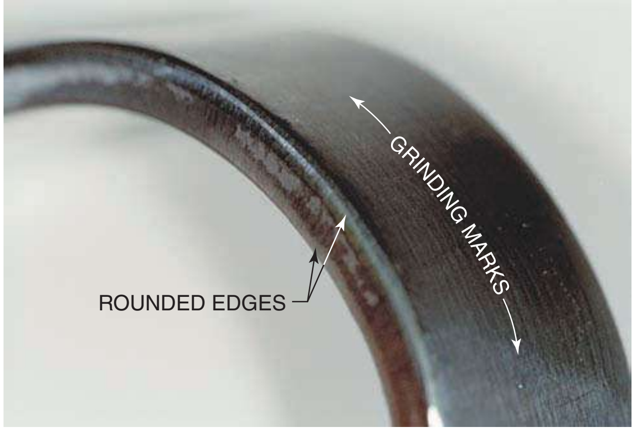

Figure 3.36
Guided bend test specimen
Courtesy of Larry Jeffus

This procedure is done to remove the heat-affected zone caused by the cut, **Figure 3.34.**

All corners must be rounded to a radius of 1/8 in. (3 mm) maximum, and all grinding or machining marks must run lengthwise on the specimen, **Figure 3.35** and **Figure 3.36.** Rounding the corners and keeping all marks running lengthwise reduce the chance of good weld specimen failure due to poor surface preparation.

The weld must pass both the face and root bends in order to be acceptable. After bending there can be no single defects larger than 1/8 in. (3 mm), and the sum of all defects larger than 1/32 in. (0.8 mm) but less than 1/8 in. (3 mm) must not exceed a total of 3/8 in. (10 mm) for each bend specimen. An exception is made for cracks that start at the edge of the specimen and do not start at a defect in the specimen.

RESTARTING A WELD BEAD

On all but short welds, the welding bead will need to be restarted after a welder stops to change electrodes. Because the metal cools as a welder changes electrodes and chips slag when restarting, the penetration and buildup may be adversely affected.

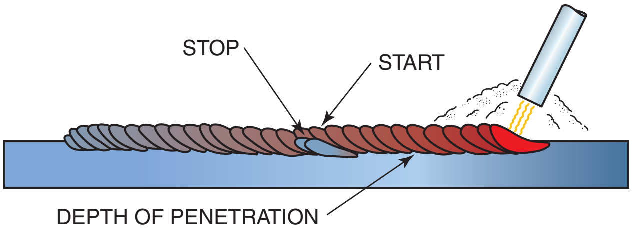

Figure 3.37
Tapering the size of the weld bead helps keep the depth of penetration uniform

When a weld bead is nearing completion, it should be tapered so that when it is restarted the buildup will be more uniform. To taper a weld bead, the travel rate should be increased just before welding stops. A 1/4-in. (6-mm) taper is all that is required. The taper allows the new weld to be started and the depth of penetration reestablished without having excessive buildup, **Figure 3.37.**

The slag should always be chipped and the weld crater should be cleaned each time before restarting the weld. This is important to prevent slag inclusions at the start of the weld.

The arc should be restarted in the joint ahead of the weld. The electrodes must be allowed to heat up so that the arc is stabilized and a shielding gas cloud is reestablished to protect the weld. Hold a long arc as the electrode heats up so that metal is not deposited. Slowly bring the electrode downward and toward the weld bead until the arc is directly on the deepest part of the crater where the crater meets the plate in the joint, **Figure 3.38.** The electrode should be low enough to start transferring metal. Next, move the electrode in a semicircle near the back edge of the weld crater. Watch

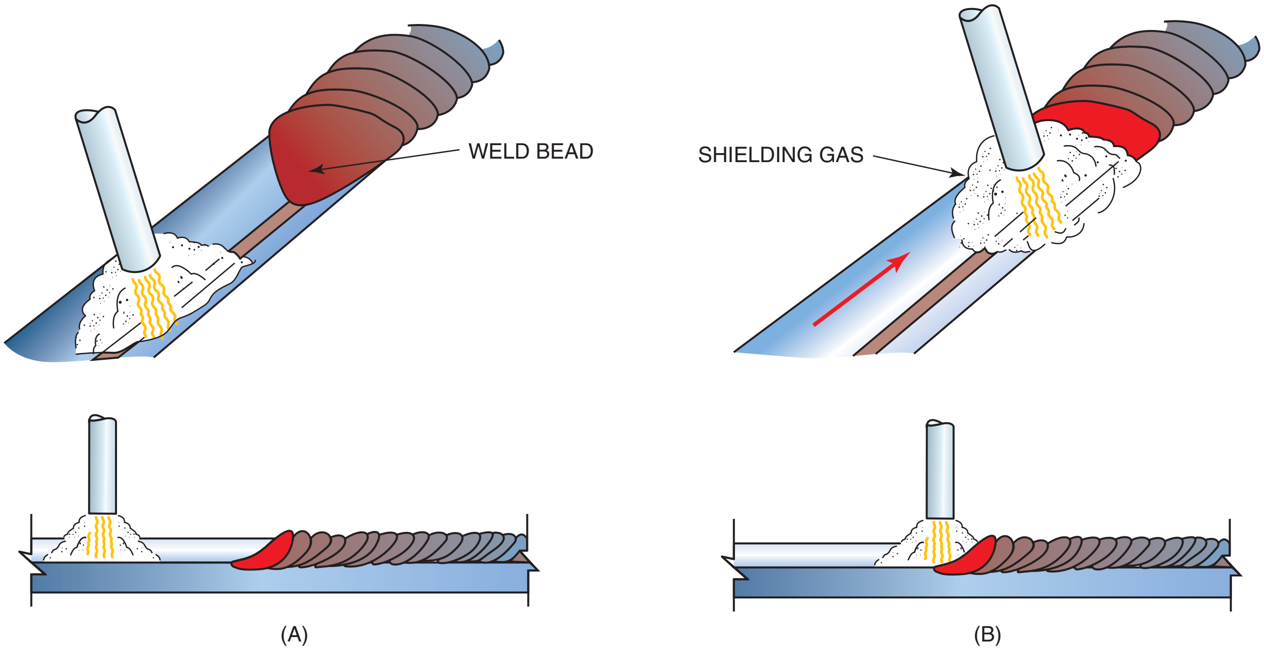

Figure 3.38
When restarting the arc, strike the arc ahead of the weld in the joint (A). Hold a long arc and allow time for the electrode to heat up, forming the protective gas envelope. Move the electrode so that the arc is focused directly on the leading edge (root) of the previous weld crater (B).

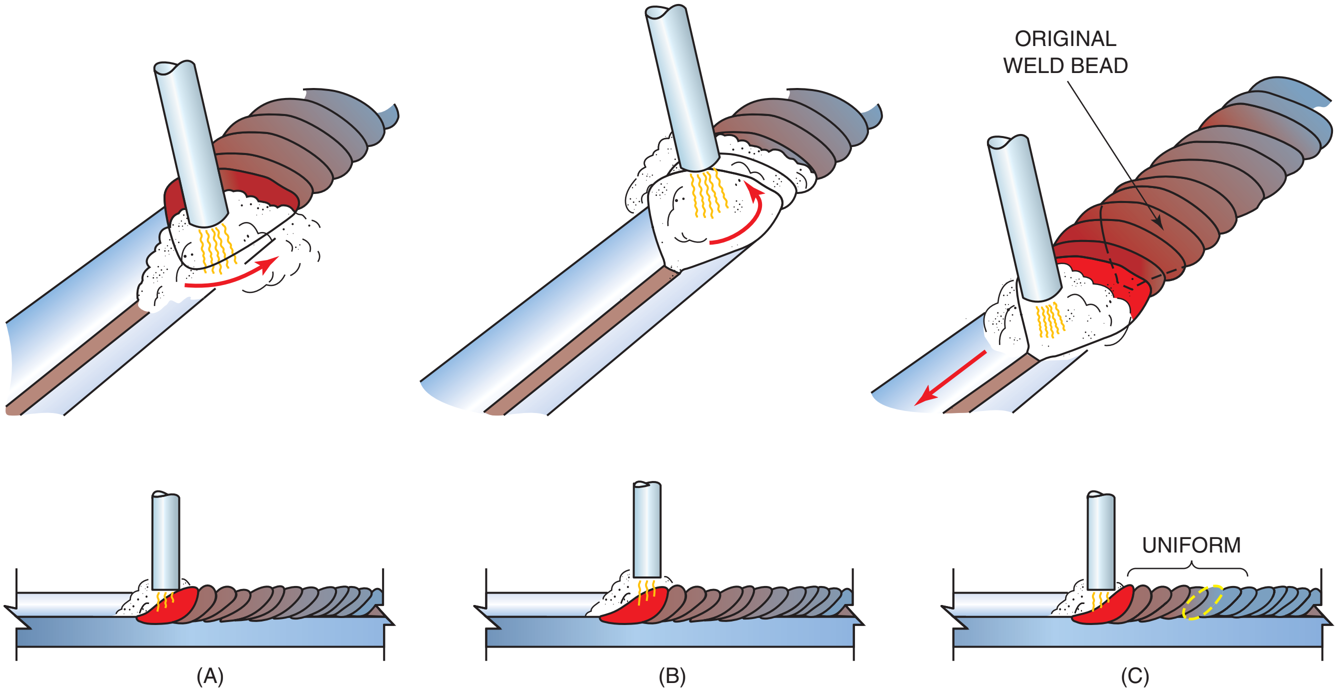

Figure 3.39
When restarting the weld pool after the root has been heated to the melting temperature, move the electrode upward along one side of the crater (A). Move the electrode along the top edge, depositing new weld metal (B). When the weld is built up uniformly with the previous weld, continue along the joint (C).

the buildup and match your speed in the semicircle to the deposit rate so that the weld is built up evenly, **Figure 3.39.** Move the electrode ahead and continue with the same weave pattern that was being used previously.

The movement to the root of the weld and back up on the bead serves both to build up the weld and reheat the metal so that the depth of penetration will remain the same. If the weld bead is started too quickly, penetration is reduced and buildup is high and narrow.

Starting and stopping weld beads in corners should be avoided. Tapering and restarting are especially difficult in corners, and this often results in defects, **Figure 3.40.**

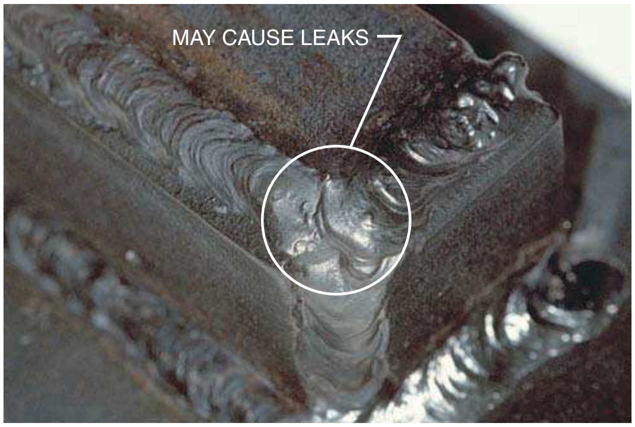

Figure 3.40
Correct method of welding through a corner. Stopping on a corner may cause leaks.
Courtesy of Larry Jeffus

Table 3.2 Preheat Temperatures for Arc Welding on Low-carbon Steels*

Plate Thickness in. (mm)	Minimum Temperature °F	°C
Up to 1/2 in. (13 mm)	70	21
1/2 in. (13 mm) to 1 in. (25 mm)	100	38
1 in. (25 mm) to 2 in. (51 mm)	200	95
Over 2 in. (51 mm)	300	150

*Metal should be above the dew point.
Allow 1 hour for each inch in order to provide uniform heating or for localized preheating. Check to ensure that preheat temperatures are sufficient to melt thermal crayons a minimum of 3 inches (76 mm) in all directions from the area to be welded prior to welding arc application.

PREHEATING AND POSTHEATING

Preheating is the application of heat to the metal before it is welded. This process helps to reduce cracking, hardness, distortion, and stresses by reducing the thermal shock from the weld and slowing the cool-down rate. Preheating is most often required on large, thick plates, when the plate is very cold, on days when the temperature is very cold, when small-diameter electrodes are used, on high-carbon or manganese steels, on complex shapes, or with fast welding speeds.

With the practices that are to be tested in this chapter, preheating should be used if the base metal to be welded is very cold. It may also be used to reduce distortion on thick sections and to reduce hardness caused by the rapid cooling of the weld, which may result in weld failure. Preheating the metal will slow the weld cooling rate, which results in a more ductile weld. **Table 3.2** lists the recommended preheat temperatures for plain carbon steels.

PRACTICE 3-7

Limited Thickness Welder Performance Qualification Test Plate with Backing and F3 and F4 Electrodes

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 1, 2, 3, 4, 6

Module 9
Key Indicator 1, 2

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 3-7. Date:

Title:

Welding SMAW of plate to plate.

Scope:

This procedure is applicable for V-groove plate with a backing strip within the range of 3/8 in. (10 mm) through 3/4 in. (20 mm). Welding may be performed in the following positions: 1G, 2G, 3G, and 4G.

Base Metal:

The base metal shall conform to M1020 or A36.
Backing material specification M1020 or A36.

Filler Metal:

The filler metal shall conform to AWS specification No. E6010 or E6011 root pass and E7018 for the cover pass from AWS specification A5.1. This filler metal falls into F-number F3 and F4 and A-number A-1.

Shielding Gas:

The shielding gas, or gases, shall conform to the following compositions and purity: N/A.

Joint Design and Tolerances:

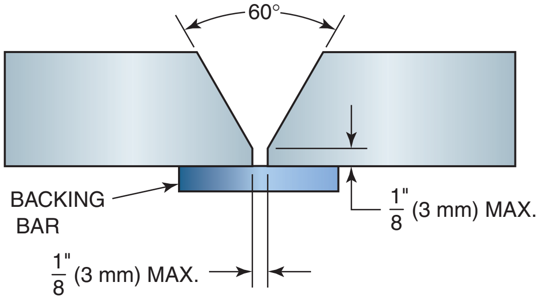

SINGLE V-GROOVE WITH BACKING

Preparation of Base Metal:

The V-groove is to be ground, flame-cut, or machined on the edge of the plate before the parts are assembled. Root faces may be prepared from a knife edge to the maximum allowable thickness of 1/8 in. All parts must be cleaned prior to welding of all contaminants, such as paints, oils, grease, or primers. Both inside and outside surfaces within 1 in. (25 mm) of the joint must be mechanically cleaned using a wire brush or grinder.

Electrical Characteristics:

The current shall be AC or DCRP.

The base metal shall be on the work lead or negative side of the line.

Preheat:

The parts must be heated to a temperature higher than 70°F (21°C) before any welding is started.

Backing Gas:

N/A

Welding Technique:

Tack weld the plates together with the backing strip. There should be about a 1/8-in. (3-mm) root gap between the plates. Use the E6010 or E6011 arc welding electrodes to make a root pass to fuse the plates and backing strip together. Clean the slag from the root pass and use either a hot pass or grinder to remove any trapped slag.

Using the E7018 arc welding electrodes, make a series of filler welds in the groove until the joint is filled.

Interpass Temperature:

The plate should not be heated to a temperature higher than 400°F (205°C) during the welding process. After each weld pass is completed, allow it to cool; the weldment must not be quenched in water.

Cleaning:

The slag can be chipped and/or ground off between passes but can only be chipped off of the cover pass.

Inspection:

Visually inspect the weld for uniformity and other discontinuities with criteria found in practice 3-8. If the weld passes the visual inspection, then it is to be prepared and guided bend tested according to criteria found in practice 3-8. Repeat each of the welds as needed until you can pass this test.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

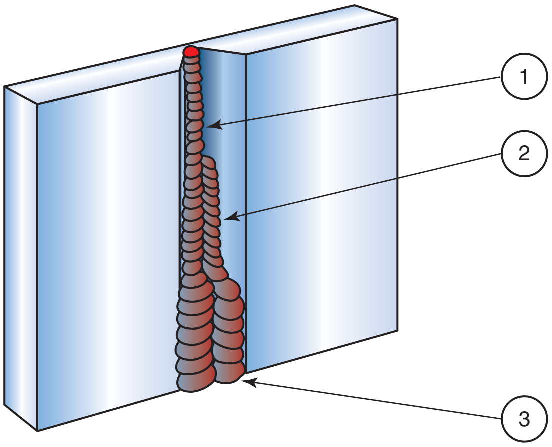

Note the sequence in depositing the beads

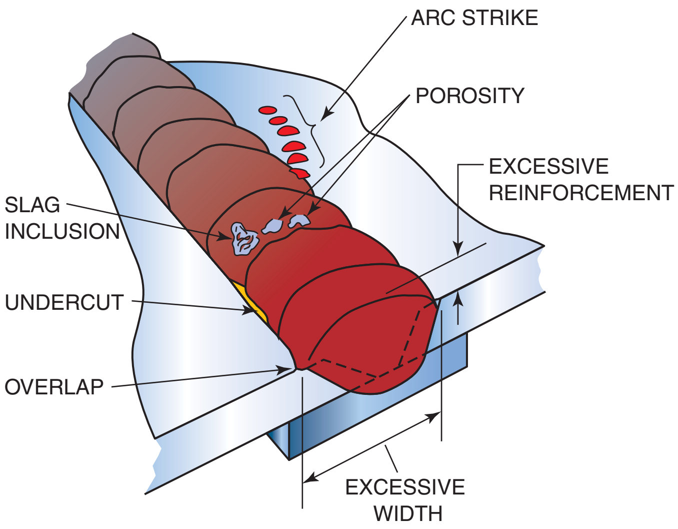
ARC STRIKE
POROSITY
EXCESSIVE REINFORCEMENT
SLAG INCLUSION
UNDERCUT
OVERLAP
EXCESSIVE WIDTH

Common discontinuities found during a visual examination

Sketches:

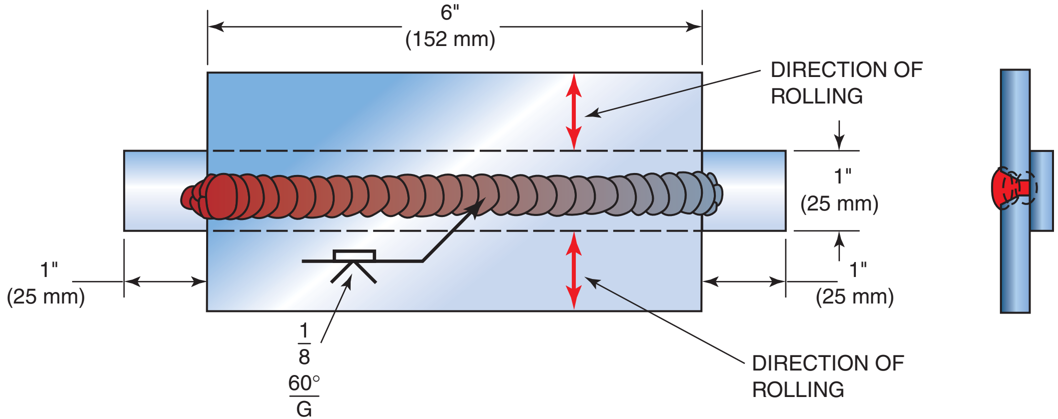
6"
(152 mm)
DIRECTION OF ROLLING
1"
(25 mm)
1"
(25 mm)
1"
(25 mm)
1/8
60°
G
DIRECTION OF ROLLING

Welding: Skills, Processes and Practices

MATERIAL:
3/8" x 6" MILD STEEL PLATE

PROCESS:
SMAW BUTT JOINT 1G

NUMBER:
PRACTICE 3–7

DRAWN BY:
RONNIE AGA

PRACTICE 3-8

SMAW Workmanship Sample and Welder Qualification Test Plate for Limited Thickness Horizontal 2G and 3G Positions with E7018 Electrodes

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 1, 2, 3, 4, 6

Module 9
Key Indicator 1, 2

Dimensional Tolerances:

2G Test plates: two (2); each 3/8 in. (10 mm) thick, 3 in. (75 mm) wide, and 7 in. (175 mm) long, one having a 45° bevel along one edge.
3G Test plates: two (2); each 3/8 in. (10 mm) thick, 3 in. (75 mm) wide, and 7 in. (175 mm) long, one having a 45° included bevel.
Backing strip: one (1); each either 1/4 in. (6 mm) or 3/8 in. (10 mm) thick, 1 in. (25 mm) wide, and 9 in. (228 mm) long.

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 3-8. Date:

Title:

Welding SMAW of plate to plate.

Scope:

This procedure is applicable for single bevel or V-groove plate with a backing strip within the range of 3/16 in. (5 mm) through 3/4 in. (20 mm). Welding may be performed in the following positions: 2G and 3G.

Base Metal:

The base metal shall conform to carbon steel M-1 or P-1, Group 1 or 2. Backing material specification carbon steel M-1 or P-1, Group 1, 2, or 3.

Filler Metal:

The filler metal shall conform to AWS specification no. E7018 from AWS specification A5.1. This filler metal falls into F-number F-4 and A-number A-1.

Shielding Gas:

The shielding gas, or gases, shall conform to the following compositions and purity: N/A.

Joint Design and Tolerances:

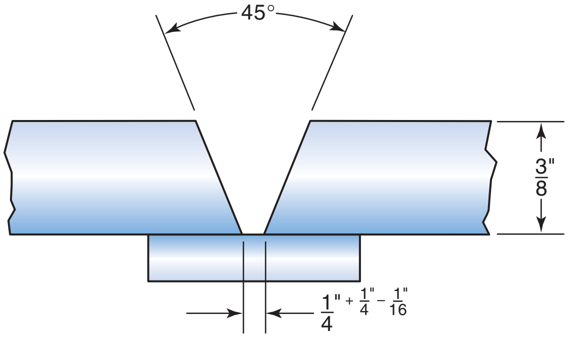

Preparation of Base Metal:

The bevel is to be flame-cut on the edge of the plate before the parts are assembled. The beveled surface must be smooth and free of notches. Any roughness or notches that are deeper than 1/64 in. (0.4 mm) must be ground smooth.

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The backing strip, groove face, and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically cleaned of slag, rust, and mill scale. Cleaning must be done with a wire brush or grinder down to bright metal.

Electrical Characteristics:

The current shall be direct current electrode positive (DCEP). The base metal shall be on the negative side of the line.

Welds	Filler Metal Diameter	Current	Amperage Range
Tack	3/32 in. (2.4 mm)	DCEP	70 to 115
Root and Filler	1/8 in. (3 mm)	DCEP	115 to 165
Filler	1/8 in. (3 mm)	DCEP	115 to 140

Preheat:

The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:

N/A

Safety:

Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:

Tack weld the plates together with the backing strip. There should be about a 1/4-in. (6-mm) root gap between the plates. Use the E7018 arc welding electrodes to make a root pass to fuse the plates and backing strip together. Clean the slag from the root pass, being sure to remove any trapped slag along the sides of the weld.

Using the E7018 arc welding electrodes, make a series of stringer or weave filler welds, no thicker than 1/4 in. (6 mm), in the groove until the joint is filled.

Interpass Temperature:

The plate should not be heated to a temperature higher than 350°F (175°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:

The slag must be cleaned off between passes. The weld beads may be cleaned by a hand wire brush, a hand-chipping hammer, a punch and hammer, or a needle scaler. All weld cleaning must be performed with the test plate in the welding position.

Visual Inspection Criteria for Entry Welders*:

There shall be no cracks, no incomplete fusion. There shall be no incomplete joint penetration in groove welds except as permitted for partial joint penetration welds.

*Courtesy of American Welding Society

The Test Supervisor shall examine the weld for acceptable appearance, and shall be satisfied that the welder is skilled in using the process and procedure specified for the text.

Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm).

Where visual examination is the only criterion for acceptance, all weld passes are subject to visual examination, at the discretion of the Test Supervisor.

The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length and the maximum diameter shall not exceed 3/32 in. (2.4 mm).

Welds shall be free from overlap.

Sketches:

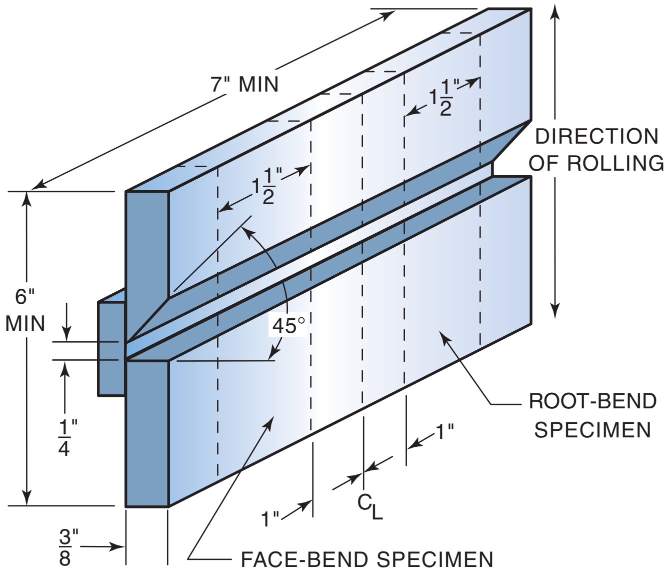

Bend Test:

The weld is to be mechanically tested only after it has passed the visual inspection. Be sure that the test specimens are properly marked to identify the welder, the position, and the process.

Specimen Preparation

For 3/8-in. (10-mm) test plates, two specimens are to be located in accordance with the requirements below. One is to be prepared for a transverse face bend, and the other is to be prepared for a transverse root bend.

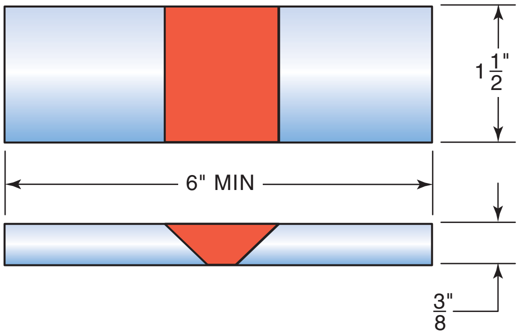

- *Transverse face bend.* The weld is perpendicular to the longitudinal axis of the specimen and is bent so that the weld face becomes the tension surface of the specimen. Transverse face bend specimens shall comply with the requirements of below.
- *Transverse root bend.* The weld is perpendicular to the longitudinal axis of the specimen and is bent so that the weld root becomes the tension surface of the specimen. Transverse root bend specimens shall comply with the requirements of above.

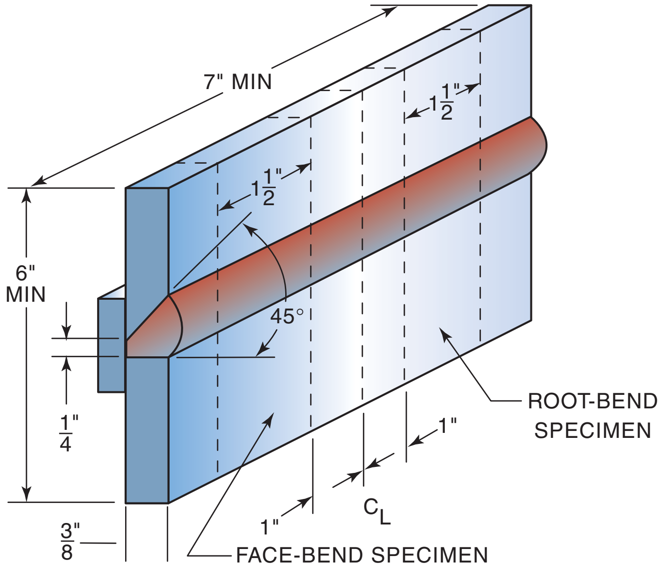

Acceptance Criteria for Bend Test*:

For acceptance, the convex surface of the face and root bend specimens shall meet both of the following requirements:

1. No single indication shall exceed 1/8 in. (3 mm), measured in any direction on the surface.
2. The sum of the greatest dimensions of all indications on the surface, which exceed 1/32 in. (0.8 mm), but are less than or equal to 1/8 in. (3 mm), shall not exceed 3/8 in. (10 mm).

*Courtesy of American Welding Society

Cracks occurring at the corner of the specimens shall not be considered unless there is definite evidence that they result from slag or inclusions or other internal discontinuities.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 1, 2, 3, 6

Module 9
Key Indicator 1, 2

PRACTICE 3-9

Limited Thickness Welder Performance Qualification Test Plate Without Backing

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 3-9. Date:

Title:

Welding SMAW of plate to plate.

Scope:

This procedure is applicable for V-groove plate without a backing strip within the range of 3/8 in. (10 mm) through 3/4 in. (20 mm).

Welding may be performed in the following positions: 1G, 2G, 3G, and 4G.

Base Metal:

The base metal shall conform to M1020 or A36.
Backing material specification: M1020 or A36.

Filler Metal:

The filler metal shall conform to AWS specification no. E6010 or E6011 root pass and E7018 for the cover pass from AWS specification A5.1. This filler metal falls into F-number F3 and F4 and A-number A-1.

Shielding Gas:

The shielding gas, or gases, shall conform to the following compositions and purity: N/A.

Joint Design and Tolerances:

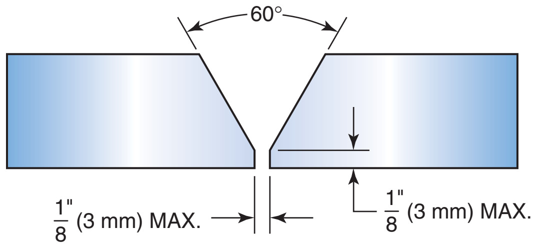

SINGLE V-GROOVE

Preparation of Base Metal:

The V-groove is to be ground, flame-cut, or machined on the edge of the plate before the parts are assembled. All parts must be cleaned prior to welding of all contaminants, such as paints, oils, grease, or primers. Both inside and outside surfaces within 1 in. (25 mm) of the joint must be mechanically cleaned using a wire brush or grinder.

Electrical Characteristics:

The current shall be DCRP. The base metal shall be on the work lead or negative side of the line.

Preheat:

The parts must be heated to a temperature higher than 70°F (21°C) before any welding is started.

Backing Gas:

N/A

Welding Technique:

Tack weld the plates together; there should be about a 1/8-in. (3-mm) root gap between the plates. Use the E6010 or E6011 arc welding electrodes to make a root pass to fuse the plates together. Clean the slag from the root pass and use either a hot pass or grinder to remove any trapped slag.

Using the E7018 arc welding electrodes, make a series of filler welds in the groove until the joint is filled.

Interpass Temperature:

The plate, outside of the heat affected zone, should not be heated to a temperature higher than 400°F (205°C) during the welding process. After each weld pass is completed, allow it to cool; the weldment must not be quenched in water.

Cleaning:

The slag can be chipped and/or ground off between passes but can only be chipped off of the cover pass.

Inspection:

Visually inspect the weld for uniformity and other discontinuities by using the criteria found in Practice 3-8. If the weld passes the visual inspection, then it is to be prepared and guided bend tested according to the "guided bend test" criteria found in Practice 3-8. Repeat each of the welds as needed until you can pass this test.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Sketches:

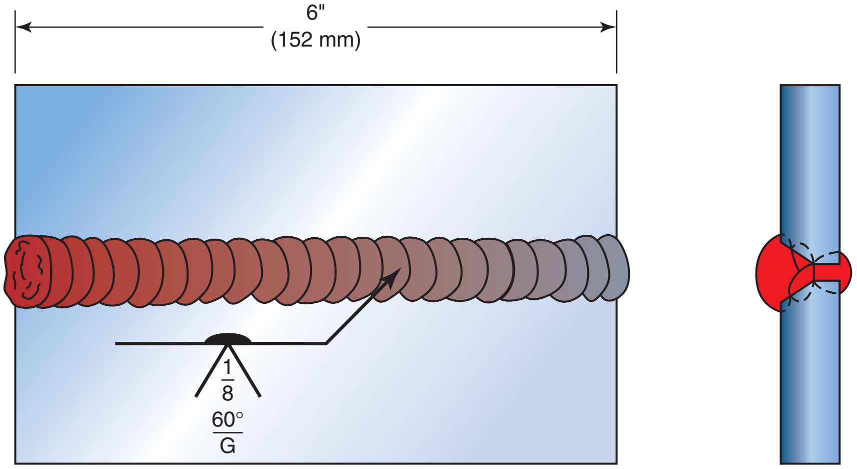

Welding: Skills, Processes and Practices	
MATERIAL: 3/8" TO 3/4" x 6" MILD STEEL PLATE	
PROCESS: SMAW BUTT JOINT 1G, 2G, 3G & 4G	
NUMBER: PRACTICE 3–9	DRAWN BY: KRISTI WEBB

Postheating is the application of heat to the metal after welding. Postheating is used to slow the cooling rate and reduce hardening.

Interpass temperature is the temperature of the metal during welding. The interpass temperature is given as a minimum and maximum. The minimum temperature is usually the same as the preheat temperature. If the plate cools below this temperature during welding, it should be reheated. The maximum temperature may be specified to keep the plate below a certain phase change temperature for the mild steel used in these practices. The maximum interpass temperature occurs when the weld bead cannot be controlled because of a slow cooling rate. When this happens, the plate should be allowed to cool down, but not below the minimum interpass temperature.

If, during the welding process, a welder must allow a practice weldment to cool so that the weld can be completed later, the weldment should be cooled slowly and then reheated before starting to weld again. A weld that is to be tested or that is done on any parts other than scrap should not be quenched in water.

POOR FITUP

Ideally, all welding will be performed on joints that are properly fitted. Most welds produced to a code or standard are properly fitted. Repair, prototype, and job shop welding, however, may not be cut and fitted properly. These welds must be performed under less than ideal conditions, but they still must be strong and have a good appearance.

To make a good weld on a poorly fitted joint requires some special skills. These welds also require a good welder, one whose skill is developed. A skilled welder can watch the molten weld pool and knows how to correct for problems before they develop into disasters. The welder must be able to read the molten weld pool correctly to make needed changes in amperage, current, electrode movement, electrode angle, and timing.

The amperage setting may have to be adjusted up or down by only a few amperes to make the necessary changes in molten weld pool size. Adjusting the machine is often preferable to lengthening the weave pattern excessively. The current may be changed from AC to DCSP or DCRP to vary the amount of heat input to the molten weld pool. Some electrodes can operate better than other electrodes with lower amperages on some currents. The current also will alter the forcefulness of some electrodes.

The “U,” “J,” and straight step patterns are usually the best to use, but they should not be moved more than required to close the gap or opening. On some poor-fitting joints, it is necessary to break and restart the arc in order to keep the molten weld pool under control. This will result in a weld with porosity, slag inclusions, and other defects. But it is often better to have a poor weld than to have no weld. A poor root weld can be capped with a sound weld to improve joint strength.

Changing the electrode angle from leading to trailing improves poor fit. Sometimes a very flat angle will also help. The time interval that the electrode is moved into and out of the molten weld pool is critical in maintaining weld control. Returning to a molten weld pool too often or too soon can cause the molten weld pool to drop out of the joint. In most cases, a welder should return to the molten weld pool only after it has started to cool.

On some joints, it is possible to make stringer beads on both sides of the joint until the gap is closed, **Figure 3.41.** Note that the beads are made alternately from the edges of the joint to the center. Welds made in this manner can have good weld soundness and strength, but they require more time to complete.

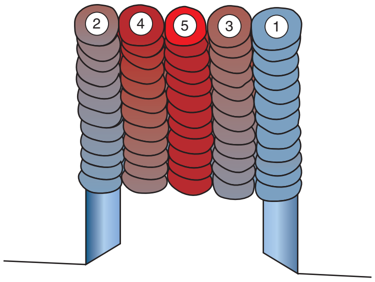

Figure 3.41
Multiple stringer beads used to close a large gap

PRACTICE 3-10

Single V-groove Open Root Butt Joint with an Increasing Root Opening

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 6

Module 9
Key Indicator 1, 2

For this practice, you will need a properly set up and adjusted arc welding machine, proper safety protection, E6010 or E6011 and E7018 arc welding electrodes having a 1/8-in. (3-mm) diameter, and two or more pieces of mild steel plate, 3/8 in. (10 mm) thick × 4 in. (102 mm) wide × 12 in. (305 mm) long. You will weld a single V-groove open root butt joint that has a poor fitup, starting from the close end.

Tack weld the plates together with a root opening of 1/8 in. (3 mm) at one end and 1/2 in. (13 mm) at the other end, **Figure 3.42.** Using the E6010 or the E6011 electrode, start the root pass at the narrow end and weld to the other end. As the root pass progresses along the widening root gap, care must be taken to maintain molten weld pool control. The “J” or “U” weave pattern works best with low current settings. Long time intervals for electrode movements will give the best weld control.

When the root pass is completed, clean the weld and make a hot pass to burn out any trapped slag. Finish the weld with filler passes using the E7018 electrode. Cool, chip, and inspect the weld for uniformity and defects. Repeat these welds until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 2, 4

Module 4
Key Indicator 6

Module 9
Key Indicator 1, 2

PRACTICE 3-11

Single V-groove Open Root Butt Joint with a Decreasing Root Opening

Using the same setup, equipment, and materials as described in Practice 3-10, you will weld a single V-groove open root butt joint that has a poor fitup, starting from the wide end.

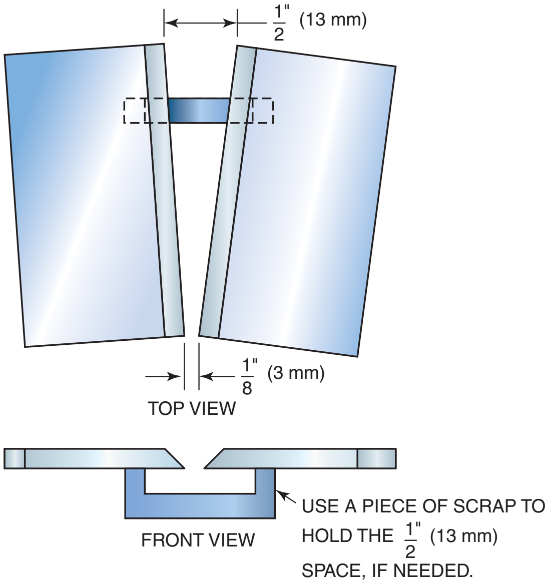
$\frac{1}{2}$" (13 mm)
$\frac{1}{8}$" (3 mm)
TOP VIEW
FRONT VIEW
USE A PIECE OF SCRAP TO HOLD THE $\frac{1}{2}$" (13 mm) SPACE, IF NEEDED.

Figure 3.42
Welding specimen with poor fitup

As in Practice 3-10, tack weld the plates together with a root opening of 1/8 in. (3 mm) at one end and 1/2 in. (13 mm) at the other end. Using the E6010 or E6011 electrode, start welding the root pass at the wide end. Both sides of the joint must be built up until it is possible to get the metal to flow together. The "J" or "U" weave pattern works best to control the bead.

When the root pass is completed, make a hot pass to clean out any trapped slag before making the filler passes with the E7018 electrode. After the weld is completed, cool, chip, and inspect it for uniformity and defects. Repeat these welds until you can consistently make welds free of defects. Turn off the welding machine and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

SUMMARY

Grooved welds on approximately 1/2-in. (13-mm)-thick plate are the most common test plates given to new welding applicants. Groove welds are used by many companies as the base welding skills performance test requirement for employment. The vertical and overhead positions are the most commonly used for the test. It is often assumed in the welding industry that a uniform, visually defect-free weld will successfully pass destructive testing. This assumption has great basis in fact because, in most cases, such a weld reflects the welder's skills required to produce quality welds, so, in many cases, applicant test welds are evaluated only by the weld shop foreman or supervisor for visual defects. For this reason, you should always attempt to make your welds as uniform in appearance as possible. Learning how to make a "pretty" groove weld can often mean the difference between successfully earning the job and losing out to another applicant.

REVIEW

1. Why are some weld joints grooved?
2. Sketch four of the standard grooves used for welded joints.
3. Why are some backing strips removed from the finished weld?
4. Why are backing tapes used on some joints?
5. Why is it very important to make a weld with a good root surface?
6. What are the two common methods of making a root pass on an open root joint?
7. How can small gaps between the weld plate and backing strip be closed?
8. What effect does changing from a trailing angle to a leading angle have on a weld?
9. What benefit would there be to the root pass if the electrode holder were rocked back and forth while keeping the electrode tip in the joint?
10. What might cause the bright color on the flux as a weld cools?

11. What can happen if the molten weld pool becomes too large on a root weld?
12. What can be done to increase the amount of high-strength welding electrode in the final weld if the root weld was made with a low-alloy electrode?
13. What is the purpose of the hot pass?
14. Why should a filler weld pass not have deep penetration?
15. Why is it important to have a good cover pass?
16. What can watching the back edge of a weld pool help you determine?
17. Other than penetration, why would thick butt joints be grooved on both sides?
18. List the things that a weld must be inspected for before it is ground for bend testing.
19. What determines the acceptance or rejection of a bend specimen?
20. What technique can be used to make restarting weld beads easier and more uniform?
21. Why should some weldments be preheated before welding starts?
22. What is postheating used for?
23. How can a wide gap in a joint be closed by welding?

CHAPTER 4

Gas Tungsten Arc Welding Equipment, Setup, Operation, and Filler Metals

OBJECTIVES

After completing this chapter, the student should be able to

- describe the gas tungsten arc welding process and list the other terms used to identify the process
- list four attributes of a quality tungsten electrode
- list four precautions taken to limit tungsten erosion
- contrast the various types of tungsten electrodes and how they are used
- shape the end of the tungsten electrode and clean it
- grind a point on a tungsten electrode using an electric grinder
- remove a contaminated tungsten end
- melt the end of the tungsten electrode into the desired shape
- list two advantages and two disadvantages of water-cooled GTA welding torches and air-cooled torches
- describe the purposes of the three hoses connecting a water-cooled torch to the welding machine
- choose an appropriate nozzle for the job
- accurately read a flowmeter
- compare and contrast the three types of welding current used for GTA welding
- list three advantages and three disadvantages of the shielding gases used in the GTA welding process
- set preflow time and postflow time on a GTA welding workstation
- list three problems that can occur as a result of an incorrect gas flow rate
- set up a GTAW workstation for carbon steel, stainless steel, and aluminum operations
- strike a GTA welding arc on carbon steel, stainless steel, and aluminum

KEY TERMS

cleaning action
collet
flowmeter
frequency
inert gas
postflow time
preflow time
rectification
spark gap oscillator
tungsten

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter

Module 1: Occupational Orientation

Key Indicator 1: Prepares time or job cards, reports or records

Key Indicator 4: Follows written instructions to complete work assignments

Module 7: Gas Tungsten Arc Welding (GTAW)

Key Indicator 1: Performs Safety inspections of GTAW equipment and accessories

Key Indicator 2: Makes minor external repairs to GTAW equipment and accessories

Carbon Steel

Key Indicator 3: Sets up for GTAW operations on carbon steel

Key Indicator 4: Operates GTAW equipment on carbon steel

Austenitic Stainless Steel

Key Indicator 8: Sets up for GTAW operations on austenitic stainless steel

Key Indicator 9: Operates GTAW equipment on austenitic stainless steel

Aluminum

Key Indicator 13: Sets up for GTAW operations on aluminum

Key Indicator 14: Operates GTAW equipment on aluminum

INTRODUCTION

The gas tungsten arc welding (GTAW) process is sometimes referred to as TIG, or heliarc. The term *TIG* is short for tungsten inert gas welding, and the much older term *heliarc* was used because helium was the first gas used for the process. The aircraft industry developed the GTAW process for welding magnesium during the late 1930s and the early 1940s. During that time, helium was the primary shielding gas used, along with DCEP welding current. These caused many of the problems that limited application of the GTA welding process. But improvements in gas composition and a better understanding of the importance of polarity improved the process's effectiveness and reduced its cost.

To use this process, an arc is established between a nonconsumable tungsten electrode and the base metal, which is called the work. Under the correct welding conditions, the electrode does not melt, although the work does at the spot where the arc impacts its surface and produces a molten weld pool. The filler metal is thin wire that is fed directly into the molten weld pool, where it melts. Since hot tungsten is sensitive to oxygen contamination, a good inert shielding gas is required to keep the air away from the hot tungsten and molten weld metal. The inert gas provides the needed arc characteristics and protects the molten weld pool. Because fluxes are not used, the welds produced are sound, free of slags, and as corrosion-resistant as the parent metal.

Before development of the GTAW process, welding aluminum and magnesium was difficult. The welds produced were porous and prone to corrosion.

When argon became plentiful and DCEN was recognized as more suitable than DCEP, the GTA process became more common. Until the development of Gas Metal Arc Welding (GMAW) in the late 1940s, GTAW was the only acceptable process for welding such reactive materials as aluminum, magnesium, titanium, and some grades of stainless steel, regardless of thickness. Reactive metals are ones that are easily contaminated when heated to their molten welding temperatures. Contamination can come from the air or can be picked up from surfaces containing oxides, oils, paints, or similar materials. Although economical for welding sheet metal, the process proved tedious and expensive for joining sections much thicker than 1/4 in. (6 mm). The eye-hand coordination required to make GTA welds is very similar to the coordination required for oxyfuel gas welding. GTA welding is often easier to learn when a person can oxyfuel gas weld.

TUNGSTEN

Tungsten, atomic symbol W, has the following properties:

- high tensile strength: 500,000 lb/in.^{2} (3447 kg/mm^2)
- hardness: Rockwell C45
- high melting temperature: 6170°F (3410°C)
- high boiling temperature: 10,700°F (5630°C)
- good electrical conductivity

Tungsten is produced mainly by reduction of its trioxide with hydrogen. Powdered tungsten is then purified to 99.95+%, compressed, and sintered (heated to a temperature below melting where grain growth can occur) to make an ingot. The ingot is heated to increase ductility and then is swaged and drawn through dies to produce electrodes. These electrodes are available in sizes varying from 0.01 in. to 0.25 in. (0.25 mm to 6 mm) in diameter. The tungsten electrode, after drawing, has a heavy black oxide that is later chemically cleaned or ground off.

The high melting temperature and good electrical conductivity make tungsten the best choice for a nonconsumable electrode. Its arc temperature, around 11,000°F (6000°C), is much higher than its melting temperature but not much higher than its boiling temperature of 10,700°F (5630°C).

As the tungsten electrode becomes hot, the arc between the electrode and the work will stabilize. Because electrons are more freely emitted from hot tungsten, the very highest temperature possible at the tungsten electrode tip is desired. Maintaining a balance between the heat necessary to have a stable arc and heat that is high enough to melt the tungsten requires an understanding of the GTA torch and electrode.

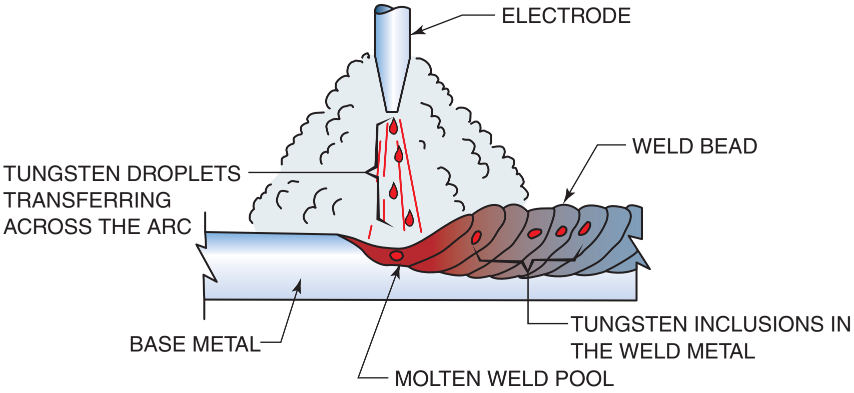
ELECTRODE
WELD BEAD
TUNGSTEN DROPLETS TRANSFERRING ACROSS THE ARC
BASE METAL
TUNGSTEN INCLUSIONS IN THE WELD METAL
MOLTEN WELD POOL

Figure 4.1
Some tungsten will erode from the electrode, be transferred across the arc, and become trapped in the weld deposit

The thermal conductivity of tungsten and the heat input are prime factors in the use of tungsten as an electrode. In general, tungsten is a good conductor of heat. This conductive property is what allows the tungsten electrode to withstand the arc temperature well above its melting temperature. The heat of the arc is conducted away from the electrode's end so fast that it does not reach its melting temperature. For example, a wooden match burns at approximately 3000°F (1647°C). Because aluminum melts at 1220°F (971°C), a match should easily melt an aluminum wire. However, a match will not even melt a 1/16-in. (2-mm) aluminum wire. The aluminum, like a tungsten electrode, conducts the heat away so quickly that it will not melt.

Because of the intense heat of the arc, some erosion of the electrode will occur. This eroded metal is transferred across the arc, **Figure 4.1.** Slow erosion of the electrode results in limited tungsten inclusions in the weld, which are acceptable. Standard codes give the size and amount of tungsten inclusions that are allowable in various types of welds. The tungsten inclusions are hard spots that cause stresses to concentrate, possibly resulting in weld failure. Although tungsten erosion cannot be completely eliminated, it can be controlled. A few ways of limiting erosion include:

- having good mechanical and electrical contact between the electrode and the collet
- using as low a current as possible
- using a water-cooled torch
- using as large a size of tungsten electrode as possible
- using DCEN current
- using as short an electrode extension from the collet as possible
- using the proper electrode end shape
- using an alloyed tungsten electrode

The torch end of the electrode is tightly clamped in a collet. The collet inside the torch is cooled by air or water. The **collet** is the cone-shaped sleeve that holds the electrode in the torch. Heat from both the arc and the tungsten electrode's resistance to the flow of current must be absorbed by the collet and torch. To ensure that the electrode is being cooled properly, be sure the collet connection is clean and tight. And for water-cooled torches, make sure water flow is adequate.

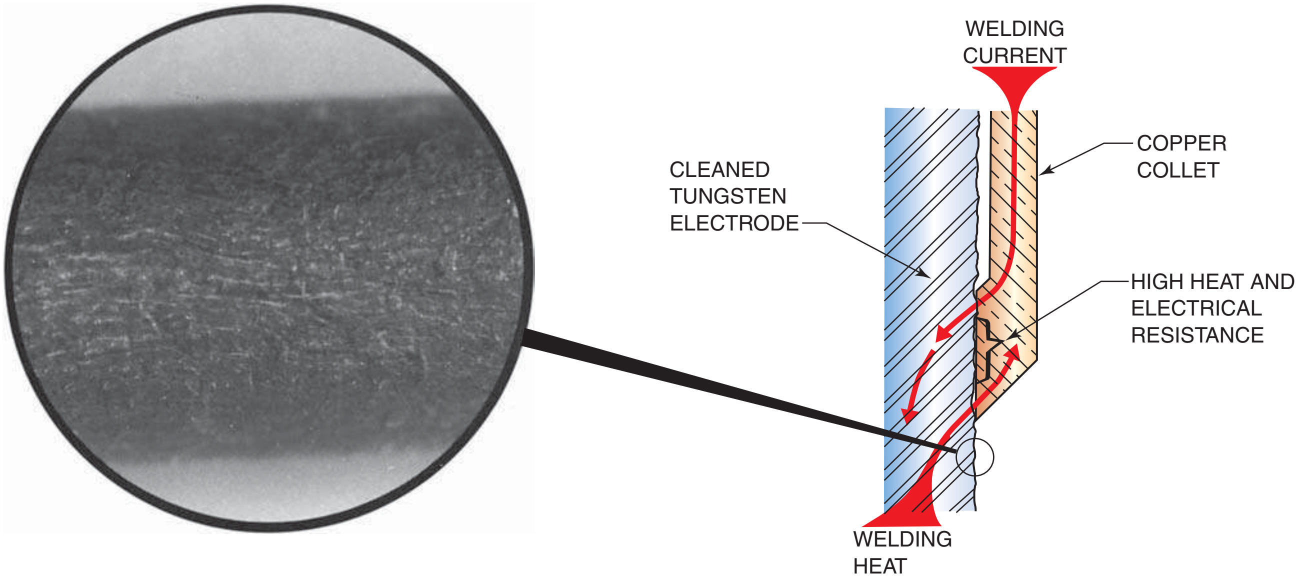
WELDING CURRENT
COPPER COLLET
CLEANED TUNGSTEN ELECTRODE
HIGH HEAT AND ELECTRICAL RESISTANCE
WELDING HEAT

Figure 4.2
Irregular surface of a cleaned tungsten electrode (poor heat transfer to collet)
Photo Courtesy of Larry Jeffus

Collet-tungsten connection efficiency is shown in **Figure 4.2** and **Figure 4.3.**

Large-diameter electrodes conduct more current because the resistance heating effects are reduced. However, excessively large sizes may result in too low a temperature for a stable arc.

In general, the current-carrying capacity at DCEN is about ten times greater than that at DCEP.

The particular electrode tip shape impacts the temperature and erosion of the tungsten. With DCEN, a pointed tip concentrates the arc as much as possible and improves arc, starting with either a high-voltage electrical

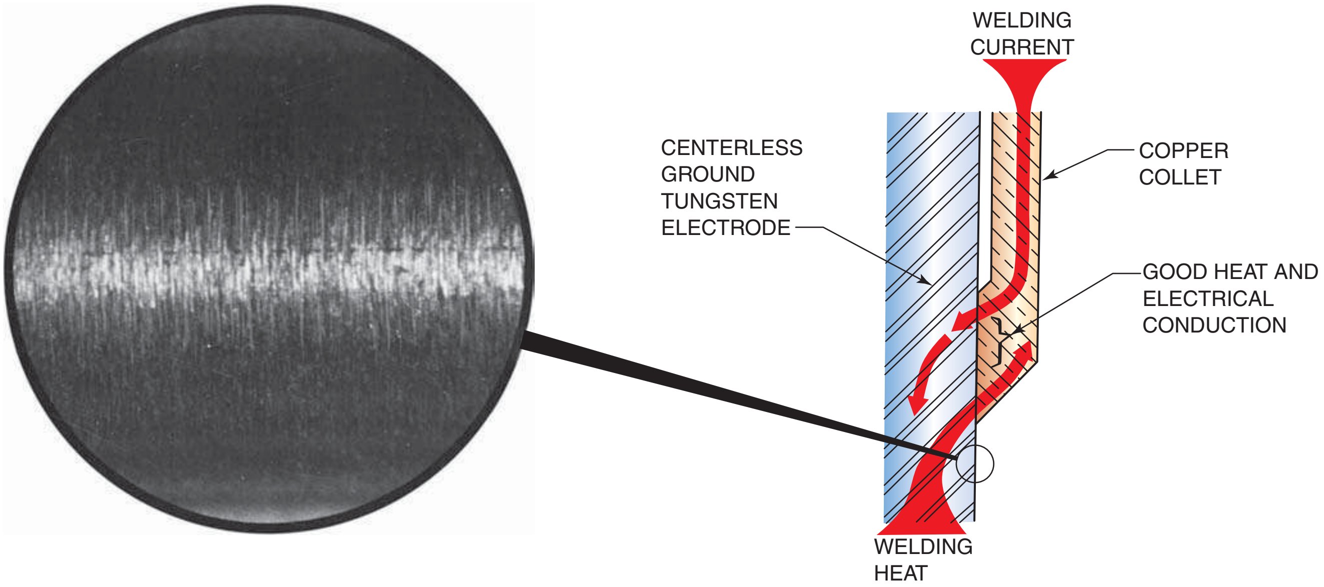
WELDING CURRENT
COPPER COLLET
CENTERLESS GROUND TUNGSTEN ELECTRODE
GOOD HEAT AND ELECTRICAL CONDUCTION
WELDING HEAT

Figure 4.3
Smooth surface of a centerless ground tungsten electrode (good heat transfer to collet)
Photo Courtesy of Larry Jeffus

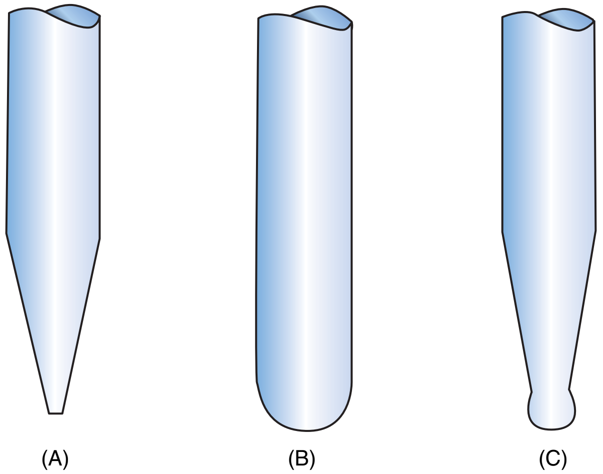

Figure 4.4
Basic tungsten electrode and shapes: pointed (A), rounded (B), and tapered with a balled end (C)

discharge or a touch start. Because DCEN does not put much heat on the tip, it is relatively cool, the point is stable, and it can survive extensive use without damage, **Figure 4.4A.**

With alternating current, the tip is subjected to more heat than with DCEN. To allow a larger mass at the tip to withstand the higher heat, the tip is rounded. The melted end must be small to ensure the best arc stability, **Figure 4.4B.**

DCEP has the highest heat concentration on the electrode tip. For this reason, a slight ball of molten tungsten is suspended at the end of a tapered electrode tip. The liquid tungsten surface tension, with the larger mass above the molten ball, holds it in place like a drop of water on your fingertip, **Figure 4.4C.**

TYPES OF TUNGSTEN ELECTRODES

Pure **tungsten** has a number of properties that make it an excellent nonconsumable electrode for the GTA welding process. These properties can be improved by adding cerium, lanthanum, thorium, or zirconium to the tungsten.

For GTA welding, tungsten electrodes are classified as the following:

- pure tungsten, EWP
- 1% thorium tungsten, EWTh-1
- 2% thorium tungsten, EWTh-2
- 1/4% to 1/2% zirconium tungsten, EWZr
- 2% cerium tungsten, EWCe-2
- 1% lanthanum tungsten, EWLa-1
- alloy not specified, EWG

See **Table 4.1.**

The type of finish on the tungsten must be specified as cleaned or ground. More information on composition and other requirements for tungsten welding electrodes is available in the AWS publication A5.12,

Table 4.1 Tungsten Electrode Types and Identification

AWS Classification	Tungsten Composition	Tip Color
EWP	Pure tungsten	Green
EWTh-1	1% thorium added	Yellow
EWTh-2	2% thorium added	Red
EWZr	1/4% to 1/2% zirconium added	Brown
EWCe-2	2% cerium added	Orange
EWLa-1	1% lanthanum added	Black
EWLa-1.5	1.5% lanthanum added	Gold
EWLa-2	2% lanthanum added	Blue
EWG	Alloy not specified	Grey

Specifications for Tungsten and Tungsten Alloy Electrodes for Arc Welding and Cutting.

Pure Tungsten Electrodes, EWP

Pure tungsten has the poorest heat resistance and electron emission characteristic of all the tungsten electrodes. It has a limited use with AC welding of metals, such as aluminum and magnesium, and is not recommended for use with inverter type welding power supplies. New alloying elements are now available with higher current carrying capacity; better emission characteristics and greater durability have replaced pure tungsten electrodes for AC welding in many shops.

Thoriated Tungsten Electrodes, EWTh-1 and EWTh-2

Thorium oxide (ThO_2), when added to tungsten in percentages of up to 0.6%, improves tungsten's current-carrying capacity. The addition of 1% to 2% of thorium oxide does not further improve current-carrying capacity. It does, however, help with electron emission. This can be observed by a reduction in the electron force (voltage) required to maintain an arc of a specific length. Thorium also increases the serviceable life of the tungsten. The improved electron emission of the thoriated tungsten allows it to carry approximately 20% more current. This also results in a corresponding reduction in electrode tip temperature, resulting in less tungsten erosion and subsequent weld contamination.

Thoriated tungsten also provides a much easier arc starting characteristic than pure or zirconiated electrodes. Thoriated tungsten works well with direct current electrode negative (DCEN). It can maintain a sharpened point well. It is very well suited for making welds on steel, steel alloys (including stainless), nickel alloys, and most other metals, other than aluminum or magnesium.

Thoriated electrodes do not work well with alternating current (AC). It is difficult to maintain a balled end, which used to be required and is still commonly used for AC welding. A thorium spike, **Figure 4.5,** may also develop on the end of a thorium electrode when it is overheated, disrupting a smooth arc. Furthermore, whether balled or sharpened to a point, when a thoriated tungsten electrode is overheated by exceeding its maximum amperage rating, small vaporized droplets of tungsten will transfer across the arc and into the work, possibly making the weld rejectable (see Figure 4.1).

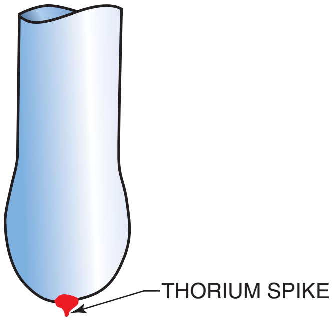

Figure 4.5
Thorium spike on a balled end tungsten electrode

CAUTION

Thorium is a very low-level radioactive oxide, but the level of radioactive contamination from a thorium electrode has not been found to be a health hazard during welding. It is, however, recommended that grinding dust be contained. Because of concern in other countries regarding radioactive contamination to the welder and welding environment, thoriated tungsten has been replaced with other alloys.

When a thorium spike becomes visible on an electrode tip, it should be assumed that the electrode was overheated and that the weld may be contaminated by vaporized droplets.

Zirconium Tungsten Electrodes, EWZr

Zirconium oxide (ZrO2) also helps tungsten emit electrons freely. The addition of zirconium to the tungsten has the same effect on the electrode characteristic as thorium, but to a lesser degree. Because zirconium tungsten is more easily melted than thorium tungsten, ZrO2 electrodes are recommended to be used with AC current. Because of the ease in forming the desired balled end on thorium versus zirconium tungsten, zirconium tungsten is normally the electrode chosen for AC welding of aluminum and magnesium alloys. Zirconiated tungsten is more resistant to weld pool contamination than pure tungsten, thus providing excellent weld qualities with minimal contamination.

Zirconiated tungsten also has the advantage over thoriated tungsten in that it is not radioactive.

Cerium Tungsten Electrodes, EWCe-2

Cerium oxide (CeO2) is added to tungsten to improve the current-carrying capacity in the same manner as does thorium. These electrodes were developed to replace thoriated tungsten electrodes because they are not made of a radioactive material. Cerium oxide electrodes have a current-carrying capacity similar to that of thoriated tungsten; however, they have an improved arc starting and arc stability characteristic, similar to that of thoriated tungsten. They can also provide a longer life than do most other electrodes, including thorium.

These electrodes have strong welder appeal for DC welding at low current settings; however, they are also used more and more with AC processes. In fact, ceriated electrodes along with lanthanum oxide electrodes are being recommended as a multipurpose electrode by the manufacturers of inverter based welding power supplies. With excellent arc starts at low amperages, ceriated tungsten has become more and more common in orbital tube and pipe manufacturing operations as well as with thin sheet and delicate work. Cerium electrodes contain approximately 2% cerium oxide.

Lanthanum Tungsten Electrodes, EWLa-1, EWLa-1.5, EWLa-2

Lanthanated tungsten is available as 1%, 1.5%, and 2%. Lanthanum trioxide has the lowest work function of any of the materials, thus it usually starts easiest and has the lowest temperature at the tip, which resists grain growth and promotes longer service life. Tests have proven that lanthanum electrodes last longer than do thorium electrodes when they are not overheated, and lanthanum electrodes are also resistant to thermal shock, which makes them a good choice for pulsed GTAW operations. Lanthanum electrodes require about 15% less power to initiate and maintain low current arcs. Lanthanum tungsten is a “rare earth” material and is not radioactive. Whereas lanthanated electrodes have been used successfully in Europe and Japan since 1993, lanthanum tungsten is a relative newcomer to the United States. This electrode has been primarily used for DC welding, but it is now being recommended, along with cerium electrodes, for AC welding, given its high current-carrying capacity and resistance to spitting during AC operations. The higher current-carrying capacity and resistance to spitting and spikes allow lanthanum and cerium electrodes to be used with a pointed tip prep on AC operations with aluminum alloys. The pointed tip allows for a more controllable arc and the ability to make smaller weld profiles in fillet joints. While slightly more expensive than thoriated electrodes, lanthanum can be used for just about any shop operation, thus eliminating confusion about electrode selection when multiple alloys are welded.

Alloy Not Specified, EWG

The EWG classification is for electrodes whose alloys have been modified by manufacturers. Such alloys have been developed and tested by manufacturers to meet specific welding criteria. The blend of alloying oxides are proprietary, meaning that specific alloy compositions are not normally available from manufacturers; however, they do provide welding characteristics for these electrodes.

SHAPING THE TUNGSTEN ELECTRODE

The desired end shape of tungsten can be obtained by grinding, breaking, remelting the end, or using chemical compounds. Tungsten is brittle and easily broken. Welders must be sure to make a smooth, square break where they want it to be located.

Grinding

A grinder or belt sander is often used to clean a contaminated tungsten electrode or to put a point at the end of the tungsten. The grinder or sander used to sharpen tungsten should have a fine, hard stone or a fine grit media for the belt (80-120 grit). It should be used for grinding tungsten only. Because of the hardness of the tungsten and its brittleness, the grinding stone chips off small particles of the electrode. A coarse grinding stone will result in more tungsten breakage and a poorer finish. If the grinder is used for metals other than tungsten, particles of these metals may become trapped on the tungsten as it is ground. The metal particles will quickly break free when the arc is started, resulting in contamination.

EXPERIMENT 4-1

Grinding the Tungsten to the Desired Shape

Module 1
Key Indicator 1, 4

Module 7
Key Indicator 2
Carbon Steel
Key Indicator 3
Austenitic Stainless Steel
Key Indicator 8
Aluminum
Key Indicator 13

Using an electric grinder with a fine grinding stone or a belt sander with 80-120 grit abrasive media, one piece of tungsten 2 in. (51 mm) long or longer, and safety glasses, you will grind a point on tungsten electrodes.

Because of the hardness of the tungsten, it will become hot. Its high thermal conductivity means that the heat will be transmitted quickly to your fingers. To prevent overheating, only light pressure should be applied

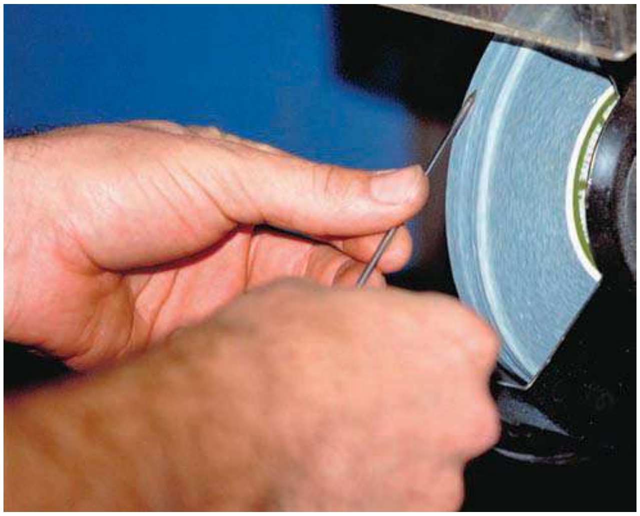

Figure 4.6
Correct method of grinding a tungsten electrode
Courtesy of Larry Jeffus

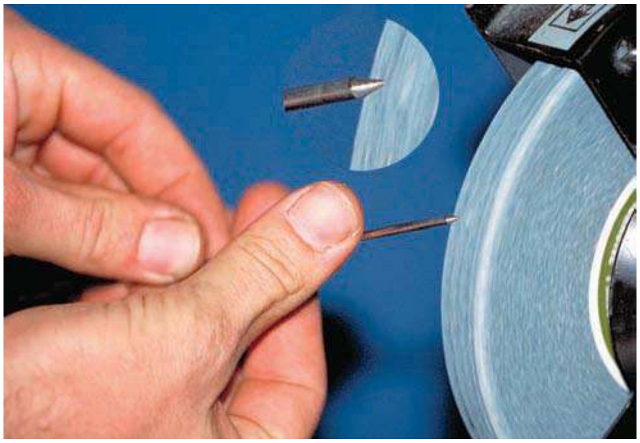

Figure 4.7
Incorrect method of grinding a tungsten electrode
Courtesy of Larry Jeffus

against the grinding wheel or belt. This will also reduce the possibility of accidentally breaking the tungsten.

Grind the tungsten so that the grinding marks run lengthwise, **Figure 4.6** and **Figure 4.7.** Lengthwise grinding reduces the amount of small particles of tungsten contaminating the weld. Move the tungsten up and down as it is twisted during grinding. This will prevent the tungsten from becoming hollow-ground.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Note

If an electrode has a large aluminum deposit at the end from contact with the work or filler metal, the deposit should be broken off before grinding on a stone wheel because nonferrous materials like aluminum, brass, and copper have a tendency to smear onto the wheel, which may cause a dangerous out-of-balance condition in the wheel.

CAUTION

When holding one end of the tungsten against the grinding wheel, the other end of the tungsten must not be directed toward the palm of your hand, **Figure 4.8.** This will prevent the tungsten from being stuck into your hand if the grinding wheel catches it and suddenly pushes it downward.

Breaking and Remelting

Tungsten is hard but brittle, resulting in a low impact strength. If tungsten is struck sharply, it will break without bending. When it is held against a sharp corner and hit, a fairly square break will result. **Figure 4.9, Figure 4.10,**

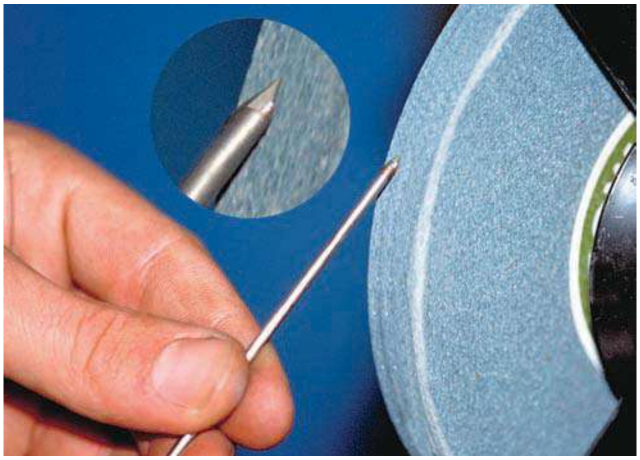

Figure 4.8
Correct way of holding a tungsten when grinding
Courtesy of Larry Jeffus

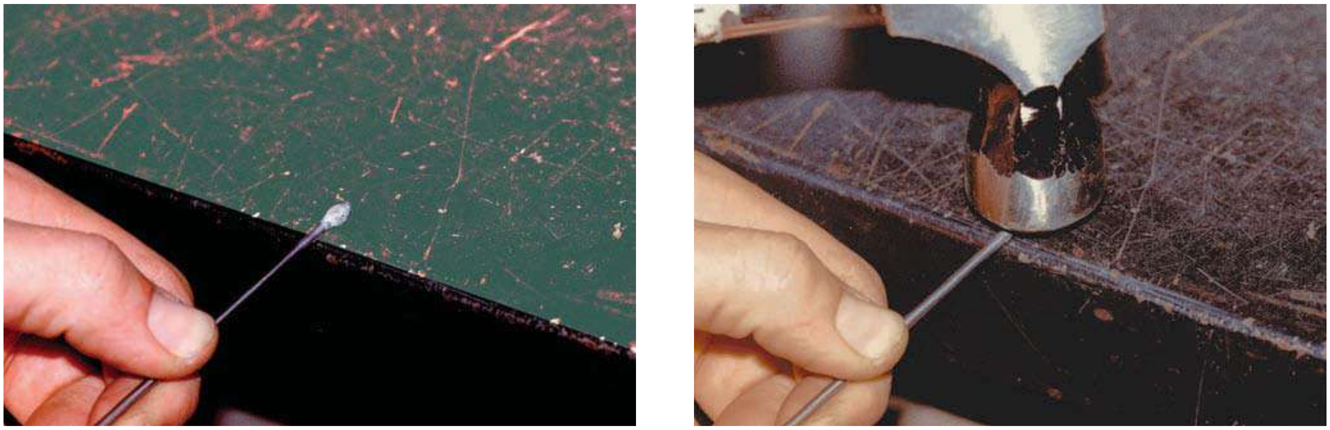

Figure 4.9
Breaking the contaminated end from a tungsten by striking it with a hammer
Courtesy of Larry Jeffus

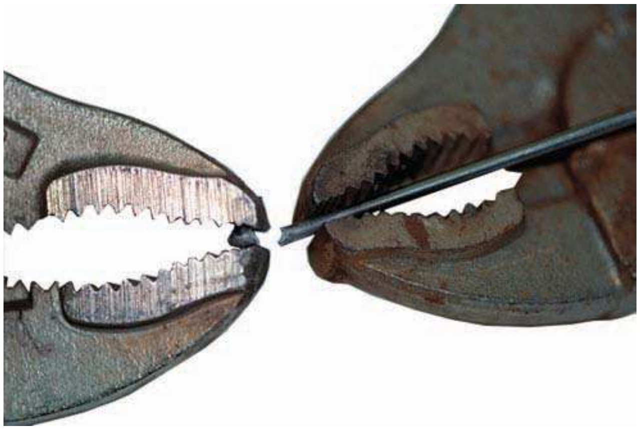

Figure 4.10
Correctly breaking the tungsten using two pairs of pliers
Courtesy of Larry Jeffus

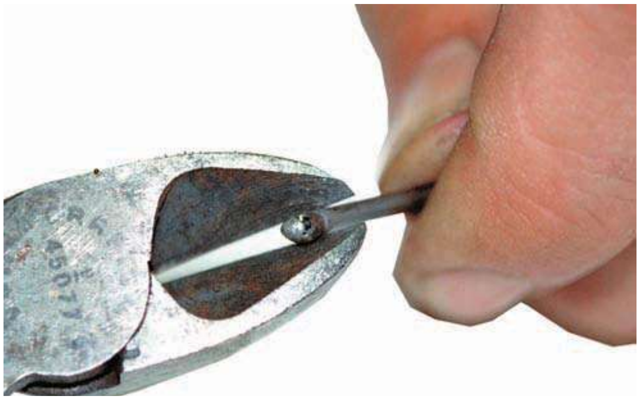

Figure 4.11
Using wire cutters to correctly break the tungsten
Courtesy of Larry Jeffus

and **Figure 4.11** show ways to break the tungsten correctly on a sharp corner using a hammer, with two pliers, and wire cutters.

Once the tungsten has been broken squarely, the end may be melted back so that it becomes somewhat rounded. The breaking and remelting technique is an older technique appropriate for pure tungsten and zirconiated tungsten electrodes due to their lower current-carrying capacity and inability to hold a pointed end. Cerium and lanthanum electrodes will maintain a pointed end much longer; therefore, this technique is rarely used with them unless the wider bead profile produced by a hemispherical tip prep is desired. This is accomplished by switching the welding current to DCEP and striking an arc under argon shielding on a piece of copper. If copper is not available, another piece of clean metal can be used. Do not use carbon, as it will contaminate the tungsten.

Note

Always wear eye protection as the end of the electrode can fly a good distance when struck.

Module 1
Key Indicator 1, 4

Module 7
Key Indicator 2
Carbon Steel
Key Indicator 3
Austenitic Stainless Steel
Key Indicator 8
Aluminum
Key Indicator 13

EXPERIMENT 4-2

Removing a Contaminated Tungsten End by Breaking

Using short scrap pieces of tungsten, pliers or wire cutters, and a light machinist's hammer, you will break the end from the tungsten.

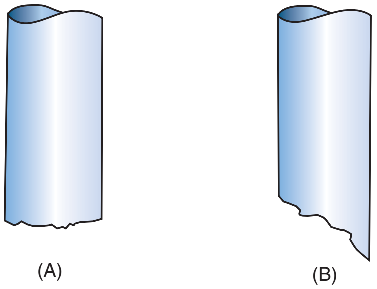

Figure 4.12
(A) Correctly broken tungsten electrode; (B) incorrectly broken tungsten electrode

Break about 1/4 in. (6 mm) from the end of the tungsten using the appropriate method, depending upon the diameter of the tungsten. Observe the break; it should be square and relatively smooth, **Figure 4.12.** The end of the tungsten should be broken only if the tungsten is badly contaminated.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Chemical Cleaning and Pointing

The tungsten can be cleaned and pointed using one of several compounds. The tungsten is heated by shorting it against the work. The tungsten is then dipped in the compound, a strong alkaline, which rapidly dissolves the hot tungsten. The chemical reaction occurs so quickly that enough additional heat is produced to keep the tungsten hot, **Figure 4.13.** When the tungsten is removed, cooled, and cleaned, the end will be tapered to a fine point. If the electrode is contaminated, the chemical compound will dissolve the tungsten, allowing the contamination to fall free.

Pointing and Remelting

The tapered tungsten with a balled end, a shape sometimes used for AC or DCEP welding, is made by first grinding or chemically pointing the electrode. Using DCEP, as in the procedure for the remelted broken end, strike an arc on some copper under argon shielding and slowly increase the current until a ball starts to form on the tungsten. The ball should be made large enough so that the color of the end stays between dull red and bright red. If the color turns white, the ball is too small and should be made larger. To increase the size of the ball, simply apply more current until the end begins to melt. Surface tension will pull the molten tungsten up onto the tapered end. Lower the current and continue welding. DCEP is seldom used for welding. If the tip is still too hot, it may be necessary to increase the size of the tungsten.

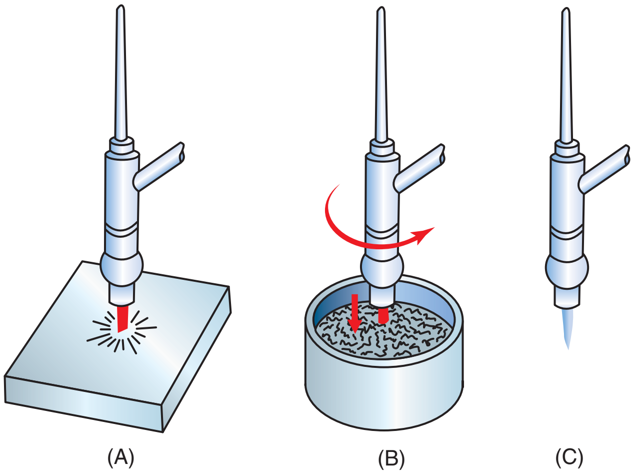

Figure 4.13
Chemically cleaning and pointing tungsten: (A) shorting the tungsten against the work to heat it to red hot, (B) inserting the tungsten into the compound and moving it around, and (C) cleaned and pointed tungsten ready for use

EXPERIMENT 4-3

Melting the Tungsten End Shape

Using a properly set up GTA welding machine, proper safety protection, one piece of copper or other clean piece of metal, and the tungsten that was sharpened and broken in Experiments 4-1 and 4-2, you will melt the end of the tungsten into the desired shape.

Properly install the tungsten, set the argon gas flow, switch the current to DCEP, and turn on the machine. Strike an arc on the copper and slowly increase the amperage. Watch the tungsten as it begins to melt and stop the current when the desired shape has been obtained, **Figure 4.14.**

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 4

Module 7
Key Indicator 2
Carbon Steel
Key Indicator 3
Austenitic Stainless Steel
Key Indicator 8
Aluminum
Key Indicator 13

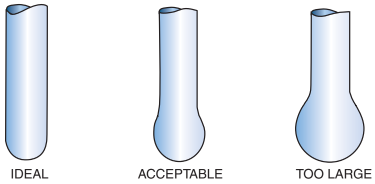

Figure 4.14
Melting the tungsten end shape

GTA WELDING EQUIPMENT

Figure 4.15 and **Figure 4.16** show two industrial applications of gas tungsten arc welding.

Torches

GTA welding torches are available water-cooled or air-cooled. The heat transfer efficiency for GTA welding may be as low as 20%. This means that 80% of the heat generated does not enter the weld. Much of this

Figure 4.15
Semiautomatic operation allows a stainless steel part to be GTA welded as it is turned past the torch
Courtesy of Lincoln Electric Company

Figure 4.16
An operator GTA welds a cap ring on a pneumatic tank
Courtesy of Lincoln Electric Company

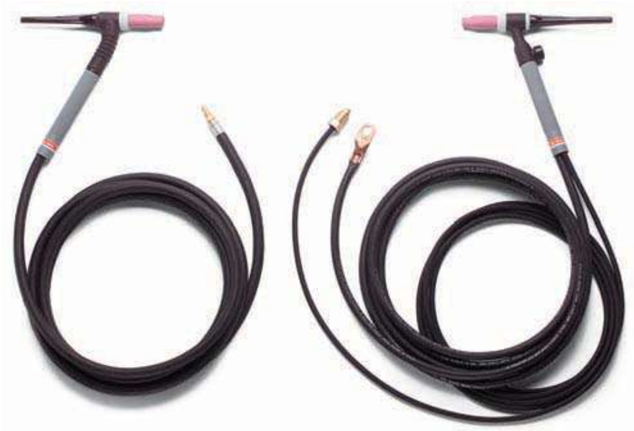

Figure 4.17
Power cable safety fuse
Courtesy of ESAB Welding & Cutting Products

heat stays in the torch and must be removed by some type of cooling method.

The water-cooled GTA welding torch is more efficient at removing waste heat than is an air-cooled torch. The water-cooled torch, as compared to the air-cooled torch, operates at a lower temperature, resulting in a lower tungsten temperature and less erosion.

The air-cooled torch is more portable because it has fewer hoses, and it may be easier to manipulate than the water-cooled torch. Also, the water-cooled torch requires a water reservoir or other system to give the needed cooling. The cooling water system should contain some type of safety device, **Figure 4.17,** to make it possible to shut off the power if the water flow is interrupted. The power cable is surrounded by the return water to keep it cool so that a smaller-size cable can be used. Without the cooling water, the cable quickly overheats and melts through the hose.

The water can become stopped or restricted for a number of reasons, such as a kink in the hose, a heavy object set on the hose, or failure to turn on the system. Water pressures higher than 35 psi (241 kg/mm^2) may cause the water hoses to burst. When an open system is used, a pressure regulator must be installed to prevent pressures that are too high from damaging the hoses.

GTA welding torch heads are available in a variety of amperage ranges and designs, **Figure 4.18.** The amperage listed on a torch is the maximum rating and cannot be exceeded without possible damage to the torch. The various head angles allow better access in tight places. Some of the heads can be swiveled easily to new angles. The back cap that both protects and tightens the tungsten can be short or long. Short caps are beneficial for getting the torch into restricted areas. Long caps can accommodate a full-length electrode, which will carry more heat away from the tip and allow slightly more current to be used before the electrode overheats. **Figure 4.19** and **Figure 4.20.**

Note

With the introduction of multipurpose electrode types, such as cerium and lanthanum, it is often no longer necessary to produce a balled end on an electrode when welding aluminum with alternating current. However, when making surfacing welds or buildups on aluminum castings, a hemispherical end preparation on an electrode will produce a wider arc and weld pool, which is desirable in these types of operations.

Hoses

A water-cooled torch has three hoses connecting it to the welding machine. The hoses are for shielding gas to the torch, cooling water to the

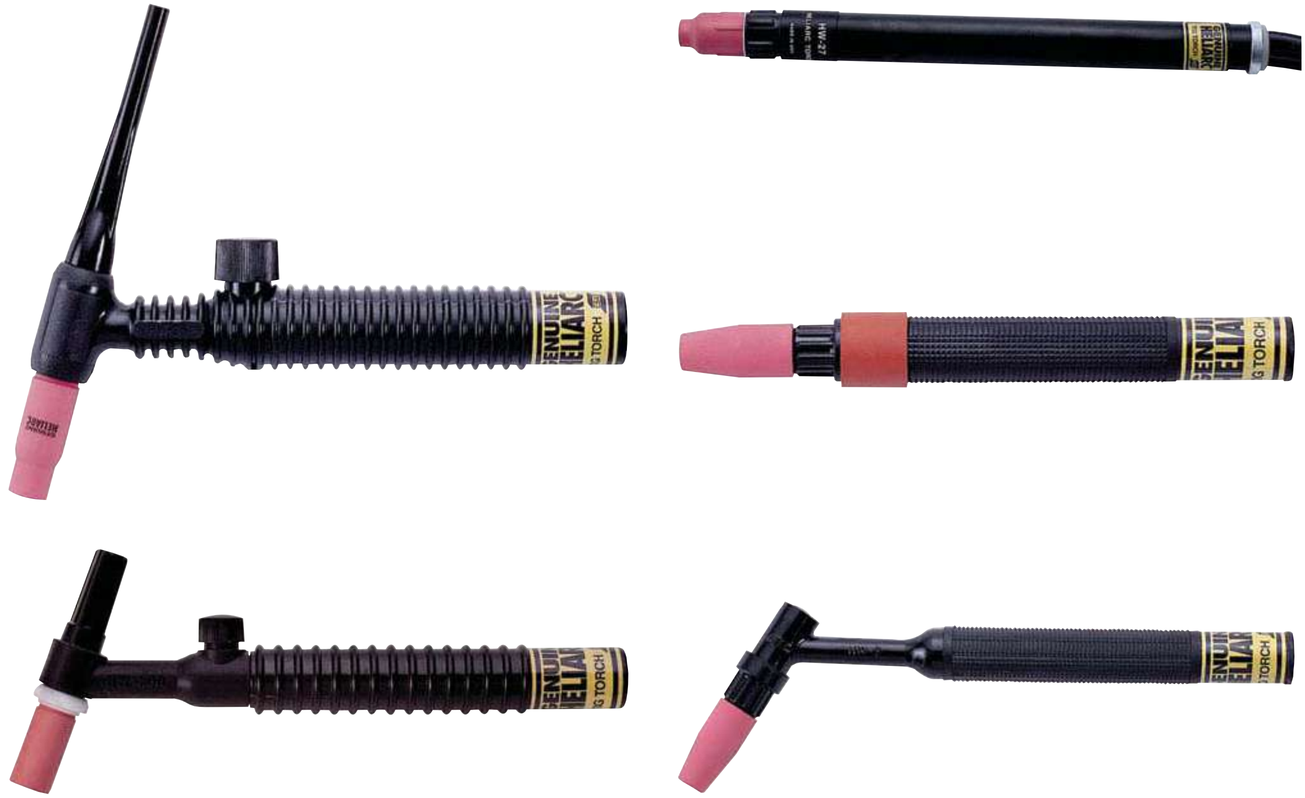

Figure 4.18
GTA welding torches
Courtesy of American Torch Tip

torch and cooling water return, and housing the power cables to the torch, **Figure 4.21.** Air-cooled torches may have one hose for shielding gas attached to the power cable, **Figure 4.22.**

The shielding gas hose must be plastic to prevent the gas from being contaminated. Rubber hoses contain oils that can be picked up by the gas, resulting in weld contamination.

The water-in hose may be made of any sturdy material. Water hose fittings have left-hand threads, and gas hose fittings have right-hand threads.

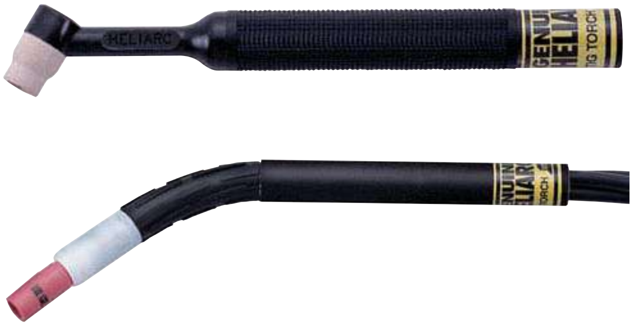

Figure 4.19
Short back caps are available for torches when space is a problem

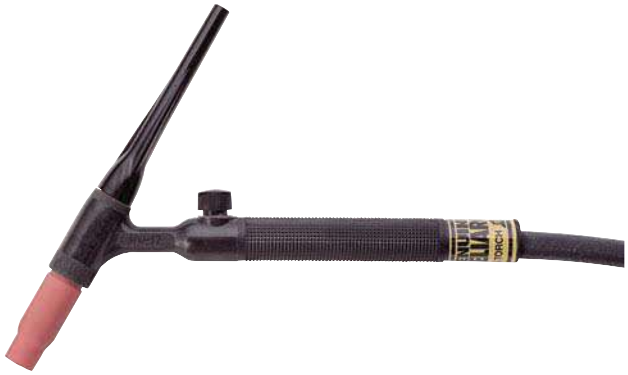

Figure 4.20
Long back caps allow tungstens that are a full 7 in. (177 mm) long to be used
Courtesy of ESAB Welding & Cutting Products

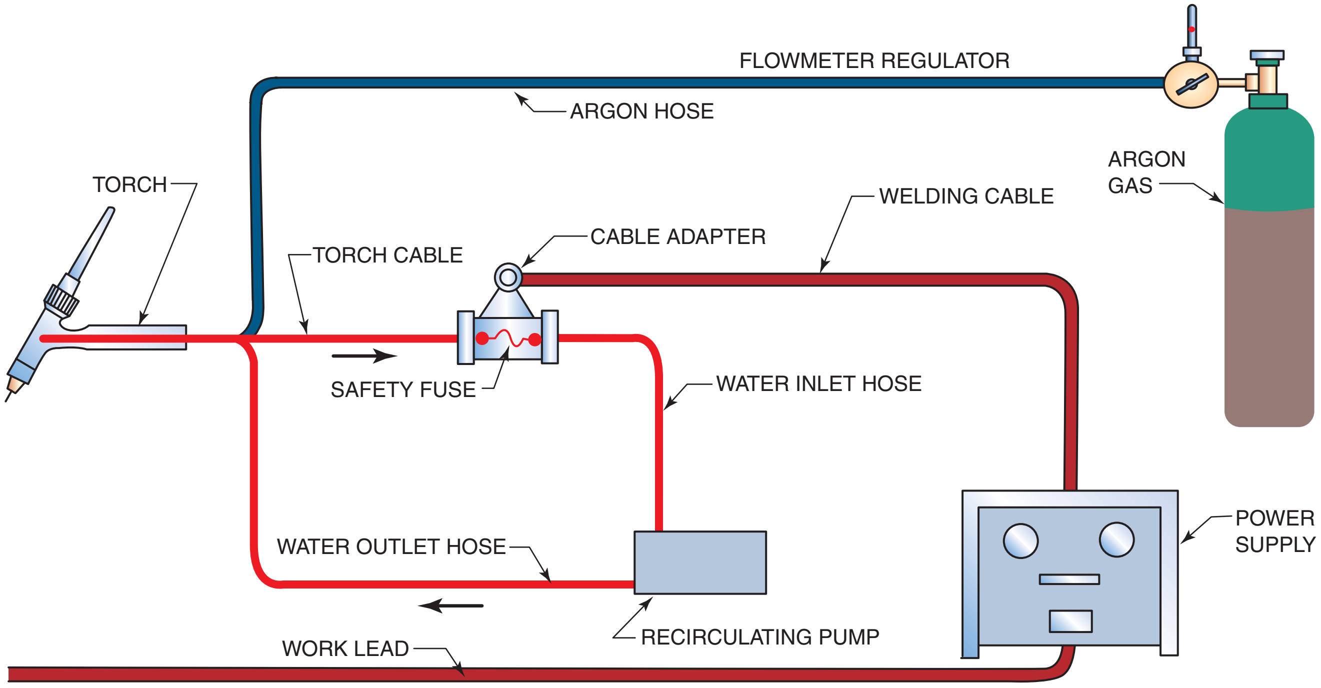
FLOWMETER REGULATOR
ARGON HOSE
ARGON GAS
TORCH
WELDING CABLE
CABLE ADAPTER
TORCH CABLE
SAFETY FUSE
WATER INLET HOSE
POWER SUPPLY
WATER OUTLET HOSE
RECIRCULATING PUMP
WORK LEAD

Figure 4.21
Schematic of a GTA welding setup with a water-cooled torch

This prevents the water and gas hoses from accidentally being reversed when they are attached to the welder. The return water hose also contains the welding power cable. This permits a much smaller-size cable to be used because the water keeps it cool.

The water must be supplied to the torch head and return around the cable. This allows the head to receive the maximum cooling from the water before the power cable warms it. Running the water through the torch first has another advantage. That is, when the water solenoid is closed, there is no water pressure in the hoses, which is particularly important. This fea-

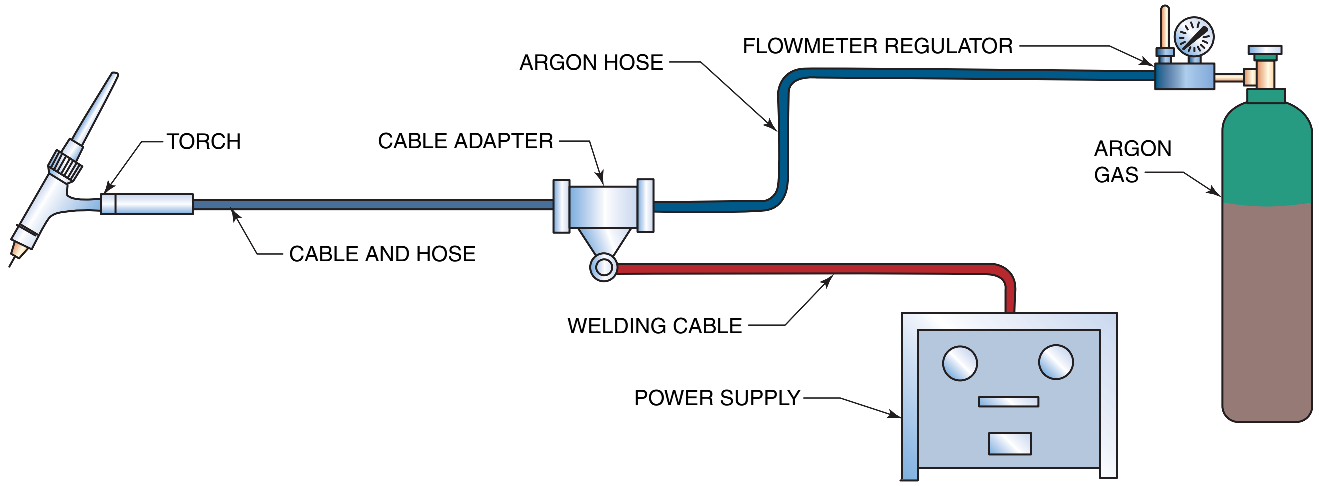
ARGON HOSE
FLOWMETER REGULATOR
TORCH
CABLE ADAPTER
ARGON GAS
CABLE AND HOSE
WELDING CABLE
POWER SUPPLY

Figure 4.22
Schematic of a GTA welding setup with an air-cooled torch

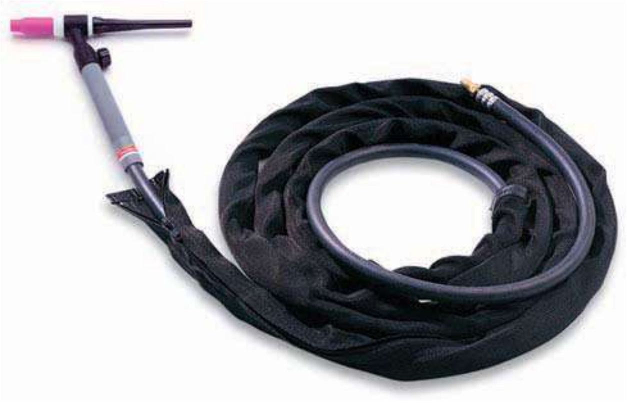

Figure 4.23
Zip-on protective covering also helps keep the hoses neat
Courtesy of ESAB Welding & Cutting Products

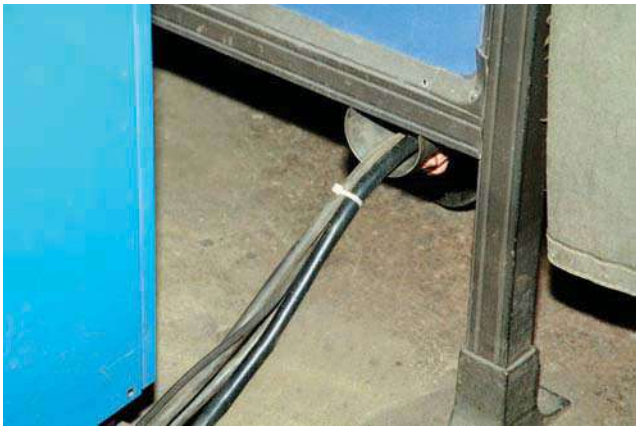

Figure 4.24
A bracket holds the leads off the floor
Courtesy of Larry Jeffus

ture also prevents condensation in the torch. If a water leak should occur during welding, the welding power is stopped, closing the water solenoid and thus stopping the leak.

A protective covering can be used to prevent the hoses from becoming damaged by hot metal, **Figure 4.23.** Even with this protection, the hoses should be supported, **Figure 4.24,** so that they are not underfoot on the floor. By supporting the hoses, the chance of their being damaged by hot sparks is reduced.

Nozzles

The nozzle, or cup, is used to direct the shielding gas directly on the welding zone. The nozzle size is determined by the diameter of the opening and its length, **Table 4.2.** Nozzles may be made from ceramic, such as alumina or silicon nitride (opaque), or from fused quartz (clear). The nozzle may also have a gas lens to improve the gas flow pattern.

The nozzle size, both length and diameter, is often the welder's personal preference. Occasionally, a specific choice must be made based upon joint design or location. Small nozzle diameters allow the welder to better see the molten weld pool and can be operated with lower gas flow rates. Larger nozzle diameters can give better gas coverage, even in drafty places.

Ceramic nozzles are heat resistant and offer a relatively long life. The useful life of a ceramic nozzle is affected by the current level and prox-

Table 4.2 **Recommended Cup Sizes**

Tungsten Electrode Diameter		Nozzle Orifice Diameter	
in.	**(mm)**	**in.**	**(mm)**
1/16	(2)	1/4 to 3/8	(6 to 10)
3/32	(2.4)	3/8 to 7/16	(10 to 11)
1/8	(3)	7/16 to 1/2	(11 to 13)
3/16	(5)	1/2 to 3/4	(13 to 20)

imity to the work. Silicon nitride nozzles will withstand much more heat, resulting in a longer useful life.

The fused quartz (glass) used in a nozzle is a special type that can withstand the welding heat. These nozzles are no more easily broken than ceramic ones but are more expensive. The added visibility with glass nozzles in tight, hard-to-reach places is often worth the added expense.

The longer a nozzle, the longer the tungsten must be extended from the collet. This can cause higher tungsten temperatures, resulting in greater tungsten erosion. When using long nozzles, it is better to use low amperages or a larger-size tungsten.

Flowmeter

The **flowmeter** may be merely a flow regulator used on a manifold system or it may be a combination flow and pressure regulator used on an individual cylinder, **Figure 4.25** and **Figure 4.26.**

The flow is metered or controlled by opening a small valve at the base of the flowmeter. The rate of flow is then read in units of cfh (cubic feet per hour), or L/min (liters per minute). The reading is taken from a fixed scale that is compared to a small ball floating on the stream of gas. Meters from various manufacturers may be read differently. For example, they may read from the top, center, or bottom of the ball, **Figure 4.27.** The ball floats on top of the stream of gas inside a tube that gradually increases in diameter in the upward direction. The increased size allows

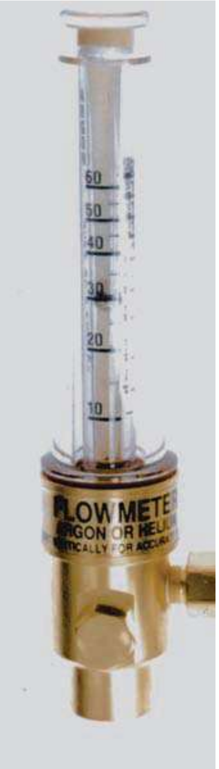

Figure 4.25
Flowmeter
Courtesy of Controls Corporation of America

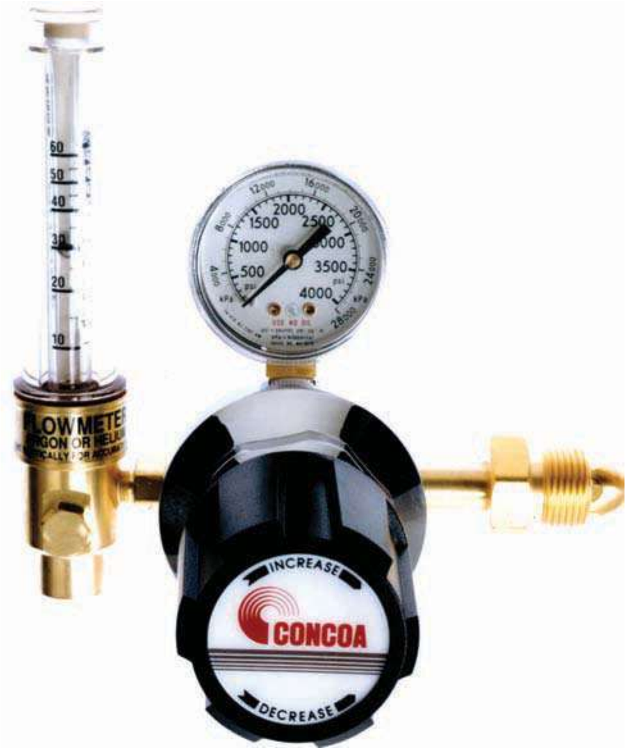

Figure 4.26
Flowmeter regulator
Courtesy of Controls Corporation of America

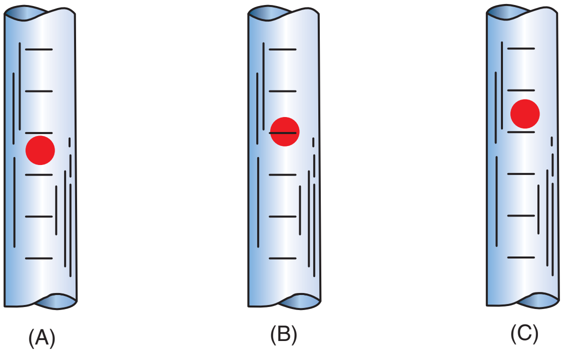

Figure 4.27
Three methods of reading a flowmeter (A) top of ball, (B) center of ball, and (C) bottom of ball

more room for the gas flow to pass by the ball. If the tube is not vertical, the reading is not accurate, but the flow is unchanged. Also, when using a line flowmeter, it is important to have the correct pressure. Changes in pressure will affect the accuracy of the flowmeter reading. In order to get accurate readings, be sure the gas being used is read on the proper flow scale. Less dense gases, such as helium and hydrogen, will not support the ball on as high a column with the same flow rate as a denser gas, such as argon.

TYPES OF WELDING CURRENT

All three types of welding current, or polarities, can be used for GTA welding. Each current has individual features that make it more desirable for specific conditions or with certain types of metals.

The major differences among the currents are in their heat distributions and the presence or degree of arc cleaning. **Figure 4.28** shows the heat distribution for each of the three types of currents.

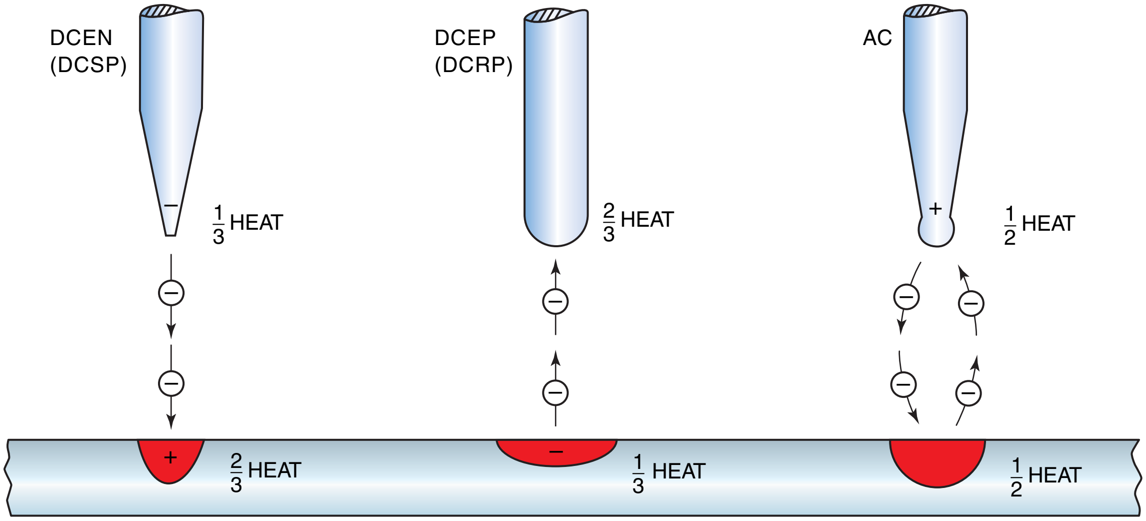

Figure 4.28
Heat distribution between the tungsten electrode and the work with each type of welding current

Direct current electrode negative (DCEN), which used to be called direct current straight polarity (DCSP), concentrates about two-thirds of its welding heat on the work and the remaining one-third on the tungsten. The higher heat input to the weld results in deep penetration. The low heat input into the tungsten means that a smaller-size tungsten can be used without erosion problems.

Direct current electrode positive (DCEP), which used to be called direct current reverse polarity (DCRP), concentrates only one-third of the arc heat on the plate and two-thirds of the heat on the electrode. This type of current produces wide welds with shallow penetration, but it has a strong cleaning action upon the base metal. The high heat input to the tungsten indicates that a large-size tungsten is required, and the end shape with a ball must be used. The low heat input to the metal and the strong cleaning action on the metal make this a good current choice for thin, heavily oxidized metals and magnesium. The metal being welded will not emit electrons as freely as does tungsten, so the arc may wander or be more erratic than DCEN. Because of the near molten state of the electrode tip during DCEP operations, it is almost always done in the flat position.

There are many theories as to why DCEP has a **cleaning action.** The most probable explanation is that the electrons accelerated from the cathode surface lift the oxides that interfere with their movement. The positive ions accelerated to the metal's surface provide additional energy. In combination, the electrons and ions cause the surface erosion needed to produce the cleaning. Although this theory is disputed, it is important to note that cleaning does occur, that it requires argon-rich shield gases and DCEP polarity, and that it can be used to advantage, **Figure 4.29.**

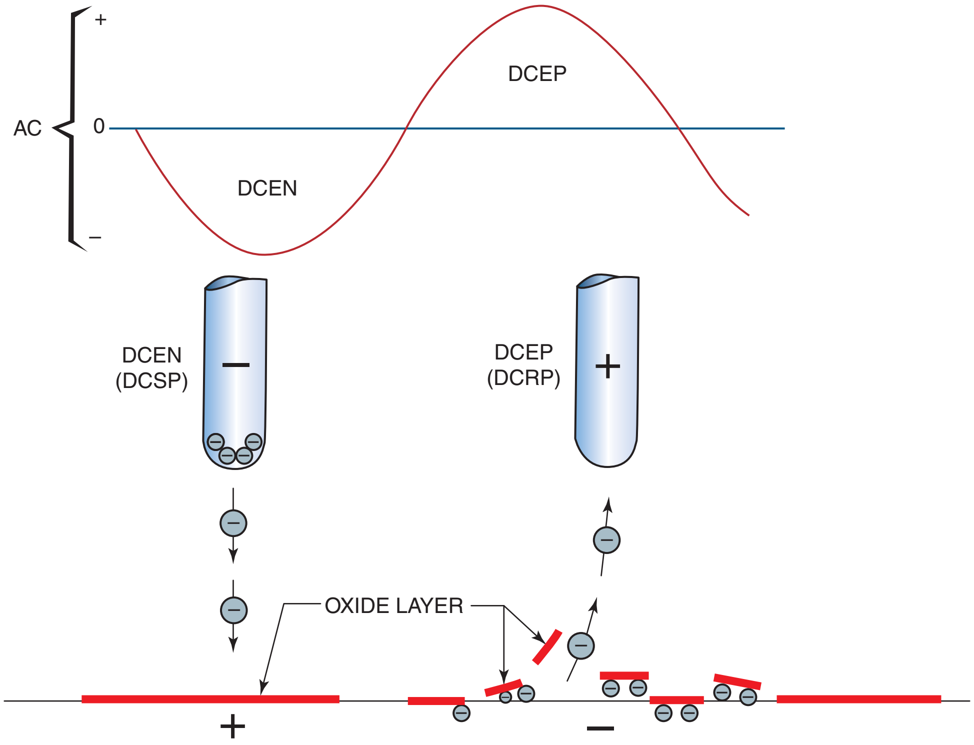

Figure 4.29
Electrons collect under the oxide layer during the DCEP portion of the cycle and lift the oxides from the surface

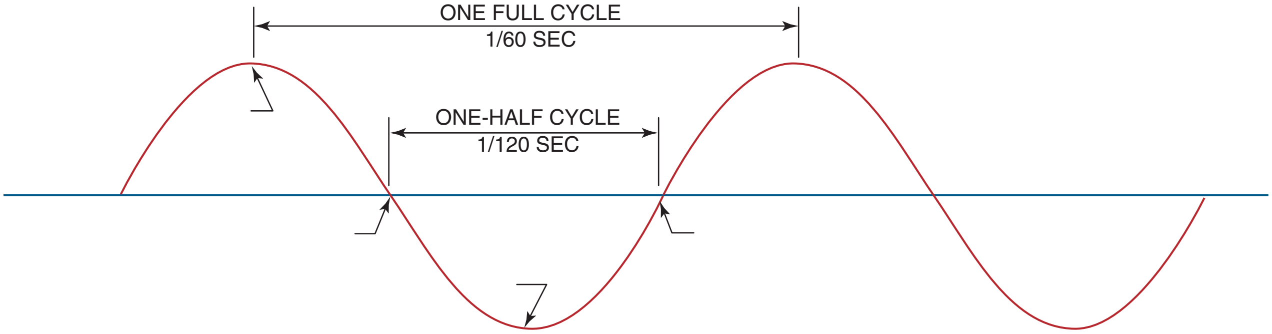

Figure 4.30
Sine wave of alternating current at 60 cycle

Alternating current (AC) concentrates about half of its heat on the work and the other half on the tungsten. Alternating current is DCEN half of the time and DCEP the other half of the time. The **frequency** at which the current cycles is the rate at which it makes a full change in direction, **Figure 4.30.** In the United States, the electric current cycles at the rate of 60 times per second, or 60 hertz (60 Hz). Referring again to Figure 4.30, the current is at its maximum peak at points A and B. The rate gradually decreases until it stops at points C and D. The arc at these points is extinguished and, as the current reversal begins, must be reestablished. This event requires the emission of electrons from the cathode to ionize the shielding gas. When the hot, emissive electrode becomes the cathode, reestablishing the arc is easy. However, it is often quite difficult to reestablish the arc when the colder and less emissive workpiece becomes the cathode. Because voltage from the power supply is designed to support a relatively low voltage arc, it may be insufficient to initiate electron flow. When the arc does not reignite consistently, it becomes destabilized and can cause poor welding performance. This phenomenon is called **rectification.** Thus, a voltage assist from another source is needed. A high-voltage but low-current **spark gap oscillator** commonly provides the assist at a relatively low cost. The high frequency ensures that a voltage peak will occur reasonably close to the current reversal in the welding arc, creating a low-resistance ionized path for the welding current to follow, **Figure 4.31A** and **Figure 4.31B.** This same device is often used to initiate direct current arcs, a particularly useful technique for mechanized welding.

The high-frequency current is established by capacitors discharging across a gap set on points inside the machine. Changing the point gap setting will change the frequency of the current. The closer the points are, the higher the frequency; the wider the spacing between the points, the lower the frequency. The voltage is stepped up with a transformer from the primary voltage supplied to the machine. The available amperage to the high-frequency circuit is very low. Thus, when the circuit is complete, the voltage quickly drops to a safe level. The high frequency is induced on the primary welding current in a coil.

The high frequency may be set so that it automatically cuts off after the arc is established, when welding with DC. It is kept on continuously with AC and transformer rectifier power supplies. When used in this manner, it is referred to as alternating current, high-frequency stabilized, or ACHF. Most of the newer inverter power supplies have advanced circuitry that

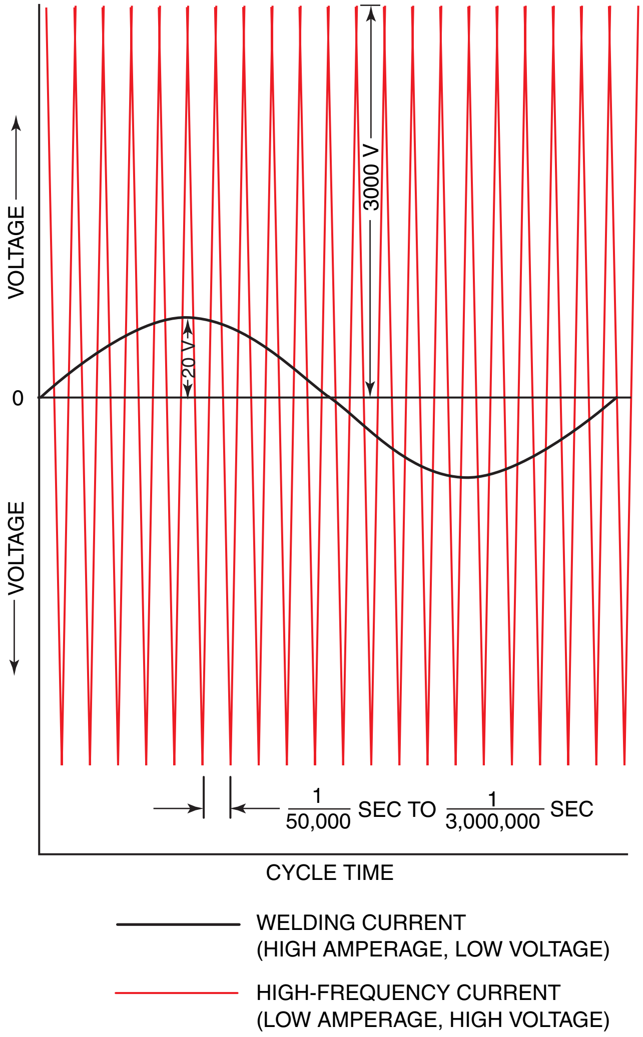

Figure 4.31A
High-frequency arc starting current shown over the low-frequency welding current

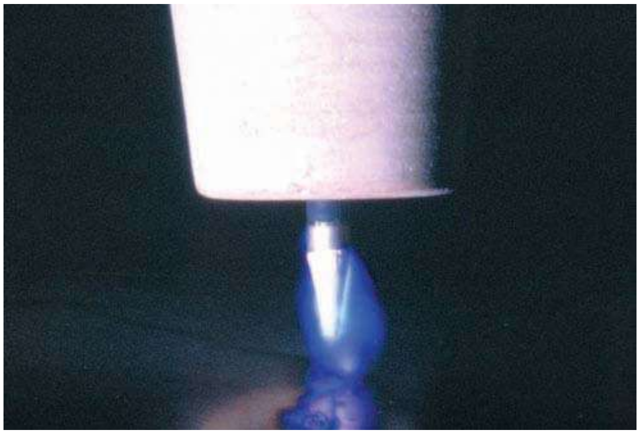

Figure 4.31B
The high frequency first appears as a blue glow around the tungsten before the welding current starts its arc
Courtesy of Larry Jeffus

switches between DCEN and DCEP so quickly that rectification cannot occur and high frequency is not used.

AC BALANCE CONTROL

While older transformer rectifier welding power supplies can only produce AC power in the form of a sine wave (Figure 4.30), newer transformer rectifiers and inverter power supplies can provide enhanced AC output by controlling the dwell time spent on each side of the AC cycle.

The two additional AC waveforms are called square wave and advanced square wave. **Figure 4.32** shows a comparison of the three waveforms. The waveforms represent how the current is measured in amplitude and time as the current moves to a high point, changes direction and polarity, and crosses over the zero point while moving to the maximum amperage on the other side. When the current has moved up and down, hitting both maximum amperages, this represents one cycle. AC electrical power in the United States is delivered at 60 cycles per second.

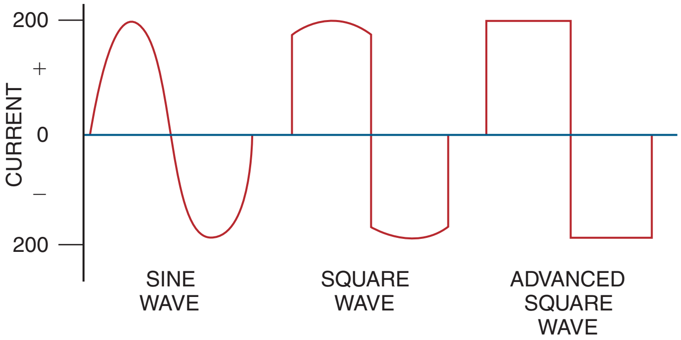

Figure 4.32
Comparison of the three different AC waveforms all representing a time balanced condition and operating at 200 amperes

Welding machines with square wave capabilities have a knob or digital readout that controls the amount of time the current spends on the EN and EP half of the AC cycle. If the machine is adjusted to have more dwell time spent on the EN side, the result will be deeper penetration and reduced cleaning action. The dial usually says "max penetration," **Figure 4.33.** An additional benefit of an AC arc balanced to favor EN is that the extra current delivered to the work will reduce the heat at the tip of the tungsten, which may allow a smaller diameter electrode, or even a pointed end, to be used.

If the machine is adjusted to have more dwell time spent on the EP side of the half cycle, the result will be a wider, shallower bead profile with greater cleaning action. A machine setting favoring more dwell time on EP (called "max cleaning") will be a benefit when it is impractical or impossible to completely remove small amounts of surface contaminants or when welding on aluminum castings. The extra EP will help to etch away more or thicker oxides. When balance controls are set toward "max cleaning," a balled end will be required on the electrode, and if the ball becomes larger than 1-1/2 times the electrode diameter, the next larger size should be selected (Figure 4.14).

Note

Increased EP balance is not a substitute for excellent surface prep, **Figure 4.34.**

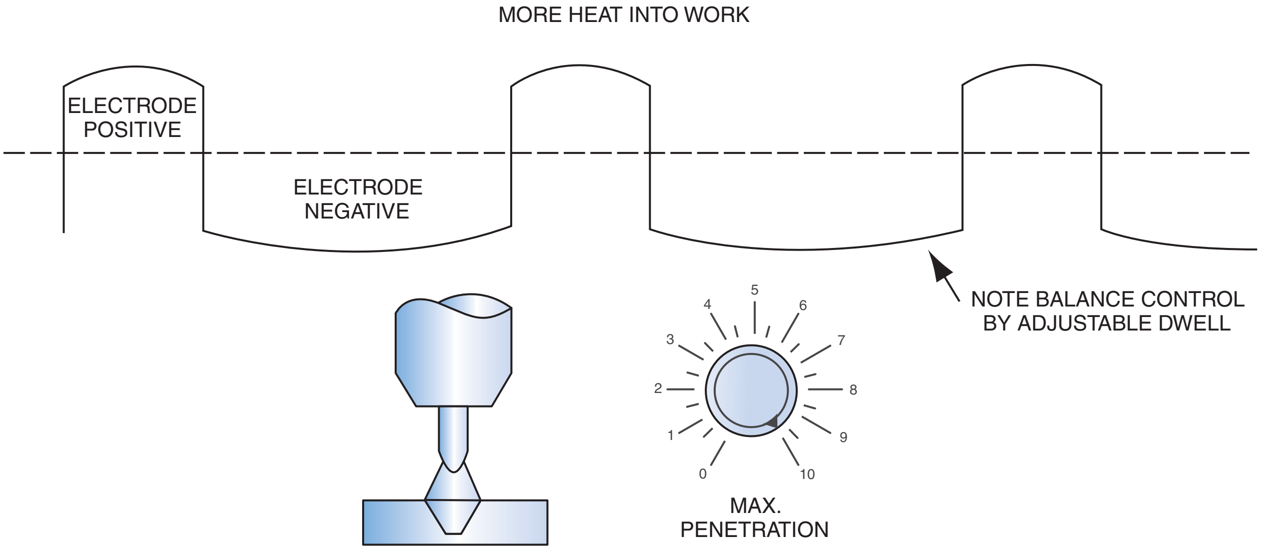

Figure 4.33
Maximum penetration balance control setting

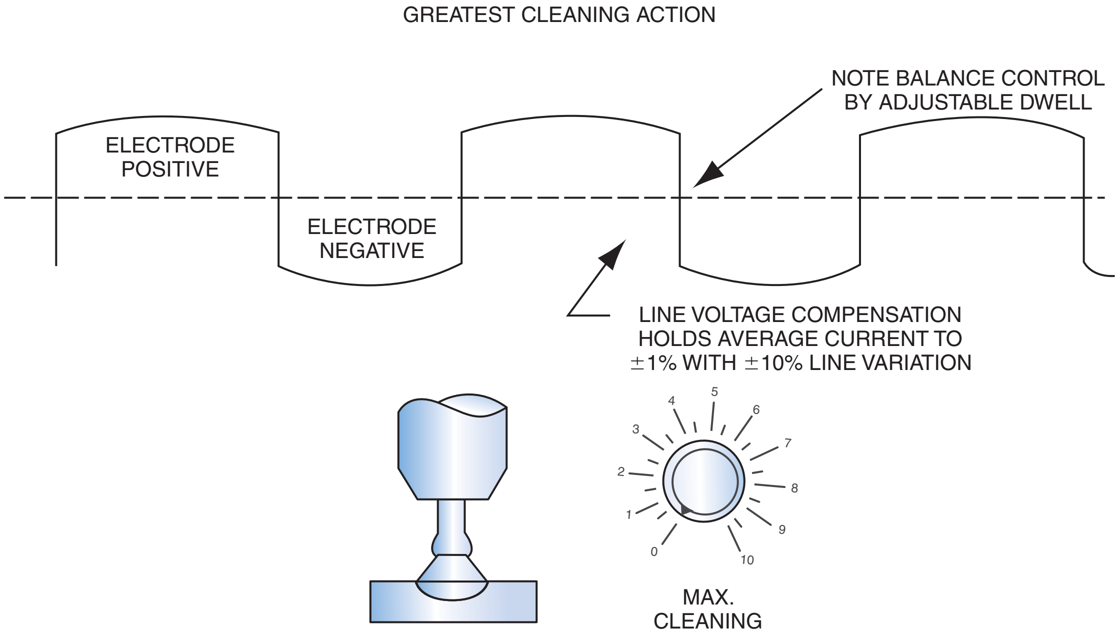
GREATEST CLEANING ACTION
NOTE BALANCE CONTROL BY ADJUSTABLE DWELL
ELECTRODE POSITIVE
ELECTRODE NEGATIVE
LINE VOLTAGE COMPENSATION HOLDS AVERAGE CURRENT TO ±1% WITH ±10% LINE VARIATION
0
1
2
3
4
5
6
7
8
9
10
MAX. CLEANING

Figure 4.34
Maximum cleaning control setting

When the control knob is set in-between max cleaning and max penetration, it is called a balanced waveform. This is a good place to start for general AC applications, **Figure 4.35.** Balance control can be increased toward max EN if greater penetration or a pointed electrode is required; however, operators must be careful not to turn the balance control too far toward EN. When this happens, the oxides on top of the aluminum will not be completely removed and poor fusion or porosity may result. It is also important to note that transformer rectifiers, when set at max penetration, have 68% dwell time on EN, while advanced square wave inverter power supplies may produce as much as 99% dwell time on EN. In essence, this

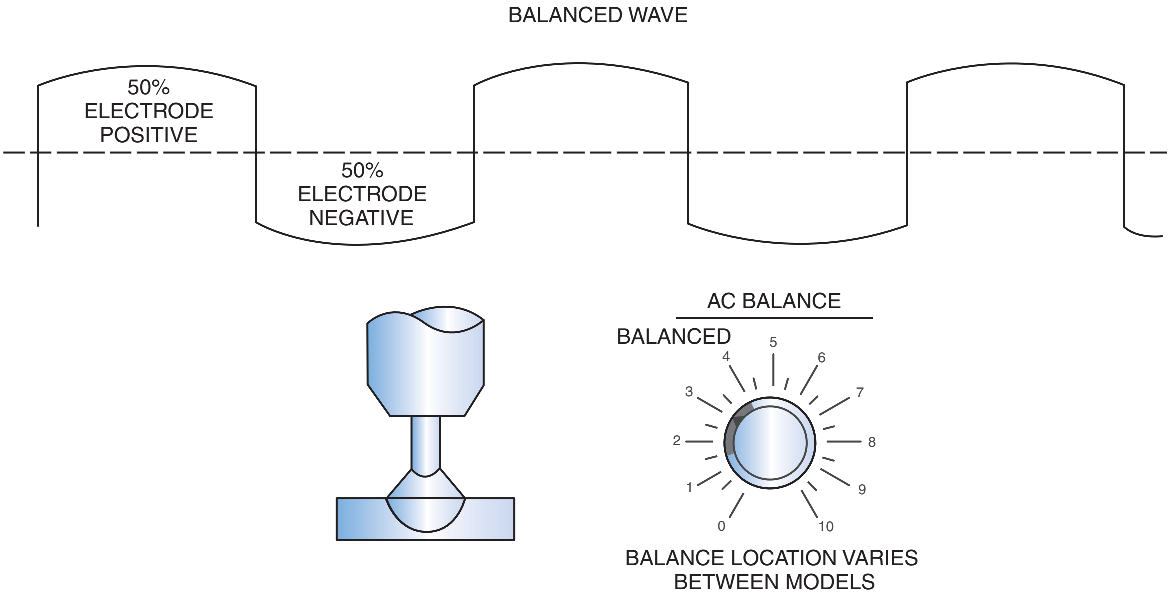
BALANCED WAVE
50% ELECTRODE POSITIVE
50% ELECTRODE NEGATIVE
AC BALANCE
BALANCED
0
1
2
3
4
5
6
7
8
9
10
BALANCE LOCATION VARIES BETWEEN MODELS

Figure 4.35
Balanced control setting

means that when a transformer square wave is set at 10 on its balance control dial, it delivers about the same amount of EN as when an advanced square wave inverter is set to 7.

If you have a square wave power supply, try adjusting the balance control toward max cleaning and max penetration when working on aluminum practices, and note the results in etching at the weld toes and the overall bead profile.

CAUTION

Never allow noninert gases, such as O_2, CO_2, or N, to come in contact with your inert gas system. Very small amounts can contaminate the inert gas, which may result in the weld failing.

SHIELDING GASES

The shielding gases used for the GTA welding process are argon (Ar), helium (He), hydrogen (H), nitrogen (N), or a mixture of two or more of these gases. The purpose of the shielding gas is to protect the molten weld pool and the tungsten electrode from the harmful effects of air. The shielding gas also affects the amount of heat produced by the arc and the resulting weld bead appearance.

Argon and helium are **inert gases.** This means that they will not combine chemically with any other material. Argon and helium may be found in mixtures but never as compounds. Because they are inert, they will not affect the molten weld pool in any way.

Argon

Argon is a by-product in air separation plants. Air is cooled to temperatures that cause it to liquefy; then its constituents are fractionally distilled. The primary products are oxygen and nitrogen. Before these gases were produced on a tonnage scale, argon was a rare gas. Now it is distributed in cylinders as gas or in bulk as a liquid.

Because argon is denser than air, it effectively shields welds in deep grooves in the flat position. However, this higher density can be a hindrance when welding overhead because higher flow rates are necessary. The argon is relatively easy to ionize and thus suitable for alternating current applications and easier starts. This property also permits fairly long arcs at lower voltages, making it virtually insensitive to changes in arc length. Argon is also the only commercial gas that produces the cleaning discussed earlier. These characteristics are most useful for manual welding, especially with filler metals added, as shown in **Figure 4.36.**

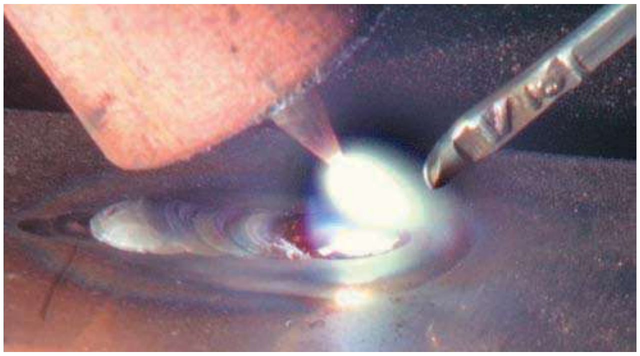

Figure 4.36
Highly concentrated ionized argon gas column
Courtesy of Larry Jeffus

Helium

Helium is a by-product of the natural gas industry. It is removed from natural gas as the gas undergoes separation (fractionation) for purification or refinement.

Helium offers the advantage of deeper penetration. The arc force with helium is sufficient to displace the molten weld pool with very short arcs. In some mechanized applications, the tip of the tungsten electrode is positioned below the workpiece surface to obtain very deep and narrow penetration. This technique is especially effective for welding aged aluminum alloys prone to overaging. It is also very effective at high welding speeds, as for tube mills. However, helium is less forgiving for manual welding. With helium, penetration and bead profile are sensitive to arc length, and the long arcs that are needed for feeding filler wires are more difficult to control.

Helium has been mixed with argon to gain the combined benefits of cathode cleaning and deeper penetration, particularly for manual welding. The most common of these mixtures is 75% helium and 25% argon.

Although the GTA process was developed with helium as the shielding gas, argon is now used whenever possible because it is much cheaper. Helium also has some disadvantages. Because it is lighter than air, helium doesn’t allow for good shielding. Its flow rates must be about twice as high as argon’s for acceptable stiffness in the gas stream, and proper protection is difficult in drafts unless high flow rates are used. It is difficult to ionize, necessitating higher voltages to support the arc and making the arc more difficult to ignite. Alternating current arcs are very unstable. However, helium is not used with alternating current because the cleaning action does not occur.

Hydrogen

Hydrogen is not an inert gas and is not used as a primary shielding gas. However, it can be added to argon in small amounts (1–3%) when deep penetration and high welding speeds are needed. It also improves the weld surface cleanliness and bead profile on some grades of stainless steel that are very sensitive to oxygen. Hydrogen additions are restricted to stainless steel because hydrogen is the primary cause of porosity in aluminum welds. It can cause porosity in carbon steel and, in highly restrained welds, underbead cracking in carbon and low alloy steel.

Nitrogen

Nitrogen is not an inert gas. Like hydrogen, nitrogen has been used as an additive to argon. But it cannot be used with some materials, such as ferritic steel, because it produces porosity. In other cases, such as with austenitic stainless steel, nitrogen is useful as an austenite stabilizer in the alloy. It is used to increase penetration when welding copper. Nitrogen is also sometimes used as a backing gas on some austenitic stainless steel pipe and tubing because of its low relative cost. When using nitrogen to protect the back side of a weld, care must be taken to be sure that all fitups are tight and that no nitrogen mixes with the shield gas at the top of the weld. Unfortunately, because of the general success with inert gas mix-

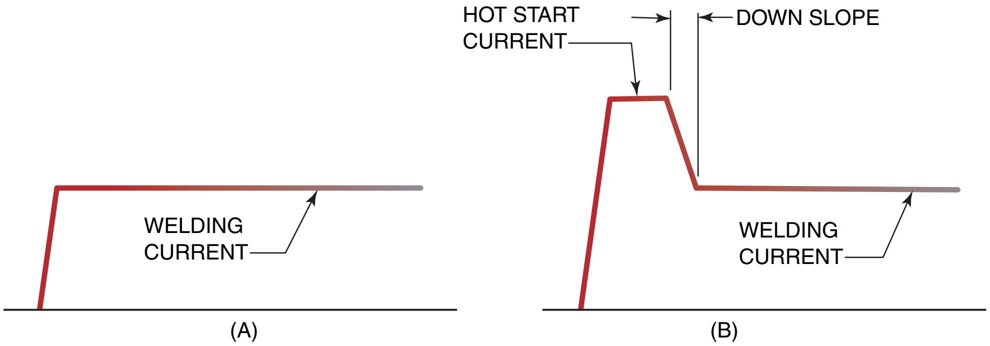

Figure 4.37
Standard method of starting welding current (A); hot start method of starting welding current (B)

tures and because of potential metallurgical problems, nitrogen has not received much attention as an additive for GTA welding.

Hot Start

The hot start allows a controlled surge of welding current as the arc is started to quickly establish a molten weld pool. Rapidly establishing a molten weld pool on metals with a high thermal conductivity is often hard without this higher than normal current. Adjustments can be made in the length of time and the percentage above the normal current, **Figure 4.37.**

Preflow Time and Postflow Time

Preflow time is the time during which gas flows to clear out any air in the nozzle or surrounding the weld zone. The operator sets the length of time that the gas flows before the welding current is started, **Figure 4.38.** Because some machines do not have preflow, many welders find it hard to hold a position while waiting for the current to start. One solution to this problem is to use the postflow for preflow. Switch on the current to engage the postflow. Now, with the current off, the gas is flowing, and the GTA torch can be lowered to the welding position. The welder's helmet should be lowered and the current restarted before the postflow stops. This allows welders to have preflow of shield gas and to start the arc when they are ready.

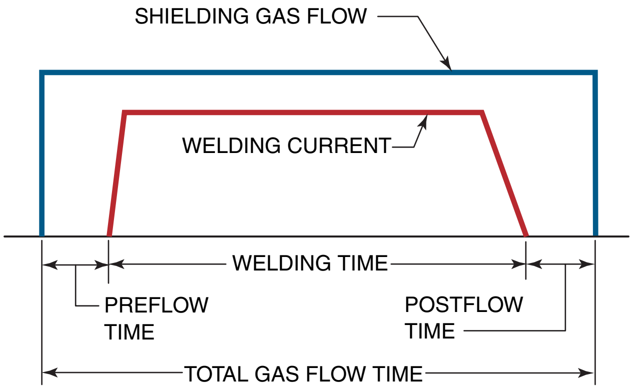

Figure 4.38
Welding time compared to shielding gas flow time

Table 4.3 **Postwelding Gas Flow Times**

Electrode Diameter in.	(mm)	Postwelding Gas Flow Time*
.01	(0.25)	5 sec
0.02	(0.5)	5 sec
0.04	(1)	5 sec
1/16	(2)	8 sec
3/32	(2.4)	10 sec
1/8	(3)	15 sec
5/32	(4)	20 sec
3/16	(5)	25 sec
1/4	(6)	30 sec

*The time may be longer if either the base metal or the tungsten electrode does not cool below the rapid oxidation temperatures within the postflow times shown.

The **postflow time** is the time during which the gas continues flowing after the welding current has stopped. This period serves to protect the molten weld pool, the filler rod, and the tungsten electrode as they cool to a temperature at which they will not oxidize rapidly. The time of the flow is determined by the welding current and the tungsten size, **Table 4.3.**

Gas Flow Rate

The shielding gas flow rate is measured in cubic feet per hour (cfh) or in metric measure as liters per minute (L/min). The rate of flow should be as low as possible and still give adequate coverage. High gas flow rates waste shielding gases and may lead to contamination. The contamination comes from turbulence in the gas at high flow rates. Air is drawn into the gas envelope by a venturi effect around the edge of the nozzle. Also, the air can be drawn in under the nozzle if the torch is held at too sharp an angle to the metal, **Figure 4.39.**

The larger the nozzle size, the higher is the flow rate permissible without causing turbulence. **Table 4.4** shows the average and maximum flow rates for most nozzle sizes. A gas lens can be used in combination with the nozzle to stabilize the gas flow, thus eliminating some turbulence. A gas lens will add to the turbulence problem if there is any spatter or contamination on its surface.

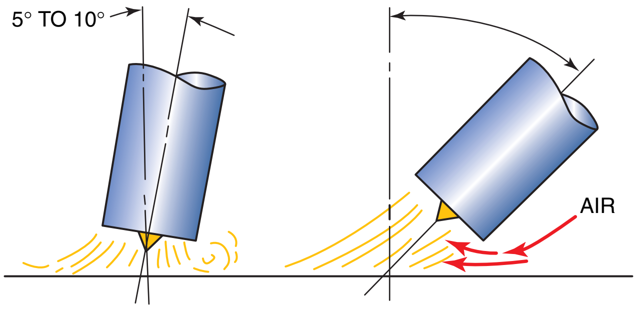

Figure 4.39
Too steep an angle between the torch and work may draw in air

Table 4.4 Suggested Argon Gas Flow Rate for Given Cup Sizes

Nozzle Inside Diameter		Gas Flow*	
in.	(mm)	cfh	(L/min)
1/4	(6)	10–14	(4.7–6.6)
5/16	(8)	11–15	(5.2–7.0)
3/8	(10)	12–16	(5.6–7.5)
7/16	(11)	13–17	(6.1–8.0)
1/2	(13)	17–20	(8.0–9.4)
5/8	(16)	17–20	(8.0–9.4)

*The flow rates may need to be increased or decreased depending upon the conditions under which the weld is to be performed.

REMOTE CONTROLS

A remote control can be used to start the weld, increase or decrease the current, and stop the weld. The remote control can be either a foot-operated or hand-operated device. The foot control works adequately if the welder can be seated. Welds that must be performed away from a welding station may use a hand or thumb control, or may not have any remote welding controls.

Most remote controls have an on-off switch that is activated at the first or last part of the control movement. A variable resistor increases the current as the control is pressed more. A variable resistor works in a manner similar to the accelerator pedal on a car to increase the power (current), **Figure 4.40.** The operating amperage range is determined by the value that has been set on the main controls of the machine.

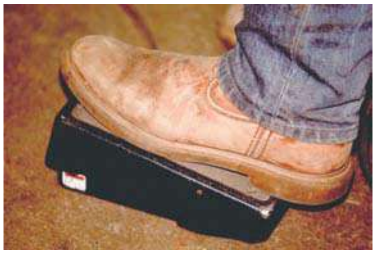

Figure 4.40
A foot-operated device can be used to increase the current
Courtesy of Larry Jeffus

EXPERIMENT 4-4

Setting Up a GTA Welder

Using a GTA welding machine; remote control welding torch; gas flowmeter; gas source (cylinder or manifold); tungsten; nozzle; collet; collet body; cap; and any other hoses, special tools, and required equipment, you will set up the machine for GTA welding, **Figure 4.41.**

Module 1
Key Indicator 1, 4

Module 7
Carbon Steel
Key Indicator 3
Austenitic Stainless Steel
Key Indicator 8
Aluminum
Key Indicator 13

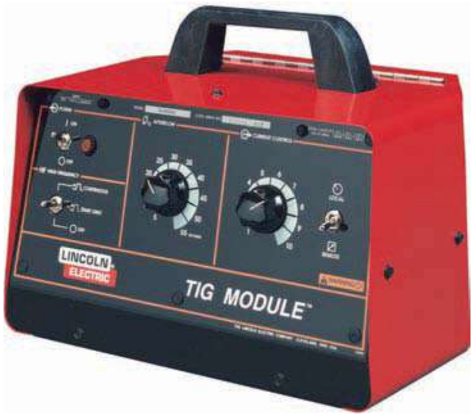

Figure 4.41
GTA welding unit that can be added to a standard power supply so that it can be used for GTA welding
Courtesy of Lincoln Electric Company

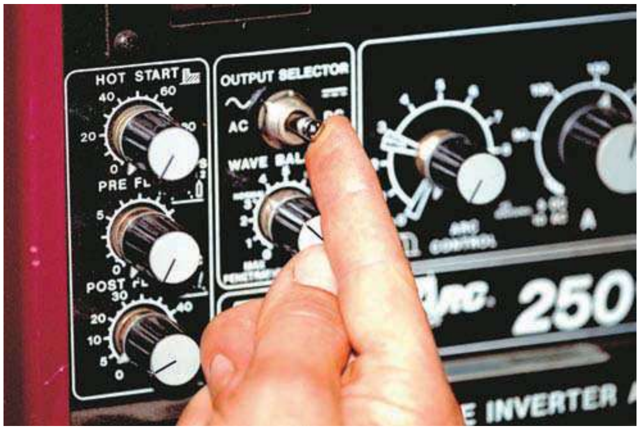

Figure 4.42
Always be sure the power is off when making machine connections
Courtesy of Larry Jeffus

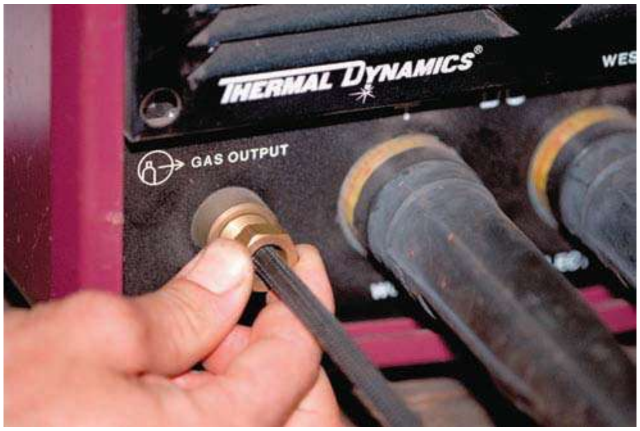

Figure 4.43
Tighten each fitting as it is connected to avoid missing a connection
Courtesy of Larry Jeffus

1. Start with the power switch off, **Figure 4.42.** Use a wrench to attach the torch hose to the machine. The water hoses should have left-hand threads to prevent incorrectly connecting them. Tighten the fittings only as tightly as needed to prevent leaks, **Figure 4.43.** Attach the cooling water "in" to the machine solenoid and the water "out" to the power block.
2. The flowmeter or flowmeter regulator should be attached next. If a gas cylinder is used, secure it in place with a safety chain. Then remove the valve protection cap and crack the valve to blow out any dirt, **Figure 4.44.** Attach the flowmeter so that the tube is vertical.
3. Connect the gas hose from the meter to the gas "in" connection on the machine.
4. With both the machine and main power switched off, turn on the water and gas so that the connection to the machine can be checked for leaks. Tighten any leaking fittings to stop the leakages.

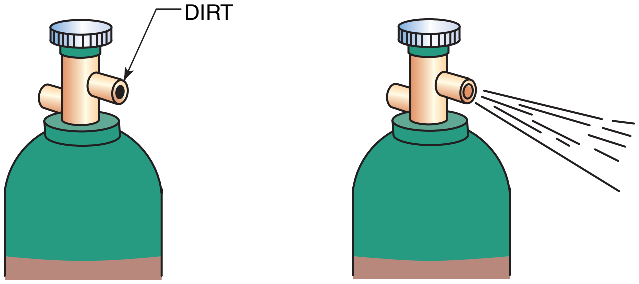

Figure 4.44
During transportation or storage, dirt may collect in the valve. Cracking the valve is the best way to remove any dirt.

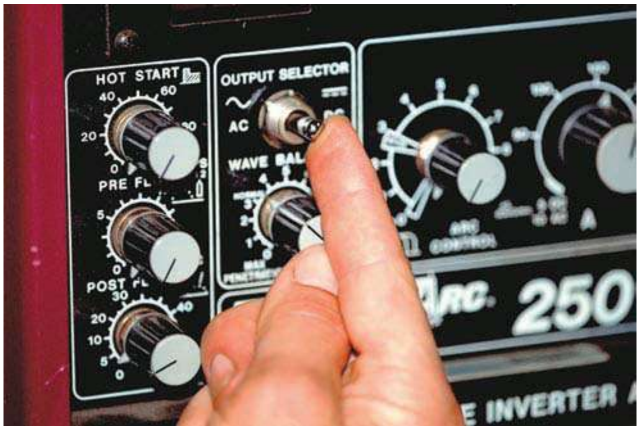

Figure 4.45
Setting the current
Courtesy of Larry Jeffus

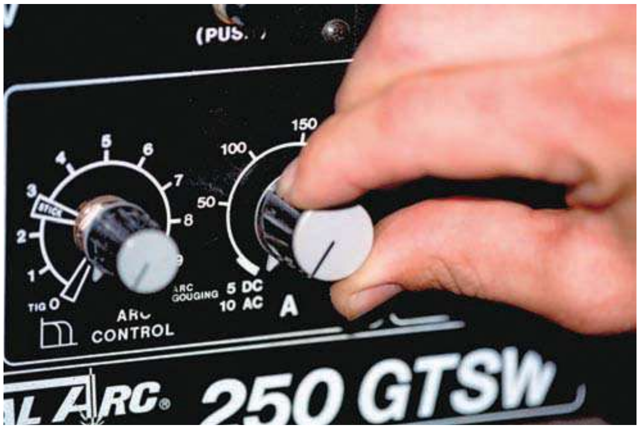

Figure 4.46
Setting the amperage range
Courtesy of Larry Jeffus

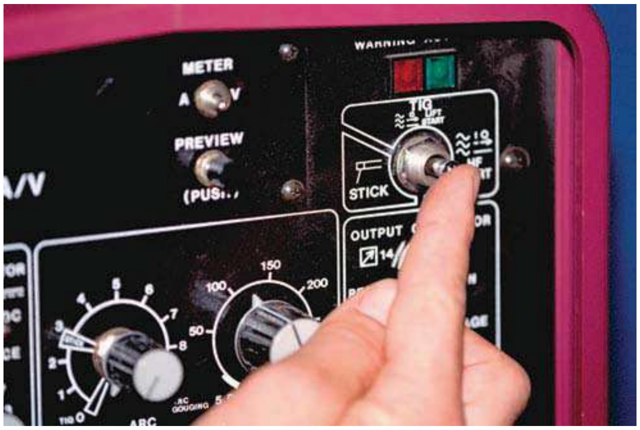

Figure 4.47
The high-frequency switch should be placed in the appropriate position
Courtesy of Larry Jeffus

5. Turn on both the machine and main power switches and watch for leaks in the torch hoses and fittings.
6. With the power off, switch the machine to the GTA welding mode.
7. Select the desired type of current and amperage range, **Figure 4.45** and **Figure 4.46.**
8. Set the fine current adjustment to the proper range, depending upon the size of tungsten used, **Table 4.5.**
9. Place the high-frequency switch in the appropriate position, auto (HF start) for DC or continuous for AC, **Figure 4.47.**
10. The remote control can be plugged in and the selector switch set, **Figure 4.48.**
11. The collet and collet body should be installed on the torch first, **Figure 4.49.**
12. On the Linde brand or copies of Linde torches, installing the back cap first will stop the collet body from being screwed into the torch fully. A poor connection will result in excessive electrical and thermal resistance, causing a heat buildup in the head.
13. The tungsten can be installed and the end cap tightened to hold the tungsten in place. Select and install the desired nozzle size. Adjust the tungsten length so that it does not stick out more than the diameter of the nozzle, **Figure 4.50.**
14. Check the manufacturer's operating manual for the machine to ensure that all connections and settings are correct.

Table 4.5 Amperage Range of Tungsten Electrodes

Electrode Diameter		DCEN	DCEP	AC
in.	(mm)			
0.04	(1)	15–60	Not recommended	10–50
1/16	(2)	70–100	10–20	50–90
3/32	(2.4)	90–200	15–30	80–130
1/8	(3)	150–350	25–40	100–200
5/32	(4)	300–450	40–55	160–300

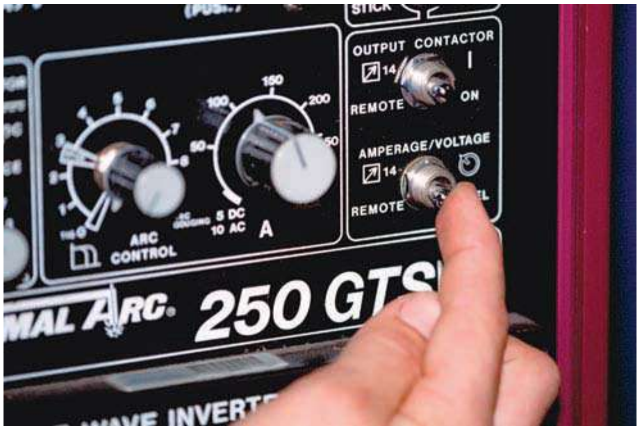

Figure 4.48
Setting the remote control switch
Courtesy of Larry Jeffus

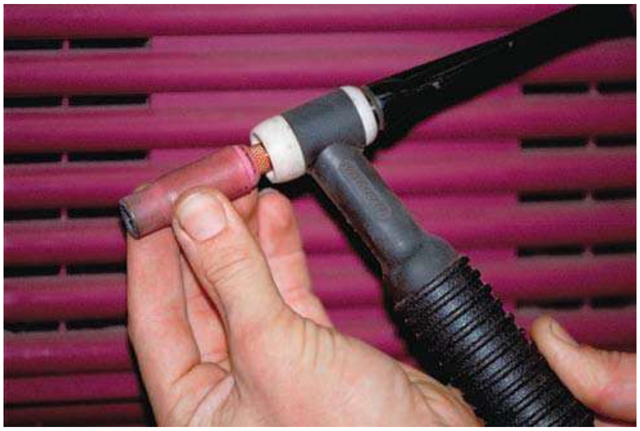

Figure 4.49
Inserting collet and collet body
Courtesy of Larry Jeffus

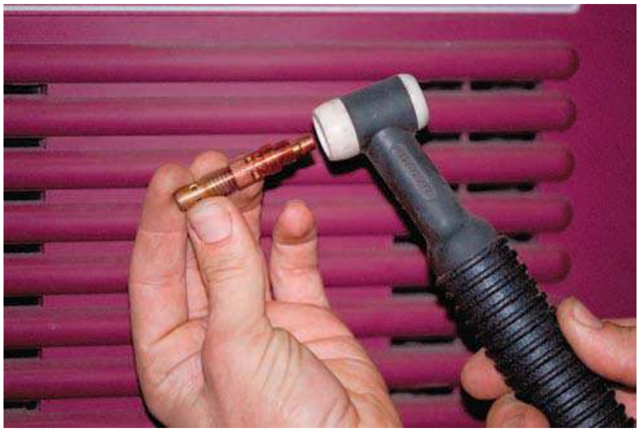

Figure 4.50
Install the nozzle (cup) to the torch body
Courtesy of Larry Jeffus

CAUTION

Turn off all power before attempting to stop any leaks in the water system.

The GTA welding system is now ready to be used.

15. Turn on the power, depress the remote control, and again check for leaks.
16. While the postflow is still engaged, set the gas flow by adjusting the valve on the flowmeter.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 4

Module 7
Key Indicator 1
Carbon Steel
Key Indicator 3, 4
Austenitic Stainless Steel
Key Indicator 8, 9
Aluminum
Key Indicator 13, 14

EXPERIMENT 4-5

Striking an Arc

Using a properly set up GTA welding machine, proper safety gear, and clean scrap metal, you will strike a GTA welding arc.

1. Position yourself so that you are comfortable and can see the torch, tungsten, and plate while the tungsten tip is held about 1/4 in. (6 mm) above the metal. Try to hold the torch at a vertical angle ranging from 0° to 15°. Too steep an angle will not give adequate gas coverage, **Figure 4.51.**

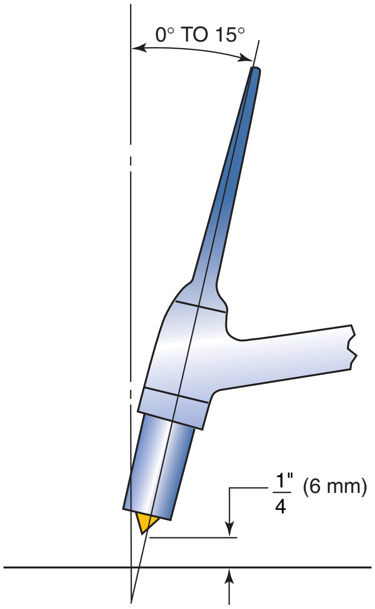

Figure 4.51
GTA torch position

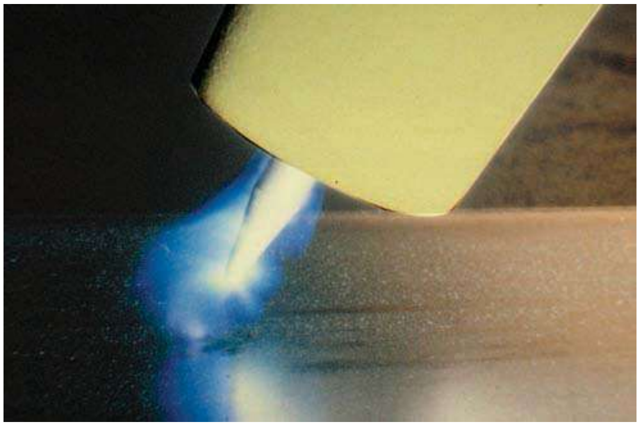

Figure 4.52
High frequency starting before arc starts
Courtesy of Larry Jeffus

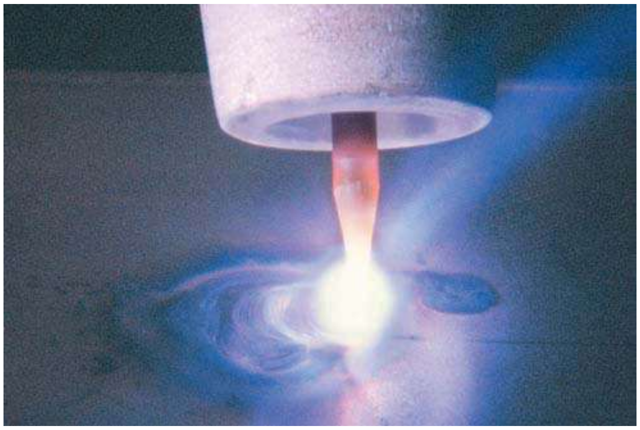

Figure 4.53
Stable gas tungsten arc
Courtesy of Larry Jeffus

2. Lower your arc welding helmet and depress the remote control. A high-pitched, erratic arc should be immediately jumping across the gap between the tungsten and the plate. If the high-frequency arc is not established, lower the torch until it appears, **Figure 4.52.**
3. Slowly increase the current until the main welding arc appears, **Figure 4.53.**
4. Observe the color change of the tungsten as the arc appears.
5. Move the tungsten around in a small circle until a molten weld pool appears on the metal.
6. Slowly decrease the current and observe the change in the molten weld pool.
7. Reduce the current until the arc is extinguished.
8. Hold the torch in place over the weld until the postflow stops.
9. Raise your hood and inspect the weld.

CAUTION

Avoid touching the metal table with any unprotected skin or jewelry. The high frequency can cause an uncomfortable shock.

Repeat this procedure until you can easily start the arc and establish a molten weld pool using both AC and DCEN currents. Turn off the welding machine, water, and shielding gas when you are finished; then clean up your work area.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

SUMMARY

One of the prime considerations for gas tungsten arc welding equipment setup is the cleanliness of the equipment, supplies, base material or material, and the welders themselves. When everything is clean, you will find that the welding process will proceed more easily and more successfully.

Another major factor affecting your ability to produce quality welds is the tungsten end or tip shape. As you practice making the various welds, you will find that keeping the tungsten electrode tip shaped appropriately will assist you in producing uniform welds.

Often, new welders feel that there is some sort of attraction between the tungsten electrode, filler metal, and base metal during the welding process because it seems to continually become contaminated. This almost continuous contamination can be very frustrating. At times it may seem overwhelming; however, with continued practice and diligence, you will be able to control this problem. Even experienced welders in the field can be plagued from time to time with tungsten contamination. At other times they can weld an entire day without contaminating the tungsten. It is often beneficial for students to realize that tungsten contamination is just part of the process, and they must therefore try to ignore the possibility of it happening and concentrate on producing the welds.

REVIEW

1. What early advancements made the GTA welding process more effective and reduced its cost?
2. What metals were weldable only by the GTAW process before GMAW was developed?
3. Which two of tungsten's properties make it the most versatile choice for GTA welding?
4. Why must the tip of the tungsten be hot?
5. Why does some tungsten erosion occur?
6. What function regarding tungsten heat do the collet and torch play?
7. What problem can an excessively large tungsten cause?
8. What holds the molten ball of tungsten in place at the tip of the electrode during DCEP welding?
9. Using **Table 4.1,** answer the following:
   a. What color identifies EWTh-2?
   b. What is the composition of EWCe-2?
   c. What color identifies EWLa-1.5
10. What does adding thorium oxide do for the tungsten electrode?

11. How can the end of a tungsten electrode be shaped?
12. Why should a grinding stone that is used for sharpening tungsten not be used for other metals?
13. Why should the grinding marks run lengthwise on the tungsten electrode end?
14. What are three ways of breaking off the contaminated end of a tungsten electrode?
15. Why should the torch be as cool as possible?
16. What will happen to a water-cooled torch cable if the flow of cooling water stops?
17. Why must shielding gas hoses not be made from rubber?
18. Why should the water solenoid be on the supply side of the water system?
19. What problem can a long nozzle cause to the tungsten?
20. Why must the tube of a flowmeter be vertical?
21. What is the heat distribution with DCEN welding current?
22. What is the heat distribution with DCEP welding current?
23. What is the heat distribution with AC welding current?
24. Why must AC welding power provided by transformer power supplies use high frequencies in order to work?
25. Why are argon and helium known as inert gases?
26. Why is argon’s ease of ionization a benefit?
27. What makes helium difficult to use for manual welding?
28. What are the benefits of adding hydrogen to argon for welding?
29. What is the purpose of a hot start?
30. Using **Table 4.3,** determine the gas postflow time for a 3/32-in. (2.4-mm) tungsten.
31. How can air be drawn into the shielding gas?
32. Using **Table 4.4,** determine the minimum gas flow rate for a 1/2-in. (13-mm) nozzle.
33. What functions can a remote control provide the welder?

This page intentionally left blank

CHAPTER 5

Gas Tungsten Arc Welding of Plate

OBJECTIVES

After completing this chapter, the student should be able to

- list four applications for which the gas tungsten arc welding process is more commonly used
- describe four effects on the weld of varying torch angles
- explain why the filler rod end must be kept inside the protective zone of the shielding gas and how to accomplish this
- list four disadvantages to welding with a contaminated tungsten electrode
- list four techniques to reduce electrode contamination
- determine the correct machine settings for the minimum and maximum welding current for the machine used, the types and sizes of tungsten, and the metal types and thicknesses
- list four factors that affect the gas preflow and postflow times required to protect the electrode and the weld
- determine the minimum and maximum gas flow settings for each nozzle size, tungsten size, and amperage setting
- compare the characteristics of low-carbon and mild steels, stainless steel, and aluminum with respect to GTA welding
- prepare carbon steel, stainless steel, and aluminum for GTA welding

KEY TERMS

chill plate
contamination
gas coverage
oxide layer
protective zone
surface tension

AWS SENSE EG2.0

Key Indicators Addressed in this Chapter

Module 1: Occupational Orientation

Key Indicator 1: Prepares time or job cards, reports or records
Key Indicator 3: Follows verbal instructions to complete work assignments
Key Indicator 4: Follows written instructions to complete work assignments

Module 2: Safety and Health of Welders

Key Indicator 1: Demonstrates proper use and inspection of personal protection equipment (PPE)
Key Indicator 2: Demonstrates safe operation practices in the work area
Key Indicator 3: Demonstrates proper use and inspection of ventilation equipment
Key Indicator 4: Demonstrates proper hot zone operation
Key Indicator 6: Understands proper use of precautionary labeling and MSDS information.
Key Indicator 7: Demonstrates proper inspection and operation of equipment for each welding or thermal cutting process used

Module 7: Gas Tungsten Arc Welding (GTAW)

Key Indicator 1: Performs Safety inspections of GTAW equipment and accessories
Key Indicator 2: Makes Minor external repairs to GTAW equipment and accessories

Carbon Steel

Key Indicator 3: Sets up for GTAW operations on carbon steel
Key Indicator 4: Operates GTAW equipment on carbon steel
Key Indicator 5: Makes GTAW fillet welds, in all positions, on carbon steel
Key Indicator 6: Makes GTAW groove welds, in all positions, on carbon steel
Key Indicator 7: Passes GTAW welder performance qualification test (workmanship sample) on carbon steel

Austenitic Stainless Steel

Key Indicator 8: Sets up for GTAW operations on austenitic stainless steel
Key Indicator 9: Operates GTAW equipment on austenitic stainless steel
Key Indicator 10: Makes GTAW fillet welds, in 1F, 2F, and 3F positions, on austenitic stainless steel
Key Indicator 11: Makes GTAW groove welds, in 1G and 2G positions, on austenitic stainless steel
Key Indicator 12: Passes GTAW welder performance qualification test (workmanship sample) on austenitic stainless steel

Aluminum

Key Indicator 13: Sets up for GTAW operations on aluminum
Key Indicator 14: Operates GTAW equipment on aluminum
Key Indicator 15: Makes GTAW fillet welds, in the 1F and 2F positions, on aluminum
Key Indicator 16: Makes GTAW groove welds, in the 1G position, on aluminum
Key Indicator 17: Passes GTAW welder performance qualification test (workmanship sample) on aluminum

Module 9: Welding Inspection and Testing

Key Indicator 1: Examines cut surfaces and edges of prepared base metal parts
Key Indicator 2: Examines tacks, root passes, intermediate layers, and completed welds.

INTRODUCTION

The gas tungsten arc welding process can be used to join nearly all types and thicknesses of metal. Welders can have a clear, unobstructed view of the molten weld pool because GTA welding is fluxless, slagless, and smokeless. The clear view of the weld allows welders to make changes in their welding technique, current, travel speed, and rate at which the filler metal is added to the weld as the weld progresses, ensuring that a quality weld is being made. This gives the welder very fine control of the welding process.

The fine control of the weld that is possible with GTA welding makes it an ideal process for very close-tolerance, high-quality welds. GTA welding is used to make critical welds, such as those on aircraft structures. If these welds fail, serious injury, death, and/or significant loss of property can result. Sometimes GTA welding is used to make the critical root pass of a weld that will be completed using another faster process. It is also used when weld appearance is important to the look of the finished part, as in some furniture, decorations, and/or sculptures.

The proper setup of GTA equipment can often affect the quality of the weld performed. Charts and graphs are available that give the correct amperage, gas flow rate, and time for various types of welds and metals. These charts are designed for optimum laboratory or classroom conditions. Actual conditions in the field will have an effect on these values. The experiments in this chapter are designed to help the welder understand the harmful effects on welding of less than ideal conditions. This will allow the welder to evaluate the appearance of a weld and make the necessary changes in technique or setup to improve the weld.

After a person has learned to weld in the lab, troubleshooting field welding problems will become much easier. The weld should be watched carefully to pick up any changes that could indicate a needed adjustment. When welders can do this, they have mastered the GTA process and have made themselves better potential employees. To make a weld is good; to solve a welding problem is better.

TORCH ANGLE

The torch should be held as close to perpendicular as possible in relation to the plate surface. The torch may be angled from 0° to 15° from perpendicular for better visibility and still have the proper shielding **gas coverage.** As the gas flows out it must form a **protective zone** around the weld. Tilting the torch changes the shape of this protective zone, **Figure 5.1.** Too much tilting of the torch will cause the protective shielding gas zone to become so distorted that the weld may not be protected from contamination from the air. The closer the torch is held to perpendicular, the better the weld is shielded.

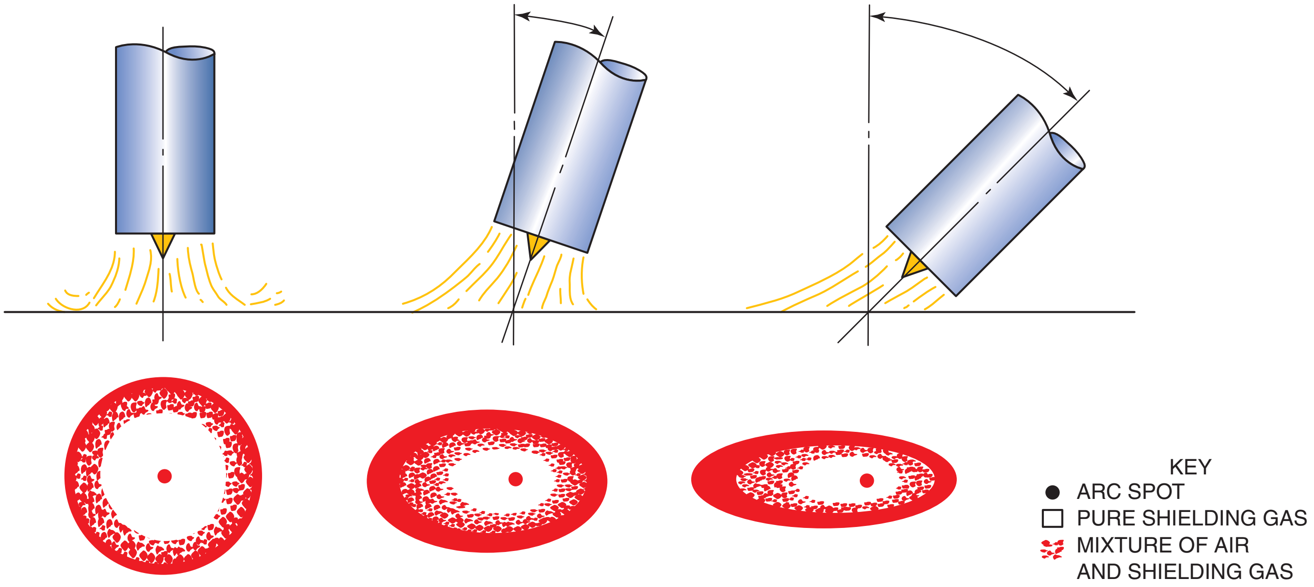

Figure 5.1
Gas coverage patterns for different GTA torch angles. Note how the area covered by the shielding gas becomes narrower and elongates as the angle of the torch increases from the perpendicular.

The velocity of the shielding gas also affects the protective zone as the torch angle changes. As the velocity increases, a low-pressure area develops behind the cup. When the low-pressure area becomes strong enough, air is pulled into the shielding gas. The sharper the angle and the higher the flow rate, the greater the possibility contamination will occur from the onset of turbulence in the gas stream. This causes air to become mixed with the shielding gas. Turbulence caused by the shielding gas striking the work will also cause air to mix with the shielding gas at high velocities.

FILLER ROD MANIPULATION

The filler rod end must be kept inside the protective zone of the shielding gas, **Figure 5.2.** The end of the filler rod is hot, and if it is removed from the gas protection, it will oxidize rapidly. The oxide will then be added to the

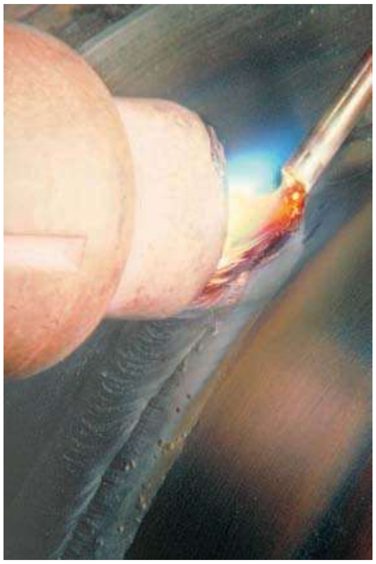

Figure 5.2
The hot filler rod end is well within the protective gas envelope
Courtesy of Larry Jeffus

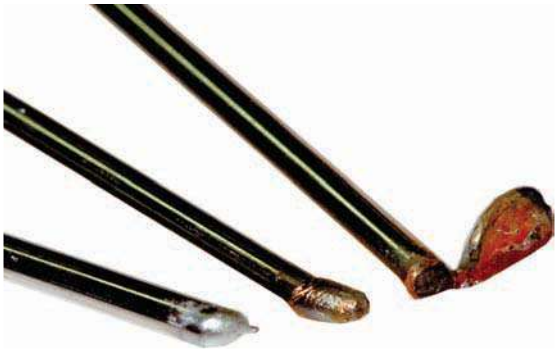

Figure 5.3
(A) Filler properly protected, (B) some oxides on filler, and (C) excessive oxides caused by improper filler rod manipulation
Courtesy of Larry Jeffus

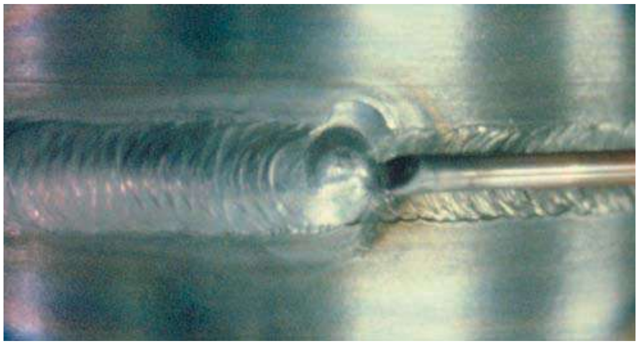

Figure 5.4
Filler being left in the molten weld pool as the arc is extinguished
Courtesy of Larry Jeffus

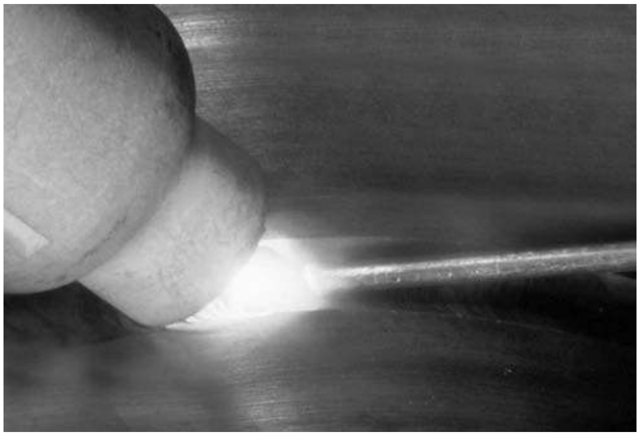

Figure 5.5
Filler being remelted as the weld is continued
Courtesy of Larry Jeffus

molten weld pool, **Figure 5.3.** When a weld is stopped so that the welder can change position, the shielding gas must be kept flowing around the rod end to protect it until it is cool. If the end of the rod becomes oxidized, it should be cut off before restarting. The following method can be used both to protect the rod end and reduce the possibility of crater cracking—that is, breaking the arc but keeping the torch over the crater while, at the same time, sticking the rod in the molten weld pool before it cools, **Figure 5.4.** When the weld is restarted, the rod is simply melted loose again, **Figure 5.5.**

The rod should enter the shielding gas as close to the base metal as possible, **Figure 5.6.** A 15° angle or less to the plate surface prevents air from being pulled into the welding zone behind the rod, **Figure 5.7.** As an example, if a rod is held in a stream of running water, air can be pulled in. The faster the water flows or the steeper the angle at which the rod is held, the more air is pulled in. The same action occurs with the shielding gas as its flow increases or as the rod angle increases.

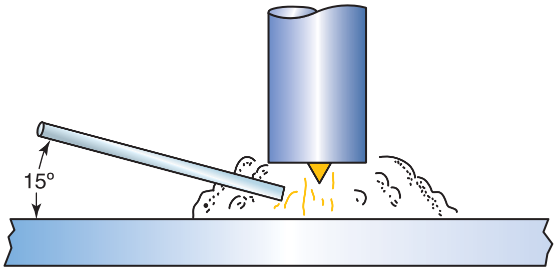

Figure 5.6
Keep the filled metal at approximately a 15° angle

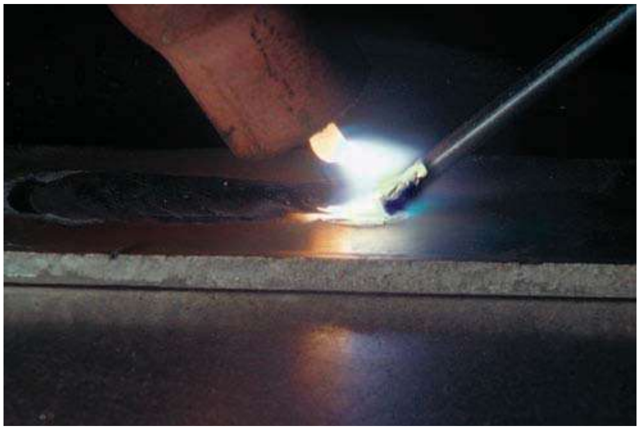

Figure 5.7
Too much filler rod angle has caused oxides to be formed on the filler rod end
Courtesy of Larry Jeffus

TUNGSTEN CONTAMINATION

For new welding students, the most frequently occurring and most time-consuming problem is tungsten **contamination.** The tungsten becomes contaminated when it touches the molten weld pool or when it is touched by the filler metal. When this happens, especially with aluminum, **surface tension** pulls the contamination up onto the hot tungsten, **Figure 5.8.** The extreme heat causes some of the metal to vaporize and form a large, widely scattered **oxide layer.** On aluminum, this layer is black. On iron (steel and stainless steel), this layer is a reddish color.

The contamination caused by the tungsten touching the molten weld pool or filler metal forms a weak weld. On a welding job, both the weld and the tungsten must be cleaned before any more welding can be done. The weld crater must be ground or chiseled to remove the tungsten contamination, and the tungsten end must be reshaped. Extremely tiny tungsten particles will show up if the weld is X-rayed. Failure to remove the contamination properly will result in the failure of the weld.

CAUTION

This procedure should never be used with heavy contaminations or when a welder is on the job in the field. It is designed only to help the new student in the first few days of training to save time and increase weld production.

When starting to weld, the beginning student may save weld practice time by burning off the contamination. On a scrap, usually copper plate, strike an arc using a higher than normal amperage setting. The arc will be erratic and discolored at first, but, as the contamination vaporizes, the arc will stabilize. Contamination can also be knocked off by quickly flipping the torch head.

CURRENT SETTING

The amperage set on a machine and the actual welding current are often not the same. The amperage indicated on the machine's control is the same as that at the arc only for the following conditions:

- The power to the machine is exactly correct.
- The lead length is very short.
- All cable connections are perfect with zero resistance.
- The arc length is exactly the right length.
- The remote current control is in the full on position.

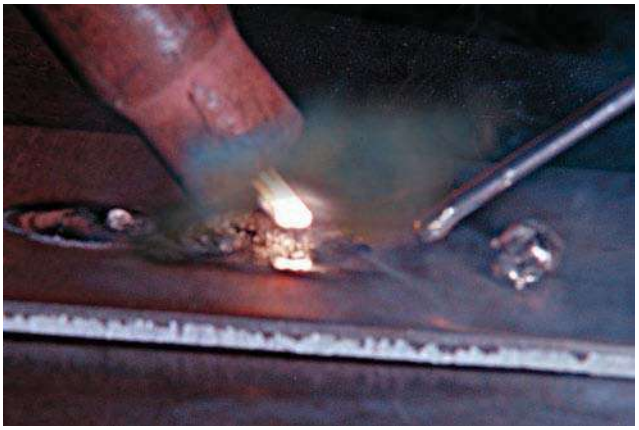

Figure 5.8
Contaminated tungsten
Courtesy of Larry Jeffus

Table 5.1 **Sample Chart Used to Record GTA Welding Machine Settings**

Current and Tungsten Electrode Size	Amperage/Machine Setting				
	Too Low	Low	Good	High	Too High

If any one of these factors changes, the actual welding amperage will change.

In addition to the difference between indicated and actual welding amperage, there is a more significant difference between amperage and welding power. The welding power, in watts, is based on the formula $W = E \times I$, or volts (E) multiplied by amperes (I) equals watts (W). Thus, the indicated power to a weld from two different types of welding machines set at 100 amperes will vary depending upon the voltage of the machine.

The welding machine setting will vary within a range from low to high (cool to hot). The range for one machine may be different from that of another machine. The setting will also be different for various types and sizes of tungstens, polarities, types and thicknesses of metal, joint position or design, and shielding gas used.

A chart, such as the one in **Table 5.1,** and a series of tests can be used to set the lower and upper limits for the amperage settings. As students' welding skills improve with practice, they will become familiar with the machine settings so that a table for these settings is no longer needed. In the welding industry, some welders will mark a line on the dial of the machine to help in resetting the machine. If a welder is required to make a number of different machine setups, a list or chart can be made and taped to the machine. This practice is more professional than marking the machine dials.

EXPERIMENTS

Experiments are designed to help new welders learn some basic skills that will help them troubleshoot welding problems. If you do the experiments listed in this chapter, you will be better able to determine what is causing a problem with your weld. As you learn more about welding, subtle changes will become more noticeable. Even experienced welders make changes in the setup, current, or welding technique as they try to resolve a problem.

Experiment 5-1 will help the welder determine the correct machine settings for the minimum and maximum welding current for the machine used, the types and sizes of tungstens, and the metal types and thicknesses. Most welding will be performed with a medium-range or mid-range machine setting. The exact setting is more important for machines without remote controls. The remote control allows changes in welding current to be made during the welding without having to stop.

EXPERIMENT 5-1

Setting the Welding Current

Module 1
Key Indicator 1, 4

Module 7
Key Indicator 1, 3, 4

Module 9
Key Indicator 1, 2

Using a properly set up GTA welding machine and torch, proper safety protection, one of each available tungsten size and type, and 16-gauge mild steel, 1/8 in. (3 mm) and 1/4 in. (6 mm) thick, you will work with a small group of students to develop a chart of the correct machine current setting for each type and size of tungsten.

Set the machine welding power switch for DCEN (DCSP) and the amperage control to its lowest setting, **Figure 5.9.** Sharpen a point on each tungsten and install one of the smaller diameter tungstens in the GTA torch. Select a nozzle with a 1/2-in. (13-mm)-diameter hole and attach it to the torch head. Set the preflow time to 0 and postflow to 10 to 15 seconds. Connect the remote control if it is available. Turn on the main power and hold the torch so that it cannot short out. Depress the remote controls to start the shielding gas so the flow rate can be set at 20 cfh (8 L/min). Switch the high frequency to start. All other functions, such as pulse, hot start, slope, and so on, should be in the off position.

Place the piece of 16-gauge sheet metal flat on the welding table. Hold the torch vertically with the tungsten about 1/4 in. (6 mm) above the metal. Lower your welding hood and fully depress the remote control. Watch the arc to see if it stabilizes and melts the metal. After a short period of time (15 to 30 seconds), stop, raise your hood, and check the plate for a melted spot. If melting occurred, note the size of the spot and depth of penetration, **Figure 5.10.** Increase the amperage setting by 5 or 10 A, note the setting on the chart, and repeat the process.

After each test, observe and record the results. The important settings to note are:

1. when the tungsten first heats up and the arc stabilizes
2. when the metal first melts
3. when 100% penetration of the metal first occurs
4. when burn-through first occurs
5. when the tungsten starts glowing white hot and/or melts

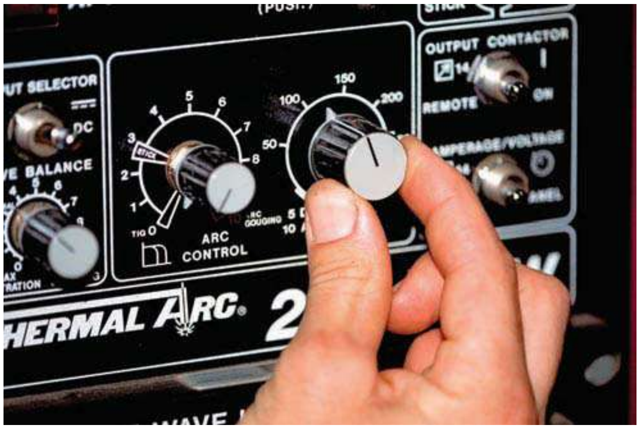

Figure 5.9
Lower the welding current to zero or as low as possible
Courtesy of Larry Jeffus

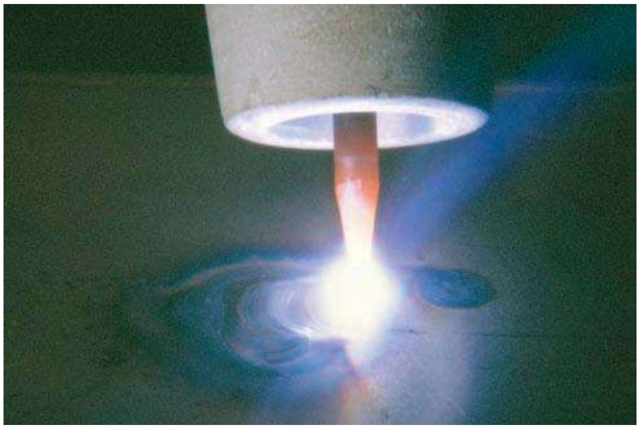

Figure 5.10
Melting first occurring
Courtesy of Larry Jeffus

The lowest (minimum) acceptable amperage setting is when the molten weld pool first appears on the base metal and the arc is stable. The highest (maximum) amperage setting is when the base metal burn-through or melting of the tungsten occurs. Any current setting between the high and low points is within the amperage range for that specific setup.

To establish the range for the next tungsten type or size, repeat the test. After each test, the metal should be cooled to prevent overheating. After each type and size of tungsten has been tested and an operating range established, repeat the procedure using the next thicker metal. Repeat this procedure until you have set up the operating ranges for all of the metals and tungstens you will be using. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

GAS FLOW

The gas preflow and postflow times required to protect both the tungsten and the weld depend upon the following factors:

- wind or draft speed
- nozzle size used
- tungsten size used
- amperage
- joint design
- welding position
- type of metal welded

The weld quality can be adversely affected by improper gas flow settings. The lowest possible gas flow rates and the shortest preflow or postflow time can help reduce the cost of welding by saving the expensive shielding gas.

In Experiment 5-2, the minimum and maximum gas flow settings for each nozzle size, tungsten size, and amperage setting will be determined. The chart a welder prepares based on experiments is to improve that welder’s skill and welding technique. Charts may differ slightly from one welder to another. As a welder’s skill improves, the chart may change. As experience is gained, a welder will learn how to set the gas flow effectively without the need for this chart.

The minimum flow rates and times must be increased when welding in drafty areas or for out-of-position welds. The rates and times can be somewhat lower for tee joints or welds made in tight areas. The maximum flow rates must never be exceeded. Exceeding these flow rates causes weld contamination and increases the rejection rate.

Module 1
Key Indicator 1, 4

Module 7
Key Indicator 1
Carbon Steel
Key Indicator 3, 4

Module 9
Key Indicator 1, 2

EXPERIMENT 5-2

Setting Gas Flow

Using a properly set up GTA welding machine and torch, proper safety protection, one of each available tungsten size, metal that is 16 gauge to 1/4 in. (6 mm) thick, and the welding current chart developed in

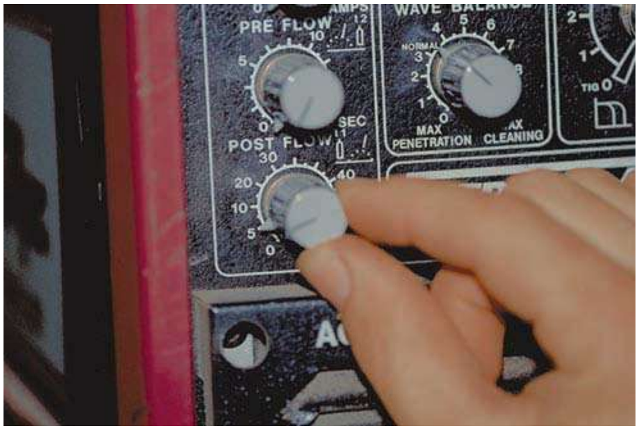

Figure 5.11
Setting the postflow timer
Courtesy of Larry Jeffus

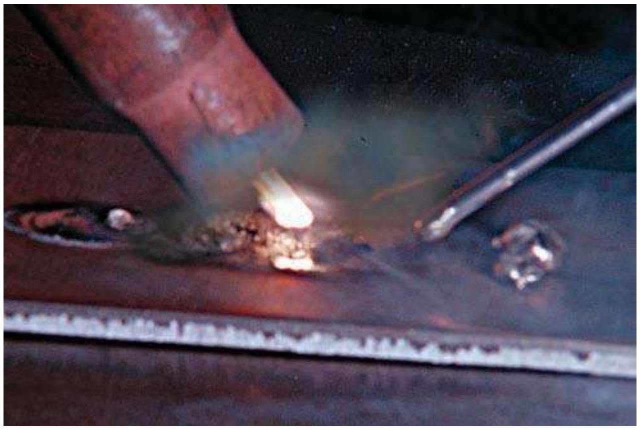

Figure 5.12
Oxides forming due to inadequate gas shielding
Courtesy of Larry Jeffus

Experiment 5-1, you will work with a small group of students to make a chart of the minimum and maximum flow rates and times for each nozzle size, tungsten size, and amperage setting. An assistant will also be needed to change and record the flow rate while you work.

Set the machine welding power switch for DCEN (DCSP). Set the amperage to the lowest setting for the size of tungsten used. Set the preflow time to 0 and postflow to 20 seconds, **Figure 5.11.** Turn on the main power. With the torch held so that it cannot short out, depress the remote control to start the shielding gas flow and set the flow at 20 cfh (9 L/min). Switch the high frequency to start. All other functions, such as pulse, hot start, slope, and so on, should be in the off position.

Starting with the smallest nozzle and tungsten size, strike an arc and establish a molten pool on a piece of metal in the flat position. Watch the molten weld pool and tungsten for signs of oxide formation as another person slowly lowers the gas flow rate. Have that person note this setting (where oxide formation begins), **Figure 5.12,** as the minimum flow rate on the chart next to the nozzle size and current setting. Now slowly increase the flow rate until the molten pool starts to be blown back or oxides start forming. This setting should be noted on the chart as the maximum flow rate for this current and nozzle size, **Table 5.2.** Lower the flow to a rate of 2 cfh or 3 cfh (1 L/min or 2 L/min) above the minimum value noted on the chart, and then stop the arc. Record the length of time from the point when the arc stops and the tungsten stops glowing as the postflow time. Repeat this test at a medium and then high current setting for this nozzle

Table 5.2 Sample Chart for Setting Shielding Gas Flow Rate and Time

Electrode and Nozzle Size	Flow Rate					Postflow Time		
	Too Low	Low	Good	High	Too High	Too Short	OK	Too Long

and tungsten size. When using high current settings, it may be necessary to move the torch or use thicker plate to prevent burn-through.

Repeat this test procedure with each available nozzle and tungsten size. Stainless steel or aluminum is preferred for this experiment because the oxides are more quickly noticeable than when mild steel is used. If aluminum is used, the welding current must be AC, and the high-frequency switch should be set on continuous.

To establish the minimum preflow time for each nozzle and tungsten size, set the amperage to a medium-high setting. Hold the torch above the metal so that an arc will be instantly started. Set the preflow timer to 0 and the gas flow to just above the minimum value noted on the chart. Quickly strike an arc on metal thin enough to cause a weld pool to form instantly at that power setting. Stop the arc and examine the weld pool and tungsten for oxides. Repeat this procedure, increasing the preflow time until no oxides are formed on either the plate or tungsten. Record this time on the chart as the minimum preflow time. Repeat this test with each available nozzle and tungsten size. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE WELDS

The practice welds are grouped according to the weld position and type of joint, and not by the type of metal. The order in which a person decides to do the welds is that person's choice. It is suggested that the stringer beads be done in each metal and position before the different joints are tried. Each metal has its own characteristics that may make one metal easier for a person to work on than another metal.

Mild steel is inexpensive and requires the least amount of cleaning. Slight changes in the metal have little effect on the welding skill required. Stainless steel is somewhat affected by cleanliness, requiring little preweld cleaning. However, the weld pool shows overheating or poor gas coverage. With aluminum, cleanliness is a critical factor. Oxides on aluminum may prevent the molten weld pool from flowing together. The surface tension helps hold the metal in place, giving excellent bead contour and appearance.

The degree of difficulty a welder encounters with each of these metals depends upon the individual's experience. Try each weld with each metal to determine which metal will be easiest to master first. The type of welding machine and materials used will also affect a welder's progress. Practice will help welders overcome any obstacle to their progress.

Low-Carbon and Mild Steels

Low-carbon and mild steel are two basic steel classifications. These steels are the most common type of steels a new GTA welding student will experience welding. Carbon is the primary alloy in these classifications of steel, and it ranges from 0.15% or less for low carbon and 0.15% to 0.30% for mild steel. The GTA welding techniques required for welding steels in both

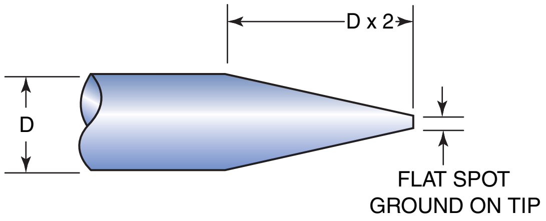

Figure 5.13
Tungsten tip shape for mild steel or stainless steel

classifications are the same. You start with EWTh-1, EWTh-2, EWCe-2, or one of the EWLa pointed tungsten electrodes, **Figure 5.13,** with the welding machine set for DCEN (DCSP) welding current. **Table 5.3** lists the types of filler metal used for both low-carbon and mild steels.

During the manufacturing process, small pockets of primarily carbon dioxide gas sometimes become trapped inside low-carbon and mild steels. There are only a few molecules of gas trapped inside the microscopic pockets within the steel, so they do not affect the steel strength. During most other types of welding, fluxes on the filler metal capture these gas pockets and they are removed. However, because GTA filler metals do not have fluxes when these gas pockets become hot during welding, they expand and can sometimes cause weld porosity. You are most likely to see porosity caused by these gases when you are not using a filler metal. Most GTA filler metals have some alloys, called deoxidizers, that can help prevent porosity caused by gases trapped in the base metal. RG45 gas welding rods do not have these deoxidizers and are not recommended for GTA welding.

Stainless Steel

The setup and manipulation techniques required for stainless steel are nearly the same as those for low-carbon and mild steels, and skills transfer is easy. The major difference is that most welds on steels do not show the effects of contamination as easily as do welds on stainless steels. To make a weld on stainless steel you must do a better job of precleaning the base metal and filler metal; make sure you have adequate shielding gas coverage and do not overheat the weld.

Table 5.3 **Filler Metals for Low-Carbon and Mild Steels**

SAE No.	Carbon %	AWS Filler Metal No.
	Low Carbon	
1006	0.08 max	RG60 or ER70S-3
1008	0.10 max	RG60 or ER70S-3
1010	0.08 to 0.15	RG60 or ER70S-3
	Mild Steel	
1015	0.11 to 0.16	RG60 or ER70S-3
1016	0.13 to 0.18	RG60 or ER70S-3
1018	0.15 to 0.20	RG60 or ER70S-3
1020	0.18 to 0.23	RG60 or ER70S-3
1025	0.22 to 0.29	RG60 or ER70S-3

Table 5.4 **Temperatures at Which Various Colored Oxide Layers Form on Steel**

Surface Color	Approximate Temperature at Which Color is Formed	
	°F	(°C)
Light Straw	400	(200)
Tan	450	(230)
Brown	525	(275)
Purple	575	(300)
Dark Blue	600	(315)
Black	800	(425)

Table 5.5 **Filler Metals for Stainless Steels**

AISI No.	AWS Filler No.	AISI No.	AWS Filler No.
303	ER308	310	ER310
304	ER308	316	ER316L
304L	ER308L	316L	ER316L
309	ER309	410	ER410

The most common sign that there is a problem with a stainless steel weld is the bead color after the weld. The greater the contamination, the darker the color. The exposure of the weld bead to the atmosphere before it has cooled will also change the bead color. It is impossible, however, to determine the extent of contamination of a weld with only visual inspection. Both light-colored and dark-colored welds may not be free from oxides. Thus, it is desirable to take the time and necessary precautions to make welds that are no darker than dark blue, **Table 5.4.** Welds with only slight oxide layers are better for multiple passes.

Using a low arc current setting with faster travel speeds is important when welding stainless steel, because some stainless steels are subject to carbide precipitation. Carbide precipitation, the combining of carbon with chromium, occurs in some stainless steels when they are kept at a temperature between 800°F and 1500°F (625°C and 815°C) for a long time. There are a number of ways of controlling carbide precipitation, including the adding of alloys to the stainless steel and the lowering of the percentage of carbon. Also, during welding, special alloy filler metals can be used to control the problem, but the most important thing a welder can do is travel fast and use as little welding heat as possible.

Black crusty spots may appear on weld beads. These spots are often caused by improper cleaning of the filler rod or failure to keep the end of the rod inside the shielding gas.

Table 5.5 lists some common types of stainless steels and the recommended filler metals.

Aluminum

Aluminum is GTA welded using EWP, EWZr, EWCe-2, or EWLa rounded tip tungsten electrodes on transformer rectifier welding power supplies, **Figure 5.14,** with the welding machine set for ACHF welding current. EWZr, EWCe-2, or EWLa electrodes with a pointed tip may be used with some inverter welding power supplies due to the heat control afforded by

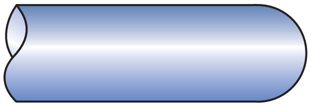

Figure 5.14
Tungsten tip shape for aluminum

Table 5.6 Filler Metals for Aluminum Alloys

AISI No.	AWS Filler No.	AISI No.	AWS Filler No.
1100	ER1100	3004	ER4043
3003	ER4043	6061	ER4043

their advanced circuitry. The alternating current provides good arc cleaning, and the continuous high frequency restarts the arc as the current changes direction.

The molten aluminum weld pool has high surface tension, which allows large weld beads to be controlled easily. **Table 5.6** lists some basic types of filler metals used for aluminum welding.

The high thermal conductivity of the metal may make starting a weld on thick sections difficult without first preheating the base metal. In most cases the preheat temperature is around 300°F (150°C) but will vary depending on metal thickness and alloy type. Specific preheat temperatures are available from the metal supplier.

The processes of cleaning and keeping the metal clean take a lot of time. Removal of the oxide layer is easy using a chemical or mechanical method. Ten minutes after cleaning, however, the oxide layer may again be thick enough to require recleaning. The oxide that forms reduces the ability of the weld pool to flow together. Keep your hands and gloves clean and oil free so the base metal or filler rods do not become recontaminated.

Although aluminum resists oxidation at room temperature, it rapidly oxidizes at welding temperatures. If the filler rod is not kept inside the shielding gas, it will quickly oxidize, but, because of the low melting temperature of the filler rod, the end will melt before it is added to the weld pool if it is held too closely to the arc, **Figure 5.15** and **Figure 5.16.**

CAUTION

The manufacturer's recommendations for using these products must be followed. Failure to do so may result in chemical burns, fires, fumes, or other safety hazards that could lead to serious injury. If anyone should come in contact with any chemicals, immediately refer to the material safety data sheet (MSDS) for the proper corrective action.

Metal Preparation

Both the base metal and the filler metal used in the GTAW process must be thoroughly cleaned before welding. Contamination left on the metal will be deposited in the weld because there is no flux to remove it. Oxides, oils, and dirt are the most common types of contaminants. They can be

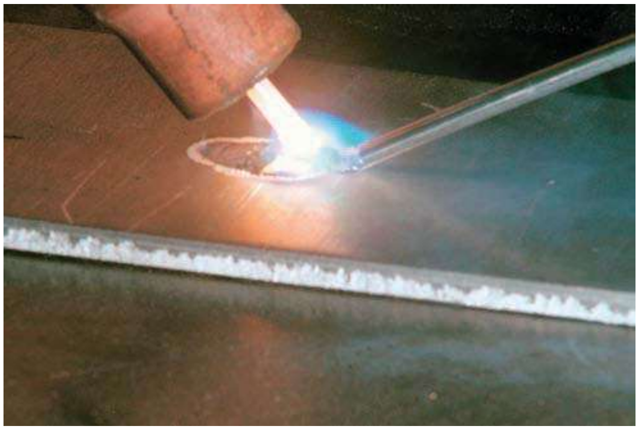

Figure 5.15
Aluminum filler being correctly added to the molten weld pool
Courtesy of Larry Jeffus

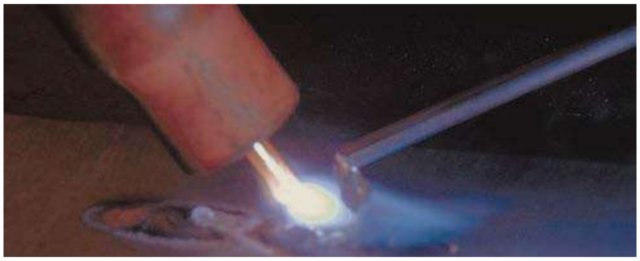

Figure 5.16
Filler rod being melted before it is added to the molten pool
Courtesy of Larry Jeffus

removed mechanically or chemically. Mechanical metal cleaning may be done by grinding, wire brushing, scraping, machining, or filing. Chemical cleaning may be done by using acids, alkalies, solvents, or detergents.

PRACTICE 5-1

Stringer Beads, Flat Position, on Mild Steel

Using a properly set up and adjusted GTA welding machine on DCEN, proper safety protection, and one or more pieces of mild steel, 6 in. (152 mm) long and 16 gauge and 1/8 in. (3 mm) thick, you will push a weld pool in a straight line down the plate, **Figure 5.17.** Maintain uniform weld pool size and penetration.

- Starting at one end of the piece of metal that is 1/8 in. (3 mm) thick, hold the torch as close as possible to a 90° angle.
- Lower your hood, strike an arc, and establish a weld pool.
- Move the torch in a stepping or circular oscillation pattern down the plate toward the other end, **Figure 5.18.**
- If the size of the weld pool changes, speed up or slow down the travel rate to keep the weld pool the same size for the entire length of the plate.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 3, 4

Module 9
Key Indicator 1,2

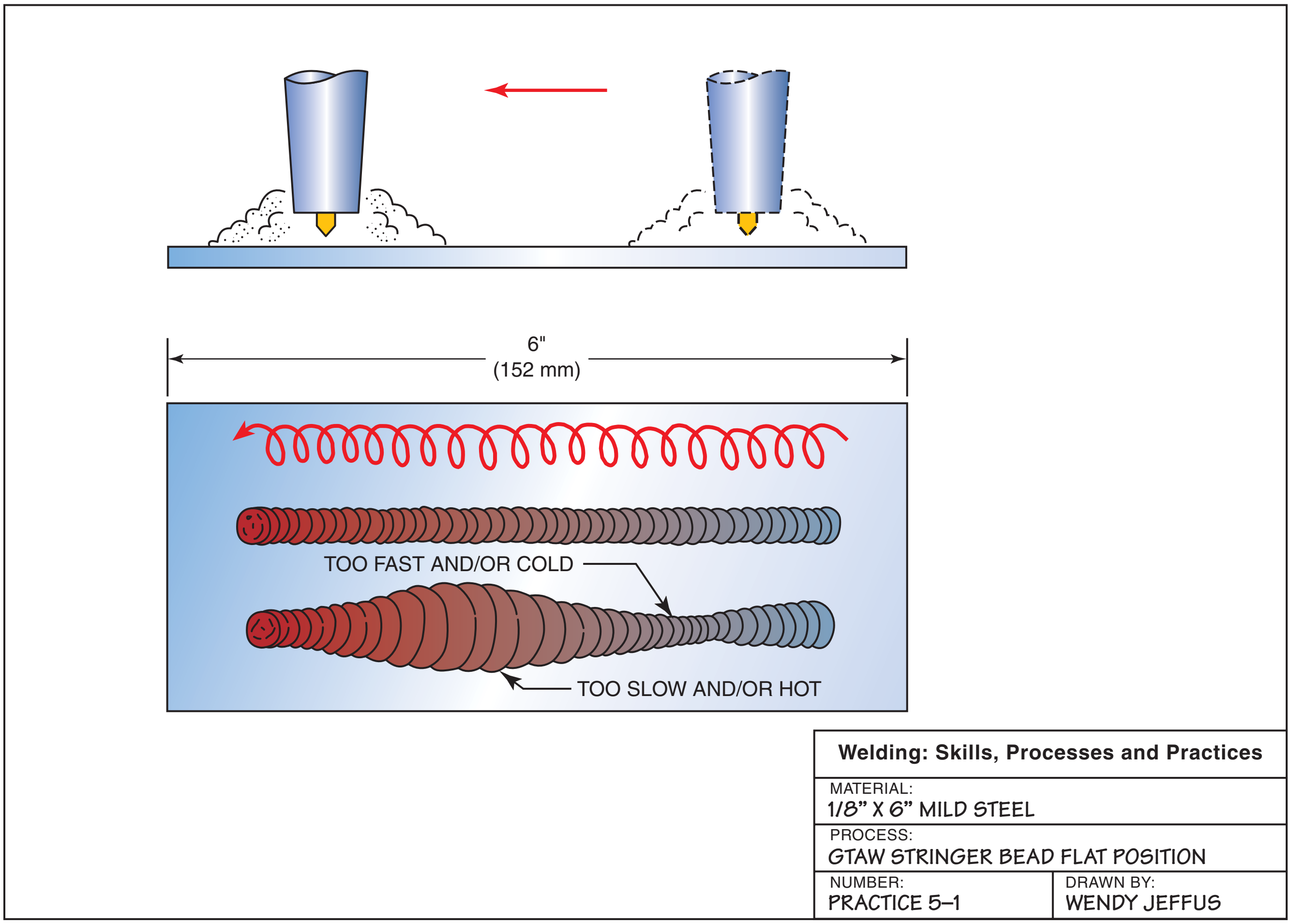

Welding: Skills, Processes and Practices
MATERIAL:
1/8" X 6" MILD STEEL
PROCESS:
GTAW STRINGER BEAD FLAT POSITION
NUMBER:
PRACTICE 5–1
DRAWN BY:
WENDY JEFFUS

Figure 5.17
Surfacing weld in the flat position

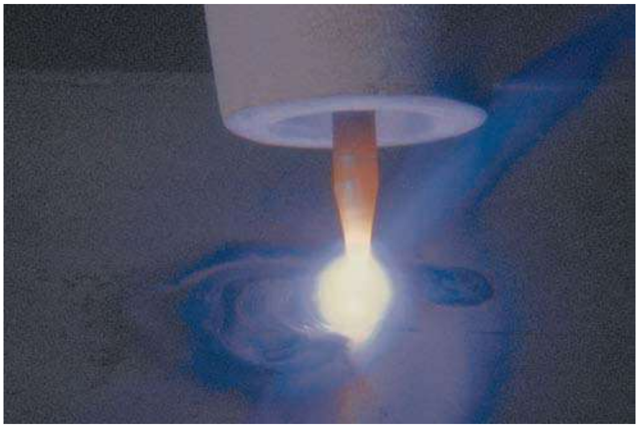

Figure 5.18
Surfacing weld
Courtesy of Larry Jeffus

The ability to maintain uniformity in width and keep a straight line increases as you are able to see more than just the weld pool. As your skill improves, you will relax, and your field of vision will increase.

Repeat the process using both thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Austenitic Stainless Steel
Key Indicator 8, 9

Module 9
Key Indicator 1, 2

PRACTICE 5-2

Stringer Beads, Flat Position, on Stainless Steel

Using the same equipment and material thicknesses as listed in Practice 5-1 and one or more pieces of stainless steel, 6 in. (152 mm) long and 1/4 in. (6 mm) thick, you will push a molten weld pool in a straight line down the plate, keeping the width and penetration uniform.

To keep the formation of oxides on the bead to a minimum, a **chill plate** (a thick piece of metal used to absorb heat) may be required. Another method is to make the bead using as low a heat input as possible. When the weld is finished, the weld bead should be no darker than dark blue.

Repeat the process using both thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Aluminum
Key Indicator 13, 14

Module 9
Key Indicator 1, 2

PRACTICE 5-3

Stringer Beads, Flat Position, on Aluminum

Using the same equipment, setup for AC, and procedure as listed in Practice 5-1 and pieces of aluminum, 6 in. (152 mm) long and 1/16 in. (2 mm), 1/8 in. (3 mm), and 1/4 in. (6 mm) thick, you will push a weld pool in a straight line, maintaining uniform width and penetration for the length of the plate.

A high current setting will allow faster travel speeds. The faster speed helps control excessive penetration. Hot cracking may occur on some types of aluminum after a surfacing weld. This is not normally a problem when filler metal is added. If hot cracking should occur during this practice, do not be concerned.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 5-4

Flat Position, Using Mild Steel, Stainless Steel, Aluminum

For this practice, you will need a properly set up and adjusted GTA welding machine, proper safety protection, and filler rods, 36 in. (0.9 m) long × 1/16 in. (2 mm), 3/32 in. (2.4 mm), and 1/8 in. (3 mm) in diameter, one or more pieces of mild steel, stainless steel, and aluminum, 6 in. (152 mm) long × 1/16 in. (2 mm) and 1/8 in. (3 mm) thick, and aluminum plate 1/4 in. (6 mm) thick, **Table 5.7, Table 5.8,** and **Table 5.9.** In this practice,

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 3, 4
Austenitic Stainless Steel
Key Indicator 8, 9
Aluminum
Key Indicator 13, 14

Module 9
Key Indicator 1, 2

Table 5.7 Suggested Setting for GTA Welding of Mild Steel

Tungsten			Welding			Shielding			Filler	
Type	Size	Tip	Amp	Current	HF	Type	Flow	Size	Type	Size
EWTh-1 or EWTh-2 or EWCe-2 or EWLa	1/16 in. (2 mm)	Point	50 to 100	DCEN DCSP	Start or auto	Argon	16 cfh 7 L/min auto	3/8 in. (10 mm)	RG60 or ER70S-3	1/16– 3/32 in. (2–2.4 mm)
	3/32 in. (2.4 mm)	Point	70 to 150	DCEN DCSP	Start or auto	Argon	16 cfh 7 L/min auto	3/8 in. (10 mm)	RG60 or ER70S-3	1/16– 3/32 in. (2–2.4 mm)
	1/8 in. (3 mm)	Point	90 to 250	DCEN DCSP	Start or auto	Argon	20 cfh 9 L/min	1/2 in. (13 mm)	RG60 or ER70S-3	3/32– 1/8 in. (2.4–3 mm)

Table 5.8 Suggested Setting for GTA Welding of Stainless Steel

Tungsten			Welding			Shielding			Filler	
Type	Size	Tip	Amp	Current	HF	Type	Flow	Size	Type	Size
EWTh-1 or EWTh-2	1/16 in. (2 mm)	Point	70 to 100	DCEN DCSP	Start or auto	Argon	16 cfh 7 L/min	3/8 in. (10 mm)	ER308 or ER316	1/16– 3/32 in. (2–2.4 mm)
EWTh-1 or EWCe-2 or EWLa	3/32 in. (2.4 mm)	Point	70 to 150	DCEN DCSP	Start or auto	Argon	16 cfh 7 L/min	3/8 in. (10 mm)	ER308 or ER316	1/16– 3/32 in. (2–2.4 mm)
	1/8 in. (3 mm)	Point	90 to 250	DCEN DCSP	Start or auto	Argon	20 cfh 9 L/min	1/2 in. (13 mm)	ER308 or ER316	3/32– 1/8 in. (2.4–3 mm)

Table 5.9 Suggested Setting for GTA Welding of Aluminum

Tungsten			Welding			Shielding			Filler	
Type	**Size**	**Tip**	**Amp**	**Current**	**HF**	**Type**	**Flow**	**Size**	**Type**	**Size**
EWZr or EWLa or EWCe-2 or EWP	1/16 in.	Round 2 mm	50 to 90	AC	Continues or on	Argon	17 cfh 8 L/min	7/16 in. (11 mm)	ER1100 or ER4043	1/16– 3/32 in. (2–2.4 mm)
	3/32 in.	Round 2.4 mm	80 to 130	AC	Continues or on	Argon	20 cfh 9 L/min	1/2 in. (13 mm)	ER1100 or ER4043	1/16– 3/32 in. (2–2.4 mm)
	1/8 in.	Round 3 mm	100 to 200	AC	Continues or on	Argon	20 cfh 9 L/min	5/8 in. (16 mm)	ER1100 or ER4043	3/32– 1/8 in. (2.4–3 mm)

Note: A pointed electrode may be preferred with AC and inverter GTAW power supplies with enhanced balance control capabilities.

you will make a straight stringer bead, 6 in. (152 mm) long, that is uniform in width, reinforcement, and penetration, **Figure 5.19.** Use DCEN current on the steel and AC current on the aluminum.

Starting with the metal that is 1/8 in. (3 mm) thick and the filler rod having a 3/32-in. (2.4-mm) diameter, strike an arc and establish a weld pool, **Figure 5.20.** Move the torch in a circle as in the practice beading. When the torch is on one side, add filler rod to the other side of the molten weld pool, **Figure 5.21.** The end of the rod can be held lightly in the leading edge of the molten weld pool, or it can be dipped into the molten weld pool. If you are using the dipping method, be sure not to allow the tip to melt and drip into the weld pool, **Figure 5.22.** Change to another size filler rod and determine its effect on the weld pool.

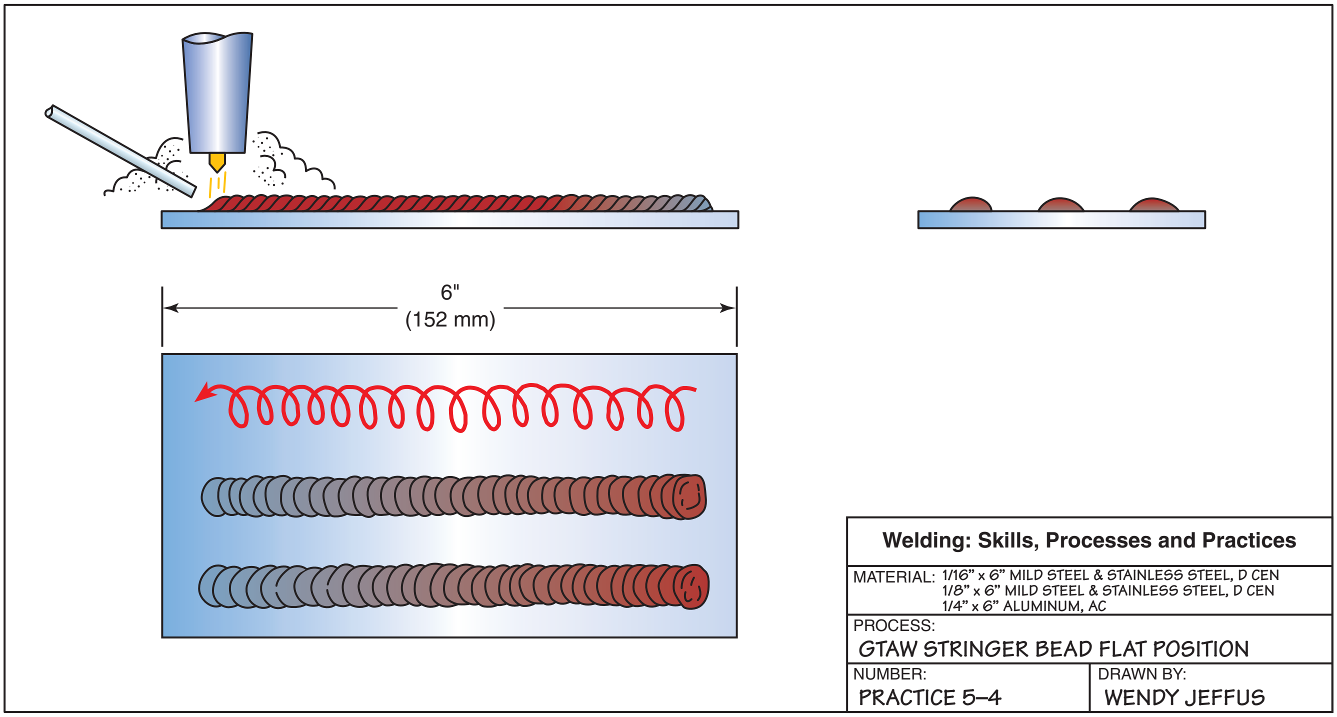

Welding: Skills, Processes and Practices

MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC

PROCESS:
GTAW STRINGER BEAD FLAT POSITION

NUMBER:
PRACTICE 5–4

DRAWN BY:
WENDY JEFFUS

Figure 5.19
Stringer beads in the flat position

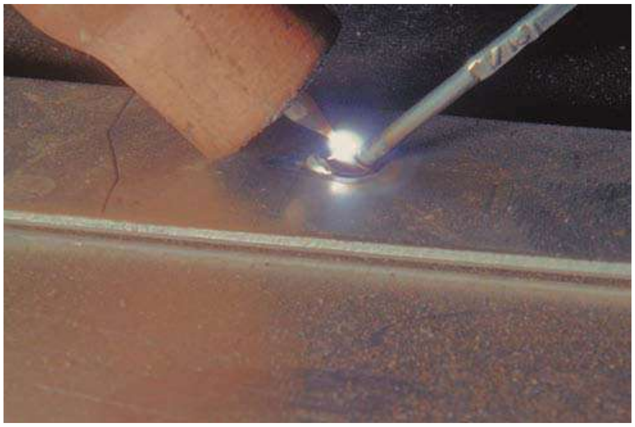

Figure 5.20
Establish a molten weld pool and dip the filler rod into it
Courtesy of Larry Jeffu

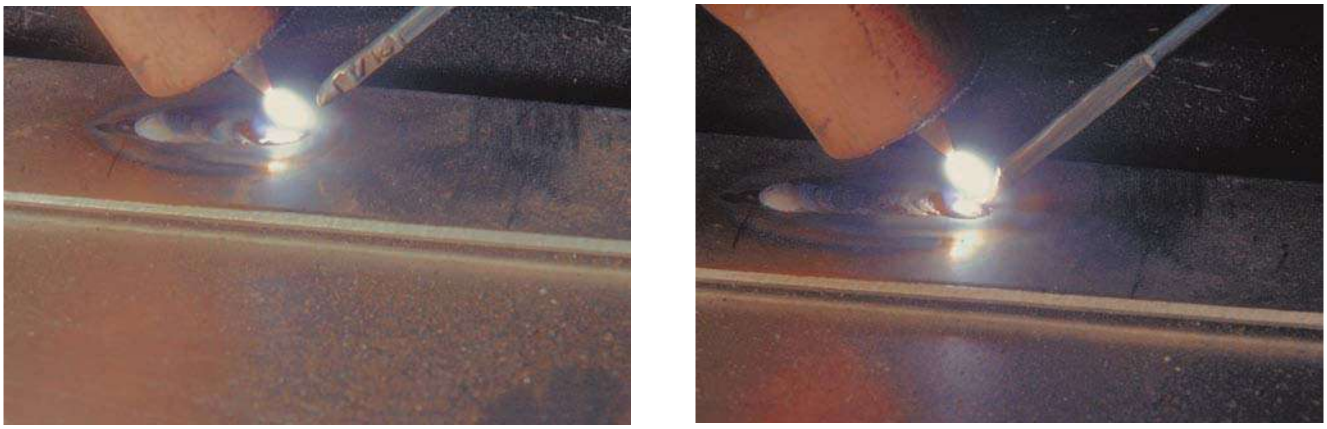

Figure 5.21
Note the difference in the weld produced when different size filler rods are used
Courtesy of Larry Jeffus

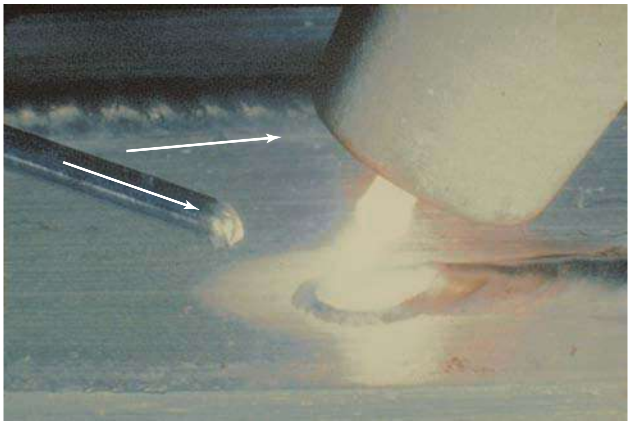

Figure 5.22
Move the electrode back as the filler rod is added
Courtesy of Larry Jeffus

Maintain a smooth and uniform rhythm as filler metal is added. This will help to keep the bead uniform. Vary the rhythms to determine which one is easiest for you. If the rod sticks, move the torch toward the rod until it melts free.

When the full 6-in. (152-mm)-long weld bead is completed, cool and inspect it for uniformity and defects. Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Key Indicator 1, 2
Carbon Steel
Key Indicator 3, 4
Austenitic Stainless Steel
Key Indicator 8, 9
Aluminum
Key Indicator 13, 14

Module 9
Key Indicator 1, 2

PRACTICE 5-5

Outside Corner Joint, 1G Position, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials listed in Practice 5-4, weld an outside corner joint in the flat position, **Figure 5.23.**

- Place one of the pieces of metal flat on the table and hold or brace the other piece of metal horizontally on it.
- Tack weld both ends of the plates together, **Figure 5.24.**
- Set the plates up and add two or three more tack welds on the joint as required, **Figure 5.25.**

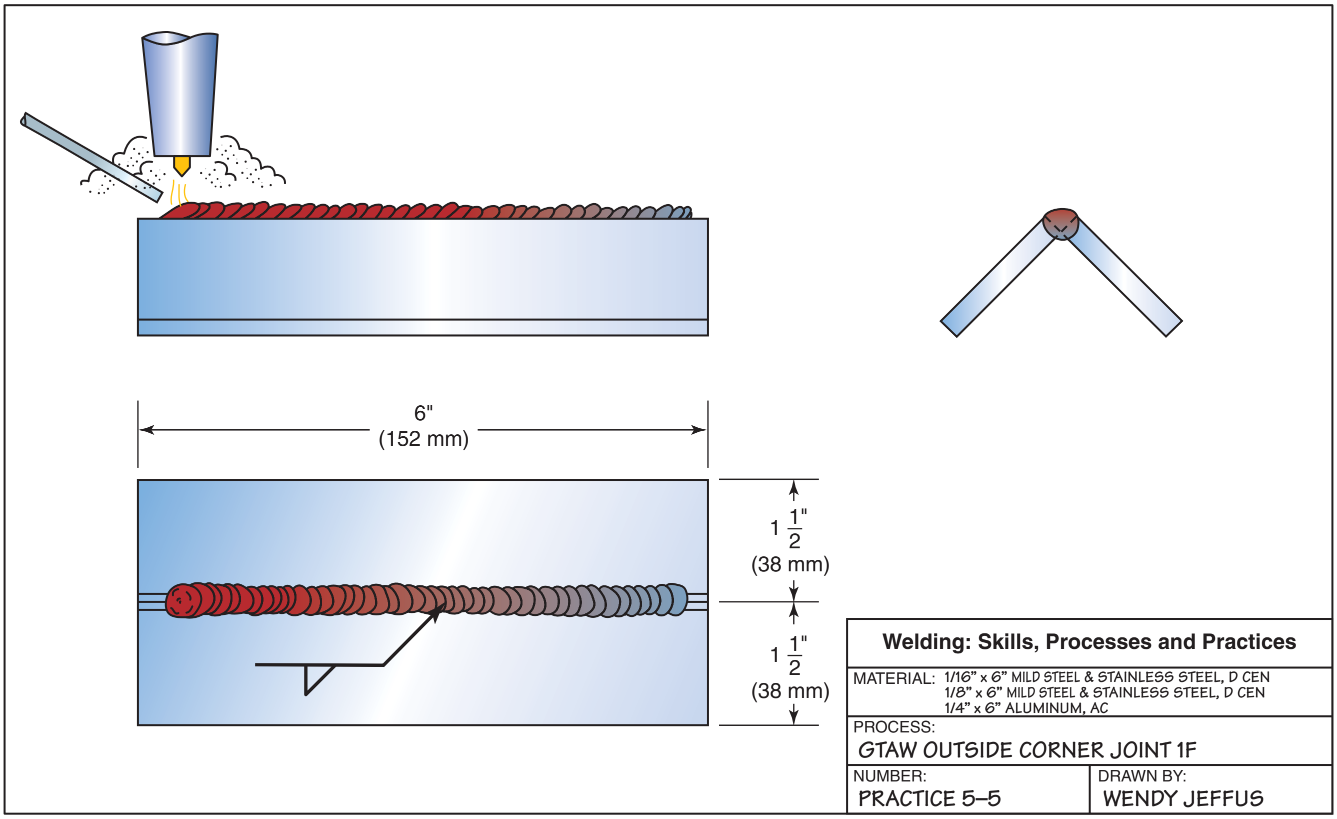

Welding: Skills, Processes and Practices
MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC
PROCESS:
GTAW OUTSIDE CORNER JOINT 1F
NUMBER:
PRACTICE 5–5
DRAWN BY:
WENDY JEFFUS

Figure 5.23
Outside corner joint in the flat position

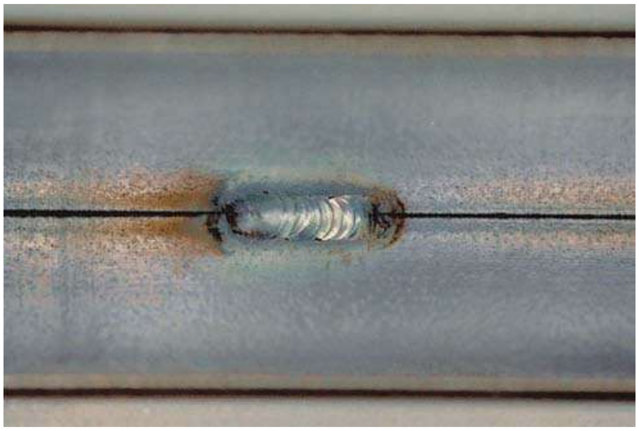

Figure 5.24
Tack weld. Note the good fusion at the start and crater fill at the end.
Courtesy of Larry Jeffus

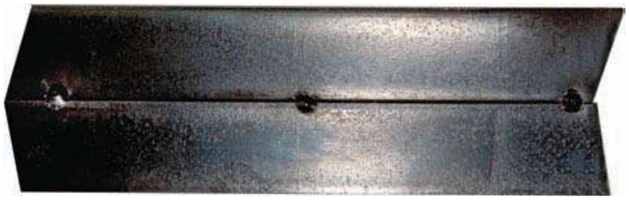

Figure 5.25
Outside corner tack welded together
Courtesy of Larry Jeffus

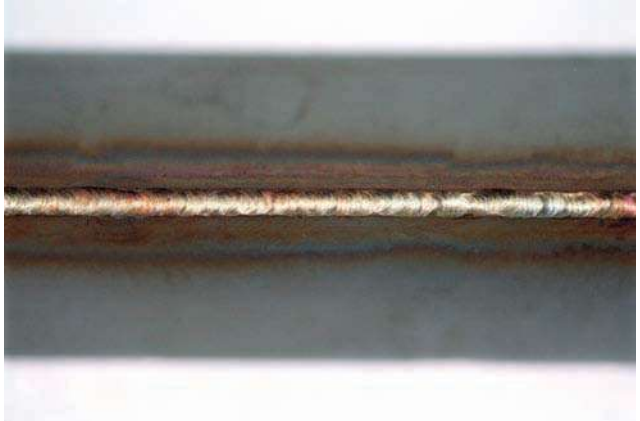

Figure 5.26
Outside corner joint. Note precleaning along weld.
Courtesy of Larry Jeffus

- Starting at one end, make a uniform weld, adding filler metal as needed. In **Figure 5.26,** note the metal areas that are precleaned before the weld is made.

Repeat each weld as needed until all are mastered. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 16

Module 9
Key Indicator 1, 2

PRACTICE 5-6

Butt Joint, 1G Position, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will weld a butt joint in the flat position, **Figure 5.27.**

Place the metal flat on the table and tack weld both ends together, **Figure 5.28.** Two or three additional tack welds can be made along the

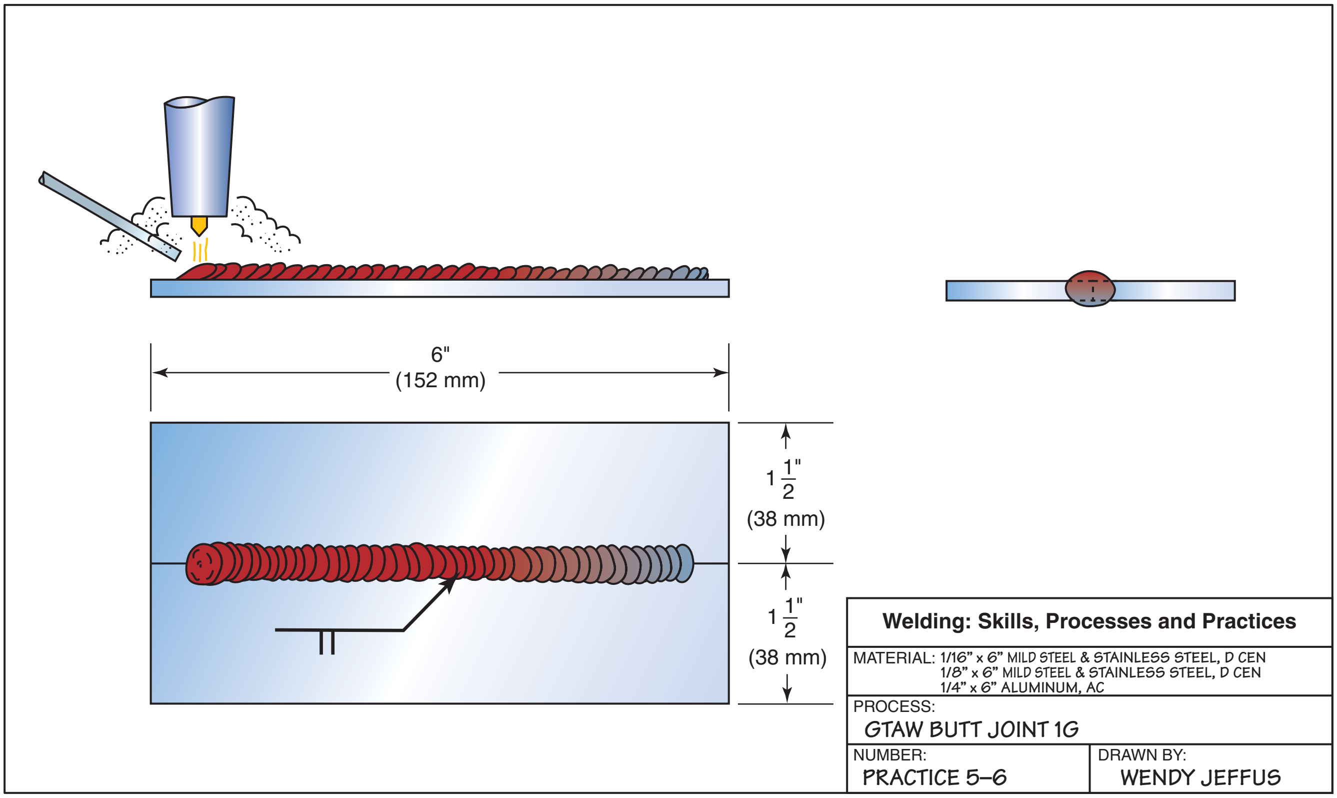

Welding: Skills, Processes and Practices

MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC

PROCESS:
GTAW BUTT JOINT 1G

NUMBER:
PRACTICE 5–6

DRAWN BY:
WENDY JEFFUS

Figure 5.27
Square butt joint in the flat position

joint as needed. Starting at one end, make a uniform weld along the joint. Add filler metal as required to make a uniform weld.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

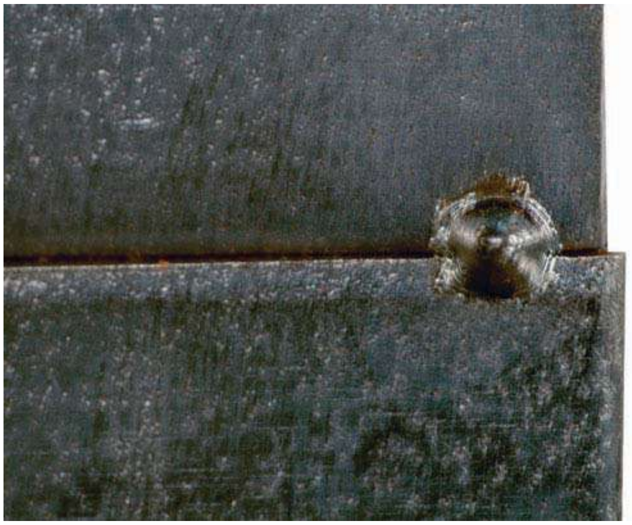

Figure 5.28
Tack weld on butt joint
Courtesy of Larry Jeffus

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-7

Butt Joint, 1G Position, with 100% Penetration, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will weld a butt joint with 100% penetration, **Figure 5.29,** along the entire 6-in. (152-mm) length of the joint. After the weld is completed, shear out strips 1 in. (25 mm) wide and bend-test them as shown in **Figure 5.30.**

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 16

Module 9
Key Indicator 1, 2

PRACTICE 5-8

Butt Joint, 1G Position, with Minimum Distortion, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will weld a flat butt joint, while controlling both distortion and penetration, **Figure 5.31.**

Tack weld the plates together as shown in **Figure 5.32.** Using a backstepping weld sequence, make a series of welds approximately 1 in. (25 mm) long along the joint. Be sure to fill each weld crater adequately to reduce crater cracking.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 16

Module 9
Key Indicator 1, 2

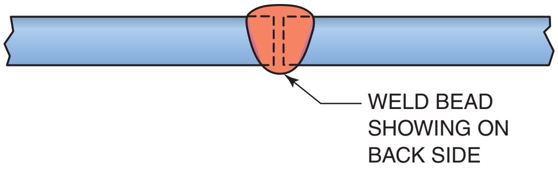

Figure 5.29
100% weld penetration

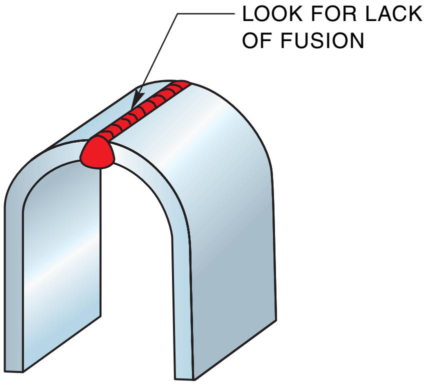

Figure 5.30
Bend the 1-in. (25-mm) strip of butt joint backward and look at the root for 100% penetration

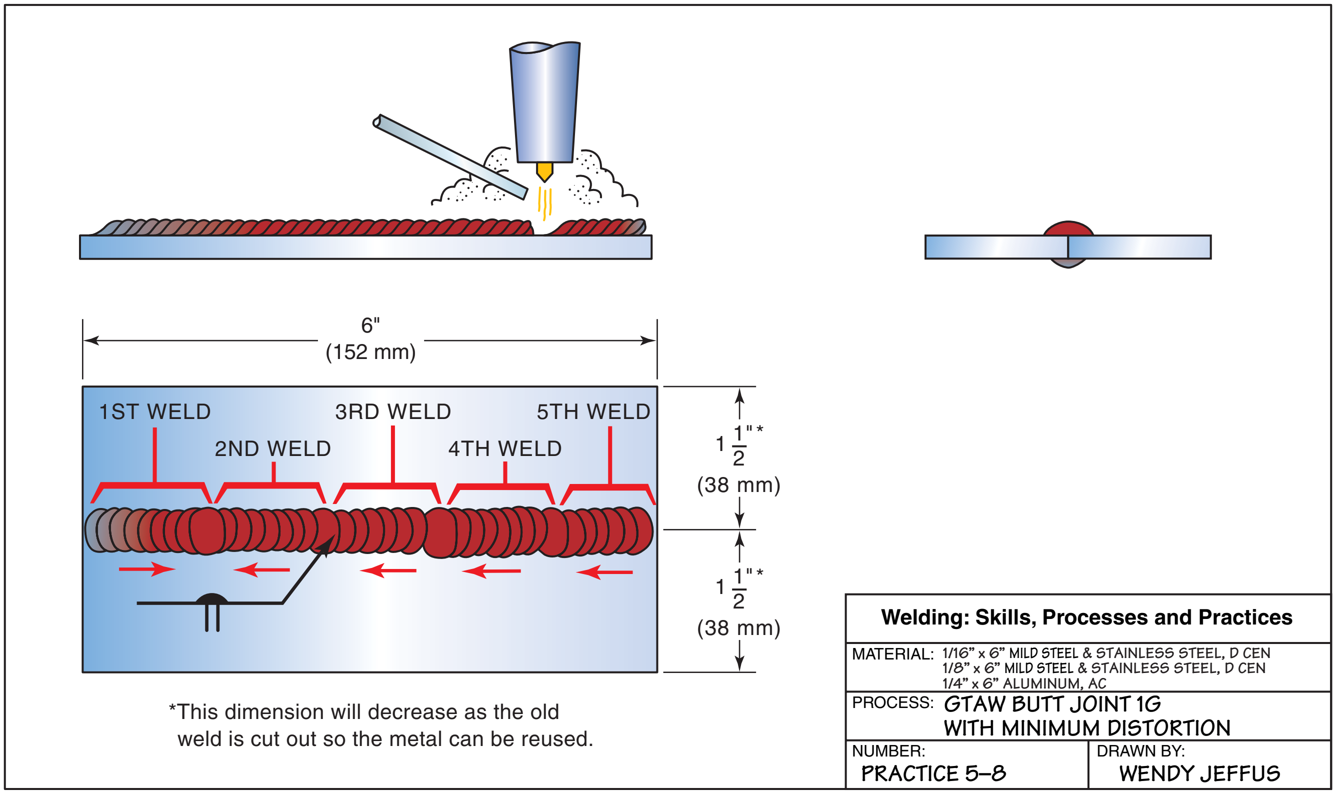
6"
(152 mm)
1ST WELD
2ND WELD
3RD WELD
4TH WELD
5TH WELD
1 1/2"*
(38 mm)
1 1/2"*
(38 mm)
*This dimension will decrease as the old weld is cut out so the metal can be reused.
Welding: Skills, Processes and Practices
MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC
PROCESS: GTAW BUTT JOINT 1G
WITH MINIMUM DISTORTION
NUMBER:
PRACTICE 5-8
DRAWN BY:
WENDY JEFFUS

Figure 5.31
Square butt joint in the flat position with minimum distortion

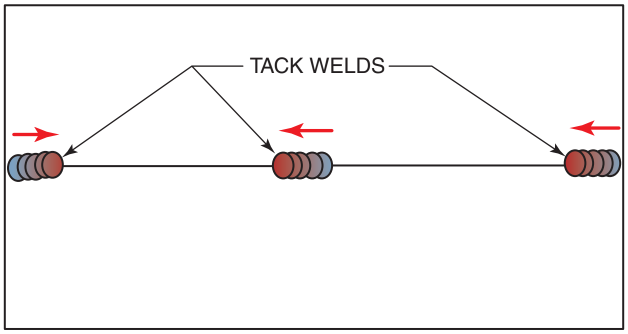

Figure 5.32
Tack welds on a butt joint

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Module 9
Key Indicator 1, 2

PRACTICE 5-9

Lap Joint, 1F Position, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will weld a lap joint in the flat position, **Figure 5.33.**

Place the two pieces of metal flat on the table with an overlap of 1/4 in. (6 mm) to 3/8 in. (10 mm). Hold the pieces of metal tightly together and tack weld them as shown in **Figure 5.34** and **Figure 5.35.** Starting at one end, make a uniform fillet weld along the joint. Both sides of the joint can be welded.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

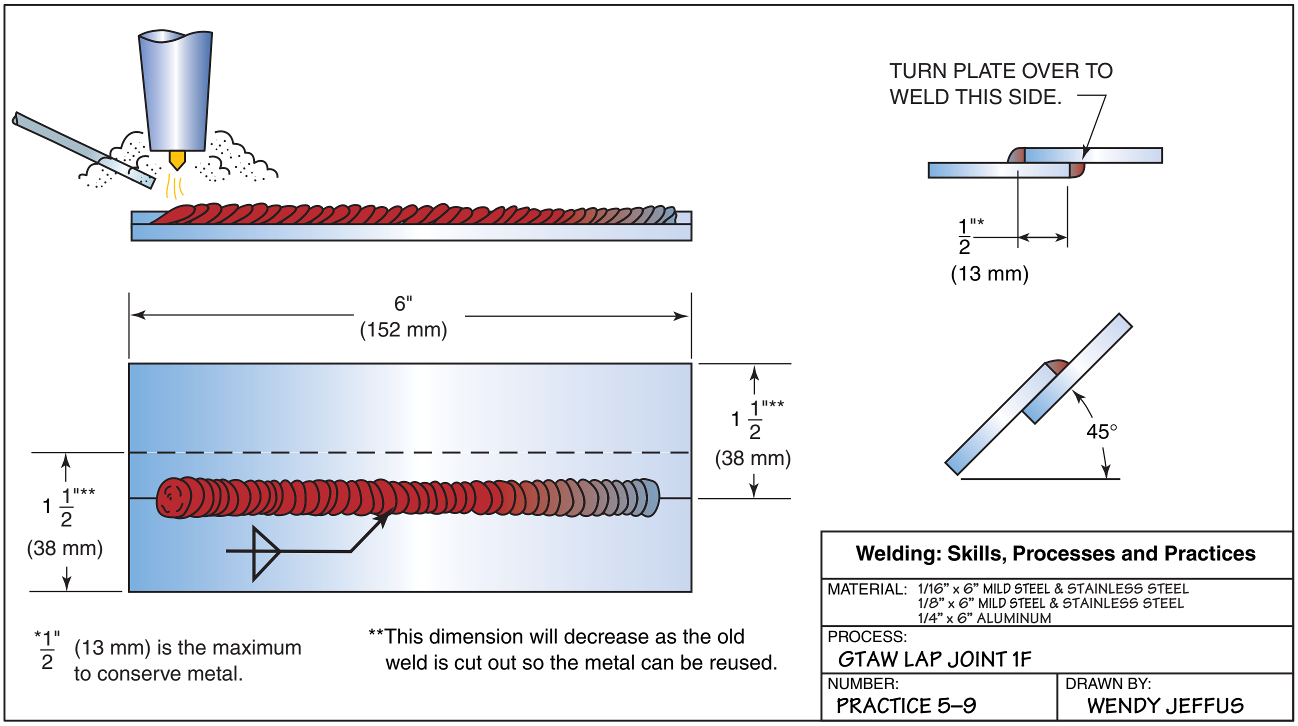

*1/2" (13 mm) is the maximum to conserve metal.

**This dimension will decrease as the old weld is cut out so the metal can be reused.

Welding: Skills, Processes and Practices

MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL
1/8" x 6" MILD STEEL & STAINLESS STEEL
1/4" x 6" ALUMINUM

PROCESS: GTAW LAP JOINT 1F

NUMBER: PRACTICE 5–9

DRAWN BY: WENDY JEFFUS

Figure 5.33
Lap joint in the flat position

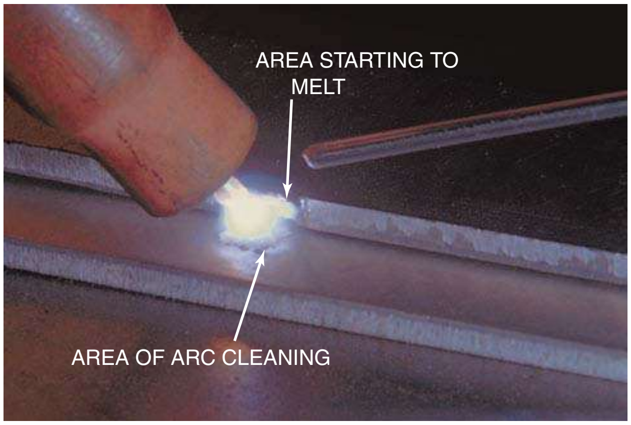

Figure 5.34
Be sure both the top and bottom pieces are melted before adding filler metal
Courtesy of Larry Jeffus

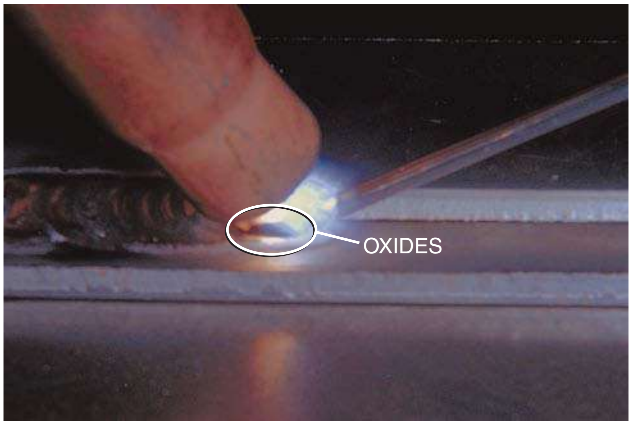

Figure 5.35
Oxides form during tack welding. Do not complete the tack welds. These oxides will become part of the finished weld if the tack is completed.
Courtesy of Larry Jeffus

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-10

Lap Joint, 1F Position, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will make a fillet weld on one side of a lap joint and test it for 100% root penetration, **Figure 5.36, Figure 5.37,** and **Figure 5.38.** After the weld is completed, shear out strips 1 in. (25 mm) wide on one side of the joint and bend-test them as shown in **Figure 5.39.**

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Module 9
Key Indicator 1, 2

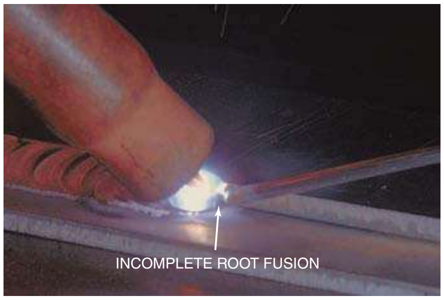
INCOMPLETE ROOT FUSION

Figure 5.36
A notch indicates that the root was not properly melted and fused
Courtesy of Larry Jeffus

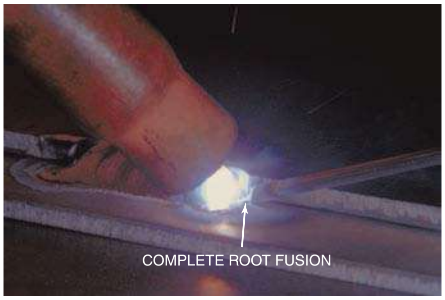
COMPLETE ROOT FUSION

Figure 5.37
Watch the leading edge of the molten weld pool to ensure that there is complete root fusion
Courtesy of Larry Jeffus

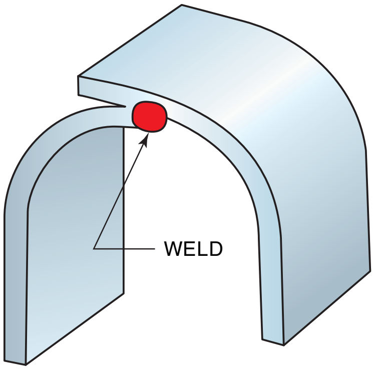

Figure 5.38
Bend the 1-in (25-mm) strip of lap joint backward and look at the root for 100% penetration

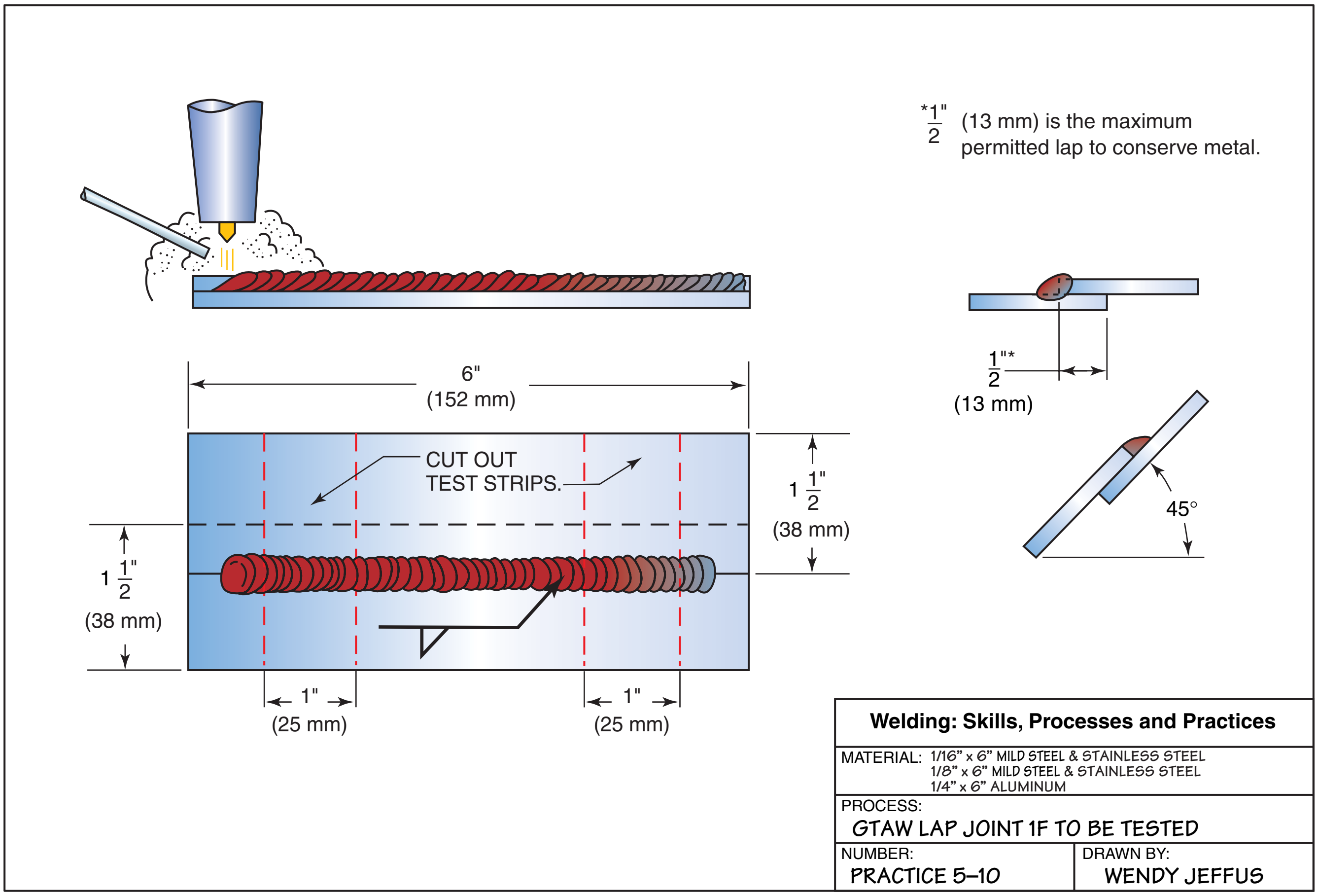
*1/2" (13 mm) is the maximum permitted lap to conserve metal.
6" (152 mm)
CUT OUT TEST STRIPS.
1 1/2" (38 mm)
1 1/2" (38 mm)
1" (25 mm)
1" (25 mm)
1/2"* (13 mm)
45°
Welding: Skills, Processes and Practices
MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL
1/8" x 6" MILD STEEL & STAINLESS STEEL
1/4" x 6" ALUMINUM
PROCESS:
GTAW LAP JOINT 1F TO BE TESTED
NUMBER:
PRACTICE 5–10
DRAWN BY:
WENDY JEFFUS

Figure 5.39
Remove strips and test for root fusion

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 5-11

Tee Joint, 1F Position, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will weld a tee joint in the flat position, **Figure 5.40.**

Place one of the pieces of metal flat on the table and hold or brace the other piece of metal horizontally on it. Tack weld both ends of the plates together, **Figure 5.41.** Set up the plates in the flat position and add two or three more tack welds to the joint as required, **Figure 5.42.**

On the metal that is 1/16 in. (1.5 mm) thick, it may not be possible to weld both sides, but on thicker material a fillet weld can usually be made on both sides. The exception to this is if carbide precipitation occurs on the stainless steel during welding.

Starting at one end, make a uniform weld, adding filler metal as needed.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Module 9
Key Indicator 1, 2

PRACTICE 5-12

Tee Joint, 1F Position, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will weld a tee joint and test it for 100% root penetration, **Figure 5.43.** After

Module 1
Key Indicator 1, 3, 4

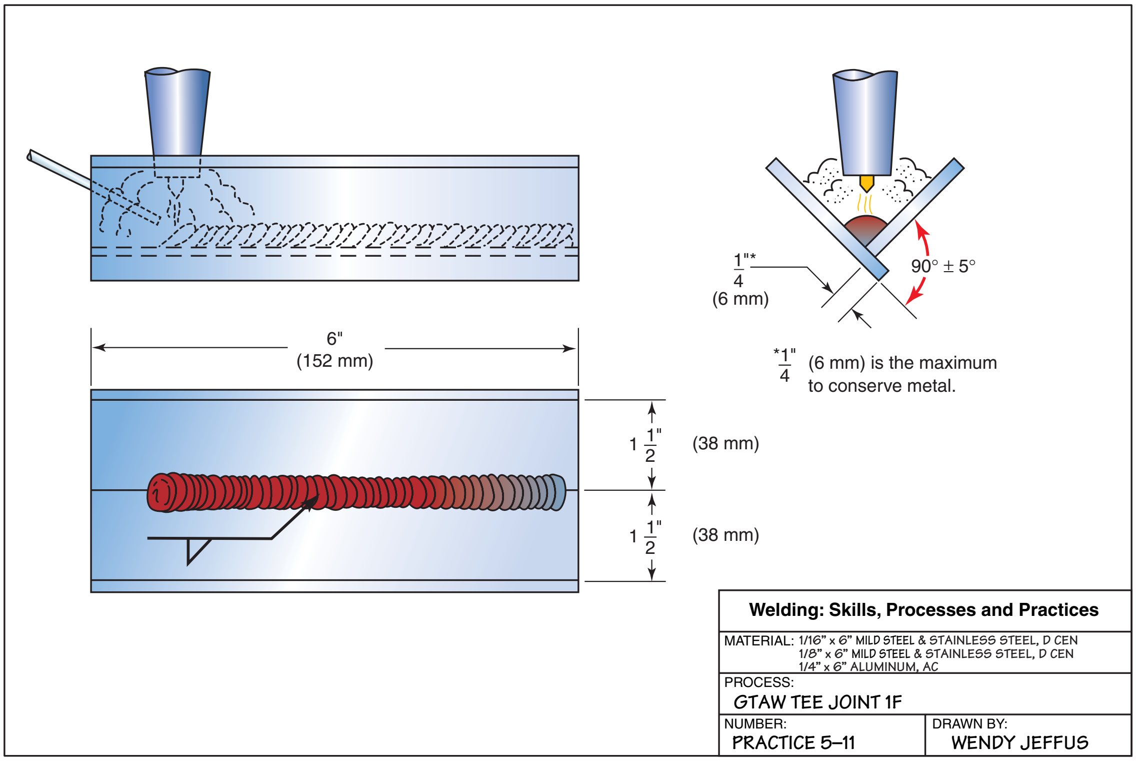

Welding: Skills, Processes and Practices
MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC
PROCESS:
GTAW TEE JOINT 1F
NUMBER:
PRACTICE 5–11
DRAWN BY:
WENDY JEFFUS

Figure 5.40
Tee joint in the flat position

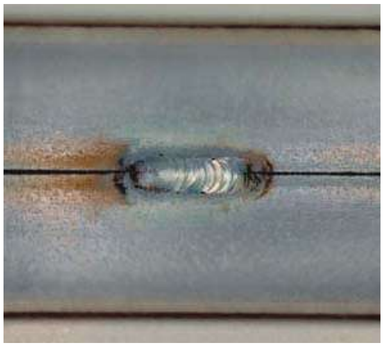

Figure 5.41
Tack weld on a tee joint
Courtesy of Larry Jeffus

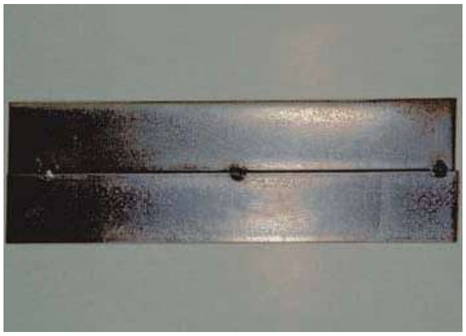

Figure 5.42
Keep the tack welds small so that they will not affect the weld
Courtesy of Larry Jeffus

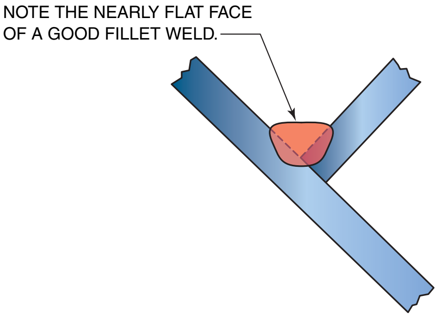
NOTE THE NEARLY FLAT FACE OF A GOOD FILLET WELD.

Figure 5.43
100% weld penetration

the weld is completed, cut or shear out strips 1 in. (25 mm) wide and bend-test them as shown in **Figure 5.44** and **Figure 5.45.**

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Module 9
Key Indicator 1, 2

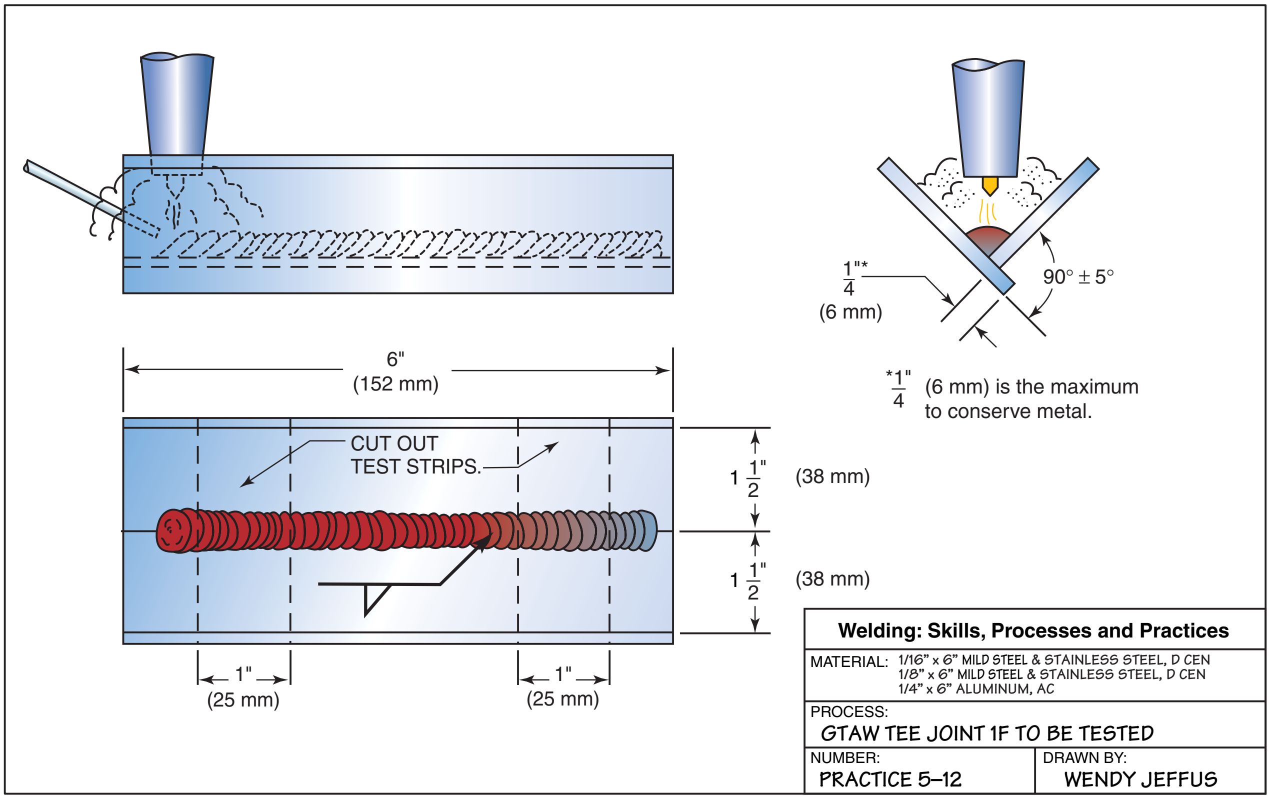
6"
(152 mm)
CUT OUT TEST STRIPS.
1 1/2"
(38 mm)
1 1/2"
(38 mm)
1"
(25 mm)
1"
(25 mm)
1/4"*
(6 mm)
90° ± 5°
*1/4" (6 mm) is the maximum to conserve metal.
Welding: Skills, Processes and Practices
MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC
PROCESS: GTAW TEE JOINT 1F TO BE TESTED
NUMBER: PRACTICE 5–12
DRAWN BY: WENDY JEFFUS

Figure 5.44
Tee joint in the flat position to be tested

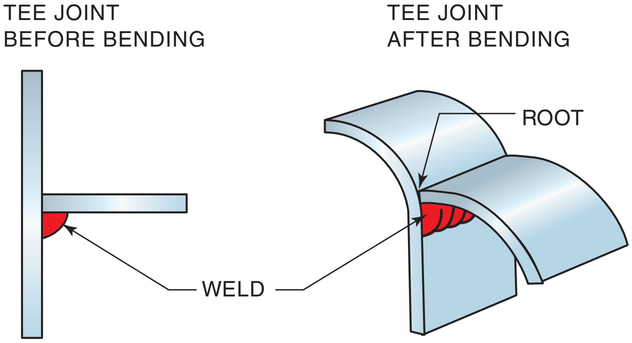

Figure 5.45
Bend the 1-in. (25-mm) strip of the tee joint backward and look at the root for 100% penetration

PRACTICE 5-13

Stringer Bead at a 45° Vertical Angle, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials listed in Practice 5-4, you will make a stringer bead in the vertical up position.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

- Starting at the bottom and welding in an upward direction, add the filler metal to the top edge of the weld pool and move the torch in a circle or "C" pattern, **Figure 5.46.** If the weld pool size starts to in-

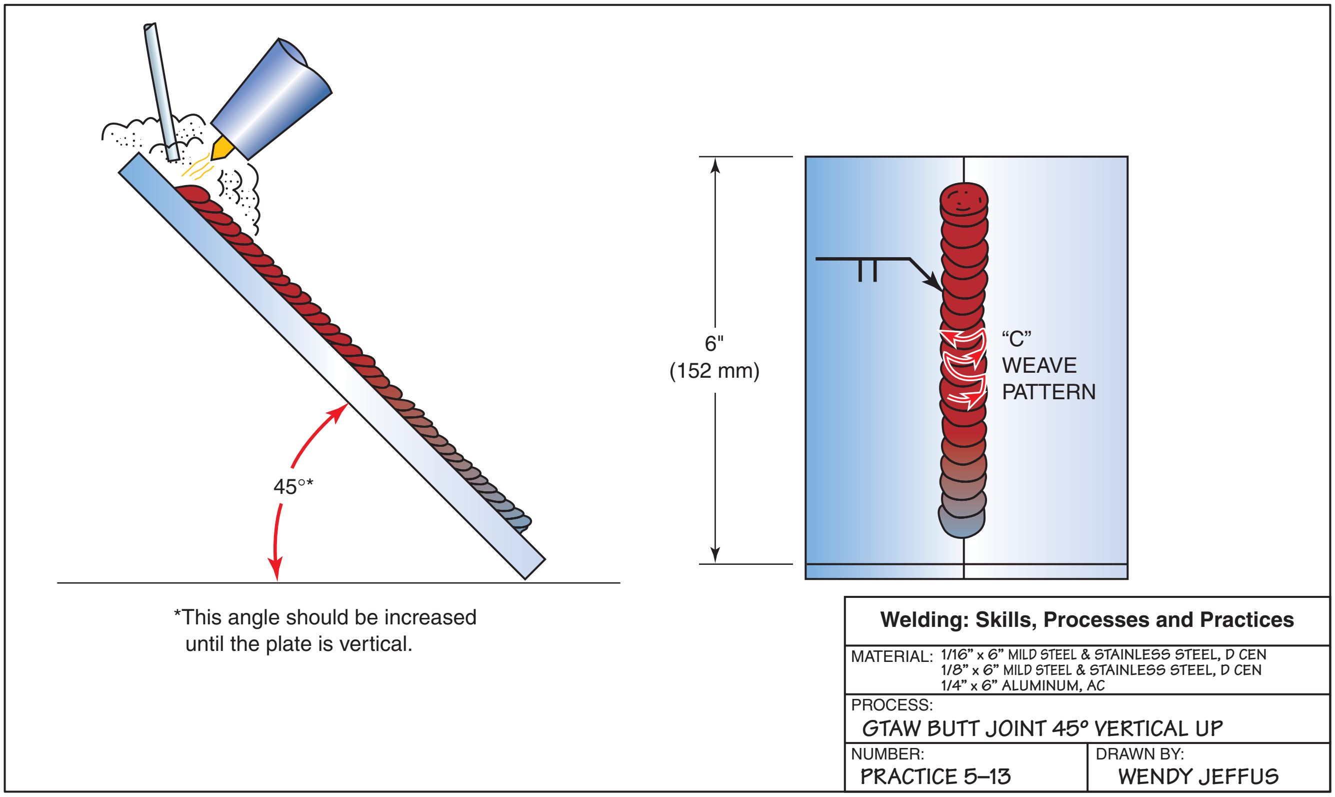

*This angle should be increased until the plate is vertical.

Welding: Skills, Processes and Practices
MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC
PROCESS: GTAW BUTT JOINT 45° VERTICAL UP
NUMBER: PRACTICE 5–13
DRAWN BY: WENDY JEFFUS

Figure 5.46
45° vertical up

crease, the “C” pattern can be increased in length or the power can be decreased.

- Watch the weld pool and establish a rhythm of torch movement and addition of rod to keep the weld uniform.

Module 7
Carbon Steel
Key Indicator 4
Austenitic Stainless Steel
Key Indicator 9
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 4
Austenitic Stainless Steel
Key Indicator 9
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

PRACTICE 5-14

Stringer Bead, 3G Position, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-13. Gradually increase the angle as you develop skill until the weld is being made in the vertical up position, **Figure 5.47.** Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

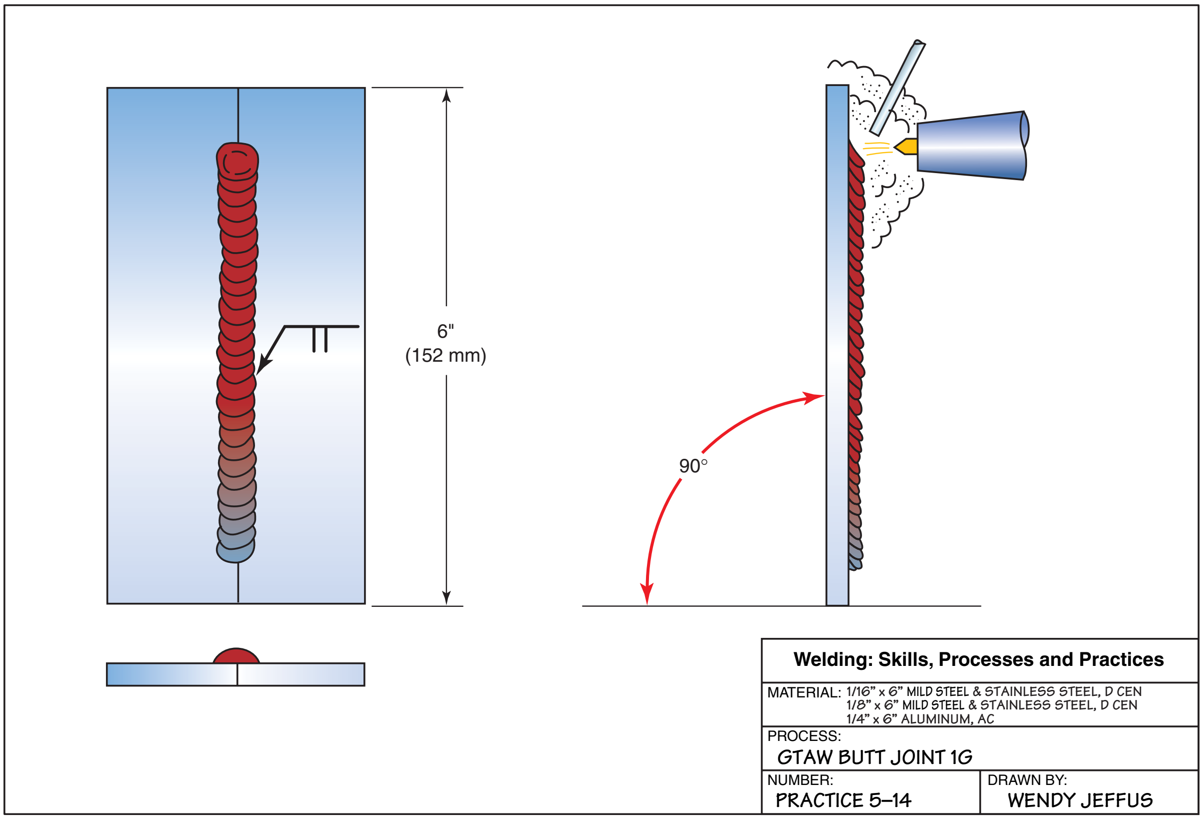
6"
(152 mm)
90°
Welding: Skills, Processes and Practices
MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC
PROCESS:
GTAW BUTT JOINT 1G
NUMBER:
PRACTICE 5–14
DRAWN BY:
WENDY JEFFUS

Figure 5.47
Vertical up position

PRACTICE 5-15

Butt Joint at a 45° Vertical Angle, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will weld a butt joint in the vertical up position.

After tack welding the plates together, start the weld at the bottom and weld in an upward direction. The same rhythmic torch and rod movement practiced for the 45° stringer bead should be used to control the weld.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 16

Module 9
Key Indicator 1, 2

PRACTICE 5-16

Butt Joint, 3G Position, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-15. Gradually increase the plate angle after each weld. As you develop skill, continue increasing the angle until the weld is being made in the vertical up position. Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 16

Module 9
Key Indicator 1, 2

PRACTICE 5-17

Butt Joint, 3G Position, with 100% Penetration, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-16. Make the needed changes in the root opening to allow 100% penetration, **Figure 5.48.** It may be necessary to provide a backing gas to protect the root from atmospheric contamination. After the weld is completed, visually inspect it for uniformity and defects. Then shear out strips 1 in. (25 mm) wide and bend-test them.

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 16

Module 9
Key Indicator 1, 2

PRACTICE 5-18

Lap Joint at a 45° Vertical Angle, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will make a vertical up fillet weld on a lap joint.

Module 1
Key Indicator 1, 3, 4

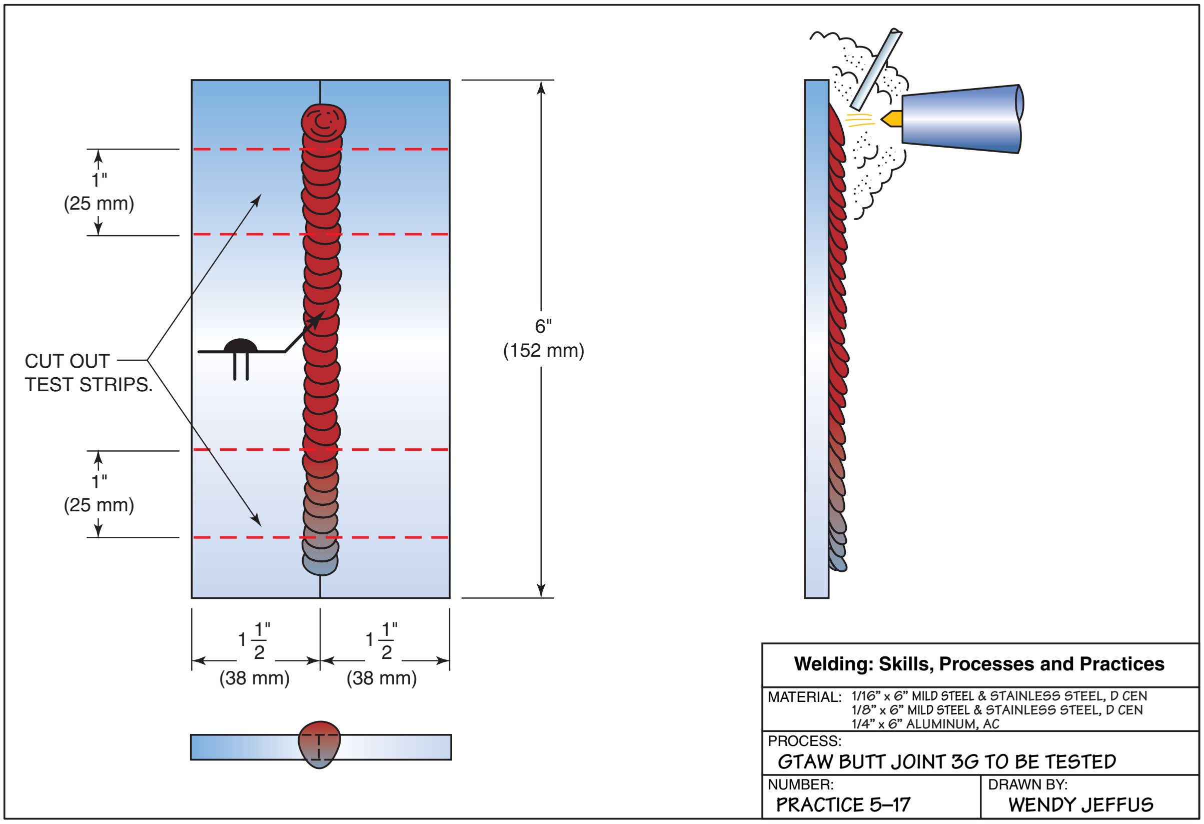
1"
(25 mm)
CUT OUT
TEST STRIPS.
1"
(25 mm)
6"
(152 mm)
$1\frac{1}{2}$"
(38 mm)
$1\frac{1}{2}$"
(38 mm)

Welding: Skills, Processes and Practices

MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC

PROCESS:
GTAW BUTT JOINT 3G TO BE TESTED

NUMBER: PRACTICE 5–17
DRAWN BY: WENDY JEFFUS

Figure 5.48
Vertical up butt joint with 100% penetration. Strips are to be cut for testing.

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Module 9
Key Indicator 1, 2

After tack welding the plates together, start the weld at the bottom and weld in an upward direction. It is important to maintain a uniform weld rhythm so that a nice-looking weld bead is formed. It may be necessary to move the torch in and around the base of the weld pool to ensure adequate root fusion, **Figure 5.49.** The filler metal should be added along the top edge of the weld pool near the top plate.

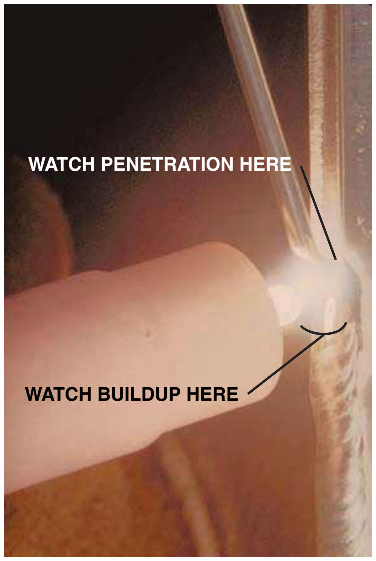
WATCH PENETRATION HERE
WATCH BUILDUP HERE

Figure 5.49
Lapjoint
Courtesy of Larry Jeffus

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-19

Lap Joint, 3F Position, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-18. Gradually increase the plate angle after each weld as you develop your skill. Increase the angle until the weld is being made in the vertical up position, **Figure 5.50.**

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Module 9
Key Indicator 1, 2

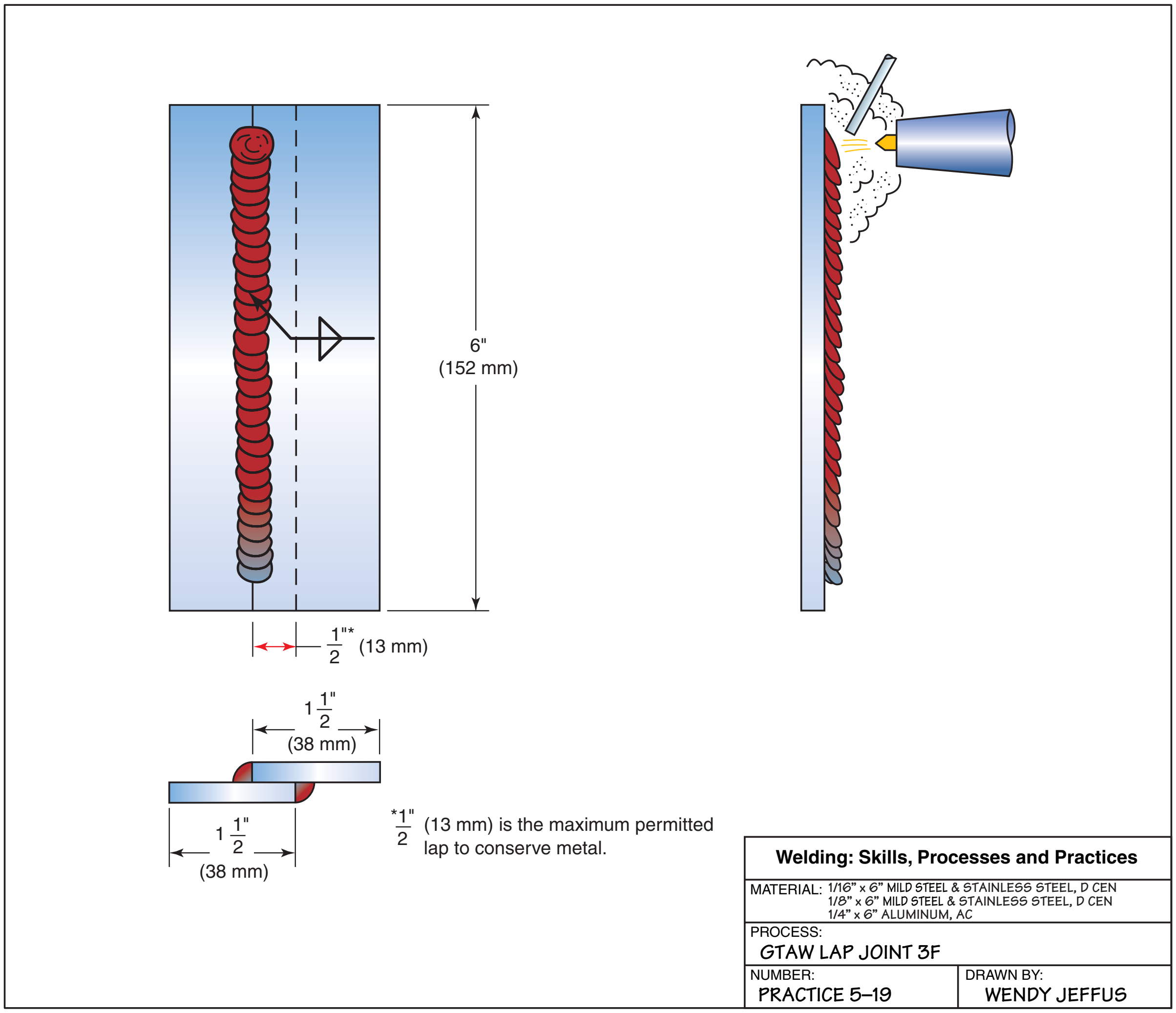
6"
(152 mm)
$\frac{1}{2}$"* (13 mm)
$1\frac{1}{2}$"
(38 mm)
$1\frac{1}{2}$"
(38 mm)
*$\frac{1}{2}$" (13 mm) is the maximum permitted lap to conserve metal.
Welding: Skills, Processes and Practices

MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC
PROCESS: GTAW LAP JOINT 3F
NUMBER: PRACTICE 5–19
DRAWN BY: WENDY JEFFUS

Figure 5.50
Vertical up lap joint

PRACTICE 5-20

Lap Joint, 3F Position, with 100% Root Penetration, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

Repeat Practice 5-19 and make the fillet weld in a lap joint with 100% root penetration. After the weld is completed, visually inspect it for uniformity and defects before shearing out strips 1 in. (25 mm) wide and bend-testing them.

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-21

Tee Joint at a 45° Vertical Angle, Using Mild Steel, Stainless Steel, Aluminum

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Module 9
Key Indicator 1, 2

Using the same equipment and materials listed in Practice 5-4, you will make a vertical up fillet weld on a tee joint.

- After tack welding the plates together, start the weld at the bottom and weld in an upward direction. The edge of the side plate, **Figure 5.51,** will heat up more quickly than the back plate. This rapid heating often leads to undercutting along this edge of the weld.
- To control undercutting, keep the arc on the back plate and add the filler metal to the weld pool near the side plate.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

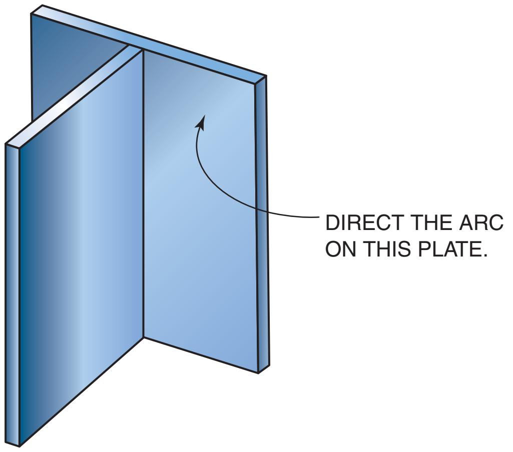

Figure 5.51
The edge of the intersecting plate will heat up faster than the base plate if the heat is not directed away from it

PRACTICE 5-22

Tee Joint, 3F Position, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-21. Gradually increase the plate angle after each weld as you develop your skill. Increase the angle until the weld is being made in the vertical up position.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

PRACTICE 5-23

Tee Joint, 3F Position, with 100% Root Penetration, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-21 and make the fillet weld in a lap joint with 100% root penetration. After the weld is completed, visually inspect it for uniformity and defects before shearing out strips 1 in. (25 mm) wide and bend-testing them.

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

PRACTICE 5-24

Stringer Bead at a 45° Reclining Angle, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will make a weld bead on a plate at a 45° reclining angle, **Figure 5.52.** Add the filler metal along the top leading edge of the weld pool. Surface tension will help hold the weld pool on the top if the bead is not too large. The weld should be uniform in width and reinforcement.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 16

Module 9
Key Indicator 1, 2

PRACTICE 5-25

Stringer Bead, 2G Position, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-24. Gradually increase the plate angle as you develop your skill until the weld is being made in the horizontal position on a vertical plate.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

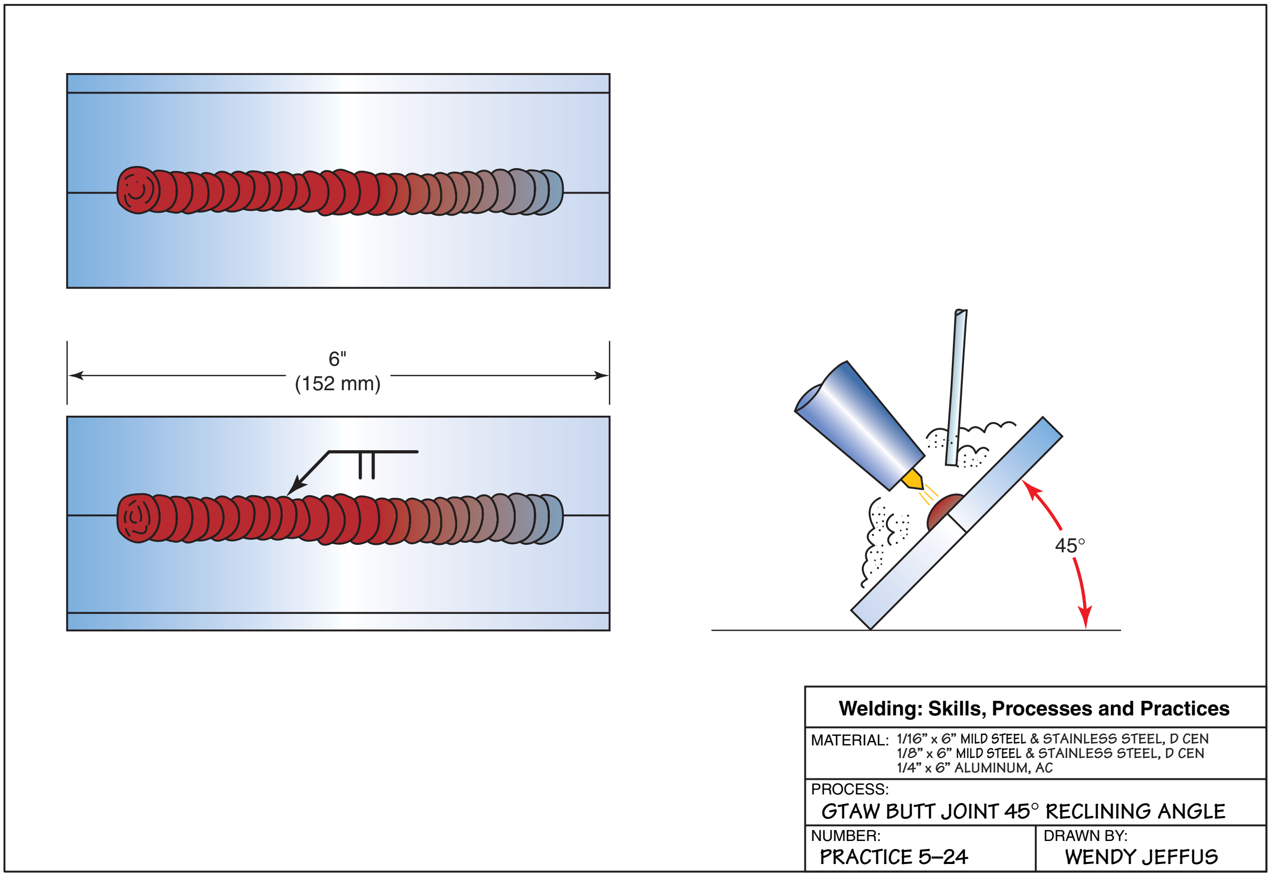
6"
(152 mm)
45°
Welding: Skills, Processes and Practices
MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC
PROCESS:
GTAW BUTT JOINT 45° RECLINING ANGLE
NUMBER:
PRACTICE 5–24
DRAWN BY:
WENDY JEFFUS

Figure 5.52
45° reclining angle

Module 7
Carbon Steel
Key Indicator 3, 4
Austenitic Stainless Steel
Key Indicator 8, 9
Aluminum
Key Indicator 13, 14

Module 9
Key Indicator 1, 2

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-26

Butt Joint, 2G Position, Using Mild Steel, Stainless Steel, Aluminum

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 16

Module 9
Key Indicator 1, 2

Using the same equipment and materials as listed in Practice 5-4, you will weld a butt joint in the horizontal position.

The welding techniques are the same as those used in Practice 5-25. Add the filler metal to the top plate, and keep the bead size small so it will be uniform.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-27

Butt Joint, 2G Position, with 100% Penetration, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-26. It may be necessary to increase the root opening to ensure 100% penetration. A backing gas may be required to prevent atmospheric contamination. Using a "J" weave pattern will help to maintain a uniform weld bead. After the weld is completed, visually inspect it for uniformity and defects. Then shear out strips 1 in. (25 mm) wide and bend-test them.

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 11
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

PRACTICE 5-28

Lap Joint, 2F Position, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, make a horizontal fillet weld on a lap joint.

- After tack welding the plates together, start the weld at one end. The bottom plate will act as a shelf to support the molten weld pool, **Figure 5.53.**
- Add the filler metal along the top edge of the weld pool to help control undercutting.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 14, 15

Module 9
Key Indicator 1, 2

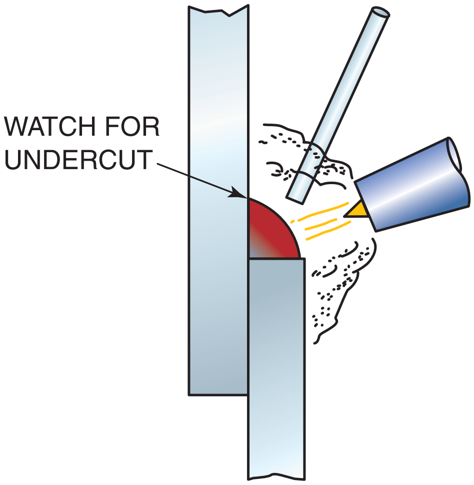

Figure 5.53
Horizontal lap joint

PRACTICE 5-29

Lap Joint, 2F Position, with 100% Root Penetration, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Module 9
Key Indicator 1, 2

Repeat Practice 5-28. Be sure the weld is penetrating the root 100%. After the weld is completed, visually inspect it for uniformity and defects before shearing out strips 1 in. (25 mm) wide and bend-testing them.

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 5-30

Tee Joint, 2F Position, Using Mild Steel, Stainless Steel, Aluminum

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 10
Aluminum
Key Indicator 15

Using the same equipment and materials as listed in Practice 5-4, you will make a horizontal fillet weld on a tee joint.

After tack welding the plates together, start the weld at one end. The bottom plate will act as a shelf to support the molten weld pool, **Figure 5.54.** As with the horizontal lap joint, add the filler metal along the top leading edge of the weld pool. This will help control undercut.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

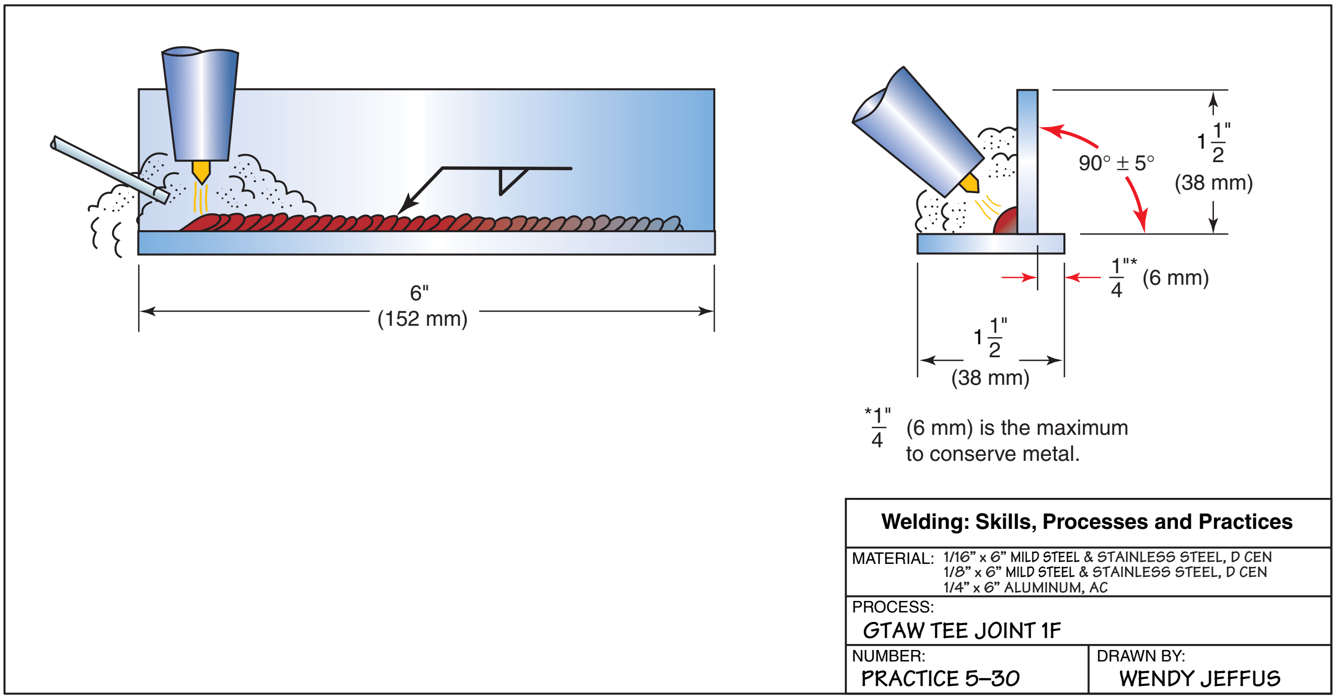

Welding: Skills, Processes and Practices
MATERIAL: 1/16” x 6” MILD STEEL & STAINLESS STEEL, D CEN
1/8” x 6” MILD STEEL & STAINLESS STEEL, D CEN
1/4” x 6” ALUMINUM, AC
PROCESS: GTAW TEE JOINT 1F
NUMBER: PRACTICE 5–30
DRAWN BY: WENDY JEFFUS

Figure 5.54
Tee joint in the horizontal position

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 9
Key Indicator 1, 2

PRACTICE 5-31

Tee Joint, 2F Position, with 100% Root Penetration, to Be Tested, Using Mild Steel, Stainless Steel, Aluminum

Repeat Practice 5-30. Be sure the weld is penetrating the root 100%. After the weld is completed, visually inspect it for uniformity and defects before shearing out strips 1 in. (25 mm) wide and bend-testing them.

Repeat each weld until all have 100% root penetration. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 4
Austenitic Stainless Steel
Key Indicator 9
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

PRACTICE 5-32

Stringer Bead, 4G Position, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will make a weld bead on a plate in the overhead position.

The surface tension on the molten metal will hold the welding bead on the plate providing that it is not too large. A wide weld with little buildup will be easier to control and less likely to undercut along the edge.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 6
Austenitic Stainless Steel
Key Indicator 9, 10
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

PRACTICE 5-33

Butt Joint, 4G Position, Using Mild Steel, Stainless Steel, Aluminum

Using the same equipment and materials as listed in Practice 5-4, you will weld a butt joint in the overhead position.

The same techniques used to make the stringer beads in Practice 5-32 are also used with the butt joint. The size of the bead should be kept small enough so that you can control the weld. Add the filler metal along the leading edge of the weld pool, **Figure 5.55.** The completed weld should be uniform and free from defects.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 9
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

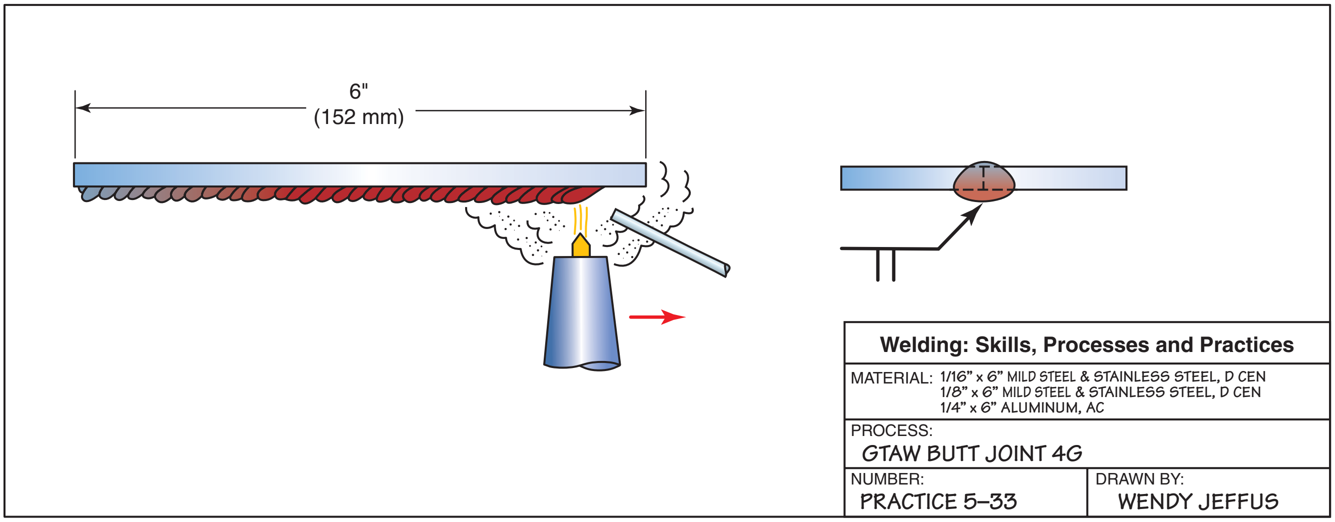

Welding: Skills, Processes and Practices

MATERIAL: 1/16" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/8" x 6" MILD STEEL & STAINLESS STEEL, D CEN
1/4" x 6" ALUMINUM, AC

PROCESS:
GTAW BUTT JOINT 4G

NUMBER:
PRACTICE 5–33

DRAWN BY:
WENDY JEFFUS

Figure 5.55
Overhead butt joint

PRACTICE 5-34

Lap Joint, 4F Position, Using Mild Steel, Stainless Steel, Aluminum

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Carbon Steel
Key Indicator 5
Austenitic Stainless Steel
Key Indicator 9
Aluminum
Key Indicator 14

Module 9
Key Indicator 1, 2

Using the same equipment and materials as listed in Practice 5-4, you will make a fillet weld on a lap joint in the overhead position.

The major concentration of heat and filler metal should be on the top plate. Gravity and an occasional sweep of the torch along the bottom plate will pull the weld pool down. Undercutting along the top edge of the weld can be controlled by putting most of the filler metal along the top edge. The completed weld should be uniform and free from defects.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-35

Tee Joint, 4F Position, Using Mild Steel, Stainless Steel, Aluminum

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Key Indicator 1, 2
Carbon Steel
Key Indicator 3, 4, 5, 6, 7

Module 9
Key Indicator 1, 2

Using the same equipment and materials as listed in Practice 5-4, you will make a fillet weld on a tee joint in the overhead position.

The same techniques used to make the overhead lap weld in Practice 5-34 are used with the tee joint. As with the lap joint, most of the heat and filler metal should be concentrated on the top plate. A "J" weave pattern will help pull down any needed metal to the side plate. The completed weld should be uniform and free from defects.

Repeat the process using all thicknesses of metal until you can consistently make the weld visually defect free. Turn off the welding machine, shielding gas, and cooling water, and clean up your work area when you are finished welding.

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-36

Gas Tungsten Arc Welding (GTAW) on Plain Carbon Steel Workmanship Sample

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 5-36. Date:

Title:

Welding GTAW of sheet to sheet.

Scope:

This procedure is applicable for square groove and fillet welds within the range of 18 gauge through 10 gauge.

Welding may be performed in the following positions: 1G and 2F.

Base Metal:

The base metal shall conform to carbon steel M-1, Group 1.
Backing material specification: none .

Filler Metal:

The filler metal shall conform to AWS specification no. 1/16 in. (2 mm) to 3/32 in. (2.4 mm) diameter. E70S-3 from AWS specification A5.18. This filler metal falls into F-number F-6 and A-number A-1.

Electrode:

The tungsten electrode shall conform to AWS specification no. EWTh-2, EWCe-2, or EWLa from AWS specification A5.12. The tungsten diameter shall be 1/8 in. (3.2 mm) maximum.

The tungsten end shape shall be tapered at two to three times its length to its diameter.

Shielding Gas:

The shielding gas, or gases, shall conform to the following compositions and purity: welding grade argon.

Joint Design and Tolerances:

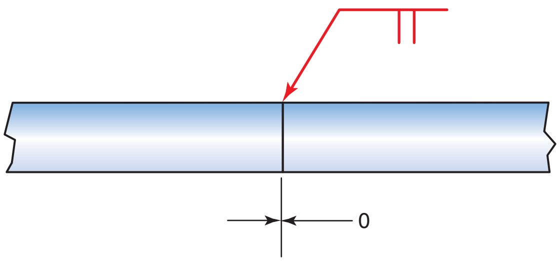

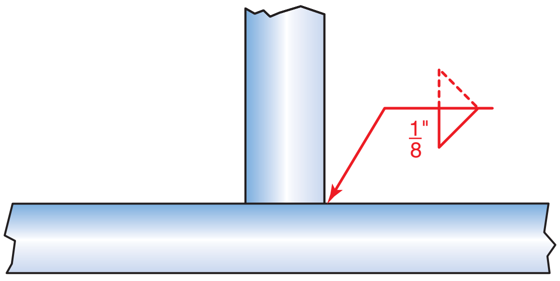

Preparation of Base Metal:

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The joint face and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically cleaned of slag, rust, and mill scale. Cleaning must be done with a wire brush or grinder down to bright metal.

Electrical Characteristics:

The current shall be direct current electrode negative (DCEN). The base metal shall be on the positive side of the line.

Metal Specifications		Gas Flow			Nozzle Size in. (mm)	Amperage Min. Max.
Thickness	Diameter of E70S-3*	Rates cfm) (L/min)	Flow Times Preflow	Postflow		
18 ga	1/16 in. (2 mm)	15 to 20 (7 to 9)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	45 to 65
17 ga	1/16 in. (2 mm)	15 to 20 (7 to 9)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	45 to 70
16 ga	1/16 in. (2 mm)	15 to 20 (7 to 9)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	50 to 75
15 ga	1/16 in. (2 mm)	15 to 20 (7 to 9)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	55 to 80
14 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	60 to 90
13 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	60 to 100
12 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	60 to 110
11 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	65 to 120
10 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	70 to 130

*Other E70S-X filler metal may be used.

Preheat:

The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:

None

Safety:

Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:

TACK WELDS: With the parts securely clamped in place with the correct root gap, the tack welds are to be performed. Holding the electrode so that it is very close to the root face but not touching, slowly increase the current until the arc starts and a molten weld pool is formed. Add filler metal as required to maintain a slightly convex weld face and a flat or slightly concave root face. When it is time to end the tack weld, lower the current slowly so that the molten weld pool can be tapered down in size. When all tack welds are complete, allow the parts to cool as needed before assembling the remaining parts. Repeat the tack welding procedure until the entire part is assembled.

SQUARE GROOVE AND FILLET WELDS: Holding the electrode so that it is very close to the metal surface but not touching, slowly increase the

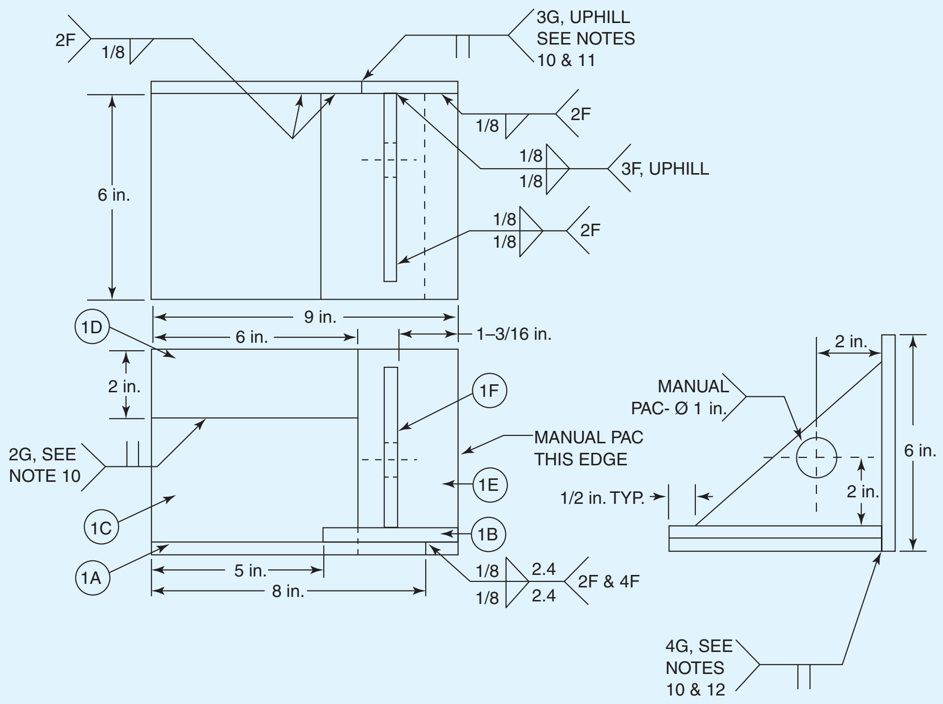

NOTES:

1. All dimensions U.S. Customary Units unless otherwise specified.
2. 10 ga -14 ga thickness carbon steel. Optional choice of thickness within range specified.
3. The welder shall prepare a bill of materials in U.S. Customary Units prior to cutting.
4. The welder shall convert the above bill of materials to S.I. Metric Units of measure.
5. All parts may be mechanically cut or machine PAC unless specified manual PAC.
6. All welds GMAW-S (Short Circuiting Transfer) or GTAW as applicable.
7. Fit and tack entire assembly on bench before attaching to positioning arm.
8. All welding to be done in position according to welding symbol.
9. Employ boxing technique where applicable.
10. Melt through not required.
11. Weld joins parts 1C and 1D to 1E.
12. Weld joins parts 1C and 1E to 1A.
13. For GMAW-S, use WPS AWS EDU GMAW-01. (See AWS QC10, Table 2.)
14. For GTAW use WPS AWS EDU GTAW-01. (See AWS QC10, Table 2.)
15. Visual examination in accordance with requirements of AWS QC10, Table 3.

ID	QTY	SIZE	METRIC CONVERSION

American Welding Society

Entry Welder Performance Qualification

GMAW-S, GTAW Carbon Steel

DATE:	SCALE:	DWG #: AWS EDU-3
DR BY:	Tolerances: (Unless otherwise specified) DRAWING NOT TO SCALE Fractions: ±1/16” Angles: +10°, −5°	
APP BY:		

Figure 5.56
GTAW Plain Carbon Steel Workmanship Qualification Test
Courtesy of the American Welding Society

current until the arc starts and a molten weld pool is formed. As the weld progresses, add filler metal as required to maintain a flat or slightly convex weld face. If it is necessary to stop the weld or to reposition yourself or if the weld is completed, the current must be lowered slowly so that the molten weld pool can be tapered down in size.

Interpass Temperature:

The plate should not be heated to a temperature higher than 120°F (49°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:

Recleaning may be required if the parts or filler metal becomes contaminated or reoxides to a degree that the weld quality will be affected. Reclean using the same procedure used for the original metal preparation. Any slag must be cleaned off between passes.

Visual Inspection:

Visual inspection criteria for entry welders*:

1. There shall be no cracks, no incomplete fusion.
2. There shall be no incomplete joint penetration in groove welds except as permitted for partial joint penetration groove welds.
3. The Test Supervisor shall examine the weld for acceptable appearance, and shall be satisfied that the welder is skilled in using the process and procedure specified for the text.
4. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm).
5. Where visual examination is the only criterion for acceptance, all weld passes are subject to visual examination, at the discretion of the Test Supervisor.
6. The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length and the maximum diameter shall not exceed 3/32 in. (2.4 mm).
7. Welds shall be free from overlap.

Sketches:

Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

PRACTICE 5-37

Gas Tungsten Arc Welding (GTAW) on Stainless Steel Workmanship Sample

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 5-37. Date:

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Key Indicator 1, 2
Austenitic Stainless Steel
Key Indicator 8, 9, 10, 11, 12

Module 9
Key Indicator 1, 2

Title:

Welding GTAW of sheet to sheet.

Scope:

This procedure is applicable for square groove and fillet welds within the range of 18 gauge through 10 gauge.

Welding may be performed in the following positions: 1G and 2F.

Base Metal:

The base metal shall conform to austenitic stainless steel M-8 or P-8. Backing material specification: none.

Filler Metal:

The filler metal shall conform to AWS specification no. ER3XX from AWS specificationA5.9.ThisfillermetalfallsintoF-numberF-6andA-numberA-8.

Electrode:

The tungsten electrode shall conform to AWS specification no. EWTh-2, EWCe-2 or EWLa from AWS specification A5.12. The tungsten diameter shall be 1/8 in. (3.2 mm) maximum. The tungsten end shape shall be tapered at two to three times its length to its diameter.

Shielding Gas:

The shielding gas, or gases, shall conform to the following compositions and purity: welding grade argon.

---

*Courtesy of the American Welding Association.

Joint Design and Tolerances:

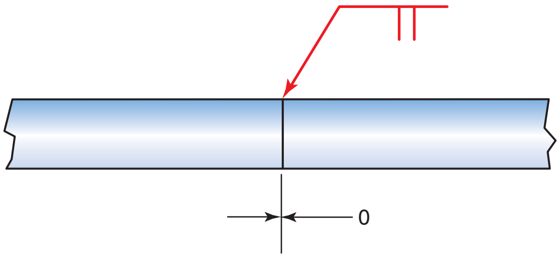

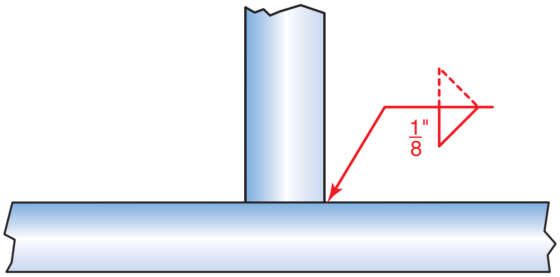

Preparation of Base Metal:

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The joint face and inside and outside plate surface within 1 in. (25 mm) of the joint must be cleaned of slag, oxide, and scale. Cleaning can be mechanical or chemical. Mechanical metal cleaning can be done by grinding, stainless steel wire brushing, scraping, machining, or filing. Chemical cleaning can be done by using acids, alkalies, solvents, or detergents. Cleaning must be done down to bright metal.

Electrical Characteristics:

The current shall be direct current electrode negative (DCEN). The base metal shall be on the positive side of the line.

Metal Specifications		Gas Flow			Nozzle Size in. (mm)	Amperage Min. Max.
Thickness	Diameter of ER3XX*	Rates cfm (L/min	Flow Times Preflow	Postflow		
18 ga	1/16 in. (2 mm)	15 to 20 (7 to 9)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	35 to 60
17 ga	1/16 in. (2 mm)	15 to 20 (7 to 9)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	40 to 65
16 ga	1/16 in. (2 mm)	15 to 20 (7 to 9)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	40 to 75
15 ga	1/16 in. (2 mm)	15 to 20 (7 to 9)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	50 to 80
14 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	50 to 90
13 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	55 to 100
12 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	60 to 110
11 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	65 to 120
10 ga	3/32 in. (2.4 mm)	20 to 25 (9 to 12)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	70 to 130

*Any ER3XX stainless steel A5.9 filler metal may be used.

Preheat:

The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Backing Gas:

None

Safety:

Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:

TACK WELDS: With the parts securely clamped in place with the correct root gap, the tack welds are to be performed. Holding the electrode so that it is very close to the root face but not touching, slowly increase the current until the arc starts and a molten weld pool is formed. Add filler metal as required to maintain a slightly convex weld face and a flat or slightly concave root face. When it is time to end the tack weld, lower the current slowly so that the molten weld pool can be tapered down in size. When all tack welds are complete, allow the parts to cool as needed before assembling the remaining parts. Repeat the tack welding procedure until the entire part is assembled.

SQUARE GROOVE AND FILLET WELDS: Holding the electrode so that it is very close to the metal surface but not touching, slowly increase the current until the arc starts and a molten weld pool is formed. As the weld progresses, add filler metal as required to maintain a flat or slightly convex weld face. If it is necessary to stop the weld or to reposition yourself or if the weld is completed, the current must be lowered slowly so that the molten weld pool can be tapered down in size.

Interpass Temperature:

The plate should not be heated to a temperature higher than 350°F (180°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

Cleaning:

Recleaning may be required if the parts or filler metal become contaminated or oxidized to a degree that the weld quality will be affected. Reclean using the same procedure used for the original metal preparation.

Visual Inspection:

Visual inspection criteria for entry welders*:

1. There shall be no cracks, no incomplete fusion.
2. There shall be no incomplete joint penetration in groove welds except as permitted for partial joint penetration groove welds.
3. The Test Supervisor shall examine the weld for acceptable appearance, and shall be satisfied that the welder is skilled in using the process and procedure specified for the text.
4. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm).
5. Where visual examination is the only criterion for acceptance, all weld passes are subject to visual examination, at the discretion of the Test Supervisor.
6. The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length and the maximum diameter shall not exceed 3/32 in. (2.4 mm).
7. Welds shall be free from overlap.

*Courtesy of the American Welding Association.

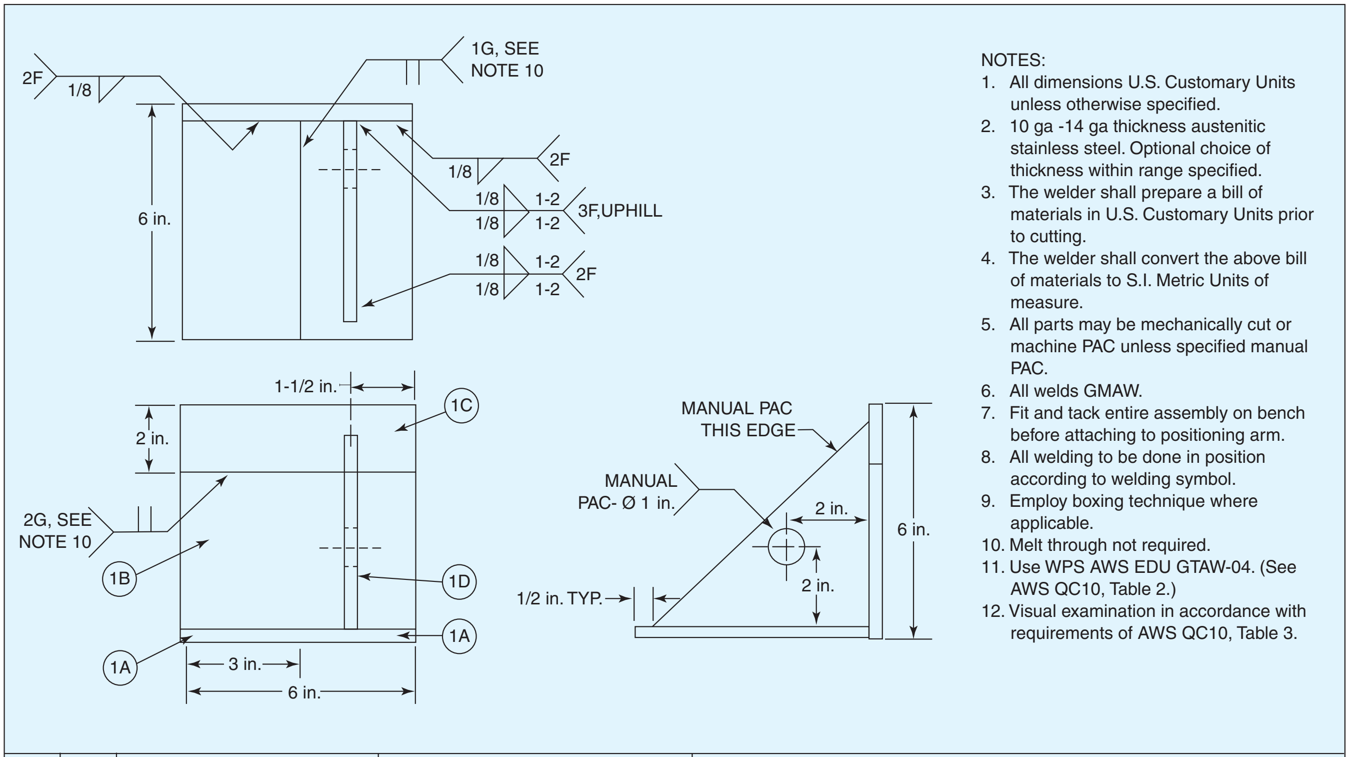

NOTES:
1. All dimensions U.S. Customary Units unless otherwise specified.
2. 10 ga -14 ga thickness austenitic stainless steel. Optional choice of thickness within range specified.
3. The welder shall prepare a bill of materials in U.S. Customary Units prior to cutting.
4. The welder shall convert the above bill of materials to S.I. Metric Units of measure.
5. All parts may be mechanically cut or machine PAC unless specified manual PAC.
6. All welds GMAW.
7. Fit and tack entire assembly on bench before attaching to positioning arm.
8. All welding to be done in position according to welding symbol.
9. Employ boxing technique where applicable.
10. Melt through not required.
11. Use WPS AWS EDU GTAW-04. (See AWS QC10, Table 2.)
12. Visual examination in accordance with requirements of AWS QC10, Table 3.

ID	QTY	SIZE	METRIC CONVERSION	American Welding Society		
				Entry Welder Performance Qualification		
				GTAW Austenitic Stainless Steel		
				DATE:	SCALE:	DWG #: AWS EDU-3
				DR BY:	Tolerances: (Unless otherwise specified) DRAWING NOT TO SCALE Fractions: ±1/16” Angles: +10°, −5°	
				APP BY:		

Figure 5.57
GTAW Stainless Steel Workmanship Qualification Test
Courtesy of the American Welding Society

Sketches:

Complete a copy of the “Student Welding Report” listed in Appendix I or provided by your instructor.

PRACTICE 5-38

Gas Tungsten Arc Welding (GTAW) on Aluminum Workmanship Sample

Welding Procedure Specification (WPS)

Welding Procedure Specification No.: Practice 5-38. Date:

Title:

Welding GTAW of sheet to sheet.

Scope:

This procedure is applicable for square groove and fillet welds within the range of 18 gauge through 10 gauge.

Welding may be performed in the following positions: 1G and 2F.

Module 1
Key Indicator 1, 3, 4

Module 2
Key Indicator 1, 2, 3, 4, 6, 7

Module 7
Key Indicator 1, 2
Aluminium
Key Indicator 13, 14, 15, 16, 17

Module 9
Key Indicator 1, 2

Base Metal:

The base metal shall conform to aluminum M-22 or P-22.
Backing material specification: none.

Filler Metal:

The filler metal shall conform to AWS specification no. ER4043 from AWS specification A5.10. This filler metal falls into F-number F-23 and A-number.

Electrode:

The tungsten electrode shall conform to AWS specification No. EWCe-2, EWZr, EWLa, or EWP from AWS specification A5.12. The tungsten diameter shall be 1/8 in. (3.2 mm) maximum. The tungsten end shape shall be rounded.

Shielding Gas:

The shielding gas, or gases, shall conform to the following compositions and purity: welding grade argon.

Joint Design and Tolerances:

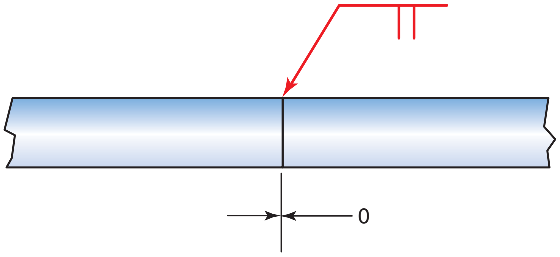

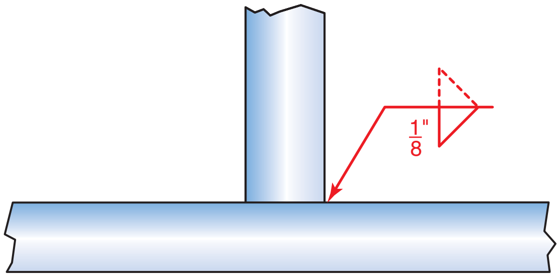

Preparation of Base Metal:

All hydrocarbons and other contaminations, such as cutting fluids, grease, oil, and primers, must be cleaned off all parts and filler metals before welding. This cleaning can be done with any suitable solvents or detergents. The joint face and inside and outside plate surface within 1 in. (25 mm) of the joint must be mechanically or chemically cleaned of oxides. Mechanical cleaning may be done by stainless steel wire brushing, scraping, machining, or filing. Chemical cleaning may be done by using acids, alkalies, solvents, or detergents. Because the oxide layer may reform quickly and affect the weld, welding should be started within 10 minutes of cleaning.

Electrical Characteristics:

The current shall be alternating current high-frequency stabilized (balanced wave preferably). The base metal shall be on the N/A side of the line.

Gas Flow

Preheat:

The parts must be heated to a temperature higher than 50°F (10°C) before any welding is started.

Metal Specifications		Gas Flow			Nozzle Size in. (mm)	Amperage Min. Max.
Thickness	Diameter of ER4043[*]	Rates cfm (L/min)	Flow Times Preflow	Postflow		
18 ga	3/32 in. (2.4 mm)	20 to 30 (9 to 14)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	40 to 60
17 ga	3/32 in. (2.4 mm)	20 to 30 (9 to 14)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	50 to 70
16 ga	3/32 in. (2.4 mm)	20 to 30 (9 to 14)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	60 to 75
15 ga	3/32 in. (2.4 mm)	20 to 30 (9 to 14)	10 to 15 sec	10 to 25 sec	1/4 to 3/8 (6 to 10)	65 to 85
14 ga	3/32 in. (2.4 mm)	20 to 30 (9 to 14)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	75 to 90
13 ga	1/8 in. (3 mm)	25 to 40 (12 to 19)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	85 to 100
12 ga	1/8 in. (3 mm)	25 to 40 (12 to 19)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	90 to 110
11 ga	1/8 in. (3 mm)	25 to 40 (12 to 19)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	100 to 115
10 ga	1/8 in. (3 mm)	25 to 40 (12 to 19)	10 to 20 sec	10 to 30 sec	3/8 to 5/8 (10 to 16)	100 to 125

[*] Other aluminum AWS A5.10 filler metal may be used if needed.

Backing Gas:

N/A

Safety:

Proper protective clothing and equipment must be used. The area must be free of all hazards that may affect the welder or others in the area. The welding machine, welding leads, work clamp, electrode holder, and other equipment must be in safe working order.

Welding Technique:

The welder's hands or gloves must be clean and oil free to prevent contamination of the metal or filler rods.

TACK WELDS: With the parts securely clamped in place with the correct root gap, the tack welds are to be performed. Holding the electrode so that it is very close to the root face but not touching, slowly increase the current until the arc starts and a molten weld pool is formed. Add filler metal as required to maintain a slightly convex weld face and a flat or slightly concave root face. When it is time to end the tack weld, lower the current slowly so that the molten weld pool can be tapered down in size. When all tack welds are complete, allow the parts to cool as needed before assembling the remaining parts. Repeat the tack welding procedure until the entire part is assembled.

SQUARE GROOVE AND FILLET WELDS: Holding the electrode so that it is very close to the metal surface but not touching, slowly increase the current until the arc starts and a molten weld pool is formed. As the weld progresses, add filler metal as required to maintain a flat or slightly convex weld face. If it is necessary to stop the weld or to reposition yourself or the weld is completed, the current must be lowered slowly so that the molten weld pool can be tapered down in size.

Interpass Temperature:

The plate should not be heated to a temperature higher than 120°F (49°C) during the welding process. After each weld pass is completed, allow it to cool but never to a temperature below 50°F (10°C). The weldment must not be quenched in water.

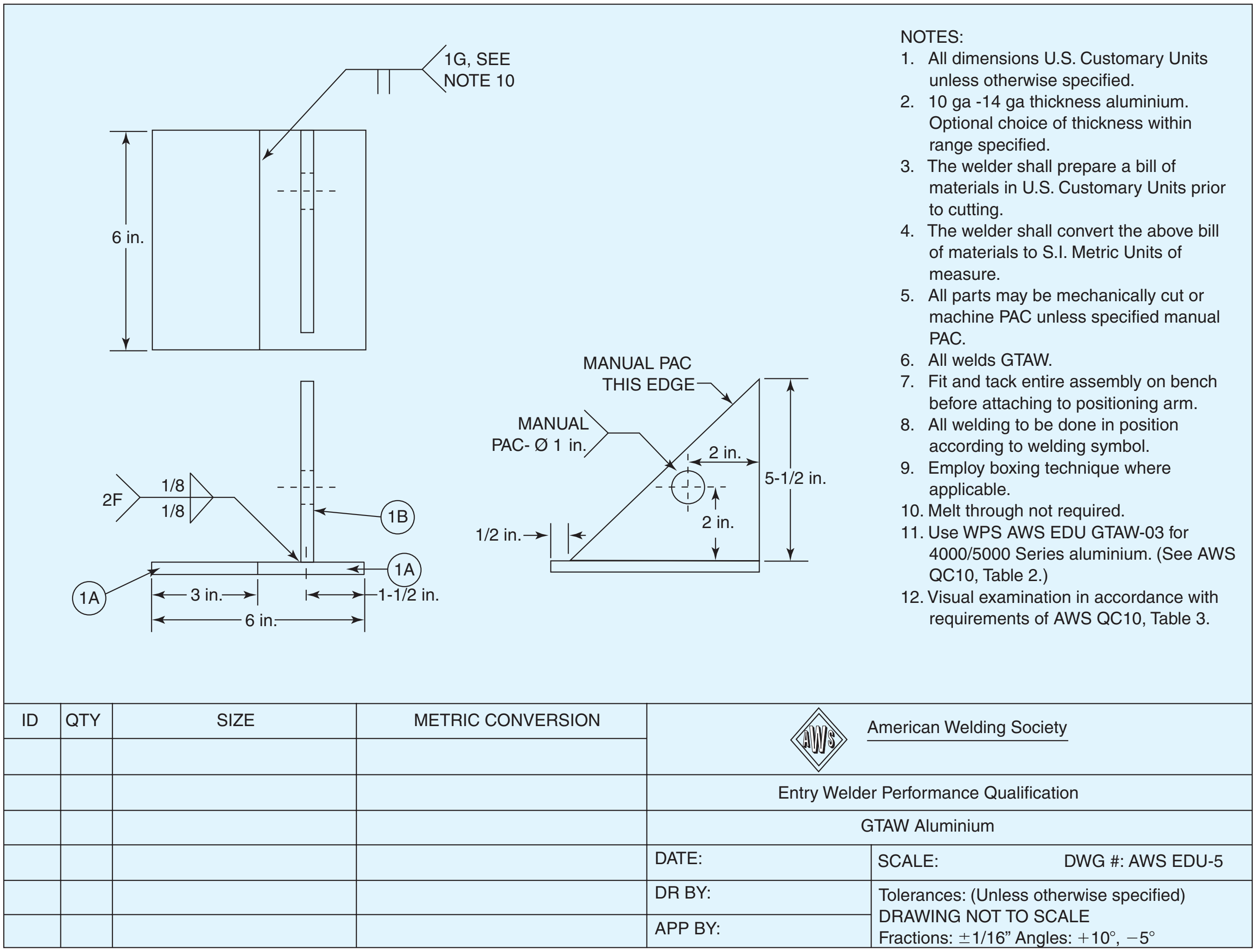

NOTES:

1. All dimensions U.S. Customary Units unless otherwise specified.
2. 10 ga -14 ga thickness aluminium. Optional choice of thickness within range specified.
3. The welder shall prepare a bill of materials in U.S. Customary Units prior to cutting.
4. The welder shall convert the above bill of materials to S.I. Metric Units of measure.
5. All parts may be mechanically cut or machine PAC unless specified manual PAC.
6. All welds GTAW.
7. Fit and tack entire assembly on bench before attaching to positioning arm.
8. All welding to be done in position according to welding symbol.
9. Employ boxing technique where applicable.
10. Melt through not required.
11. Use WPS AWS EDU GTAW-03 for 4000/5000 Series aluminium. (See AWS QC10, Table 2.)
12. Visual examination in accordance with requirements of AWS QC10, Table 3.

ID	QTY	SIZE	METRIC CONVERSION

American Welding Society

Entry Welder Performance Qualification

GTAW Aluminium

DATE:

SCALE: DWG #: AWS EDU-5

DR BY:

APP BY:

Tolerances: (Unless otherwise specified)
DRAWING NOT TO SCALE
Fractions: ±1/16” Angles: +10°, −5°

Figure 5.58
GTAW Aluminum Workmanship Qualification Test
Courtesy of the American Welding Society

Cleaning:

Recleaning may be required if the parts or filler metal becomes contaminated or oxidizes to a degree that the weld quality will be affected. Reclean using the same procedure used for the original metal preparation.

Visual Inspection:

Visual inspection criteria for entry welders*:

1. There shall be no cracks, no incomplete fusion.
2. There shall be no incomplete joint penetration in groove welds except as permitted for partial joint penetration groove welds.
3. The Test Supervisor shall examine the weld for acceptable appearance, and shall be satisfied that the welder is skilled in using the process and procedure specified for the text.
4. Undercut shall not exceed the lesser of 10% of the base metal thickness or 1/32 in. (0.8 mm)
5. Where visual examination is the only criterion for acceptance, all weld passes are subject to visual examination, at the discretion of the Test Supervisor.

*Courtesy of the American Welding Association.

6. The frequency of porosity shall not exceed one in each 4 in. (100 mm) of weld length and the maximum diameter shall not exceed 3/32 in. (2.4 mm).
7. Welds shall be free from overlap.

Sketches:
Complete a copy of the "Student Welding Report" listed in Appendix I or provided by your instructor.

SUMMARY

One of the most difficult aspects of learning to produce gas tungsten arc welds is positioning yourself so that you can control the electrode filler metal and see the joint both at the same time. Beginning welders assume they must see the tungsten tip as they make the weld. Experienced welders, however, realize that they need to see only the leading edge of the molten weld pool to know how the weld is being produced. A good view of the leading edge will tell you how the base metal is being melted, or depth of penetration. You can even tell from this small portion whether the filler metal is being added at an appropriate rate. As you learn how to control the weld and develop your skills, it is a good idea to gradually reduce your need for seeing 100% of the molten weld pool. Increasing this part of your skill will be a significant advantage in the field because, unlike welding in the classroom, which is typically done on a comfortable-height table, welding in the field may have to be done out of position.

REVIEW

1. What effect does torch angle have on the shielding gas protective zone?
2. Why must the end of the filler rod be kept in the shielding gas protective zone?
3. What can cause tungsten contamination?
4. What determines the correct current setting for a GTA weld?
5. What is the lowest acceptable amperage setting for GTA welding?
6. List the factors that affect the gas flow setting for GTA welding.
7. When should the minimum gas flow rates be increased?
8. What is the minimum gas flow rate for a nozzle size?
9. What is the maximum gas flow rate for a nozzle size?
10. Which incorrect welding parameters does stainless steel show clearly?
11. Using **Table 5.4,** determine the approximate temperature of metal that has formed a dark blue color.
12. Using **Table 5.3, Table 5.5,** and **Table 5.6,** list the filler metals for the following metals.
    a. 1020 low-carbon steel
    b. 309 stainless steel
13. Why is it possible to control a large aluminum weld bead?
14. What may happen to the end of the aluminum welding rod if it is held too close to the arc?
15. What should be done if someone comes in contact with a cleaning chemical?

16. Using **Table 5.7,** determine the suggested setting for GTA welding of mild steel using a 3/32-in. (2.4-mm) tungsten.
17. What can be done to limit oxide formation on stainless steel?
18. How should the filler metal be added to the molten weld pool?
19. How can the rod be freed if it sticks to the plate?
20. How is an outside corner joint assembled?
21. What must be done with the weld craters when back stepping a weld? Why?
22. How is the lap joint tested for 100% penetration?
23. What can prevent both sides of a stainless steel tee joint from being welded?
24. How is the filler metal added for a 3F weld?
25. What can cause undercutting on a 3F tee joint?
26. What helps hold the weld in place on a 2F lap joint?
27. What helps hold the weld in place on a 4G weld?

APPENDIX I—STUDENT WELDING REPORT

I. STUDENT WELDING REPORT

Student Name: ______________________________ Date: __________

Instructor: ______________________________ Class: __________

Experiment or Practice #: ______________________________ Process: __________

Briefly describe task: ______________________________

____

____

____

INSPECTION REPORT

Inspection	Pass/Fail	Inspector's Name	Date
Safety:			
Equip. Setup:			
Equip. Operation:			
Welding	Pass/Fail	Inspector's Name	Date
Accuracy:			
Appearance:			
Overall Rating:			
Comments:			

Student Grade: __________ Instructor Initials: __________ Date: __________

Jeffus: Welding Principles and Applications, Sixth Edition. Copyright © 2008 by Delmar Cengage Learning, Inc.

This page intentionally left blank

Glossary

amperage A measurement of the rate of flow of electrons; amperage controls the size of the arc.

amperaje Una medida de la proporción de la corriente de electrones; el amperaje controla el tamaño del arco.

amperage range The lower and upper limits of welding power, in amperage, that can be produced by a welding machine or used with an electrode or by a process.

rango de amperaje Los límites máximos y mínimos de poder de soldadura (en amperaje) que puede tener una máquina para soldar o que pueden usarse con un electrodo o a través de un proceso.

anode Material with a lack of electrons; thus, it has a positive charge.

ánodo Un material que carece electrones; por eso tiene una carga positiva.

arc length The length from the tip of the welding electrode to the adjacent surface of the weld pool.

largura del arco La distancia de la punta del electrodo a la superficie que colinda con el charco de la soldadura.

back gouging The removal of weld metal and base metal from the weld root side of a welded joint to facilitate complete fusion and complete joint penetration upon subsequent welding from that side.

gubia trasera Quitar el metal soldado y el metal base del lado de la raíz de una junta soldada para facilitar una fusión completa y penetración completa de la junta soldada subsecuente a soldar de ese lado.

burn-through Burning out of molten metal on the back side of the plate.

metal quemado que pasa al otro lado Metal derretido que se quema en el lado de atrás del plato.

cathode A natural curve material with an excess of electrons, thus having a negative charge.

cátodo Un material de curva natural con un exceso de electrones, por eso tiene una carga negativa.

cellulose-based fluxes Fluxes that use an organic-based cellulose ($C_6H_{10}O_5$) (a material commonly used to make paper) held together with a lime binder. When this flux is exposed to the heat of the arc, it burns and forms a rapidly expanding gaseous cloud of CO_2 that protects the molten weld pool from oxidation. Most of the fluxing material is burned, and little slag is deposited on the weld. E6010 is an example of an electrode that uses this type of flux.

fundentes para electrodos celulósicos Fundentes que usan celulosa de base orgánica ($C_6H_{10}O_5$) (un material normalmente utilizado para fabricar papel), y que se mantienen unidos con un aglomerante de cal. Cuando a este fundente se lo expone al calor del arco, se consume y forma una nube gaseosa de CO_2 que se expande rápidamente y protege de la oxidación al charco de soldadura derretido. La mayor parte del material del fundente se consume, y se deposita poca escoria en la soldadura. El E6010 es un ejemplo de un electrodo que utiliza este tipo de fundente.

chill plate A large piece of metal used in welding to correct overheating.

plato desalentador Una pieza de metal grande que se usa para corregir el sobrecalentamiento.

cleaning action A phenomenon occurring during DCEP in which oxides are removed from the surface to be welded by accelerated ions or electrons. These particles cause surface erosion, also called etching, that assist in the cleaning of the surface to be welded. Cleaning action requires an argon-rich atmosphere.

acción limpiadora Fenómeno que se presenta durante el proceso de DCEP (corriente directa electrodo positivo) en el que los óxidos se eliminan de la superficie para ser soldados por iones acelerados o electrones. Estas partículas provocan que la superficie se erosione (un proceso también conocido como decapado), lo cual auxilia en la limpieza de la superficie que se va a soldar. La acción limpiadora requiere una atmósfera con cantidades abundantes de argón.

collet A cone-shaped sleeve that holds a GTAW electrode in place within a GTA welding torch.

collarín Un manguito cónico que mantiene al electrodo en su lugar dentro del soplete para soldar por arco de tungsteno con gas (en inglés, GTAW).

contamination Undesirable foreign substances found on base metal surfaces, on filler wires, or in gas shielding atmospheres that inhibit welding operations.

contaminación Sustancias extrañas e indeseables en las superficies de metales básicos, alambres de relleno o atmósferas de protección por gas que inhiben las operaciones de soldadura.

cover pass The last layer of weld beads on a multiple pass weld. The final bead should be uniform in width and reinforcement, not excessively wide, and free of any visual defects.

pasada para cubrir La última capa de cordónes soldadura de pasadas múltiples. La pasada final debe ser uniforme en anchura y refuerzo, no excesivamente ancha, y libre de defectos visuales.

duty cycle The percentage of time during an arbitrary test period that a power source or its accessories can be operated at rated output without overheating.

ciclo de trabajo El porcentaje de tiempo durante un período a prueba arbitraria de una fuente de poder y sus accesorios que pueden operarse a la capacidad de carga de salida sin sobrecalentarse.

electrode angle The angle between the electrode and the surface of the metal; also known as the direction of travel (leading angle or trailing angle); leading angle pushes molten metal and slag ahead of the weld; trailing angle pushes the molten metal away from the leading edge of the molten weld pool toward the back, where it solidifies.

ángulo del electrodo El ángulo en medio del electrodo y la superficie del metal; también conocido como la dirección de avance (apuntado hacia adelante o apuntado hacia atras); el ángulo apuntado empuja el metal derretido y la escoria enfrente de la soldadura; y el ángulo apuntado hacia atrás empuja el metal derretido lejos de la orilla delantera del charco del metal derretido hacia atrás, donde se solidifica.

electrons A negatively charged component of an atom. Electrons exist outside of and surrounding the atom's nucleus. Each electron carries one unit of negative charge and has a very small mass as compared with that of a neutron or proton.

electrones Uno de los componentes con carga negativa de un átomo. Los electrones existen en el exterior y alrededor del núcleo del átomo. Cada electrón cuenta con una unidad de carga negativa así como con una masa pequeña en comparación con la de un neutrón o protón.

filler pass One or more weld beads used to fill the groove with weld metal. The bead must be cleaned after each pass to prevent slag inclusions.

pasada para rellenar Uno o más cordones de soldadura usados para llenar la ranura con el metal de soldadura. El cordón debe ser limpiado después de cada pasada para prevenir inclusiones de escoria.

flowmeter A control device used to regulate shield and backing gas flow-in. A flowmeter may control flow on a manifold system or it may be part of a combination flow and pressure regulator device.

flujómetro Un dispositivo de control empleado para regular el flujo del gas de protección y respaldo. Un flujómetro puede controlar el flujo en un sistema colector o bien puede ser parte de un dispositivo con regulación combinada de flujo y presión.

frequency The rate at which alternating electrical current (AC) is measured as it switches back and forth between polarities. Each cycle back and forth is counted as one hertz (Hz).

frecuencia La velocidad registrada por la corriente eléctrica alterna (CA) durante los cambios de polaridad alternados. Cada ciclo de intercambio se cuenta como un hertz (Hz).

gas coverage The protective atmosphere delivered on top of or underneath a weld. Gas coverage may be supplied from the welding torch or from a secondary trailing or backup device.

cobertura de gas La atmósfera protectora aplicada a la parte superior o inferior de una soldadura. La cobertura de gas puede provenir del soplete para soldar o bien de un dispositivo de escape secundario o de respaldo.

guided bend specimen Any bend specimen that will be bend-tested in a fixture that controls the bend radii, such as the AWS bend-test fixture.

probeta de dobléz guiada Cualquier probeta de dobléz en la cual se va a hacer un dobléz guiado en una máquina que controla el radio del dobléz, como la máquina de dobléz guiado del AWS.

hot pass The welding electrode is passed over the root pass at a higher than normal amperage setting and travel rate to reshape an irregular bead and turn out trapped slag. A small amount of metal is deposited during the hot pass so the weld bead is convex, promoting easier cleaning.

pasada caliente El electrodo de soldadura se pasa sobre la pasada de raíz poniendo el amperaje más alto que lo normal y proporción de avance para reformar un cordón irregular y sacar la escoria atrapada. Una cantidad pequeña de metal es depositada durante la pasada caliente para que el cordón soldado sea convexo, promoviendo más fácil la limpieza.

inert gas A gas that normally does not combine chemically with materials.

gas inerte Un gas que normalmente no se combina químicamente con materiales.

interpass temperature In a multiple pass weld, the temperature of the weld area between weld passes.

temperatura de pasada interna En una soldadura de pasadas multiples, la temperatura en la área de la soldadura entre pasadas de soldaduras.

inverter A welding machine that is much smaller than other types of machines of the same amperage range.

inversor Una máquina soldadora más pequeña que otros tipos de máquinas similares con el mismo rango de amperaje.

keyhole A welding technique in which a concentrated heat source penetrates completely through a workpiece, forming a hole at the leading edge of the weld pool. As the heat source progresses, the molten metal fills in behind the hole to form the weld bead.

soldadura con pocillo Una técnica en la cual una fuente de calor concentrado se penetra completamente a través de la pieza de trabajo, formando un agujero en la orilla del frente del charco de la soldadura. Asi como progresa la potencia de calor, el metal derretido rellena detrás del agujero para formar un cordón de soldadura.

lap joint A joint between two overlapping members.

junta de solape Una junta entre dos miembros traslapadas.

magnetic flux lines Parallel lines of force that always go from the north pole to the south pole in a magnet and surround a wire carrying DC current.

líneas magnéticas de flujo Líneas paralelas de fuerza que siempre van del polo norte al polo sur en un magneto, y rodea un alambre que lleva corriente DC.

mineral-based fluxes Fluxes that use inorganic compounds, such as the rutile-based flux (titanium dioxide, TiO_2). These mineral compounds do not contain hydrogen, and electrodes that use these fluxes are often referred to as low hydrogen electrodes. Less smoke is generated with this welding electrode than with cellulose-based fluxes, but a thicker slag layer is deposited on the weld. E7018 is an example of an electrode that uses this type of flux.

fundentes para electrodos de base mineral Fundentes que usan compuestos inorgánicos, como por ejemplo, el fundente a base de rutilo (bióxido de titanio, TiO_2). Estos compuestos minerales no contienen hidrógeno, y a los electrodos que usan estos fundentes se los llama con frecuencia electrodos de bajo hidrógeno. En la soldadura con electrodos se producen menos humos que en la que se realiza con fundentes celulósicos, pero se deposita una capa de escoria más gruesa en la soldadura. El F7018 es un ejemplo de un electrodo que usa este tipo de fundente.

molten weld pool The liquid state of a weld prior to solidification as weld material.

charco de soldadura derretido El estado líquido de una soldadura antes de solidificarse como material de soldadura.

multiple pass weld A weld requiring more than one pass to ensure complete and satisfactory joining of the metal pieces.

soldadura de pasadas múltiples Una soldadura que requiere más de una pasada para asegurar una completa y satisfactoria unión de las piezas de metal.

open circuit voltage The voltage between the output terminals of the power source when no current is flowing to the torch or gun.

voltaje de circuito abierto El voltaje entre los terminales de salida de una fuente de poder cuando la corriente no está corriendo a la antorcha o pistola.

operating voltage Also called closed circuit voltage, it is the voltage measured during welding.

voltaje operativo También conocido como voltaje de circuito cerrado, es el voltaje registrado durante el proceso de soldadura.

output The welding current a particular welding power supply is rated to produce, or the measure of current being produced during welding operations.

potencia de salida La capacidad de la corriente de soldadura de un suministro eléctrico para soldar, o bien la cantidad de corriente producida durante las operaciones de soldadura.

oxide layer An oxide layer is a thin coating that forms on the surface of aluminum alloys. Once the oxide layer has formed, the progress of corrosion is significantly reduced. However, the oxide layer must be removed prior to welding operations because it inhibits penetration. The melting point of aluminum oxide is approximately 3500°F (1926°C), while the melting point of aluminum itself is about 1200°F (650°C).

capa de óxido Una capa de óxido es una película delgada presente en la superficie de las aleaciones de aluminio. Una vez que se forma la capa de óxido el avance de la corrosión se reduce de manera significativa. Sin embargo, la capa de óxido debe eliminarse antes de realizar la soldadura ya que evita la penetración. El punto de fusión del óxido de aluminio es aproximadamente a 3500°F (1926°C), mientras que el punto de fusión del aluminio es a unos 1200°F (650°C).

postflow time The time interval from current shut-off to shielding gas and/or cooling water shut-off.

tiempo de poscorriente El intervalo de tiempo de cuando se cierra la corriente a cuando se cierra el gas de protección y o cuando se cierra el agua para enfriar.

postheating The application of heat to an assembly after welding, brazing, soldering, thermal spraying, or thermal cutting.

poscalentamiento La aplicación de calor a una asamblea después de la soldadura, soldadura fuerte, soldadura blanda, rociado termal, o corte termal.

preflow time The time interval that shield gas is delivered to a welding gun or torch before current is applied.

tiempo de preflujo El intervalo en el que el gas de protección se deposita en un soplete antes de aplicar corriente eléctrica.

preheating The heat applied to the base metal or substrate to attain and maintain preheat temperature.

precalentamiento El calor aplicado al metal base o substrato para obtener y mantener temperatura de precalentamiento.

protective zone The area covered by shield gas during welding operations.

zona protectora El área cubierta con gas de protección durante las operaciones de soldadura.

rectification Arc rectification is a phenomenon that occurs when the surface oxide of a nonferrous metal, such as aluminum, acts as a barrier that obstructs the flow of electrons between the electrode and the work during alternating current GTAW operations. Traditional square wave power supplies are equipped with superimposed high-frequency current, which helps establish and maintain the arc. Inverter power supplies with advanced circuitry switch back and forth so quickly between DCEN and DCEP that rectification is rarely experienced and superimposed high frequency is not needed.

rectificación Un arco de rectificación es un fenómeno que se presenta cuando el óxido superficial de un metal no ferroso, tal como el aluminio, actúa como una barrera que obstruye el flujo de electrones entre el electrodo y el proyecto durante la realización de la soldadura TIG (en inglés, GTAW) con corriente alterna. Los suministros eléctricos tradicionales de ondas cuadradas están equipados con una corriente de alta frecuencia superpuesta que ayuda a establecer y mantener el arco. El suministro eléctrico de los inversores con circuitos avanzados cambia a una velocidad tal entre DCEN y DCEP que rara vez se presenta la rectificación, por lo que el uso de alta frecuencia superpuesta no es necesaria.

rectifier Allows current to flow in one direction only.

rectificador Permite que la corriente fluya en una sola dirección.

root pass The first weld of a multiple pass weld. The root pass fuses the two pieces together and establishes the depth of weld metal penetration.

pasada de raíz La primera soldadura de una soldadura de pasadas múltiples. La pasada de raíz funde las dos piezas juntas y establece la profundidad de la penetración del metal soldado.

rutile-based fluxes This flux system produces a smooth and stable medium-penetrating arc with easily removeable slag. Rutile-based fluxes are used in several types of electrodes and in all welding positions for SMAW and FCAW.

flujos con base de rutilo Este sistema de flujo produce un arco homogéneo y estable para la penetración de medios con escoria fácil de eliminar. Los flujos con base de rutilo se utilizan en varios tipos de electrodos y en todas las posiciones de soldadura SMAW y FCAW.

spark gap oscillator A component in GTAW power supplies that helps to deliver superimposed high-frequency current that assists in arc initiation and alternating current GTAW welding of aluminum alloys.

oscilador de chispa Un componente en suministros eléctricos para soldar por arco de tungsteno con gas (en inglés, GTAW) que ayuda a administrar una corriente de alta frecuencia superpuesta que auxilia en la iniciación del arco y en el intercambio de corriente en soldaduras GTAW de aleaciones de aluminio.

square butt joint A joint made when two flat pieces of metal face each other with no edge preparation.

junta escuadra de tope Una junta hecha cuando dos piezas planas de metal se enfrentan una a la otra sin preparación de orilla.

step-down transformer A transformer-type welding machine that is quieter, more energy efficient, requires less maintenance, and is less expensive than electric-motor or gas-powered engine welding power supplies.

transformador reductor Máquina para soldar tipo transformador que es más silenciosa, consume menos energía, requiere menos mantenimiento y es menos costosa que los dispositivos para soldar con motor eléctrico o de gas.

stringer bead A type of weld bead made without appreciable weaving motion.

cordón encordador Un tipo de cordón de soldadura sin movimiento del tejido apreciable.

surface tension Surface tension is an attractive property of the surface of a liquid. Surface tension causes the surface portion of liquid to be attracted to another surface, such as that of another portion of liquid (as in connecting bits of water or a drop of mercury that forms a cohesive ball). In welding, surface tension keeps the molten weld bead in the proper shape during overhead joining operations.

tensión superficial La tensión superficial es una propiedad de atracción en la superficie de un líquido. La tensión superficial provoca que la superficie de un líquido se anexe a otra superficie, por ejemplo, la porción de otro líquido (como gotas de agua o una gota de mercurio, las cuales forman una esfera cohesiva). En soldaduras, la tensión superficial mantiene la forma correcta del cordón de soldadura fundido durante la realización de operaciones de acoplamiento elevado.

tee joint A joint between two members located approximately at right angles to each other in the form of a "T."

junta en T Una junta en medio de dos miembros que están localizados aproximadamente a ángulos rectos de uno al otro en la forma de "T."

tungsten The primary element used in making GTAW electrodes.

tungsteno El elemento primario empleado en la creación de electrodos GTAW.

voltage The measurement of electrical pressure.

voltaje La medición de presión eléctrica.

wagon tracks A pattern of trapped slag inclusions in the weld that show up as discontinuities in X rays of the weld.

huellas de carreta Una muestra de inclusiones de escoria atrapadas en la soldadura que enseña que hay discontinuidades en los rayos-x de la soldadura.

wattage A measurement of the amount of power in the arc; the wattage of the arc controls the width and depth of the weld bead.

número de vatios Una medida de la cantidad de poder en el arco; el número de vatios del arco controla lo ancho y hondo del cordón de la soldadura.

weave pattern The movement of the welding electrode as the weld progresses; common weave patterns include circular, square, zigzag, stepped, "C," "J," "T," and figure eight.

muestra de tejido El movimiento del electrodo para soldar a como progresa la soldadura; las muestras de tejidos comunes incluyen circular, de cuadro, zigzag, de pasos, "C," "J," "T" y la figura ocho.

weld groove A channel in the surface of a workpiece or an opening between two joint members that provides space to contain a weld.

soldadura de ranura Un canal en la superficie de una pieza de trabajo o una abertura entre dos miembros de junta que provee espacio para contener una soldadura.

welding cables The work cable and electrode cable of an arc welding circuit. Refer to Figure 1-7 for direct current electrode positive.

cables para soldar Los cables de pieza de trabajo y el portelectrodo de un circuito de soldadura de arco. Refierase al dibujo orriente directa con el electrodo positivo.

welding leads The work lead and electrode lead of an arc welding circuit.

cables para soldar Los cables de pieza de trabajo y el portelectrodo de un circuito de soldadura de arco.

weld specimen A welded component or section of a welded component that is to be inspected.

muestra de soldadura El componente o sección soldado de un componente también soldado que se va a analizar.

This page intentionally left blank

Index

Italic page numbers indicate material in tables or figures.

AC (alternating current), 5, *6,* 8–9, 111–113, 123–129
advanced square wave, 126–127
air-cooled torches, 117–119
alloy not specified tungsten electrodes (EWG), 113
alternating current (AC), 5, *6,* 8–9, 111–113, 123–129
alternators, 13–15
aluminum base metal, 107, 112, 146, 151–154, 158t
aluminum deposits, 114
American Bureau of Ships, 87
American Society of Mechanical Engineers (ASME), 87, 88
American Welding Society (AWS), 71, 87, 88
amperage, 4, 32, 146–147
  estimating and calculating, 11–12
  range charts, *29, 135,* 147
angle, torch, 143–144
angle, trailing, 35–36
anode, 6
arc, *3,* 7
  cleaning, 165
  reestablishing the, 125
  striking the, 15, 27–29, 136–138
  temperature of, 4–5
arc blow, 7–8, *9,* 21
arc length, 32–34, 78
arc strike defects, 28–29
argon, 107, 129
ASME Boiler and Pressure Vessel (BPV) Section IX, 87, 88
ASTM standards, 87, 88
AWS Certified Welder program, 71
AWS standards, 88

back caps, 118–119
back edge, 52, 81, 85
back gouging, 72, *73,* 86, 87f
backing gas, 130, 171, 177
backing strips, 72, 74f
backing tapes, 72
bead color, 153
belt sander, 113
breaking and remelting tungsten electrodes, 114–115
bridge rectifiers, 16, 17f
burn back, 83f
burn-through, 31, 45, 72, 75–77, 148–149

cathode, 6
CC (constant current), 6, 7f
cellulose-based fluxes, 41
cerium, 110, 112
cerium tungsten electrodes (EWCe-2), *111,* 112, 152, 153
checklist, portable welder, 16t
chemical cleaning and pointing, 116
chill plate, 156
clamps, 21
cleaning action, 124, 129
cold lap, 34–35
collet, 108–109, 136f
condensation, 121
constant current (CC), 6, 7f
constant voltage (CV), 6, 7f
contamination, 107
  and max cleaning setting, 127
  atmospheric, 143–144, 171, 177
  carbon, 115
  from dirt, 154, 181, 182, 189
  from grinding, 113–114
  from high gas flow rate, 132, 149
  from oxygen, 106
  from vaporized droplets, 112
  from weld pool, 112, 146
  of inert gas, 129
  of shielding gas, 120
  of stainless steel, 152–153
  on tungsten electrode, 115–116, 138, 146
  radioactive, 112
copper, 18–19, 116, 130
cover passes, 84–85
cracking the valve, 134
crater cracking, 145
cups (nozzles), 121–122, 132, *133,* 136f
current settings, 29–31, 146–149
CV (constant voltage), 6, 7f

DAV (dropping-arc voltage), 6
DC (direct current), 14–16, *17,* 112–113
DCEN (direct current electrode negative), 5, 107, 108–110, 123–126
DCEP (direct current electrode positive), 5, *6,* 106–107, 109–110, 116, 123–126
DCRP (direct current reverse polarity). *See* DCEP (direct current electrode positive)
DCSP (direct current straight polarity). *See* DCEN (direct current electrode negative)
defects
  arc strikes, 28–29
  overlaps, 73
  root surface, 72–73
  undercuts, 39, 63, 73, 79, 84–85, 88
deoxidizers, 152
deposition rate, 31
direct current (DC), 14–16, *17,* 112–113
direct current electrode negative (DCEN), 5, 107, 108–110, 123–126
direct current electrode positive (DCEP), 5, *6,* 106–107, 109–110, 116, 123–126
direct current reverse polarity (DCRP). *See* direct current electrode positive (DCEP)
direct current straight polarity (DCSP). *See* direct current electrode negative (DCEN)
drag electrodes (drag rods), 33
dropping-arc voltage (DAV), 6
duty cycles, 17–18

E6010 electrodes, 41
E6011 electrodes, 41
E6012 electrodes, 41
E6013 electrodes, 41
E7016 electrodes, 41
E7018 electrodes, *35,* 41, *44,* 71

E7024 electrodes, 33, 40–41
E7028 electrodes, 40–41
edge weld, 49
electrode amperage range table, *29, 135,* 147
electrode angle, 34–37
electrode holders, *5,* 20–21, 28
electrode lead, 5f
electrode manipulation, 37–39
electrodes, 2, 3f. *See also* tungsten electrodes
  and arc length, 32–33
  F-number classes, 40–41
  size and heat of, 31, 32f
electrons, 3
ER1100 filler metal, 154t
ER308, ER308L filler metal, 153t
ER309 filler metal, 153t
ER310 filler metal, 153t
ER316L filler metal, 153t
ER4043 filler metal, 154t
ER410 filler metal, 153t
ER70S-3 filler metal, 152t
erosion, electrode, 108–110
EWCe-2 (cerium tungsten electrodes), *111,* 112, 152, 153
EWG (alloy not specified tungsten electrodes), 113
EWLa-X (lanthanum tungsten electrodes), *111,* 112–113, 152, 153
EWP (pure tungsten electrodes), 111, 153
EWTh-X (thoriated tungsten electrodes), 111–112, 152
EWZr (zirconium oxide tungsten electrodes), 112, 153

F-number electrode classes, 40–41
face bends, 88–89
filler metals, 106, 114, 129, 152–153
filler passes, 81–84
filler rod manipulation, 144–145
fitup, 101–102
flame-cut specimens, 88–89
flowmeters, 122–123
flux, 2, 28–29, 33, 40–41
frequency, 125

gas coverage, 143
gas flow rate, 132, *133,* 149–151
gas metal arc welding (GMAW), 107
gas tungsten arc welding (GTAW). *See also* tungsten electrodes
  metal preparation, 154–155
  overview, 106–107, 143
  setup, *120,* 133–136
  types of current, 123–126
generators, 13–15
gouging, back, 72, *73,* 86, 87f
grain growth, 107, 112
grinding, 113–114
grinding stone wheel, 114
guided bend specimen, 89f

hard spatter, 30, 33, 41, 132
heat. *See* amperage
heliarc welding, 106
helium, 106, 130
high frequency arc starting, 125–126, 137
hoses, 118, 119–121
hot pass, 71, 79–81
hot start, 131
hydrogen additives, 130

icicles, 53, 54f
inclusions, 82, *94,* 108
inert gases, 129–130
interpass temperatures, 100–101
inverters, 13, *14,* 125–126
iron core, in transformer, 9, 13
iron-powder-based fluxes, 40–41

keyhole, 73–74
knobs, adjustment, 11–12

lanthanum, 110
lanthanum tungsten electrodes (EWLa-X), *111,* 112–113, 152, 153
lap joints, 58–63
leading angle, 34–35
lugs, 19, 20f

magnesium base metal, 107, 112
magnetic flux lines, 7–8
max cleaning setting, 127–128
max penetration setting, 127–128
metals, reactive, 107
mild steel, 151–152
mineral-based fluxes, 41
molten weld pool, 2–4
movable-coil machines, 12, 13f
multiple pass weld, 71
multiple-coil machines, 10–11

nameplate, 18
nickel alloy base metal, 111
nitrogen additives, 130–131
nozzles, 121–122, 132, *133,* 136f

open circuit voltage, 7
open root, 72–73
operating voltage, 7
output, 7
outside corner joints, 54–58
overlaps, 73
oxide layers, 146, 153–154, 165f

patterns, weave, 37–39
penetration, minimizing, 35
plasma arc welding (PAW), 13
point gap setting, 125
pointing tungsten electrodes, 116
poor fitup, 101–102
porosity, 41, 78, *94,* 130, 152
portable welders, 15, 16t
postflow time, 131–132, 149
postheating temperatures, 92, 100
power cable safety fuse, 118
power supplies, 6–7, 9–13
practice welds, 40–41, 151–155
preflow time, 131, 149
preheating temperatures, 92
protective zone, 143–144
pure tungsten electrodes (EWP), 111, 153

reactive metals, 107
rectification, 125
rectifier, 15–16, *17,* 126–128
remelting
  filler, 145
  tungsten electrodes, 116
remote controls, *120,* 133, 147
return water hose, 120
RG45 gas welding rods, 152
RG60 filler metal, 152t
root bends, 88–89
root fusion, *59,* 166f
root opening (root gap), 44, 45f
root pass, 71–78, 143
root penetration problems, 33–34
root surface defects, 72–73
run-off tabs, 83f
rutile-based fluxes, 41

safety
  setup considerations for, 22
  SMAW, 27
  when grinding tungsten electrodes, 114
  when using chemical cleaners, 154
safety fuse, 118

semiautomatic operation, 117f
settings, suggested
  aluminum, 158t
  mild steel, 157t
  stainless steel, 157t
setup
  gas tungsten arc welding (GTAW), *120,* 133–136
  shielded metal arc welding (SMAW), 22
shelf, support, *38,* 39, *44,* 47
shielded metal arc welding (SMAW)
  overview, 2–3
  power supply requirements, 7
  safety, 27
  setup, 22
  to join plate, 27
shielding gas, *3, 90,* 129–131
  argon, 107, 129
  coverage, 143
  effect of low amperage, 30f
  flow rate, 132, 133t
  helium, 106, 130
  hydrogen additives, 130
  nitrogen additives, 130–131
  protective zone, 143–144
shielding gas hose, 120
slag, 2, *3, 47,* 79–81
slag inclusions, 30, 35
spark gap oscillator, 125
spatter, hard, 30, 33, 41, 132
splices (welding cables), 19, 20f
square butt joints, 44–49
square wave, 126–127
stabilizers, arc, 4
steel base metal
  low-carbon and mild, 92, 151–152
  stainless, 107, 111, 130, 151–153, 157t
step-down transformers, 9–10
stick welding. *See* shielded metal arc welding (SMAW)
sticking, 66–67
stone grinding wheel, 114
striking the arc, 27–29, 136–138
stringer beads, 41–43
surface tension, 146

tack welding, 44–45, *161f–162f*
tap-type machines, 10–11
tee joints, 63–66
testing, 82, 88–89
thermal conductivity, 108, 113, 131, 154
thoriated tungsten electrodes (EWTh-X), 111–112, 152
thorium, 110–111
tip color chart for tungsten electrodes, 111t
titanium base metal, 107
torches, 108, 117–119, 143–144
trailing angle, 35–36
trailing edge, 31, *52,* 59, *61,* 85
transformers, 9–10
tungsten (W), properties of, 107
tungsten electrodes
  amperage chart for, 135t
  erosion of, 108
  properties of, 106–110
  shapes and shaping of, *110,* 113–117
  types of, 110–113
tungsten inert gas welding (TIG). *See* gas tungsten arc welding (GTAW)

under-fill, *83,* 88f
undercuts, 39, 63, 73, 79, 84–85, 88
valve, cracking the, 134
  voltage, 4
  closed circuit, 7
  constant, 4, 6, 7f
  open circuit, 7
  operating, 7
  rising-arc, 6, 7f

wagon tracks, 80
water-cooled torches, 117–119
water-in hose, 120
wattage, 4
waveforms, AC, 126–129
weave pattern, 37–39
weld beads, 89–92
weld failure from
  backing strip, 72
  inclusions, 82, 108, 146
  inert gas contamination, 129
  poor surface preparation, 89
  rapid cooling, 92
weld groove, 85, 86f
weld pool, 2–4
weld specimen, 88–89
weld toes, 76f
welders, vision of, 42
welding cables (leads), 13, 18–19, 22
welding current, types of, 5–6
welding position, 40
welds, practice, 40–41, 151–155
work (base metal), 106
work clamps, 21
work lead cable, 8–9f

zip-on hose protection, 121
zirconium, 110–112
zirconium oxide tungsten electrodes (EWZr), 112, 153

针灸

Claudia Focks

Atlas of Acupuncture

CHURCHILL
LIVINGSTONE
ELSEVIER

Dedication

To my parents Paula and Alfons Focks who have always supported my siblings and me in finding and pursuing our own paths, with love and gratitude

Commissioning Editor: Karen Morley, Claire Wilson
Development Editor: Kerry McGechie, Martin Mellor
Project Manager: Morven Dean
Designer: Stewart Larking
Photographs: Anja Messerschmidt
Graphics: Henriette Rintelen

CHURCHILL
LIVINGSTONE
ELSEVIER

CHURCHILL LIVINGSTONE
An imprint of Elsevier Limited

First and second edition published in German under the title *Atlas Akupunktur*
1. First edition 1998
2. Second edition 2006
© Elsevier GmbH, München

First edition published in English
© 2008, Elsevier Limited. All rights reserved.

The right of Claudia Focks to be identified as editor of this work has been asserted by her in accordance with the Copyright, Designs and Patents Act 1988

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the Publishers. Permissions may be sought directly from Elsevier's Health Sciences Rights Department, 1600 John F. Kennedy Boulevard, Suite 1800, Philadelphia, PA 19103-2899, USA: phone: (+1) 215 239 3804; fax: (+1) 215 239 3805; or, e-mail: healthpermissions@elsevier.com. You may also complete your request on-line via the Elsevier homepage (http://www.elsevier.com), by selecting 'Support and contact' and then 'Copyright and Permission'.

First edition	**1998**
Second edition	**2006**
English edition	**2008**

ISBN-13: 9780443100284
ISBN-10: 0443100284

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book is available from the Library of Congress

Notice
Neither the Publisher nor the Editor assumes any responsibility for any loss or injury and/or damage to persons or property arising out of or related to any use of the material contained in this book. It is the responsibility of the treating practitioner, relying on independent expertise and knowledge of the patient, to determine the best treatment and method of application for the patient.

The Publisher

ELSEVIER your source for books, journals and multimedia in the health sciences
www.elsevierhealth.com

Working together to grow libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

Printed in China

The Publisher's policy is to use **paper manufactured from sustainable forests**

Preface

'An acupuncture point is located where it is located' – like veins, arteries or nerves, the location of acupuncture points may vary. Consequently, locating and stimulating acupuncture points is an individual process in each patient, similar to locating veins or arteries for puncture. The goal of this atlas is therefore to help acupuncture students and therapists find the correct location of acupuncture points. Ultimately, the exact location of an 'acupuncture hole' will be where it can be palpated, addressed energetically in a safe way, and – most importantly – where it will be therapeutically effective.

Acupuncture therapy requires not only the knowledge to locate individual points but also an understanding of the context of these points. This atlas therefore includes both a description of the individual points as well as the central idea – the channel system.

The core chapters 4, 5 and 6 contain single page portraits of all the channel points as well as extra points used in body acupuncture. The clearly structured page layout, complemented by pictograms, allows for quick access to the clinically relevant information for the location and indication of individual points. Clear step-by-step instructions guide the reader through the surrounding surface anatomy to the correct location of the point. Figures with the relevant anatomical structures, as well as drawings showing channel pathways pertaining to the relevant region of the body, provide a further practical aid for correct point location. The text and the photo details also contain information regarding other points located in the vicinity or points in comparable locations in other parts of the body. This not only draws attention away from the individual points to the larger anatomical orientation, but also fosters an understanding of the context between the location and the action of a given point.

Chapter 7 presents the channel points according to the anatomical region – including illustrations of anatomical overview as well as text/picture details. This format will deepen the understanding of the anatomical relationship between the channel points and complete this academic concept. Chapter 2 'Location Methods and Cun Measurements' and Chapter 3 'Anatomical Orientation' provide further support in locating the individual points. The modified illustrations in these chapters are based on the well-known Sobotta Atlas.

Chinese medicine practitioners work with the Qi and its flow. In this context the individual 'acupuncture hole' has both an anatomical and an energetic component and is an intrinsic part of the channel and vessel network, which provides an exchange between the body's Exterior and Interior. In this way Chapter 1 describes the network of channels and vessels, and Chapter 8 the point categories and point combinations, explaining them in terms of their effects with relation to channel energetics, and presenting them in a clearly defined and visually organised manner. Chapter 9 contains up-to-date information and data about the scientifically proven effects or otherwise of individual points.

I hope the new edition of this atlas will provide a valuable help to students and therapists alike in studying acupuncture and applying the information in clinical practice. I look forward to hearing your constructive criticism and ideas.

Claudia Focks, March 2008

Acknowledgements

This Acupuncture Picture Atlas is a joint project of many very helpful contributors who have helped to complete it.

First of all my thanks go to the models and especially to the photographer Anja Messerschmidt. Her understanding and photographic skills provided the visual foundation for the Atlas to also be an artistic ‘feast for the eyes’.

I would like to thank the graphic designer Henriette Rintelen for the successful graphic design and converting my ideas into visual images.

Thanks go to my colleague Ulrich März for the succint arrangement and editing of the chapter on Anatomical Orientation as well as his contributions to the core chapters. I would like to thank my colleague Ingolf Hosbach for his contributions to the chapter on Research as well as his idea about the visually concise pictograms and for partially converting them by computer.

I want to thank all the staff at Elsevier involved in this project for their, once again, very productive co-operation. My thanks go particularly to Rolf Lenzen for agreeing to do the atlas, Christine Kosel for dealing with the gruelling organisation during production, as well as Petra Münzel-Kaiser and Christl Kiener for editing. My thanks extend particularly to Christl Kiener for her numerous ideas and fruitful exchange of concepts.

My heartfelt gratitude also includes my teachers who have sown the nourishing spark of enthusiasm for Chinese and energetic medicine. I thank my patients for their trust and for allowing me to learn from them.

I would also like to thank the strong women in my family and my circle of friends for their inspiration and being a source of vibrant female power. Thank you to my sister Annette Focks for her friendship and wonderful music, which helped me overcome some of the hurdles of this project. I extend particular thanks to Anneliese Solenski and Christa Skopp for their helpful support with my private affairs.

Above all I want to thank my inspirational husband Christoph Ranzinger and our wonderful children Tabea and Lasse for their love and generous tolerance, without which the Atlas would not have been possible.

For the English edition I would like to extend my thanks to the staff at Elsevier in London, particularly Karen Morley for accepting the atlas and to Kerry McGechie and Martin Mellor for the extensive editing work. Thank you also to Johanna Schuster who has translated the Atlas into English with great care and knowledge.

Claudia Focks, Rottweil, March 2008

About the editor and contributors

Dr med. Claudia Focks, Rottweil: anaesthesiologist, additional qualification in naturopathy and homeopathy; training in Chinese Medicine in China, Austria, and Germany, postgraduate training at the ETHNOMED Institut für Ethnomedizin e.V. (Institute for Ethnomedicine, registered charity) in Munich. Clinical focus on Chinese Medicine and energy therapies. Co-editor and co-author of *Leitfaden Chinesische Medizin* (*Chinese Medicine: An Introduction*), co-author of *Leitfaden Akupunktur* (*Acupuncture: An Introduction*), both published by Elsevier Munich. Lectures on Chinese Medicine, both in Germany and abroad.

Dr med. Ulrich März, Ulm: General practitioner with his own practice, focus on acupuncture and Traditional Chinese Medicine. Adjunct lecturer at the University of Ulm for acupuncture and TCM. Project manager for the postgraduate programme for acupuncture and TCM at the Akademie für Wissenschaft, Wirtschaft und Technik (Academy for Science, Economy and Technology) at the University of Ulm. Co-author of *Leitfaden Akupunktur* (*Acupuncture: An Introduction*), published by Elsevier Munich.

Dr med. Ingolf Hosbach, Bochum: Anaesthesiologist, additional qualification in palliative therapy. Employed at the University Clinic Bergmannsheil Bochum. Adjunct lecturer at the Universities of Witten/Herdecke and Bochum. Scientific publications and lectures on palliative therapy, acupuncture and TCM research. Webmaster of the homepage of the Department for Chinese Medicine at the University Witten/Herdecke and of the DWGTCM e.V. (www.dwgtcm.com).

Pictogrammes

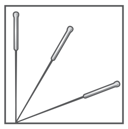

Angle of insertion: The angle of insertion recommended for a particular point (see figure below) is shown in colour. (Note that occasionally more than one angle is possible – this is signified by colouring as appropriate).

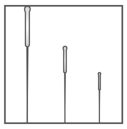

Depth of insertion: Shallow: up to 0.5 cun, **Medium:** 0.5–1 cun, **Deep:** >1 cun. (In some cases more than one depth is possible depending on the angle of insertion – this is signified by colouring as appropriate).

Caution! Organs and structures such as the eyes, nerves, blood vessels, peritoneum (symbolised by intestinal tissue) etc. that might be injured by needling a particular point are shown by a symbol together with a blue exclamation mark. A lightning bolt signifies a point that is painful when needled.

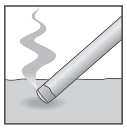

Moxibustion: In principle all points can be treated with moxibustion. The following options are pointed out: **Blue colour below moxa cigar:** moxibustion particularly recommended, **Question mark next to moxa cigar:** moxibustion questionable (contraindicated according to some classical texts, but contradictory information), **Moxa cigar crossed out:** moxibustion contraindicated.

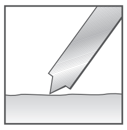

Bloodletting: In principle bloodletting may be applied to all points. **Neutral pictogramme:** bloodletting possible, **Blue droplet of blood:** bloodletting particularly recommended, **Blue question mark:** bloodletting questionable

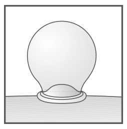

Blue cup: Cupping particularly recommended.

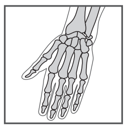

Anatomical orientation: A blue hand signifies cases where anatomical structures are of particular importance for point location (for more detail → chapter 3).

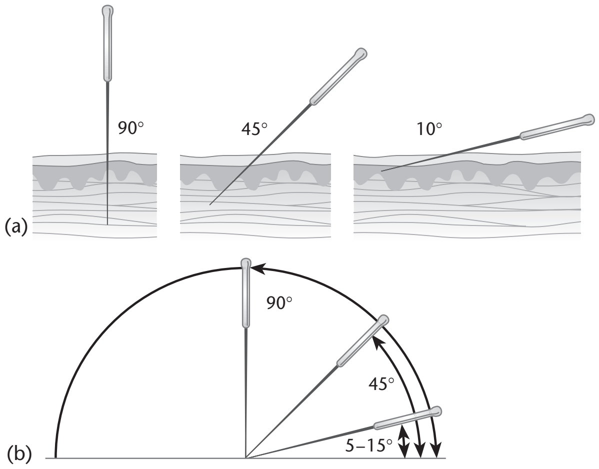

Angle of insertion
a) Angle of insertion and layers of tissue: All needles are inserted to the same 'depth' (=the same proportion of the needle is inserted into the tissue), but reaching different levels of tissue.
b) Angle of insertion: transverse (subcutaneously): 5–15°; oblique: 15–45°; perpendicular: 90°.

Point hierarchy (Chapters 4–6)

In the introductions to the channels (chapters 4–6) the importance of points are marked with the following symbols:

■■: very important, universal point
■■: important point

This hierarchy is of course subjective and influenced by the authors' clinical experience but has proved very helpful for beginners.

1 *Jing luo* (Channel and Network Vessel System)

Claudia Focks

1.1 Introduction

1.1.1 Two Models of Qi Flow in the Channels – Historical Overview

There exist essentially two views about the direction of the Qi flow in the channels, which are based on different understandings of the Qi movement within the body (→ Pirog 1995, Manaka 1995/2004). For a better understanding of either model, imagine a person extending his/her arms towards the sky (→ Figs. 1.1 and 1.2).

Centripetal circulation model

The centre of classical Chinese cosmology, which describes humans as standing between heaven (Yang) and earth (Yin), includes the mutual relationship, influence, and dependence between cosmos and man. Accordingly, early records of the pathways (→ see appendix) depict the (11/12) channels as pathways that originate at the extremities, flow towards the centre of the body and terminate either on the head or the torso. This is the distinguishing feature of the **centripetal circulation model** (→ Fig. 1.1).

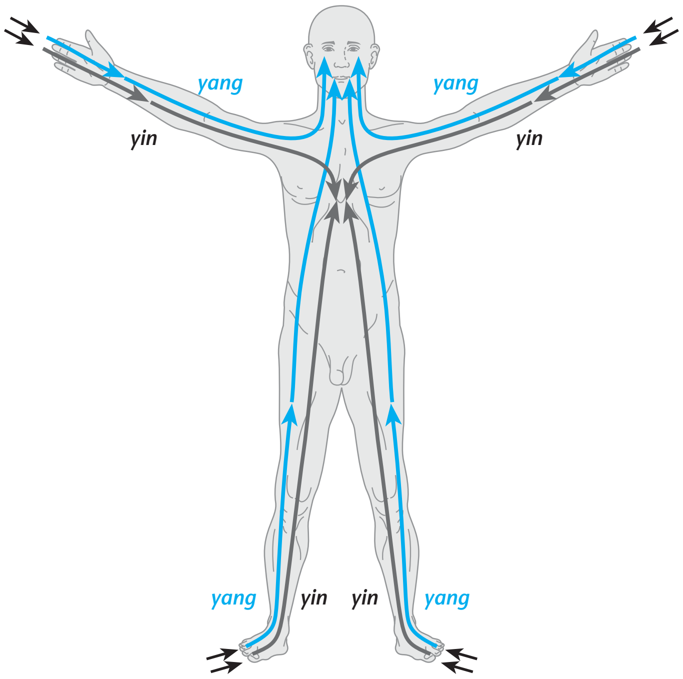

Fig. 1.1 Centripetal circulation model (modified according to Pirog 1996)

This model is also present in the theory of the five *shu*-Transporting points (→ 8.1.6). The macrocosmic Qi enters the body at the tips of the extremities and can be compared to the course of a river. It begins very dynamically as a well, spring and stream, expands to form a river and flows into the wide, deep sea at the elbows and the knees and then further to the internal organs via the channels.

According to this model, the functions of the channels can be compared to antennae, which receive the cosmic influence, transmitting it into the body. The Qi flow in the channels is always from distal (coming from the Exterior, entering at the tips of the extremities) to proximal (towards the centre, flowing towards the internal organs). Each channel connects man with a different part of the cosmos, which can be identified by a numerological structure. Thus a relationship was assumed between the eight extraordinary vessels and the eight trigrams of the *Yijing* (→ 1.7). The primary channels reflect the 12 earthly branches and the 10 heavenly branches. Since the early records only mention 11 channels (→ Appendix 2), they were counted as 10 channels on the arms and 12 channels on the legs. Later, the 12 primary channels were mainly associated with the 12 earthly branches – the 10 heavenly branches rather representing the Five Phases model (*note*: the earthly branches can be included in the latter as well).

Self-contained circulation model

In the course of the development of modern Chinese society, the idea of a close relationship between man and cosmos (as a heaven-earth-man model) became weaker. The channel system was now increasingly compared to phenomena created by man himself such as **canals** and **drainage ditches**. The importance of the original connection with the cosmos declined while that of relationships within society was on the rise. Chinese society became more complex, forming a self-contained unit. Similarly, the concept of the channels as a connection to the macrocosm was partially abandoned and the Qi flow is described as independent and self-contained (→ Fig. 1.2). According to this **self-contained circulation model**, Qi can flow forward as well as backward – from the outer extremities to the Interior of the body and from the Interior to the Exterior.

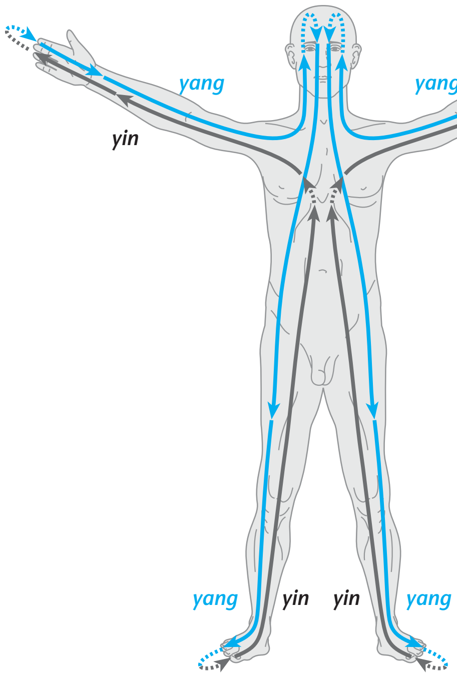

Fig. 1.2 Self-contained circulation model (modified according to Pirog 1996)

Thus the direction of the Qi flow in the primary channels can be from proximal to distal and vice versa, depending on the Yin/Yang polarity of the respective channel, as well as the respective extremity.

This second, probably more recent concept describes the Qi **as circulating continuously** through the body (→ Appendix 2): from the thorax to the hand, to the head, to the foot and back to the thoracic region. These ideas reflect the development of the Chinese civilisation, its agriculture and in particular its **water control** and water **storage** in reservoirs, drainage canals, ditches, etc., which largely form the foundation of the theory of the **channels as water conduits** circulating Qi and Blood. The connections among the channels are considered to be anastomoses (→ 1.2.2), which facilitate the ceaseless, circular flow of Qi from one channel to the next, allowing the exchange of the Qi flow with the Interior.

The self-contained circulation model can thus be described as follows:

- Hand Yin channels run from the thorax to the hand: **LU, HE, P**
- Hand Yang channels run from the hand to the head: **L.I., S.I., T.B.**
- Foot Yang channels run from the head to the foot: **ST, BL, G.B.**
- Foot Yin channels run from the foot to the thorax: **SP, KID, LIV**

More importance is attached to this model of Qi flow in the Western acupuncture tradition, which can in part explain the use of numbers in naming the acupuncture points, rather than names as in China.

However, the structure of this new circulatory model was probably also too rigid to sufficiently explain some of the effects of acpuncture. According to Pirog (1996), this might be the reason why the secondary channels, for example the sinew channels (→ 1.4) and the divergent channels (→ 1.3), with their rather primitive and natural pathways, were integrated into the channel and network vessel system (*jing luo*) in accordance with the **centripetal** circulation model.

Comparison of the two circulation models (modified according to Pirog 1996)

	Centripetal model	Self-contained model
Direction of Qi-flow	Always from distal to proximal	Either from proximal to distal or vice versa, depending on the polarity of the channel (Yin/Yang)
Origin of Qi	From outside the body. The distal end of the channel is open in order to receive cosmic Qi	Originates in the inside of the body. The distal end of the channels is connected to the channel that follows next.
Function of the channels	Transporting the Qi from the outer cosmos to the Interior of the body. Supporting the relationship between man and nature/ cosmos	Circulation of the Qi inside the body. Supporting man's relationship with himself

1.1.2 Overview of the Channel and Network Vessel System (*jing luo* System)

In *Ling Shu*, chapter 11, it says: 'Man lives, diseases occur … both the beginner and the experienced master always have to start with the channels and network vessels (*jing luo*).'

In Chinese Medicine, the *jing luo* are considered to be a network of channels and vessels in which the Qi and Blood (*xue*) flow. They are connected to the Organ systems (*zangfu*) and 'water' the whole organism, supplying the body with Qi and Blood (*xue*) on the surface (Exterior) and deep inside the body (Interior), above as well as below.

From a functional point of view, the channels and network vessels (*jing luo*) govern the distribution of Qi and Blood (*xue*), they regulate Yin and Yang and they protect the body. However, they also enable the spreading of diseases. Reactions to any disorders may therefore manifest along these channel pathways. These could be either disorders of the channels themselves, or external reflections of *zangfu* disorders. In clinical practice, the channels and network vessels (*jing luo*) can be utilised in order to send Qi to the diseased parts of the body (for an overview of the classification and nomenclature of the *jing luo* system → Fig. 1.3).

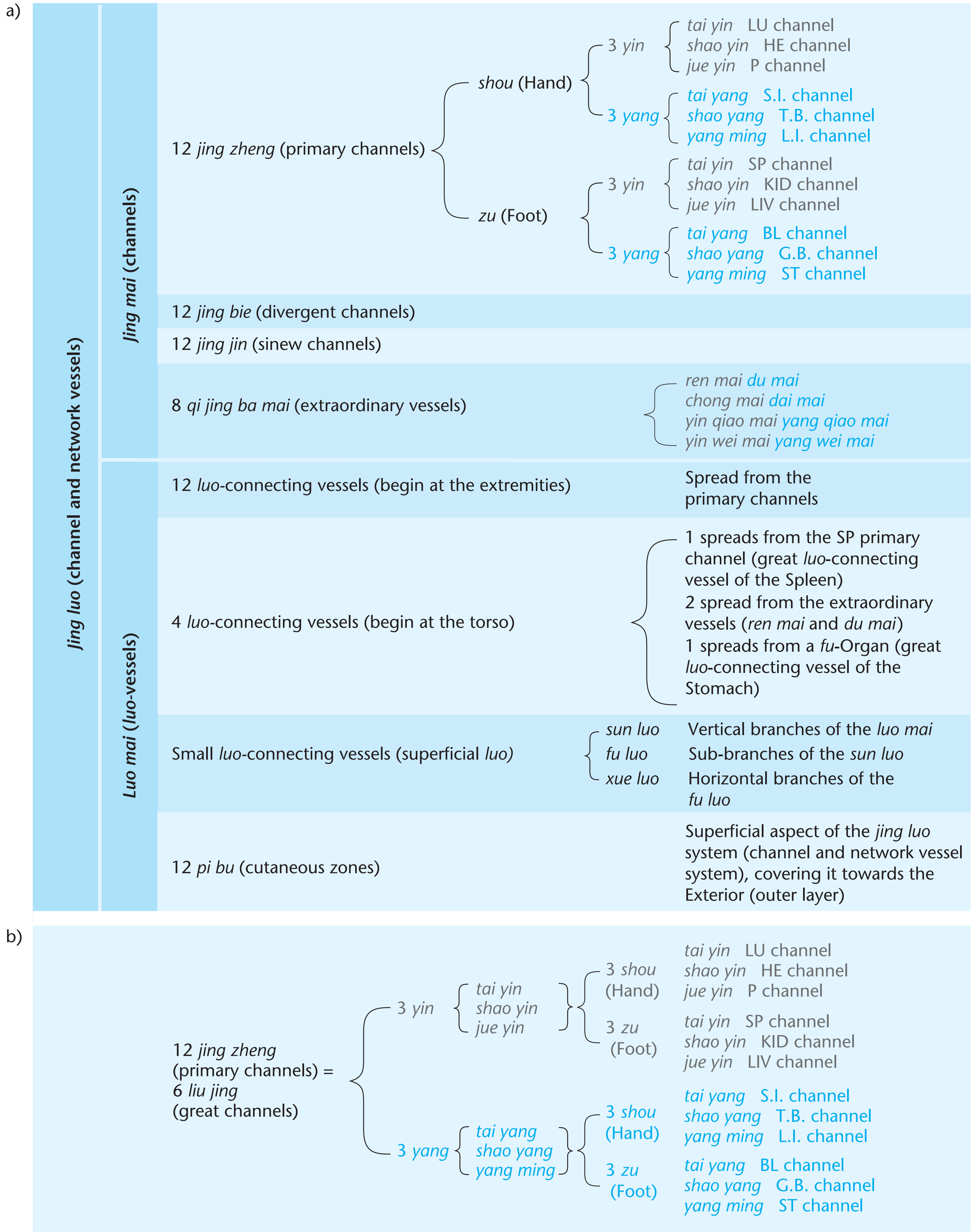
a)
Jing luo (channel and network vessels)
Jing mai (channels)
12 jing zheng (primary channels)
shou (Hand)
3 yin
tai yin LU channel
shao yin HE channel
jue yin P channel
3 yang
tai yang S.I. channel
shao yang T.B. channel
yang ming L.I. channel
zu (Foot)
3 yin
tai yin SP channel
shao yin KID channel
jue yin LIV channel
3 yang
tai yang BL channel
shao yang G.B. channel
yang ming ST channel
12 jing bie (divergent channels)
12 jing jin (sinew channels)
8 qi jing ba mai (extraordinary vessels)
ren mai du mai
chong mai dai mai
yin qiao mai yang qiao mai
yin wei mai yang wei mai
Luo mai (luo-vessels)
12 luo-connecting vessels (begin at the extremities)
Spread from the primary channels
4 luo-connecting vessels (begin at the torso)
1 spreads from the SP primary channel (great luo-connecting vessel of the Spleen)
2 spread from the extraordinary vessels (ren mai and du mai)
1 spreads from a fu-Organ (great luo-connecting vessel of the Stomach)
Small luo-connecting vessels (superficial luo)
sun luo
fu luo
xue luo
Vertical branches of the luo mai
Sub-branches of the sun luo
Horizontal branches of the fu luo
12 pi bu (cutaneous zones)
Superficial aspect of the jing luo system (channel and network vessel system), covering it towards the Exterior (outer layer)
b)
12 jing zheng (primary channels) = 6 liu jing (great channels)
3 yin
tai yin
shao yin
jue yin
3 shou (Hand)
tai yin LU channel
shao yin HE channel
jue yin P channel
3 zu (Foot)
tai yin SP channel
shao yin KID channel
jue yin LIV channel
3 yang
tai yang
shao yang
yang ming
3 shou (Hand)
tai yang S.I. channel
shao yang T.B. channel
yang ming L.I. channel
3 zu (Foot)
tai yang BL channel
shao yang G.B. channel
yang ming ST channel

Fig. 1.3 (a) Overview and classification of the *jing luo* system (channel and network vessel system); (b) the 12 primary channels and the six great channels *(liu jing)*

1.1.3 Distribution and Organisation of the Channel and Network Vessel (*jing luo*) System

According to the law of the Exterior–Interior (*biao-li*) the 'Exterior' communicates with the 'Interior'.

Exterior (*biao*) corresponds to the skin, the muscles and the pathways of the channel and network vessel superficial (*jing luo*) system. The deeper pathways of the channels and the Organ systems (*zangfu*) are attributed to the Interior (*li*). A particular organisational structure within the *jing luo* system is necessary to safeguard the circulation of Qi and the communication between the Exterior and the Interior. In this respect, the extraordinary vessels play a special role. While they play a major role in coordinating and regulating the primary channels and the *jing luo* system in general, they do not directly connect the Interior and the Exterior. There is also no direct connection between the extraordinary vessels and the *zangfu*-Organs (see → 1.7 and Chapter 5).

Depth organisation of the *jing luo* system

Differing ideas exist regarding the depth at which the various channels and vessels are located within the body. The table below and Fig. 1.4 show the ideas according to Solinas et al. (1998) and Deadman et al. (1998) in a modified way. While these authors regard the deep, Interior pathways of the primary and divergent channels as the deepest channel structures, other authors attribute this role to the eight extraordinary vessels (for the eight extraordinary vessels, see → 1.7).

Overview over the possible location with regard to depth of the *jing luo* system

Levels	Channel system
Superficial levels (skin and muscles)	• 12 cutaneous regions • Superficial *luo*-connecting vessels (*xue luo, fu luo, sun luo*) • 12 sinew channels (*jing jin*)
Mid-levels	• 15/16 *luo*-connecting vessels (*luo mai*) • External pathways of the 12 primary channels • External pathways of the 12 divergent channels (*jing bie*) • Eight extraordinary vessels
Deep levels (*zangfu*-Organs)	• Deep (internal) pathways of the 12 primary channels • Deep (internal) pathways of the 12 divergent channels

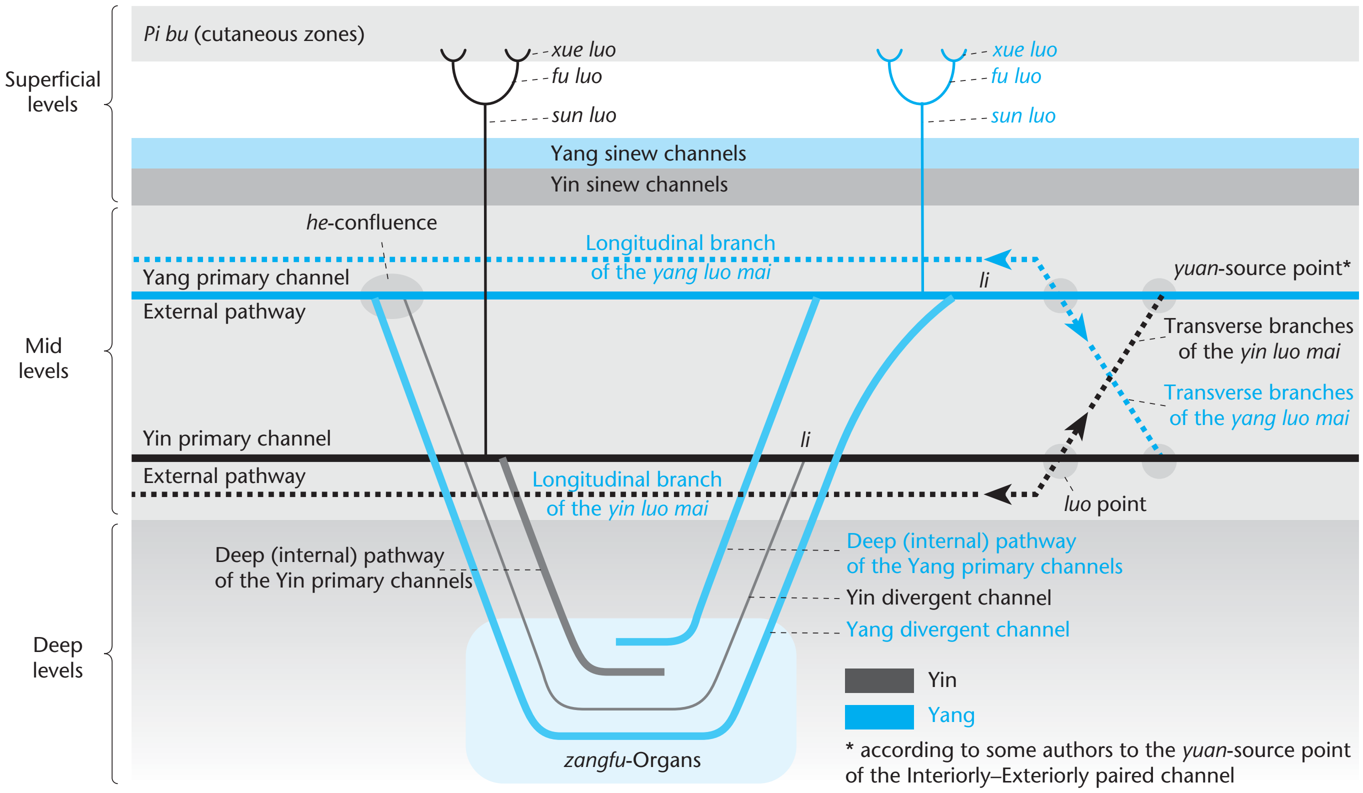
Superficial levels
Mid levels
Deep levels
Pi bu (cutaneous zones)
xue luo
fu luo
sun luo
xue luo
fu luo
sun luo
Yang sinew channels
Yin sinew channels
he-confluence
Longitudinal branch of the yang luo mai
Yang primary channel
External pathway
li
yuan-source point*
Transverse branches of the yin luo mai
Transverse branches of the yang luo mai
Yin primary channel
li
External pathway
Longitudinal branch of the yin luo mai
luo point
Deep (internal) pathway of the Yin primary channels
Deep (internal) pathway of the Yang primary channels
Yin divergent channel
Yang divergent channel
Yin
Yang
zangfu-Organs
* according to some authors to the yuan-source point of the Interiorly–Exteriorly paired channel

Fig. 1.4 Illustration of the possible organisation of the *jing luo* system (channel and network vessel system)

Channel systems

Each primary channel, together with its pertaining 'secondary channels', forms a complex, multi-layered organisational unit, a 'system'. These systems comprise the various relationships and connections among each other as well as with the eight extraordinary vessels (→ 1.7, → Chapter 5). They support the harmonious regulation within the body. There are 12 channel systems (hand *taiyin*, hand *yangming*, etc.), each of which includes the following:

- a **primary channel** (*jing zheng*) (→ 1.2) with an external and internal pathway, which enters its pertaining *zangfu*-Organ
- a **divergent channel** (*jing bie*) (→ 1.3)
- a **sinew channel** (*jing jin*) (→ 1.4)
- a great ***luo*-connecting vessel** (*luo mai*) (→ 1.5).

In turn, the **cutaneous zones** (*pi bu*) (→ 1.6) wrap around the whole channel system.

1.1.4 Qi Circulation and the Channel System

What moves with and within the channels?

The channel system represents the 'pathway of the Qi'. In acupuncture many schools of thought work with the **True Qi (*zhen qi*)**, which, according to Maciocia (1989), represents the final stage of a process of Qi refinement and transformation: the Gathering Qi (*zong qi*), acted upon by the Original Qi (*yuan qi*) as a catalyst, becomes True Qi (*zhen qi*), which manifests in its two aspects, the Defensive Qi (*wei qi*) and the Nutritive Qi (*ying qi*). According to Larre and Rochat de la Vallée (1986), the True Qi is regarded as the **sum of all Qi mechanisms and/or forms of Qi** in the body. In other words: if there is a balanced and harmonious flow in the channels, this can be called True Qi (*zhen qi*). It means that which circulates in the 'here and now'. The sum total of all forms of Qi in their correct alignment is called **Upright Qi (*zheng qi*)** and forms the counterpart to Evil Qi (*xie qi*), for example to pathogenic factors or counterflow Qi.

Forms of Qi

Essence (Qi) (*jing qi*)

The Essence (Qi) acts within the organism and, according to some authors, to some extent also in the extraordinary vessels (→ 1.7). It represents the combination of Early Heaven Essence inherited from the parents, which, according to many schools of thought, is stored in the Kidneys and has a relationship to the Life Gate (*mingmen*), and the Later Heaven Essence of the Middle Burner. They support and complement each other.

Original Qi (*yuan qi*)

The Original Qi (*yuan qi*) is often described as the **active form of Essence (*jing*)**, circulating in the channels and spreading to the whole organism with the help of the Triple Burner. Like a catalyst, it is the dynamic driving force which wakens and sustains the functional activity of all organs and structures. It has a pre- and a postnatal component. Therefore its proper functioning is dependent on the supply of acquired Qi derived from water and food supplied by the Middle Burner. The Original Qi can be directly accessed and influenced at the *yuan*-source points (→ 8.1.1) or at the points **Ren-17** (*danzhong/shanzhong*), **Ren-12** (*zhongwan*) and **Ren-6** (*qihai*).

Gathering Qi (*zong qi*)

According to Larre and Rochat de la Vallée (1986), the Gathering Qi (*zong qi*) is formed in the centre of the thorax when the essences derived from respiration (Great Qi, *da qi*) meet with the food (*shui gu*, from Grain Qi (*gu qi*) and water) and are set in motion through ancestral mechanisms by the first postnatal breath. Like an 'engine', it dictates the rhythm and circulation, it controls respiration and it regulates the heartbeat. It does not circulate itself, but is like a 'Sea of Qi' that 'collects' in the centre of the thorax behind **Ren-17** (*danzhong/shanzhong*). This sea contains the water from all the rivers (e.g. the vessels), resembling an inexhaustible reservoir that does not overflow but redistributes all its water supplies.

Defensive Qi (*wei qi*)

The *wei qi* is the body's **Defensive Qi**. Thus it defends the body at the level of the skin, the fascia and the muscles. In relation to the Nutritive Qi (*Ying qi*), it is a more slippery and less pure form of Qi that moves dynamically and quickly through the body like a **guard**. Its root lies in the **Lower Burner**, where it is produced by the ***mingmen***-Fire (Life Gate fire), so that it has a considerable genetic or constitutional component. Essence (*jing*) and Original Qi (*yuan qi*), which are stored in the Lower Burner (according to many schools in the Kidneys), are involved in the formation of Defensive Qi (*wei qi*), therefore also playing a role in the defence against Exterior pathogens. Additionally, the *wei qi* is constantly being replenished with pure essence, formed from food by the Spleen and Stomach in the **Middle Burner**. It is dispersed throughout the body by the **Upper Burner**. Proper functioning of the *wei qi* therefore depends on all three Burners.

Nutritive Qi (*ying qi*)

The Nutritive Qi (*ying qi*) is the result of a purification and/or distillation process of pure, clear origin. Compared to Defensive Qi (*wei qi*), it is more Yin in nature. It **nourishes** the whole body and is its '**building substance**', thus often translated as 'constructive Qi' or 'building energy'.

Circulation of Defensive Qi (*wei qi*) and Nutritive Qi (*ying qi*)

Circulation of Defensive Qi (*wei qi*)

In chapter 43 of the *Su Wen* the *wei qi* is described as flowing 'outside of the *mai*' (the channels and vessels). However, it circulates in part along the channel pathways, moving in the space between the skin and the muscles – the ***cou li***.

The *cou li* – often unsatisfactorily translated as ‘pores’ – are the striae or compartments between the skin and the muscles. They have the function of a gateway for the entry and exit of Qi and fluids and serve as protection against the invasion of Exterior pathogenic factors. According to Larre and Rochat de la Vallée (1986), the *cou li*, as the outermost of the wrappings of the Triple Burner, cover the whole body. They connect the surface of the body with its inner organs. This explains the effectiveness of acupuncture and other manual therapies for the treatment of internal disorders as well as showing up the presence of diseases of the inner organs on the Exterior of the body.

In the **superficial layers of the body**, the Defensive Qi (*wei qi*) circulates through the skin and the superficial musculature, warming, nourishing and strengthening them. Through these actions it supports the general defence against Exterior pathogenic factors, acting mainly in the realm of the sinew channels (*jing jin*) (→ 1.4). In the **deep layers of the body**, it plays an important role in the functioning of the ‘diaphragm’. According to Larre and Rochat de la Vallée (1986), this represents not only a membranous barrier between the thorax and the abdomen, but can be regarded as a ‘sac of membranes’, including and connecting the peritoneum, the pleura and the pericardium. Nielsen (1995) interprets this ‘network of caves’ as the **inner aspect of the Triple Burner**, which connects with its **outer aspect, the *cou li***. Following this interpretation, the Defensive Qi (*wei qi*) would also be involved with the **mesenteric defence** and the protection of the inner organs.

According to chapter 75 of the *Ling Shu*, the **circulation of the Defensive Qi (*wei qi*)** is cyclical, changing from day to night and vice versa. At dawn, when the Yin Qi is exhausted, the Yang Qi pours itself from the eyes and the eyes are opened. As a result, the *wei qi* rises from the heel via the (Yin) *qiao mai* upwards to the eye to **BL-1** (*jingming*) and flows into the whole body like a waterfall by following the six great Yang channels – ‘it moves 25 times in the Yang’. At dusk, when the Yang Qi is exhausted, the *wei qi* enters the Interior of the body and moves ‘25 times in the Yin’, following the controlling cycle (*ke* cycle → 8.2.5): from the Kidneys to the Heart, from the Heart to the Lungs, from the Lungs to the Liver, from the Liver to the Spleen, and back to the Kidneys. Therefore the *wei qi* has a relationship with the sleep–wake rhythm: during sleep it withdraws deeper into the body – during daytime it circulates in the outer layers of the body.

The two extraordinary vessels, the *yin qiao mai* and the *yang qiao mai* (→ 1.7, → Chapter 5), play an important role in this respect. If their cycle is blocked, the circulation of the *wei qi* will be interrupted and disorders will arise. The *yin qiao mai* ascends, while the *yang qiao mai* descends, both meeting at the eyes at **BL-1** (*jingming*) and forming a cycle, similar to the small heavenly cycle of the *ren mai* and *du mai*. If there is an excess of Yang Qi, this will be drained from the channels into the *yang qiao mai*, ‘the eyes cannot close’ and insomnia will occur. If there is an excess of Yin Qi, this will be drained from the channels into the *yin qiao mai*, ‘the eyes cannot open’ and somnolence will occur. In both cases, a relative disharmony develops between the two opponents.

Circulation of the Nutritive Qi (*ying qi*)

The Nutritive Qi (*ying qi*) circulates in the *mai*. The *mai* include both the primary channels (*jing mai*) and the secondary channels such as the Connecting Vessels (*luo mai*, *sun luo*), the divergent channels (*jing bie*), the extraordinary vessels and the Blood Vessels. Wherever there are channels and vessels, the *ying qi* will circulate. According to chapter 16 and 18 of the *Ling Shu*, the *ying qi* circulates constantly and continuously.

First circulation of the *ying qi* (→ Fig. 1.5)

The first circulation of the *ying qi* encompasses the 12 primary channels. The cycle **starts** at the **primary LU channel** and **ends** at the **primary LIV channel**, which in turn connects with the primary LU channel, thus closing the cycle (→ Fig. 1.5). The **Middle Burner** provides and **distributes** the *ying qi* throughout the whole body. According to some of the classics, the Grain Qi (*gu qi*) derived from food is only transformed into Nutritive Qi (*ying qi*) in the Lungs, not already in the Middle Burner. Since the internal pathway of the primary Lung channel begins in the Middle Burner, these two statements do not contradict each other in principle.

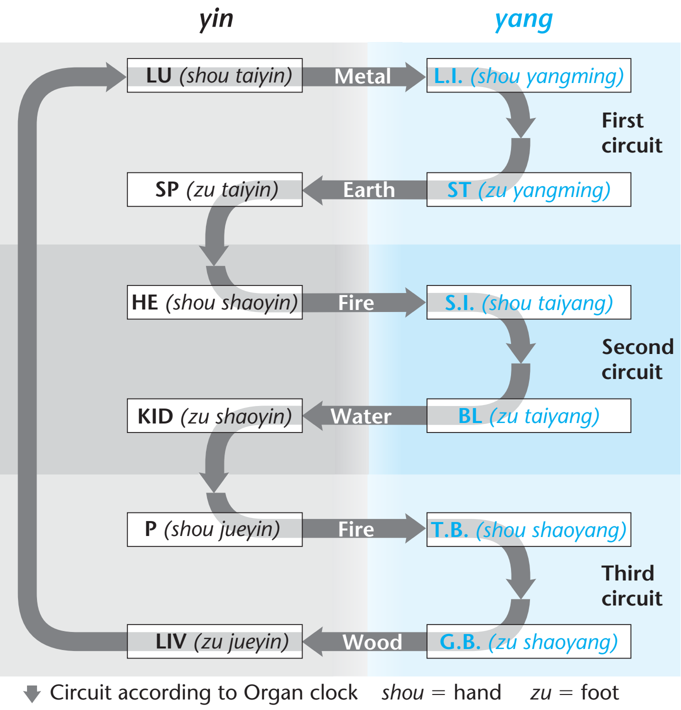

Fig. 1.5 First circulation of the Nutritive Qi (*ying qi*)

The **changes in Yin/Yang polarisation** – the changes from a Yin channel to a Yang channel and vice versa – always take place in the hand or the foot (also → 1.2.2). For example, the Yang energy potential rises from the Yin channel to the Yang channel, and then descends again from the Yang channel to the Yin channel (→ Fig. 1.6).

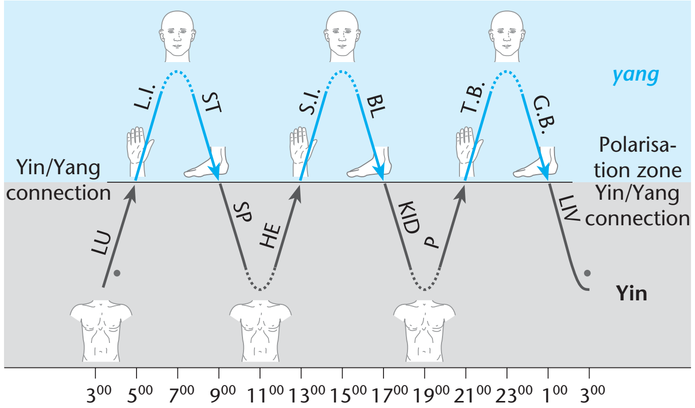
yang
L.I.
ST
S.I.
BL
T.B.
G.B.
Polarisation zone
Yin/Yang connection
Yin/Yang connection
LU
SP
HE
KID
P
LIV
Yin
3^{00} 5^{00} 7^{00} 9^{00} 11^{00} 13^{00} 15^{00} 17^{00} 19^{00} 21^{00} 23^{00} 1^{00} 3^{00}

Fig. 1.6 Changes in Yin/Yang polarisation, shown as a sinus curve

First circulation of the Nutritive Qi (*ying qi*) and the Organ clock

The Nutritive Qi (*ying qi*) circulates in the 12 primary channels in a **circadian 24-hour rhythm** (→ Fig. 1.7) with each **Chinese hour**, named after one of the 'earthly branches', corresponding to two Western hours (→ expanded Organ clock, Fig. 1.8).

Each two-hour slot favours a particular channel, in other words: for two (Western) hours per day, **energy flow peaks** in a particular channel. During this period, the respective channel Qi increases, falling again during the following two hours. However, it will never fall below a particular level, so that there is always a certain portion of energy flowing continuously through the channels.

Example: The channel Qi of the primary Stomach channel will begin to appear more 'powerful', when the channel Qi of the Large Intestine is very strong. Therefore the 'high tide' of the Stomach channel takes place between 5–7 a.m., its peak is between 7–9 a.m. and its 'low tide' between 9–11 a.m. During the peak period of the Stomach channel, the channel on the opposite side of the Organ clock is at its lowest point (→ 8.3.7); while the Stomach channel is peaking, the Pericardium channel is at its lowest point.

In figure 1.8 the 'expanded Organ clock' with its 24-hour circulation through the 12 primary channels is shown in relation to the earthly branches and the hexagrams of the *Yijing*.

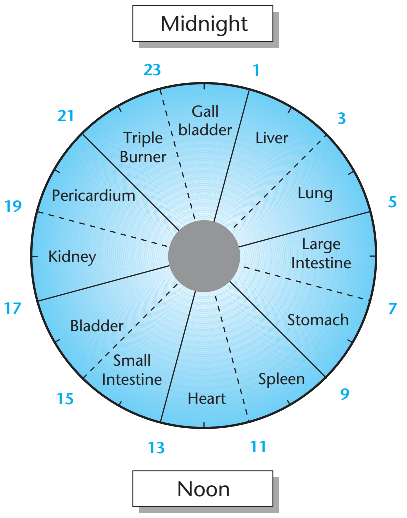
Midnight
23
1
21
3
19
5
17
7
15
9
13
11
Gall bladder
Liver
Lung
Large Intestine
Stomach
Spleen
Heart
Small Intestine
Bladder
Kidney
Pericardium
Triple Burner
Noon

Fig 1.7 Organ Clock

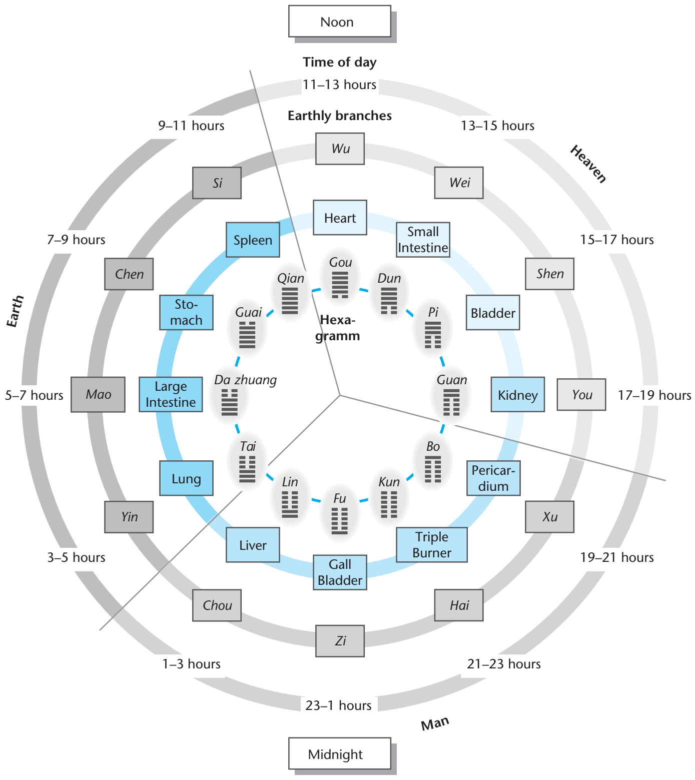
Noon
Time of day
11–13 hours
Earthly branches
9–11 hours
13–15 hours
Heaven
Wu
Si
Wei
Heart
Spleen
Small Intestine
7–9 hours
15–17 hours
Chen
Shen
Gou
Qian
Dun
Sto-mach
Guai
Pi
Bladder
Hexa-gramm
Earth
5–7 hours
Mao
Large Intestine
Da zhuang
Guan
Kidney
You
17–19 hours
Tai
Bo
Lung
Lin
Fu
Kun
Pericar-dium
Yin
Xu
Liver
Triple Burner
Gall Bladder
3–5 hours
19–21 hours
Chou
Hai
Zi
1–3 hours
21–23 hours
23–1 hours
Man
Midnight

Fig. 1.8 Expanded Organ clock

Second circulation of the Nutritive Qi (*ying qi*)

According to chapter 16 of the *Ling Shu*, the Nutritive Qi (*ying qi*), after having circulated through the 12 primary channels, takes a deep pathway of the primary Liver channel, which originates at **LIV-14** (*qimen*), passes the Lungs, the posterior aspect of the neck, the inner nasal passages and reaches **Du-20** (*baihui*), then follows along the *du mai* on the back and the *ren mai* along the midline of the abdomen. At the approximate level of **Ren-22** (*tiantu*) the *ying qi* again reaches the Lung channel, which traverses the supraclavicular fossa (near **ST-12**). Here, a new circulation of the *ying qi* through the 12 primary channels starts at the beginning of the primary LU channel.

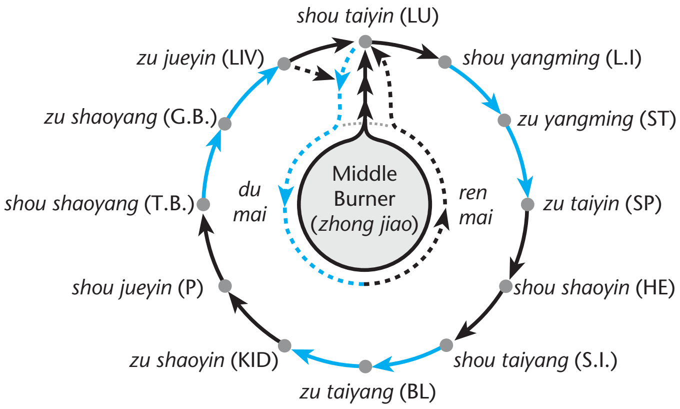

Fig. 1.9 First and second circulation of the Nutritive Qi (*ying qi*) with the Middle Burner

1.2 The 12 Primary Channels (*jing zheng*)

1.2.1 Background Information

Synonyms: Meridians, regular channels. '*zheng*' can be translated as 'mainly' (primary channels), but also relates to terms such as 'straight' and 'direct'.

Chinese terms for the channels

Meanings (also → 1.2.4, the six great channels (*liu jing*)

Tai: means greater, highest. The polarised energy (Yang or Yin) develops and reaches the maximum extent of its activity in the *taiyang* and *taiyin* channels. Here, the maximum extent of their respective polarity is reached and begins to fall again.

Shao: means less or younger. The polarised energy (Yang or Yin) is less in the *shaoyang* and *shaoyin* channels than in the *taiyang* or *taiyin* channels.

Ming: means clear, bright, radiant, shining (for more detail see → 1.2.4, the six great channels).

Jue: means 'absolute', 'at the end', 'exhausted', and according to Wiseman also 'inverted'. The polarised energy (Yin) in the *jueyin* channel represents the terminal phase of the Yin; here polarity changes to Yang.

Composition of the channel names

In the Chinese language, the name for each channel comprises

- the nature or **polarity of the energy** (Yin or Yang) that flows in the channel
- the **quality** or **intensity** (*tai*, *shao*, *jue*, *ming*) of the Yin or Yang energy and
- the **extremity** at which the channel originates or terminates.

Thus the Yin or Yang channels that begin or terminate at the hand carry the name of the primary hand channels (***shou*** *jing zheng*). The Yin and Yang channels that terminate or begin at the feet carry the name of the foot primary channels (***zu*** *jing zheng*) (→ Fig. 1.2).

1.2.2 Communication and Connections

Principle of the primary channel system

The 12 primary channels cover the body **bilaterally**. Each channel has its **individual regular course** with a deep, **internal** and a more superficial **external** pathway.

One can distinguish between Yin and Yang channels, which are Interiorly/Exteriorly paired. While each channel is connected with its pertaining *zang* or *fu* Organ, it is also connected with the Organ of its Interiorly/Exteriorly coupled channel.

All hand Yin channels begin in the region of the thorax and flow to the hand. All hand Yang channels start at the hands and travel to the head, where they meet the foot Yang channels. These descend down to the toes, where they meet the foot Yin channels, which ascend to the thorax and there meet the hand Yin channels.

For a better understanding of this model, imagine a person standing with his/her hands raised to the sky. In this position, all **Yin channels are ascending** (a **Yang phenomenon**), while all **Yang channels are descending** (a **Yin phenomenon**) (→ Fig. 1.10).

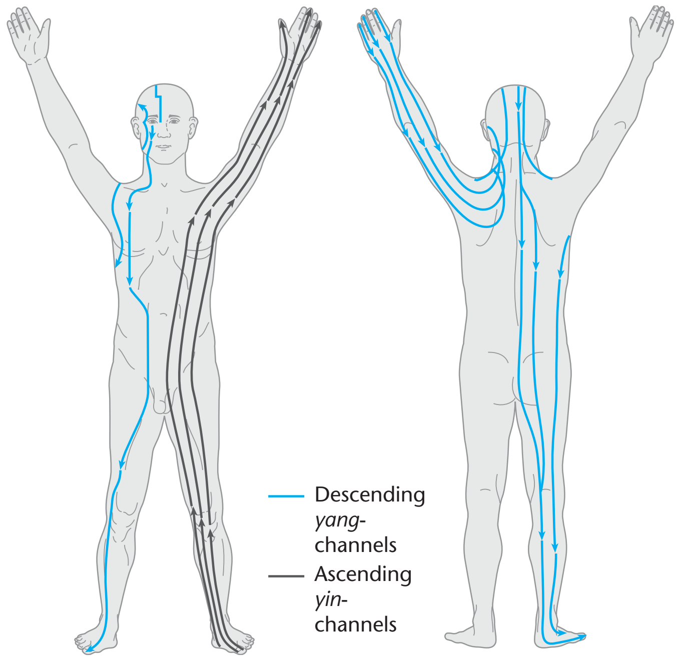

Fig. 1.10 Ascending and descending primary channels (according to Pirog 1996)

Each channel is associated with its own specific pathological symptoms, which are an important diagnostic tool in clinical practice (for channel-specific pathologies see → chapter 4).

Communication between channels and *zangfu*-Organs

The **connections (anastomoses)** between the channels have the function of allowing communication within the channel system as well as between the channels and the Organ systems. This external/internal (*biao-li*) relationship comprises the following aspects:

Communication from the Interior to the Exterior

- The **internal pathways of the hand Yin channels** connect with their pertaining Organ as well as with their Yin/Yang-paired *fu*-Organ. They flow from the thoracic region to the Exterior and connect with their paired Yang channel at the finger tips or on the hand (→ Fig. 1.11 a).
- The **internal pathways of the foot Yang channels** connect with their pertaining *fu*-Organ as well as with their Yin/Yang-paired *zang*-Organ. They then flow from the centre of the body to its Exterior and connect with their paired Yin–channel at the foot (→ Fig. 1.11 b).

Communication from the Exterior to the Interior

- **Each hand Yang channel** begins on the hand, runs to the thoracic region, penetrates the thorax and connects in the Interior of the body with its pertaining *fu*-Organ and its Yin/Yang-paired *zang*-Organ (→ Fig. 1.11 c).
- **Each foot Yin channel** starts at the foot, penetrates the abdomen, reaches the Interior of the body and its pertaining *zang*-Organ as well as its Yin/Yang-paired *fu*-Organ (→ Fig. 1.11 d).

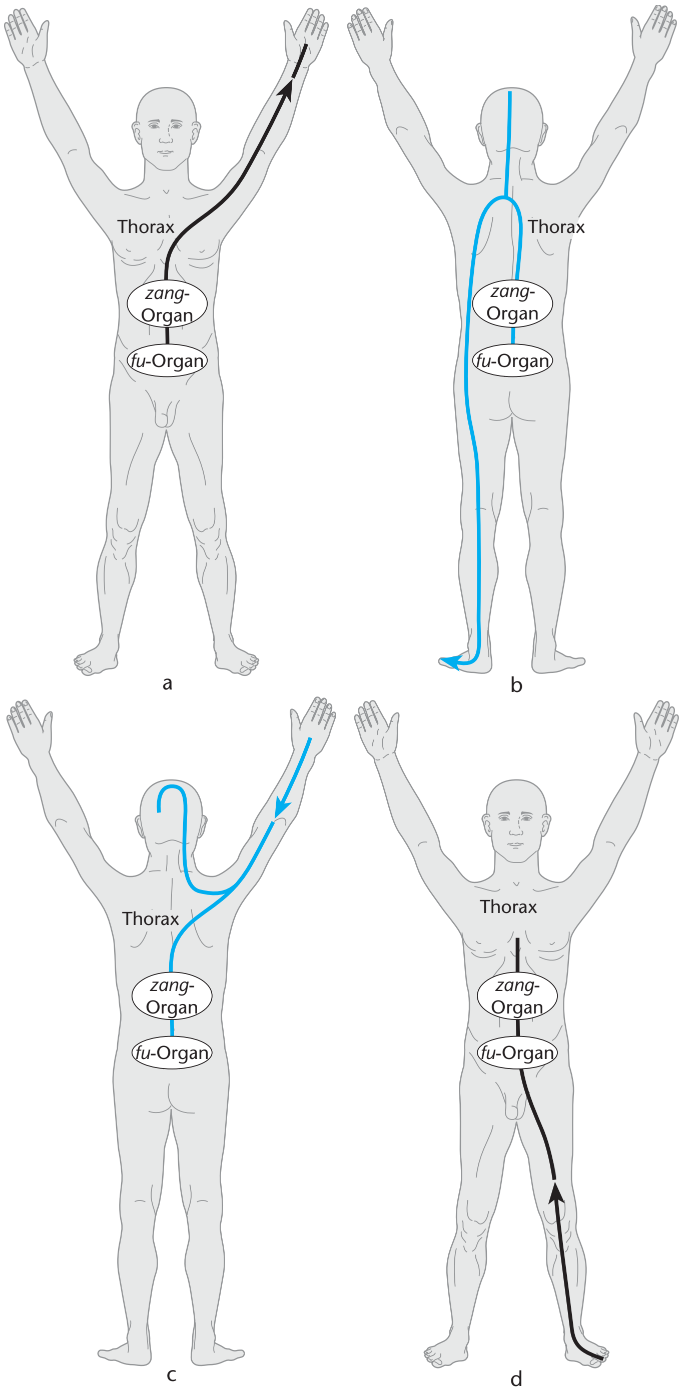

Fig. 1.11 Communication Interior → Exterior, hand-Yin channels (a). Communication Interior → Exterior, foot-Yang channels (b). Communication Exterior → Interior, hand-Yang channels (c). Communication Exterior → Interior, foot-Yin channels (d).

Channel circuits

Within the primary channel system we can distinguish three circuits, each circuit comprising four primary channels. The trajectories of two of the channels in one circuit are located on the Yin side of the body, the more anterior side (Interior), and two are found on the Yang side, the more posterior side of the body (Exterior). The Stomach channel (foot *yangming*, for more detail ➞ 1.2.3) presents the exception to this rule. Despite its major part running along the anterior side of the body, it is considered a Yang channel (➞ Fig. 1.12 and 1.3).

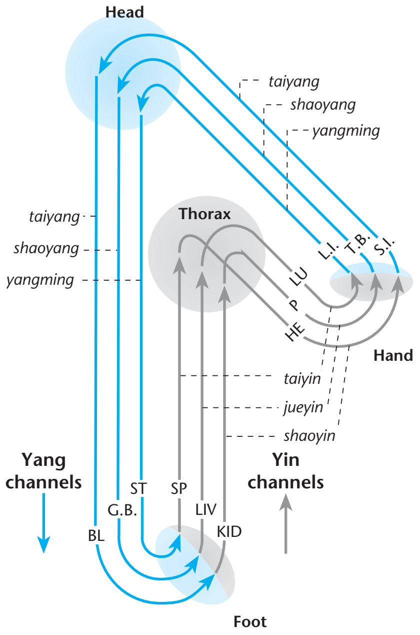
Head
taiyang
shaoyang
yangming
taiyang
shaoyang
yangming
Thorax
L.I.
T.B.
S.I.
LU
P
HE
Hand
taiyin
jueyin
shaoyin
Yang channels
Yin channels
ST
SP
G.B.
LIV
BL
KID
Foot

Fig. 1.12 Principle of the channel circuits

		First circuit			
Channel	**Time**	**Starts**	**Terminates**	**Hand/foot**	
LU	3–5 a.m.	Thorax	Finger tips	Hand (*shou*)	*taiyin*
L.I.	5–7 a.m.	Finger tips	Face	Hand (*shou*)	*yangming*
ST	7–9 a.m.	Face	Tips of the toes	Foot (*zu*)	*yangming*
SP	9–11 a.m.	Tips of the toes	Thorax (Heart)	Foot (*zu*)	*taiyin*
		Second circuit			
H	11 a.m.–1 p.m.	Thorax	Finger tips	Hand (*shou*)	*shaoyin*
S.I.	1–3 p.m.	Finger tips	Face	Hand (*shou*)	*taiyang*
BL	3–5 p.m.	Face	Tips of the toes	Foot (*zu*)	*taiyang*
KID	5–7 p.m	Tips of the toes	Thorax (Pericardium)	Foot (*zu*)	*shaoyin*
		Third circuit			
P	7–9 p.m.	Thorax	Finger tips	Hand (*shou*)	*jueyin*
T.B.	9–11 p.m.	Finger tips	Face	Hand (*shou*)	*shaoyang*
G.B.	11 p.m.–1 a.m.	Face	Tips of the toes	Foot (*zu*)	*shaoyang*
LIV	1–3 a.m.	Tips of the toes	Thorax (Lung)	Foot (*zu*)	*jueyin*

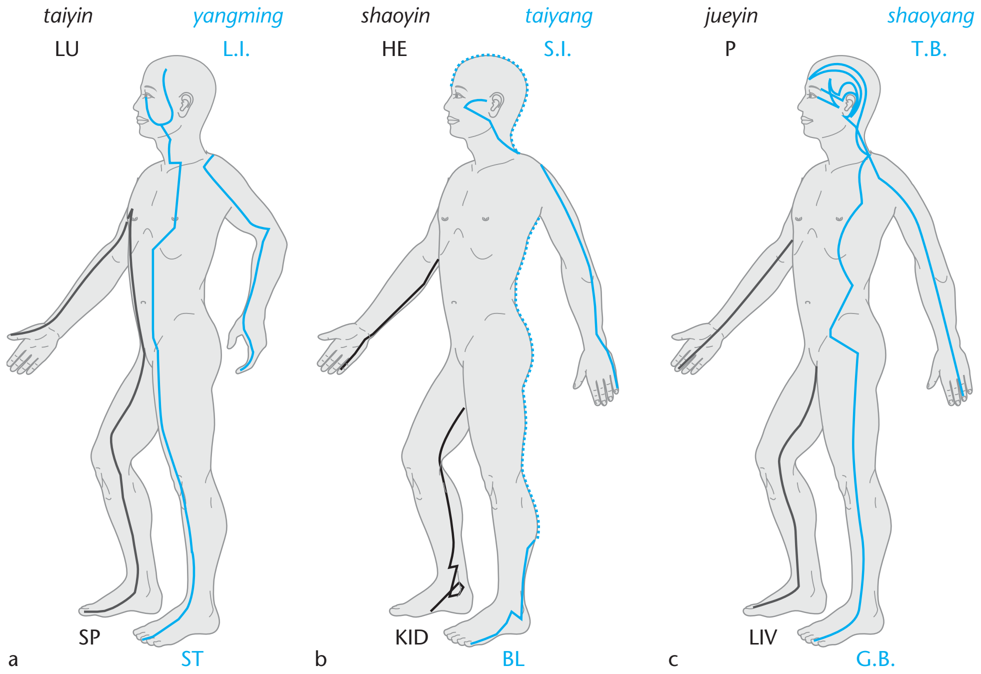

Fig. 1.13 Overview of the channel circuits

Connections among the primary channels

In order to ensure the continuous circulation (also see → 1.1.4) within the primary channel system, there need to be connections among the individual primary channels. Fig. 1.5 illustrates these in a schematic graphic overview.

According to some schools, the connections among the primary channels are facilitated by so-called entry/exit points. Various authors (for example Worsley, Jarrett, Pirog, Hicks et al., for more detail → 8.1.16) describe these as the shunting points between successive channels according to the Organ clock. The exit point marks the point on a channel from which the (internal) flow moves to a point (the entry point) on the channel that succeeds it on the Organ clock. There are some discrepancies regarding the location of the connections as described by the above authors and those described by Solinas et al. (1998), which are marked below as appropriate.

Yin–Yang connections

The connections – anastomoses – between the Yin and Yang channels are located on the hands and feet. The channels are most commonly connected at their respective final points, but sometimes also by branches separating from the main channel. These Yin–Yang connections are places where the Qi can change its polarity: Yin Qi becomes Yang Qi, and Yang Qi becomes Yin Qi. This creates a dynamic balance and allows for a better circulation of the Qi. The Yin–Yang connections follow the order of the basic structure of the circadian rhythm (24-hour cycle).

Hand Yin–Yang connections (→ Fig. 1.14)

The **hand Yin channels** are in most cases connected to the **hand Yang channels** at the tips of the fingers; sometimes they are connected on the hand by a branch that separates from the primary Yin channel and runs to the hand Yang channel:

FIRST CIRCUIT

LU channel (*shou taiyin*) ➡ L.I. channel (*shou yangming*)

LU-7 (*lieque*) ➡ **L.I.-1** (*shangyang*)

A branch separates from the primary LU channel at → **LU-7** (*lieque*) and travels to the primary L.I. channel at → **L.I.-1** (*shangyang*).

Note: According to Hicks et al. (2004, → 8.1.16) **L.I.-4** (*hegu*) is considered to be an entry point.

SECOND CIRCUIT

HE channel (*shou shaoyin*) ➡ S.I. channel (*shou taiyang*)

HE-9 (*shaoyang*) ➡ **S.I.-1** (*shaoze*)

THIRD CIRCUIT

P channel (*shou jueyin*) ➡ T.B. channel (*shou shaoyang*)

P-8 (*laogong*) ➡ **T.B.-1** (*guanchong*)

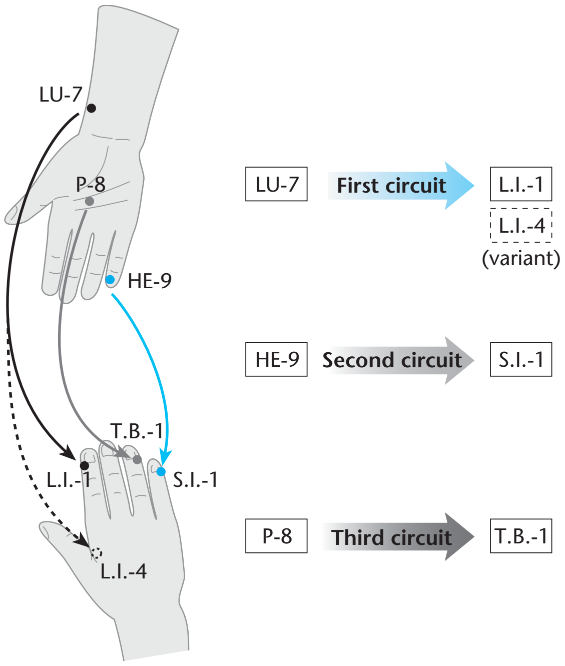
LU-7
P-8
HE-9
T.B.-1
L.I.-1
S.I.-1
L.I.-4
LU-7
First circuit
L.I.-1
L.I.-4
(variant)
HE-9
Second circuit
S.I.-1
P-8
Third circuit
T.B.-1

Fig. 1.14 Hand Yin–Yang connections (primary channels)

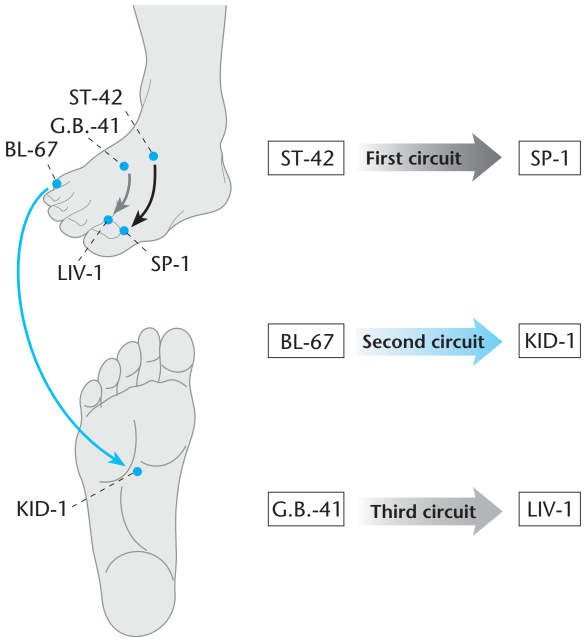
ST-42
G.B.-41
BL-67
LIV-1
SP-1
KID-1
ST-42
First circuit
SP-1
BL-67
Second circuit
KID-1
G.B.-41
Third circuit
LIV-1

Fig. 1.15 Foot Yang–Yin connections (primary channels)

Foot Yang–Yin connections (→ Fig. 1.15)

The **foot Yang channels** are connected to the **foot Yin channels** at the toes by branches that separate from the primary Yang channels and run to the foot Yin channels (→ Fig. 1.15):

First circuit

ST channel (*zu yangming*) ➡ SP channel (*zu taiyin*)

ST-42 (*chongyang*) ➡ **SP-1** (*yinbai*)

At → **ST-42** a branch separates from the primary ST channel and travels to → **SP-1**.

Second circuit

BL channel (*zu taiyang*) ➡ KID channel (*zu shaoyin*)

BL-67 (*zhiyin*) ➡ **KID-1** (*yongquan*)

Third circuit

G.B. channel (*zu shaoyang*) ➡ LIV channel (*zu jueyin*)

G.B.-41 (*zulinqi*) ➡ **LIV-1** (*dadun*)

On the instep of the foot, at → **G.B.-41**, a branch separates from the G.B. channel and travels to → **LIV-1**

Yang–Yang connections (→ Fig. 1.16)

The Yang–Yang connections between the primary channels are located on the **head**. They are rather superficial and follow the order of the channels within the **circadian cycle** (according to the Organ clock → Fig. 1.7). These connections support the communication between the hand Yang and foot Yang primary channels in the cranial region, which convey Qi of the same nature and quality (*yangming*, *taiyang*, *shaoyang*). They also mark the connections between the Yang axes, also referred to as hand–foot pairing (→ 1.2.3), and serve as a link between **above** and **below**.

First circuit

L.I. channel (*shou yangming*) ➡ ST channel (*zu yangming*)

L.I.-20 (*yingxiang*) ➡ **ST-1** (*chengqi*)

Second circuit

S.I. channel (*shou taiyang*) ➡ BL channel (*zu taiyang*)

S.I.-18 (*quanliao*) ➡ **BL-1** (*yingming*)

Third circuit

T.B. channel (*shou shaoyang*) ➡ G.B. channel (*zu shaoyang*)

T.B.-23 (*sizhukong*) ➡ **G.B.-1** (*tongziliao*). Comment: according to Hicks et al. (2004, → entry/exit points, → 8.1.16) **T.B.-22** (*erheliao*) is considered an exit point.

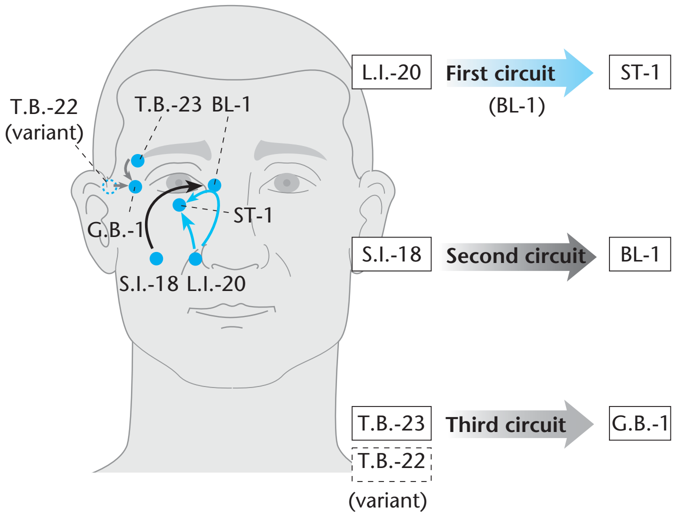
T.B.-22
(variant)
T.B.-23 BL-1
G.B.-1
ST-1
S.I.-18 L.I.-20
L.I.-20
First circuit
(BL-1)
ST-1
S.I.-18
Second circuit
BL-1
T.B.-23
T.B.-22
(variant)
Third circuit
G.B.-1

Fig. 1.16 Yang–Yang connections of the primary channels

Yin–Yin connections

The Yin–Yin connections between the primary channels are located in the **thoracic region** (Fig. 1.17).

In contrast to the more superficial Yang–Yang connections, they are located deep inside the body (internal pathways of the channels).

There are two types of Yin–Yin connection:

- the Yin axes connections or hand–foot pairings (*taiyin*, *jueyin*, *shaoyin*), which are **non-circadian** (do not flow according to the Organ clock) and
- the (deep) Yin–Yin connections, whose Qi flow follows the circadian rhythm of the Organ clock.

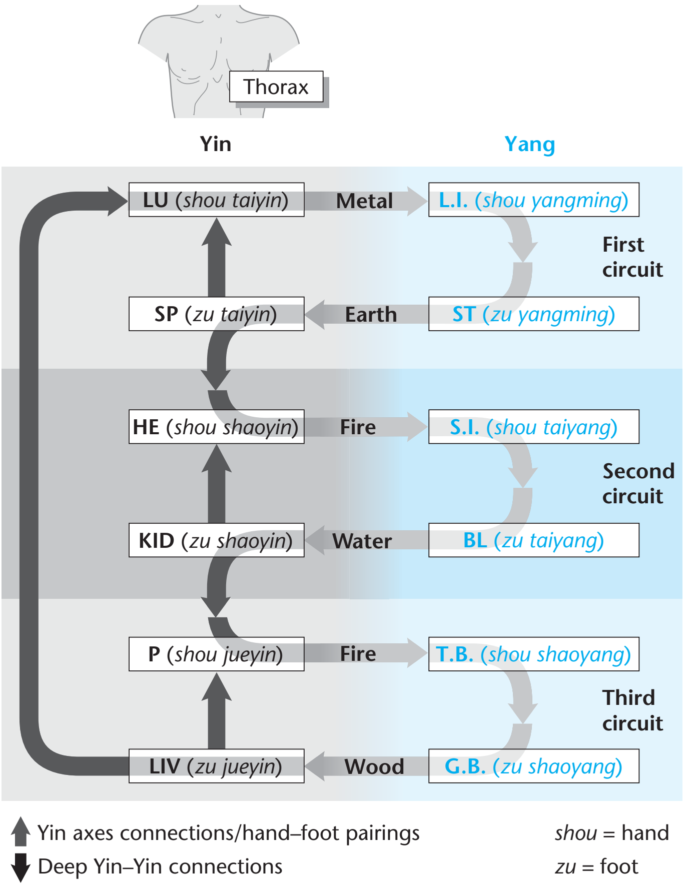
Thorax
Yin
Yang
LU (shou taiyin)
Metal
L.I. (shou yangming)
First circuit
SP (zu taiyin)
Earth
ST (zu yangming)
HE (shou shaoyin)
Fire
S.I. (shou taiyang)
Second circuit
KID (zu shaoyin)
Water
BL (zu taiyang)
P (shou jueyin)
Fire
T.B. (shou shaoyang)
Third circuit
LIV (zu jueyin)
Wood
G.B. (zu shaoyang)
Yin axes connections/hand–foot pairings
Deep Yin–Yin connections
shou = hand
zu = foot

Fig. 1.17 Overview of the Yin–Yin connections (primary channels)

Connections of the Yin axes (hand–foot pairings)

These Yin–Yin connections support the communication between the hand Yin and the foot Yin channels, which convey Qi of the same quality (*taiyin*, *jueyin*, *shaoyin*) and serve as a link between **above** and **below**. Their Qi flow does **not follow the circadian rhythm** of the Organ clock (→ Fig. 1.17).

FIRST CIRCUIT

SP channel (*zu taiyin*) ➡ LU channel (*shou taiyin*)

SP-20 (*zhourong*) ➡ **LU-1** (*zhongfu*)

According to Solinas et al. (1998), a superficial branch separates from the primary SP channel at → **SP-20** and travels to the primary LU channel at → **LU-1**.

SECOND CIRCUIT

LIV channel (*zu jueyin*) ➡ P channel (*shou jueyin*)

LIV ➡ P

An internal branch ascends from the Liver, passes through the diaphragm, connects with the primary P channel below → **P-1** (*tianchi*).

THIRD CIRCUIT

KID channel (*zu shaoyin*) ➡ HE channel (*shou shaoyin*)

KID ➡ HE

An internal branch of the primary KID channel runs to the Liver, penetrates the diaphragm and spreads into the Lungs. From the

Lung a branch travels to the Heart, where it connects with the primary HE channel.
Many authors give **HE-1** (*jiquan*) as the site of connection.

Deep Yin–Yin connections

These Yin–Yin connections facilitate the **circadian cycle** according to the Organ clock and mark the transitions of the channel circuits. They are located deep in the thorax area on the level of the *zang*-Organs.

Connection of the first and second circuit

SP channel (*zu taiyin*) ➡ HE channel (*shou shaoyin*)

SP ➡ HE

An internal branch of the primary SP channel spreads in the Heart and connects with the HE channel. Note: according to Hicks et al. (2004, → entry/exit points, → 8.1.16) **SP-21** is considered to be the exit point of the SP channel and **HE-1** the entry point of the HE channel.

Connection of the second and third circuit

KID channel (*zu shaoyin*) ➡ P channel (*shou jueyin*)

KID ➡ P

An internal branch of the primary KID channel travels to the Kidneys, then to the Liver, penetrates the diaphragm and spreads into the Lung. From the Lung, an internal branch runs to the Heart, where it meets the primary P channel and also reaches → **Ren-17** (*shanzhong/danzhong*). Note: according to Hicks et al. (2004, → entry/exit points → 8.1.16) **KID-22** is considered to be the exit point of the KID channel and **P-1** the entry point of the P channel.

Connection of the third and first circuit

LIV channel (*zu jueyin*) ➡ LU channel (*shou taiyin*)

LIV ➡ P

An internal branch of the LIV channel originates in the Liver, passes the diaphragm, spreads in the Lungs and connects with the LU channel through a network of branches. Note: according to Hicks et al. (2004, → entry/exit points → 8.1.16) **LIV-14** is considered to be the exit point of the LIV channel and **LU-1** the entry point of the LU channel.

1.2.3 The Six Great Channels (*liu jing*)

Based on the Yang–Yang and Yin–Yin connections (→ 1.2.2, → Fig. 1.16, → Fig. 1.17), primary channels of the same nature and quality can be divided into **six great channels** (*liu jing*) or **hand–foot pairings (or axes).** They represent the three great Yang channels and the three great Yin channels.

Channels	**Trajectory**		**Amount of Qi and Blood**
Exterior			
Yang channels (hand–foot pairings)			
taiyang **(S.I., BL)**	Latero-posterior aspect of the four extremities, posterior aspect of the head and body	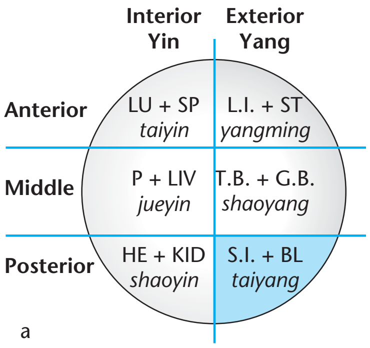 Interior Yin Exterior Yang Anterior LU + SP *taiyin* L.I. + ST *yangming* Middle P + LIV *jueyin* T.B. + G.B. *shaoyang* Posterior HE + KID *shaoyin* S.I. + BL *taiyang* a	Less Qi, more Blood **Qi < Blood**
shaoyang **(T.B., G.B.)**	Lateral aspect of the four extremities and torso	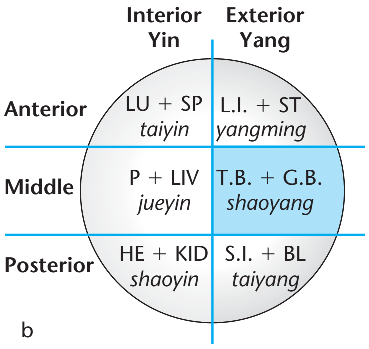 Interior Yin Exterior Yang Anterior LU + SP *taiyin* L.I. + ST *yangming* Middle P + LIV *jueyin* T.B. + G.B. *shaoyang* Posterior HE + KID *shaoyin* S.I. + BL *taiyang* b	More Qi, less Blood **Qi > Blood**

Fig. 1.18 Overview of the six great channels (*liu jing*)

Yang channels (hand–foot pairings)			
yangming **(ST, L.I.)**	Latero-anterior aspect of the four extremities, frontal aspect of the head (face, forehead), anterior aspect of the trunk	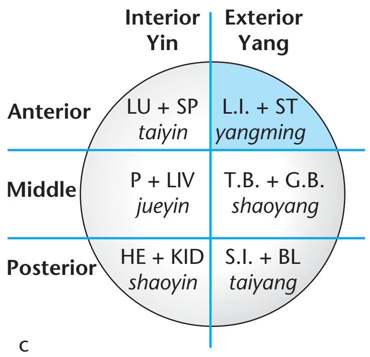 Interior Yin Exterior Yang Anterior Middle Posterior LU + SP taiyin L.I. + ST yangming P + LIV jueyin T.B. + G.B. shaoyang HE + KID shaoyin S.I. + BL taiyang c	More Qi, more Blood **Qi = Blood**
Yin channels (hand–foot pairings)			
taiyin **(LU, SP)**	Medio-anterior aspect of the four extremities as well as the thorax and abdomen	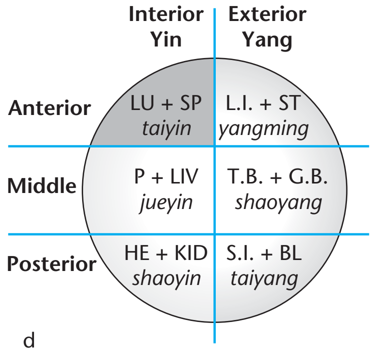 Interior Yin Exterior Yang Anterior Middle Posterior LU + SP taiyin L.I. + ST yangming P + LIV jueyin T.B. + G.B. shaoyang HE + KID shaoyin S.I. + BL taiyang d	More Qi, less Blood **Qi > Blood**
jueyin **(P, LIV)**	Middle of the medial aspect of the four extremities as well as the thorax and abdomen	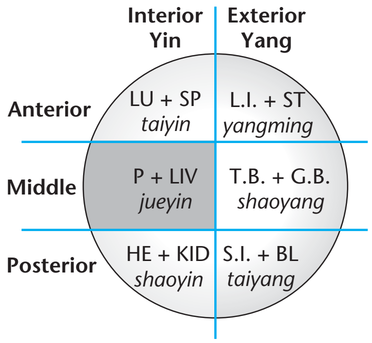 Interior Yin Exterior Yang Anterior Middle Posterior LU + SP taiyin L.I. + ST yangming P + LIV jueyin T.B. + G.B. shaoyang HE + KID shaoyin S.I. + BL taiyang e	Less Qi, more Blood **Qi < Blood**
shaoyin **(HE, KID)**	Medio-anterior aspect of the four extremities as well as the thorax and abdomen	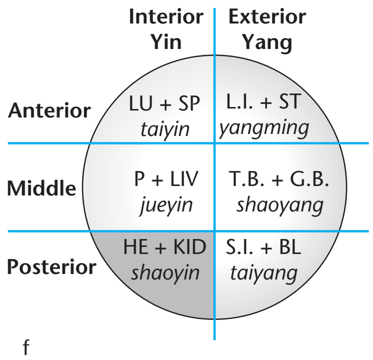 Interior Yin Exterior Yang Anterior Middle Posterior LU + SP taiyin L.I. + ST yangming P + LIV jueyin T.B. + G.B. shaoyang HE + KID shaoyin S.I. + BL taiyang f	More Qi, less Blood **Qi > Blood**
Interior			

Fig. 1.18 (cont'd) Overview of the six great channels (*liu jing*)

Names and functions of the six great channels (*liu jing*)

Tai means 'greater', 'highest'

The polarised energy (Yang or Yin) develops and reaches its maximum activity in the *taiyang* or *taiyin* channels. Here, **the maximum level of the respective polarity** is reached and begins to fall again. '*Taiyang* opens to the outside' means that it spreads to the Exterior (➞ see comments below under *yangming*, also ➞ Fig. 1.19).

Shao means 'less' or 'younger'

The polarised energy (Yang or Yin) is less in the *shaoyang* or *shaoyin* channels than in the *taiyang* or *taiyin* channels. This also manifests in their location: the *shaoyang* channels are located midway between the **inside** (or anterior aspect) and the **outside** (or posterior aspect) of the body, acting like a hinge. According to Larre and Rochat de la Vallée (1986), *shaoyang* should be translated as 'young Yang'. In their view it is positioned between the outside (the other two Yang channels) and the inside (the other three Yin channels).

Depending on the interpretation, the *shaoyin* channels are considered either to be located between the *taiyin* and the *jueyin* channels or to be the deepest of the six channels. Thus they are either seen as a hinge or as a fixed central point (for more detail see ➞ Depth organisation of the channels).

Ming means 'clear', 'bright', 'radiant', 'shining'

'*Yangming* closes towards the inside – *yangming* contracts towards the inside.'

The *yangming* channel has some unique features. Based on its course alone, it is **not 'Yang'** like the other two Yang channels, but it develops as a Yang phenomenon within the Yin aspect of the body (partially anterior pathway of the ST channel).

The character for *ming* is composed of the radicals for the sun and the moon. The character for Yang also contains the sun radical. According to the *Su Wen*, the *yangming* is where the two Yang shine together. Thus, the two suns can be symbolically interpreted as a doubling of Yang energy. *Taiyang* and *shaoyang* are 'Yang' simply due to their physical location. *Yangming*, however, is 'Yang' due to the content of its stored energy. This raises the question as to why the *yangming* channel is so abundant, why it carries so much Qi and Blood. Pirog (1996) explains this fact based on embryonic development. The 'closing' of the *yangming*' can be equated with the image of a curled-up **embryo** (➞ Fig. 1.19). Because of its contracting, 'closing' position, the energy of the *yangming* channel can be compared to the compressed energy in a pressure cooker. Thus the 'Yin' process of contracting towards the Interior results in producing Yang or, in other words, compression of the energy in the Interior. This naturally compressed energy is utilised in clinical practice. The *yangming* channels, but especially the ST channel, can be accessed, either by massage or by acupuncture, to supply the whole body with the energy stored and accumulated here. By the same token, the *yangming* channels can be needled in order to drain excess Heat – a procedure comparable to a controlled release of steam from a pressure cooker.

Jue means 'absolute', 'at the end', 'exhausted' (also, according to Wiseman, 'to revert')

The polarised energy (Yin) in the *jueyin* channel represents the terminal phase of the Yin; here polarity changes to Yang (➞ see below for more detail).

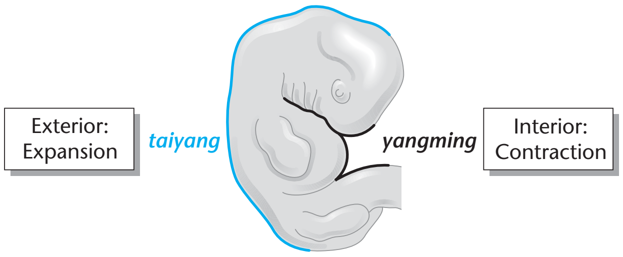

Fig. 1.19 Embryonic development of the *yangming* and *taiyang* channels (according to Pirog 1996)

The six great channels (*liu jing*): relationships and depth organisation

The systemic division of the channels into six levels or stages is mainly associated with the *Shang Han Lun*, a herbal classic. But according to Pirog (1996), this system can also be used as a paradigm for acupuncture, since it provides fundamental ideas regarding the functions of the points and channels as well as the relationships among the channels. It focuses particularly on changes of characteristics of the channel Qi, depending on which level it is acting within the body. The channels can be described as being stacked one upon each other, similar to layered rocks in a quarry. The depth of a channel or a point will be indicative of its function and significance. The postulate is that the more deeply a channel is located in the body, the more deep-seated its functions are and the more deep-seated illnesses it will be able to treat. They are not static, however. Depending on their location in the human body, they are indeed living structures with specific activities such as openers, closers, hinges or pivots, reflected in the meaning of the Chinese names. Fig 1.20 illustrates the arrangement of the six great channels from the outside to the inside, resembling the layers of an onion (➞ Fig. 1.20).

Nguyen Van Nghi (1996) compares the three Yin and three Yang channels to doors. The door (*taiyang* or *taiyin*) is closed to prevent the entry of unwelcome guests such as external pathogenic factors. But the door can also be opened to let in friends. Whether the door opens and closes well depends largely on the hinges (*shaoyang*, *shaoyin*). Only if the hinges function smoothly can the energy circulate adequately. The doorframe (or closing mechanism) (*yangming*, *jueyin*) closes towards the Interior; in other words, it protects and supports what is contained on the inside. (→ Fig. 1.21).

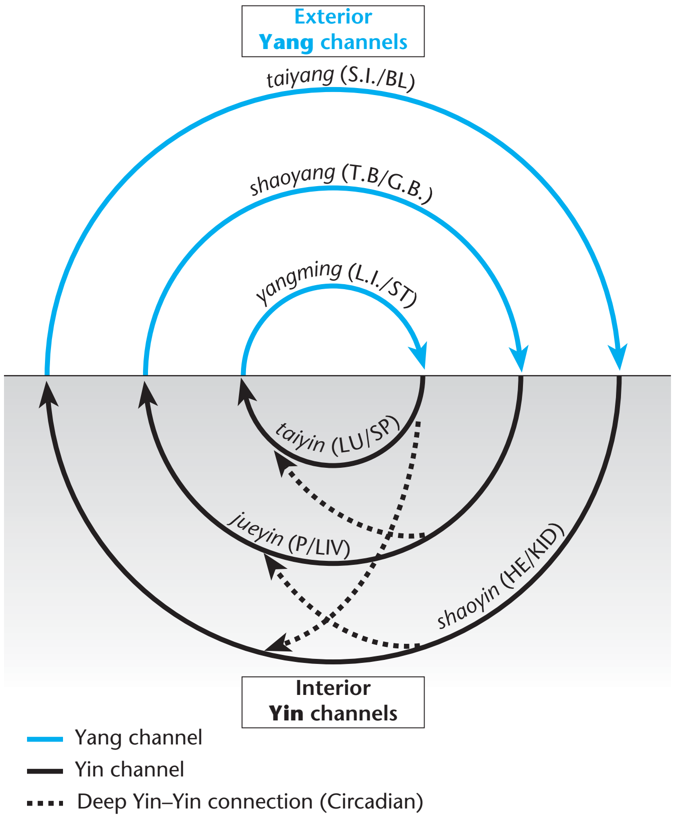
Exterior
Yang channels
taiyang (S.I./BL)
shaoyang (T.B/G.B.)
yangming (L.I./ST)
taiyin (LU/SP)
jueyin (P/LIV)
shaoyin (HE/KID)
Interior
Yin channels
Yang channel
Yin channel
Deep Yin–Yin connection (Circadian)

Fig. 1.20 Relationships of the three Yin and three Yang channels (according to Nguyen Van Nghi, 1996)

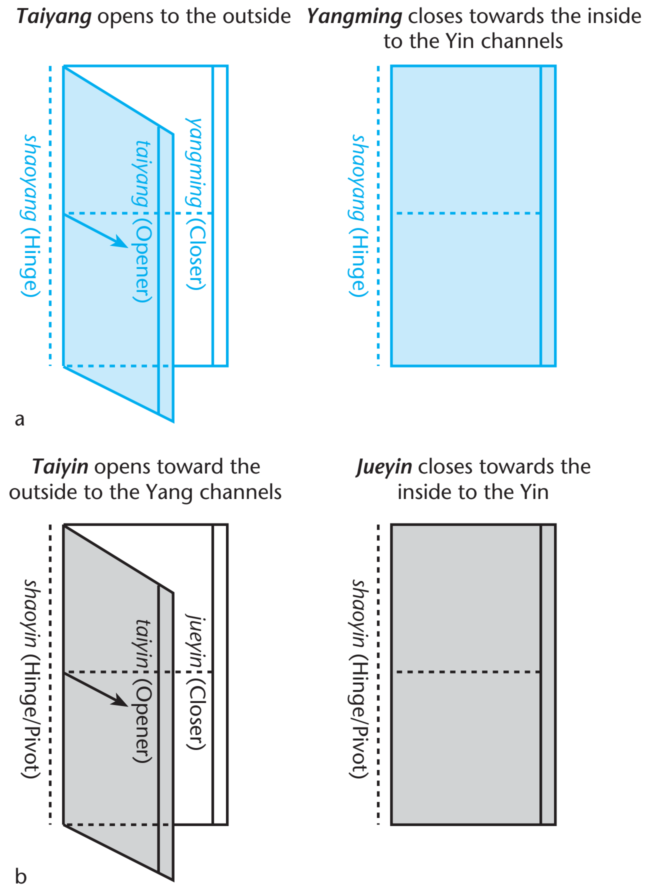
Taiyang opens to the outside
Yangming closes towards the inside to the Yin channels
shaoyang (Hinge)
taiyang (Opener)
yangming (Closer)
shaoyang (Hinge)
a
Taiyin opens toward the outside to the Yang channels
Jueyin closes towards the inside to the Yin
shaoyin (Hinge/Pivot)
taiyin (Opener)
jueyin (Closer)
shaoyin (Hinge/Pivot)
b

Fig. 1.21 The six great channels as opener, hinge and frame of a door

The literature is somewhat contradictory about which are the deeper of the six great channels, the *shaoyin* or the *jueyin* channels. Pirog (1996) has contributed an interesting theory in suggesting that the *shaoyin* channels occupy the deepest position in the body. They are located in such a medial location (near the middle of the body) that, when a person is standing in a normal position, the *shaoyin* channels are covered and not directly visible. According to Pirog, the *shaoyin* does not function like a hinge, but rather like a fixed pivot, similar to the axis of a mill wheel, which itself is immobile, but is able to cause movement. When the Qi has reached the deepest point of the *shaoyin*, and the 'Yin of the *jueyin* is exhausted', it has to revert. In other words: it has to return to a more superficial level on the outside or it has to transform into Yang. In that sense, it is the P and LIV channels that again start the flow of the Qi towards the Yang, from the inside to the outside and from below to above (→ Fig. 1.22).

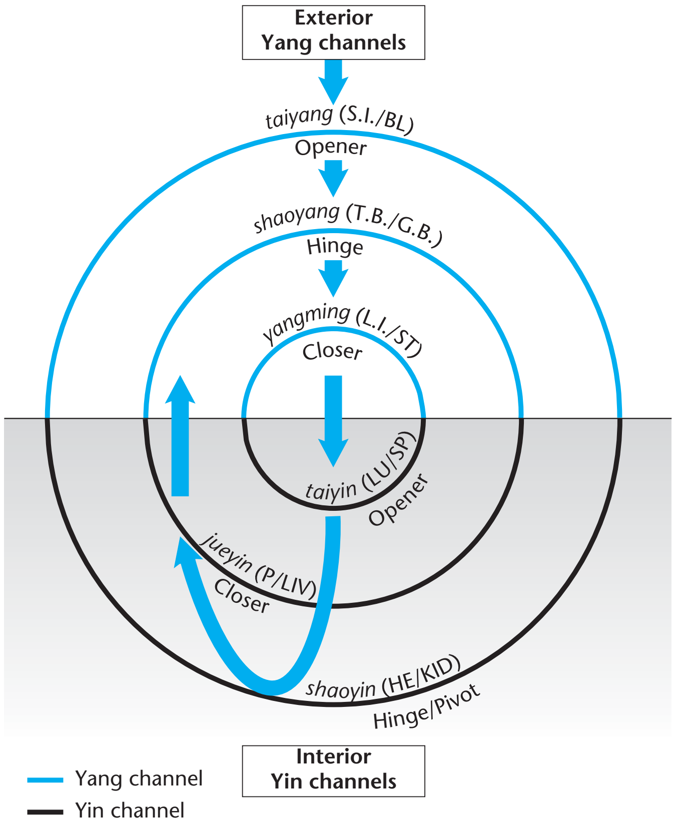
Exterior
Yang channels
taiyang (S.I./BL)
Opener
shaoyang (T.B./G.B.)
Hinge
yangming (L.I./ST)
Closer
taiyin (LU/SP)
Opener
jueyin (P/LIV)
Closer
shaoyin (HE/KID)
Hinge/Pivot
Interior
Yin channels
Yang channel
Yin channel

Fig. 1.22 Depth organisation of the six great channels (according to Pirog, 1996)

Confluent or connecting points of the six great channels (*liu jing*)

The **primary Yin channels** (*liu jing*) connect in the **thoracic region** (see also → Yin axes or hand–foot pairings → 1.2.2, Fig. 1.17). In contrast, the **primary Yang channels** connect in the **cranial region** (see also → Yang axes or hand–foot pairings → 1.2.2, Fig. 1.16). Many schools of thought suggest the connections of the channels of the same nature and quality at the following **six confluent points**:

Yin channels	**LU-1** (*zhongfu*), **P-1** (*tianchi*), **H-1** (*jiquan*)
Yang channels	**BL-1** (*jingming*), **G.B.-1** (*tongziliao*), **ST-1** (*chengqi*)

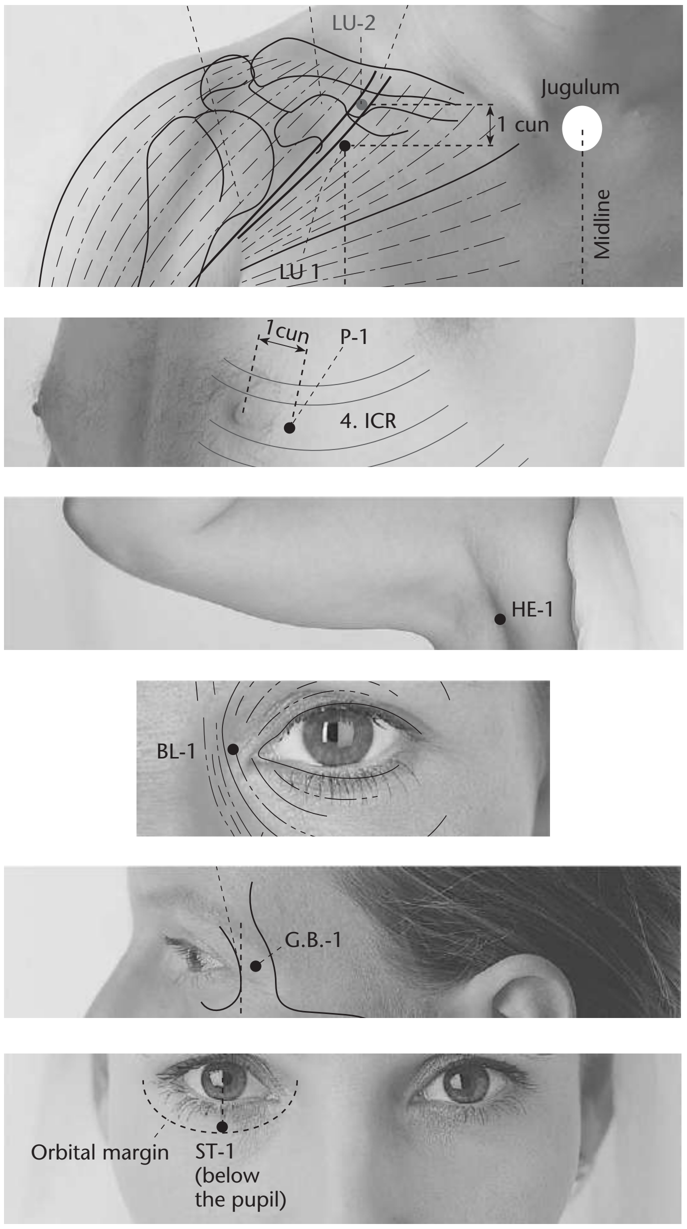
LU-2
Jugulum
1 cun
Midline
LU 1
1cun
P-1
4. ICR
HE-1
BL-1
G.B.-1
Orbital margin
ST-1
(below
the pupil)

Fig. 1.23 Confluent points of the six great channels

Starting and end points of the six great channels (*liu jing*)

Fig. 1.24 shows the starting and end points of the six great channels on the feet or hands respectively.

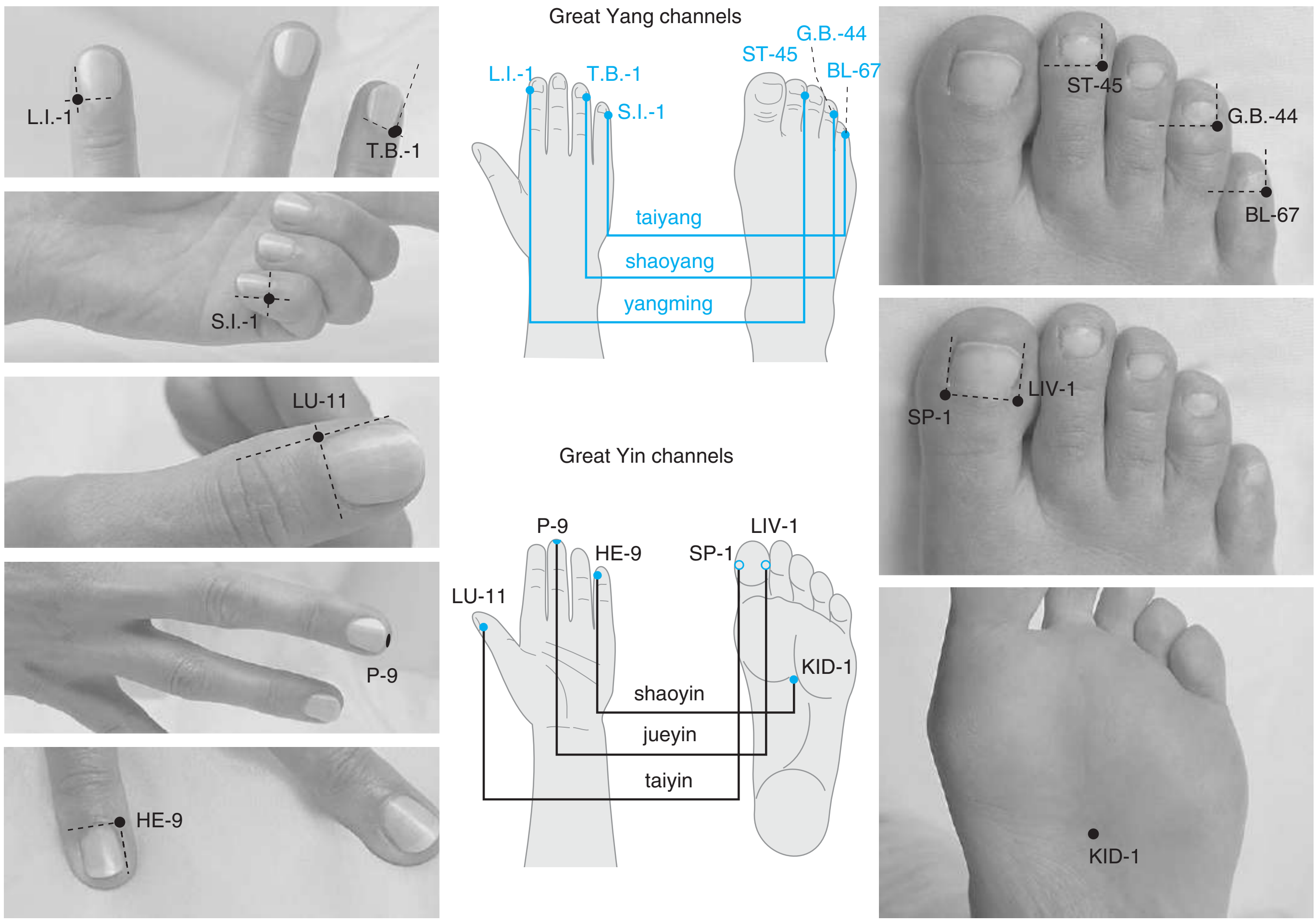
L.I.-1
T.B.-1
S.I.-1
LU-11
P-9
HE-9
Great Yang channels
G.B.-44
ST-45
BL-67
L.I.-1
T.B.-1
S.I.-1
taiyang
shaoyang
yangming
Great Yin channels
P-9
LIV-1
HE-9
SP-1
LU-11
KID-1
shaoyin
jueyin
taiyin
ST-45
G.B.-44
BL-67
LIV-1
SP-1
KID-1

Fig. 1.24 Starting and end points of the *jiu jing*

Root points and binding points of the six great channels (*liu jing*)

According to Chapter 5 of the *Ling Shu*, each of the six great channels (*liu jing*) has a root point (*gen*) and a binding point (*jie*).

Root points (*gen*)

At the root point, the polarised energy (Yin/Yang) of a great channel is at its minimum – this point marks the end of the Yang or the beginning of the Yin. The root points are all located on the foot (→ Fig. 1.24).

Root points of the three Yang channels (→ Fig. 1.24)

These always represent the **terminal point** of each great **Yang** channel.

taiyang (S.I. ➡ BL)	**BL-67** (*zhiyin*)
shaoyang (T.B. ➡ G.B.)	**G.B.-44** (*zuqiaoyin*)
yangming (L.I. ➡ ST)	**ST-45** (*lidui*)

Root points of the three Yin channels (→ Fig. 1.24)

These always represent the **first point** of each great **Yin** channel.

taiying (SP ➡ LU)	**SP-1** (*yinbai*)
jueyin (LIV ➡ P)	**LIV-1** (*dadun*)
shaoyin (KID ➡ H)	**KID-1** (*yongquan*)

Binding points (*jie*)

The binding point is the point at which the polarised energy (Yin/Yang) of a great channel reaches its maximum level. It is at this point that the hand channel and the foot channel are connected.

Binding points of the three Yang channels

These are the **switching points** of the great Yang channels (hand to foot) and are located in the **cranial region**.

taiyang (BL ➡ S.I.)	**BL-1** (*jingming*)
shaoyang (G.B. ➡ T.B.)	**T.B.-21** (*ermen*)*
yangming (ST ➡ L.I.)	**ST-1** (*chengqi*)**

Binding points of the three Yin channels

These are the **switching points** of the great Yin channels (foot to hand) and are located in the **thoracic or abdominal region**.

taiying (SP ➡ LU)	**Ren-12** (*zhongwan*)
jueyin (LIV ➡ P)	**Ren-18** (*yutang*)
shaoyin (KID ➡ H)	**Ren-23** (*lianqian*)

1.3 Divergent Channels (*jing bie*)

Synonyms: Channel divergence (Wiseman), separate pathways (Larre and Rochat de la Vallée).
Character: *bie* means diverging, separating, branching off. In this sense, the 12 divergent channels can be seen as branches of the primary channels.

General pathways of the divergent channels (→ Fig. 1.25)

The 12 bilateral divergent channels, which have no specific points of their own, are named after and are dependent on their corresponding primary channel.
They can be divided into **six great Yin and six great Yang pairs** and have their own system of six confluences (reunions): BL/KID, G.B./LIV, ST/SP, S.I./H, T.B./P, L.I./LU. According to Ramakers (course material, 2003) the divergent channels follow their own (non-cosmic) daily biorhythm:
BL ➡ G.B. ➡ ST ➡ S.I. ➡ T.B. ➡ L.I. ➡ KID ➡ LIV ➡ SP ➡ H ➡ P ➡ LU.
All divergent channels separate (*li*) from their corresponding primary channel on the extremities, except for the divergent T.B. channel, which branches off on the head. But many authors suspect that their pathways begin before the actual branching off. Based on this assumption, the divergent channels begin at the distal ends of the primary channels, running parallel to them, but as separate entities. The circulation in the divergent channels is of a centripetal nature (→ 1.1.1); in other words, the flow is always from a distal in a proximal direction, towards the trunk and the head. After separating from the primary channels, they enter deeper (*ru*) into the body and, after completing their internal pathway, they re-emerge close to the surface of the body (*chu*). The divergent channel pairs (Yin/Yang) pass together through the Interior of the body to join (*he*) their pertaining primary Yang channel in the occipital or cranial region. It is this joining of the courses that is referred to as the six confluences (or reunions) (*liu he*).

Pathways of the divergent Yin channels

After separating from their primary channels, they connect with their pertaining *zang*-Organ, then with the Interiorly-Exteriorly paired *fu*-Organ. They continue their flow in the upper half of the body together with their paired divergent Yang channel to join the paired primary Yang channel. In contrast to the divergent Yang channels, the divergent Yin channels will not return to their pertaining primary Yin channel after having completed their internal pathway (→ Fig. 1.25).

Pathways of the divergent Yang channels

After separating from their primary channels, the divergent Yang channels travel to their pertaining *fu*-Organ, then to the Interiorly-Exteriorly paired *zang*-Organ. They connect again with their pertaining primary channel and also with their Interiorly-Exteriorly paired divergent Yin channel in the upper half of the body (→ Fig. 1.25).

The divergent channels and the Heart

All divergent channels pass the thoracic region and the Heart, except for the divergent channels of the LU, L.I. and KID. According to Shima and Chase (2001), this means that all divergent channels passing the thorax will communicate with the True Qi (*zhen qi*), representing the synthesis of the Gathering Qi (*zong qi*) and Original Qi (*yuan qi*) (→ 1.1.4). They suggest that this hypothesis is in accordance with all the information available about the divergent channels, and explains furthermore why the divergent channels are so useful in the treatment of deep-seated Organ disorders as well as for disharmonies of the Defensive Qi (*wei qi*). It supports the image of a divergent channel as a fundamental axis between the Interior and the Exterior of the body. In the opinion of Shima and Chase (2001), the divergent channels are an invaluable connection to facilitate the return of the Defensive Qi (*wei qi*) to the Interior of the body and to transfer the fundamental inner Original Qi (*yuan qi*) to the periphery. In that respect the Original Yang (*yuan yang*) not only touches upon, but also strengthens the Defensive Yang (*wei-yang*).

---

* According to Nguyen Van Nghi (1996) and Solinas et al. (1998), this is G.B.-2 (*tinghui*)

** According to Nguyen Van Nghi (1996) and Solinas et al. (1998), this is ST-8 (*touwei*)

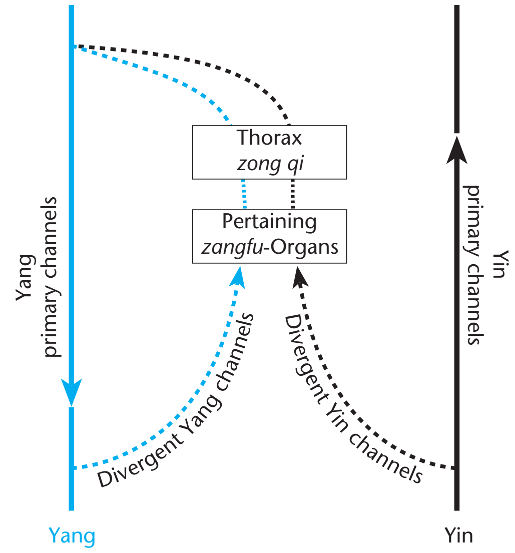

Fig. 1.25 Schematic illustration of the divergent channel pathways (according to Shima and Chase 2001)

Functions

The divergent channels reinforce the circulation of and communication between the primary channels in the Interior of the body. On the one hand, this is achieved due to their connection with the primary Yang channels (➞ pathways). But their organisational structure also supplements the primary channels in a functional way – for example they also control areas of the body not covered by the primary channels. Thus points on the primary Yin channels can affect the head and face, in spite of their 'external' pathways ending in the thoracic region. This effect can be explained by the fact that the Yin divergent channels, after branching off from their pertaining primary Yin channel, pour into the Yang divergent channels flowing to the head. A deeper understanding of the trajectory and functions of the divergent channels will in turn lead to a better understanding of the clinical application of some commonly used acupuncture points. Since their pathway passes the Heart and thorax, the divergent channels further strengthen the connection to the Heart (➞ explanations above). According to some schools, they can therefore also be used for the treatment of psychosomatic disorders. Many interpretations and variants exist regarding the pathways and clinical application of the divergent channels (for a good overview see ➞ Shima and Chase 2001).

1.4 The 12 Sinew Channels (*jing jin*)

Synonyms and translations: muscle channels, muscular meridians/the muscles in connection with the meridians (Larre and Rochat de la Vallée), muscle regions, channel sinews.

Principles of the sinew channel system

The sinew channels mainly represent the muscles, sinews and ligaments located along the pathways of the primary channels and their pertaining connecting vessels. They do not have their own specific points and also no direct connection to the *zangfu*-Organs. The 12 bilateral sinew channels are connected to and named after their pertaining primary channel. Each sinew channel has its own pathway, which generally follows the trajectory of its pertaining primary channel. However, in contrast to the primary channels, the sinew channels always begin at the extremities, mostly separating from the primary channels at the *jing*-well points (➞ 8.1.6) on the tips of the fingers and the toes. Their pathways always run from the extremities to the trunk or further to the head and face (➞ centripetal circulation ➞ 1.1.1). They cover a larger area than the primary channels, in the shape of a band and binding (*jie*) like a spindle at the joints and other areas of the body. This allows them to cover areas of the body which are not reached by the primary channels or the divergent channels. It further explains the actions of some of the points located on their pertaining primary channel. Under pathological conditions, the sinew channels will manifest as so-called *ashi*-points (*a shi xue*) or as trigger points.

Terms

Binding (*jie*) refers to spindle-shaped concentrations of the sinew channels, which tend to be located on the bigger muscles and joints (for more details on specific sinew channels see ➞ Chapter 4).
The term *ju* (point of confluence, gathering) is used to signify an area where two or more sinew channels are bundled together. In the larger muscles, the Qi disperses or spreads (*san*), so that they can be divided into smaller muscle groups. The sinew channels connect (*luo*) the muscles, sinews and ligaments with the joints in order to facilitate mobility.

Rhythm of Qi flow

According to Ramakers (course material, 2003) the sinew channels are subject to their own daily rhythm of Qi flow:
BL ➡ G.B. ➡ ST ➡ S.I. ➡ T.B. ➡ L.I. ➡ SP ➡ LU ➡ LIV ➡ P ➡ KID ➡ HE.

General pathways of the sinew channels (→ Fig. 1.26)

- The three hand Yang sinew channels (L.I./S.I./T.B.) flow from the finger tips towards the scapula, ascend to the ear and intersect with **G.B.-13** (*benshen*)
- The three foot Yang sinew channels (ST/BL/G.B.) run from the tips of the toes towards the head, join each other on the cheek and meet with **S.I.-18** (*quanliao*)
- The three hand Yin sinew channels (HE/P/LU) flow from the finger tips towards the thorax, joining within the thoracic musculature and meeting at **G.B.-22** (*yuanye*)
- The three foot Yin sinew channels (LIV/KID/SP) run from the tips of the feet towards the thorax, join each other in the genital region and, depending on the author, intersect in the area of **Ren-2** (*qugu*) or **Ren-3** (*zhongji*)

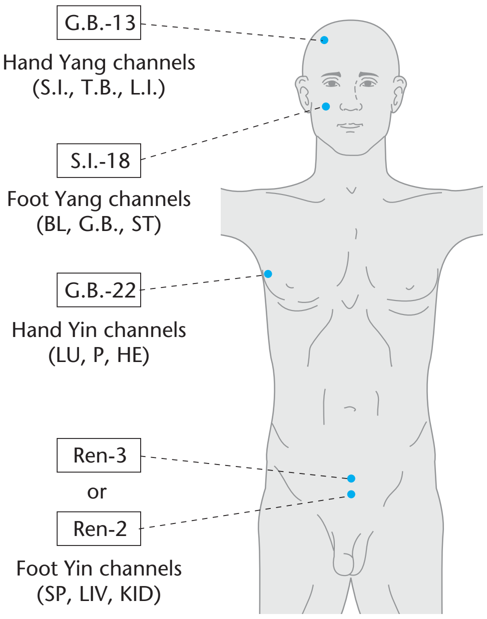

Fig. 1.26 Meeting points of the sinew channels

Functions

The main function of the sinew channels is to distribute Qi and Blood (*xue*) over the surface of the body as well as to integrate the muscles, sinews and ligaments with the joints. Generally, the sinew channels join groups of synergistic muscles, supporting their mobility and natural movement while also serving the function of integrating the surface of the body. At the same time they protect the bones and skeletal structure, linking the structures of the body.

Proper functioning of the muscles is dependent on the Spleen (*pi*) and the Liver (*gan*) and particularly on their Blood (*xue*). While the Yin (or material) aspect of the musculature, the 'flesh', is governed by the nourishing function of the Spleen, the functioning of the muscles and sinews is governed by the Liver. But the Liver is also connected to the mesenchymal structures of the body, especially those surrounding the organs. The sinew channels are not only composed of the muscular structures, but also of the mesenchymal formations, so that according to Larre and Rochat de la Vallée they are also involved in the structural aspect of the diaphragm as well as the thoracoabdominal serous membranes, in particular the pleura, peritoneum and pericardium. In the deeper parts of the body they act through the deep muscular fascia. One of the main functions of the sinew channels is that of a 'protective coat' that shields the body against attacks from the Exterior. In this respect the Defensive Qi (*wei qi*) is of particular importance ((→ 1.1.4): the muscle spindles are crossed by loose and diffuse flows of Qi, which also contain Defensive Qi (*wei qi*). Nutritive Qi (*ying qi*) is also present, particularly at the outer ends of the spindles (→ 1.1.4).

The sinew channels also have the task of barring access to the deeper channels. In other words, in a pathological situation, a deeper channel can only be reached if the resistance of the muscular layers can be broken. An important aim of Tuina therapy is to remove muscular tension. According to Larre and Rochat de la Vallée (1996), the sinew channels also play a role in psychosomatic disorders since they very effectively constitute the muscular carapace. The authors refer here to a bioenergetic interpretation of muscular tension.

Clinical and therapeutic importance

Disorders of the sinew channels will show the following symptoms along their pathway: muscular tension, pain, cramping, spasms, stiffness, but also paralyses, weakness and swellings of the muscles, tendons and ligaments. Any impairment of range of motion of the joints also indicates that the sinew channels are involved. In addition, each sinew channel also has its specific symptoms (→ chapter 4, individual channel descriptions).

The following factors can contribute to disorders of the sinew channels:

- *Bi* syndromes – external invasion of pathogenic Wind, Cold, Damp or Heat
- Traumatic injuries
- Muscle strain through over-use
- Muscle strain and contraction due to longstanding emotional and mental stress

While the sinew channels do not have their own specific points, they can easily be accessed by needling *ashi*-points, as well as through adjunctive therapies such as cupping, Tuina massage, plum blossom needling, moxibustion and *gua sha* (scraping technique). Distal points can also be used in treating disorders of the sinew channels.

For traumatic injuries with limited range of motion, needling of distal points with simultaneous movement of the affected joint has proven to be very effective. The distal points (→ 8.2.1) can be chosen by first determining the affected area, then needling points in the corresponding area on the contralateral side.

For example, for a knee injury, the exact location of the pain is determined, then the corresponding point or area on the contralateral elbow is needled or massaged. Fig. 1.27 illustrates the areas that have been found to be effective. These areas are also referred to as reciprocal areas (for more detail also see → 8.2, 8.3).

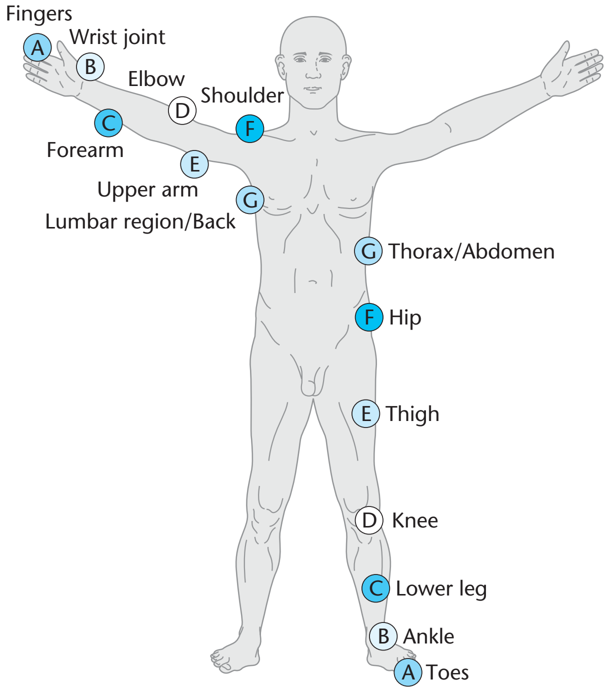

Fig. 1.27 Corresponding areas of the body

1.5 *Luo*-Connecting Vessels

Synonyms: secondary pathways (Larre and Rochat de la Vallée), network vessels, collaterals

Organisation

The *luo*-connecting vessels can be divided into 15/16 great *luo*-connecting vessels as well as into many smaller and more superficial *luo*-connecting vessels. Looking at the *luo*-connecting vessels altogether, they form a network that spans the entire body. And all *luo*-connecting vessels in turn connect with the great *luo*-connecting vessel of the Spleen (at → **SP-21**), which plays a major role within the network of the *luo*-connecting vessels (*Ling Shu*, Chapter 10). Together with the primary channel system, they form a network of vessels which distribute Qi and Blood not only within the body, but also to its surface (→ Fig. 1.3, 1.4).

The 16 great *luo*-connecting vessels

Traditionally there are 15 *luo*-connecting vessels, but if one includes the great *luo*-connecting vessel of the Stomach (*wei zhi da luo* or *xu li*), it raises the number to 16:

- 12 *luo*-connecting vessels (*luo mai*), related to the 12 primary channels and beginning at the extremities
- 4 *luo*-connecting vessels beginning at the thorax:
  - 2 *luo*-connecting vessels of the *ren mai* and *du mai*
  - 1 great *luo*-connecting vessel of the Spleen
  - 1 great *luo*-connecting vessel of the Stomach.

Except for the Great *luo*-connecting vessel of the Stomach (*xu li*) which originates in a *fu*-Organ, all other 15 *luo*-connecting vessels emerge at a specific point on their pertaining primary channel, the *luo*-connecting point (→ 8.1.2). 13 *luo*-connecting points are located on primary channels, two on extraordinary vessels (*ren mai* and *du mai*). The *luo*-connecting points are located on areas of the body, where the *luo*-connecting vessels that connect the Interiorly-Exteriorly pertaining channels with each other divide into smaller vessels. The 15/16 *luo*-connecting vessels thus fulfil a controlling function over all the other, smaller *luo*-connecting vessels. Each *luo*-connecting vessel has its own pathology according to its pertaining primary channel (→ introductions to the primary channels → Chapters 4 and 5).

Superficial *luo*-connecting vessels

The smaller superficial *luo*-connecting vessels are sub-branches of the ‘great’ *luo*-connecting vessels. They, in turn, are divided into the *sun luo*, vertical branches which are sometimes also referred to as tertiary branches or grandchild-vessels. The *sun luo* not only spread over the surface of the body, but also cover the internal Organs. The *sun luo* further divide into the more superficial *fu luo*. The *fu luo* again divide horizontally into small branches called the *xue luo* (blood vessel network).

Pathogenic factors tend to enter the body through the superficial *luo*-connecting vessels. By the same token, disorders such as Qi or Blood stagnation in a primary channel or an Organ may manifest in the *xue luo* as skin discolorations, spider veins, etc.

Clinical importance

The *luo*-connecting vessels can be accessed through superficial needling techniques as well as skin-sensitising techniques such as plum blossom needling, bloodletting, cupping, *gua sha* and moxibustion.

> New diseases are located in the primary channels; old (chronic) diseases lodge in the *luo*-connecting channels.

1.6 Cutaneous Regions (*pi bu*)

Characteristics

The cutaneous regions form the most superficial aspect or outer cover of the primary channel system (*pi* = skin). According to Ramakers (course material, 2003), they can be compared to a 'big bag that holds everything together' and – besides the extraordinary vessels – they represent the oldest tissue in the body, which developed from the ectoderm. The cutaneous zones are not channels *per se*, but rather areas of skin connected to and covering the network of superficial channels and vessels. They are named after the six great channels (*liu jing* → 1.2.3) (Fig. 1.28).

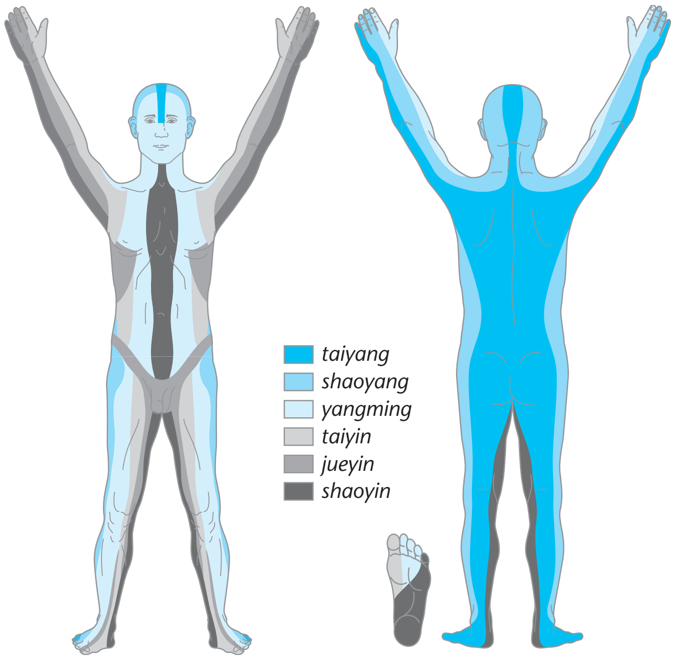

Fig. 1.28 Cutaneous zones (anterior and posterior aspect)

Functions

The cutaneous zones circulate Blood and Qi, especially Defensive Qi (*wei qi*), to the surface of the body and regulate the skin and pores, thus strengthening the body's defences. Proper functioning depends on the Lung Qi, which governs the skin, and the Defensive Qi (*wei qi*) (→ 1.1.4). Pathogenic factors can enter and exit the body through the cutaneous zones.

Clinical importance

Disorders of the primary channels and their pertaining *zangfu*-Organs may be reflected on the outside of the body by the cutaneous zones, for example pathological skin sensations, lesions or discolorations may occur. A bluish-greenish discoloration may indicate pain or obstructions, a reddish colour Heat and a pale-whitish colour deficiency or Cold. Therefore, the cutaneous zones can serve as a diagnostic tool for skin disorders or pain, specifically for chronic pain. Treatment techniques include superficial needling, plum blossom needling, cupping, massage, and *gua sha* (Chinese scraping technique). The author had good results in the treatment of chronic pain with the somatopic wrist/ankle acupuncture developed by Dr Zhang Xin Shu (→ JCM, No. 37,11/1991, → Focks and Hillenbrand 2003). The strictly subcutaneous needling used in this technique influences, of course, the cutaneous zones.

1.7 The Eight Extraordinary Vessels (*qi jing ba mai*)

Synonyms: Extraordinary meridians, Wondrous meridians
Meaning of *qi jing ba mai*:
Qi: extraordinary, strange, not paired, miraculous, wonderful
Jing: channels, meridians
Ba: eight (first differentiation)
Mai: vessels

The eight extraordinary vessels are not channels such as the *jing mai*, but they are – as the name implies – extraordinary. While they are often mentioned together with the primary channels (*jing mai* → Fig. 1.3), they still play a special role.

1.7.1 Theories about the Origin of the Extraordinary Vessels

The eight extraordinary vessels are often considered to be the deepest and most original structures of the body and channel system, developing very early on during the embryonic phase. They are considered to 'cover all of human activity' (Larre et al 1986). They also initiate all secondary functions and always remain operative. Therefore, their pathologies manifest with complex symptoms and include pre- and postnatal Essence problems. The *Nanjing* (Unschuld's translation, 1986) is the first to describe the extraordinary vessels as a complete system. Earlier records only mention them individually in the context of their respective functions.

Origins of the extraordinary vessels

In order for the manifestations of life to develop, the body requires structures that keep the rhythm of the Qi and its related processes in harmony and balance. The *Nanjing* introduces the concept of an energetic centre in the human body, the 'moving Qi between the Kidneys', which is often seen as connected with the *mingmen* and *qihai dantian* and is subject to the influences of the cosmic as well as of the prenatal and postnatal Qi.

According to Matsumoto and Birch (1986), the 'moving Qi between the Kidneys' represents the *Taiji* of the body, the Great One, the origin of Yin and Yang. It is considered to be the root and origin of all human systems and structures and, according to some authors, it has a connection to the Original Qi (*yuan qi*) (→ 1.1.4). The *chong mai*, as the oldest vessel, is sometimes equated with the 'moving Qi'. The *ren mai* and *du mai* represent the basic division of the Yin and Yang in the body; in other words, they originate from the 'moving Qi' or *chong mai* as Yin and Yang branches. During the 'division', one movement ascends the front of the body (*ren mai*), the other movement the back (*du mai*).

Heaven and Earth have noon and midnight, the body has the *ren mai* and *du mai* as its polar axes. The *du mai*, *ren mai* and *chong mai* have different names, but in the end they are the same and have the same significance. The *chong mai* ensures the inseparability or unity of the *du mai* and *ren mai* – of Yin and Yang. If we tried to separate Yin and Yang, we would have to realise that they are an 'inseparable whole, a unit'.

Therefore, the extraordinary vessels function as a kind of matrix, which is the source of the energetic functioning and the cellular divisions in the body. This matrix is present from the moment of conception and organises the structuring of the embryonic development from the earliest stages.

Explanation of Fig. 1.29 (after Matsumoto and Birch)

In the beginning, there is the Great One, the *Taiji* (first level). At this level of 'no-form', the Great One begins to separate (second level) and becomes Yin and Yang (third level). At the fourth level, Yin and Yang separate twice (two branches):

- One branch **separates in a threefold way** to become the **Triple** Burner (*san jiao*) and its associated forms of Qi. Further separations lead to the formation of the **six** great channels (*liu jing* → 1.2), the **Five** Phases and **their source**, as well as the 12 primary channels.
- The other branch separates at the fourth level to form the **four** body areas (left, right, above, below) and then the **eight** extraordinary vessels.

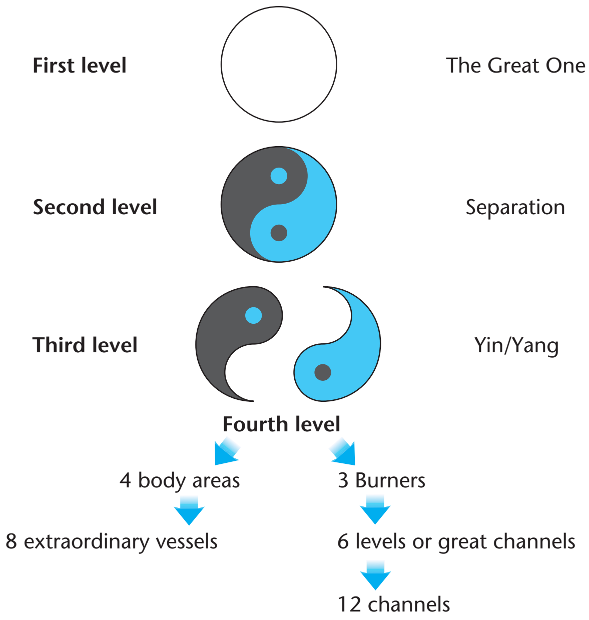

Fig. 1.29 Origin of the channels

Scars of the first cell division

Matsumoto and Birch (1986) comment on the work of Manaka and others, who interpret the extraordinary vessels with regard to embryonic development, regarding them as scars or 'lines of separation of the first cell divisions'.

The *du mai* and *ren mai* are the result of the first division of the fertilised egg, while the *dai mai* develops during the second division. At this stage, the embryo is essentially a torso, while the extremities and the more 'peripheral' extraordinary vessels – the *yang qiao mai* and the *yin qiao mai* as well as the *yang wei mai* and the *yin wei mai* – only form during subsequent divisions.

The eight extraordinary vessels and body structure

- According to Manaka et al (1995), a close connection exists between the eight extraordinary vessels and **body** structure/symmetry.
- The ***du mai*** and the ***ren mai*** separate the body **vertically** into **left** and **right**: the *ren mai* marks the anterior midline, the *du mai* ascends the posterior midline.
- The ***dai mai*** circles the waist **horizontally,** forming a **horizontal axis**. It separates **above** from **below** (→ Fig. 1.30 a).

In a further development of this idea, Manaka describes the division of the body into **eight areas,** using the structure of a topological octahedron for illustration. (→ Fig. 1.30 b). In this model, the eight extraordinary vessels directly relate to the lines 'separating' the faces of the octahedron:

- Due to their location on the anterior/posterior midline, the ***ren mai*** and the ***du mai*** separate the body into its right and left section.
- The ***dai mai*** separates above and below.
- The **G.B. and T.B. channels** compare to the anterior and posterior dividing lines at the lateral edges of the Yang aspect of the body.

- The **P** and **SP channels** compare to the anterior and posterior dividing lines at the lateral edges of the Yin portion of the body.

The opening points of the extraordinary vessels, with the exception of **SP-4** (a point on the LIV channel would be more appropriate), represent this division:

- **P-6** and **T.B.-5** are the opening points of the *yin wei mai* and *yang wei mai* respectively
- **SP-4** and **G.B.-41** are the opening points of the *chong mai* and *dai mai* respectively.

These four channels divide the **anterior** and the **lateral** aspects of the body and connect the **upper** and **lower** parts, the four associated extraordinary vessels functioning accordingly. The *yin qiao mai* and the *yang qiao mai*, due to their connection to the KID and BL channels (opening points **KID-6** and **BL-62**) have a close relationship to the *ren mai* and *du mai*. Furthermore, the pathway of the KID channel is located close to the *ren mai* while the BL channel is close to the *du mai*.

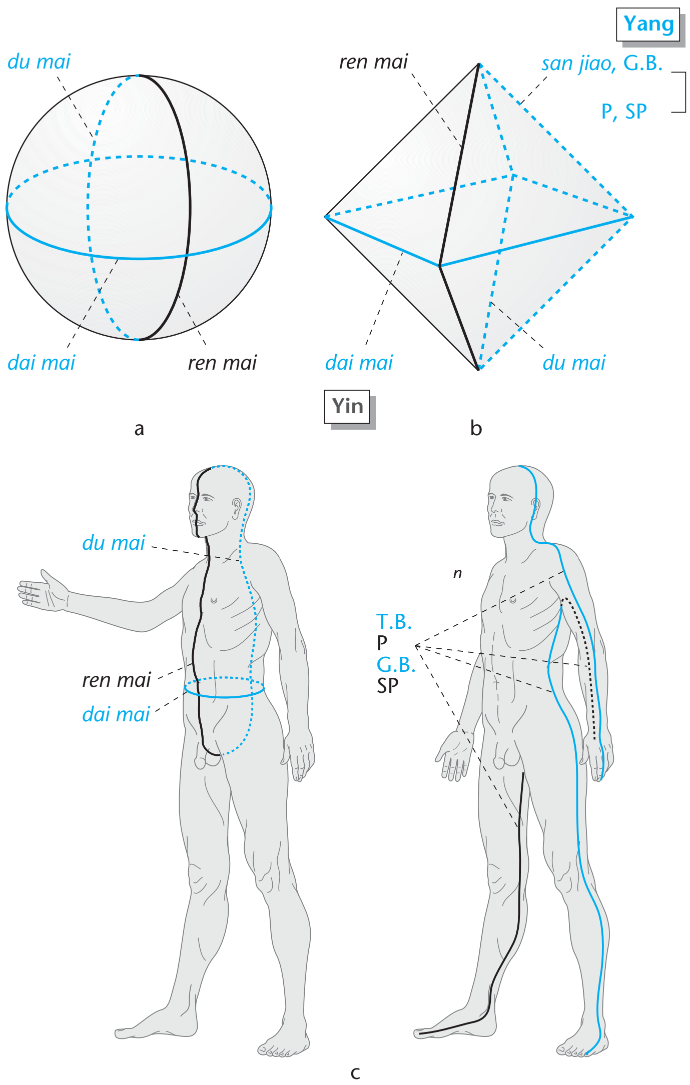

Fig. 1.30 Development of the octahedral structure based on body symmetry

1.7.2 Principles and Functions

'The primary channels are the rivers, the extraordinary vessels are the lakes.'

In some of the classics, the extraordinary vessels are compared to reservoirs, which store excess Yang, Yin, Qi or Blood, releasing it into the primary channel system in times of need. With the exception of the *ren mai* and *du mai* they do not have their own points, but they can be accessed through and are connected by points on the primary channels (see pathways of the primary channels → Chapter 5). All extraordinary vessels (with the exception of the *dai mai* with its horizontal trajectory) begin at the lower extremities or in the lower part of the torso and travel to the head. They can be activated by needling their opening and coupled points (differential therapy → 8.1.8).

Reservoir and compensation function

'*…when the network vessels (the primary channels) are filled to overflowing, none of the… conduits could seize any of their contents, and it is only then that the surplus contents of these vessels flow into the single conduit vessels (the extraordinary vessels)*' (*Nanjing*, translation by Unschuld 1986).

The *Nanjing* compares the extraordinary vessels to reservoirs, which are able to absorb a surplus of Qi and Blood in the primary channels, releasing them again during periods of deficiency. For example, if pathogenic factors penetrate the primary channel system, this excess can be absorbed by the extraordinary vessels. This will ensure that the *zangfu*-Organs remain unaffected. In that sense, the extraordinary vessels have a regulating compensation function.

Reproduction and fertility

The *chong mai* and the *ren mai* in particular, but also the *du mai*, play an important role in reproduction and fertility.

Connecting the primary channels

- The *chong mai*, also known as the '**sea of the 12 primary channels**', connects the primary ST and KID channels. It also strengthens the connection between the *ren mai* and the *du mai*.
- The *ren mai* as the **sea of Yin** connects all Yin channels
- The *du mai* as the **sea of Yang** connects all Yang channels at **Du-14** (*dazhui*).
- The *dai mai* circles the body at the waist and binds the vertical trajectories of all 12 primary channels in general, but especially the *chong mai*, the *ren mai* and the KID, LIV and SP channels.
- The *yin wei mai* **dominates the Interior** of the body and connects the KID, SP and LIV channels and the *ren mai*.
- The *yang wei mai* **controls the Exterior** of the body and connects the BL, G.B., T.B., S.I. and ST channels and the *du mai*.
- The *yin qiao mai* connects the KID and BL channels and controls balance.
- The *yang qiao mai* connects the BL, G.B., S.I., L.I. and ST channels and controls activity.

In addition, the eight extraordinary vessels are regarded as **connecting channels for the extraordinary *fu***-Organs.

Protection of the body

The *chong mai*, *ren mai* and *du mai* circulate the Defensive Qi (*wei qi*) in the thorax, abdomen and the back, and thus help to protect the body against external pathogenic factors. Due to their origin, they are also directly linked to the Kidneys, the pre-Heaven Qi and to the constitutional forces.

1.7.3 Comparison of the Eight Extraordinary Vessels with the Primary Channels

	Primary channels	Eight extraordinary vessels
Points, trajectory	Specific, bilateral points	Only the *ren mai* and *du mai* have their own points (and internal pathways); all other extraordinary vessels 'borrow' points from the primary channels. The four central vessels are unilateral (except for the *chong mai*, which also has a pathway on the extremities); the four peripheral vessels are bilateral.
Circulation	Continuously circulate *Nutritive Qi* and Blood	Circulate Essence Qi, Nutritive Qi and Defensive Qi, but according to some authors they do not have a continuous flow of Qi
Depth	Superficial, external pathway and a deeper, internal pathway; directly connected to their pertaining *zangfu*-Organs	Compared to the internal pathways of the primary channels they are more superficial, compared to the external pathways of the primary channels they are deeper. They have **no direct connection** to the *zangfu*-Organs, but their origin – especially of the four central vessels – lies deep within the body.
Connections	They connect the Exterior and Interior (through internal/ external pathways and through Yin/Yang channels) as well as above and below (hand–foot pairing, axes)	They do not connect the Interior with the Exterior, nor do they connect Yin and Yang.
Functions	Guide excess to the eight extraordinary vessels or to other secondary channels	Reservoir function: excess of Qi and Blood in the primary channels flows to the eight vessels and are stored there. In times of deficiency they are transported back into the primary channels.
Diagnosis	Diagnosis based on channel theory (meridian therapy), in TCM according to the *ba gang* (*zangfu*, pulse and tongue)	Diagnosis is characterised by specific arrays of symptoms and specific pulses.

1.7.4 Pairing and Special Functions

The eight extraordinary vessels can be paired in several ways (→ Fig. 1.31):

- The central vessels and the peripheral vessels can be coupled to form Yin/Yang pairs
- Or a central vessel can be paired with a peripheral vessel. These pairs support particular regions and are very important in acupuncture therapy (→ Opening points 8.1.8).

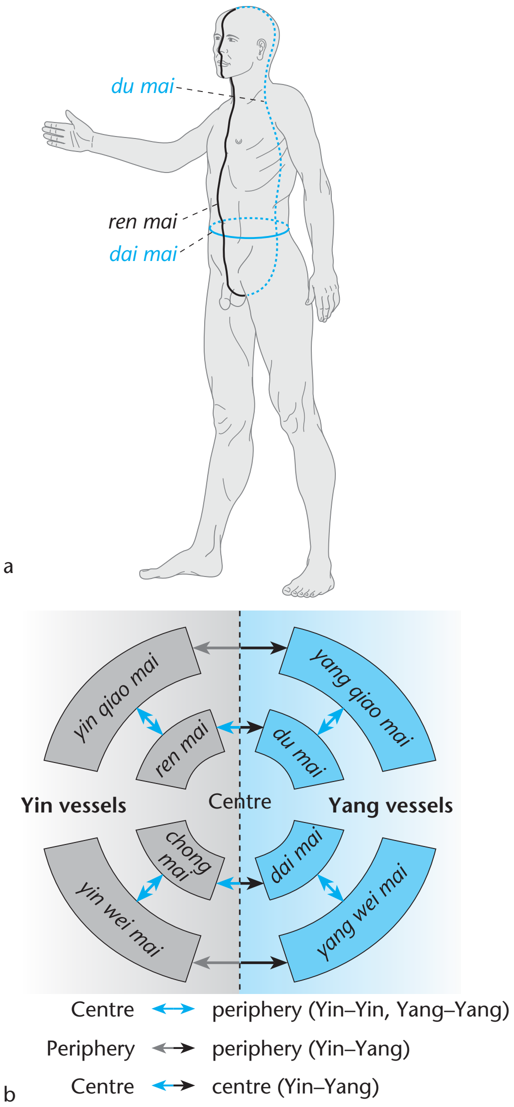
du mai
ren mai
dai mai
a
yin qiao mai
ren mai
du mai
yang qiao mai
Yin vessels
Centre
Yang vessels
chong mai
dai mai
yin wei mai
yang wei mai
Centre ↔ periphery (Yin–Yin, Yang–Yang)
Periphery ↔ periphery (Yin–Yang)
Centre ↔ centre (Yin–Yang)
b

Fig. 1.31 Pairing of the extraordinary vessels (b, graphic overview modified according to Pirog 1996)

Pairing according to Yin/Yang polarity

Pairing of the extraordinary vessels according to Yin/Yang polarity

Yin vessels	Yang vessels	Functions of the Yin/Yang pair	General functions – Centre/periphery
Central vessels (pathways on the head and torso)			
ren mai	*du mai*	The *du mai* can be considered to be the first manifestation of the Qi and Yang, the *ren mai* the first manifestation of the Blood and Yin. This pair **stores the Yin Qi and Yang Qi**.	**Functions:** Storage and distribution (of Yin and Yang). Foundation of the eight extraordinary vessels, stores Essence Qi, is responsible for the support of the internal organs and extraordinary *fu* organs, as well as for the fundamental body structures
chong mai	*dai mai*	The *chong mai* is like a harmonious connection between Yin and Yang, it expresses vitality and sets impulses. The *dai mai* encircles; it is like a container and prepares the circulation of Yin and Yang. The pair **circulates the Essence (*jing*) and Qi between the Yin and Yang regions of the body**.	
Peripheral vessels (pathways from the lower extremities to the torso)			
yin qiao mai	*yang qiao mai*	This pair **creates balance between Yin and Yang** (for example balance between Essence (*jing*) and Qi, Blood and Qi, left and right, Nutritive Qi and Defensive Qi, also on the outside – between day and night), controls and regulates the ascending (*yin qiao mai*) and descending (*yang qiao mai*) movement in the body, is responsible for balance and rhythm between Yin and Yang.	**Functions:** Balance, circulation and connection (of the Yin and Yang aspects). Provides contact between the distal regions of the body and the torso and head.
yin wei mai	*yang wei mai*	The *yin wei mai* governs the Interior (*li*), it organises and creates connections within the *Yin* aspects of the body. The *yang wei mai* governs the Exterior (*biao*), it organises within the Yang aspects of the body. This pair **connects the Yin and Yang respectively.** 'That which moves and circulates between the *Yin* channels is called the *yin wei* – that which moves between the Yang channels is called the *yang wei*.'	

Central–peripheral pairing

Here, a central vessel is combined with a peripheral vessel. This pairing probably refers to regions of the body which are traversed and supplied by the pair.

Central–peripheral pairing of the extraordinary vessels

Pair	Vessel	Relationship	Opening point	Coupling point	Supported regions
Pair 1	*chong mai*	father	SP-4	P-6	Heart, thorax, Stomach
	yin wei mai	mother	P-6	SP-4	
Pair 2	*du mai*	husband	S.I.-3	BL-62	Inner canthus of the eye, occiput, shoulders and back, S.I. and BL channels
	ren mai	wife	BL-62	S.I.-3	
Pair 3	*dai mai*	man	G.B.-41	T.B.-5	Lateral eye and temple region, ears, cheeks, occiput and shoulders
	yang wei mai	woman	T.B.-5	G.B.-41	
Pair 4	*ren mai*	host	LU-7	KID-6	Face, throat, thorax, Lungs, diaphragm, abdomen
	yin qiao mai	guest	KID-6	LU-7	

2 Location Methods and Cun Measurements

Claudia Focks

One of the basic prerequisites for any successful acupuncture treatment, besides a differential Chinese Medicine diagnosis and the relevant point selection, is the correct point location.

2.1 Location Methods for Body Acupuncture

There are various methods for locating acupuncture points:

Anatomical approach: many acupuncture points are situated at clearly defined anatomical locations, for example in depressions, at muscle and tendon insertions, in grooves, at joint clefts, at bony prominences, etc. With some experience, the palpating finger will often automatically ‘fall’ into the right depressions and holes.

Various anatomical structures and zones form the basis for point location. These are described in more detail in Chapter 3. We can distinguish between the following:

- **Fixed markers** are not affected by the body position or by movement. These include bony landmarks such as depressions or prominences of the skeletal system, but also the finger- and toenails, the nipples, the umbilicus, etc. Most of the proportional cun measurements (→ 2.2) rely on fixed markers.
- **Movable markers** often become more clearly defined with a particular body posture or movement (→ 2.3.2). For example, by flexing the elbow, the cubital crease becomes much more clearly visible to locate **L.I.-11** precisely. Equally, the most distal palmar crease becomes much more visible when making a loose fist, in order to locate **S.I.-3**. Other parameters that might determine the location of a point include changed skin consistency, sensitivity to pressure, puffiness or a resistant sensation when palpating gently with a finger.

Proportional measurements: When trying to locate points not situated at any prominent structures, Chinese Medicine applies the proportional cun measurement (→ 2.2).

Electric tools: These measure the electric resistance of the skin in order to find the correct location of the points. Generally, electric resistance is lower in the immediate area around the point. This method is especially used in auricular acupuncture, but it has not proved to be a practical method for body acupuncture.

Very point technique (after Gleditsch 2005): The needle is held loosely and guided in a ‘dancing’ fashion over the skin. When the correct point – the ‘very point’ – is found, the needle will ‘catch’ on to it and penetrate the skin as if out of its own volition.

2.2 Cun Measurements of the Body

In Chinese medicine, distances on the body are traditionally measured in cun. In contrast to the official Chinese cun (1 cun = 2.5 cm), the cun used in a medical context is a **proportional unit of measurement** that takes the individual proportions of each body into account.

This relative unit of measurement is defined either by using the width of a particular finger or fingers (finger cun) or by the distances between clearly defined landmarks of the body (body cun).

In clinical practice, many points can be located by using **finger cun** measurements (→ Fig. 2.1). It is important, however, to use the **patient’s fingers**, not the practitioner’s fingers, for reference. If the fingers of both the practitioner and the patient have approximately the same width, the width of the practitioner’s fingers can be applied for point location. Otherwise, measurements have to be adapted to correct minor deviations, for example by adjusting the spaces (narrow or wide) between the practitioner’s fingers when using finger cun measurements.

The **body or proportional cun** is based on the proportions of particular sections of the body, which are divided into a specific number of units (→ Fig. 2.3). In clinical practice, the measurement of these divisions can be quite cumbersome. Therefore, the spreading hands technique is commonly applied (Koenig and Wancura 1979/1983). It allows the practitioner to quickly determine the midpoint of a given distance (→ 2.3.3) and, combined with finger cun measurements and palpation, is in most cases sufficient for correctly locating a point. For point location on the abdomen, a prepared elastic tape can be useful for measuring proportional cun (→ 2.3.1).

Finger cun

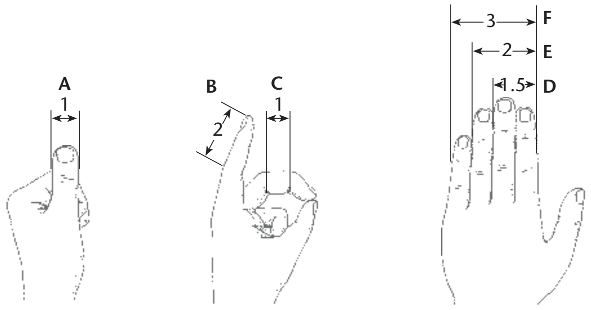
A
1
B
2
C
1
3
F
2
E
1.5
D

A:
1 cun:
thumb

B:
2 cun:
index finger

C:
1 cun:
middle finger

D:
1.5 cun:
finger-width

E:
2 cun:
finger-width

F:
3 cun:
finger-width

Fig 2.1

A	1 cun as thumb measure	The width of the interphalangeal joint of the thumb at its widest point
B	2 cun as index finger measure	The distance of the two distal phalanges of the index finger
C	1 cun as middle finger measure	With the tip of the middle finger placed on the thumb, the distance between the ends of the creases of the interphalangeal joints of the middle finger
D	1.5 cun as two-finger measure	The width of the index and middle finger when placed closely together, at the level of the distal interphalangeal joints
E	2 cun as index, middle and ring finger measure	The width of the index, middle and ring finger when placed closely together, at the level of the distal interphalangeal joints
F	3 cun as four finger measure	The width of the four fingers when placed closely together, at the level of the distal interphalangeal joints (=1 hand-breadth)

Body cun or proportional cun – the most commonly used proportional distances

Tip for locating the end of the axillary fold: The anterior or posterior axillary fold defines the end of the axilla, forming a fold in the skin (both anteriorly and posteriorly), where the arm joins the thorax. The axillary fold is the reference point for any cun measurements on the upper arm (→ Figs 2.2, 2.3, LU-3 and LU-4), which can be found by asking the patient to adduct his/her arm and to insert his/her flat hand into the axilla. The end of the axillary fold points in a superior direction, terminating at the lower border of the pectoralis major muscle, which in this area extends to its insertion in the humerus, below the deltoid muscle. The end of the axillary fold may vary considerably, depending on the individual proportions, positioning of the arm and flexing of the muscles. It should therefore be located with the patient's arms hanging down in a relaxed way.

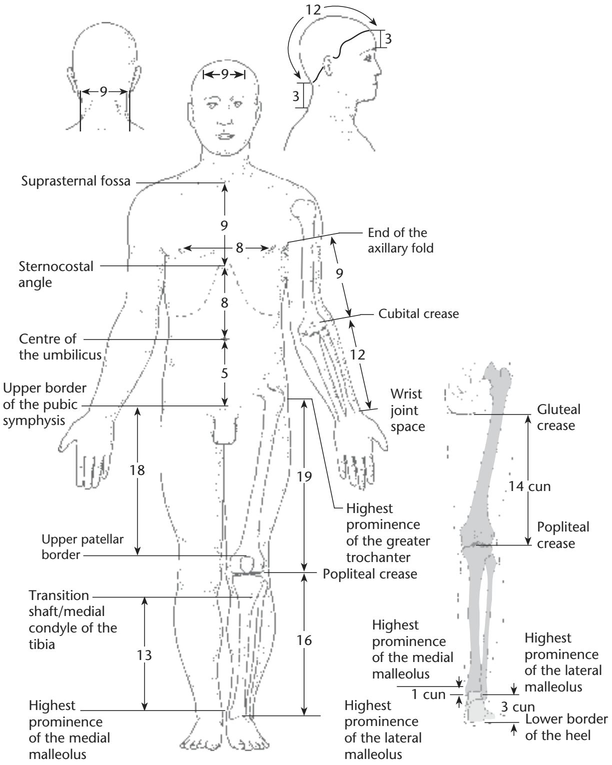
9
9
12
3
3
Suprasternal fossa
9
8
End of the axillary fold
Sternocostal angle
9
8
Cubital crease
Centre of the umbilicus
12
5
Upper border of the pubic symphysis
Wrist joint space
Gluteal crease
18
19
14 cun
Highest prominence of the greater trochanter
Popliteal crease
Upper patellar border
Popliteal crease
Transition shaft/medial condyle of the tibia
16
13
Highest prominence of the medial malleolus
Highest prominence of the lateral malleolus
1 cun
3 cun
Highest prominence of the medial malleolus
Highest prominence of the lateral malleolus
Lower border of the heel

Fig. 2.2

Head		
Longitudinal	12 cun	From the anterior (→ 3.1.1) to the posterior hairline (→ 3.1.5)
	3 cun	From the midpoint of the eyebrows to the anterior hairline (→ 3.1.1)
	3 cun	From the lower border of the spinous process of C7 (→ 3.4.1) to the posterior hairline (→ 3.1.5)
	18 cun	From **Ex-HN-3** (***yintang***) to **Du-14**
Horizontal	9 cun	The distance between the mastoid processes (→ 3.1.4)
	9 cun	From corner to corner on the forehead (**ST-8** to **ST-8**)
Thorax and abdomen		
Longitudinal	9 cun	From the suprasternal depression (**Ren-22**) to the sternocostal angle (→ 3.5, xiphosternal synchondrosis)
	8 cun	From the sternocostal angle (→ 3.5, xiphosternal synchondrosis) to the centre of the umbilicus
	5 cun	From the centre of the umbilicus to the upper border of the pubic symphysis (→ 3.5)
Horizontal	8 cun	The distance between the two nipples
Laterally longitudinal	12 cun	From the end of the axillary fold to the tip of the free end of the 11th rib (**LIV-13**)
Back and lumbar region		
Both the spinous processes and the intercostal spaces are used for orientation. Main reference points are: the spinous process of C7, the inferior angle of the scapula (level with T7), the iliac crest (level with the spinous process of L4), the posterior superior iliac spine (PSIS) (level with S2 or the 2nd sacral foramen). Please note: the position of the patient (standing, sitting, lying) as well as individual anatomy and spinal structure (kyphosis, lordosis etc.) all influence the level of the vertebrae, which consequently may vary considerably in relation to other anatomical structures. For more detail → Chapter 3.		
Longitudinal	30 cun	From the spinous process of the T1 to the tip of the coccyx
Horizontal	3 cun	From the medial border of the scapula to the posterior midline (spinous processes), with the shoulder relaxed and hanging arms
Upper extremity		
Lateral	9 cun	From the end of the anterior axillary fold to the transverse cubital crease
	12 cun	From the transverse cubital crease to the wrist joint space (→ 3.3.3, transverse wrist crease)
Lower extremity		
Lateral	19 cun	From the highest prominence of the greater trochanter (→ 3.6) to the popliteal crease
	16 cun	From the popliteal crease to the highest prominence of the lateral malleolus (→ 3.6.2)
	14 cun	From **G.B.-34** (anterior and inferior to the head of the fibula) to the highest prominence of the lateral malleolus (→ 3.6.2)
	3 cun	From the highest prominence of the lateral malleolus (→ 3.6.2) to the lower border of the heel
Medial	18 cun	From the upper border of the pubic symphysis (3.5) to the upper border of the patella (→ 3.6.1)
	15 cun	From the medial end of the popliteal crease to the highest prominence of the medial malleolus (→ 3.6.2)
	13 cun	From the medial condyle of the tibia (→ 3.6.1, junction of the head and the shaft of the tibia) to the highest prominence of the medial malleolus (→ 3.6.2)
Dorsal	14 cun	From the gluteal crease to the popliteal crease
Difference between the medial and lateral malleolus		
	1 cun	Difference between the level of the highest prominence of the lateral and medial malleolus respectively (→ 3.6.2)

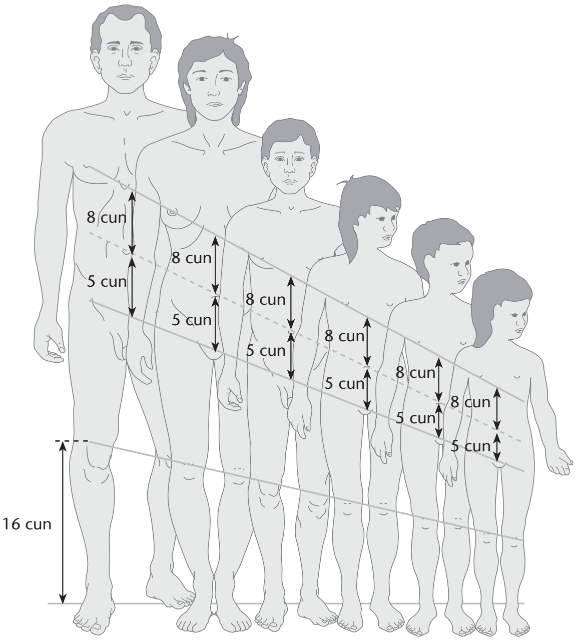
8 cun
5 cun
8 cun
5 cun
8 cun
5 cun
8 cun
5 cun
8 cun
5 cun
8 cun
5 cun
16 cun

Fig. 2.3

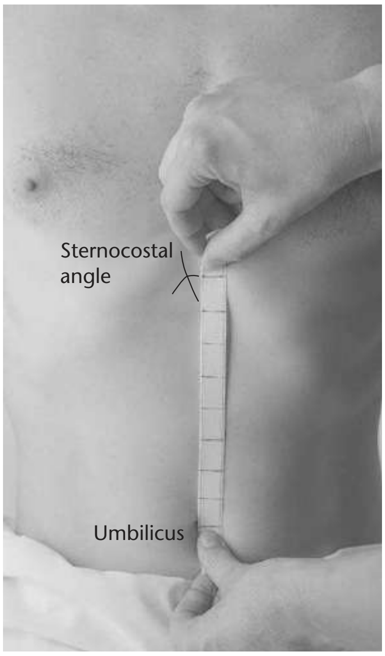
Sternocostal angle
Umbilicus

Fig. 2.4

2.3 Tips for Point Location

2.3.1 Elastic Tape

A marked elastic tape has proven a very useful tool for simplifying cun measurements and facilitating correct point location. It is especially recommended for measurements on the abdomen, the shape of which can vary considerably from patient to patient (Kitzinger 1995). An elastic tape, 1–2 cm wide and approximately 40 cm long, is marked at regular intervals of 2 cm, which can optionally be numbered from 1–20. The tape is placed onto the distance to be measured, stretching it according to the correct number of cun for that particular distance. This method allows the use of proportional cun for each individual patient.
Examples:

- For locating points on the upper abdomen, stretch the tape so that eight units cover the distance between the sternocostal angle and the centre of the umbilicus (→ Fig. 2.4).
- For locating points on the lower abdomen, stretch the tape so that five units cover the distance between the upper border of the pubic symphysis and the centre of the umbilicus (→ Fig. 2.5).

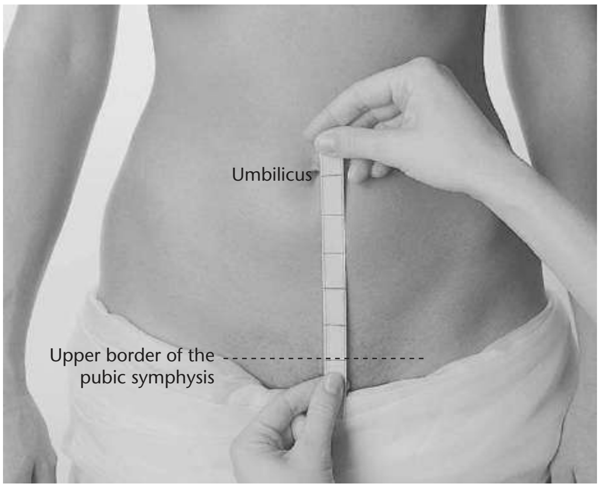
Umbilicus
Upper border of the pubic symphysis

Fig. 2.5

2.3.2 Patient Positioning and Body Movements

To locate certain points, it can be very helpful to ask the patient to assume a particular position or to perform a specific movement.

Examples:

- For locating points on the **P** channel on the forearm, ask the patient to tightly press the thumb and little finger together. This will make the tendons more visible (➞ Fig. 2.6)
- 'Tiger mouth grip' as location help for **LU-7** and **L.I.-6** (➞ Fig. 2.7)
- For locating **L.I.-4**, ask the patient to press his/her thumb and index finger together. A muscle bulge will form; L.I.-4 is located on its highest point (➞ Fig. 2.8).
- When turning the wrist joint from a pronated to a half-supinated position, the palpating finger will glide into a bony cleft just proximal to the styloid process of the ulna (**S.I.-6**, ➞ fig 2.9).
- As the palpating finger glides along the skin, the skin bulge forming in front of it will 'arrest' the finger at the point to be located. This method is very useful for quick location of certain points such as **T.B.-5** or **P-6** (➞ Fig. 2.10).

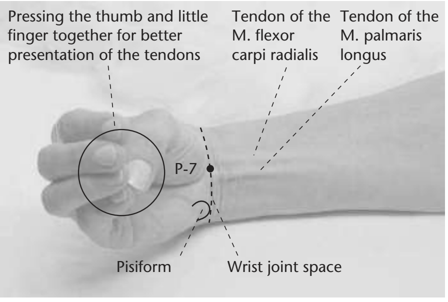
Pressing the thumb and little finger together for better presentation of the tendons
Tendon of the M. flexor carpi radialis
Tendon of the M. palmaris longus
P-7
Pisiform
Wrist joint space

Fig. 2.6

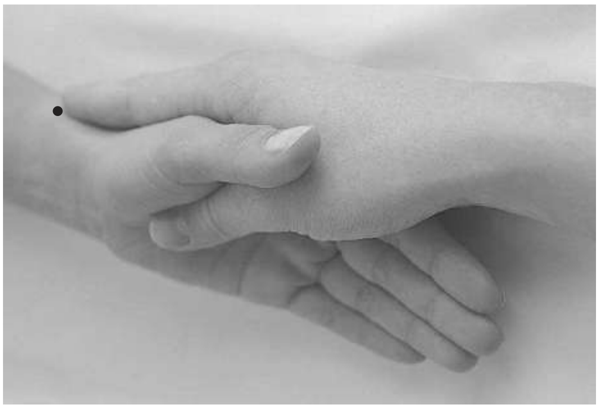

Fig. 2.7

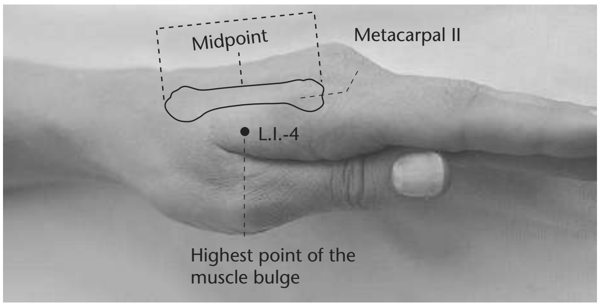
Midpoint
Metacarpal II
L.I.-4
Highest point of the muscle bulge

Fig. 2.8

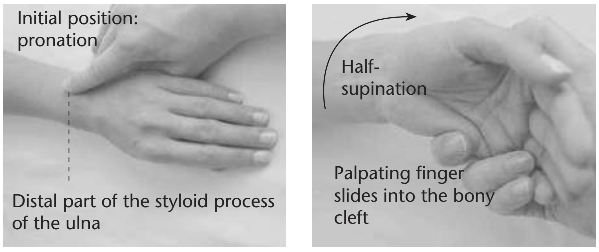
Initial position: pronation
Distal part of the styloid process of the ulna
Half-supination
Palpating finger slides into the bony cleft

Fig. 2.9

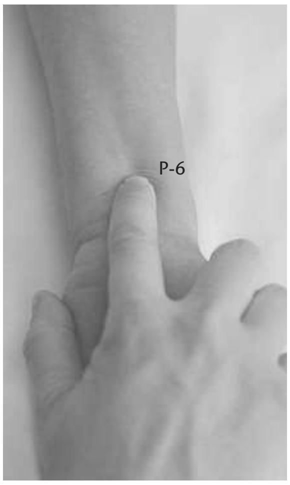

Fig. 2.10

2.3.3 Spreading Hands Technique According to König and Wancura

The spreading hands technique is particularly useful for quickly determining the midpoint of a given distance. For this purpose, the little fingers of each hand are placed on the endpoints of the distance to be measured. Both hands are then evenly spread over the distance. The thumbs are joined in the middle, indicating the midpoint of the distance.

Examples:

- To determine the midpoint of the 16 cun-distance between the popliteal crease and the prominence of the lateral malleolus, the little fingers are placed on both endpoints of this distance (→ Fig. 2.11)
- To locate **Du-20**, the hands are placed on the lateral aspects of the patient's head, with the little fingers touching the apices of the ears. The thumbs are then joined on the vertex of the head, and will be pointing to **Du-20** (→ Fig. 2.12).

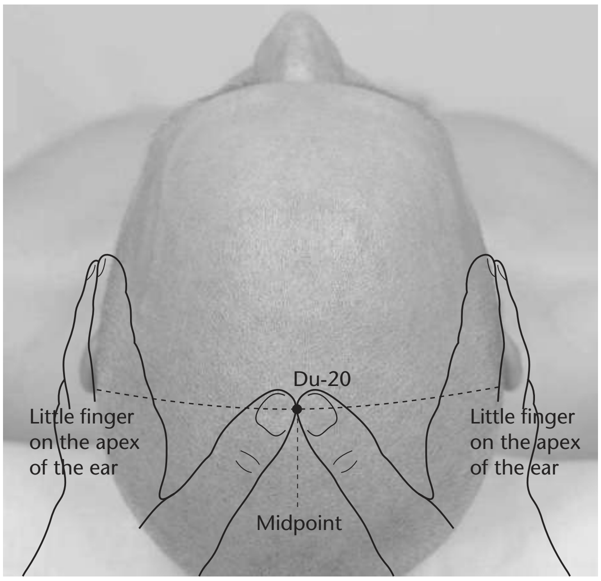
Du-20
Little finger on the apex of the ear
Little finger on the apex of the ear
Midpoint

Fig. 2.12

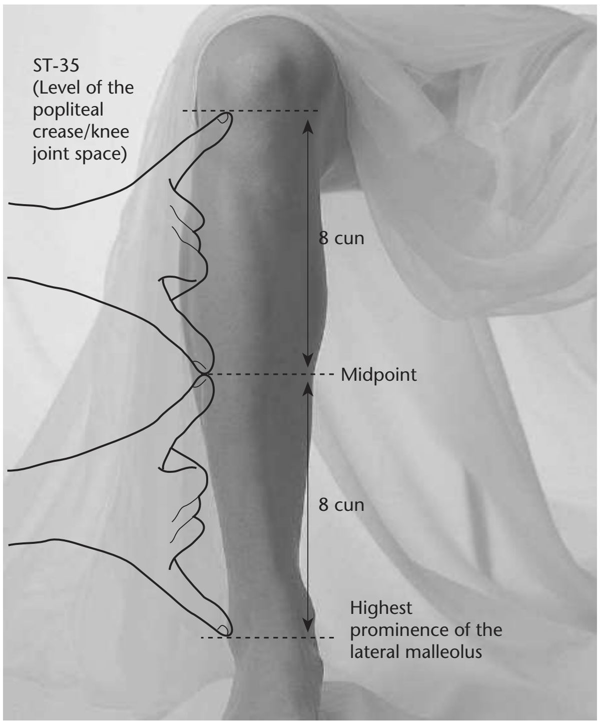
ST-35 (Level of the popliteal crease/knee joint space)
8 cun
Midpoint
8 cun
Highest prominence of the lateral malleolus

Fig. 2.11

3 Anatomical Orientation

Ulrich März

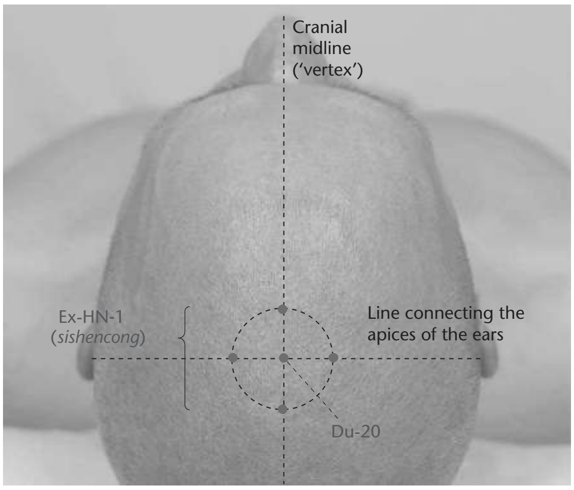
Cranial
midline
('vertex')
Ex-HN-1
(*sishencong*)
Line connecting the
apices of the ears
Du-20

Fig. 3.1

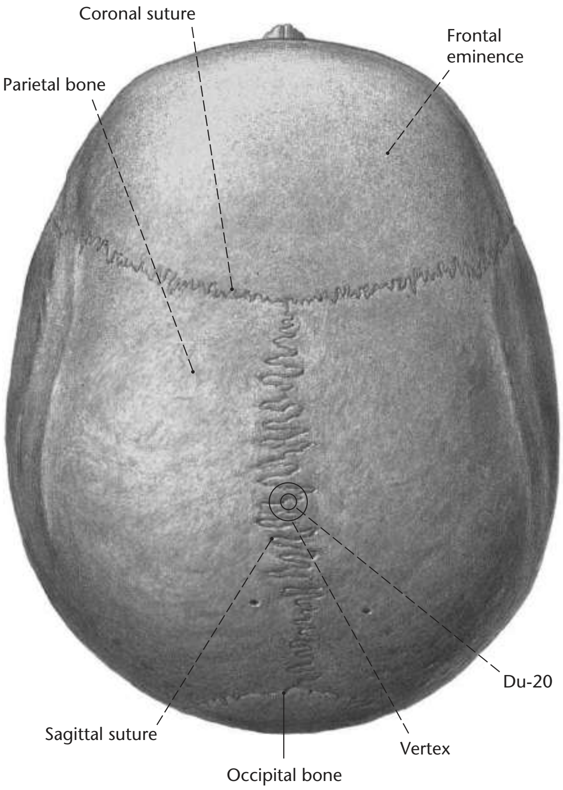
Coronal suture
Frontal
eminence
Parietal bone
Du-20
Sagittal suture
Vertex
Occipital bone

Fig. 3.2

3.1 Head

3.1.1 Vertex, Forehead and Supraorbital Region

Vertex (→ Figs 3.1, 3.2)

→ **Du-20** (*baihui*) is located at the junction of the cranial midline and an imaginary line connecting the apices of the ears. This point is at a distance of 5 cun from the anterior hairline and 7 cun from the posterior hairline, marking the highest point of the cranium ('vertex').

Please note: When locating **Du-20**, the patient's head should be in an upright position. Spreading hands technique (after König and Wancura, → Chapter 2): Place the hands on the lateral aspects of the patient's head, with the little fingers touching the apices of the ears. The thumbs are then joined at the cranial midline and will be pointing to → **Du-20** (*baihui*).

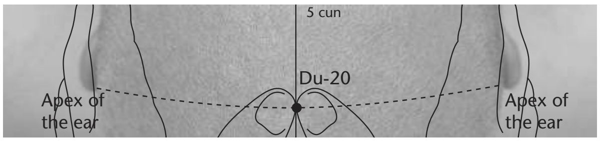
5 cun
Du-20
Apex of
the ear
Apex of
the ear

Fig. 3.3

Anterior hairline (→ Fig. 3.4)

The anterior hairline forms the superior border of the forehead and can vary considerably from person to person. The original line of hair growth is generally considered as the anterior hairline.
In men especially, the hairline may begin to recede early in life due to progressive baldness, so that the existing hairline may not be identical with the original line, which will now be located in an area not covered by hair. In these cases ask the patient to frown: the original hairline will be located superior to the wrinkled forehead, at the junction of the still slightly creased skin with the completely smooth skin. Points in this area are → **Du-20** to **Du-24**, **BL-3** to **BL-7, G.B.-4** to **G.B.-7, G.B.-13** to **G.B.-15, ST-8.**

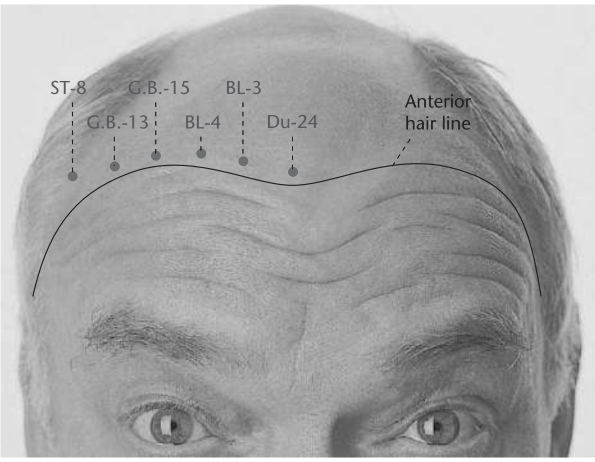
ST-8
G.B.-15
BL-3
G.B.-13
BL-4
Du-24
Anterior hair line

Fig. 3.4

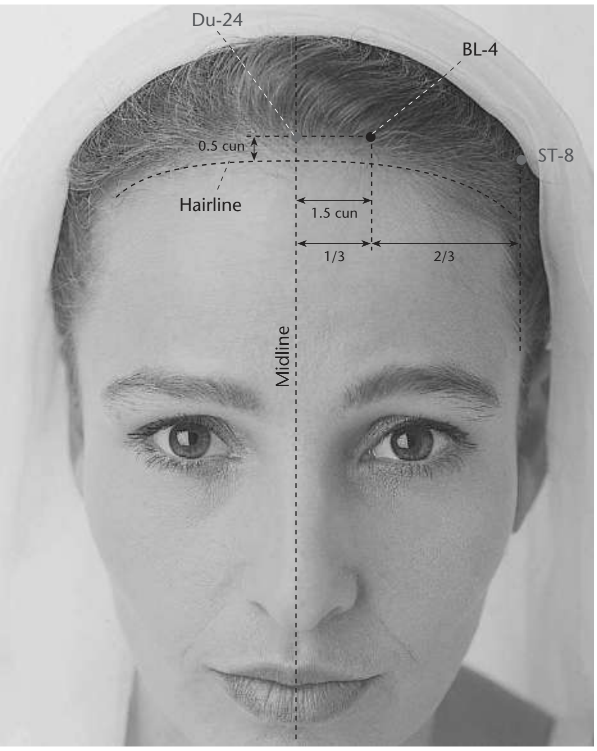
Du-24
BL-4
0.5 cun
ST-8
Hairline
1.5 cun
1/3
2/3
Midline

Fig. 3.5

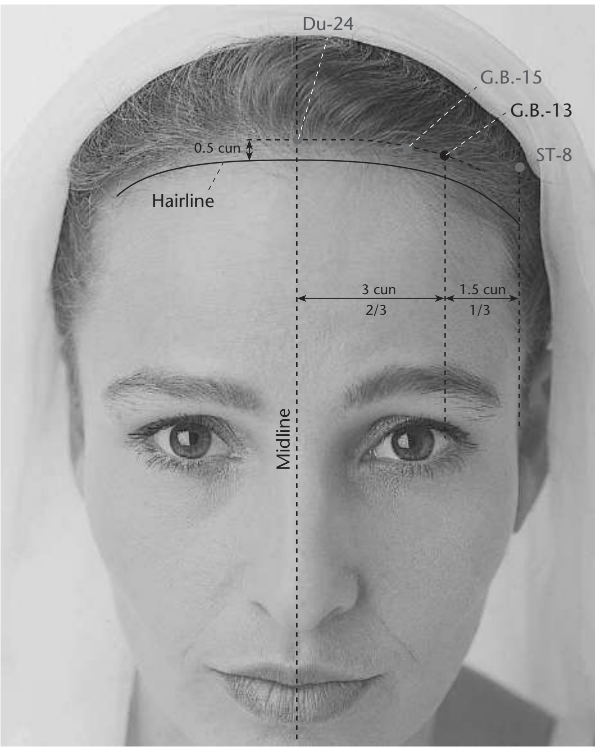
Du-24
G.B.-15
G.B.-13
0.5 cun
ST-8
Hairline
3 cun
2/3
1.5 cun
1/3
Midline

Fig. 3.6

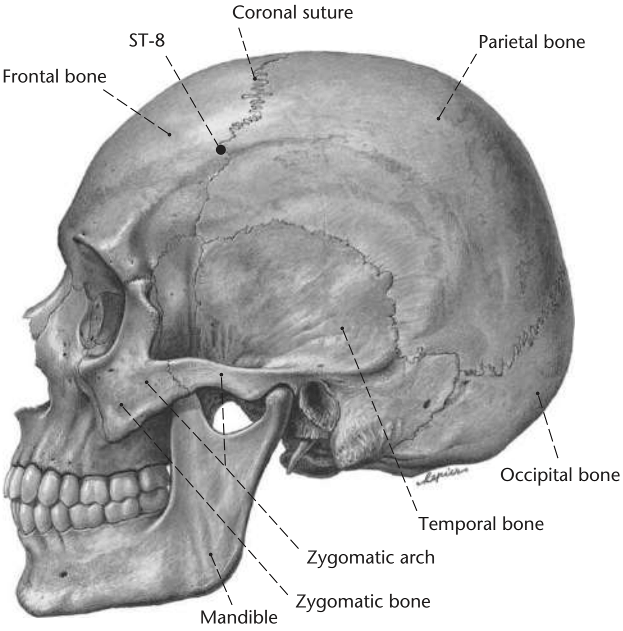
Coronal suture
ST-8
Parietal bone
Frontal bone
Occipital bone
Temporal bone
Zygomatic arch
Zygomatic bone
Mandible

Fig. 3.7

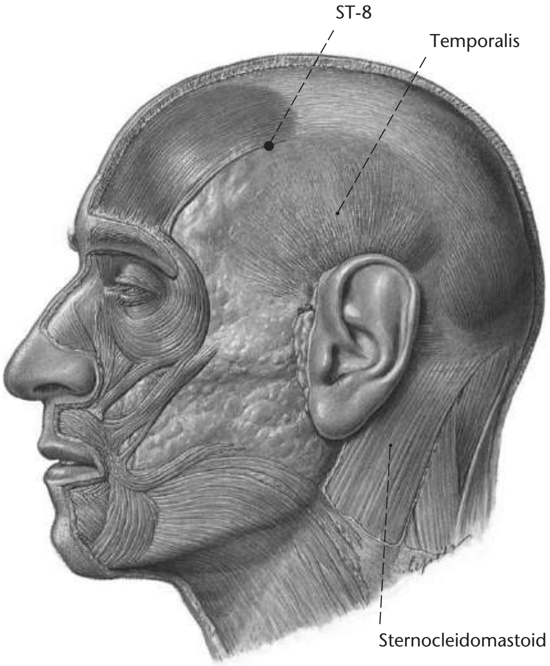
ST-8
Temporalis
Sternocleidomastoid

Fig. 3.8

Temporalis muscle and temporal corner of the forehead (→ Figs 3.7–3.9)

By asking the patient to firmly clench and release his/her teeth, the temporalis muscle can be palpated on the lateral aspect of the forehead. The corner between the forehead and the temples is marked by the junction of the anterior hairline with the temporal hairline, which tends to become more pronounced with a receding hairline.

→ **ST-8** is located in the centre of this area, at the insertion point of the temporalis muscle and 0.5 cun within the natural hairline. Often, the coronal suture can be palpated here.

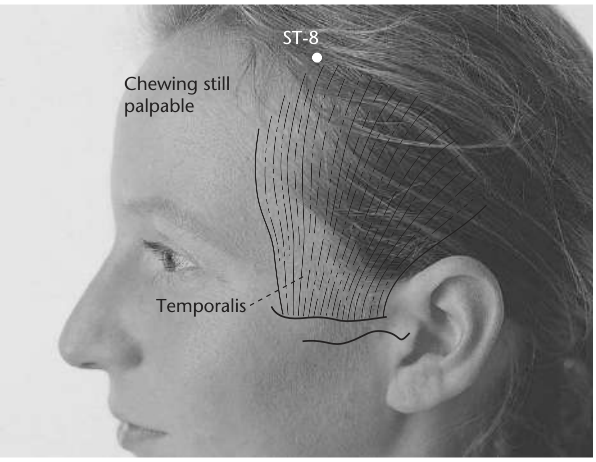
ST-8
Chewing still palpable
Temporalis

Fig. 3.9

Forehead (→ Figs 3.10, 3.11)

The more or less prominent bulge of the frontal eminences can be palpated on both sides of the forehead. → **G.B.-14** is located at their lower border, in a shallow depression at the junction of the eminences and the superciliary arches. With the patient looking straight ahead, G.B.-14 is located directly above the pupil, approximately 1 cun superior to the supraorbital margin.

Glabella, superciliary arches

The glabella is the central bony reference point on the forehead, forming a slightly convex, level area superior to the root of the nose and between the superciliary arches.

The extra point *yintang* (Ex-HN-3) is located at the centre of the glabella. Several points are located along the superciliary arches (from medial to lateral): → **BL-2, *yuyao*** (Ex-HN-4) and **T.B.-23.**

Please note: → **BL-2** is located superior to the inner canthus of the eye, in the area of the frontal (not the supraorbital) incisure, when this is present.

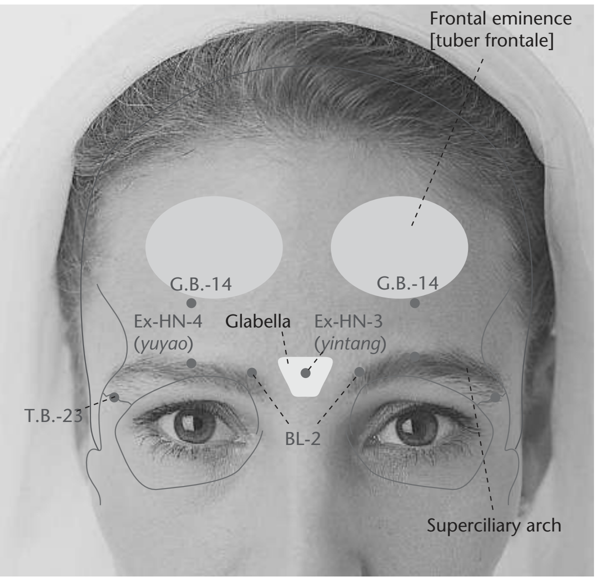
Frontal eminence
[tuber frontale]
G.B.-14
G.B.-14
Ex-HN-4
(yuyao)
Glabella
Ex-HN-3
(yintang)
T.B.-23
BL-2
Superciliary arch

Fig. 3.10

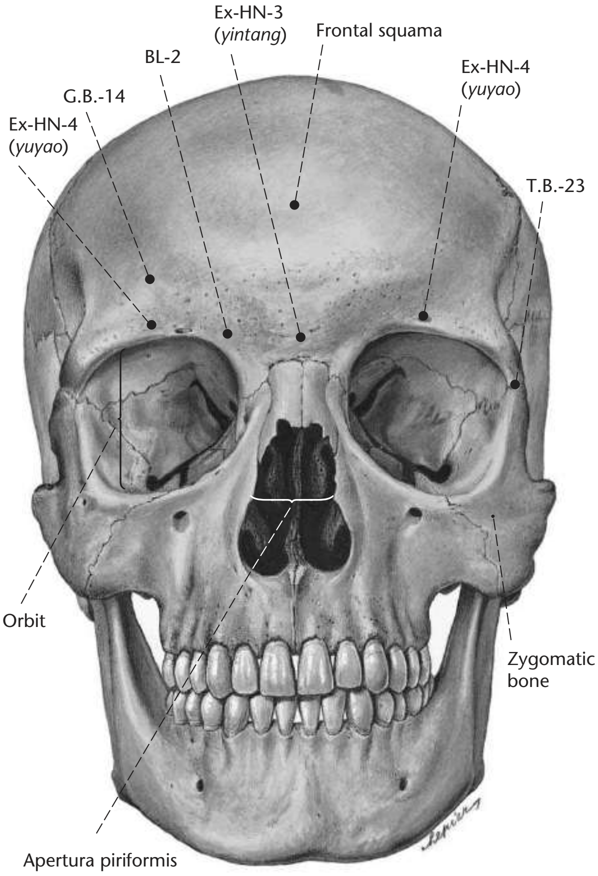
Ex-HN-3
(yintang)
Frontal squama
BL-2
Ex-HN-4
(yuyao)
G.B.-14
Ex-HN-4
(yuyao)
T.B.-23
Orbit
Zygomatic
bone
Apertura piriformis

Fig. 3.11

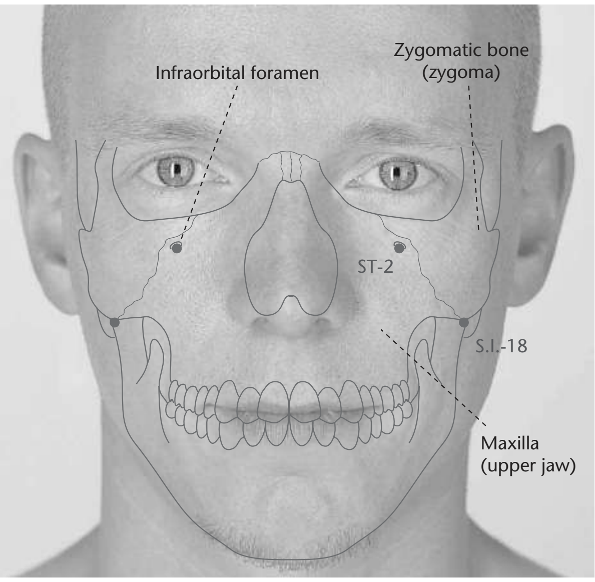
Infraorbital foramen
Zygomatic bone
(zygoma)
ST-2
S.I.-18
Maxilla
(upper jaw)

Fig. 3.12

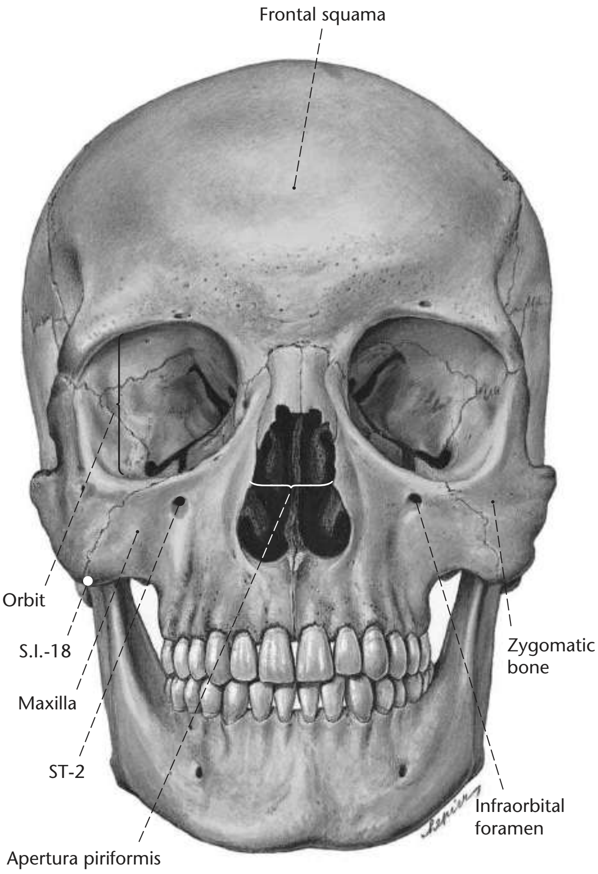
Frontal squama
Orbit
S.I.-18
Maxilla
ST-2
Apertura piriformis
Zygomatic
bone
Infraorbital
foramen

Fig. 3.13

3.1.2 Centre of the Face and Nasal Region

Cheek bones (→ Figs 3.12, 3.13)

Both the maxilla and the zygoma form the cheek bones below the eyes. Important anatomical landmarks are the infraorbital foramen, the lower border of the zygoma and the zygomatic arch of the temporal bone.

Infraorbital foramen (→ Figs 3.12, 3.13)

With the patient looking straight ahead, the infraorbital foramen with → **ST-2** is located on the maxilla, directly below the pupil. The foramen lies 1 finger-width below the infraorbital ridge. By palpating inferiorly from there, it can be felt below the maximal curvature of the ridge, in a small depression that is often sensitive to pressure.

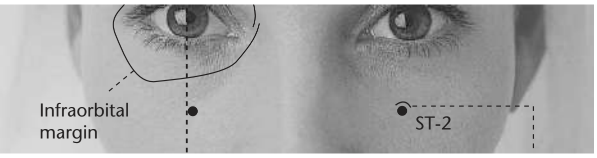
Infraorbital
margin
ST-2

Fig. 3.14

Zygoma (→ Figs 3.15, 3.16)

The bony structures below the lateral orbit are formed by the zygoma. → **S.I.-18** is located directly below the outer canthus of the eye at the lower border of the zygoma, at the junction of the zygoma and the maxilla, on the anterior border of the masseter muscle. (→ Lower face and chin).

Zygomatic arch (→ Figs 3.15, 3.16)

On the lateral aspect of the face, the temporal process of the zygoma articulates with the zygomatic process of the temporal bone to form the zygomatic arch. This is an approximately horizontal bony structure which can be palpated laterally to the ear. → **G.B.-3** and → **T.B.-22** are both located on the upper border of the zygomatic arch, while → **ST-7** and → **T.B.-21** are situated at its lower border.

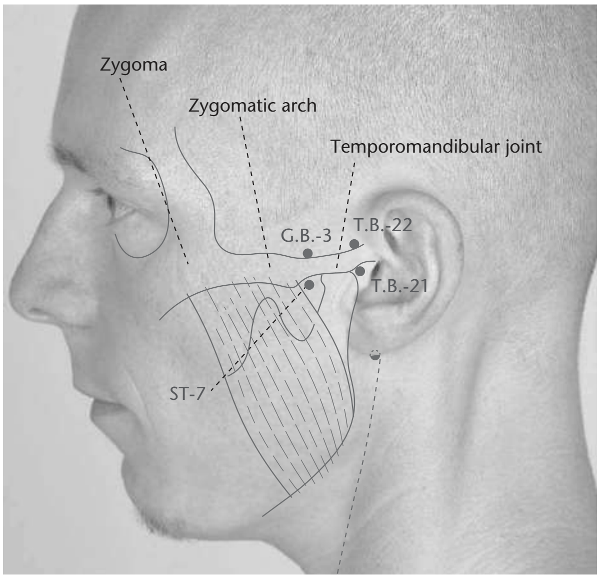
Zygoma
Zygomatic arch
Temporomandibular joint
G.B.-3
T.B.-22
T.B.-21
ST-7

Fig. 3.15

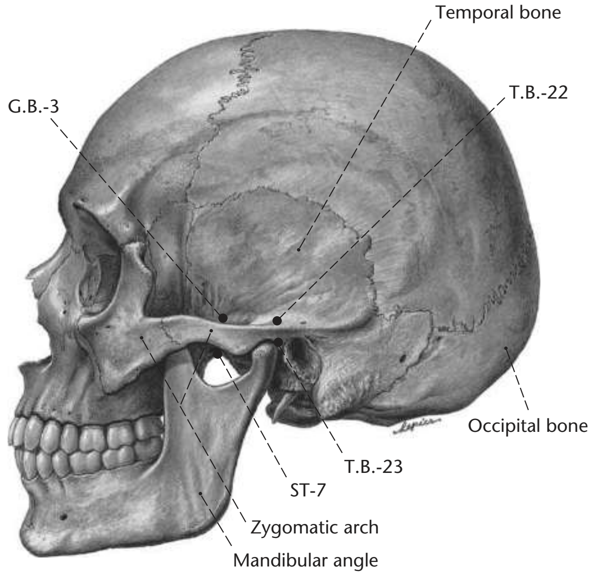
Temporal bone
T.B.-22
G.B.-3
Occipital bone
T.B.-23
ST-7
Zygomatic arch
Mandibular angle

Fig. 3.16

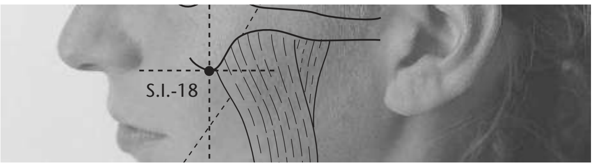

Fig. 3.17

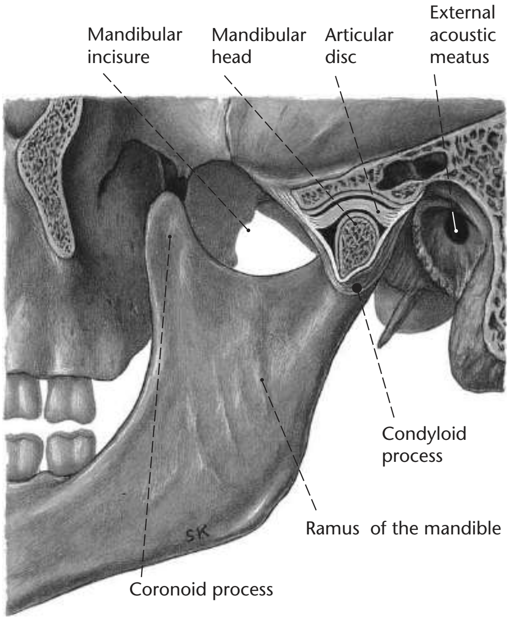
Mandibular incisure
Mandibular head
Articular disc
External acoustic meatus
Condyloid process
Ramus of the mandible
Coronoid process
SK

Fig. 3.18

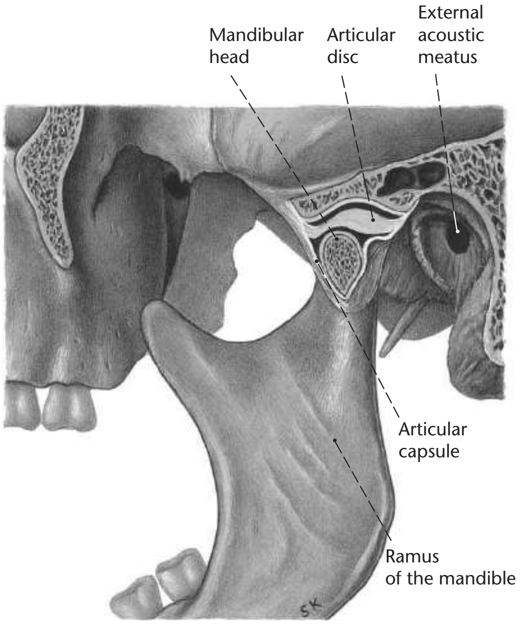
Mandibular head
Articular disc
External acoustic meatus
Articular capsule
Ramus of the mandible
SK

Fig. 3.19

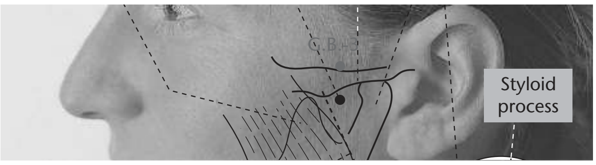
G.B.-3
Styloid process

Fig. 3.20

Temporomandibular joint (→ Figs 3.15, 3.18, 3.19)

The temporomandibular joint is located inferior to the zygomatic arch and anterior to the tragus of the ear. Its movement can easily be palpated by gently opening and closing the mouth. → **ST-7** is located in a depression anterior to the joint and directly inferior to the zygomatic arch. This depression is situated in the mandibular incisure, which is formed by the coronoid and condylar processes of the mandible. By opening the mouth, the palpating finger resting on → **ST-7** is 'lifted' out of the depression by the condyloid process sliding anteriorly.

Nasolabial groove (→ Fig. 3.21)

The nasolabial groove begins lateral to the nose and superior to the ala nasi, at the junction of the cartilage and the bone. From there it runs to the corners of the mouth. → **Ex-HN-8** (*shang-yingxiang, bitong*) is located at its upper end, at the junction of the upper jaw and the nasal cavities. → **L.I.-20** is located slightly below *bitong*, between the lateral border of the side of the nose and the nasolabial groove.

3.1.3 Lower Face and Chin

Mandibular angle (→ Fig. 3.22)

The mandibular angle corresponds to the area where the lower jaw line changes its direction from horizontal to vertical. It is located markedly inferior and anterior to the ear lobe. This angle marks the location of → **ST-6** (1 finger-width anterior and superior to the mandibular angle) as well as of → **S.I.-17** and → **T.B.-16** (level with the mandibular angle).

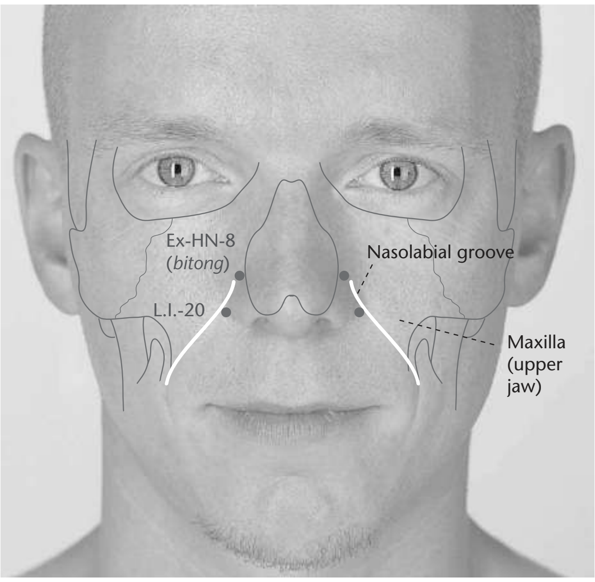
Ex-HN-8
(*bitong*)
Nasolabial groove
L.I.-20
Maxilla
(upper
jaw)

Fig. 3.21

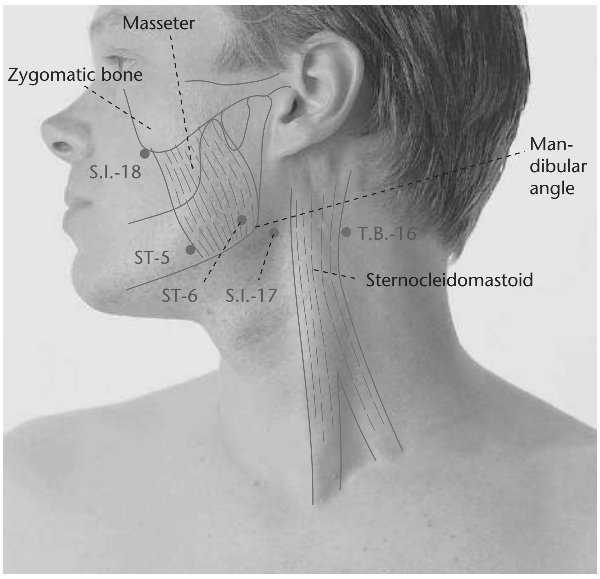
Masseter
Zygomatic bone
Man-
dibular
angle
S.I.-18
T.B.-16
ST-5
Sternocleidomastoid
ST-6
S.I.-17

Fig. 3.22

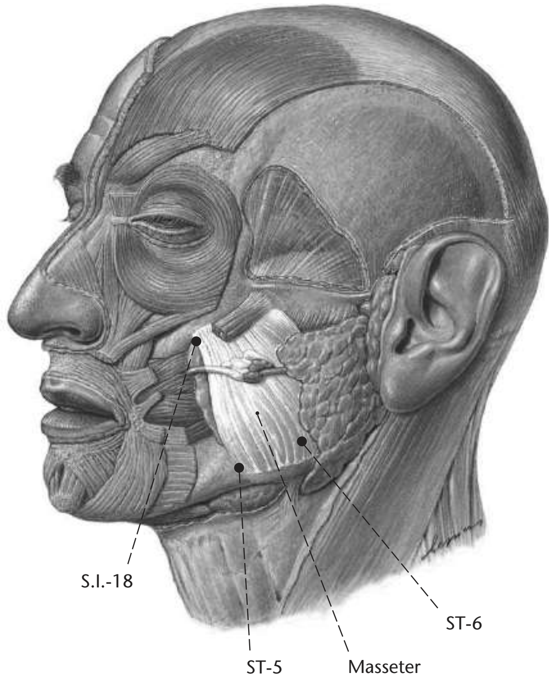
S.I.-18
ST-6
ST-5
Masseter

Fig. 3.23

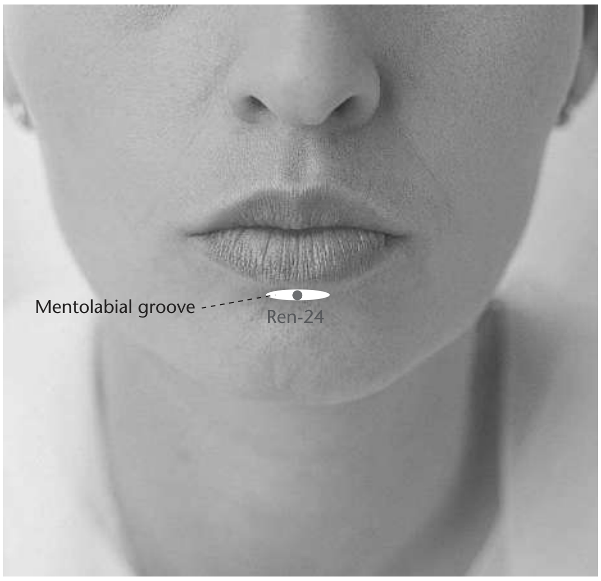
Mentolabial groove
Ren-24

Fig. 3.24

Masseter muscle (→ Fig. 3.23)

By firmly clenching the jaw, the contours of the masseter muscle become well defined, stretching from the lateral aspect of the lower jaw to the zygomatic bone. It originates at the zygomatic arch (→ Fig. 3.15) and inserts at the mandibular angle (→ Fig. 3.22).

Mentolabial groove (→ Fig. 3.24)

This forms a transverse groove at the junction of the chin and the lower lip. → **Ren-24** is located in its centre.

3.1.4 Ear Region

Note: points used in auricular acupuncture are not discussed here.

Helix root (→ Figs 3.25, 3.26)

The helix root borders the upper part of the ear. This is the location of → **T.B.-22**.

Supratragic notch (→ Figs 3.25, 3.26)

The supratragic notch forms an incisure within the cartilage of the ear, separating the helix root from the tragus. → **T.B.-21** is located anterior to it.

Tragus (→ Figs 3.25, 3.26)

→ **S.I.-19** is located anterior to the midpoint of the tragus.

Intertragic notch (→ Figs 3.25, 3.26)

The intertragic notch forms an incisure in the auricular cartilage, separating the tragus from the lobe and antitragus. → **G.B.-2** is located anterior to this notch.

Ear lobe (→ Figs 3.25, 3.26)

Inferior to the earlobe and the ear canal is a depression, which is bordered posteriorly by the mastoid process and anteriorly by the lower jaw. The lower border is formed by the transverse process of the 1st cervical vertebra (atlas). → **T.B.-17** is located in this depression.

Border between the auricle and the face (→ Fig. 3.26)

The border between the auricle and the temple/cheek generally forms a more or less vertical line, which will become more clearly defined by bending the auricle and tragus towards the anterior. Several points are located along this line (from top to bottom): → **T.B.-22, T.B.-21, S.I.-19, G.B.-2.**

Mastoid process (→ Figs 3.26, 3.27)

The mastoid process is a cone-shaped bony structure which can be palpated posterior to the ear. → **G.B.-12** is located at its tip, while → ***anmian*** (Ex-HN) can be found at the border of the mastoid process and the occiput. → ***yiming*** (Ex-HN-14) is located somewhat more inferiorly.

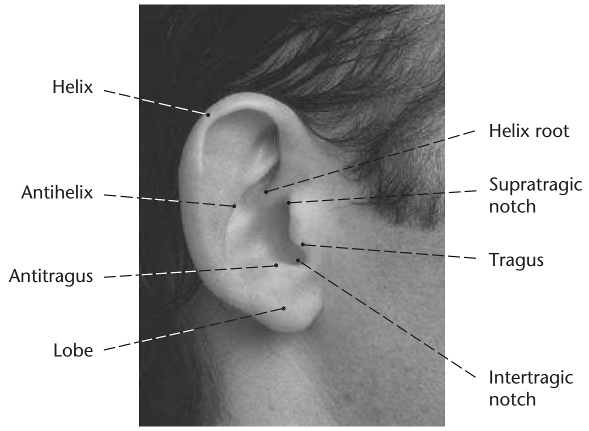
Helix
Antihelix
Antitragus
Lobe
Helix root
Supratragic notch
Tragus
Intertragic notch

Fig. 3.25

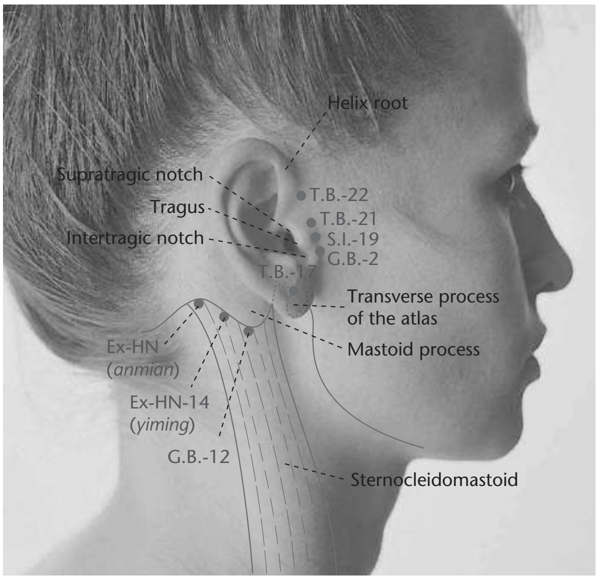
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Transverse process of the atlas
Mastoid process
Ex-HN (anmian)
Ex-HN-14 (yiming)
G.B.-12
Sternocleidomastoid

Fig. 3.26

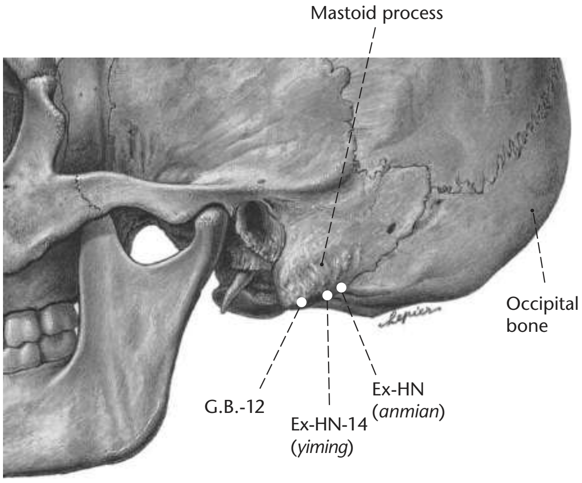
Mastoid process
Occipital bone
G.B.-12
Ex-HN-14 (yiming)
Ex-HN (anmian)

Fig. 3.27

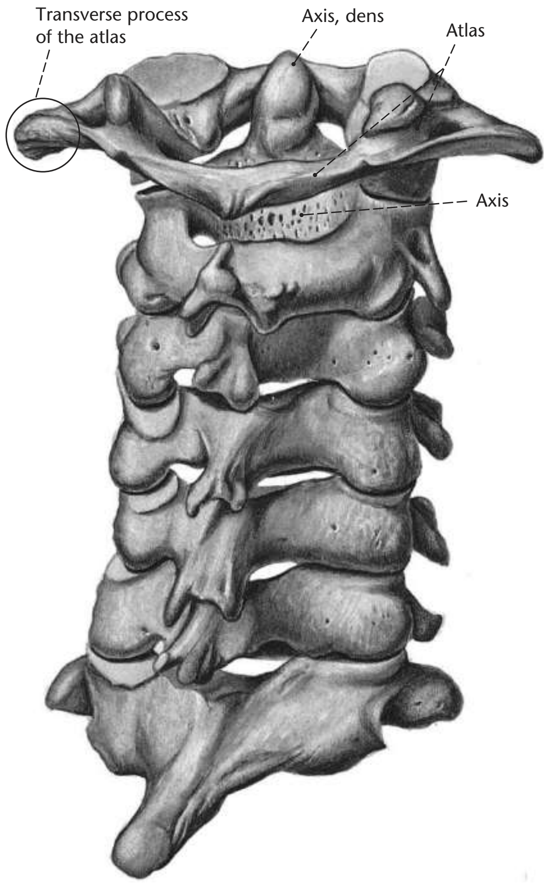
Transverse process of the atlas
Axis, dens
Atlas
Axis

Fig. 3.28

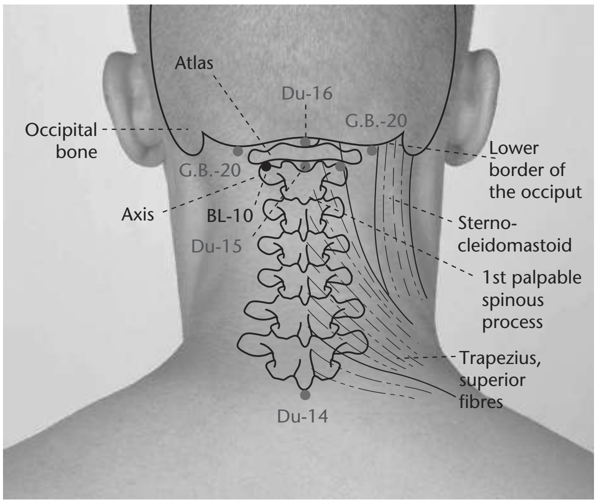
Atlas
Du-16
G.B.-20
Occipital bone
G.B.-20
Lower border of the occiput
Axis
BL-10
Sterno-cleidomastoid
Du-15
1st palpable spinous process
Trapezius, superior fibres
Du-14

Fig. 3.29

Transverse process of the 1st cervical vertebra (atlas) (→ Fig. 3.28)

The transverse process of the atlas can be palpated inferior to the earlobe as a deep bony structure, which in most cases is very sensitive to pressure.

3.1.5 Occiput

Craniocervical transition (→ Figs 3.28–3.32)

The transition from the head to the neck is formed by the mastoid process, the adjacent dorsal musculature and the occiput. The following points are located in this area (from lateral to medial) → **G.B.-12,** ***yiming*** **(Ex-HN-14),** ***anmian*** **(Ex-HN), G.B.-20, BL-10, Du-15** and **Du-16.**

External occipital protuberance (→ Figs 3.32, 3.33)

The external occipital protuberance forms a flat projection on the posterior midline of the occiput, which can be palpated slightly superior to the craniocervical transition. It defines the location of → **Du-16, Du-17** and **BL-9.** Sometimes, and more frequently in women, the protuberance may be defined only poorly or not at all.

Posterior hairline (no figure)

The posterior hairline is used as a reference point for locating points on the occiput, but due to its variable position it is not a very reliable landmark.

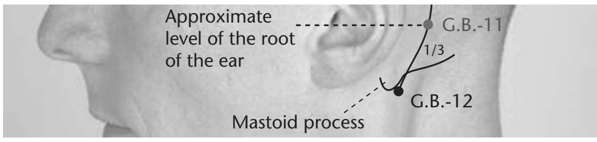
Approximate level of the root of the ear
G.B.-11
1/3
G.B.-12
Mastoid process

Fig. 3.30

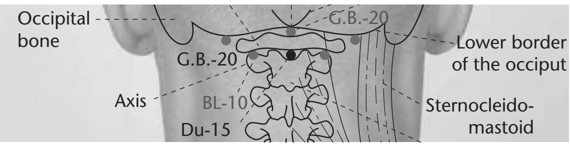
Occipital bone
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid
Du-15

Fig. 3.31

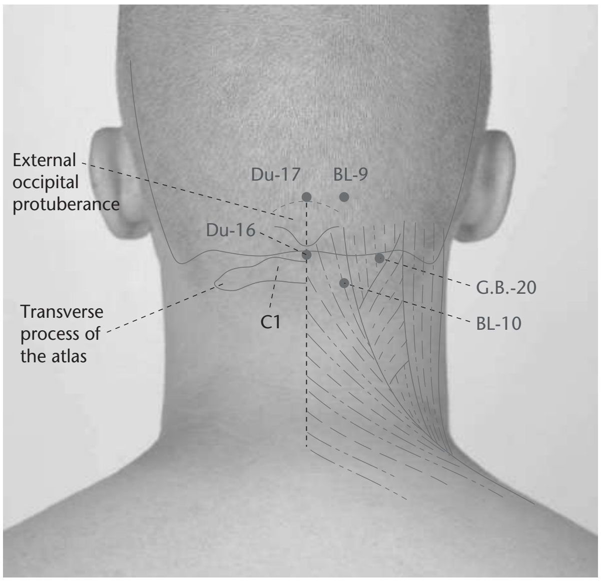
External occipital protuberance
Du-17
BL-9
Du-16
G.B.-20
BL-10
Transverse process of the atlas
C1

Fig. 3.32

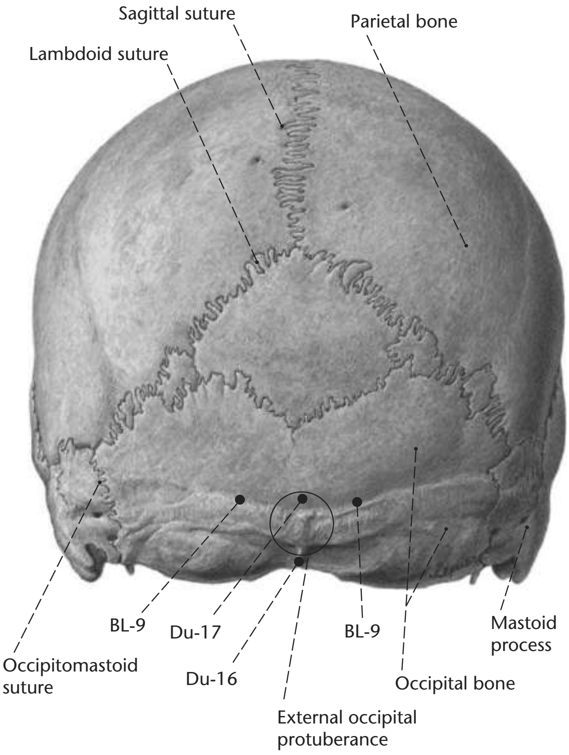
Sagittal suture
Parietal bone
Lambdoid suture
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

Fig. 3.33

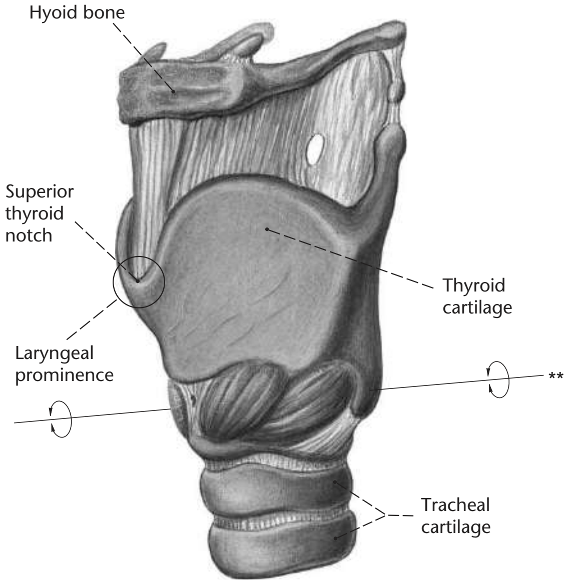
Hyoid bone
Superior thyroid notch
Laryngeal prominence
Thyroid cartilage
**
Tracheal cartilage

Fig. 3.34

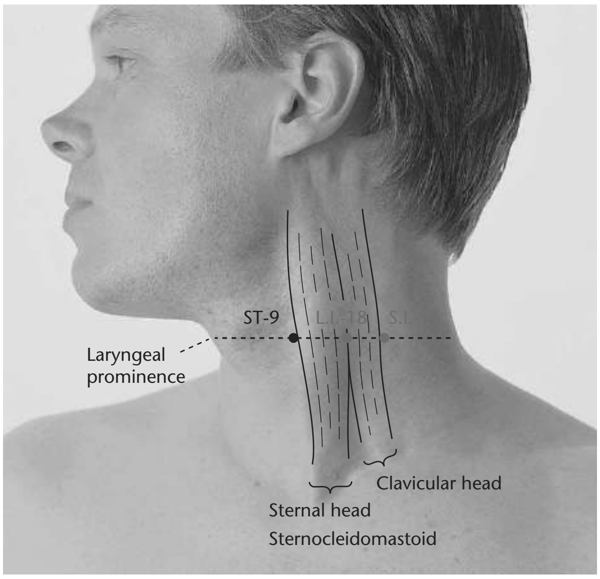
ST-9
L.I.-18
S.I.-16
Laryngeal prominence
Clavicular head
Sternal head
Sternocleidomastoid

Fig. 3.35

3.2 Neck

Laryngeal prominence (→ Figs 3.34, 3.35)

In men, the laryngeal prominence as the highest point of the larynx ('Adam's apple') tends to form a clearly palpable and visible structure. In women, visual identification is more difficult, so that palpation is necessary in order to locate the V-shaped incisure on the upper border of the thyroid cartilage on the laryngeal midline. Located at the level of the laryngeal prominence are → **ST-9, L.I.-18** and **S.I.-16**.

Sternocleidomastoid (SCM) muscle (→ Figs 3.36, 3.37)

This muscle forms a distinct, easily identifiable structure on the anterior and lateral aspect of the neck, which becomes visible and palpable by rotating the neck in the opposite direction. This effect can be increased by rotating the head against resistance (by pushing against the chin for example). The more slender clavicular head originates from the anterior aspect of the manubrium, while the more flat clavicular head originates from the medial third of the clavicle, revealing a small triangle between the two heads near the sternoclavicular joint. This is the location of → **ST-11**. By palpating along the SCM muscle in a superior direction, the two heads can be distinguished for the major part of the muscle, before they merge to insert at the posterior part of the mastoid process and the lateral superior nuchal line of the occiput. Besides → **ST-11,** the following points are located alongside the muscle belly of the SCM: → **L.I.-17, L.I.-18, S.I.-16, ST-9, G.B-12** and ***anmian*** **(Ex-HN)**.

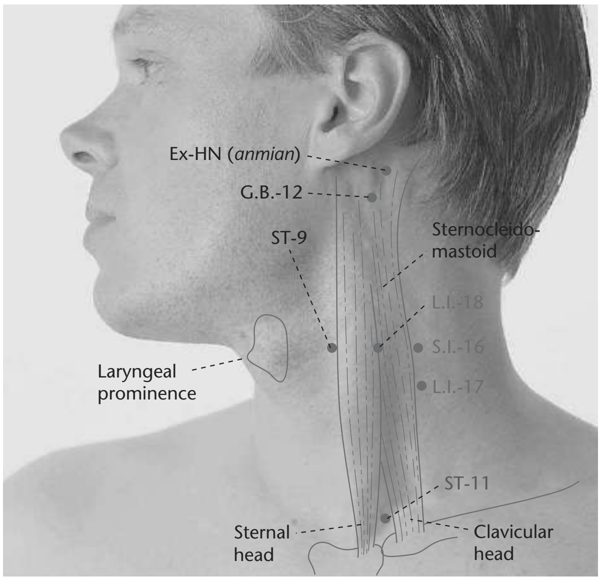
Ex-HN (*anmian*)
G.B.-12
ST-9
Sternocleido-mastoid
L.I.-18
S.I.-16
L.I.-17
Laryngeal prominence
ST-11
Sternal head
Clavicular head

Fig. 3.36

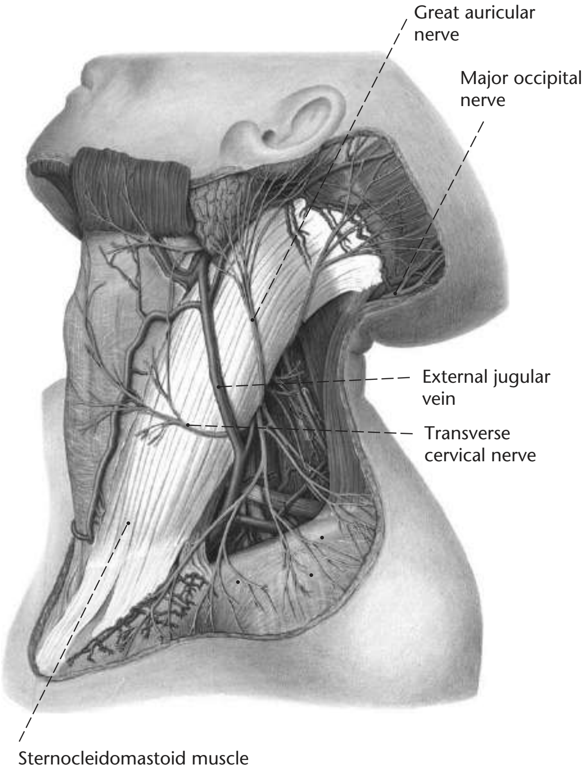
Great auricular nerve
Major occipital nerve
External jugular vein
Transverse cervical nerve
Sternocleidomastoid muscle

Fig. 3.37

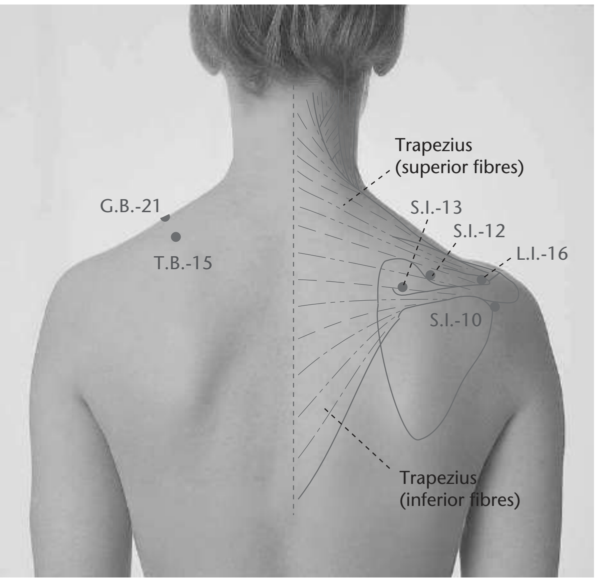
Trapezius
(superior fibres)
G.B.-21
S.I.-13
S.I.-12
L.I.-16
T.B.-15
S.I.-10
Trapezius
(inferior fibres)

Fig. 3.38

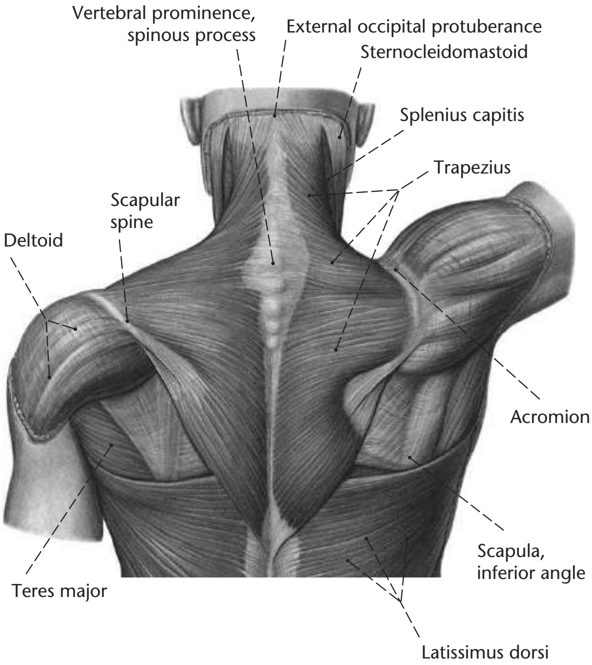
Vertebral prominence,
spinous process
External occipital protuberance
Sternocleidomastoid
Splenius capitis
Trapezius
Scapular
spine
Deltoid
Acromion
Scapula,
inferior angle
Teres major
Latissimus dorsi

Fig. 3.39

3.3. Shoulder and Arm

3.3.1 Shoulder and Upper Arm

Trapezius muscle (→ Figs 3.38, 3.39)

This muscle covers the superior aspect of the shoulder, its superior portions extending from the cervical spine to the acromion at the lateral end of the scapula. Located at the midpoint of the superior border of the muscle are → **G.B.-21** and, slightly inferiorly, → **T.B.-15.**

Acromion (→ Figs 3.39, 3.42, 3.43)

When palpating the shoulder joint from a lateral direction, the acromion can be felt as a bony extension with a flat surface, slightly superior to the laterally protruding head of the humerus.

Scapular spine (→ Figs 3.39–3.44)

The scapular spine forms a bony ridge that originates at the acromion and runs in an oblique angle across the scapula, ending at its medial border in an upward opening curve. This is the location of → **S.I.-13**, → **L.I.-16** is located on the most lateral portion of the trapezius and supraspinatus muscles, in the angle between the acromial origin of the scapular spine and the acromioclavicular joint.

Head of the humerus, shoulder hollows (→ Figs 3.42, 3.44)

The head of the humerus is located inferior to the acromion and protrudes slightly on the lateral aspect of the shoulder. When abducting the arm horizontally, two hollows will form at the transition from the arm to the shoulder. The anterior hollow is the location of → **L.I.-15**, while → **T.B.-14** is located in the posterior one. The hollows mark the border between the clavicular and acromial portion and the border between the acromial and spinal portion of the deltoid muscle.

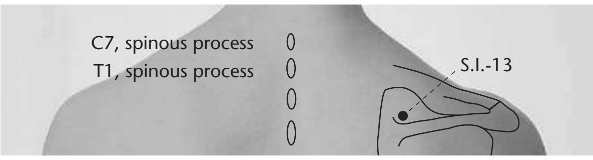
C7, spinous process
T1, spinous process
S.I.-13

Fig. 3.40

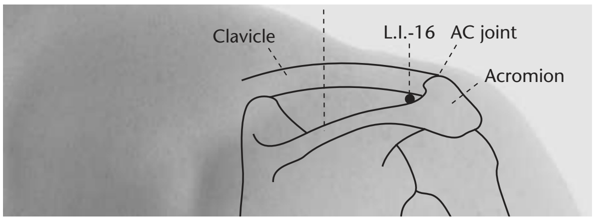
Clavicle
L.I.-16
AC joint
Acromion

Fig. 3.41

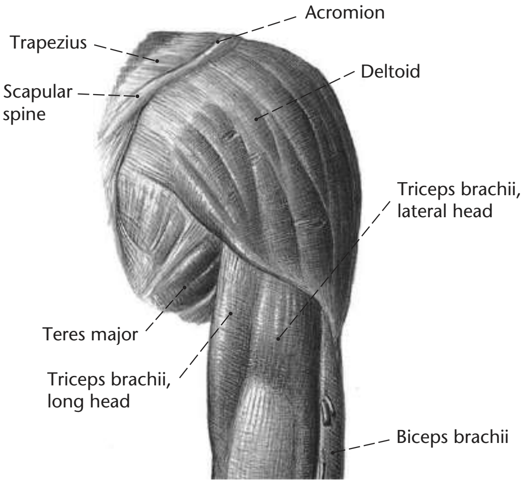
Acromion
Trapezius
Deltoid
Scapular spine
Triceps brachii, lateral head
Teres major
Triceps brachii, long head
Biceps brachii

Fig. 3.42

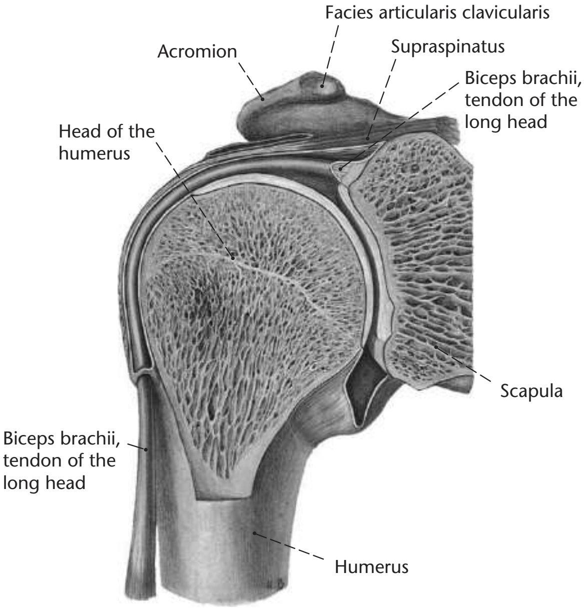
Facies articularis clavicularis
Acromion
Supraspinatus
Biceps brachii, tendon of the long head
Head of the humerus
Scapula
Biceps brachii, tendon of the long head
Humerus

Fig. 3.43

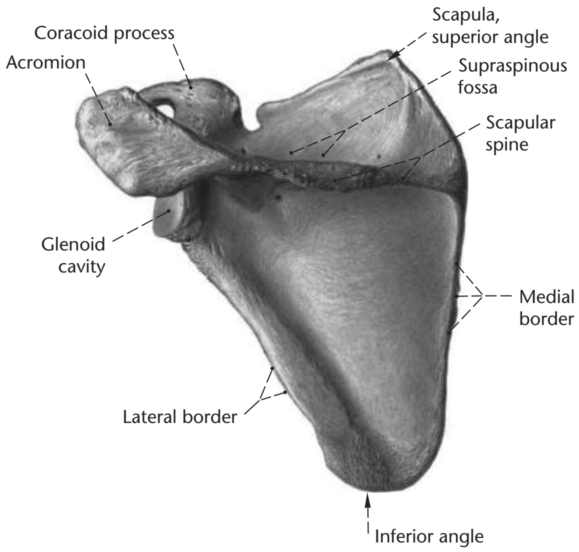
Coracoid process
Acromion
Scapula, superior angle
Supraspinous fossa
Scapular spine
Glenoid cavity
Medial border
Lateral border
Inferior angle

Fig. 3.44

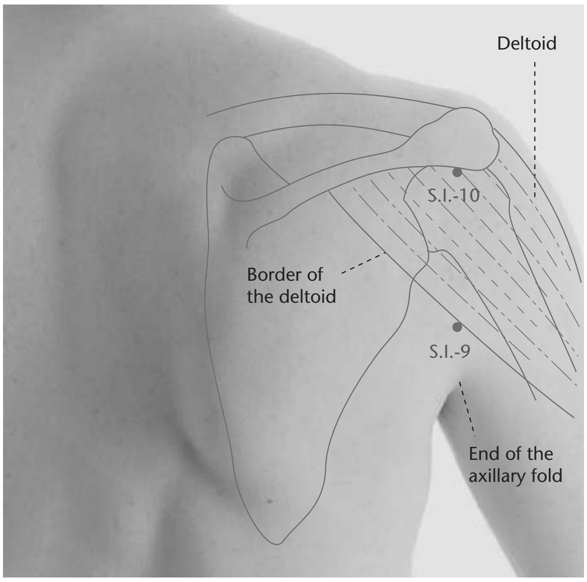
Deltoid
S.I.-10
Border of the deltoid
S.I.-9
End of the axillary fold

Fig. 3.45

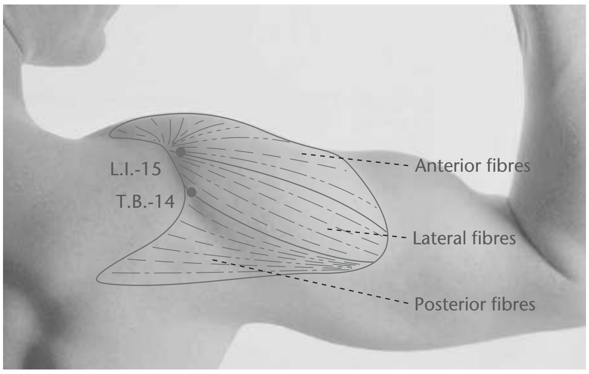
L.I.-15
T.B.-14
Anterior fibres
Lateral fibres
Posterior fibres

Fig. 3.46

Deltoid muscle (→ Figs 3.42, 3.45–3.51)

This muscle covers the anterior, posterior and lateral aspect of the head of the humerus. The following points are located either on the muscle or on its margins: → **S.I.-9, S.I.-10, T.B.-14, L.I.-15,** ***jianquan*** **(Ex-UE), LU-1, LU-2, LU-3, L.I.-14**.

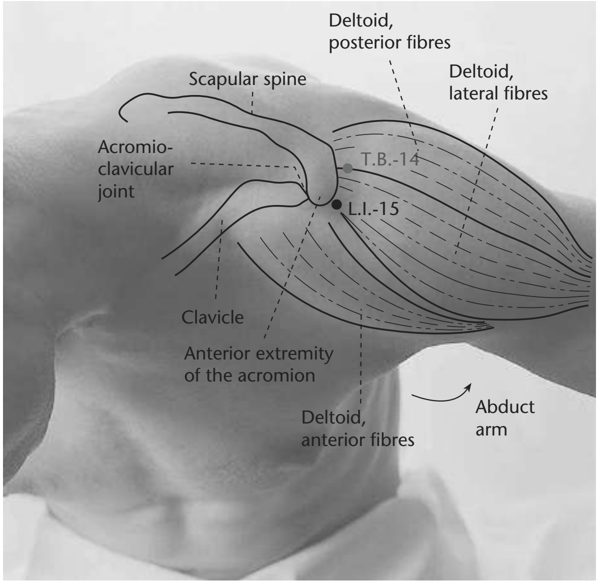
Deltoid, posterior fibres
Scapular spine
Deltoid, lateral fibres
Acromio-clavicular joint
T.B.-14
L.I.-15
Clavicle
Anterior extremity of the acromion
Deltoid, anterior fibres
Abduct arm

Fig. 3.47

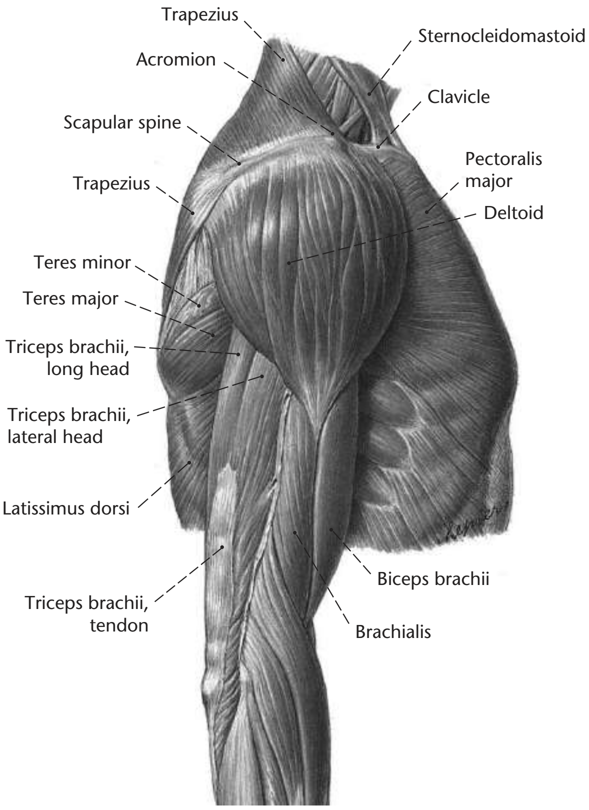
Trapezius
Acromion
Scapular spine
Trapezius
Teres minor
Teres major
Triceps brachii, long head
Triceps brachii, lateral head
Latissimus dorsi
Triceps brachii, tendon
Sternocleidomastoid
Clavicle
Pectoralis major
Deltoid
Biceps brachii
Brachialis

Fig. 3.48

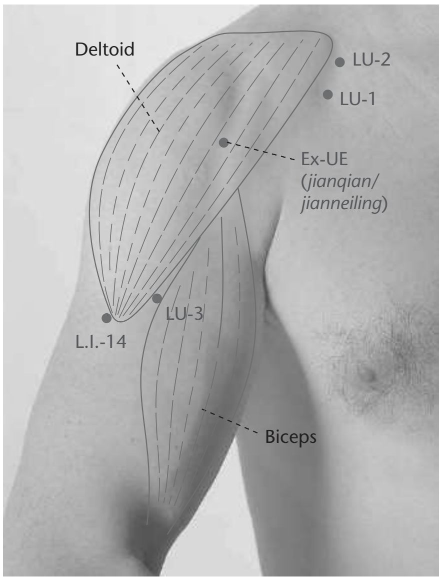
Deltoid
LU-2
LU-1
Ex-UE (jianqian/ jianneiling)
LU-3
L.I.-14
Biceps

Fig. 3.50

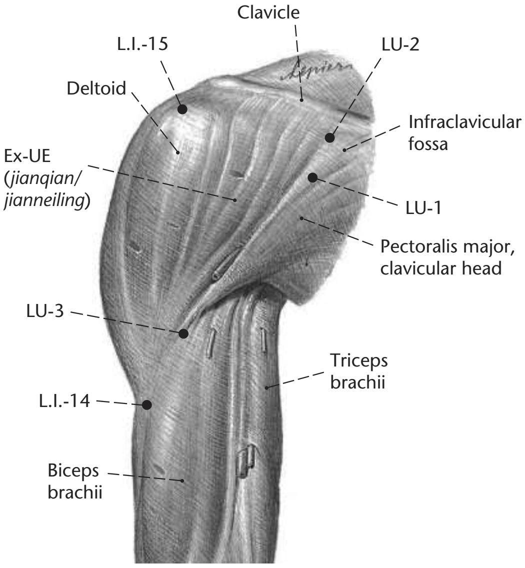
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
Ex-UE (jianqian/ jianneiling)
LU-1
Pectoralis major, clavicular head
LU-3
Triceps brachii
L.I.-14
Biceps brachii

Fig. 3.51

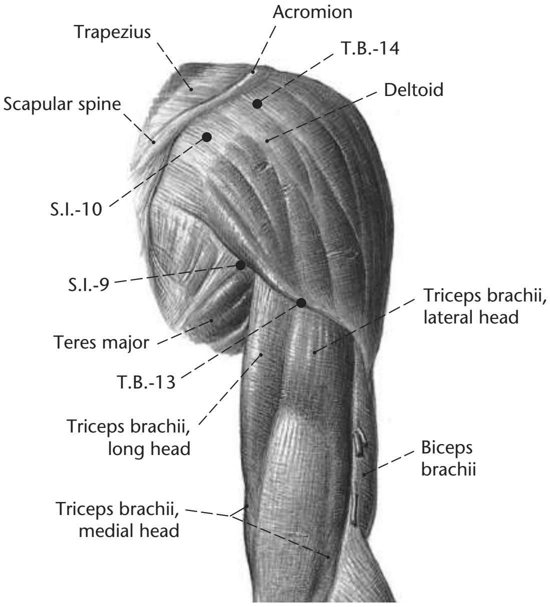
Trapezius
Acromion
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

Fig. 3.49

3.3.2 Elbow

Elbow, tendon of the biceps muscle (→ Figs 3.52, 3.53)

The anterior aspect of the elbow is marked by the cubital crease, extending from the lateral to the medial epicondyle of the humerus. It is divided by the tendon of the biceps muscle. → **L.I.-11** and → **LU-5** are located lateral to the tendon while → **P-3** and → **HE-3** are located medial to it.

Lateral and medial epicondyles of the humerus (→ Figs 3.53, 3.55, 3.56)

The lateral and medial epicondyles of the humerus form the lateral aspects of the elbow. The olecranon is located between them, on the posterior aspect of the elbow. The medial epicondyle is the origin of the extensor muscles of the wrist joint, while the flexor muscles originate at the medial epicondyle.

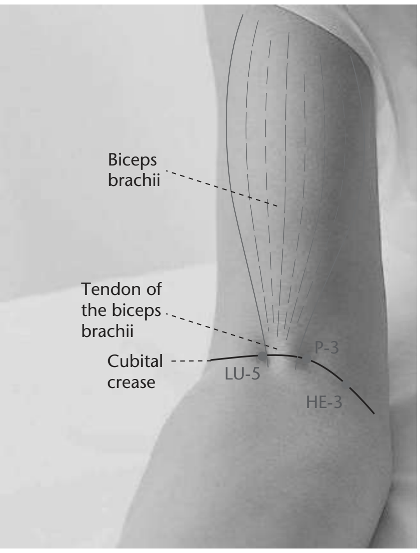
Biceps brachii
Tendon of the biceps brachii
Cubital crease
LU-5
P-3
HE-3

Fig. 3.52

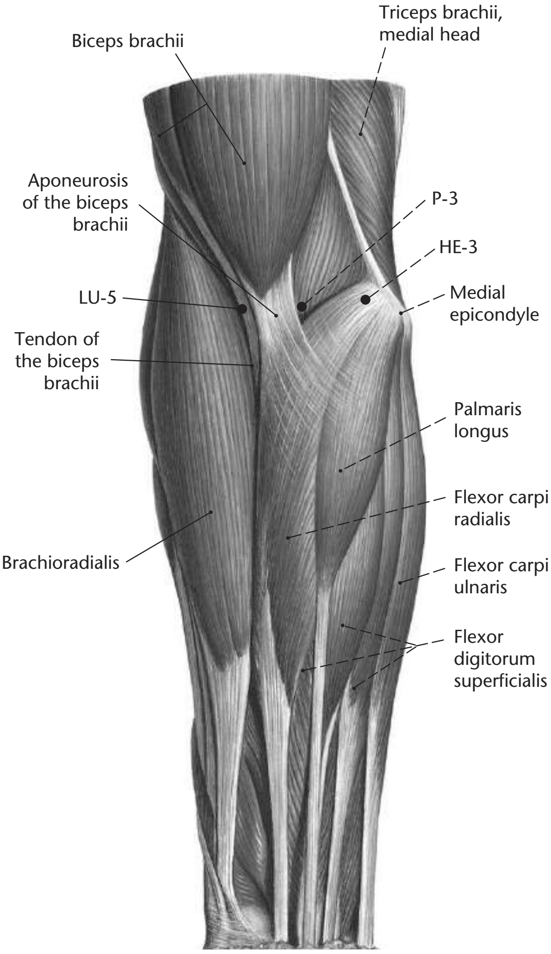
Biceps brachii
Triceps brachii, medial head
Aponeurosis of the biceps brachii
P-3
HE-3
LU-5
Medial epicondyle
Tendon of the biceps brachii
Palmaris longus
Flexor carpi radialis
Brachioradialis
Flexor carpi ulnaris
Flexor digitorum superficialis

Fig. 3.53

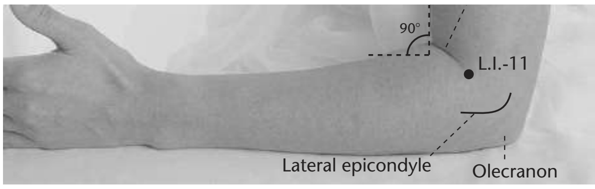
90°
L.I.-11
Lateral epicondyle
Olecranon

Fig. 3.54

Olecranon (→ Figs 3.55, 3.56)

The olecranon forms the tip of the elbow. Located in the ulnar sulcus, between the medial epicondyle and the tip of the olecranon, is → **S.I.-8**, while → **T.B.-10** is located in a depression proximal to the olecranon.

3.3.3 Forearm and Hand

For locating points on the forearm, it is important to note whether the arm is in a pronated or supinated position, since this can have a considerable effect on the relative location of some points (see also location details of specific points). For example, in supination the line connecting → **L.I.-5** and → **L.I.-11** runs along the lateral border of the forearm, while in pronation it traverses the forearm.

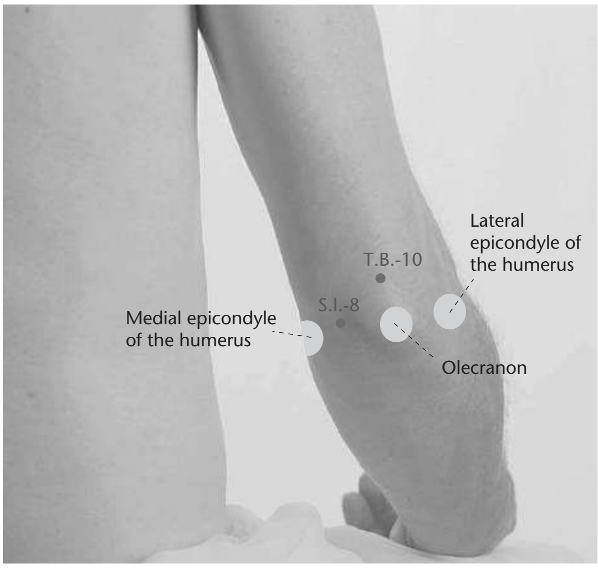
T.B.-10
Lateral epicondyle of the humerus
S.I.-8
Medial epicondyle of the humerus
Olecranon

Fig. 3.55

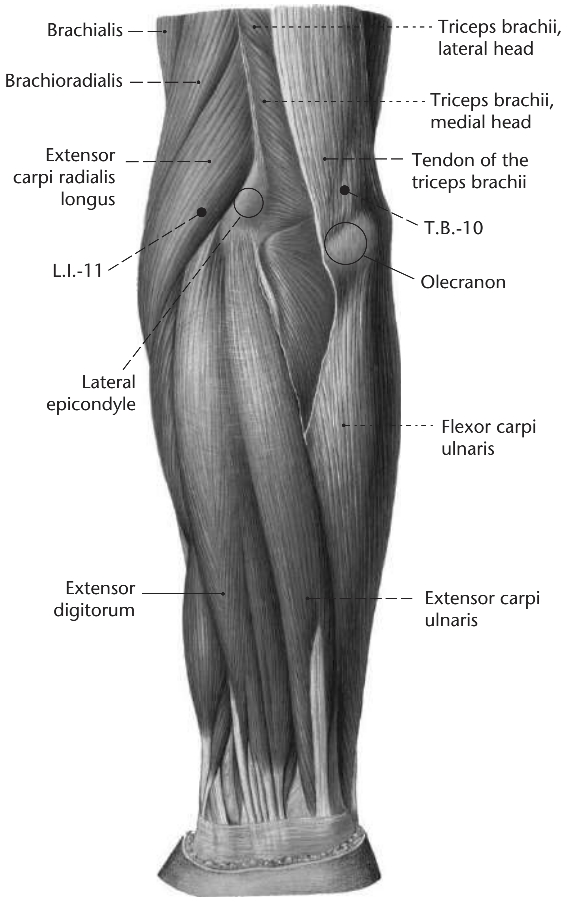
Brachialis
Triceps brachii, lateral head
Brachioradialis
Triceps brachii, medial head
Extensor carpi radialis longus
Tendon of the triceps brachii
T.B.-10
L.I.-11
Olecranon
Lateral epicondyle
Flexor carpi ulnaris
Extensor digitorum
Extensor carpi ulnaris

Fig. 3.56

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Lumbosacral junction

Following orientation in the lumbar region with the help of the iliac crests (→ Iliac crest), the lumbosacral joint can be located by counting the spinous processes in an inferior direction. It can generally be felt as a marked ‘kink’ or depression, followed by the smaller spinous processes of the slightly convex median sacral crest directly below.

Caution: Note that (according to the literature in up to 15% of cases) there can be lumbarization of the 1st sacral vertebra or sacralization of L5. In these cases there will be either four or six lumbar vertebrae.

Posterior superior iliac spine (PSIS) (→ Figs 3.73, 3.74)

Another important landmark on the lower back is the posterior superior iliac spine (PSIS), which forms the posterior end of the iliac crest lateral to the upper sacral region. The PSIS can often be identified by two superficially visible dimples. It is best located by palpating from an inferior in a superior direction. If the dimples are not visible, they can be located by palpating 3 cun from the rima ani in a superior and lateral direction at an angle of 45°, until a marked bony ridge can be felt.

Generally, the PSIS is located at the level of the space between the 1st and 2nd sacral foramina, so that → **BL-27** is located directly medial or slightly superior and medial to the PSIS, while → **BL-28** is located just inferior and medial to the PSIS.

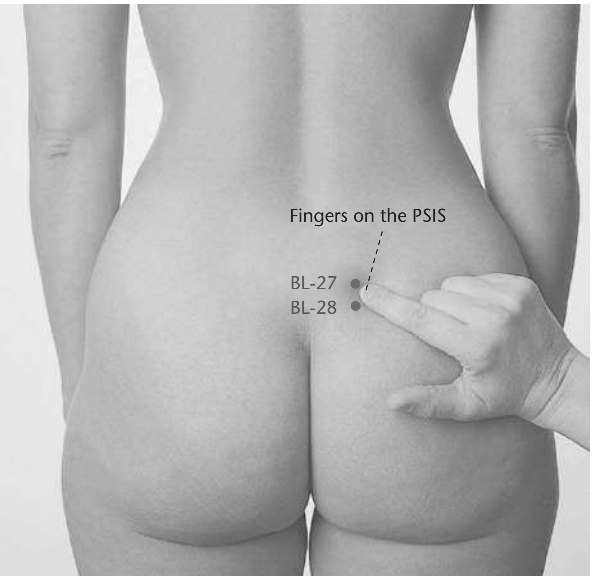

Fig. 3.73

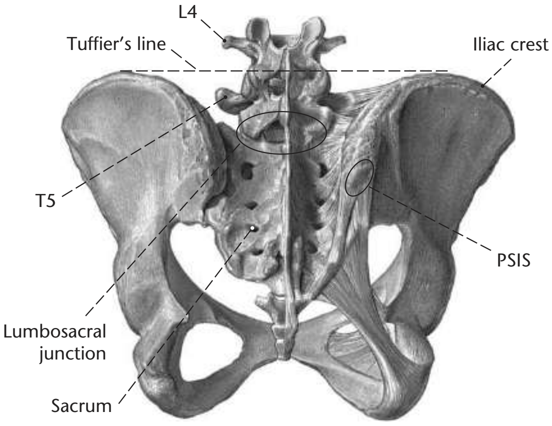

Fig. 3.74

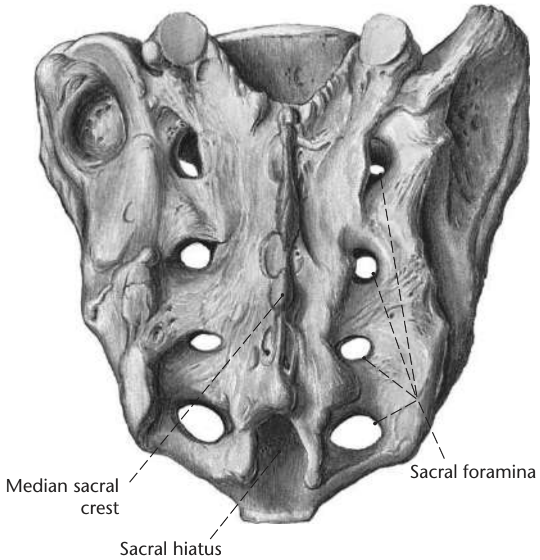
Median sacral crest
Sacral foramina
Sacral hiatus

Fig. 3.75

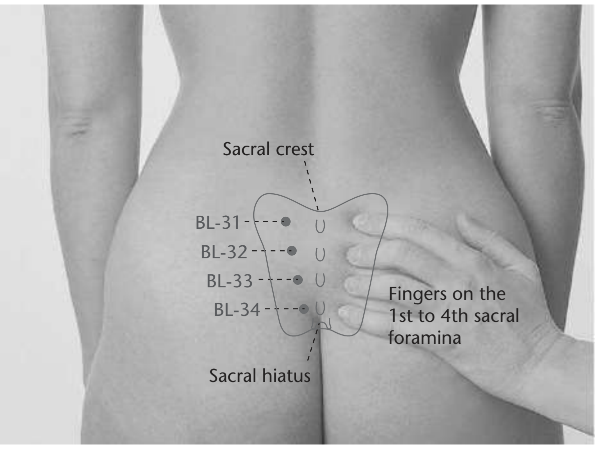
Sacral crest
BL-31
BL-32
BL-33
BL-34
Fingers on the 1st to 4th sacral foramina
Sacral hiatus

Fig. 3.76

3.4.4 Sacrum and Sacroiliac Joint

Sacrum (→ Figs 3.75, 3.76)

On the sacrum, landmarks for palpation are the median sacral crest and the sacral hiatus as well as the sacral foramina on its lateral aspects.

Sacral crest (→ Figs 3.74, 3.75)

The sacral crest is formed by the narrow, irregular spinous processes, which can be palpated on the sacral midline.

Sacral hiatus (→ Figs 3.75, 3.76)

The sacral hiatus can be palpated as a U-shaped depression at the inferior end of the sacral crest. → **Du-2** is located on the midline, just inferior to it. Both the sacral hiatus and the greater trochanter (→ Lower extremity) are reference points for locating → **G.B.-30.**

Sacral foramina (→ Figs 3.75, 3.76)

The sacral foramina are located between the lumbosacral joint and the sacral hiatus and can usually be palpated. They are in most cases evenly spaced and lie approximately 1 finger-width lateral to the midline, with the distance decreasing in a distal direction. The foramina are the location of → **BL-31–BL-34.**

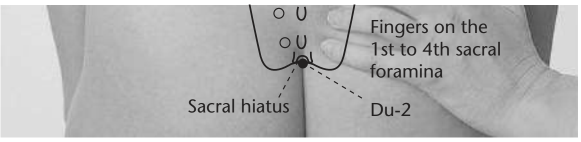
Fingers on the 1st to 4th sacral foramina
Sacral hiatus
Du-2

Fig. 3.77

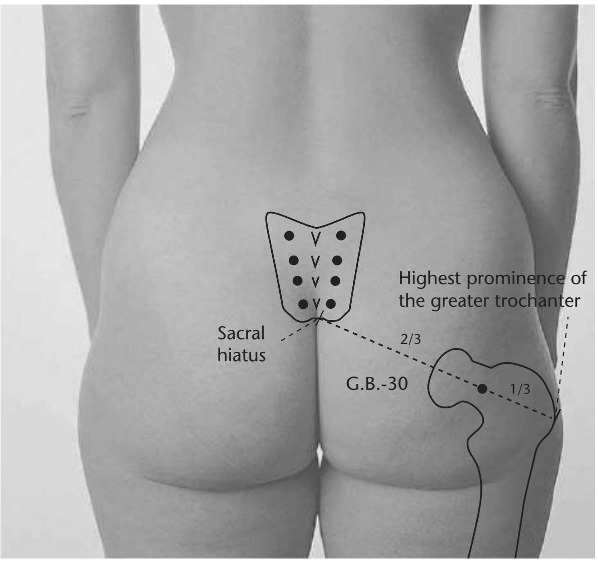
Highest prominence of the greater trochanter
Sacral hiatus
2/3
G.B.-30
1/3

Fig. 3.78

Iliosacral joint (→ Figs 3.80, 3.81)

The iliosacral joint can be palpated bilaterally as a shallow groove between the sacrum, the PSIS (see above) and more inferiorly located ilial structures. → **BL-26–BL-29** are located in this area.

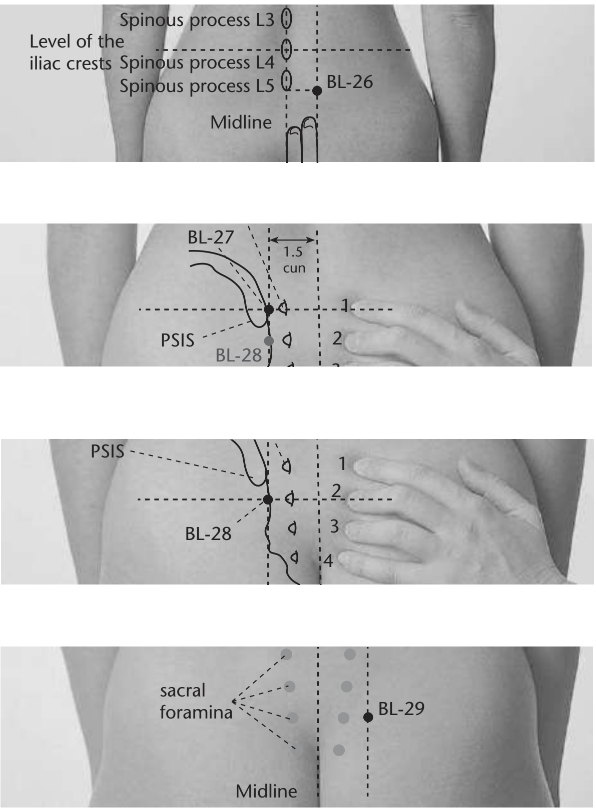
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
BL-26
Midline
BL-27
1.5 cun
1
PSIS
2
BL-28
PSIS
1
2
BL-28
3
4
sacral foramina
BL-29
Midline

Fig. 3.79

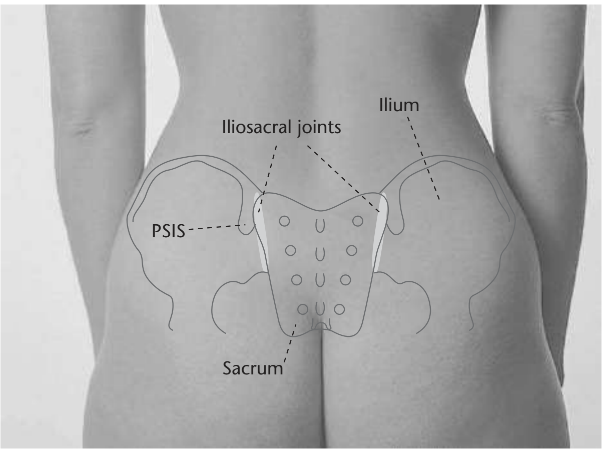
Ilium
Iliosacral joints
PSIS
Sacrum

Fig. 3.80

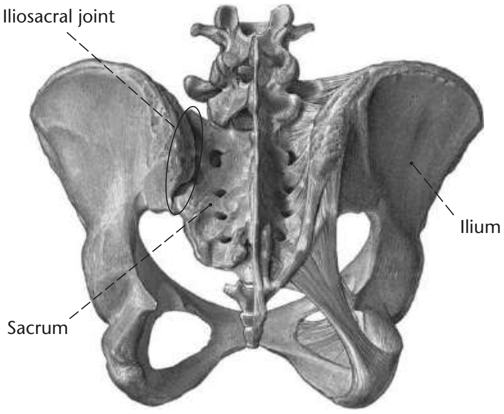
Iliosacral joint
Ilium
Sacrum

Fig. 3.81

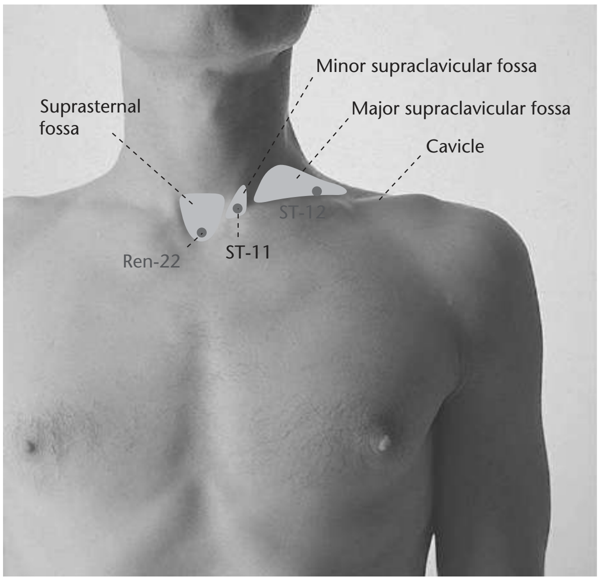
Minor supraclavicular fossa
Suprasternal fossa
Major supraclavicular fossa
Cavicle
ST-12
Ren-22
ST-11

Fig. 3.82

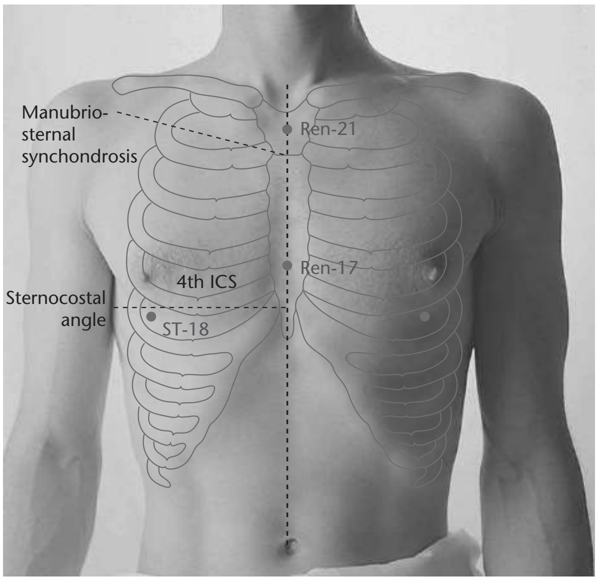
Manubrio-sternal synchondrosis
Ren-21
4th ICS
Ren-17
Sternocostal angle
ST-18

Fig. 3.83

3.5 Anterior and Lateral Aspect of the Thorax and Abdomen

Clavicle (→ Figs 3.82, 3.86)

The clavicle forms the connection between the shoulder girdle and the thorax. It separates the major and minor supraclavicular fossae with → **ST-11, ST-12** and ***jingbi*** (Ex-HN) from the ribcage. Located on the lower border of the clavicle are → **ST-13** (at its midpoint) and → **KID-27** (2 cun lateral to the midline).

Suprasternal fossa (→ Fig. 3.82)

The suprasternal fossa forms a depression superior to the sternum between the sternoclavicular joints. It is the location of → **Ren-22**.

Ribs (→ Figs 3.83, 3.86)

For anatomical orientation of the ribs, the midsection of the clavicle is a good starting point. The first rib lies directly below the clavicle, forming a narrow bow. It articulates with the manubrium just inferior to the sternoclavicular joint.

Please note the following when palpating the ribs and intercostal spaces (ICS) and locating the respective acupuncture points:

- From their articulation with the sternum, the ribs run initially horizontally or in a gentle curve in an inferior direction, while at the lateral aspect of the thorax, they assume a superior direction towards the posterior aspect of the body.
- Inferior to the 4th/5th intercostal spaces, the medial sections of the ribs merge with the costal cartilage, so that the intercostal spaces do not extend very far medially. Therefore it is necessary to palpate somewhat more laterally.

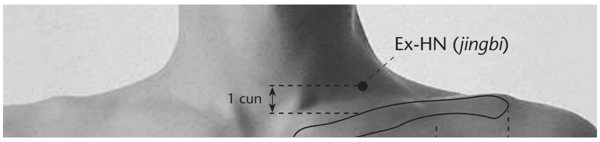
Ex-HN (*jingbi*)
1 cun

Fig. 3.84A

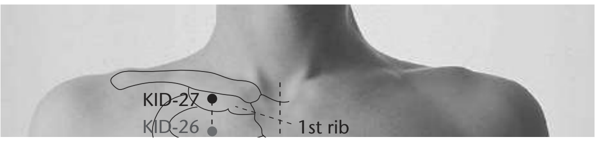
KID-27
KID-26
1st rib

Fig. 3.84C

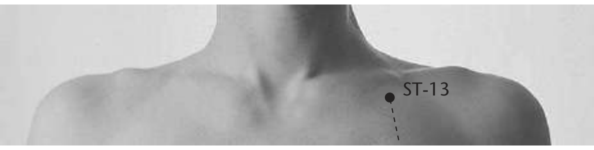

Fig. 3.84B

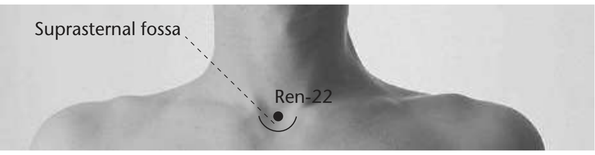
Suprasternal fossa
Ren-22

Fig. 3.84D

1st rib, 1st intercostal space (→ Fig. 3.86)

By palpating parasternally in an inferior direction, the first rib can generally be felt directly below the clavicle. Here, → **KID-27** is located 2 cun lateral to the anterior midline. Below follows the first intercostal space, then the second rib and so on. → **Ren-20** is located on the anterior midline, at the level of the first intercostal space, → **Ren-21** is located slightly superior to it.

2nd rib, manubriosternal synchondrosis (manubriosternal junction) (→ Figs 3.83, 3.86)

The synchondrosis is usually a clearly defined transverse bony structure on the superior section of the sternum, level with the articulation of the second rib. Directly below the synchondrosis is the 2nd intercostal space from where subsequent ribs and intercostal spaces can be counted by palpation.

Laterosuperior aspect of the thorax and coracoid process (→ Fig. 3.87)

The laterosuperior aspect of the thorax is bordered by the clavicle and the deltoid muscle with the coracoid process deep to the latter. The deltopectoral triangle can be located by palpating from the end of the anterior axillary fold along the border of the deltoid muscle towards the shoulder and the lower border of the clavicle. Here, in the centre of the triangle, lies → **LU-2**. Laterally, the triangle is bordered by a clearly palpable bony structure, the coracoid process. In order to distinguish the process from the lesser tubercle of the humerus, the arm is rotated externally with the elbow flexed at the same time. While the process will remain fixed, the tubercle will follow the movement. → **LU-1** is located slightly superior and lateral to → **LU-2** and medial to the lower border of the coracoid process. The contours of the deltoid muscle and the deltopectoral triangle can be emphasised through muscle flexion, for example by pressing the hands against each other in front of the chest.

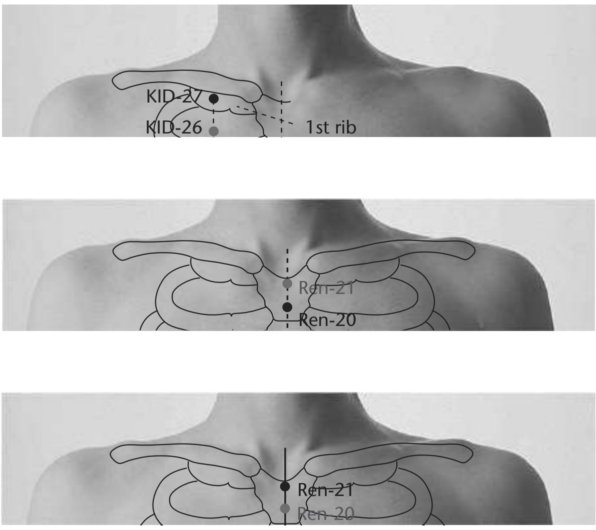
KID-27
KID-26
1st rib
Ren-21
Ren-20
Ren-21
Ren-20

Fig. 3.85

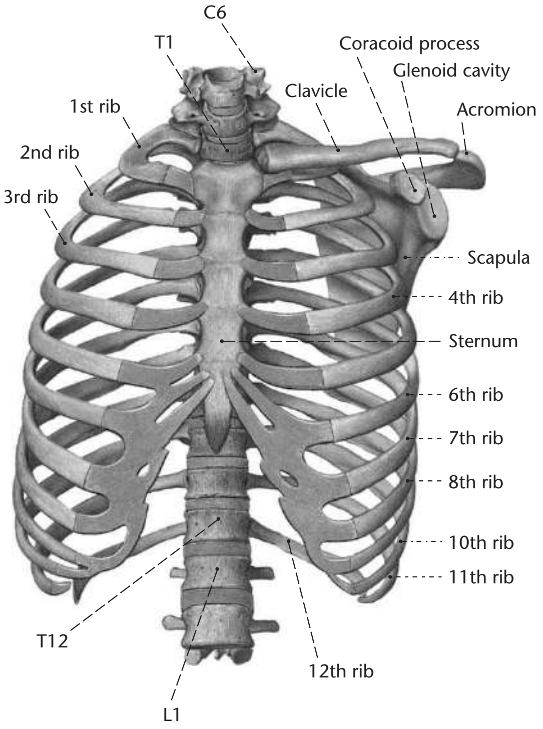
C6
T1
Coracoid process
Glenoid cavity
Clavicle
Acromion
1st rib
2nd rib
3rd rib
Scapula
4th rib
Sternum
6th rib
7th rib
8th rib
10th rib
11th rib
T12
12th rib
L1

Fig. 3.86

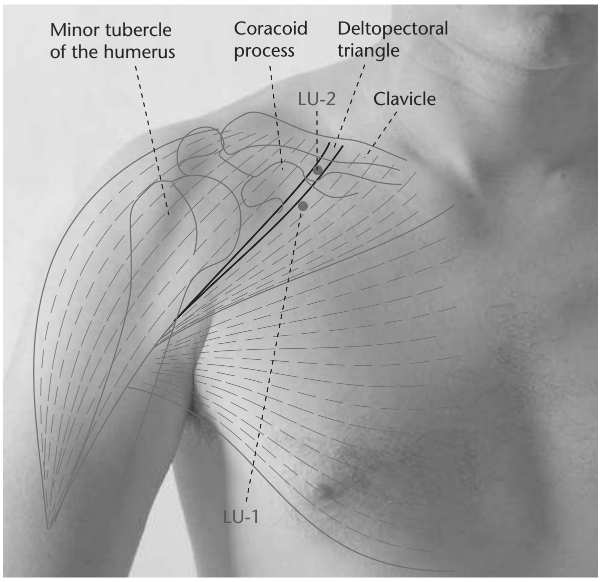
Minor tubercle of the humerus
Coracoid process
Deltopectoral triangle
LU-2
Clavicle
LU-1

Fig. 3.87

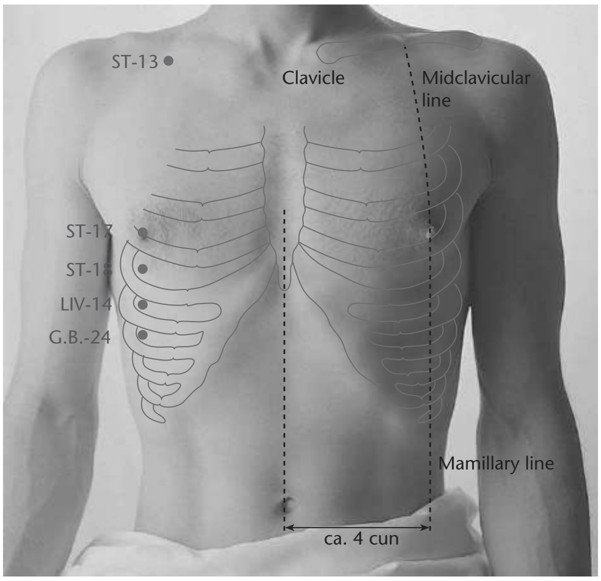

Fig. 3.88

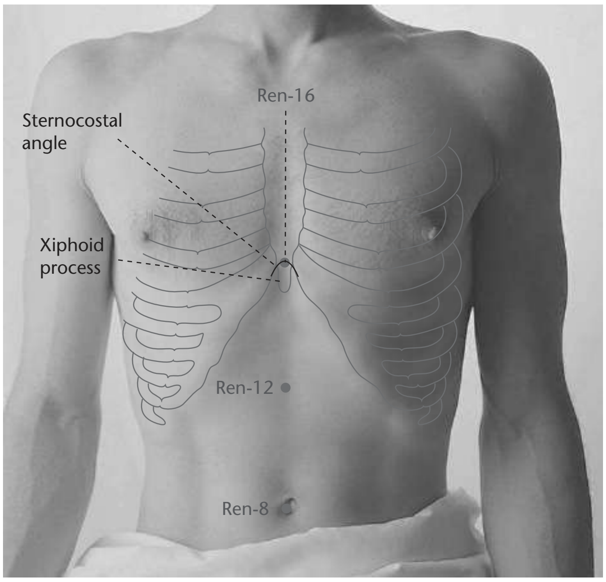

Fig. 3.89

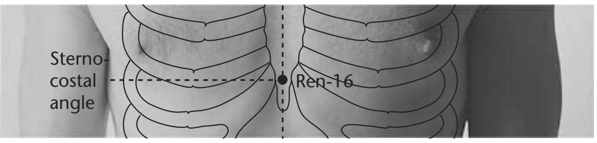

Fig. 3.90

Midclavicular line (→ Fig. 3.88)

The midclavicular line provides a further tool for reference on the upper thorax. It is an imaginary line that runs from the midpoint of the clavicle to the nipple. Since in most cases the nipple lies somewhat more laterally than the clavicular midpoint, the line tends to have a slightly oblique pathway. → **ST-13** to **ST-16** are all located on the line.

Nipple (→ Fig. 3.88)

In men, the nipple is located in the 4th intercostal space, lateral to the midline, while in women, the location can vary depending on the shape of the breast. → **ST-17** is located on the nipple.

Mamillary line (→ Fig. 3.88)

The mamillary line is an imaginary line for orientation on the mediolateral thorax and abdomen. It runs vertically from the nipple in an inferior direction. The points → **ST-18, LIV-14** and **G.B.-24** are all located on this line.

Sternocostal angle, xiphoid process and umbilicus (→ Figs 89, 3.91)

The lower borders of the ribcage unite at the lower sternal border, forming the sternocostal angle. This angle and the umbilicus are the two major anatomical landmarks for point location on the upper abdomen.
Caution: It is important not to mistake the sternocostal angle for the tip of the xiphoid process. The latter is attached to the sternum at the sternocostal angle, projecting inferiorly towards the abdomen. For definite identification, palpate along the costal cartilage in a superior direction to where it meets the sternocostal angle. This is the location of → **Ren-16**.

Axillary line (→ Fig. 3.92)

The axillary line is an imaginary line for orientation on the lateral aspect of the thorax and abdomen. It runs vertically from the middle of the axilla in an inferior direction. Both → **G.B.-22** and **SP-21** are located on this line.

Free end of the 11th rib (→ Figs 3.91, 3.92)

The free end of the 11th rib can be palpated below the costal cartilage, at the transitional zone from the thorax to the abdomen. Location tip: the patient places his/her flexed elbow (90° angle) alongside his/her thorax. The tip of the olecranon will then usually be close to the free end of the 11th rib. → **LIV-13** is located below its anterior lower border.

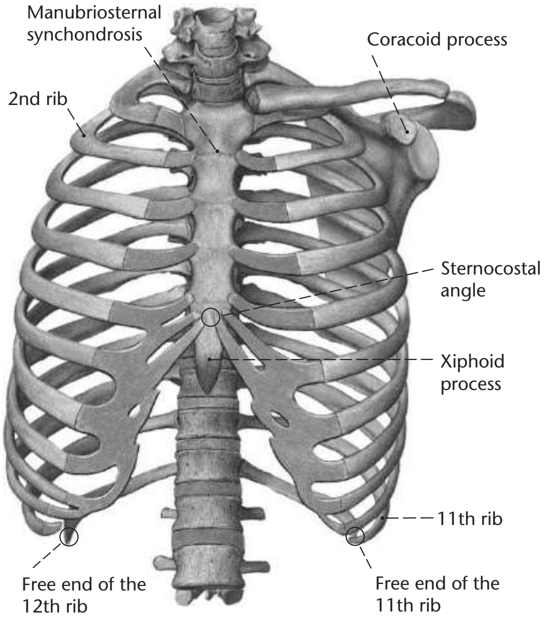
Manubriosternal synchondrosis
Coracoid process
2nd rib
Sternocostal angle
Xiphoid process
11th rib
Free end of the 12th rib
Free end of the 11th rib

Fig. 3.91

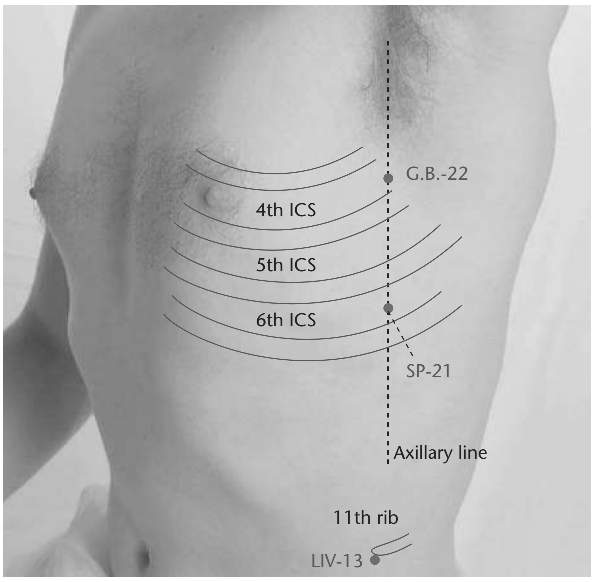
G.B.-22
4th ICS
5th ICS
6th ICS
SP-21
Axillary line
11th rib
LIV-13

Fig. 3.92

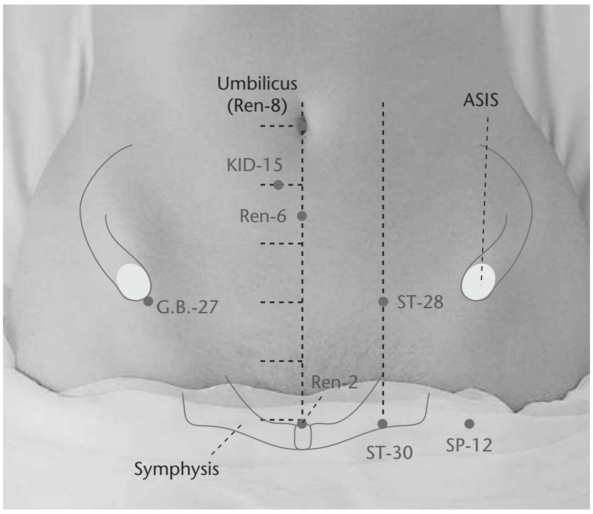
Umbilicus
(Ren-8)
ASIS
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

Fig. 3.93

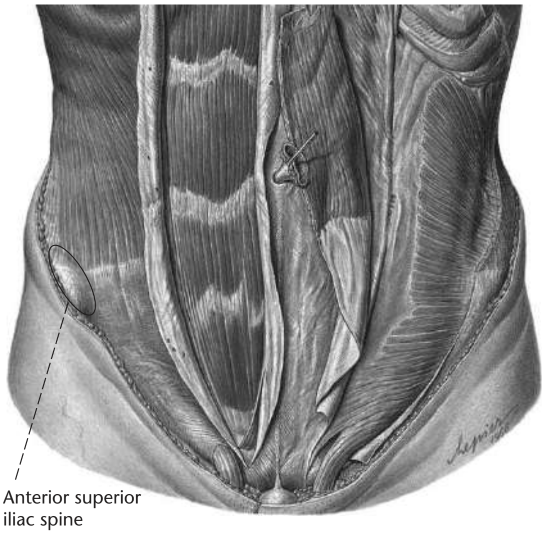
Anterior superior
iliac spine

Fig. 3.94

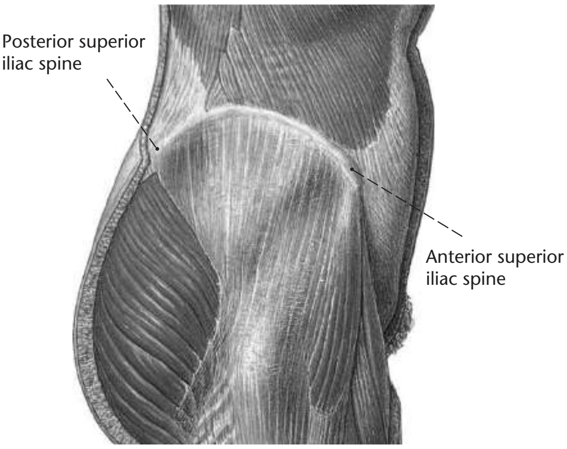
Posterior superior
iliac spine
Anterior superior
iliac spine

Fig. 3.95

Anterior superior iliac spine (ASIS) (→ Figs 3.93–3.95)

By palpating along the iliac crest in an anterior and inferior direction, the anterior superior iliac spine (ASIS) is reached at the lateral end of the inguinal groove. The ASIS is felt as a distinct bony ridge. → **G.B.-27** is located anterior to it, while → **G.B.-28** is located slightly more inferiorly and medially.

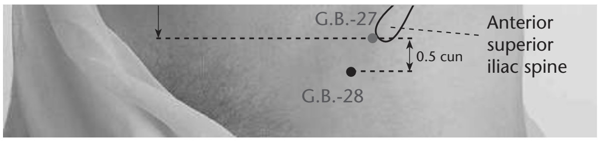
G.B.-27
0.5 cun
Anterior
superior
iliac spine
G.B.-28

Fig. 3.96

Pubic symphysis, umbilicus (→ Fig. 3.98)

The upper border of the pubic symphysis is an important orientation point for point location on the lower abdomen. The distance between the centre of the umbilicus and the upper border of the symphysis can be divided into 5 proportional cun, which can differ significantly from the patient's finger cun (→ Chapter 2). → **Ren-2** is located on the anterior midline and on the upper border of the pubic symphysis.

3.6 Lower Extremity

Greater trochanter (→ Fig. 3.99)

The greater trochanter forms a distinct, projecting bony structure on the hip area. It is one of the endpoints on an imaginary line connecting it with the sacral hiatus. Located on this line is → **G.B.-30**, a third of the distance from the greater trochanter. Another line connecting the greater trochanter and the anterior superior iliac spine (ASIS) is used for locating → **G.B.-29**.

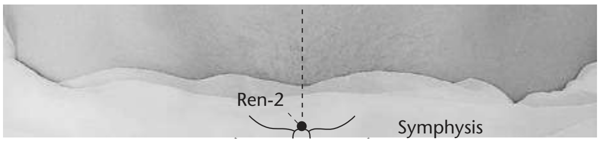
Ren-2
Symphysis

Fig. 3.97

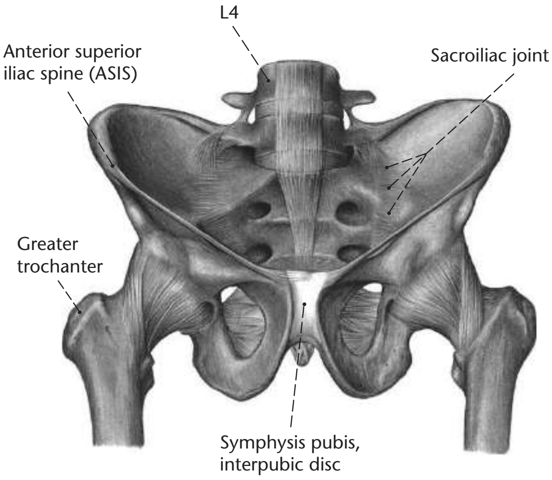
L4
Anterior superior iliac spine (ASIS)
Sacroiliac joint
Greater trochanter
Symphysis pubis, interpubic disc

Fig. 3.98

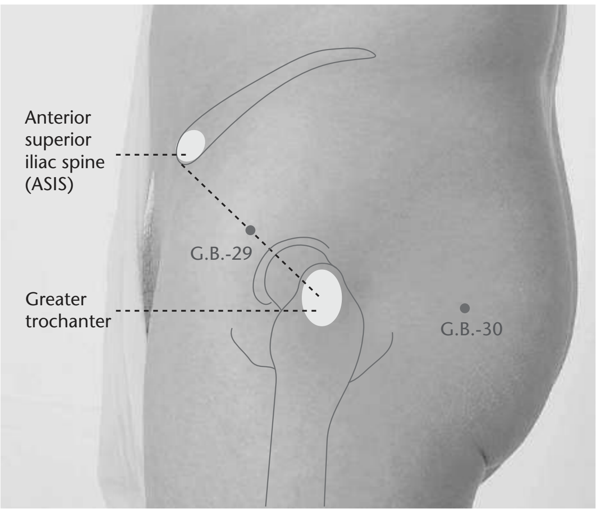
Anterior superior iliac spine (ASIS)
G.B.-29
Greater trochanter
G.B.-30

Fig. 3.99

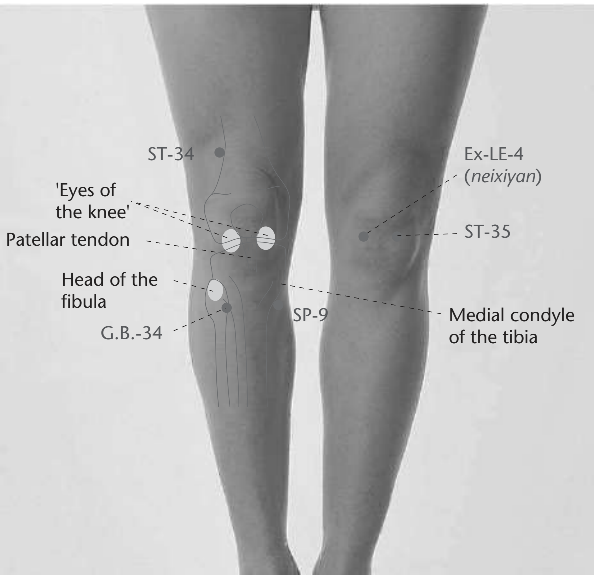
ST-34
'Eyes of the knee'
Patellar tendon
Head of the fibula
G.B.-34
SP-9
Ex-LE-4 (*neixiyan*)
ST-35
Medial condyle of the tibia

Fig. 3.100

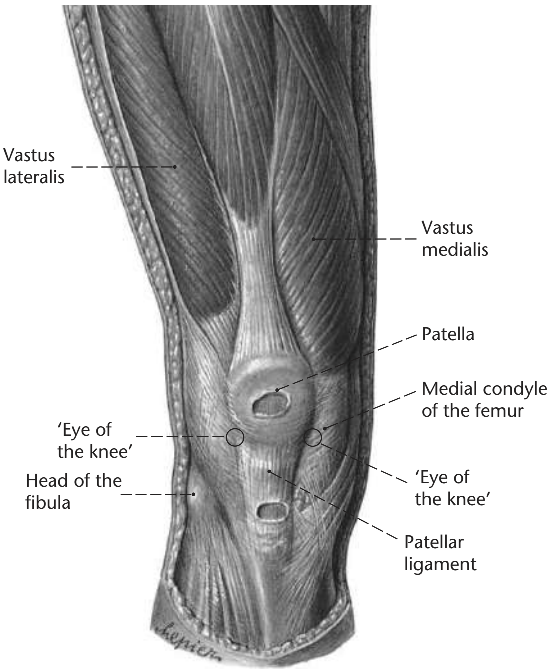
Vastus lateralis
Vastus medialis
Patella
Medial condyle of the femur
'Eye of the knee'
Head of the fibula
'Eye of the knee'
Patellar ligament

Fig. 3.101

3.6.1 Knee Region

For locating and needling points in the knee region, the patient should be in a relaxed supine position, with the knees slightly flexed and supported by pillows. This opens up the knee joint and allows for better palpation of the eyes of the knees. A prone position with a support below the feet and ankles is recommended for treating points in the popliteal crease.

Patella (→ Figs 3.100, 3.101)

The upper patellar border marks the location of → **SP-10** as well as of → **ST-32** to **ST-34**.

Eyes of the knee (→ Figs 3.100, 3.101)

The eyes of the knee are formed by the depressions on either side of the patellar tendon at the level of the lower patellar border. The lateral eye corresponds to the location of → **ST-35**, the medial eye is the location of the extra point → **Ex-LE-4** (*neixiyan*).

Head of the fibula (→ Figs 3.100, 3.101)

The head of the fibula forms a further important reference point inferior to the lateral aspect of the knee joint. For locating → **G.B.-34**, find the head of the fibula (approximately where the seam of a pair of trousers would be) and hold it with two fingers. By sliding downward, the anterior finger will 'drop' into the depression anterior and inferior to the head of the fibula, which marks the location of → **G.B.-34**.

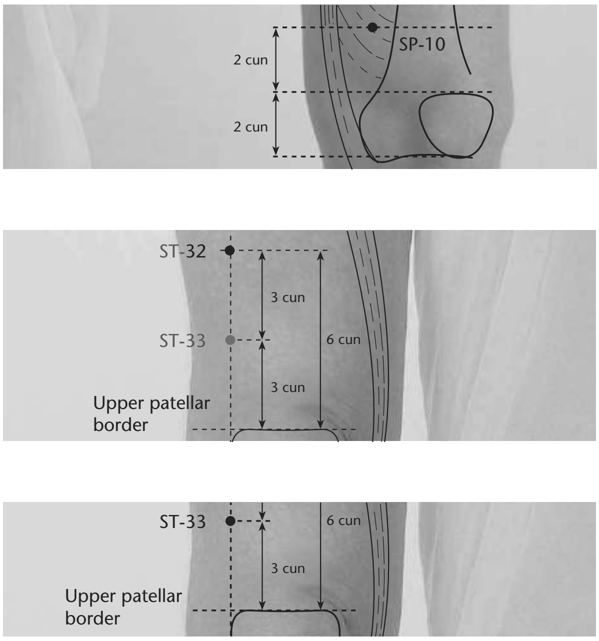
2 cun
2 cun
SP-10
ST-32
3 cun
ST-33
6 cun
3 cun
Upper patellar border
ST-33
6 cun
3 cun
Upper patellar border

Fig. 3.102

Tibial shaft and medial condyle of the tibia (➞ Fig. 3.103)

Below the medial aspect of the knee joint, the transition from the medial condyle of the tibia to the tibial shaft marks the location of ➞ **SP-9**. This transition is best located by palpating in a superior direction along the posterior border of the medial aspect of the tibia. When using a knee roll, be aware that the soft tissue around this area may be pushed anteriorly, hindering palpation.

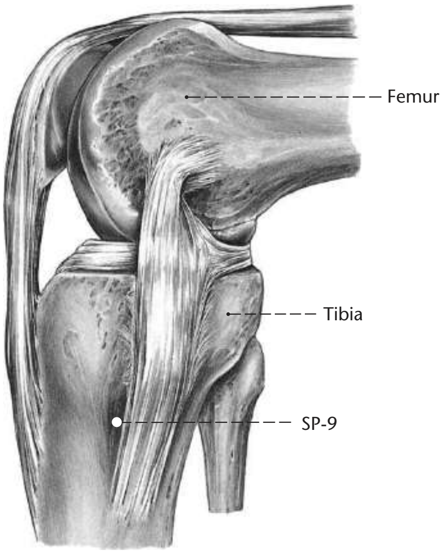
Femur
Tibia
SP-9

Fig. 3.103

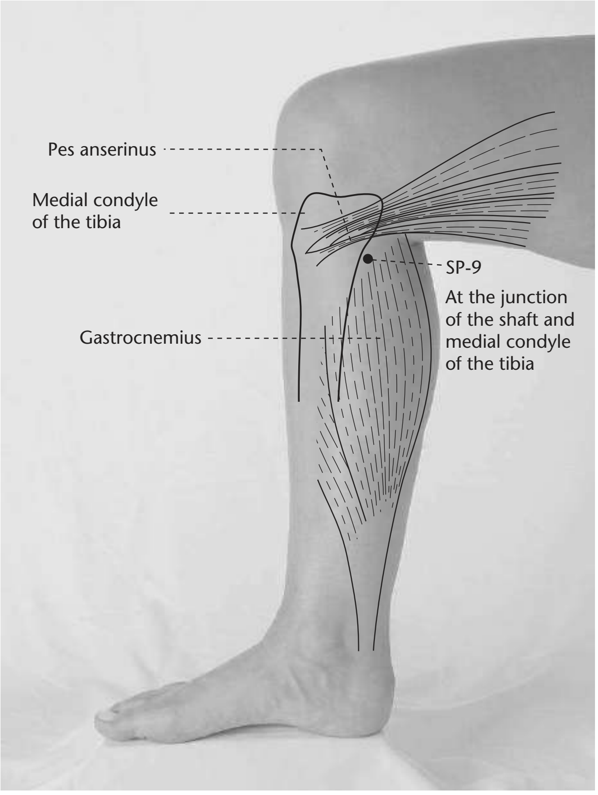
Pes anserinus
Medial condyle of the tibia
Gastrocnemius
SP-9
At the junction of the shaft and medial condyle of the tibia

Fig. 3.104

3.6.2 Ankle and Foot

Highest prominence of the medial/lateral malleolus (→ Figs 3.105, 3.106)

Please note: The highest prominence of the malleolus does not refer to its lower border, but literally to its most protruding tip. The prominences are important reference points for the lower leg and ankle. → **Ex-LE-8** (*neihuajian*) is located on the highest prominence of the lateral malleolus, while → **Ex-LE-9** (*waihuajian*) is its medial counterpart.

Please also note: Contrary to some textbooks, the lateral and medial prominences are anatomically not located at the same level. There is approximately a 1 cun (→ Chapter 1, Fig. 1.2) difference in level. Thus, → **BL-60** and → **KID-3** are not located directly opposite each other, but rather occupy a comparable position.

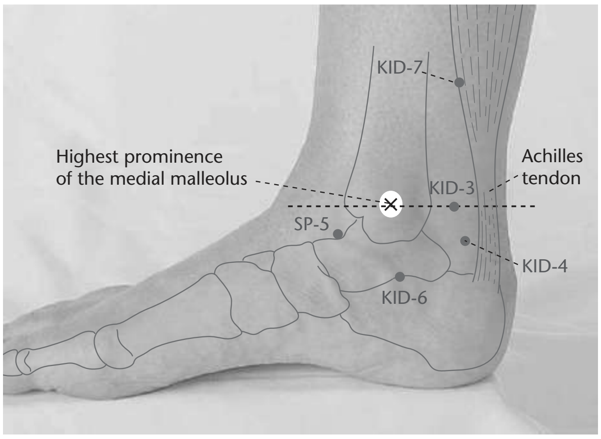
KID-7
Highest prominence
of the medial malleolus
Achilles
tendon
KID-3
SP-5
KID-4
KID-6

Fig. 3.105

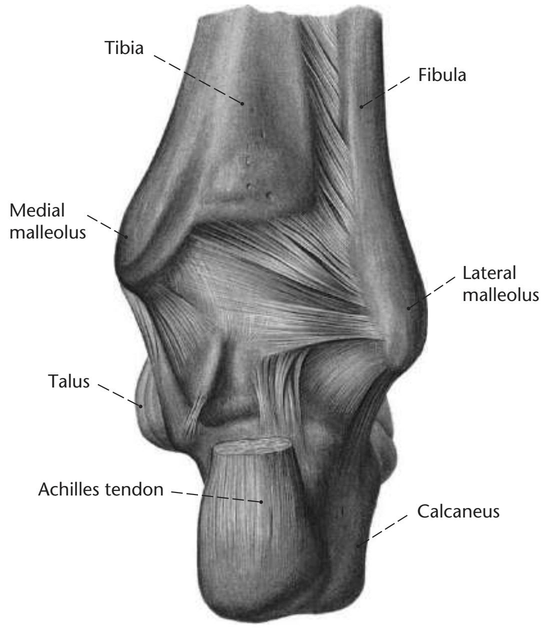
Tibia
Fibula
Medial
malleolus
Lateral
malleolus
Talus
Achilles tendon
Calcaneus

Fig. 3.106

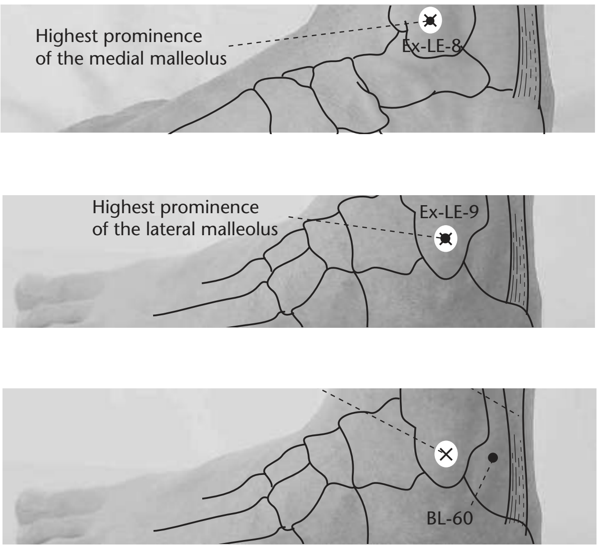
Highest prominence
of the medial malleolus
Ex-LE-8
Highest prominence
of the lateral malleolus
Ex-LE-9
BL-60

Fig. 3.107

Joint space on the anterior ankle (→ Figs 3.108, 3.109)

The joint space on the anterior aspect of the ankle (between the malleoli) is best palpated by passively flexing and extending the foot. → **LIV-4** and → **ST-41** are both located here.

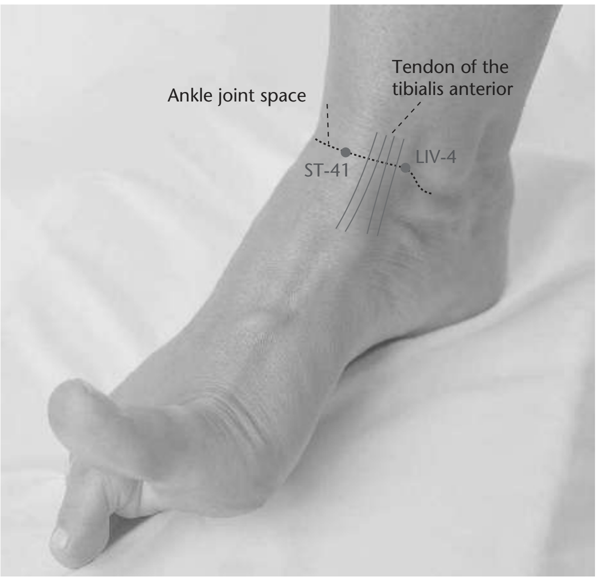
Tendon of the tibialis anterior
Ankle joint space
ST-41
LIV-4

Fig. 3.108

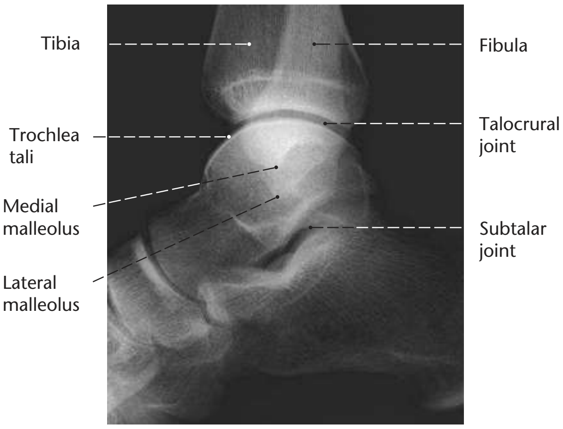
Tibia
Fibula
Trochlea tali
Talocrural joint
Medial malleolus
Subtalar joint
Lateral malleolus

Fig. 3.109

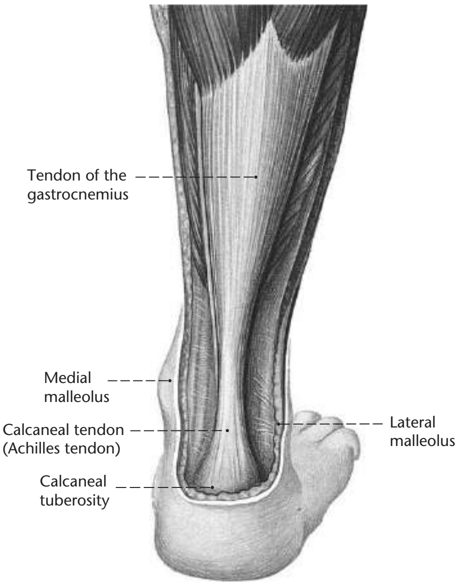
Tendon of the gastrocnemius
Medial malleolus
Calcaneal tendon (Achilles tendon)
Calcaneal tuberosity
Lateral malleolus

Fig. 3.110

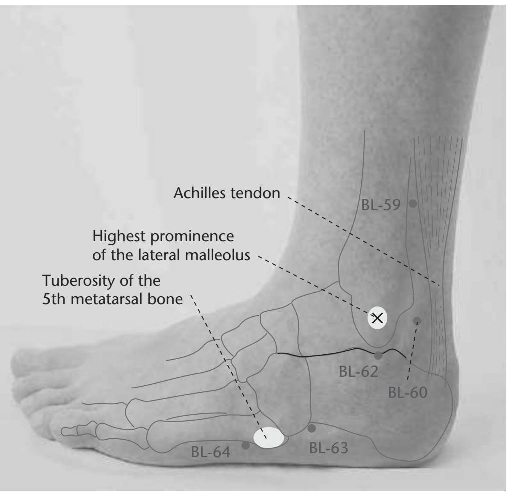
Achilles tendon
Highest prominence of the lateral malleolus
Tuberosity of the 5th metatarsal bone
BL-59
BL-62
BL-60
BL-64
BL-63

Fig. 3.111

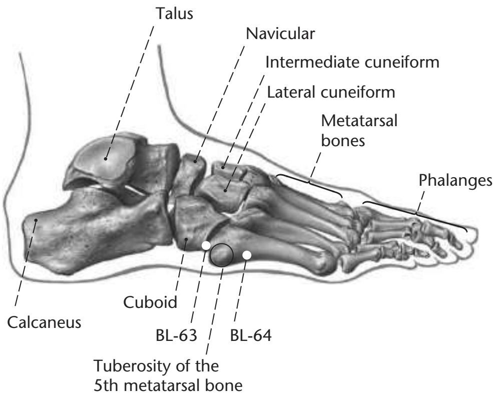
Talus
Navicular
Intermediate cuneiform
Lateral cuneiform
Metatarsal bones
Phalanges
Calcaneus
Cuboid
BL-63
BL-64
Tuberosity of the 5th metatarsal bone

Fig. 3.112

Achilles tendon (→ Fig. 3.110)

On the distal aspect of the lower leg and on the heel, the Achilles tendon defines the location of → **KID-3, KID-4**, and **KID-7** on its medial aspect and of → **BL-59** and **BL-60** on its lateral aspect.

Tuberosity of the 5th metatarsal bone (→ Figs 3.111, 3.112)

The tuberosity at the proximal end of the 5th metatarsal bone is the most palpable orientation point on the lateral aspect of the midfoot. → **BL-63** is located in the depression proximal to the tuberosity, while → **BL-64** is located in the depression distal to it, at the junction of the head and the shaft.

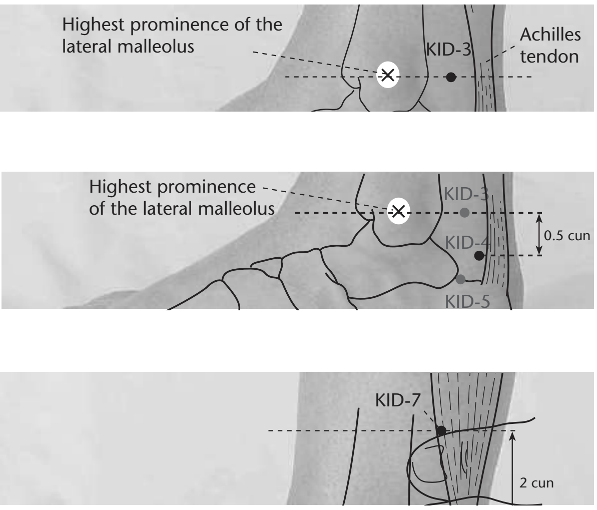
Highest prominence of the lateral malleolus
KID-3
Achilles tendon
Highest prominence of the lateral malleolus
KID-3
KID-4
KID-5
0.5 cun
KID-7
2 cun

Fig. 3.113

Base of the 1st metatarsal bone (→ Figs 3.114, 3.115)

When palpating along the medial aspect of the midfoot in a proximal direction, the proximal end (base) of the 1st metatarsal bone forms the first distinct prominence. → **SP-4** is located distal to this bony orientation point, at the junction of the shaft and the base.

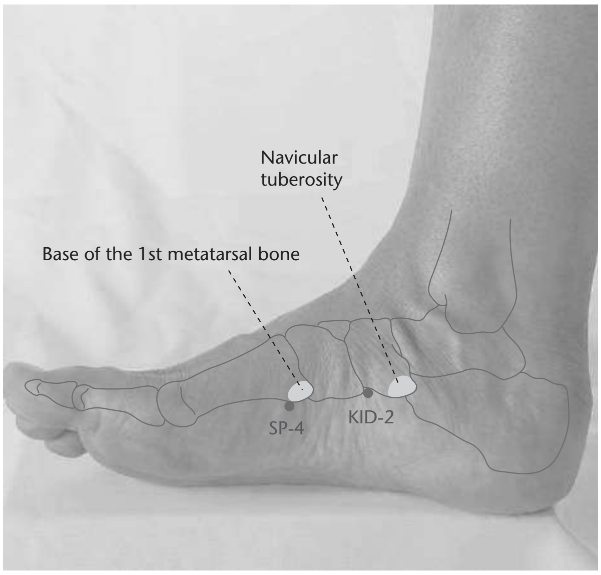
Navicular tuberosity
Base of the 1st metatarsal bone
SP-4
KID-2

Fig. 3.114

Navicular tuberosity (→ Fig. 3.115)

When palpating along the medial side of the foot, past → **SP-4** and the base of the 1st metatarsal bone, a further bony structure can be clearly felt. This is the navicular tuberosity, where → **KID-2** is located.

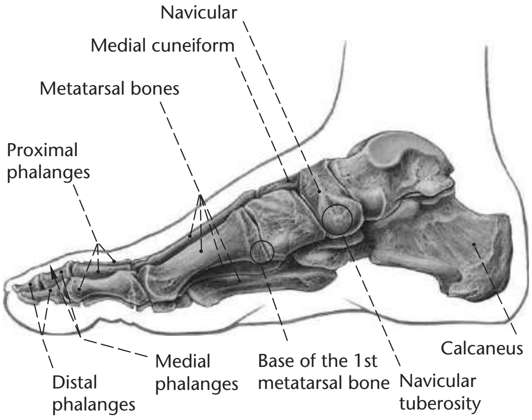
Navicular
Medial cuneiform
Metatarsal bones
Proximal phalanges
Distal phalanges
Medial phalanges
Base of the 1st metatarsal bone
Calcaneus
Navicular tuberosity

Fig. 3.115

4 Acupuncture Points of the Twelve Primary Channels

Claudia Focks, Ulrich März

4.1 The Lung Channel System – Hand-*taiyin* (*shou tai yin jing luo*)

4.1.1 The Lung Primary Channel (*shou tai yin jing*)

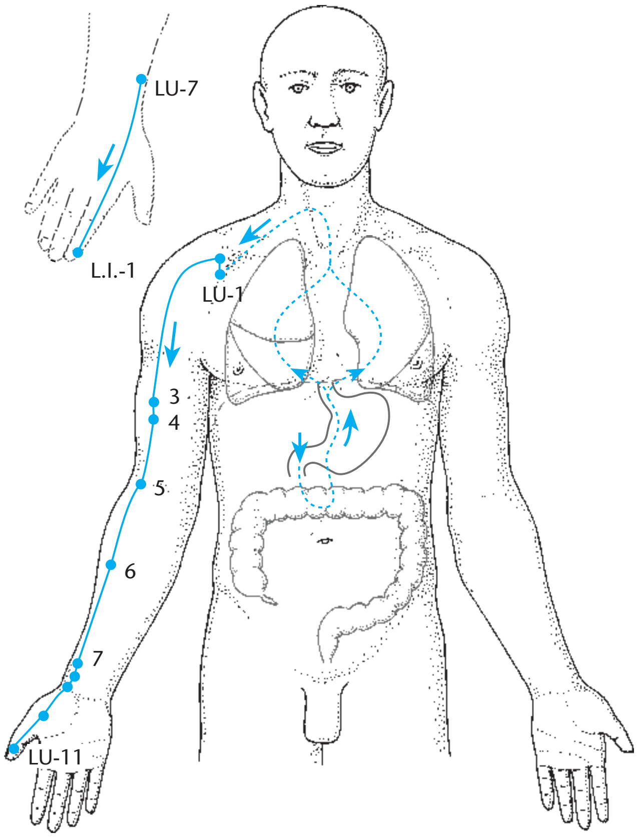
LU-7
L.I.-1
LU-1
3
4
5
6
7
LU-11

Pathway

The internal section of the Lung primary channel originates in the Middle Burner in the region of the Stomach (*wei*)

- ➡ descends to connect with the Large Intestine (*dachang*), its paired *fu*-Organ
- ➡ curves upwards again to the Stomach (*wei*)
- ➡ penetrates the diaphragm
- ➡ enters the Lung (*fei*), its pertaining *zang*-Organ
- ➡ ascends to the throat region
- ➡ and curves towards the lateral thoracic region.

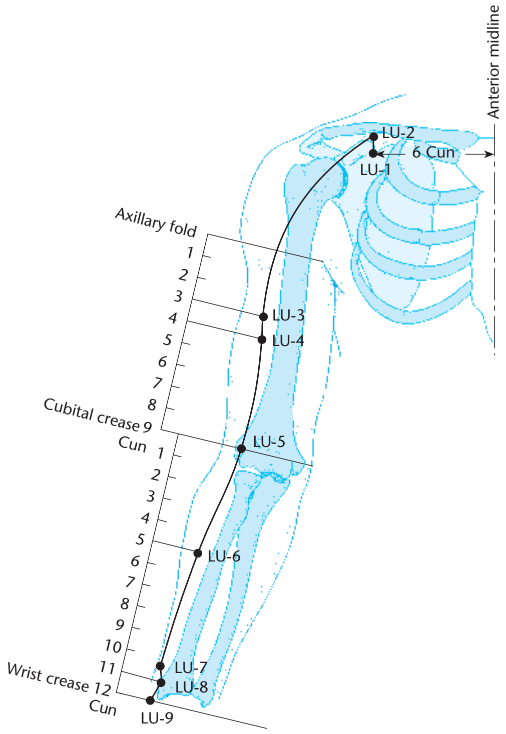
Anterior midline
LU-2
6 Cun
LU-1
Axillary fold
1
2
3
4
LU-3
5
LU-4
6
7
8
Cubital crease 9
Cun
LU-5
1
2
3
4
5
LU-6
6
7
8
9
10
11
LU-7
Wrist crease 12
LU-8
Cun
LU-9

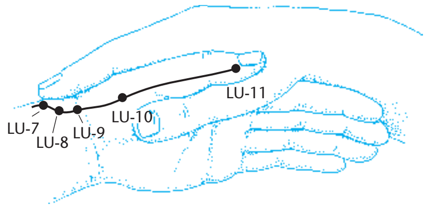
LU-11
LU-10
LU-7
LU-8
LU-9

The Lung primary channel emerges at LU-1 (*zhongfu*) on the lateral thoracic wall at the level of the first intercostal space

- ➡ travels along the anterolateral aspect of the upper arm and forearm
- ➡ ends at LU-11 (*shaoshang*) on the thumb, at the radial aspect of the corner of the nail.

A branch separates from the primary channel at LU-7 (at the styloid process of the radius) and travels along the radial side of the index finger to its tip, where it connects with → **L.I.-1** (*shangyang*) (hand Yin–Yang connection of the first great circuit).

Clinical importance (→ 1.2)

Exterior (***biao***) signs and symptoms: Fever, aversion to cold, nasal blockage, headache, pain in the region of the thorax, shoulder and back, pain with cold sensations along the channel
Interior (***li***) or *zangfu*-Organ signs and symptoms: Cough, asthma, wheezing, dyspnoea, shortness of breath, fullness and tightness of the chest, expectoration of phlegm, dry throat, changed colour of the urine, irritability, heat sensation in the palms, abdominal fullness and distension, diarrhoea
Excess (***shi***): Shoulder and back pain, Wind-Cold attack with spontaneous sweating, frequent urination, yawning
Deficiency (***xu***): Shoulder and back pain with aversion to cold, shortness of breath, changed colour of urine

Connections of the Lung primary channel (→ 1.2)

Connections with other channels

Large Intestine primary channel (*shou yang ming jing*)

Connection: Hand Yin–Yang connection of the first great circuit
Location: LU-7 → L.I.-1 (on the hand). A branch of the Lung primary channel separates at **LU-7** (*lieque*) and connects with the Large Intestine primary channel at **L.I.-1** (*shangyang*), according to some authors also with → **L.I.-4** (*hegu*).
Circulation: Circadian (according to the Organ clock)
Importance: Interior–Exterior relationship

Spleen primary channel (*zu tai yin jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *tai yin* (Yin axes of the first great circuit)
Location: SP-20 → LU-1 (on the thorax). A branch of the Spleen primary channel separates at → **SP-20** (*zhourong*) and connects with the Lung primary channel at → **LU-1** (*zhongfu*).
Circulation: Non-circadian (not according to the Organ clock)
Importance: Above–below relationship

Liver primary channel (*zu jue yin jing*)

Connection: Deep Yin–Yin connection
Location: LIV → LU (on the thorax). An internal branch of the Liver primary channel emerges at the Liver, passes the diaphragm and spreads inside the Lung (*fei*), forming a network with the Lung primary channel.
Circulation: Circadian (according to the Organ clock)
Importance: Closes the circle of the first great circuit of the Nutritive Qi (*ying qi*, → 1.1.4)

Ren mai

Connection: *Ren mai* → Lung primary channel
Importance: Closes the circle of the second circuit of the Nutritive Qi (*ying qi*, → 1.1.4). The Lung primary channel as well as the great circuit (circulation through the 12 primary channels) receives Qi from the small circuit (*ren mai* – *du mai*).

Connections to the *zangfu*-Organ systems

Lung (***fei***), **Large Intestine** (***dachang***), Stomach (*wei*)

4.1.2 The Lung Divergent Channel (*shou tai yin jing bie*)

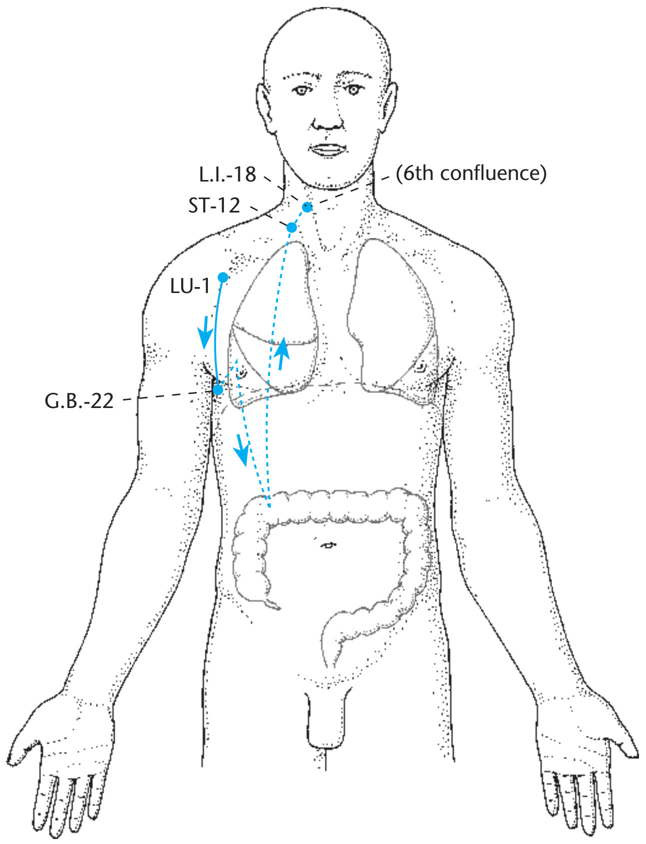

Pathway

The Lung divergent channel separates from the Lung primary channel in the axillary region

- travels anteriorly to → **G.B.-22** (*yuanye*) on the midaxillary line, 3 cun inferior to the axilla
- enters the thoracic region
- disperses in the Lung (*fei*), its pertaining *zang*-Organ
- descends to the Large Intestine (*dachang*), its paired *fu*-Organ
- ascends again and emerges in the region of the supraclavicular fossa at → **ST-12** (*quepen*)
- traverses the neck in a cranial direction and connects with the Large Intestine primary channel at → **L.I.-18** (*futu*) at one of the six *he*-confluences* (L.I./LU as 6th confluence. → 1.3).

There exist variations regarding the channel pathway due to differing interpretations of the *Ling Shu* (→ Solinas et al. 1998).

---

* According to some authors the 6th confluence is localised at → **ST-12** (*quepen*).

Clinical importance

- Strengthens the connection between the Lung and the Large Intestine (*zangfu*-Organ systems). Points on the Large Intestine primary channel can therefore be used for disorders of the Lung, and vice versa points on the Lung primary channel can treat disorders of the Large Intestine.
- Creates a network between the Lung primary channel and the neck region at the confluent point → **L.I.-18**: thus LU points such as → **LU-10** or → **LU-11** may be used to treat the neck region.

4.1.3 The Lung Sinew Channel (*shou tai yin jing jin*)

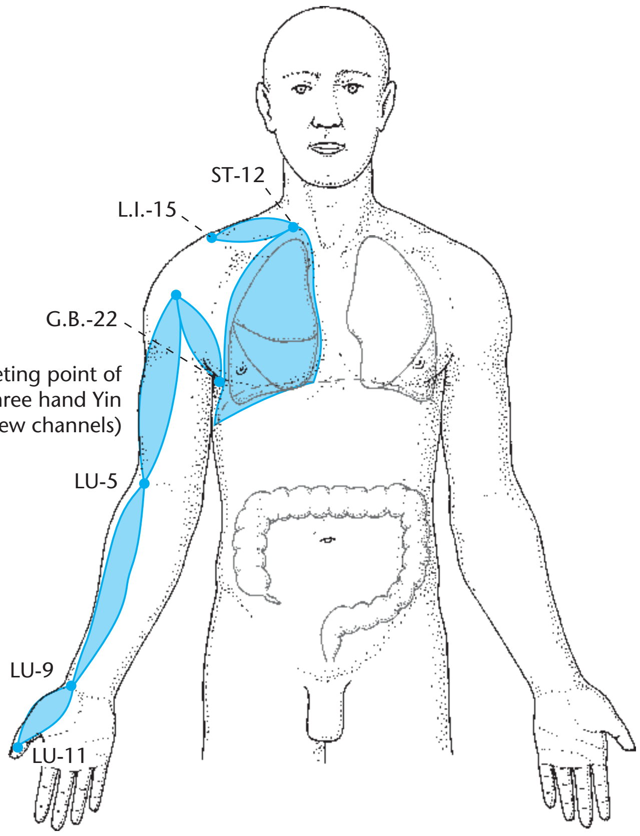

Pathway

The Lung sinew channel

➡ begins on the thumb at **LU-11** (*shaoshang*), at the radial aspect of the corner of the nail

➡ spreads along the radial aspect of the thumb and the 1st metacarpal bone as well as the thenar eminence

➡ meets **LU-9** (*taiyuan*) and binds (*jie*) at the radial wrist joint space

➡ travels proximally along the anterolateral aspect of the forearm

➡ reaches the elbow at **LU-5** (*chize*) and binds (*jie*) at the biceps tendon

➡ follows the biceps brachii muscle and the lateral portion of the deltoid muscle

➡ binds (*jie*) at the anterior shoulder region near the acromioclavicular joint

➡ continues to the axilla and meets the other hand Yin sinew channels (P, HE) at → **G.B.-22** (*yuanye*) and binds (*jie*) there

➡ continues deep to the pectoralis major muscle and enters the supraclavicular fossa at → **ST-12** (*quepen*)

➡ courses to → **L.I.-15** (*jianyu*) and returns to → **ST-12** (*quepen*)

➡ penetrates the supraclavicular fossa and disperses in the thoracic and hypochondriac region and the diaphragm (→ Fig.)

Clinical importance

Pathology: Stiffness, cramps and pain along the sinew channel. *Xi fen*: spasms and pain in the thorax and hypochondriac region. In severe cases formation of nodules below the right lateral costal region and distension along the lateral costal region.

Indication: Mainly for *bi*-syndromes (painful obstruction syndrome) along the LU channel.

The area covered by the Lung sinew channel is larger than that covered by the Lung primary channel. This explains why the indications of points on the Lung primary channel include disorders and diseases of the thoracic and hypochondriac region.

4.1.4 The Lung *luo*-Connecting Vessel System (*shou tai yin luo mai*)

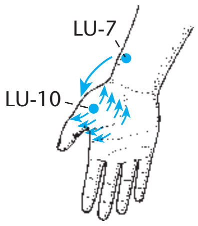

Pathway

The Lung *luo*-connecting channel system separates from the Lung primary channel at **LU-7** (*lieque*) (→ 8.1.2) and forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

➡ Horizontal divisions course to the Interiorly–Exteriorly paired Large Intestine primary channel; according to some schools (for example Van Nghi, → Appendix) they travel as a transverse Lung *luo* vessel to the *yuan*-source point → **L.I.-4** (*hegu*).

➡ A longitudinally orientated division spreads on the palm and thenar eminence.

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*)**: Hot and burning sensations in the palms and the wrist joints
- **Deficiency (*xu*)**: Yawning, shortness of breath, frequent urination, enuresis

4.1.5 Cutaneous Region (*tai yin pi bu*)

See description and figures → 1.6.

4.1.6 Points of the Lung Primary Channel (Overview)

Specific points according to function

- ***Yuan*-source point (→ 8.1.1): LU-9** (*taiyuan*) 🟦🟦
- ***Luo*-connecting point (→ 8.1.2.): LU-7** (*lieque*) 🟦🟦
- ***Xi*-cleft point (→ 8.1.3): LU-6** (*kongzui*)
- **Pertaining Back-*shu* point (→ 8.1.4): BL-13** (*feishu*) 🟦■
- **Pertaining Front-*mu* point (→ 8.1.5): LU-1** (*zhongfu*) 🟦■
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood): **LU-11** (*shaoshang*) 🟦■
  ying-spring point (Fire): **LU-10** (*yuji*)
  shu-stream point (Earth), tonification point: **LU-9** (*taiyuan*) 🟦🟦
  jing-river point (Metal), *ben* point: **LU-8** (*jingqu*) 🟦■
  he-sea point (Water), sedation point: **LU-5** (*chize*) 🟦■
- **Influential *hui*-meeting point (→ 8.1.7) of the vessels: LU-9** (*taiyuan*) 🟦🟦
- **Opening point (→ 8.1.8) of the *ren mai*: LU-7** (*lieque*) 🟦🟦
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - with the SP primary channel: **LU-1** (*zhongfu*) 🟦■
  - of other channels with the LU channel: –
- **Gao Wu Command point (→ 8.1.11) for the occiput and nape: LU-7** (*lieque*) 🟦🟦
- **Window of Heaven point (→ 8.1.12): LU-3** (*tianfu*)
- **Ma Dan Yang Heavenly Star points (→ 8.1.14): LU-7** (*lieque*) 🟦🟦
- **Sun Si Miao Ghost point (→ 8.1.15): LU-11** (*shaoshang*) 🟦■

Points according to region

- **Local points (→ 8.2.1):** Lung – **LU-1** (*zhongfu*) 🟦■; elbow – **LU-5** (*chize*) 🟦■; wrist – **LU-7** (*lieque*) 🟦🟦
- **Adjacent points (→ 8.2.1):** hands – **LU-7** (*lieque*) 🟦🟦
- **Distal points (→ 8.2.1):** occiput – **LU-7** (*lieque* 🟦🟦); throat – **LU-10** (*yuji*) – **LU-11** (*shaoshang*); Lung – **LU-5** (*chize*) 🟦■, **LU-7** (*lieque*) 🟦🟦; shoulder – **LU-7** (*lieque*) 🟦🟦

Specific points according to the course of the channel (in numeric order)

- **LU-1** (*zhongfu* 🟦■): associated Front-*mu* point (→ 8.1.5), *jiaohui*-meeting point with the SP channel (→ 8.1.10)
- **LU-3** (*tianfu*): Window of Heaven point (→ 8.1.2)
- **LU-5** (*chize*) 🟦■: *he*-sea point (Water) (→ 8.1.6); sedation point; distal point of the Lung (→ 8.2.1)
- **LU-6** (*kongzui*): *xi*-cleft point (→ 8.13)
- **LU-7** (*lieque*) 🟦🟦: *luo*-connecting point (→ 8.1.2); opening point (→ 8.1.8) of the *ren mai*; Gao Wu Command point (→ 8.1.11) of the nape and occiput; Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the occiput, shoulder and Lungs (→ 8.2.1); adjacent point for the hands (→ 8.2.1)
- **LU-8** (*jingqu*): *jing*-river point (Metal) (→ 8.1.6); *ben* point (Five Phase point)
- **LU-9** (*taiyuan*) 🟦🟦: *yuan*-source point (→ 8.1.1), hui-meeting point (→ 8.1.7) of the vessels; *shu*-stream point (Earth) (→ 8.1.6); tonification point
- **LU-10** (*yuji*): *ying*-spring point (Fire) (→ 8.1.6); distal point for the neck (→ 8.2.1)
- **LU-11** (*shaoshang*) 🟦■: Sun Si Miao Ghost point (→ 8.1.15); *jing*-well point (Wood) (→ 8.1.6); distal point for the neck

Central Residence ZHONGFU LU-1

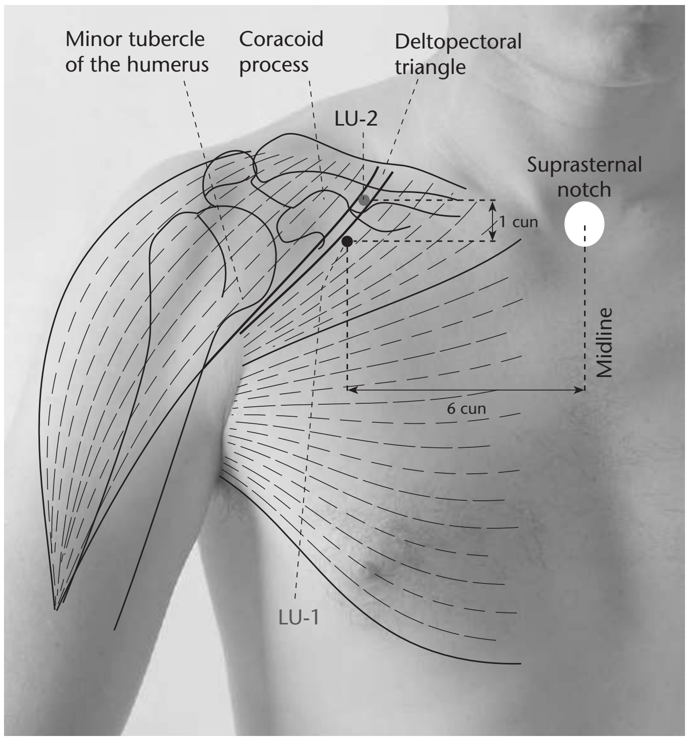
Minor tubercle of the humerus
Coracoid process
Deltopectoral triangle
LU-2
Suprasternal notch
1 cun
Midline
6 cun
LU-1

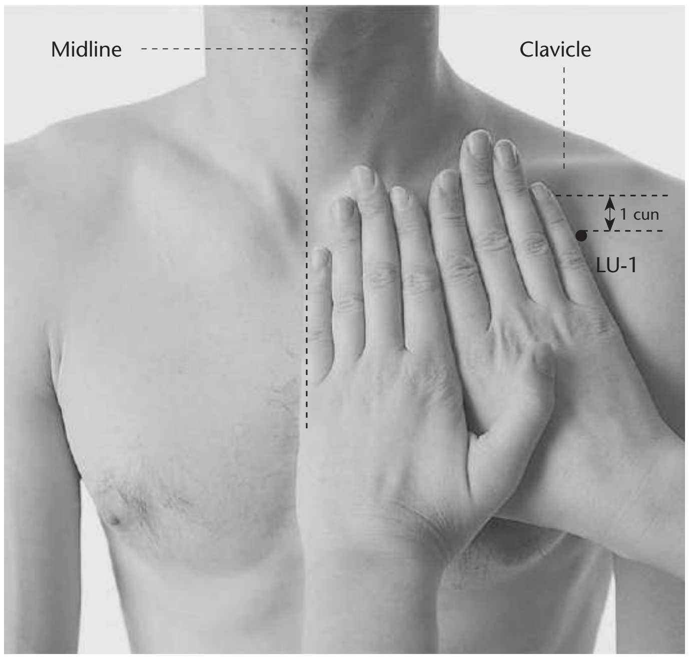
Midline
Clavicle
1 cun
LU-1

Location

6 cun lateral of the anterior midline and approximately 1 cun below → **LU-2,** slightly medial to the lower border of the coracoid process.

How to find

First, locate → **LU-2** in the deltopectoral triangle, which is bordered superiorly by the clavicle and laterally by the coracoid process (within the deltoid muscle). From there, palpate along the border of the deltoid for 1 cun and there locate **LU-1** slightly medial to the lower border of the deltoid on the thoracic wall, approximately 6 cun lateral to the midline. For orientation: when externally rotating the arm with the elbow flexed at the same time, the coracoid process will remain static, while the minor tuberculum of the humerus will follow the movement. → **SP-20** is also located 6 cun lateral to the midline on the level of the 2nd intercostal space. Located approximately on the same level are → **Ren-20,** → **KID-26** and → **ST-14**.

Needling

0.5–1 cun obliquely in a cranial (approximately 45°) and lateral direction towards the coracoid process.

Actions/Indications

- Regulates and descends the Lung Qi
- Clears Heat in the Upper Burner, transforms Phlegm
- Regulates the water passages
- Moves Qi locally and in the sinew channel

Special features

Front-*mu* point of the Lung, meeting point with the Spleen channel, entry point. Important point for disorders of the Lung.

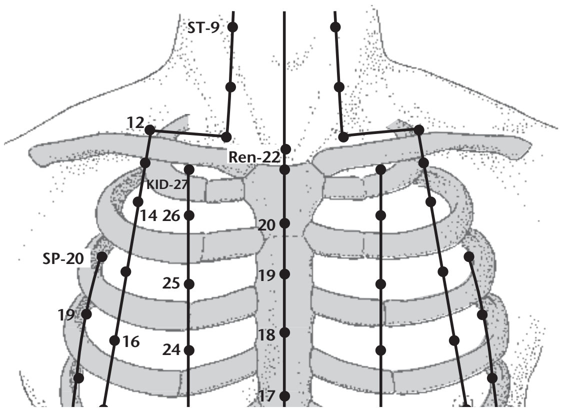
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17

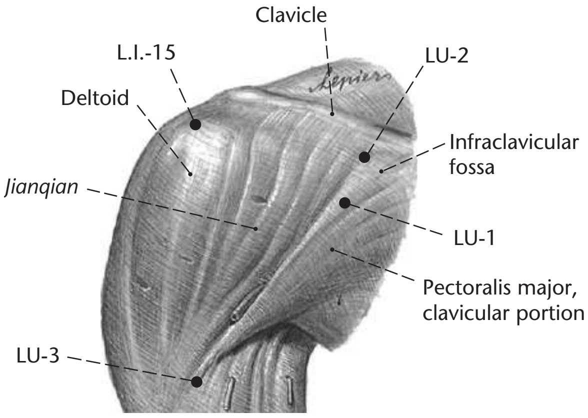
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
Jianqian
LU-1
Pectoralis major, clavicular portion
LU-3

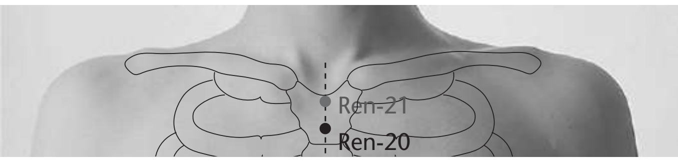
Ren-21
Ren-20

LU-2 *Cloud Gate* YUNMEN

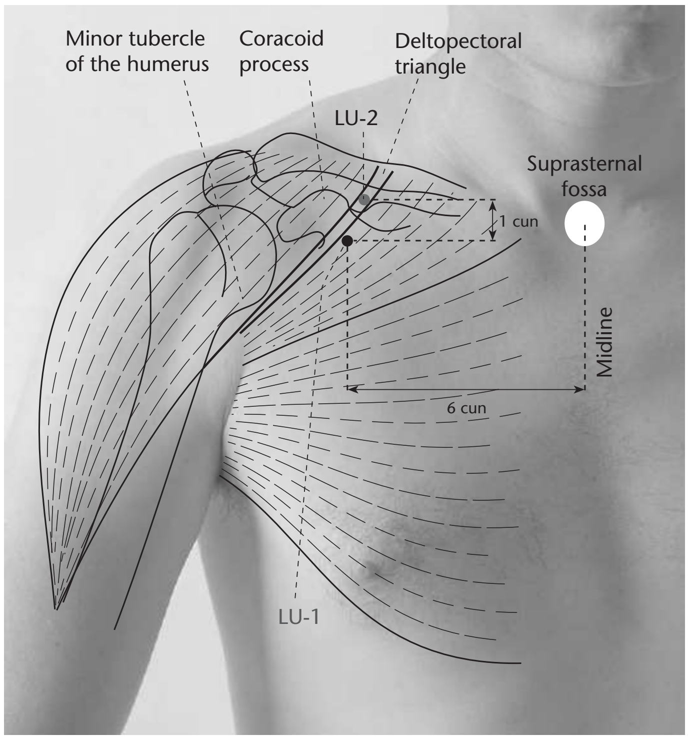
Minor tubercle of the humerus
Coracoid process
Deltopectoral triangle
LU-2
Suprasternal fossa
1 cun
Midline
6 cun
LU-1

Location

6 cun lateral to the anterior midline, below the clavicle, in the centre of the deltopectoral triangle.

How to find

Find the angle formed by the lateral lower border of the clavicle and the coracoid process (within the deltoid muscle). Locate **LU-2** at the junction of the clavicular insertions of the deltoid and pectoralis muscles (deltopectoral triangle), in the centre of a clearly palpable depression in the thoracic wall, 6 cun lateral to the anterior midline.

→ **ST-13** is also located at the lower border of the clavicle, 4 cun lateral to the anterior midline or at the midpoint of the clavicle. Also located on approximately the same level are → **KID-27** (2 cun lateral to the midline) and → **Ren-21** (on the midline). → **LU-1** is located 1 cun below and in most cases slightly lateral to **LU-2**.

Needling

Obliquely to a maximal depth of 0.8 cun. No needling in a medial direction! Risk of pneumothorax!

Actions/Indications

- Clears Lung Heat, decends the Lung Qi
- Opens the channel and sinew channel
- Clears Heat from the extremities

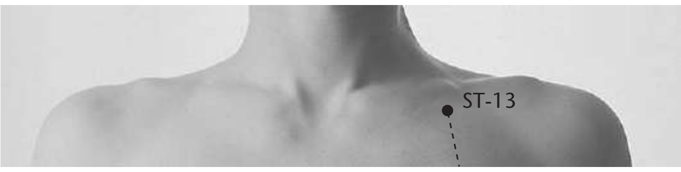

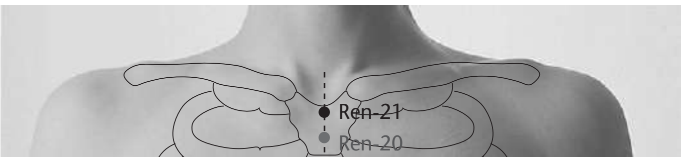
Ren-21
Ren-20

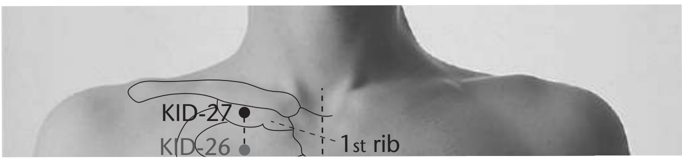
KID-27
KID-26
1st rib

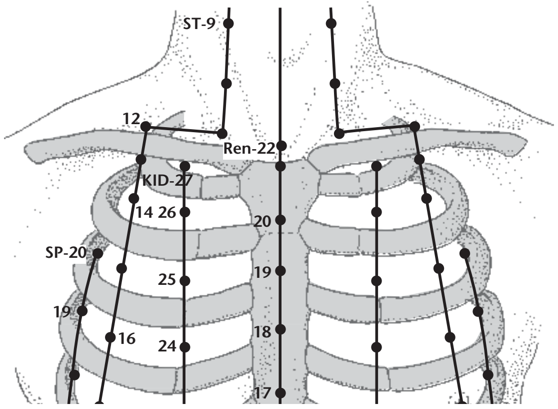
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17

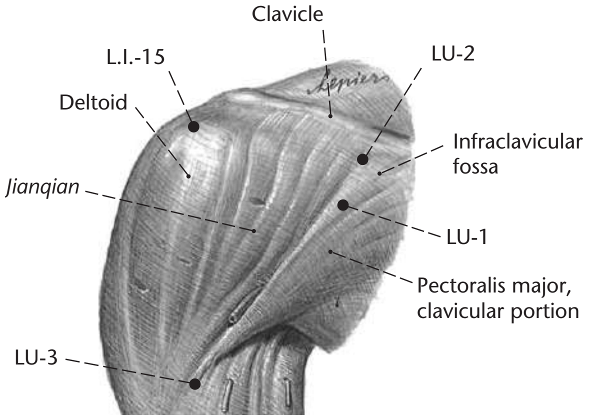
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
Jianqian
LU-1
Pectoralis major, clavicular portion
LU-3

Heavenly Residence TIANFU LU-3

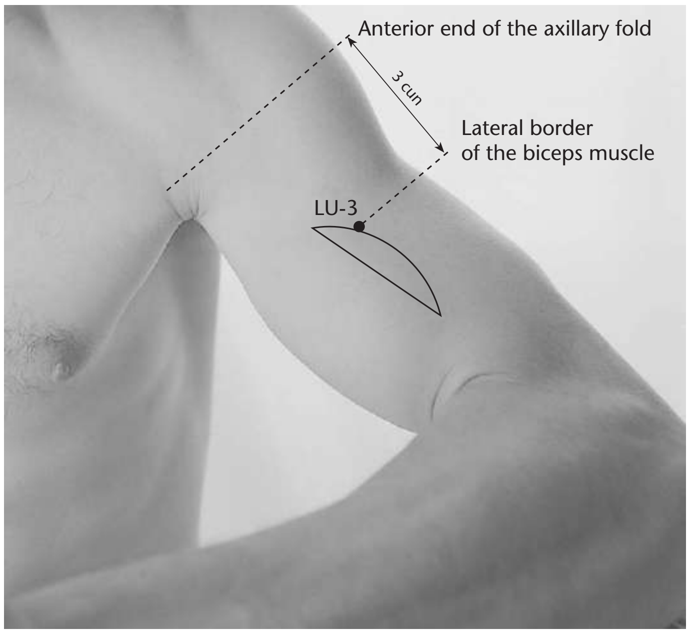
Anterior end of the axillary fold
3 cun
Lateral border
of the biceps muscle
LU-3

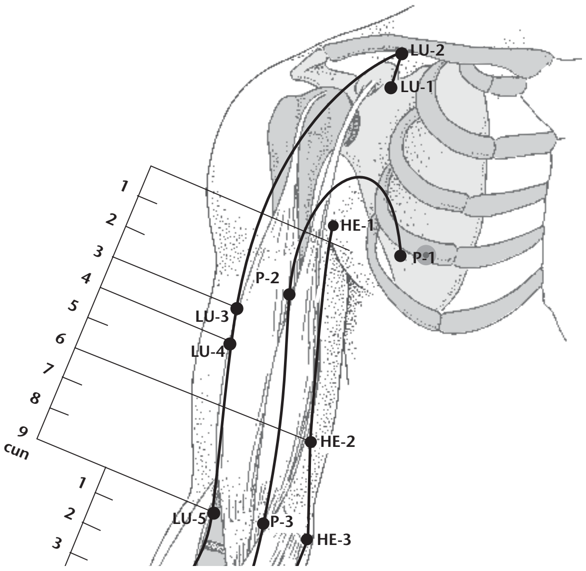
LU-2
LU-1
HE-1
P-1
P-2
LU-3
LU-4
HE-2
LU-5
P-3
HE-3
1
2
3
4
5
6
7
8
9
cun
1
2
3

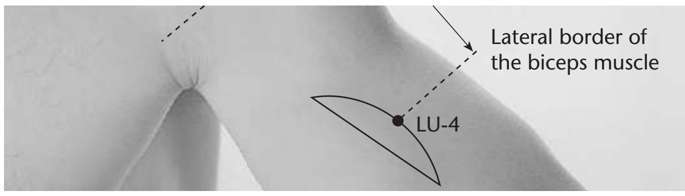
Lateral border of
the biceps muscle
LU-4

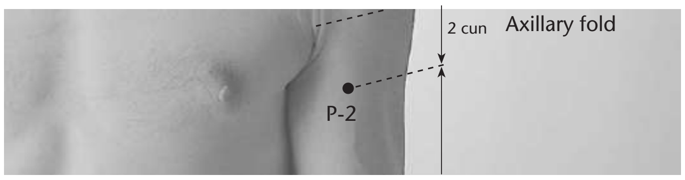
2 cun
Axillary fold
P-2

Location

On the medial aspect of the upper arm, 3 cun distal to the end of the anterior axillary fold, in the groove on the lateral border of biceps brachii.

How to find

Ask the patient to flex the biceps muscle against resistance. Locate **LU-3** 3 cun below the anterior end of the axillary fold (→ 2.2) in a groove on the lateral border of biceps brachii. Sometimes, you can feel the pulse of the brachial artery here. **Or:** Divide the distance of 9 cun (→ 2.2) between the end of the axillary fold and the cubital crease (→ **LU-5)** into thirds and locate **LU-3** a third of the distance from the axillary fold.
→ **LU-4** is located in the sulcus, 1 cun distal to **LU-3**. → **P-2** is located 2 cun below the axillary fold, between the two heads of the biceps brachii muscle.

Needling

Vertically 0.5–1 cun. Moxibustion is contraindicated according to some classics.

Actions/Indications

- Regulates and descends the Lung Qi
- Clears Lung Heat
- Cools the Blood and stops bleeding
- Calms the *po* (corporeal soul)
- Opens the channel and alleviates pain

Special features

Window of Heaven point

LU-4 *Clasping the White* XIABAI

Location

On the medial aspect of the upper arm, 4 cun distal to the end of the anterior axillary fold, in the depression on the lateral border of sulcus bicipitalis lateralis.

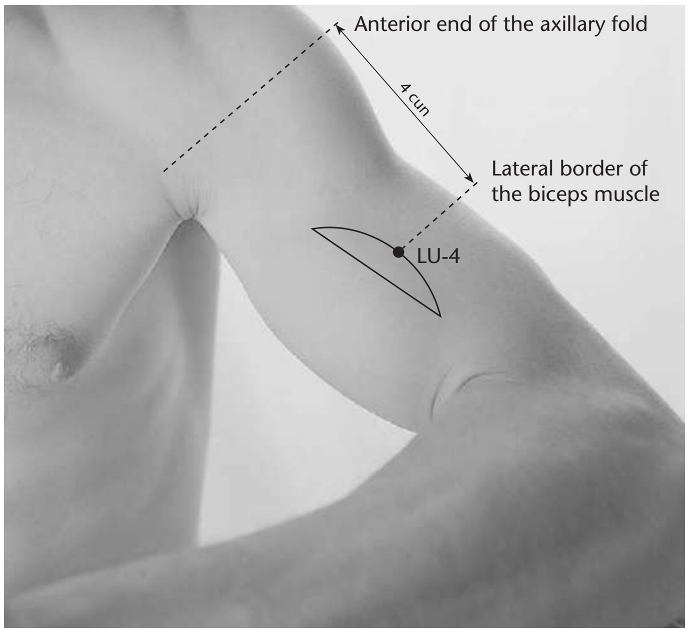
Anterior end of the axillary fold
4 cun
Lateral border of the biceps muscle
LU-4

How to find

Ask the patient to flex their biceps muscle against resistance. Locate **LU-4** 4 cun below the anterior end of the axillary fold (→ 2.2) in the depression on the lateral border of the sulcus bicipitalis muscle. Sometimes you can feel the pulse of the brachial artery here. **Or quick method:** Spreading hands technique (→ 2.3.3): place the little fingers on the axillary fold and the cubital crease (at → **LU-5**) respectively (this distance corresponds to 9 proportional cun → 2.2) and determine the midpoint of this distance by joining the thumbs. Locate **LU-4** 0.5 proximal to the midpoint in the sulcus on the lateral border of the biceps muscle. → **LU-3** is also located in the sulcus, 1 cun proximal to **LU-4**.

Needling

Vertically 0.5–1 cun

Actions/Indications

- Regulates and moves Qi and Blood in the chest
- Descends the Lung Qi
- Opens the channel and alleviates pain

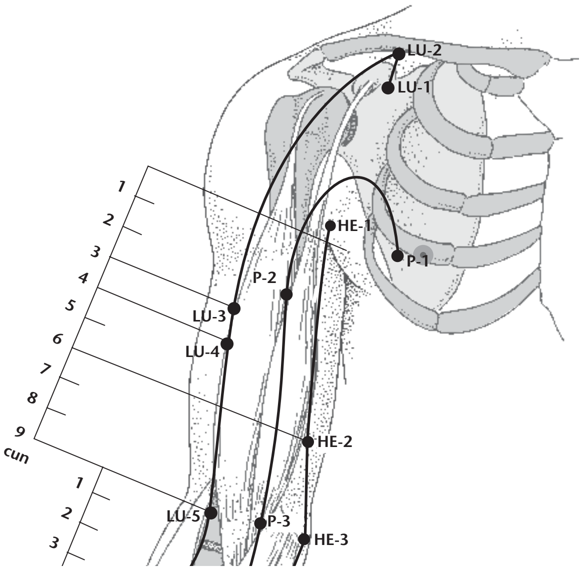
LU-2
LU-1
HE-1
P-1
P-2
LU-3
LU-4
HE-2
LU-5
P-3
HE-3
1
2
3
4
5
6
7
8
9
cun
1
2
3

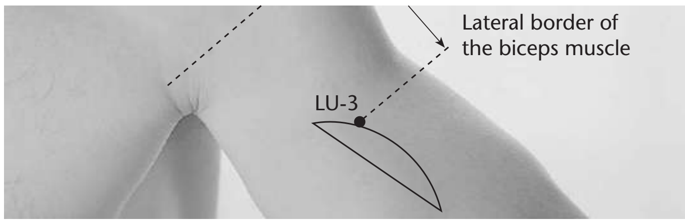
Lateral border of the biceps muscle
LU-3

Cubit Marsh CHIZE LU-5

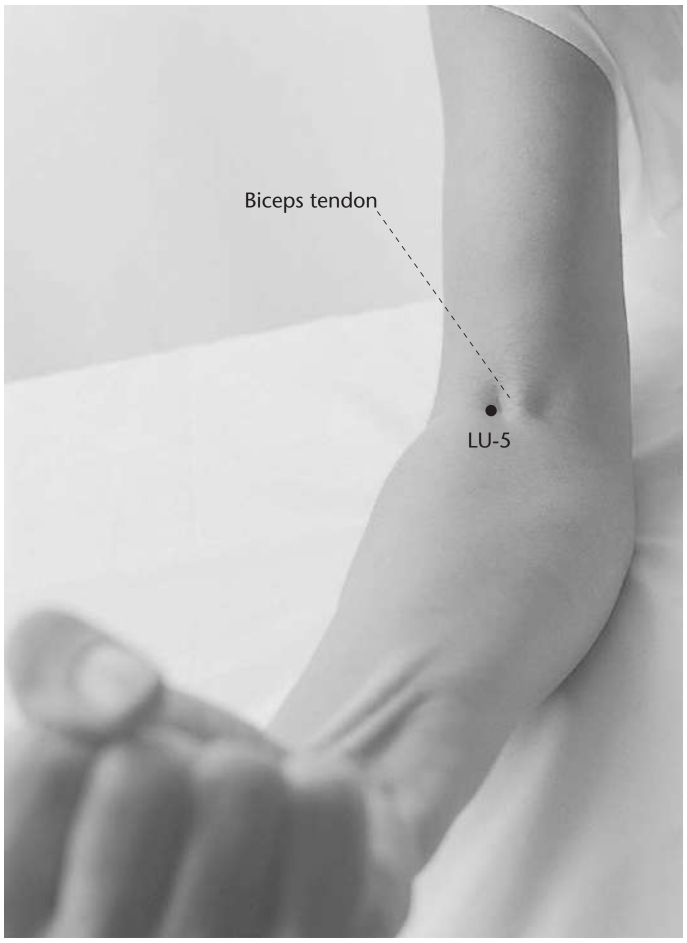
Biceps tendon
LU-5

Right arm

Location

In the cubital crease, on the radial aspect of the tendon of the biceps.

How to find

By slightly flexing the forearm, the biceps tendon becomes more visible. **LU-5** is located on its radial aspect in the cubital crease. Also located in the area of the cubital crease are → **P-3** (ulnar to the tendon), → **HE-3** (at the ulnar end of the cubital crease) and → **L.I.-11** (between the radial end of the cubital crease and the lateral epicondyle).

Needling

Vertically 0.5–1 cun. Needle with the arm slightly flexed. Use bloodletting for Heat disorders. Caution: Avoid cubital vein.

Actions/Indications

- Clears Heat in the Upper Burner
- Descends the Lung Qi
- Regulates the water passages
- Opens the channel and *luo*-connecting vessels, alleviates pain

Special features

He-sea point, Water point, sedation point, distal point for the Lung, local point for the elbow. Use bloodletting for disorders of the skin and Lung due to excess and/or Heat.

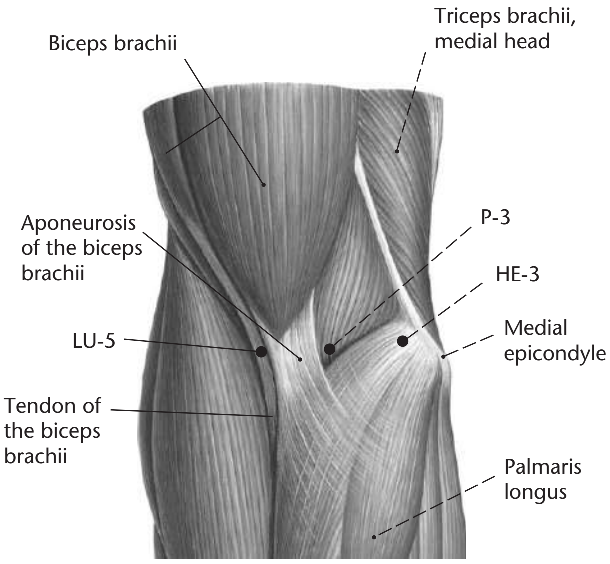
Biceps brachii
Triceps brachii, medial head
Aponeurosis of the biceps brachii
P-3
HE-3
LU-5
Medial epicondyle
Tendon of the biceps brachii
Palmaris longus

Left arm, anterior aspect

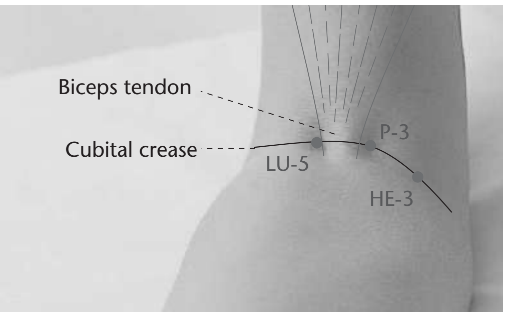
Biceps tendon
Cubital crease
P-3
LU-5
HE-3

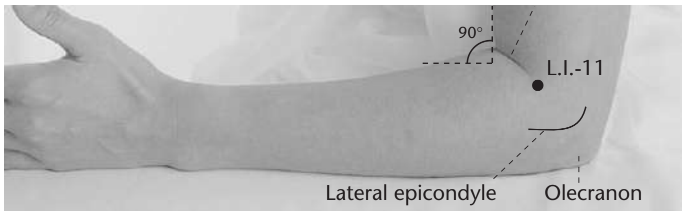
90°
L.I.-11
Lateral epicondyle
Olecranon

LU-6 *Biggest Hole* KONGZUI

Location

On the line connecting → **LU-5** (in the cubital crease) and → **LU-9** (on the wrist crease/joint space), 5 cun distal to → **LU-5** or 7 cun proximal to → **LU-9**.

How to find

The biceps tendon becomes more visible with the arm flexed. First, locate → **LU-5** radial to the biceps tendon in the cubital crease and → **LU-9** radial to the radial artery on the wrist joint space (→ 3.3.3) between the proximal row of carpals and the radius/ulna. By moving the hand, the joint space becomes more easily palpable. Next, find the midpoint of the distance between → **LU-5** and → **LU-9** using the spreading hands technique (→ 2.3.3). **LU-6** is located 1 cun proximal to the midpoint of this distance.

→ **P-4** is located more medially (between the tendons) and distally (1 cun distal to the midpoint of the distance).

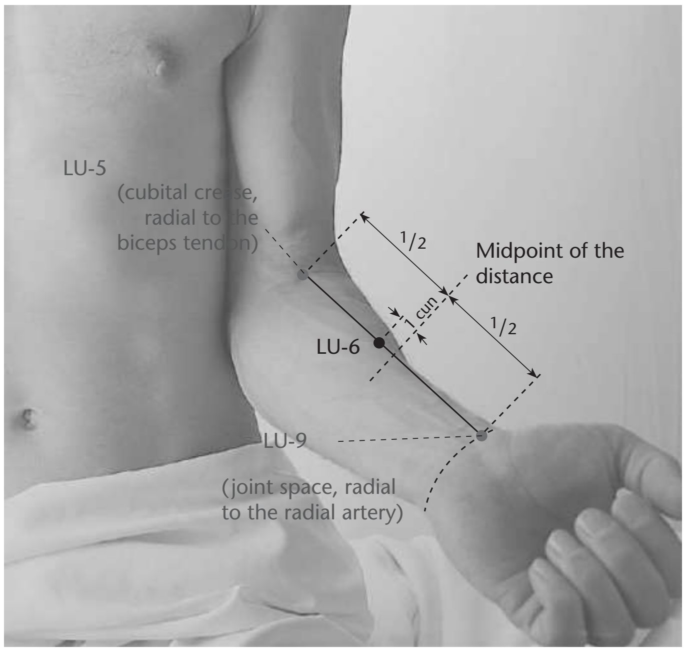
LU-5
(cubital crease,
radial to the
biceps tendon)
1/2
Midpoint of the
distance
1 cun
1/2
LU-6
LU-9
(joint space, radial
to the radial artery)

Needling

Vertically 0.5–1 cun

Actions/Indications

- Descends the Lung Qi
- Clears Lung Heat
- Cools Blood Heat, stops bleeding
- Expels Exterior pathogenic factors
- Opens the channel

Special features

Xi-cleft point

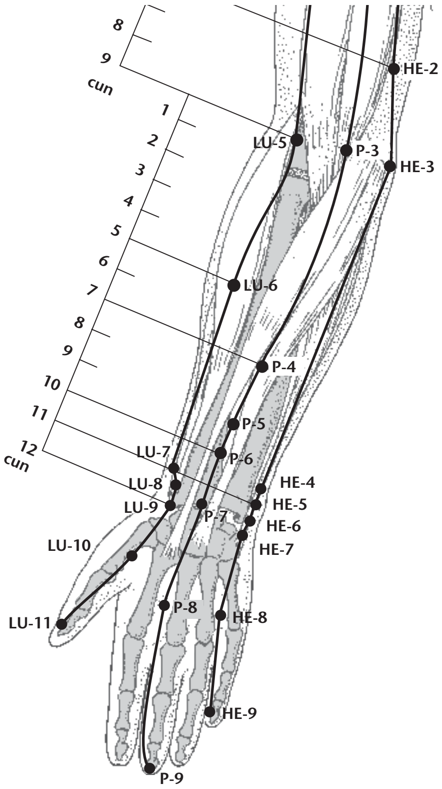
8
9
cun
1
2
3
4
5
6
7
8
9
10
11
12
cun
HE-2
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

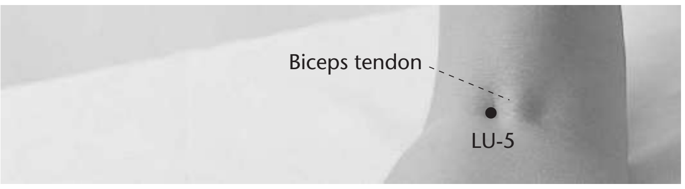
Biceps tendon
LU-5

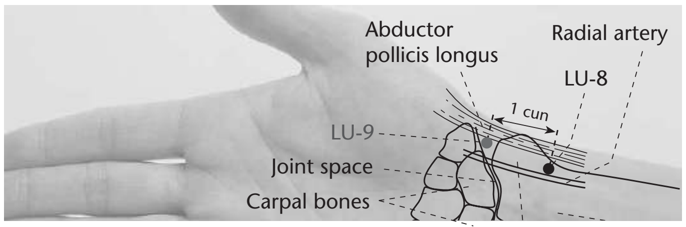
Abductor
pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones

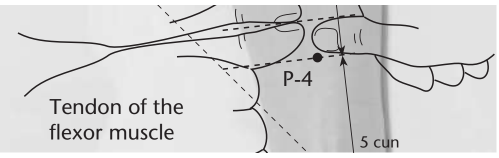
P-4
Tendon of the
flexor muscle
5 cun

Broken Sequence LIEQUE LU-7

Location

On the radial aspect of the forearm, directly above the styloid process of the radius, approximately 1.5 cun proximal to the wrist joint space (wrist crease) in a V-shaped groove.

How to find

Place the palpating forefinger on the anatomical snuffbox (➞ **L.I.-5**). From there, glide proximally over the styloid process of the radius and locate **LU-7** where the finger can palpate a cleft between two tendons (brachioradialis/abductor pollicis longus). **Or:** 'Tiger mouth grip' (➞ Fig. 2.6): spread the thumb and forefinger of both hands, crossing them so that the forefinger of one hand comes to rest on the styloid process of the other, but avoiding bending the wrist. **LU-7** is located directly below the tip of the forefinger in a V-shaped groove

Needling

Lift the skin above the styloid process by pinching it. Insert the needle 0.5–1 cun obliquely and proximally (reducing) or distally (tonifying). Caution: Cephalic vein.

Actions/Indications

- Supports the occiput and head
- Releases the Exterior, expels Wind, descends the Lung Qi
- Opens and regulates the *ren mai*
- Regulates the water passages
- Opens the channel and *luo*-connecting channel, alleviates pain

Special features

Luo-connecting point, opening point of the *ren mai*, Ma Dan Yang Heavenly Star point, Gao Wu command point of the occiput, exit point.

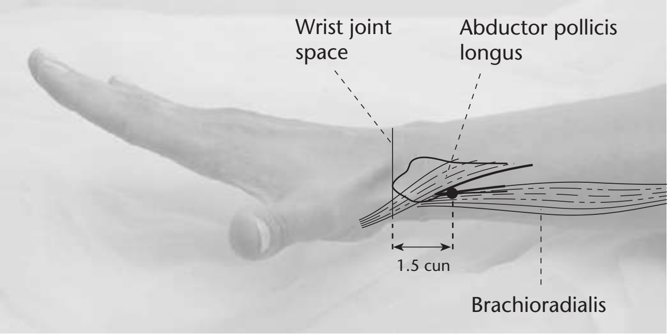
Wrist joint space
Abductor pollicis longus
1.5 cun
Brachioradialis

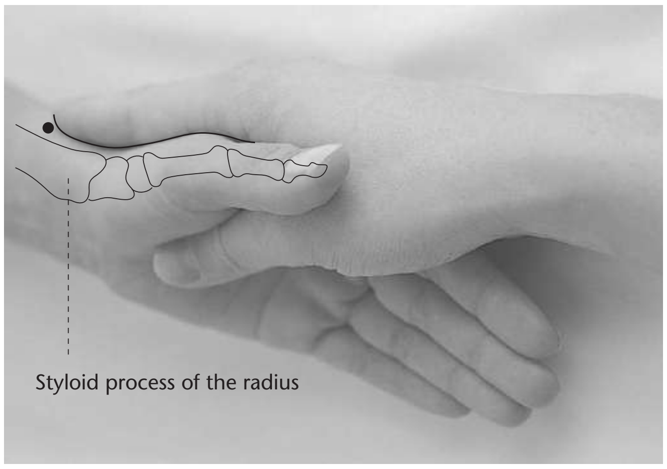
Styloid process of the radius

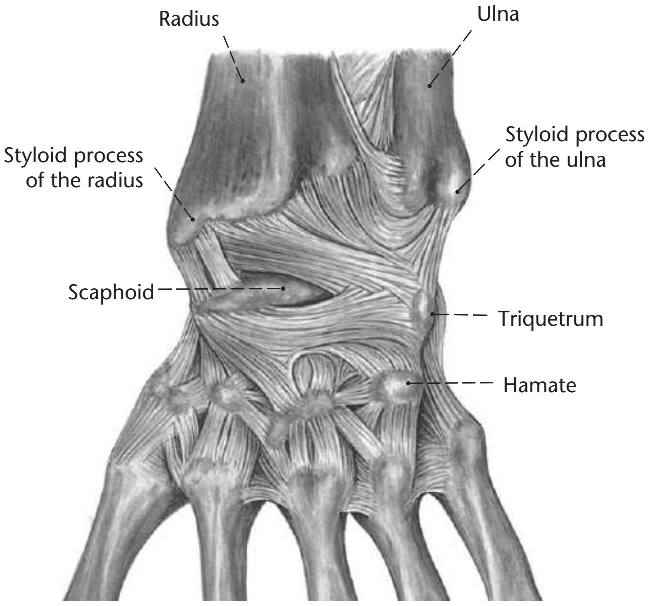
Radius
Ulna
Styloid process of the radius
Styloid process of the ulna
Scaphoid
Triquetrum
Hamate

Left hand, dorsal aspect

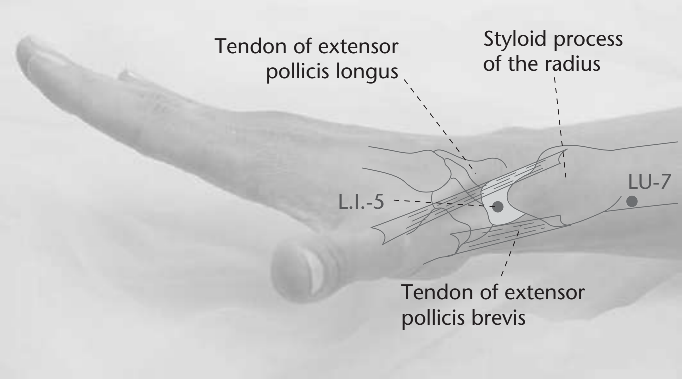
Tendon of extensor pollicis longus
Styloid process of the radius
L.I.-5
LU-7
Tendon of extensor pollicis brevis

LU-8 *Channel Gutter* JINGQU

Location

Lateral to the radial artery, 1 cun proximal to the ventral wrist joint space (most distal wrist crease).

How to find

As the location of the wrist crease varies, use the joint space (→ 3.3.3) between the proximal row of carpals and the radius/ulna for orientation. By loosely moving the hand, the joint space is easily palpable. This is the level for palpating the radial pulse (using the finger tips, not the finger pads). → **LU-9** is located on the radial side of the artery, while **LU-8** is located 1 cun proximal to → **LU-9**.

→ **L.I.-5** is located close to **LU-8**, but on the dorsal aspect of the wrist joint in the anatomical snuffbox.

Needling

0.3–0.5 cun proximally or vertically. Avoid the radial artery by gently pushing it to the side when palpating for the point. Moxibustion is contraindicated according to some classics.

Actions/Indications

- Descends the Lung Qi, alleviates cough and wheezing
- Opens the pores and expels Wind in particular
- Opens the channel, especially locally

Special features

Jing-river point, Metal point, *ben* point (Five Phases).

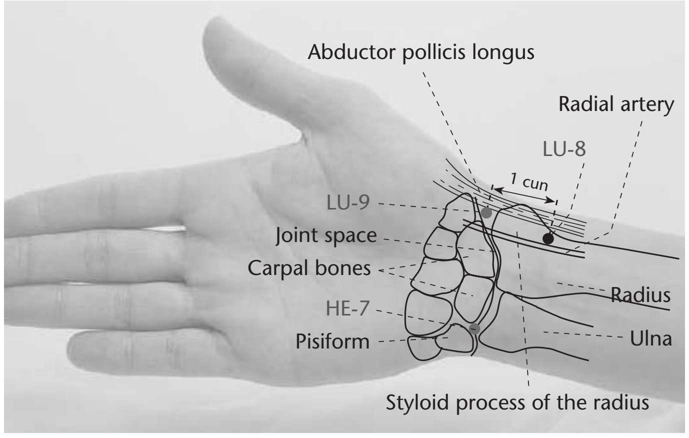
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones
HE-7
Pisiform
Radius
Ulna
Styloid process of the radius

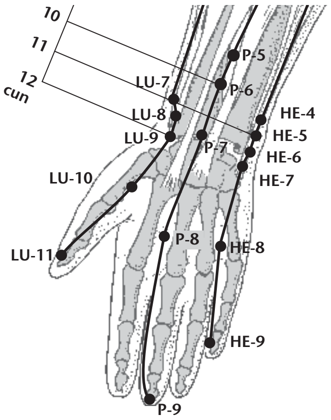
10
11
12
cun
P-5
P-6
LU-7
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9

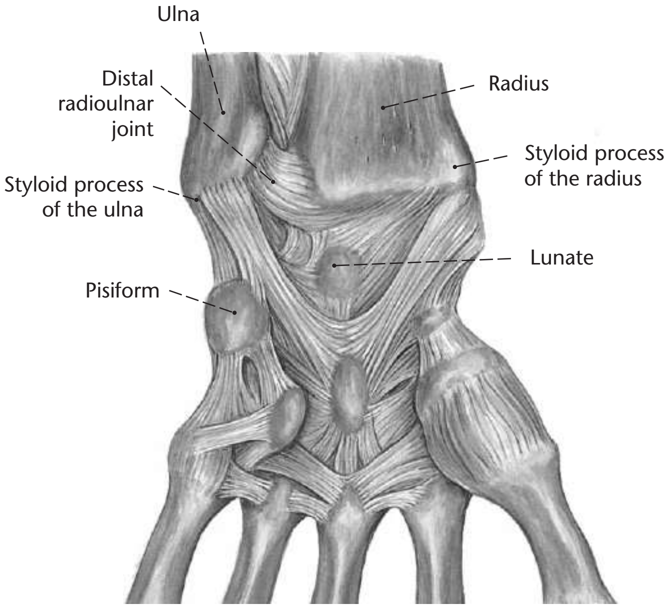
Ulna
Distal radioulnar joint
Radius
Styloid process of the radius
Styloid process of the ulna
Lunate
Pisiform

Left hand, palmar aspect

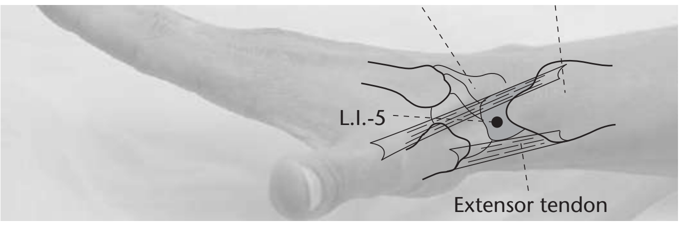
L.I.-5
Extensor tendon

Supreme Abyss TAIYUAN LU-9

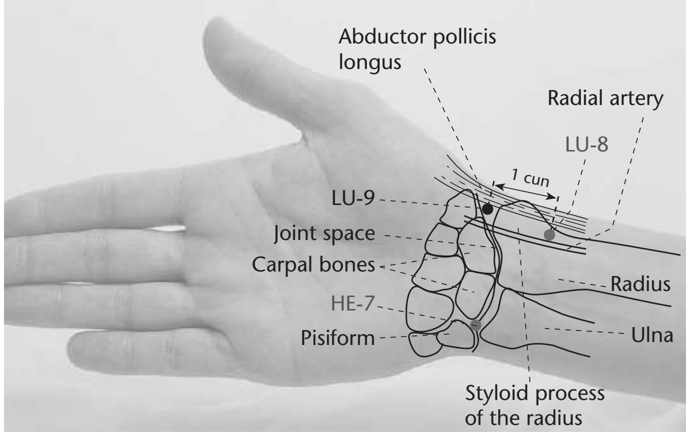
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones
Radius
HE-7
Pisiform
Ulna
Styloid process of the radius

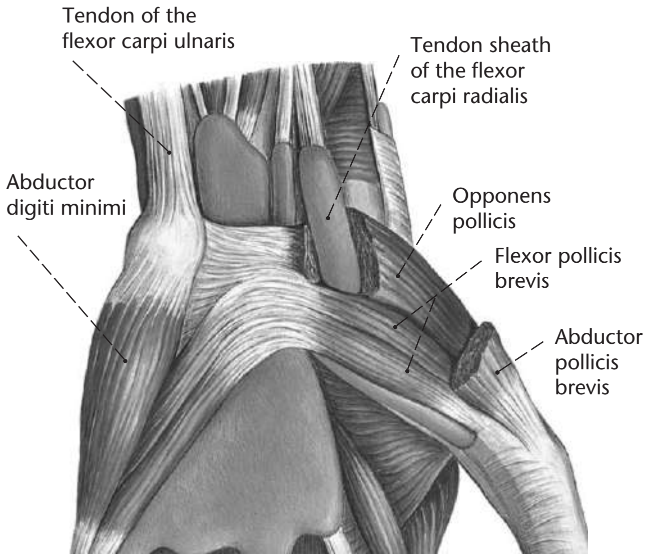
Tendon of the flexor carpi ulnaris
Tendon sheath of the flexor carpi radialis
Abductor digiti minimi
Opponens pollicis
Flexor pollicis brevis
Abductor pollicis brevis

Left hand, palmar aspect

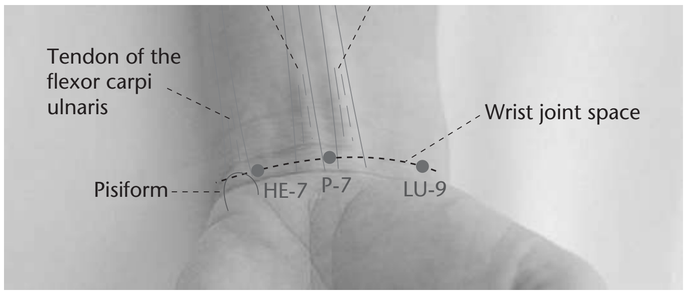
Tendon of the flexor carpi ulnaris
Wrist joint space
Pisiform
HE-7
P-7
LU-9

Location

On the ventral aspect of the wrist, at the level of the wrist joint space (most distal wrist crease), on the radial aspect of the radial artery and ulnar to the tendon of the abductor pollicis longus muscle.

How to find

As the location of the wrist crease varies, the ventral joint space (→ 3.3.3) between the proximal row of carpals and the radius/ulna is a more reliable landmark for orientation. By loosely moving the hand, the joint space is easily palpable. Palpate the radial artery (pulsation) on the radial side and locate **LU-9** on its radial aspect. Located on the same level are → **P-7** (between two tendons) and → **HE-7** (radial to the insertion of the tendon of the flexor carpi ulnaris on the pisiform bone).

Needling

Vertically 0.2–0.5 cun. Caution: Avoid radial artery. If the needle is pulsating (desired effect), no additional stimulation!

Actions/Indications

- Strengthens the Lung, transforms Phlegm, descends the Lung Qi
- Regulates and harmonises the vessels (relationship between *zong qi* and Blood circulation)
- Opens the channel, alleviates pain

Special features

Yuan-source point, *shu*-stream point, Earth point, tonification point, *hui*-meeting point of the vessels. Important point for tonifying the Lung Qi and Yin, especially with chronic disorders.

LU-10 *Fish Border* YUJI

Location

At the midpoint of the palmar border of the 1st metacarpal bone.

How to find

With the thumb relaxed, palpate the ‘belly’ of the thenar eminence from palmar to lateral (radial), then palpate the first metacarpal bone. **LU-10** is located at the midpoint of its palmar ‘border’.
→ **L.I.-4** is located on the dorsal aspect of the hand, close to the 2nd metacarpal bone.

Needling

Vertically 0.5–1 cun. Caution: Painful!

Actions/Indications

- Regulates the Lung Qi
- Cools Blood Heat
- Expels pathogenic factors from the throat
- Descends rebellious Qi
- Harmonises the Stomach and Heart

Special features

Ying-spring point, Fire point, important distal point for sore throat caused by Wind-Heat.

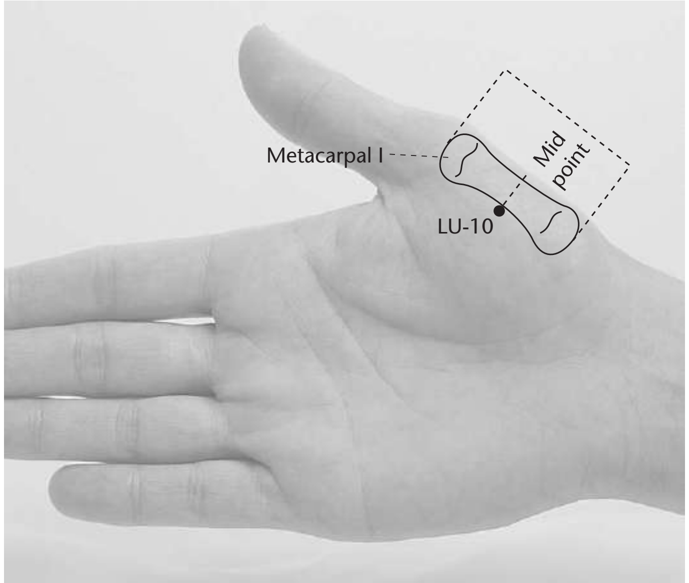
Metacarpal I
Mid point
LU-10

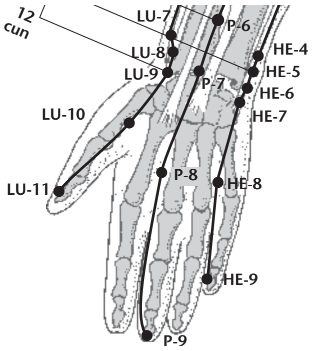
12 cun
LU-7
P-6
LU-8
LU-9
HE-4
HE-5
P-7
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

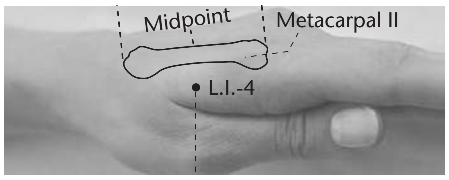
Midpoint
Metacarpal II
L.I.-4

Lesser Shang SHAOSHANG **LU-11**

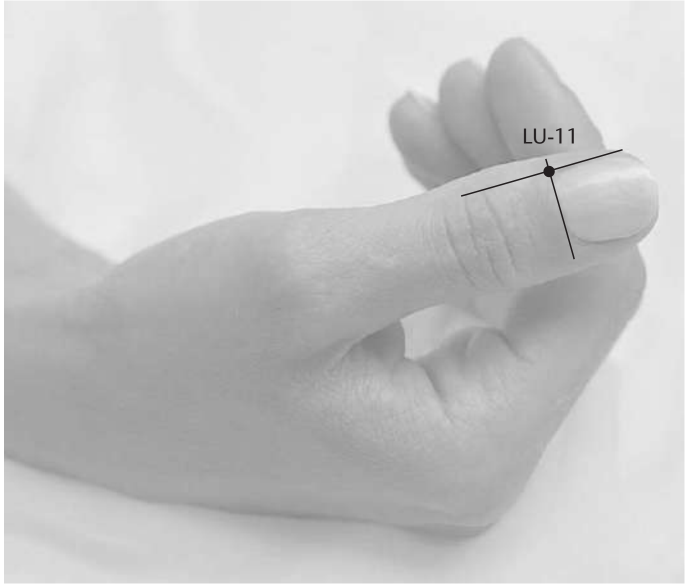

Location

On the thumb, 0.1 cun from the radial corner of the nail.

How to find

The point is located at the junction of two tangents running along the proximal and radial border of the thumb nail, approximately 0.1 cun from the actual border of the nail.

Needling

0.1–0.2 cun vertically or obliquely in a proximal direction or prick to bleed. Avoid needling the perionychium.

Actions/Indications

- Frees the senses
- Regulates the Lung Qi
- Clears Heat (especially from the throat)
- Opens the channel

Special features

Jing-well point, Wood point, Sun Si Miao Ghost point, important distal point for sore throat caused by Wind-Heat.

4.2 The Large Intestine Channel System – Hand *yangming* (*shou yang ming jing luo*)

4.2.1 The Large Intestine Primary Channel (*shou yang ming jing*)

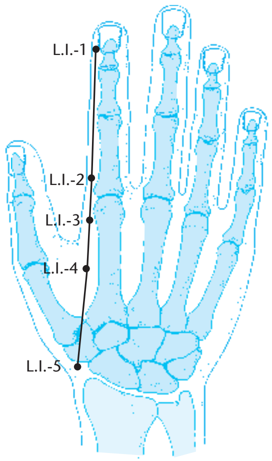
L.I.-1
L.I.-2
L.I.-3
L.I.-4
L.I.-5

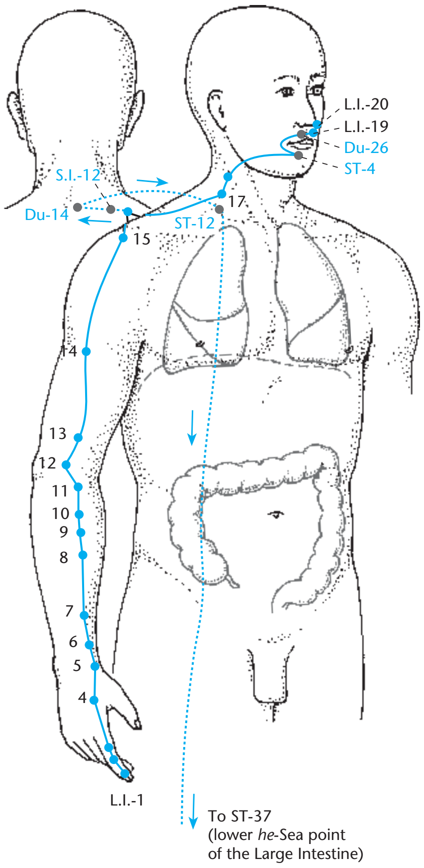
L.I.-20
L.I.-19
Du-26
ST-4
S.I.-12
Du-14
ST-12
17
15
14
13
12
11
10
9
8
7
6
5
4
L.I.-1
To ST-37
(lower *he*-Sea point
of the Large Intestine)

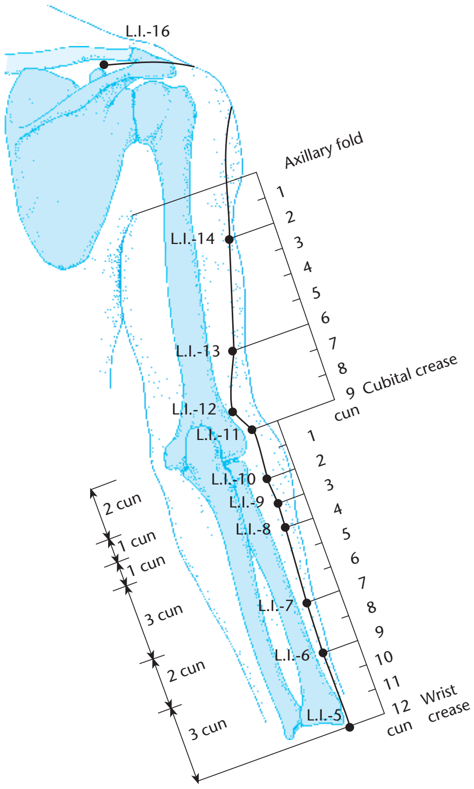
L.I.-16
Axillary fold
1
2
3
4
5
6
7
8
9 Cubital crease
cun
L.I.-14
L.I.-13
L.I.-12
L.I.-11
L.I.-10
L.I.-9
L.I.-8
L.I.-7
L.I.-6
L.I.-5
2 cun
1 cun
1 cun
3 cun
2 cun
3 cun
1
2
3
4
5
6
7
8
9
10
11
12 Wrist
cun crease

Pathway

The **external** part of the Large Intestine primary channel begins on the index finger at ➞ **L.I.-1** (*shangyang*, at the radial aspect of the corner of the nail). This point is reached by a branch which separates from the Lung primary channel at ➞ **LU-7** (*lieque*; hand Yin–Yang connection of the first great circuit).

The Large Intestine primary channel runs between the first and second metacarpal bones,

- ➡ then runs proximally along the radial aspect of the forearm towards the lateral aspect of the elbow
- ➡ travels in a proximal direction along the lateral aspect of the upper arm towards the shoulder
- ➡ crosses the Small Intestine primary channel at ➞ **S.I.-12** (*bingfeng*) in the centre of the suprascapular fossa
- ➡ continues to ➞ **Du-14** (*dazhui*) inferior to the spinous process of the 7th cervical vertebra, where it meets with the other five Yang primary channels
- ➡ crosses the neck and enters the supraclavicular fossa at ➞ **ST-12** (*quepen*).

Here the **internal** pathway branches off and connects with the Lung (*fei*), its paired *zang*-Organ, penetrates the diaphragm and enters the Large Intestine (*dachang*), its pertaining *fu*-Organ. From the Large Intestine (according to some authors: the Large Intestine primary channel) an **internal** branch runs in an inferior direction to ➞ **ST-37** (*shangjuxu*) on the lateral aspect of the lower leg, the lower *he*-sea point of the Large Intestine.

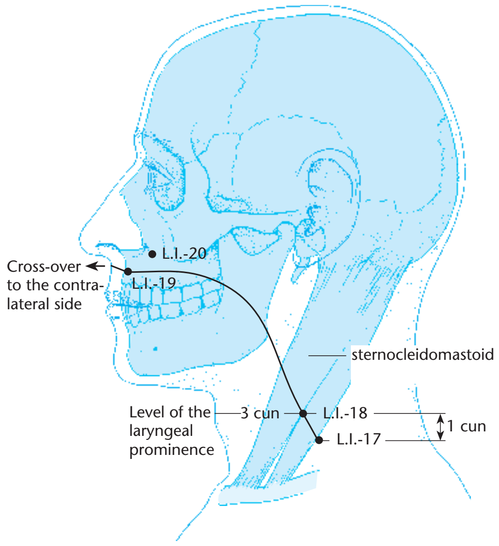
L.I.-20
Cross-over to the contra-lateral side
L.I.-19
sternocleidomastoid
Level of the laryngeal prominence
3 cun
L.I.-18
L.I.-17
1 cun

The **external** branch ascends from the supraclavicular fossa, travels along the lateral aspect of the neck and traverses the cheek.

- ➡ From here another **internal** branch enters the lower gums, where it extends further.
- ➡ The **external** course of the primary channel passes → **ST-4** (*dicang*) and curves around the mouth to the upper lip, crossing to the contralateral side of the body at → **Du-26** (*renzhong*), located at the philtrum. On the contralateral side of the body, it reaches → **L.I.-19** (*kouheliao*) and terminates **contralaterally** at → **L.I.-20** (*yingxiang*) on the side of the nose.

Note: Deadman et al. (1998) mention → **Ren-24** (*chengjiang*) as additional meeting point (variant → figure). They also mention → **G.B.-5** (*xuanlu*), → **G.B.-6** (*xuanli*) and → **G.B.-14** (*yangbai*) as classic meeting points with the Large Intestine channel but these are not normally shown in illustrations of the channel.

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Fever, dryness of the mouth, thirst, a sore and swollen throat, nosebleed, toothache, red, painful eyes, stiff fingers, painful cold or painful hot swellings of the upper arm and shoulder

Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain in the lower abdomen, borborygmus, diarrhoea or yellow stools with mucus, shortness of breath, dyspnoea

Excess (*shi*): Heat sensations along the channel

Deficiency (*xu*): Cold sensations and chills along the channel

Connections of the Large Intestine primary channel (→ 1.2)

With other channels

Lung primary channel (*shou tai yin jing*)

Connection: Hand Yin–Yang connection of the first great circuit

Location: LU-7 → L.I.-1 (on the hand). A branch of the Lung primary channel connects with the Large Intestine primary channel at **L.I.-1** (*shangyang*)

Circulation: Circadian (according to the Organ clock)

Importance: Exterior–Interior relationship

Stomach primary channel (*zu yang ming jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *yangming* (Yang axes of the first great circuit)

Location: L.I.-20 → (BL-1) → ST-1 (on the head)

Circulation: Circadian (according to the Organ clock)

Importance: Above–below relationship

With the *zangfu*-Organ systems

Lung (*fei*), **Large Intestine (*dachang*)**, Stomach (*wei*)

4.2.2 The Large Intestine Divergent Channel (*shou yang ming jing bie*)

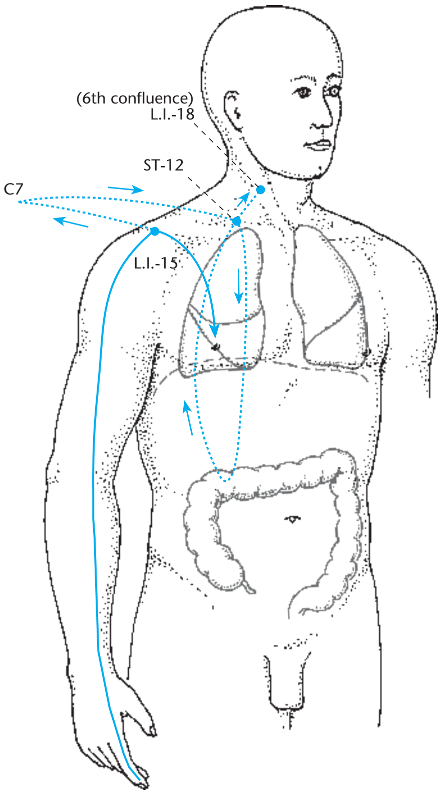

Pathway

The Large Intestine divergent channel separates from the Large Intestine primary channel on the hand,

➡ travels along the arm to the shoulder to → **L.I.-15** (*jianyu*)

➡ from **L.I.-15** a **branch** runs to the thorax and branches out into the breasts

➡ **another branch** runs to the cervical spine at the 7th cervical vertebra, returns from there to the supraclavicular fossa, descends to the Large Intestine (*dachang*), ascends to the Lung (*fei*) and **emerges** to the Exterior at → **ST-12** (*quepen*).*

It then travels along the anterolateral aspect of the neck and connects with the Large Intestine primary channel and the Lung divergent channel at → **L.I.-18** (*futu*) at one of the six *he*-confluences† (here: L.I./LU as 6th confluence → 1.3).

Clinical importance

- Strengthens the connection between the Lung and the Large Intestine (*zangfu*-Organs). Points on the Large Intestine primary channel can therefore be used for disorders of the Lung, and vice versa points on the Lung primary channel can treat disorders of the Large Intestine.
- A branch of the Large Intestine divergent channel spreads (from → **L.I.-15**) across the thorax and the breasts, thus extending the actions of the points on the Large Intestine primary channel to these regions, for example for the treatment of mastitis, myalgias and thoracic pain.

---

* Some authors describe the divergent channel as travelling directly from the spine to the Large Intestine, without meeting → **ST-12** (variant → Fig.).

† According to some authors the 6th *he*-Confluence is located at → **ST-12** (*quepen*).

4.2.3 The Large Intestine Sinew Channel (*shou yang ming jing jin*)

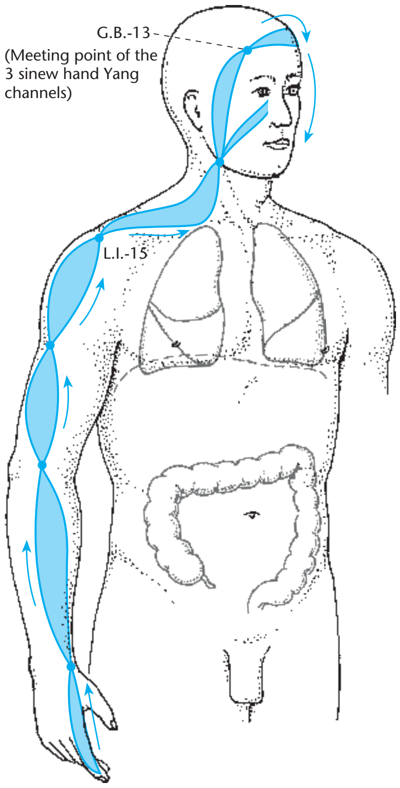

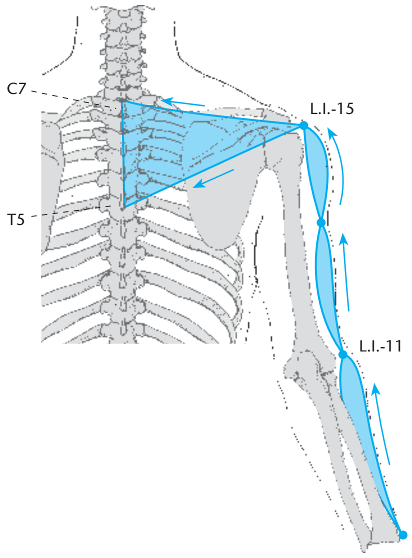

Pathway

The Large Intestine sinew channel

- begins on the index finger at → **L.I.-1** (*shangyang*)
- runs along the index finger and the 2nd metacarpal bone towards the wrist, where it binds (*jie*) at → **L.I.-5** (*yangxi*)
- travels along the posterolateral aspect of the forearm
- reaches the elbow and binds (*jie*) at → **L.I.-11** (*quchi*)
- continues along the lateral aspect of the upper arm and the deltoid muscle to the shoulder, where it binds (*jie*) again at → **L.I.-15** (*jianyu*)
- from **L.I.-15** a branch spreads across the scapula and then attaches to the spine between the 7th cervical vertebra and the 5th thoracic vertebra (or between → **Du-14** (*dazhui*) and → **Du-11** (*shendao*).
- The **main branch** ascends from **L.I.-15**, crosses the supraclavicular fossa and continues from there along the lateral aspect of the neck towards the angle of the jaw where it binds (*jie*).

Here it forks into two branches:

- **One branch** runs across the zygomatic bone towards the lateral side of the nose
- **The other branch** ascends the lateral aspect of the face and anterior to the Small Intestine sinew channel, connects at → **G.B.-13** (*benshen*) with the other sinew Yang channels, crosses over the head and ends at the lower jaw on the opposite side.

Note: According to some schools of thought, the sinew channel only ends at → **L.I.-17** (*tianding*; Solinas et al. 1998).

Clinical importance

Pathology: cramps, pain, tension and stiffness along the Large Intestine sinew channel. Limited range of motion of the upper limbs and shoulders. Occipital pain, stiffness and limited range of motion.

Indication: Mainly for *bi*-syndrome (painful obstruction syndrome) along the Large Intestine channel. The area covered by the Large Intestine sinew channel is larger than that covered by the Large Intestine primary channel. This explains why the indications of points on the Large Intestine primary channel include disorders and diseases of the upper neck and thoracic region as well as of the whole face (for example frontal and temporal headaches).

4.2.4 The Large Intestine *luo*-Connecting Vessel System (*shou yang ming luo mai*)

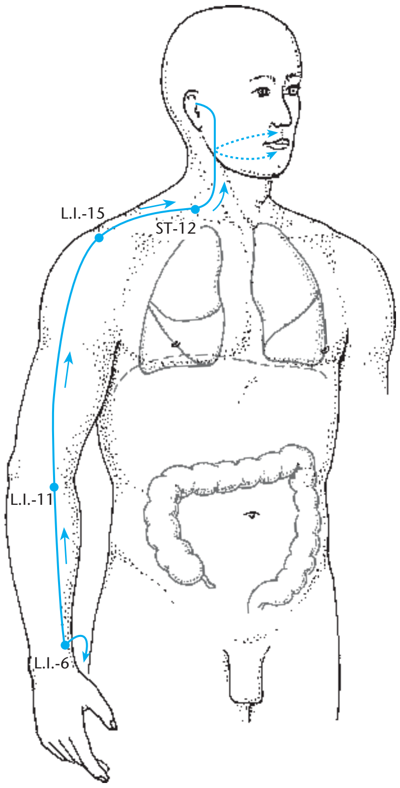

Pathway

The **Large Intestine** *luo*-connecting channel system begins at the *luo*-connecting point → **L.I.-6** (*pianli*) (→ 8.1.2), where it branches off from the primary channel and forms a three-dimensional reticular network, dividing into numerous branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

- ➡ Horizontal divisions course to the Interiorly–Exteriorly paired Lung primary channel; according to some schools (for example, Van Nghi → Appendix) they travel as a **transverse** LU *luo*-connecting vessel to the *yuan*-source point → **LU-9** (*taiyuan*).
- ➡ A **longitudinally** orientated division follows the Large Intestine primary channel to **L.I.-15** (*jianyu*), continues via the supraclavicular fossa to → **ST-12** (*quepen*) and along the neck to the mandibular angle, where it spreads. One branch runs to the teeth; another branch reaches the ear, where it communicates with all channels reaching the ear.

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*):** Disorders of the teeth and gums, disorders of the ears
- **Deficiency (*xu*):** Sensitive teeth and gums (such as sensitivity to cold), sensation of tightness in the chest and diaphragm

4.2.5 Cutaneous Region (*yang ming pi bu*)

See description and figures → 1.6

4.2.6 Points of the Large Intestine Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): L.I.-4** (*hegu*) 🟦🟦
- ***Luo*-Connecting point (→ 8.1.2.): L.I.-6** (*pianli*) 🟦■
- ***Xi*-cleft point (→ 8.1.3): L.I.-7** (*wenliu*)
- **Associated Back-*shu* point (→ 8.1.4): BL-25** (*dachangshu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): ST-25** (*tianshu*)
- **Five *shu*-transport point (→ 8.1.6):**
  jing-well point (Metal), *ben* point:. **L.-I.1** (*shangyang*)
  ying-spring point (Water), sedation point: **L.I.-2** (*erjian*)
  shu-stream point (Wood): **L.I.-3** (*sanjian*)
  jing-river point (Fire): **L.I.-5** (*yangxi*)
  he-sea point (Earth), tonification point: **L.I.-11** (*quchi*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8): –**
- **Lower *he*-sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - with the Small Intestine primary and BL channel, *yang wei mai*: **L.I.-14** (*binao*)*
  - with the *yang qiao mai*: **L.I.-15** (*jianyu*) 🟦■, **L.I.-16** (*jugu*)
  - with the primary ST channel: **L.I.-20** (*yingjiang*) 🟦■
  - of other channels with the L.I. channel: **ST-4, ST-12, S.I.-12, Du-14, Du-26, Ren-24***; according to some classics: **G.B.-5, G.B.-6, G.B.-14**
- **Gao Wu command point (→ 8.1.11) for the face and mouth: L.I.-4** (*hegu*) 🟦🟦
- **Window of Heaven point (→ 8.1.12): L.I.-18** (*futu*)
- **Point of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star points (→ 8.1.14): L.I.-4** (*hegu*) 🟦🟦, **L.I.-11** (*quchi*) 🟦🟦
- **Sun Si Miao Ghost point (→ 8.1.15): L.I.-11** (*quchi*) 🟦🟦

---

* Mentioned by only some authors

Points according to region

- **Local points (→ 8.2.1):** nose – **L.I.-20** (*yingxiang*) 🟦⬛; shoulder – **L.I.-15** (*jianyu*) 🟦⬛; elbow – **L.I.-11** (*quchi*) 🟦🟦; hand – **L.I.-5** (*yangxi*); hand and fingers – **L.I.-4** (*hegu*) 🟦🟦; fingers (numbness and pain) – **L.I.-3** (*sanjian*)
- **Regional points (→ 8.2.1):** elbow – **L.I.-10** (*shousanli*), **L.I.-13** (*shouwuli*); shoulder and elbow – **L.I.-14** (*binao*)
- **Distal points (→ 8.2.1):** forehead, eyes, mouth/teeth – **L.I.-4** (*hegu*) 🟦🟦; nose – **L.I.-3** (*sanjian*), **L.I.-4** (*hegu*) 🟦🟦; neck – **L.I.-4** (*hegu*) 🟦🟦, **L.I.-11** (*quchi*) 🟦🟦; mouth/teeth – **L.I.-2** (*erjian*), **L.I.-4** (*hegu*) 🟦🟦; shoulder, elbow, hand, fingers – **L.I.-4** (*hegu*) 🟦🟦, **L.I.-1** (*shangyang*); knee and foot – **L.I.-5** (*yangxi*); foot – **L.I.-2** (*erjian*), **L.I.-3** (*sanjian*); toes – **L.I.-4** (*hegu*) 🟦🟦

Specific points according to the course of the channel (in numerical order)

- **L.I.-1** (*shangyang*): *jing*-well point (Metal) (→ 8.1.6); *ben* point (Five Phase point); distal point for the elbow and the hand (→ 8.2.1)
- **L.I.-2** (*erjian*): *ying*-spring point (Water) (→ 8.1.6); sedation point; distal point for the mouth/teeth (→ 8.2.1); distal point for the foot (→ 8.2.1)
- **L.I.-3** (*sanjian*): *shu*-stream point (Wood) (→ 8.1.6); local point for the fingers (numbness and pain) (→ 8.2.1); distal point for the nose (→ 8.2.1); distal point for the feet (→ 8.2.1)
- **L.I.-4** (*hegu*) 🟦🟦: *yuan*-source point (→ 8.1.1); Ma Dan Yang Heavenly Star point (→ 8.1.14); Gao Wu command point (→ 8.1.11) for the face and mouth; important point for acupuncture anaesthesia; distal point for the forehead, eyes, nose, mouth/teeth, jaw, shoulder, elbow, hand, fingers (→ 8.2.1); local point for the hand and fingers (stiffness) (→ 8.2.1)
- **L.I.-5** (*yangxi*): *shu*-stream point (Fire) (→ 8.1.6); local point for the hand (→ 8.2.1); distal point for the knees and feet (→ 8.2.1)
- **L.I.-6** (*pianli*): *luo*-connecting point (→ 8.1.2)
- **L.I.-7** (*wenliu*): *xi*-cleft point (→ 8.1.3)
- **L.I.-10** (*shousanli*): adjacent point for the elbow (→ 8.2.1)
- **L.I.-11** (*qichi*) 🟦🟦: Sun Si Miao Ghost point (→ 8.1.15); *he*-sea point (Earth) (→ 8.1.6); tonification point; Ma Dan Yang Heavenly Star point; distal point for the neck (→ 8.2.1); local point for the elbow (→ 8.2.1)
- **L.I.-13** (*shouwuli*): regional point for the elbow (→ 8.2.1)
- **L.I.-14** (*binao*)*: *jiaohui*-meeting point with the S.I. and BL channel and the *yang wei mai** (→ 8.1.10); regional point for the shoulder and elbow (→ 8.2.1)
- **L.I.-15** (*jianyu*) 🟦⬛: *jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10); local point for the shoulder (→ 8.2.1)
- **L.I.-16** (*jugu*): *jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10)
- **L.I.-18** (*futu*): Window of Heaven point (→ 8.1.12)
- **L.I.-20** (*yingxiang*) 🟦⬛: local point for disorders of the nose (→ 8.2.1); *jiaohui*-meeting point with the ST channel (→ 8.1.10)

General location help

- The course of the line connecting **L.I.-5** and **L.I.-11** (12 cun) changes depending on whether the forearm is in a supinated or a pronated position. For location purposes the forearm should be in a midposition with the elbow slightly flexed.
- Since the L.I. channel crosses to the opposite side of the body at **Du-26** (nasolabial groove), locate **L.I.-19** and **L.I.-20** contralaterally.

---

* Mentioned by only a few authors

Shang Yang SHANGYANG L.I.-1

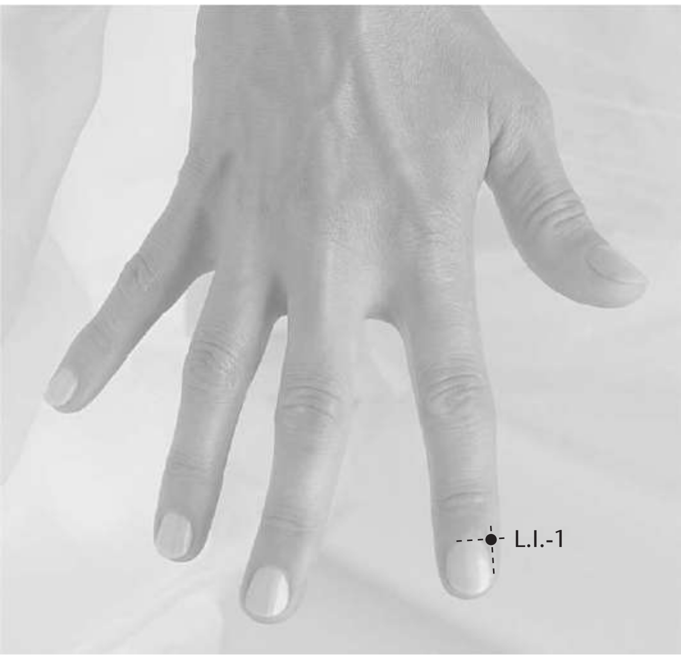

Location

On the index finger, 0.1 cun from the radial corner of the nail.

How to find

This point is located where two tangents bordering the nail of the index finger medially and proximally cross, approximately 0.1 cun from the corner of the nail.

Needling

0.2–0.3 cun vertically or obliquely in a proximal direction. Avoid needling the perionychium. For acute disorders (pain, inflammation), prick to bleed with needle, lancet or three-edged needle.

Actions/Indications

- Expels pathogenic factors and opens the channel
- Frees the senses

Special features

Jing-well point, Metal point, *ben* point (Five Phases), entry point according to some authors.

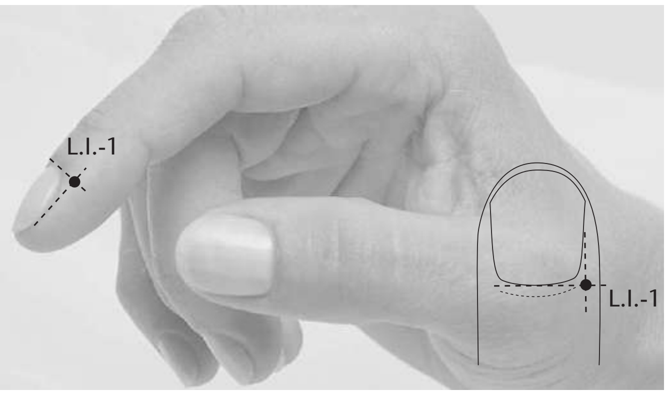

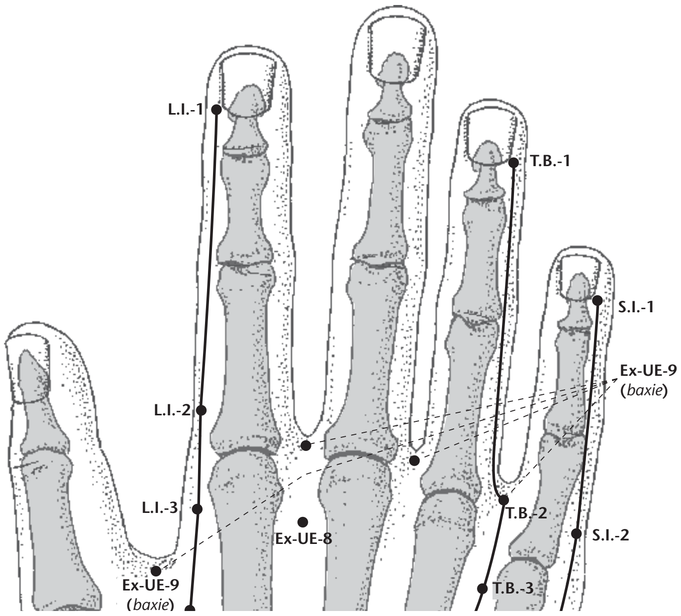

L.I.-2 *Second Space* ERJIAN

Location

On the radial aspect of the index finger, distal to the metacarpophalangeal joint, at the junction of the shaft and the basis of the proximal phalanx.

How to find

With the hand relaxed, palpate the border between the 'red to white' skin along the shaft of the second proximal phalangeal bone towards the metacarpo-phalangeal joint until the angle between the base and the shaft of the bone can be clearly felt. **L.I.-2** is located at the junction of the shaft and the base and slightly below (palmar to) the most external curvature of the bone.

→ **S.I.-2** is located in a similar position at the proximal phalanx of the little finger.

Needling

0.2–0.5 cun obliquely in a proximal direction or distal in a slightly palmar direction.

Actions/Indications

- Clears Heat and expels Wind

Special features

Ying-spring point, Water point, sedation point, distal point for the mouth and teeth

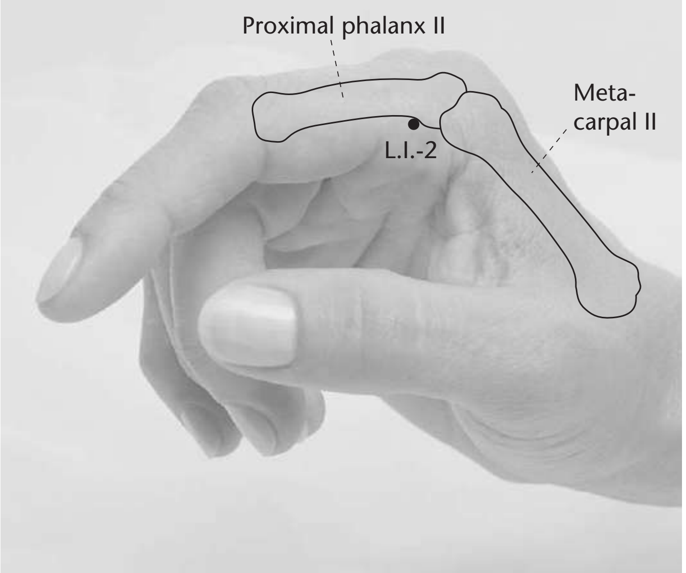

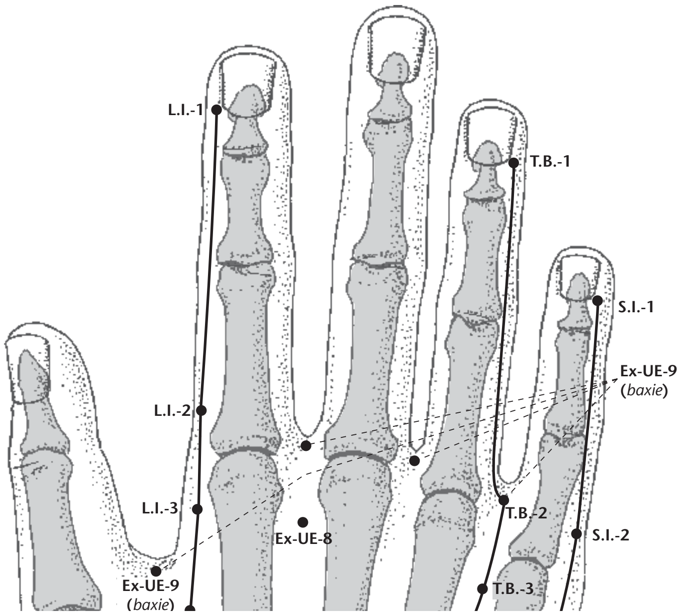

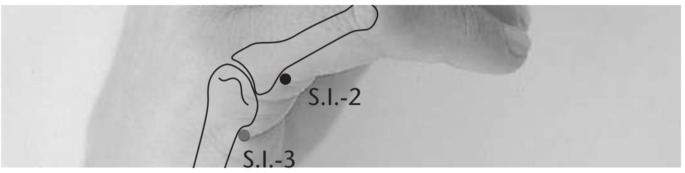

Third Space SANJIAN L.I.-3

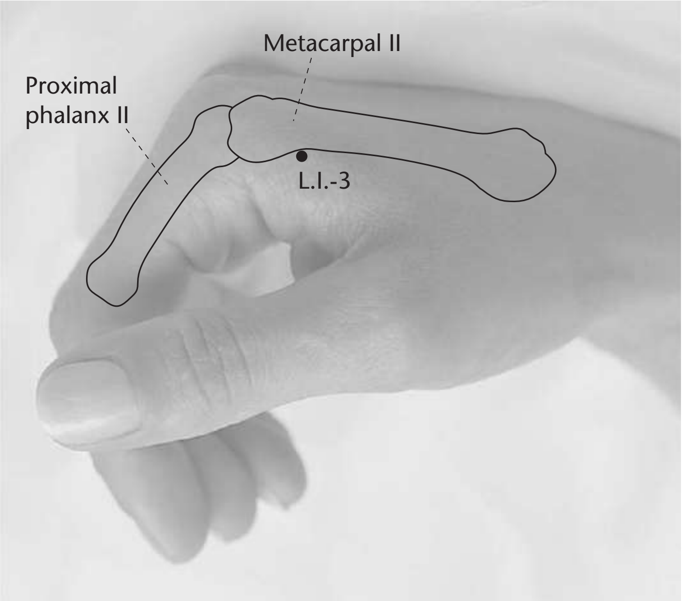
Metacarpal II
Proximal phalanx II
L.I.-3

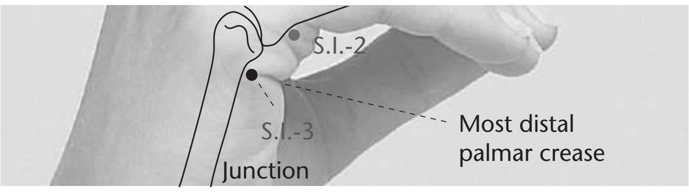
S.I.-2
S.I.-3
Most distal palmar crease
Junction

Location

On the radial aspect of the index finger, proximal to the metacarpo-phalangeal joint, at the junction of the shaft and the head of the 2nd metacarpal bone.

How to find

With the hand relaxed, palpate along the radial aspect of the 2nd metacarpal bone in a distal direction until the angle formed between the head and the shaft can be clearly felt. **L.I-3** is located at the junction of the shaft and the head and slightly below (palmar to) its most external curvature. → **S.I.-3** is located in a comparable position on the ulnar aspect of the 5th metacarpal bone (on the ulnar border of the hand). Located in a comparable location on the feet are → **SP-3** and → **BL-65**.

Needling

With the hand relaxed, 0.3–1 cun vertically directly below the lower border of the metacarpal bone towards → **S.I.-3**.

Actions/Indications

- Clears Heat and expels Wind
- Benefits the throat and teeth
- Alleviates pain

Special features

Shu-stream point, Wood point. Important point for pain in the fingers and the metacarpus.

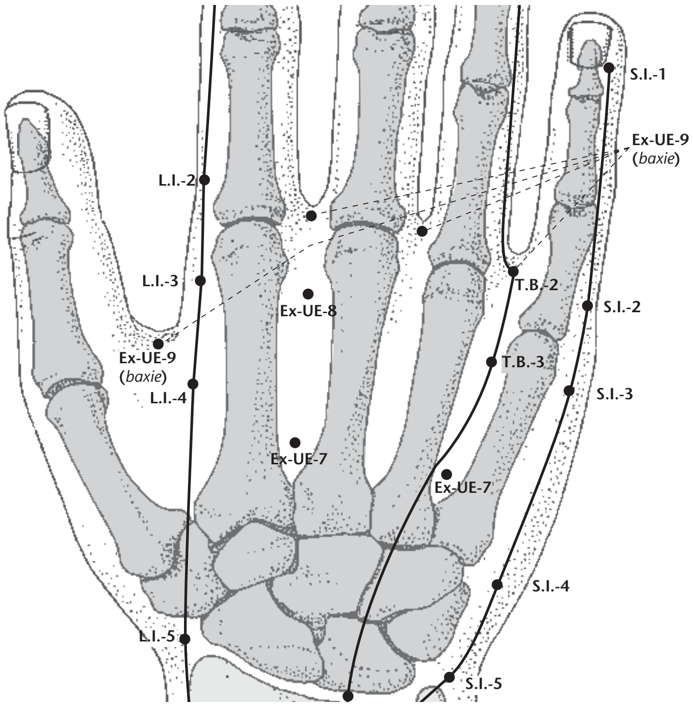
S.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5

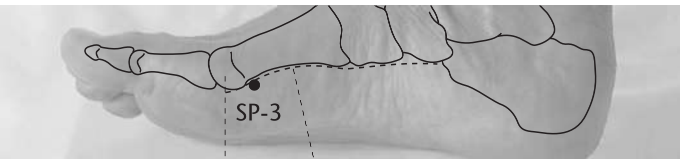

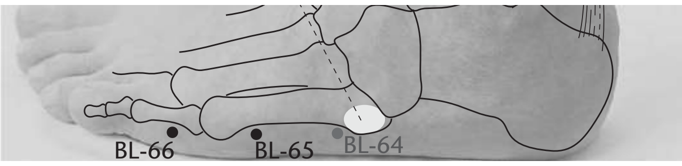
BL-66
BL-65
BL-64

L.I.-4 *Joining Valley* HEGU

Location

On the radial aspect of the hand, between the 1st and 2nd metacarpal bones, closer to the 2nd metacarpal bone and approximately at its midpoint.

How to find

When pressing the thumb and index finger together, the adductor pollicis and the interosseus dorsalis muscles will form a bulge. Needle **L.I.-4** at the highest point of the bulge and push it further towards the inferior aspect of the second metacarpal bone. **Or**: Spread the thumb and forefinger and place the distal phalanx of the thumb of the other hand on the webbed border. When flexing the thumb, its tip will point to **L.I.-4. Or**: Insertion on the angle bisector between the 1st and 2nd metacarpal bones, needling in the direction of the midpoint of the second metacarpal bone.

Needling

0.5–1 cun vertically or slightly obliquely towards the palmar aspect of the hand. Caution: Reducing needle techniques are contraindicated during pregnancy; exception: to induce labour.

Actions/Indications

- Releases the Exterior (main point!)
- Regulates the face and head
- Regulates the Defensive Qi (*wei qi*) and sweating
- Opens the channel and *luo*-connecting vessels
- Alleviates pain
- Promotes labour

Special features

Yuan-source point, Gao Wu command point for the head and mouth, Ma Dan Yang Heavenly Star point, entry point according to some authors. Most important analgesic point for the whole body. Most important single distal point for disorders of the face and sensory organs. It is the most commonly used point in clinical practice. The combined bilateral needling of **L.I.-4** and → **LIV-3** (between the 2nd and 3rd metatarsal bones) is known as *si guan* (Four Gates): they strongly regulate the Qi and Blood, stop pain and relieve spasms.

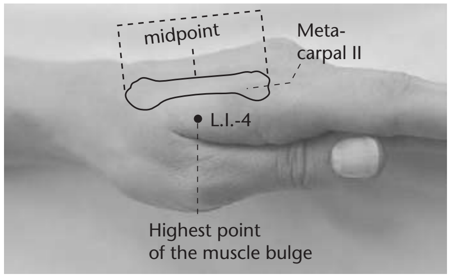
midpoint
Meta-
carpal II
L.I.-4
Highest point
of the muscle bulge

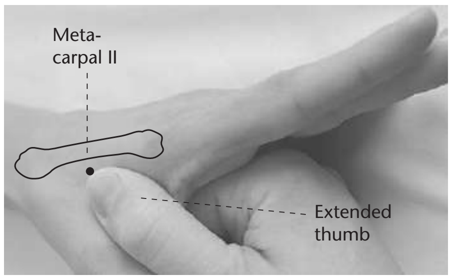
Meta-
carpal II
Extended
thumb

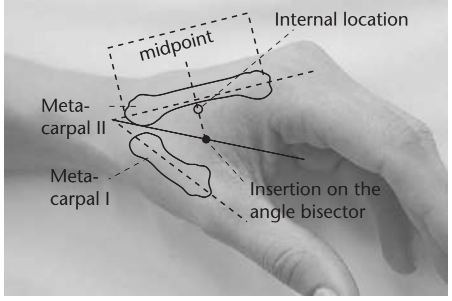
Internal location
midpoint
Meta-
carpal II
Meta-
carpal I
Insertion on the
angle bisector

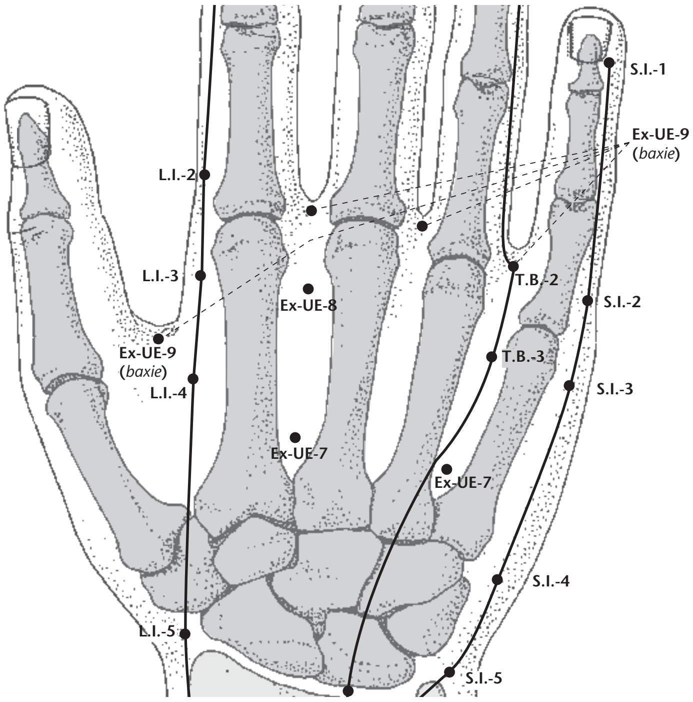
S.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
T.B.-2
S.I.-2
Ex-UE-8
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5

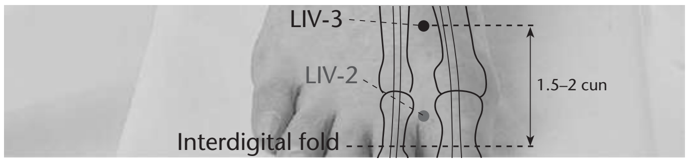
LIV-3
LIV-2
1.5–2 cun
Interdigital fold

Yang Stream YANGXI **L.I.-5**

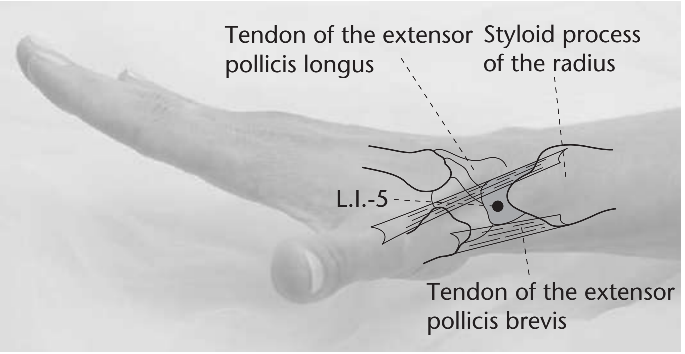
Tendon of the extensor pollicis longus
Styloid process of the radius
L.I.-5
Tendon of the extensor pollicis brevis

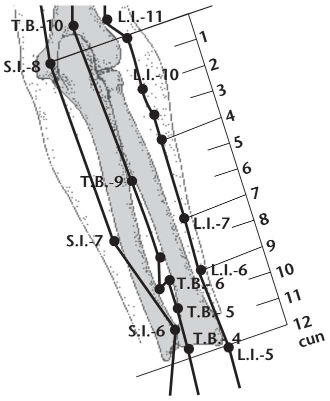
T.B.-10
L.I.-11
S.I.-8
L.I.-10
1
2
3
4
5
6
7
8
9
10
11
12
cun
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5

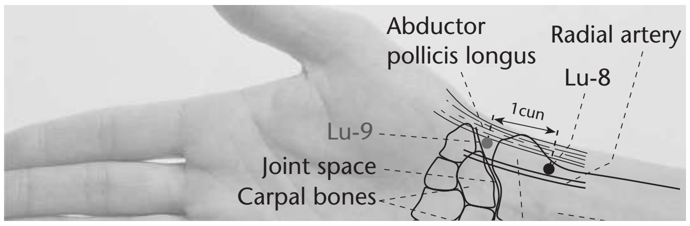
Abductor pollicis longus
Radial artery
Lu-8
1cun
Lu-9
Joint space
Carpal bones

Location

With the thumb abducted, in a depression between the tendons of the extensor pollicis longus and brevis muscles (‘anatomical snuffbox’), on the radial aspect of the wrist.

How to find

The anatomical snuffbox is a depression on the radial aspect of the wrist formed when abducting the thumb. Holding the hand in a horizontal position, it opens towards the body. **L.I.-5** is located on the wrist joint space, which becomes palpable by moving the wrist. → **LU-9** is also located on the wrist joint space and can be found by moving anteriorly from **L.I.-5** across the short tendon of the extensor pollicis muscle.

Needling

Vertically 0.5–1 cun. Caution: Avoid the superficial cephalic vein.

Actions/Indications

- Supports the wrist joint
- Clears Fire and expels Wind
- Clears Fire from the *yangming* (L.I., ST) and calms the *shen*

Special features

Jing-river point, Fire point. Important local point for disorders of the wrist, together with **T.B.-4, S.I.-5, HE-7, P-7** and **LU-9**.

L.I.-6 *Veering Passage* PIANLI

Location

3 cun proximal to → **L.I.-5** (in the centre of the anatomical snuffbox), on the line connecting → **L.I.-5** and → **L.I.-11**, between the abductor pollicis longus and the extensor pollicis brevis muscles, at the level of the junction between the tendon and the muscle.

How to find

Please note: The course of the line connecting → **L.I.-5** and → **L.I.-11** depends on the position of the forearm. For location purposes, a midposition of the forearm with flexed elbow is recommended. To locate **L.I.-6**, measure 3 cun (1 handsbreath) from → **L.I.-5** (centre of the anatomical snuffbox) on the connecting line. **Or**: 'Tiger mouth grip' (→ Fig. 2.6): Spread thumb and index finger of both hands, crossing them so that the forefinger of one hand comes to rest on the styloid process of the other, but avoiding bending the wrist. Place the middle finger against the forefinger: **L.I.-6** is located directly below the tip of the middle finger.

Located at the same level (3 cun proximal to the wrist joint space) on the dorsal side of the forearm are → **T.B.-6** (in the depression between the radius and the muscle) and → **T.B.-7** (in the depression between the ulna and the muscle) as well as → **P-5** (on the ventral side of the forearm between the tendons).

Needling

Obliquely or transversely (subcutaneously) 0.5–1 cun

Actions/Indications

- Expels Wind, clears Heat
- Regulates and opens the water passages
- Opens the channel and *luo* vessels

Special features

Luo-connecting point. Major point of the L.I. channel for regulating the water passages.

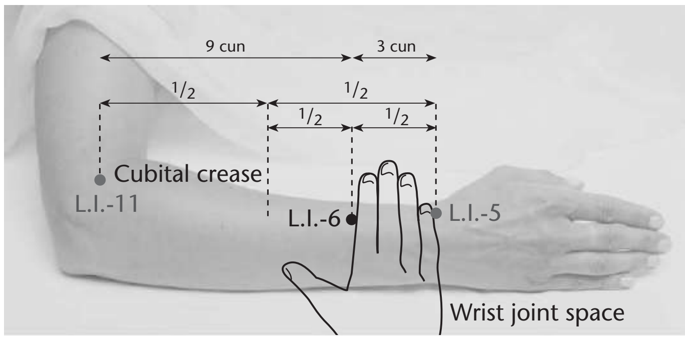

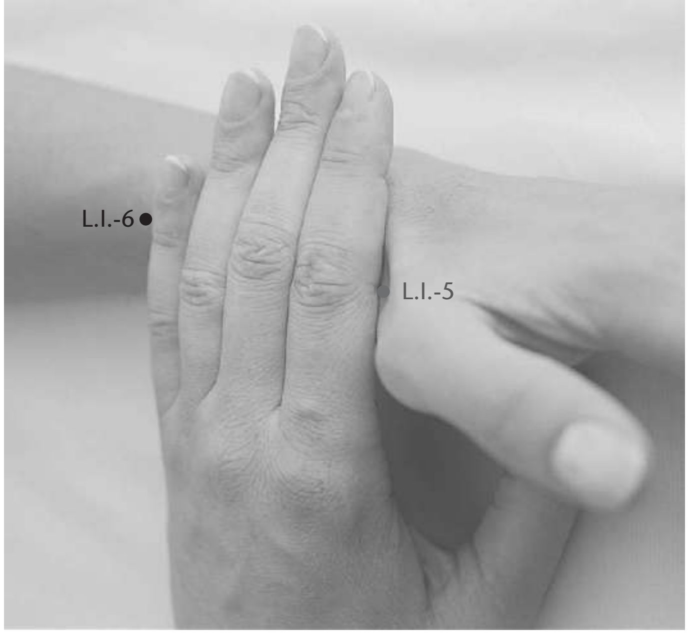

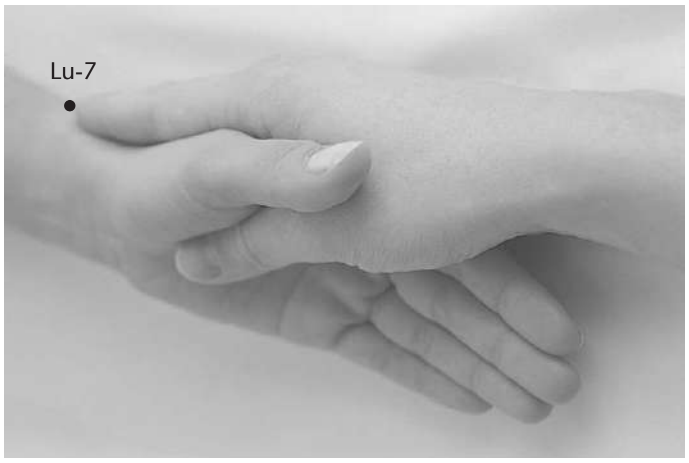

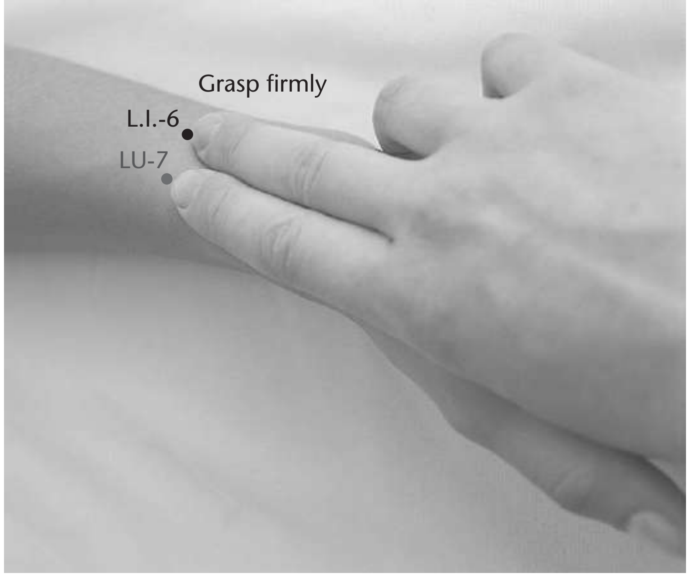

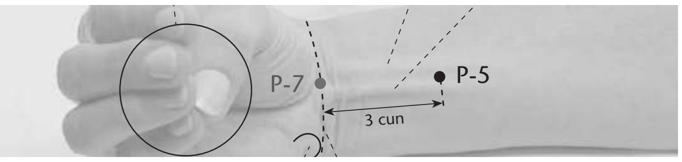

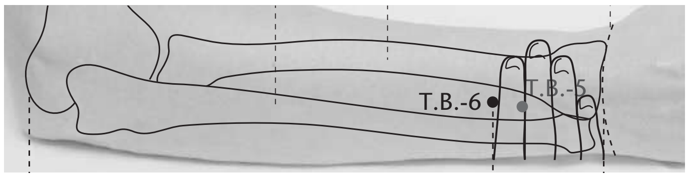

Warm Flow WENLIU L.I.-7

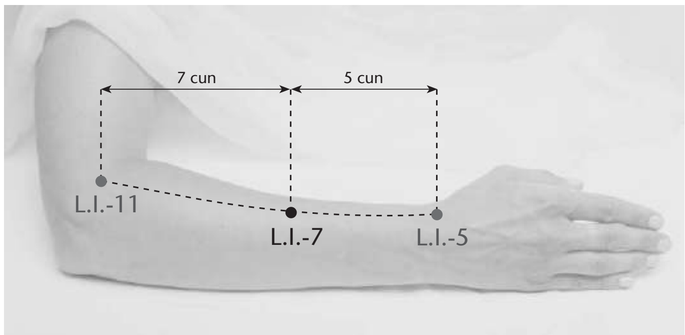
7 cun
5 cun
L.I.-11
L.I.-7
L.I.-5

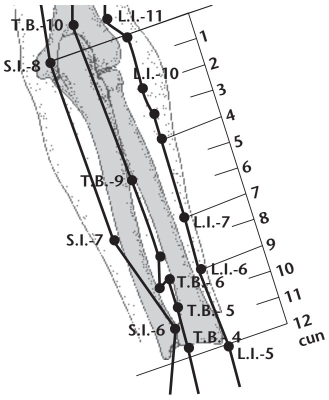
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
1
2
3
4
5
6
7
8
9
10
11
12
cun

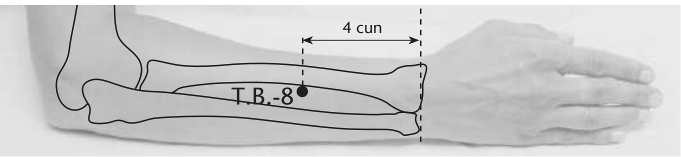

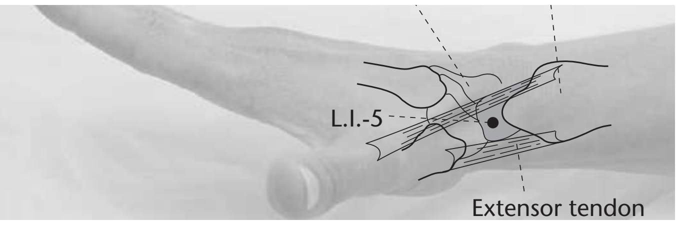
L.I.-5
Extensor tendon

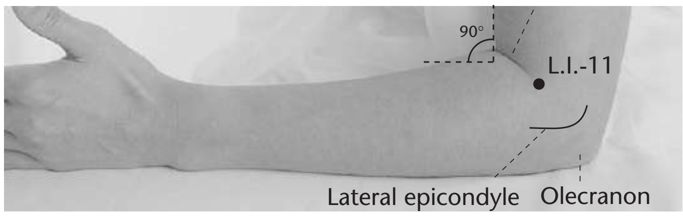
90°
L.I.-11
Lateral epicondyle
Olecranon

Location

5 cun proximal to the anatomical snuffbox in the direction of the lateral end of the elbow crease or 1 cun distal to the midpoint of the line connecting → **L.I.-5** and → **L.I.-11**.

How to find

Please note: With the hand supinated, the line connecting → **L.I.-5** and → **L.I.-11** runs alongside the radial margin of the forearm. However, with the arm pronated, it transverses the forearm. **L.I.-7** lies on the dorso-lateral aspect of the radius and is best located with the forearm in midposition and the elbow flexed. Spreading hands technique: Place the little fingers on → **L.I.-5** and → **L.I.-11** and span the hands evenly so that the thumbs join at the midpoint of the line. From there, measure 1 cun distally. → **T.B.-8** is located 4 cun proximal to the wrist crease between the radius and the ulna in the centre of the dorsal forearm or 1 cun distal to **L.I.-7**.

Needling

Vertically 0.5–1 cun

Actions/Indications

- Treats acute disorders
- Clears Heat and toxic Heat
- Regulates the Stomach and Large Intestine

Special features

Xi-cleft point

L.I.-8 *Lower Angle* XIALIAN

Location

4 cun distal to the lateral end of the elbow crease in the direction of the anatomical snuffbox and on a line connecting → **L.I.-5** and → **L.I.-11**.

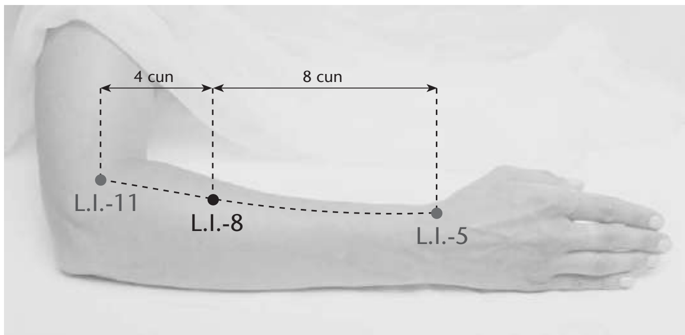
4 cun
8 cun
L.I.-11
L.I.-8
L.I.-5

How to find

Please note: In supination, the line connecting → **L.I.-5** and → **L.I.-11** runs along the radial margin of the forearm; in pronation, it transverses the forearm. **L.I.-8** lies on the dorsal aspect of the radius and is best located with the forearm in mid-position and the elbow flexed. **L.I.-8** can be found either 4 cun distal to → **L.I.-11** or 2 cun proximal to the midpoint of the line connecting → **L.I.-5** and → **L.I.-11** (spreading hands technique, → 2.3.3).

→ **T.B.-9** is 5 cun distal to the olecranon between the radius and the ulna, about 1 cun distal to **L.I.-8** in the middle of the dorsal forearm.

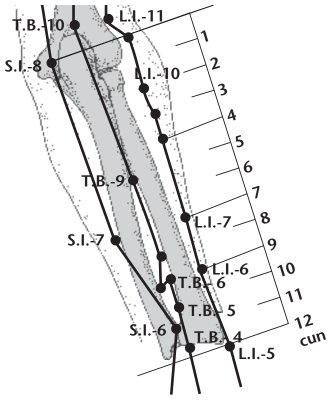
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
1
2
3
4
5
6
7
8
9
10
11
12
cun

Needling

Vertically or obliquely 0.5–1 cun

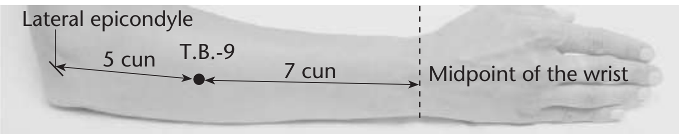
Lateral epicondyle
5 cun
T.B.-9
7 cun
Midpoint of the wrist

Actions/Indications:

- Opens the channel
- Cools and expels Wind
- Harmonises the Small Intestine

Upper Angle SHANGLIAN L.I.-9

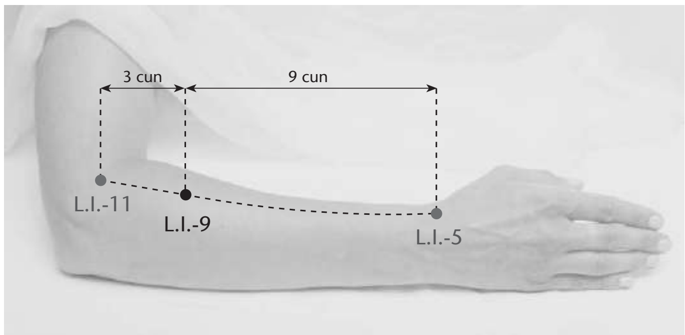

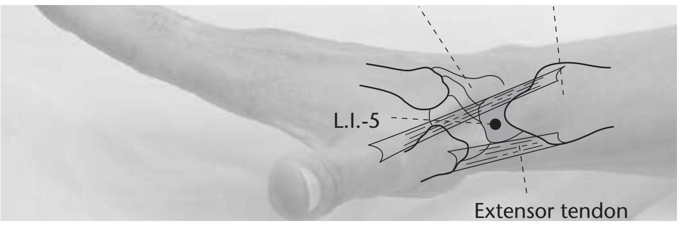

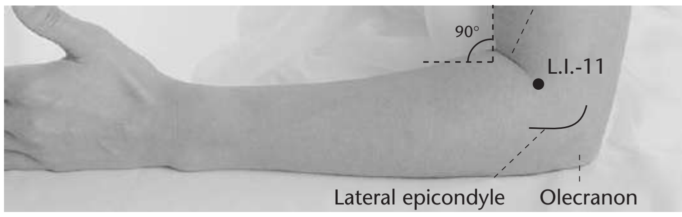

Location

3 cun distal to the lateral end of the elbow crease in the direction of the anatomical snuffbox, on the line connecting → **L.I.-5** and → **L.I.-11**.

How to find

Please note: With the hand supinated, the line connecting → **L.I.-5** and → **L.I.-11** runs along the radial margin of the forearm. However, with the arm pronated, it transverses the forearm. **L.I.-9** lies on the dorsal aspect of the radius and is best located with the forearm in midposition and the elbow flexed. From → **L.I.-11**, measure 3 cun distally. The point is located in an easily palpable depression between two muscle bellies and is often tender with pressure.

Needling

Vertically or obliquely 0.5–1 cun

Actions/Indications

- Opens the channel, cools and expels Wind
- Harmonises the Large Intestine

L.I.-10 *Arm Three Miles* SHOUSANLI

Location

2 cun distal to ➞ **L.I.-11**, on the line connecting ➞ **L.I.-5** and ➞ **L.I.-11**, on the extensor carpi radialis longus muscle; a deeper insertion will reach the supinator muscle.

How to find

Please note: With the hand supinated, the line connecting ➞ **L.I.-5** and ➞ **L.I.-11** runs along the radial margin of the forearm. However, with the arm pronated, it transverses the forearm. **L.I.-10** lies on the dorsal aspect of the radius and is best located with the forearm in midposition and with the elbow flexed. First, palpate ➞ **L.I.-11** in the depression lateral to the radial end of the cubital crease. **L.I.-10**, usually sensitive to pressure, is found by measuring 2 cun distally from ➞ **L.I.-11** on the connecting line.

Needling

Vertically 1–2 cun

Actions/Indications

- Regulates the Qi and Blood (in the upper extremity), opens the channel and *luo* vessels, alleviates pain
- Regulates the Stomach and Intestines

Special features

Important local point, often used in chain-and-lock method with other channel points

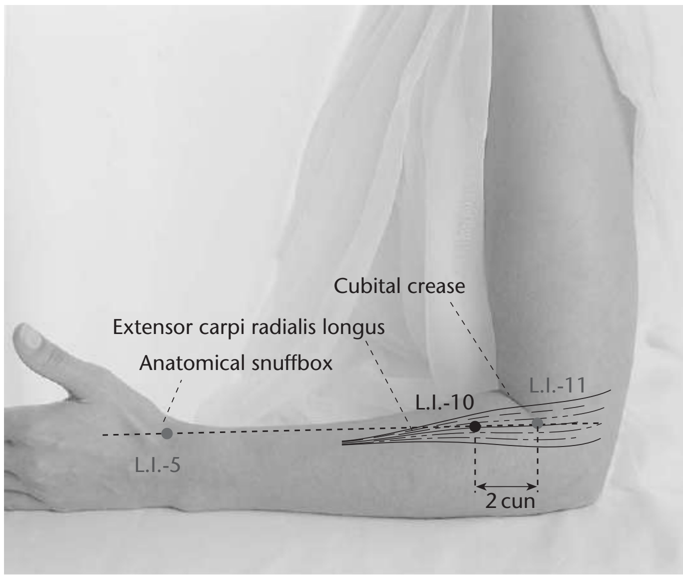
Cubital crease
Extensor carpi radialis longus
Anatomical snuffbox
L.I.-11
L.I.-10
L.I.-5
2 cun

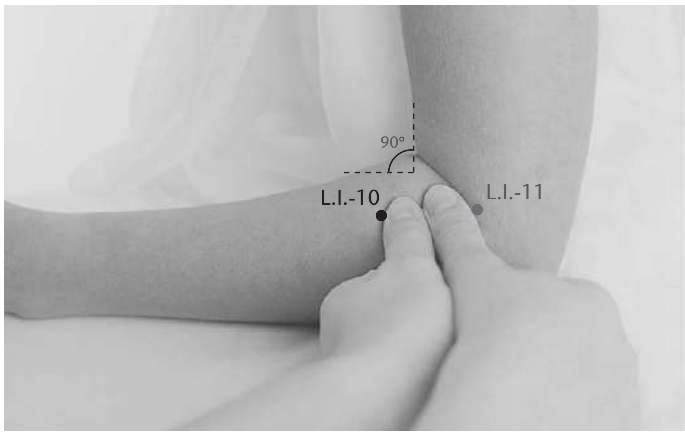
90°
L.I.-10
L.I.-11

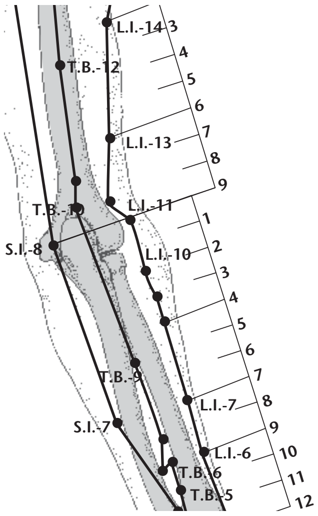
L.I.-14
T.B.-12
L.I.-13
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
3
4
5
6
7
8
9
1
2
3
4
5
6
7
8
9
10
11
12

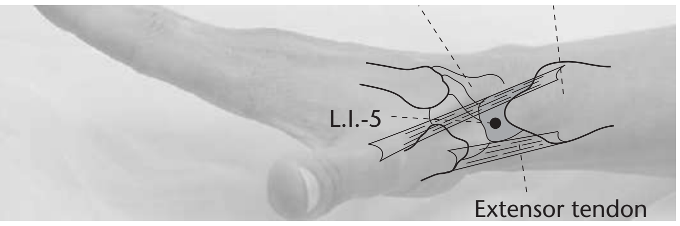
L.I.-5
Extensor tendon

Pool at the Crook QUCHI L.I.-11

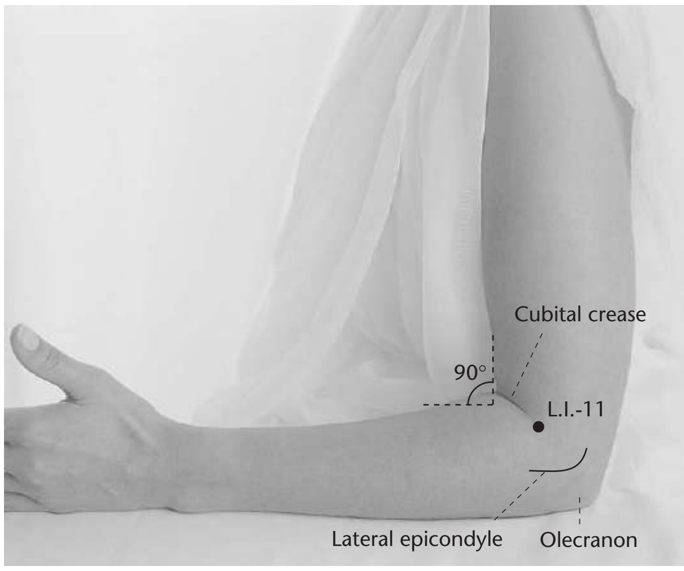
Cubital crease
90°
L.I.-11
Lateral epicondyle
Olecranon

Location

With the elbow flexed, on the lateral end of the elbow crease, in a depression between the end of the crease and the lateral epicondyle of the humerus, on the extensor carpi radialis longus muscle.

How to find

First, with the elbow completely flexed, locate the lateral end of the elbow crease. Next, with the elbow flexed at approximately 90°, palpate for a pressure-sensitive depression in this area on the extensor carpi radialis longus muscle. **L.I.-11** is located close to the border of the proximal aspect of the ulna.
Also located at the level of the elbow crease are → **LU-5,** → **P-3** (radial/ulnar of the tendon of the biceps) and → **HE-3** (with the elbow completely flexed at the medial end of the cubital crease).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Clears Heat, expels Wind
- Clears *yangming* – Fire
- Cools the Blood, drains Dampness, alleviates itching
- Opens the channel

Special features

He-sea point, Earth point, tonification point, Sun Si Miao Ghost point, Ma Dan Yang Heavenly Star point. Important point for Heat conditions and disorders of the upper extremity.

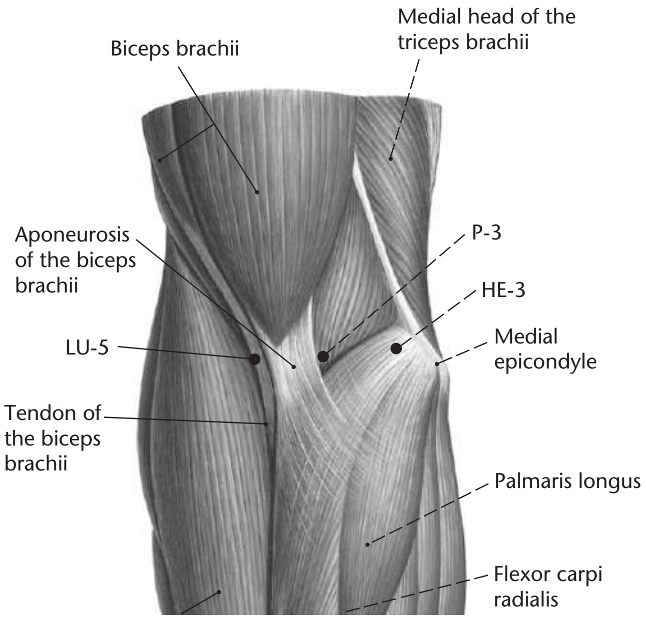
Biceps brachii
Medial head of the triceps brachii
Aponeurosis of the biceps brachii
P-3
HE-3
LU-5
Medial epicondyle
Tendon of the biceps brachii
Palmaris longus
Flexor carpi radialis

Left arm, anterior aspect

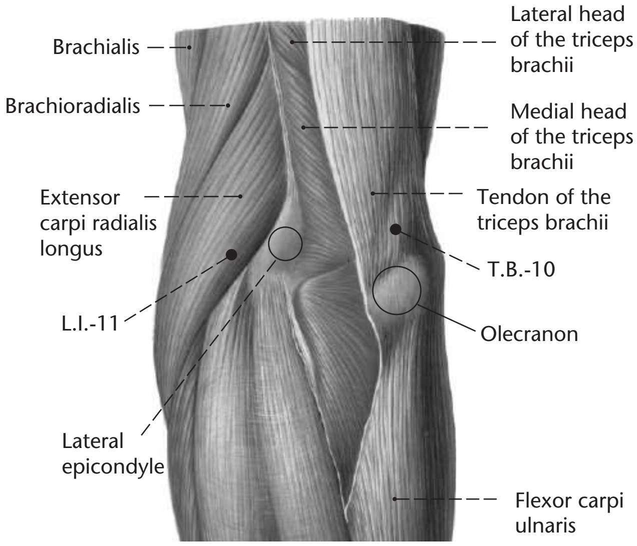
Brachialis
Lateral head of the triceps brachii
Brachioradialis
Medial head of the triceps brachii
Extensor carpi radialis longus
Tendon of the triceps brachii
T.B.-10
L.I.-11
Olecranon
Lateral epicondyle
Flexor carpi ulnaris

Left arm, posterior aspect

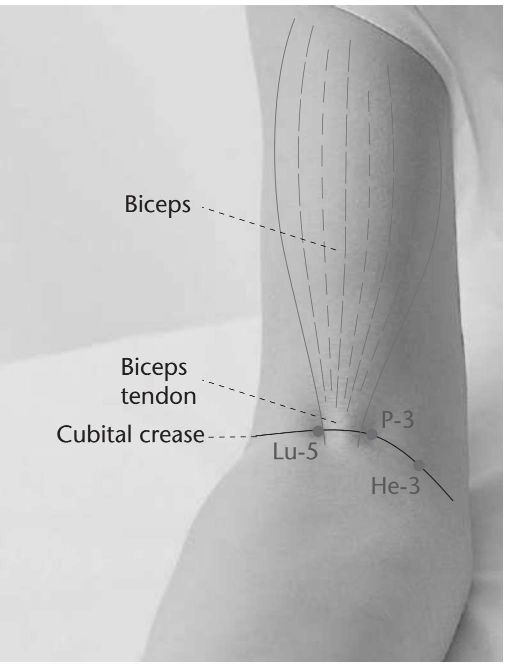
Biceps
Biceps tendon
Cubital crease
P-3
Lu-5
He-3

Right arm, anterior aspect

L.I.-12 *Elbow Crevice* ZHOULIAO

Location

1 cun proximal to the lateral end of the elbow crease (→ **L.I.-11**), on the anterior border of the humerus.

How to find

L.I.-12 is best located with the elbow flexed at 90°. From → **L.I.-11** (on the lateral end of the elbow crease), palpate 1 cun towards proximal. There, palpate for the anterior border of the humerus. **L.I.-12** is located at the junction of the shaft with the lateral epicondyle of the humerus and in front of the anterior border of the bone.

Needling

Vertically 0.5–1 cun, needle insertion between the border of the humerus and the flexor muscles in a medial direction.

Actions/Indications

- Opens the channel

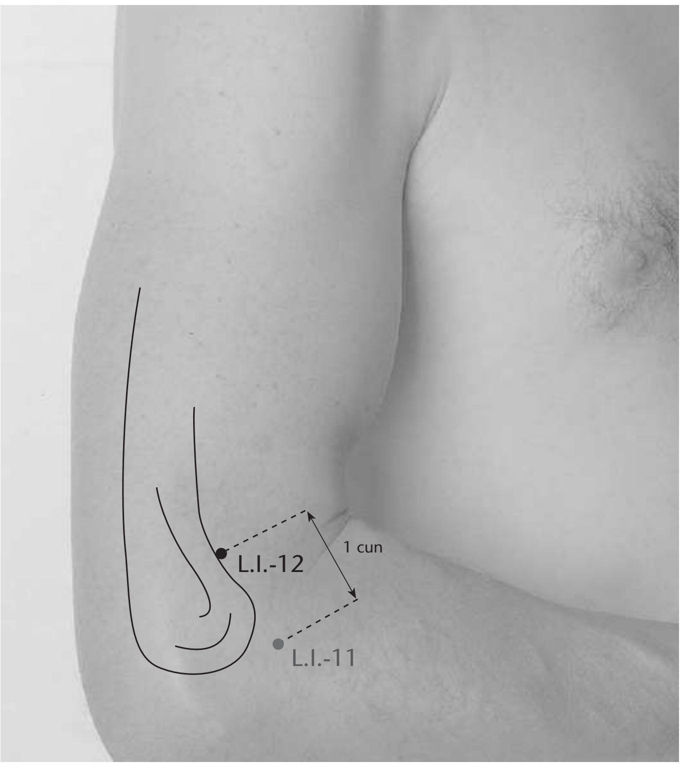
1 cun
L.I.-12
L.I.-11

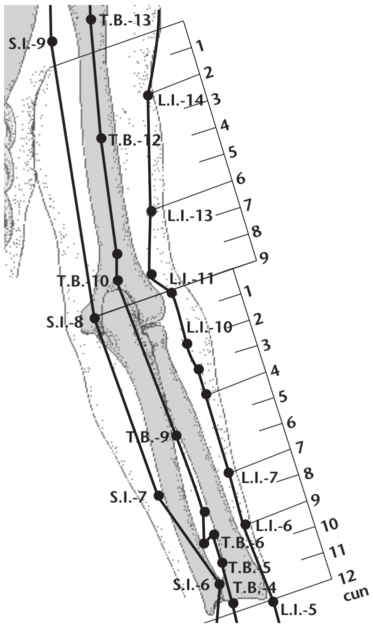
T.B.-13
S.I.-9
L.I.-14
T.B.-12
L.I.-13
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
S.I.-7
L.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
1
2
3
4
5
6
7
8
9
1
2
3
4
5
6
7
8
9
10
11
12
cun

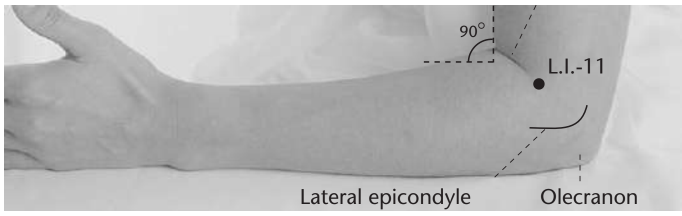
90°
L.I.-11
Lateral epicondyle
Olecranon

Arm Five Miles SHOUWULI L.I.-13

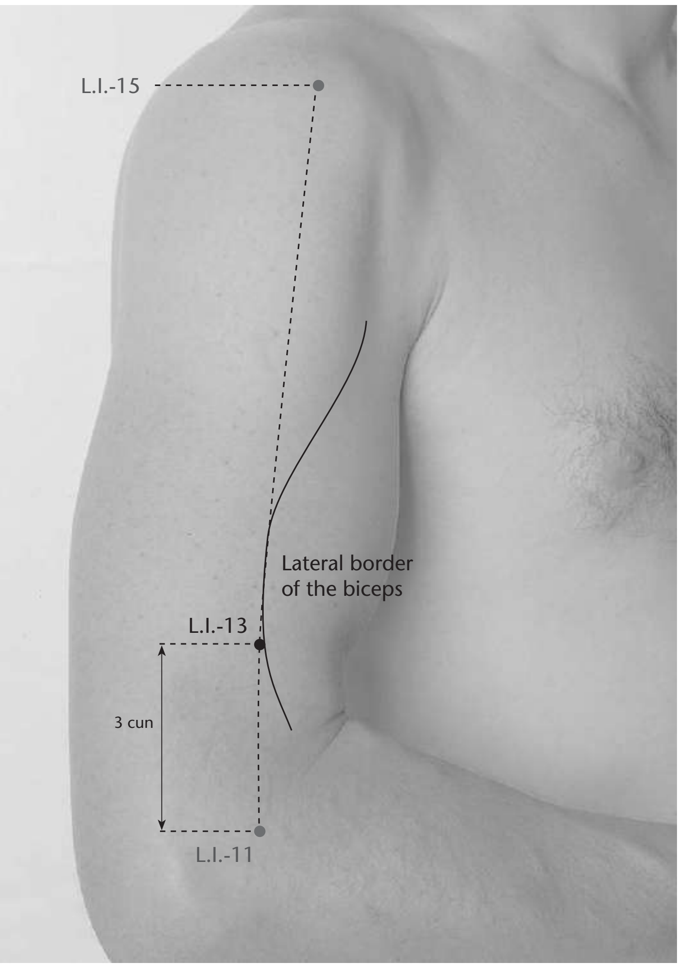
L.I.-15
Lateral border
of the biceps
L.I.-13
3 cun
L.I.-11

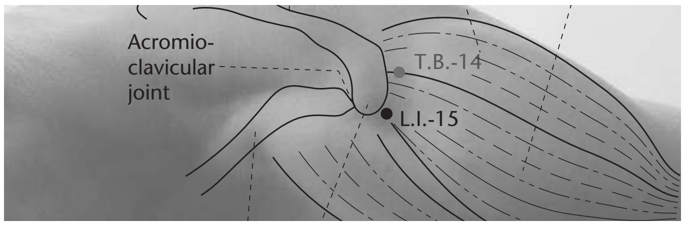
Acromio-
clavicular
joint
T.B.-14
L.I.-15

Location

On the lateral aspect of the upper arm, 3 cun proximal to the lateral end of the elbow crease (→ **L.I.-11**), in the direction of the head of the humerus.

How to find

This point is located on a line connecting → **L.I.-11** (lateral end of the elbow crease) and → **L.I.-15** (in a depression anterior and inferior to the acromion), on the lateral border of the biceps, in a groove between the biceps and the brachialis muscles. With the elbow flexed at 90°, measure 3 cun from → **L.I.-11** towards proximal. Ask the patient to flex their biceps.

Needling

Vertically 0.5–1 cun, alongside the anterior border of the humerus.

Actions/Indications

- Opens the channel

Ch04.2-F10028.qxd 2/23/08 7:06 PM Page 114

L.I.-14 *Upper Arm* BINAO

Location

On the lateral aspect of the upper arm, on a line connecting → **L.I.-11** and → **L.I.-15**, 7 cun proximal to → **L.I.-11** and slightly superior to the pointed insertion of the deltoid muscle.

How to find

This point is best located with the elbow flexed and the deltoid muscle flexed against resistance, making the lateral border of the muscle belly more visible. **L.I.-14** is located in a depression on the lower pointed insertion area of the deltoid. 'By moving the upper arm towards the tip of the nose, the patient will touch the approximate location of this point.' **For orientation:** This point lies approximately 2 cun inferior to the anterior end of the axillary fold (→ 2.2).

Needling

Vertically 0.5–1 cun; also obliquely towards the shoulder for disorders of the eyes.

Actions/Indications

- Opens the channel and *luo* vessels, alleviates pain
- Benefits the eyes

Special features

Important local point. According to some authors, meeting point with the *yang wei mai* and the S.I. and BL channels.

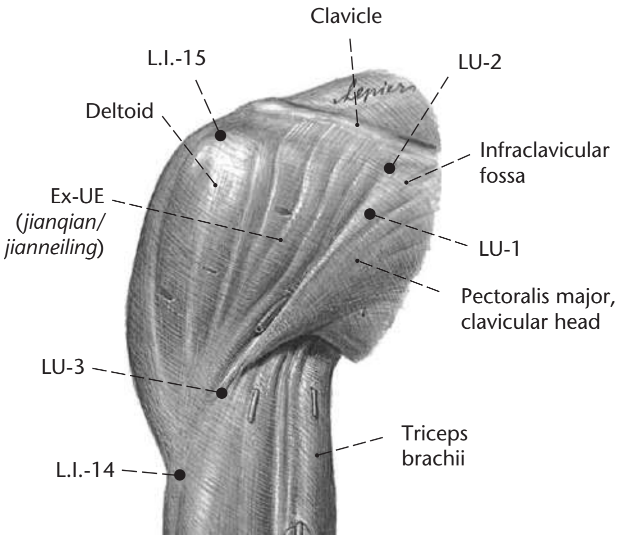
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
Ex-UE (*jianqian/ jianneiling*)
LU-1
Pectoralis major, clavicular head
LU-3
Triceps brachii
L.I.-14

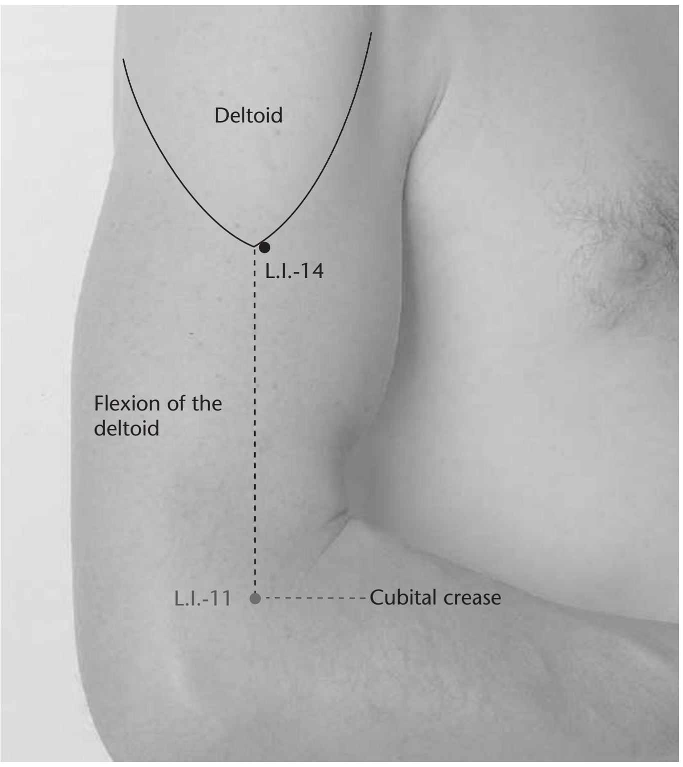
Deltoid
L.I.-14
Flexion of the deltoid
L.I.-11
Cubital crease

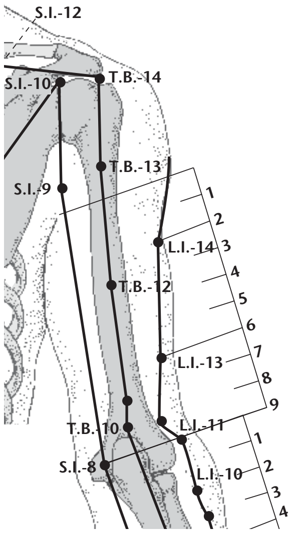
S.I.-12
S.I.-10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I.-14
3
4
T.B.-12
5
6
7
L.I.-13
8
9
T.B.-10
L.I.-11
S.I.-8
1
L.I.-10
2
3
4

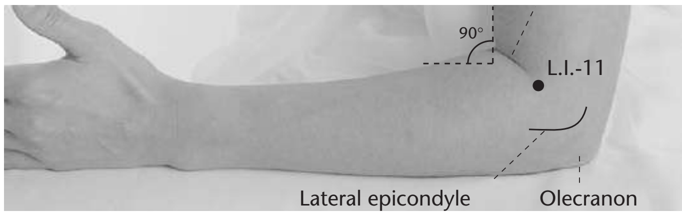
90°
L.I.-11
Lateral epicondyle
Olecranon

Shoulder Bone JIANYU L.I.-15

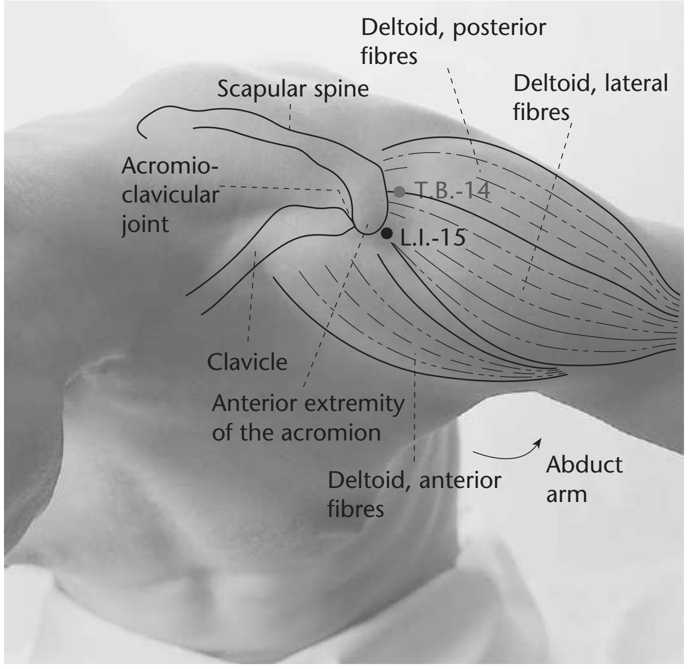
Deltoid, posterior fibres
Scapular spine
Deltoid, lateral fibres
Acromio-clavicular joint
T.B.-14
L.I.-15
Clavicle
Anterior extremity of the acromion
Abduct arm
Deltoid, anterior fibres

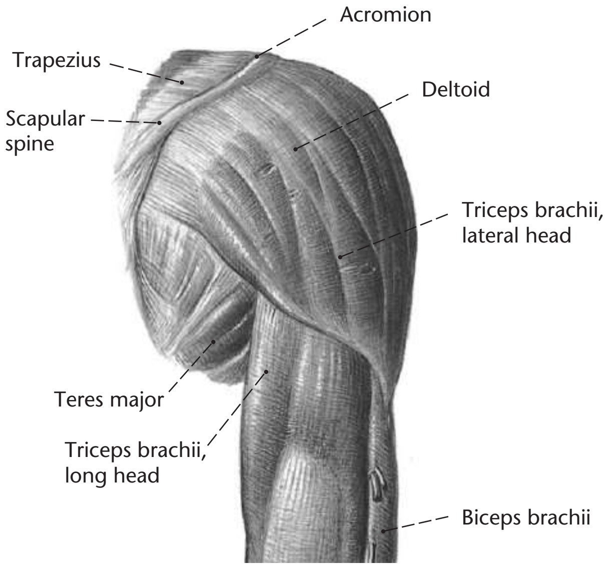
Acromion
Trapezius
Deltoid
Scapular spine
Triceps brachii, lateral head
Teres major
Triceps brachii, long head
Biceps brachii

Location

In the depression distal and anterior to the acromion, between the clavicular and acromial portions of the deltoid muscle.

How to find

Ask the patient to abduct their arm. With the arm in a horizontal position, two depressions will form in the insertion area of the deltoid, distally to the acromion. By placing the thumb and index finger in those depressions in a pincer-like grip (with thumb and finger 1 thumb's breadth apart), the anterior finger will be in the anterior depression, where **L.I.-15** is located.
→ **T.B.-14** is located under the posterior finger in the dorsal depression.

Needling

With the arm slightly adducted, 1–1.5 cun obliquely in the bony cleft and in a posterior direction; with the arm abducted, 0.5–1 cun vertically. Caution: Shoulder joint.

Actions/Indications

- Expels Wind-Damp, opens the channel, alleviates pain, supports the shoulder joint
- Expels Wind
- Regulates the Qi, resolves Phlegm

Special features

Meeting point with the *yang qiao mai*. Important local point for disorders of the shoulder, often used with chain-and-lock method (→ 8.3.6) with other channel points of the upper extremity.

L.I.-16 *Great Bone* JUGU

Location

In a depression between the acromial extremity of the clavicle and the junction of the scapular spine and the acromion.

How to find

This point is located on the shoulder, at the angle between the acromio-clavicular joint and the junction of the scapular spine and the acromion (→ 3.3.1), above the most lateral portion of the trapezius and supraspinatus muscles. At this point, the tendon of the supraspinatus delves below the acromion, where it often causes problems owing to the cramped anatomical structure (for example, impingement syndrome).

Needling

Vertically 0.5–1 cun

Actions/Indications

- Opens the channel, supports the shoulder joint

Special features

Meeting point with the *yang qiao mai*

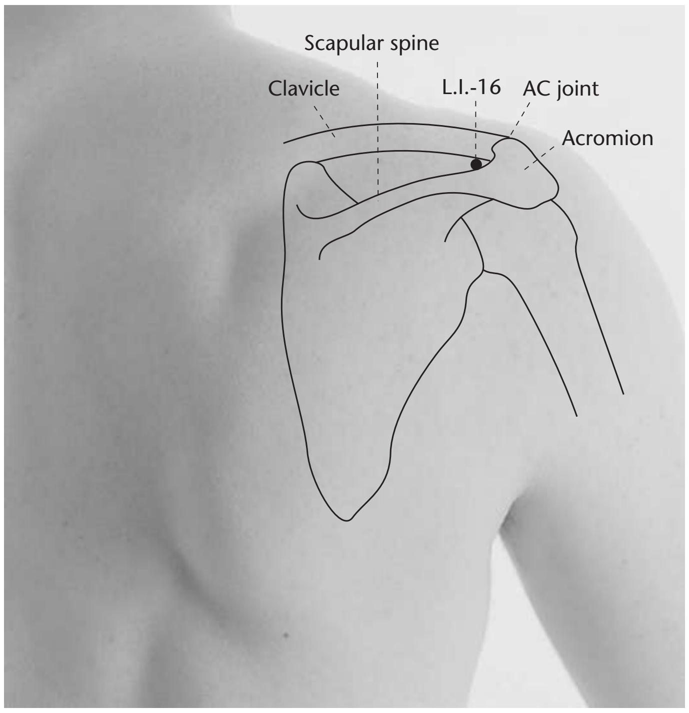
Scapular spine
Clavicle
L.I.-16
AC joint
Acromion

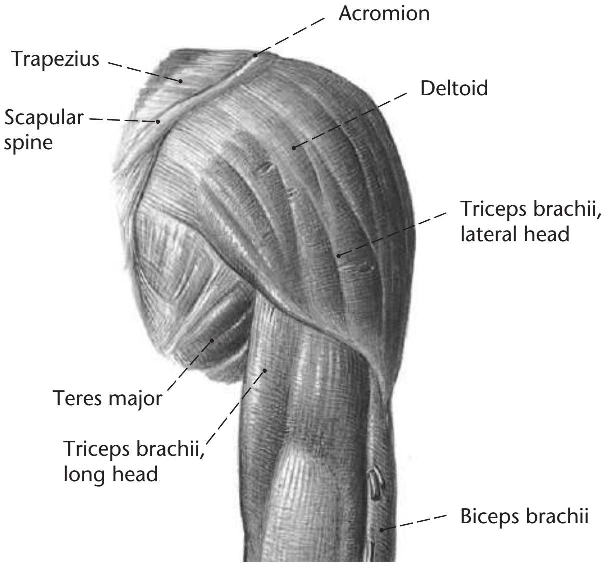
Acromion
Trapezius
Deltoid
Scapular spine
Triceps brachii, lateral head
Teres major
Triceps brachii, long head
Biceps brachii

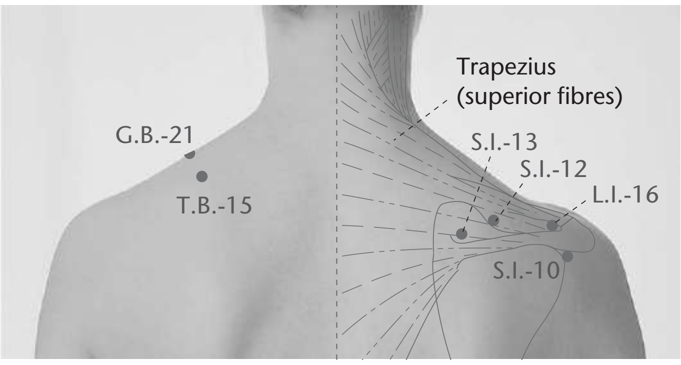
Trapezius (superior fibres)
G.B.-21
S.I.-13
S.I.-12
L.I.-16
T.B.-15
S.I.-10

Heaven's Tripod TIANDING **L.I.-17**

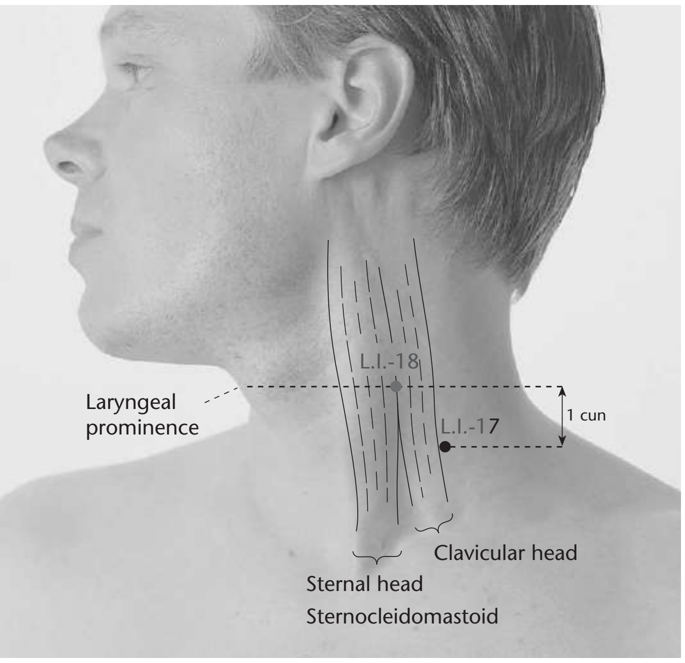
L.I.-18
Laryngeal prominence
L.I.-17
1 cun
Clavicular head
Sternal head
Sternocleidomastoid

Location

On the posterior border of the sternocleidomastoid muscle, 1 cun below the laryngeal prominence (→ 3.2).

How to find

This point is located on the lateral musculature of the neck, directly posterior to the sternocleidomastoid muscle. By rotating the patient's head to the opposite side, this muscle will become more visible and palpable. → **L.I.-18** is located 1 cun superior to **L.I.-17**, between the two heads of the sternocleidomastoid.

Needling

Rotate the head back to its normal position before needle insertion. Vertically 0.3–0.5 cun or obliquely up to 0.8 cun. Caution: Carotid artery, jugular vein.

Actions/Indications

- Benefits the throat and larynx

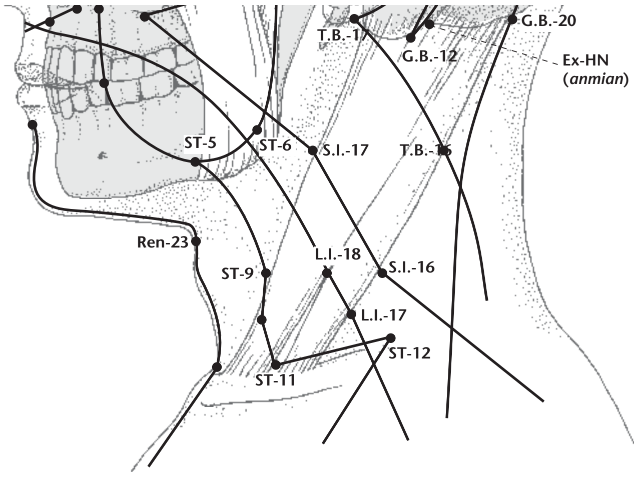
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

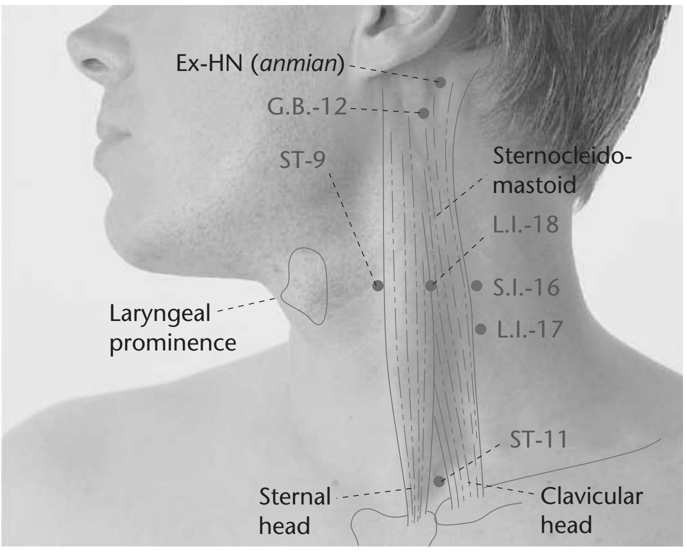
Ex-HN (*anmian*)
G.B.-12
ST-9
Sternocleido-
mastoid
L.I.-18
S.I.-16
L.I.-17
Laryngeal
prominence
ST-11
Sternal
head
Clavicular
head

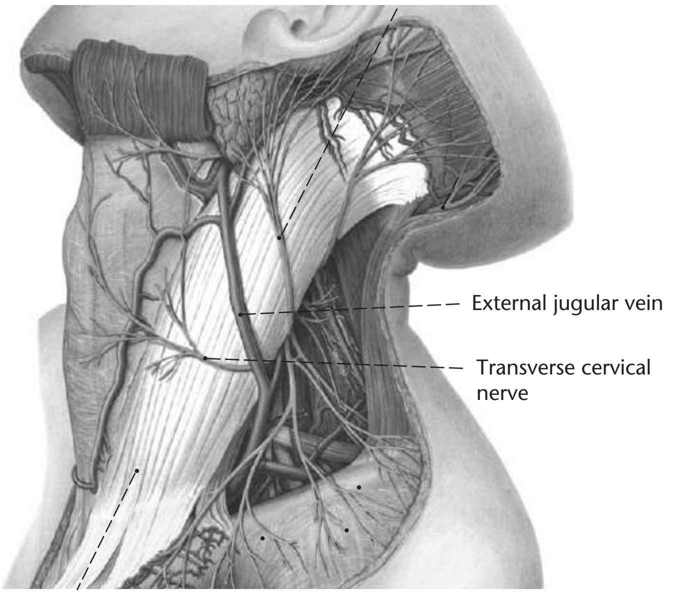
External jugular vein
Transverse cervical
nerve

L.I.-18 *Support the Prominence* FUTU

Location

On the lateral aspect of the neck, at the level of the laryngeal prominence, between the two heads of the sternocleidomastoid muscle (→ 3.2).

How to find

From the laryngeal prominence (tip of the Adam's apple), draw a line across the sternocleidomastoid muscle to its posterior border. **L.I.-18** is located between the two heads of this muscle, which becomes more visible and palpable by rotating the head to the opposite side.

Located on the same level are → **ST-9** on the anterior border and → **S.I.-16** on the posterior border of the sternocleidomastoid muscle.

Needling

Rotate the head back to its normal position before inserting the needle. Vertically 0.3–0.5 cun or obliquely up to 0.8 cun. Caution: Carotid artery, jugular vein.

Actions/Indications

- Benefits the throat and larynx

Special features

Window of Heaven point. Often used when swallowing is difficult, for example after a stroke.

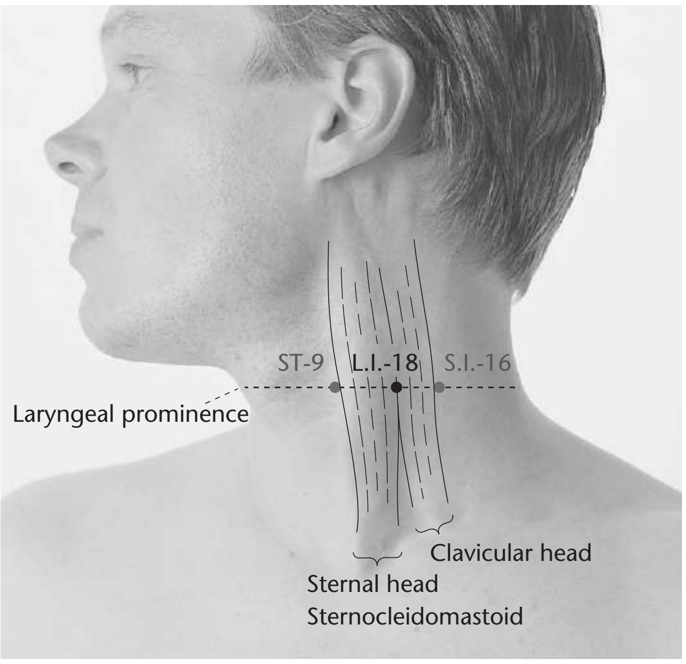
ST-9
L.I.-18
S.I.-16
Laryngeal prominence
Clavicular head
Sternal head
Sternocleidomastoid

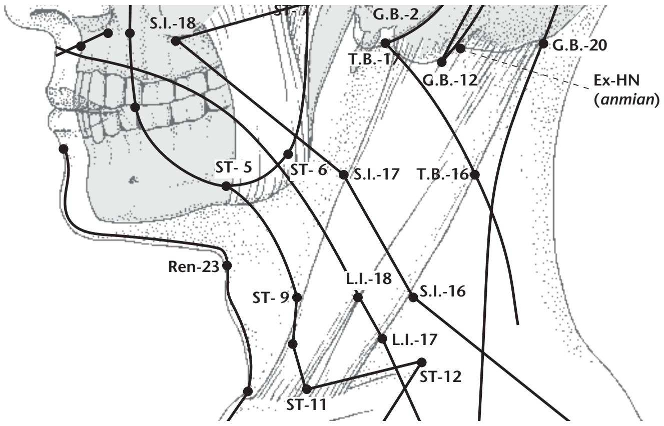
S.I.-18
ST-7
G.B.-2
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST- 5
ST- 6
S.I.-17
T.B.-16
Ren-23
ST- 9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

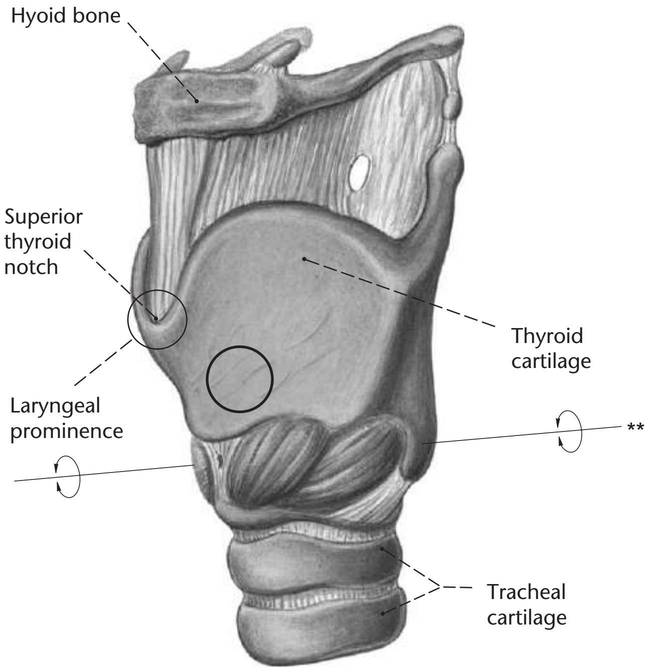
Hyoid bone
Superior
thyroid
notch
Thyroid
cartilage
Laryngeal
prominence
**
Tracheal
cartilage

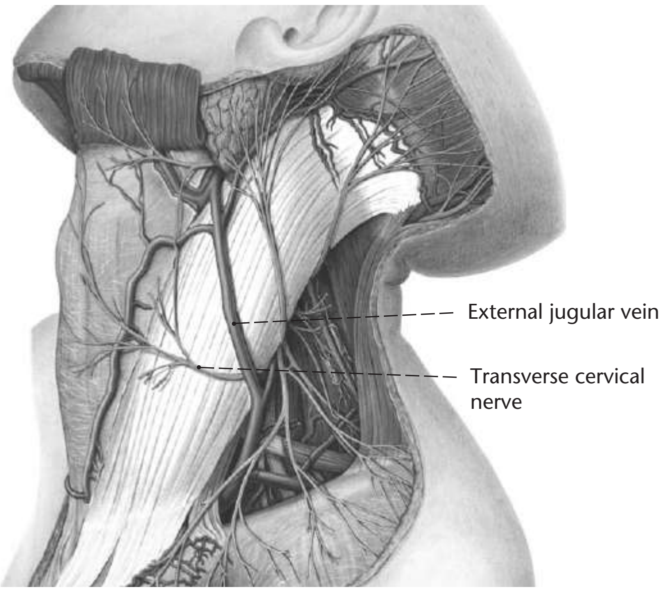
External jugular vein
Transverse cervical
nerve

Mouth Grain Crevice KOUHELIAO **L.I.-19**

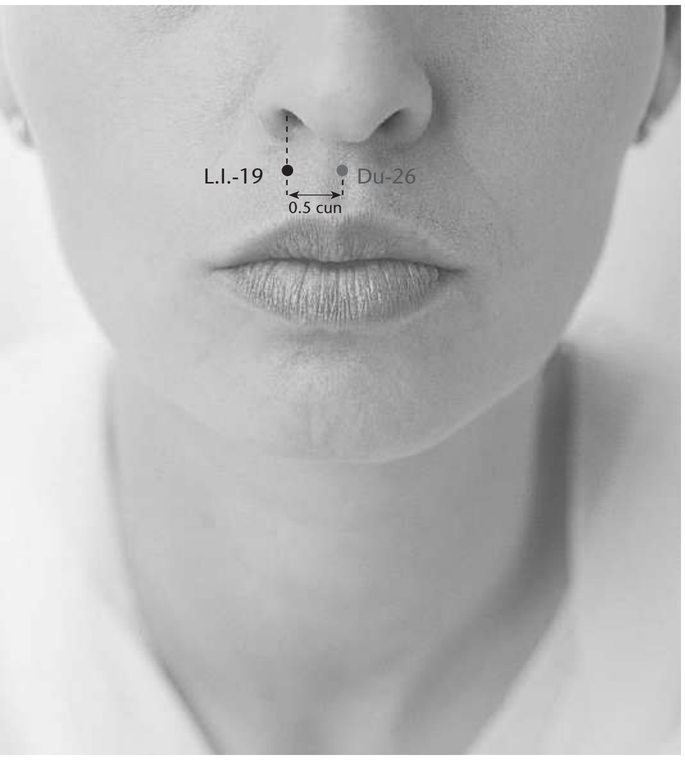
L.I.-19
Du-26
0.5 cun

Location

On the maxilla, slightly below the lateral margin of the nostril.

How to find

By dividing the distance between the margin of the nostril and the margin of the upper lip into three sections, **L.I.-19** is located at the junction of the upper and the middle third.

→ **Du-26** is located on the same level, on the anterior midline and at a distance of approximately 0.5 cun.

Needling

Obliquely 0.3–0.5 cun. Painful point! According to some classics, moxibustion is contraindicated.

Actions/Indications

- Expels Wind and opens the nasal passages

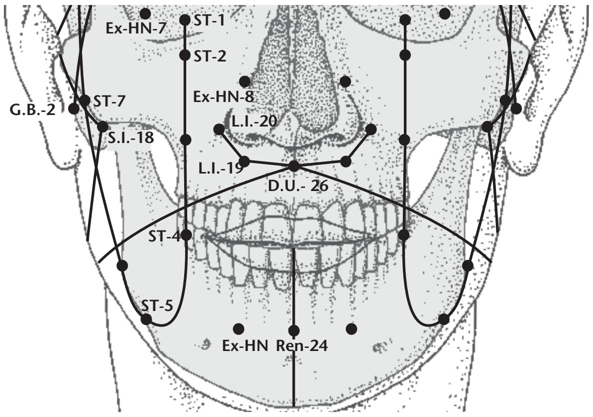
Ex-HN-7
ST-1
ST-2
Ex-HN-8
G.B.-2
ST-7
L.I.-20
S.I.-18
L.I.-19
D.U.- 26
ST-4
ST-5
Ex-HN
Ren-24

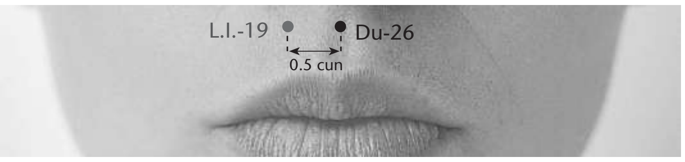
L.I.-19
Du-26
0.5 cun

L.I.-20 *Welcome Fragrance* YINGXIANG

Location

In the nasolabial groove, on the level of the midpoint of the lateral border of the ala nasi.

How to find

L.I.-20 is located by finding the midpoint of the lateral border of the ala nasi (➞ 3.1.2) and following it to the nasolabial groove. Tip: The nasolabial groove becomes more pronounced if you ask the patient to smile.

Needling

0.3–0.5 cun vertically, obliquely or transversely (subcutaneously) in a medial and superior direction towards the extra point ➞ **EX-HN-8** (*shangyingxiang* or *bitong*: at the upper end of the nasolabial groove, at the transition between the nasal bone and the cartilage). According to some authors, moxibustion is contraindicated.

Actions/Indications

- Opens the nasal passages, expels Wind, clears Heat

Special features

Meeting point with the ST channel, exit point. Most important local point for disorders of the nose.

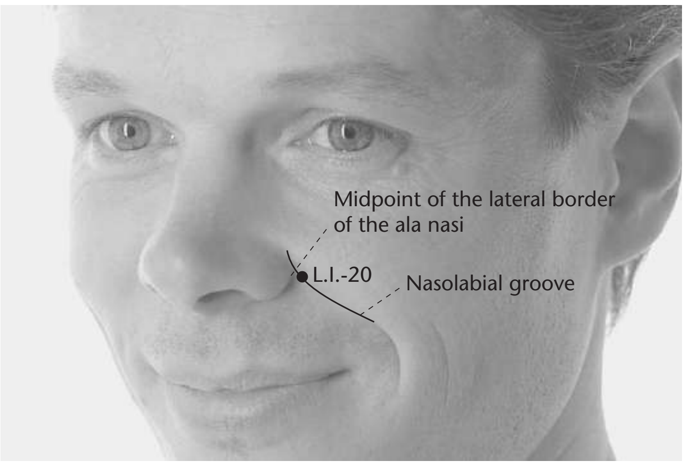
Midpoint of the lateral border of the ala nasi
L.I.-20
Nasolabial groove

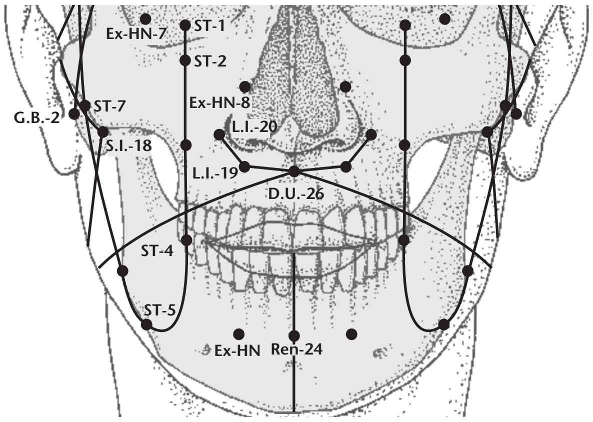
Ex-HN-7
ST-1
ST-2
Ex-HN-8
ST-7
G.B.-2
L.I.-20
S.I.-18
L.I.-19
D.U.-26
ST-4
ST-5
Ex-HN
Ren-24

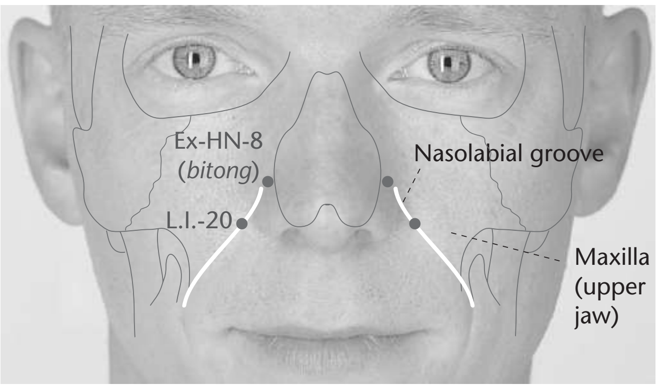
Ex-HN-8
(*bitong*)
L.I.-20
Nasolabial groove
Maxilla
(upper
jaw)

4.3 The Stomach Channel System – Foot *yangming* (*zu yang ming jing luo*)

4.3.1 The Stomach Primary Channel (*zu yang ming jing*)

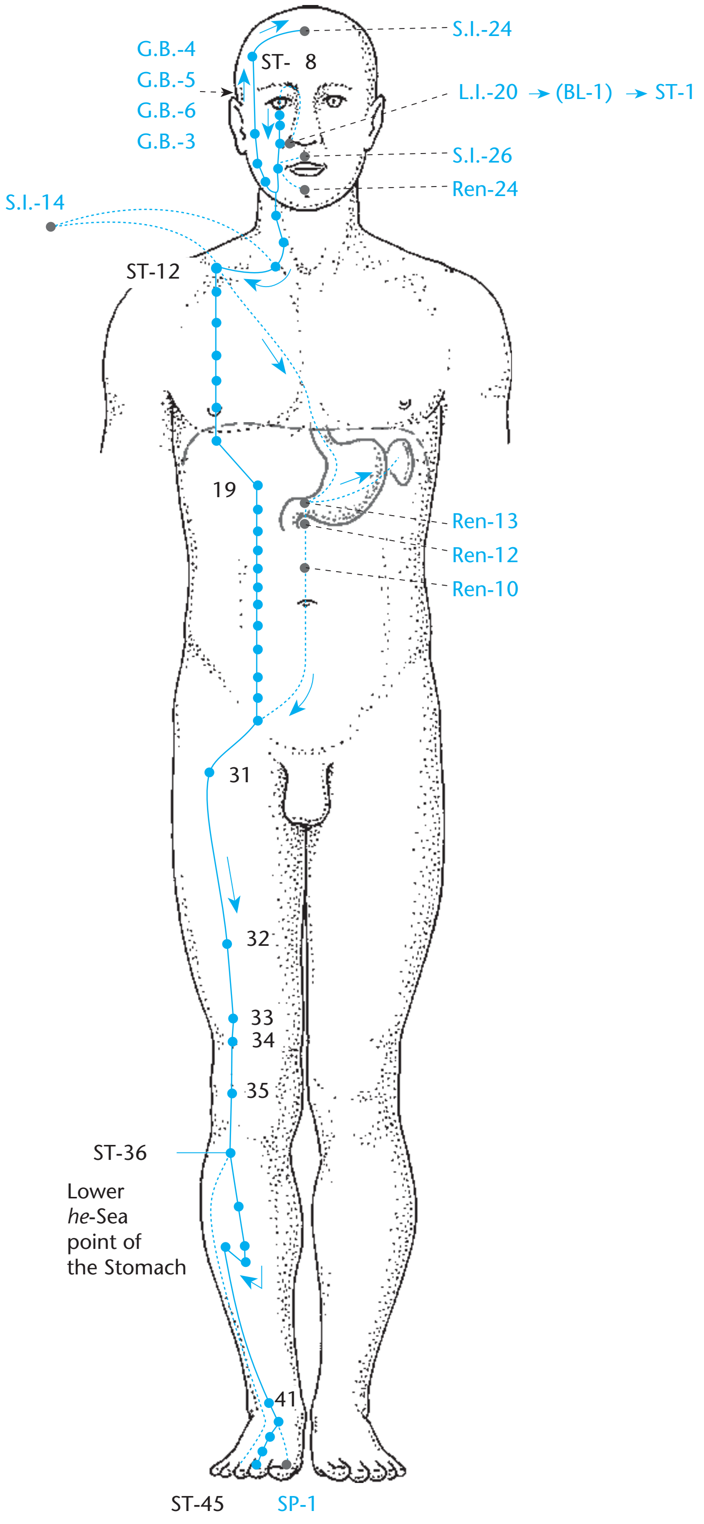
G.B.-4
G.B.-5
G.B.-6
G.B.-3
S.I.-24
ST- 8
L.I.-20 → (BL-1) → ST-1
S.I.-26
Ren-24
S.I.-14
ST-12
19
Ren-13
Ren-12
Ren-10
31
32
33
34
35
ST-36
Lower *he*-Sea point of the Stomach
41
ST-45
SP-1

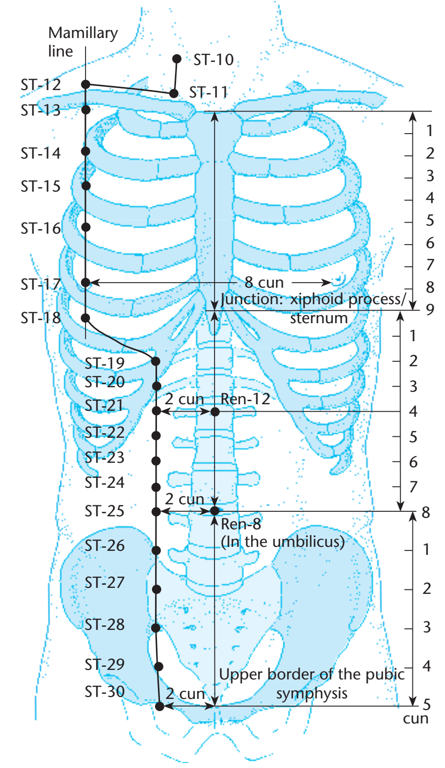
Mamillary line
ST-10
ST-12
ST-11
ST-13
ST-14
ST-15
ST-16
ST-17
8 cun
Junction: xiphoid process/ sternum
ST-18
ST-19
ST-20
ST-21
2 cun
Ren-12
ST-22
ST-23
ST-24
ST-25
2 cun
Ren-8 (In the umbilicus)
ST-26
ST-27
ST-28
ST-29
ST-30
2 cun
Upper border of the pubic symphysis
1
2
3
4
5
6
7
8
9
1
2
3
4
5
6
7
8
1
2
3
4
5
cun

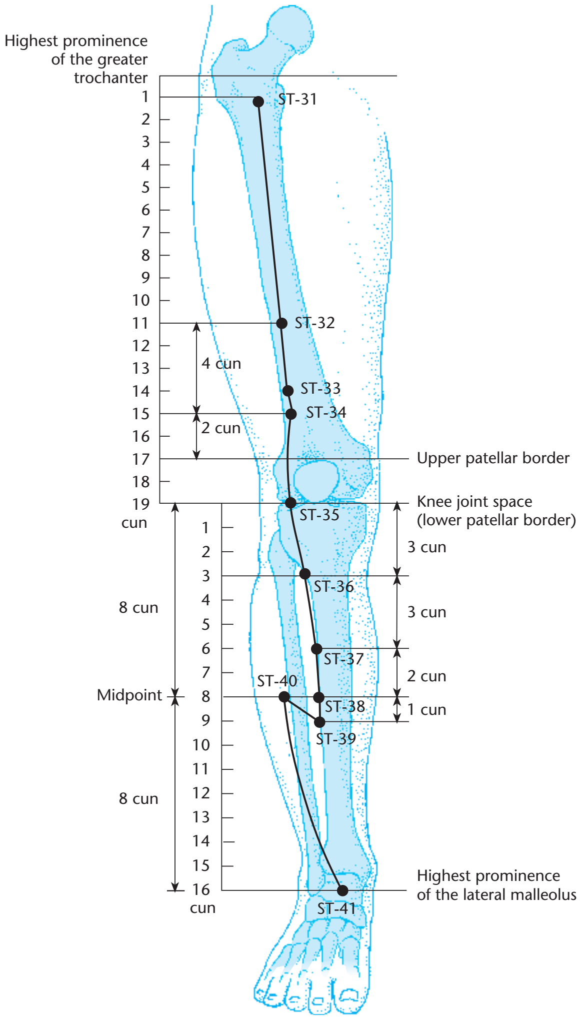
Highest prominence of the greater trochanter
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
ST-31
ST-32
4 cun
ST-33
ST-34
2 cun
Upper patellar border
Knee joint space (lower patellar border)
ST-35
3 cun
ST-36
3 cun
ST-37
2 cun
ST-38
1 cun
ST-39
ST-40
8 cun
Midpoint
8 cun
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus
ST-41

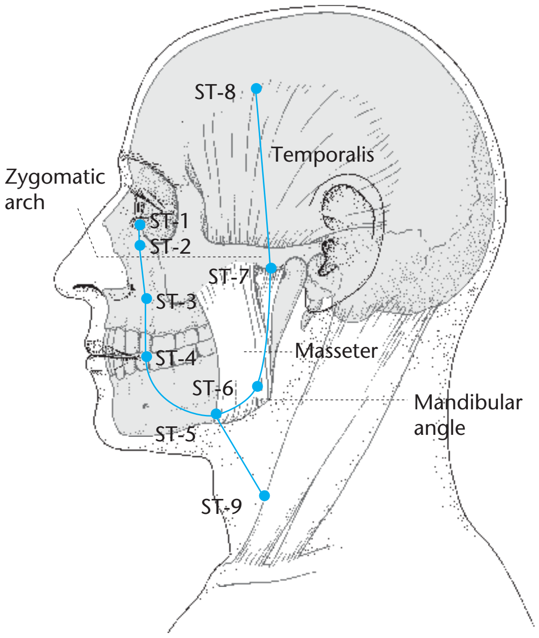

Pathway

The Stomach primary channel begins with an **internal** branch at the Yang axes connection of the first great circuit (*yangming*, hand–foot pairing) at → **LI-20** (*yingxiang*) lateral to the ala nasi

- travels **internally** to → **BL-1** (*jingming*) at the inner canthus of the eye
- emerges at the infraorbital ridge at **ST-1** (*chengqi*), the beginning of the external part of the primary channel
- descends laterally along the nose and **enters** the gums of the upper jaw
- meets the contralateral Stomach primary channel and the *du mai* at → **Du-26** (*renzhong*) and emerges again
- curves around the lips to meet the *ren mai* and again the contralateral Stomach channel at → **Ren-24** (*chengjiang*)
- traverses the lower cheek and winds around the lower angle of the mandible.

Here, at → **ST-5** (*daying*), the external channel **divides into two branches**:

- **One branch** ascends anterior to the ear, travels via → **G.B.-3** (*shangguan*), → **G.B.-6** (*xuanli*), → **G.B.-5** (*xuanlu*) and → **G.B.-4** (*hanyan*) to the temporal region to → **ST-8** (*touwei*) and continues to the forehead to meet → **Du-24** (*shenting*), where it terminates.
- The **other external branch** descends along the lateral aspect of the neck to the supraclavicular fossa and travels posteriorly over the back of the neck to → **Du-14** (*dazhui*) below the spinous process of the 7th cervical vertebra, where it meets the other five Yang primary channels.

An internal branch descends from the supraclavicular fossa, passes the diaphragm, enters first its pertaining *fu*-Organ, the Stomach (*wei*) and then connects with its paired *zang*-Organ, the Spleen (*pi*). Minor branches of the **internal** channel intersect with → **Ren-13** (*shangwan*), → **Ren-12** (*zhongwan*) and → **Ren-10** (*xiawan*). The internal branch then descends towards the inguinal region where it emerges slightly superior to **ST-30** (*qichong*) and reconnects with the **external** branch.

- The **external** channel descends from the supraclavicular fossa, crosses the nipple and laterally passes the umbilicus, travels in a curve to the leg and knee, continues along the anterior crest of the tibia and the dorsum of the foot and terminates on the second toe at → **ST-45** (*lidui*) at the lateral corner of the nail.
- An **internal** branch originates below the knee area at → **ST-36** (*zusanli*) and travels to the lateral aspect of the middle toe.
- A further branch separates from the primary channel at → **ST-42** (*chongyang*) on the dorsum of the foot and connects at → **SP-1** (*yinbai*) with its paired SP channel on the medial aspect of the big toe (foot Yin–Yang connection of the first great circuit).

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: High fever, malaria, flushed face, sweating, confused mental state, aversion to cold, eye pain, dry nasal passages, nosebleeds, dry lips and mouth, lesions of the lips and mouth, sore and swollen throat, thoracic pain, red and swollen legs

Interior (*li*) or *zangfu*-Organ signs and symptoms: Abdominal distension, sensation of fullness, oedema, irritability during work and rest, mania and epilepsy, hyperpepsinia, constant hunger, yellow urine

Excess (*shi*): Heat sensations on the anterior aspect of the body, constant hunger, yellow urine

Deficiency (*xu*): Cold sensations on the anterior aspect of the body, chills, Stomach Cold with distension and fullness

Connections and meeting points of the Stomach primary channel (→ 1.2)

Connections with other channels

Spleen primary channel (*zu tai yin jing*)

Connection: Foot Yin–Yang connection of the first great circuit
Location: ST-42 → SP-1 (on the foot).
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Large Intestine primary channel (*shou yang ming jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *yang ming* (Yang axes of the first great circuit)
Location: L.I.-20 → (BL-1) → ST-1 (on the head)
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with other *zangfu*-Organ systems

Stomach (*wei*), Spleen (*pi*)

4.3.2 The Stomach Divergent Channel (*zu yang ming jing bie*)

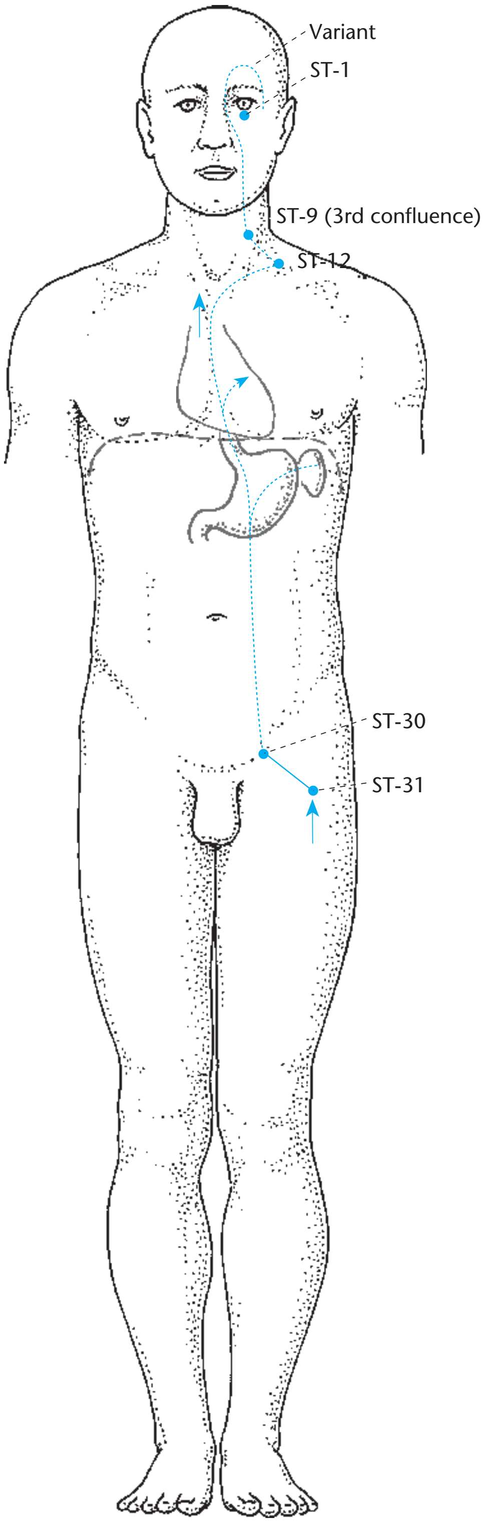

Pathway

The Stomach divergent Channel separates from the Stomach primary channel at → **ST-31** (*biguan*) on the antero-lateral aspect of the thigh

➡ penetrates the body at the inguinal region at **ST-30** (*qichong*)

➡ continues to the Stomach (*wei*) and disperses in the Spleen (*pi*)

➡ ascends and penetrates the Heart (*xin*)

➡ ascends along the oesophagus to the suprasternal fossa and travels laterally to the supraclavicular fossa at **ST-12** (*quepen*).

From there it rises along the anterior border of the sternocleidomastoid muscle and meets its primary channel as well as the divergent Spleen channel at **ST-9** (*renying*) at one of the six *he*-confluences (here: ST/SP as third confluence → 1.3).

It then **emerges** at the mouth, ascends along the nose, reaches the inner canthus of the eye and connects with the eye, circles the eye region and meets the Stomach primary channel at **ST-1** (*chengqi*) where it terminates (→ fig.).

According to some authors, the Stomach divergent channel connects with the Stomach primary channel at **BL-1** (→ variant, fig.). After intersecting with **BL-1**, the divergent channel could enter the head, disperse in the brain and re-emerge at **ST-1.**

Clinical importance

- Strengthens the relationship between the Stomach and the Spleen (*zangfu*-Organs). Points on the Stomach primary channel can therefore be used for disorders of the Spleen, and vice versa points on the Spleen primary channel can treat disorders of the Stomach.
- Spreads Qi to the face and sensory organs: many points on the ST channel treat disorders of the head and face.
- Strengthens the relationship between the Stomach and the eyes: Heat and excess in this region can be directed downward by using points on the ST channel.

Ch04.3-F10028.qxd 2/22/08 8:41 PM Page 124

4.3.3 The Stomach Sinew Channel (*zu yang ming jing jin*)

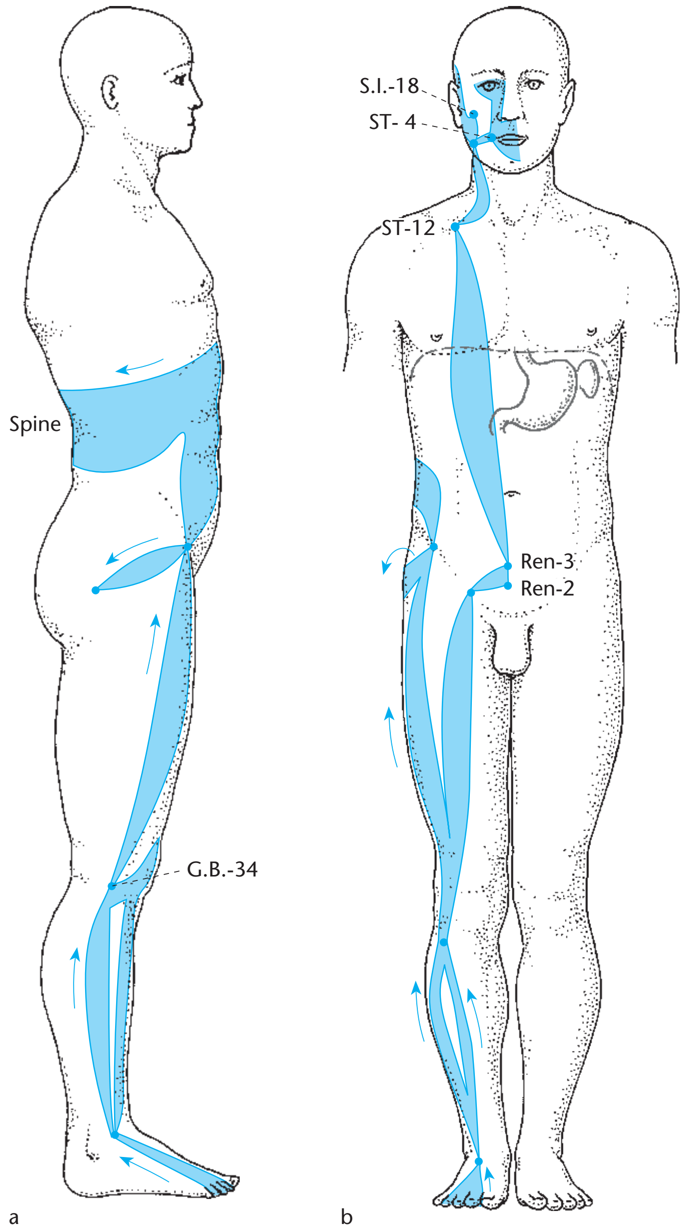

Pathway

The Stomach sinew channel begins as a wider band, encompassing the second, third and fourth toe, traverses the dorsum of the foot, binds (*jie*) at the tibial depression and divides into two branches:

- **One branch** obliquely ascends the **anterolateral aspect** of the leg and binds (*jie*) at the lateral aspect of the knee. It then ascends the anterolateral aspect of the thigh and binds (*jie*) at the anterior hip region. From there it sends **a branch** to → **G.B.-30** (*huantiao*) in the direction of the greater trochanter. The sinew channel travels from the anterior hip region across the lateral abdomen and spreads across the lower ribcage posteriorly to the spine.
- **Another branch** follows the tibia on the **anterior aspect** of the lower leg and binds (*jie*) at the level of the patella. From there a branch runs to the head of the fibula where it meets the G.B. sinew channel at → **G.B.-34** (*yanglingquan*).
- From the patella the channel ascends along the anterior aspect of the thigh, binds (*jie*) in the inguinal region, passes → **Ren-2** (*qugu*) and binds (*jie*) at → **Ren-3** (*zhongji*). It then ascends across the abdominal and thoracic region and reaches the supraclavicular fossa to bind (*jie*) at **ST-12** (*quepen*). It then continues along the anterolateral aspect of the neck and binds (*jie*) at the angle of the mandible.

At the angle of the mandible the channel divides into three branches:

- **One branch** terminates anterior to the ear.
- **A further branch** ascends to the zygomatic bone where it meets the other Yang sinew channels at → **S.I.-18** (*quanliao*).
- **The third branch** circles the mouth, ascends to the upper nose region, there connects with the Bladder sinew channel and then spreads along the lower eyelid. The Bladder sinew channel spreads along the upper eyelid, so that the two channels together form a network around the eyes.

Clinical importance

Pathology: Stiffness and aching of the toes, leg cramps (gastrocnemius and quadriceps muscles), stiffness and pain on the dorsum of the foot (at **ST-41**), swelling and tension in the inguinal region, *shan*-disorders, cramping in the abdomen as well as in the supraclavicular fossa and the face, facial paralysis, weakness and paralysis of the superior rectus muscle.

Indication: Mainly used for *bi*-syndromes (painful obstruction syndromes) along the Stomach channel. The area covered by the Stomach sinew channel is larger than that covered by the Stomach primary channel. This explains why the indications of points on the Stomach primary channel include disorders and diseases of the external genitalia (for example urinary tract disorders such as cystitis, hernia and orchitis, which are often treated in combination with LIV points) as well as disorders of the eyelids and of all sensory organs.

4.3.4 The Stomach *luo*-Connecting Vessel System (*zu yang ming luo mai*)

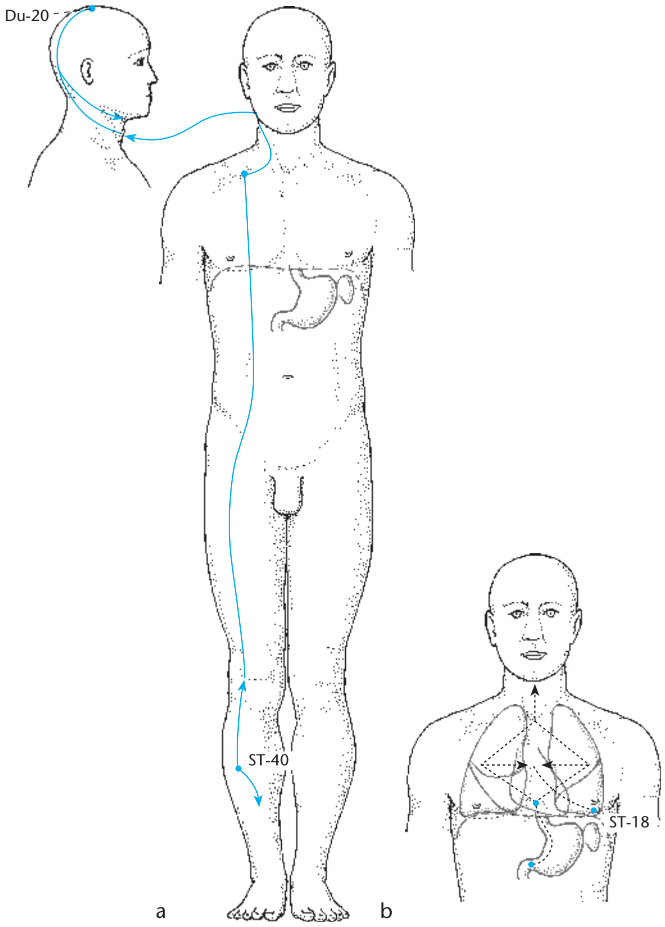

Pathway

The ST *luo*-connecting vessel separates from the Stomach primary channel at its *luo*-connecting point **ST-40** (*fenglong*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches *(sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- ➡ Horizontal divisions run to the Interiorly–Exteriorly paired primary SP channel; according to some schools of thought (for example Van Nghi, → Appendix) they travel as a **transverse** Stomach *luo*-connecting vessel to the *yuan* point → **SP-3** (*taibai*).
- ➡ A **longitudinal** division ascends along the anterolateral aspect of the leg to the torso and reaches the lateral aspect of the neck at → **ST-12**, where it divides into two branches: **one branch** traverses the neck, the **other branch** runs to the occiput and ascends the head to reach → **Du-20** (*baihui*).

Clinical importance (→ 8.1.2)

Pathology

- **Counterflow Qi:** Swelling and pain of the throat, sudden loss of voice, sensation of tightness in the throat
- **Excess (*shi*):** Psychological disorders, also epilepsy and mania
- **Deficiency (*xu*):** *Wei*-syndrome of the legs (atrophy syndrome with muscle weakness and paralysis)

The great *luo*-connecting vessel of the Stomach (*wei zhi dao luo* or *xu li*)

Pathway

The great Stomach *luo*-connecting vessel begins at the Stomach (*wei*), passes through the diaphragm, intersects with → **Ren-17** (*shanzhong*) and spreads in the Lungs (*fei*), trachea and larynx. From the Lung it travels to the Heart (*xin*) and emerges on the left aspect of the thorax near → **ST-18** (*rugen*) which is located where the heartbeat is visible (→ 8.1.2).

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*):** Dyspnoea
- **Deficiency (*xu*):** Chest tightness, for example with asthma, coughing, angina pectoris, etc.

4.3.5 Cutaneous Region (*yang ming pi bu*)

See description and figures → 1.6

4.3.6 Points of the Stomach Primary Channel (Overview)

Specific points according to their function (→ 8.1)

- ***Yuan*-source point (→ 8.1.1): ST-42** (*chongyang*) 🟦■
- ***Luo*-connecting point** (→ **8.1.2.): ST-40** (*fenglong*) 🟦🟦
- ***Xi*-cleft point (→ 8.1.3): ST-34** (*liangqiu*) 🟦■
- **Associated Back-*shu* point (→ 8.1.4): BL-21** (*weishu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): Ren-12** (*zhongwan*) 🟦■
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Metal), sedation point: **ST-45** (*lidui*) 🟦■
  ying-spring point (Water): **ST-44** (*neiting*) 🟦■
  shu-stream point (Wood): **ST-43** (*xiangu*) 🟦■
  jing-river point (Fire), tonification point: **ST-41** (*jiexi*) 🟦■
  he-sea point (Earth), *ben* point: **ST-36** (*zusanli*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8): –**
- **Lower *he*-sea points (→ 8.1.9):**
  of the Stomach: **ST-36** (*zusanli*) 🟦🟦
  of the Large Intestine: **ST-37** (*shangjuxu*) 🟦■
  of the Small Intestine: **ST-39** (*xiajushu*) 🟦■
- ***jiaohui*-meeting points (→ 8.1.10):**

  - with the *ren mai, yang qiao mai, du mai**: **ST-1** (*chengqi*) 🟦■
  - with the *yang qiao mai**: **ST-2** (*sibai*) 🟦■
  - with the *yang qiao mai*: **ST-3** (*juliao*)
  - with the primary L.I. channel, *yang qiao mai, du mai*, ren mai*:* **ST-4** (*dicang*)
  - with the G.B. channel*: **ST-5** (*daying*), **ST-6** (*jiache*)
  - with the G.B. channel: **ST-7** (*xiaguan*) 🟦■**, ST-9** (*renying*)
  - with the G.B. channel, *yang wei mai*: **ST-8** (*touwei*) 🟦■
  - with the primary L.I., S.I., T.B., G.B. channels: **ST-12** (*quepen*)
  - with the *chong mai*, G.B. channel*: **ST-30** (*qichong*) 🟦■
  - other channels with the ST channel: **L.I.-20, BL-1, G.B.-3, G.B.-4, G.B.-5, G.B.-6; Du-14, Du-24; Du-26, Du-28*; Ren-10*, Ren-12, Ren-13**, **Ren-24; G.B.-14*, G.B.-21***
- **Gao Wu command point** (**→ 8.1.11**) **for the abdomen: ST-36** (*zusanli*) ■■
- **Window of Heaven point** (**→ 8.1.12**)**: ST-9** (*renying*)
- **Points of the Four Seas** (**→ 8.1.13**)**:**
  Sea of Qi point: **ST-9** (*renying*)
  Sea of Grain: **ST-30** (*qichong*) 🟦■, **ST-36** (*zusanli*) 🟦🟦
  Sea of Blood: **ST-37** (*shangjuxu*) 🟦■, **ST-39** (*xiajushu*) 🟦■
- **Ma Dan Yang Heavenly Star points** (**→ 8.1.14**)**: ST-36** (*zusanli*) 🟦🟦, **ST-44** (*neiting*) 🟦🟦
- **Sun Si Miao Ghost point** (**→ 8.1.15**)**: ST-6** (*jiache*) 🟦■
- **Other functional points:**
  - Front-*mu* point of the Large Intestine: **ST-25** (*tianshu*) 🟦🟦
  - most important acupuncture 'Phlegm'-point: **ST-40** (*fenglong*) 🟦🟦
- Points according to region
- **Local points** (**→ 8.2.1**)**:** eyes – **ST-1** (*chengqi*) 🟦■; mouth and teeth – **ST-4** (*dicang*), **ST-6** (*jiache*) 🟦■; jaw – **ST-7** (*xiaguan*) 🟦■; Large Intestine – **ST-25** (*tianshu*) 🟦🟦; Small Intestine and Bladder – **ST-28** (*shuidao*); urogenital region – **ST-30** (*qichong*) 🟦■; knee – **ST-34** (*liangqiu*) 🟦■, **ST-35** (*dubi*) 🟦■, **ST-36** (*zusanli*) 🟦🟦; foot – **ST-41** (*jiexi*) 🟦■
- **Adjacent points** (**→ 8.2.1**)**:** forehead – **ST-8** (*touwei*) 🟦■; eyes – **ST-2** (*sibai*) 🟦■; mouth/teeth – **ST-7** (*xiaguan*) 🟦■; throat – **ST-10** (*shuitu*); Spleen, Stomach, Gall Bladder – **ST-21** (*liangmen*); Kidney – **ST-29** (*guilai*); abdominal region – **ST-25** (*tianshu*) 🟦🟦, **ST-26** (*wailing*), **ST-30** (*qichong*) 🟦■; foot – **ST-34** (*liangqiu*) 🟦■, **ST-36** (*zusanli*) 🟦🟦; toes – **ST-41** (*jiexi*) 🟦■
- **Distal points** (**→ 8.2.1**)**:** frontal headaches – **ST-44** (*neiting*) 🟦■; nose – **ST-44** (*neiting*) 🟦■; mouth/teeth – **ST-44** (*neiting*) 🟦■; throat – **ST-44** (*neiting*) 🟦■; shoulder region – **ST-38** (*tiakou*) 🟦■; Lung – **ST-40** (*fenglong*) 🟦🟦; Spleen/Stomach – **ST-36** (*zusanli*) 🟦🟦; Large Intestine – **ST-37** (*shangjuxu*) 🟦■; Small Intestine – **ST-39** (*xiajushu*) 🟦■; epigastrium – **ST-36** (*zusanli*) 🟦🟦; shoulder – **ST-36** (*zusanli*) 🟦🟦, **ST-38** (*tiakou*) 🟦■

Specific points according to the channel pathway (in numerical order)

- **ST-1** (*chengqi*) 🟦■: meeting point with the *ren mai, yang qiao mai*, du mai* (→ 8.1.10); local point for the eyes (→ 8.2.1)
- **ST-2** (*sibai*) 🟦■: meeting point with the *yang qiao mai** (→ 8.1.10); adjacent point for the eyes (→ 8.2.1)
- **ST-3** (*juliao*): meeting point with the *yang qiao mai* (→ 8.1.10)
- **ST-4** (*dicang*): meeting point with the Large Intestine channel, *yang qiao mai, du mai*, ren mai** (→ 8.1.10); local point for the mouth/teeth (→ 8.2.1)
- **ST-5** (*daying*): meeting point with the Gall Bladder channel
- **ST-6** (*jiache*) 🟦■*:* Sun Si Miao Ghost point (→ 8.1.15); meeting point with the Gall Bladder channel* (→ 8.1.10); local point for the mouth/teeth (→ 8.2.1)
- **ST-7** (*xiaguan*) 🟦■: meeting point with the Gall Bladder channel (→ 8.1.10); adjacent point for the teeth and jaw (8.2.1); local point for the jaw (→ 8.2.1)
- **ST-8** (*touwei*) 🟦■: meeting point with the Gall Bladder channel, *yang wei mai* (→ 8.1.10); adjacent point for frontal headaches (→ 8.2.1)
- **ST-9** (*renying*): meeting point with the Gall Bladder channel (→ 8.1.10); Sea of Qi point (→ 8.1.13); Window of Heaven point (→ 8.1.12)
- **ST-10** (*shuitu*): adjacent point for the neck (→ 8.2.1)
- **ST-12** (*quepen*): meeting point with the L.I., S.I., T.B. and G.B. channels (→ 8.1.10)
- **ST-21** (*liangmen*): adjacent point for the Spleen, Stomach and Gall Bladder (→ 8.2.1)
- **ST-25** (*tianshu*) 🟦🟦: Front-*mu* point of the Large Intestine (→ 8.1.5); adjacent point for the abdominal region (→ 8.2.1)
- **ST-26** (*wailing*): adjacent point for the abdominal region (→ 8.2.1)
- **ST-28** (*shuidao*): local point for the Small Intestine and Bladder (→ 8.2.1)
- **ST-29** (*guilai*): local point for the urogenital region (→ 8.2.1)
- **ST-30** (*qichong*) 🟦■: meeting point with the *chong mai*, G.B. channel* (→ 8.1.10); Sea of Water and Grain point (→ 8.1.13); local point for the urogenital region (8.2.1); adjacent point for the abdominal region (→ 8.2.1)
- **ST-34** (*liangqiu*) 🟦■: *xi*-cleft point, important adjacent point for disorders of the knee (→ 8.2.1)
- **ST-35** (*dubi*) 🟦■: local point for the knee region (→ 8.2.1)
- **ST-36** (*zusanli*) 🟦🟦: *he*-sea point (Earth) (→ 8.1.6), *ben*-point (Five Phase point); Lower *he*-sea point of the Stomach (→ 8.1.9); *Gao Wu* command point (→ 8.1.11) for the abdomen; Sea of Water and Grain point (→ 8.1.13); *Ma Dan Yang* Heavenly Star point (→ 8.1.14); distal point for the Spleen/Stomach and epigastrium (→ 8.2.1); local point for the knee (→ 8.2.1); distal point for the shoulder (→ 8.2.1); adjacent point for the foot (→ 8.2.1)

---

* Mentioned by only some authors.

Ch04.3-F10028.qxd 2/22/08 8:41 PM Page 127

- **ST-37** (*shangjuxu*) 🟦■: Lower *he*-sea point of the Large Intestine (→ 8.1.9); Sea of Blood point (→ 8.1.13); distal point for the Large Intestine (→ 8.2.1)
- **ST-38** (*tiakou*) 🟦■: distal point for the shoulder region (→ 8.2.1)
- **ST-39** (*xiajushu*) 🟦■: Lower *he*-sea point of the Small Intestine (→ 8.1.9); Sea of Blood point (→ 8.1.13); distal point for the Small Intestine (→ 8.2.1)
- **ST-40** (*fenglong*) 🟦■: *luo*-connecting point (→ 8.1.2); main 'Phlegm'-point; distal point for the Lung (→ 8.2.1)
- **ST-41** (*jiexi*) 🟦■: *jing*-river point (Fire) (→ 8.1.6); tonification point; local point for the foot (8.2.1); adjacent point for the toes (→ 8.2.1)
- **ST-42** (*chongyang*) 🟦■: *yuan*-source point (→ 8.1.1)
- **ST-43** (*xiangu*) 🟦■: *shu*-stream point (Wood) (→ 8.1.6)
- **ST-44** (*neiting*) 🟦🟦: *ying*-spring point (Water) (→ 8.1.6); distal point for frontal headaches, as well as for the nose, mouth/teeth and throat (→ 8.2.1); Ma Dan Yang Heavenly Star point (→ 8.1.14)
- **ST-45** (*lidui*) 🟦■: *jing*-well point (Metal) (→ 8.1.6); sedation point.

General location help

- **ST-1** to **ST-4**: with the patient looking straight ahead, these points are located on the vertical pupil line
- **ST12** to **ST-18** are all located on the midclavicular/mamillary line, 4 cun lateral to the anterior midline
    - **ST-12**: in the supraclavicular fossa
    - **ST-13**: on the lower border of the clavicle
    - **ST-14** to **ST-18**: in the first to fifth intercostal spaces
- **ST-19** to **ST-30** are all located on a vertical line midway between the mamillary line and the anterior midline, 2 cun lateral to the anterior midline:
    - **ST-19** to **ST-25** are located on the 8 cun section (→ 2.2) between the sternocostal angle and the umbilicus (**ST-19** to **ST-25** are evenly spread on this line, from 6 cun above the umbilicus to the same level as the umbilicus)
    - **ST-26** to **ST-30** are located on the 5 cun section (→ 2.2) between the umbilicus and the upper border of the pubic symphysis (1 cun steps from 4 cun superior to the upper border of the pubic symphysis to its upper border).

Container of Tears CHENGQI ST-1

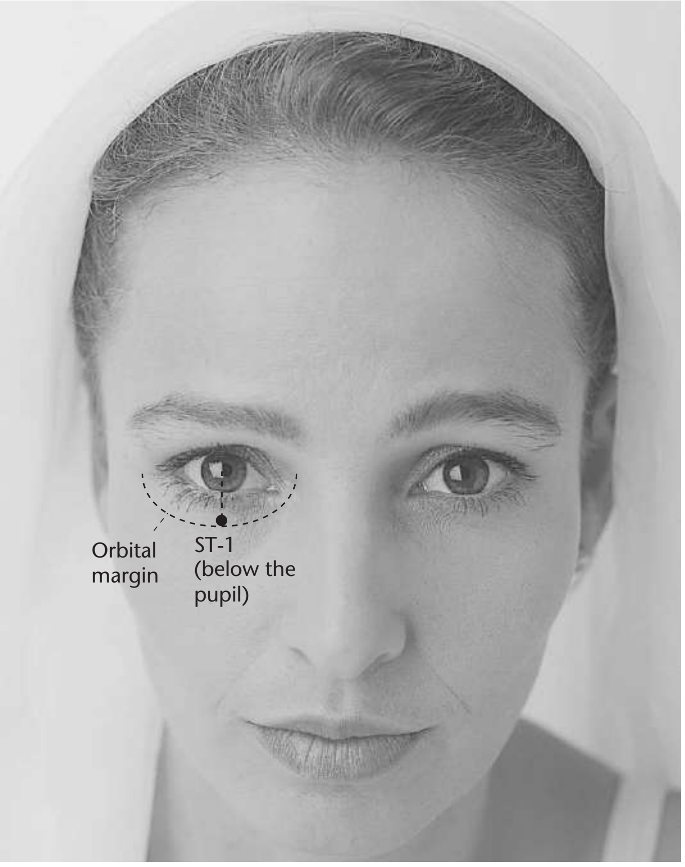
Orbital margin
ST-1 (below the pupil)

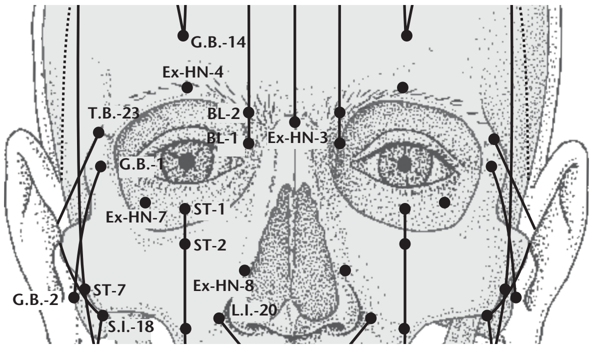
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN-8
G.B.-2
ST-7
L.I.-20
S.I.-18

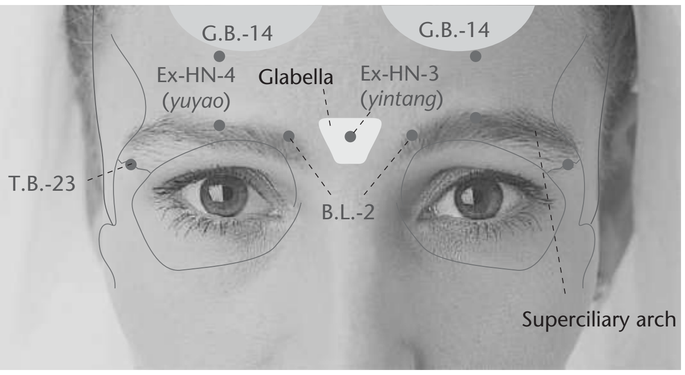
G.B.-14
G.B.-14
Ex-HN-4 (*yuyao*)
Glabella
Ex-HN-3 (*yintang*)
T.B.-23
B.L.-2
Superciliary arch

Location

With the eyes looking straight ahead, directly below the centre of the pupil, between the eyeball and the infraorbital ridge.

How to find

With the patient looking straight ahead, the first four points of the ST channel are located on a vertical line drawn through the centre of the pupil. Within the area of the lower eyelid, the infraorbital ridge presents itself as a distinct bony structure. **ST-1** is located on the pupil line, directly superior to the infraorbital ridge.

Needling

Ask the patient to look upward, gently push the eyeball upwards and insert the needle vertically along the orbital ridge in a dorsal direction. Caution: Venous plexus and arteries, avoid injuring the eyeball and periost! Needling recommended only by experienced practitioners! Pay attention to needling pain! No needle manipulation! After removing the needle, press firmly on the insertion site for about one minute to prevent haematoma. Inform the patient that it may still be possible for a haematoma to develop. Moxibustion is contraindicated! Points for disorders of the eyes with less risk of complications include **BL-2**, **T.B.-21**, **G.B.-1**, **ST-2**, **Ex-HN-5** (*taiyang*), **EX-HN-4** (*yuyao*).

Actions/Indications

- Expels Wind and Heat, clears and benefits the eyes

Special features

Meeting point with the *yang qiao mai* and *ren mai*, also with the *du mai* according to some authors, important point for the eyes, entry point.

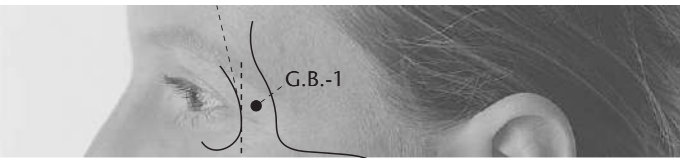

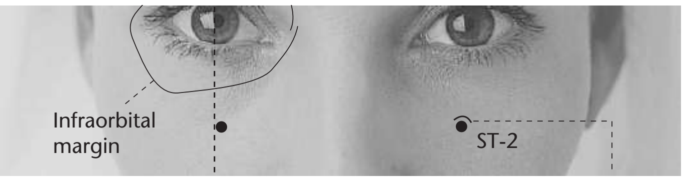
Infraorbital margin
ST-2

Ch04.3-F10028.qxd 2/22/08 8:41 PM Page 130

ST-2 *Four Whites* SIBAI

Location

With the eyes looking straight ahead, directly below the centre of the pupil, in the depression at the infraorbital foramen.

How to find

With the patient looking straight ahead, the first four points of the ST channel are located on a vertical line drawn through the centre of the pupil. Within the area of the lower eyelid, the infraorbital ridge presents itself as a distinct bony structure. Starting at the infraorbital ridge (➞ 3.1.2), palpate inferiorly until you can feel the depression of the infraorbital foramen (this is often located slightly medial to the pupil line), where **ST-2** is located.

Needling

0.3–0.5 cun vertically or transversely (subcutaneously) towards ➞ **S.I.-18** or ➞ **L.I.-20** (for example, in cases of facial paralysis). Deep needling (in an oblique cranial direction) is contraindicated. Caution: Infraorbital nerve, eye injury. According to some classic texts, moxibustion is contraindicated.

Actions/Indications

- Benefits the eyes, clears Heat, eliminates Wind

Special features

Important local point for disorders of the eyes, pain and paralysis of the face; meeting point with the *yang qiao mai* according to some authors. A less risky substitute for ➞ **ST-1**.

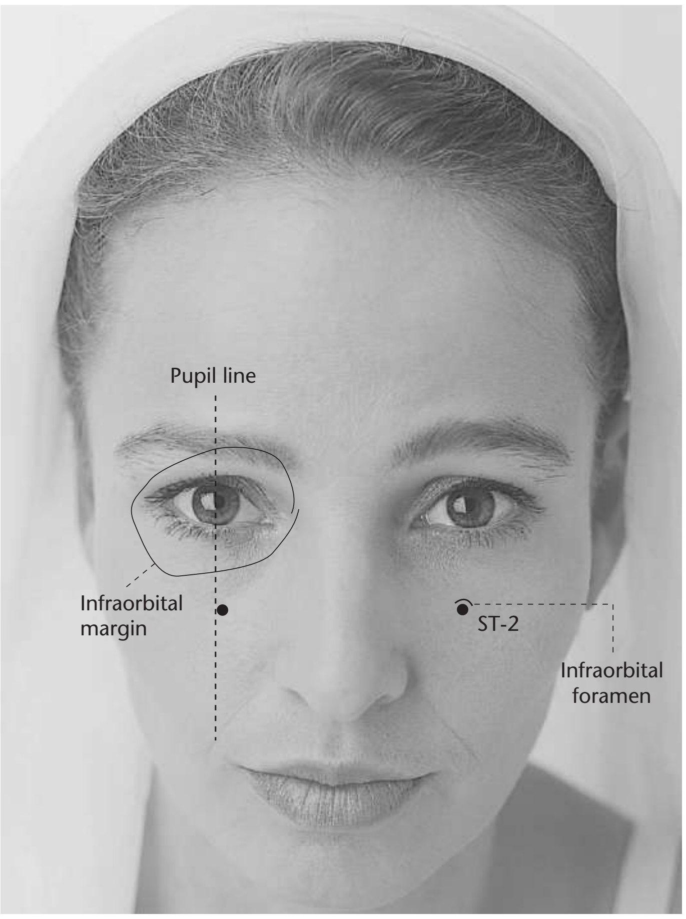

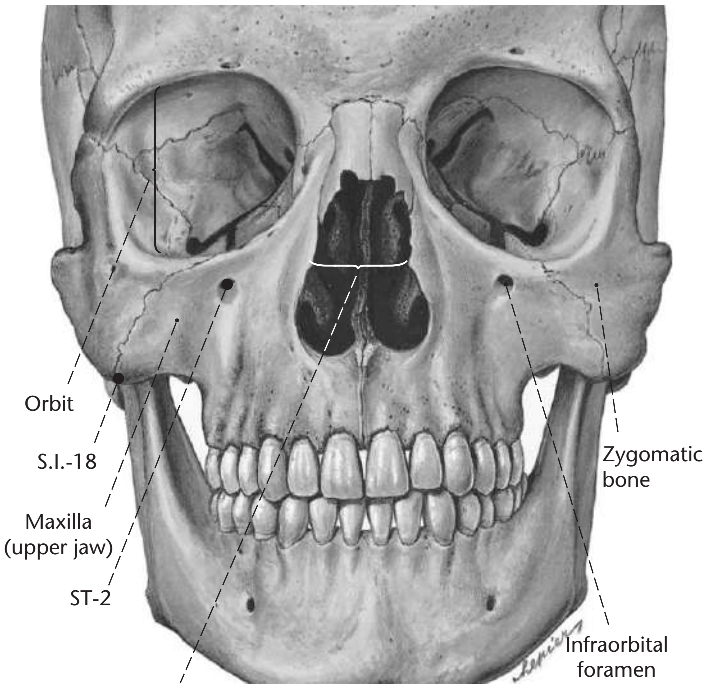

Great Crevice JULIAO ST-3

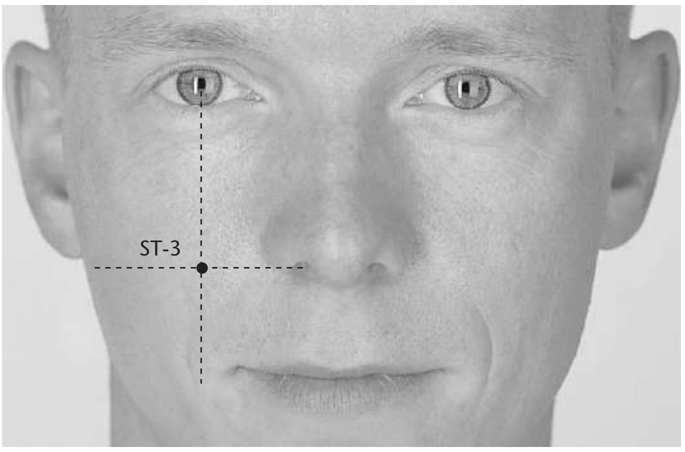

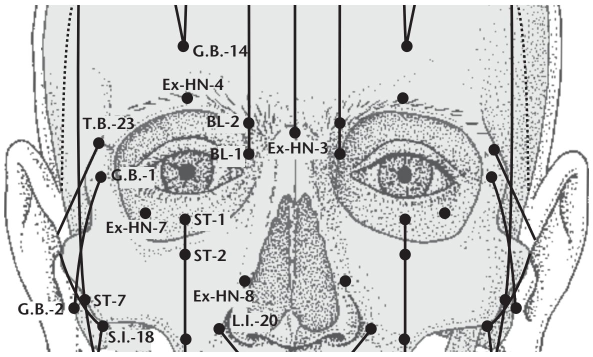
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
G.B.-2
ST-7
Ex-HN-8
L.I.-20
S.I.-18

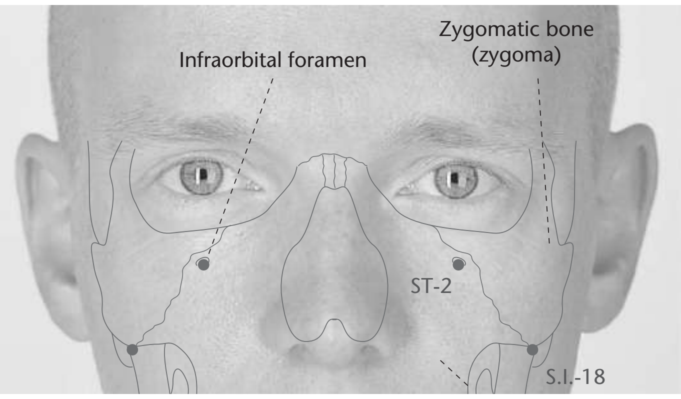
Infraorbital foramen
Zygomatic bone
(zygoma)
ST-2
S.I.-18

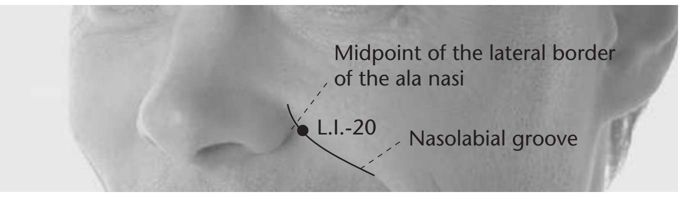
Midpoint of the lateral border
of the ala nasi
L.I.-20
Nasolabial groove

Location

With the patient looking straight ahead, directly below the centre of the pupil, on the level of the lower border of the ala nasi.

How to find

With the patient looking straight ahead, the first four points of the ST channel are located on a vertical line drawn through the centre of the pupil (pupil line). **ST-3** is located at the crossing point of the pupil line and a horizontal line on the level of the lower border of the ala nasi.

Needling

Vertically or obliquely 0.3–0.5 cun or penetration method: transversely (subcutaneously) towards → **ST-4,** → **S.I.-18**, etc, for example, in cases of peripheral facial paralysis.

Actions/Indications

- Expels Wind, opens the channel, disperses accumulations

Special features

Meeting point with the *yang qiao mai*

ST-4 *Earth Granary* DICANG

Location

With the patient looking straight ahead, directly below the centre of the pupil and approximately 0.4 cun lateral to the corner of the mouth.

How to find

With the patient looking straight ahead, the first four points of the ST channel are located on a vertical line drawn through the centre of the pupil (pupil line). **ST-4** is located at the crossing point of the pupil line and a horizontal line originating at the corner of the mouth. The point is located in the nasolabial groove (→ 3.1.2), approximately 0.4 cun lateral to the corner of the mouth. Ask the patient to smile in order to make the line more clearly visible.

Needling

Obliquely in a lateral direction or vertically 0.3–0.5 cun or penetration method: transversely (subcutaneously) approximately 1.5–2 cun in cases of facial paralysis towards → **ST-6**, for trigeminal neuralgia towards → **L.I.-20**. Caution: Facial artery/vein.

Actions/Indications

- Expels Wind (from the face), opens the channel, alleviates pain, relaxes the facial musculature
- Sometimes as distal point for disorders of the leg

Special features

Meeting point with the L.I. channel, the *ren mai* and *yang qiao mai* (also with the *du mai* according to some authors). Important local point for neuralgias and pareses affecting the mouth and cheeks.

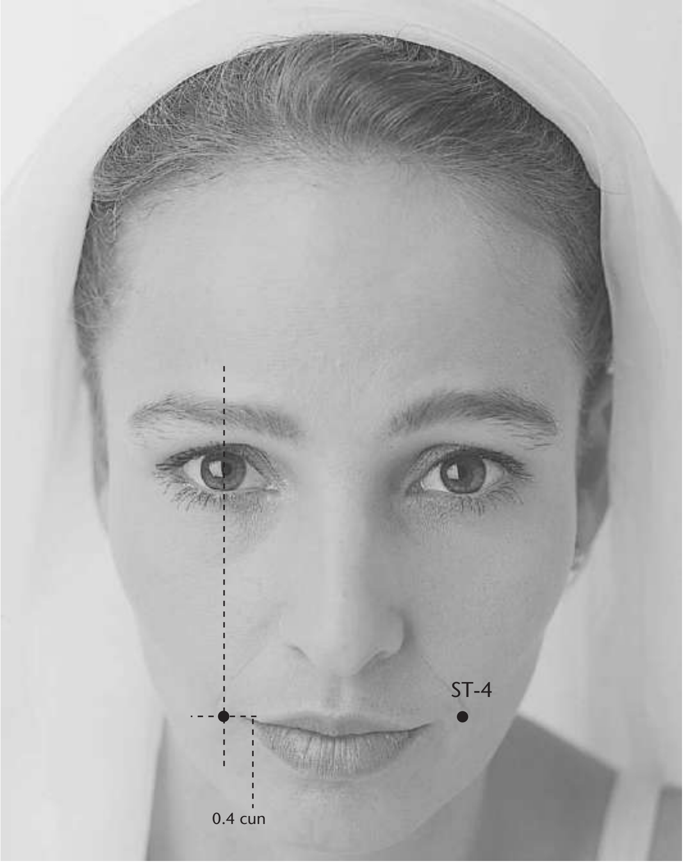

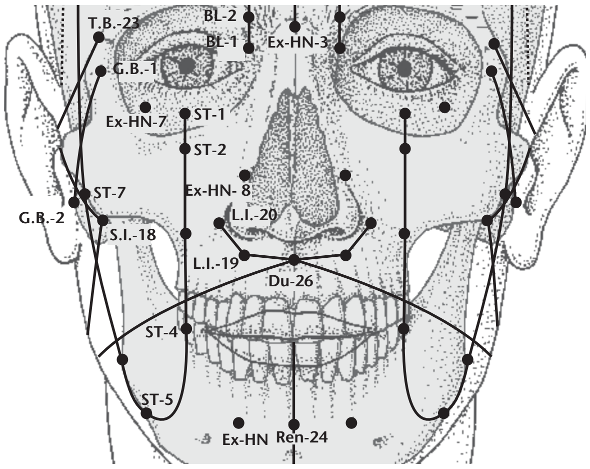

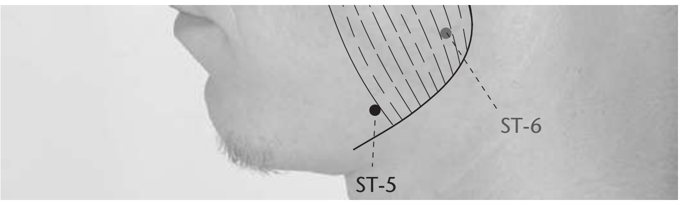

Great Welcome DAYING ST-5

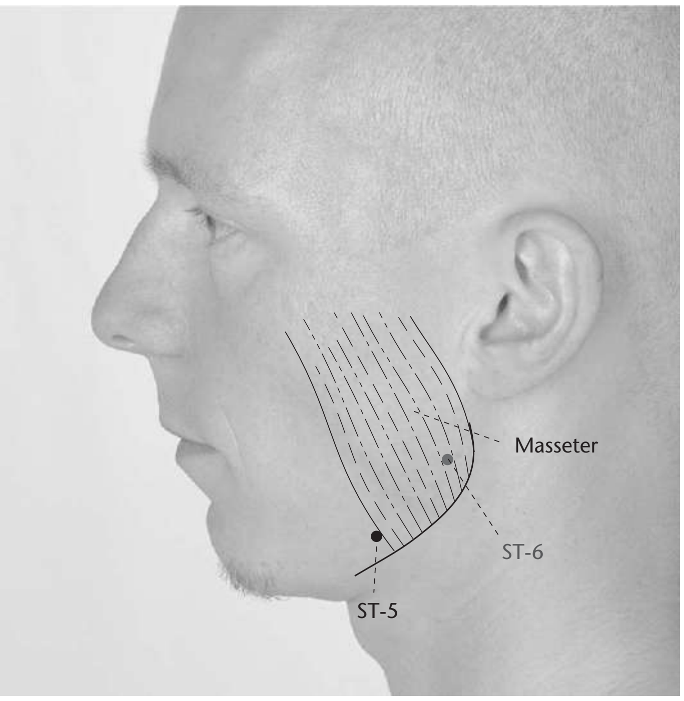
Masseter
ST-6
ST-5

Location

On the lateral mandible, anterior to the masseter muscle.

How to find

By asking the patient to clench their teeth, the anterior border of the masseter muscle becomes clearly palpable. **ST-5** is located directly anterior to the border of the muscle, in a shallow depression on the mandible and slightly above the angle of the jaw. Here, the facial artery can be palpated.

Needling

0.3–0.5 cun obliquely or transversely (subcutaneously) in the direction of → **ST-6**. Caution: Facial artery/vein.

Actions/Indications

- Expels Wind, opens the channels

Special features

Meeting point with the G.B. channel according to some authors

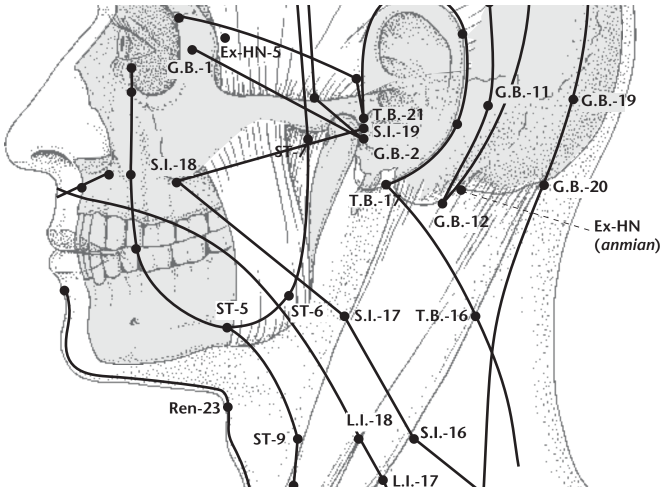
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
ST-7
T.B.-17
G.B.-20
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
S.I.-16
ST-9
L.I.-17

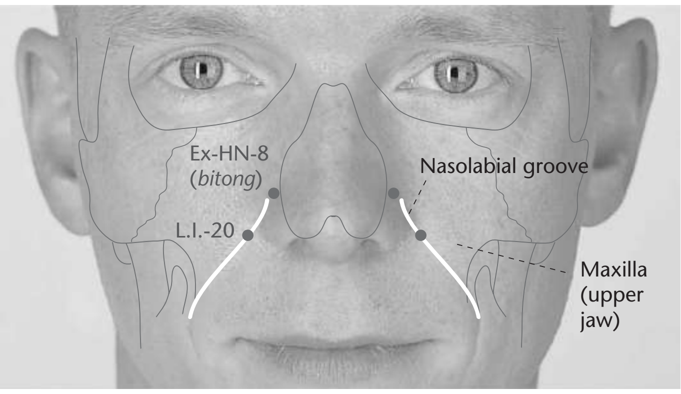
Ex-HN-8
(*bitong*)
Nasolabial groove
L.I.-20
Maxilla
(upper
jaw)

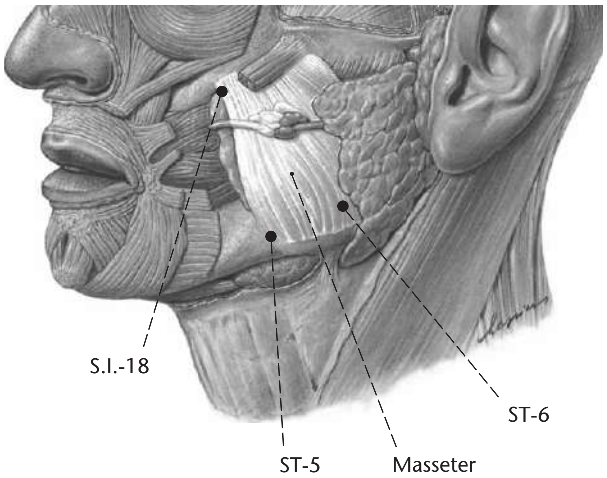
S.I.-18
ST-6
ST-5
Masseter

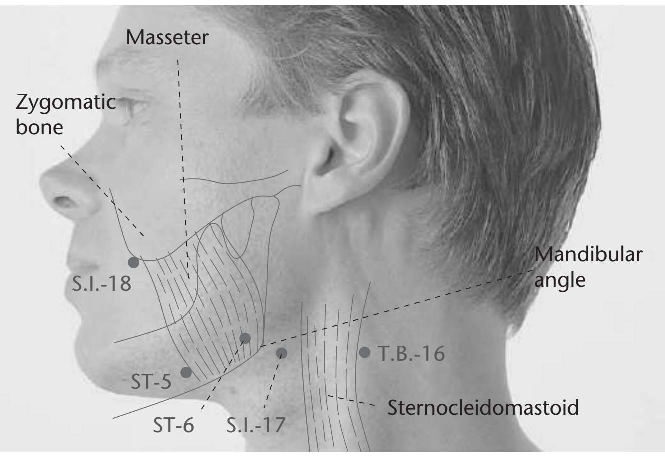
Masseter
Zygomatic
bone
S.I.-18
Mandibular
angle
T.B.-16
ST-5
Sternocleidomastoid
ST-6
S.I.-17

Ch04.3-F10028.qxd 2/22/08 8:41 PM Page 134

ST-6 *Jaw Bone* JIACHE

Location

With the teeth clenched, on the highest prominence of the masseter muscle, approximately one finger-width (middle finger) anterior and superior to the angle of the jaw.

How to find

Ask the patient to clench their teeth, which will allow the muscle belly of the masseter muscle to bulge. **ST-6** is located on its prominence, a little anterior and superior to the angle of the jaw (mandibular angle → 3.1.3). With chewing movements of the jaw ('biting and relaxing'), the palpating finger will naturally fall into the right depression when the muscle is relaxed.

→ **ST-5** is located more medially in the depression at the anterior border of the masseter muscle.

Needling

Insertion with the patient's mouth slightly opened (relaxed masseter muscle). Vertically 0.3–0.5 cun or obliquely or transversely (subcutaneously) 1–1.5 cun in the direction of → **ST-4**, **ST-5**, **ST-7**, for example, in cases of facial paralysis.

Actions/Indications

- Eliminates Wind, benefits the jaw and teeth, opens the channel and *luo* vessels, alleviates pain

Special features

Sun Si Miao Ghost point. Important local point for the jaw region. Its location corresponds to a commonly used trigger point on the masseter muscle. According to some authors, **ST-6** is a meeting point with the G.B. channel.

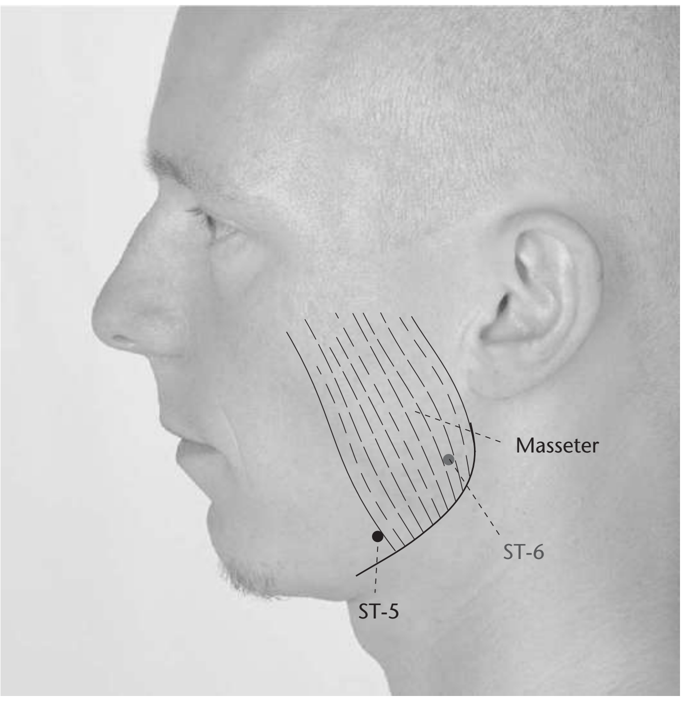
Masseter
ST-6
ST-5

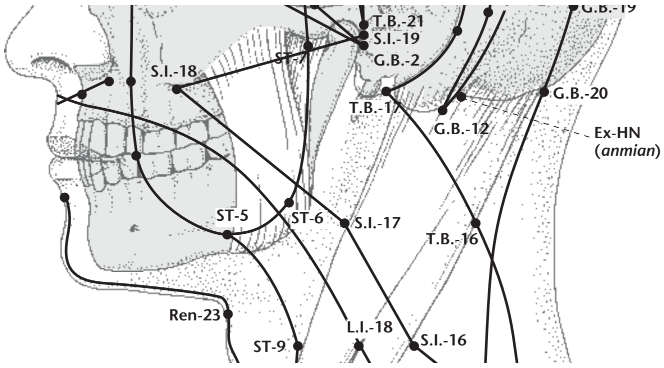
G.B.-19
T.B.-21
S.I.-19
G.B.-2
ST-7
S.I.-18
T.B.-17
G.B.-20
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16

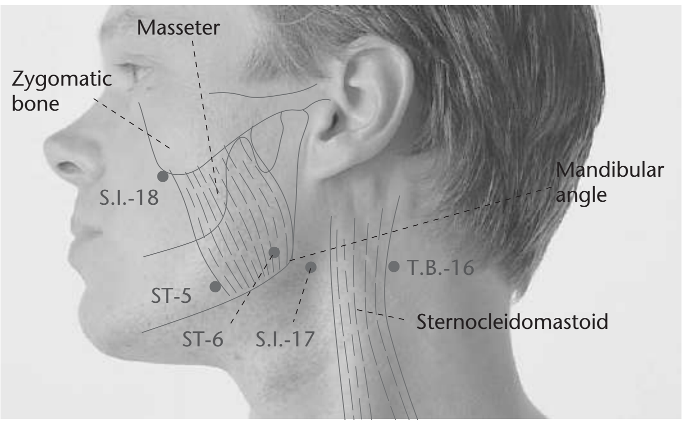
Masseter
Zygomatic bone
S.I.-18
Mandibular angle
T.B.-16
ST-5
ST-6
S.I.-17
Sternocleidomastoid

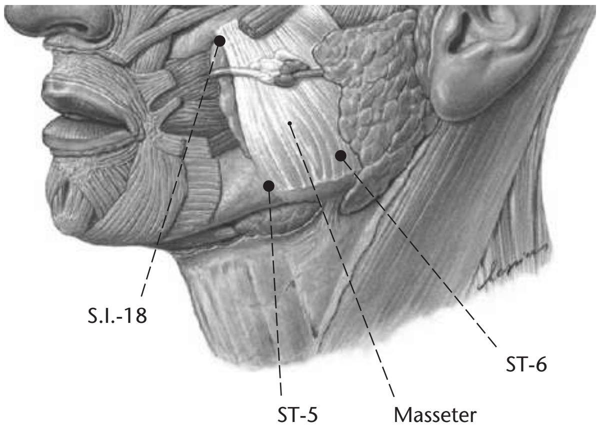
S.I.-18
ST-6
ST-5
Masseter

Below the Joint XIAGUAN ST-7

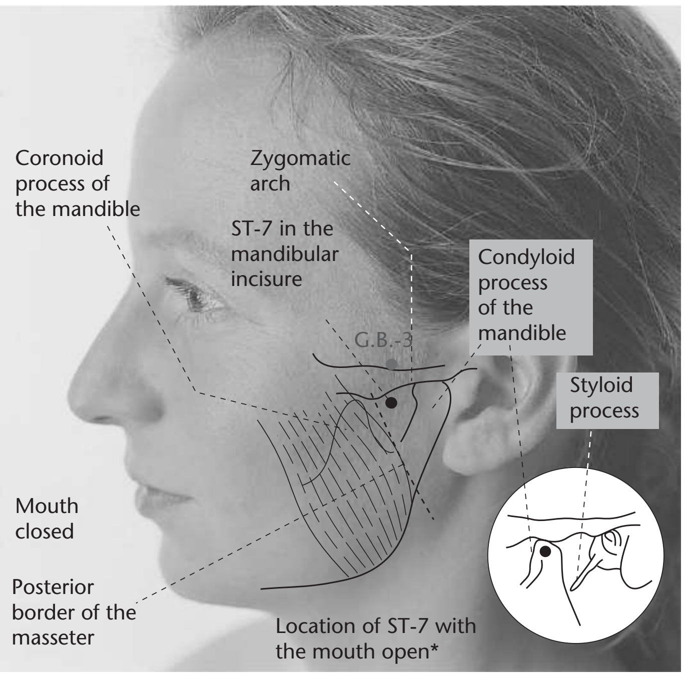
Coronoid process of the mandible
Zygomatic arch
ST-7 in the mandibular incisure
G.B.-3
Condyloid process of the mandible
Styloid process
Mouth closed
Posterior border of the masseter
Location of ST-7 with the mouth open*

Location

With the mouth closed, at the lower border of the zygomatic arch (→ 3.1.2) in the centre of the depression of the mandibular fossa, between the coronoid process and the condyloid process of the mandible.

How to find

Palpate along the lower border of the zygomatic arch (→ 3.1.2) towards the ear. **ST-7** is located in a clearly palpable depression just before the temporomandibular joint and at the posterior border of the masseter muscle (clench teeth). To confirm: When the patient's mouth is fully opened, the condyloid process of the lower jaw slides anteriorly and the depression disappears (figure modified after Deadman et al).

→ **G.B.-3** is located directly superior to **ST-7**, on the upper border of the zygomatic arch.

Needling

Insert the needle with the patient's mouth closed. 0.3–0.5 cun vertically or 1–1.5 cun transversely (subcutaneously) in the direction of → **ST-6**, → **S.I.-19**, → **S.I.-18**.

Actions/Indications

- Opens the channel, benefits the jaw and teeth, alleviates pain
- Benefits the ears

Special features

Meeting point with the G.B. channel. Important local point and frequently used trigger point.

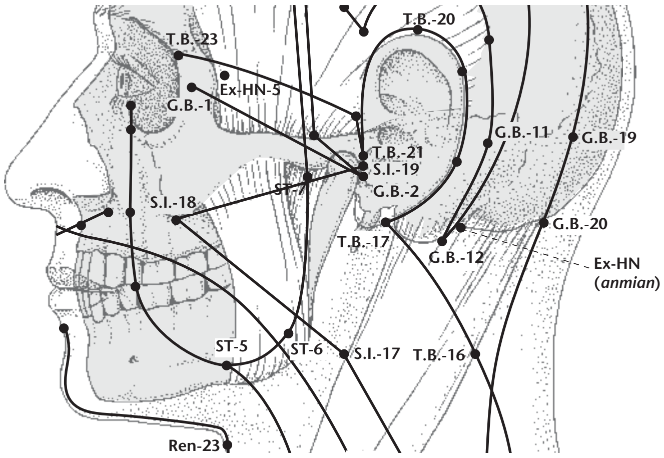
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
ST-7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23

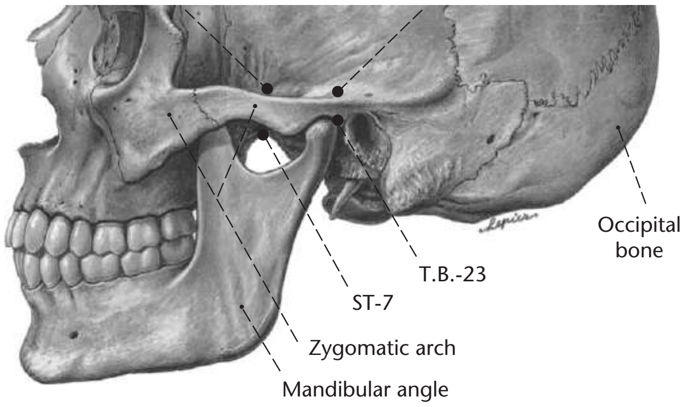
Occipital bone
T.B.-23
ST-7
Zygomatic arch
Mandibular angle

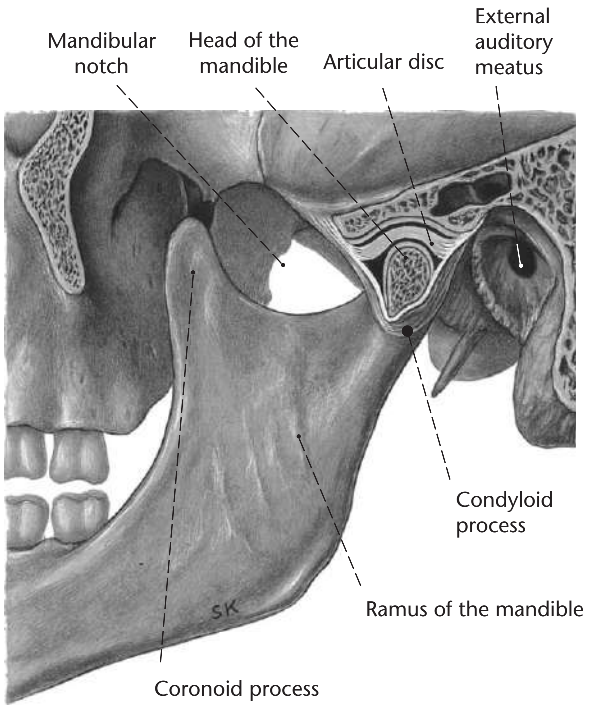
Mandibular notch
Head of the mandible
Articular disc
External auditory meatus
Condyloid process
Ramus of the mandible
Coronoid process

---

* According to Deadman et al 1998.

ST-8 *Head's Binding* TOUWEI

Location

At the temporal corner of the forehead, on the border of the temporalis muscle and 0.5 cun within the anterior hairline or 4.5 cun lateral to the anterior midline (→ **Du-24**).

How to find

First, find the anterior hairline (→ 3.1.1). Then locate **ST-8** at the temporal corner of the forehead 0.5 cun within the hairline. Chewing movements of the temporalis muscle will still be palpable here.

Also located 0.5 cun superior to the anterior hairline are → **Du-24/BL-3/BL-4/G.B.-15/G.B.-13** (anterior midline/superior to the inner canthus of the eye/1.5 cun lateral to the anterior midline/3 cun lateral to the anterior midline). These distances refer to the distance between → **Du-24** and **ST-8** (=4.5 cun, → 2.2). **ST-8** and → **G.B.-7** (a depression on the level of the apex of the ear, within the temporal hairline) mark the endpoints of a curved line, on which **ST-8 and** → **G.B.-4–7** are located at regular intervals (by dividing the line into five parts).

Needling

0.5–1 cun transversely (subcutaneously) or obliquely in the direction of the disorder. Moxibustion possible (contraindicated according to some classical texts).

Actions/Indications

- Eliminates Wind from the head and eyes, benefits the eyes, alleviates pain

Special features

Meeting point with the *yang wei mai* and the G.B. channel. Important local point for headaches and disorders of the eyes.

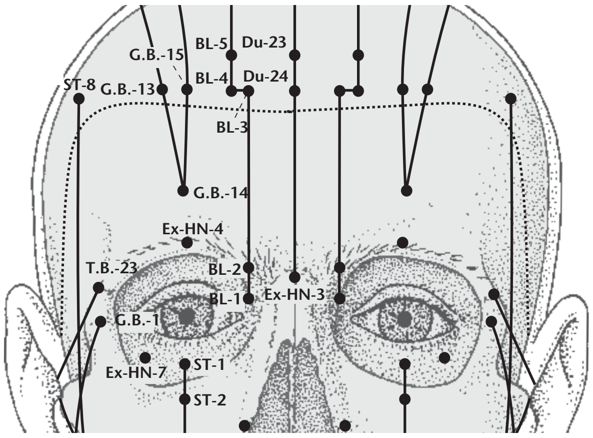
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2

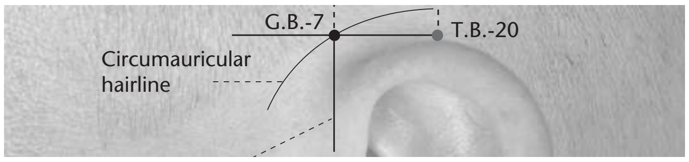
G.B.-7
T.B.-20
Circumauricular
hairline

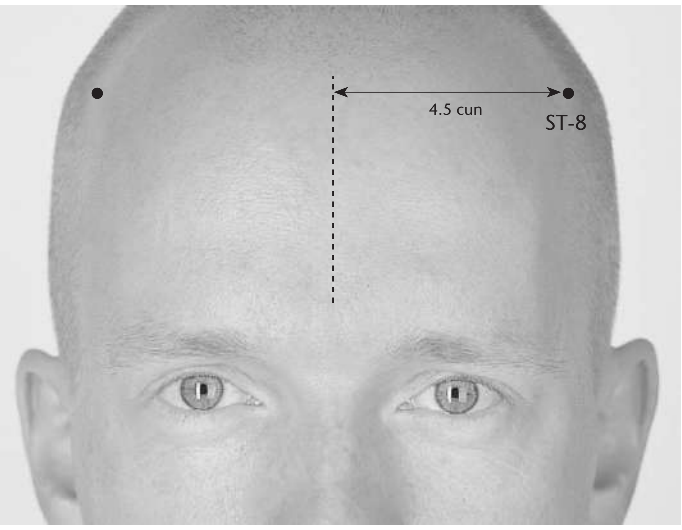

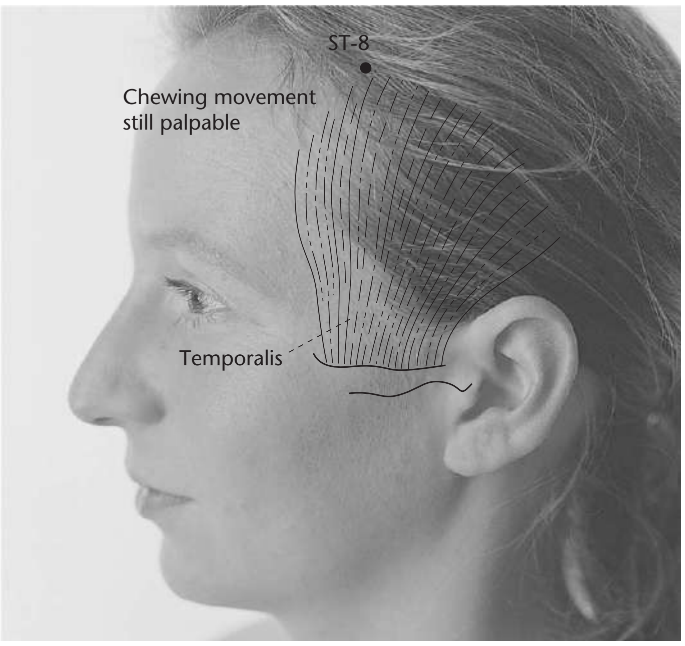
ST-8
Chewing movement
still palpable
Temporalis

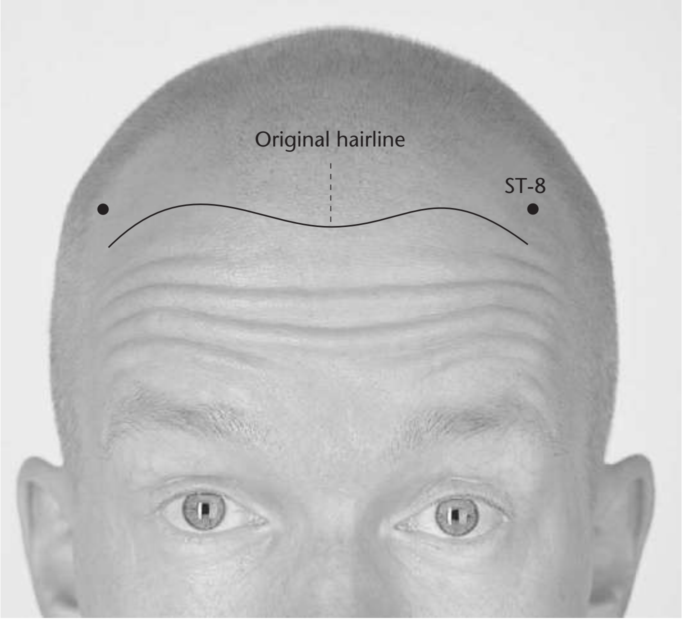
Original hairline
ST-8

Man's Welcome **RENYING** **ST-9**

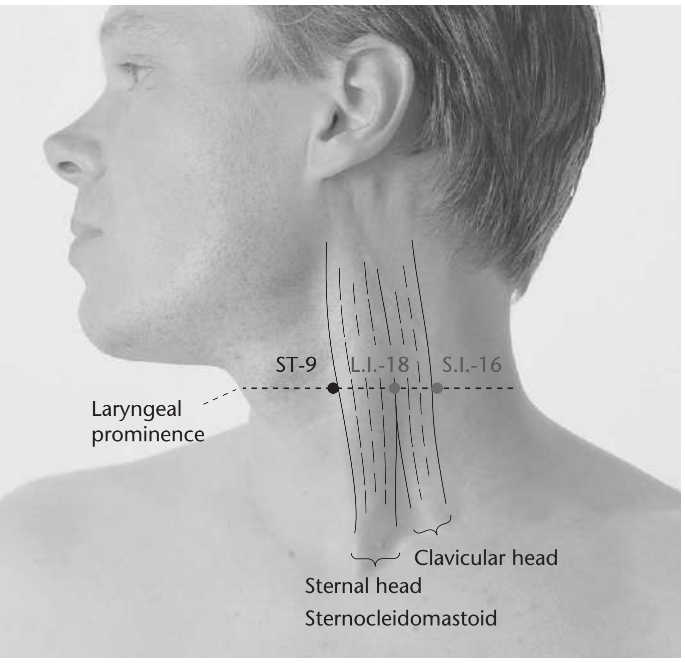
ST-9
L.I.-18
S.I.-16
Laryngeal prominence
Clavicular head
Sternal head
Sternocleidomastoid

Location

1.5 cun lateral to the anterior midline, on the level of the laryngeal prominence and at the anterior border of the sternocleidomastoid muscle.

How to find

From the laryngeal prominence, palpate posteriorly to the anterior border of the sternocleidomastoid muscle (→ 3.2). **ST-9** is located at the anterior border of the muscle, which becomes more clearly visible and palpable if the patient's head is turned in the opposite direction.

Located on the same level are → **S.I.-16** on the posterior border of the sternocleidomastoid muscle and **L.I.-18** between the two heads of the sternocleidomastoid muscle.

Needling

If necessary, move the head to its normal position before needling. Caution: Carotid artery, carotid sinus, jugular veins. The carotid artery is located directly at the anterior border of the sternocleidomastoid muscle. The needle should be inserted anterior to where you can feel the pulsating of the carotid artery. Vertically 0.5–1 cun between the artery and the thyroid cartilage. For reasons of safety, the artery should be held in a lateral position with one finger of the palpating hand above and one finger below the insertion site. Irritation of the carotid sinus can lead to a lowering of the blood pressure and fainting!

Actions/Indications

- Regulates the Qi
- Opens the channel
- Harmonises excess or deficiency in the thorax

Special features

Window of Heaven point, Sea of Qi point, meeting point with the G.B. channel.

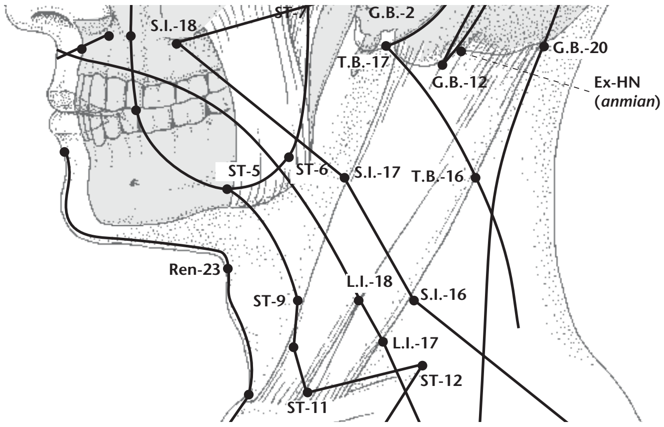
S.I.-18
ST-7
G.B.-2
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
S.I.-16
ST-9
L.I.-17
ST-12
ST-11

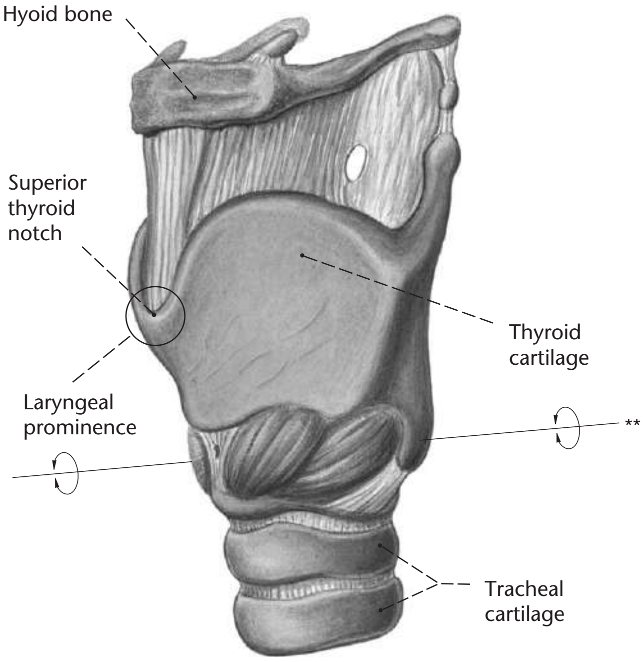
Hyoid bone
Superior thyroid notch
Thyroid cartilage
**
Laryngeal prominence
Tracheal cartilage

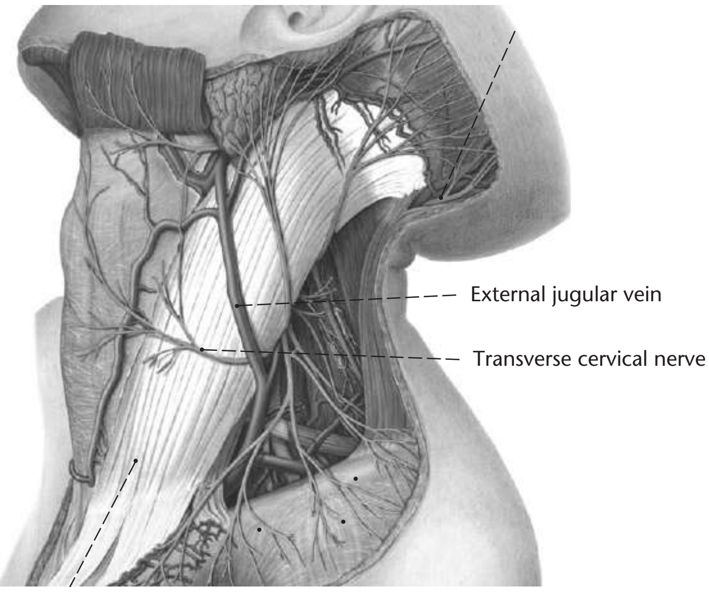
External jugular vein
Transverse cervical nerve

ST-10 *Water Prominence* SHUITU

Location

On the anterior border of the sternocleidomastoid muscle, at the midpoint of a line connecting → **ST-9** and **ST-11**.

How to find

→ **ST-9** and **ST-10** are both located on the anterior border of the sternocleidomastoid muscle (→ 3.2). → **ST-11** is located posterior to the sternal head. Place the middle finger of the left hand onto → **ST-9** on the level of the laryngeal prominence and the thumb on → **ST-11** posterior to the sternal insertion. **ST-10** is located at the midpoint between the two fingers. With the patient's head rotated in the opposite direction, the muscle will be more visible and can be more easily palpated.

Needling

0.5–1 cun obliquely in a medial direction, away from the artery. Caution: Carotid artery, jugular veins. If necessary, rotate the head back to a middle position before needling.

Actions/Indications

- Regulates the Lung Qi
- Benefits the throat

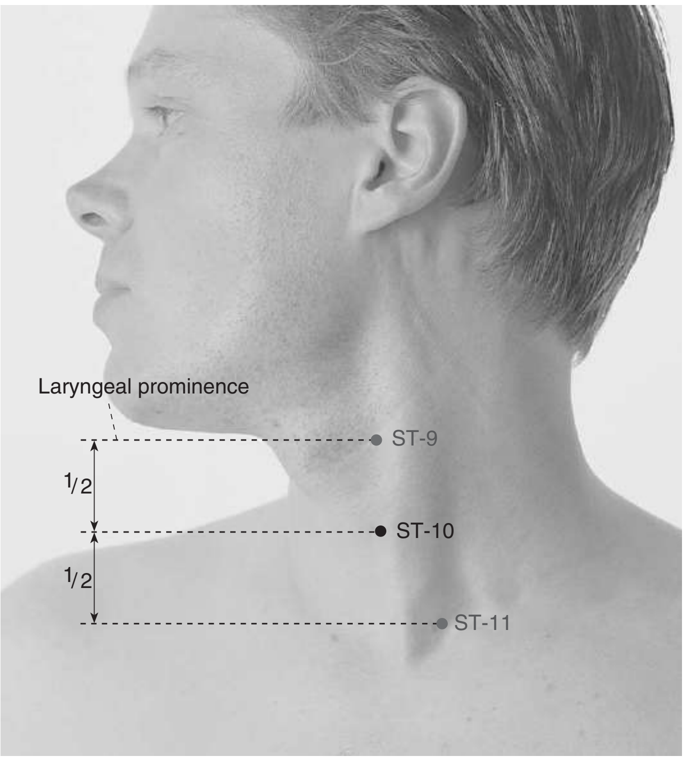
Laryngeal prominence
ST-9
1/2
ST-10
1/2
ST-11

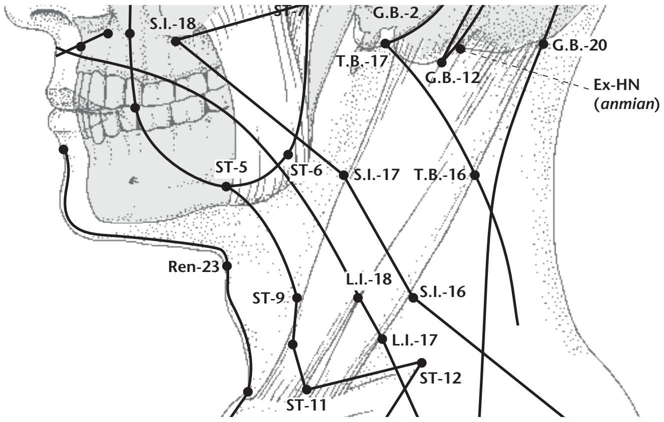
S.I.-18
ST-7
G.B.-2
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(anmian)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

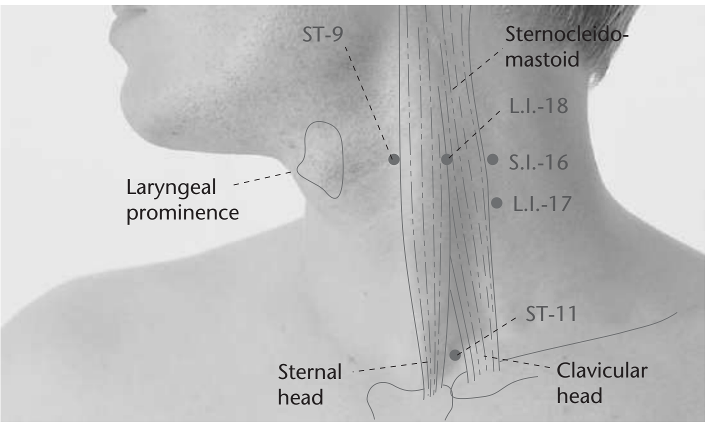
ST-9
Sternocleido-
mastoid
L.I.-18
S.I.-16
L.I.-17
Laryngeal
prominence
ST-11
Sternal
head
Clavicular
head

Qi Abode QISHE ST-11

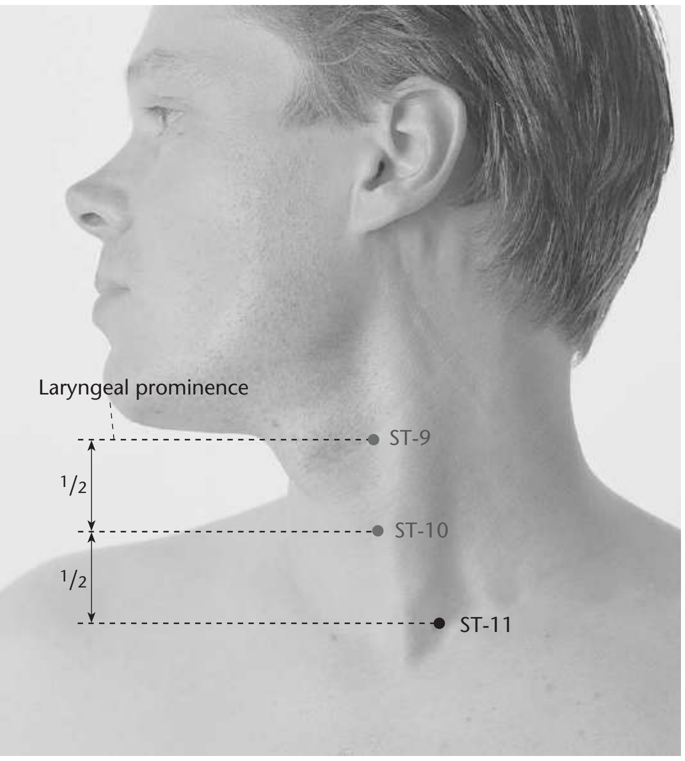
Laryngeal prominence
ST-9
1/2
ST-10
1/2
ST-11

Location

On the upper border of the clavicle, between the tendons of the sternal and clavicular heads of the sternocleidomastoid muscle.

How to find

By rotating the patient's head in the opposite direction (especially against resistance), the ropey sternal insertion of the sternocleidomastoid muscle (→ 3.2) becomes clearly visible. **ST-11** is located in a shallow depression, the minor supraclavicular fossa, between the sternal insertion and the more extended insertion of the clavicular head.

Needling

Vertically 0.3–0.5 cun. Caution: Pneumothorax, jugular veins.

Actions/Indications

- Benefits the throat and neck
- Regulates counterflow Qi

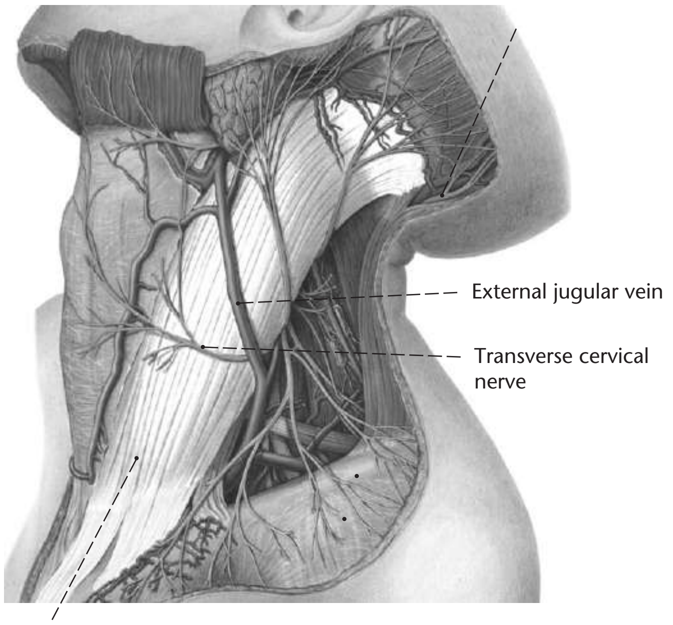
External jugular vein
Transverse cervical nerve

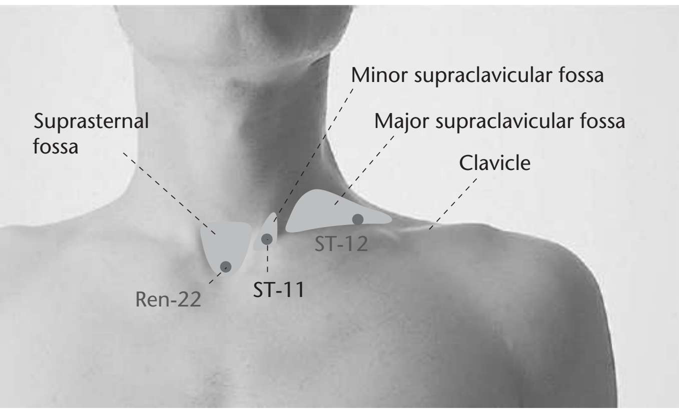
Minor supraclavicular fossa
Suprasternal fossa
Major supraclavicular fossa
Clavicle
ST-12
Ren-22
ST-11

ST-12 *Empty Basin* QUEPEN

Location

In the supraclavicular fossa, superior to the midpoint of the clavicle, approximately 4 cun lateral to the anterior midline.

How to find

From the anterior midline, palpate to the midpoint of the clavicle (=4 cun in a lateral direction). There, an imaginary vertical line marks the midclavicular line. **ST-12** is located on the midclavicular line, at the upper border of the clavicle, in the supraclavicular fossa, lateral to the clavicular head of the sternocleidomastoid muscle.

Needling

Vertically 0.3–0.5 cun. The needle should be guided along the posterior border of the clavicle. Caution: Pneumothorax, especially with emphysema; superficial and deep cervical arteries. According to some authors, this point is contraindicated during pregnancy.

Actions/Indications

- Opens the channel
- Disperses and descends the Lung Qi
- Clears Heat in the throat

Special features

Meeting point with the L.I., S.I., G.B. and T.B. channel

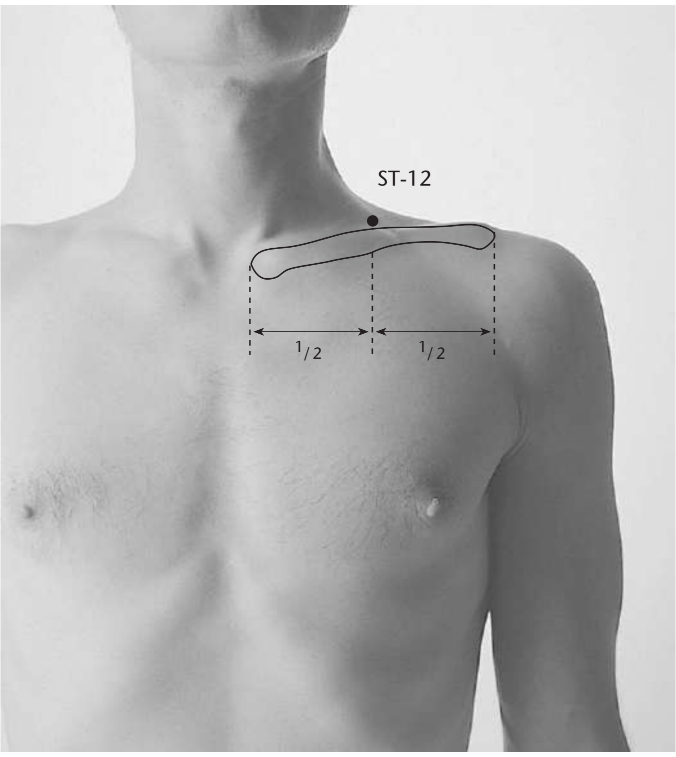

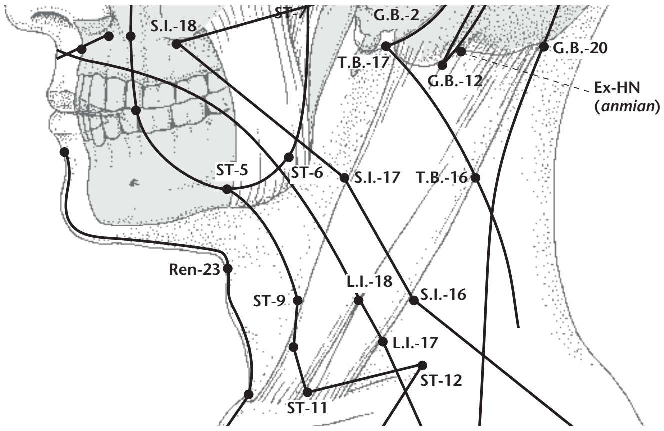
S.I.-18
G.B.-2
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(anmian)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
ST-9
S.I.-16
L.I.-17
ST-12
ST-11

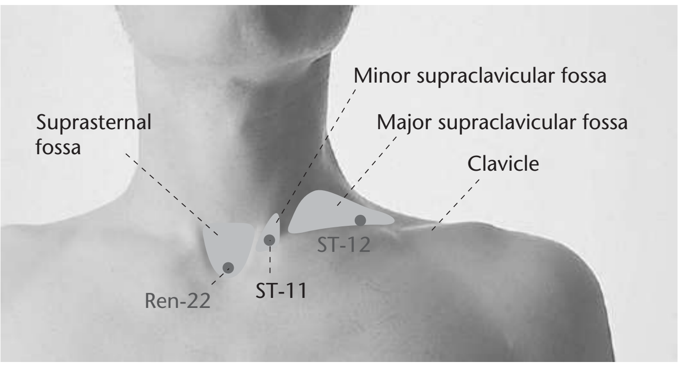
Minor supraclavicular fossa
Suprasternal fossa
Major supraclavicular fossa
Clavicle
ST-12
ST-11
Ren-22

Qi Door **QIHU** **ST-13**

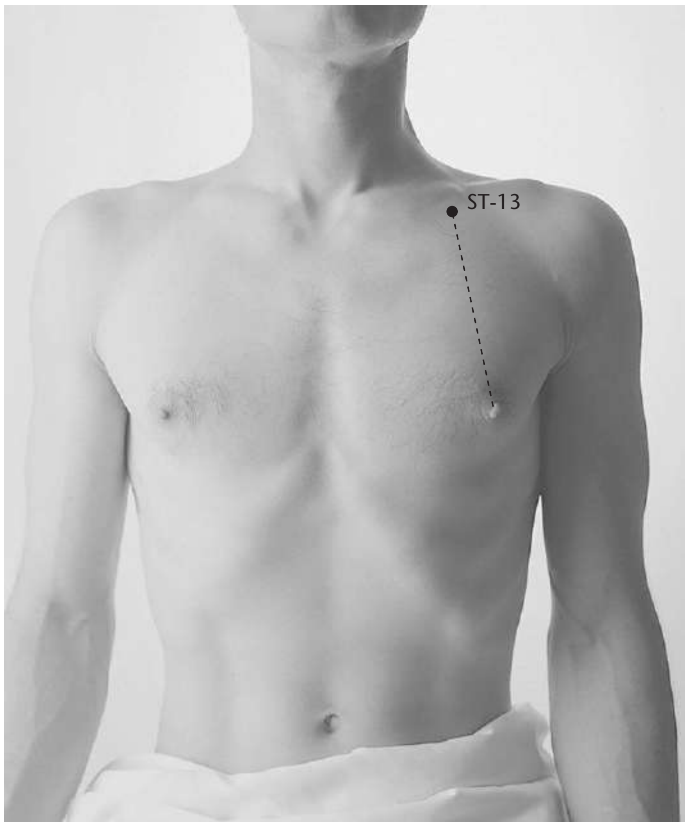

Location

At the midpoint of the clavicle and on its inferior border, 4 cun lateral to the anterior midline.

How to find

On the upper thorax region, the midclavicular line (4 cun lateral to the anterior midline) runs slightly obliquely from the clavicular midpoint to the generally more lateral nipple. Find the midpoint of the clavicle (or measure 4 cun lateral to the midline) and locate **ST-13** at the lower border of the clavicle. Since the 1st rib curves posteriorly in a short bow, **ST-13** is generally situated lateral to the 1st rib.

→ **KID-27** is located medial to **ST-13**, but inferior to the lower border of the clavicle and 2 cun lateral to the anterior midline. → **Ren-21** is located on the midline. → **ST-14** is located only slightly inferior to **ST-13** in the 1st intercostal space.

Needling

0.3–0.5 cun obliquely in a lateral or medial direction or transversely (subcutaneously) along the channel pathway. Caution: Pneumothorax, subclavicular artery/vein.

Actions/Indications

- Descends counterflow Qi
- Clears Heat
- Unbinds the chest

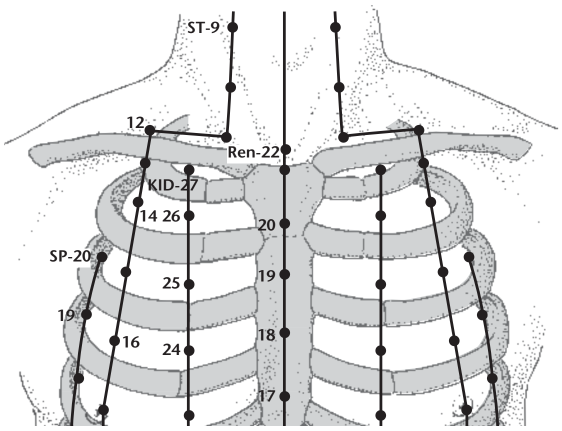

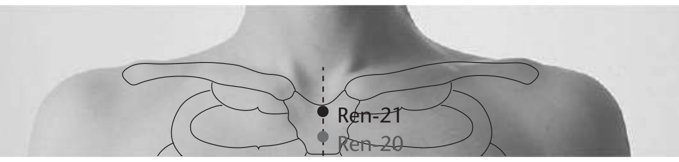

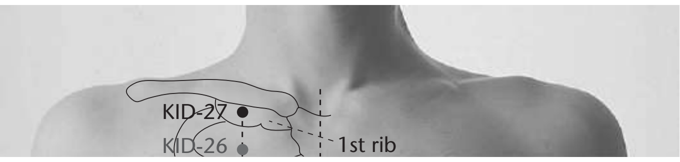

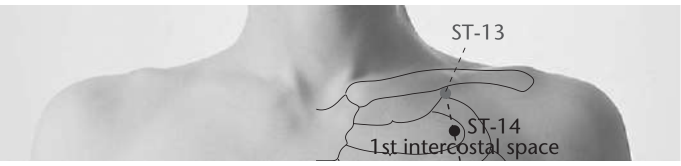

ST-14 *Store House* KUFANG

Location

On the midclavicular line, in the first intercostal space, 4 cun lateral to the anterior midline.

How to find

On the upper thorax, the midclavicular line (4 cun lateral to the anterior midline) runs slightly obliquely from the clavicular midpoint to the generally more lateral nipple. By palpating across the clavicle from superior to inferior, the first rib will generally be felt directly below the clavicle, although, in some cases, the 1st rib might be completely hidden under the clavicle. The 1st intercostal space follows directly inferior to the 1st rib, where **ST-14** is located.

→ **ST-13** is located only slightly superior to **ST-14** at the lower border of the clavicle. Also located on the level of the first intercostal space are → **Ren-20** (anterior midline), → **KID-26** (2 cun lateral to the anterior midline) and → **LU-1** (6 cun lateral to the midline). For more detail about orientation regarding intercostal spaces → 3.5.

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the course of the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Qi
- Unbinds the chest
- Benefits the breasts

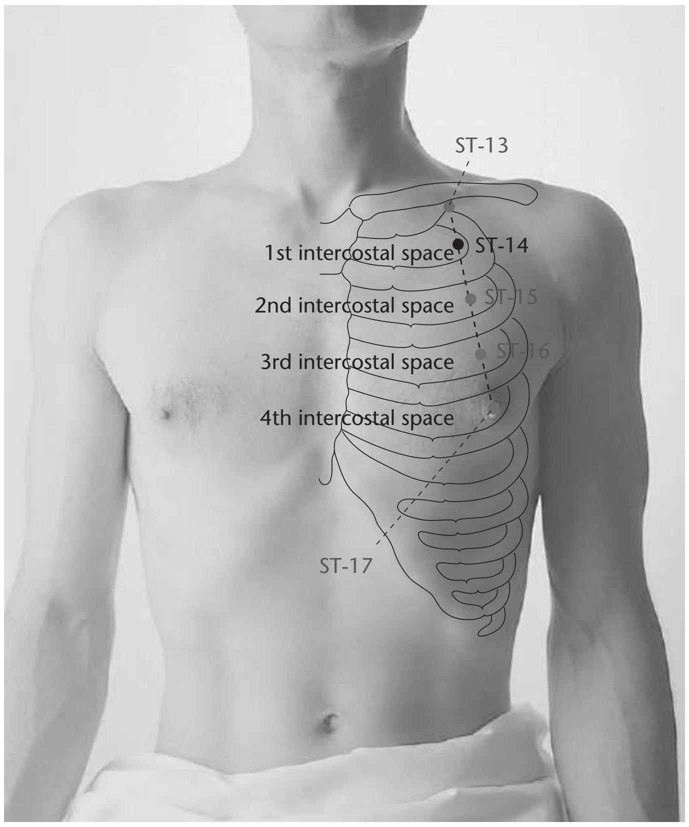
ST-13
1st intercostal space
ST-14
2nd intercostal space
ST-15
3rd intercostal space
ST-16
4th intercostal space
ST-17

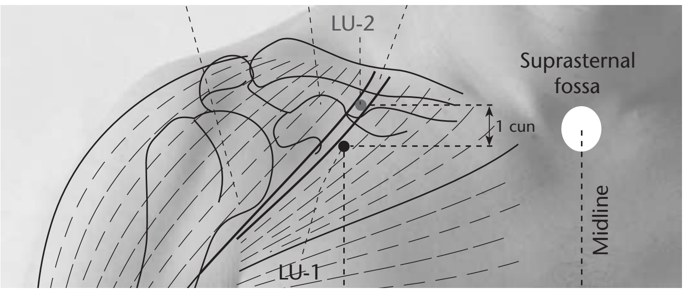
LU-2
Suprasternal fossa
1 cun
Midline
LU-1

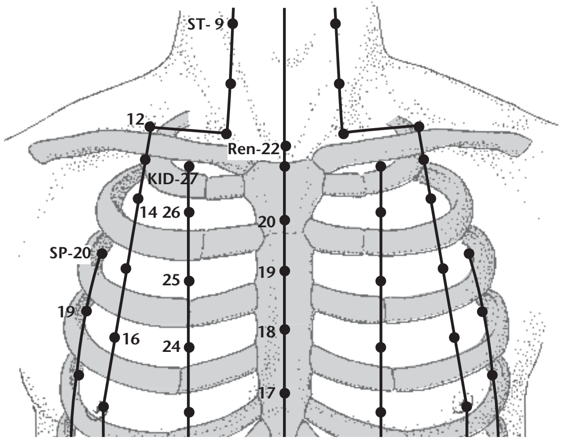
ST- 9
12
Ren-22
KID-27
14 26
20
SP-20
25
19
19
16
24
18
17

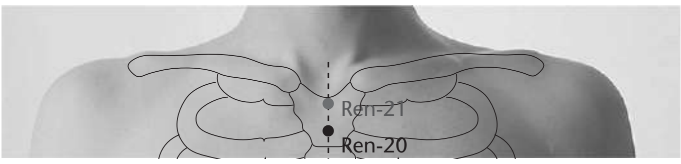
Ren-21
Ren-20

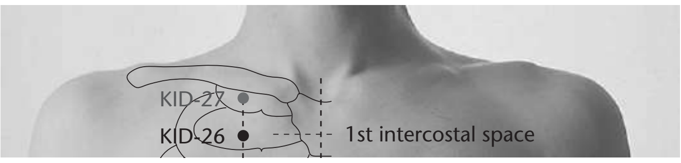
KID-27
KID-26
1st intercostal space

Room Screen WUYI ST-15

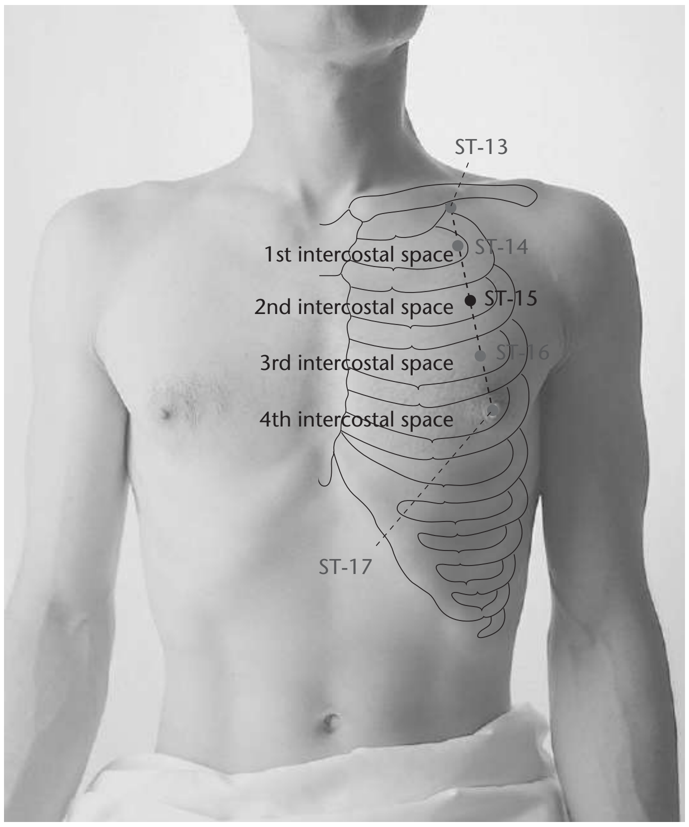
ST-13
1st intercostal space
ST-14
2nd intercostal space
ST-15
3rd intercostal space
ST-16
4th intercostal space
ST-17

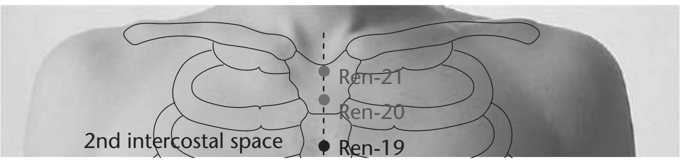
Ren-21
Ren-20
2nd intercostal space
Ren-19

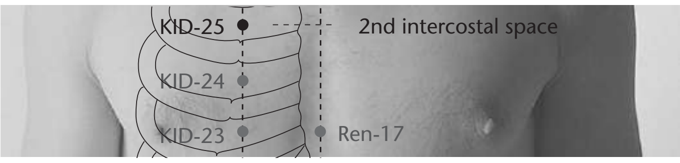
KID-25
2nd intercostal space
KID-24
KID-23
Ren-17

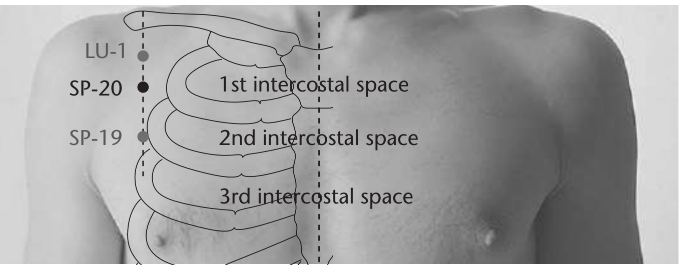
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

Location

In the 2nd intercostal space, on the midclavicular line, 4 cun lateral to the anterior midline.

How to find

On the upper thorax, the midclavicular line (4 cun lateral to the anterior midline) runs slightly obliquely from the clavicular midpoint to the generally more lateral nipple. Locate the 2nd intercostal space by palpating downward from the clavicle or by palpating from the manubriosternal synchondrosis (second rib) (→ 3.5). Then palpate along its course to the midclavicular line, where **ST-15** is located. Note the ascending course of the intercostal space! Located on the same level are → **Ren-19** (anterior midline), → **KID-25** (2 cun lateral to the anterior midline) and → **SP-20** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the course of the intercostal space or with or against the channel pathway. Caution: pneumothorax.

Actions/Indications

- Regulates the Lung Qi
- Unbinds the chest
- Benefits the breasts and clears Heat

ST-16 *Breast Window* YINGCHUANG

Location

In the 3rd intercostal space, on the midclavicular line, 4 cun lateral to the anterior midline.

How to find

On the upper thorax, the midclavicular line (4 cun lateral to the anterior midline) runs slightly obliquely from the clavicular midpoint to the generally more lateral nipple. Locate the 3rd intercostal space by palpating downward from the clavicle or by palpating from the manubriosternal synchondrosis (second rib) (→ 3.5). Then palpate along its course to the midclavicular line, where **ST-16** is located. Note the ascending course of the intercostal space! In male patients, this point can also be located by counting one intercostal space upward from the level of the nipple.

Located on the same level are → **Ren-18** (anterior midline), → **KID-24** (2 cun lateral to the anterior midline) and → **SP-19** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), along the course of the intercostal space or with or against the flow of the channel. Caution: pneumothorax.

Actions/Indications

- Regulates the Qi
- Unbinds the chest
- Benefits the breasts

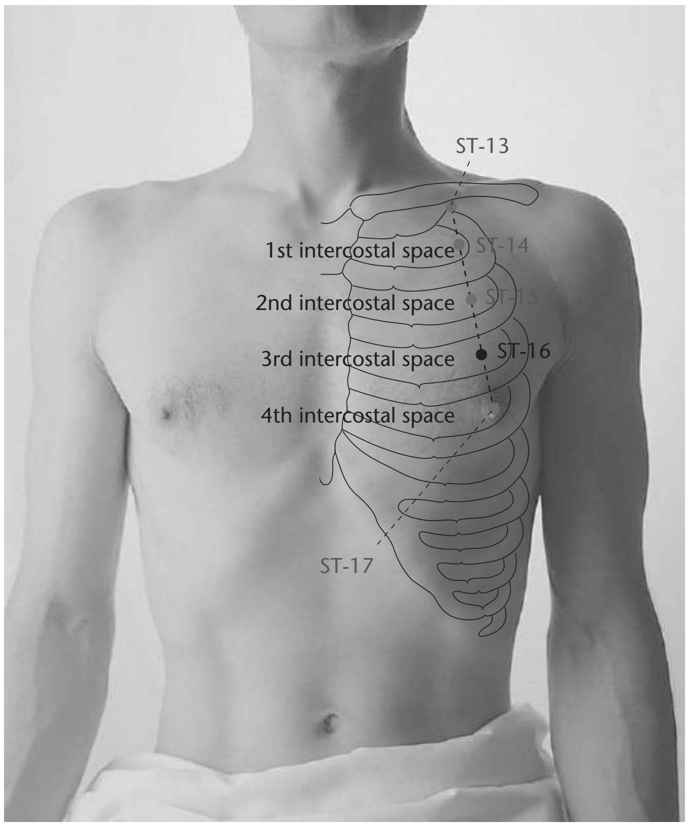
ST-13
1st intercostal space
ST-14
2nd intercostal space
ST-15
3rd intercostal space
ST-16
4th intercostal space
ST-17

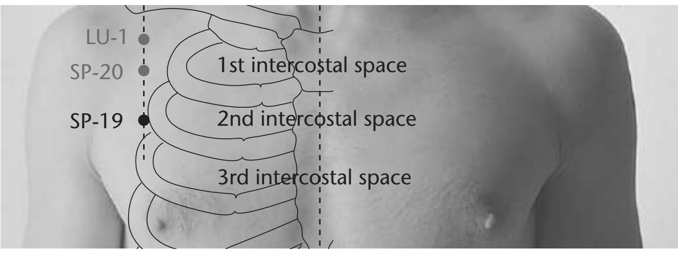
LU-1
SP-20
SP-19
1st intercostal space
2nd intercostal space
3rd intercostal space

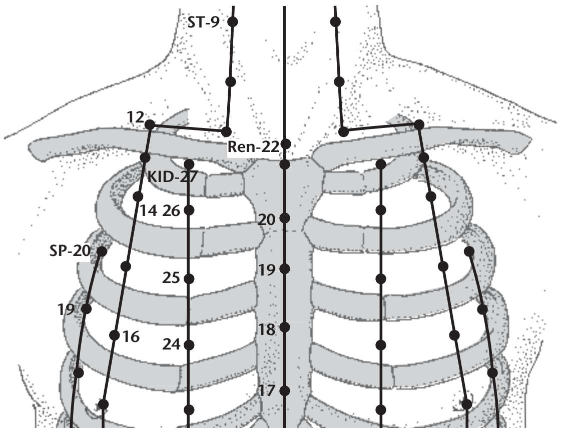
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17

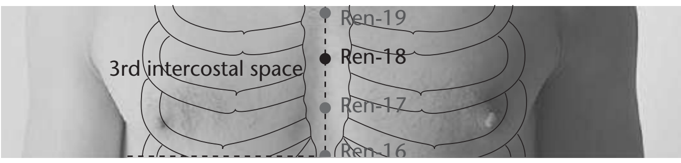
Ren-19
3rd intercostal space
Ren-18
Ren-17
Ren-16

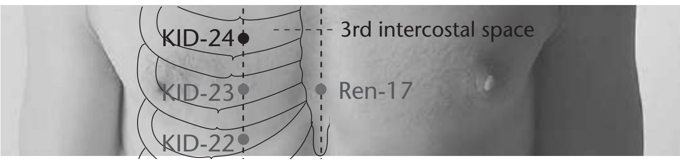
KID-24
3rd intercostal space
KID-23
Ren-17
KID-22

Breast Centre RUZHONG ST-17

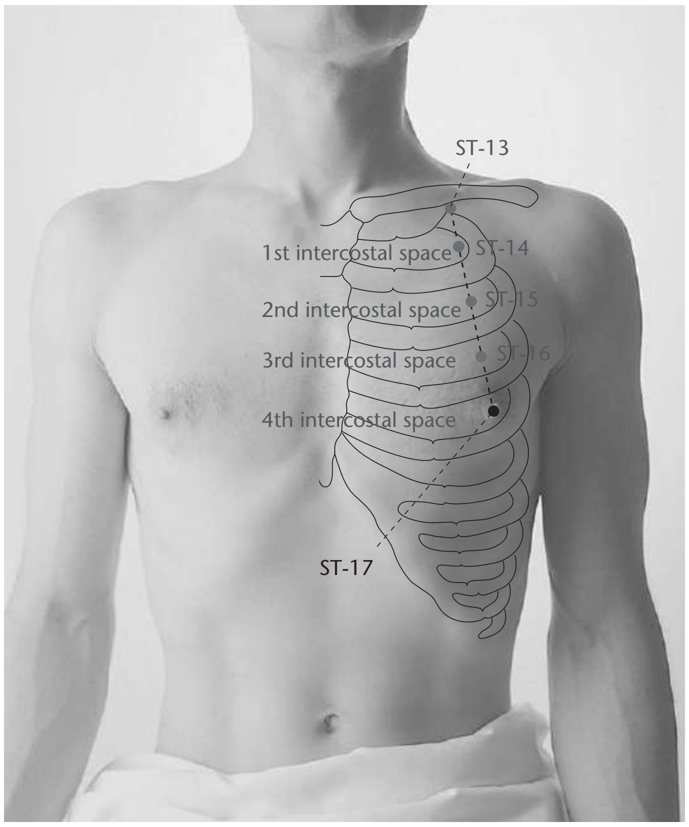
ST-13
1st intercostal space
ST-14
2nd intercostal space
ST-15
3rd intercostal space
ST-16
4th intercostal space
ST-17

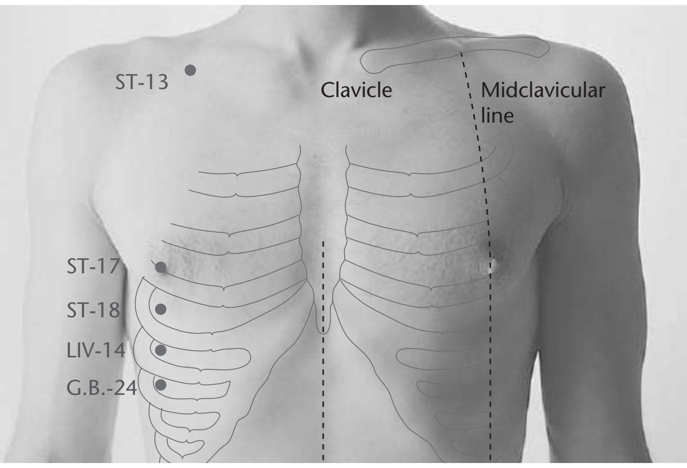
ST-13
Clavicle
Midclavicular line
ST-17
ST-18
LIV-14
G.B.-24

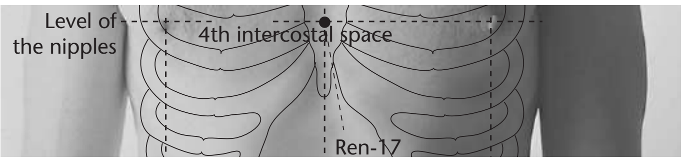
Level of the nipples
4th intercostal space
Ren-17

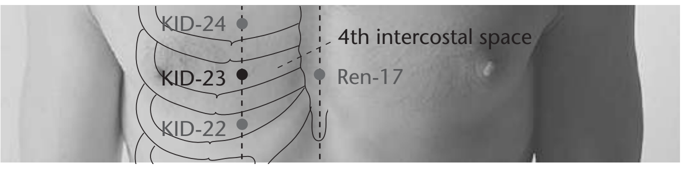
KID-24
4th intercostal space
KID-23
Ren-17
KID-22

Location

In the centre of the nipple.

How to find

In the centre of the nipple. In men, this is usually located 4 cun lateral to the anterior midline, at the level of the 4th intercostal space and is also used for reference.

In men, also located on the same level are → **Ren-17** (anterior midline), → **KID-23** (2 cun lateral to the anterior midline), → **SP-18** (6 cun lateral to the midline), → **P-1** (1 cun lateral to the nipple), → **G.B.-22** and → **G.B.-23**. In women, the location of the nipples varies considerably and should therefore not be used for reference.

Needling

No therapy; this point is for reference only.

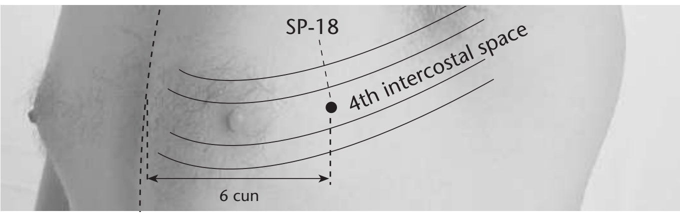
SP-18
4th intercostal space
6 cun

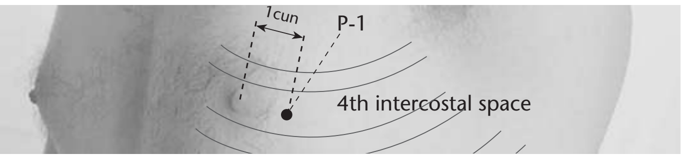
1cun
P-1
4th intercostal space

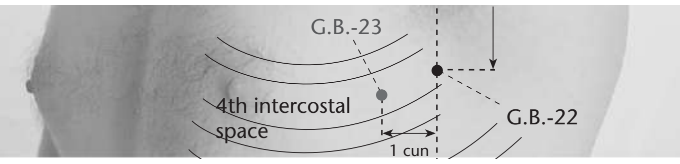
G.B.-23
4th intercostal space
G.B.-22
1 cun

ST-18 *Breast Root* RUGEN

Location

In the 5th intercostal space, on the mamillary line, 4 cun lateral to the anterior midline.

How to find

The mamillary line, a vertical line used for reference, is located 4 cun lateral to the anterior midline (→ 3.5). In men, the nipple is generally located on the level of the 4th intercostal space, and **ST-18** is located in the intercostal space directly below. The point is often located at the lower border of the pectoralis major muscle; in women, it is often found at the root of the breast. Location help in women: Find the 5th intercostal space by palpating downward from the clavicle or from the manubriosternal synchondrosis (second rib) (→ 3.5). Then follow the course of the intercostal space laterally to the mamillary line and there locate **ST-18**. Note the ascending course of the intercostal spaces!

Located on the same level are → **Ren-16** (anterior midline), → **KID-22** (2 cun lateral to the anterior midline) and → **SP-17** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the course of the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Benefits the breasts, moves the Blood
- Unbinds the chest
- Regulates the Lung Qi

Special features

Important local point

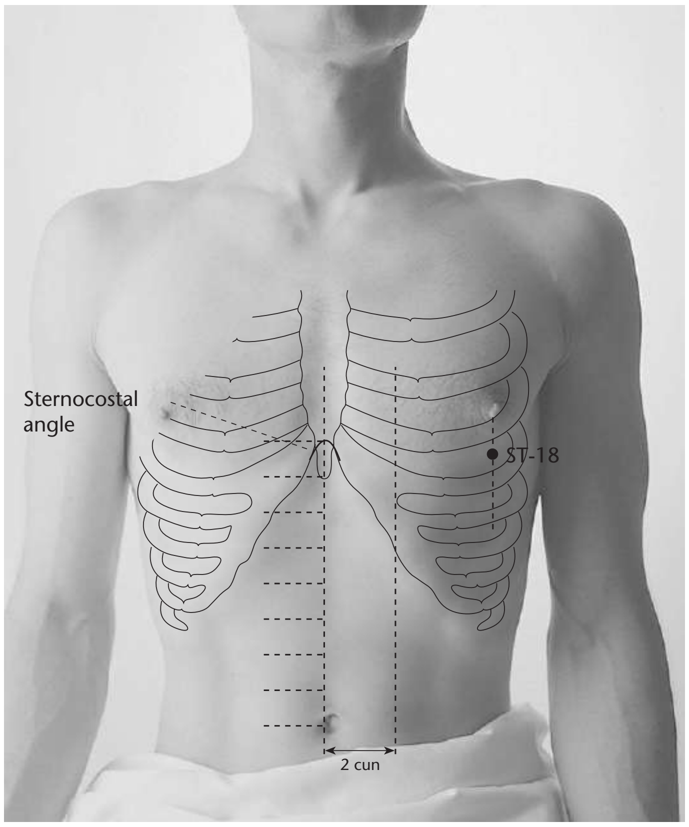
Sternocostal angle
ST-18
2 cun

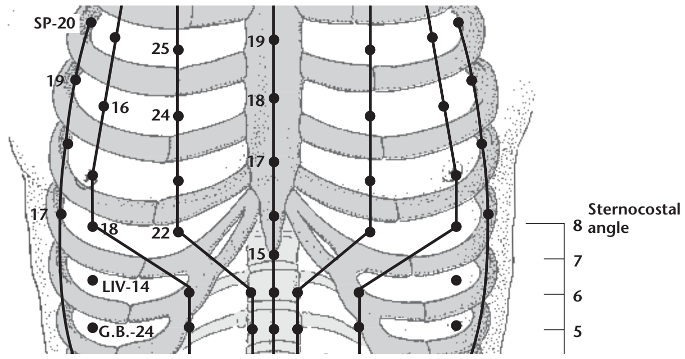
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
Sternocostal angle
8
7
6
5

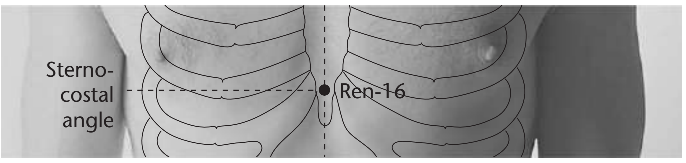
Sterno-costal angle
Ren-16

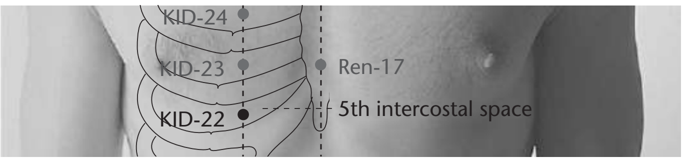
KID-24
KID-23
KID-22
Ren-17
5th intercostal space

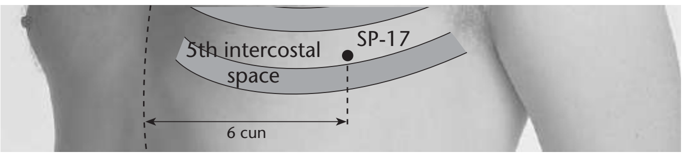
5th intercostal space
SP-17
6 cun

Not Contained BURONG ST-19

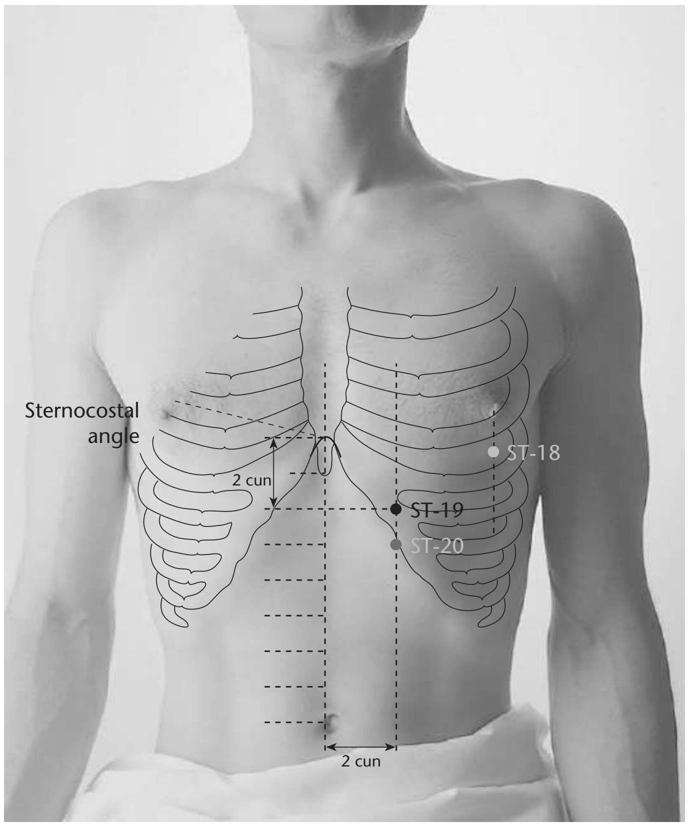

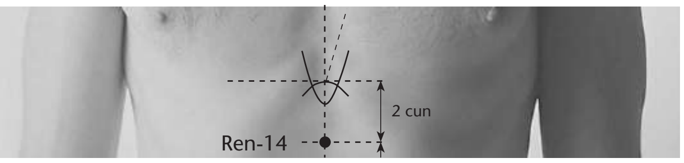

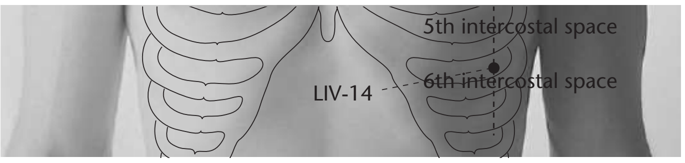

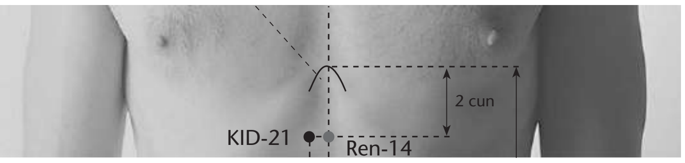

Location

2 cun below the sternocostal angle and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape ➞ 2.3.1). First, measure 2 cun down from the sternocostal angle. **Or:** Measure the midpoint of the upper half of the distance between the umbilicus and the subcostal angle (➞ **Ren-12**: spreading hands technique ➞ 2.3.3). From there, measure 2 cun in a lateral direction to locate **ST-19**. As its location depends on the shape of the thorax, it might be located either on the rib or on the abdomen. Located on the same level are ➞ **Ren-14** (anterior midline), ➞ **KID-21** (0.5 cun lateral to the anterior midline) and ➞ **LIV-14** (on the mamillary line and in the 6th intercostal space). Please note: From **ST-19** onwards, the ST channel is located only 2 cun lateral to the anterior midline.

Needling

Vertically 0.5–1 cun. If the correct location projects this point on the lower ribcage, needle transversely on the rib or locate the point more medially (or choose an alternative point) Caution: Peritoneum, pericardium.

Actions/Indications

- Regulates the Stomach, Intestines and Middle Burner
- Regulates the Lung Qi

ST-20 *Supporting Fullness* CHENGMAN

Location

3 cun below the sternocostal angle (or 5 cun above the umbilicus) and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). First, measure 3 cun down from the sternocostal angle. **Or:** Measure 1 cun up from the midpoint of the distance between the umbilicus and the sternocostal angle (→ **Ren-12**: spreading hands technique → 2.3.3). From there, measure 2 cun in a lateral direction to locate **ST-20**. Depending on the shape of the thorax, **ST-20** might be located on the lower border of the ribcage in some cases (narrow subcostal angle).

Located on the same level are → **Ren-13** (anterior midline) and → **KID-20** (0.5 cun lateral to the anterior midline) and → **G.B.-24** (on the mamillary line, in the 7th intercostal space).

Needling

Vertically 0.5–1 cun. If the correct location projects this point on the lower ribcage, needle transversely on the rib or locate the point more medially (or choose an alternative point). Caution: Peritoneum.

Actions/Indications

- Regulates the Middle Burner and the Stomach Qi
- Regulates the Lung Qi

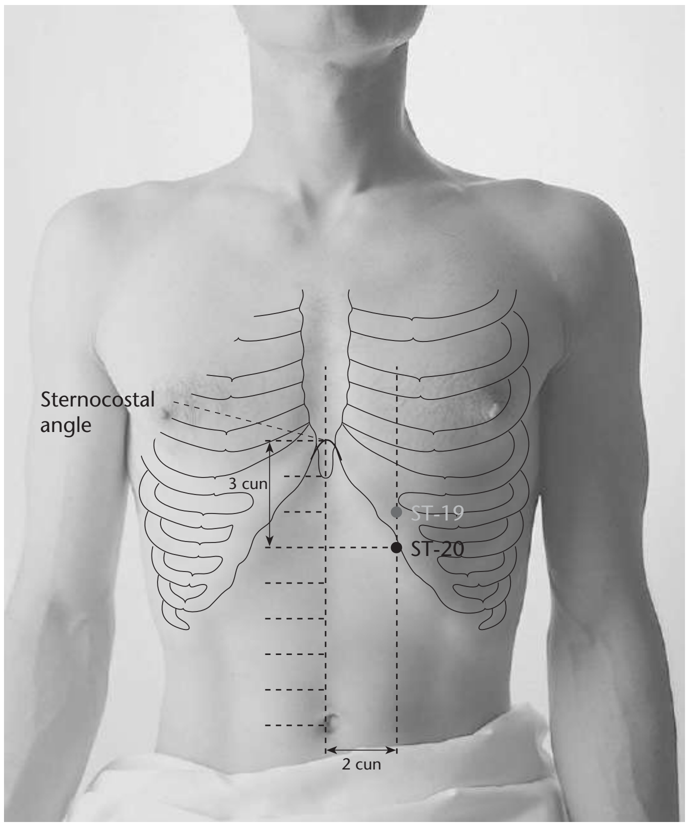
Sternocostal angle
3 cun
ST-19
ST-20
2 cun

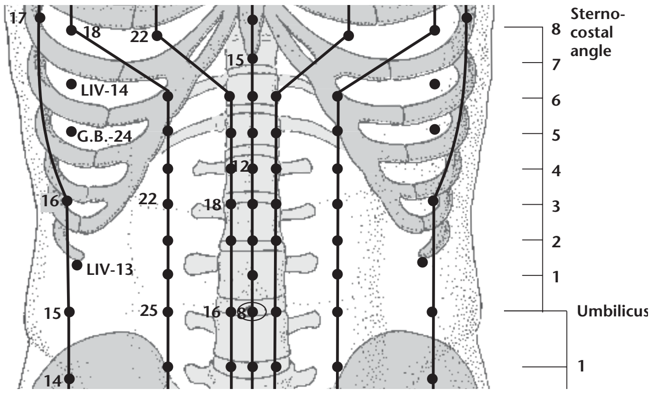
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
Sterno-costal angle
8
7
6
5
4
3
2
1
Umbilicus
1

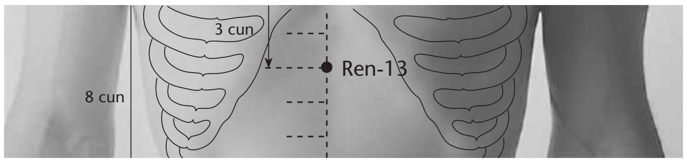
3 cun
Ren-13
8 cun

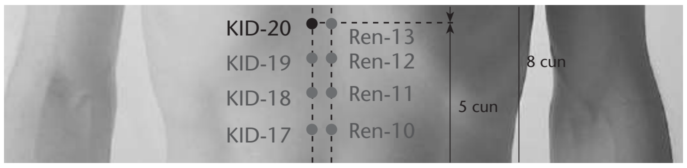
KID-20
KID-19
KID-18
KID-17
Ren-13
Ren-12
Ren-11
Ren-10
5 cun
8 cun

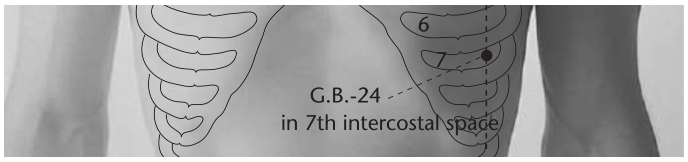
6
7
G.B.-24
in 7th intercostal space

Beam Gate LIANGMEN ST-21

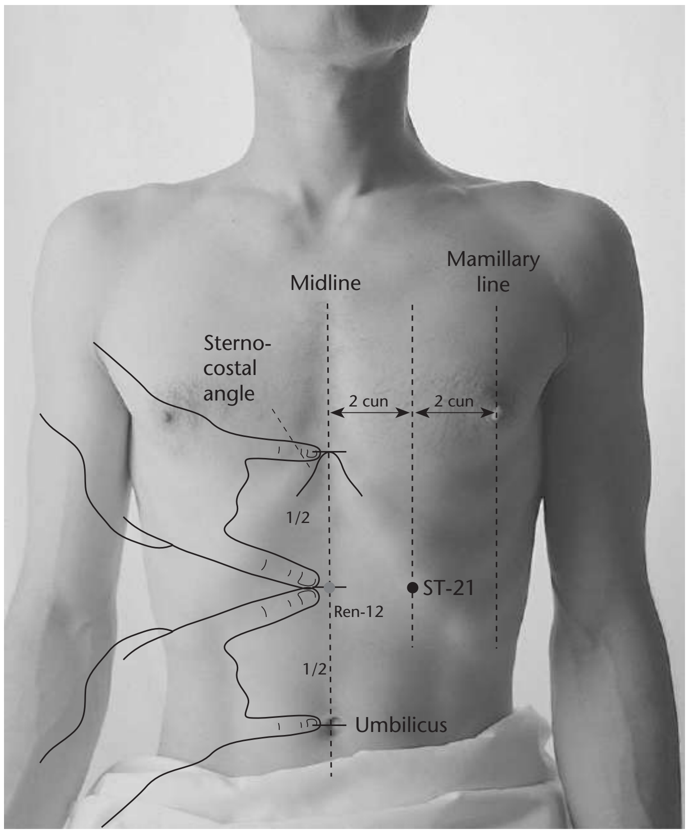
Midline
Mamillary line
Sterno-costal angle
2 cun
2 cun
1/2
ST-21
Ren-12
1/2
Umbilicus

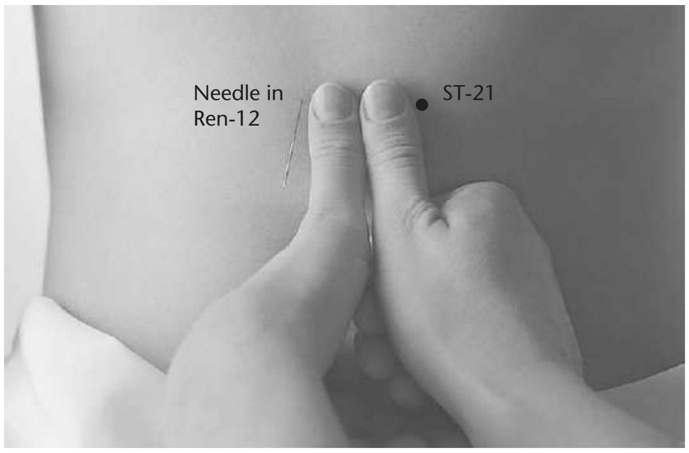
Needle in Ren-12
ST-21

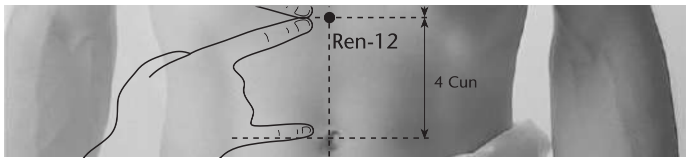

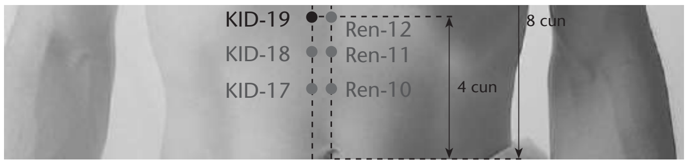
KID-19
Ren-12
8 cun
KID-18
Ren-11
KID-17
Ren-10
4 cun

Location

2 cun lateral to the anterior midline, 4 cun above the umbilicus (or 4 cun below the subcostal angle).

How to find

The distance between the sternocostal angle and the centre of the umbilicus can be divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). **Quick method:** Find the midpoint between the umbilicus and the sternocostal angle (→ **Ren-12**) by using the spreading hands technique (→ 2.3.3), then measure 2 cun in a lateral direction to locate **ST-21**. Depending on the shape of the thorax, this can be located on the lower rib cage.
Located on the same level are → **Ren-12** (anterior midline) and → **KID-19** (0.5 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. If the correct location projects this point on the lower ribcage, needle transversely on the rib or locate the point more medially (or choose an alternative point). Caution: Peritoneum, enlarged liver (on the right), pregnancy.

Actions/Indications

- Regulates the Qi, harmonises the Middle Burner, eliminates stagnation, raises the Qi and stops diarrhoea

Special features

Important local point for distension and pain in the epigastrium and the lateral abdomen

ST-22 *Pass Gate* GUANMEN

Location

3 cun above the umbilicus (or 5 cun inferior to the sternocostal angle) and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus can be divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). First, measure 3 cun up from the umbilicus. **Or:** Measure 1 cun downward from the midpoint of the distance between the umbilicus and the sternocostal angle (→ **Ren-12**: spreading hands technique → 2.3.3). **ST-22** is located 2 cun lateral from here. Located on the same level are → **Ren-11** (anterior midline), → **KID-18** (0.5 cun lateral to the anterior midline) and → **SP-16** (4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Stomach and Intestines
- Regulates the water passages

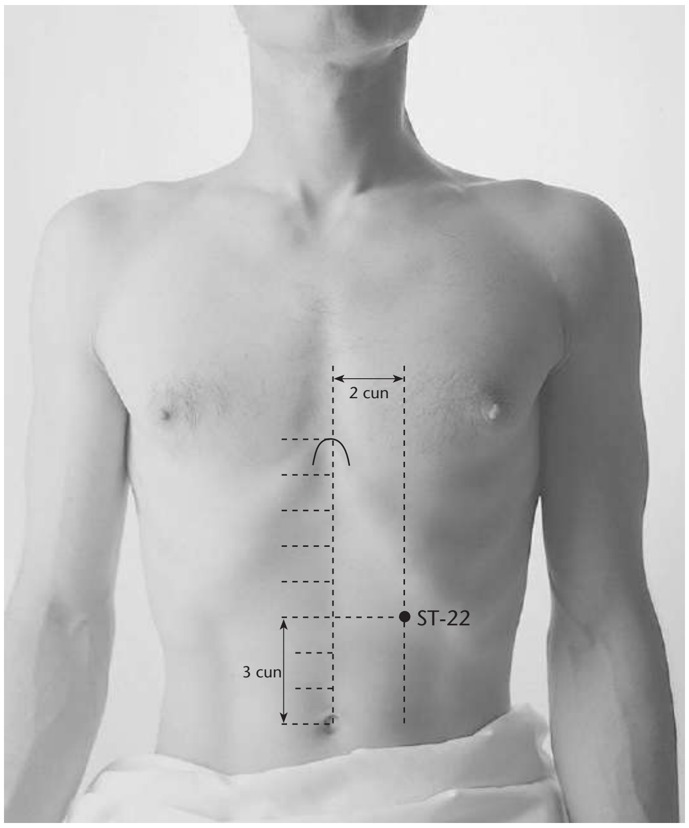
2 cun
ST-22
3 cun

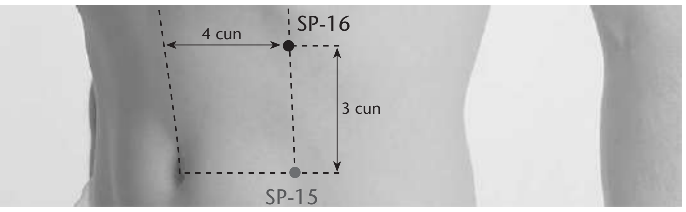
4 cun
SP-16
3 cun
SP-15

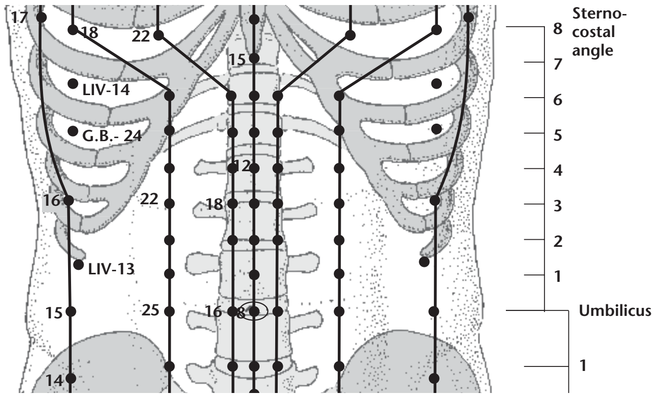
17
18
22
15
LIV-14
G.B.- 24
12
16
22
18
LIV-13
15
25
16
8
14
Sterno-
costal
angle
8
7
6
5
4
3
2
1
Umbilicus
1

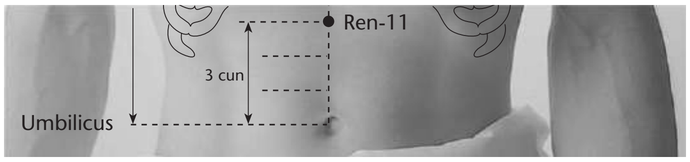
Ren-11
3 cun
Umbilicus

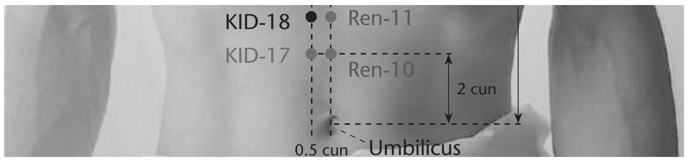
KID-18
Ren-11
KID-17
Ren-10
2 cun
0.5 cun
Umbilicus

Supreme Unity TAIYI ST-23

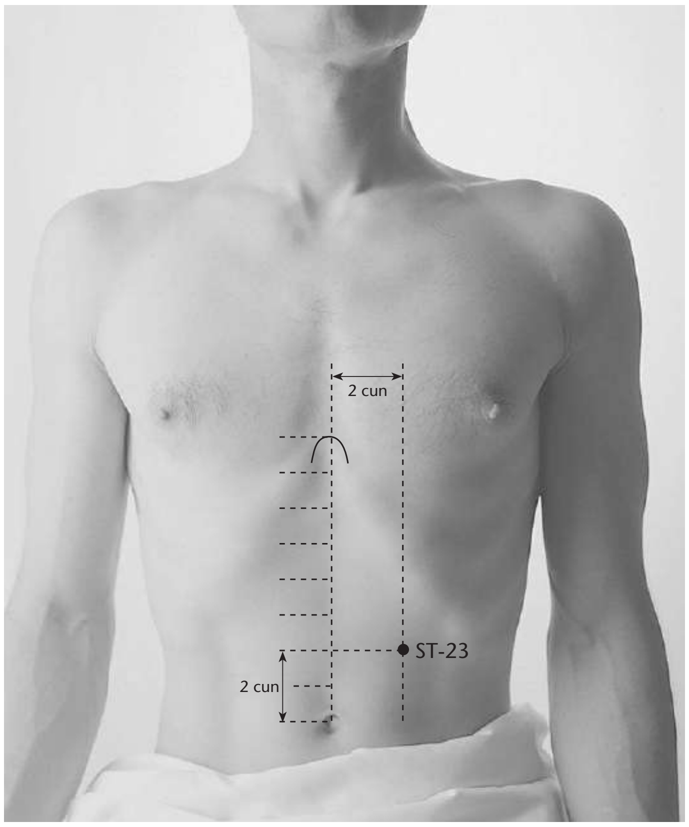

Location

2 cun above the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape ➞ 2.3.1). First, measure 2 cun up from the umbilicus. **Or:** Measure the midpoint of the lower half of the distance between the umbilicus and the subcostal angle (➞ **Ren-12**: spreading hands technique ➞ 2.3.3). **ST-23** is located 2 cun lateral from here.

Located on the same level are ➞ **Ren-10** (anterior midline), ➞ **KID-17** (0.5 cun lateral to the anterior midline) and ➞ **Ex-CA** (*weishang*; 4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Harmonises the Middle Burner, strengthens the Spleen
- Benefits the Heart and calms the *shen*

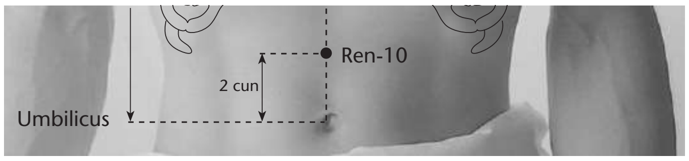

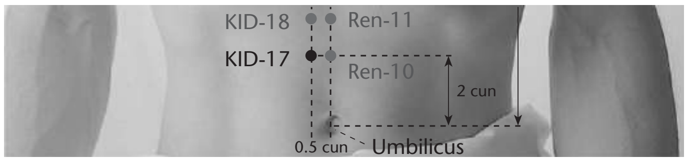

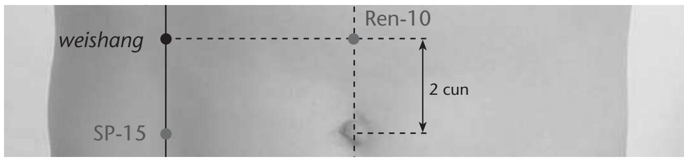

ST-24 *Slippery Flesh Gate* HUAROUMEN

Location

1 cun above the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). **ST-24** is located 1 cun above the umbilicus and 2 cun lateral to the anterior midline.

Located on the same level is → **Ren-19** (anterior midline); often, also → **LIV-13** (at the free end of the 11th rib) will be located at this level.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Stomach
- Transforms Phlegm and calms the *shen*

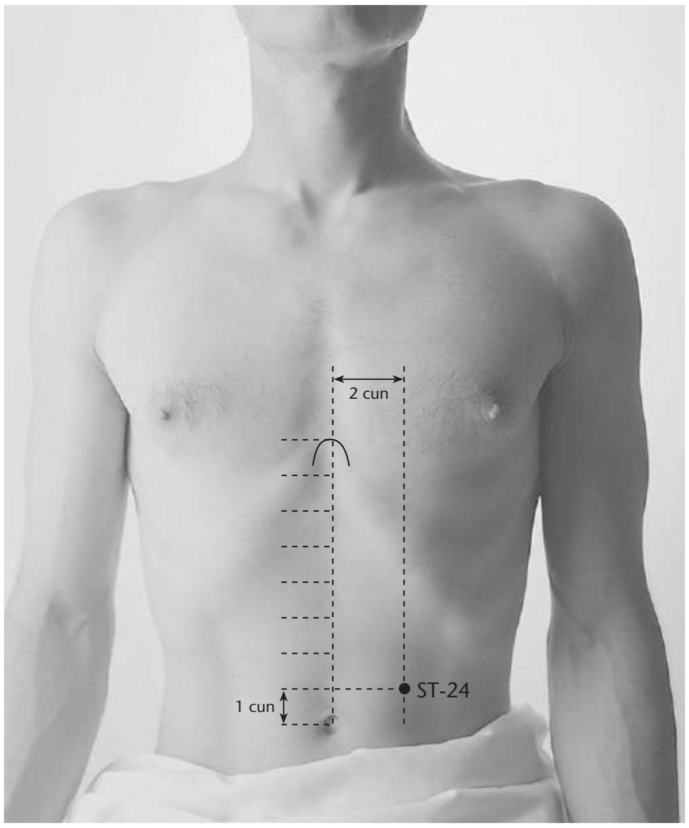
2 cun
ST-24
1 cun

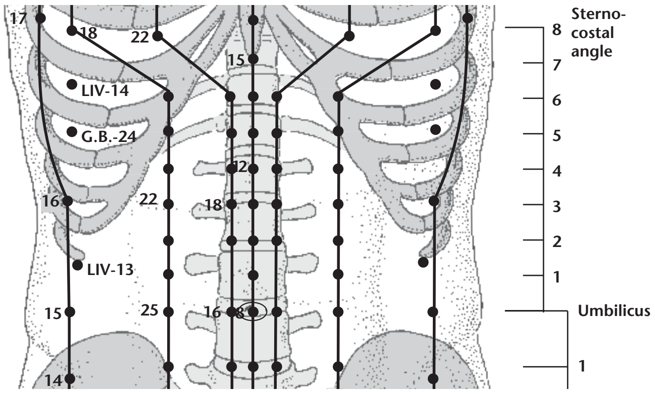
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
8
Sterno-costal angle
7
6
5
4
3
2
1
Umbilicus
1

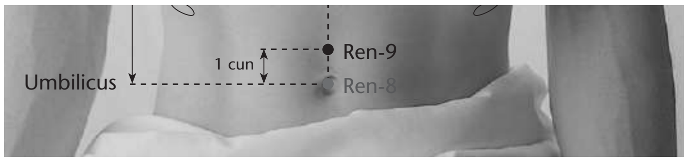
Ren-9
Umbilicus
1 cun
Ren-8

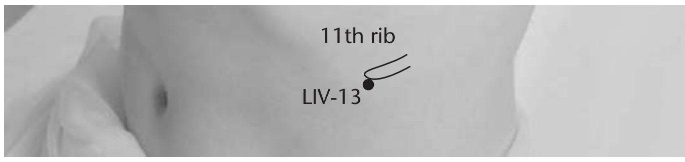
11th rib
LIV-13

Heaven's Pivot TIANSHU ST-25

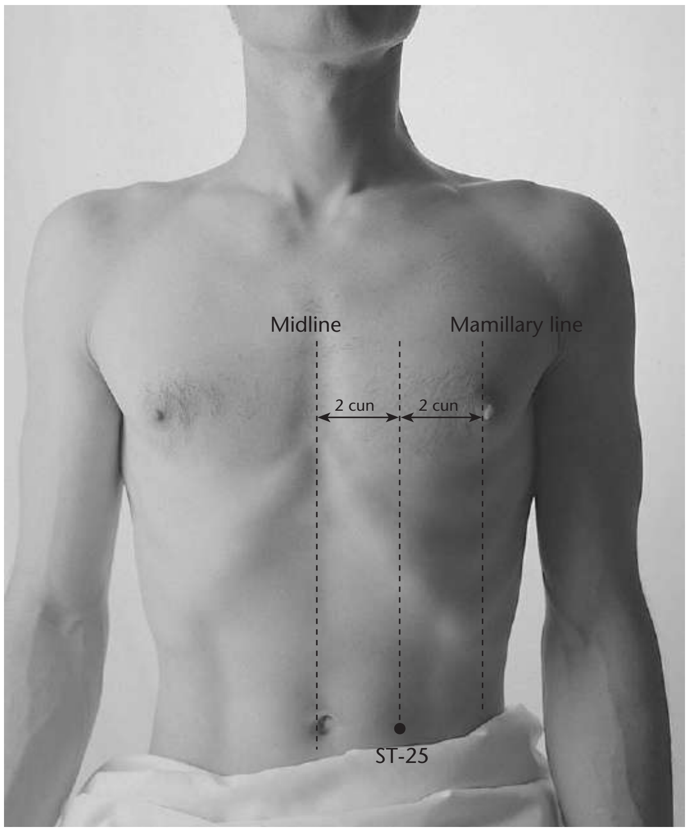

Location

2 cun lateral to the umbilicus.

How to find

Measure 2 cun lateral from the centre of the umbilicus. **For reference:** On the abdomen, 2 cun lateral corresponds to the midpoint between the anterior midline and the mamillary line (→ 3.5). Located on the same level are → **Ren-8/KID-16/SP-15/G.B.-26** (umbilicus/0.5 cun lateral to the umbilicus/mamillary line or 4 cun lateral to the umbilicus/at the level of the umbilicus and directly inferior to the free end of the 11th rib).

Needling

Vertically 0.5–1.5 cun. For gynaecological disorders, obliquely in an inferior direction towards the uterus. Caution: Peritoneum, pregnancy. The needle may reach the aponeuroses of the oblique abdominal musculature at the transition to the sheath of the rectus abdominis muscle.

Actions/Indications

- Regulates the Spleen, Stomach and Intestines, eliminates Dampness and clears Damp-Heat
- Eliminates stagnation

Special features

Front-*mu* point of the Large Intestine. Important point for disorders of the gastrointestinal tract.

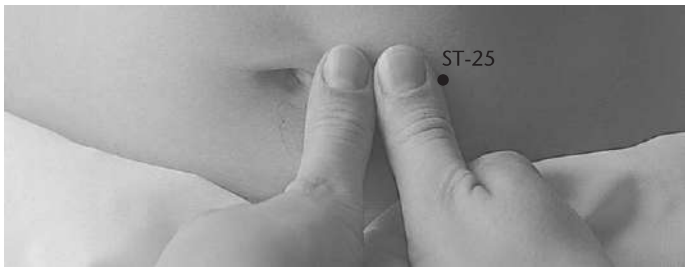

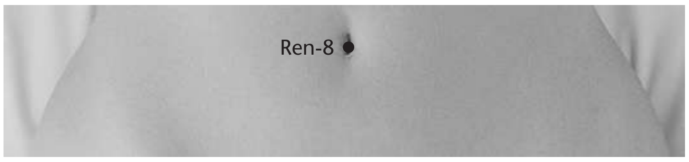

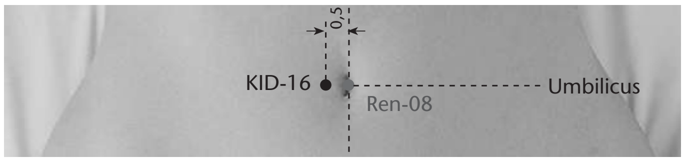

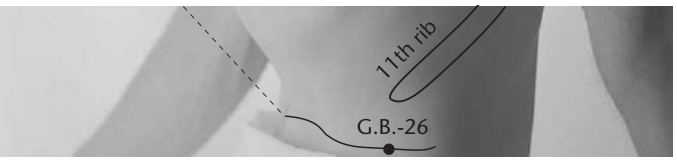

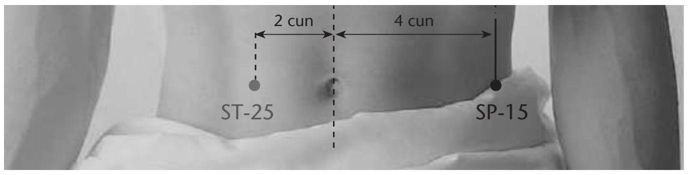

ST-26 *Outer Mound* WAILING

Location

1 cun below the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). From the umbilicus, measure 1 cun down: **ST-26** is located 2 cun laterally from there.
Located on the same level are → **Ren-7** (anterior midline) and → **KID-15** (0.5 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Qi, alleviates pain and Cold

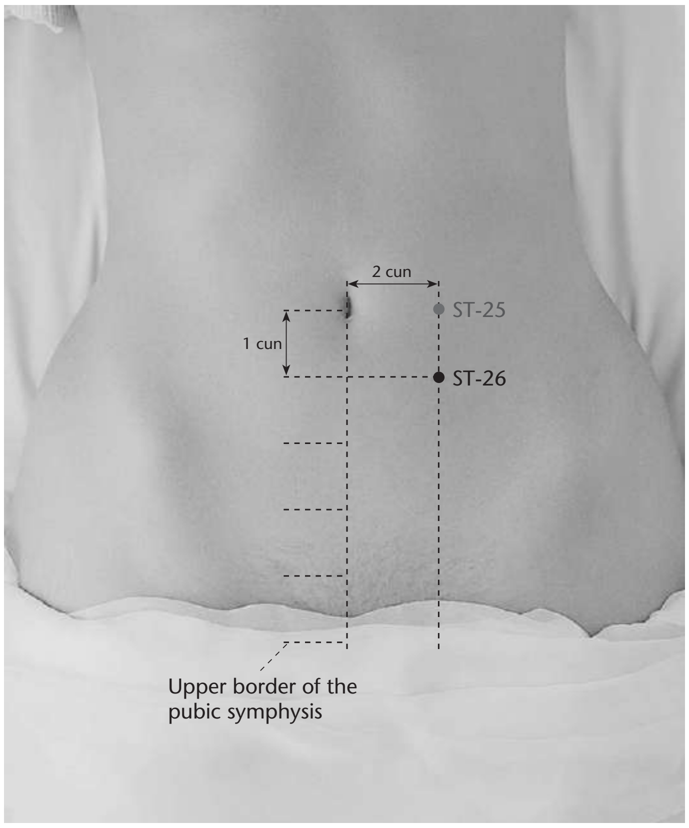
2 cun
ST-25
1 cun
ST-26
Upper border of the pubic symphysis

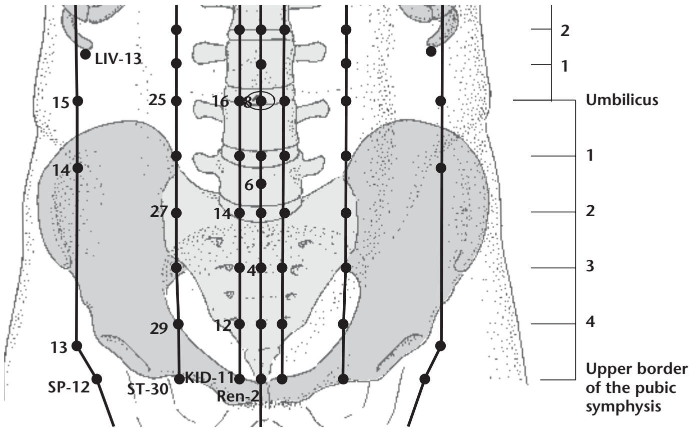
LIV-13
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
2
1
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

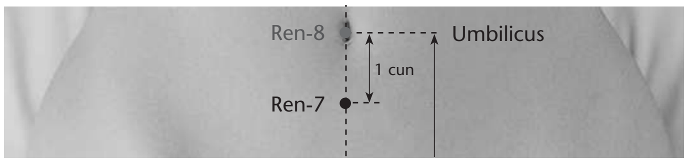
Ren-8
Umbilicus
1 cun
Ren-7

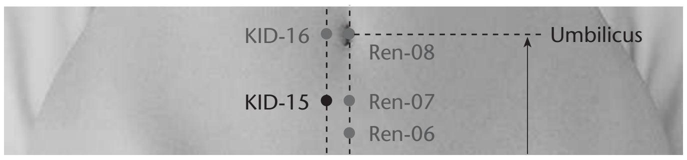
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06

The Great DAJU ST-27

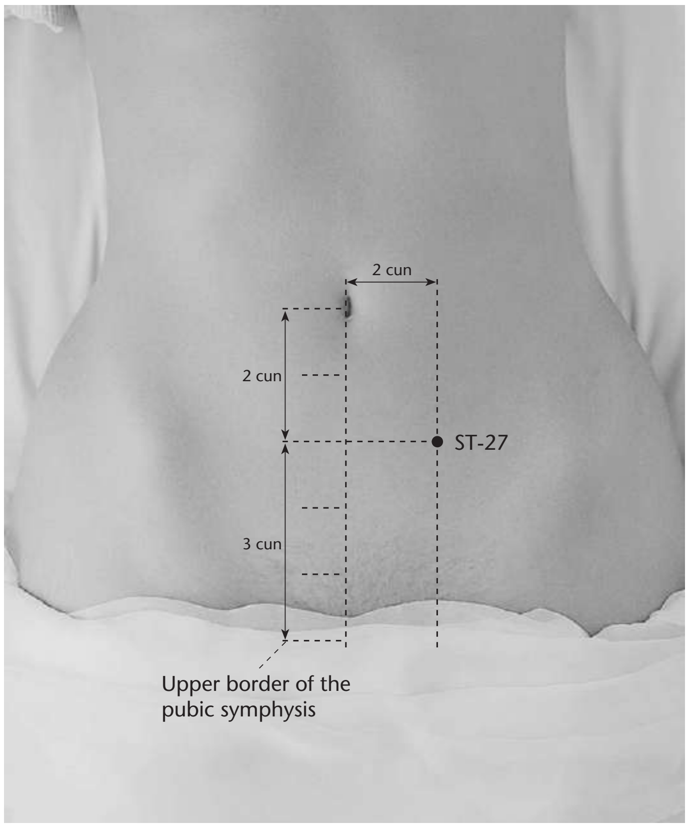

Location

2 cun below the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). From the umbilicus, measure 2 cun down and 2 cun laterally, where **ST-27** is located.

Located on the same level are → **Ren-5** (anterior midline) and → **KID-14** (0.5 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates and strengthens the Qi
- Benefits the Kidneys and promotes diuresis, strengthens the Yang and Essence (*jing*)

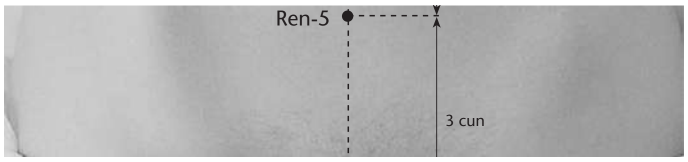

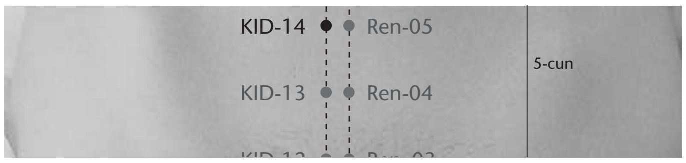

ST-28 *Water Passage* SHUIDAO

Location

3 cun below the umbilicus or 2 cun above the upper border of the pubic symphysis and 2 cun lateral to the anterior midline.

How to find

The distance between the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun up or measure 3 cun down from the umbilicus. **ST-28** is located 2 cun laterally from there.

Located on the same level are → **Ren-4** (anterior midline), → **KID-13** (0.5 cun lateral to the anterior midline) and → **Ex-CA** (*tituo/quimen/yijing*; 4/3/1 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, urinary bladder, pregnancy.

Actions/Indications

- Clears Damp-Heat, benefits the Lower Burner and eliminates stagnation

Special features

Important local point for the urogenital tract

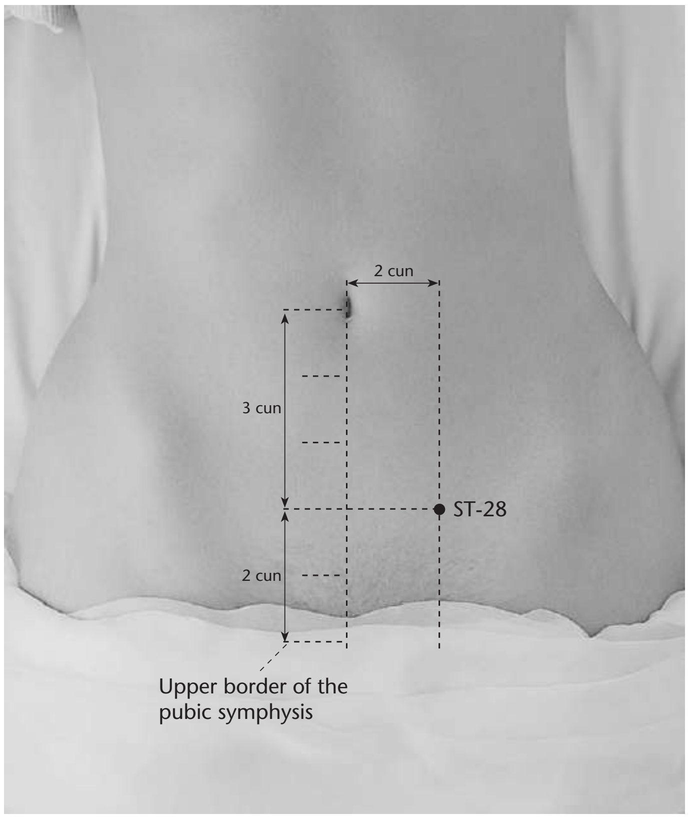
2 cun
3 cun
ST-28
2 cun
Upper border of the pubic symphysis

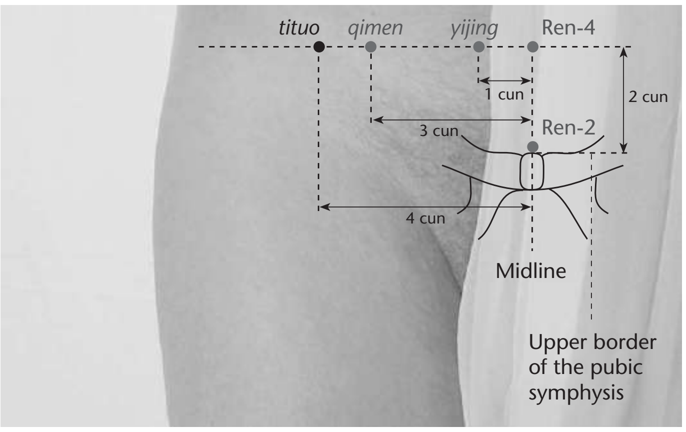
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2
4 cun
Midline
Upper border of the pubic symphysis

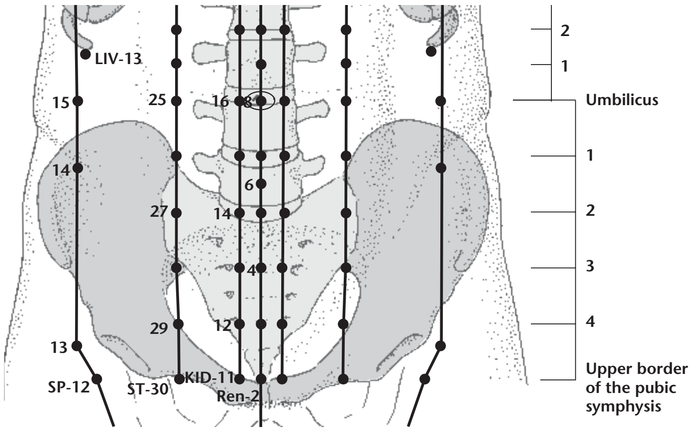
LIV-13
15
25
16
8
14
6
27
14
4
13
29
12
SP-12
ST-30
KID-11
Ren-2
2
1
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

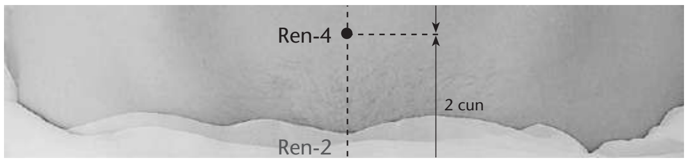
Ren-4
2 cun
Ren-2

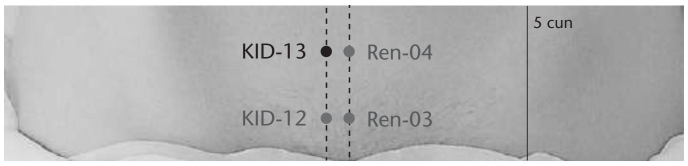
5 cun
KID-13
Ren-04
KID-12
Ren-03

Return GUILAI ST-29

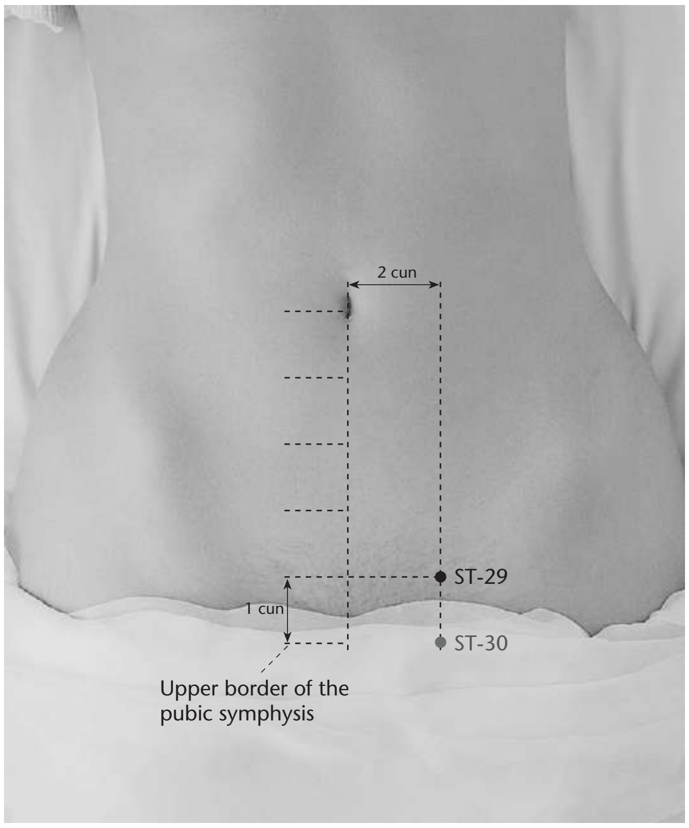
2 cun
ST-29
1 cun
ST-30
Upper border of the
pubic symphysis

Location

1 cun above the upper border of the pubic symphysis or 4 cun below the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 1 cun up and 2 cun lateral to **ST-29**.

Located on the same level are → **Ren-3** (anterior midline), → **KID-12** (0.5 cun lateral to the anterior midline) and → **Ex-CA** (*zigong*; 3 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, urinary bladder, pregnancy.

Actions/Indications

- Warms and regulates the Lower Burner and menstruation

Special features

Commonly treated with moxibustion, the combination of acupuncture and moxibustion being particularly beneficial. Important local point for the urogenital tract.

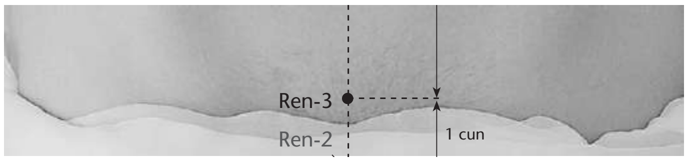
Ren-3
Ren-2
1 cun

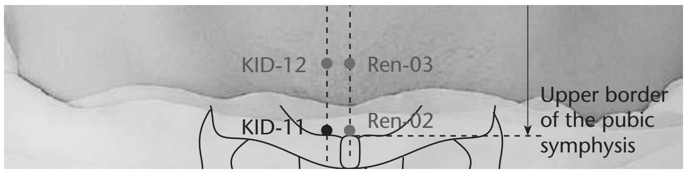
KID-12
Ren-03
KID-11
Ren-02
Upper border
of the pubic
symphysis

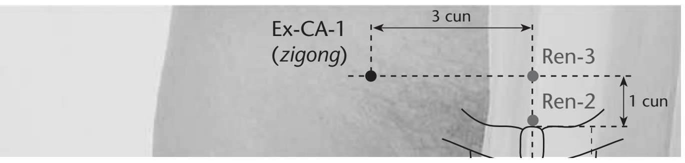
Ex-CA-1
(*zigong*)
3 cun
Ren-3
Ren-2
1 cun

ST-30 *Rushing Qi* QICHONG

Location

2 cun lateral to the upper border of the pubic symphysis and medial to the femoral artery and vein, approximately 1 cun superior to the inguinal groove.

How to find

First, locate the upper border of the pubic symphysis (location of → **Ren-2**). Then locate **ST-30** 2 cun lateral from there (on the vertical line halfway between the anterior midline and the mamillary line). This often corresponds with a location 1 cun superior to the inguinal groove and medial to the femoral artery/vein.
Located on the same level are → **Ren-2/KID-11/SP-12** (anterior midline/0.5 cun lateral to the anterior midline/3.5 cun lateral to the anterior midline). → **LIV-12** is located in the inguinal groove, 1 cun inferior and 0.5 cun lateral to **ST-30**.

Needling

0.5–1 cun vertically or 1.5 cun slightly obliquely towards the external genitalia for disorders of the urogenital tract. Caution: Femoral artery, urinary bladder (quite superficial in a superior direction); in men, the spermatic ducts (quite superficial in an inferior direction), pregnancy. Ask the patient to empty their bladder before needling this point.

Actions/Indications

- Regulates the Qi in the Lower Burner
- Strengthens the Sea of Grain (later Heaven Qi)

Special features

Meeting point with the *chong mai* (according to some authors, also with the G.B. channel), Sea of Grain point. Wide usage for disorders of the Lower Burner.

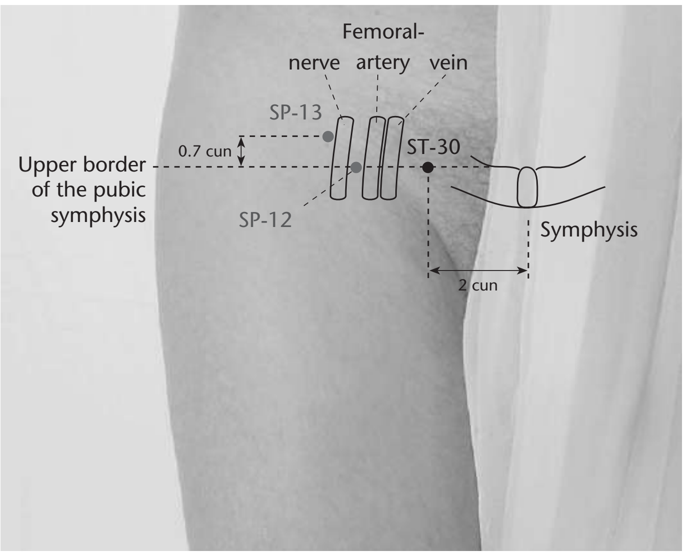
Femoral-
nerve
artery
vein
SP-13
0.7 cun
ST-30
Upper border of the pubic symphysis
SP-12
Symphysis
2 cun

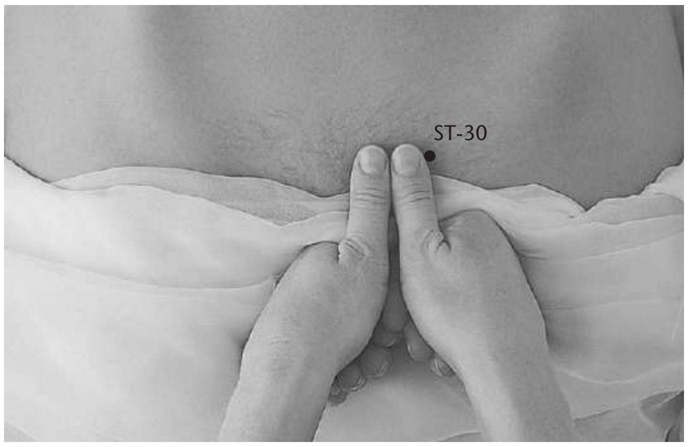

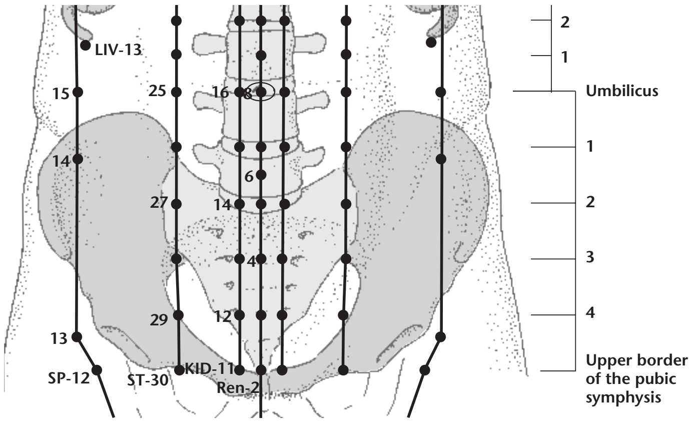
LIV-13
15
25
16
8
14
6
27
14
4
13
29
12
SP-12
ST-30
KID-11
Ren-2
2
1
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

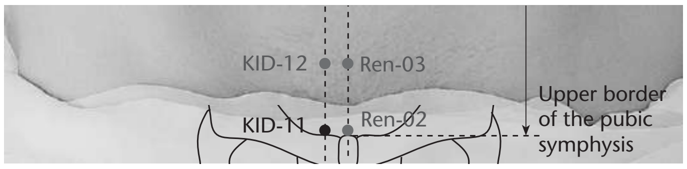
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

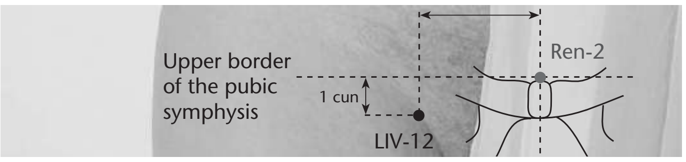
Ren-2
Upper border of the pubic symphysis
1 cun
LIV-12

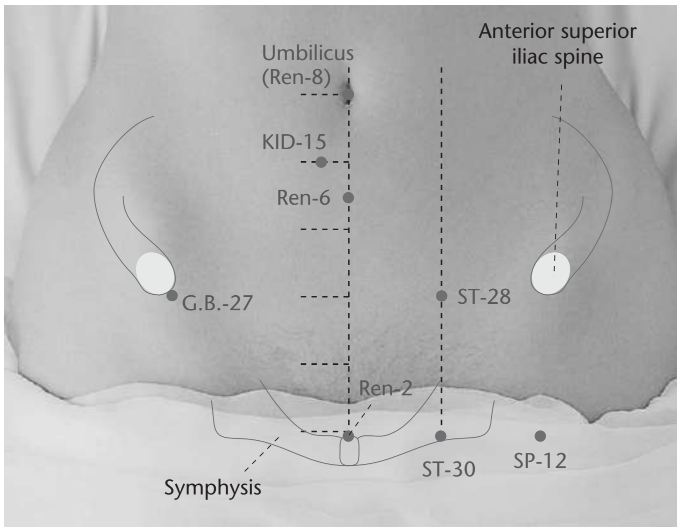
Anterior superior iliac spine
Umbilicus (Ren-8)
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

Thigh Gate BIGUAN ST-31

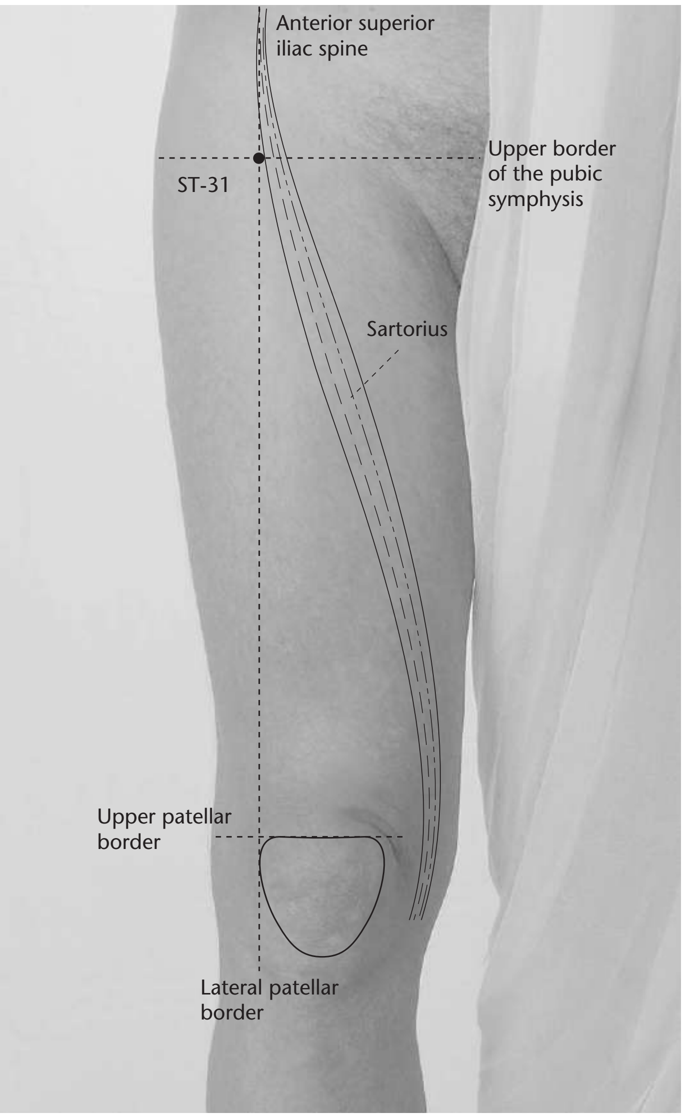
Anterior superior iliac spine
ST-31
Upper border of the pubic symphysis
Sartorius
Upper patellar border
Lateral patellar border

Location

Inferior to the anterior superior iliac spine and lateral to the sartorius muscle, at the level of the lower border of the pubic symphysis.

How to find

The sartorius muscle originates at the anterior superior iliac spine (ASIS) and crosses the thigh towards the medial aspect of the knee. With the knee and hip slightly flexed and with active lateral rotation, it can be isolated easily. **ST-31** is located on the lateral border of the muscle, at the crossing point of a vertical line connecting the ASIS and the lateral upper border of the patella and a horizontal line at the level of the lower border of the pubic symphysis.

Needling

Vertically 0.5–2 cun

Actions/Indications

- Opens the channel, warms and expels pathogenic factors

Special features

Particularly useful for Cold *bi*-syndrome, disorders of the knees and lower back. The chain-and-lock method in combination (→ 8) with → **ST-36** and → **ST-41** is often used for paraesthesia and pain in the hip and leg radiating along the channel.

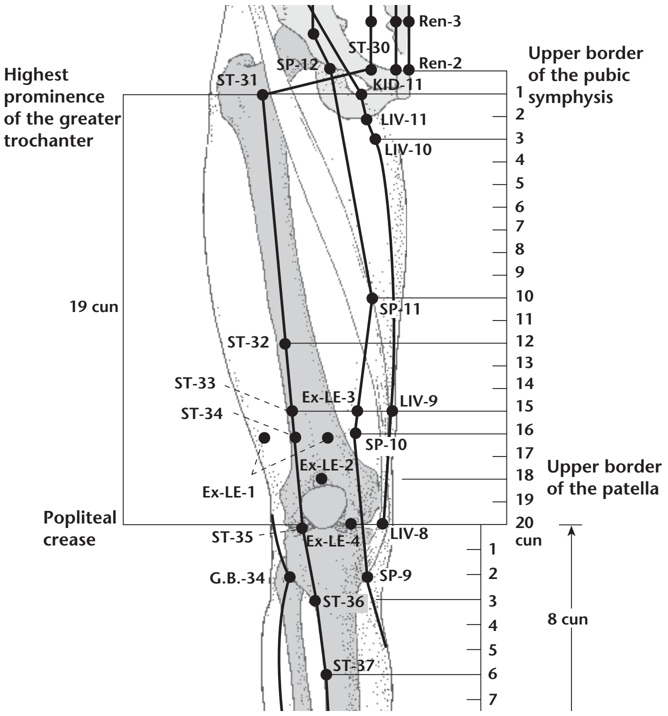
Ren-3
ST-30
SP-12
Ren-2
Upper border of the pubic symphysis
Highest prominence of the greater trochanter
ST-31
KID-11
LIV-11
LIV-10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
cun
19 cun
SP-11
ST-32
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Upper border of the patella
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
1
2
3
4
5
6
7
8 cun

ST-32 *Crouching Rabbit* FUTU

Location

On the thigh, on a line connecting the lateral patellar border and the anterior superior iliac spine, 6 cun superior to the upper, lateral border of the patella.

How to find

Divide the distance between the tip of the greater trochanter and the lateral upper border of the patella into three sections. **ST-32** is located at the junction of the lower and the middle third of the thigh in a depression on the quadriceps muscle.

Needling

Vertically 0.5–2 cun

Actions/Indications

- Opens the channel and expels Wind, Dampness and Cold

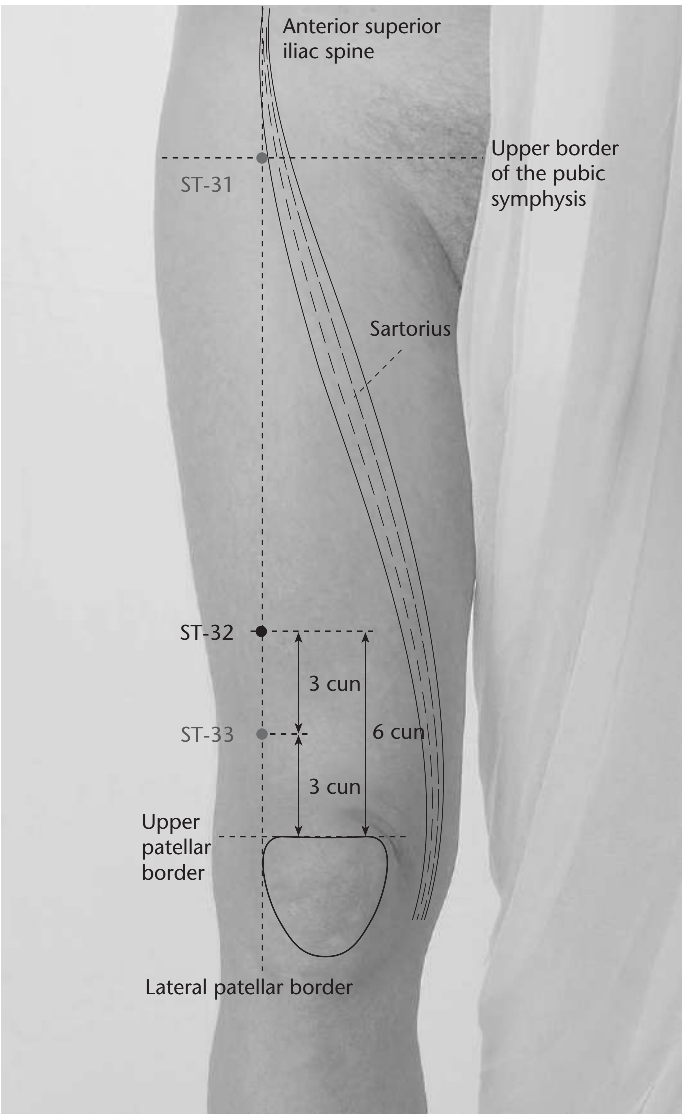
Anterior superior iliac spine
Upper border of the pubic symphysis
ST-31
Sartorius
ST-32
3 cun
ST-33
6 cun
3 cun
Upper patellar border
Lateral patellar border

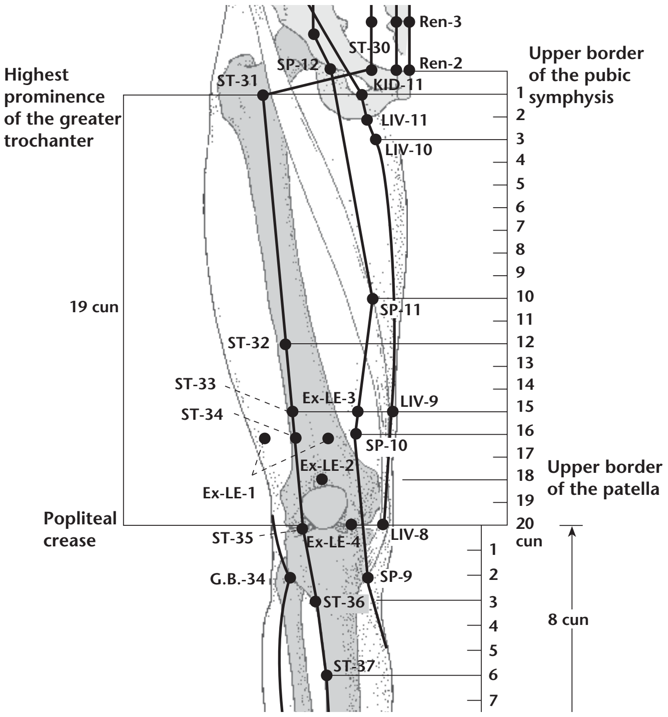
Ren-3
ST-30
Ren-2
SP-12
Upper border of the pubic symphysis
Highest prominence of the greater trochanter
ST-31
KID-11
LIV-11
LIV-10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
cun
19 cun
SP-11
ST-32
ST-33
Ex-LE-3
LIV-9
ST-34
SP-10
Ex-LE-2
Upper border of the patella
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
1
2
3
4
5
6
7
G.B.-34
SP-9
ST-36
8 cun
ST-37

Ch04.3-F10028.qxd 2/22/08 8:43 PM Page 161

Yin Market YINSHI ST-33

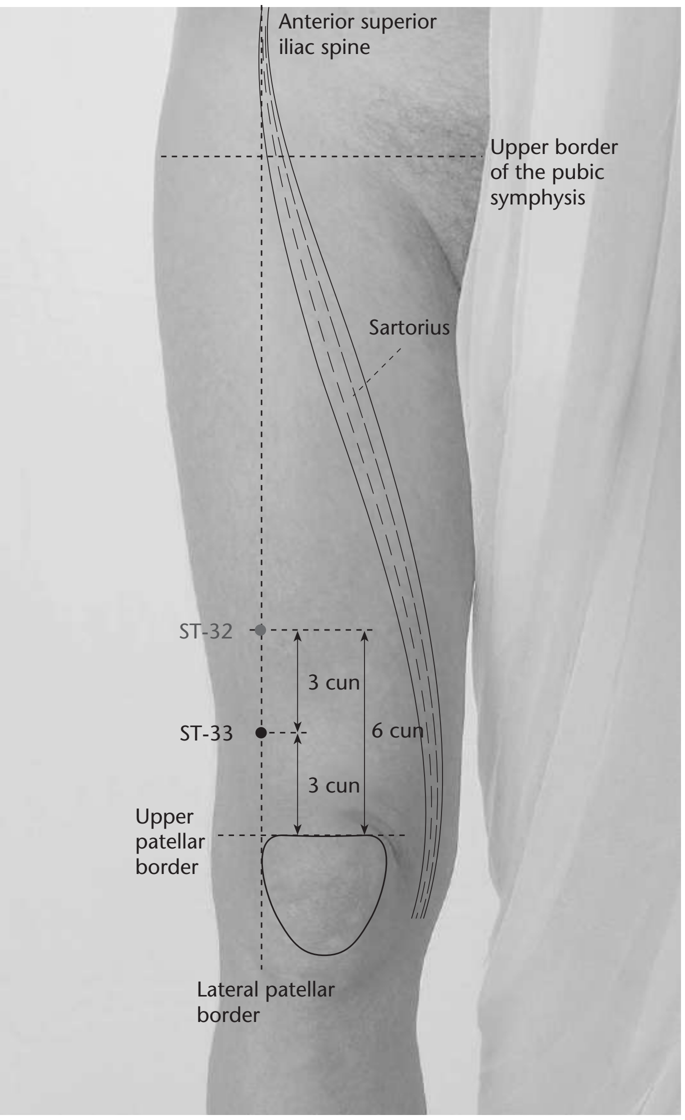
Anterior superior iliac spine
Upper border of the pubic symphysis
Sartorius
ST-32
3 cun
ST-33
6 cun
3 cun
Upper patellar border
Lateral patellar border

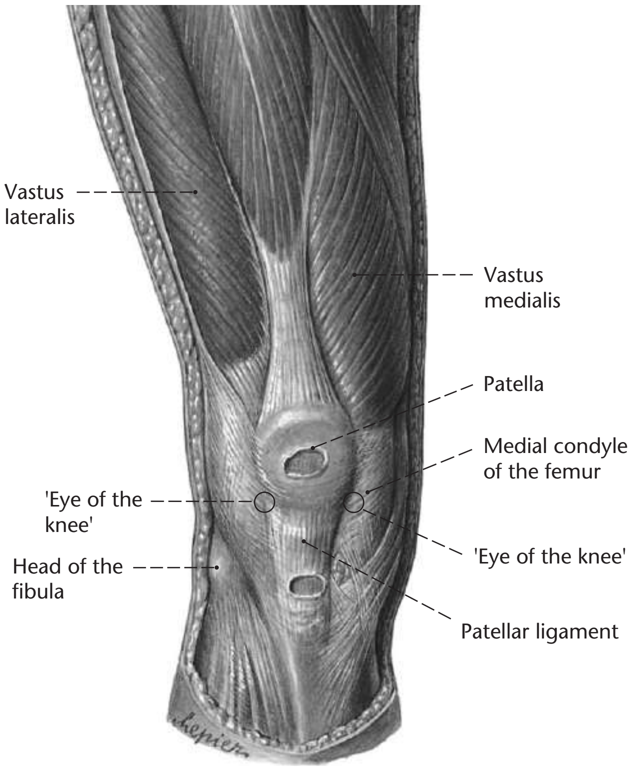
Vastus lateralis
Vastus medialis
Patella
Medial condyle of the femur
'Eye of the knee'
Head of the fibula
'Eye of the knee'
Patellar ligament

Location

On a line joining the lateral patellar border and the anterior superior iliac spine, 3 cun superior to the upper lateral border of the patella.

How to find

Measure 3 cun up from the upper lateral border of the patella, on the connecting line to the anterior superior iliac spine. **ST-33** is located in a groove at the junction of the rectus femoris and vastus lateralis muscles.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Opens the channel and expels pathogenic factors

ST-34 *Ridge Mound* LIANGQIU

Location

On a line joining the lateral patellar border and the anterior superior iliac spine, 2 cun proximal to the upper lateral border of the patella, in a groove of the vastus lateralis muscle.

How to find

Locate the lateral upper border of the patella and from there measure 2 cun proximally. **ST-34** is located in a palpable groove on the vastus lateralis muscle of the quadriceps femoris group. This point projects on the line joining the lateral upper border of the patella and the anterior superior iliac spine (ASIS → 3.5). Located on the same level are the points → **Ex-LE-1** (each 1.5 cun lateral and medial to **ST-34**). → **SP-10** is located in a comparable position on the medial aspect of the thigh (2 cun proximal to the medial upper border of the patella and in a groove on the vastus medialis muscle).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Regulates the Stomach Qi, alleviates acute conditions
- Opens the channel and alleviates pain

Special features

Xi-cleft point. Important local point for disorders of the knee. Modern application for acute epigastric pain – strongly reducing needling techniques have an immediate and calming effect on gastrointestinal peristalsis.

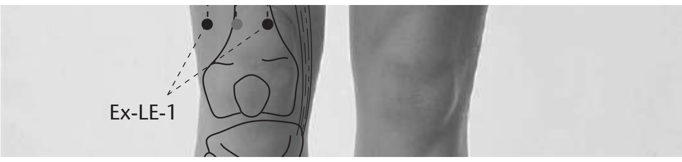

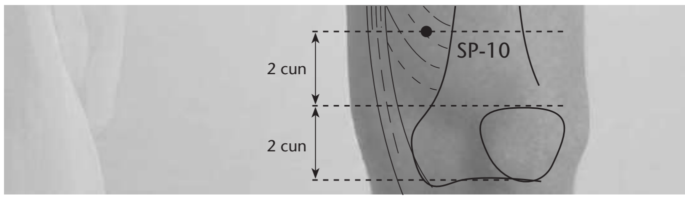
2 cun
2 cun
SP-10

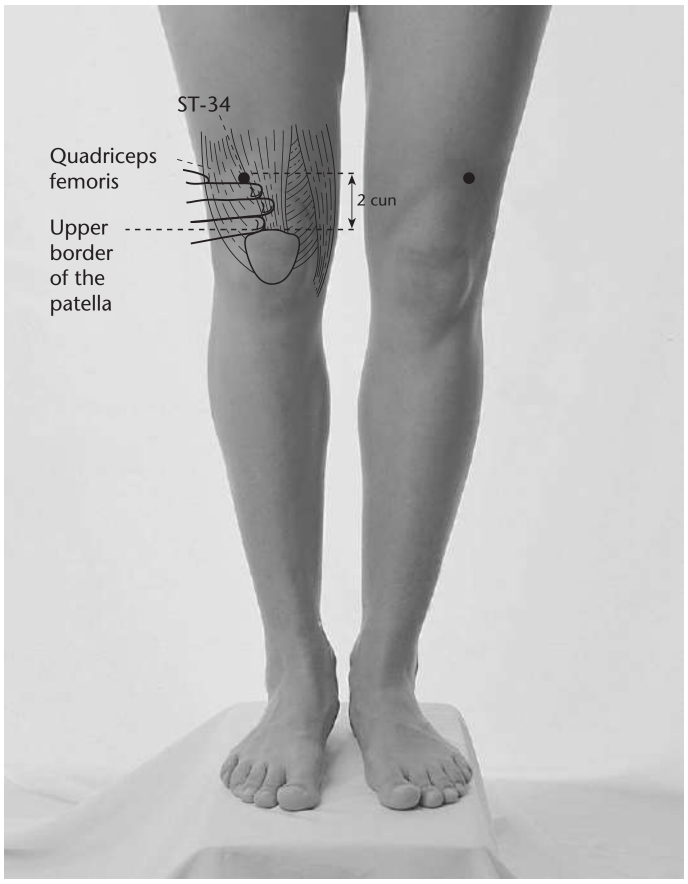
ST-34
Quadriceps femoris
2 cun
Upper border of the patella

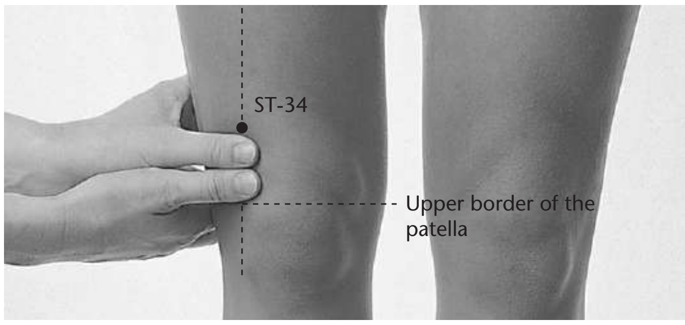
ST-34
Upper border of the patella

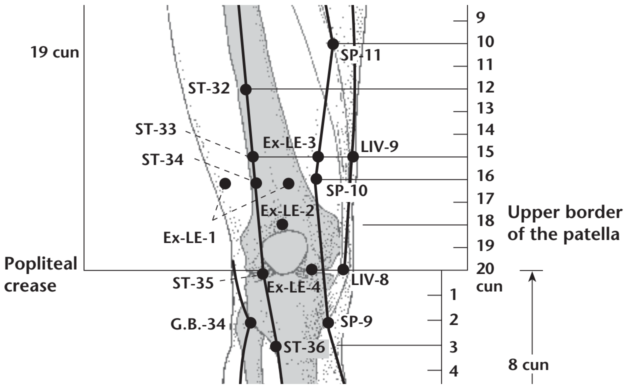
19 cun
SP-11
ST-32
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
9
10
11
12
13
14
15
16
17
18
19
20
cun
Upper border of the patella
1
2
3
4
8 cun

Calf's Nose DUBI ST-35

Location

With the knee flexed, in the depression inferior to the patella and lateral to the patellar ligament.

How to find

This point is best located and needled with the knee slightly flexed (use knee support). **ST-35** is located in a depression on the lower lateral border of the patella, in a groove lateral to the patellar ligament. It projects onto the level of the knee joint space and more or less corresponds with the arthroscopic access to the joint. **ST-35** is also known as 'lateral eye of the knee' and is part of the extra point ➞ **Ex-LE-5** (*xiyan*).

Needling

0.5–1 cun slightly obliquely in a medial direction, towards the medial eye of the knee ➞ **Ex-LE-4** (*neixiyan*). Caution: Knee joint (the location of this point corresponds approximately with the arthroscopic access to this joint!).

Actions/Indications

- Expels Wind and Dampness, reduces swellings, opens the channel, alleviates pain

Special features

Important local point for all disorders of the knee

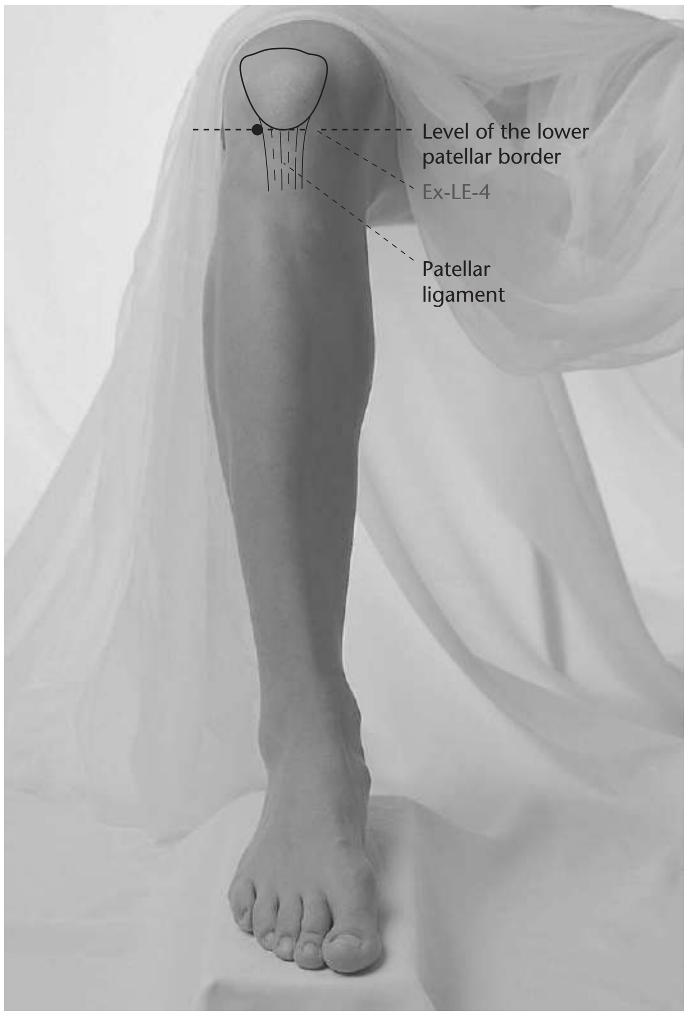
Level of the lower patellar border
Ex-LE-4
Patellar ligament

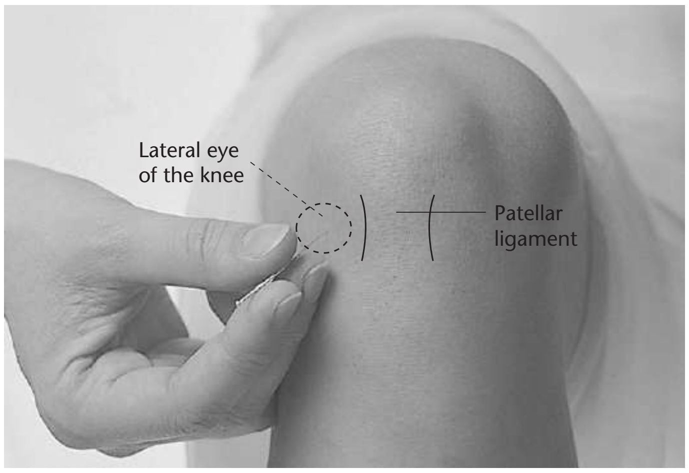
Lateral eye of the knee
Patellar ligament

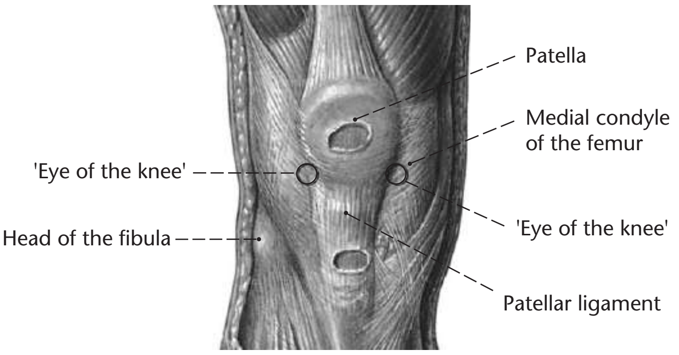
Patella
Medial condyle of the femur
'Eye of the knee'
'Eye of the knee'
Head of the fibula
Patellar ligament

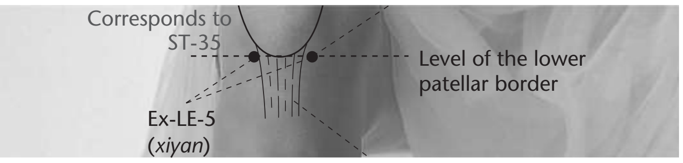
Corresponds to ST-35
Ex-LE-5 (xiyan)
Level of the lower patellar border

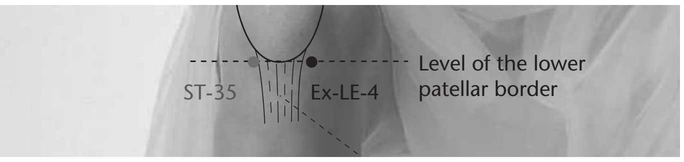
ST-35
Ex-LE-4
Level of the lower patellar border

ST-36 *Leg Three Miles* ZUSANLI

Location

3 cun distal to → **ST-35** ('lateral eye of the knee') and 1 fingerbreadth lateral to the anterior crest of the tibia, on the tibialis anterior muscle.

How to find

First, locate → **ST-35** ('lateral eye of the knee', at the level of the knee joint space). From → **ST-35**, palpate 3 cun (1 handbreadth) down and 1 fingerbreadth towards lateral by placing the middle finger adjacent to the tibial crest. There, **ST-36** is located in a depression, which can be palpated 'dynamically' (choose the point according to pressure sensitivity). **Or:** Palpate the lower border of the tibial tuberosity and 'dynamically' locate **ST-36** 1 fingerbreadth lateral to it.

Needling

Vertically 1–1.5 cun. Moxibustion is often used on this point.

Actions/Indications

- Regulates the Stomach, strengthens the Spleen, transforms Dampness
- Tonifies Qi and Yang, nourishes the Blood and Yin
- Calms the *shen*
- Opens the channel

Special features

He-sea point, Earth point, *ben* point (Five Phases point), lower *he*-sea point of the Stomach, *Gao Wu* command point for disorders of the abdomen, Ma Dan Yang point, Sea of Water and Grain point. Main point to tonify Qi and Blood for general weakness. Very important point.

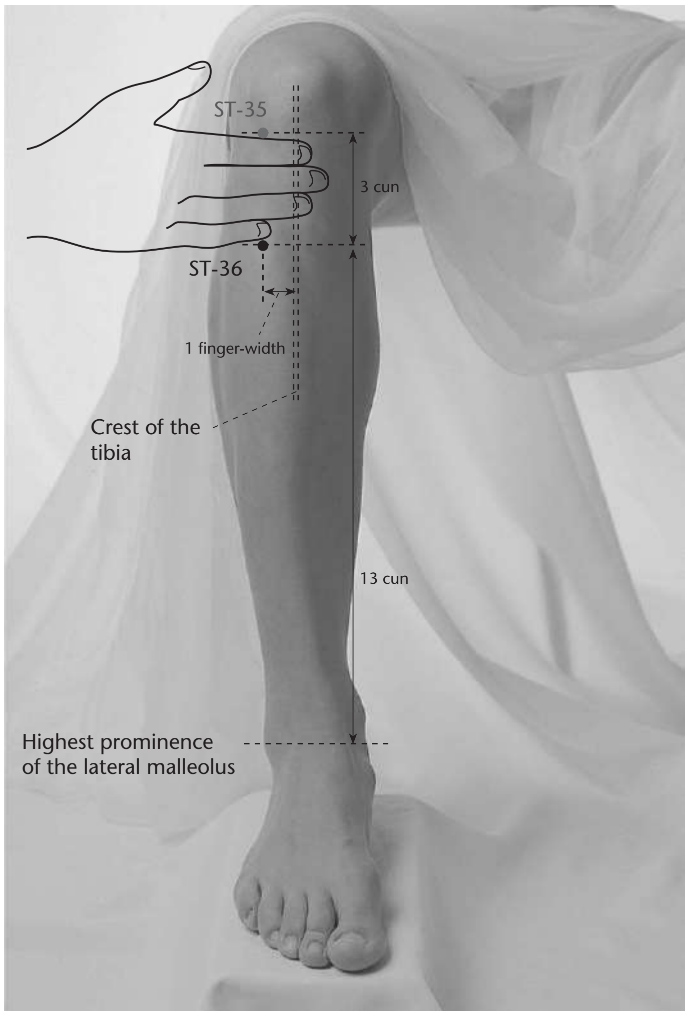
ST-35
3 cun
ST-36
1 finger-width
Crest of the tibia
13 cun
Highest prominence of the lateral malleolus

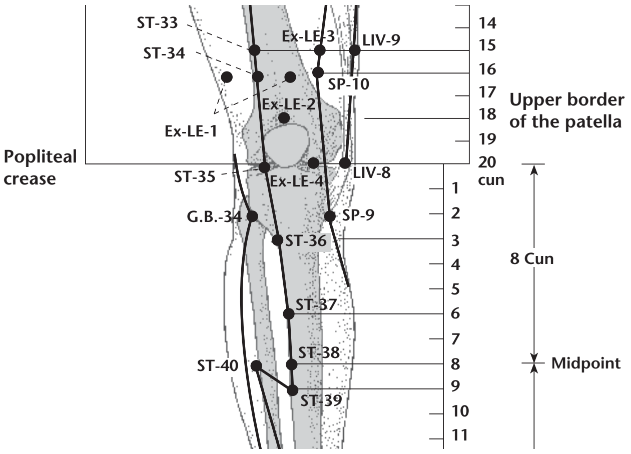
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
14
15
16
17
18
19
20
cun
Upper border of the patella
1
2
3
4
5
6
7
8
9
10
11
8 Cun
Midpoint

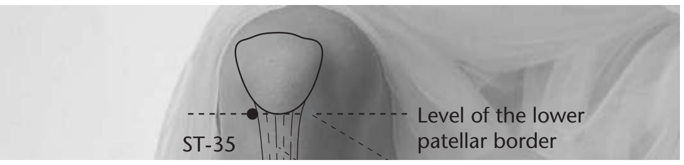
ST-35
Level of the lower patellar border

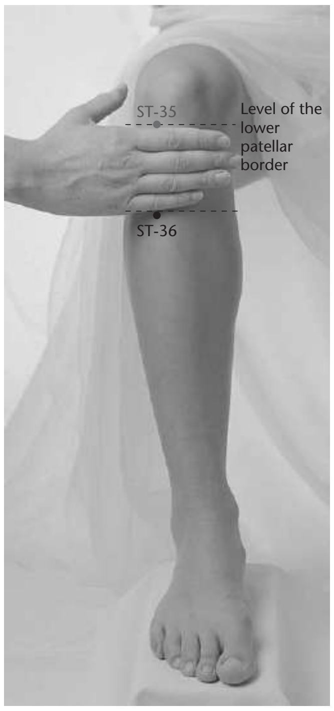
ST-35
Level of the lower patellar border
ST-36

Upper Great Emptiness SHANGJUXU ST-37

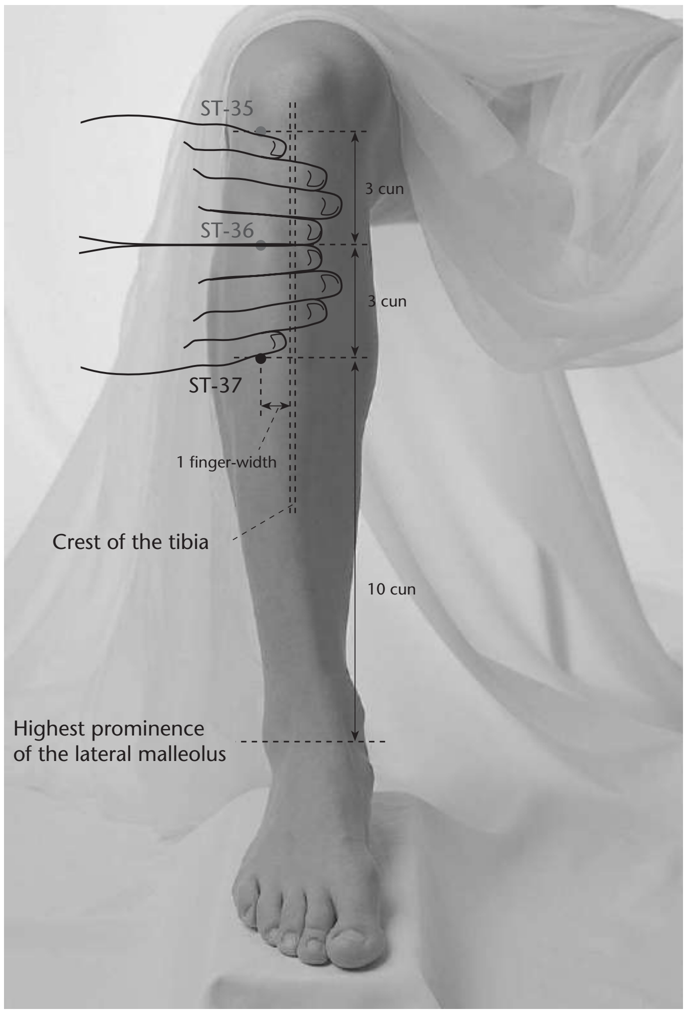
ST-35
3 cun
ST-36
3 cun
ST-37
1 finger-width
Crest of the tibia
10 cun
Highest prominence of the lateral malleolus

Location

6 cun distal to ➞ **ST-35** (at the knee joint space) or 3 cun distal to ➞ **ST-36** and 1 fingerbreadth lateral to the anterior crest of the tibia.

How to find

First, find ➞ **ST-35** ('lateral eye of the knee', level with the knee joint space). From ➞ **ST-35**, measure 6 cun (2 handbreadths) in a distal direction and, at that level, 1 fingerbreadth towards lateral (place the middle finger onto the anterior crest of the tibia). There, **ST-37** is located in a depression, which can be 'dynamically' palpated (choose the point according to pressure sensitivity). **Or:** Find the midpoint of the line joining ➞ **ST-35** and ➞ **ST-41** (in the depression between the tendons of the extensor digitorum longus and the hallucis longus, level with the highest prominence of the lateral malleolus) by using the spreading hands technique (➞ 2.3.3). **ST-37** is located 2 cun proximal and 1 fingerbreadth lateral to this point.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Regulates the Spleen, Stomach and Intestines, eliminates stagnation, clears Damp-Heat
- Opens the channel, alleviates pain

Special features

Lower *he*-sea point of the Large Intestine, Sea of Blood point.

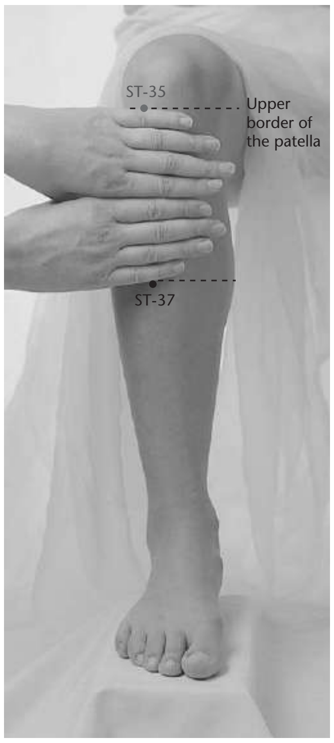
ST-35
Upper border of the patella
ST-37

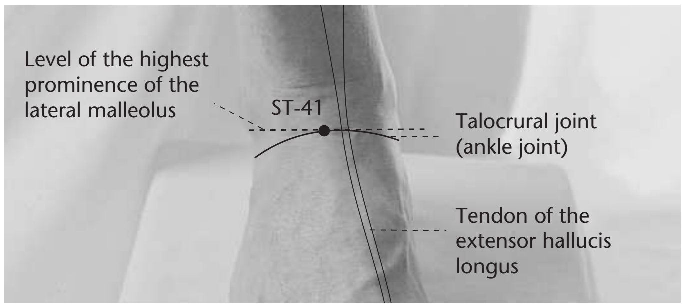
Level of the highest prominence of the lateral malleolus
ST-41
Talocrural joint (ankle joint)
Tendon of the extensor hallucis longus

ST-38 *Lines Opening* TIAOKOU

Location

At the midpoint of the line joining → **ST-35** and → **ST-41** and 1 fingerbreadth lateral to the anterior crest of the tibia.

How to find

Quick method: Spreading-hands technique (→ 2.3.3): place the small fingers on → **ST-35** ('lateral eye of the knee', level with the knee joint space) and → **ST-41** (in the depression between the tendons of the extensor digitorum longus and the hallucis longus, level with the highest prominence of the lateral malleolus) and determine the midpoint of this line, using the thumbs. **ST-38** is located in a depression 1 fingerbreadth (middle fingerbreadth, measured from the anterior crest of the tibia) lateral to the midpoint. **ST-38** may be palpated 'dynamically' (according to pressure sensitivity). **Or:** From → **ST-35**, measure 8 cun (→ 2 handbreadths and 3 fingerbreadths) distally and 1 fingerbreadth laterally.

Located on the same level are → **ST-40** (1 fingerbreadth more laterally) and → **BL-57** (dorsal aspect of the lower leg, on the midpoint of the line joining → **BL-40** and → **BL-60**).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Expels Wind-Damp, opens the channel, alleviates pain, benefits the shoulder

Special features

Very effective distal point for disorders of the shoulder. For limited range of motion of the shoulder joint: apply strong reducing needling techniques while the patient moves their shoulder joint.

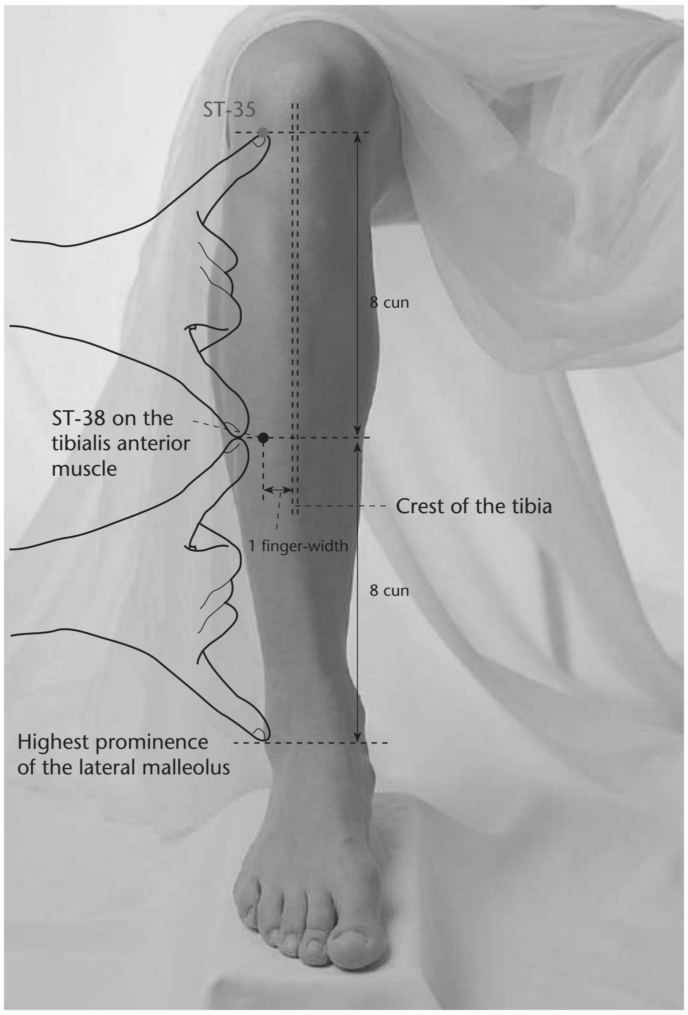
ST-35
8 cun
ST-38 on the tibialis anterior muscle
Crest of the tibia
1 finger-width
8 cun
Highest prominence of the lateral malleolus

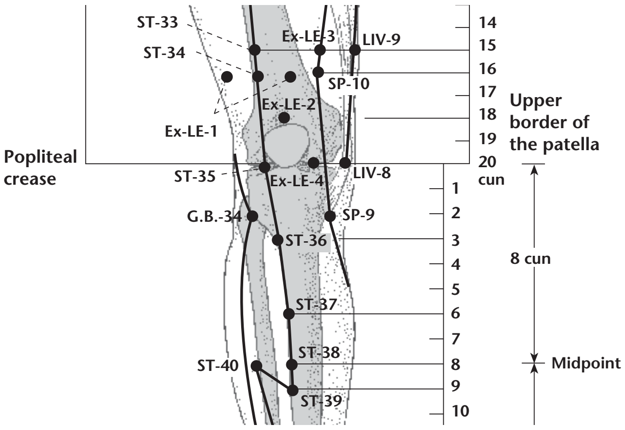
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
14
15
16
17
18
19
20
cun
1
2
3
4
5
6
7
8
9
10
Upper border of the patella
8 cun
Midpoint

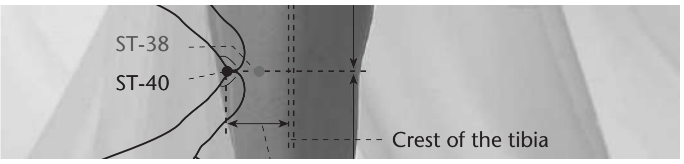
ST-38
ST-40
Crest of the tibia

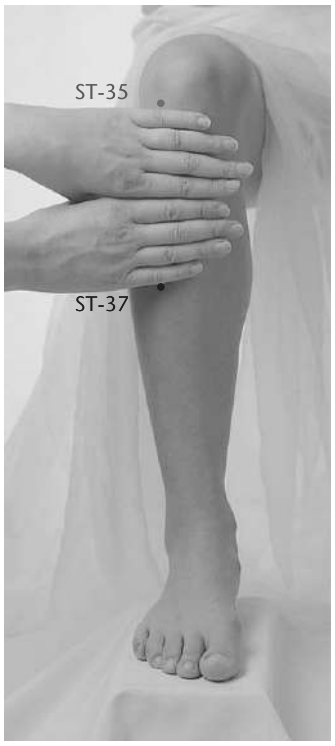

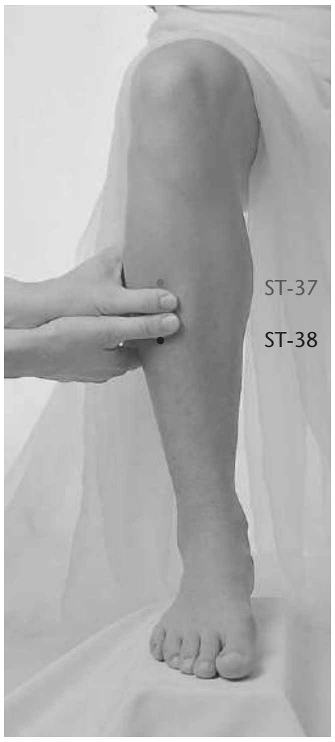

Lower Great Void XIAJUXU ST-39

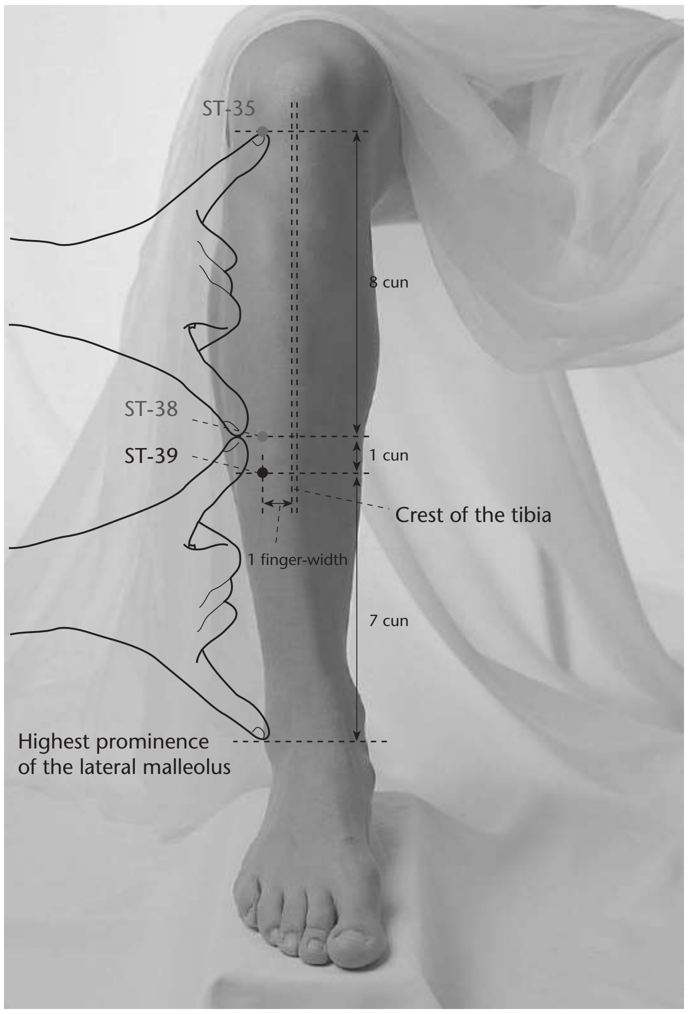
ST-35
8 cun
ST-38
ST-39
1 cun
Crest of the tibia
1 finger-width
7 cun
Highest prominence of the lateral malleolus

Location

1 cun distal to → **ST-38** (midpoint of the line joining → **ST-35** and → **ST-41**) and 1 fingerbreadth lateral to the anterior crest of the tibia.

How to find

Quick method: Spreading-hands technique (→ 2.3.3): Place the little fingers on → **ST-35** ('lateral eye of the knee', level with the knee joint space) and → **ST-41** (in the depression between the tendons of the extensor digitorum longus and the hallucis longus, level with the highest prominence of the lateral malleolus) and determine the midpoint of this line (level with → **ST-38**), using the thumbs. **ST-39** is located in a depression 1 cun distal and 1 fingerbreadth (middle fingerbreadth, measured from the anterior crest of the tibia) lateral to this point. **ST-39** may be palpated 'dynamically' (according to pressure sensitivity).

Located on the same level (7 cun proximal to the lateral malleolus) are → **G.B.-35** (on the posterior border of the fibula), → **G.B.-36** (on the anterior border of the fibula) and → **BL-58** (on the lateral border of the gastrocnemius muscle).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Moves the Small Intestine Qi, eliminates stagnation, clears Damp-Heat
- Opens the channel, alleviates pain

Special features

Lower *he*-sea point of the Small Intestine, Sea of Blood point.

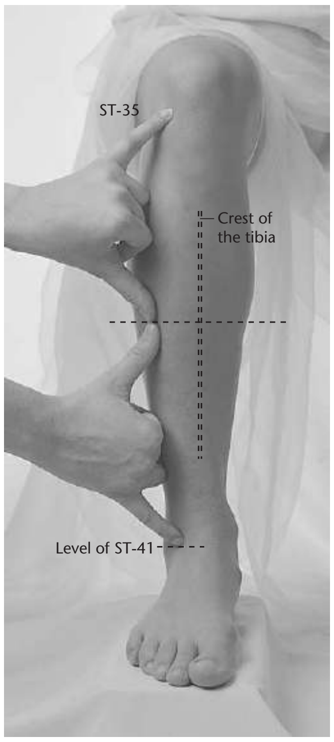
ST-35
Crest of the tibia
Level of ST-41

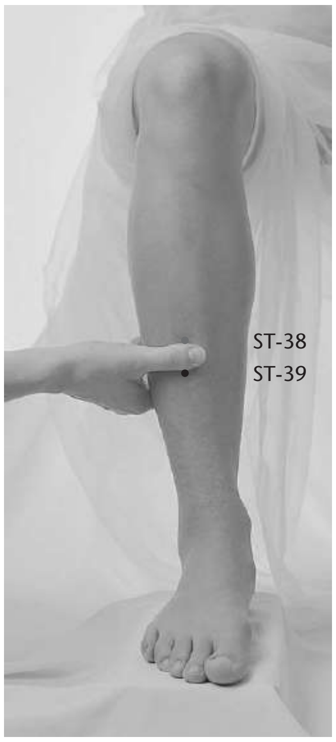

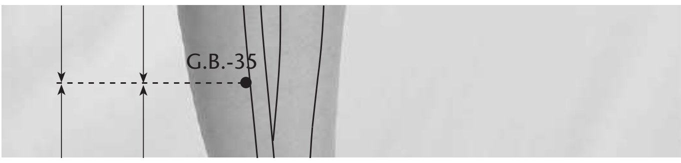

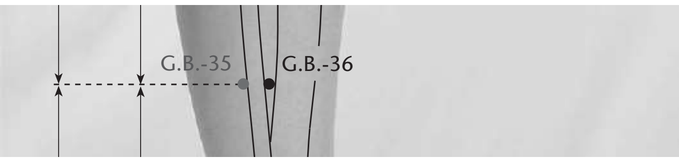

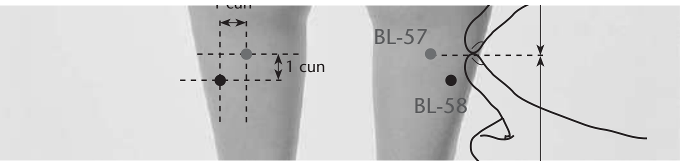
1 cun
BL-57
BL-58

ST-40 *Abundant Bulge* FENGLONG

Location

At the midpoint of the line joining ➞ **ST-35** and ➞ **ST-41** and 2 fingerbreadths lateral to the anterior crest of the tibia or 1 fingerbreadth lateral to ➞ **ST-38**, between the extensor digitorum longus and peroneus brevis muscles.

How to find

Quick method: Spreading-hands technique (➞ 2.3.3): Place the little fingers on ➞ **ST-35** ('lateral eye of the knee', level with the knee joint space) and ➞ **ST-41** (in the depression between the tendons of the extensor digitorum longus and the hallucis longus, level with the highest prominence of the lateral malleolus) and determine the midpoint of this line, using the thumb. **ST-40** is located in a depression 2 fingerbreadths (2 middle fingerbreadths, measured from the anterior crest of the tibia) lateral to this point. **ST-40** may be palpated 'dynamically' (according to pressure sensitivity). Located on the same level (8 cun proximal to the highest prominence of the lateral malleolus) are ➞ **ST-38** (1 fingerbreadth lateral to the tibia) and ➞ **BL-57** (on the dorsal aspect of the lower leg, halfway between ➞ **BL-40** and ➞ **BL-60**).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Transforms Dampness and Phlegm, clears Phlegm in the Lung and Heart, calms the *shen*

Special features

Luo-connecting point. Main acupuncture point for eliminating Phlegm.

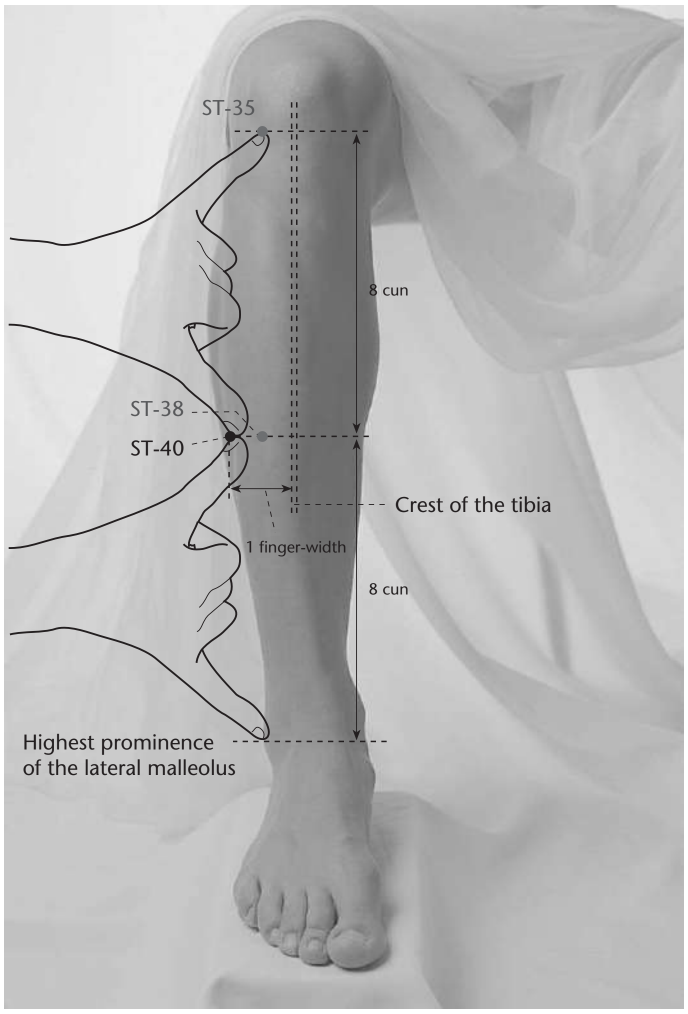
ST-35
8 cun
ST-38
ST-40
Crest of the tibia
1 finger-width
8 cun
Highest prominence of the lateral malleolus

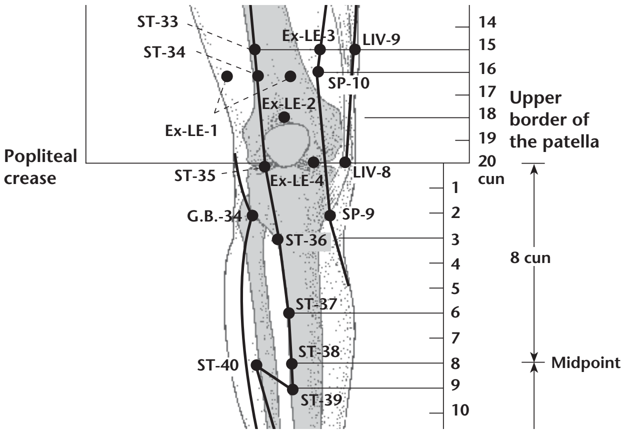
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
14
15
16
17
18
19
20 cun
Upper border of the patella
1
2
3
4
5
6
7
8
9
10
8 cun
Midpoint

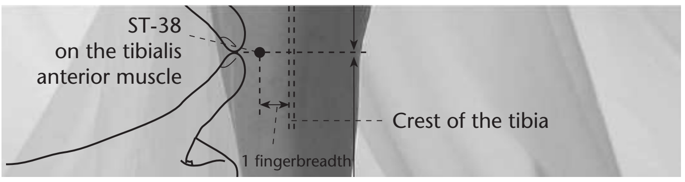
ST-38 on the tibialis anterior muscle
Crest of the tibia
1 fingerbreadth

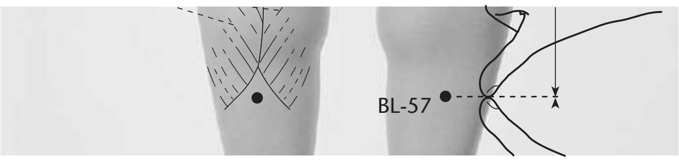

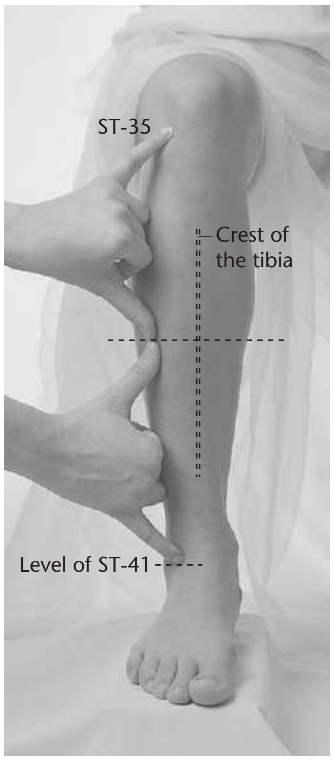
ST-35
Crest of the tibia
Level of ST-41

Stream Divide JIEXI ST-41

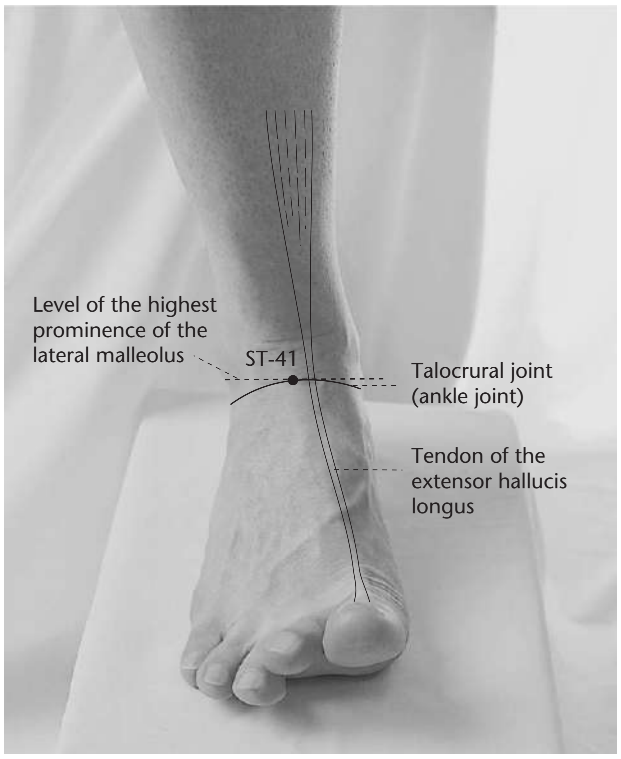
Level of the highest prominence of the lateral malleolus
ST-41
Talocrural joint (ankle joint)
Tendon of the extensor hallucis longus

Location

On the ankle, on the level of the highest prominence of the lateral malleolus, in the depression between the tendons of the extensor digitorum and the extensor hallucis longus.

How to find

With the patient extending their big toe against resistance, the tendon of the extensor hallucis longus muscle will become more distinct. **ST-41** is located lateral to this tendon, in an obvious depression on the level of the highest prominence of the lateral malleolus (→ 3.6.2) and ankle joint space.
Located on the same level is → **LIV-4** (medial to the tendon, halfway between → **SP-5** and **ST-41**).

Needling

0.5–1 cun vertically or obliquely deep to the tendons in a medial direction towards → **SP-5** or in a lateral direction towards → **G.B.-40**. Caution: Tibial artery/vein/nerve lie deep to this point.

Actions/Indications

- Clears Heat from the Stomach and Stomach channel
- Calms the *shen*
- Opens the channel and *luo* vessels, alleviates pain

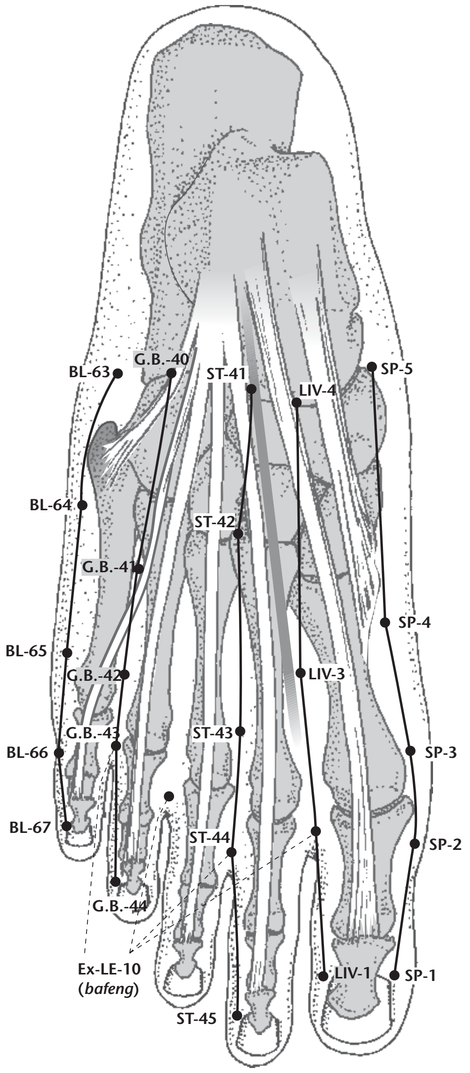
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

Special features

Jing-river point, Fire point, tonification point. Important local point for disorders of the ankle joint. Important distal point for frontal headaches caused by Stomach Heat or Stomach Fire.

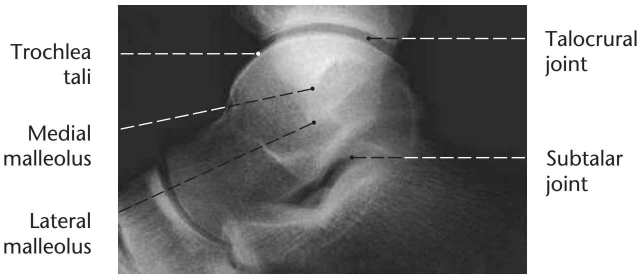
Trochlea tali
Medial malleolus
Lateral malleolus
Talocrural joint
Subtalar joint

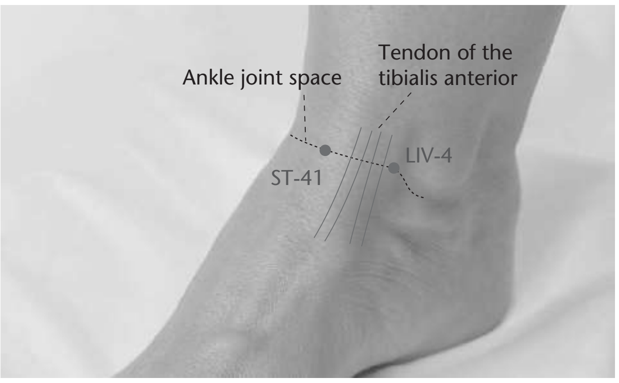
Ankle joint space
Tendon of the tibialis anterior
LIV-4
ST-41

ST-42 *Rushing Yang* CHONGYANG

Location

On the highest point of the dorsum of the foot, between the tendons of the extensor hallucis longus and the extensor digitorum longus, directly lateral to the point where the dorsalis pedis artery may be palpated. The point is bordered proximally by the 2nd and 3rd metatarsal bones and distally by the 2nd and 3rd cuneiform bones. **Alternative location:** Sometimes, this point may be located lateral to the medial portion of the extensor digitorum longus tendon (joining the 2nd toe).

How to find

With the palpating finger, glide from → **ST-43** (in the depression between the base of the 2nd and 3rd metatarsal bones) proximally towards → **ST-41** (in the depression between the tendons of extensor digitorum longus and hallucis longus). Locate **ST-42** at the highest point of the dorsum of the foot, in a depression lateral to the pulse. In most cases, **ST-42** is located lateral to the tendon of the extensor hallucis longus, which becomes more distinct by extending the big toe.

Needling

Vertically 0.2–0.5 cun. Caution: Dorsalis pedis artery.

Actions/Indications

- Clears Heat from the Stomach channel, regulates the Stomach
- Calms the *shen*
- Opens the channel and *luo* vessels

Special features

Yuan-source point, exit point.

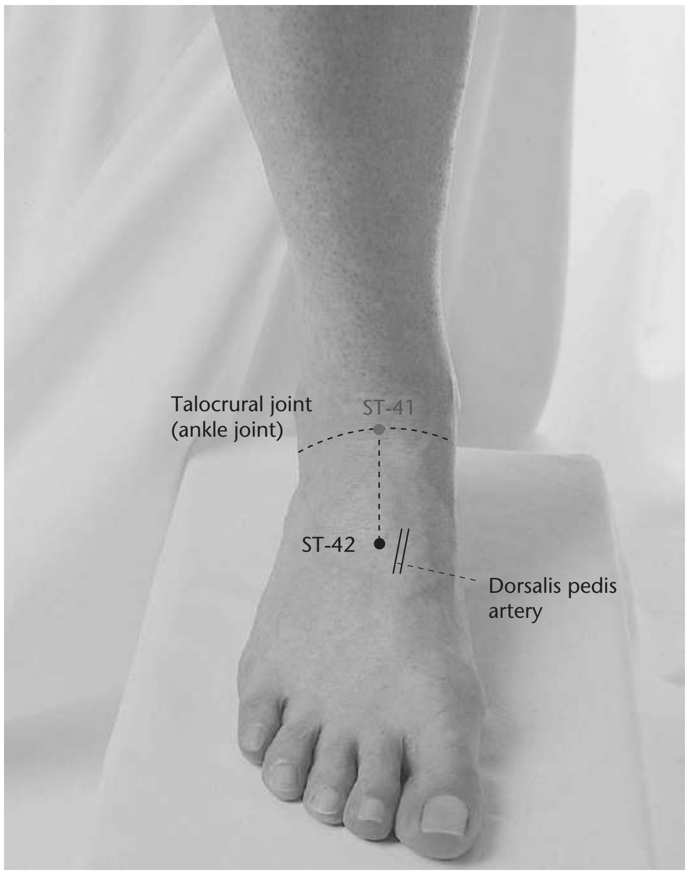
Talocrural joint
(ankle joint)
ST-41
ST-42
Dorsalis pedis
artery

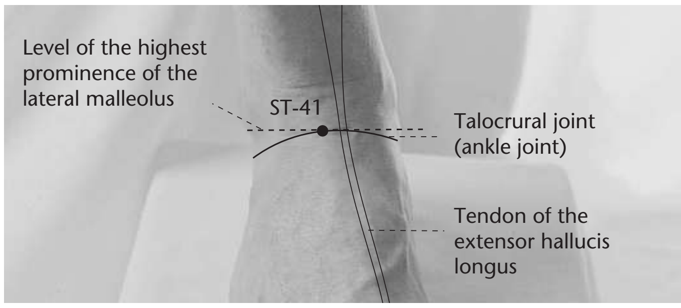
Level of the highest
prominence of the
lateral malleolus
ST-41
Talocrural joint
(ankle joint)
Tendon of the
extensor hallucis
longus

Sunken Valley XIANGU ST-43

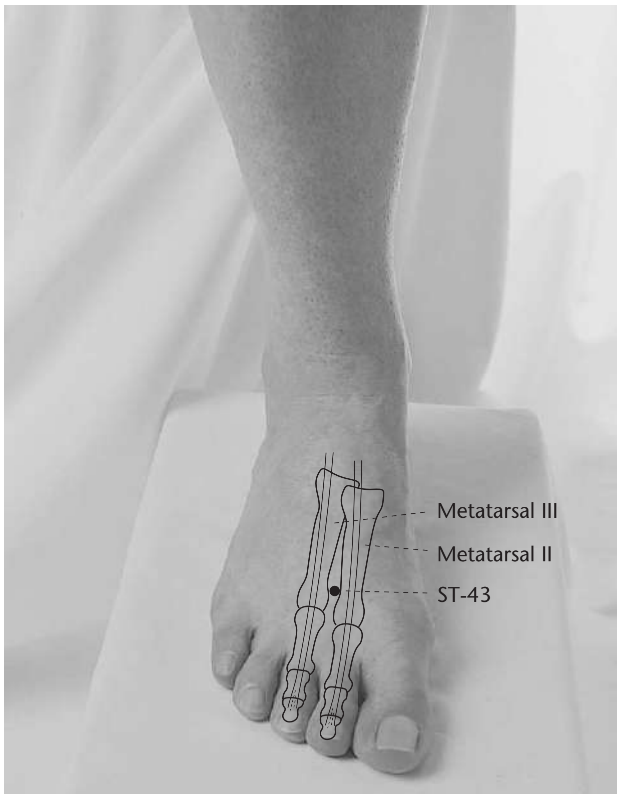
Metatarsal III
Metatarsal II
ST-43

Location

On the dorsum of the foot, in the depression between the 2nd and 3rd metatarsal bones, at the level of the junction of the shafts and the heads of these metatarsal bones.

How to find

Palpate from distal (toes) towards proximal (ankle) between the 2nd and 3rd metatarsal bones, until the finger glides into a depression distal to the metatarsophalangeal joint, where **ST-43** is located. It is approximately on the same level as the junction of the shafts and the heads of the two metatarsal bones.

Needling

Vertically or obliquely 0.5–1 cun

Actions/Indications

- Regulates the Spleen, Stomach and Intestines, reduces oedema

Special features

Shu-stream point, Wood point, generally used for 'inflammatory' joint disorders (*bi*-syndromes).

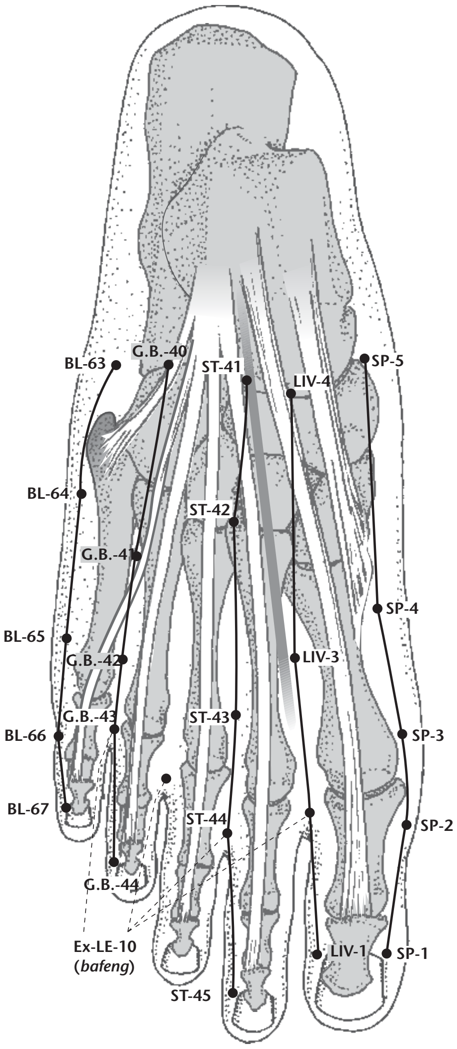
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
LIV-3
G.B.-42
G.B.-43
ST-43
SP-3
BL-66
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

ST-44 *Inner Court* NEITING

Location

Between the 2nd and the 3rd toes, proximal to the interdigital fold.

How to find

Locate **ST-44** just proximal to the interdigital fold between the 2nd and the 3rd toes.

Needling

0.5–1 cun vertically or obliquely in a proximal direction

Actions/Indications

- Opens the channel, expels Wind and clears Heat, alleviates pain
- Regulates the Intestines, clears Damp-Heat
- Calms the *shen*

Special features

Ying-spring point, Water point, Ma Dan Yang Heavenly Star point. Major point for clearing Heat from the head and face.

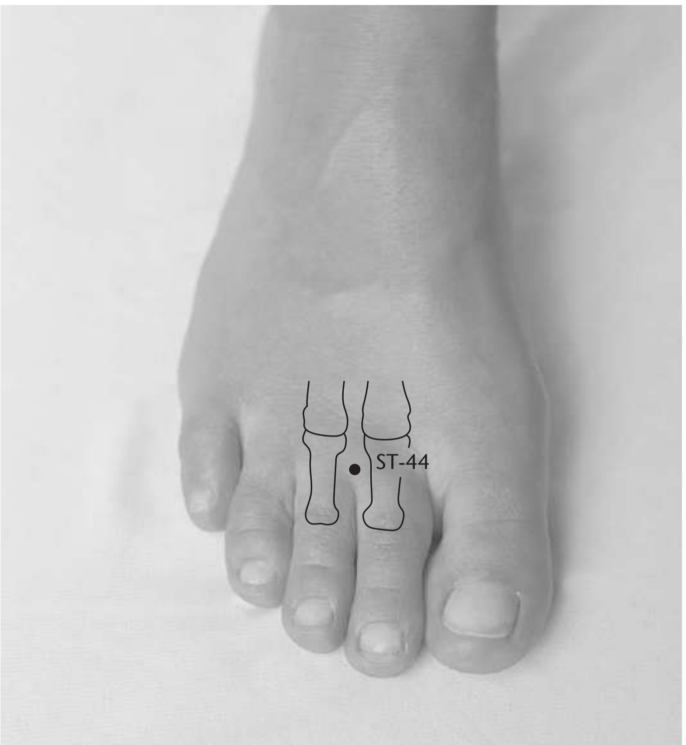

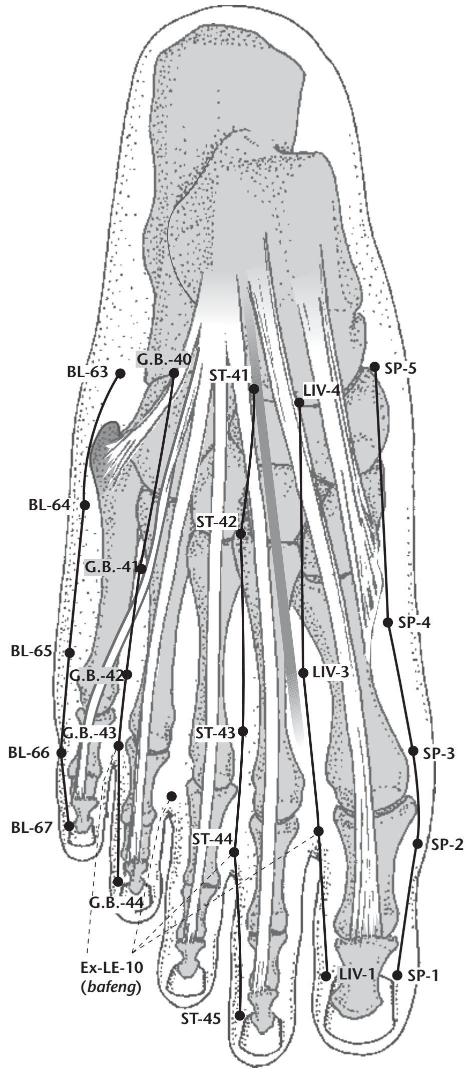
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

Strict Exchange LIDUI ST-45

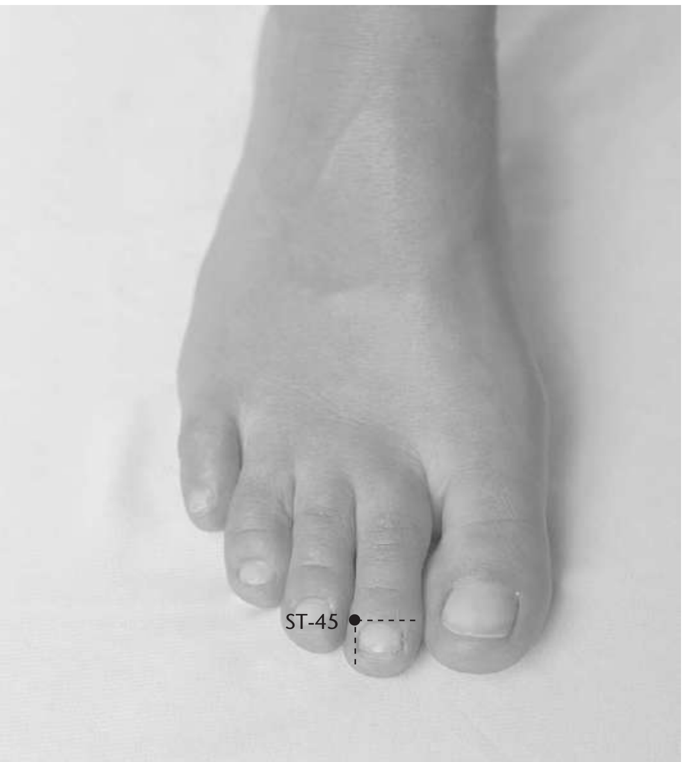

Location

On the 2nd toe, 0.1 cun from the lateral corner of the nail.

How to find

ST-45 is located at the junction of two tangents drawn along the lateral and proximal border of the nail of the 2nd toe, approximately 0.1 cun from the margin of the nail.

Needling

0.1 cun vertically or obliquely in a proximal direction or prick to bleed.

Actions/Indications

- Opens the channel and clears Heat from the *yangming* channels (L.I./ST)
- Calms and regulates the *shen*

Special features

Jing-well point, Metal point, sedation point, important distal point for all Heat disorders affecting the head. For severe insomnia due to Fire or Phlegm-Heat, apply moxibustion ('small Fire attracting big Fire' technique).

4.4 The Spleen Channel System – Foot *taiyin* (*zu tai yin jing luo*)

4.4.1 The Spleen Primary Channel (*zu tai yin jing*)

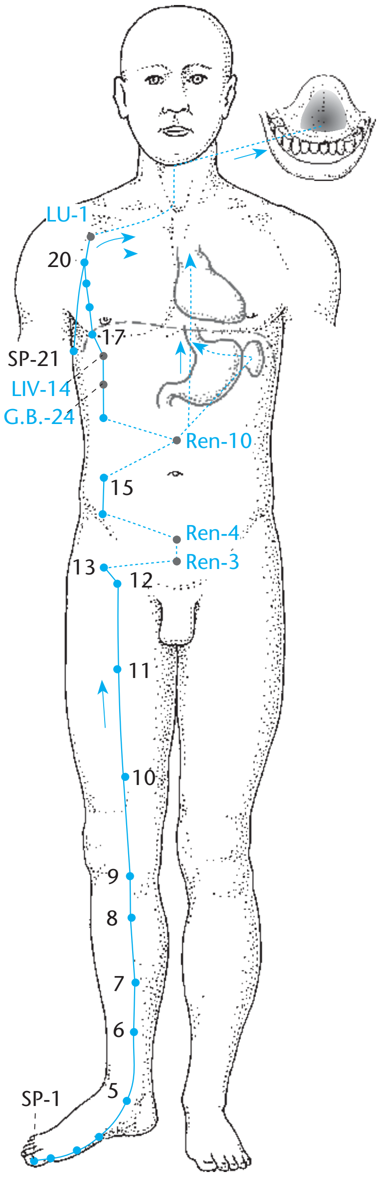
LU-1
20
17
SP-21
LIV-14
G.B.-24
Ren-10
15
Ren-4
Ren-3
13
12
11
10
9
8
7
6
5
SP-1

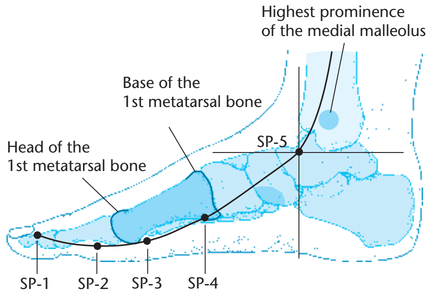
Highest prominence of the medial malleolus
Base of the 1st metatarsal bone
Head of the 1st metatarsal bone
SP-5
SP-1
SP-2
SP-3
SP-4

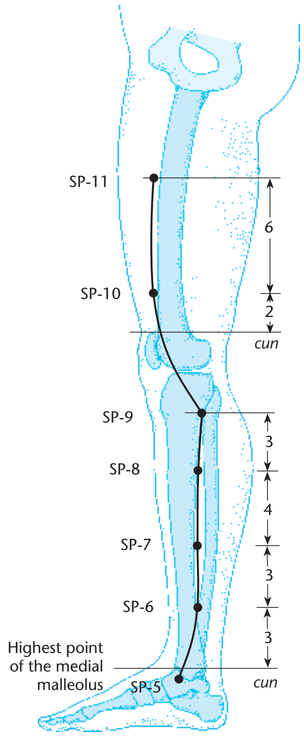
SP-11
6
SP-10
2
cun
SP-9
3
SP-8
4
SP-7
3
SP-6
3
Highest point of the medial malleolus
SP-5
cun

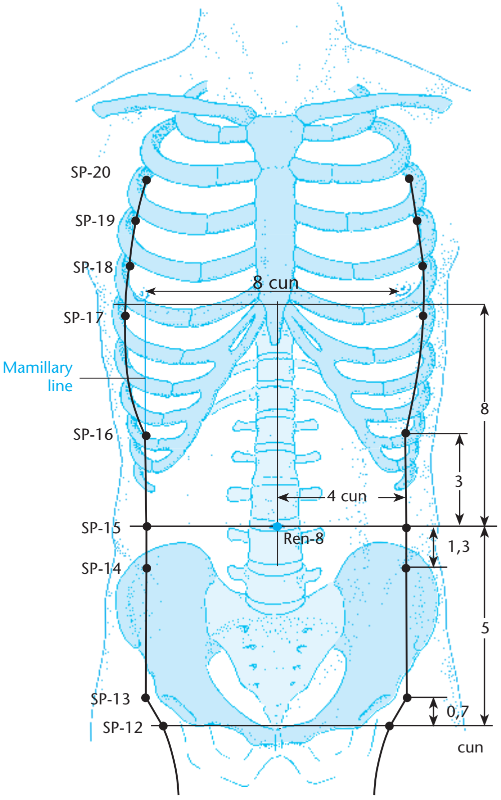
SP-20
SP-19
SP-18
8 cun
SP-17
Mamillary line
8
SP-16
3
4 cun
SP-15
Ren-8
1,3
SP-14
5
SP-13
0,7
SP-12
cun

Pathway

The **external pathway** of the Spleen primary channel begins at **SP-1** (*yinbai*) on the big toe at the medial corner of the nail. This point is reached by a branch that separates from the Stomach primary channel at → **ST-42** (*chongyang*) (foot Yin–Yang connection of the first great circuit).

➡ From **SP-1** the channel courses along the medial aspect of the foot

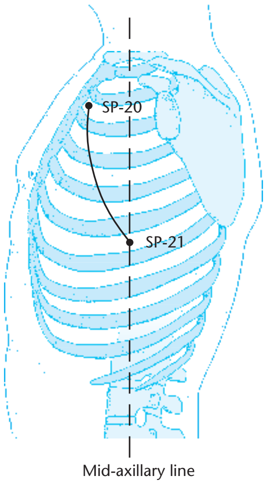

- ➡ ascends anterior to the medial malleolus and continues along the posterior border of the medial crest of the tibia
- ➡ passes the anteromedial region of the knee and thigh
- ➡ ascends the abdomen lateral to the midline to the second intercostal space, passing on its way → **Ren-3** (*zhongji*), → **Ren-4** (*guanyuan*), → **Ren-10** (*xiawan*), → **G.B.-24** (*riyue*) and → **LIV-14** (*qimen*)
- ➡ curves downward again and terminates 6 cun inferior to the axilla on the midaxillary line at **SP-21** (*dabao*).

According to Solinas et al (1998), its **inner pathway** branches off at → **Ren-10** (*xiawan*) and enters the Spleen (*pi*), its pertaining *zang*-Organ, and connects with the Stomach (*wei*), its paired *fu*-Organ. From the Stomach (*wei*), an **Interior branch** passes the diaphragm, ascends part of the oesophagus and enters the Heart (*xin*), where it connects with the Heart primary channel (*shou shao yin jing*) (deep Yin–Yin connection → 1.2). From **SP-20** (*zhourong*), a **further branch** ascends to → **LU-1** *(zhongfu)* at the level of the first intercostal space to form the hand–foot pairing (Yin axes of the first great circuit, *taiyin*). The branch ascends further to the tongue, reaches the root of the tongue and disperses on its lower surface.

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Heavy sensations of the head and body, sensation of weakness in the legs, general fever, pain in the lower cheek, muscular atrophy and weakness of the legs and medial hip region, cold knees, feet and legs, oedema

Interior (*li*) or *zangfu*-Organ signs and symptoms: Abdominal pain, diarrhoea or loose stools with undigested food, borborygmus, belching and dizziness, abdominal masses, loss of appetite, a sallow complexion, retention of urine

Excess (*shi*) → Spasms, foot pain

Deficiency (*xu*) → Abdominal fullness, borborygmus, diarrhoea with undigested food

Connections and meeting points of the Spleen primary channel (→ 1.2)

Connections with other channels

Stomach primary channel (*zu yang ming jing*)

Connection: Foot Yin–Yang connection of the first great circuit
Location: ST-42 SP-1 (on the foot).
Circulation: Circadian (according to the Organ clock)
Importance: Interior–Exterior relationship

Lung primary channel (*shou tai yin jing*)

Connection: Paired according to the six channel theory (hand–foot pairing) → *tai yin* (Yin axes of the first great circuit)
Location: SP-20 → LU-1 (on the thorax). A branch of the Spleen primary channel ascends from **SP-20** (*zhourong*) to **LU-1** (*zhongfu*)
Circulation: Non-circadian (**not** according to the Organ clock)
Importance: Above–below relationship

Heart primary channel (*shou shao yin jing*)

Connection: Deep Yin–*Yin* connection
Location: SP → HE (on the thorax). An internal branch of the Spleen primary channel spreads inside the Heart (*xin*) and connects with the Heart primary channel.
Circulation: Circadian (according to the Organ clock)
Importance: The Heart primary channel receives Nutritive Qi (*ying qi*) from the Spleen primary channel (first circulation of the *ying qi* → 1.1.4)

Connections with other *zangfu*-Organs

Spleen (*pi*), Stomach (*wei*), Heart (*xin*)

4.4.2 The Spleen Divergent Channel (*zu tai yin jing bie*)

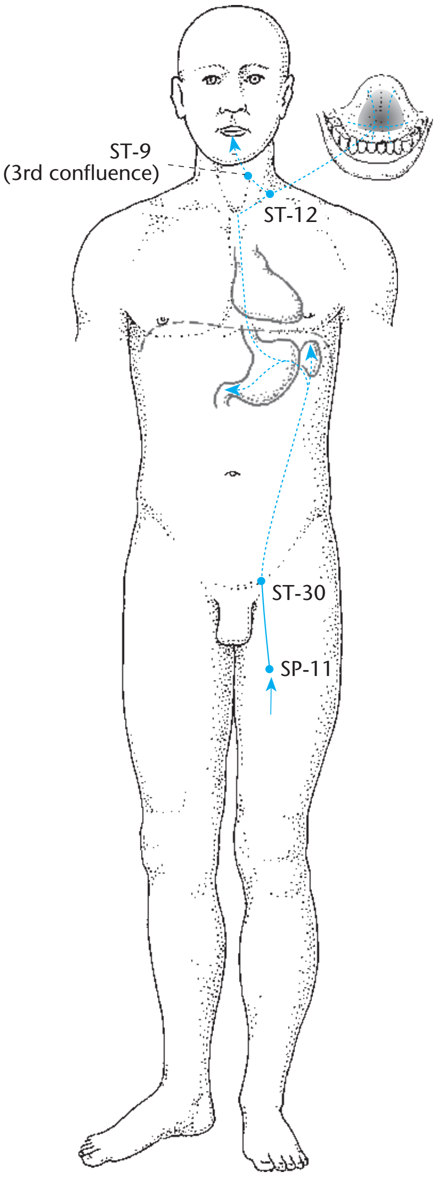

Pathway

The Spleen divergent channel separates from the Spleen primary channel near **SP-11** (*jimen*) on the anteromedial aspect of the thigh, at the medial border of the sartorius muscle. It

- ➡ penetrates the abdomen in the inguinal region at → **ST-30** (*qichong*)
- ➡ connects with the Stomach (*wei*), Spleen (*pi*) and Heart (*xin*)
- ➡ ascends along the oesophagus to the suprasternal fossa and travels laterally to → **ST-12** (*quepen*) in the supraclavicular fossa
- ➡ crosses the Stomach primary channel and the Stomach divergent channel at → **ST-9** (*renying*) where they meet to connect at one of the six *he*-confluences (here: ST/SP as third confluence → 1.3).

Clinical importance

- Strengthens the relationship between the Spleen and the Stomach (*zangfu*-Organs). Points on the Spleen primary channel can therefore be used for disorders of the Stomach, and vice versa points on the Stomach primary channel can treat disorders of the Spleen. In clinical practice, Spleen points are used for the whole gastrointestinal tract.
- Meets the Stomach primary channel at → **ST-30**; this is also a Sea of Water and Grain point, which can be used for many digestive disorders.
- Links the Spleen primary channel with the Heart; in combination with other points, SP points can be used for Qi and Blood deficiency as well as for psycho-emotional disorders.
- Connects the Spleen primary channel and the neck region and can therefore be used for chronic disorders of the neck and thyroid caused by retention of Phlegm (for example **SP-3** with **ST-40**, **P-5**).
- Connects the Spleen primary channel with the mouth and tongue and can therefore be used for disorders/ulcers of the tongue and mouth caused by retention of Phlegm (for example **SP-2, SP-3, SP-6**).

4.4.3 The Spleen Sinew Channel (*zu tai yin jing jin*)

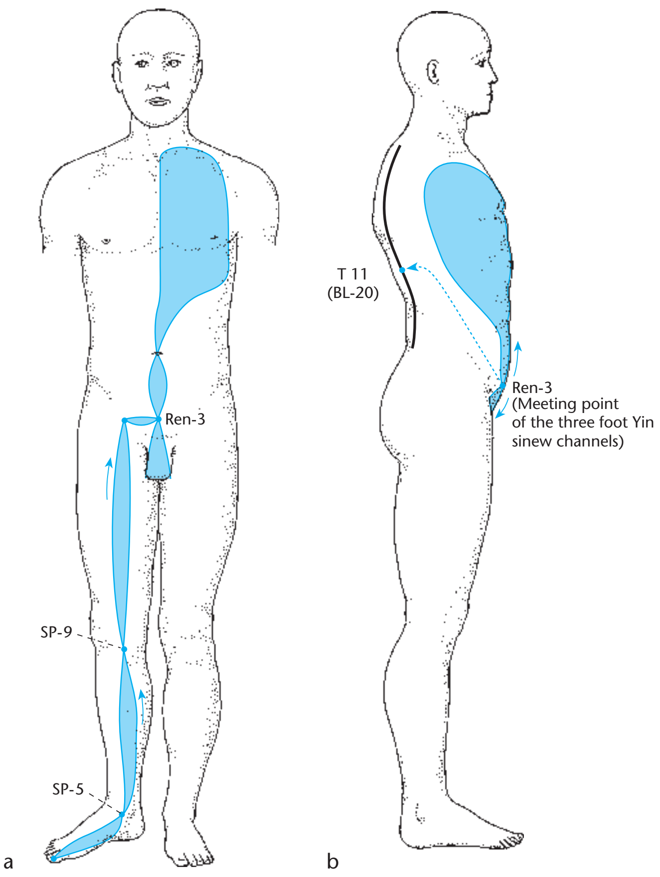

Pathway

The Spleen sinew channel begins on the medial aspect of the big toe at **SP-1** (*yinbai*)

➡ runs along the medial border of the foot

➡ binds (*jie*) anterior to the medial malleolus at **SP-5** (*shangqiu*)

➡ ascends along the medial aspect of the tibia to **SP-9** (*yinlingquan*) and binds (*jie*) anterior to the medial head of the tibia

➡ runs along the medial aspect of the thigh and binds (*jie*) on the medial aspect of the inguinal region

➡ continues to → **Ren-3** (*zhongji*) where it meets the other sinew channels of the foot Yin channels

At → **Ren-3** the sinew channel separates into three branches:

- **One branch** runs to the genital region
- **A further branch** ascends along the anterior midline, covering the *ren mai* and binding (*jie*) at the umbilical region. It penetrates the abdomen and spreads in the costal and thoracic region.
- **A third branch** penetrates deeper into the body and ends at the spinal region at → **BL-20** (*pishu*).

Clinical importance

- **Pathology:** Distending pain of the big toe, pain at the medial malleolus, cramps and pain in the calves, pain of the medial aspect of the knees and thighs, pain in the inguinal and genital region (radiating to the umbilicus, thorax and back).
- **Indication:** Mainly for *bi*-syndromes (Painful Obstruction Syndrome) and *wei*-syndrome (atrophy syndrome) along the Spleen channel, especially along the medial aspect of the leg. The area covered by the Spleen sinew channel is larger than that covered by the Spleen primary channel. This explains why Spleen primary channel points can also be used for disorders and diseases of the external genitalia and the umbilical region.

4.4.4 The Spleen *luo*-Connecting Vessel System (*zu tai yin luo mai*)

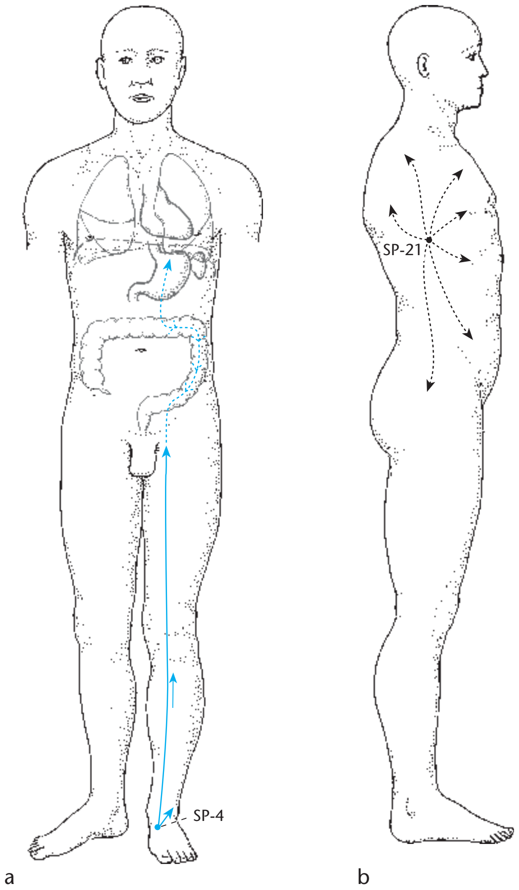

Pathway

The Spleen *luo*-connecting channel separates from the Spleen primary channel at its *luo*-connecting point **SP-4** (*gongsun*) (→ 8.1.2). It forms a three-dimensional reticulate network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

➡ Horizontal divisions run to the Interiorly–Exteriorly paired Stomach primary channel, according to some schools of thought (for example Ngyen Van Nghi → Appendix) they travel as a **transverse** Spleen *luo*-connecting vessel to the *yuan*-source point → **ST-42** (*chongyang*).

- A **longitudinal** division follows the pathway of the Spleen primary channel, enters the abdomen in the inguinal region and connects with the Large Intestine (*dachang*) and the Stomach (*wei*).

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*)**: Colicky pain in the Stomach and Large Intestine
- **Deficiency (*xu*)** → *Gu zhang* (drum-like abdominal distension) such as digestive disorders in children, abdominal fullness and distension, ascites
- **Counterflow Qi:** Acute vomiting, diarrhoea, severe abdominal pain

The great *luo*-connecting vessel of the Spleen (*zu tai yin luo mai*)

Pathway

The great *luo*-connecting vessel of the Spleen begins on the lateral aspect of the thorax at → **SP-21** (*dabao*). It contains the Blood of the *luo*-connecting vessel and disperses to the vessels of the whole body, covering them in a net-like fashion (→ 1.5, → 8.1.2).

Clinical importance (→ 8.1.2)

- **Excess (*shi*)**: Pain in the whole body, polyarthritis, *bi*-Syndrome (Painful Obstruction Syndrome)
- **Deficiency (*xu*)**: Muscular atrophy and weakness, weakness of the joints, weakness in the whole body.

4.4.5 Cutaneous Region (*tai yin pi bu*)

See description and figures → 1.6

4.4.6 Points of the Spleen Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): SP-3** (*taibai*) 🟦⬛
- ***Luo*-connecting point (→ 8.1.2.): SP-4** (*gongsun*) 🟦⬛/ SP-21 (*dabao*) 🟦🟦
- ***Xi*-cleft point (→ 8.1.3): SP-8** (*diji*) 🟦⬛
- **Associated Back-*shu* point (→ 8.1.4): BL-20** (*pishu*) 🟦🟦
- **Associated Front-*mu* point (→ 8.1.5): LIV-13** (*zhangmen*) 🟦🟦
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood): **SP-1** (*yinbai*) 🟦⬛
  ying-spring point (Fire), tonification point: **SP-2** (*dadu*)
  shu-stream point (Earth), *ben* point: **SP-3** (*taibai*) 🟦⬛
  jing-river point (Metal), sedation point: **SP-5** (*shangqiu*) 🟦⬛
  he-sea point (Water): **SP-9** (*yinlingquan*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8) of the *chong mai*: SP-4** (*gongsun*) 🟦⬛
- **Lower *he*-sea points (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - with the Kidney and Liver primary channels (meeting point of the three foot *Yin* channels and the *yin qiao mai**: **SP-6** (*sanyinjiao*) 🟦🟦
  - with the LIV and T.B.* channels and the *yin wei mai:* **SP-12** (*chongmen*)
  - with the LIV channel and the *yin wei mai*: **SP-13** (*fushe*)
  - with the *yin wei mai*: **SP-15** (*daheng*), **SP-16** (*fuai*)
  - of other channels with the SP channel: **LU-1**, **G.B.-24**, **LIV-14**, **Ren-3**, **Ren-4**, **Ren-10**, **Ren-17***
- **Gao Wu command point (→ 8.1.11):** –
- **Window of Heaven point (→ 8.1.12):** –
- **Points of the Four Seas (→ 8.1.13):** –
- **Ma Dan Yang Heavenly Star point(→ 8.1.14):** –
- **Sun Si Miao Ghost point (→ 8.1.15): SP-1** (*yinbai*) 🟦⬛
- **Other functional points:**
  – important point for diarrhoea: **SP-4** (*gongsun*) 🟦⬛
  – important point for Damp *bi*-syndrome: **SP-5** (*shangqiu*) 🟦⬛
  – important point for eliminating Dampness: **SP-9** (*yinlingquan*) 🟦🟦
  – important point for regulating the Blood: **SP-10** (*xuehai*) 🟦⬛
  – important point for general pain: **SP-21** (*dabao*) 🟦🟦

Points according to region

- **Local points (→ 8.2.1):** knee region – **SP-9** (*yinlingquan*) 🟦🟦; Large Intestine – **SP-15** (daheng) 🟦⬛; knee – **SP-9** (*yinlingquan*) 🟦🟦; foot – **SP-5** (*shangqiu*) 🟦⬛; toes – **SP-3** (*taibai*) 🟦⬛
- **Adjacent points (→ 8.2.1):** Spleen/Stomach – **SP-15** (*daheng*) 🟦⬛; hypochondrium – **SP-21** (*dabao*) 🟦🟦; knee – **SP-10** (*xuehai*) 🟦⬛; toes – **SP-4** (*gongsun*) 🟦⬛, **SP-5** (*shangqiu*) 🟦⬛, **SP-9** (*yinlingquan*) 🟦🟦
- **Distal points (→ 8.2.1):** for the Spleen/Stomach, upper abdominal region – **SP-4** (*gongsun*) 🟦⬛; for the female genital organs, urogenital region – **SP-6** (*sanyinjiao*) 🟦🟦; lower abdomen – **SP-1** (*yinbai*) 🟦⬛; for the hand and knee – **SP-5** (*shangqiu*) 🟦⬛

Specific points according to the channel pathway (in numerical order):

- **SP-1** (*yinbai*) 🟦⬛: Sun Si Miao Ghost point (→ 8.1.15); distal point for the lower abdomen (→ 8.2.1); *jing*-well point (Wood) (→ 8.1.6)
- **SP-2** (*dadu*): *ying*-spring point (Fire) (→ 8.1.6), tonification point
- **SP-3** (*taibai*) 🟦⬛: *yuan*-source point (→ 8.1.1), *shu*-stream point (→ 8.1.6), *ben* point (Five Phase point) (→ 8.1.6)

---

* Mentioned by only some authors.

- **SP-4** (*gongsun*) 🟦⬛: *luo*-connecting point of the Spleen (→ 8.1.2); opening point of the *chong mai* (→ 8.1.8); distal point for the Spleen/Stomach, upper abdomen (→ 8.2.1); important point for diarrhoea; adjacent point for the toes (→ 8.2.1)
- **SP-5** (*shangqiu*) 🟦⬛: *jing*-river point (Metal) (→ 8.1.6); Sedation point; important point for Damp *bi*-syndrome; local point for the foot (→ 8.2.1); distal point for the hand and knee (→ 8.2.1)
- **SP-6** (*sanyinjiao*) 🟦🟦: *jiaohui*-meeting point with the KID and LIV channels and the *yin qiao mai* (→ 8.1.10) (meeting point of the three foot Yin channels); distal point for the female genital system, the urogenital region (→ 8.2.1)
- **SP-8** (*diji*) 🟦⬛: *xi*-cleft point (→ 8.1.3)
- **SP-9** (*yinlingquan*) 🟦🟦: *he*-sea point (Water) (→ 8.1.6); important point for eliminating Dampness; local point for the knee region (→ 8.2.1); local point for the knees (→ 8.2.1)
- **SP-10** (*xuehai*) 🟦⬛: important point for regulating the Blood; adjacent point for the knees (→ 8.2.1)
- **SP-12** (*chongmen*): *jiaohui*-meeting point with the LIV and T.B.* channels and the *yin wei mai* (→ 8.1.10)
- **SP-13** (*fushe*): *jiaohui*-meeting point with the LIV channel and the *yin wei mai* (→ 8.1.10)
- **SP-15** (*daheng*) 🟦⬛: local point for the Large Intestine (→ 8.2.1); adjacent point for the Spleen/Stomach (→ 8.2.1); *jiaohui*-meeting point with the *yin wei mai* (→ 8.1.10)
- **SP-16** (*fuai*): *jiaohui*-meeting point with the *yin wei mai* (→ 8.1.10)
- **SP-21** (*dabao*) 🟦🟦: adjacent point for the hypochondrium (→ 8.2.1); Main-*luo*-connecting point of the Spleen (→ 8.1.2); important point for generalised pain

General location help

- **SP-13 – SP-16**: these points are all located 4 cun lateral to the anterior midline (= mamillary line) → **SP-13**: 0.7 cun superior to **SP-12; SP-14**: 1.3 cun inferior to the umbilicus; **SP-15**: at the level of the umbilicus; **SP-16**: 3 cun superior to the umbilicus.
- **SP-17 – SP-20**: these points are all located 6 cun lateral to the anterior midline in the fifth to second intercostal spaces.

---

*Mentioned by only some authors.

Ch04.4-F10028.qxd 2/23/08 7:08 PM Page 181

Hidden White YINBAI **SP-1**

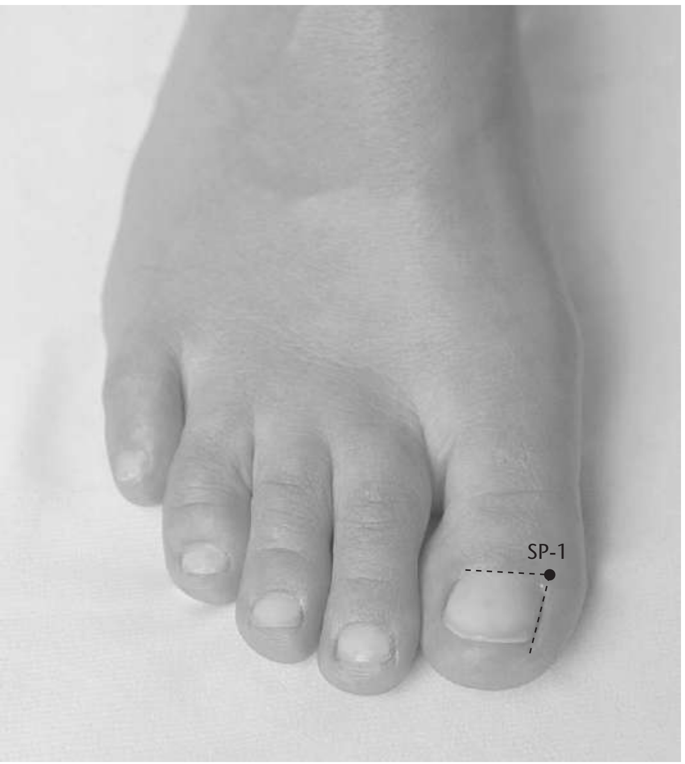

Location

On the big toe, 0.1 cun from the medial corner of the nail.

How to find

SP-1 is located at the junction of two tangents drawn along the medial and proximal border of the nail of the big toe, approximately 0.1 cun from the margin of the nail.

Needling

0.1–0.2 cun vertically or obliquely in a proximal direction or prick to bleed with a needle, lancet or three-edged needle.

Actions/Indications

- Regulates the Blood
- Regulates the Spleen
- Unbinds the chest
- Calms the *shen*
- Revives consciousness

Special features

Jing-well point, Wood point, Sun Si Miao Ghost point, entry point, important distal point for the lower abdomen.

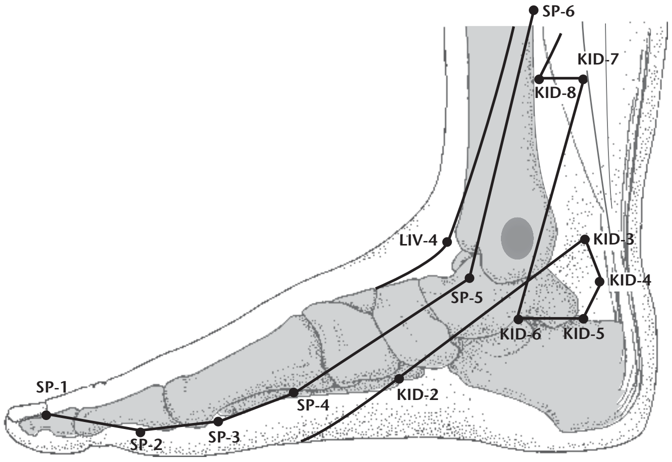
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-2
SP-3

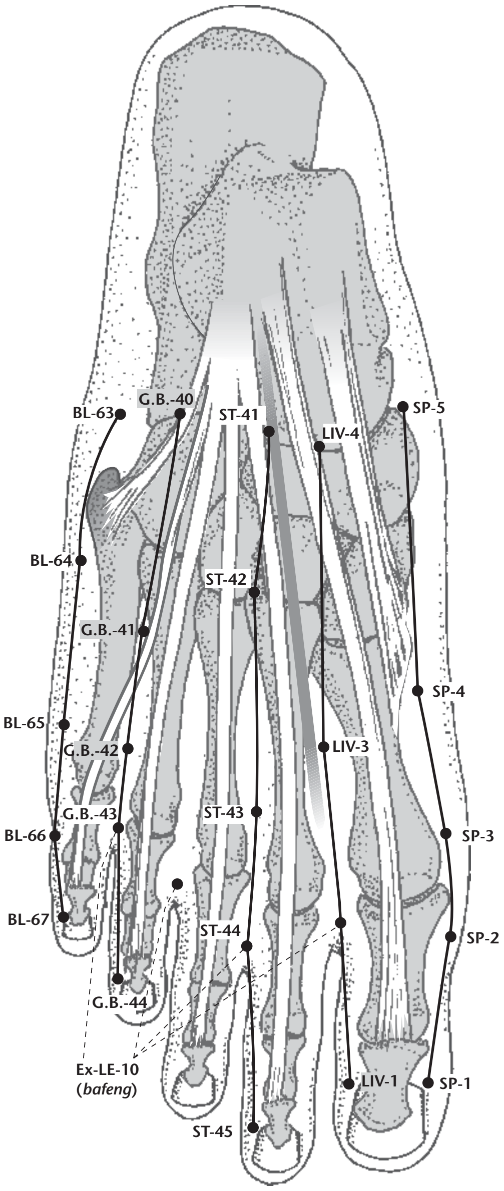
BL-63
G.B.-40
ST-41
SP-5
LIV-4
BL-64
ST-42
G.B.-41
SP-4
BL-65
LIV-3
G.B.-42
G.B.-43
ST-43
SP-3
BL-66
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

SP-2 *Great Metropolis* DADU

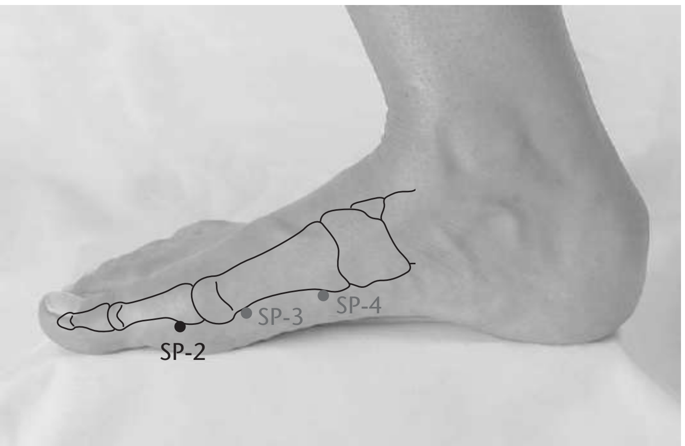

Location

On the medial aspect of the big toe, at the junction of the shaft and the base of the proximal phalanx, distal to the 1st metatarsophalangeal joint.

How to find

On the medial aspect of the big toe, palpate from distal to proximal towards the metatarsophalangeal joint, until you can feel the junction of the shaft with the base of the proximal phalanx. **SP-2** is located distally to the base of the bone, which forms a distinct step, and slightly inferiorly to its outer curvature, at the border of the red and white skin (sole/dorsum of the foot). Located in comparable positions are → **BL-66** on the lateral border of the foot, → **L.I.-2** on radial aspect and → **S.I.-2** on the ulnar aspect of the hand.

Needling

0.2–0.5 cun vertically or obliquely, just below the lower border of the bone. Caution: Painful point.

Actions/Indications

- Regulates and supports the Spleen, harmonises the Lower Burner and clears Heat

Special features

Ying-spring point, Fire point, tonification point.

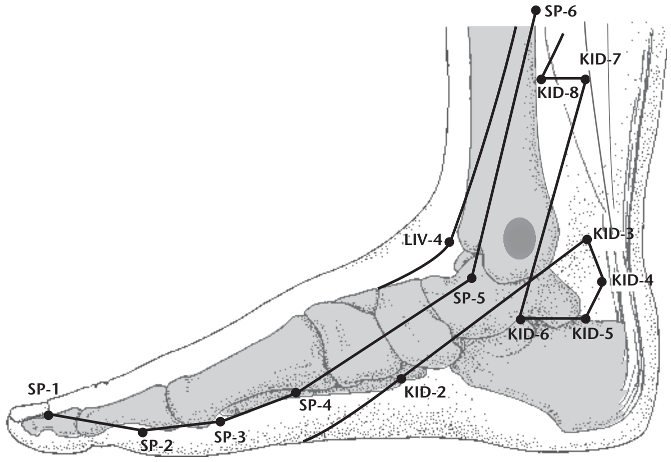
SP-6
KID-7
KID-8
KID-3
LIV-4
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

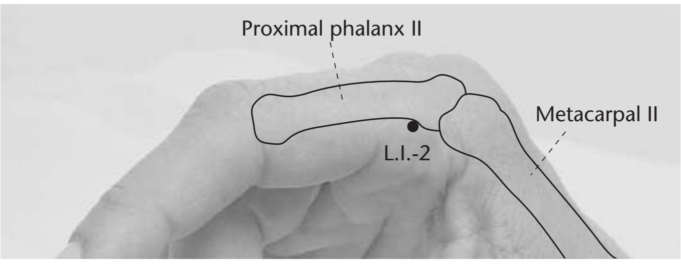
Proximal phalanx II
Metacarpal II
L.I.-2

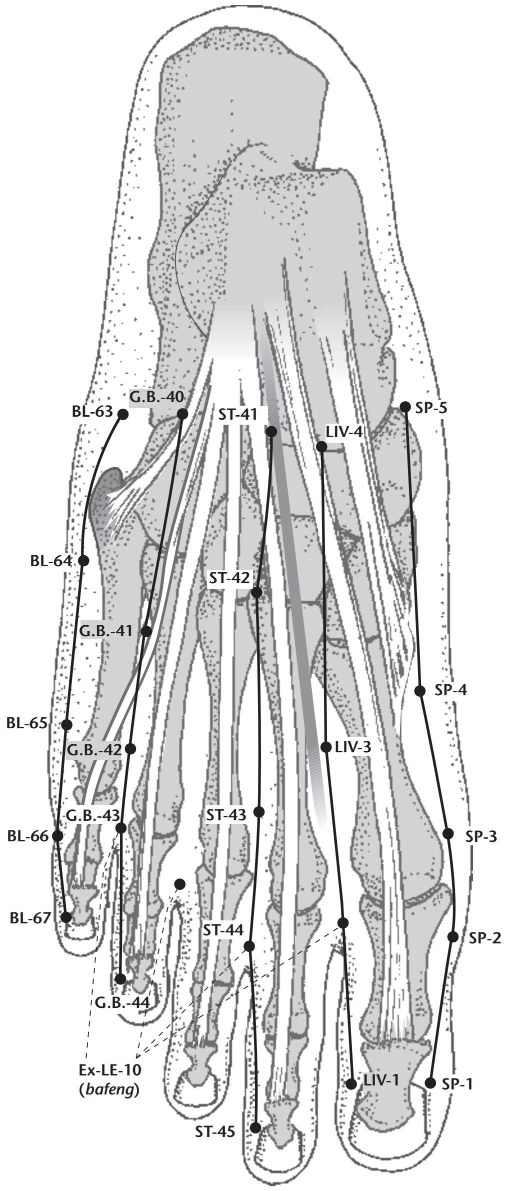
BL-63
G.B.-40
ST-41
SP-5
LIV-4
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

Supreme White TAIBAI **SP-3**

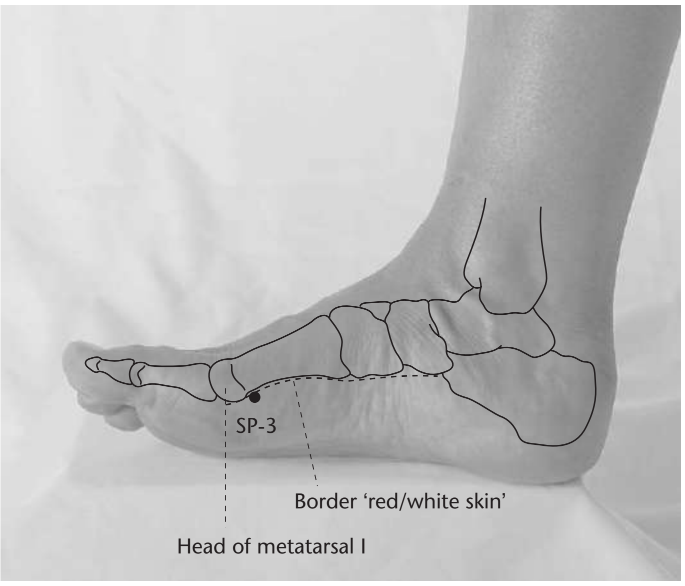
SP-3
Border 'red/white skin'
Head of metatarsal I

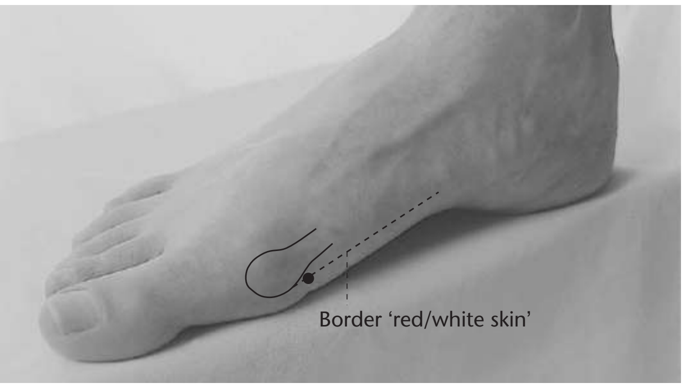
Border 'red/white skin'

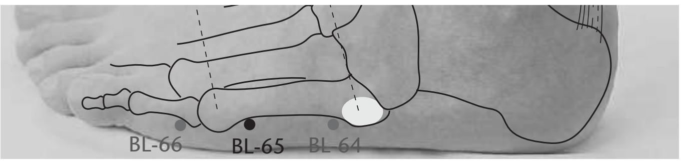
BL-66
BL-65
BL-64

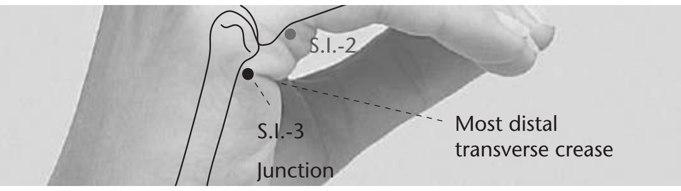
S.I.-2
S.I.-3
Junction
Most distal transverse crease

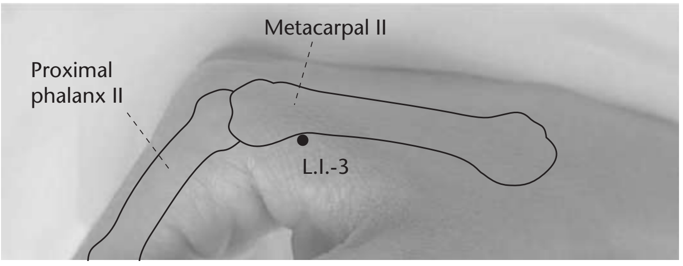
Metacarpal II
Proximal phalanx II
L.I.-3

Location

On the medial aspect of the foot, in the depression proximal to the head of the 1st metatarsal bone, at the border of the red and white skin.

How to find

When palpating along the border of the red and white skin on the medial aspect of the foot from distal (toes) to proximal (ankle), the metatarsophalangeal joint forms a prominent landmark. **SP-3** is located just proximal to this joint in a clearly palpable depression, which marks the junction of the head and the shaft of the 1st metatarsal bone.

Located in a comparable position is → **BL-65** (proximal to the head of the 5th metatarsal bone) on the lateral border of the foot. In comparable locations on the borders of the hand are → **L.I.-3** on the radial aspect and → **S.I.-3** on the ulnar aspect of the hand.

Needling

0.5–1 cun vertically or obliquely from medial towards lateral. Caution: Needling this point tends to be painful.

Actions/Indications

- Tonifies the Spleen, harmonises the Lower Burner
- Eliminates Dampness and Damp-Heat
- Regulates Qi locally

Special features

Yuan-source point, *shu*-stream point, Earth point, *ben* point (Five Phases). Important point for tonifying the Spleen.

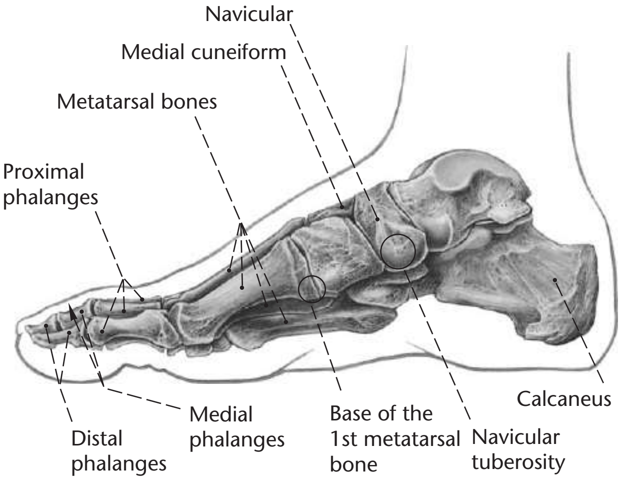
Navicular
Medial cuneiform
Metatarsal bones
Proximal phalanges
Distal phalanges
Medial phalanges
Base of the 1st metatarsal bone
Navicular tuberosity
Calcaneus

SP-4 *Grandfather Grandson* GONGSUN

Location

In the depression distal to the base of the 1st metatarsal bone, on the border of the red and white skin.

How to find

When palpating along the border of the red and white skin (sole/dorsum of the foot) on the medial aspect of the midfoot from proximal towards medial, the base of the 1st metatarsal bone (→ 2.6.2) forms a distinct bony prominence. **SP-4** is located in the depression distal to this prominence, at the junction of the shaft and the base of the bone.

→ **SP-3** is located slightly more distally at the junction of the head and the shaft of the 1st metatarsal bone. → **BL-64** is located in a comparable position on the lateral aspect of the midsection of the foot (at the junction of the shaft and the base of the 5th metatarsal bone).

Needling

0.5–1 cun vertically or obliquely from medial towards lateral. Caution: Needling may be painful.

Actions/Indications

- Strengthens and regulates the Spleen Qi, eliminates Dampness
- Calms the *shen*
- Benefits the Heart and chest
- Regulates the *chong mai*
- Local point, for example for metatarsalgia

Special features

Luo-connecting point, opening (master) point of the *chong mai*. Important point for distension and pain in the upper and lower abdomen caused by Qi stagnation, Blood stasis or obstruction due to Dampness.

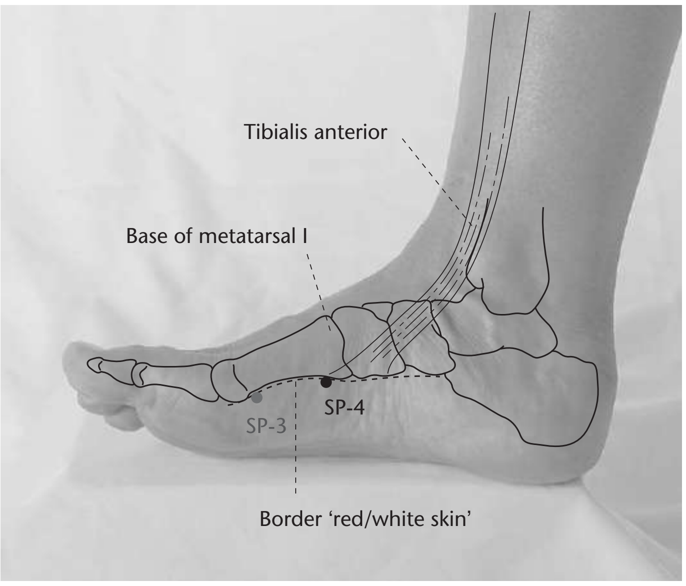
Tibialis anterior
Base of metatarsal I
SP-3
SP-4
Border 'red/white skin'

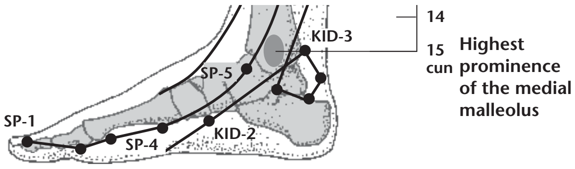
14
15
cun
Highest prominence of the medial malleolus
KID-3
SP-5
SP-1
SP-4
KID-2

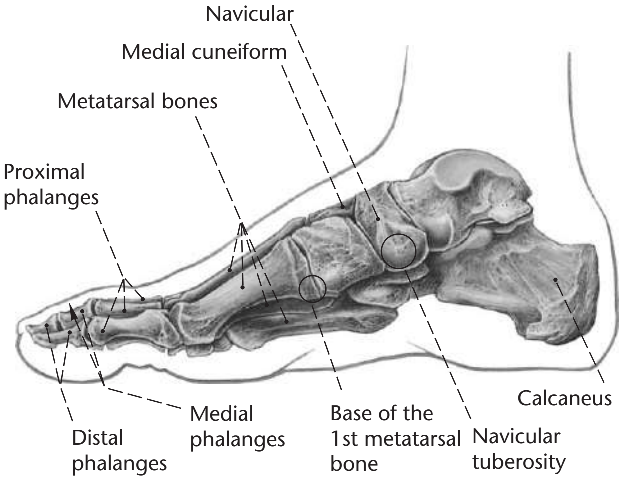
Navicular
Medial cuneiform
Metatarsal bones
Proximal phalanges
Calcaneus
Distal phalanges
Medial phalanges
Base of the 1st metatarsal bone
Navicular tuberosity

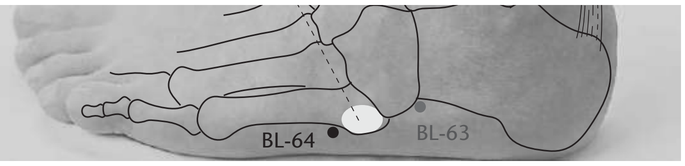

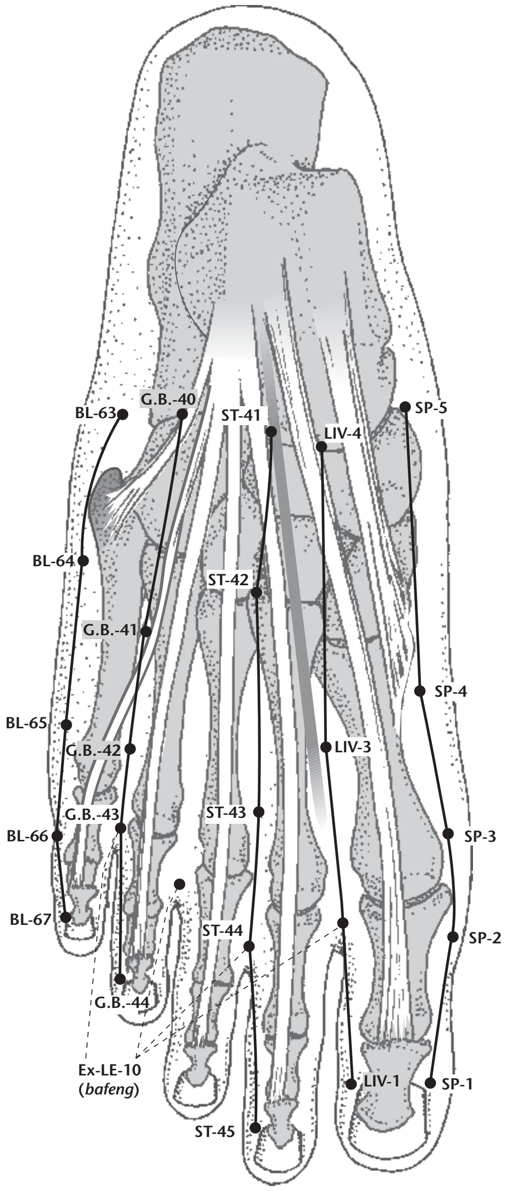
BL-63
G.B.-40
ST-41
SP-5
LIV-4
BL-64
ST-42
G.B.-41
SP-4
BL-65
LIV-3
G.B.-42
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

Shang Mound SHANGQIU **SP-5**

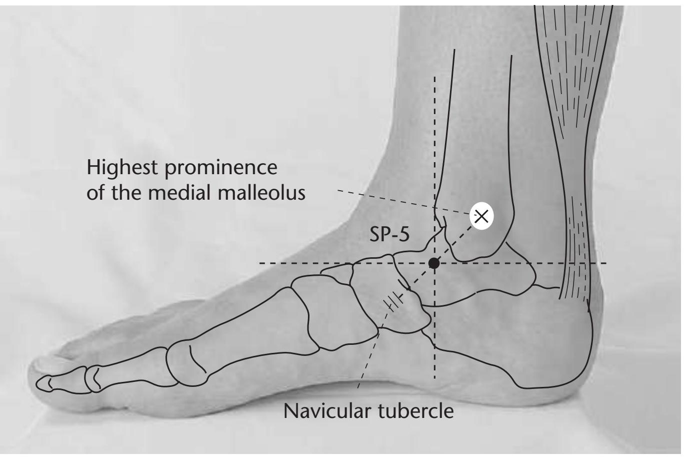
Highest prominence
of the medial malleolus
SP-5
Navicular tubercle

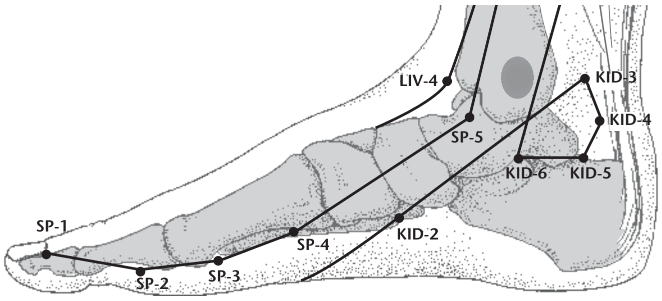
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-2
SP-3

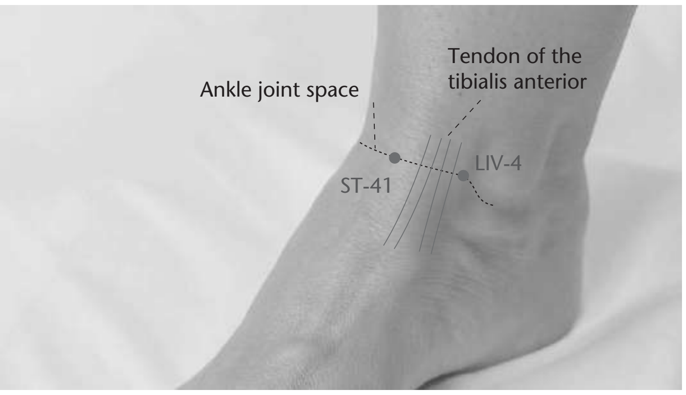
Tendon of the
tibialis anterior
Ankle joint space
LIV-4
ST-41

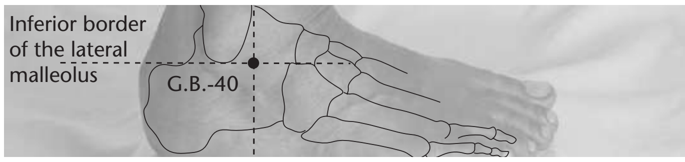
Inferior border
of the lateral
malleolus
G.B.-40

Location

In the depression at the junction of a vertical line drawn along the anterior border and a horizontal line drawn along the lower border of the medial malleolus. **Or**: In the depression halfway between the highest prominence of the medial malleolus and the tubercle of the navicular bone.

How to find

SP-5 is located in a depression anterior and inferior to the medial malleolus, at the junction of a vertical line drawn along the anterior border and a horizontal line drawn along the inferior border of the medial malleolus. **SP-5** is located distal to the tendon of the tibialis anterior muscle.

Located in a comparable position on the lateral aspect of the foot is → **G.B.-40** (in a depression at the junction of a vertical line drawn along the anterior border and a horizontal line drawn along the inferior border of the lateral malleolus).

Needling

0.3–0.6 cun vertically, obliquely or transversely (subcutaneously) deep to the tendons towards → **ST-41**.

Actions/Indications

- Tonifies the Spleen, eliminates Dampness, benefits the tendons and bones
- Moves Qi locally
- Calms the *shen*

Special features

Jing-river point, Metal point, sedation point. Important point for treating Damp *bi*-syndrome.

SP-6 *Three Yin Intersection* SANYINJIAO

Location

3 cun proximal to the highest prominence of the medial malleolus, on the posterior border of the medial crest of the tibia.

How to find

From the highest prominence of the medial malleolus (→ 3.6.2), measure 3 cun (1 handbreadth) in a superior direction, where **SP-6** is located in a depression at the posterior border of the medial crest of the tibia. This point is often sensitive to pressure. Sometimes, this depression is located more anteriorly, and locating **SP-6** should be based on pressure-sensitivity.

Located in a comparable position is → **G.B.-39** (3 cun proximal to the prominence of the lateral malleolus, on the anterior border of the fibula, intersecting point of the three foot Yang channels).

Needling

Vertically or obliquely 1–2 cun. Caution: Contraindicated during pregnancy, particularly with reducing needle techniques; exception: to promote labour.

Actions/Indications

- Tonifies the Spleen and the Stomach, eliminates Dampness
- Nourishes the Blood and Yin, regulates menstruation, promotes labour
- Regulates urination, benefits the genitals, harmonises the Lower Burner
- Calms the *shen*

Special features

Meeting point with the Liver and Kidney channels (group *luo*-connecting point) and the *yin qiao mai*. Very important point for disorders of the Lower Burner (especially for gynaecological disorders, birth and urinary disorders), as well as for many digestive disorders and psychological disorders (caused by Blood and Yin deficiency).

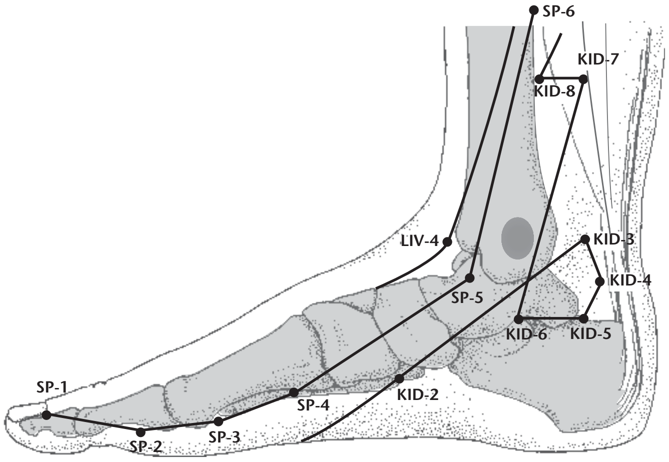
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
SP-4
KID-2
SP-3
SP-2

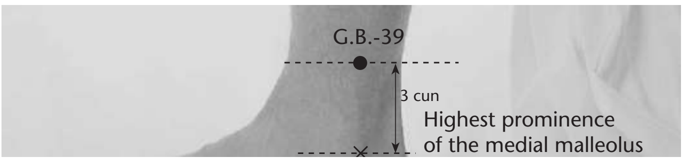
G.B.-39
3 cun
Highest prominence
of the medial malleolus

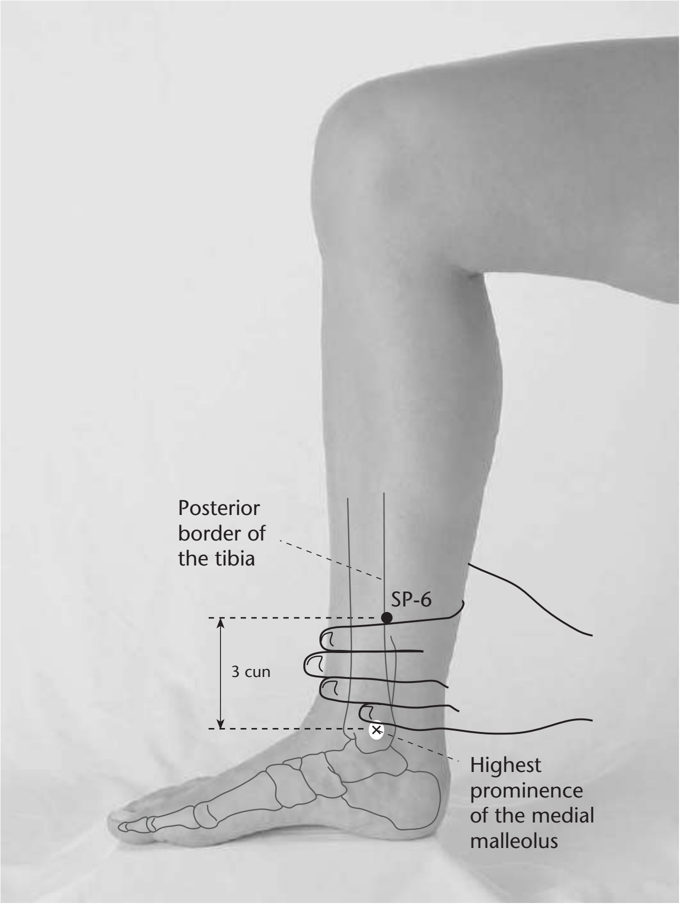
Posterior
border of
the tibia
SP-6
3 cun
Highest
prominence
of the medial
malleolus

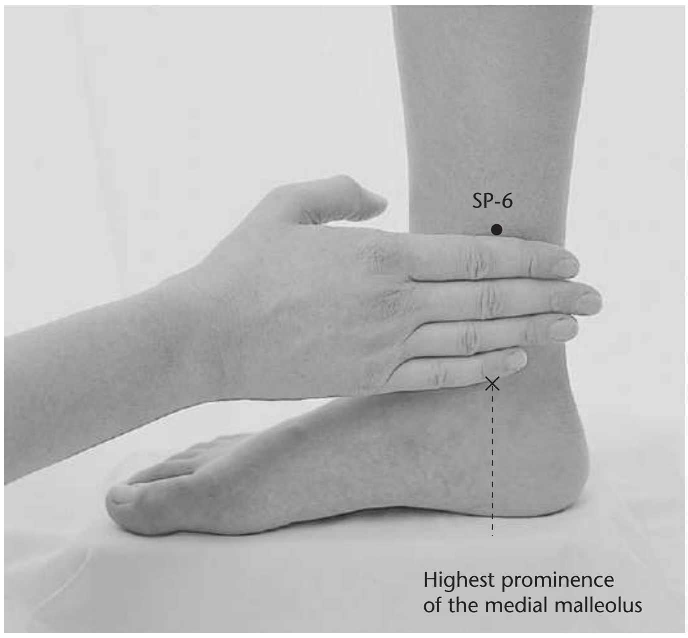
SP-6
Highest prominence
of the medial malleolus

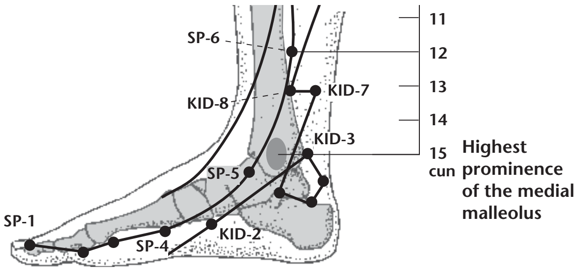
SP-6
KID-8
KID-7
KID-3
SP-5
SP-1
SP-4
KID-2
11
12
13
14
15
cun
Highest
prominence
of the medial
malleolus

Dripping Valley LOUGU SP-7

Location

6 cun proximal to the highest prominence of the medial malleolus, on the posterior border of the medial crest of the tibia.

How to find

From the highest prominence of the medial malleolus, measure 6 cun in a superior direction, where **SP-7** is located at the posterior border of the medial crest of the tibia. **Or**: Spreading hands technique (→ 2.3.3): **SP-7** is located 0.5 cun distal to the midpoint of a line connecting → **SP-9** (distal to the tibial condyle) and the highest prominence of the medial malleolus.
→ **SP-6** is located midway between the tip of the malleolus and **SP-7** (3 cun superior/inferior respectively).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Strengthens the Spleen, drains Dampness, opens the channel
- Promotes urination

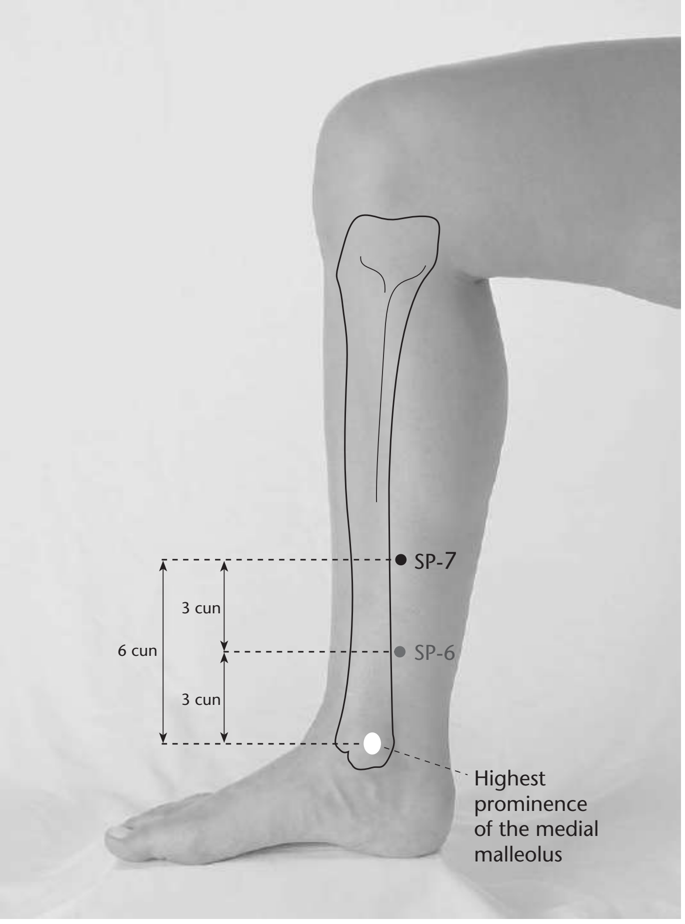
SP-7
3 cun
6 cun
SP-6
3 cun
Highest prominence of the medial malleolus

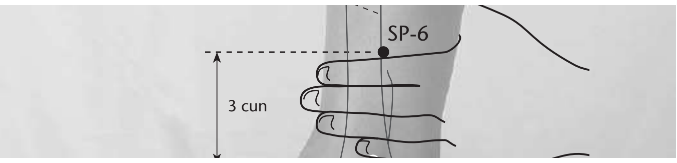

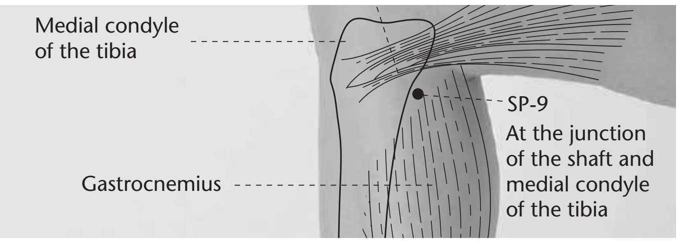
Medial condyle of the tibia
Gastrocnemius
SP-9
At the junction of the shaft and medial condyle of the tibia

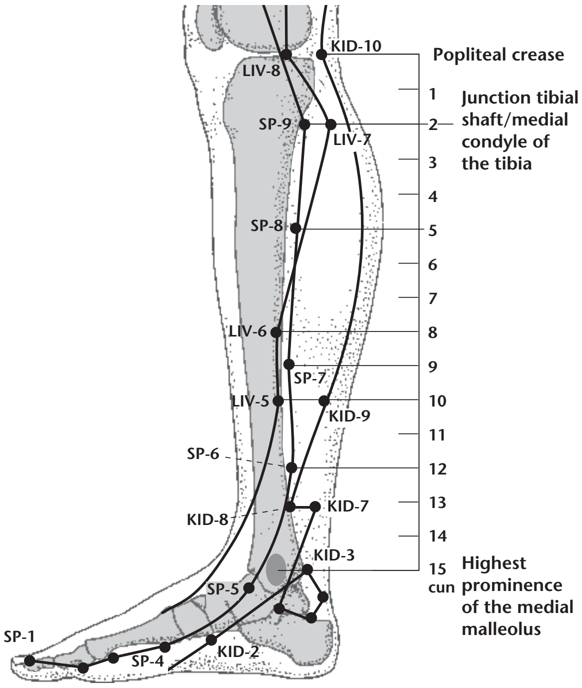
KID-10
Popliteal crease
LIV-8
1
Junction tibial shaft/medial condyle of the tibia
SP-9
2
LIV-7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
10
LIV-5
KID-9
11
SP-6
12
KID-8
KID-7
13
14
KID-3
15 cun
Highest prominence of the medial malleolus
SP-5
SP-1
SP-4
KID-2

SP-8 *Earth Pivot* DIJI

Location

3 cun distal to the junction of the shaft and the medial condyle of the tibia (→ **SP-9**), at the posterior border of the medial crest of the tibia.

How to find

Measure 1 handbreadth directly inferior to the medial condyle of the tibia (→ **SP-9**). **SP-8** is located 3 cun distal to → **SP-9** on the posterior border of the medial crest of the tibia, on a line connecting → **SP-9** and the highest prominence of the medial malleolus.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Regulates the uterus and the Blood
- Regulates the Spleen and eliminates Dampness

Special features

Xi-cleft point. Important point for acute dysmenorrhoea, often used in combination with → **L.I.-4**.

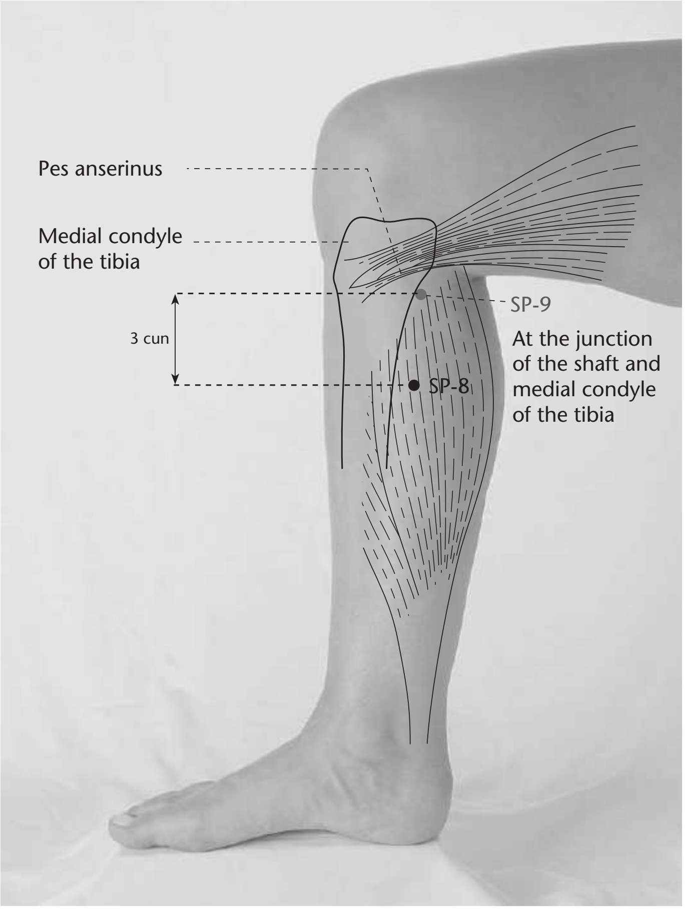
Pes anserinus
Medial condyle of the tibia
3 cun
SP-9
At the junction of the shaft and medial condyle of the tibia
SP-8

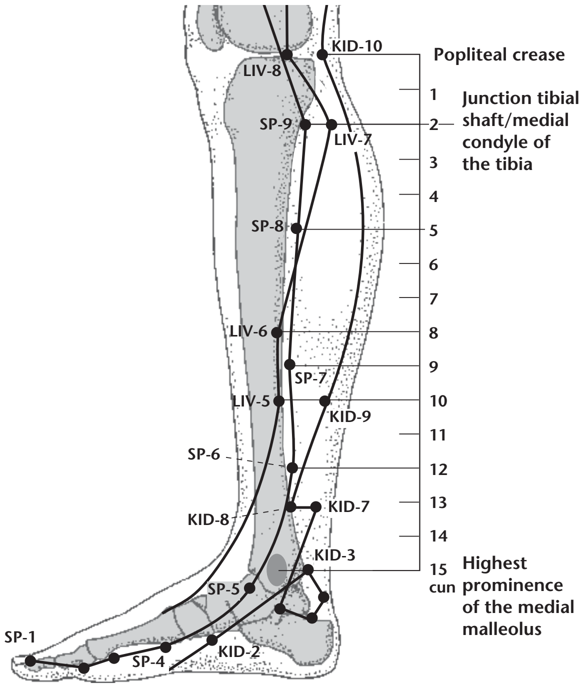
KID-10
Popliteal crease
LIV-8
1
Junction tibial shaft/medial condyle of the tibia
SP-9
2
LIV-7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
LIV-5
10
KID-9
11
SP-6
12
KID-7
13
KID-8
14
KID-3
15 cun
Highest prominence of the medial malleolus
SP-5
SP-1
SP-4
KID-2

Yin Mound Spring YINLINGQUAN SP-9

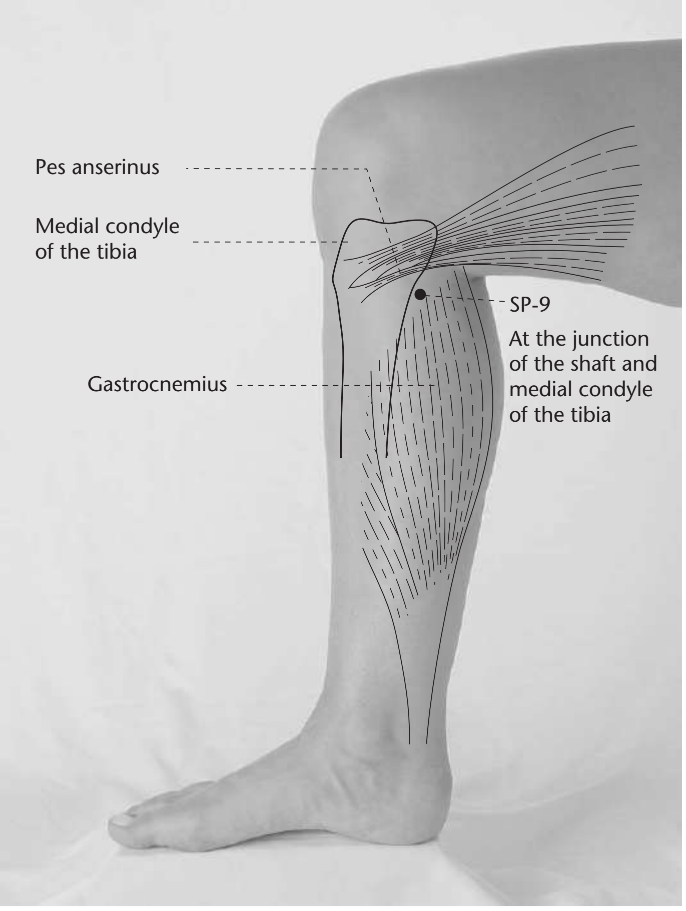
Pes anserinus
Medial condyle of the tibia
Gastrocnemius
SP-9
At the junction of the shaft and medial condyle of the tibia

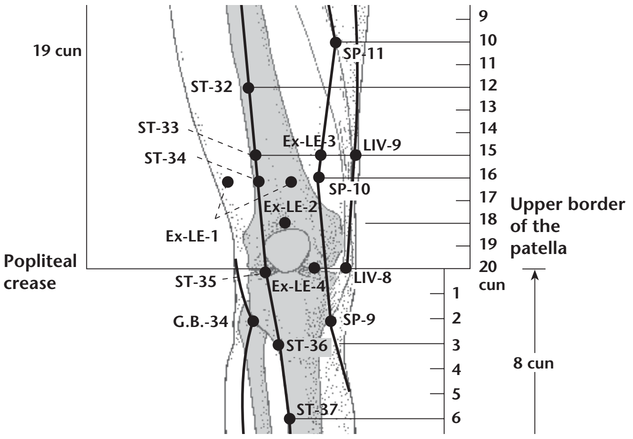
19 cun
Popliteal crease
ST-32
ST-33
ST-34
Ex-LE-1
ST-35
G.B.-34
SP-11
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-4
LIV-8
SP-9
ST-36
ST-37
9
10
11
12
13
14
15
16
17
18
19
20
cun
1
2
3
4
5
6
Upper border of the patella
8 cun

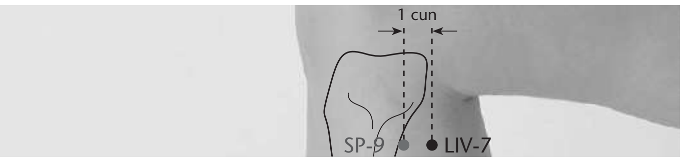
1 cun
SP-9
LIV-7

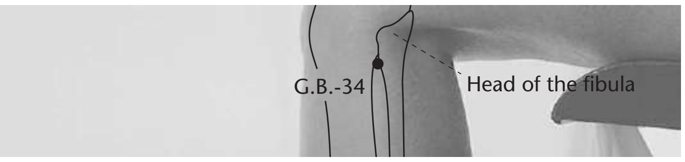
G.B.-34
Head of the fibula

Location

With the knee flexed, this point is located in a depression distal to the medial condyle of the tibia, at the junction of the shaft and the medial condyle.

How to find

Patient's position: Preferably locate and needle **SP-9** with the knee bent and the hip slightly externally rotated, so that the knee slightly lifts from the surface, or rest the knee in an externally rotated position on a knee support. **Location**: Palpate along the posterior border of the medial crest of the tibia towards the head of the tibia to locate **SP-9** in a pressure-sensitive depression at the junction of the shaft and the head of the tibia. This depression is located between the pes anserinus superficialis and the belly of the gastrocnemius muscle (➞ 3.6.1).

Located on the same level but 1 cun posterior to **SP-9** is ➞ **LIV-7**.

Located on the lateral aspect of the lower leg is ➞ **G.B.-34** (anterior and inferior to the head of the fibula).

Needling

1–1.5 cun vertically, along the posterior border of the tibia.

Actions/Indications

- Regulates the Spleen, transforms Dampness, regulates the water passages, benefits the Lower Burner
- As local point for disorders of the knee, especially if there are swellings

Special features

He-sea point, Water point. One of the major points for eliminating Dampness anywhere in the body. Important local point for disorders of the knee. **SP-9** is located approximately on the same level as ➞ **G.B.-34**. If indicated, this point tends to be very tender upon pressure.

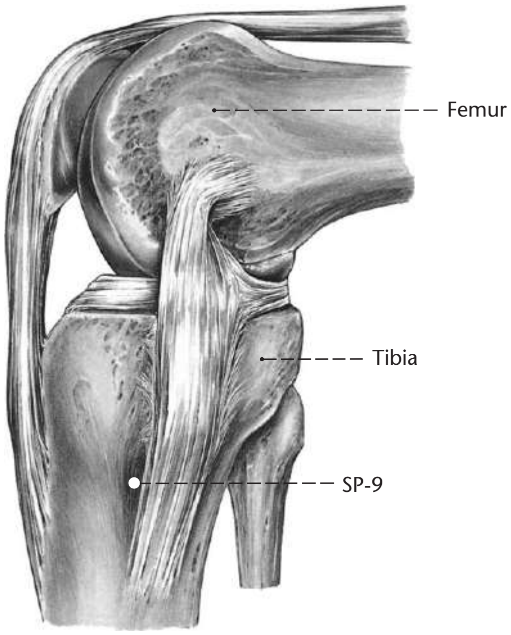
Femur
Tibia
SP-9

SP-10 *Sea of Blood* XUEHAI

Location

With the knee flexed, **SP-10** is located 2 cun proximal and slightly medial to the medial superior border of the patella, in a depression on the vastus medialis muscle.

How to find

Measure 2 cun proximal to the medial superior border of the patella and locate **SP-10** by palpating slightly medially for a soft depression. **Or**: The practitioner places their left hand on the patient's right patella, with the heel of the palm level with the inferior patellar border and the extended fingers pointing in a superior direction. With the thumb and index finger forming an angle of 45°, the tip of the thumb will point to **SP-10**.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Strengthens the Blood, dispels Blood stasis, cools the Blood, stops bleeding, regulates menstruation, benefits the skin
- Local point for disorders of the knee

Special features

Important point for regulating the Blood

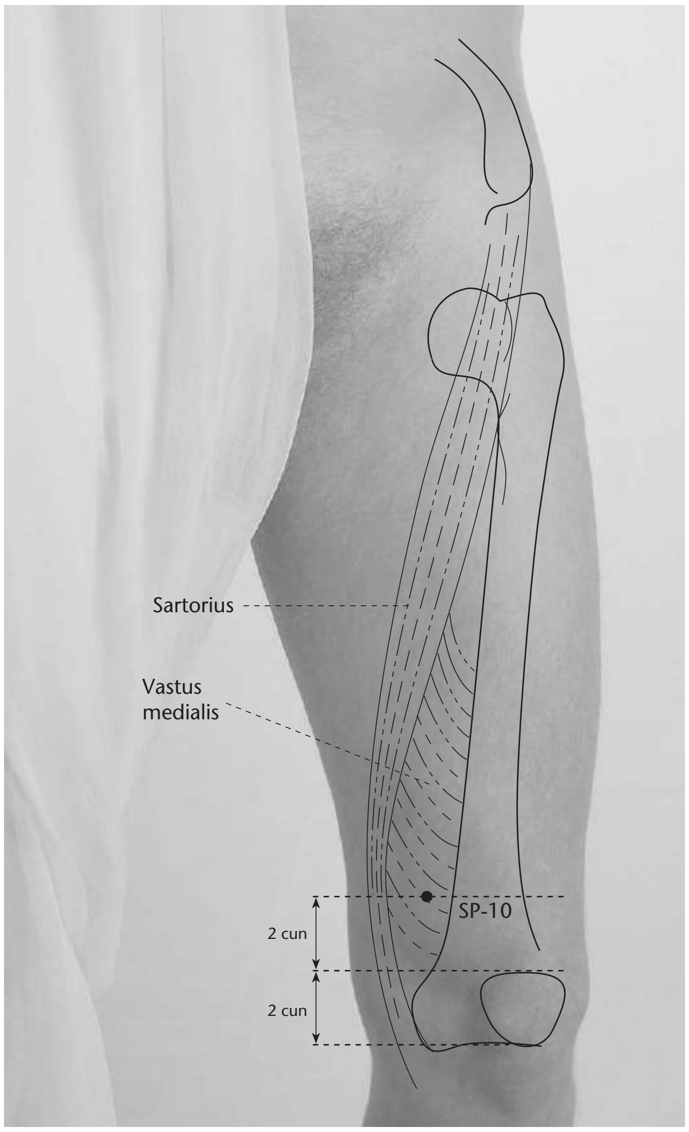

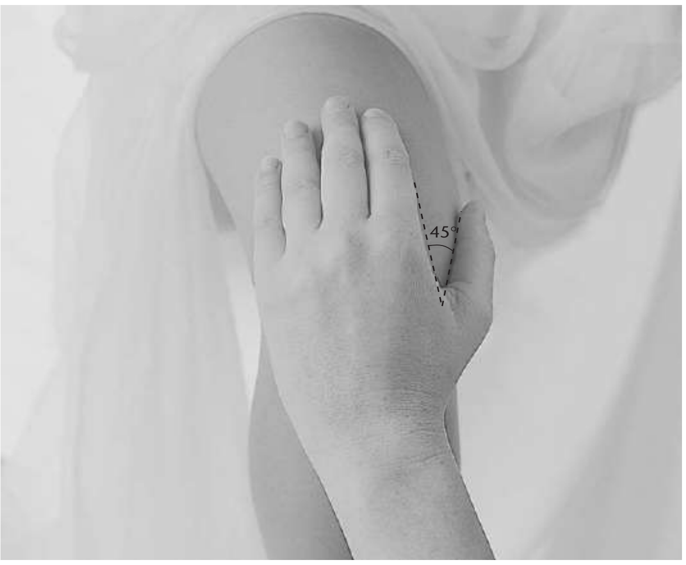

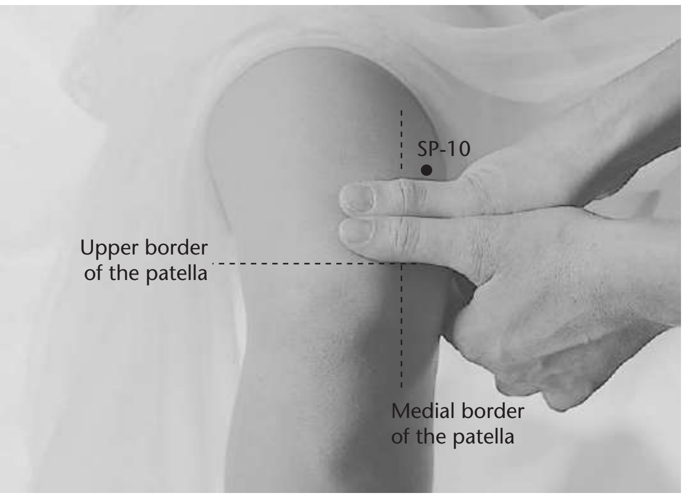

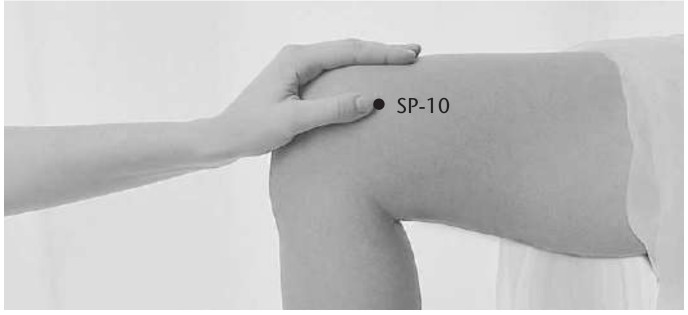

Winnowing Gate **JIMEN** **SP-11**

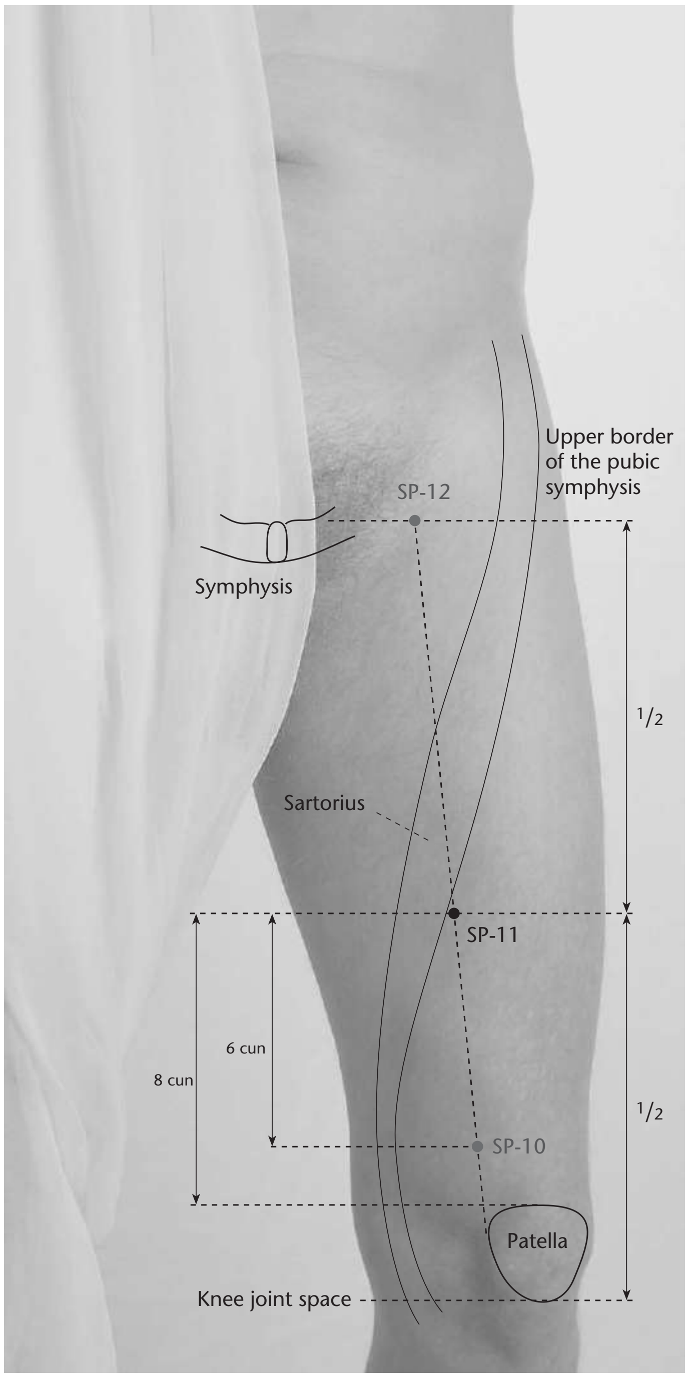
Upper border
of the pubic
symphysis
SP-12
Symphysis
1/2
Sartorius
SP-11
6 cun
8 cun
SP-10
1/2
Patella
Knee joint space

Location

6 cun proximal to → **SP-10** or 8 cun proximal to the medial upper border of the patella, at the midpoint of the femur between the sartorius and vastus lateralis muscles.

How to find

First, locate → **SP-10** by measuring 2 cun proximal and 1 cun medial from the medial upper border of the patella. From there, measure 6 cun vertically in a proximal direction to locate **SP-11** in a depression on the lateral border of the sartorius muscle. **Or**: Spreading hands technique (→ 2.3.3): **SP-11** is located midway between the knee joint space and → **SP-12** (2 cun lateral to the anterior midline, on a level with the upper border of the pubic symphysis).

Needling

Vertically 0.5–1 cun. Caution: Femoral artery/vein.

Actions/Indications

- Regulates the Lower Burner and urination, drains Dampness, clears Heat

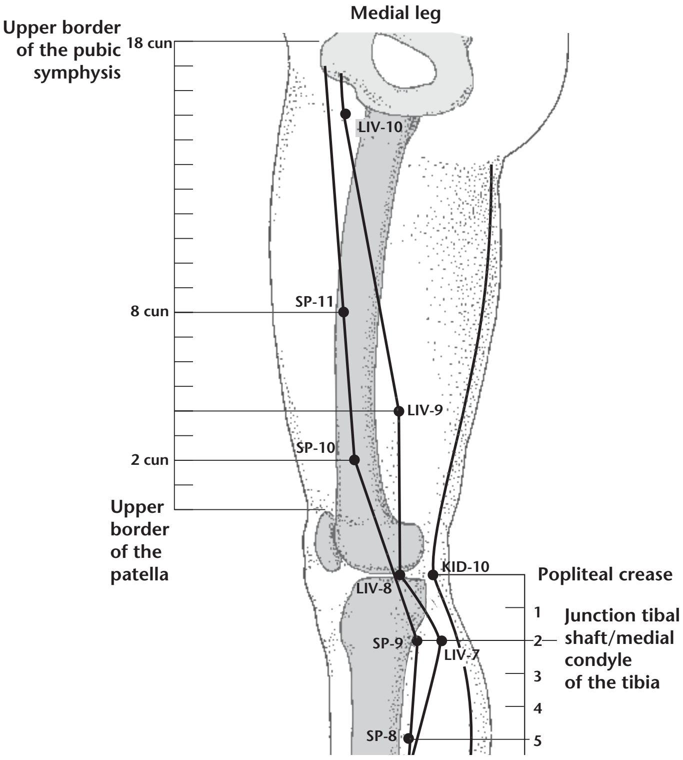
Medial leg
Upper border
of the pubic
symphysis
18 cun
LIV-10
SP-11
8 cun
LIV-9
SP-10
2 cun
Upper
border
of the
patella
KID-10
Popliteal crease
LIV-8
1
Junction tibial
shaft/medial
condyle
of the tibia
SP-9
2
LIV-7
3
4
SP-8
5

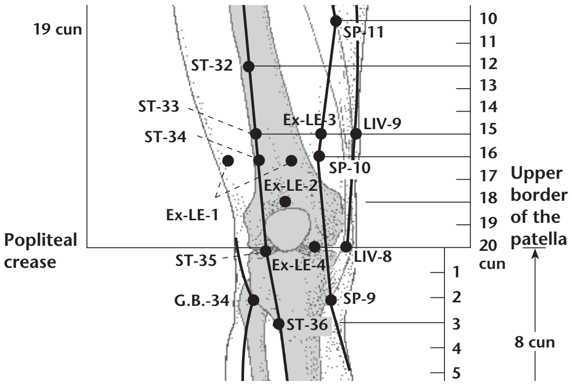
19 cun
SP-11
10
11
ST-32
12
13
ST-33
14
Ex-LE-3
LIV-9
15
ST-34
16
SP-10
17
Upper
border
of the
patella
Ex-LE-2
18
Ex-LE-1
19
Popliteal
crease
20
cun
ST-35
Ex-LE-4
LIV-8
1
G.B.-34
SP-9
2
ST-36
3
8 cun
4
5

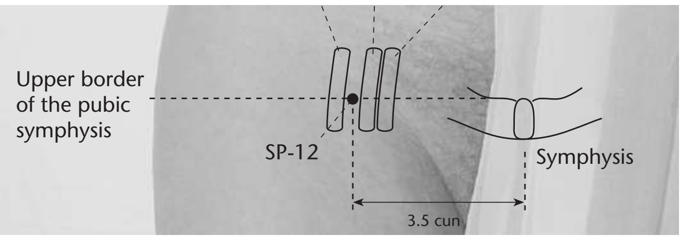
Upper border
of the pubic
symphysis
SP-12
Symphysis
3.5 cun

SP-12 *Rushing Gate* CHONGMEN

Location

3.5 cun lateral to the midline, at the level of the upper border of the pubic symphysis, lateral to the femoral artery.

How to find

From the upper border of the pubic symphysis (→ **Ren-2**), measure 3.5 cun in a lateral direction. There, you can feel the pulsation of the femoral artery. **SP-12** is located in a depression immediately lateral to the artery.

Located on the same level are **Ren-2** (on the midline), → **KID-11** (0.5 cun lateral to the midline) and → **ST-30** (2 cun lateral to the midline). → **LIV-12** is located 2.5 cun lateral to the midline, but 1 cun inferior to the upper border of the pubic symphysis.

Needling

Vertically 0.5–1 cun. Caution: Needling in a medial direction may puncture the artery; needling in a lateral direction may injure the femoral nerve.

Actions/Indications

- Regulates Qi and Blood in the Lower Burner
- Descends rising fetal Qi
- Clears Heat and drains Dampness

Special features

Meeting point with the LIV and T.B.* channels and the *yin wei mai*

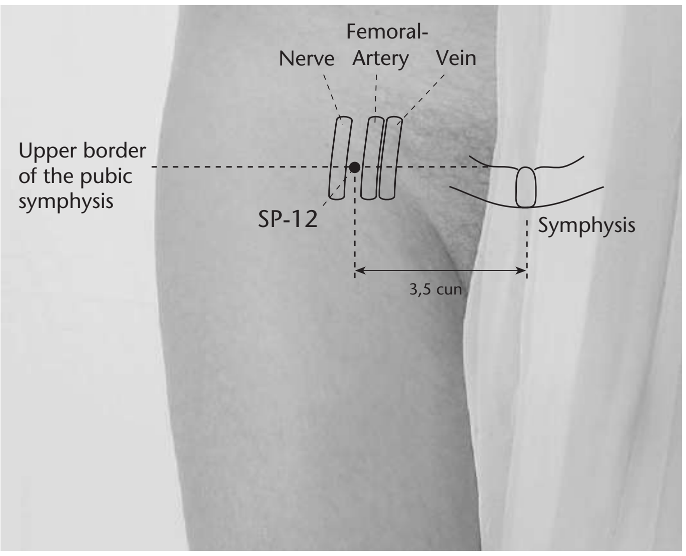
Femoral-
Nerve
Artery
Vein
Upper border of the pubic symphysis
SP-12
Symphysis
3,5 cun

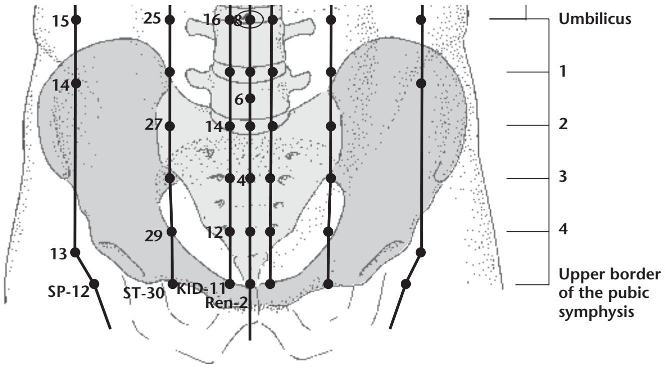
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

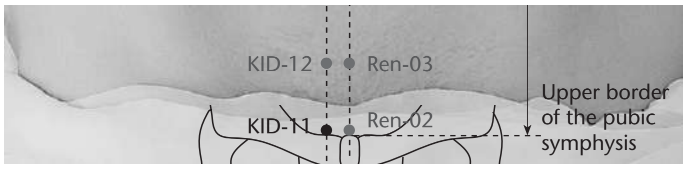
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

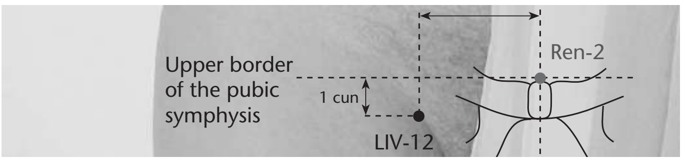
Ren-2
Upper border of the pubic symphysis
1 cun
LIV-12

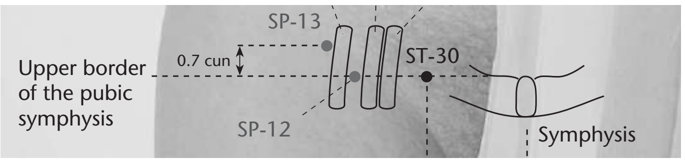
SP-13
0.7 cun
ST-30
Upper border of the pubic symphysis
SP-12
Symphysis

---

* According to some authors

Abode of the Fu FUSHE SP-13

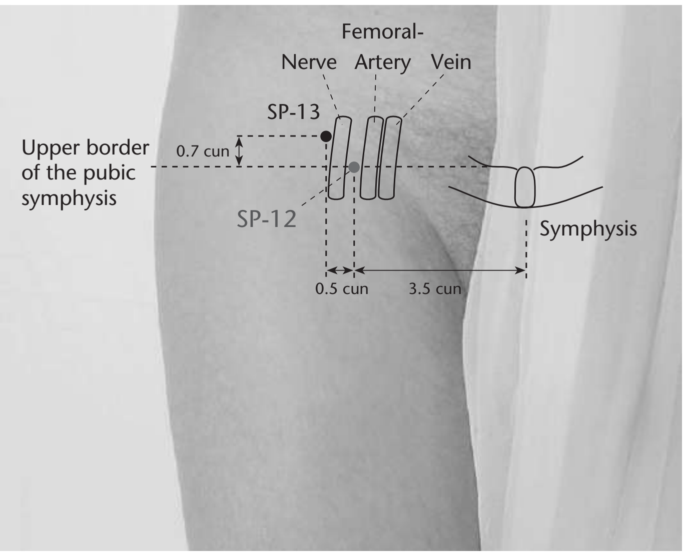
Femoral-
Nerve
Artery
Vein
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

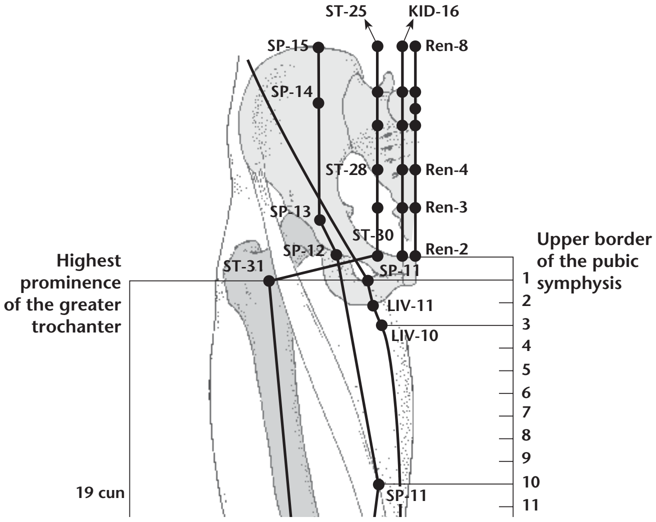
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
SP-13
ST-30
SP-12
Ren-2
Highest prominence of the greater trochanter
ST-31
SP-11
Upper border of the pubic symphysis
LIV-11
LIV-10
19 cun
SP-11
1
2
3
4
5
6
7
8
9
10
11

Location

4 cun lateral to the anterior midline and 0.7 cun superior to the upper border of the pubic symphysis.

How to find

First, measure 4 cun in a lateral direction from the upper border of the pubic symphysis and from there 0.7 cun in a superior direction to locate **SP-13** in a depression in the inguinal groove. **SP-13** is located 0.5 cun lateral and 0.7 cun superior to → **SP-12** (→ 3.5 cun lateral to the upper border of the pubic symphysis).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum.

Actions/Indications

- Disperses accumulations
- Regulates the Qi

Special features

Meeting point with the *yin wei mai* and the Liver channel

SP-14 *Abdomen Knot* FUJIE

Location

4 cun lateral to the anterior midline, on the mamillary line, 3 cun superior to ➞ **SP-13** or 1.3 cun inferior to ➞ **SP-15**.

How to find

First, measure 4 cun from the umbilicus in a lateral direction in order to locate the mamillary line. This is the point ➞ **SP-15**. From there, measure 1.3 cun in an inferior direction to locate **SP-14**.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Moves Qi and Blood, warms the middle, dispels Cold
- Regulates the Lower Burner and descends counterflow Qi

Special features

Sensitive to pressure with appendicitis

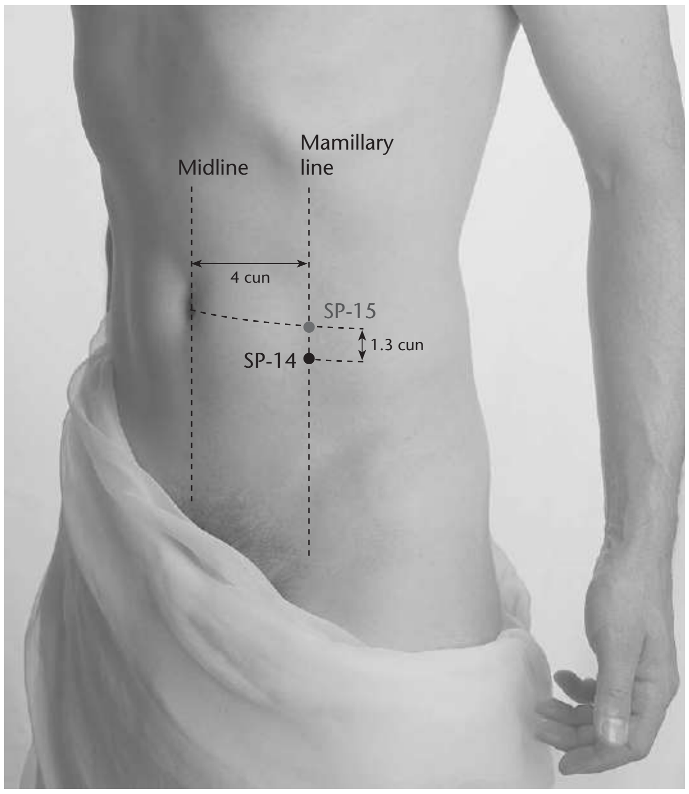
Midline
Mamillary line
4 cun
SP-15
1.3 cun
SP-14

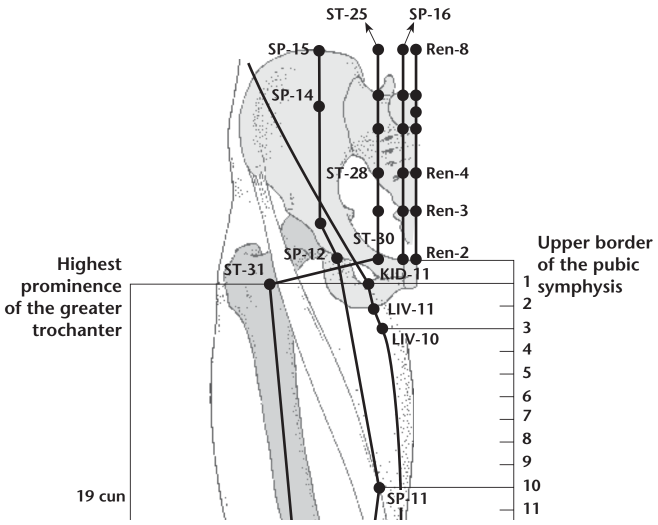
ST-25
SP-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
ST-30
Ren-2
SP-12
Highest prominence of the greater trochanter
ST-31
KID-11
Upper border of the pubic symphysis
LIV-11
LIV-10
19 cun
SP-11
1
2
3
4
5
6
7
8
9
10
11

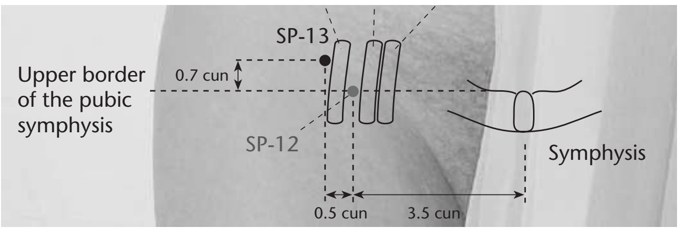
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

Great Horizontal DAHENG SP-15

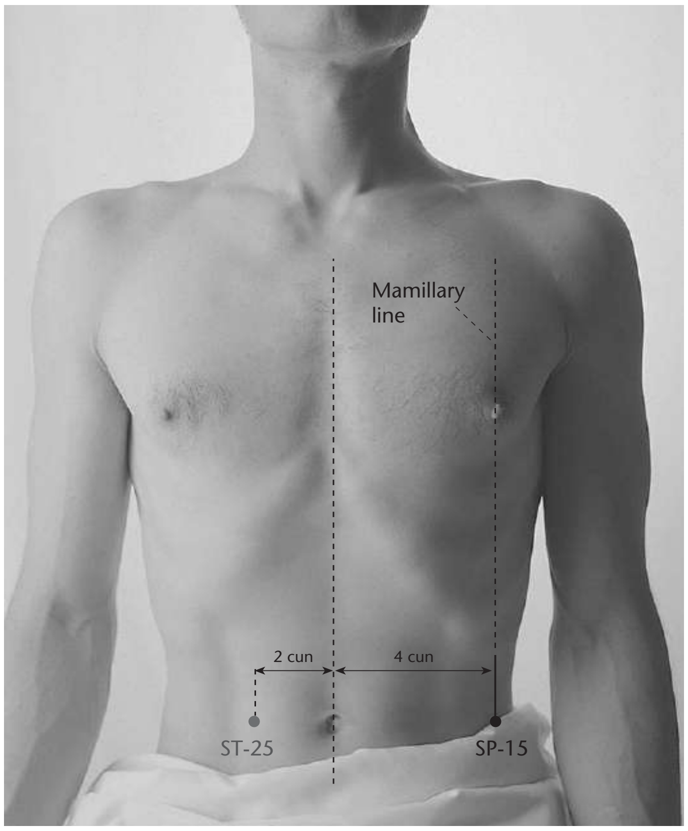
Mamillary line
2 cun
4 cun
ST-25
SP-15

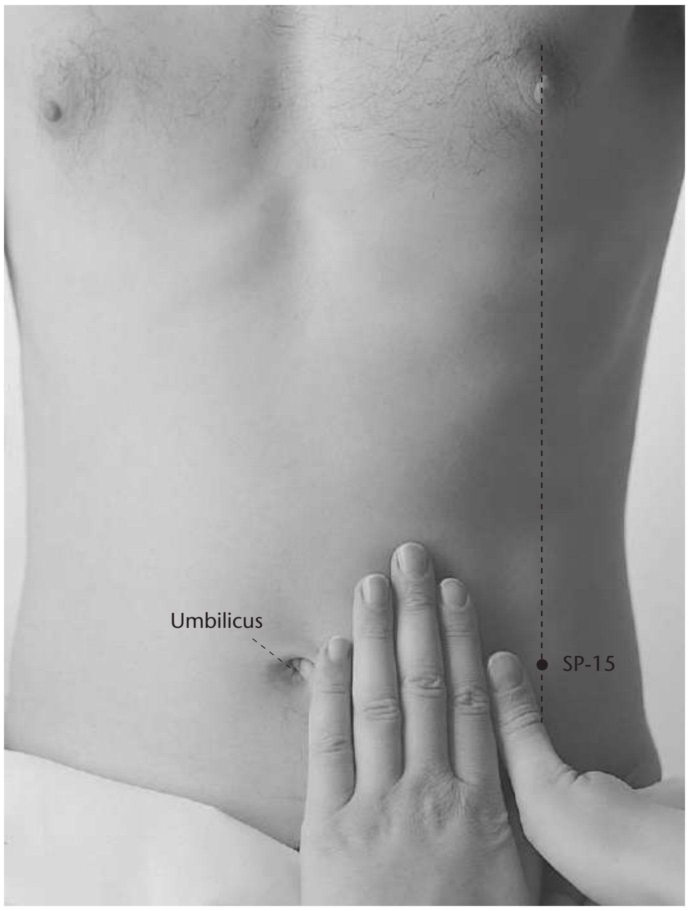
Umbilicus
SP-15

Location

4 cun lateral to the centre of the umbilicus, on the mamillary line.

How to find

Locate **SP-15** 4 cun lateral to the umbilicus, using the mamillary line for reference, especially in male patients.
Located on the same level are → **Ren-8/KID-16/ST-25/G.B.-26** (umbilicus/0.5 cun lateral to the umbilicus/2 cun lateral to the umbilicus/below the free end of the 11th rib at the level of the umbilicus).

Needling

Vertically 1–1.5 cun. Caution: Peritoneum, pregnancy. Deep needling may reach the aponeuroses of the transverse abdominal musculature.

Actions/Indications

- Regulates and moves Qi in the Intestines (promotes bowel movement by stimulating peristalsis)

Special features

Meeting point with the *yin wei mai*. Important point for regulating the Large Intestine Qi.

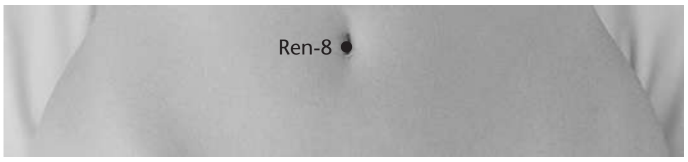

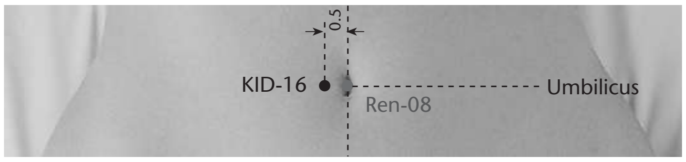
0.5
KID-16
Umbilicus
Ren-08

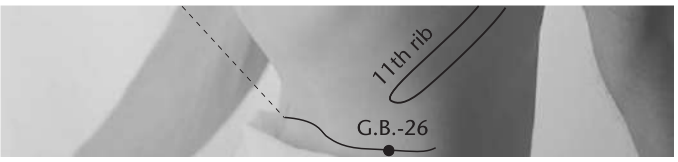
11th rib
G.B.-26

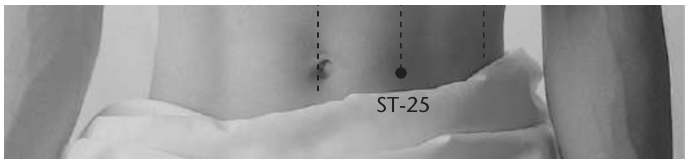

SP-16 *Abdomen Sorrow* FUAI

Location

3 cun above the centre of the umbilicus and 4 cun lateral to the anterior midline, on the mamillary line.

How to find

The distance between the sternocostal angle and the centre of the umbilicus can be divided into 8 proportional cun, which can differ considerably from the patient's finger cun; therefore, only proportional cun measurements should be applied (elastic tape → 2.3.1). First, measure 3 cun from the umbilicus in a superior direction, then 4 cun in a lateral direction, where **SP-16** is located on the mamillary line. Depending on the shape of the thorax, this point will either be located on the abdomen or it will project onto the lower border of the ribcage.

Located on the same level are → **Ren-11** (anterior midline), → **KID-18** (0.5 cun lateral to the anterior midline) and → **ST-22** (2 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy. Depending on the shape of the thorax, the correct location of this point might be on the costal cartilage. In this case, either transverse (subcutaneous) needling or locate the point more medially (or choose an alternative point).

Actions/Indications

- Regulates Qi in the Intestines
- Drains Dampness

Special features

Meeting point with the *yin wei mai*

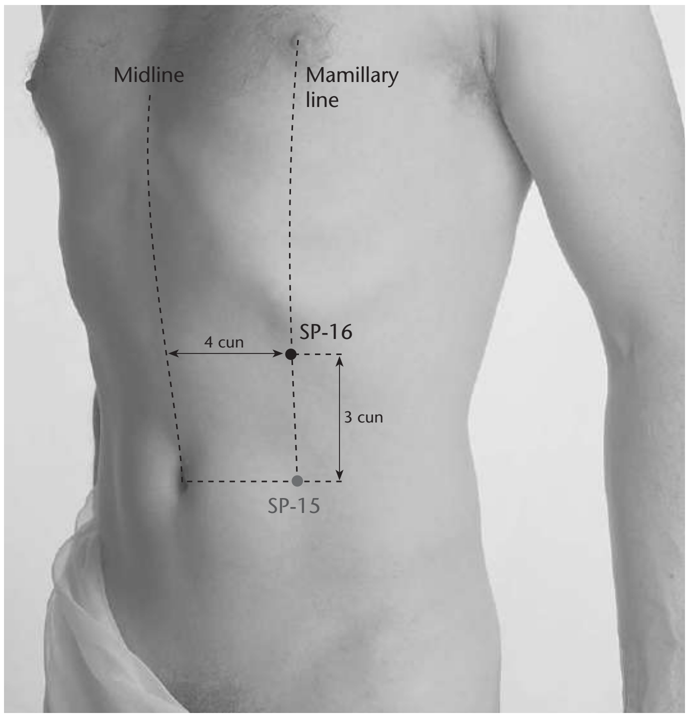
Midline
Mamillary line
SP-16
4 cun
3 cun
SP-15

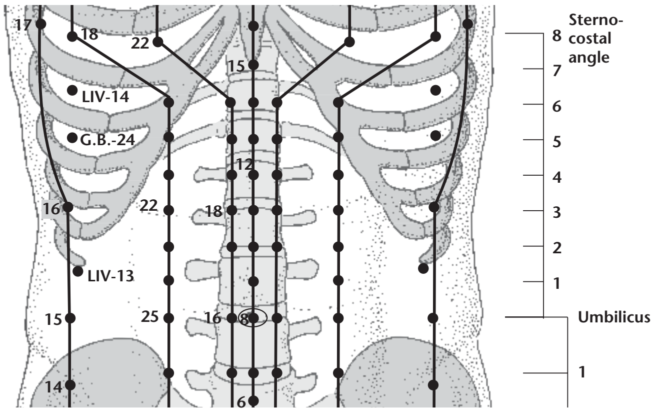
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
6
Sterno-costal angle
8
7
6
5
4
3
2
1
Umbilicus
1

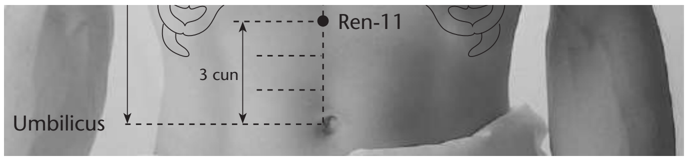
Ren-11
3 cun
Umbilicus

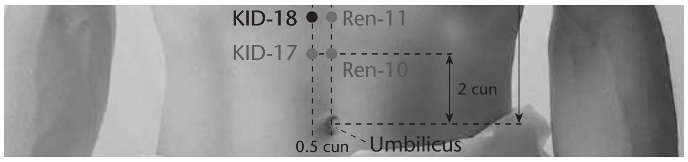
KID-18
Ren-11
KID-17
Ren-10
2 cun
0.5 cun
Umbilicus

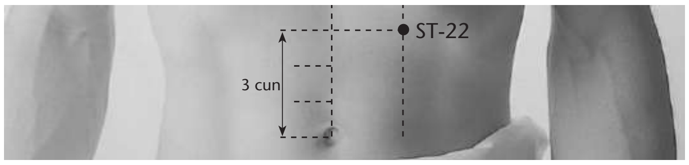

***Food Cavity* SHIDOU SP-17**

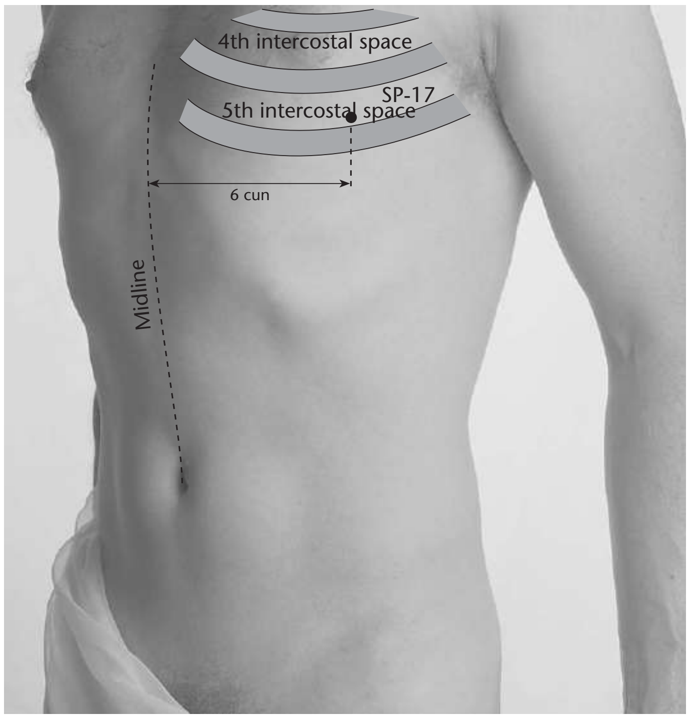
4th intercostal space
SP-17
5th intercostal space
6 cun
Midline

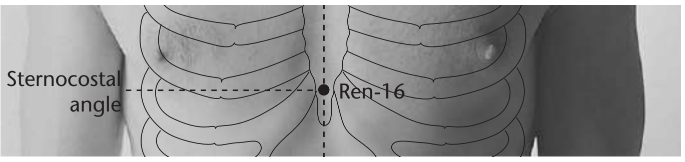
Sternocostal angle
Ren-16

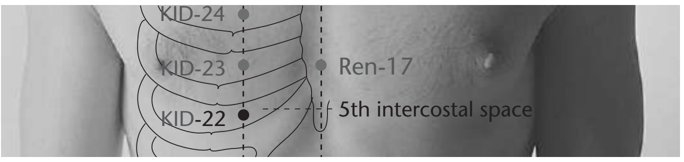
KID-24
KID-23
KID-22
Ren-17
5th intercostal space

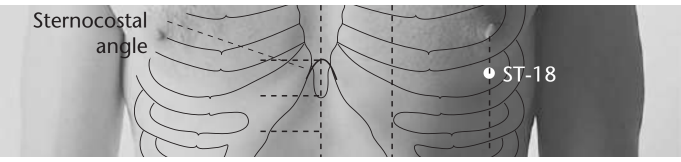
Sternocostal angle
ST-18

Location

In the 5th intercostal space, 6 cun lateral to the anterior midline (in male patients, 2 cun lateral to the nipple).

How to find

In men, the nipple is generally located on the level of the 4th intercostal space, while in women, when in a supine position, it is not always at this level. Locate **SP-17** in the 5th intercostal space, 2 cun lateral to the mamillary line (note that the intercostal space laterally curves upward). Both the clavicle and the manubriosternal synchondrosis are good reference points for locating the 5th intercostal space. **SP-17** can be located by following the course of the intercostal space in a lateral direction and measuring 2 cun lateral to the mamillary line.

Also located on the level of the 5th intercostal space are → **Ren-16** (anterior midline), → **KID-22** (2 cun lateral to the anterior midline) and → **ST-18** (4 cun lateral to the anterior midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Qi
- Harmonises the Stomach and eliminates stagnation
- Drains Dampness

SP-18 *Heavenly Stream* TIANXI

Location

In the 4th intercostal space, 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line.

How to find

In men, the nipple is generally located on the level of the 4th intercostal space, while in women, when in a supine position, it is not always at this level. Follow the course of the 4th intercostal space in a lateral direction and locate **SP-18** 2 cun lateral to the nipple (note that the intercostal space laterally curves upward). In women, either the clavicle or the manubriosternal synchondrosis are good reference points for locating the 4th intercostal space (➞ 3.5). **SP-18** can be located by following the course of the intercostal space in a lateral direction while measuring 6 cun laterally from the anterior midline or 2 cun laterally from the nipple.

Also located on the level of the 4th intercostal space are ➞ **Ren-17** (anterior midline), ➞ **KID-23** (2 cun lateral to the anterior midline), ➞ **ST-17** (nipple), ➞ **P-1** (1 cun lateral to the nipple) and ➞ **G.B.-23** (1 cun anterior to **G.B.-22**).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications:

- Unbinds the chest
- Descends the Qi
- Benefits the breasts

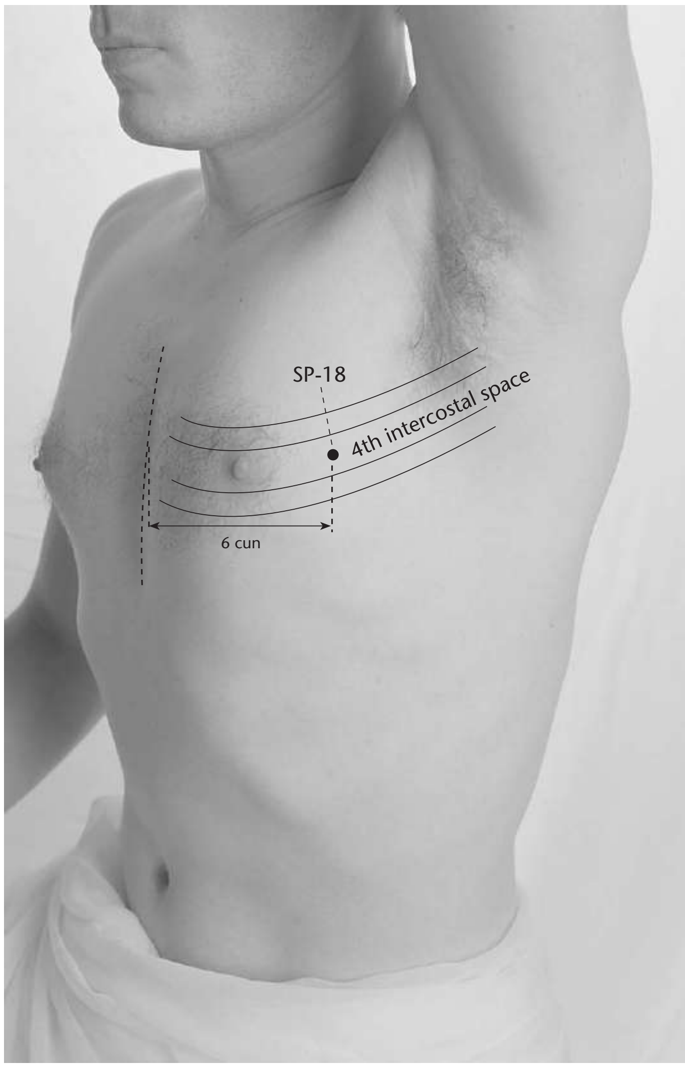
SP-18
4th intercostal space
6 cun

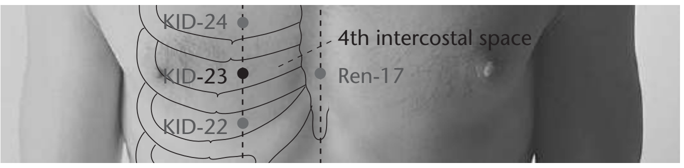
KID-24
KID-23
KID-22
4th intercostal space
Ren-17

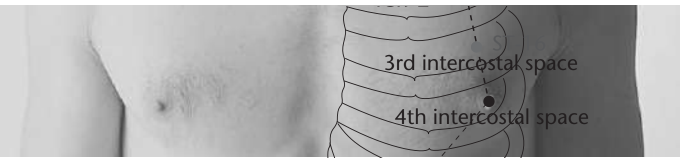
3rd intercostal space
4th intercostal space

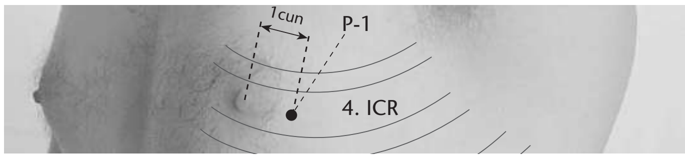
1cun
P-1
4. ICR

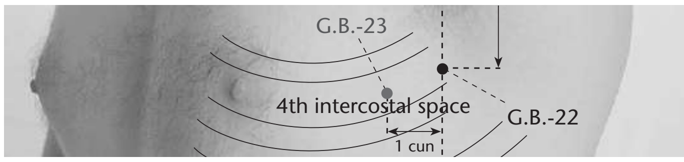
G.B.-23
4th intercostal space
G.B.-22
1 cun

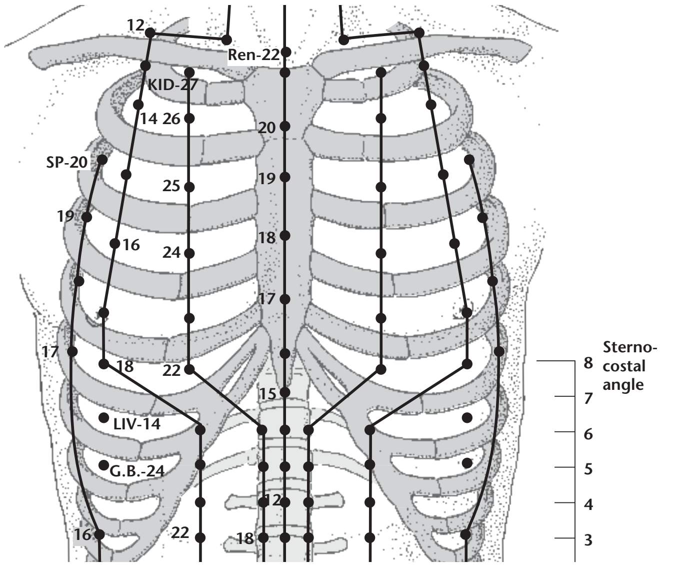
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
Sterno-costal angle
8
7
6
5
4
3

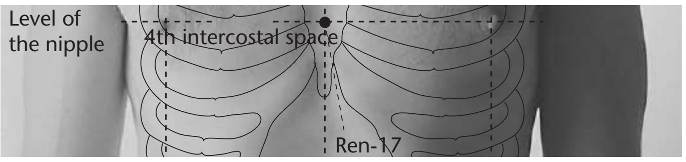
Level of the nipple
4th intercostal space
Ren-17

Chest Village XIONGXIANG SP-19

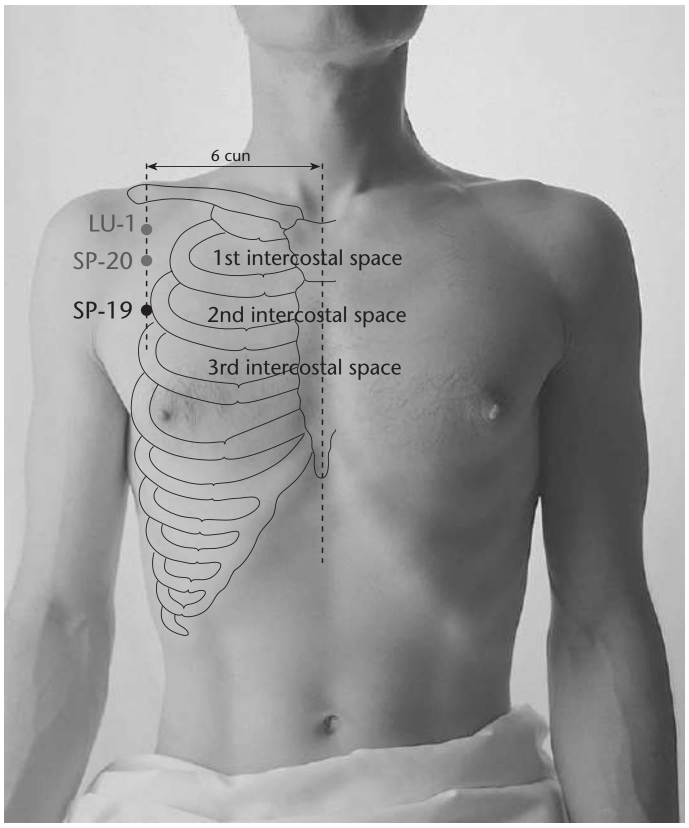
6 cun
LU-1
SP-20
SP-19
1st intercostal space
2nd intercostal space
3rd intercostal space

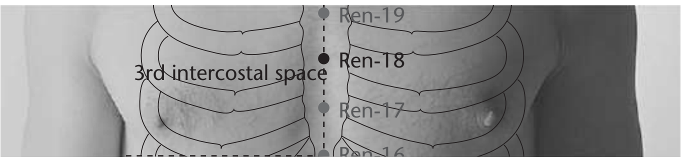
Ren-19
Ren-18
3rd intercostal space
Ren-17
Ren-16

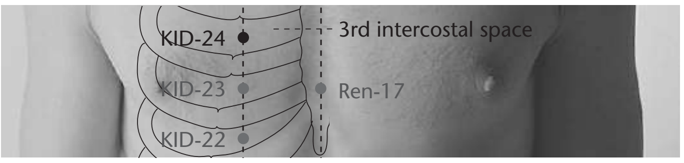
KID-24
3rd intercostal space
KID-23
Ren-17
KID-22

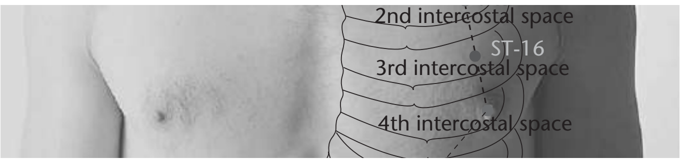
2nd intercostal space
ST-16
3rd intercostal space
4th intercostal space

Location

In the 3rd intercostal space, 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line.

How to find

Locate the 3rd intercostal space by counting parasternally from the clavicle or from the manubriosternal synchondrosis (➞ 3.5). Follow the course of the intercostal space in a lateral direction and locate **SP-19** 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line (note: the intercostal space laterally curves upward).

Located on the same level (note: course of the intercostal space) are ➞ **Ren-18** (anterior midline), ➞ **KID-24** (2 cun lateral to the midline) and ➞ **ST-16** (4 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space, or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Disperses and descends the Lung Qi
- Unbinds the chest

SP-20 *Encircling Glory* ZHOURONG

Location

In the 2nd intercostal space, 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line.

How to find

Locate the 2nd intercostal space by counting parasternally from the clavicle or from the manubriosternal synchondrosis (→ 3.5). Follow the course of the intercostal space in a lateral direction and locate **SP-20** 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line (note: the intercostal space laterally curves upward). **SP-20** is located 1 cun inferior to → **LU-1**. Located on the same level (note: course of the intercostal space) are → **Ren-19** (anterior midline), → **KID-25** (2 cun lateral to the midline) and → **ST-15** (4 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Unbinds the chest
- Regulates the Lung Qi

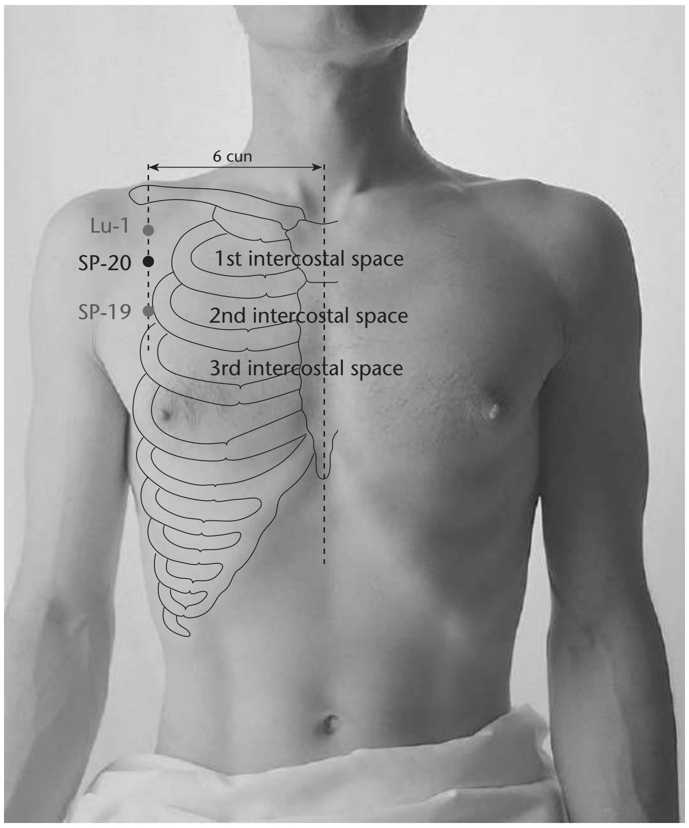
6 cun
Lu-1
SP-20
SP-19
1st intercostal space
2nd intercostal space
3rd intercostal space

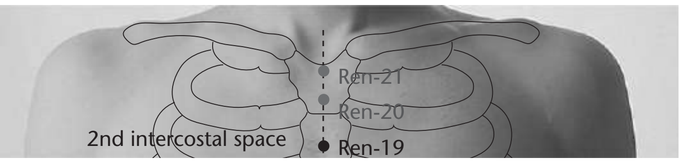
Ren-21
Ren-20
2nd intercostal space
Ren-19

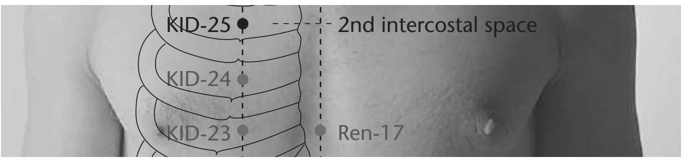
KID-25
2nd intercostal space
KID-24
KID-23
Ren-17

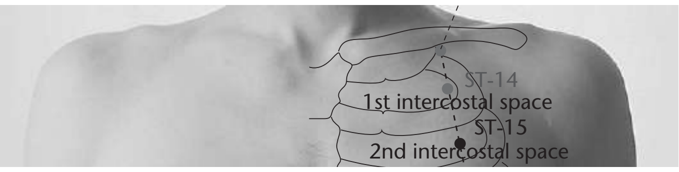
ST-14
1st intercostal space
ST-15
2nd intercostal space

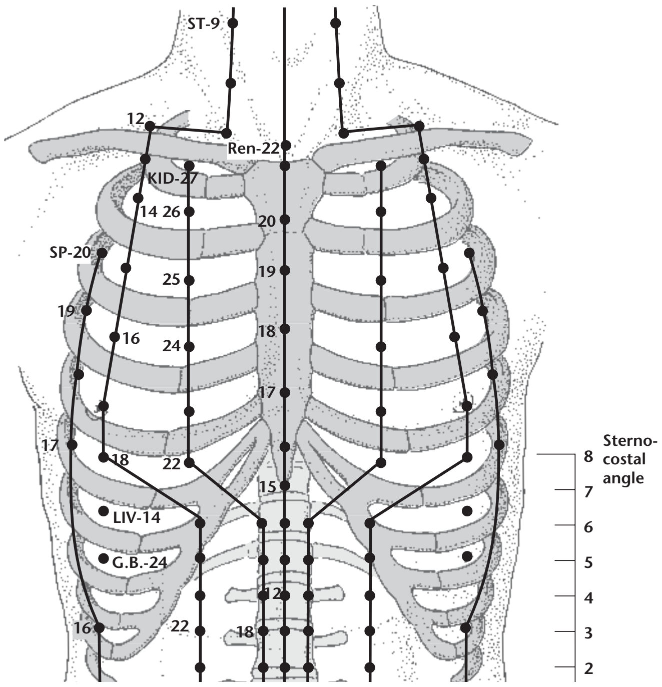
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
8
7
6
5
4
3
2
Sterno-costal angle

Great Wrapping DABAO **SP-21**

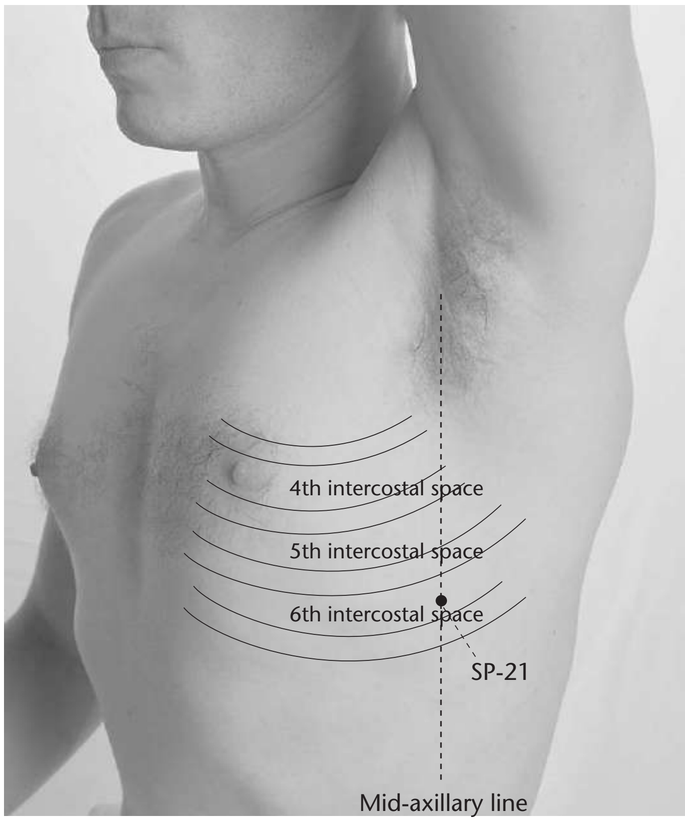
4th intercostal space
5th intercostal space
6th intercostal space
SP-21
Mid-axillary line

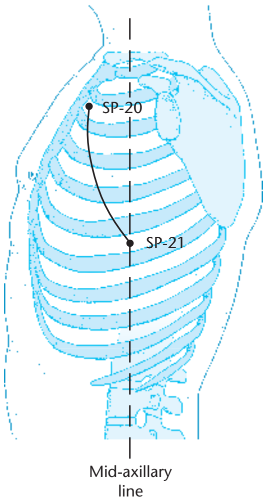
SP-20
SP-21
Mid-axillary line

Location

On the midaxillary line, in the 6th intercostal space. Note: According to some texts, this point is located in the 7th intercostal space.

How to find

Locate the 6th intercostal space by counting parasternally from the clavicle or from the manubriosternal synchondrosis (→ 3.5). Follow the course of the intercostal space in a lateral direction and locate **SP-21** on the axillary midline (note: the intercostal space laterally curves upward). **SP-21** is often sensitive to pressure. Or: Locate **SP-21** on the midaxillary line, midway between the axilla and the end of the 11th rib.

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates Qi and Blood, governs and wraps all the *luo*-connecting vessels of the body
- Unbinds the chest

Special features

Great *luo*-connecting point of the spleen (→ 8.1.2), exit point.

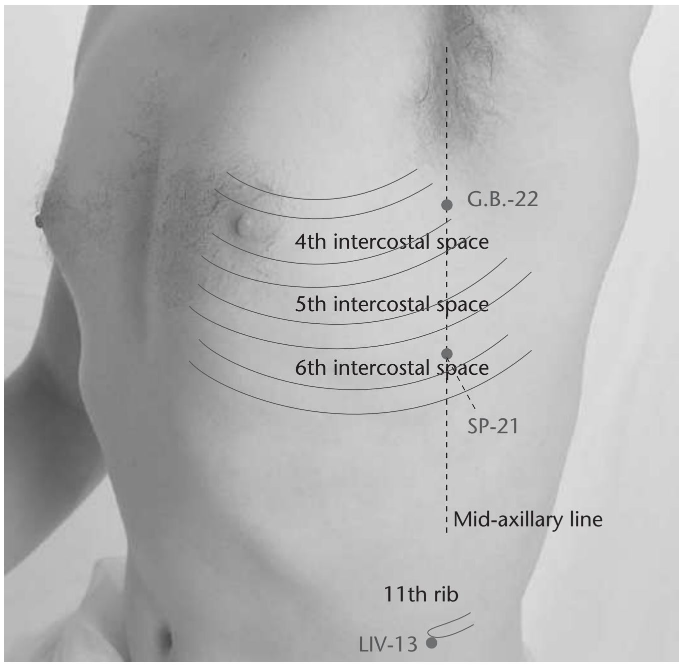
G.B.-22
4th intercostal space
5th intercostal space
6th intercostal space
SP-21
Mid-axillary line
11th rib
LIV-13

4.5 The Heart Channel System – Hand *shaoyin* (*shou shao yin jing luo*)

4.5.1 The Heart Primary Channel (*shou shao yin jing*)

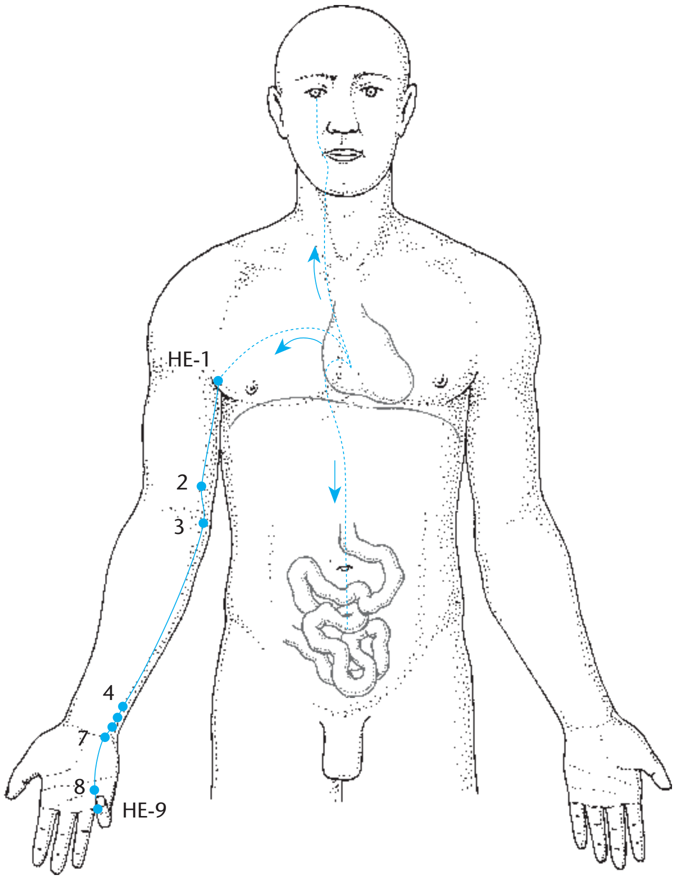
HE-1
2
3
4
7
8
HE-9

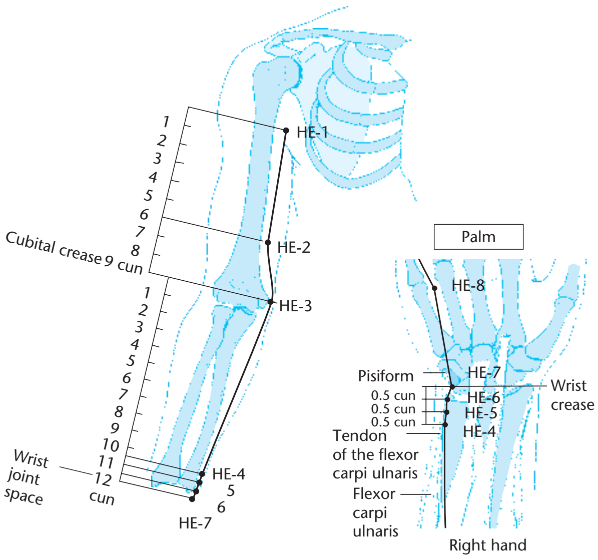
1
2
3
4
5
6
7
8
Cubital crease 9 cun
1
2
3
4
5
6
7
8
9
10
11
12
Wrist joint space
12 cun
HE-1
HE-2
HE-3
HE-4
5
6
HE-7
Palm
HE-8
Pisiform
0.5 cun
0.5 cun
0.5 cun
Tendon of the flexor carpi ulnaris
Flexor carpi ulnaris
HE-7
HE-6
HE-5
HE-4
Wrist crease
Right hand

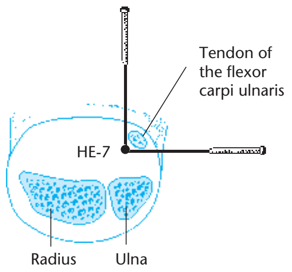
Tendon of the flexor carpi ulnaris
HE-7
Radius
Ulna

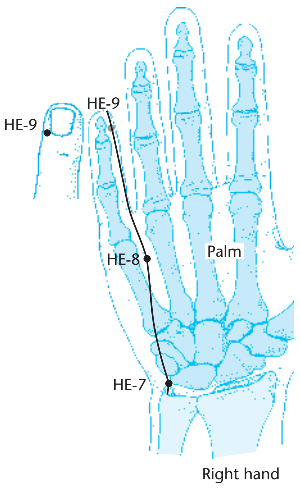
HE-9
HE-9
HE-8
Palm
HE-7
Right hand

Pathway

The **internal** pathway of the Heart primary channel originates in its related *zang*-Organ, the Heart (*xin*). An **internal branch** of the Spleen primary channel reaches the Heart and connects with the Heart primary channel (deep Yin–Yin connection ➞ 1.2).

The **internal pathway** of the Heart primary channel splits into three branches:

- ➡ **the first branch** penetrates the diaphragm and connects with the internally–externally related *fu*-Organ, the Small Intestine (*xiaochang*)
- ➡ **the second branch** ascends along the oesophagus and the neck region, disperses in the tissues around the eyes and, according to some authors, also spreads inside the brain.
- ➡ **the main branch** runs from the Heart to the Lung (*fei*) and descends again to emerge in the centre of the axilla at ➞ **HE-1** (*jiquan*) (beginning of the external pathway).

From here the **external pathway** runs along the anteromedial aspect of the upper arm and forearm, traverses the palm between the fourth and fifth metacarpal bones, continues along the radial aspect of the little finger and ends at the radial corner of its nail at **HE-9** (*shaoyang*).

From **HE-9** (*shaoyang*) a branch crosses to the ulnar border of the little finger's nail to ➞ **S.I.-1** (*shaoze*) (hand Yin–Yang connection of the second great circuit).

Clinical importance (➞ 1.2)

Exterior (*biao*) signs and symptoms: General fever, headaches, pain in the thorax and back, thirst with a desire to drink, Heat sensation in the palms, Cold invasion in the extremities, pain in the shoulder and the medial aspect of the upper arm

Interior (*li*) or *zangfu*-Organ signs and symptoms: Heart pain, pain and fullness in the thorax and lateral costal region, hypochondriac pain, irritability, shortness of breath, restlessness, dizziness, mental confusion

Connections of the Heart primary channel (→ 1.2)

Connections with other channels

Small Intestine primary channel (*shou tai yang jing*)

Connection: Hand Yin–Yang connection of the second great circuit
Location: H-9 → S.I.-1 (on the hand).
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Kidney primary channel (zu shao yin jing)

Connection: Paired according to the six channel theory (hand–foot pairing): *shao yin* (Yin axes of the second great circuit)
Location: KID → HE (on the thorax). An internal branch of the Kidney primary channel travels to the Liver (*gan*), penetrates the diaphragm and spreads in the Lung (*fei*). From the Lung a branch runs to Heart (*xin*) and there connects with the Heart primary channel.
Circulation: Non-circadian (**not** according to the Organ clock)
Importance: Above–below relationship

Spleen primary channel (*zu tai yin jing*)

Connection: Deep Yin–Yin connection
Location: SP → HE (on the thorax). An **internal** branch of the Spleen primary channel spreads in the Heart (*xin*) and connects with the Heart primary channel.
Circulation: Circadian (according to the Organ clock)
Importance: The Heart primary channel receives Nutritive Qi (*ying qi*) from the Spleen primary channel (first circulation of the *ying qi* → 1.1.4)

Connections with other *zangfu*-Organ systems

Small Intestine (*xiaochang*), **Heart (*xin*)**, **Lung (*fei*)**

4.5.2 The Heart Divergent Channel (*zu shao yin jing bie*)

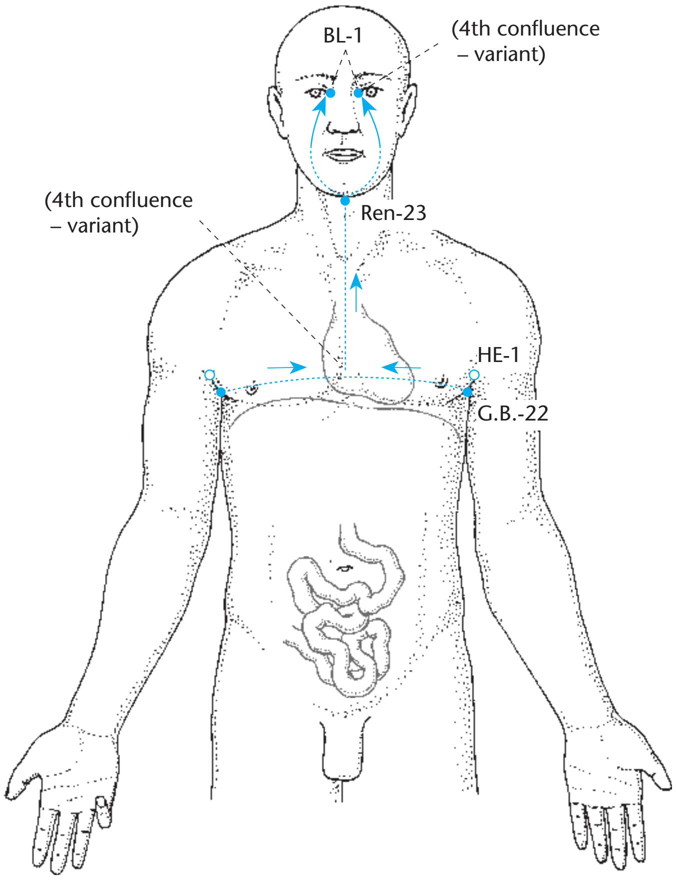

Pathway

The Heart divergent channel separates from the Heart primary channel at **HE-1** (*jiquan*) in the axilla

➡ penetrates the thorax at → **G.B.-22** (*yuanyue*) on the mid-axillary line and 3 cun distal to the axilla
➡ and reaches the Heart (*xin*) (see commentary below).
➡ It then ascends to the neck and crosses → **Ren-23** (*lianquan*)
➡ disperses in the cheek and reaches the medial canthus of the eye at → **BL-1** (*jingming*), where, according to most authors, it connects with a small branch coming from → **S.I.-18** (*quanliao*) to form one of the six he-confluences (here: S.I./BL as fourth confluence → 1.3).

Commentary: according to Solinas et al (1998) the Heart divergent channel meets the Small Intestine divergent channel and the Small Intestine primary channel at the Heart (*xin*) in order to bind with one of the six he-confluences.

Clinical importance

- Strengthens the relationship between the Heart and the thoracic region. Points on the Heart primary channel mainly treat disorders of the Heart and the thorax. There is no connection with the Small Intestine (*xiaochang*).

- Connects the Heart primary channel with the neck. Points on the Heart primary channel can be used to treat disorders of the throat and vocal chords such as laryngitis, pharyngitis, etc. Use points such as **HE-5, HE-8**.
- Strengthens the connection between the Heart and the inner canthus of the eye. Heart points can thus be used for disorders of the eyes.

4.5.3 The Heart Sinew Channel (*shou shao yin jing jin*)

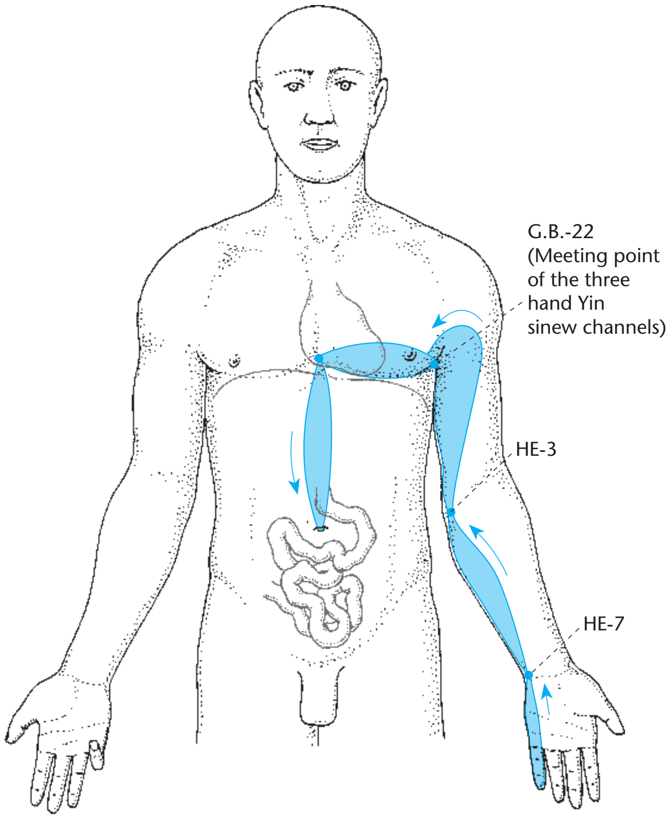

Pathway

The Heart sinew channel begins on the radial aspect of the little finger

- ➡ continues along the anterior and radial aspect of the little finger and the hand
- ➡ binds (*jie*) at the pisiform bone at **HE-7** (*shenmen*)
- ➡ ascends the anterior and ulnar aspect of the forearm, binds (*jie*) at the medial epicondyle at **HE-3** (*shaohai*) and continues to the axillary region, where it meets the Lung sinew channel, binds (*jie*) in the axilla and meets the other hand Yin sinew channels near → **G.B.-22** (*yuanye*)
- ➡ from here it continues at a deeper level, crosses the thoracic region, binds (*jie*) at → **Ren-17** (*shanzhong*), descends to the upper cardiac orifice, penetrates the diaphragm and terminates in the umbilical region.

Clinical importance

Pathology: Stiffness, cramps and pain along the pathway of the Heart sinew channel. Internal cramping in the cardiac region, for example disorders with angina-like pain due to stress, hiatal hernia and disorders of the stomach.

Indication: Mainly for *bi*-syndromes (painful obstruction syndrome) along the pathway of the Heart channel. The area covered by the Heart sinew channel is larger than that covered by the Heart primary channel. This explains why Heart primary channel points can also be used for disorders and diseases of the thorax and the chest, for example pain and tightness in the chest, mastitis and mastopathy. Since the Heart sinew channel also penetrates the diaphragm, it is also indicated for disorders such as hiccups, acid reflux and oesophagitis. Due to its connection with the umbilical region, the Heart sinew channel has a direct link to the Original Qi (*yuan qi*): if the Heart (*shen*) has suffered a shock, for example due to fainting or collapse, direct moxibustion at **Ren-8** can be very helpful.

4.5.4 The Heart *luo*-Connecting Vessel System (*shou shao yin luo mai*)

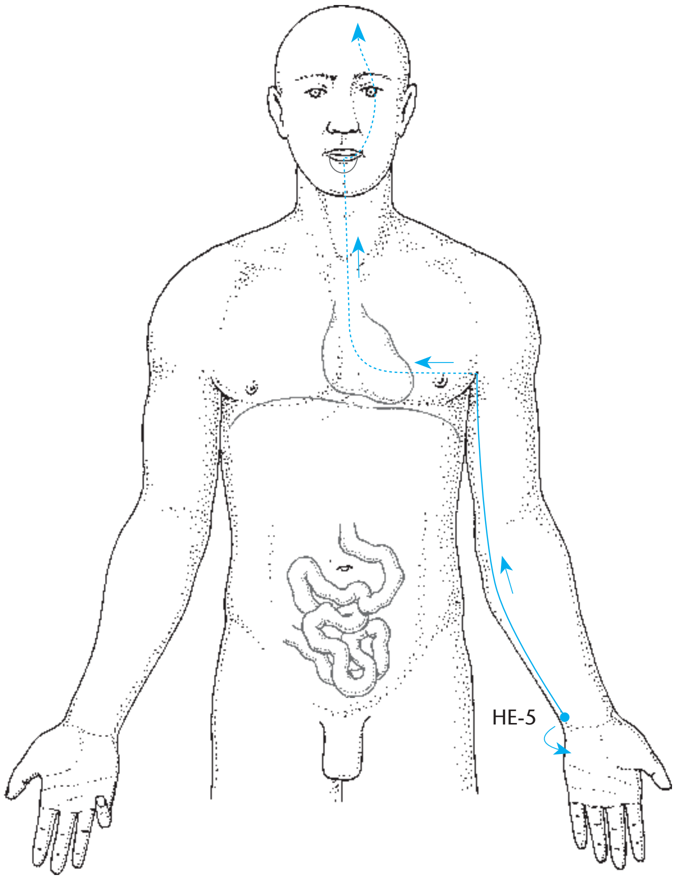

Pathway

The HE *luo*-connecting vessel separates from the Heart primary channel at its *luo*-connecting point **HE-5** (*tongli*) (→ 8.1.2). It forms a three-dimensional reticulate network, dividing into

Ch04.5-F10028.qxd 2/22/08 4:37 PM Page 206

multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Small Intestine primary channel; according to some schools of thought (for example Ngyen Van Nghi → Appendix) they travel as a **transverse** HE *luo*-connecting vessel to the *yuan*-source point → **S.I.-4** (*wangu*).
- A **longitudinal** division follows the pathway of the Heart primary channel along the anteromedial aspect of the arm, ascends to the axilla, reaches the Heart, ascends to the root of the tongue and terminates in the eye while further branches penetrate the brain.

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*)**: Distension and fullness of the thorax and diaphragm
- **Deficiency (*xu*)**: Aphasia, disorders of the vocal cords

4.5.5 Cutaneous Region (*shao yin pi bu*)

See description and figures → 1.6.

4.5.6 Points of the Heart Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): HE-7** (*shenmen*) 🟦🟦
- ***Luo*-connecting point (→ 8.1.2.): HE-5** (*tongli*) 🟦■
- ***Xi*-cleft point (→ 8.1.3): HE-6 (*yinxi*)** 🟦■
- **Associated Back-*shu* point (→ 8.1.4): BL-15** (*xinshu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): Ren-14** (*juque*) 🟦■
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood): **HE-9** (*shaochong*) 🟦■
  ying-spring point (Fire), *ben* point (Five Phases): **HE-8** (*shaofu*) 🟦■
  shu-stream point (Earth), sedation point: **HE-7** (*shenmen*) 🟦🟦
  jing-river point (Metal): **HE-4** (*lingdao*)
  he-sea point (Water): **HE-3** (*shaohai*) 🟦■
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8): –**
- **Lower *he*-sea points (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10): –**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): –**
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): HE-5** (*tongli*) 🟦■
- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points: –**

Points according to region

- **Local point (→ 8.2.1):** elbow – **HE-3** (*shaohai*) 🟦■
- **Distal points (→ 8.2.1):** tongue – **HE-5** (*tongli*) 🟦■; Heart – **HE-7** (*shenmen*) 🟦🟦, **HE-5** (*tongli*) 🟦■, **HE-3** (*shaohai*) 🟦■

Specific points according to the channel pathway (in numeric order)

- **HE-3** (*shaohai*) 🟦■: *he*-sea point (Water); local point for the elbow (→ 8.2.1); distal point for the Heart (→ 8.2.1)
- **HE-4** (*lingdao*): *jing*-river point (Metal)
- **HE-5** (*tongli*) 🟦■: *luo*-connecting point (→ 8.1.2); Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the tongue and Heart (→ 8.2.1)
- **HE-6** (*yinxi*) 🟦■: *xi*-cleft point (→ 8.1.3)
- **HE-7** (*shenmen*) 🟦🟦: *yuan*-source point (→ 8.1.1); distal point for the Heart (→ 8.2.1); *shu*-stream point (Earth), sedation point
- **HE-8** (*shaofu*) 🟦■: *ying*-spring point (Fire); *ben* point (Five Phases)
- **HE-9** (*shaochong*) 🟦■: *jing*-well point (Wood), tonification point

Summit Spring JIQUAN **HE-1**

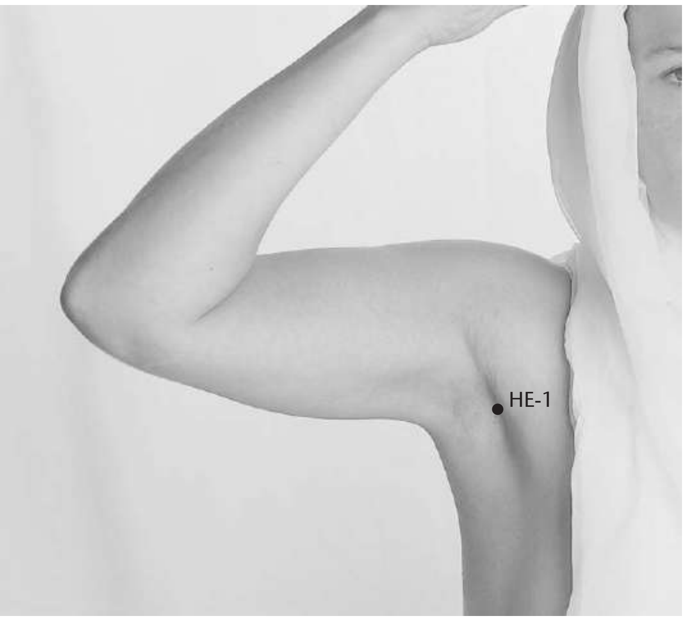

Location

With the arm abducted, in the centre of the axilla, medial to the axillary artery.

How to find

With the arm abducted, palpate along the thoracic wall in a superior direction towards the centre of the axilla until you reach the highest point.

Needling

Palpate for the artery before needling! 0.5–1 cun vertically in the direction of → **G.B.-21** (border of the trapezius, middle of the shoulder). Caution: Axillary plexus and artery. Needling in a medial direction may puncture the lung!

Actions/Indications

- Unbinds the chest
- Opens the channel
- Nourishes the Heart Yin

Special features

Entry point. Important point in Qigong practice: During practice (and in everyday life), the axillary region should always be slightly open to allow for the free circulation of Qi and Blood to and from the upper extremity.

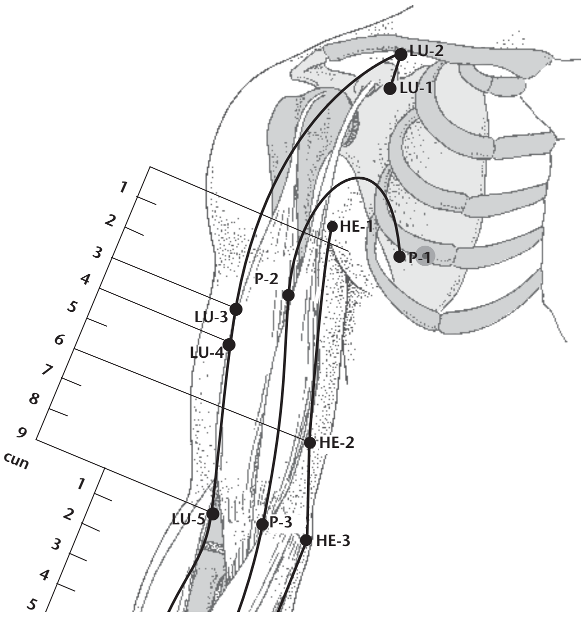

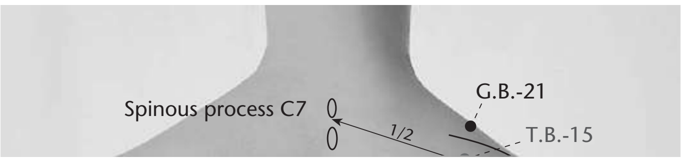

HE-2 *Green Spirit* QINGLING

Location

3 cun proximal to the cubital crease, on the medial border of the biceps brachii muscle.

How to find

With the elbow flexed, measure 3 cun from the ulnar end of the cubital crease (→ **HE-3**) towards the axilla. **HE-2** is located in a groove on the medial border of the biceps brachii muscle. Ask the patient to flex this muscle for easier location.

Needling

Vertically or obliquely 0.5–1 cun. Caution: Brachial artery! According to classic texts, needling is contraindicated, only moxibustion is permitted.

Actions/Indications

- Opens the channel

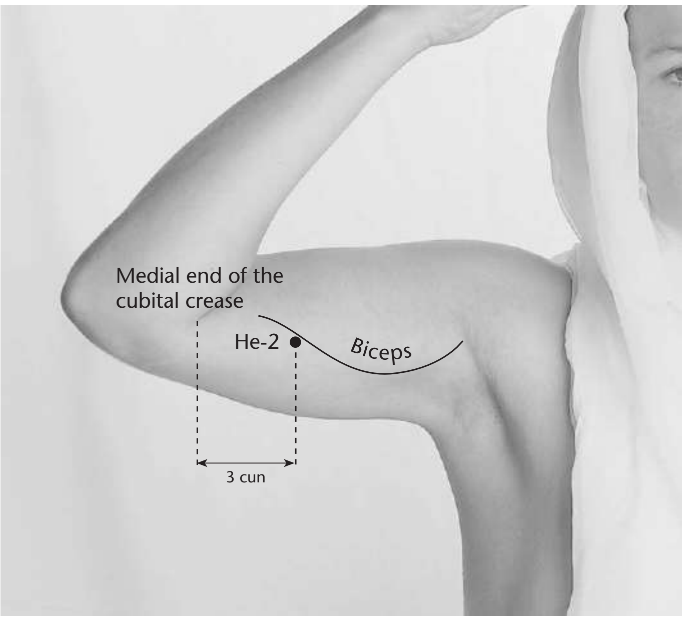
Medial end of the cubital crease
He-2
Biceps
3 cun

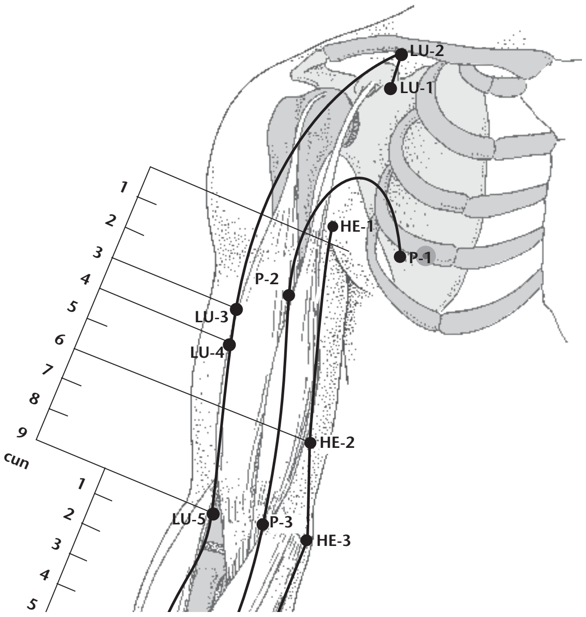
LU-2
LU-1
1
2
3
4
5
6
7
8
9
cun
HE-1
P-1
P-2
LU-3
LU-4
HE-2
1
2
3
4
5
LU-5
P-3
HE-3

Lesser Sea SHAOHAI **HE-3**

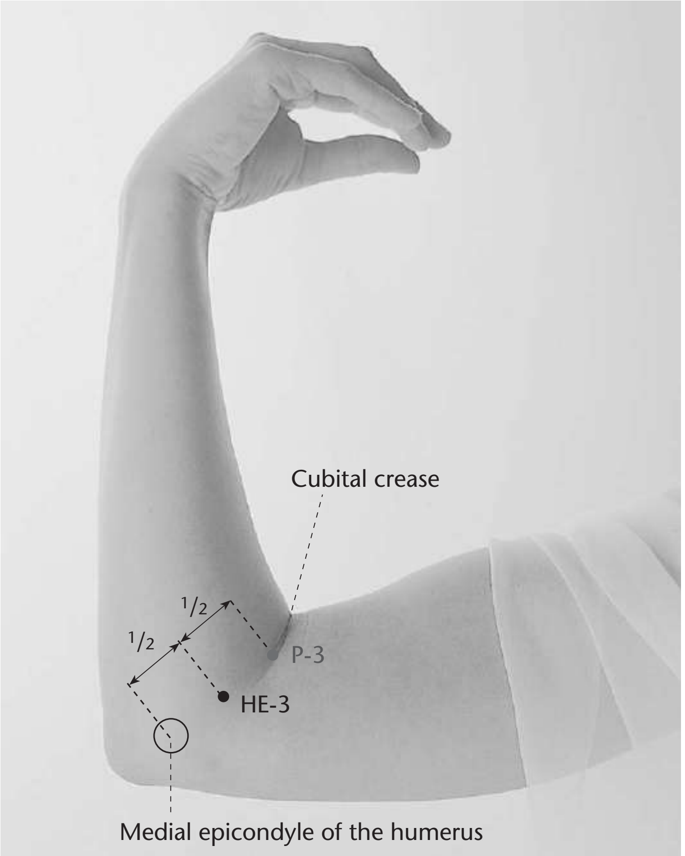
Cubital crease
1/2
1/2
P-3
HE-3
Medial epicondyle of the humerus

Location

With the elbow flexed, this point is in a depression between the ulnar end of the transverse cubital crease and the medial epicondyle of the humerus.

How to find

This point should be located with the elbow slightly flexed and the forearm in a supinated position. Find the ulnar end of the transverse cubital crease and palpate towards the medial epicondyle of the humerus (→ 3.2.2). **HE-3** is located in a depression approximately midway between → **P-3** and the medial epicondyle. **Or**: With the elbow fully flexed, **HE-3** is located directly at the medial end of the transverse crease.

Located on the same level are → **P-3** (ulnar to the tendon of the biceps muscle), → **LU-5** (radial to the tendon of the biceps muscle) and → **L.I.-11** (at the radial end of the crease).

Needling

0.5–1.5 cun obliquely in a distal or proximal direction or vertically towards → **L.I.-11**.

Actions/Indications

- Transforms Phlegm, clears Heat (from the heart), calms the *shen*
- Opens the channel and *luo*-vessels

Special features

He-sea point, Water point. Important local point, important calming point.

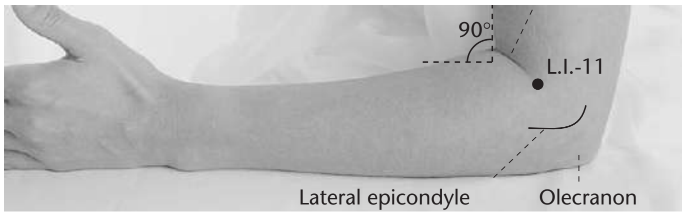
90°
L.I.-11
Lateral epicondyle
Olecranon

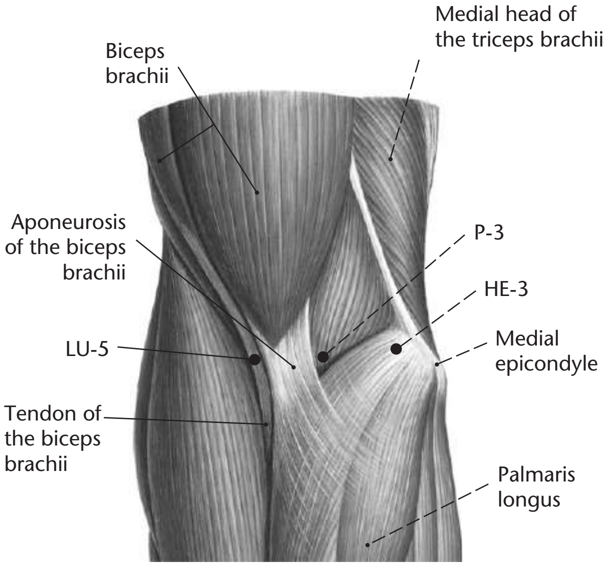
Medial head of the triceps brachii
Biceps brachii
Aponeurosis of the biceps brachii
P-3
HE-3
LU-5
Medial epicondyle
Tendon of the biceps brachii
Palmaris longus

Right arm anterior aspect

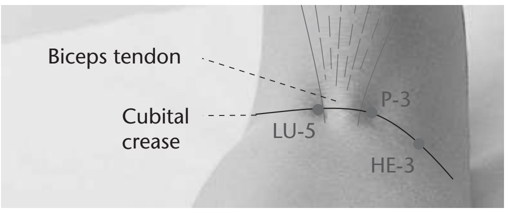
Biceps tendon
Cubital crease
P-3
LU-5
HE-3

Right arm

HE-4 *Spirit Path* LINGDAO

Location

1.5 cun proximal to the transverse wrist crease, on the radial side of the tendon of the flexor carpi ulnaris muscle.

How to find

The flexor carpi radialis muscle has a distinct tendon which can be easily palpated ulnar and proximal to the wrist crease. It attaches to the pisiform bone (➞ 3.3.3). **HE-4** is located directly on the radial side of the tendon.

➞ **HE-5**, ➞ **HE-6** and ➞ **HE-7** are located at 0.5 cun intervals distal to **HE-4**. ➞ **LU-7** is located approximately 1.5 cun proximal to the wrist crease on the radial aspect of the forearm.

Needling

0.3–1 cun vertically or obliquely in a proximal or distal direction.

Actions/Indications

- Opens the channel
- Nourishes the Heart, calms the *shen*
- Strengthens the voice

Special features

Jing-river point, Metal point.

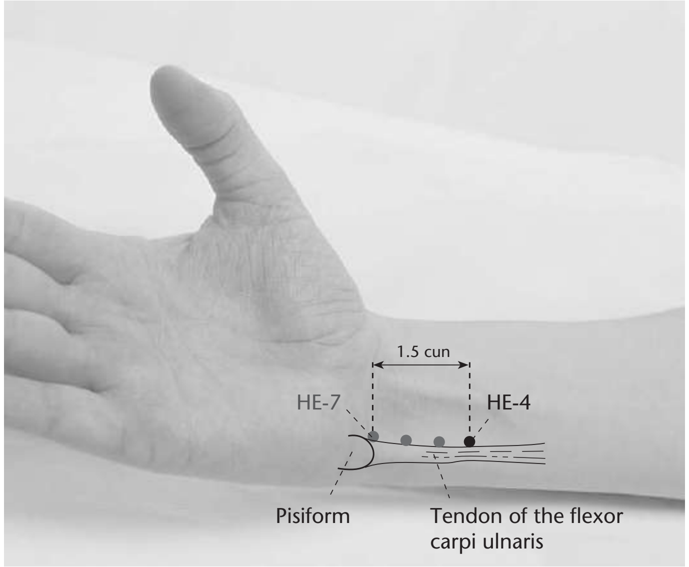
1.5 cun
HE-7
HE-4
Pisiform
Tendon of the flexor carpi ulnaris

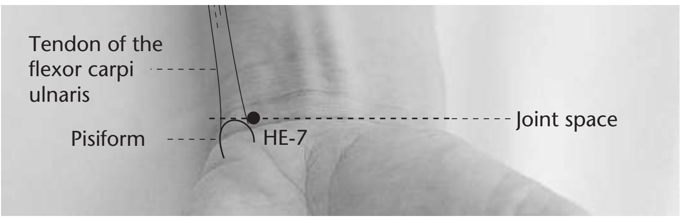
Tendon of the flexor carpi ulnaris
Pisiform
HE-7
Joint space

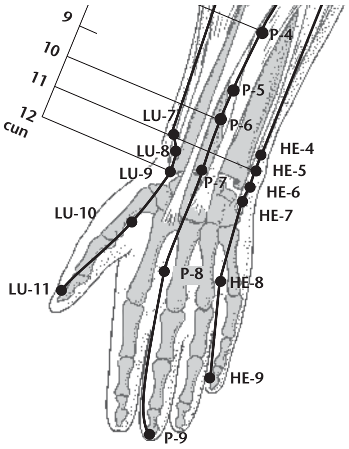
9
10
11
12 cun
P-4
P-5
P-6
LU-7
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9

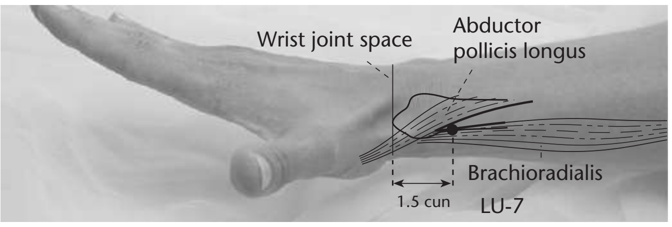
Wrist joint space
Abductor pollicis longus
Brachioradialis
1.5 cun
LU-7

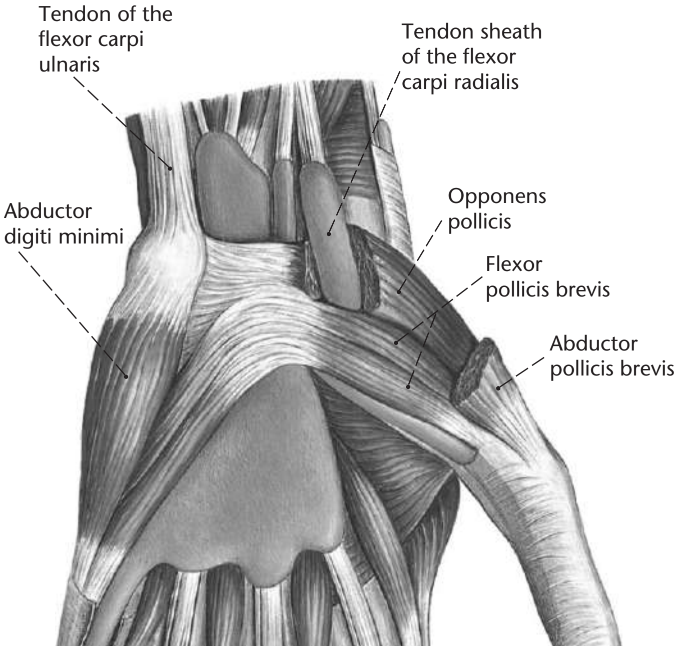
Tendon of the flexor carpi ulnaris
Tendon sheath of the flexor carpi radialis
Opponens pollicis
Abductor digiti minimi
Flexor pollicis brevis
Abductor pollicis brevis

Penetrating the Interior TONGLI **HE-5**

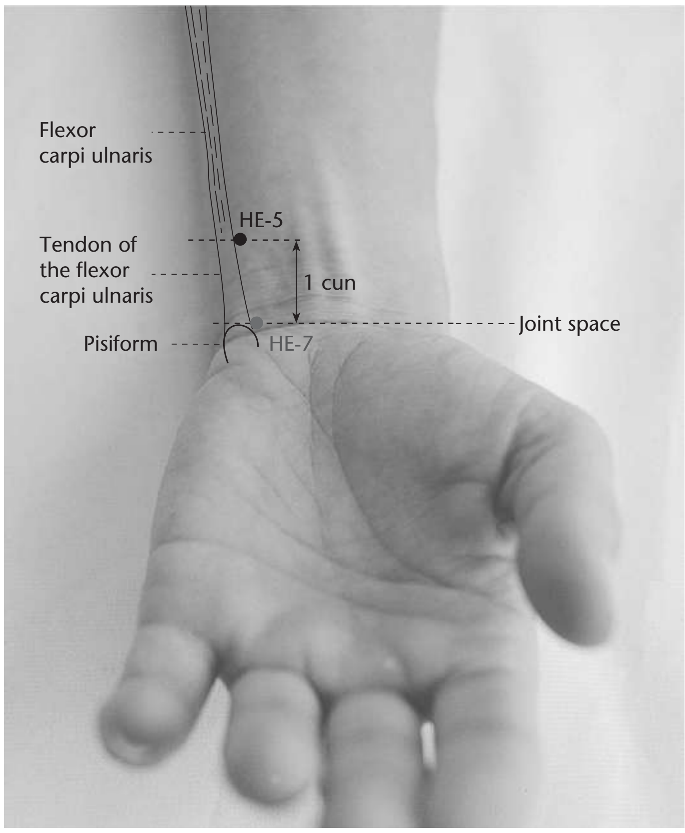
Flexor carpi ulnaris
HE-5
Tendon of the flexor carpi ulnaris
1 cun
Joint space
Pisiform
HE-7

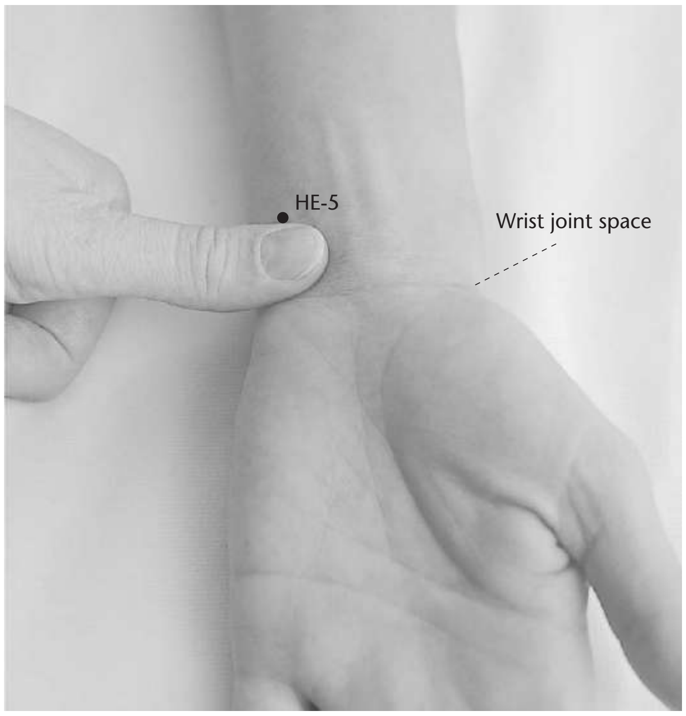
HE-5
Wrist joint space

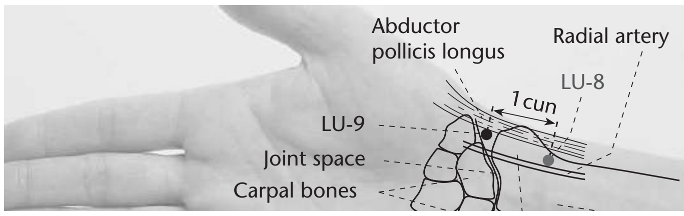
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones

Location

1 cun proximal to the ventral wrist joint space ('most distal transverse wrist crease'), on the radial side of the tendon of the flexor carpi ulnaris muscle.

How to find

Position the forearm in a relaxed fashion. The location of the ventral wrist crease is variable; therefore, the proximal border of the pisiform bone (→ 3.3.2) provides a more reliable landmark. → **HE-7** is located on the wrist joint space, radial to the distinct tendon of the flexor carpi ulnaris muscle. Locate **HE-5** 1 cun proximal to → **HE-7**, also on the radial side of the tendon. Located along a line on the radial aspect of the tendon are → **HE-6** (0.5 cun proximal to → **HE-7**) and → **HE-4** (1.5 cun proximal to **HE-7**). → **LU-8** is also located 1 cun proximal to the ventral wrist joint space, but on the radial aspect of the forearm.

Needling

Vertically 0.3–0.5 cun or obliquely 0.5–1 cun in a proximal or distal direction. Caution: Needling may be painful.

Actions/Indications

- Regulates and strengthens the Heart Qi (main point!) and the Heart rhythm
- Calms the *shen*
- Benefits the tongue
- Benefits the Bladder (via the *taiyang* channel)
- Opens the channel and *luo*-connecting vessels

Special features

Luo-connecting point, Ma Dan Yang Heavenly Star point. Main point for regulating and strengthening the Heart Qi.

HE-6 *Yin Cleft* YINXI

Location

0.5 cun proximal to the transverse wrist crease, on the radial side of the tendon of the flexor carpi ulnaris muscle.

How to find

The flexor carpi ulnaris muscle has a distinct tendon on the ventral aspect of the forearm that can be easily palpated ulnar and proximal to the wrist. **HE-6** is located on the radial aspect of the tendon, 0.5 cun proximal to → **HE-7** (located on the joint space).

→ **HE-5** and → **HE-4** are located proximally on the same line at 0.5 cun intervals.

Needling

0.3–1 cun vertically or obliquely in a proximal or distal direction. Caution: Ulnar artery/nerve. Painful point!

Actions/Indications

- Tonifies and regulates the Heart Yin and Blood, clears empty Heat in the Heart, calms the *shen*

Special features

Xi-cleft point

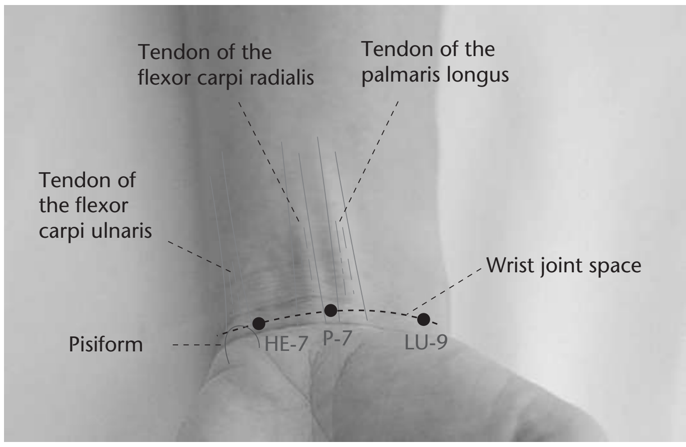
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
Tendon of the flexor carpi ulnaris
Wrist joint space
Pisiform
HE-7
P-7
LU-9

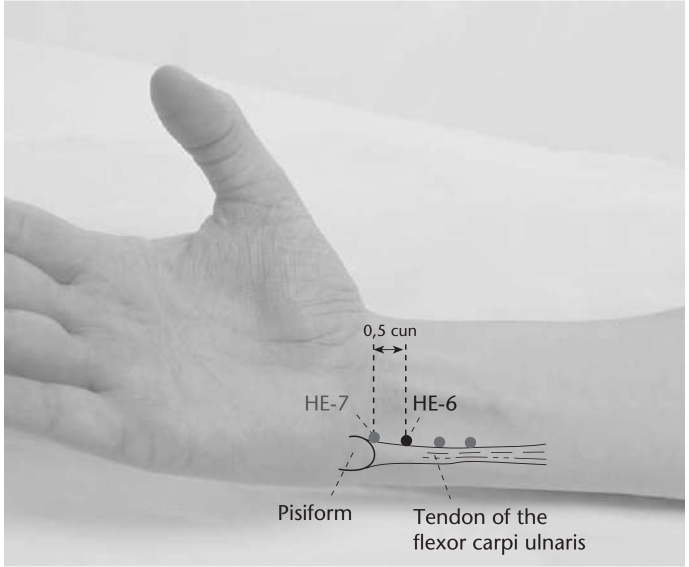
0,5 cun
HE-7
HE-6
Pisiform
Tendon of the flexor carpi ulnaris

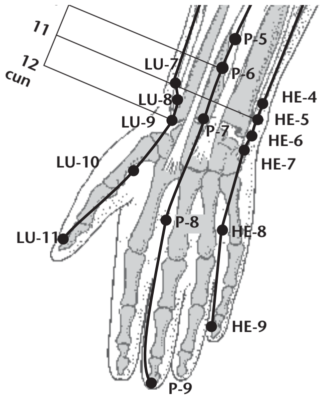
11
12
cun
P-5
LU-7
P-6
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

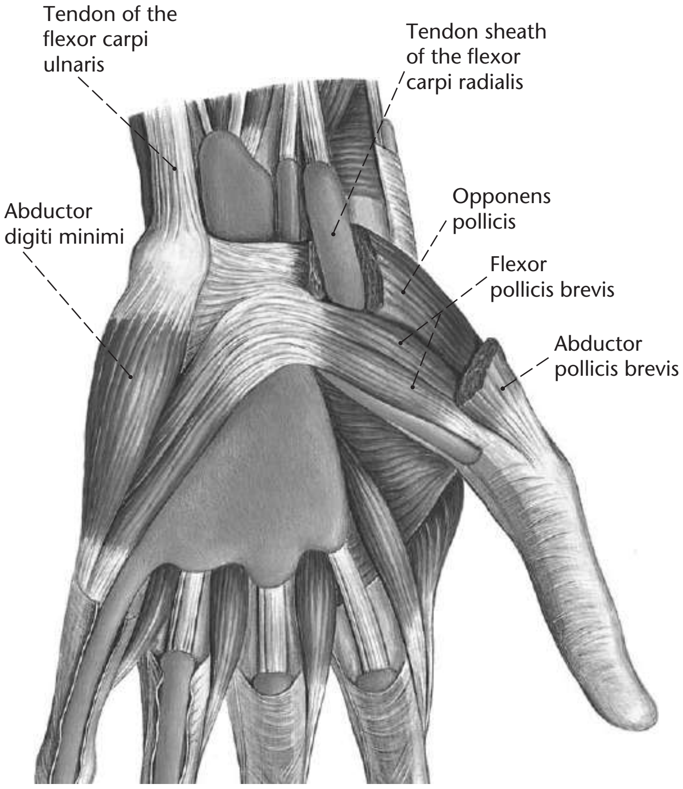
Tendon of the flexor carpi ulnaris
Tendon sheath of the flexor carpi radialis
Opponens pollicis
Abductor digiti minimi
Flexor pollicis brevis
Abductor pollicis brevis

Spirit Gate SHENMEN HE-7

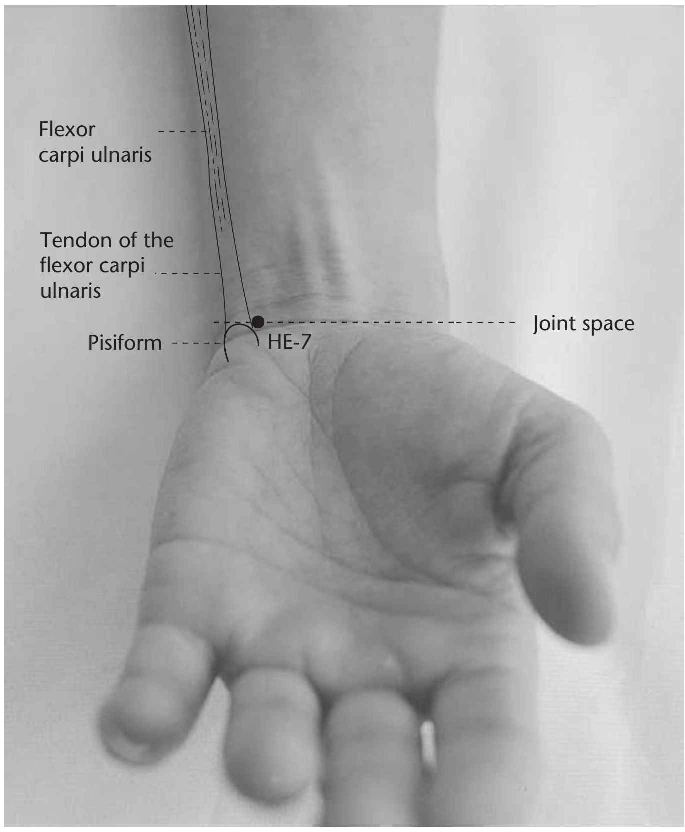
Flexor carpi ulnaris
Tendon of the flexor carpi ulnaris
Pisiform
HE-7
Joint space

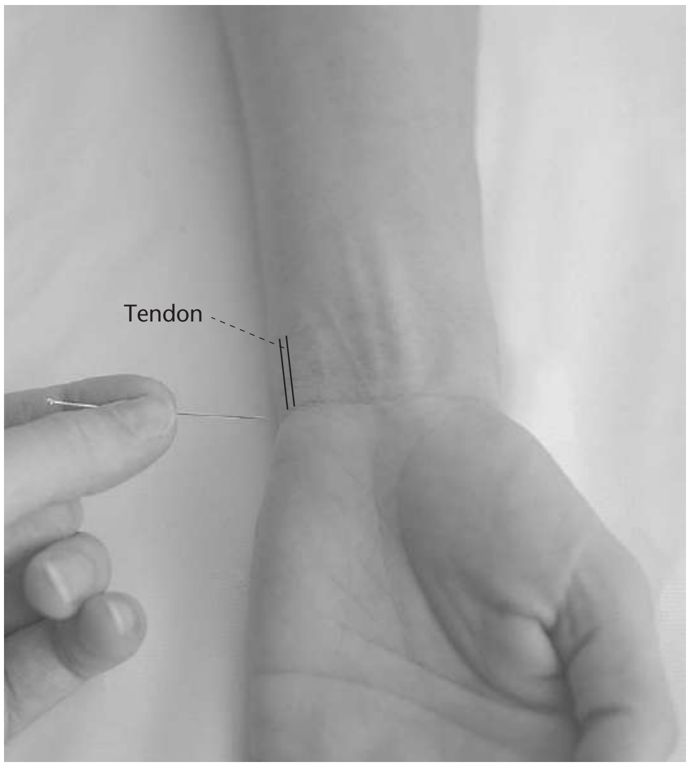

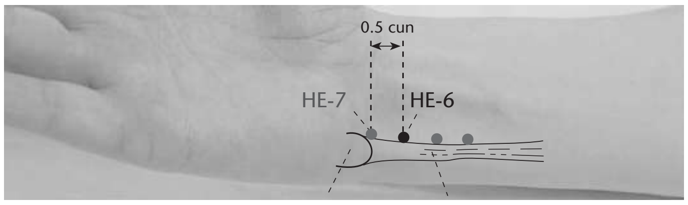
0.5 cun
HE-7
HE-6

Location

On the ventral aspect of the wrist joint space ('most distal wrist crease'), radial to the insertion of the tendon of the flexor carpi ulnaris muscle.

How to find

Position the forearm in a relaxed supinated position. As the location of the wrist crease can vary, the proximal border of the pisiform bone (→ 3.3.3) should be used for reference instead. **HE-7** is located radial to the insertion of the tendon of the flexor carpi ulnaris muscle, in a depression proximal to the pisiform bone.

→ **HE-6**, → **HE-5** and → **HE-4** are all located on a line on the radial aspect of the tendon at 0.5 cun intervals. Also located on the wrist crease are → **P-7** (between the tendons of the palmaris longus and the flexor carpi radialis muscles) and → **LU-9** (on the radial aspect, lateral to the radial artery).

Needling

0.3–0.5 cun vertically or transversely deep to the tendon (parallel to the wrist crease 0.5–0.8 cun). Caution: Ulnar artery/nerve. Painful point!

Actions/Indications

- Regulates and strengthens the Heart (Blood and Yin), calms the *shen*
- Clears Heat from the Heart channel
- Moves Qi locally for disorders of the wrist

Special features

Yuan-source point, *shu*-stream point, Earth point, sedation point. Major point for calming the *shen*.

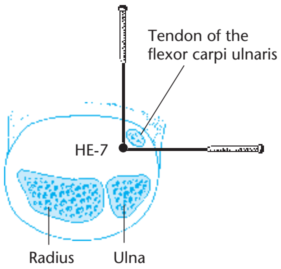
Tendon of the flexor carpi ulnaris
HE-7
Radius
Ulna

HE-8 *Lesser Palace* SHAOFU

Location

On the palm, between the 4th and 5th metacarpal bones.

How to find

When making a fist, **HE-8** is generally located below the tip of the little finger, in the depression between the 4th and 5th metacarpal bones. **HE-8** often lies between the two transverse palmar creases. It is often sensitive to pressure.

→ **P-8** is also located on the palm, generally below the tip of the middle finger when making a fist (between the 2nd and 3rd metacarpal bones).

Needling

Vertically 0.5–1 cun. Caution: Needling this point may be painful. Provide a competitive stimulus by pressing the palm with the fingernail close to the point before inserting the needle. Ask the patient to breathe out during needle insertion.

Actions/Indications

- Clears Heat in the Heart and Small Intestine
- Regulates the Heart Qi for Liver Qi stagnation
- Calms the *shen* and strengthens the Heart Qi
- Opens the channel, especially for contractures of the fingers

Special features

Ying-spring point, Fire point, *ben* point (Five Phases).

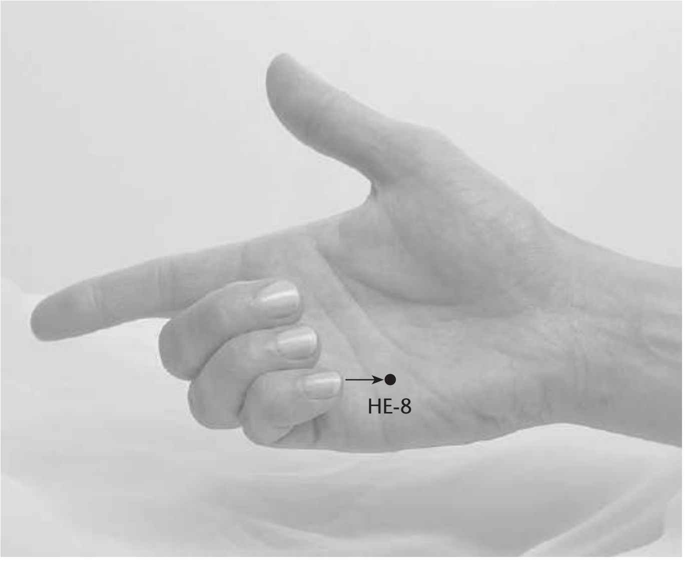

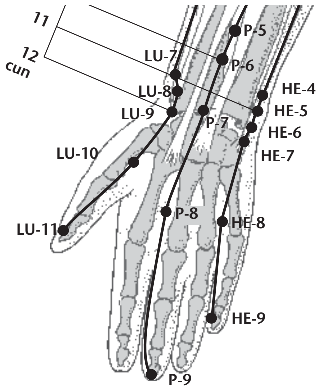

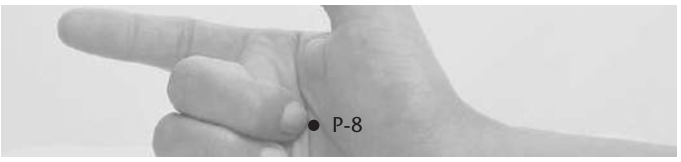

Lesser Rushing SHAOCHONG **HE-9**

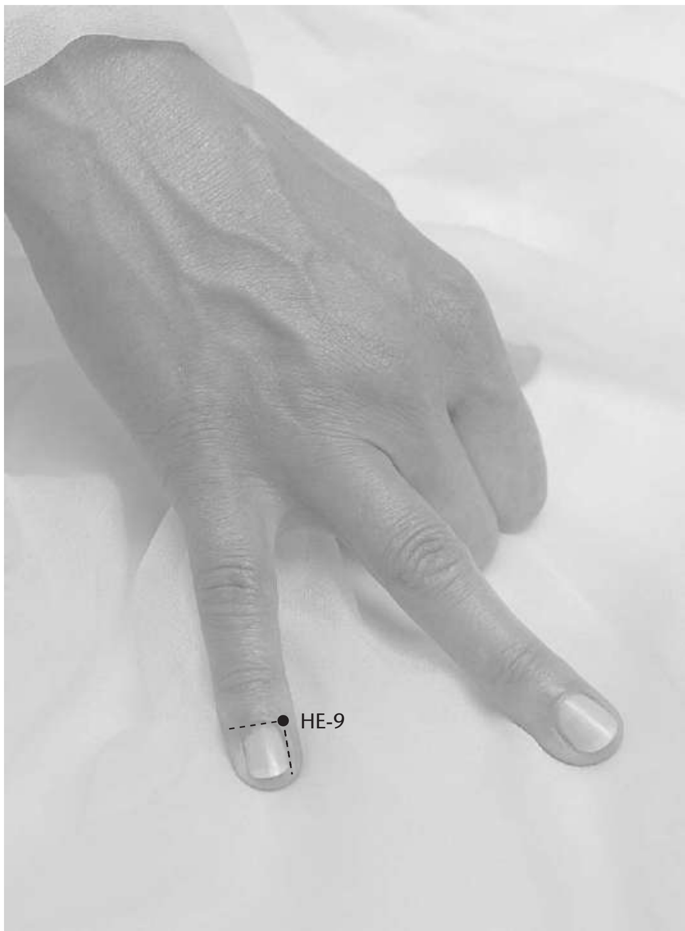

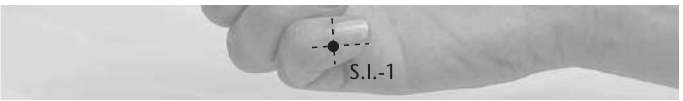

Location

On the little finger, 0.1 cun lateral to the radial corner of the nail.

How to find

This point is located at the junction of two tangents that border the radial and proximal aspect of the nail, approximately 0.1 cun from the actual corner of the nail. → **S.I.-1** is located on the opposite (ulnar) aspect of the nail.

Needling

0.2–0.3 cun vertically or obliquely in a proximal direction. For acute disorders (pain, inflammation), prick to bleed with a needle, lancet or three-edged needle. Caution: Painful point!

Actions/Indications

- Revives consciousness
- Clears Heat from the opposite end of the channel
- Regulates Qi in the chest and clears Heat in the Heart
- Opens the channel

Special features

Jing-well point, Wood point, tonification point, exit point.

4.6 The Small Intestine Channel System – Hand *taiyang* (*shou tai yang jing luo*)

4.6.1 The Small Intestine Primary Channel (*shou tai yang jing*)

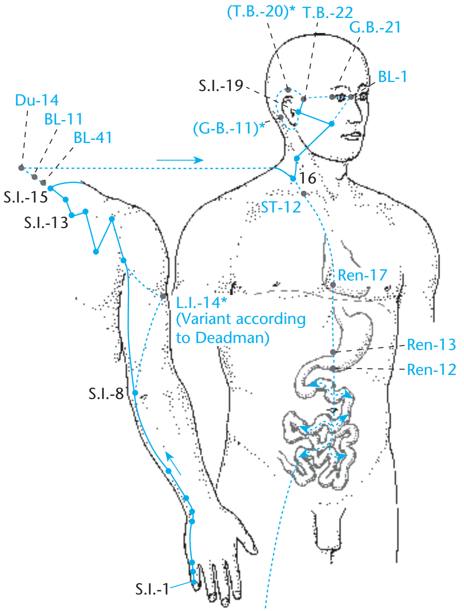
(T.B.-20)*
T.B.-22
G.B.-21
BL-1
S.I.-19
(G-B.-11)*
Du-14
BL-11
BL-41
S.I.-15
S.I.-13
16
ST-12
Ren-17
L.I.-14*
(Variant according to Deadman)
Ren-13
Ren-12
S.I.-8
S.I.-1

*meeting points only mentioned by some authors

To ST-39
Lower *he*-Sea point of the Small Intestine

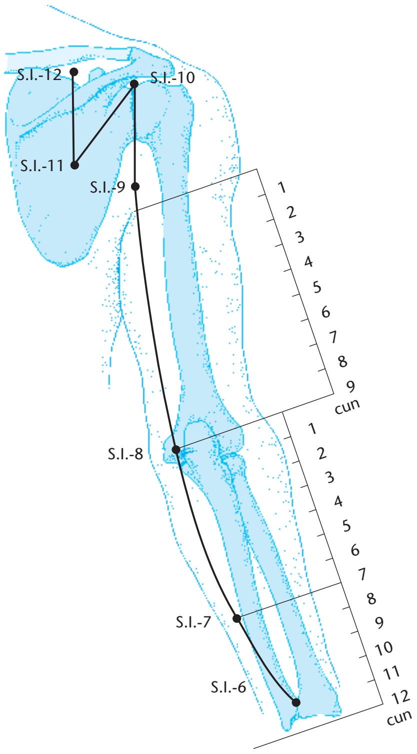
S.I.-12
S.I.-10
S.I.-11
S.I.-9
1
2
3
4
5
6
7
8
9
cun
S.I.-8
1
2
3
4
5
6
7
8
9
10
11
12
cun
S.I.-7
S.I.-6

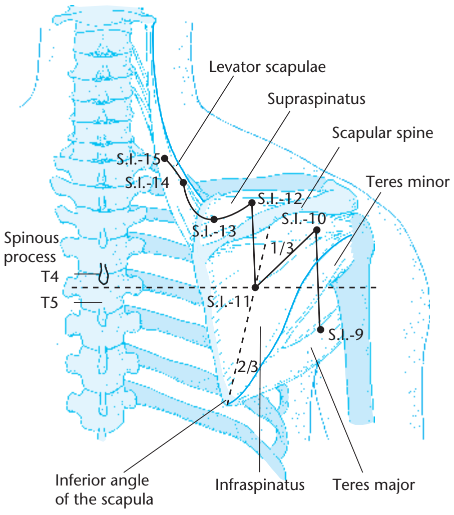
Levator scapulae
Supraspinatus
Scapular spine
S.I.-15
S.I.-14
Teres minor
S.I.-12
S.I.-13
S.I.-10
Spinous process
1/3
T4
T5
S.I.-11
S.I.-9
2/3
Inferior angle of the scapula
Infraspinatus
Teres major

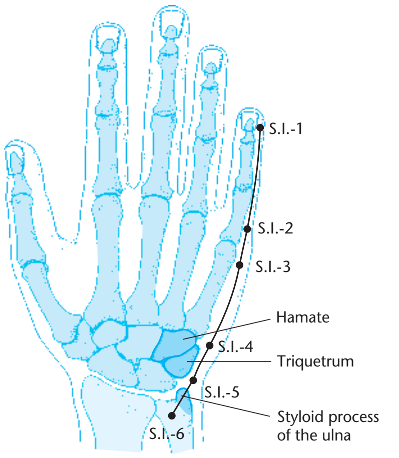
S.I.-1
S.I.-2
S.I.-3
Hamate
S.I.-4
Triquetrum
S.I.-5
Styloid process of the ulna
S.I.-6

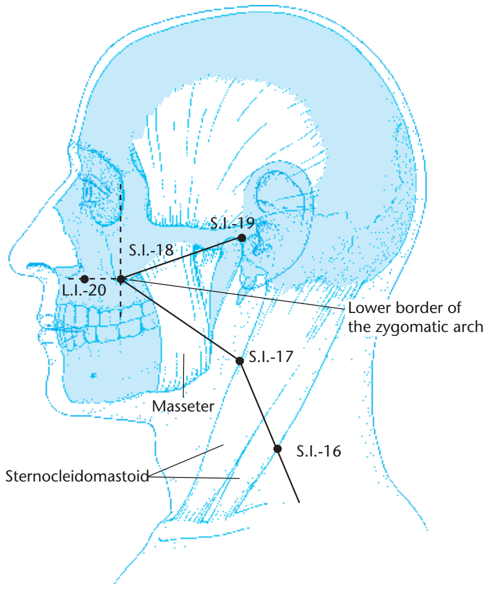
S.I.-19
S.I.-18
L.I.-20
Lower border of the zygomatic arch
S.I.-17
Masseter
S.I.-16
Sternocleidomastoid

Pathway

The **external** pathway of the Small Intestine primary channel begins on the little finger, on the ulnar corner of the nail, at **S.I.-1** (*shaoze*). This point is reached by a branch that separates from the Heart primary channel at → **HE-9** (*shaochong*) (hand Yin–Yang connection of the second great circuit).

The channel runs along the ulnar border of the hand, passes the wrist and continues along the ulna to the medial aspects of the elbow.

- passes between the olecranon and the medial epicondyle, ascends the posterior aspect of the humerus to the shoulder joint and, according to some authors, meets → **L.I.-14**
- zig-zags across the scapula, passes → **BL-41** (*fufen*) and → **BL-11** (*dazhu*) and meets the extraordinary vessel *du mai* and the other Yang primary channels at → **Du-14** (*dazhui*), below the spinous process of C7
- continues anteriorly to the supraclavicular fossa to → **ST-12** (*quepen*).

An **internal** branch originates at → **ST-12** and runs inferiorly to the Heart (*xin*), the paired *zang*-Organ, continues to → **Ren-17** and descends along the oesophagus, penetrates the diaphragm and continues to the Stomach (*wei*). Here it crosses → **Ren-13** (*shangwan*) and → **Ren-12** (*zhongwan*), further descends and spreads in its pertaining *fu*-Organ, the Small Intestine (*xiaochang*).

According to the *Neijing Ling Shu*, an **internal** branch courses from the Small Intestine (according to some authors from the Small Intestine primary channel) to → **ST-39** (*xiajuxu*), the lower *he*-sea point of the Small Intestine.

The **external** pathway of the Small Intestine primary channel continues from the supraclavicular fossa along the neck and ascends to the cheek to **S.I.-18** (*quanliao*).

From **S.I.-18**

- **a branch** courses to → **BL-1** (*jingming*) at the inner canthus of the eye, where it connects with the Bladder primary channel (six channel theory, Yang axes: *taiyang*)
- **a further branch** travels to the outer canthus of the eye, intersects with → **G.B.-1** (*tongziliao*), according to some authors also with → **T.B.-20** and → **G.B.-11**, continues to → **T.B-22** (*erheliao*) anterior to the ear and terminates anterior to the ear deep to **S.I.-19** (*tinggong*)

Clinical importance (➡ 1.2)

Exterior (*biao*) signs and symptoms: Mouth and tongue ulcers, cheek pain, neck pain, occipital stiffness, pain on the lateral aspect of the shoulder and the upper arm

Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain and distension of the upper abdomen, radiating to the genital region, diarrhoea, epigastric pain with constipation

Connections of the Small Intestine primary channel

Connections with other channels

Heart primary channel (*shou shao yin jing*)

Connection: Hand Yin–Yang connection of the second great circuit
Location: **H-9** → **S.I.-1** (on the hand)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Bladder primary channel (*zu tai yang jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *tai yang* (Yang axes of the second great circuit)
Location: **S.I.-18** → **BL-1** (on the head)
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with other *zangfu*-Organ systems

Heart (*xin*), Stomach (*wei*), **Small Intestine (*xiaochang*)**

4.6.2 The Small Intestine Divergent Channel (*shou tai yang jing bie*)

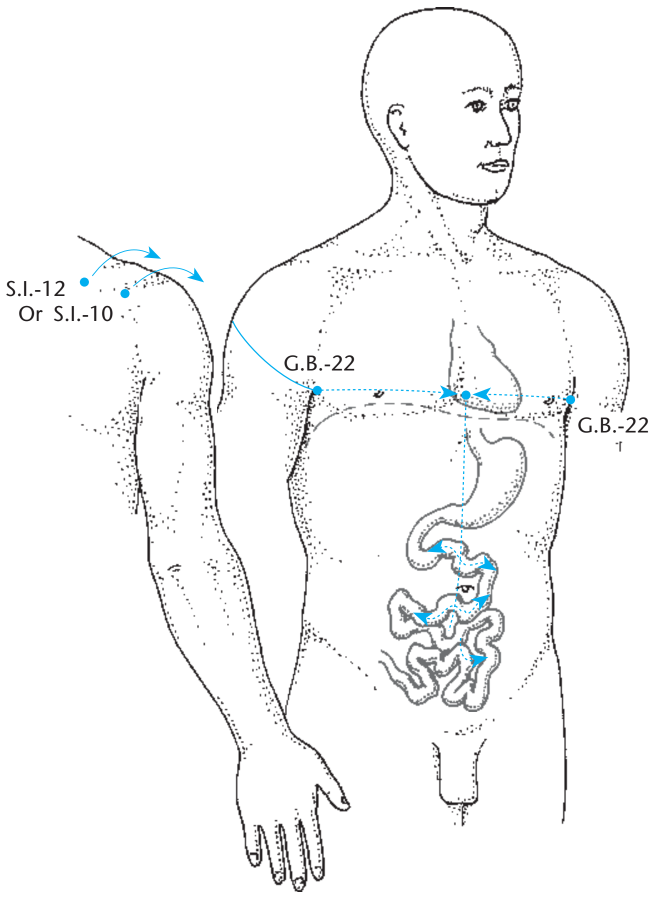

Pathway

The Small Intestine divergent channel separates from the Small Intestine primary channel at the shoulder, depending on the school of thought in the area of → **S.I.-12** (*bingfeng*) or in the area of → **S.I.-10** (*naoshu*)

- descends towards the axilla and continues to → **G.B.-22** (*yuanye*)
- penetrates the thorax and spreads in the Heart (*xin*) where, according to most authors, it forms the fourth confluence
- penetrates the diaphragm to connect with its pertaining *fu*-Organ, the Small Intestine

Clinical importance

- Strengthens the relationship between the Heart and the Small Intestine (*zangfu*-Organ system). It provides an explanation for the pattern of Heart Fire transferring to the Small Intestine and the Bladder.

4.6.3 The Small Intestine Sinew Channel (*shou tai yang yin jing jin*)

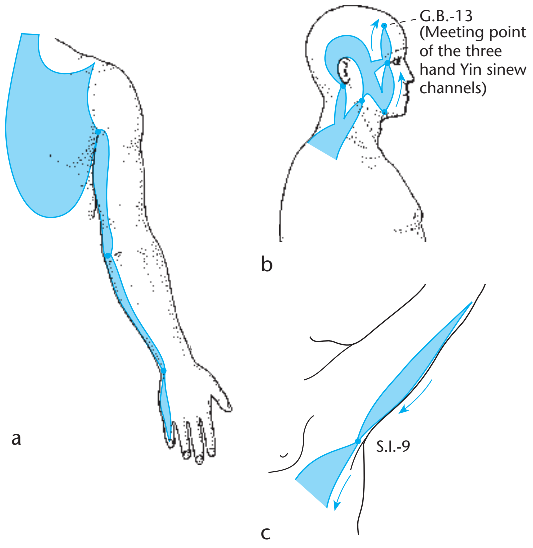
G.B.-13 (Meeting point of the three hand Yin sinew channels)
b
a
S.I.-9
c

Pathway

The Small Intestine sinew channel begins on the ulnar aspect of the little finger at ➞ **S.I.-1** (*shaoze*)

- ➡ continues along the ulnar aspect of the little finger and the ulnar border of the hand
- ➡ runs to **S.I.-5** (*yanggu*) distal to the styloid process of the ulna and binds (*jie*) at the wrist
- ➡ travels along the ulnar aspect of the forearm to the elbow, continues to the axilla, binding at **S.I.-8** (*xiaohai*) and the posterior aspect of the axilla.

From here a branch courses into the axilla where it binds (*jie*). Another branch spreads across the scapula and ascends anterior to the Bladder and Gall Bladder sinew channels and posterior to the Stomach sinew channel to the lateral aspect of the neck, where it meets the Triple Burner sinew channel.

At the neck it divides into **two branches:**

- ➡ the **anterior branch** runs to the mandibular angle where it binds (*jie*), continues along the mandible to the ear and the zygomatic bone to bind (*jie*) at the lateral aspect of the orbit
- ➡ the **posterior branch** binds (*jie*) at the mastoid process (from here a small branch runs to the ear). The branch curves around the ear, descends to the cheek and binds (*jie*) at the insertion of the masseter muscle on the mandible, continues to the lateral aspect of the orbit and reaches ➞ **G.B.-13** (*benshen*) on the frontoparietal aspect of the head, where it meets the other hand Yang sinew channels.

Clinical importance

Pathology: Distending sensations, stiffness and pain of the little finger, impaired range of motion of the upper extremity and shoulder. Pain, stiffness and impaired range of motion of the cervical spine and occiput. Tinnitus and ear pain radiating to the mandible.

Indication: Mainly for *bi*-syndromes (painful obstruction syndrome) along the pathway of the S.I. channel. The area covered by the Small Intestine sinew channel is larger than that covered by the Small Intestine primary channel. This explains why Small Intestine primary channel points can also be used for disorders and diseases of the cervical and thoracic spine and the whole face (for example frontal and lateral headaches).

4.6.4 The Small Intestine *luo*-Connecting Vessel System (*shou tai yang luo mai*)

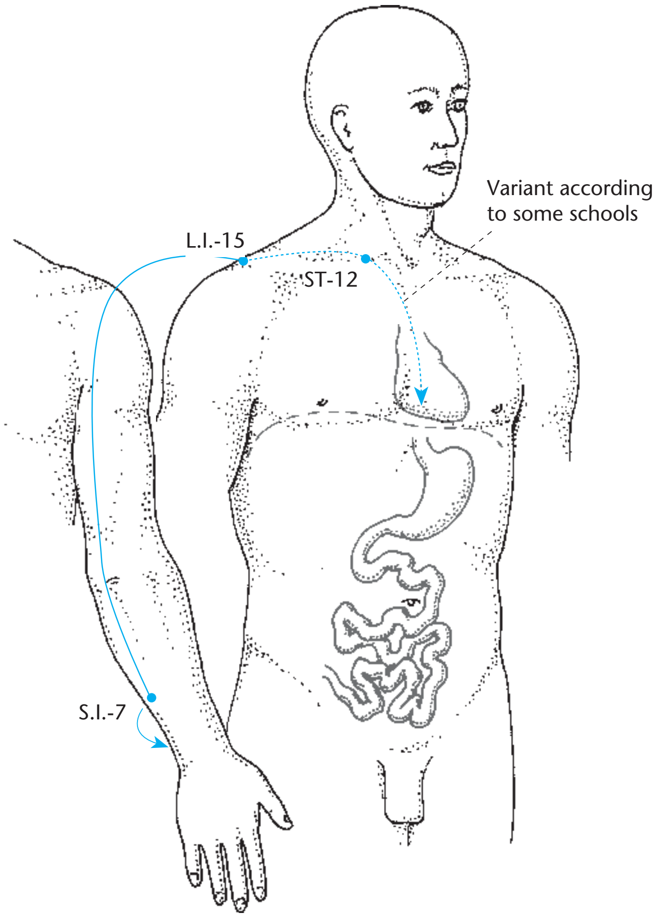

Pathway

The S.I. *luo*-connecting vessel separates from the Small Intestine primary channel at its *luo*-connecting point **S.I.-7** (*zhizheng*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Heart primary channel, according to some schools of thought (for example Ngyen Van Nghi → Appendix) they travel as a **transverse** S.I. *luo*-connecting vessel to the *yuan*-source point → **HE-7** (*shenmen*).
- A **longitudinal** division runs to the medial condyle of the humerus, ascends to the shoulder to → **L.I.-15** (*jianyu*). According to some schools of thought the vessel spreads and terminates here, while according to other schools it traverses the acromion, continues to → **ST-12** (*quepen*) in the supraclavicular fossa, penetrating the thorax in order to connect with the Heart.

Clinical importance (➡ 8.1.2)

Pathology

- **Excess** (***shi***): Weak and unstable joints, weakness and paralysis of the elbow and the arm
- **Deficiency** (***xu***): Warts

4.6.5 Cutaneous Region (*tai yang pi bu*)

See description and figures → 1.6.

4.6.6 Points of the Small Intestine Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): S.I.-4** (*wangu*)
- ***Luo*-connecting point (→ 8.1.2.): S.I.-7** (*zhizheng*)
- ***Xi*-cleft point (→ 8.1.3): S.I.-6** (*yanglao*) 🟦■
- **Associated Back-*shu* point (→ 8.1.4): BL-27** (*xiaochangshu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): Ren-4** (*guanyuan*) 🟦🟦
- **Five *shu*-transporting point (→ 8.1.6):**
  jing-well point (Metal): **S.I.-1** (*shaoze*)
  ying-spring point (Water): **S.I.-2** (*qiangu*)
  shu-stream point (Wood), tonification point: **S.I.-3** (*houxi*) 🟦■
  jing-river point (Fire), *ben* point: **S.I.-5** (*yanggu*)
  he-sea point (Earth), tonification point: **S.I.-8** (*xiaohai*)
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8) of the *du mai*: S.I.-3** (*houxi*) 🟦🟦
- **Lower *he*-sea points (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the BL channel, the *yang wei mai, yang qiao mai*: **S.I.-10** (*naoshu*)
  – with the L.I., T.B., G.B. channels: **S.I.-12** (*bingfeng*)
  – with the T.B. channel: **S.I.-18** (*quanlao*) 🟦■
  – with the T.B., G.B.: **S.I.-19** (*tinggong*) 🟦■
  – of other channels with the S.I. channel: **L.I.-14*, Du-14, BL-41, BL-11, ST-12, Ren-17*, Ren-13, Ren-12, G.B.-1, G.B.-11*, T.B.-20*, T.B.-22, BL-1**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): S.I.-16** (*tianchuang*), **S.I.-17** (*tianrong*) 🟦■
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): –**
- **Sun Si Miao Ghost point (→ 8.1.15): –**

Points according to region

- **Local points (→ 8.2.1):** face and cheek – **S.I.-18** (*quanlao*) 🟦■; jaw – **S.I.-19** (*tinggong*) 🟦■; neck – **S.I.-17** (*tianrong*) 🟦■; ears – **S.I.-19** (*tinggong*) 🟦■; shoulder – **S.I.-9** (*jianzhen*), **S.I.-10** (*naoshu*), **S.I.-11** (*tianzong*), **S.I.-12** (*bingfeng*) ■■, **S.I.-13** (*quyuan*), **S.I.-14** (*jianwaishu*) 🟦■; elbow – **S.I.-8** (*xiaohai*); hand – **S.I.-4** (*wangu*), **S.I.-5** (*yanggu*); finger (numbness and pain) – **S.I.-3** (*houxi*) 🟦■; finger (stiffness) – **S.I.-3** (*houxi*) 🟦■, **S.I.-5** (*yanggu*)
- **Adjacent points (→ 8.2.1):** shoulder – **S.I.-9–S.I.-15**; finger – **S.I.-5** (*yanggu*)
- **Distal points (→ 8.2.1):** for the occiput – **S.I.-3** (*houxi*) 🟦■; for the eyes – **S.I.-6** (*yanglao*) 🟦■; for the cervical and thoracic spine – **S.I.-3** (*houxi*) 🟦■; for the thoracic spine – **S.I.-6** (*yanglao*) 🟦■; for the lumbar region – **S.I.-3** (*houxi*) 🟦■, **S.I.-6** (*yanglao*) 🟦■

Specific points according to the channel pathway (in numerical order)

- **S.I.-1** (*shaoze*): *jing*-well point (Metal)
- **S.I.-2** (*qiangu*): *ying*-spring point (Water)
- **S.I.-3** (*houxi*) 🟦■: *shu*-stream point (Wood), tonification point, opening point (→ 8.1.8) of the *du mai*, distal point for the occiput and the lumbar region (→ 8.2.1); local point for the fingers (→ 8.2.1); distal point for the cervical and thoracic spine (→ 8.2.1)
- **S.I.-4** (*wangu*): *yuan*-source point (→ 8.1.1); local point for the hand (→ 8.2.1)
- **S.I.-5** (*yanggu*): *jing*-river point (Fire), *ben* point (Five Phases); local point for the hand and fingers (stiffness) (→ 8.2.1); regional point for the fingers (→ 8.2.1)
- **S.I.-6** (*yanglao*) 🟦■: *xi*-cleft point (→ 8.1.3); distal point for the eyes, the thoracic and lumbar spine (→ 8.2.1)
- **S.I.-7** (*zhizheng*): *luo*-connecting point (→ 8.1.2)
- **S.I.-8** (*xiaohai*): *he*-sea point (Earth), sedation point; local point for the elbow (→ 8.2.1)
- **S.I.-9** *(jianzhen):* local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-9 – S.I.-15**: adjacent points for the shoulder (often sensitive to pressure) (→ 8.2.1)

---

* According to Deadman et al 1998.

- **S.I.-10** (*naoshu*): *jiaohui*-meeting point with the BL channel, the *yang wei mai, yang qiao mai* (→ 8.1.10); local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-11** (*tianzong*) 🟦⬛: local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-12** (*bingfeng*): *jiaohui*-meeting point with the L.I., T.B. and G.B. channels (→ 8.1.10); local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-13** (*quyuan*): local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-14** (*jianwaishu*) 🟦⬛: local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-16** (*tianchuang*): Window of Heaven point (→ 8.1.12)
- **S.I.-17** (*tianrong*) 🟦⬛: Window of Heaven point (→ 8.1.12); local point for the neck (→ 8.2.1)
- **S.I.-18** (*quanlao*) 🟦⬛: local point for the face and cheek (→ 8.2.1); *jiaohui*-meeting point with the T.B. channel (→ 8.1.10)
- **S.I.-19** (*tinggong*) 🟦⬛: *jiaohui*-meeting point with the T.B., G.B. channels (→ 8.1.10); local point for disorders of the ear and jaw (→ 8.2.1)

Lesser Marsh SHAOZE S.I.-1

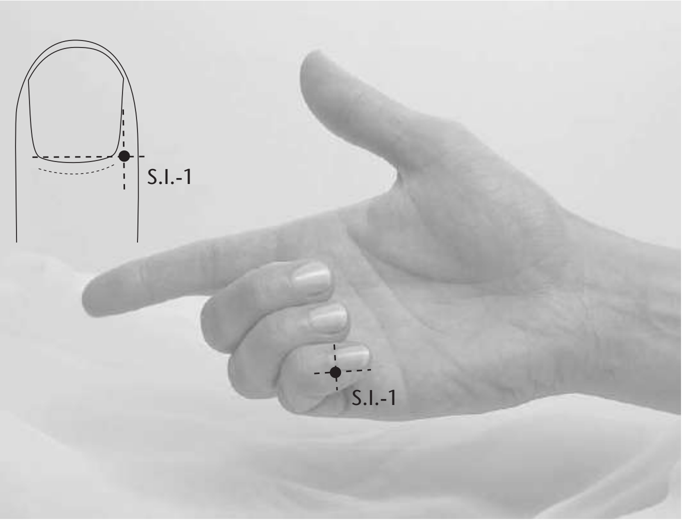

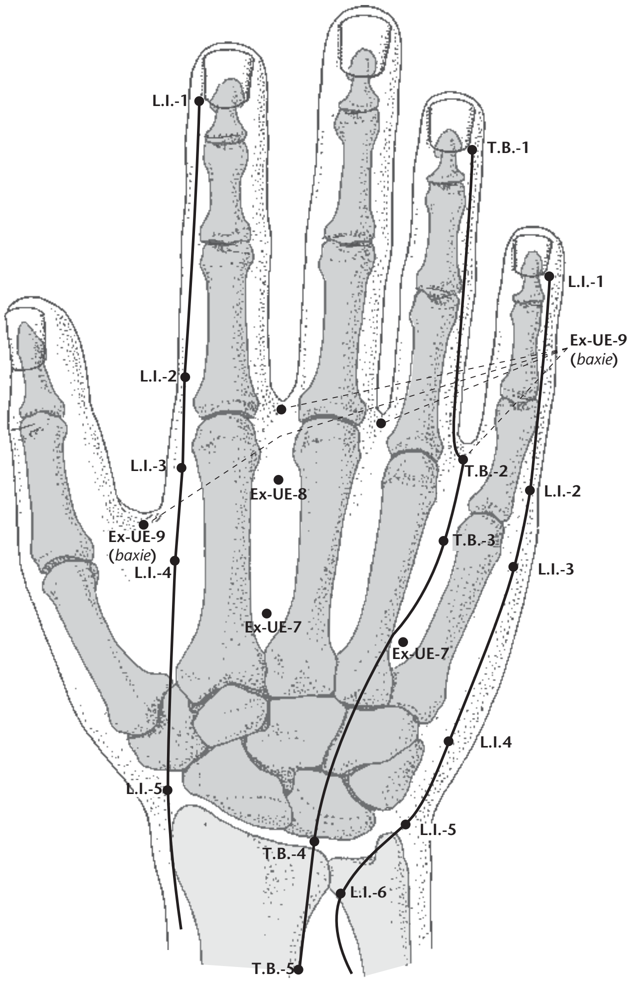
L.I.-1
T.B.-1
L.I.-1
Ex-UE-9
(*baxie*)
L.I.-2
L.I.-3
T.B.-2
Ex-UE-8
L.I.-2
Ex-UE-9
(*baxie*)
L.I.-4
T.B.-3
L.I.-3
Ex-UE-7
Ex-UE-7
L.I.4
L.I.-5
L.I.-5
T.B.-4
L.I.-6
T.B.-5

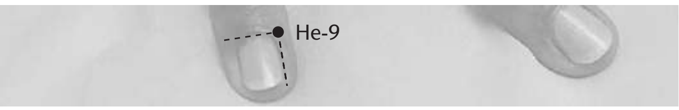

Location

On the little finger, 0.1 cun from the ulnar corner of the nail.

How to find

This point is located at the junction of two tangents drawn along the lateral and proximal borders of the nail, approximately 0.1 cun from the actual corner of the nail. → **HE-9** is located at the radial corner of the nail.

Needling

0.1–0.2 cun vertically or obliquely in a proximal direction or prick to bleed with a needle, lancet or three-edged needle. Caution: Painful point.

Actions/Indications

- Expels Wind and Heat
- Opens the orifices
- Revives consciousness
- Opens the channel
- Benefits the breasts

Special features

Jing-well point, Metal point, entry point.

Ch04.6-F10028.qxd 2/22/08 9:47 PM Page 224

S.I.-2 *Front Valley* QIANGU

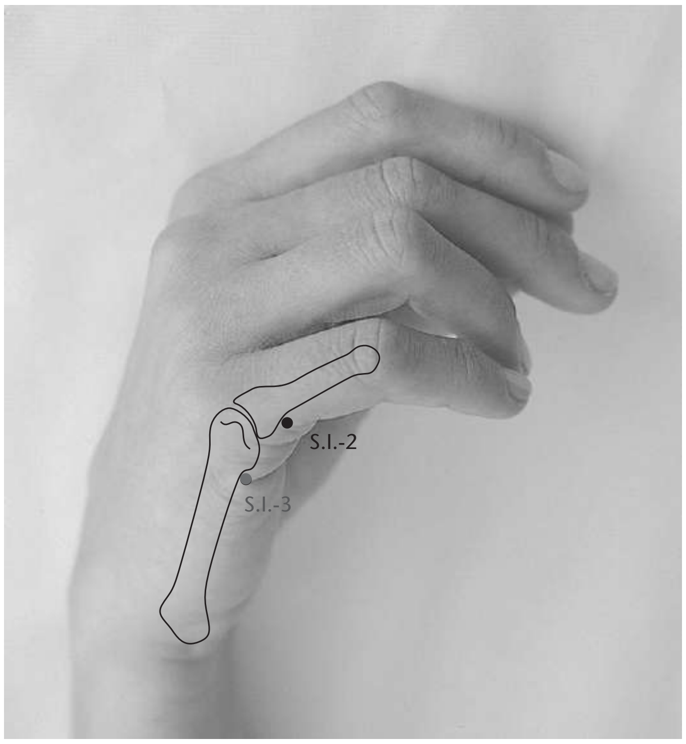

Location

On the ulnar aspect of the little finger, distal to the metacarpophalangeal joint, at the junction of the shaft and the base of the proximal phalanx.

How to find

On the ulnar aspect of the little finger, palpate along the border of the 'red and white' skin in the direction of the metacarpophalangeal joint, until you can clearly feel the junction between the shaft and the base. **S.I.-2** is located at the junction of the shaft and the base and slightly inferior (palmar) to the exterior curvature of the bone. → **L.I.-2** is located in a comparable position on the proximal phalanx of the index finger. In a comparable location on the foot are → **BL-66** (on the lateral aspect) and → **SP-2** (on the medial aspect).

Needling

0.2–0.5 cun obliquely in a proximal or distal direction and slightly towards the palm. Caution: Painful point.

Actions/Indications

- Drains Heat, expels Wind
- Disperses accumulations

Special features

Ying-spring point, Water point.

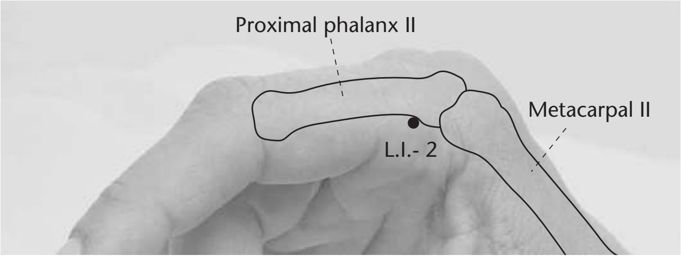
Proximal phalanx II
Metacarpal II
L.I.- 2

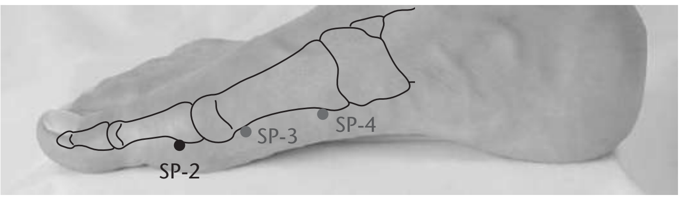

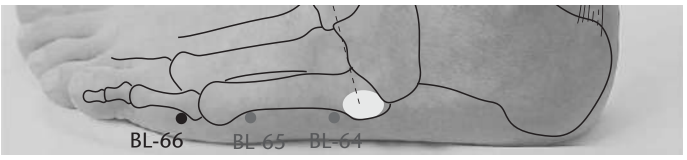
BL-66
BL-65
BL-64

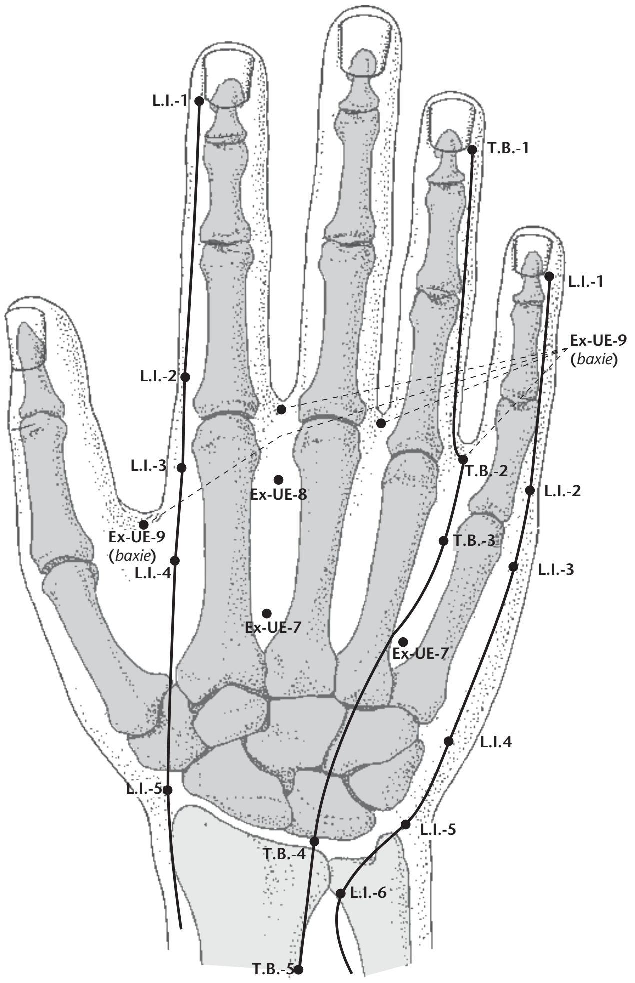
L.I.-1
T.B.-1
L.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
T.B.-2
L.I.-2
Ex-UE-8
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
L.I.-3
Ex-UE-7
Ex-UE-7
L.I.4
L.I.-5
L.I.-5
T.B.-4
L.I.-6
T.B.-5

Back Stream HOUXI S.I.-3

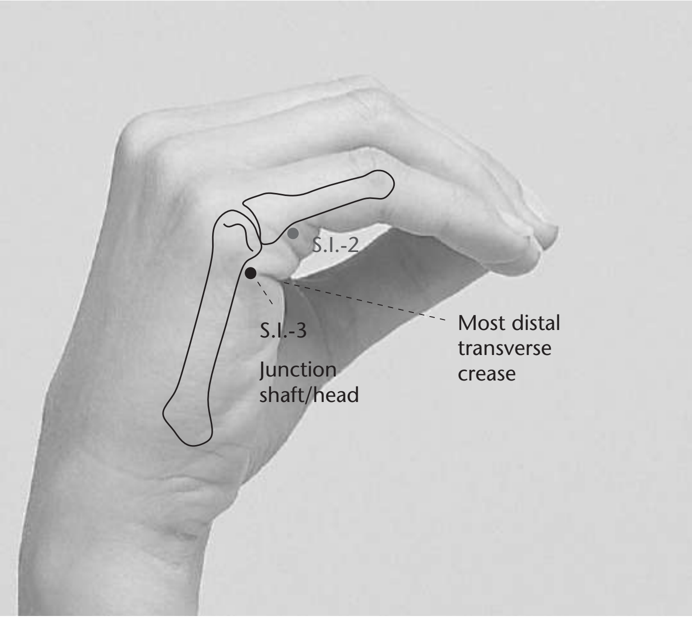

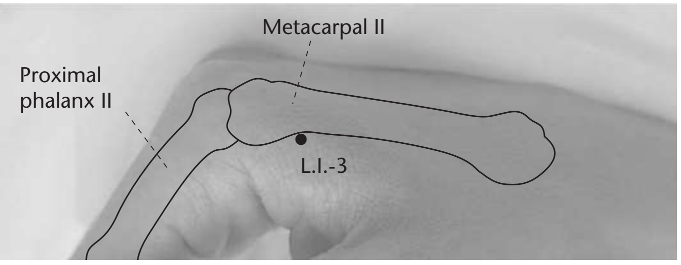

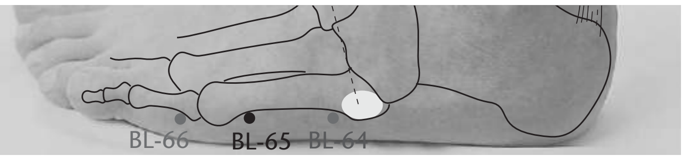

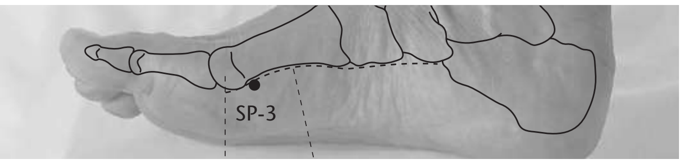

Location

On the ulnar border of the hand, in the depression proximal to the metacarpophalangeal joint, at the border of the ‘red and white’ skin.

How to find

Quick method: Slide from proximal (wrist) to distal (little finger) along the ulnar border of the hand, until the palpating finger comes to rest at the junction between the base and the head of the 5th metacarpal bone proximal to the metacarpophalangeal joint. Here, locate **S.I.-3. Or:** Locate, with the patient making a loose fist. Locate **S.I.-3** at the end of the most distal transverse crease (this often originates between the index and middle fingers and runs to the little finger). When making a fist, a small bulge forms at the end of the crease. There, locate **S.I.-3** in a palpable depression slightly proximal and dorsal to the bulge.
→ **L.I.-3** is located in a comparable position on the radial border of the hand. Located in a comparable position on the foot are → **SP-3** (on the medial aspect) and → **BL-65** (on the lateral aspect).

Needling

0.5–1 cun horizontally below the bone – that is, vertically to the skin towards the palm – or 0.2–0.3 cun obliquely in a distal direction. Caution: Painful point.

Actions/Indications

- Eliminates Wind and Heat from the *taiyang* (S.I., BL)
- Opens the channel and the *luo* vessels
- Clears Heat and benefits the orifices
- Regulates the *du mai*, calms the *shen*

Special features

Shu-stream point, Wood point, tonification point, opening (master) point of the *du mai*. Important distal point for pain, stiffness and contractures along the course of the channel and for disorders of the cervical spine.

S.I.-4 *Wrist Bone* WANGU

Location

On the ulnar border of the hand, between the 5th metacarpal bone and the carpal bones, on the border of the red and white skin.

How to find

On the ulnar aspect of the hand, palpate along the shaft of the 5th metacarpal bone in a proximal direction, until the finger glides over the curvature of the base and into the cleft between the metacarpal and carpal bones.

→ **S.I.-5** is located on the same line, proximal to the pisiform bone and at the level of the wrist crease.

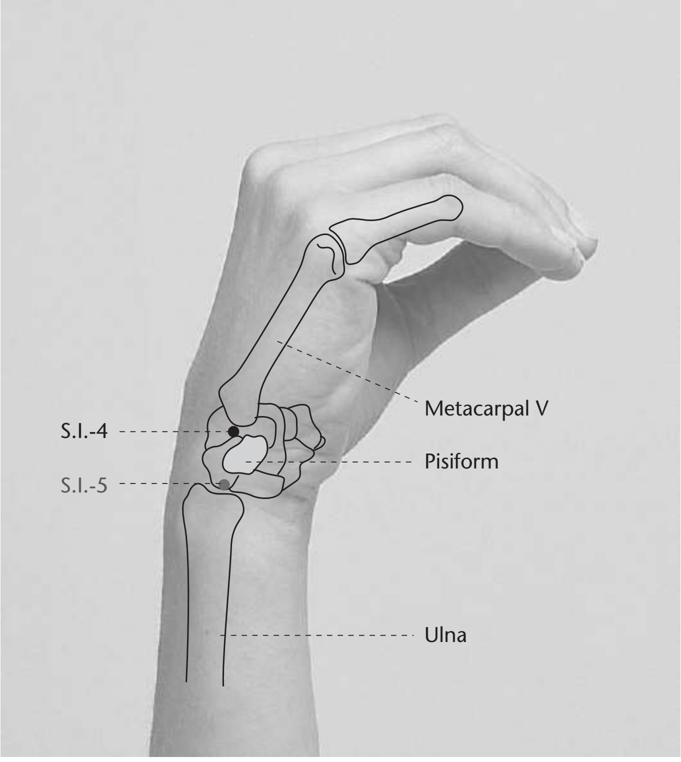
S.I.-4
S.I.-5
Metacarpal V
Pisiform
Ulna

Needling

Vertically 0.3–0.5 cun. Caution: Painful point.

Actions/Indications

- Clears pathogenic factors from the *taiyang* channel
- Clears Damp-Heat in the Small Intestine

Special features

Yuan-source point

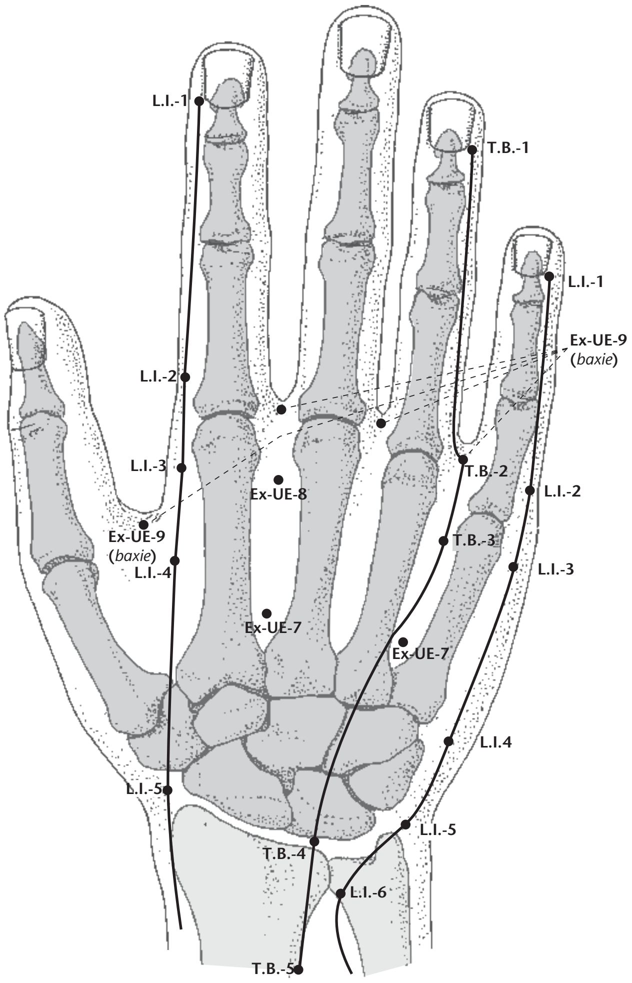
L.I.-1
T.B.-1
L.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
L.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
L.I.-3
Ex-UE-7
Ex-UE-7
L.I.4
L.I.-5
L.I.-5
T.B.-4
L.I.-6
T.B.-5

Yang Valley YANGGU S.I.-5

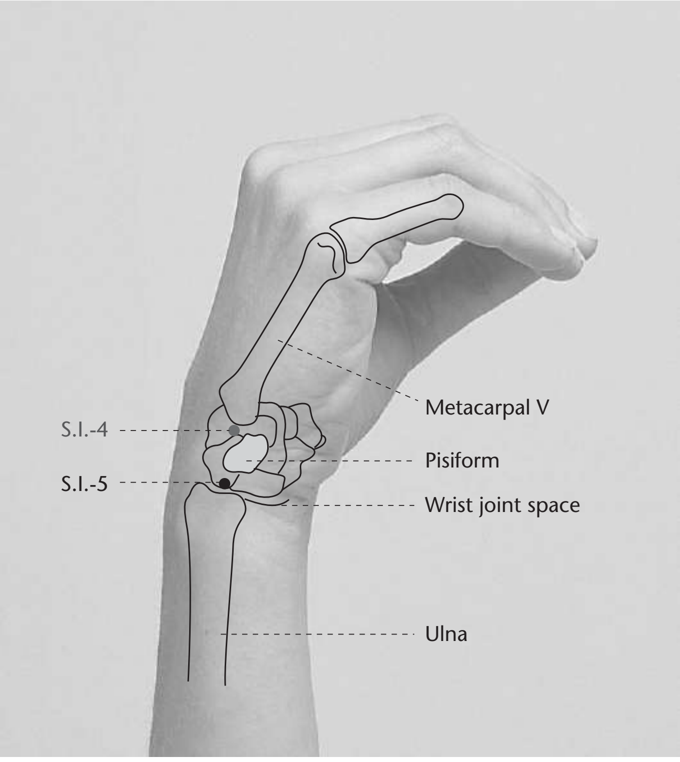
S.I.-4
S.I.-5
Metacarpal V
Pisiform
Wrist joint space
Ulna

Location

On the ulnar aspect of the wrist, at the level of the lateral joint space.

How to find

By moving the wrist in a relaxed way, the wrist joint space can be palpated on the ulnar side. **S.I.-5** is located directly distal to the styloid process of the ulna (➞ 3.3.3), on a line running along the ulnar border of the hand and continuing to the wrist. Located on the same line and distal to the prominence of the pisiform bone is ➞ **S.I.-4**.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Clears Heat and reduces swellings

Special features

Jing-river point, Fire point, *ben* point (Five Phases point).

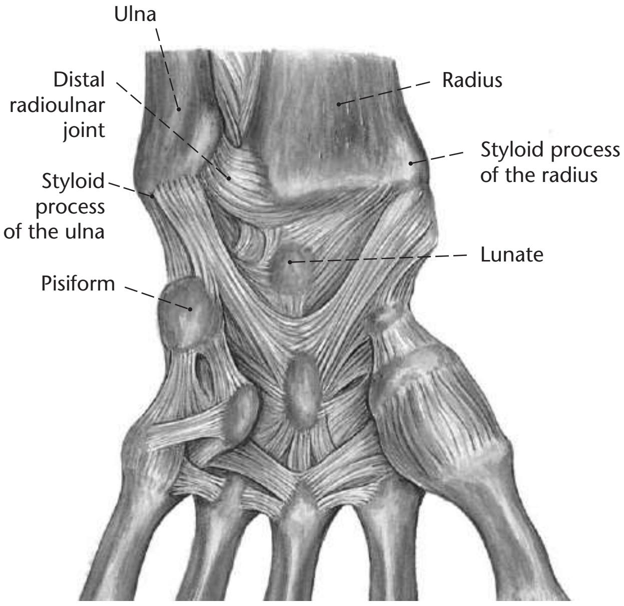
Ulna
Distal radioulnar joint
Styloid process of the ulna
Pisiform
Radius
Styloid process of the radius
Lunate

S.I.-6 *Support the Aged* YANGLAO

Location

In the depression radial and proximal to the styloid process of the ulna, which forms when the hand is turned from a pronated to a supinated position.

How to find

This point is best located with the elbow slightly flexed. Place the palpating finger on the distal portion of the styloid process of the ulna (→ 3.3.3). When turning the hand from a pronated to a supinated position (mnemonic: cut **br**ead and eat **soup**), the palpating finger will glide into a bony cleft at the proximal aspect of the styloid process (the tendon of the extensor carpi ulnari muscle runs along this groove). This cleft also becomes palpable when the patient rests the hand on the chest (half-supinated position). Locate **S.I.-6** in this bony groove.

Needling

0.3–0.8 cun vertically or slightly obliquely in a proximal direction towards → **P-6** or along the tendon towards the elbow joint.

Actions/Indications

- Opens the channel, alleviates pain, benefits the shoulder and arm, alleviates acute conditions
- Benefits the eyes

Special features

Xi-cleft point. Important distal point for pain and limited range of motion of the occiput and shoulder (for example, acute torticollis) and the lumbar region (for example, acute sciatica). For acute and painful impairment of motion, use strong reducing needling technique while the patient gently moves the affected region.

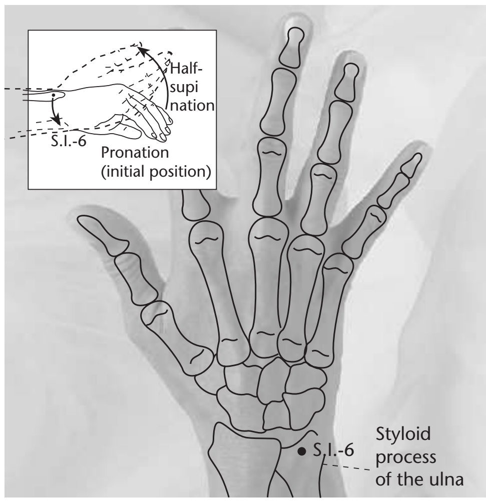
Half-
supi
nation
S.I.-6
Pronation
(initial position)
S.I.-6
Styloid
process
of the ulna

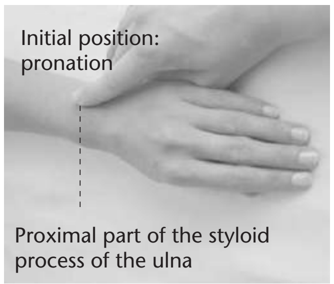
Initial position:
pronation
Proximal part of the styloid
process of the ulna

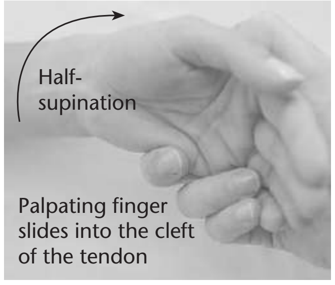
Half-
supination
Palpating finger
slides into the cleft
of the tendon

Branch of the Upright ZHIZHENG S.I.-7

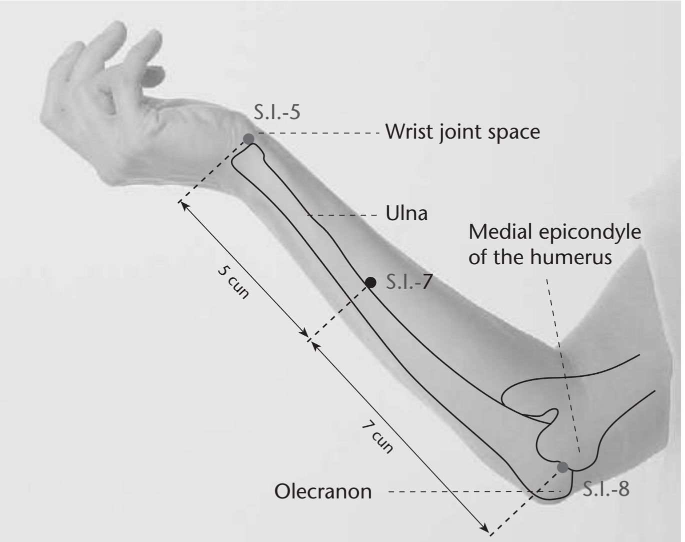
S.I.-5
Wrist joint space
Ulna
Medial epicondyle of the humerus
S.I.-7
5 cun
7 cun
Olecranon
S.I.-8

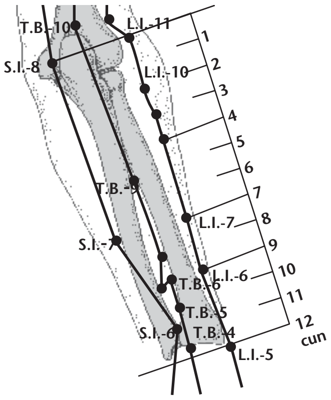
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
T.B.-6
L.I.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
1
2
3
4
5
6
7
8
9
10
11
12
cun

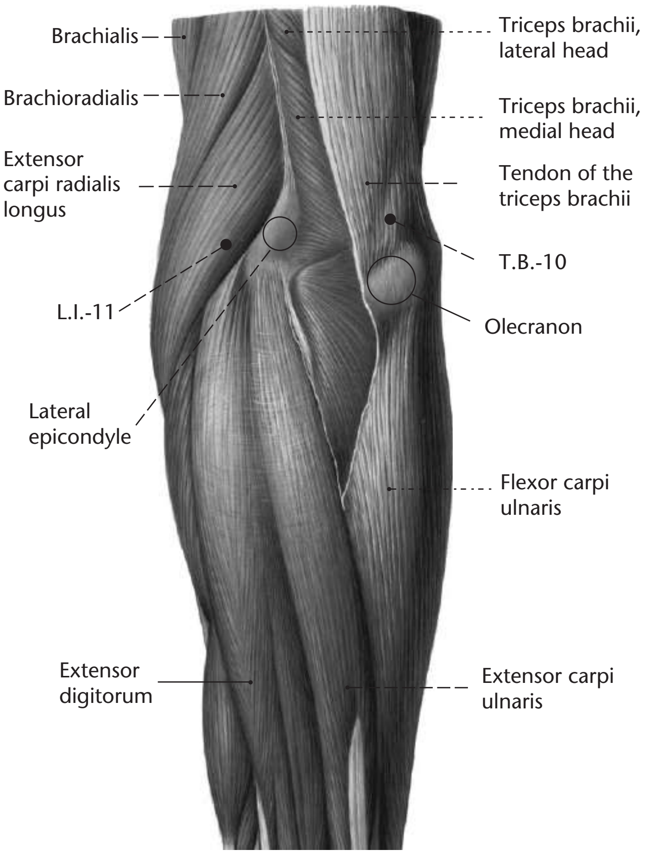
Brachialis
Triceps brachii, lateral head
Brachioradialis
Triceps brachii, medial head
Extensor carpi radialis longus
Tendon of the triceps brachii
T.B.-10
L.I.-11
Olecranon
Lateral epicondyle
Flexor carpi ulnaris
Extensor digitorum
Extensor carpi ulnaris

Location

5 cun proximal to the wrist joint, on the line connecting the ulnar wrist joint space with the ulnar sulcus of the elbow (➞ **S.I.-5** to ➞ **S.I.-8**) or 1 cun distal to the midpoint of this line.

How to find

Find the midpoint of the distance between ➞ **S.I.-5** and ➞ **S.I.-8** by using the spreading hands technique (➞ 2.3.3): Place the little fingers on ➞ **S.I.-5** and ➞ **S.I.-8** respectively and span the hands evenly, with the thumbs joining at the midpoint. From the midpoint, measure 1 cun in a distal direction. The point lies between the palpable border of the ulna and the more anteriorly located flexor carpi ulnaris muscle.

Needling

Vertically 0.5–1 cun

Actions/Indications

- Opens the channel
- Releases the Exterior
- Calms the *shen*

Special features

Luo-connecting point

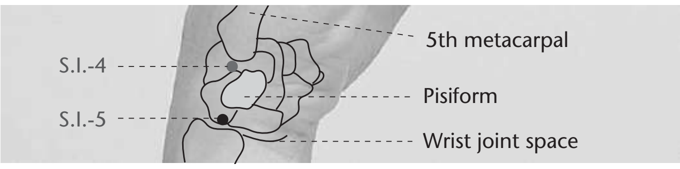
5th metacarpal
S.I.-4
Pisiform
S.I.-5
Wrist joint space

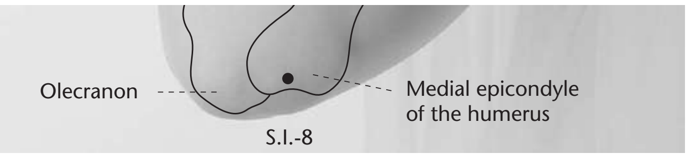
Olecranon
Medial epicondyle of the humerus
S.I.-8

S.I.-8 *Small Sea* XIAOHAI

Location

With the elbow flexed, in the depression between the olecranon process of the ulna and the medial epicondyle of the humerus.

How to find

S.I.-8 is located on a line connecting the tip of the olecranon and the tip of the medial epicondyle, at the deepest point of the groove between those two bony protrusions.

Needling

0.3–0.5 cun vertically or obliquely with or against the flow of the channel. Caution: The ulnar nerve is located in this groove.

Actions/Indications

- Opens the channel

Special features

He-sea point, Earth point, sedation point.

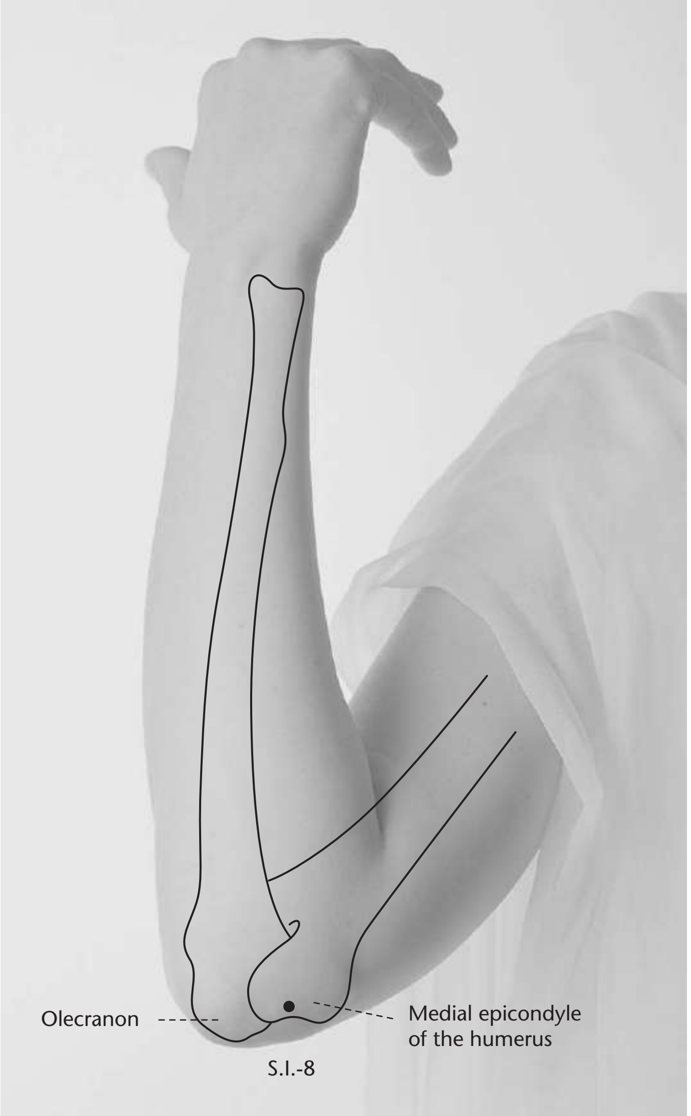
Olecranon
Medial epicondyle of the humerus
S.I.-8

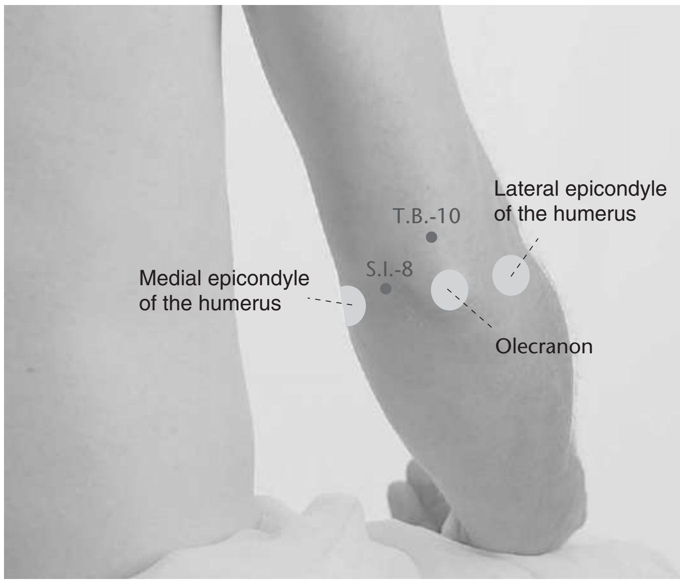
T.B.-10
Lateral epicondyle of the humerus
Medial epicondyle of the humerus
S.I.-8
Olecranon

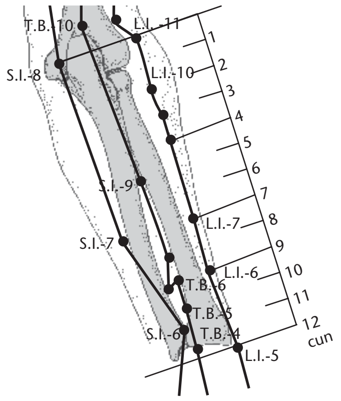
T.B.-10
L.I.-11
S.I.-8
L.I.-10
S.I.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
1
2
3
4
5
6
7
8
9
10
11
12
cun

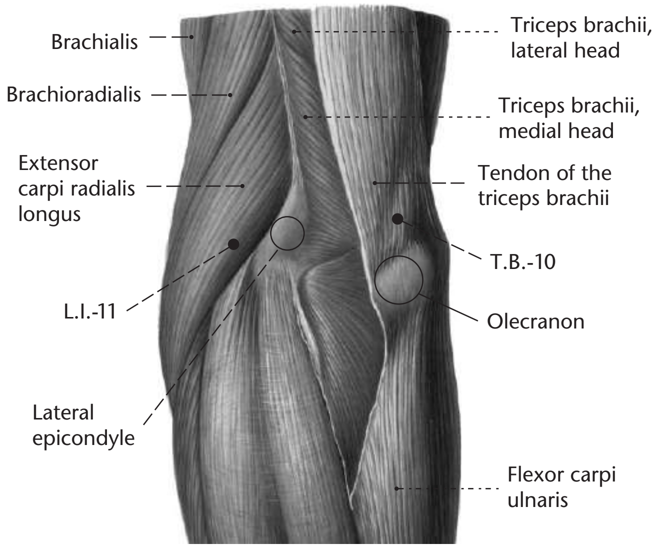
Brachialis
Brachioradialis
Extensor carpi radialis longus
L.I.-11
Lateral epicondyle
Triceps brachii, lateral head
Triceps brachii, medial head
Tendon of the triceps brachii
T.B.-10
Olecranon
Flexor carpi ulnaris

Ch04.6-F10028.qxd 2/22/08 9:48 PM Page 231

True Shoulder JIANZHEN S.I.-9

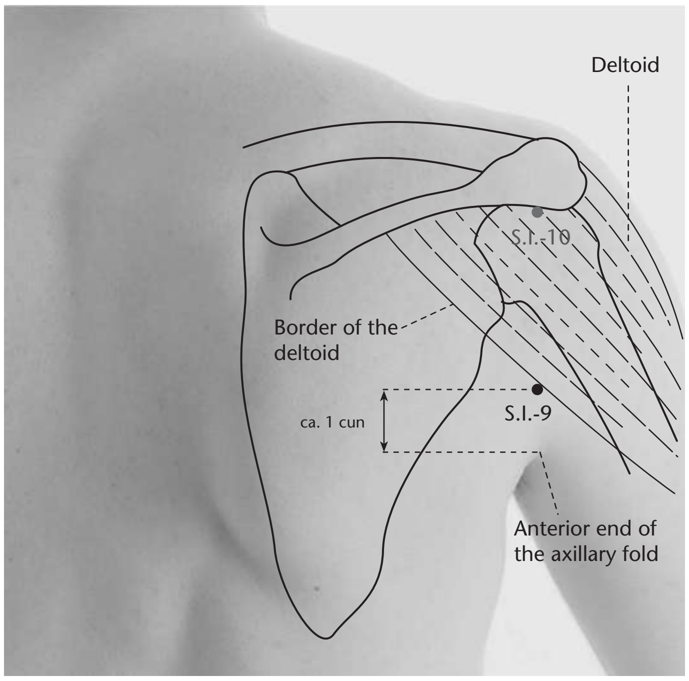
Deltoid
S.I.-10
Border of the deltoid
ca. 1 cun
S.I.-9
Anterior end of the axillary fold

Location

With the arm adducted, 1 cun superior to the posterior axillary fold, on the lower border of the deltoid muscle.

How to find

With the patient sitting upright, palpate upward from the end of the posterior axillary fold, until you can feel the lower border of the deltoid muscle. If in doubt, ask the patient to flex their muscle. **S.I.-9** is located on its lower border.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Opens the channel, expels Wind and benefits the shoulder

Special features

S.I.-9 corresponds to a trigger point that is often sensitive to pressure.

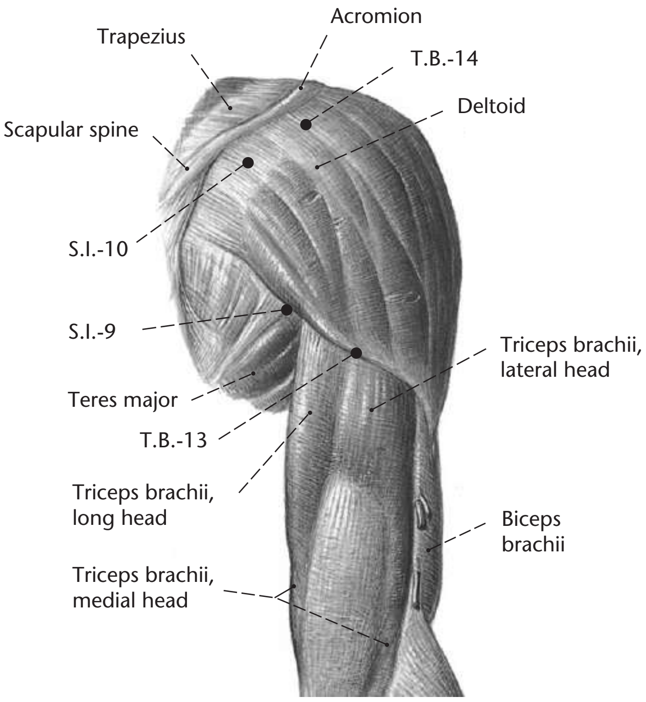
Acromion
Trapezius
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

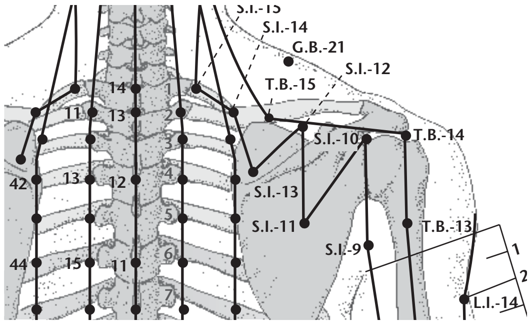
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
1
11
13
2
S.I.-10
T.B.-14
3
42
13
12
4
S.I.-13
5
S.I.-11
T.B.-13
S.I.-9
44
15
11
6
7
1
2
L.I.-14

S.I.-10 *Upper Arm* Shu NAOSHU

Location

With the arm adducted, on an imaginary line extending in a superior direction from the posterior axillary fold, on the lower border of the scapular spine.

How to find

With the patient sitting upright, palpate from the posterior axillary fold upward, past the lower border of the deltoid muscle (→ **S.I.-9**) to the bony resistance of the scapular spine and acromion, which here form a curve that opens towards the inferior. **S.I.-10** is located below its bony border.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Opens the channel and relaxes the sinews

Special features

Meeting point of the *yang wei mai* and the *yang qiao mai*. S.I.10 corresponds to a trigger point that is often sensitive to pressure. It also corresponds to the posterior access for puncturing the shoulder joint.

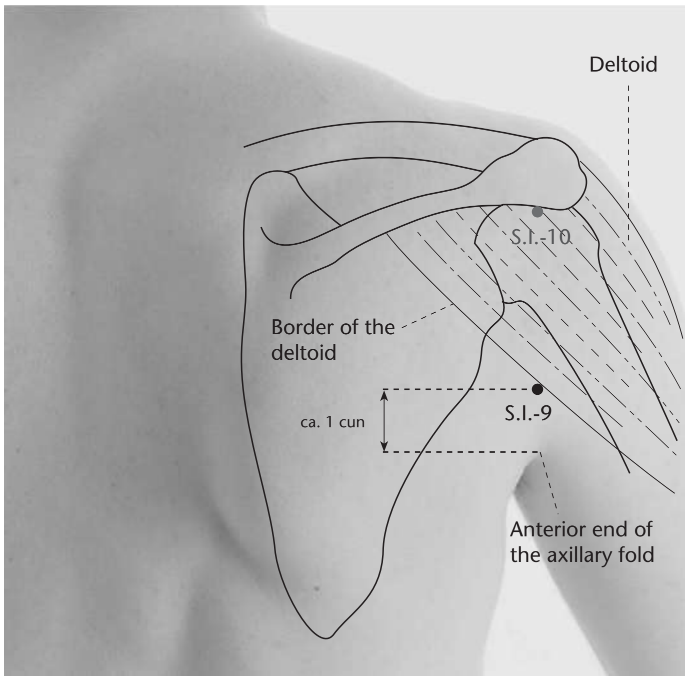
Deltoid
S.I.-10
Border of the deltoid
ca. 1 cun
S.I.-9
Anterior end of the axillary fold

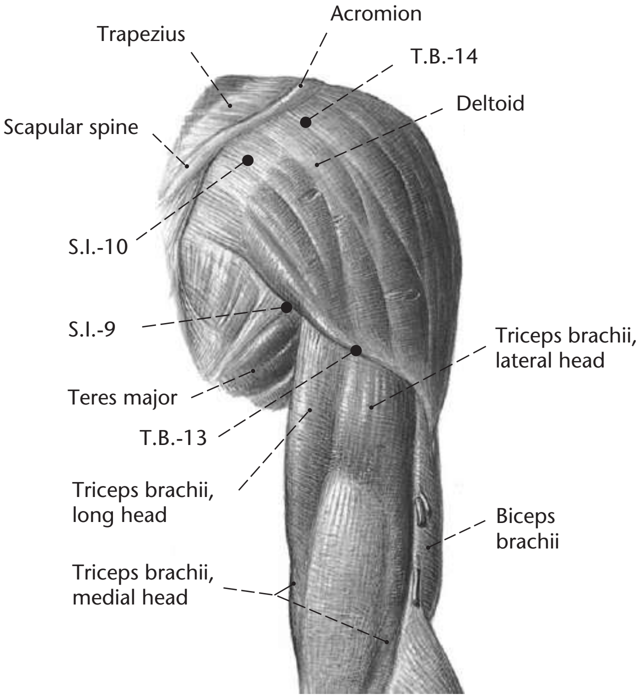
Trapezius
Acromion
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

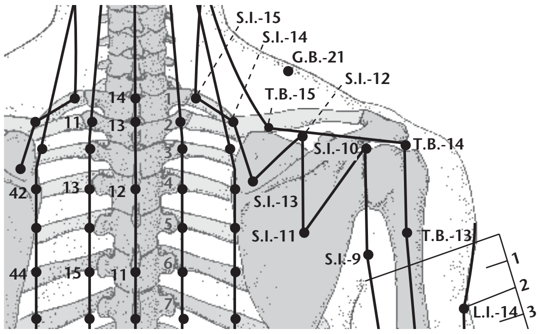
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
1
11
13
2
3
T.B.-14
S.I.-10
42
13
12
4
S.I.-13
5
S.I.-11
T.B.-13
S.I.-9
1
44
15
11
6
2
7
L.I.-14
3

Heavenly Gathering TIANZONG S.I.-11

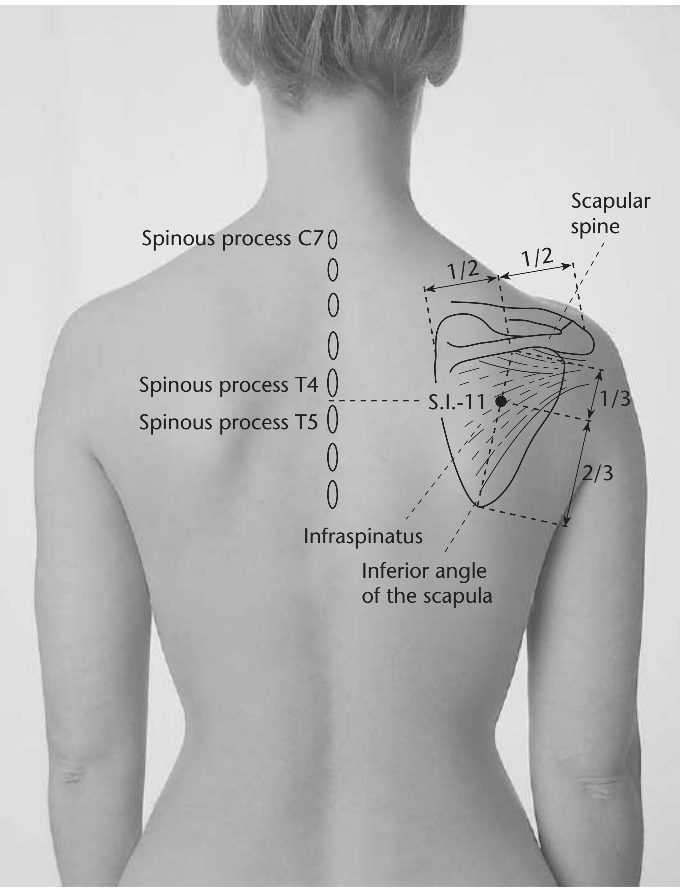
Scapular spine
Spinous process C7
1/2
1/2
Spinous process T4
Spinous process T5
S.I.-11
1/3
2/3
Infraspinatus
Inferior angle of the scapula

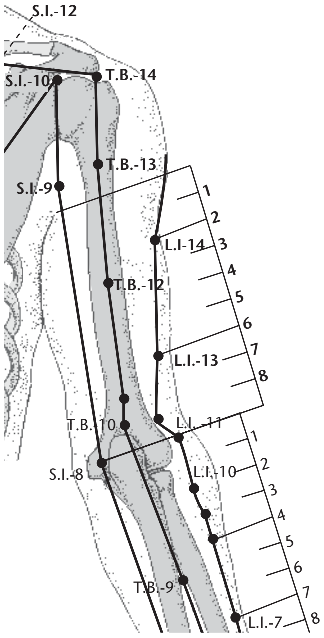
S.I.-12
S.I.-10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I-14
3
4
T.B.-12
5
6
7
L.I.-13
8
T.B.-10
L.I.-11
1
S.I.-8
L.I.-10
2
3
4
5
6
7
T.B.-9
L.I.-7
8

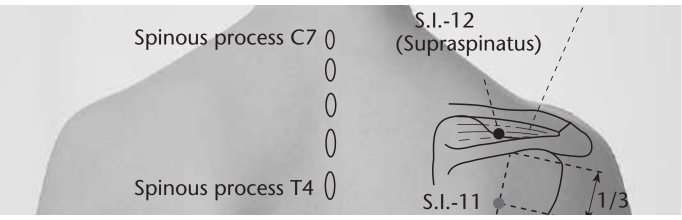
Spinous process C7
S.I.-12
(Supraspinatus)
Spinous process T4
S.I.-11
1/3

Location

On the scapula, in a depression on the infraspinatus muscle, one third of the distance from the midpoint of the scapular spine and the inferior angle of the scapula.

How to find

Patient's position: Prone position or better seated with relaxed shoulders. Draw an imaginary line from the midpoint of the scapular spine (→ 3.3.1) to the inferior angle of the scapula. **S.I.-11** is located at the junction of the upper and middle third of this line, in a depression on the infraspinatus muscle, which is often sensitive to pressure. In a seated or standing position with the arms hanging down, it will generally be level with the spinous process of the 4th thoracic vertebra or the body of the 5th thoracic vertebra.

→ **S.I.-12** is located directly above **S.I.-11**, in the centre of the supraspinous fossa.

Needling

0.5–1.5 cun vertically or obliquely into the muscle tissue

Actions/Indications

- Opens the channel, alleviates pain, moves Qi locally, unbinds the chest and lateral costal region
- Benefits the breasts

Special features

S.I.-11 is often a trigger point on the infraspinatus muscle.

S.I.-12 *Grasping the Wind* BINGFENG

Location

Directly above → **S.I.-11**, in the centre of the supraspinous fossa.

How to find

Patient's position: Prone position or better seated with relaxed shoulders. First, locate → **S.I.-11** (one third of the distance between the midpoint of the scapular spine and the inferior angle of the scapula). **S.I.-12** is located directly above → **S.I.-11** in the centre of the supraspinous fossa.

Needling

0.5–1 cun vertically or obliquely into the supraspinatus muscle in the direction of the shoulder. Caution: Pneumothorax.

Actions/Indications

- Eliminates Wind, benefits the shoulder and scapula, opens the channel and the *luo* vessels

Special features

Meeting point with the Gall Bladder, Triple Burner and Large Intestine channels. **S.I.-12** is often a trigger point in the supraspinatus muscle.

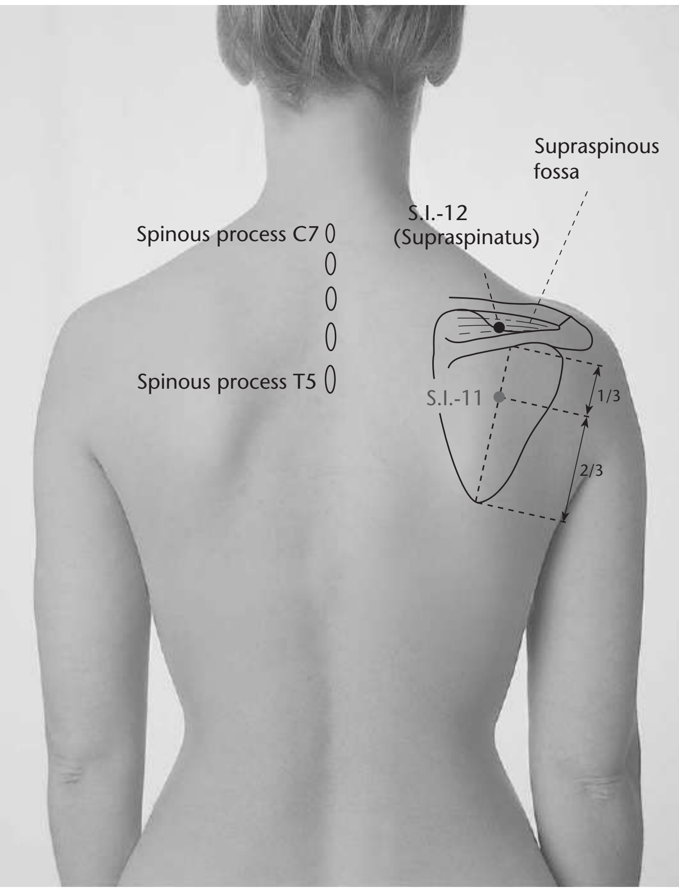
Supraspinous fossa
S.I.-12 (Supraspinatus)
Spinous process C7
Spinous process T5
S.I.-11
1/3
2/3

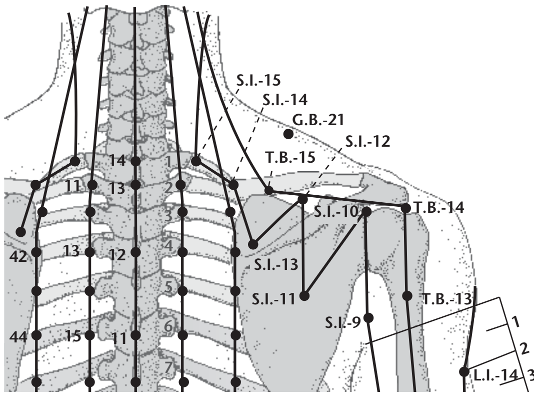
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
1
11
13
2
3
S.I.-10
T.B.-14
42
13
12
4
S.I.-13
5
S.I.-11
T.B.-13
S.I.-9
1
44
15
11
6
2
7
L.I.-14
3

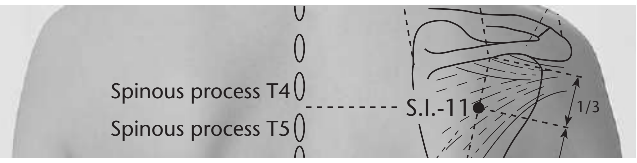
Spinous process T4
Spinous process T5
S.I.-11
1/3

Crooked Wall QUYUAN S.I.-13

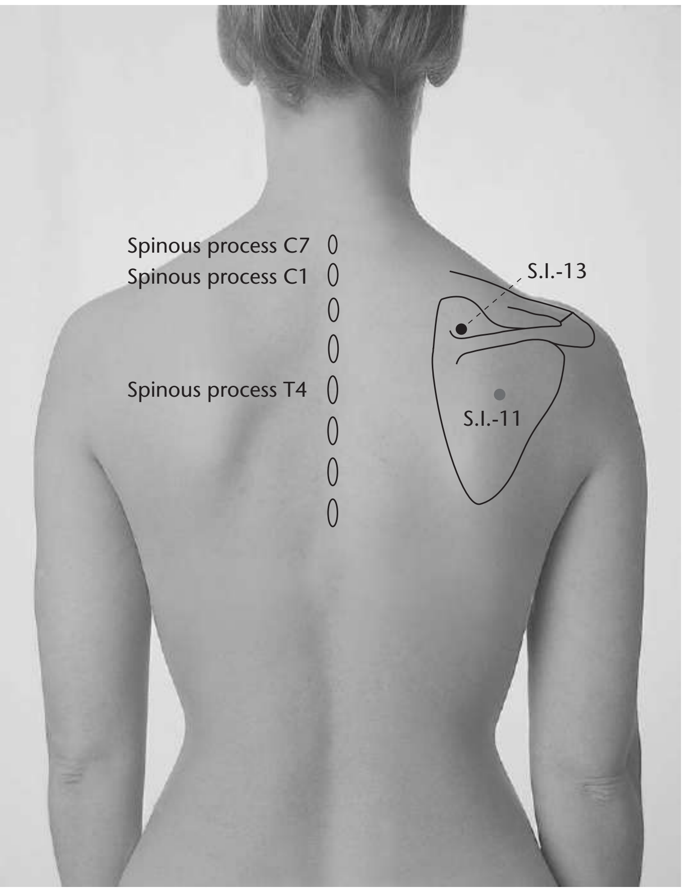
Spinous process C7
Spinous process C1
Spinous process T4
S.I.-13
S.I.-11

Location

At the medial end of the supraspinous fossa.

How to find

At its medial end, the scapular spine fans out in a deltoid shape, curving round in a superior direction. When palpating, this feels like a ‘crooked wall’, hence its name. **S.I.-13** is located on the scapula, in the slightly concave bend, at the medial origin of the supraspinatus muscle.

Needling

0.5–1 cun vertically or obliquely in a lateral direction. No danger of puncturing the pleura if located correctly.

Actions/Indications

- Opens the channel

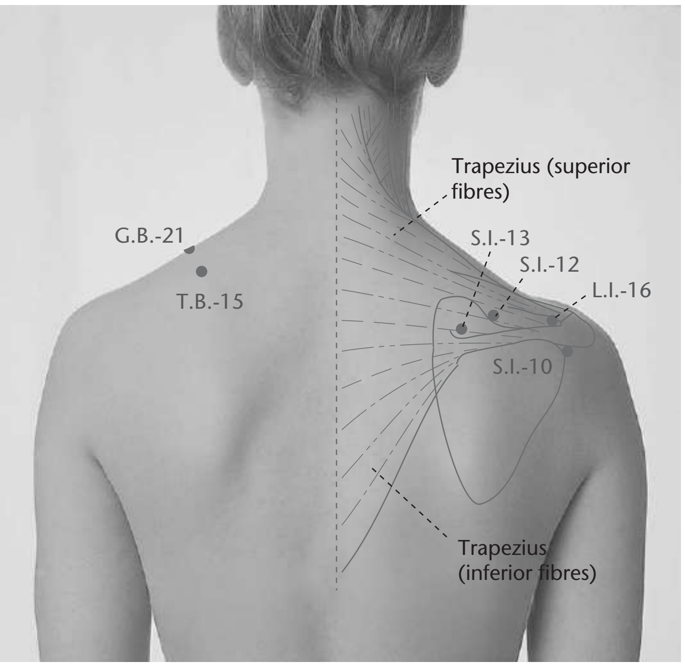
Trapezius (superior fibres)
G.B.-21
S.I.-13
S.I.-12
L.I.-16
T.B.-15
S.I.-10
Trapezius (inferior fibres)

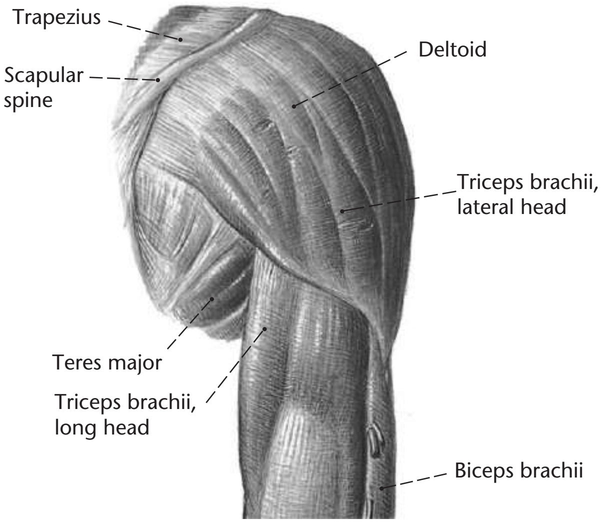
Trapezius
Scapular spine
Deltoid
Triceps brachii, lateral head
Teres major
Triceps brachii, long head
Biceps brachii

S.I.-14 *Outer Shoulder Shu* JIANWAISHU

Location

3 cun lateral to the lower border of the spinous process of the 1st thoracic vertebra, at the insertion of the levator scapulae muscle.

How to find

The levator scapulae muscle originates on the upper aspect of the medial border and the superior angle of the scapula. In this area, it is often myogelotic and therefore well defined and often sensitive to pressure (trigger point). **S.I.-14** is located just superior to the actual muscle insertion; in contrast to → **S.I.-13**, it is not located on the scapula.

Located on approximately the same level are → **Du-13** (below the spinous process of the 1st thoracic vertebra) and → **BL-11** (1.5 cun distal to the posterior midline).

Needling

0.5–1 cun vertically or obliquely in a medial direction. Caution: Pneumothorax.

Actions/Indications

- Opens the channel
- Expels Wind-Cold

Special features

Trigger point on the levator scapulae muscle

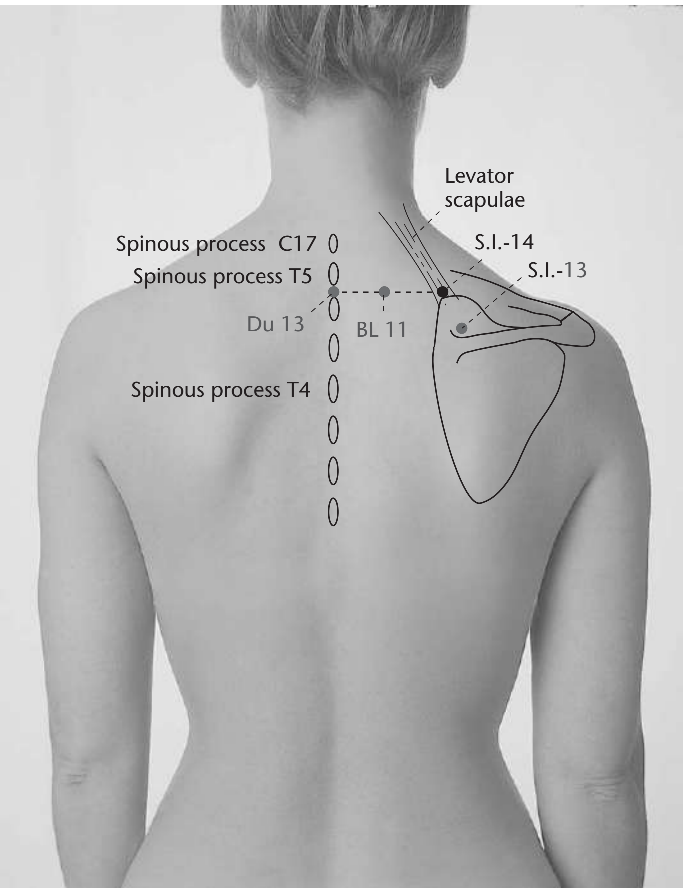
Levator scapulae
S.I.-14
S.I.-13
Spinous process C17
Spinous process T5
Du 13
BL 11
Spinous process T4

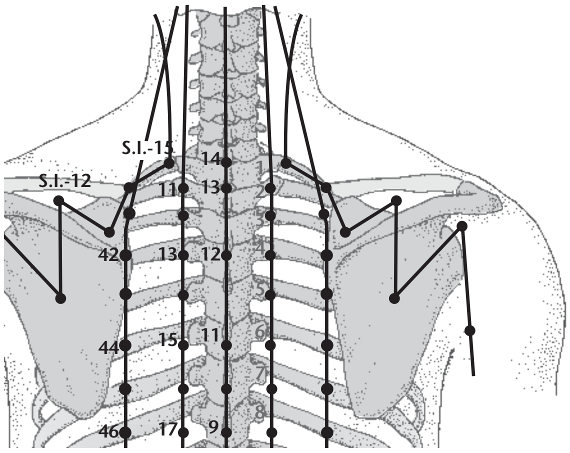
S.I.-15
S.I.-12
14
11
13
42
13
12
44
15
11
46
17
9
1
2
3
4
5
6
7
8

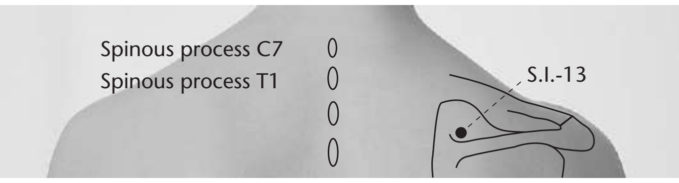
Spinous process C7
Spinous process T1
S.I.-13

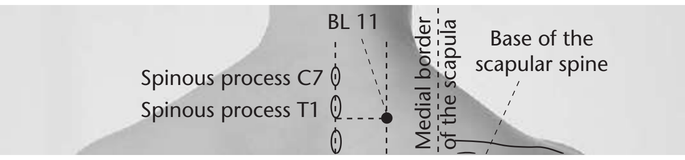
BL 11
Spinous process C7
Spinous process T1
Medial border of the scapula
Base of the scapular spine

Middle Shoulder Shu JIANZHONGSHU S.I.-15

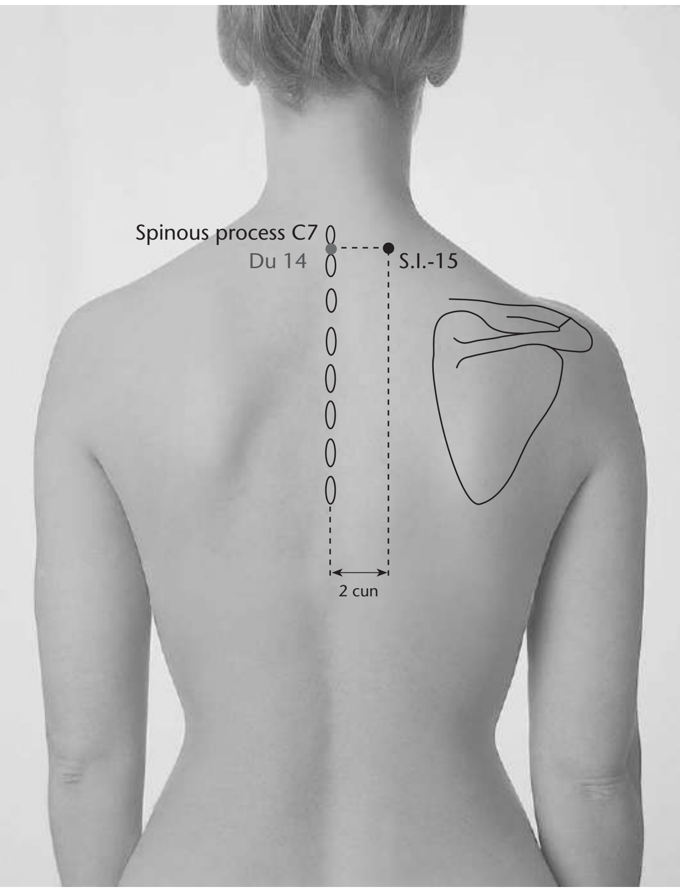
Spinous process C7
Du 14
S.I.-15
2 cun

Location

2 cun lateral to the lower border of the spinous process of the 7th thoracic vertebra.

How to find

For locating the 7th cervical vertebra, see ➞ 3.4.
Located on the same level is ➞ **Du-14** below the spinous process of the 7th cervical vertebra. Locate **S.I.-15** by measuring 2 cun in a lateral direction.

Needling

0.5–1 cun vertically or obliquely in a medial direction. Caution: Pneumothorax.

Actions/Indications

- Opens the channel

Special features

Trigger point on the levator scapulae muscle, effective point for disorders of the cervicodorsal area.

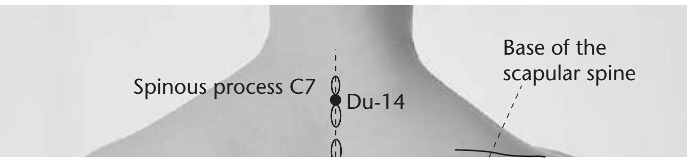
Spinous process C7
Du-14
Base of the scapular spine

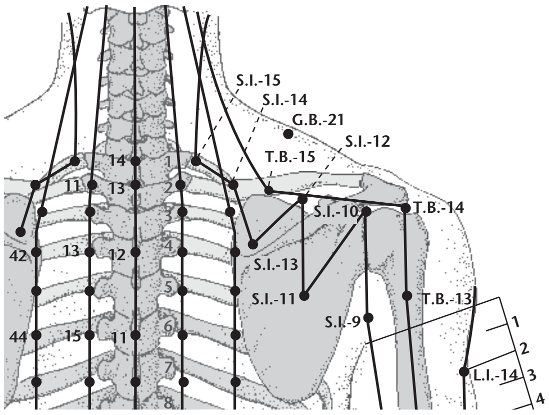
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
11
13
S.I.-10
T.B.-14
42
13
12
S.I.-13
S.I.-11
T.B.-13
1
S.I.-9
2
44
15
11
L.I.-14
3
4

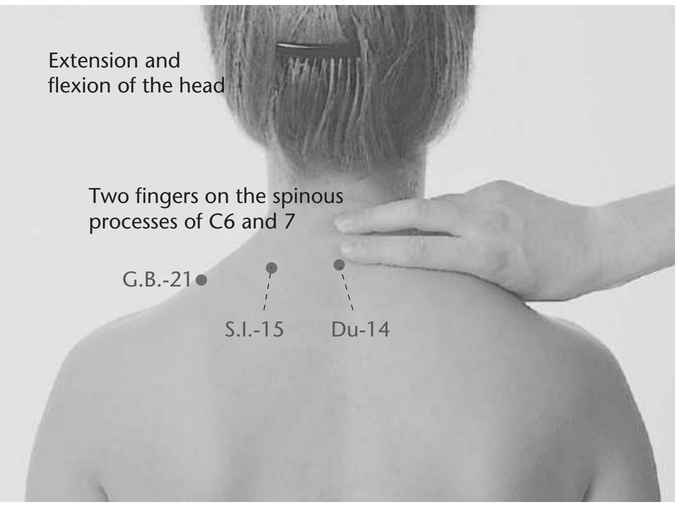
Extension and flexion of the head
Two fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

S.I.-16 *Heavenly Window* TIANCHUANG

Location

Approximately 3.5 cun lateral to the anterior midline, at the level of the laryngeal prominence, on the posterior border of the sternocleidomastoid muscle.

How to find

From the laryngeal prominence (→ 3.2), draw a line to the posterior border of the sternocleidomastoid muscle. **S.I.-16** is located directly at its posterior border, which can be made more clearly visible and palpable by rotating the patient's head in the opposite direction.

Located on the same level are → **ST-9** (on the anterior border of the sternocleidomastoid muscle) and → **L.I.-18** (between the two heads of the SCM muscle).

Needling

Vertically 0.5–1 cun. If necessary, rotate the patient's head back to its normal position before needle insertion.

Actions/Indications

- Expels Wind and Heat, benefits the throat, ears and voice
- Opens the channel

Special features

Window of Heaven point

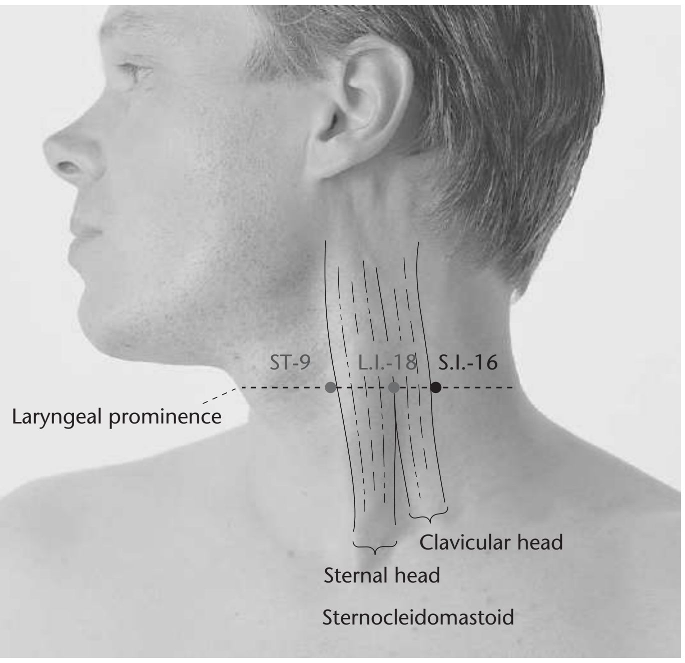
ST-9
L.I.-18
S.I.-16
Laryngeal prominence
Clavicular head
Sternal head
Sternocleidomastoid

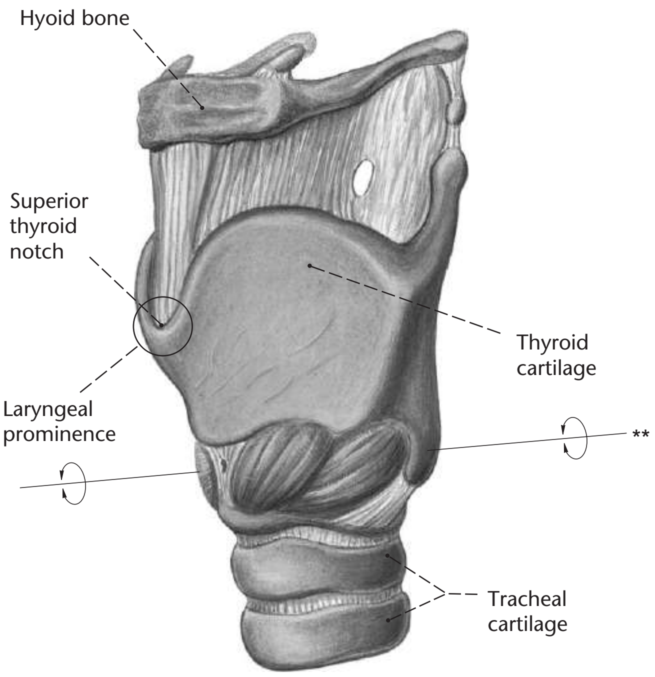
Hyoid bone
Superior thyroid notch
Thyroid cartilage
Laryngeal prominence
**
Tracheal cartilage

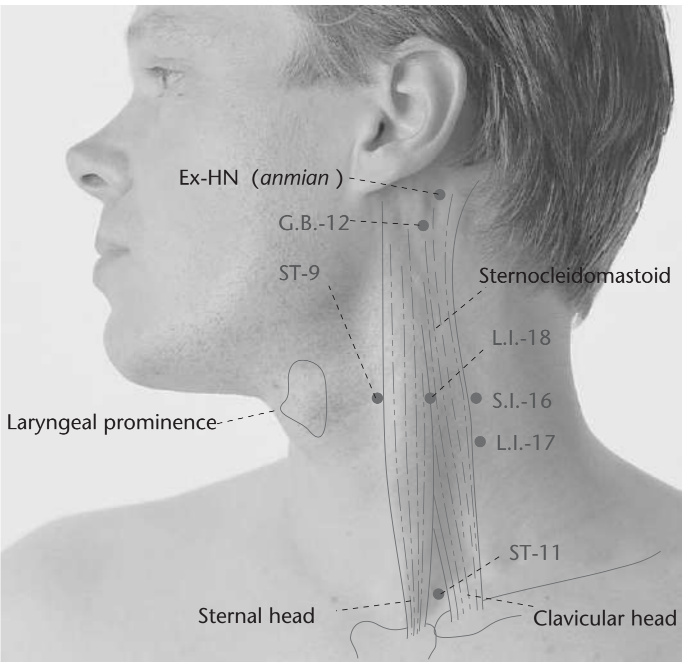
Ex-HN (*anmian*)
G.B.-12
ST-9
Sternocleidomastoid
L.I.-18
S.I.-16
L.I.-17
Laryngeal prominence
ST-11
Sternal head
Clavicular head

Heavenly Appearance TIANRONG S.I.-17

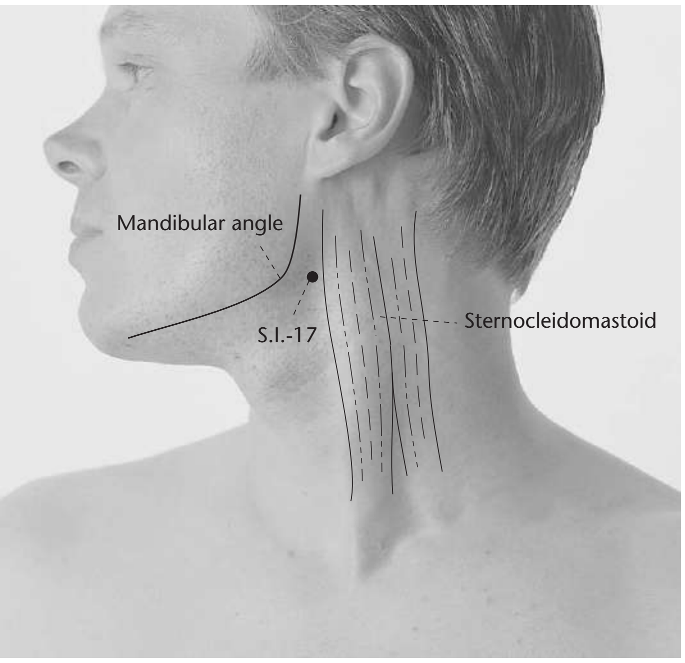
Mandibular angle
S.I.-17
Sternocleidomastoid

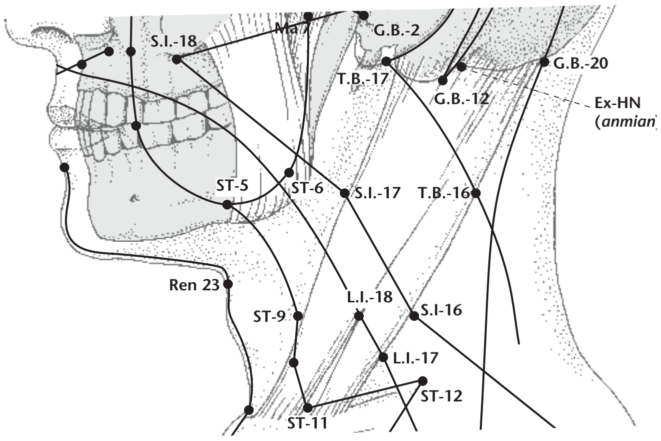
S.I.-18
Ma 7
G.B.-2
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren 23
L.I.-18
ST-9
S.I-16
L.I.-17
ST-12
ST-11

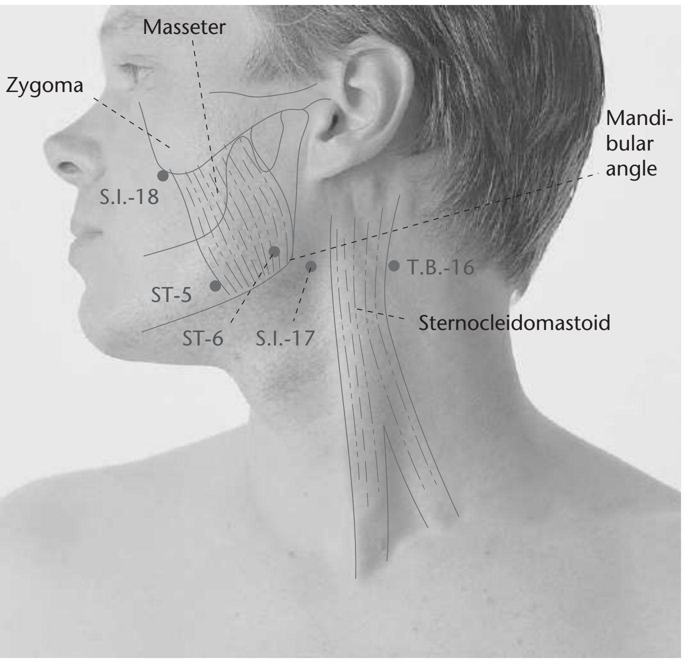
Masseter
Zygoma
Mandi-
bular
angle
S.I.-18
T.B.-16
ST-5
Sternocleidomastoid
ST-6
S.I.-17

Location

Posterior to the mandibular angle, on the anterior border of the sternocleidomastoid muscle.

How to find

By rotating the patient's head in the opposite direction, the border of the sternocleidomastoid muscle becomes more pronounced. **S.I.-17** is located halfway between the angle of the mandible and the anterior border of the sternocleidomastoid muscle.

Needling

Vertically 0.5–1 cun. Caution: Internal carotid artery, internal and external jugular veins.

Actions/Indications

- Expels Wind, opens the channel, disperses accumulations, benefits the ears and throat

Special features

Window of Heaven point

S.I.-18 *Cheekbone Crevice* QUANLIAO

Location

At the intersection of a vertical line through the outer canthus of the eye with the lower border of the zygomatic bone, on the anterior border of the masseter muscle.

How to find

Draw an imaginary vertical line from the outer canthus of the eye to the lower border of the zygomatic bone. **S.I.-18** is located in a depression on the anterior border of the masseter muscle, approximately on the level of the lateral border of the ala nasi (→ **L.I.-20**). By making chewing movements, the anterior border of the masseter muscle becomes easily palpable.

Needling

0.3–0.5 cun vertically or 1–1.5 cun transversely (subcutaneously) in the direction of → **ST-4,** → **ST-7** and → **L.I.-20**. According to some texts, moxibustion is contraindicated.

Actions/Indications

- Eliminates Wind, alleviates pain, clears Heat, reduces swellings

Special features

Meeting point with the Triple Burner channel. Important local point for disorders of the face (caused by internal or external Wind).

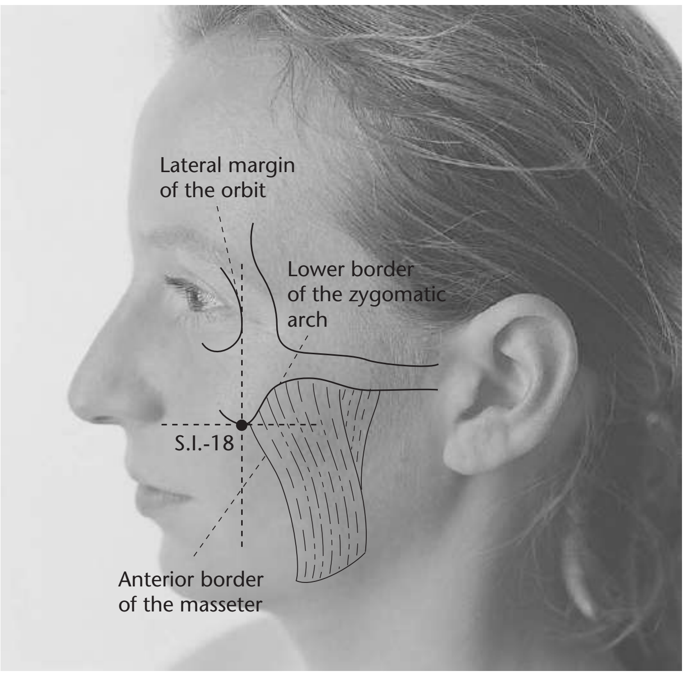
Lateral margin of the orbit
Lower border of the zygomatic arch
S.I.-18
Anterior border of the masseter

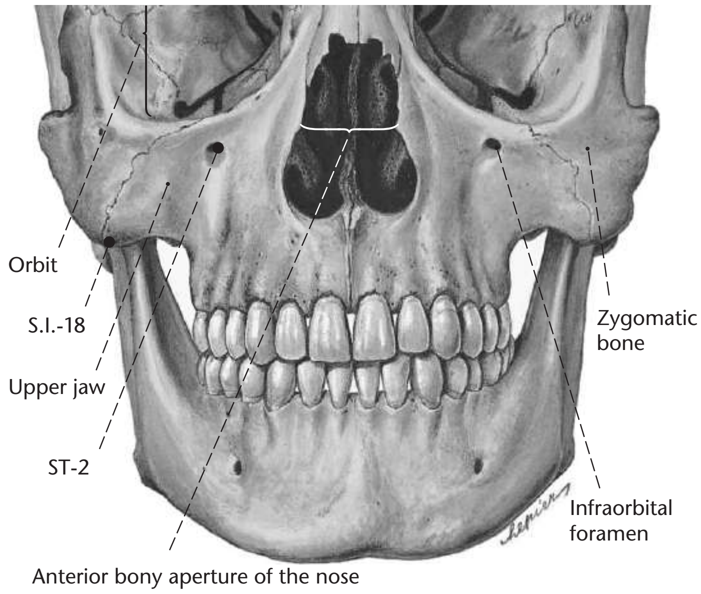
Orbit
S.I.-18
Upper jaw
ST-2
Zygomatic bone
Infraorbital foramen
Anterior bony aperture of the nose

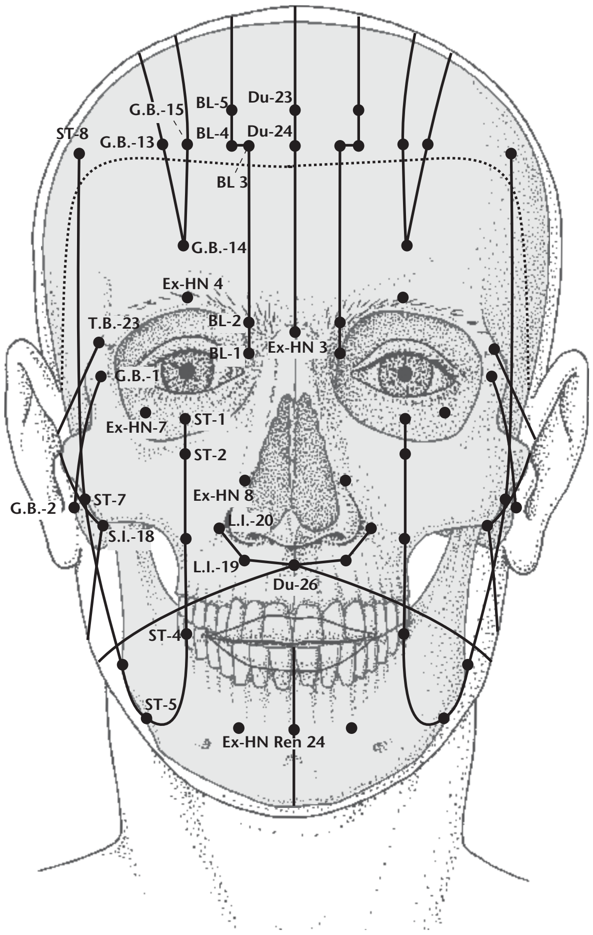
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL 3
G.B.-14
Ex-HN 4
T.B.-23
BL-2
BL-1
Ex-HN 3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN 8
G.B.-2
ST-7
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4
ST-5
Ex-HN Ren 24

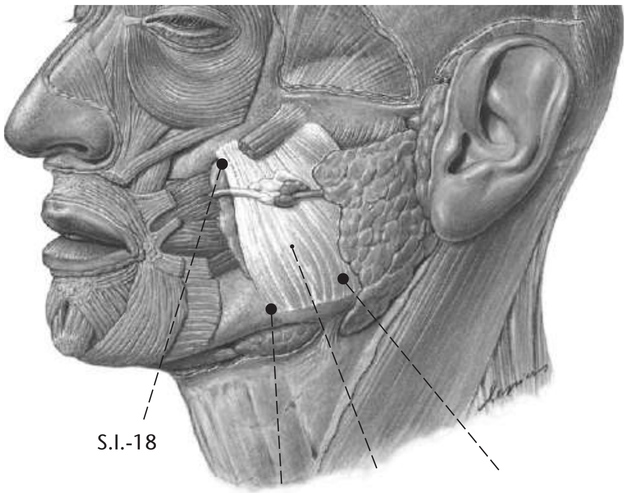

Palace of Hearing TINGGONG **S.I.-19**

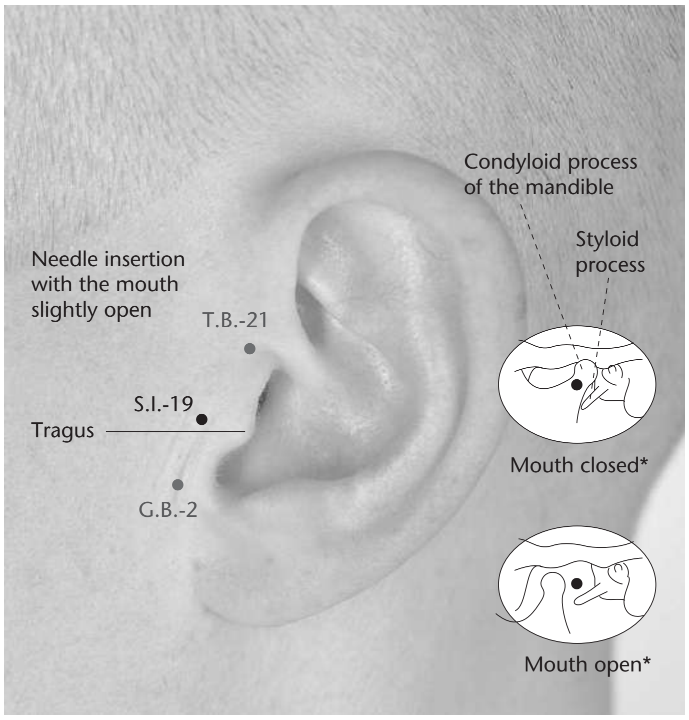
Condyloid process
of the mandible
Styloid process
Needle insertion with the mouth slightly open
T.B.-21
S.I.-19
Tragus
G.B.-2
Mouth closed*
Mouth open*

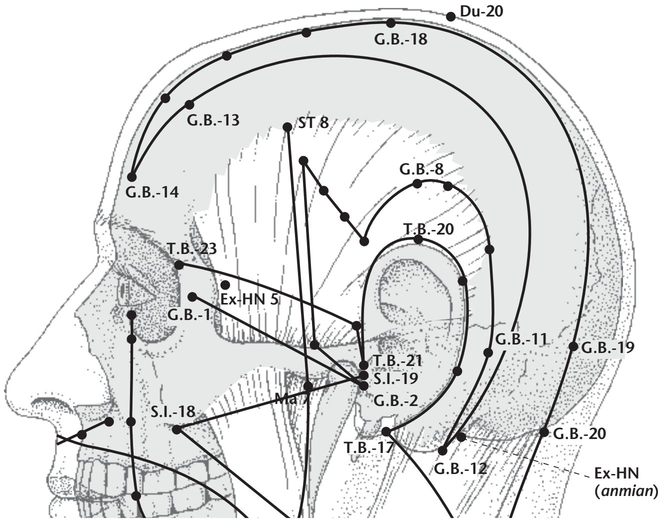
Du-20
G.B.-18
G.B.-13
ST 8
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN 5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
Ma
T.B.-17
G.B.-12
G.B.-20
Ex-HN (anmian)

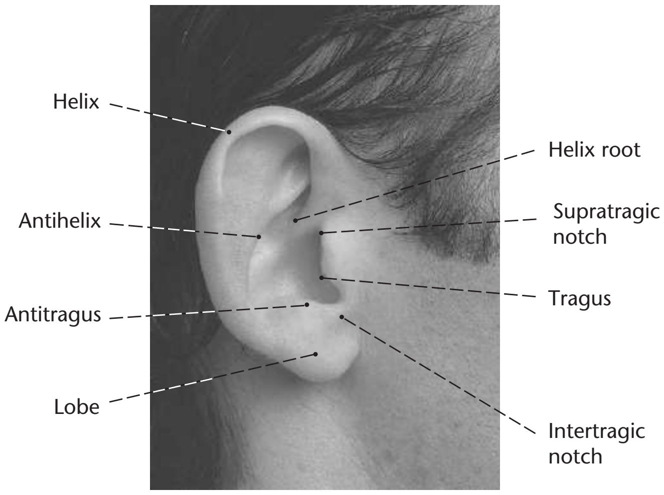
Helix
Antihelix
Antitragus
Lobe
Helix root
Supratragic notch
Tragus
Intertragic notch

Location

Anterior to the ear, with the mouth open at the level of the middle of the tragus, in a depression between the tragus and condyloid process of the mandible.

How to find

Locate and needle this point with the mouth open. This allows the condyloid process of the mandible to slide anteriorly and reveals the depression. Locate the vertical sulcus at the junction between the ear and the cheek (this is not always very pronounced). Then locate **S.I.-19** on the level of the tragus on the sulcus. If you cannot clearly define the sulcus (it becomes more pronounced with increasing age), an (ear) point locator can be used to locate the point (the patient should have their mouth open in order to reveal the depression).

S.I.-19 is the centre point of three points located anteriorly to the ear (→ **T.B.-21** is located proximal and → **G.B.-2** is located distal to **S.I.-19**).

Needling

0.5–1 cun vertically or slightly obliquely in an inferior direction. Needle with the mouth open (to avoid intra-articular needling). Caution: This point is located close to the superficial temporal artery and the auriculotemporal nerve (as are → **T.B.-21** and → **G.B.-2**).

Actions/Indications

- Benefits the ears
- Calms the *shen*

Special features

Meeting point with the Gall Bladder and Triple Burner channels, exit point. Important local point for disorders of the ears. For temporomandibular joint dysfunction and pain, rather use → **G.B.-2** as a local point.

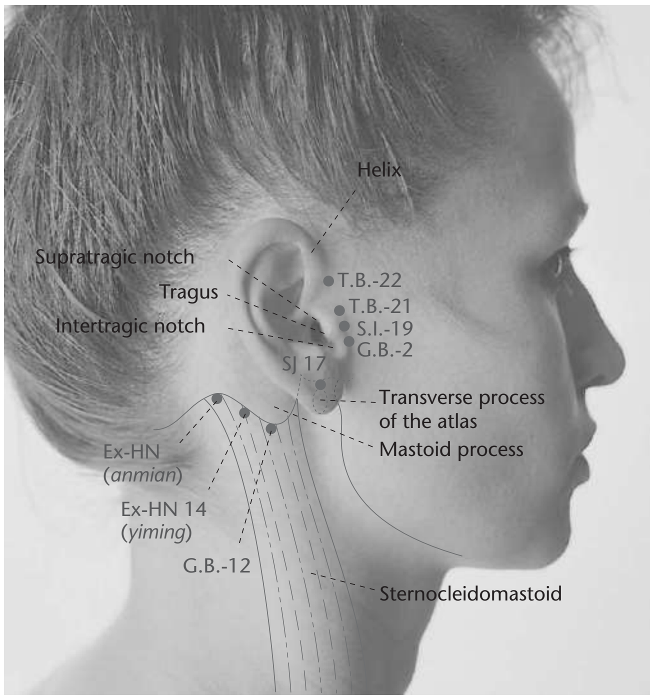
Helix
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
SJ 17
Transverse process of the atlas
Mastoid process
Ex-HN (anmian)
Ex-HN 14 (yiming)
G.B.-12
Sternocleidomastoid

---

* According to Deadman et al 1998.

4.7 The Bladder Channel System – Foot *taiyang* (*zu tai yang jing luo*)

4.7.1 The Bladder Primary Channel (*zu tai yang jing*)

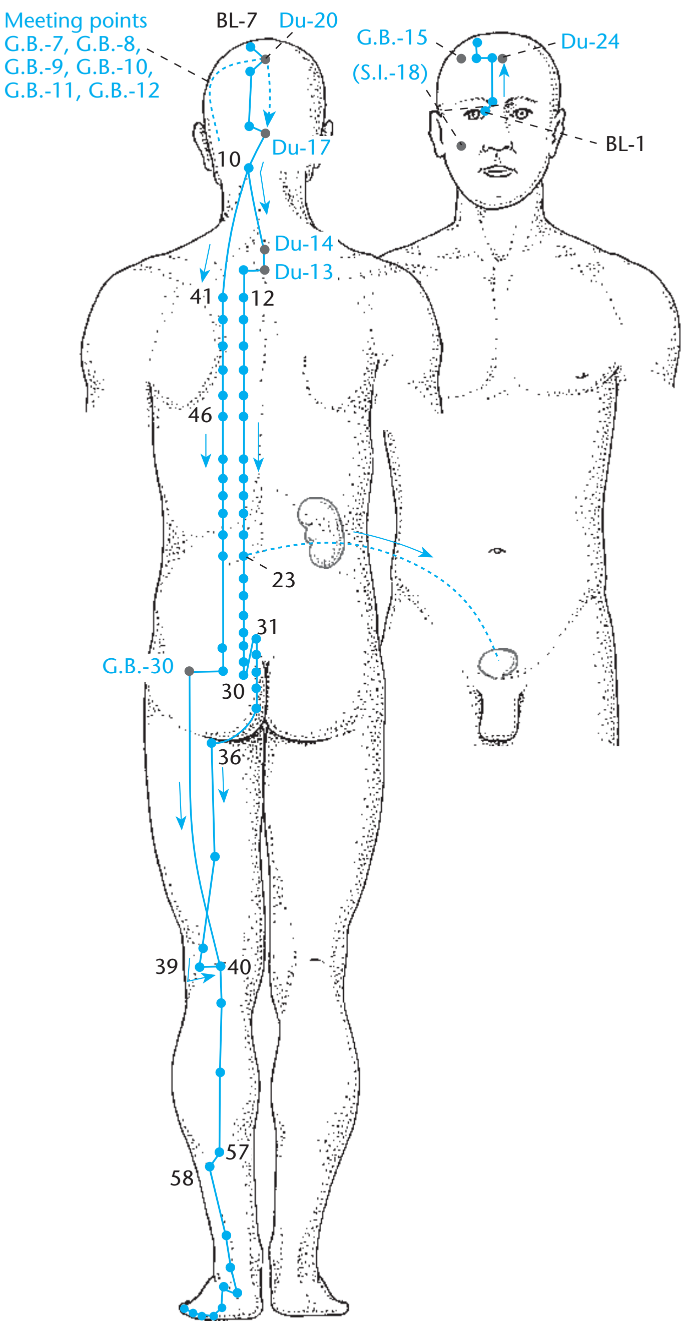
Meeting points
G.B.-7, G.B.-8,
G.B.-9, G.B.-10,
G.B.-11, G.B.-12
BL-7
Du-20
G.B.-15
(S.I.-18)
Du-24
BL-1
Du-17
10
Du-14
Du-13
41
12
46
23
31
G.B.-30
30
36
39
40
57
58

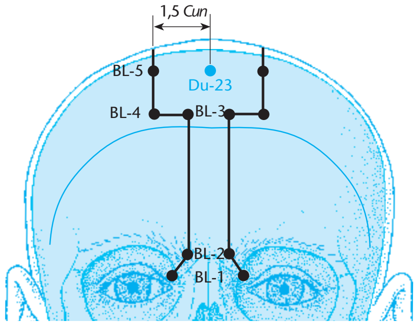
1,5 Cun
BL-5
Du-23
BL-4
BL-3
BL-2
BL-1

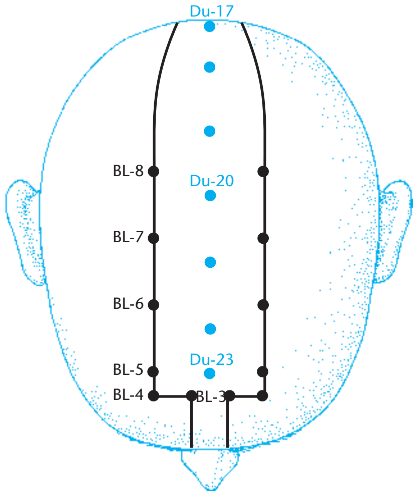
Du-17
BL-8
Du-20
BL-7
BL-6
BL-5
Du-23
BL-4
BL-3

Pathway

The **external** pathway of the Bladder primary channel begins at the inner canthus of the eye at → **BL-1** (*jingming*). This point is reached by a branch that separates from the Small Intestine primary channel at → **S.I.-18** (*quanliao*) (hand–foot pairing of the second great circuit, Yang axes: *taiyang*).

The **external** pathway ascends the forehead, there intersects with → **Du-24** (*shenting*) and → **G.B.-15** (*toulinqi*), continues to the vertex and there connects with the *du mai* at → **Du-20** (*baihui*).

At → **Du-20 two branches** separate from the primary channel:

- **one branch** travels to the temporal region and intersects the Gall Bladder primary channel at **G.B.-8** (*shuaigu*), **G.B.-7** (*qubin*), **G.B.-9** (*tianchong*), **G.B.-10** (*fubai*), **G.B.-11** (*touqiaoyin*) and **G.B.-12** (*wangu*)
- the **other branch** penetrates deeper into the brain and re-emerges at → **Du-17** (*naohu*) or, according to some authors, at **BL-8** (*luoque*).

From → **Du-20**, the **external pathway** runs to **BL-8** (*luoque*) and **BL-9** (*yuzhen*), passes → **Du-17** (*naohu*) and continues to **BL-10** (*tianzhu*) in the occipital region. Here, the **external pathway** of the channel separates into **two branches**:

- the **medial** branch runs to → **Du-14** (*dazhui*) below the spinal process of C7 where it meets the other Yang primary channels, then continues to → **Du-13** (*taodao*). From T1 it runs parallel to the posterior midline, 1.5 cun lateral to it.

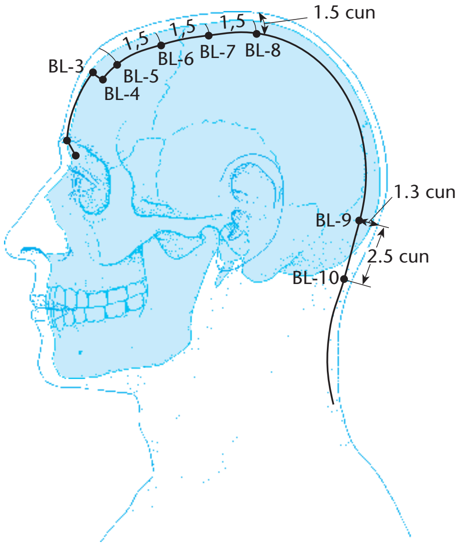
1.5 cun
1.5
1.5
1.5
BL-3
BL-6
BL-7
BL-8
BL-5
BL-4
1.3 cun
BL-9
2.5 cun
BL-10

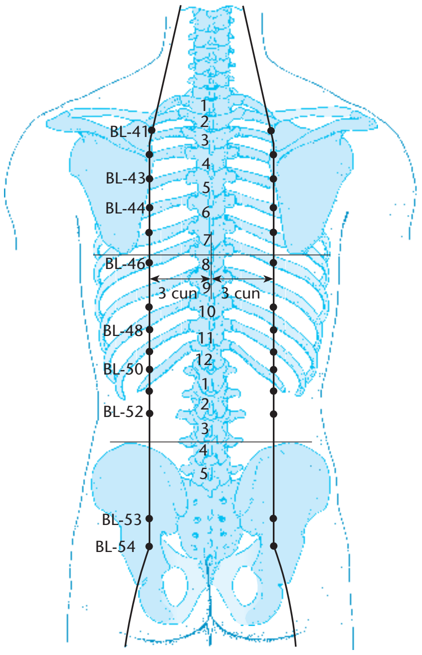
BL-41
1
2
3
BL-43
4
5
BL-44
6
7
BL-46
8
9
3 cun
3 cun
10
BL-48
11
12
BL-50
1
2
BL-52
3
4
5
BL-53
BL-54

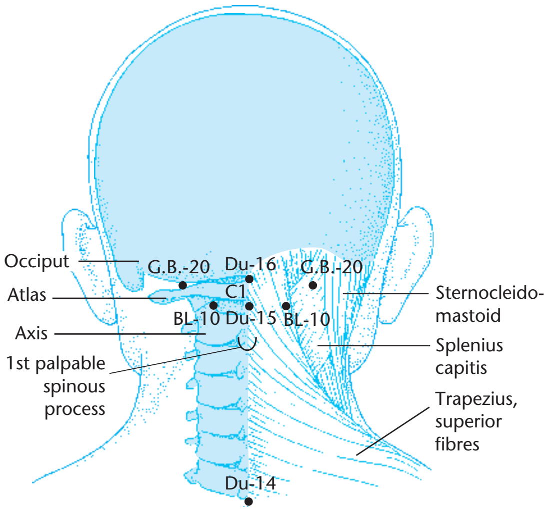
Occiput
G.B.-20
Du-16
G.B.-20
Atlas
C1
Sternocleido-mastoid
BL-10
Du-15
BL-10
Axis
Splenius capitis
1st palpable spinous process
Trapezius, superior fibres
Du-14

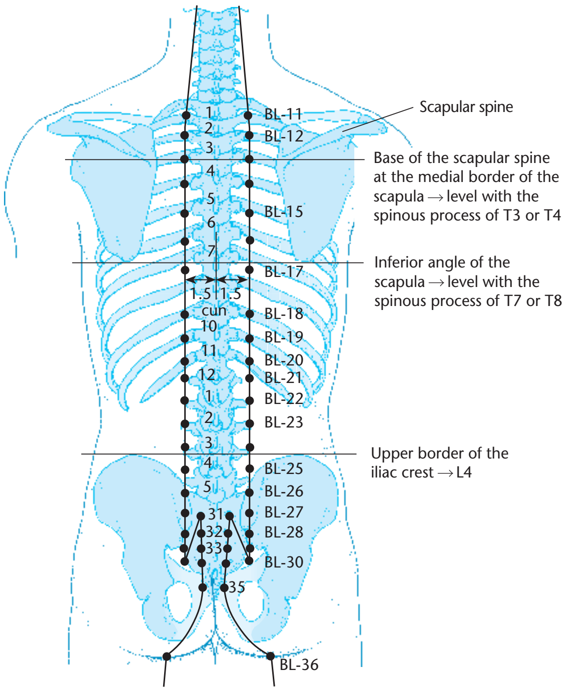
1
BL-11
2
BL-12
3
Scapular spine
4
Base of the scapular spine at the medial border of the scapula → level with the spinous process of T3 or T4
5
BL-15
6
7
BL-17
Inferior angle of the scapula → level with the spinous process of T7 or T8
1.5
1.5
cun
BL-18
10
BL-19
11
BL-20
12
BL-21
1
BL-22
2
BL-23
3
Upper border of the iliac crest → L4
4
BL-25
5
BL-26
31
BL-27
32
BL-28
33
BL-30
35
BL-36

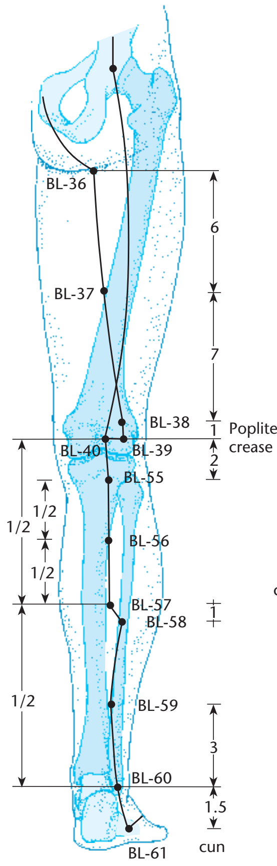
BL-36
6
BL-37
7
BL-38
1
Popliteal crease
BL-40
BL-39
2
BL-55
1/2
1/2
1/2
BL-56
BL-57
BL-58
1
1/2
BL-59
3
BL-60
1.5
cun
BL-61

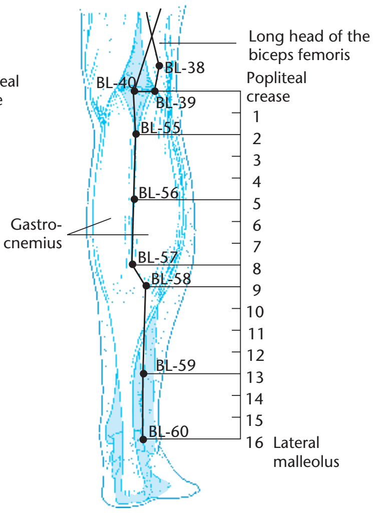
Long head of the biceps femoris
BL-38
BL-40
Popliteal crease
BL-39
1
BL-55
2
3
4
BL-56
5
Gastro-cnemius
6
7
BL-57
8
BL-58
9
10
11
12
BL-59
13
14
15
BL-60
16
Lateral malleolus

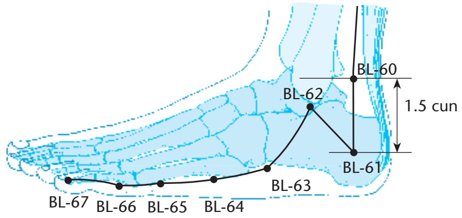
BL-60
BL-62
1.5 cun
BL-61
BL-63
BL-67
BL-66
BL-65
BL-64

At the level of L2, it **enters** the abdomen and reaches its paired *zang*-Organ, the Kidney (*shen*) and its pertaining *fu*-Organ, the Bladder (*pangguang*). The **external** branch runs through the gluteal region and the thigh to the popliteal fossa (**BL-40**), where it re-unites with the lateral external branch.

- the **lateral branch** runs from **BL-10** (*tianzhu*) to **BL-41** (*fufen*) at the level of T2, from where it descends down the back, 3 cun lateral to the midline to the level of the fourth sacral foramen. From there it traverses the gluteal region to → **G.B.-30** (*huantiao*), continues along the posterolateral aspect of the thigh and meets the medial external branch at the popliteal fossa.

From the popliteal fossa (**BL-40**) the channel descends along the posterolateral aspect of the foreleg to the foot

- passes the lateral malleolus
- travels along the lateral border of the foot to the lateral corner of the little toenail to **BL-67** (*zhiyin*). From here, a branch travels to → **KID-1** (*yongquan*) located on the Interiorly–Exteriorly paired Kidney primary channel (foot Yin–Yang pairing of the second great circuit).

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Chills and fever, headaches, occipital stiffness, pain in the lumbar region, obstruction of the nasal passages, eye pain with tearing, pain along the posterior aspect of the hip, knee and foreleg, foot pain

Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain and tension in the lower abdomen, retention of urine, enuresis, psycho-emotional disorders, opisthotonus

Connections and meeting points of the Bladder primary channel (→ 1.2)

Connections with other channels

Kidney primary channel (*zu shao yin jing*)

Connection: Foot Yin–Yang pairing of the second great circuit
Location: BL-67 → KID-1 (on the foot).
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Small Intestine primary channel (*shou tai yang jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *tai yang* (Yang axes of the second great circuit)
Location: S.I.-18 → BL-1 (on the head).
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with other *zangfu*-Organ systems

Kidney (*shen*), Bladder (*pangguang*)

4.7.2 The Bladder Divergent Channel (*zu tai yang jing bie*)

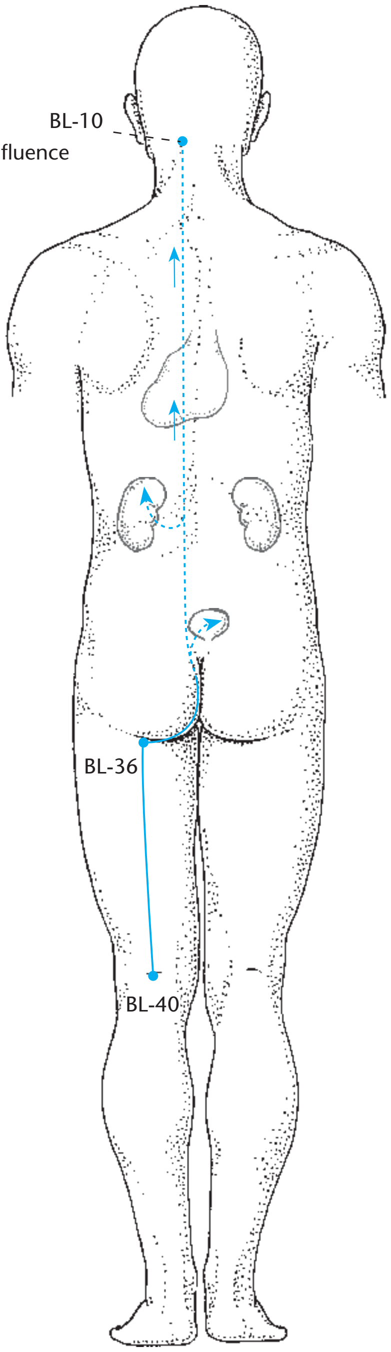

Pathway

The Bladder divergent channel separates from the Bladder primary channel in the popliteal fossa at **BL-40** (*weizhong*)

- ascends to the centre of the gluteal crease at **BL-36** (*chengfu*)
- ascends further and penetrates the anus, continues to the Bladder (*pangguang*), its pertaining *fu*-Organ, and the Kidney (*shen*), its paired *zang*-Organ
- rises along the spine and spreads in the Heart (*xin*)

Ch04.7a-F10028.qxd 2/23/08 3:53 PM Page 246

➡ from here it ascends to the occiput, where it unites at **BL-10** (*tianzhu*) with the Kidney divergent channel and the Bladder primary channel to form one of the six *he*-confluences (here: KID/BL as first confluence → 1.3)

Clinical importance

- Strengthens the relationship between the Bladder and the Kidney (*zangfu*-Organs). Points on the Bladder primary channel can therefore be used for disorders of the Kidney, and vice versa points on the Kidney primary channel can treat disorders of the Bladder.
- Supports the connection between the Bladder primary channel and the Heart
- Supports the connection between the Bladder primary channel and the anus
- Creates a connection between the popliteal region (**BL-40**) and the sacral region (**BL-32**, **BL-31**) as well as with the occipital region (**BL-10**).

4.7.3 The Bladder Sinew Channel (*zu tai yang jing jin*)

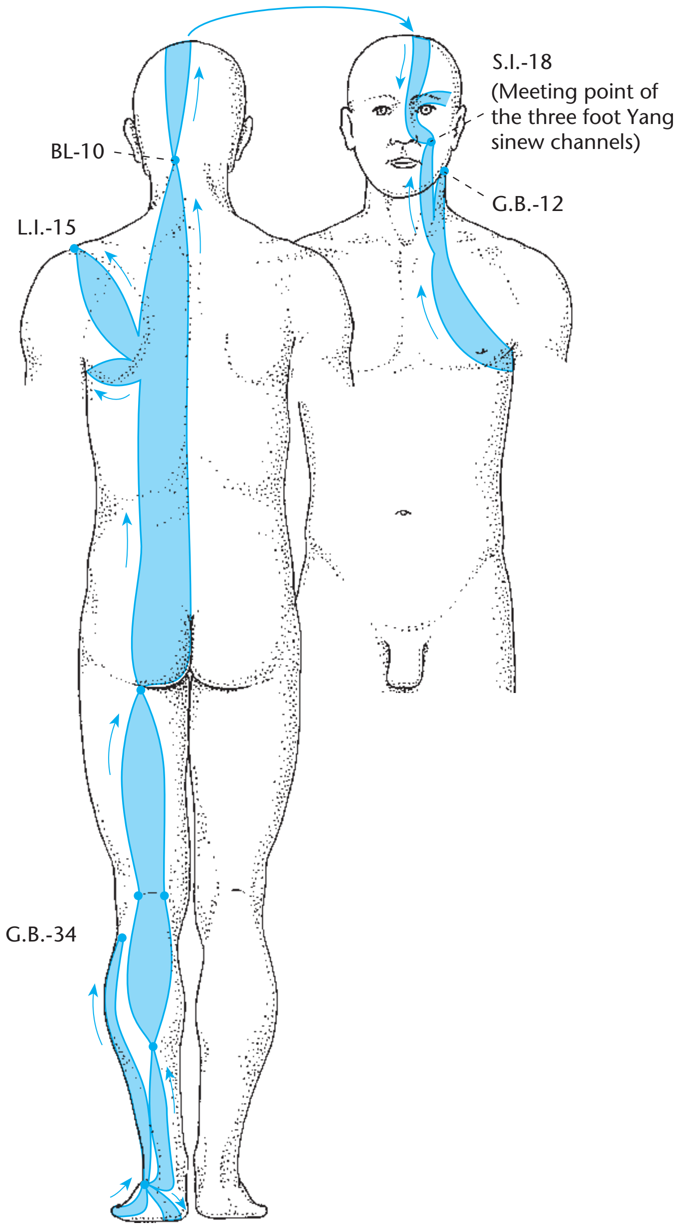

Pathway

The Bladder sinew channel begins on the lateral aspect of the little toe and travels along the lateral border of the foot to the lateral malleolus, where it binds (*jie*).

At the lateral malleolus **three branches** separate from the sinew channel:

➡ **one branch** runs to the calcaneus where it binds (*jie*) and disperses over the heel

➡ **a further branch** ascends the lateral aspect of the lower leg and binds (*jie*) at the head of the fibula, where it also meets the Gall Bladder and Stmach sinew channels at → **G.B.-34** (*yanglingquan*)

➡ **the main branch** travels to the Achilles tendon, binds (*jie*) in the area of the two heads of the gastrocnemius muscle, ascends the posterior aspect of the leg along the gastrocnemius and soleus muscles and binds (*jie*) on either side of the popliteal crease, ascends the posterior aspect of the thigh and binds (*jie*) at the centre of the gluteal crease, ascends as a broad band along the spine and binds (*jie*) at **BL-10** (*tianzhu*).

Two branches separate from the main branch in the thoracic region at the level of the axilla:

- ➡ **one branch** ascends to the shoulder and reaches → **L.I.-15** (*jianyu*)
- ➡ **the other branch** penetrates the axilla, ascends the thorax and runs to the supraclavicular fossa, where again **two sub-branches** separate:
    - **one sub-branch** ascends to the mastoid process to → **G.B.-12** (*wangu*)
    - **the other sub-branch** crosses the cheek to → **S.I.-18** (*quanliao*) where it meets with the other foot Yang sinew channels

The **main branch** divides at **BL-10** (*tianzhu*) into **two branches:**

- ➡ **one branch** penetrates deeper to the root of the tongue
- ➡ **the other branch** crosses the head lateral to the midline, reaches the inner canthus of the eye and separates into **two branches:**
    - **one branch** binds (*jie*) lateral to the nose and descends the cheek
    - **the other branch** crosses the upper eyelid and connects with the Stomach sinew channel, together forming a network around the eye. It then penetrates deeper into the upper orbital region and ends at the roof of the mouth.

Clinical importance

Pathology: Muscular cramping in the region of the little toe. Tension and pain in the heel. Stiffness and limited range of motion of the joints. Tension and stiffness in the back and along the spine, and difficulty when bending forward. Inability to lift the arms above the shoulder region. Stiffness or distending pain in the axillary region. Pain and tense musculature in the supraclavicular fossa. Limited range of motion of the cervical spine.

Indication: Mainly with *bi*-syndromes (painful obstruction syndrome) along the Bladder channel. The area covered by the Bladder sinew channel is larger than that covered by the Bladder primary channel. This explains why Bladder primary channel points can also be used for disorders of the scapula and the shoulder region as well as for disorders of the lower eyelid and the nasal region.

4.7.4 The Bladder *luo*-Connecting Vessel System (*zu tai yang luo mai*)

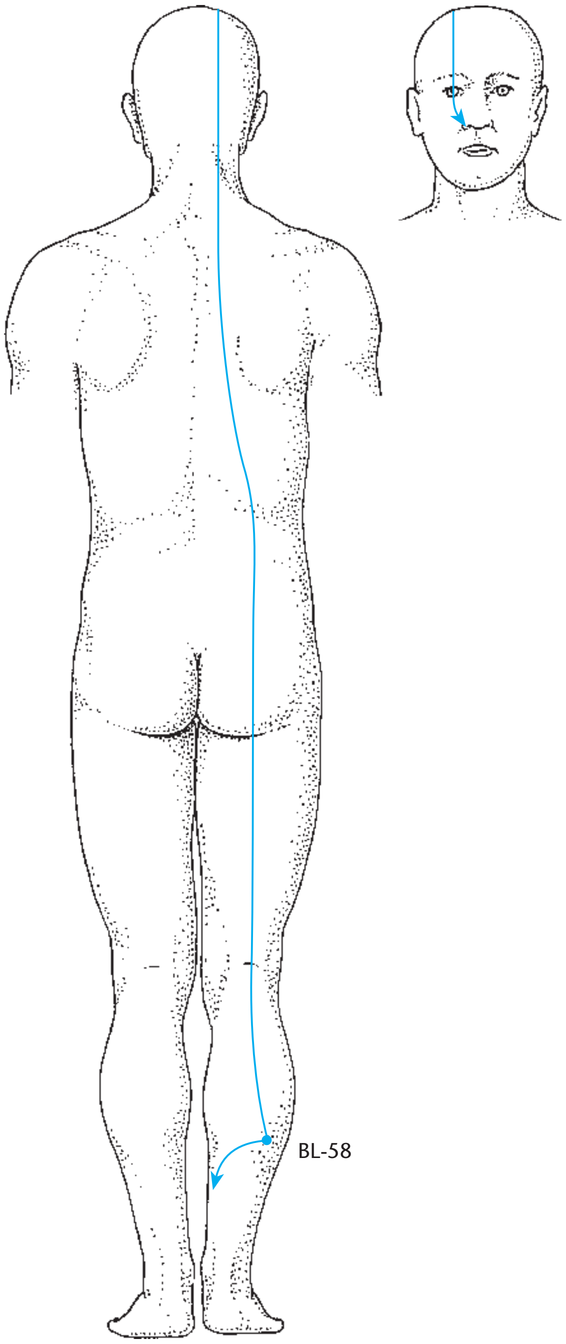

Pathway

The Bladder *luo*-connecting channel separates from the Bladder primary channel at **BL-58** (*feiyang*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Kidney primary channel; according to some schools of thought (for example Nguyen Van Nghi → Appendix) they travel as a **transverse** BL *luo*-connecting vessel to the *yuan*-source point → **KID-3** (*taixi*).

- Solinas et al. (1998, variant → fig.) describe a **longitudinal** division following the pathway of the Bladder primary channel, reaching the head and penetrating the nose. While this division is not mentioned in the classics, according to Solinas et al it becomes evident based on the classic descriptions of the pathologies of the *luo*-connecting vessels.

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*):** Obstruction of the nasal passages, clear watery nasal discharge, headaches, back pain
- **Deficiency (*xu*):** Nosebleeds, chronic clear nasal discharge, sinusitis

4.7.5 Cutaneous Region (*tai yang pi bu*)

See description and figures → 1.6.

4.7.6 Points of the Bladder Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): BL-64** (*jinggu*)
- ***Luo*-connecting point (→ 8.1.2): BL-58** (*feiyang*)
- ***Xi*-cleft point (→ 8.1.3): BL-63** (*jinmen*)
- ***Xi*-cleft point of the *yang qiao mai*: BL-59** (*fuyang*)
- **Associated Back-*shu* point (→ 8.1.4): BL-28** (*pangguangshu*) 🟦🟦
  – Back-*shu* point of the Lung: **BL-13** (*feishu*) 🟦🟦
  – Back-*shu* point of the Pericardium: **BL-14** (*jueyinshu*)
  – Back-*shu* point of the Heart: **BL-15** (*xinshu*) 🟦🟦
  – Back-*shu* point of the *du mai*: **BL-16** (*dushu*)
  – Back-*shu* point of the diaphragm: **BL-17** (*geshu*) 🟦🟦
  – Back-*shu* point of the Liver: **BL-18** (*ganshu*) 🟦🟦
  – Back-*shu* point of the Gall Bladder: **BL-19** (*danshu*) 🟦⬛
  – Back-*shu* point of the Spleen: **BL-20** (*pishu*) 🟦🟦
  – Back-*shu* point of the Stomach: **BL-21** (*weishu*) 🟦🟦
  – Back-*shu* point of the Triple Burner: **BL-22** (*sanjiaoshu*) 🟦⬛
  – Back-*shu* point of the Kidneys: **BL-23** (*shenshu*) 🟦🟦
  – Back-*shu* point of the Large Intestine: **BL-25** (*dachangshu*) 🟦⬛
  – Back-*shu* point of the Small Intestine: **BL-27** (*xiaochangshu*)
  – Back-*shu* point of the Bladder: **BL-28** (*pangguangshu*) 🟦🟦
- **Associated Front-*mu* point (→ 8.1.5): Ren-3** (*zhongji*) 🟦🟦
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Metal), tonification point: **BL-67** (*zhiyin*) 🟦⬛
  ying-spring point (Water), *ben* point (Five Phases): **BL-66** (*zutonggu*)
  shu-stream point (Wood), sedation point: **BL-65** (*shugu*)
  jing-river point (Fire): **BL-60** (*kunlun*) 🟦🟦
  he-sea point (Earth): **BL-40** (*weizhong*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7):**
  – of the Blood: **BL-17** (*geshu*) 🟦🟦
  – of the Bones: **BL-11** (*dazhu*) 🟦⬛
- **Opening point (→ 8.1.8) of the *yang qiao mai*: BL-62** (*shenmai*) 🟦🟦
- **Lower *he*-sea points (→ 8.1.9):**
  of the Triple Burner: **BL-39** (*weiyang*) 🟦⬛
  of the Bladder: **BL-40** (*weizhong*) 🟦🟦
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the ST, T.B.* channels, the *yin qiao mai*, *yang qiao mai*, S.I., G.B.* channels, *du mai*: **BL-1** (*jingming*) 🟦⬛
  – with the S.I. channel, (T.B., G.B. channels, *du mai**): **BL-11** (*dazhu*) 🟦⬛
  – with the *du mai*: **BL-12** (*fengmen*) 🟦⬛
  – with the *du mai**: **BL-23** (*shenshu*) 🟦🟦
  – with the G.B. channel: **BL-31–BL-34** 🟦⬛
  – with the S.I. channel: **BL-41** (*fufen*)
  – with the *yang qiao mai*: (**BL-59** (*fuyang*)*), **BL-61** (*pucan*), **BL-62*** (*shenmai*) 🟦🟦
  – with the *yang wei mai*: **BL-63** (*jinmen*)
  – of other channels with the Bladder channel: **Du-24, G.B.-15, Du-20, (G.B.-6*), G.B.-7, G.B.-8, G.B.-9, G.B.-10, G.B.-11, G.B.-12, Du-17, Du-16*, Du-14, Du-13, G.B.-30**
- **Gao Wu command point (→ 8.1.11) for the Back and lumbar region: BL-40** (*weizhong*) 🟦🟦
- **Window of Heaven point (→ 8.1.12): BL-10** (*tianzhu*) 🟦🟦
- **Points of the Four Seas (→ 8.1.13):**
  – Sea of Blood point: **BL-11** (*dazhu*) 🟦⬛
- **Ma Dan Yang Heavenly Star points (→ 8.1.14): BL-40** (*weizhong*) 🟦🟦, **BL-57** (*chengshan*) 🟦⬛, **BL-60** (*kunlun*) 🟦🟦
- **Sun Si Miao Ghost point (→ 8.1.15): BL-62** (*shenmai*) 🟦🟦
- **Other functional points:**
  – *xi*-cleft point of the *yang qiao mai*: **BL-59** (*fuyang*)

Points according to region

- **Local points (→ 8.2.1):** occiput – **BL-10** (*tianzhu*) 🟦🟦; eyes – **BL-1** (*jingming*) 🟦⬛; Stomach/Spleen – **BL-20** (*pishu*) 🟦🟦, **BL-21** (*weishu*) 🟦🟦; Kidney – **BL-23** (*shenshu*) 🟦🟦, **BL-62** (*shenmai*) 🟦🟦; rectum – **BL-35** (*huiyang*); cervical spine – **BL-10** (*tianzhu*) 🟦🟦; thoracic spine – **BL-17** – **BL-23**; lumbar spine – **BL-23** – **BL-26**, **BL-52** (*zhishi*); iliosacral region – **BL-27** (*xiaochangshu*), **BL-28** (*pangguangshu*); hip – **BL-49** (*yishe*), **BL-50** (*weicang*), **BL-54** (*zhibian*); knee – **BL-40** (*weizhong*) 🟦🟦; foot – **BL-60** (*kunlun*) 🟦🟦, **BL-62** (*shenmai*) 🟦🟦
- **Adjacent points (→ 8.2.1):** eyes – **BL-2** (*zanzhu*) 🟦⬛; nose – **BL-7** (*tongtian*); Lung – **BL-13** (*feishu*) 🟦🟦, **BL-43** (*gaohuang*); Heart – **BL-14** (*jueyinshu*), **BL-15** (*xinshu*) 🟦🟦; Liver – **BL-18** (*ganshu*) 🟦🟦, **BL-20** (*pishu*) 🟦🟦; Gall Bladder – **BL-19** (*danshu*) 🟦⬛; Large Intestine – **BL-25** (*dachangshu*) 🟦🟦; Small Intestine – **BL-27** (*xiaochangshu*) 🟦⬛; rectum – **BL-30** (*baihuanshu*), **BL-34** (*xialiao*) 🟦⬛; Bladder – **BL-28** (*pangguangshu*) 🟦🟦, **BL-23** (*shenshu*) 🟦🟦; urogenital region – **BL-23** (*shenshu*) 🟦🟦, **BL-32** (*ciliao*); cervical spine – **BL-11** (*dazhu*) 🟦⬛;

---

* Mentioned by only some authors

iliosacral region – **BL-23** (*shenshu*) 🟦🟦; thoracic spine – **BL-38** – **BL-47**; toes – **BL-60** (*kunlun*) 🟦🟦, **BL-65** (*shugu*)

- **Distal points (→ 8.2.1):** occipital region – **BL-65** (*shugu*), **BL-60** (*kunlun*); rectum – **BL-57** (*chengshan*) 🟦🟦; thoracic spine – **BL-57** (*chengshan*) 🟦⬛; cervical, thoracic, lumbar spine – **BL-60** (*kunlun*) 🟦🟦; shoulder and iliosacral region – **BL-58** (*feiyang*); for the thoracic spine, lumbar spine and iliosacral region – **BL-40** (*weizhong*) 🟦🟦; lumbar spine – **BL-40** (*weizhong*) 🟦🟦, **BL-59** (*fuyang*), **BL-60** (*kunlun*) 🟦🟦; lumbar spine and hip – **BL-62** (*shenmai*) 🟦🟦

Specific points according to the channel pathway (in numerical order):

- **BL-1** (*jingming*) 🟦⬛: *jiaohui*-meeting point with the ST, T.B.* channels, the *yin qiao mai, yang qiao mai*, S.I., G.B.* channels, *du mai* (→ 8.1.10): local point for the eyes (→ 8.2.1)
- **BL-2** (*zanzhu*) 🟦⬛: adjacent point for the eyes (→ 8.2.1)
- **BL-7** (*tongtian*): adjacent point for the nose (→ 8.2.1)
- **BL-10** (*tianzhu*) 🟦🟦: Window of Heaven point (→ 8.1.12); important point for eliminating Wind; local point for the occiput and the cervical spine (→ 8.2.1)
- **BL-11** (*dazhu*) 🟦⬛: *hui*-meeting point (→ 8.1.7) of the bones; Sea of Blood point (→ 8.1.13); *jiaohui*-meeting point with the S.I. channel (T.B., G.B. channels, *du mai**) (→ 8.1.10); adjacent point for the cervical spine (→ 8.1.10)
- **BL-12** (*fengmen*) 🟦⬛: *jiaohui*-meeting point with the *du mai* (→ 8.1.10)
- **BL-13** (*feishu*) 🟦🟦: Back-*shu* point of the Lung (→ 8.1.4); adjacent point for the Lung (→ 8.2.1)
- **BL-14** (*jueyinshu*): Back-*shu* point of the Pericardium (→ 8.1.4); adjacent point for the Heart (→ 8.2.1)
- **BL-15** (*xinshu*) 🟦🟦: Back-*shu* point of the Heart (→ 8.1.4); adjacent point for the Heart (→ 8.2.1)
- **BL-16** (*dushu*): Back-*shu* point of the *du mai* (→ 8.1.4);
- **BL-17** (*geshu*) 🟦🟦: Back-*shu* point of the diaphragm (→ 8.1.4); *hui*-meeting point of the Blood (→ 8.1.7)
- **BL-17 to BL-23**: local points for the thoracic spine (→ 8.2.1)
- **BL-18** (*ganshu*) 🟦🟦: Back-*shu* point of the Liver (→ 8.1.4); adjacent point for the Liver (→ 8.2.1)
- **BL-19** (*danshu*) 🟦⬛: Back-*shu* point of the Gall Bladder (→ 8.1.4); adjacent point for the Gall Bladder (→ 8.2.1)
- **BL-20** (*pishu*) 🟦🟦: Back-*shu* point of the Spleen (→ 8.1.4); adjacent point for the Liver (→ 8.2.1)
- **BL-21** (*weishu*) 🟦🟦: local point for the Stomach (→ 8.2.1); Back-*shu* point of the Stomach (→ 8.1.4)
- **BL-22** (*sanjiaoshu*) 🟦⬛: Back-*shu* point of the Triple Burner (→ 8.1.4)
- **BL-23** (*shenshu*) 🟦🟦: *jiaohui*-meeting point with the *du mai** (→ 8.1.10); local point for the Kidneys and lumbar region (→ 8.2.1); adjacent point for the Bladder, urogenital region and iliosacral region (→ 8.2.1); Back-*shu* point of the Kidney (→ 8.1.4)
- **BL-24** (*qihaishu*): local point for the lumbar spine (→ 8.2.1)
- **BL-25** (*dachangshu*) 🟦🟦: Back-*shu* point of the Large Intestine (→ 8.1.4); adjacent point for the lumbar spine (→ 8.2.1)
- **BL-26** (*guanyuanshu*): local point for the lumbar spine (→ 8.2.1)
- **BL-27** (*xiaochangshu*) 🟦⬛: Back-*shu* point of the Small Intestine (→ 8.1.4); adjacent point for the Small Intestine (→ 8.2.1); local point for the iliosacral region (→ 8.2.1)
- **BL-28** (*pangguangshu*) 🟦🟦: Back-*shu* point of the Bladder (→ 8.1.4); adjacent point for the Bladder (→ 8.2.1); local point for the iliosacral region (→ 8.2.1)
- **BL-30** (*baihuanshu*): regional point for the rectum (→ 8.2.1)
- **BL-31** – **BL-34**: *jiaohui*-meeting points with the G.B. channel (→ 8.1.10)
- **BL-32** (*ciliao*): regional point for the urogenital region (→ 8.2.1); local point for the iliosacral region (→ 8.2.1)
- **BL-34** (*xiaoliao*): regional point for the rectum (→ 8.2.1)
- **BL-35** (*huiyang*): local point for the rectum (→ 8.2.1)
- **BL-38** – **BL-47**: adjacent points for the thoracic spine (→ 8.2.1)
- **BL-39** (*weiyang*) 🟦⬛: lower *he*-sea point of the Triple Burner (→ 8.1.9)
- **BL-40** (*weizhong*) 🟦🟦: *he*-sea point (Earth); lower *he*-sea point of the Bladder (→ 8.1.9); Gao Wu command point (→ 8.1.11) of the back and lumbar region; Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the thoracic and lumbar spine as well as for the iliosacral region (→ 8.2.1); local point for the knee (→ 8.2.1)
- **BL-41** (*fufen*): *jiaohui*-meeting point with the S.I. channel (→ 8.1.10)
- **BL-43** (*gaohuang*): adjacent point for the Lung (→ 8.2.1)
- **BL-49** (*yishe*): local point for the hip (→ 8.2.1)
- **BL-50** (*weicang*): local point for the hip (→ 8.2.1)
- **BL-52** (*zhishi*): local point for the lumbar region (→ 8.2.1)
- **BL-54** (*zhibian*): local point for the hip (→ 8.2.1)
- **BL-57** (*chengshan*) 🟦⬛: Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the rectum (→ 8.2.1); distal point for the thoracic spine (→ 8.2.1)
- **BL-58** (*feiyang*): *luo*-connecting point (→ 8.1.2); distal point for the shoulder and iliosacral region (→ 8.2.1)
- **BL-59** (*fuyang*): *xi*-cleft point of the *yang qiao mai; jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10)*; distal point for the lumbar spine (→ 8.2.1)
- **BL-60** (*kunlun*) 🟦🟦: *jing*-river point (Fire); Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the cervical, thoracic and lumbar spine (→ 8.2.1); distal point for the occipital and lumbar region (→ 8.2.1); local point for the foot (→ 8.2.1); regional point for the toes (→ 8.2.1)
- **BL-61** (*pucan*): *jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10)
- **BL-62** (*shenmai*) 🟦🟦:Opening point of the *yang qiao mai* (→ 8.1.8); Sun Si Miao Ghost point (→ 8.1.15); *jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10); local point

---

* Mentioned by only some authors

for the Kidney (→ 8.2.1); local point for the foot (→ 8.2.1); distal point for the lumbar spine and hip (→ 8.2.1)

- **BL-63** (*jinmen*): *xi*-cleft point (→ 8.1.3); *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **BL-64** (*jinggu*): *yuan*-source point (→ 8.1.1)
- **BL-65** (*shugu*): *shu*-stream point (Wood); sedation point; distal point for the occiput (→ 8.2.1); adjacent point for the toes (→ 8.2.1)
- **BL-66** (*zutongu*): *ying*-spring point (Water); *ben* point (Five Phases point)
- **BL-67** (*zhiyin*) 🟦■: *jing*-well point; tonification point

General location help

- **The medial branch of the Bladder channel** runs from **BL-11** (below the spinal process of the T1) to **BL-30** (level with the fourth sacral foramen), 1.5 cun lateral to the midline

Memorisation help for the medial branch of the BL channel:

- **BL11 to BL-17**: the second digit of the point number corresponds to the level of the respective thoracic vertebra. Example: **BL-13** is located below the spinal process of T3.
- **BL-18 to BL-21**: second digit of the point number + 1. *Example*: **BL-19** (9 + 1) is located below the spinal process of the T10.

The outer branch of the Bladder channel runs from **BL-41** (below the spinal process of the T2) to **BL-54** (below the spinal process of S4/level with the 4th sacral foramen) 3 cun lateral to the midline.

Memorisation help for the lateral branch of the BL channel:

- **BL41 to BL-46**: second digit of the point number + 1. *Example*: **BL-42** (2 + 1) corresponds to T3
- **BL-47 to BL-49**: second digit of the point number + 2. *Example*: **BL-47** (7 + 2) corresponds to the level of T9.
- **BL-50**: second digit of the point number + 12.

Needling: The Back-*shu* points are also commonly needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape).

Bright Eyes JINGMING BL-1

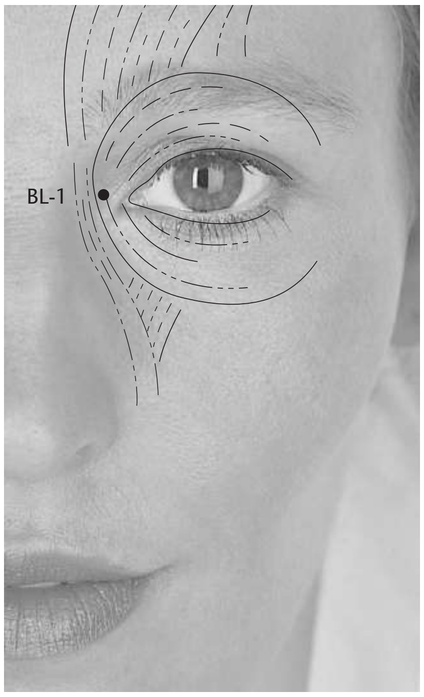

Location

0.1 cun superior and medial to the inner canthus of the eye, in a depression.

How to find

BL-1 is located in a small depression near the inner canthus of the eye, superior to the insertion of the upper eyelid. Medially, it is bordered by the bony root of the nose. → **BL-2** is located directly above **BL-1** in a depression at the medial end of the eyebrow. → **ST-1** is located on the pupil line between the eyeball and the border of the infraorbital ridge, → **G.B.-1** and → **T.B.-23** are located on the outer canthus of the eye.

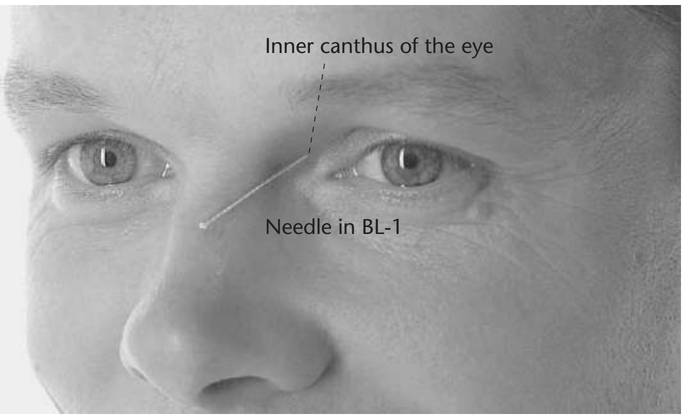
Inner canthus of the eye
Needle in BL-1

Needling

Ask the patient to look downward and outward, with their eyes closed. Gently push the eyeball downward and outward from above. Needle perpendicularly along the supraorbital ridge in a posterior direction. Caution: Angular artery and vein, eyeball, periosteum. This point should only be needled by experienced practitioners. After removing the needle, press on the insertion site for 10 minutes. Despite this precaution, a bruise may still form (inform the patient). **No** needle stimulation! Moxibustion is contraindicated. Alternative point: → **BL-2** or transverse (subcutaneous) insertion at **BL-1** in the direction of → **BL-2**.

Actions/Indications

- Expels Wind and clears Heat, benefits the eyes

Special features

Meeting point with the *du mai*, *yin qiao mai*, *yang qiao mai* as well as the ST, S.I., G.B. channels and, according to some authors, also the T.B. channel. Entry point.

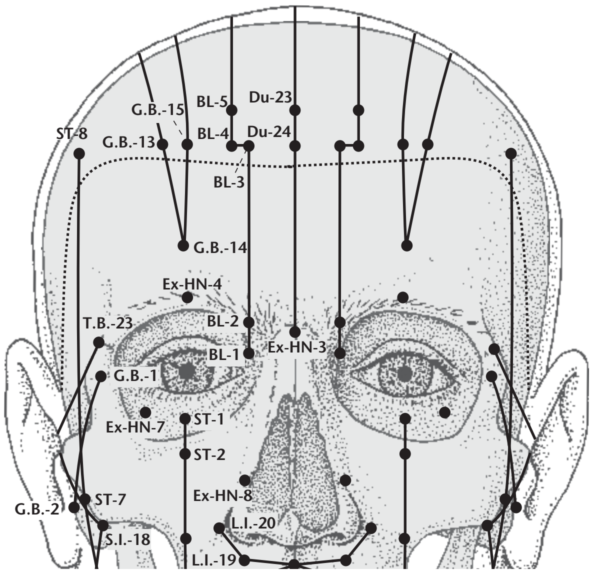
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
Ex-HN-3
BL-1
G.B.-1
Ex-HN-7
ST-1
ST-2
G.B.-2
ST-7
Ex-HN-8
L.I.-20
S.I.-18
L.I.-19

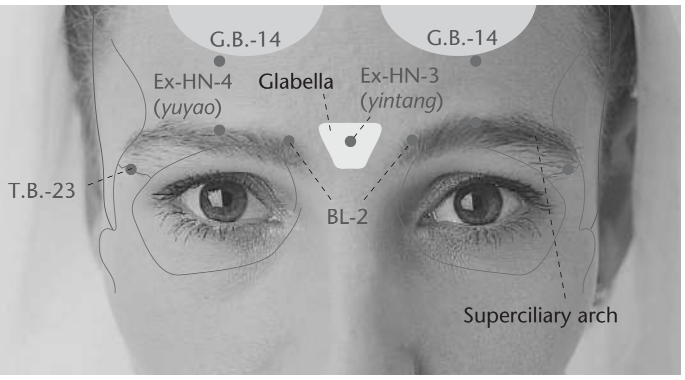
G.B.-14
G.B.-14
Ex-HN-4
(yuyao)
Glabella
Ex-HN-3
(yintang)
T.B.-23
BL-2
Superciliary arch

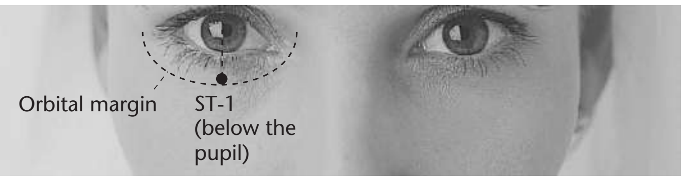
Orbital margin
ST-1
(below the
pupil)

BL-2 *Gathered Bamboo* ZANZHU

Location

In a depression at the medial end of the eyebrow, directly above the inner canthus of the eye.

How to find

As the location of the eyebrow can vary, the medial canthus of the eye should be used for reference (→ **BL-1**). Palpate for a generally pressure-sensitive depression in this area of the orbital ridge. Caution: According to Sobotta, except in rare cases, **BL-2** is not located on the supraorbital foramen (mainly located more laterally), but in the area where the supratrochlear artery and the supraorbital nerve emerge (frontal notch).

→ **BL-1** is located inferior to **BL-2**.

Needling

0.3–0.5 cun obliquely from lateral to medial or transversely (subcutaneously) in the direction of the disorder. Caution: Branches of the frontal nerve, facial nerve, blood vessels! According to the majority of authors, moxibustion is contraindicted. Considered a less risky alternative to → **BL-1**.

Actions/Indications

- Benefits the eyes
- Expels external pathogenic factors such as Wind, clears Heat
- Clears the head
- Moves Qi in the Bladder divergent channel (for example, for painful haemorrhoids)

Special features

Important and frequently used local point for the eye region. Good point for excess and Heat syndromes of the eye region: prick to bleed (puncture the point and let it bleed), but avoid facial haematoma.

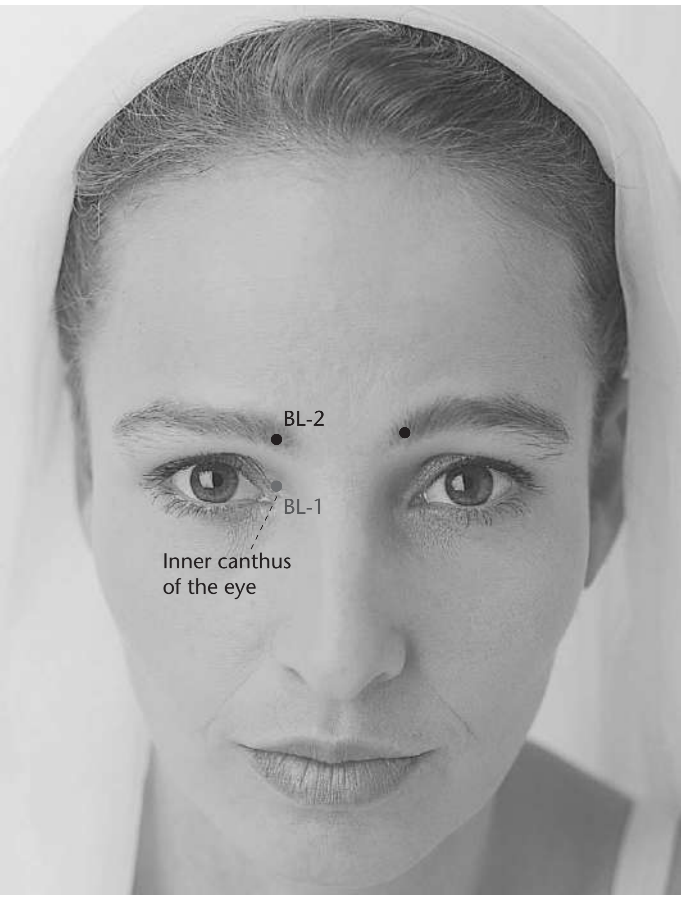
BL-2
BL-1
Inner canthus of the eye

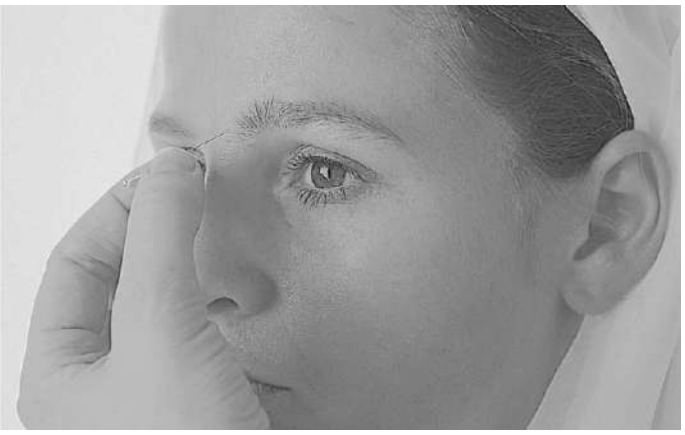

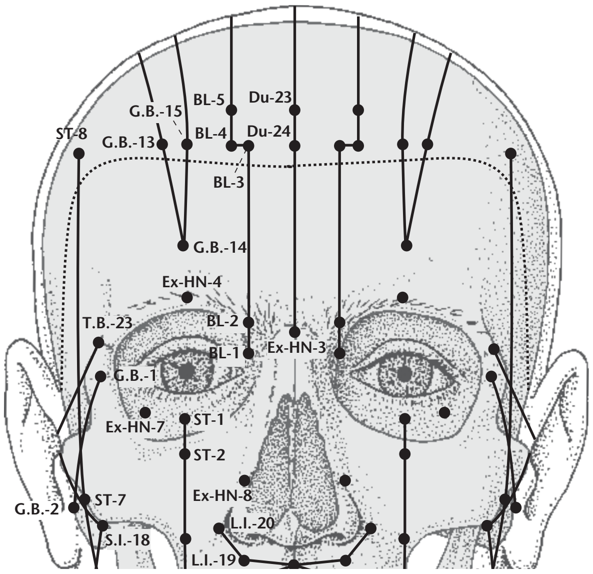

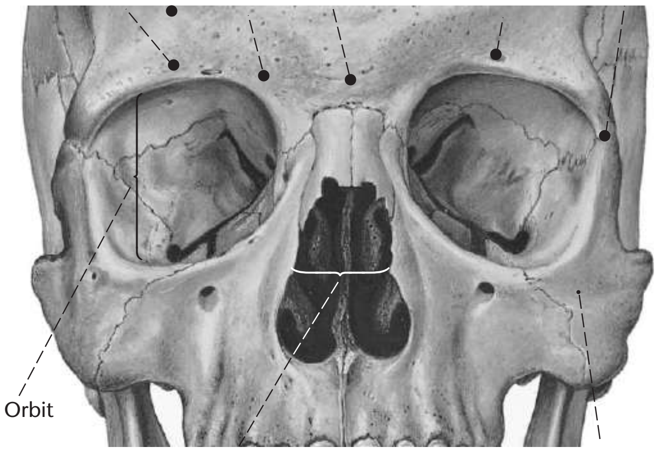

Eyebrows' Pouring MEICHONG BL-3

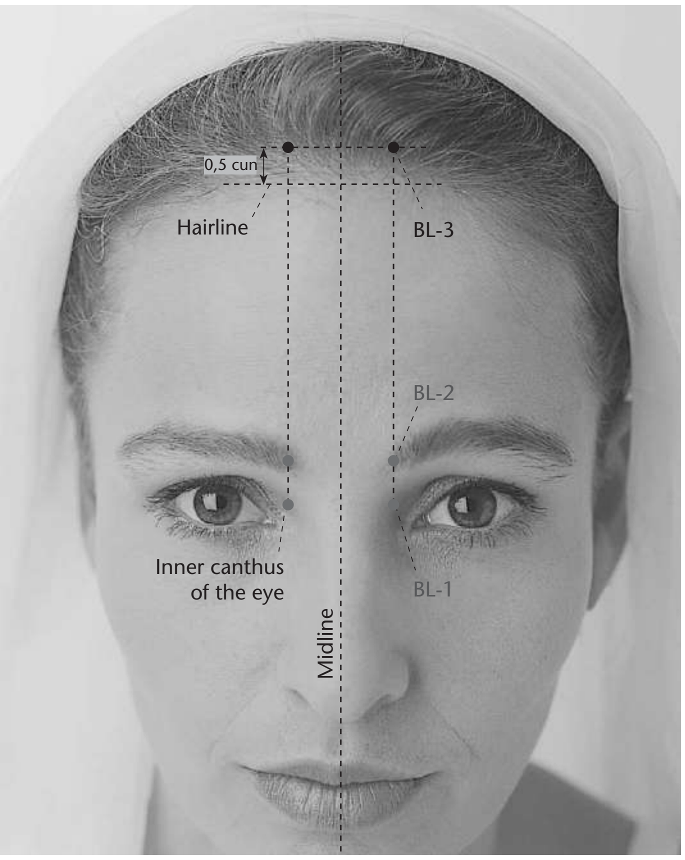
0,5 cun
Hairline
BL-3
BL-2
Inner canthus of the eye
BL-1
Midline

Location

0.5 cun within the anterior hairline, vertically above the medial canthus of the eye.

How to find

BL-3 is located directly above the medial canthus of the eye (→ **BL-1**) and 0.5 cun within the anterior hairline (→ 3.1.1). Located on the same level (0.5 cun within the anterior hairline) are → **Du-24** (on the midline), → **G.B.-15** (on the pupil line), → **G.B.-13** (3 cun lateral to the midline) and → **ST-8** (on the corner of the forehead).

Needling

0.5–1 cun obliquely in a dorsal direction or transversely (subcutaneously). According to some classic texts, moxibustion is contraindicated.

Actions/Indications

- Expels Wind and clears Heat from the head and forehead

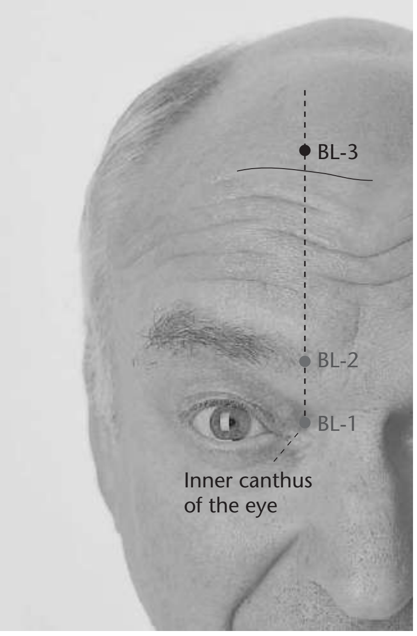
BL-3
BL-2
BL-1
Inner canthus of the eye

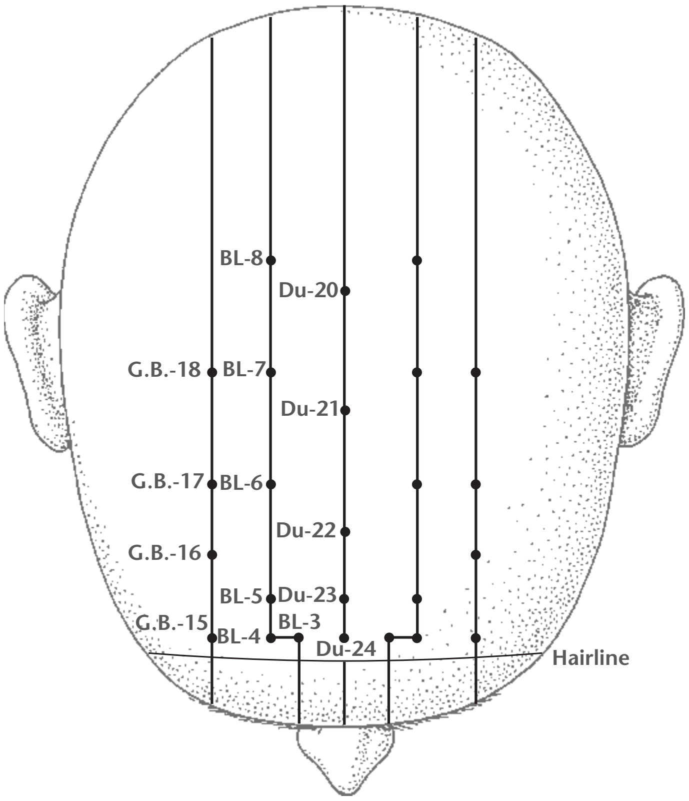
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

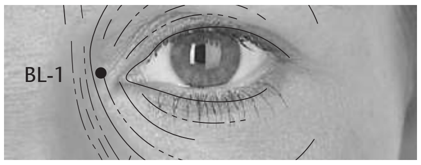

BL-4 *Crooked Curve* QUCHA

Location

0.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline (or one third of the distance between → **Du-24** and → **ST-8**).

How to find

While → **BL-3** is still located on a vertical line through the inner canthus of the eye, **BL-4** is located slightly more laterally (hence its name). The distance of 1.5 cun from the midline refers to the proportional cun measurements along the anterior hairline (→ 3.1.1): the distance between → **Du-24** (on the midline) and → **ST-8** (on the corner of the forehead) is 4.5 proportional cun; therefore, the distance between **BL-4** and the midline is one third of this distance.

Located on the same level (0.5 cun within the anterior hairline) are → **Du-24** (on the midline), → **BL-3** (superior to the inner canthus of the eye), → **G.B.-15** (on the pupil line), → **G.B.-13** (3 cun lateral to the midline) and → **ST-8** (on the corner of the forehead).

Needling

Transversely (subcutaneously) 0.5–1 cun

Actions/Indications

- Expels Wind and Wind-Heat, especially from the eyes and the nose
- Clears the head, opens the orifices, alleviates pain

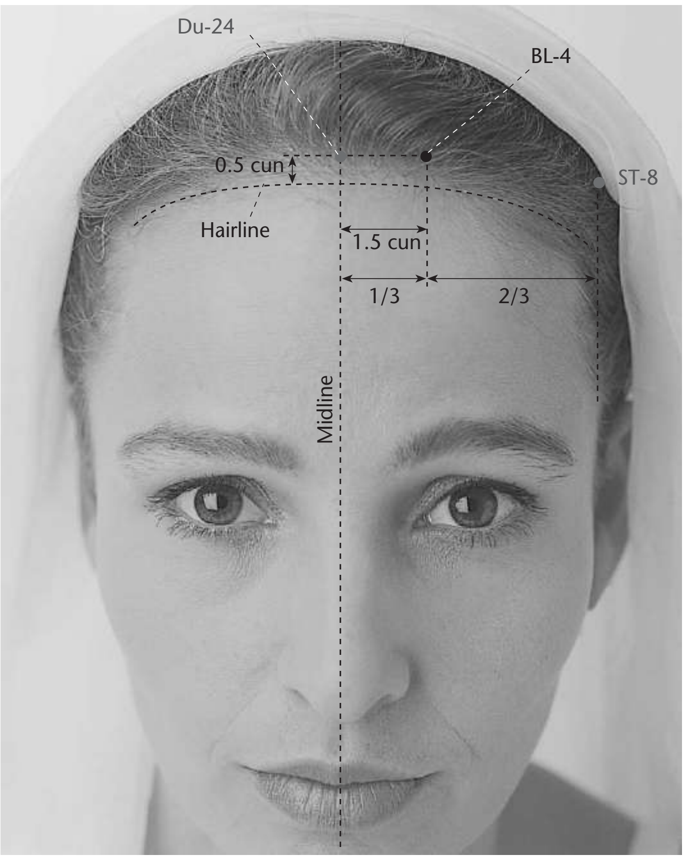
Du-24
BL-4
0.5 cun
ST-8
Hairline
1.5 cun
1/3
2/3
Midline

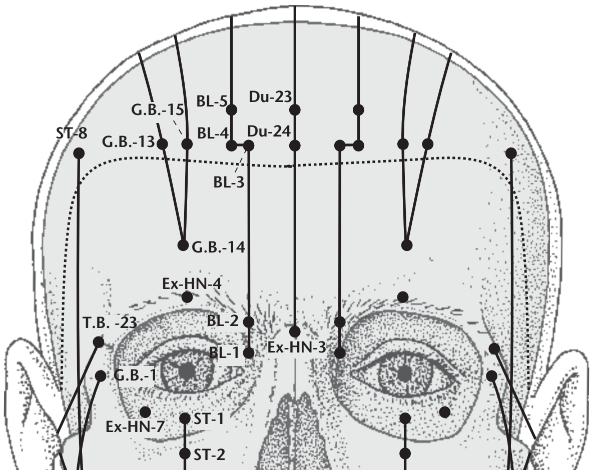
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B. -23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2

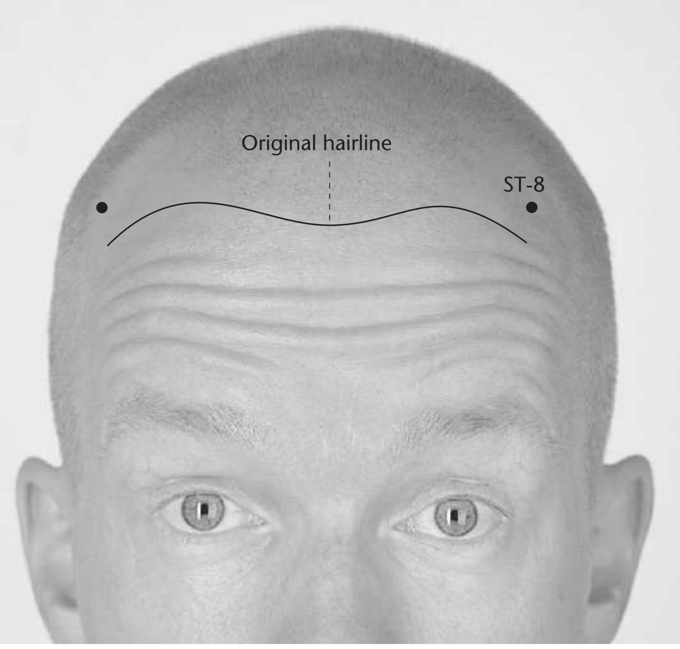
Original hairline
ST-8

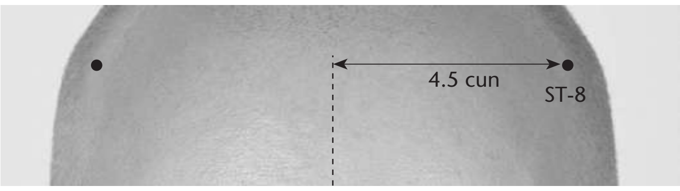

Fifth Place WUCHU BL-5

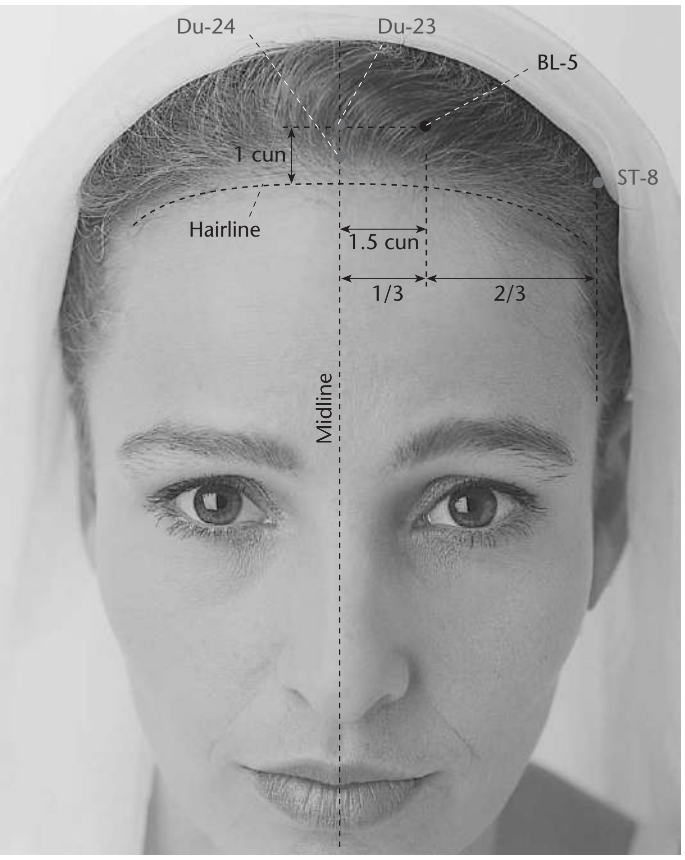
Du-24
Du-23
BL-5
1 cun
ST-8
Hairline
1.5 cun
1/3
2/3
Midline

Location

1 cun superior to the anterior hairline and 1.5 cun lateral to the midline or one third of the distance between → **Du-24** and → **ST-8**.

How to find

The distance of 1.5 cun refers to the proportional cun measurement for the distance of 4.5 cun between → **Du-24** (midline, 0.5 cun superior to the anterior hairline) and → **ST-8** (corner of the forehead), (→ 2.2). By dividing this distance into thirds, **BL-5** can be located one third of the distance from → **Du-24** and 1 cun superior to the anterior hairline.
Located on the same level is → **Du-23** (on the midline).

Needling

Transversely (subcutaneously) 0.5–1 cun. Moxibustion is contraindicated according to some texts.

Actions/Indications

- Releases the Exterior, expels Wind and Heat from the head and eyes
- Opens the channel and the *luo* vessels
- Opens the orifices (especially the eyes) and revives consciousness

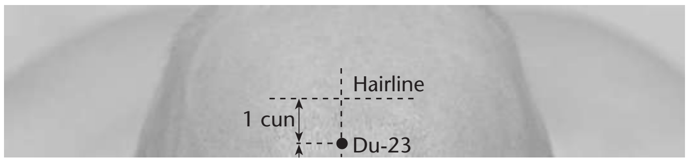
Hairline
1 cun
Du-23

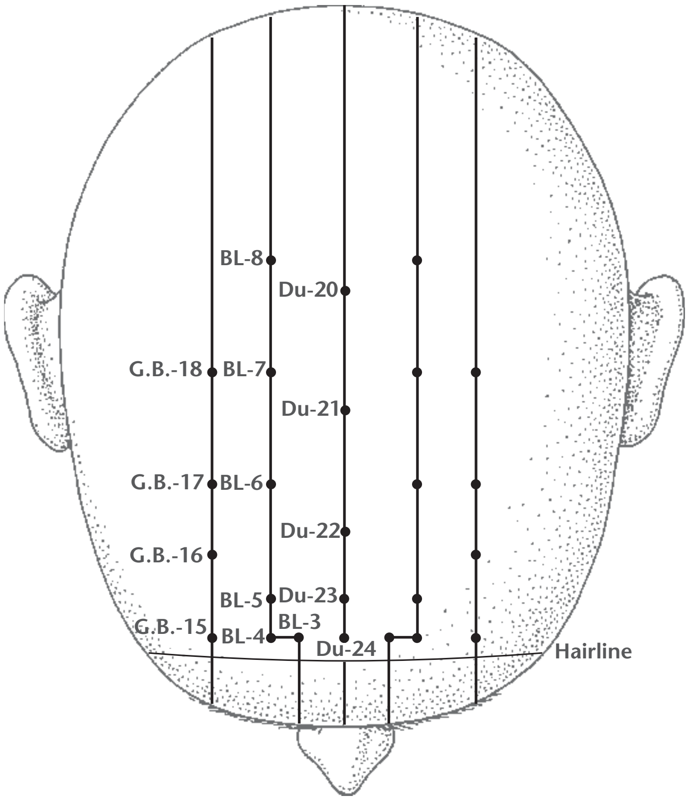
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
G.B.-15
BL-3
BL-4
Du-24
Hairline

BL-6 *Receiving Light* CHENGGUAN

Location

2.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline or one third of the distance between ➞ **Du-24** and ➞ **ST-8**.

How to find

Spreading hands technique (➞ 2.3.3): First, locate the anterior midline at the level of the anterior hairline (➞ 3.1.1) as well as ➞ **Du-20** (5 cun superior to the anterior hairline, on the vertex, at the junction of the midline with a line connecting the apices of the ears). Next, determine the midpoint between those two reference points (➞ 2.5 cun superior to the anterior hairline). From this point, measure 1.5 cun laterally (or one third of the distance between ➞ **Du-24** and ➞ **ST-8**) to locate **BL-6**.

Located on the same level (➞ 2.5 cun superior to the anterior hairline), but more laterally (pupil line), is ➞ **G.B.-17**.

Needling

Transversely (subcutaneously) 0.5–1 cun. Moxibustion only following careful diagnosis.

Actions/Indications

- Expels Wind, especially from the head and eyes
- Clears Heat from the head, especially from the eyes and nose

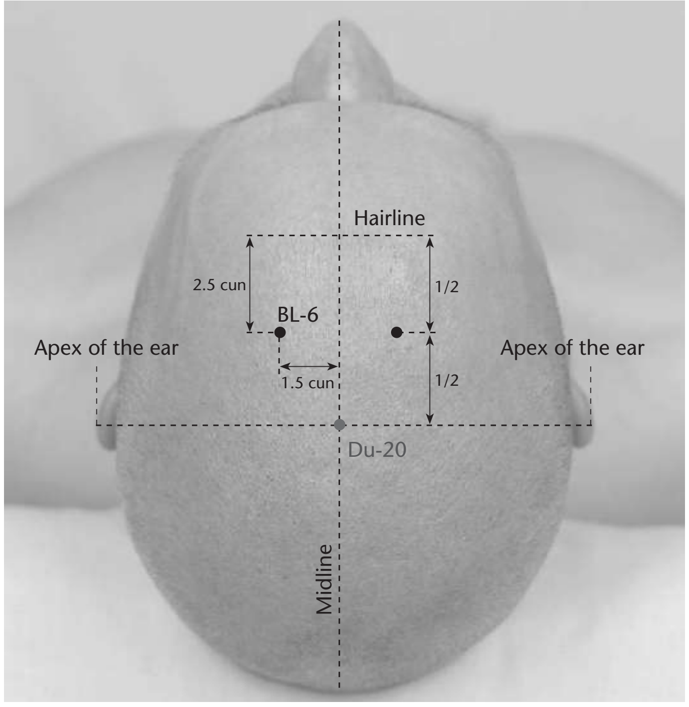
Hairline
2.5 cun
1/2
BL-6
Apex of the ear
Apex of the ear
1.5 cun
1/2
Du-20
Midline

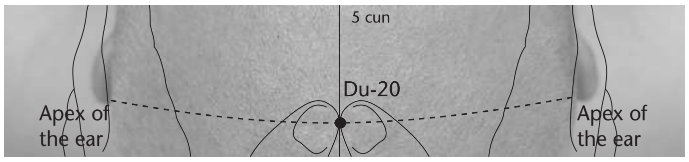
5 cun
Du-20
Apex of the ear
Apex of the ear

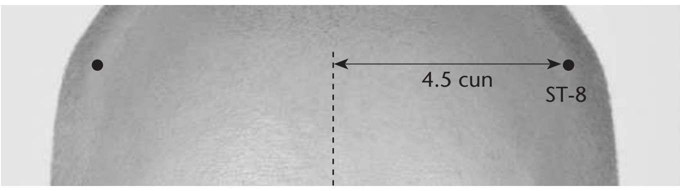

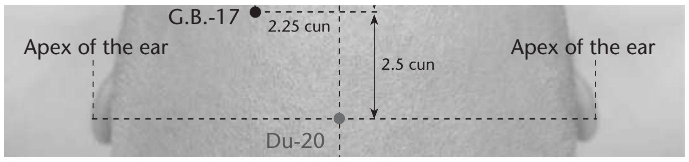
G.B.-17
2.25 cun
Apex of the ear
Apex of the ear
2.5 cun
Du-20

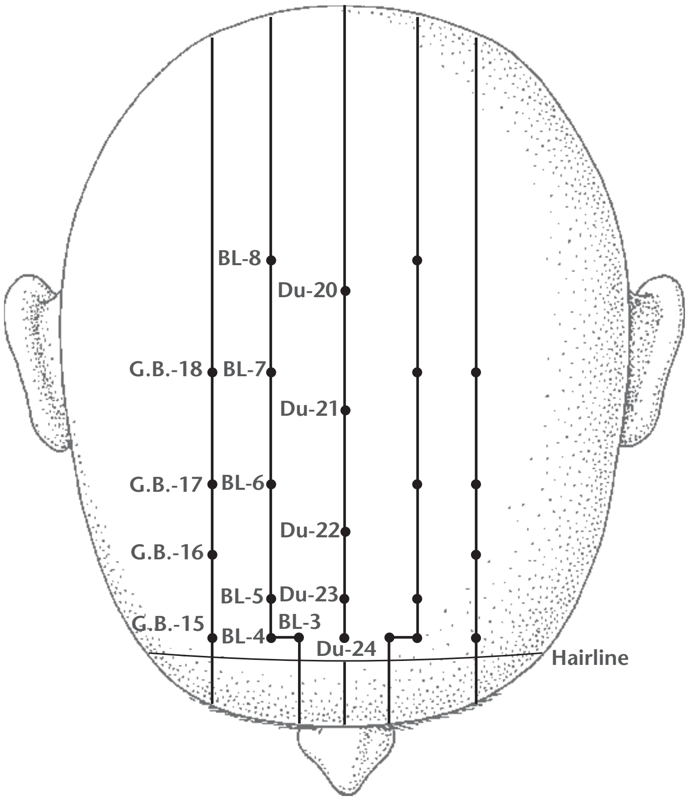
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
G.B.-15
BL-3
BL-4
Du-24
Hairline

Heavenly Connection TONGTIAN **BL-7**

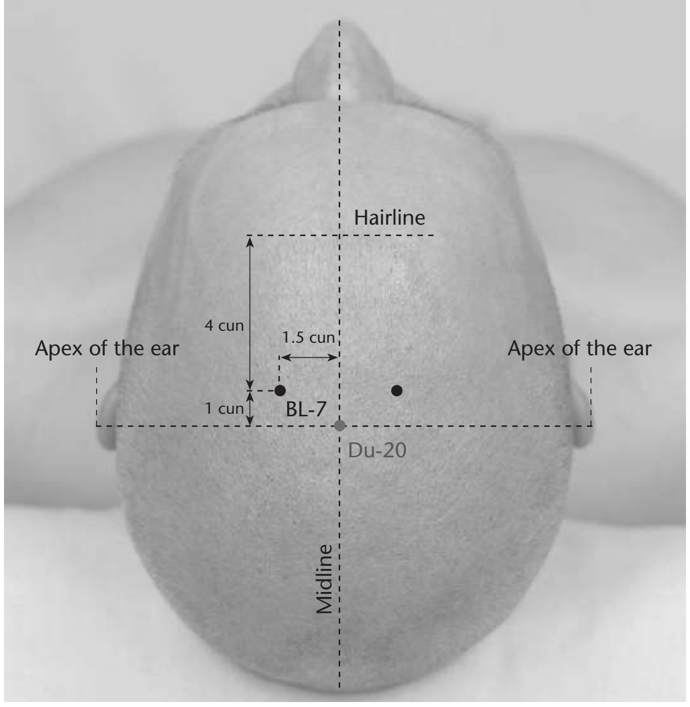
Hairline
4 cun
1.5 cun
Apex of the ear
Apex of the ear
1 cun
BL-7
Du-20
Midline

Location

1.5 cun lateral to the midline and 4 cun superior to the anterior hairline or 1 cun anterior to → **Du-20**.

How to find

BL-7 is best located by using → **Du-20** (5 cun superior to the anterior hairline, on the vertex, at the junction of the midline with a line connecting the apices of the ears) as a reference point. Measure 1 cun anterior from → **Du-20** and 1.5 cun lateral to the midline (or one third of the distance between → **Du-24** and **ST-8,** → 2.2) to locate **BL-7**. Located on the same level (4 cun superior to the anterior hairline), but more laterally (pupil line), is → **G.B.-18**. → **Du-21** is located on the midline, 1.5 cun anterior to → **Du-20**.

Needling

Transversely (subcutaneously) 0.5–1 cun

Actions/Indications

- Clears the head, benefits especially the nose

Special features

Important point for treating disorders of the nose

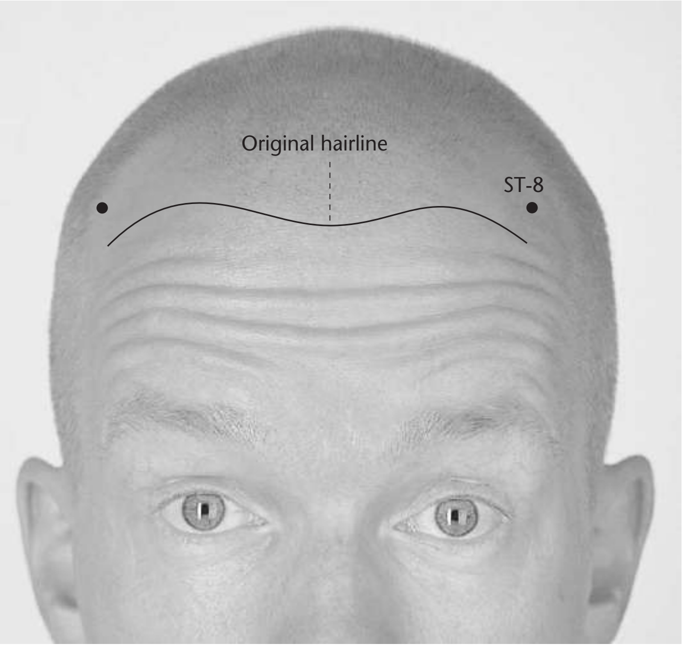
Original hairline
ST-8

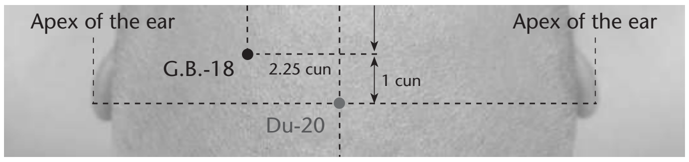
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
Du-20

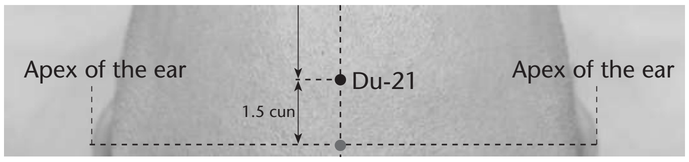
Apex of the ear
Du-21
Apex of the ear
1.5 cun

BL-8 *Declining Connection* LUOQUE

Location

1.5 cun lateral to the midline and 5.5 cun superior to the anterior hairline or 0.5 cun posterior to → **Du-20**.

How to find

BL-8 is best located by using → **Du-20** as a reference point. → **Du-20** is located at the junction of the vertical midline with a line connecting the apices of the ears, 5 cun superior to the anterior hairline (→ 3.1.1). To locate **BL-8**, measure 0.5 cun from → **Du-20** in a posterior direction and 1.5 cun in a lateral direction.

Needling

Transversely (subcutaneously) 0.5–1 cun. Moxibustion possible.

Actions/Indications

- Opens the orifices
- Subdues Wind
- Transforms Phlegm
- Calms the *shen*

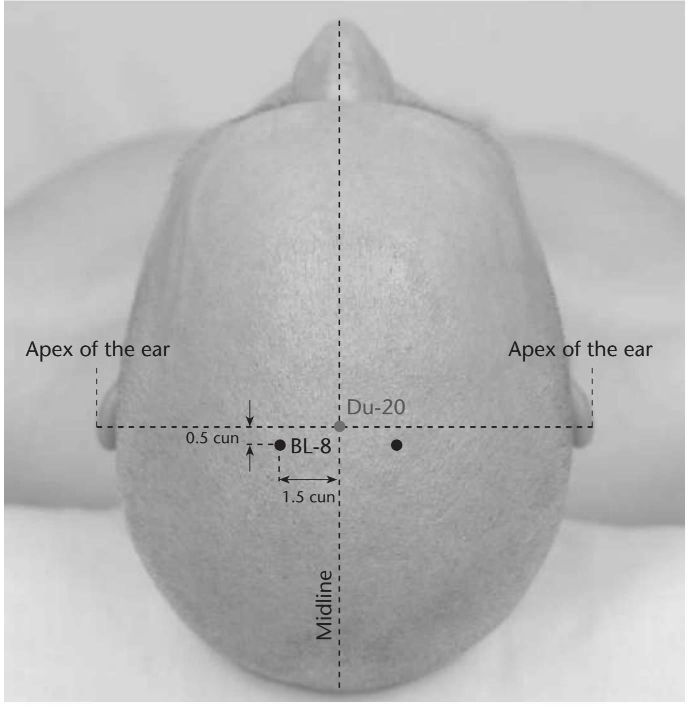
Apex of the ear
Apex of the ear
Du-20
0.5 cun
BL-8
1.5 cun
Midline

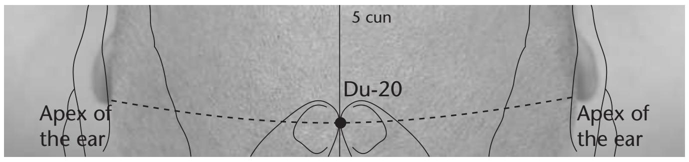
5 cun
Du-20
Apex of the ear
Apex of the ear

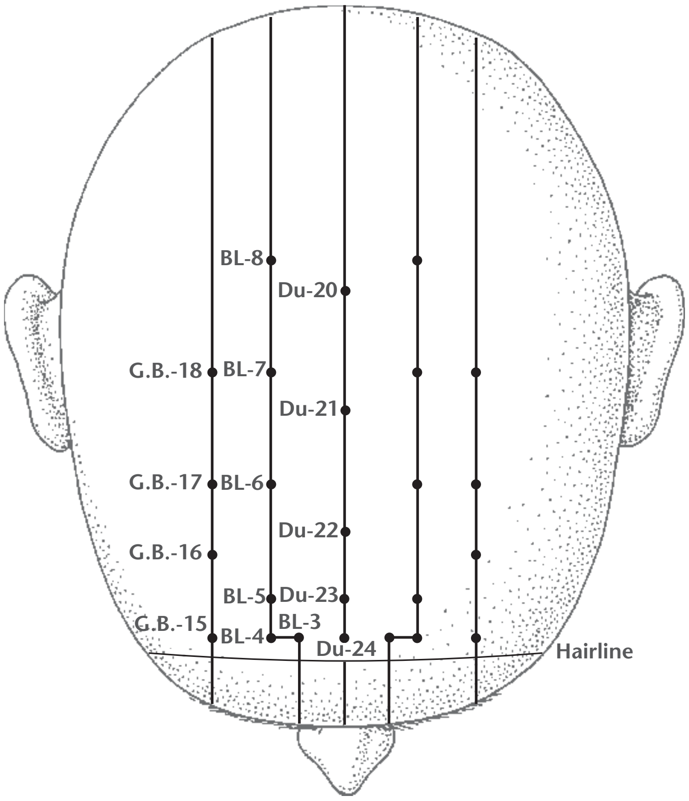
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

Jade Pillow YUZHEN BL-9

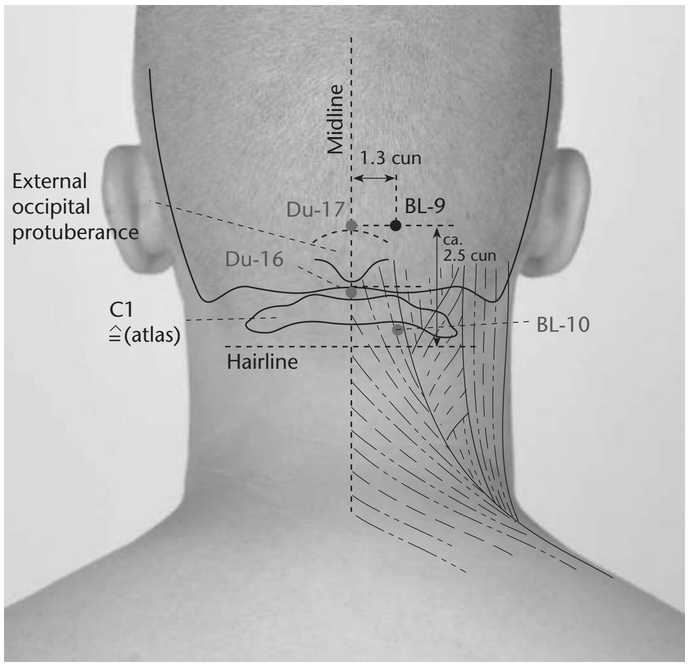
Midline
1.3 cun
External occipital protuberance
Du-17
BL-9
ca. 2.5 cun
Du-16
C1 ≙(atlas)
BL-10
Hairline

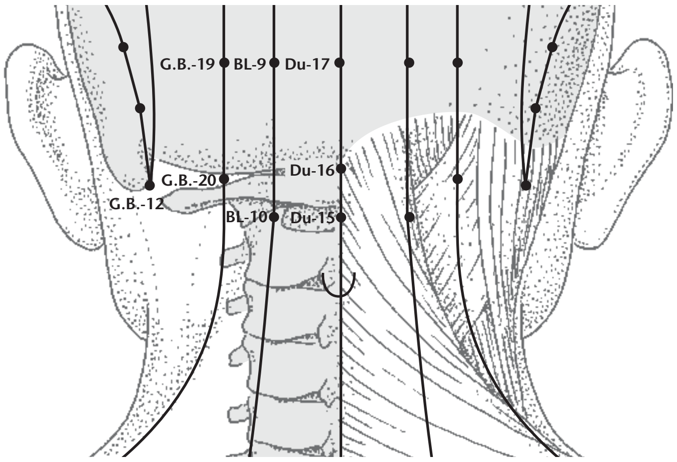
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

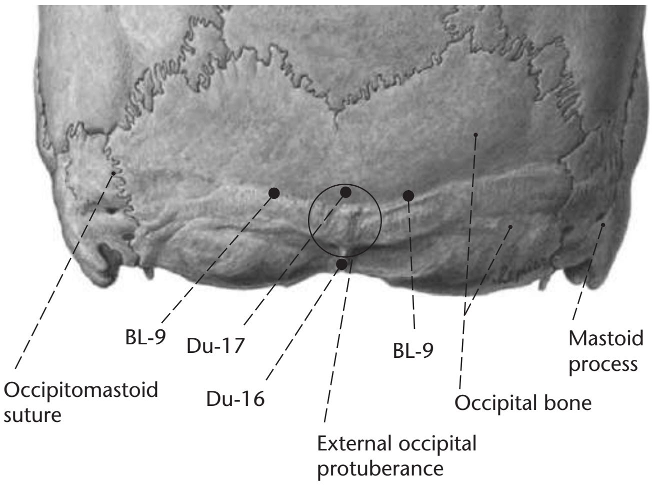
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

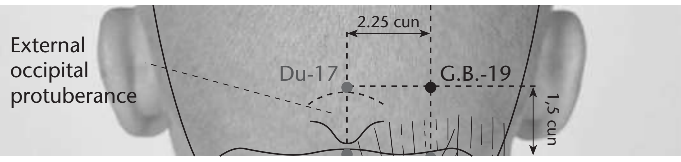
External occipital protuberance
2.25 cun
Du-17
G.B.-19
1,5 cun

Location

2.5 cun superior to the posterior hairline and 1.3 cun lateral to the midline or → **Du-17** (directly superior to the external occipital protuberance).

How to find

To locate **BL-9,** → **Du-17** is a good reference point, located on the midline, in the depression directly superior to the upper border of the external occipital protuberance (→ 3.1.5). From there, measure 1.3 cun in a lateral direction to locate **BL-9**, which is generally located on an imaginary vertical line through → **BL-10.**
Or: Locate the posterior hairline (often not very accurate → 3.1.5) and, on the midline, measure 2.5 cun in a superior direction, then 1.3 cun towards lateral.
Located on the same level are → **Du-17** (midline, in the depression at the upper border of the external occipital protuberance) and → **G.B.-19** (more laterally, on a vertical line through → **G.B.-20**).

Needling

Transversely (subcutaneously) 0.5–1 cun. Moxibustion possible.

Actions/Indications

- Expels Wind and Cold
- Opens the channel and *luo* vessels, alleviates pain
- Benefits the eyes and nose

Special features

In Qigong practice, **BL-9** plays an important role as one of the three gates (*san guan*).

BL-10 *Heavenly Pillar* TIANZHU

Location

Approximately 1.3 cun lateral to → **Du-15**, where the trapezius muscle inserts on the lower border of the occiput, close to where the major occipital nerve emerges.

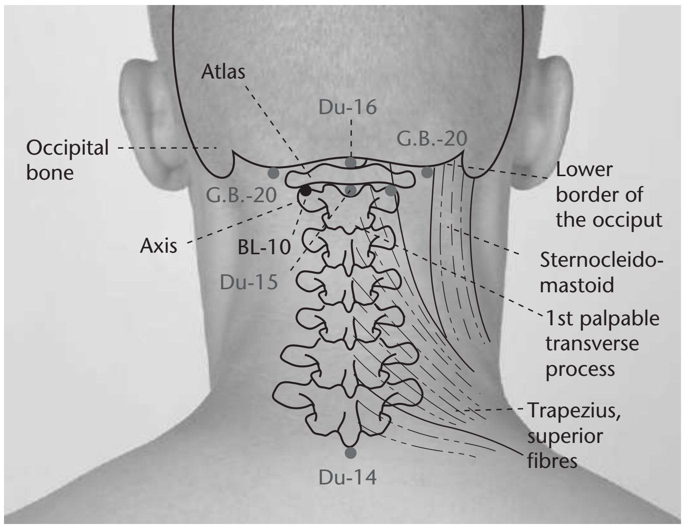
Atlas
Du-16
G.B.-20
Occipital bone
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid
Du-15
1st palpable transverse process
Trapezius, superior fibres
Du-14

How to find

With the patient's head in an upright and relaxed position, first locate → **G.B.-20** below the lower border of the occiput and between the insertions of the sternocleidomastoid and trapezius muscles. From there, palpate in a medial and inferior direction at a 45° angle, until the finger comes to a halt against the bulge of the trapezius muscle. This is the location of **BL-10**. For reference: If the practitioner places their left middle finger on → **G.B.-20**, the (shorter) index finger will be pointing to **BL-10**. Located on the same level is → **Du-15** (approximately 1.3 cun lateral to the midline and superior to the axis (first palpable spinous process)).

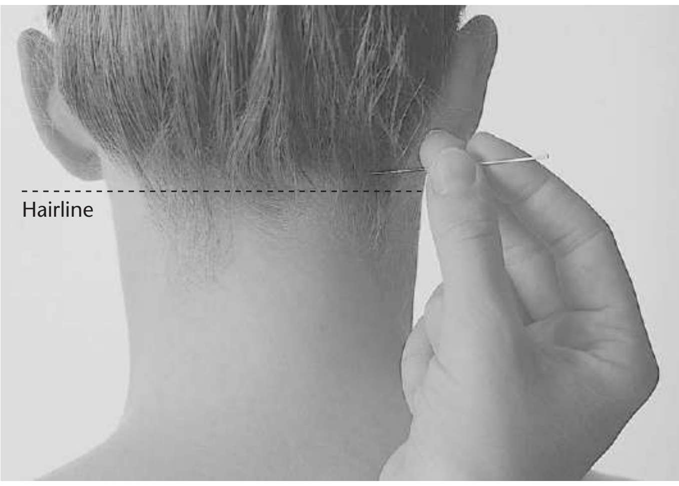

Needling

0.5–1 cun vertically or slightly obliquely in a dorsal direction. Caution: Do not needle in a superior direction.

Actions/Indications

- Regulates the Qi, calms Wind and the *shen*, benefits the head and sensory orifices
- Opens the channel
- Strengthens the lower back
- Expels Wind

Special features

Window of Heaven point. Important point for expelling (internal and external) Wind. In contrast to → **G.B.-20**, **BL-10** is used in clinical practice mainly for pacifiying internal Wind.

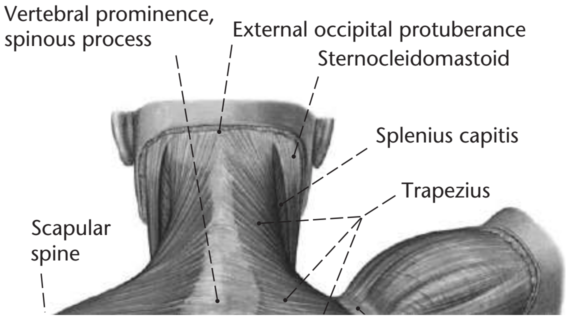
Vertebral prominence, spinous process
External occipital protuberance
Sternocleidomastoid
Splenius capitis
Trapezius
Scapular spine

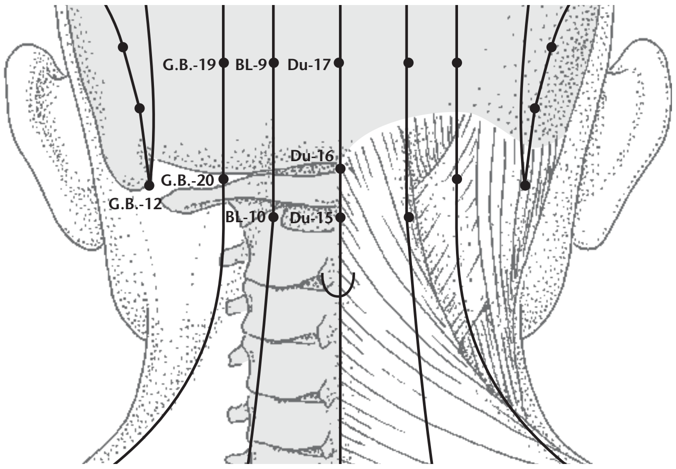
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

Great Shuttle DAZHU BL-11

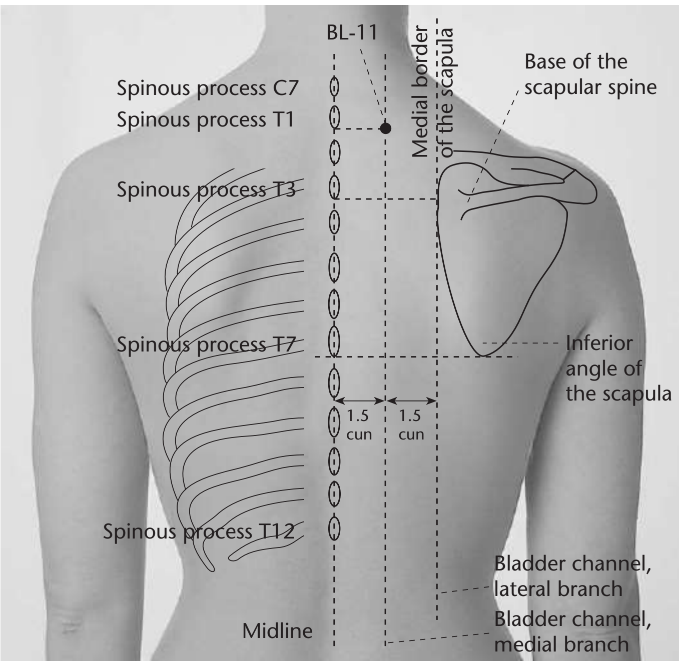
BL-11
Medial border of the scapula
Base of the scapular spine
Spinous process C7
Spinous process T1
Spinous process T3
Spinous process T7
Inferior angle of the scapula
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

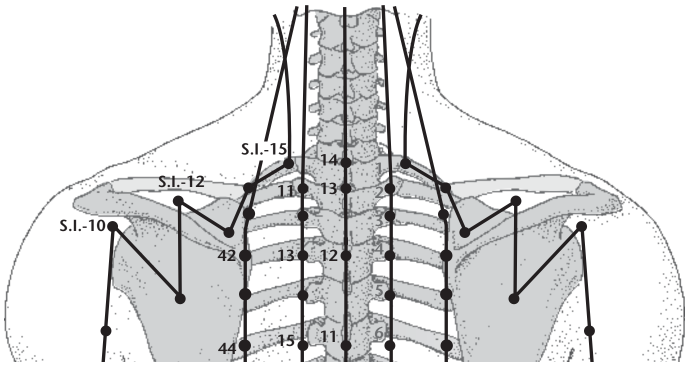
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11

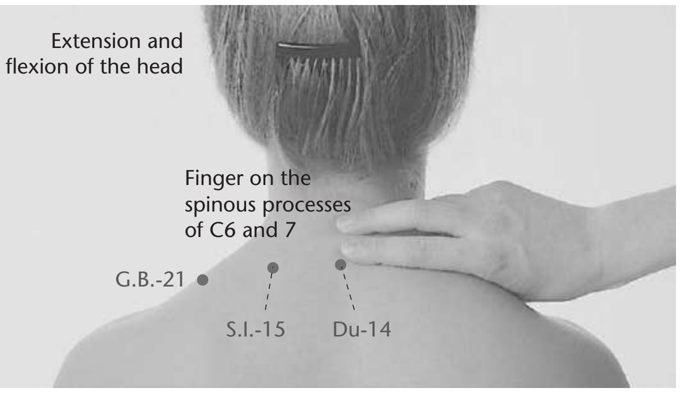
Extension and flexion of the head
Finger on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

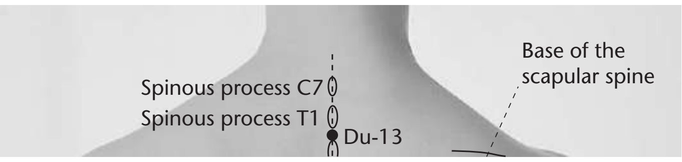
Base of the scapular spine
Spinous process C7
Spinous process T1
Du-13

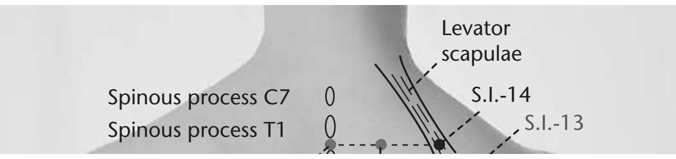
Levator scapulae
Spinous process C7
Spinous process T1
S.I.-14
S.I.-13

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 1st thoracic vertebra (T1, with hanging shoulders on the level of the acromion).

How to find

First, locate the spinous process of C7 (→ 3.4.1). The next spinous process inferior to this belongs to the T1. Locate **BL-11** 1.5 cun lateral to the lower border of T1, at the highest point of the paraspinal musculature.

Located on the same level are → **Du-13** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **S.I.-14** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine or transversely (subcutaneously). Caution: Pneumothorax.

Actions/Indications

- Expels pathogenic factors (Wind-Heat)
- Regulates the Lung Qi
- Opens the channel, benefits the bones and joints

Special features

Hui-meeting point with the bones; meeting point with the S.I. channel, according to some authors also with the T.B., G.B. channels and the *du mai*; Sea of Blood point.

BL-12 *Wind Gate* FENGMEN

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd thoracic vertebra (T2).

How to find

First, locate the spinous process of C7 (→ 3.4.1). From there, count down two spinous processes to the lower border of the spinous process of T2. Measure 1.5 cun laterally to locate **BL-12** on the highest point of the paraspinal musculature.
Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline), → **BL-41** (3 cun lateral to the midline) and → **S.I.-13** (on the scapula, on the medial end of the scapular spine).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax. For acute febrile colds, use reducing needling techniques or cupping. For susceptibility to catching colds, use tonifying needling techniques and/or moxibustion.

Actions/Indications

- Expels (external) Wind and releases the Exterior
- Spreads and descends the Lung Qi
- Tonifies the defensive Qi (*wei qi*), stabilises the Exterior
- Benefits the nose
- Opens the channel

Special features

Meeting point with the *du mai*. Important point for expelling pathogenic factors, especially Wind.

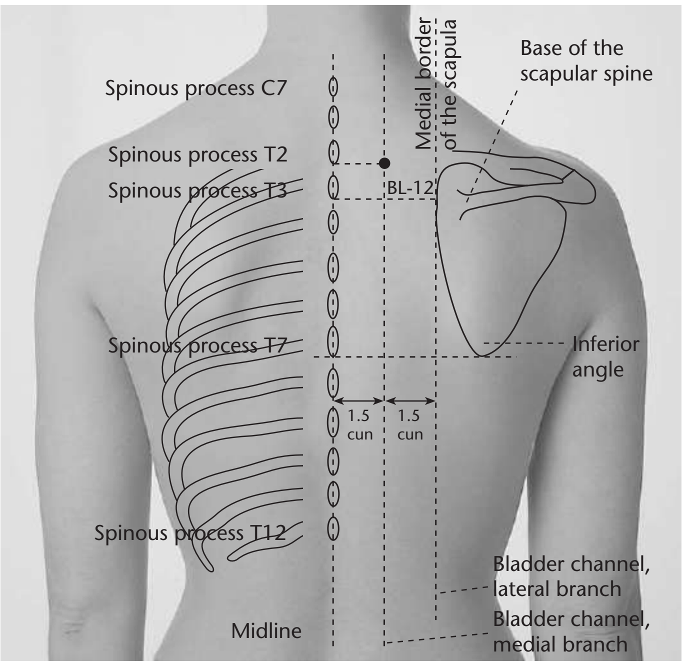
Spinous process C7
Spinous process T2
Spinous process T3
Medial border of the scapula
Base of the scapular spine
BL-12
Spinous process T7
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

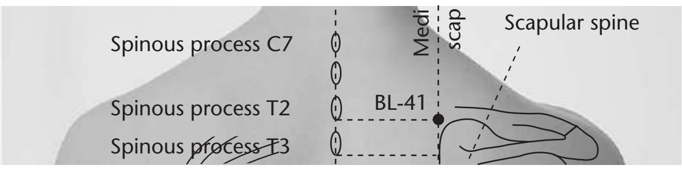
Spinous process C7
Spinous process T2
Spinous process T3
Medi
scap
Scapular spine
BL-41

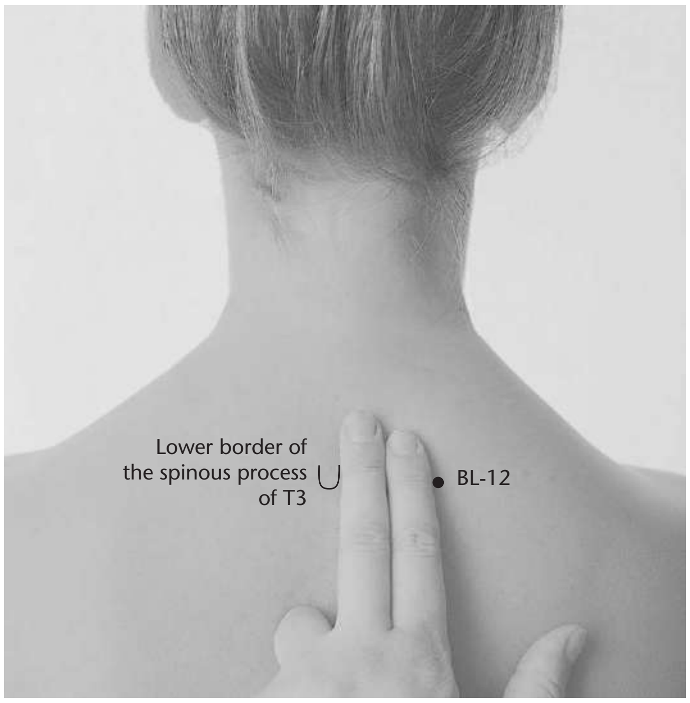
Lower border of the spinous process of T3
BL-12

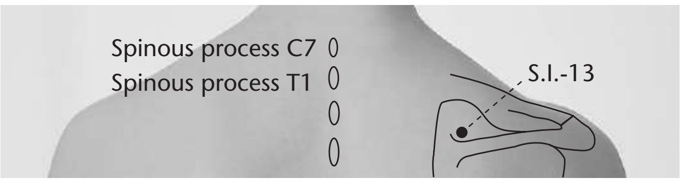
Spinous process C7
Spinous process T1
S.I.-13

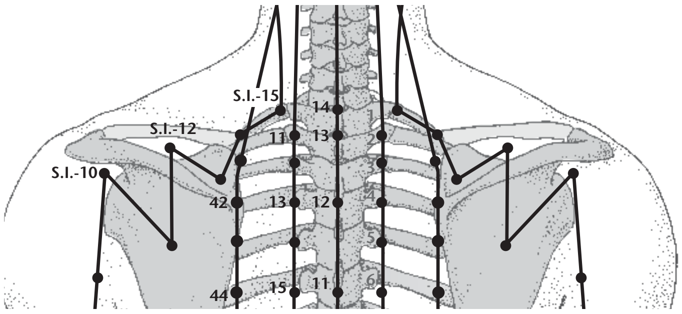
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11

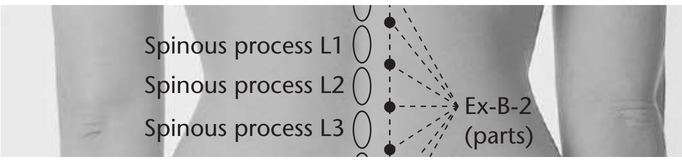
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Lung Shu FEISHU BL-13

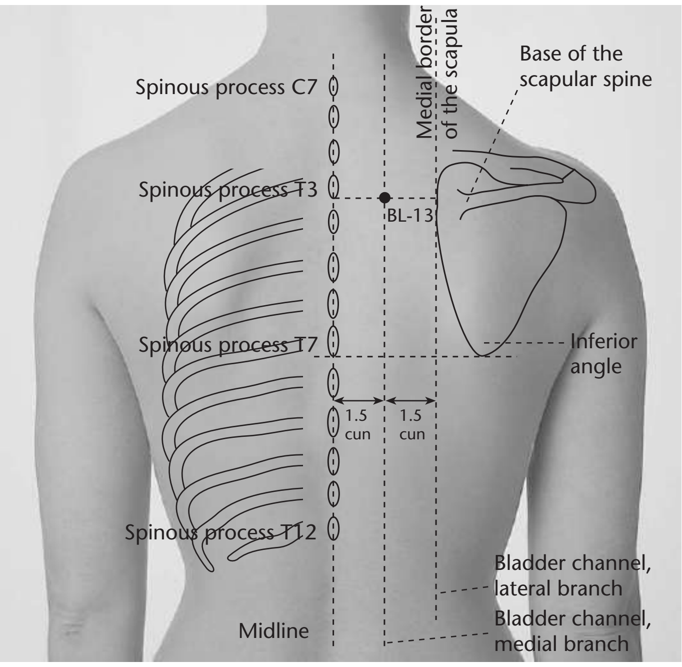
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
BL-13
Spinous process T7
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

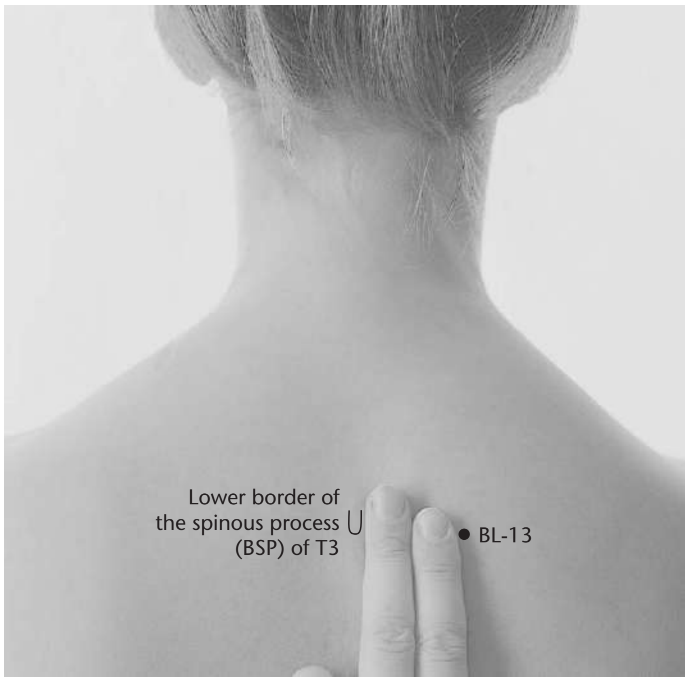
Lower border of the spinous process (BSP) of T3
BL-13

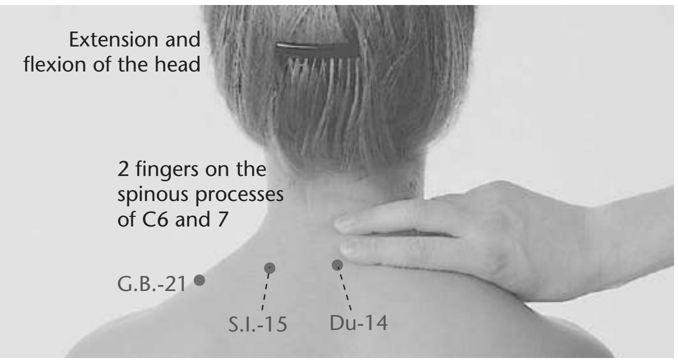
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 3rd thoracic vertebra (T3).

How to find

First, locate the spinous process of C7 (→ 3.4.1). From there, count down three spinous processes to the lower border of the spinous process of the T3. Measure 1.5 cun laterally to locate **BL-13** on the highest point of the paraspinal musculature. **Or:** With the patient seated and their arms hanging down, the spinous process of the T3 is generally on the level of the medial end of the scapular spine (→ 3.4.2). Locate **BL-13** 1.5 cun lateral to the midline.

Located on the same level are → **Du-12** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and **BL-42** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax. For chronic Lung Qi deficiency without Heat signs, use moxibustion.

Actions/Indications

- Tonifies, spreads and descends the Lung Qi, nourishes the Lung Yin
- Clears Heat from the Lung
- Releases the Exterior
- Opens the channel, alleviates pain

Special features

Back-*shu* point of the lung. Major point for all disorders of the Lung (Lung deficiency and excess patterns).

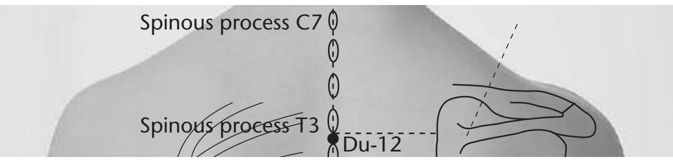
Spinous process C7
Spinous process T3
Du-12

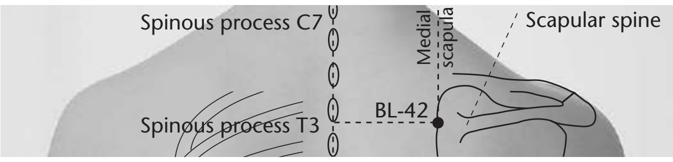
Spinous process C7
Medial scapula
Scapular spine
Spinous process T3
BL-42

BL-14 Jueyin Shu JUEYINSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 4th thoracic vertebra (T4).

How to find

First, locate the spinous process of C7 (➞ 3.4.1). From there, count down four spinous processes to the lower border of the spinous process of T4. Measure 1.5 cun laterally to locate **BL-14** on the highest point of the paraspinal musculature. **Or:** With the patient seated and their arms hanging down, the spinous process of T3 is generally on the level of the medial end of the scapular spine (➞ 3.4.2). The next spinous process in an inferior direction belongs to the T4. Locate **BL-14** 1.5 cun lateral to the midline. Located on the same level are a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-43** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine or transversely (subcutaneously). Caution: Pneumothorax.

Actions/Indications

- Regulates and tonifies the Heart
- Unbinds the chest, regulates the Qi
- Calms the *shen*

Special features

Back-*shu* point of the Pericardium

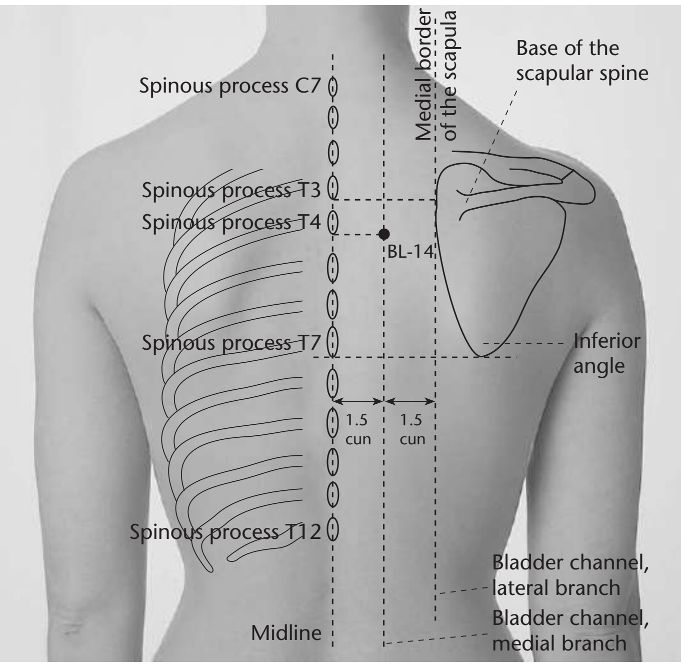
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T4
BL-14
Spinous process T7
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

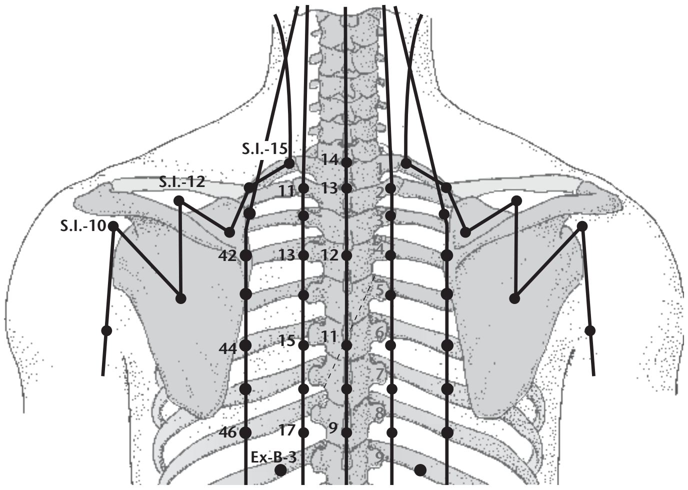
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3

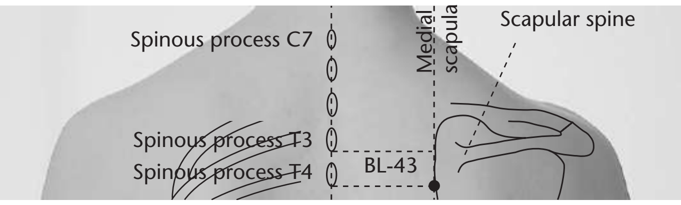
Spinous process C7
Medial scapula
Scapular spine
Spinous process T3
Spinous process T4
BL-43

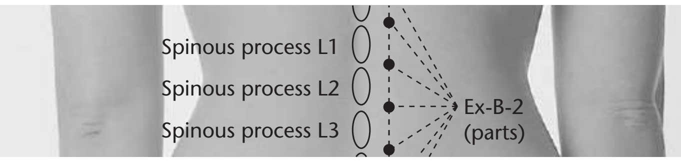
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

***Heart* Shu XINSHU BL-15**

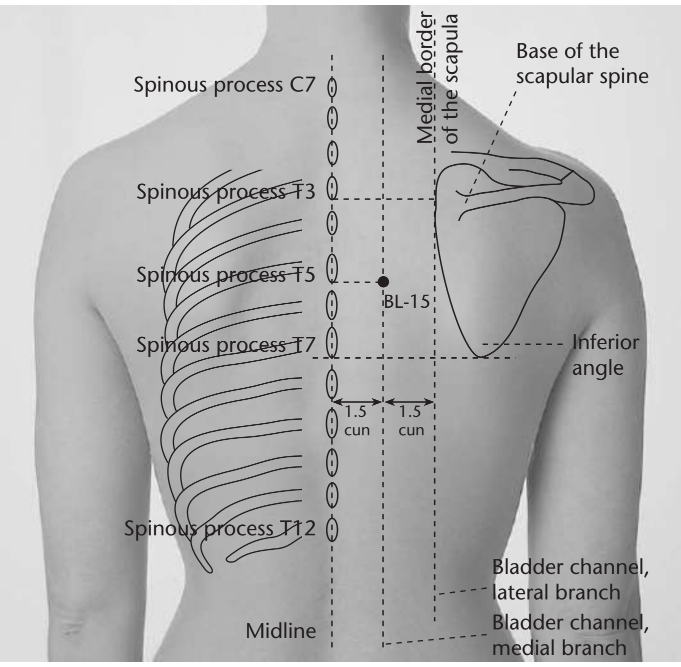
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T5
BL-15
Spinous process T7
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

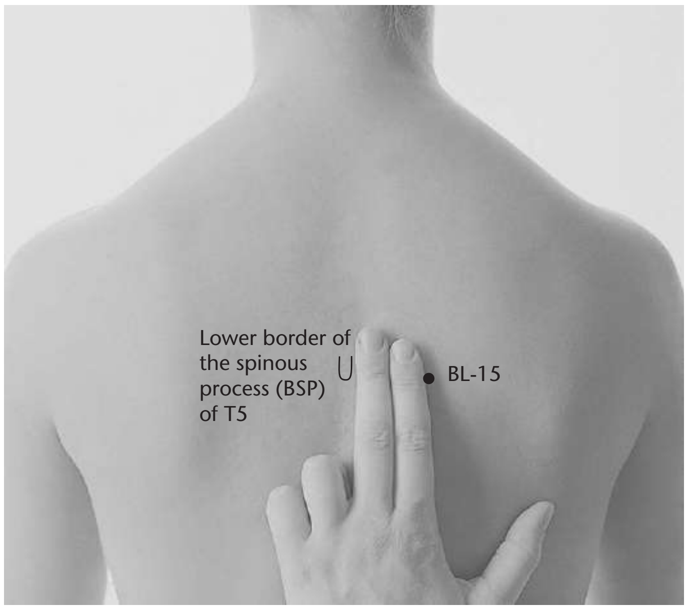
Lower border of the spinous process (BSP) of T5
BL-15

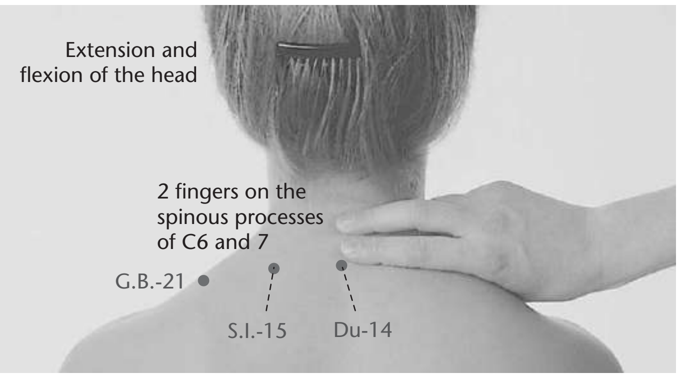
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 5th thoracic vertebra (T5).

How to find

First, locate the spinous process of C7 (→ 3.4.1). From there, count down five spinous processes to the lower border of the spinous process of T5. Measure 1.5 cun laterally to locate **BL-15** on the highest point of the paraspinal musculature. **Or:** With the patient seated and their arms hanging down, the spinous process of the T3 is generally on the level of the easily palpable medial end of the scapular spine (→ 3.4.2). From there, count down two spinous processes in an inferior direction to the T5. Locate **BL-15** 1.5 cun lateral to the midline.

Located on the same level are → **Du-11** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-44** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Tonifies and nourishes the Heart, regulates the Heart Qi, unbinds the chest and removes Blood stasis, calms the *shen*, clears Fire, moves Qi locally and along the channel pathway: for example, for disorders of the thoracic spine and intercostal region

Special features

Back-*shu* point of the Heart

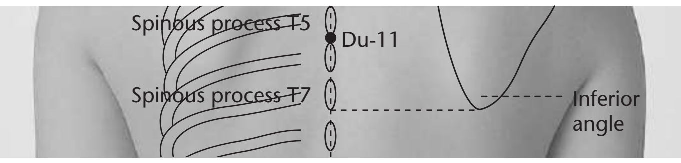
Spinous process T5
Du-11
Spinous process T7
Inferior angle

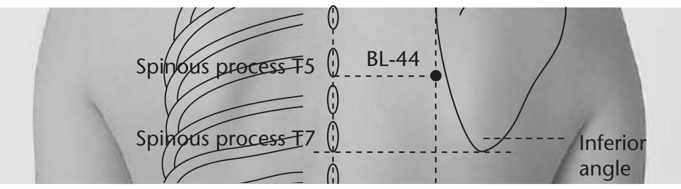
Spinous process T5
BL-44
Spinous process T7
Inferior angle

BL-16 *Governing* Shu DUSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 6th thoracic vertebra (T6).

How to find

First, locate the spinous process of the C7 (→ 3.4.1). From there, count down six spinous processes to the lower border of the spinous process of T6. Measure 1.5 cun laterally to locate **BL-16** on the highest point of the paraspinal musculature. **Or:** With the patient standing and their arms hanging down, the spinous process of T7 is level with the inferior angle of the scapula (→ 3.4.2). The next spinous process in a superior direction belongs to T6. Locate **BL-16** 1.5 cun lateral to the midline.

Located on the same level are → **Du-10** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-45** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine or transversely (subcutaneously). Caution: Pneumothorax.

Actions/Indications

- Unbinds the chest and regulates the Qi movement in the thorax and the abdomen

Special features

Despite its designation as Back-*shu* point of the *du mai*, this point is hardly mentioned in classical or modern texts.

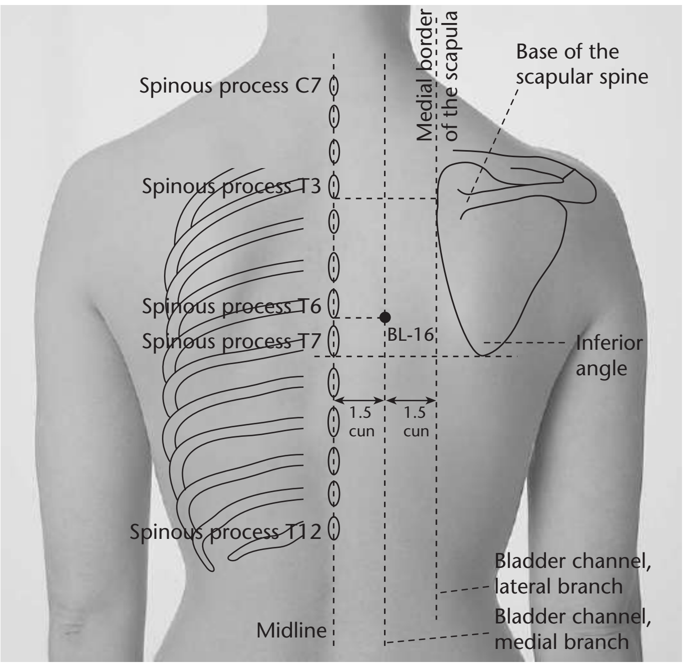
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T6
Spinous process T7
BL-16
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

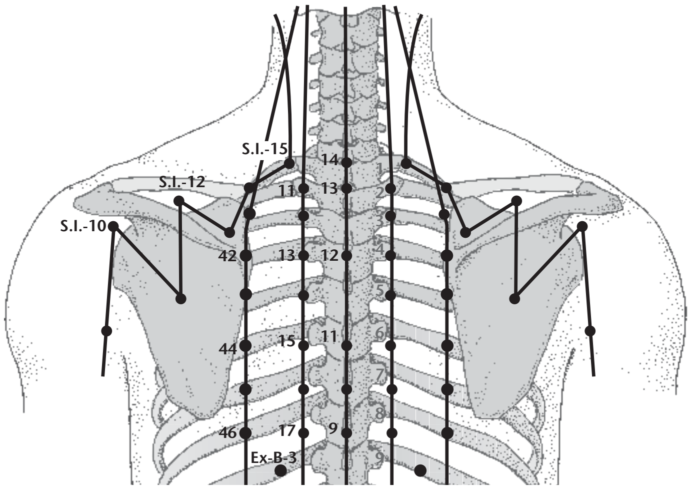
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3

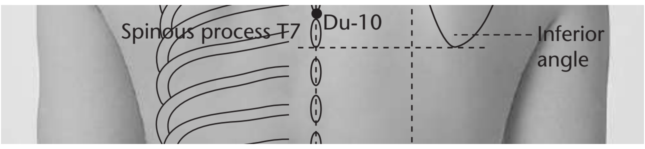
Spinous process T7
Du-10
Inferior angle

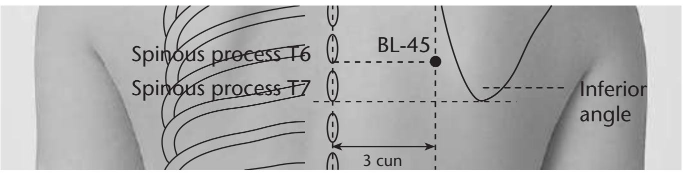
Spinous process T6
Spinous process T7
BL-45
Inferior angle
3 cun

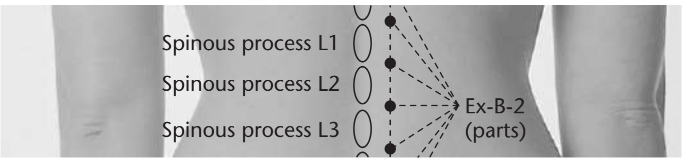
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Diaphragm Shu GESHU BL-17

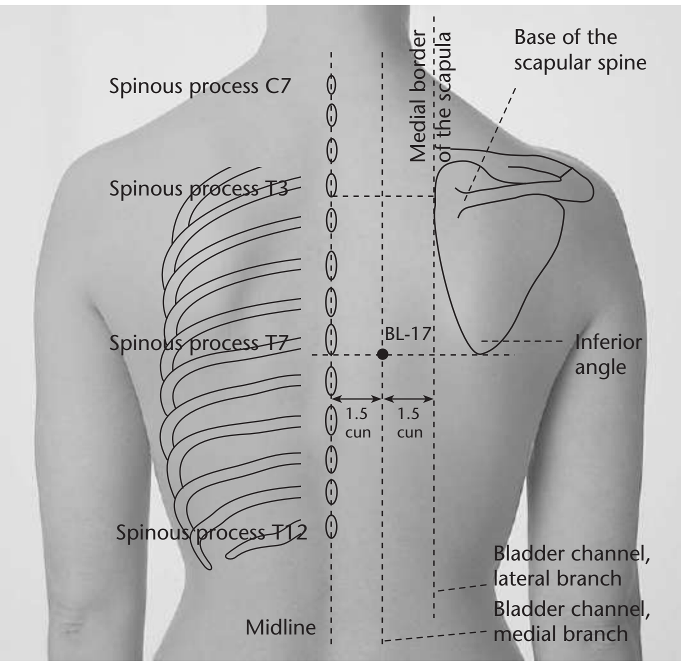
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T7
BL-17
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

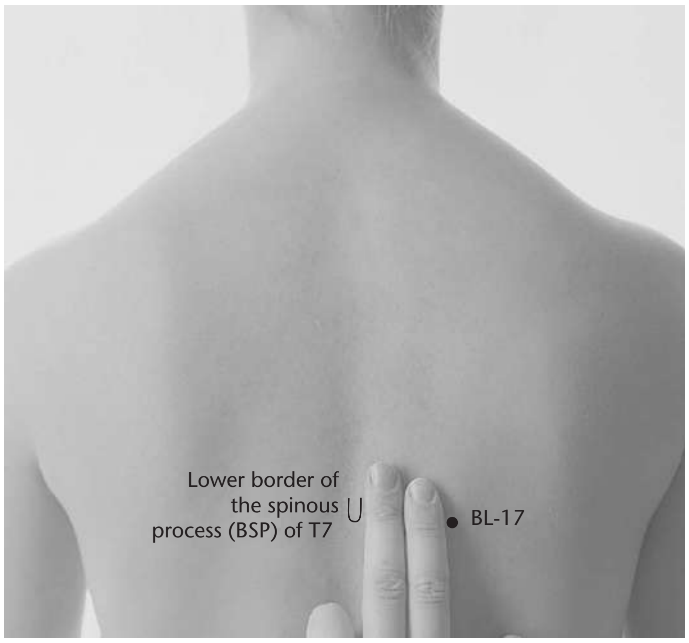
Lower border of the spinous process (BSP) of T7
BL-17

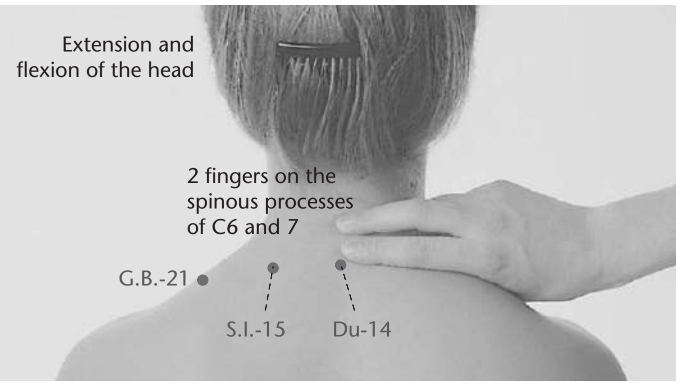
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 7th thoracic vertebra (T7).

How to find

The spinous process of the T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). **BL-17** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Use the spinous process of the T3 for reference (generally on the level of the medial end of the scapular spine) (→ 3.4.2). **Or:** Use C7 as a reference point (→ 3.4.1).
Located on the same level are → **Du-9** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-46** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Cools Blood-Heat, stops bleeding, removes Blood stasis, nourishes and harmonises Blood (and Yin)
- Regulates the diaphragm, descends counterflow Qi
- Moves Qi locally and along the channel pathway

Special features

Back-*shu* point of the Diaphragm, *hui*-meeting point of the Blood.

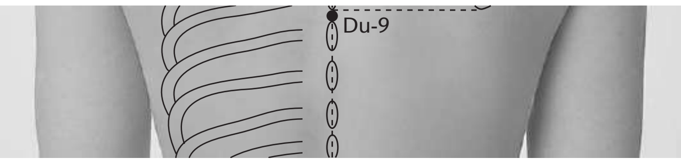

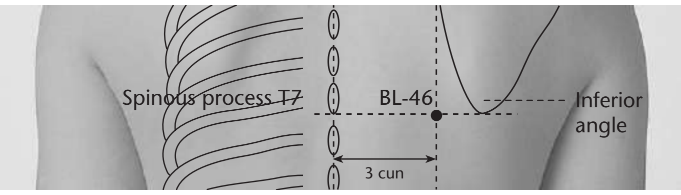
Spinous process T7
BL-46
Inferior angle
3 cun

BL-18 *Liver* Shu GANSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 9th thoracic vertebra (T9).

How to find

In a standing position with the arms hanging down, the spinous process of T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down two spinous processes to the spinous process of T9. **BL-18** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Use the C7 as a reference point (→ 3.4.1). From there, count down nine spinous processes to T9 and locate **BL-18** 1.5 cun lateral to the lower border of its spinous process.

Located on the same level are → **Du-8** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-47** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Spreads Liver Qi, cools Fire, clears Damp-Heat, regulates and nourishes Liver Blood, subdues (internal) Wind
- Benefits the eyes
- Benefits the sinews

Special features

Back-*shu* point of the Liver

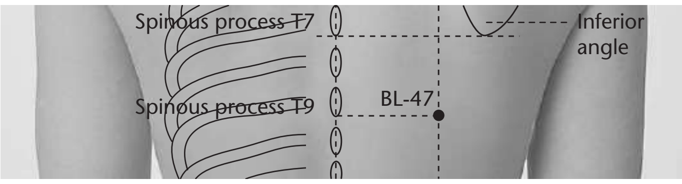
Spinous process T7
Inferior angle
Spinous process T9
BL-47

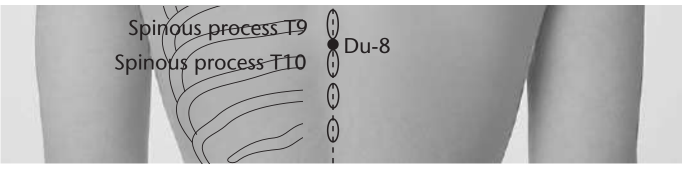
Spinous process T9
Spinous process T10
Du-8

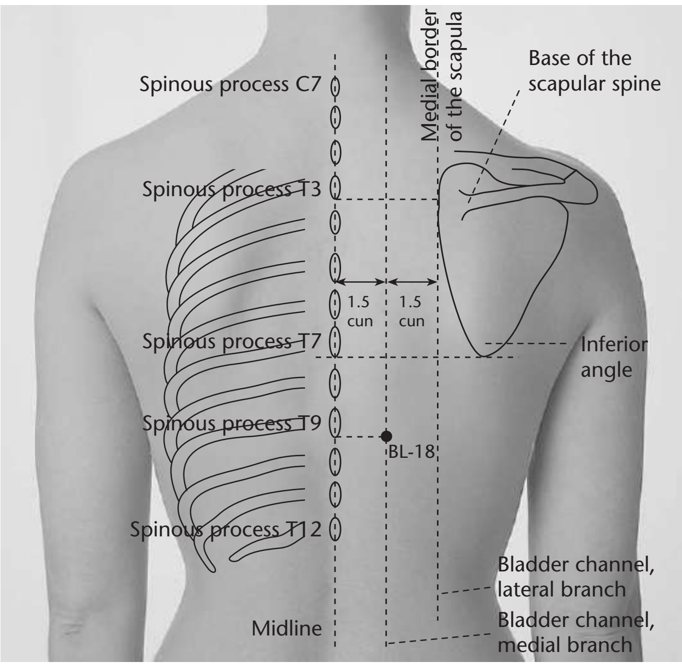
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T9
BL-18
Spinous process T12
Bladder channel, lateral branch
Midline
Bladder channel, medial branch

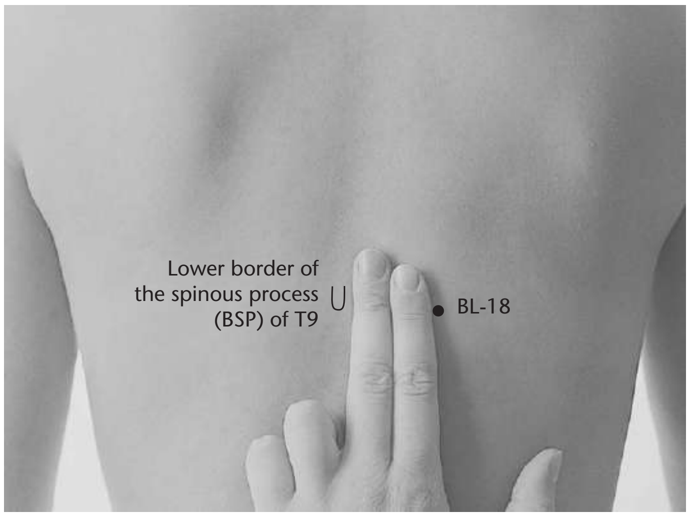
Lower border of the spinous process (BSP) of T9
BL-18

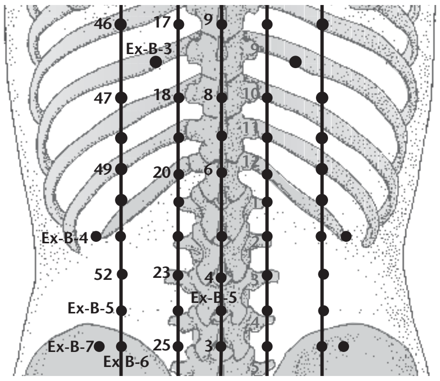
46
17
9
Ex-B-3
9
47
18
8
10
11
49
20
6
12
1
Ex-B-4
2
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex B-6
5

Gall Bladder Shu DANSHU BL-19

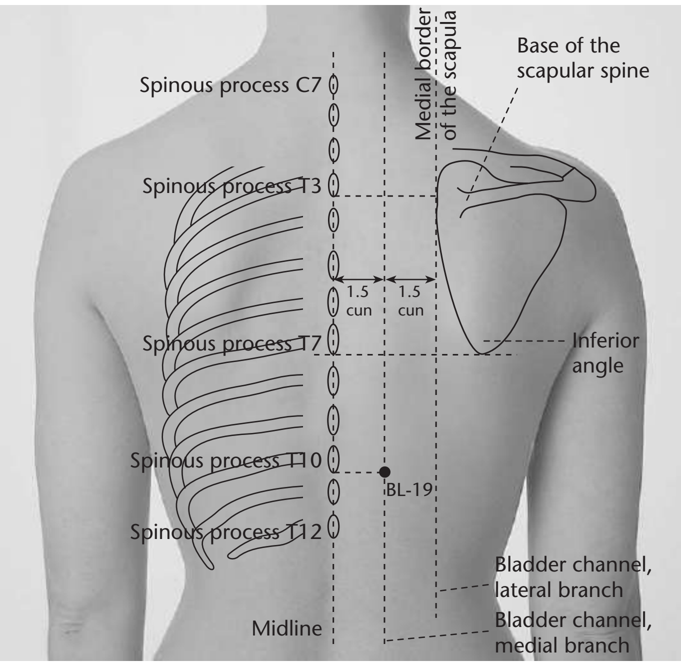
Spinous process C7
Base of the scapular spine
Medial border of the scapula
Spinous process T3
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T10
BL-19
Spinous process T12
Bladder channel, lateral branch
Midline
Bladder channel, medial branch

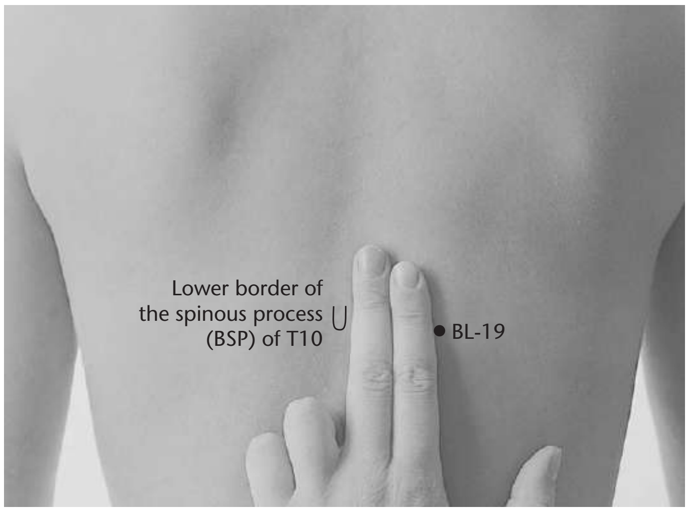
Lower border of the spinous process (BSP) of T10
BL-19

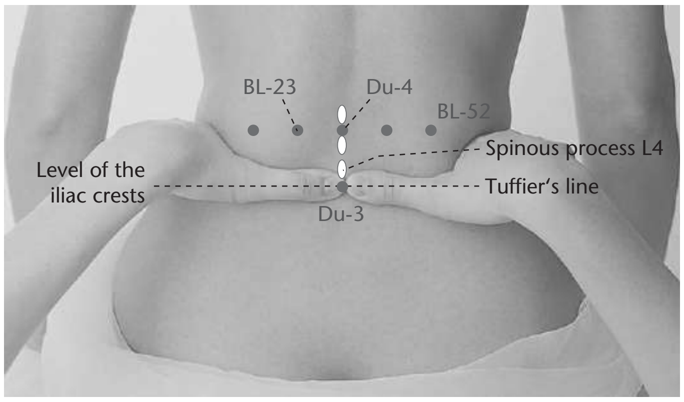
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

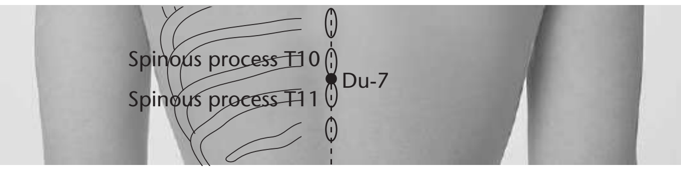
Spinous process T10
Du-7
Spinous process T11

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 10th thoracic vertebra (T10).

How to find

In a standing position with the patient's arms hanging down, the spinous process of T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down three spinous processes to the spinous process of T10. **BL-19** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Use C7 as a reference point (→ 3.4.1). From there, count down 10 spinous processes to T10 and locate **BL-19** 1.5 cun lateral to the lower border of its spinous process. **Or:** Use the lumbar spine as a reference point (→ 3.4.3).

Located on the same level are → **Du-7** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-48** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Clears Damp-Heat from the Liver and Gall Bladder
- Expels pathogenic factors from the *shaoyang*
- Tonifies and regulates the Gall Bladder Qi
- Unbinds the chest and relaxes the pathway of the channel

Special features

Back-*shu* point of the Gall Bladder. Important point for clearing Damp-Heat from the Liver and Gall Bladder.

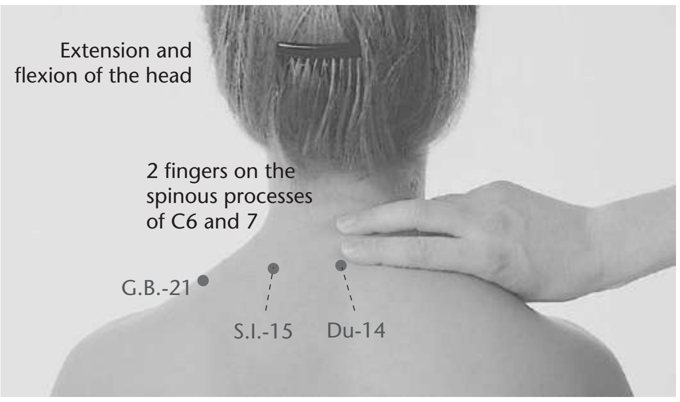
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

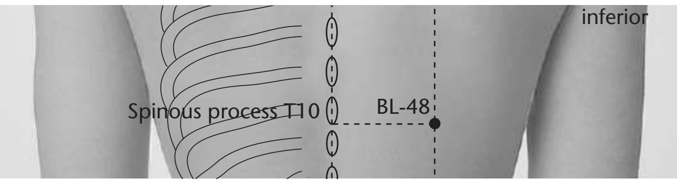
inferior
Spinous process T10
BL-48

BL-20 *Spleen* Shu PISHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 11th thoracic vertebra (T11).

How to find

In a seated position with the arms hanging down, the spinous process of T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down four spinous processes to the spinous process of T11. **BL-20** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Locate the articulation of the most inferior rib (T12) and palpate in a superior direction to T11. **Or:** Use the lumbar spine for reference (→ 3.4.3).

Located on the same level are → **Du-6** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and **BL-49** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Tonifies the Spleen Qi and Yang, regulates Qi in the Middle, raises Qi
- Transforms Dampness
- Holds and nourishes Blood

Special features

Back-*shu* point of the spleen. Major point for tonifying the Middle, especially for Spleen deficiency syndromes. Use moxibustion or tonifying needle techniques for nourishing the Blood.

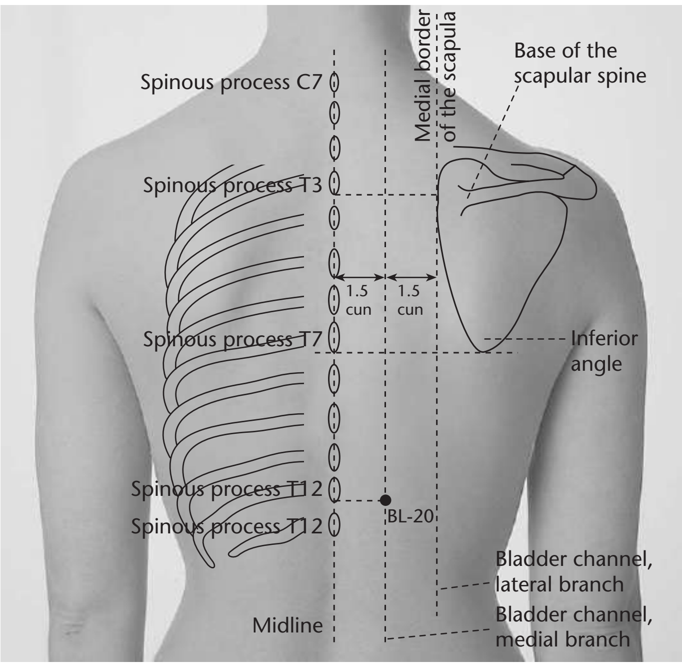
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T12
Spinous process T12
BL-20
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

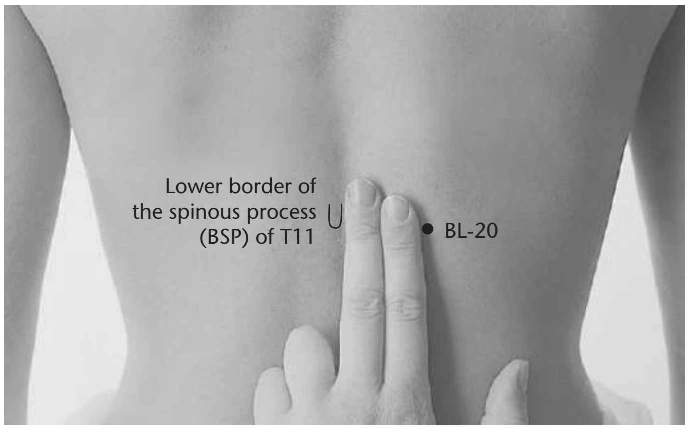
Lower border of the spinous process (BSP) of T11
BL-20

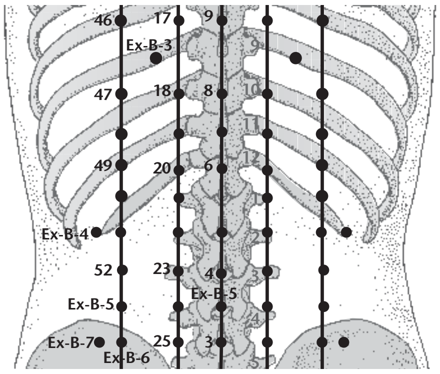
46
17
9
Ex-B-3
9
47
18
8
10
11
49
20
6
12
1
Ex-B-4
2
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex-B-6
5

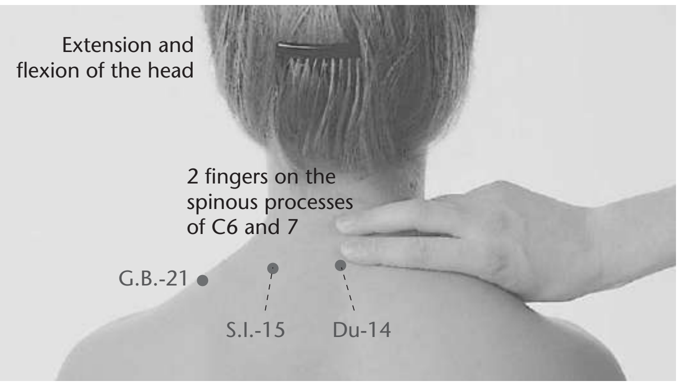
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

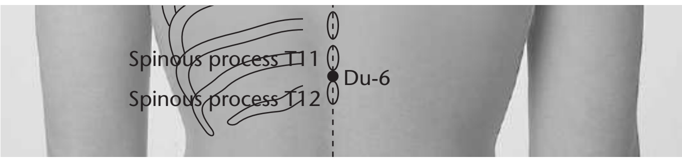
Spinous process T11
Du-6
Spinous process T12

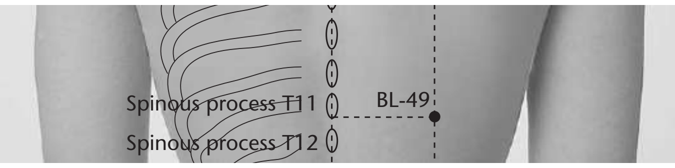
Spinous process T11
BL-49
Spinous process T12

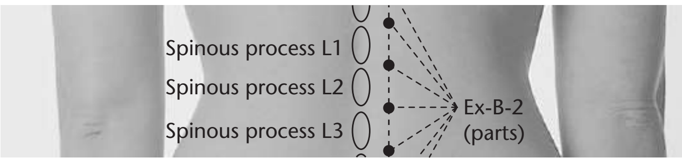
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Stomach Shu WEISHU BL-21

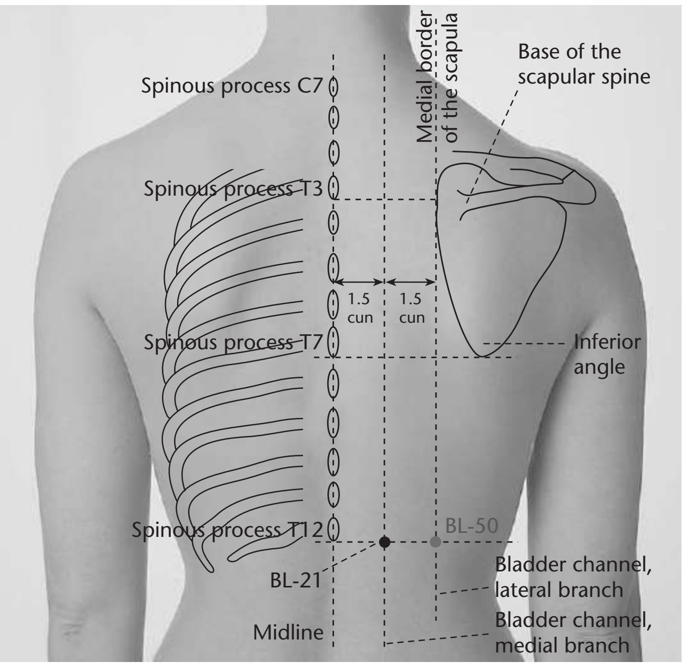
Medial border of the scapula
Base of the scapular spine
Spinous process C7
Spinous process T3
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T12
BL-50
BL-21
Bladder channel, lateral branch
Midline
Bladder channel, medial branch

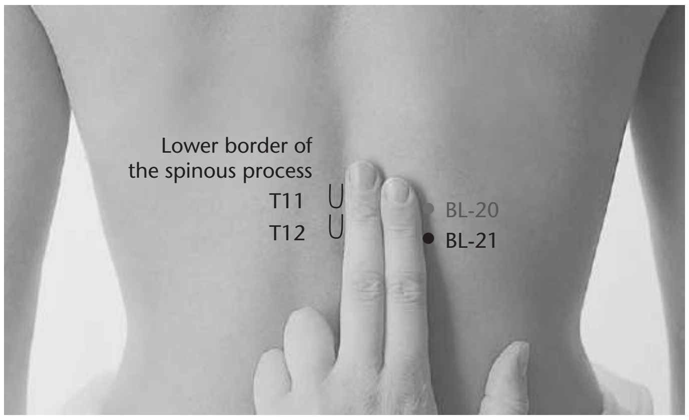
Lower border of the spinous process
T11
T12
BL-20
BL-21

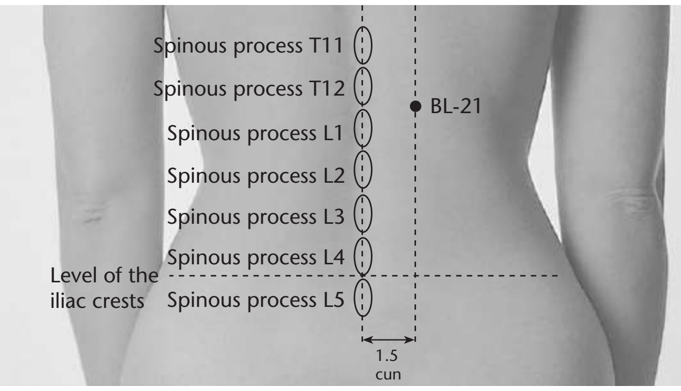
Spinous process T11
Spinous process T12
BL-21
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
Level of the iliac crests
Spinous process L5
1.5 cun

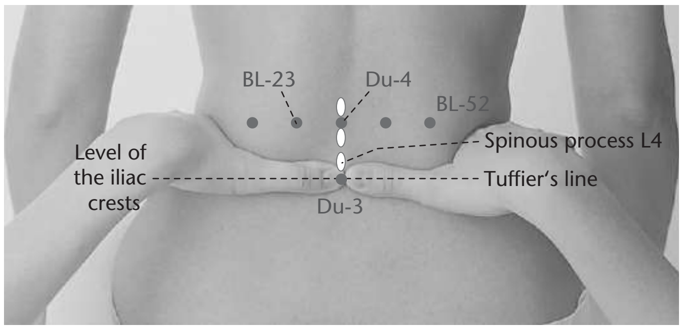
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 12th thoracic vertebra (T12).

How to find

In a seated position with the arms hanging down, the spinous process of T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down five spinous processes to the spinous process of T12. **BL-21** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Palpate for the articulation of the most inferior rib with T12 and locate **BL-21** 1.5 cun lateral to the lower border of its spinous process. **Or:** Use the lumbar spine for reference (→ 3.4.3). Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-50** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Regulates the Stomach, descends counterflow Qi, harmonises the Middle Burner, eliminates Dampness and food stagnation
- Opens the channel

Special features

Back-*shu* point of the Stomach, main point for regulating any dysfunction of the Stomach.

BL-22 San Jiao Shu SANJIAOSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 1st lumbar vertebra (L1).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line (a line connecting the highest points of the iliac crests), which in most cases intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.3). From there, count upward to the lower border of the spinous process of L1 and, on this level, locate **BL-22** 1.5 cun in a lateral direction, on the highest point of the paraspinal musculature. Located on the same level are → **Du-5** (on the midline), a point of → **Ex-B-2/BL-51/Ex-B-4** (0.5 cun/3 cun/3.5cun lateral of the midline).

Needling

Obliquely 0.5–1.5 cun. Caution: Kidneys.

Actions/Indications

- Regulates the Triple Burner
- Opens the water passages and promotes urination
- Has an effect on the *shaoyang*

Special features

Back-*shu* point of the Triple Burner. Especially regulates the relationship between the Middle and Lower Burner.

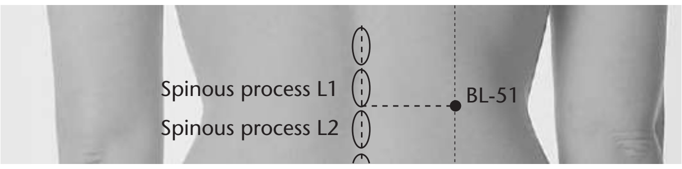
Spinous process L1
Spinous process L2
BL-51

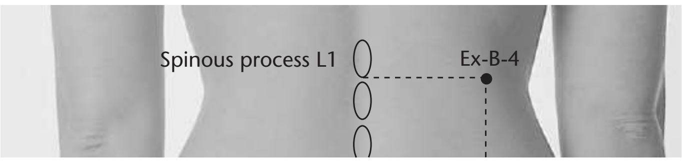
Spinous process L1
Ex-B-4

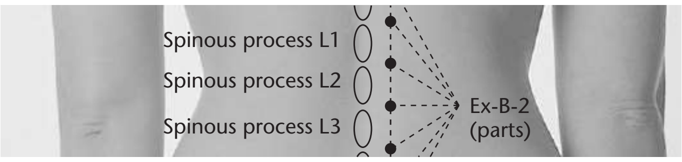
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

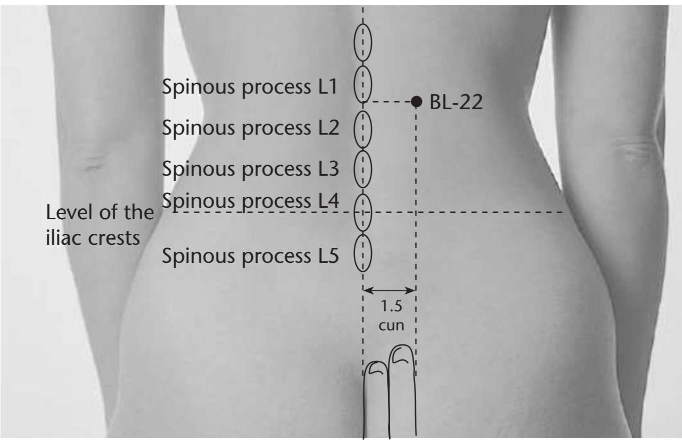
Spinous process L1
BL-22
Spinous process L2
Spinous process L3
Spinous process L4
Level of the
iliac crests
Spinous process L5
1.5
cun

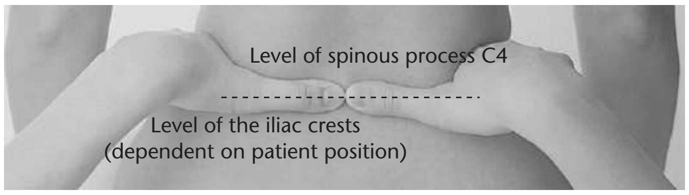
Level of spinous process C4
Level of the iliac crests
(dependent on patient position)

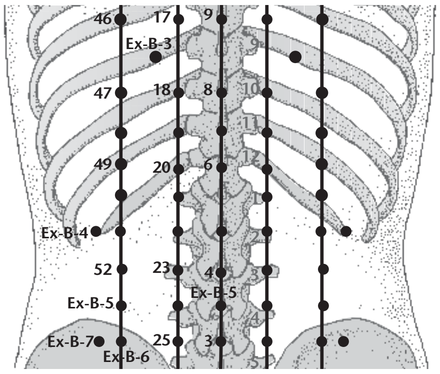
46
17
9
Ex-B-3
47
18
8
10
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6

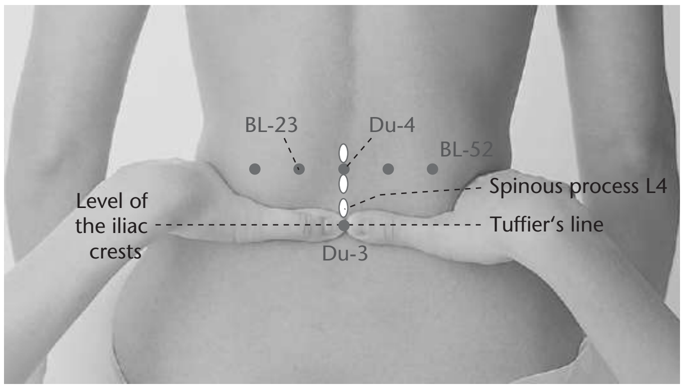
BL-23
Du-4
BL-52
Spinous process L4
Level of
the iliac
crests
Tuffier's line
Du-3

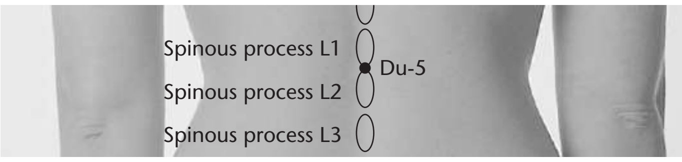
Spinous process L1
Du-5
Spinous process L2
Spinous process L3

Kidney Shu SHENSHU BL-23

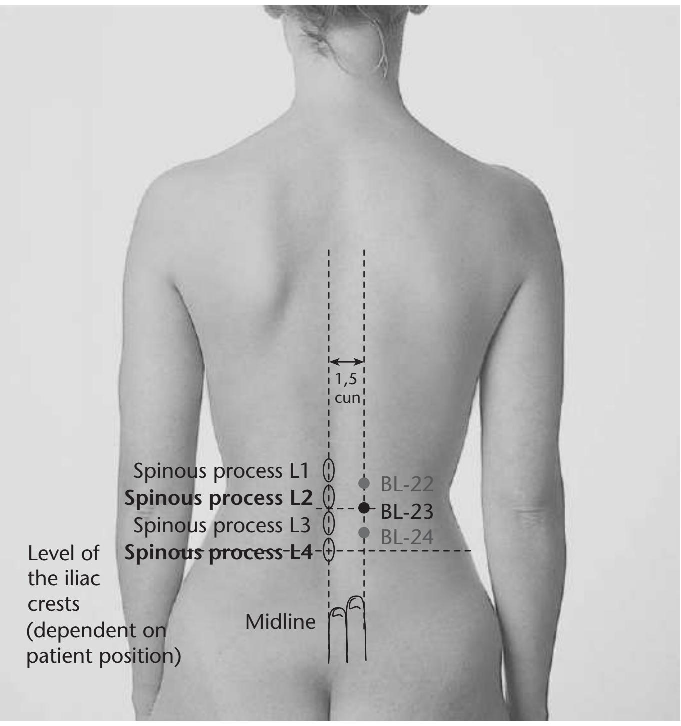
1,5 cun
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
BL-22
BL-23
BL-24
Level of the iliac crests (dependent on patient position)
Midline

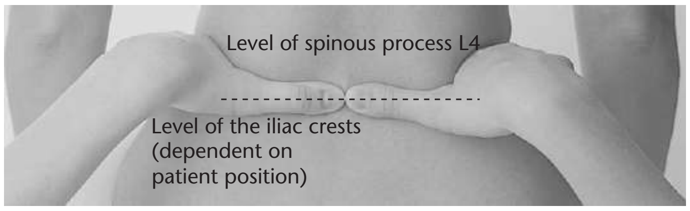
Level of spinous process L4
Level of the iliac crests (dependent on patient position)

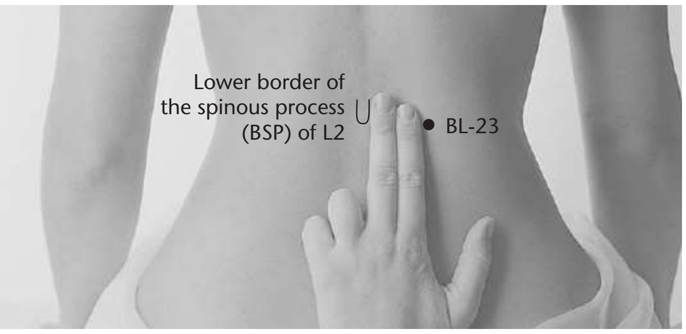
Lower border of the spinous process (BSP) of L2
BL-23

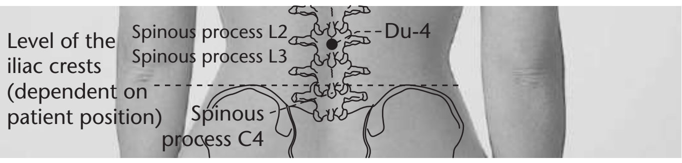
Level of the iliac crests (dependent on patient position)
Spinous process L2
Spinous process L3
Du-4
Spinous process C4

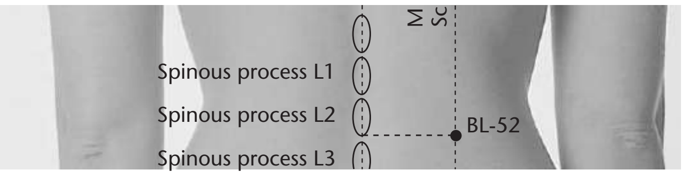
Spinous process L1
Spinous process L2
Spinous process L3
BL-52

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd lumbar vertebra (L2).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line (a line connecting the highest points of the iliac crests), which in most cases intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.3). From there, count upward to the lower border of the spinous process of L2 and, on this level, locate **BL-23** 1.5 cun in a lateral direction, on the highest point of the paraspinal musculature. Located on the same level are → **Du-4** (on the midline), a point of → **Ex-B-2/BL-52** (0.5 cun/3 cun lateral of the midline).

Needling

Vertically or obliquely 0.5–1.5 cun. Caution: Kidneys.

Actions/Indications

- Strengthens the Kidneys, tonifies the Kidney Qi and Yang, benefits the Essence (*jing*), nourishes the Kidney Yin
- Regulates the Lower Burner, benefits the uterus
- Benefits the bones and the marrow
- Benefits the eyes and ears
- Strengthens the lower back

Special features

Back-*shu* point of the Kidneys, according to some authors, meeting point with the *du mai*. Major point for tonifying the Kidneys. Use tonifying needle techniques for all Kidney deficiency syndromes. Moxibustion tonifies especially the Kidney Yang. **Caution**: Moxibustion is contraindicated for Kidney Yin deficiency with empty Heat (Heat in the Five Hearts (Heat in the thorax, the palms and the soles of the feet), night sweats, restlessness, often a red tongue).

BL-24 *Sea of Qi* QIHAISHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 3rd lumbar vertebra (L3).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line. Place both hands directly superior or lateral to the highest points of the iliac crests, joining the thumbs on the midline at the lumbar spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.3). From there, count upward to the lower border of the spinous process of L3 and, on this level, locate **BL-24** 1.5 cun in a lateral direction.

Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **Ex-B-5** (*xiazhishi*; 3 cun lateral to the midline according to the WHO location notes).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Benefits the back and strengthens the knees
- Regulates and tonifies Qi and Blood of the Lower Burner

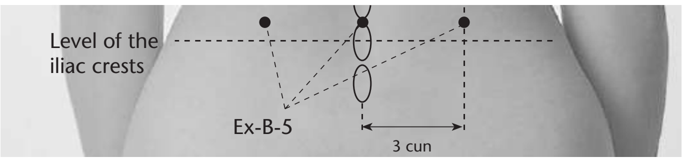
Level of the iliac crests
Ex-B-5
3 cun

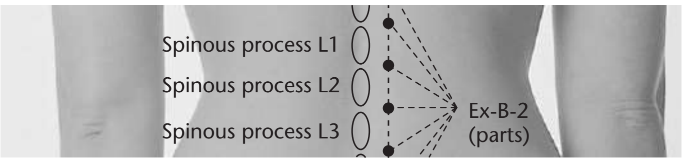
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

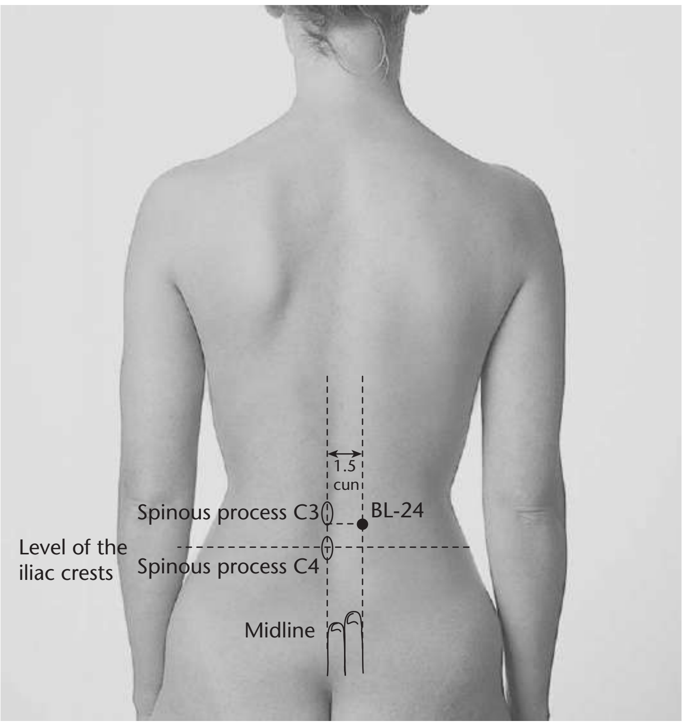
1.5 cun
Spinous process C3
BL-24
Level of the iliac crests
Spinous process C4
Midline

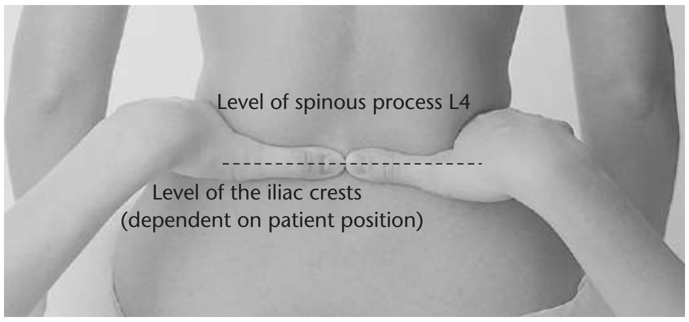
Level of spinous process L4
Level of the iliac crests (dependent on patient position)

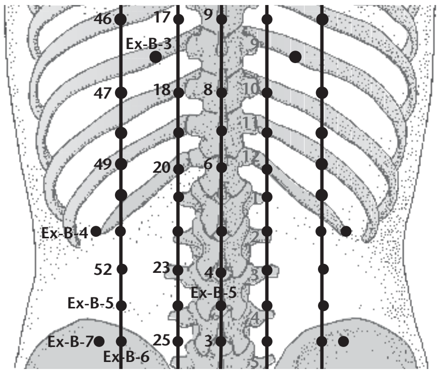
46
17
9
Ex-B-3
9
47
18
8
10
11
49
20
6
12
1
Ex-B-4
2
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex-B-6
5

Large Intestine Shu DACHANGSHU BL-25

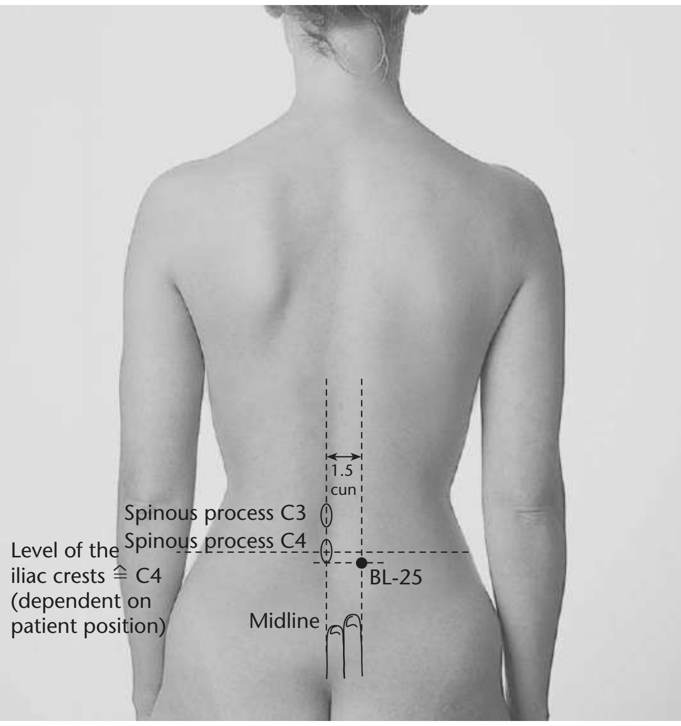
1.5 cun
Spinous process C3
Spinous process C4
Level of the iliac crests ≙ C4 (dependent on patient position)
BL-25
Midline

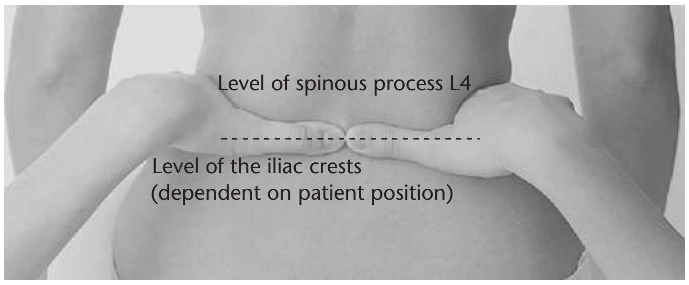
Level of spinous process L4
Level of the iliac crests (dependent on patient position)

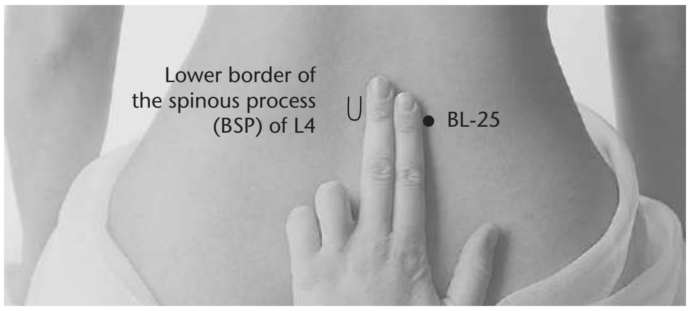
Lower border of the spinous process (BSP) of L4
BL-25

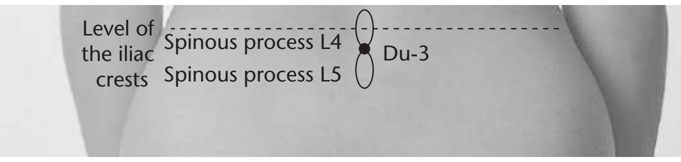
Level of the iliac crests
Spinous process L4
Spinous process L5
Du-3

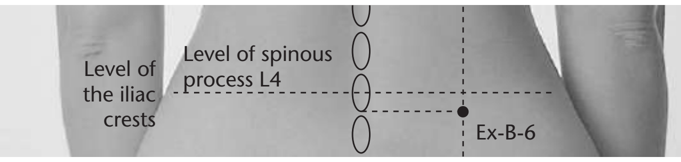
Level of the iliac crests
Level of spinous process L4
Ex-B-6

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 4th lumbar vertebra (L4).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line. Place both hands directly superior or lateral to the highest points of the iliac crests, joining the thumbs on the midline at the lumbar spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.3). On this level, locate **BL-25** 1.5 cun in a lateral direction.

Located on the same level are → **Du-3** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline), → **Ex-B-6** (3 cun lateral to the midline) and → **Ex-B-7** (3.5 cun lateral to the midline).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Regulates the intestines and promotes the Qi flow
- Strengthens the lower back

Special features

Back-*shu* point of the Large Intestine. Important point for regulating the Large Intestine Qi. Important local point for acute and chronic lumbar disorders.

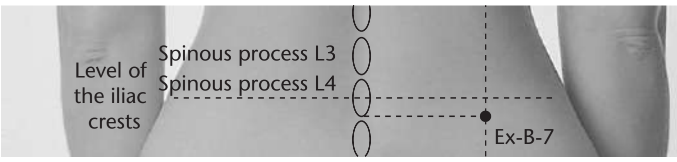
Level of the iliac crests
Spinous process L3
Spinous process L4
Ex-B-7

BL-26 *Gate of Origin* Shu GUANYUANSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 5th lumbar vertebra (L5).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line. Place both hands directly superior or lateral to the highest points of the iliac crests, joining the thumbs on the midline at the lumbar spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position; for more detail on orientation in the lumbar/sacral region → 3.4.3, 3.4.4). Next, locate the lower border of the spinous process of L5 and, on this level, locate **BL-26** 1.5 cun in a lateral direction.

Located on the same level are → **Ex-B-8** (on the midline) and a point of → **Ex-B-2** (0.5 cun lateral to the midline).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Strengthens the back, especially for Kidney deficiency syndromes
- Regulates the Lower Burner

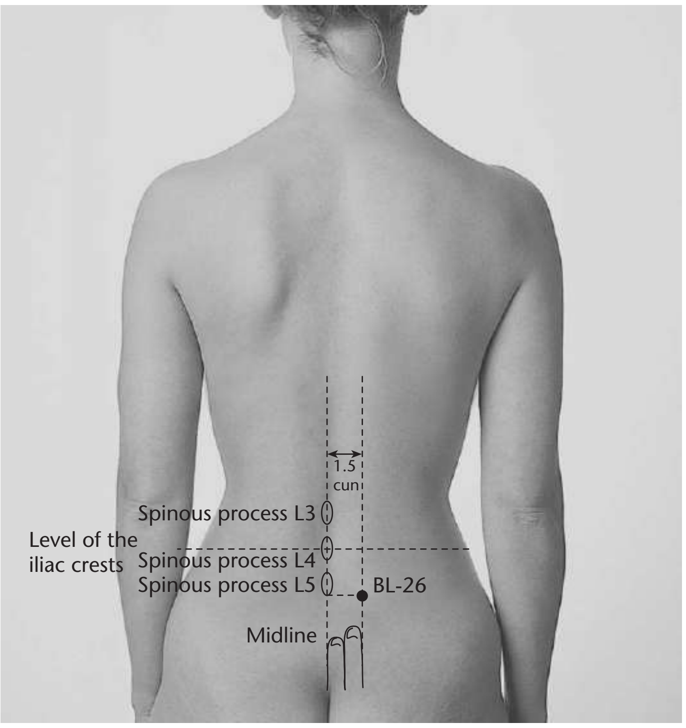
1.5 cun
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
BL-26
Midline

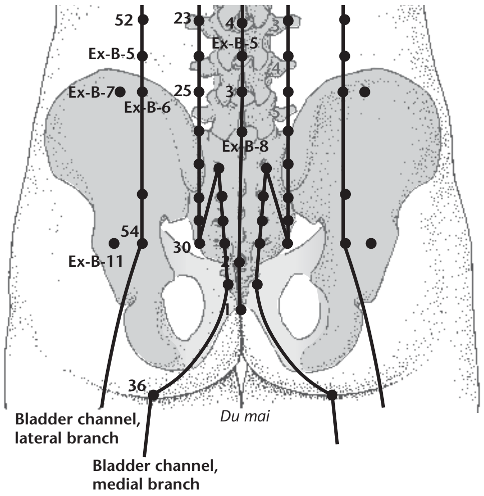
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
36
Bladder channel, lateral branch
Du mai
Bladder channel, medial branch

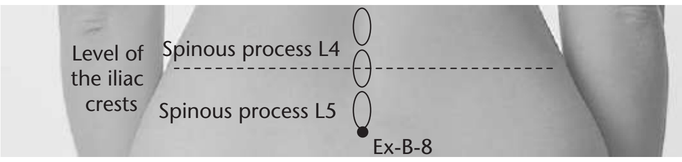
Level of the iliac crests
Spinous process L4
Spinous process L5
Ex-B-8

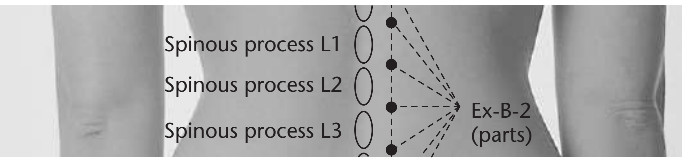
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

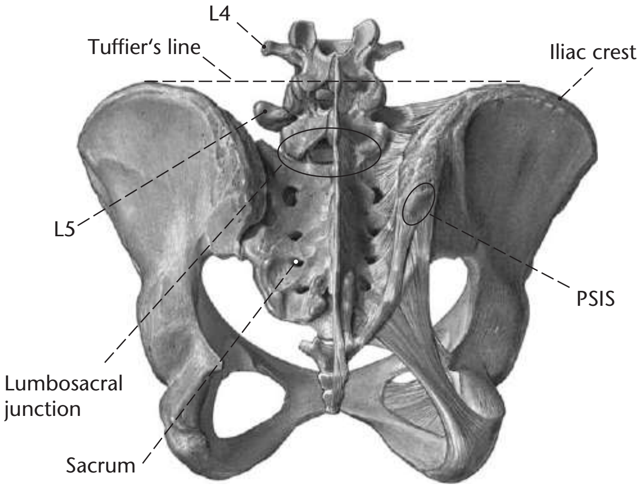
L4
Tuffier's line
Iliac crest
L5
PSIS
Lumbosacral junction
Sacrum

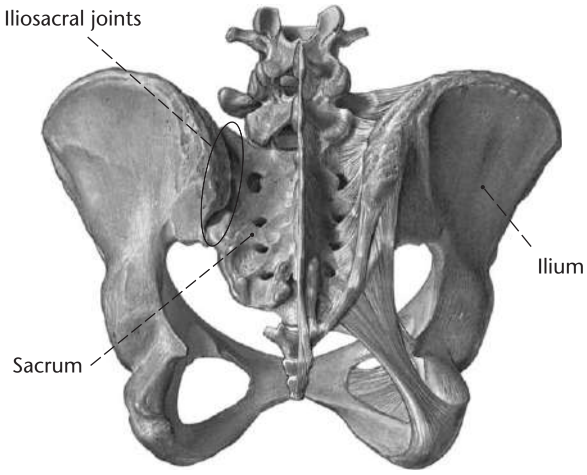
Iliosacral joints
Ilium
Sacrum

Small Intestine Shu XIAOCHANGSHU BL-27

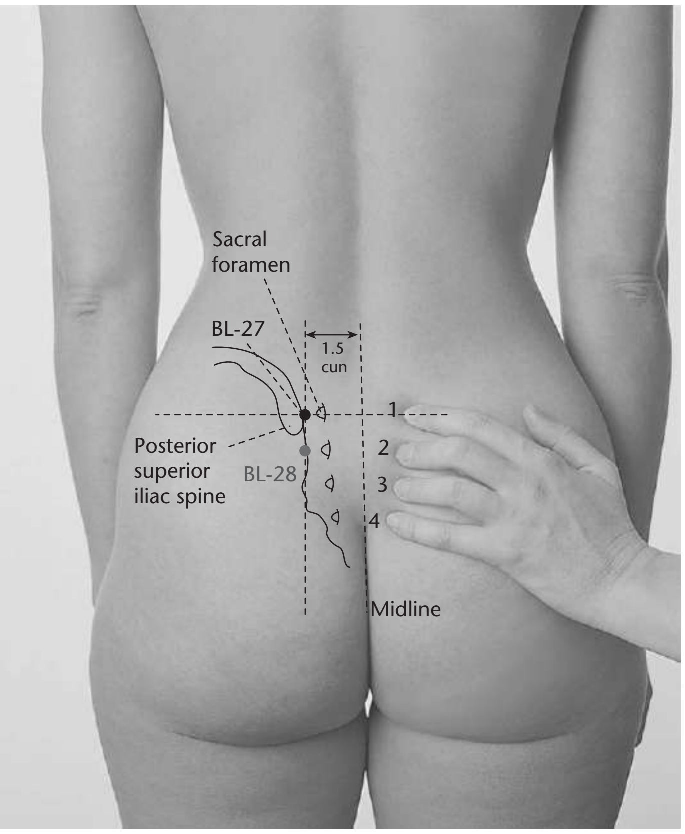
Sacral
foramen
BL-27
1.5
cun
1
2
3
4
Posterior
superior
iliac spine
BL-28
Midline

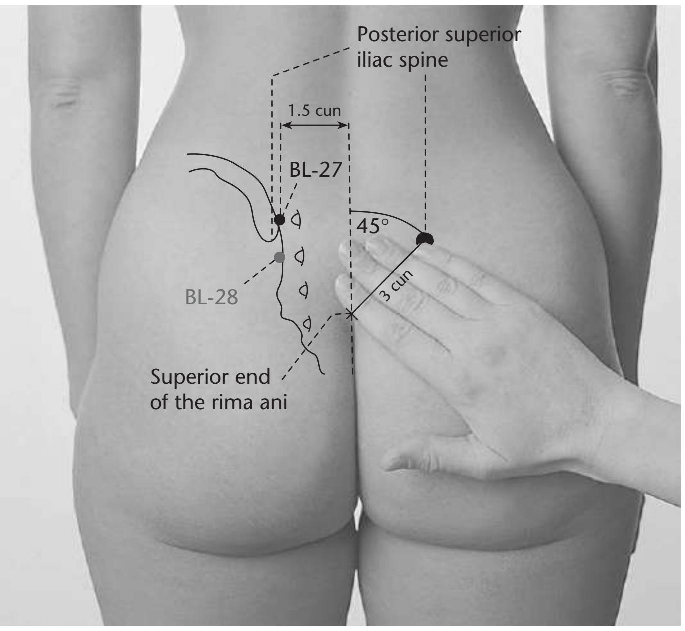
Posterior superior
iliac spine
1.5 cun
BL-27
45°
3 cun
BL-28
Superior end
of the rima ani

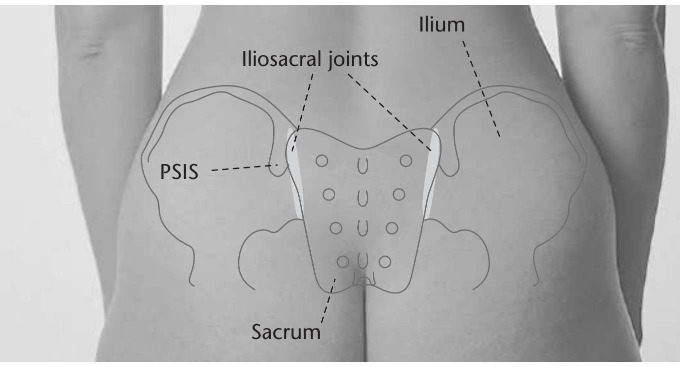
Ilium
Iliosacral joints
PSIS
Sacrum

Location

1.5 cun lateral to the posterior midline, on the level of the first sacral foramen.

How to find

Quick method: Locate the posterior superior iliac spine (PSIS → 3.4.3): from the rima ani, palpate 3 cun (1 handbreadth) in a superior and lateral direction at a 45° angle, until you can feel a distinct bony ridge (often superficially visible by a dimple). **BL-27** is located slightly superiorly and medially to the PSIS, 1.5 cun lateral to the midline, on the level of the first sacral foramen. For more detail on orientation in the lumbar and sacral region → 3.4.3, 3.4.4.

Located on the same level are → **BL-31** (over the 1st sacral foramen) and → **BL-28** (slightly below and medial to the PSIS at the level of the 2nd sacral foramen).

Needling

Vertically 0.5–1 cun

Actions/Indications

- Tonifies the Small Intestine Qi
- Drains Dampness and Damp-Heat
- Regulates the water passages

Special features

Back-*shu* point of the Small Intestine

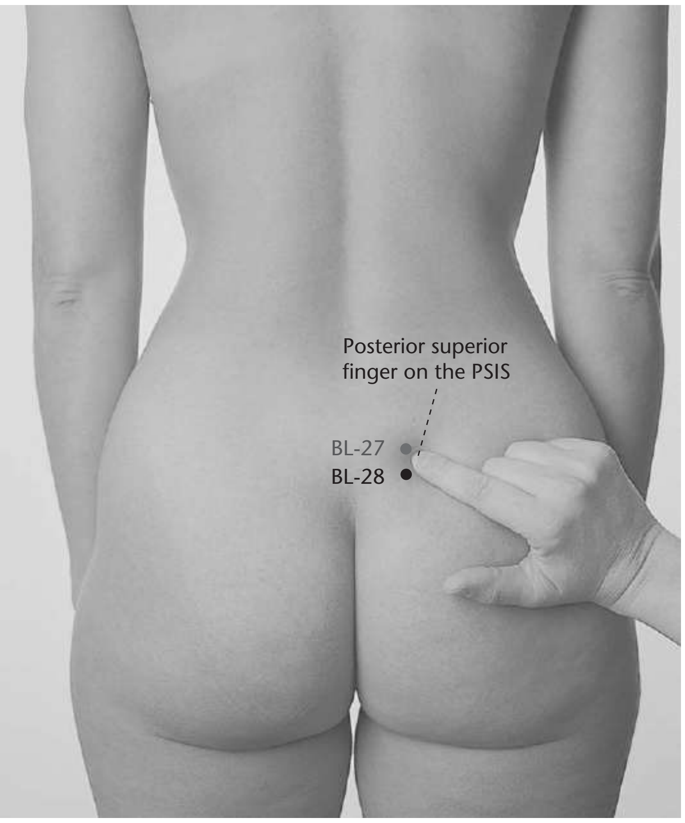
Posterior superior
finger on the PSIS
BL-27
BL-28

Ch04.7a-F10028.qxd 2/23/08 3:54 PM Page 278

BL-28 *Bladder* Shu PANGGUANGSHU

Location

1.5 cun lateral to the posterior midline, on the level of the second sacral foramen.

How to find

Quick method: Locate the posterior superior iliac spine (PSIS → 3.4.3): from the rima ani, palpate 3 cun (1 handbreadth) in a superior and lateral direction at a 45° angle, until you can feel a distinct bony ridge (often superficially visible by a dimple). **BL-28** is located slightly inferiorly and medially to the PSIS, 1.5 cun lateral to the midline on the level of the 2nd sacral foramen. For more detail on orientation in the lumbar and sacral region → 3.4.3, 3.4.4.

Located on the same level are → **BL-32** (over the 2nd sacral foramen) and → **BL-53** (3 cun lateral to the midline). → **BL-27** is located slightly superiorly and medially to the PSIS on the level of the 1st sacral foramen.

Needling

Vertically 0.5–1.5 cun. For local disorders, slightly oblique insertion towards the iliosacral joint.

Actions/Indications

- Regulates the Bladder and the Lower Burner, clears Damp-Heat from the Lower Burner, removes stagnation and transforms masses
- Benefits the lumbar region and the legs

Special features

Back-*shu* point of the Bladder

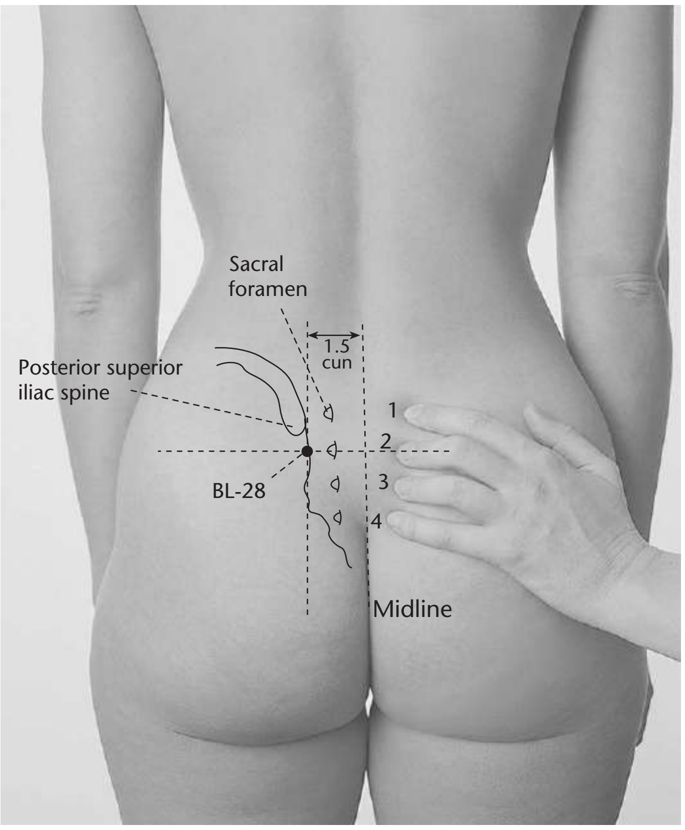
Sacral
foramen
1.5
cun
Posterior superior
iliac spine
1
2
3
4
BL-28
Midline

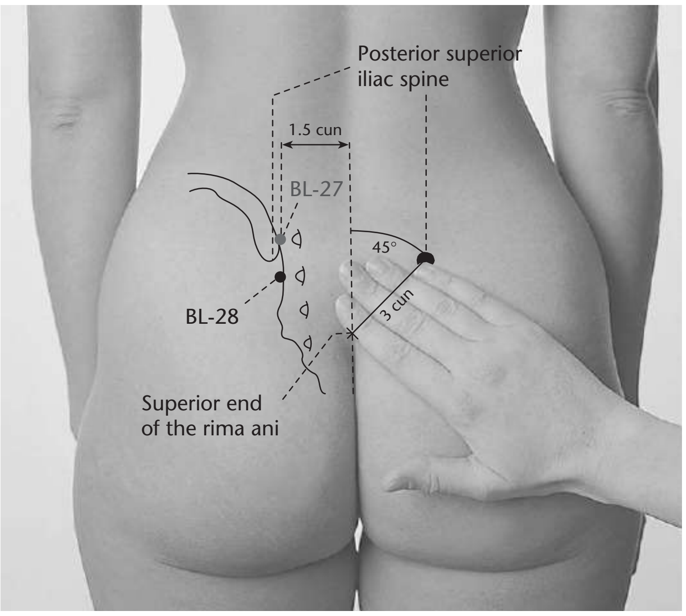
Posterior superior
iliac spine
1.5 cun
BL-27
45°
3 cun
BL-28
Superior end
of the rima ani

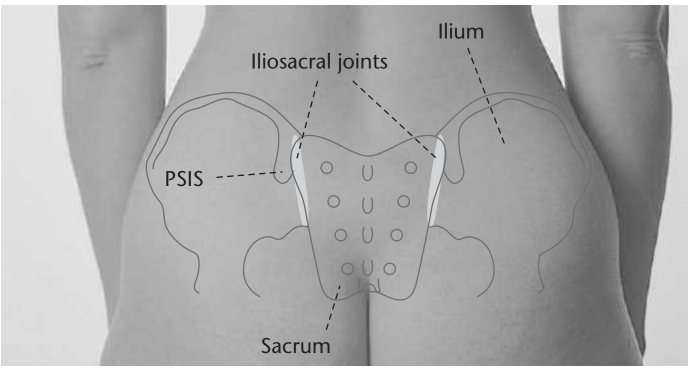
Ilium
Iliosacral joints
PSIS
Sacrum

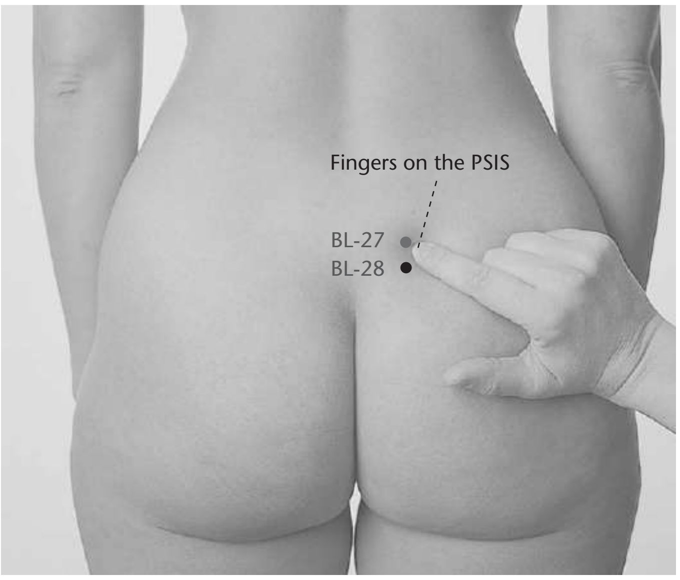
Fingers on the PSIS
BL-27
BL-28

Mid Spine Shu ZHONGLÜSHU BL-29

Location

1.5 cun lateral to the posterior midline, on the level of the 3rd sacral foramen.

How to find

For more detail on orientation in the lumbar and sacral region, see → 3.4.3, 3.4.4. First of all find L5. The next palpable structure in an inferior direction is the most superior sacral spinous process. Palpate in an inferior direction for two further sacral processes and locate the 3rd sacral foramen in a depression less than 1 cun lateral to the 3rd process. **BL-29** is located on the same level (3rd sacral foramen) and 1.5 cun lateral to the midline. Also located on this level is → **BL-33** (over the 3rd sacral foramen).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Strengthens the lower back and the Kidneys
- Expels Cold
- Regulates the Lower Burner and stops diarrhoea

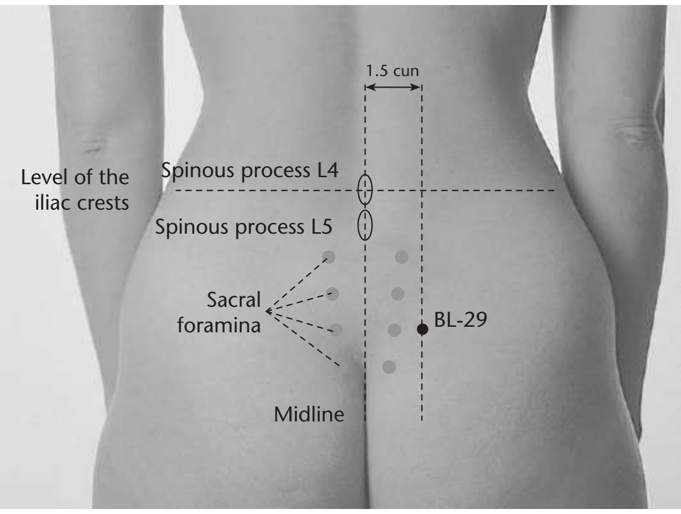
1.5 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
BL-29
Midline

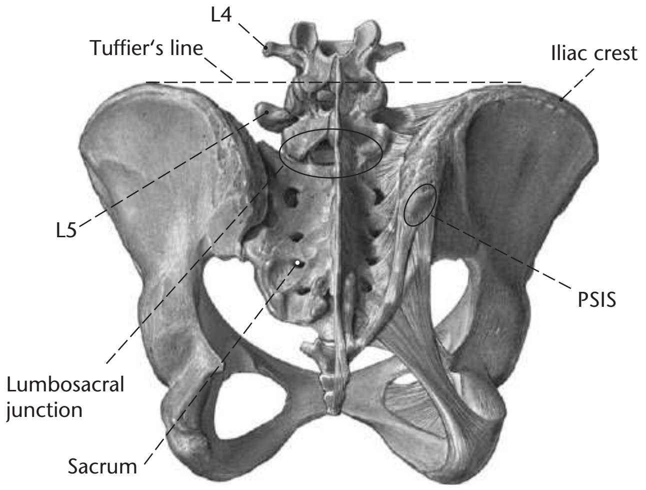
L4
Tuffier's line
Iliac crest
L5
PSIS
Lumbosacral junction
Sacrum

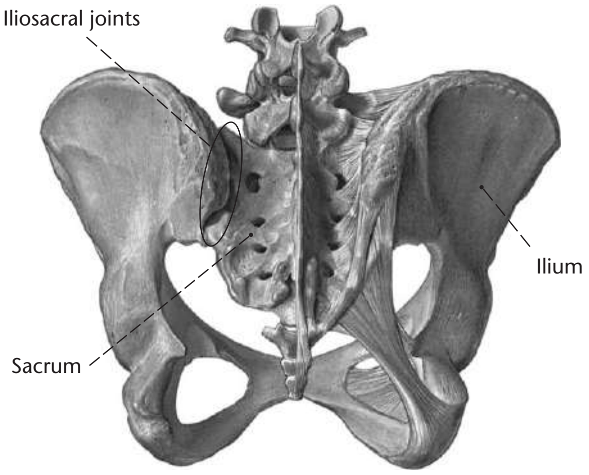
Iliosacral joints
Ilium
Sacrum

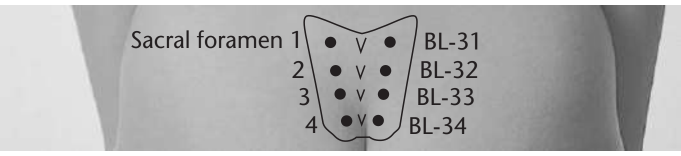
Sacral foramen 1
2
3
4
BL-31
BL-32
BL-33
BL-34

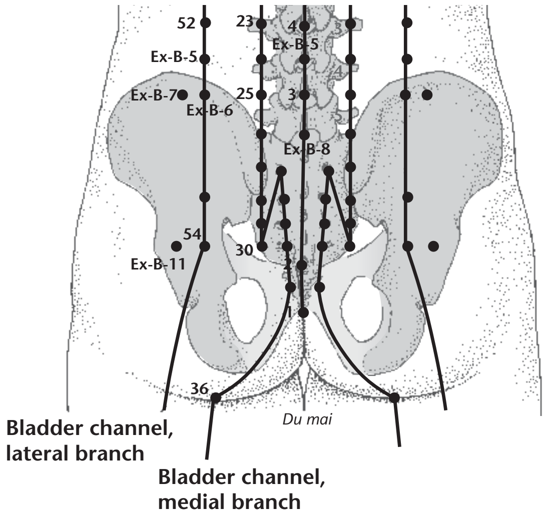
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

BL-30 *White Ring* Shu BAI HUAN SHU

Location

1.5 cun lateral to the posterior midline, on the level of the 4th sacral foramen.

How to find

For more detail on orientation in the lumbar and sacral region, see → 3.4.3, 3.4.4. First, find L5. The next palpable structure in an inferior direction is the most superior sacral spinous process. Palpate in an inferior direction for three further sacral processes and locate the 4th sacral foramen in a depression less than 1 cun lateral to the 4th process. **BL-30** is located on the level of the 4th foramen, 1.5 cun lateral to the midline.

Located on the same level are → **BL-34** (over the 4th sacral foramen) and → **BL-54** (3 cun lateral to the midline).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Strengthens the lower back and the knees
- Eliminates Damp-Heat from the Lower Burner
- Regulates menstruation

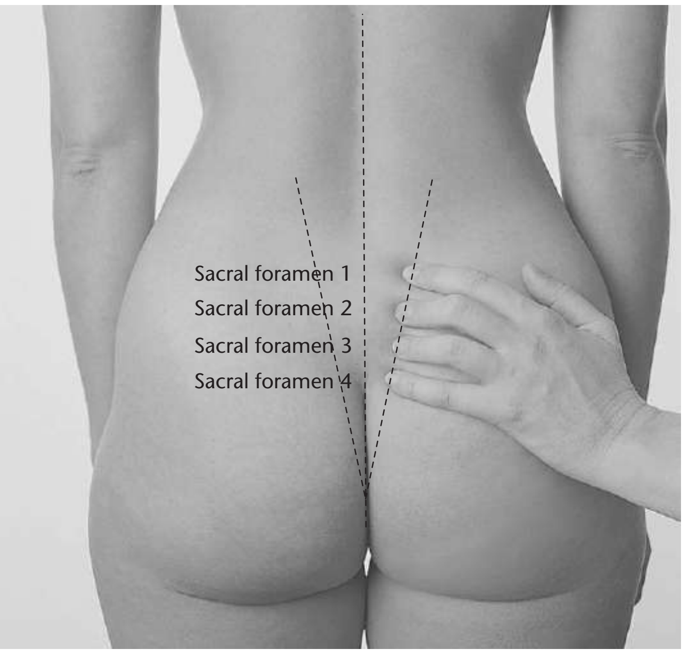
Sacral foramen 1
Sacral foramen 2
Sacral foramen 3
Sacral foramen 4

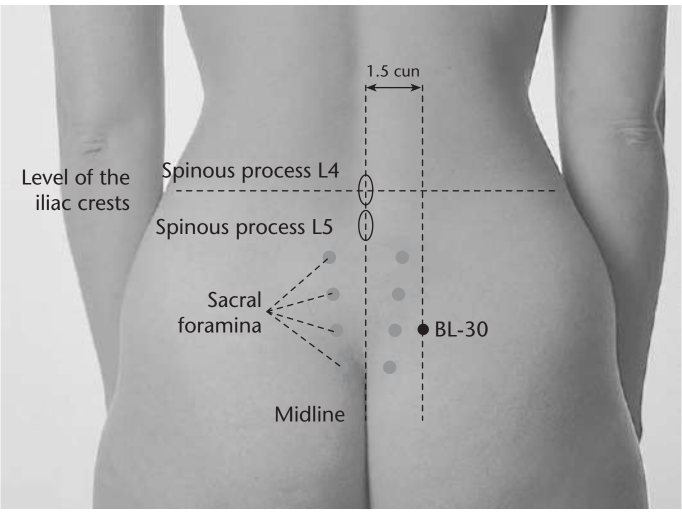
1.5 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
BL-30
Midline

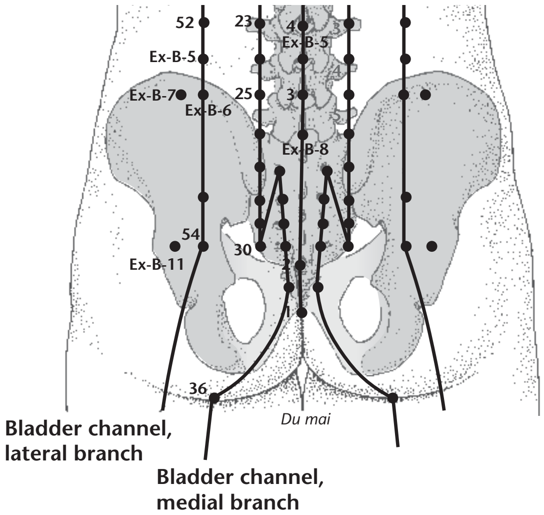
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex-B-6
5
Ex-B-8
54
30
Ex-B-11
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

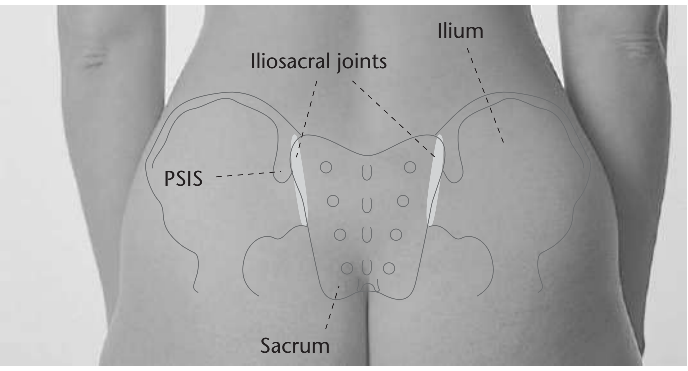
Ilium
Iliosacral joints
PSIS
Sacrum

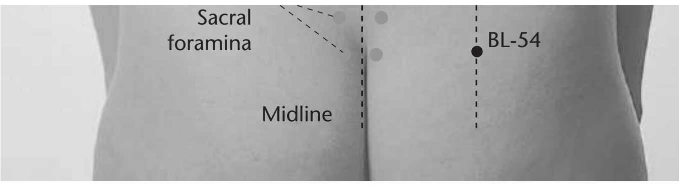
Sacral foramina
BL-54
Midline

Eight Liao BALIAO BL 31 – BL 34

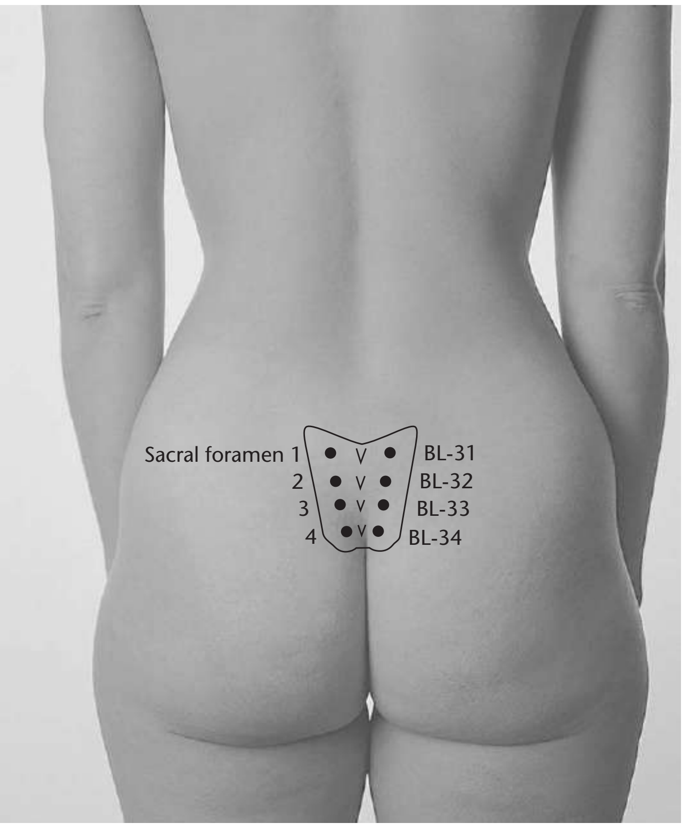

These points are described together because of their similar actions and indications.

BL-31 (*shangliao*) 'Upper Crevice', 1st sacral foramen
BL-32 (*ciliao*) 'Second Crevice', 2nd sacral foramen
BL-33 (*zhongliao*) 'Middle Crevice', 3rd sacral foramen
BL-34 (*xialiao*) 'Lower Crevice', 4th sacral foramen

Location

BL-31: Over the 1st sacral foramen
BL-32: Over the 2nd sacral foramen
BL-33: Over the 3rd sacral foramen
BL-34: Over the 4th sacral foramen

How to find

For more detail on orientation in the lumbar and sacral region, ➞ 3.4.3, 3.4.4. First, locate L5. The next palpable structure in an inferior direction is the most superior sacral spinous process. Palpate in a lateral direction on the level of the sacral processes for the depressions of the sacral foramina. **Or:** Locate the 4th sacral foramen on the level of the superior end of the rima ani (sometimes a bit more superiorly or inferiorly). Place the little finger on the 4th foramen, then the ring, middle and index fingers on a slightly V-shaped line in a superior direction, palpating for the depressions of the sacral foramina. With the fingers in this position, the index finger will be resting on the 1st foramen (**BL-31**), the middle finger on the 2nd foramen (**BL-32**), the ring finger on the 3rd foramen (**BL-33**) and the little finger on the 4th foramen (**BL-34**). Located on the same level as **BL-31–BL-34** are the points ➞ **BL-27–BL-30** on the inner branch of the BL channel (1.5 cun lateral to the midline) as well as ➞ **BL-53** (on the level of **BL-32**) and ➞ **BL-54** (on the level of **BL-34**), both 3 cun lateral to the midline. Also on the level of **BL-34** is ➞ **Ex-B-11** (3.5 cun lateral to the midline).

Needling

Vertically 0.7–1.5 cun. Caution: Pregnancy (especially reducing needling techniques are contraindicated); exception: to promote labour.

Actions/Indications

- Regulates the Lower Burner, promotes urination (all eight points) (**BL-32** and **BL-33** have the strongest action in cases of urinary disorders, while **BL-34** has the strongest action in cases of genital disorders)
- Strengthens the Kidneys and Essence (*jing*)
- Benefits the Intestines (**BL-34** has the strongest action in cases of intestinal disorders)
- Promotes labour (for example, use **BL-32** with electro-acupuncture)
- Benefits the lumbar region (**BL-32** and **BL-33** have the strongest action)

Special features

According to some authors, meeting point with the G.B. channel. Clinically, **BL-32** is the most commonly used point owing to its widest range of indications. Electro-acupuncture at **BL-32** is useful for alleviating pain during labour: after needling this point, bend the handle of the needle in a superior direction at a 90° angle and tape it to the skin. Now attach the electro-stimulation. This method allows the woman to lie down if desired.
Moxibustion of the *baliao* is indicated for difficult urination or defaecation due to Kidney deficiency in elderly people.

BL-35 *Meeting of Yang* HUIYANG

Location

0.5 cun lateral to the posterior midline, on the level of the tip of the coccyx.

How to find

Locate the coccyx superior to the anus. In contrast to the sacrum, the coccyx is movable. **BL-35** is located 0.5 cun lateral to the midline on the level of the tip of the coccyx.

→ **Du-2** is located on the midline, on the level of the sacral hiatus (→ 3.4.4) and superior to **BL-35** and the tip of the coccyx.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Clears Damp-Heat from the Lower Burner
- Treats haemorrhoids and rectal prolapse
- Benefits the coccyx

Special features

Especially in patients with a weakened immune system, this point should be needled only if strictly necessary and after using a disinfecting wipe three times.

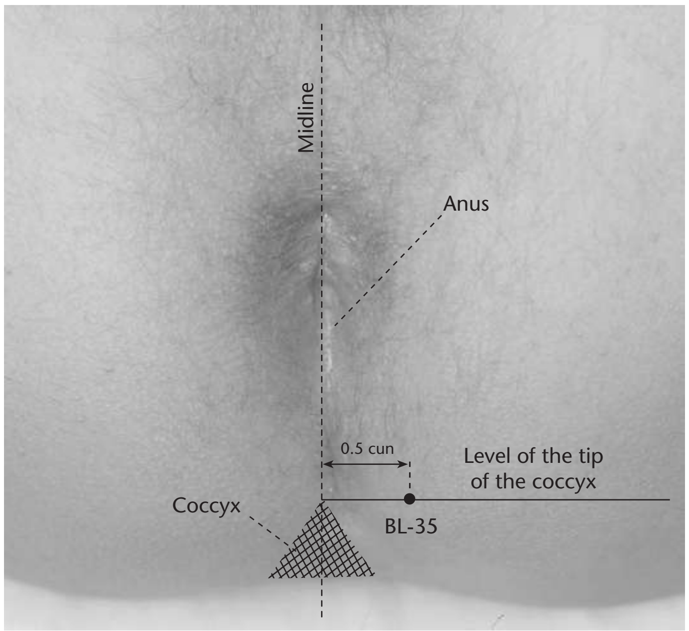
Midline
Anus
0.5 cun
Level of the tip of the coccyx
Coccyx
BL-35

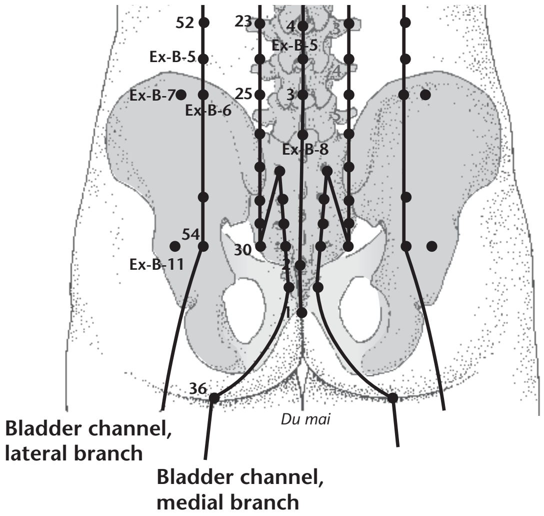
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
Ex-B-6
25
3
5
Ex-B-8
54
Ex-B-11
30
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

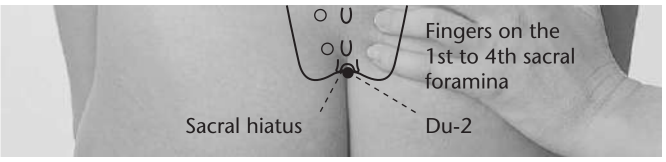
Fingers on the 1st to 4th sacral foramina
Sacral hiatus
Du-2

Hold and Support CHENGFU BL-36

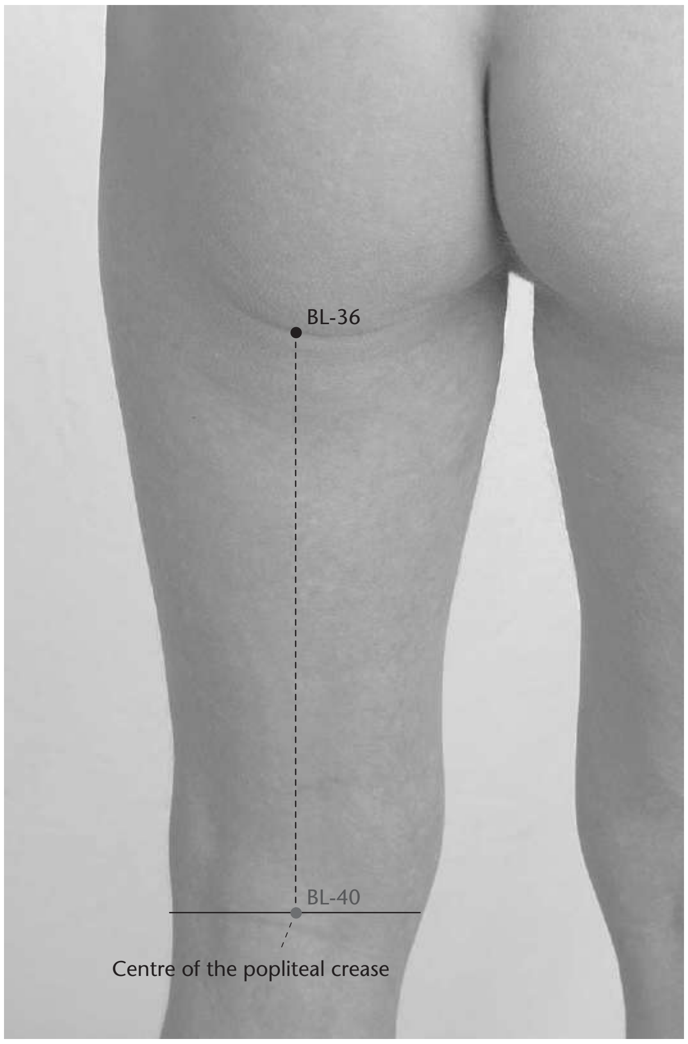
BL-36
BL-40
Centre of the popliteal crease

Location

In the gluteal crease, superior to the midpoint of the popliteal crease (→ **BL-40**).

How to find

Locate the midpoint of the popliteal crease (in obese patients, use bony/muscular structures to find the midpoint). **BL-36** is located directly vertically above the midpoint of the popliteal crease, in the gluteal crease at the transition between the buttocks and the posterior aspect of the thigh (palpate for a depression).

Needling

Vertically 1–2 cun

Actions/Indications

- Opens the channel, alleviates pain
- Relaxes the tendons
- Regulates the Lower Burner and treats haemorrhoids

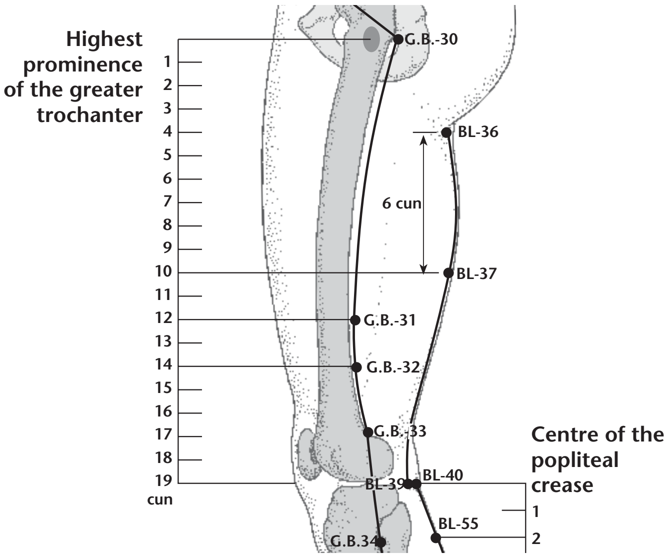
Highest prominence of the greater trochanter
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-40
BL-39
G.B.34
BL-55
Centre of the popliteal crease
1
2

BL-37 *Gate of Abundance* YINMEN

Location

6 cun distal to → **BL-36** (gluteal crease), on a line connecting → **BL-36** and → **BL-40** (in the popliteal crease), in a gap in the musculature.

How to find

First, locate → **BL-36** in the centre of the gluteal crease and, from there, palpate 6 cun along the posterior aspect of the thigh in an inferior direction towards the centre of the popliteal crease (→ **BL-40**). This is the location of **BL-37**, which in slim patients can be located where the long head of the biceps femoris muscle and the semitendinosus muscle converge (usually 8 cun superior to the centre of the popliteal crease. **Or:** Spreading hands technique (→ 2.3.3 or elastic tape): Place the little fingers on the centre of the popliteal crease (→ **BL-40**) and the centre of the gluteal crease (→ **BL-36**). Locate **BL-37** 1–2 cun proximal to the midpoint of the distance, in a depression between the two muscle bellies.

Needling

Vertically 0.5–2 cun

Actions/Indications

- Relaxes the tendons
- Opens the channel and *luo* vessels
- Benefits the lower back

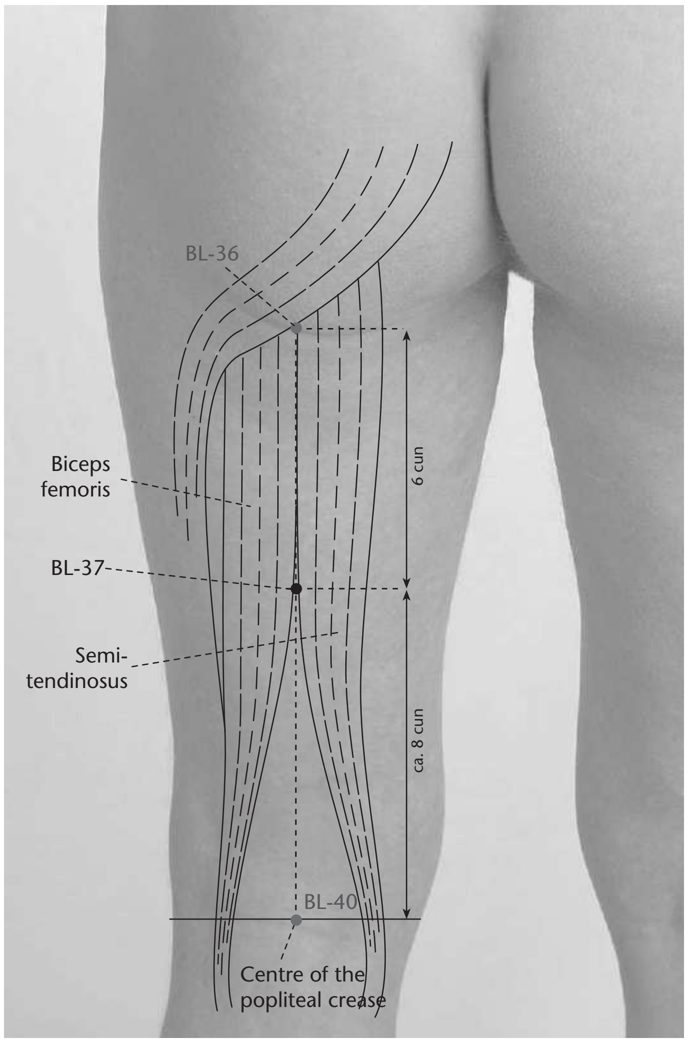
BL-36
Biceps femoris
BL-37
Semi-tendinosus
6 cun
ca. 8 cun
BL-40
Centre of the popliteal crease

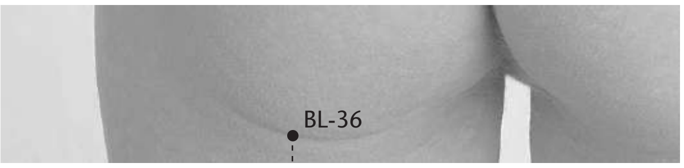

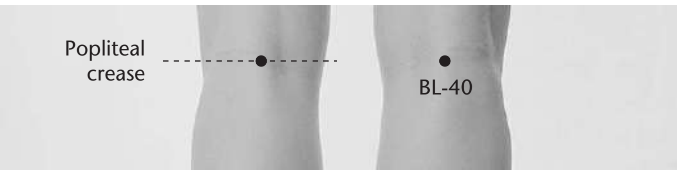
Popliteal crease
BL-40

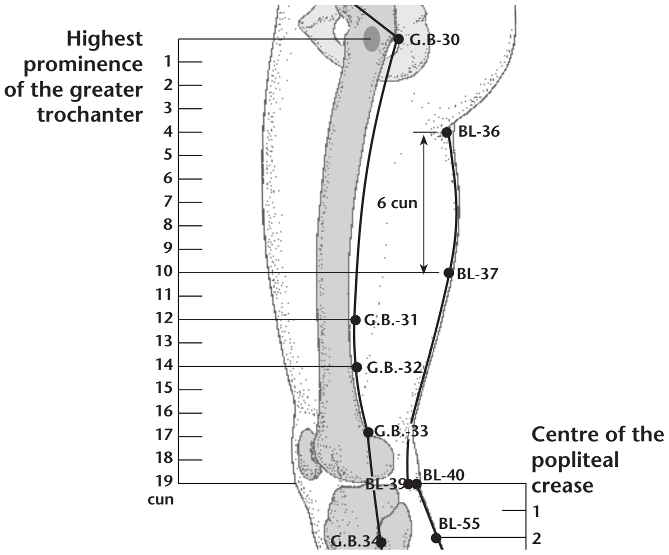
Highest prominence of the greater trochanter
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
G.B-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-39
BL-40
BL-55
G.B.34
Centre of the popliteal crease
1
2

Floating Cleft FUXI BL-38

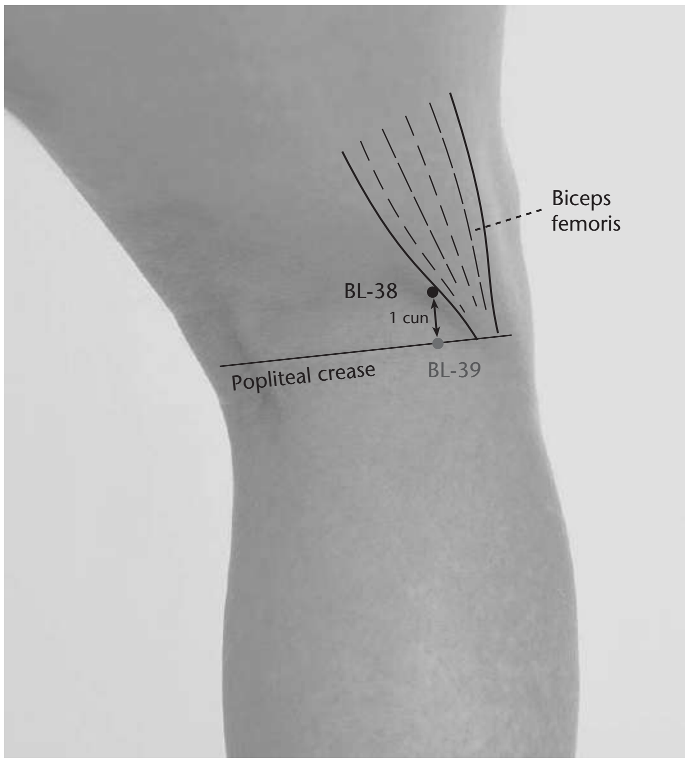

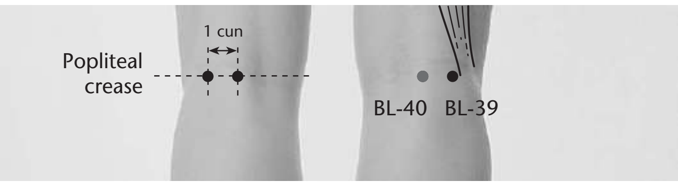

Location

On the laterodorsal aspect of the knee, 1 cun superior and lateral to the centre of the popliteal crease (→ **BL-40**), medial to the biceps femoris muscle, or 1 cun proximal to → **BL-39**.

How to find

To locate the popliteal crease, ask the patient to flex their knee slightly and palpate for the knee joint space. By flexing the knee against resistance, the tendon of the biceps femoris muscle will become more pronounced and visible. Locate **BL-38** 1 cun superior to the knee joint on the medial border of the tendon.
→ **BL-39** is located 1 cun distal to **BL-38** on the level of the popliteal crease.

Needling

Vertically 1–2 cun

Actions/Indications

- Relaxes the muscles and tendons
- Opens the channel and *luo* vessels
- Clears Heat (from the Small Intestine)

BL-39 *Outside of the Crook* WEIYANG

Location

At the lateral end of the popliteal crease, on the medial side of the tendon of the long head of the biceps femoris muscle, 1 cun lateral to → **BL-40** (in the centre of the popliteal crease).

How to find

This point is best located with the patient's knee slightly flexed. Find the midpoint of the popliteal crease (→ **BL-40**) and locate **BL-39** approximately 1 cun lateral in a depression medial to the tendon of the long head of the biceps femoris muscle. Located on the same level is → **BL-40** at the centre of the popliteal crease.

Needling

Vertically 0.5–1.5 cun. Caution: N. fibularis communis.

Actions/Indications

- Harmonises the Triple Burner and regulates the water passages
- Opens the channel and the *luo* vessels, alleviates pain

Special features

Lower *he*-sea point of the Triple Burner

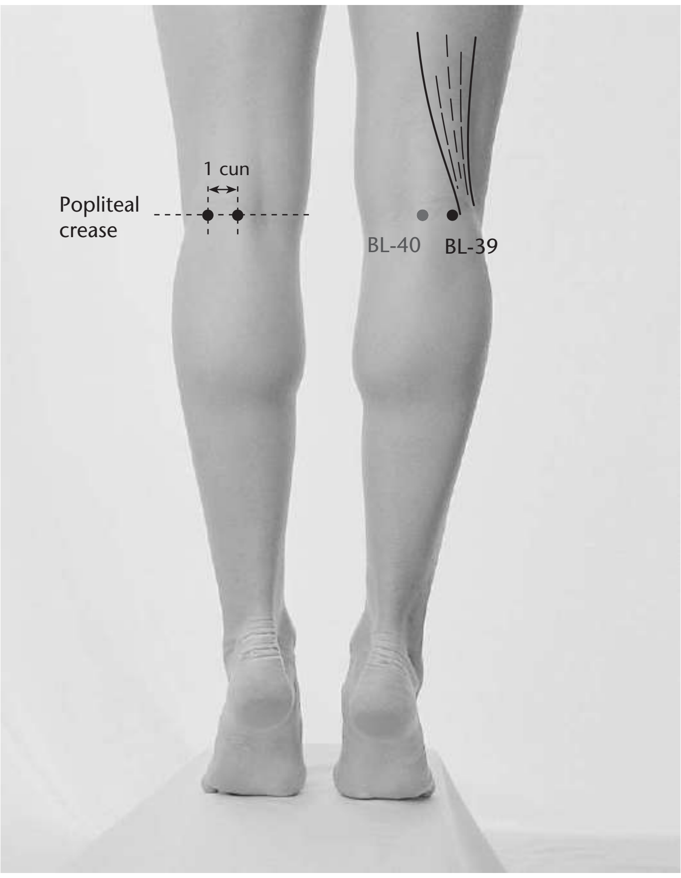
1 cun
Popliteal crease
BL-40
BL-39

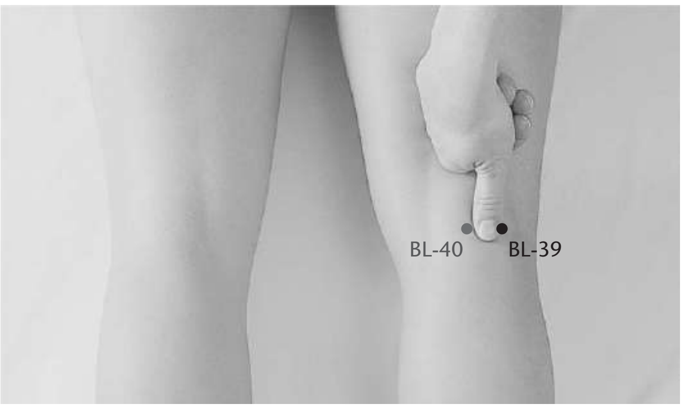

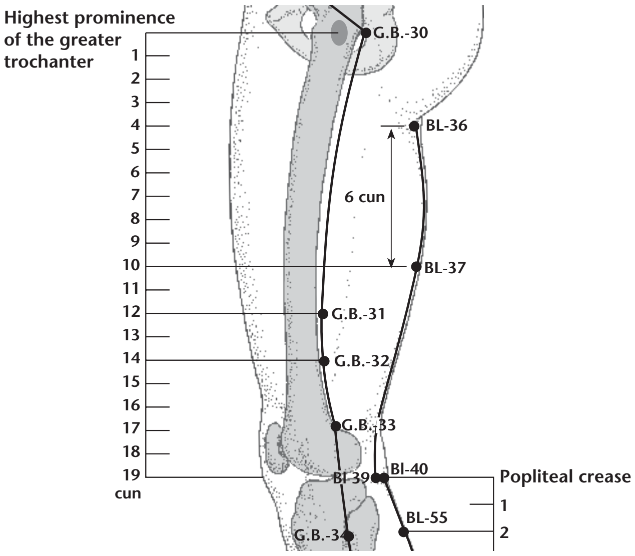
Highest prominence of the greater trochanter
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
Bl-40
Bl-39
BL-55
G.B.-34
Popliteal crease
1
2

Middle of the Crook WEIZHONG BL-40

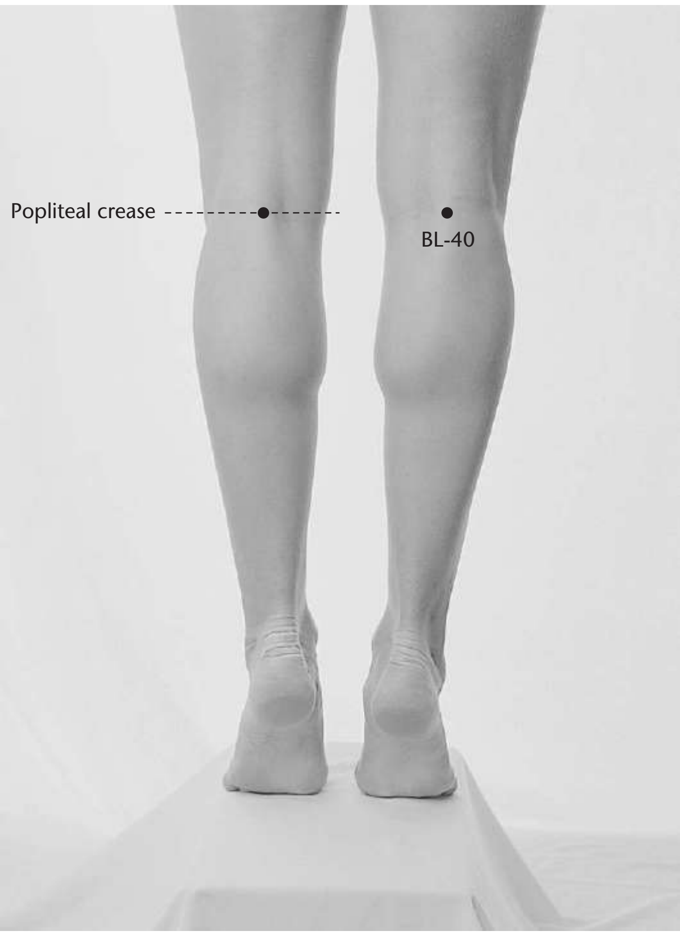

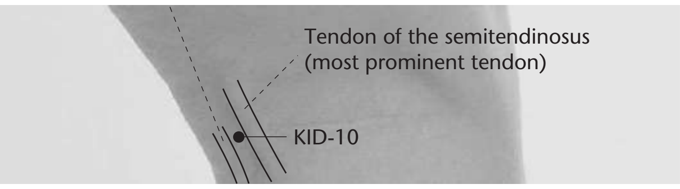

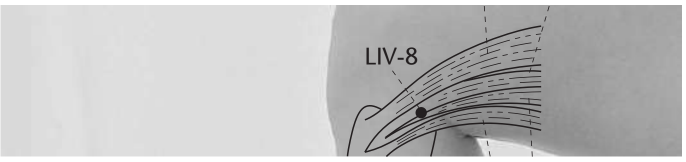

Location

In the centre of the popliteal crease, between the tendons of the biceps femoris and semitendinosus muscles.

How to find

This point is best located with the patient's knee slightly flexed. Locate the centre of the popliteal crease (a pulse may be palpable) and there locate **BL-40**.
Located on the same level are → **BL-39** (1 cun laterally), → **KID-10** (more medially, between the tendons of the semimembranosus and semitendinosus muscles) and → **LIV-8** (more medially, anterior to the tendons of the semimembranosus and semitendinosus muscles).

Needling

Vertically 0.5–1.5 cun. Caution: Popliteal nerve, artery and vein lie deep to this point. Use moxibustion with caution (contraindicated according to some authors). To clear Heat and eliminate stasis: bleed the superficial veins, possibly followed by brief cupping. Caution: Only use for excess conditions and in constitutionally robust patients. **BL-40** is also indicated as a distal point for the lower back: needle with the patient standing (for safety and balance, use a treatment couch or chair as a support for the hands) and ask them to move their lower back gently while manipulating the needle (only short needle retention). When needling this point with the patient in a prone position, place a supporting roll or pillow under the ankles to relax the popliteal crease.

Actions/Indications

- Clears (summer) Heat, stops vomiting and diarrhoea
- Cools the Blood
- Benefits the lower back and knees, opens the channel and *luo* vessels, alleviates pain

Special features

He-sea point, Earth point, lower *he*-sea point of the Bladder, Ma Dan Yang Heavenly Star point, Gao Wu command point of the lumbar region. Important distal point for the lower back, point with a wide range of indications.

BL-41 *Attached Branch* FUFEN

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd thoracic vertebra (T2).

How to find

After locating the spinous process of C7 (→ 3.4.1), count down 2 spinous processes to the lower border of the spinous process of T2. Locate **BL-41** 3 cun lateral from the midline (with the shoulders relaxed and the arms hanging loosely down, 3 cun corresponds on this level to the distance from the midline to the medial border of the scapula). Note: From **BL-41** to → **BL-54**, all points on the BL channel are located on its outer branch.

Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline), → **BL-12** (1.5 cun lateral to the midline) and → **S.I.-13** (more laterally, on the scapula).

Needling

Obliquely 0.3–0.5 cun. Caution: Pneumothorax.

Actions/Indications

- Expels pathogenic factors such as Wind and Cold
- Opens the channel and the *luo* vessels, alleviates pain

Special features

Meeting point with the S.I. channel

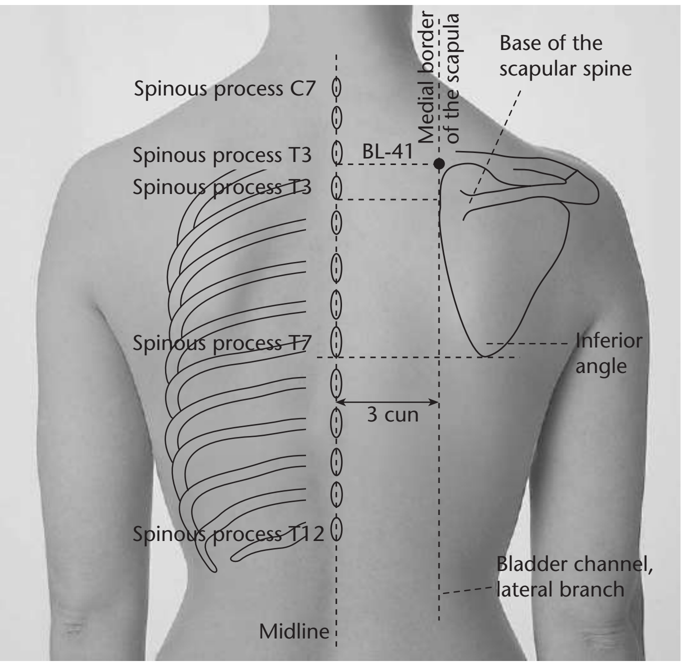
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
BL-41
Spinous process T3
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

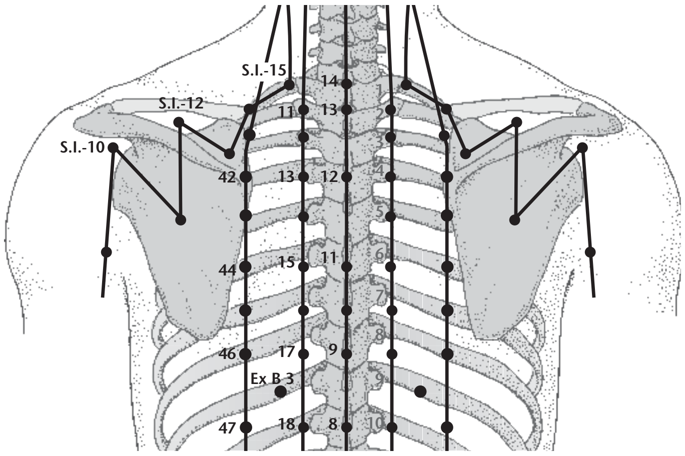
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex B 3
47
18
8

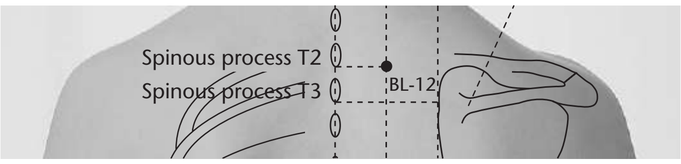
Spinous process T2
Spinous process T3
BL-12

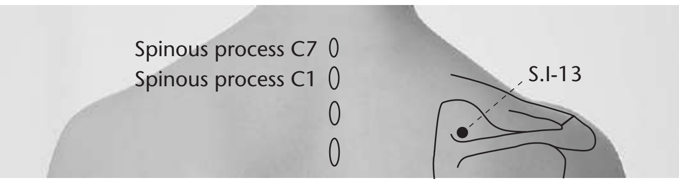
Spinous process C7
Spinous process C1
S.I-13

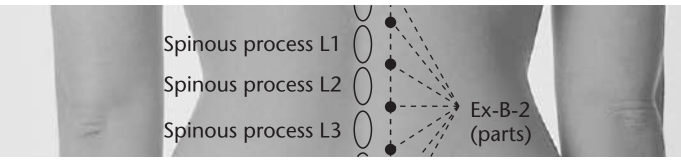
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

Door of the Corporeal Soul POHU BL-42

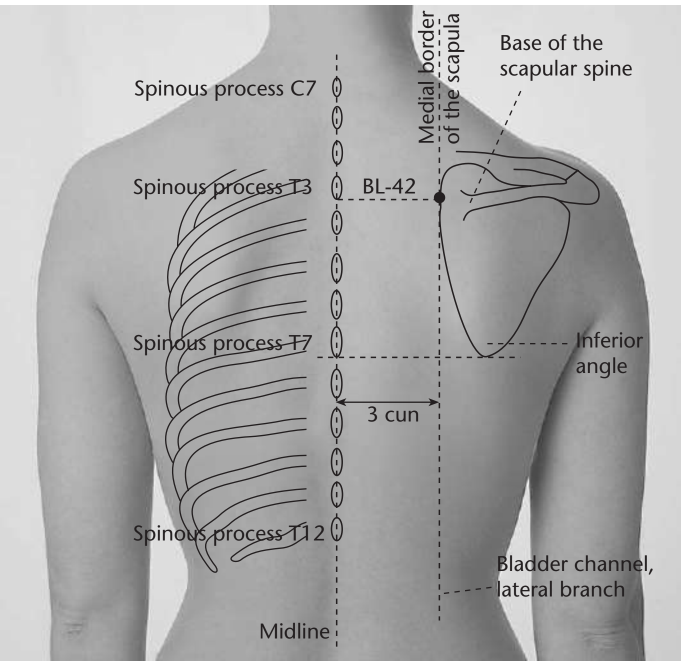
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
BL-42
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

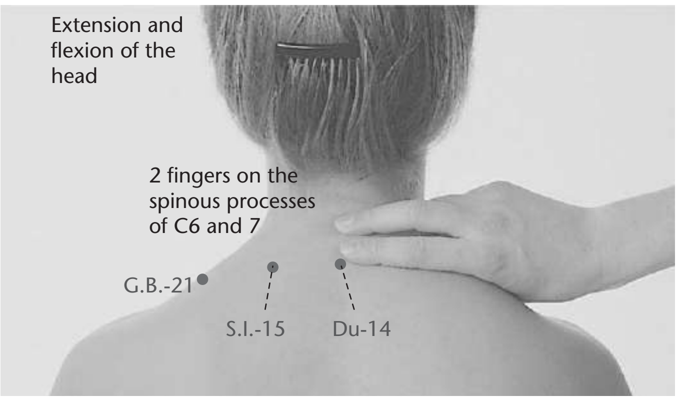
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

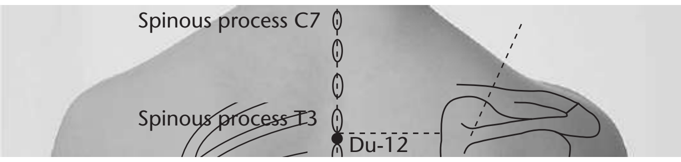
Spinous process C7
Spinous process T3
Du-12

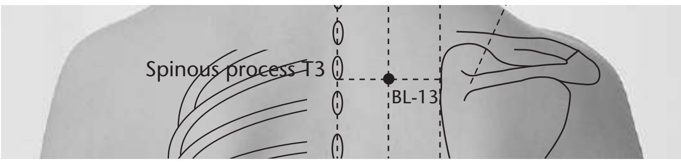
Spinous process T3
BL-13

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 3rd thoracic vertebra (T3).

How to find

After locating the spinous process of C7 (→ 3.4.1), count down 3 spinous processes to the lower border of the spinous process of T3. Located on this level and 3 cun lateral to the midline is **BL-42.**

Or: With the patient seated and their shoulders in a relaxed position, the spinous process of T3 is generally on the level of the medial end of the scapular spine, an easily palpable structure (→ 3.4.2). Locate **BL-42** on the level of the lower border of the spinous process of T3, 3 cun lateral to the midline. With the patient's shoulders relaxed, 3 cun corresponds on this level to the distance from the midline to the medial border of the scapula.

Located on the same level are → **Du-12** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-13** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.5 cun. Caution: Pneumothorax.

Actions/Indications

- Tonifies and nourishes the Lung
- Alleviates wheezing and coughing, calms the corporeal soul (*po*)
- Opens the channel and alleviates pain
- Clears Heat from the Lung

BL-43 *Vital Region* Shu GAOHUANG

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 4th thoracic vertebra (T4).

How to find

After locating the spinous process of C7 (➞ 3.4.1), count down 4 spinous processes to the lower border of the spinous process of T4. Located on this level and 3 cun lateral to the midline is **BL-43. Or:** Locate **BL-43** by using the spinous process of T3 (generally, on the level of the easily palpable medial end of the scapular spine, 3.4.2) as a reference point.

Located on the same level are a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-14** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.5 cun. Moxibustion is often used on this point. Caution: Pneumothorax.

Actions/Indications

- Tonifies and nourishes the five *zang*-Organs: Lung, Heart, Kidneys, Stomach and Spleen
- Nourishes Yin, clears Heat
- Calms the *shen*
- Strengthens the Original Qi (*yuan qi*)
- Eliminates Phlegm

Special features

Important point for deficiency syndromes (classic indication)

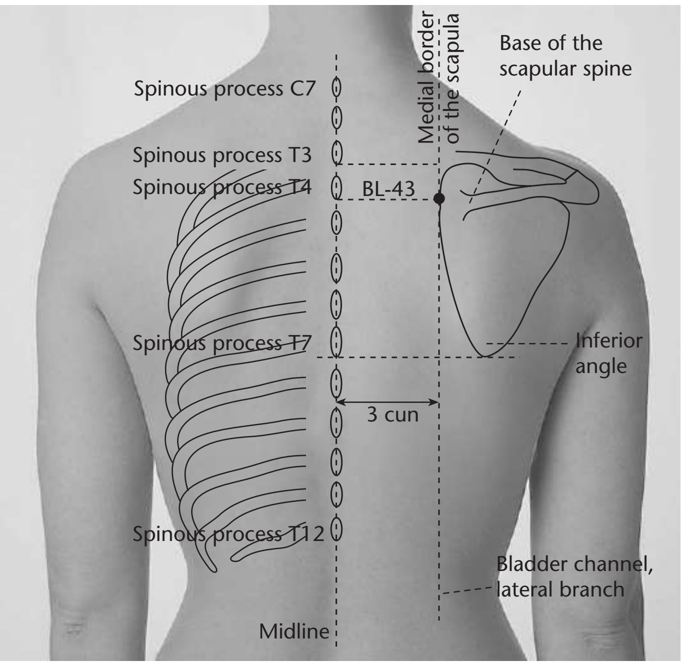
Medial border of the scapula
Base of the scapular spine
Spinous process C7
Spinous process T3
Spinous process T4
BL-43
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

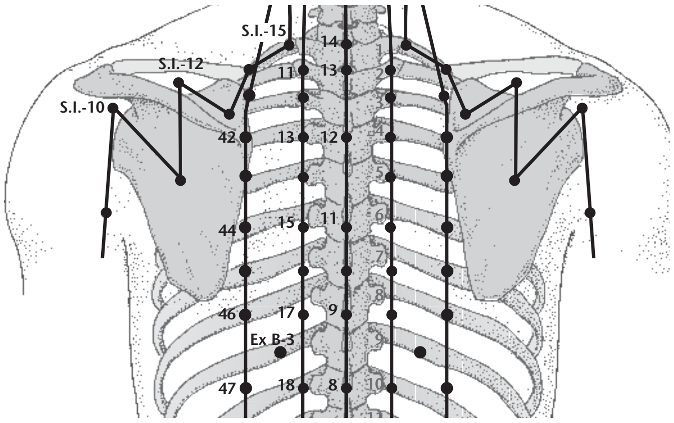
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
6
7
8
46
17
9
Ex B-3
47
18
8
10

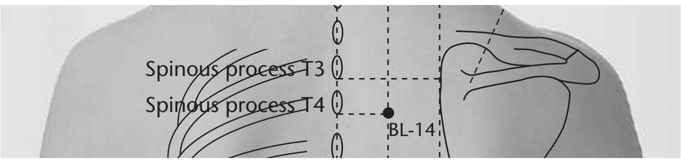
Spinous process T3
Spinous process T4
BL-14

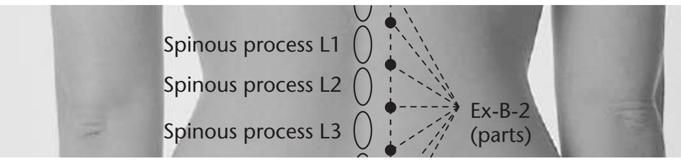
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Spirit Hall SHENTANG BL-44

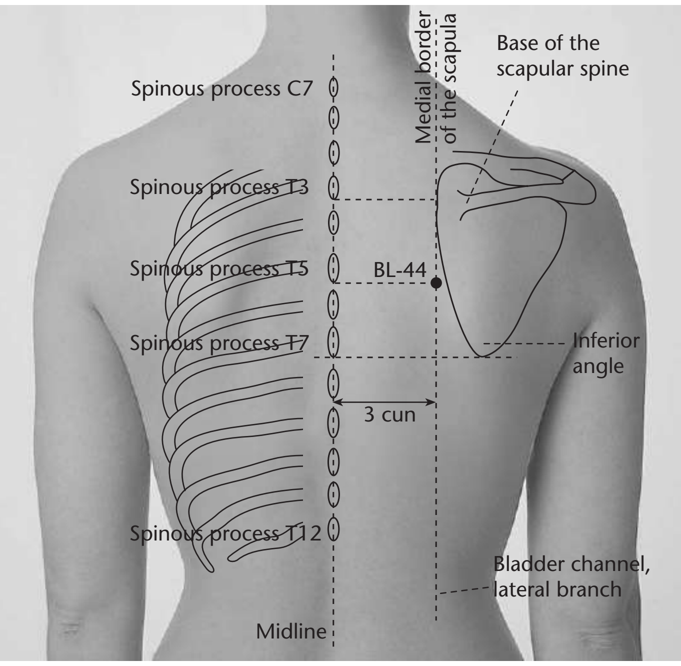
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T5
BL-44
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

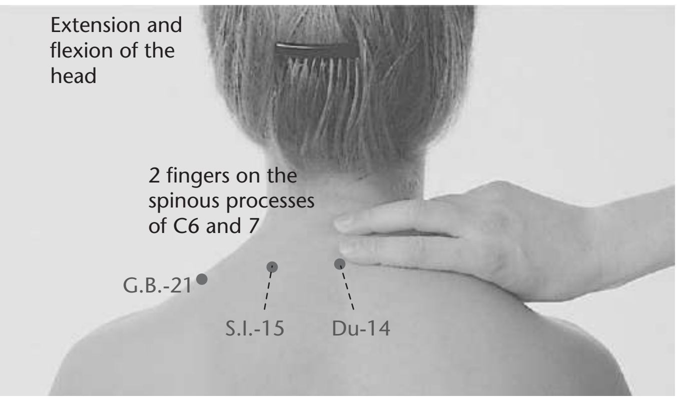
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

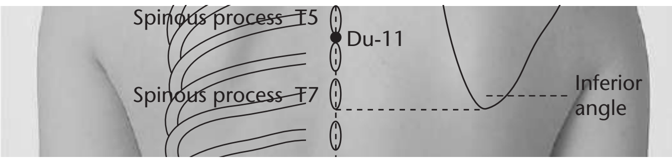
Spinous process T5
Du-11
Spinous process T7
Inferior angle

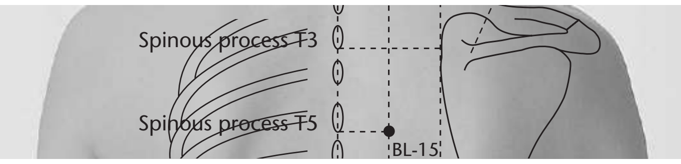
Spinous process T3
Spinous process T5
BL-15

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 5th thoracic vertebra (T5).

How to find

After locating the spinous process of C7 (→ 3.4.1), count down 5 spinous processes to the lower border of the spinous process of T5. Located on this level and 3 cun lateral to the midline is **BL-44. Or:** With the patient seated and their shoulders in a relaxed position, the spinous process of T3 is generally on the level of the medial end of the scapular spine, an easily palpable structure (→ 3.4.2). Locate **BL-44** by counting down to the level of the lower border of the spinous process of T5 and measure 3 cun lateral to the midline.

Located on the same level are → **Du-11** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-15** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Regulates the Upper Burner, unbinds the chest
- Opens the channel and alleviates pain

BL-45 *Cry of Pain* YIXI

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 6th thoracic vertebra (T6).

How to find

After locating the spinous process of C7 (→ 3.4.1), count down 6 spinous processes to the lower border of the spinous process of T6. Located on this level and 3 cun lateral to the midline is **BL-45. Or:** With the patient seated and their shoulders in a relaxed position, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). Locate **BL-45** on the level of the spinous process of T6, 3 cun lateral to the midline. Located on the same level are → **Du-10** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-16** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Expels pathogenic factors, especially Wind and Heat
- Descends the Lung Qi and unbinds the chest
- Strengthens the Qi and Blood and alleviates pain

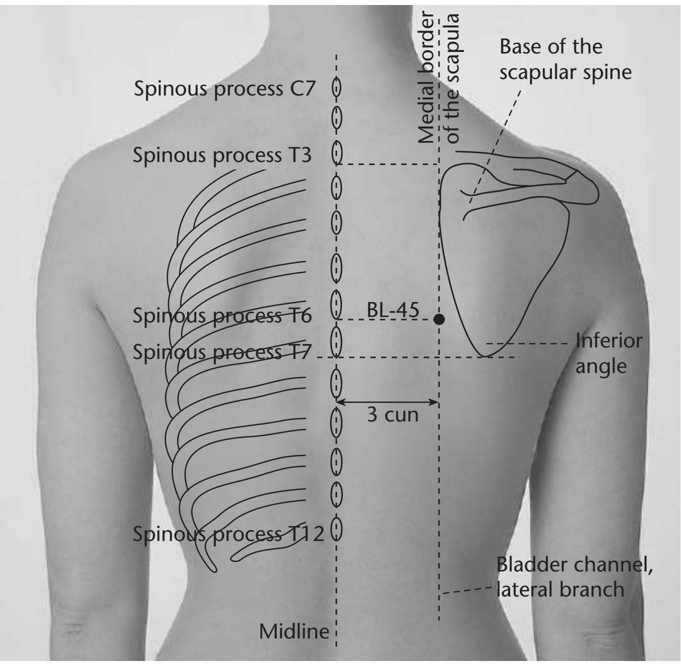
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T6
BL-45
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

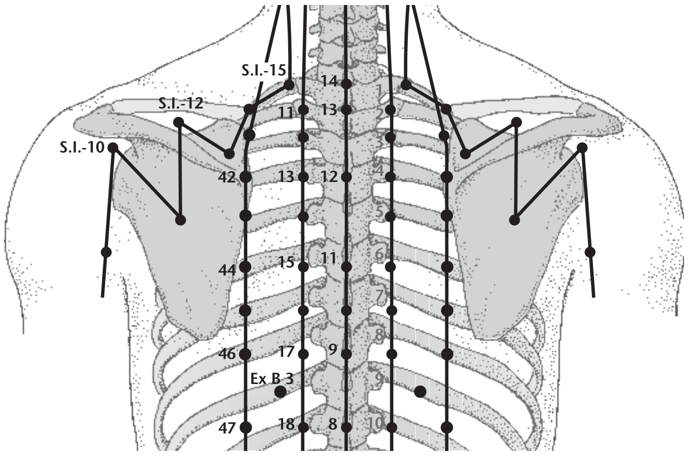
S.I.-15
S.I.-12
S.I.-10
14
1
11
13
2
3
42
13
12
4
5
44
15
11
6
7
8
46
17
9
Ex B 3
9
47
18
8
10

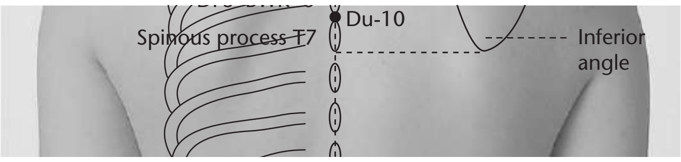
Du-10
Spinous process T7
Inferior angle

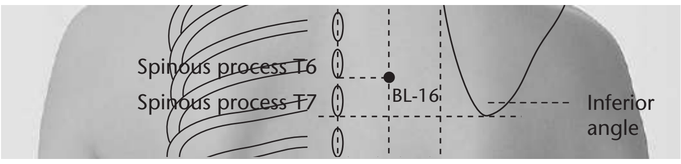
Spinous process T6
Spinous process T7
BL-16
Inferior angle

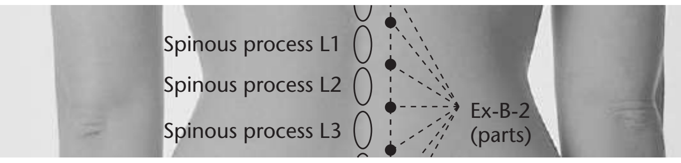
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Diaphragm's Gate GEGUAN BL-46

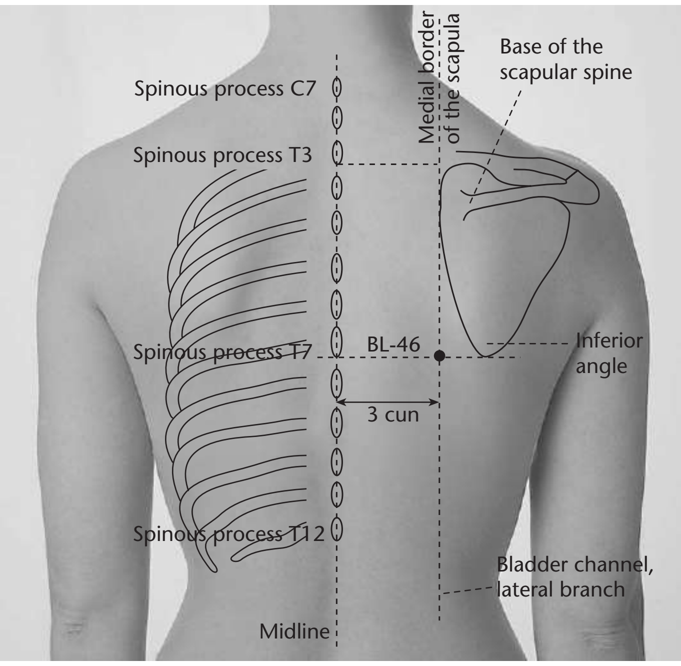
Medial border of the scapula
Base of the scapular spine
Spinous process C7
Spinous process T3
Spinous process T7
BL-46
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

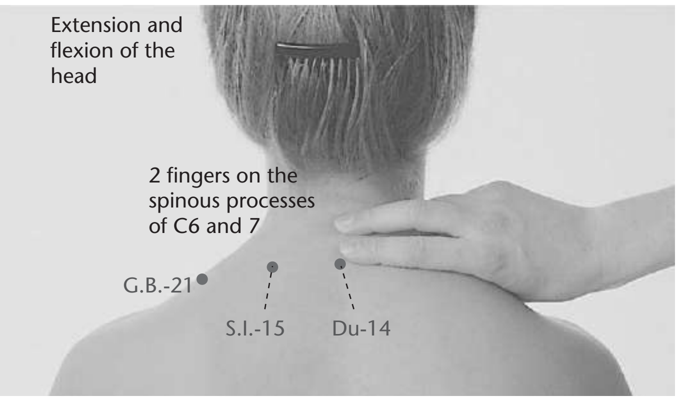
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

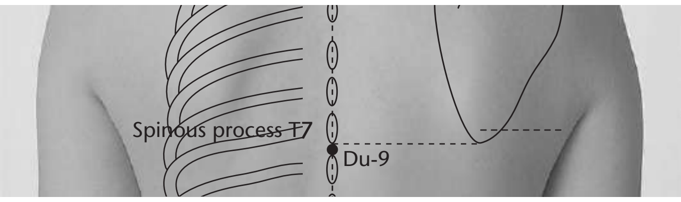
Spinous process T7
Du-9

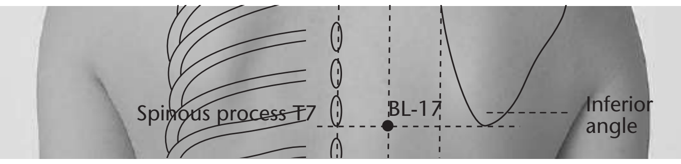
Spinous process T7
BL-17
Inferior angle

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 7th thoracic vertebra (T7).

How to find

With the patient standing and their shoulders in a relaxed position, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). Locate **BL-46** on the level of the spinous process of T7, 3 cun lateral to the midline. **Or:** With the patient seated, the spinous process of T3 is generally on the level of the medial end of the scapular spine, an easily palpable bony protrusion at the medial border of the scapula (→ 3.4.2). From there, count down 4 spinous processes to the spinous process of T7. Locate **BL-46** on this level, 3 cun lateral to the midline. **Or:** Find the spinous process of C7 (→ 3.4.1). From there, count down 7 spinous processes in an inferior direction to the spinous process of T7. Locate **BL-46** on this level, 3 cun lateral to the midline.

Located on the same level are → **Du-9** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-17** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Regulates the diaphragm and descends counterflow Qi
- Harmonises the Middle Burner
- Opens the channel and alleviates pain

BL-47 *Gate of the Ethereal Soul* HUNMEN

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 9th thoracic vertebra (T9).

How to find

In a sitting position with the arms hanging down, the spinous process of T7 is generally on the level of the inferior angle of the scapula (➞ 3.4.2). From there, count down 2 processes to the lower border of the spinous process of T9. **BL-47** is located 3 cun lateral to this point. **Or:** Locate the spinous process of C7 (➞ 3.4.1). From there, count down 9 processes to the lower border of the spinous process of T9 and locate **BL-47** 3 cun lateral to it.
Located on the same level are ➞ **Du-8** (on the midline), a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-18** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.5 cun. Caution: Pneumothorax.

Actions/Indications

- Spreads the Liver Qi
- Harmonises the Spleen and the Stomach
- Relaxes the tendons

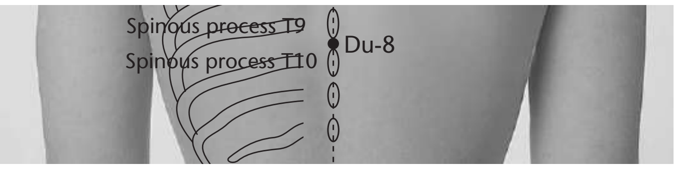
Spinous process T9
Spinous process T10
Du-8

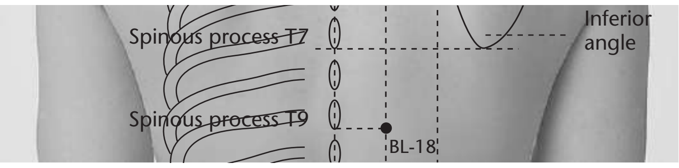
Spinous process T7
Spinous process T9
BL-18
Inferior angle

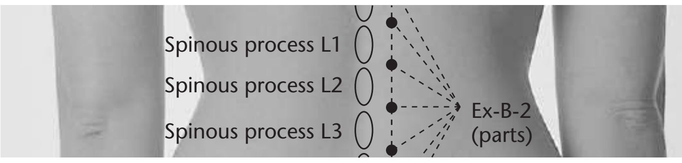
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

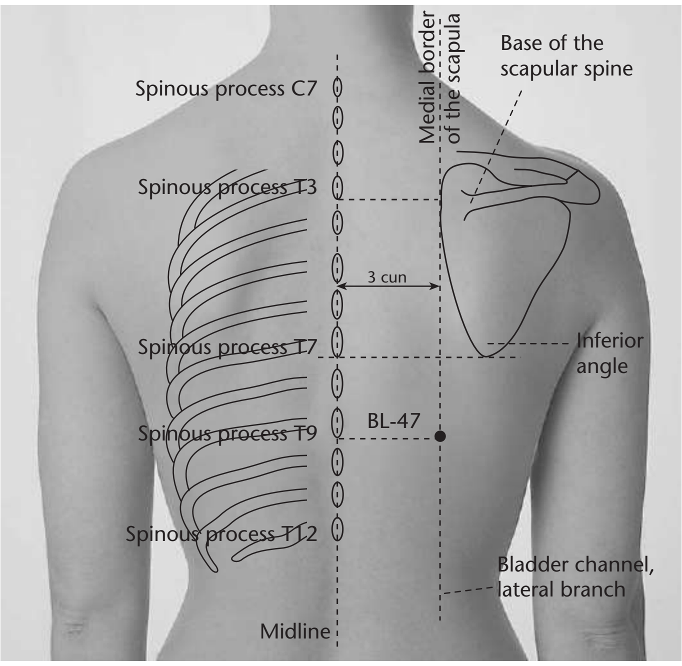
Spinous process C7
Spinous process T3
Spinous process T7
Spinous process T9
Spinous process T12
Medial border of the scapula
Base of the scapular spine
3 cun
Inferior angle
BL-47
Bladder channel, lateral branch
Midline

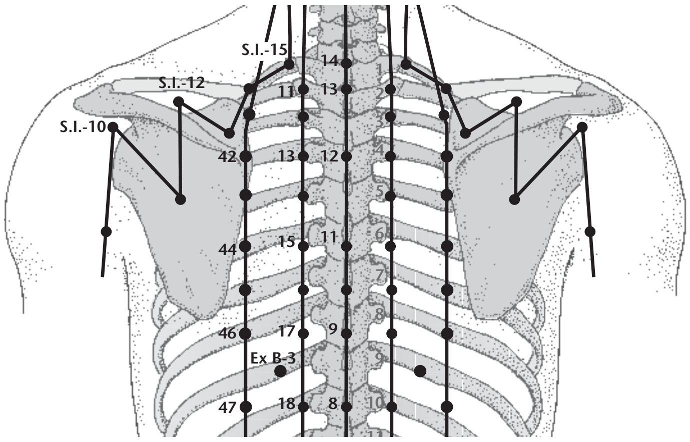
S.I.-15
S.I.-12
S.I.-10
14
11
13
42
13
12
44
15
11
46
17
9
Ex B-3
47
18
8
1
2
3
4
5
6
7
8
9
10
11

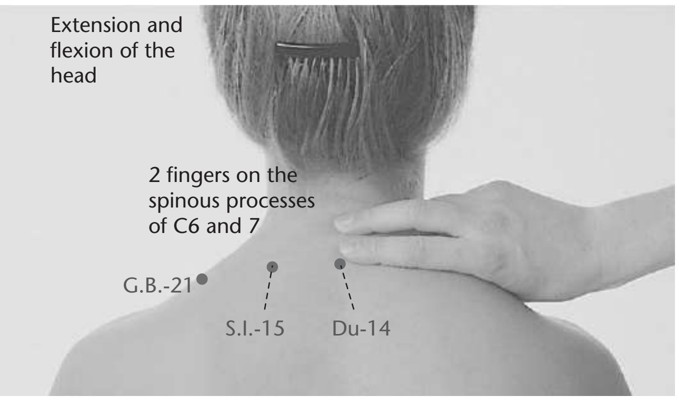
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Yang's Key Link YANGGANG **BL-48**

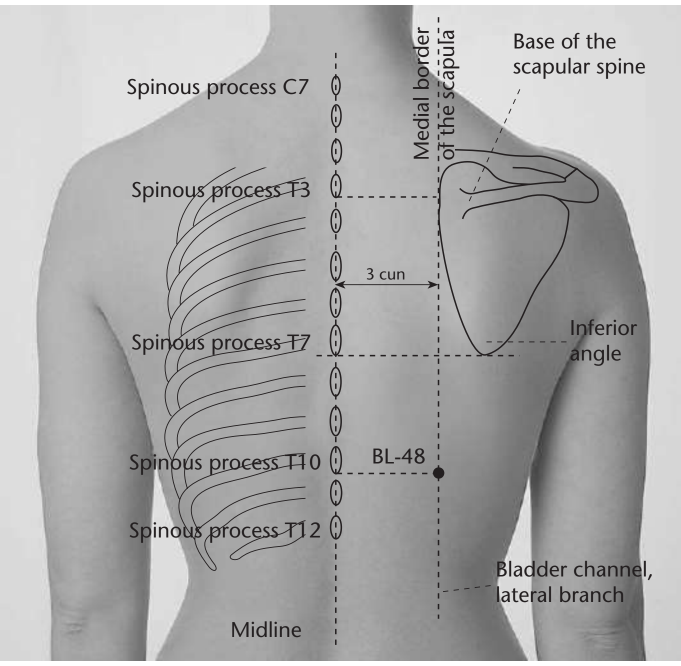
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
3 cun
Spinous process T7
Inferior angle
Spinous process T10
BL-48
Spinous process T12
Bladder channel, lateral branch
Midline

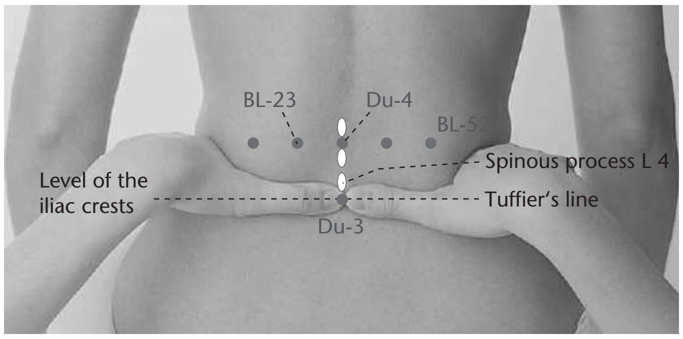
BL-23
Du-4
BL-52
Spinous process L 4
Level of the iliac crests
Tuffier's line
Du-3

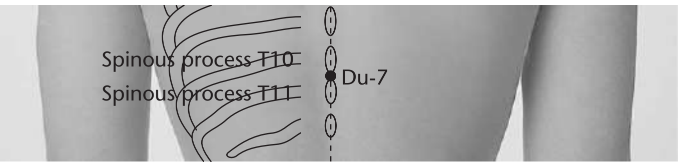
Spinous process T10
Spinous process T11
Du-7

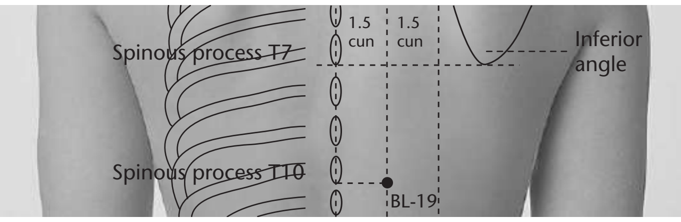
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T10
BL-19

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 10th thoracic vertebra (T10).

How to find

In a sitting position with the arms hanging down, the spinous process of T7 is generally on the level of the inferior angle of the scapula (➞ 3.4.2). From there, count down 3 processes to the lower border of the spinous process of T10. **BL-48** is located 3 cun lateral to this point. **Or:** Locate the spinous process of C7 (➞ 3.4.1). From there, count down 10 processes to the lower border of the spinous process of T10 and locate **BL-48** 3 cun lateral to it.

Located on the same level are ➞ **Du-7** (on the midline), a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-19** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Regulates the Gall Bladder
- Clears Damp-Heat
- Harmonises the Middle Burner

BL-49 *Abode of Thought* YISHE

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 11th thoracic vertebra (T11).

How to find

In a sitting position with the arms hanging down, the spinous process of T7 is generally on the level of the inferior angle of the scapula (➞ 3.4.2). From there, count down 4 spinous processes to the lower border of the spinous process of T11 and locate **BL-49** 3 cun lateral to this point. **Or:** From the articulation of the most inferior rib, palpate in a superior direction to T11. **Or:** Use the lumbar spine for reference (➞ 3.4.3).

Located on the same level are ➞ **Du-6** (on the midline), a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-20** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Clears Damp-Heat
- Harmonises the Spleen and the Stomach

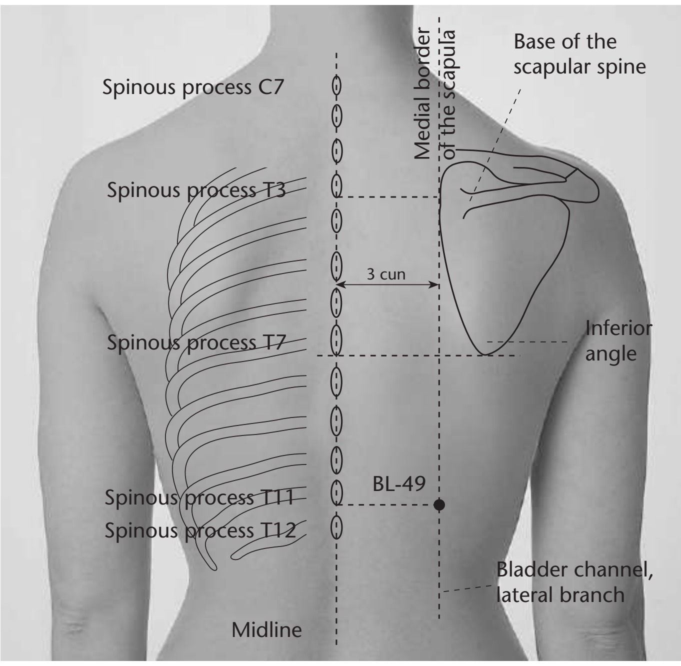
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
3 cun
Spinous process T7
Inferior angle
BL-49
Spinous process T11
Spinous process T12
Bladder channel, lateral branch
Midline

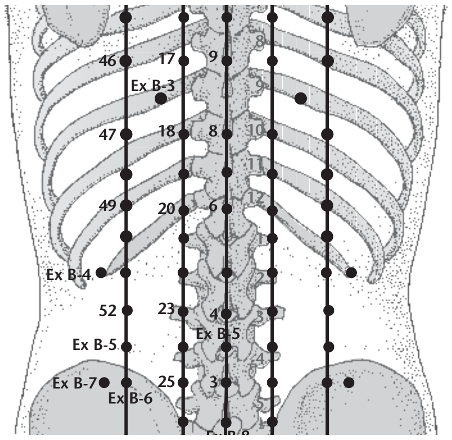
46
17
9
Ex B-3
47
18
8
49
20
6
Ex B-4
52
23
4
Ex B-5
Ex B-5
Ex B-7
25
3
Ex B-6

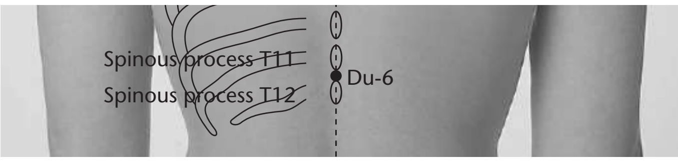
Spinous process T11
Spinous process T12
Du-6

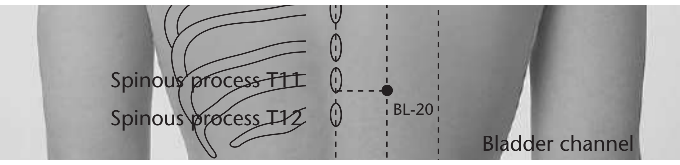
Spinous process T11
Spinous process T12
BL-20
Bladder channel

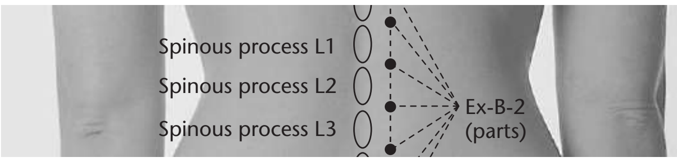
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Stomach Granary WEICANG BL-50

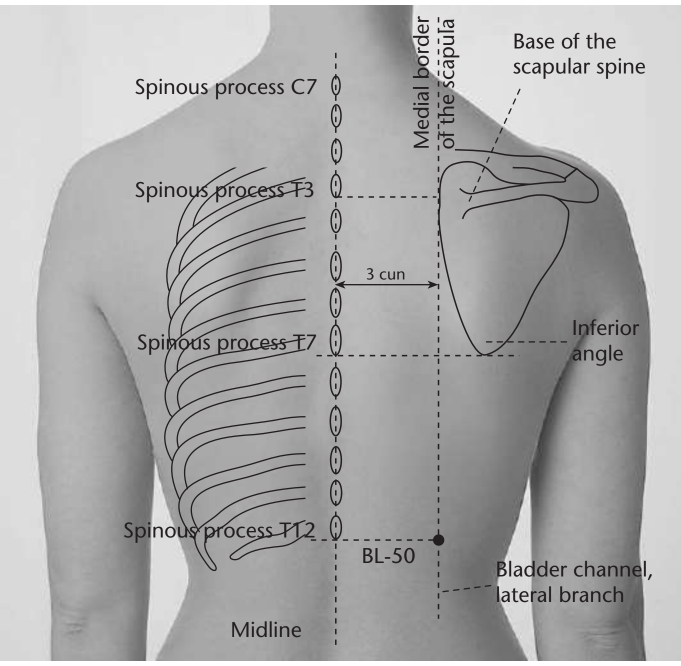
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
3 cun
Spinous process T7
Inferior angle
Spinous process T12
BL-50
Bladder channel, lateral branch
Midline

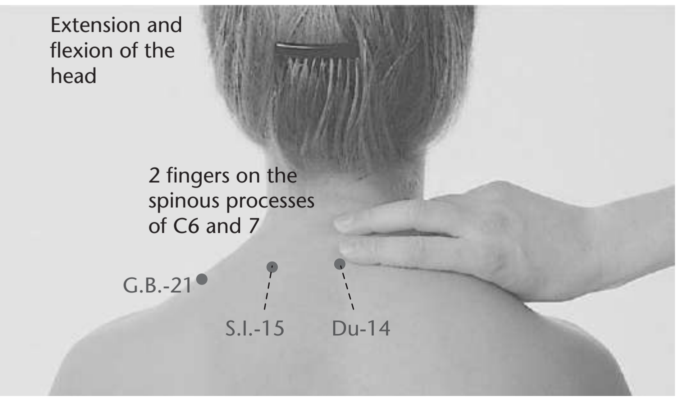
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

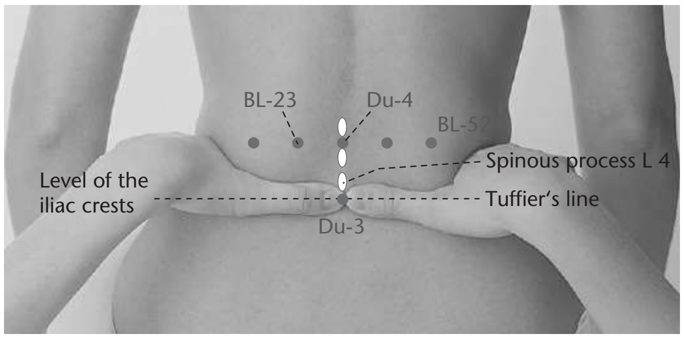
BL-23
Du-4
BL-52
Spinous process L 4
Level of the iliac crests
Tuffier's line
Du-3

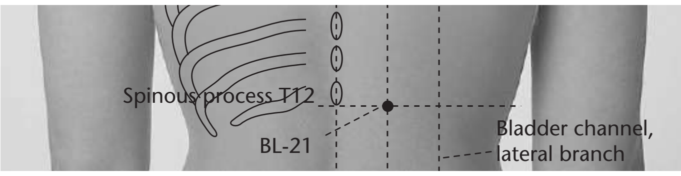
Spinous process T12
BL-21
Bladder channel, lateral branch

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 12th thoracic vertebra (T12).

How to find

In a sitting position with the arms hanging down, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down 5 spinous processes to the lower border of the spinous process of the T12 and locate **BL-50** 3 cun lateral to this point. **Or:** Locate the articulation of the most inferior rib with T12 and palpate 3 cun in a lateral direction for **BL-50. Or:** Orientation from the lumbar spine (→ 3.4.3). Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-21** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax, especially with asthenic patients or patients with emphysema.

Actions/Indications

- Harmonises the Middle Burner
- Opens the channel

BL-51 *Vitals Gate* HUANGMEN

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 1st lumbar vertebra (L1).

How to find

For orientation in the lumbar region (→ 3.4.3), first, the Tuffier's line: place both hands superior or lateral to the highest points of the iliac crest, joining the thumbs in the middle above the spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.). From there, count upward to the lower border of the spinous process of L1 and, on this level, locate **BL-51** 3 cun in a lateral direction. **Or:** With the patient seated and their arms hanging down in a relaxed way, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down six spinous processes and locate **BL-51** 3 cun lateral to the spinous process of L1.

Located on the same level are → **Du-5** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline), → **BL-22** (1.5 cun lateral to the midline) and → **Ex-B-4** (→ 3.5 cun lateral to the midline).

Needling

Vertically or obliquely 0.5–1 cun. Caution: Kidneys.

Actions/Indications

- Moves the Qi and resolves stagnation
- Benefits the breasts

Special features

Distal point for disorders of the breasts

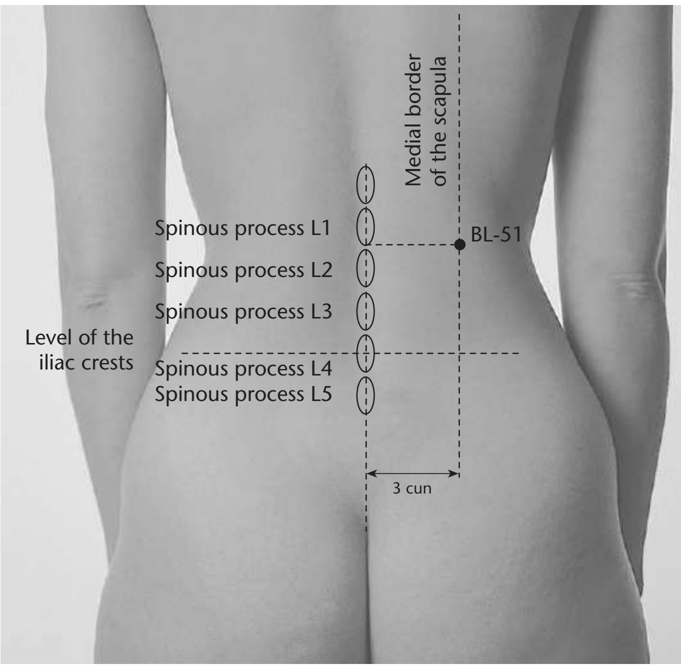
Medial border of the scapula
Spinous process L1
BL-51
Spinous process L2
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
3 cun

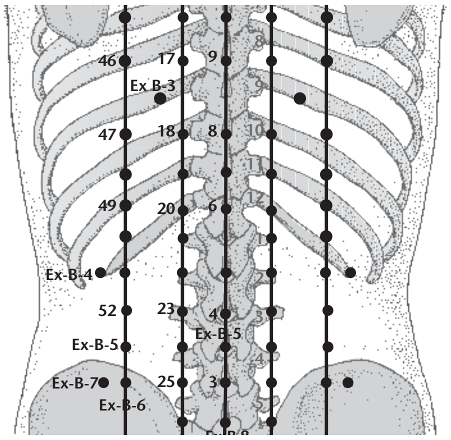
46
17
9
8
Ex B-3
9
47
18
8
10
11
49
20
6
12
1
Ex-B-4
2
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex-B-6
5

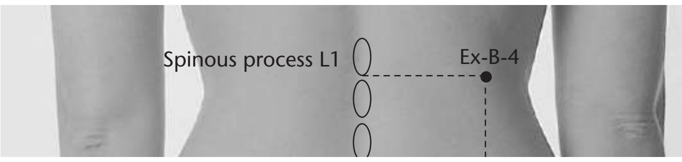
Spinous process L1
Ex-B-4

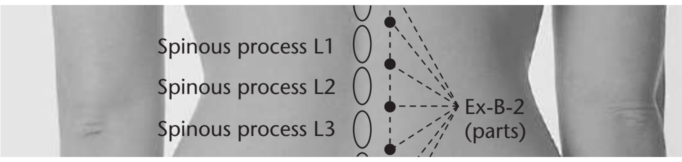
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

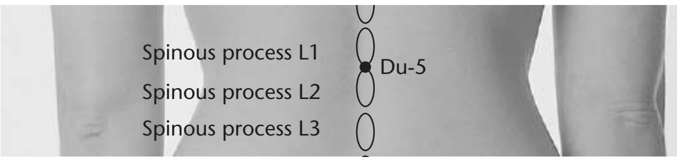
Spinous process L1
Du-5
Spinous process L2
Spinous process L3

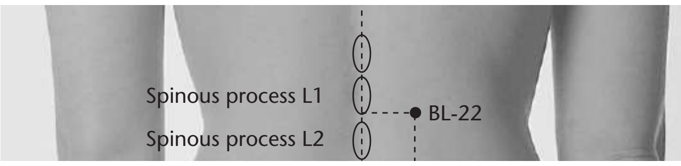
Spinous process L1
BL-22
Spinous process L2

Residence of the Will ZHISHI BL-52

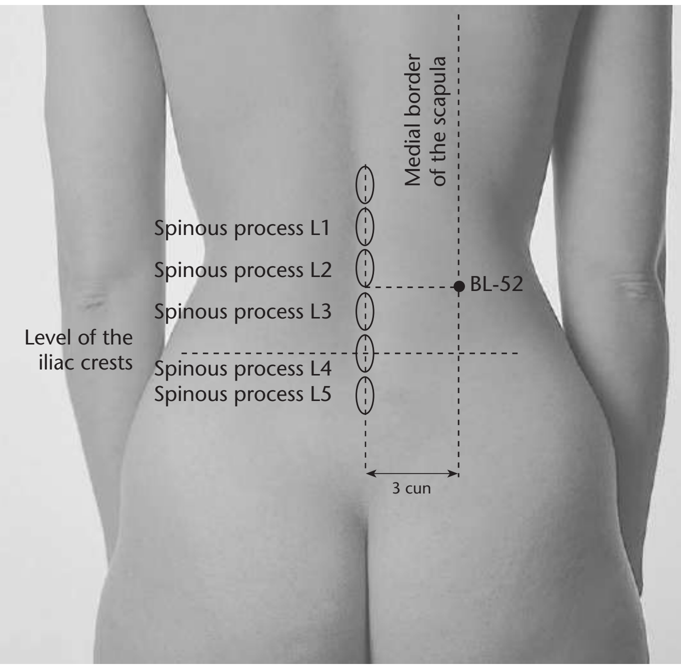
Medial border of the scapula
Spinous process L1
Spinous process L2
Spinous process L3
BL-52
Level of the iliac crests
Spinous process L4
Spinous process L5
3 cun

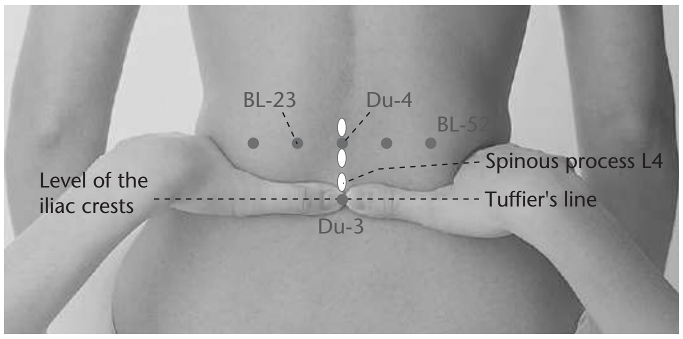
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

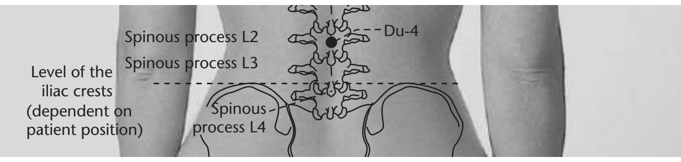
Spinous process L2
Du-4
Spinous process L3
Level of the iliac crests (dependent on patient position)
Spinous process L4

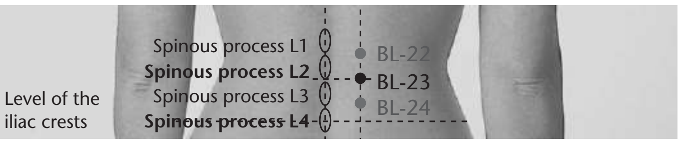
Spinous process L1
BL-22
Spinous process L2
BL-23
Spinous process L3
BL-24
Level of the iliac crests
Spinous process L4

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd lumbar vertebra (L2).

How to find

For orientation in the lumbar region (→ 3.4.3), first, locate the Tuffier's line: place both hands superior or lateral to the highest points of the iliac crest, joining the thumbs in the middle above the spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.). From there, count upward to the lower border of the spinous process of L2 and, on this level, locate **BL-52** 3 cun in a lateral direction.

Located on the same level are → **Du-4** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-23** (1.5 cun lateral to the midline).

Needling

Vertically or obliquely 0.5–1 cun. Caution: Kidneys.

Actions/Indications

- Tonifies the Kidneys and Essence (*jing*), regulates urination
- Benefits the lumbar region

BL-53 *Bladder Vitals* BAOHUANG

Location

3 cun lateral to the posterior midline, on the level of the 2nd sacral foramen.

How to find

For details on orientation in the lumbar and sacral region, → 3.4.3, 3.4.4. First, locate L5. The next palpable bony structure inferior to L5 is the highest process of the medial sacral crest. Palpate inferiorly for the 2nd process and locate the 2nd sacral foramen slightly less than 1 cun lateral to the process. Next, locate **BL-53** 3 cun lateral to the midline on the level of the 2nd foramen. **Or:** Use the sacral foramina as a reference point (→ 3.4.4). **BL-53** is located on the level of the 2nd foramen, 3 cun lateral to the midline.

Located on the same level are → **BL-32** (over the 2nd sacral foramen) and → **BL-28** (1.5 cun lateral to the midline).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Benefits the lumbar region, opens the channel and alleviates pain
- Regulates the Lower Burner

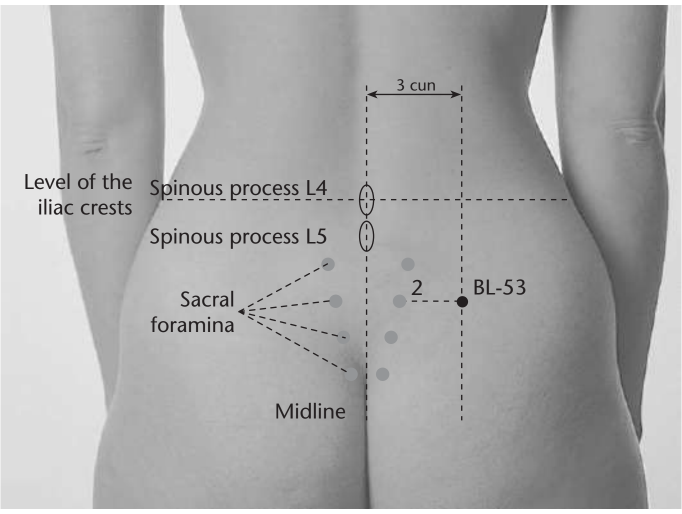
3 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
2
BL-53
Midline

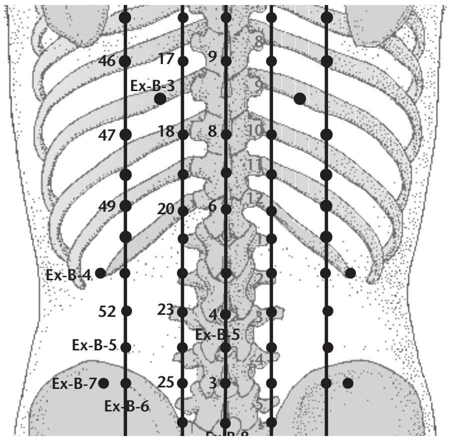
46
17
9
Ex-B-3
47
18
8
10
11
49
20
6
12
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6

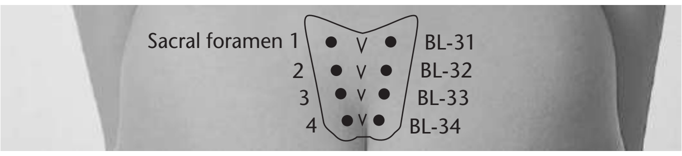
Sacral foramen 1
2
3
4
BL-31
BL-32
BL-33
BL-34

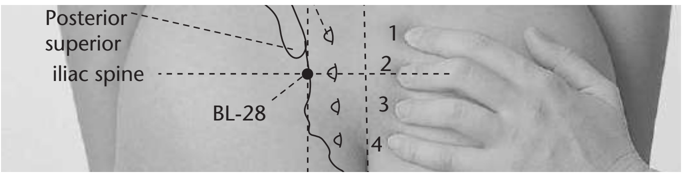
Posterior superior iliac spine
BL-28
1
2
3
4

Order's Limit ZHIBIAN BL-54

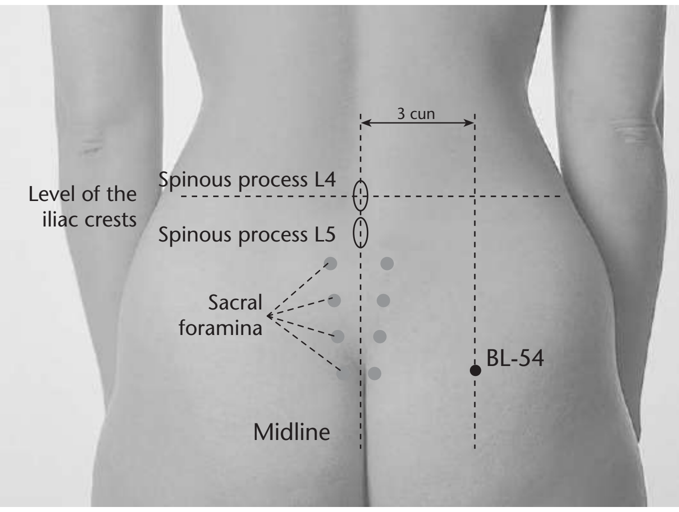
3 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
BL-54
Midline

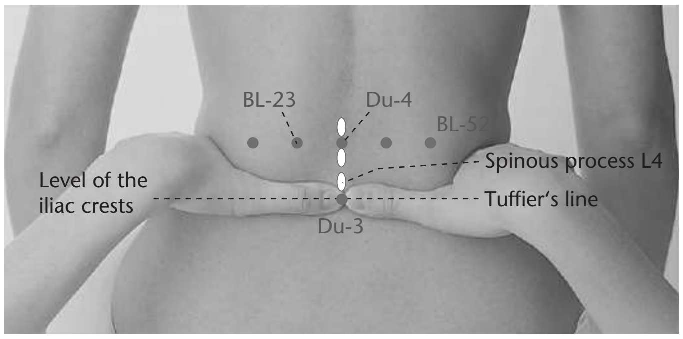
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

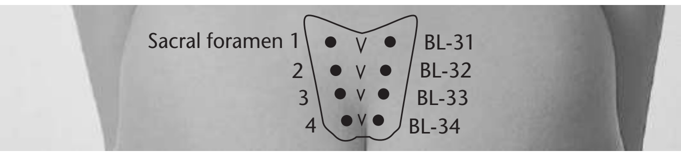
Sacral foramen 1
2
3
4
BL-31
BL-32
BL-33
BL-34

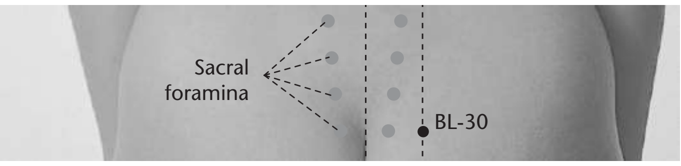
Sacral foramina
BL-30

Location

3 cun lateral to the posterior midline, on the level of the 4th sacral foramen, approximately at the centre of the buttock.

How to find

For details on orientation in the lumbar and sacral region, see → 3.4.3, 3.4.4. First, locate L5. Inferior to it, the next palpable bony structure is the highest point on the medial sacral crest (= sacral spinous processes). Palpate downward for 3 more bony elevations and locate the depression of the 4th sacral foramen less than 1 cun lateral to it. This is on the level of **BL-54**, located 3 cun lateral to the midline. For further reference: The 4th sacral foramen is mostly level with the superior end of the rima ani. Located on the same level are → **BL-34** (over the 4th sacral foramen) and → **BL-30** (1.5 cun lateral to the midline).

Needling

Vertically 1.5–2 cun or 2–3 cun in the direction of the anus or the genital region, depending on the indication.

Actions/Indications

- Opens the channel and *luo* vessels, eliminates stagnation
- Regulates urination
- Eliminates haemorrhoids

Special features

Important local point

BL-55 *Confluence of Yang* HEYANG

Location

2 cun inferior to the midpoint of the popliteal crease, in a depression between the two bellies of the gastrocnemius muscle.

How to find

Locate the popliteal crease by flexing the knee. In slim patients, palpate for the joint space. From the centre of the popliteal crease, palpate 2 cun in an inferior direction and locate **BL-55** in a depression between the two bellies of the gastrocnemius muscle.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Opens the channel, alleviates pain (also in the urogenital region)
- Stops uterine bleeding

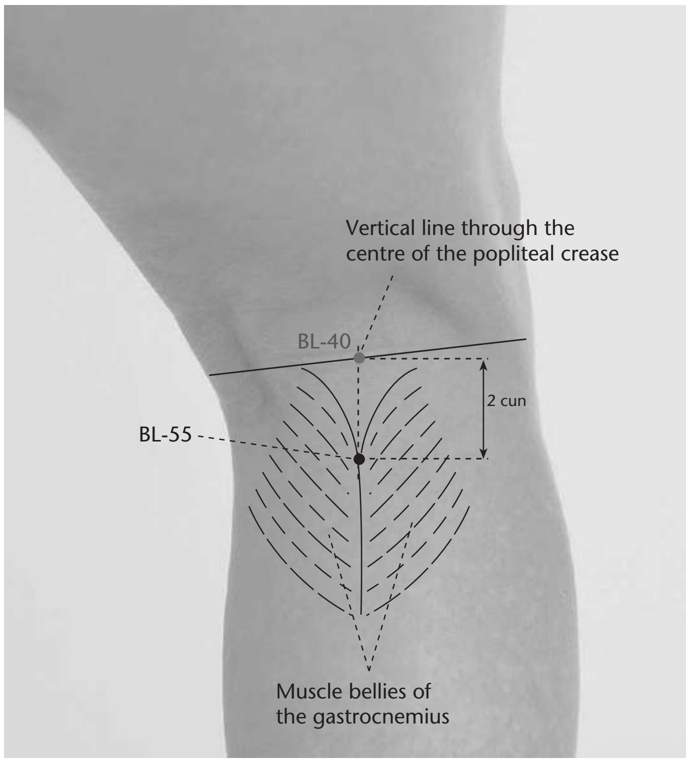
Vertical line through the centre of the popliteal crease
BL-40
2 cun
BL-55
Muscle bellies of the gastrocnemius

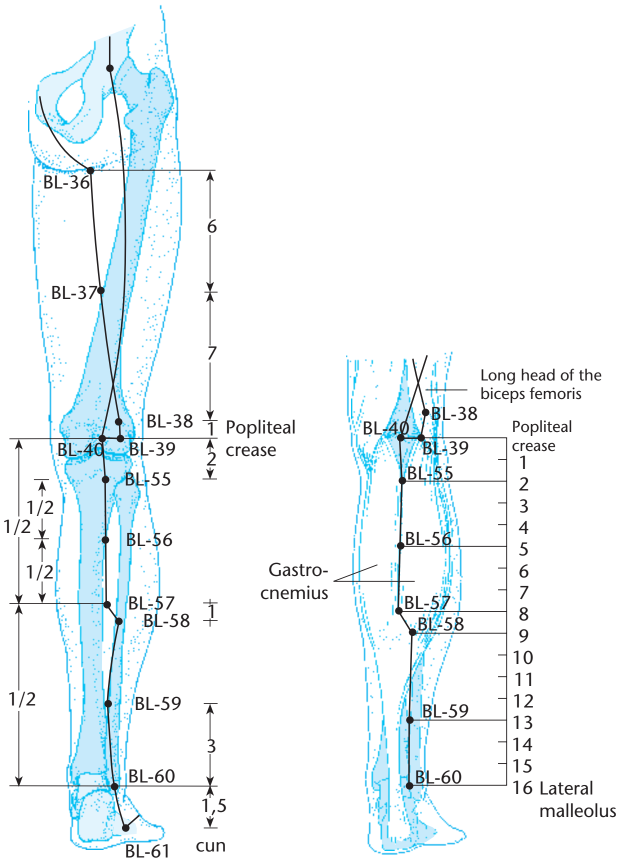
BL-36
6
BL-37
7
BL-38
Popliteal crease
BL-40
BL-39
2
BL-55
1/2
1/2
1/2
BL-56
BL-57
BL-58
1
1/2
BL-59
3
BL-60
1,5
BL-61
cun
Long head of the biceps femoris
BL-38
BL-40
Popliteal crease
BL-39
BL-55
BL-56
Gastro-cnemius
BL-57
BL-58
BL-59
BL-60
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16 Lateral malleolus

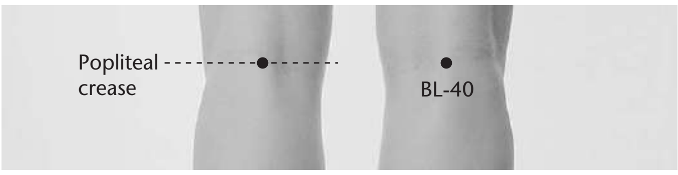
Popliteal crease
BL-40

Support the Sinews CHENGJIN BL-56

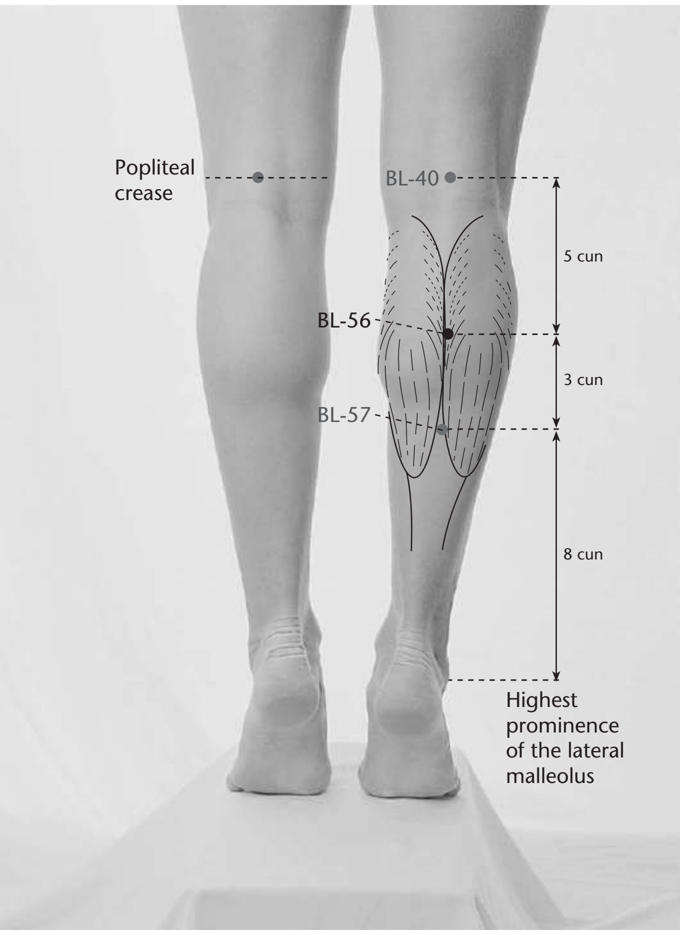
Popliteal crease
BL-40
5 cun
BL-56
3 cun
BL-57
8 cun
Highest prominence of the lateral malleolus

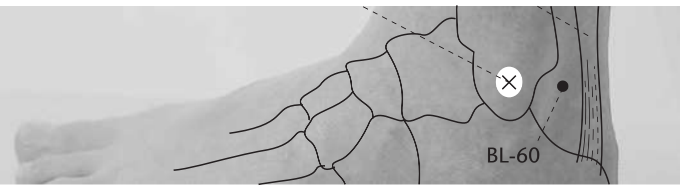

Location

5 cun inferior to the midpoint of the popliteal crease, between the two bellies of the gastrocnemius muscle.

How to find

Locate the popliteal crease by flexing the knee. In slim patients, palpate for the joint space. **BL-56** is located 5 cun distal to the popliteal crease, in a depression between the two bellies of the gastrocnemius muscle.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Relaxes the muscles and tendons
- Opens the channel
- Eliminates haemorrhoids

BL-57 *Supporting Mountain* CHENGSHAN

Location

In the middle of the calf, between the two heads of the gastrocnemius muscle, on a line connecting ➞ **BL-40** and ➞ **BL-60**, approximately 8 cun distal to ➞ **BL-40**.

How to find

Slide along the dorsal aspect of the Achilles tendon in a superior direction until you can feel the depression between the two bellies of the gastrocnemius muscle. This will become more defined when the patient's muscles are flexed. **Or:** Spreading hands technique (➞ 2.3.3): Place the little fingers on ➞ **BL-40** (centre of the popliteal crease) and ➞ **BL-60** (depression between the Achilles tendon and the highest prominence of the lateral malleolus). The midpoint of this distance, where the thumbs meet, marks the location of **BL-57**.

Located on the same level on the anterolateral aspect of the lower leg are ➞ **ST-38** (1 finger-width lateral to the tibia) and ➞ **ST-40** (2 fingerbreadths lateral to the tibia).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Relaxes the muscles and tendons, opens the channel
- Eliminates haemorrhoids (pathway of the Bladder divergent channel)

Special features

Ma Dan Yang Heavenly Star point. Important local point with distal effect on the lumbar and anal region.

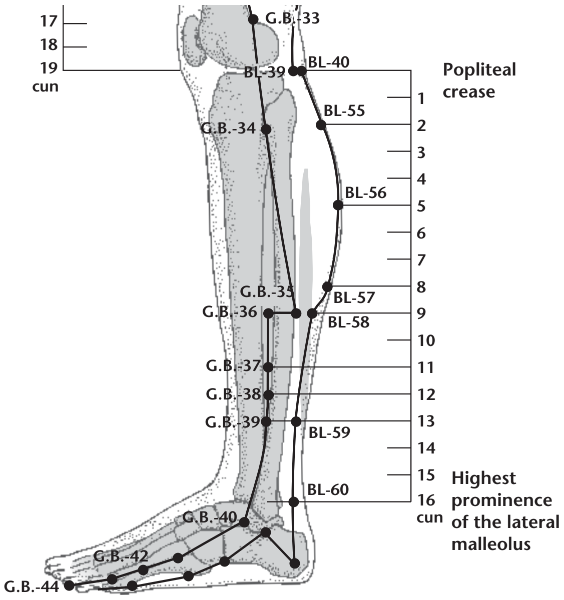
17
18
19
cun
G.B.-33
BL-40
BL-39
Popliteal crease
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun
BL-55
G.B.-34
BL-56
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
Highest prominence of the lateral malleolus
G.B.-40
G.B.-42
G.B.-44

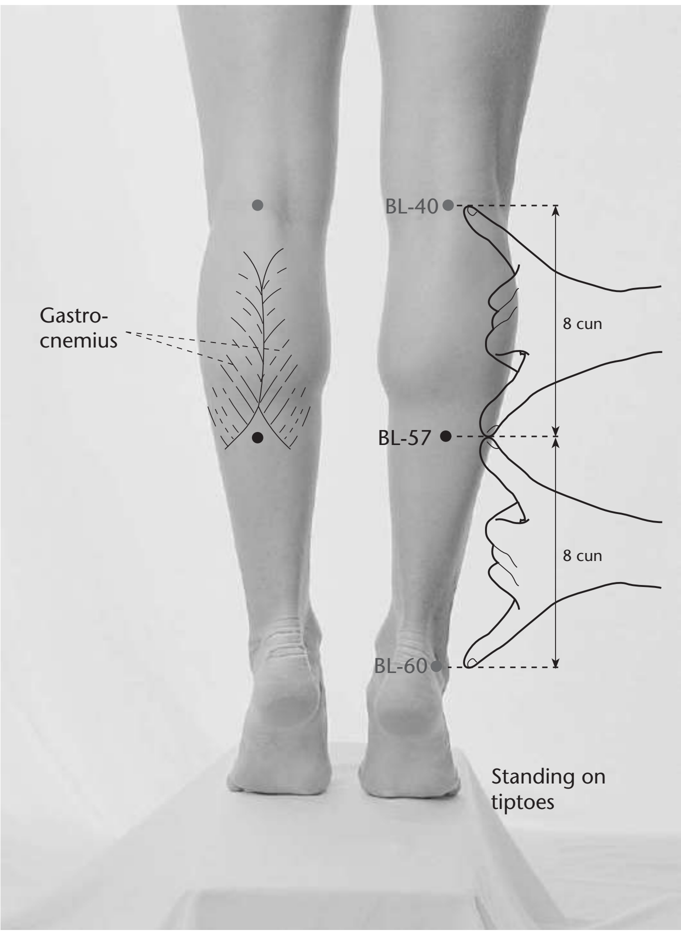
Gastro-
cnemius
BL-40
8 cun
BL-57
8 cun
BL-60
Standing on
tiptoes

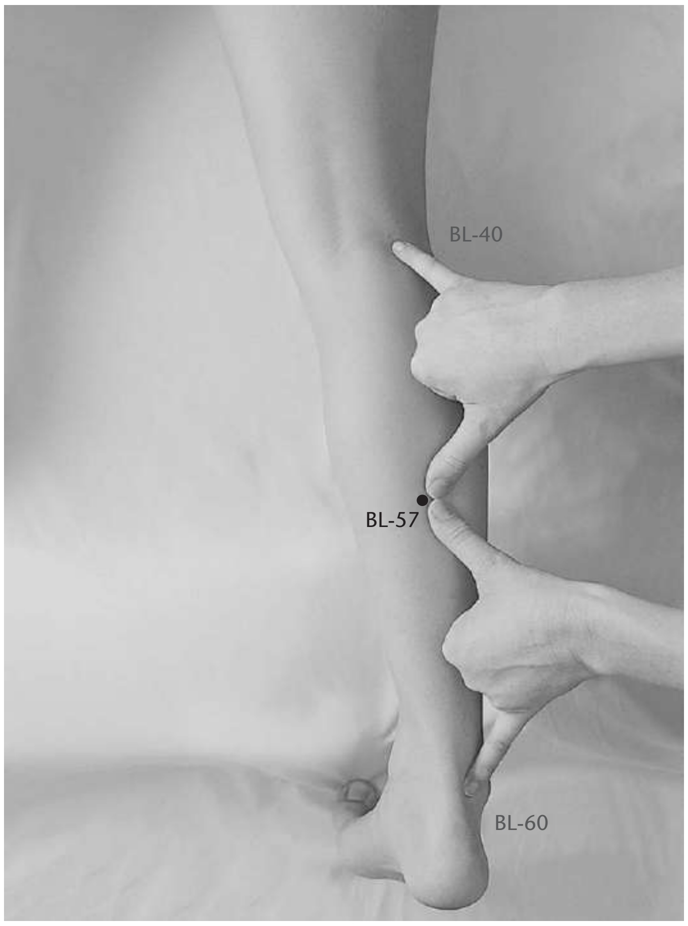
BL-40
BL-57
BL-60

Soaring Upward FEIYANG BL-58

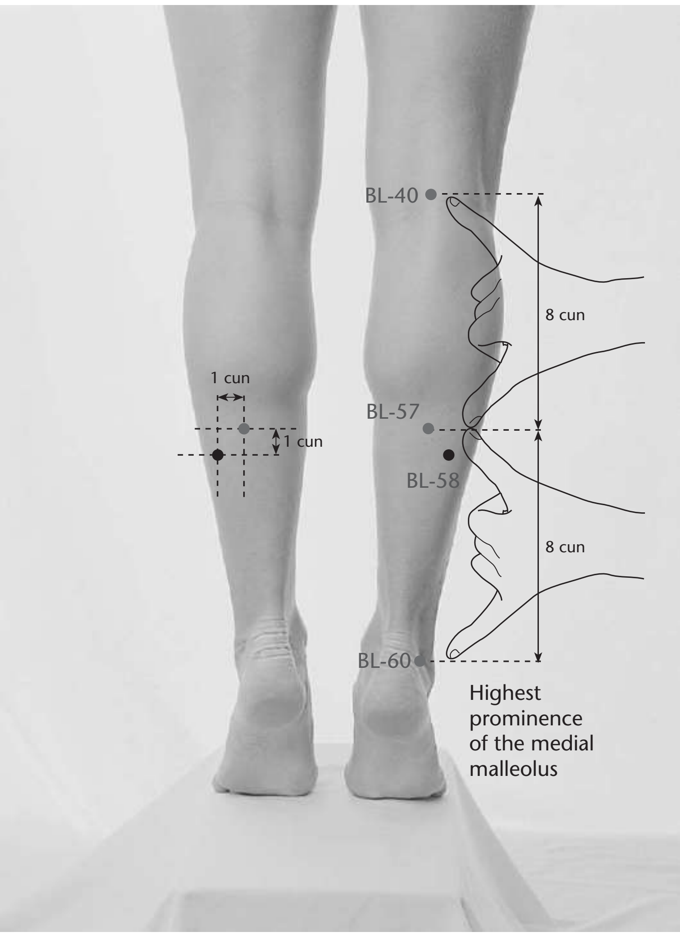
BL-40
8 cun
1 cun
BL-57
1 cun
BL-58
8 cun
BL-60
Highest prominence of the medial malleolus

Location

1 cun distal and 1 cun lateral to → **BL-57** or 7 cun proximal to → **BL-60**, at the posterior border of the fibula and at the lower border of the gastrocnemius muscle.

How to find

First, locate → **BL-57** (in the centre of the calf, between the two bellies of the gastrocnemius muscle). From → **BL-57**, palpate 1 cun in a distal direction and 1 cun in a lateral direction. There, locate **BL-58** on the lower border of the gastrocnemius muscle. Located on the same level (7 cun superior to the highest prominence of the lateral malleolus) are → **G.B.-35** (at the posterior border of the fibula), → **G.B.-36** (at the anterior border of the fibula) and → **ST-39** (1 cun distal to the midpoint of the line connecting → **ST-35** and → **ST-41** and 1 fingerbreadth lateral to the anterior crest of the tibia).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Opens the channel and *luo* vessels, alleviates pain
- Expels pathogenic factors from the *taiyang* channels (S.I., BL)
- Harmonises above and below
- Treats haemorrhoids

Special features

Luo-connecting point

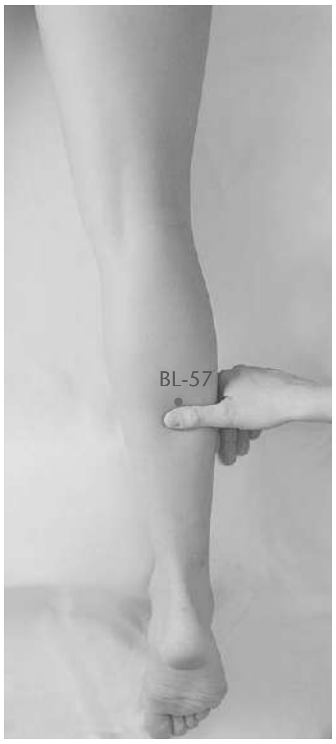

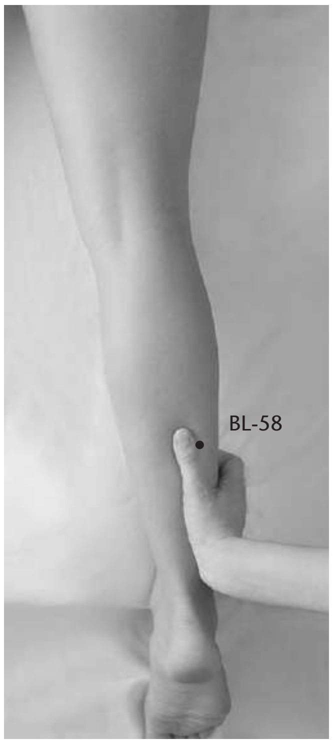

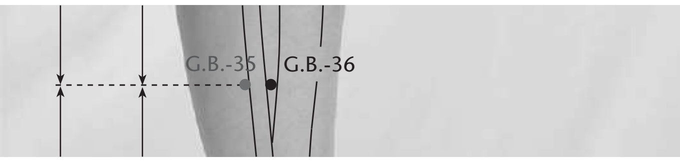

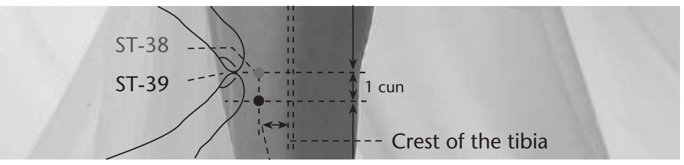
ST-38
ST-39
1 cun
Crest of the tibia

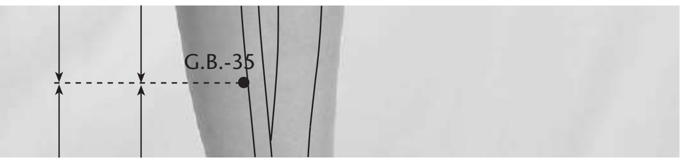

BL-59 *Instep Yang* FUYANG

Location

On the lateral aspect of the lower leg, 3 cun superior to → **BL-60** (in the depression between the highest prominence of the lateral malleolus and the Achilles tendon).

How to find

First, locate → **BL-60** in the depression between the highest prominence of the lateral malleolus and the Achilles tendon. With oedema of the lower leg, this point is not visible but palpable. From → **BL-60**, measure 3 cun (1 handbreadth) in a superior direction. There, locate **BL-59** in a depression between the Achilles tendon and the tendons of the peroneous longus and brevis muscles.

Located on the same level is → **G.B.-39**, 3 cun directly above the prominence of the lateral malleolus.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Opens the channel and *luo* vessels
- Benefits the lower back
- Benefits the head, activates the *yang qiao mai*

Special features

Xi-cleft point of the *yang qiao mai*, meeting point with the *yang qiao mai* according to some authors.

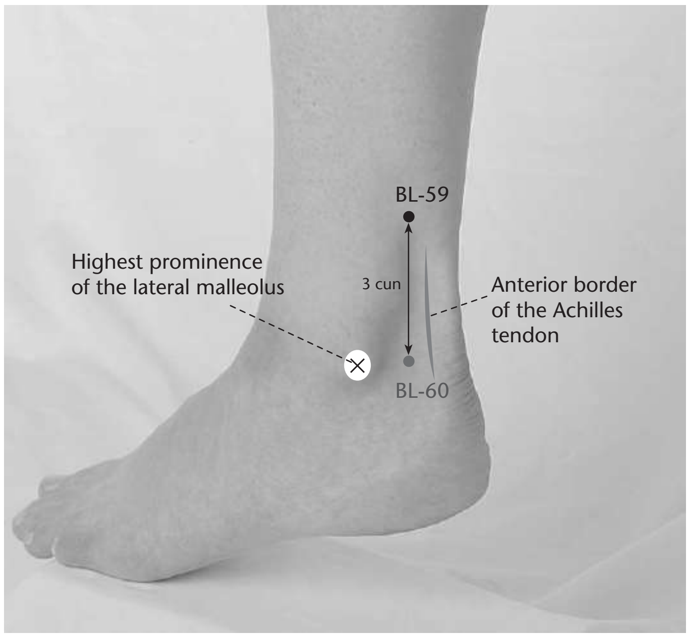
BL-59
Highest prominence of the lateral malleolus
3 cun
Anterior border of the Achilles tendon
BL-60

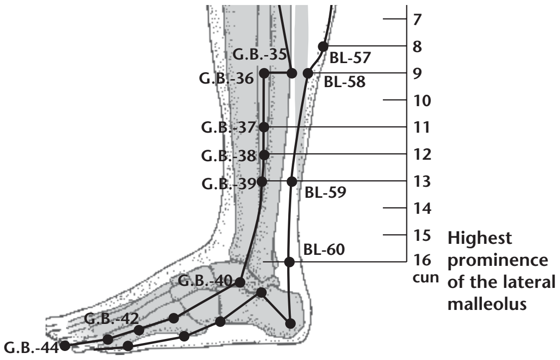
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
G.B.-40
G.B.-42
G.B.-44
7
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus

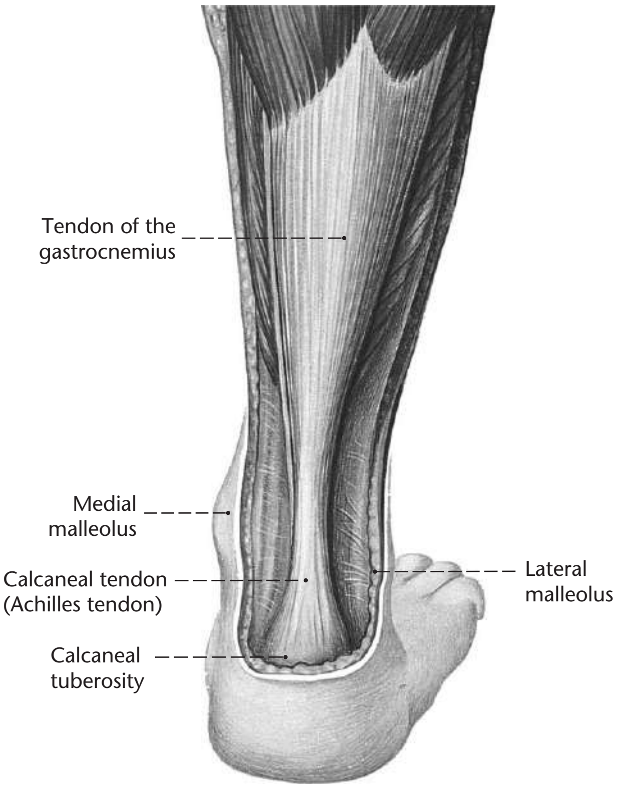
Tendon of the gastrocnemius
Medial malleolus
Calcaneal tendon (Achilles tendon)
Calcaneal tuberosity
Lateral malleolus

Kunlun Mountains KUNLUN BL-60

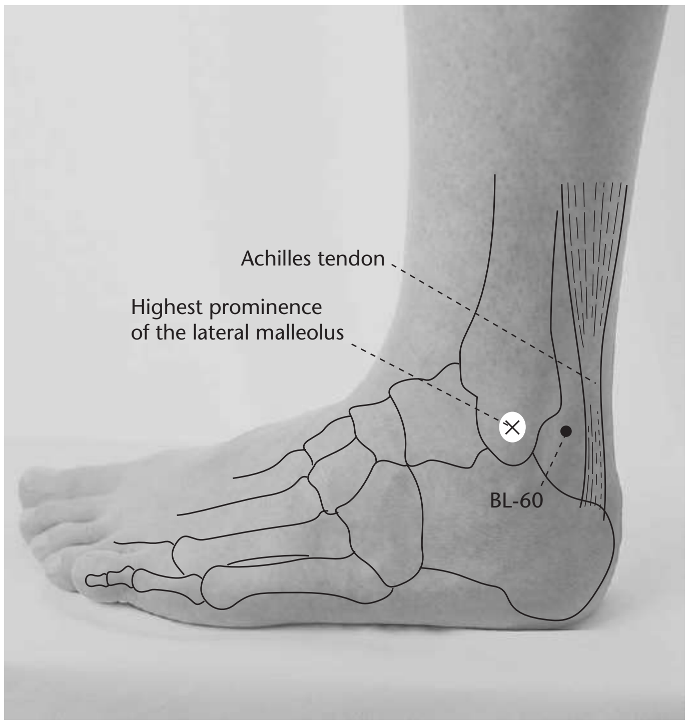
Achilles tendon
Highest prominence
of the lateral malleolus
BL-60

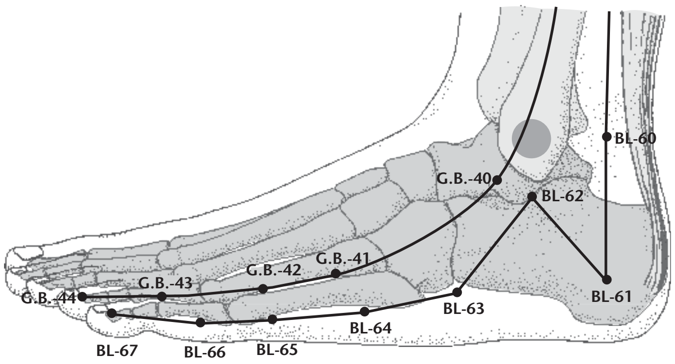
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

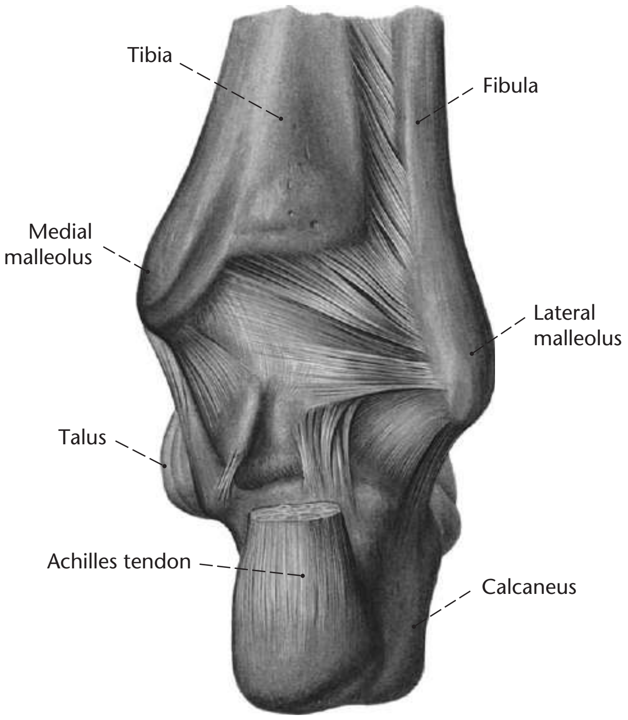
Tibia
Fibula
Medial
malleolus
Lateral
malleolus
Talus
Achilles tendon
Calcaneus

Location

In the depression on the line connecting the Achilles tendon and the highest prominence of the lateral malleolus.

How to find

Locate the highest prominence of the lateral malleolus (→ 3.6.2). From there, palpate horizontally towards the Achilles tendon and locate **BL-60** in a depression anterior to the tendon.

Needling

Vertically 0.5–1 cun. Caution: Reducing needle techniques contraindicated during pregnancy. Exception: A supportive point during labour. If the patient has cold feet, the therapeutic effect can be increased by using a heatlamp or a hot cherry-pit bag under the soles of the feet.

Actions/Indications

- Clears Heat, subdues Yang and Wind and eliminates excess, especially in the head
- Opens the channel, alleviates pain, relaxes the tendons, strengthens the lumbar region
- Promotes labour

Special features

Jing-river point, Fire point, Ma Dan Yang Heavenly Star point. Important distal point for the cervical, thoracic and lumbar spine, especially for chronic cases.

BL-61 *Servant's Respect* PUCAN

Location

On the lateral aspect of the heel, 1.5 cun inferior to → **BL-60** (in the depression between the highest prominence of the lateral malleolus and the Achilles tendon), in a depression on the calcaneus.

How to find

First, locate → **BL-60** in the depression between the highest prominence of the lateral malleolus and the Achilles tendon. With oedema of the lower leg, the depression is not visible but palpable. From → **BL-60**, measure 1.5 cun in a distal direction. There, locate **BL-61** in a depression on the calcaneus, approximately on the midpoint of a line connecting the sole of the foot and → **BL-60**.

Needling

Vertically or obliquely 0.3–0.5 cun

Actions/Indications

- Relaxes the tendons
- Opens the channel, alleviates pain

Special features

Meeting point with the *yang qiao mai*

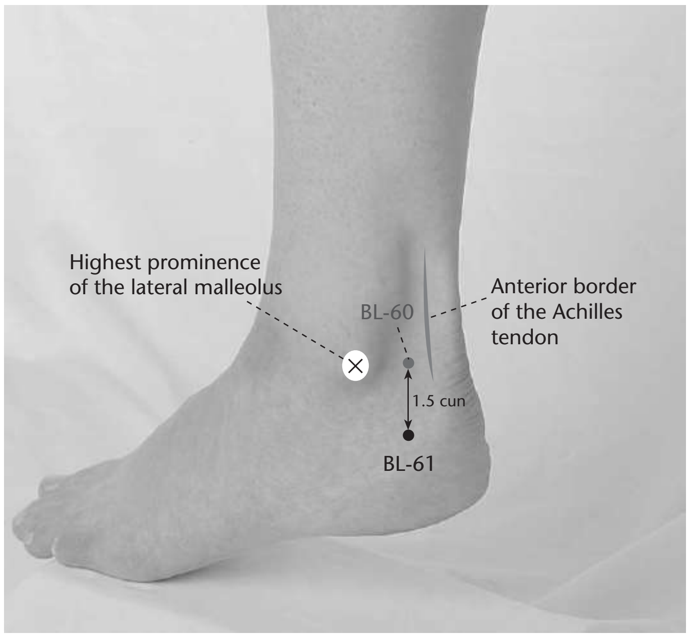
Highest prominence
of the lateral malleolus
BL-60
Anterior border
of the Achilles
tendon
1.5 cun
BL-61

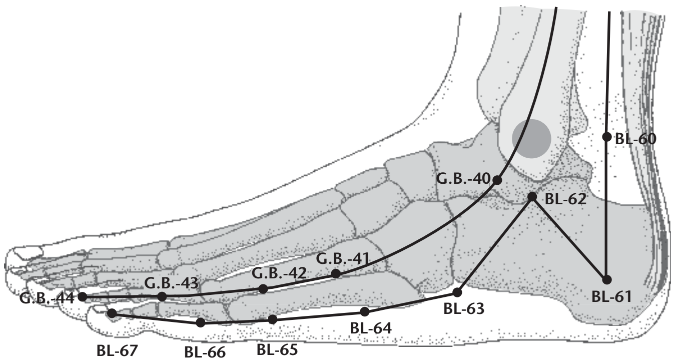
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

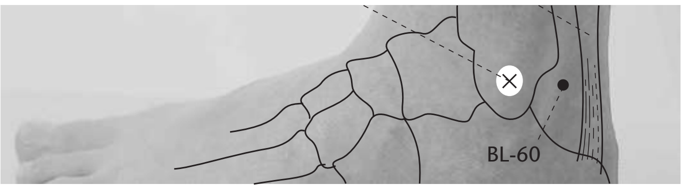

Extending Vessel SHENMAI **BL-62**

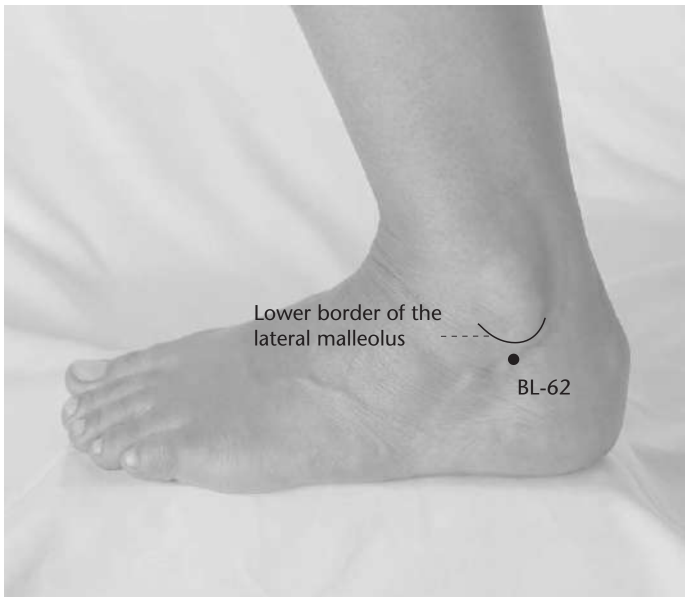
Lower border of the lateral malleolus
BL-62

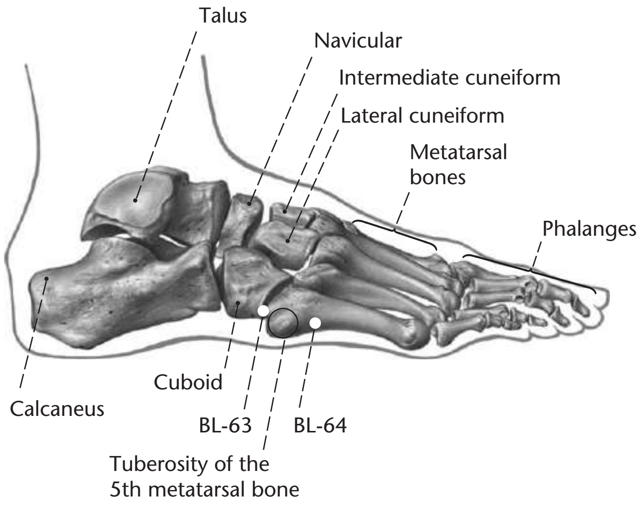
Talus
Navicular
Intermediate cuneiform
Lateral cuneiform
Metatarsal bones
Phalanges
Calcaneus
Cuboid
BL-63
BL-64
Tuberosity of the 5th metatarsal bone

Location

In a depression directly inferior to the highest prominence of the lateral malleolus, above the joint space between the talus and calcaneus.

How to find

First, locate the highest prominence of the lateral malleolus (→ 3.6.2). **BL-62** is located directly below it, in a depression inferior to the lower border of the malleolus and superior to the tendons of the peroneus muscle, in the joint space between the talus and calcaneus.

Needling

Vertically or obliquely 0.3–0.5 cun. The needle may reach the calcaneofibular ligaments superior to the tendons of the peroneus longus and brevis muscles, possibly also the joint space.

Actions/Indications

- Subdues (internal) Wind, clears Heat from the head, calms the *shen*, benefits the head and eyes
- Expels external Wind
- Opens and regulates the *yang qiao mai*
- Opens the channel, alleviates pain
- Moves the Qi in the Bladder sinew channel

Special features

Opening (master) point of the *yang qiao mai*, Sun Si Miao Ghost point.

BL-63 Golden Gate JINMEN

Location

On the lateral aspect of the foot, proximal to the tuberosity of the 5th metatarsal bone, in a depression anterior and inferior to → **BL-62** between the calcaneus and the cuboid bone. Note: Some authors locate **BL-63** between the cuboid bone and the tuberosity of the 5th metatarsal bone (see figure on **p. 311**); the more tender point should be selected.

How to find

When palpating along the lateral aspect of the foot on the level of the border of the 'red and white' skin, a bony structure can be felt approximately at the midpoint of the foot. This is the tuberosity of the 5th metatarsal point (→ 3.6.2). Slightly proximal to it (towards the heel), you can palpate a depression between the calcaneus and the cuboid bone. There, locate **BL-63**.

→ **BL-64** is located distal (towards the toes) to the tuberosity of the 5th metatarsal bone.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Opens the channel, alleviates pain and eliminates stagnation (*xi*-cleft point)
- Subdues internal Wind and calms the *shen*

Special features

Xi-cleft point, meeting point with the *yang wei mai*.

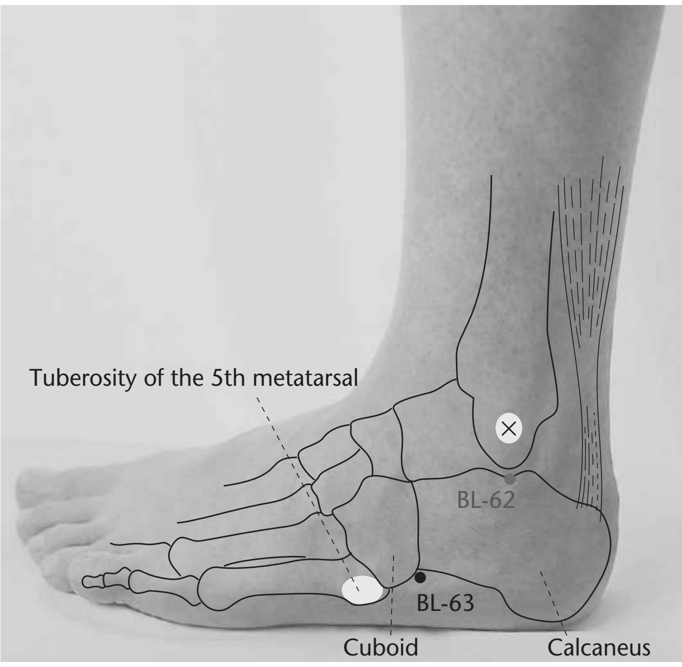
Tuberosity of the 5th metatarsal
BL-62
BL-63
Cuboid
Calcaneus

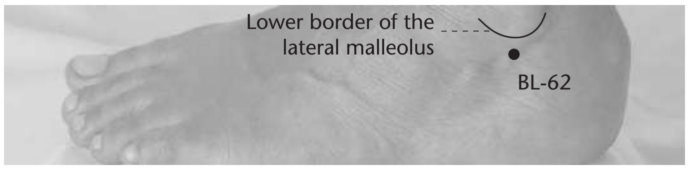
Lower border of the lateral malleolus
BL-62

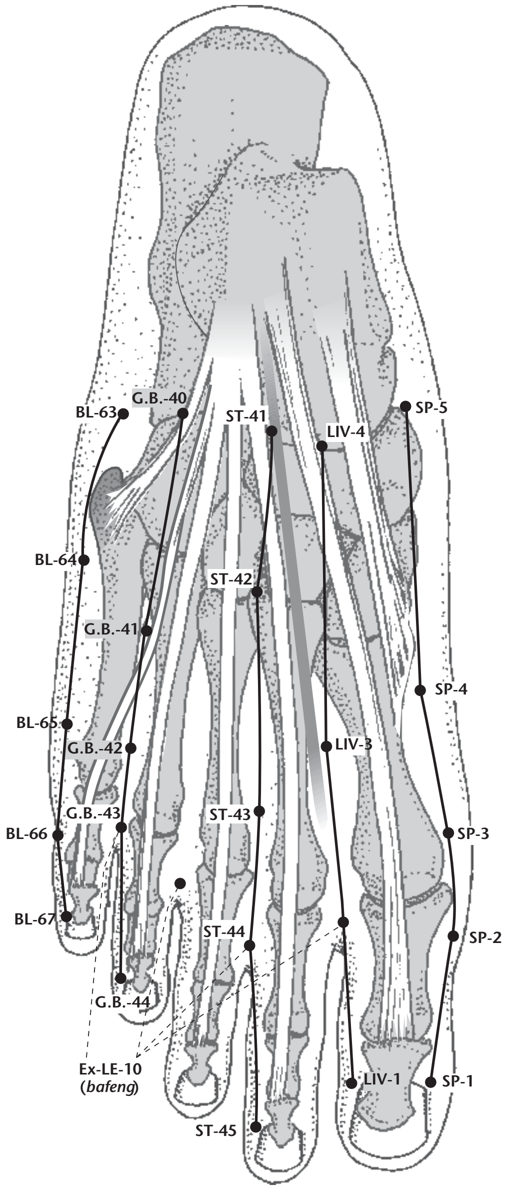
BL-63
G.B.-40
ST-41
SP-5
LIV-4
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

Capital Bone JINGGU BL-64

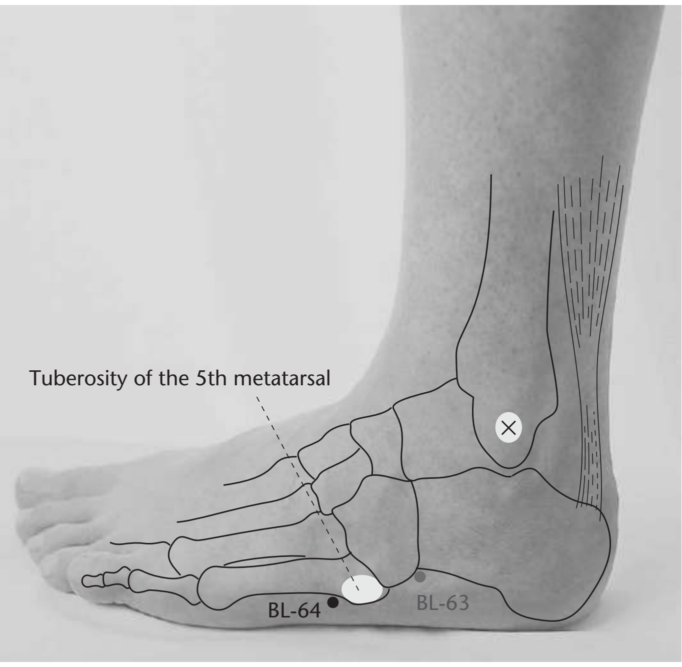
Tuberosity of the 5th metatarsal
BL-64
BL-63

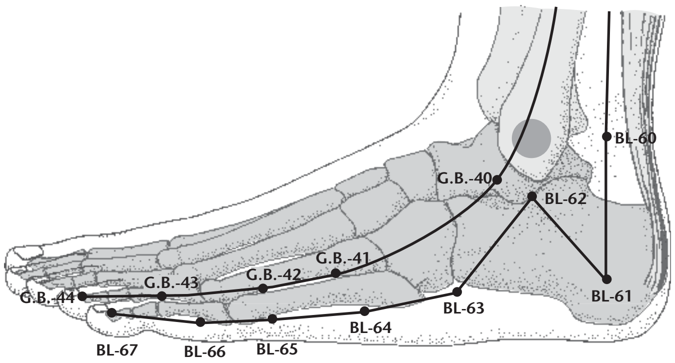
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

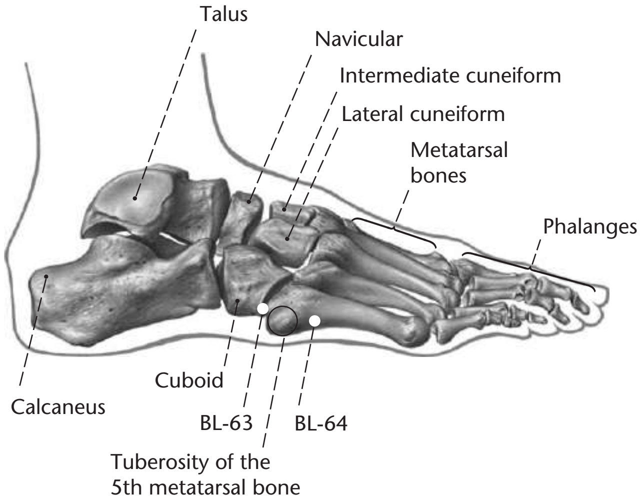
Talus
Navicular
Intermediate cuneiform
Lateral cuneiform
Metatarsal bones
Phalanges
Calcaneus
Cuboid
BL-63
BL-64
Tuberosity of the 5th metatarsal bone

Location

On the lateral aspect of the foot, distal to the tuberosity of the 5th metatarsal bone.

How to find

When palpating along the lateral aspect of the foot on the level of the border of the 'red and white' skin, a bony structure can be felt approximately at the midpoint of the foot. This is the tuberosity of the 5th metatarsal bone (➞ 3.6.2). **BL-64** is located directly distal to it (towards the toes), at the junction between the base and the shaft of the 5th metatarsal bone.

➞ **BL-63** is located proximal to the tuberosity of the 5th metatarsal bone, in a depression between the calcaneus and cuboid bone (or between the tuberosity and the cuboid bone). ➞ **SP-4** is located in a comparable position on the medial aspect of the foot, in a depression distal to the base of the 1st metatarsal bone.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Clears Wind and Heat from the head and eyes
- Calms the *shen*
- Opens the channel

Special features

Yuan-source point

BL-65 *Restraining Bone* SHUGU

Location

On the lateral aspect of the foot, in the depression proximal to the head of the 5th metatarsal bone.

How to find

When palpating along the bony structures on the lateral aspect of the foot, a prominent bony landmark can be felt approximately at the midpoint of the foot. This is the tuberosity of the 5th metatarsal bone. Distal to it (towards the toes), you can feel a further bony structure, the head of the 5th metatarsal bone. Directly proximal to it, at the border of the 'red and white' skin (border dorsum/sole of the foot), **BL-65** can be palpated in a depression. → **BL-66** is located in a depression distal to the head of the 5th metatarsal bone (distal to the metatarsophalangeal joint), at the junction of the base and the shaft of the proximal phalanx of the little toe.

→ **SP-3** is located in a comparable position on the medial aspect of the foot (proximal to the head of the 1st metatarsal bone). → **S.I.-3** and → **L.I.-3** are located in comparable positions at the ulnar/lateral border of the hand.

Needling

Vertically 0.3–0.5 cun. Caution: Painful point.

Actions/Indications

- Expels pathogenic factors (especially from the head)
- Clears Heat
- Opens the channel and *luo* vessels, alleviates pain
- Calms the *shen*

Special features

Shu-stream point, Wood point, sedation point.

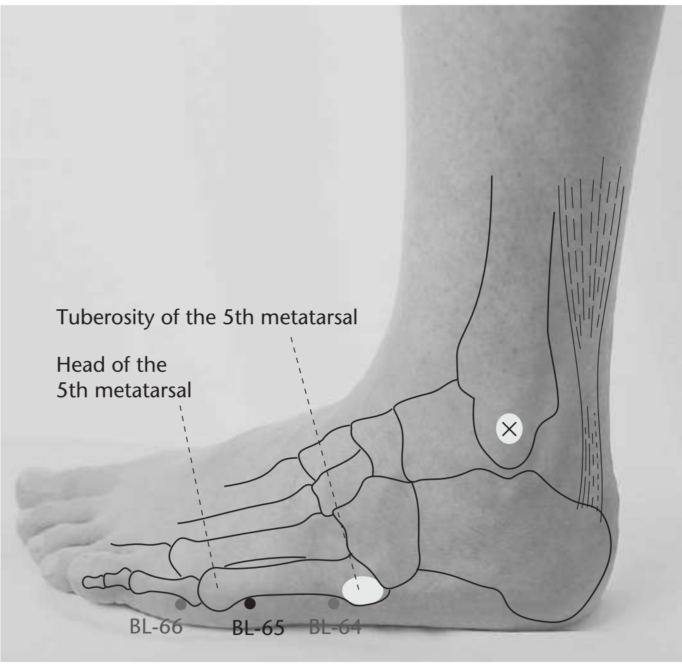
Tuberosity of the 5th metatarsal
Head of the
5th metatarsal
BL-66
BL-65
BL-64

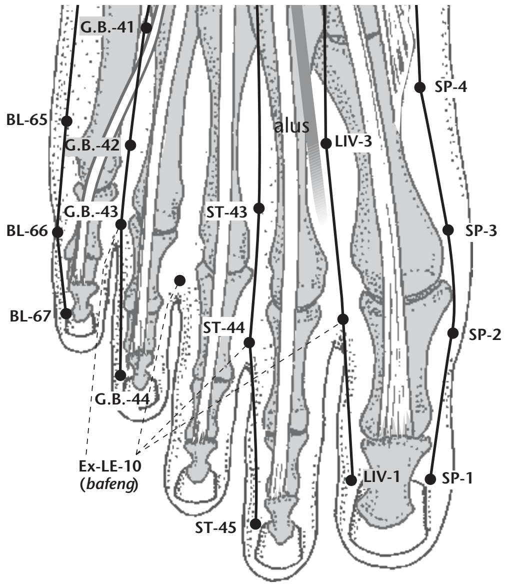
G.B.-41
BL-65
G.B.-42
G.B.-43
BL-66
BL-67
G.B.-44
Ex-LE-10
(bafeng)
ST-43
ST-44
ST-45
alus
LIV-3
LIV-1
SP-4
SP-3
SP-2
SP-1

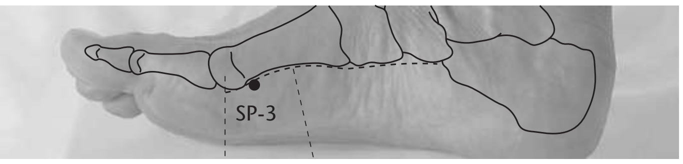

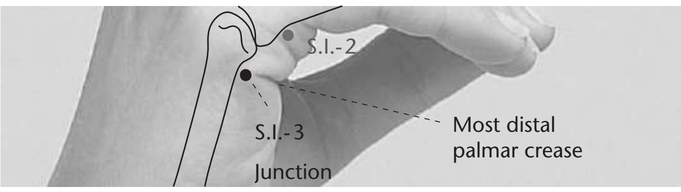
S.I.-2
S.I.-3
Junction
Most distal
palmar crease

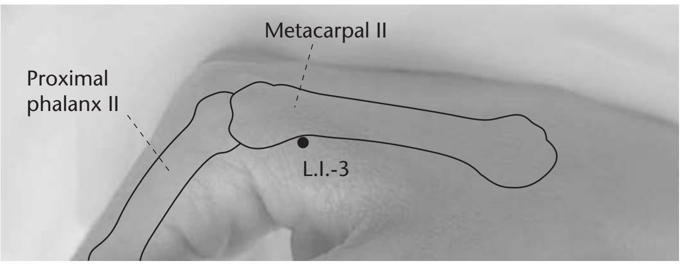
Metacarpal II
Proximal
phalanx II
L.I.-3

Foot Connecting Valley ZUTONGGU BL-66

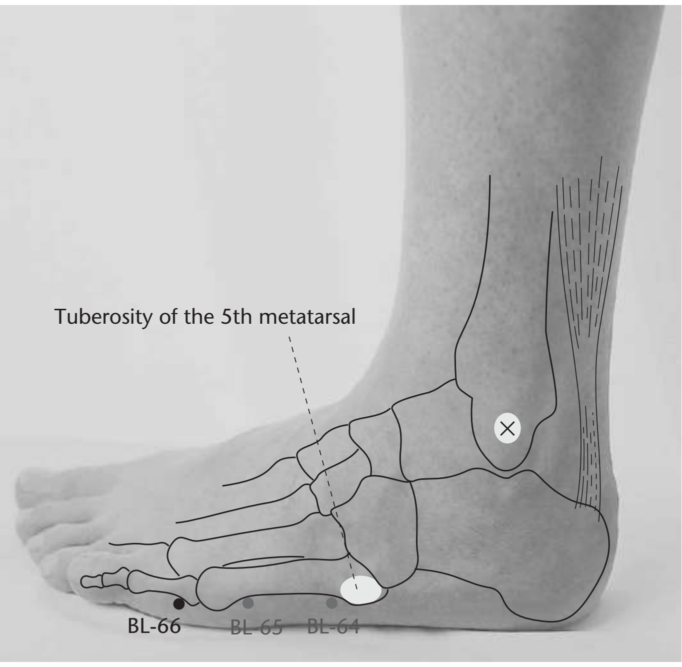
Tuberosity of the 5th metatarsal
BL-66
BL-65
BL-64

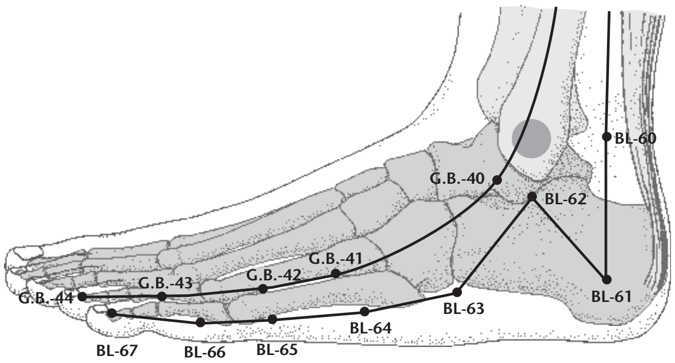
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

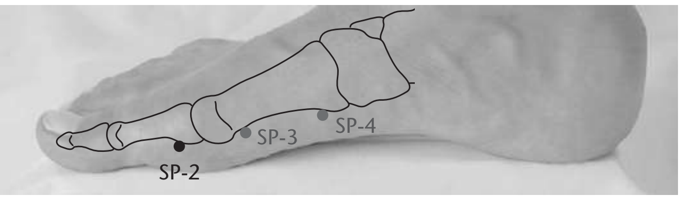

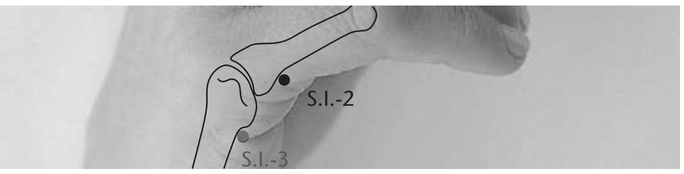

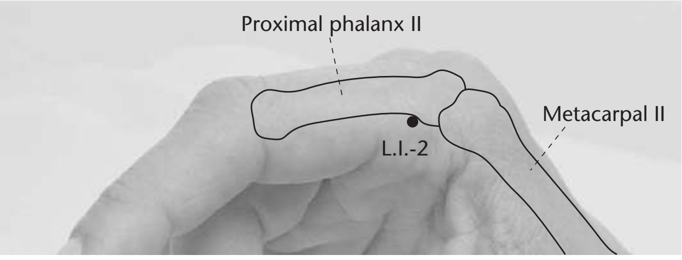
Proximal phalanx II
Metacarpal II
L.I.-2

Location

At the lateral border of the foot, in the depression distal to the metatarsophalangeal joint of the little toe.

How to find

When palpating along the border of the 'red and white' skin on the lateral aspect of the foot, at approximately the midpoint of the length of the foot, a clearly defined bony structure can be felt. This is the tuberosity of the 5th metatarsal bone. Distal to it (towards the toes), you can feel a further prominent bony structure, the head of the 5th metatarsal bone (or the 5th metatarsophalangeal joint). **BL-66** is located directly distal to it, in a depression at the junction of the base and the shaft of the proximal phalanx of the little toe.

→ **BL-65** is located proximal to the head of the 5th metatarsal bone. → **SP-2** is located in a comparable position on the medial aspect of the foot. Located in comparable positions on the hand are → **S.I.-2** (on the ulnar aspect) and → **L.I.-2** (on the radial aspect).

Needling

Vertically 0.3–0.5 cun. Caution: Painful point.

Actions/Indications

- Clears the head
- Descends the Lung and Stomach Qi
- Calms the *shen*

Special features

Ying-spring point, Water point, *ben* point (Five Phases point).

Ch04.7b-F10028.qxd 2/22/08 5:01 PM Page 314

BL-67 *Reaching Yin* ZHIYIN

Location

On the dorsal aspect of the little toe, 0.1 cun from the lateral corner of the nail.

How to find

This point is located at the junction of two tangents that border the nail of the little toe laterally and proximally, approximately 0.1 cun from the actual corner of the nail.

Needling

Vertically 0.1 cun or obliquely 0.2 cun in the direction of the foot. Prick to bleed. Caution: Painful point.

Actions/Indications

- Scientifically established: conversion of breech presentations prior to birth: moxibustion and electro-acupuncture at this point seem to be most effective (approximately 70–80%)
- Facilitates labour
- Expels pathogenic factors from the opposite end of the channel
- Regulates Yin and Yang of the Water element

Special features

Jing-Well point, Metal point, tonification point, exit point. An isolated case has been reported regarding a pregnant woman who was treated with moxibustion at **BL-67**. This resulted in a fetomaternal transfusion of 300 ml of blood, endangering the fetus. For this reason, treatments should be controlled by cardiotocography.

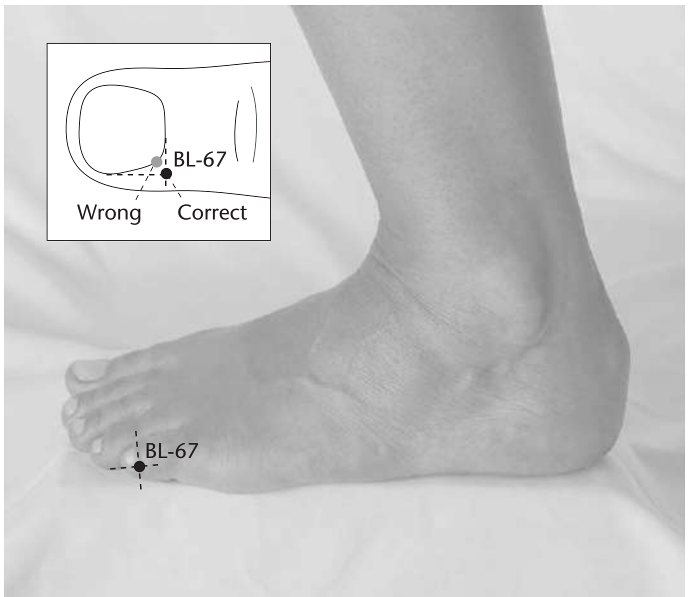
BL-67
Wrong
Correct
BL-67

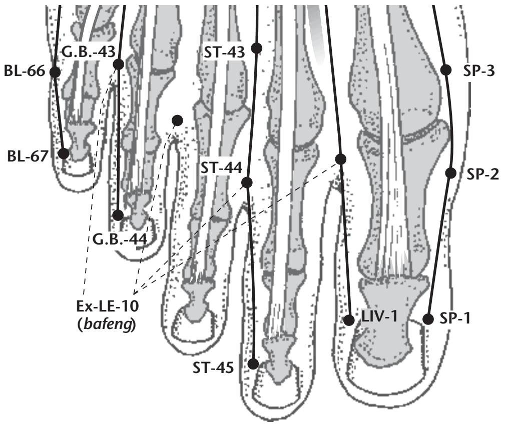
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

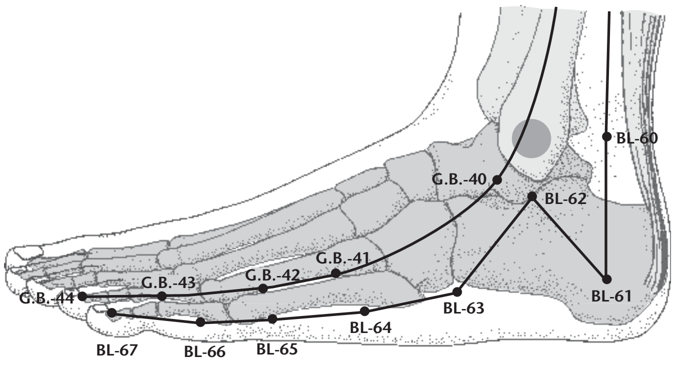
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

4.8 The Kidney Channel System – Foot *shaoyin* (*zu shao yin jing luo*)

4.8.1 The Kidney Primary Channel (*zu shao yin jing*)

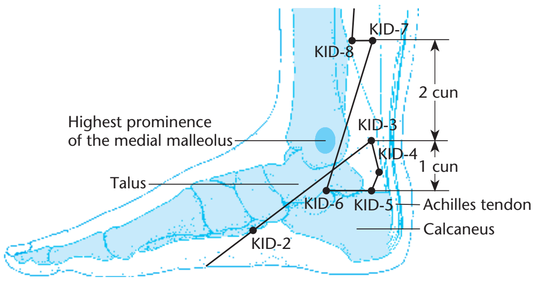
KID-7
KID-8
2 cun
Highest prominence of the medial malleolus
KID-3
KID-4
1 cun
Talus
KID-6
KID-5
Achilles tendon
Calcaneus
KID-2

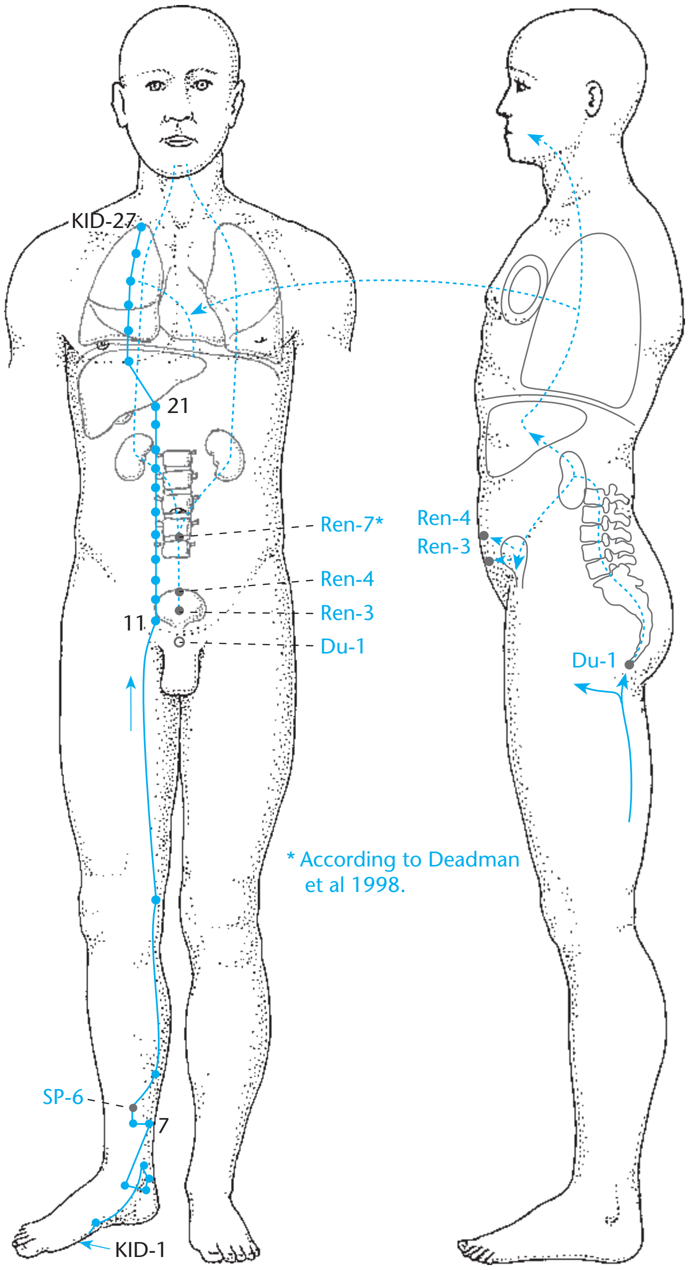
KID-27
21
Ren-7*
Ren-4
Ren-3
Du-1
11
Ren-4
Ren-3
Du-1
SP-6
7
KID-1

* According to Deadman et al 1998.

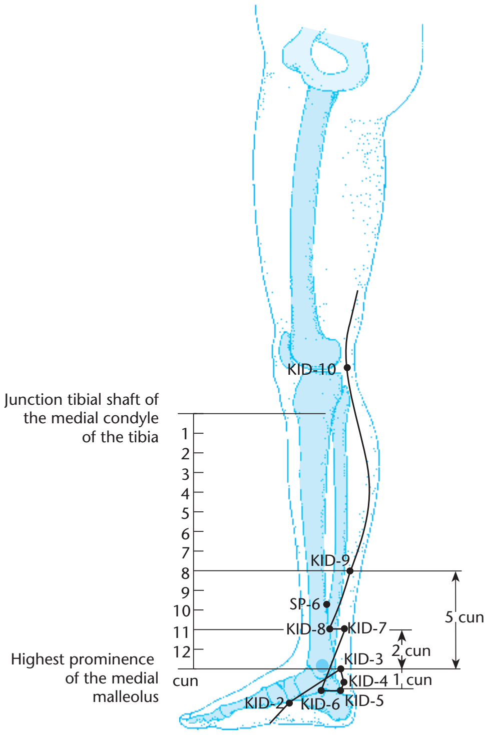
KID-10
Junction tibial shaft of the medial condyle of the tibia
1
2
3
4
5
6
7
8
9
10
11
12
KID-9
SP-6
KID-8
KID-7
5 cun
2 cun
KID-3
Highest prominence of the medial malleolus
cun
KID-4
1 cun
KID-2
KID-6
KID-5

Pathway

The Kidney primary channel begins beneath the little toe, which is reached by **a branch** separating from the Bladder primary channel at its endpoint at → **BL-67** (*zhiyin*) (foot Yin–Yang connection of the second great circuit).

The Kidney primary channel diagonally crosses the sole of the foot to **KID-1** (*yongquan*) and to **KID-2** (*rangu*) below the navicular tuberosity, continues to the posterior aspect of the medial malleolus and descends to enter the heel. From here it ascends the medial aspect of the lower leg, intersecting with → **SP-6** (*yinlingquan*) and continuing to ascend the posteromedial aspect of the thigh towards the perineum.

In the perineal region the primary channel divides into **two branches**:

- the **deep, internal** branch travels to → **Du-1** (*changqiang*) and ascends alongside the spine to connect with its pertaining *zang*-Organ, the Kidneys (*shen*), and its paired *fu*-Organ, the Bladder (*pangguang*). Here it divides into smaller branches which connect with → **Ren-4** (*guanyuan*) and → **Ren-3** (*zhongji*), and, according to some authors, also with → **Ren-7** (*yinjiao*). From the Kidney (*shen*) a branch ascends to the Liver (*gan*), runs to and spreads in the Lung (*fei*), continues to the trachea and terminates at the root of the tongue. From the Lung (*fei*), an **inner branch** travels to the Heart (*xin*), where it meets the Pericardium primary channel (deep Yin–Yin connection). It then disperses in the chest and reaches → **Ren-17** (*shanzhong*).
- the **external** branch separates in the perineal region and runs to the lower abdomen at **KID-11** (*henggu*). From the pubic

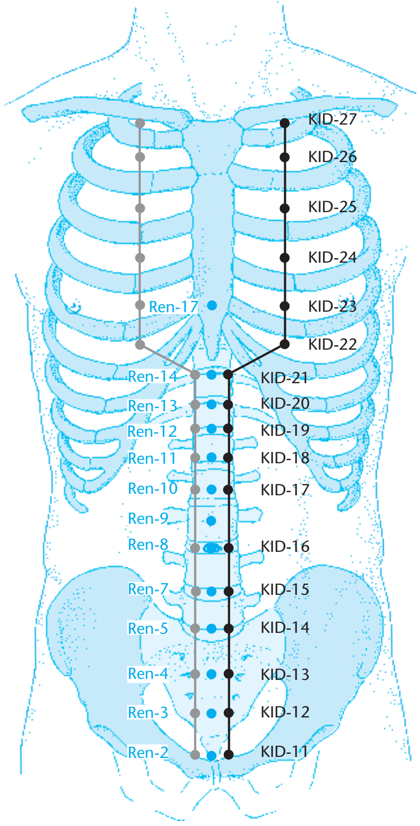

bone it ascends initially 0.5 cun lateral to the anterior midline, from the 5th intercostal space (**KID-22**) 2 cun lateral to the midline to the infraclavicular fossa.

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Pain in the lumbar region, counterflow Cold or weakness of the legs, dry mouth, sore throat, pain in the lateral gluteal region and on the posterior thigh, pain on the soles of the feet.
Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain, facial oedema, dark rings under the eyes, shortness of breath, somnolence, restlessness, diarrhoea, thin or dry stools, abdominal distension, nausea and vomiting, impotence.

Connections of the Kidney primary channel

Connections with other channels

Bladder primary channel (*zu tai yang jing*)

Connection: Foot Yin–Yang connection of the second great circuit
Location: BL-67 → KID-1 (on the foot)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Heart primary channel (*shou shao yin jing*)

Connection: Paired according to the Six-Channel theory (hand–foot pairing): *shao yin* (Yin axes of the second great circuit)
Location: KID → HE (on the thorax). A branch of the internal pathway of the Kidney primary channel travels from the Kidney to the Liver, penetrates the diaphragm and spreads in the Lung (*fei*). From the Lung a branch runs to the Heart (*xin*), there connecting with the Heart primary channel.
Circulation: Non-circadian (**not** according to the Organ clock)
Importance: Above–below relationship

Pericardium primary channel (*shou jue yin jing*)

Connection: Deep Yin–Yin connection
Location: KID → P. A branch of the internal pathway of the Kidney primary channel travels from the Kidney to the Liver, penetrates the diaphragm and spreads in the Lung (*fei*). From the Lung an internal branch courses to the Heart (*xin*), where it joins the Pericardium primary channel (deep Yin–Yin connection) and also reaches → **Ren-17** (*shanzhong*).
Circulation: Circadian (according to the Organ clock)
Importance: The Pericardium primary channel receives Nutritive Qi (*ying qi*) from the Kidney primary channel (first circulation of the *ying qi* → 1.1.4).

Connections with other *zangfu*-Organ systems

Kidney (*shen*), **Bladder (*pangguang*)**, Liver (*gan*), Lung (*fei*), Pericardium (*xin bao*), Heart (*xin*)

4.8.2 The Kidney Divergent Channel (*zu shao yin jing bie*)

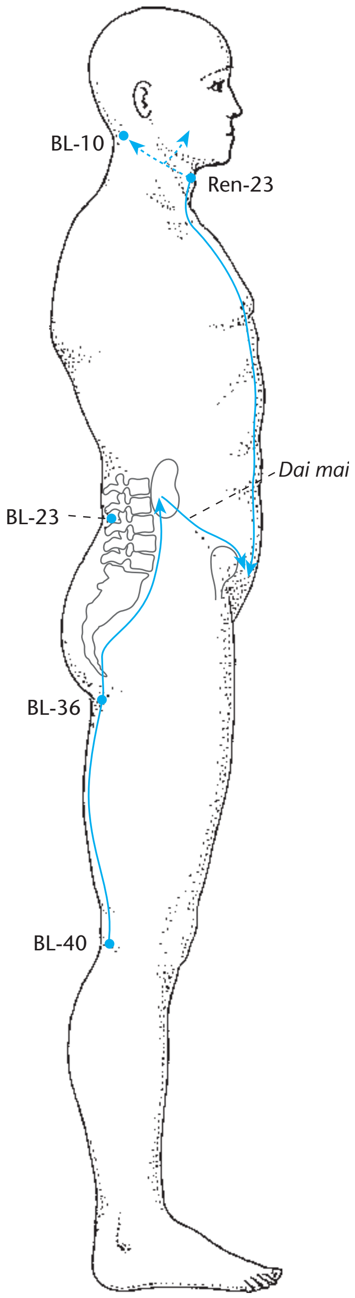

Pathway

The Kidney divergent channel separates from the Kidney primary channel in the popliteal fossa at → **KID-10** (*yinggu*). It

➡ meets → **BL-40** (*weizhong*) in the popliteal fossa

➡ ascends to → **BL-36** (*chengfu*) in the centre of the gluteal crease

➡ penetrates the anus and reaches the Kidney (*shen*) and the Bladder (*pangguang*)

➡ ascends to → **BL-23** (*shenshu*) at the level of the lower border of the spinous process of L2, where it also meets the *dai mai*, following it round the waist to the abdomen

➡ ascends lateral to the midline to the infraclavicular fossa

➡ reaches the neck at → **Ren-23** (*liangquan*)

➡ emerges at the nape of the neck at → **BL-10** (*tianzhu*) and converges with the Bladder primary channel and the Bladder divergent channel to form one of the 6 *he*-confluences (here: BL/KID as first confluence → 1.3).

Clinical importance

- Strengthens the relationship between the Kidneys and the Bladder (*zangfu*-Organ systems). Points on the Kidney primary channel can therefore be used for disorders of the Bladder, and vice versa points on the Bladder primary channel can treat disorders of the Kidneys.
- Through the connection with the *dai m*ai, KID points can be used for disorders of the *dai mai*.
- Kidney Qi and Essence (*jing*) are sent to the Brain through the Kidney divergent channel via the spinal canal, thus strengthening the connection between the Kidneys, the Marrow and the Brain.
- Supports the connection to the tongue.

4.8.3 The Kidney Sinew Channel (*zu shao yin jing jin*)

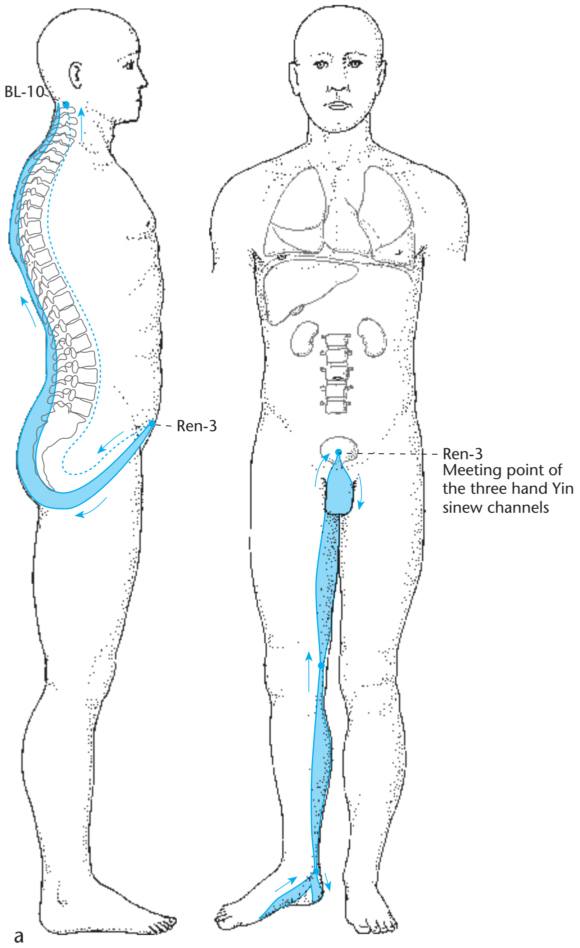

a

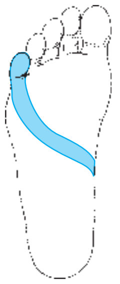

b

Pathway

The Kidney sinew channel originates beneath the little toe

- diagonally crosses the sole of the foot, meets the Spleen sinew channel and continues to the medial malleolus where it binds (*jie*). From here a **minor branch** travels to the medial portion of the calcaneus. The **main branch** ascends the posteromedial aspect of the leg, binds (*jie*) at the medial condyle of the femur, crosses the Bladder sinew channel, ascends the posteromedial aspect of the thigh to the genital region to → **Ren-3** (*zhongji*) and → **Ren-2** (→ 1.3), where it meets the other Yin sinew channels. It then penetrates the abdomen, runs along the spine to the nape of the neck and there meets the Bladder sinew channel.

From the genital region **an internal branch** traverses the gluteal region and follows the deep musculature along the spine to the occiput, where it meets the Bladder sinew channel.

Clinical importance

Pathology: Spasms on the sole of the foot, pain or spasms along the posteromedial aspect of the lower extremity, the back and the occiput. Chronic *bi*-syndromes. Epileptic seizures, limited range of motion with spinal flexion and extension. Degenerative disorders of the bones and joints, spondylosis, osteoarthritis.

Indication: Mainly for chronic pain, spasms and Cold sensations of the lower back and spine, accompanied by limited range of motion.

4.8.4 The Kidney *luo*-Connecting Vessel System (*zu shao yin luo mai*)

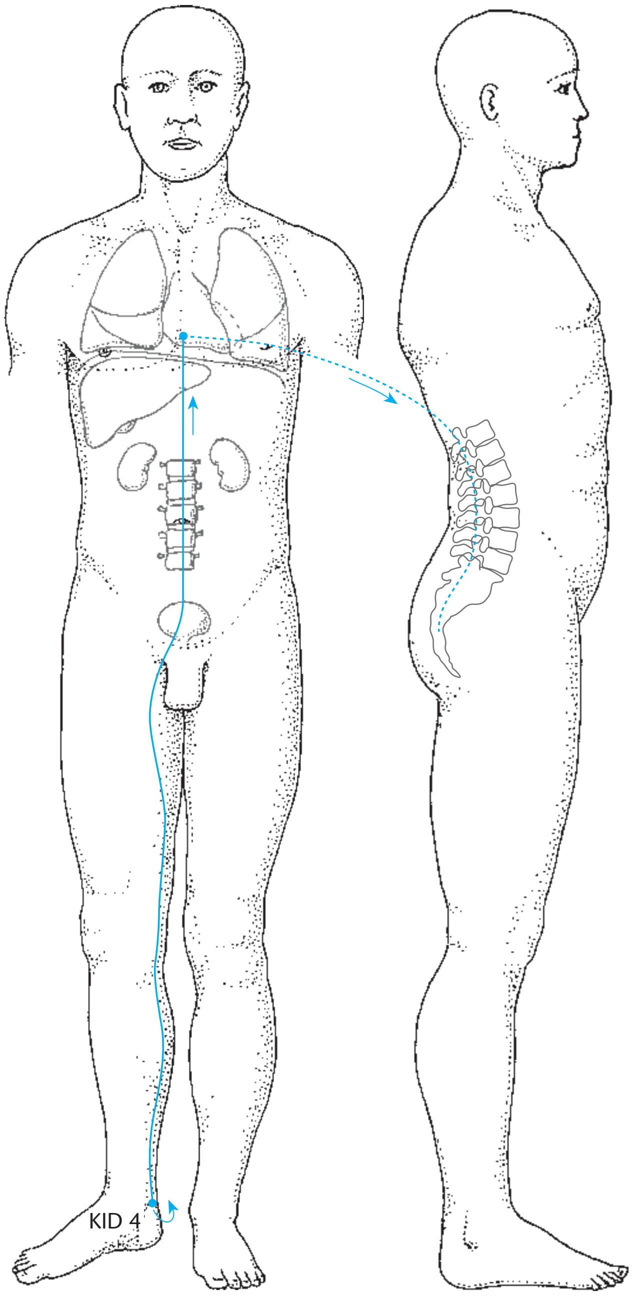

Pathway

The Kidney *luo*-connecting channel separates from the Kidney primary channel at **KID-4** (*dazhong*). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

➡ Horizontal divisions run to the Interiorly–Exteriorly paired Bladder primary channel; according to some schools of thought (for example Ngyen Van Nghi → Appendix) they travel as a **transverse** KID *luo*-connecting vessel to the *yuan*-source point **BL-64** (*jinggu*).

➡ A **longitudinal** division follows the Kidney primary channel to → **Ren-17** (*shanzhong*), penetrates the thorax and reaches the spine in the lumbar region.

Clinical importance (→ 8.1.2)

Pathology
Excess (*shi*): Retention of urine
Deficiency (*xu*): Pain in the lumbar region
Counterflow Qi: Restlessness, anxiety, fear, depression, fullness in the chest and epigastrium.

4.8.5 Cutaneous Region (*shao yin pi bu*)

See description and figures → 1.6.

4.8.6 Points of the Kidney Primary Channel (Overview)

Specific points according to their functions

- ***Yuan*-source point (→ 8.1.1): KID-3** (*taixi*) 🟦🟦
- ***Luo*-connecting point (→ 8.1.2.): KID-4** (*dazhong*)
- ***Xi*-cleft point (→ 8.1.3): KID-5** (*shuiquan*)
- **Associated Back-*shu* point (→ 8.1.4): BL-23** (*shenshu*) 🟦🟦
- **Associated Front-*mu* point (→ 8.1.5): G.B.-25** (*jingmen*) ■■
- **Five *shu*-transporting point (→ 8.1.6):**
  jing-well point (Wood), sedation point: **KID-1** (*yongquan*)
  ying-spring point (Fire): **KID-2** (*rangu*)
  shu-stream point (Earth): **KID-3** (*taixi*) 🟦■
  jing-river point (Metal), tonification point: **KID-7** (*fuliu*) 🟦■
  he-sea point (Water), *ben* point (Five Phases): **KID-10** (*yingu*) 🟦■
- *Hui*-meeting point (→ 8.1.7): –
- Opening point (→ 8.1.8) of the *yin qiao mai*: **KID-6** (*zhaohai*) 🟦■
- Lower *he*-sea points (→ 8.1.9): –
- *Jiaohui*-meeting points (→ 8.1.10):
  - with the *yin qiao mai*: **KID-2** (*rangu*)*, (**KID-6** (*zhaohai*)*, **KID-8** (*jiaoxin*)*)
  - with the *yin wei mai*: **KID-9** (*zhubin*)
  - with the *chong mai*: **KID-11–KID-21**
  - of other channels with the Kidney channel: **SP-6**, **Du-1**, **Ren-4**, **Ren-3**, **Ren-7***, **Ren-17**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): –**
- **Points of the Four Seas (→ 8.1.13): –**

---

* Mentioned by only some authors

- **Ma Dan Yang Heavenly Star point (→ 8.1.14): –**
- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points:**
  - *xi*-cleft point of the *yin qiao mai*: **KID-8** (*jiaoxin*)
  - *xi*-cleft point of the *yin wei mai*: **KID-9** (*zhubin*)

Points according to region

- **Local points** (→ 8.2.1): foot – **KID-6** (*zhaohai*) 🟦🟦; knee – **KID-10** (*yingu*) 🟦⬛
- **Adjacent points** (→ 8.2.1): foot – **KID-7** (*fuliu*) 🟦⬛; toes – **KID-6** (*zhaohai*) 🟦🟦
- **Distal points** (→ 8.2.1): for the throat – **KID-6** (*zhaohai*) 🟦🟦; for the tongue – **KID-6** (*zhaohai*) 🟦🟦; for the Kidneys and the Bladder – **KID-3** (*taixi*) 🟦⬛, **KID-7** (*fuliu*) 🟦🟦; for the genitourinary region – **KID-3** (*taixi*) 🟦🟦

Specific points according to the channel pathway (in numerical order)

- **KID-1** (*yongquan*) 🟦⬛: *jing*-well point (Wood); sedation point
- **KID-2** (*rangu*): *ying*-spring point (Fire); *jiaohui*-meeting point with the *yin qiao mai* (→ 8.1.10)
- **KID-3** (*taixi*) 🟦🟦: *shu*-stream point (Earth); distal point for the Kidneys, Bladder, urogenital region (→ 8.2.1); *yuan*-source point (→ 8.1.1)
- **KID-4** (*dazhong*): *luo*-connecting point (→ 8.1.2)
- **KID-5** (*shuiquan*): *xi*-cleft point (→ 8.1.3)
- **KID-6** (*zhaohai*) 🟦🟦: opening point (→ 8.1.8) of the *yin qiao mai*; distal point for the tongue and throat (→ 8.2.1); *jiaohui*-meeting point with the *yin qiao mai** (→ 8.1.10); local point for the foot (→ 8.2.1)
- **KID-7** (*fuliu*) 🟦⬛: distal point for the Kidneys and the Bladder (→ 8.2.1); *jing*-river point (Metal); tonification point; regional point for the foot (→ 8.2.1)
- **KID-8** (*jiaoxin*): *xi*-cleft point of the *yin qiao mai*; *jiaohui*-meeting point with the *yin qiao mai** (→ 8.1.10)
- **KID-9** (*zhubin*): *xi*-cleft point of the *yin wei mai*; *jiaohui*-meeting point with the *yin wei mai** (→ 8.1.10)
- **KID-10** (*yingu*) 🟦⬛: *he*-sea point (Water), *ben* point (Five Phases); local point for the knee (→ 8.2.1)
- **KID-11** to **KID-21**: *jiaohui*-meeting points with the *chong mai* (→ 8.1.10)

General location help

- **KID-11** to **KID-21** are all located 0.5 cun lateral to the midline
  - **KID-11** to **KID-15** are spread on the 5 cun distance (→ 1.2) between the upper border of the pubic symphysis and the umbilicus (they are spaced at 1 cun intervals from the upper border of the pubic symphysis to 4 cun above it)
  - **KID-16** to **KID-21** are spread on the 8 cun distance (→ 1.2) between the umbilicus and the sternocostal angle (they are spaced at 1 cun intervals from the level of the umbilicus to 6 cun superior to it).
- **KID-22** to **KID-27** are all located 2 cun lateral to the midline
  - **KID-22** to **KID-26** are located in the 5th to 1st intercostal spaces
  - **KID-27** is located on the lower border of the sternoclavicular joint

---

* Mentioned by only some authors

Gushing Spring YONGQUAN **KID-1**

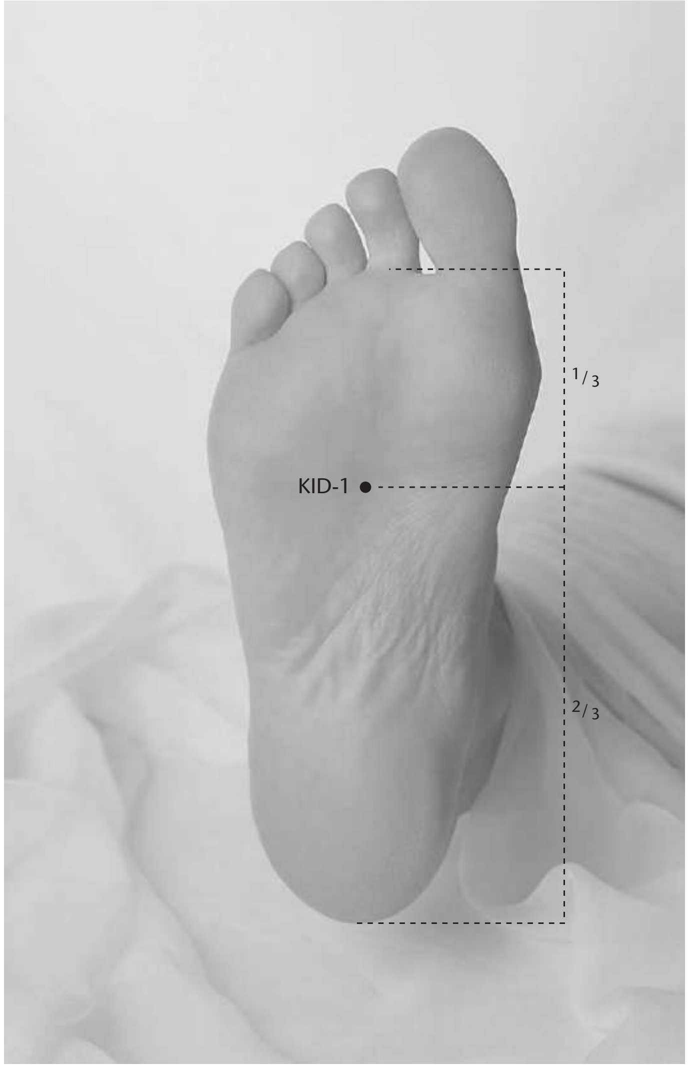

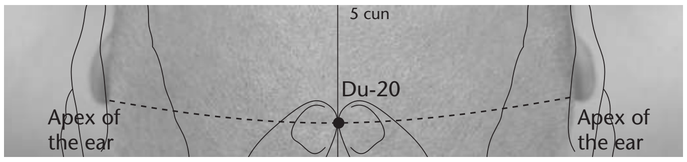

Location

On the sole of the foot, in a depression between the 2nd and 3rd metatarsal bones, at the junction of the anterior third and the posterior two-thirds of the sole.

How to find

Measure one third from the anterior border of the sole of the foot. Palpate for a pressure-sensitive point dorsal to the balls of the foot, at the midpoint of the width of the foot.

Needling

Needling this point is very painful and is therefore only recommended for severe disorders or in an emergency. Acupressure is commonly used instead; moxibustion is possible. For the treatment of collapse, unconsciousness, shock or a severe excess condition, this point should be strongly stimulated; in pronounced deficiency conditions, stimulation should be applied more cautiously.

Actions/Indications

- Revives collapsed Yang
- Clears excess, Heat and Wind from the head, lowers Yang
- Calms the *shen*

Special features

Jing-well point, Wood point, sedation point, entry point. **KID-1** is the lowest acupuncture point on the body and the only point on the sole of the foot. For this reason, it plays an important role in Qigong, forming a grounding point in the centre of the foot, where the energies of the Earth and Man unite. During practice, the centre of gravity should be above this point.

KID-2 *Blazing Valley* RANGU

Location

At the medial border of the foot, in a depression at the anterior border of the navicular bone, at the border of the 'red and white' skin.

How to find

Palpate from distal to proximal along the tarsal section of the medial aspect of the foot, past the shaft and the head of the 1st metatarsal bone (→ **SP-4**) and the medial cuneiform bone, finally reaching the prominent navicular bone. Locate **KID-2** anterior to the navicular bone, at the inferior angle of the joint between the medial cuneiform bone and the navicular bone.

Needling

From the medial aspect of the foot, 0.5–1 cun vertically below the border of the bone. Use reducing needling techniques for empty Heat. Use tonifying techniques for Kidney Yang/Qi deficiency.

Actions/Indications

- Clears empty Heat
- Regulates the Lower Burner and the Kidneys
- Local point

Special features

Ying-spring point, Fire point, meeting point with the *yin qiao mai*.

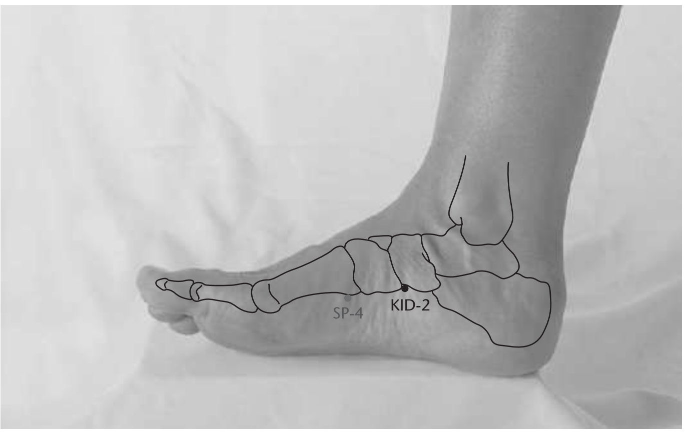

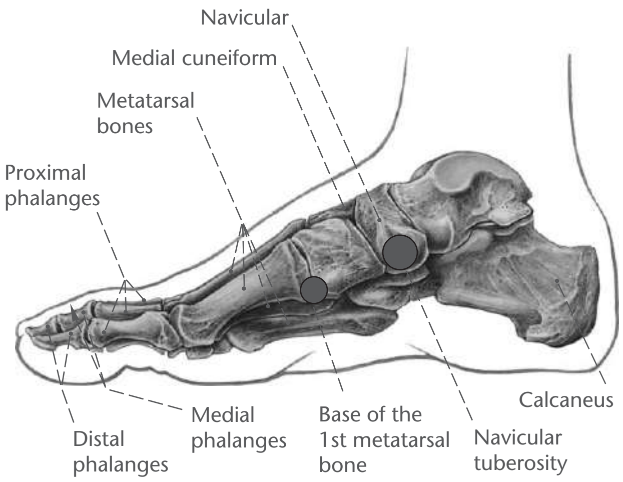
Navicular
Medial cuneiform
Metatarsal bones
Proximal phalanges
Calcaneus
Distal phalanges
Medial phalanges
Base of the 1st metatarsal bone
Navicular tuberosity

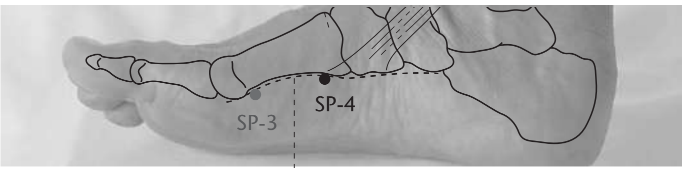

Supreme Stream TAIXI KID-3

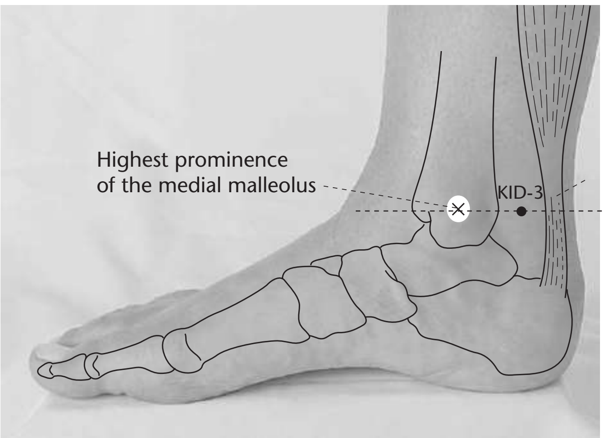
Highest prominence
of the medial malleolus
KID-3

Location

In the depression between the highest prominence of the medial malleolus and the Achilles tendon.

How to find

Locate the highest prominence of the medial malleolus (→ 3.6.2). From there, palpate horizontally towards the Achilles tendon. **KID-3** is located in a depression that can be palpated anterior to the tendon.

Needling

Vertically 0.3–1 cun

Actions/Indications

- Nourishes Kidney Yin and clears deficiency Heat, tonifies Kidney Yang, stabilises the Kidney Qi and the Lungs (helping the Kidneys absorb the Qi), regulates menstruation
- Strengthens the lower back, local point

Special features

Yuan-source point, *shu*-stream point, Earth point. Important point for tonifying the Kidneys (especially Kidney Yin).

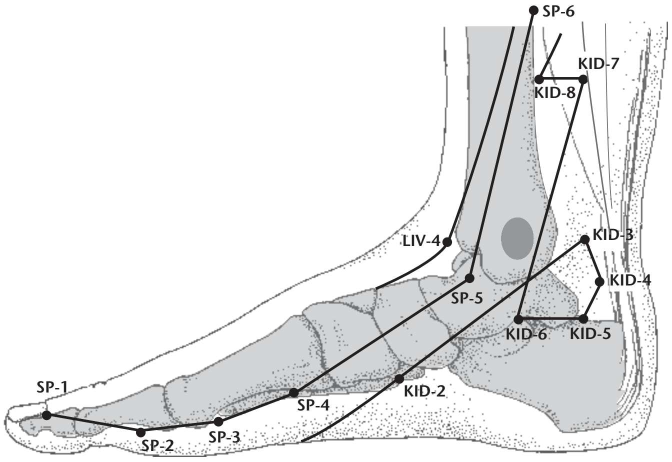
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

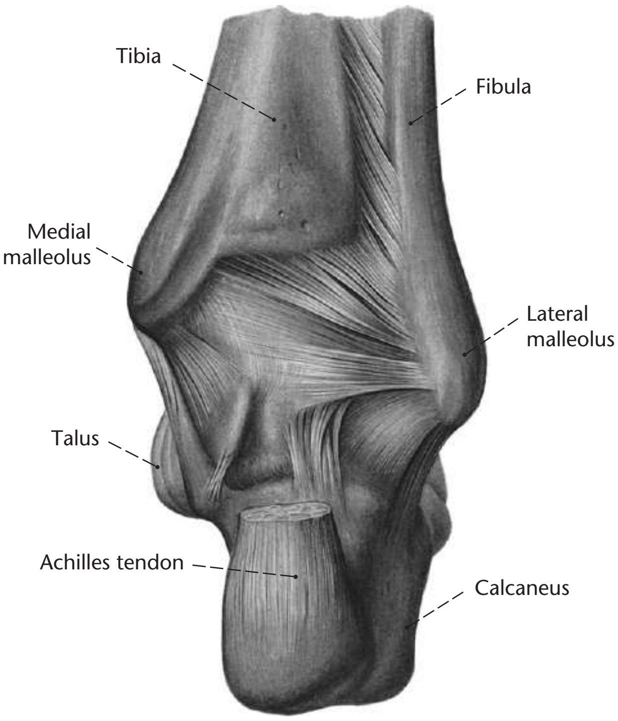
Tibia
Fibula
Medial
malleolus
Lateral
malleolus
Talus
Achilles tendon
Calcaneus

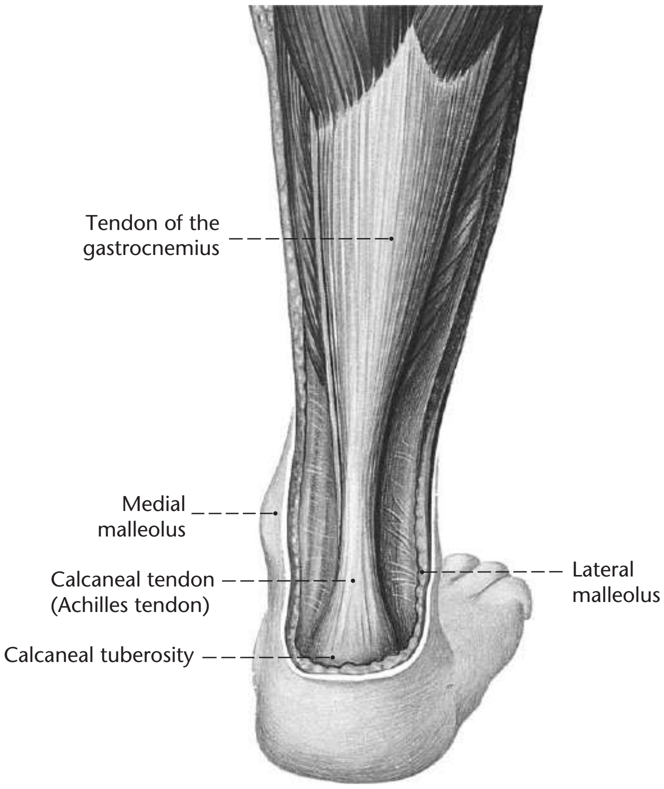
Tendon of the
gastrocnemius
Medial
malleolus
Calcaneal tendon
(Achilles tendon)
Lateral
malleolus
Calcaneal tuberosity

KID-4 *Great Bell* DAZHONG

Location

Anterior to the medial border of the Achilles tendon, superior to its insertion at the calcaneus.

How to find

From the highest prominence of the medial malleolus, draw a horizontal line to the medial border of the Achilles tendon. From there, measure 0.5 cun in a distal direction. **KID-4** is located in a depression anterior to the Achilles tendon, slightly superior to its insertion at the calcaneus. **Or: KID-4** is located posterior to the midpoint of a line connecting → **KID-3** and → **KID-5** anterior to the Achilles tendon.

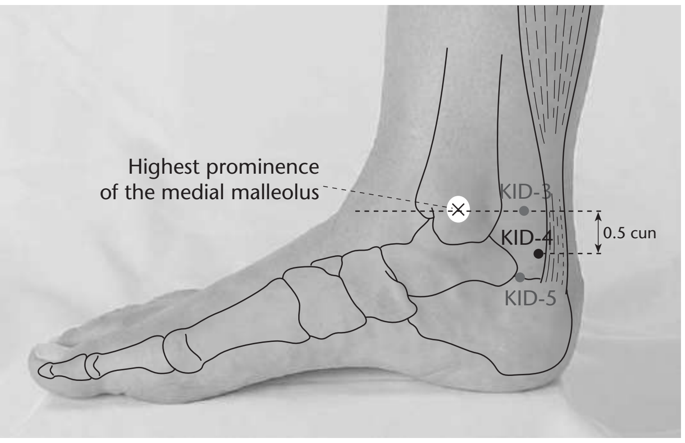
Highest prominence
of the medial malleolus
KID-3
KID-4
KID-5
0.5 cun

Needling

Vertically 0.3–0.5 cun. Avoid puncturing the tendon.

Actions/Indications

- Firms the Kidney Qi (helping the Kidneys absorb the Qi) and supports the Lungs
- Strengthens the Kidneys and cools empty Heat
- Strengthens the will, dispels fear
- Local point

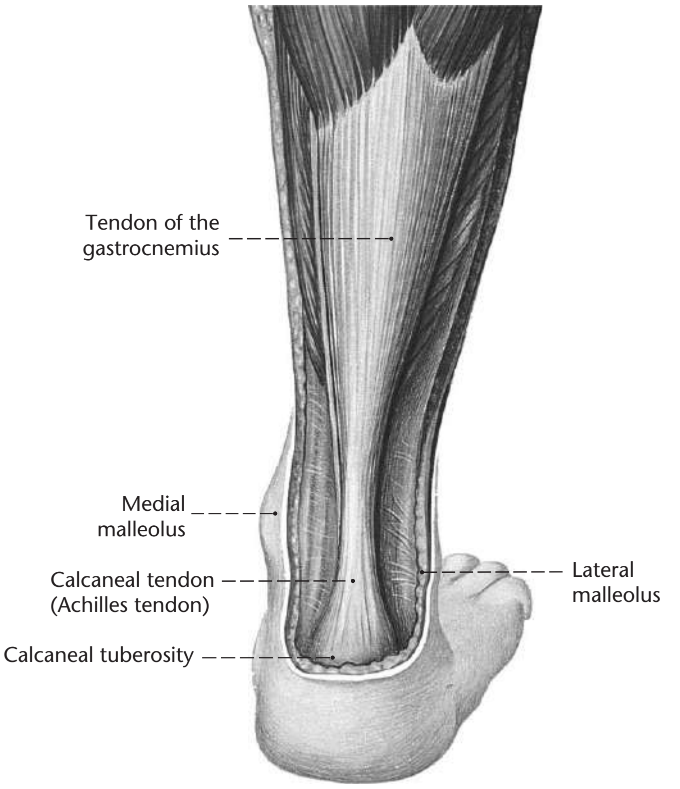
Tendon of the
gastrocnemius
Medial
malleolus
Calcaneal tendon
(Achilles tendon)
Calcaneal tuberosity
Lateral
malleolus

Special features

Luo-connecting point

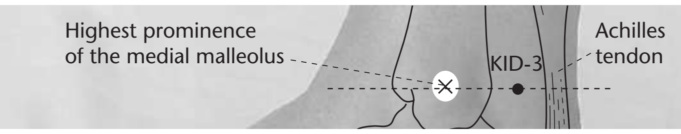
Highest prominence
of the medial malleolus
KID-3
Achilles
tendon

Ch04.8-F10028.qxd 2/22/08 6:14 PM Page 325

Water Spring SHUIQUAN KID-5

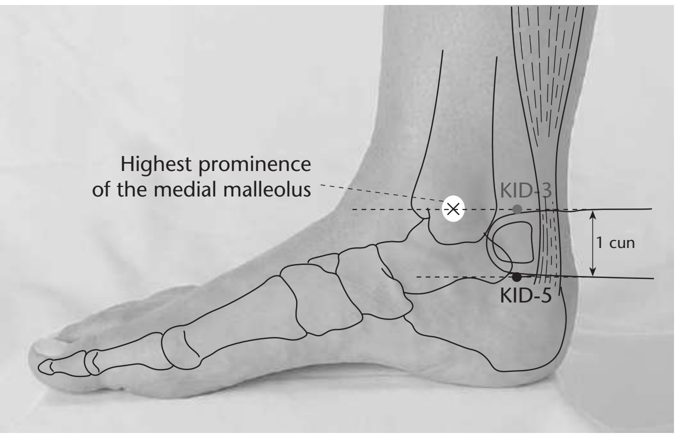
Highest prominence
of the medial malleolus
KID-3
1 cun
KID-5

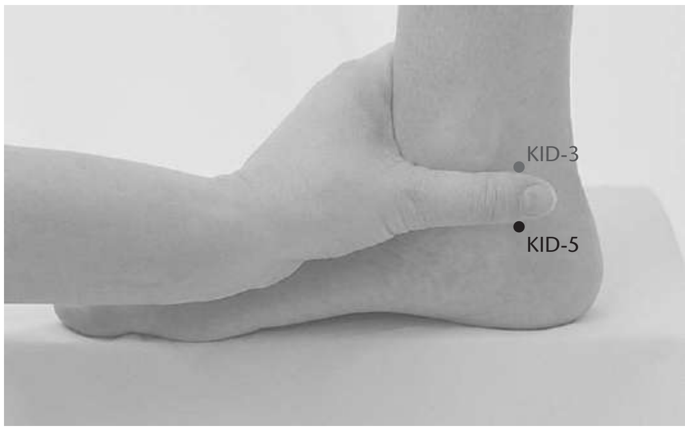

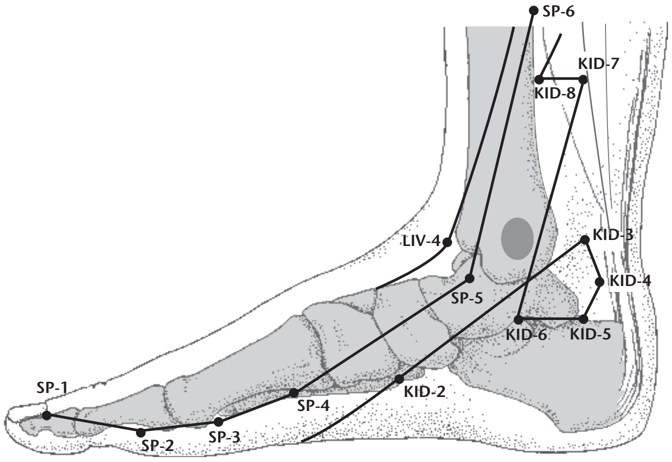
SP-6
KID-7
KID-8
LIV-4
KID-3
SP-5
KID-4
KID-6
KID-5
SP-1
KID-2
SP-4
SP-2
SP-3

Location

1 cun distal to ➞ **KID-3**, in a depression over the joint space between the talus and calcaneus.

How to find

First, locate ➞ **KID-3** on the level of the highest prominence of the medial malleolus (➞ 3.6.2), in a depression between the malleolus and the Achilles tendon. From **KID-3**, palpate 1 cun in an inferior direction to a palpable depression over the joint space between the talus and calcaneus. This is the location of **KID-5**.

Needling

0.3–0.5 cun obliquely to vertically superior to the margin of the bone.

Actions/Indications

- Regulates the *chong mai* and *ren mai*, benefits menstruation and urination

Special features

Xi-cleft point. Especially indicated for acute disorders and painful conditions.

KID-6 *Shining Sea* ZHAOHAI

Location

Approximately 1 cun inferior to the highest prominence of the medial malleolus, over the joint space between the talus and the calcaneus.

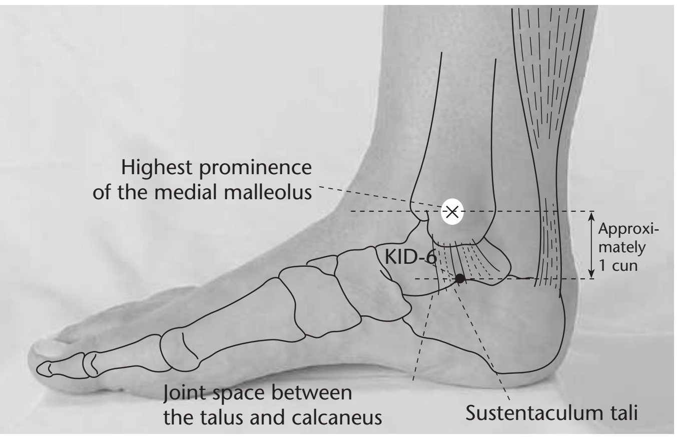
Highest prominence
of the medial malleolus
KID-6
Approxi-
mately
1 cun
Joint space between
the talus and calcaneus
Sustentaculum tali

How to find

From the highest prominence of the medial malleolus (→ 3.6.2), palpate distally, until the palpating finger can feel the depression of the joint space between the talus and the calcaneus. (**KID-6** is often located between the tendons of the tibialis and flexor digitorum longus muscles.) Directly distal to this point, you can palpate a small bony prominence, the sustentaculum tali. With the patient's foot in a supinated position, a visible skin crease forms on the level of the point. Note: There are several variations regarding the exact location of this point. However, the determining aspect for its location is not the cun measurement but the joint space below the prominence.

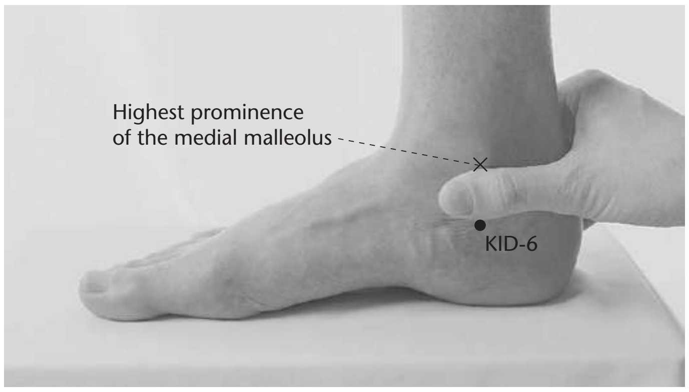
Highest prominence
of the medial malleolus
KID-6

Needling

0.3–0.5 cun vertically or obliquely in a proximal direction. The needle may reach the deltoid ligament.

Actions/Indications

- Nourishes Kidney Yin, clears empty Heat, benefits the throat, regulates the Lower Burner
- Calms the *shen*
- Moves Qi locally

Special features

Opening (master) point of the *yin qiao mai*. Important point, especially for strengthening the Kidney Yin.

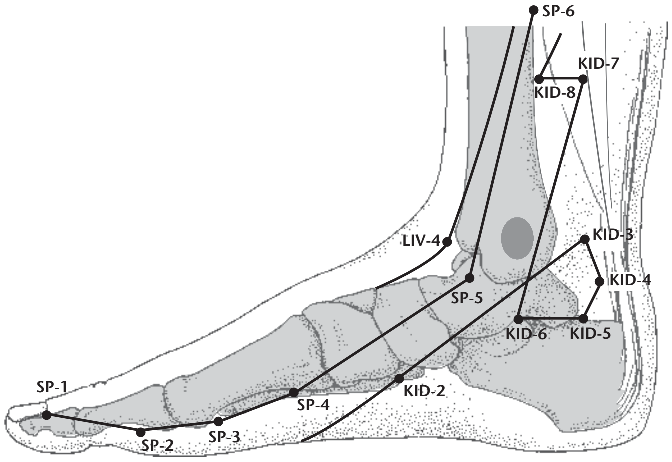
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

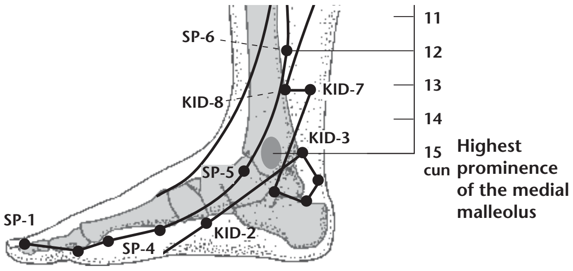
SP-6
KID-8
KID-7
KID-3
SP-5
SP-1
KID-2
SP-4
11
12
13
14
15
cun
Highest
prominence
of the medial
malleolus

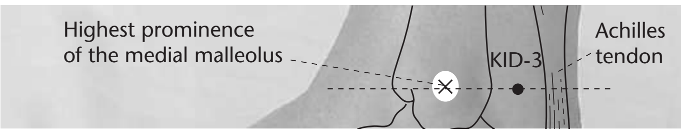
Highest prominence
of the medial malleolus
KID-3
Achilles
tendon

Returning Current FULIU **KID-7**

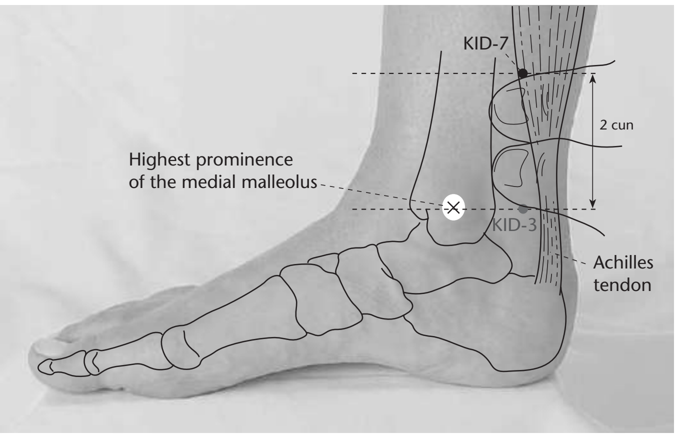
KID-7
2 cun
Highest prominence of the medial malleolus
KID-3
Achilles tendon

Location

2 cun proximal to → **KID-3**, at the anterior border of the Achilles tendon.

How to find

First, locate → **KID-3** on the level of the higest prominence of the medial malleolus (→ 3.6.2), in the depression between the malleolus and the Achilles tendon. From → **KID-3**, measure 2 cun in a proximal direction (towards the knee joint) and locate **KID-7** in a depression at the anterior border of the Achilles tendon.

Located on the same level, but more medially, is → **KID-8** (2 cun directly proximal to the medial malleolus).

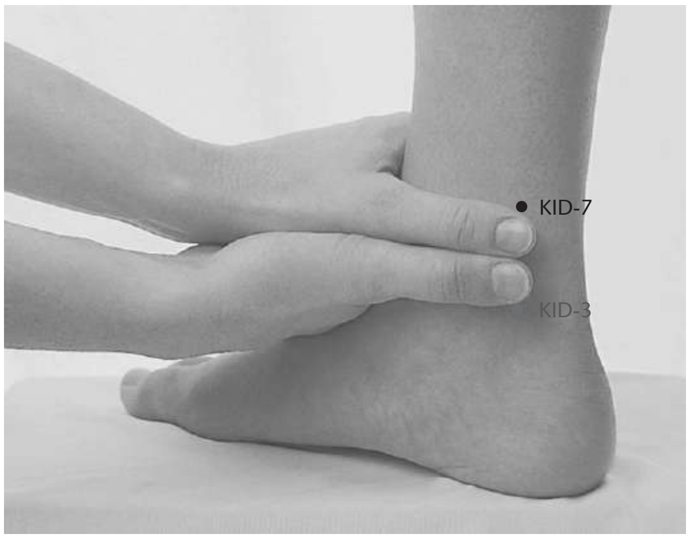

Needling

Vertically 0.5–1 cun

Actions/Indications

- Regulates the water passages and treats oedema, strengthens the Kidneys (especially Kidney Yang), drains Dampness and clears Damp-Heat
- For disorders of the Intestines due to Damp-Heat
- Strengthens the lumbar region

Special features

Jing-river point, Metal point, tonification point

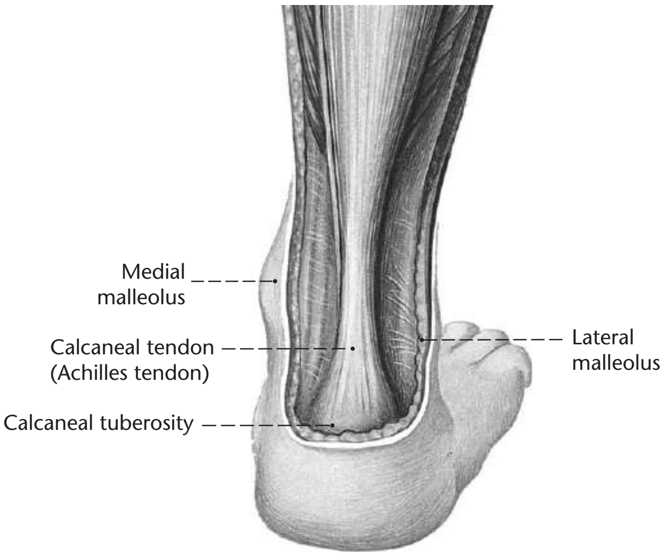
Medial malleolus
Calcaneal tendon (Achilles tendon)
Calcaneal tuberosity
Lateral malleolus

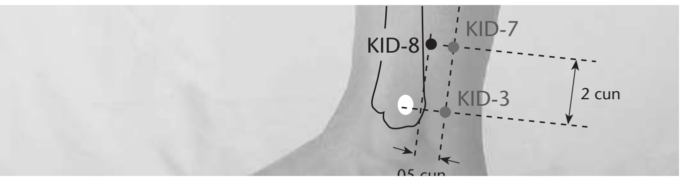
KID-8
KID-7
KID-3
2 cun

KID-8 *Exchange Belief* JIAOXIN

Location

2 cun proximal to the highest prominence of the medial malleolus, posterior to the medial border of the tibia.

How to find

KID-8 is located on the same level as → **KID-7**, but slightly closer to the border of the tibia (approximately 0.5 cun).

Needling

Vertically 0.5–1 cun

Actions/Indications

- Regulates menstruation, regulates the *ren mai* and *chong mai*
- Clears Heat and eliminates Dampness from the Lower Burner
- Treats blockages of the *yin qiao mai*

Special features

Xi-cleft point of the *yin qiao mai*

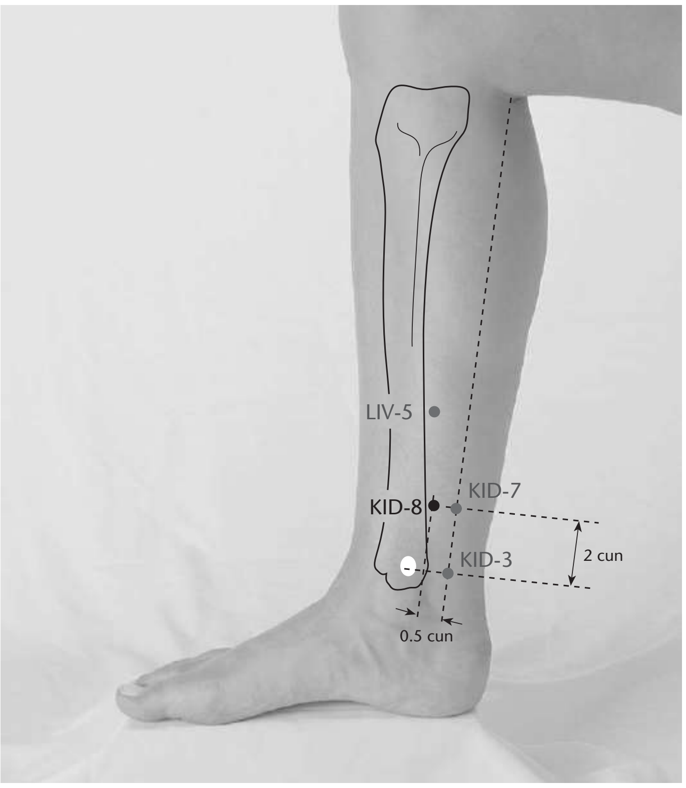
LIV-5
KID-7
KID-8
KID-3
2 cun
0.5 cun

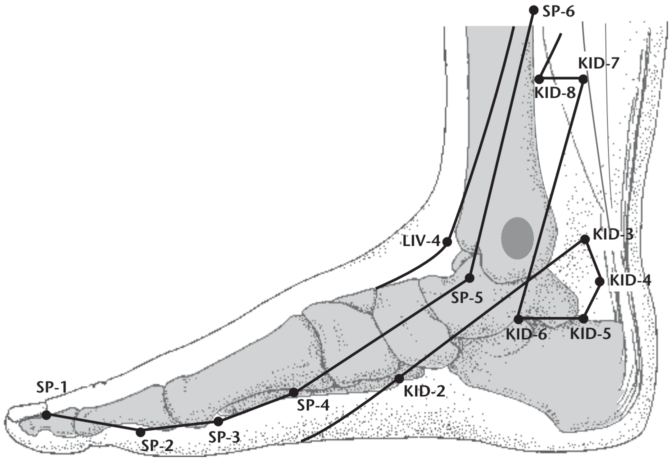
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

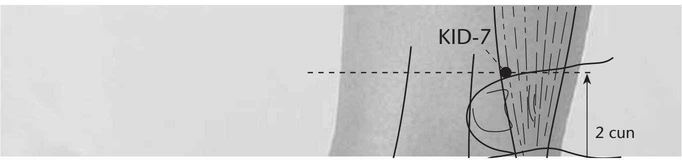

Guesthouse ZHUBIN KID-9

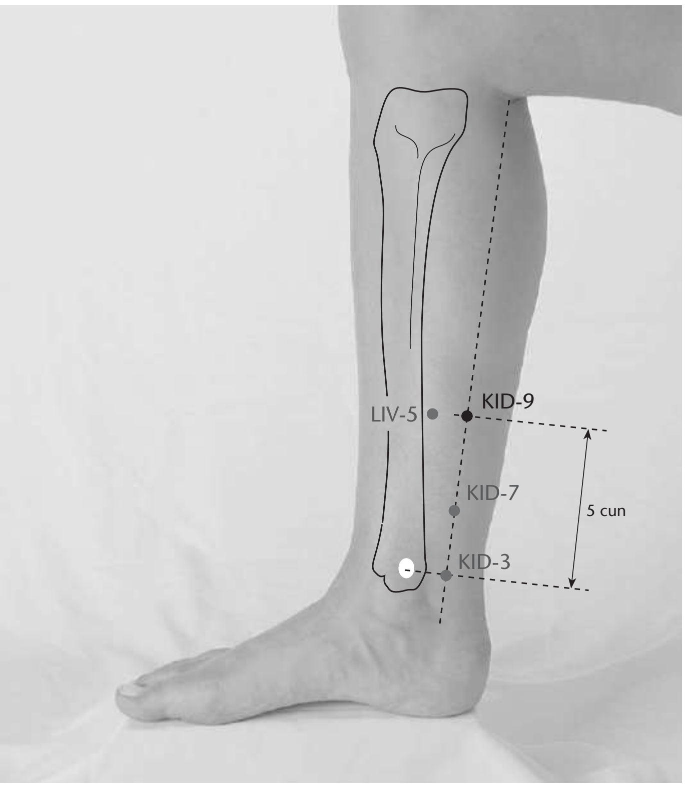
LIV-5
KID-9
KID-7
KID-3
5 cun

Location

5 cun proximal to the highest prominence of the medial malleolus, 2 cun posterior to the medial border of the tibia.

How to find

On a line connecting → **KID-3** (posterior to the medial malleolus) and → **KID-10** (medial aspect of the popliteal crease), measure 5 cun from the level of the prominence of the medial malleolus. Here, **KID-9** is located 2 cun posterior to the medial border of the tibia.
Located on the same level, but directly posterior to the medial border of the tibia, is → **LIV-5**.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Clears the Heart and transforms Phlegm
- Regulates the Qi and alleviates pain

Special features

Xi-cleft point of the *yin wei mai*

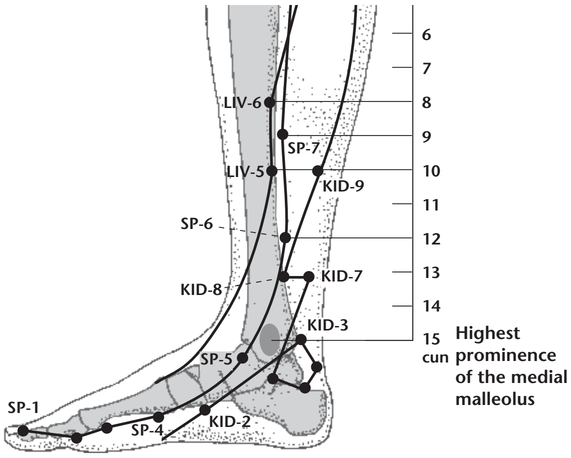
6
7
8
9
10
11
12
13
14
15
cun
LIV-6
SP-7
LIV-5
KID-9
SP-6
KID-8
KID-7
KID-3
SP-5
SP-1
SP-4
KID-2
Highest prominence of the medial malleolus

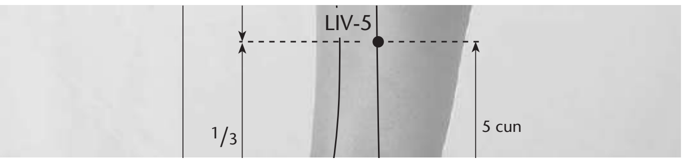
LIV-5
1/3
5 cun

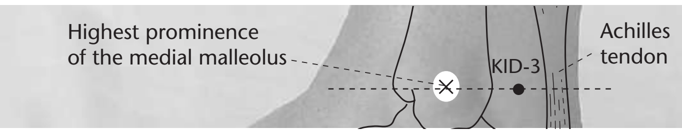
Highest prominence of the medial malleolus
KID-3
Achilles tendon

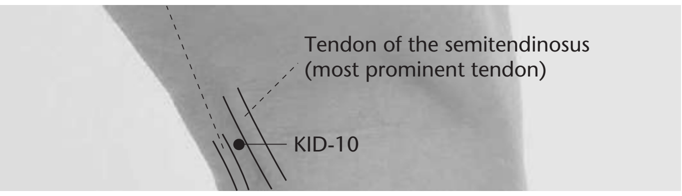
Tendon of the semitendinosus (most prominent tendon)
KID-10

KID-10 *Yin Valley* YINGU

Location

At the medial end of the popliteal crease, between the tendons of the semimembranosus and semitendinosus muscles, on the level of the knee joint space.

How to find

With the knee flexed less than 90°, ask the patient to press their heel against the treatment couch so that the two tendons will become more pronounced in the popliteal crease. From a medial direction, **KID-10** is located in a small gap anterior to the more prominent tendon of the semitendinosus muscle and posterior to the less well-defined tendon of the semimembranosus muscle.
→ **BL-40** is located on the same level, in the centre of the popliteal crease.

Needling

Needle 1–1.5 cun vertically from a dorsomedial aspect towards the tibial tuberosity. Avoid puncturing the tendon.

Actions/Indications

- Drains Damp-Heat from the Lower Burner and benefits the Kidneys
- Opens the channel

Special features

He-sea point, *ben* point (Five Phases point); for recurring urinary disorders with Kidney Yin deficiency.

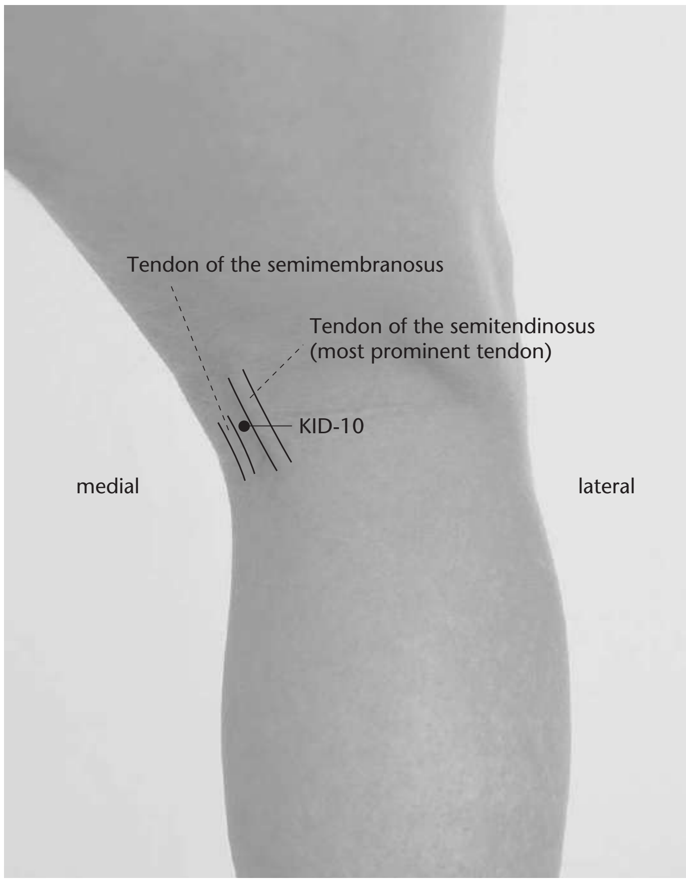
Tendon of the semimembranosus
Tendon of the semitendinosus (most prominent tendon)
KID-10
medial
lateral

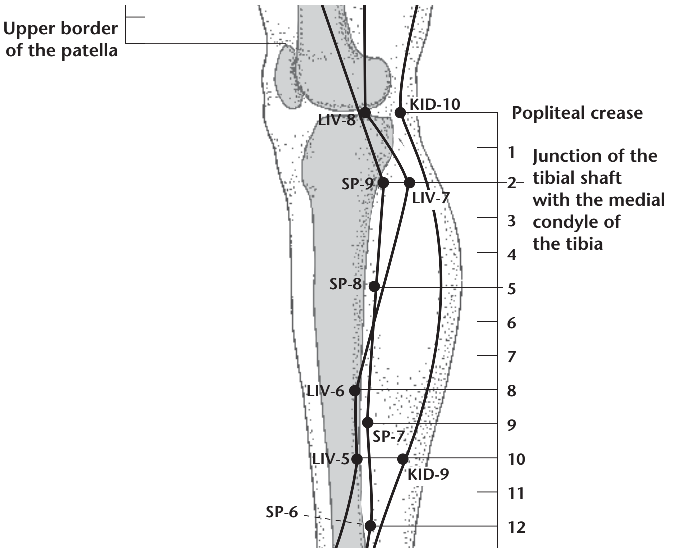
Upper border of the patella
KID-10
Popliteal crease
LIV-8
1
Junction of the tibial shaft with the medial condyle of the tibia
SP-9
2
LIV-7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
LIV-5
10
KID-9
11
SP-6
12

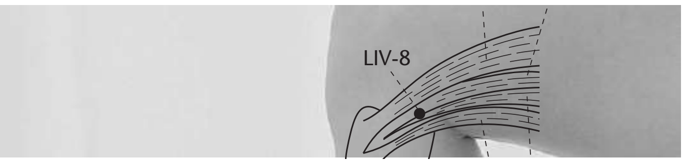

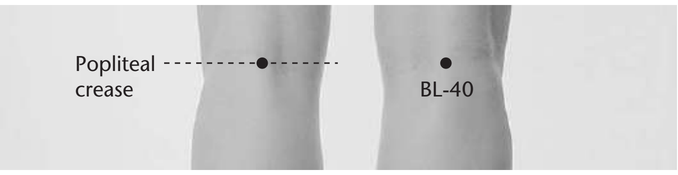
Popliteal crease
BL-40

Pubic Bone HENGGU KID-11

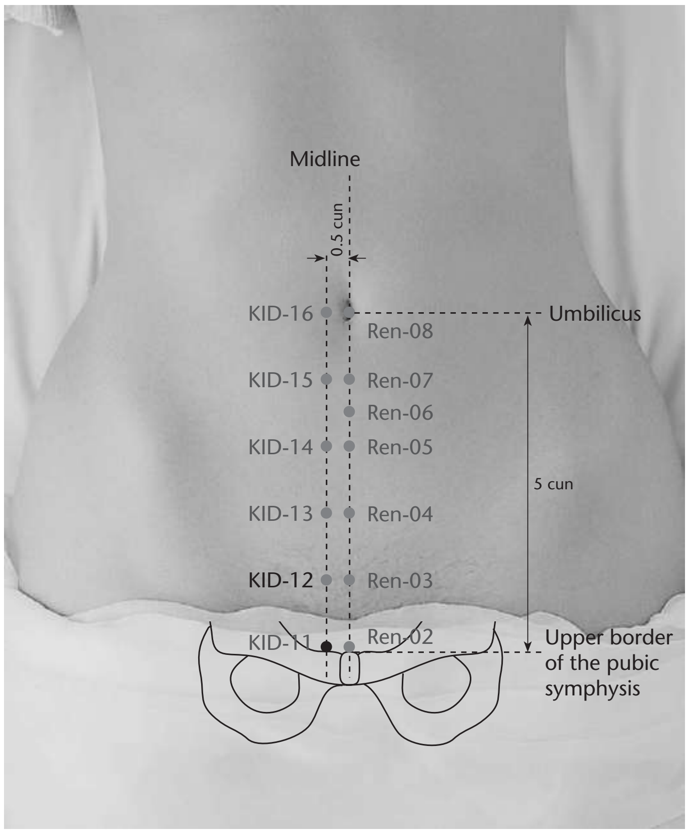
Midline
0.5 cun
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

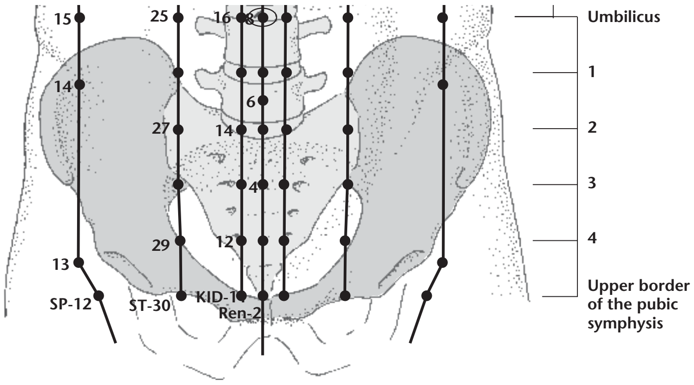
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

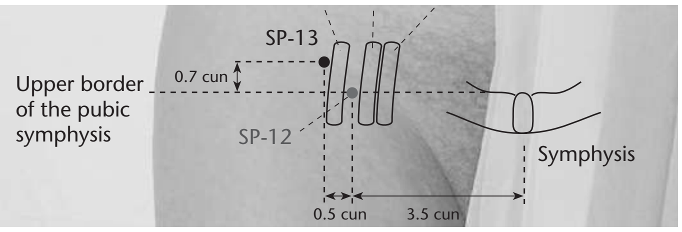
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

Location

At the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline.

How to find

Close to the midline, the upper border of the pubic symphysis is generally easily palpable. **KID-11** is located directly superior to its bony margin and 0.5 cun lateral to the midline.
Located on the same level are → **Ren-2** (on the midline), → **ST-30** (2 cun lateral to the midline) and → **SP-12** (3.5 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

- Regulates the Lower Burner and the water passages, strengthens the Kidneys

Special features

Meeting point with the *chong mai*

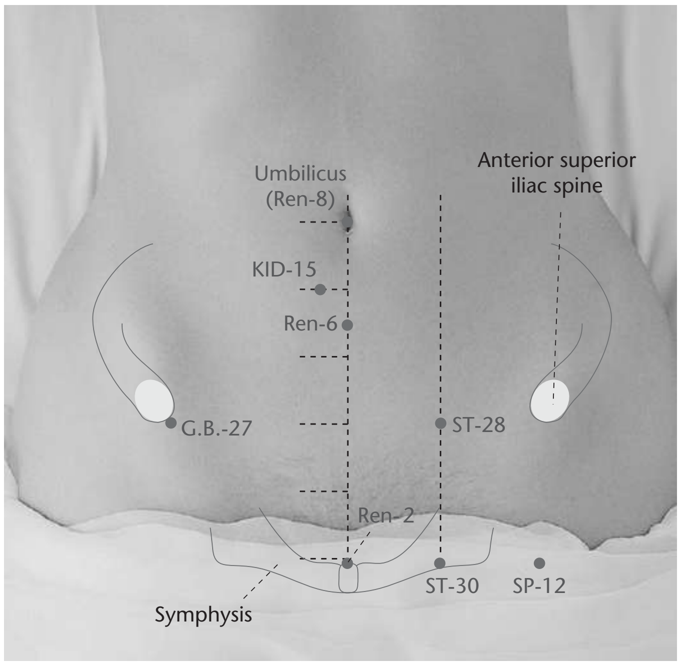
Umbilicus (Ren-8)
Anterior superior iliac spine
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
ST-30
SP-12
Symphysis

Ch04.8-F10028.qxd 2/22/08 6:14 PM Page 332

KID-12 *Great Luminance* DAHE

Location

1 cun superior to the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun. As these can vary considerably from finger cun measurements, only proportional cun measurements should be applied (elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 1 cun in a superior direction and on this level locate **KID-12** 0.5 cun lateral to the midline. Located on the same level are → **Ren-3** (on the midline), → **ST-29** (2 cun lateral to the midline) and → **Ex-CA-1** (*zigong*; 3 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

- Regulates the Lower Burner, strengthens the Kidneys and the Essence (*jing*)

Special features

Meeting point with the *chong mai*

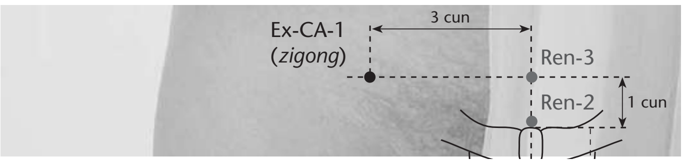
Ex-CA-1
(zigong)
3 cun
Ren-3
Ren-2
1 cun

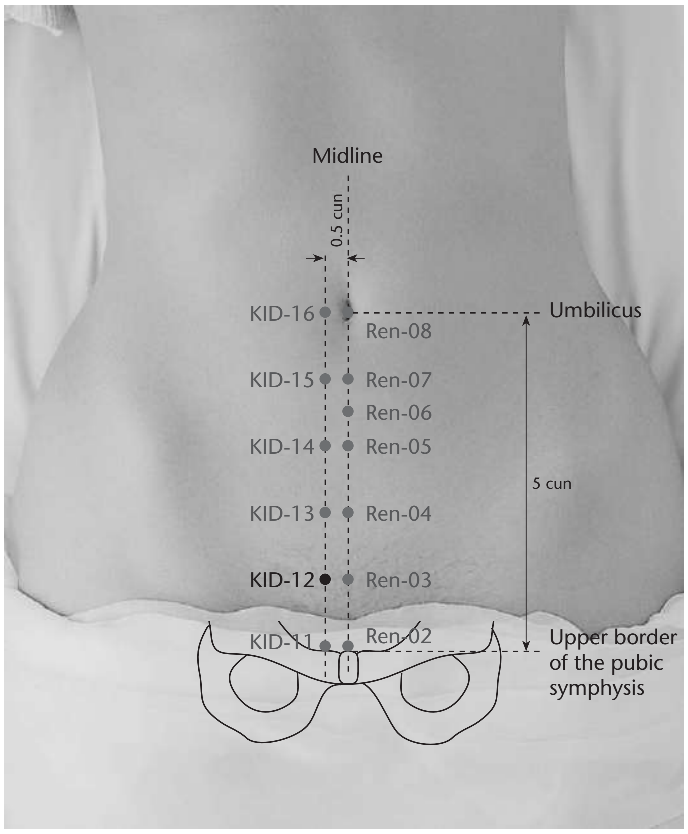
Midline
0.5 cun
KID-16
Ren-08
Umbilicus
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border
of the pubic
symphysis

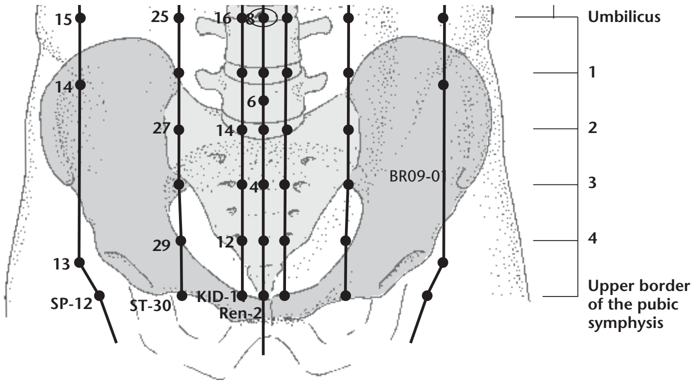
15
25
16
8
Umbilicus
1
14
6
27
14
2
BR09-01
4
3
29
12
4
13
SP-12
ST-30
KID-11
Ren-2
Upper border
of the pubic
symphysis

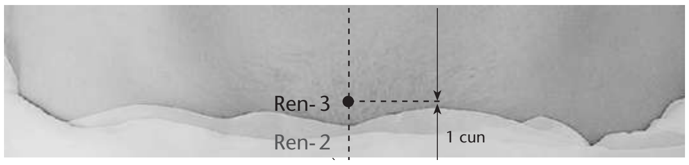
Ren-3
Ren-2
1 cun

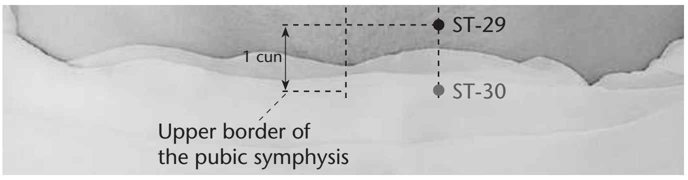
ST-29
1 cun
ST-30
Upper border of
the pubic symphysis

Qi Cave QIXUE KID-13

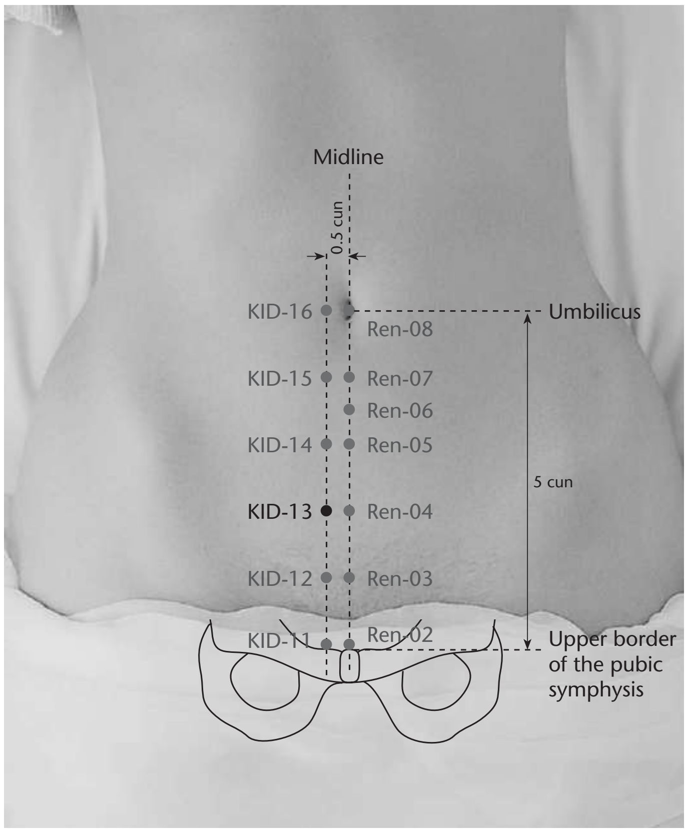
Midline
0.5 cun
KID-16
Ren-08
Umbilicus
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

Location

2 cun superior to the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun. As these can vary considerably from finger cun measurements, only proportional cun measurements should be applied (elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun in a superior direction and on this level locate **KID-13** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-4** (on the midline), → **ST-28** (2 cun lateral to the midline), → **Ex-CA** (*tituo*/*qimen*/*yijing*: 4/3/1 cun lateral to the midline) and → **G.B.-27** (anterior and medial to the ASIS).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

- Regulates the Lower Burner, regulates the *chong mai* and *ren mai*, strengthens the Kidneys and the essence (*jing*)

Special features

Meeting point with the *chong mai*

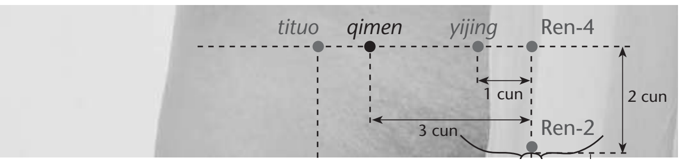
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

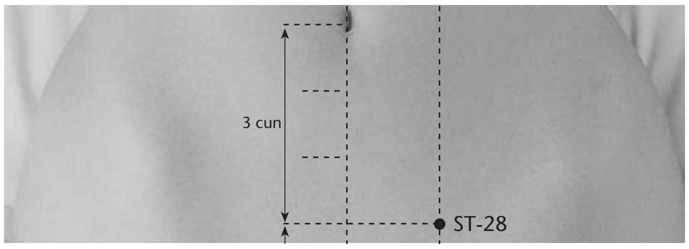

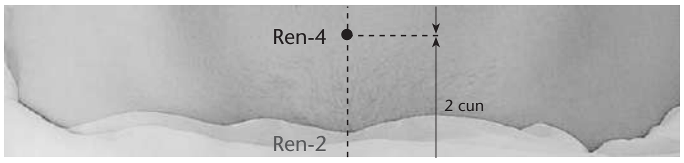
Ren-4
2 cun
Ren-2

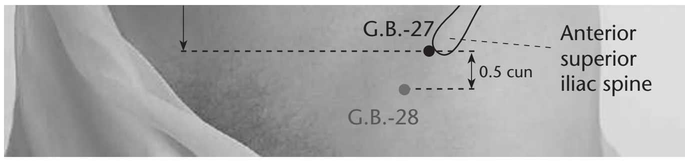
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

Ch04.8-F10028.qxd 2/22/08 6:14 PM Page 334

KID-14 *Four Fullnesses* SIMAN

Location

2 cun inferior to the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun. As these can vary considerably from finger cun measurements, only proportional cun measurements should be applied (elastic tape → 2.3.1). From the centre of the umbilicus, measure 2 cun in an inferior direction and on this level locate **KID-14** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-5** (on the midline) and → **ST-27** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Lower Burner and benefits the Kidneys
- Regulates the water passages
- Regulates the Qi and moves Blood stasis

Special features

Meeting point with the *chong mai*

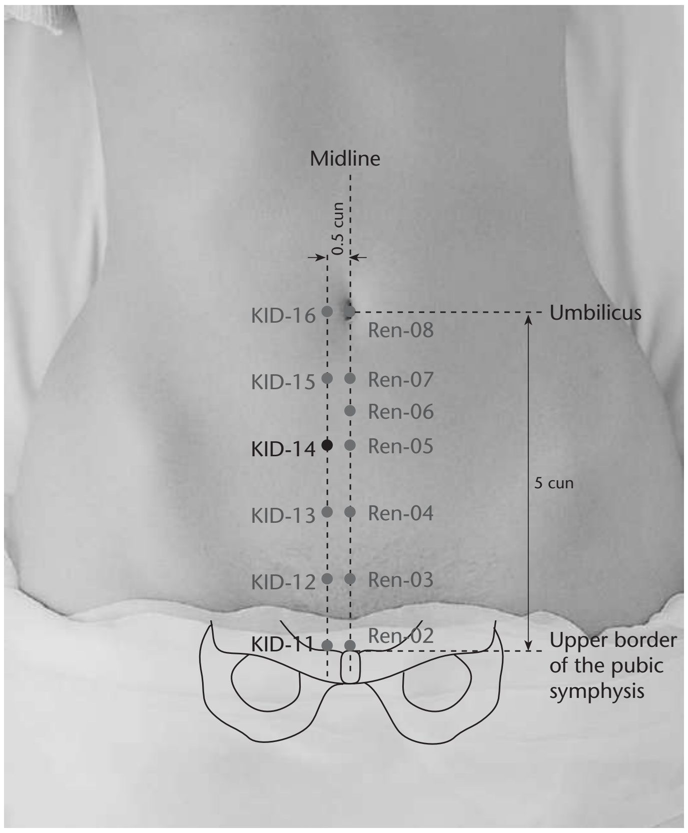
Midline
0.5 cun
KID-16
Ren-08
Umbilicus
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

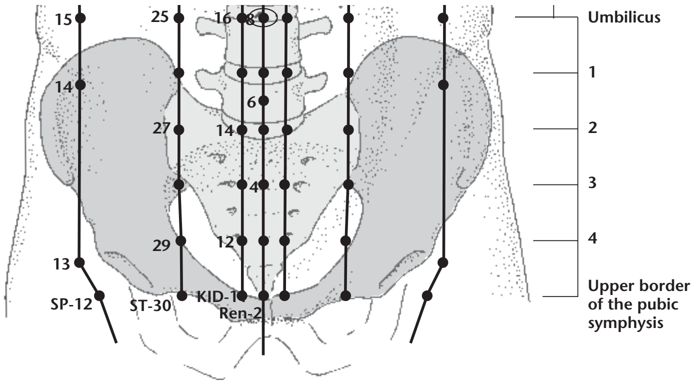
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

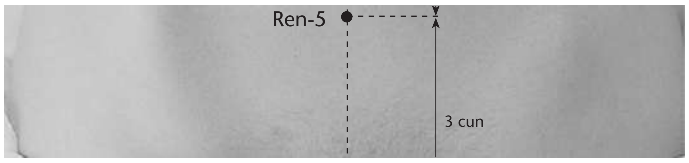

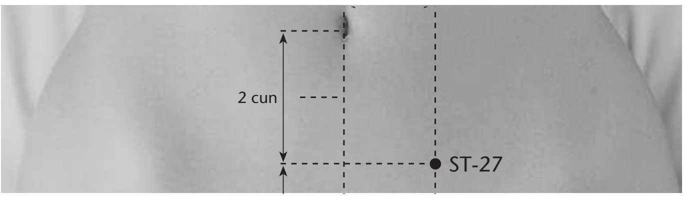

Middle Flow ZHONGZHU KID-15

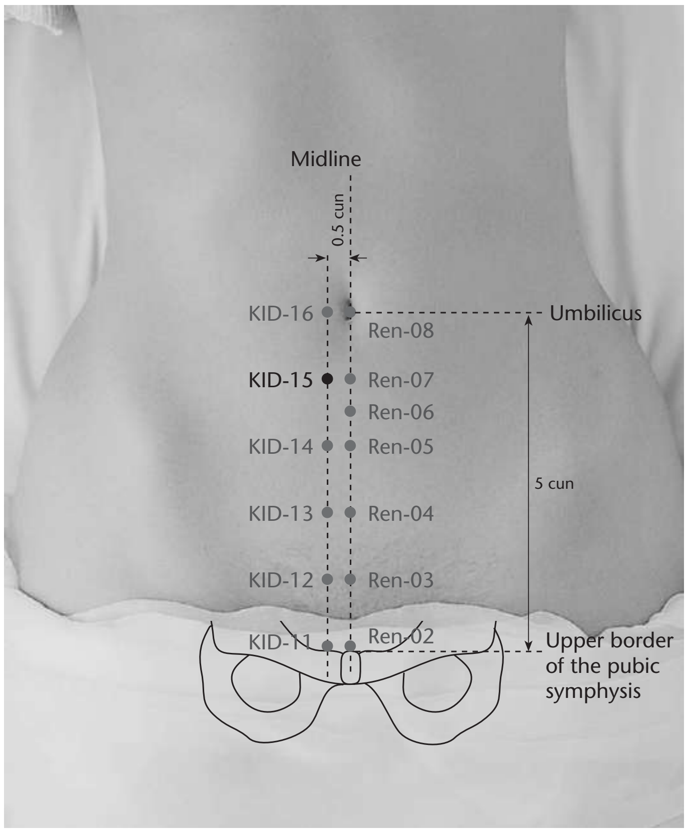
Midline
0.5 cun
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

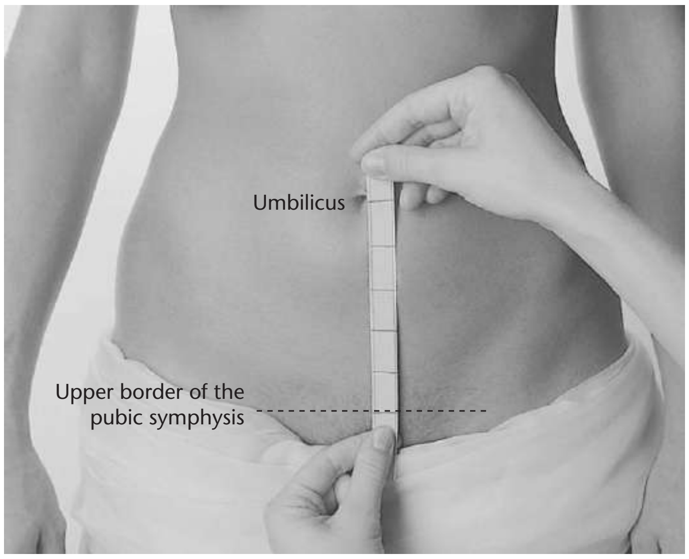
Umbilicus
Upper border of the pubic symphysis

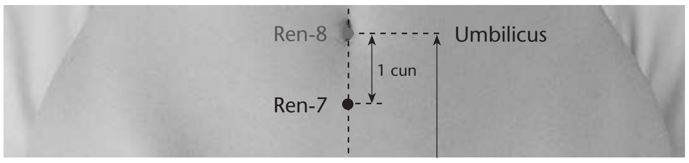
Ren-8
Umbilicus
1 cun
Ren-7

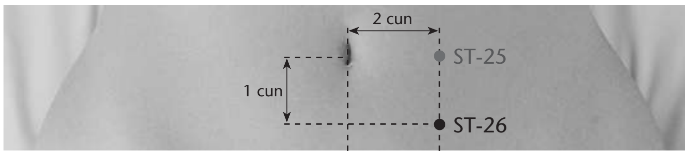
2 cun
ST-25
1 cun
ST-26

Location

1 cun inferior to the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun. As these can vary considerably from finger cun measurements, only proportional cun measurements should be applied (elastic tape → 2.3.1). From the centre of the umbilicus, measure 1 cun in an inferior direction and on this level locate **KID-15** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-7** (on the midline) and → **ST-26** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Intestines
- Regulates the Lower Burner

Special features

Meeting point with the *chong mai*

KID-16 *Vitals* Shu HUANGSHU

Location

0.5 cun lateral to the centre of the umbilicus.

How to find

From the centre of the umbilicus, measure 0.5 cun in a lateral direction. Located on the same level are → **Ren-8** (in the centre of the umbilicus), → **ST-25** (2 cun lateral to the midline), → **SP-15** (4 cun lateral to the midline) and → **G.B.-26** (on the level of umbilicus, on a vertical line drawn through the free end of the 11th rib).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy. Avoid the umbilicus.

Actions/Indications

- Regulates the Qi, regulates and warms the Stomach and Intestines

Special features

Meeting point with the *chong mai*

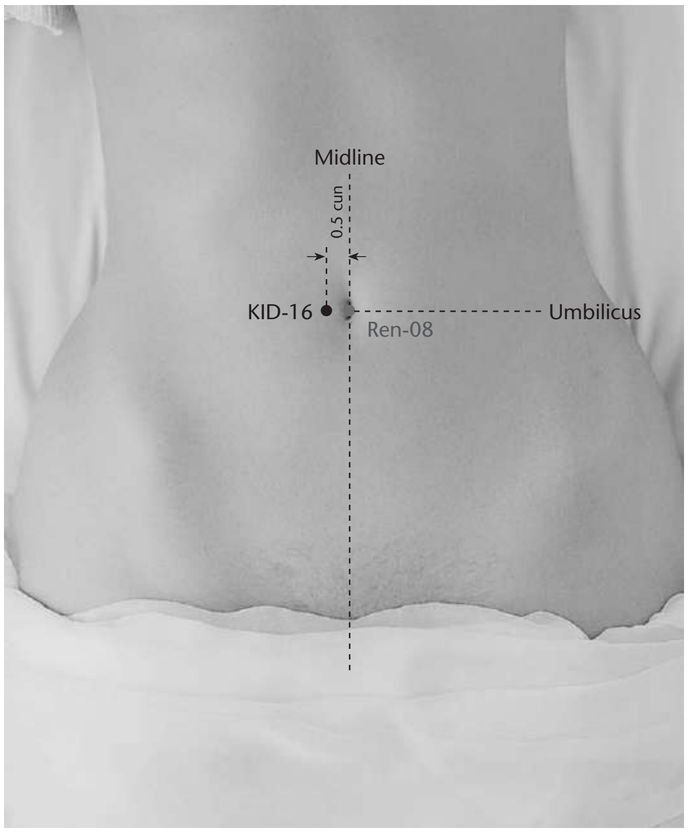
Midline
0.5 cun
KID-16
Umbilicus
Ren-08

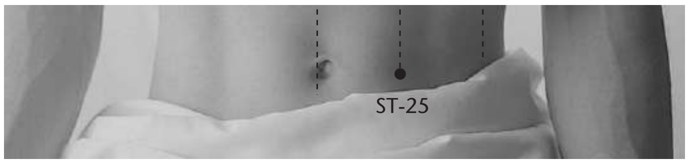

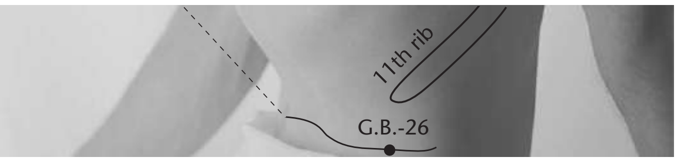
11th rib
G.B.-26

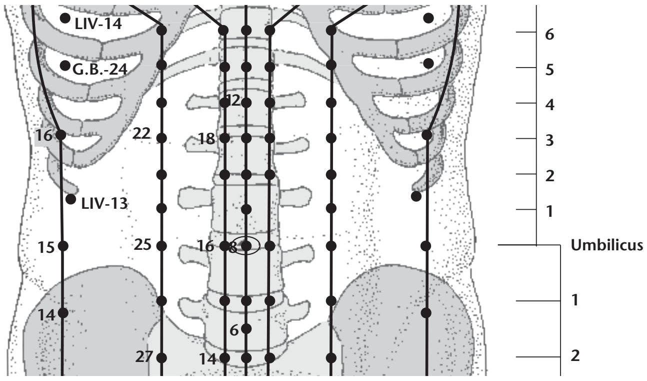
LIV-14
G.B.-24
16
22
12
18
LIV-13
15
25
16
8
14
6
27
14
6
5
4
3
2
1
Umbilicus
1
2

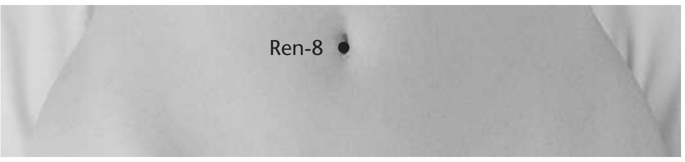

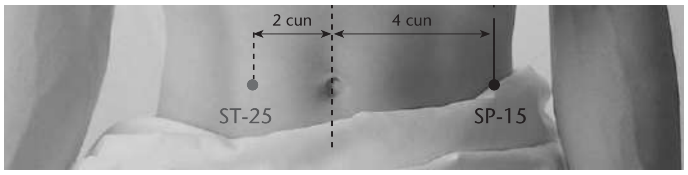
2 cun
4 cun
ST-25
SP-15

Shang Bend SHANGQU KID-17

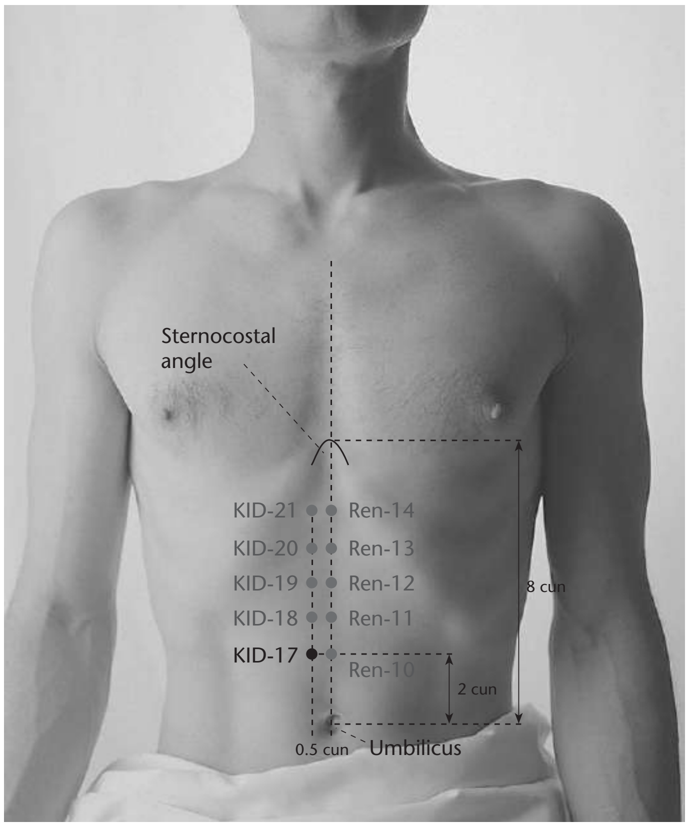
Sternocostal angle
KID-21
Ren-14
KID-20
Ren-13
KID-19
Ren-12
KID-18
Ren-11
8 cun
KID-17
Ren-10
2 cun
0.5 cun
Umbilicus

Location

2 cun superior to the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun should be used (elastic tape → 2.3.1). From the centre of the umbilicus, measure 2 cun in a superior direction and on this level locate **KID-17** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-10** (on the midline), → **ST-23** (2 cun lateral to the midline) and **Ex-CA** (*weishang*, 4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Stomach and Intestines, eliminates stagnation

Special features

Meeting point with the *chong mai*

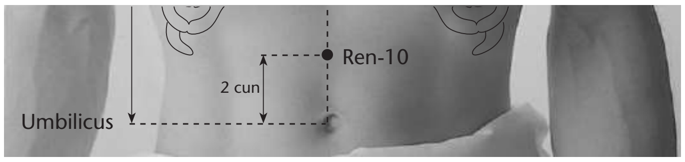
Ren-10
2 cun
Umbilicus

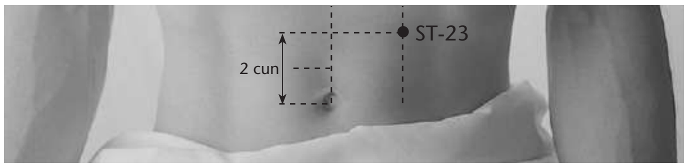

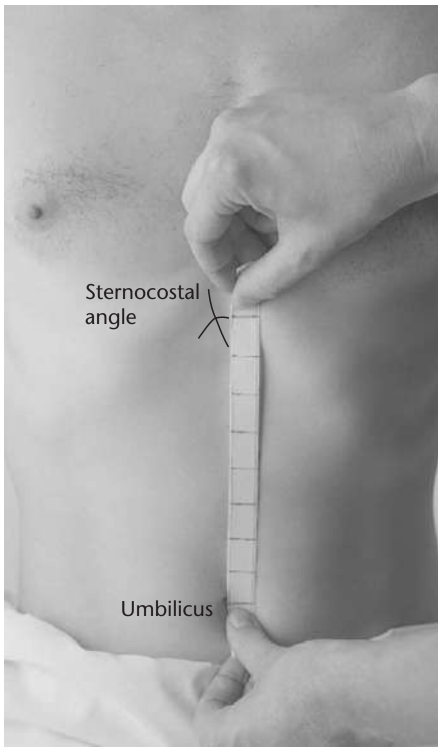
Sternocostal angle
Umbilicus

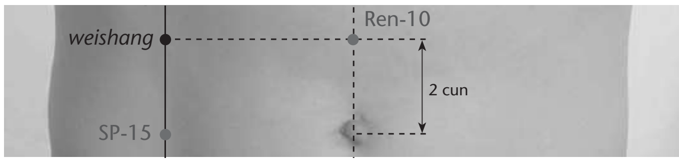
Ren-10
weishang
2 cun
SP-15

KID-18 Stone Gate SHIGUAN

Location

3 cun superior to the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (elastic tape → 2.3.1). From the centre of the umbilicus measure 3 cun in a superior direction and on this level locate **KID-18** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-11** (on the midline), → **ST-22** (2 cun lateral to the midline) and → **SP-16** (on the mamillary line).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Strengthens the Middle, regulates the Stomach and Intestines, removes Blood stagnation

Special features

Meeting point with the *chong mai*

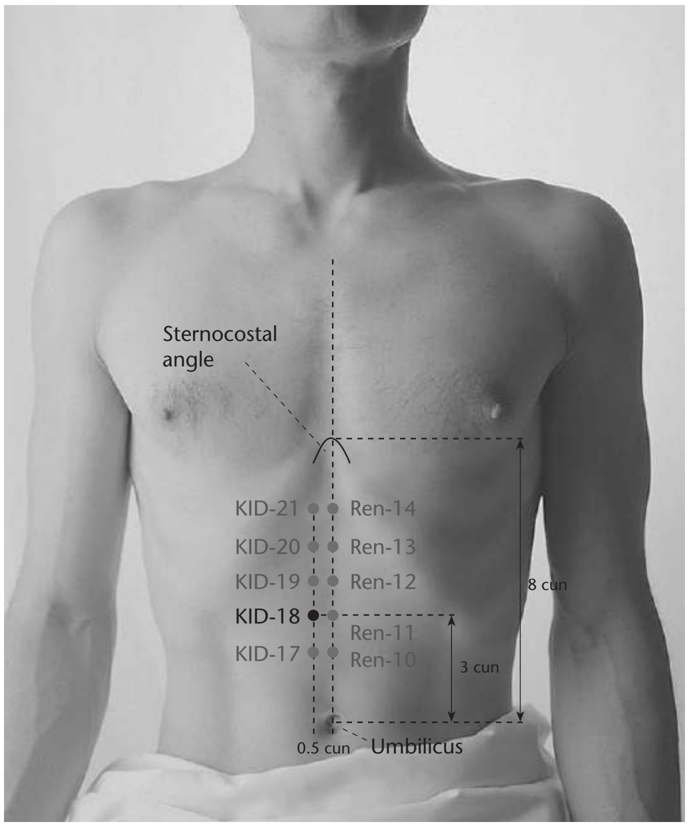
Sternocostal angle
KID-21
Ren-14
KID-20
Ren-13
KID-19
Ren-12
KID-18
Ren-11
KID-17
Ren-10
8 cun
3 cun
0.5 cun
Umbilicus

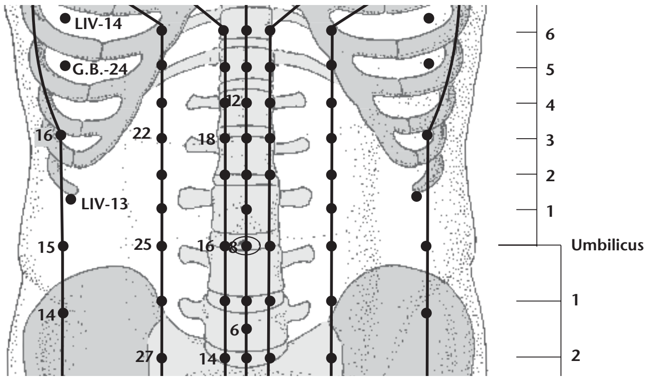
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
27
6
14
6
5
4
3
2
1
Umbilicus
1
2

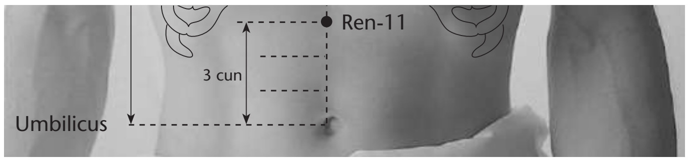
Ren-11
3 cun
Umbilicus

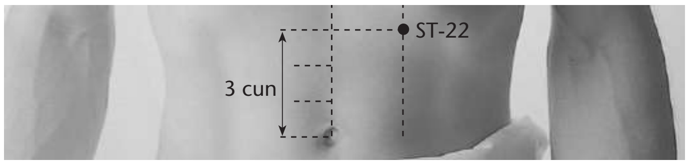

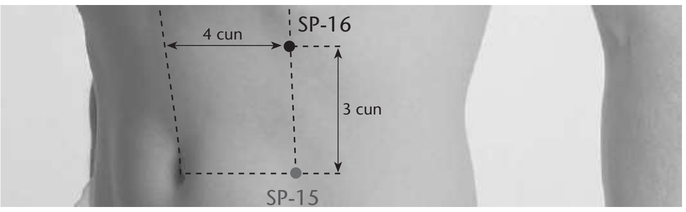
4 cun
SP-16
3 cun
SP-15

Yin Metropolis YINDU KID-19

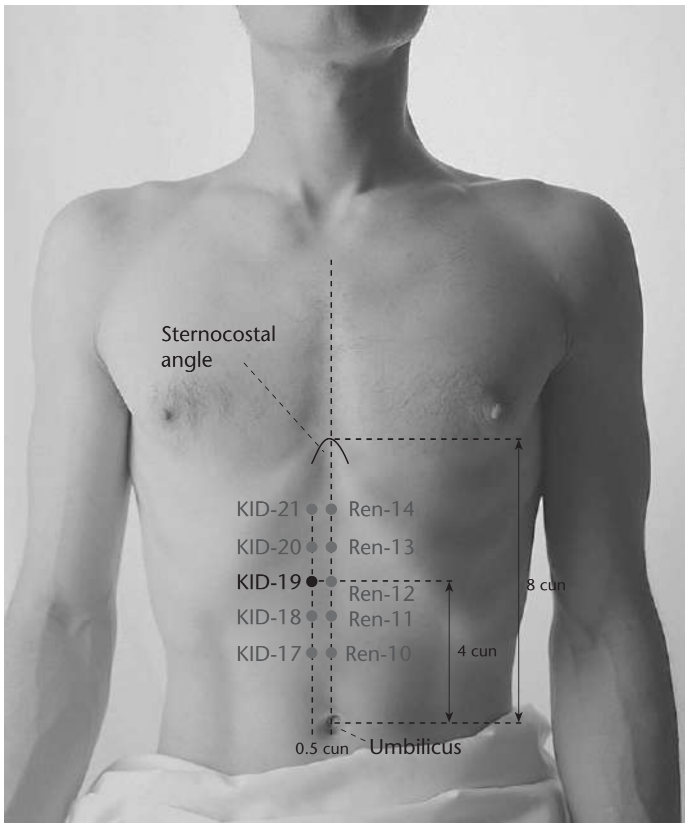
Sternocostal angle
KID-21
Ren-14
KID-20
Ren-13
KID-19
Ren-12
8 cun
KID-18
Ren-11
KID-17
Ren-10
4 cun
0.5 cun
Umbilicus

Location

Midway between the sternocostal angle and the centre of the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun should be used (elastic tape → 2.3.1). Spreading hands technique (→ 2.3.3): Place the little fingers on the centre of the umbilicus and the sternocostal angle respectively and join the thumbs at the midpoint of this distance by evenly spreading both hands. **KID-19** is located on this level, 0.5 cun lateral to the midline.

Located on the same level are → **Ren-12** (on the midline) and → **ST-21** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Qi, regulates the Stomach and Intestines
- Regulates counterflow Qi

Special features

Meeting point with the *chong mai*

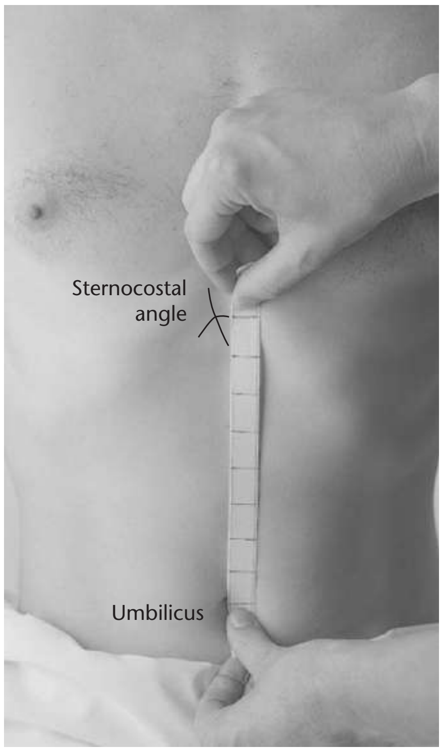
Sternocostal angle
Umbilicus

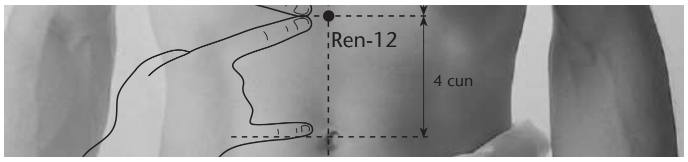

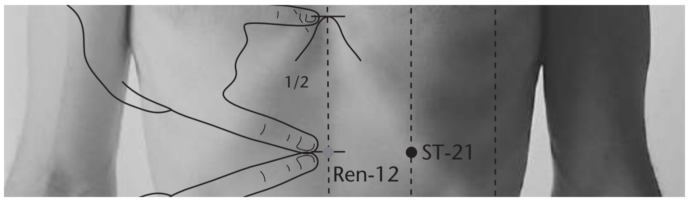
1/2
ST-21
Ren-12

KID-20 *Abdomen Connecting Valley* FUTONGGU

Location

5 cun superior to the umbilicus or 3 cun inferior to the sternocostal angle, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (elastic tape → 2.3.1). From the sternocostal angle, measure 3 cun in an inferior direction and on this level locate **KID-20** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-13** (on the midline), → **ST-20** (2 cun lateral to the midline) and → **G.B.-24** (on the mamillary line, in the 7th intercostal space).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Strengthens the Spleen, harmonises the Stomach and Intestines, regulates counterflow Qi

Special features

Meeting point with the *chong mai*

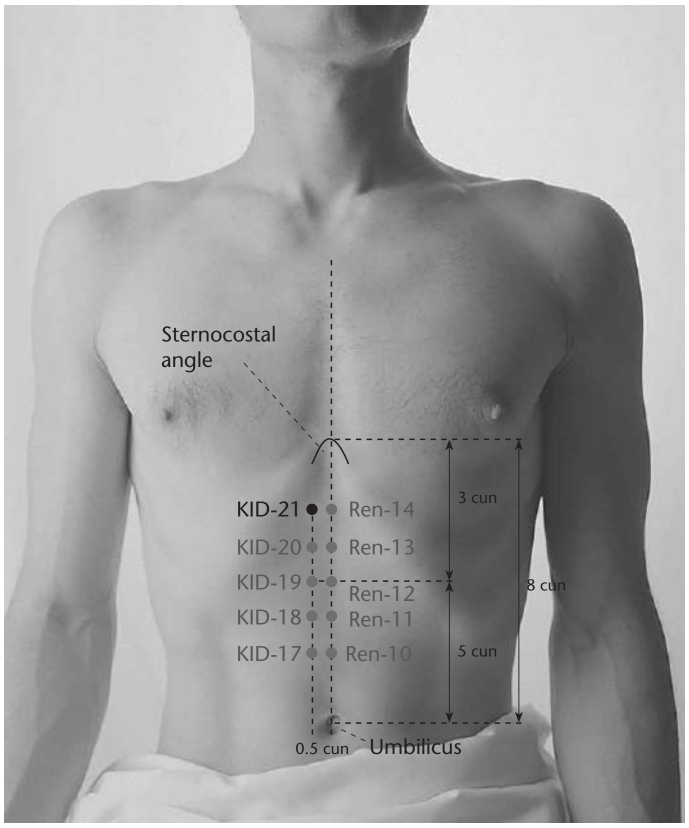
Sternocostal angle
KID-21
KID-20
KID-19
KID-18
KID-17
Ren-14
Ren-13
Ren-12
Ren-11
Ren-10
3 cun
8 cun
5 cun
0.5 cun
Umbilicus

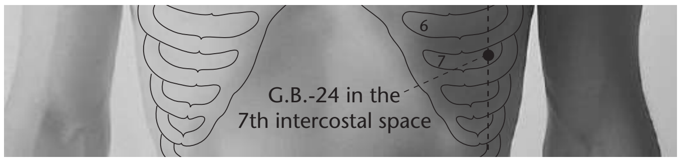
6
7
G.B.-24 in the
7th intercostal space

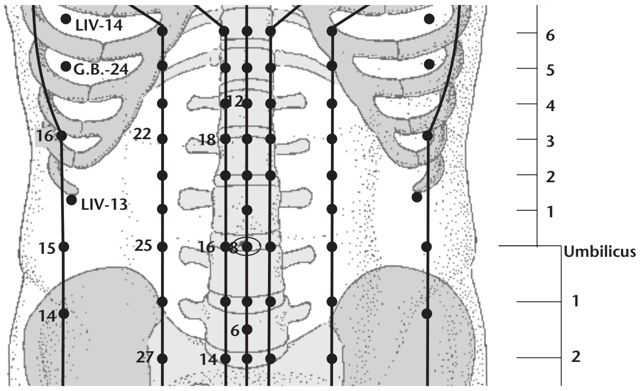
LIV-14
G.B.-24
16
22
18
21
LIV-13
15
25
16
8
14
6
27
14
6
5
4
3
2
1
Umbilicus
1
2

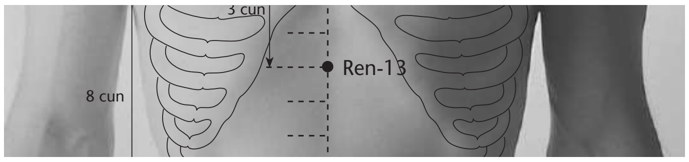
3 cun
8 cun
Ren-13

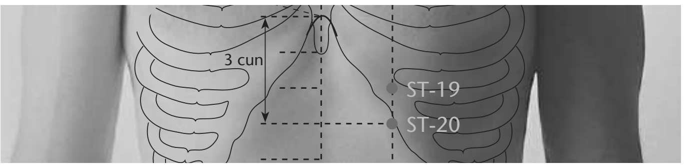
3 cun
ST-19
ST-20

Hidden Gate YOUMEN KID-21

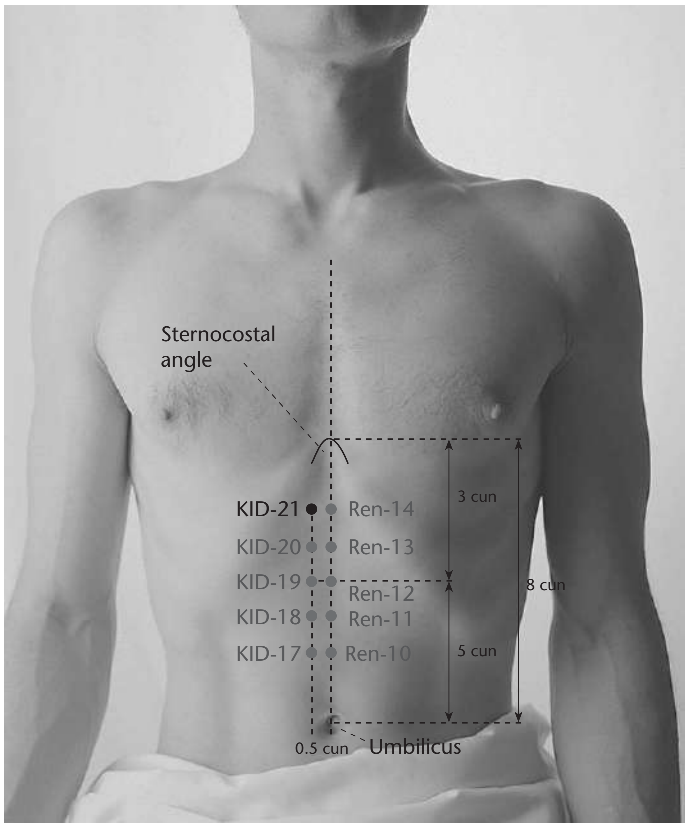
Sternocostal angle
3 cun
KID-21
Ren-14
KID-20
Ren-13
KID-19
Ren-12
8 cun
KID-18
Ren-11
KID-17
Ren-10
5 cun
0.5 cun
Umbilicus

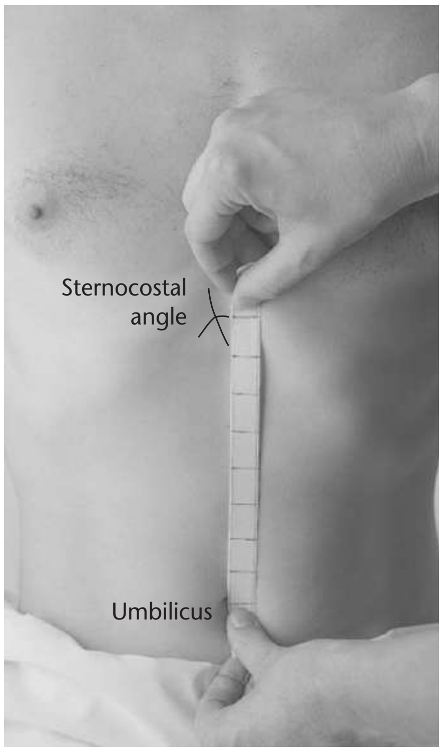
Sternocostal angle
Umbilicus

Location

2 cun inferior to the sternocostal angle, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun should be used (elastic tape → 2.3.1). From the sternocostal angle, measure 2 cun in an inferior direction and on this level locate **KID-21** 0.5 cun lateral to the midline.
Located on the same level are → **Ren-14** (on the midline), → **ST-19** (2 cun lateral to the midline) and → **LIV-14** (on the mamillary line, in the 6th intercostal space).

Needling

Vertically 0.5–1 cun. Caution: The liver is located on the right, the peritoneum on the left. Caution during pregnancy.

Actions/Indications

- Harmonises the Stomach, regulates counterflow Qi

Special features

Meeting point with the *chong mai*. Last point of the *chong mai* located on the Kidney channel.

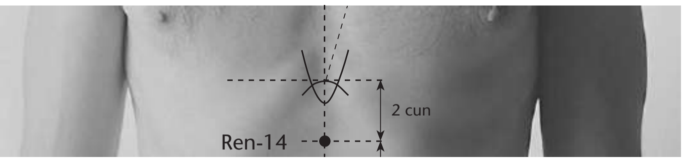

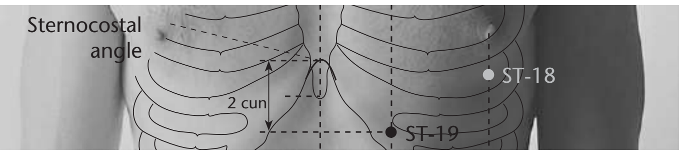
Sternocostal angle
ST-18
2 cun
ST-19

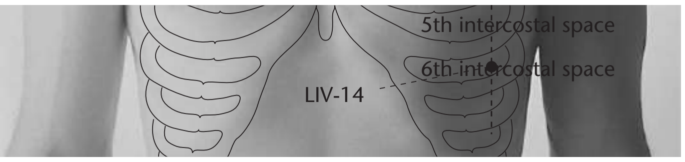
5th intercostal space
6th intercostal space
LIV-14

KID-22 *Walking Corridor* BULANG

Location

In the 5th intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Locate the 5th ICS either by counting parasternally from the clavicle or from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5), then locate **KID-22** 2 cun lateral to the anterior midline. Located on the same level (5th ICS) are → **Ren-16** (on the midline), → **ST-18** (on the mamillary line) and → **SP-17** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) with or against the channel pathway. Caution: Avoid injuring the liver (on the right side), the heart (on the left side) and the pleura.

Actions/Indications

- Regulates the Lung Qi and counterflow Qi, unbinds the chest

Special features

Exit point. The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

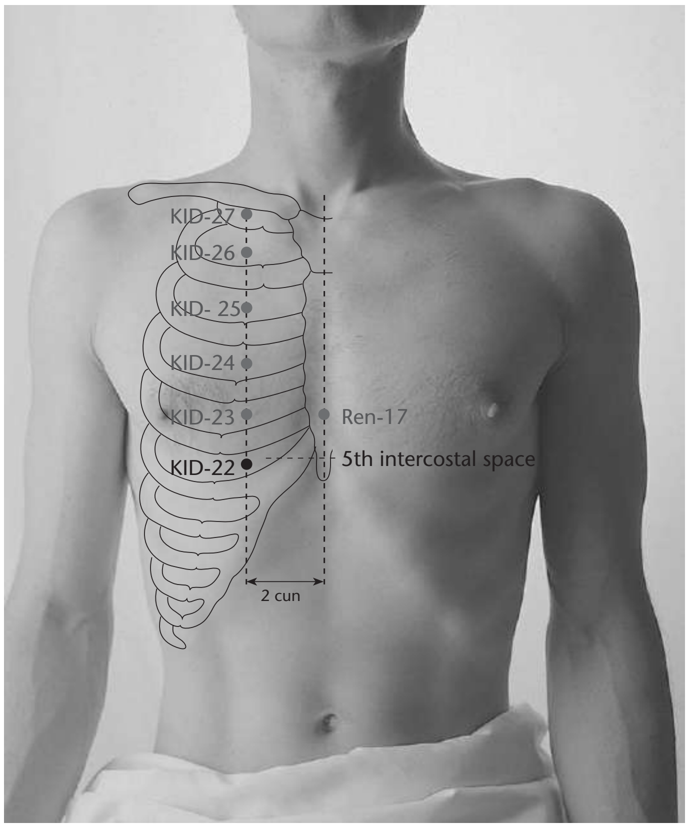
KID-27
KID-26
KID- 25
KID-24
KID-23
KID-22
Ren-17
5th intercostal space
2 cun

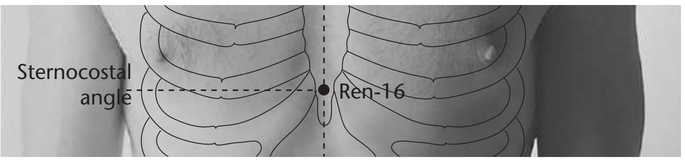
Sternocostal angle
Ren-16

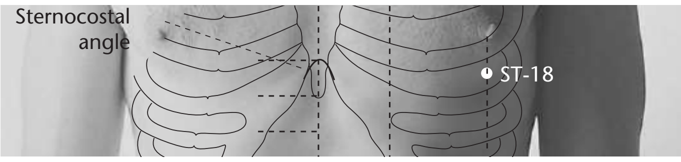
Sternocostal angle
ST-18

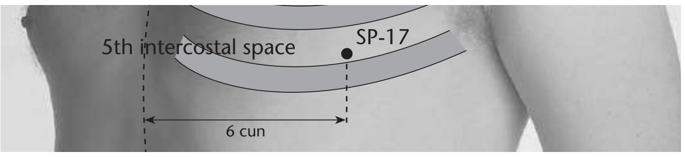
5th intercostal space
SP-17
6 cun

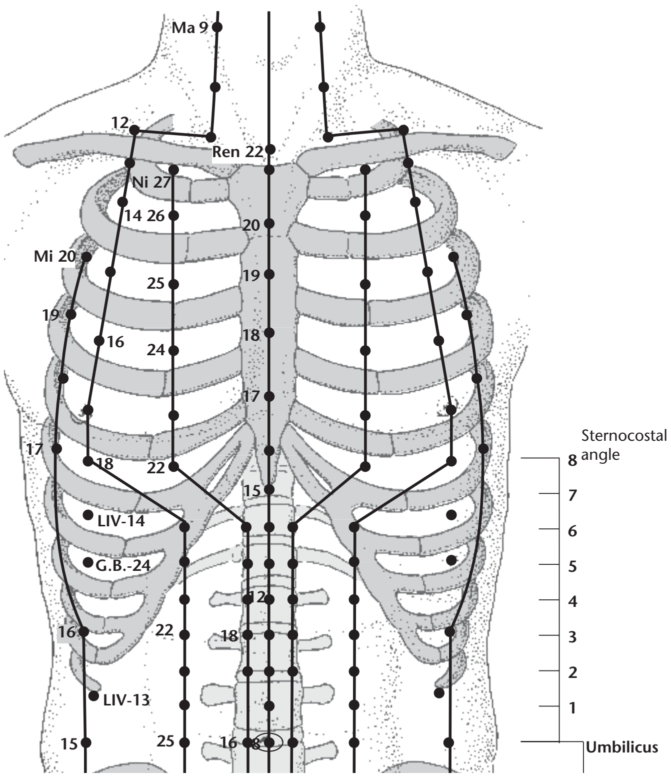
Ma 9
12
Ren 22
Ni 27
14 26
20
Mi 20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sternocostal angle
8
7
6
5
4
3
2
1
Umbilicus

Spirit Seal SHENFENG KID-23

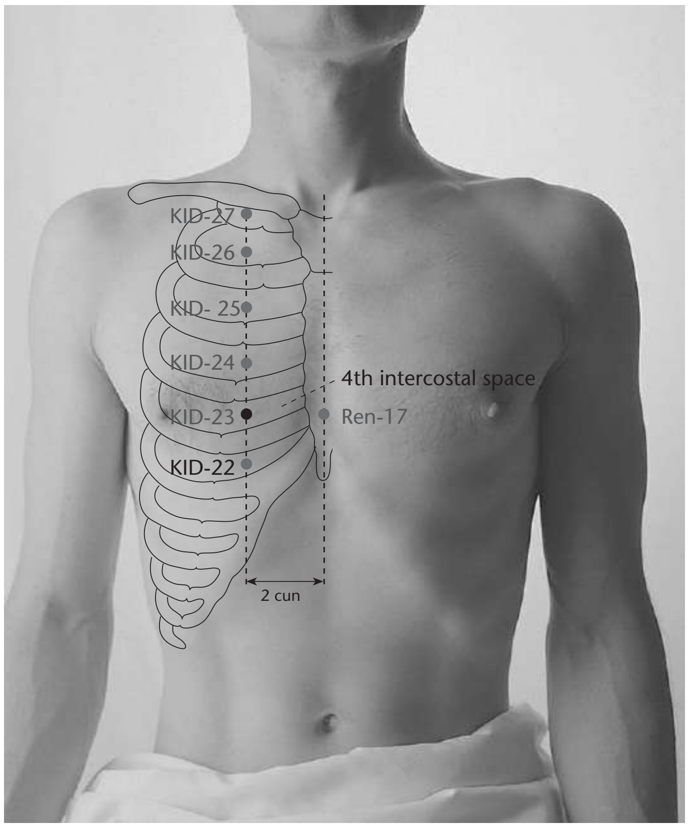
KID-27
KID-26
KID-25
KID-24
KID-23
KID-22
4th intercostal space
Ren-17
2 cun

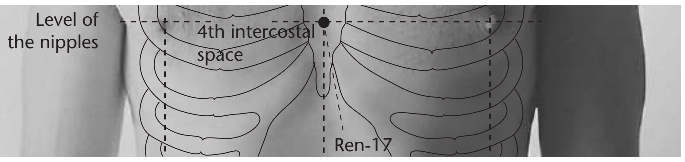
Level of the nipples
4th intercostal space
Ren-17

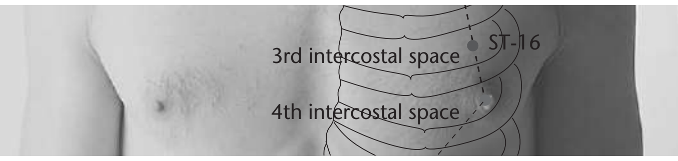
ST-16
3rd intercostal space
4th intercostal space

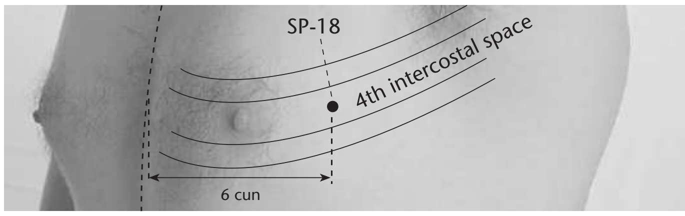
SP-18
4th intercostal space
6 cun

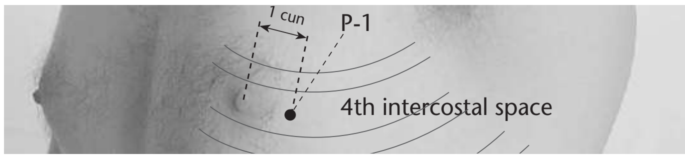
1 cun
P-1
4th intercostal space

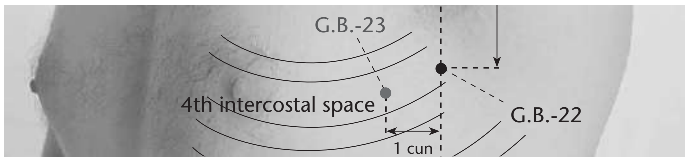
G.B.-23
4th intercostal space
G.B.-22
1 cun

Location

In the 4th intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Locate the 4th ICS either by palpating parasternally from the clavicle or by counting from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5), then locate **KID-23** 2 cun lateral to the anterior midline.

Located on the same level (4th ICS) are → **Ren-17** (on the midline), → **ST-17** (on the nipple), → **SP-18** (6 cun lateral to the midline), → **P-1** (1 cun lateral to the nipple) and → **G.B.-22** (3 cun inferior to the axilla) as well as → **G.B.-23** (1 cun anterior to **G.B.-22**).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Heart, pneumothorax.

Actions/Indications

- Regulates the Lung Qi and counterflow Qi, unbinds the chest
- Harmonises the Stomach
- Benefits the breasts

Special features

The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

KID-24 *Spirit Ruin* LINGXU

Location

In the 3rd intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Locate the 3rd ICS by palpating parasternally from the clavicle or by counting from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5), then locate **KID-24** 2 cun lateral to the anterior midline. Located on the same level (→ 3rd ICS) are → **Ren-18** (on the midline), → **ST-16** (4 cun lateral to the midline) and → **SP-19** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Lung and Stomach Qi
- Unbinds the chest
- Benefits the breasts

Special features

The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

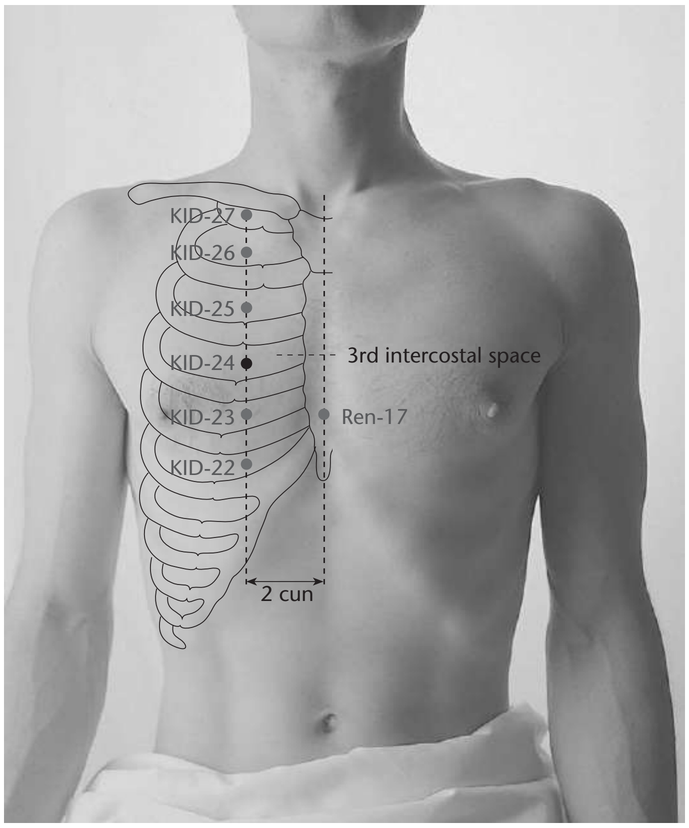
KID-27
KID-26
KID-25
KID-24
3rd intercostal space
KID-23
Ren-17
KID-22
2 cun

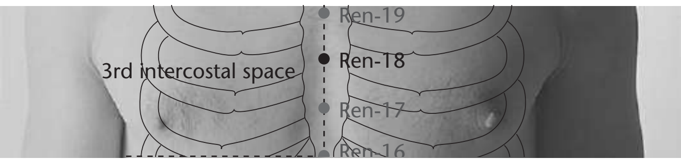
Ren-19
3rd intercostal space
Ren-18
Ren-17
Ren-16

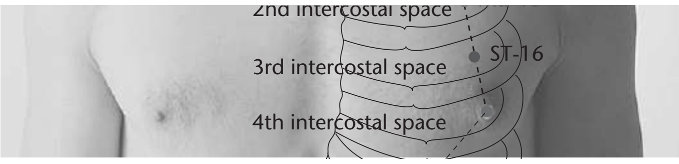
2nd intercostal space
ST-16
3rd intercostal space
4th intercostal space

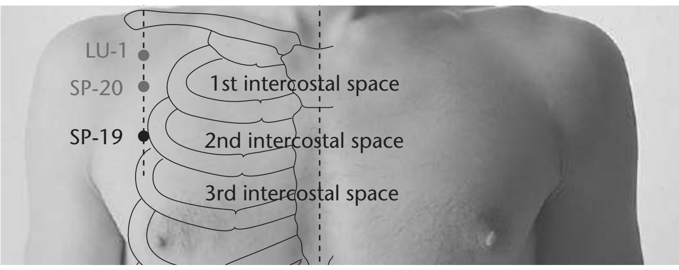
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

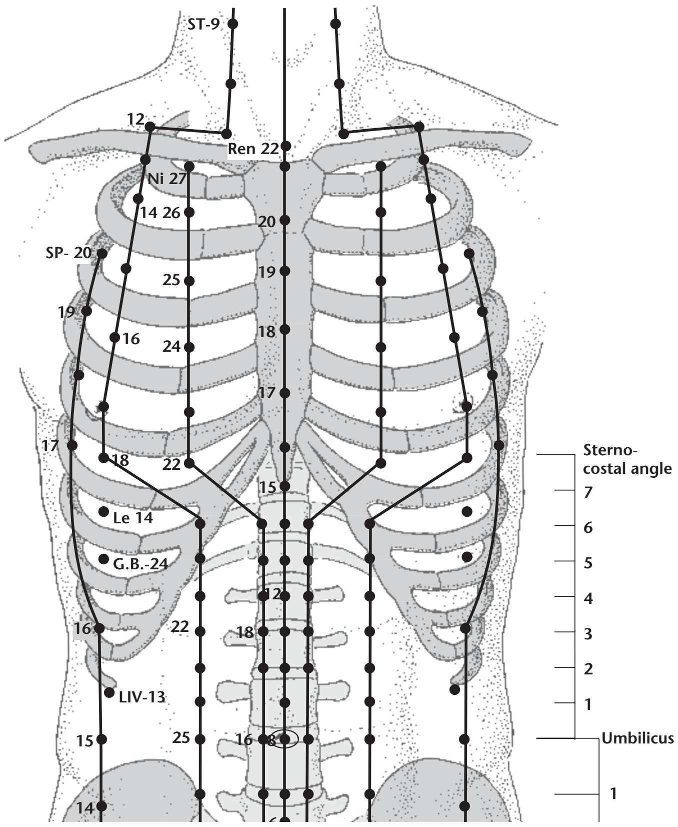
ST-9
12
Ren 22
Ni 27
14 26
20
SP- 20
25
19
19
16
24
18
17
17
18
22
15
Le 14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
Sterno-
costal angle
7
6
5
4
3
2
1
Umbilicus
1

Ch04.8-F10028.qxd 2/22/08 6:14 PM Page 345

Spirit Storehouse SHENCANG KID-25

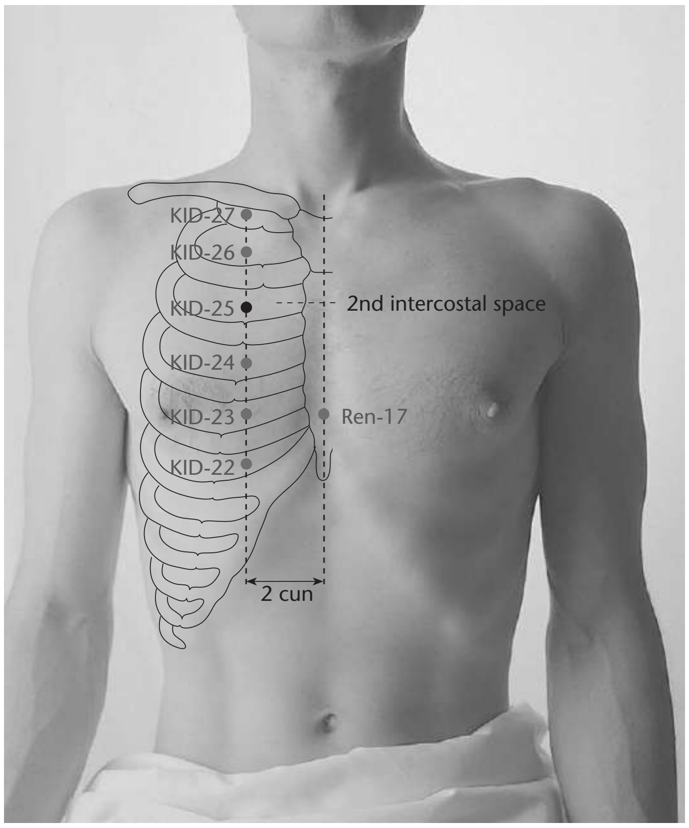
KID-27
KID-26
KID-25
2nd intercostal space
KID-24
KID-23
Ren-17
KID-22
2 cun

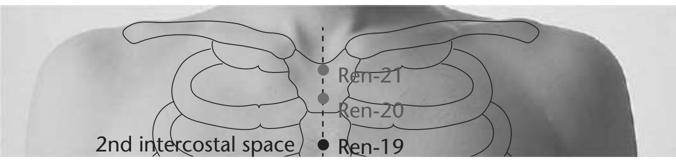
Ren-21
Ren-20
2nd intercostal space
Ren-19

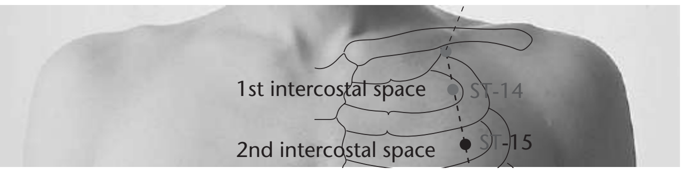
1st intercostal space
ST-14
2nd intercostal space
ST-15

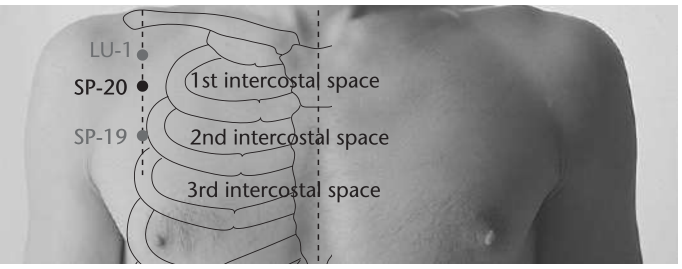
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

Location

In the 2nd intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Locate the 2nd ICS by palpating parasternally from the clavicle or by counting from the manubriosternal synchondrosis (➞ 2nd rib) (➞ 3.5), then locate **KID-24** 2 cun lateral to the anterior midline. Located on the same level (➞ 2nd ICS) are ➞ **Ren-19** (on the midline), ➞ **ST-15** (4 cun lateral to the midline) and ➞ **SP-20** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Lung Qi
- Unbinds the chest

Special features

The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to ‘excess above, deficiency below’, which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

KID-26 *Comfortable Chest* YUZHONG

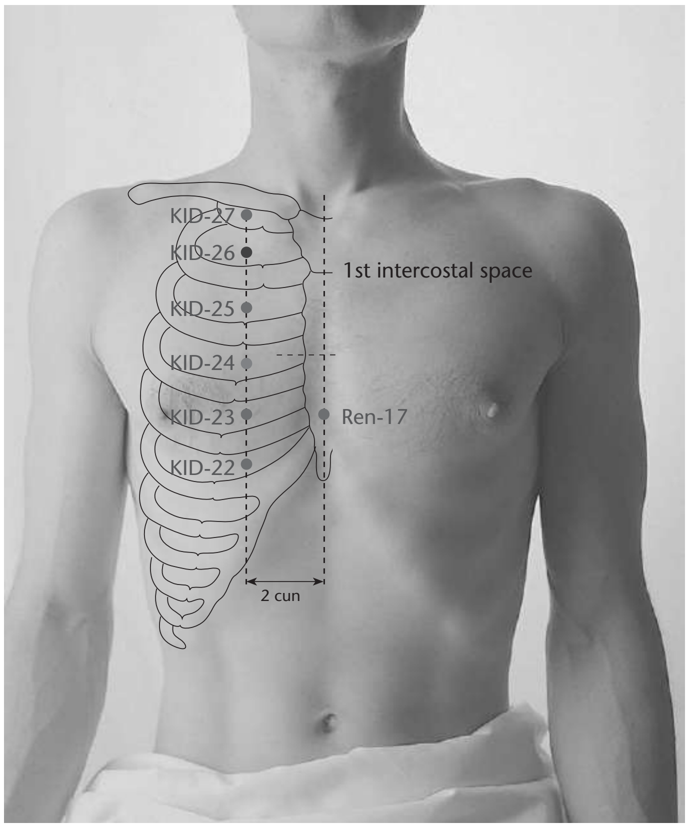
KID-27
KID-26
KID-25
KID-24
KID-23
KID-22
1st intercostal space
Ren-17
2 cun

Location

In the 1st intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Palpating parasternally from superior to inferior, the 1st rib can be palpated directly below the clavicle; in some cases, it can be completely hidden under the clavicle. Locate **KID-26** below the 1st rib in the 1st ICS, 2 cun lateral to the midline. Located on the same level (1st ICS) are → **Ren-20** (on the midline), → **ST-14** (4 cun lateral to the midline) and → **LU-1** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Lung Qi and transforms Phlegm
- Unbinds the chest
- Benefits the breasts

Special features

The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

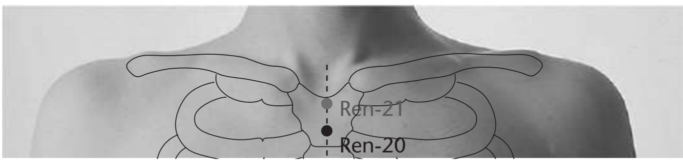
Ren-21
Ren-20

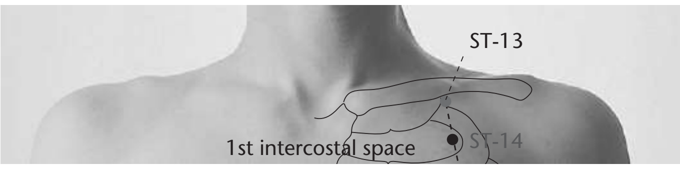
ST-13
1st intercostal space
ST-14

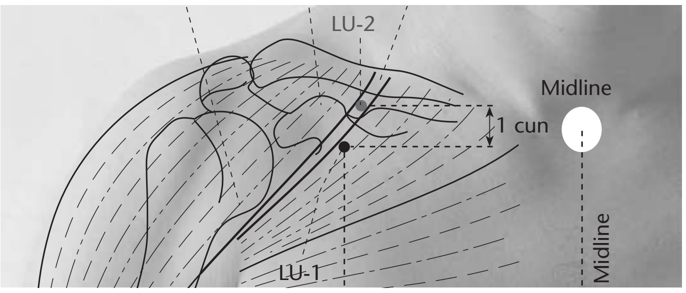
LU-2
Midline
1 cun
LU-1
Midline

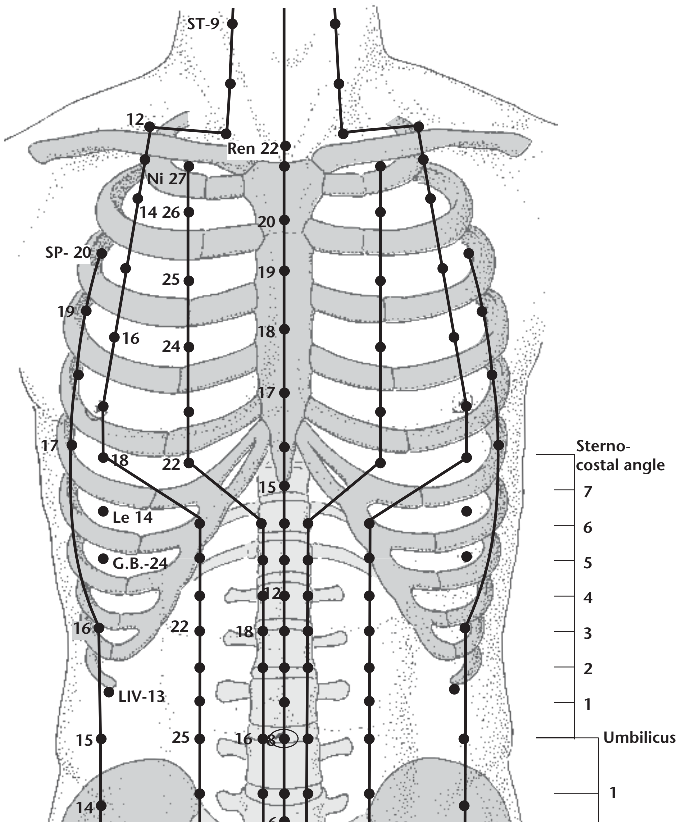
ST-9
12
Ren 22
Ni 27
14 26
20
SP- 20
25
19
19
16
24
18
17
17
18
22
15
Le 14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
Sterno-
costal angle
7
6
5
4
3
2
1
Umbilicus
1

Ch04.8-F10028.qxd 2/22/08 6:15 PM Page 347

Shu *Mansion* SHUFU KID-27

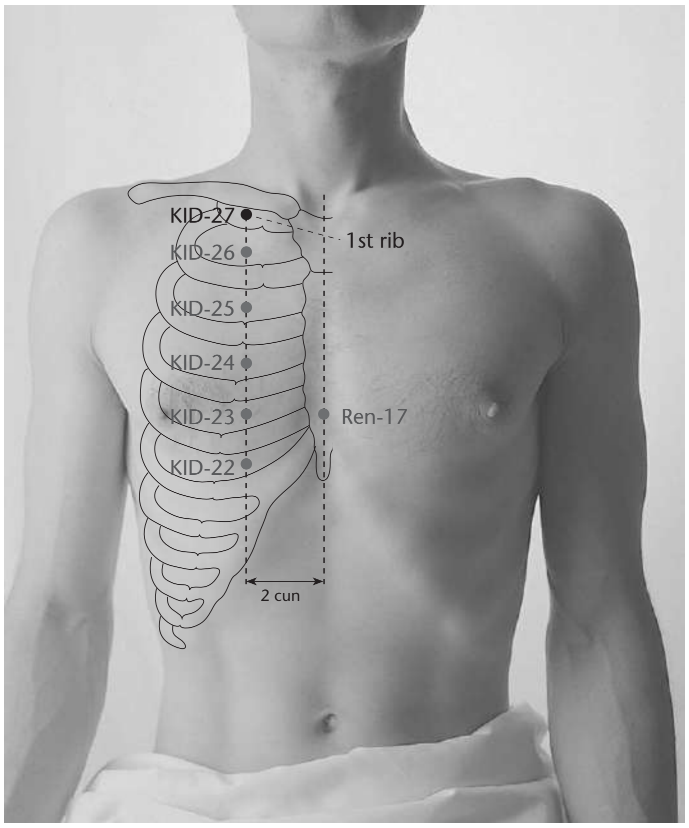
KID-27
1st rib
KID-26
KID-25
KID-24
KID-23
Ren-17
KID-22
2 cun

Location

At the lower border of the clavicle, 2 cun lateral to the anterior midline

How to find

When palpating parasternally from superior to inferior, the 1st rib can be felt directly below the clavicle; in some cases, it can be completely hidden under the clavicle. Therefore, **KID-27** will be located either on the 1st rib or, in some cases, at its lower border. ➞ **ST-13** is located lateral to the short curvature of the 1st rib, below the midpoint of the clavicle.

Located on approximately the same level are ➞ **Ren-21** (on the midline) and ➞ **LU-2** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Lung Qi and transforms Phlegm
- Unbinds the chest

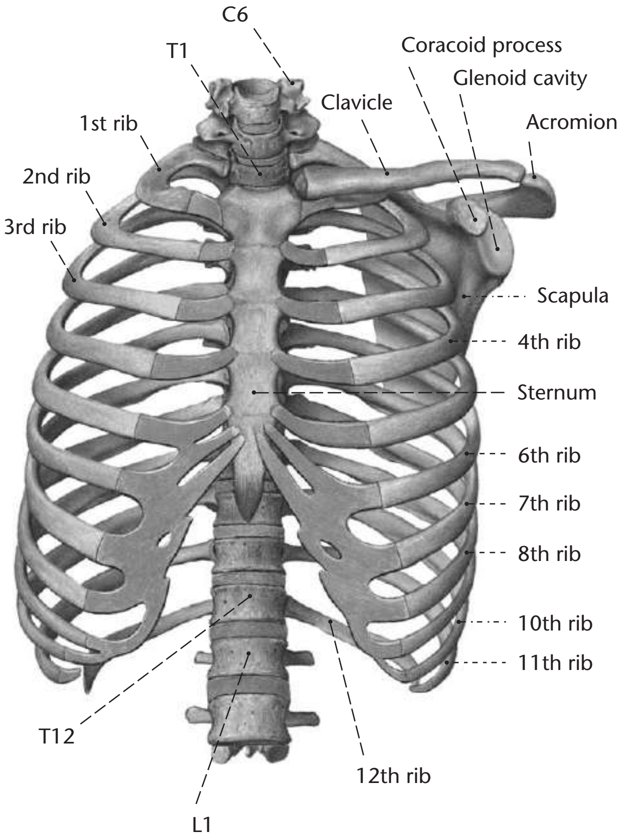
C6
T1
Coracoid process
Glenoid cavity
Clavicle
Acromion
1st rib
2nd rib
3rd rib
Scapula
4th rib
Sternum
6th rib
7th rib
8th rib
10th rib
11th rib
T12
12th rib
L1

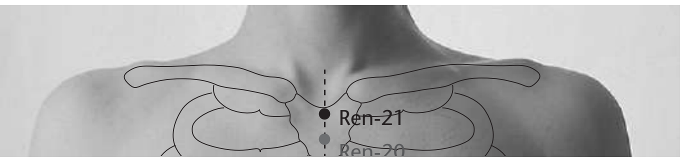
Ren-21
Ren-20

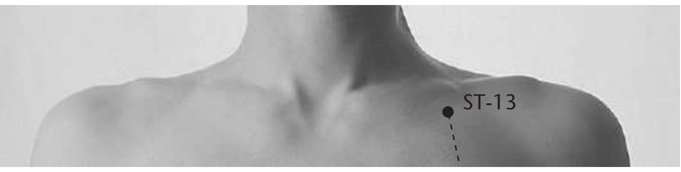

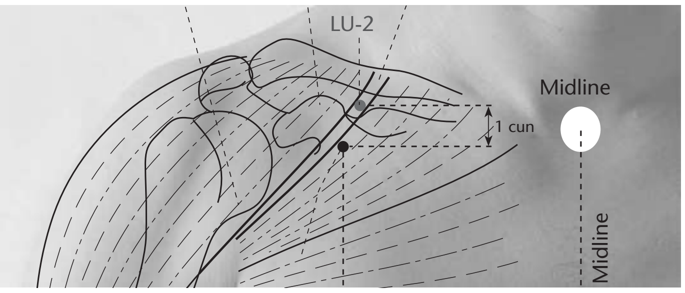
LU-2
Midline
1 cun
Midline

4.9 The Pericardium Channel System – Hand *jueyin* (*shou jue yin jing luo*)

4.9.1 Primary Pericardium Channel (*shou jue yin jing*)

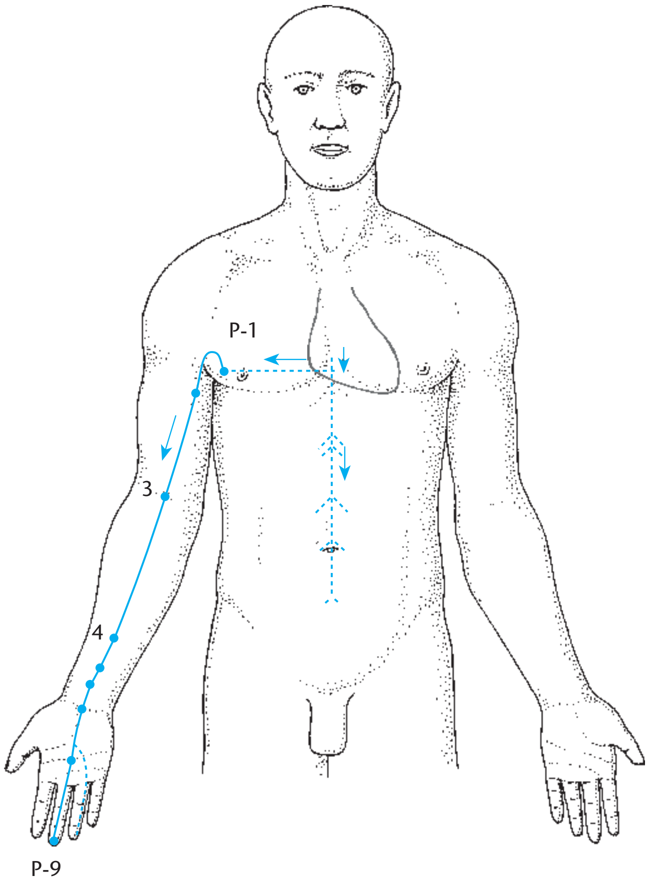

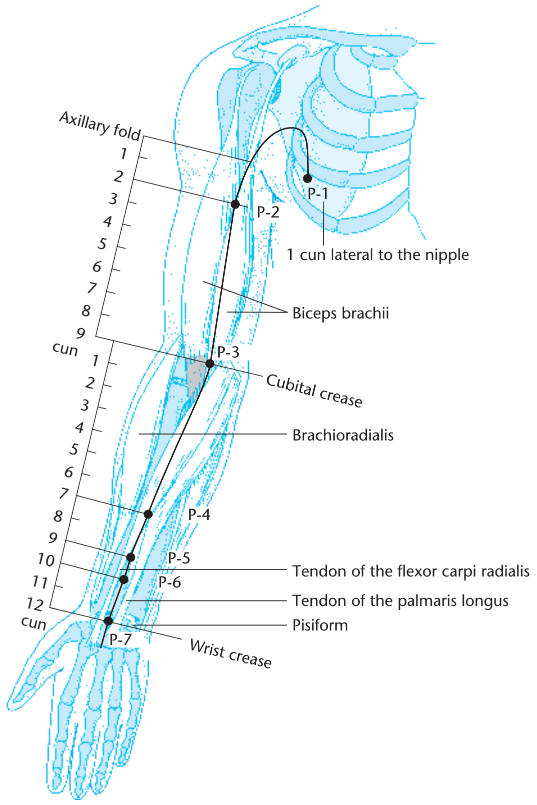
Axillary fold
1
2
3
4
5
6
7
8
9
cun
1
2
3
4
5
6
7
8
9
10
11
12
cun
P-1
P-2
1 cun lateral to the nipple
Biceps brachii
P-3
Cubital crease
Brachioradialis
P-4
P-5
P-6
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
Pisiform
P-7
Wrist crease

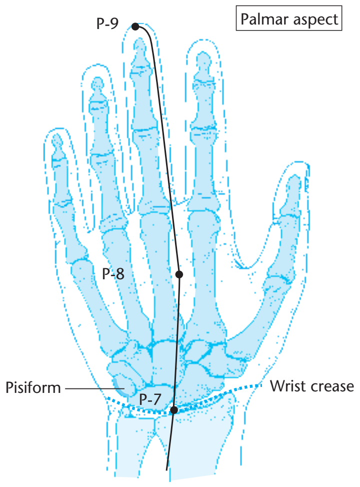
P-9
Palmar aspect
P-8
Pisiform
P-7
Wrist crease

Pathway

The Pericardium primary channel begins in the centre of the thorax on the level of → **Ren-17** (*shanzhong*), where an internal branch of the Kidney primary channel ends, having also connected with the Liver, Lung and Heart (deep Yin–Yin connection).

The Pericardium primary channel then divides into two branches:

- ➡ The **first branch** penetrates the diaphragm, descends to the abdomen and connects with the Upper, Middle and Lower Burner.
- ➡ A **further internal branch** also originates in the thoracic region, continues to the breast and emerges approximately 1 cun lateral to the nipple in the 4th intercostal space at **P-1** (*tianchi*). Beneath **P-1**, an internal branch of the Liver primary channel emerges from the Liver and joins the Pericardium primary channel (Yin axes of the third great circuit: *jueyin*).

The **external** pathway ascends to the axilla, follows the medial aspect of the upper arm between the Lung and Heart primary channels, passes the elbow and continues along the forearm between the tendons of the palmaris longus and flexor carpi ulnaris muscles, ending at the tip of the middle finger.

Another **internal** branch originates in the centre of the palm at **P-8** (*laogong*), runs along the ring finger and connects at its tip with the Triple Burner primary channel (hand Yin–Yang connection of the third circuit).

Clinical importance (→ 1.2)

Exterior *(biao)* signs and symptoms: Occipital stiffness, leg cramps, red complexion, eye pain, swellings in the subaxillary region, hypertonic musculature with limited range of motion of the arm and elbow, Heat sensations in the palms

Interior (*li*) **or *zangfu*-Organ signs and symptoms**: Delirious speech, confusion, irritability, fullness and distension in the thorax and lateral costal region, aphasia, palpitations

Connections of the Pericardium primary channel (→ 1.2)

Connections with other channels

Triple Burner primary channel (*shou shao yang jing*)

Connection: Hand Yin–Yang pairing of the third great circuit
Location: P-8 → T.B.-1 (on the hand)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Liver primary channel (*zu jue yin jing*)

Connection: Six-channel pairing of the third great circuit (Yin axes): *jue yin*
Location: LIV → P. An internal branch of the Liver primary channel emerges at the Liver, penetrates the diaphragm and connects with the Pericardium primary channel beneath **P-1** (*tianchi*).
Circulation: Non-circadian (**not** according to the Organ clock)
Importance: Above–below relationship

Kidney primary channel (*zu shao yin jing*)

Connection: Deep Yin–Yin connection
Location: KID → P (on the thorax). A branch of the internal pathway of the Kidney primary channel travels from the Kidney to the Liver, penetrates the diaphragm and spreads in the Lung (*fei*). From the Lung, an internal branch runs to the Heart (*xin*), where it joins the Pericardium primary channel (deep Yin–Yin connection) and further reaches → **Ren-17** (*shanzhong*).
Circulation: Circadian (according to the Organ clock)
Importance: The Pericardium primary channel receives Nutritive Qi (*ying qi*) from the Kidney primary channel (first circulation of the *ying qi* → 1.1.4).

Connections with *zangfu*-Organ systems

Pericardium (*xin bao*), Triple Burner (*san jiao*)

4.9.2 The Pericardium Divergent Channel (*shou jue yin jing bie*)

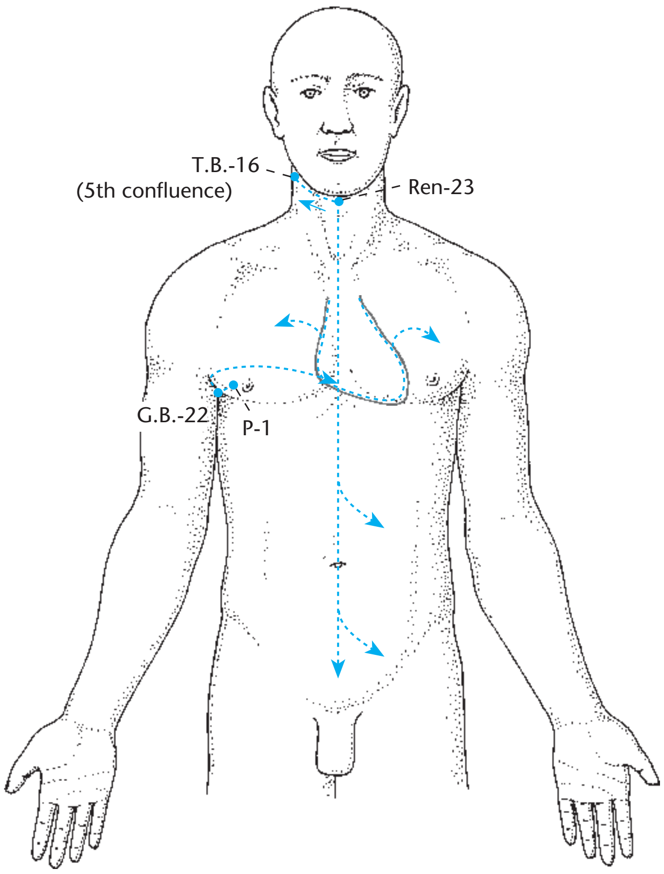

Pathway

The Pericardium divergent channel separates from the Pericardium primary channel near **P-1** (*tianchi*), 1 cun lateral to the nipple on the level of the 4th intercostal space. It

- ➡ runs in a lateral direction to → **G.B.-22** (*yuanye*)
- ➡ penetrates the thorax, continues to the Heart (*xin*) and divides into **two branches:**
  - **one branch** connects with the Upper, Middle and Lower Burner
  - the **other branch** ascends to the neck to → **Ren-23** (*lianquan*) and continues to the lateral aspect of the neck, emerges at → **T.B.-16** (*tianyou*) on the posterior border of the sternocleidomastoid muscle on the level of the mandibular angle, connecting with the Triple Burner primary channel and the Triple Burner divergent channel to form one of the 6 *he*-confluences (here: P/T.B. as 5th Confluence → 1.3)

Clinical importance

- Strengthens the relationship between the Pericardium and the Triple Burner (*zangfu*-Organ systems). Points on the Pericardium primary channel can therefore be used for disorders of the Triple Burner; vice versa, points on the Triple Burner primary channel can treat disorders of the Pericardium.

- In contrast to the Pericardium primary channel, the Pericardium divergent channel reaches the occiput and the neck. This explains why several points on the P channel can be used for disorders of the neck and ears, chronic throat pain, etc.
- By reaching the thorax, breast and hypochondrium, the Pericardium divergent channel supports the relationship between the Pericardium primary channel and these regions.

4.9.3 The Pericardium Sinew Channel (*shou jue yin jing jin*)

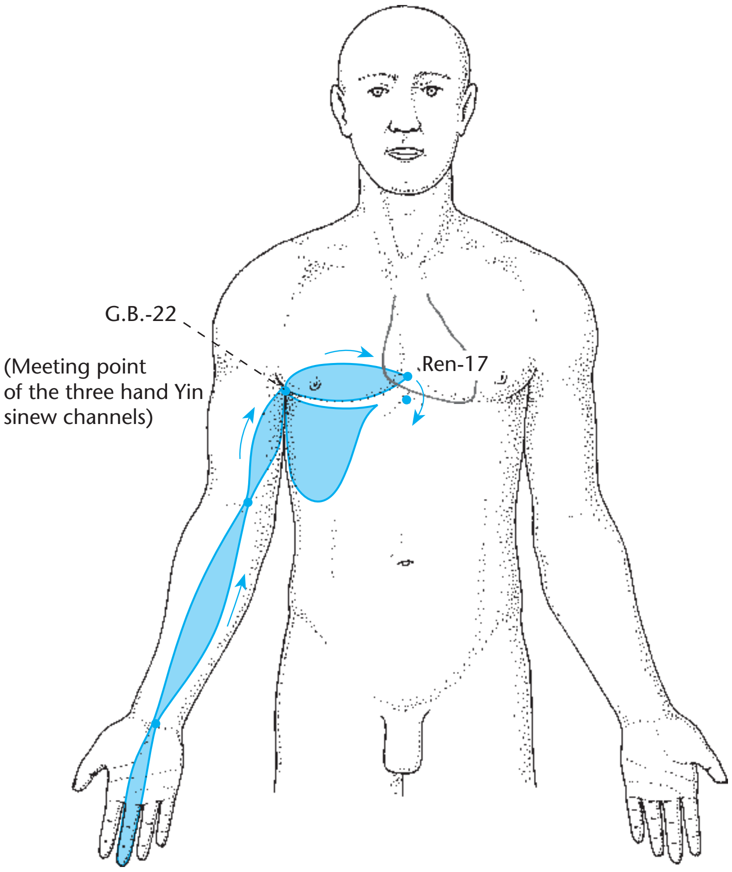

Pathway

The Pericardium sinew channel begins at the tip of the middle finger

➡ runs along the anterior aspect of the middle finger and along the 3rd metacarpal bone

➡ spreads in the centre of the palm and meets the Lung sinew channel

➡ binds (*jie*) in the centre of the wrist

➡ ascends the anterior aspect of the forearm as well as the anteromedial aspect of the upper arm to the axilla, binding (*jie*) at **P-3** (*quze*) on the elbow and at → **G.B.-22** (*yuanye*), where it meets the other hand Yin sinew channels.

At → **G.B.-22**, the channel separates into two branches:

- **one branch** disperses on the anterior, lateral and posterior aspects of the costal region
- **the other branch** penetrates the thorax below the axilla and reaches the midthoracic region on the level of → **Ren-17** (*shanzhong*), from where it spreads in the thoracic cavity to bind (*jie*) at the diaphragm.

Clinical importance

Pathology: Stiffness, pain, spasms and distending sensations along the pathway of the Pericardium sinew channel. Thoracic pain and tightness, *xi fen* syndrome (one of the five *ji*-syndromes, an 'accumulation syndrome') due to Lung Qi stagnation with accumulation of Phlegm and Heat manifesting with symptoms such as masses in the right hypochondriac region, possibly accompanied by thoracic or back pain, haemoptysis, chills and fever, shortness of breath and dyspnoea, etc.

Indication: Mainly for pain, spasms and distending sensations along the pathway of the sinew channel, limited range of motion along the medial aspect of the arm and elbow, swellings and pain in the axilla and the breast.

4.9.4 The Pericardium *luo*-Connecting Vessel System (*shou jue yin luo mai*)

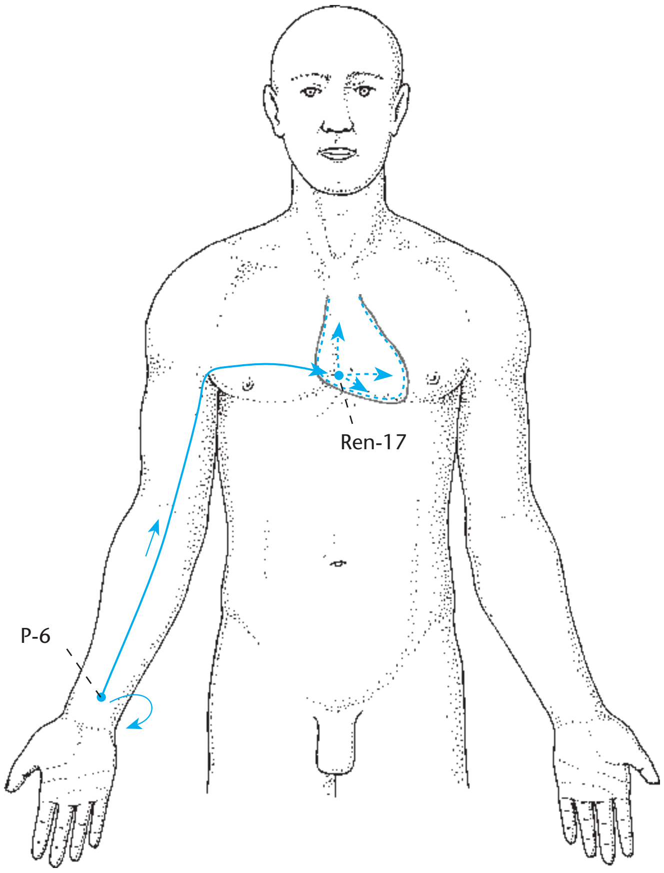

Pathway

The *luo*-connecting Pericardium vessel system separates from the Pericardium primary channel at its associated *luo*-connecting point **P-6** (*neiguan*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Triple Burner primary channel; according to some schools of thought (for example, Ngyen Van Nghi → Appendix), they travel as a **transverse** Pericardium *luo*-connecting vessel to the *yuan*-source point → **T.B.-4** (*yangchi*).
- **A longitudinal division** follows the Pericardium primary channel along the anterior aspect of the arm, crosses → **Ren-17** (*shanzhong*) on the midline on the level of the 4th intercostal space, penetrates the thorax and terminates at the Pericardium (*xin bao*) and the Heart (*xin*).

Clinical importance (→ 8.1.2)

Pathology

Excess (*shi*): Heart pain, angina pain, thoracic pain
Deficiency (*xu*): Restlessness, irritability

4.9.5 Cutaneous Region (*jue yin pi bu*)

See description and figures → 1.6.

4.9.6 Points of the Primary Pericardium Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): P-7** (*daling*) 🟦■
- ***Luo*-connecting point (→ 8.1.2.): P-6** (*neiguan*) 🟦🟦
- ***Xi*-cleft point (→ 8.1.3): P-4** (*ximen*) 🟦■
- **Associated Back-*shu* point (→ 8.1.4): BL-14** (*jueyinshu*)
- **Associated Front-*mu* point (→ 8.1.5): Ren-17** (*shanzhong*) 🟦🟦
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood), tonification point: **P-9** (*zhongchong*) 🟦■
  ying-spring point (Fire), *ben* point (Five Phases): **P-8** (*laogong*) 🟦■
  shu-stream point (Earth), sedation point: **P-7** (*daling*) 🟦■
  jing-river point (Metal): **P-5** (*jianshi*) 🟦■
  he-sea point (Water): **P-3** (*quze*) 🟦■
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8) of the *yin wei mai*: P-6** (*neiguan*) 🟦🟦
- **Lower *he*-sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the LIV, T.B., G.B.* channels: **P-1** (*tianchi*)
  – of other channels with the Pericardium channel: –
- **Gao Wu command point (→ 8.1.11) for the thoracic region: P-6** (*neiguan*) 🟦🟦
- **Window of Heaven point (→ 8.1.12): P-1** (*tianchi*)
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): –**
- **Sun Si Miao Ghost point (→ 8.1.15): P-7** (*daling*) 🟦■, **P-8** (*laogong*) 🟦■
- **Other functional points:**
  – important point for eliminating Phlegm misting the orifices: **P-5** (*jianshi*) 🟦■
  – important point for calming the *shen*: **P-6** (*neiguan*) 🟦🟦

Points according to region

- **Local points (→ 8.2.1):** hand – **P-7** (*daling*) 🟦■
- **Adjacent points (→ 8.2.1): –**
- **Distal points (→ 8.2.1):** tongue – **P-8** (*laogong*) 🟦■; Lung – **P-6** (*neiguan*) 🟦🟦; Heart – **P-4** (*ximen*) 🟦■; **P-6** (*neiguan*) 🟦🟦; Spleen – **P-6** (*neiguan*) 🟦🟦; thorax and epigastrium – **P-6** (*neiguan*) 🟦🟦; foot – **P-7** (*daling*) 🟦■

Specific points according to the channel pathway (in numeric order)

- **P-1** (*tianchi*) 🟦■: *jiaohui*-meeting point with the LIV, T.B., G.B.* channels (→ 8.1.10); Window of Heaven point (→ 8.1.12)
- **P-3** (*quze*) 🟦■: lower *he*-sea point (Water) (→ 8.1.6)
- **P-4** (*ximen*) 🟦■: *xi*-cleft point (→ 8.1.3); distal point for the Heart (→ 8.2.1)
- **P-5** (*jianshi*) 🟦■: *jing*-river point (Metal) (→ 8.1.6); important point for eliminating Phlegm misting the orifices
- **P-6** (*neiguan*) 🟦🟦: *luo*-connecting point (→ 8.1.2); Gao Wu command point (→ 8.1.11) for the thorax; Opening point (→ 8.1.8) of the *yin wei mai*; distal point for the Heart, Spleen, thorax and epigastrium (→ 8.2.1); important point for calming the *shen*
- **P-7** (*daling*) 🟦■: *yuan*-source point (→ 8.1.1); *shu*-stream point (Earth) (→ 8.1.6); sedation point; Sun Si Miao Ghost point (→ 8.1.15); local point for the hand (→ 8.2.1); distal point for the foot (→ 8.2.1)
- **P-8** (*laogong*) 🟦■: *ying*-spring point (Fire) (→ 8.1.6); *ben* point (Five Phases); Sun Si Miao Ghost point (→ 8.1.15); distal point for the tongue (→ 8.2.1)
- **P-9** (*zhongchong*) 🟦■: *jing*-well point (Wood) (→ 8.1.6); tonification point

---

* Mentioned by only some authors

Heavenly Pool TIANCHI **P-1**

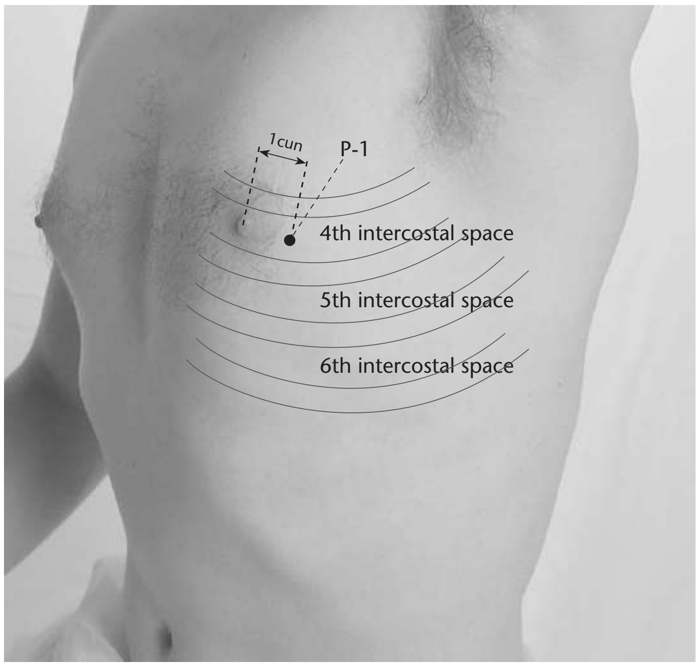
1cun
P-1
4th intercostal space
5th intercostal space
6th intercostal space

Location

In the 4th intercostal space (ICS), 1 cun lateral to the nipple.

How to find

Quick method, especially in male patients: The 4th ICS is located on the same level as the nipple. Locate **P-1** 1 cun lateral to the nipple in the 4th ICS (note: ascending course of the ICS towards lateral). **Or:** For better orientation in the intercostal region, first, locate the manubriosternal synchondrosis (→ 3.4.2.). This is on the level of the costal cartilage of the 2nd rib, followed by the 2nd ICS below it. Count down to ICS to the 4th ICS and measure 5 cun in a lateral direction. There, locate **P-1**.

Located on the same level (4th ICS) are → **Ren-17** (on the midline), → **KID-23/ST-17/SP-18** (2/4/6 cun lateral to the midline), as well as → **G.B.-22** and → **G.B.-23**.

Needling

0.5–1 cun transversely (subcutaneously) along the course of the ICS. Especially in women, transverse needling only, avoid puncturing the mamillary tissue. Caution: Pneumothorax.

Actions/Indications

- Descends counterflow Qi
- Eliminates excess (especially Heat and Phlegm) from the thorax
- Promotes lactation and eliminates local Qi stagnation

Special features

Meeting point with the LIV and T.B. channels, also with the G.B. channel according to some authors. Window of Heaven point, entry point.

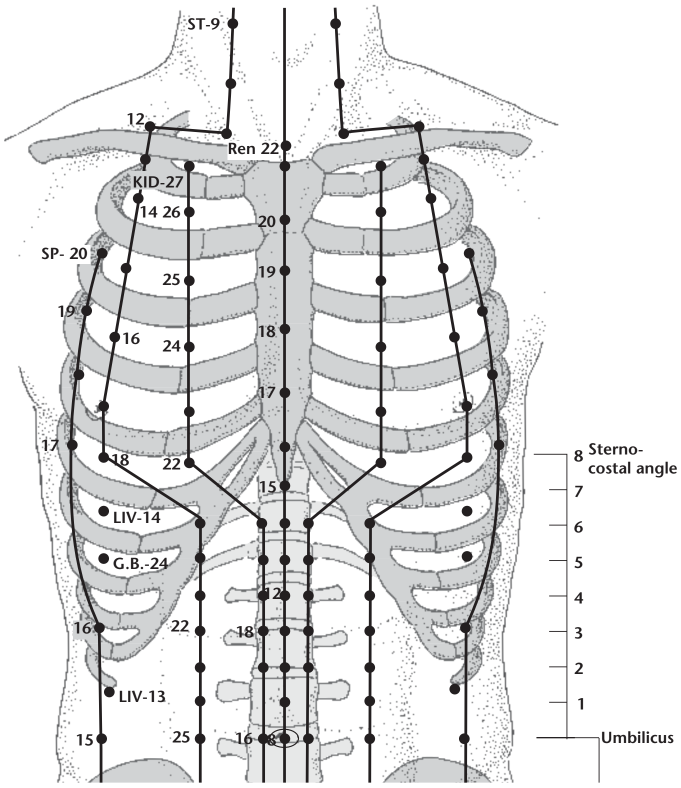
ST-9
12
Ren 22
KID-27
14 26
20
SP- 20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
16
22
12
18
LIV-13
15
25
16
8 Sterno-costal angle
7
6
5
4
3
2
1
Umbilicus

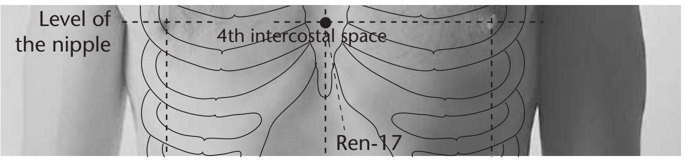
Level of the nipple
4th intercostal space
Ren-17

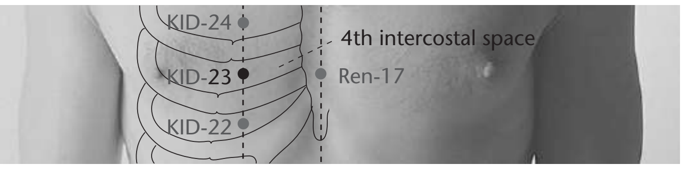
KID-24
KID-23
KID-22
4th intercostal space
Ren-17

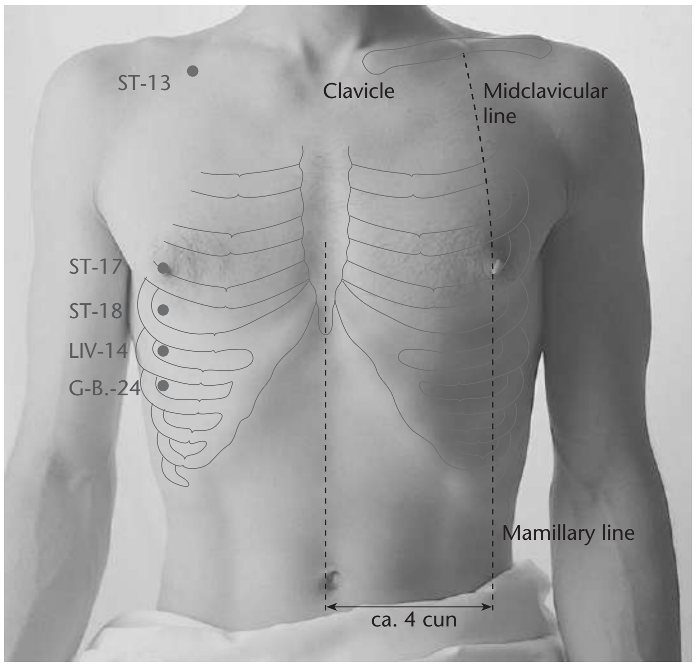
ST-13
Clavicle
Midclavicular line
ST-17
ST-18
LIV-14
G-B.-24
Mamillary line
ca. 4 cun

P-2 *Heavenly Spring* TIANQUAN

Location

Between the two heads of the biceps brachii muscle, 2 cun inferior to the axillary fold.

How to find

Locate the biceps brachii muscle through slight external rotation; in patients with less muscle definition, ask the patient to flex their elbow against resistance. **P-2** is located between the two heads of the biceps brachii muscle, 2 cun inferior to the upper end of the anterior axillary fold (➞ 2.2).

➞ **LU-3** is located radial to **P-2,** on the lateral aspect of the biceps brachii muscle and 1 cun more distally.

Needling

1–1.5 cun obliquely in a proximal or lateral direction.

Actions/Indications

- Unbinds the chest.
- Nourishes and invigorates the Blood
- Calms the *shen*

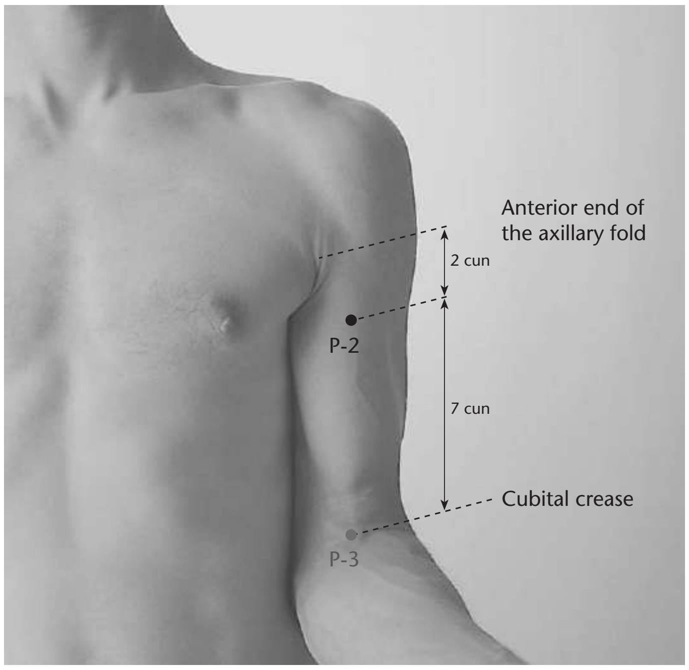
Anterior end of the axillary fold
2 cun
P-2
7 cun
Cubital crease
P-3

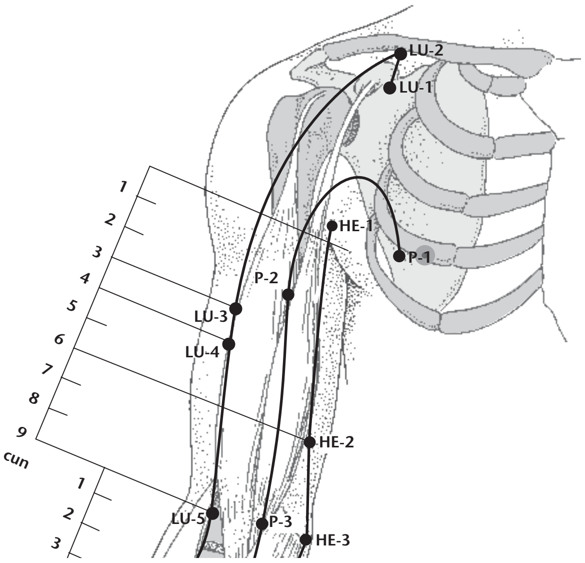
LU-2
LU-1
HE-1
P-1
P-2
LU-3
LU-4
HE-2
LU-5
P-3
HE-3
1
2
3
4
5
6
7
8
9
cun
1
2
3

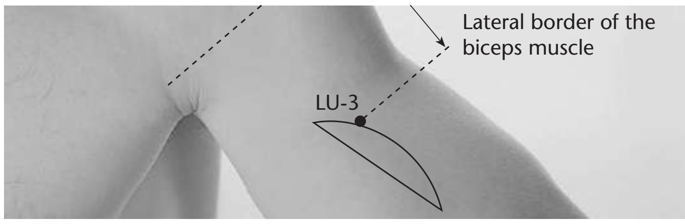
Lateral border of the biceps muscle
LU-3

Marsh at the Crook QUZE **P-3**

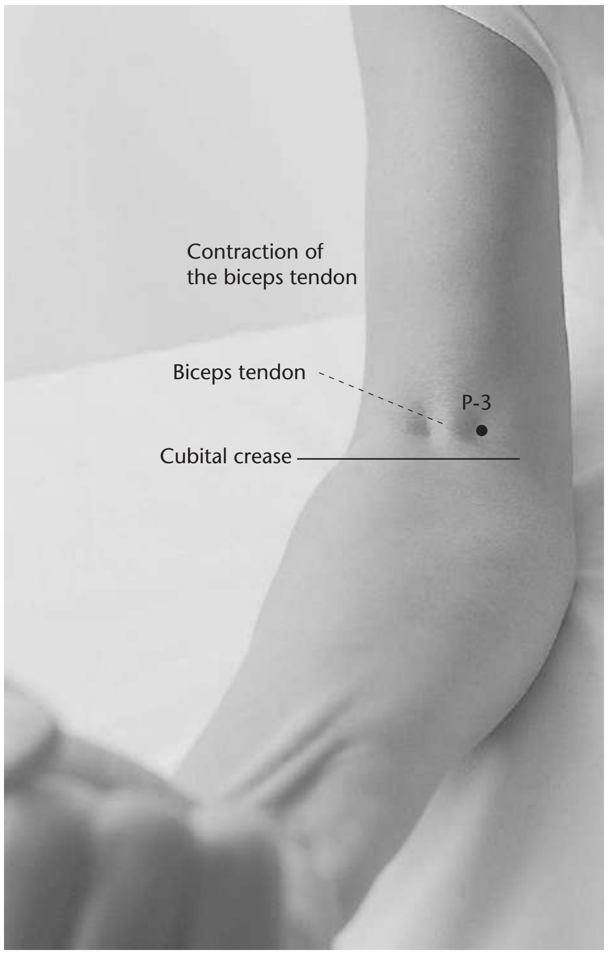
Contraction of
the biceps tendon
Biceps tendon
P-3
Cubital crease

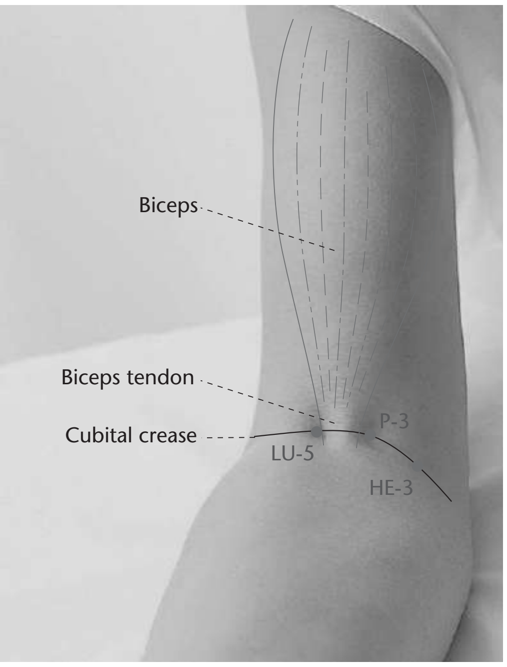
Biceps
Biceps tendon
P-3
Cubital crease
LU-5
HE-3

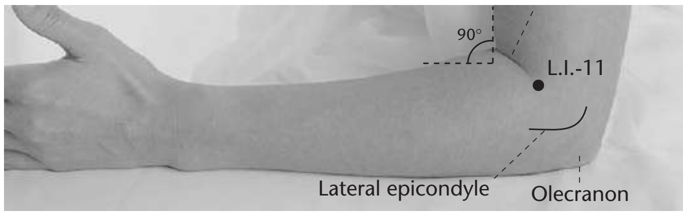
90°
L.I.-11
Lateral epicondyle
Olecranon

Location

On the cubital crease, on the ulnar side of the biceps tendon, between the tendon and the brachial artery.

How to find

Place the patient's arm in a relaxed supinated position. This point is best located with the elbow bent and the biceps flexed in order to make the tendon and elbow crease more clearly visible and palpable. **P-3** is located on the cubital crease, on the ulnar aspect of the tendon.

Also located in the area of the cubital crease are → **LU-5** (on the radial side of the tendon), → **L.I.-11** (between the radial end of the crease and the lateral epicondyle of the humerus) and → **HE-3** (with the elbow fully flexed, at the ulnar end of the cubital crease).

Needling

0.5–1 cun vertically or prick to cause bleeding of the vein. Caution: Brachial artery/vein. Palpate for the pulse before needling.

Actions/Indications

- Clears Heat, drains Fire
- Harmonises the Stomach and Intestines, stops vomiting
- Opens the channel, alleviates pain

Special features

He-sea point, Water point. Important local point and important point to drain Heat. For Heat conditions, apply bloodletting.

P-4 Xi *Cleft Gate* XIMEN

Location

5 cun proximal to the anterior space of the wrist joint (most distal wrist crease), between the tendons of the palmaris longus and flexor carpi radialis muscles.

How to find

Spreading hands technique (→ 2.3.3): Place the little fingers in the centre of the anterior wrist joint space (→ 3.3.3, → **P-7**) and the cubital crease respectively (→ **P-3**). From the midpoint of this distance, measure 1 cun in a distal direction. On this level, locate **P-4** between the two tendons, which become more pronounced when the fingertips are pressed together. If only one tendon is visible, this will be the tendon of the flexor carpi radialis muscle. **P-4** is located on its ulnar aspect.

Needling

0.5–1 cun vertically or 1–1.5 cun obliquely in a proximal direction.

Actions/Indications

- Tonifies the Blood, dispels Blood stasis
- Calms the *shen*
- Cools the Blood
- Opens the channel

Special features

Xi-cleft point. Major point for acute, painful Blood stasis in the region of the thorax and the Heart

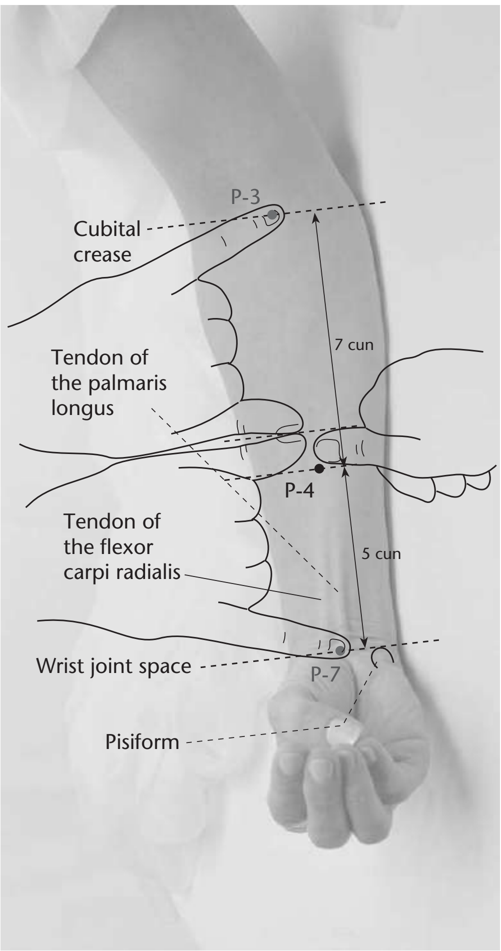
P-3
Cubital crease
7 cun
Tendon of the palmaris longus
P-4
Tendon of the flexor carpi radialis
5 cun
Wrist joint space
P-7
Pisiform

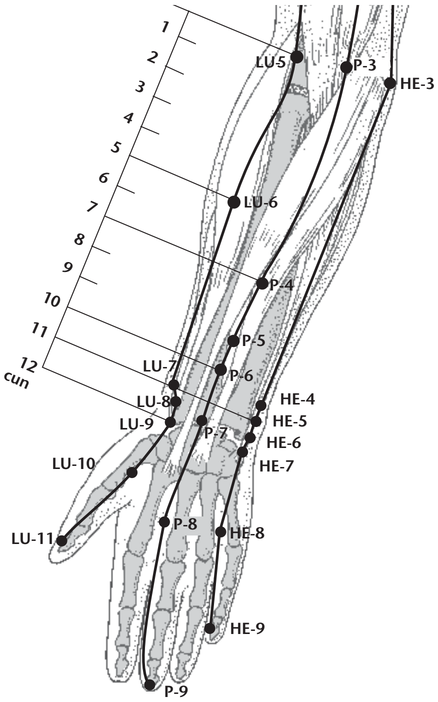
1
2
3
4
5
6
7
8
9
10
11
12 cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
LU-8
HE-4
LU-9
P-7
HE-5
HE-6
LU-10
HE-7
P-8
HE-8
LU-11
HE-9
P-9

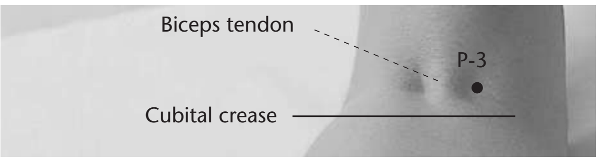
Biceps tendon
P-3
Cubital crease

Intermediary Messenger JIANSHI P-5

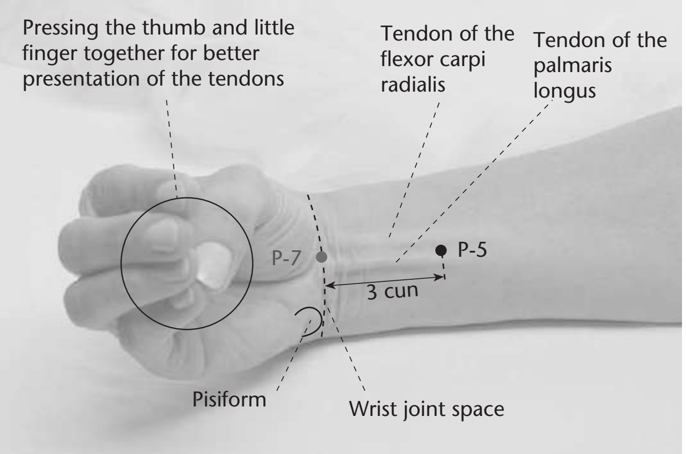
Pressing the thumb and little finger together for better presentation of the tendons
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
P-7
P-5
3 cun
Pisiform
Wrist joint space

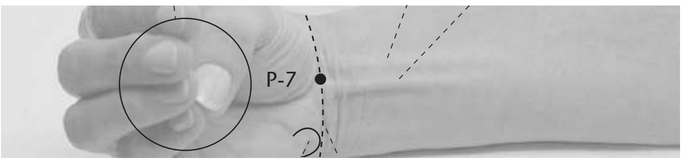

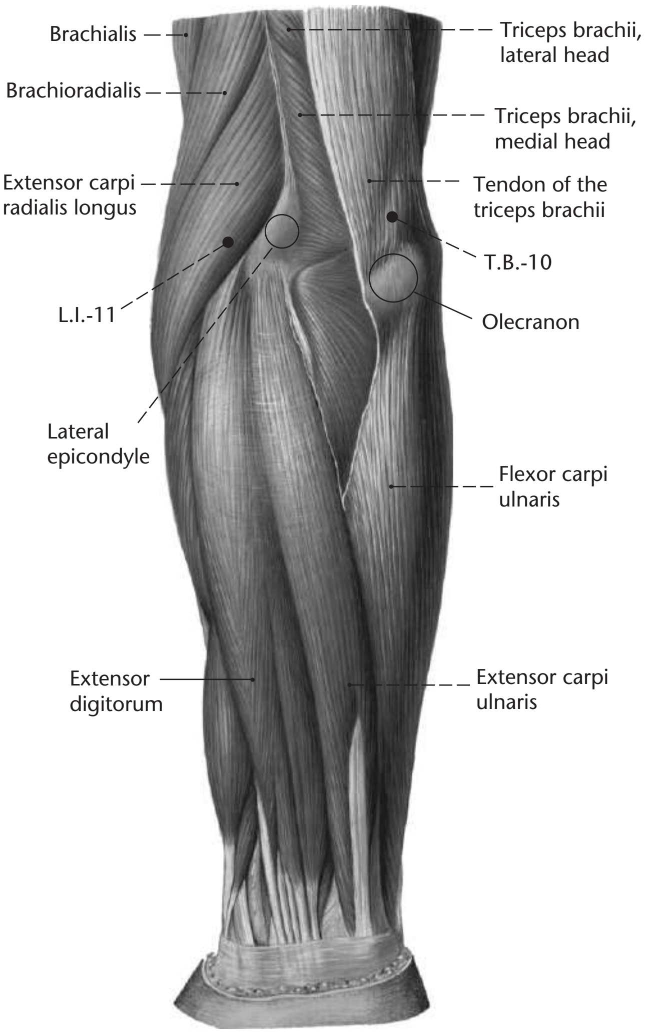
Brachialis
Brachioradialis
Extensor carpi radialis longus
L.I.-11
Lateral epicondyle
Extensor digitorum
Triceps brachii, lateral head
Triceps brachii, medial head
Tendon of the triceps brachii
T.B.-10
Olecranon
Flexor carpi ulnaris
Extensor carpi ulnaris

Location

3 cun proximal to the anterior wrist joint space (‘most distal wrist crease’), between the tendons of the palmaris longus and flexor carpi radialis muscles.

How to find

As the location of the wrist crease varies, the wrist joint space is a more reliable reference point (→ 3.3.3). It can easily be palpated by moving the patient’s hand in a relaxed way. From the centre of the joint space (→ **P-7**), measure 3 cun in a proximal direction and locate **P-5** between the two tendons, which become more pronounced when the fingertips are pressed together. If only one tendon is visible, this will be the tendon of the flexor carpi radialis muscle. **P-5** is located at its ulnar aspect.

Needling

0.5–1 cun vertically or obliquely in a proximal direction. Caution: The median nerve here is only covered by the radial head of the flexor digitorum superficialis. Vertical needling should not exceed 0.5 cun; no strong needle stimulation.

Actions/Indications

- Calms the *shen* and the Heart
- Harmonises the Middle Burner
- Clears Heart Fire
- Unbinds the chest
- Transforms invisible Phlegm

Special features

Jing-river point, Metal point. Important point for transforming Phlegm blocking the orifices.

P-6 *Inner Gate* NEIGUAN

Location

2 cun proximal to the anterior wrist joint space (‘most distal wrist crease’), between the tendons of the palmaris longus and flexor carpi radialis muscles.

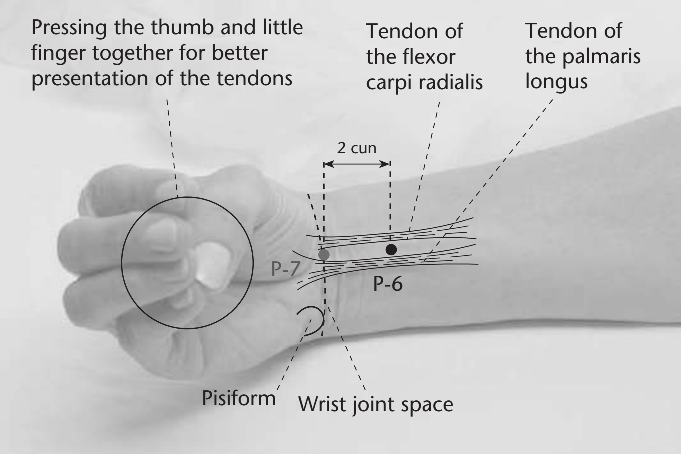
Pressing the thumb and little finger together for better presentation of the tendons
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
2 cun
P-7
P-6
Pisiform
Wrist joint space

How to find

As the location of the wrist crease varies, the wrist joint space is a more reliable reference point (→ 3.3.3). It can easily be palpated by moving the patient’s hand in a relaxed way. From the centre of the wrist joint space (→ **P-7**), measure 2 cun in a proximal direction and locate **P-6** between the two tendons, which become more pronounced when the fingertips are pressed together. If only one tendon is visible, this will be the tendon of the flexor carpi radialis muscle. **P-6** is located on its ulnar aspect.

Or quick method: From the centre of the wrist joint space, glide in a proximal direction between the two tendons. The skin will begin to form a bulge in front of the sliding finger, which will come to a halt at **P-6.**

→ **T.B.-5** (2 cun proximal to the dorsal wrist joint space) is located opposite **P-6**.

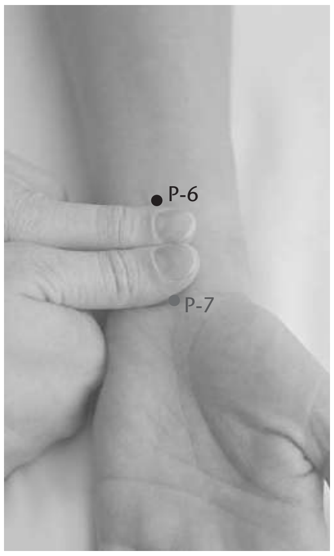

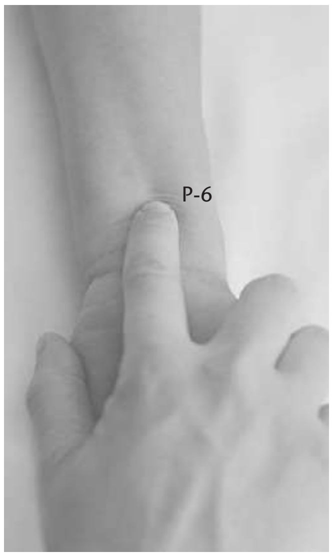

Needling

0.5–1 cun vertically or obliquely in a proximal or lateral direction. Caution: The median nerve is located directly below this point.

Actions/Indications

- Calms the *shen*
- Unbinds the chest
- Harmonises the Stomach
- Opens the channel

Special features

Luo-connecting point, opening (master) point of the *yin wei mai*, Gao Wu command point for the thorax. Main point for nausea and vomiting. Very important point.

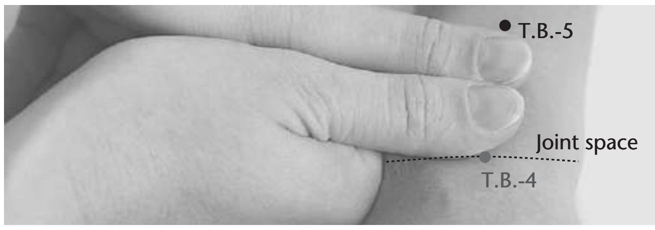
T.B.-5
Joint space
T.B.-4

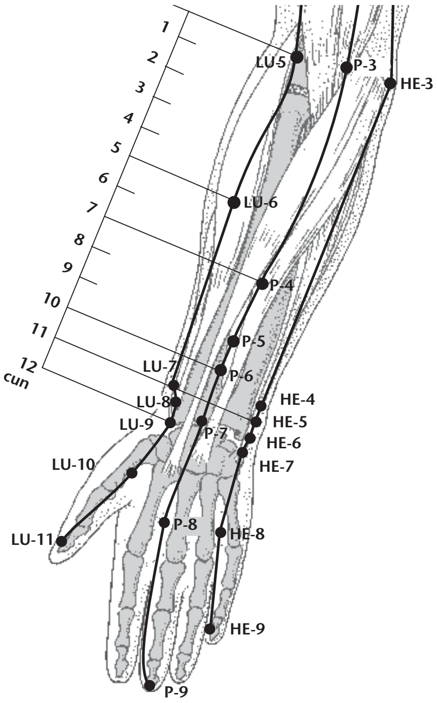
1
2
3
4
5
6
7
8
9
10
11
12
cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9

Great Mound DALING P-7

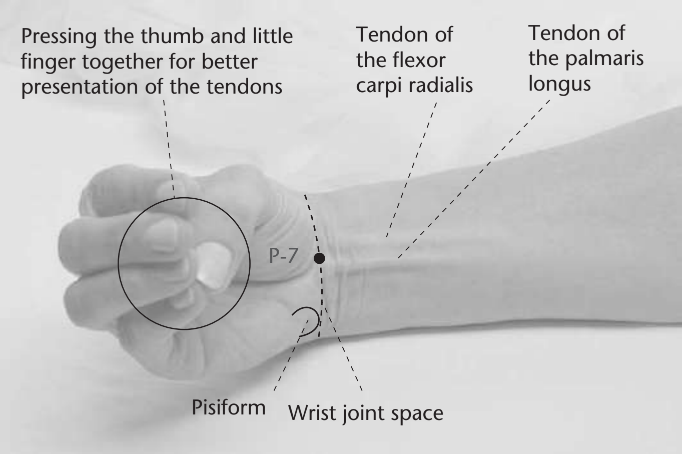
Pressing the thumb and little finger together for better presentation of the tendons
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
P-7
Pisiform
Wrist joint space

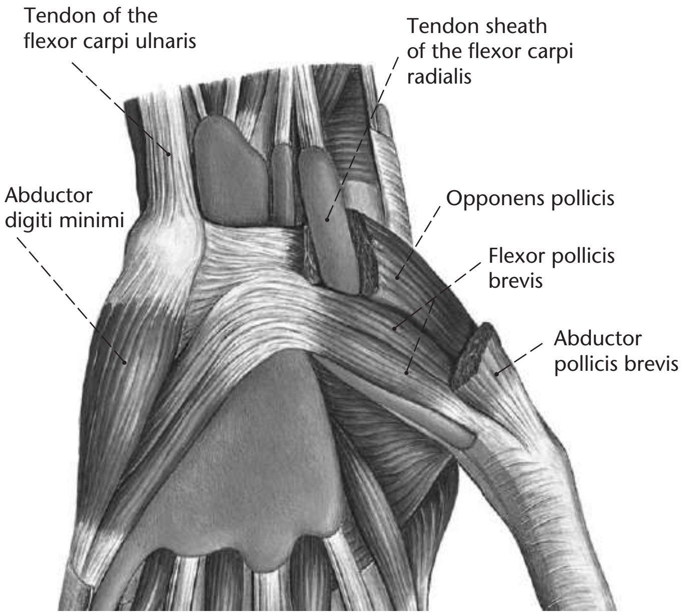
Tendon of the flexor carpi ulnaris
Tendon sheath of the flexor carpi radialis
Abductor digiti minimi
Opponens pollicis
Flexor pollicis brevis
Abductor pollicis brevis

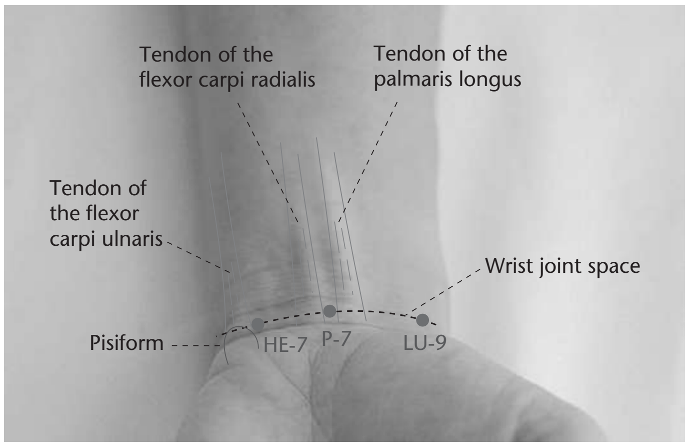
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
Tendon of the flexor carpi ulnaris
Wrist joint space
Pisiform
HE-7
P-7
LU-9

Location

On the anterior aspect of the wrist joint space ('most distal wrist crease'), between the tendons of the palmaris longus and flexor carpi radialis muscles.

How to find

As the location of the wrist crease varies, the wrist joint space is a more reliable reference point (→ 3.3.3). By moving the patient's hand in a relaxed way, it can easily be palpated. Locate **P-7** on this level between the two tendons, which become more pronounced when the fingertips are pressed together. If only one tendon is visible, this will be the tendon of the flexor carpi radialis muscle. **P-7** is located on its ulnar aspect.

Located on the same level are → **HE-7** (on the ulnar aspect of the wrist joint, radial to the tendon of the flexor carpi ulnaris muscle) and → **LU-9** (on the radial aspect of the wrist joint, lateral to the radial artery.

Needling

Vertically 0.3–0.5 cun, for carpal tunnel syndrome also 0.5–1 cun obliquely in a lateral direction along the tunnel. Caution: The median nerve is located directly under this point, thus needling can cause a significant electric sensation; in this case, no further needle manipulation to avoid possible injury.

Actions/Indications

- Clears Heat from the Heart, calms the *shen*
- Harmonises the Stomach and Intestines
- Cools Heart Fire affecting the Bladder
- Unbinds the chest
- Cools the Blood
- Opens the channel, alleviates pain

Special features

Yuan-source point, *shu*-stream point, Earth point, sedation point, Sun Si Miao Ghost point. Important point for calming, main point for carpal tunnel syndrome.

P-8 *Palace of Toil* LAOGONG

Location

In the centre of the palm, between the 2nd and 3rd metacarpal bones, slightly closer to the 3rd metacarpal bone.

How to find

When making a loose fist, the tip of the middle finger will rest on **P-8,** between the 2nd and 3rd metacarpal bones, slightly closer to the 3rd metacarpal bone.

→ **HE-8** is located more radially, between the 4th and 5th metacarpal bones.

Needling

Vertically 0.3–0.5 cun. Thinner needles should be used as needling this point can be painful.

Actions/Indications

- Clears Heart Fire and calms the *shen*
- Harmonises the Middle Burner
- Clears the *ying*-nutritive level and cools the Blood

Special features

Ying-spring point, Fire point, *ben* point (Five Phases), Sun Si Miao Ghost point, exit point.

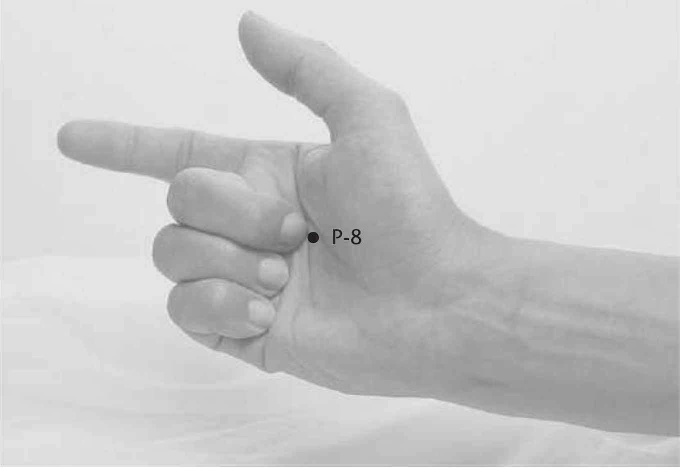

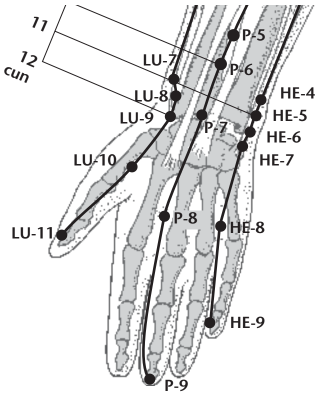

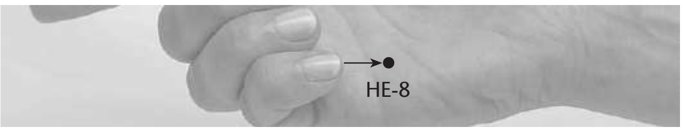

Middle Rushing ZHONGCHONG **P-9**

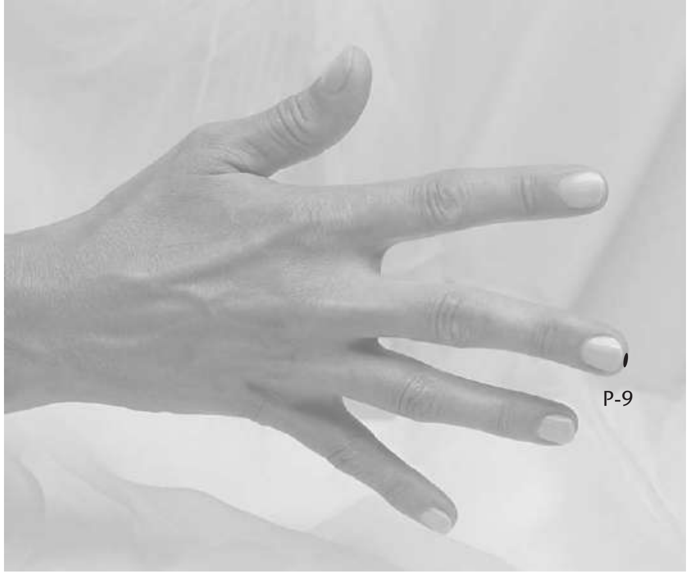

Location

On the most distal point of the middle finger.

How to find

Find the most distal point on the tip of the middle finger and there locate **P-9**.

Needling

Vertically 0.1–0.2 cun or prick to bleed. Caution: Painful point.

Actions/Indications

- Clears Heat, especially from the Heart and Pericardium
- Strengthens collapsed Yang
- Expels internal Wind

Special features

Jing-well point, Wood point, tonification point.

4.10 The Triple Burner Channel System – Hand *shaoyang* (*shou shao yang jing luo*)

4.10.1 Triple Burner Primary Channel (*shou shao yang jing*)

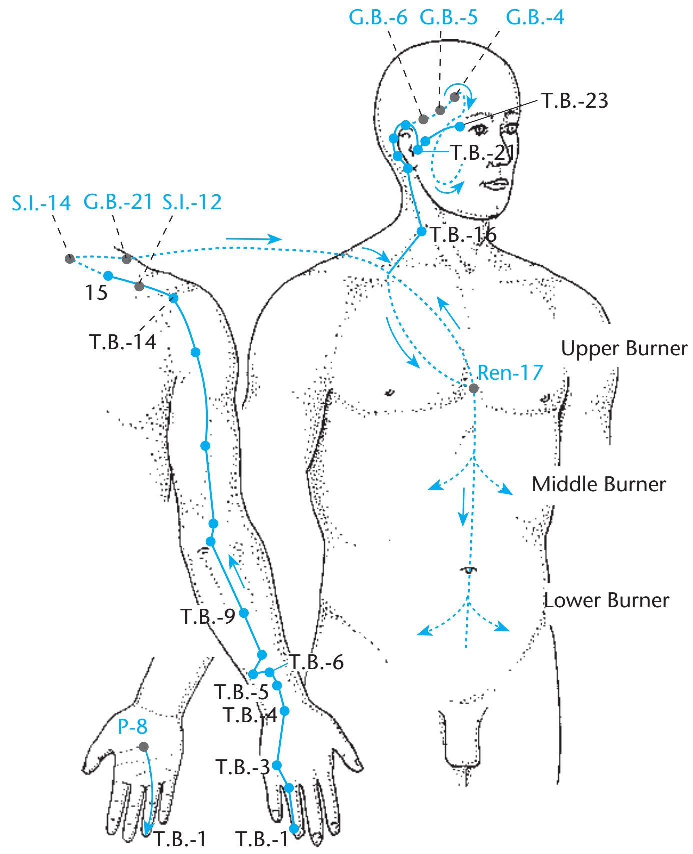

Pathway

The **external pathway** of the Triple Burner primary channel begins on the ring finger, at the ulnar corner of the nail at **T.B.-1** (*guanchong*). **T.B.-1** is reached by a branch of the Pericardium primary channel that separates on the palm of the hand at → **P-8** (*laogong*) (hand Yin–Yang connection of the third great circuit). The channel runs along the ulnar side of the ring finger

- continues between the 4th and 5th metacarpal bones
- travels along the posterior aspect of the forearm between the radius and the ulna towards the lateral aspect of the arm
- passes the olecranon and continues along the upper arm to the shoulder
- traverses the shoulder, intersects with → **S.I.-12** (*bingfeng*) and meets the other Yang primary channels at → **Du-14** (*dazhu*)
- runs to the anterior aspect of the body, crossing → **G.B.-21** (*jianjing*) and → **ST-12** (*quepen*), the latter in the supraclavicular fossa.

At **ST-12,** the channel enters the body to begin its **internal** section. It meets the Pericardium (*xin bao*) at → **Ren-17** (*shanzhong*), penetrates the diaphragm and enters the abdomen, connecting with the Upper, Middle and Lower Burners.

From the Lower Burner, an **internal branch** (according to some authors, the T.B. channel) descends to the popliteal fossa to → **BL-39** (*weiyang*), the lower *he*-sea point of the Triple Burner.

A further **inner branch** originates in the thorax at → **Ren-17** (*shanzhong*), emerges at the supraclavicular fossa and ascends to the posterior border of the sternocleidomastoid muscle to **T.B.-16**, follows via **T.B.-17, T.B.-18, T.B.-19** the temporal bone along its border behind the ear and reaches **T.B.-20** (*jiasun*) above the apex of the ear. It crosses the Gall Bladder primary channel at **G.B.-6** (*xuanli*), **G.B.-5** (*xuanlu*) and **G.B.-4** (*hanyan*), according to Deadman et al 1998) also at **G.B.-11** (*touqiaoyin*) and **G.B.-14** (*yangbai*), descends the cheek to the lower jaw and ascends again to intersect with → **S.I.-18** (*quanliao*) on the lower border of the zygomatic arch and continues to the infraorbital region.

From **T.B.-17** (*yifeng*), a branch penetrates the ear, emerges to the **Exterior** anterior to the tragus, intersects with → **S.I.-19** (*tinggong*), continues past **T.B.-21** (*ermen*) and **T.B.-22** (*erheliao*), meets → **G.B.-3** (*shangguan*), crosses its own cheek branch and ends at **T.B.-23** (*sizhukong*) on the frontozygomatic suture.

From **T.B.-23**, or, according to some authors from **T.B.-22**, a small branch runs to → **G.B.-1** (*tongziliao*) at the lateral canthus of the eye (Yang axes of the third great circuit: *shao yang*).

Clinical importance (→ 1.2)

Exterior (***biao***) signs and symptoms: Neck pain, cheek pain, red and painful eyes, deafness, pain behind the ears, pain on the posterior aspect of the shoulder and the upper arm

Interior (***li***) or *zangfu*-Organ signs and symptoms: Abdominal fullness and distension, difficult urination, skin oedema and skin swellings, enuresis

Connections of the Triple Burner primary channel (→ 1.2)

Connections with other channels

Pericardium primary channel (*shou jue yin jing*)

Connection: Hand Yin–Yang pairing of the third great circuit
Location: P-8 → **T.B.-1** (on the hand).
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Gall Bladder primary channel (*zu shao yang jing*)

Connection: Paired according to the six-channel theory (hand–foot pairing): *shao yang* (Yang axes of the third great circuit)
Location: T.B.-23 (or **T.B.-22**) → **G.B.-1** (on the head).
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with *zangfu*-Organ systems
Pericardium (*xin bao*), Triple Burner (*san jiao*)

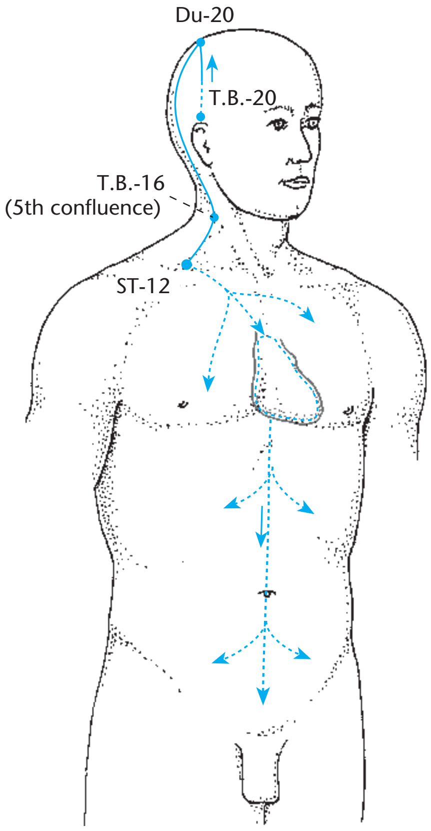

4.10.2 The Triple Burner Divergent Channel (*shou shao yang jing bie*)

Pathway

The Triple Burner divergent channel separates from the Triple Burner primary channel near **T.B.-20** (*jioasun*)

- ➡ ascends to → **Du-20** (*baihui*)
- ➡ descends along the posterior aspect of the ear to the neck to **T.B.-16** (*tianyou*) on the posterior border of the sternocleidomastoid muscle, where it connects with the Triple Burner primary channel and the Pericardium primary channel to form one of the six *he*-confluences (here: P/T.B. as fifth confluence → 1.3)
- ➡ The Triple Burner divergent channel further descends to → **ST-12** (*quepen*) in the supraclavicular fossa, penetrates the thorax, disperses in the Pericardium (*xin bao*) and reaches the Upper, Middle and Lower Burners.

Clinical importance

- Strengthens the relationship between the Upper, Middle and Lower Burners. T.B. points can be used for disorders affecting these areas and Organs.
- The Triple Burner divergent channel reaches the vertex at → **Du-20**. Therefore, T.B. points can be used for headaches, dizziness, etc, but also to raise the Yang Qi for prolapse, etc.
- The Triple Burner divergent channel spreads in the thorax, so that T.B. points can be used to support the treatment of thoracic problems such as pain, coughing etc.

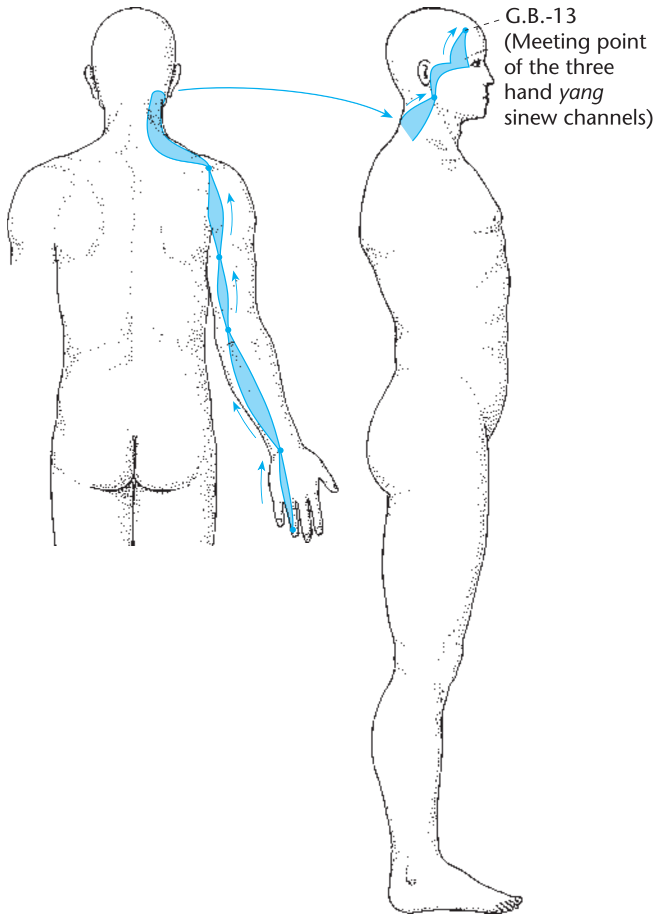

4.10.3 The Triple Burner Sinew Channel (*shou shao yang jing jin*)

Pathway

The Triple Burner sinew channel begins on the ring finger, at the ulnar corner of the nail at **T.B.-1** (*guanchong*)

- ➡ travels between the 4th and 5th metacarpal bones to the wrist
- ➡ binds (*jie*) at **T.B.-4** (*yangchi*)
- ➡ continues along the posterior aspect of the forearm between the Small Intestine and Large Intestine sinew channels, binds (*jie*) at the olecranon, ascends the posterior aspect of the upper arm to bind (*jie*) at the posterior corner of the acromion, traverses the posterior aspect of the shoulder and the lateral aspect of the occiput and there meets the Small Intestine sinew channel. It then reaches the mandibular angle, where it divides into **two branches**:
- ➡ **One branch** enters internally and penetrates the root of the tongue

- ➡ **The other branch** ascends in front of the ear, binds (*jie*) at the outer canthus of the eye, crosses the temple and terminates at ➞ **G.B.-13** (*benshen*) in the frontoparietal region. There, it also meets the other hand Yang sinew channels.

Clinical importance

Pathology: Stiffness, swellings and distending sensations along the course of the sinew channel. Curling or contraction of the tongue.

Indication: Predominantly for pain, stiffness, spasms, tension and distending sensations along the course of the channel, including the lateral aspects of the head and the body as well as the lateral sections of the extremities. T.B. points can also be used for muscle tension caused by emotional disorders. Owing to the connection of the sinew channel with the tongue, some T.B. points are also indicated for motor disorders of the tongue.

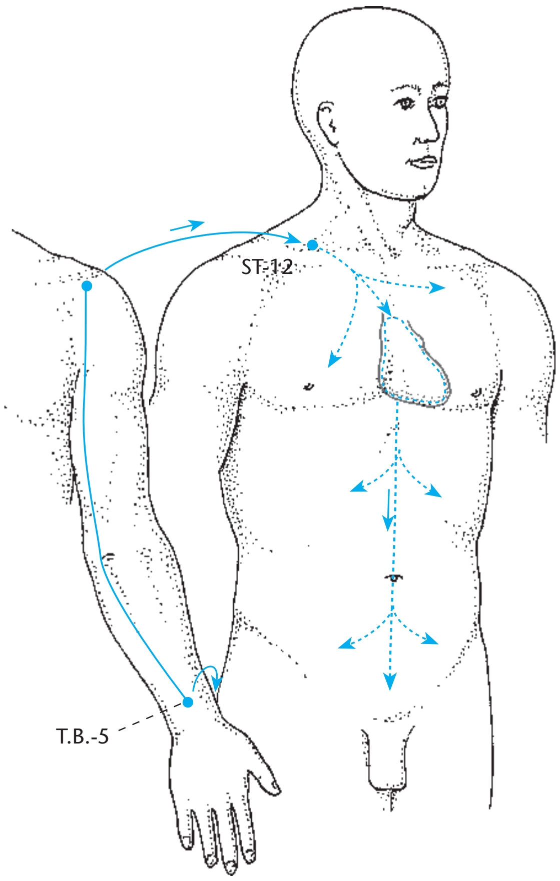

4.10.4 The Triple Burner *luo*-Connecting Vessel System (*shou shao yang luo mai*)

Pathway

The Triple Burner *luo*-connecting vessel system separates from the Triple Burner primary channel at its associated *luo*-connecting point **T.B.-5** (*waiguan*) (➞ 8.1.2). It forms a three-dimensional reticulate network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* ➞ 1.5) within the surrounding tissue.

- ➡ Horizontal divisions run to the Interiorly–Exteriorly paired Pericardium primary channel; according to some schools of thought (for example Ngyen Van Nghi ➞ Appendix), they travel as a **transverse** Triple Burner *luo*-connecting vessel to the *yuan*-source point **P-7** (*daling*).
- ➡ **A longitudinal division** follows the posterior aspect of the arm, crosses the posterior shoulder region, continues to the front of the body and, according to some authors (for example, Solinas et al 1998), reaches ➞ **ST-12** (*quepen*) in the supraclavicular fossa, penetrates the thorax and reaches the Pericardium primary channel, both dispersing in the Upper, Middle and Lower Burners.

Clinical importance

Pathology

Excess (*shi*): Tension of the elbow joint

Deficiency (*xu*): Muscle weakness in the arm, difficulty flexing the elbow

4.10.5 Cutaneous Region (*shao yang pi bu*)

See description and figures ➞ 1.6.

4.10.6 Points of the Triple Burner Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (➞ 8.1.1): T.B.-4** (*yangchi*)
- ***Luo*-connecting point (➞ 8.1.2.): T.B.-5** (*waiguan*) 🟦🟦
- ***Xi*-cleft point (➞ 8.1.3): T.B.-7** (*huizong*)
- **Associated Back-*shu* point (➞ 8.1.4): BL-22** (*sanjiaoshu*)
- **Associated Front-*mu* point (➞ 8.1.5): Ren-5** (*shimen*)
- **Five *shu*-transporting points (➞ 8.1.6):**
  jing-well point (Metal): **T.B.-1** (*guanchong*)
  ying-spring point (Water): **T.B.-2** (*yemen*)
  shu-stream point (Wood), tonification point: **T.B.-3** (*zhongzhu*)
  jing-river point (Fire), *ben point* (Five Phases): **T.B.-6** (*zhigou*)
  he-sea point (Earth), sedation point: **T.B.-10** (*tianjing*)
- ***Hui*-meeting point (➞ 8.1.7): –**
- **Opening point (➞ 8.1.8) of the *yin wei mai*: T.B.-5** (*waiguan*) 🟦🟦
- **Lower *xiahe*-sea points (➞ 8.1.9): –**
- ***Jiaohui*-meeting points (➞ 8.1.10):**
  - with the *yang wei mai* and the G.B. channel: **T.B.-13*** (*naohui*)
  - with the *yang wei mai* and the G.B. channel: **T.B.-15** (*tianliao*) 🟦■
  - with the G.B. channel: **T.B.-17** (*yifeng*) 🟦■
  - with the S.I.* and G.B. channel: **T.B.-20** (*jiaosun*)
  - with the S.I. and G.B. channels: **T.B.-22** (*erheliao*)

---

* Mentioned by only some authors

  - – of other channels with the T.B. channel: **S.I.-12, BL-11, (S.I.-18, S.I.-19*), Du-14, Ren-12, Ren-17, ST-12, G.B.-21, G.B.-14, G.B.-11, G.B.-6, G.B.-5, G.B.-4, G.B.-3, G.B.-1, (S.I.-9*)**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): T.B.-16** (*tianyou*)
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): –**
- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points: –**

Points according to region

- **Local points (→ 8.2.1):** ears – **T.B.-21** (*ermen*) 🟦■; shoulder – **T.B.-14** (*jianliao*) 🟦■, **T.B.-15** (*tianliao*) 🟦■
- **Adjacent points (→ 8.2.1):** temporal aspect of the head – **T.B.-17** (*yifeng*) 🟦■; eyes – **T.B.-23** (*sizhukong*) 🟦■; ears – **T.B.-17** (*yifeng*) 🟦■; jaw – **T.B.-17** (*yifeng*) 🟦■; shoulder – **T.B.-13** (*naohui*), **T.B.-15** (*tianliao*) 🟦■
- **Distal points (→ 8.2.1):** temporal aspect of the head – **T.B.-3** (*zhongzhu*) 🟦■, **T.B.-5** (*waiguan*) 🟦🟦; ears – especially **T.B.-3** (*zhongzhu*), **T.B.-5** (*waiguan*) 🟦🟦; hypochondrium – **T.B.-5** (*waiguan*) 🟦🟦, **T.B.-6** (*zhigou*); cervical spine – **T.B.-8** (*sanyangluo*); cervical spine and shoulder – **T.B.-5** (*waiguan*) 🟦🟦; shoulder – **T.B.-1** (*guanchong*)

Specific points according to the channel pathway (in numeric order)

- **T.B.-1** (*guanchong*): *jing*-well point (Metal) (→ 8.1.6); distal point for the shoulder (→ 8.2.1)
- **T.B.-2** (*yemen*): *ying*-spring point (Water) (→ 8.1.6)
- **T.B.-3** (*zhongzhu*): *shu*-stream point (Wood) (→ 8.1.6); tonification point; distal point for the temporal region and the ears (→ 8.2.1); local point for the fingers (numbness and pain) (→ 8.2.1)
- **T.B.-4** (*yangchi*): *yuan*-source point (→ 8.1.1); local point for the hand (→ 8.2.1)
- **T.B.-5** (*waiguan*) 🟦🟦: *luo*-connecting point (→ 8.1.2); Opening point (→ 8.1.8) of the *yang wei mai*; distal point for the temporal region, the ears, the cervical spine, the shoulder, the elbow and the hypochondrium (→ 8.2.1); regional point for the hand and fingers (→ 8.2.1)
- **T.B.-6** (*zhigou*): *jing*-river point (Fire) (→ 8.1.6); *ben* point (Five Phases); distal point for the hypochondrium (→ 8.2.1)
- **T.B.-7** (*huizong*): *xi*-cleft point (→ 8.1.3)
- **T.B.-8** (*sanyangliao*): distal point for the cervical spine (→ 8.2.1)
- **T.B.-10** (*tianjing*): *he*-sea point (Earth) (→ 8.1.6); sedation point; local point for the elbow (→ 8.2.1)
- **T.B.-13** (*naohui*): *jiaohui*-meeting point (→ 8.1.10) of the *yang wei mai** and the G.B. channel*; adjacent point for the shoulder (→ 8.2.1)
- **T.B.-14** (*jianliao*) 🟦■: local point for the shoulder (→ 8.2.1)
- **T.B.-15** (*tianliao*) 🟦■: *jiaohui*-meeting point of the *yang wei mai* and the G.B. channel; local point for the shoulder (→ 8.2.1)
- **T.B.-16** (*tianyou*): Window of Heaven point (→ 8.1.12)
- **T.B.-17** (*yifeng*) 🟦■: *jiaohui*-meeting point of the G.B. channel; adjacent point for the temporal region, the ears and the jaw (→ 8.2.1)
- **T.B.-20** (*jiaosun*): *jiaohui*-meeting point of the S.I.* and G.B. channels
- **T.B.-21** (*ermen*) 🟦■: local point for the ears (→ 8.2.1)
- **T.B.-22** (*erheliao*): *jiaohui*-meeting point with the S.I. and G.B. channels
- **T.B.-23** (*sizhukong*) 🟦■: adjacent point for the eyes (→ 8.2.1)

---

*Mentioned only by some authors

Rushing Pass GUANCHONG T.B.-1

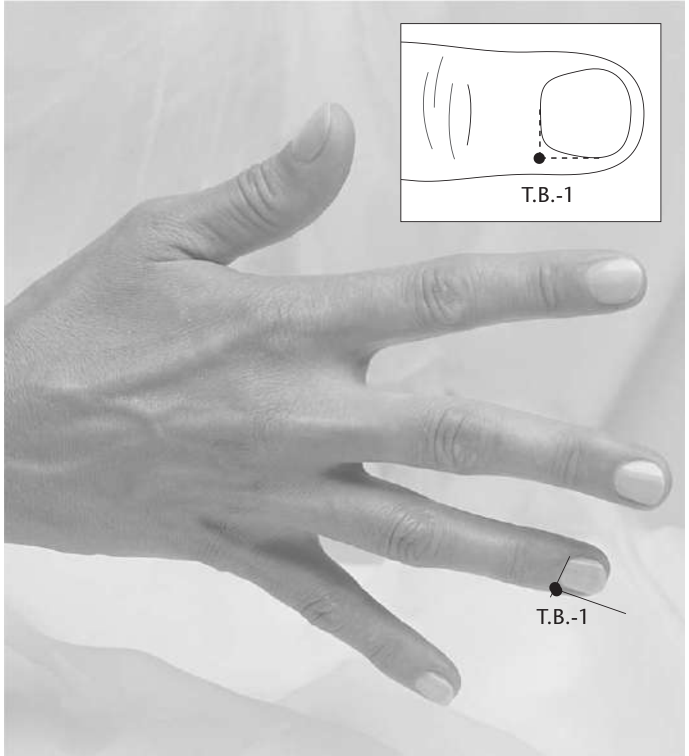

Location

On the ring finger, 0.1 cun from the ulnar corner of the nail.

How to find

This point is located at the junction of two tangents bordering the ulnar and proximal margins of the nail, approximately 0.1 cun from the corner of the actual nail.

Needling

Approximately 0.2 cun vertically or obliquely in a proximal direction or prick to bleed. Caution: Avoid needling too close to the nail; painful point.

Actions/Indications

- Clears Heat from the Upper Burner
- Benefits the ears and tongue
- Opens the channel and alleviates pain

Special features

Jing-well point, Metal point, entry point. Especially indicated for acute disorders.

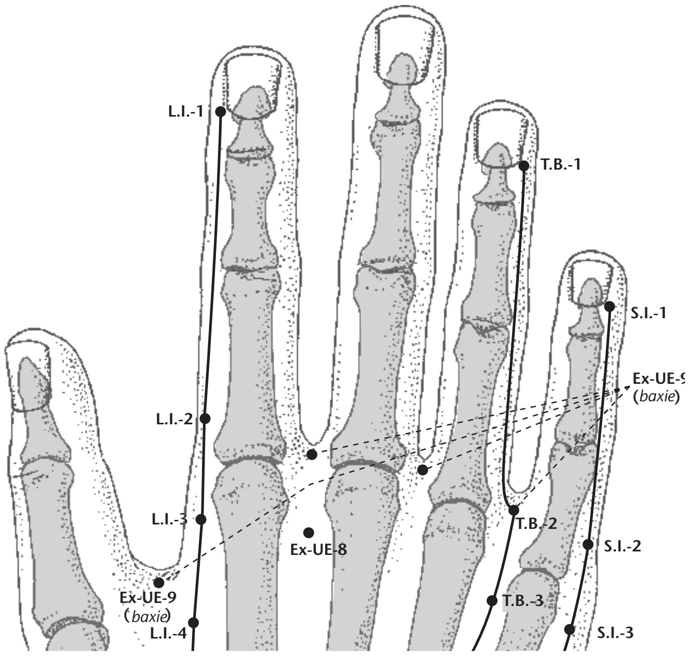

T.B.-2 *Fluid Gate* YEMEN

Location

Between the little finger and ring finger, proximal to the margin of the web.

How to find

This point is best located when making a loose fist. Locate the web between the little finger and ring finger (4th and 5th fingers) and locate **T.B.-2** proximal to its margin. **T.B.-2** is part of → **Ex-UE-9** (*baxie*: proximal to the webs between the fingers). Located in a comparable position on the foot is → **G.B.-43** (on the web between the 4th and 5th toes), which is also part of → **Ex-LE-10** (*bafeng*).

Needling

Vertically up to 0.5 cun

Actions/Indications

- Clears Heat from the Upper Burner
- Benefits the ears and calms the *shen*
- Opens the channel and alleviates pain

Special features

Ying-spring point, Water point. Especially indicated for psycho-neurological disorders caused by Heat and local channel problems.

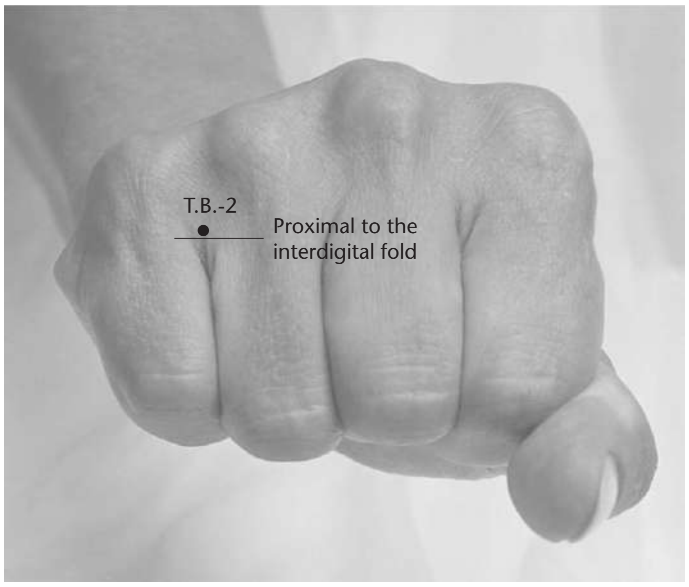
T.B.-2
Proximal to the interdigital fold

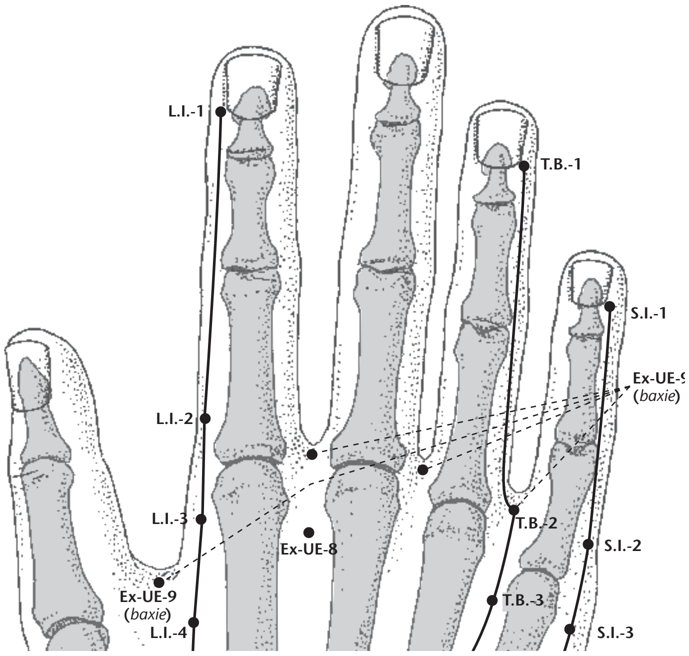
L.I.-1
T.B.-1
S.I.-1
Ex-UE-
(baxie)
L.I.-2
L.I.-3
T.B.-2
Ex-UE-8
S.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3

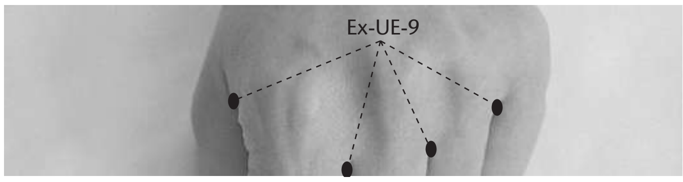

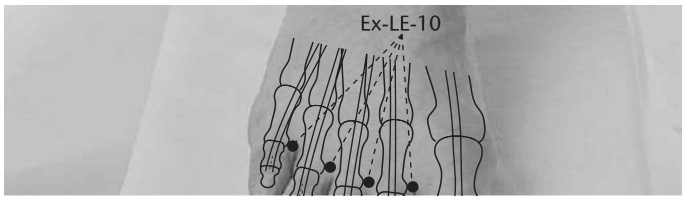

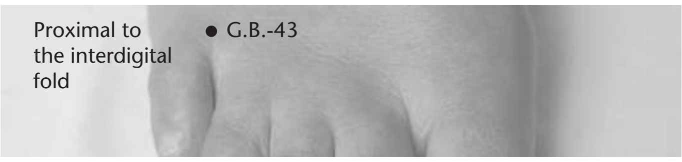
Proximal to the interdigital fold
G.B.-43

Central Islet ZHONGZHU **T.B.-3**

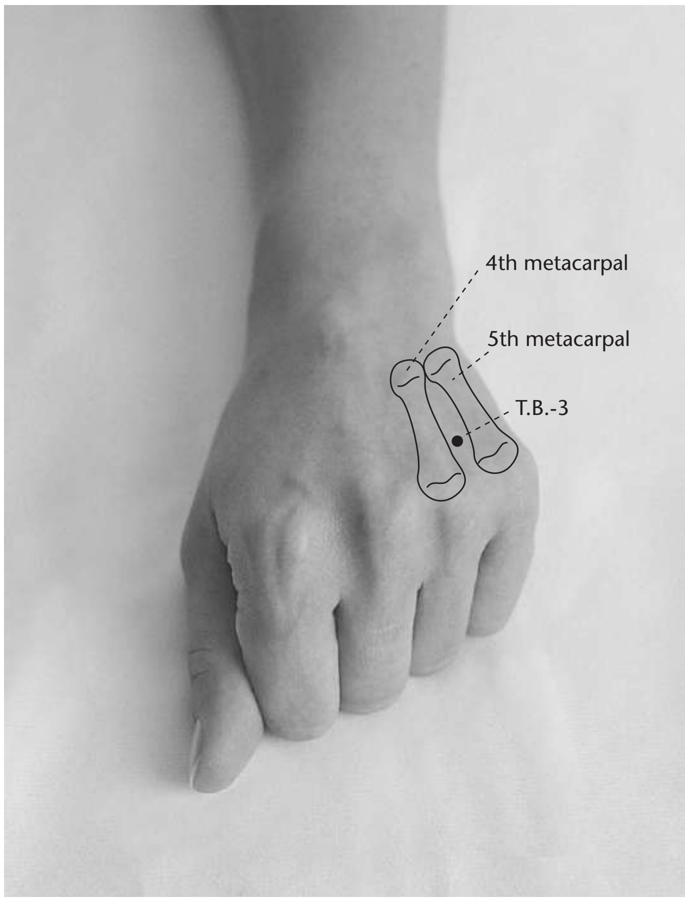

Location

On the dorsum of the hand, in a depression between the 4th and 5th metacarpal bones, proximal to the metacarpophalangeal joints. **T.B.-3** is located at the junction between the heads and shafts of the two metacarpal bones.

How to find

The hand should be relaxed or a loose fist should be made. Palpate from the metacarpophalangeal joints along the groove between the 4th and 5th metacarpal bones towards proximal. **T.B.-3** is located at the widest/deepest point of the groove, slightly distal to the carpometacarpal joints.
Located in comparable positions are → **Ex-UE-8** (*wailaogong*: between the 2nd and 3rd metacarpal bones), → **S.I.-3** (on the ulnar aspect of the 5th metacarpal bone) and → **L.I.-3** (on the radial aspect of the 2nd metacarpal bone).

Needling

0.5–1 cun vertically or obliquely in a proximal direction.

Actions/Indications

- Benefits the ears, clears Heat (and drains Fire), clears the head and the eyes
- Opens the channel and *luo*-connecting vessels, alleviates pain

Special features

Shu-stream point, Wood point, tonification point. One of the most important distal points for any disorder of the ears.

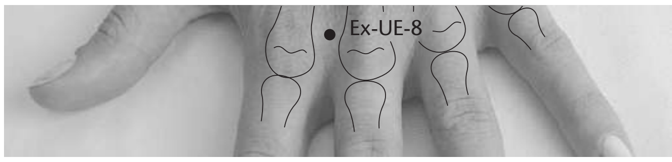

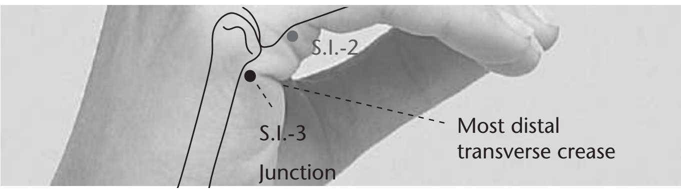

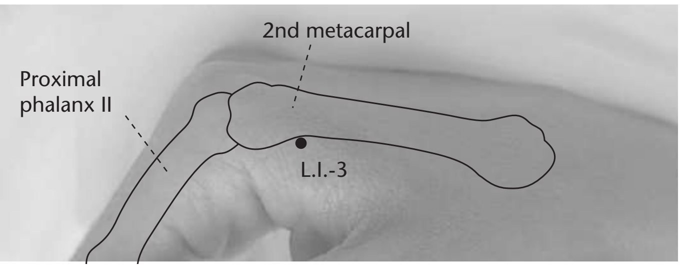

T.B.-4 *Yang Pool* YANGCHI

Location

On the dorsum of the wrist (above the wrist joint space, 'dorsal wrist crease'), in the gap between the tendons of the extensor digitorum (on the ulnar aspect) and the extensor digiti minimi muscles (on the radial aspect).

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (→ 3.3.3). By moving the patient's wrist in a relaxed way, the joint space becomes well defined. **T.B.-4** is located slightly laterally to the midpoint of the joint space, in the depression between the tendons of the extensor digiti minimi muscle (continuing to the little finger) on its radial aspect and the tendon of the extensor digitorum muscle on its ulnar aspect. The tendons of the extensor digitorum muscle will become more pronounced by moving the three middle fingers.

Located on the same level but radial to the tendons of the extensor digitorum muscle is → **Ex-UE-3** (*zhongquan*).

Needling

0.3–0.5 cun vertically or slightly obliquely in a proximal direction or transversely (subcutaneously) deep to the tendons towards the radial aspect of the wrist.

Actions/Indications

- Clears Heat, opens the channel, relaxes the tendons. In Japanese acupuncture tradition, **T.B.-4**, in its function as *yuan*-source point, is used for deficiency syndromes.

Special features

Yuan-source point

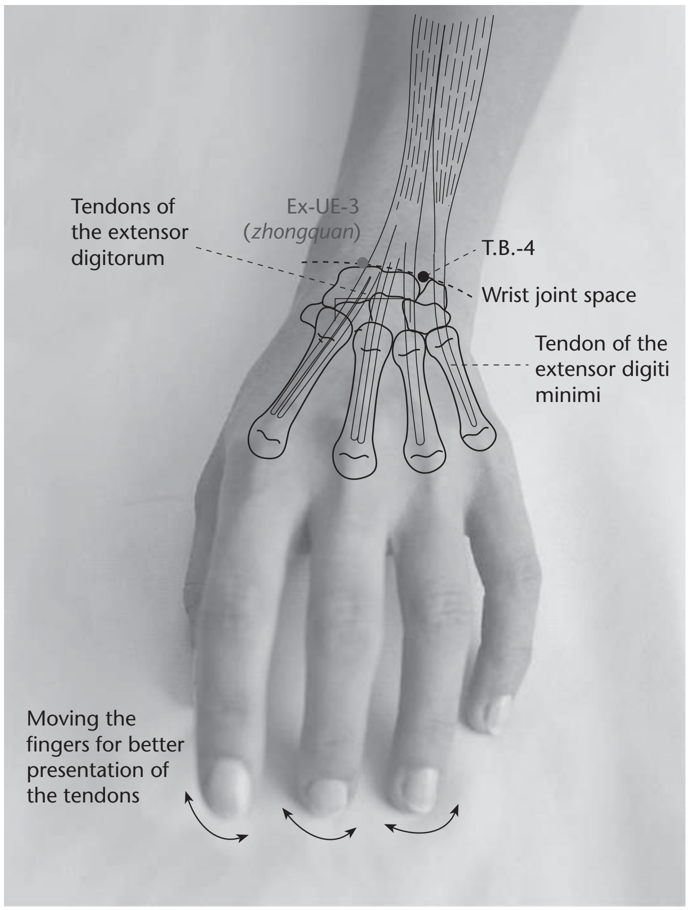
Tendons of the extensor digitorum
Ex-UE-3 (*zhongquan*)
T.B.-4
Wrist joint space
Tendon of the extensor digiti minimi
Moving the fingers for better presentation of the tendons

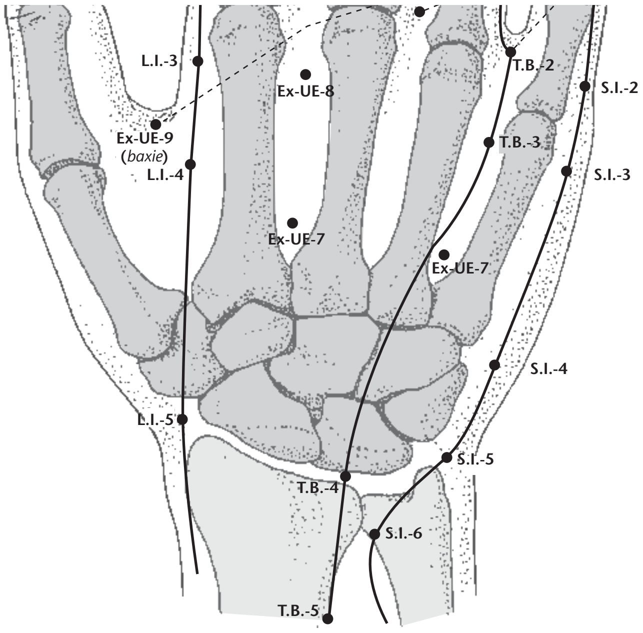
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(*baxie*)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

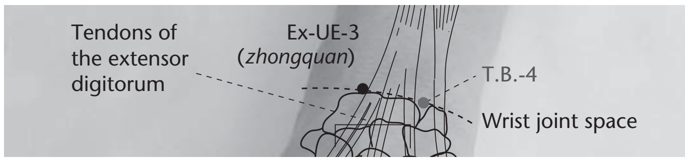
Tendons of the extensor digitorum
Ex-UE-3
(*zhongquan*)
T.B.-4
Wrist joint space

Outer Pass WAIGUAN T.B.-5

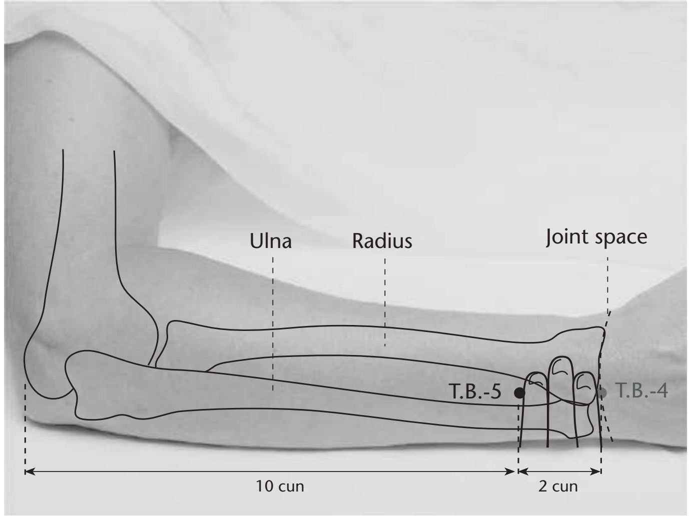

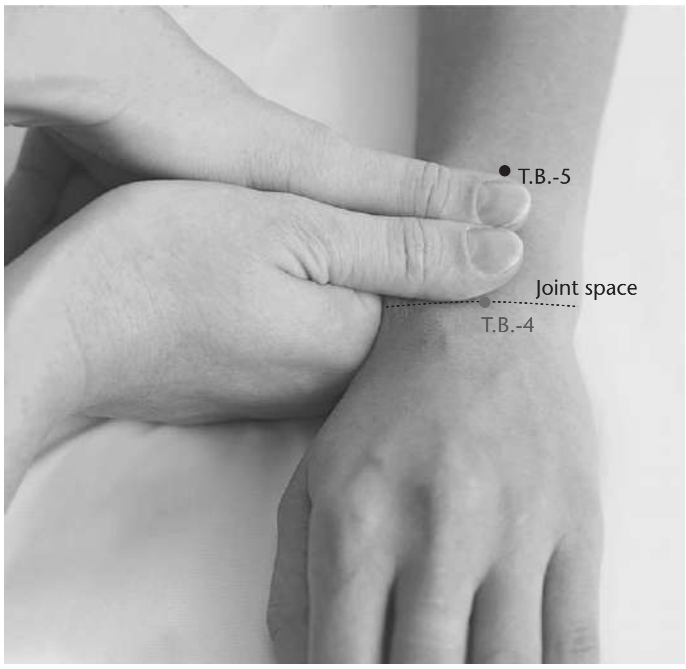

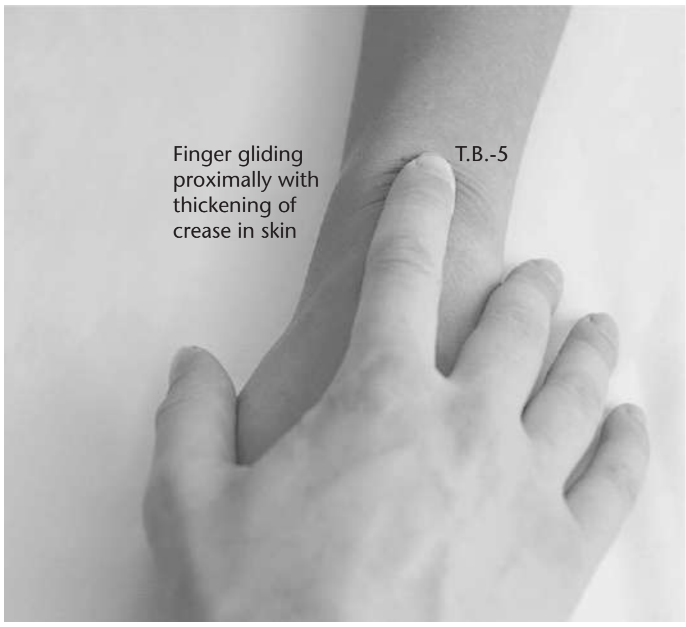

Location

2 cun proximal to the dorsal wrist joint space (‘dorsal wrist crease’), between the radius and the ulna.

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (→ 3.3.3). By moving the patient’s wrist in a relaxed way, the joint space becomes more easily palpable. **T.B.-5** is located 2 cun proximal to the midpoint of the joint space, in the depression midway between the radius and the ulna. **Or quick method:** With the palpating finger, slide with slight pressure from the dorsal aspect of the wrist joint in a proximal direction between the radius and the ulna, until the finger is brought to rest at **T.B.-5** by the increasing skin bulge building up in front of it.

→ **P-6** is located opposite **T.B.-5,** on the anterior aspect of the forearm.

Needling

Vertically or obliquely 0.5–1.5 cun. Caution: Movement of the hand/arm during needle retention may cause bending of the needle.

Actions/Indications

- Expels Wind and benefits the head and ears, clears Heat, opens and regulates the *yang wei mai*
- Opens the channel and alleviates pain

Special features

Luo-connecting point, opening (master) point of the *yang wei mai*. Major point for expelling Wind-Heat, analgesic point for the upper extremity.

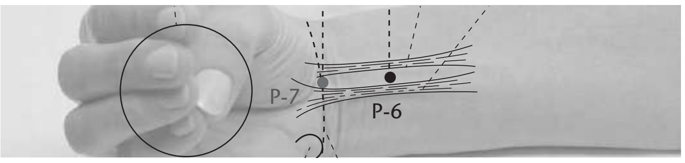

T.B.-6 *Branching Ditch* ZHIGOU

Location

3 cun proximal to the dorsal wrist joint space ('dorsal wrist crease'), in a depression between the radius and the ulna, radial to the tendon of the extensor digitorum communis muscle.

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (➞ 3.3.3). By moving the patient's wrist in a relaxed way, the joint space becomes more easily palpable. From the joint space, measure 3 cun in a proximal direction. At this point, the extensor digitorum communis muscle will often lie midway between the ulna and the radius. **T.B.-6** is located in a depression close to the border of the radius, on the radial side of the muscle.

Located on the same level is ➞ **T.B.-7,** in a depression between the ulna and the extensor digitorum communis muscle. ➞ **P-5** is also located on this level, but on the anterior aspect of the forearm.

Needling

0.5–1.5 cun slightly obliquely towards the ulna or obliquely in a proximal (elbow) or distal (wrist) direction. Or through-needling to **P-5** (*jianshi*). Caution: Movement of the hand/arm during needle retention may cause bending of the needle.

Actions/Indications

- Regulates the Qi, clears Heat in the Triple Burner, benefits the lateral costal region, promotes bowel movements
- Benefits the voice
- Opens the channel

Special features

Jing-river point, Fire point, *ben* point (Five Phases).

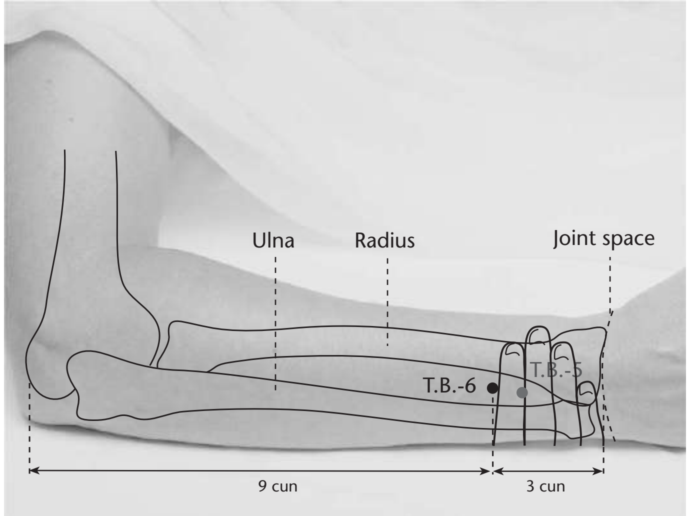

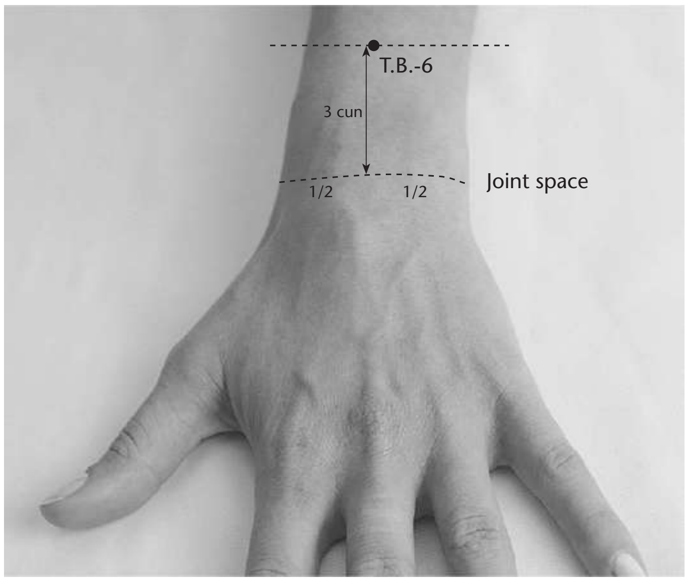

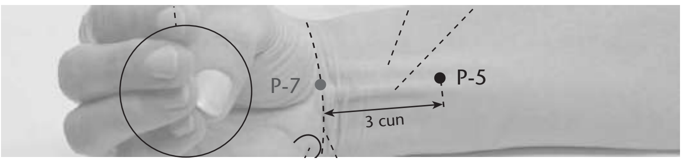

Ch04.10-F10028.qxd 2/23/08 2:27 PM Page 373

Ancestral Meeting HUIZONG T.B.-7

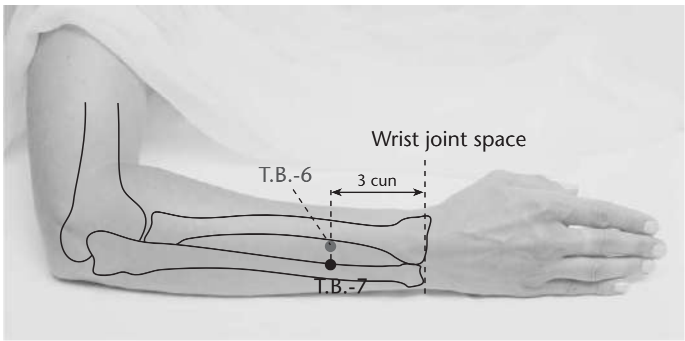

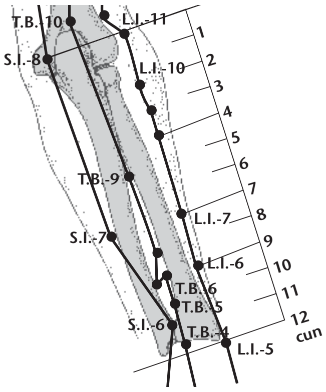

Location

3 cun proximal to the dorsal wrist joint space (‘dorsal wrist crease’) and 0.5 cun ulnar to the centre of the forearm.

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (→ 3.3.3). By moving the patient’s wrist in a relaxed way, the joint space becomes more easily palpable. From the joint space, measure 3 cun in a proximal direction and locate **T.B.-7** in a depression between the ulna and the extensor digitorum communis muscle. **Or:** Spreading hands technique (→ 2.3.3): Place the little fingers on the elbow crease and the wrist joint space respectively (this distance corresponds to 12 cun). Divide this distance into quarters and locate **T.B.-7** one quarter of the distance from the wrist joint and 0.5 cun ulnar to the centre of the forearm (where → **T.B.-6** is located) on the border of the ulna.

Located on the same level is → **T.B.-6** (0.5 cun towards the radius).

Needling

Vertically or obliquely up to 1.5 cun

Actions/Indications

- Opens the channel and alleviates pain
- Benefits the ears

Special features

Xi-cleft point

T.B.-8 *Three Yang Connection* SANYANGLUO

Location

4 cun proximal to the dorsal wrist joint space ('dorsal wrist crease'), between the radius and the ulna, radial to the tendon of the extensor digitorum communis muscle.

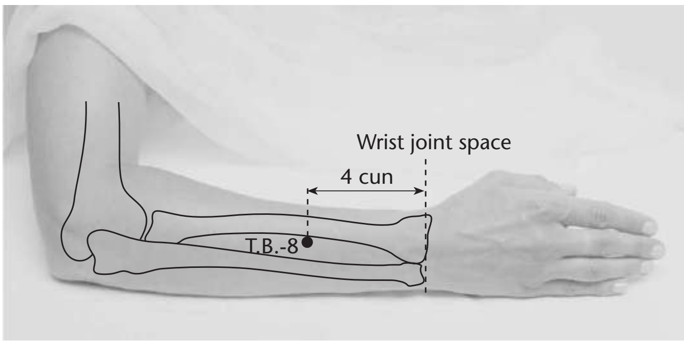

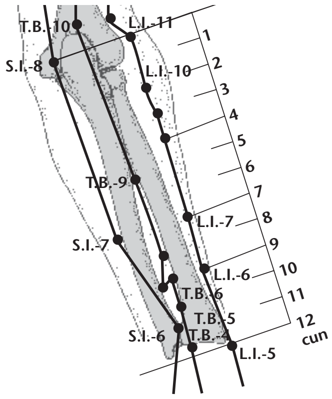

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (➞ 3.3.3). By moving the patient's wrist in a relaxed way, the joint space becomes more easily palpable. From the joint space, measure 4 cun in a proximal direction and locate **T.B.-8** in a depression between the ulna and radial to the extensor digitorum communis muscle. **Or:** Spreading hands technique (➞ 2.3.3): Divide the distance between the elbow crease and the wrist joint space (=12 cun) into thirds and locate **T.B.-8** one third of the distance from the wrist joint between the radius and the ulna.

Needling

Vertically or obliquely up to 1.5 cun

Actions/Indications

- Benefits the ears and the voice
- Opens the channel and the *luo*-connecting vessels, alleviates pain

Four Rivers SIDU T.B.-9

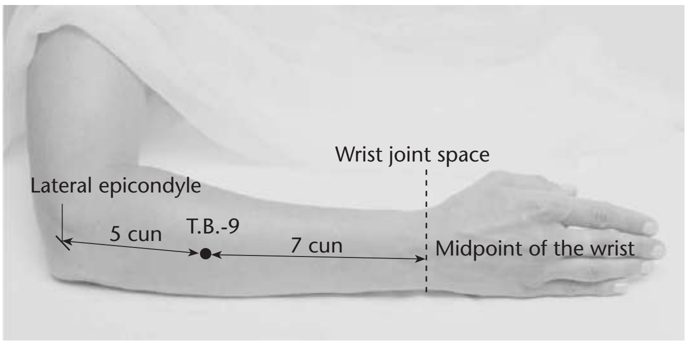

Location

7 cun proximal to the dorsal wrist joint space (‘dorsal wrist crease’), between the radius and the ulna.

How to find

From the wrist joint space (→ 3.3.3, more accurate reference point than the variable dorsal wrist crease), measure 7 cun in a proximal direction and locate **T.B.-9** in a depression between the extensor digitorum communis muscle and the extensor carpi ulnaris muscle. Alternatively, measure 5 cun distally from the lateral epicondyle of the humerus. **T.B.-9** is located on a line connecting the midpoint of the dorsal wrist joint space and the lateral epicondyle of the humerus. **Or:** Spreading hands technique (→ 2.3.3): Place the little fingers on the cubital crease and the wrist joint space (→ 2.2) respectively and join the thumbs at the midpoint of this distance (=12 cun). From the midpoint of this line, measure 1 cun in a proximal direction and there locate **T.B.-9** between the radius and the ulna.

Needling

Vertically or obliquely up to 2 cun

Actions/Indications

- Opens the channel and *luo*-connecting vessels
- Benefits the ears and throat

T.B.-10 *Heavenly Well* TIANJING

Location

On the lateral aspect of the upper arm, with the elbow flexed in a depression approximately 1 cun proximal to the olecranon.

How to find

This point is best located with the patient's elbow flexed to 90°, which will reveal an easily palpable depression 1 cun superior to the olecranon. **T.B.-10** is located in this depression, on the tendon of the triceps brachii muscle.

Needling

Up to 1 cun vertically or obliquely in a proximal direction.

Actions/Indications

- Transforms Phlegm and disperses accumulations
- Regulates and descends the Qi
- Calms the *shen*
- Clears Heat from the channel
- Opens the channel and *luo*-connecting vessels, alleviates pain

Special features

He-sea point, Earth point, sedation point.

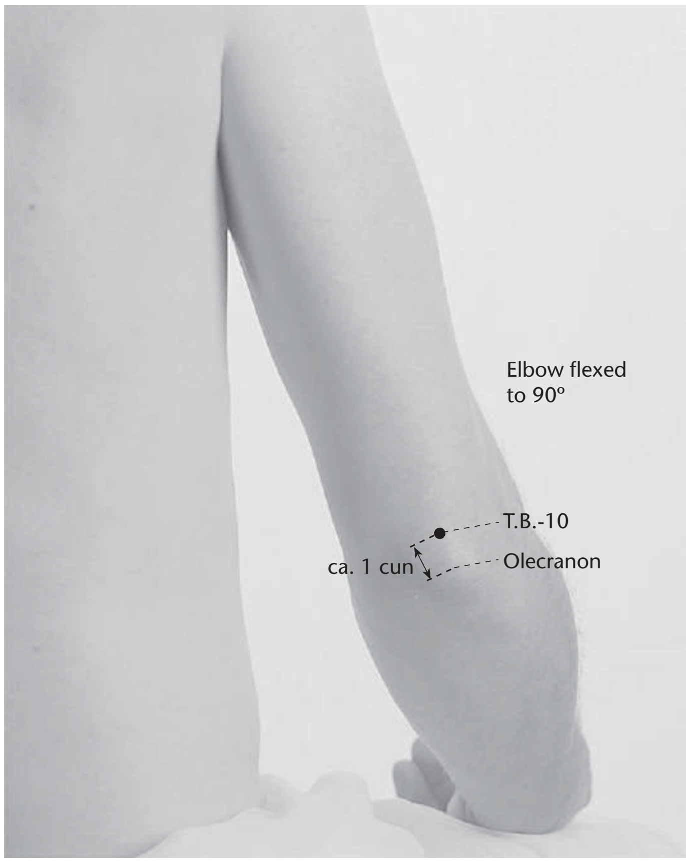
Elbow flexed to 90°
T.B.-10
ca. 1 cun
Olecranon

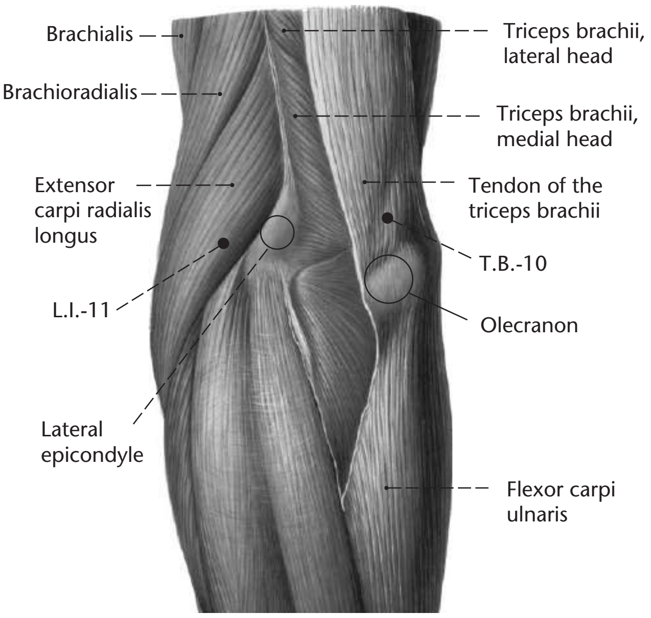
Brachialis
Brachioradialis
Extensor carpi radialis longus
L.I.-11
Lateral epicondyle
Triceps brachii, lateral head
Triceps brachii, medial head
Tendon of the triceps brachii
T.B.-10
Olecranon
Flexor carpi ulnaris

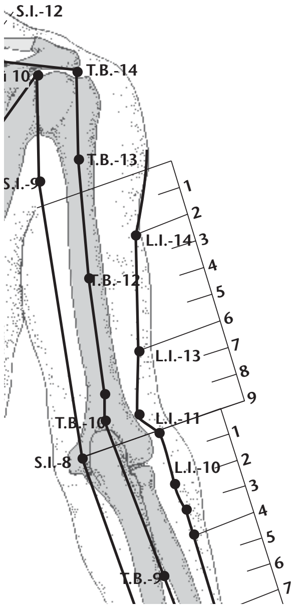
S.I.-12
10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I.-14
3
4
T.B.-12
5
6
L.I.-13
7
8
9
L.I.-11
T.B.-10
S.I.-8
1
L.I.-10
2
3
4
5
6
T.B.-9
7

Clear Cold Abyss QINGLENGYUAN T.B.-11

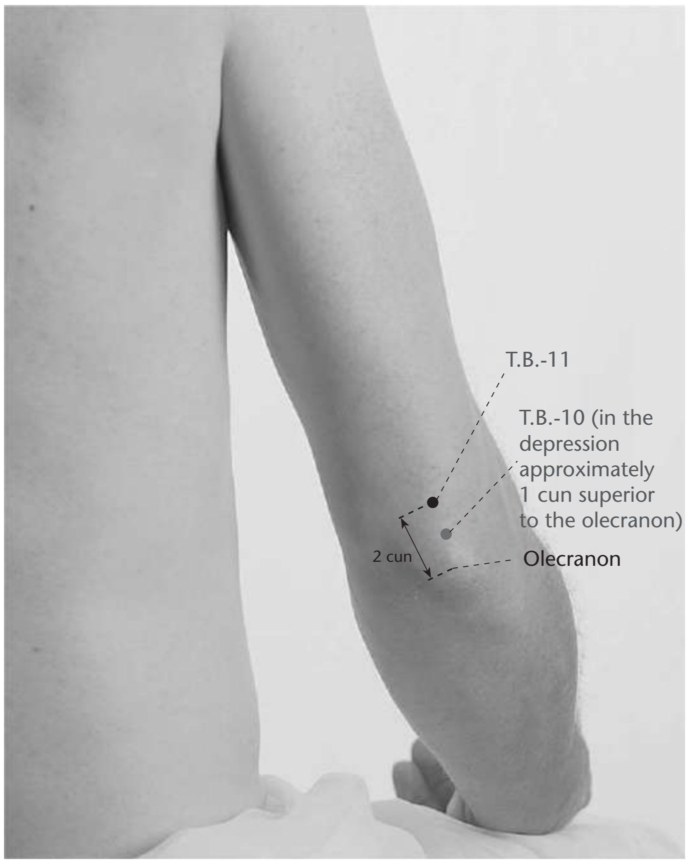
T.B.-11
T.B.-10 (in the depression approximately 1 cun superior to the olecranon)
2 cun
Olecranon

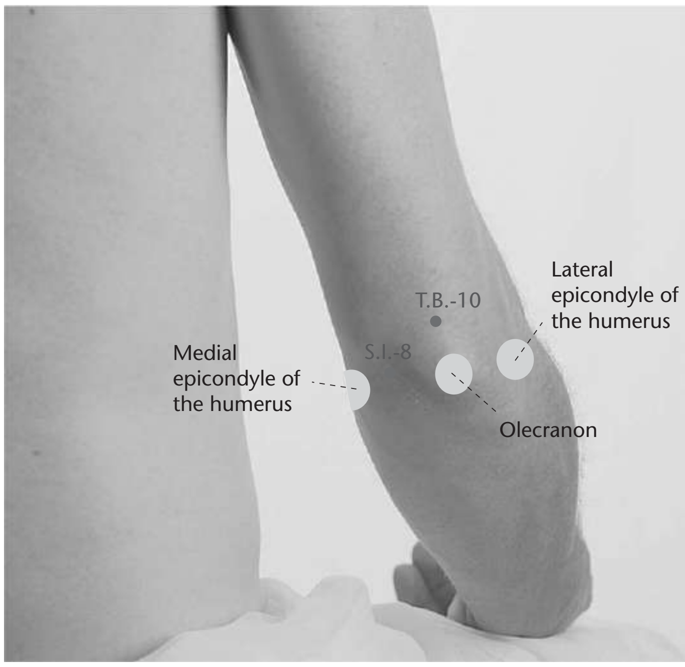
Lateral epicondyle of the humerus
T.B.-10
S.I.-8
Medial epicondyle of the humerus
Olecranon

Location

On the lateral aspect of the upper arm, with the elbow flexed 1 cun proximal to → **T.B.-10** or 2 cun proximal to the olecranon, on the triceps brachii muscle.

How to find

This point is best located with the patient's elbow flexed to 90°, using → **T.B.-10** as a reference point. This is located in an easily palpable depression 1 cun superior to the olecranon. From → **T.B.-10**, palpate approximately 1 cun in a proximal direction and locate **T.B.-11** in a depression on the triceps brachii muscle.

Needling

Vertically up to 1 cun

Actions/Indications

- Dispels Wind-Damp
- Opens the channel and the *luo*-connecting vessels
- Clears Damp-Heat

T.B.-12 *Dispersing Luo River* XIAOLUO

Location

4 cun proximal to ➞ **T.B.-10** (with the elbow flexed in a depression superior to the olecranon) or 5 cun proximal to the olecranon on a line connecting the olecranon and the lateral extremity of the acromion (location of ➞ **T.B.-14**).

How to find

First, locate the acromion (➞ 3.3.1). When the patient's arm is in abduction, two depressions will form below the acromion, in the area of origin of the deltoid muscle. ➞ **T.B.-14** is located in the posterior depression, below the posterior aspect of the lateral extremity of the acromion. Next, on a line connecting ➞ **T.B.-14** and the olecranon, measure 4 cun from ➞ **T.B.-10** (1 cun proximal to the olecranon, in the depression that forms when the elbow is flexed) in a proximal direction and there locate **T.B.-12**. **Or quick method:** Spreading hands technique (➞ 2.3.3): Place the little fingers on ➞ **T.B.-14** and ➞ **T.B.-10** respectively and determine the midpoint of this distance with the thumbs. From the midpoint, measure 1 cun in a distal direction and there locate **T.B.-12. For orientation: T.B.-12** is midway between ➞ **T.B.-11** (1 cun proximal to **T.B.-10** or 2 cun proximal to the olecranon) and ➞ **T.B.-13** (3 cun distal to ➞ **T.B.-14**, on the margin of the deltoid muscle).

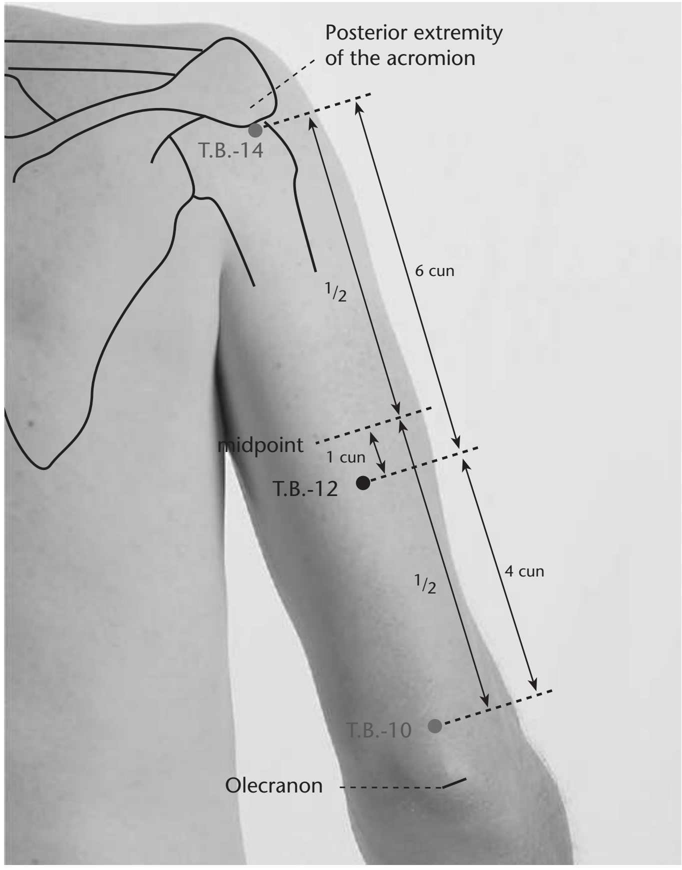
Posterior extremity
of the acromion
T.B.-14
6 cun
1/2
midpoint
1 cun
T.B.-12
1/2
4 cun
T.B.-10
Olecranon

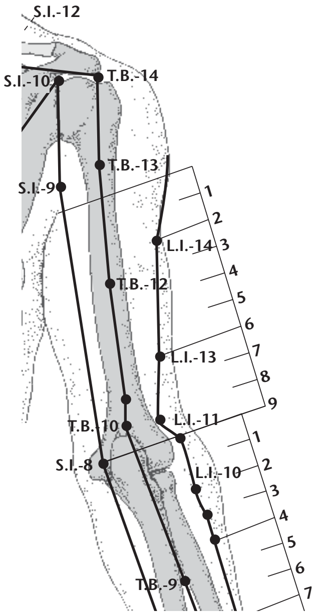
S.I.-12
S.I.-10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I.-14
3
4
T.B.-12
5
6
L.I.-13
7
8
9
L.I.-11
T.B.-10
1
S.I.-8
2
L.I.-10
3
4
5
6
T.B.-9
7

Needling

Vertically or obliquely up to 2 cun

Actions/Indications

- Opens the channel and alleviates pain

Special features

T.B.-12 represents an area extending from 5–7 cun proximal to the tip of the elbow rather than a discrete acupuncture point. Determine the exact location by tenderness.

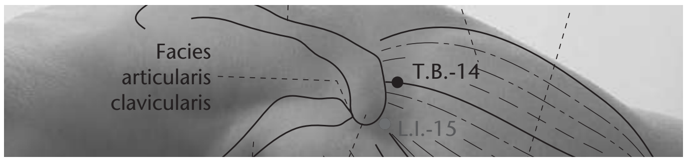
Facies
articularis
clavicularis
T.B.-14
L.I.-15

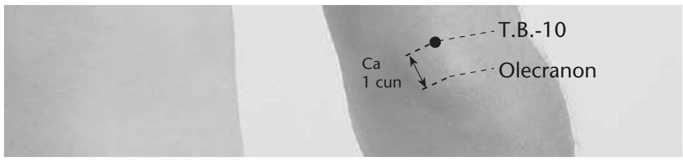
T.B.-10
Ca
1 cun
Olecranon

Upper Arm Meeting NAOHUI **T.B.-13**

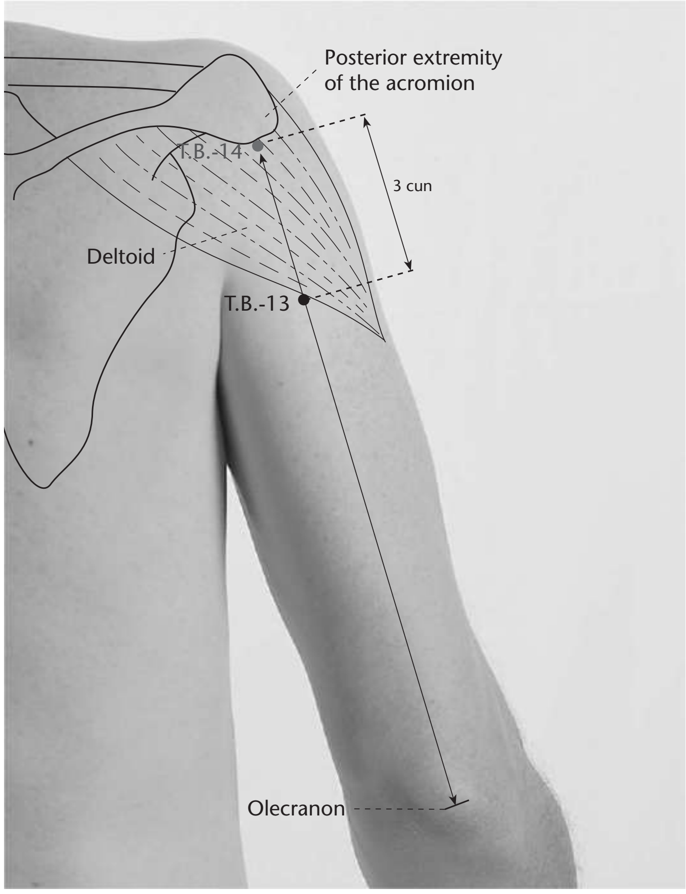
Posterior extremity of the acromion
T.B.-14
3 cun
Deltoid
T.B.-13
Olecranon

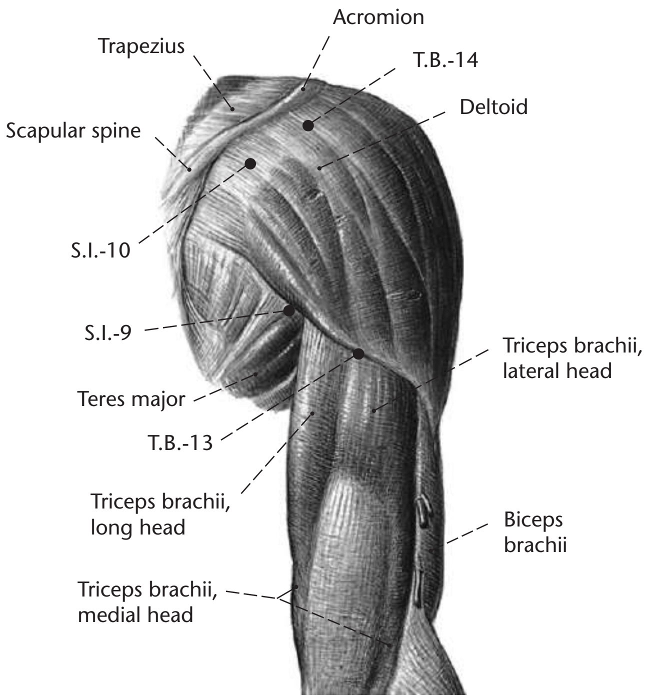
Acromion
Trapezius
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

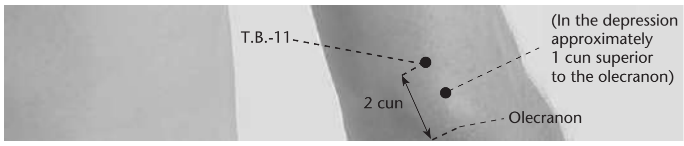
T.B.-11
(In the depression approximately 1 cun superior to the olecranon)
2 cun
Olecranon

Location

3 cun distal to the lateral extremity of the acromion (location of → **T.B.-14**), on a line connecting → **T.B.-14** and the olecranon, at the junction of this line with the margin of the deltoid muscle.

How to find

First, locate the acromion (→ 3.3.1). When the patient's arm is in abduction, two depressions will form inferior to the acromion, in the area of origin of the deltoid muscle. → **T.B.-14** is located in the posterior depression, below the dorsal aspect of the lateral extremity of the acromion. Next, locate **T.B.-13** 3 cun distal to → **T.B.-14**, on a line connecting → **T.B.-14** and the olecranon. **T.B.-13** is located at the junction of this line with the margin of the deltoid muscle, approximately on the level of the end of the posterior axillary fold.

Needling

Vertically or obliquely up to 2 cun

Actions/Indications

- Opens the channel and the *luo*-connecting vessels, alleviates pain
- Regulates the Qi flow and transforms Phlegm

Special features

According to some authors, meeting point with the *yang wei mai* and the G.B. channel

T.B.-14 *Shoulder Crevice* JIANLIAO

Location

Inferior to the lateral extremity of the acromion between the acromial and spinal portions of the deltoid muscle or, with the arm abducted, in the posterior of the two depressions on the shoulder joint.

How to find

Ask the patient to abduct their arm to a 90° angle, which will reveal two depressions anterior and posterior to the acromion. **T.B.-14** is located in the posterior depression, directly below the dorsal aspect of the lateral extremity of the acromion, between the acromial and spinal fibres of the deltoid muscle. The dorsal aspect of the lateral extremity of the acromion becomes more easily palpable when the upper arm is gently pulled (extension of the upper arm), allowing the head of the humerus to glide anteriorly.

→ **L.I.-15** is located in the depression anterior to **T.B.-14.**

Needling

With the arm adducted, 0.5–1 cun perpendicularly towards the centre of the axilla or 1–2 cun transversally to obliquely towards the elbow.

Actions/Indications

- Disperses Wind and Dampness, benefits the shoulder joint, opens the channel and the *luo*-connecting vessels, alleviates pain

Special features

Important local point for shoulder pain, especially for the dorsal aspect of the shouler

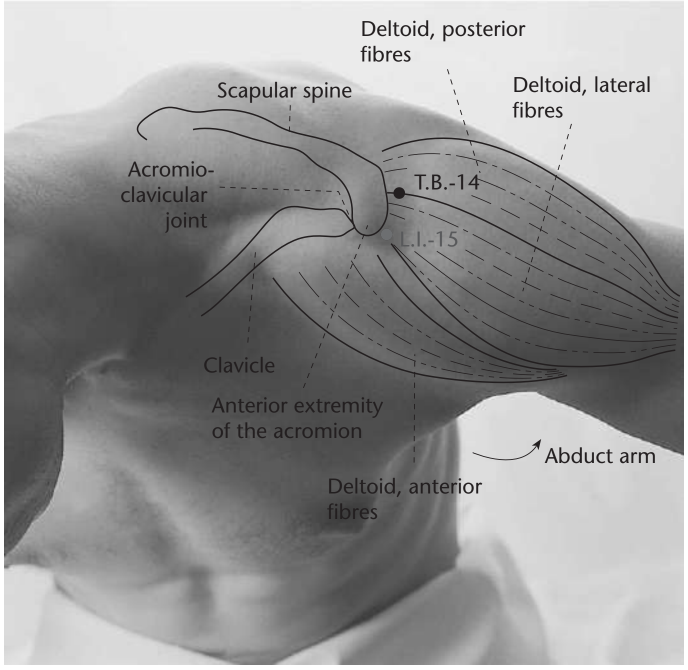
Deltoid, posterior fibres
Scapular spine
Deltoid, lateral fibres
Acromio-clavicular joint
T.B.-14
L.I.-15
Clavicle
Anterior extremity of the acromion
Abduct arm
Deltoid, anterior fibres

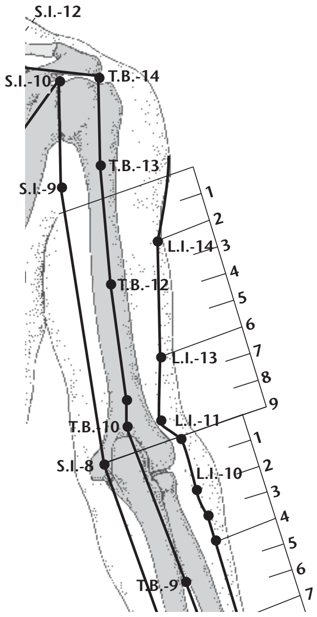
S.I.-12
S.I.-10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I.-14
3
4
T.B.-12
5
6
7
L.I.-13
8
9
L.I.-11
T.B.-10
1
S.I.-8
2
L.I.-10
3
4
5
6
T.B.-9
7

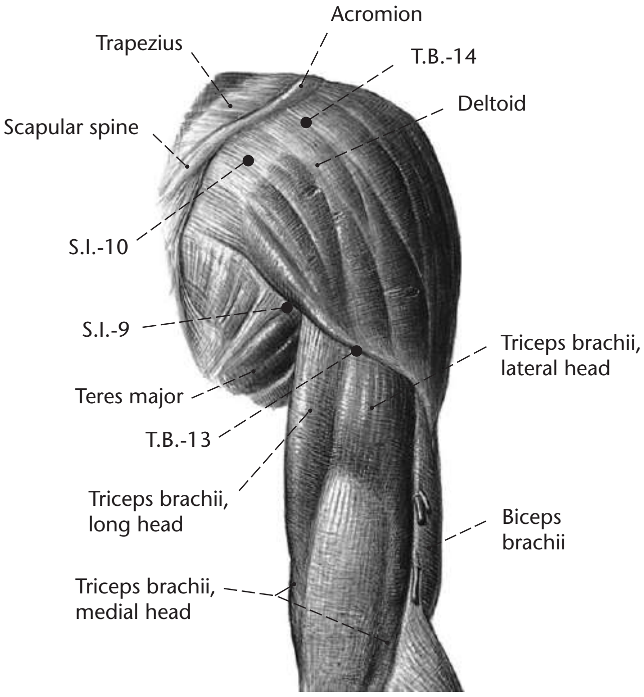
Acromion
Trapezius
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

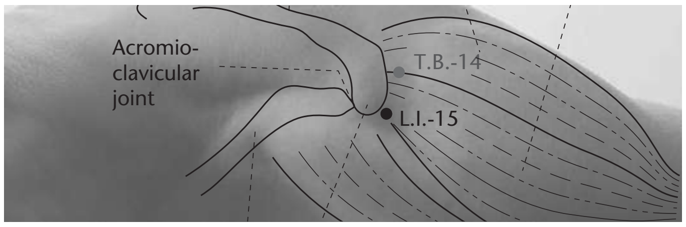
Acromio-clavicular joint
T.B.-14
L.I.-15

Heavenly Crevice TIANLIAO T.B.-15

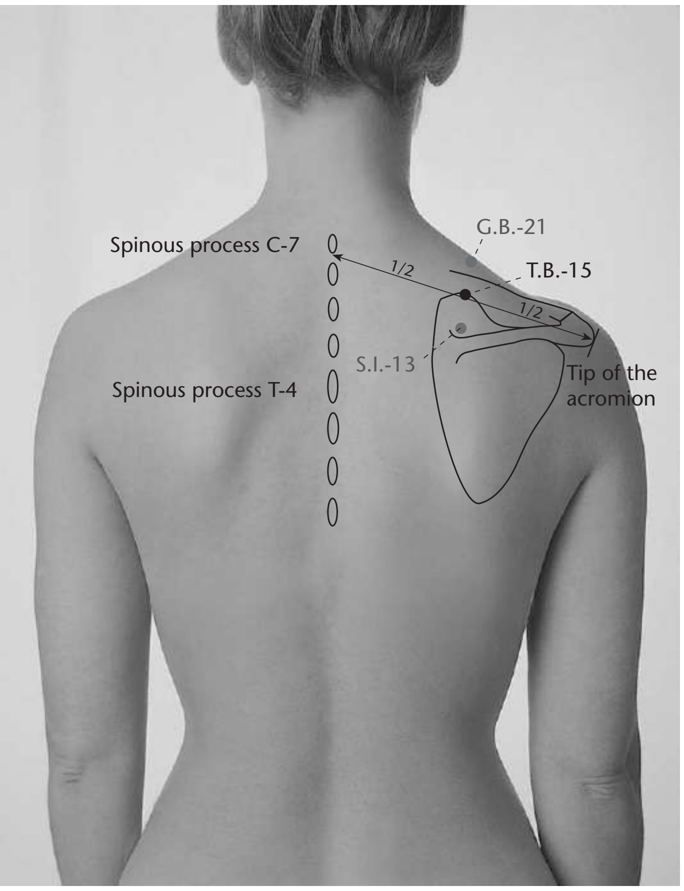
G.B.-21
Spinous process C-7
1/2
T.B.-15
1/2
S.I.-13
Tip of the acromion
Spinous process T-4

Location

At the midpoint of an imaginary line between the spinous process of the 7th cervical vertebra (C7) and the tip of the acromion (lateral extremity of the acromion).

How to find

First, locate C7 (→ 3.4.1) and the acromion (→ 3.3.1), which represents the lateral extension of the scapular spine (→ 3.3.1) and, at its most lateral end, forms a flat surface superior to the head of the humerus (acromial angle). Using, for example, the spreading hands technique (→ 2.3.3), determine the midpoint between C7 and the acromial angle. Located at the highest point of the shoulder is → **G.B.-21. T.B.-15** is located approximately 1 cun inferior to → **G.B.-21**, on the superior angle of the scapula.
→ **S.I.-13** is located inferior to **T.B.-15**, directly superior to the medial end of the scapular spine.

Needling

0.5–1 cun vertically or obliquely towards the neck or shoulder. Caution: Pneumothorax.

Actions/Indications

- Opens the channel and alleviates pain
- Dispels especially Wind-Damp, unbinds the chest

Special features

Meeting point with the *yang wei mai* and the G.B. channel

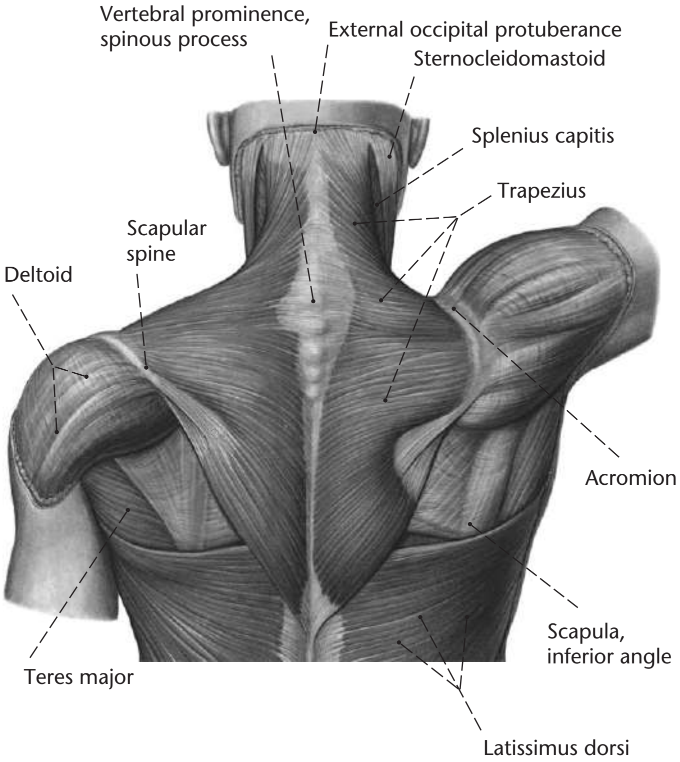
Vertebral prominence, spinous process
External occipital protuberance
Sternocleidomastoid
Splenius capitis
Trapezius
Scapular spine
Deltoid
Acromion
Scapula, inferior angle
Teres major
Latissimus dorsi

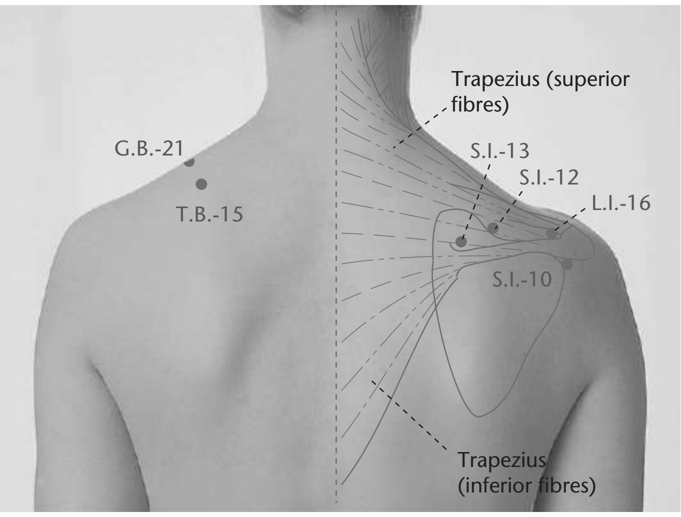
Trapezius (superior fibres)
G.B.-21
S.I.-13
S.I.-12
L.I.-16
T.B.-15
S.I.-10
Trapezius (inferior fibres)

T.B.-16 *Window of Heaven* TIANYOU

Location

At the posterior border of the sternocleidomastoid muscle, directly inferior to the mastoid process, on the level of the mandibular angle.

How to find

For a better definition of the sternocleidomastoid muscle, ask the patient to rotate their head against resistance towards the side to be needled. **T.B.-16** is located on the level of the mandibular angle, on the dorsal aspect of the sternocleidomastoid muscle. This point is located inferior to the posterior border of the mastoid (→ 3.1.4). → **G.B.-12** is located directly posterior and inferior to the mastoid; → **S.I.-17** is also located on the level of the mandibular angle, but on the anterior border of the sternocleidomastoid muscle.

Needling

Vertically 0.5–1 cun. Caution: Carotid artery.

Actions/Indications

- Benefits the head and sensory organs
- Descends the Qi, expels Wind
- Opens the channel and alleviates pain

Special features

Window of Heaven point

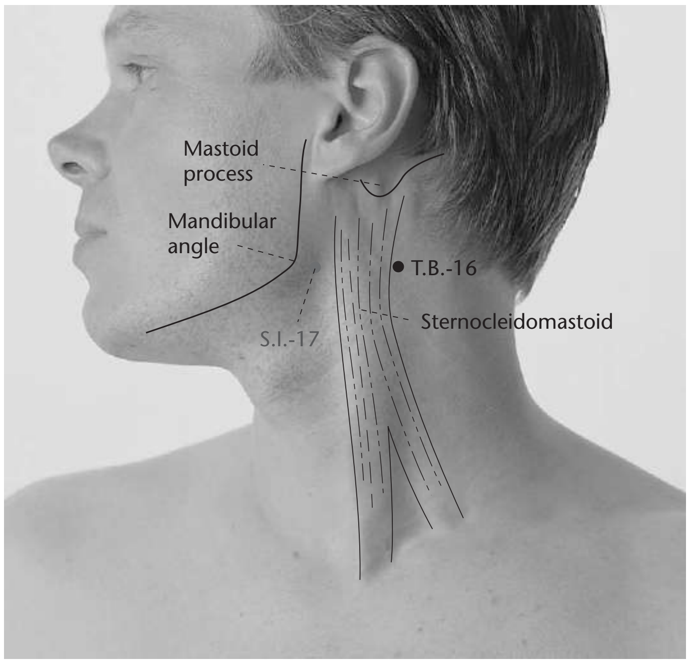
Mastoid process
Mandibular angle
T.B.-16
Sternocleidomastoid
S.I.-17

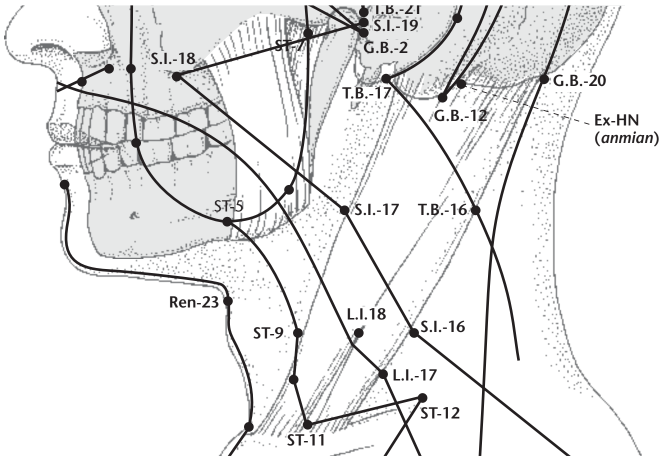
T.B.-21
S.I.-19
G.B.-2
S.I.-18
ST-7
T.B.-17
G.B.-20
G.B.-12
Ex-HN
(anmian)
ST-5
S.I.-17
T.B.-16
Ren-23
L.I.18
ST-9
S.I.-16
L.I.-17
ST-12
ST-11

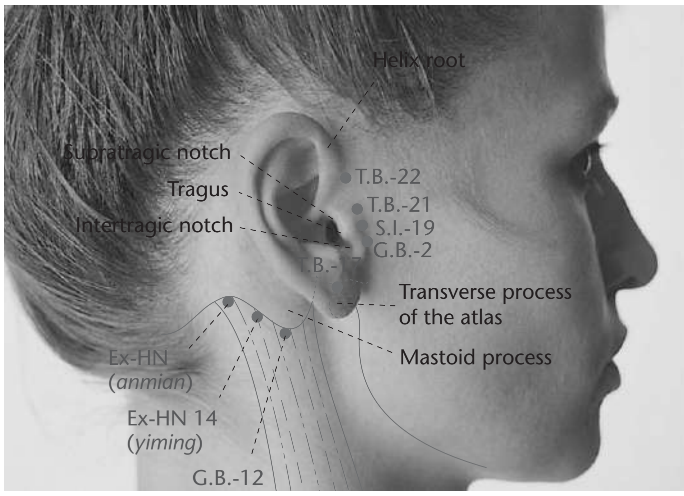
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
B.-
Transverse process of the atlas
Mastoid process
Ex-HN
(anmian)
Ex-HN 14
(yiming)
G.B.-12

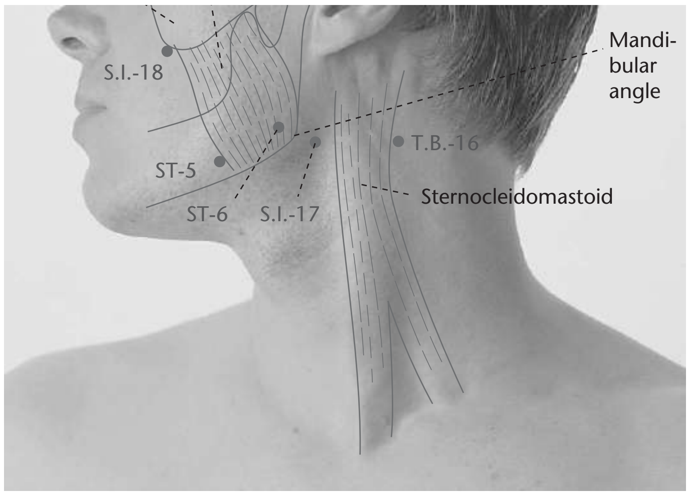
S.I.-18
Mandibular angle
T.B.-16
ST-5
Sternocleidomastoid
ST-6
S.I.-17

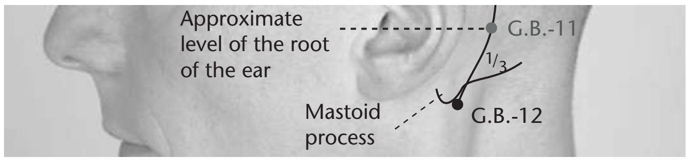
Approximate level of the root of the ear
G.B.-11
1/3
Mastoid process
G.B.-12

Wind Screen YIFENG T.B.-17

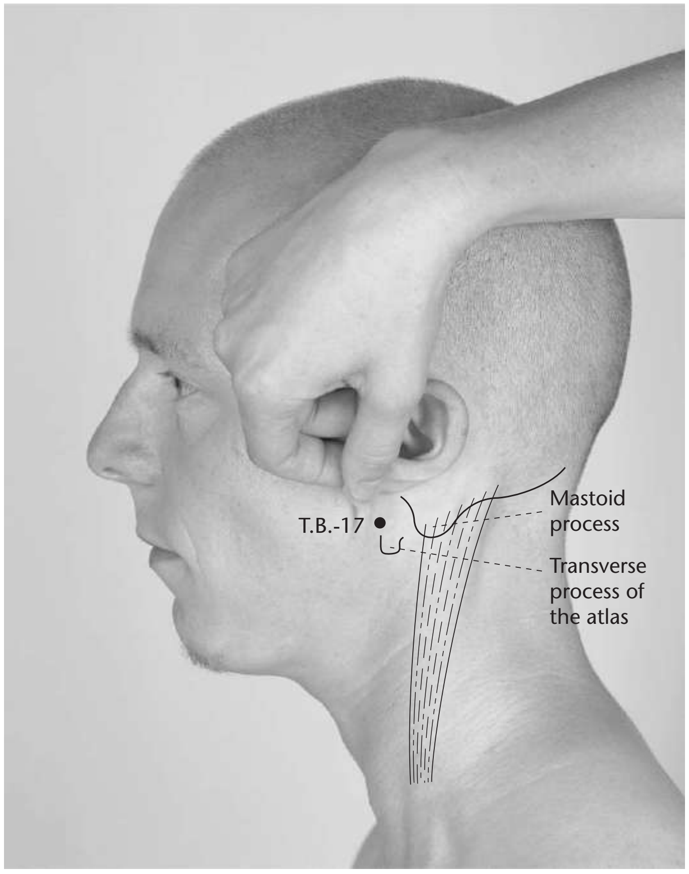
T.B.-17
Mastoid process
Transverse process of the atlas

Location

With the patient's mouth open, in the depression behind the earlobe, between the mastoid process and the mandible.

How to find

This point should be located with the patient's mouth open. **Quick method:** When pressing onto the ear lobe, it will touch **T.B.-17**. **Or:** Gently fold the earlobe anteriorly. **T.B.-17** is located posterior to the lobe, in a pressure-sensitive depression which can be palpated between the mandible and the mastoid process. The tip of the needle will get close to the transverse process of the 1st cervical vertebra (transverse process of the atlas, → 3.1.4), which, below the earlobe, can be palpated as a deep bony structure often sensitive to pressure.

Needling

This point should be needled with the patient's mouth slightly open. Caution: This point is close to the facial nerve, therefore no deep needling. Superior to the transverse process, the vertebral artery curves round posteriorly, therefore strictly needle 0.5–1 cun in an anterior direction.

Actions/Indications

- Expels (external) Wind, benefits the ears, clears Heat, opens the channel and *luo*-connecting vessels, alleviates pain

Special features

Meeting point with the G.B. channel. Important local point for disorders of the ears and the temporomandibular joint.

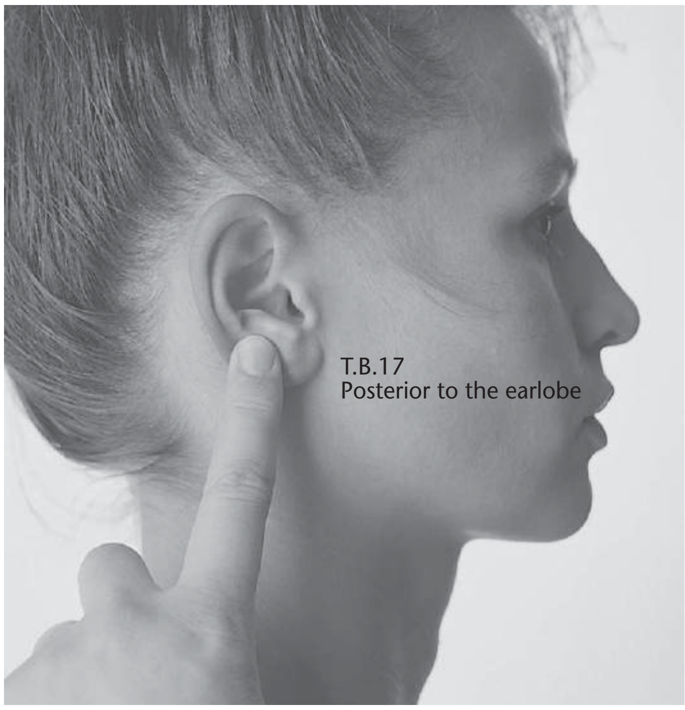
T.B.17
Posterior to the earlobe

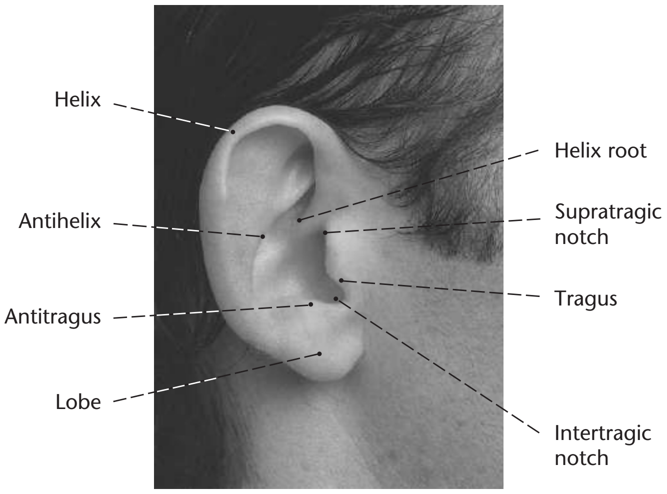
Helix
Antihelix
Antitragus
Lobe
Helix root
Supratragic notch
Tragus
Intertragic notch

T.B.-18 *Spasm Vessel* QIMAI

Location

Posterior to the ear, in a well-defined depression in the centre of the mastoid.

How to find

On an imaginary clock face superimposed over the auricle (12 o'clock = apex of the ear, 6 o'clock = earlobe), you can find an easily palpable depression at 8 o'clock (on the right side) and at 4 o'clock (on the left side). It is located directly behind the margin of the auricle and is the location of **T.B.-18**.

→ **G.B.-11** is located slightly superior to **T.B.-18** and approximately 0.3 cun from the ear; → **G.B.-12** is located inferior to **T.B.-18**, directly posterior to the mastoid.

Needling

0.5 cun transversely (subcutaneously) in an inferior direction or prick to bleed.

Actions/Indications

- Dispels Wind
- Opens the channel, alleviates pain
- Benefits the ears

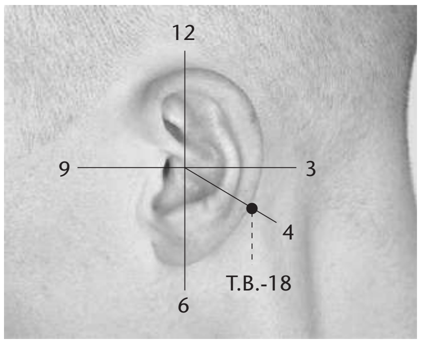
12
9
3
4
T.B.-18
6

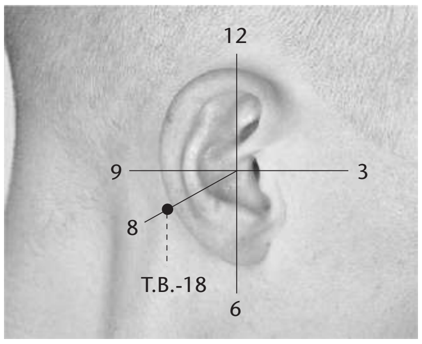
12
9
3
8
T.B.-18
6

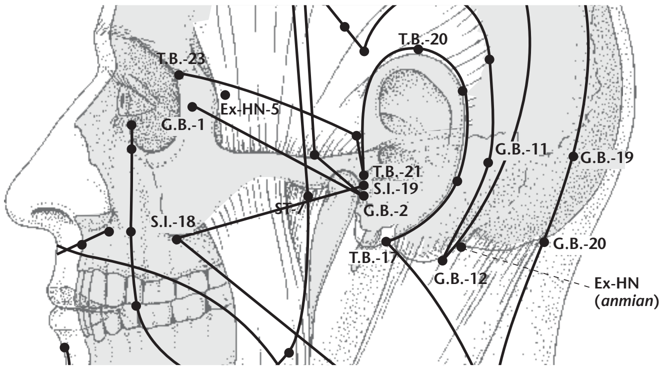
T.B.-23
T.B.-20
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
ST-7
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(anmian)

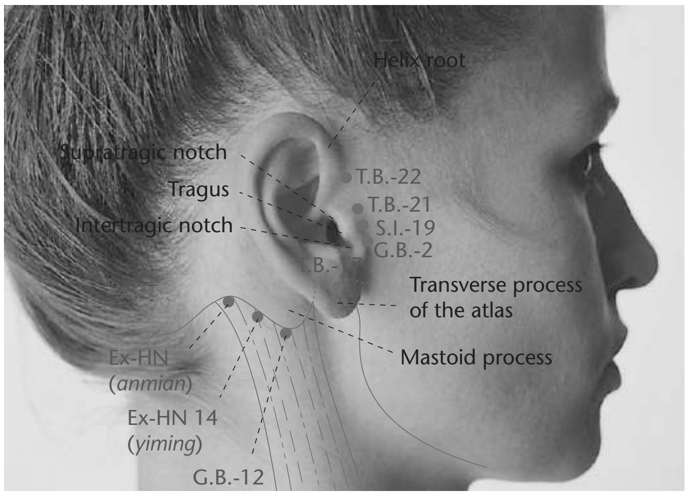
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
Transverse process of the atlas
Mastoid process
Ex-HN (anmian)
Ex-HN 14 (yiming)
G.B.-12

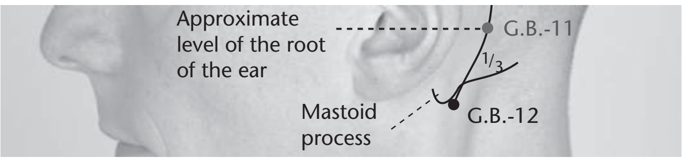
Approximate level of the root of the ear
G.B.-11
1/3
Mastoid process
G.B.-12

Skull's Rest LUXI T.B.-19

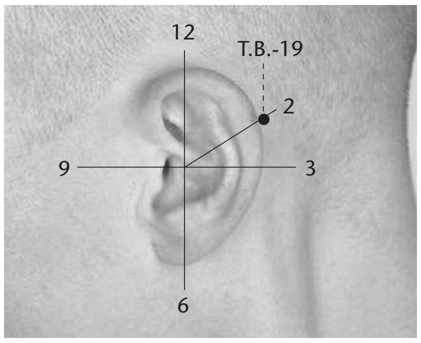

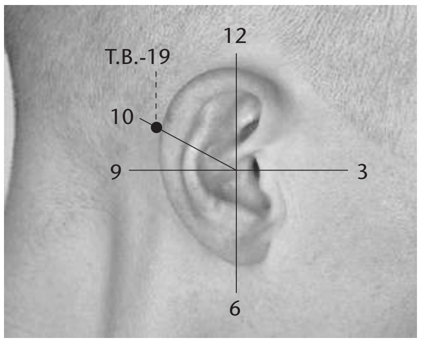

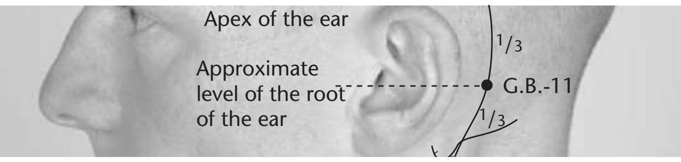

Location

Posterior to the ear, in a well-defined depression superior to the centre of the ear.

How to find

On an imaginary clock face superimposed over the auricle (12 o'clock = apex of the ear, 6 o'clock = earlobe), you can find an easily palpable depression at approximately 10 o'clock (on the right side) and at 2 o'clock (on the left side). It is located directly behind the margin of the auricle and is the location of **T.B.-19**.
→ **G.B.-11** is located slightly inferior to **T.B.-19** and approximately 0.3 cun from the margin of the ear.

Needling

Up to 0.5 cun transversely (subcutaneously) in an inferior direction or prick to bleed.

Actions/Indications

- Dispels Wind
- Opens the channel
- Benefits the ear, clears Heat

T.B.-20 *Minute Angle* JIAOSUN

Location

Directly superior to the apex of the ear.

How to find

Directly superior to the apex of the ear. → **G.B.-7** is also located on the apex of the ear, but anterior to it, within the circumauricular hairline. → **G.B.-8** is located 1.5 cun superior to the apex of the ear.

Needling

Up to 1.5 cun transversely (subcutaneously) in the direction of the disorder.

Actions/Indications

- Benefits the ears and the eyes
- Clears Heat, especially from the mouth

Special features

Meeting point with the G.B. channel, also with the S.I. channel according to some authors.

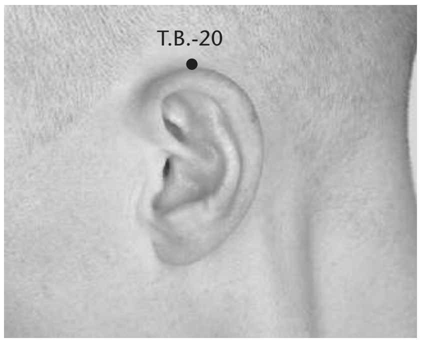

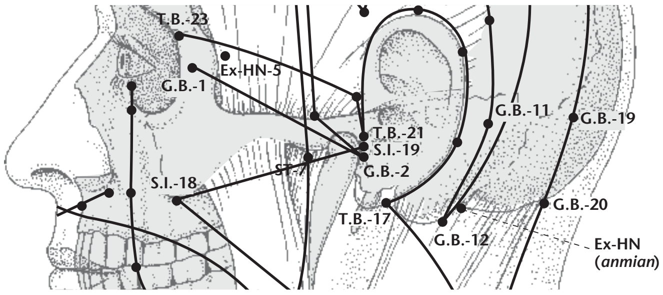
T.B.-23
Ex-HN-5
G.B.-1
T.B.-21
S.I.-19
G.B.-2
S.I.-18
ST-7
T.B.-17
G.B.-11
G.B.-19
G.B.-20
G.B.-12
Ex-HN
(*anmian*)

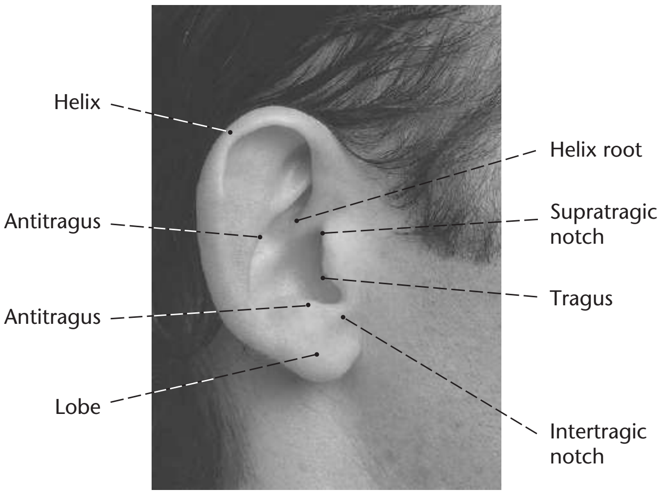
Helix
Antitragus
Antitragus
Lobe
Helix root
Supratragic notch
Tragus
Intertragic notch

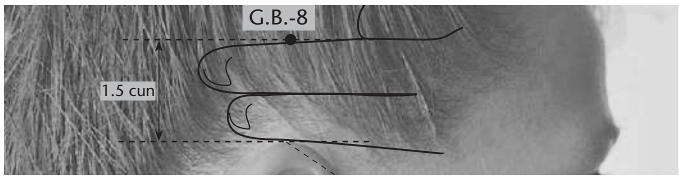

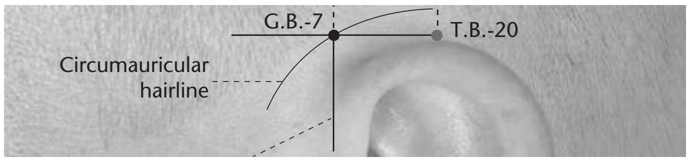
G.B.-7
T.B.-20
Circumauricular hairline

Ch04.10-F10028.qxd 2/23/08 2:27 PM Page 387

Ear Gate ERMEN T.B.-21

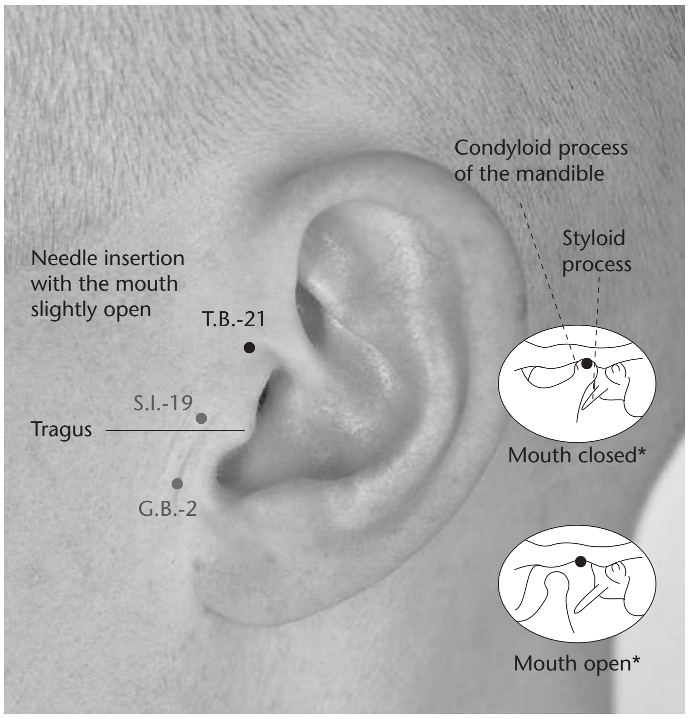
Condyloid process of the mandible
Styloid process
Needle insertion with the mouth slightly open
T.B.-21
S.I.-19
Tragus
G.B.-2
Mouth closed*
Mouth open*

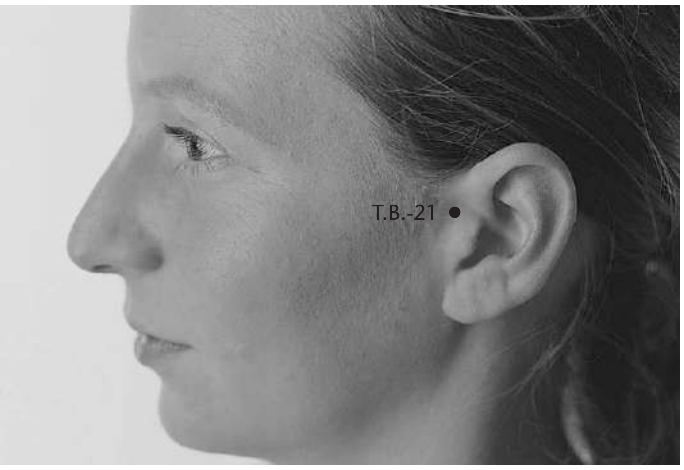

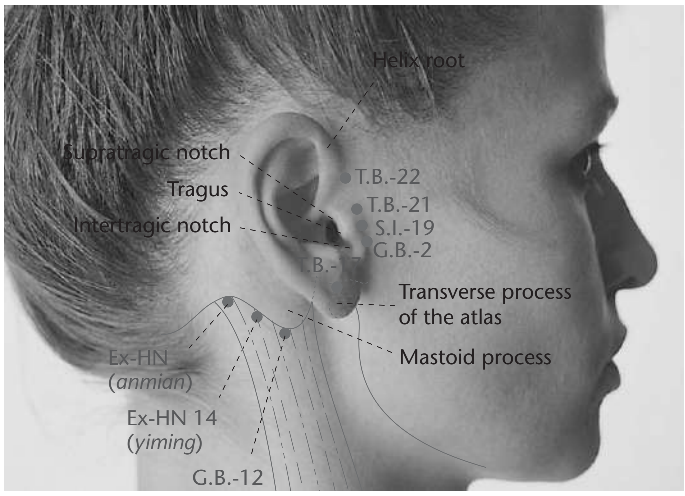
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
B.-
Transverse process of the atlas
Mastoid process
Ex-HN (*anmian*)
Ex-HN 14 (*yiming*)
G.B.-12

Location

Anterior to the ear, with the patient's mouth open in the depression on the level of the supratragic notch and slightly superior to the condyloid process of the mandible.

How to find

This point should be both located and needled with the patient's mouth open. This allows the condyloid process of the mandible to slide anteriorly to reveal the depression where **T.B.-21** is located. Locate the sulcus between the auricle and the cheek anterior to the tragus (this can be more or less well defined). Then locate **T.B.-21** on the level of the supratragic notch on the sulcus. If the latter cannot be clearly identified (it becomes more pronounced with increasing age), the depression can be located with an (ear) point locator when the patient's mouth is opened. **T.B.-21** is the most superior of three points located anterior to the ear (**→ S.I.-19** and **→ G.B.-2** are located more distally).

Needling

0.5–1 cun vertically or slightly obliquely in an inferior direction. Needle with the patient's mouth open (to avoid intra-articular insertion). After insertion, patients can close their mouth again. Caution: Like S.I.-19 and G.B.-2, this point is close to the superficial temporal artery and the auriculotemporal nerve.

Actions/Indications

- Clears Heat, benefits the ears

Special features

Important local point for disorders of the ears

---

* Accoring to Deadman et al 1998.

T.B.-22 *Ear Harmony Crevice* ERHELIAO

Location

In a depression at the border of the circumauricular temporal hairline, anterior to and on the level of the root of the auricle.

How to find

First, locate the anterior aspect of the root of the ear. From there, palpate approximately the breadth of the little finger in the direction of the eye. There, **T.B.-22** is located superior to the zygomatic arch (→ 3.1.2), which forms a bony ridge when sliding inferiorly from **T.B.-22**.
Located inferior to the zygomatic arch is → **T.B.-21**.

Needling

Transversely (subcutaneously) 0.5 cun. Caution: Superficial temporal artery.

Actions/Indications

- Dispels Wind
- Opens the channel and the *luo*-connecting vessels

Special features

Meeting point with the S.I. and G.B. channels, exit point according to some authors.

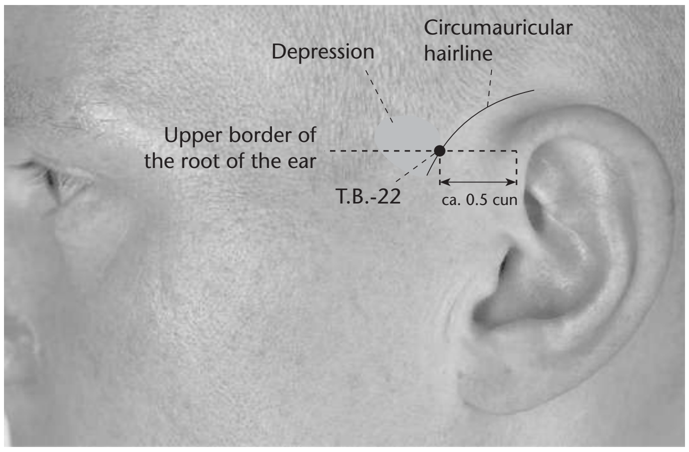
Depression
Circumauricular hairline
Upper border of the root of the ear
T.B.-22
ca. 0.5 cun

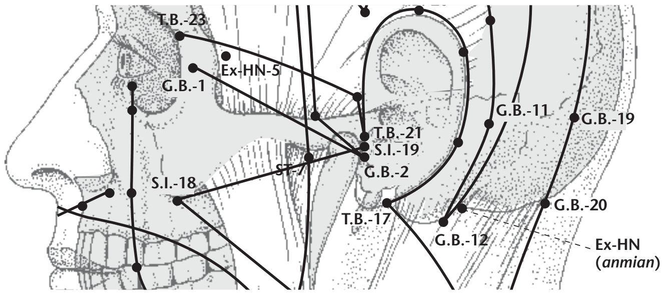
T.B.-23
Ex-HN-5
G.B.-1
T.B.-21
S.I.-19
G.B.-2
S.I.-18
ST-7
T.B.-17
G.B.-11
G.B.-19
G.B.-20
G.B.-12
Ex-HN
(anmian)

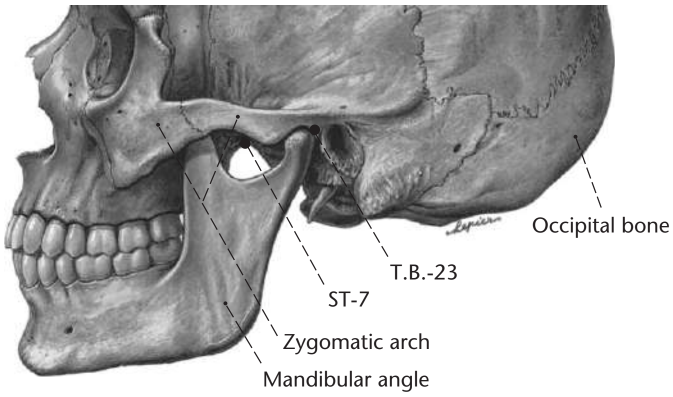
Occipital bone
T.B.-23
ST-7
Zygomatic arch
Mandibular angle

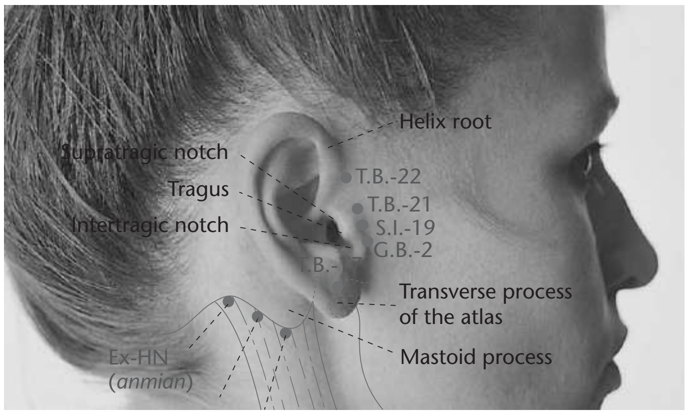
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
Transverse process of the atlas
Mastoid process
Ex-HN
(anmian)

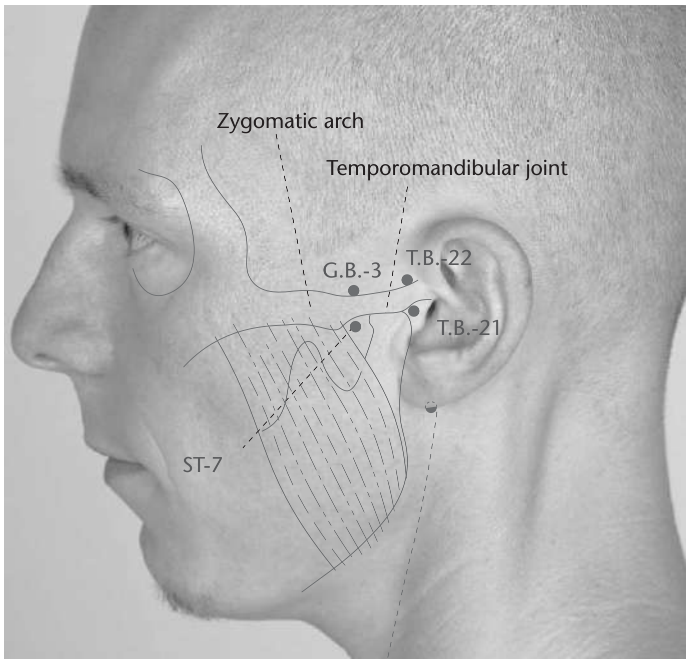
Zygomatic arch
Temporomandibular joint
G.B.-3
T.B.-22
T.B.-21
ST-7

Silken Bamboo Hollow SIZHUKONG T.B.-23

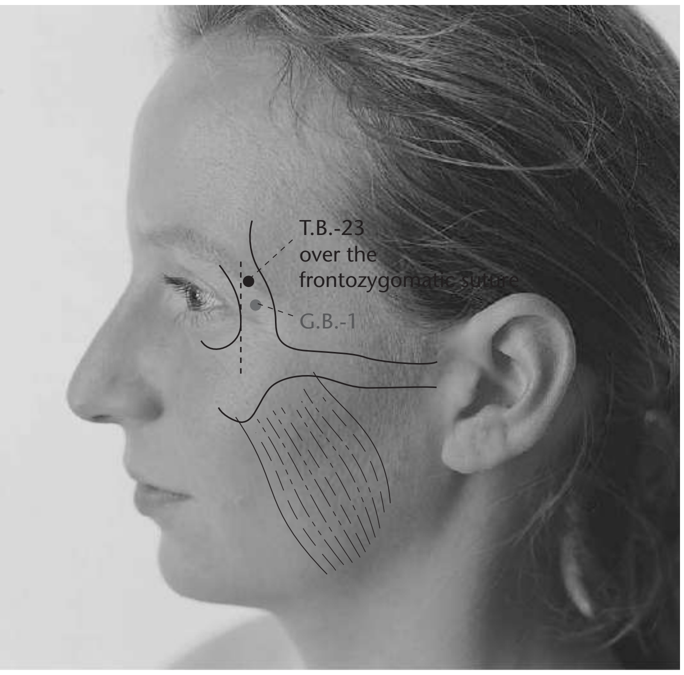
T.B.-23
over the
frontozygomatic suture
G.B.-1

Location

On the lateral end of the eyebrow, in the bony depression of the frontozygomatic suture, between the frontal and zygomatic bones.

How to find

The frontozygomatic suture (between the zygoma and the frontal bone) is generally located at the lateral end of the eyebrow. As the position of the latter can vary considerably, the suture is a more suitable reference point. In order to locate it, palpate from the outer canthus of the eye superiorly along the orbital margin, from its zygomatic to the frontal section, until you can feel a bony depression in the area of the suture. Locate **T.B.-23** in this mainly pressure-sensitive depression.

Needling

0.5–1 cun obliquely or transversely (subcutaneously) in a posterior direction towards **Ex-HN-5** (*taiyang*). In China, this point is also needled along the eyebrow to **Ex-HN-4** (*yuyao*: in the centre of the eyebrow). Moxibustion is contraindicated according to some classic texts.

Actions/Indications

- Benefits the eyes, alleviates pain, eliminates Wind

Special features

Important local point for headaches and disorders of the eyes

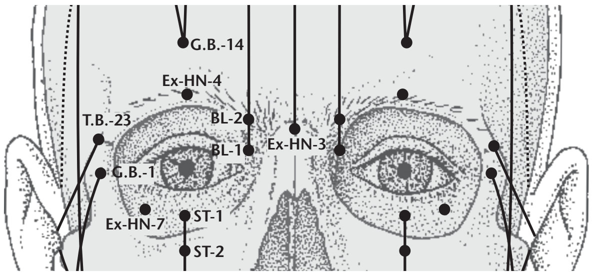
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2

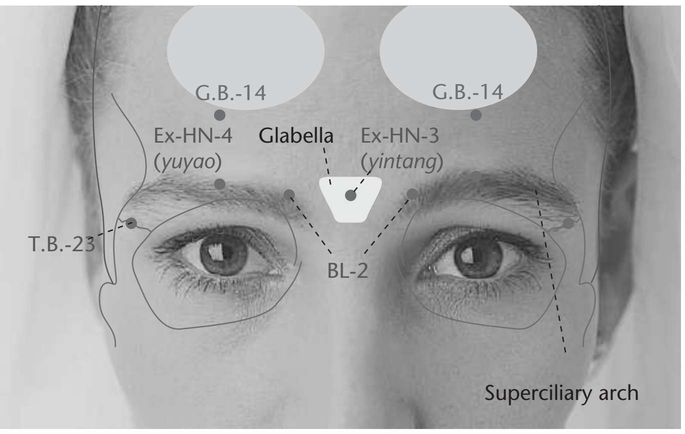
G.B.-14
G.B.-14
Ex-HN-4
(yuyao)
Glabella
Ex-HN-3
(yintang)
T.B.-23
BL-2
Superciliary arch

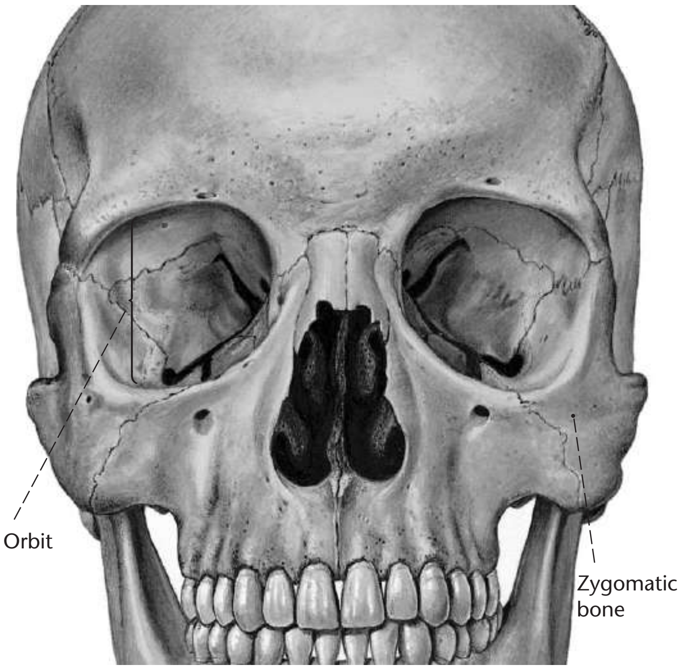
Orbit
Zygomatic
bone

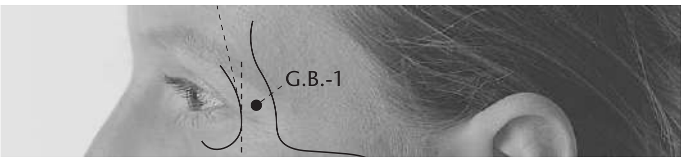

4.11 Gall Bladder Channel System – Foot *shaoyang* (*zu shao yang jing luo*)

4.11.1 The Gall Bladder Primary Channel (*zu shao yang jing*)

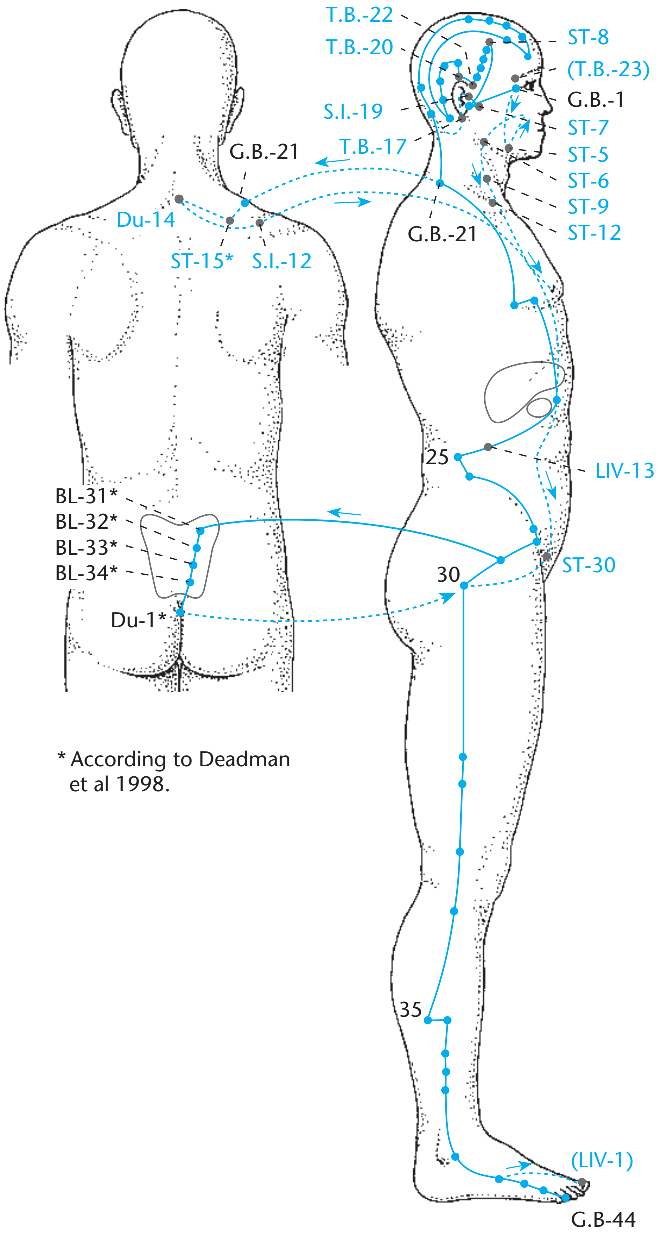
T.B.-22
T.B.-20
S.I.-19
T.B.-17
G.B.-21
Du-14
ST-15*
S.I.-12
G.B.-21
ST-8
(T.B.-23)
G.B.-1
ST-7
ST-5
ST-6
ST-9
ST-12
25
LIV-13
BL-31*
BL-32*
BL-33*
BL-34*
Du-1*
30
ST-30
35
(LIV-1)
G.B-44

* According to Deadman et al 1998.

Pathway

The Gall Bladder primary channel begins at **G.B.-1** (*tongziliao*) at the outer canthus of the eye. This point is reached by a **small branch** which separates from the Triple Burner primary channel at → **T.B.-23** (*sizhukong*) on the frontozygomatic suture, according to some authors at → **T.B.-22** (hand–foot pairing of the third great circuit, Yang axes: *shaoyang*).

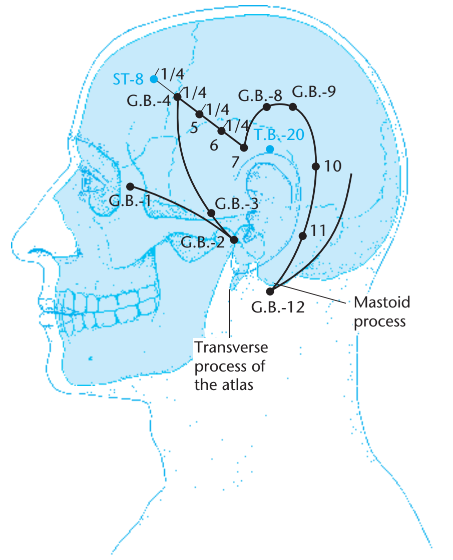
ST-8
1/4
G.B.-4
1/4
1/4
5
1/4
6
7
G.B.-8
G.B.-9
T.B.-20
10
G.B.-1
G.B.-3
G.B.-2
11
G.B.-12
Mastoid process
Transverse process of the atlas

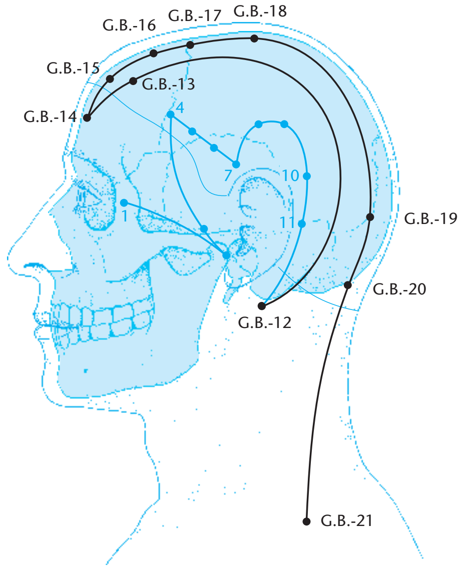
G.B.-16
G.B.-17
G.B.-18
G.B.-15
G.B.-13
G.B.-14
4
7
10
1
11
G.B.-19
G.B.-20
G.B.-12
G.B.-21

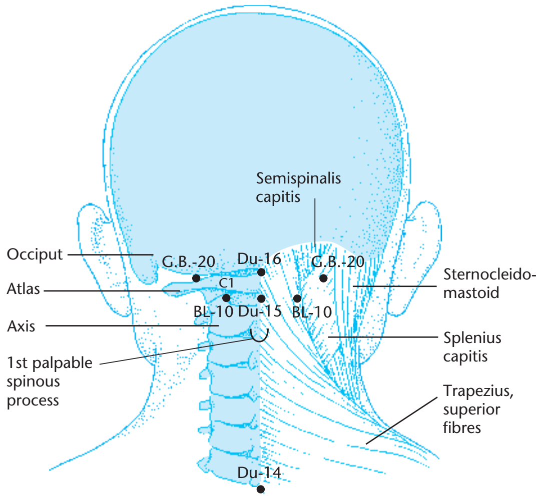
Semispinalis capitis
Occiput
Atlas
Axis
1st palpable spinous process
G.B.-20
Du-16
G.B.-20
C1
BL-10
Du-15
BL-10
Sternocleido-mastoid
Splenius capitis
Trapezius, superior fibres
Du-14

From **G.B.-1**, the **external pathway** runs to **G.B.-2** (*tinghui*) in front of the ear

➡ ascends to the forehead and **ST-8** (*touwei*)

➡ traverses the temporal region in a slight curve, passing

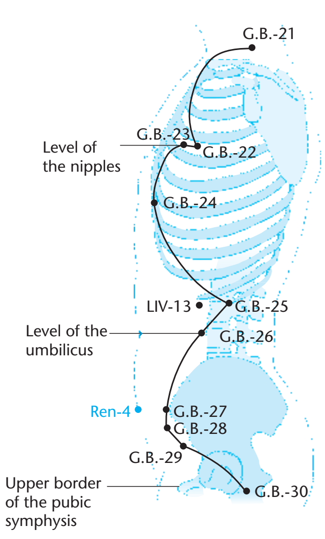
G.B.-21
G.B.-23
G.B.-22
Level of the nipples
G.B.-24
LIV-13
G.B.-25
Level of the umbilicus
G.B.-26
Ren-4
G.B.-27
G.B.-28
G.B.-29
Upper border of the pubic symphysis
G.B.-30

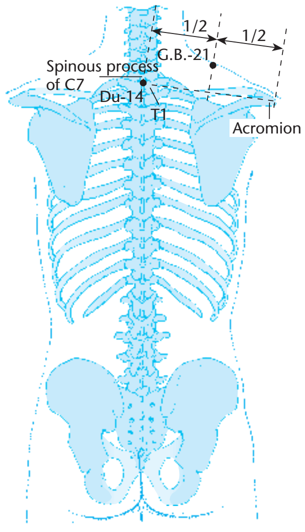
1/2
1/2
G.B.-21
Spinous process of C7
Du-14
T1
Acromion

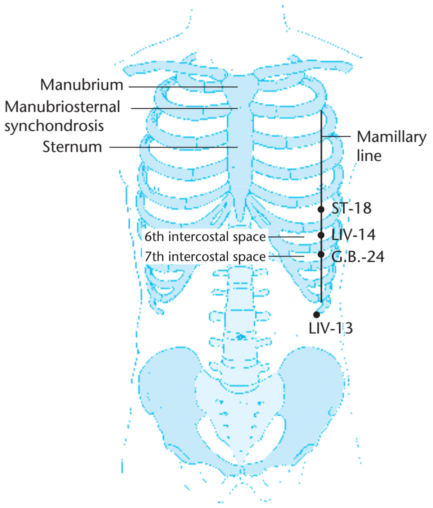
Manubrium
Manubriosternal synchondrosis
Sternum
Mamillary line
ST-18
6th intercostal space
LIV-14
7th intercostal space
G.B.-24
LIV-13

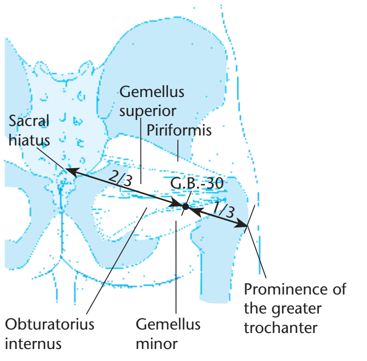
Gemellus superior
Sacral hiatus
Piriformis
2/3
G.B.-30
1/3
Prominence of the greater trochanter
Obturatorius internus
Gemellus minor

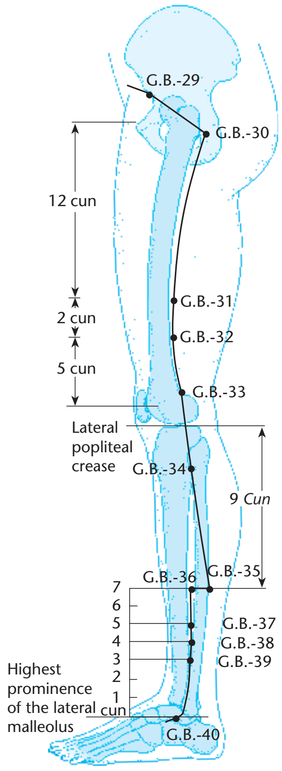
G.B.-29
G.B.-30
12 cun
G.B.-31
2 cun
G.B.-32
5 cun
G.B.-33
Lateral popliteal crease
G.B.-34
9 Cun
G.B.-36
G.B.-35
7
6
5
4
3
2
1
G.B.-37
G.B.-38
G.B.-39
Highest prominence of the lateral malleolus
cun
G.B.-40

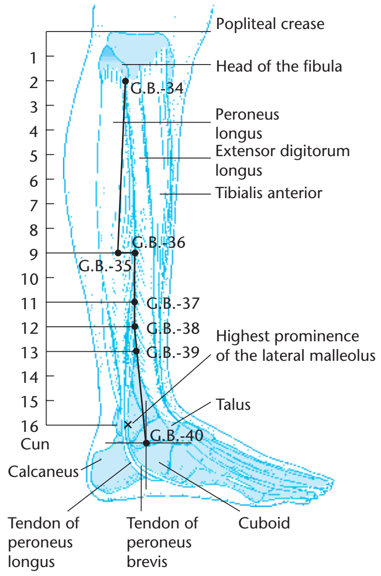
Popliteal crease
Head of the fibula
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Cun
G.B.-34
Peroneus longus
Extensor digitorum longus
Tibialis anterior
G.B.-36
G.B.-35
G.B.-37
G.B.-38
G.B.-39
Highest prominence of the lateral malleolus
Talus
G.B.-40
Calcaneus
Tendon of peroneus longus
Tendon of peroneus brevis
Cuboid

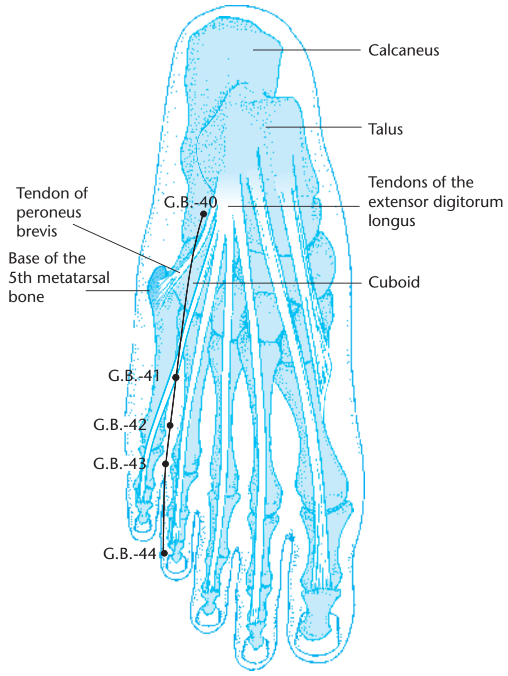
Calcaneus
Talus
Tendon of peroneus brevis
G.B.-40
Tendons of the extensor digitorum longus
Base of the 5th metatarsal bone
Cuboid
G.B.-41
G.B.-42
G.B.-43
G.B.-44

G.B.-4 to **G.B.-7**, and continues to → **T.B.-22** (*erheliao*) anterior to the root of the ear

- passes → **T.B.-20** (*jiaosun*) superior to the apex of the ear and curves posterior to the ear from **G.B.-8** to **G.B.-12**
- traverses the lateral aspect of the head to **G.B.-13** (*benshen*) and **G.B.-14** (*yangbai*) on the forehead
- again traverses the side of the head to **G.B.-20** (*fengchi*)
- crosses the superior aspect of the shoulder and passes **G.B.-21** (*jianjing*) and → **T.B.-15** (*tianliao*) to reach → **Du-14** (*dazhu*), where it meets the other Yang primary channels
- continues to **S.I.-12** (*bingfeng*), according to some authors (Deadman et al 1998) previously passing → **BL-11** (*dashu*) and **ST-12** (*quepen*) in the supraclavicular fossa.

One branch runs from **G.B.-20** to → **T.B.-17** (*yifeng*), enters the ear and travels to → **S.I.-19** (*tinggong*), passes → **ST-7** (*xiaguan*) and continues to **G.B.-1** (*tongziliao*) at the lateral canthus of the eye, descends to → **ST-5** (*daying*), again ascends to the infraorbital region, where it meets the Triple Burner primary channel; then descends to → **ST-6** (*jiache*) on the lower jaw, traverses the lateral aspect of the neck, passing → **ST-9** (*renying*) at the sternocleidomastoid, and meets the main branch again at

→ **ST-12** (*quepen*) in the supraclavicular fossa, where the latter divides into **2 branches**:

- **The inner branch** enters the thorax, meets the Pericardium primary channel at → **P-1** (*tianchi*), runs across the diaphragm, connects with the Liver (*gan*) and the Gall Bladder (*dan*), traverses the hypochondrium and continues to the lower abdomen, enters the lumbar region near the femoral artery, according to Solinas et al (1998) via → **ST-30** (*qichong*), and continues along the pubic hairline to the hip.
- **The outer branch** descends superficially from the supraclavicular fossa, crosses the midaxillary line and continues along the lateral thoracic wall, passing → **LIV-13** (*zhangmen*) at the free end of the 11th rib and continuing to the hip to **G.B.-29** (*juliao*). It then traverses the sacral region, passing → **BL-31** (*shangliao*), → **BL-32** (*ciliao*), → **BL-33** (*zhongliao*) and → **BL-34** (*xialiao*) over the sacral foramina and continues to → **Du-1** (*changqiang*). From there, it travels laterally to **G.B.-30** (*huantiao*), where it reconnects with the main branch (→ see comment below*).

The **external pathway** descends along the lateral aspect of the thigh to the knee and further along the anterior/posterior border of the fibula, passing the anterior aspect of the lateral malleolus and ending at the ulnar corner of the nail of the fourth toe.

A further branch originates at **G.B.-41** (*zulinqi*) on the dorsum of the foot and travels between the 1st and 2nd metatarsal bones to the tip of the big toe and **LIV-1** (*dadun*), where it joins the LIV channel.

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Alternating fevers and chills, headaches, eye pain, pain in the cheek and on the chin, subaxillary swellings, deafness, lateral knee and leg pain

Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain in the lateral costal region, vomiting, bitter taste in the mouth, thoracic pain

Connections of the Gall Bladder primary channel (→ 1.2)

Connections with other channels (→ 1.2)

Liver primary channel (*zu jue yin jing*)

Connection: Foot Yin–Yang pair of the third great circuit
Location: G.B.-41 → LIV-1 (on the foot)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Triple Burner primary channel (*shou shao yang jing*)

Connection: Hand–foot pairing of the third great circuit: *shaoyang* (Yang axes)
Location: T.B.-23 (or **T.B.-22**, according to some authors) → **G.B.-1** (on the head)
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with *zangfu*-Organ systems

Liver (*gan*), Gall Bladder (*dan*)

4.11.2 The Gall Bladder Divergent Channel (*zu shao yang jing bie*)

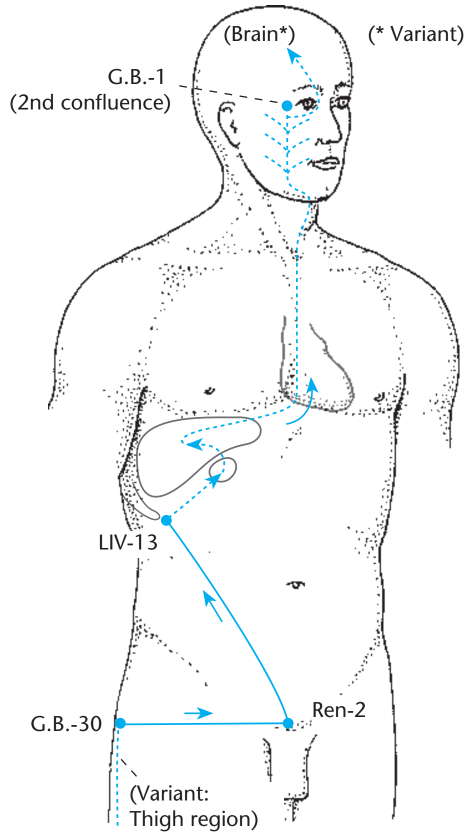

Pathway

The Gall Bladder Divergent channel separates from the Gall Bladder primary channel near **G.B.-30** (*huantiao*); according to some authors, it separates on the thigh (→ variant in the illustration)

- traverses the hip region to the anterior aspect of the body
- reaches the inguinal region and crosses the Liver divergent channel at → **Ren-2** (*qugu*)
- ascends the abdomen to **LIV-13** (*zhangmen*) below the free end of the 11th rib
- continues internally along the thoracic wall
- connects with the Gall Bladder (*dan*) and the Liver (*gan*) and passes the Heart (*xin*)
- ascends to the neck
- emerges on the lower jaw

---

*Not all authors describe the pathway of the G.B. channel as passing the points **BL-31** to **BL-34** and **Du-1** (see figure for variant pathway).

➡ disperses on the face

➡ connects with the Gall Bladder primary channel and the Liver divergent channel to form one of the six *he*-confluences (here: G.B./LIV as 2nd confluence → 1.3)

➡ continues to the eye system and ends in the Brain.

Clinical importance

- Strengthens the relationship between the Gall Bladder and the Liver (*zangfu*-Organ systems). Points on the Gall Bladder primary channel can therefore be used to treat disorders of the Liver and vice versa.
- Supports the relationship between the oesophagus and the Heart.
- The Gall Bladder Divergent channel covers the area around the eyes, thus supporting the action of some G.B. points for disorders of the eyes.

4.11.3 The Gall Bladder Sinew Channel (*zu shao yang jing jin*)

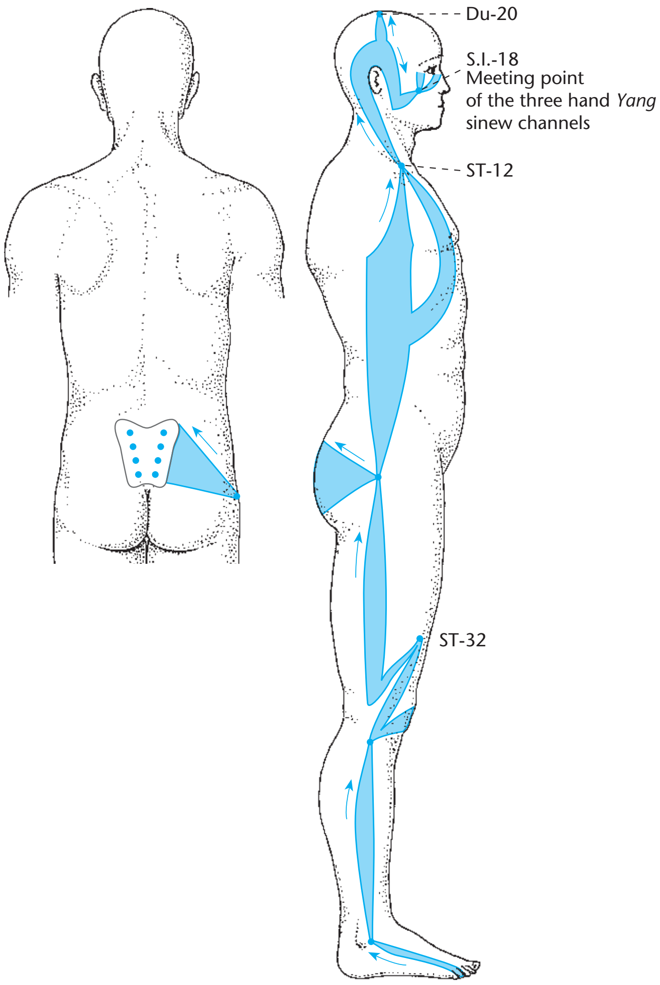

Pathway

The Gall Bladder sinew channel begins at the 4th toe, binds (*jie*) anterior and inferior to the lateral malleolus in the area of **G.B.-40** (*qiuxu*), follows the lateral aspect of the leg, binds (*jie*) at the fibula and the lateral aspect of the knee.

From the fibula, it continues to ascend the lateral aspect of the thigh

➡ while a **branch** runs obliquely to → **ST-32** (*futu*), where it binds (*jie*).

The main branch further ascends the leg to bind (*jie*) at the greater trochanter. Here, a **branch** separates and disperses over the gluteal and sacral regions.

From the hip, the channel continues to ascend the flanks to the lower costal region, where it divides into **two branches**:

➡ **one branch** traverses the lateral aspect of the thorax and ascends to bind (*jie*) at → **ST-12** (*quepen*)

➡ **the other branch** follows the midaxillary line to the thorax, meeting the other branch in the supraclavicular fossa.

From **ST-12**, the channel ascends the lateral aspect of the neck

➡ curves around the ear

➡ a branch continues to the apex of the ear and to → **Du-20** (*baihui*)

➡ from the temporal region, **a branch** descends and traverses the cheek to the lower jaw

➡ crosses the zygomatic arch to reach **S.I.-18** (*quanliao*), where it meets the other foot Yang sinew channels and divides into **two branches**. One branch travels laterally to the root of the nose, the other to the outer canthus of the eye.

Clinical importance

Pathology: Stiffness and distending sensations in the area of the 4th toe as well as on the lateral aspect of the knee. Limited range of motion of the knee joint. Pain, tension and distending sensations in the popliteal crease radiating to the thigh and the sacral region and vice versa. Pain and distending sensations in the sacral region radiating to the hypochondrium and vice versa. Pain and tension in the supraclavicular fossa, thorax, breast and neck region. Pain along the left side of the channel and inability to open the right eye and vice versa.

Indication: Predominantly for pain, muscle tension, stiff joints and distending sensations on the lateral aspect of the body. Used for disorders of the anterior, lateral and posterior aspects of the legs, especially those affecting the knees and the thighs. The Gall Bladder Sinew channel spreads in the thorax and the breasts so that points on the G.B. channel can be used for disorders in those areas. Also for headaches on the vertex and temporal regions. The channel also reaches the lateral aspect of the nose, which supports the relationship with the nose. Therefore, G.B. points can be used for chronic nasal disorders, chronic sinusitis, etc.

Ch04.11-F10028.qxd 2/23/08 11:23 AM Page 395

4.11.4 The Gall Bladder *luo*-Connecting Channel (*zu shao yang luo mai*)

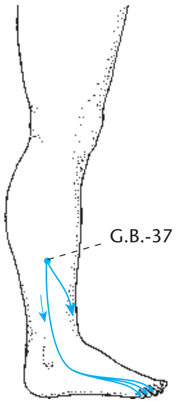

Pathway

The Gall Bladder *luo*-connecting channel separates from the Gall Bladder primary channel at its *luo*-connecting point **G.B.-37** (*guangming*) (→ 8.1.2). It forms a three-dimensional reticulate network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Liver primary channel; according to some schools of thought (for example, Ngyen Van Nghi → Appendix), they travel as a **transverse** G.B. *luo* vessel to the *yuan*-source point **LIV-3** (*taichong*).
- A **longitudinal** division descends to the dorsum of the foot and spreads to the 3rd, 4th and 5th toes.

Clinical importance (→ 8.1.2)

Pathology

Excess (*shi*): Cold sensations in the feet

Deficiency (*xu*): Weakness, weak musculature of the foot, including difficulty in standing, paralysis of the lower extremity

4.11.5 Cutaneous Region (*shao yang pi bu*)

See description and figures → 1.6.

4.11.6 Points of the Gall Bladder Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): G.B.-40** (*qiuxu*) 🟦■
- ***Luo*-connecting point (→ 8.1.2.): G.B.-37** (*guangming*) 🟦■
- ***Xi*-cleft point (→ 8.1.3): G.B.-36** (*waiqiu*)
- **Associated Back-*shu* point (→ 8.1.4): BL-19** (*danshu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): G.B.-24** (*riyue*) 🟦■
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Metal): **G.B.-44** (*zuqiaoyin*) 🟦■
  ying-spring point (Water), tonification point: **G.B.-43** (*xiaxi*)
  shu-stream point (Wood), *ben* point (Five Phases): **G.B.-41** (*zulinqi*) 🟦■
  jing-river point (Fire), sedation point: **G.B.-38** (*yangfu*)
  he-sea point (Earth): **G.B.-34** (*yanglingquan*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7)**
  – with the sinews: **G.B.-34** (*yanglingquan*) 🟦🟦
  – with the Marrow: **G.B.-39** (*xuanzhong*) 🟦■
- **Opening point (→ 8.1.8) of the *dai mai:* G.B.-41** (*zulinqi*) 🟦■
- **Lower *he*-sea point (→ 8.1.9) of the Gall Bladder: G.B.-34** (*yanglingquan*) 🟦🟦
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the S.I. and T.B. channels: **G.B.-1** (*tongziliao*)
  – with the T.B. and ST channels: **G.B.-3** (*shangguan*), **G.B.-4** (*hanyan*)
  – with the T.B., L.I. and ST channels: **G.B.-5** (*xuanlu*), **G.B.-6** (*xuanli*)
  – with the BL channel: **G.B.-7** (*qubin*); **G.B.-8** (*shuaigu*) 🟦■, **G.B.-9** (*tianchong*), **G.B.-10** (*fubai*)
  – with the BL, (S.I., T.B.*) channels: **G.B.-11** (*touqiaoyin*)
  – with the BL channel: **G.B.-12** (*wangu*) 🟦■
  – with the *yang wei mai*: **G.B.-13** (*benshen*) 🟦■
  – with the *yang wei mai*, (T.B., ST, L.I. channels*): **G.B.-14** (*yangbai*) 🟦■
  – with the *yang wei mai* and the BL channel: **G.B.-15** (*toulinqi*)
  – with the *yang wei mai*: **G.B.-16** (*muchuang*). **G.B.-17** (*zhengying*), **G.B.-18** (*chengling*), **G.B.-19** (*naokong*)
  – with the *yang wei mai, yang qiao mai* and (T.B. channel*): **G.B.-20** (*fengchi*) 🟦🟦
  – with the *yang wei mai*, T.B. and (ST*) channels: **G.B.-21** (*jianjing*) 🟦■
  – with the BL channel*: **G.B.-23** (*zhejin*)
  – with the SP channel and (*yang wei mai**): **G.B.-24** (*riyue*) 🟦■
  – with the *dai mai*: **G.B.-26** (*daimai*), **G.B.-27** (*wushu*), **G.B.-28** (*weidao*)
  – with the *yang qiao mai* and (*dai mai**): **G.B.-29** (*juliao*)
  – with the BL channel: **G.B.-30** (*huantiao*) 🟦■
  – with the *yang wei mai**: **G.B.-35** (*yangjiao*)
  – of other channels with the G.B. channel: **ST-7, ST-8, ST-9, ST-12, S.I.12, S.I.-19, (BL-1, BL-11*), BL-31–34, P-1, (T.B.-15*), T.B-17, T.B.-20, T.B.-22, LIV-13, Du-1, Du-14, (Du-20, ST-5, ST-6, ST-30*)**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): G.B.-9** (*tianchong*)
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star points (→ 8.1.14): G.B.-30** (*huantiao*) 🟦■, **G.B.-34** (*yanglingquan*) 🟦🟦

---

* Mentioned by only some authors.

- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points:**
  - – main point for all disorders of the head, sensory organs, Brain: **G.B.-20** (*fengchi*) 🟦🟦
  - – front-*mu* point of the Kidneys (→ 8.1.5): **G.B.-25** (*jingmen*) 🟦■
  - – *xi*-cleft point (→ 8.1.3) of the *yang wei mai:* **G.B.-35** (*yangjiao*)

Points according to region

- **Local points (→ 8.2.1):** forehead – **G.B.-14** (*yangbai*) 🟦■; head, temporal region – **G.B.-8** (*shuaigu*) 🟦■; head, occipital region – **G.B.-20** (*fengchi*) 🟦🟦; ears – **G.B.-2** (*tingshui*) 🟦■; cervical spine – **G.B.-20** (*fengchi*) 🟦🟦; neck and shoulder region – **G.B.-20** (*fengchi*) 🟦■, **G.B.-21** (*jianjing*) 🟦■; Gall Bladder – **G.B.-24** (*riyue*) 🟦■; hypochondrium – **G.B.-25** (*jingmen*) 🟦■, **G.B.-26** (*daimai*), **G.B.-27** (*wushu*); hip – **G.B.-29** (*juliao*), **G.B.-30** (*huantiao*) 🟦■; knee – **G.B.-34** (*yanglingquan*) 🟦🟦, **G.B.-35** (*yangjiao*); foot – **G.B.-40** (*qiuxu*) 🟦■
- **Adjacent points (→ 8.2.1):** head, temporal region – **G.B.-20** (*fengchi*) 🟦🟦; ears – **G.B.-8** (*shuaigu*) 🟦■, **G.B.-20** (*fengchi*) 🟦🟦; cervical spine and shoulders – **G.B.-21** (*jianjing*) 🟦■; lumbar region – **G.B.-25** (*jingmen*) 🟦■, **G.B.-30** (*huantiao*) 🟦■; hip – **G.B.-31** (*fengshi*); knee – **G.B.-33** (*xiyangguan*); foot – **G.B.-34** (*yanglingquan*) 🟦🟦; toes – **G.B.-38** (*yangfu*)
- **Distal points (→ 8.2.1):** head, temporal region – **G.B.-43** (*xiaxi*), **G.B.-41** (*zulinqi*) 🟦■; eyes – **G.B.-37** (*guangming*) 🟦■; ears – **G.B.-41** (*zulinqi*) 🟦■; Liver – **G.B.-34** (*yanglingquan*) 🟦🟦; Gall Bladder – **G.B.-34** (*yanglingquan*) 🟦🟦, **G.B.-40** (*qiuxu*) 🟦■; hypochondrium – **G.B.-38** (*yangfu*), **G.B.-34** (*yanglingquan*) 🟦🟦, **G.B.-43** (*xiaxi*); cervical spine and hip – **G.B.-39** (*xuanzhong*) 🟦■; hip – **G.B.-41** (*zulinqi*) 🟦■

Specific points according to the channel pathway (in numeric order)

- **G.B.-1** (*tongziliao*): *jiaohui*-meeting point with the S.I. and T.B. channels (→ 8.1.10)
- **G.B.-2** (*tinghui*) 🟦■: local point for the ears (→ 8.2.1)
- **G.B.-3** (*shangguan*): *jiaohui*-meeting point with the T.B. and ST channels (→ 8.1.10)
- **G.B.-4** (*hanyan*): *jiaohui*-meeting point with the T.B. and ST channels (→ 8.1.10)
- **G.B.-5** (*xuanlu*): *jiaohui*-meeting point with the T.B., L.I. and ST channels (→ 8.1.10)
- **G.B.-6** (*xuanli*):): *jiaohui*-meeting point with the T.B., L.I. and ST channels (→ 8.1.10)
- **G.B.-7** (*qubin*):): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **G.B.-8** (*shuaigu*) 🟦■: *jiaohui*-meeting point with the BL channel (→ 8.1.10); local point for the lateral aspect of the head (→ 8.2.1); adjacent point for the ears (→ 8.2.1)
- **G.B.-9** (*tianchong*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)*; Window of Heaven point (→ 8.1.12)*
- **G.B.-10** (*fubai*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **G.B.-11** (*touqiaoyin*): *jiaohui*-meeting point with the BL, (S.I., T.B.)* channels (→ 8.1.10)
- **G.B.-12** (*wangu*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **G.B.-13** (*benshen*) 🟦■: *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-14** (*yangbai*) 🟦■: *jiaohui*-meeting point with the *yang wei mai* and (T.B., ST, L.I.*) channels (→ 8.1.10); local point for the forehead (→ 8.2.1)
- **G.B.-15** (*toulinqi*): *jiaohui*-meeting point with the *yang wei mai* and BL channel (→ 8.1.10)
- **G.B.-16** (*muchuang*): *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-17** (*zhengying*): *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-18** (*chengling*): *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-19** (*naokong*): *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-20** (*fengchi*) 🟦🟦: *jiaohui*-meeting point with the *yang wei mai, yang qiao mai* and (T.B. channel*) (→ 8.1.10); local point for the occiput (→ 8.2.1); local point for the neck and shoulders (→ 8.2.1); local point for all disorders of the head, sensory organs, Brain; adjacent point for the lateral aspect of the head and the ears (→ 8.2.1)
- **G.B.-21** (*jianjing*) 🟦■: *jiaohui*-meeting point with the *yang wei mai*, T.B. and (ST*) channels (→ 8.1.10); adjacent point for the neck and shoulders (→ 8.2.1)
- **G.B.-23** (*zhejin*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **G.B.-24** (*riyue*) 🟦■: associated Front-*mu* point (→ 8.1.5); *jiaohui*-meeting point with the SP channel and (*yang wei mai**) (→ 8.1.10); local point for the Gall Bladder (→ 8.2.1)
- **G.B.-25** (*jingmen*) 🟦■: Front-*mu* point of the Kidneys (→ 8.1.5); local point for the hypochondrium (→ 8.2.1); adjacent point for the lumbar region (→ 8.2.1)
- **G.B.-26** (*daimai*): *jiaohui*-meeting point with the *dai mai* (→ 8.1.10); local point for the hypochondrium (→ 8.2.1)
- **G.B.-27** (*wushu*): *jiaohui*-meeting point with the *dai mai* (→ 8.1.10); local point for the hypochondrium (→ 8.2.1)
- **G.B.-28** (*weidao*): *jiaohui*-meeting point with the *dai mai* (→ 8.1.10);
- **G.B.-29** (*juliao*): *jiaohui*-meeting point with the *yang qiao mai* and (*dai mai**) (→ 8.1.10); local point for the hip (→ 8.2.1)
- **G.B.-30** (*huantiao*) 🟦■: *jiaohui*-meeting point with the BL channel (→ 8.1.10); Ma Dan Yang Heavenly Star point (→ 8.1.14); adjacent point for the lumbar region and the hip (→ 8.2.1); local point for the hip (→ 8.2.1)
- **G.B.-31** (*fengshi*): adjacent point for the hip (→ 8.2.1)
- **G.B.-33** (*xiyangguan*): adjacent point for the knee (→ 8.2.1)

- **G.B.-34** (*yanglingquan*) 🟦🟦: *he*-sea point (Earth) (→ 8.1.6); *hui*-meeting point with the sinews (→ 8.1.7); lower *he*-sea point of the Gall Bladder (→ 8.1.9); Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the Liver, Gall Bladder and hypochondrium (→ 8.2.1); local point for the knee (→ 8.2.1); adjacent point for the foot (→ 8.2.1)
- **G.B.-35** (*yangjiao*): *xi*-cleft point of the *yang wei mai* (→ 8.1.3); *jiaohui*-meeting point with the *yang wei mai** (→ 8.1.10); local point for the knee (→ 8.2.1)
- **G.B.-36** (*waiqiu*): *xi*-cleft point (→ 8.1.3)
- **G.B.-37** (*guangming*) 🟦■: *luo*-connecting point (→ 8.1.2); distal point for the eyes (→ 8.2.1)
- **G.B.-38** (*yangfu*): *jing*-river point (Fire) (→ 8.1.6); sedation point; distal point for the hypochondrium (→ 8.2.1)
- **G.B.-39** (*xuanzhong*) 🟦■: *hui*-meeting point with the Marrow (→ 8.1.7); distal point for the cervical spine and the hip (→ 8.2.1)
- **G.B.-40** (*qiuxu*) 🟦■: *yuan*-source point (→ 8.1.1); local point for the foot (→ 8.2.1); distal point for the Gall Bladder (→ 8.2.1)
- **G.B.-41** (*zulinqi*) 🟦■: *shu*-stream point (Wood); *ben* point (Five phases) (→ 8.1.6); Opening point (→ 8.1.8) of the *dai mai*; distal point for the temporal region and the ears (→ 8.2.1); distal point for the hip (→ 8.2.1)
- **G.B.-43** (*xiaxi*): *ying*-spring point (Water) (→ 8.1.6); tonification point; distal point for the lateral aspect of the head and the hypochondrium (→ 8.2.1)
- **G.B.-44** (*zuqiaoyin*) 🟦■: *jing*-well point (Metal) (→ 8.1.6)

Pupil Crevice TONGZILIAO G.B.-1

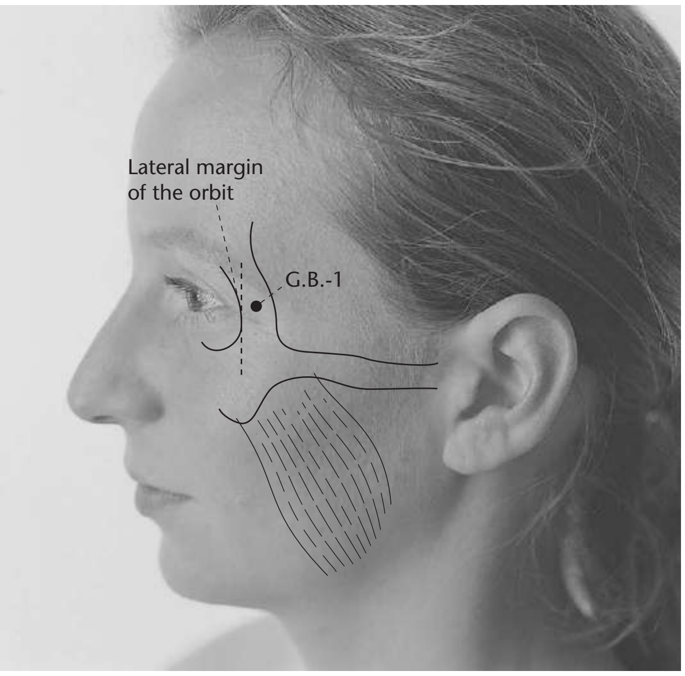
Lateral margin
of the orbit
G.B.-1

Location

In a bony depression on the lateral aspect of the orbital margin, on the level of the outer canthus of the eye.

How to find

From the outer canthus of the eye, palpate in a lateral direction. Locate **G.B.-1** on the outer aspect of the orbit (diagonally inferior to the temple), where a bony depression can be palpated.
→ **T.B.-23** is located more superiorly on the lateral end of the eyebrow, in a depression on the frontozygomatic suture. → **BL-1** is located at the inner canthus of the eye.

Needling

0.2–0.3 cun obliquely in a posterior direction or up to 1 cun transversely (subcutaneously) towards → **Ex-HN-5** (*taiyang*). According to some texts, moxibustion is contraindicated.

Actions/Indications

- Expels Wind and Heat from the eyes

Special features

Meeting point with the S.I. and T.B. channels, entry point.

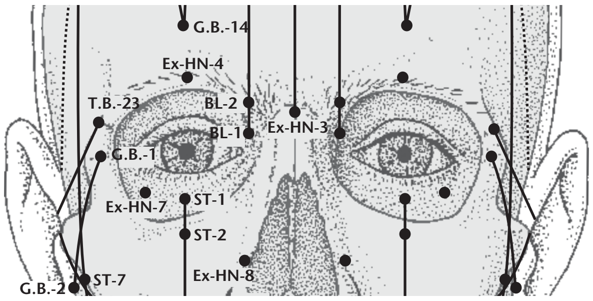
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
G.B.-2
ST-7
Ex-HN-8

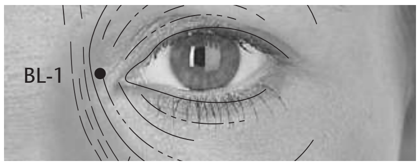

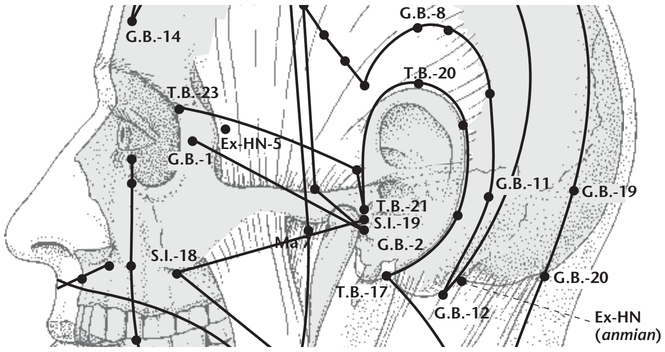
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
Ma 7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)

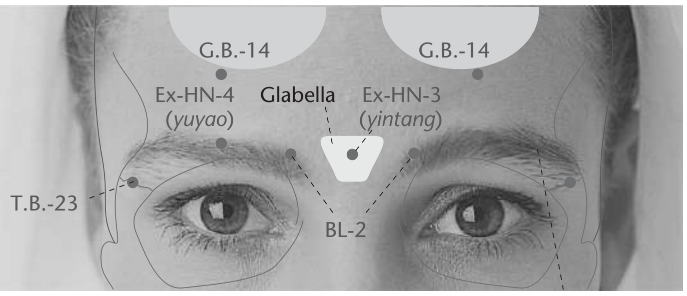
G.B.-14
G.B.-14
Ex-HN-4
(*yuyao*)
Glabella
Ex-HN-3
(*yintang*)
T.B.-23
BL-2

G.B.-2 *Meeting of Hearing* TINGHUI

Location

Anterior to the ear, with the patient's mouth open in the depression on the level of the intertragic notch, on the lower border of the condyloid process of the mandible.

How to find

Locate and needle with the patient's mouth open as this allows the condyloid process of the mandible to slide towards the anterior, revealing the depression where **G.B.-2** is located. Find the vertical sulcus anterior to the ear (at the ear/cheek junction), which may be more or less pronounced. Then locate **G.B.-2** on the sulcus, on the level of the intertragic notch. If the sulcus is not clearly defined (it becomes more clearly visible with increasing age), the depression can be located by using an (ear) point locator. **G.B.-2** is the most distal of three points located anterior to the ear (→ **S.I.-19** and → **T.B.-21** are both located more superiorly).

Needling

0.5–1 cun vertically or slightly obliquely in an inferior direction. Needle with the patient's mouth open to avoid intra-articular insertion. Caution: Like **T.B.-21** and **S.I.-19**, this point is located close to the superficial temporal artery and the auriculotemporal nerve.

Actions/Indications

- Eliminates Wind, clears Heat, benefits the ears and the temporomandibular joint
- Opens the channel and the *luo*-connecting vessel, alleviates pain

Special features

Important local point for disorders of the ears and the jaw. This point is often used alternately with **T.B.-21** and **S.I.-19**.

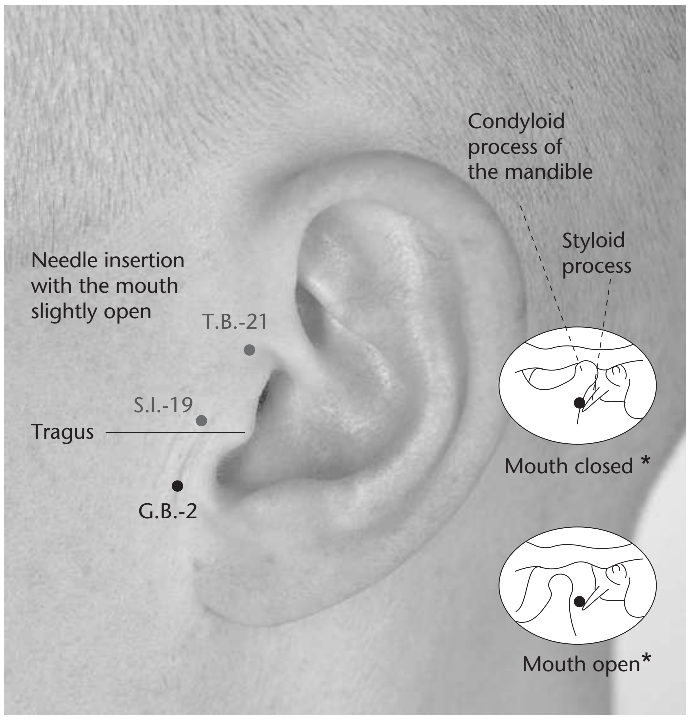
Condyloid process of the mandible
Styloid process
Needle insertion with the mouth slightly open
T.B.-21
S.I.-19
Tragus
G.B.-2
Mouth closed *
Mouth open*

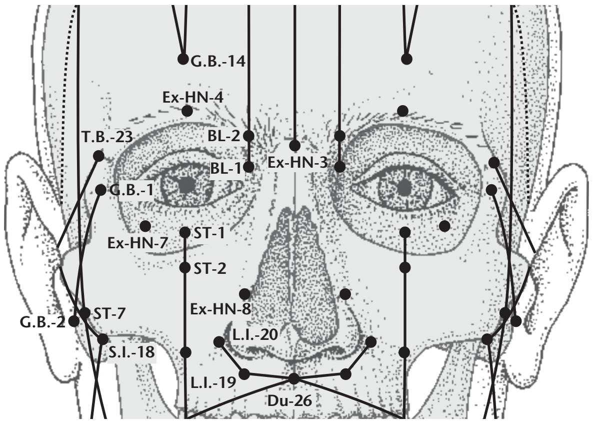
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN-8
G.B.-2
ST-7
L.I.-20
S.I.-18
L.I.-19
Du-26

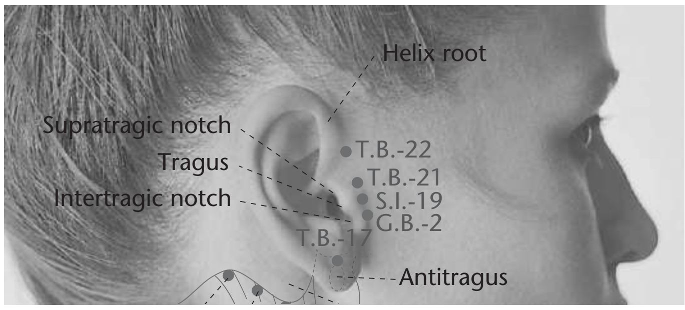
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Antitragus

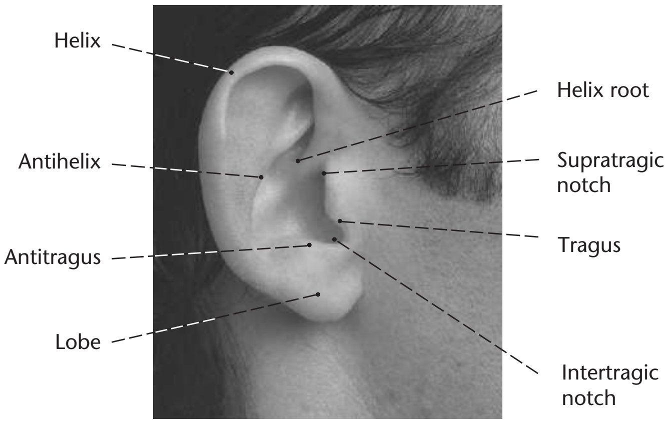
Helix
Antihelix
Antitragus
Lobe
Helix root
Supratragic notch
Tragus
Intertragic notch

---

* According to Deadman et al 1998.

Above the Joint SHANGGUAN G.B.-3

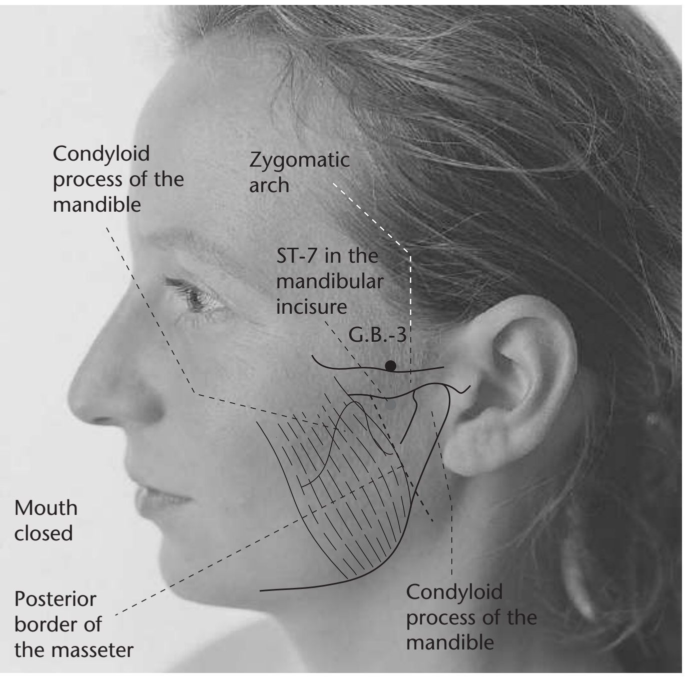
Condyloid
process of the
mandible
Zygomatic
arch
ST-7 in the
mandibular
incisure
G.B.-3
Mouth
closed
Posterior
border of
the masseter
Condyloid
process of the
mandible

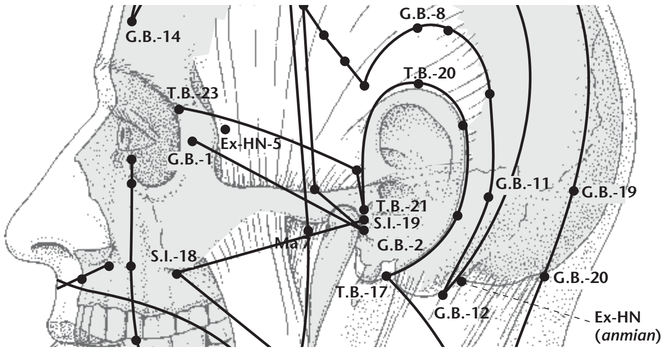
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
Ma
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)

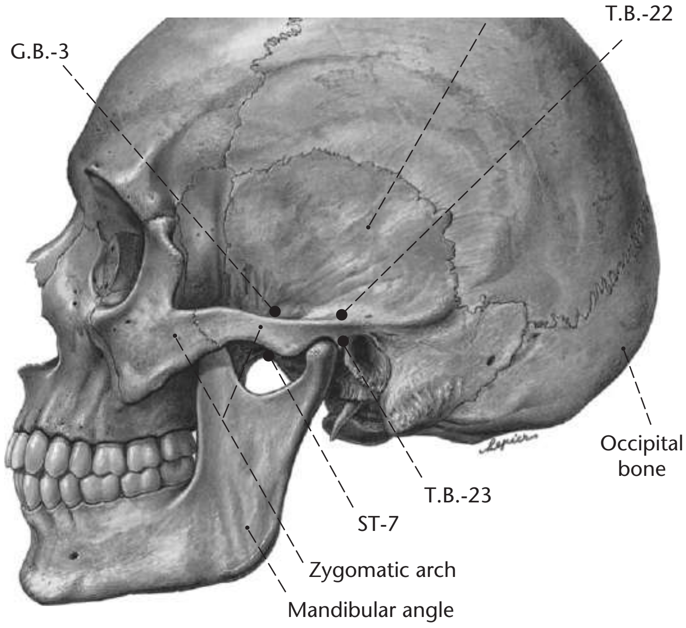
T.B.-22
G.B.-3
Occipital
bone
T.B.-23
ST-7
Zygomatic arch
Mandibular angle

Location

In a depression on the upper border of the zygomatic arch, approximately 1 cun anterior to the root of the ear, superior to → **ST-7**.

How to find

Locate the zygomatic arch (→ 3.1.2) by palpating from the root of the ear (root of the helix) approximately 1 cun towards the orbit. Follow its course by placing one finger above and the other finger below the arch. As soon as the lower finger has reached a clearly palpable depression anterior to the temporomandibular joint and posterior to the masseter muscle (→ **ST-7**), the upper finger will be resting on **G.B.-3**, which is located directly superior to → **ST-7** in a shallow depression on the superior border of the zygomatic arch.

Needling

Vertically 0.3–0.5 cun, no strong stimulation. Caution: Branches of the temporal, transverse facial and masseteric arteries! Traditionally, deep needling is prohibited.

Actions/Indications

- Opens the channel, alleviates pain
- Eliminates Wind
- Benefits the ears

Special features

Meeting point with the T.B. and ST channels

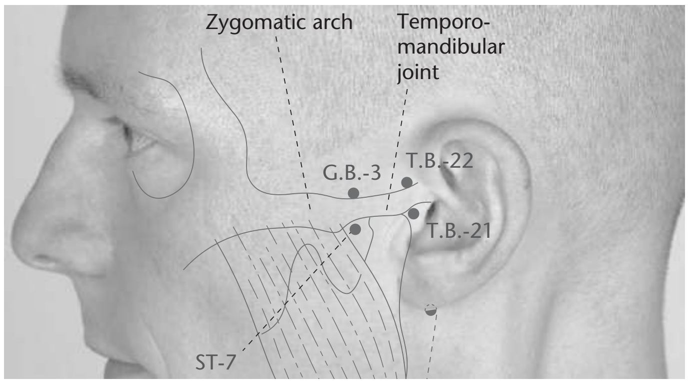
Zygomatic arch
Temporo-
mandibular
joint
G.B.-3
T.B.-22
T.B.-21
ST-7

G.B.-4 *Jaw Serenity* HANYAN

Location

At the junction of the upper quarter and the second quarter of a line connecting ➞ **ST-8** and ➞ **G.B.-7**.

How to find

First, locate the two reference points: ➞ **ST-8** (4.5 cun lateral to the midline and 0.5 cun within the anterior hairline, at the corner of the forehead) and ➞ **G.B.-7** (in the depression on the level of the apex of the ear, within the circumauricular temporal hairline). Then divide the slightly curved line between these two points into quarters and locate **G.B.-4** at the junction of the upper quarter with the lower three quarters. Generally, **G.B.-4** is located on the temporal hairline and the anterior portion of the temporalis muscle, which can be felt when chewing.

Needling

0.3–1.5 cun transversely (subcutaneously), tangentially along the skull, in the direction of the occiput or towards the disorder/pain.

Actions/Indications

- Eliminates Wind and Heat
- Opens the channel

Special features

Meeting point with the T.B. and ST channel

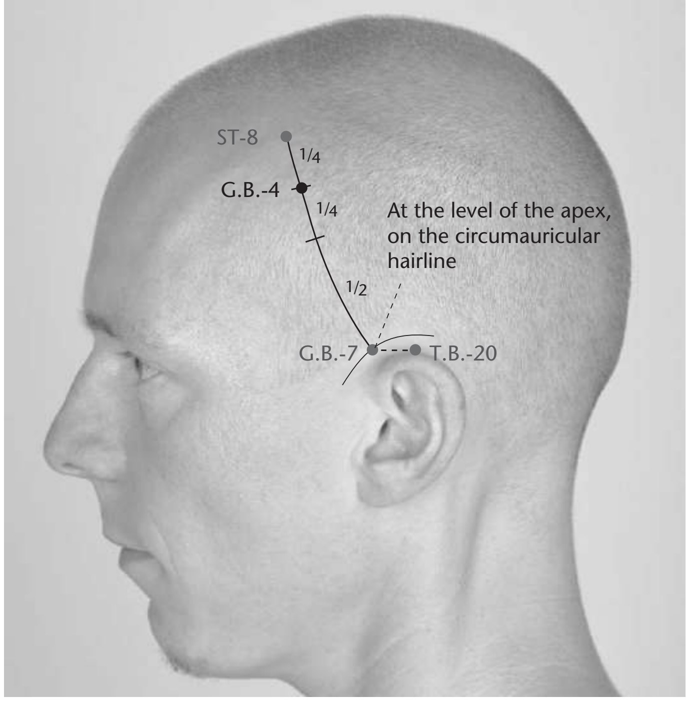
ST-8
1/4
G.B.-4
1/4
At the level of the apex, on the circumauricular hairline
1/2
G.B.-7
T.B.-20

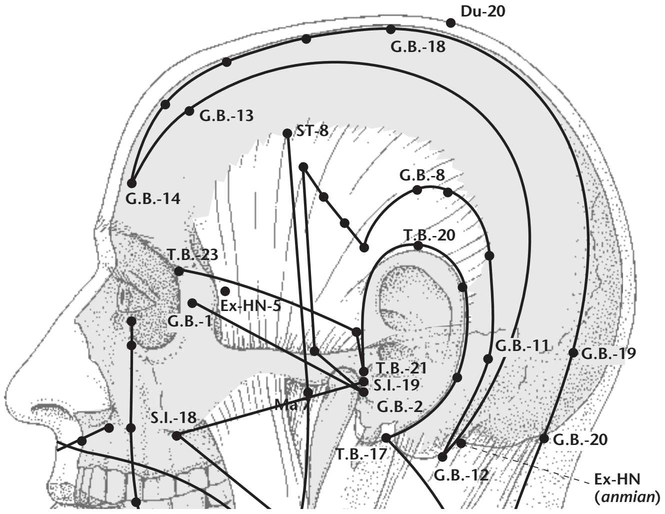
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
Ma
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)

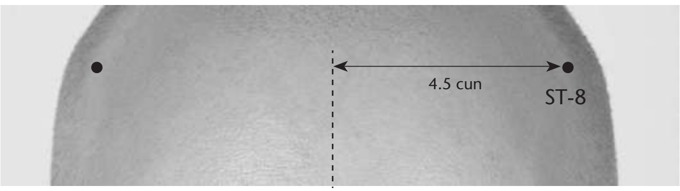

Suspended Skull XUANLU **G.B.-5**

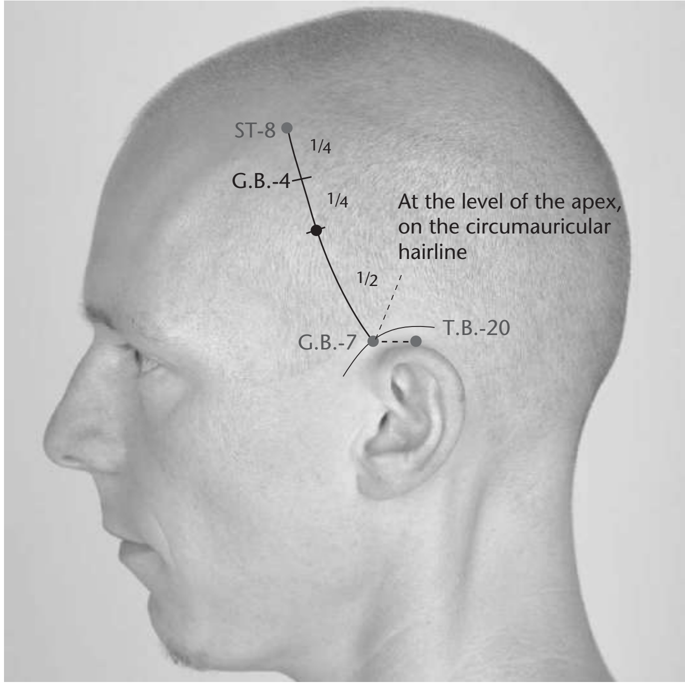
ST-8
1/4
G.B.-4
1/4
At the level of the apex, on the circumauricular hairline
1/2
G.B.-7
T.B.-20

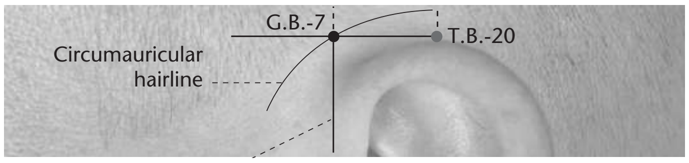
G.B.-7
T.B.-20
Circumauricular hairline

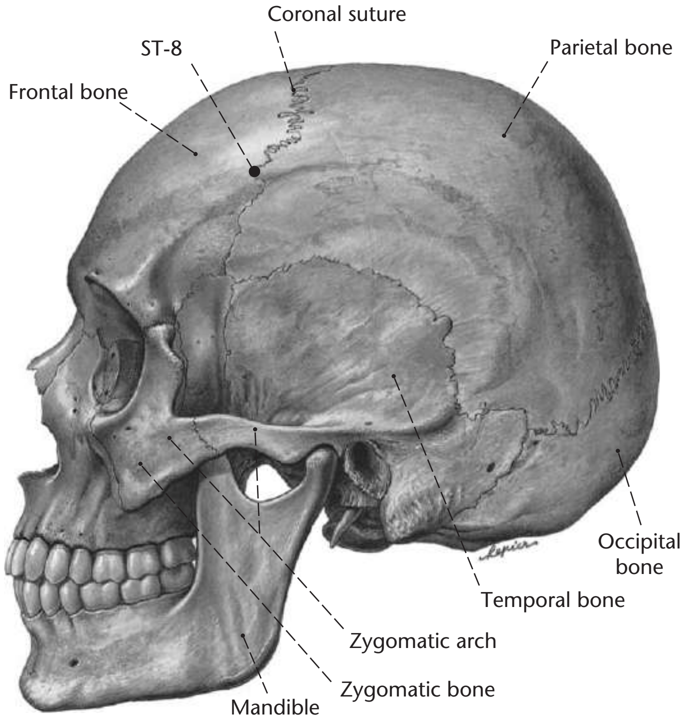
Coronal suture
ST-8
Parietal bone
Frontal bone
Occipital bone
Temporal bone
Zygomatic arch
Zygomatic bone
Mandible

Location

At the junction of the 2nd and 3rd quarter of an imaginary line connecting → **ST-8** and → **G.B.-7**.

How to find

First, locate the two reference points: → **ST-8** (4.5 cun lateral to the midline and 0.5 cun within the anterior hairline, at the corner of the forehead) and → **G.B.-7** (in the depression on the level of the apex of the ear, within the circumauricular temporal hairline). Then divide the slightly curved line between these two points into quarters and locate **G.B.-5** at the junction of the upper two quarters with the lower two quarters. Generally, this point is located on the level of the parietal suture, just within the hairline.

Needling

0.3–1.5 cun transversely (subcutaneously), tangentially along the skull, in the direction of the occiput or towards the disorder/pain.

Actions/Indications

- Expels Wind and Heat
- Opens the channel

Special features

Meeting point with the T.B., L.I. and ST channels

Ch04.11-F10028.qxd 2/23/08 11:24 AM Page 404

G.B.-6 *Suspended Hair* XUANLI

Location

At the junction of the lower quarter and the upper three quarters of a line connecting → **ST-8** and → **G.B.-7**.

How to find

First, locate the two reference points: → **ST-8** (4.5 cun lateral to the midline and 0.5 cun within the anterior hairline, at the corner of the forehead) and → **G.B.-7** (in the depression on the level of the apex of the ear, within the circumauricular temporal hairline). Then divide the slightly curved line between these two points into quarters and locate **G.B.-6** at the junction of the upper three quarters with the lower quarter.

Needling

0.3–1.5 cun transversely (subcutaneously) in the direction of the occiput or towards the disorder/pain.

Actions/Indications

- Expels Wind and Heat
- Opens the channel

Special features

Meeting point with the L.I., ST and T.B. channels

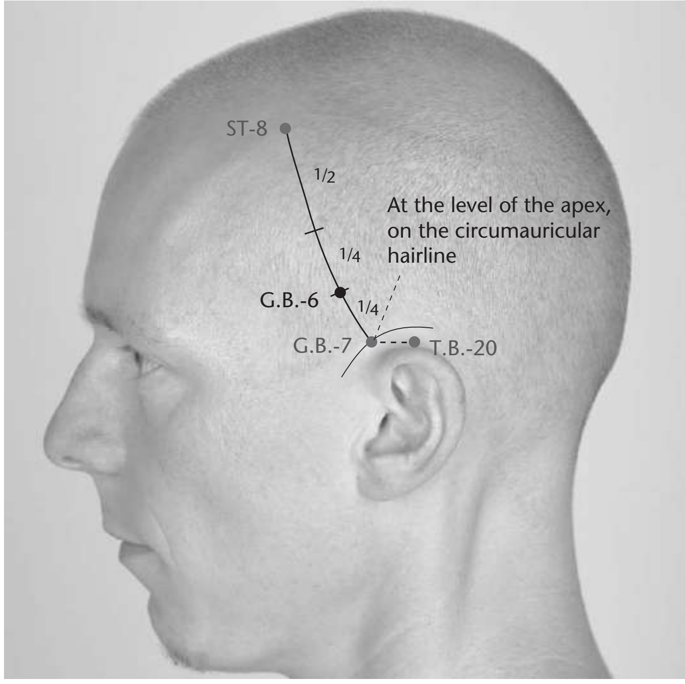
ST-8
1/2
1/4
1/4
At the level of the apex, on the circumauricular hairline
G.B.-6
G.B.-7
T.B.-20

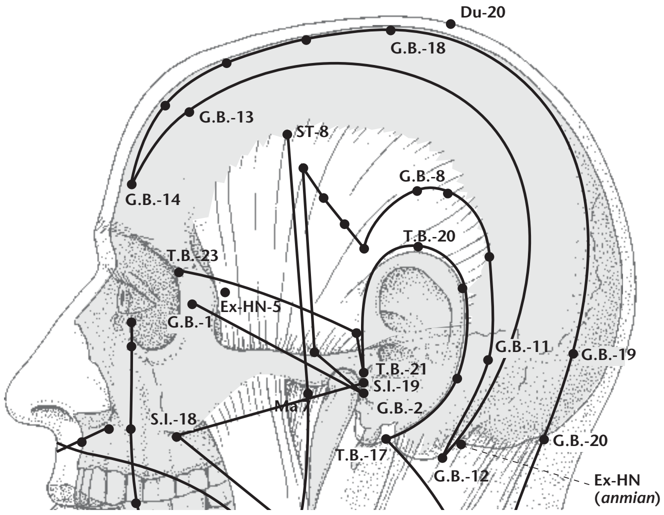
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
Ma 7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(anmian)

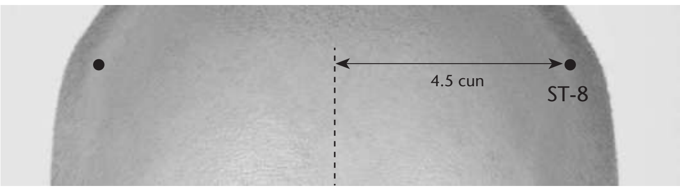

Crook of the Temple QUBIN G.B.-7

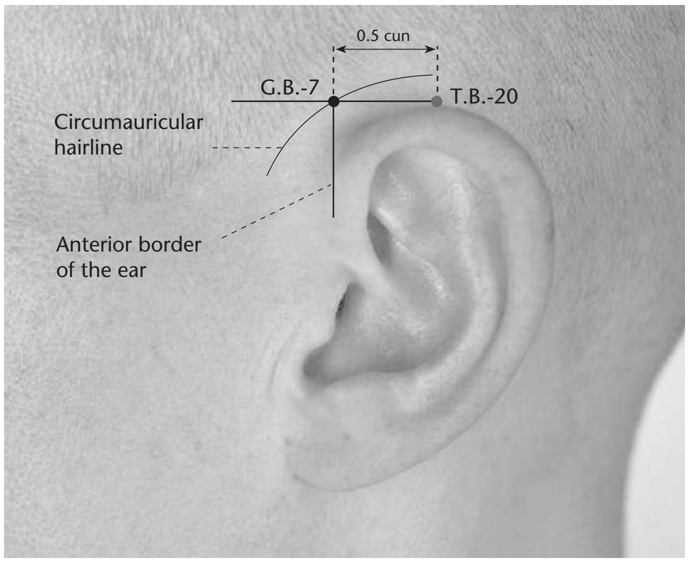

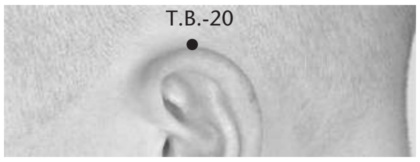

Location

In a depression on the level of the apex of the ear, within the circumauricular temporal hairline, approximately at the junction of a horizontal line through the apex of the ear and a vertical line along the posterior border of the temple anterior to the ear.

How to find

Locate **G.B.-7** by palpating for a small depression anterior to the apex of the ear at the circumauricular temporal hairline. The point is located approximately at a junction between a horizontal line through the apex of the ear and a vertical line along the posterior border of the temple anterior to the ear.
For orientation: → **T.B.-20** is located directly superior to the apex of the ear.

Needling

0.3–1.5 cun transversely (subcutaneously) in the direction of the occiput or towards the site of the disorder/pain.

Actions/Indications

- Expels Wind
- Benefits the mouth and jaw

Special features

Meeting point with the BL channel

G.B.-8 *Leading Valley* SHUAIGU

Location

1.5 cun directly superior to the apex of the ear, in a depression on the upper border of the temporalis muscle.

How to find

Locate the apex of the ear, which becomes more clearly defined by folding the auricle towards the anterior so that the posterior part of the upper helix covers its anterior part. From the apex of the ear, measure 1.5 cun (2 fingerbreadths) in a superior direction. There, the palpating finger will glide into a bony depression (**G.B.-8**) which is often sensitive to pressure. For orientation: When a chewing movement is made, this can just about be felt at this point.

→ **G.B.-9** is located on the same level, 0.5 cun dorsal to **G.B.-8**.

Needling

0.3–1.5 cun transversely (subcutaneously), mostly from anterior to posterior or towards the site of the pain.

Actions/Indications

- Expels Wind, benefits the head and ears, alleviates pain, harmonises the diaphragm and Stomach

Special features

Meeting point with the BL channel. Important local point for parietal and temporal headaches.

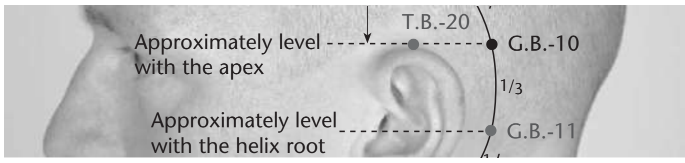
Approximately level with the apex
T.B.-20
G.B.-10
1/3
Approximately level with the helix root
G.B.-11

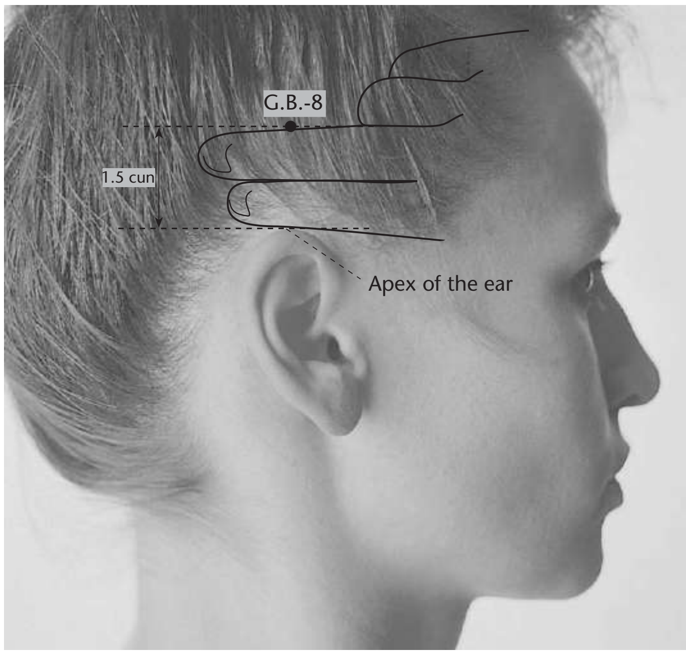
G.B.-8
1.5 cun
Apex of the ear

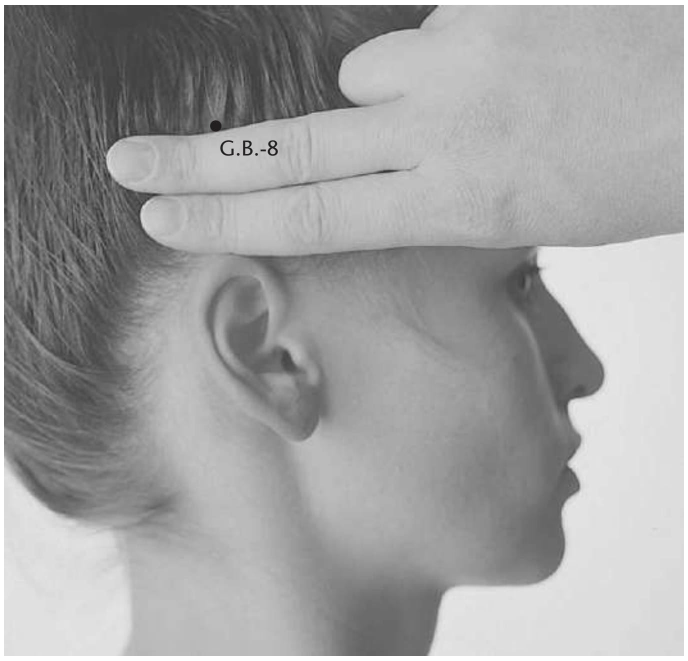

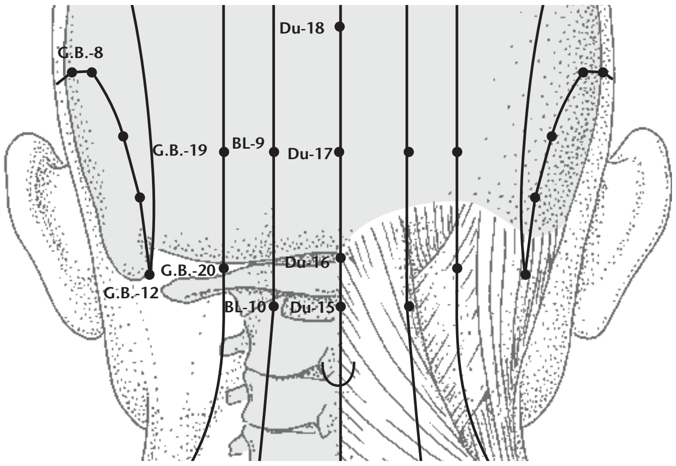
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

Heavenly Rushing TIANCHONG G.B.-9

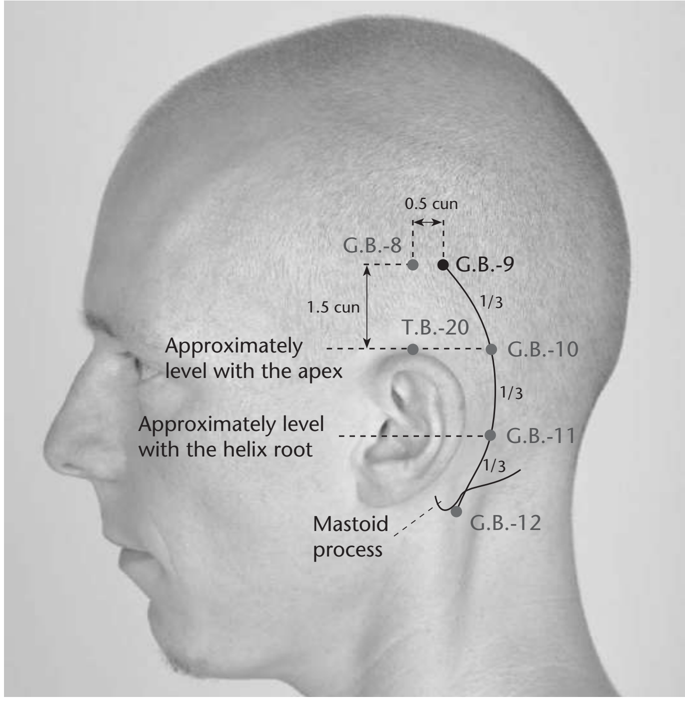
0.5 cun
G.B.-8
G.B.-9
1.5 cun
1/3
T.B.-20
Approximately level with the apex
G.B.-10
1/3
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

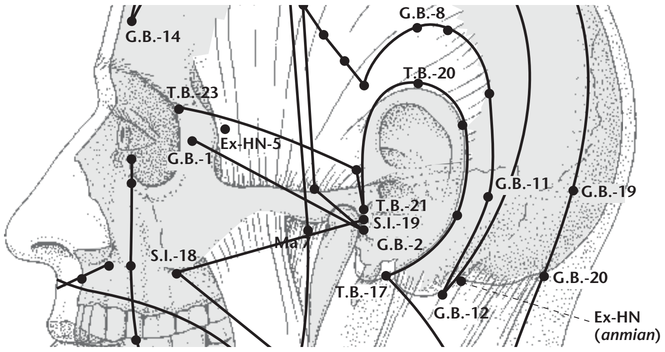
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
Ma7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)

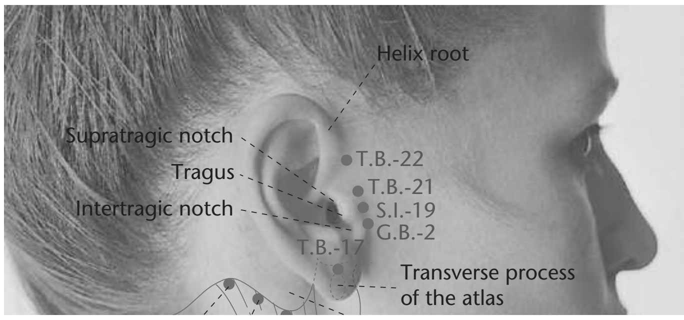
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Transverse process of the atlas

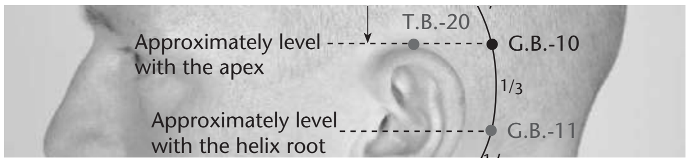
T.B.-20
Approximately level with the apex
G.B.-10
1/3
Approximately level with the helix root
G.B.-11

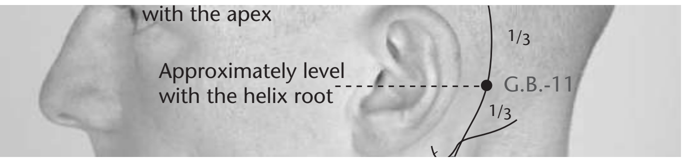
with the apex
1/3
Approximately level with the helix root
G.B.-11
1/3

Location

1.5 cun directly superior to the apex of the ear (→ **T.B.-20**) and 0.5 cun posterior to → **G.B.-8**. The point is approximately superior to the posterior margin of the ear.

How to find

First, locate → **G.B.-8** 1.5 cun superior to the apex of the ear. The latter becomes more clearly defined by folding the auricle towards the anterior so that the posterior part of the upper helix covers its anterior part. From → **G.B.-8**, measure 0.5 cun in a posterior direction and there locate **G.B.-9**. Similar to → **G.B.-8**, **G.B.-9** is located in a slight 'dip' in the bone.

Both **G.B.-9** and → **G.B.-12** (in the depression posterior and inferior to the mastoid process) are reference points for a curved line that runs approximately parallel to the posterior margin of the ear within the hairline. When dividing this line into thirds, → **G.B.-10** and → **G.B.-11** are located at the junctions of the thirds.

Needling

0.5–1.5 cun transversely (subcutaneously) towards the occiput or the site of the pain.

Actions/Indications

- Expels Wind and Heat
- Opens the *luo*-connecting vessels
- Calms the *shen*

Special features

Meeting point with the BL channel; Window of Heaven point according to some authors.

G.B.-10 *Floating White* FUBAI

Location

Posterior to the ear, at the junction of the upper third with the two lower thirds of the curved line connecting → **G.B.-9** and → **G.B.-12**.

How to find

First, locate → **G.B.-9** (1.5 cun superior to the apex of the ear and 0.5 cun in a posterior direction). Then locate → **G.B.-12** in a depression posterior and inferior to the mastoid process (→ 3.1.4). These two points are the endpoints of a curved line that runs approximately parallel to the posterior margin of the ear within the hairline. Divide this line into thirds and locate **G.B.-10** at the junction of the upper and the middle third, where often a small 'dip' can be felt on the bone. **G.B.-10** tends to be on the level of the apex of the ear.
→ **T.B.-20** is located directly above the apex of the ear.

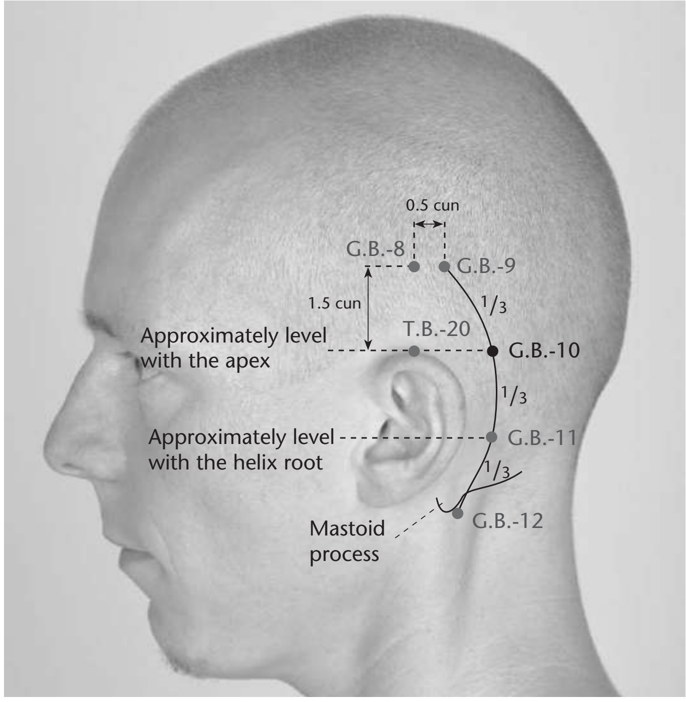
0.5 cun
G.B.-8
G.B.-9
1.5 cun
1/3
Approximately level with the apex
T.B.-20
G.B.-10
1/3
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

Needling

Transversely (subcutaneously) 0.5–0.8 cun

Actions/Indications

- Expels Wind
- Opens the *luo*-connecting vessels

Special features

Meeting point with the BL channel

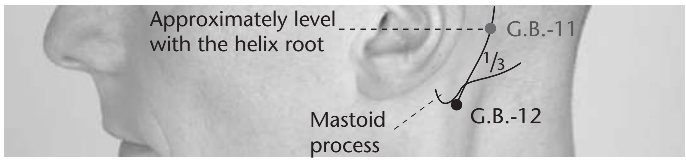
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

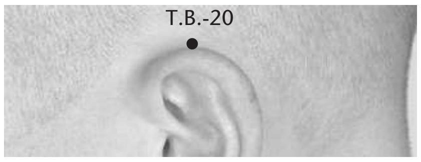

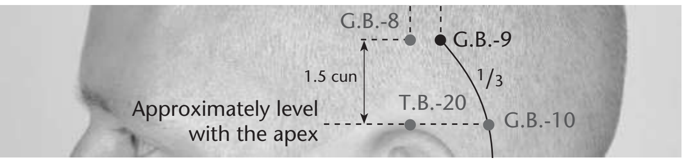
G.B.-8
G.B.-9
1.5 cun
1/3
Approximately level with the apex
T.B.-20
G.B.-10

Head Portal Yin TOUQIAOYIN G.B.-11

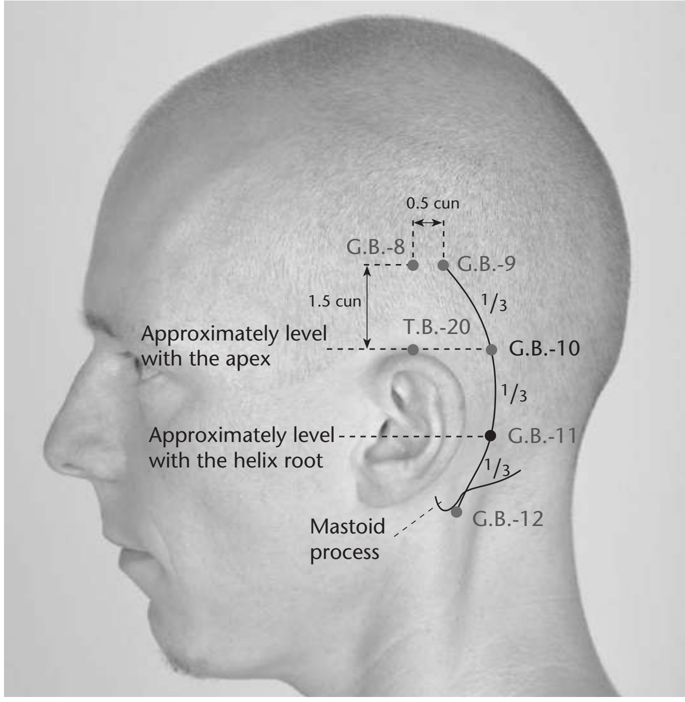
0.5 cun
G.B.-8
G.B.-9
1.5 cun
1/3
Approximately level with the apex
T.B.-20
G.B.-10
1/3
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

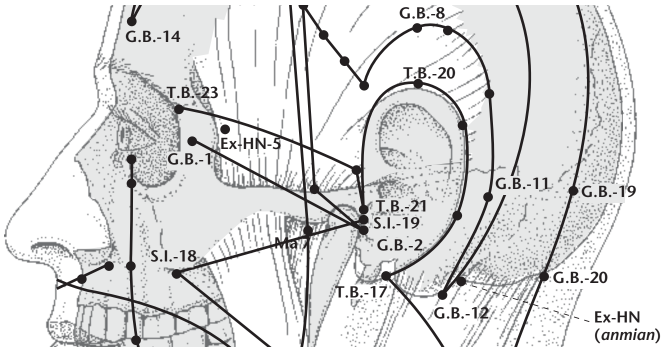
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
Ma 7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)

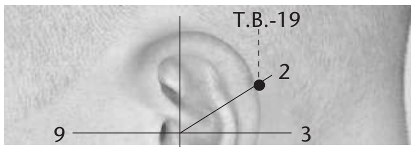

Location

Posterior to the ear, at the junction of the lower third with the two upper thirds of the curved line connecting → **G.B.-9** and → **G.B.-12**.

How to find

First, locate → **G.B.-9** (1.5 cun superior to the apex of the ear and 0.5 cun in a posterior direction). Then locate → **G.B.-12** in a depression posterior and inferior to the mastoid process (→ 3.1.4). These two points are the endpoints of a curved line that runs approximately parallel to the posterior margin of the ear within the hairline. Divide this line into thirds and locate **G.B.-11** at the junction of the middle and the lower third. At this point, a small 'dip' often can be felt in the bone. For reference: **G.B.-11** is located at the midpoint between a line connecting → **G.B.-10** and → **G.B.-12** and is generally on the level of the root of the helix.

→ **T.B.-19** is located on the same level, directly posterior to the helix.

Needling

Transversely (subcutaneously) 0.5–0.8 cun

Actions/Indications

- Expels Wind and Damp-Heat
- Opens the ears and the eyes
- Moves the (Liver) Qi and Blood

Special features

Meeting point with the BL channel, also with the S.I. and T.B. channels according to some authors.

G.B.-12 *Mastoid Process* WANGU

Location

In a depression directly posterior and inferior to the mastoid process.

How to find

First, locate the mastoid process posterior to the ear (→ 3.1.4). This can be palpated at the junction of the cranium and the neck as a cone-shaped, bony structure. With the palpating finger, find the inferior aspect of the mastoid process and locate **G.B.-12** on its lower border posterior to the tip of the process.
→ **T.B.-17** is located more anteriorly, in the depression posterior to the earlobe, between the mastoid process and the mandible.

Needling

0.5–1 cun obliquely in an inferior direction.

Actions/Indications

- Expels Wind, Heat and Dampness/Phlegm
- Opens the ears
- Calms the *shen*

Special features

Meeting point with the BL channel

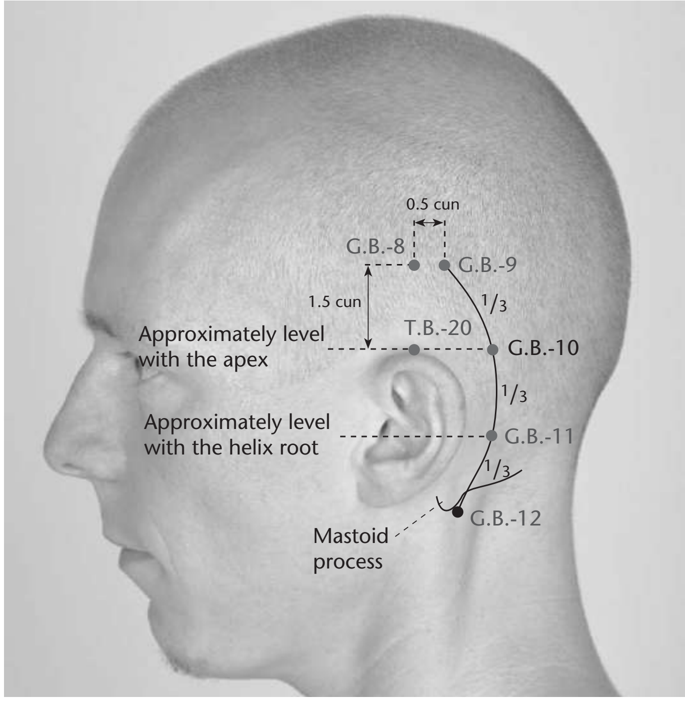
0.5 cun
G.B.-8
G.B.-9
1.5 cun
1/3
Approximately level with the apex
T.B.-20
G.B.-10
1/3
Approximately level with the helix root
G.B.-11
1/3
G.B.-12
Mastoid process

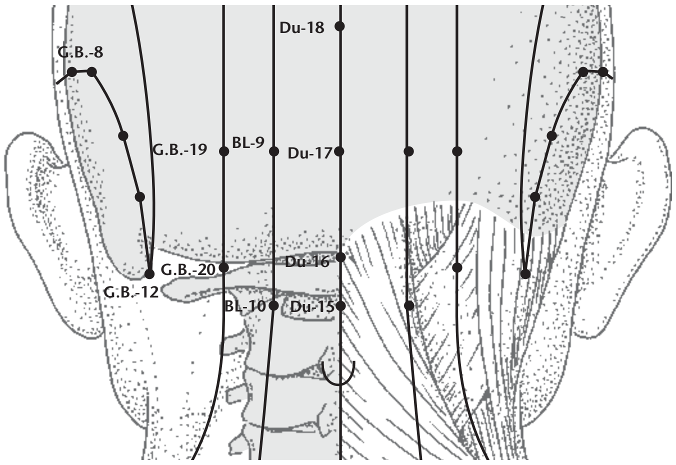
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
G.B.-20
Du-16
G.B.-12
BL-10
Du-15

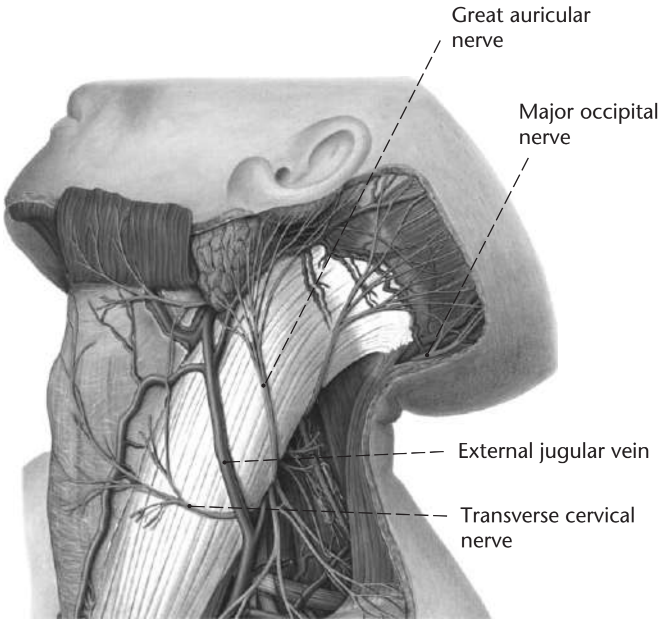
Great auricular nerve
Major occipital nerve
External jugular vein
Transverse cervical nerve

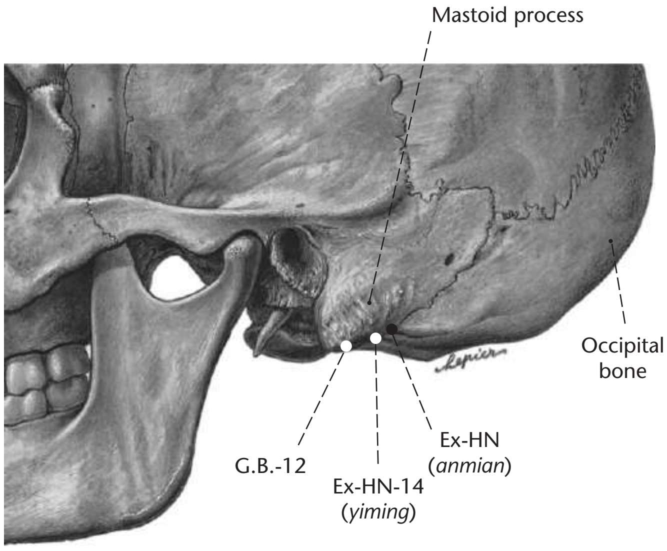
Mastoid process
Occipital bone
G.B.-12
Ex-HN (anmian)
Ex-HN-14 (yiming)

Spirit Root BENSHEN **G.B.-13**

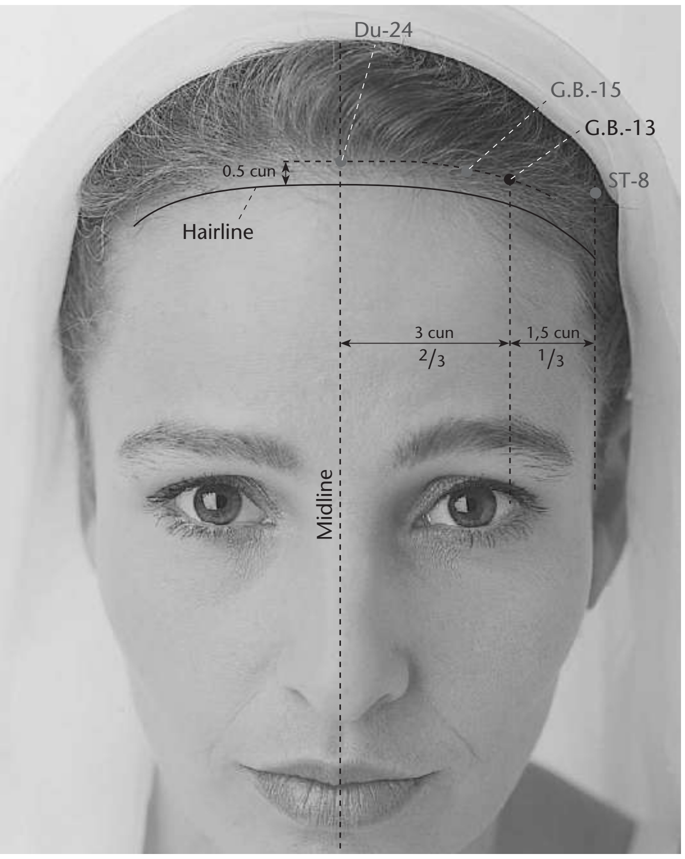
Du-24
G.B.-15
G.B.-13
0.5 cun
ST-8
Hairline
3 cun
2/3
1,5 cun
1/3
Midline

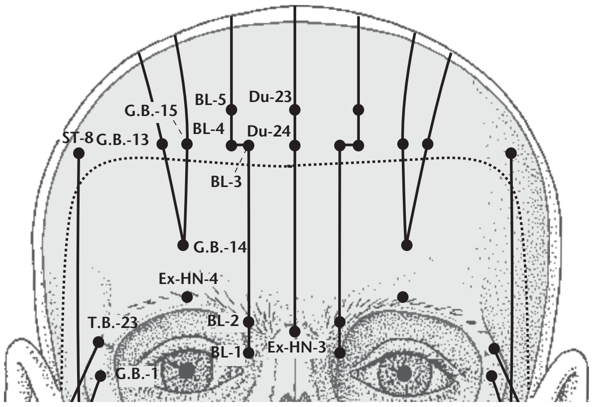
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1

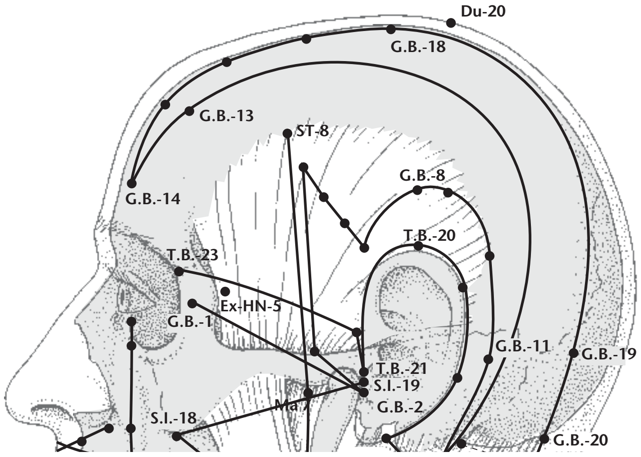
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
Ma
G.B.-20

Location

3 cun lateral to → **Du-24** (on the midline, 0.5 cun superior to the anterior hairline).

How to find

First, locate the anterior hairline (→ 3.1.1, with hairloss: when the patient frowns, the anterior hairline is marked by the border between the creased forehead and the smooth skin above) and there locate → **Du-24** on the midline and 0.5 cun within the hairline. **G.B.-13** is located on the same level, 3 cun lateral to the midline, measured in proportional cun based on the distance → **ST-8** to → **Du-24** (=4.5 cun, 2.2). Divide this distance into thirds and locate **G.B.-13** one third of the distance from **ST-8**. Located at the same level (0.5 cun within the anterior hairline) are → **Du-24/BL-3/BL-4/G.B.-15/ST-8** (on the midline/superior to the inner canthus of the eye/1.5 cun lateral to the midline/on the pupil line or 2.25 cun lateral to the midline/on the corner of the forehead).

Needling

0.5–0.8 cun transversely (subcutaneously) towards the occiput.

Actions/Indications

- Expels Wind
- Calms the *shen*
- Benefits the eyes

Special features

Meeting point with the *yang wei mai*

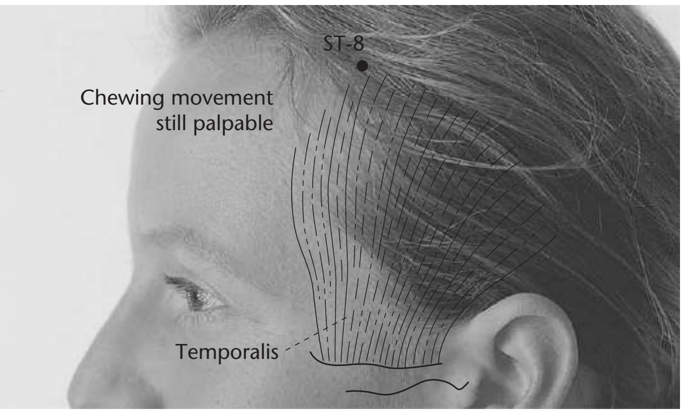
ST-8
Chewing movement
still palpable
Temporalis

G.B.-14 *Yang White* YANGBAI

Location

With the patient looking straight ahead, on the pupil line, approximately 1 cun superior to the midpoint of the eyebrow, at the junction of the frontal eminence and the superciliary arch.

How to find

On the pupil line, palpate from the anterior hairline in an inferior direction, past the frontal eminence, and locate **G.B.-14** at the deepest point between the frontal eminence and the superciliary arch. The distance between the midpoint of the eyebrow and the anterior hairline (➞ 3.1.1) is 3 proportional cun (➞ 2.2). **G.B.-14** is located one third of the distance or 1 cun superior to the midpoint of the eyebrow. With the patient looking straight ahead, **G.B.-14** is located on the pupil line.

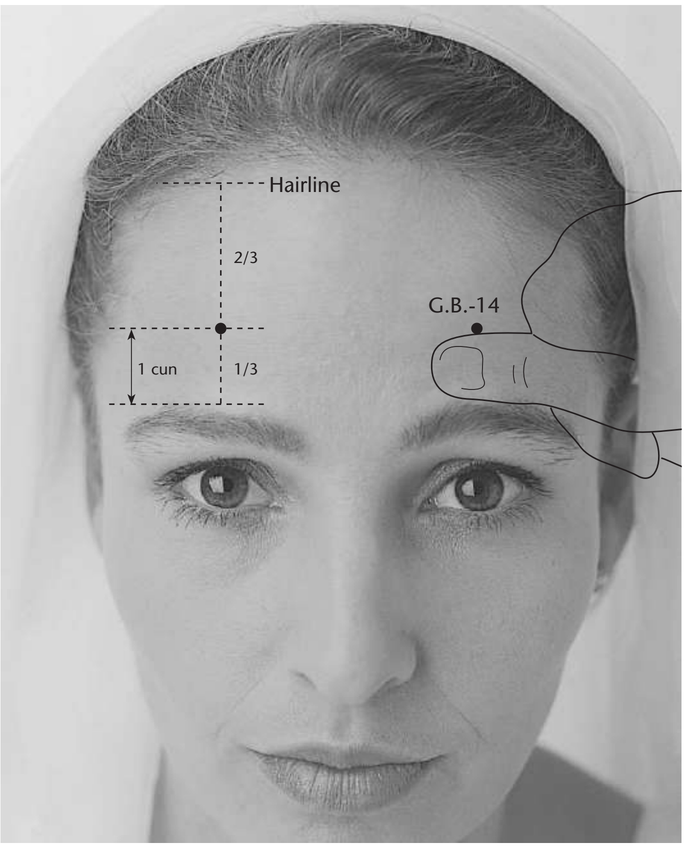
Hairline
2/3
G.B.-14
1 cun
1/3

Needling

0.3–1 cun transversely (subcutaneously) towards the centre of the eyebrow or towards the site of the pain. The pinching-skin method might be used for needling: pinch the skin between the thumb and index finger so that a skin fold forms. Insert the needle subcutaneously into this fold, directing it towards the site of the pain.

Actions/Indications

- Expels (internal and external) Wind and Wind-Heat, benefits the head, alleviates pain
- Benefits the eyes

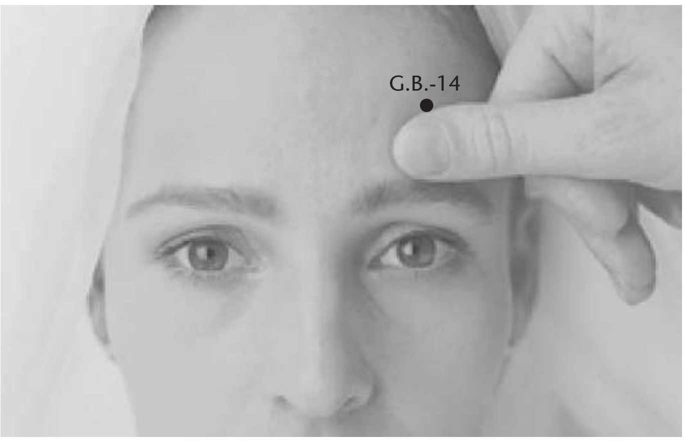

Special features

Meeting point with the *yang wei mai*, also with the T.B., L.I. and ST channels according to some authors. Important local point for frontal headache regardless of the pathology.

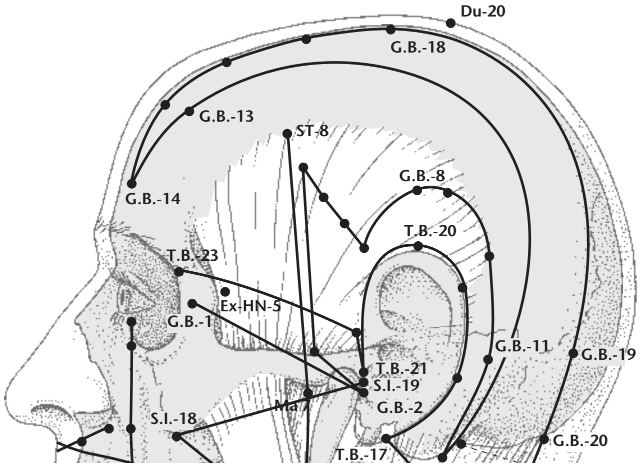
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
Ma 7
T.B.-17
G.B.-20

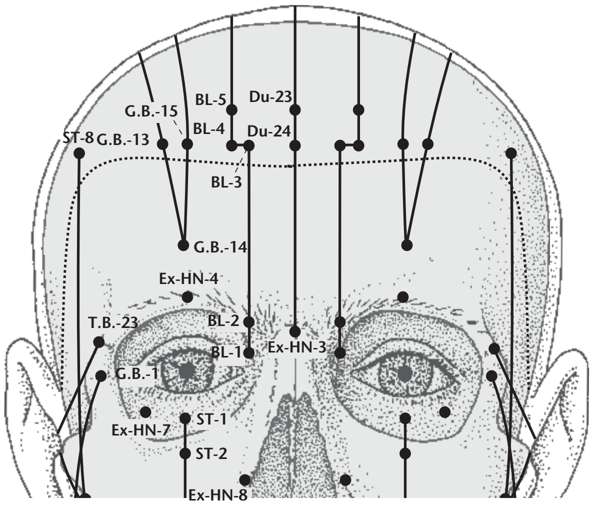
G.B.-15
BL-5
Du-23
BL-4
Du-24
ST-8
G.B.-13
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
Ex-HN-8

Head Governor of Tears TOULINQI G.B.-15

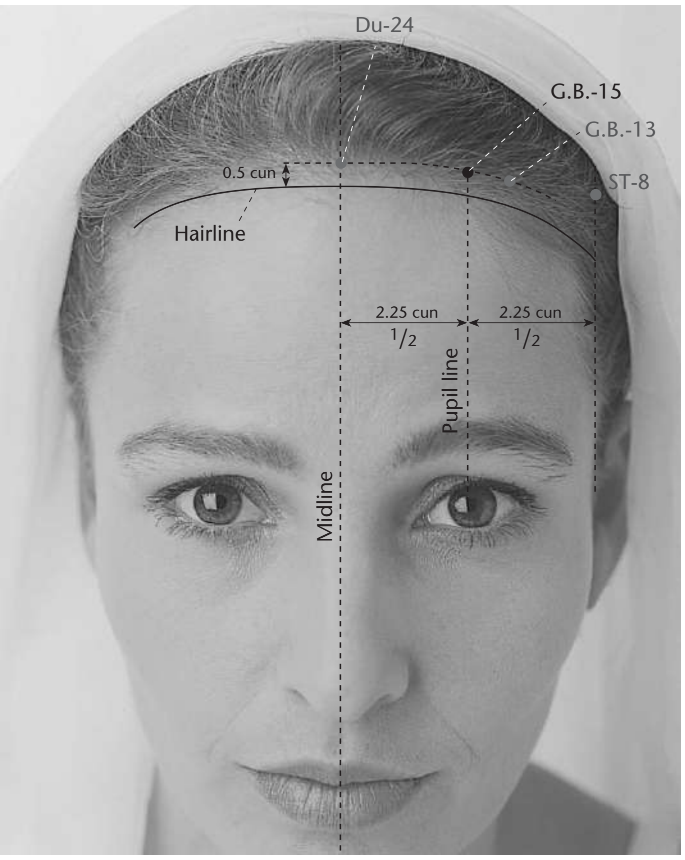
Du-24
G.B.-15
G.B.-13
0.5 cun
ST-8
Hairline
2.25 cun
1/2
2.25 cun
1/2
Pupil line
Midline

Location

With the patient looking straight ahead, directly superior to the pupil and 0.5 cun superior to the anterior hairline.

How to find

First, locate the anterior hairline (→ 3.1.1, with hairloss: when the patient frowns, the anterior hairline is marked by the border between the creased forehead and the smooth skin above). With the patient looking straight ahead, **G.B.-15** is located 0.5 cun superior to the anterior hairline and directly superior to the pupil. **G.B.-15** can also be described as being located midway between → **Du-24** (on the midline) and → **ST-8** (4.5 cun lateral to the midline, on the corner of the forehead).

Located on the same level (0.5 cun within the anterior hairline) are → **Du-24/BL-3/BL-4/G.B.-13/ST-8** (on the midline/ superior to the inner canthus of the eye/1.5 cun lateral to the midline/3 cun lateral to the midline or one third of the distance from → **ST-8** (on the corner of the forehead) to → **Du-24**.

Needling

Transversely (subcutaneously) 0.3–0.5 cun

Actions/Indications

- Expels Wind
- Opens the eyes and nasal passages
- Calms the *shen*

Special features

Meeting point with the *yang wei mai*, also with the BL channel according to some authors.

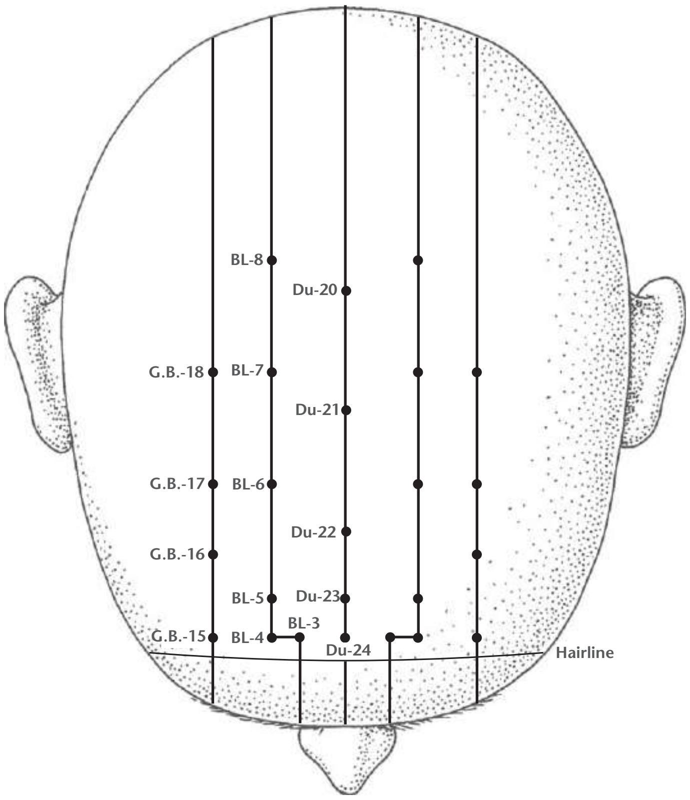
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

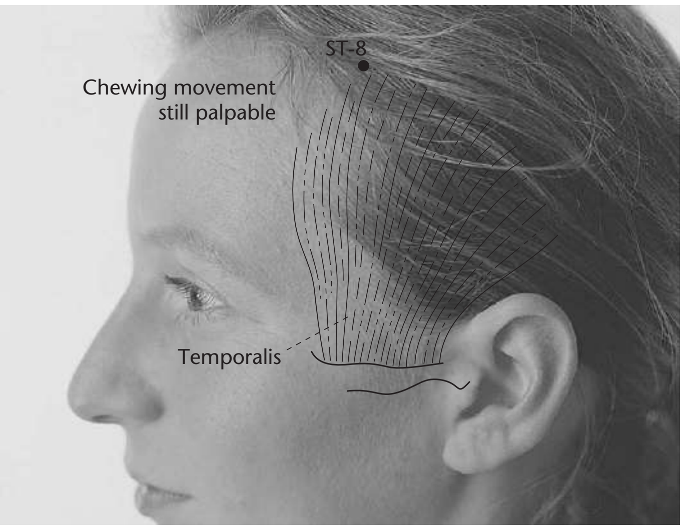
ST-8
Chewing movement
still palpable
Temporalis

G.B.-16 *Window of the Eye* MUCHUANG

Location

1.5 cun superior to the anterior hairline, on the pupil line or 2.25 cun lateral to the midline (midway between → **Du-24** and → **ST-8**).

How to find

Locate the anterior hairline (→ 3.1.1, with hairloss: when the patient frowns, the anterior hairline is marked by the border between the creased forehead and the smooth skin above). With the patient looking straight ahead, **G.B.-16** is located 1.5 cun superior to the anterior hairline, on a vertical line through the centre of the pupil. This line is located 2.25 cun lateral to the midline (this corresponds to the midpoint of the distance between → **Du-24** and → **ST-8**). For further orientation: The distance between the anterior hairline and → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears) is 5 cun. → **G.B.-16** is located 3.5 cun anterior to → **Du-20**, on the curved line connecting → **G.B.-15** and → **G.B.-20**.
→ **G.B.-15** is located on the same vertical line, but 0.5 cun superior to the anterior hairline.

Needling

0.3–1.5 cun transversely (subcutaneously) towards the occiput or the site of the disorder/pain.

Actions/Indications

- Expels pathogenic factors (especially Wind) from the head and eyes
- Opens the *luo*-connecting vessels

Special features

Meeting point with the *yang wei mai*

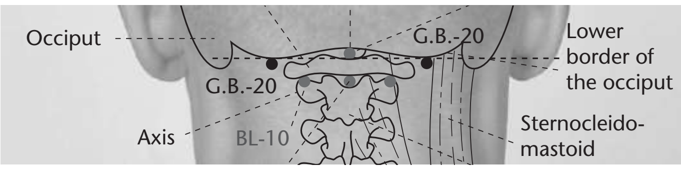
Occiput
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid

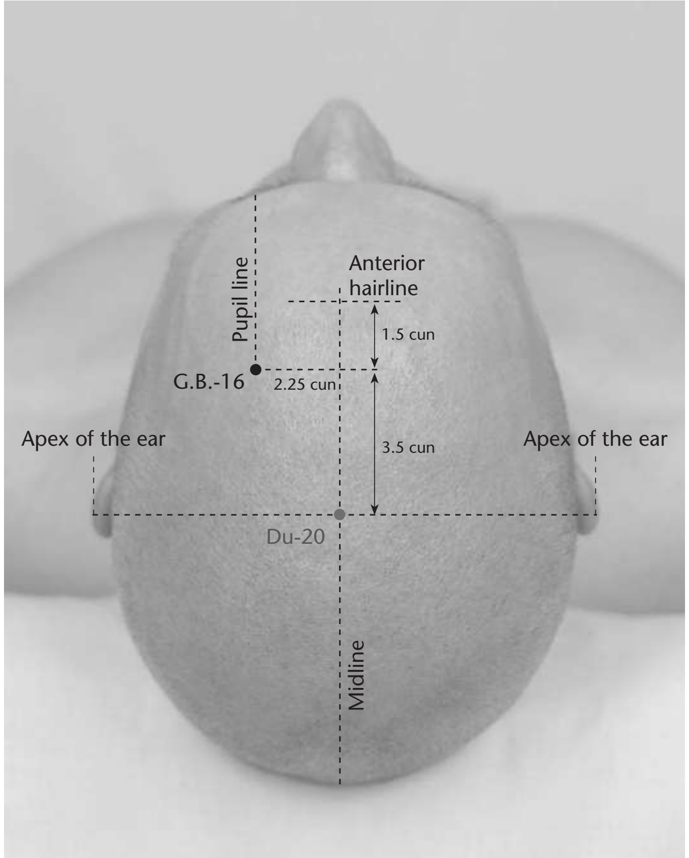
Pupil line
Anterior hairline
1.5 cun
G.B.-16
2.25 cun
Apex of the ear
3.5 cun
Apex of the ear
Du-20
Midline

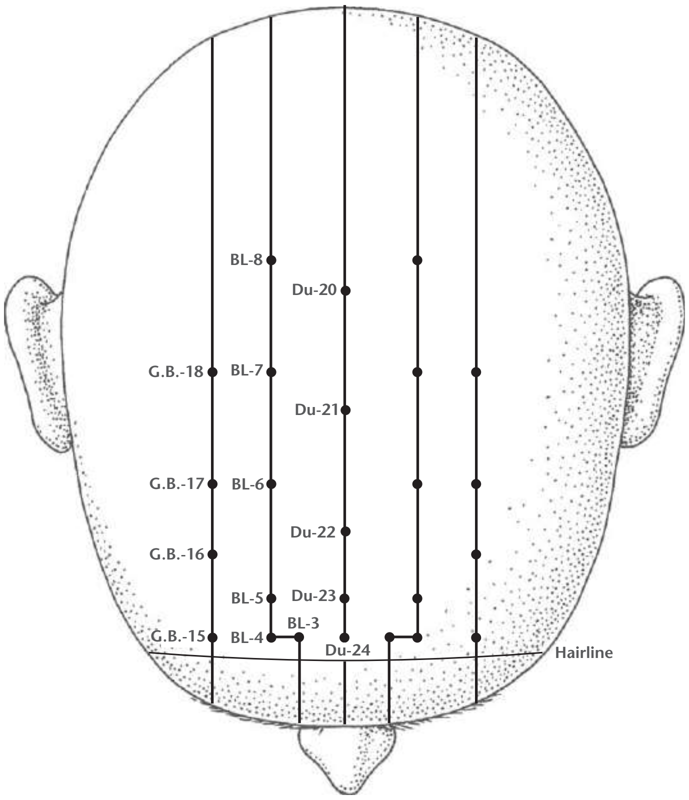
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

Upright Nutrition ZHENGYING G.B.-17

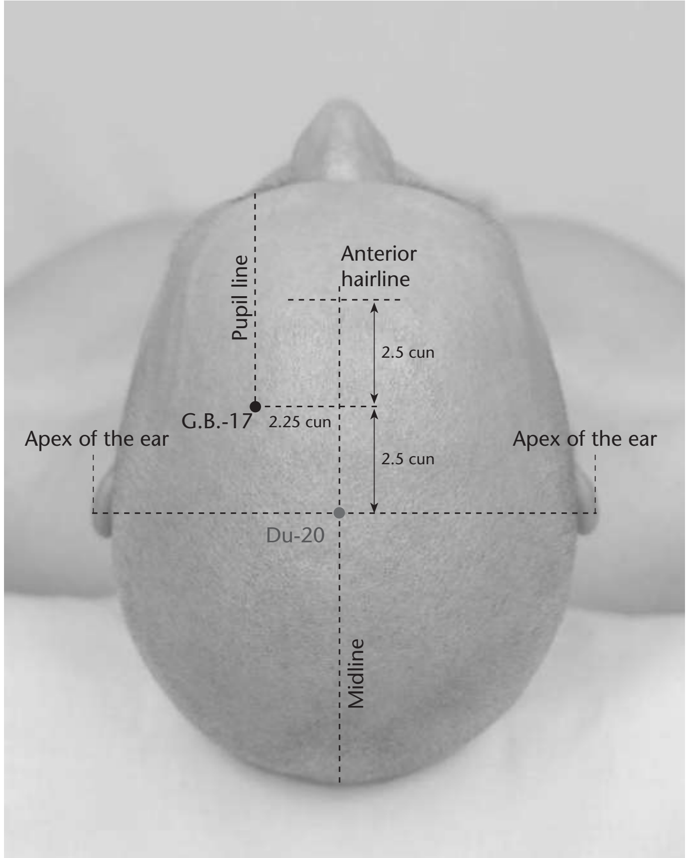
Pupil line
Anterior hairline
2.5 cun
G.B.-17
2.25 cun
Apex of the ear
2.5 cun
Apex of the ear
Du-20
Midline

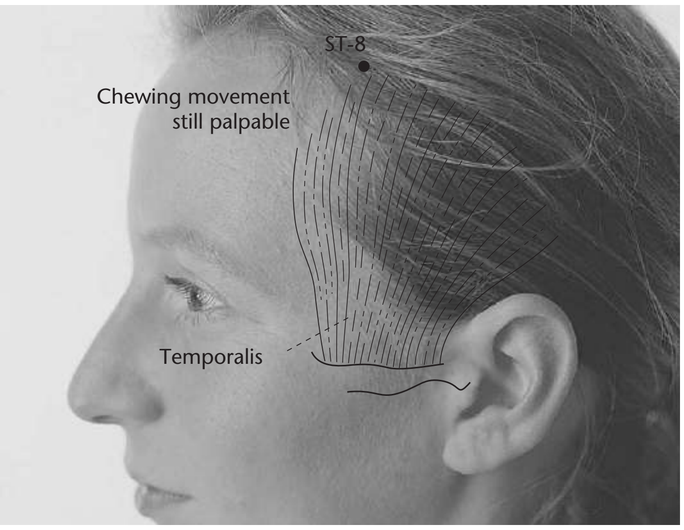
ST-8
Chewing movement still palpable
Temporalis

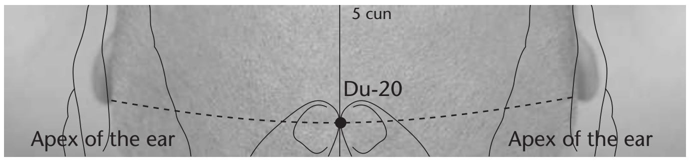
5 cun
Du-20
Apex of the ear
Apex of the ear

Location

2.5 cun superior to the anterior hairline and 2.25 cun lateral to the midline.

How to find

First, on the midline, locate the anterior hairline (→ 3.1.1, with hairloss: when the patient frowns, the anterior hairline is marked by the border between the creased forehead and the smooth skin above) and → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears). The distance between these two points is 5 proportional cun (→ 2.2). Locate the midpoint (for example by using the spreading hands technique → 2.3.3). **G.B.-17** is located on the curved line connecting → **G.B.-15** and → **G.B.-20** (an extension of the pupil line, 2.25 cun lateral to the midline).

→ **BL-6** is also located 2.5 cun anterior to → **Du-20**, but 1.5 cun lateral to the midline.

Needling

0.3–1.5 cun transversely (subcutaneously) towards the occiput or the site of the disorder/pain.

Actions/Indications

- Expels Wind
- Opens the *luo*-connecting vessels
- Descends counterflow Stomach Qi

Special features

Meeting point with the *yang wei mai*

G.B.-18 *Support Spirit* CHENGLING

Location

4 cun superior to the anterior hairline or 1 cun anterior to → **Du-20** and 2.25 cun lateral to the midline.

How to find

First, locate → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears, distance to the anterior hairline = 5 cun). Next, locate **G.B.-18** 1 cun anterior to → **Du-20** on the curved line connecting → **G.B.-15** and → **G.B.-20** (= extension of the pupil line, midway between → **ST-8** and → **Du-24**).
→ **BL-7** is located on the same level (1 cun anterior to → **Du-20**), but 1.5 cun lateral to the midline.

Needling

0.3–1.5 cun transversely (subcutaneously) towards the occiput or the site of the disorder/pain.

Actions/Indications

- Expels Wind, especially from the nose

Special features

Meeting point with the *yang wei mai*

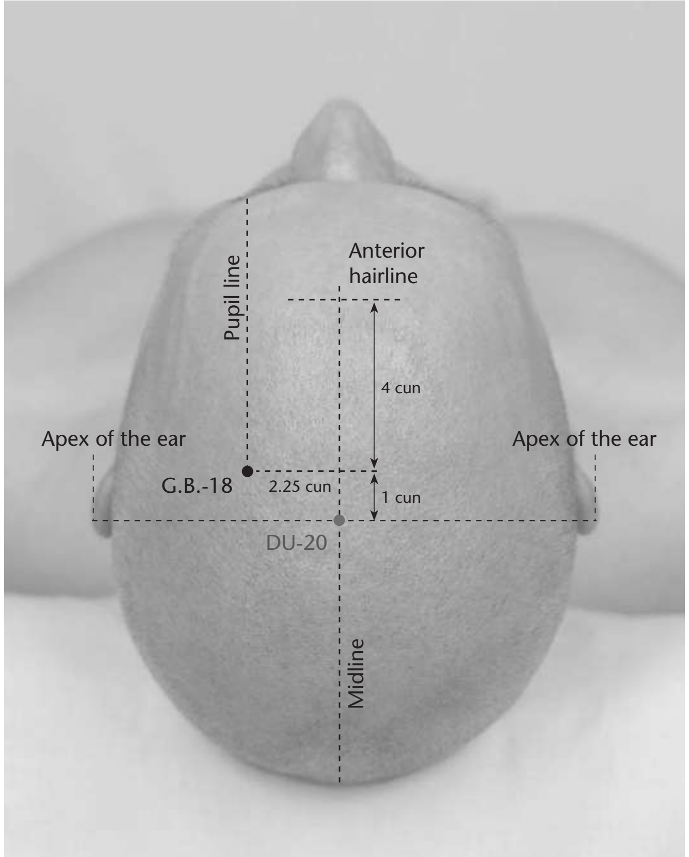
Pupil line
Anterior hairline
4 cun
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
DU-20
Midline

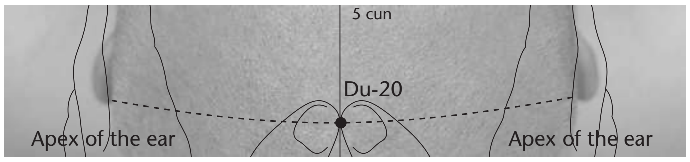
5 cun
Du-20
Apex of the ear
Apex of the ear

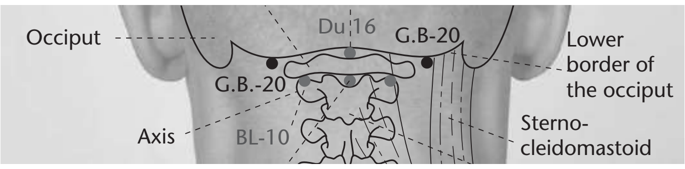
Occiput
Du 16
G.B-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Sterno-cleidomastoid

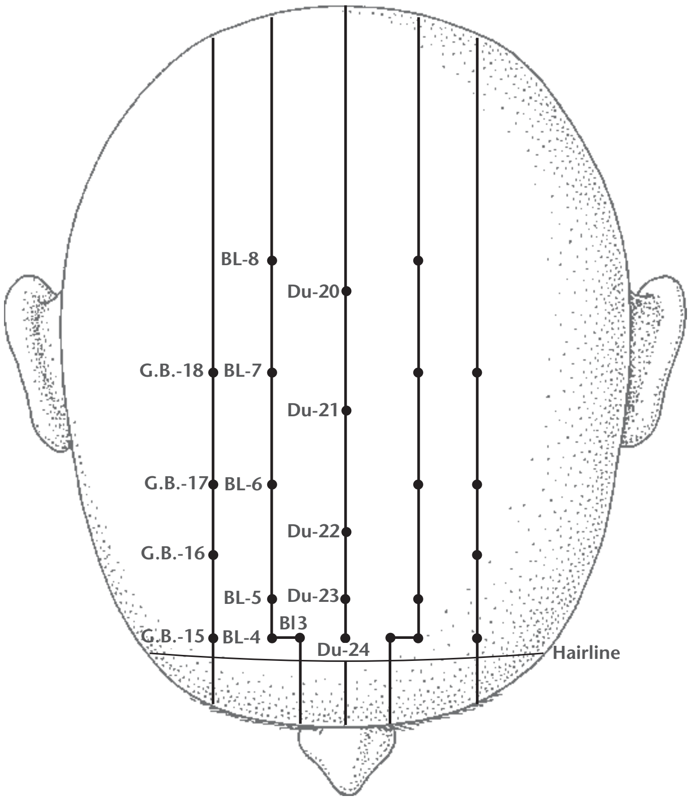
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
Bl3
G.B.-15
BL-4
Du-24
Hairline

Brain Hollow NAOKONG G.B.-19

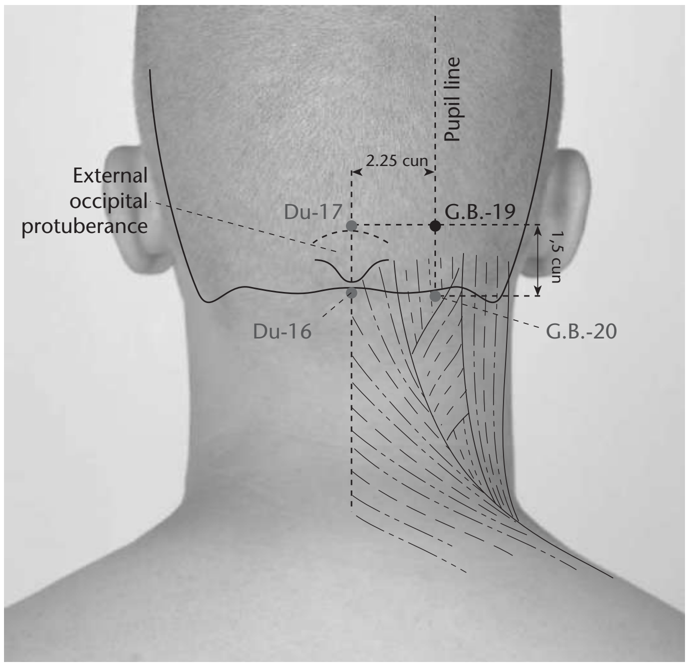
Pupil line
2.25 cun
External occipital protuberance
Du-17
G.B.-19
1.5 cun
Du-16
G.B.-20

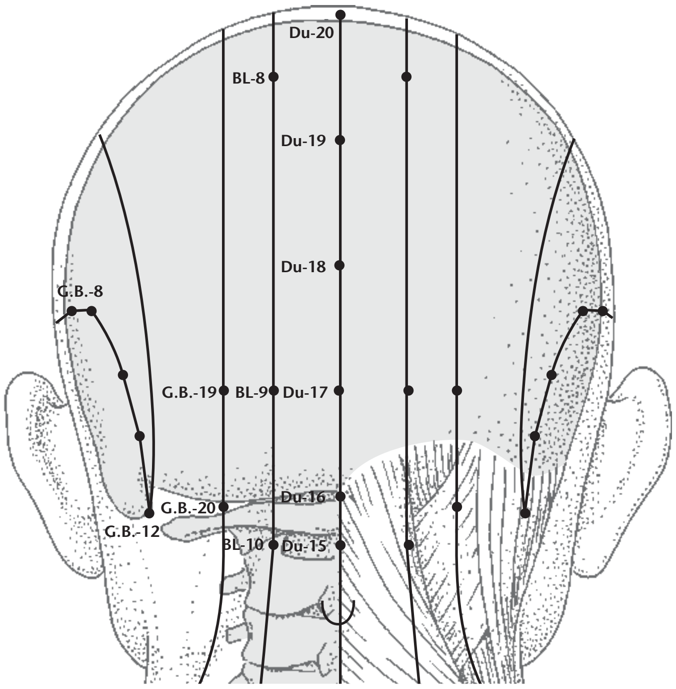
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

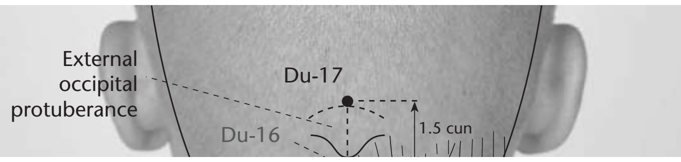
External occipital protuberance
Du-17
Du-16
1.5 cun

Location

On the posterior aspect of the head, at the upper border of the external occipital protuberance (→ **Du-17**) and 2.25 cun lateral to the midline.

How to find

First, locate the external occipital protuberance (→ 3.1.5), which can be palpated as a flat protruding area on the midline. Next, find → **Du-17** in the depressoin just superior to the protuberance. From there, measure 2.25 cun in a lateral direction and locate **G.B.-19** on the curved line connecting → **G.B.-15** and → **G.B.-20**, an extension of the pupil line (→ 2.25 cun lateral to the midline or midway between → **Du-24** and → **ST-8**). **For orientation: G.B.-19** is located approximately 2.5 cun superior to the posterior hairline (→ 3.1.5) and approximately 1.5 cun superior to → **G.B.-20** (on the lower border of the occiput, between the origins of the sternocleidomastoid and trapezius muscles).

Needling

0.3–1.5 cun transversely (subcutaneously) towards the occiput or the site of the disorder/pain.

Actions/Indications

- Expels Wind
- Clears the sensory orifices
- Opens the *luo*-connecting vessels

Special features

Meeting point with the *yang wei mai*

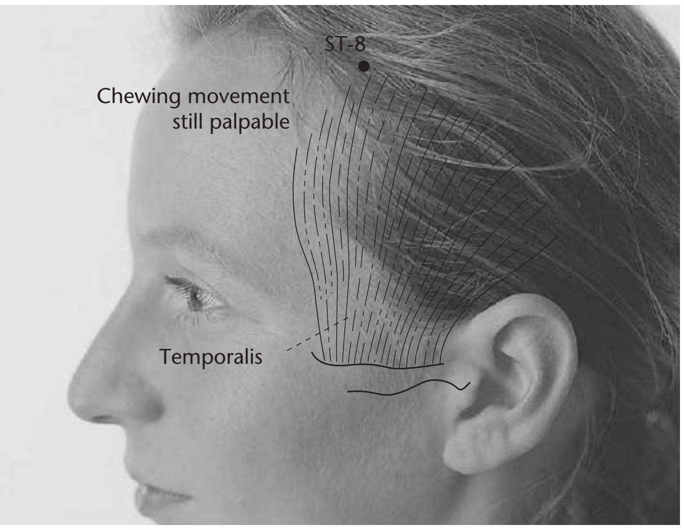
ST-8
Chewing movement still palpable
Temporalis

G.B.-20 *Wind Pool* FENGCHI

Location

At the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius muscles.

How to find

Patient's position: Prone, sitting or supine (supported by a pillow, so that the occipital region is accessible). Starting at the midline, glide with the palpating finger along the lower border of the occiput, crossing the bulge of the origin of the trapezius muscle, until you reach a depression the size of a finger pad. Locate **G.B.-20** in its centre.

Located on the same level is → **Du-16**. → **BL-10** is located more medially and inferiorly.

Needling

Needle with the patient's head bent forward and the tip of the needle pointing inferiorly. Depending on the position of the head, insertion towards the tip of the nose or the contralateral orbit, 0.5–1.2 cun. Caution: In slim patients, do not needle deeper than 2 cm (the vertebral artery is located at a depth of approximately 4 cm).

Actions/Indications

- Eliminates Wind, benefits the head, clears the sensory organs
- Opens the channel
- With tonifying needle technique: strengthens the Marrow and the Brain (according to Maciocia)

Special features

Meeting point with the T.B. channel*, *yang wei mai, yang qiao mai*. Major point for all 'Wind disorders', very important point for disorders of the head and eyes.

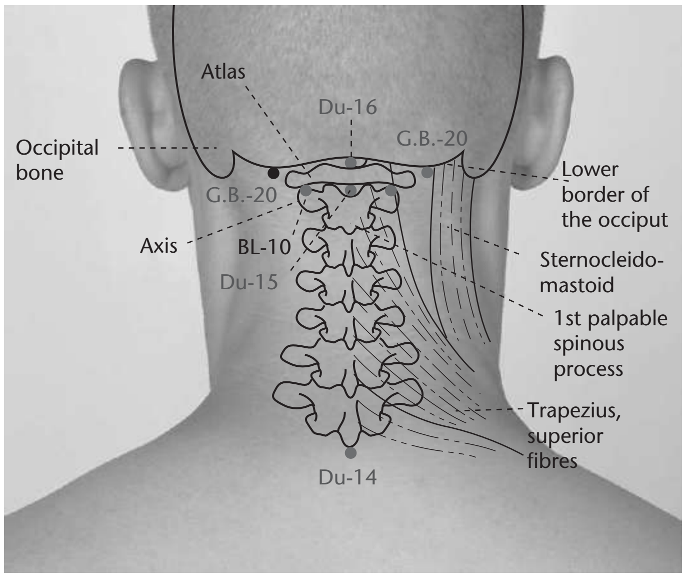
Atlas
Du-16
G.B.-20
Occipital bone
Lower border of the occiput
G.B.-20
Axis
BL-10
Du-15
Sternocleido-mastoid
1st palpable spinous process
Trapezius, superior fibres
Du-14

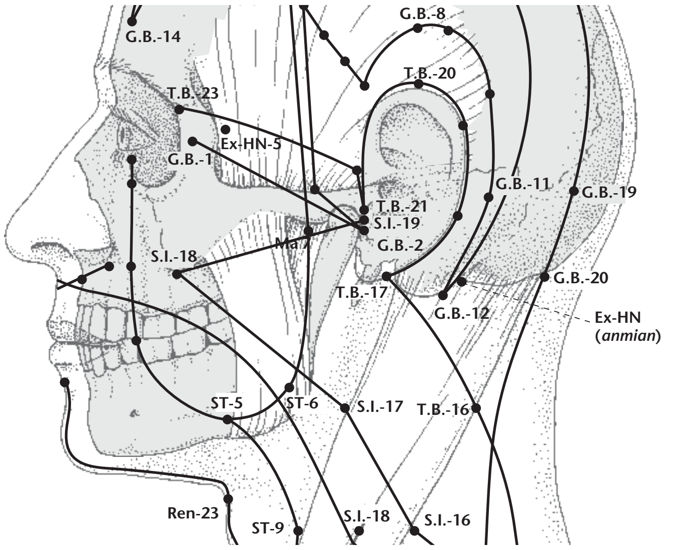
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
Ma 7
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
S.I.-18
S.I.-16

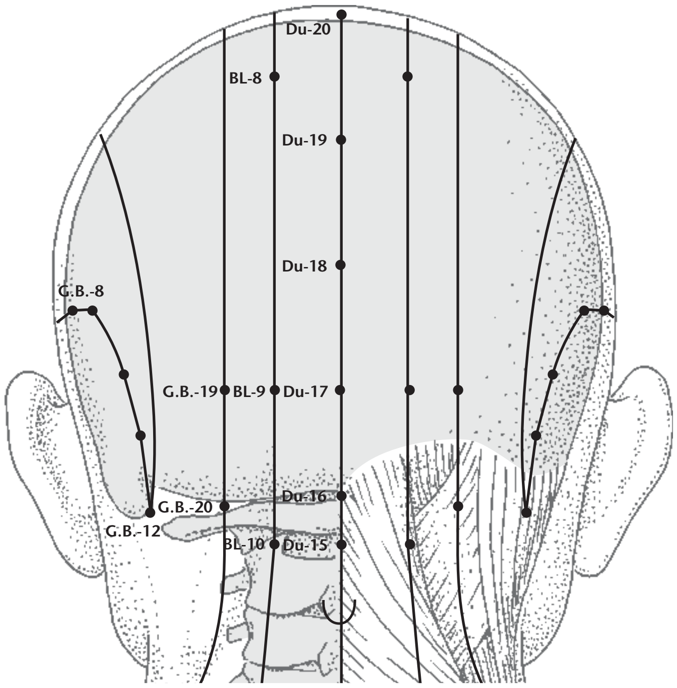
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

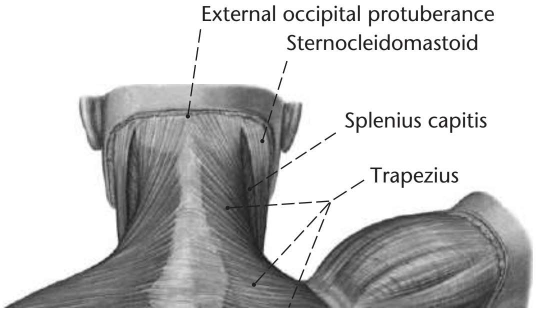
External occipital protuberance
Sternocleidomastoid
Splenius capitis
Trapezius

---

*According to Deadman et al 1998.

Shoulder Well JIANJING G.B.-21

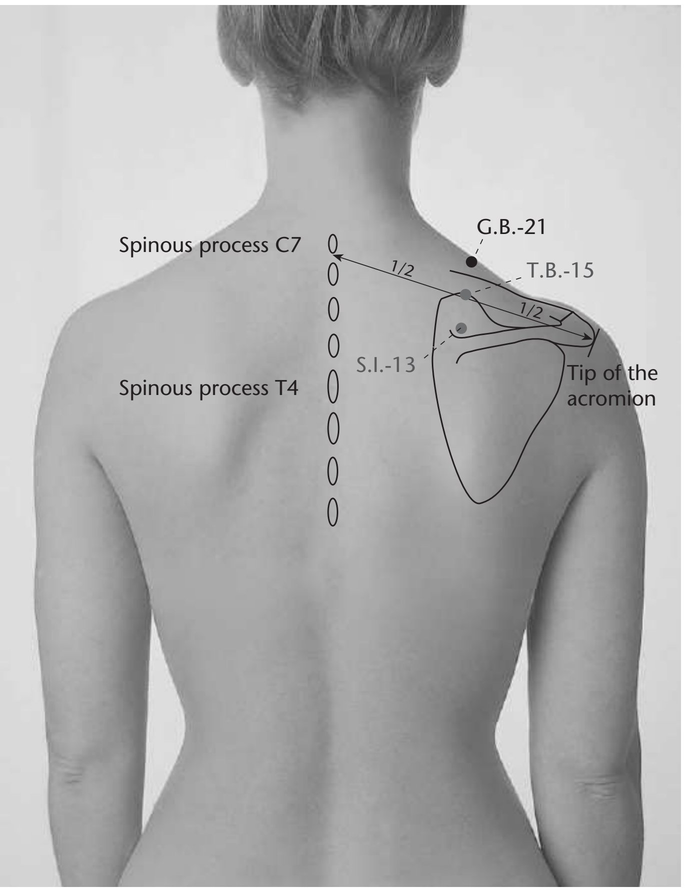

Location

At the highest point of the shoulder, at the midpoint of a line connecting the 7th cervical vertebra (C7) and the lateral extremity of the acromion.

How to find

First, locate C7 (→ 3.4.1) and the lateral extremity of the acromion (→ 3.3.1). Next, locate **G.B.-21** at the midpoint of a line connecting these two reference points, on the highest point of the trapezius muscle on the sagittal plane.
→ **T.B.-15** is located 1 cun inferior to **G.B.-21**, midway between **G.B.-21** and → **S.I.-13** (medial to the supraspinous fossa).

Needling

Vertically approximately 0.3–0.5 cun (lift the muscle). An alternative, safer method: Lift the muscle and insert the needle approximately 1 cun anteriorly or posteriorly into the muscle belly. Caution: Contraindicated during pregnancy! Pneumothorax.

Actions/Indications

- Opens the channel
- Descends the Qi
- Regulates the Qi flow, transforms Phlegm, dissipates nodules
- Promotes labour, benefits the breasts

Special features

Meeting point with the *yang wei mai* and the T.B. and ST* channels. Important local point. Trigger point for the shoulder with a distal effect on the Uterus.

---

*According to Deadman et al 1998.

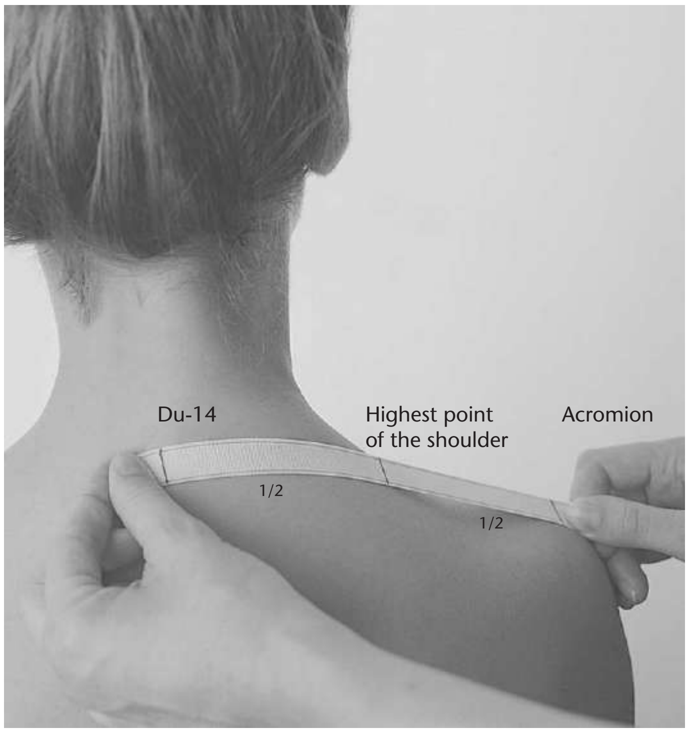

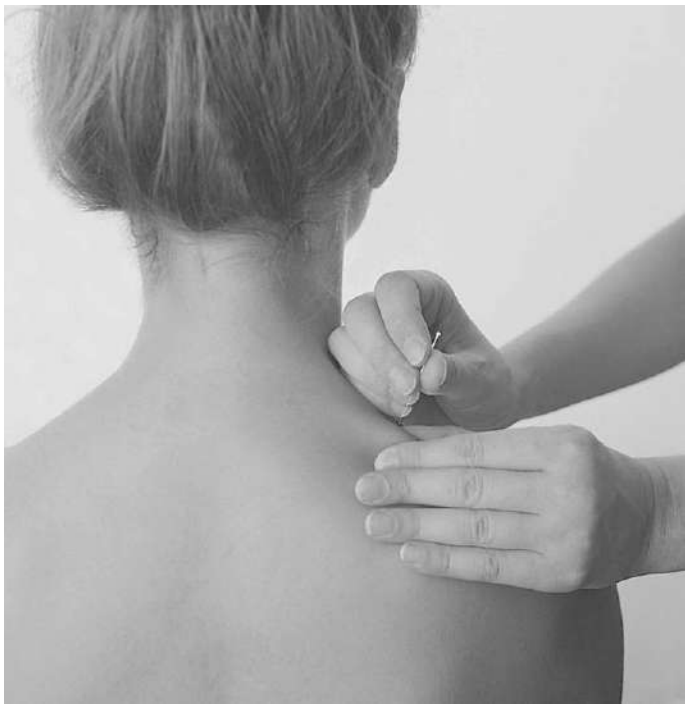

G.B.-22 *Armpit Abyss* YUANYE

Location

On the midaxillary line, with the arm abducted approximately 3 cun inferior to the apex of the axilla, in the 4th intercostal space (according to some authors, in the 5th intercostal space).

How to find

With the patient's arm slightly abducted, locate the apex of the axilla. Locate **G.B.-22** 3 cun inferior to it in the 4th intercostal space. **For orientation:** In men, the nipple is located in the 4th intercostal space; in women, in a supine position, its location may vary. Therefore, in women, the manubriosternal synchondrosis (→ 3.5) is a more reliable reference point for the costal region. Note: The intercostal space curves in a superior direction towards lateral.

Also located on the level of or in the 4th intercostal space, but more medially, are → **Ren-17/KID-23/ST-17/P-1/SP-18/G.B.-23** (on the midline/2 cun lateral to the midline/in the centre of the nipple or 4 cun lateral to the midline/1 cun lateral to the nipple or 5 cun lateral to the midline/6 cun lateral to the midline/1 cun anterior to **G.B.-22**).

Needling

0.5–1 cun obliquely or transversely (subcutaneously) along the intercostal space. Caution: Pneumothorax.

Actions/Indications

- Regulates the Qi in the Upper Burner
- Opens the *luo*-connecting vessels of the axilla

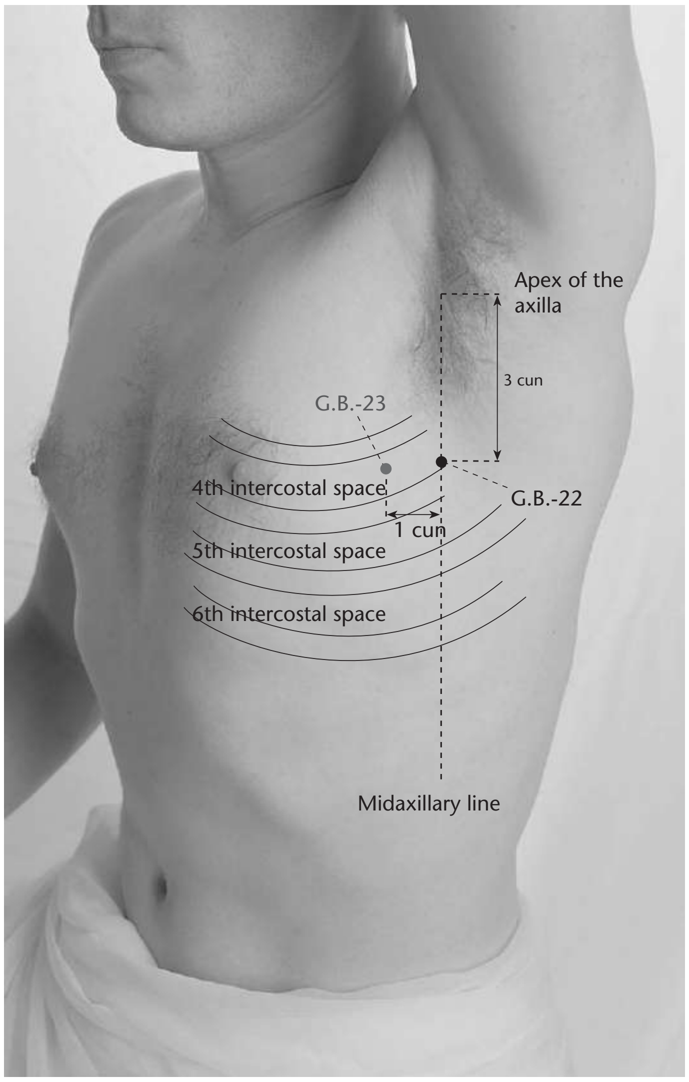
Apex of the axilla
3 cun
G.B.-23
4th intercostal space
1 cun
G.B.-22
5th intercostal space
6th intercostal space
Midaxillary line

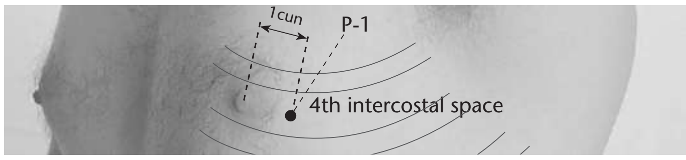
1cun
P-1
4th intercostal space

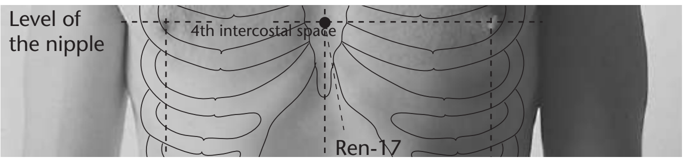
Level of the nipple
4th intercostal space
Ren-17

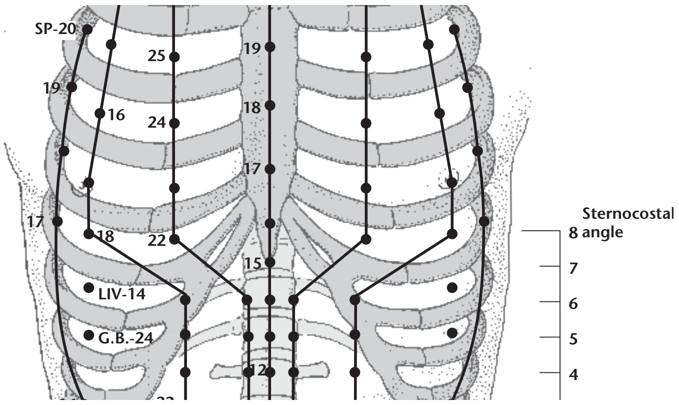
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
Sternocostal angle
8
7
6
5
4

Flank Sinews ZHEJIN G.B.-23

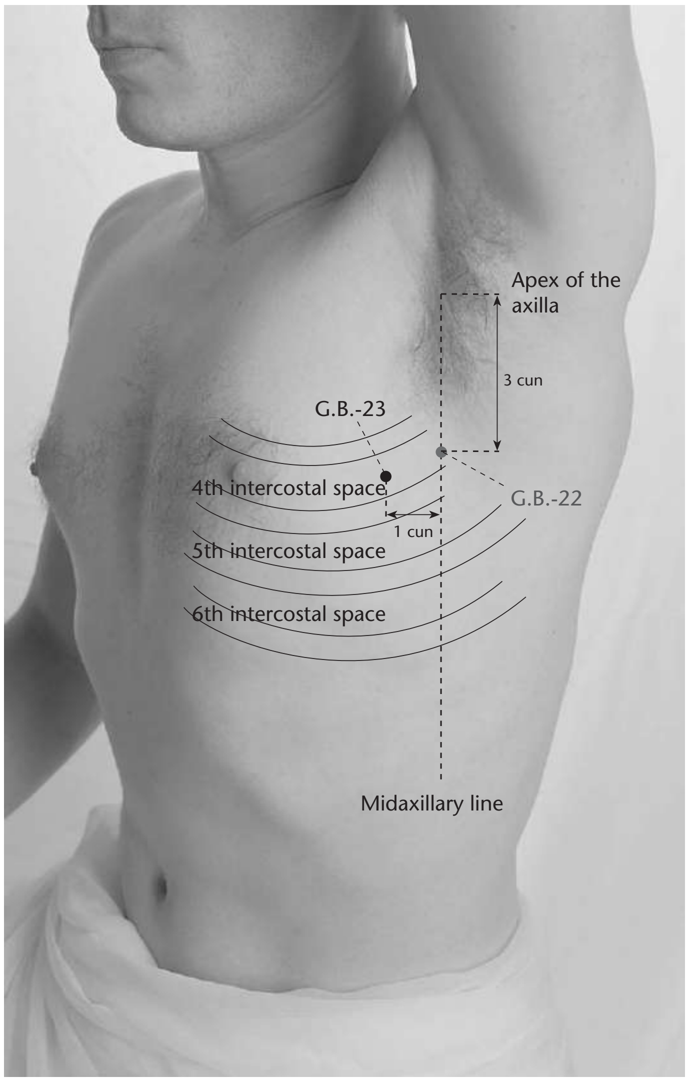
Apex of the axilla
3 cun
G.B.-23
4th intercostal space
G.B.-22
1 cun
5th intercostal space
6th intercostal space
Midaxillary line

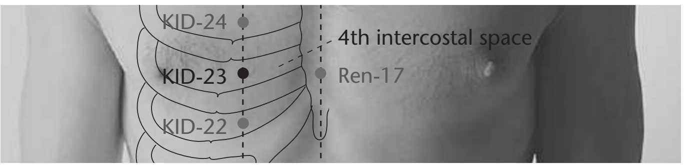
KID-24
4th intercostal space
KID-23
Ren-17
KID-22

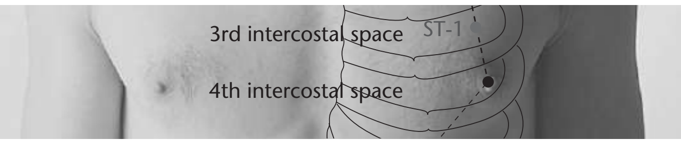
3rd intercostal space
ST-1
4th intercostal space

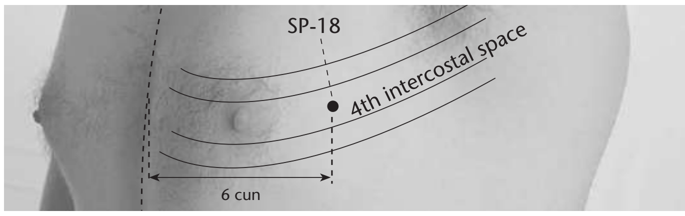
SP-18
4th intercostal space
6 cun

Location

1 cun anterior to → **G.B.-22** (on the midaxillary line, 3 cun inferior to the apex of the axilla in the 4th intercostal space).

How to find

With the patient's arm slightly abducted, locate the apex of the axilla. First, locate → **G.B.-22** 3 cun inferior to it in the 4th intercostal space, then locate **G.B.-23** 1 cun anterior to → **G.B.-22** and also in the 4th intercostal space. **For orientation:** In men, the nipple is located in the 4th intercostal space; in women, in a supine position, its location may vary. Therefore, in women, the manubriosternal synchondrosis (→ 3.5) is a more reliable reference point for the costal region. Note: The intercostal space curves in a superior direction towards lateral.

Also located on the level of or in the 4th intercostal space, but more medially, are → **Ren-17/KID-23/ST-17/P-1/SP-18** (on the midline/2 cun lateral to the midline/in the centre of the nipple or 4 cun lateral to the midline/1 cun lateral to the nipple or 5 cun lateral to the midline/6 cun lateral to the midline).

Needling

0.5–1 cun obliquely or transversely (subcutaneously) along the 4th intercostal space. Caution: Pneumothorax.

Actions/Indications

- Regulates the Qi between the Upper and Middle Burners
- Opens the channel

Special features

Meeting point with the BL channel according to some authors

G.B.-24 *Sun and Moon* RIYUE

Location

In the 7th intercostal space, on the mamillary line (4 cun lateral to the anterior midline).

How to find

The mamillary line, a vertical line 4 cun lateral to the midline, is used for reference in the thoracic region (→ 3.5). **Quick method, especially in men:** In men, the nipple is usually located on the level of the 4th intercostal space. From the nipple, count downward to the 7th intercostal space. **Or:** A more reliable reference, especially in women, is the manubriosternal synchondrosis, a horizontal bony structure on the sternum. The costal cartilage of the second rib is lateral to the synchondrosis, with the 2nd intercostal space below. From there, count downward to the 7th intercostal space and locate **G.B.-24** on the mamillary line. → **LIV-14** is located directly superior to **G.B.-24** in the 6th intercostal space.

Located approximately on the same level (1 cun inferior to the sternocostal angle) are → **Ren-13/KID-20/ST-20** (on the midline/0.5 cun lateral to the midline/2 cun lateral to the midline).

Needling

0.3–0.8 cun obliquely in a lateral direction along the intercostal space. Caution: Pneumothorax.

Actions/Indications

- Benefits the Gall Bladder, eliminates Damp-Heat, regulates and spreads Liver Qi, descends counterflow Qi, harmonises the Middle Burner

Special features

Front-*mu* point of the Gall Bladder, meeting point with the SP channel, also with the *yang wei mai* according to some authors. Major point for disorders of the Gall Bladder.

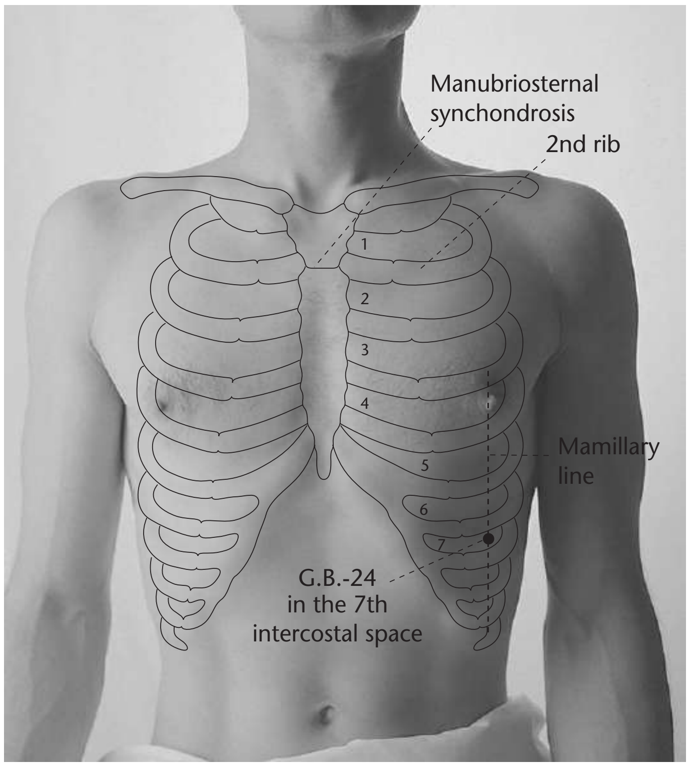
Manubriosternal synchondrosis
2nd rib
1
2
3
4
5
6
7
Mamillary line
G.B.-24 in the 7th intercostal space

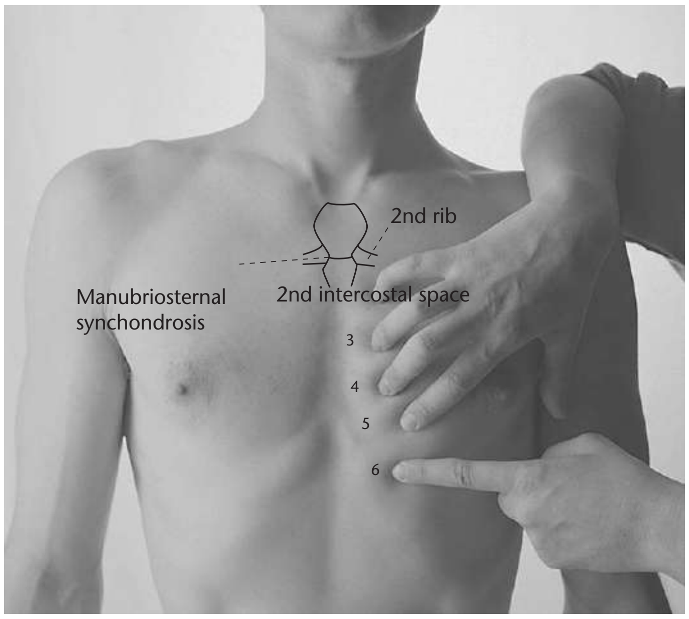
2nd rib
Manubriosternal synchondrosis
2nd intercostal space
3
4
5
6

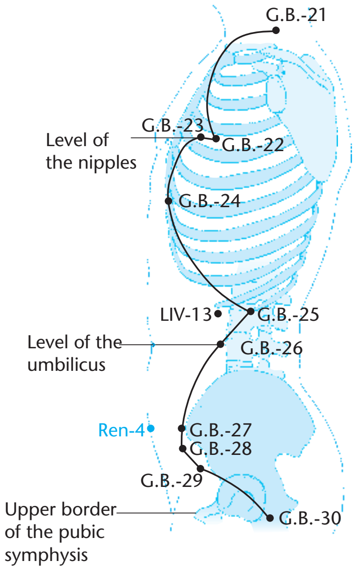
G.B.-21
G.B.-23
Level of the nipples
G.B.-22
G.B.-24
LIV-13
G.B.-25
Level of the umbilicus
G.B.-26
Ren-4
G.B.-27
G.B.-28
G.B.-29
Upper border of the pubic symphysis
G.B.-30

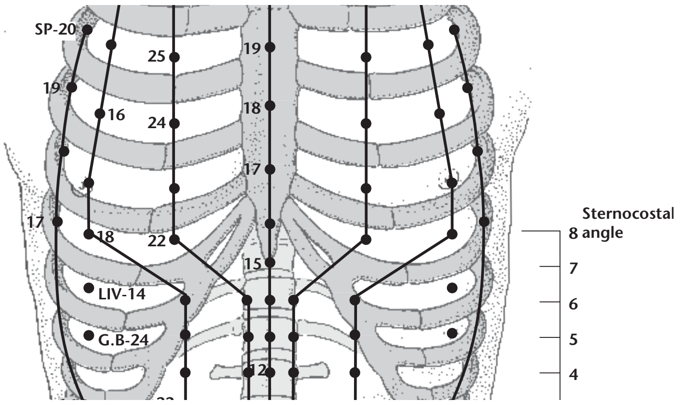
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B-24
12
Sternocostal angle
8
7
6
5
4

Capital Gate JINGMEN G.B.-25

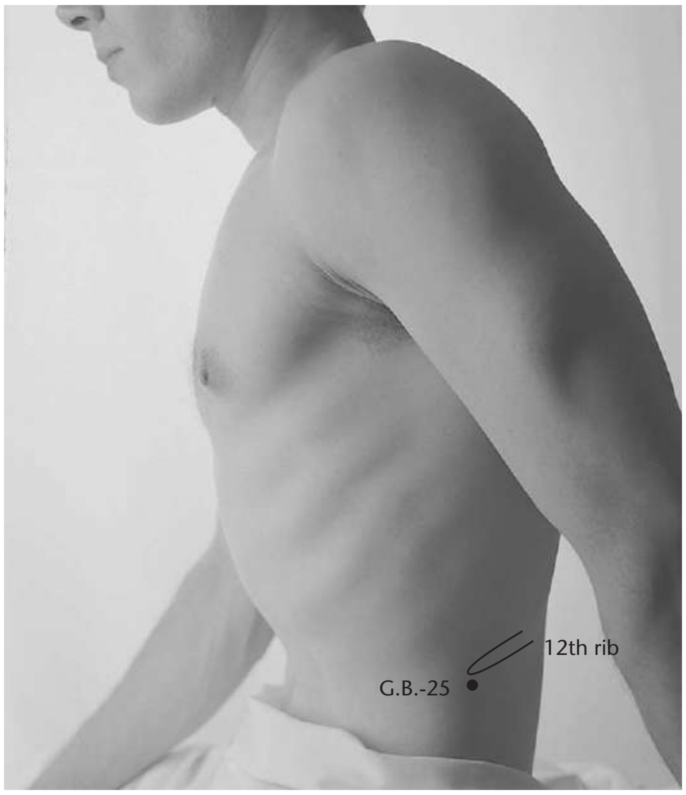

Location

On the lateral aspect of the ribcage, at the lower border of the free end of the 12th rib.

How to find

Place the hand on the upper abdomen and, with gentle pressure, glide along the lower border of the ribcage until you can feel the free end of the 11th rib (→ **LIV-13**) slightly superior to the umbilicus. By continuing to palpate along the lower border of the ribcage, you will feel the free end of the 12th rib on the lateral aspect of the waist. Locate **G.B.-25** on its lower border. **For orientation:** When pressing the flexed elbow onto the thorax, the tip of the olecranon will rest on the area around the free end of the 11th rib (→ **LIV-13**).

Needling

Vertically or obliquely 0.3–1 cun. Caution: Peritoneum; the needle should be inserted into the obliquus externus or internus muscles.

Actions/Indications

- Tonifies the Kidneys and regulates the water passages
- Strengthens the Spleen, regulates the Intestines
- Benefits the lumbar region

Special features

Front-*mu* point of the Kidneys

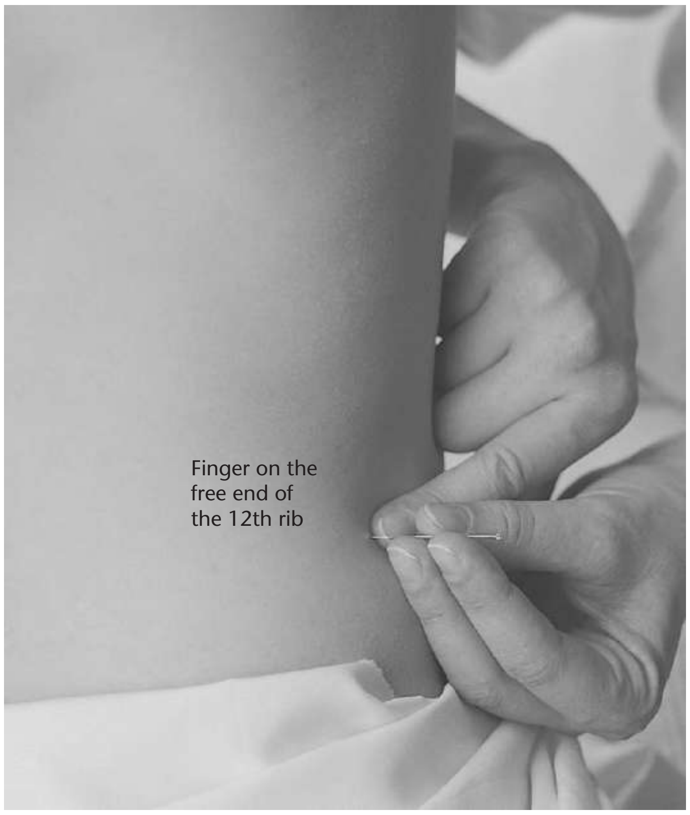

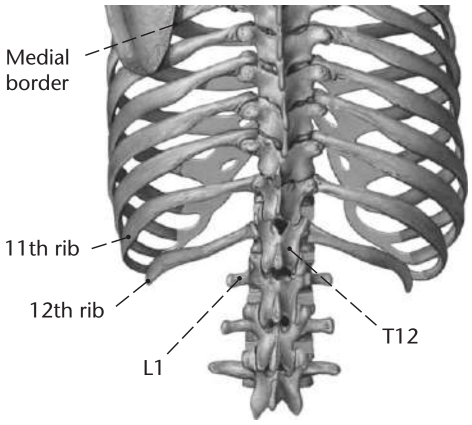

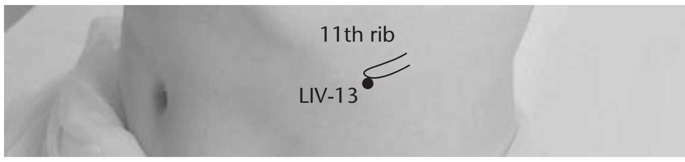

G.B.-26 *Girdling Vessel* DAIMAI

Location

On the lateral aspect of the waist, at the junction of a vertical line through the free end of the 11th rib and a horizontal line through the umbilicus, approximately 1.8 cun inferior to → **LIV-13**.

How to find

First, palpate the lower border of the ribcage to locate the free end of the 11th rib (→ **LIV-13**). Next, locate **G.B.-26** inferior to the free end of the 11th rib, level with the umbilicus. **Quick method for locating → LIV-13:** When pressing the flexed elbow onto the thorax, the tip of the olecranon will rest on the area around the free end of the 11th rib (→ **LIV-13**).

Located on the same level are → **Ren-8** (in the centre of the umbilicus), → **KID-16/ST-25/SP-15** (0.5/2/4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Be careful with slim patients.

Actions/Indications

- Regulates the *dai mai*
- Regulates the Uterus
- Drains Damp-Heat

Special features

Meeting point with the *dai mai*

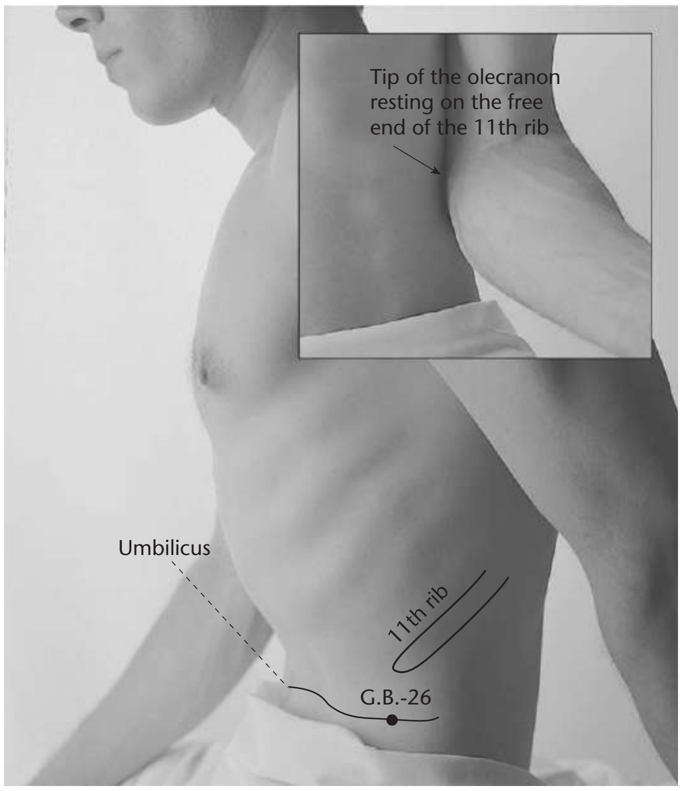
Tip of the olecranon resting on the free end of the 11th rib
Umbilicus
11th rib
G.B.-26

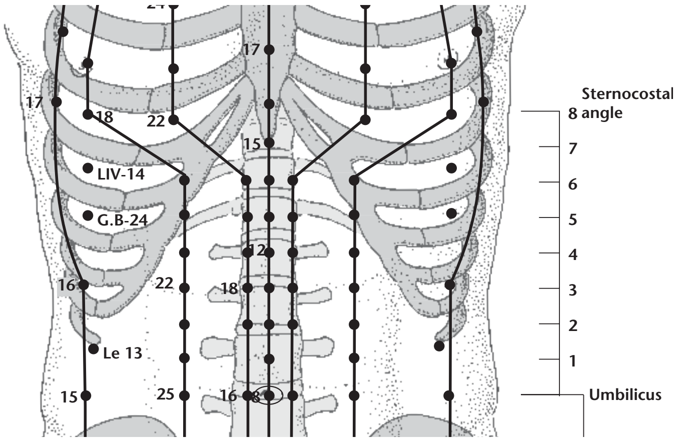
17
17
18
22
15
LIV-14
G.B-24
12
16
22
18
Le 13
15
25
16
8
Sternocostal angle
8
7
6
5
4
3
2
1
Umbilicus

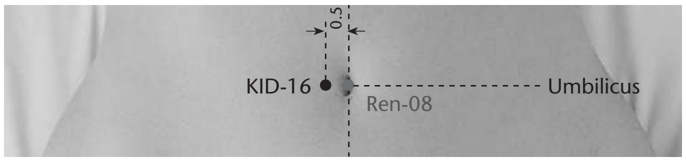
0.5
KID-16
Ren-08
Umbilicus

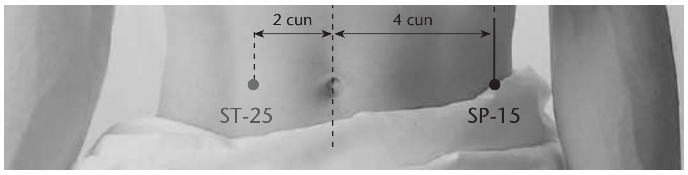
2 cun
4 cun
ST-25
SP-15

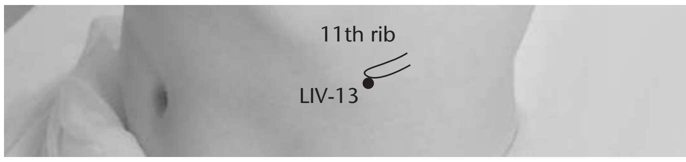
11th rib
LIV-13

Fifth Pivot WUSHU G.B.-27

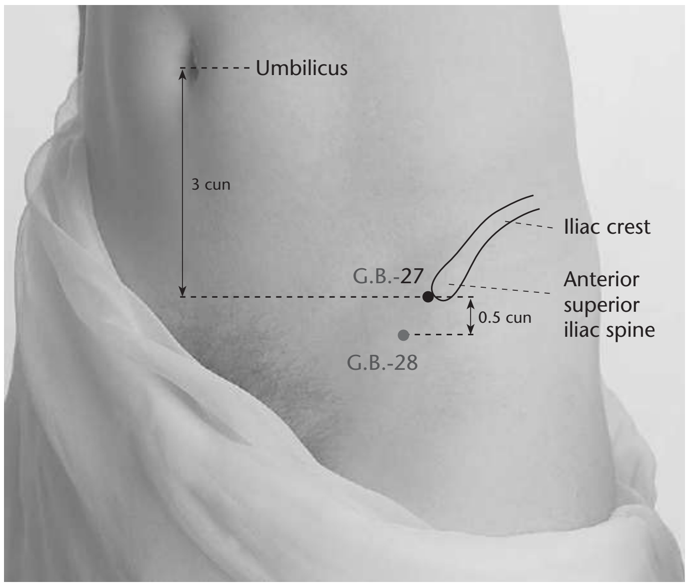
Umbilicus
3 cun
Iliac crest
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

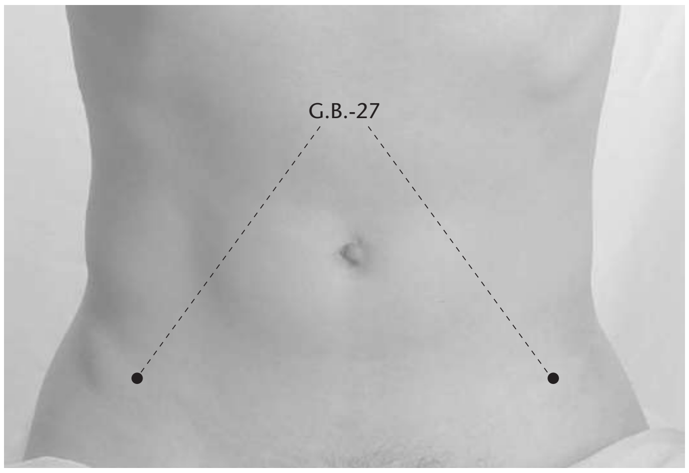

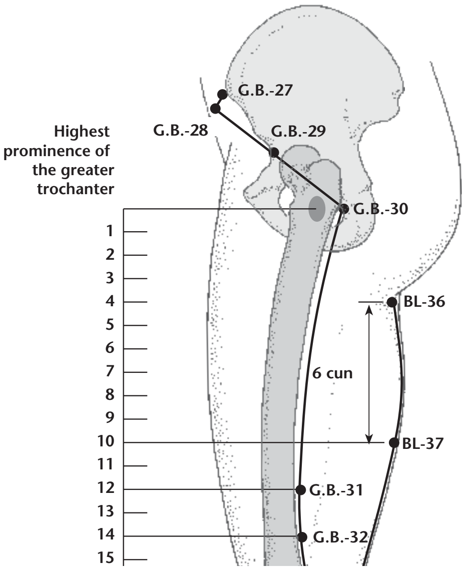
G.B.-27
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
BL-36
6 cun
BL-37
G.B.-31
G.B.-32

Location

In the depression medial to the anterior superior iliac spine (ASIS), approximately 3 cun inferior to the umbilicus.

How to find

First, locate the anterior superior iliac spine (ASIS → 3.5) by palpating along the upper border of the iliac crest in an anterior and inferior direction. At its anterior end, the ASIS can be palpated as a bony ridge on the lateral aspect of the lower abdomen. Locate **G.B.-27** in a depression anterior and medial to the ASIS. → **G.B.-28** is located approximately 0.5 cun inferior and medial to **G.B.-27**.

Located approximately on the same level (3 cun inferior to the umbilicus) are → **Ren-4/KID-13/ST-28** (on the midline/0.5 cun lateral to the midline/2 cun lateral to the midline).

Needling

Vertically 1–1.5 cun. Caution during pregnancy.

Actions/Indications

- Strengthens the Kidneys
- Regulates the *dai mai*
- Regulates the Triple Burner

Special features

Meeting point with the *dai mai*

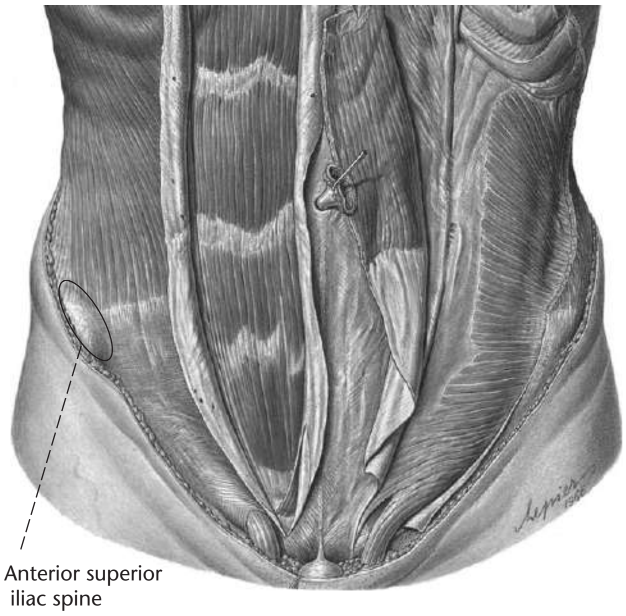
Anterior superior iliac spine

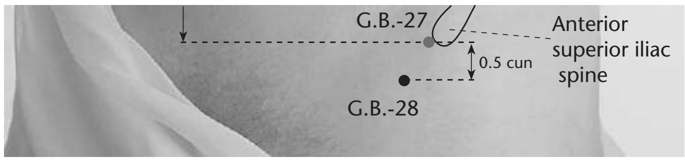
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

G.B.-28 *Linking Path* WEIDAO

Location

On the lateral aspect of the abdomen, anterior and inferior to the anterior superior iliac spine (ASIS), approximately 0.5 cun anterior and inferior to ➞ **G.B.-27**.

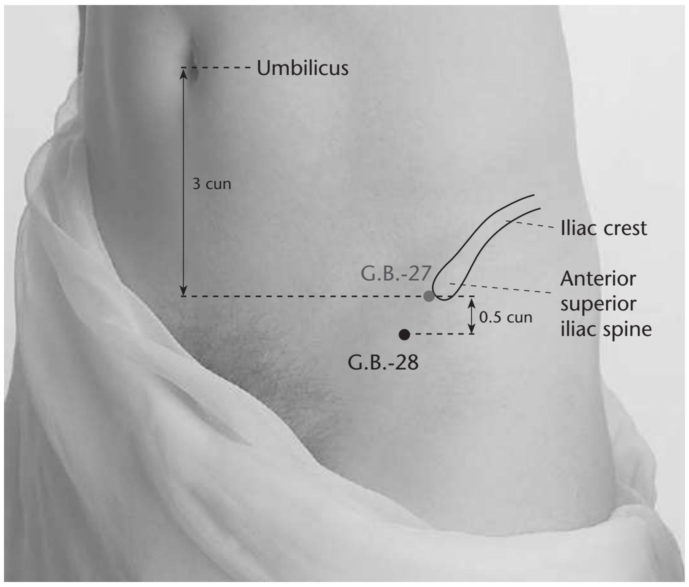
Umbilicus
3 cun
Iliac crest
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

How to find

First, locate the ASIS (➞ 3.5) by palpating along the upper border of the iliac crest in an anterior and inferior direction. On its anterior border, the ASIS can be palpated as a bony ridge on the lateral aspect of the lower abdomen, with ➞ **G.B.-27** located in a depression anterior and medial to the ASIS. Locate **G.B.-28** approximately 0.5 cun anterior and inferior to ➞ **G.B.-27**.

Needling

Vertically 1–1.5 cun. Caution during pregnancy.

Actions/Indications

- Regulates the *dai mai*
- Regulates the Triple Burner
- Eliminates Dampness

Special features

Meeting point with the *dai mai*

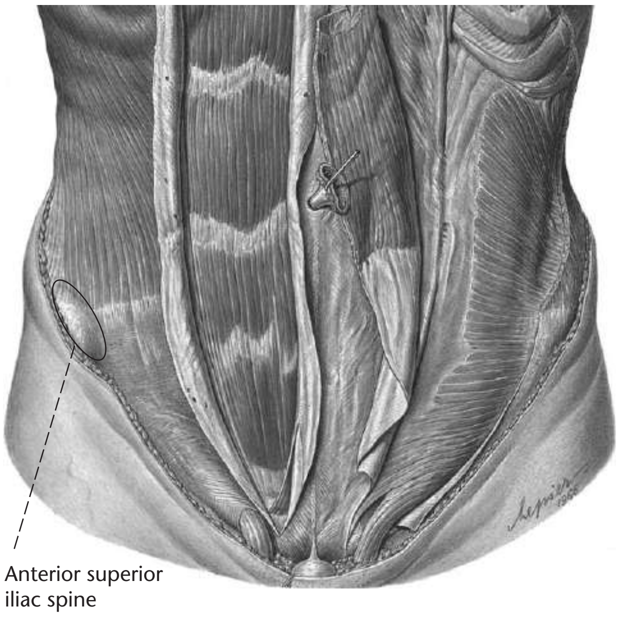
Anterior superior iliac spine

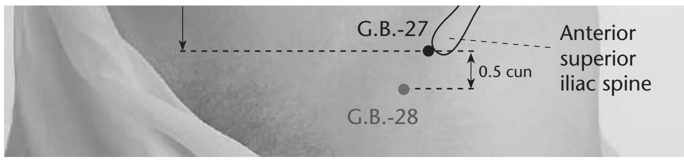
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

Stationary Crevice JULIAO **G.B.-29**

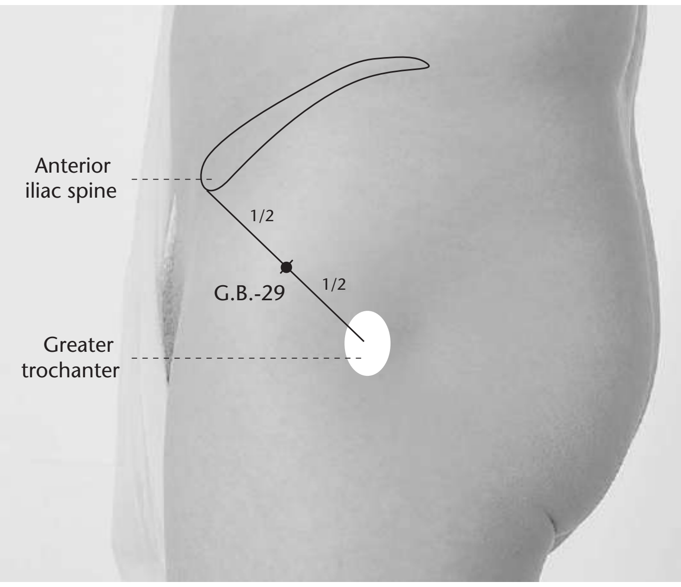
Anterior iliac spine
1/2
G.B.-29
1/2
Greater trochanter

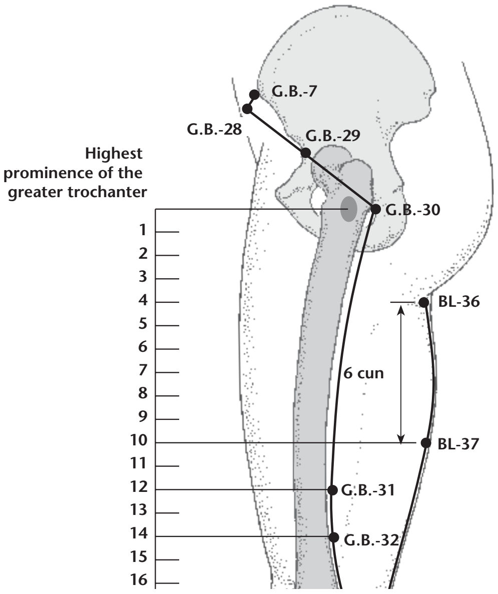
G.B.-7
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
BL-36
6 cun
BL-37
G.B.-31
G.B.-32

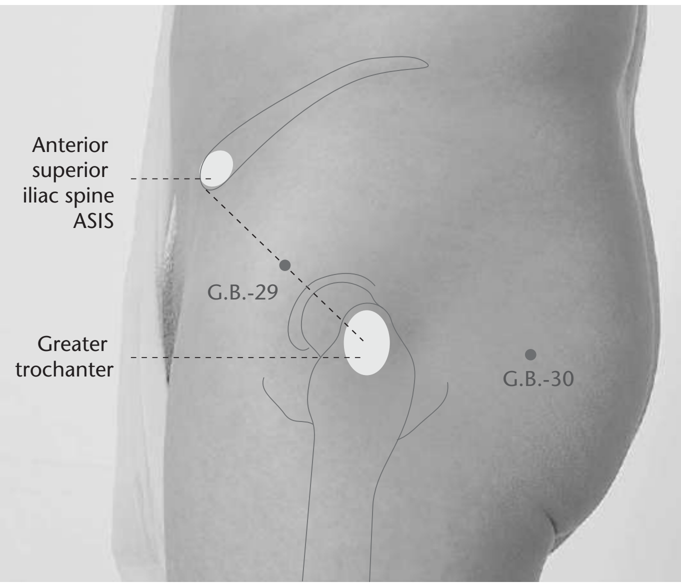
Anterior superior iliac spine ASIS
G.B.-29
Greater trochanter
G.B.-30

Location

At the midpoint of a line connecting the anterior superior iliac spine (ASIS) and the greater trochanter, at the anterior border of the iliac crest.

How to find

First, locate the ASIS (→ 3.5), the highest point on the anterior aspect of the iliac crest, by palpating along the upper border of the iliac crest in an anterior and inferior direction. At its anterior end, the ASIS can be palpated as a bony ridge on the lateral aspect of the lower abdomen. With the patient lying on their side and their leg slightly flexed, locate the greater trochanter (→ 3.6), a clearly marked bony structure in the region of the hip joint. **G.B.-29** is located at the midpoint of a line connecting these two reference points.

Needling

Vertically 1–2 cun

Actions/Indications

- Expels Cold, Dampness and Wind, opens the channel

Special features

Meeting point with the *yang qiao mai*, also with the *dai mai* according to some authors.

G.B.-30 *Jumping Circle* HUANTIAO

Location

With the patient lying on their side, the point is at the junction between the medial two-thirds and the lateral third of a line connecting the greater trochanter and the sacral hiatus.

How to find

Patient's position: Supine or, better, lying on their side, preferably with the hip and knee joints of the side to be needled flexed and the lower leg straight. Use pillows, etc, for a comfortable position. Reference points: the sacral hiatus (→ 3.4.4) and the lateral prominence of the greater trochanter (→ 3.6). **G.B.-30** is located on a line connecting these two points, one third of the distance from the greater trochanter.

Needling

1.5–3 cun vertically towards the genital region. The needle may reach the fascia of the obturator internus muscle as well as intermuscular connective tissue. Long 3 cun needles (50mm) should be used. Caution: Needling is often painful. Needling can result in an electric *de qi* sensation radiating to the toes. This sensation is particularly common with sciatic disorders and if **G.B.-30** is needled 1 cun inferior to its normal location.

Actions/Indications

- Opens the channel and the *luo*-connecting vessels, alleviates pain, benefits the hip and legs, eliminates Wind-Damp.

Special features

Meeting point with the BL channel, Ma Dan Yang Heavenly Star point. Important point for disorders of the hips.

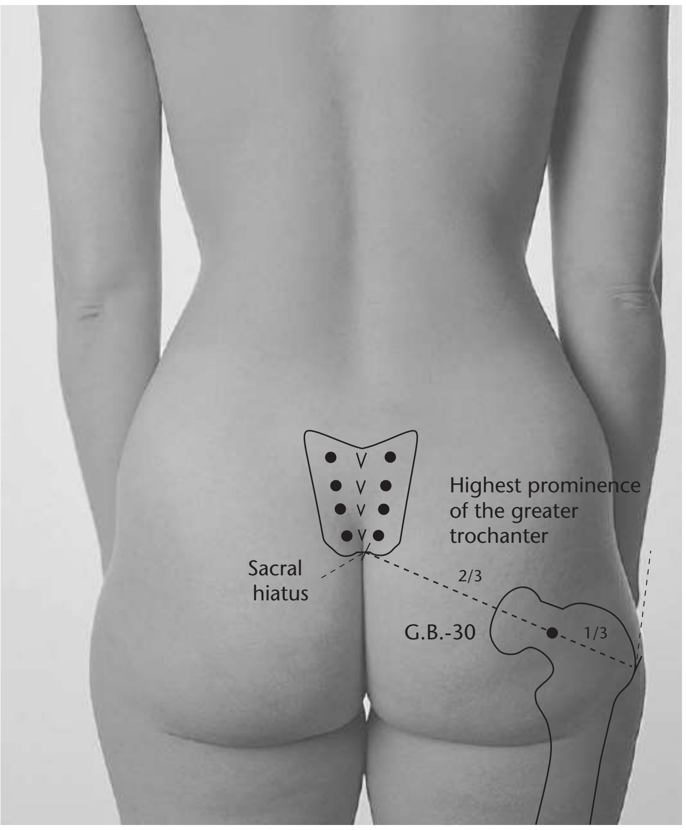

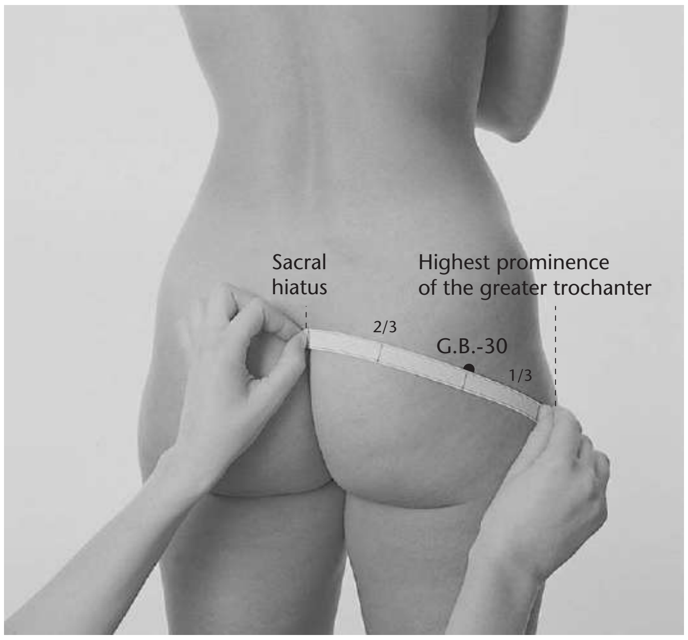

Wind Market FENGSHI G.B.-31

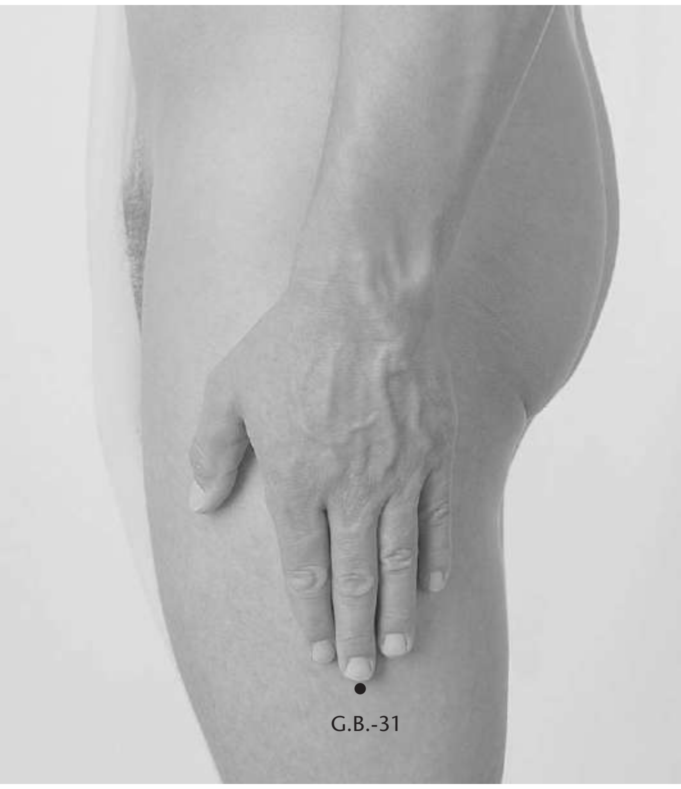

Location

On the lateral aspect of the thigh, inferior to the greater trochanter, approximately 7 cun proximal to the popliteal crease.

How to find

Ask the patient to place their hands on the imaginary seam of their trousers (this is best done with the patient standing). **G.B.-31** is located where the middle finger touches the lateral aspect of the thigh. **Or:** The distance between the highest prominence of the greater trochanter (→ 3.6) to the popliteal crease is 19 cun (→ 2.2). Divide this distance into thirds and locate **G.B.-31** 1 cun proximal to a third of the distance from the popliteal crease. Sensitivity to pressure should help determine the location of this point.

Needling

Vertically 1–2 cun

Actions/Indications

- Eliminates Wind, Dampness and Heat
- Opens the channel and alleviates pain

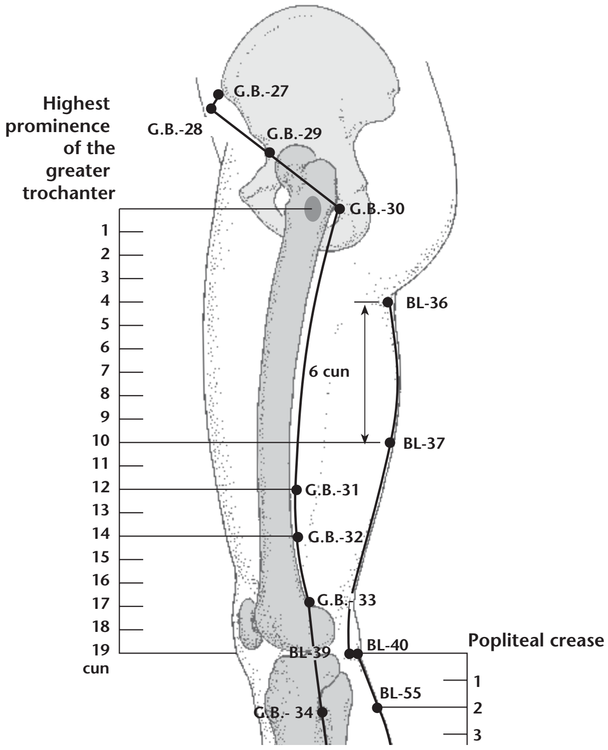
Highest prominence of the greater trochanter
G.B.-27
G.B.-28
G.B.-29
G.B.-30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-39
BL-40
Popliteal crease
BL-55
G.B.- 34
1
2
3

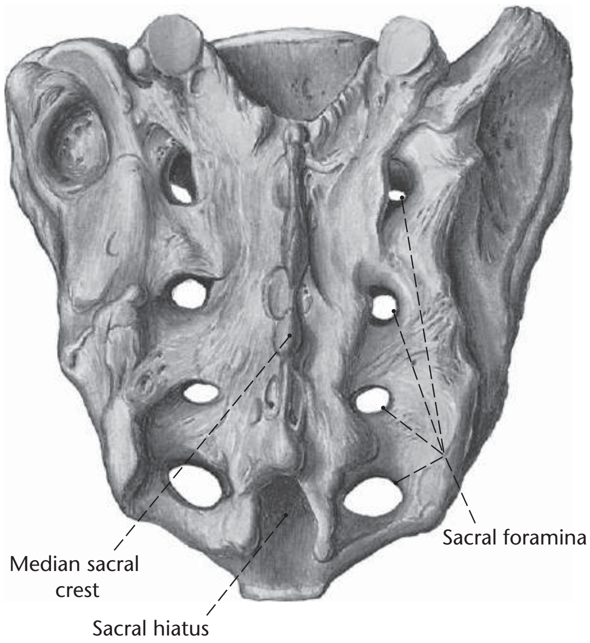
Median sacral crest
Sacral hiatus
Sacral foramina

G.B.-32 *Middle Ditch* ZHONGDU

Location

On the lateral aspect of the thigh, 5 cun proximal to the popliteal crease, between the vastus lateralis and biceps femoris muscles.

How to find

First, ask the patient to place their hands on the imaginary seam of their trousers (this is best done with the patient standing) and locate → **G.B.-31** where the middle finger touches the thigh. **G.B.-32** can now be located 2 cun distal to → **G.B.-31. Or:** The distance between the highest prominence of the greater trochanter (→ 3.6) to the popliteal crease is 19 cun (→ 2.2). Divide this distance into quarters (use an elastic tape or the spreading hands technique) and locate **G.B.-32** slightly proximal to a quarter of the distance from the popliteal crease. Sensitivity to pressure should help determine the location of this point.

Needling

Vertically 1–2 cun

Actions/Indications

- Eliminates Wind, Dampness and Cold, opens the channel

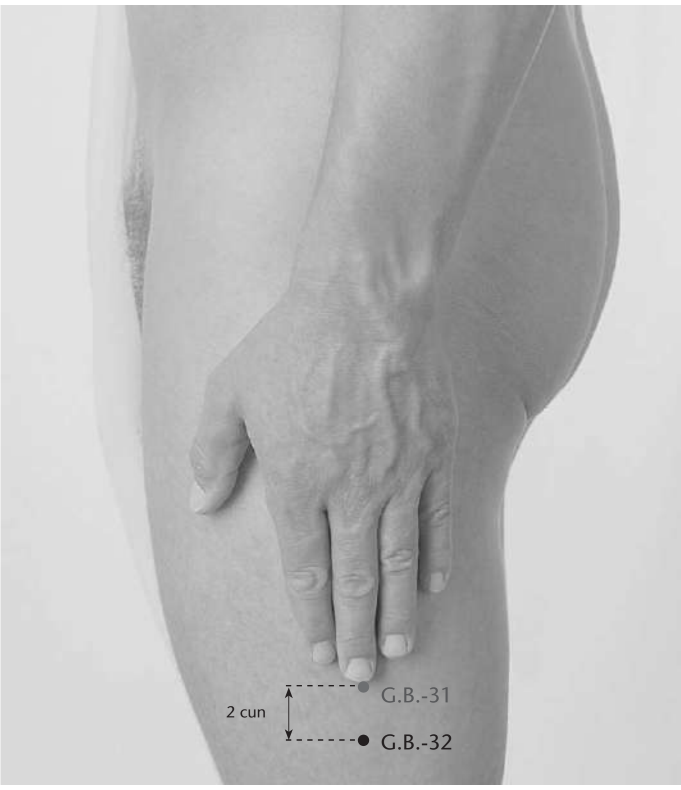
2 cun
G.B.-31
G.B.-32

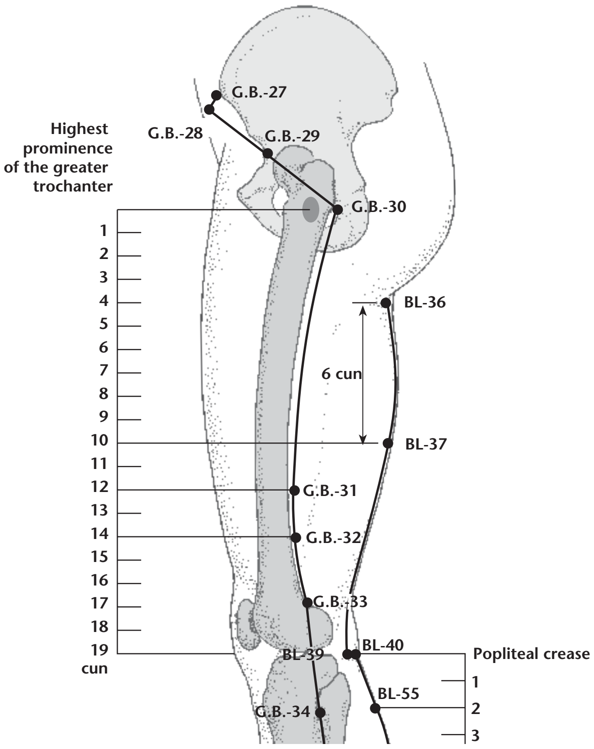
G.B.-27
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-40
BL-39
Popliteal crease
1
2
3
BL-55
G.B.-34

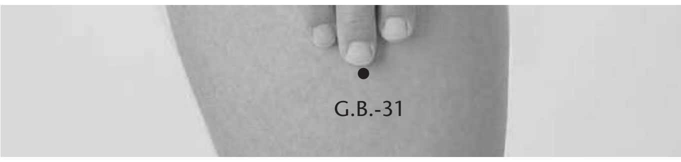

Knee Yang Gate XIYANGGUAN G.B.-33

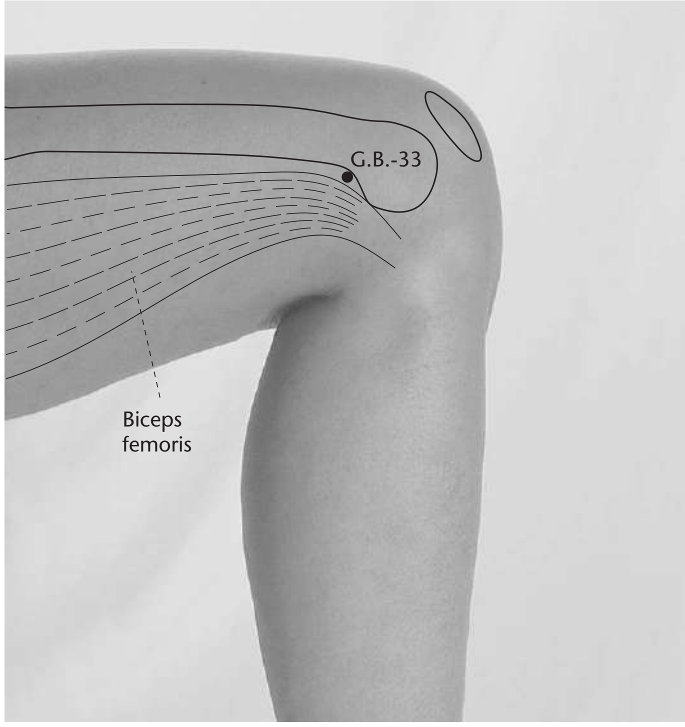

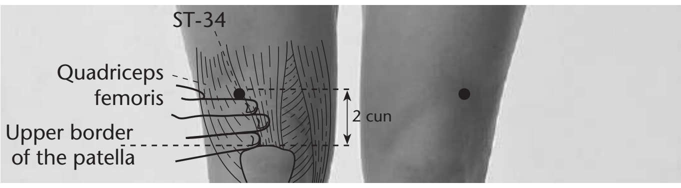

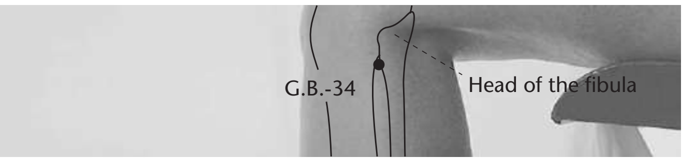

Location

On the lateral aspect of the knee. With the knee flexed, in the depression between the shaft and the lateral epicondyle and the tendon of the biceps femoris muscle, approximately 3 cun proximal to → **G.B.-34**.

How to find

This point is best located with the patient's knee flexed. On the level of the lateral upper border of the patella, palpate in a lateral direction towards the thigh and palpate for the distal end of the lateral epicondyle of the femur. Then locate **G.B.-33** in a clearly palpable depression between the epicondyle and the tendon of the biceps femoris muscle. The latter runs along the lateral aspect of the leg (along the line of an imaginary trouser seam) and attaches inferior to the knee joint at the head of the fibula.
→ **G.B.-34** is located 3 cun more distal, → **ST-34** is located 2 cun proximal to the lateral upper border of the patella.

Needling

Vertically 1–2 cun

Actions/Indications

- Opens the channel
- Relaxes the tendons

G.B.-34 *Yang Mound Spring* YANGLINGQUAN

Location

In the depression anterior and inferior to the head of the fibula, between the peroneus longus and extensor digitorum longus muscles.

How to find

This point is best located with the patient's knee flexed (use knee support). Palpate for the head of the fibula on the lateral aspect of the leg and hold it with the index and middle fingers in a tweezer-like way. Gliding distally with both fingers, the more medial finger will drop into a depression directly anterior and inferior to the head of the fibula, the location of **G.B.-34**.
Located on the same level but on the medial aspect of the lower leg is → **SP-9** (at the junction of the shaft and the medial condyle of the tibia).

Needling

1–1.5 cun vertical insertion between the tibia and fibula towards the interosseous membrane. Caution: Deep peroneal nerve with deep needling, in some cases also common peroneal nerve. The needle may reach the interosseous membrane as well as the epineural tissue of the peroneal nerve.

Actions/Indications

- Benefits the tendons and joints
- Opens the channel, alleviates pain, benefits the lateral costal region
- Clears Damp-Heat from the Liver and Gall Bladder
- Spreads Liver Qi

Special features

He-sea point, Earth point, *hui*-meeting point of the tendons (coordination/movement), lower *he*-sea point of the Gall Bladder, Ma Dan Yang Heavenly Star point. Major point for disorders of the tendons and musculature.

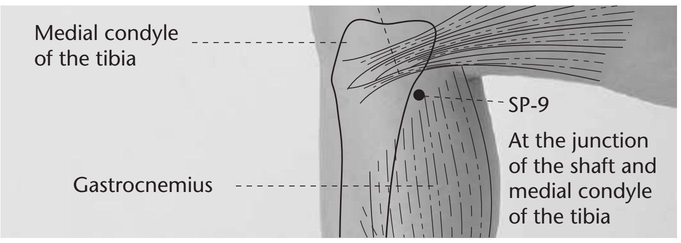
Medial condyle
of the tibia
SP-9
At the junction
of the shaft and
medial condyle
of the tibia
Gastrocnemius

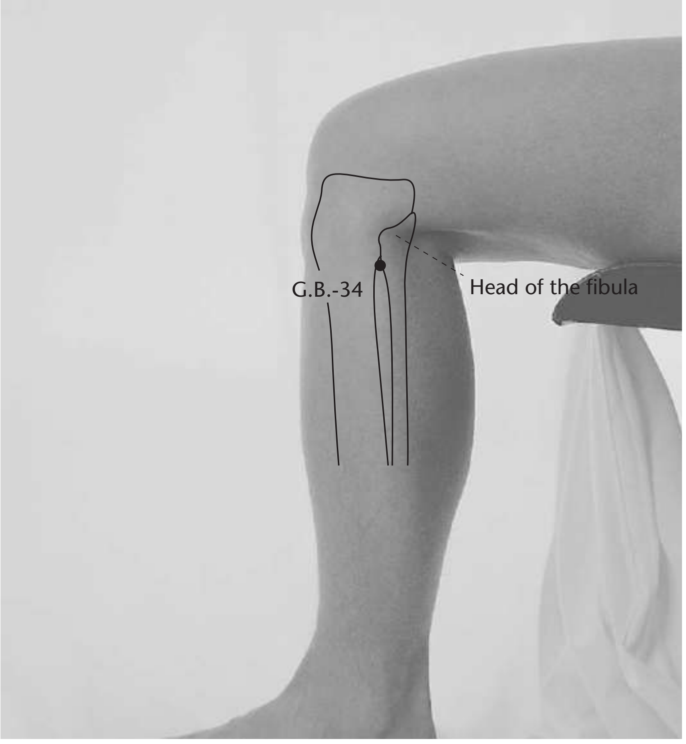
G.B.-34
Head of the fibula

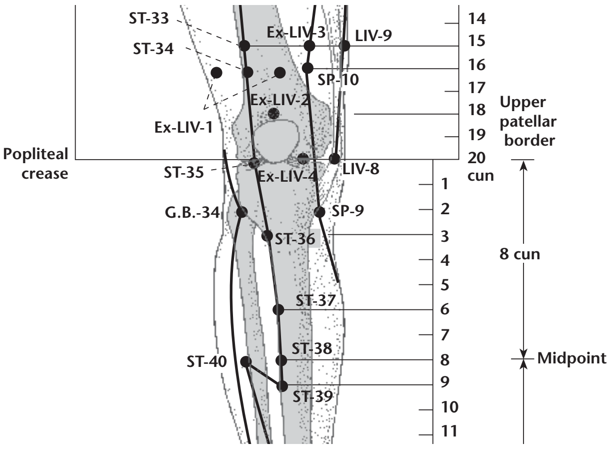
ST-33
ST-34
Ex-LIV-3
LIV-9
SP-10
Ex-LIV-2
Ex-LIV-1
Popliteal
crease
ST-35
Ex-LIV-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
14
15
16
17
18
19
20
cun
Upper
patellar
border
1
2
3
4
5
6
7
8
9
10
11
8 cun
Midpoint

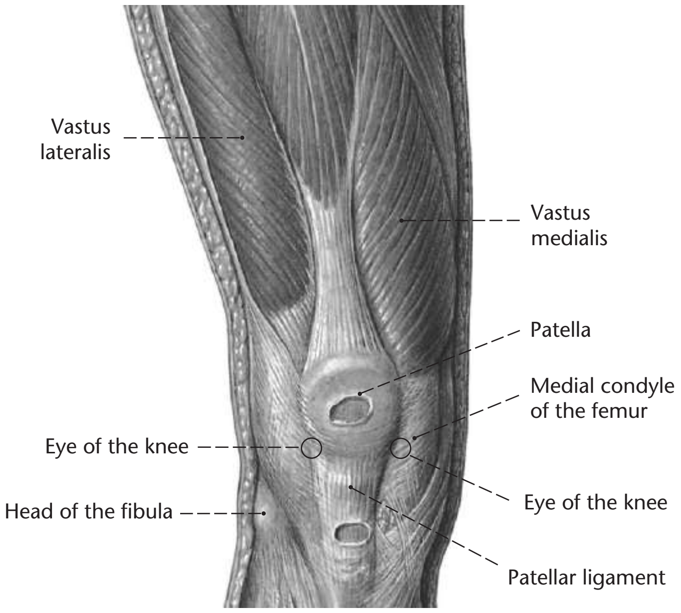
Vastus
lateralis
Vastus
medialis
Patella
Medial condyle
of the femur
Eye of the knee
Eye of the knee
Head of the fibula
Patellar ligament

Yang Intersection YANGJIAO **G.B.-35**

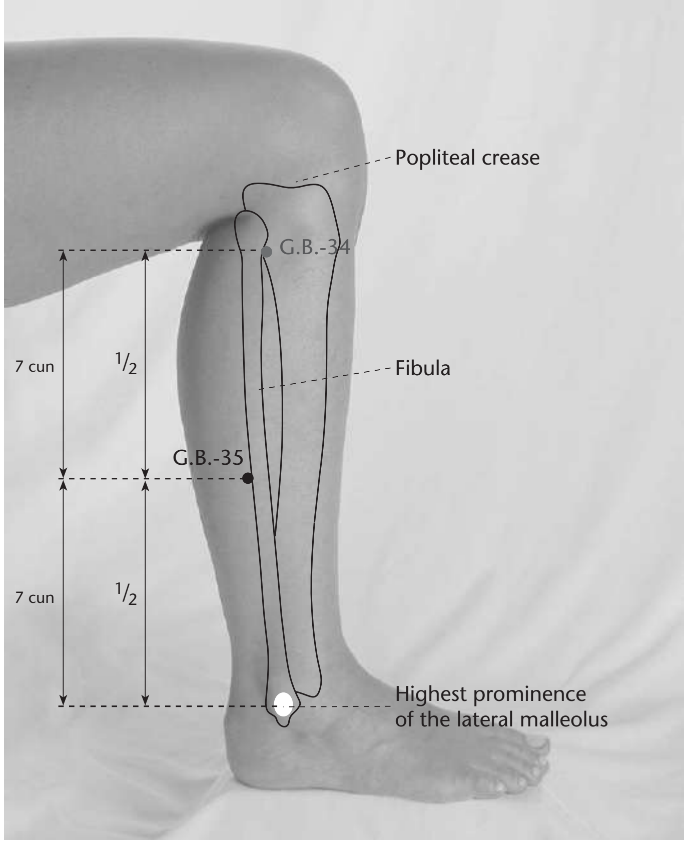
Popliteal crease
G.B.-34
7 cun
1/2
Fibula
G.B.-35
7 cun
1/2
Highest prominence of the lateral malleolus

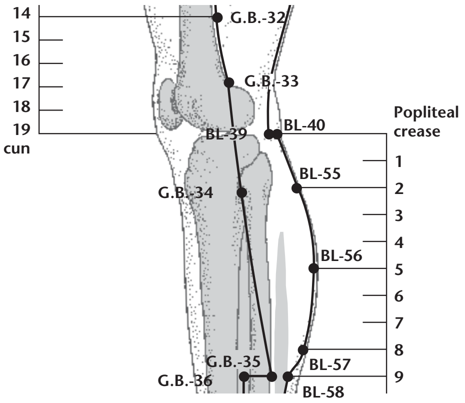
14
15
16
17
18
19
cun
G.B.-32
G.B.-33
BL-39
BL-40
Popliteal crease
BL-55
G.B.-34
BL-56
G.B.-35
G.B.-36
BL-57
BL-58
1
2
3
4
5
6
7
8
9

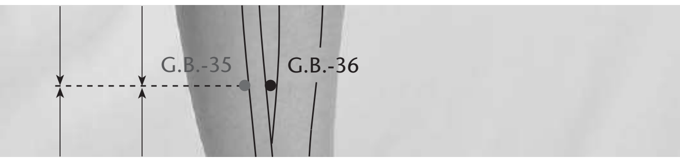

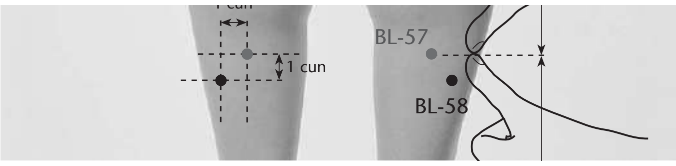
1 cun
BL-57
BL-58

Location

7 cun proximal to the highest prominence of the lateral malleolus, on the **posterior** border of the fibula.

How to find

Quick method: Spreading hands technique (→ 2.3.3): Locate **G.B.-35** at the midpoint of a line connecting → **G.B.-34** (in the depression anterior and inferior to the head of the fibula) and the highest prominence of the lateral malleolus (→ 3.6.2), on the **posterior** border of the fibula (distance = 14 cun). **G.B.-35** is located 7 cun proximal to the prominence of the lateral malleolus. **Or:** Spreading hands technique (→ 2.3.3): Locate **G.B.-35** 1 cun distal to the midpoint of a line connecting the popliteal crease and the highest prominence of the lateral malleolus, on the **posterior** border of the fibula (distance = 16 cun, → 2.2). **For orientation:** The borders of the fibula are deep to the peroneus brevis muscle and are often not easily palpable. For this reason, it is suggested to palpate the posterior border of the fibula just superior to the lateral malleolus and then locate the point on an imaginary line running to the head of the fibula.

Located on the same level are → **G.B.-36** (on the **anterior** border of the fibula) and → **BL-58** (7 cun proximal to → **BL-60**).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Clears Heat, opens the channel, relaxes the tendons
- Regulates Qi of the Gall Bladder
- Calms the *shen*

Special features

Xi-cleft point of the *yang wei mai*, meeting point with the *yang wei mai* according to some authors.

G.B.-36 *Outer Mound* WAIQIU

Location

7 cun proximal to the highest prominence of the lateral malleolus, on the **anterior** border of the fibula.

How to find

Quick method: Spreading hands technique (→ 2.3.3): Locate → **G.B.-35** at the midpoint of a line connecting → **G.B.-34** (in the depression anterior and inferior to the head of the fibula) and the highest prominence of the lateral malleolus (→ 3.6.2), on the **anterior** border of the fibula (distance = 14 cun). **G.B.-36** is located 7 cun proximal to the prominence of the lateral malleolus. **Or:** Spreading hands technique (→ 2.3.3): locate **G.B.-36** 1 cun distal to the midpoint of a line connecting the popliteal crease and the highest prominence of the lateral malleolus, on the **anterior** border of the fibula (distance = 16 cun, → 2.2). **For orientation:** The borders of the fibula are deep to the peroneus brevis muscle and are often not easily palpable. For this reason, it is suggested to palpate the anterior border of the fibula just superior to the ankle and then locate the point on an imaginary line running to the head of the fibula.

Located on the same level are → **G.B.-35** (on the **posterior** border of the fibula) and → **BL-58** (7 cun proximal to → **BL-60**).

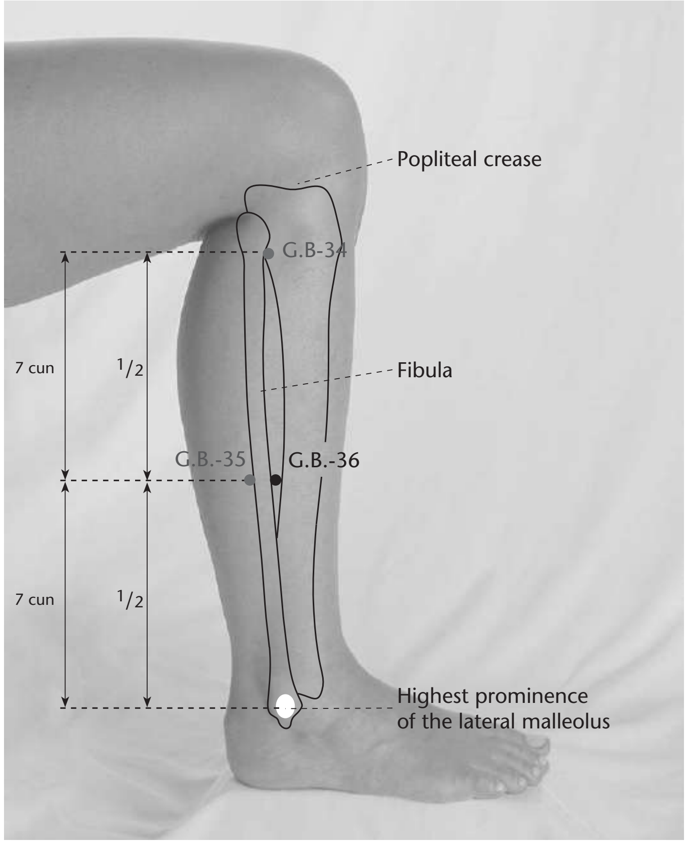

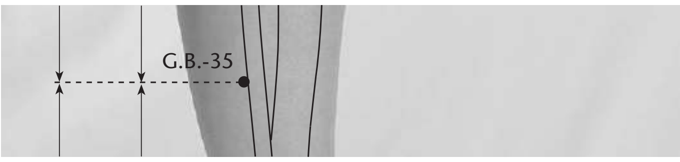

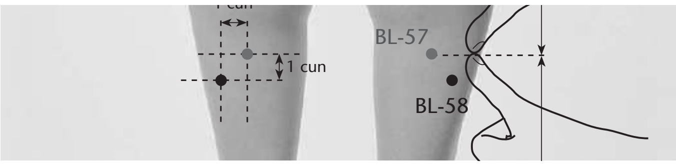

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Regulates Gall Bladder and Liver Qi
- Drains Damp-Heat
- Relaxes the tendons and muscles
- Calms the *shen*

Special features

Xi-cleft point

Bright Light GUANGMING **G.B.-37**

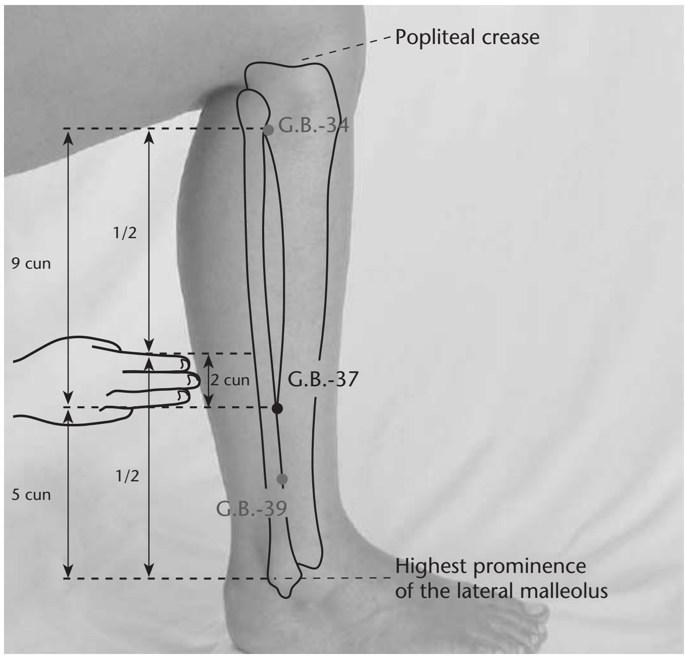
Popliteal crease
G.B.-34
1/2
9 cun
2 cun
G.B.-37
1/2
5 cun
G.B.-39
Highest prominence of the lateral malleolus

Location

5 cun proximal to the highest prominence of the lateral malleolus, on the anterior border of the fibula, between the peroneus longus and extensor digitorum longus muscles.

How to find

Quick method: Spreading hands technique (➞ 2.3.3): Place the little fingers on ➞ **G.B.-34** (in the depression anterior and inferior to the head of the fibula) and the highest prominence of the lateral malleolus (➞ 3.6.2). This distance is 14 cun. From the midpoint of this distance, measure 2 cun in a distal direction and locate **G.B.-37** on this level in a depression on the **anterior** border of the fibula. This depression is located 5 cun proximal to the prominence of the lateral malleolus (the width of 1 hand and 2 thumbs). **For orientation:** The borders of the fibula are deep to the peroneus brevis muscle and are often not easily palpable. For this reason, it is suggested to palpate the anterior border of the fibula just superior to the ankle and then locate the point on an imaginary line running to the head of the fibula.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Benefits the eyes
- Eliminates Wind-Damp, opens the channel, alleviates pain

Special features

Luo-connecting point. Important distal point for disorders of the eyes.

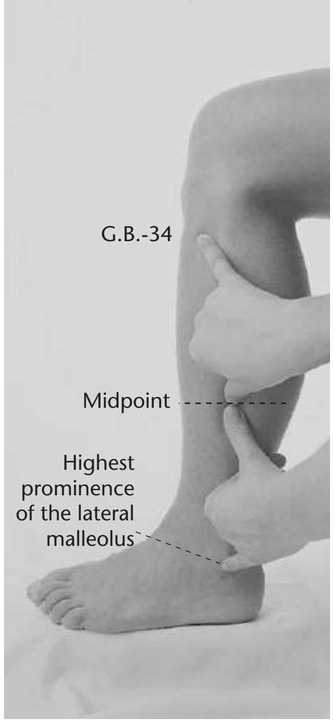
G.B.-34
Midpoint
Highest prominence of the lateral malleolus

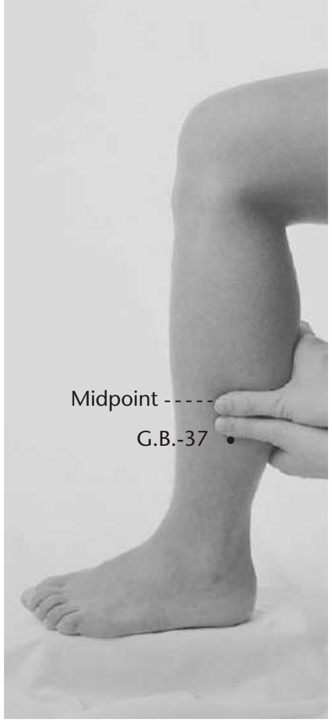
Midpoint
G.B.-37

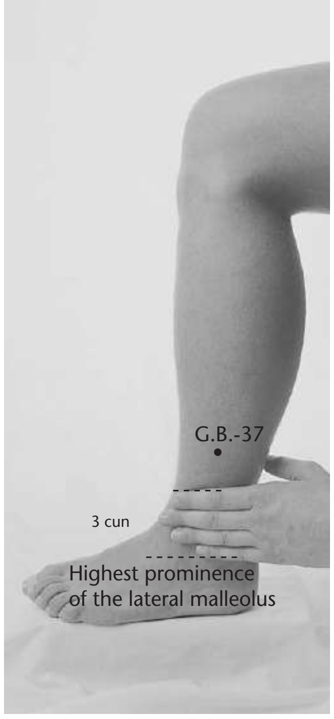
G.B.-37
3 cun
Highest prominence of the lateral malleolus

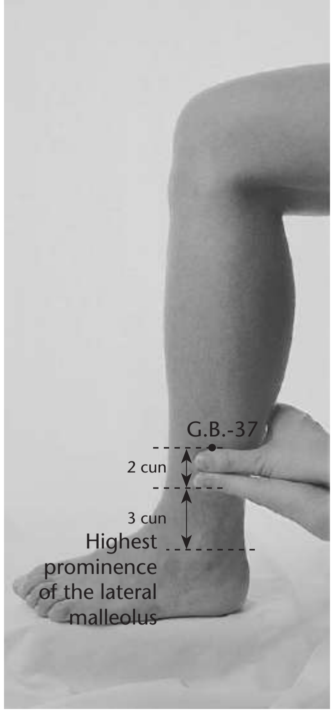
G.B.-37
2 cun
3 cun
Highest prominence of the lateral malleolus

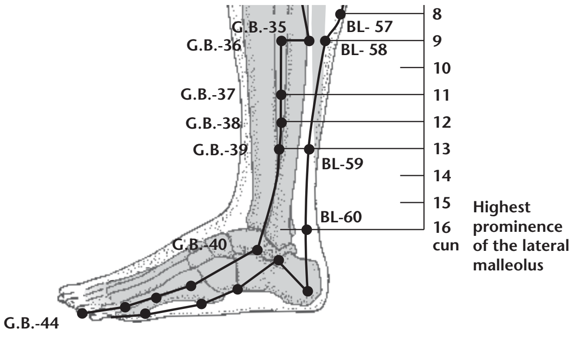
G.B.-35
BL- 57
G.B.-36
BL- 58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
G.B.-40
G.B.-44
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus

Ch04.11-F10028.qxd 2/23/08 11:25 AM Page 436

G.B.-38 *Yang Assistance* YANGFU

Location

4 cun proximal to the highest prominence of the lateral malleolus, on the **anterior** border of the fibula.

How to find

First, locate the prominence of the lateral malleolus (→ 3.6.2) and, from there, measure 4 cun in a proximal direction. There, locate **G.B.-38** on the **anterior** border of the fibula. **Or:** On the lateral aspect of the lower leg, the distance between the prominence of the lateral malleolus (→ 3.6.2) and the popliteal crease is 16 cun (→ 2.2). Divide this distance into quarters and locate **G.B.-38** one quarter of the distance from the lateral malleolus (aid: spreading hands technique or elastic tape → 2). **For orientation:** The borders of the fibula are deep to the peroneus brevis muscle and are often not easily palpable. For this reason, it is suggested to palpate the anterior border of the fibula just superior to the ankle and then locate the point on an imaginary line running to the head of the fibula.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Opens the channel
- Clears Wind and Heat

Special features

Jing-river point, Fire point, sedation point.

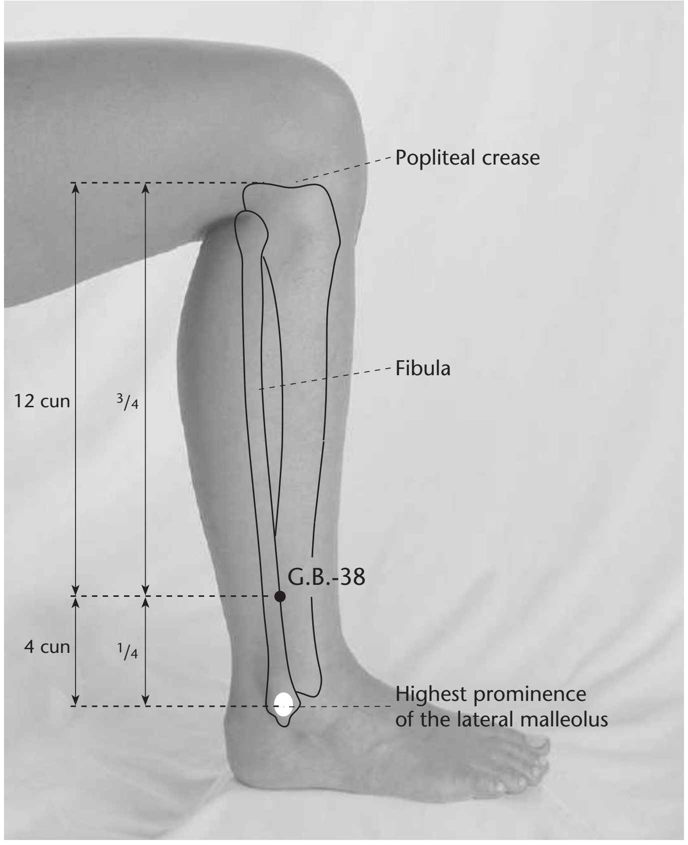
Popliteal crease
Fibula
12 cun
$^3/_4$
G.B.-38
4 cun
$^1/_4$
Highest prominence of the lateral malleolus

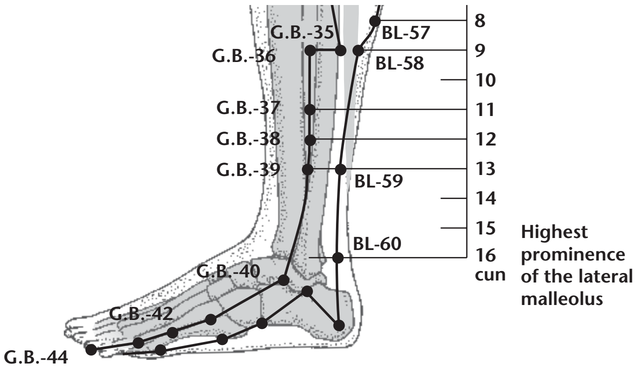
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
G.B.-40
G.B.-42
G.B.-44
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus

Hanging Bell XUANZHONG G.B.-39

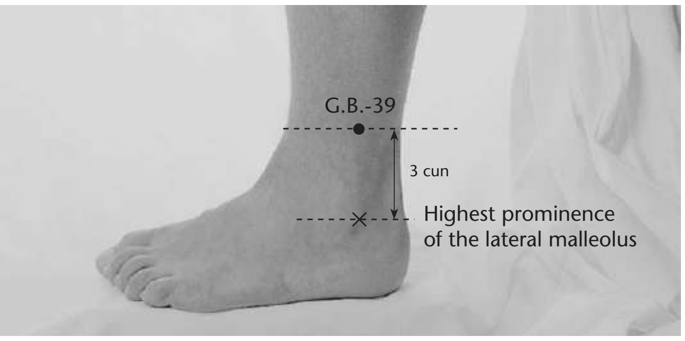
G.B.-39
3 cun
Highest prominence
of the lateral malleolus

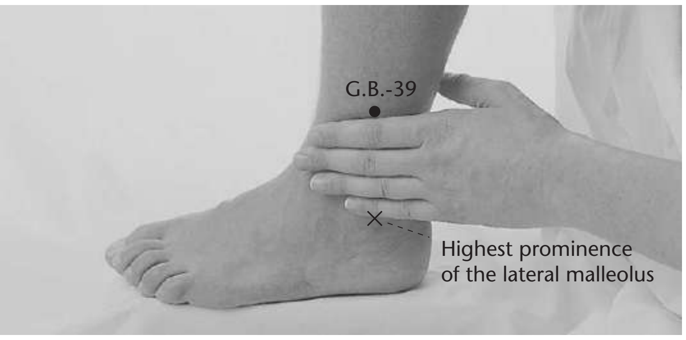
G.B.-39
Highest prominence
of the lateral malleolus

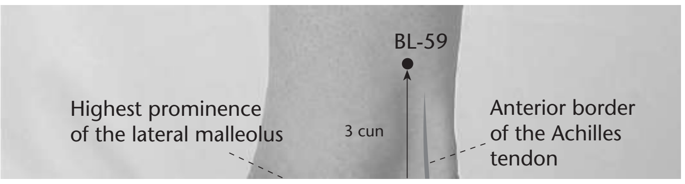
BL-59
Highest prominence
of the lateral malleolus
3 cun
Anterior border
of the Achilles
tendon

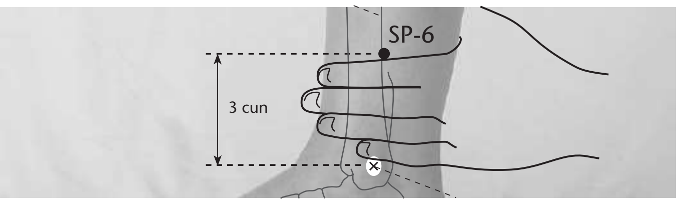

Location

3 cun proximal to the highest prominence of the lateral malleolus, on the **anterior** border of the fibula.

How to find

From the highest prominence of the lateral malleolus (→ 3.6.2), measure 3 cun (1 handbreadth) in a proximal direction and locate **G.B.-39** in a depression on the **anterior** border of the fibula. According to some authors, it is located between the posterior border of the fibula and the tendons of the peroneus longus and brevis musles. When in doubt, choose the more pressure-sensitive point.

Located on the same level but 3 cun proximal to → **BL-60** (depression between the malleolus and the Achilles tendon) is → **BL-59**. In a comparable position but on the medial aspect of the leg is → **SP-6** (3 cun proximal to the highest prominence of the medial malleolus, meeting point of the three leg Yin channels).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Opens the channel, benefits the tendons and bones
- Benefits the Marrow, clears Wind-Dampness
- Clears Heat from the Gall Bladder

Special features

Hui-meeting point of the Marrow. Important distal point for disorders of the cervical spine.

G.B.-40 *Mound of Ruins* QIUXU

Location

In the depression anterior and inferior to the lateral malleolus, lateral to the tendons of the extensor digitorum longus muscle.

How to find

For easier location, ask the patient to flex their ankle at a 90° angle. From the lateral malleolus (→ 3.6.2), let the palpating finger glide into a well-defined depression anterior and inferior to the malleolus. By lifting the toes, the tendons of the extensor digitorum longus muscle as well as the depression (**G.B.-40**) will become more pronounced. **For orientation: G.B.-40** is located at the junction of a vertical line along the anterior border of the lateral malleolus and a horizontal line along its lower border.

→ **SP-5** is located in a comparable position but on the medial aspect of the ankle (in the depression anterior and inferior to the medial malleolus). On a line connecting → **SP-5** and **G.B.-40**, → **LIV-4** is located medial to and → **ST-41** is located lateral to the tendon of the extensor hallucis longus muscle, which runs to the big toe.

Needling

0.5–1.5 cun vertically or slightly obliquely in the area of the fibulotarsal ligaments towards the medial arch of the foot.

Actions/Indications

- Opens the channel, benefits the joints
- Spreads Liver Qi, clears Heat and Dampness from the Gall Bladder

Special features

Yuan-source point

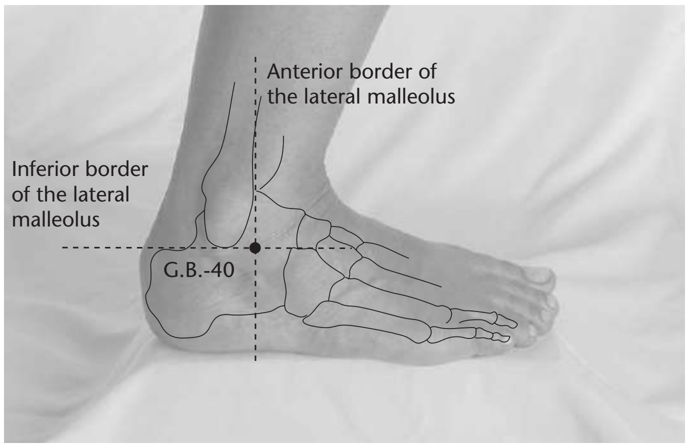
Anterior border of
the lateral malleolus
Inferior border
of the lateral
malleolus
G.B.-40

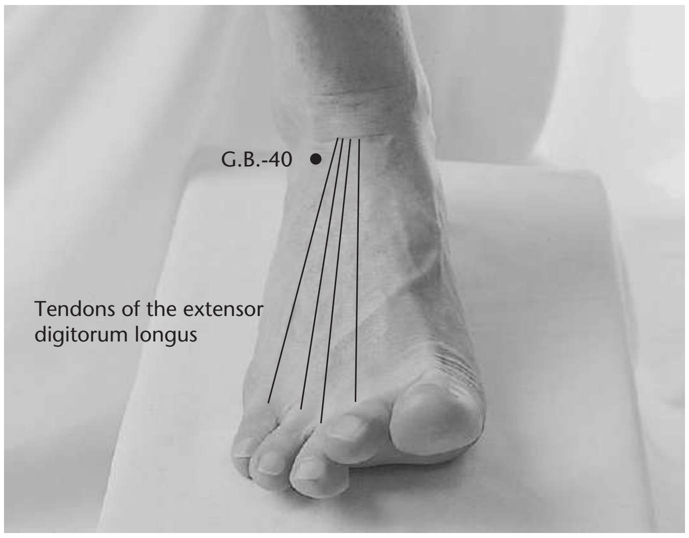
G.B.-40
Tendons of the extensor
digitorum longus

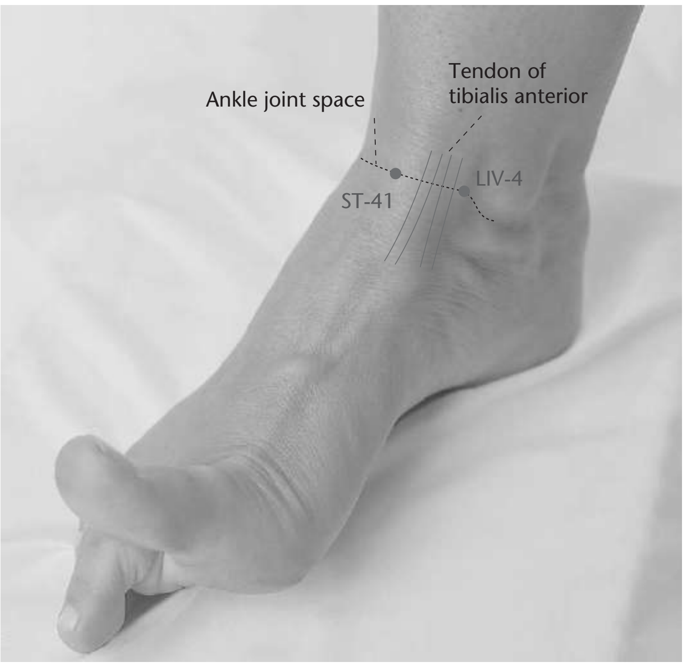
Ankle joint space
Tendon of
tibialis anterior
ST-41
LIV-4

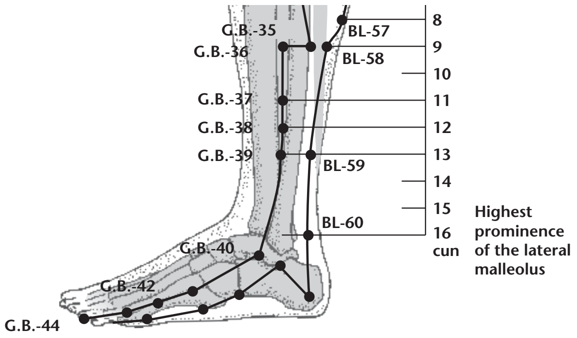
G.B.-35
G.B.-36
G.B.-37
G.B.-38
G.B.-39
G.B.-40
G.B.-42
G.B.-44
BL-57
BL-58
BL-59
BL-60
8
9
10
11
12
13
14
15
16
cun
Highest
prominence
of the lateral
malleolus

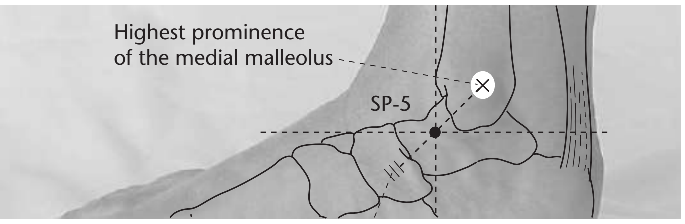
Highest prominence
of the medial malleolus
SP-5

Foot Governor of Tears ZULINQI **G.B.-41**

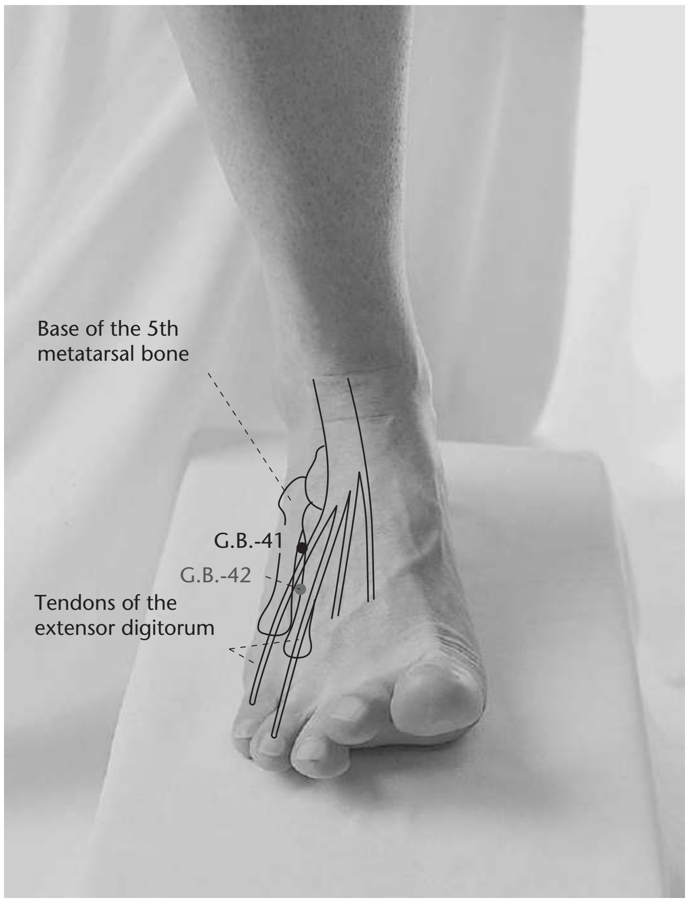
Base of the 5th metatarsal bone
G.B.-41
G.B.-42
Tendons of the extensor digitorum

Location

In the depression at the junction of the shafts and the bases of the 4th and 5th metatarsal bones, lateral to the tendon of the extensor digitorum longus muscle.

How to find

Palpate for the distinct tuberosity of the 5th metatarsal bone (→ 3.6.2) on the lateral aspect of the midfoot. With the palpating finger, glide on from the tuberosity to the dorsum of the foot, into the groove between the 4th and 5th metatarsal bones. There, palpate in a distal direction. **G.B.-41** is located at the junction of the heads and the shafts of the two bones. **Or:** Ask the patient to abduct their toes so that the branch of the tendon of the extensor digitorum longus muscle extending to the little toe becomes more pronounced. Then palpate in the groove between the 4th and 5th metatarsal bones from distal to proximal to where the tendon crosses the groove. → **G.B.-42** is located medial (or anterior) to it, while **G.B.-41** is located in the depression lateral (or posterior) to the tendon.

Needling

Perpendicularly or obliquely approximately 0.3–0.8 cun

Actions/Indications

- Spreads Liver Qi, benefits the thorax and the lateral costal region, transforms Phlegm, dissipates nodules, benefits the breasts
- Clears the head, benefits the eyes

Special features

Shu-stream point, Wood point, opening point of the *dai mai*, *ben* point (Five Phases), exit point.

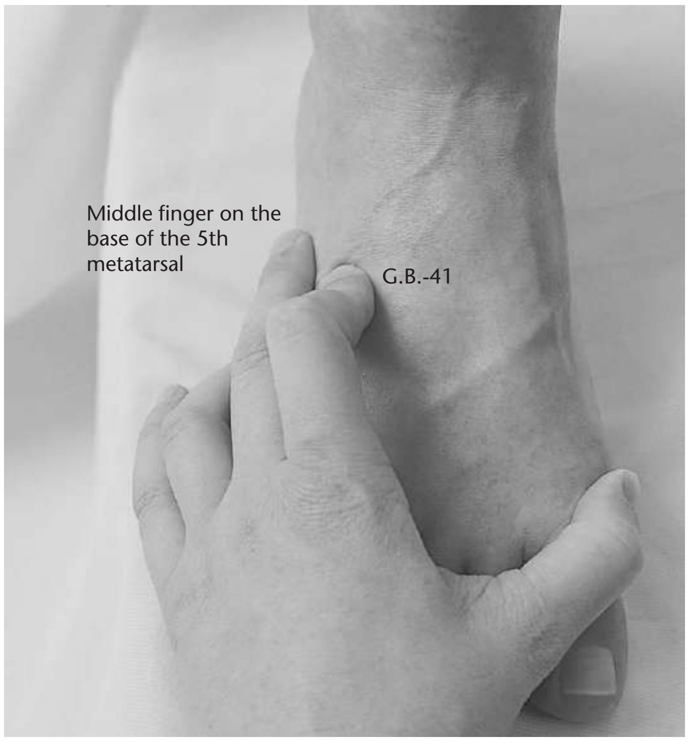
Middle finger on the base of the 5th metatarsal
G.B.-41

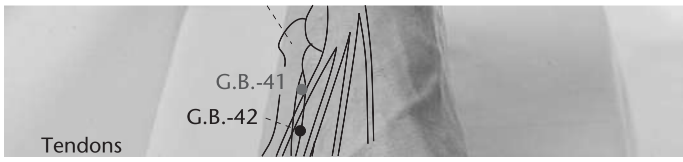
G.B.-41
G.B.-42
Tendons

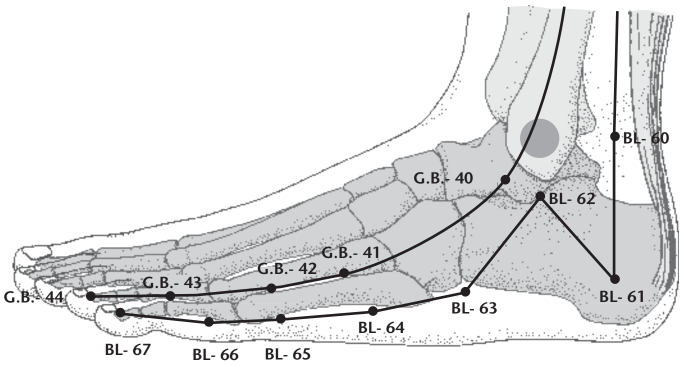
BL- 60
G.B.- 40
BL- 62
G.B.- 41
G.B.- 42
G.B.- 43
G.B.- 44
BL- 61
BL- 63
BL- 64
BL- 67
BL- 66
BL- 65

G.B.-42 *Earth Five Meetings* DIWUHUI

Location

Between the 4th and 5th metatarsal bones, proximal to the metatarsophalangeal joints and medial to the tendon of the extensor digitorum longus muscle.

How to find

From the space between the 4th and 5th toes, palpate towards the ankle. Locate **G.B.-42** in the groove between the two bones, proximal to the metatarsophalangeal joints (and proximal to the heads of the two metatarsal bones). **Or:** Ask the patient to abduct their toes so that the branch of the tendon of the extensor digitorum longus muscle extending to the little toe becomes more pronounced. Then palpate from the space between the toes along the groove between the 4th and 5th metatarsal bones from distal to proximal to where the tendon crosses the groove. **G.B.-42** is located medial (or anterior) to the tendon.

→ **G.B.-41** is located lateral (or posterior) to the tendon when continuing to palpate along the groove. → **BL-65** is located approximately on the same level on the lateral border of the foot, proximal to the head of the 5th metatarsal bone.

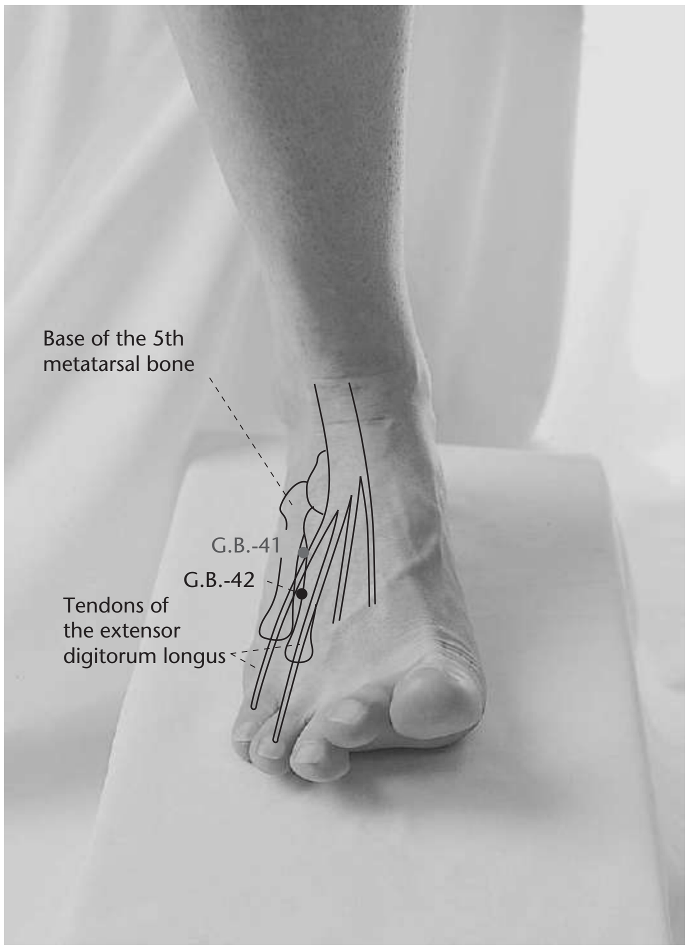
Base of the 5th metatarsal bone
G.B.-41
G.B.-42
Tendons of the extensor digitorum longus

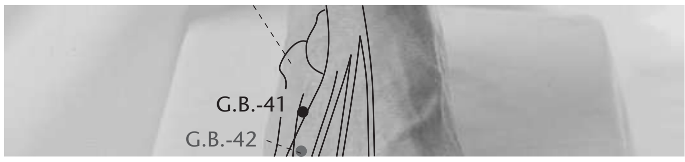

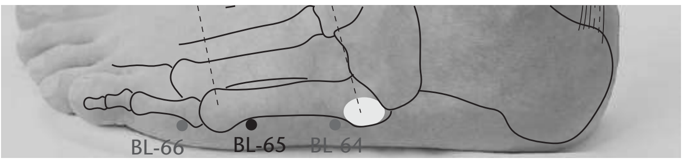

Needling

Vertically or obliquely 0.3–0.8 cun

Actions/Indications

- Moves Liver Qi
- Clears Heat from the Gall Bladder

Clamped Stream JIAXI/XIAXI G.B.-43

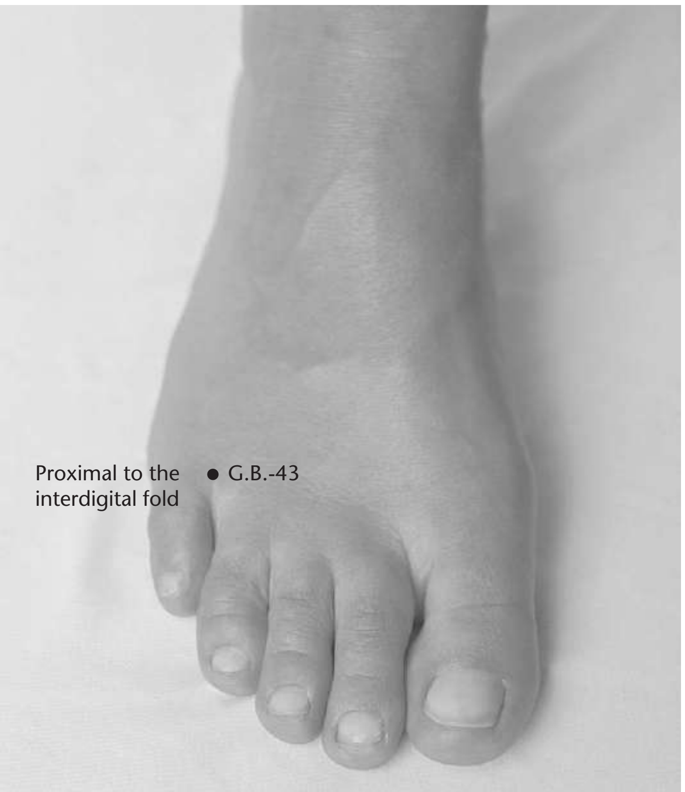
Proximal to the interdigital fold
G.B.-43

Location

Between the 4th and 5th toes, proximal to the margin of the interdigital web.

How to find

Locate the interdigital web between the 4th and 5th toes. Then locate **G.B.-43** slightly proximal to the margin of the web. **G.B.-43** as well as → **LIV-2** and → **ST-44** are part of → **Ex-UE-10** (*bafeng*: proximal to the margins of the webs between the toes). → **T.B.-2** is located in a comparable position on the hand, between the ring finger and little finger. It is also part of → **Ex-UE-9** (*baxie*).

Needling

Up to 0.5 cun vertically or 1 cun obliquely in a proximal direction.

Actions/Indications

- Clears Heat, Damp-Heat and Wind, especially from the opposite end of the channel
- Calms Liver Yang

Special features

Ying-spring point, Water point, tonification point.

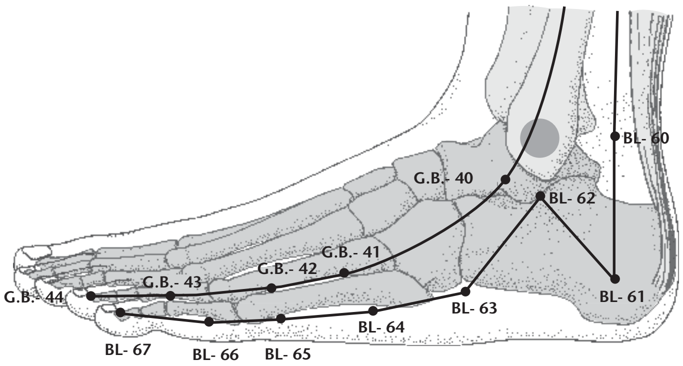
BL- 60
G.B.- 40
BL- 62
G.B.- 41
G.B.- 42
G.B.- 43
G.B.- 44
BL- 61
BL- 63
BL- 64
BL- 67
BL- 66
BL- 65

T.B.-2
Proximal to the interdigital fold

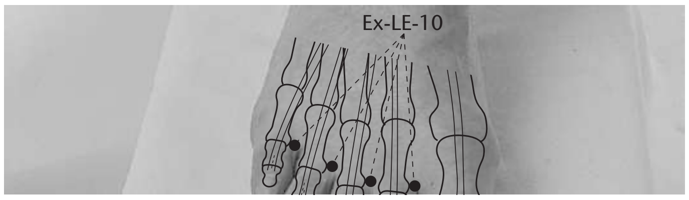

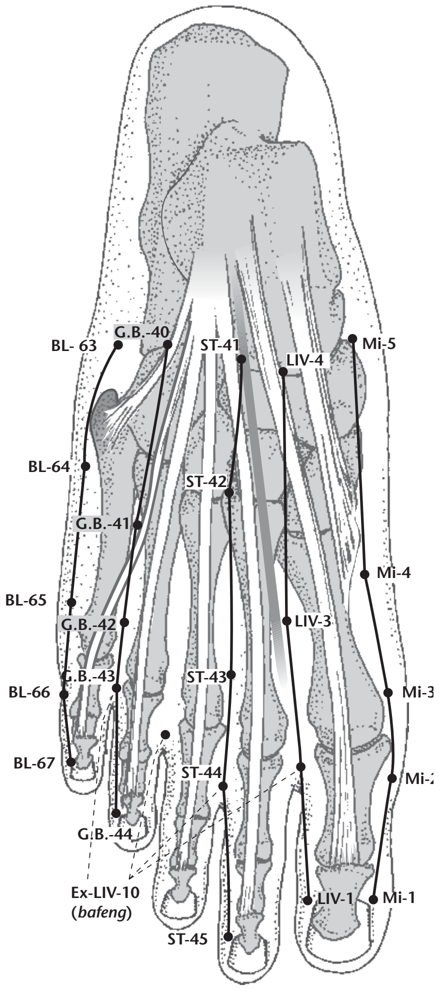
BL- 63
G.B.-40
ST-41
LIV-4
Mi-5
BL-64
ST-42
G.B.-41
Mi-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
Mi-3
BL-67
ST-44
Mi-
G.B.-44
Ex-LIV-10 (bafeng)
LIV-1
Mi-1
ST-45

G.B.-44 *Foot Portal Yin* ZUQIAOYIN

Location

On the 4th toe, 0.1 cun from the lateral corner of the nail.

How to find

G.B.-44 is located at the junction of two tangents along the proximal and lateral borders of the 4th toe.

→ **BL-67** is located at the lateral corner of the little toe.

Needling

0.1 cun vertically or 0.2 cun obliquely in a proximal direction or prick to bleed. Caution: Painful point.

Actions/Indications

- Clears Wind, Heat and Fire
- Harmonises the Liver and Gall Bladder
- Calms the *shen*

Special features

Jing-well point, Metal point

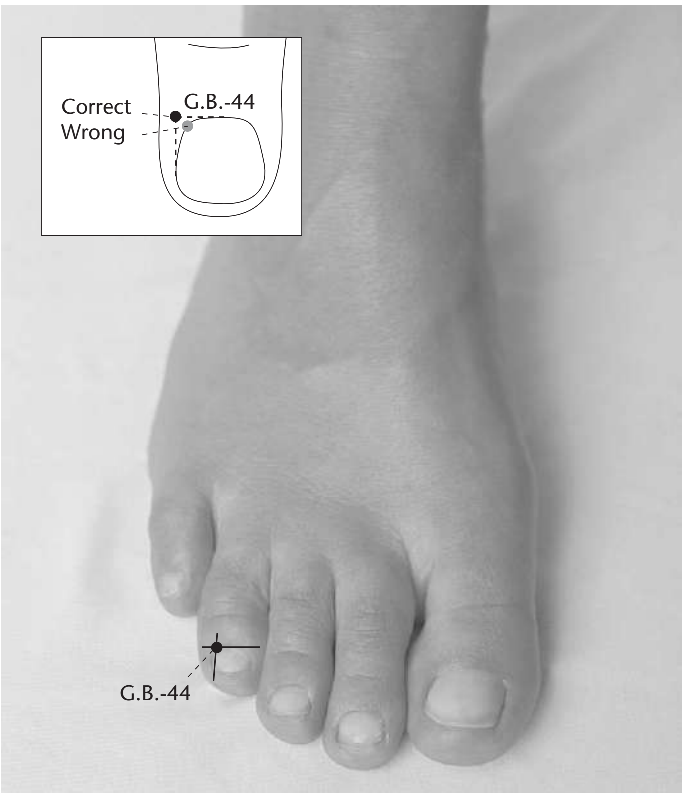
Correct
Wrong
G.B.-44
G.B.-44

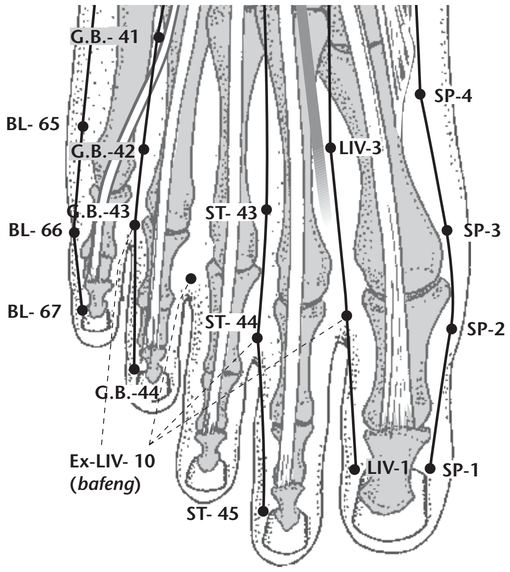
G.B.- 41
SP-4
BL- 65
G.B.-42
LIV-3
G.B.-43
ST- 43
BL- 66
SP-3
BL- 67
ST- 44
SP-2
G.B.-44
Ex-LIV- 10
(*bafeng*)
LIV-1
SP-1
ST- 45

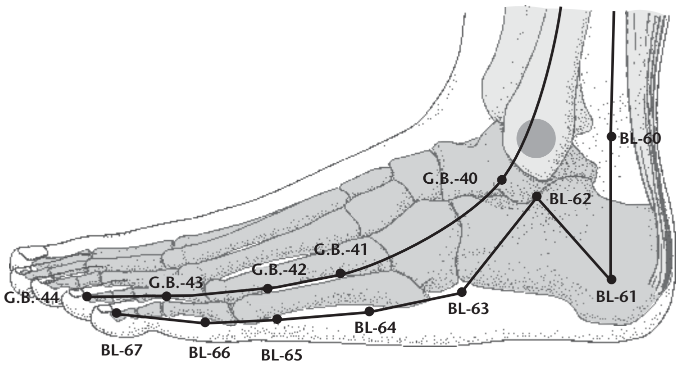
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

4.12 The Liver Channel System – Foot *jueyin* (*zu jue yin jing luo*)

4.12.1 The Liver Primary Channel (*zu jue yin jing*)

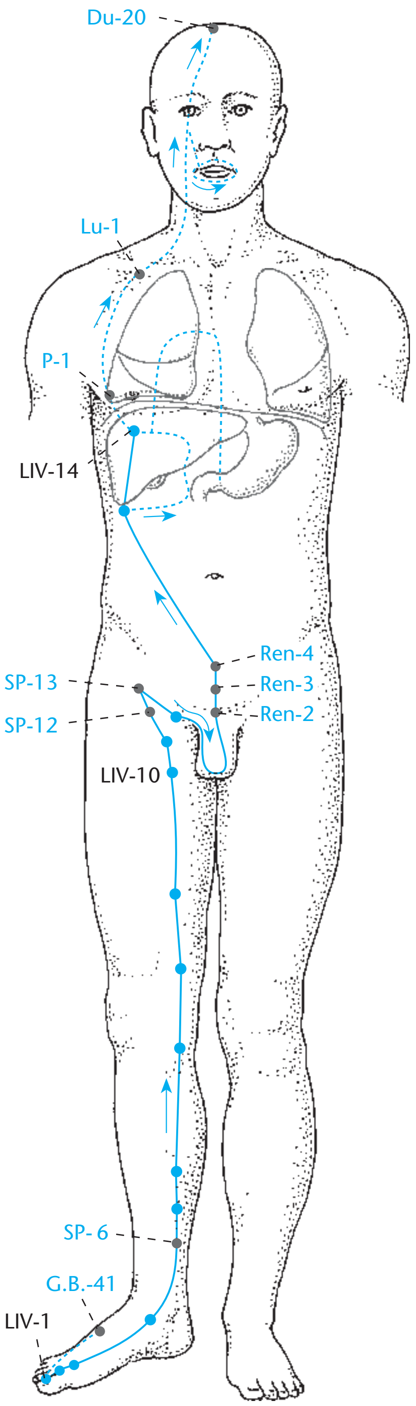
Du-20
Lu-1
P-1
LIV-14
Ren-4
Ren-3
Ren-2
SP-13
SP-12
LIV-10
SP- 6
G.B.-41
LIV-1

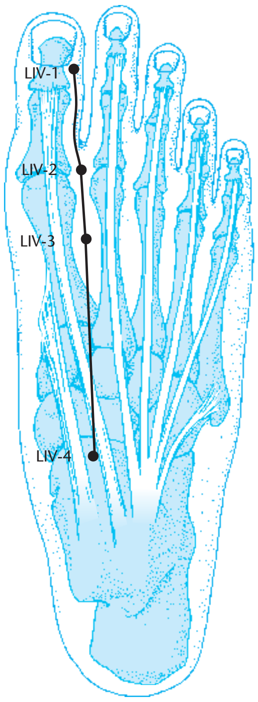
LIV-1
LIV-2
LIV-3
LIV-4

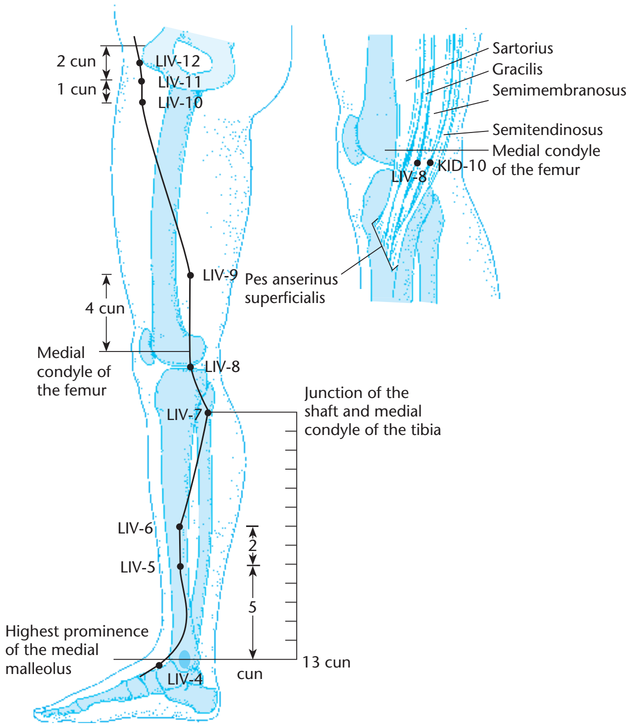
2 cun
1 cun
LIV-12
LIV-11
LIV-10
LIV-9
Pes anserinus superficialis
4 cun
Medial condyle of the femur
LIV-8
LIV-7
Junction of the shaft and medial condyle of the tibia
LIV-6
LIV-5
2
5
Highest prominence of the medial malleolus
LIV-4
cun
13 cun
Sartorius
Gracilis
Semimembranosus
Semitendinosus
Medial condyle of the femur
KID-10
LIV-8

Pathway

The Liver primary channel begins at **LIV-1** (*dadun*) on the lateral corner of the nail of the big toe. This point is reached by a small branch of the G.B. channel, which separates from the Gall Bladder primary channel on the dorsum of the foot at → **G.B.-41** (*zulinqi*) (foot Yin–Yang connection of the third great circuit).

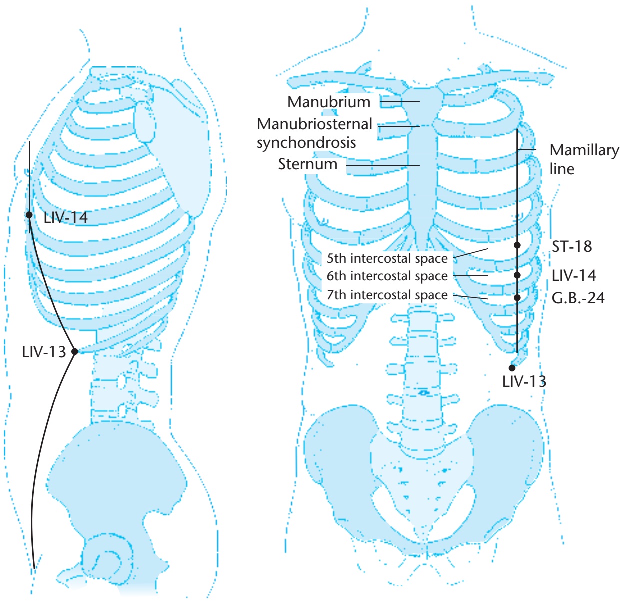

The **external pathway** runs proximally between the 1st and 2nd metatarsal bones

- ascends anterior to the medial malleolus
- crosses the SP channel at ➞ **SP-6** (*yinlingquan*) on the medial aspect of the lower leg
- ascends the medial aspect of the leg towards the knee anterior to the SP channel
- continues along the medial aspect of the thigh to the pubic region, where it passes ➞ **SP-12** (*chongmen*) and ➞ **SP-13** (*fushe*)
- circles the external genitalia
- ascends the lower abdomen, passing ➞ **Ren-2** (*qugu*), ➞ **Ren-3** (*zhongji*) and ➞ **Ren-4** (*guanyuan*)
- obliquely traverses the abdomen to **LIV-13** (*zhangmen*) at the free end of the 11th rib and to **LIV-14** (*qimen*), where the **external pathway** terminates.

At **LIV-13**, the channel enters the abdomen, marking the beginning of the **internal pathway**. It

- circles the Stomach (*wei*)
- connects with its pertaining *zang*-Organ, the Liver (*gan*) and its paired *fu*-Organ, the Gall Bladder (*dan*)
- penetrates the diaphragm
- spreads in the lateral hypochondrium and thoracic region.

The channel then ascends along the posterior aspect of the trachea to the throat and nasopharynx and connects with the eye system and the brain. It crosses the forehead and ascends to the vertex, where it connects with the extraordinary vessel *du mai* at ➞ **Du-20** (*baihui*).

An **internal branch** descends from the maxillary sinus to the cheek and circles the **inner surface** of the lips.

An **internal branch** emerges from the Liver, penetrates the diaphragm, disperses in the Lung (*fei*) and meets the Lung primary channel (deep Yin–Yin connection), closing the first circuit of the Nutritive Qi (*ying qi*) (➞ 1.1.4). This branch further connects with the Pericardium primary channel beneath ➞ **P-1** (*tianchi*) (hand–foot pairing of the third great circuit: Yin axes, *jue yin*).

Clinical importance (➞ 1.2)

Exterior (*biao*) signs and symptoms: Headaches, dizziness, blurred vision, tinnitus, fever, spastic extremities

Interior (*li*) or *zangfu*-Organ signs and symptoms: Distension, fullness and pain in the costal region, tightness and fullness in the chest, abdominal pain, vomiting, jaundice, diarrhoea, *shan* disorder, enuresis, urinary retention, yellow urine

Connections of the Liver primary channel (➞ 1.2)

Connections with other channels

Gall Bladder primary channel (*zu shao yang jing*)

Connection: Foot Yin–Yang pairing of the third great circuit
Location: G.B.-41 ➞ LIV-1 (on the foot)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Pericardium primary channel (*shou jue yin jing*)

Connection: Paired according to the six-channel theory (hand–foot pairing): *jue yin* (Yin axes of the third great circuit)
Location: LIV ➞ P. An internal branch originating in the Liver penetrates the diaphragm and connects with the Pericardium primary channel beneath ➞ P-1.
Circulation: Non-circadian (not according to the Organ clock)
Importance: Above–below relationship

Lung primary channel (*shou tai yin jing*)

Connection: Deep Yin–Yin connection
Location: LIV ➞ LU (in the thorax). An internal branch originating in the Liver penetrates the diaphragm and disperses in the Lung to connect with the Lung primary channel.
Circulation: Circadian (according to the Organ clock)
Importance: The Lung primary channel receives part of its Nutritive Qi (*ying qi*) from the Liver primary channel (first circulation of the *ying qi* ➞ 1.1.4).

Connections with zangfu-Organ systems

Stomach (*wei*), **Liver (*gan*)**, **Gall Bladder (*dan*)**, Lung (*fei*)

4.12.2 The Liver Divergent Channel (*zu jue yin jing bie*)

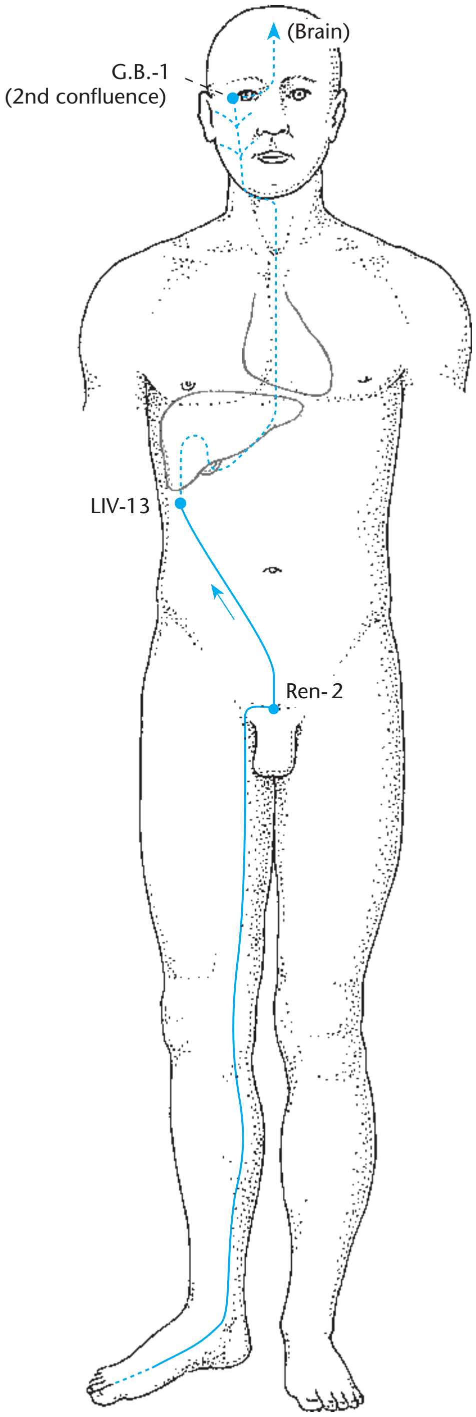

Pathway

The Liver divergent channel separates from the Liver primary channel on the dorsum of the foot

➡ ascends the medial aspect of the leg to the inguinal region

➡ meets the Gall Bladder divergent channel near → **Ren-2** (*qugu*)

➡ traverses the abdomen and enters internally at **LIV-13** (*zhangmen*)

➡ continues inside the thoracic cavity, spreads in the Liver (*gan*), connects with the Gall Bladder (*dan*) and runs to the Heart (*xin*)

➡ ascends to the neck

➡ emerges at the mandibular angle and disperses over the face

➡ connects with the Gall Bladder primary channel and the Gall Bladder divergent channel at → **G.B.-1** (*tongziliao*) on the outer canthus of the eye to form one of the six *he*-confluences (here: G.B./LIV as 2nd confluence → 1.3)

➡ continues to the eye system and ends in the Brain.

Clinical importance

- Strengthens the relationship between the Liver and the Gall Bladder (*zangfu*-Organ systems). Points on the Liver primary channel can therefore be used to treat disorders of the Gall Bladder and vice versa.
- Based on the pathways of the Gall Bladder and Liver divergent channels, points on both channels can be used for disorders of the hips and the lower extremities as well as for disorders of the eye system.

4.12.3 The Liver Sinew Channel (*zu jue yin jing jin*)

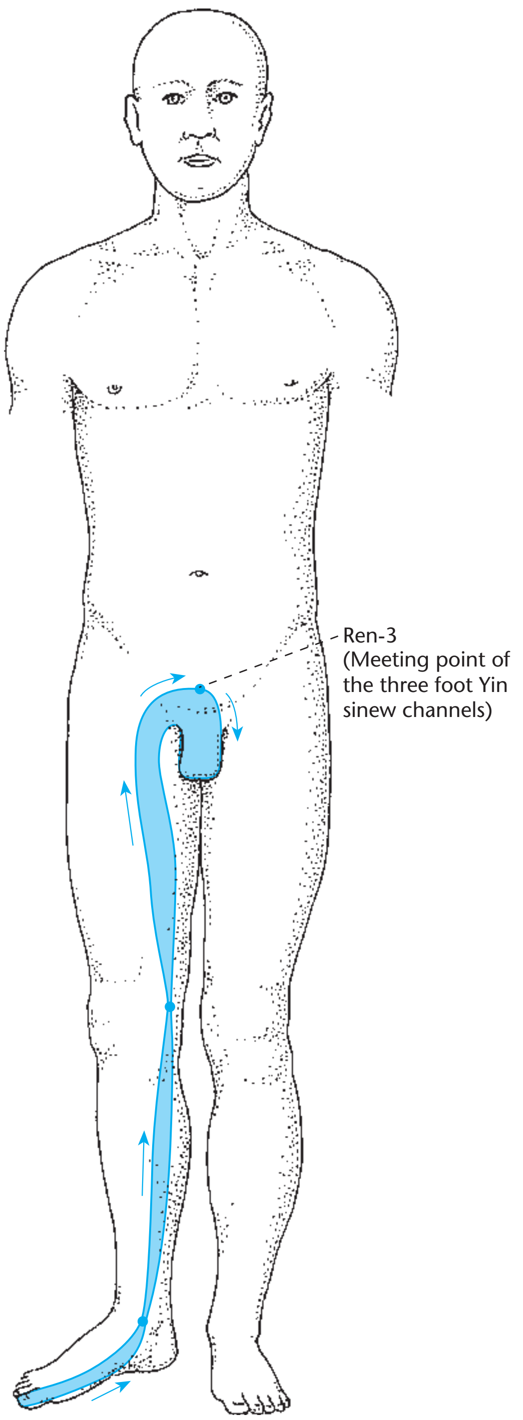

Pathway

The Liver sinew channel begins on the dorsum of the big toe

- ➡ follows the dorsolateral aspect of the big toe to the anterior aspect of the medial malleolus, where it binds (*jie*)
- ➡ ascends the medial aspect of the tibia and binds (*jie*) at the medial aspect of the knee
- ➡ further ascends the medial aspect of the thigh between the Kidney and Spleen sinew channels
- ➡ reaches the inguinal region, meeting the other foot Yin sinew channels at → **Ren-3** (*zhongji*) and continues to the genitalia.

Clinical importance

Pathology: Stiffness, tension and distending sensations along the big toe. Pain around the medial malleolus and medial aspect of the knee. Pain and muscular tension along the medial aspect of the thigh. Disorders of the genital region.

Indication: Pain, tension, muscle contractions and distending sensations along the pathway of the channel. Disorders of the genital region.

4.12.4 The Liver *luo*-Connecting Vessel System (*zu jue yin luo mai*)

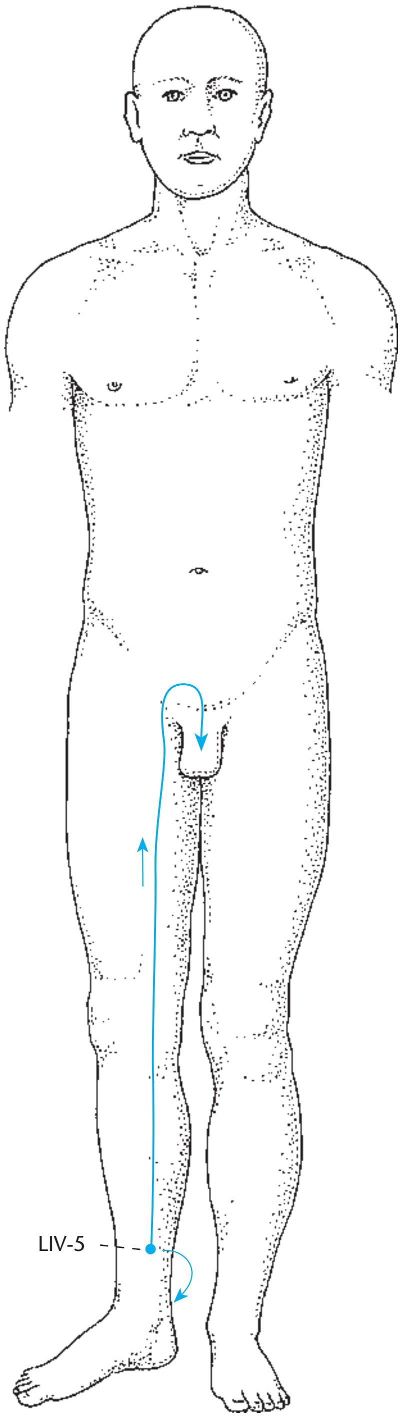

Pathway

The Liver *luo*-connecting vessel system separates from the Liver primary channel at its *luo* point **LIV-5** (*ligou*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Gall Bladder primary channel; according to some schools of thought (for example, Ngyen Van Nghi → Appendix), they travel as a **transverse** Liver *luo*-connecting vessel to the *yuan*-source point → **G.B.-40** (*qiuxu*).
- **A longitudinal division** ascends the medial aspect of the leg to the genital region, where it branches out.

Clinical importance (→ 8.1.2)

Pathology

Counterflow Qi: Swellings and pain of the testicles and scrotum, hernia

Excess (*shi*): Priapism

Deficiency (*xu*): Itching of the external genitalia

4.12.5 Cutaneous Region (*jue yin pi bu*)

See description and figures → 1.6.

4.12.6 Points of the Liver Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): LIV-3** (*taichong*) 🟦🟦
- ***Luo*-connecting point (→ 8.1.2.): LIV-5** (*ligou*) 🟦⬛
- ***Xi*-cleft point (→ 8.1.3): LIV-6** (*zhongdu*) 🟦⬛
- **Associated Back-*shu* point (→ 8.1.4): BL-18** (*ganshu*) 🟦⬛
- **Associated Front-*mu* point (→ 8.1.5): LIV-14** (*qimen*) 🟦⬛
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood): **LIV-1** (*dadun*) 🟦⬛, *ben* point (Five Phases)
  ying-spring point (Fire): **LIV-2** (*xingjian*) 🟦🟦
  shu-stream point (Earth): **LIV-3** (*taichong*) 🟦🟦
  jing-river point (Metal): **LIV-4** (*zhongfeng*)
  he-sea point (Water): **LIV-8** (*ququan*) 🟦⬛, tonification point
- ***Hui*-meeting point (→ 8.1.7) of the *zang*-Organs: LIV-13** (*zhangmen*) 🟦🟦
- **Opening point (→ 8.1.8): –**
- **Lower *he*-sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - with the G.B. channel and the *dai mai**: **LIV-13** (*zhangmen*) 🟦🟦

---

*Mentioned by only some authors

  - with the *yin wei mai* and the ST channel: **LIV-14** (*qimen*) 🟦⬛
  - of other channels with the LIV channel: **SP-6, SP-12, SP-13, Ren-2, Ren-3, Ren-4, P-1, (Du-20*)**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): –**
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): LIV-3** (*taichong*) 🟦🟦
- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points:**
  - Front-*mu* point of the Spleen: **LIV-13** (*zhangmen*) 🟦🟦

Points according to region

- **Local points** (→ 8.2.1): Liver – **LIV-13** (*zhangmen*) 🟦🟦; knee – **LIV-7** (*xiguan*), **LIV-8** (*ququan*) 🟦⬛; foot – **LIV-4** (*zhongfeng*)
- **Adjacent points (→ 8.2.1):** Stomach/Spleen – **LIV-13** (*zhangmen*) 🟦🟦; hypochondrium – **LIV-13** (*zhangmen*) 🟦🟦, **LIV-14** (*qimen*)
- **Distal points** (→ 8.2.1): vertex – **LIV-3** (*taichong*) 🟦🟦; eyes – **LIV-2** (*xingjian*) 🟦🟦, **LIV-3** (*taichong*) 🟦🟦; Liver – **LIV-3** (*taichong*) 🟦🟦; Gall Bladder – **LIV-3** (*taichong*) 🟦🟦; urogenital region – **LIV-3** (*taichong*) 🟦🟦; genital region – **LIV-5** (*ligou*) 🟦⬛; lower abdomen – **LIV-8** (*ququan*) 🟦⬛

Specific points according to the channel pathway (in numeric order)

- **LIV-1** (*dadun*) 🟦⬛: *jing*-well point (Wood), *ben* point (Five Phases) (→ 8.1.6)
- **LIV-2** (*xingjian*) 🟦🟦: *ying*-spring point (Fire) (→ 8.1.6), sedation point; distal point for the eyes (→ 8.2.1)
- **LIV-3** (*taichong*) 🟦🟦: *yuan*-source point (→ 8.1.1); *shu*-stream point (Earth) (→ 8.1.6); Ma Dan Yang Heavenly Star point (→ 8.1.4); distal point for the vertex, eyes, Liver, Gall Bladder and urogenital region (→ 8.2.1)
- **LIV-4** (*zhongfeng*): *jing*-river point (Metal) (→ 8.1.6); local point for the foot (→ 8.2.1)
- **LIV-5** (*ligou*) 🟦⬛: *luo*-connecting point (→ 8.1.2); distal point for the genital region (→ 8.2.1)
- **LIV-6** (*zhongdu*) 🟦⬛: *xi*-cleft point (→ 8.1.3)
- **LIV-7** (*xiguan*): local point for the knee (→ 8.2.1)
- **LIV-8** (*ququan*) 🟦⬛: *he*-sea point (Water) (→ 8.1.6); tonification point; distal point for the abdomen (→ 8.2.1); local point for the knee (→ 8.2.1)
- **LIV-13** (*zhangmen*) 🟦🟦: Front-*mu* point of the Spleen (→ 8.1.5); *hui*-meeting point of the *zang*-Organs (→ 8.1.7); *jiaohui*-meeting point with the G.B. channel and the (*dai mai**) (→ 8.1.10); adjacent point for the Spleen/Stomach and the hypochondrium (→ 8.2.1)
- **LIV-14** (*qimen*) 🟦⬛: associated Front-*mu* point(→ 8.1.5); *jiaohui*-meeting point with the *yin wei mai* and the SP channel (→ 8.1.10); regional point for the hypochondrium (→ 8.2.1)

---

* Mentioned by only some authors

Big Mound DADUN LIV-1

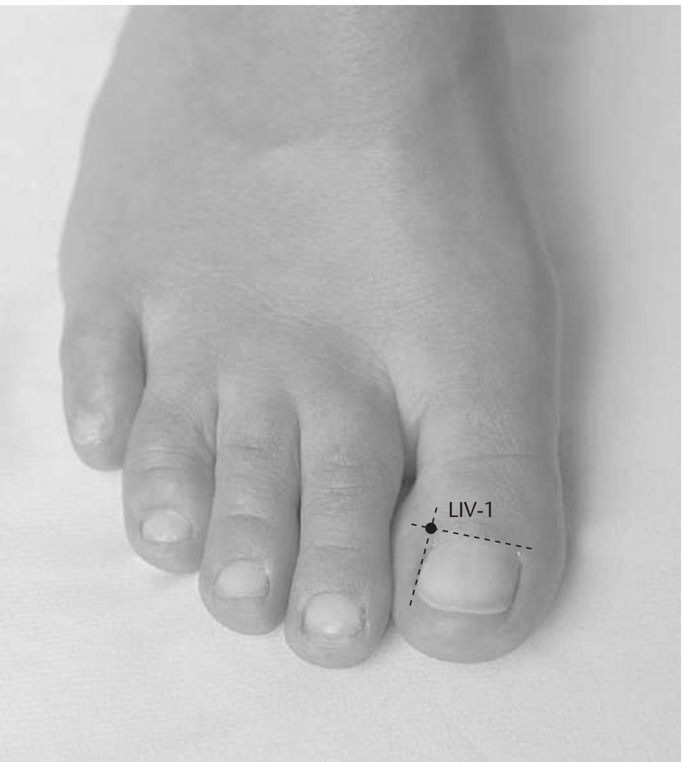

Location

On the lateral aspect of the big toe, 0.1 cun proximal and lateral to the corner of the nail.

How to find

LIV-1 is located at the junction of two tangents along the proximal and lateral borders of the big toe, 0.1 cun from the actual margin of the nail. → **SP-1** is located at the medial corner of the big toe.

Needling

Vertically or obliquely 0.1–0.2 cun. Avoid needling into the perionychium. For excess conditions, prick to bleed. Caution: Painful point.

Actions/Indications

- Regulates the Lower Burner, the genitals and Liver Qi
- Eliminates Dampness
- Opens the sensory orifices

Special features

Jing-well point, Wood point, *ben* point (Five Phases), entry point.

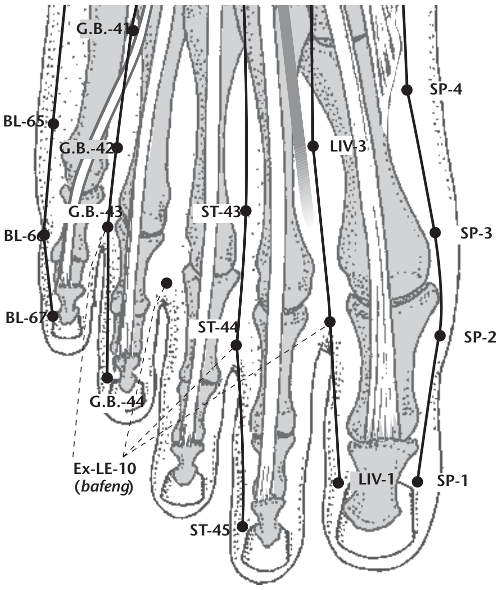
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

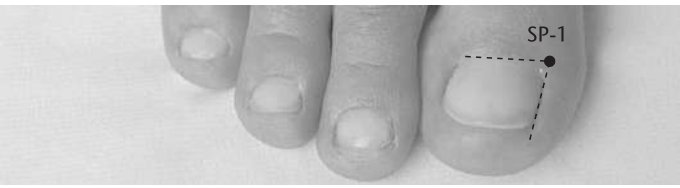

LIV-2 *Moving Between* XINGJIAN

Location

Between the 1st and 2nd toes, proximal to the margin of the interdigital web.

How to find

Locate **LIV-2** slightly proximal to the margin of the interdigital web between the 1st and 2nd toes.
Located in comparable positions are → **ST-44** (between the 2nd/3rd toes) and → **G.B.-43** (between the 4th/5th toes). → **LIV-2,** → **ST-44** and → **G.B.-43** are all part of the extra point → **Ex-LE-10** (*bafeng*). Its counterpart on the hand is the extra point → **Ex-UE-9** (*baxie*).

Needling

0.3–0.5 cun obliquely towards the heel or vertically.

Actions/Indications

- Clears Liver Fire, spreads Liver Qi, calms (internal) Liver Wind, clears Heat and Blood Heat, stops bleeding, benefits the Lower Burner

Special features

Ying-spring point, Fire point, sedation point. Major point for excess syndromes of the Liver (especially Liver Fire).

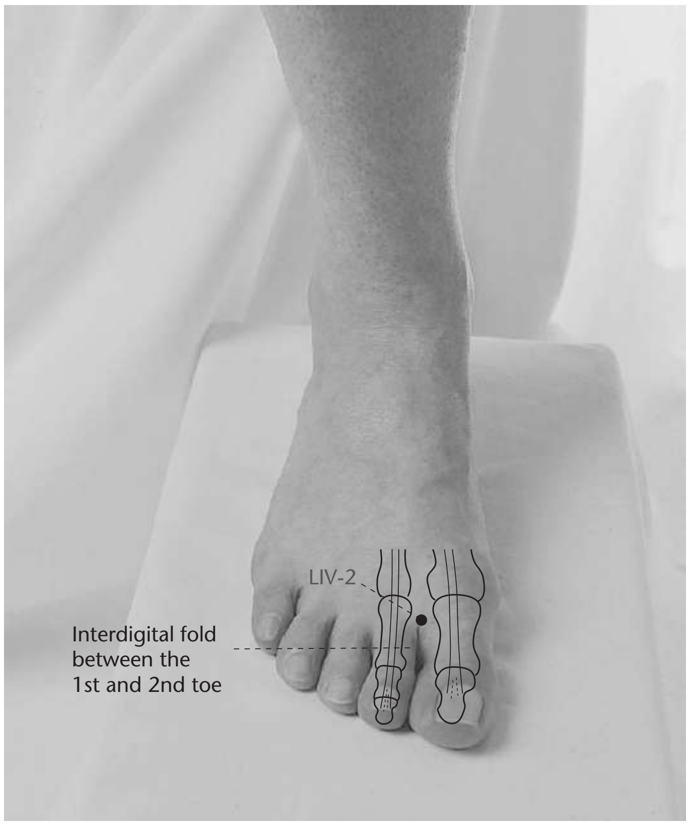

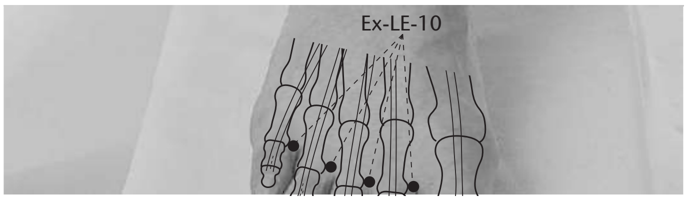

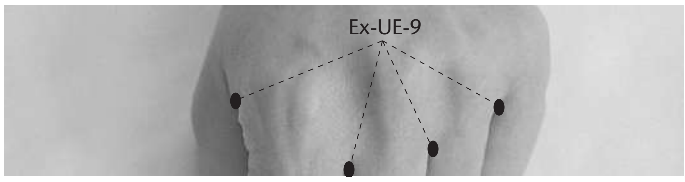

Great Rushing TAICHONG LIV-3

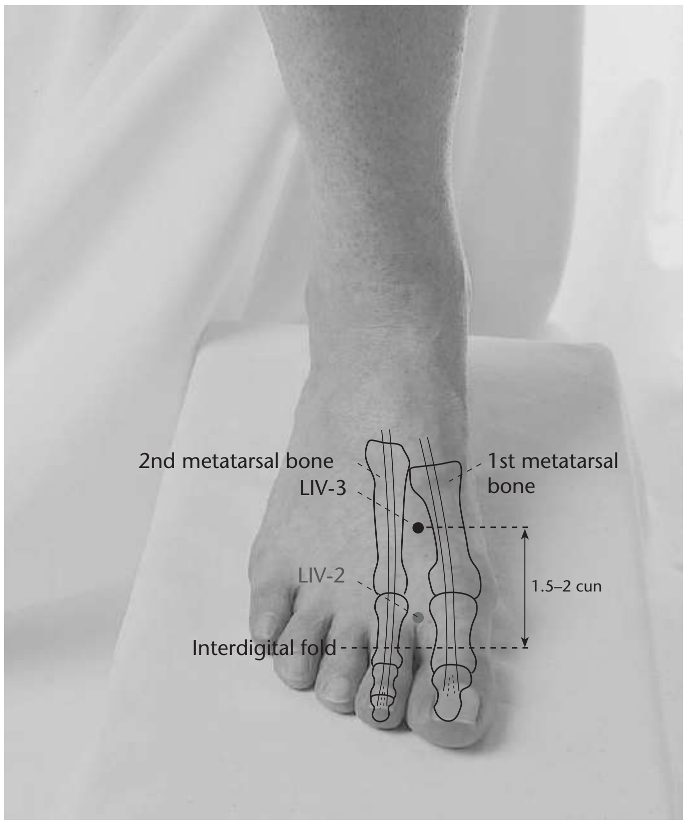
2nd metatarsal bone
LIV-3
1st metatarsal bone
LIV-2
1.5–2 cun
Interdigital fold

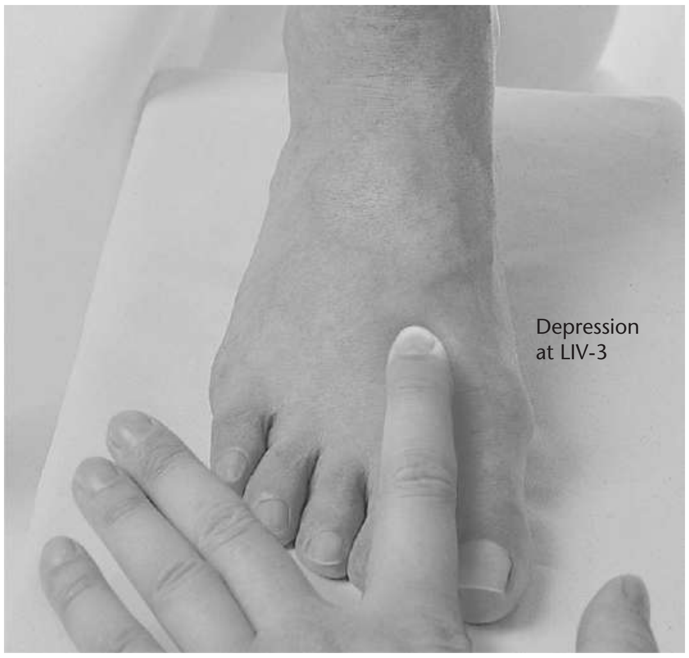
Depression at LIV-3

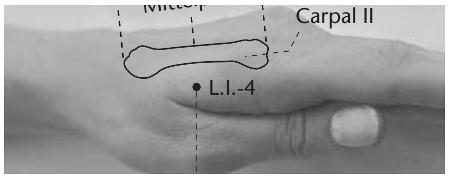
Carpal II
L.I.-4

Location

On the dorsum of the foot, between the 1st and 2nd metatarsal bones, in the depression proximal to the metatarsophalangeal joints and the proximal angle between the two bones.

How to find

From the web between the 1st and 2nd toes, palpate proximally past the metatarsophalangeal joints and along the groove between the 1st and 2nd metatarsal bones, until you reach the widest and deepest part of the groove. There, locate **LIV-3**, which is often sensitive to pressure. To double-check: When continuing to palpate along the groove, it will become more narrow and shallow again.

Located in a similar position on the hand is → **L.I.-4** (between the 1st and 2nd metacarpal bones).

Needling

Vertically or slightly obliquely 0.5–1 cun. Caution during pregnancy.

Actions/Indications

- Spreads Liver Qi, clears the head, benefits the eyes, nourishes Liver Blood and Liver Yin, calms spasms, alleviates pain, regulates menstruation and the Lower Burner
- Eliminates (internal) Wind, pacifies Liver Yang

Special features

Yuan-source point, *shu*-stream point, Earth point, Ma Dan Yang Heavenly Star point (later added by Xu Feng). Major point for spreading Liver Qi.

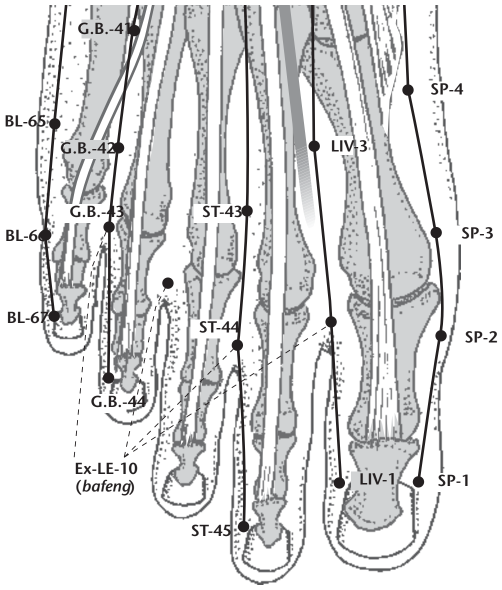
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

LIV-4 Middle Seal ZHONGFENG

Location

1 cun anterior to the prominence of the medial malleolus, medial to the tendon of the tibialis anterior muscle.

How to find

Even with slight dorsiflexion of the ankle, the tendon of the tibialis anterior muscle will become more pronounced on the medial aspect of the ankle. **LIV-4** is located between the tendon and the prominence of the medial malleolus, in a depression over the palpable joint space.

→ **ST-41** is also located over the joint space (on the midpoint of its anterior aspect); → **SP-5** lies only slightly distal to → **LIV-4** at the junction of a vertical line along the anterior border and a horizontal line along the lower border of the medial malleolus.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Opens the channel, promotes the flow of Liver Qi, regulates the Lower Burner

Special features

Jing-river point, Metal point.

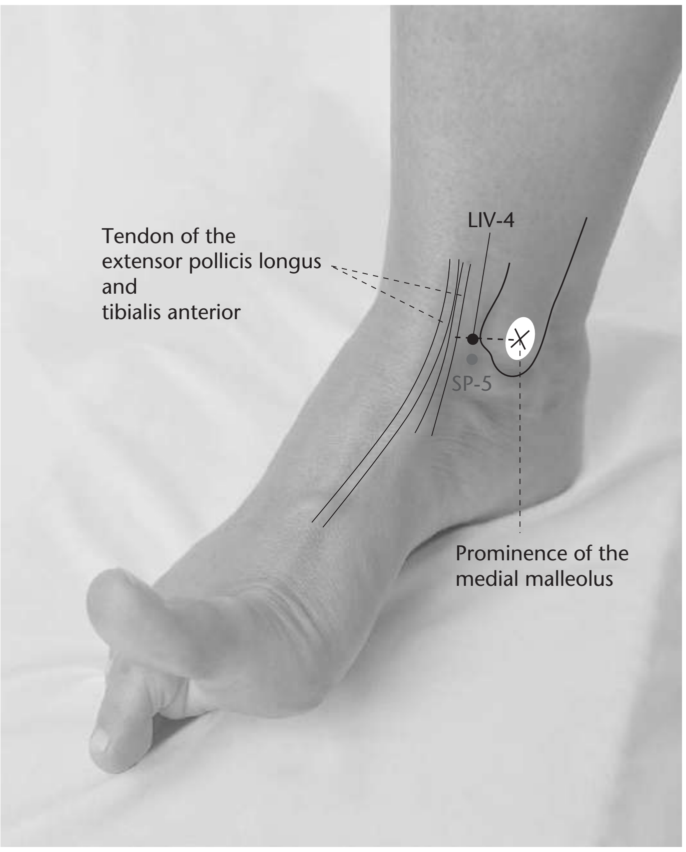
Tendon of the
extensor pollicis longus
and
tibialis anterior
LIV-4
SP-5
Prominence of the
medial malleolus

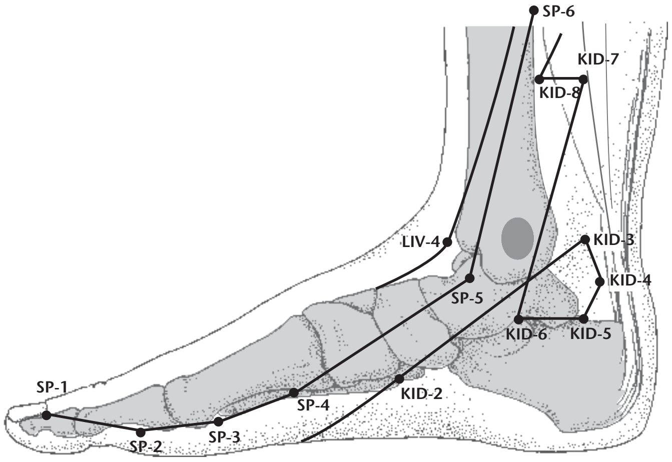
SP-6
KID-7
KID-8
LIV-4
KID-3
SP-5
KID-4
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

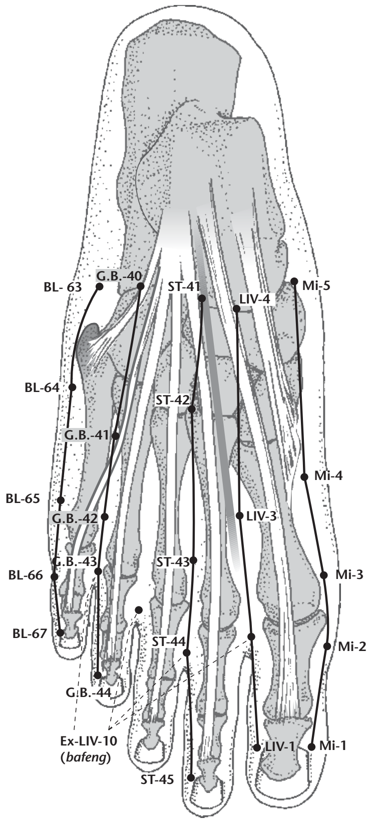
BL-63
G.B.-40
ST-41
LIV-4
Mi-5
BL-64
ST-42
G.B.-41
Mi-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
Mi-3
BL-66
BL-67
ST-44
Mi-2
G.B.-44
Ex-LIV-10
(bafeng)
LIV-1
Mi-1
ST-45

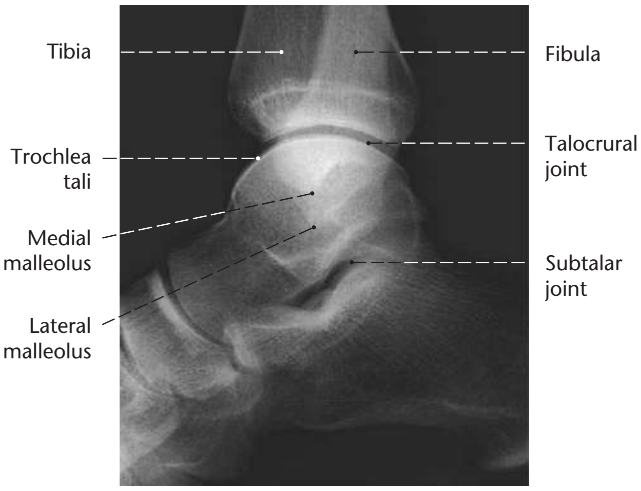
Tibia
Fibula
Trochlea tali
Talocrural joint
Medial malleolus
Subtalar joint
Lateral malleolus

Woodworm Canal LIGOU LIV-5

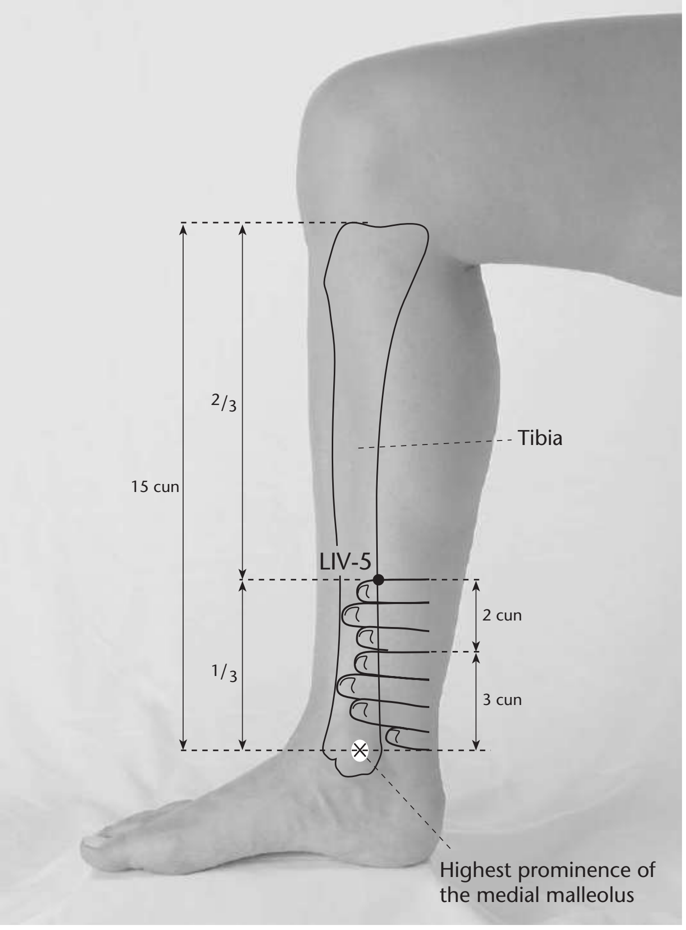

Location

On the medial aspect of the lower leg, 5 cun proximal to the highest prominence of the medial malleolus, just posterior to the medial crest of the tibia, between the crest of the tibia and the gastrocnemius muscle.

How to find

First, locate the highest prominence of the medial malleolus (→ 3.6.2). From there, palpate 5 cun in a proximal direction and there locate **LIV-5** in a depression directly posterior to the medial crest of the tibia. **Or quick method:** Divide the distance between the medial end of the popliteal crease and the highest prominence of the medial malleolus (=approximately 15 cun, → 2.2) into thirds and locate **LIV-5** a third of the distance from the medial malleolus. Note: Some texts locate **LIV-5 on** the tibia, others on the **posterior** border of the tibia; sensitivity to pressure should be the determining factor. → **KID-9** is located on the same level but directly superior to → **KID-3** (between the medial malleolus and the Achilles tendon).

Needling

0.5–1 cun vertically or obliquely in a posterior direction towards the fibula or transversely (subcutaneously) along the tibia towards the abdomen.

Actions/Indications

- Regulates Liver Qi, benefits the genitals, clears Damp-Heat from the Lower Burner, regulates menstruation
- Eliminates 'plum-pit' Qi
- Moves Qi locally

Special features

Luo-connecting point. Important distal point for the urogenital region (especially with Liver Qi stagnation and Damp-Heat).

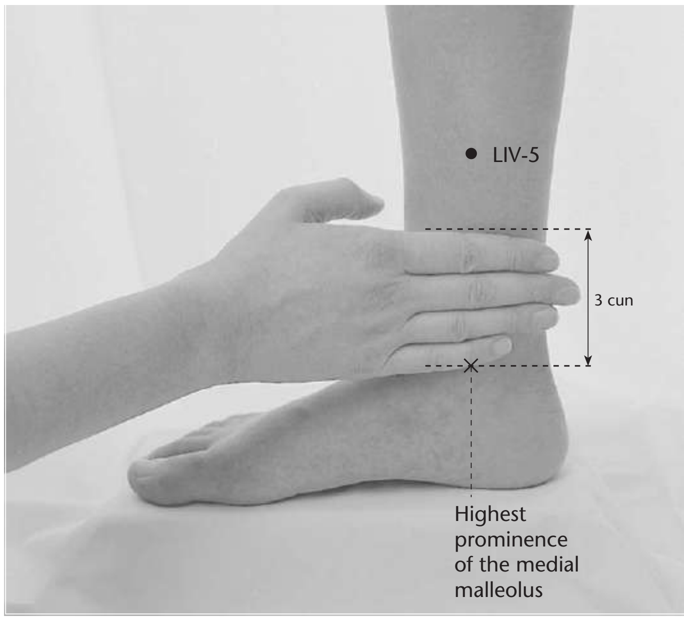

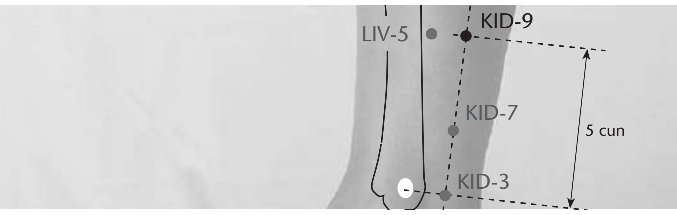

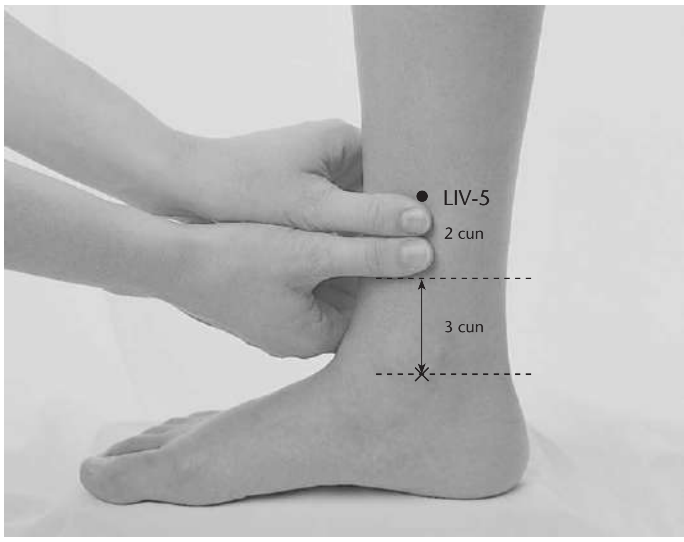

LIV-6 *Central Capital* ZHONGDU

Location

7 cun proximal to the highest prominence of the medial malleolus, just posterior to the medial crest of the tibia.

How to find

First, locate the highest prominence of the medial malleolus (→ 3.6.2). From there, palpate 7 cun (= the width of 2 hands and 1 thumb) in a proximal direction and there locate **LIV-6** in a depression directly posterior to the medial crest of the tibia. **Or quick method:** Spreading hands technique (→ 2.3.3): Place the little fingers on the medial end of the popliteal crease and the highest prominence of the medial malleolus respectively and join the thumbs on the posterior border of the tibia at the midpoint of this distance (= approximately 15 cun, 2.2). From there, measure 0.5 cun in a distal direction and here locate **LIV-6** posterior to the medial crest of the tibia. Note: Some texts locate **LIV-6 on** the tibia, others on the **posterior** border of the tibia; sensitivity to pressure should be the determining factor.

Needling

0.5–1 cun vertically or obliquely in a posterior direction towards the fibula or transversely (subcutaneously) in a proximal direction along the tibia.

Actions/Indications

- Spreads Liver Qi, regulates the Lower Burner, regulates Blood, eliminates Dampness

Special features

Xi-cleft point

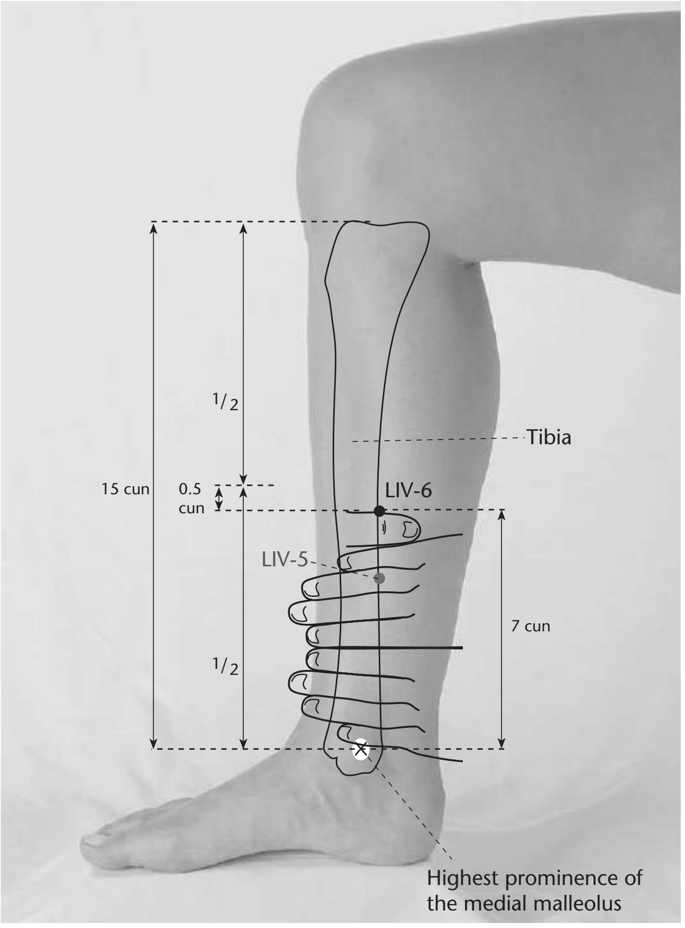

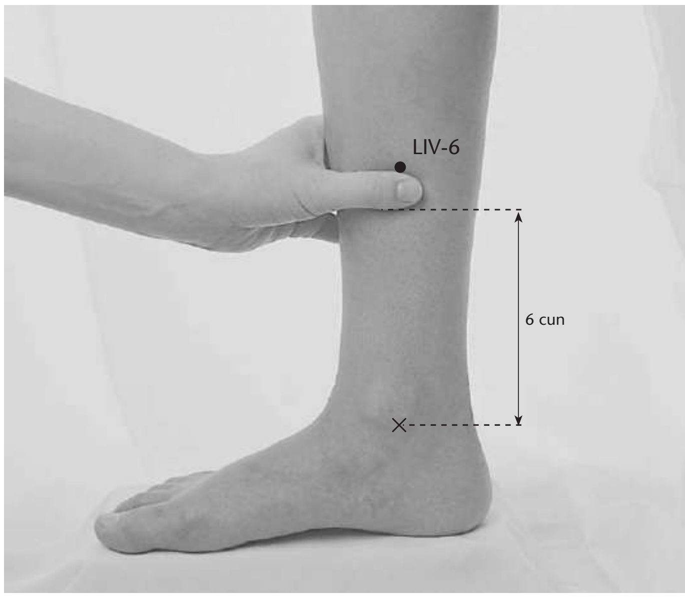

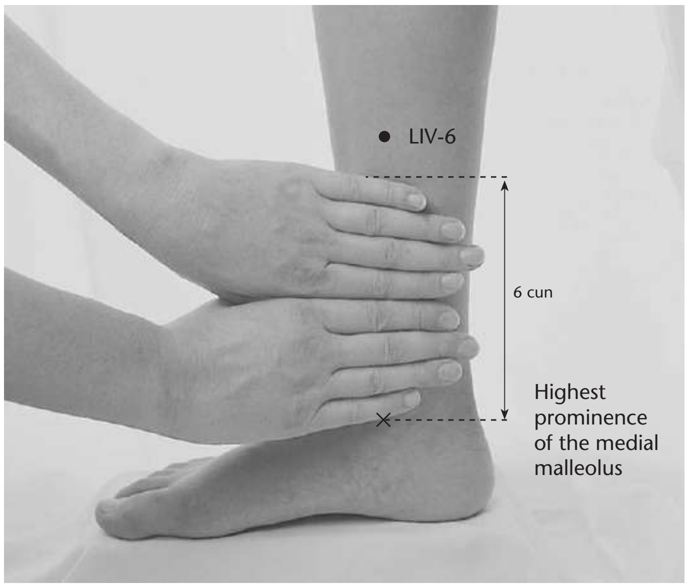

Knee Joint XIGUAN **LIV-7**

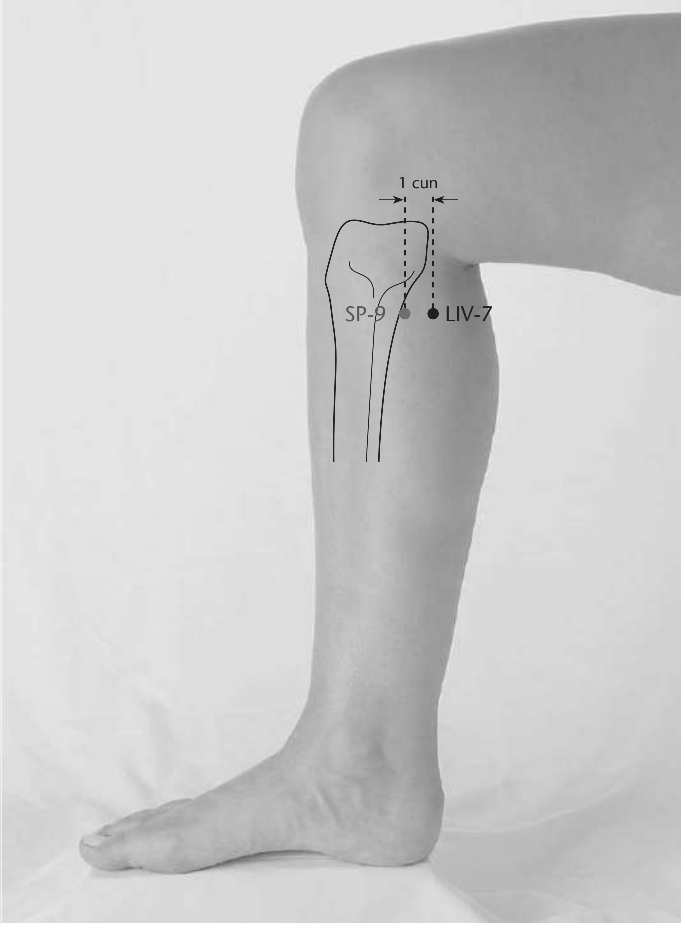
1 cun
SP-9
LIV-7

Location

At the junction of the shaft and the medial condyle of the tibia, 1 cun posterior to → **SP-9**.

How to find

The junction of the shaft and the medial condyle of the tibia can be easily palpated. There, locate → **SP-9** and **LIV-7** 1 cun posterior to it. Both points tend to be sensitive to pressure if indicated.

Needling

Vertically 1–2 cun

Actions/Indications

- Opens the channel
- Expels Wind and Dampness

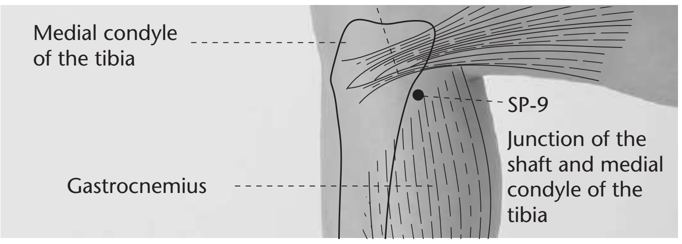
Medial condyle of the tibia
Gastrocnemius
SP-9
Junction of the shaft and medial condyle of the tibia

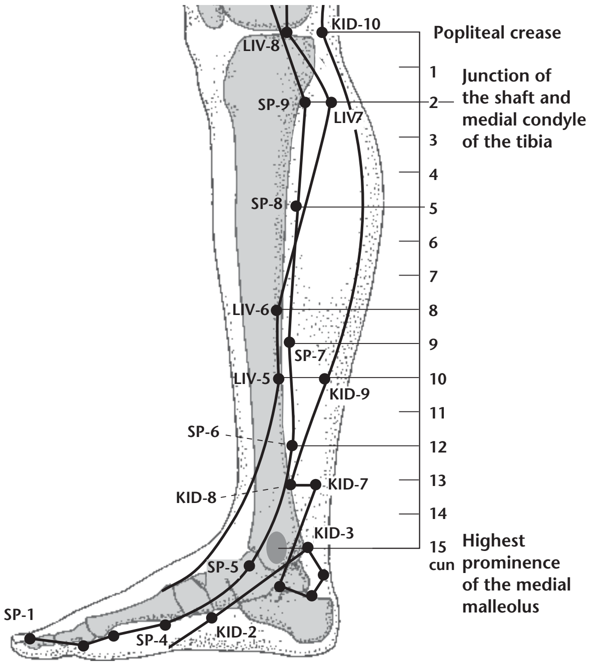
KID-10
Popliteal crease
LIV-8
1
Junction of the shaft and medial condyle of the tibia
SP-9
2
LIV7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
LIV-5
10
KID-9
11
SP-6
12
KID-8
KID-7
13
14
KID-3
15 cun
Highest prominence of the medial malleolus
SP-5
SP-1
SP-4
KID-2

LIV-8 *Spring at the Crook* QUQUAN

Location

With the knee flexed, **LIV-8** is located directly proximal to the medial end of the popliteal crease, in a depression anterior to the tendons of the semitendinosus and semimembranosus muscles.

How to find

LIV-8 is best located with the patient's knee flexed and slight external rotation of the hip. First, palpate the prominent, ropey tendon of the semitendinosus muscle, which becomes more pronounced with the knee flexed. The tendon of the semimembranosus muscle lies deep to the semitendinosus and is often more difficult to isolate. With the palpating finger, glide from the tendon of the semitendinosus towards the patella. Approximately after 1 cun, you can feel a depression between the muscle bellies. **LIV-8** is located anterior to both tendons. Some authors locate **LIV-8** 1 cun superior and anterior to → **KID-10**. Sensitivity to pressure should determine the location.
→ **KID-10** is located approximately 1 cun in the direction of the popliteal crease and, from a medial perspective, posterior to the tendon of the semimembranosus and anterior to the tendon of the semitendinosus.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Clears and eliminates Dampness and Damp-Heat from the Lower Burner (major action), benefits the genitals and the Uterus
- Nourishes Blood and Yin
- Moves Qi locally

Special features

he-sea point, Water point, tonification point.

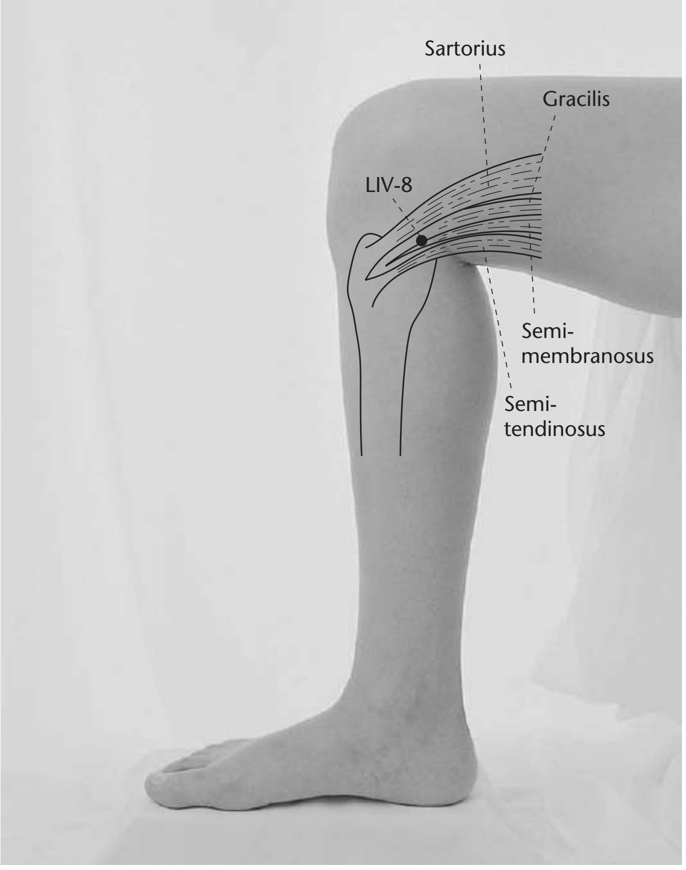

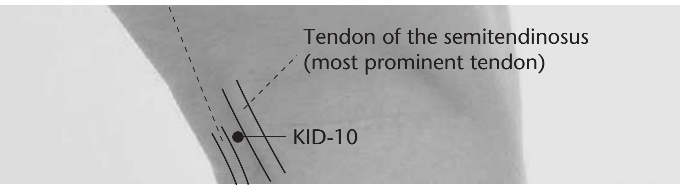

Yin Wrapping YINBAO LIV-9

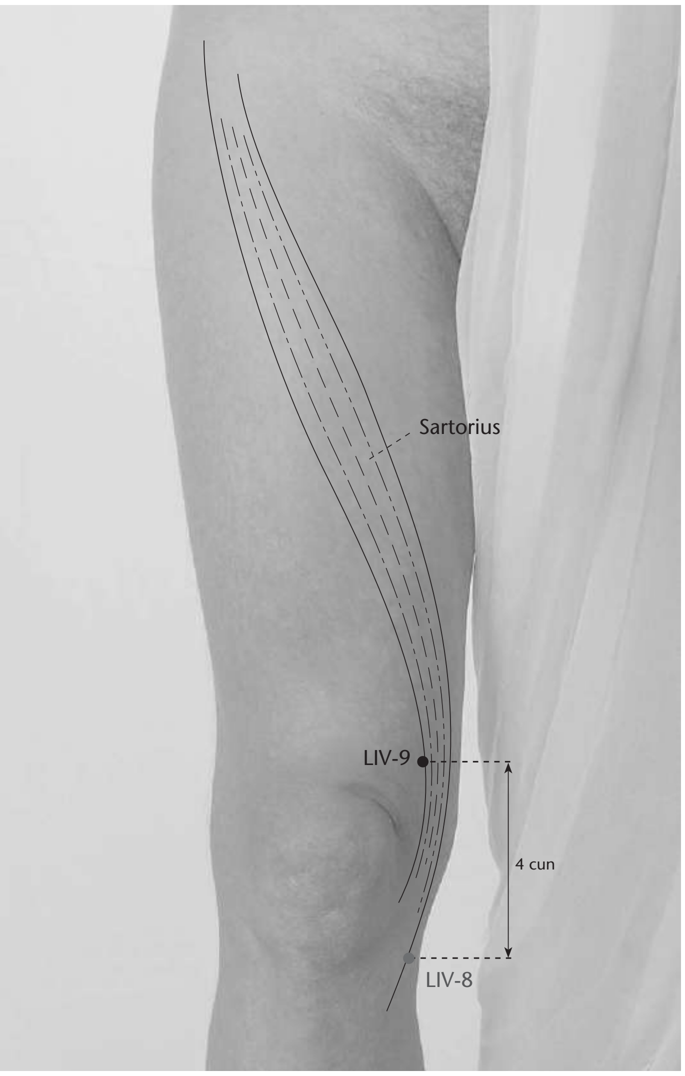
Sartorius
LIV-9
4 cun
LIV-8

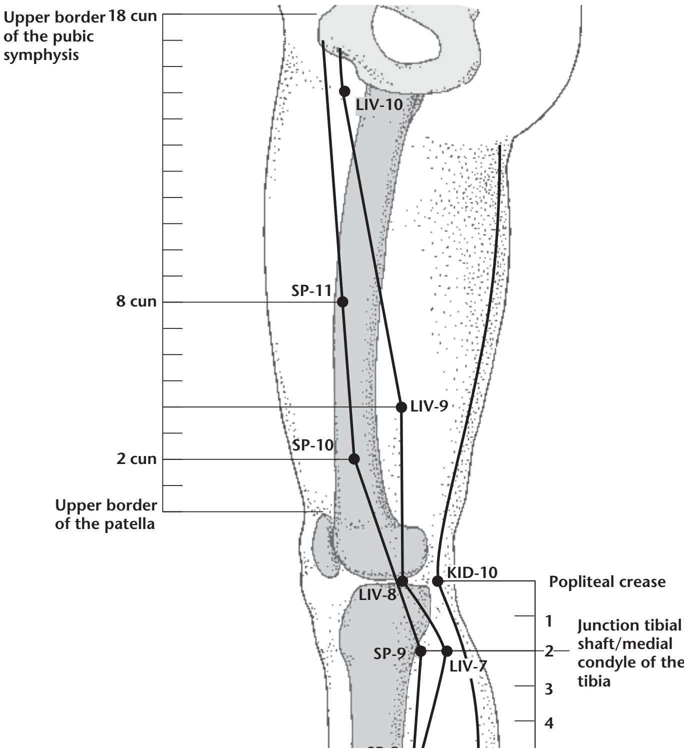
Upper border of the pubic symphysis
18 cun
LIV-10
8 cun
SP-11
LIV-9
2 cun
SP-10
Upper border of the patella
KID-10
Popliteal crease
LIV-8
1
Junction tibial shaft/medial condyle of the tibia
2
SP-9
LIV-7
3
4

Location

4 cun proximal to the medial condyle of the femur, between the sartorius and vastus medialis muscles. Alternative locations: 4 cun proximal to ➞ **LIV-8** (for example, Deadman et al 1998, ➞ photo), 4 cun proximal to the junction of the shaft and the condyle of the femur (for example, Hecker et al ➞ see line drawing) or 4 cun proximal to the midpoint of the condyle of the femur (for example, Ogal and Stör 1999, ➞ illustration on p. 443).

How to find

From ➞ **LIV-8**, measure 4 cun in a proximal direction. There, locate **LIV-9** on the border between the more medial, narrow sartorius muscle and the anteromedial vastus medialis muscle. Sensitivity to pressure should determine the location of this point.

Needling

Vertically 1–2 cun

Actions/Indications

- Benefits the Lower Burner
- Opens the channel

LIV-10 *Leg Five Miles* ZUWULI

Location

On the anterior aspect of the thigh, 3 cun inferior to the upper border of the symphysis, on the lateral border of the adductor longus muscle.

How to find

Ask the patient to flex their adductors by pressing their feet against each other. **LIV-10** is located in a groove on the medial border of the adductor longus muscle, where it forms an angle with the sartorius muscle.

Needling

Vertically 0.5–1.5 cun. Caution: Great saphenous vein, femoral artery/nerve.

Actions/Indications

- Relaxes the tendons and muscles
- Clears Damp-Heat from the Lower Burner
- Regulates the water passages

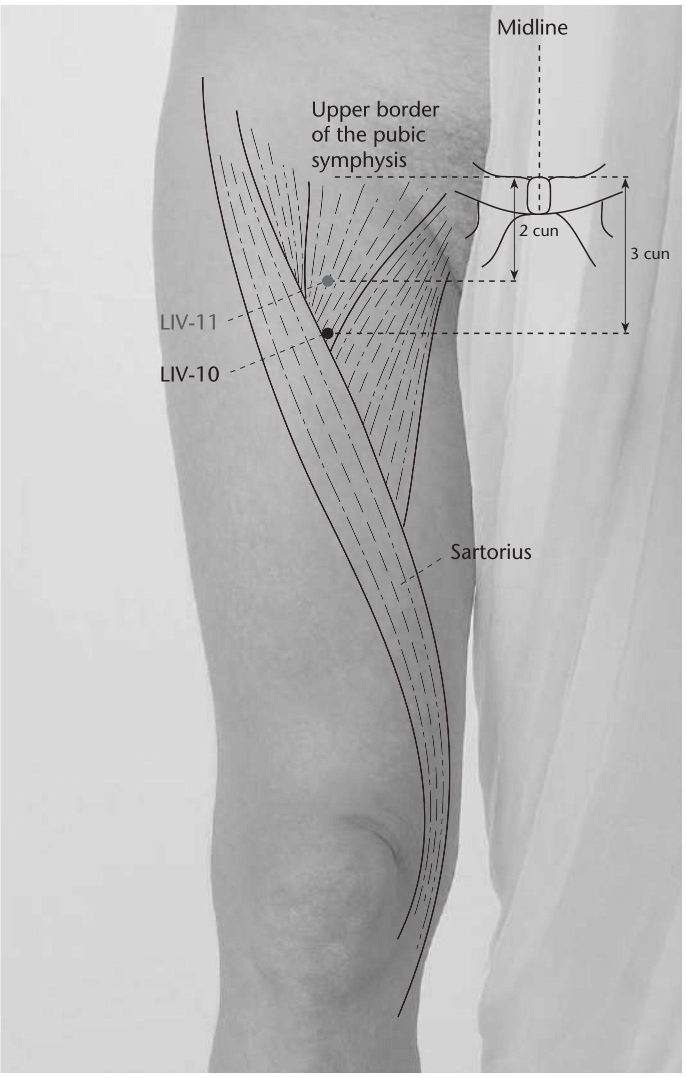
Midline
Upper border of the pubic symphysis
2 cun
3 cun
LIV-11
LIV-10
Sartorius

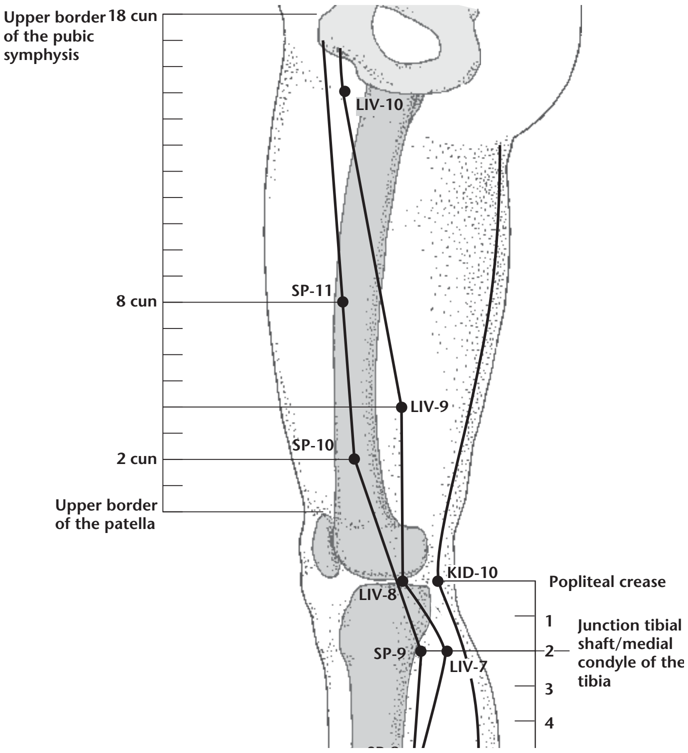
Upper border of the pubic symphysis
18 cun
LIV-10
8 cun
SP-11
LIV-9
SP-10
2 cun
Upper border of the patella
KID-10
Popliteal crease
LIV-8
1
Junction tibial shaft/medial condyle of the tibia
2
SP-9
LIV-7
3
4

Yin Corner YINLIAN **LIV-11**

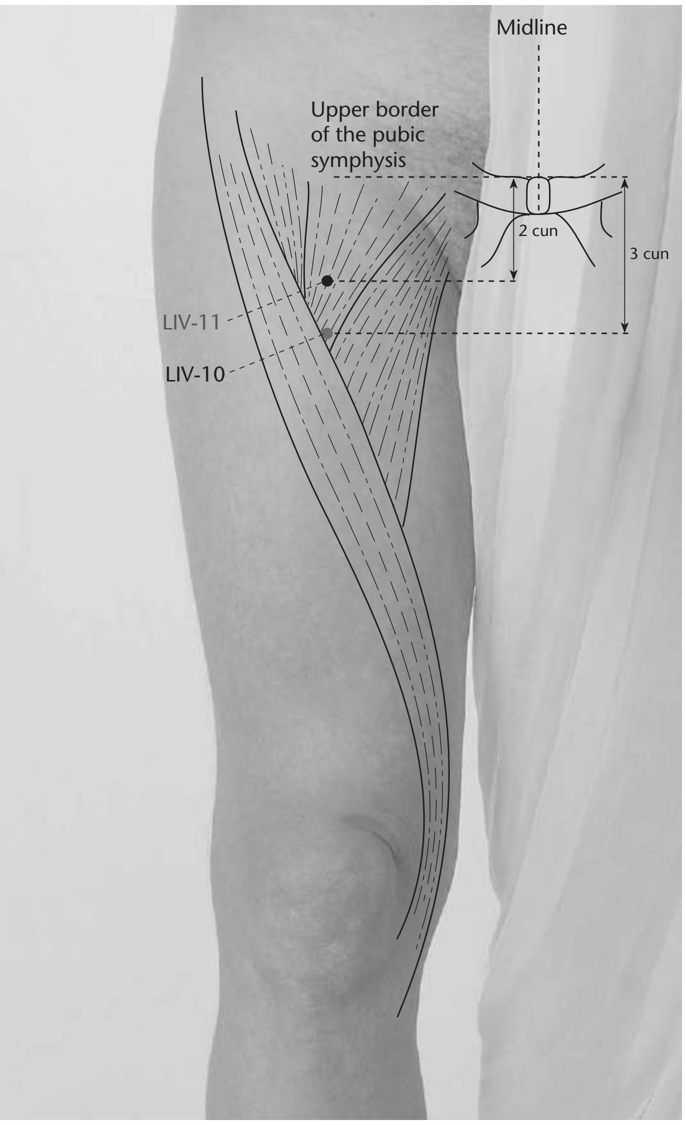
Midline
Upper border
of the pubic
symphysis
2 cun
3 cun
LIV-11
LIV-10

Location

On the anterior aspect of the thigh, 2 cun inferior to the upper border of the symphysis and on the lateral border of the adductor longus muscle.

How to find

Ask the patient to flex their adductors by pressing their feet against each other. **LIV-11** is located 2 cun inferior to the upper border of the pubic symphysis, in a groove on the medial border of the adductor longus muscle.

LIV-11 is located approximately 1 cun inferior to where the femoral artery passes under the inguinal ligament.

Needling

Vertically 0.5–1.5 cun. Caution: Great saphenous vein, femoral artery/vein/nerve. Classical texts recommend moxibustion for female sterility.

Actions/Indications

- Relaxes the tendons
- Benefits the Uterus and regulates menstruation
- For disorders of the lumbar spine

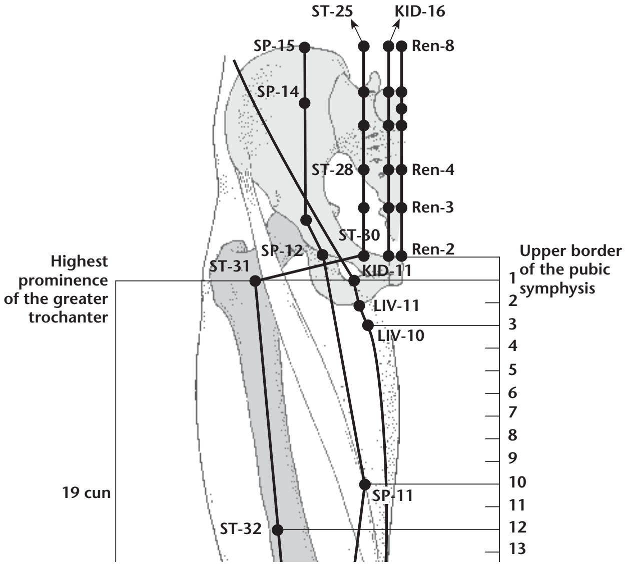
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
ST-30
SP-12
Ren-2
Highest
prominence
of the greater
trochanter
ST-31
KID-11
Upper border
of the pubic
symphysis
LIV-11
LIV-10
1
2
3
4
5
6
7
8
9
10
11
12
13
19 cun
SP-11
ST-32

LIV-12 *Urgent Pulse* JIMAI

Location

In the inguinal groove, 2.5 cun lateral to the anterior midline and 1 cun inferior to the upper border of the symphysis.

How to find

From the upper border of the pubic symphysis, measure 2.5 cun in a lateral and 1 cun in an inferior direction. In this area, you can palpate the pulse of the femoral artery. **LIV-12** is located medial to the pulsating of the artery. As the femoral vein also runs medial to the artery, **LIV-12** should be needled in a medial direction, at least 1 fingerbreadth from the artery.

→ **ST-30** is located 1 cun superior and 0.5 cun medial to **LIV-12**.

Needling

Slightly obliquely 0.5–0.8 cun. Caution: Femoral artery/vein. The femoral vein is located medial to the artery and is approximately the breadth of a finger. To avoid injury to the vein, **LIV-12** should not be needled medial to the artery. Owing to its tricky location, classical texts recommend moxibustion only, while in modern texts moxibustion is contraindicated for the same reason.

Actions/Indications

- Dispels Cold from the Liver channel and benefits the Lower Burner

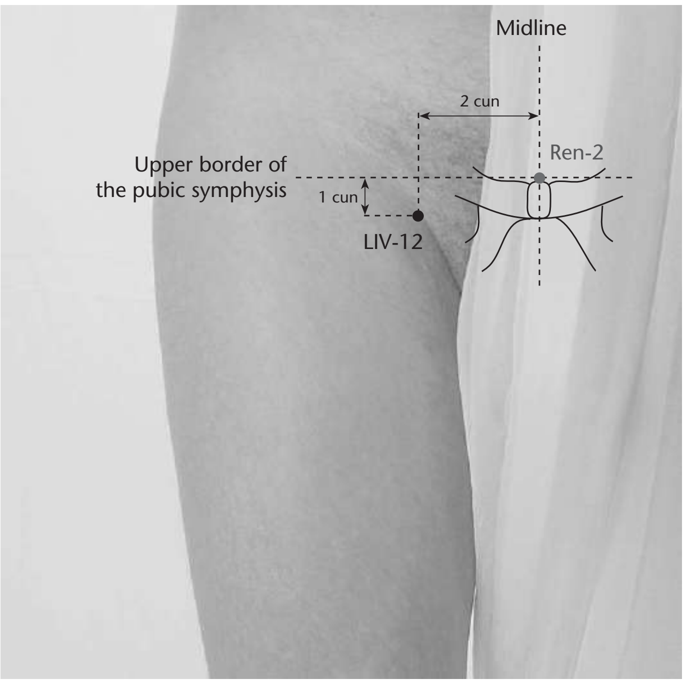
Midline
2 cun
Ren-2
Upper border of the pubic symphysis
1 cun
LIV-12

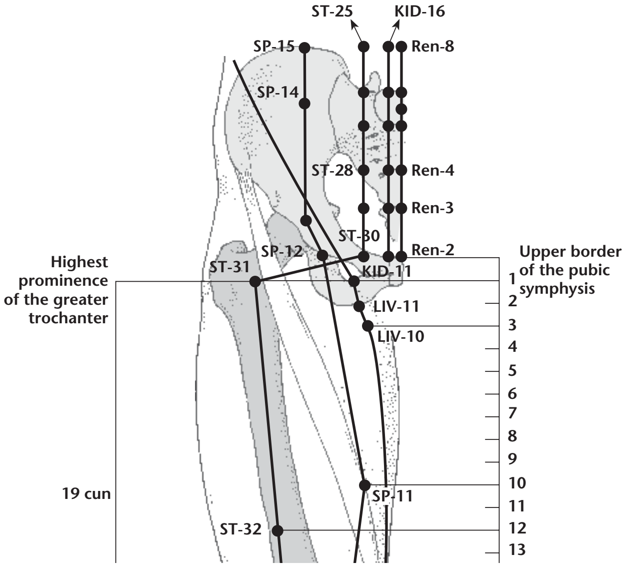
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
ST-30
SP-12
Ren-2
Highest prominence of the greater trochanter
ST-31
KID-11
Upper border of the pubic symphysis
LIV-11
LIV-10
1
2
3
4
5
6
7
8
9
10
11
12
13
19 cun
SP-11
ST-32

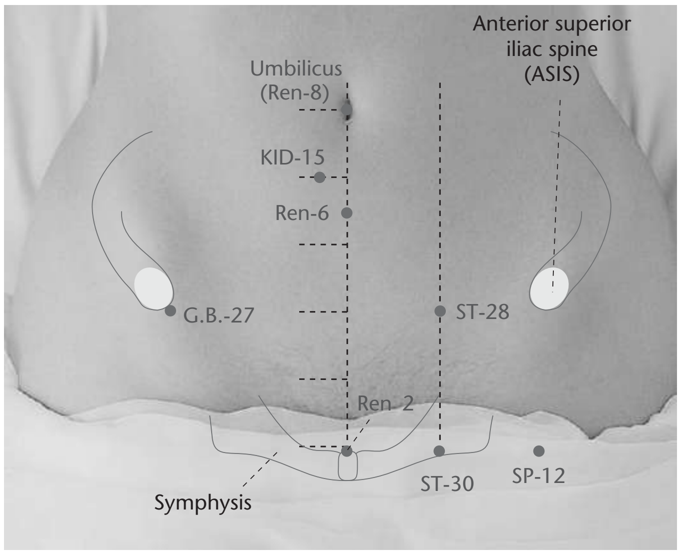
Anterior superior iliac spine (ASIS)
Umbilicus (Ren-8)
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

Completion Gate ZHANGMEN LIV-13

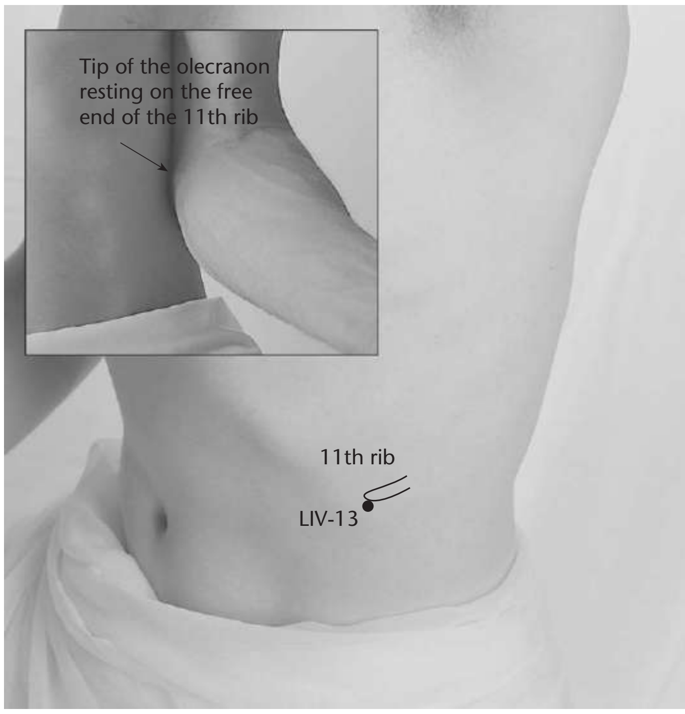
Tip of the olecranon resting on the free end of the 11th rib
11th rib
LIV-13

Location

Anterior and inferior to the free end of the 11th rib, approximately 1 cun superior to the umbilicus.

How to find

Quick method: Ask the patient to press their flexed elbow (90°) against the lateral aspect of the thorax and the tip of the olecranon will be pointing to the free end of the 11th rib. **LIV-13** can then be located on its anterior and inferior border. **Or:** Place the hand on the upper abdomen and, with gentle pressure, palpate along the lower border of the rib cage until you can feel the free end of the 11th rib just superior to the umbilicus. Locate **LIV-13** on its anterior and inferior border.

Located on approximately the same level (1 cun superior to the umbilicus) are → **Ren-9** (on the midline) and → **ST-24** (2 cun lateral to the midline). → **G.B.-25** is located more laterally and inferiorly, at the free end of the 12th rib.

Needling

0.5–1 cun vertically or better obliquely in a lateral direction. Caution: Peritoneum, hypertrophied organs (right: liver, left: spleen).

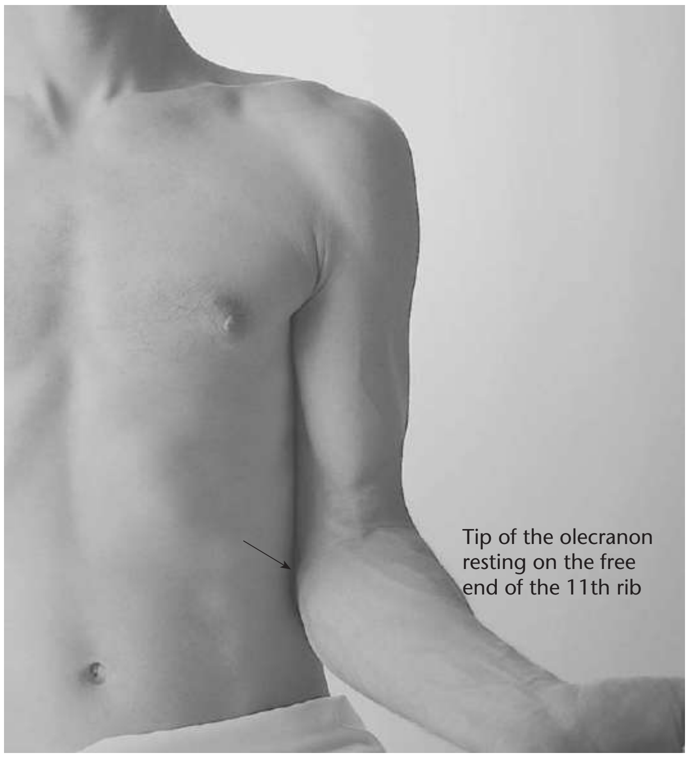
Tip of the olecranon resting on the free end of the 11th rib

Actions/Indications

- Harmonises the Liver and the Spleen, regulates Liver Qi (especially in the Middle and Lower Burner), strengthens the Spleen
- Opens the channel

Special features

Front-*mu* point of the Spleen, meeting point with the *dai mai* and the G.B. channel. *Hui*-meeting point of the *zang*-Organs.

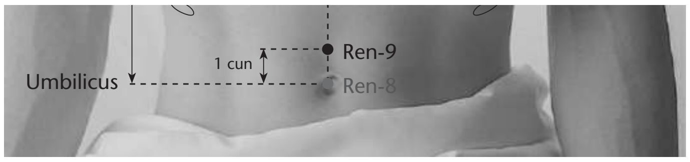
Umbilicus
1 cun
Ren-9
Ren-8

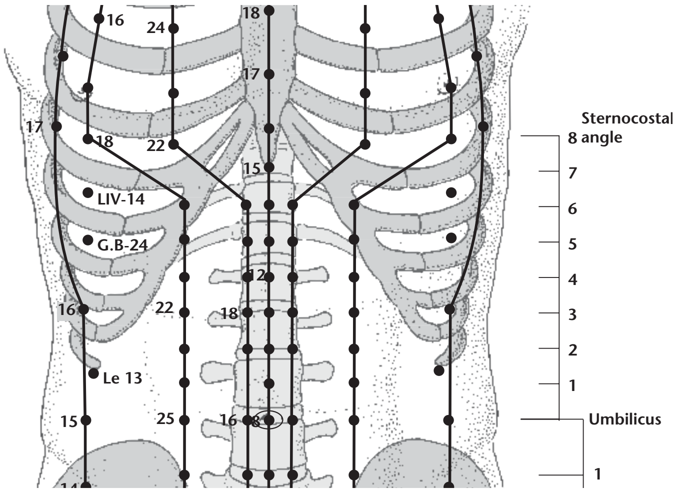
16
24
18
17
17
18
22
15
LIV-14
G.B-24
12
16
22
18
Le 13
15
25
16
8
14
Sternocostal angle
8
7
6
5
4
3
2
1
Umbilicus
1

LIV-14 *Cycle Gate* QIMEN

Location

In the 6th intercostal space, on the mamillary line or 4 cun lateral to the midline.

How to find

Quick method, especially in men: The nipple is generally located on the level of the 4th intercostal space. From there, palpate downward 2 intercostal spaces and locate **LIV-14** in the 6th intercostal space. **Or:** For more precise orientation on the ribcage (→ 3.5), palpate for the manubriosternal synchondrosis, a well-defined horizontal bony structure on the sternum. Lateral to it, locate the 2nd rib and below it the 2nd intercostal space. From there, count downward 4 intercostal spaces to the 6th intercostal space and there locate **LIV-14** on the mamillary line (4 cun lateral to the midline).

→ **G.B.-24** is located in the 7th intercostal space. Located on the same level (2 cun inferior to the sternocostal angle → 3.5) are → **Ren-14** (on the midline), → **KID-21** (0.5 cun lateral to the midline) and → **ST-19** (2 cun lateral to the midline).

Needling

0.5–1 cun obliquely, either in a lateral or medial (along the course of the ICS) direction. The latter is recommended for disorders of the breasts. Caution: Pneumothorax.

Actions/Indications

- Regulates Liver Qi and Liver Blood (especially in the Upper and Middle Burner), cools the Blood, disperses accumulations, harmonises the Liver and the Stomach

Special features

Front-*mu* point of the Liver, meeting point with the *yin wei mai* and the Spleen channel, exit point.

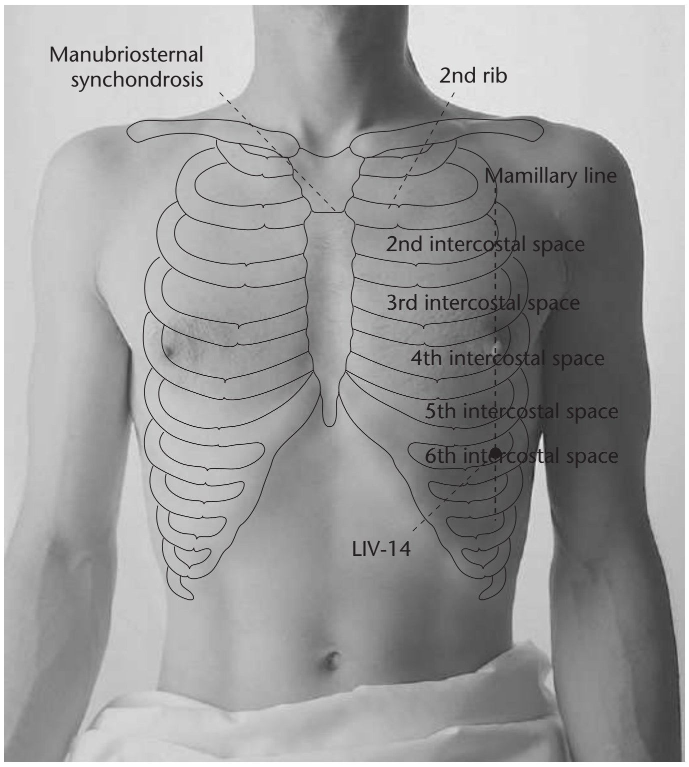
Manubriosternal synchondrosis
2nd rib
Mamillary line
2nd intercostal space
3rd intercostal space
4th intercostal space
5th intercostal space
6th intercostal space
LIV-14

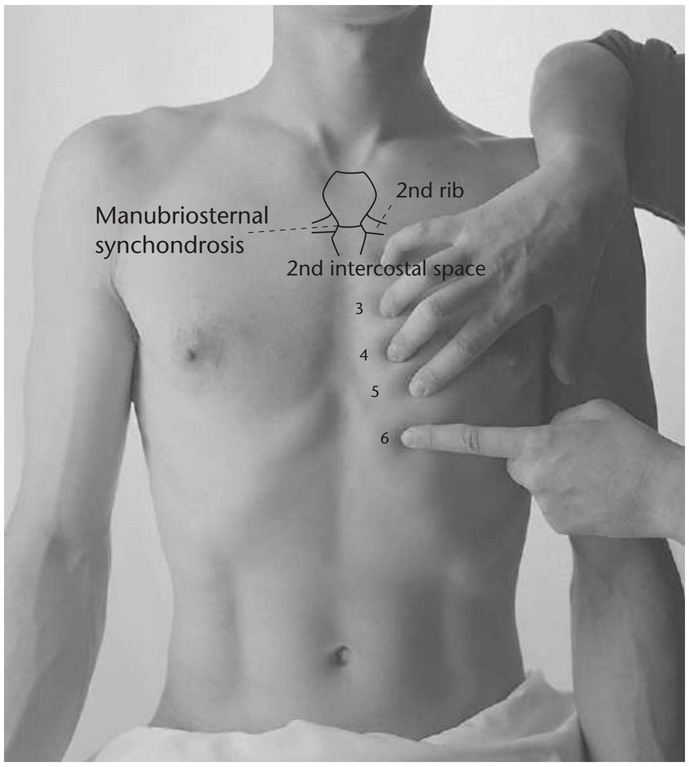
Manubriosternal synchondrosis
2nd rib
2nd intercostal space
3
4
5
6

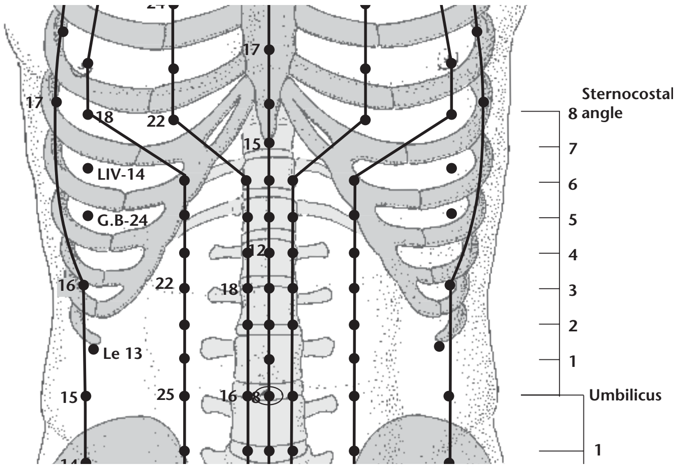
17
17
18
22
15
LIV-14
G.B-24
12
16
22
18
Le 13
15
25
16
8
Sternocostal angle
8
7
6
5
4
3
2
1
Umbilicus
1

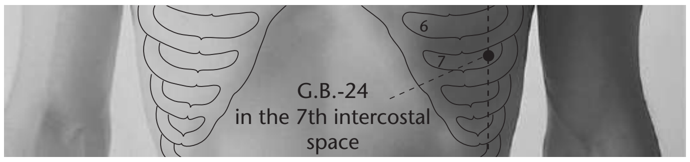
6
7
G.B.-24
in the 7th intercostal space

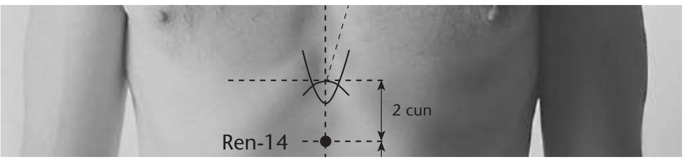

5 The Eight Extraordinary Vessels (*qi jing ba mai*) with Points

Claudia Focks, Ulrich März

For an overview of the theoretical background of the eight extraordinary vessels → 1.7.

5.1 *Chong mai*

Synonym: Penetrating Vessel

Relationships (→ 1.7.3)

- **Yin/Yang:** *chong mai*/*dai mai*
- **Central/peripheral:** *chong mai*/*yin wei mai*. Regions supported by this pairing: Heart, thorax, Stomach. Opening point: SP-4 (*gongsun*), coupled point: P-6 (*neiguan*)

Pathway

The extraordinary vessel *chong mai* originates like the *du mai* and *ren mai* in the lower abdomen at *bao zhong* (in women in the Uterus) or, according to some authors, at the Kidneys. It crosses the genital area and divides into two branches:

- ➡ A **deep branch** reaches the perineum at → **Ren-1** (*huiyin*), travels to the spine, penetrates the spinal canal and ascends to the level of T1 and T2. (According to some authors the vessel ascends the anterior aspect of the spine.)
- ➡ The **other branch** emerges at → **ST-30** (*qichong*) to the surface of the body and divides into **two external branches**, one descending and one ascending branch:
- the **ascending** branch reaches → **KID-11** (*henggu*) 0.5 cun lateral to the midline at the level of the pubic symphysis and from there follows the KID channel to → **KID-21** (*youmen*). The KID channel initially ascends the abdomen at a distance of 0.5 cun from the midline, passing → **KID-12** (*dahe*), → **KID-13** (*qixue*), → **KID-14** (*siman*) and → **KID-15** (*zhongzhu*), runs to the midline, meeting → **Ren-7** (*yinjiao*) 1 cun inferior to the umbilicus, and again returns to its pathway 0.5 cun lateral to the midline, passing → **KID-16** (*huangshu*), → **KID-17** (*shangqu*), → **KID-18** (*shiguan*), → **KID-19** (*yindu*), → **KID-20** (*futonggu*) and → **KID-21** (*youmen*). The vessel then disperses into the intercostal spaces, ascends the neck, circles the inner surface of the lips and spreads in the upper pharynx and the nasal cavities.

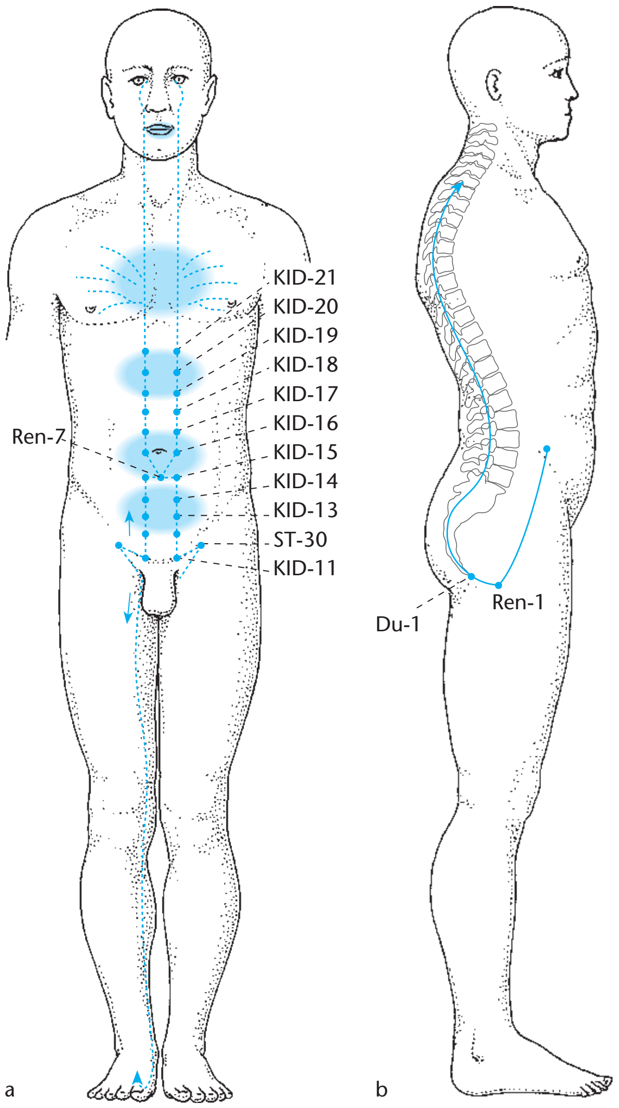

- the **descending branch** meets the Kidney primary channel and descends the medial aspect of the thigh (according to the *Ling Shu* it divides into the three foot Yin channels), penetrates the popliteal fossa, follows the medial border of the tibia and emerges posterior to the medial malleolus, where it divides into two sub-branches: one branch travels along the anteromedial aspect of the foot to the big toe, the other branch reaches the sole of the foot.

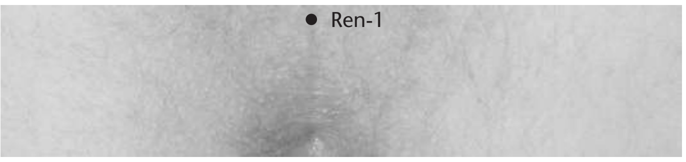

Ren-1 (*huiyin*): In the centre of the perineum.

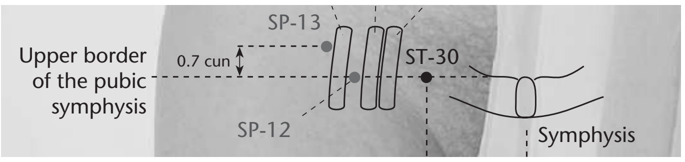
SP-13
Upper border of the pubic symphysis
0.7 cun
ST-30
SP-12
Symphysis

ST-30 (*qichong*): 2 cun lateral to the upper border of the pubic symphysis, medial to the femoral artery; at the level of Ren-2

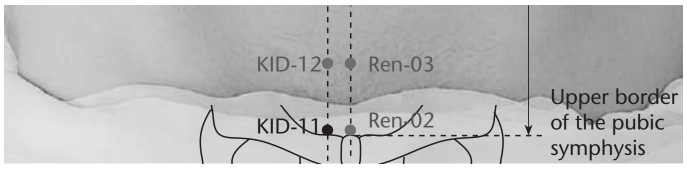
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

KID-11 (*henggu*): On the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline

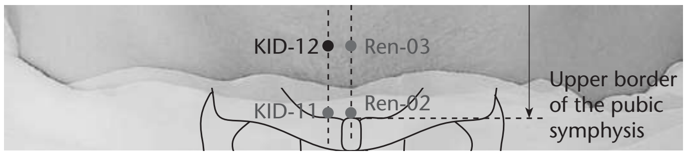
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

KID-12 (*dahe*): 1 cun superior to the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline

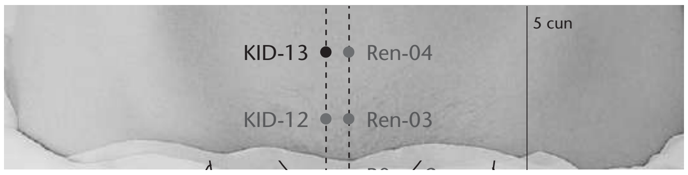
5 cun
KID-13
Ren-04
KID-12
Ren-03

KID-13 (*qixue*): 2 cun superior to the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline

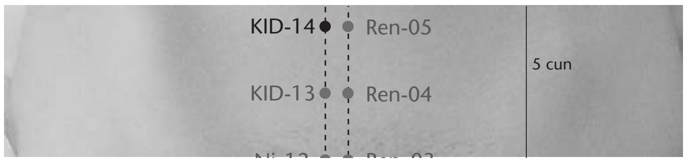
KID-14
Ren-05
5 cun
KID-13
Ren-04

KID-14 (*siman*): 2 cun inferior to the umbilicus, 0.5 cun lateral to the anterior midline

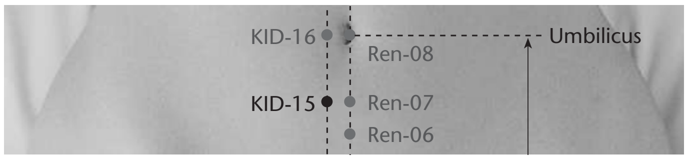
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06

KID-15 (*zhongzhu*): 1 cun inferior to the umbilicus, 0.5 cun lateral to the anterior midline

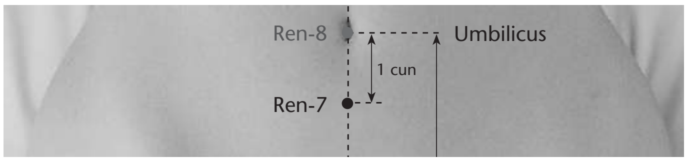
Ren-8
Umbilicus
1 cun
Ren-7

Ren-7 (*yinjiao*): On the anterior midline, 1 cun inferior to the umbilicus

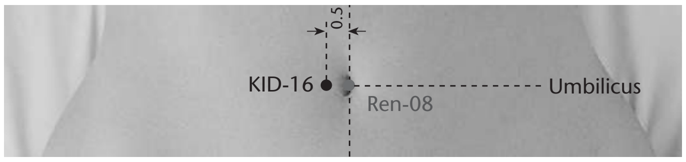
0.5
KID-16
Umbilicus
Ren-08

KID-16 (*huangshu*): 0.5 cun lateral to the centre of the umbilicus

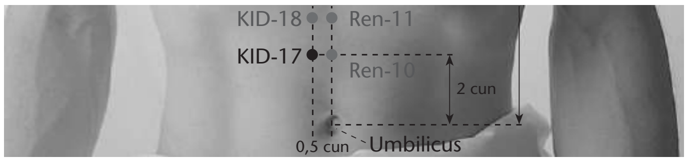
KID-18
Ren-11
KID-17
Ren-10
2 cun
0,5 cun
Umbilicus

KID-17 (*shangqu*): 2 cun superior to the umbilicus, 0.5 cun lateral to the anterior midline

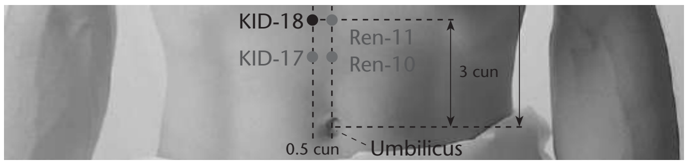
KID-18
Ren-11
KID-17
Ren-10
3 cun
0.5 cun
Umbilicus

KID-18 (*shiguan*): 3 cun superior to the umbilicus, 0.5 cun lateral to the anterior midline

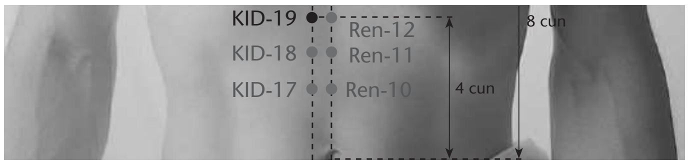
KID-19
Ren-12
8 cun
KID-18
Ren-11
KID-17
Ren-10
4 cun

KID-19 (*yindu*): Midway between the sternocostal angle and the centre of the umbilicus, 0.5 cun lateral to the anterior midline

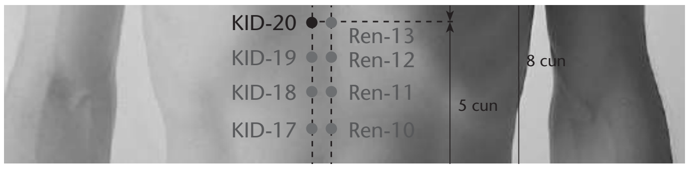

KID-20 (*futonggu*): 5 cun superior to the umbilicus or 3 cun inferior to the sternocostal angle, 0.5 cun lateral to the anterior midline

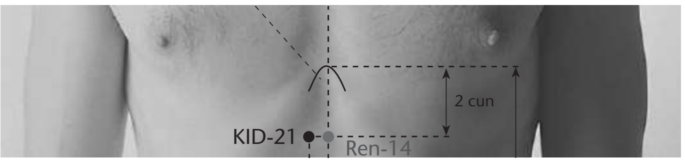

KID-21 (*youmen*): 2 cun inferior to the sternocostal angle, 0.5 cun lateral to the anterior midline

Connections with other channels and organs

- Kidney primary channel, *ren mai*, *du mai*, Stomach and Liver primary channels
- Kidney, Uterus

Clinical importance (→ 1.7.2, 1.7.3)

- **Sea of Blood**, strong connection to the various functions of the Blood
- Strong relationship to the Uterus, regulates the menses
- Regulates the 12 primary channels, effectively regulates the circulation of the Blood (*xue*) in the whole body
- Regulates counterflow Qi
- Connects the Stomach and Kidney primary channels and strengthens the relationship between the *ren mai* and *du mai*

5.2 *Ren mai*

Synonym: Conception Vessel

Relationships (→ 1.7.3)

- **Yin/Yang:** *ren mai/du mai*
- **Central/peripheral:** *ren mai/yin qiao mai*. Regions supported by this pairing: face, throat, thorax, Lungs, diaphragm, abdomen. Opening point: **LU-7** (*lieque*); coupled point: **KID-6** (*zhaohai*)

Pathway

The extraordinary vessel *ren mai* originates like the *chong mai* and *du mai* in the lower abdomen (*bao zhong*) or, according to some authors, at the Kidneys, connects with the urogenital area and emerges at the perineum at **Ren-1** (*huiyin*). From there the **external** pathway ascends the anterior midline, passing the abdomen, the chest, and the throat. It terminates on the chin in the mentolabial groove at **Ren-24** (*chengjiang*), where it meets the Stomach primary channel and the extraordinary vessel *du mai*. From here the vessel continues internally, encircles the inner surface of the lips and connects with the *du mai* at → **Du-28** (*yinjiao*) beneath the upper frenulum. It divides into **two branches**, which both reach the infraorbital ridge at → **ST-1** (*chengqi*), where they cross the Stomach primary channel as well as the extraordinary vessel *yin qiao mai*, and terminate near the eyes.

- **An internal branch** runs from the perineum to the spine, penetrates the spinal canal at **Du-1** (*changqiang*) and ascends in the spine.

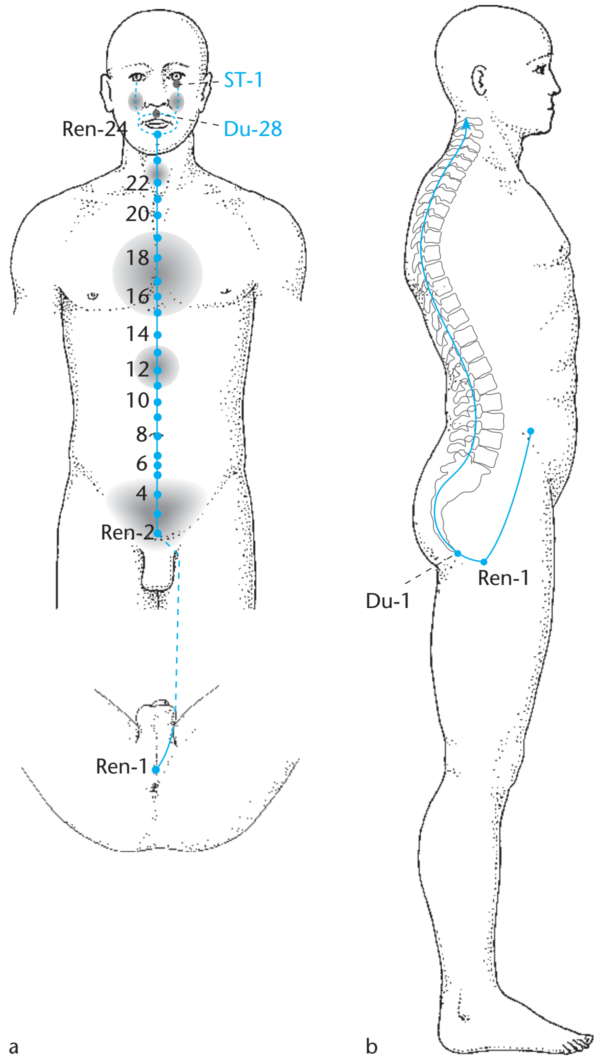

Meeting Points of other channels with the *ren mai*

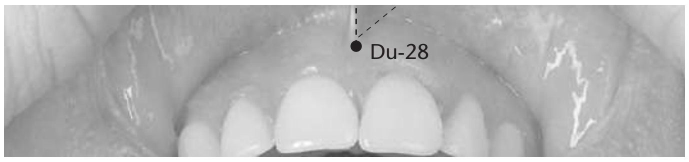

Du-28 (*yinjiao*): On the inside of the upper lip, at the junction of the frenulum and the upper gum

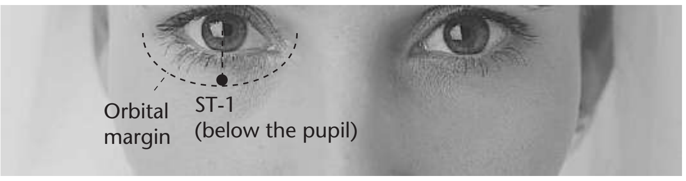

ST-1 (*chengqi*): With the eyes looking straight ahead, directly below the centre of the pupil, between the eyeball and the infraorbital ridge

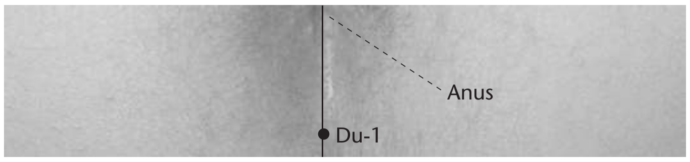

Du-1 (*changqiang*): On the midline, midway between the tip of the coccyx and the anus

Connections with other channels and organ systems

- *Du mai*, *chong mai*, Kidney and Stomach primary channels
- Uterus, Kidney

Clinical importance (→ 1.7.2, 1.7.3)

- **Sea of the Yin Qi**, regulates and circulates the Yin Qi (including the Blood, Essence (*jing*) and body fluids) in order to nourish and moisten the body
- Connects all Yin channels, especially those of the lower abdomen
- Controls the anterior thoracic region and the abdomen
- Close relationship to the functions of the abdominal organs, particularly the Kidney, Bladder, Uterus and Lung
- Regulates reproduction
- Stores the Essence Qi (*jing qi*) overflowing from the Yin channels
- Absorbs excess from the Yin channels, especially stagnant Qi and Blood

Notes about the *ren mai*

Both the *ren mai* and *du mai* have a unique position among the eight extraordinary vessels. They are the only two extraordinary vessels with their own specific points – all other extraordinary vessels can only be accessed through points on the primary channels. For this reason the *ren mai* and the *du mai* together with the primary channels are often categorised as 14 channels. However, from the point of view of channel energetics this is not correct, since the extraordinarily vessels occupy a special position, both evolutionary as well as functionally (→ 1.7).

The pathway of the *ren mai*, like that of the *du mai*, is limited to the torso and the head. While it does not course along the extremities and does not have any points on the limbs (neither general points nor specific points such as the five *shu*-transporting points, *yuan*-source point, etc.) it does have an opening point (or Master point) (→ 8.1.8). The *ren mai* is also referred to as the Sea of Yin and governs the Yin in the whole body. Along its pathway are located the anterior aspects of the energy centres (or chakras according to Yogi tradition). Daoist tradition describes three energy centres, the three cinnabar fields (*dantian*) on the anterior aspect of the body, which play an important role in *Qigong* practice as centres where the Qi pools and originates. While all three centres are located on the pathway of the *ren mai*, they do not correspond to specific points but rather to certain areas.

The *ren mai* runs along the Yin side of the body. It contains not only several Front-*mu* points acting directly on the *zangfu*-Organs, but also multiple meeting points with other channels. This clearly denotes the outstanding importance of the *ren mai*.

Specific points according to their function

- ***Yuan*-Source point (→ 8.1.1): –**
- ***Luo*-Connecting point (→ 8.1.2.): Ren-15** (*jiuwei*) 🟦■
- ***Xi*-Cleft point (→ 8.1.3): –**
- **Associated Back-*shu* point (→ 8.1.4): –**
- **Associated Front-*mu* point (→ 8.1.5): –**
- **Five *shu*-transporting points (→ 8.1.6): –**
- **Hui-meeting point (→ 8.1.7)**
  - – of the Qi: **Ren-17** (*danzhong*) 🟦🟦
  - – of the *fu*-Organs: **Ren-12** (*zhongwan*) 🟦■
- **Pertaining opening point (→ 8.1.8): LU-7** (*lieque*)
- **Pertaining coupled point (→ 8.1.8): KID-6** (*zhaohai*)
- **Lower *he*-Sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - – with the *chong mai* and the *du mai*: **Ren-1** (*huiyin*)
  - – with the LIV channel: **Ren-2** (*qugu*) 🟦■
  - – with the SP, LIV, KID channels: **Ren-3** (*zhongji*) 🟦🟦, **Ren-4** (*guanyuan*) 🟦🟦
  - – with the *chong mai* and the KID channel*: **Ren-7** (*yinjiao*)
  - – with the SP and ST* channels: **Ren-10** (*xiawan*)
  - – with the S.I., T.B.* and ST channels: **Ren-12** (*zhongwan*) 🟦🟦
  - – with the ST and S.I. channels: **Ren-13** (*shangwan*)
  - – with the SP*, KID, S.I.* and T.B. channels: **Ren-17** (*danzhong*) 🟦🟦
  - – with the *yin wei mai*: **Ren-22** (*tiantu*) 🟦🟦, **Ren-23** (*lianquan*)

---

* Mentioned by only some authors

  –468 with the *du mai*, the L.I. and ST channels: **Ren-24** (*chengjiang*) 🟦⬛
  - of other channels with the *ren mai*: **Du-28, ST-1, Du-1**
- **Gao Wu command point (→ 8.1.11):** –
- **Window of Heaven point (→ 8.1.12): Ren-22** (*tiantu*) 🟦🟦
- **Points of the Four Seas (→ 8.1.13) Qi: Ren-17** (*danzhong*) 🟦🟦
- **Ma Dan Yang Heavenly Star point (→ 8.1.14):** –
- **Sun Si Miao Ghost points (→ 8.1.15): Ren-1** (*huiyin*), **Ren-24** (*chengjiang*) 🟦⬛
- **Other functional points:**
  - Front-*mu* point of the Bladder: **Ren-3** (*zhongji*) 🟦🟦
  - Front-*mu* point of the Small Intestine: **Ren-4** (*guanyuan*) 🟦🟦
  - Front-*mu* point of the Triple Burner: **Ren-5** (*shimen*)
  - important, general tonification point: **Ren-6** (*qihai*) 🟦🟦
  - Front-*mu* point of the Stomach: **Ren-12** (*zhongwan*) 🟦🟦
  - Front-*mu* point of the Heart: **Ren-14** (*juque*) 🟦🟦
  - important calming point: **Ren-15** (*jiuwei*) 🟦⬛
  - Front-*mu* point of the Pericardium, point of the Sea of Qi: **Ren-17** (*danzhong*) 🟦🟦

Points according to region

- **Local points (→ 8.2.1):** throat – **Ren-23** (*lianquan*); Lungs – **Ren-22** (*tiantu*) 🟦🟦, **Ren-17** (*danzhong*) 🟦🟦; Heart – **Ren-17** (*danzhong*) 🟦🟦; Spleen/Stomach – **Ren-12** (*zhongwan*) 🟦🟦; Small Intestine – **Ren-4** (*guanyuan*) 🟦🟦, **Ren-9** (*shuifen*); Bladder – **Ren-3** (*zhongji*) 🟦🟦; urogenital region – **Ren-3** (*zhongji*), **Ren-4** (*guanyuan*) 🟦🟦; epigastrium – **Ren-12** (*zhongwan*) 🟦🟦, **Ren-13** (*shangwan*); abdomen – **Ren-3** (*zhongji*) 🟦🟦, **Ren-4** (*guanyuan*) 🟦🟦, **Ren-6** (*qihai*) 🟦🟦
- **Adjacent points (→ 8.2.1):** throat – **Ren-22** (*tiantu*) 🟦🟦; tongue – **Ren-23** (*lianquan*); Heart – **Ren-14** (*juque*) 🟦🟦; Gall Bladder – **Ren-11** (*jianli*); Kidney – **Ren-4** (*guanyuan*) 🟦🟦; epigastrium – **Ren-8** (*shenque*)
- **Distal points (→ 8.2.1):** –

Specific points according to the vessel pathway (in numerical order)

- **Ren-1** (*huiyin*): *jiaohui*-meeting point with the *chong mai* and *du mai* (→ 8.1.10); Sun Si Miao Ghost point (→ 8.1.15)
- **Ren-2** (*qugu*) 🟦⬛: *jiaohui*-meeting point with the LIV channel (→ 8.1.10)
- **Ren-3** (*zhongji*) 🟦🟦: *jiaohui*-meeting point with the SP, LIV, KID channels (→ 8.1.10); Front-*mu* point of the Bladder; local point for the Bladder, urogenital region and lower abdomen (→ 8.2.1)
- **Ren-4** (*guanyuan*) 🟦🟦: Front-*mu* point of the Small Intestine (→ 8.1.5); *jiaohui*-meeting point with the SP, LIV, KID channels (→ 8.1.10); local point for the Small Intestine, urogenital region and lower abdomen (→ 8.2.1); adjacent point for the Kidneys
- **Ren-5** (*shimen*): Front-*mu* point of the Triple Burner
- **Ren-6** (*qihai*) 🟦🟦: important general tonification point; local point for the lower abdomen (→ 8.2.1); adjacent point for the Lung (→ 8.2.1)
- **Ren-7** (*yinjiao*): *jiaohui*-meeting point with the *chong mai* and the KID channel (→ 8.1.10)
- **Ren-8** (*shenque*): adjacent point for the epigastrium (→ 8.2.1)
- **Ren-9** (*shuifen*): local point for the Small Intestine (→ 8.2.1)
- **Ren-10** (*xiawan*): *jiaohui*-meeting point with the SP and ST* channels (→ 8.1.10)
- **Ren-11** (*jianli*): adjacent point for the Gall Bladder (→ 8.2.1)
- **Ren-12** (*zhongwan*) 🟦🟦: *hui*-meeting point (→ 8.1.7) of the *fu*-Organs; *jiaohui*-meeting point with the S.I., T.B.* and ST channels (→ 8.1.10); Front-*mu* point of the Stomach (→ 8.1.5); local point for the SP/ST and the epigastrium (→ 8.2.1)
- **Ren-13** (*shangwan*): *jiaohui*-meeting point with the ST and S.I. channels (→ 8.1.10); local point for the epigastrium (→ 8.2.1)
- **Ren-14** (*juque*) 🟦🟦: Front-*mu* point of the Heart (→ 8.1.5); adjacent point for the Heart
- **Ren-15** (*jiuwei*) 🟦⬛: *luo*-connecting point (→ 8.1.2); important point for calming
- **Ren-17** (*danzhong*) 🟦🟦: *hui*-meeting point (→ 8.1.7) of the Qi; *jiaohui*-meeting point with the SP, KID, S.I. and T.B. channels (→ 8.1.10); Front-*mu* point of the Pericardium, point of the Sea of Qi; local point for the Lung and Heart (→ 8.2.1)
- **Ren-22** (*tiantu*) 🟦🟦: *jiaohui*-meeting point with the *yin wei mai* (→ 8.1.10); Window of Heaven point (→ 8.1.12); local point for the Lungs (→ 8.2.1); adjacent point for the throat (→ 8.2.1)
- **Ren-23** (*lianquan*): *jiaohui*-meeting point with the *yin wei mai* (→ 8.1.10); local point for the throat (→ 8.2.1); adjacent point for the tongue (→ 8.2.1)
- **Ren-24** (*chengjiang*) 🟦⬛: *jiaohui*-meeting point with the *du mai*, the L.I. and ST channels (→ 8.1.10); Sun Si Miao Ghost point

---

* Mentioned by only some authors

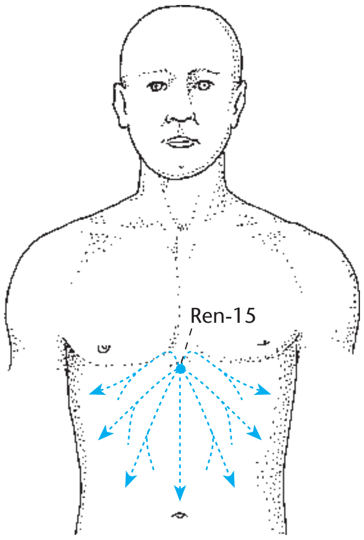

Luo-Connecting Vessel of the *Ren Mai* (→ 1.6)

Pathway

The *luo*-connecting vessel of the *ren mai* originates at **Ren-15** (*jiuwei*), below the xiphoid process (→ 3) and disperses over the skin of the abdomen.

Clinical importance

Pathology

Excess (*shi*): pain of the abdominal skin
Deficiency (*xu*): itching of the abdominal skin

Ch05a-F10028.qxd 2/22/08 9:43 PM Page 471

Meeting of the Yin HUIYIN Ren-1

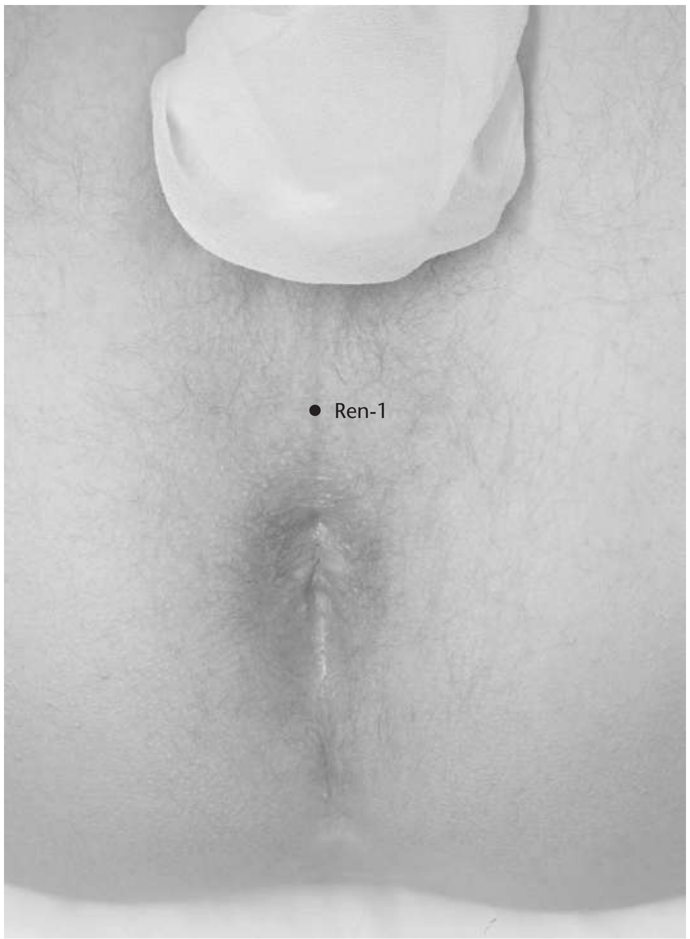

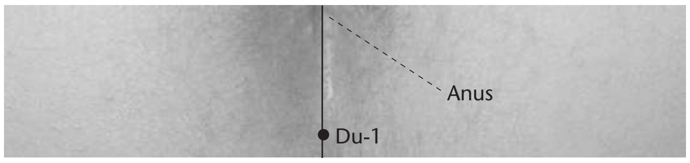

Location

In the centre of the perineum.

How to find

In women: midway between the anus and the posterior labial commissure. In men: midway between the anus and the scrotum.

Needling

Vertically 0.5–1 cun. Caution: Contraindicated during pregnancy. Some classic texts prohibit needling this point. Particular attention should be paid to the prevention of infection.

Actions/Indications

- Clears Heat, strengthens the Kidneys and drains Dampness
- Regulates menstruation

Special features

Meeting point with the *du mai* and the *chong mai*. As a Sun Si Miao Ghost point, **Ren-1** is traditionally indicated for treating epilepsy and psychoses.

Ren-2 *Curved Bone* QUGU

Location

On the anterior midline, on the upper border of the pubic symphysis.

How to find

In the pubic area, locate the upper border of the pubic symphysis. **Ren-2** is located directly superior to its upper border, on the anterior midline.

Located on the same level are → **KID-11** (0.5 cun lateral to the midline), → **ST-30** (2 cun lateral to the midline) and → **SP-12** (3.5 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum. Also caution with a full bladder and during pregnancy.

Actions/Indications

- Benefits urination, warms the Yang and strengthens the Kidneys
- Regulates the Lower Burner

Special features

Meeting point with the LIV channel

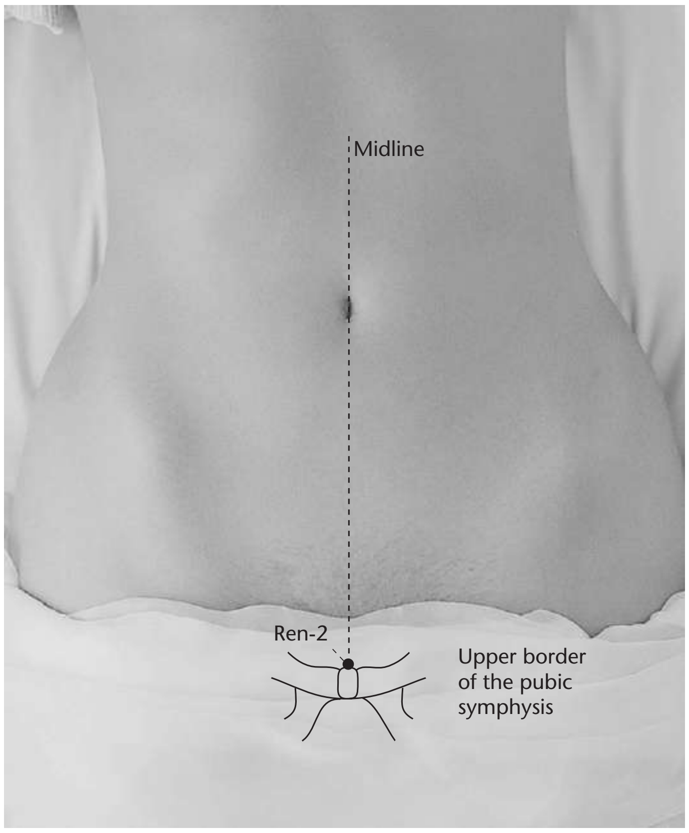
Midline
Ren-2
Upper border of the pubic symphysis

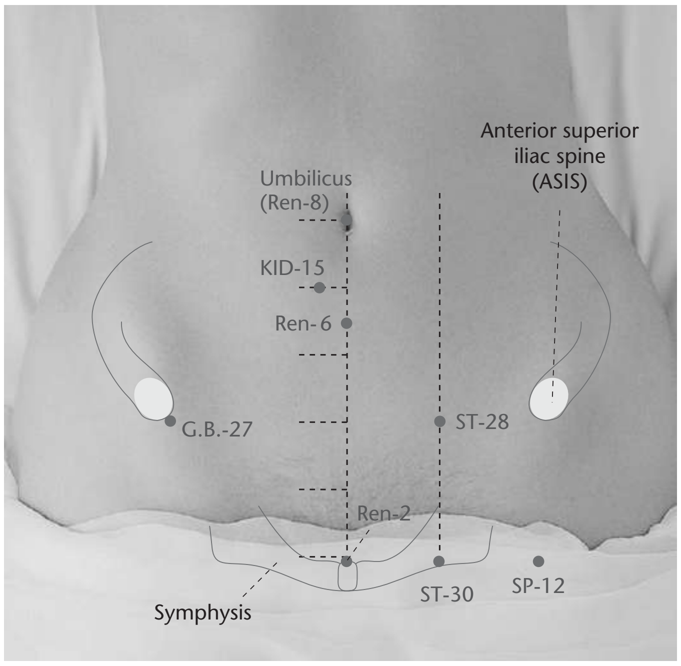
Anterior superior iliac spine (ASIS)
Umbilicus (Ren-8)
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

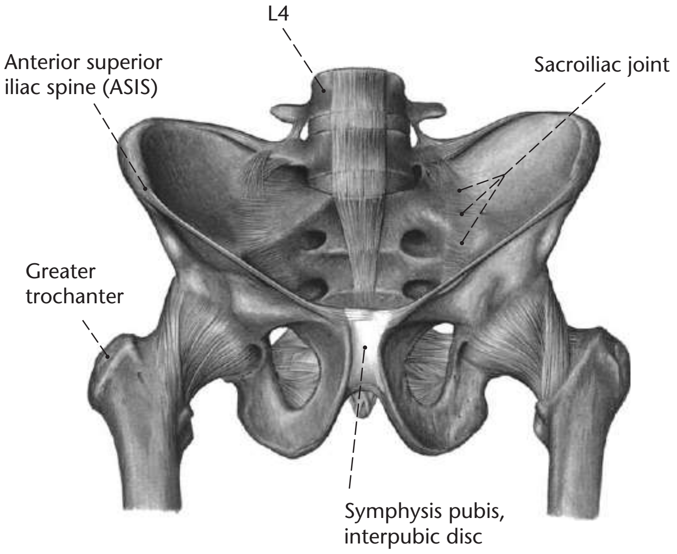
L4
Anterior superior iliac spine (ASIS)
Sacroiliac joint
Greater trochanter
Symphysis pubis, interpubic disc

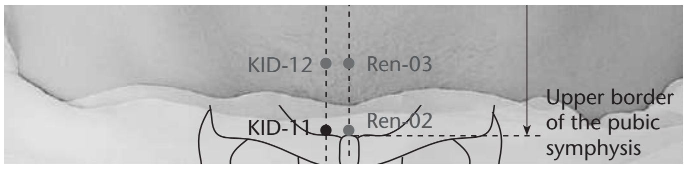
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

Middle Pole ZHONGJI **Ren-3**

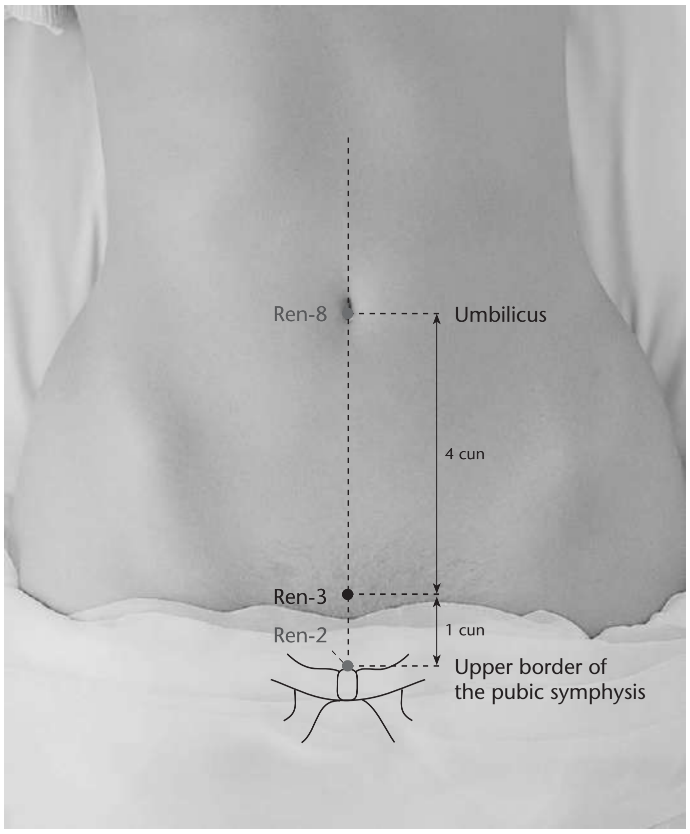
Ren-8
Umbilicus
4 cun
Ren-3
Ren-2
1 cun
Upper border of the pubic symphysis

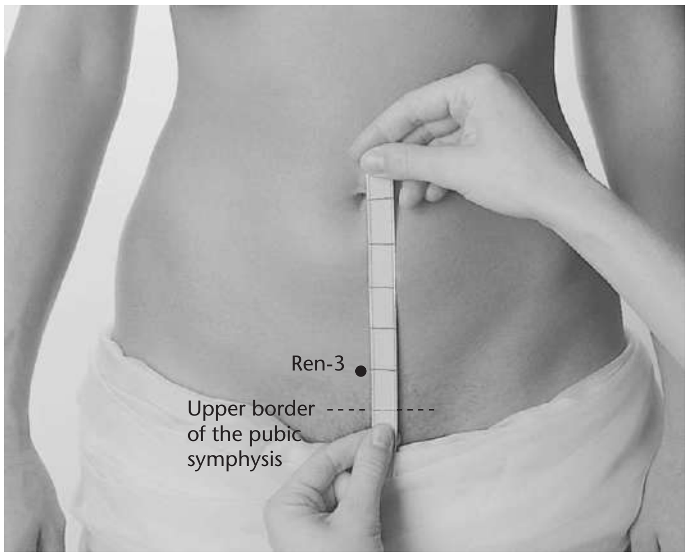
Ren-3
Upper border of the pubic symphysis

Location

On the anterior midline, 1 cun superior to the upper border of the pubic symphysis or 4 cun inferior to the umbilicus.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun. As these can vary considerably from the patient's finger cun, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). Locate **Ren-3** on the midline, 1 cun superior to the midpoint of the upper border of the pubic symphysis.

Located on the same level are → **KID-12** (0.5 cun lateral to the midline) and → **ST-29** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, bladder. Contraindicated during pregnancy. Ask the patient to void their bladder before needling.

Actions/Indications

- Benefits the Bladder, drains Dampness and Damp-Heat, dispels stagnation, benefits the Lower Burner
- Regulates menstruation
- Strengthens the Kidneys (→ **Ren-4** is more commonly used for this purpose)

Special features

Front-*mu* point of the Bladder, meeting point with the LIV, KID and SP channels. Important point for disorders of the urogenital tract, with its main action on the Bladder (especially for acute cases).

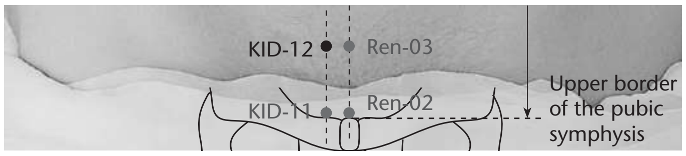
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

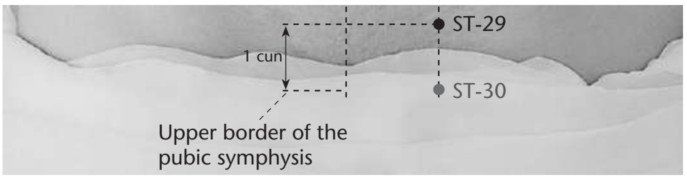
ST-29
1 cun
ST-30
Upper border of the pubic symphysis

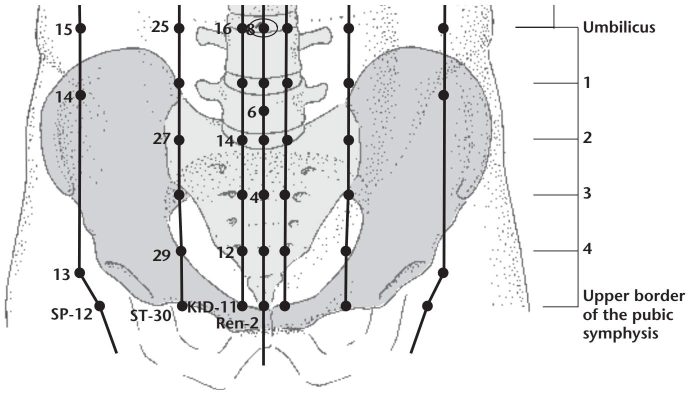
15
25
16
8
14
6
27
14
4
13
29
12
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

Ren-4 *Gate of the Source* GUANYUAN

Location

On the anterior midline, 2 cun superior to the upper border of the pubic symphysis or 3 cun inferior to the umbilicus.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). **Ren-4** is located on the midline, either 2 cun superior to the midpoint of the upper border of the pubic symphysis or 3 cun inferior to the umbilicus.

Located on the same level are → **KID-13** (0.5 cun lateral to the midline), → **ST-28** (2 cun lateral to the midline) and three extra points → **Ex-CA** (*yijing*: 1 cun lateral to the midline; *qimen*: 3 cun lateral to the midline; *tituo*: 4 cun lateral to the midline). Also located on approximately this level is → **G.B.-27** (medial to the ASIS).

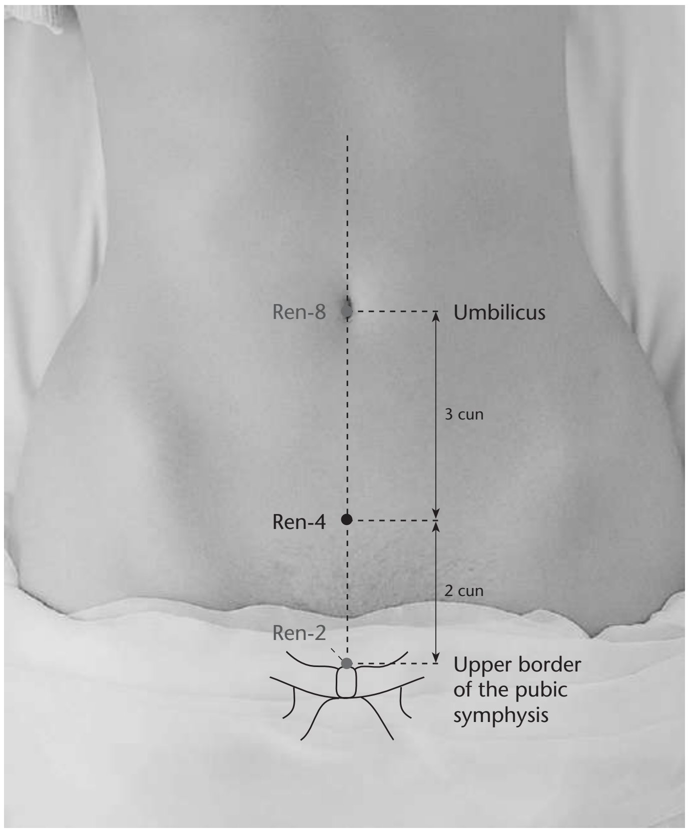
Ren-8
Umbilicus
3 cun
Ren-4
2 cun
Ren-2
Upper border of the pubic symphysis

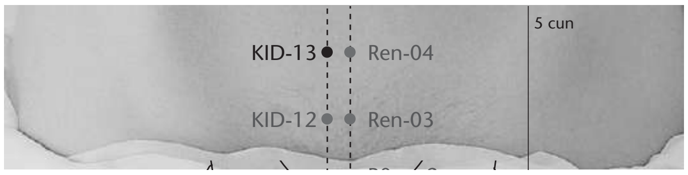
5 cun
KID-13
Ren-04
KID-12
Ren-03

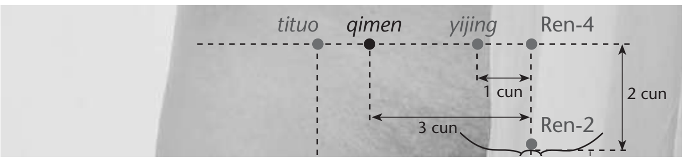
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

Needling

Vertically 0.5–1.2 cun. Caution: Peritoneum. Needling contraindicated during pregnancy.

Actions/Indications

- Tonifies the Original Qi (*yuan qi*), benefits the Essence (*jing*), strengthens and nourishes the Kidneys as well as the Blood and Yin
- Eliminates Cold and Dampness from the Lower Burner (especially with moxibustion)

Special features

Front-*mu* point of the Small Intestine, meeting point with the SP, KID and LIV channels. Important point for tonification. Major point for urogenital disorders (especially gynaecological disorders).

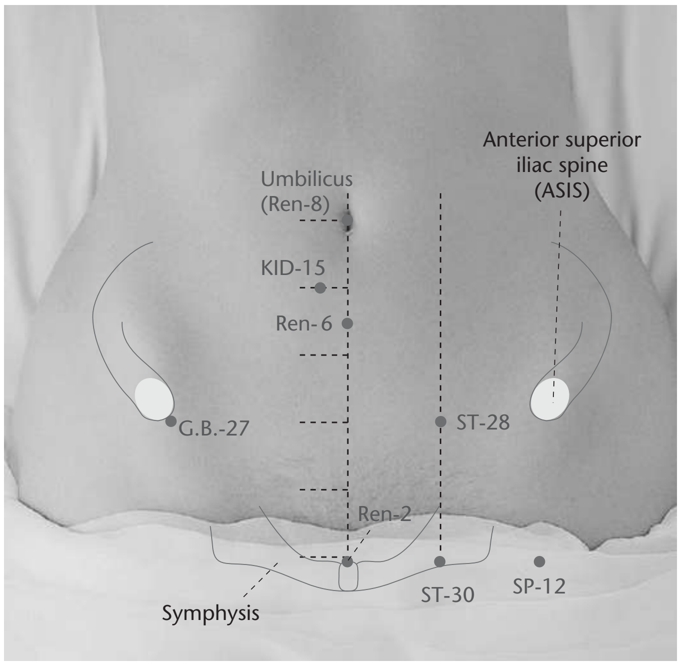
Anterior superior iliac spine (ASIS)
Umbilicus (Ren-8)
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

Stone Gate SHIMEN **Ren-5**

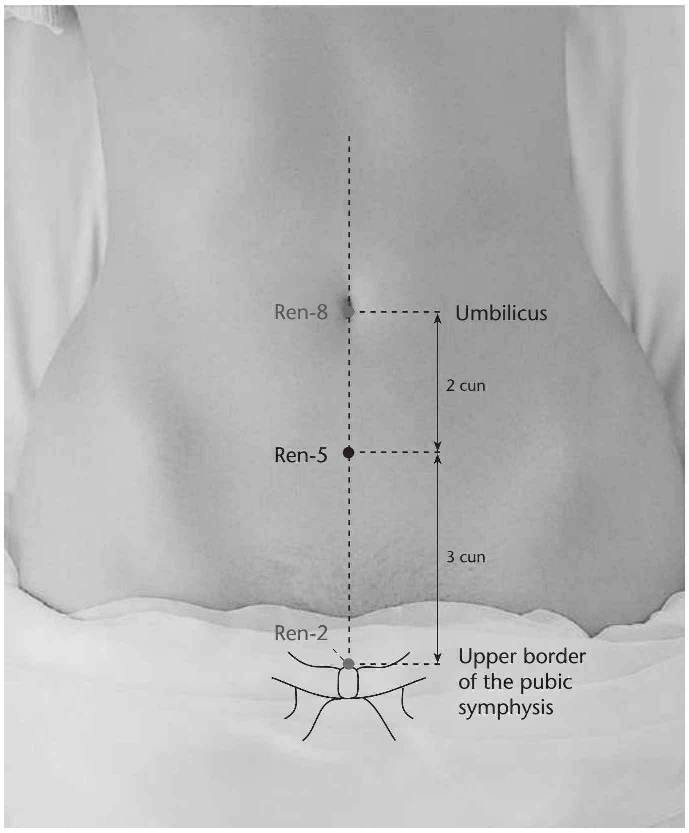
Ren-8
Umbilicus
2 cun
Ren-5
3 cun
Ren-2
Upper border of the pubic symphysis

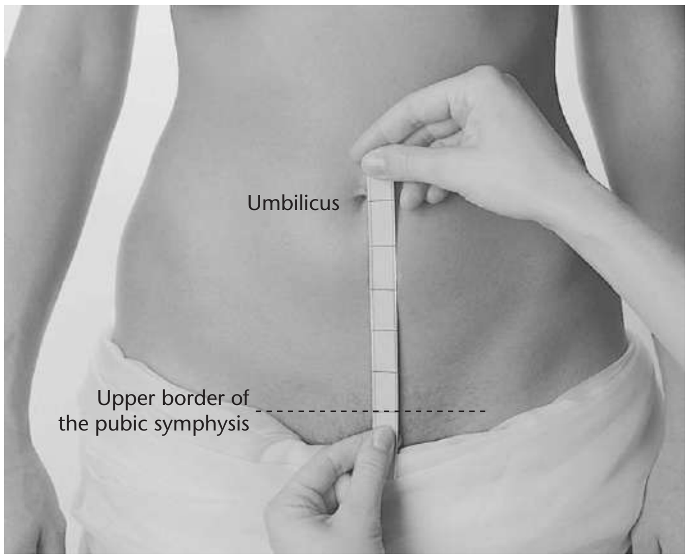
Umbilicus
Upper border of the pubic symphysis

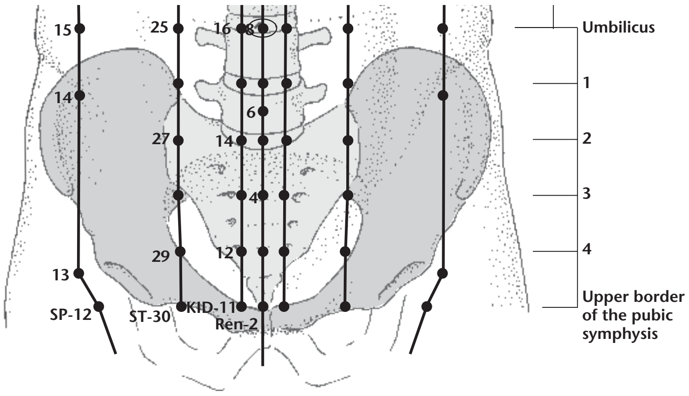
15
25
16
8
14
6
27
14
4
13
29
12
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

Location

On the anterior midline, 2 cun inferior to the umbilicus or 3 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). **Ren-5** is located on the midline, 2 cun inferior to the umbilicus. Located on the same level are → **KID-14** (0.5 cun lateral to the midline) and → **ST-27** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Moves and benefits the water passages
- Regulates the Qi and Blood in the Lower Burner
- Regulates the Uterus

Special features

Front-*mu* point of the Triple Burner. Some classic texts warn that needling this point in women can lead to infertility (hence the name: infertile women were referred to as 'stone women'). There are no such warnings in modern texts.

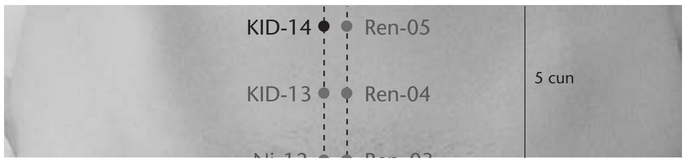
KID-14
Ren-05
KID-13
Ren-04
5 cun

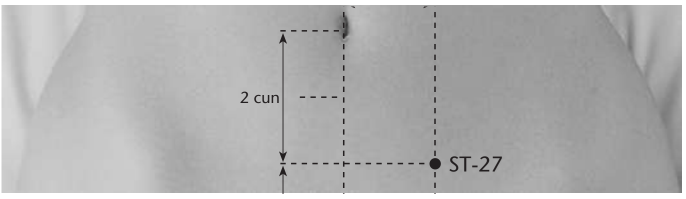

Ch05a-F10028.qxd 2/22/08 9:44 PM Page 476

Ren-6 *Sea of Qi* QIHAI

Location

On the anterior midline, 1.5 cun inferior to the umbilicus or 3.5 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). **Ren-6** is located on the midline, 1.5 cun inferior to the umbilicus or 3.5 cun superior to the upper border of the pubic symphysis.

Needling

Vertically 0.8–1.5 cun. Caution: Peritoneum, pregnancy. According to some authors, this point should never be needled with strong reducing techniques owing to its strong Qi tonifying action. Moxibustion and cupping are recommended for this point.

Actions/Indications

- Tonifies the Original Qi (*yuan qi*) as well as Qi in general, strengthens the Kidneys (especially Kidney Yang), regulates Qi, harmonises Blood and eliminates Dampness

Special features

Important tonification point for mental and physical exhaustion, strengthens especially the Qi and Yang with moxibustion. One of the most important acupuncture points.

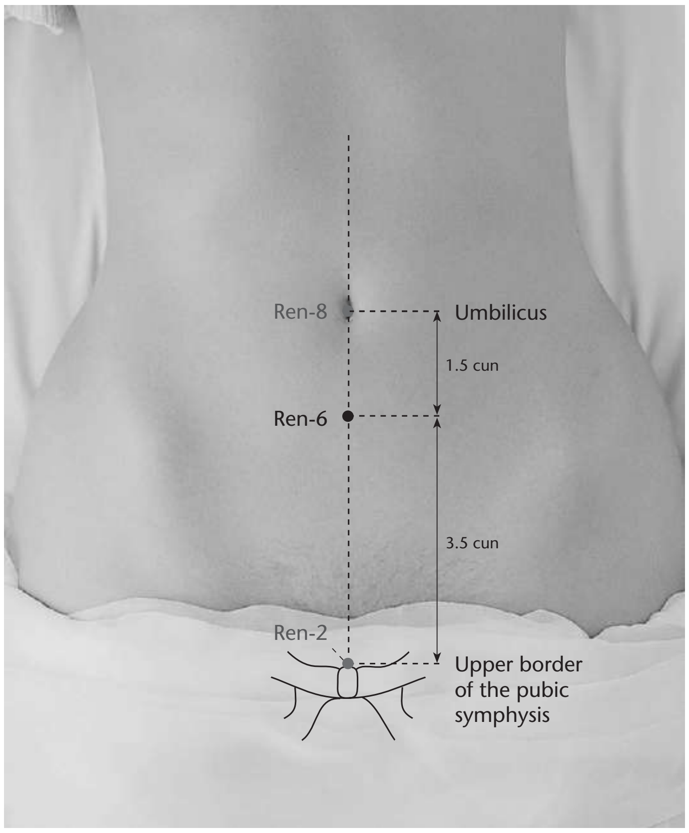
Ren-8
Umbilicus
1.5 cun
Ren-6
3.5 cun
Ren-2
Upper border of the pubic symphysis

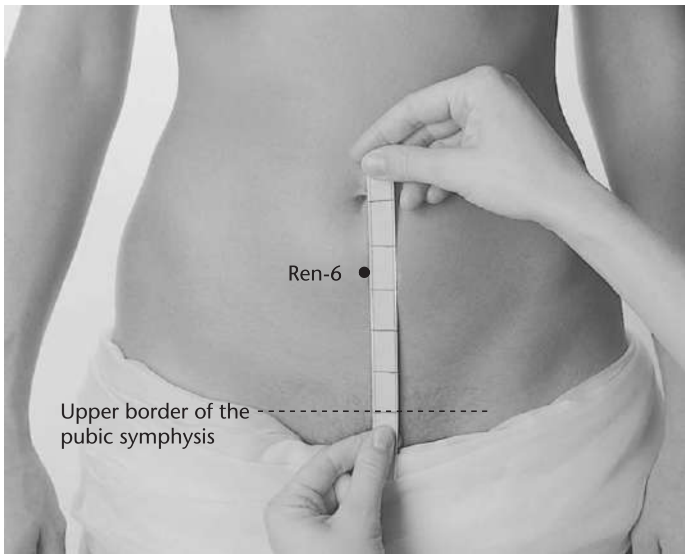
Ren-6
Upper border of the pubic symphysis

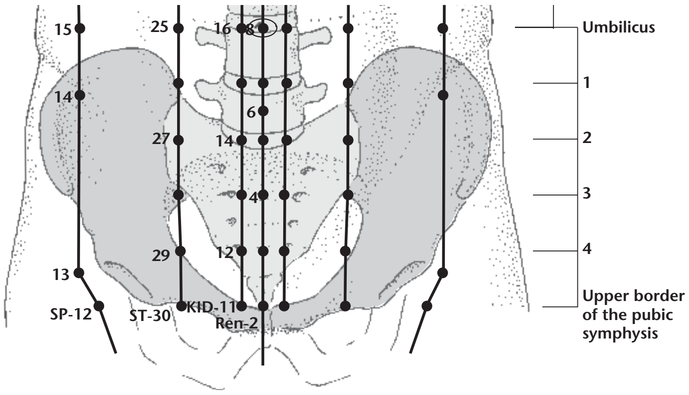
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

Yin Intersection YINJIAO Ren-7

Location

On the anterior midline, 1 cun inferior to the umbilicus.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). **Ren-7** is located on the midline, 1 cun inferior to the umbilicus.

Located on the same level are → **KID-15** (0.5 cun lateral to the midline) and → **ST-26** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates menstruation
- Regulates the Qi flow in the abdomen and the genital area

Special features

Meeting point with the *chong mai* and the KID channel

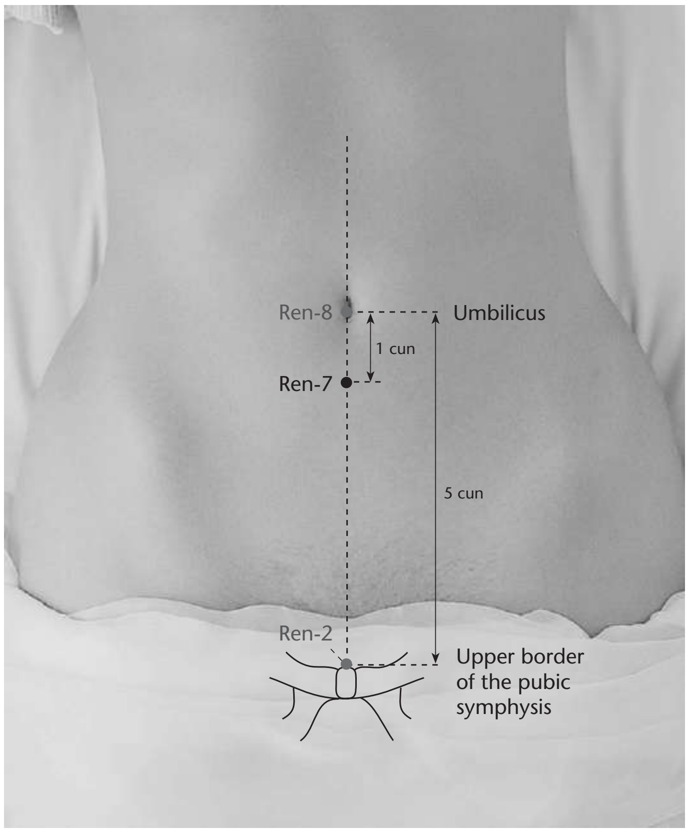
Ren-8
Umbilicus
1 cun
Ren-7
5 cun
Ren-2
Upper border of the pubic symphysis

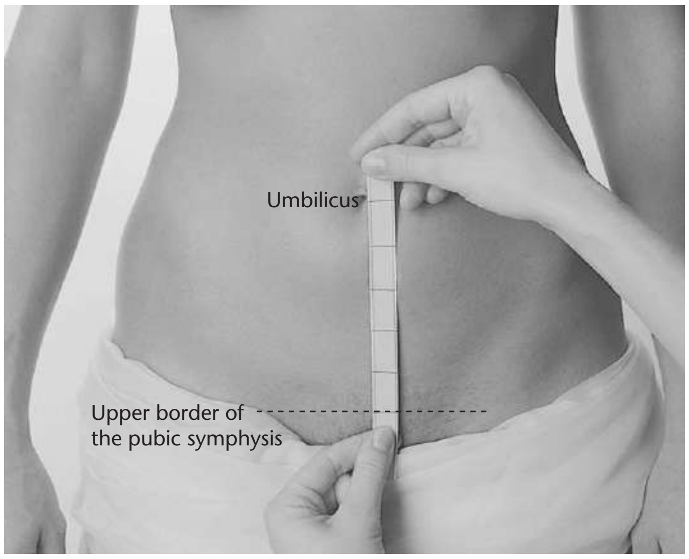
Umbilicus
Upper border of the pubic symphysis

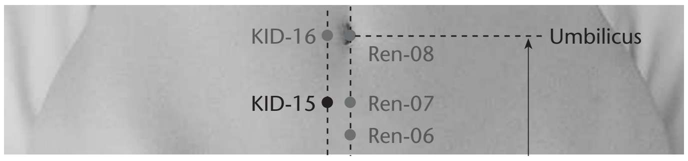
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06

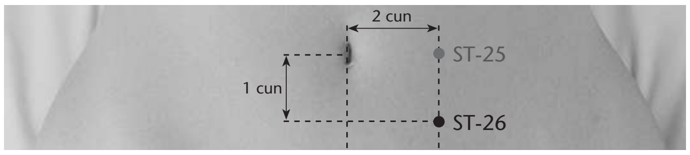
2 cun
ST-25
1 cun
ST-26

Ren-8 *Spirit Gateway* SHENQUE

Location

In the centre of the umbilicus.

How to find

Located on the same level are ➞ **KID-16** (0.5 cun lateral to the midline), ➞ **ST-25** (2 cun lateral to the midline), ➞ **SP-15** (4 cun lateral to the midline) as well as ➞ **G.B.-26** (on a vertical line through the free end of the 11th rib).

Needling

Needling contraindicated. Only indirect moxibustion (on salt or ginger, moxa pole, moxa box).

Actions/Indications

- Warms and stabilises the Yang and the Intestines

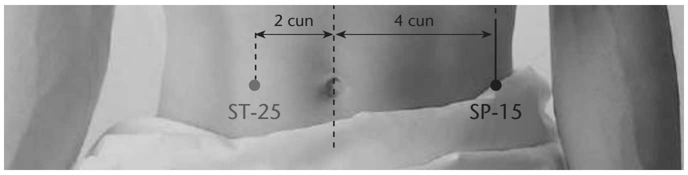
2 cun
4 cun
ST-25
SP-15

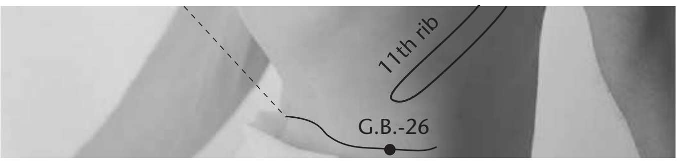
11th rib
G.B.-26

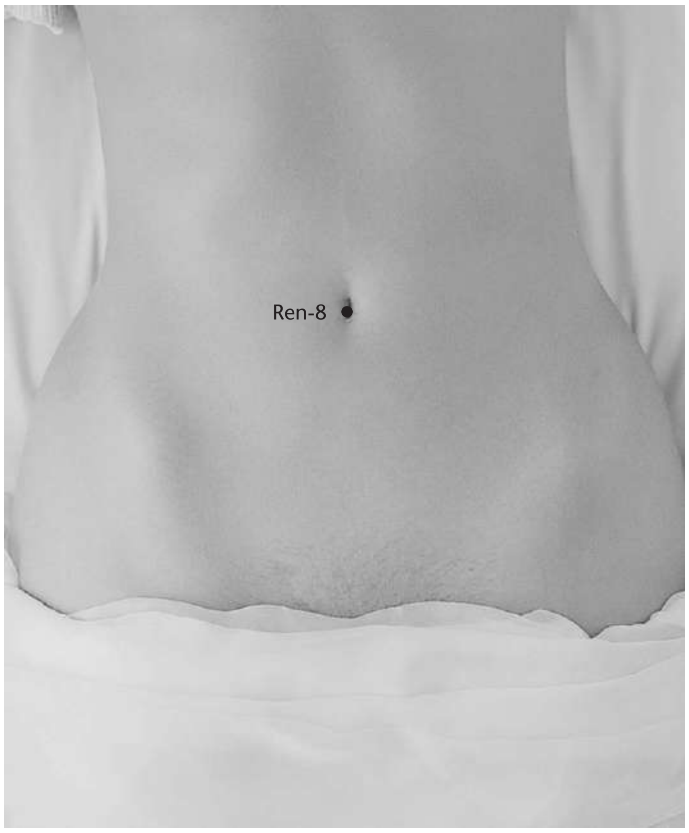

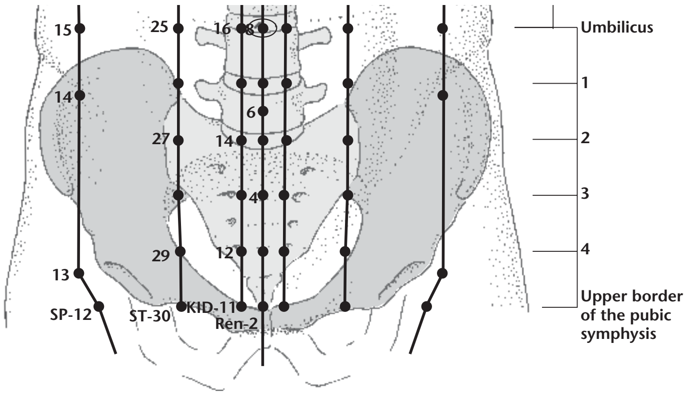
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

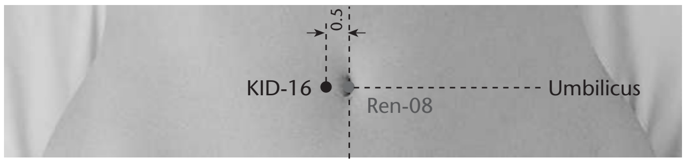
0.5
KID-16
Umbilicus
Ren-08

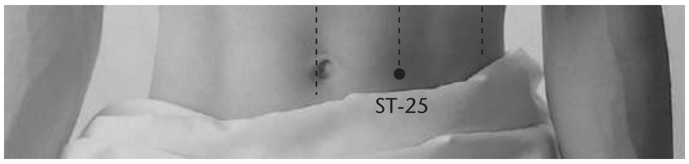

Water Separation SHUIFEN Ren-9

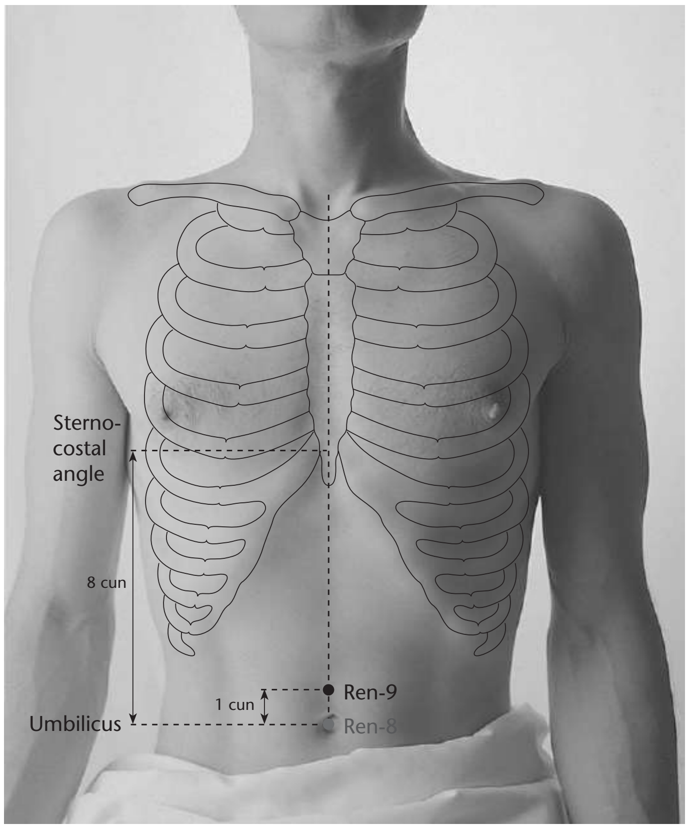
Sterno-
costal
angle
8 cun
1 cun
Ren-9
Umbilicus
Ren-8

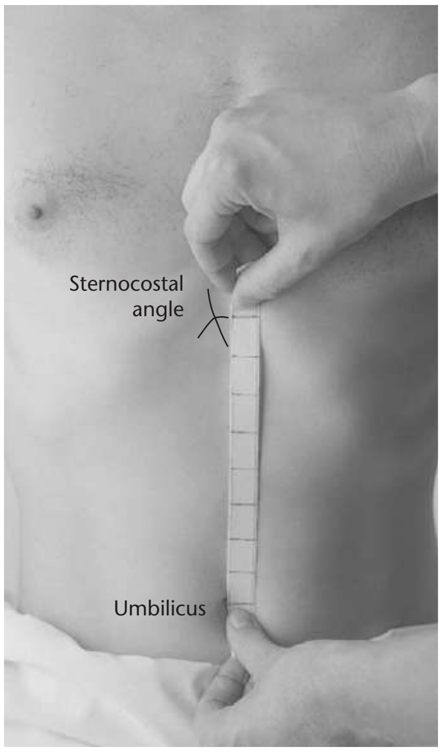
Sternocostal
angle
Umbilicus

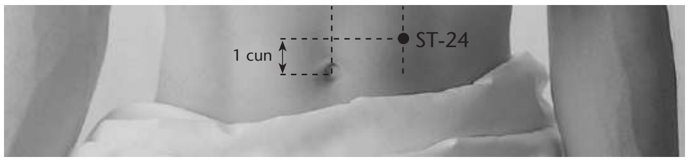

Location

On the anterior midline, 1 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the centre of the umbilicus, measure 1 cun in a superior direction and there locate **Ren-9** on the anterior midline.

Located on the same level is → **ST-24** (2 cun lateral to the midline). → **LIV-13** is also located on approximately this level (at the free end of the 11th rib).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy

Actions/Indications

- Eliminates water accumulations, drains Dampness and controls the water passages Regulates the Intestines and disperses accumulations
- Delayed closure of the fontanelles

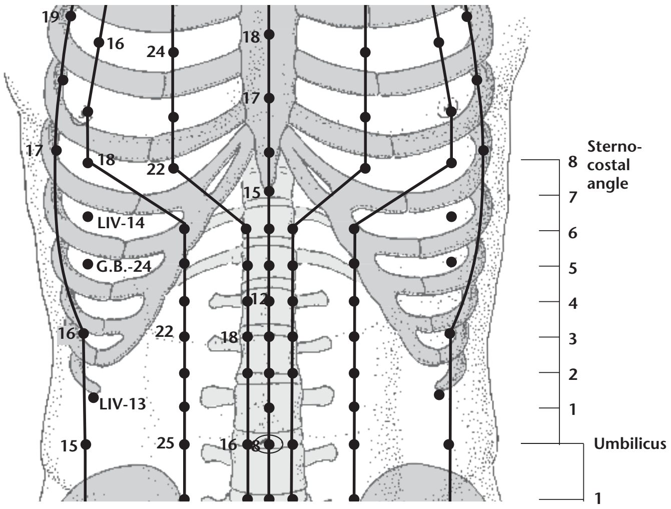
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-
costal
angle
7
6
5
4
3
2
1
Umbilicus
1

Ren-10 *Lower Cavity* XIAWAN

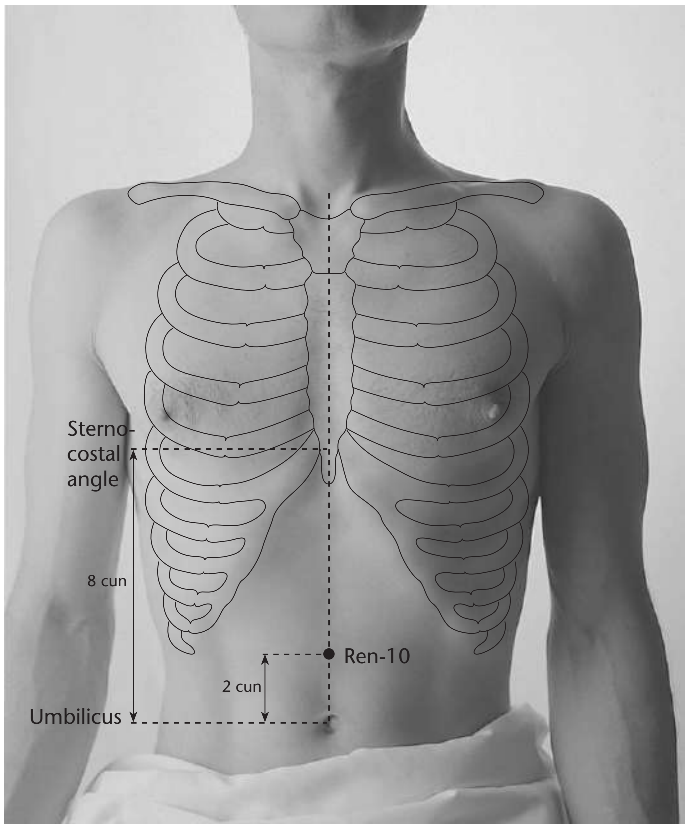
Sterno-
costal
angle
8 cun
Ren-10
2 cun
Umbilicus

Location

On the anterior midline, 2 cun superior to the centre of the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the centre of the umbilicus, measure 2 cun in a superior direction and there locate **Ren-10** on the anterior midline.

Located on the same level are → **KID-17** (0.5 cun lateral to the midline), → **ST-23** (2 cun lateral to the midline) and **Ex-CA** (*weishang*; 4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Tonifies and regulates the Spleen and Stomach Qi, dispels food stagnation

Special features

Meeting point with the SP channel and according to some authors with the ST channel. According to some classical texts, this point is contraindicated during pregnancy (for both needling and moxibustion).

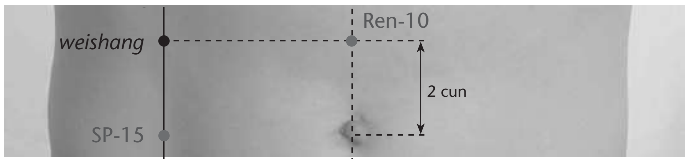
weishang
Ren-10
2 cun
SP-15

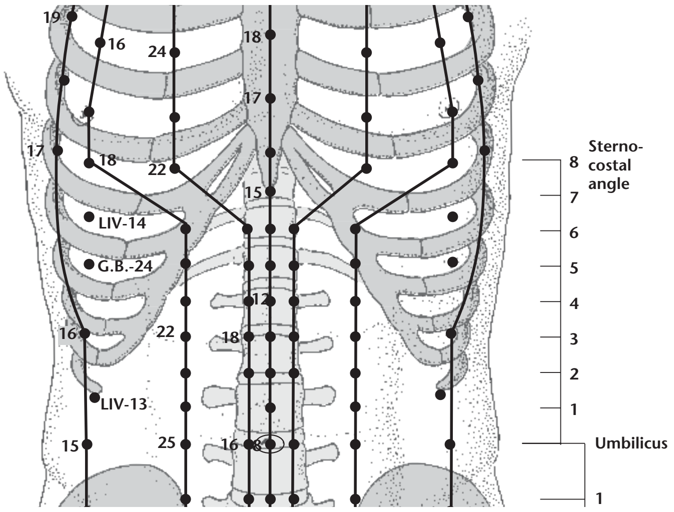
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-
costal
angle
8
7
6
5
4
3
2
1
Umbilicus
1

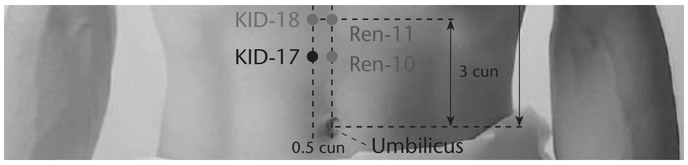
KID-18
Ren-11
KID-17
Ren-10
3 cun
0.5 cun
Umbilicus

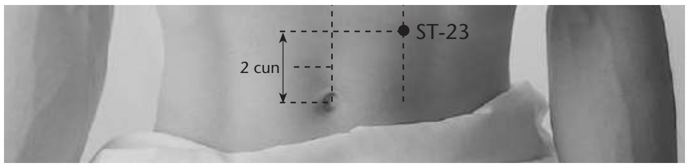

Interior Strengthening JIANLI **Ren-11**

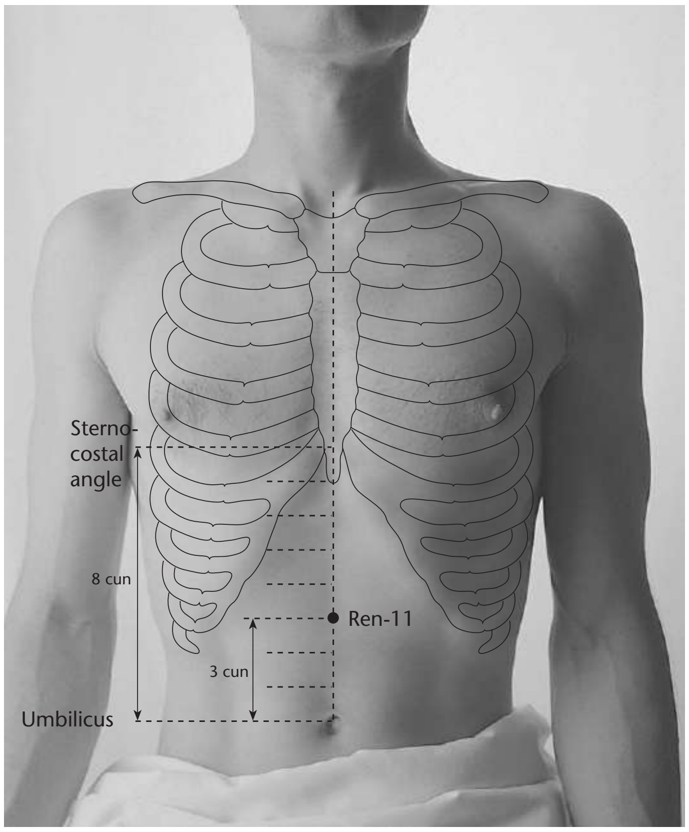
Sterno-
costal
angle
8 cun
Ren-11
3 cun
Umbilicus

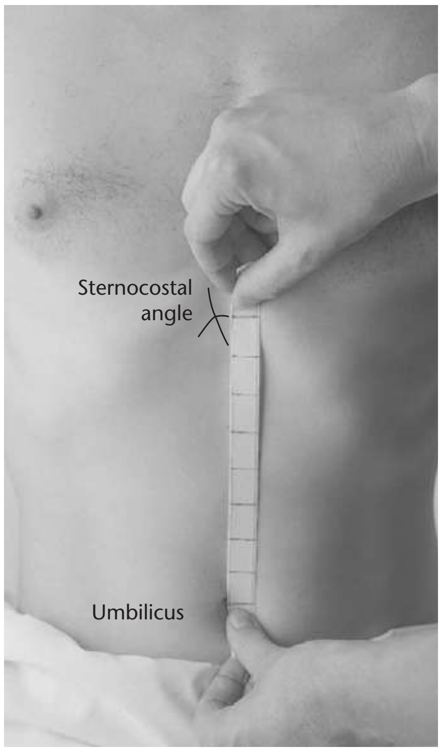
Sternocostal
angle
Umbilicus

Location

On the anterior midline, 3 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the centre of the umbilicus, measure 3 cun in a superior direction and there locate **Ren-11** on the anterior midline. Located on the same level are → **KID-18** (0.5 cun lateral to the midline), → **ST-22** (2 cun lateral to the midline) and → **SP-16** (4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Harmonises the Middle Burner, regulates the Qi flow

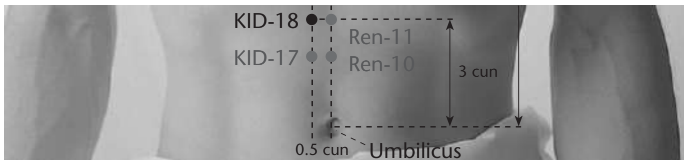
KID-18
KID-17
Ren-11
Ren-10
3 cun
0.5 cun
Umbilicus

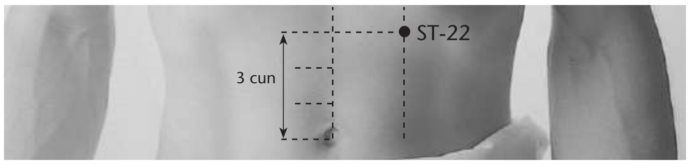

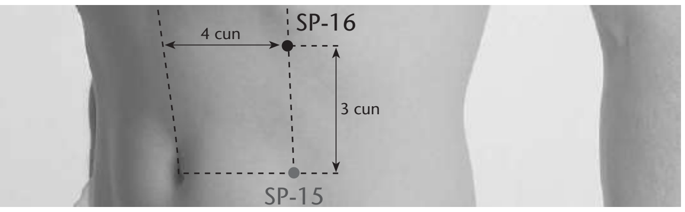
4 cun
SP-16
3 cun
SP-15

Ren-12 *Middle Cavity* ZHONGWAN

Location

On the anterior midline, 4 cun superior to the umbilicus or 4 cun inferior to the sternocostal angle.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus can be divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the centre of the umbilicus, measure either 4 cun in a superior direction or from the sternocostal angle 4 cun in an inferior direction. There, locate **Ren-12** on the anterior midline. **Or:** Spreading hands technique (→ 2.3.3): Place the little fingers on the sternocostal angle and the centre of the umbilicus respectively and determine the midpoint (=**Ren-12**) of this distance by joining the index fingers. Located on the same level are → **KID-19** (0.5 cun lateral to the midline) and → **ST-21** (2 cun lateral to the midline).

Needling

0.8–1.5 cun vertically or obliquely towards surrounding points (→ **ST-21**, **Ren-10**, **Ren-15**). Caution: Peritoneum, pregnancy. If indicated, moxa is recommended.

Actions/Indications

- Harmonises and tonifies the Middle Burner, descends counterflow Qi, regulates the Qi flow and alleviates pain, transforms Dampness

Special features

Front-*mu* point of the stomach, meeting point with the S.I., T.B.* and ST channels, *hui*-meeting point of the *fu*-Organs. Major point for (functional) stomach disorders.

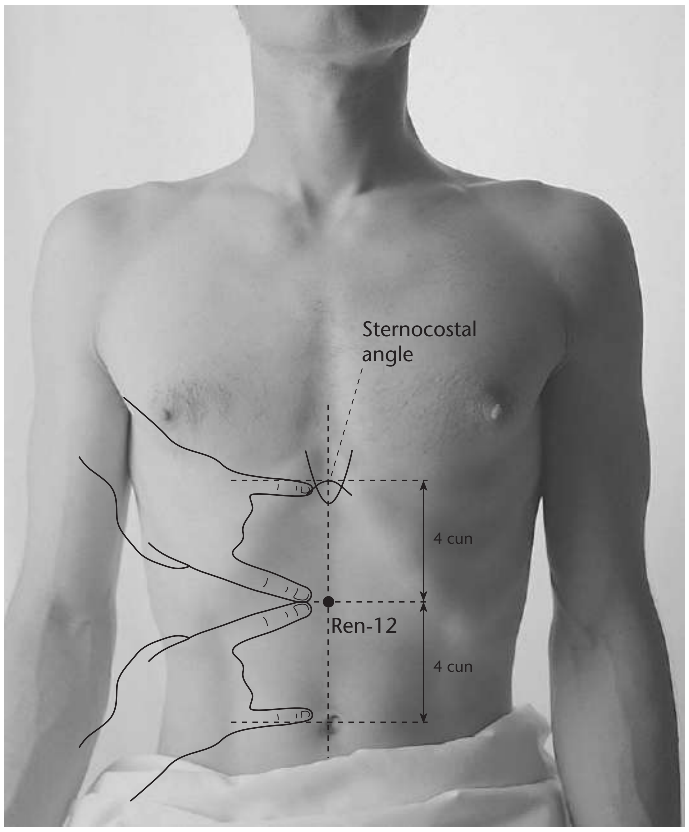
Sternocostal angle
4 cun
Ren-12
4 cun

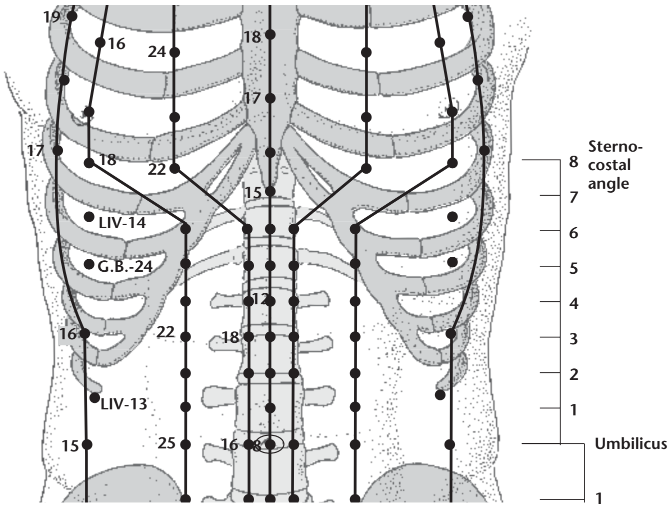
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-costal angle
8
7
6
5
4
3
2
1
Umbilicus
1

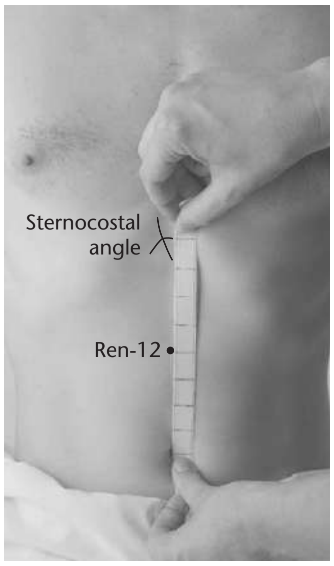
Sternocostal angle
Ren-12

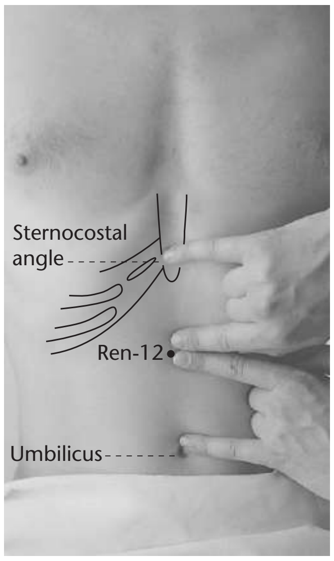
Sternocostal angle
Ren-12
Umbilicus

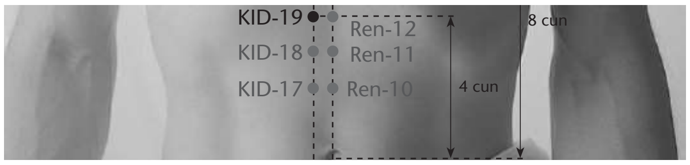
KID-19
KID-18
KID-17
Ren-12
Ren-11
Ren-10
8 cun
4 cun

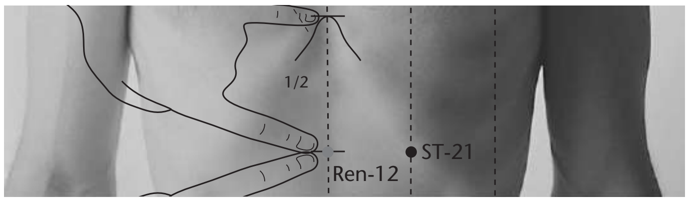
1/2
ST-21
Ren-12

---

* Mentioned by only some authors

Ch05a-F10028.qxd 2/22/08 9:44 PM Page 483

Upper Cavity SHANGWAN **Ren-13**

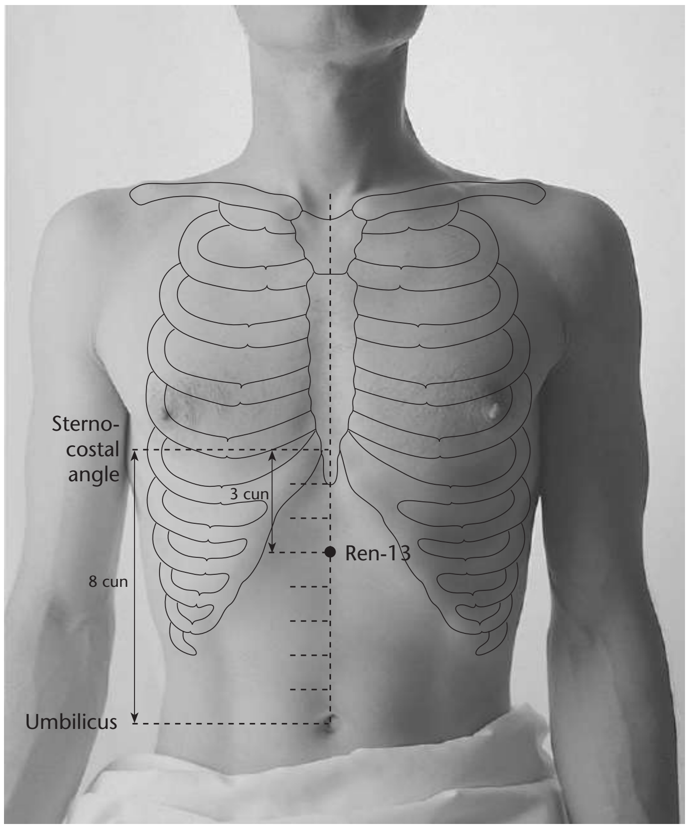
Sterno-
costal
angle
3 cun
Ren-13
8 cun
Umbilicus

Location

On the anterior midline, 3 cun inferior to the sternocostal angle.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the sternocostal angle, measure 3 cun in an inferior direction and there locate **Ren-13** on the anterior midline.

Located on the same level are → **KID-20** (0.5 cun lateral to the midline), → **ST-20** (2 cun lateral to the midline) and → **G.B.-24** (approximately on this level, in the 7th intercostal space, on the mamillary line).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Stomach and counterflow Qi
- Regulates the Heart

Special features

Meeting point with the ST and S.I. channel

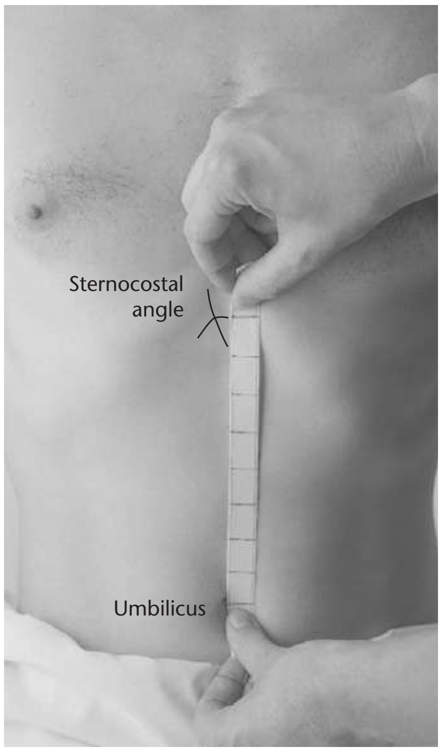
Sternocostal
angle
Umbilicus

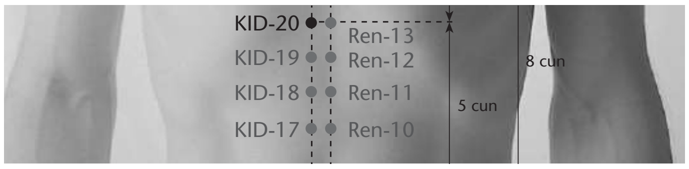
KID-20
KID-19
KID-18
KID-17
Ren-13
Ren-12
Ren-11
Ren-10
5 cun
8 cun

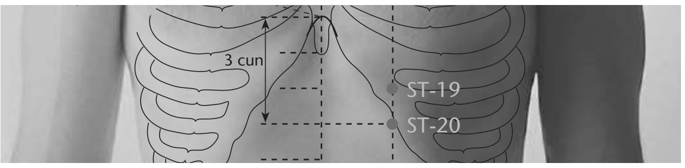
3 cun
ST-19
ST-20

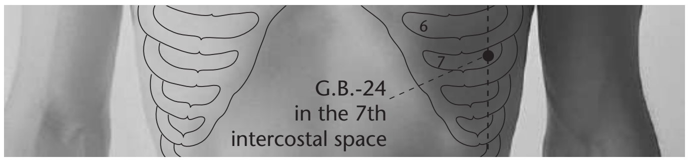
6
7
G.B.-24
in the 7th
intercostal space

Ren-14 *Great Gateway* JUQUE

Location

On the anterior midline, 2 cun inferior to the sternocostal angle or 6 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun (note: these are proportional cun → 2.2). **Ren-14** is located on the midline, 2 cun inferior to the sternocostal angle or 6 cun superior to the umbilicus. **Or:** Use spreading hands technique (→ Chapter 2): Place the little fingers on the sternocostal angle and the umbilicus respectively and determine the midpoint of this distance (→ **Ren-12**). Then locate **Ren-14** on the midpoint of the distance between → **Ren-12** and the sternocostal angle.
Located on the same level are → **KID-21** (0.5 cun lateral to the midline), → **ST-19** (2 cun lateral to the midline) and → **LIV-14** (approximately on this level, in the 6th intercostal space, on the mamillary line).

Needling

0.3–0.8 cun vertically or often obliquely in an inferior direction or up to 1.5 cun transversely (subcutaneously). Depending on individual anatomy, this point can be located on the xiphoid process. Caution: Peritoneum, injury to hypertrophied organs (left: heart, right: liver) depending on the thickness of the abdominal wall.

Actions/Indications

- Regulates the Heart, disperses stagnant Phlegm, alleviates pain, unbinds the thorax, descends the Qi
- Harmonises the Stomach, descends Stomach Qi
- Transforms Phlegm, calms the *shen*

Special features

Front-*mu* point of the Heart. Major local point for thoracic pain caused by Blood stasis or Phlegm blockages.

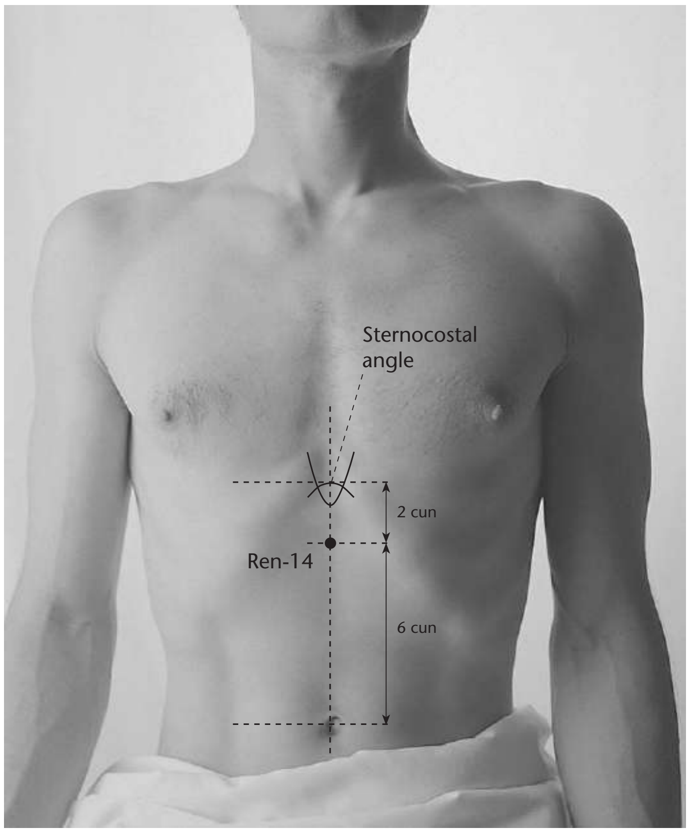
Sternocostal angle
2 cun
Ren-14
6 cun

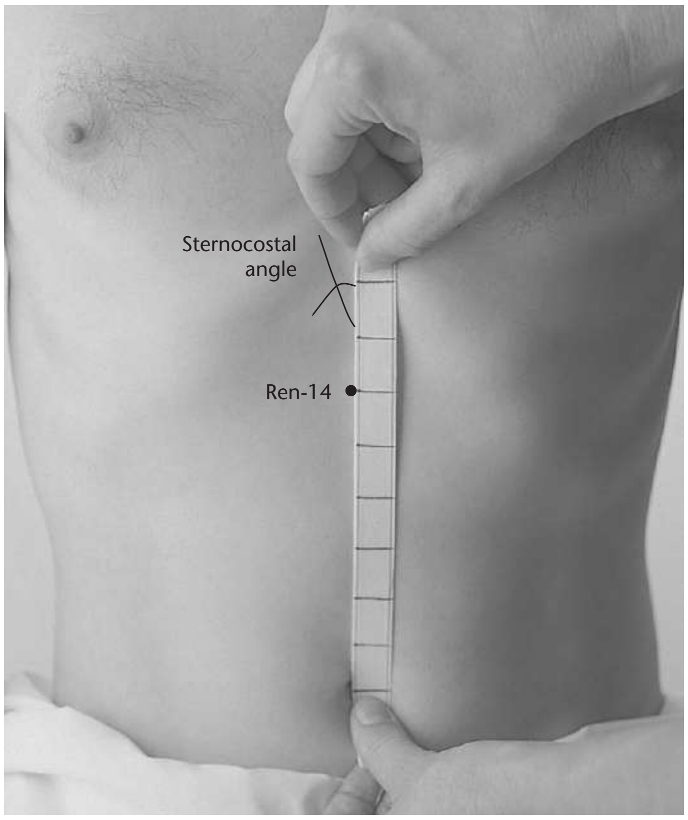
Sternocostal angle
Ren-14

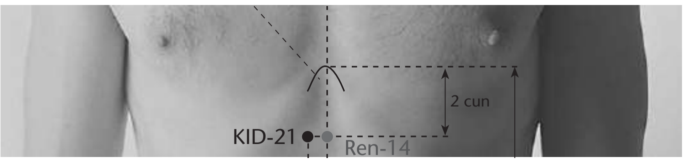
2 cun
KID-21
Ren-14

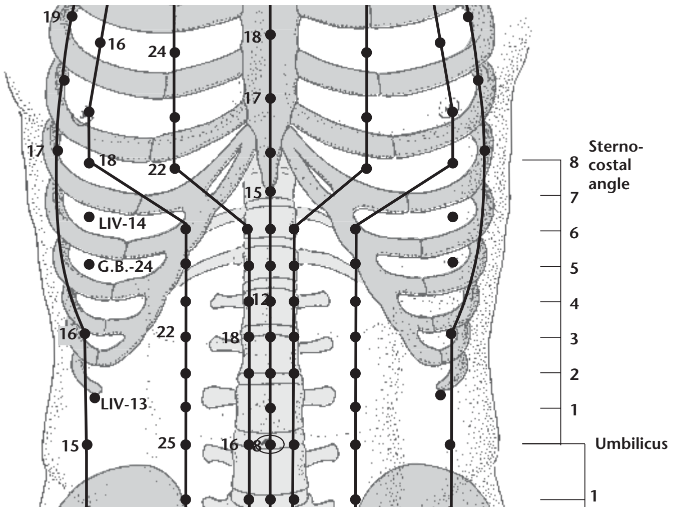
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
8
7
6
5
4
3
2
1
Sterno-
costal
angle
Umbilicus
1

Turtledove Tail **JIUWEI** **Ren-15**

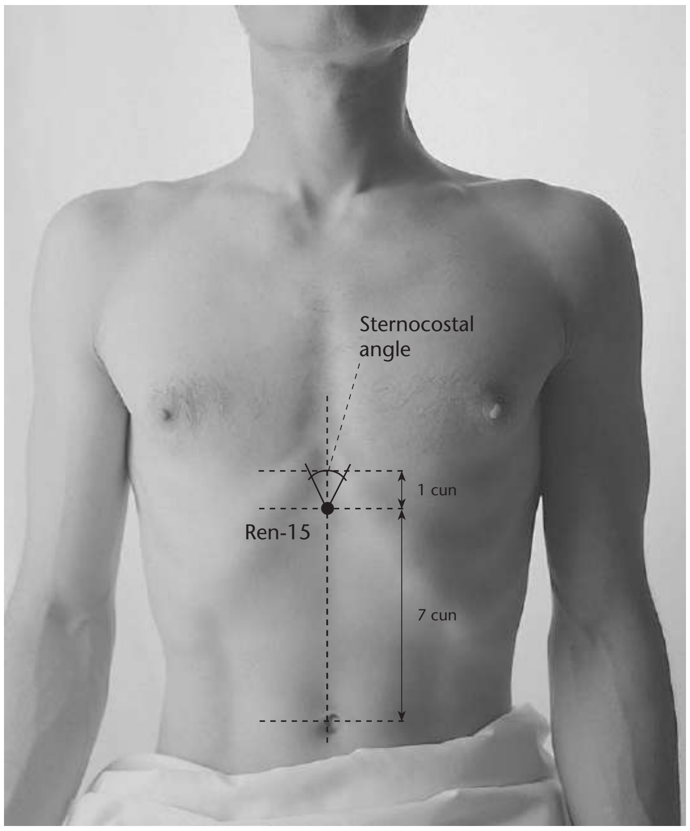
Sternocostal angle
1 cun
Ren-15
7 cun

Location

On the anterior midline, 1 cun inferior to the sternocostal angle or 7 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the sternocostal angle, measure 1 cun in an inferior direction and there locate **Ren-15** on the anterior midline. It is often located directly below or, depending on individual anatomy, on the xiphoid process.

Needling

0.5–1 cun obliquely in an inferior direction. The patient should be in an extended position. Caution: Possible injury to organs with deep needling or oblique insertion in a superior direction, especially in patients with hypertrophied organs (left: Liver, right: Heart). According to some classical texts, moxibustion is contraindicated.

Actions/Indications

- Regulates the Heart, calms the *shen*
- Descends counterflow lung and stomach Qi, unbinds the chest
- Regulates the *luo*-connecting vessel of the *ren mai*

Special features

Luo-connecting point of the *ren mai*. Important point for calming.

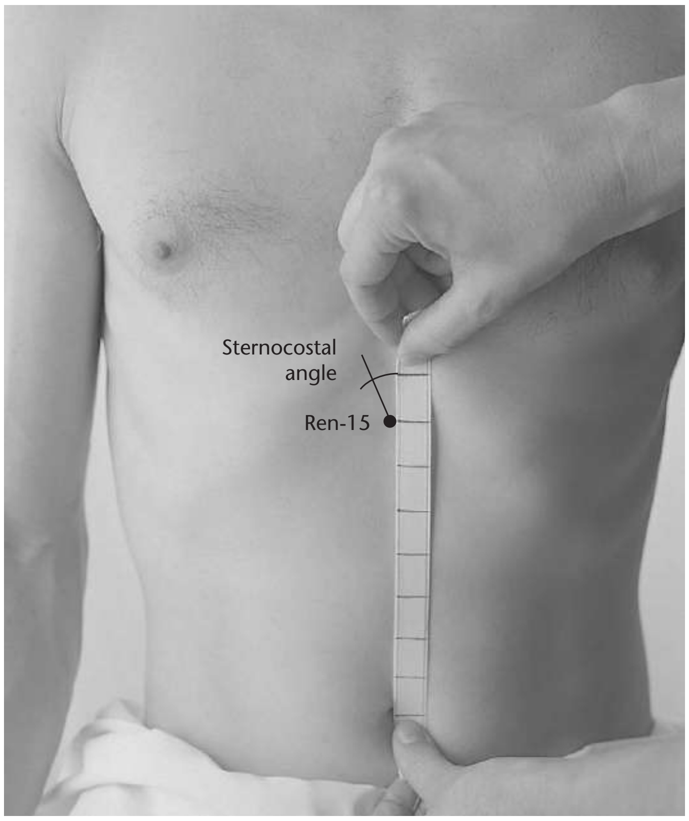
Sternocostal angle
Ren-15

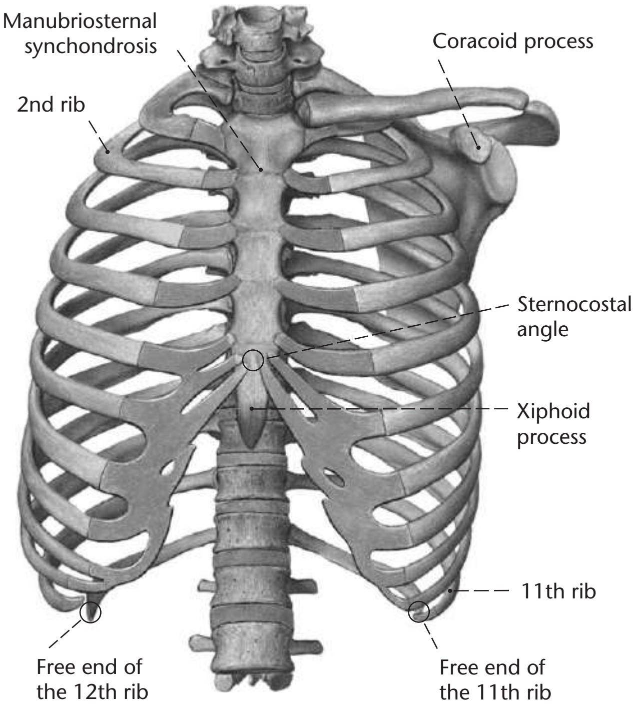
Manubriosternal synchondrosis
Coracoid process
2nd rib
Sternocostal angle
Xiphoid process
11th rib
Free end of the 12th rib
Free end of the 11th rib

Ren-16 *Central Courtyard* ZHONGTING

Location

On the anterior midline, on the level of the sternocostal angle.

How to find

Ren-16 is located on the level of the sternocostal angle, where the two lower borders of the ribcage (→ 3.5) meet. To locate this point, palpate upward along the lower borders of the ribcage with one hand on either side until the palpating fingers meet at the sternocostal angle. Here, locate **Ren-16** in a shallow depression between the sternum and the xiphoid process.

Located on the same level (approximately the 5th intercostal space) are → **KID-22** (2 cun lateral to the midline), → **ST-18** (4 cun lateral to the midline) and → **SP-17** (6 cun lateral to the midline).

Needling

0.5 cun transversely (subcutaneously) in a superior or inferior direction. Caution: Danger of heart injury in patients with a split xiphoid process or a sternal foramen. Therefore, only transverse needling is recommended.

Actions/Indications

- Unbinds the chest
- Regulates the Stomach and counterflow Stomach Qi

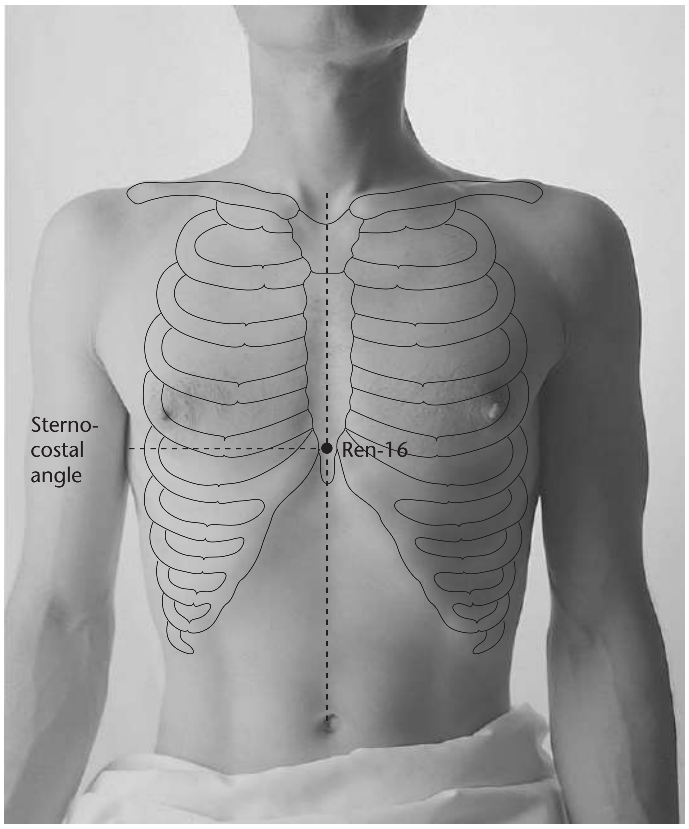
Sterno-
costal
angle
Ren-16

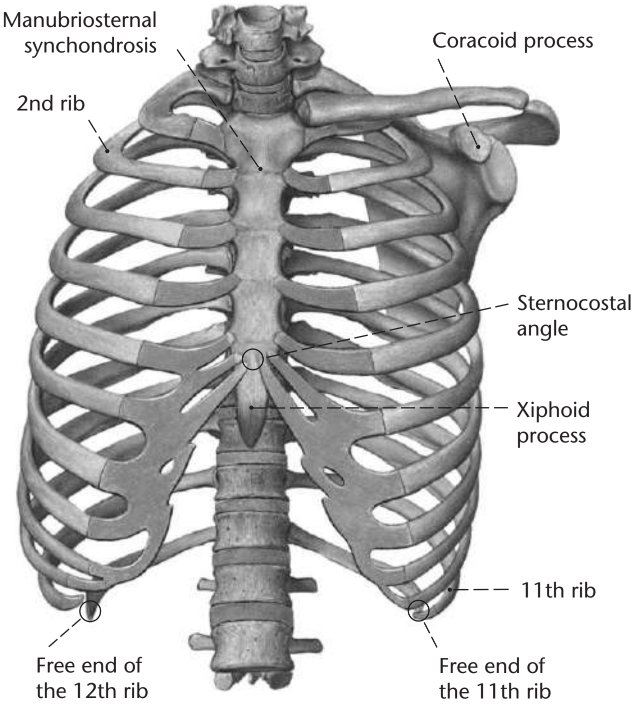
Manubriosternal synchondrosis
2nd rib
Coracoid process
Sternocostal angle
Xiphoid process
11th rib
Free end of the 12th rib
Free end of the 11th rib

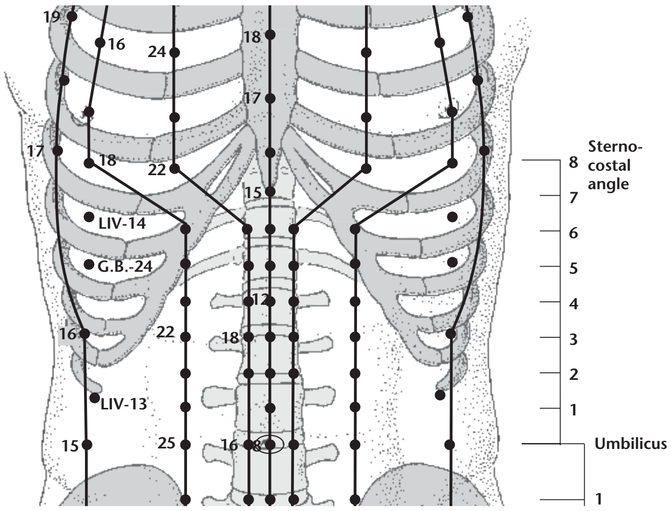
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-costal angle
8
7
6
5
4
3
2
1
Umbilicus
1

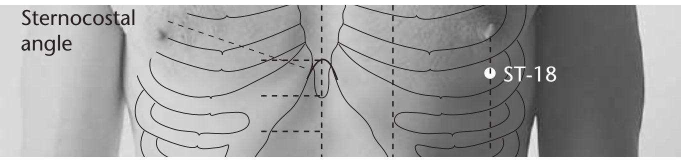
Sternocostal angle
ST-18

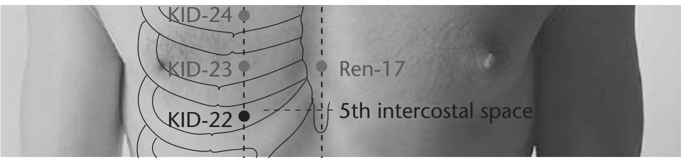
KID-24
KID-23
KID-22
Ren-17
5th intercostal space

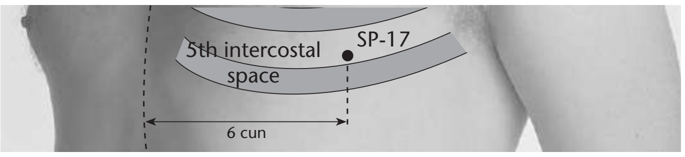
5th intercostal space
SP-17
6 cun

Chest Centre DANZHONG Ren-17

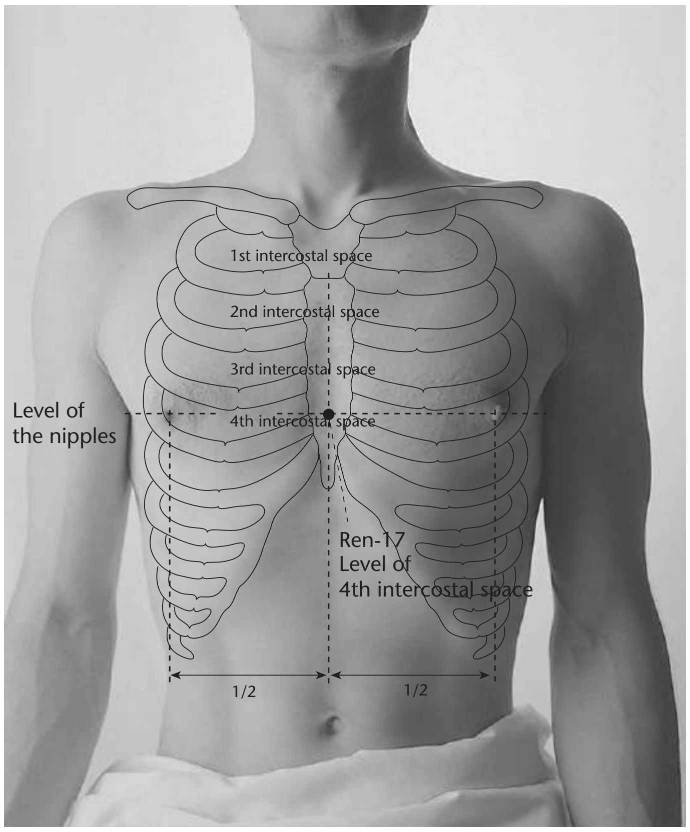
1st intercostal space
2nd intercostal space
3rd intercostal space
Level of the nipples
4th intercostal space
Ren-17
Level of
4th intercostal space
1/2
1/2

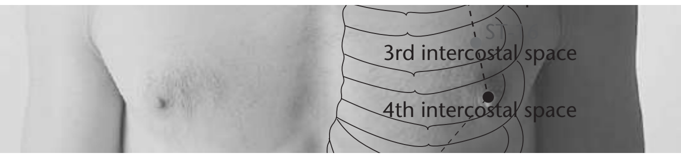
ST
3rd intercostal space
4th intercostal space

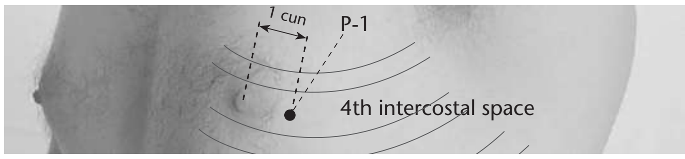
1 cun
P-1
4th intercostal space

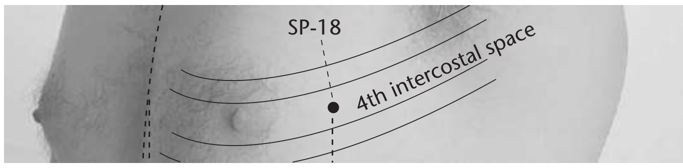
SP-18
4th intercostal space

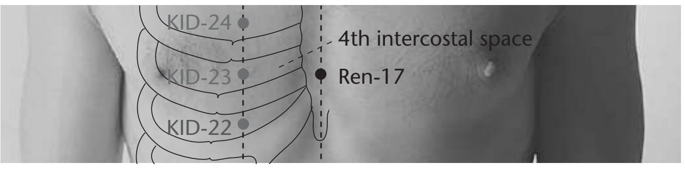
KID-24
4th intercostal space
KID-23
Ren-17
KID-22

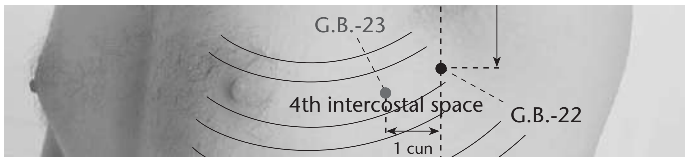
G.B.-23
4th intercostal space
G.B.-22
1 cun

Location

On the anterior midline, on the level of the 4th intercostal space.

How to find

Quick method, especially in men: Locate **Ren-17** on the sternum, between the two nipples. For more detailed orientation on the intercostal region (→ 3.5), first, locate the manubriosternal synchondrosis, which can be felt on the sternum as a well-defined horizontal bony structure on the level of the costal cartilage of the 2nd rib. The 2nd intercostal space is below it. From there, count in an inferior direction to the 4th intercostal space. Locate **Ren-17** on this level on the midline of the sternum. Located on the same level in the 4th intercostal space are **→ KID-23/ST-17/P-1/SP-18/G.B.-22/G.B.-23** (2 cun lateral to the midline/on the nipple/1 cun lateral to the nipple/6 cun lateral to the midline/on the axillary line/1 cun anterior to **→ G.B.-22**).

Needling

Depending on the indication, 0.5–1 cun transversely (subcutaneously) in a superior or inferior direction or towards the nipples (for mamillary disorders). Caution: Intracardiac needling, pneumothorax. The sternum can be very thin in this area owing to embryonic ossification disorders! Moxibustion is recommended in order to tonify the Qi, but caution with asthmatic patients.

Actions/Indications

- Regulates and benefits the Qi, unbinds the chest, descends counterflow Lung and Stomach Qi
- Benefits the breasts

Special features

Front-*mu* point of the Pericardium, meeting point with the SP, KID, S.I. and T.B. channels. *Hui*-meeting point of the Qi. Sea of Qi. Important point for respiratory disorders.

Ren-18 *Jade Hall* YUTANG

Location

On the anterior midline, on the level of the 3rd intercostal space.

How to find

Locate the 3rd intercostal space either by parasternal palpation starting at the clavicle or by counting downward from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5).

Located on the same level (→ 3rd intercostal space) are → **KID-24/ST-16/SP-19** (2/4/6 cun lateral to the midline).

Needling

0.5–1 cun transversely (subcutaneously) in a superior or inferior direction. Caution: Danger of injury to the heart or the pleura in patients with a sternal foramen, therefore only transverse needling with or against the flow of the channel.

Actions/Indications

- Unbinds the chest, regulates and descends the Qi

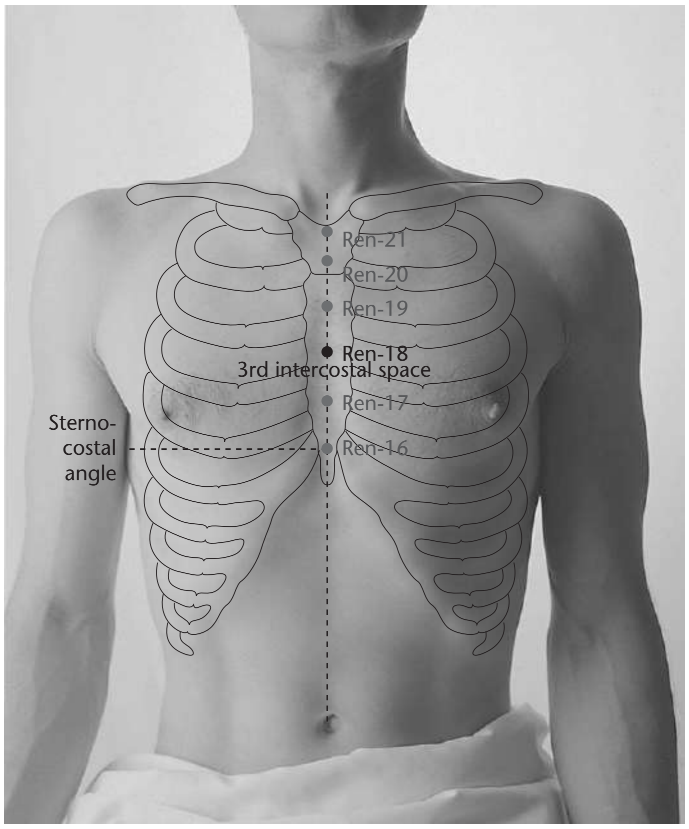
Ren-21
Ren-20
Ren-19
Ren-18
3rd intercostal space
Ren-17
Sterno-
costal
angle
Ren-16

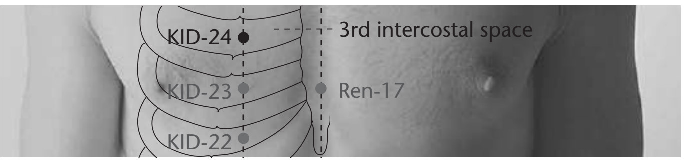
KID-24
3rd intercostal space
KID-23
Ren-17
KID-22

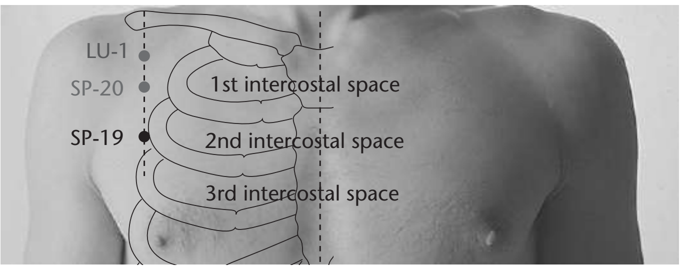
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

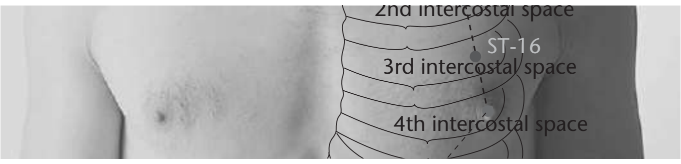
2nd intercostal space
ST-16
3rd intercostal space
4th intercostal space

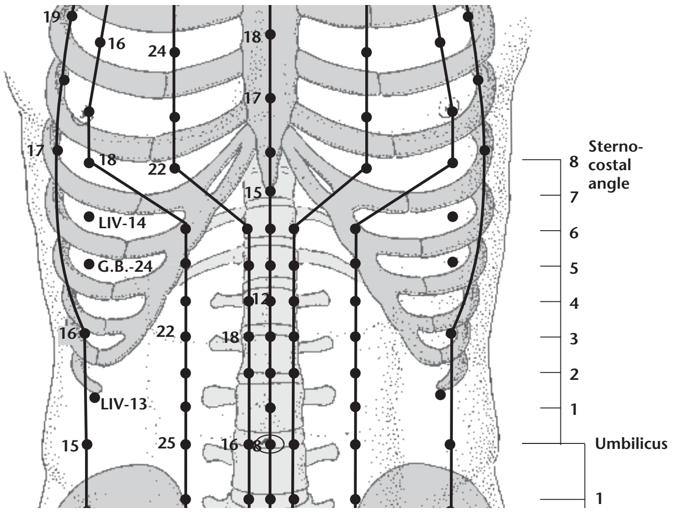
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-
costal
angle
8
7
6
5
4
3
2
1
Umbilicus
1

Purple Palace **ZIGONG** **Ren-19**

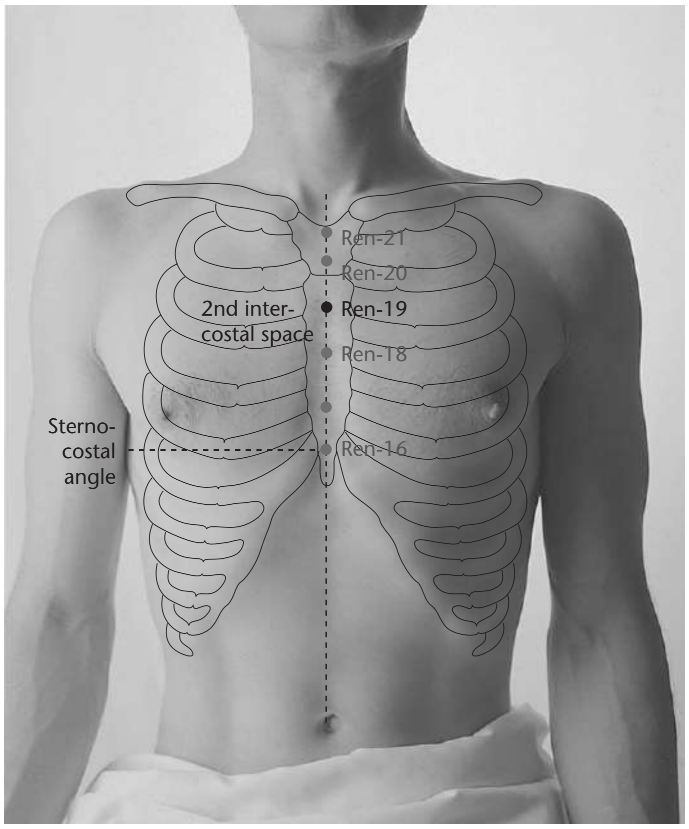
Ren-21
Ren-20
2nd inter-
costal space
Ren-19
Ren-18
Sterno-
costal
angle
Ren-16

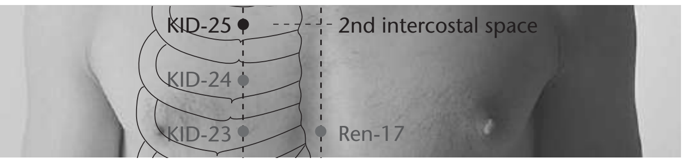
KID-25
2nd intercostal space
KID-24
KID-23
Ren-17

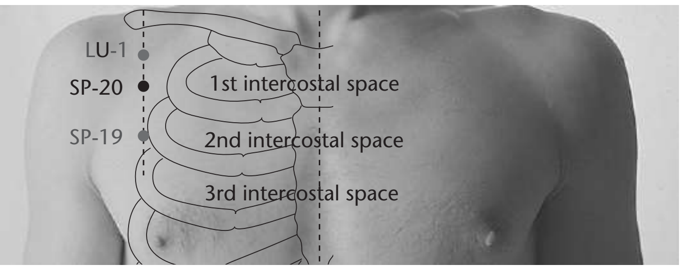
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

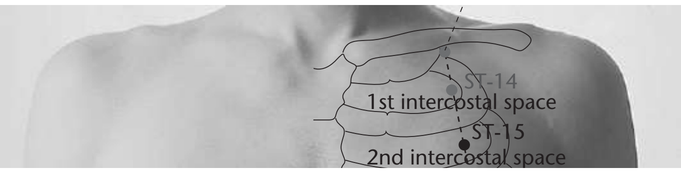
ST-14
1st intercostal space
ST-15
2nd intercostal space

Location

On the anterior midline, on the level of the 2nd intercostal space.

How to find

Locate the 2nd intercostal space by parasternal palpation starting at the clavicle or by palpating from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5).

Located at the same level (2nd intercostal space) are → **KID-25/ST-15/SP-20** (2/4/6 cun lateral to the midline).

Needling

0.5 cun transversely in a superior or inferior direction. Caution: In patients with a sternal foramen, danger of injury to the pleura or the mediastinum, therefore only transverse needling with or against the flow of the channel.

Actions/Indications

- Unbinds the chest
- Regulates and descends the Qi

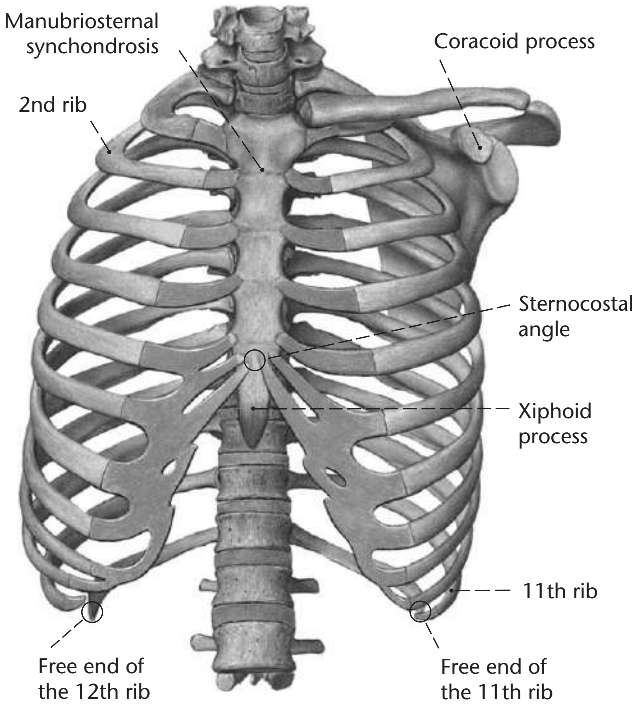
Manubriosternal
synchondrosis
Coracoid process
2nd rib
Sternocostal
angle
Xiphoid
process
11th rib
Free end of
the 12th rib
Free end of
the 11th rib

Ren-20 *Magnificent Canopy* HUAGAI

Location

On the anterior midline, on the lower part of the manubrium sterni, on the level of the 1st intercostal space.

How to find

The manubriosternal synchondrosis generally forms a distinct horizontal bony structure on the upper part of the sternum. Locate **Ren-20** slightly superior to the synchondrosis, on the midline and on the level of the 1st intercostal space.

Located at the same level (1st intercostal space) are → **KID-26/ST-14/LU-1** (2/4/6 cun lateral to the midline).

Needling

0.5 cun transversely in a superior or inferior direction. Caution: In patients with a sternal foramen, danger of injury to the pleura or the mediastinum, therefore only transverse needling with or against the flow of the channel.

Actions/Indications

- Unbinds the chest
- Regulates and descends the Qi

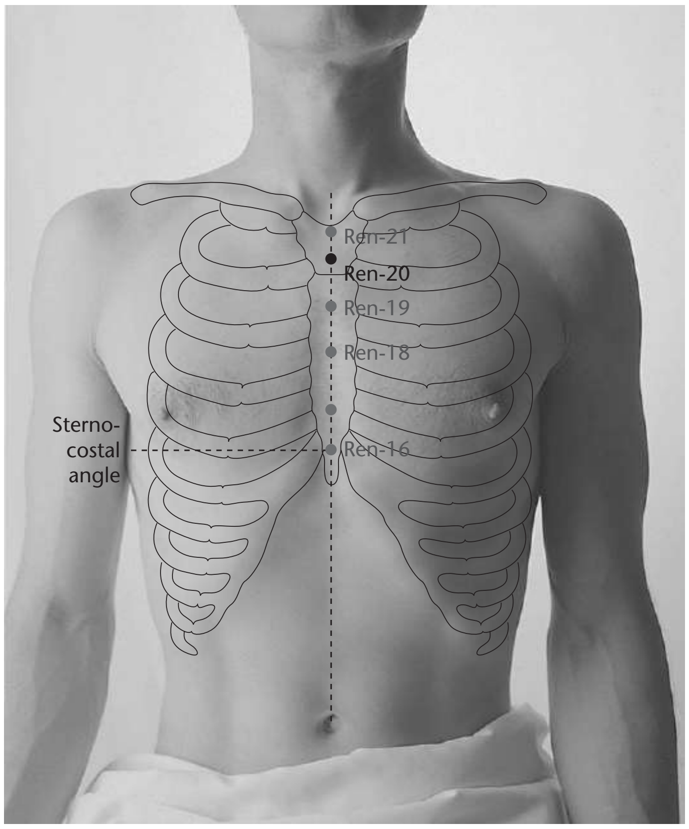
Ren-21
Ren-20
Ren-19
Ren-18
Sterno-costal angle
Ren-16

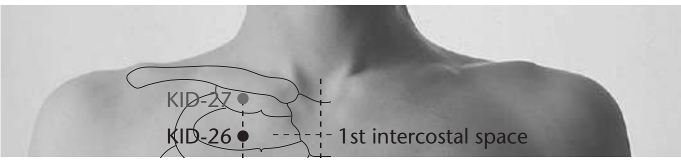
KID-27
KID-26
1st intercostal space

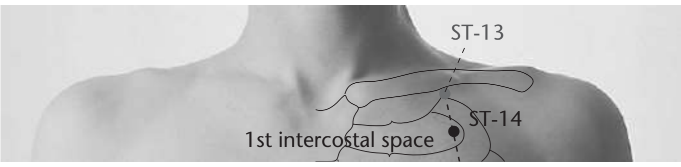
ST-13
ST-14
1st intercostal space

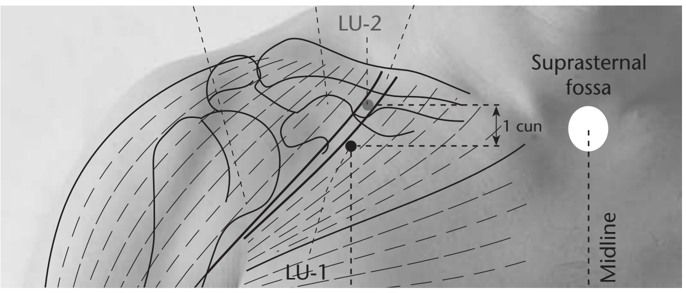
LU-2
Suprasternal fossa
1 cun
LU-1
Midline

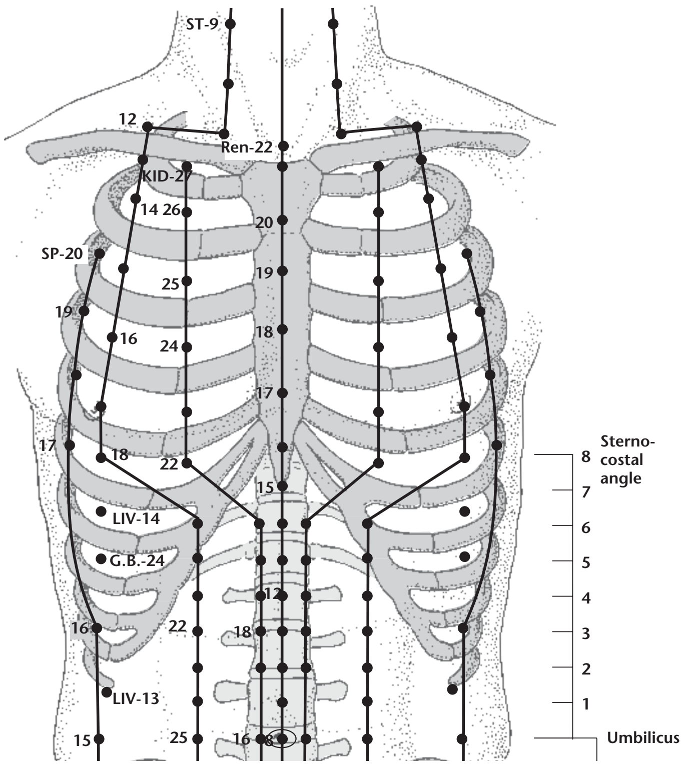
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-costal angle
8
7
6
5
4
3
2
1
Umbilicus

Jade Pivot XUANJI **Ren-21**

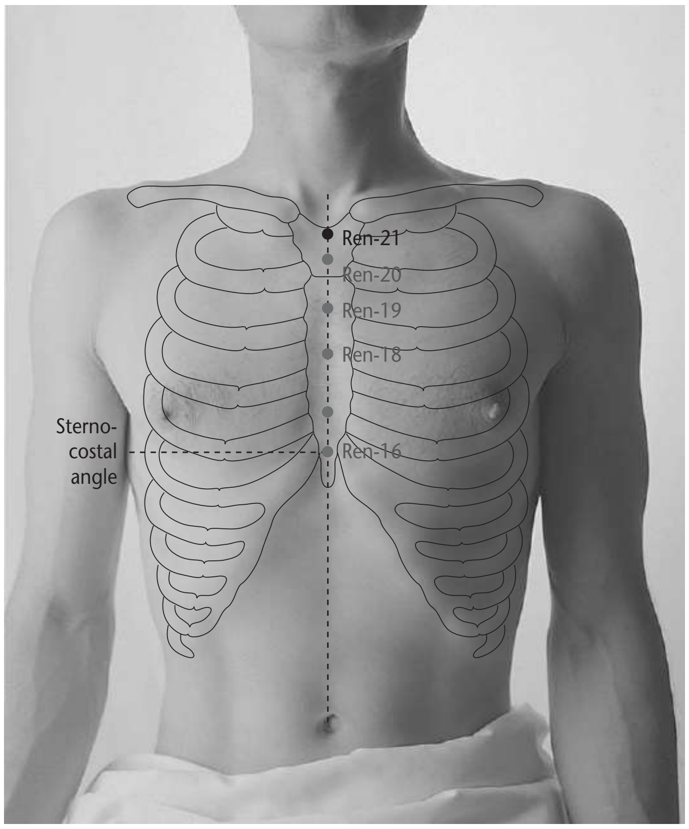
Ren-21
Ren-20
Ren-19
Ren-18
Sterno-
costal
angle
Ren-16

Location

On the anterior midline, below the upper border of the manubrium sterni.

How to find

The upper border of the manubrium sterni forms the lower border of the suprasternal fossa. Locate **Ren-21** on the midline, inferior to the bony margin of the sternum.

Located on approximately the same level (below the clavicle) are ➞ **KID-27/ST-13/LU-2** (2/4/6 cun lateral to the midline).

Needling

0.5 cun transversely in a superior or inferior direction. Caution: In patients with a sternal foramen, danger of injury to the pleura or the mediastinum, therefore only transverse needling with or against the flow of the channel.

Actions/Indications

- Unbinds the chest
- Benefits the throat
- Descends counterflow Qi

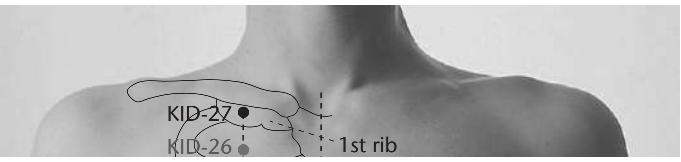
KID-27
KID-26
1st rib

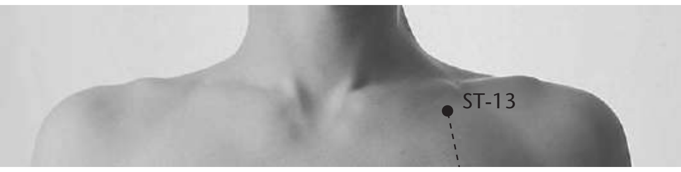

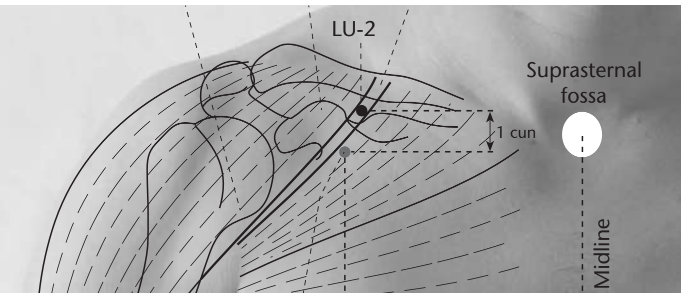
LU-2
Suprasternal
fossa
1 cun
Midline

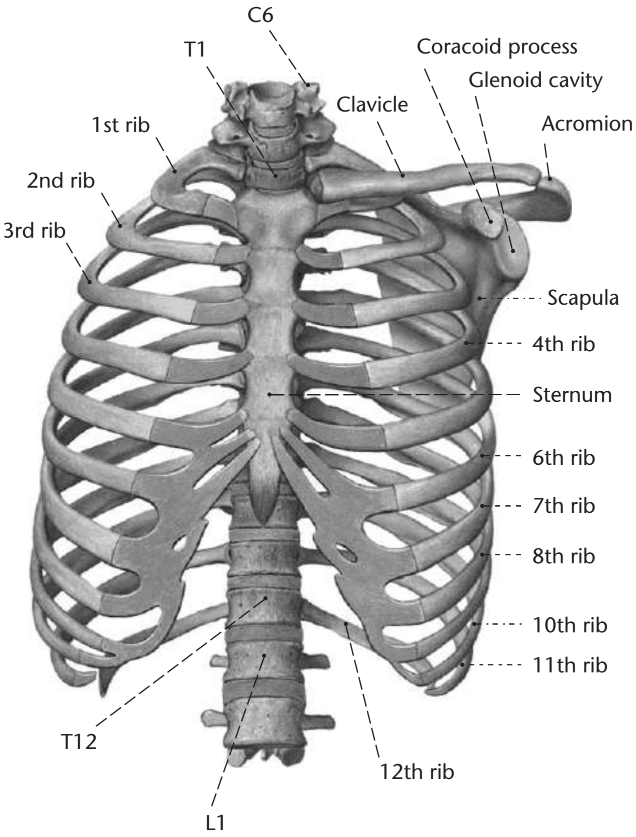
C6
T1
Coracoid process
Glenoid cavity
Clavicle
Acromion
1st rib
2nd rib
3rd rib
Scapula
4th rib
Sternum
6th rib
7th rib
8th rib
10th rib
11th rib
T12
12th rib
L1

Ren-22 *Heavenly Prominence* TIANTU

Location

0.5 cun superior to the sternum, in the centre of the suprasternal fossa.

How to find

This point is best located with the patient in a supine position and a pillow under the shoulders or seated with a comfortable and safe head support. Locate the centre of the suprasternal fossa, approximately 0.5 cun superior to the sternum and there locate **Ren-22**.

Needling

Initial vertical insertion 0.2 cun (in order to penetrate the skin), then direct the tip of the needle inferiorly and continue to insert the needle parallel to the posterior aspect of the sternum up to 0.5–1 cun. Caution: Dangerous point. Wrong insertion technique (for example, vertical or oblique retrosternal needling) may lead to injury of the big vessels and organs inside the mediastinum. This point should only be needled by experienced practitioners.

Actions/Indications

- Descends counterflow Lung Qi, alleviates coughing and dyspnoea, benefits the throat

Special features

Meeting point with the *yin wei mai*, Window of Heaven point. Important emergency point for acute asthma attacks.

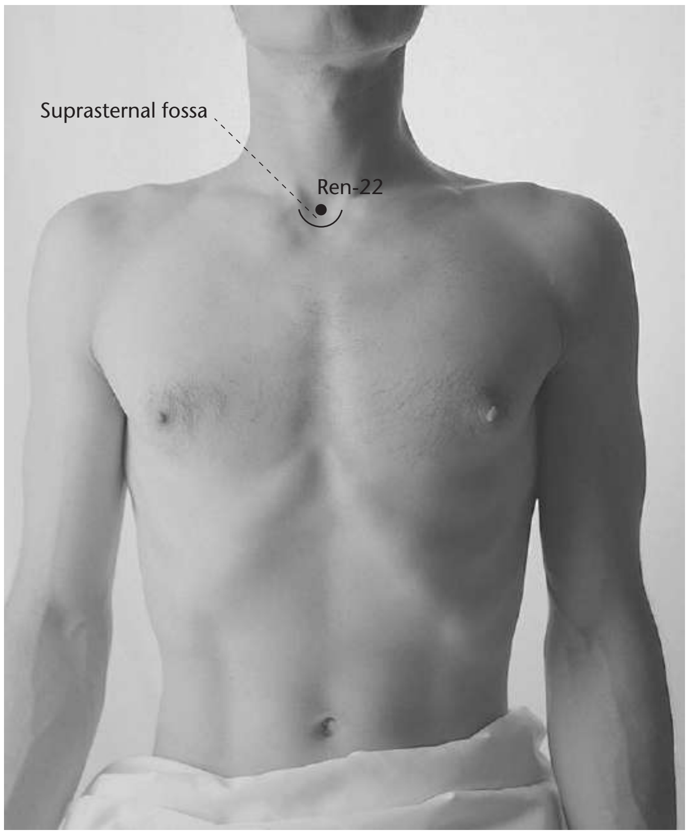
Suprasternal fossa
Ren-22

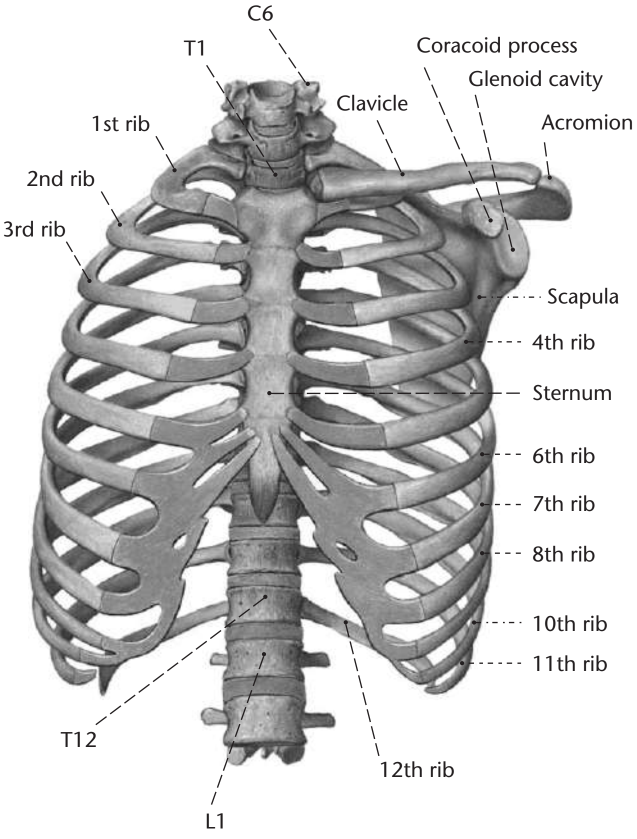
C6
T1
Coracoid process
Glenoid cavity
Clavicle
Acromion
1st rib
2nd rib
3rd rib
Scapula
4th rib
Sternum
6th rib
7th rib
8th rib
10th rib
11th rib
T12
12th rib
L1

Corner Spring LIANQUAN Ren-23

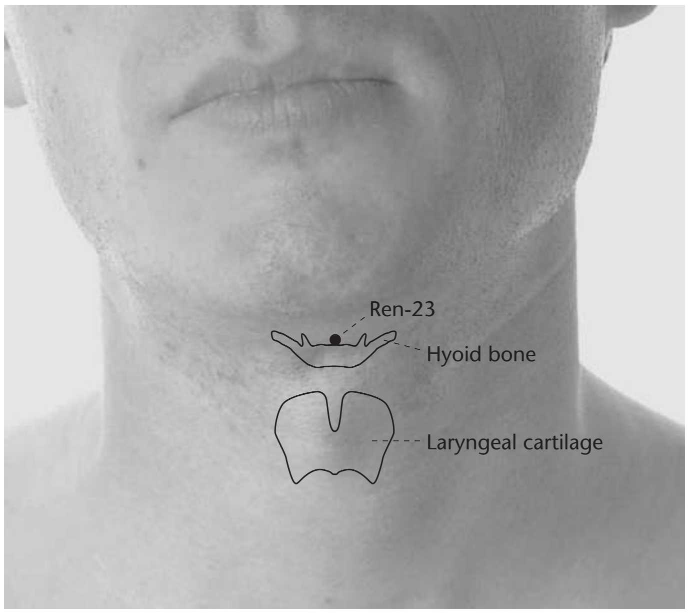
Ren-23
Hyoid bone
Laryngeal cartilage

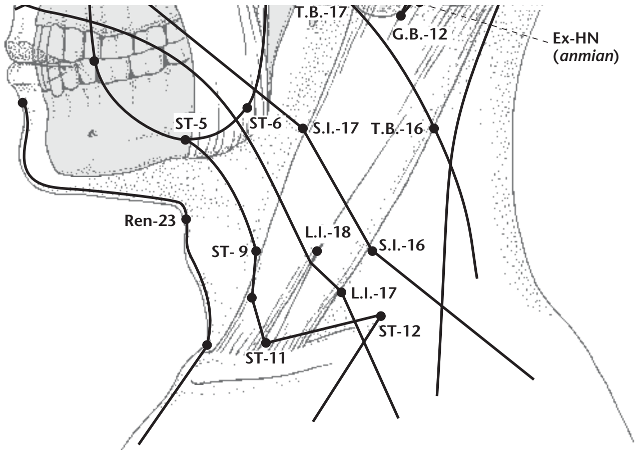
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST- 9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

Location

On the anterior midline, superior to the upper border of the hyoid bone.

How to find

The hyoid bone can be palpated as a delicate bony structure slightly superior to the upper border of the laryngeal prominence. The patient should not recline their head too much, as this will cause the soft tissue to become stretched, making palpation of the bone difficult. **Ren-23** is located on the midline, on the upper border of the hyoid bone and at the junction of the vertical aspect of the neck to the horizontal floor of the mouth (with normal tissue distribution).

Needling

0.3–1 cun obliquely in a superior direction towards the vertex (➞ **Du-20**). Moxibustion is contraindicated according to some texts.

Actions/Indications

- Benefits the tongue
- Benefits the throat, eliminates Phlegm and Heat
- Strengthens the voice

Special features

Meeting point with the *yin wei mai*

Ren-24 *Container of Fluids* CHENGJIANG

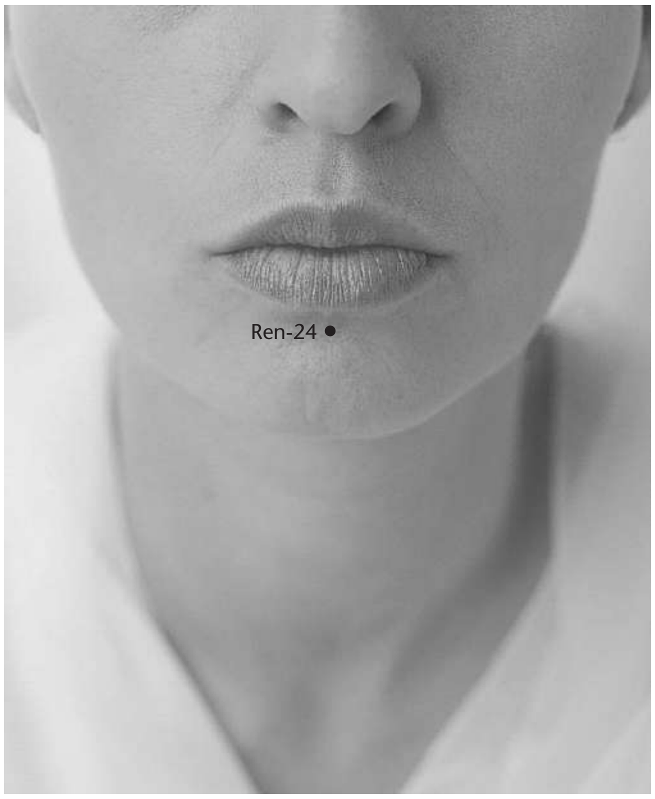

Location

On the anterior midline, below the lower lip, in the mentolabial groove.

How to find

On the chin, locate the mentolabial groove (→ 3.1.3), which forms a horizontal depression at the transition from the chin to the lower lip. Locate **Ren-24** in this groove, on the midline.

Needling

0.2–0.3 cun obliquely in a superior direction.

Actions/Indications

- Eliminates (external) Wind, opens the *luo*-connecting vessels, alleviates pain and swellings, regulates the *ren mai*
- As Sun Si Miao Ghost point for epilepsy and mania

Special features

Meeting point with the *du mai*, ST and L.I. channels, Sun Si Miao Ghost point. Important local point.

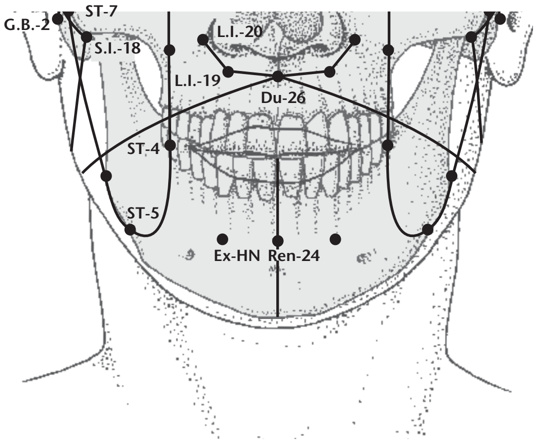

5.3 *Du mai*

Synonym: Governing Vessel

Relationships (→ 1.7.3)

- **Yin/Yang:** *ren mai*/*du mai*
- **Central/peripheral:** *du mai*/*yang qiao mai*. Regions supported by this pairing: inner canthus of the eye, occiput, shoulders, back, S.I. and BL channels, opening point: **S.I.-3** (*houxi*), coupled point: **BL-62** (*shenmai*)

Pathway

The extraordinary vessel *du mai* originates like the *chong mai* and *ren mai* in the lower abdomen (*bao zhong*), according to some authors at the Kidneys, emerges at the perineum and divides into several branches:

➡ The **main pathway** of the *du mai* begins at → **Ren-1** (*huiyin*) at the perineum, reaches **Du-1** (*changqiang*) between the anus and the coccyx and ascends the posterior midline along the spine. At **Du-12** (*shenzhu*), below the spinous process of T3, a branch separates on either side to reach → **BL-12** (*fengmen*), rejoining the main branch at **Du-13** (*taodao*) below the spinous process of T1. It then continues to **Du-16** (*fengfu*) on the lower border of the occipital bone. Here, an **internal branch** enters the Brain. The external pathway continues to ascend the head to **Du-20** (*baihui*), from there beginning its descent along the midline to the forehead and the nose to **Du-26** (*renzhong*) where it crosses the ST and L.I. channels. It terminates inside the mouth below the frenulum at **Du-28** (*yinjiao*) where it meets the ST channel and the extraordinary vessel *ren mai*.

➡ **one branch** of the *du mai* (the spinal branch) originates at → **Ren-1** (*huiyin*), courses to the tip of the coccyx where it meets the Kidney primary channel, ascends **within** the spinal canal to the space between L2 and L3 and enters the Kidneys.

➡ **The second branch** of the *du mai* (the abdominal branch) also originates at → **Ren-1** (*huiyin*), encircles the external genitalia and ascends to the umbilicus, passes through the Heart (*xin*), continues to the throat, circles the lips, travels to the cheek, passes → **ST-1** (*chengqi*) at the infraorbital ridge and reaches → **BL-1** (*jingming*).

➡ **The third branch** of the *du mai* travels bilaterally alongside the BL channel from → **BL-1** (*jingming*) to the vertex. From here, it branches out **internally** into the Brain. The **external** pathway follows the medial branches of the BL channel in an inferior direction and enters the body at the level of the lower border of the spinous process of L2 at → **BL-23** (*shenshu*) to terminate at the Kidneys (*shen*).

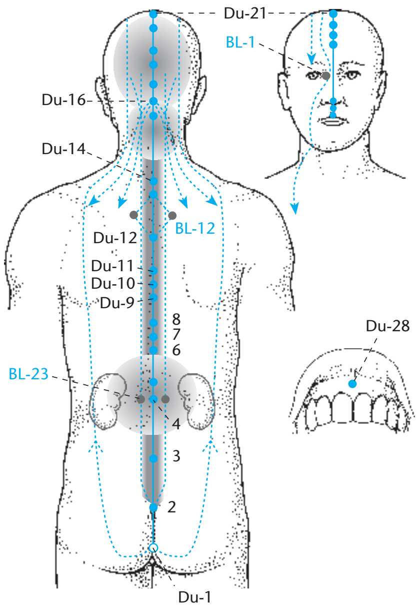

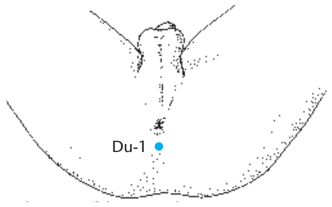

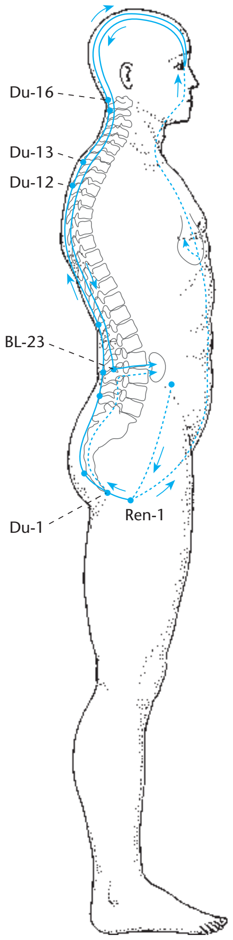

Meeting points of other channels with the *du mai*

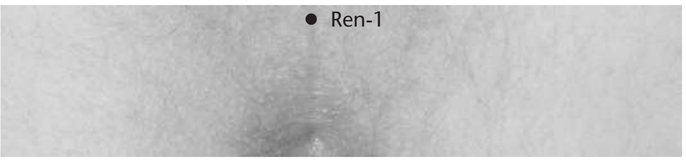

Ren-1 (*huiyin*): In the centre of the perineum

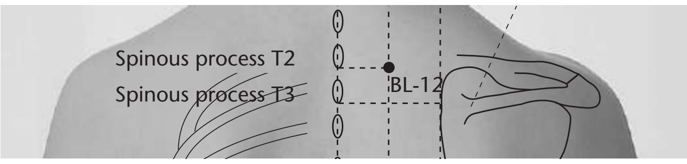

BL-12 (*fengmen*): 1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd thoracic vertebra (T2)

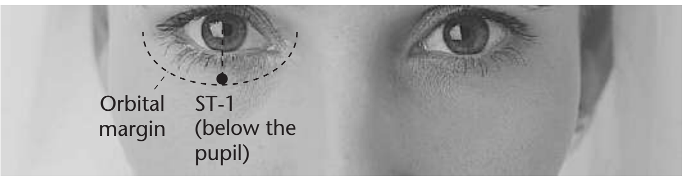

ST-1 (*chengqi*): With the eyes looking straight ahead, directly below the centre of the pupil, between the eyeball and the infraorbital ridge

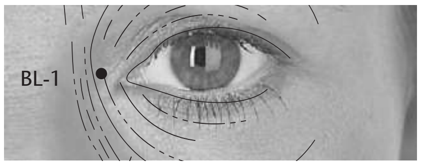

BL-1 (*jingming*): 0.1 cun superior and medial to the inner canthus of the eye, in a depression

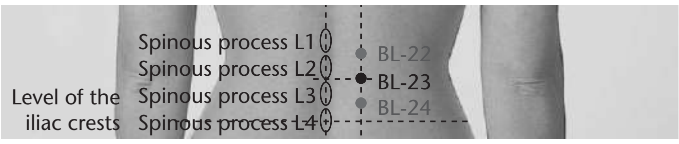

BL-23 (*shenshu*): 1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd lumbar vertebra (L2)

Connections with other channels/Organs

- *Ren mai*, *chong mai*, Bladder and Kidney primary channels
- Uterus, Kidneys, Brain

Clinical importance (→ 1.7.2, 1.7.3)

- **Sea of the Yang Qi**, supplies Yang Qi in order to warm the Organs and channels
- Supports and regulates all Yang channels, especially those of the upper back and head
- Influences and strengthens the functions of the Brain, Marrow and sensory organs
- Strengthens the body constitutionally and supports the immune system (circulation of the Defensive Qi (*wei qi*))
- Stores the Essence Qi (*jing qi*) overflowing from the Yang channels
- Absorbs excess from the Yang channels, especially excess Heat and internal Wind

Notes about the *du mai*

Both the *du mai* and *ren mai* have a unique position among the eight extraordinary vessels. They are the only two extraordinary vessels with their own specific points – all the other extraordinary vessels can only be accessed through points on the primary channels. For this reason the *ren mai* and the *du mai*, together with the primary channels, are often categorised as 14 channels. However, from the point of view of channel energetics this is not correct, since the extraordinary vessels occupy a special position, both evolutionarily and functionally (→ 1.7).

The pathway of the *du mai*, like that of the *ren mai*, is limited to the torso and the head. While it does not course along the extremities and does not have any points on the limbs (neither general points nor specific points such as the five *shu*-transporting points, *yuan*-source points, etc.) it does have a pertaining opening point (or master point) (→ 8.1.8).

The *du mai* is also referred to as the Sea of Yang and governs the Yang in the whole body. Located along its pathway are the posterior aspects of the energy centres (or chakras according to Yogi tradition) while their anterior aspects are located along the pathway of the *ren mai*. In Daoist tradition the *du mai* plays an important role in the cultivation of Qi. This concerns the development of a 'spirit child', which can be influenced by opening and activating energy centres located along the spine.

Furthermore, the *du mai* contains many meeting points with other channels. This clearly denotes the outstanding importance of the *du mai*.

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): –**
- ***Luo*-Connecting point (→ 8.1.2.): Du-1** (*changqiang*)
- ***Xi*-cleft point (→ 8.1.3): –**
- **Back-*shu* point (→ 8.1.4): –**
- **Front-*mu* point (→ 8.1.5): –**
- **Five *shu*-transporting points (→ 8.1.6): –**
- ***Hui*-meeting point (→ 8.1.7): –**
- **Pertaining opening point (→ 8.1.8): S.I.-3** (*houxi*)
- **Pertaining coupled point (→ 8.1.8): BL-62** (*shenmai*)
- **Lower *he*-Sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**

- with the G.B. and KID channels and the *ren mai*: **Du-1** (*changqiang*)
- with the BL channel: **Du-13** (*taodao*)
- with all Yang channels: **Du-14** (*dazhui*) 🟦🟦
- with the *yang wei mai*: **Du-15** (*yamen*) 🟦⬛
- with the *yang wei mai* and the *yang qiao mai**: **Du-16** (*fengfu*) 🟦🟦
- with the BL channel: **Du-17** (*naohu*)
- with the BL, G.B., T.B. and LIV channels: **Du-20** (*baihui*) 🟦🟦
- with the BL and ST channels: **Du-24** (*shenting*) 🟦⬛
- with the L.I. and ST channels: **Du-26** (*renzhong*) 🟦🟦
- with the *ren mai* and ST channel*: **Du-28** (*yinjiao*)
- of other channels with the *du mai*: **Ren-1, BL-12, ST-1, BL-1, BL-23**

- **Gao Wu command point (→ 8.1.11):** –
- **Window of Heaven point (→ 8.1.12): Du-16** (*fengfu*) 🟦🟦
- **Points of the Four Seas (→ 8.1.13):**
  - Sea of Qi: **Du-14** (*dazhui*) 🟦🟦, **Du-15** (*yamen*) 🟦⬛
  - Sea of Marrow: **Du-16** (*fengfu*) 🟦🟦; **Du-20** (*baihui*) 🟦🟦
- **Ma Dan Yang Heavenly Star point (→ 8.1.14):**
- **Sun Si Miao Ghost point (→ 8.1.15): Du-16** (*fengfu*) 🟦🟦, **Du-26** (*renzhong*) 🟦🟦
- **Other functional points:**
  - very important point for strengthening the Yang (with moxibustion): **Du-4** (*mingmen*) 🟦🟦
  - important emergency point: **Du-26** (*renzhong*) 🟦🟦

Points according to region

- **Local points (→ 8.2.1):** Temporal region – **Du-20** (*baihui*) 🟦🟦; nose – **Du-25** (*suliao*); rectum – **Du-1** (*changqiang*); lumbar spine – **Du-3** (*yaoyangguan*), **Du-4** (*mingmen*) 🟦🟦
- **Adjacent points (→ 8.2.1):** occipital region – **Du-14** (*dazhui*) 🟦🟦; temporal region, eyes – **Du-23** (*shangxing*); nose – **Du-23** (*shangxing*), **Du-24** (*shenting*); Lungs – **Du-14** (*dazhui*) 🟦🟦; Heart – **Du-11** (*shendao*); cervical spine – **Du-14** (*dazhui*) 🟦🟦
- **Distal points (→ 8.2.1):** tongue – **Du-15** (*yamen*) 🟦⬛; rectum – **Du-20** (*baihui*) 🟦🟦; lumbar spine – **Du-12** (*shenzhu*), – **Du-26** (*renzhong*) 🟦🟦

Specific points according to the vessel pathway (in numerical order)

- **Du-1** (*changqiang*): *luo*-connecting point (→ 8.1.2); *jiaohui*-meeting point with the G.B., KID channels and the *ren mai* (→ 8.1.10); local point for the rectum (→ 8.2.1)
- **Du-3** (*yaoyangguan*): local point for the lumbar spine (→ 8.2.1)
- **Du-4** (*mingmen*) 🟦🟦: very important point for strengthening the Yang (with moxibustion)
- **Du-11** (*shendao*): adjacent point for the Heart (→ 8.2.1)
- **Du-12** (*shenzhu*): distal point for the lumbar spine (→ 8.2.1)
- **Du-13** (taodao): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **Du-14** (*dazhui*) 🟦🟦: *jiaohui*-meeting point with all Yang channels (→ 8.1.10); point of the Sea of Qi (→ 8.1.13); adjacent point for the occiput and Lung (→ 8.2.1); adjacent point for the cervical spine (→ 8.2.1)
- **Du-15** (*yamen*) 🟦⬛: *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10); Sea of Qi point (→ 8.1.13); distal point for the tongue (→ 8.2.1)
- **Du-16** (*fengfu*) 🟦🟦: point of the Sea of Marrow (→ 8.1.13); Window of Heaven point (→ 8.1.12); Sun Si Miao Ghost point (→ 8.1.15); *jiaohui*-meeting point with the *yang wei mai* and the *yang qiao mai** (→ 8.1.10)
- **Du-17** (*naohu*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **Du-20** (*baihui*) 🟦🟦 *jiaohui*-meeting point with the BL, G.B., T.B. and LIV channels (→ 8.1.10); point of the Sea of Marrow (→ 8.1.13); local point for the temporal aspect of the head (→ 4.2.1); distal point for the rectum (→ 8.2.1)
- **Du-23** (*shangxing*): regional point for the temporal region, the eyes and the nose (→ 8.2.1)
- **Du-24** (*shenting*) 🟦⬛: *jiaohui*-meeting point with the BL and ST channels (→ 4.1.10); adjacent point for the nose (→ 8.2.1)
- **Du-25** (*suliao*): local point for the nose (→ 4.2.1)
- **Du-26** (*renzhong*) 🟦🟦: *jiaohui*-meeting point with the L.I. and ST channels (→ 8.1.10); Sun Si Miao Ghost point (→ 8.1.15); distal point for the lumbar spine (→ 8.2.1); important emergency point
- **Du-28** (*yinjiao*): *jiaohui*-meeting point with the *ren mai* and ST channel (→ 8.1.10)

---

* Mentioned by only some authors

luo-Connecting Vessel of the *Du Mai*

Pathway

The *luo*-connecting vessel of the *du mai* originates at **Du-1** (*changqiang*) between the anus and the tip of the coccyx, bilaterally ascends the spine to the neck and disperses in the occipital region, where it connects with the Bladder primary channel. Both course together to the shoulder region. There the *luo*-connecting vessel enters deeper into the body at the level of the scapula and spreads in the paravertebral musculature.

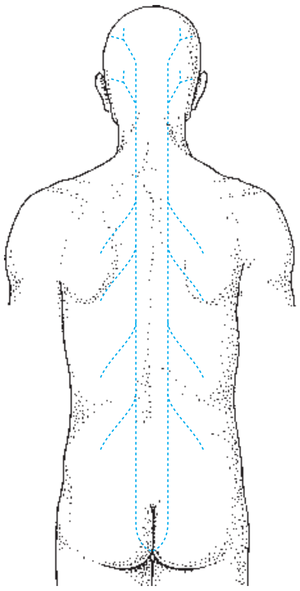

Clinical importance

Pathology

Excess (***shi***): limited range of motion and stiffness of the spine

Deficiency (***xu***): heavy sensations of the head, dizziness, balance disorders

Long Strong CHANGQIANG **Du-1**

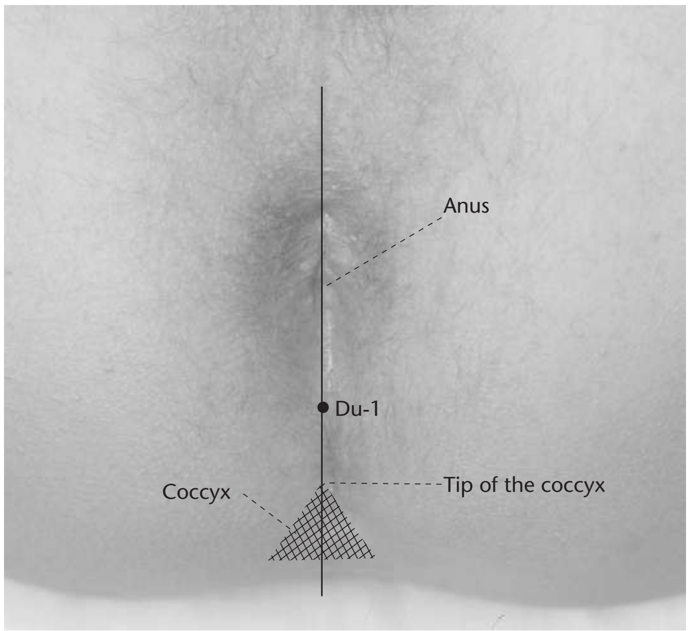
Anus
Du-1
Coccyx
Tip of the coccyx

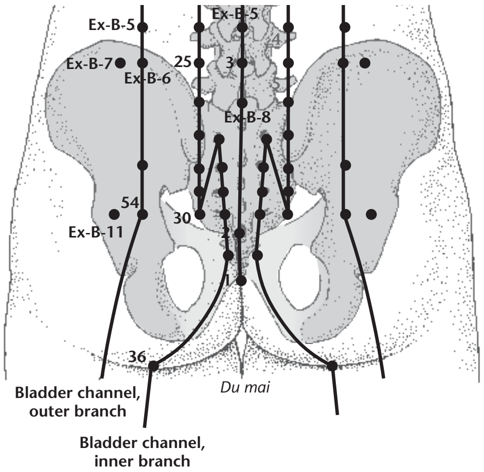
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
36
Bladder channel, outer branch
Du mai
Bladder channel, inner branch

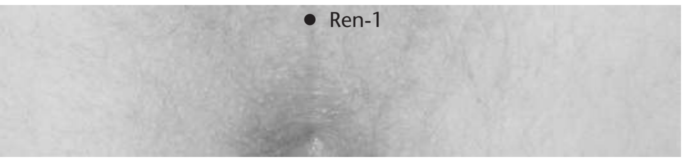

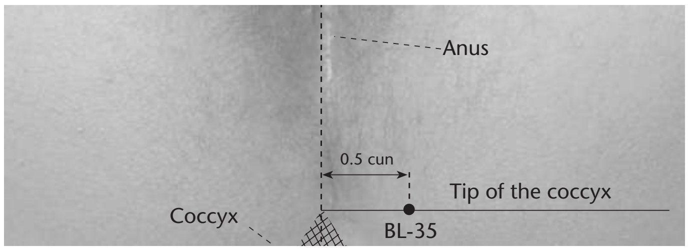
Anus
0.5 cun
Tip of the coccyx
Coccyx
BL-35

Location

On the midline, midway between the tip of the coccyx and the anus.

How to find

First, locate the coccyx superior to the anus, then palpate for its tip. From the tip of the coccyx, palpate with pressure towards the anus. **Du-1** can often be located based on a strong *de* Qi sensation.

→ **Ren-1** is located anterior to the anus, posterior to the genitals.

→ **BL-35** is located 0.5 cun lateral to the midline, at the level of the tip of the coccyx.

Needling

After thorough disinfection, vertically 0.5–1 cun. Avoid puncturing the rectum.

Actions/Indications

- Clears Damp-Heat, lifts the Qi
- Opens the channel and alleviates pain
- Calms the *shen*

Special features

Luo-connecting point, meeting point with the *ren mai*, G.B. and KID channels.

Du-2 *Lumbar Shu* YAOSHU

Location

On the midline, in the sacral hiatus.

How to find

Palpate for the sacral hiatus (→ 3.4.4) from the sacrum. It can be felt at the inferior end of the iliac median crest as a U-shaped depression opening towards the inferior. **Du-2** is located on the midline, directly below the hiatus.
→ **BL-35** is located slightly more inferiorly, 0.5 cun lateral to the midline, on the level of the tip of the coccyx.

Needling

Up to 1 cun vertically or obliquely in a superior direction.

Actions/Indications

- Strengthens the lumbar region and the legs
- Dispels Wind-Dampness
- Warms the Lower Burner

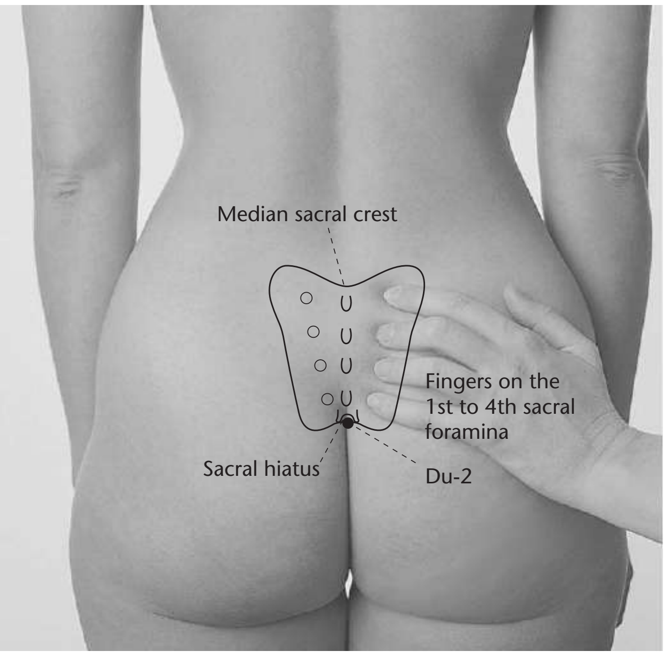
Median sacral crest
Fingers on the 1st to 4th sacral foramina
Sacral hiatus
Du-2

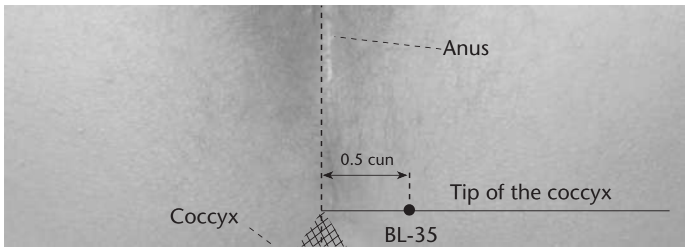
Anus
0.5 cun
Tip of the coccyx
Coccyx
BL-35

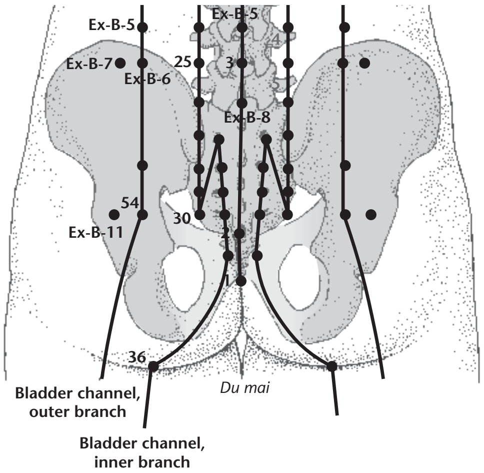
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
36
Bladder channel, outer branch
Du mai
Bladder channel, inner branch

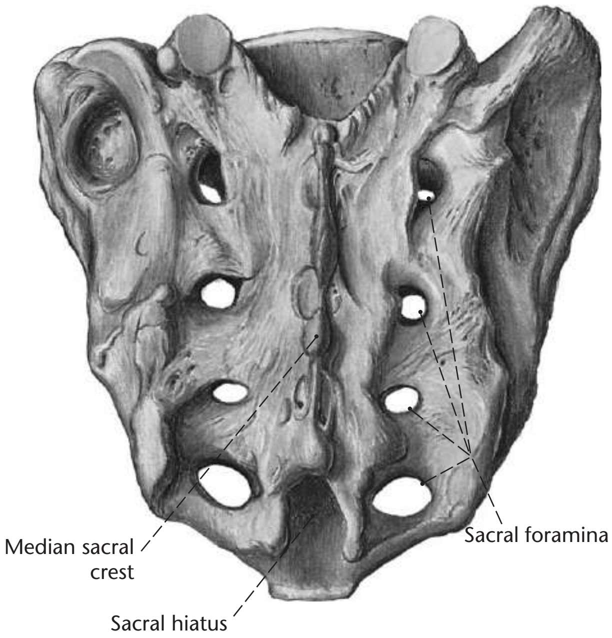
Median sacral crest
Sacral foramina
Sacral hiatus

Lumbar Yang Gate YAOYANGGUAN Du-3

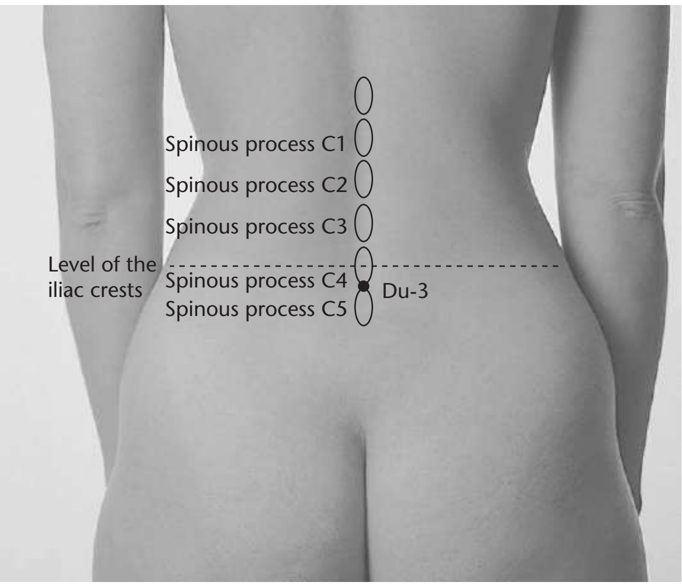
Spinous process C1
Spinous process C2
Spinous process C3
Level of the iliac crests
Spinous process C4
Du-3
Spinous process C5

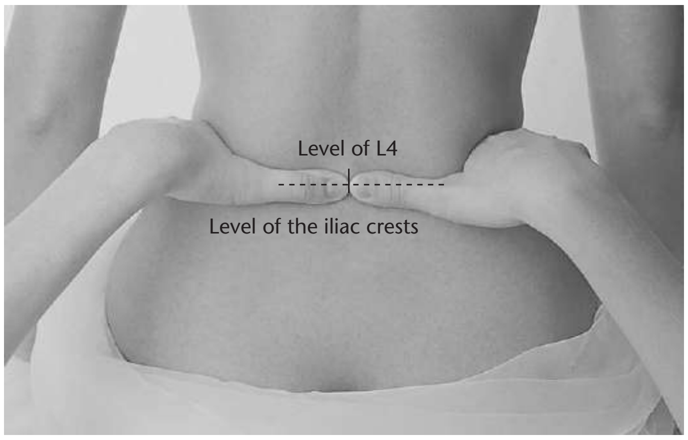
Level of L4
Level of the iliac crests

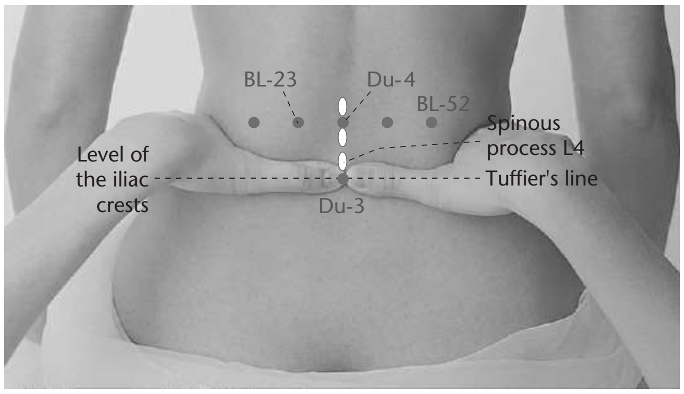
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

Location

On the midline, below the spinous process of the 4th lumbar vertebra (L4).

How to find

For orientation in the lumbar region (→ 3.4.3), first, locate the Tuffier's line (a line connecting the two most superior points on the iliac crest), which generally intersects with the spinous process of L4. (Note: Point location in the lumbar region is dependent on the patient's position; for more detail on orientation → 3.4). Locate **Du-3** on the midline, below the spinous process of L4.

Located on the same level are a point of → **Ex-B-2/BL-25/ Ex-B-6/Ex-B-7** (0.5/1.5/3/3.5 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction (generally, there is only danger of accidental puncture of the spinal canal from the level of L2 in a cranial direction).

Actions/Indications

- Dispels Wind, Cold and Dampness
- Regulates Qi in the Lower Burner, strengthens the Original Qi (*yuan qi*)

Special features

Important local point for combined back and leg pain

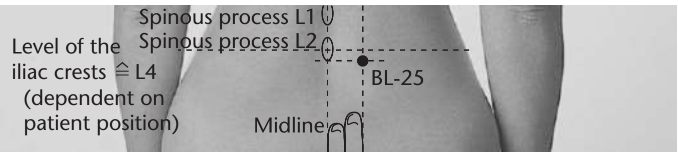
Spinous process L1
Spinous process L2
Level of the iliac crests ≙ L4 (dependent on patient position)
BL-25
Midline

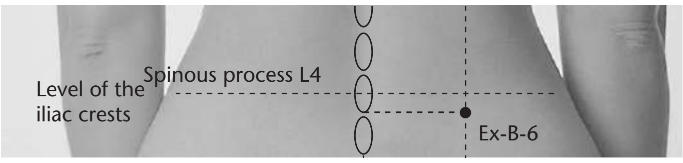
Spinous process L4
Level of the iliac crests
Ex-B-6

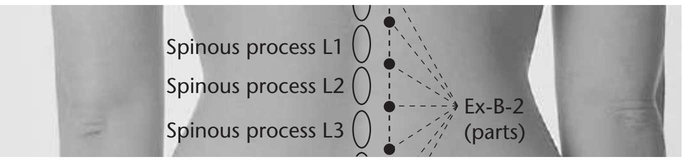
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

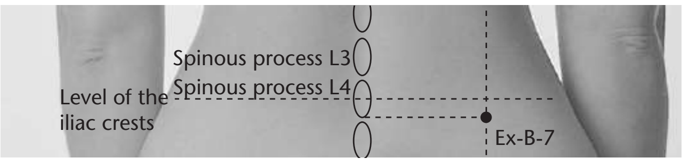
Spinous process L3
Spinous process L4
Level of the iliac crests
Ex-B-7

Du-4 *Gate of Life* MINGMEN

Location

On the posterior midline, below the spinous process of the 2nd lumbar vertebra (L2).

How to find

For orientation in the lumbar region (→ 3.4.3), first, locate the Tuffier's line by placing both hands directly superior and lateral to the two highest points on the iliac crest, joining the thumbs on the midline. The line will generally intersect with the spinous process of L4. (Note: Orientation on the lumbar region depends on the patient's position; for more detail on differential orientation → 3.4). From there, count up to the spinous process of L2 and locate **Du-4** on the midline, below the spinous process. Located on the same level are a point of → **Ex-B-2/BL-23/BL-52** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun vertically to slightly obliquely in an inferior direction, with the patient's back straight or flexed. The spinal cord generally ends between the 1st and 2nd lumbar vertebrae; still, there is hardly any mention in the literature of spinal punctures with deep needling and/or insertions in a superior direction.

Actions/Indications

- Tonifies Kidney Yang (especially with moxibustion), warms the *mingmen*, strengthens the Kidneys, regulates the *du mai*, benefits the lumbar region
- Pacifies Wind in the *du mai*

Special features

Major point for tonifying Yang, especially Kidney Yang

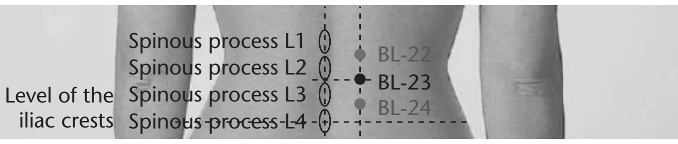
Level of the
iliac crests
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
BL-22
BL-23
BL-24

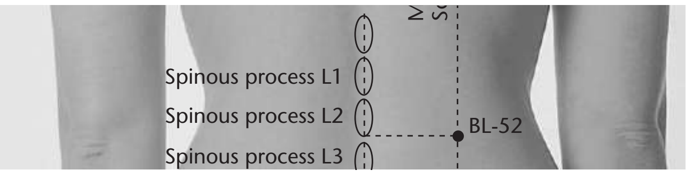
Spinous process L1
Spinous process L2
Spinous process L3
BL-52

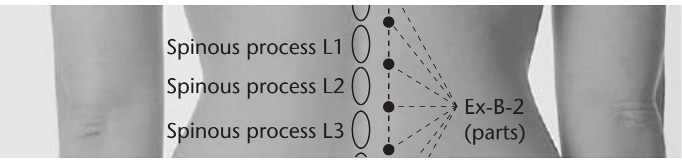
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

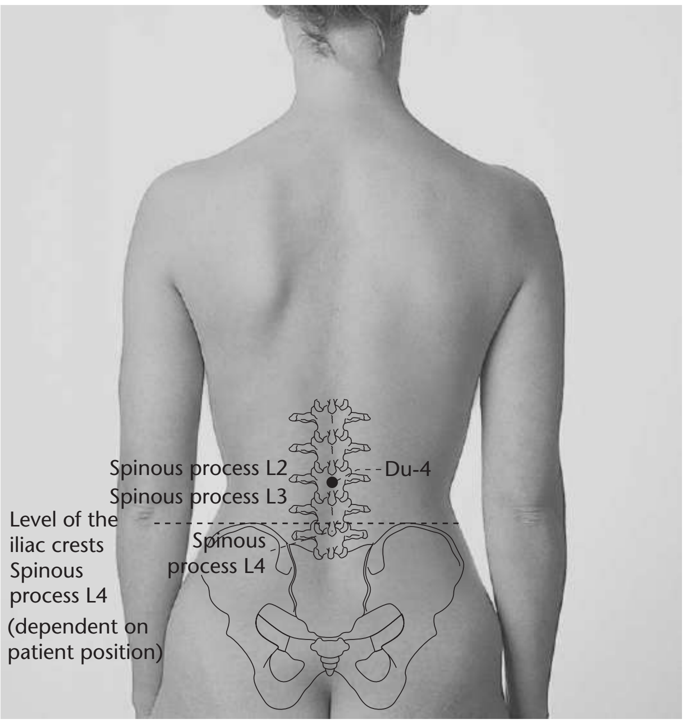
Spinous process L2
Du-4
Spinous process L3
Level of the
iliac crests
Spinous
process L4
Spinous
process L4
(dependent on
patient position)

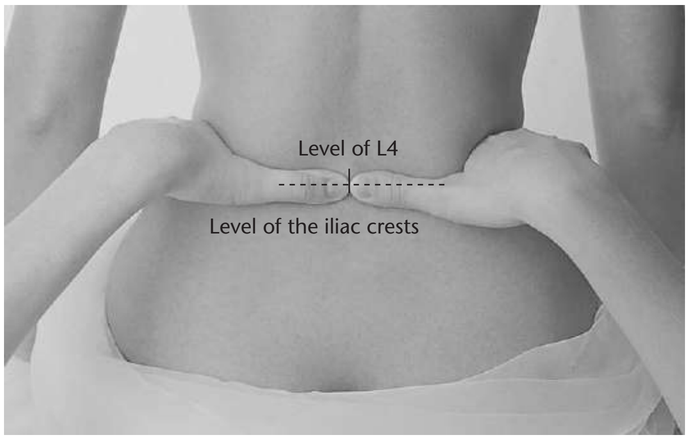
Level of L4
Level of the iliac crests

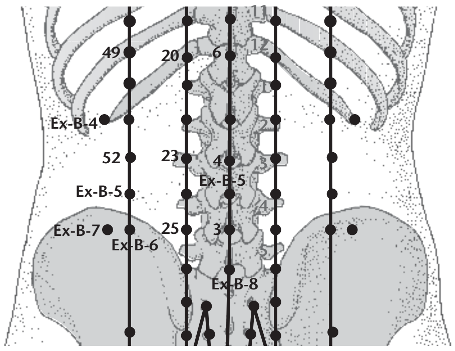
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
Ex-B-6
25
3
Ex-B-8

Ch05b-F10028.qxd 2/22/08 9:55 PM Page 503

Suspended Pivot XUANSHU Du-5

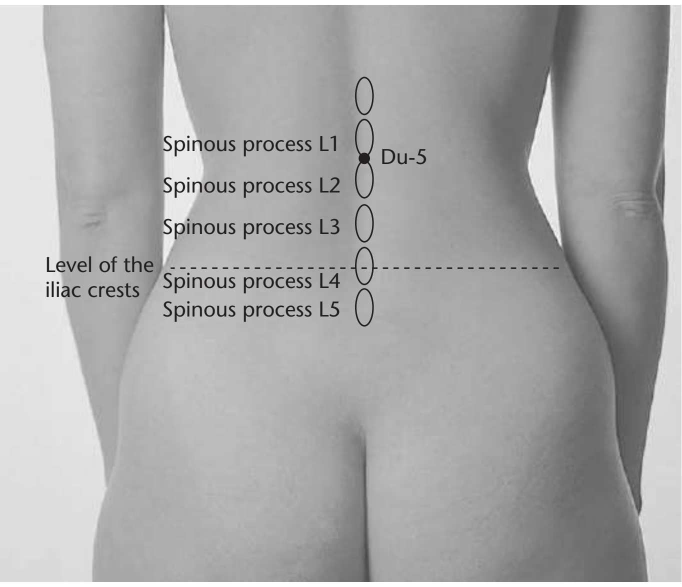
Spinous process L1
Du-5
Spinous process L2
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5

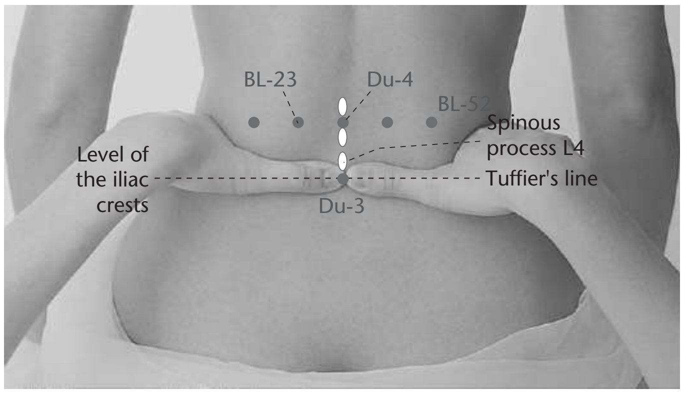
BL-23
Du-4
BL-5
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

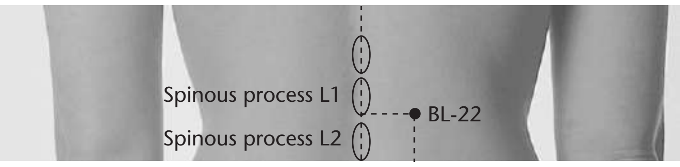
Spinous process L1
BL-22
Spinous process L2

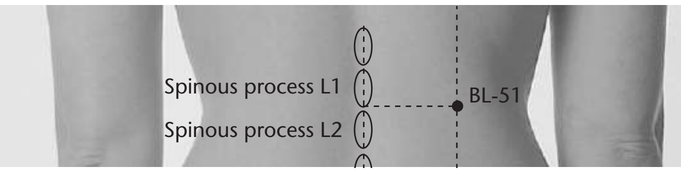
Spinous process L1
BL-51
Spinous process L2

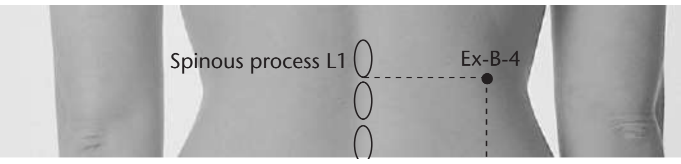
Spinous process L1
Ex-B-4

Location

On the midline, below the spinous process of the 1st lumbar vertebra (L1).

How to find

For orientation in the lumbar region (→ 3.4.3), first, locate the Tuffier's line by placing both hands directly superior and lateral to the two highest points on the iliac crest, joining the thumbs on the midline. The Tuffier's line generally intersects with the spinous process of L4. (Note: Anatomical orientation in the lumbar region is dependent on the patient's position; for more detail on differential orientation → 3.4.) From there, count up to the spinous process of L1 and locate **Du-5** on the midline, below its spinous process. **Or:** Use the spinous process of T7 (inferior angle of the scapula → 3.4.2) for reference.

Located on the same level are a point of → **Ex-B-2/BL-22/BL-51/Ex-B-4** (0.5/1.5/3/3.5 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Opens the channel
- Regulates Qi in the Lower Burner

Du-6 *Centre of the Spine* JIZHONG

Location

On the midline, below the spinous process of the 11th thoracic vertebra (T11).

How to find

Locate T7 (→ 3.4.2) and, from there, count down 4 spinous processes to T11. Locate **Du-6** on the midline, below its spinous process. **Or:** Use the Tuffier's line for reference (→ 3.4.3). Locate T11 by counting up 5 spinous processes from L4 and there locate **Du-6** on the midline, below the spinous process.

Located on the same level are a point of → **Ex-B-2/BL-20/BL-49** (0.5/1.5/3 cun lateral to the midline).

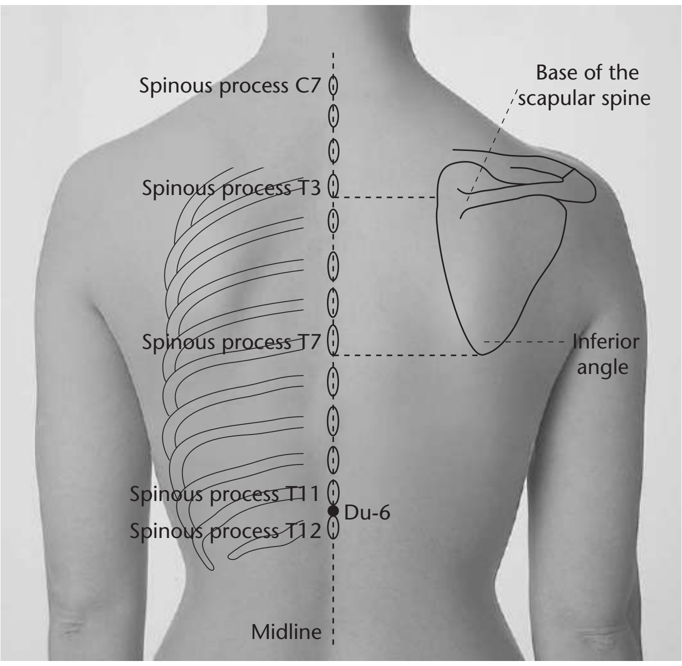
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T7
Inferior angle
Spinous process T11
Du-6
Spinous process T12
Midline

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Strengthens the Spleen and its function of transforming Dampness
- Benefits the spine

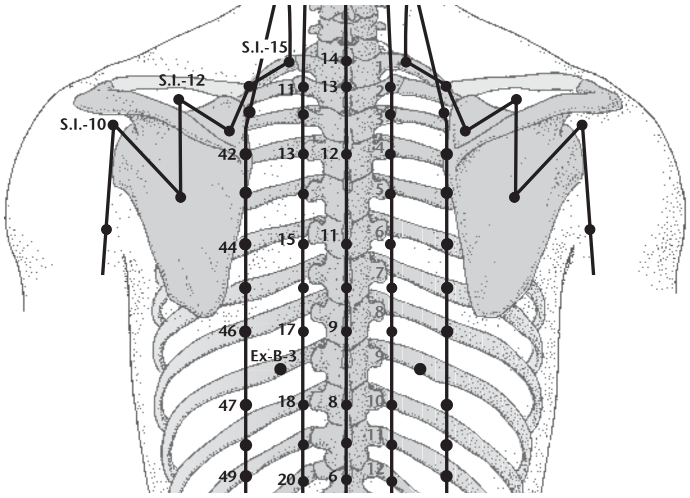
S.I.-15
S.I.-12
S.I.-10
14
13
11
1
2
3
42
13
12
4
5
44
15
11
6
7
46
17
9
8
Ex-B-3
9
47
18
8
10
11
49
20
6
12

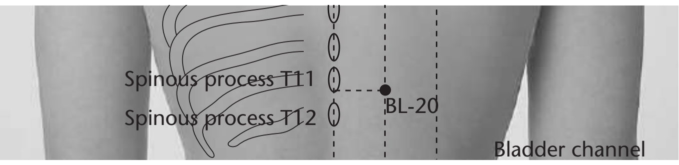
Spinous process T11
Spinous process T12
BL-20
Bladder channel

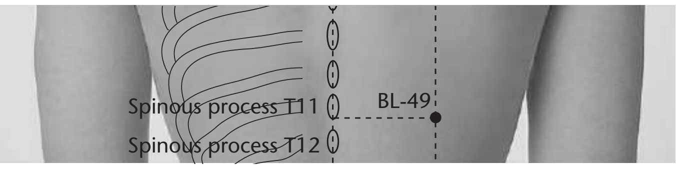
Spinous process T11
BL-49
Spinous process T12

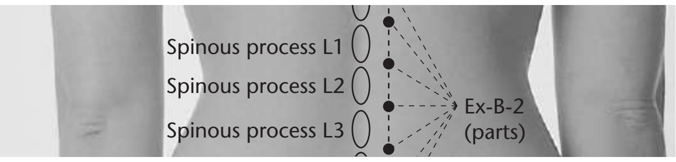
Spinous process L1
Spinous process L2
Ex-B-2
(parts)
Spinous process L3

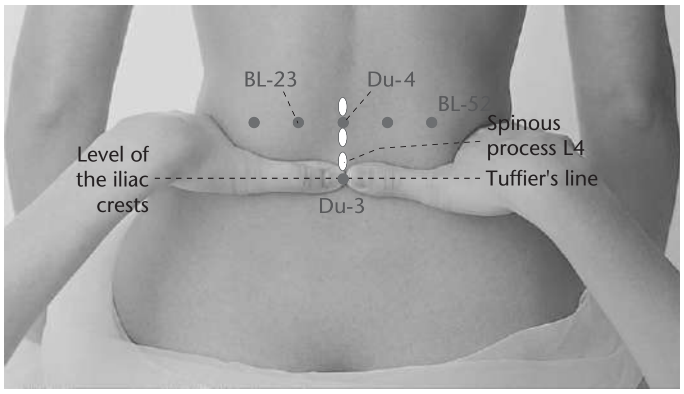
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

Central Pivot ZHONGSHU **Du-7**

Location

On the midline, below the spinous process of the 10th thoracic vertebra (T10).

How to find

Locate T7 (➞ 3.4.2) and, from there, count down 3 spinous processes to T10. Locate **Du-7** on the midline, below its spinous process. **Or:** Use the Tuffier's line for reference (➞ 3.4.3). Locate T11 by counting up 5 spinous processes from L4 and there locate **Du-7** on the midline, below the spinous process.
Located on the same level are a point of ➞ **Ex-B-2/BL-19/BL-48** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Tonifies the Middle Burner
- Benefits the spine

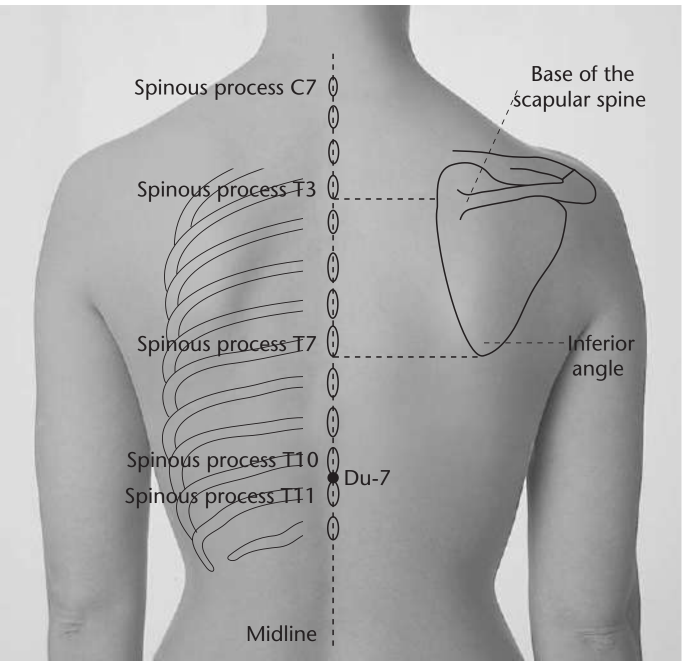
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T7
Inferior angle
Spinous process T10
Du-7
Spinous process T11
Midline

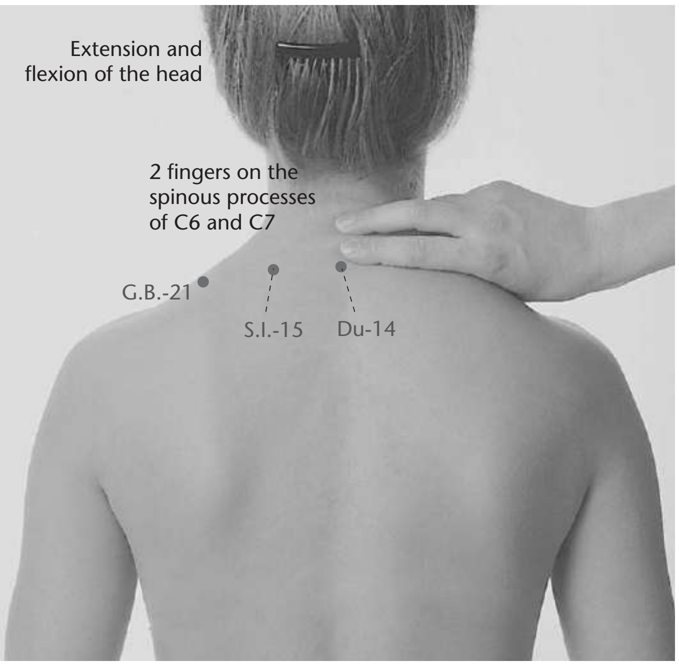
Extension and flexion of the head
2 fingers on the spinous processes of C6 and C7
G.B.-21
S.I.-15
Du-14

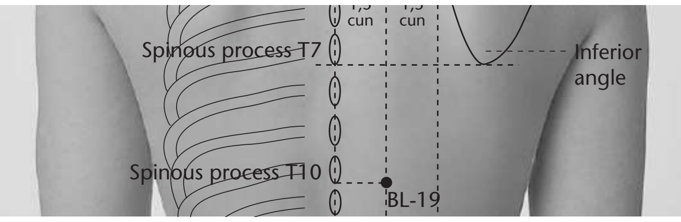
cun
cun
Spinous process T7
Inferior angle
Spinous process T10
BL-19

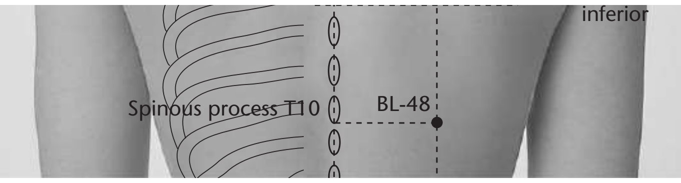
inferior
Spinous process T10
BL-48

Du-8 *Sinew Contraction* JINSUO

Location

On the midline, below the spinous process of the 9th thoracic vertebra (T9).

How to find

Locate T7 (→ 3.4.2) and, from there, count down 2 spinous processes to T9. Locate **Du-8** on the midline, below its spinous process. **Or:** Use the Tuffier's line for reference (→ 3.4.3). Locate T9 by counting up from L4 and there locate **Du-8** on the midline, below the spinous process.
Located on the same level are a point of → **Ex-B-2/BL-18/BL-47** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Soothes the Liver
- Dispels Wind
- Calms the *shen*

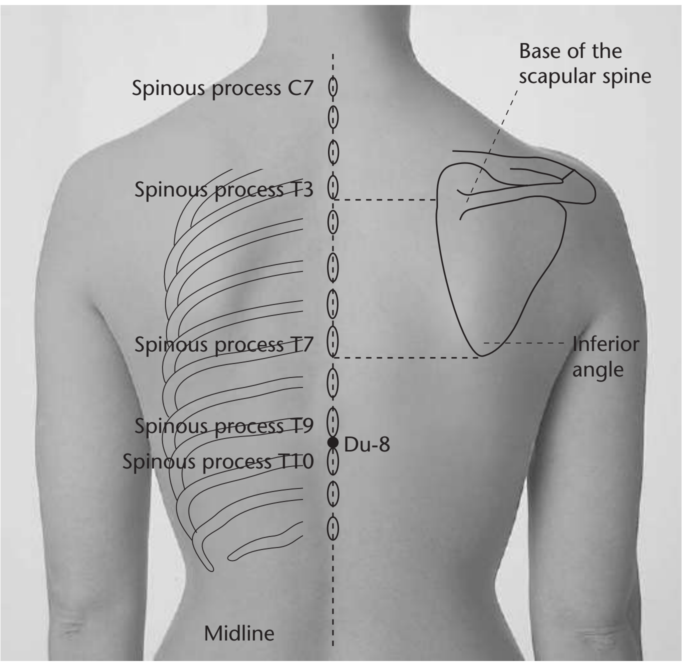
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T7
Inferior angle
Spinous process T9
Du-8
Spinous process T10
Midline

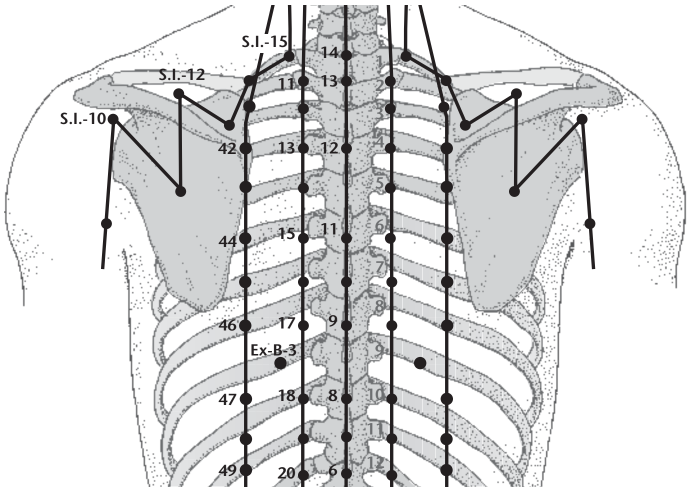
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3
47
18
8
49
20
6

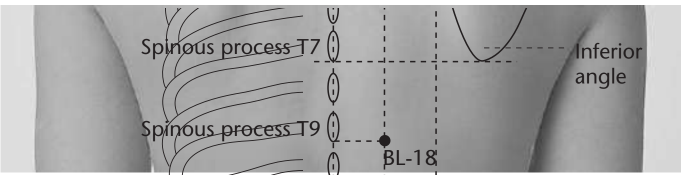
Spinous process T7
Inferior angle
Spinous process T9
BL-18

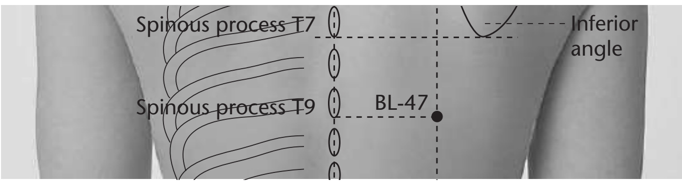
Spinous process T7
Inferior angle
Spinous process T9
BL-47

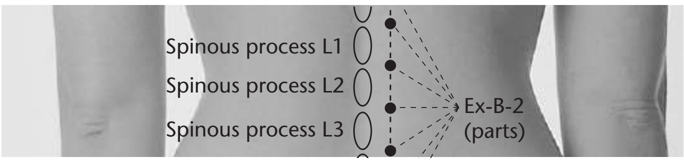
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Reaching Yang ZHIYANG **Du-9**

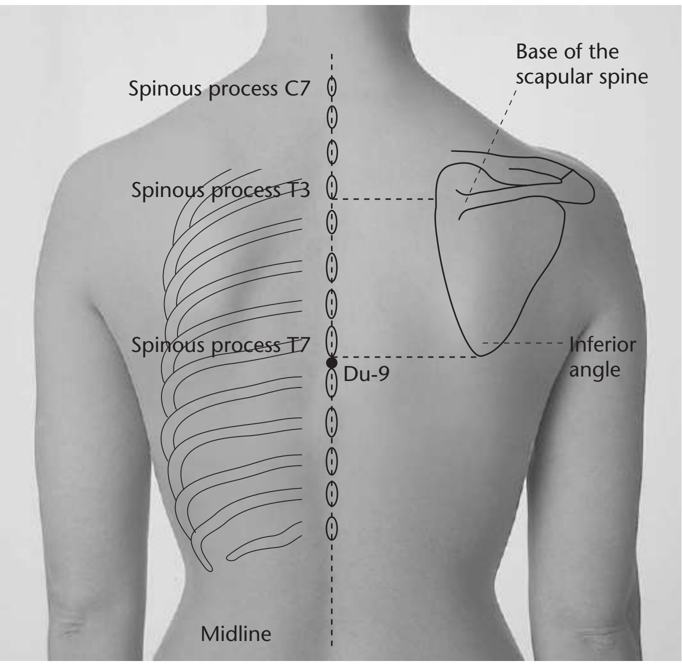
Base of the scapular spine
Spinous process C7
Spinous process T3
Spinous process T7
Du-9
Inferior angle
Midline

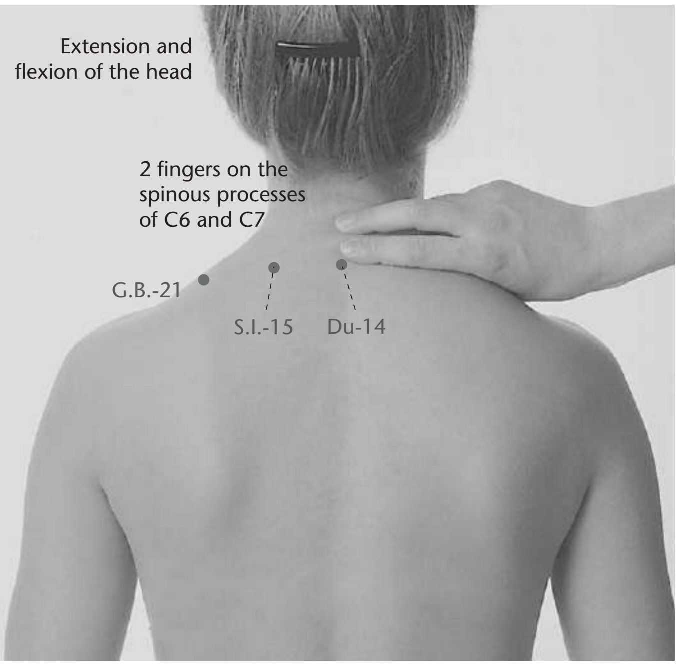
Extension and flexion of the head
2 fingers on the spinous processes of C6 and C7
G.B.-21
S.I.-15
Du-14

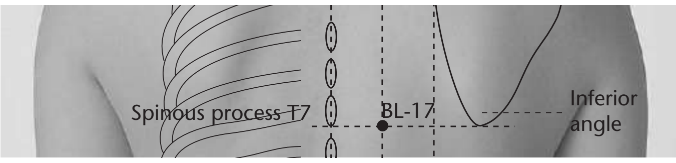
Spinous process T7
BL-17
Inferior angle

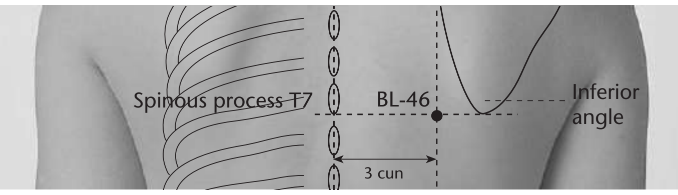
Spinous process T7
BL-46
Inferior angle
3 cun

Location

On the midline, below the spinous process of the 7th thoracic vertebra (T7).

How to find

Find T7 (→ 3.4.2) and locate **Du-9** on the midline, below its spinous process. **Or:** Use the Tuffier's line for reference (→ 3.4.3). Locate T7 by counting up from L4 and there locate **Du-9** on the midline, below the spinous process.
Located on the same level are a point of → **Ex-B-2/BL-17/BL-46** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Strengthens the Spleen and eliminates Dampness
- Clears Damp-Heat from the Middle Burner
- Harmonises the Upper Burner

Du-10 *Spirit Tower* LINGTAI

Location

On the midline, below the spinous process of the 6th thoracic vertebra (T6).

How to find

Locate T7 (→ 3.4.2). **Du-10** is located on the midline, below the spinous process of the T6. **Or:** Locate C7 (→ 3.4.1). **Du-10** can then be located by counting down 6 spinous processes (T1–T6). Located on the same level are a point of → **Ex-B-2/BL-16/BL-45** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Harmonises the Upper Burner
- Clears Heat and Fire toxins

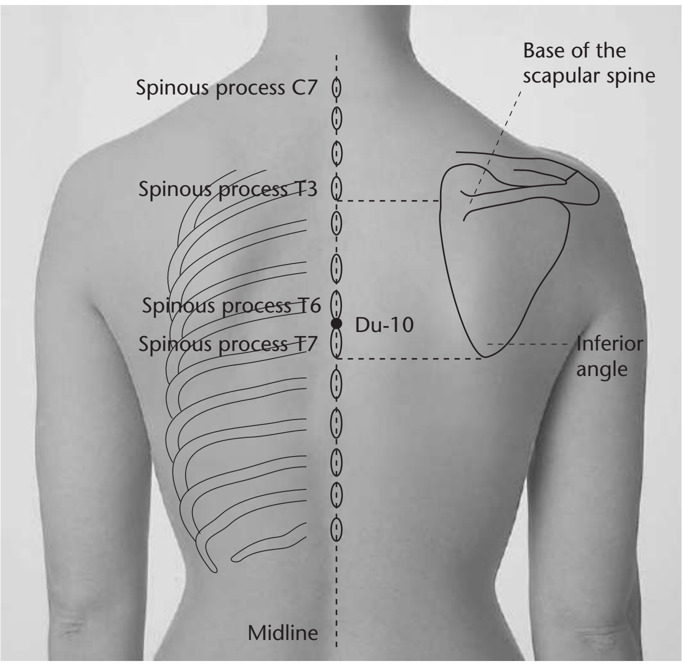
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T6
Du-10
Spinous process T7
Inferior angle
Midline

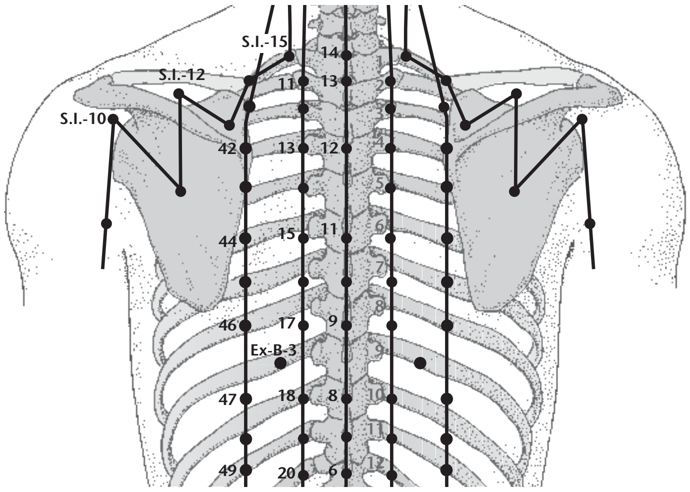
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
5
44
15
11
6
7
46
17
9
8
Ex-B-3
47
18
8
10
49
20
6
12

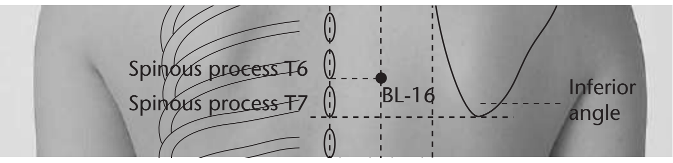
Spinous process T6
Spinous process T7
BL-16
Inferior angle

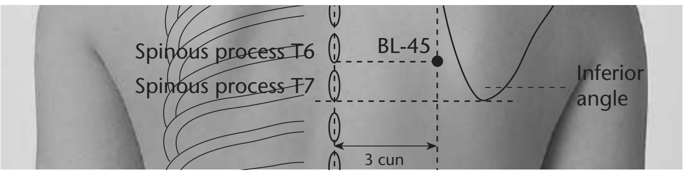
Spinous process T6
Spinous process T7
BL-45
Inferior angle
3 cun

Spirit Pathway SHENDAO Du-11

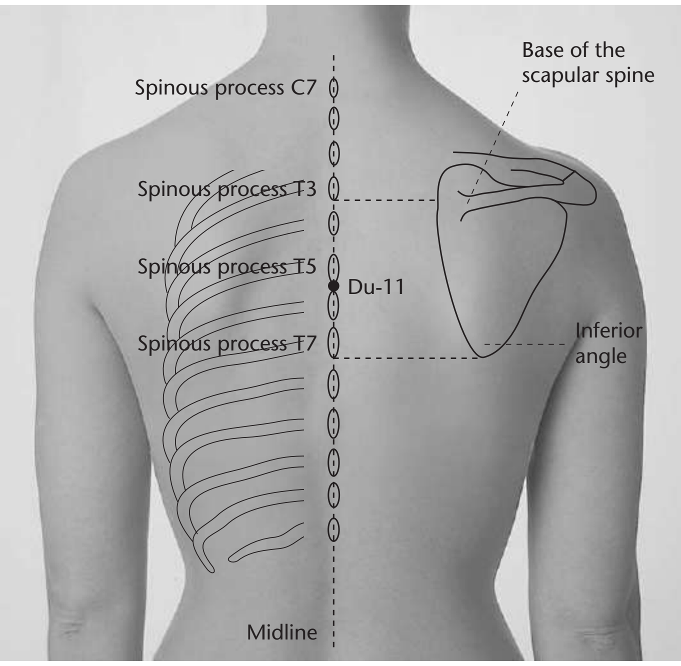
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T5
Du-11
Spinous process T7
Inferior angle
Midline

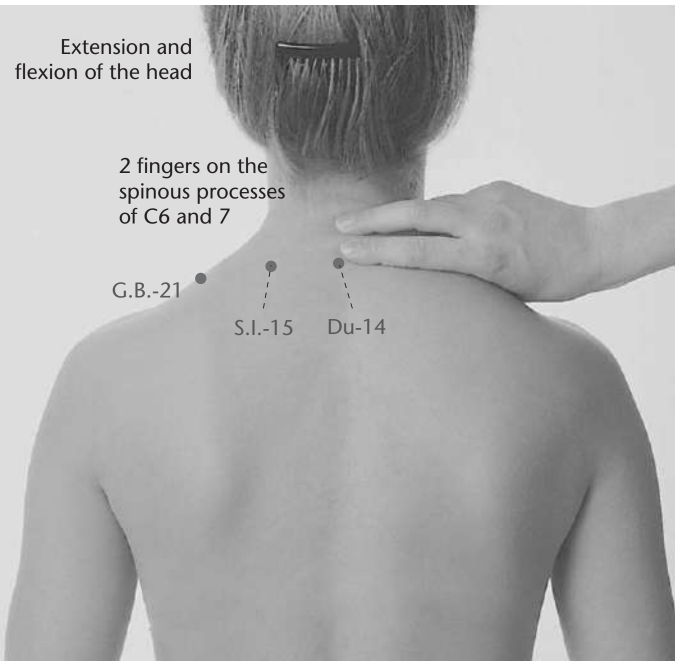
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

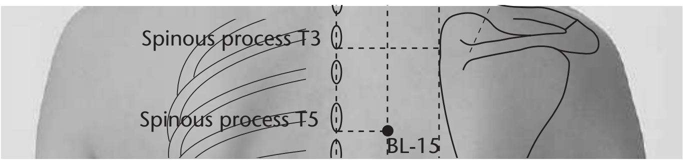
Spinous process T3
Spinous process T5
BL-15

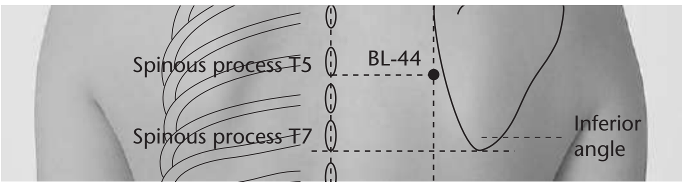
Spinous process T5
BL-44
Spinous process T7
Inferior angle

Location

On the midline, below the spinous process of the 5th thoracic vertebra (T5).

How to find

Locate T7 (→ 3.4.2). From there, count up 2 spinous processes and find **Du-11** on the midline, below the spinous process of T5.
Or: Locate C7 (→ 3.4.1). **Du-11** can then be located by counting down 5 spinous processes (T1–T5).
Located on the same level are a point of → **Ex-B-2/BL-15/BL-44** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Tonifies the Heart and the Lungs
- Calms the *shen*
- Dispels Wind and Heat

Du-12 *Body Pillar* SHENZHU

Location

On the midline, below the spinous process of the 3rd thoracic vertebra (T3).

How to find

Locate T7 (➞ 3.4.2). From there, count up 4 spinous processes and locate **Du-12** on the midline, below the spinous process of T3. **Or:** Locate C7 (➞ 3.4.1). **Du-12** can then be located by counting down 3 spinous processes (T1–T3).

Located on the same level are a point of ➞ **Ex-B-2/BL-13/BL-42** (0.5/1.5/3 cun lateral to the midline).

Needling

0.3–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Dispels internal and external Wind
- Calms the *shen*
- Tonifies the Lungs

Special features

Meeting point with the BL channel

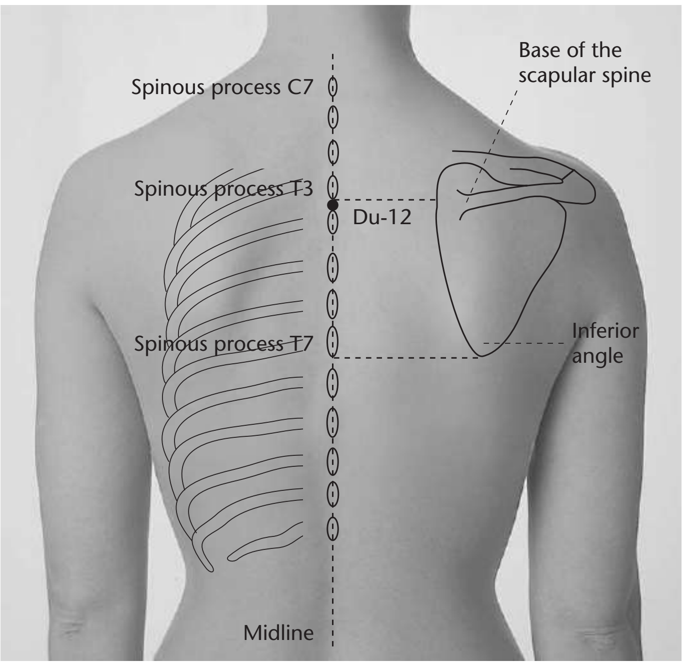
Spinous process C7
Base of the scapular spine
Spinous process T3
Du-12
Spinous process T7
Inferior angle
Midline

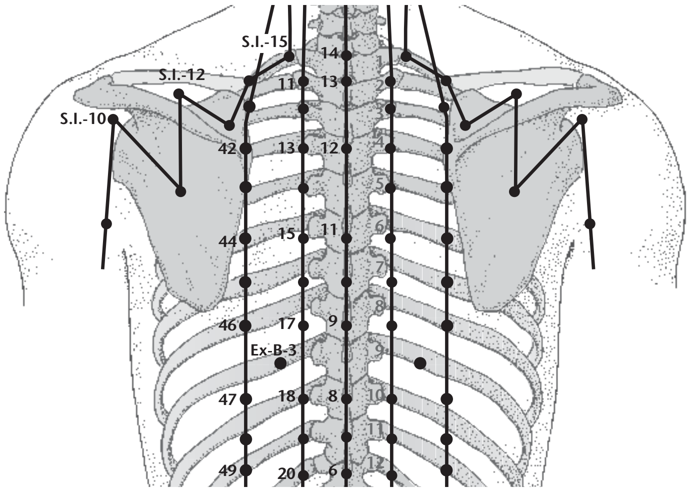
S.I.-15
S.I.-12
S.I.-10
14
13
11
1
2
3
42
13
12
4
5
44
15
11
6
7
46
17
9
8
Ex-B-3
9
47
18
8
10
11
49
20
6
12

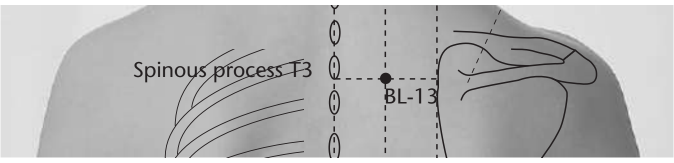
Spinous process T3
BL-13

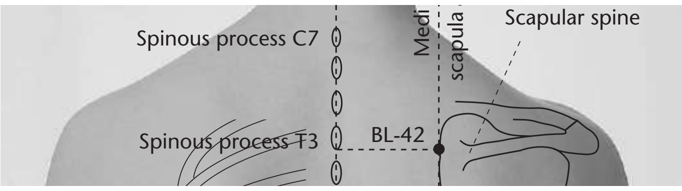
Spinous process C7
Medi
scapula
Scapular spine
Spinous process T3
BL-42

Way of Happiness TAODAO Du-13

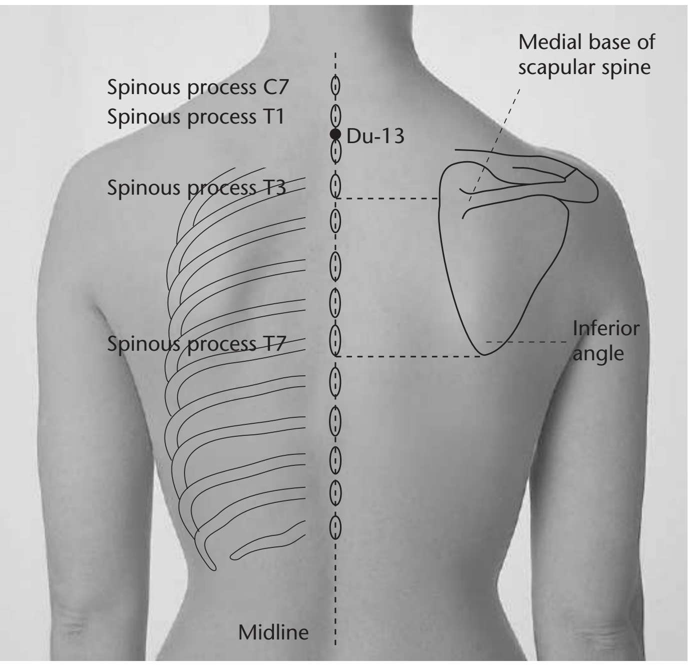
Medial base of scapular spine
Spinous process C7
Spinous process T1
Du-13
Spinous process T3
Spinous process T7
Inferior angle
Midline

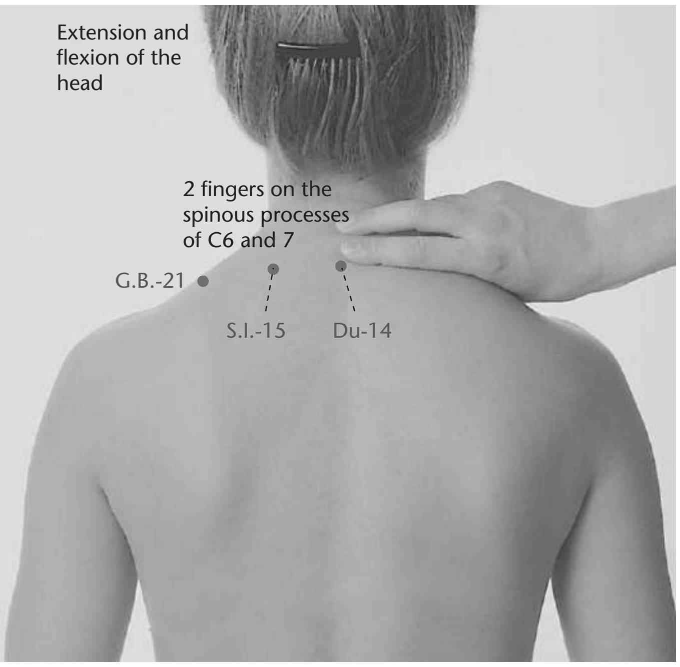
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

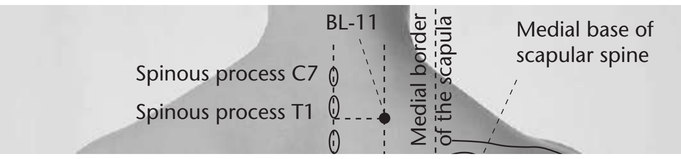
BL-11
Spinous process C7
Spinous process T1
Medial border of the scapula
Medial base of scapular spine

Location

On the midline, below the spinous process of the 1st thoracic vertebra (T1).

How to find

Locate C7 (→ 3.4.2). Palpate downward to the next spinous process (T1) and locate **Du-13** on the midline, below its spinous process.

Located on the same level are a point of → **Ex-B-2/BL-11/S.I.-14** (0.5/1.5/3 cun lateral to the midline).

Needling

0.3–1 cun strictly perpendicularly to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Dispels pathogenic factors
- Opens the channel and *luo*-connecting vessels

Special features

Meeting point with the BL channel

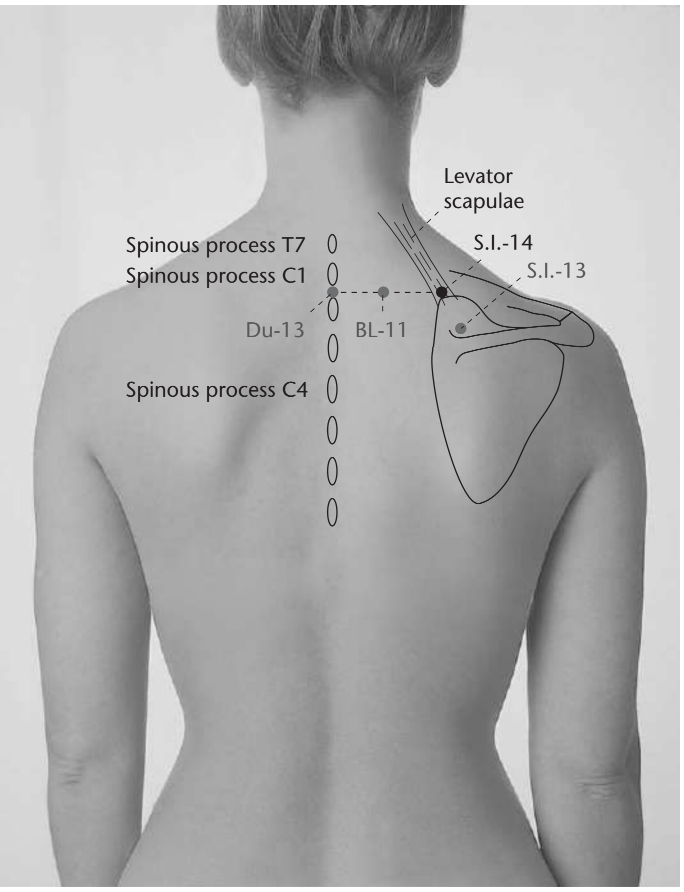
Levator scapulae
Spinous process T7
Spinous process C1
S.I.-14
S.I.-13
Du-13
BL-11
Spinous process C4

Du-14 *Great Vertebra* DAZHUI

Location

On the midline, below the spinous process of the 7th cervical vertebra (C7).

How to find

First, identify C7 (→ 3.4.1): Place 2 fingers on the spinous processes believed to belong to C6 and C7 and ask the patient to flex and extend their head. In a fully functional spine and with correct finger placement, C6 will glide anteriorly with neck extension, while C7 will remain fixed. However, if the vertebra below the upper finger does not move with neck extension, the fingers will most probably have been placed on C7 and T1. Locate **Du-14** below the spinous process of C7.

Located on the same level are → **Ex-B-1** (*dingchuan*)/ **S.I.-15/Ex-B** (*jiehexue*) (0.5/2/3.5 cun lateral to **Du-14**).

Needling

0.3–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun. For persistent epistaxis, place a cold metal object (for example, a spoon) or a cold cloth on **Du-14**. Cupping on this point for acute colds.

Actions/Indications

- Clears Wind, firms and regulates the surface, dispels pathogenic factors
- Clears Heat
- Calms (internal) Wind and the *shen*
- Strengthens Qi and Yang in deficiency conditions
- Benefits the spine, especially the cervical spine

Special features

Meeting point with all Yang channels, Sea of Qi point.

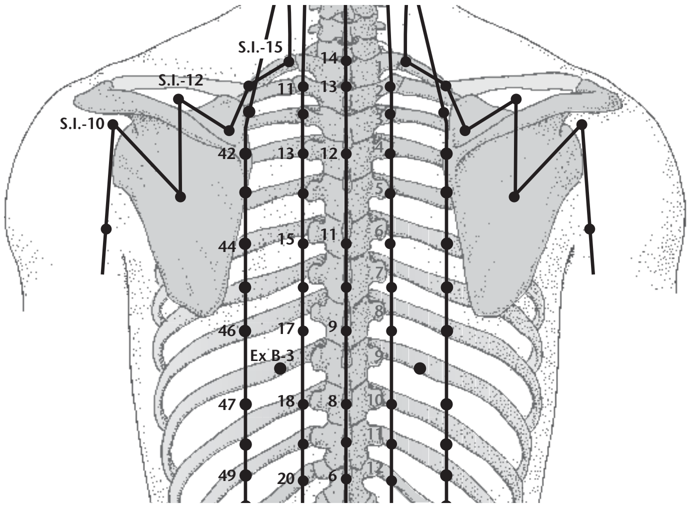
S.I.-15
S.I.-12
S.I.-10
14
13
11
42
13
12
44
15
11
46
17
9
Ex B-3
47
18
8
49
20
6

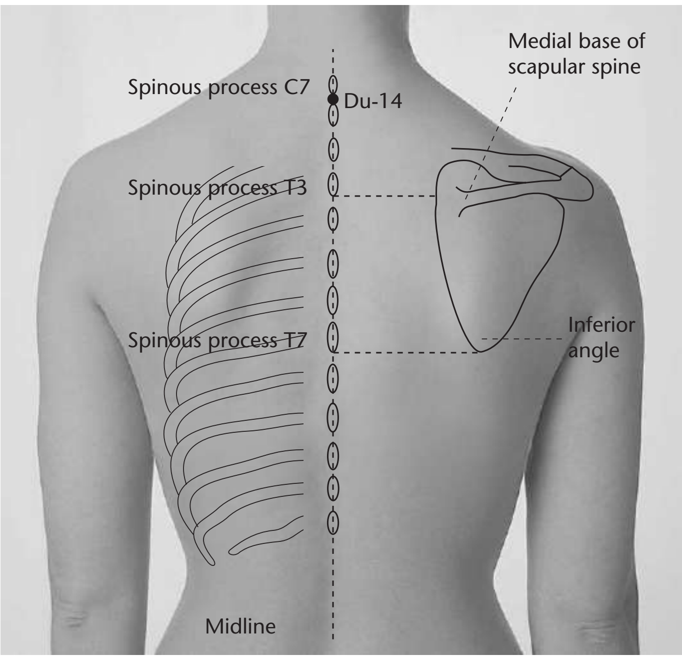
Spinous process C7
Du-14
Medial base of scapular spine
Spinous process T3
Spinous process T7
Inferior angle
Midline

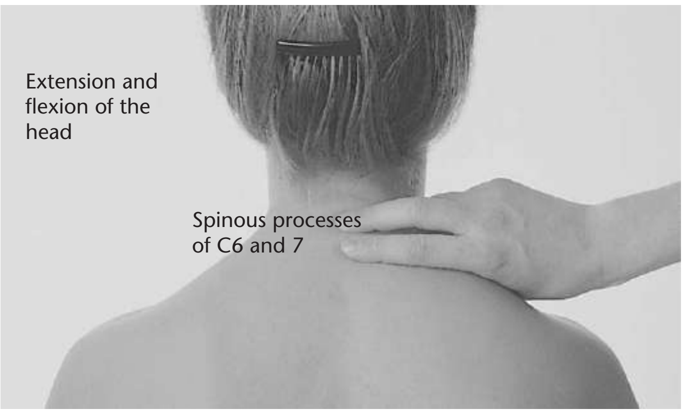
Extension and flexion of the head
Spinous processes of C6 and 7

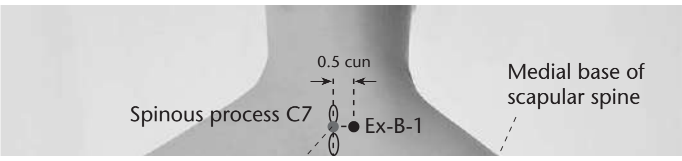
0.5 cun
Spinous process C7
Ex-B-1
Medial base of scapular spine

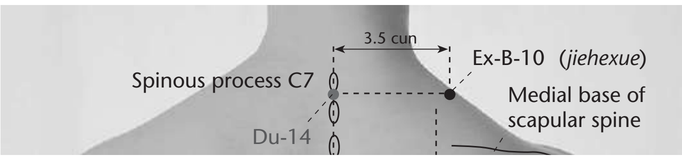
3.5 cun
Ex-B-10 (jiehexue)
Spinous process C7
Medial base of scapular spine
Du-14

Gate of Muteness YAMEN **Du-15**

Location

On the occiput, on the posterior midline, in the depression between the 1st (atlas) and 2nd (axis) cervical vertebrae, approximately 0.5 cun inferior to ➞ **Du-16** (directly below the external occipital protuberance).

How to find

The external occipital protuberance (➞ 3.1.5) is a hump-shaped projection on the posterior surface of the occipital bone, on the midline and slightly superior to the craniocervical junction (➞ 3.1.5). ➞ **Du-16** is located in a depression on the posterior midline that you can palpate directly inferior to the protuberance. From there, glide 0.5 cun in an inferior direction and locate **Du-15** superior to the first palpable spinous process (belonging to the axis; the atlas has no spinous process) and approximately 0.5 cun superior to the posteror hairline.

Located on the same level is ➞ **BL-10** (on the lateral aspect of the trapezius muscle).

Needling

0.5–1 cun strictly perpendicularly in an inferior direction. Caution: Injury to the cervical marrow. According to classic texts, moxibustion is contraindicated.

Actions/Indications

- Benefits the tongue, ears, neck and spine
- Dispels Wind
- Clears Yang Heat/excess

Special features

Meeting point with the *yang wei mai*. Sea of Qi point.

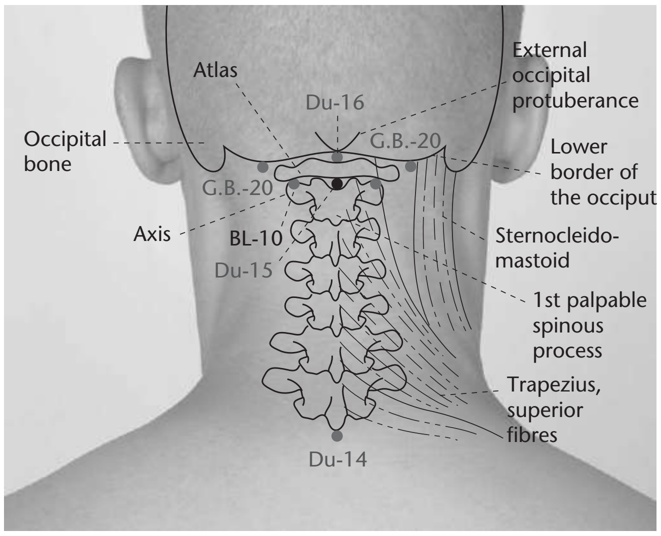
Atlas
External occipital protuberance
Du-16
Occipital bone
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid
Du-15
1st palpable spinous process
Trapezius, superior fibres
Du-14

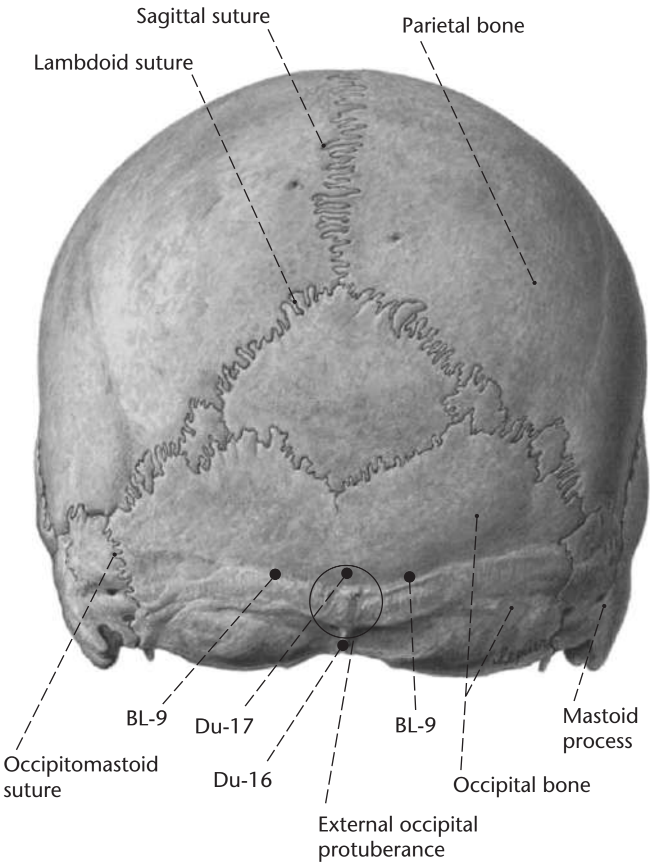
Sagittal suture
Parietal bone
Lambdoid suture
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

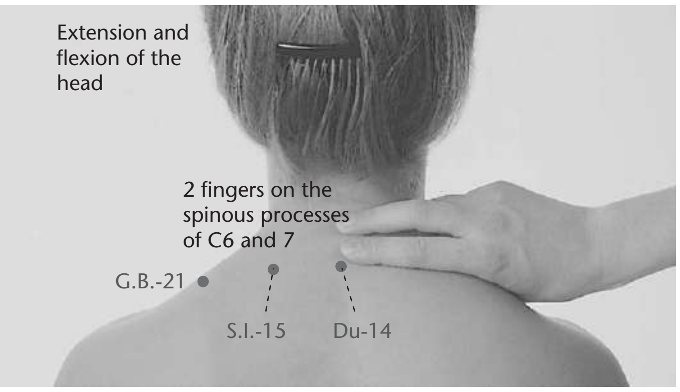
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

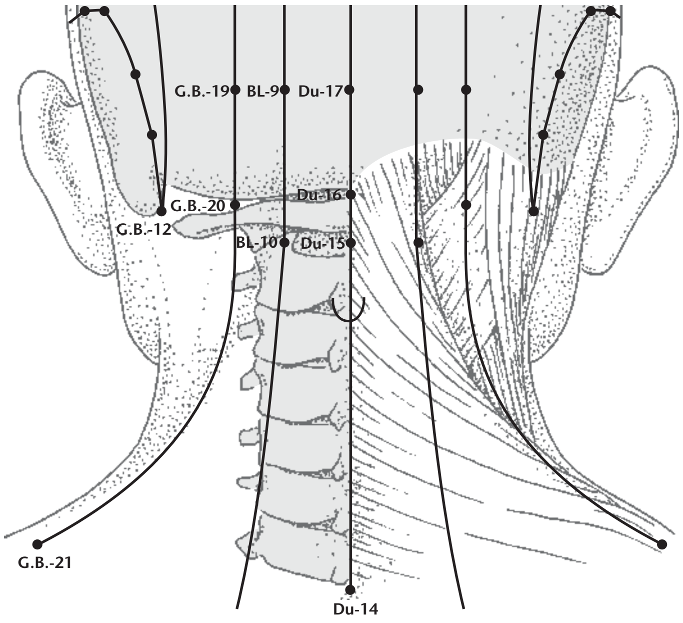
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15
G.B.-21
Du-14

Du-16 *Palace of Wind* FENGFU

Location

On the posterior midline, directly below the external occipital protuberance, in the depression between the origins of the trapezius muscle.

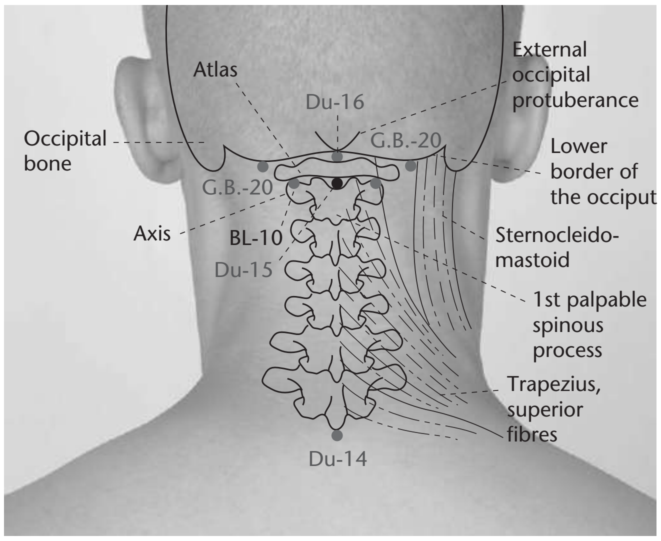
Atlas
External occipital protuberance
Du-16
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Du-15
Sternocleido-mastoid
1st palpable spinous process
Trapezius, superior fibres
Du-14

How to find

The external occipital protuberance (→ 3.1.5) is a hump-shaped projection on the posterior surface of the occipital bone, on the midline and slightly superior to the craniocervical junction (→ 3.1.5). Locate **Du-16** in a depression on the posterior midline that can be palpated directly inferior to the protuberance and between the two origins of the trapezius muscle. Generally, **Du-16** is located approximately 1 cun superior to the posterior hairline. → **G.B.-20** is located on the same level below the occiput, in a depression between the origins of the trapezius and sternocleidomastoid muscles.

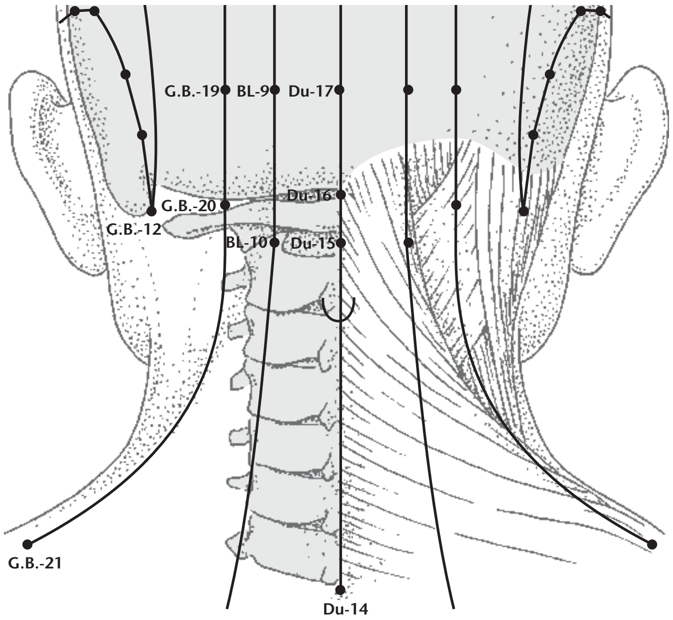
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15
G.B.-21
Du-14

Needling

0.5–1 cun slightly obliquely in an inferior direction. Bend the patient's head slightly forward when needling. Caution: Dangerous point with deep needling! No stimulation. Strictly no needling in a superior direction, danger of puncturing the cerebellomedullary cistern (also site for suboccipital puncture). The needle should be inserted into the nuchal ligament. According to some authors, moxibustion is contraindicated.

Actions/Indications

- Dispels (external) Wind
- Calms internal Wind and the *shen*
- Nourishes the Sea of Marrow, lightens the *shen*

Special features

Meeting point with the *yang wei mai* and, according to some authors, the *yang qiao mai*. Sea of Marrow point, Window of Heaven point, Sun Si Miao Ghost point. Used for all disorders which are caused by 'external or internal' Wind and which are characterised by sudden onset, changing location and/or intensity of the disorder, involuntary movements.

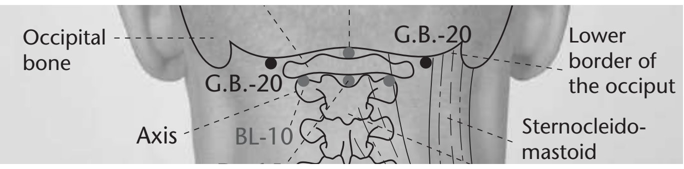
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Sternocleido-mastoid

Brain's Door **NAOHU** **Du-17**

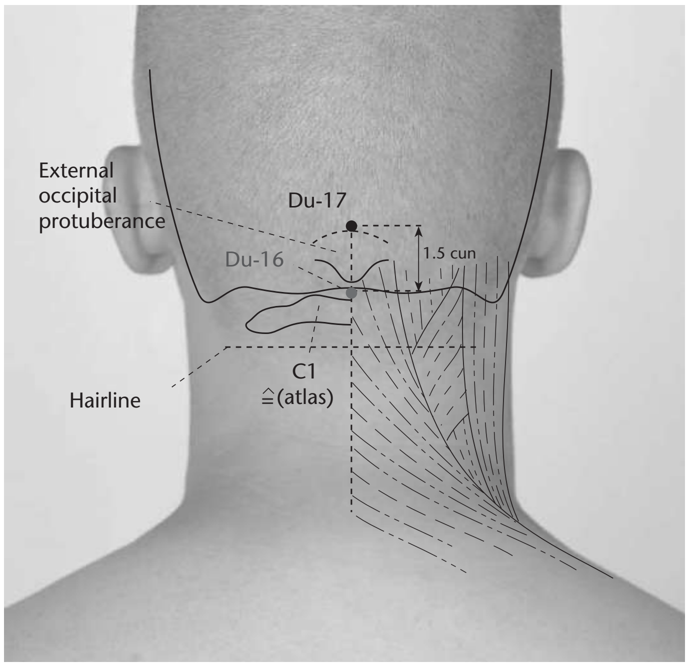
External occipital protuberance
Du-17
1.5 cun
Du-16
C1
≙(atlas)
Hairline

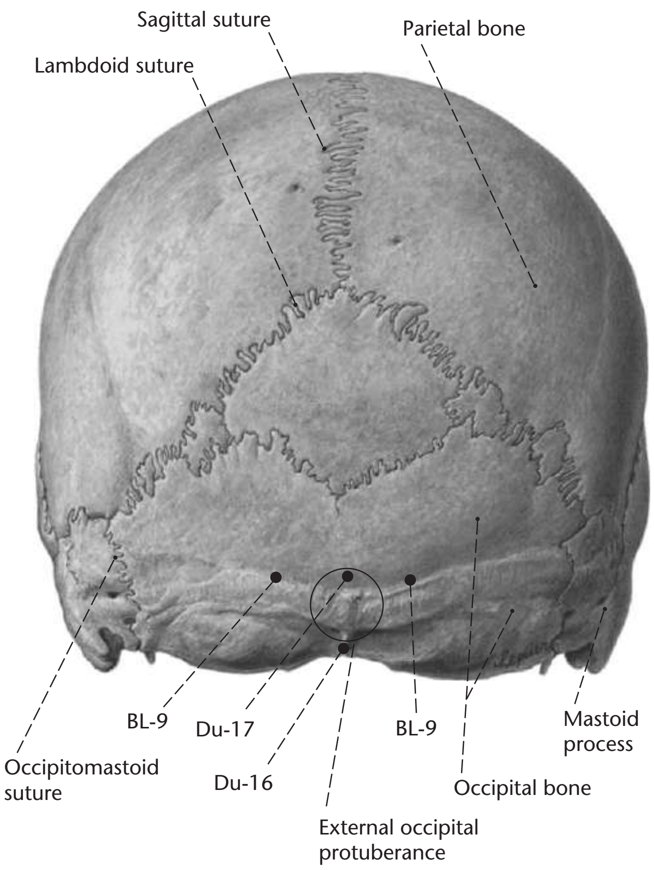
Sagittal suture
Parietal bone
Lambdoid suture
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

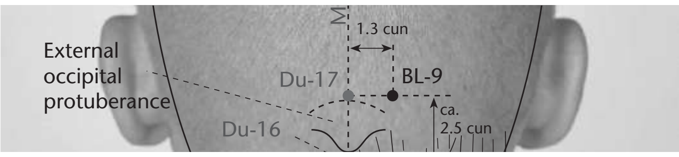
M
1.3 cun
External occipital protuberance
Du-17
BL-9
Du-16
ca. 2.5 cun

Location

In a depression superior to the external occipital protuberance, approximately 2.5 cun superior to the posterior hairline or 1.5 cun superior to ➞ **Du-16**.

How to find

First, locate the external occipital protuberance (➞ 3.1.5), which forms a hump-shaped projection on the posterior surface of the occipital bone. Next, locate **Du-17** on the posterior midline, in a depression directly superior to the protuberance. This corresponds to a distance of approximately 2.5 cun superior to the posterior hairline (➞ 3.1.5).

➞ **Du-16** is located directly inferior to the protuberance, whereas ➞ **BL-9** is located 1.3 cun lateral to **Du-17**.

Needling

0.5–1 cun transversely (subcutaneously) in an inferior or superior direction. Moxibustion controversial.

Actions/Indications

- Dispels Wind
- Opens the channel
- Benefits the eyes
- Calms the *shen*

Special features

Meeting point with the BL channel

Du-18 *Unyielding Space* QIANGJIAN

Location

On the posterior midline, 1.5 cun superior to **Du-17** (directly superior to the external occipital protuberance) or 3 cun inferior to → **Du-20** (on the vertex).

How to find

First, locate → **Du-17** in a depression directly superior to the external occipital protuberance (→ 3.1.5), which forms a hump-shaped projection on the posterior surface of the occipital bone. From there, palpate 1.5 cun in a superior direction and there locate **Du-18** on the posterior midline.

→ **Du-20** is located 3 cun in a superior direction, on the vertex.

Needling

Transversely (subcutaneously) 0.5–1 cun

Actions/Indications

- Dispels Wind, especially internal Wind
- Opens the channel
- Calms the *shen*

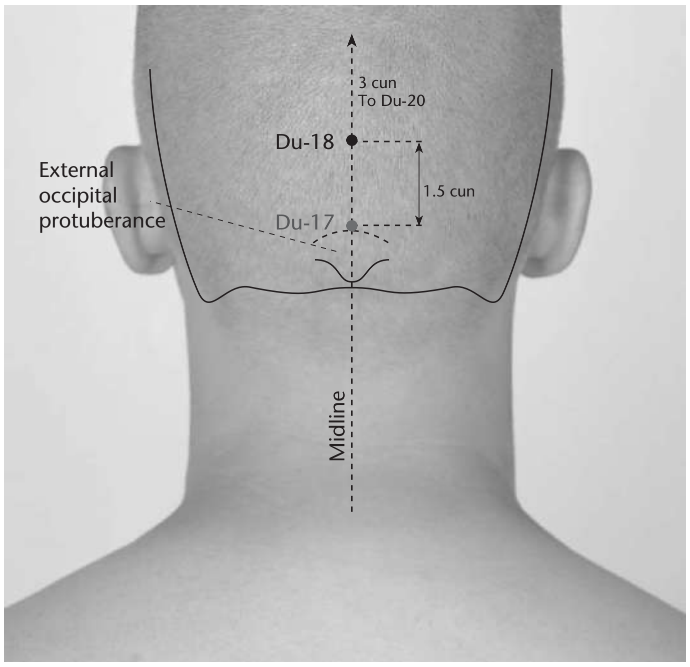
3 cun
To Du-20
Du-18
External occipital protuberance
1.5 cun
Du-17
Midline

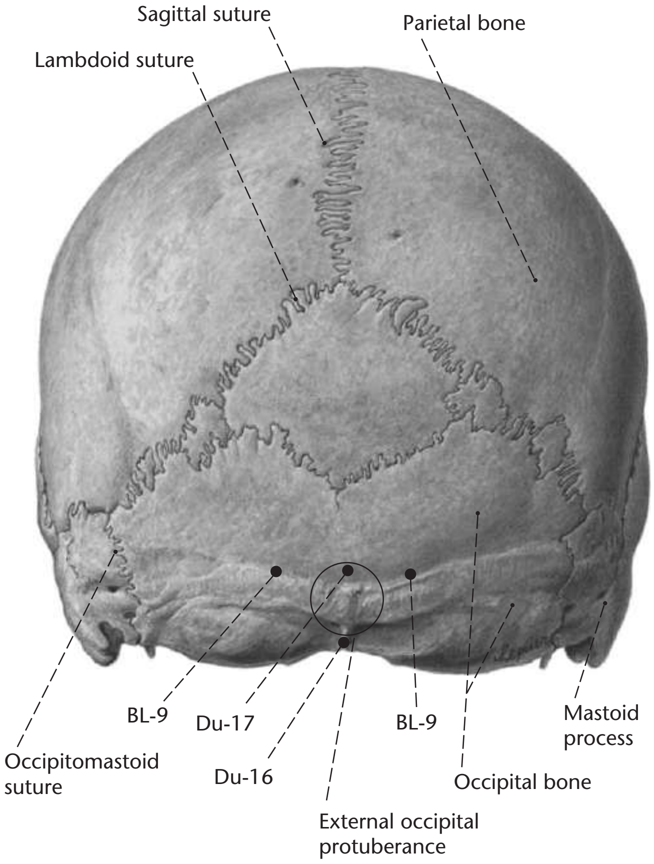
Sagittal suture
Parietal bone
Lambdoid suture
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

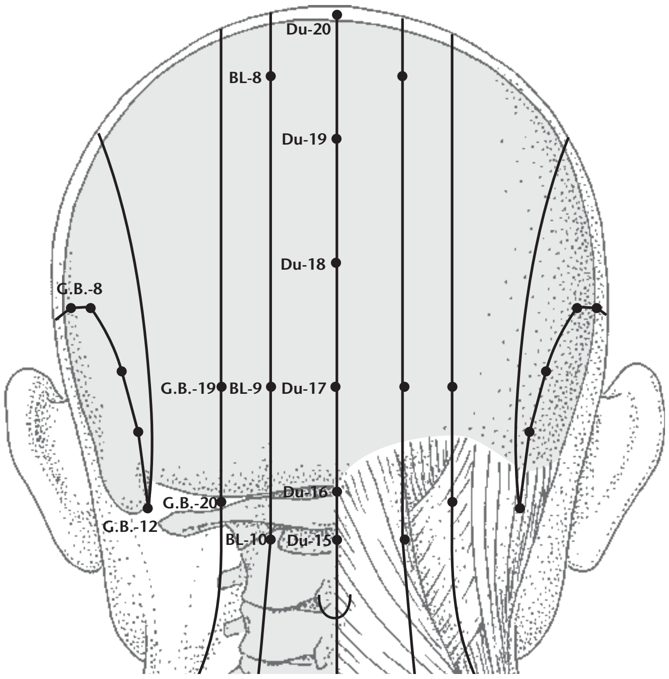
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

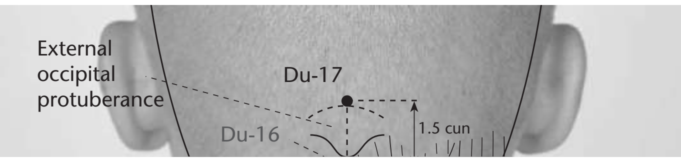
External occipital protuberance
Du-17
Du-16
1.5 cun

Behind the Crown HOUDING Du-19

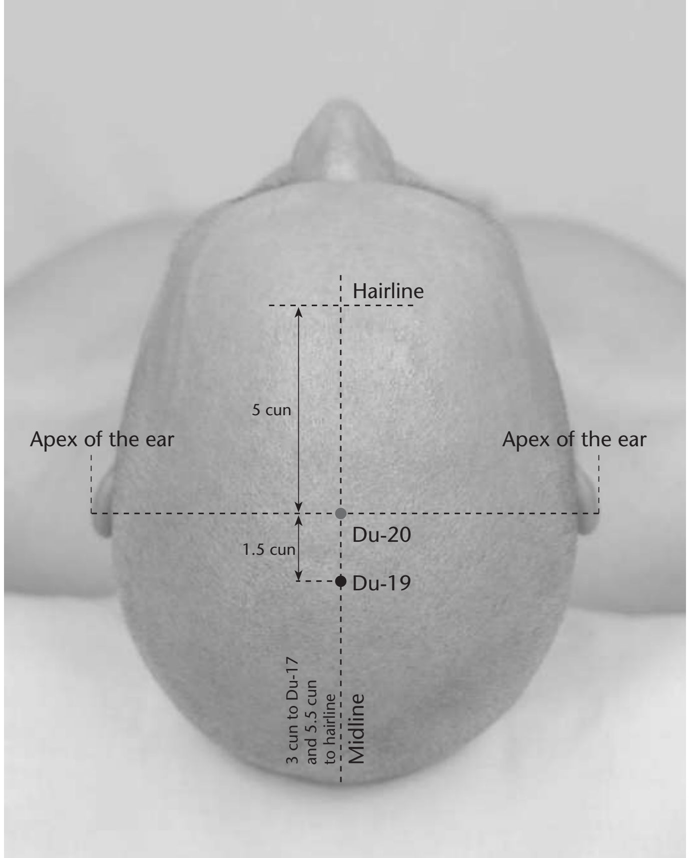
Hairline
5 cun
Apex of the ear
Apex of the ear
Du-20
1.5 cun
Du-19
3 cun to Du-17
and 5.5 cun
to hairline
Midline

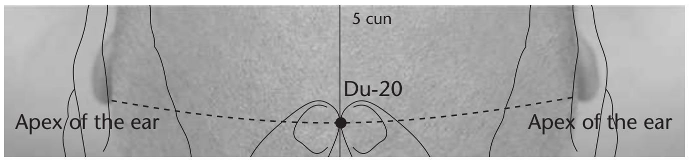
5 cun
Du-20
Apex of the ear
Apex of the ear

Location

On the midline, 3 cun superior to → **Du-17** (directly superior to the external occipital protuberance) or 1.5 cun posterior to → **Du-20**.

How to find

This point is best located by using → **Du-20** as a reference point (at the junction of the vertical midline and a line connecting the apices of the ears). **Du-19** can then be located by palpating 1.5 cun in a posterior direction from → **Du-20**. **Or:** First, locate → **Du-17** superior the external occipital protuberance (→ 3.1.5), then locate **Du-19** on the midline, 3 cun superior to → **Du-17**.

Needling

Transversely (subcutaneously) 0.5–1 cun.

Actions/Indications

- Dispels Wind
- Opens the channel
- Calms the *shen*

Du-20 *Hundred Meetings* BAIHUI

Location

At the junction of a line connecting the apices of the ears and the midline, 5 cun from the anterior or 7 cun from the posterior hairline respectively.

How to find

Spreading hands technique (→ 2.3.3): Place the hands on both sides of the head, with the little fingers touching the apices of the ears. Join the thumbs on the midline and locate **Du-20** in shallow depression on the vertex of the head (alternative location method: use an elastic tape, with the midpoint marked on it). The symbolic counterpart to **Du-20** (as the highest point on the body and in contact with heaven) is → **KID-1**, the lowest part of the body and in contact with the earth.

Needling

0.5–1 cun transversely (subcutaneously) towards the posterior (reducing effect) or the anterior (tonifying effect). Caution: Apply pressure to head points after needle removal to avoid bleeding. Remember to remove the needle after treatment.

Actions/Indications

- Calms Wind, pacifies Yang, benefits the Brain and sensory organs, calms the *shen*
- Nourishes the Sea of Marrow; applying moxibustion raises the Yang

Special features

Meeting point with the BL, G.B., T.B. and LIV channels; Sea of Marrow point. Important point for descending the Yang (reducing needle techniques) or raising the Yang (tonifying needle techniques, moxibustion).

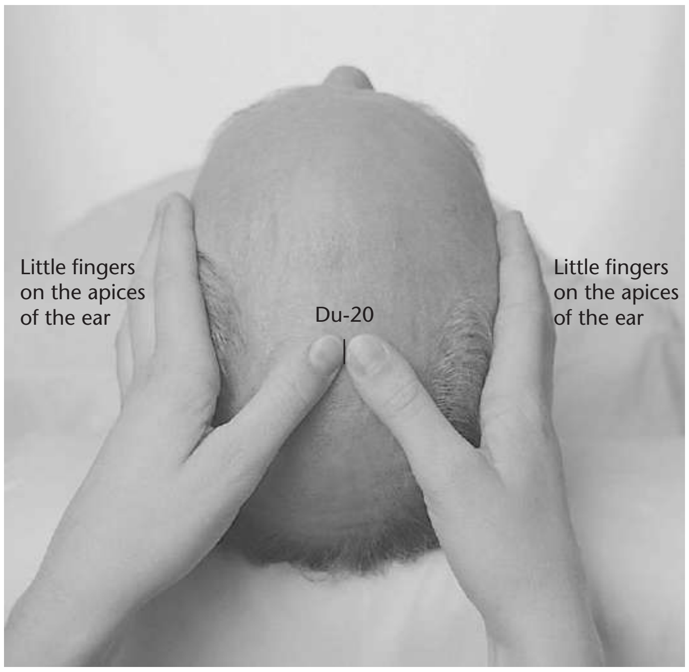
Little fingers on the apices of the ear
Du-20
Little fingers on the apices of the ear

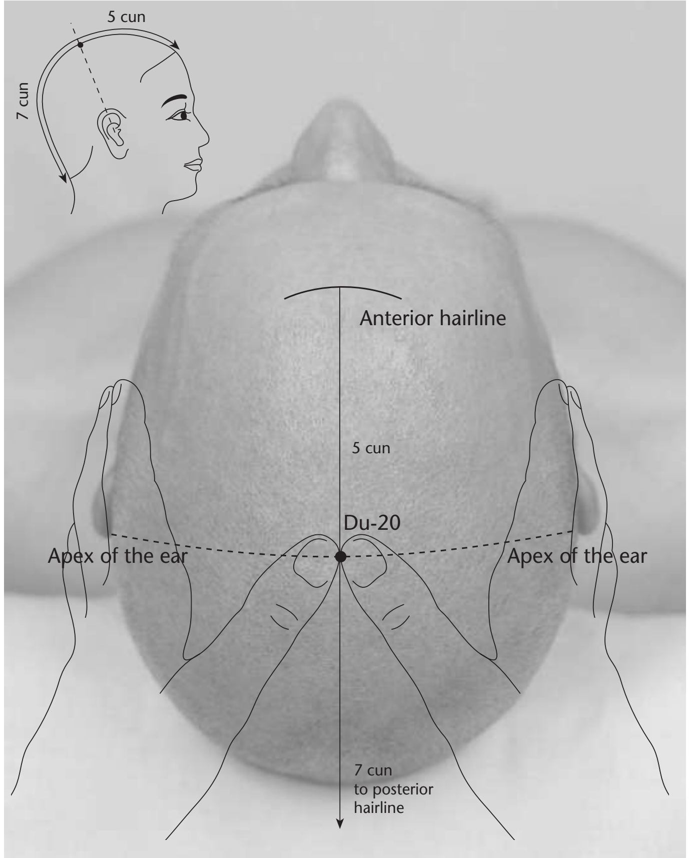
5 cun
7 cun
Anterior hairline
5 cun
Du-20
Apex of the ear
Apex of the ear
7 cun to posterior hairline

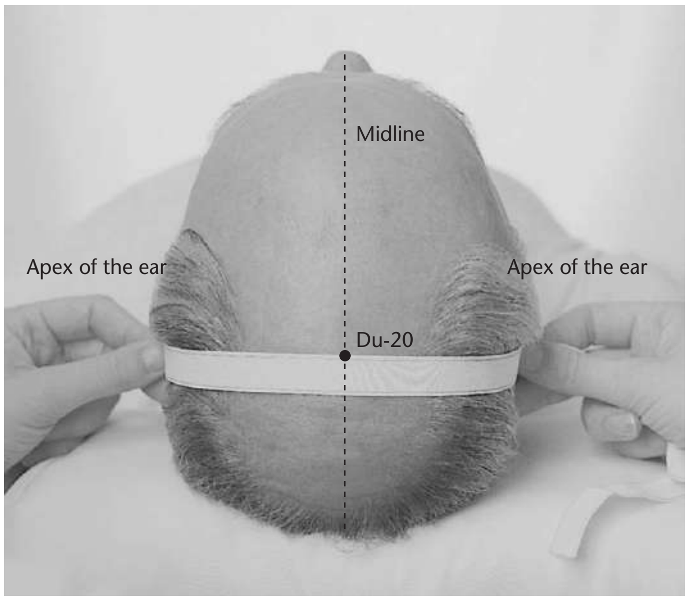
Midline
Apex of the ear
Apex of the ear
Du-20

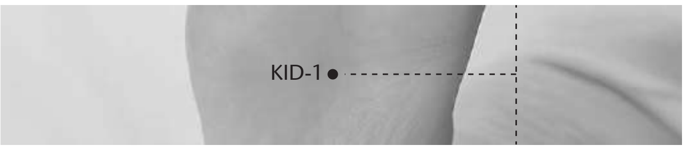

In Front of the Crown QIANDING **Du-21**

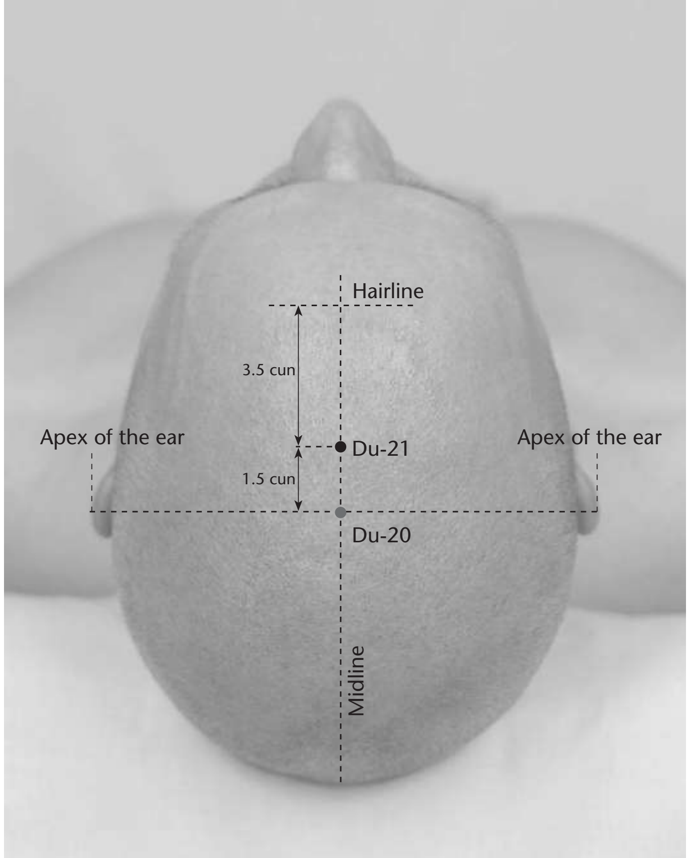
Hairline
3.5 cun
Apex of the ear
Du-21
Apex of the ear
1.5 cun
Du-20
Midline

Location

On the midline, 3.5 cun superior to the anterior hairline or 1.5 cun anterior to → **Du-20**.

How to find

First, locate → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears) and, from there, measure 1.5 cun in an anterior direction. **Or:** Locate the anterior hairline (→ 3.1.1; the distance from the anterior hairline to → **Du-20** is 5 cun) and palpate 3.5 cun in a superior/posterior direction. There, locate **Du-21** on the midline.

Needling

Transversely (subcutaneously) 0.5–1 cun. Caution with infants whose fontanelle has not yet closed. Moxibustion possible.

Actions/Indications

- Dispels Wind and Dampness
- Opens the channel
- Calms the *shen*

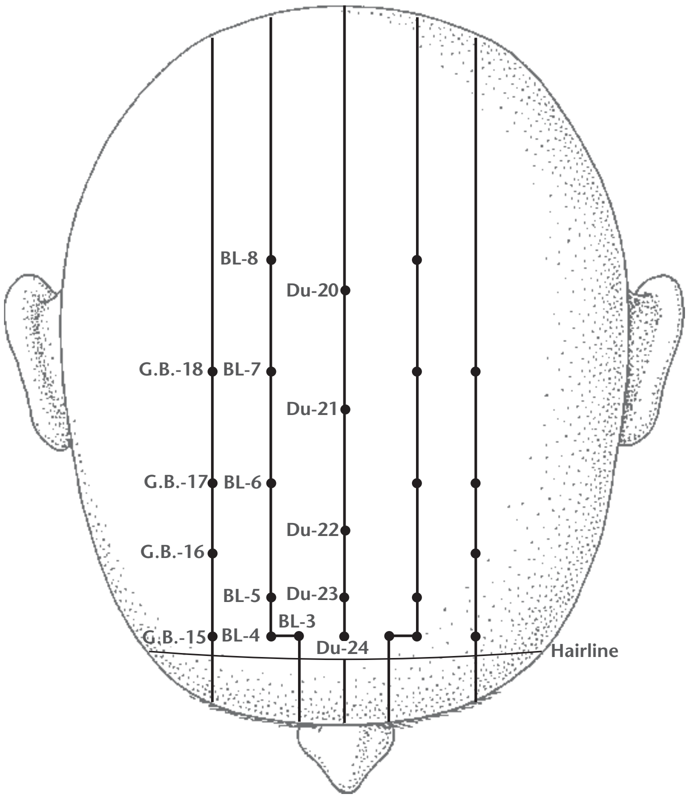
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

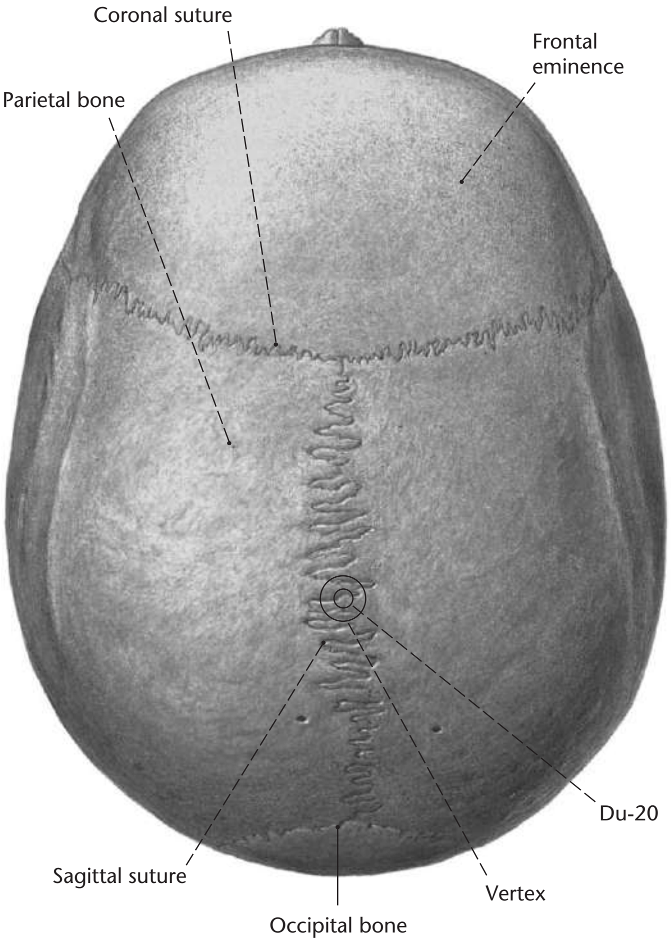
Coronal suture
Frontal eminence
Parietal bone
Du-20
Sagittal suture
Vertex
Occipital bone

Du-22 *Fontanelle Meeting* XINHUI

Location

On the midline, 2 cun superior to the anterior hairline.

How to find

First, locate the anterior hairline (→ 3.1.1) and, from there, palpate 2 cun in a superior direction. There, locate **Du-22** on the midline. For reference: The distance from the anterior hairline to → **Du-20** is 5 cun. **Or:** From → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears), measure 3 cun in an anterior direction and there locate **Du-22** on the midline.

Needling

Transversely (subcutaneously) 0.5–1 cun. Caution with infants whose fontanelle has not yet closed (**Du-22** is located on the anterior border of the fontanelle). Moxibustion possible.

Actions/Indications

- Dispels Wind
- Calms the *shen*

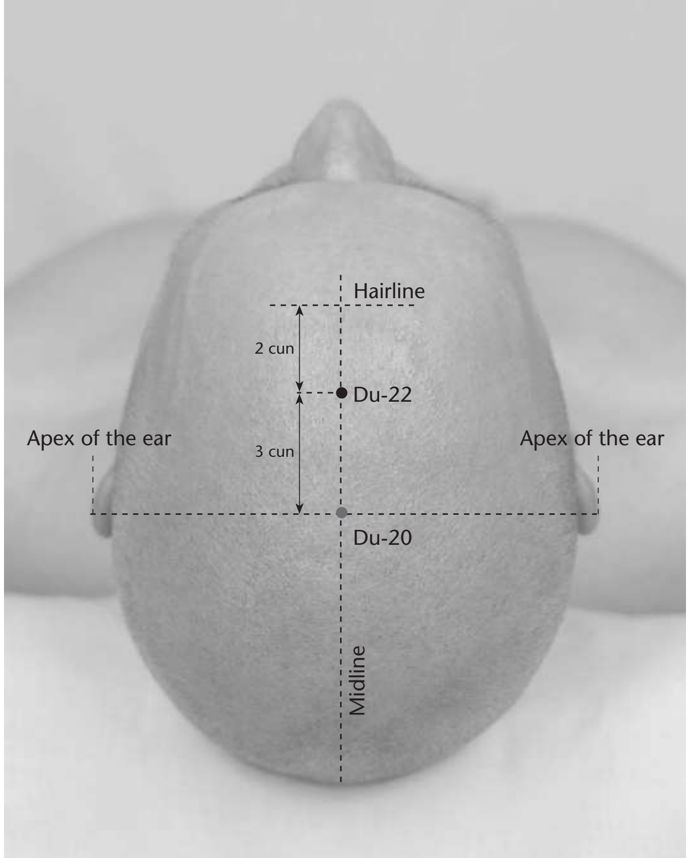
Hairline
2 cun
Du-22
3 cun
Apex of the ear
Apex of the ear
Du-20
Midline

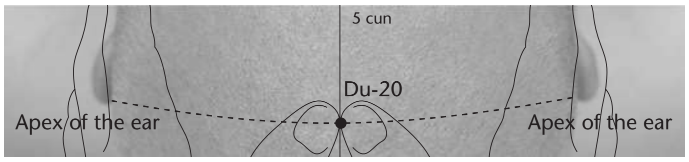
5 cun
Du-20
Apex of the ear
Apex of the ear

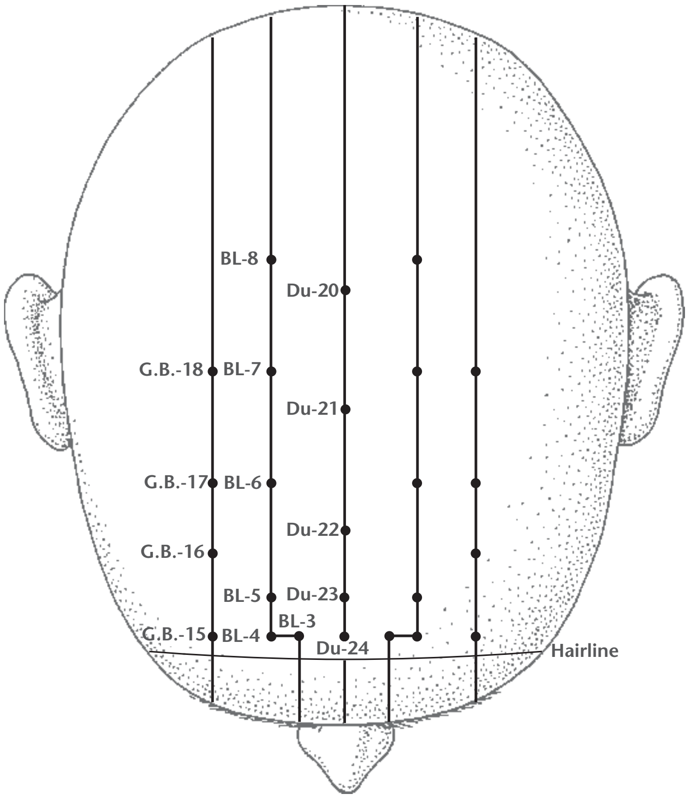
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

Upper Star SHANGXING Du-23

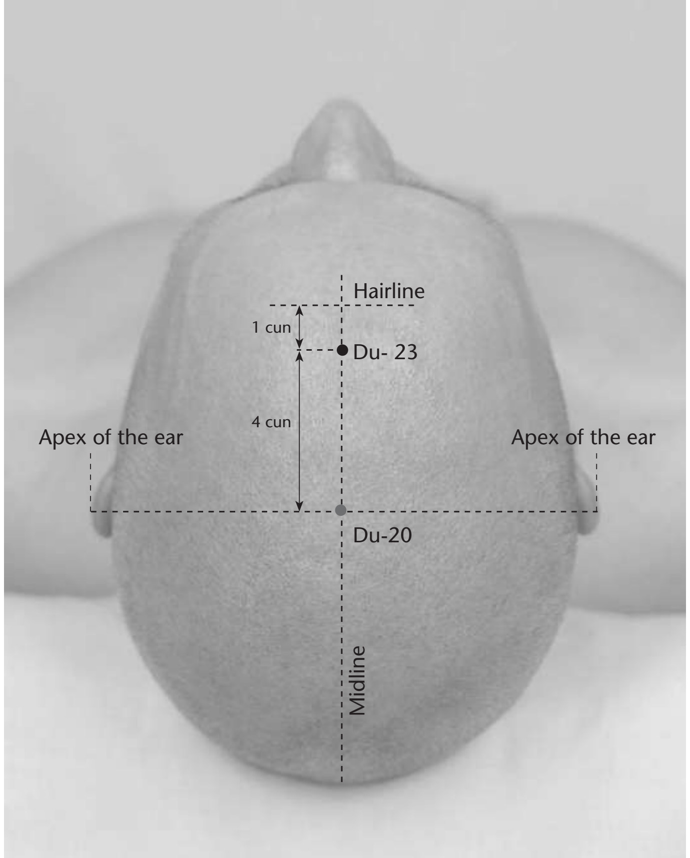
Hairline
1 cun
Du- 23
4 cun
Apex of the ear
Apex of the ear
Du-20
Midline

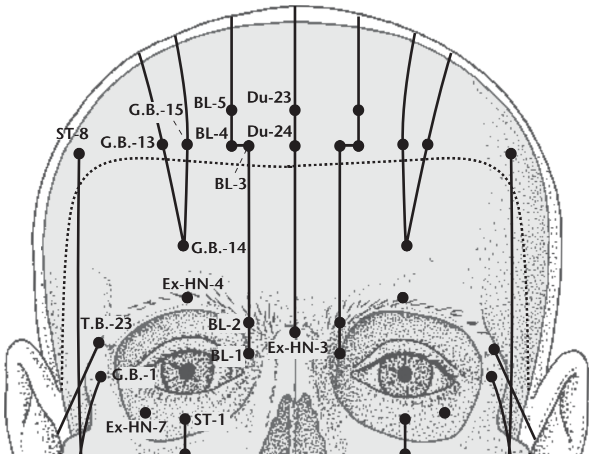
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1

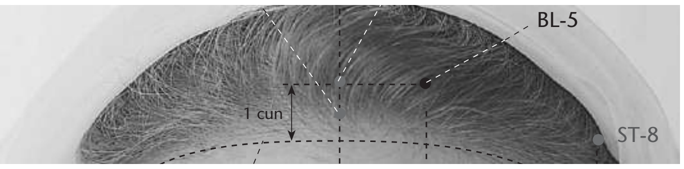
BL-5
1 cun
ST-8

Location

On the midline, 1 cun superior to the anterior hairline or 4 cun anterior to → **Du-20**.

How to find

First, locate the anterior hairline (→ 3.1.1) and, from there, palpate 1 cun in a superior direction. There, locate **Du-23** on the midline. For reference: The distance from the anterior hairline to → **Du-20** is 5 cun.

Located on the same level is → **BL-5** (1.5 cun lateral to the midline).

Needling

Transversely (subcutaneously) 0.3–0.5 cun. Moxibustion possible.

Actions/Indications

- Dispels pathogenic factors (especially Wind and Heat) from the nose
- Dispels pathogenic factors (especially Heat) from the face
- Calms the *shen*

Du-24 *Courtyard of the Spirit* SHENTING

Location

On the midline, 0.5 cun superior to the anterior hairline or 4.5 cun anterior to → **Du-20**.

How to find

First, locate the anterior hairline (→ 3.1.1) and, from there, palpate 0.5 cun in a superior direction. There, locate **Du-24** on the midline. For reference: The distance from the anterior hairline to → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears) is 5 cun.
Located on the same level (0.5 cun superior to the anterior hairline) are → **BL-3/BL-4/G.B.-15/G.B.-13** (superior to the inner canthus of the eye/1.5 cun lateral to the midline/on the pupil line or 2.25 cun lateral to the midline/3 cun lateral to the midline). These distances refer to the proportional cun distance (→ 2.2) between **Du-24** and → **ST-8** (on the corner of the forehead), which equals 4.5 cun (1.5 cun lateral to the midline).

Needling

Transversely (subcutaneously) 0.3–0.5 cun. Moxibustion possible.

Actions/Indications

- Calms the *shen*
- Dispels (internal) Wind
- Clears pathogenic factors from the eyes and nose

Special features

Meeting point with the BL and ST channels. One of the major points for calming the *shen*.

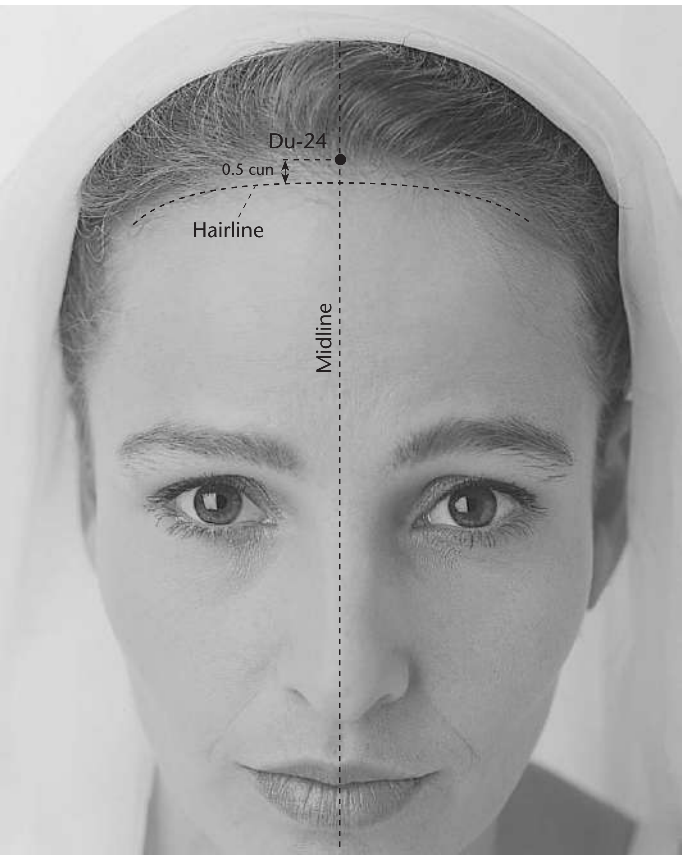
Du-24
0.5 cun
Hairline
Midline

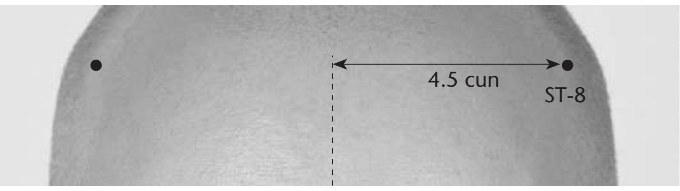

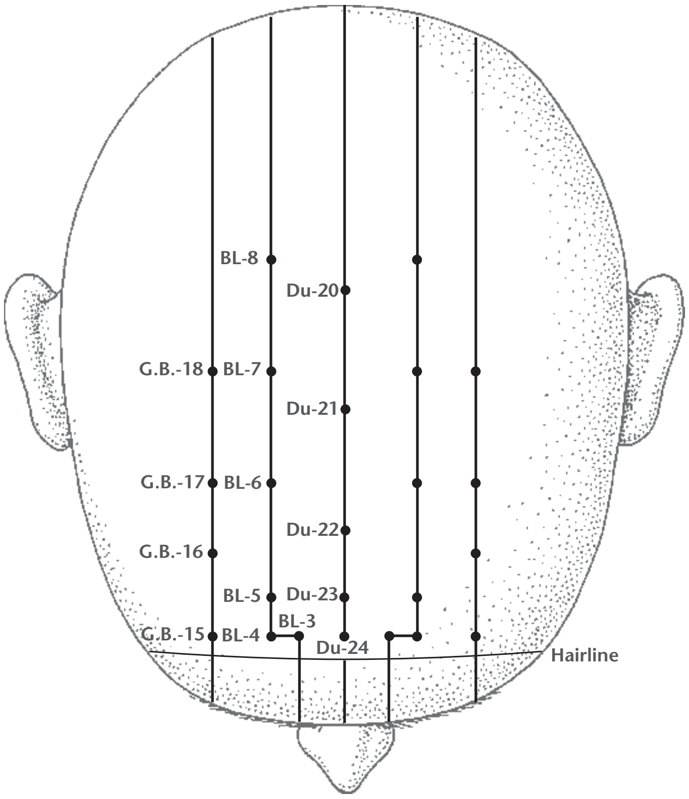
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

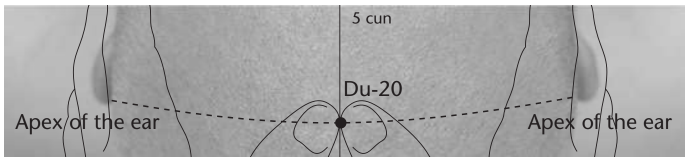
5 cun
Du-20
Apex of the ear
Apex of the ear

Ch05b-F10028.qxd 2/22/08 9:56 PM Page 523

White Crevice SULIAO Du-25

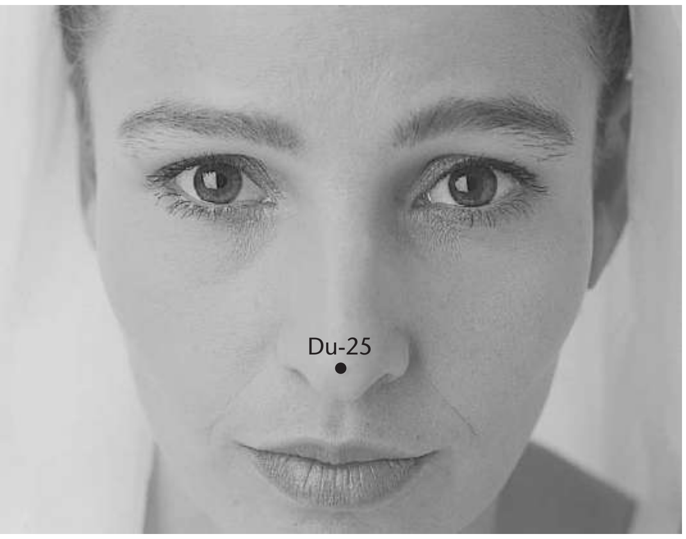

Location

In a depression on the tip of the nose.

How to find

As the name implies, **Du-25** can be located in a depression on the tip of the nose, which generally can be easily palpated.

Needling

0.2 cun vertically or transversely (subcutaneously) in a superior direction up to 1 cun. Or prick to bleed. Moxibustion controversial.

Actions/Indications

- Benefits the nose

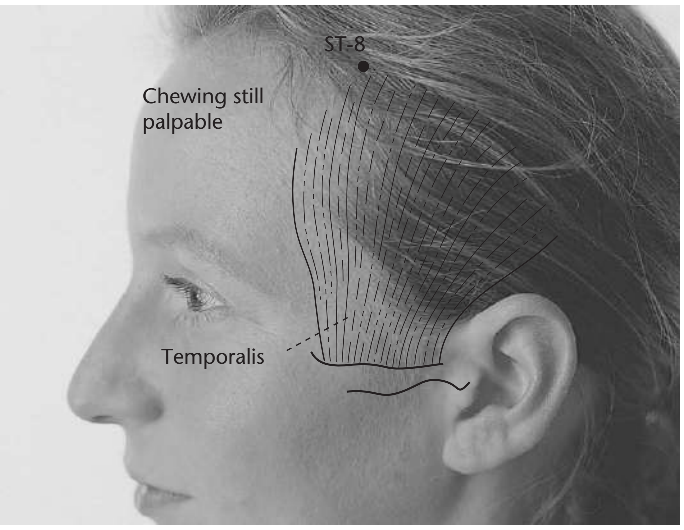
ST-8
Chewing still palpable
Temporalis

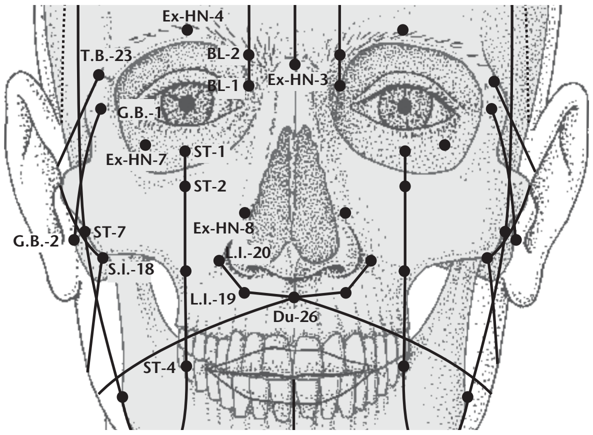
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
G.B.-2
ST-7
Ex-HN-8
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4

Du-26 *Man's Middle* RENZHONG

Location

Below the nose, on the upper third of the philtrum. Note: *shuigou* (Water Grave) is an alternative name for this point.

How to find

Locate the philtrum, which forms a distinct groove on the midline between the root of the nose and the margin of the upper lip. Locate **Du-26** sligthly superior to the midpoint of the philtrum. → **L.I.-19** is located on the same level, 0.5 cun lateral to the midline.

Needling

0.3–0.5 cun obliquely in a superior direction. Caution: Needling this point may be painful.

Actions/Indications

- Revives consciousness
- Benefits the face and nose, eliminates (external) Wind
- Benefits the spine

Special features

Meeting point with the L.I. and ST channels; Sun Si Miao Ghost point. Major point for acute emergencies: Needle with strongly reducing technique. If no acupuncture needle is available, use a syringe or perform strong acupressure with the nail of the thumb or index finger, supporting the patient's chin with the same hand. Can be used in cases of needle collapse – first remove all other needles.

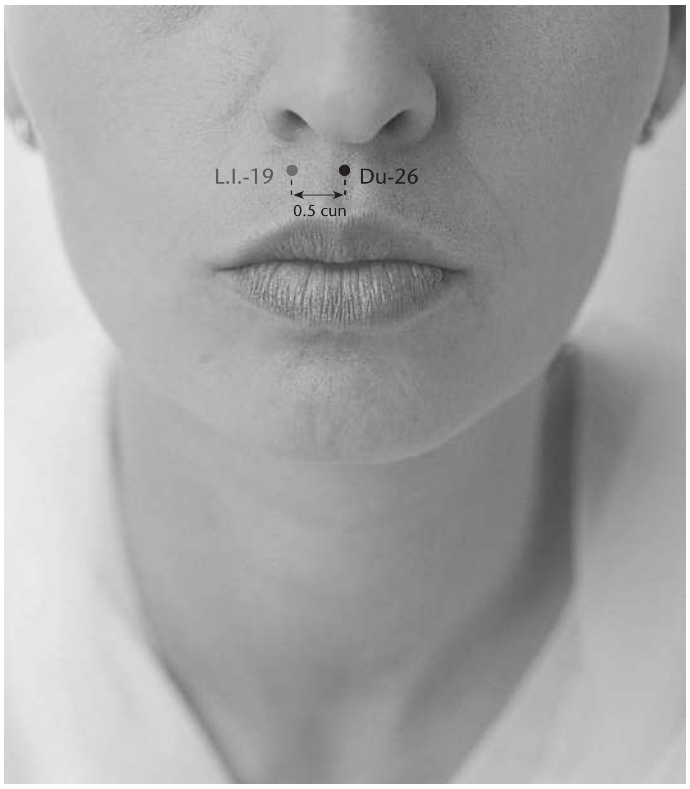

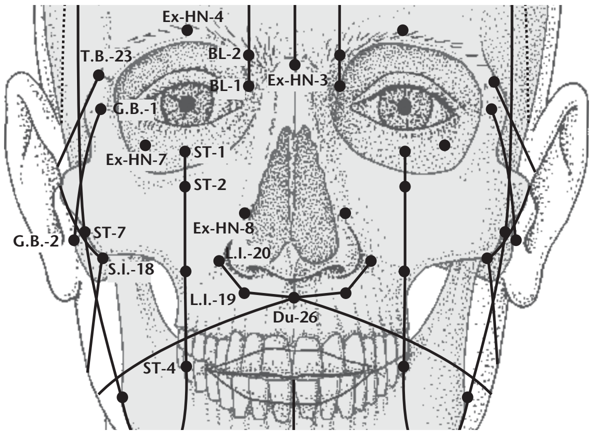

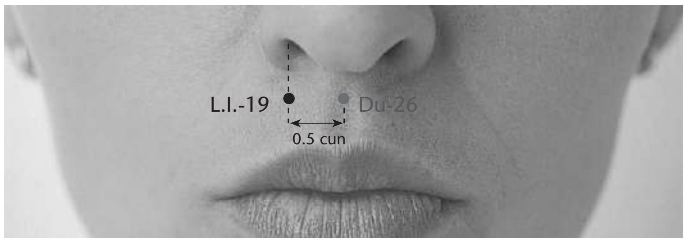

Extremity of the Mouth DUIDUAN **Du-27**

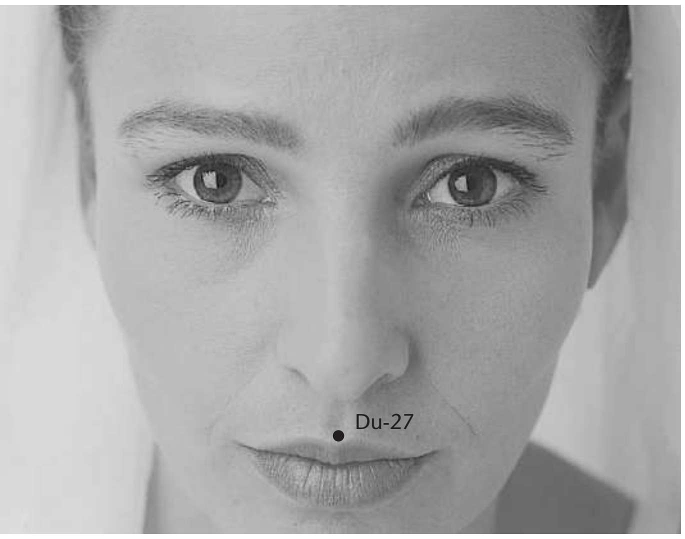

Location

On the midline, on the margin of the upper lip and the philtrum.

How to find

Location help: As the name implies, **Du-27** is located on the midline, at the junction of the upper lip and the philtrum.

Needling

Up to 0.3 cun obliquely in a superior direction. No moxibustion.

Actions/Indications

- Clears Heat
- Moistens the body
- Local point for the mouth
- Calms the *shen*

Du-28 *Gum Intersection* YINJIAO

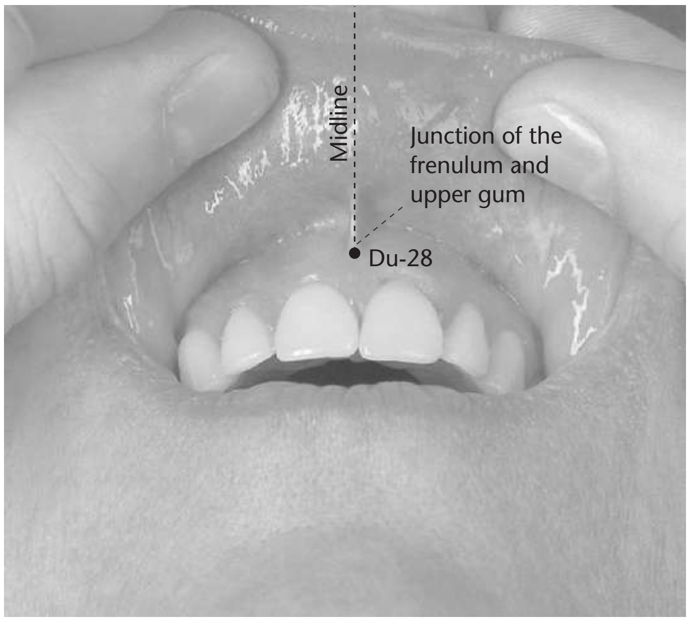

Location

On the inside of the upper lip, at the junction of the frenulum and the upper gum.

How to find

Hold the upper lip in a raised position in order to reveal the frenulum and the upper gum. Locate **Du-28** on the midline, at the junction of the frenulum with the upper gum.

Needling

0.1–0.2 cun obliquely in a superior direction or prick to bleed. Do not puncture the frenulum.

Actions/Indications

- Clears Heat (especially from the eyes and the mouth)

Special features

As the name implies, meeting point with the *ren mai* and the ST channel

5.4 *Dai mai*

Synonym: Girdling Vessel
The extraordinary vessel *dai mai* is the only vessel (or channel) with a horizontal pathway (except for *luo*-connecting vessels).

Relationships (→ 1.7.3)

- **Yin/Yang:** *chong mai/dai mai*
- **Central/peripheral:** *dai mai/yang wei mai*. Regions supported by this pairing: lateral eye region as well as temporal region, ears, cheeks, occiput and shoulders. Opening point: **G.B.-41** (*zulinqi*), coupled point: **T.B.-5** (*waiguan*).

Pathway

The extraordinary vessel *dai mai* originates in the hypochondrium at the level of L2; according to many other authors it originates at → **LIV-13** (*zhangmen*) at the free end of the 11th rib. It encircles the waist of the body like a belt. As it does so it passes → **G.B.-26** (*dai mai*) at the level of the umbilicus and below the free end of the 11th rib, → **G.B.-27** (*wushu*) and → **G.B.-28** (*weidao*).

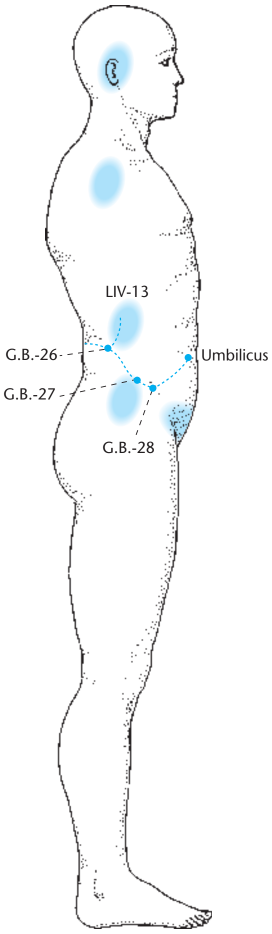

Meeting points with other channels

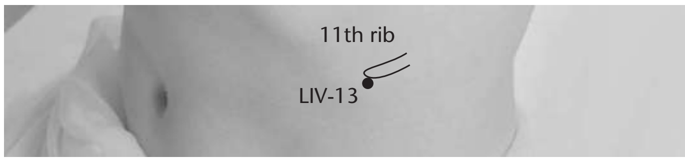

LIV-13 (*zhangmen*): Anterior and inferior to the free end of the 11th rib

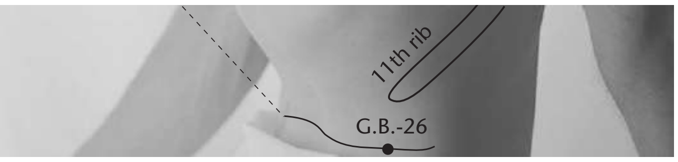

G.B.-26 (*daimai*): On the lateral aspect of the waist, at the junction of a vertical line through the free end of the 11th rib and a horizontal line through the umbilicus, approximately 1.8 cun inferior to → **LIV-13**

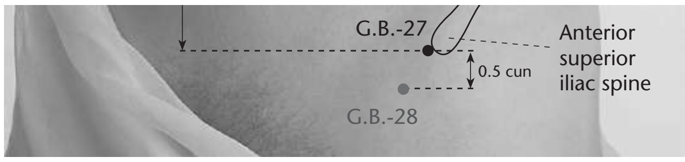

G.B.-27 (*wushu*): In the depression medial to the anterior superior iliac spine, approximately level with Ren-4

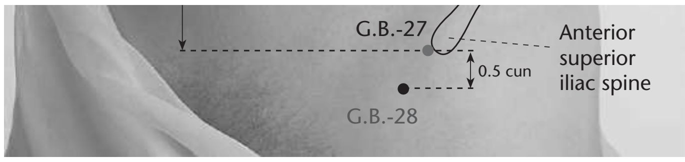

G.B.-28 (*weidao*): Anterior and inferior to the anterior superior iliac spine, 0.5 cun anterior and inferior to G.B.-27

Connections with other channels/Organs

- Gall Bladder and Liver primary channels, Kidney divergent channel
- Kidney, Uterus

Clinical importance (→ 1.7.2, 1.7.3)

- **Controls** and **encircles** the horizontal channels like a belt
- **Connects** the upper and lower halves of the body at the waist
- **Regulates** vaginal discharge
- **Eliminates** Damp-Heat from the Lower Burner, especially from the genitalia
- **Controls** the *shaoyang* channels, especially the G.B. channel

5.5 *Yin wei mai*

Synonym: Yin Motility Vessel (Deadman), **Yin Heel Vessel** (Maciocia)

Relationships (→ 1.7.3)

- **Yin/Yang:** *yin wei mai*/*yang wei mai*
- **Central/peripheral:** *chong mai*/*yin wei mai*. Regions supported by this pairing: Heart, thorax, Stomach. Opening point: **P-6** (*neiguan*), coupled point: **SP-4** (*gongsun*).

Pathway

The extraordinary vessel *yin wei mai* originates in the hypochondrium at → **KID-9** (*zhubin*), 5 cun proximal to the prominence of the medial malleolus, ascends the medial aspect of the leg to the inguinal region, follows the Spleen and Liver primary channels, passing the points → **SP-12** (*chongmen*) and → **SP-13** (*fushe*). It then passes → **SP-15** (*daheng*) and → **SP-16** (*fuai*) and again meets with the Spleen and Liver primary channels at → **LIV-14** (*qimen*). The vessel then traverses the thoracic region, passes → **Ren-22** (*tiantu*) in the suprasternal fossa and follows the extraordinary vessel *ren mai* to → **Ren-23** (*lianqiao*) where it terminates.

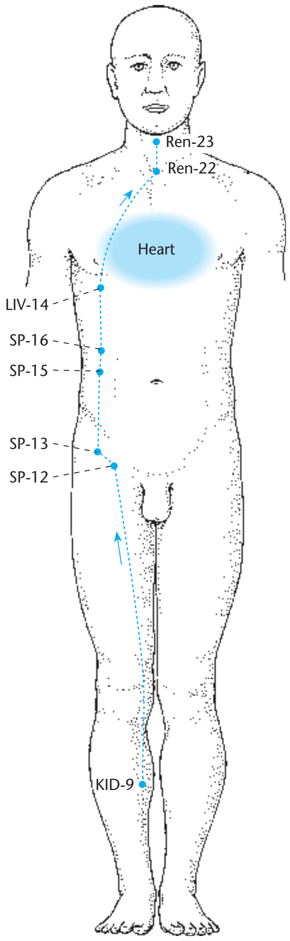
Ren-23
Ren-22
Heart
LIV-14
SP-16
SP-15
SP-13
SP-12
KID-9

Meeting points with other channels

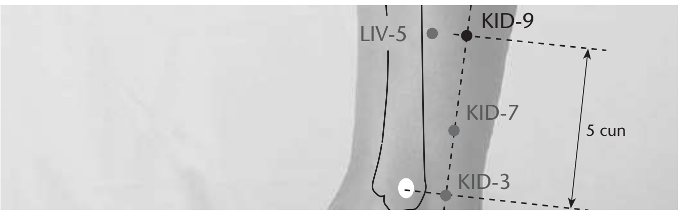
LIV-5
KID-9
KID-7
5 cun
KID-3

KID-9 (*zhubin*): 5 cun proximal to KID-3 and 2 cun posterior to the medial border of the tibia

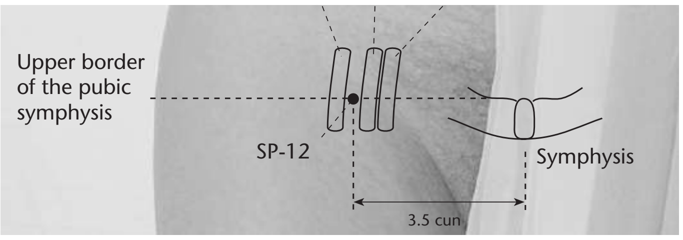
Upper border of the pubic symphysis
SP-12
Symphysis
3.5 cun

SP-12 (*chongmen*): 3.5 cun lateral to the midline, level with the upper border of the pubic symphysis, lateral to the femoral artery

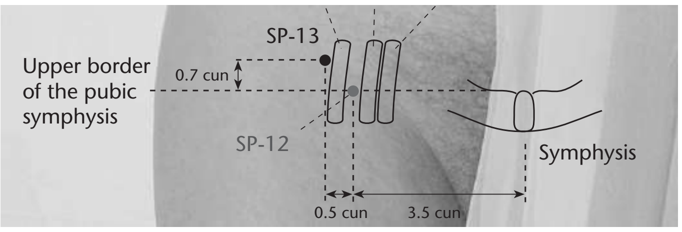
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

SP-13 (*fushe*): 4 cun lateral to the anterior midline (mamillary line) and 0.7 cun superior to the upper border of the pubic symphysis

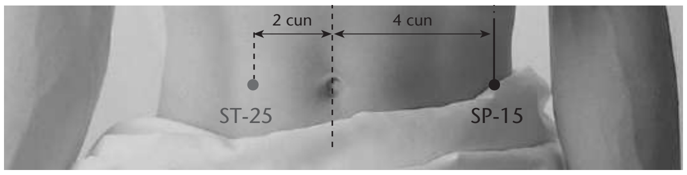
2 cun
4 cun
ST-25
SP-15

SP-15 (*daheng*): 4 cun lateral to the centre of the umbilicus

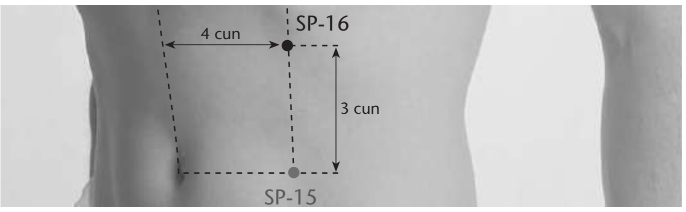
4 cun
SP-16
3 cun
SP-15

SP-16 (*fuai*): 3 cun superior to the centre of the umbilicus and 4 cun lateral to the anterior midline, on the mamillary line

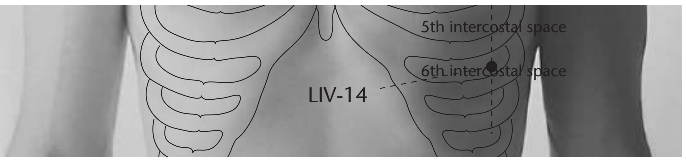
5th intercostal space
6th intercostal space
LIV-14

LIV-14 (*qimen*): In the 6th intercostal space, on the mamillary line, 4 cun lateral to the midline.

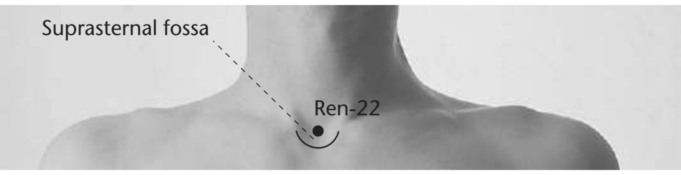
Suprasternal fossa
Ren-22

Ren-22 (*tiantu*): 0.5 cun superior to the sternum, in the centre of the suprasternal fossa

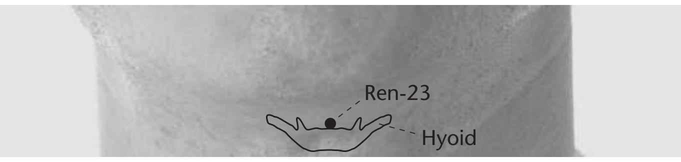
Ren-23
Hyoid

Ren-23 (*lianquan*): On the anterior midline, superior to the upper border of the hyoid bone

Connections with other channels/organs

- Kidney, Spleen and Liver primary channels, *ren mai*
- Internal organs

Clinical importance (→ 1.7.2, 1.7.3)

- **Connects** and **regulates** all **Yin channels** of the body. It connects the KID, SP and LIV channels and the *ren mai* and dominates the **Interior** of the body (the internal organs).
- **Strengthens** Yin and Heart-Blood, especially in women
- **Balances** the emotions

5.6 *Yang wei mai*

Synonym: Yang Motility Vessel (Deadman), **Yang Heel Vessel** (Maciocia)

Relationships (→ 1.7.3)

Yin/Yang: *yin wei mai*/*yang wei mai*
Central/peripheral: *dai mai*/*yang wei mai*. Regions supported by this pairing: Lateral eye region, temporal region, ears, cheeks, neck and shoulders. Opening point: **T.B.-5** (*waiguan*), coupled point: **G.B.-41** (*zulinqi*).

Pathway

The extraordinary vessel *yang wei mai* originates at → **BL-63** (*jinmen*) in the depression posterior to the tuberosity of the 5th metatarsal bone, runs anterior to the lateral malleolus, passes → **G.B.-35** (*yangjiao*), ascends the lateral aspect of the leg, passes the hip, ascends the posterior aspect of the costal region to the shoulder, passes → **S.I.-10** (*naoshu*), → **T.B.-15** (*tianliao*), → **G.B.-21** (*jianjing*) and ascends the neck. The vessel then ascends anterior to the ear (according to some authors, for example Solinas et al. (1998), posterior to the ear) to the forehead, passing → **ST-8** (*touwei*), → **G.B.-13** (*benshen*) and → **G.B.-14** (*yangbai*). It then runs with the G.B. channel from **G.B.-15** to **G.B.-20** on the occiput, passing → **G.B.-15** (*toulinqi*), → **G.B.-16** (*muchuang*), → **G.B.-17** (*zhengying*), → **G.B.-18** (*chengling*), → **G.B.-19** (*naokong*) and → **G.B.-20** (*fengchi*). The vessel then continues from **G.B.-20** to the posterior midline, passing → **Du-16** (*fengfu*) and → **Du-15** (*yamen*), where the vessel terminates.

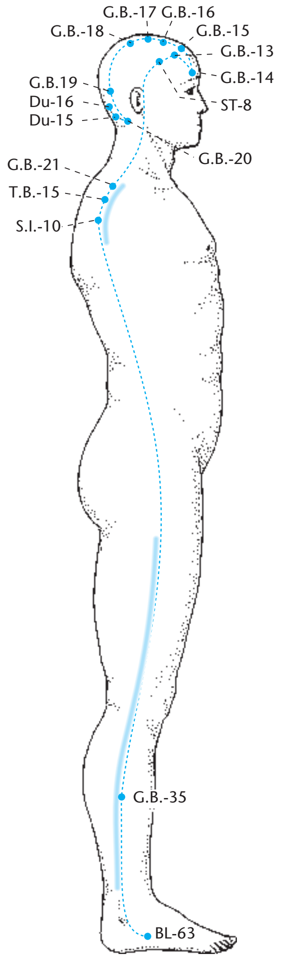

Meeting points with other channels

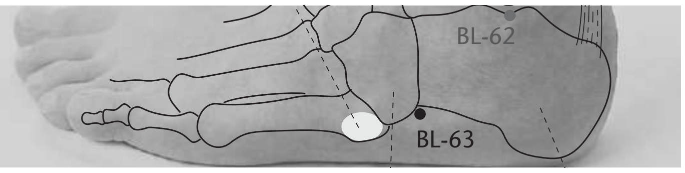

BL-63 (*jinmen*): In a depression anterior and inferior to → **BL-62** between the calcaneus and the cuboid bone, proximal to the tuberosity of the 5th metatarsal bone

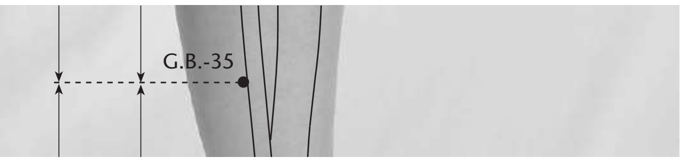

G.B.-35 (*yangjiao*): 7 cun proximal to the highest prominence of the lateral malleolus, on the posterior border of the fibula

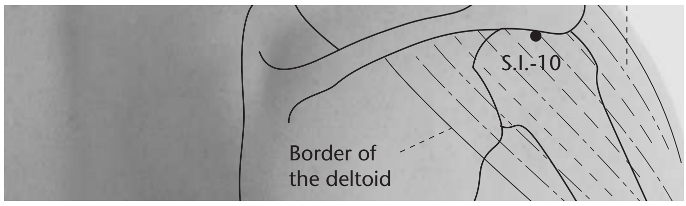
S.I.-10
Border of
the deltoid

S.I.-10 (*naoshu*): With the arm adducted on a line extending in a superior direction from the posterior axillary fold, on the lower border of the scapular spine

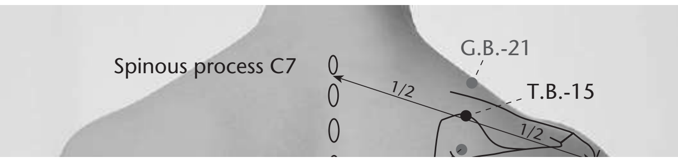
Spinous process C7
G.B.-21
T.B.-15
1/2
1/2

T.B.-15 (*tianliao*): At the midpoint of an imaginary line between the spinous process of the C7 and the lateral extremity of the acromion

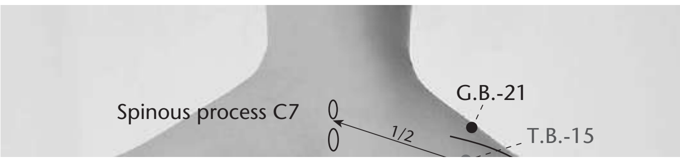
Spinous process C7
G.B.-21
1/2
T.B.-15

G.B.-21 (*jianjing*): At the highest point of the shoulder, midway between the lower border of the spinous process of C7 and the lateral extremity of the acromion

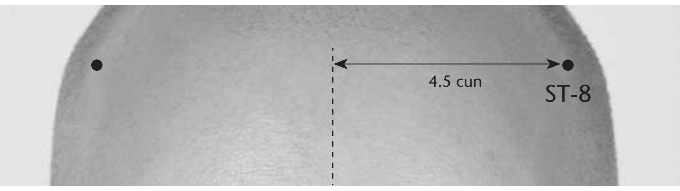

ST-8 (*touwei*): At the temporal corner of the forehead, on the border of the temporalis muscle and 0.5 cun within the anterior hair line or 4.5 cun lateral to the anterior midline (Du-24)

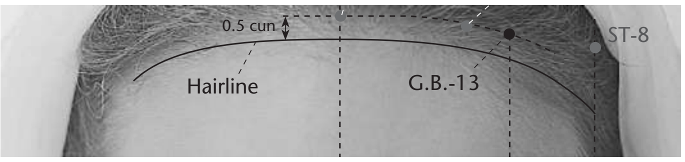
0.5 cun
ST-8
Hairline
G.B.-13

G.B.-13 (*benshen*): 3 cun lateral to → **Du-24** (on the midline, 0.5 cun superior to the anterior hairline)

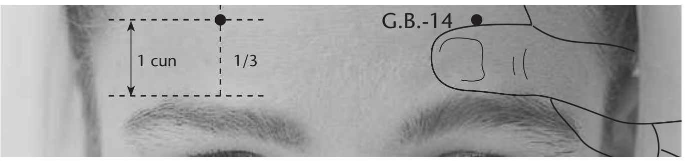
G.B.-14
1 cun
1/3

G.B.-14 (*yangbai*): When looking straight ahead on the pupil line, approximately 1 cun superior to the midpoint of the eyebrow, at the junction of the eminence and the superciliary arch

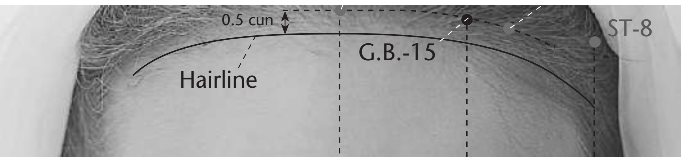
0.5 cun
ST-8
G.B.-15
Hairline

G.B.-15 (*toulinqi*): When looking straight ahead superior to the pupil, 0.5 cun superior to the anterior hairline

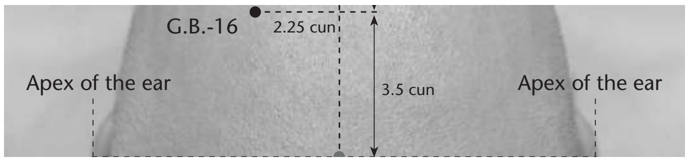
G.B.-16
2.25 cun
Apex of the ear
3.5 cun
Apex of the ear

G.B.-16 (*muchuang*): 1.5 cun superior to the anterior hairline, on the pupil line, or 2.25 cun lateral to the midline (= midway between → **Du-24** and → **ST-8**)

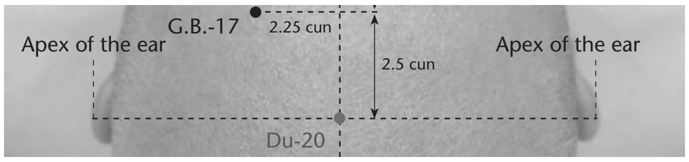
G.B.-17
2.25 cun
Apex of the ear
2.5 cun
Apex of the ear
Du-20

G.B.-17 (*zhengying*): 2.5 cun superior to the anterior hairline and 2.25 cun lateral to the midline

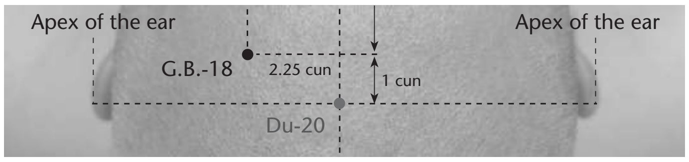
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
Du-20

G.B.-18 (*chengling*): 4 cun superior to the anterior hairline or 1 cun anterior to → **Du-20**, 2.25 cun lateral to the midline

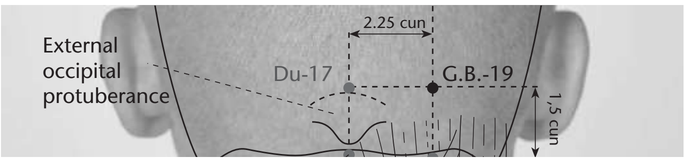
External occipital protuberance
2.25 cun
Du-17
G.B.-19
1.5 cun

G.B.-19 (*naokong*): On the occiput, level with the upper border of the external occipital protuberance (→ **Du-17**), 2.25 cun lateral to the midline

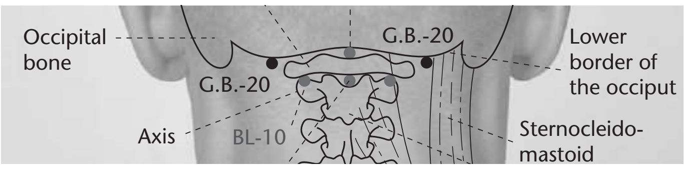
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Sternocleido-mastoid

G.B.-20 (*fengchi*): On the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius

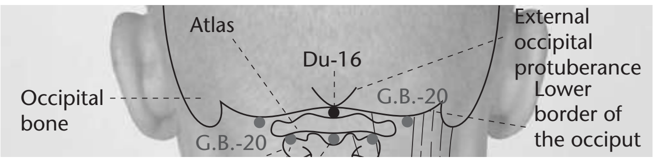
Atlas
Du-16
External occipital protuberance
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput

Du-16 (*fengfu*): On the posterior midline, directly below the external occipital protuberance, in the depression between the origins of the trapezius

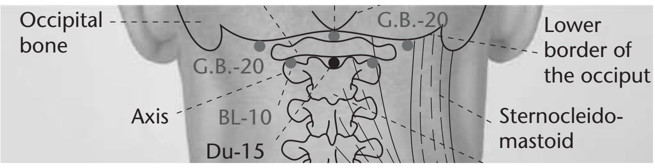
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Du-15
Sternocleido-mastoid

Du-15 (*yamen*): On the occiput, on the posterior midline, in the depression between the 1st (atlas) and 2nd cervical vertebrae, approximately 0.5 cun below → **Du-16**

Connections with other channels/Organs

- BL, G.B., ST, S.I., T.B. and L.I. primary channels, *du mai*

Clinical importance

- **Connects** and **regulates** all **Yang channels** of the body. It connects the BL, G.B., ST, S.I., T.B. and L.I. channels as well as the *du mai* and controls the Exterior of the whole body (especially the *taiyang* and *shaoyang* channels)
- Harmonises the Nutritive Qi (*ying qi*) and the Defensive Qi (*wei qi*)
- Expels Exterior Wind-Cold, especially with *shaoyang*-syndrome.

5.7 *Yin qiao mai*

Synonym: Yin Motility Vessel (Deadman), **Yin Heel Vessel** (Maciocia)

Relationships (→ 1.7.2, 1.7.3)

- **Yin/Yang:** *yin qiao mai*/*yang qiao mai*
- **Central/peripheral:** *ren mai*/*yin qiao mai*. Regions supported by this pairing: Face, throat, thorax, Lung, diaphragm, abdomen. Opening point: **KID-6** (*zhaohai*), coupled point: **LU-7** (*lieque*).

Pathway

The extraordinary vessel *yin qiao mai* originates at → **KID-2** (*rangu*) inferior to the navicular tuberosity (according to some authors it only begins at → **KID-6**), runs to → **KID-6** (*zhaohai*) and → **KID-8** (*jiaoxin*), ascends the posteromedial aspect of the leg to the external genital region, ascends the abdomen to the thorax, crosses the supraclavicular fossa and continues to the throat and face. It then runs to the inner canthus of the eye where it connects with the Bladder primary channel and the *yang qiao mai* at → **BL-1** (*jingming*). It then ascends to enter the Brain.

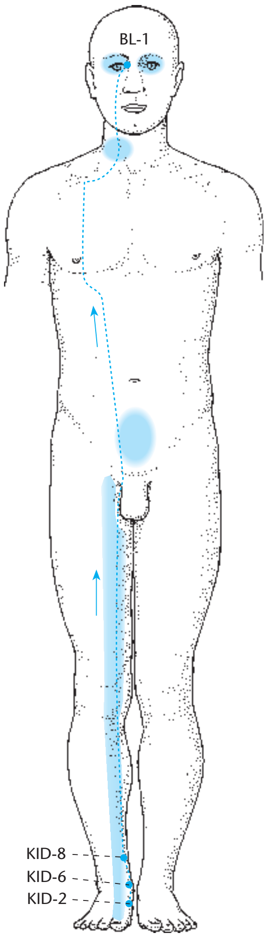

Meeting points with other channels

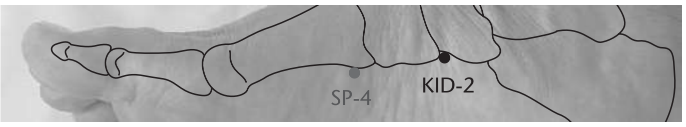

KID-2 (*rangu*): In a depression at the anterior/inferior border of the navicular bone, at the border of the 'red and white' skin

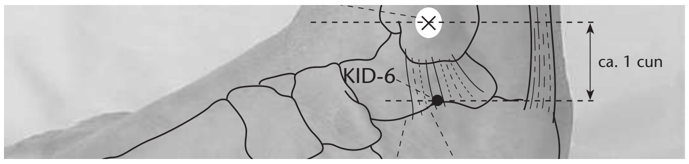

KID-6 (*zhaohai*): In the depression inferior to the highest prominence of the medial malleolus, over the joint space between the talus and calcaneus

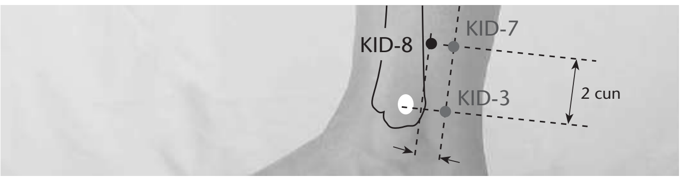

KID-8 (*jiaoxin*): 2 cun proximal to the highest prominence of the medial malleolus, posterior to the border of the tibia

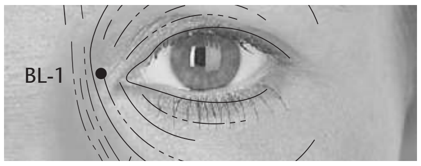

BL-1 (*jingming*): In a depression 0.1 cun superior and medial to the inner canthus of the eye

Connections with other channels/Organs

- Kidney and Bladder primary channels
- Brain

Clinical importance (→ 1.7.2, 1.7.3)

- Together with the *yang qiao mai* controls the musculature of the lower extremities
- Together with the *yang qiao mai* regulates the opening and closing of the eyes
- Regulates the functioning of the Brain
- Removes stagnation (of Qi, Blood or Dampness) in the Lower Burner, especially in the Uterus in women

5.8 *Yang qiao mai*

Synonym: Yang Motility Vessel (Deadman), **Yang Heel Vessel** (Maciocia)

Relationships (→ 1.7.3)

- **Yin/Yang:** *yin qiao mai*/*yang qiao mai*
- **Central/peripheral:** *du mai*/*yin qiao mai*. Regions supported by this pairing: Inner canthus of the eye, occiput, shoulders and back. Opening point: **BL-62** (*shenmai*), coupled point: **S.I.3** (*houxi*)).

Pathway

The extraordinary vessel *yang qiao mai* originates at → **BL-62** (*shenmai*) below the prominence of the lateral malleolus, curves briefly around the malleolus to → **BL-61** (*pucan*) and ascends anterior to the Achilles tendon to → **BL-59** (*fuyang*), travels up the lateral aspect of the leg, crosses the hip, passing → **G.B.-29** (*juliao*), ascends the flanks and the posterior aspect of the shoulders, passes → **S.I.10** (*naoshu*), continues to → **L.I.-15** (*jianyu*) and runs in a curve to → **L.I.-16** (*jugu*), crosses the supraclavicular fossa, ascends the throat, according to some authors passing → **ST-9** (*renying*), and reaches the face. Here the vessel passes the points → **ST-4** (*dicang*) and → **ST-3** (*juliao*); according to some authors also → **ST-2** (*sibai*) and → **ST-1** (*chengqi*) in the infraorbital region. It then reaches → **BL-1** (*jingming*), where it meets the extraordinary vessel *yin qiao mai*, ascends the forehead to the vertex and descends posteriorly to → **G.B.-20** (*fengchi*); according to the *Nan Jing* and the *Nei Jing* it also reaches → **Du-16** (*fengfu*), where it enters the Brain.

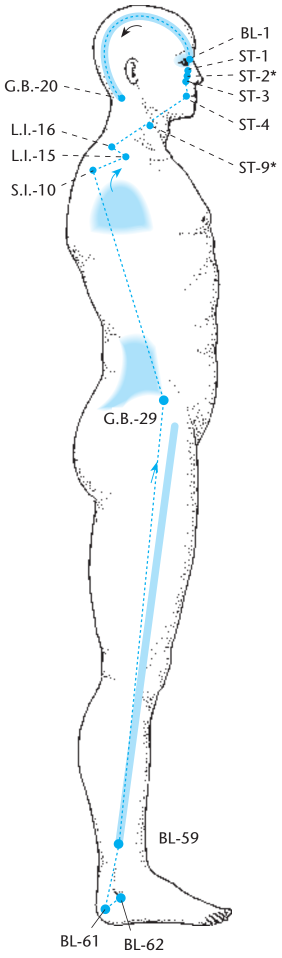

* According to Deadman et al 1998.

Intersection points with other channels

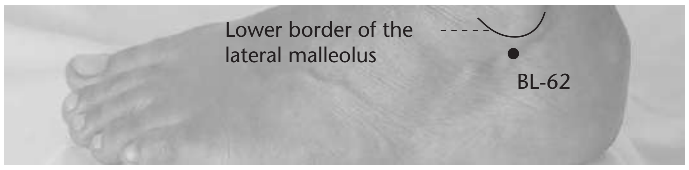

BL-62 (*shenmai*): In a depression directly inferior to the highest prominence of the lateral malleolus, over the joint space between the talus and calcaneus

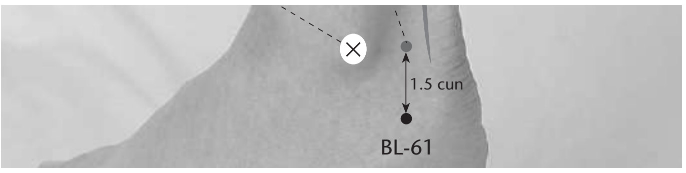

BL-61 (*pucan*): In a depression on the calcaneus, approximately 1.5 cun inferior to → **BL-60**

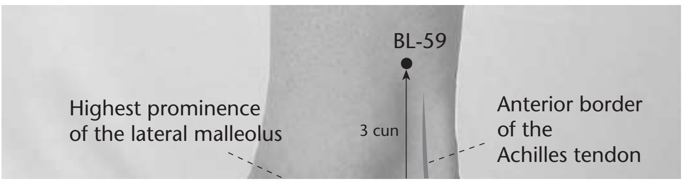
BL-59
Highest prominence of the lateral malleolus
3 cun
Anterior border of the Achilles tendon

BL-59 (*fuyang*): 3 cun superior to → BL-60 (in the depression between the highest prominence of the lateral malleolus and the Achilles tendon)

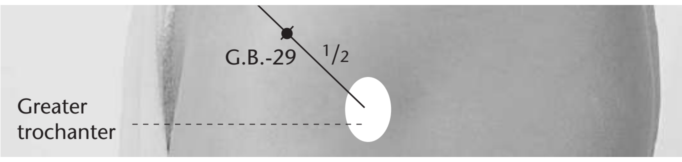
G.B.-29
1/2
Greater trochanter

G.B.-29 (*juliao*): At the midpoint of a line connecting the anterior superior iliac spine and the greater trochanter, at the anterior border of the iliac crest

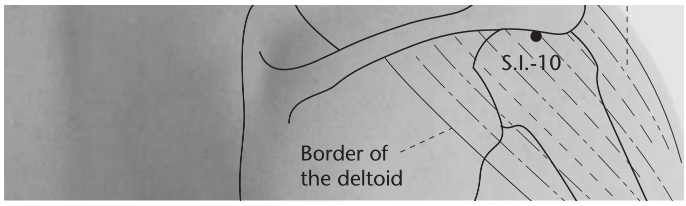
S.I.-10
Border of the deltoid

S.I.-10 (*naoshu*): With the arm adducted on a line extending in a superior direction from the posterior axillary fold, on the lower border of the scapular spine

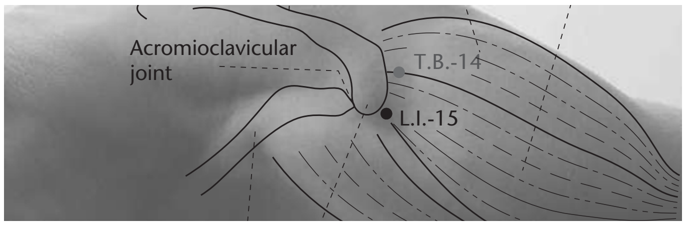
Acromioclavicular joint
T.B.-14
L.I.-15

L.I.-15 (*jianyu*): In the depression distal and anterior to the acromion, between the clavicular and acromial portions of the deltoid muscle

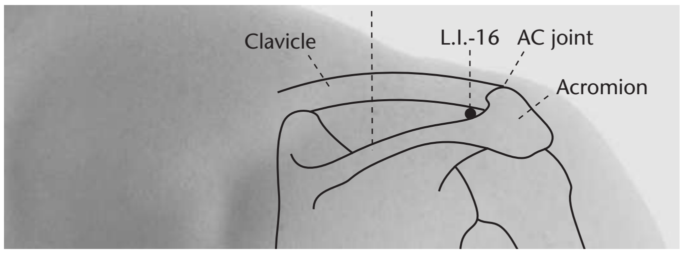
Clavicle
L.I.-16
AC joint
Acromion

L.I.-16 (*jugu*): In a depression between the acromial extremity of the clavicle and the junction of the scapular spine and the acromion

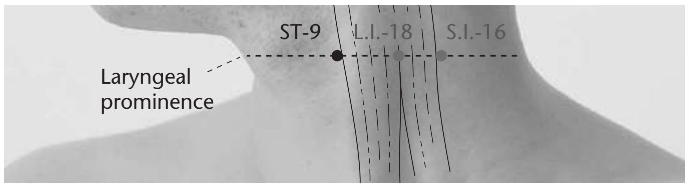
ST-9
L.I.-18
S.I.-16
Laryngeal prominence

ST-9 (*renying*): Approximately 1.5 cun lateral to the anterior midline, level with the laryngeal prominence, at the anterior border of the sternocleidomastoid

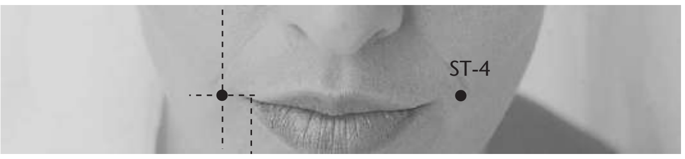

ST-4 (*dicang*): When looking straight ahead on the pupil line, approximately 0.4 cun lateral to the corner of the mouth

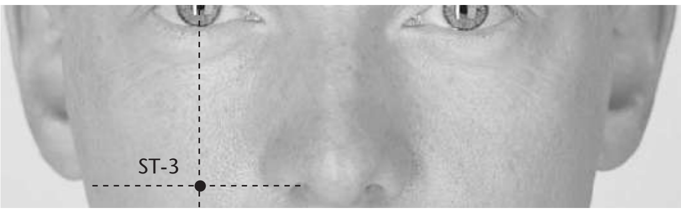

ST-3 (*juliao*): When looking straight ahead, directly below the centre of the pupil, level with the lower border of the ala nasi

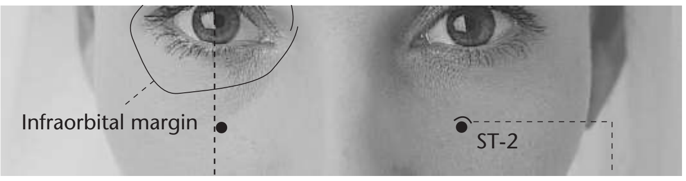
Infraorbital margin
ST-2

ST-2 (*sibai*): When looking straight ahead, directly below the centre of the pupil, in the depression at the infraorbital foramen

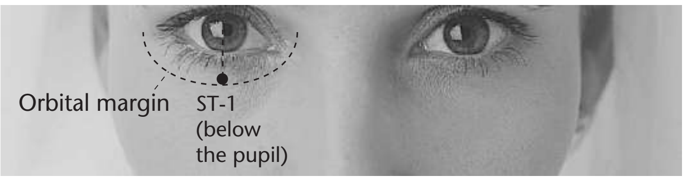
Orbital margin
ST-1
(below
the pupil)

ST-1 (*chengqi*): With the eyes looking straight forward, directly below the centre of the pupil, between the eyeball and the infraorbital ridge

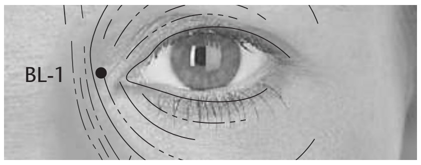

BL-1 (*jingming*): In a depression 0.1 cun superior and medial to the inner canthus of the eye

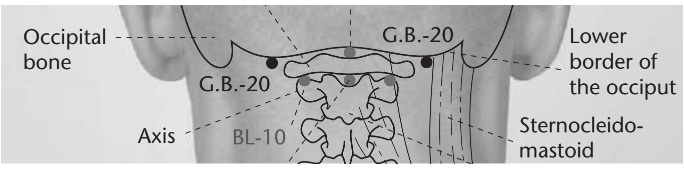
Occipital bone
G.B.-20
G.B.-20
Lower border of the occipit
Axis
BL-10
Sternocleido-mastoid

G.B.-20 (*fengchi*): On the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius

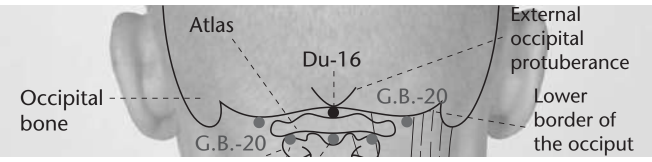
Atlas
Du-16
External occipital protuberance
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput

Du-16 (*fengfu*): On the posterior midline, directly below the external occipital protuberance, in the depression between the origins of the trapezius

Connections with other channels/Organs

- Kidney, Bladder, Gall Bladder, Stomach, Large Intestine and Small Intestine primary channels
- Brain

Clinical importance (→ 1.7.2, 1.7.3)

- Together with the *yin qiao mai* controls the musculature of the lower extremities
- Together with the *yang qiao mai* regulates the opening and closing of the eyes
- Regulates the functioning of the Brain, eliminates internal and external Wind from the head
- Removes obstructions and stagnation from the spine, especially after traumatic injuries

6 Extra Points

Claudia Focks, Ulrich März

Besides the 361 classic acupuncture points located on the channels there are a number of **Ex**tra Points (**Ex**), located with a few exceptions outside the pathways of the primary channels. In 1991, a new nomenclature for these extra points was confirmed in the People's Republic of China, listing 48 points (*The Location of Acupoints, State Standard of the People's Republic of China*, Foreign Languages Press, Beijing, 1990).

In the past, authors have developed ways of identifying the extra points by names or numbers:

- Nguyen Van Nghi; König and Wancura: '**P**unkte **a**ußerhalb der **M**eridiane' (points not on meridians) (PaM) and '**N**eu-**P**unkte' (new points) (NP); Schnorrenberger (wall charts translated from the Chinese language) uses the same numbering system as Nguyen Van Nghi but refers to the **PaM** as '**Z**usatz**p**unkte' (additional points) (ZP) and also as '**N**eu-**P**unkte' (NP)
- Shanghai College for Traditional Medicine (*Acupuncture – a Comprehensive Text*, English translation by O'Connor J, Bensky D, which also forms the foundation for Deadman et al. 1998 and Ellis and Wiseman 1991): **M**iscellaneous (M) and **N**ew (N) Points
- Hempen 1995: '**E**xtrapunkte' (extra points) (Ex).

The following tables give an overview of the most commonly used extra points also listed in this atlas.

6.1 Extra Points: Head and Neck (EX-HN)

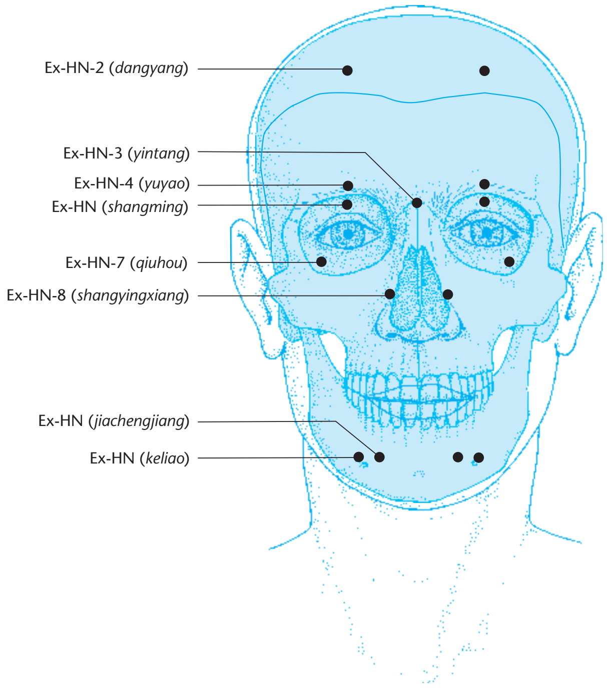
Ex-HN-2 (dangyang)
Ex-HN-3 (yintang)
Ex-HN-4 (yuyao)
Ex-HN (shangming)
Ex-HN-7 (qiuhou)
Ex-HN-8 (shangyingxiang)
Ex-HN (jiachengjiang)
Ex-HN (keliao)

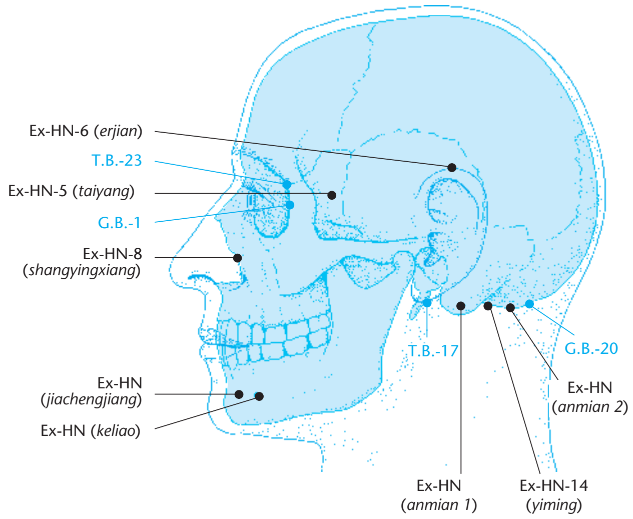
Ex-HN-6 (erjian)
T.B.-23
Ex-HN-5 (taiyang)
G.B.-1
Ex-HN-8 (shangyingxiang)
Ex-HN (jiachengjiang)
Ex-HN (keliao)
T.B.-17
G.B.-20
Ex-HN (anmian 2)
Ex-HN (anmian 1)
Ex-HN-14 (yiming)

English abbreviation (standard)	***Pinyin* name**	**Nguyen Van Nghi, König and Wancura, Schnorrenberger**	**Shanghai College**	**Ex (Hempen)**
Ex-HN-1	*sishencong*	PaM or ZP1	M-HN-1	Ex6
Ex-HN-2	*dangyang*		M-HN	
Ex-HN-3	*yintang*	PaM or ZP3	M-HN-3	Ex1
Ex-HN-4	*yuyao*	PaM or ZP6	M-HN-6	
Ex-HN-5	*taiyang*	PaM or ZP9	M-HN-9	Ex2
Ex-HN-6	*erjian*	PaM or ZP10	M-HN-10	
Ex-HN-7	*qiuhou*	PaM or ZP8	M-HN-8	
Ex-HN-8	*shangyingxiang/bitong*	NP12 (PaM or ZP14[*])	M-HN-14	Ex3
Ex-HN-9	*neiyingxiang*		M-HN-35	
Ex-HN-10	*juquan*		M-HN-36	
Ex-HN-11	*haiquan*		M-HN-37	
Ex-HN-12	*jinjin*[†]	PaM or ZP20	M-HN-20	
Ex-HN-13	*yuye*[†]	PaM or ZP20	M-HN-20	
Ex-HN-14	*yiming*	PaM or ZP13	M-HN-13	Ex4
Ex-HN-15	*jing(bailao)*	PaM or ZP30	M-HN-30	
Further extra points				
Ex-HN	*shangming*		N-HN-4	
Ex-HN	*anmian*[‡]	NP27 and 28 (*anmian*[‡]1 und 2)	N-HN-54 M-HN-54 (Deadman)	Ex5
Ex-HN	*jiabi*	PaM or ZP15		
Ex-HN	*jiachengjiang*	PaM or ZP18 (*heliao/keliao*)	M-HN-18	Ex7
Ex-HN	*chonggu/zhuidong*	PaM or ZP31	M-HN-31	
Ex-HN	*jingbi*		M-HN-41	

---

[*] Nguyen Van Nghi and Schnorrenberger describe a different location for the point with this *pinyin* name (location: 0.5 cun inferior to the inner canthus of the eye); the 'standard' location is described under NP12 as '*bitong*' or '*bicong*'.

[†] The Shanghai nomenclature and also Nguyen Van Nghi, König and Wancura and Schnorrenberger describe the extra points *jinjin* and *yuye* as a pair of points.

[‡] There exists differing information regarding the extra point *anmian*; for example the Shanghai College (and Wiseman) describe *anmian* as N-HN-54 (in Deadman: M-HN-54) with the following location: midway between G.B.-20 and T.B.-17; Nguyen Van Nghi and Schnorrenberger describe *anmian* as two new points with the following locations: *anmian* 1 is located midway between T.B.-17 and *yiming* (PaM13 or Ex-HN-14), *anmian* 2 midway between G.B.-20 and *yiming* (*anmian* according to the Shanghai College is shown on the point page; the location of *anmian* 1 and 2 according to Nguyen and Schnorrenberger in figure 6.1).

Four Alert Spirit SISHENCONG Ex-HN-1

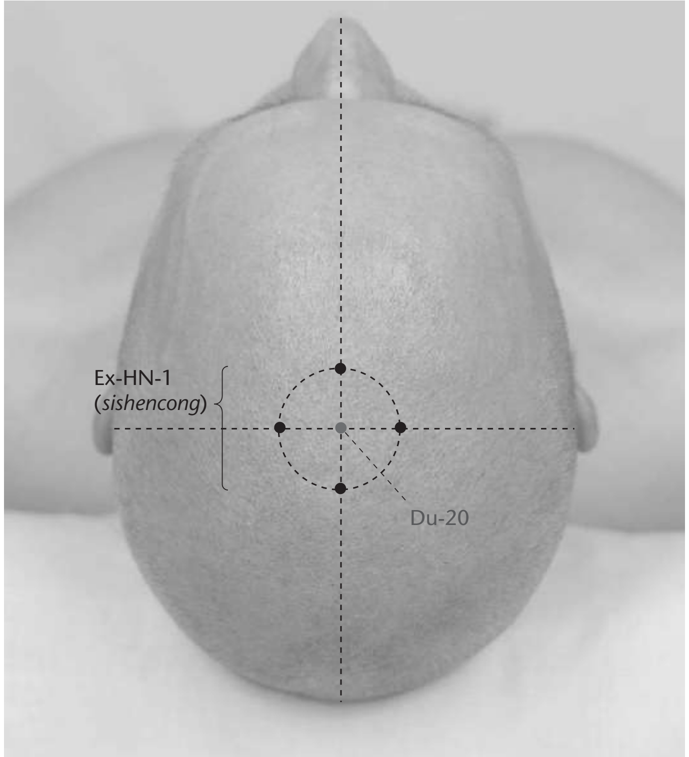
Ex-HN-1
(*sishencong*)
Du-20

Location

A group of four points, each located 1 cun from → **Du-20** (anterior, posterior and lateral) (→ 3.1.1).

How to find

First, locate → **Du-20** on the highest point of the head, at the junction of a line connecting the apices of the ears and the vertical midline. Alternative location method: Locate → **Du-20** 5 cun superior to the anterior and 7 cun superior to the posterior hairline. **Ex-HN-1** (*sishencong*) forms a star-shaped group of points around → **Du-20**: two points are located at a distance of 1 cun each on the *du mai*, whereas the other two are located 1 cun each from → **Du-20** in a lateral direction.

Needling

0.5–1 cun obliquely towards → **Du-20**

Actions/Indications

- Calm the *shen*
- Alleviate pain
- Calm Wind
- Benefit the eyes and ears

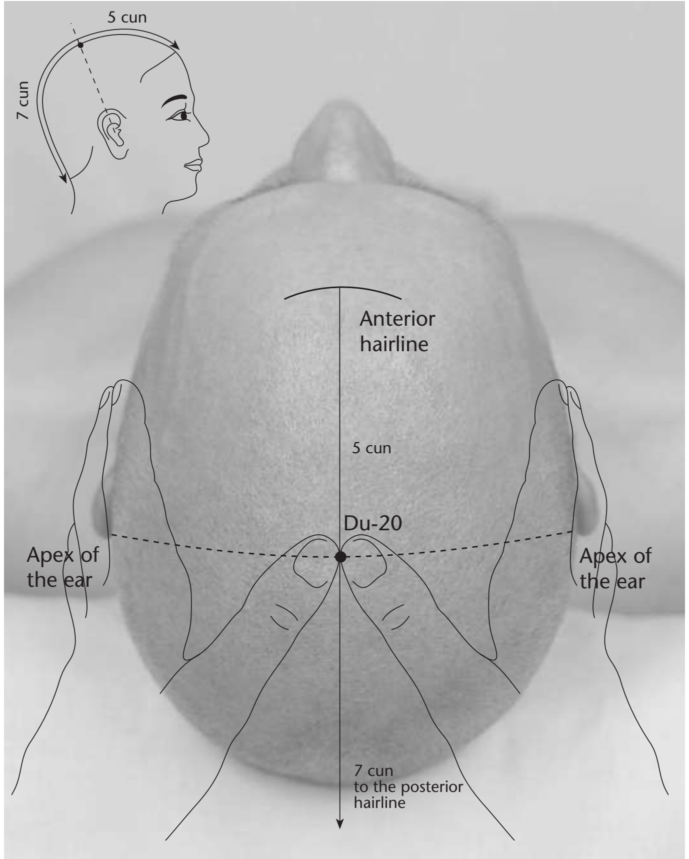
5 cun
7 cun
Anterior hairline
5 cun
Du-20
Apex of the ear
Apex of the ear
7 cun to the posterior hairline

Ex-HN-2 *Above the Yang* DANGYANG

Location

With the patient looking straight ahead, vertically above the pupil, 1 cun above the anterior hairline.

How to find

In patients with a receding hairline, the original hairline can often be recognized by a change in skin texture. Location help (→ 3.1.1): Ask the patient to frown, as this often reveals the site of the original hairline. Ask the patient to look straight ahead, then locate **Ex-HN-2** by measuring 1 cun from the hairline, on a vertical line through the pupil.

→ **G.B.-14** is also located on the pupil line (on the lower border of the frontal eminence). Located on the same level (1 cun above the hairline) are → **BL-5** (more medial, 1.5 cun from the anterior midline) and → **Du-23** (on the anterior midline).

Needling

0.5 cun transversely (subcutaneously) towards the site of the disorder/pain.

Actions/Indications

- Dispels Wind and Heat, alleviates pain

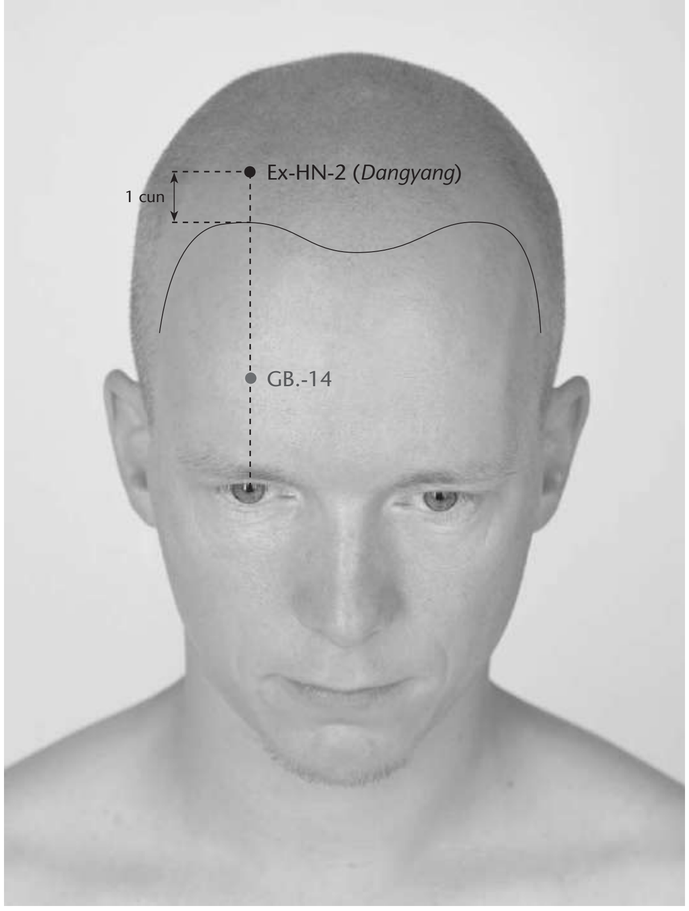
1 cun
Ex-HN-2 (*Dangyang*)
GB.-14

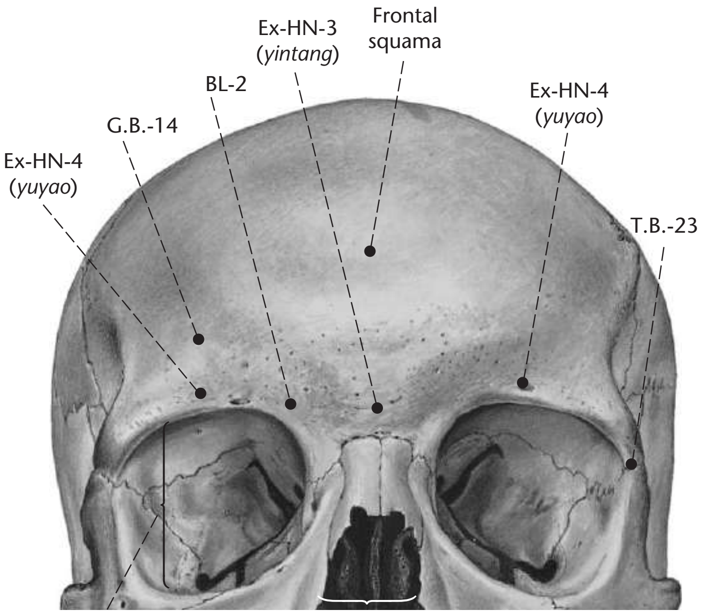
Ex-HN-3
(*yintang*)
Frontal
squama
BL-2
G.B.-14
Ex-HN-4
(*yuyao*)
Ex-HN-4
(*yuyao*)
T.B.-23

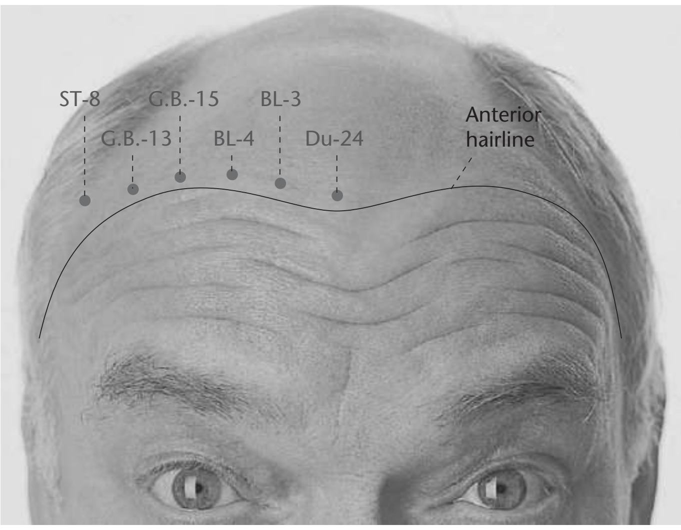
ST-8
G.B.-15
BL-3
G.B.-13
BL-4
Du-24
Anterior
hairline

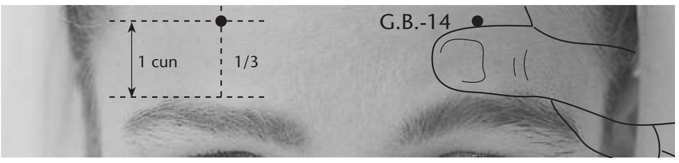
1 cun
1/3
G.B.-14

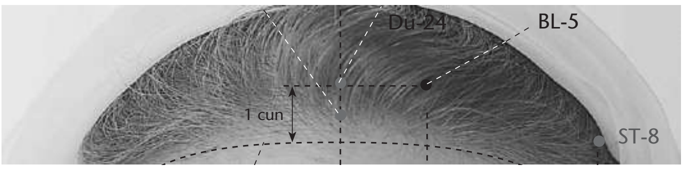
Du-24
BL-5
1 cun
ST-8

Hall of Impression YINTANG Ex-HN-3

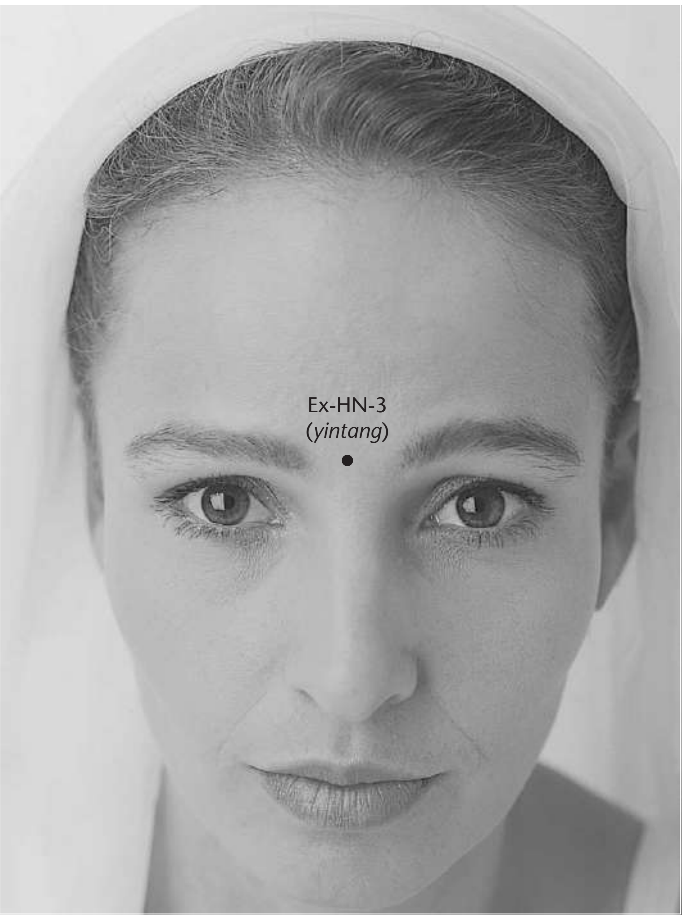
Ex-HN-3
(*yintang*)

Location

On the anterior midline (*du mai*), between the eyebrows.

How to find

Locate the glabella (→ 3.1.1), a smooth surface on the frontal bone, directly above the root of the nose, between the superciliary arches. Then locate **Ex-HN-3** (*yintang*) in its centre, on the midline and between the medial extremities of the eyebrows. Located along the superciliary arches (from medial to lateral) are → **BL-2,** → **Ex-HN-4** (*yuyao*) and → **T.B.-23**.

Needling

Pinching-skin method: With the thumb and index finger, form a skin fold above the point and insert the needle 0.3–0.5 cun transversely from superior to inferior into the fold, towards the root of the nose. Alternative methods: Oblique or transverse (subcutaneous) insertion towards the eyebrows (→ **BL-2**) or prick to bleed.

Actions/Indications

- Calms the *shen*
- Benefits the nose
- Eliminates (internal) Wind
- Opens the channel, alleviates pain

Special features

Commonly used extra point with a harmonising effect

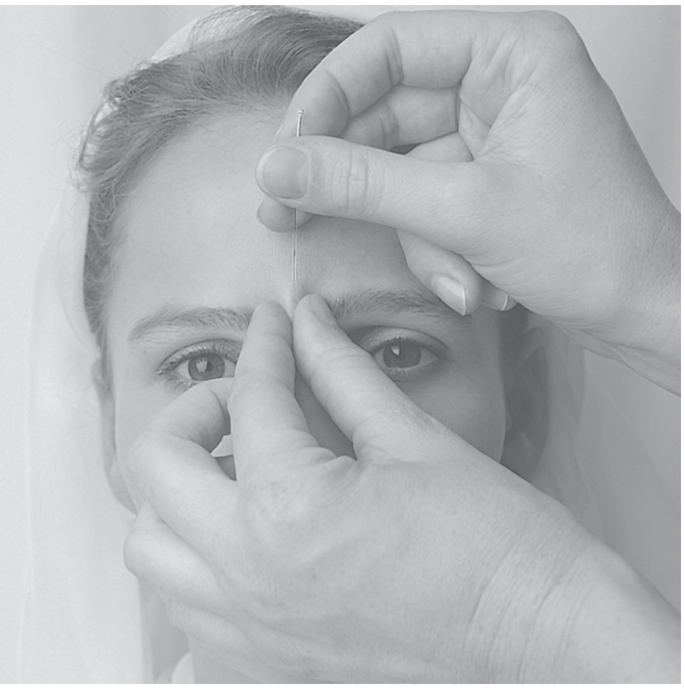

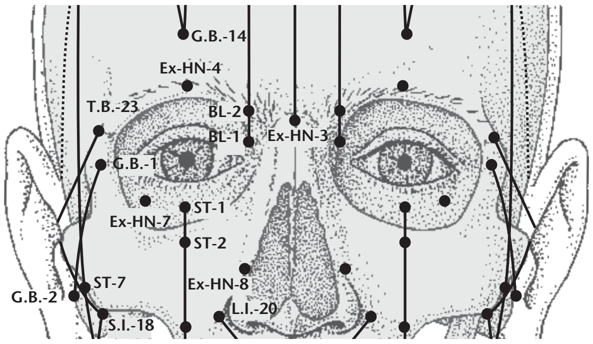
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
ST-7
Ex-HN-8
G.B.-2
L.I.-20
S.I.-18

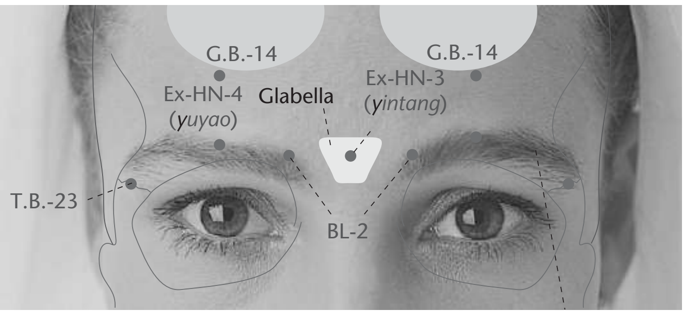
G.B.-14
G.B.-14
Ex-HN-4
(*yuyao*)
Glabella
Ex-HN-3
(*yintang*)
T.B.-23
BL-2

Ex-HN-4 ***Fish Waist*** **YUYAO**

Location

In the centre of the eyebrow, with the eyes looking straight ahead directly above the pupil.

How to find

With the patient looking straight ahead, palpate the eyebrow directly above the pupil, superior to the supraorbital ridge. There, locate **Ex-HN-4** (*yuyao*) in a small depression on the superciliary arch.

Needling

0.3–0.5 cun vertically or obliquely towards the eye.

Actions/Indications

- Benefits the eyes
- Dispels pain

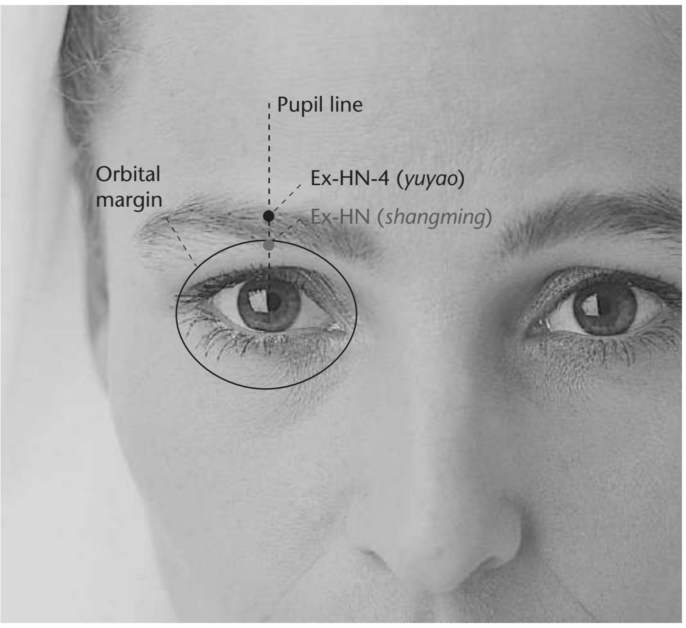
Pupil line
Orbital margin
Ex-HN-4 (*yuyao*)
Ex-HN (*shangming*)

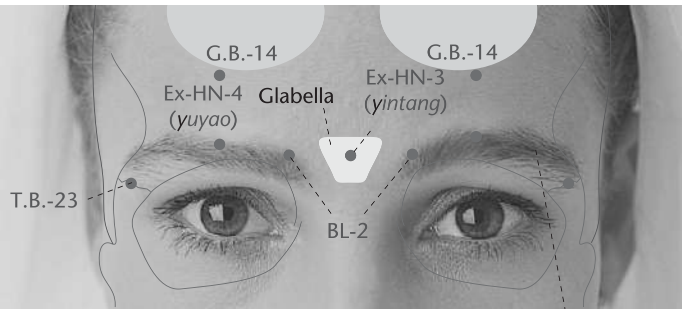
G.B.-14
G.B.-14
Ex-HN-4 (*yuyao*)
Glabella
Ex-HN-3 (*yintang*)
T.B.-23
BL-2

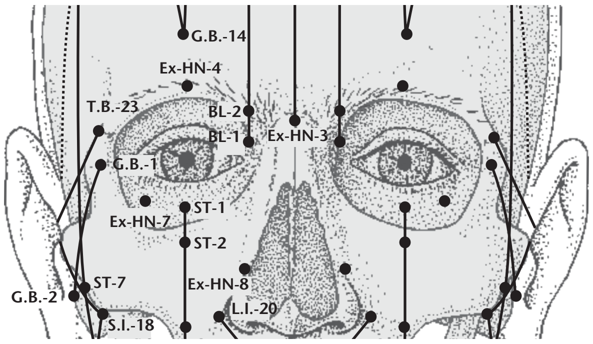
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
Ex-HN-8
G.B.-2
L.I.-20
S.I.-18

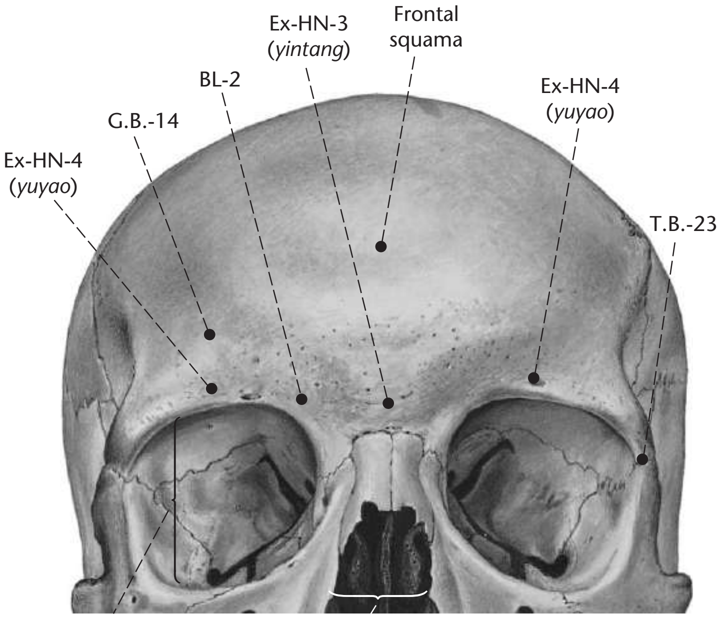
Ex-HN-3 (*yintang*)
Frontal squama
BL-2
G.B.-14
Ex-HN-4 (*yuyao*)
Ex-HN-4 (*yuyao*)
T.B.-23

Supreme Yang TAIYANG **Ex-HN-5**

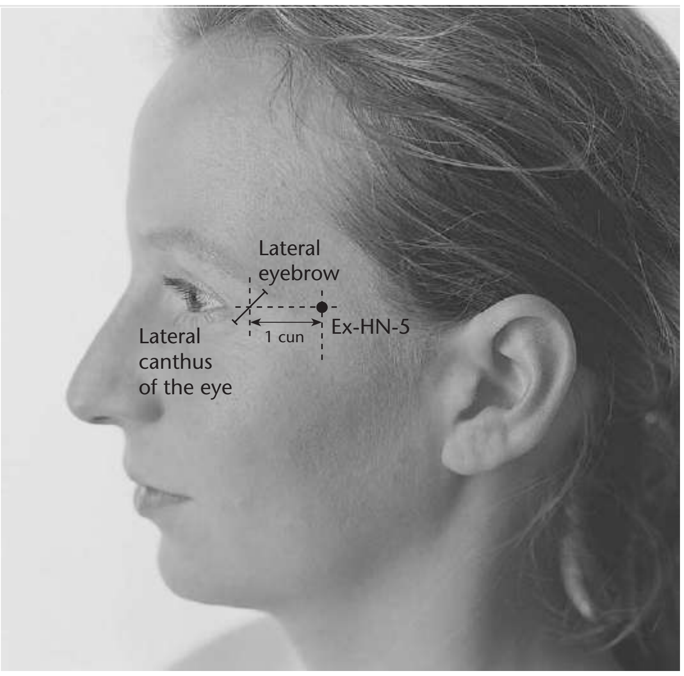
Lateral
eyebrow
Ex-HN-5
1 cun
Lateral
canthus
of the eye

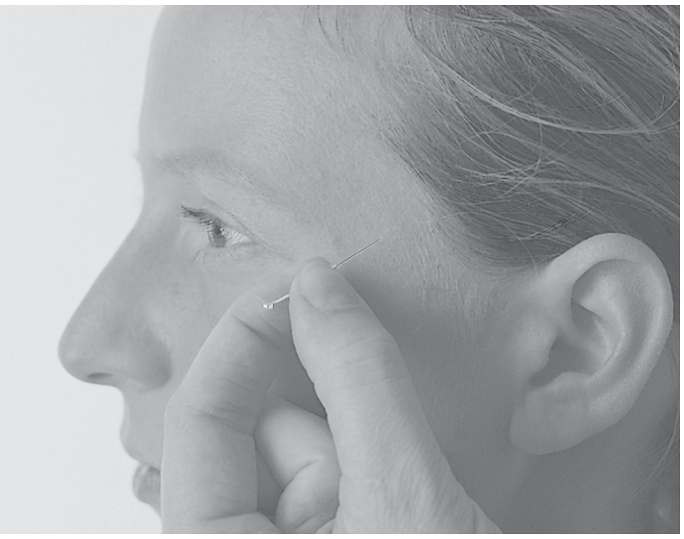

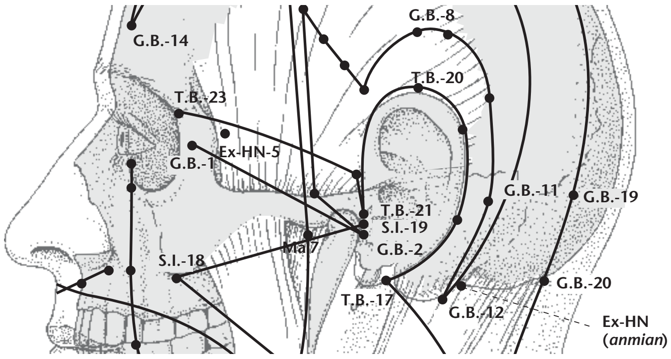
G.B.-14
G.B.-8
T.B.-23
T.B.-20
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
Ma 7
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)

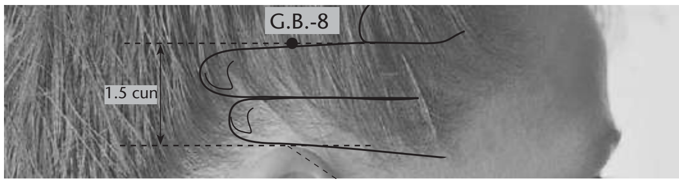

Location

On the temple, in a depression approximately 1 cun lateral to the midpoint of a line connecting the lateral extremity of the eyebrow and the outer canthus of the eye.

How to find

On the temple, first, draw a line between the lateral extremity of the eyebrow and the outer canthus of the eye. Next, palpate from the midpoint of this line in a lateral direction, until you can feel a distinct depression in the temporal bone. **Ex-HN-5** (*taiyang*) is located in the centre of this depression. This point tends to be pressure-sensitive, especially with temporal headaches. Patients often get relief by using acupressure on this point.

Needling

0.5–0.6 cun vertical insertion into the temporal muscle or obliquely in a lateral direction or transversely (subcutaneously) towards → **G.B.-8.** This point often bleeds when needled. Prick to bleed to drain Heat.

Actions/Indications

- Eliminates Wind, clears Heat, reduces swellings, opens the channel, alleviates pain

Special features

One of the most important and most commonly used extra points, especially with headaches

Ex-HN-6

Tip of the Ear ERJIAN

Location

On the apex of the ear, on the helix.

How to find

Fold the ear anteriorly, so that the posterior portion of the upper helix covers its anterior part. The point is located at the highest point of the fold. This is also the location of the auricular point **78**.

Needling

Vertically 0.1–0.2 cun. Some authors advise to bleed this point for Heat conditions, but this is not recommended owing to the danger of causing an otic haematoma. Indirect moxibustion for vision disorders.

Actions/Indications

- Clears Heat, alleviates pain, benefits the eyes and the throat

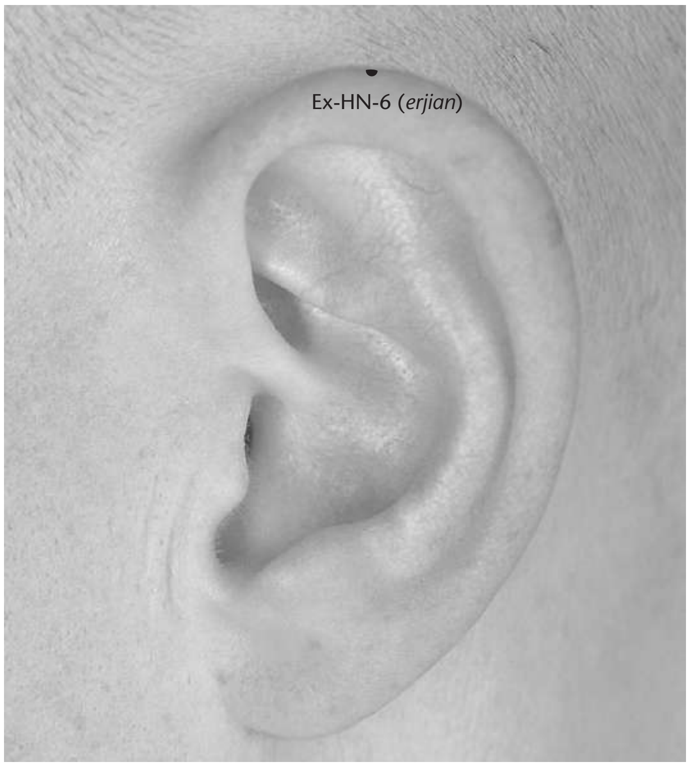

Behind the Ball QIUHOU **Ex-HN-7**

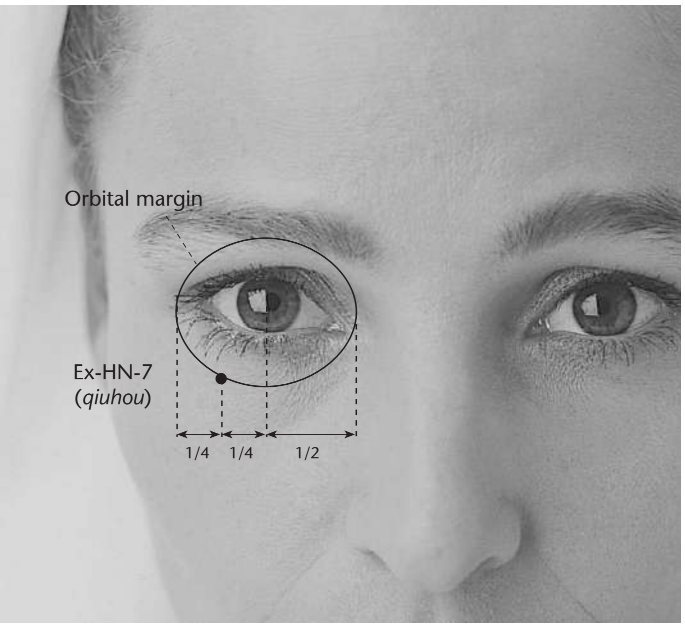
Orbital margin
Ex-HN-7
(*qiuhou*)
1/4
1/4
1/2

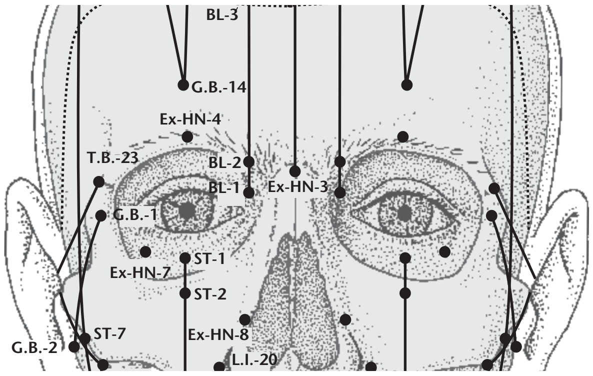
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
G.B.-2
Ex-HN-8
L.I.-20

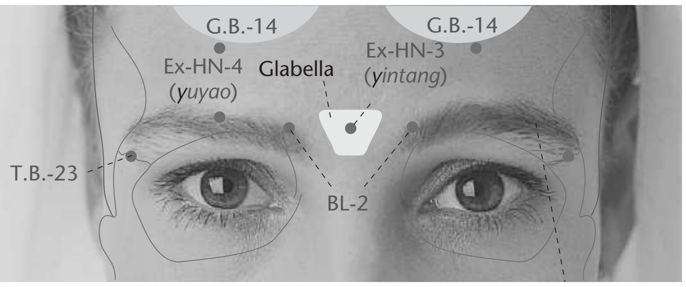
G.B.-14
G.B.-14
Ex-HN-4
(*yuyao*)
Glabella
Ex-HN-3
(*yintang*)
T.B.-23
BL-2

Location

On the lower border of the orbit, at the junction of the lateral quarter and the medial three quarters.

How to find

Divide the horizontal extension of the diameter of the orbit in quarters. Then locate the point at the junction of the first and second lateral quarters, slightly superior to the border of the orbit.

Needling

Ask the patient to look upward and gently push the eyeball upward, away from the lower eyelid. Slowly insert the needle 0.5–1 cun into the fatty tissue immediately above the bone (orbit). Caution: Do not injure the eyeball and the periosteum. Avoid the venous plexus and arteries! Pay attention to any pain from needling. No needle manipulation! After removal of the needle, compress the site for 10 minutes. Haematomas may still occur (inform the patient prior to needling). Moxibustion is contraindicated. Alternative, less difficult points for eye disorders: **BL-2, T.B.-23**, **G.B.-1**, **ST-2, Ex-HN-5** (*taiyang*), **Ex-HN-4** (*yuyao*).

Actions/Indications

- Disorders of the eyes

Ex-HN-8 *Upper* Yingxiang SHANGYINGXIANG

Location

At the upper end of the nasiolabial groove, at the junction of the maxilla and the nasal cavity. Note: *bitong* (Clear Nose) is an alternative name for this point.

How to find

At the upper end of the nasiolabial groove, palpate for the bony margin of the nasal cavity. Here, **Ex-HN-8** (*shangyingxiang* or *bitong*) is located at the junction of the nose and the cheek.

Needling

0.3–0.5 cun vertically towards the centre of the nasal cavity.

Actions/Indications

- Clears Heat, opens the nasal passages

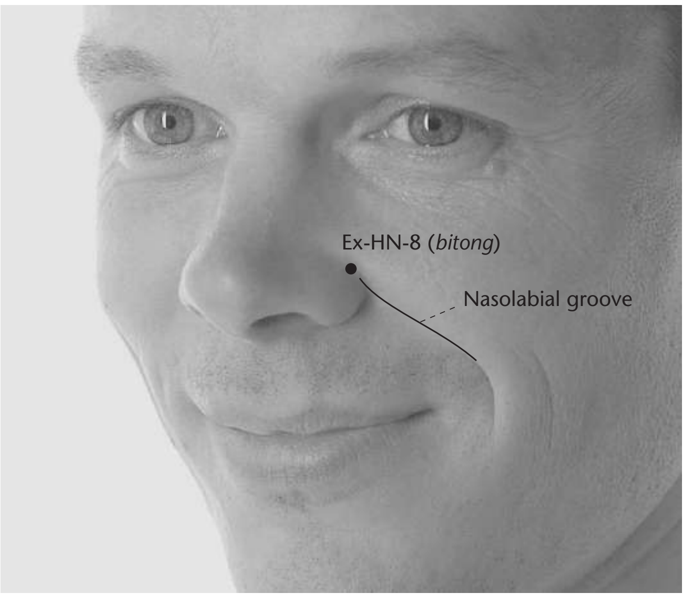
Ex-HN-8 (*bitong*)
Nasolabial groove

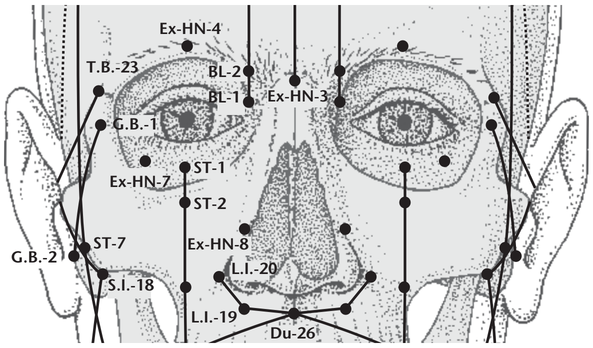
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
Ex-HN-8
G.B.-2
L.I.-20
S.I.-18
L.I.-19
Du-26

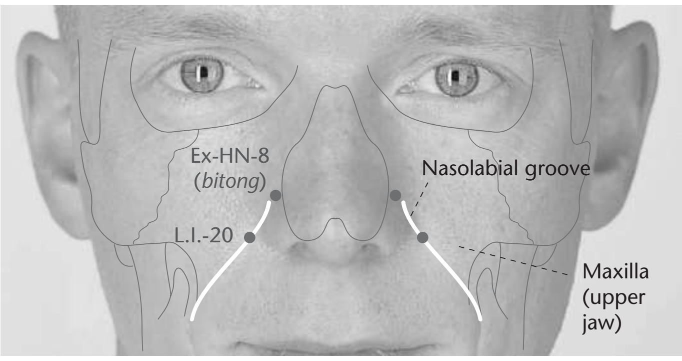
Ex-HN-8
(*bitong*)
Nasolabial groove
L.I.-20
Maxilla
(upper
jaw)

Inner Yingxiang NEIYINGXIANG Ex-HN-9

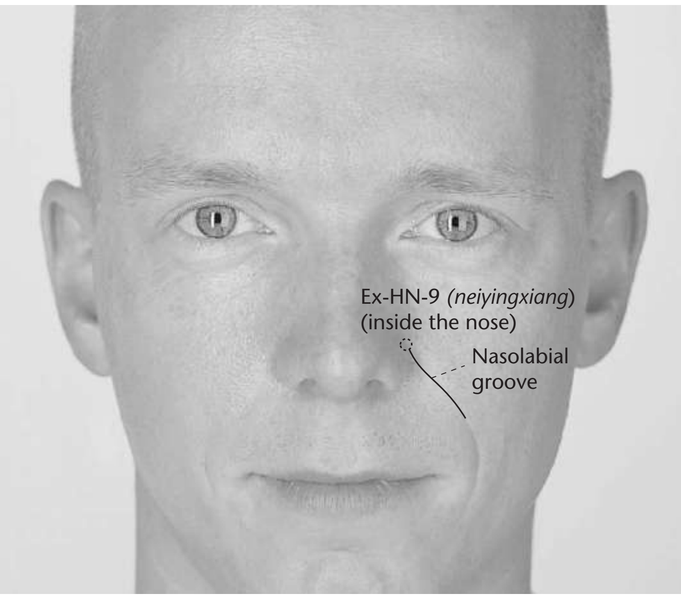
Ex-HN-9 (*neiyingxiang*)
(inside the nose)
Nasolabial groove

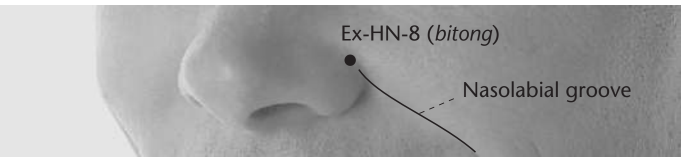
Ex-HN-8 (*bitong*)
Nasolabial groove

Location

In the nasal cavity, at the junction of the nasal bone and the nasal cartilage.

How to find

This point lies ‘opposite’ the external point → **Ex-HN-8** (*shangyingxiang* or *bitong*), which is located at the upper end of the nasolabial groove.

Needling

Prick to bleed with a needle, lancet or three-edged needle. Caution: Contraindicated in patients with bleeding disorders (or taking anticoagulants). Needling may be painful!

Actions/Indications

- Clears Heat, drains Fire

Ex-HN-10 *Gathering Spring* JUQUAN

Location

With maximal extension of the tongue, in the centre of the tongue body.

How to find

With maximal extension of the tongue in the centre of the tongue body.

Needling

Vertically 0.2 cun. Needling may be painful!

Actions/Indications

- Deviation of the tongue, impaired mobility or atrophy of the musculature of the tongue: for example, after a stroke or with loss of sense of taste

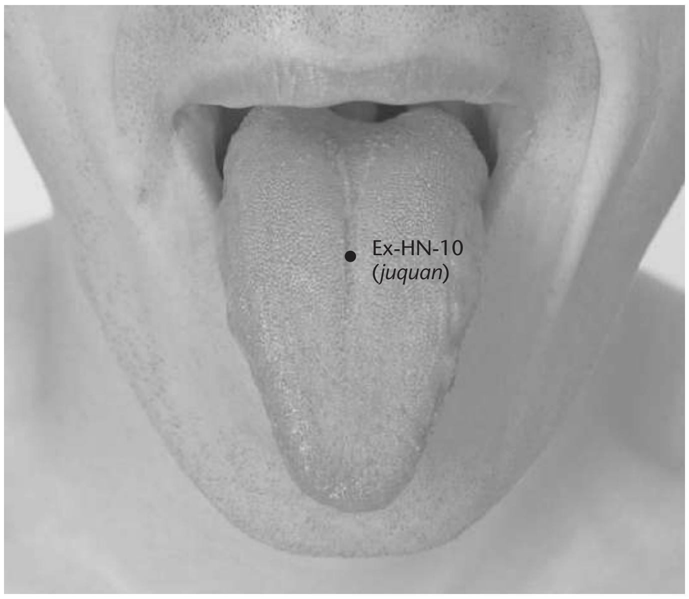

Sea Spring HAIQUAN Ex-HN-11

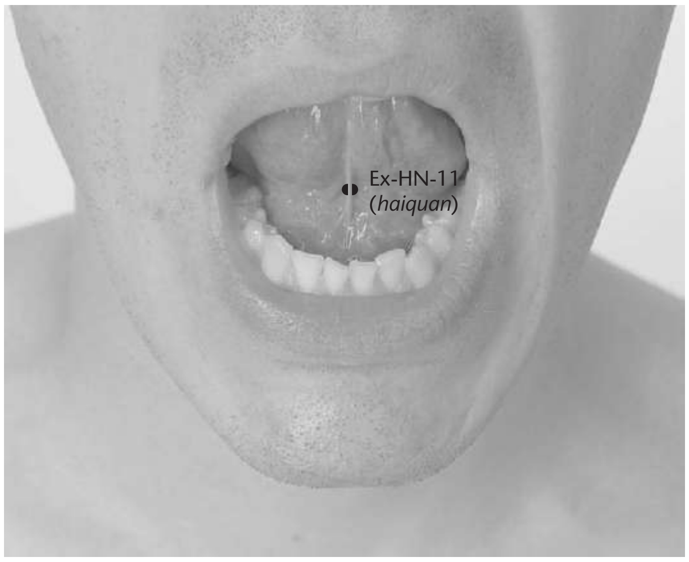
Ex-HN-11
(haiquan)

Location

Below the tongue, in the centre of the frenulum, between the points ➞ **Ex-HN-12** (*jinjin*) and ➞ **Ex-HN-13** (*yuye*).

How to find

Ask the patient to roll their tongue upward to reveal the frenulum. To avoid injury to the frenulum, this point should be located at the base of the frenulum.

Needling

Vertically 0.2 cun. Remove needle immediately after short stimulation. Needling may be painful!

Actions/Indications

- Mouth and tongue ulcers, hiccups

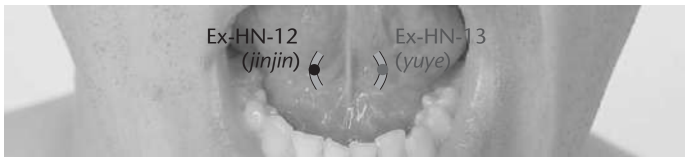
Ex-HN-12
(jinjin)
Ex-HN-13
(yuye)

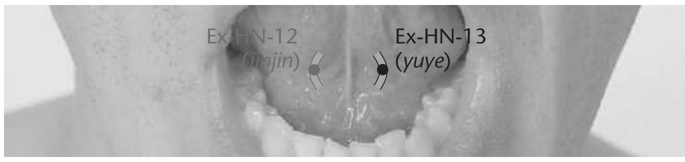
Ex-HN-12
(jinjin)
Ex-HN-13
(yuye)

Ex-HN-12 *Golden Liquid* JINJIN

Location

On the underside of the tongue, on the lingual vein to the left of the frenulum.

How to find

Ask the patient to roll their tongue upward to reveal the lingual veins.

Needling

Prick to bleed: Briefly puncture with a needle, lancet or three-edged needle and let bleed. Often used together with → **Ex-HN-13** (*yuye*). Caution: Contraindicated with bleeding disorders or in patients taking anticoagulant medication.

Actions/Indications

- Dispels Heat and Wind

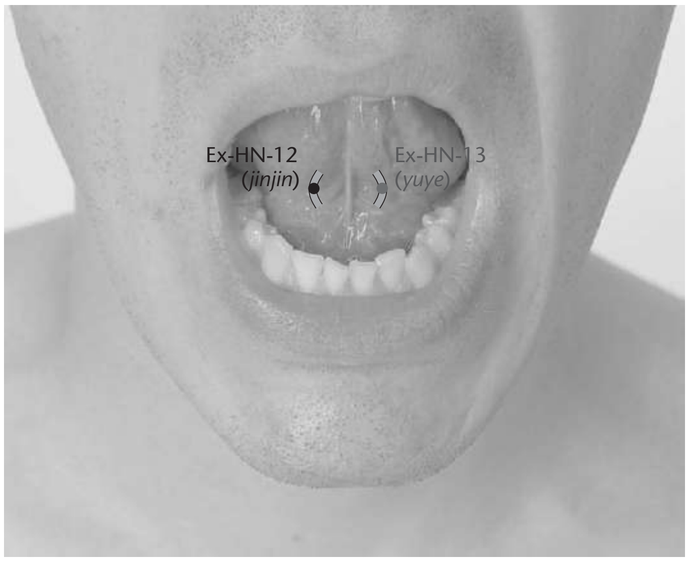

Jade Fluid YUYE Ex-HN-13

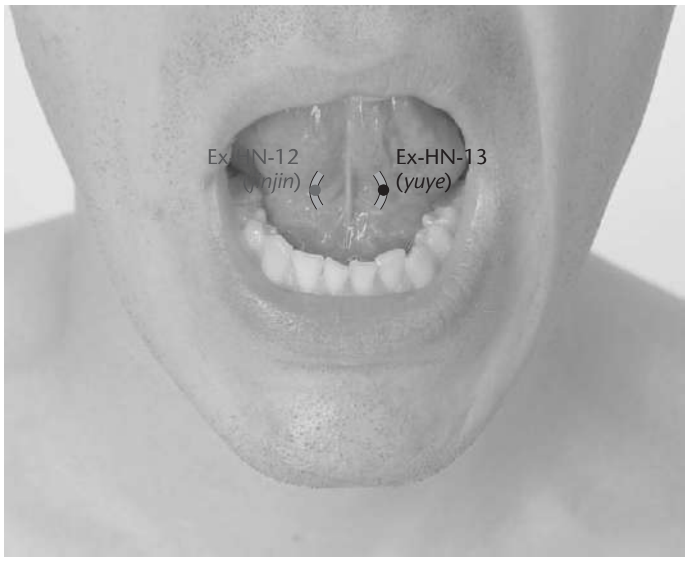

Location

On the underside of the tongue, on the great lingual vein to the right of the frenulum.

How to find

Ask the patient to roll their tongue upward to reveal the lingual veins.

Needling

Prick to bleed: Briefly puncture with a needle, lancet or three-edged needle and let bleed. Often used together with → **Ex-HN-12** (*jinjin*). Caution: Contraindicated with bleeding disorders or in patients taking anticoagulant medication.

Actions/Indications

- Dispels Heat and Wind

Ex-HN-14 *Eye Brightening* YIMING

Location

At the junction of the head and the occiput, posterior to the mastoid bone.

How to find

Ex-HN-14 (*yiming*) is located on a line connecting → **T.B.-17** (below the earlobe, between the maxilla and the mastoid) and → **G.B.-20** (at the junction of the occiput and the neck, in the centre of the dorsolateral depression). From → **T.B.-17**, measure 1 cun towards the posterior and there locate *yiming* posterior to the mastoid.

Located slightly superior and posterior to *yiming* is → **Ex-HN** (*anmian*), in the corner between the mastoid and the occiput.

Needling

Vertically 0.5–1 cun

Actions/Indications

- Disorders of the eyes and ears, dizziness, insomnia

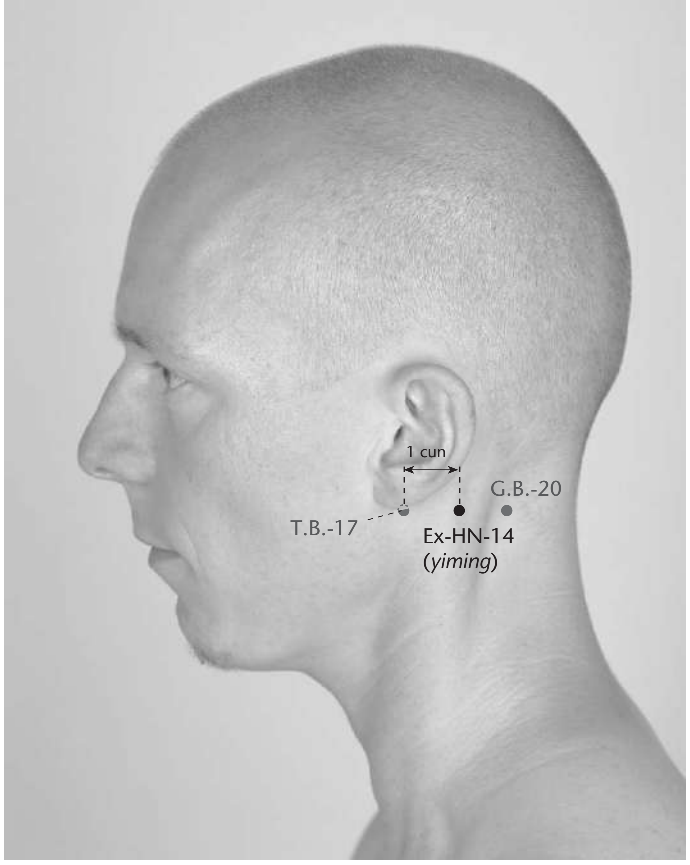
1 cun
G.B.-20
T.B.-17
Ex-HN-14
(*yiming*)

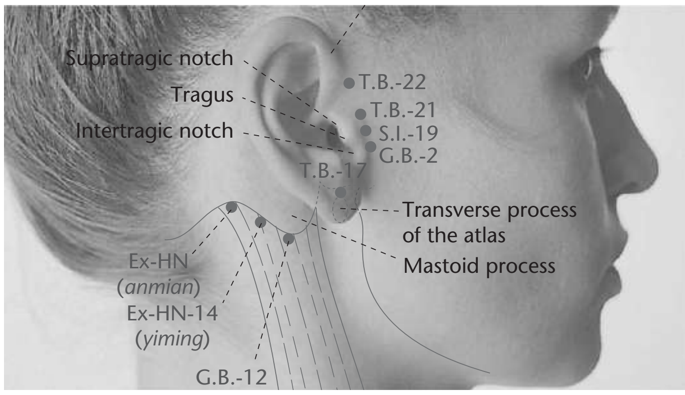
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Transverse process of the atlas
Mastoid process
Ex-HN (*anmian*)
Ex-HN-14 (*yiming*)
G.B.-12

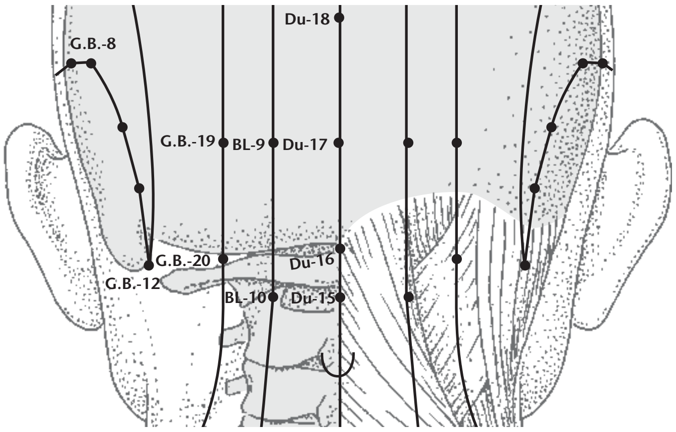
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
G.B.-20
Du-16
G.B.-12
BL-10
Du-15

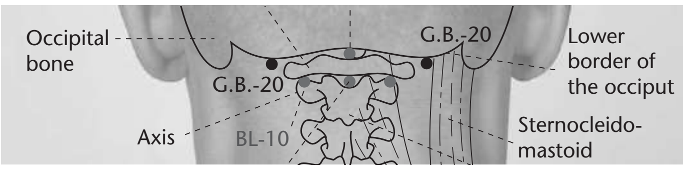
Occipital bone
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid

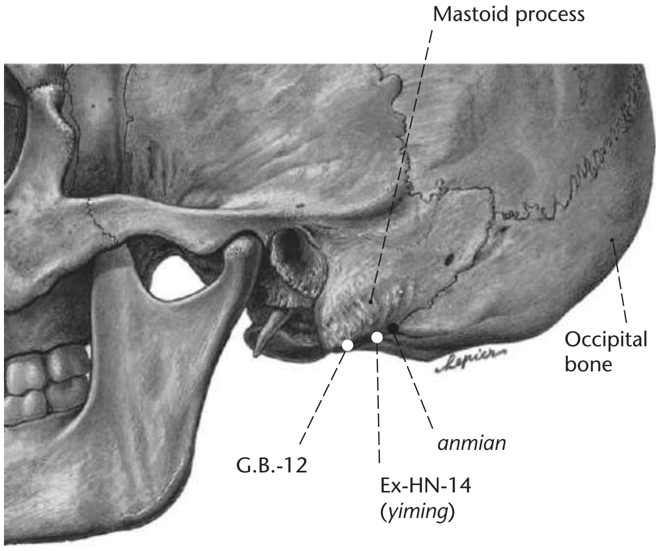
Mastoid process
Occipital bone
anmian
G.B.-12
Ex-HN-14
(*yiming*)

Hundred Taxations JINGBAILAO/BAILAO Ex-HN-15

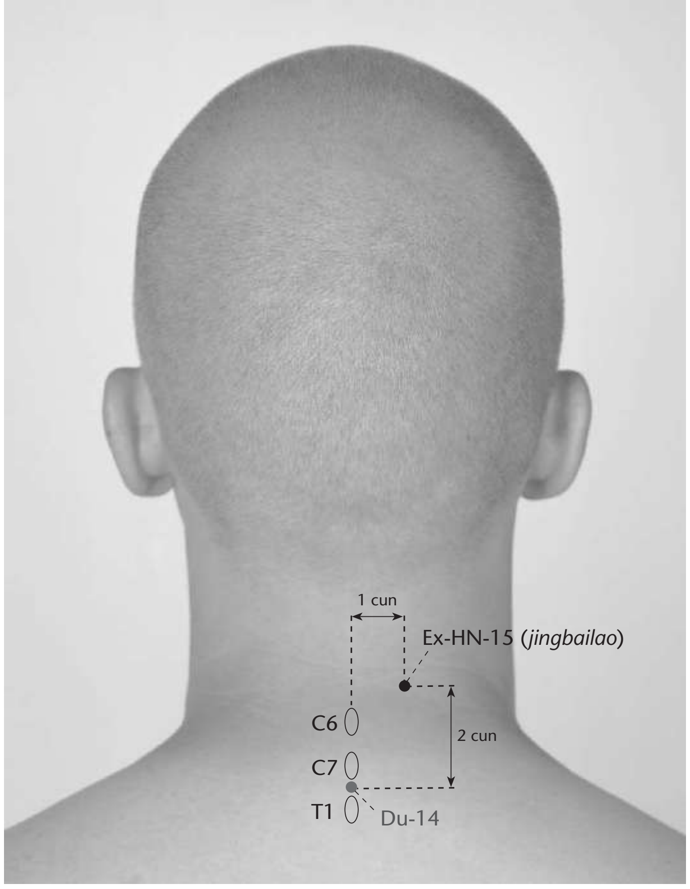
1 cun
Ex-HN-15 (*jingbailao*)
C6
2 cun
C7
T1
Du-14

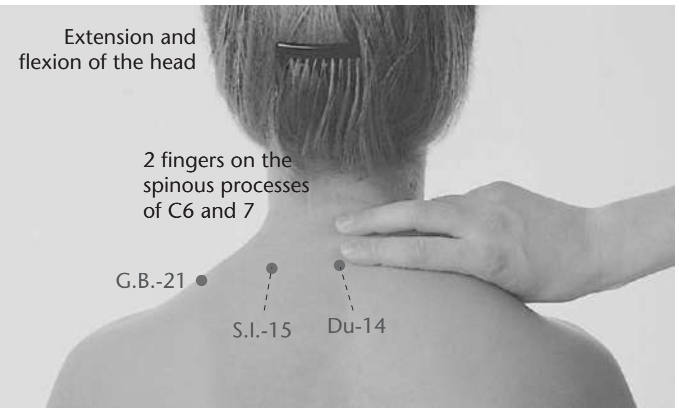
Extension and
flexion of the head
2 fingers on the
spinous processes
of C6 and 7
G.B.-21
S.I.-15
Du-14

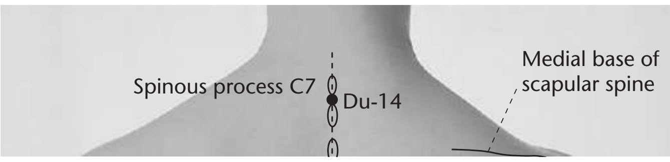
Spinous process C7
Du-14
Medial base of
scapular spine

Location

2 cun superior to the lower border of the spinous process of the 7th cervical vertebra (C7) and 1 cun lateral to the midline.

How to find

For locating C7, see → 3.4.1. → **Du-14** (*dazhui*) is located below the spinous process of C7. From there, measure 2 cun in a superior and 1 cun in a lateral direction.

Needling

Vertically 0.5–0.8 cun

Actions/Indications

- Dispels Wind and Dampness
- Opens the channel

Ex-HN *Upper Brightness* SHANGMING

Location

Directly superior to the pupil, below the margin of the orbit.

How to find

With the patient looking straight ahead, palpate for the margin of the orbit directly above the pupil. **Ex-HN** (*shangming*) is located between the upper margin of the orbit and the eyeball.

Needling

Gently push the eyeball downward. Slowly insert the needle 0.5–1 cun vertically into the fatty tissue immediately below the bone (orbit). Caution: Pay attention to any pain from needling. No manipulation! After removal of the needle, compress the site for 10 minutes. Haematomas may still occur (inform the patient prior to needling). Alternative points with less risk of complications are **BL-2, T.B.-23**, **G.B.-1, ST-2, Ex-HN-5** (*taiyang*), **Ex-HN-4** (*yuyao*).

Actions/Indications

- Disorders of the eyes

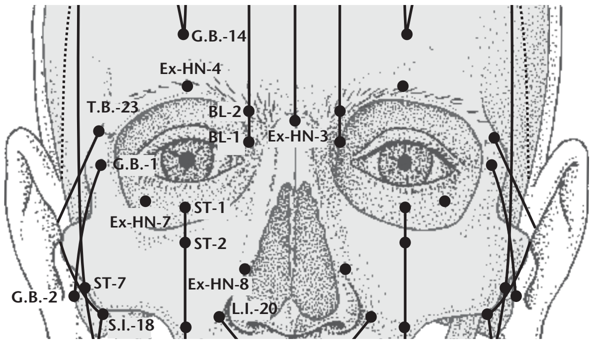
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
G.B.-2
Ex-HN-8
L.I.-20
S.I.-18

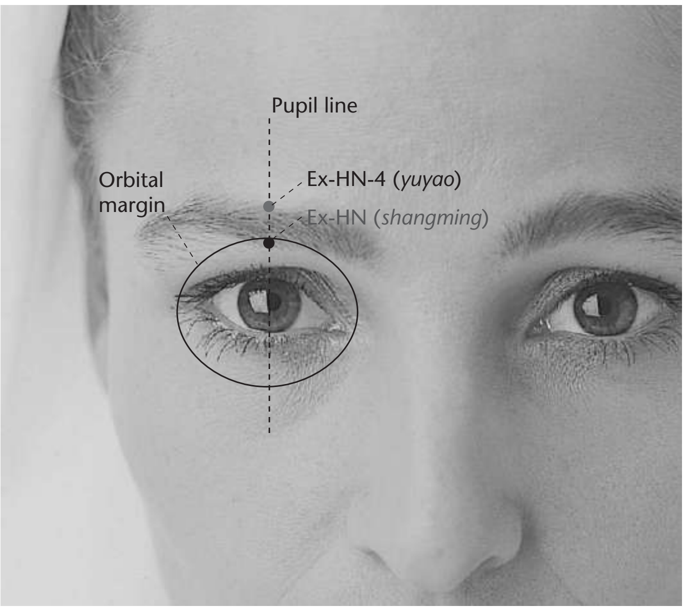
Pupil line
Orbital margin
Ex-HN-4 (*yuyao*)
Ex-HN (*shangming*)

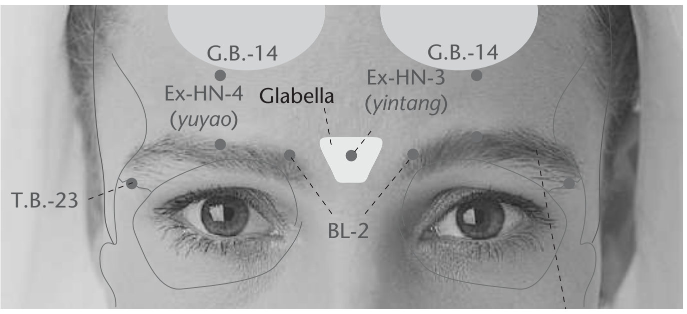
G.B.-14
G.B.-14
Ex-HN-4 (*yuyao*)
Glabella
Ex-HN-3 (*yintang*)
T.B.-23
BL-2

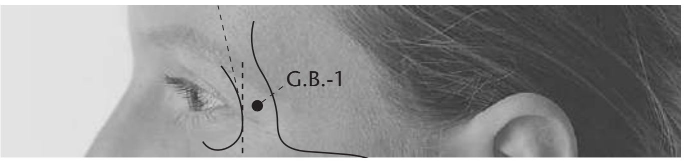

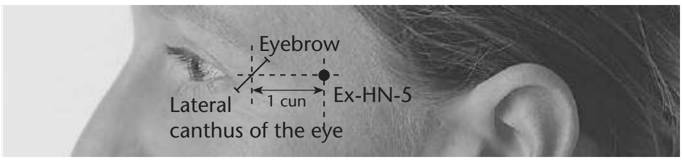
Eyebrow
Lateral canthus of the eye
1 cun
Ex-HN-5

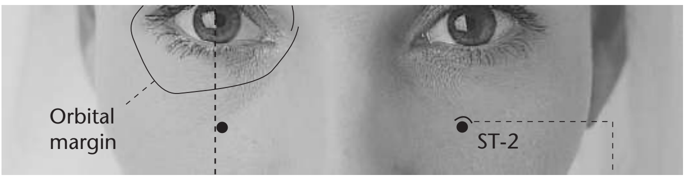
Orbital margin
ST-2

Peaceful Sleep ANMIAN Ex-HN

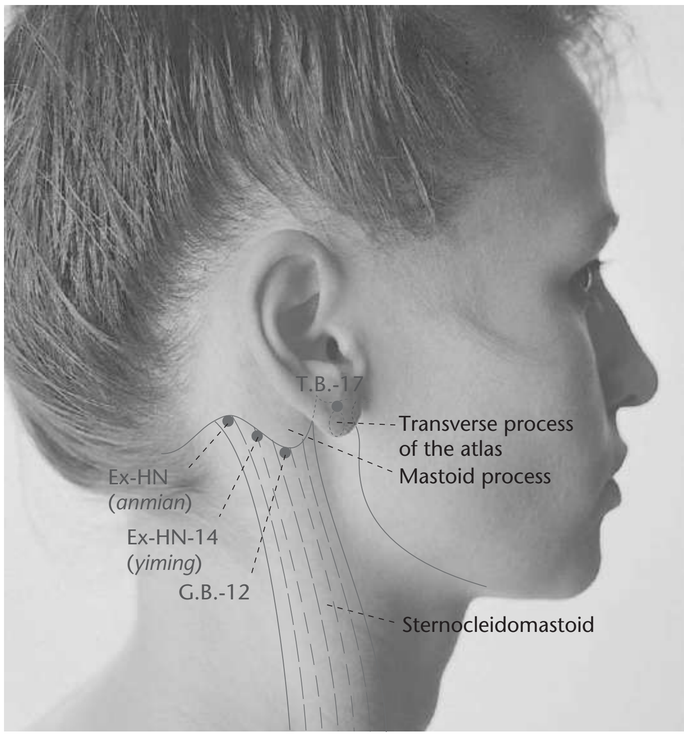
T.B.-17
Transverse process
of the atlas
Mastoid process
Ex-HN
(*anmian*)
Ex-HN-14
(*yiming*)
G.B.-12
Sternocleidomastoid

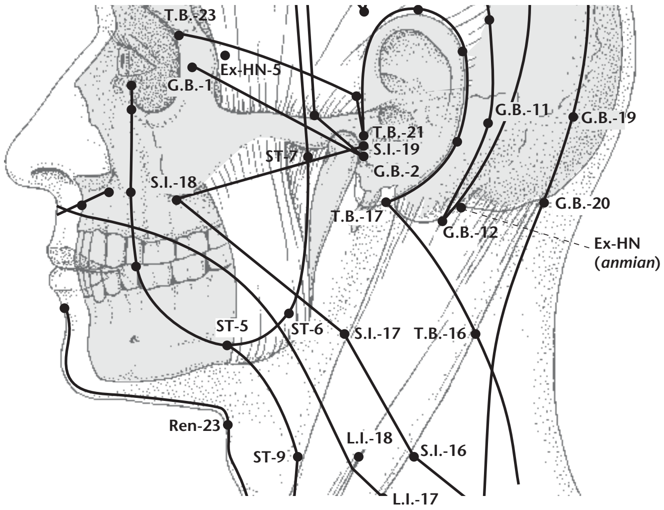
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
ST-7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17

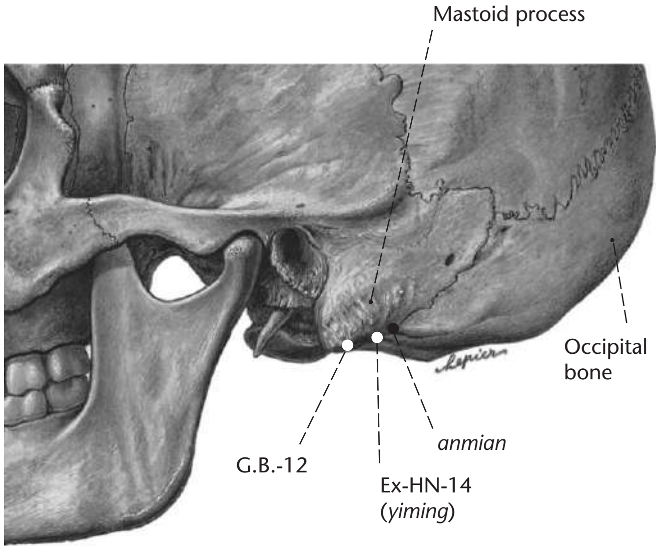
Mastoid process
Occipital
bone
anmian
G.B.-12
Ex-HN-14
(*yiming*)

Location

Posterior to the ear, between → **T.B.-17** and → **G.B.-20**, posterior to the mastoid process.

How to find

First, locate → **T.B.-17** (directly below the earlobe, in the depression between the lower jaw and the mastoid process (→ 3.1) that forms when the mouth is open). Next, locate → **G.B.-20** on the lower border of the occiput, in the depression between the origins of the sternocleidomastoid and trapezius muscles. Locate **Ex-HN** (*anmian*) approximately midway between these two points, slightly posterior to the mastoid bone and superior to → **G.B.-12**, in a depression on the lower border of the occiput.

Needling

0.5–1 cun vertically or obliquely towards → **T.B.-17** (*yifeng*) or → **G.B.-20** (*fengchi*).

Actions/Indications

- Calms the *shen*

Special features

Important calming point for sleeping disorders. This extra point is not part of the WHO classification of extra points.

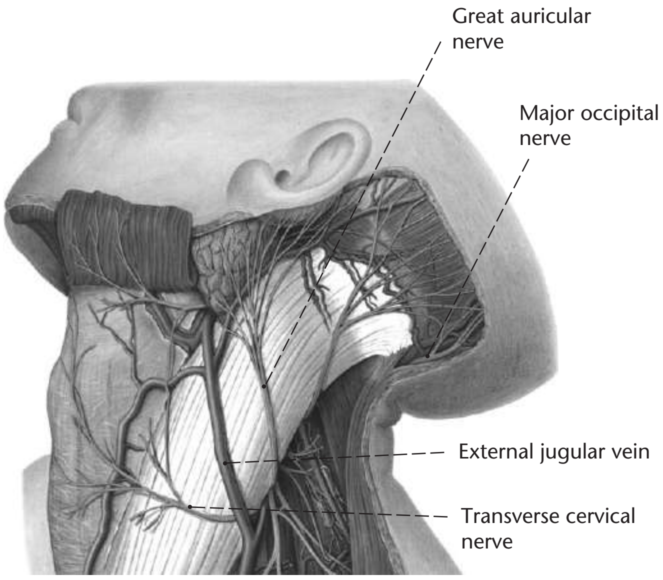
Great auricular
nerve
Major occipital
nerve
External jugular vein
Transverse cervical
nerve

Ex-HN

Cheek Centre JIALI

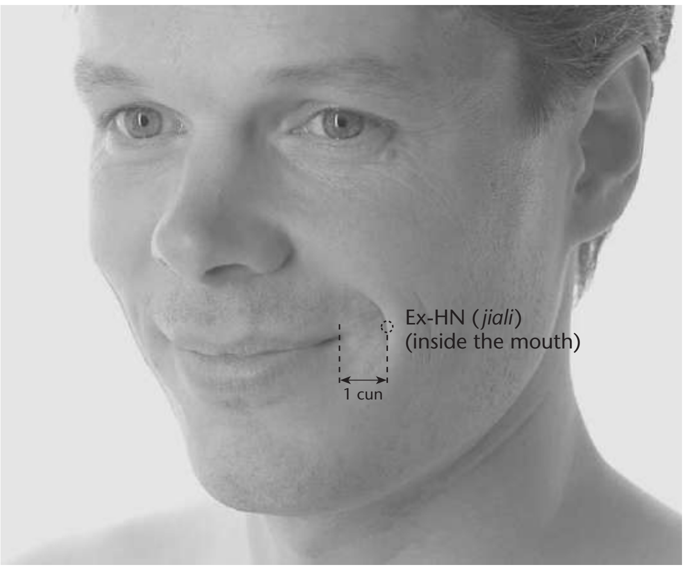

Location

Inside the mouth, on the mucosa of the cheek, 1 cun posterior to the corner of the mouth.

How to find

Locate this point on the mucosa of the cheek, 1 cun posterior to the corner of the mouth.

Needling

0.3–0.5 cun obliquely in a posterior direction, prick to bleed.

Actions/Indications

- Clears Heat
- For inflammations of the mouth and throat, for gastritis

Adjacent to Container of Fluids JIACHENGJIANG Ex-HN

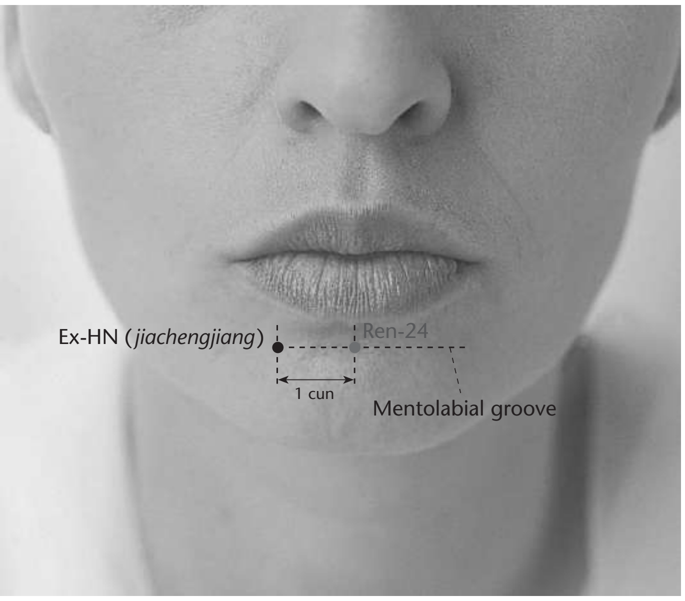
Ex-HN (*jiachengjiang*)
Ren-24
1 cun
Mentolabial groove

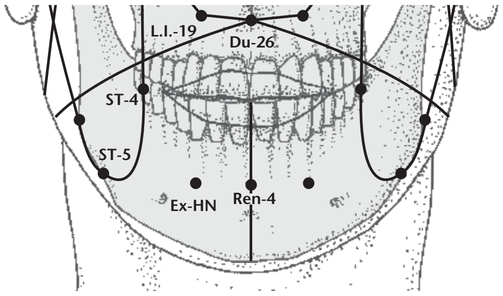
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren-4

Location

1 cun lateral to the centre of the mentolabial groove.

How to find

This point is located on the mental foramen. → **Ren-24** (*chengjiang*) is located in the centre of the mentolabial groove.

Needling

0.3–0.5 cun vertically or up to 1 cun obliquely towards the mental foramen.

Actions/Indications

- Dispels Wind, opens the channel

Ex-HN *Prominent Bone* CHONGGU ZHUIDONG

Location

Below the spinous process of the 6th cervical vertebra (C6).

How to find

This point is located on the posterior midline and thus on the *du mai*. In fact, the ‘prominent bone’ (vertebra prominens) inferior to *chonggu zhuidong* is the spinous process of either C7 or T1. → **Du-14** is located between the two. For the correct method of locating C6 and C7 → 3.4.1.

Needling

0.5–1 cun obliquely in a superior direction.

Actions/Indications

- Dispels external pathogenic factors
- Harmonises the *shen*

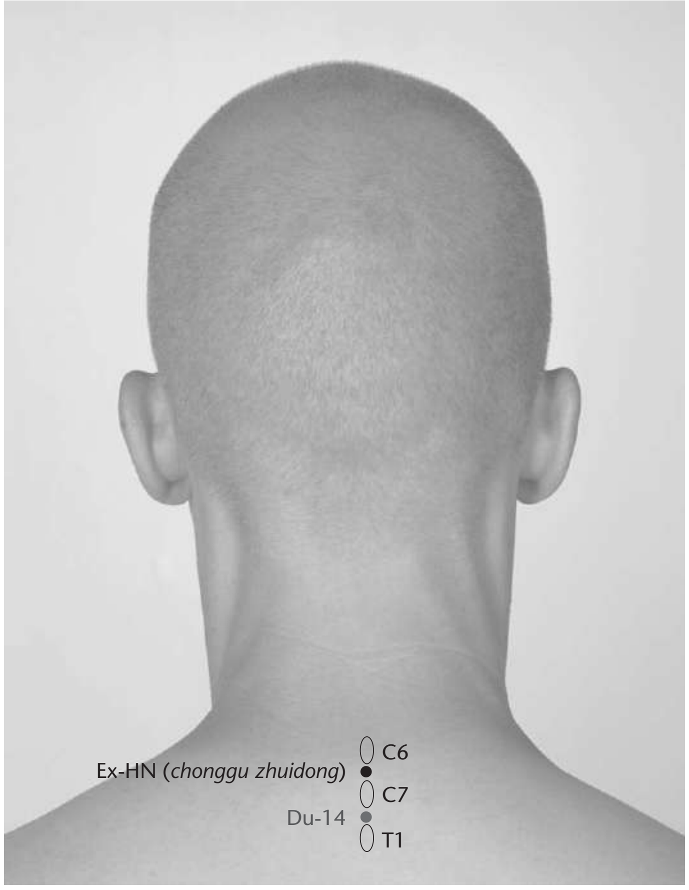
Ex-HN (*chonggu zhuidong*)
C6
C7
Du-14
T1

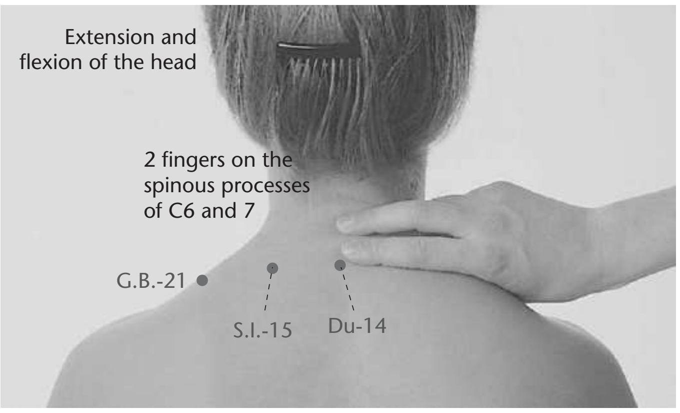
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Upper Arm JINGBI Ex-HN

Location

1 cun superior to the junction of the proximal and middle third of the clavicle.

How to find

Measure a third of the clavicle from its medial end, then measure 1 cun in a superior direction and there locate *jingbi* in the supraclavicular fossa. The point is located above the brachial plexus. → **ST-12** is very close to this point, directly superior to the midpoint of the clavicle.

Needling

Vertically 0.3–0.5 cun. During insertion or stimulation, a tingling or warm sensation should be felt radiating to the fingers. Caution: Pneumothorax.

Actions/Indications

- Paraesthesia and paralysis of the upper extremity

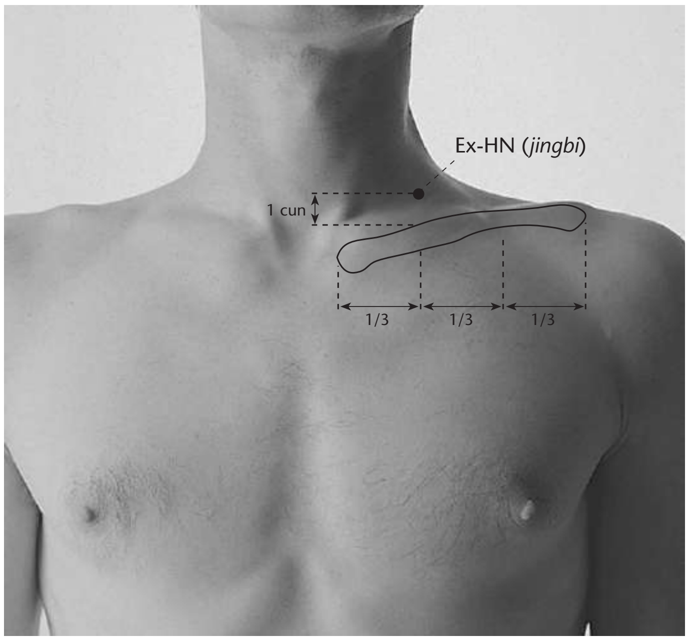
Ex-HN (*jingbi*)
1 cun
1/3
1/3
1/3

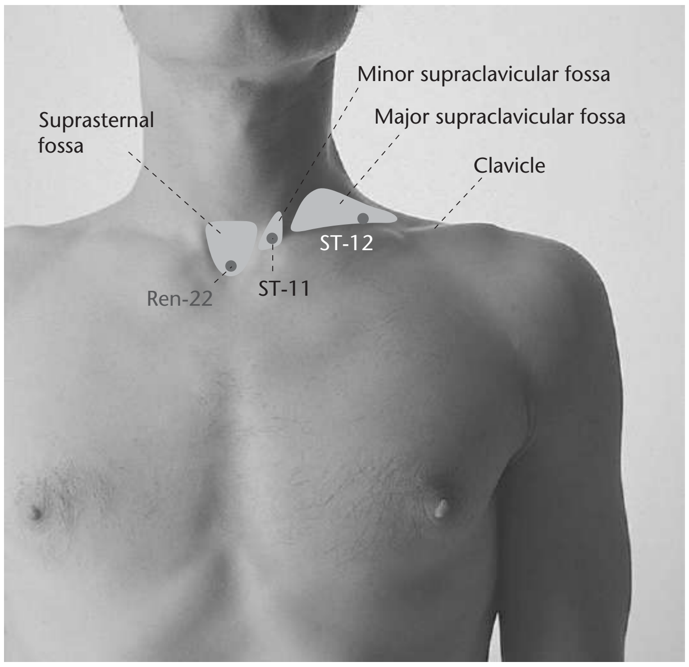
Minor supraclavicular fossa
Suprasternal fossa
Major supraclavicular fossa
Clavicle
ST-12
Ren-22
ST-11

6.2 Extra Points: Chest and Abdomen (EX-CA)

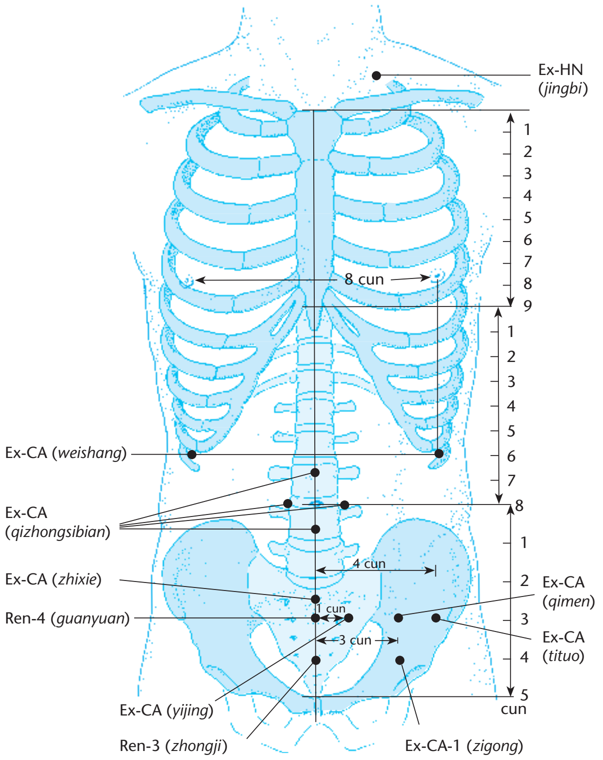
Ex-HN
(*jingbi*)
1
2
3
4
5
6
7
8
9
8 cun
1
2
3
4
5
6
7
8
Ex-CA (*weishang*)
Ex-CA
(*qizhongsibian*)
Ex-CA (*zhixie*)
Ren-4 (*guanyuan*)
4 cun
1 cun
3 cun
1
2
3
4
5
cun
Ex-CA
(*qimen*)
Ex-CA
(*tituo*)
Ex-CA (*yijing*)
Ren-3 (*zhongji*)
Ex-CA-1 (*zigong*)

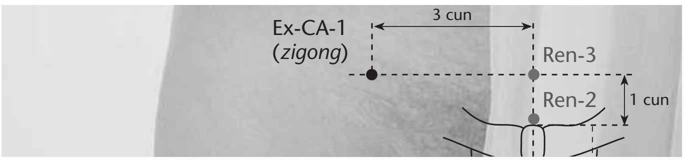
Ex-CA-1
(*zigong*)
3 cun
Ren-3
Ren-2
1 cun

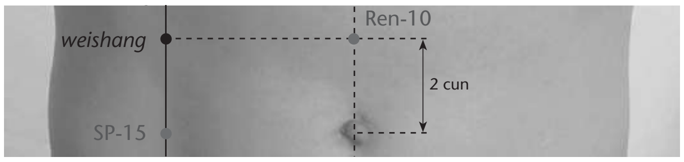
weishang
Ren-10
2 cun
SP-15

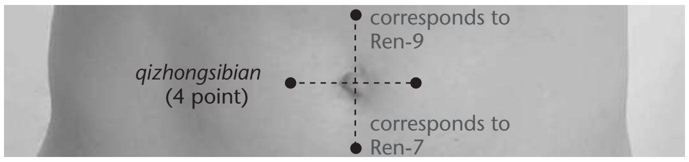
corresponds to
Ren-9
qizhongsibian
(4 point)
corresponds to
Ren-7

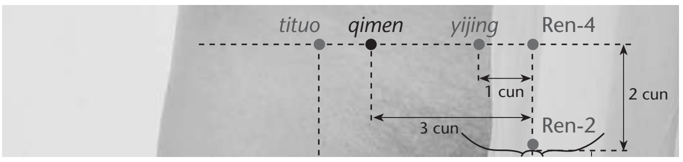
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

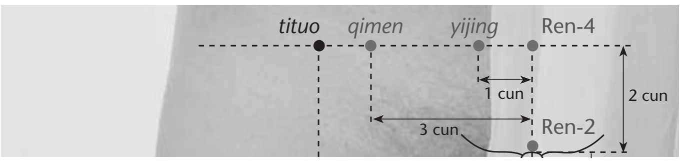
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

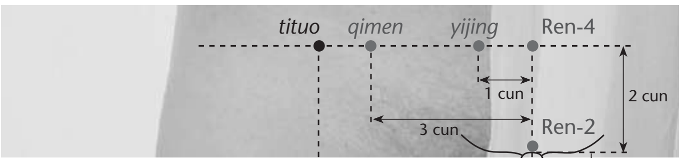
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

English abbreviation (standard)	***Pinyin* name**	**Nguyen Van Nghi, König and Wancura, Schnorrenberger**	**Shanghai College**	**Ex (Hempen)**
Ex-CA-1	*zigong*	PaM or ZP49	M-CA-18	Ex9
Further extra points				
Ex-CA	*weishang*			
Ex-CA	*qizhongsibian*			
Ex-CA	*yijing*	PaM or ZP45		
Ex-CA	*qimen*	PaM or ZP46		
Ex-CA	*tituo*	NP39	M-CA-4	
Ex-CA	*zhixie*	NP38	N-CA31	

Palace of the Child ZIGONG Ex-CA-1

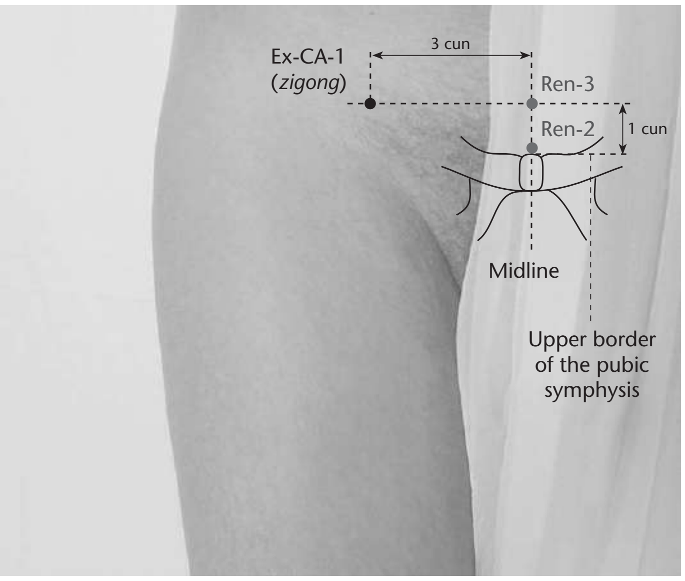
Ex-CA-1
(*zigong*)
3 cun
Ren-3
Ren-2
1 cun
Midline
Upper border of the pubic symphysis

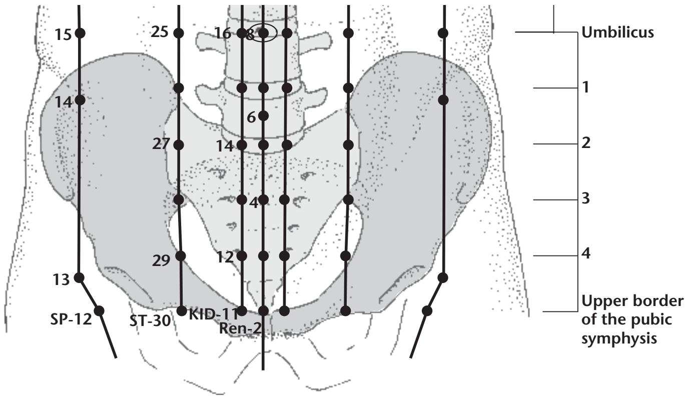
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

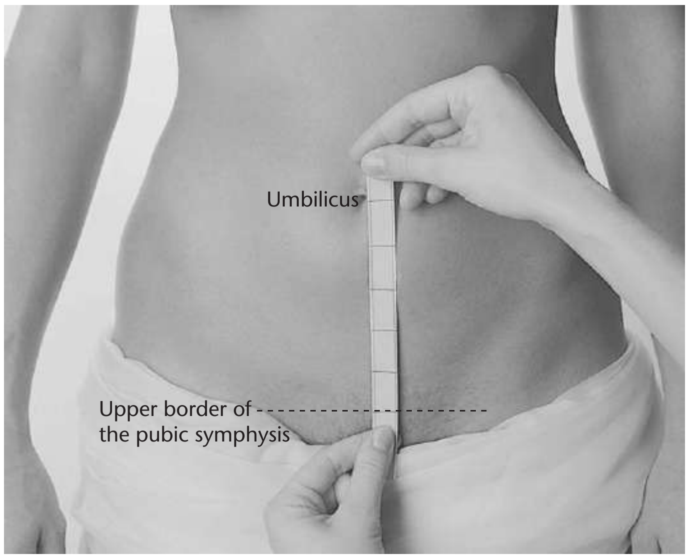
Umbilicus
Upper border of the pubic symphysis

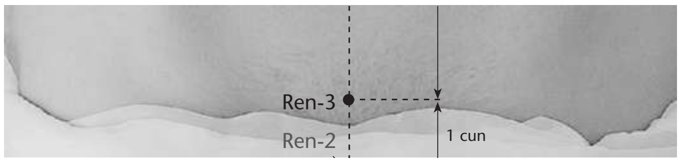
Ren-3
Ren-2
1 cun

Location

3 cun lateral to the anterior midline and 1 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 1 cun in a superior direction – this is the location of → **Ren-3**. From there, measure 3 cun in a lateral direction to locate *zigong*.

Located on the same level are → **KID-12** (0.5 cun from the midline) and → **ST-29** (2 cun lateral to the midline).

Needling

0.5–1 cun vertically or 1–2 cun obliquely towards the upper border of the pubic symphysis. Caution: Peritoneum, pregnancy, full bladder.

Actions/Indications

- Strengthens and raises Qi
- Regulates menstruation
- Alleviates pain in the abdomen and the lumbar region

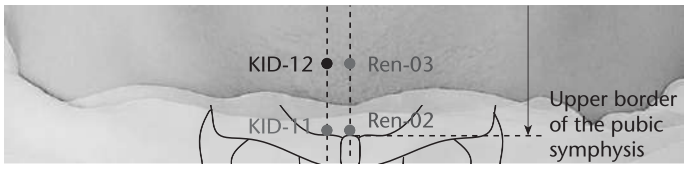
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

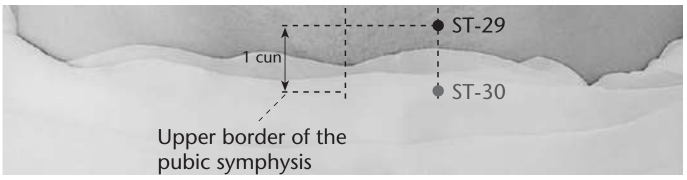
ST-29
1 cun
ST-30
Upper border of the pubic symphysis

Ex-CA Lifting the Stomach WEISHANG

Location

On the Spleen channel, 4 cun lateral and 2 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape ➞ 2.3.1). From the centre of the umbilicus, measure 2 cun in a superior direction (location of ➞ **Ren-10**) and then locate *weishang* 4 cun lateral to the midline. 1 cun superior to ➞ **Ex-CA** (*weishang*) is ➞ **SP-16**.
Located on the same level are ➞ **Ren-10/KID-17/ST-23** (on the midline/0.5 cun lateral to the midline/2 cun lateral to the midline).

Needling

2–3 cun obliquely in the direction of the umbilicus. Caution: Peritoneum, pregnancy.

Actions/Indications

- Gastroptosis, abdominal pain

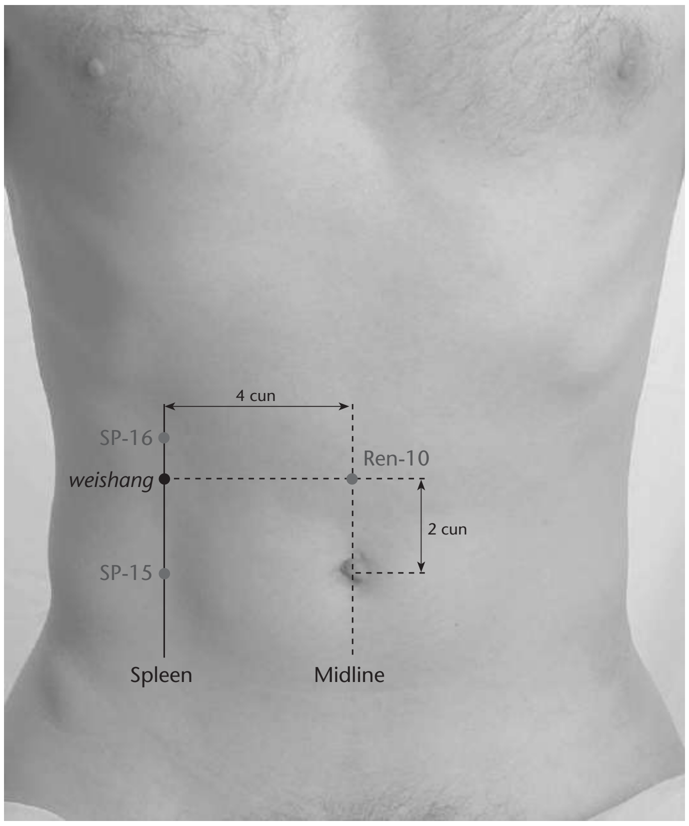
4 cun
SP-16
weishang
Ren-10
2 cun
SP-15
Spleen
Midline

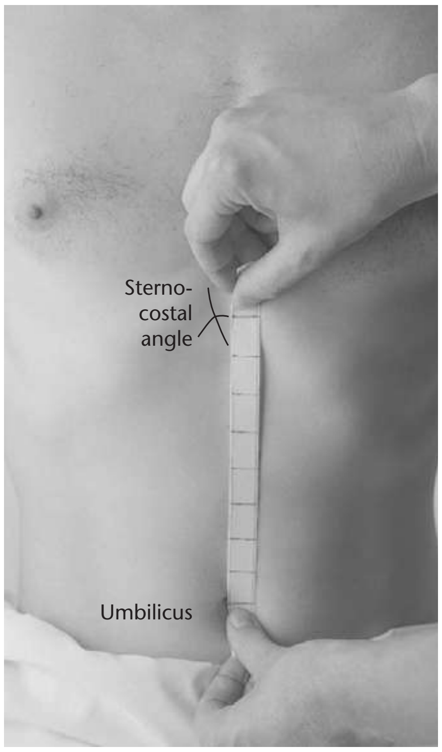
Sterno-
costal
angle
Umbilicus

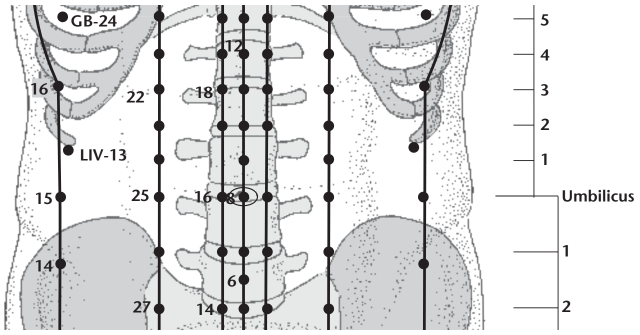
GB-24
12
16
22
18
LIV-13
15
25
16
8
14
27
6
14
5
4
3
2
1
Umbilicus
1
2

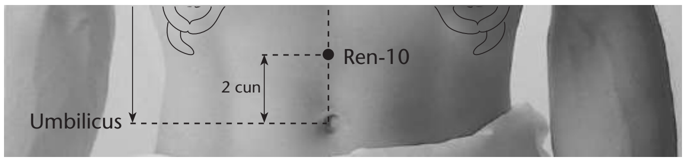
Ren-10
2 cun
Umbilicus

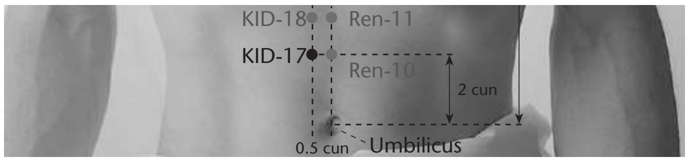
KID-18
Ren-11
KID-17
Ren-10
2 cun
0.5 cun
Umbilicus

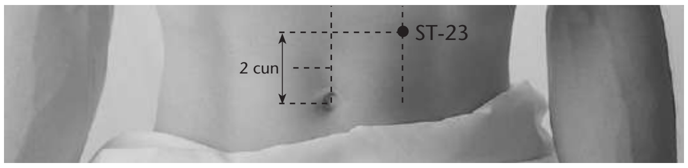

Four Points Around the Umbilicus QIZHONGSIBIAN Ex-CA

Location

Four points, 1 cun lateral, superior and inferior to the umbilicus.

How to find

These four points are arranged in a star shape around the umbilicus.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Distension, diarrhoea, dyspepsia, dysmenorrhoea

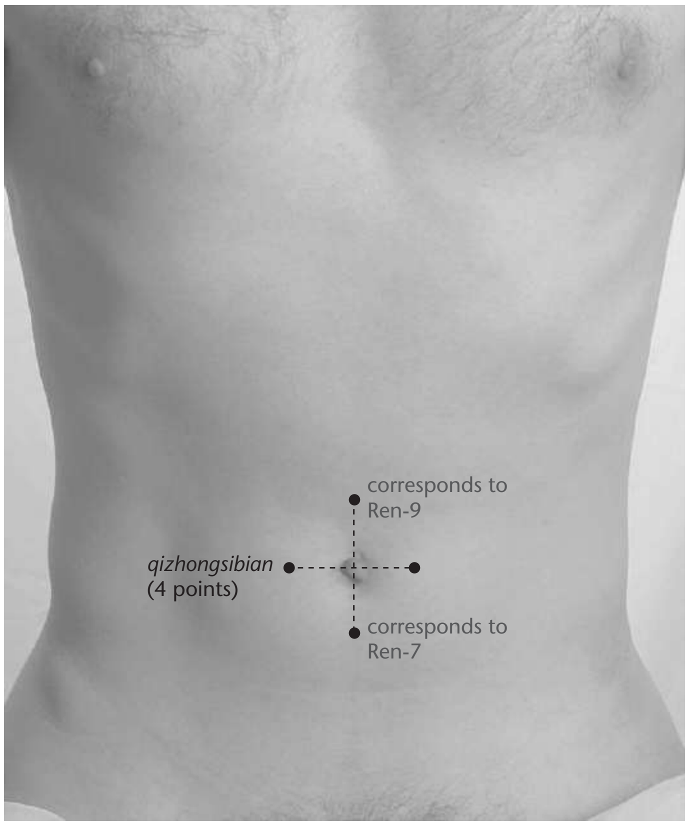
corresponds to Ren-9
qizhongsibian (4 points)
corresponds to Ren-7

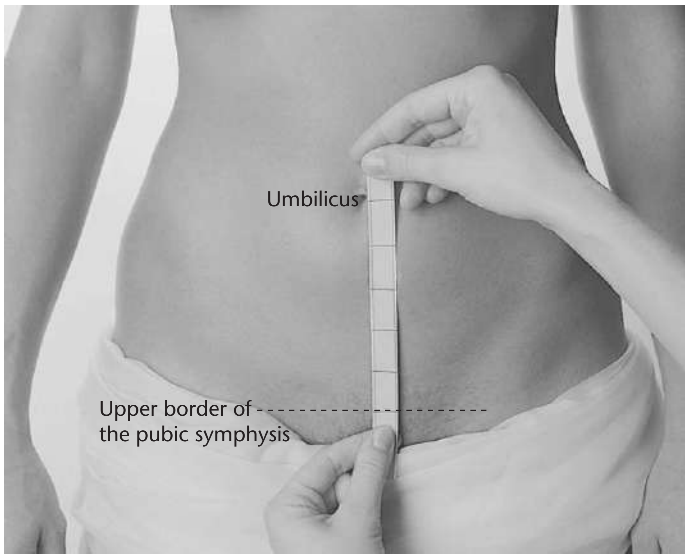
Umbilicus
Upper border of the pubic symphysis

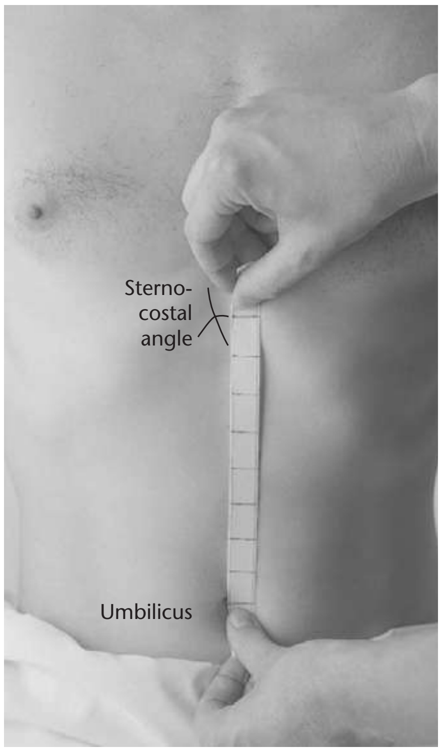
Sterno-costal angle
Umbilicus

Ex-CA Loss of Semen YIJING

Location

1 cun lateral to the anterior midline and 2 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun in a cranial direction (→ **Ren-4**) and, from there, measure 1 cun in a lateral direction. There, locate *yijing*.

Located on the same level are → **KID-13** (0.5 cun lateral to the midline), **Ex-CA** (*qimen*, 3 cun lateral to the midline), → **ST-28** (2 cun lateral to the midline) and **Ex-CA** (*tituo*, 4 cun lateral to the midline). → **G.B.-27** is located medial to the ASIS, approximately on the same level.

Needling

0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

Ejaculation disorders, impotence, scrotal eczema

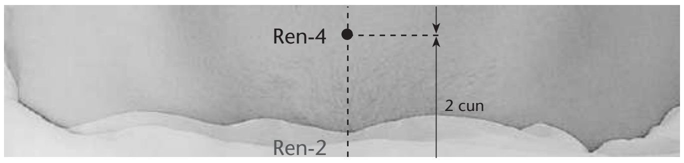
Ren-4
2 cun
Ren-2

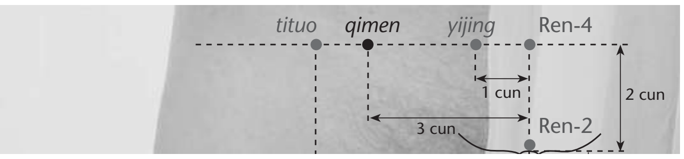
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

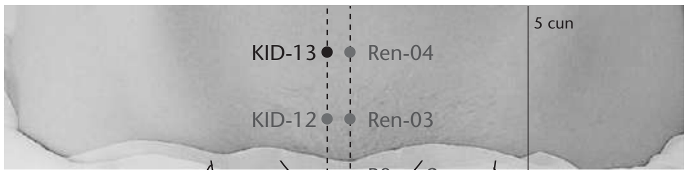
5 cun
KID-13
Ren-04
KID-12
Ren-03

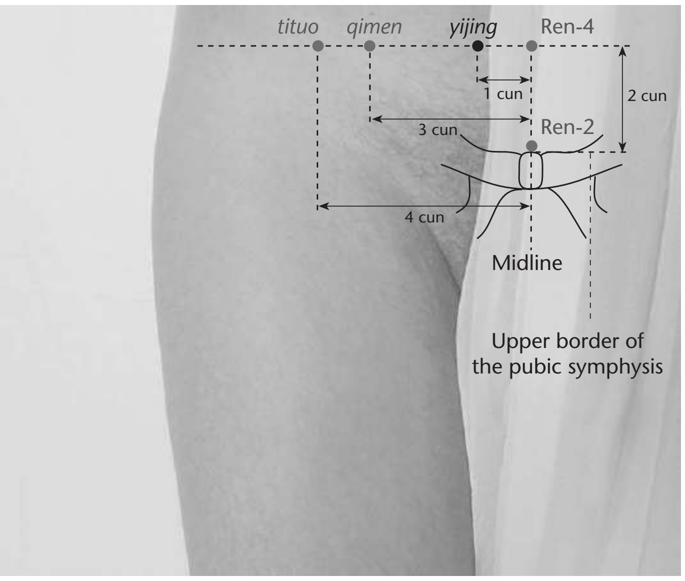
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2
4 cun
Midline
Upper border of the pubic symphysis

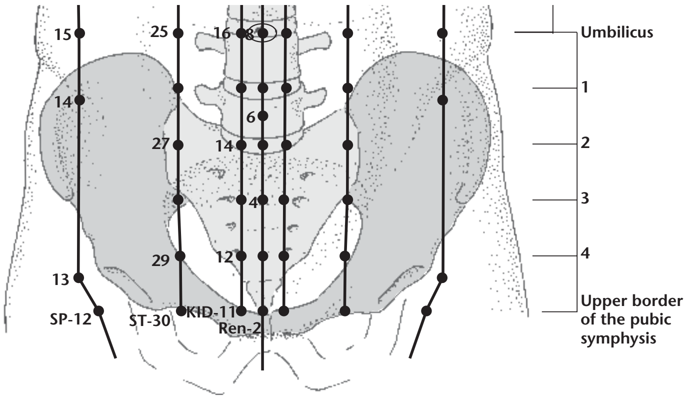
15
25
16
8
Umbilicus
14
1
6
27
14
2
4
3
29
12
4
13
SP-12
ST-30
KID-11
Ren-2
Upper border of the pubic symphysis

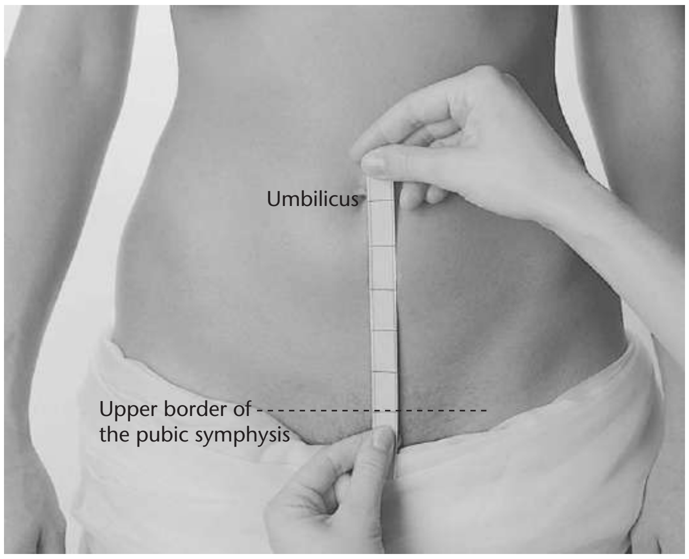
Umbilicus
Upper border of the pubic symphysis

Qi Gate QIMEN Ex-CA

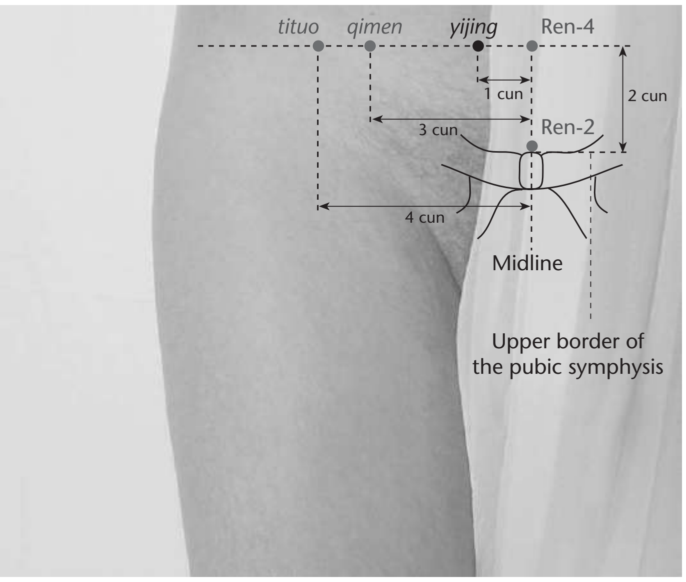
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2
4 cun
Midline
Upper border of the pubic symphysis

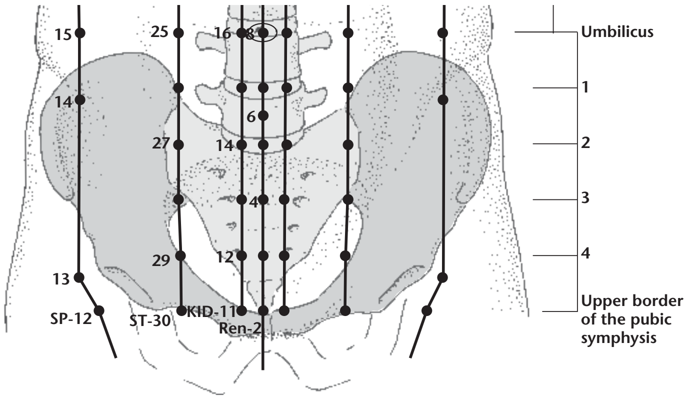
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

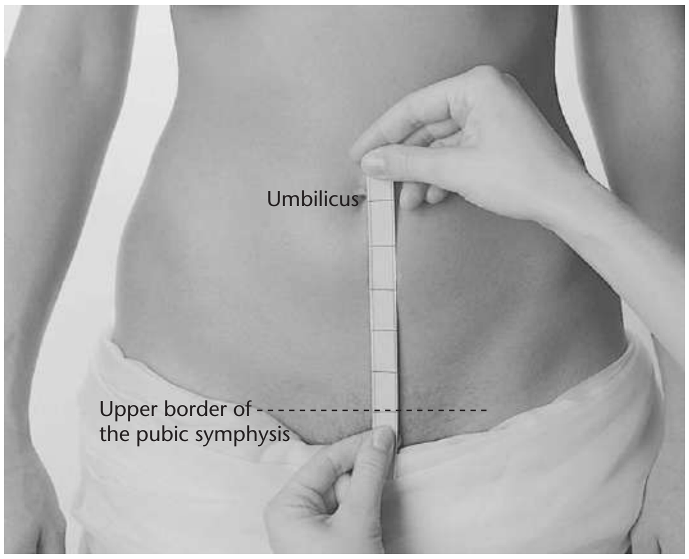
Umbilicus
Upper border of the pubic symphysis

Location

3 cun lateral to the anterior midline and 2 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun in a cranial direction (→ **Ren-4**) and, from there, measure 3 cun in a lateral direction. There, locate *qimen*.

Located on the same level are → **KID-13** (0.5 cun lateral to the midline), **Ex-CA** (*yijing*, 1 cun lateral to the midline), → **ST-28** (2 cun lateral to the midline) and **Ex-CA** (*tituo*, 4 cun lateral to the midline). → **G.B.-27** is located medial to the ASIS, approximately on the same level.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Metrorrhagia, female infertility, orchitis, urinary tract infections, persistent lochial discharge

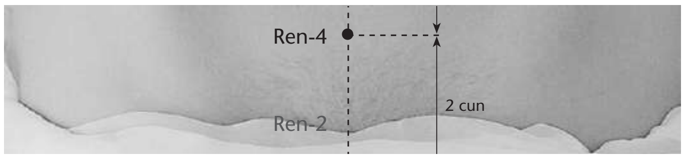
Ren-4
2 cun
Ren-2

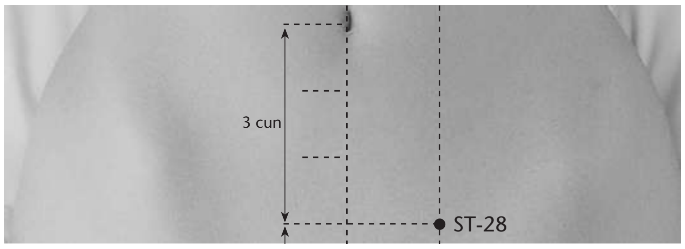

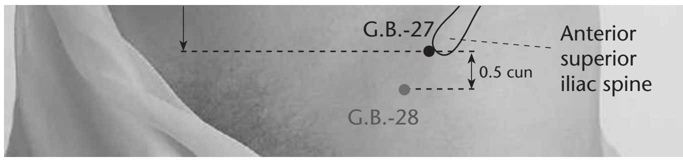
G.B.-27
0.5 cun
G.B.-28
Anterior superior iliac spine

Ex-CA Lift and Support TITUO

Location

4 cun lateral to the anterior midline and 2 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun in a cranial direction (→ **Ren-4**) and from there measure 4 cun in a lateral direction. There, medial to the ASIS, locate *tituo*.
Located on the same level are → **KID-13** (0.5 cun lateral to the midline), → **Ex-CA** (*yijing*, 1 cun lateral to the midline), → **ST-28** (2 cun lateral to the midline) and → **Ex-CA** (*qimen*, 3 cun lateral to the midline). → **G.B.-27** is located medial to the ASIS, approximately on the same level.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Strengthens the rising Qi and alleviates organ prolapse

Special features

In modern TCM, *tituo* is considered one of the most important points for prolapse of the uterus.

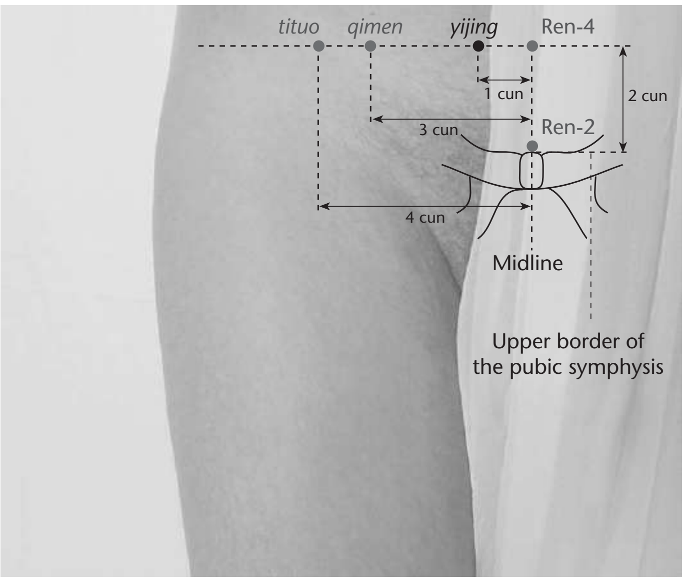
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2
4 cun
Midline
Upper border of the pubic symphysis

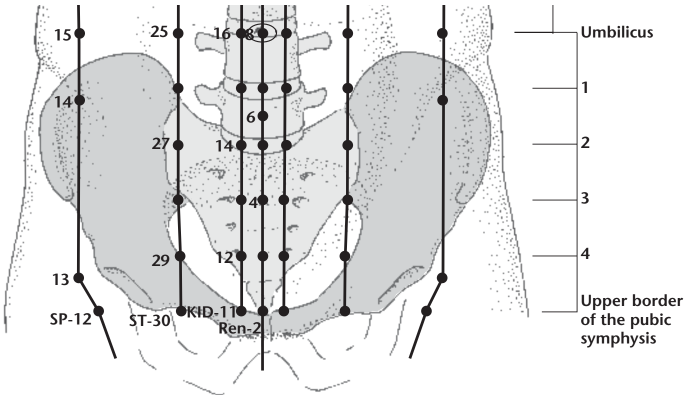
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

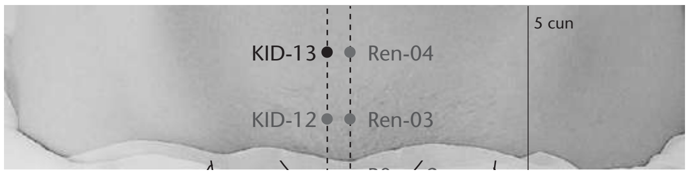
5 cun
KID-13
Ren-04
KID-12
Ren-03

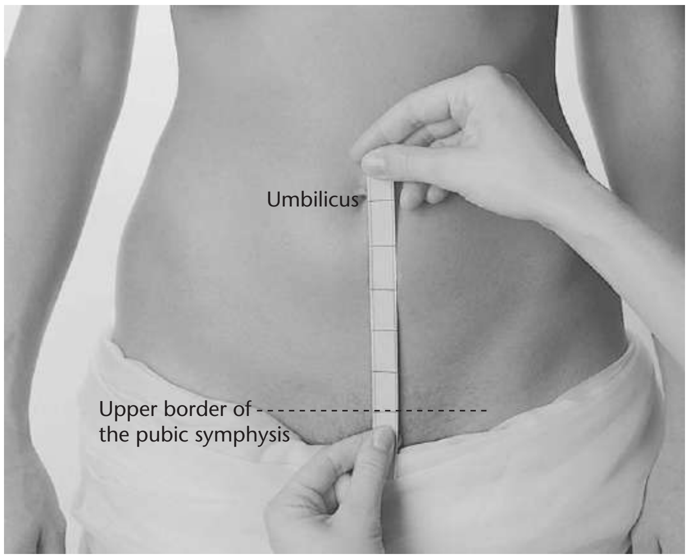
Umbilicus
Upper border of the pubic symphysis

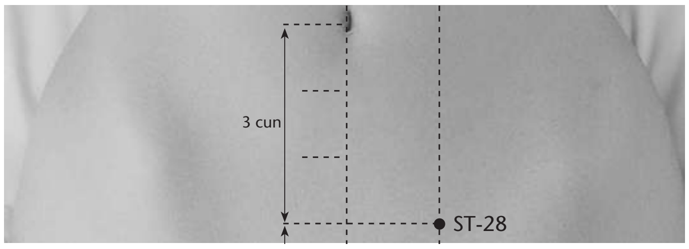

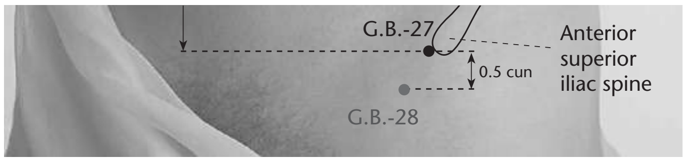
G.B.-27
0.5 cun
Anterior superior iliac spine
G.B.-28

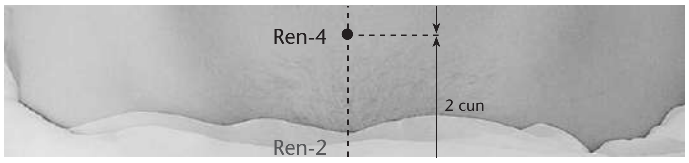
Ren-4
2 cun
Ren-2

End Diarrhoea ZHIXIE Ex-CA

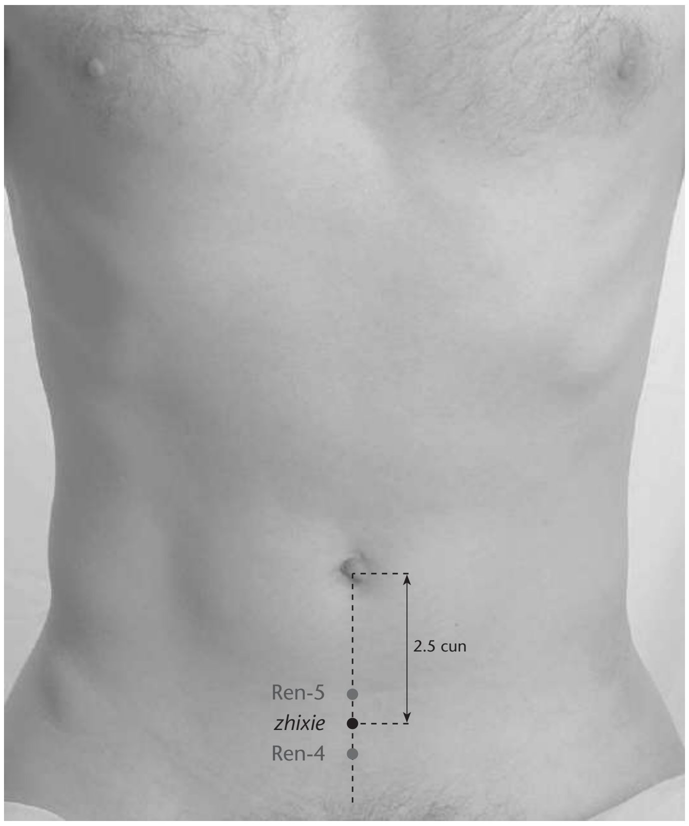

Location

On the anterior midline, 2.5 cun inferior to the umbilicus.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). *Zhixie* is located at the midpoint of this distance, between → **Ren-4** (2 cun superior to the upper border of the pubic symphysis) and → **Ren-5** (2 cun inferior to the umbilicus). **Or:** Spreading hands technique (→ 2.3.3): Place the little fingers on the umbilicus and the upper border of the pubic symphysis respectively and spread the hands evenly, joining the thumbs at the midpoint of this distance.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

- Stops diarrhoea

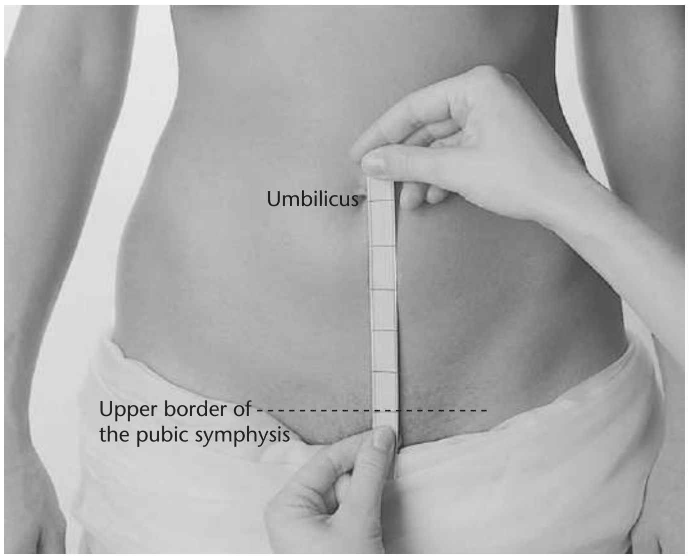

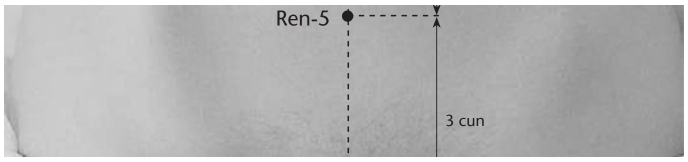

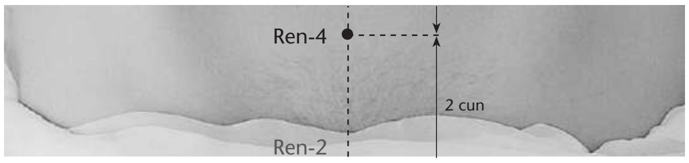

Ex-CA *Triangle Moxibustion* SANJIAOJIU

Location

These three points are located on the corners of an equilateral triangle, the apex of which is formed by the umbilicus, while the base forms a horizontal line on the abdomen. The sides are equal to the patient's smile.

How to find

It is helpful to explain the location of this point to the patient. Invariably, this will make them smile, allowing the practitioner to determine the length of the sides of the triangle.

Needling

Moxibustion only, no needling.

Actions/Indications

- Regulate Qi and stop diarrhoea.

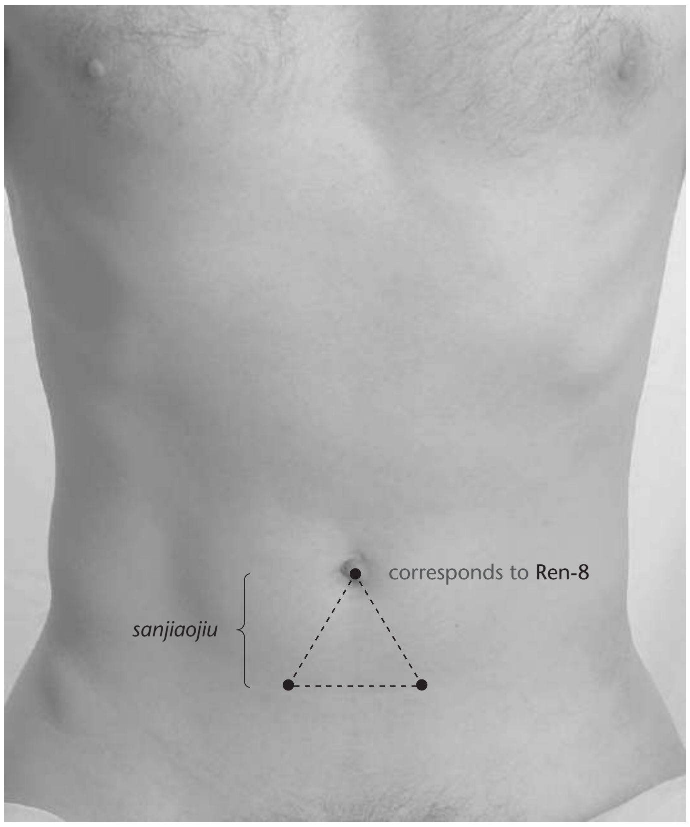

6.3 Extra Points: Back (EX-B)

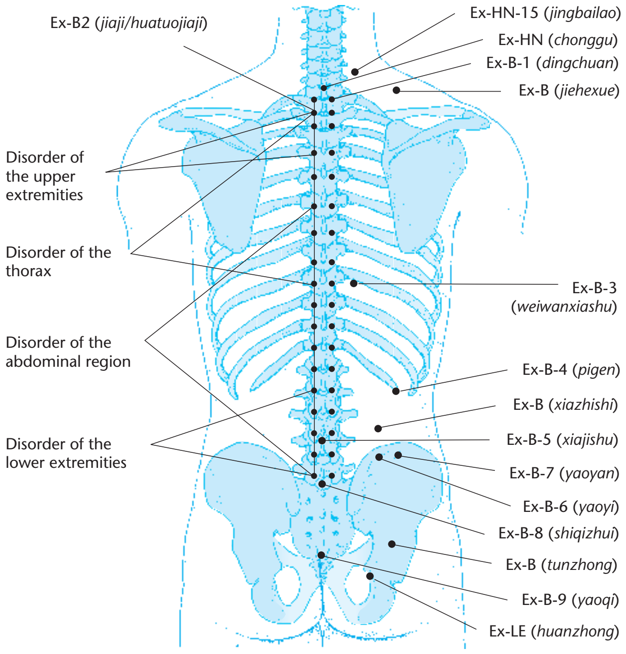
Ex-B2 (*jiaji/huatuojiaji*)
Disorder of the upper extremities
Disorder of the thorax
Disorder of the abdominal region
Disorder of the lower extremities
Ex-HN-15 (*jingbailao*)
Ex-HN (*chonggu*)
Ex-B-1 (*dingchuan*)
Ex-B (*jiehexue*)
Ex-B-3 (*weiwanxiashu*)
Ex-B-4 (*pigen*)
Ex-B (*xiazhishi*)
Ex-B-5 (*xiajishu*)
Ex-B-7 (*yaoyan*)
Ex-B-6 (*yaoyi*)
Ex-B-8 (*shiqizhui*)
Ex-B (*tunzhong*)
Ex-B-9 (*yaoqi*)
Ex-LE (*huanzhong*)

English abbreviation (standard)	***Pinyin* name**	**Nguyen Van Nghi, König and Wancura, Schnorrenberger**	**Shanghai College**	**Ex (Hempen)**
Ex-B-1	*dingchuan*	NP45	M-BW-1	Ex10
Ex-B-2	*huatuojiaji/jiaji*	PaM or ZP85	M-BW-35	Ex12
Ex-B-3	*weiwanxiashu/* *waiguanxiashu/* *bashu* or *yishu*	PaM or ZP62	M-BW-12	
Ex-B-4	*pigen*	PaM or ZP66	M-BW-16	
Ex-B-5	*xiajishu*	PaM or ZP71	M-BW-21	
Ex-B-6	*yaoyi*	PaM or ZP73	M-BW-23	
Ex-B-7	*yaoyan*	PaM or ZP74	M-BW-24	
Ex-B-8	*shiqizhui/shiqizhuixia*	PaM or ZP75	M-BW-25	Ex11
Ex-B-9	*yaoqi*	PaM or ZP79	M-BW-29	
Further extra points				
Ex-B	*jiehexue*	NP47	N-BW-6	
Ex-B	*tunzhong*	PaM or ZP83	M-BW-33	

Ch06a-F10028.qxd 2/23/08 1:26 PM Page 573

Calm Dyspnoea DINGCHUAN **Ex-B-1**

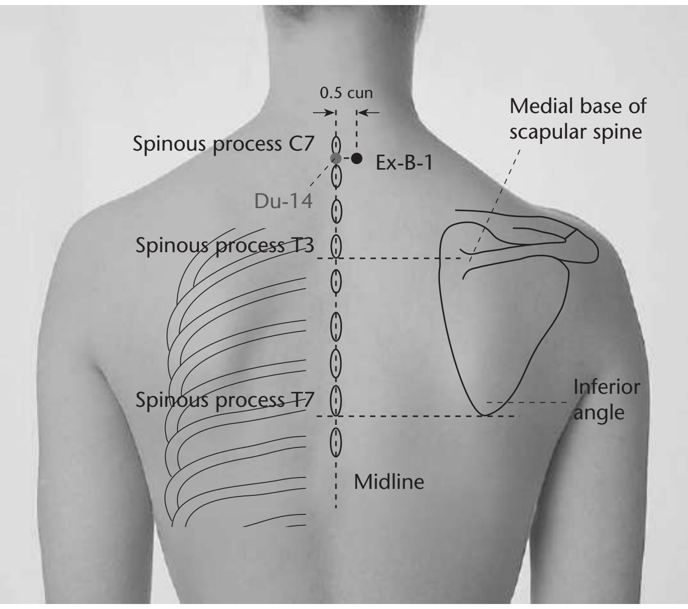
0.5 cun
Medial base of scapular spine
Spinous process C7
Ex-B-1
Du-14
Spinous process T3
Spinous process T7
Inferior angle
Midline

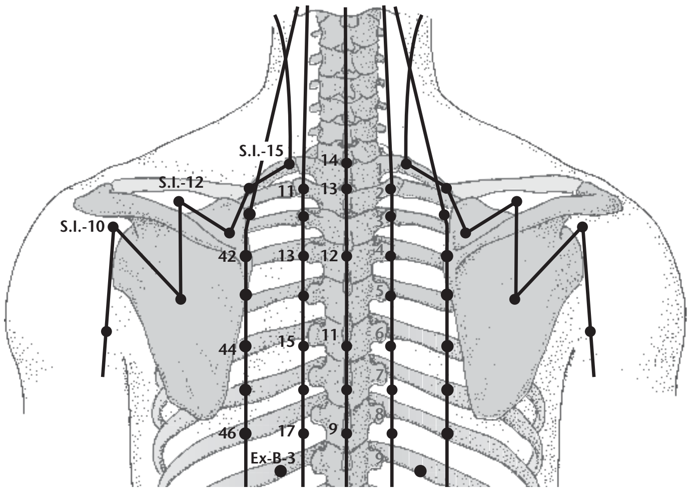
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3

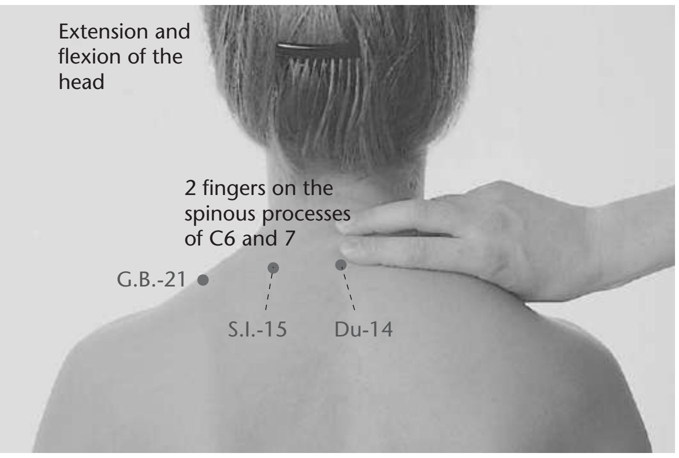
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

0.5 cun lateral to the lower border of the spinous process of the 7th cervical vertebra (C7).

How to find

First, identify C7 (→ 3.4.1). Next, measure 0.5 cun from the lower border of its spinous process in a lateral direction and there locate **Ex-B-1.**

Located on the same level are → **Du-14** (on the midline), → **S.I.-15** (2 cun lateral to the midline) and **Ex-B** (*jiehexue*, 3.5 cun lateral to the midline).

Needling

0.5–1 cun in a medial direction.

Actions/Indications

- Harmonises Lung Qi
- Local point for the neck and shoulders

Special features

Modern main point for dyspnoea

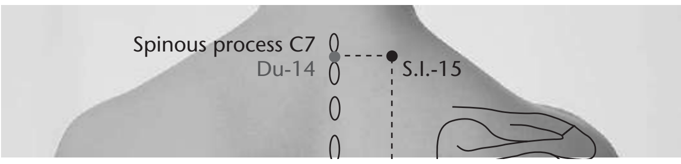
Spinous process C7
Du-14
S.I.-15

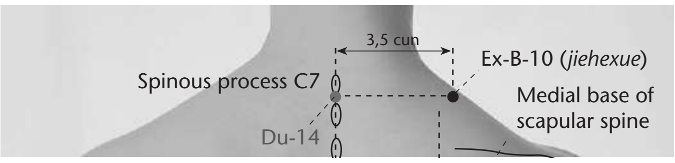
3,5 cun
Ex-B-10 (jiehexue)
Spinous process C7
Medial base of scapular spine
Du-14

Ch06a-F10028.qxd 2/23/08 1:26 PM Page 574

Ex-B-2 *Hua Tuo's Paravertebral Points* HUATUOJIAJI

Location

17 point pairs, 0.5 cun lateral to the lower borders of the spinous processes, close to the spinal facet joints:

- 12 thoracic point pairs (*xiongjiaji*): between T1 and T12
- 5 lumbar point pairs (*yaojiaji*): between L1 and L5, Depending on the school of thought, corresponding points lateral to the cervical spine are described as 'additional *huatuojiaji*'.

How to find

Select points of **Ex-B-2** depending on the disorder. Determine the relevant spinous process and locate the point pairs 0.5 cun lateral to its lower border.

Needling

0.5–1 cun vertically or better obliquely in a medial direction towards the spine, up to 1.5 cun in the lumbar region. The purpose of needling these points is to stimulate the local nerve roots. This means that the insertion angle has to be adapted to the patient's anatomy. Never needle in a lateral direction. Once *deqi* is obtained, do not further stimulate the needle; it is possible to apply electrostimulation.

Actions/Indications

- Regulate and harmonise the five *zang* and six *fu* Organs, depending on the points selected

Special features

With disorders of the facet joints of the cervical spine, there are often tender points 0.5 cun lateral to the lower borders of the spinous processes. In China, these are needled for chronic disorders of the cervical spine. If oblique insertions (approximately 45°) in a medial direction are used for points on the inner branch of the Bladder channel, the tip of the needle will reach the area of the *huatuojiaji* points. This enhances the therapeutic effect.

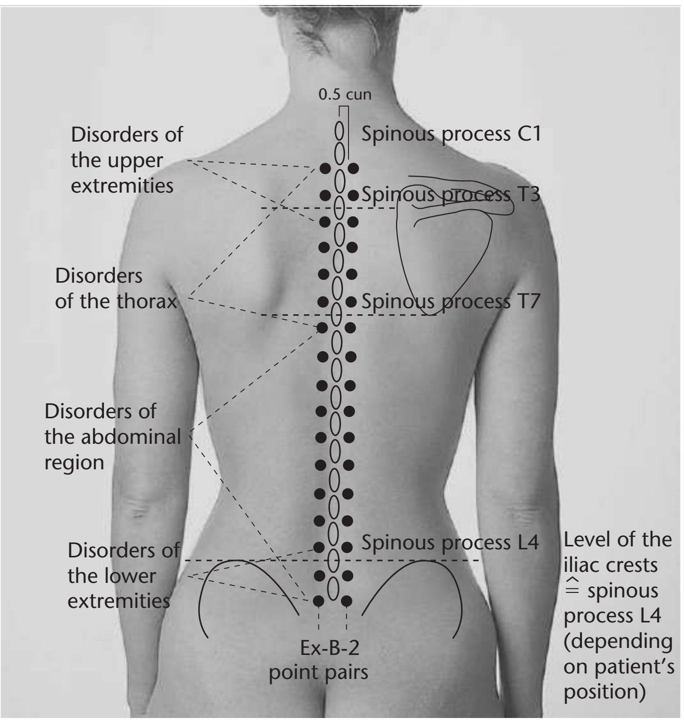
0.5 cun
Disorders of the upper extremities
Spinous process C1
Spinous process T3
Disorders of the thorax
Spinous process T7
Disorders of the abdominal region
Disorders of the lower extremities
Spinous process L4
Level of the iliac crests ≙ spinous process L4 (depending on patient's position)
Ex-B-2 point pairs

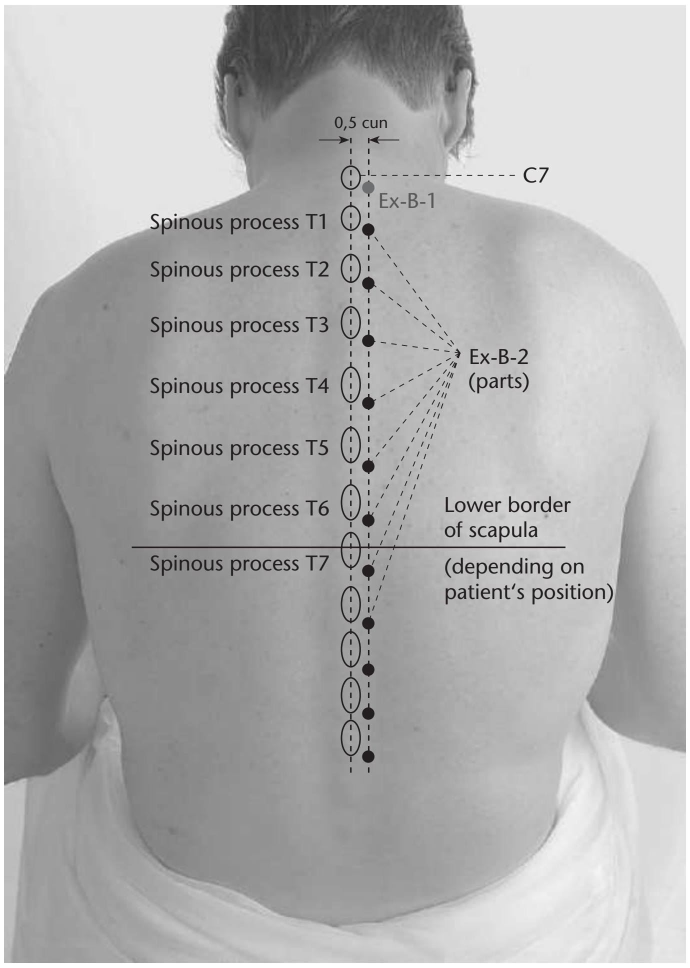
0,5 cun
C7
Ex-B-1
Spinous process T1
Spinous process T2
Spinous process T3
Spinous process T4
Spinous process T5
Spinous process T6
Spinous process T7
Ex-B-2 (parts)
Lower border of scapula
(depending on patient's position)

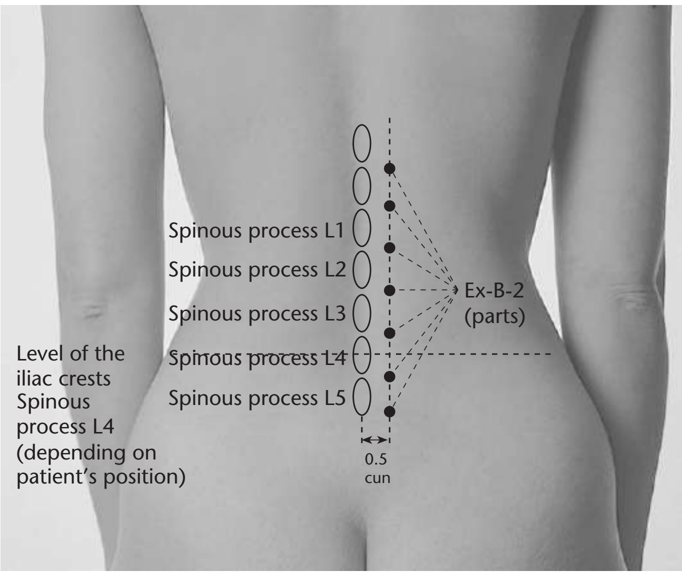
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
Spinous process L5
Ex-B-2 (parts)
Level of the iliac crests Spinous process L4 (depending on patient's position)
0.5 cun

Stomach Controller Lower Shu WEIWANXIASHU Ex-B-3

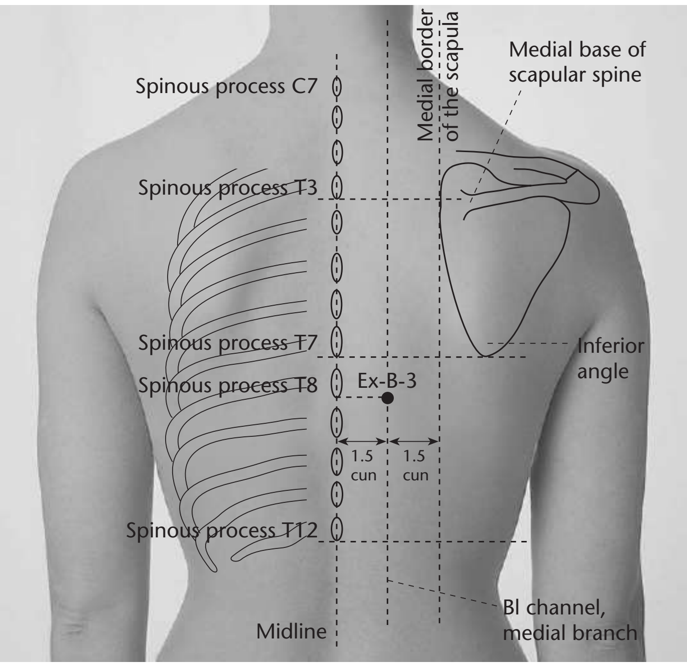
Medial border of the scapula
Medial base of scapular spine
Spinous process C7
Spinous process T3
Spinous process T7
Inferior angle
Spinous process T8
Ex-B-3
1.5 cun
1.5 cun
Spinous process T12
Bl channel, medial branch
Midline

Location

1.5 cun lateral to the lower border of the spinous process of the 8th thoracic vertebra (T8).

How to find

Locate T7 (→ 3.4.2). From there, count down one spinous process (to T8) and locate **Ex-B-3** 1.5 cun lateral from the midline. Use the Tuffier's line (L4, 3.4.3) to confirm the location of this point: from L4, count up the spinous processes of L3 to L1 and of T12 to T8.

Located on the same level is a point of → **Ex-B-2** (0.5 cun lateral to the midline).

Needling

0.5–1 cun vertically or up to 1.5 cun obliquely in a medial direction. Do not needle in a lateral direction. Caution: Pneumothorax.

Actions/Indications

- Opens the channel locally
- Moistens the body and clears Heat

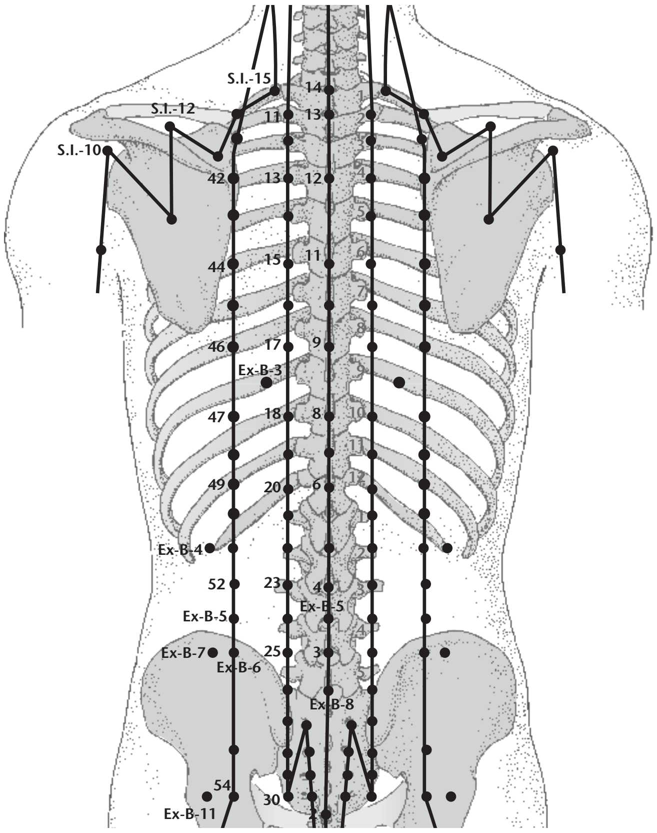
S.I.-15
S.I.-12
S.I.-10
14
13
11
13
42
13
12
44
15
11
46
17
9
Ex-B-3
47
18
8
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11

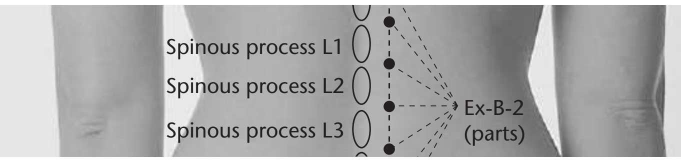
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Ex-B-4 *Fullness Root* PIGEN

Location

3.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 1st lumbar vertebra (L1).

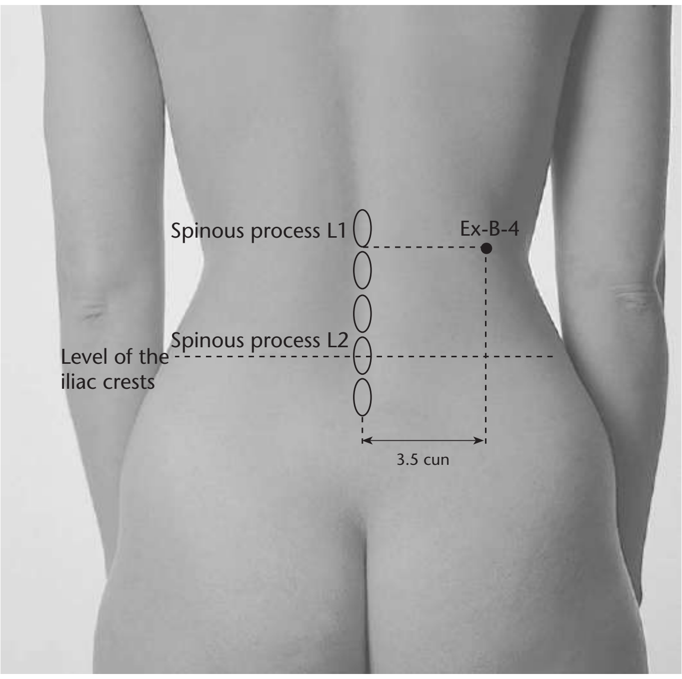
Spinous process L1
Ex-B-4
Spinous process L2
Level of the iliac crests
3.5 cun

How to find

Locate the T7 (→ 3.4.2) and count down 6 spinous processes to L1. Locate **Ex-B-4** on the level of the lower border of its spinous process, 3.5 cun lateral to the midline. Use the Tuffier's line (L4, → 3.4.3) to confirm the location of this point: from L4, count up the spinous processes of L3 to L1.

Located on the same level are → **Du-5** (on the midline), a point of → **Ex-B-2/BL-22/BL-51** (0.5/1.5/3 cun lateral to the midline) as well as → **G.B.-25** (on the lower border of the free end of the 12th rib).

Needling

0.8–1 cun obliquely in a medial direction. Caution: Injury to the kidneys!

Actions/Indications

- Improves the Qi flow between the Upper and Middle Burner
- Eliminates local Qi and Blood stagnation

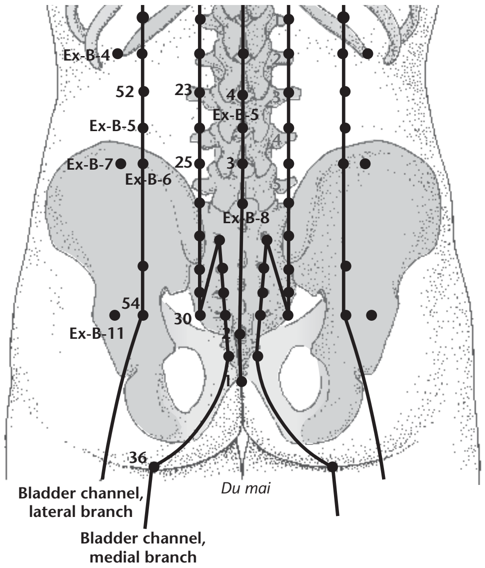
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
1
2
3
4
5
54
30
Ex-B-11
2
1
36
Bladder channel, lateral branch
Du mai
Bladder channel, medial branch

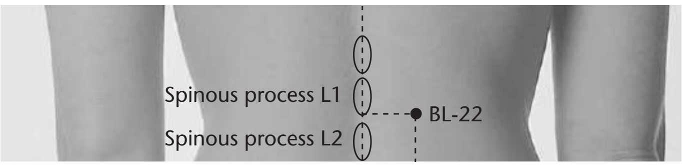
Spinous process L1
BL-22
Spinous process L2

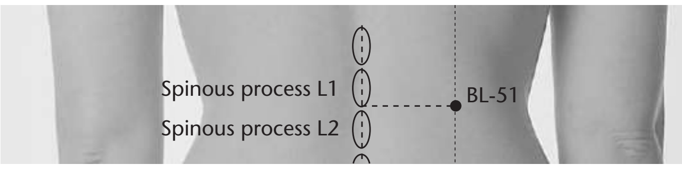
Spinous process L1
BL-51
Spinous process L2

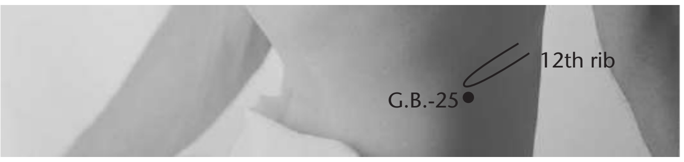
12th rib
G.B.-25

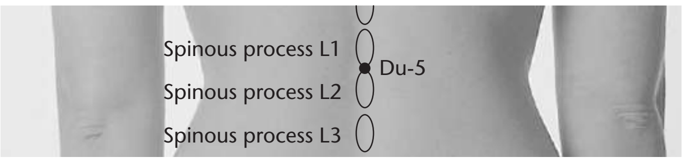
Spinous process L1
Du-5
Spinous process L2
Spinous process L3

Lower Shu Point XIAJISHU **Ex-B-5**

Location

Some authors consider **Ex-B-5** (*xiajishu*) to comprise three points: one point on the midline below the spinous process of the 3rd lumbar vertebra (L3) (this is the more common and single location of this extra point), complemented by two lateral points 3 cun lateral to the centre point on the midline.

How to find

Locate the Tuffier's line and L4 (→ 3.4.3). Locate **Ex-B-5** on the midline, on the lower border of the spinous process of L3. Then measure 3 cun in a lateral direction and there locate the two lateral points of **Ex-B-5.**

Located on the same level are a point of → **Ex-B-2/BL-24** (0.5/1.5 cun lateral to the midline).

Needling

0.5–0.8 cun vertically to the skin or obliquely in an inferior direction. The patient's back should be straight or overextended to avoid a spinal puncture. Oblique insertions in a superior direction should only be carried out by experienced practitioners, as in small persons (regardless of their weight), the spinal canal can be reached after only 1.25 cun. Caution: In pregnant women about to go into labour, this point might have a labour-promoting effect.

Actions/Indications

- Tonifies Kidney Yang
- Opens the channel locally

Special features

Strictly speaking, this extra point should be considered as a single point and a pair of points with differing indications: depending on the school of thought and the clinical requirements, the lateral points are used for localised Qi and Blood stagnation, whereas the central point is indicated for Kidney Yang deficiency.

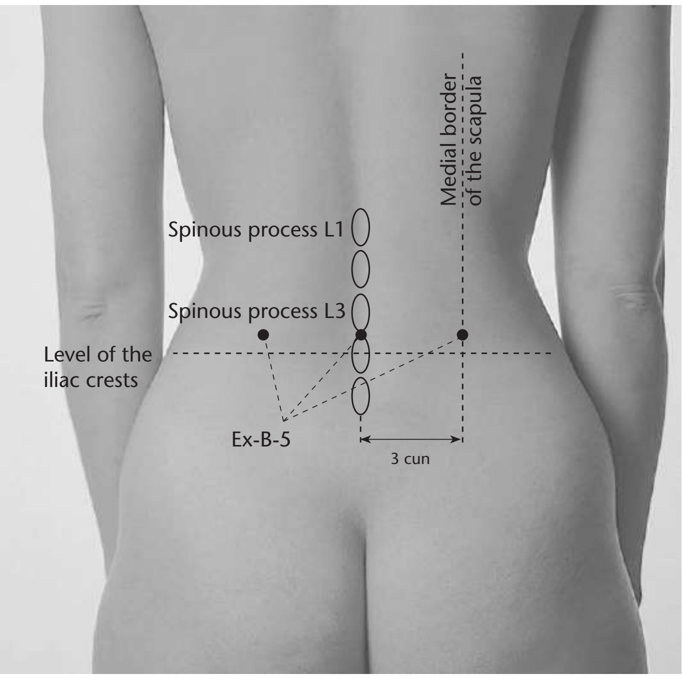
Medial border of the scapula
Spinous process L1
Spinous process L3
Level of the iliac crests
Ex-B-5
3 cun

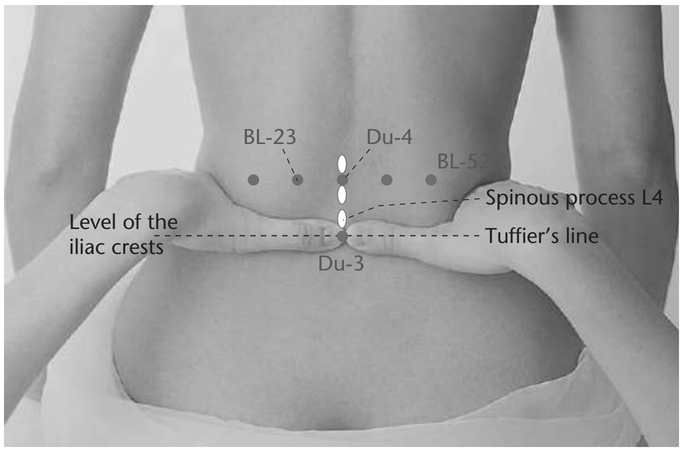
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

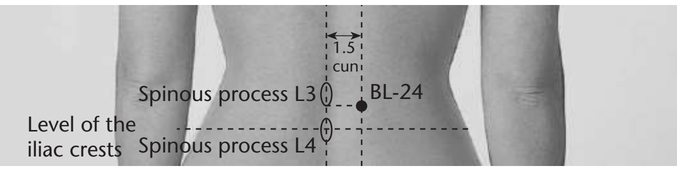
1.5 cun
Spinous process L3
BL-24
Level of the iliac crests
Spinous process L4

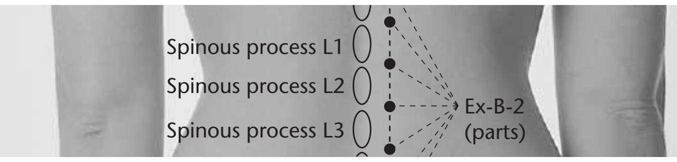
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Ex-B-6 Back Pain Point YAOYI

Location

3 cun lateral to the midline, on the level of the lower border of the spinous process of the 4th lumbar vertebra (L4).

How to find

Locate the Tuffier's line and L4 (→ 3.4.3). On the level of the lower border of the spinous process of L4, measure 3 cun in a lateral direction and there locate **Ex-B-6.** Located on the same level are → **Du-3** (on the midline), a point of → **Ex-B-2/BL-25/ Ex-B-7** (0.5/1.5/3.5 cun lateral to the midline).

Needling

Vertically 0.5–0.8 cun

Actions/Indications

- Opens the channel and *luo* vessels

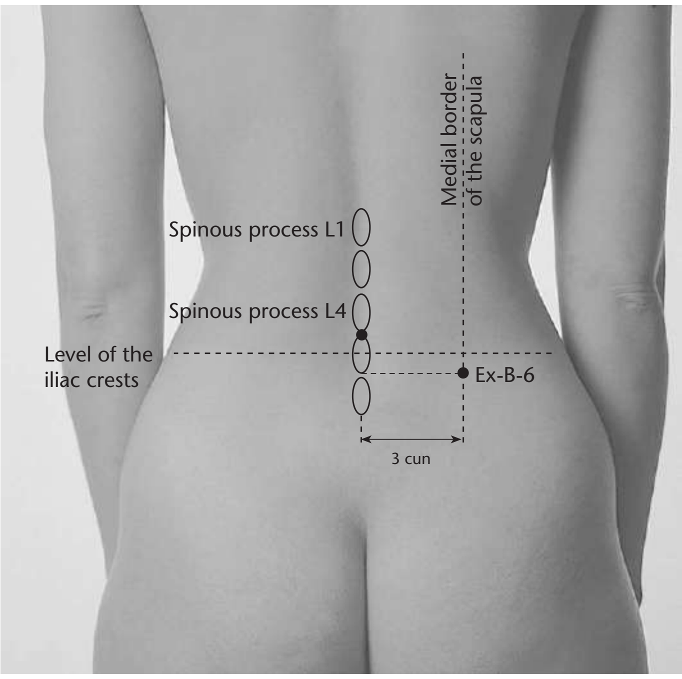
Medial border of the scapula
Spinous process L1
Spinous process L4
Level of the iliac crests
Ex-B-6
3 cun

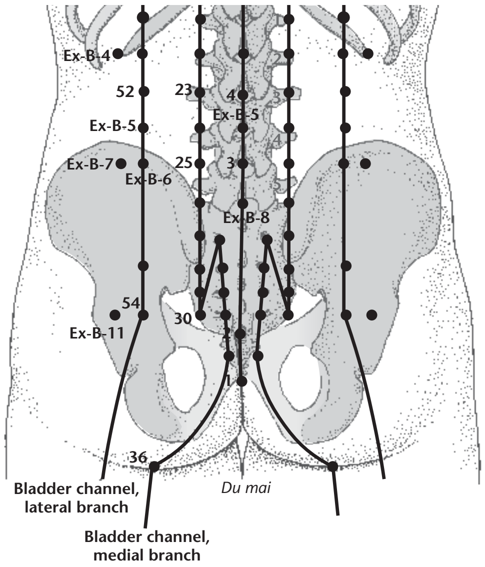
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
Ex-B-11
30
2
1
36
Bladder channel, lateral branch
Du mai
Bladder channel, medial branch
1
2
3
4
5

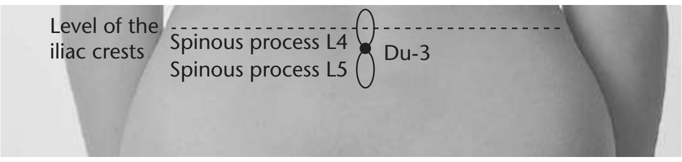
Level of the iliac crests
Spinous process L4
Spinous process L5
Du-3

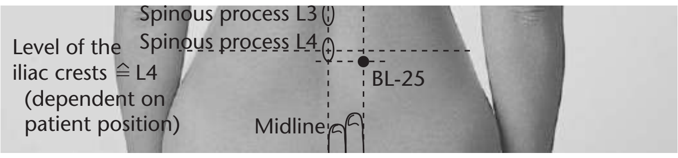
Spinous process L3
Spinous process L4
Level of the iliac crests ≙ L4 (dependent on patient position)
BL-25
Midline

Lumbar Eyes YAOYAN Ex-B-7

Location

3.5 cun lateral to the midline, on the level of the lower border of the spinous process of the 4th lumbar vertebra (L4).

How to find

Locate the Tuffier's line and L4 (→ 3.4.3). **Ex-B-7** is located 3.5 cun lateral to it. Located on the same level are → **Du-3** (on the midline), a point of → **Ex-B-2/BL-25/Ex-B-6** (0.5/1.5/3 cun lateral to the midline).

Needling

Vertically 0.5–0.8 cun

Actions/Indications

- Opens the channel locally
- Tonifies the Kidneys

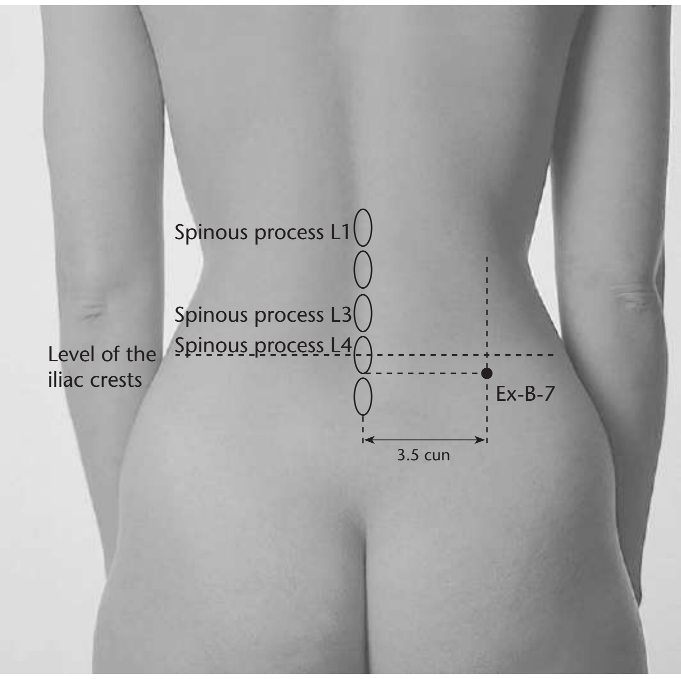
Spinous process L1
Spinous process L3
Spinous process L4
Level of the
iliac crests
Ex-B-7
3.5 cun

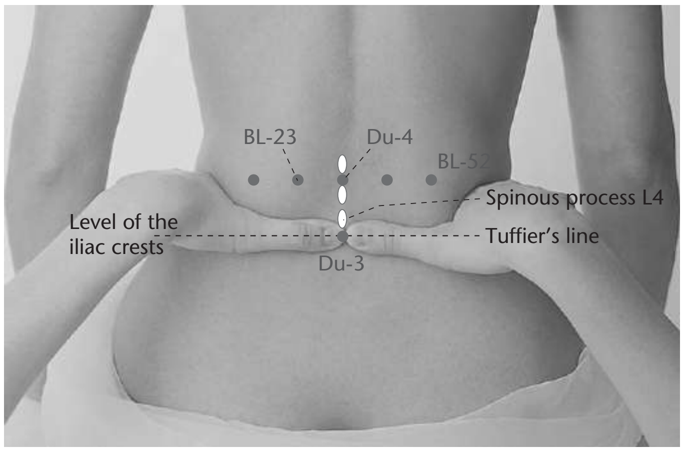
BL-23
Du-4
BL-52
Spinous process L4
Level of the
iliac crests
Tuffier's line
Du-3

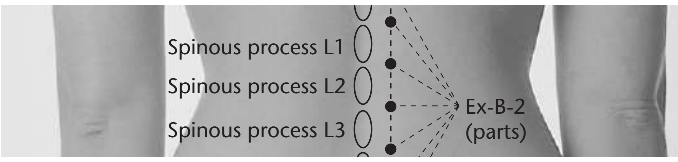
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

Ex-B-8 *Below the 17th Vertebra* SHIQIZHUI/SHIQIZHUIXIA

Location

On the midline, below the spinous process of the 5th lumbar vertebra (L5).

How to find

Locate the Tuffier's line and L4 (→ 3.4.3). Next, palpate in an inferior direction to the next vertebra (L5) and locate **Ex-B-8** below the border of its spinous process, at the lumbosacral joint.
Located on the same level are a point of → **Ex-B-2/BL-26** (0.5/1.5 cun lateral to the midline).

Needling

0.5–1 cun vertically to the skin or obliquely in an inferior direction. The patient's back should be straight or slightly flexed to avoid a spinal puncture. Oblique insertions in a superior direction should only be carried out by experienced practitioners, as in small persons (regardless of their weight), the spinal canal can be reached after only 1.25 cun. Caution: In pregnant women about to go into labour, this point might have a labour-promoting effect.

Actions/Indications

- Tonifies the Kidneys
- Opens the channel locally

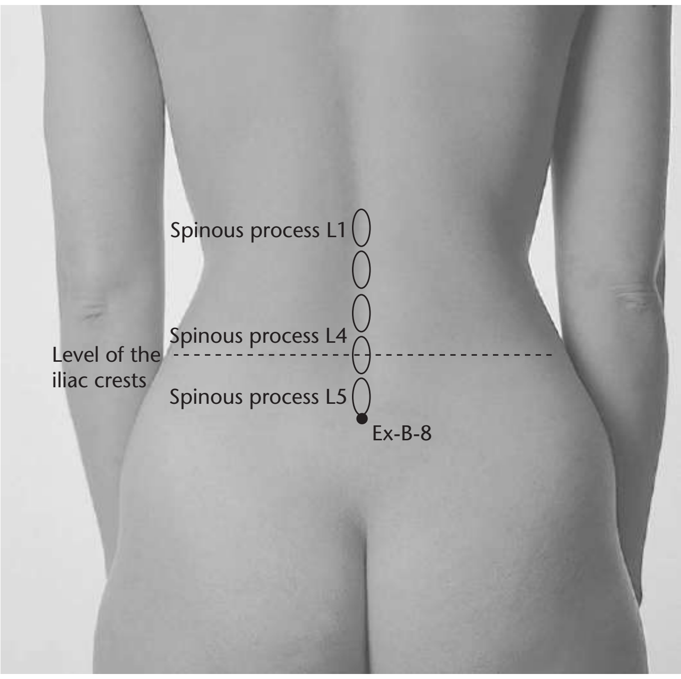
Spinous process L1
Spinous process L4
Level of the iliac crests
Spinous process L5
Ex-B-8

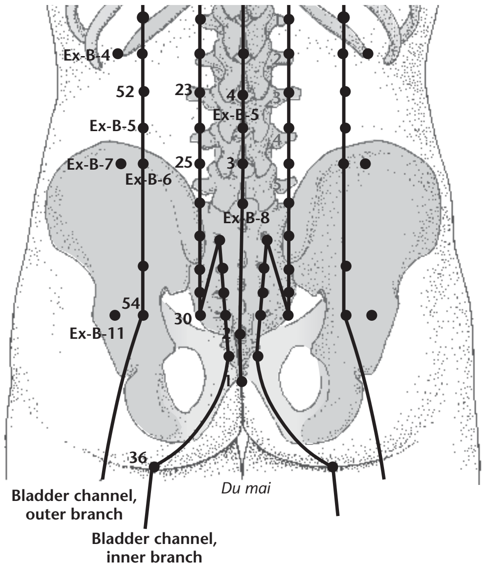
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
36
Du mai
Bladder channel, outer branch
Bladder channel, inner branch

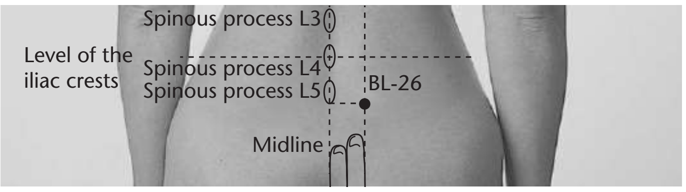
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
BL-26
Midline

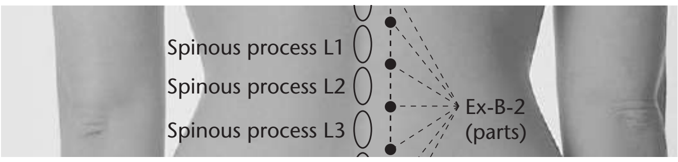
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

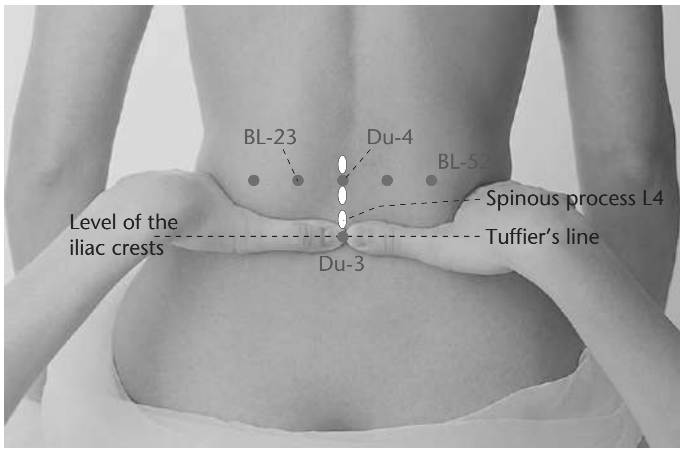
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

Miraculous Lumbar Point YAOQI Ex-B-9

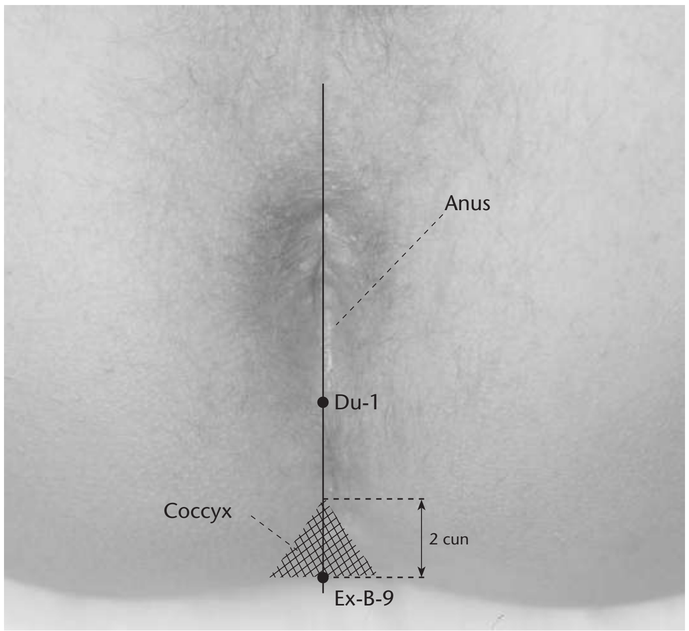

Location

2 cun superior to the tip (inferior end) of the coccyx.

How to find

Locate the tip of the coccyx superior to the anus and palpate for the sacral hiatus. From there, palpate 2 cun in a superior direction and locate **Ex-B-9** in a depression between the sacral processes.
→ **Du-1** is located on the midline, between the tip of the coccyx and the anus.

Needling

Up to 1.5 cun obliquely in a superior direction.

Actions/Indications

- Calms the *shen*

Ex-B-10 *Tuberculosis Point* JIEHEXUE

Location

3.5 cun lateral to the lower border of the spinous process of the 7th cervical vertebra (C7).

How to find

Locate the spinous process of C7 and, on this level, palpate 3.5 cun in a lateral direction. There, locate **Ex-B** (*jiehexue*).
Located on the same level are → **Du-14** (on the midline), → **Ex-B-2/S.I.-15** (0.5/2 cun lateral to the midline).

Needling

Vertically 0.5–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Tonifies the Lung
- Opens the channel locally

Special features

This is an unofficial extra point.

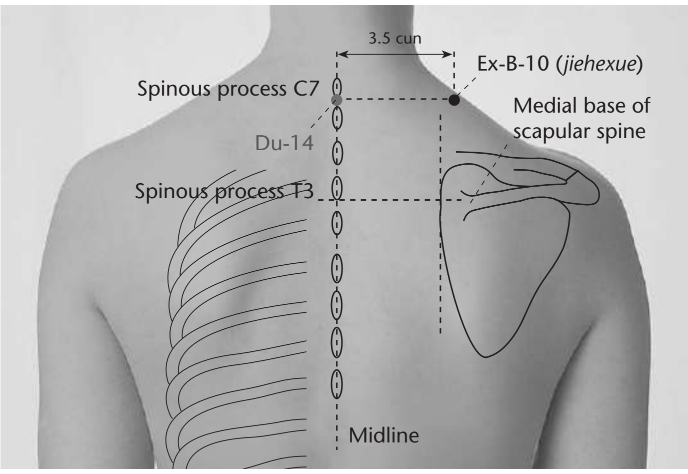
3.5 cun
Ex-B-10 (*jiehexue*)
Spinous process C7
Medial base of scapular spine
Du-14
Spinous process T3
Midline

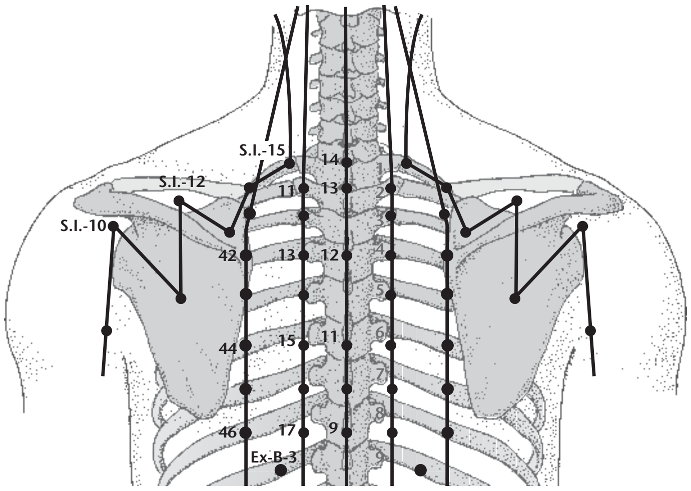
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
6
46
17
9
Ex-B-3

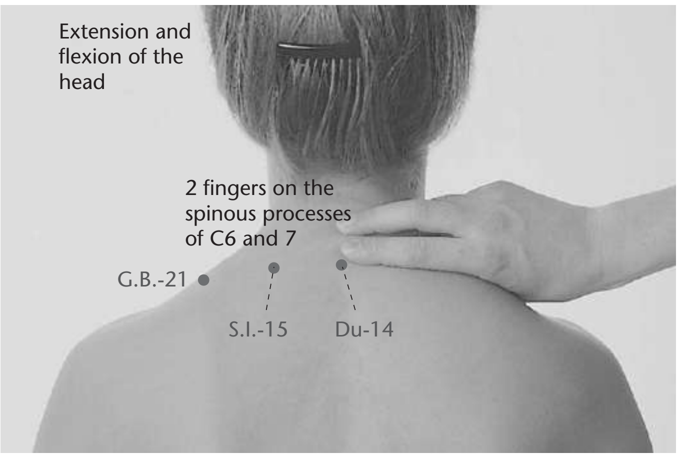
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Spinous process C7
Du-14
Medial base of scapular spine

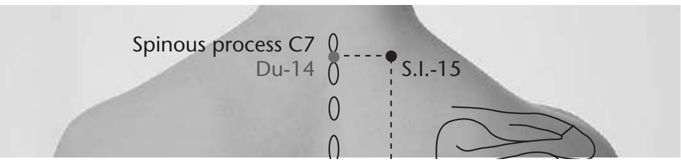
Spinous process C7
Du-14
S.I.-15

Buttock Centre TUNZHONG Ex-B

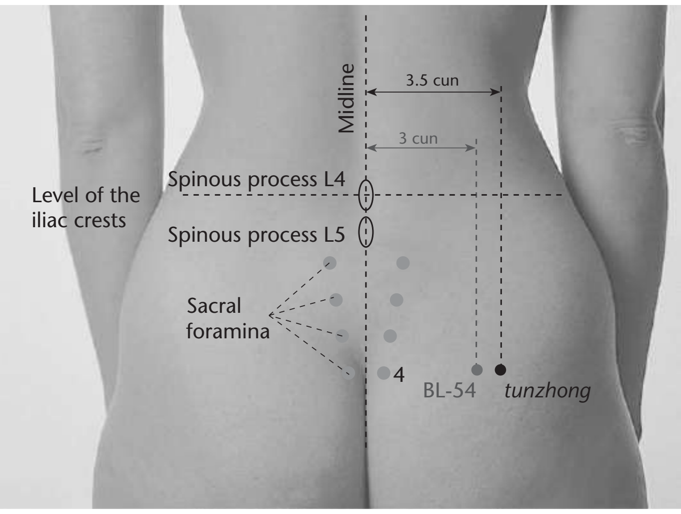
Midline
3.5 cun
3 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
4
BL-54
tunzhong

Location

At the centre of the buttock, 3.5 cun lateral to the posterior midline, on the level of the 4th sacral foramen.

How to find

Locate the lumbosacral joint, which often forms an easily palpable 'kink' between the L5 and the sacrum. Alternatively, locate the Tuffier's line (→ 3.4.3) for reference. Then palpate the inferior aspect of the sacrum for the sacral hiatus (→ 3.4.4). Between these two reference points, the four pairs of sacral foramina are arranged on either side of the midline in a slight V-shaped form and can be palpated as shallow depressions. **Ex-B** (*tunzhong*) is located 3.5 cun lateral to the lowest depression (=4th foramen), approximately at the centre of the buttock.

Located on the same level are → **BL-34/BL-30/BL-54** (over the 4th foramen/1.5/3 cun lateral to the midline).

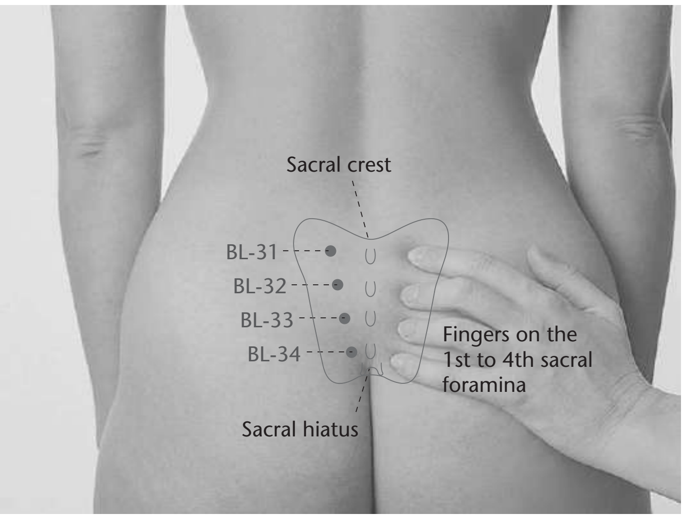
Sacral crest
BL-31
BL-32
BL-33
BL-34
Fingers on the 1st to 4th sacral foramina
Sacral hiatus

Needling

Vertically 2–3 cun

Actions/Indications

- Opens the channel and *luo* vessels locally

Special features

This is an unofficial extra point.

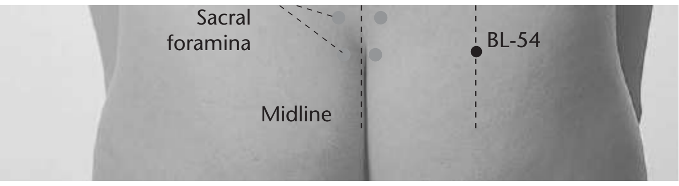
Sacral foramina
BL-54
Midline

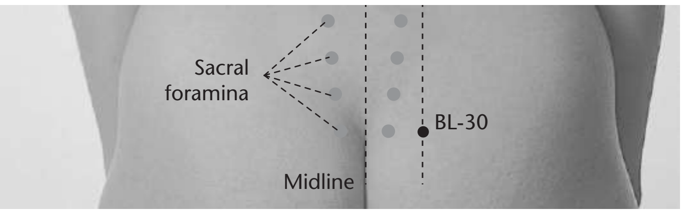
Sacral foramina
BL-30
Midline

Ch06b-F10028.qxd 2/22/08 10:04 PM Page 585

6.4 Extra Points: Upper Extremities (EX-UE)

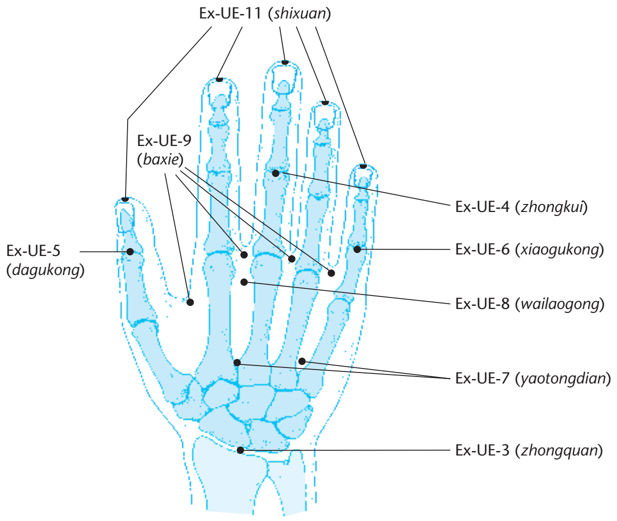
Ex-UE-11 (*shixuan*)
Ex-UE-9 (*baxie*)
Ex-UE-5 (*dagukong*)
Ex-UE-4 (*zhongkui*)
Ex-UE-6 (*xiaogukong*)
Ex-UE-8 (*wailaogong*)
Ex-UE-7 (*yaotongdian*)
Ex-UE-3 (*zhongquan*)

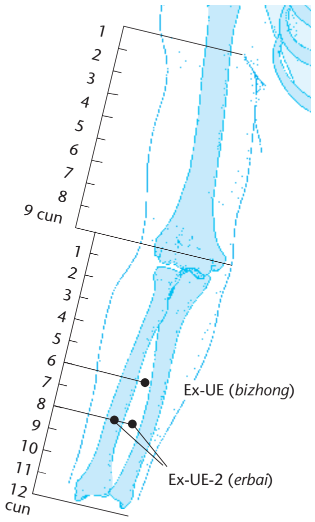
1
2
3
4
5
6
7
8
9 cun
1
2
3
4
5
6
7
8
9
10
11
12 cun
Ex-UE (*bizhong*)
Ex-UE-2 (*erbai*)

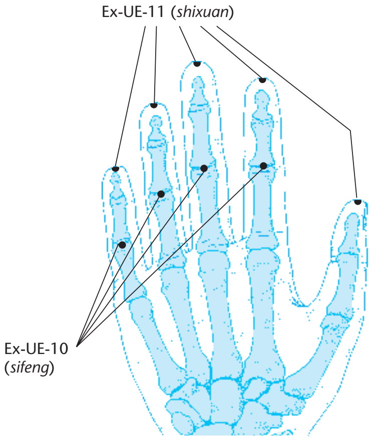
Ex-UE-11 (*shixuan*)
Ex-UE-10 (*sifeng*)

English abbreviation (standard)	***Pinyin* name**	**Nguyen Van Nghi, König and Wancura, Schnorrenberger**	**Shanghai College**	**Ex (Hempen)**
Ex-UE-1	*zhoujian*		M-UE-46	
Ex-UE-2	*erbai*	PaM or ZP114	M-UE-29	
Ex-UE-3	*zhongquan*	PaM or ZP118	M-UE-33	
Ex-UE-4	*zhongkui*	PaM101	M-UE-16	
Ex-UE-5	*dagukong*	PaM or ZP100	M-UE-15	
Ex-UE-6	*xiaogukong*	PaM or ZP102	M-UE-17	
Ex-UE-7	*yaotongxue/yaotongdian*	Handpoint 2 PaM or ZP110/111[*]	M-UE	Ex18
Ex-UE-8	*wailaogong*[†]*/luozhen*	PaM or ZP108[†]	M-UE-24	Ex17
Ex-UE-9	*baxie*	PaM or ZP107	M-UE-22	Ex14
Ex-UE-10	*sifeng*		M-UE-9	Ex13
Ex-UE-11	*shixuan*	PaM or ZP86	M-UE-1	Ex15
Further Extra points				
Ex-UE	*jianqian/jianneiling*	Part of NP74[‡] (*sanjian*): *jianyu* (L.I.-15), *jianqian*, *jianhou*	M-UE-48[§]	Ex16
Ex-UE	*bizhong*	PaM or ZP115	M-UE-30	

---

[*] The 'standard' point Ex-UE-7 (*yaotongdian* or *yaotongxue*) is described by Nguyen Van Nghi and Schnorrenberger as two separate points with different *pinyin* names: PaM or ZP110 (*weiling*) and 111 (*jingling*). Their location also corresponds to handpoint 2.

[†] Nguyen Van Nghi and Schnorrenberger distinguish between *luozhen* (PaM or ZP108) and *wailaogong* (PaM or ZP109). They describe the location of *luozhen* as identical with the 'standard' location of Ex-UE-8 (*wailaogong*): on the dorsum of the hand, between the 2nd and 3rd metacarpal bones, approximately 0.5 cun proximal to the metacarpophalangeal joint. 'Their' *wailaogong* is located at the midpoint of a line connecting the wrist and the head of the 3rd metacarpal bone, between the metacarpals, directly opposite to the point *laogong* (P-8) – hence the name *wailaogong* ('outer' *laogong*).

[‡] **NP74** (*jiansanzhen*/Three pinpricks at the Shoulder) according to Schnorrenberger: a set of three points: the first is identical with **L.I.-15** (*jianyu*), the second (*jianquan*/ Anterior to the Shoulder) is located 1 cun superior to the end of the anterior axillary fold, and the third (*jianhou*/Posterior to the Shoulder) is located 1 cun superior to the end of the posterior axillary fold.

[§] According to Deadman et al. (1998), **M-UE-48** (*jianqian*) is located midway between the end of the anterior axillary fold and **L.I.-15** (this point is not mentioned by the Shanghai College). See the point page for the location of this point.

Elbow Tip ZHOUJIAN **Ex-UE-1**

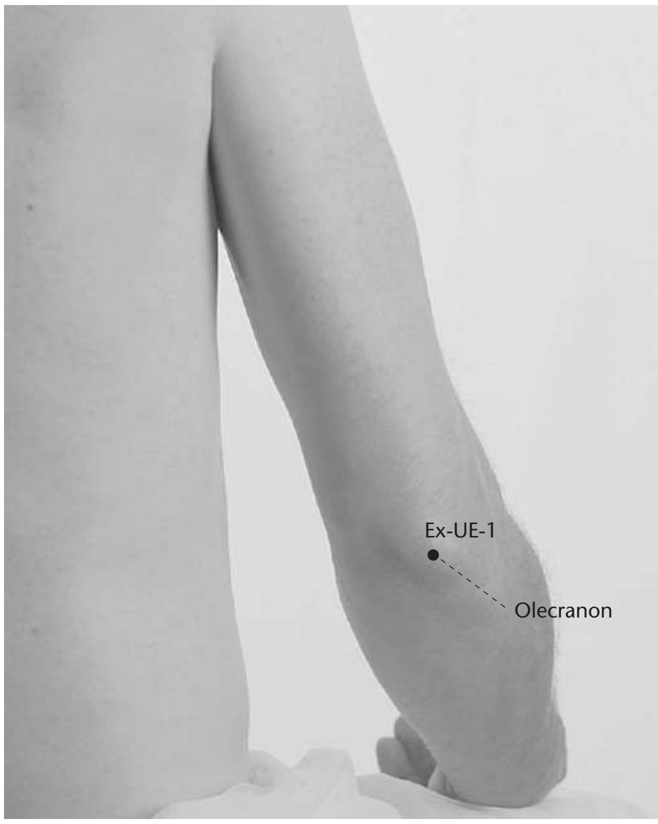
Ex-UE-1
Olecranon

Location

On the tip of the olecranon.

How to find

On the tip of the olecranon. This point is best located with the elbow flexed.

Needling

Moxibustion only!

Actions/Indications

- Opens the channel and *luo* vessels locally
- Disperses Phlegm locally

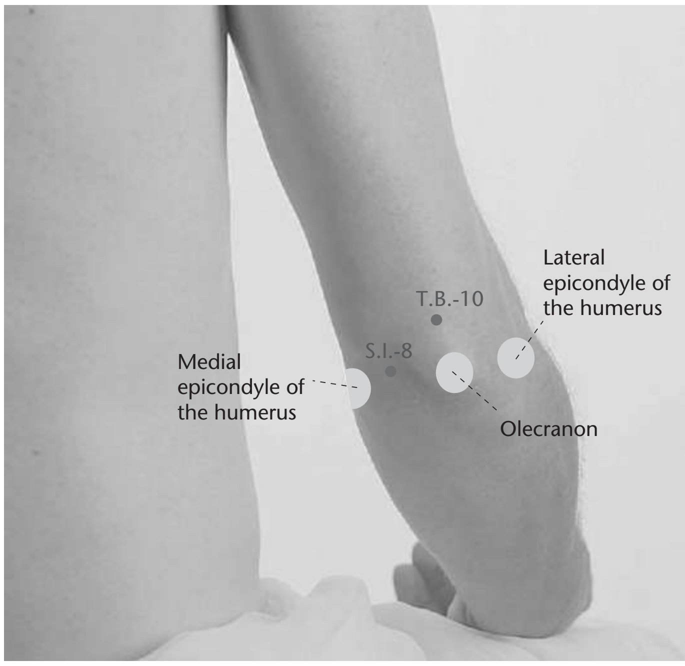
Lateral epicondyle of the humerus
T.B.-10
S.I.-8
Medial epicondyle of the humerus
Olecranon

Ex-UE-2 *Two Whites* ERBAI

Location

A pair of points on the palmar aspect of the forearm, 4 cun proximal to the wrist joint space (most distal wrist crease), on either side of the tendon of the flexor carpi radialis muscle.

How to find

Place the patient's arm in a relaxed supinated position. As the location of the wrist crease can vary, the wrist joint space provides a more reliable point of reference (→ 3.3.3). From the centre of the wrist joint space, measure 4 cun in a proximal direction and, on this level, locate one of the points lateral to and the other point medial to the tendon of the flexor carpi radialis muscle. **Or:** Divide the distance between the cubital crease (at → **P-3**) and the wrist joint space (at → **P-7**) into thirds, then locate → **Ex-UE-2** one third of the distance from the wrist joint (helpful tool: elastic tape → 2).

Needling

Up to 1 cun vertically or up to 1.5 cun obliquely towards proximal.

Actions/Indications

- Raises Qi to treat prolapse of the rectum and haemorrhoids

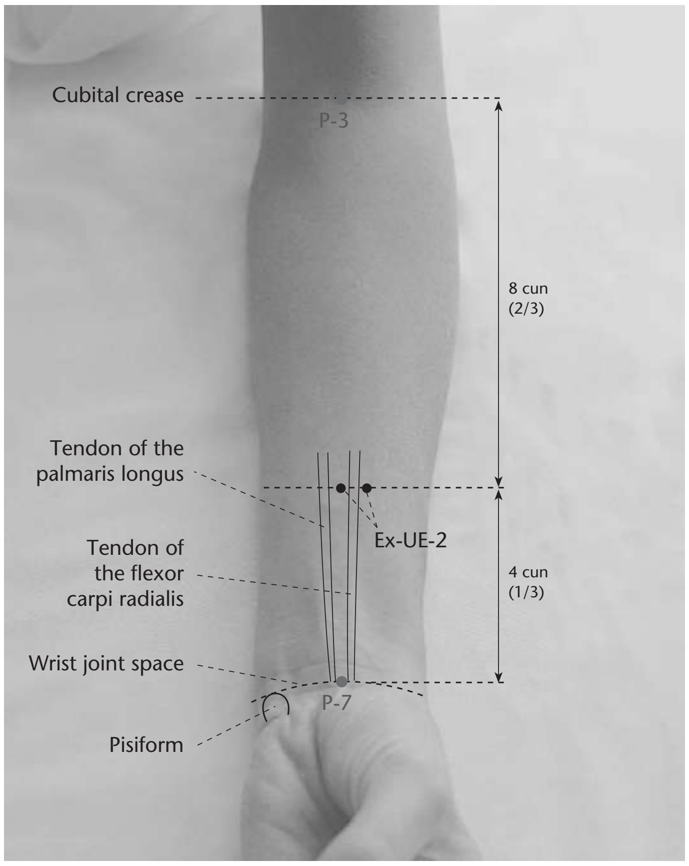

Cubital crease
P-3
8 cun
(2/3)
Tendon of the
palmaris longus
Ex-UE-2
Tendon of
the flexor
carpi radialis
4 cun
(1/3)
Wrist joint space
P-7
Pisiform

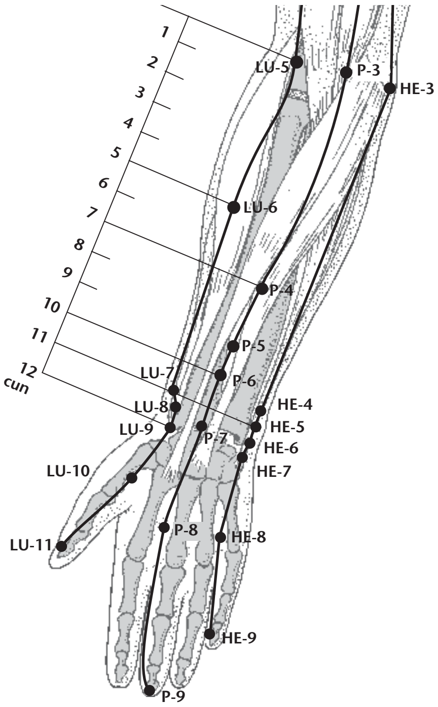

1
2
3
4
5
6
7
8
9
10
11
12
cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
LU-8
HE-4
LU-9
HE-5
P-7
HE-6
LU-10
HE-7
P-8
HE-8
LU-11
HE-9
P-9

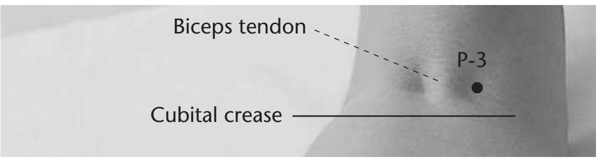

Biceps tendon
P-3
Cubital crease

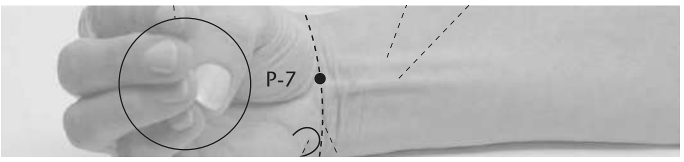

Posterior Spring ZHONGQUAN **Ex-UE-3**

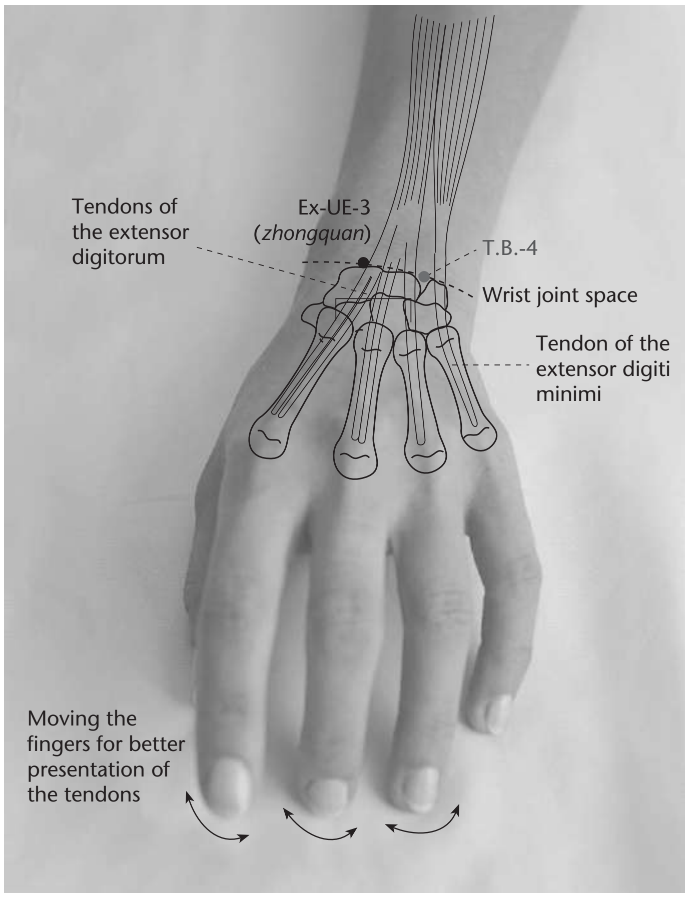
Tendons of
the extensor
digitorum
Ex-UE-3
(*zhongquan*)
T.B.-4
Wrist joint space
Tendon of the
extensor digiti
minimi
Moving the
fingers for better
presentation of
the tendons

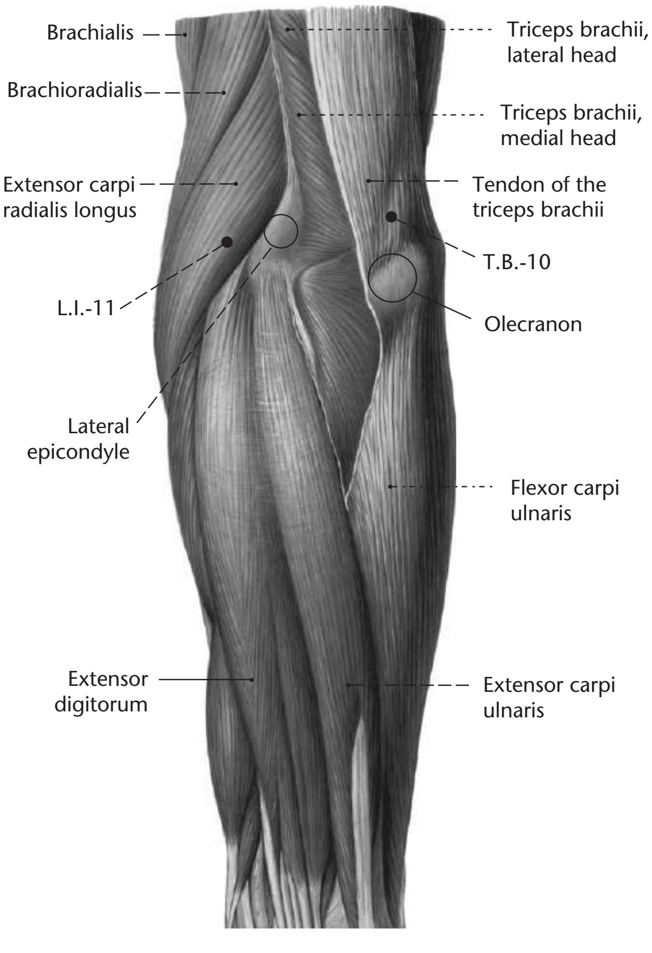
Brachialis
Brachioradialis
Extensor carpi
radialis longus
L.I.-11
Lateral
epicondyle
Extensor
digitorum
Triceps brachii,
lateral head
Triceps brachii,
medial head
Tendon of the
triceps brachii
T.B.-10
Olecranon
Flexor carpi
ulnaris
Extensor carpi
ulnaris

Location

On the dorsal aspect of the wrist joint space (dorsal wrist crease), radial to the tendon of the extensor digitorum communis muscle.

How to find

Place the patient’s arm in a relaxed pronated position. As the location of the dorsal wrist crease can vary, the joint space between the proximal row of carpals and the radius/ulna provides a more reliable point of reference (→ 3.3.3). By moving the wrist in a relaxed way, you can easily palpate the joint space. Locate **Ex-UE-3** in the joint space, radial to the tendon of the extensor digitorum communis muscle. The tendons become more pronounced by moving and overextending the three medial fingers. → **T.B.-4** is located on the same level, ulnar to the tendon of the extensor digitorum communis muscle.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Harmonises the Qi flow between the Upper and Middle Burner

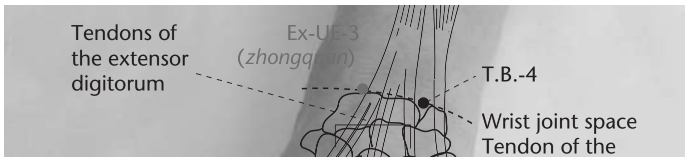
Tendons of
the extensor
digitorum
Ex-UE-3
(zhongquan)
T.B.-4
Wrist joint space
Tendon of the

Ex-UE-4 *Back of the Middle Finger* ZHONGKUI

Location

As the name implies, on the dorsal aspect of the middle finger, in the centre of the transverse creases of the proximal interphalangeal joint (PIP).

How to find

Slightly bend the PIP of the middle finger. The point is located at the vertex of the flexed joint.

Located in a comparable position on the little finger is ➞ **Ex-UE-6** (*xiaokugong*, in the centre of the proximal PIP of the little finger). Also located in a comparable position is ➞ **Ex-UE-10** (*sifeng*, four points on the palmar aspect of the fingers of each hand, in the centre of the creases of the PIP joints of the 2nd to 5th fingers).

Needling

Prick to bleed or moxibustion.

Actions/Indications

- Descends Stomach Qi
- Clears Heat

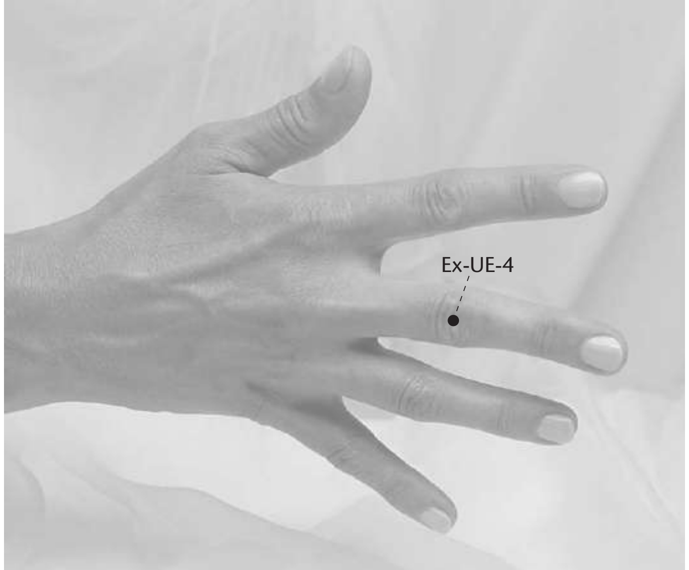

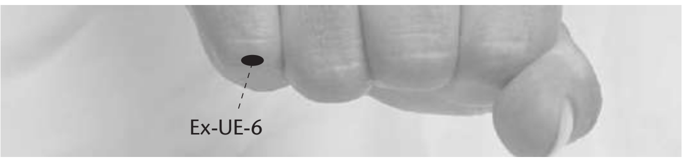

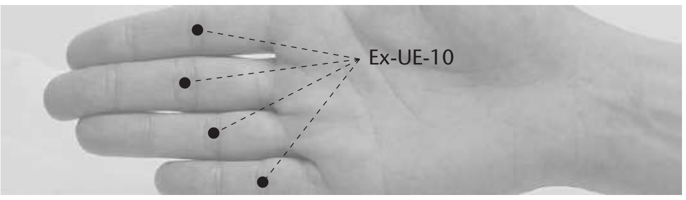

Thumb Joint DAGUKONG **Ex-UE-5**

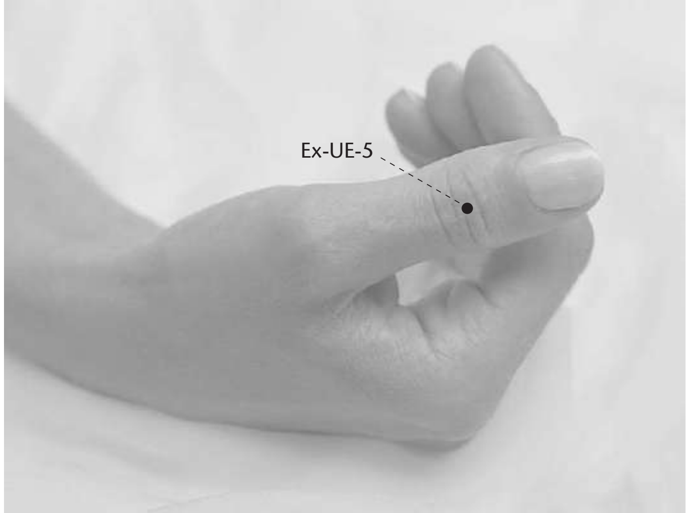

Location

As the name implies, on the dorsal aspect of the thumb, in the centre of the transverse creases of the interphalangeal joint.

How to find

Ask the patient to slightly flex the interphalangeal joint of the thumb. The point is located at the vertex of the bent joint.

Needling

Prick to bleed or moxibustion.

Actions/Indications

- Clears Heat
- Harmonises the Middle Burner

Ex-UE-6 *Little Finger Joint* XIAOGUKONG

Location

As the name implies, on the dorsal aspect of the little finger, in the centre of the transverse creases of the proximal interphalangeal joint (PIP).

How to find

Slightly bend the PIP of the little finger. The point is located at the vertex of the flexed joint.

Located in a comparable position on the middle finger are → **Ex-UE-4** (in the centre of the proximal PIP of the middle finger) and, on the palmar aspect of the little finger, → **Ex-UE-10** (four points on the palmar aspect of the fingers of each hand, in the centre of the creases of the PIP joints of the 2nd to 5th fingers).

Needling

Prick to bleed

Actions/Indications

- Clears Heat

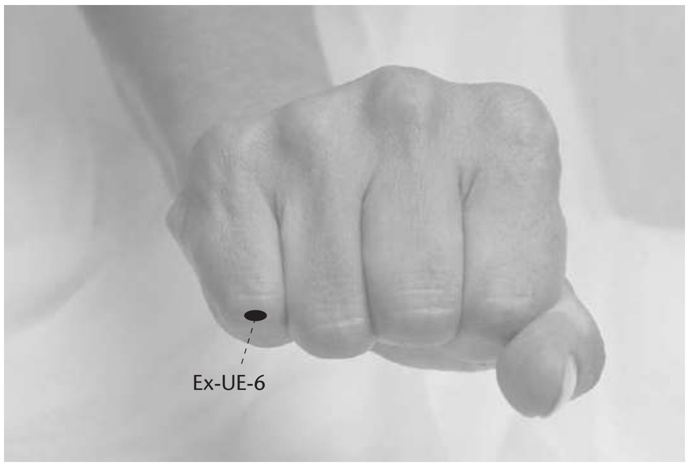

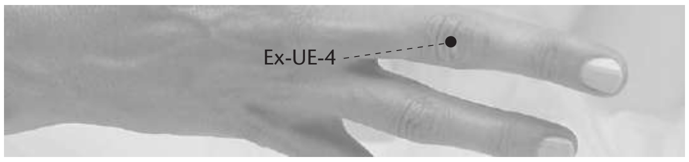

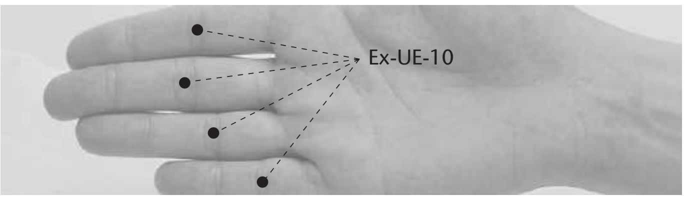

Lumbar Pair Point YAOTONGDIAN Ex-UE-7

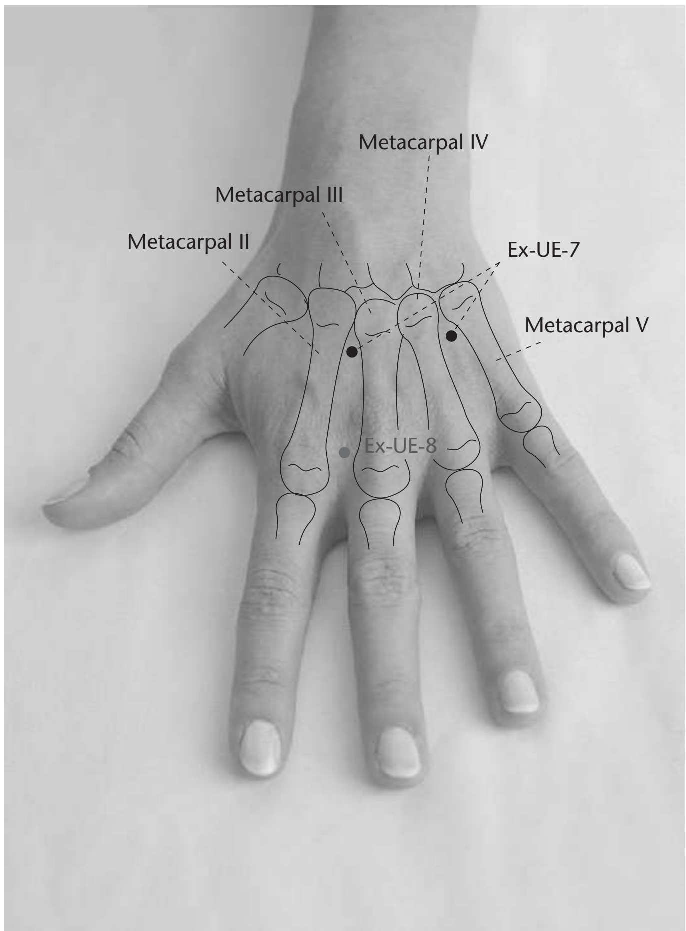
Metacarpal IV
Metacarpal III
Metacarpal II
Ex-UE-7
Metacarpal V
Ex-UE-8

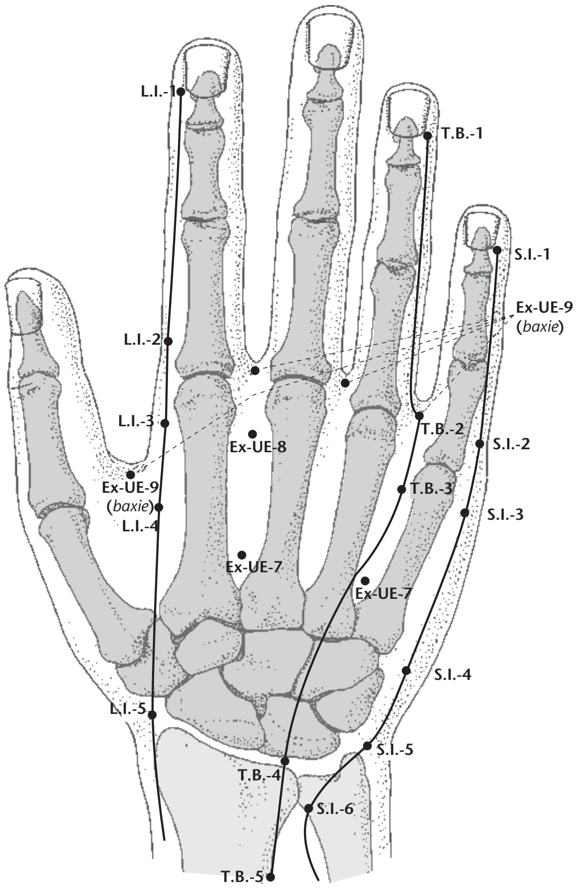
L.I.-1
T.B.-1
S.I.-1
Ex-UE-9
(*baxie*)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(*baxie*)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

Location

Two points on the dorsal aspect of the hand, between the 2nd/3rd and 4th/5th metacarpal bones, on the level of the junctions of the shaft and the base of the respective metacarpal bones.

How to find

On the dorsal aspect of the hand, glide with the palpating fingers in the grooves between the 2nd and 3rd as well as between the 4th and 5th metacarpal bones towards the wrist joint, until the fingers come to rest in the depressions just distal to the bases of the metacarpal bones. Here, locate the points of **Ex-UE-7** (*yaotongdian* or *yaotongxue*).

Needling

0.5–0.8 cun vertically or slightly obliquely towards the centre of a loose fist. Caution: Painful point! Do not stimulate these distal points in patients with a weak constitution: danger of needle collapse!

Actions/Indications

- Strengthen the Qi and Blood in the lumbar region, alleviate pain

Special features

Very effective distal point for acute disorders of the lumbar spine

Ch06b-F10028.qxd 2/22/08 10:04 PM Page 594

Ex-UE-8 *Stiff Neck* WAILAOGONG/LUOZHEN/XIANQIANG

Location

On the dorsal aspect of the hand, between the 2nd and 3rd metacarpal bones, proximal to the metacarpophalangeal joints, at the junction of the heads and the shafts of the metacarpal bones.

How to find

This point is best located with the patient making a loose fist. With the palpating finger, glide in the groove between the 2nd and 3rd metacarpal bones from the wrist joint towards the fingers until the finger comes to rest in the depression just proximal to the heads of the two metacarpal bones. Here, locate **Ex-UE-8** (*wailaogong*). Located in a comparable position between the 4th and 5th metacarpal bones is → **T.B.-3**.

Needling

0.5–1 cun vertically or slightly obliquely. Caution: Painful point! Do not stimulate this point in patients with a weak constitution: danger of needle collapse!

Actions/Indications

- Strengthens the Qi and Blood of the neck and alleviates pain

Special features

Very effective distal point for acute disorders of the cervical spine. Use strongly reducing needling techniques (for 1–2 minutes) on the affected or contralateral side while asking the patient to gently move their neck/cervical spine. The practitioner should keep the patient's hand in a fixed position and ensure that the hand and forearm are sufficiently supported.

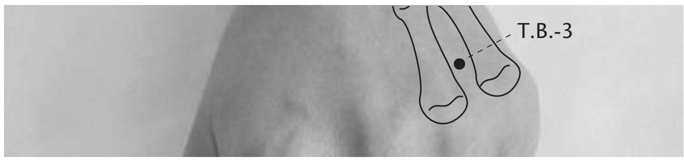

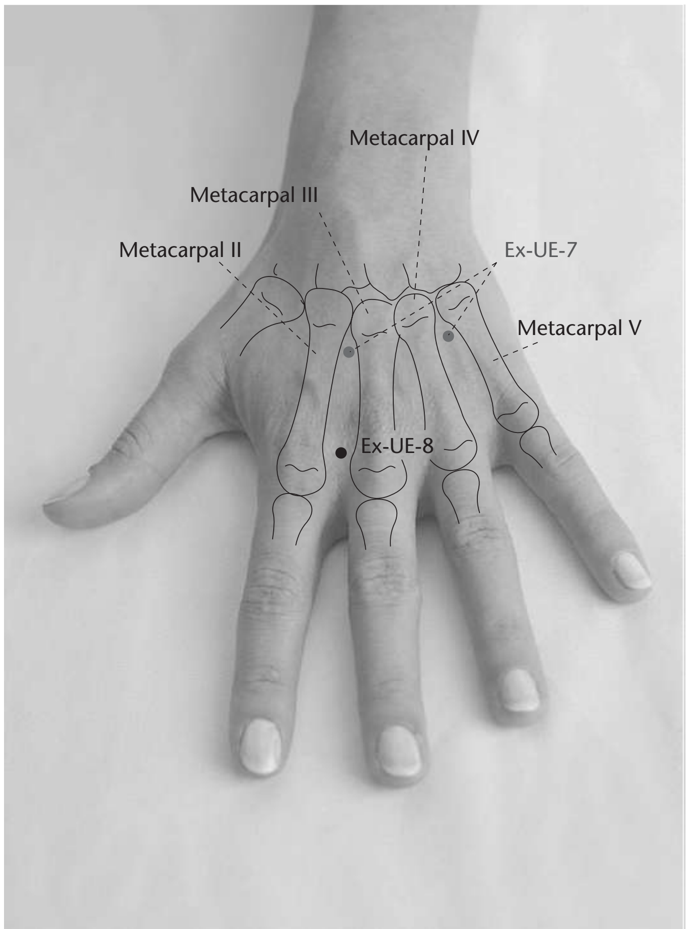
Metacarpal IV
Metacarpal III
Metacarpal II
Ex-UE-7
Metacarpal V
Ex-UE-8

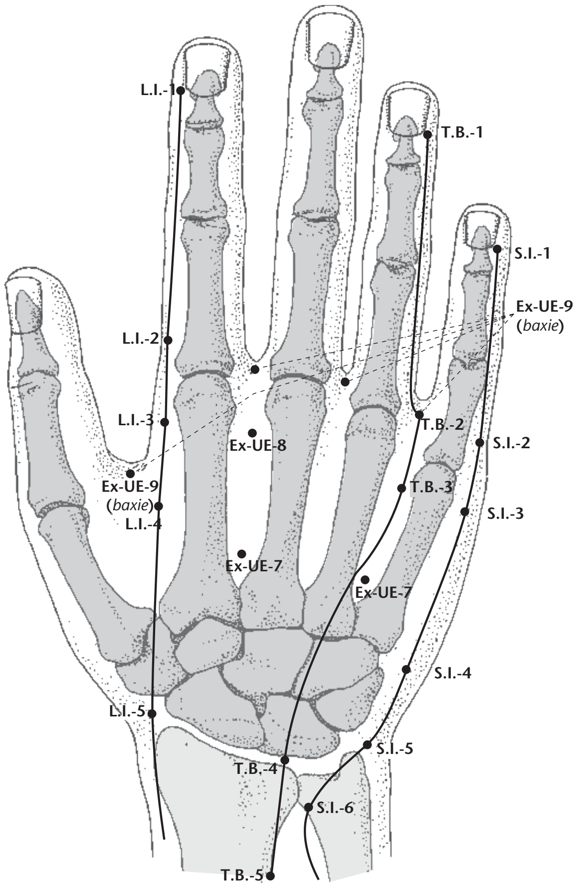
L.I.-1
T.B.-1
S.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
T.B.-2
S.I.-2
Ex-UE-8
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

Eight Pathogens BAXIE Ex-UE-9

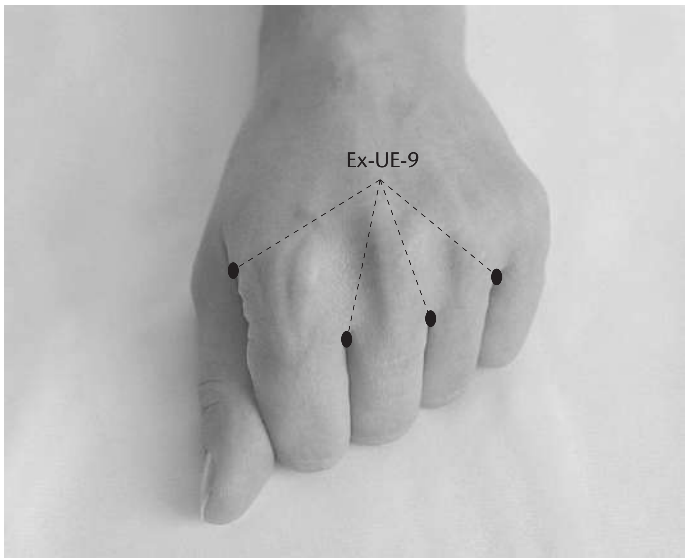

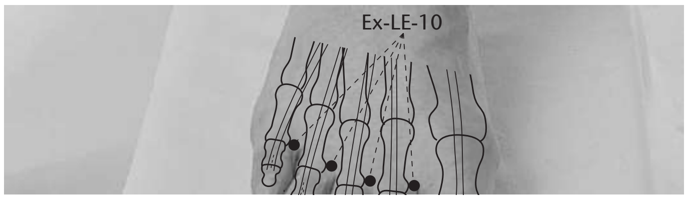

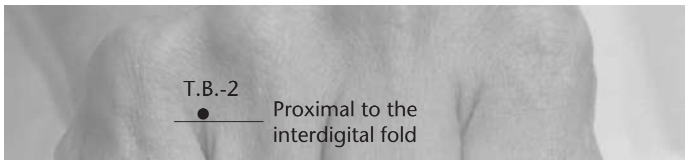

Location

Slightly proximal to the margins of the webs between the fingers, on the border of the red and white skin.

How to find

These points are best located when a loose fist is made. Locate the points on the border between the red and white skin. → **T.B.-2** (between the 4th and 5th fingers) is a point of **Ex-UE-9** (*baxie*). The points of → **Ex-LE-10** (*bafeng*) are located in a comparable position on the foot (proximal to the margins of the webs between the toes). → **LIV-2,** → **ST-44** and → **G.B.-43** are all part of → **Ex-LE-10**.

Needling

Up to 1 cun parallel to the metacarpal bones.

Actions/Indications

- Clears Heat
- Expels Cold and Dampness

Ex-UE-10 *Four Seams* SIFENG

Location

On the palmar aspect of the 2nd to 5th fingers, at the midpoint of the transverse creases of the proximal interphalangeal joints (PIP).

How to find

Locate these points on the palmar aspect of the 2nd to 5th fingers, at the midpoint of the transverse creases of the proximal interphalangeal joints (PIP).
Located in a similar position on the dorsal aspect of the little and middle fingers are → **Ex-UE-6** and → **Ex-UE-4** respectively.

Needling

Prick to bleed

Actions/Indications

- Harmonises the Qi flow between the Upper and Middle Burner

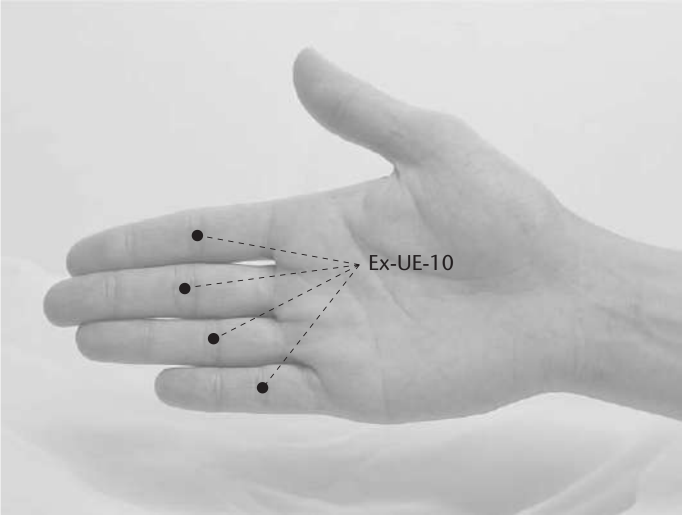

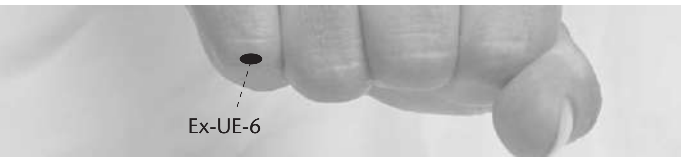

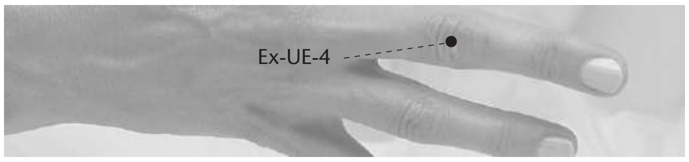

Ten Diffusions SHIXUAN Ex-UE-11

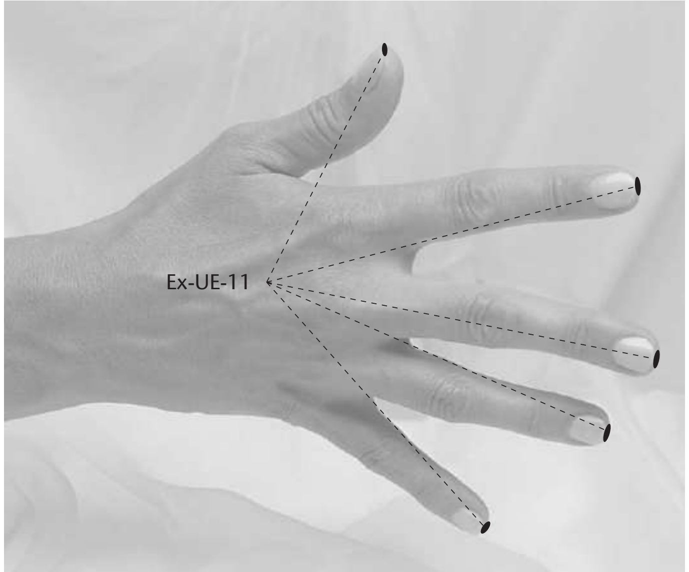

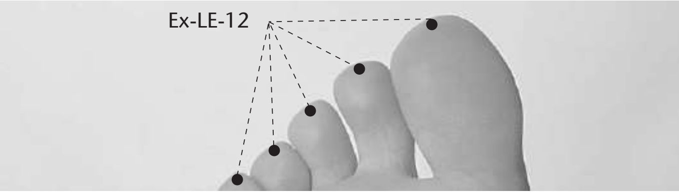

Location

On the tips of the 10 fingers.

How to find

These 10 points are located on the tips of the 10 fingers, approximately 0.1 cun from the free margin of the fingernail.

Located in a comparable position on the foot are the points of → **Ex-LE-12** (*qiduan*, on the tips of the 10 toes).

Needling

Prick to bleed

Actions/Indications

- Expels pathogenic factors

Front of the Shoulder JIANQIAN/JIANNEILING

Location

On the midpoint of a line connecting the end of the anterior axillary fold and → **L.I.-15**.

How to find

To locate → **L.I.-15**: Ask the patient to actively abduct their arm and locate **L.I.-15** in the depression anterior and inferior to the acromion (=ventrolateral extremity of the acromion). The end of the anterior axillary fold is formed by the easily palpable lower border of the pectoralis major muscle. *Jianqian* is located on the anterior aspect of the shoulder joint, midway between these two reference points.

Needling

Vertically up to 1.5 cun

Actions/Indications

- Opens the channel and *luo* vessels locally

Addendum: For more details, → Table 6.4

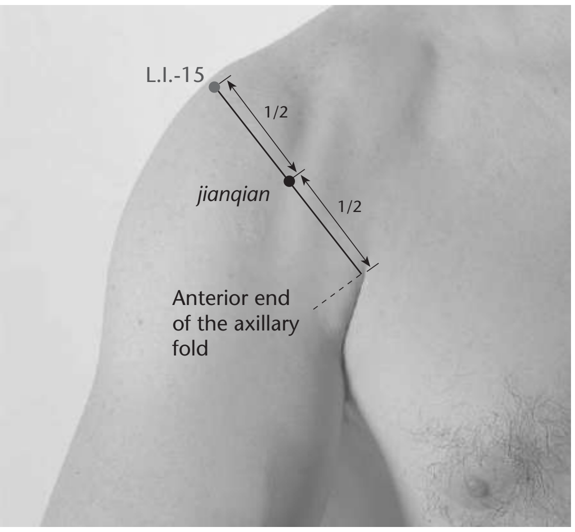
L.I.-15
1/2
jianqian
1/2
Anterior end of the axillary fold

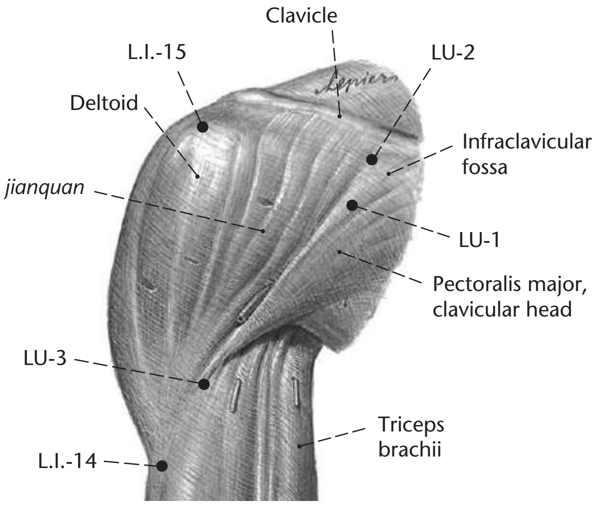
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
jianquan
LU-1
Pectoralis major, clavicular head
LU-3
Triceps brachii
L.I.-14

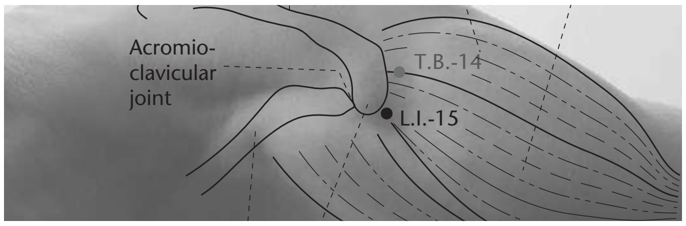
Acromio-clavicular joint
T.B.-14
L.I.-15

Arm Centre BIZHONG Ex-UE

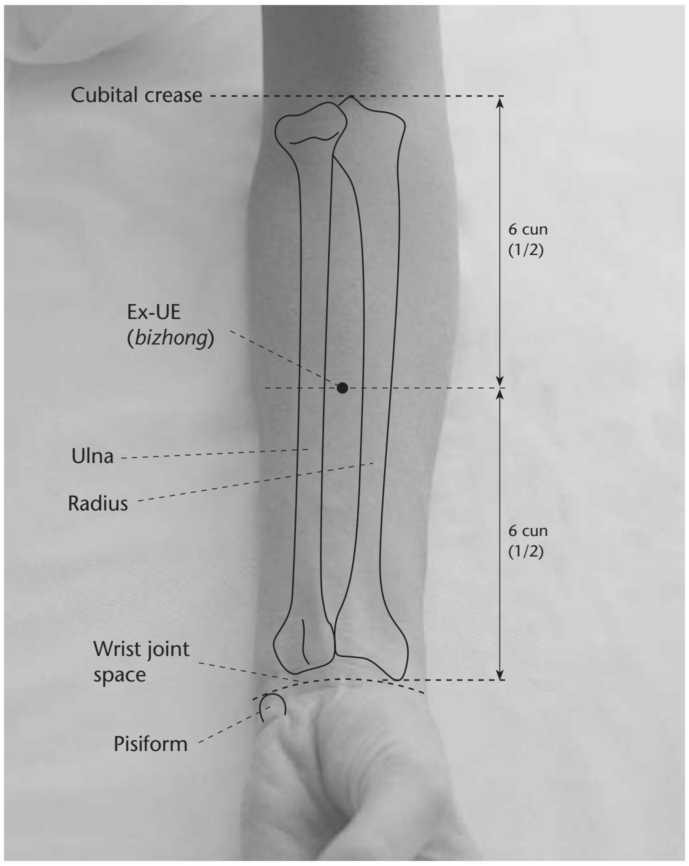
Cubital crease
6 cun
(1/2)
Ex-UE
(*bizhong*)
Ulna
Radius
6 cun
(1/2)
Wrist joint
space
Pisiform

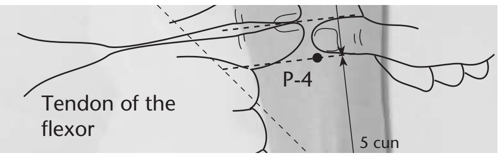
P-4
Tendon of the
flexor
5 cun

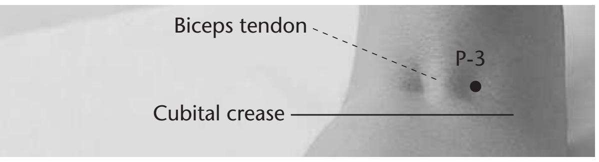
Biceps tendon
P-3
Cubital crease

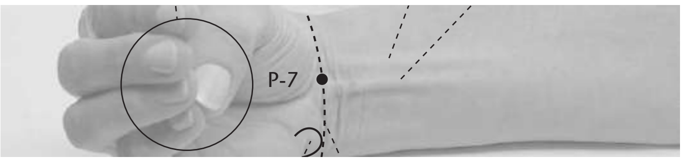

Location

On the palmar aspect of the forearm, midway between the cubital crease and the wrist joint space ('most distal wrist crease').

How to find

Place the patient's forearm in a relaxed supinated position. As the location of the wrist crease may vary, the wrist joint space presents a more reliable reference point (→ 3.3.3). The spreading hands technique is the best method for locating this point (→ 2.3.3): Place the little fingers on the centre of the wrist joint (→ **P-7**) and the cubital crease (→ **P-3**) respectively and determine the midpoint of this distance by joining the thumbs. Locate **Ex-UE** (*bizhong*) on this level between the radius and the ulna.

→ **P-4** is located 1 cun distal to the midpoint of the distance between → **P-3** and → **P-7** or 5 cun proximal to → **P-7**.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Opens the channel and the *luo* vessels locally

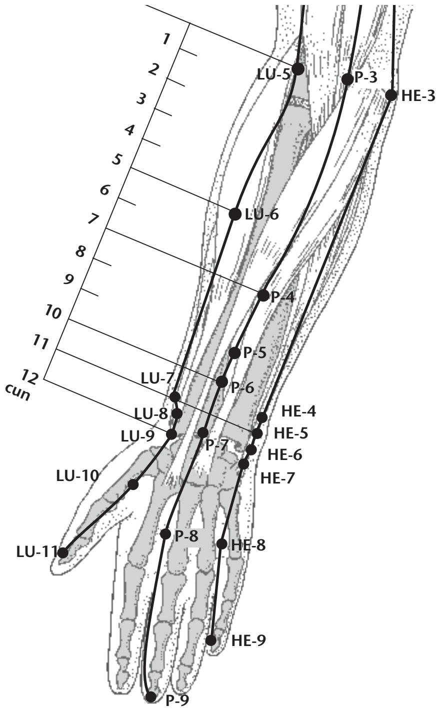
1
2
3
4
5
6
7
8
9
10
11
12
cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
LU-8
HE-4
LU-9
HE-5
P-7
HE-6
LU-10
HE-7
P-8
HE-8
LU-11
HE-9
P-9

6.5 Extra Points: Lower Extremities (EX-LE)

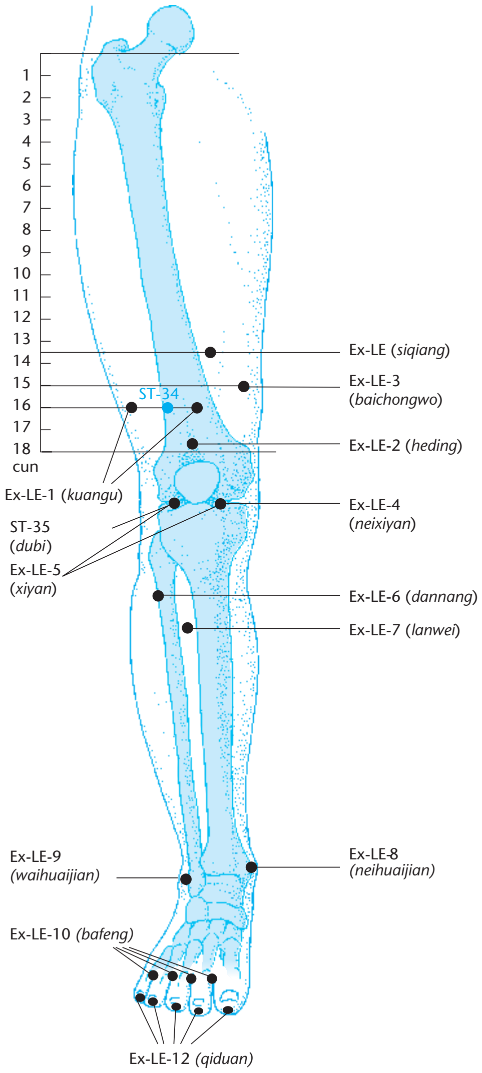
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
cun
ST-34
Ex-LE (*siqiang*)
Ex-LE-3
(*baichongwo*)
Ex-LE-2 (*heding*)
Ex-LE-1 (*kuangu*)
ST-35
(*dubi*)
Ex-LE-5
(*xiyan*)
Ex-LE-4
(*neixiyan*)
Ex-LE-6 (*dannang*)
Ex-LE-7 (*lanwei*)
Ex-LE-9
(*waihuaijian*)
Ex-LE-8
(*neihuaijian*)
Ex-LE-10 (*bafeng*)
Ex-LE-12 (*qiduan*)

English abbreviation (standard)	*Pinyin* name	Nguyen Van Nghi, König and Wancura, Schnorrenberger	Shanghai College	Ex (Hempen)
Ex-LE-1	*kuangu*	PaM or ZP165 (*changgu*)	M-LE-28	
Ex-LE-2	*heding/xiding*	PaM or ZP156	M-LE-27	
Ex-LE-3	*baichongwo*	PaM or ZP163	M-LE-34	Ex21
Ex-LE-4	*neixiyan*	(PaM or ZP145)	M-LE	(Ex23)
Ex-LE-5	*xiyan*	PaM or ZP145	M-LE	Ex23
Ex-LE-6	*dannangxue*	PaM or ZP152	M-LE-23	
Ex-LE-7	*lanweixue*	PaM or ZP142	M-LE-13	Ex22
Ex-LE-8	*neihuaijian*	PaM or ZP146	M-LE-17	
Ex-LE-9	*waihuaijian*	PaM or ZP151	M-LE-22	
Ex-LE-10	*bafeng*	PaM or ZP137	M-LE-8	Ex19
Ex-LE-11	*duyin*		M-LE	
Ex-LE-12	*qiduan*		M-LE-6	
Further extra points				
Ex-LE	*huanzhong*	PaM or ZP84	M-**BW-34**	Ex20
Ex-LE	*siqiang*	NP94	**N**-LE-19	
Ex-LE	*lineiting*	PaM or ZP130	M-LE-1	

Hip Bone KUANGU **Ex-LE-1**

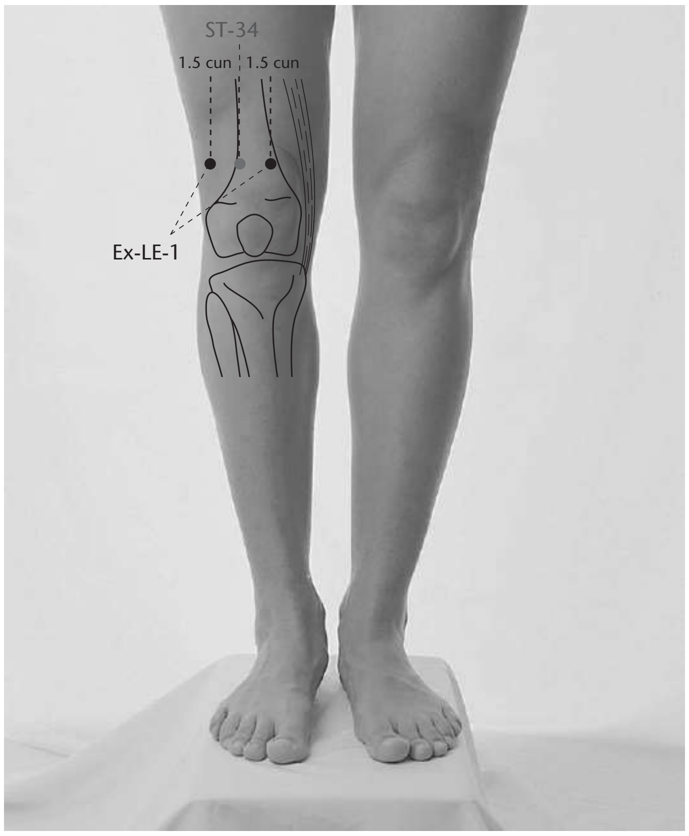
ST-34
1.5 cun
1.5 cun
Ex-LE-1

Location

A pair of points 2 cun superior to the patella and 1.5 cun lateral and medial to → **ST-34** (*liangqiu*).

How to find

First, locate → **ST-34**: measure 2 cun superiorly from the lateral upper border of the patella and there locate → **ST-34** in a depression on the vastus lateralis muscle. Then locate **Ex-LE-1** 1.5 cun lateral and medial to → **ST-34**.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Alleviate pain

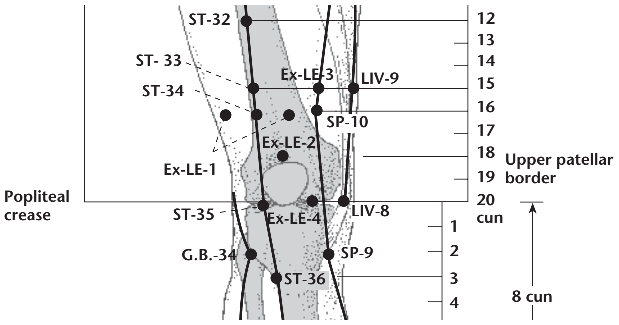
ST-32
ST- 33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
12
13
14
15
16
17
18
19
20
cun
Upper patellar border
1
2
3
4
8 cun

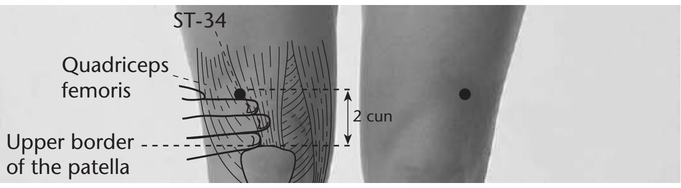
ST-34
Quadriceps femoris
Upper border of the patella
2 cun

Ex-LE-2 *Crane's Summit* HEDING/XIDING

Location

In the centre of the upper border of the patella.

How to find

This point is best located and needled with the patient's knee slightly flexed (knee support). Locate the midpoint of the upper patellar border and there locate **Ex-LE-2** (*heding*).

Needling

Vertically 0.5–0.8 cun

Actions/Indications

- Benefits the knee joint

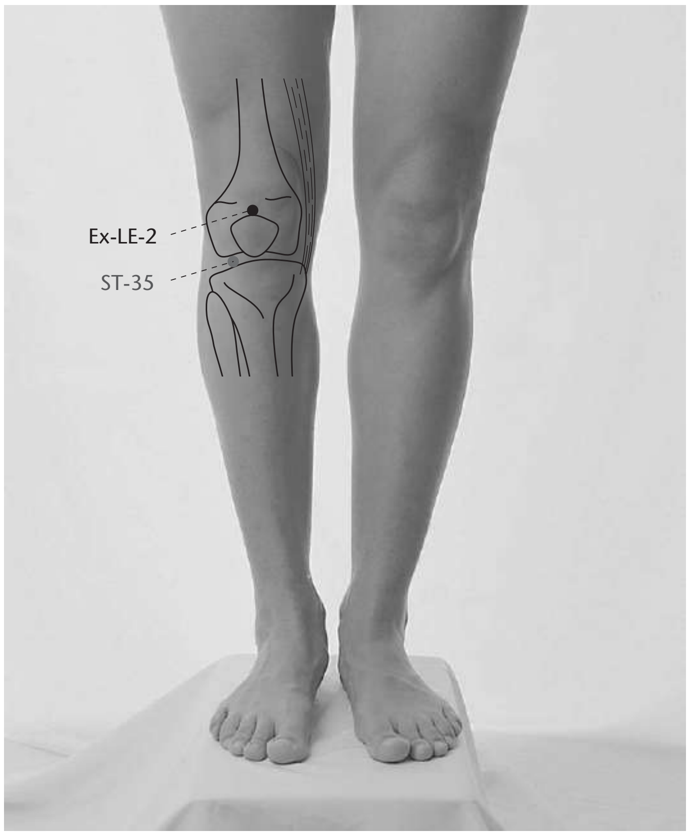

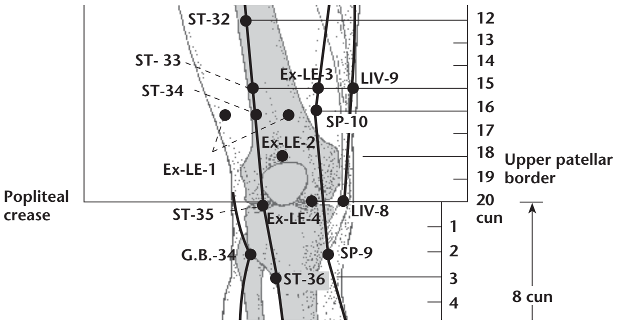
ST-32
ST- 33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
12
13
14
15
16
17
18
19
20 cun
Upper patellar border
1
2
3
4
8 cun

Hundred Insect Burrow BAICHONGWO Ex-LE-3

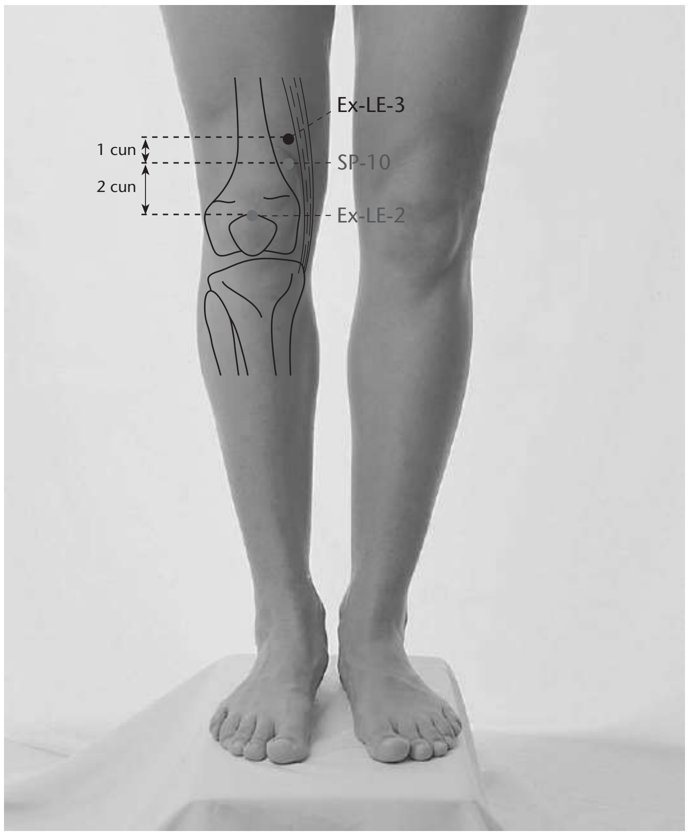

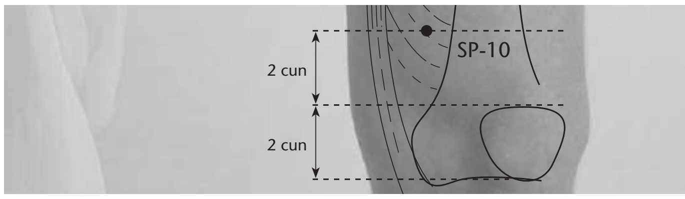

Location

3 cun superior and 1 cun medial to the upper medial border of the patella, in a small depression on the vastus medialis muscle or 1 cun superior to ➞ **SP-10**.

How to find

The patient should be in a supine or preferably in a seated position, with their knee flexed. Locate the upper medial border of the patella and, from there, measure 3 cun (1 handsbreadth) towards proximal. There, locate **Ex-LE-3** in a small depression on the vastus lateralis muscle.

Needling

Vertically 1–2 cun

Actions/Indications

- Cools Blood-Heat, eliminates Wind, drains Dampness

Ex-LE-4 *Inner Eye of the Knee* NEIXIYAN

Location

With the knee flexed, inferior to the patella, in a depression medial to the patellar ligament.

How to find

This point is best located with the patient's knee flexed (use knee support). The medial eye of the knee can be located in a distinct depression on the level of the lower border of the patella, medial to the patellar ligament. The lateral eye of the knee corresponds to → **ST-35** (*dubi*). Together, these two points form the extra point → **Ex-LE-5** (*xiyan*).

Needling

0.5–1 cun vertically or obliquely towards the lateral eye of the knee. Caution: Knee joint.

Actions/Indications

- Disorders of the knee joint

Special features

Important local point

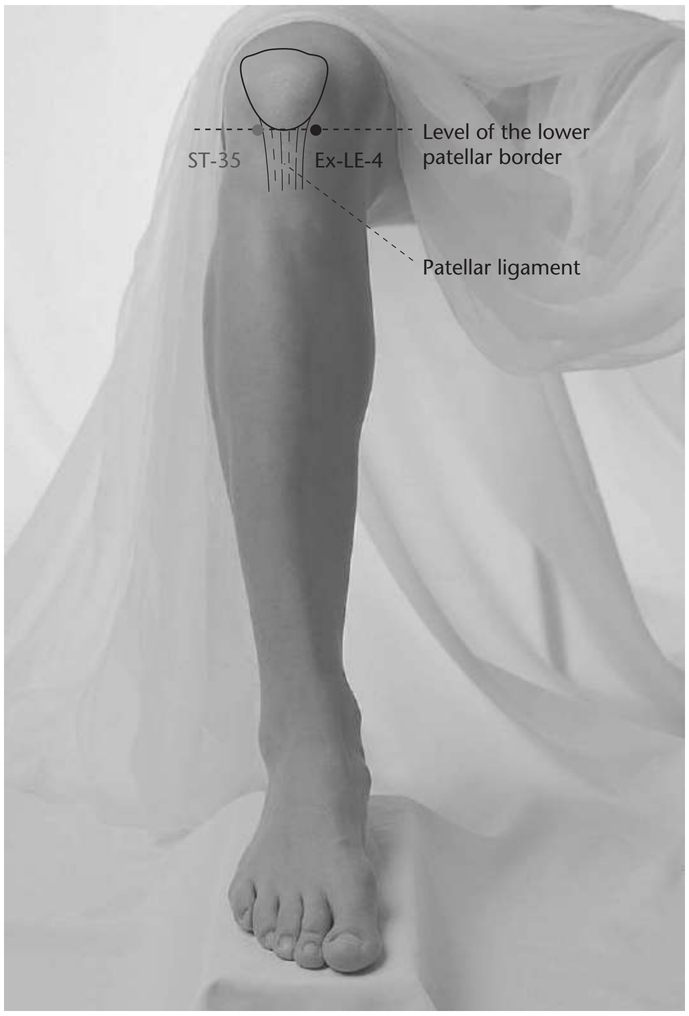
ST-35
Ex-LE-4
Level of the lower patellar border
Patellar ligament

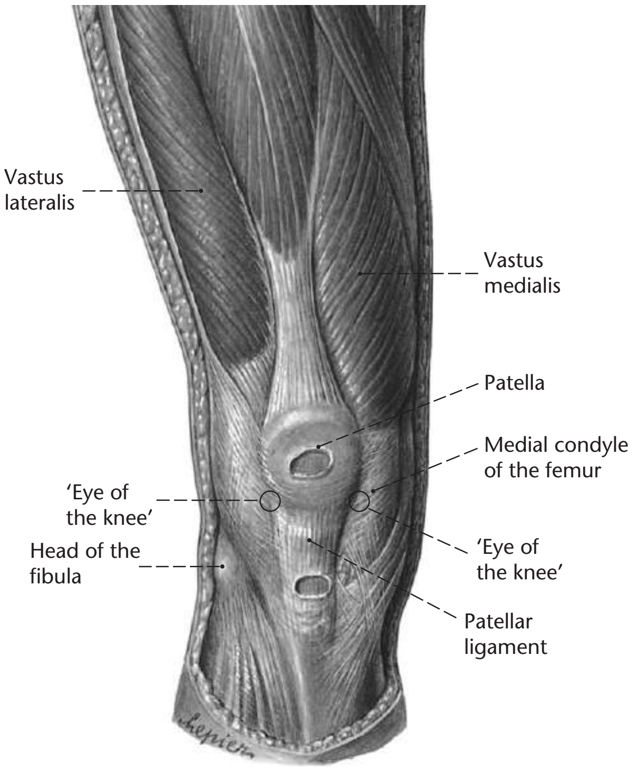
Vastus lateralis
Vastus medialis
Patella
Medial condyle of the femur
'Eye of the knee'
Head of the fibula
'Eye of the knee'
Patellar ligament

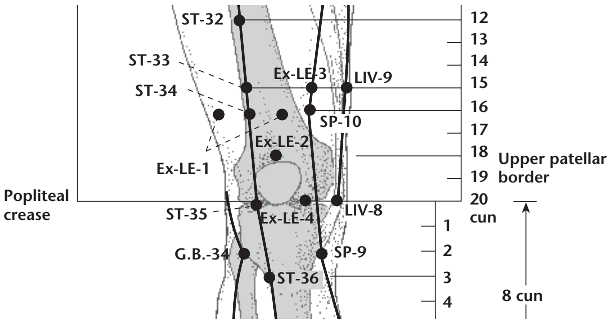
ST-32
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
12
13
14
15
16
17
18
19
20
cun
Upper patellar border
1
2
3
4
8 cun

Eyes of the Knee XIYAN **Ex-LE-5**

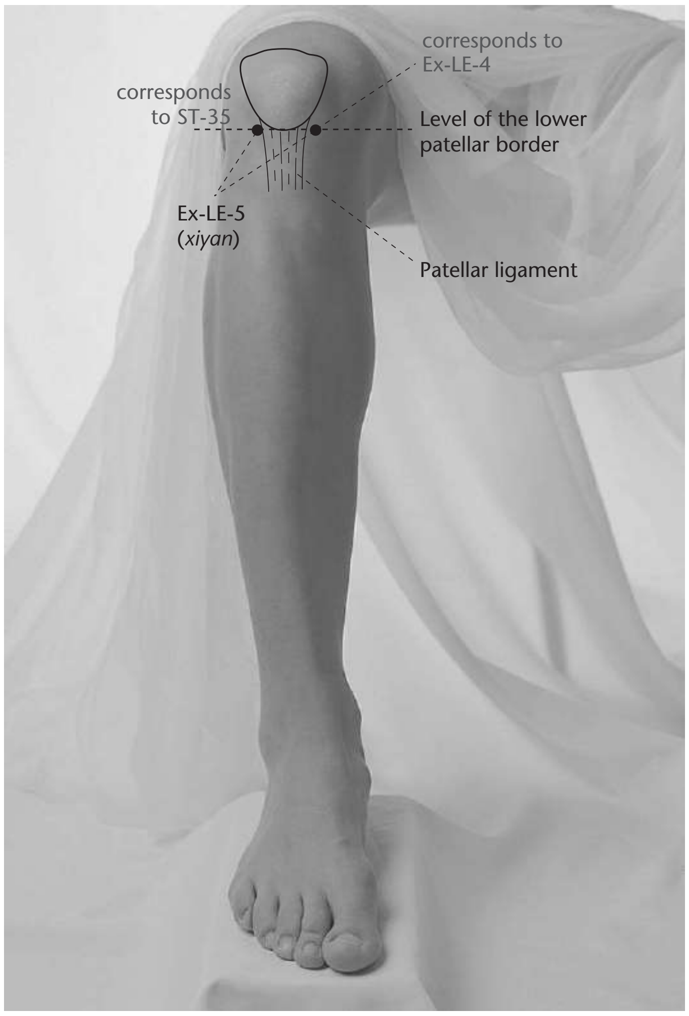

Location

With the knee flexed, this pair of points is located inferior to the patella, medial and lateral to the patellar ligament. **Ex-LE-5** (*xiyan*) includes two points: the medial eye of the knee corresponds to → **Ex-LE-4** (*neixiyan*), the lateral eye of the knee to **ST-35** (*dubi*).

How to find

This point is best located with the patient's knee flexed (use knee support). The eyes of the knee can be located on the level of the lower border of the patella, lateral and medial to the patellar ligament. The lateral eye of the knee corresponds to → **ST-35** (*dubi*), the medial eye of the knee to → **Ex-LE-4** (*neixiyan*). Together, these two points form the extra point **Ex-LE-5** (*xiyan*).

Needling

Medial eye of the knee: 0.5–1 cun vertically or obliquely towards the lateral eye of the knee; lateral eye of the knee: 0.5–1 cun vertically or obliquely towards the medial eye of the knee. Caution: Do not needle too deeply to avoid intra-articular puncture (these points correspond to the incision sites for knee arthroscopy).

Actions/Indications

- Eliminate Wind-Dampness, reduce swellings and alleviate pain

Special features

Important local points for disorders of the knee

Ex-LE-6 *Gall Bladder Point* DANNANG/DANNANGXUE/DANNANGDIAN

Location

On the G.B. channel of the **right** leg, approximately 1–2 cun distal to → **G.B.-34**, palpate for the most tender point.

How to find

First, locate → **G.B.-34** on the right leg: locate the head of the fibula on the lateral aspect of the leg (where the seam of the trousers would be) and hold it with the index and middle fingers (as with tweezers). By gliding distally with both fingers, the index finger will come to rest in a depression directly inferior and anterior to the head of the fibula (→ **G.B.-34**). From there, continue to palpate slowly along the G.B. channel, until you find the most tender point approximately 1–2 cun inferior to → **G.B.-34**. This is the location of **Ex-LE-6** (*dannangxue*).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Clears Heat and drains Dampness for acute and chronic disorders of the Gall Blader such as cholecystitis and cholelithiasis (tenderness on this point is diagnostically significant).

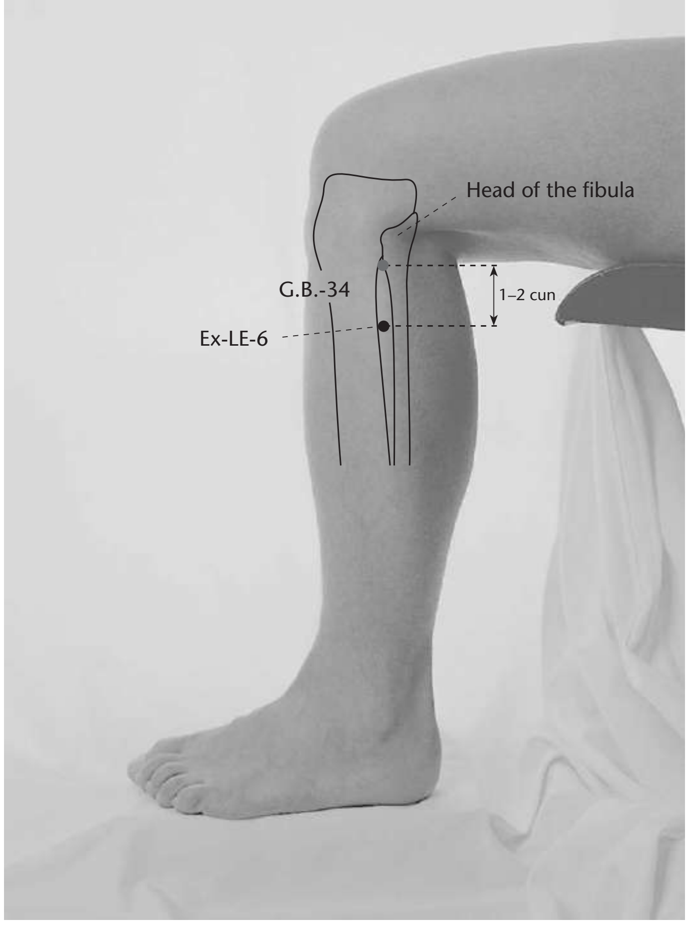

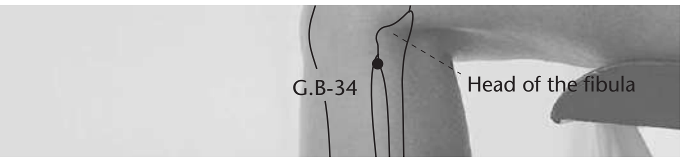

Appendix Point LANWEI/LANWEIXUE Ex-LE-7

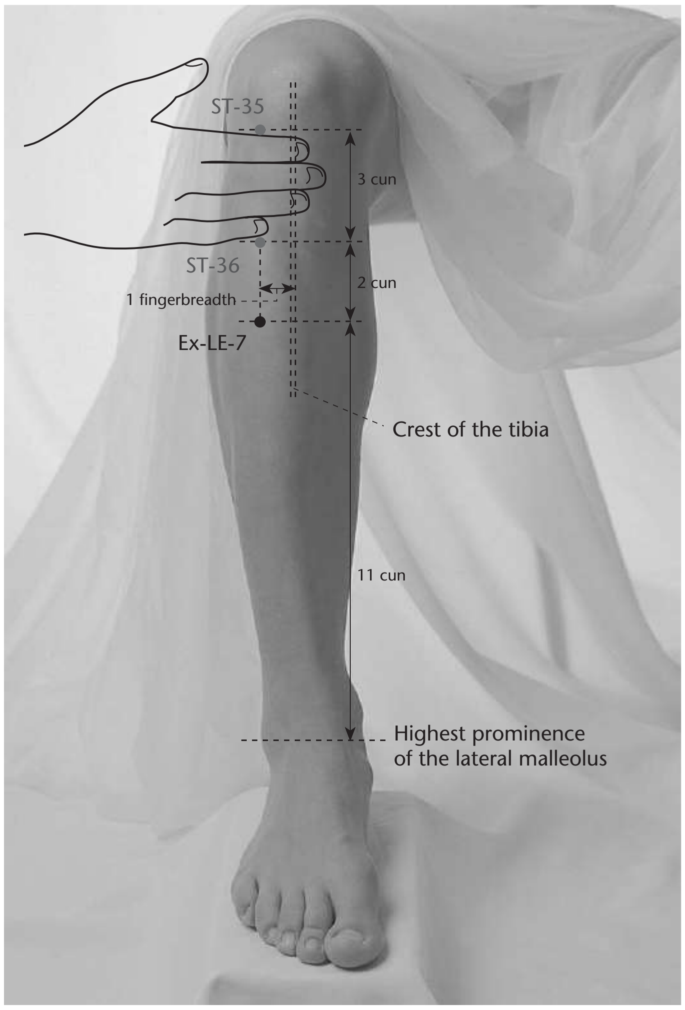
ST-35
3 cun
ST-36
2 cun
1 fingerbreadth
Ex-LE-7
Crest of the tibia
11 cun
Highest prominence
of the lateral malleolus

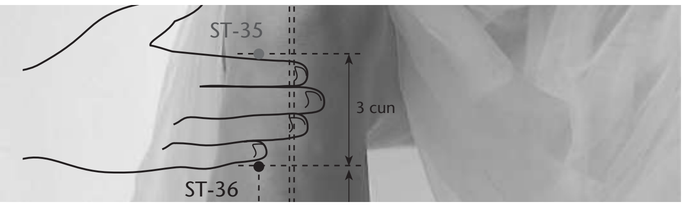
ST-35
3 cun
ST-36

Location

On the ST channel of the **right** leg, the most tender point approximately 2 cun distal to → **ST-36**.

How to find

Locate this point on the right leg only (this corresponds to the unilateral location of the appendix on the abdomen). First, locate → **ST-36** 3 cun inferior to the knee joint space (→ **ST-35**) and 1 cun lateral to the anterior crest of the tibia. Then continue to palpate along the ST channel for approximately 2 cun (towards → **ST-37**, 3 cun inferior to → **ST-36**) and locate **Ex-LE-7** (*lanweixue*) on the most tender point.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Activates the Qi and Blood flow in the Large Intestine, clears Heat and Heat toxins from the Large Intestine

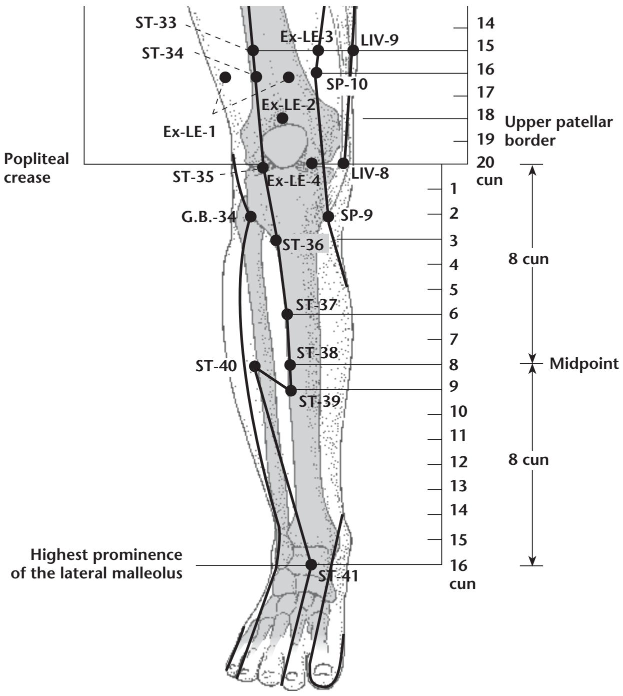
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-40
ST-38
ST-39
Highest prominence of the lateral malleolus
ST-41
14
15
16
17
18
19
20
cun
Upper patellar border
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun
8 cun
Midpoint
8 cun

Ex-LE-8 *Medial Malleolus Tip* NEIHUAIJIAN

Location

On the highest point of the medial malleolus.

How to find

Find the highest prominence of the medial malleolus and there locate **Ex-LE-8** (*neihuaijian*). Located in a comparable position but on the lateral malleolus is → **Ex-LE-9** (*waihuaijian*).

Needling

0.1 cun transversely (subcutaneously) or prick to bleed.

Actions/Indications

- Pain in the medial ankle region

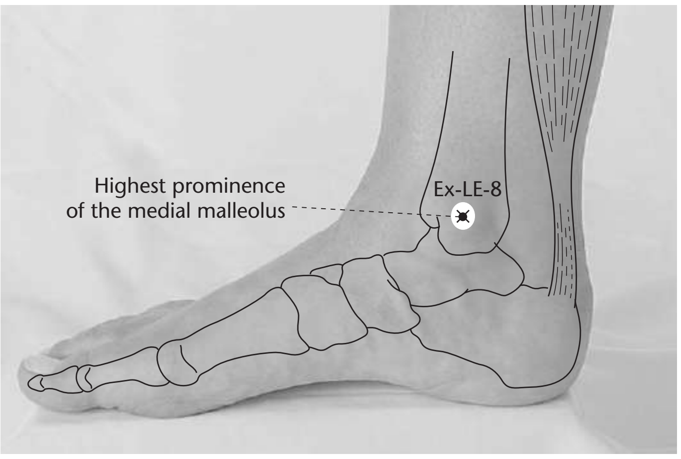
Highest prominence
of the medial malleolus
Ex-LE-8

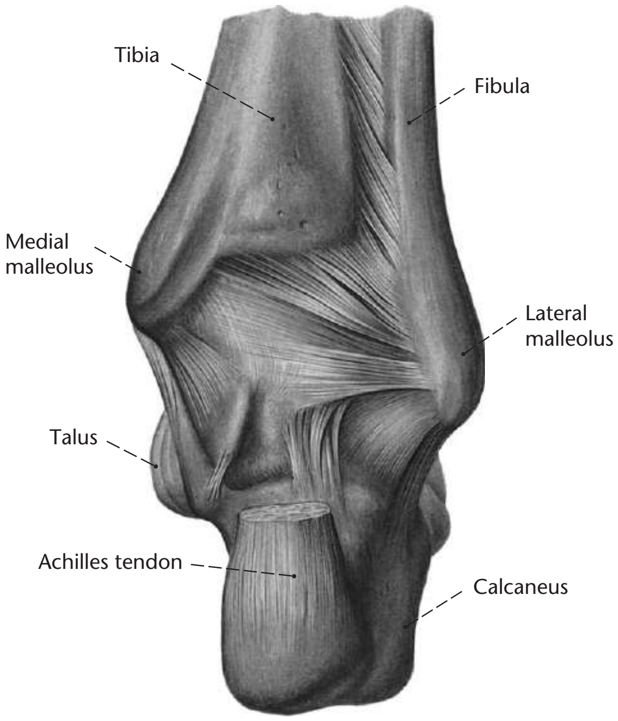
Tibia
Fibula
Medial
malleolus
Lateral
malleolus
Talus
Achilles tendon
Calcaneus

Outer Malleolus Tip WAIHUAIJIAN Ex-LE-9

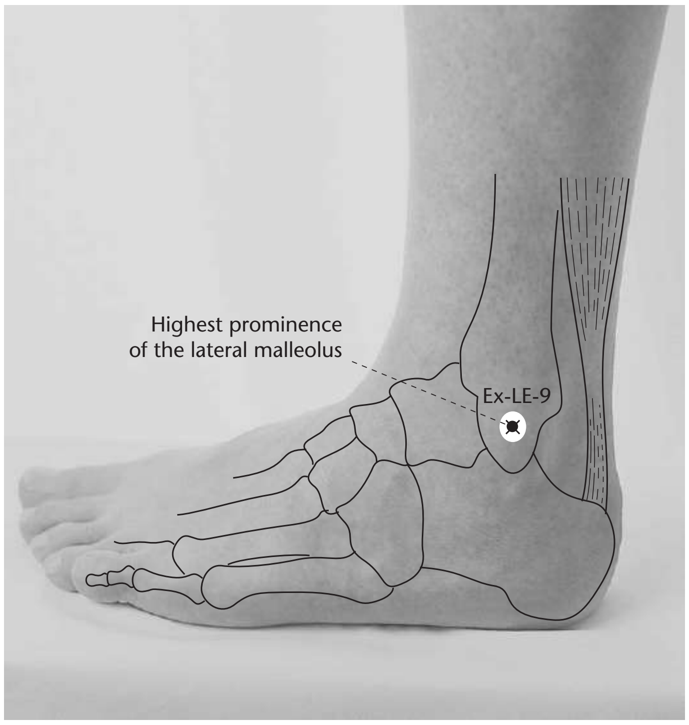

Location

On the highest prominence of the lateral malleolus

How to find

Find the highest prominence of the lateral malleolus and there locate **Ex-LE-9** (*waihuaijian*). Located in a comparable position but on the medial malleolus is → **Ex-LE-8** (*neihuaijian*).

Needling

0.1 cun transversely (subcutaneously) or prick to bleed.

Actions/Indications

- Pain in the lateral ankle region

Ex-LE-10 *Eight Winds* BAFENG

Location

Eight points located on the dorsum of the foot, slightly proximal to the margins of the interdigital webs.

How to find

Locate the points of **Ex-LE-10** (*bafeng*) on the dorsum of the foot, proximal to the margins of the webs between the toes, on the border of the red and white skin. → **LIV-2** (*xingjian*), → **ST-44** (*neiting*) and → **G.B.-43** (*xiabai*) are all part of the *bafeng* points. The extra points → **Ex-UE-9** (*baxie*) are located in a comparable position on the dorsum of the hand, proximal to the margins of the webs between the fingers (→ **T.B.-2** is part of the → **Ex-UE-9**).

Needling

0.3–1 cun obliquely in a proximal direction or prick to bleed.

Actions/Indications

- Clears Heat and reduces swelling

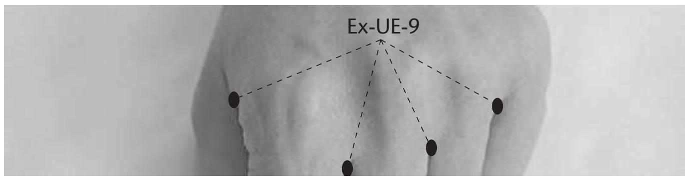

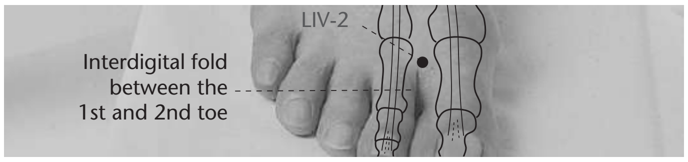
LIV-2
Interdigital fold between the 1st and 2nd toe

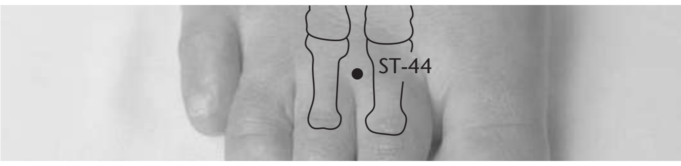

Proximal to the interdigital fold
G.B.-43

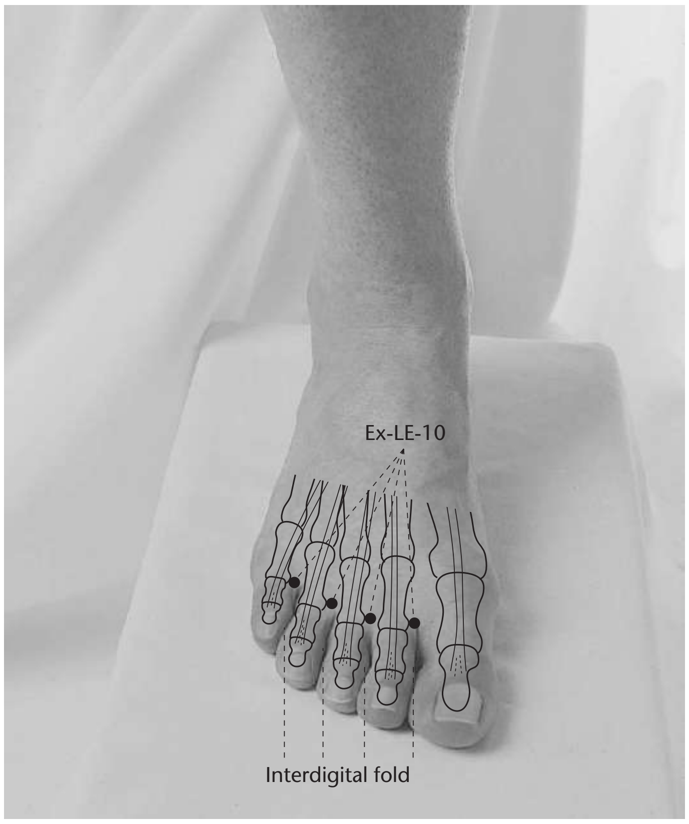
Ex-LE-10
Interdigital fold

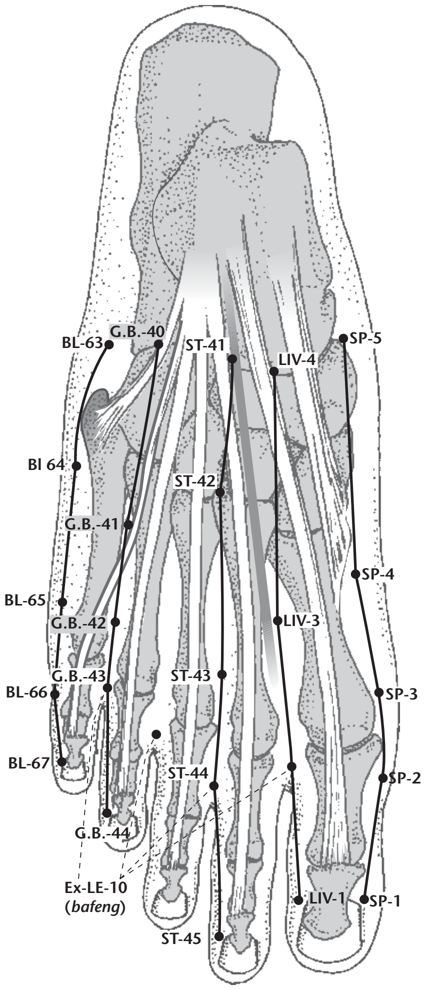
BL-63
G.B.-40
ST-41
LIV-4
SP-5
Bl 64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10 (bafeng)
LIV-1
SP-1
ST-45

Solitary Yin **DUYIN** **Ex-LE-11**

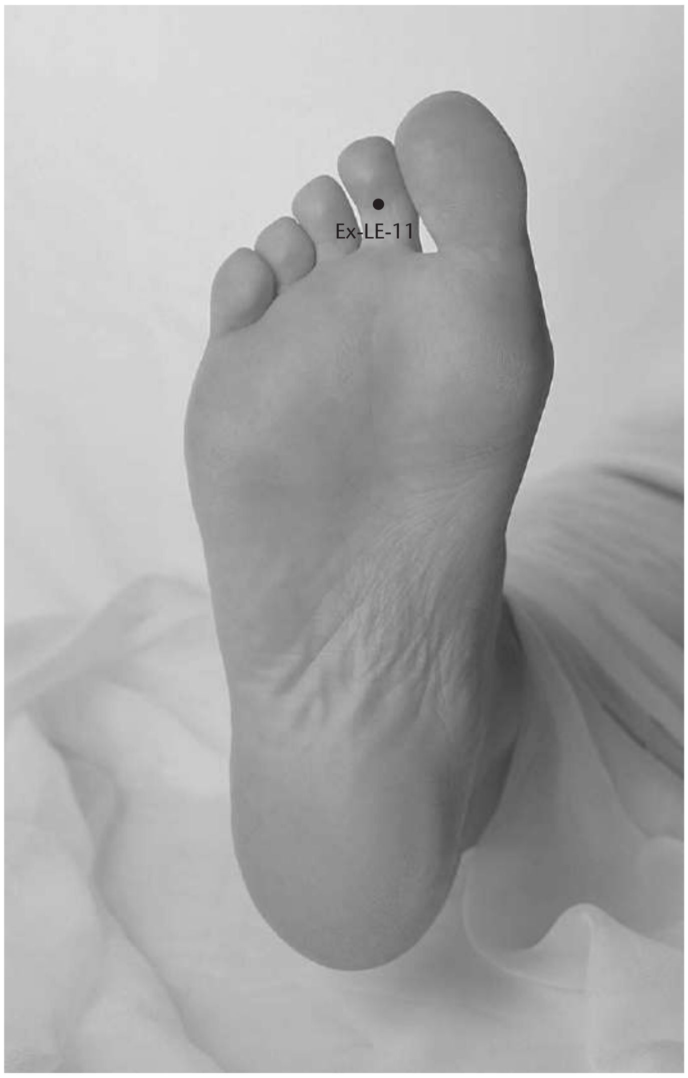

Location

On the plantar aspect of the 2nd toe, at the midpoint of the transverse crease of the distal interphalangeal joint.

How to find

Locate the plantar aspect of the 2nd toe. There, locate **Ex-LE-11** (*duyin*) at the midpoint of the transverse crease of the distal interphalangeal joint.

Needling

0.2–0.3 cun vertically or transversely (subcutaneously) or prick to bleed or moxibustion.

Actions/Indications

- Acute angina, thoracic and hypochondriac pain, nausea, vomiting, retention of the lochia, irregular menstruation, inguinal hernia

Ch06b-F10028.qxd 2/22/08 10:05 PM Page 614

Ex-LE-12 *End of Qi* QIDUAN

Location

On the tips of the 10 toes.

How to find

Locate the points of **Ex-LE-12** (*qiduan*) on the tips of the 10 toes, 0.1 cun from the free margin of the nail.
The points of → **Ex-UE-11** (*shixuan*) are located in a comparable position on the hand, on the tips of the 10 fingers.

Needling

Prick to bleed

Actions/Indications

- Syncope, oedema of the feet, acute abdominal pain

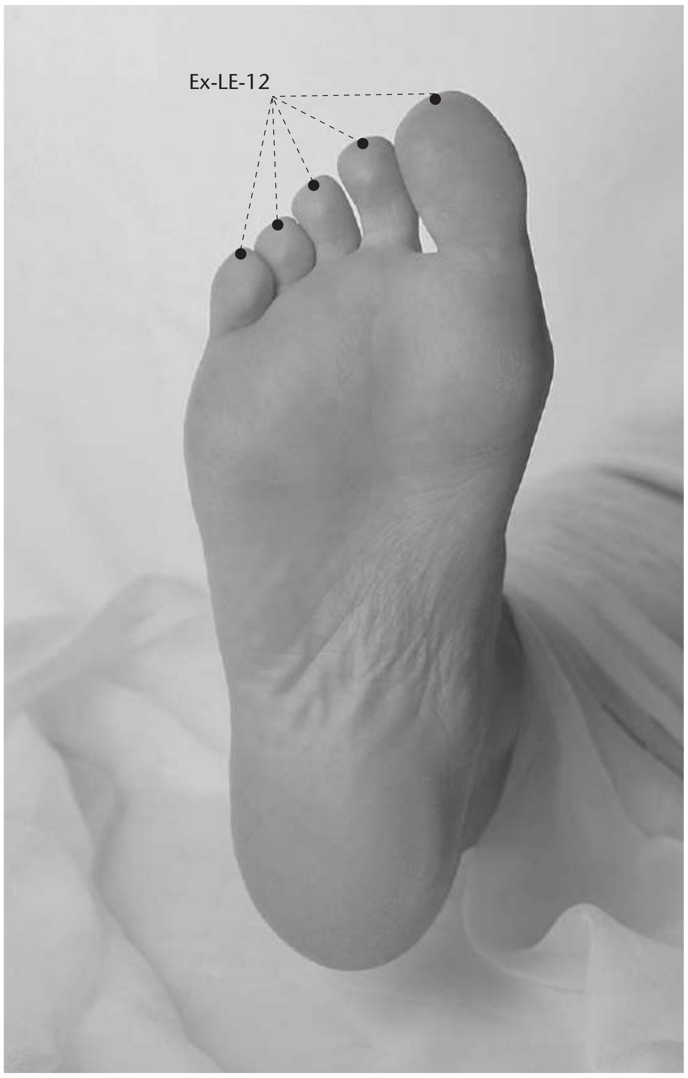

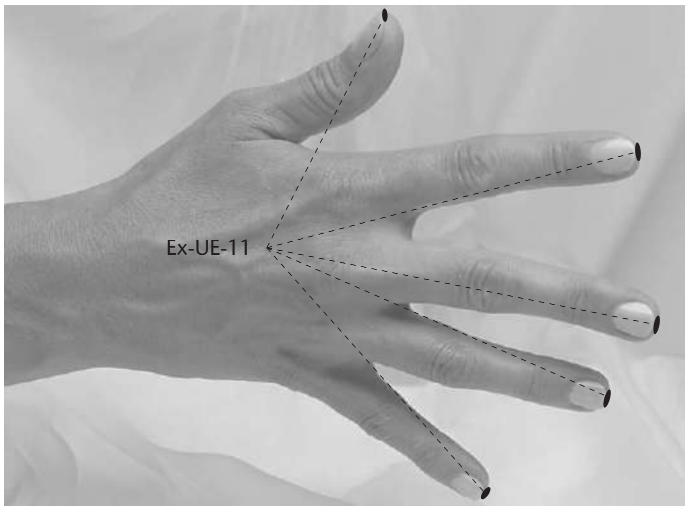

Circle Centre HUANZHONG Ex-LE

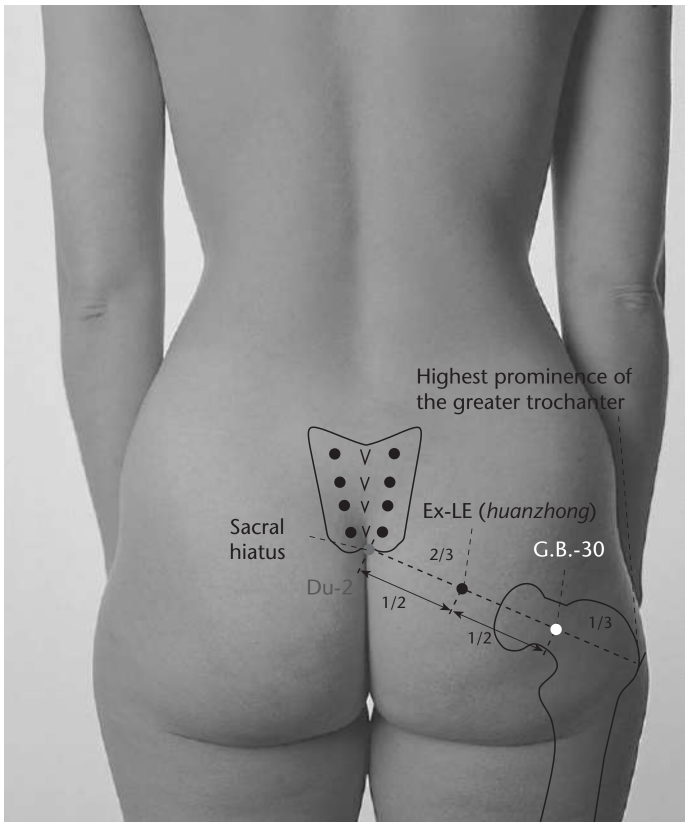
Highest prominence of the greater trochanter
Sacral hiatus
Ex-LE (*huanzhong*)
2/3
G.B.-30
Du-2
1/2
1/2
1/3

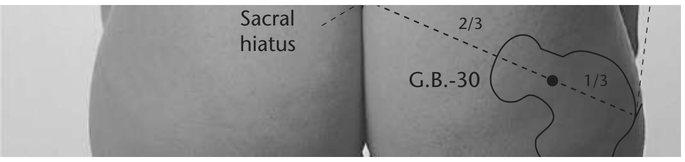
Sacral hiatus
2/3
G.B.-30
1/3

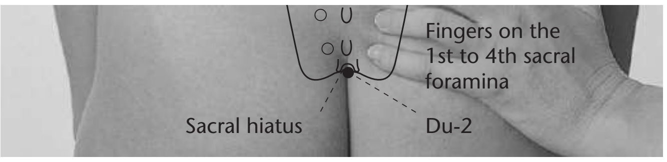
Fingers on the 1st to 4th sacral foramina
Sacral hiatus
Du-2

Location

Midway between → **G.B.-30** (*huantiao*) and → **Du-2** (*yaoshu*).

How to find

First, locate → **G.B.-30**: divide the distance between the prominence of the greater trochanter and the sacral hiatus into thirds (→ 3.4.4) and locate → **G.B.-30** a third of the distance from the greater trochanter. Then locate → **Du-2** directly below the sacral hiatus. Next, locate **Ex-LE** (*huanzhong*) midway between these two points.

Needling

Vertically 2–2.5 cun

Actions/Indications

- Sciatica, urinary tract infection, haemorrhoids, paralysis of the lower extremities

Special features

- This point is not part of the WHO classification of extra points.

Ex-LE *Four Muscles Strengthening Point* SIQIANG

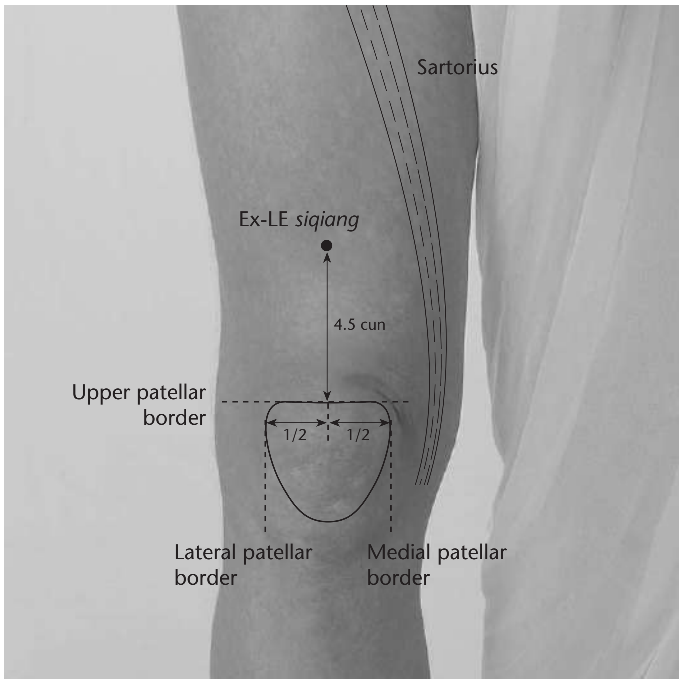
Sartorius
Ex-LE *siqiang*
4.5 cun
Upper patellar border
1/2
1/2
Lateral patellar border
Medial patellar border

Location

4.5 cun superior to the centre of the upper patellar border.

How to find

Find the centre of the upper patellar border. From there, measure 4.5 cun (1 handbreadth and 2 fingerbreadths) in a superior direction. **Or:** The distance between the upper border of the pubic symphysis and the upper border of the patella is 18 cun (➞ 2.2). Divide this distance into quarters (spreading hands technique, elastic tape ➞ Chapter 2) and **Ex-LE** (*siqiang*) is located on the centre of the thigh, one quarter of the distance from the upper patellar border.

Needling

Vertically 1–2 cun

Actions/Indications

- Paralysis and atrophy of the muscles of the lower extremities, especially of the quadriceps femoris muscle

Special features

'Four muscles' refers to the four portions of the quadriceps femoris muscle. This point is not part of the WHO classification of extra points.

Plantar Inner Space LINEITING Ex-LE

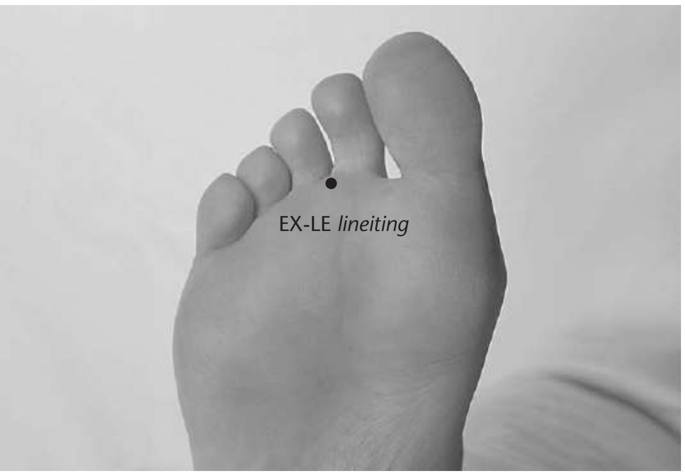

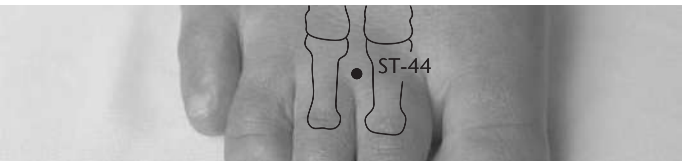

Location

On the plantar aspect of the foot, between the 2nd and 3rd metatarsal bones, opposite → **ST-44** (*neiting*).

How to find

Locate the space between the 2nd and 3rd toes on the plantar aspect of the foot. The extra point *lineiting* lies opposite → **ST-44** (on the dorsal aspect of the foot, proximal to the margin of the web between the 2nd and 3rd toe).

Needling

0.2–0.3 cun vertically towards → **ST-44**. Caution: Painful point.

Actions/Indications

- Acute epigastric pain, local pain, epilepsy, restlessness

Special features

This point is not part of the WHO classification of extra points.

7 Important Points According to Region

Claudia Focks

7.1 Anterior Aspect of the Head

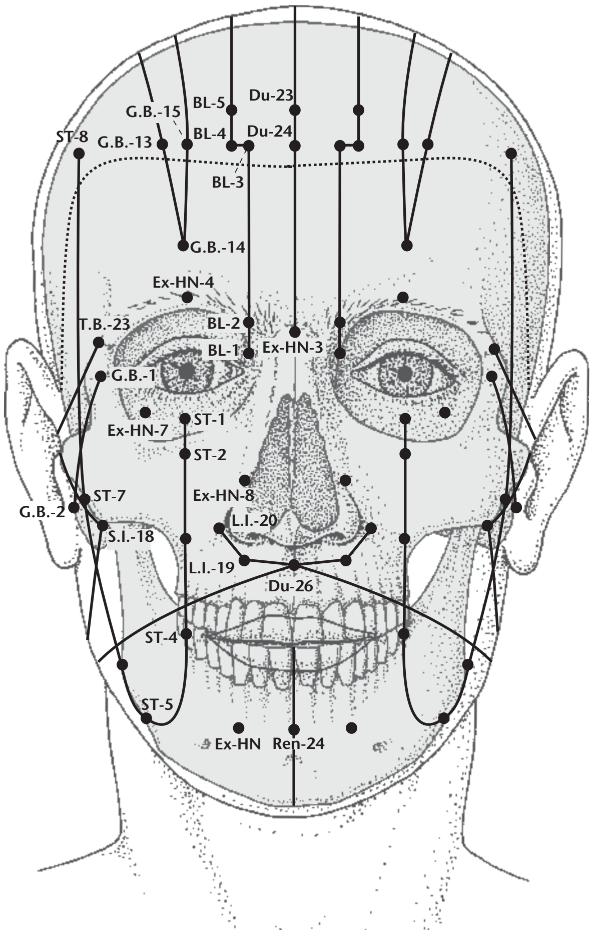
Du-23
BL-5
G.B.-15
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
ST-7
G.B.-2
Ex-HN-8
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren-24

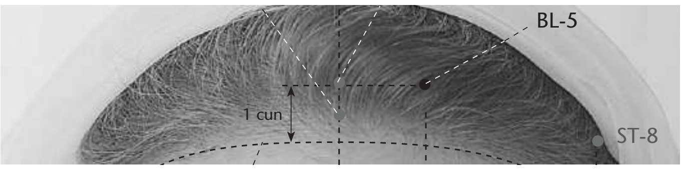
BL-5
1 cun
ST-8

BL-5 (*wuchu*) 'Fifth Place' 1 cun superior to the anterior hairline and 1.5 cun lateral to the midline (directly above **BL-4**, level with **Du-23**)

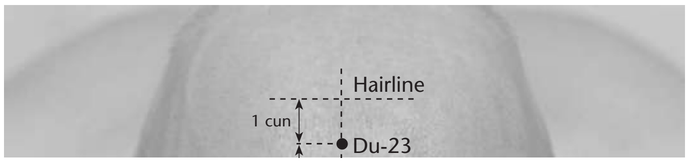
Hairline
1 cun
Du-23

Du-23 (*shangxing*) 'Upper Star' On the midline, 1 cun superior to the anterior hairline

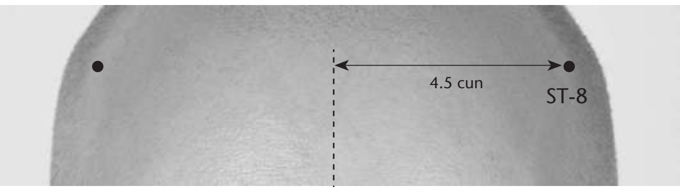

ST-8 (*touwei*) 'Head's Binding' At the temporal corner of the forehead, on the border of the temporalis muscle and 0.5 cun within the anterior hair line or 4.5 cun lateral to the anterior midline (**Du-24**)

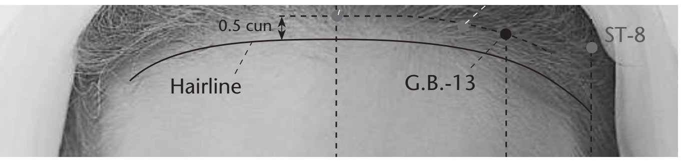
0.5 cun
ST-8
Hairline
G.B.-13

G.B.-13 (*benshen*) 'Root of the Spirit' 0.5 cun superior to the anterior hairline and 3 cun lateral to the midline, on the line between **Du-24** and **ST-8** (= 4.5 cun) at the junction of the lateral and medial third

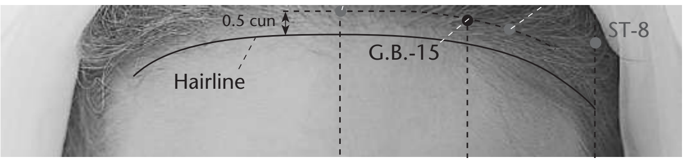
0.5 cun
ST-8
G.B.-15
Hairline

G.B.-15 (*toulinqi*) 'Head Governor of Tears' On the pupil line when looking straight ahead, 0.5 cun superior to the anterior hairline, at the midpoint of the line between **Du-24** and **ST-8** (= 4.5 cun)

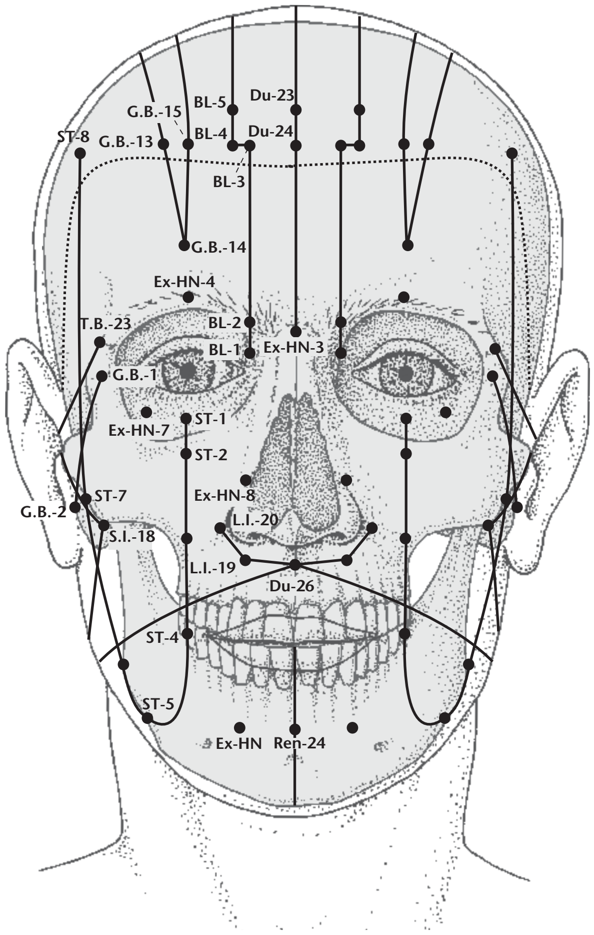
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
Ex-HN-8
ST-7
G.B.-2
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren-24

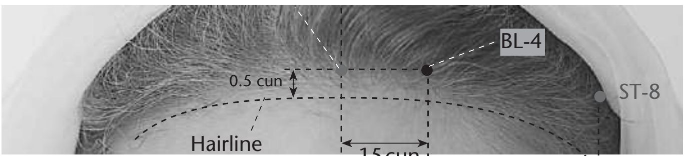
BL-4
ST-8
0.5 cun
Hairline
1.5 cun

BL-4 (*qucha*) 'Crooked Curve' 0.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline, or one third of the distance between **Du-24** and **ST-8** (= 4.5 cun)

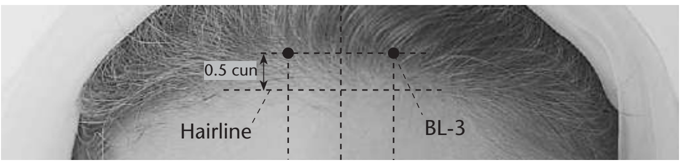
0.5 cun
Hairline
BL-3

BL-3 (*meichong*) 'Eyebrows' Pouring' 0.5 cun within the anterior hairline, superior to the medial canthus of the eye (**BL-1**)

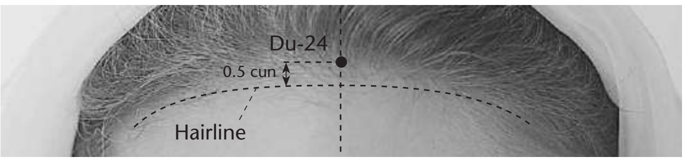
Du-24
0.5 cun
Hairline

Du-24 (*shenting*) 'Courtyard of the Spirit' On the midline, 0.5 cun superior to the anterior hairline

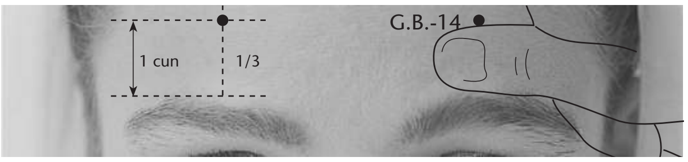
1 cun
1/3
G.B.-14

G.B.-14 (*yangbai*) 'Yang White' When looking straight ahead on the pupil line, approximately 1 cun superior to the midpoint of the eyebrow (centre of the eyebrow to anterior hairline = 3 cun)

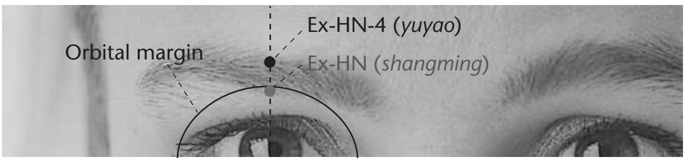
Orbital margin
Ex-HN-4 (yuyao)
Ex-HN (shangming)

Ex-HN-4 (*yuyao*) 'Fish Waist' In the centre of the eyebrow, with the eyes looking straight ahead directly above the pupil

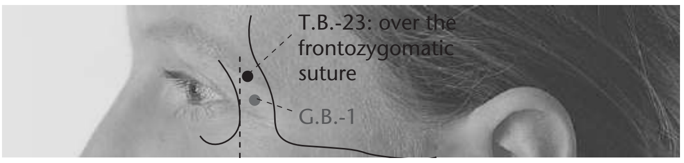
T.B.-23: over the frontozygomatic suture
G.B.-1

T.B.-23 (*sizhukong*) 'Silken Bamboo Hollow' In the depression (frontozygomatic suture) at the lateral end of the eyebrow

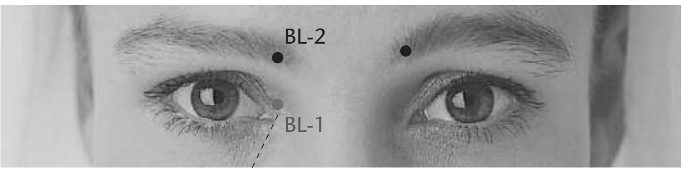

BL-2 (*zanzhu*) 'Gathered Bamboo' In a depression at the medial end of the eyebrow, directly superior to the inner canthus of the eye

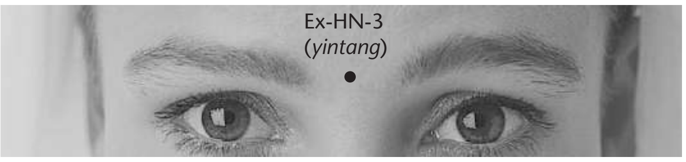
Ex-HN-3
(yintang)

Ex-HN-3 (*yintang*) 'Hall of Impression' On the anterior midline between the eyebrows

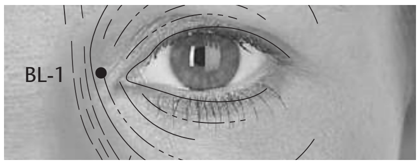

BL-1 (*jingming*) 'Bright Eyes' 0.1 cun superior to the inner canthus of the eye, where spectacles tend to rest

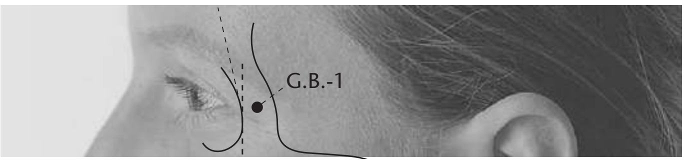

G.B.-1 (*tongziliao*) 'Pupil Crevice' 0.5 cun lateral to the outer canthus of the eye

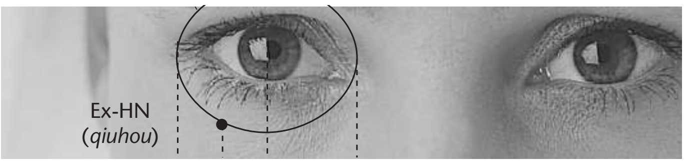

Ex-HN-7 (*qiuhou*) 'Behind the Ball' On the lower border of the orbit, at the junction of the lateral quarter and the medial three-quarters of the orbital margin

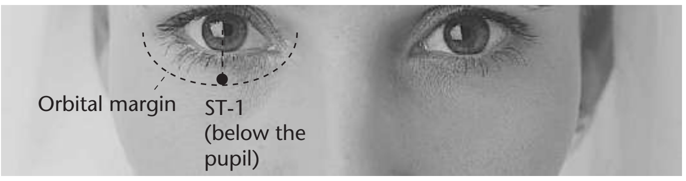

ST-1 (*chengqi*) 'Container of Tears' With the eyes looking straight forward, directly below the centre of the pupil, between the eyeball and the infraorbital ridge

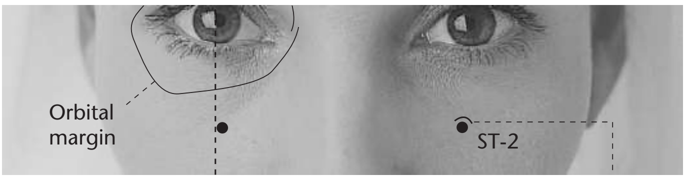

ST-2 (*sibai*) 'Four Whites' When looking straight ahead, directly below the centre of the pupil, in the depression at the infraorbital foramen

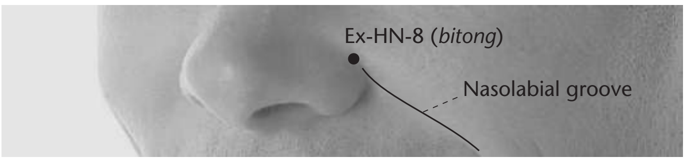

Ex-HN-8 (*shangyingxiang or bitong*) 'Upper Yingxiang' or 'Free Nose' At the upper end of the nasolabial groove, at the junction of the maxilla and the nasal cavity (at the border between the nasal bone and cartilage)

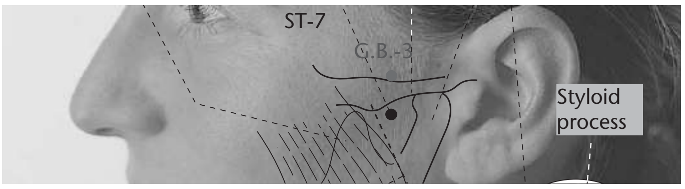

ST-7 (*xiaguan*) 'Below the Joint' With the mouth closed in the depression of the mandibular incisure, between the coronoid and condyloid process of the mandible

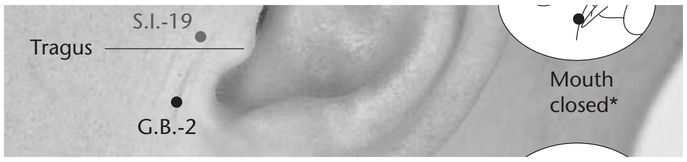

G.B.-2 (*tinghui*) 'Meeting of Hearing' Anterior to the ear, with the mouth open in the depression at the level of the intertragic notch, on the lower border of the condyloid process of the mandible.

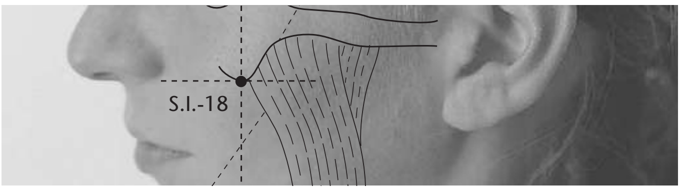

S.I.-18 (*quanliao*) 'Cheek Bone Crevice' In a depression at the intersection of a vertical line through the outer canthus of the eye with the lower border of the zygomatic bone

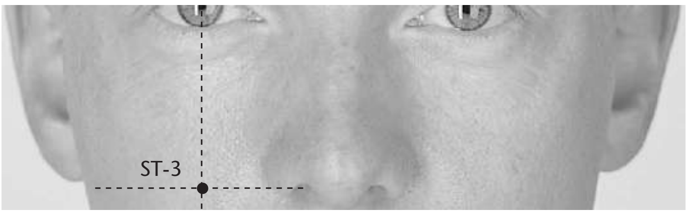

ST-3 (*juliao*) 'Great Crevice' When looking straight ahead, directly below the centre of the pupil, level with of lower border of the ala nasi.

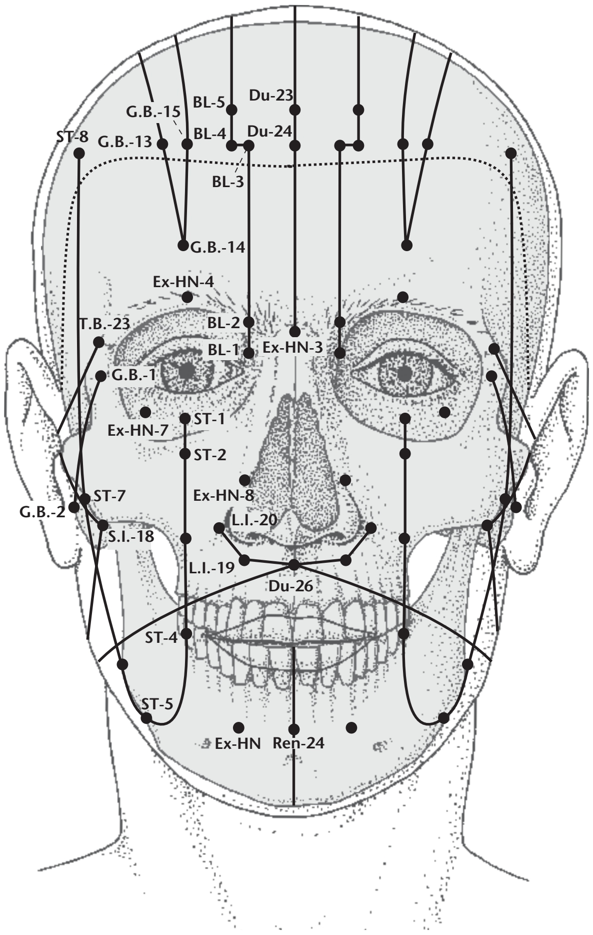
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
Ex-HN-8
ST-7
G.B.-2
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren-24

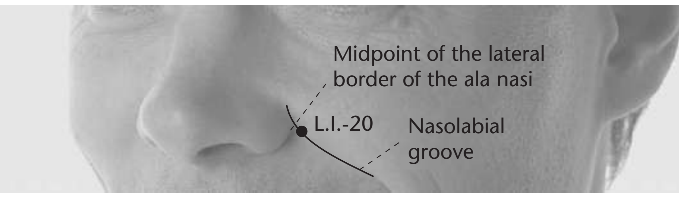
Midpoint of the lateral border of the ala nasi
L.I.-20
Nasolabial groove

L.I.-20 (*yingxiang*) 'Welcome Fragrance' In the nasolabial groove, at the level of the midpoint of the lateral border of the ala nasi.

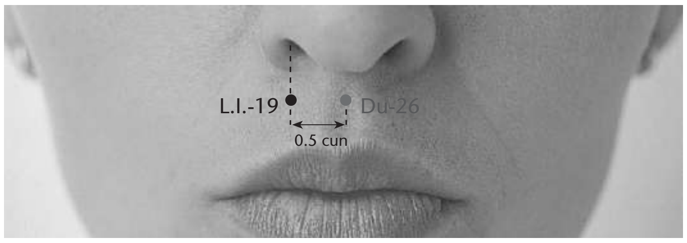
L.I.-19
Du-26
0.5 cun

L.I.-19 (*kouheliao*) 'Mouth Grain Crevice' Directly inferior to the lateral margin of the nostril, level with **Du-26**, which is located slightly superior to the midpoint of the philtrum (meeting point of the L.I. channel)

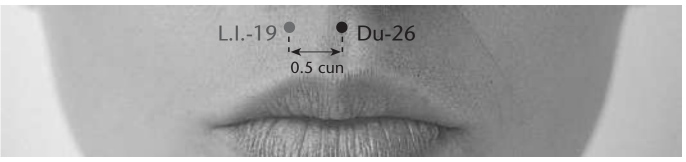
L.I.-19
Du-26
0.5 cun

Du-26 (*renzhong* or *shuigou*) 'Man's Middle' Below the nose, on the midline, at the junction of the upper and middle third of the philtrum

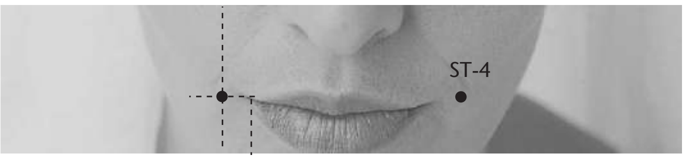

ST-4 (*dicang*) 'Earth Granary' When looking straight ahead on the pupil line, approximately 0.4 cun lateral to the corner of the mouth

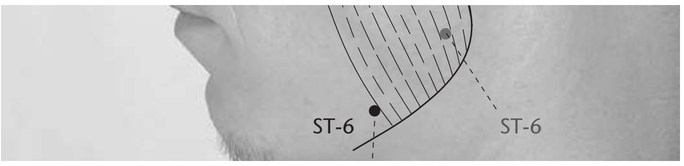

ST-5 (*daying*) 'Great Welcome' On the lateral mandible, anterior to the masseter (ask the patient to clench the teeth); a branch of the facial artery is palpable here

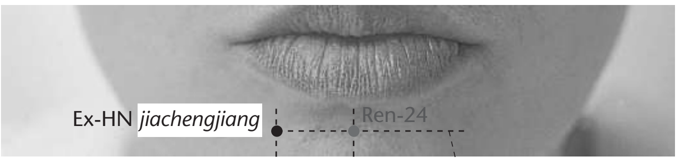
Ex-HN *jiachengjiang*
Ren-24

Ex-HN (*jiachengjiang*) 'Adjacent to Container of Fluids' At the same level and 1 cun lateral to **Ren-24** (chengjiang)

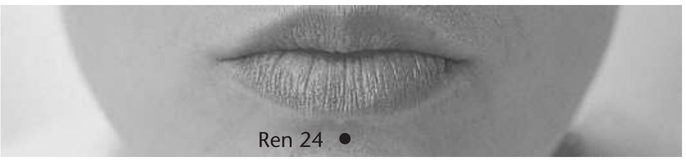

Ren-24 (*chengjiang*) 'Container of Fluids' On the midline, in the mentolabial groove below the lower lip

7.2 Head and Lateral Aspect of the Neck

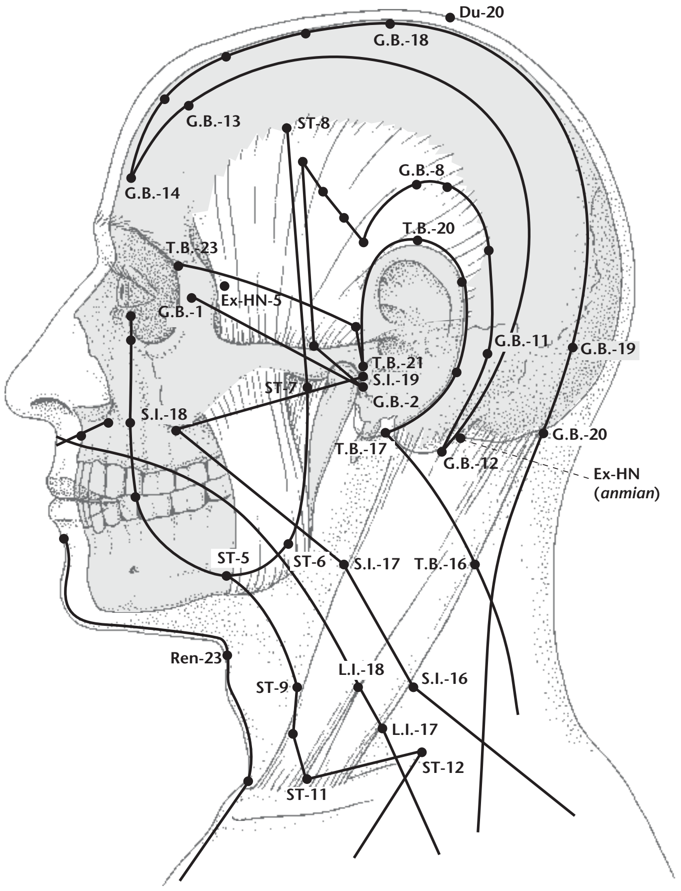
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
ST-7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(anmian)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
S.I.-16
ST-9
L.I.-17
ST-12
ST-11

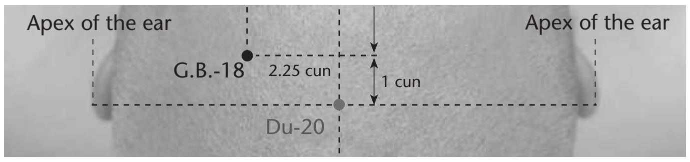
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
Du-20

G.B.-18 (*chengling*) 'Support Spirit' 4 cun superior to the anterior hairline or 1 cun anterior to **Du-20**, on the line connecting **G.B.-15** and **G.B.-20**, or 2.25 cun lateral to the midline (midpoint of the line connecting **ST-8** and **Du-24**)

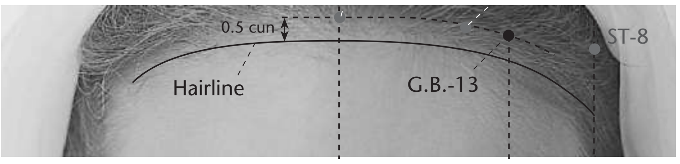
0.5 cun
ST-8
Hairline
G.B.-13

G.B.-13 (*benshen*) 'Root of the Spirit' 0.5 cun superior to the anterior hairline and 3 cun lateral to the midline, on the line between **Du-24** and **ST-8** (= 4.5 cun) at the junction of the lateral and medial third

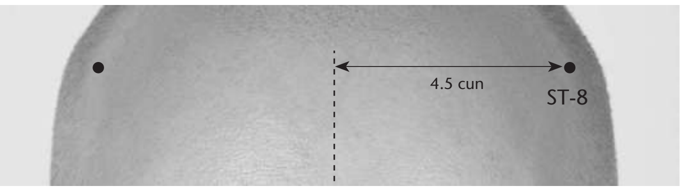

ST-8 (*touwei*) 'Head's Binding' At the temporal corner of the forehead, on the border of the temporalis muscle and 0.5 cun within the anterior hair line or 4.5 cun lateral to the anterior midline (**Du-24**)

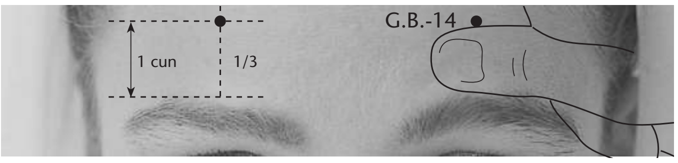
G.B.-14
1 cun
1/3

G.B.-14 (*yangbai*) 'Yang White' When looking straight ahead on the pupil line, approximately 1 cun superior to the midpoint of the eyebrow (centre of the eyebrow to the anterior hairline = 3 cun)

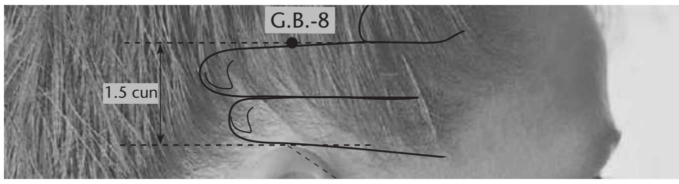

G.B.-8 (*shuaigu*) 'Leading Valley' 1.5 cun directly superior to the apex of the ear

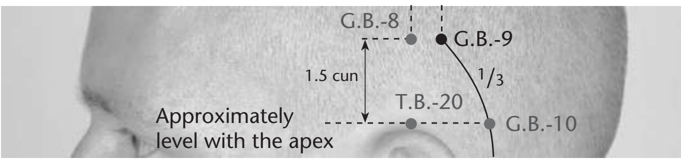
G.B.-8
G.B.-9
1.5 cun
1/3
T.B.-20
G.B.-10
Approximately
level with the apex

G.B.-9 (*tianchong*) 'Heavenly Rushing' 1.5 cun superior and 0.5 cun posterior to the apex of the ear or 0.5 cun posterior to **G.B.-8**

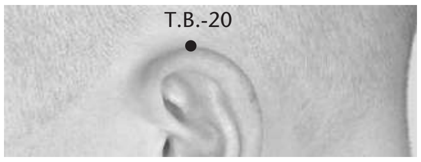

T.B.-20 (*jiaosun*) 'Minute Angle' Directly superior to the apex of the ear

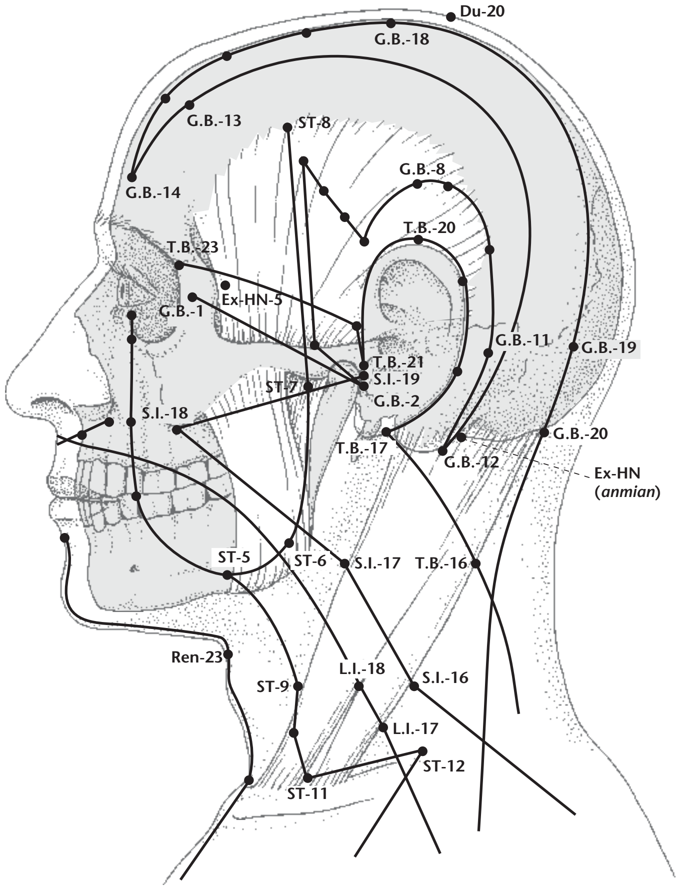
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
ST-7
G.B.-2
S.I.-18
T.B.-17
G.B.-20
G.B.-12
Ex-HN
(anmian)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
S.I.-16
ST-9
L.I.-17
ST-12
ST-11

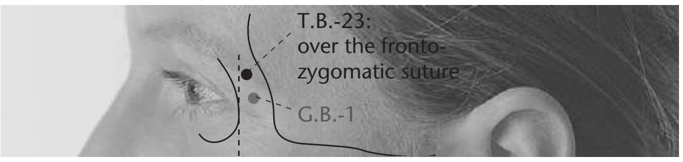
T.B.-23:
over the fronto-
zygomatic suture
G.B.-1

T.B.-23 (*sizhukong*) 'Silken Bamboo Hollow' In the depression (frontozygomatic suture) at the lateral end of the eyebrow

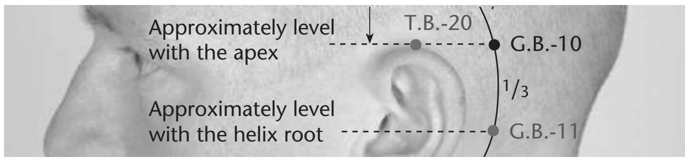
Approximately level
with the apex
T.B.-20
G.B.-10
1/3
Approximately level
with the helix root
G.B.-11

G.B.-10 (*fubai*) 'Floating White' At the junction of the upper third with the lower two-thirds of the curved line connecting **G.B.-9** and **G.B.-12**

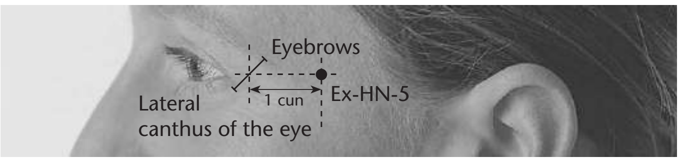
Eyebrows
Lateral
canthus of the eye
1 cun
Ex-HN-5

Ex-HN-5 (*taiyang*) 'Supreme Yang' In the depression approximately 1 cun posterior to the midpoint between the lateral end of the eyebrow and the outer canthus of the eye

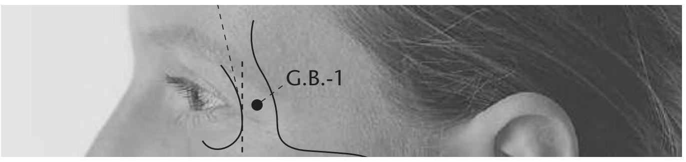

G.B.-1 (*tongziliao*) 'Pupil Crevice' 0.5 cun lateral to the outer canthus of the eye

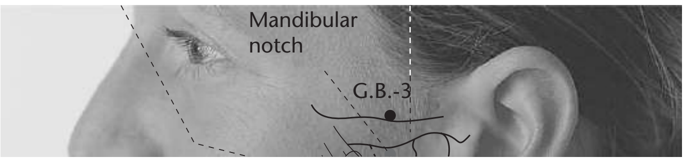
Mandibular
notch
G.B.-3

G.B.-3 (*shangguan*) 'Above the Joint' In a depression on the upper border of the zygomatic arch, approximately 1 cun anterior to the root of the ear, directly superior to **ST-7** (lower border of the zygomatic arch)

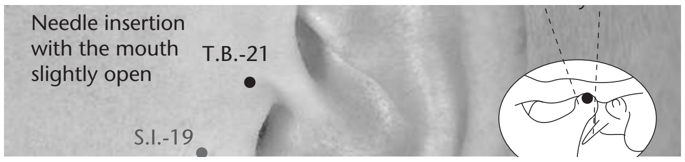
Needle insertion
with the mouth
slightly open
T.B.-21
S.I.-19

T.B.-21 (*ermen*) 'Ear Gate' Anterior to the ear, with the mouth open in the depression at the level of the supratragic notch, slightly superior to the condyloid process of the mandible

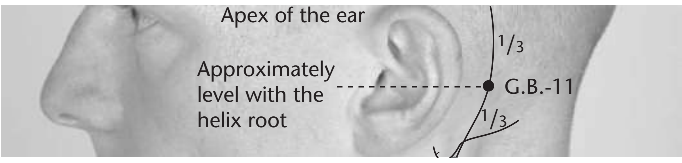
Apex of the ear
1/3
Approximately
level with the
helix root
G.B.-11
1/3

G.B.-11 (*touqiaoyin*) 'Yin Portals of the Head' At the junction of the lower third with the upper two-thirds of the curved line connecting **G.B.-9** and **G.B.-12**

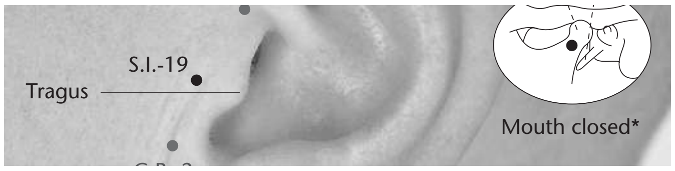
S.I.-19
Tragus
Mouth closed*

S.I.-19 (*tinggong*) 'Palace of Hearing' Anterior to the ear with the mouth open, at the level of the middle of the tragus, in a depression between the tragus and condyloid process of the mandible

---

* According to Deadman et al 1998.

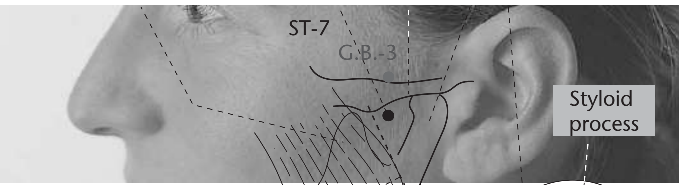
ST-7
G.B.-3
Styloid
process

ST-7 (*xiaguan*) 'Below the Joint' With the mouth closed in the depression of the mandibular notch, between the coronoid and condyloid process of the mandible

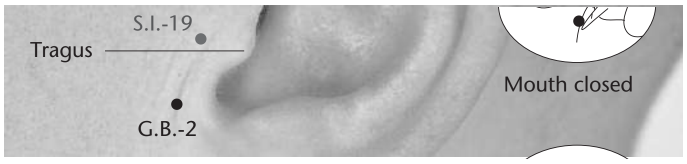
S.I.-19
Tragus
G.B.-2
Mouth closed

G.B.-2 (*tinghui*) 'Meeting of Hearing' Anterior to the ear with the mouth open, in the depression at the level of the intertragic notch, on the lower border of the condyloid process of the mandible

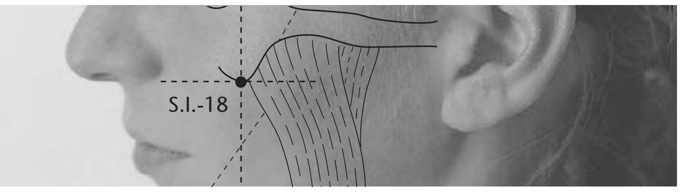

S.I.-18 (*quanliao*) 'Cheek Bone Crevice' In a depression at the intersection of a vertical line through the outer canthus of the eye with the lower border of the zygomatic bone

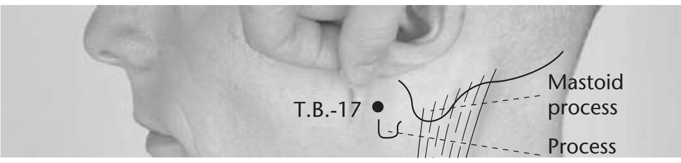
T.B.-17
Mastoid
process
Process

T.B.-17 (*yifeng*) 'Wind Screen' With the mouth open in the depression behind the earlobe, between the mastoid process and the mandible

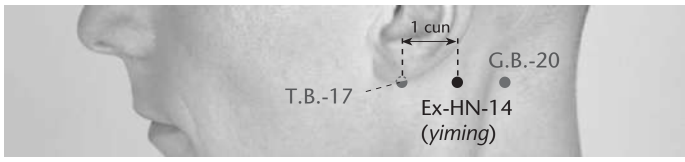
1 cun
G.B.-20
T.B.-17
Ex-HN-14
(yiming)

Ex-HN-14 (*yiming*) 'Eye Brightening' On the occiput, 1 cun posterior to **T.B.-17**, on a line connecting **T.B.-17** and **G.B.-20**

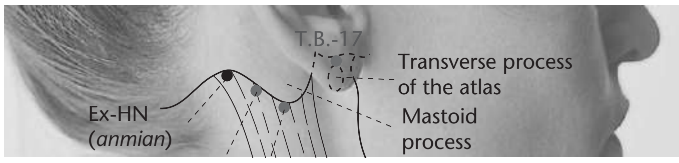
T.B.-17
Transverse process
of the atlas
Ex-HN
(anmian)
Mastoid
process

Ex-HN (*anmian*) 'Peaceful Sleep' Posterior to the ear, midway between **T.B.-17** and **G.B.-20**, posterior to the mastoid process

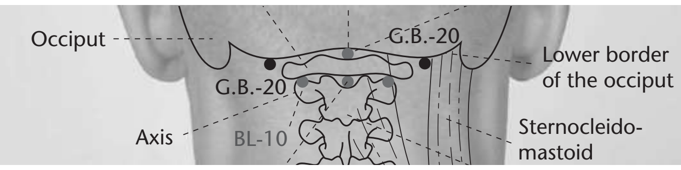
Occiput
G.B.-20
Lower border
of the occiput
G.B.-20
Axis
BL-10
Sternocleido-
mastoid

G.B.-20 (*fengchi*) 'Wind Pool' On the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius

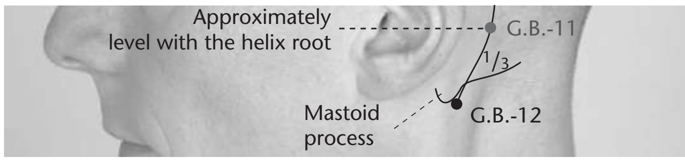
Approximately
level with the helix root
G.B.-11
1/3
Mastoid
process
G.B.-12

G.B.-12 (*wangu*) 'Mastoid Process' In a depression posterior and inferior to the mastoid process

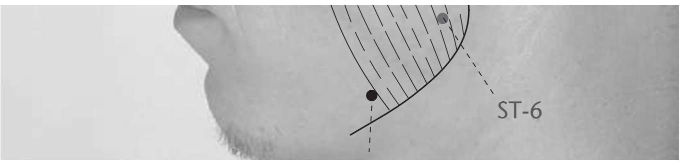

ST-6 (*jiache*) 'Jaw Bone' With the teeth clenched, on the highest prominence of the masseter, approximately one fingerbreadth anterior and superior to the mandibular angle

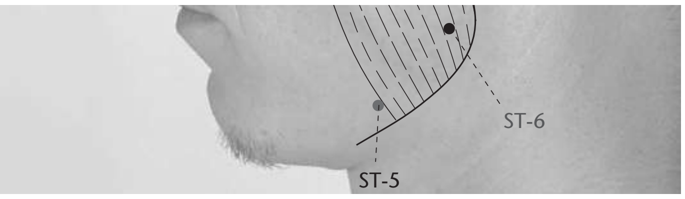

ST-5 (*daying*) 'Great Welcome' On the lateral mandible, anterior to the masseter (ask the patient to clench the teeth); a branch of the facial artery is palpable here

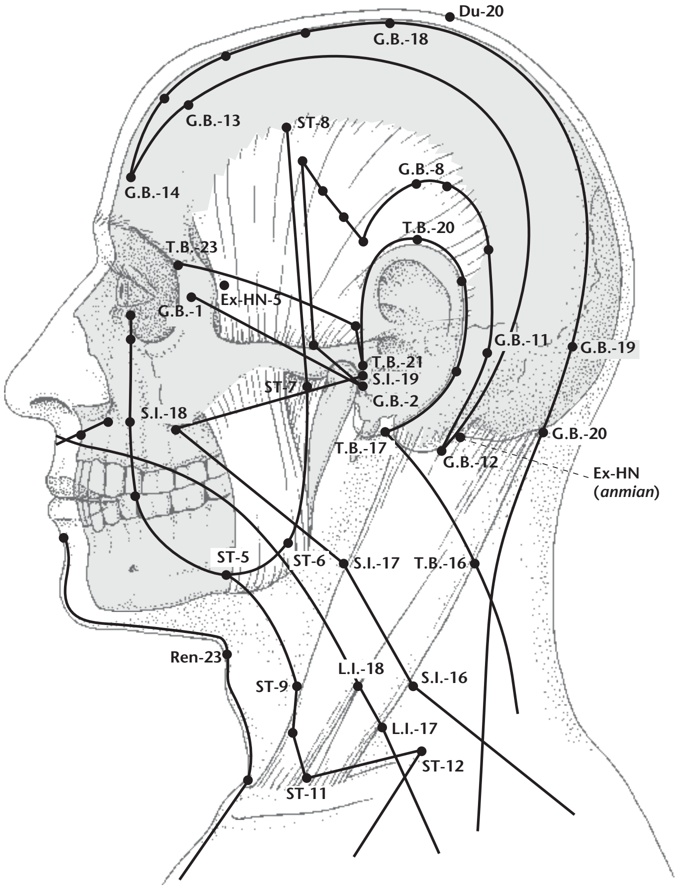
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
ST-7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
S.I.-16
ST-9
L.I.-17
ST-12
ST-11

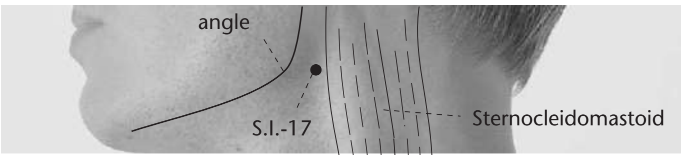
angle
S.I.-17
Sternocleidomastoid

S.I.-17 (*tianrong*) 'Heavenly Appearance' Level with and posterior to the mandibular angle, on the anterior border of the sternocleidomastoid

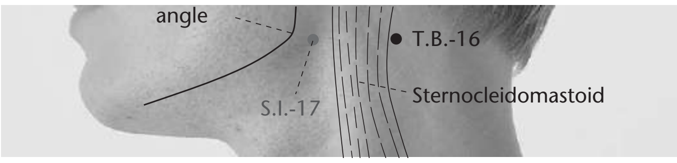
angle
T.B.-16
S.I.-17
Sternocleidomastoid

T.B.-16 (*tianyou*) 'Window of Heaven' posterior and inferior to the mastoid process, on the posterior border of the sternocleidomastoid, directly level with the mandibular angle

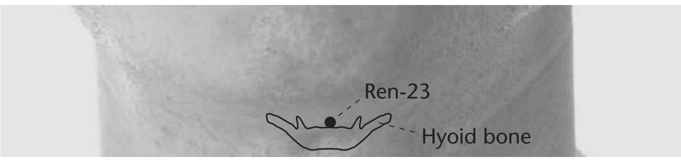
Ren-23
Hyoid bone

Ren-23 (*lianquan*) 'Corner Spring' On the anterior midline, superior to the upper border of the hyoid bone

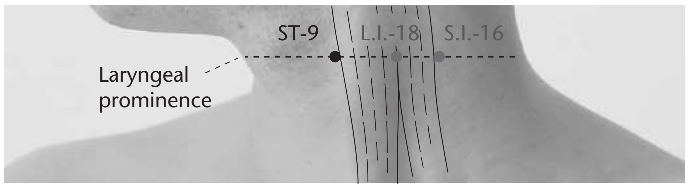
ST-9
L.I.-18
S.I.-16
Laryngeal
prominence

ST-9 (*renying*) Man's Welcome' Approximately 1.5 cun lateral to the anterior midline, level with the laryngeal prominence, at the anterior border of the sternocleidomastoid

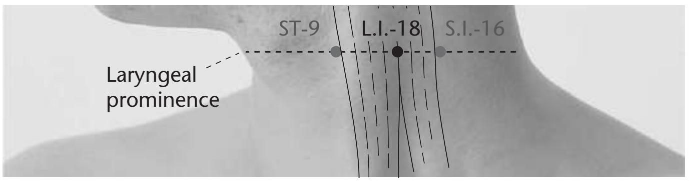
ST-9
L.I.-18
S.I.-16
Laryngeal
prominence

L.I.-18 (*futu*) 'Support the Prominence' At the level of the laryngeal prominence, between the sternal and clavicular heads of the sternocleidomastoid muscle

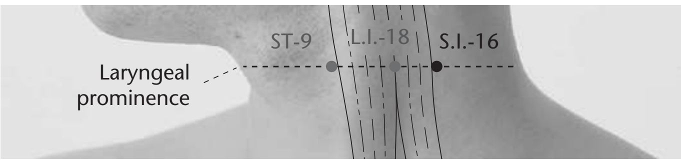
ST-9
L.I.-18
S.I.-16
Laryngeal
prominence

S.I.-16 (*tianchuang*) 'Heavenly Window' Approximately 3.5 cun lateral to the anterior midline, level with the laryngeal prominence, on the posterior border of the sternocleidomastoid

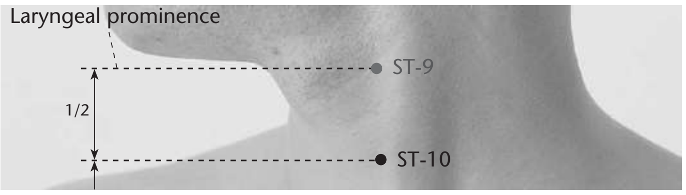
Laryngeal prominence
ST-9
1/2
ST-10

ST-10 (*shuitu*) 'Water Prominence' On the anterior border of the sternocleidomastoid, at the midpoint of a line connecting **ST-9** and **ST-11**

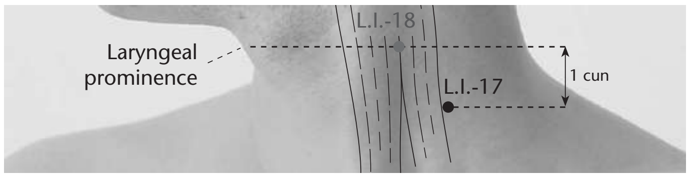
L.I.-18
Laryngeal prominence
L.I.-17
1 cun

L.I.-17 (*tianding*) 'Heaven's Tripod' 1 cun inferior to the laryngeal prominence, on the posterior border of the sternocleidomastoid

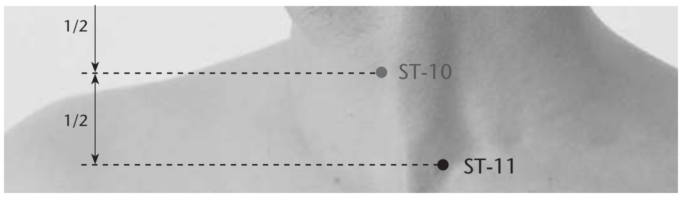
1/2
ST-10
1/2
ST-11

ST-11 (*qishe*) 'Abode of Qi' On the upper border of the clavicle, between the tendons of the sternal and clavicular heads of the sternocleidomastoid

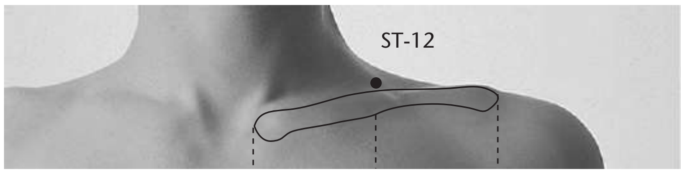

ST-12 (*quepen*) 'Empty Basin' In the supraclavicular fossa, superior to the midpoint of the clavicle, approximately 4 cun lateral to the midline

7.3 Occipital Region

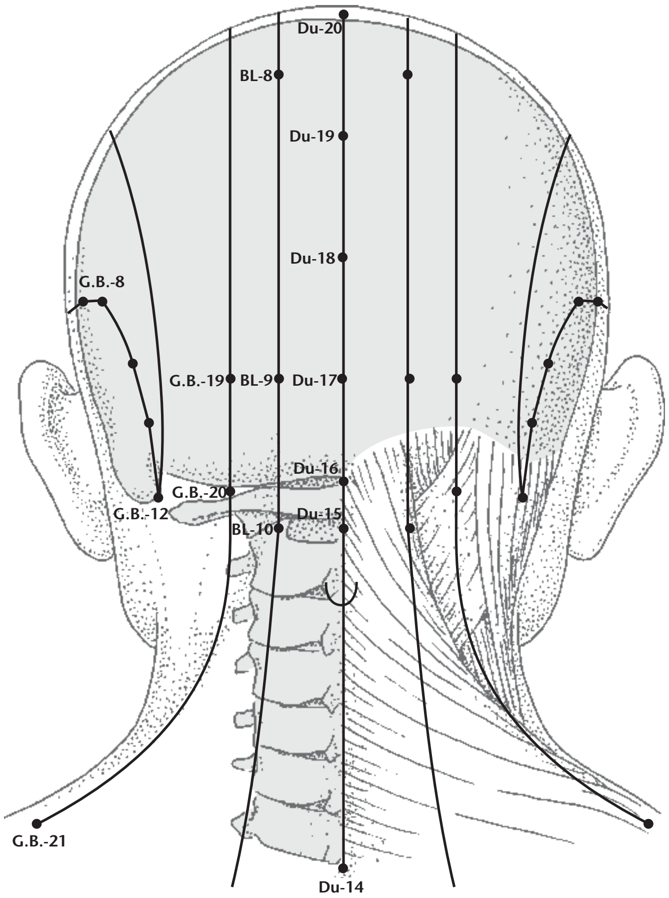
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
Du-15
BL-10
G.B.-21
Du-14

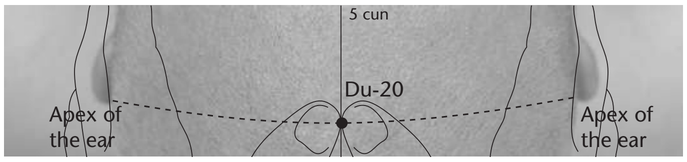
5 cun
Du-20
Apex of the ear
Apex of the ear

Du-20 (*baihui*) 'Hundred Meetings' At the junction of a line connecting the apices of the ears and the midline, 5 cun from the anterior or 7 cun from the posterior hairline respectively

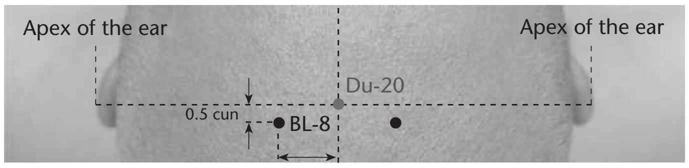
Apex of the ear
Apex of the ear
Du-20
0.5 cun
BL-8

BL-8 (*luoque*) 'Declining Connection' 1.5 cun lateral to the midline and 5.5 cun superior to the anterior hairline, or 0.5 cun posterior to **Du-20**

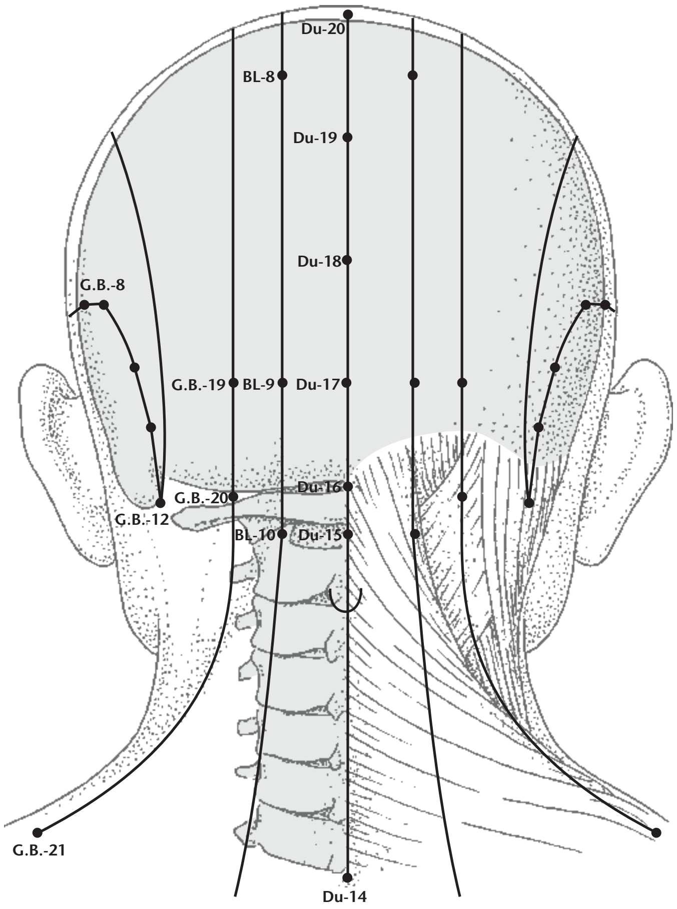
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15
G.B.-21
Du-14

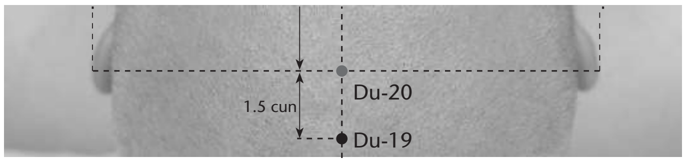
1.5 cun
Du-20
Du-19

Du-19 (*houding*) 'Behind the Crown' On the midline, 1.5 cun posterior to **Du-20** or 3 cun superior to **Du-17**

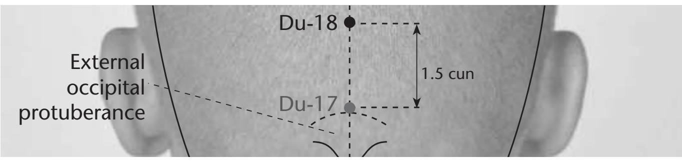
Du-18
1.5 cun
External occipital protuberance
Du-17

Du-18 (*qiangjian*) 'Unyielding Space' On the posterior midline, 1.5 cun superior to **Du-17** or 3 cun posterior to **Du-20**

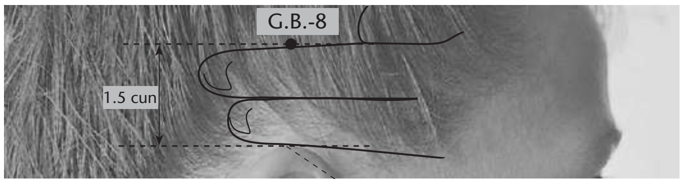

G.B.-8 (*shuaigu*) 'Leading Valley' 1.5 cun directly superior to the apex of the ear

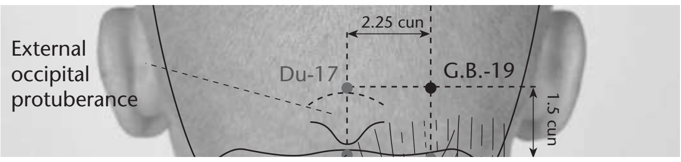
External occipital protuberance
2.25 cun
Du-17
G.B.-19
1.5 cun

G.B.-19 (*naokong*) 'Brain Hollow' Level with the upper border of the external occipital protuberance (**Du-17**), on the line connecting **G.B.-15** and **G.B.-20**, and 2.25 cun lateral to the midline

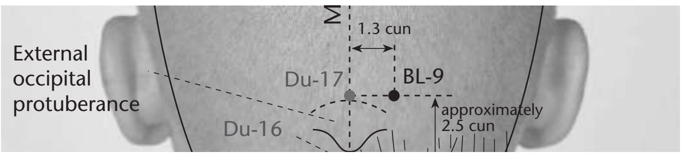
External occipital protuberance
1.3 cun
Du-17
BL-9
Du-16
approximately 2.5 cun

BL-9 (*yuzhen*) 'Jade Pillow' 2.5 cun superior to the posterior hairline and 1.3 cun lateral to the midline or **Du-17** (directly superior to the external occipital protuberance)

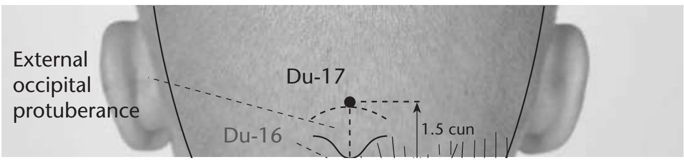
External occipital protuberance
Du-17
Du-16
1.5 cun

Du-17 (*naohu*) 'Brain's Door' On the midline, in a depression superior to the external occipital protuberance, approximately 2.5 cun superior to the posterior hairline, 1.5 cun superior to **Du-16**

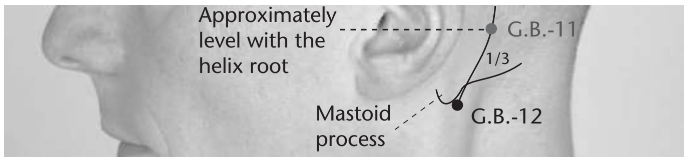
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

G.B.-12 (*wangu*) 'Mastoid Process' In a depression posterior and inferior to the mastoid process

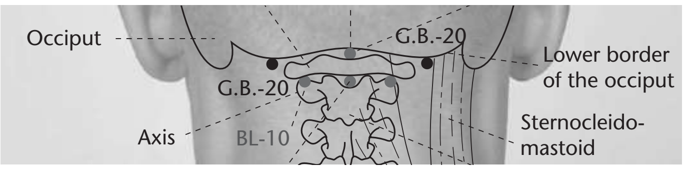
Occiput
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid

G.B.-20 (*fengchi*) 'Wind Pool' On the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius

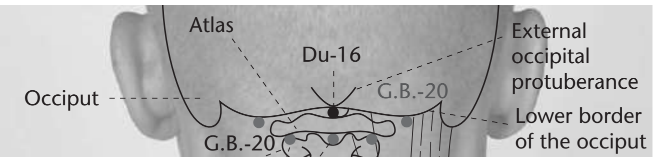
Atlas
Du-16
External occipital protuberance
G.B.-20
Occiput
Lower border of the occiput
G.B.-20

Du-16 (*fengfu*) 'Palace of Wind' On the posterior midline, directly below the external occipital protuberance, in the depression between the origins of the trapezius

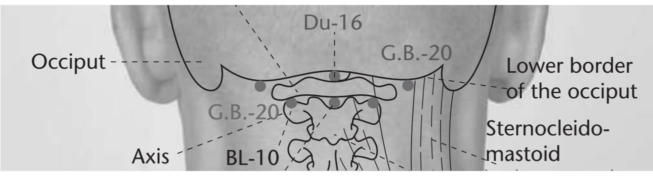
Du-16
G.B.-20
Occiput
Lower border of the occiput
G.B.-20
Sternocleido-mastoid
Axis
BL-10

BL-10 (*tianzhu*) 'Celestial Pillar' Below the lower border of the occiput, approximately 1.3 cun lateral to the midline (**Du-15**), at the insertion of the trapezius

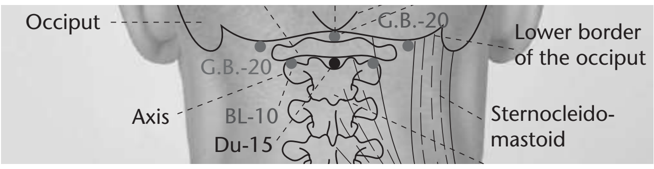
Occiput
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Du-15
Sternocleido-mastoid

Du-15 (*yamen*) 'Gate of Muteness' On the posterior midline, in the depression between the 1st (atlas) and 2nd cervical vertebrae (axis = first palpable spinous process), approximately 0.5 cun below **Du-16**

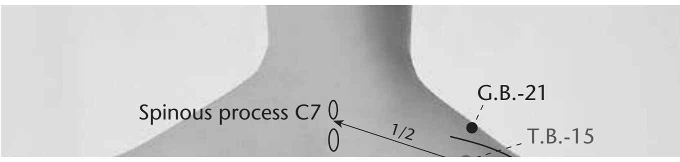
G.B.-21
Spinous process C7
1/2
T.B.-15

G.B.-21 (*jianjing*) 'Shoulder Well' At the highest point of the shoulder, midway between the lower border of the spinous process of C7 and the lateral extremity of the acromion

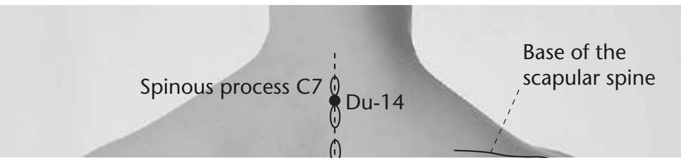
Base of the scapular spine
Spinous process C7
Du-14

Du-14 (*dazhui*) 'Great Vertebra' Below the spinous process of C7

7.4 Vertex

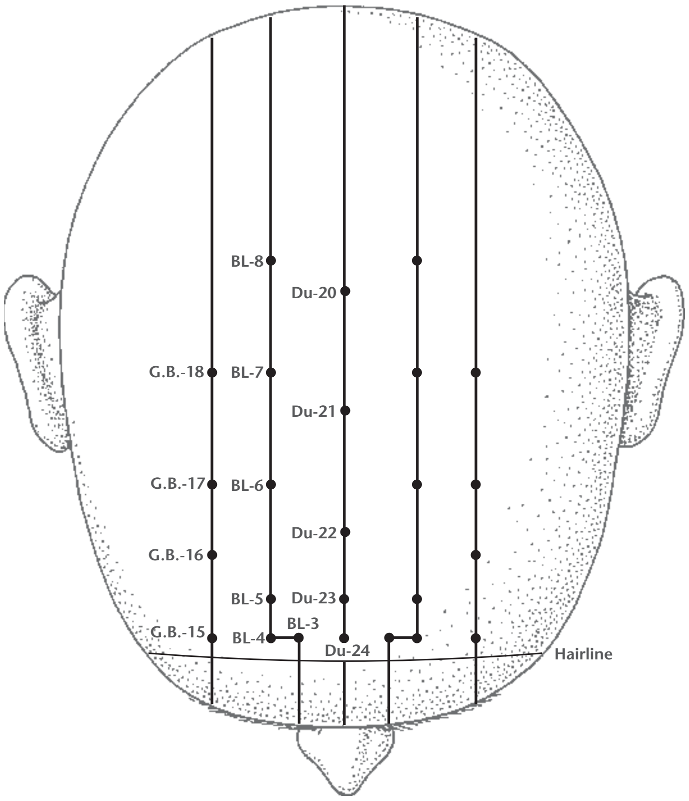
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

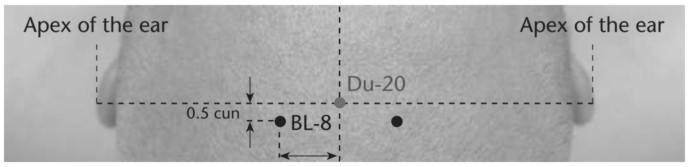
Apex of the ear
Apex of the ear
Du-20
0.5 cun
BL-8

BL-8 (*luoque*) 'Declining Connection' 1.5 cun lateral to the midline and 5.5 cun superior to the anterior hairline, or 0.5 cun posterior to **Du-20**

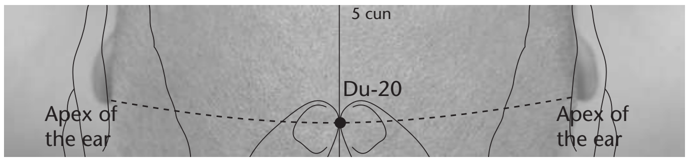
5 cun
Du-20
Apex of the ear
Apex of the ear

Du-20 (*baihui*) 'Hundred Meetings' At the junction of a line connecting the apices of the ears and the midline, 5 cun from the anterior or 7 cun from the posterior hairline respectively

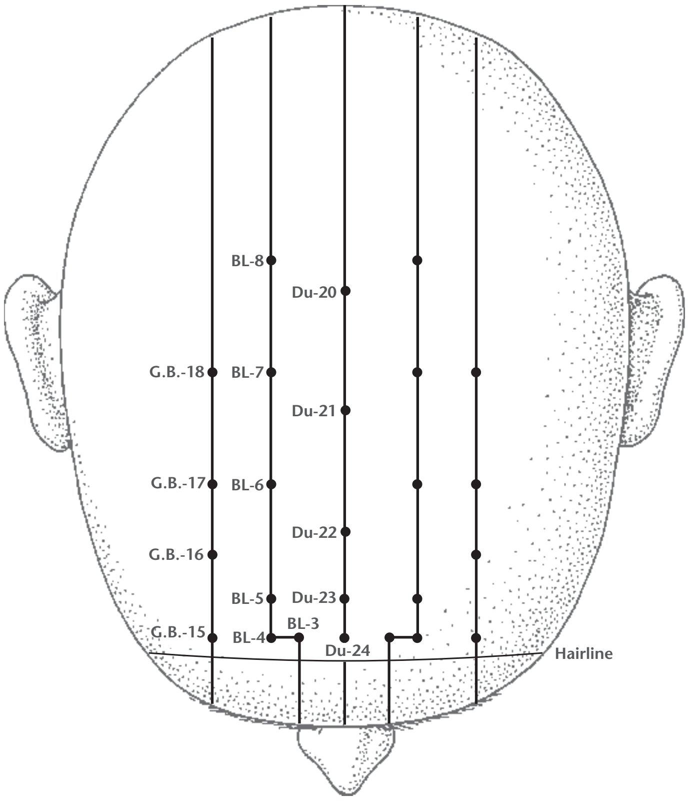
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

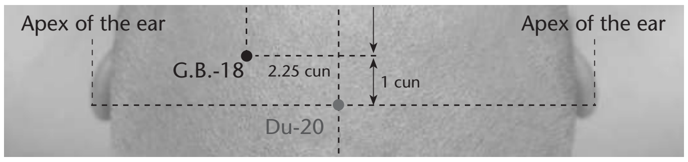
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
Du-20

G.B.-18 (*chengling*) 'Support Spirit' 4 cun superior to the anterior hairline or 1 cun anterior to **Du-20**, on the line connecting **G.B.-15** with **G.B.-20**, or 2.25 cun lateral to the midline (midpoint of the line connecting **ST-8** and **Du-24**)

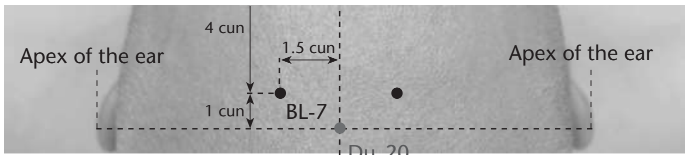
4 cun
1.5 cun
Apex of the ear
Apex of the ear
1 cun
BL-7
Du-20

BL-7 (*tongtian*) 'Heavenly Connection' 1.5 cun lateral to the midline and 4 cun superior to the anterior hairline or 1 cun anterior to **Du-20**

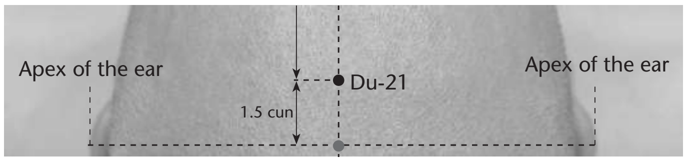
Apex of the ear
Apex of the ear
Du-21
1.5 cun

Du-21 (*qianding*) 'In Front of the Crown' On the midline, 1.5 cun anterior to **Du-20** or 3.5 cun above the anterior hairline

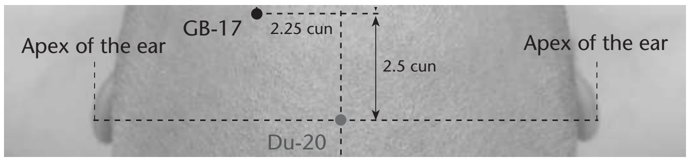
GB-17
2.25 cun
Apex of the ear
Apex of the ear
2.5 cun
Du-20

G.B.-17 (*zhengying*) 'Upright Nutrition' 2.5 cun superior to the anterior hairline, on the pupil line, on the line connecting **G.B.-15** and **G.B.-20** (= 2.25 cun lateral to the midline or midway between **Du-23** and **ST-8**)

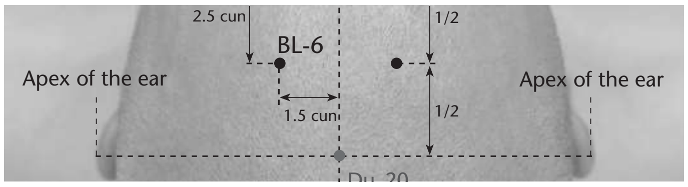
2.5 cun
1/2
BL-6
Apex of the ear
Apex of the ear
1.5 cun
1/2
Du-20

BL-6 (*chengguang*) 'Receiving Light' 2.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline, or one-third of the distance between **Du-24** and **ST-8**

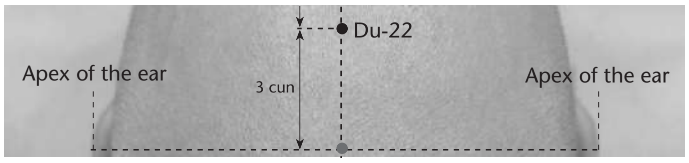
Du-22
Apex of the ear
Apex of the ear
3 cun

Du-22 (*xinhui*) 'Fontanelle Meeting' On the midline, 2 cun superior to the anterior hairline or 3 cun anterior to **Du-20**

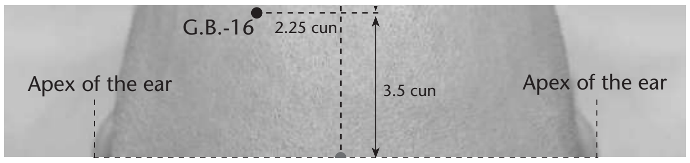
G.B.-16
2.25 cun
Apex of the ear
3.5 cun
Apex of the ear

G.B.-16 (*muchuang*) 'Window of the Eye' 1.5 cun superior to the anterior hairline, on the pupil line, on the line connecting **G.B.-15** and **G.B.-20** (= 2.25 cun lateral to the midline or midway between **Du-24** and **ST-8**)

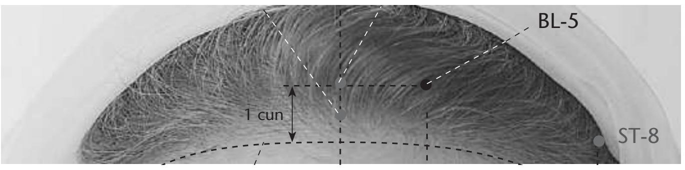
BL-5
1 cun
ST-8

BL-5 (*wuchu*) 'Fifth Place' 1 cun superior to the anterior hairline and 1.5 cun lateral to the midline (directly above **BL-4**, level with **Du-23**)

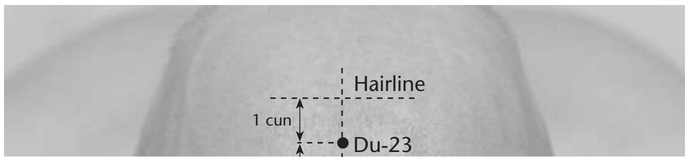
Hairline
1 cun
Du-23

Du-23 (*shangxing*) 'Upper Star' On the midline, 1 cun superior to the anterior hairline

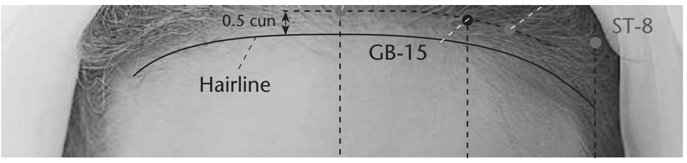
0.5 cun
ST-8
GB-15
Hairline

G.B.-15 (*toulinqi*) 'Head Governor of Tears' On the pupil line when looking straight ahead, 0.5 cun superior to the anterior hairline, at the midpoint of the line between **Du-24** and **ST-8** (= 4.5 cun)

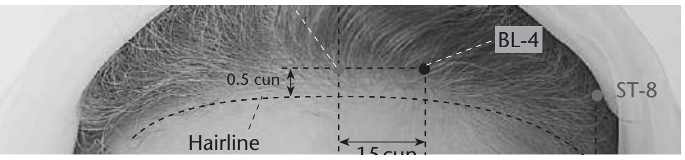
BL-4
0.5 cun
ST-8
Hairline
1.5 cun

BL-4 (*qucha*) 'Crooked Curve' 0.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline, or one-third of the distance between **Du-24** and **ST-8** (= 4.5 cun)

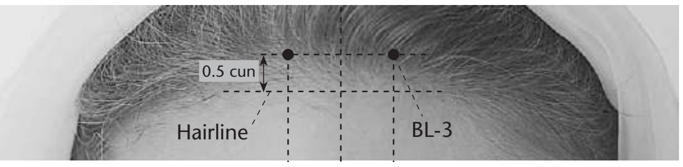
0.5 cun
Hairline
BL-3

BL-3 (*meichong*) 'Eyebrows' Pouring' 0.5 cun within the anterior hairline, superior to the medial canthus of the eye (**BL-1**)

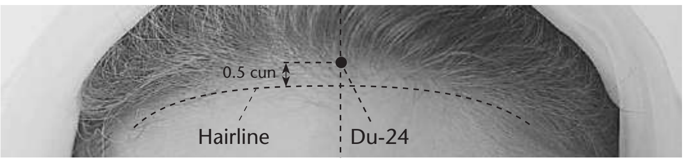
0.5 cun
Hairline
Du-24

Du-24 (*shenting*) 'Courtyard of the Spirit' On the midline, 0.5 cun superior to the anterior hairline

7.5 Back

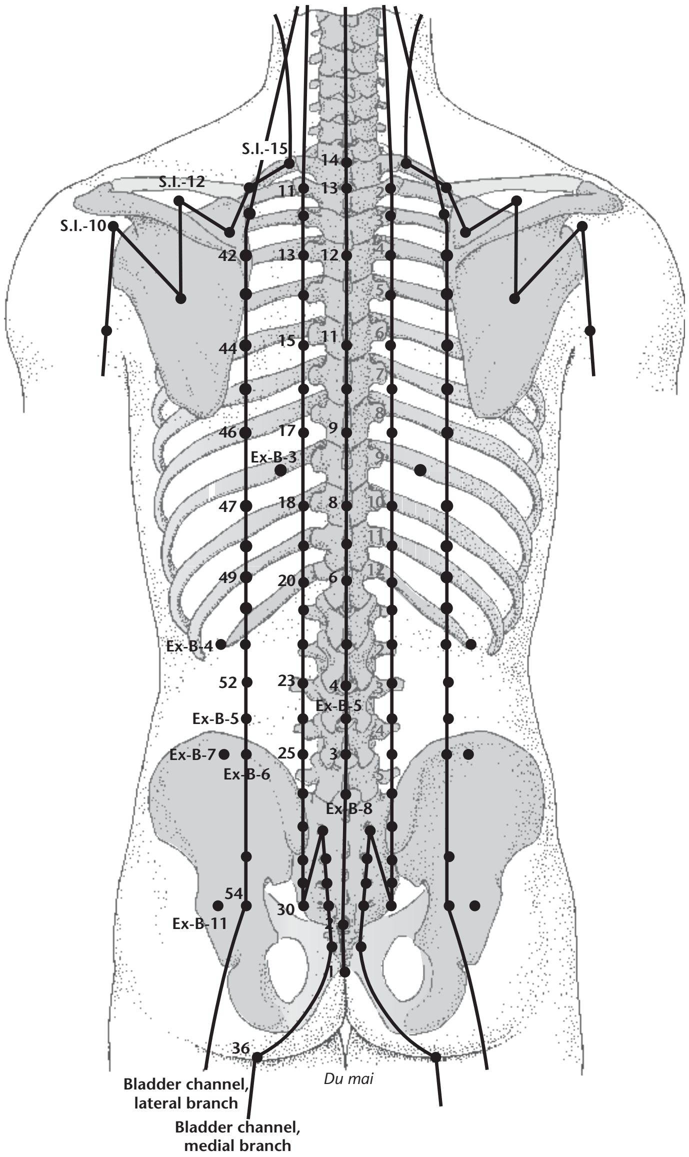
S.I.-15
S.I.-12
S.I.-10
14
11
13
42
13
12
44
15
11
46
17
9
Ex-B-3
47
18
8
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

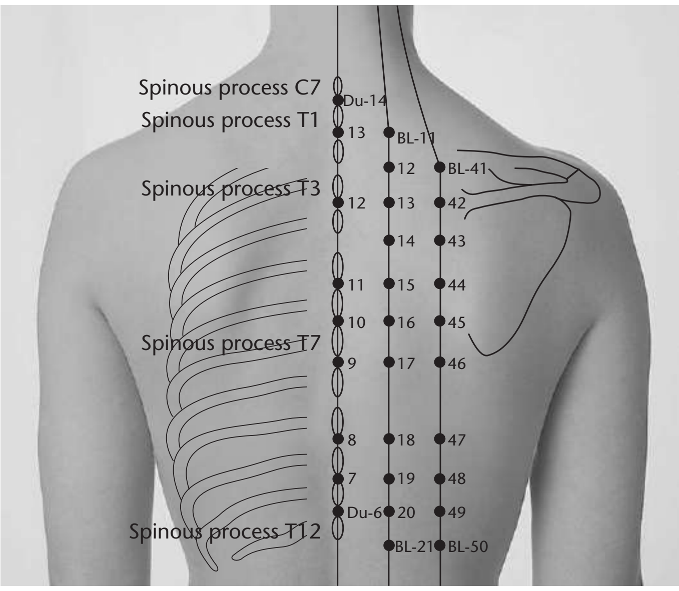
Spinous process C7
Spinous process T1
Spinous process T3
Spinous process T7
Spinous process T12
Du-14
13
12
11
10
9
8
7
Du-6
BL-11
12
13
14
15
16
17
18
19
20
BL-21
BL-41
42
43
44
45
46
47
48
49
BL-50

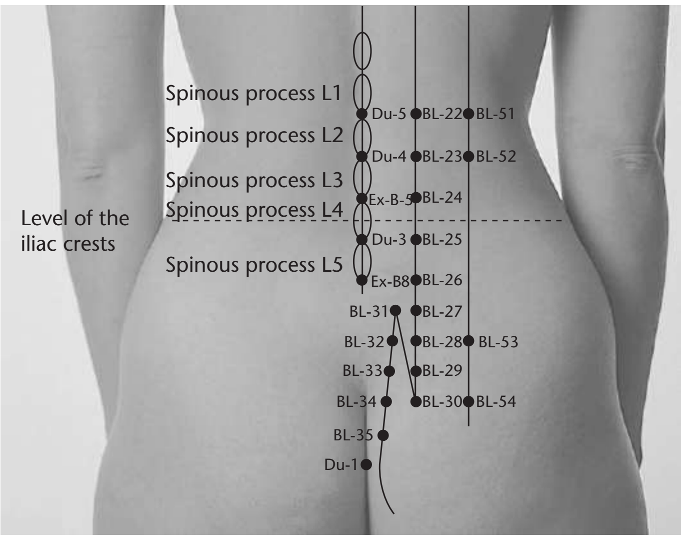
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
Spinous process L5
Level of the iliac crests
Du-5
Du-4
Ex-B-5
Du-3
Ex-B8
BL-22
BL-23
BL-24
BL-25
BL-26
BL-27
BL-28
BL-29
BL-30
BL-51
BL-52
BL-53
BL-54
BL-31
BL-32
BL-33
BL-34
BL-35
Du-1

Vertebra	**Below the spinous process or in the sacral foramen**	**0.5 cun lateral to the midline, level with the lower border of the spinous process or sacral foramen**	**1.5 cun lateral to the midline, level with the lower border of the spinous process or sacral foramen**	**3 cun lateral to the midline, level with the lower border of the spinous process or sacral foramen**
C7	Du-14	Ex-B-1 *(dingchuan)*	**2 cun lateral** S.I.-15	**3.5 cun lateral** Ex-B (*jiehexue*)
T1	Du-13	A point of Ex-B-2 (*huatuojiaji*)	BL-11	S.I.-14
T2		A point of Ex-B-2 (*huatuojiaji*)	BL-12	BL-41
T3	Du-12	A point of Ex-B-2 (*huatuojiaji*)	BL-13	BL-42
T4		A point of Ex-B-2 (*huatuojiaji*)	BL-14	BL-43
T5	Du-11	A point of Ex-B-2 (*huatuojiaji*)	BL-15	BL-44
T6	Du-10	A point of Ex-B-2 (*huatuojiaji*)	BL-16	BL-45
T7	Du-9	A point of Ex-B-2 (*huatuojiaji*)	BL-17	BL-46
T8		A point of Ex-B-2 (*huatuojiaji*)	Ex-B-3 (*weiwanxiashu*)	
T9	Du-8	A point of Ex-B-2 (*huatuojiaji*)	BL-18	BL-47
T10	Du-7	A point of Ex-B-2 (*huatuojiaji*)	BL-19	BL-48
T11	Du-6	A point of Ex-B-2 (*huatuojiaji*)	BL-20	BL-49
T12		A point of Ex-B-2 (*huatuojiaji*)	BL-21	BL-50
L1	Du-5	A point of Ex-B-2 (*huatuojiaji*)	BL-22	BL-51/ Ex-B-4 (**3.5 cun lateral**)
L2	Du-4	A point of Ex-B-2 (*huatuojiaji*)	BL-23	BL-52
L3	Ex-B-5 (*xiajishu*)	A point of Ex-B-2 (*huatuojiaji*)	BL-24	Ex-B-5 (according to WHO a set of 3 points)
L4	Du-3	A point of Ex-B-2 (*huatuojiaji*)	BL-25	Ex-B-6/Ex-B-7 (**3.5 cun lateral**)
L5	Ex-B-8	A point of Ex-B-2 (*huatuojiaji*)	BL-26	
1st sacral foramen	BL-31		BL 27	
2nd sacral foramen	BL-32		BL 28	BL-53
3rd sacral foramen	BL-33		BL 29	
4th sacral foramen	BL-34		BL 30	BL-54/Ex-B-11 (3.5 cun lateral)
Sacral hiatus	Du- 2			
		BL-35 (0.5 cun lateral and inferior to Du-2)		
Between the coccyx and the anus	Du-1			

7.6 Frontal and Lateral Thorax and Abdomen

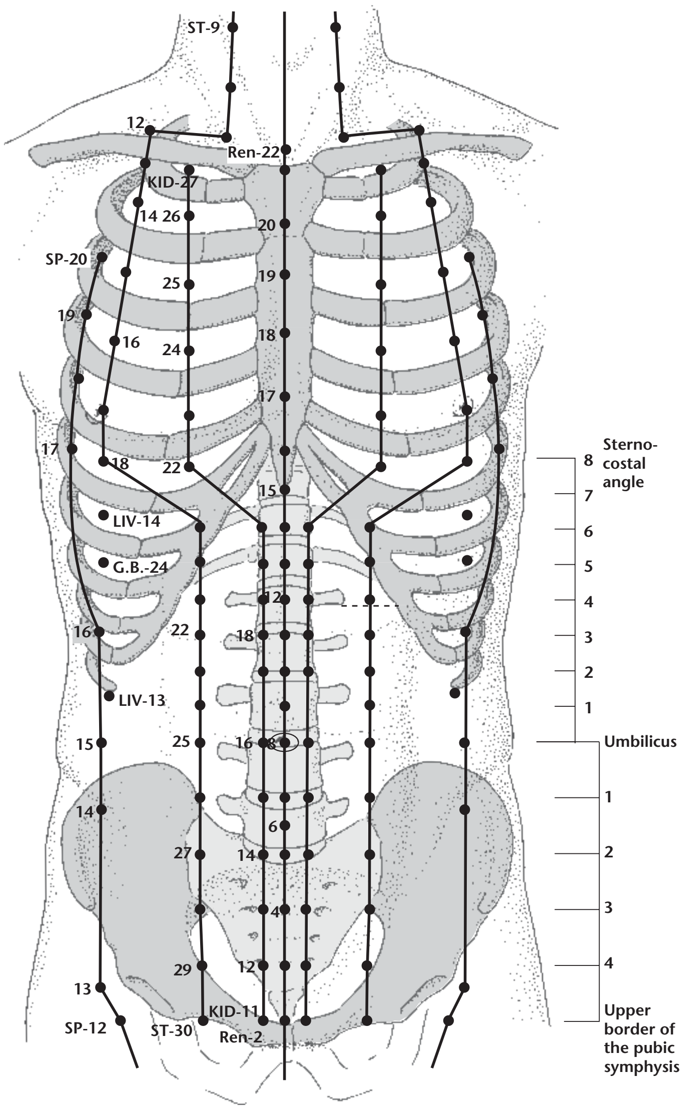
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17
17
18
22
15
8
Sterno-costal angle
7
6
5
4
3
2
1
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Umbilicus
1
14
27
6
14
2
4
3
29
12
4
13
KID-11
SP-12
ST-30
Ren-2
Upper border of the pubic symphysis

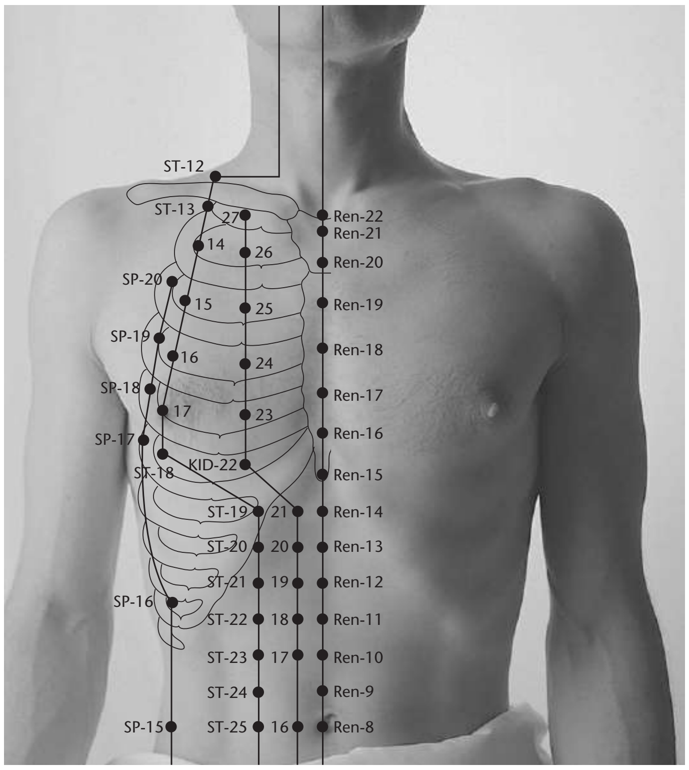
ST-12
ST-13
27
14
26
Ren-22
Ren-21
Ren-20
SP-20
15
25
Ren-19
SP-19
16
24
Ren-18
SP-18
17
23
Ren-17
SP-17
Ren-16
ST-18
KID-22
Ren-15
ST-19
21
Ren-14
ST-20
20
Ren-13
ST-21
19
Ren-12
SP-16
ST-22
18
Ren-11
ST-23
17
Ren-10
ST-24
Ren-9
SP-15
ST-25
16
Ren-8

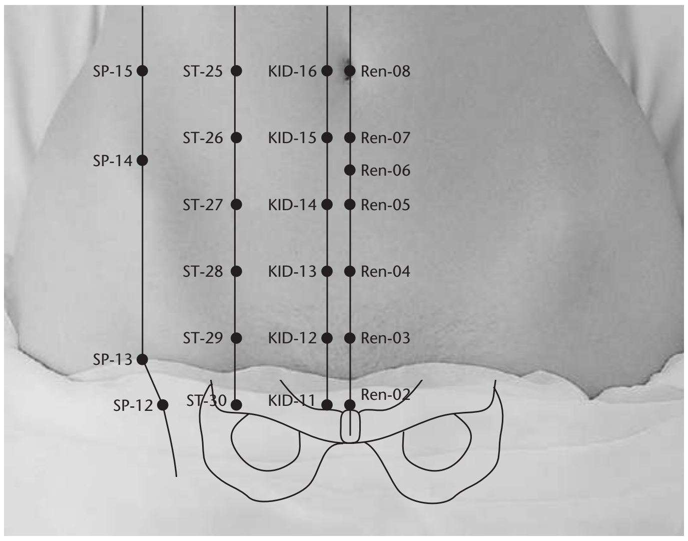
SP-15
ST-25
KID-16
Ren-08
ST-26
KID-15
Ren-07
SP-14
Ren-06
ST-27
KID-14
Ren-05
ST-28
KID-13
Ren-04
ST-29
KID-12
Ren-03
SP-13
SP-12
ST-30
KID-11
Ren-02

Anatomical level	***Ren Mai* (midline)**	**KID channel (2 cun lateral to the midline)**	**ST channel (4 cun lateral to the midline)**	**Other (6 cun lateral to the midline)**
Below the clavicle	≈ Ren-21	KID-27	ST-13	LU-2
1st ICS	Ren-20	KID-26	ST-14	LU-1
2nd ICS	Ren-19	KID-25	ST-15	SP-20
3rd ICS	Ren-18	KID-24	ST-16	SP-19
4th ICS	Ren-17	KID-23	ST-17 (nipple)	SP-18/P-1 (1 cun lateral to the nipple)/G.B.-22 (3 cun inferior to the apex of the axilla)/G.B.-23 (1 cun anterior to G.B.-22)
5th ICS	Ren 16	KID-22	ST-18	SP-17
Anatomical level	***Ren mai* (midline)**	**KID-channel (0.5 cun lateral to the midline)**	**ST channel (2 cun lateral to the midline)**	**Other (4 cun lateral to the midline)**
6 cun superior/6th ICS	Ren-14	KID-21	ST-19	LIV-14 (6th ICS)
5 cun superior/7th ICS	Ren-13	KID-20	ST-20	G.B.-24 (7th ICS)
4 cun superior	Ren-12	KID-19	ST-21	
3 cun superior	Ren-11	KID-18	ST-22	SP-16
2 cun superior	Ren-10	KID-17	ST-23	Ex-CA (*weishang*)
1 cun superior	Ren-9		ST-24	≈ LIV-13 (free end of the 11th rib)
Level of umbilicus	Ren-8	KID-16	ST-25	SP-15, G.B.-26 (inferior to the free end of the 12th rib, level with the umbilicus)
1 cun inferior	Ren-7	KID-15	ST-26	
(1.5 cun inferior)	Ren-6			
2 cun inferior	Ren-5	KID-14	ST-27	
(2.5 cun inferior)	Ex-CA (*zhishi*)			
3 cun inferior	Ren-4	KID-13	ST-28	Ex-CA (*tituo*)/Ex-CA (*qimen*)/ Ex-CA (*yijing*)/(4/3/1 cun lateral to the midline)/≈ G.B.-27 (anterior and medical to the ASIS ➞ 3)
4 cun inferior	Ren-3	KID-12	ST-29	Ex-CA (*zigong*), 3 cun lateral to the midline
5 cun inferior **symphysis**	Ren-2	KID-11	ST-30	SP-12 (3.5 cun lateral to the midline)

Note: The intercostal spaces and the ribs curve upwards laterally!

7.7 Posterior Shoulder and Arm

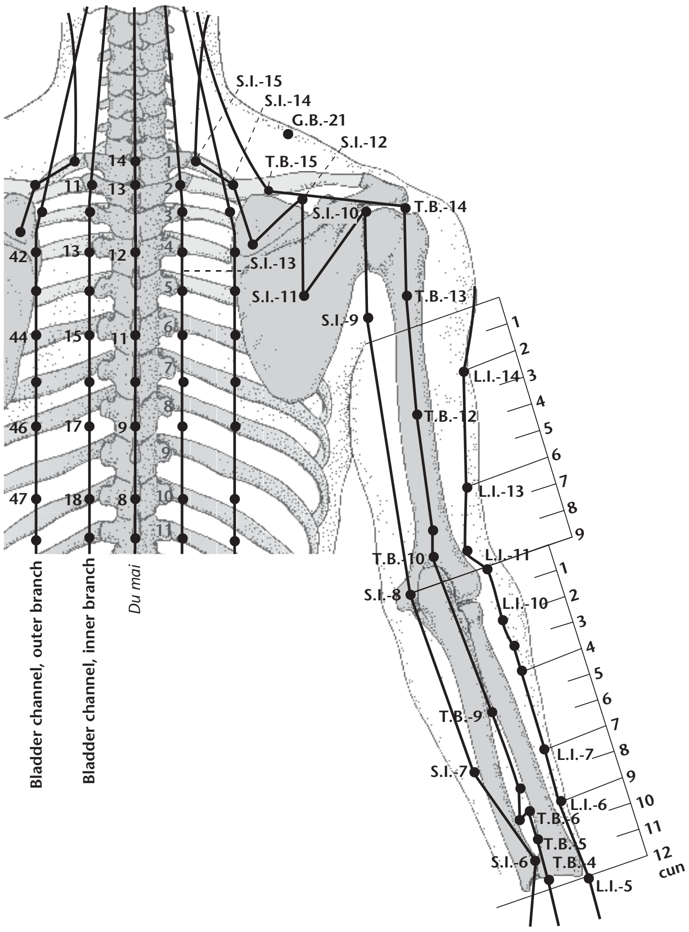
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
T.B.-14
S.I.-10
S.I.-13
S.I.-11
S.I.-9
T.B.-13
L.I.-14
T.B.-12
L.I.-13
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
11
13
14
42
12
44
15
46
17
9
47
18
8
10
1 2 3 4 5 6 7 8 9 10 11 12
cun
Bladder channel, outer branch
Bladder channel, inner branch
Du mai

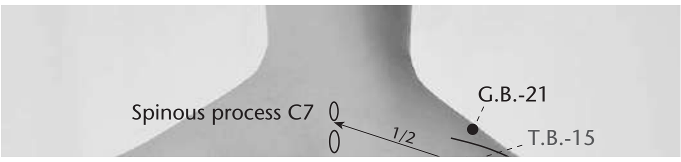
Spinous process C7
G.B.-21
1/2
T.B.-15

G.B.-21 (*jianjing*) 'Shoulder Well' At the highest point of the shoulder, midway between the lower border of the spinous process of C7 and the lateral extremity of the acromion

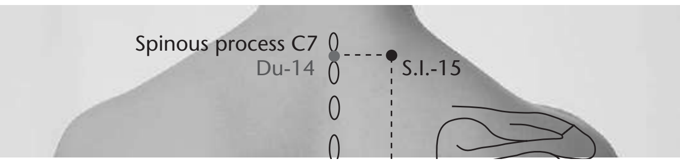
Spinous process C7
Du-14
S.I.-15

S.I.-15 (*jianzhongshu*) 'Middle Shoulder Shu' 2 cun lateral to the lower border of the spinous process of the T7 (**Du-14**)

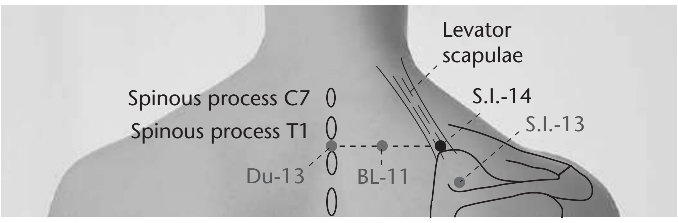
Levator scapulae
Spinous process C7
Spinous process T1
S.I.-14
S.I.-13
Du-13
BL-11

S.I.-14 (*jianwaishu*) 'Outer Shoulder Shu' 3 cun lateral to the lower border of the spinous process of the T1 (**Du-13**), at the insertion of the levator scapulae

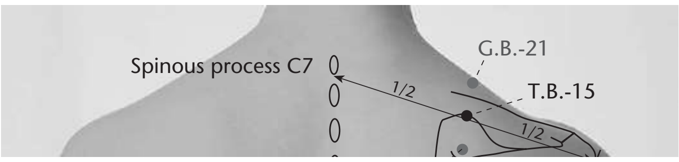
Spinous process C7
G.B.-21
1/2
T.B.-15
1/2

T.B.-15 (*tianliao*) 'Heavenly Crevice' At the midpoint of an imaginary line between the spinous process of the C7 and the lateral extremity of the acromion

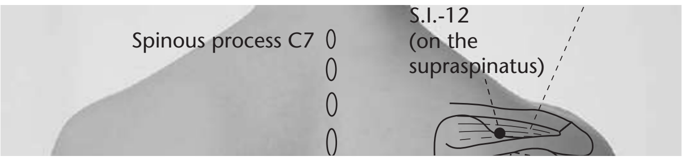
Spinous process C7
S.I.-12
(on the
supraspinatus)

S.I.-12 (*bingfeng*) 'Grasping the Wind' Directly superior to **S.I.-11**, in the centre of the supraspinous fossa

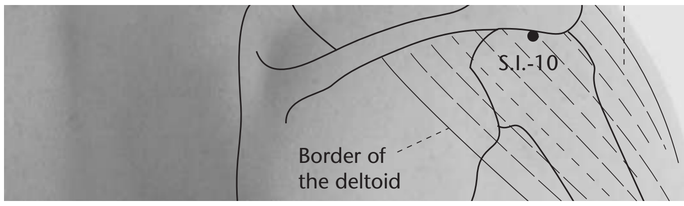
S.I.-10
Border of
the deltoid

S.I.-10 (*naoshu*) 'Upper Arm Shu' With the arm adducted on a line extending in a superior direction from the posterior axillary fold, on the lower border of the scapular spine

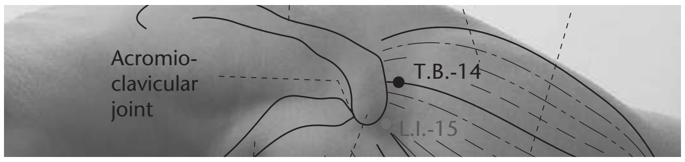
Acromio-
clavicular
joint
T.B.-14
L.I.-15

T.B.-14 (*jianliao*) 'Shoulder Crevice' Inferior to the lateral extremity of the acromion between the acromial and spinal portions of the deltoid, or with the arm abducted in the posterior of the two depressions on the shoulder joint

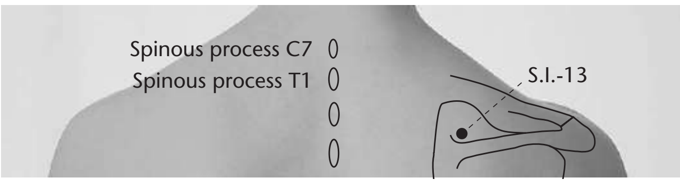
Spinous process C7
Spinous process T1
S.I.-13

S.I.-13 (*quyuan*) 'Crooked Wall' At the medial end of the supraspinous fossa

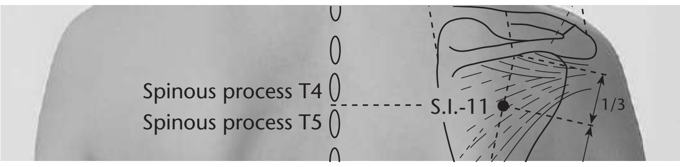
Spinous process T4
Spinous process T5
S.I.-11
1/3

S.I.-11 (*tianzong*) 'Heavenly Gathering' On the scapula, in a depression on the infraspinatus, one third of the distance from the midpoint of the scapular spine and the inferior angle of the scapula

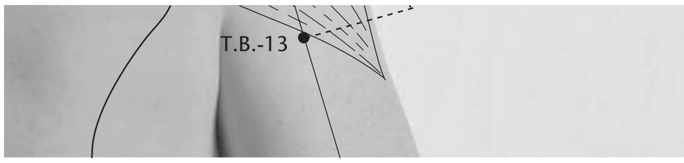

T.B.-13 (*naohui*) 'Upper Arm Meeting' 3 cun distal to the lateral extremity of the acromion (location of **T.B.-14**), on a line connecting **T.B.-14** and the olecranon, at the junction of this line with the margin of the deltoid

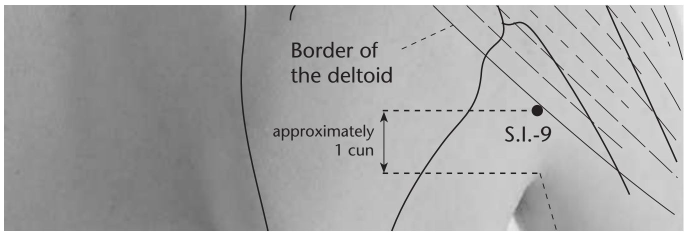
Border of the deltoid
approximately 1 cun
S.I.-9

S.I.-9 (*jianzhen*) 'True Shoulder' With the arm adducted on a line extending in a superior direction from the posterior axillary fold, on the lower border of the deltoid, 1 cun superior to the posterior axillary fold

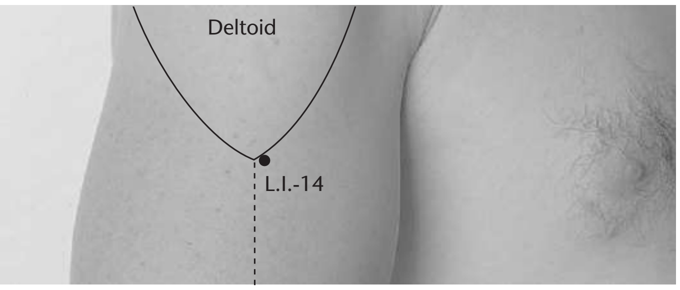
Deltoid
L.I.-14

L.I.-14 (*binao*) 'Upper Arm' 7 cun proximal to **L.I.-11**, at the distal end of the deltoid, on a line connecting **L.I.-11** and **L.I.-15**

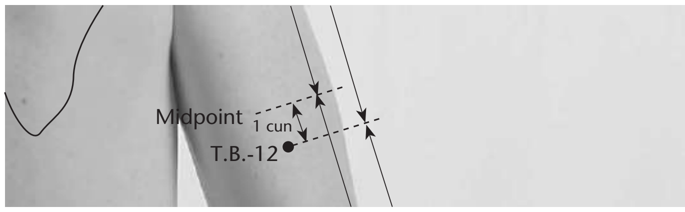
Midpoint
1 cun
T.B.-12

T.B.-12 (*xiaoluo*) 'Dispersing Luo River' 4 cun proximal to **T.B.-10** on a line connecting the olecranon and the lateral extremity of the acromion (location of **T.B.-14**)

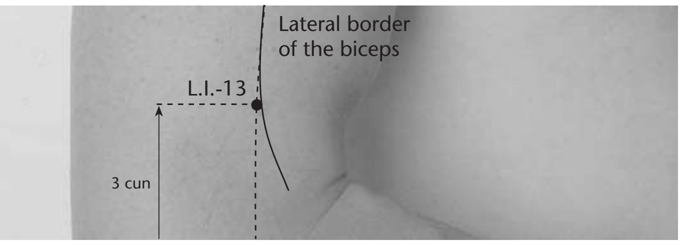
Lateral border of the biceps
L.I.-13
3 cun

L.I.-13 (*shouwuli*) 'Arm Five Miles' On the lateral aspect of the upper arm, 3 cun proximal to **L.I.-11**

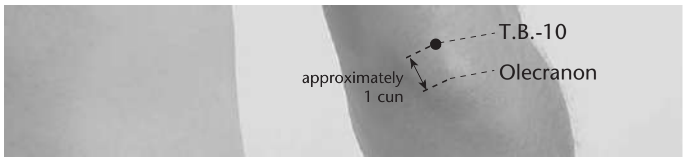
T.B.-10
Olecranon
approximately 1 cun

T.B.-10 (*tianjing*) 'Heavenly Well' With the elbow flexed in a depression approximately 1 cun proximal to the olecranon

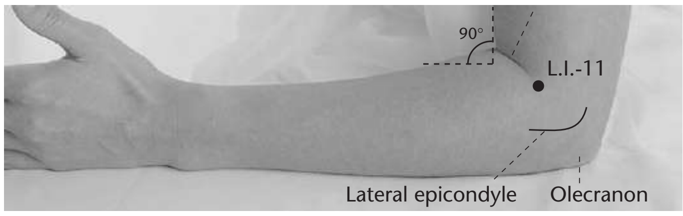
90°
L.I.-11
Lateral epicondyle
Olecranon

L.I.-11 (*quchi*) 'Pool at the Crook' With the elbow flexed, on the lateral end of the elbow crease, in a depression between the end of the crease and the lateral epicondyle of the humerus, on the extensor carpi radialis longus muscle

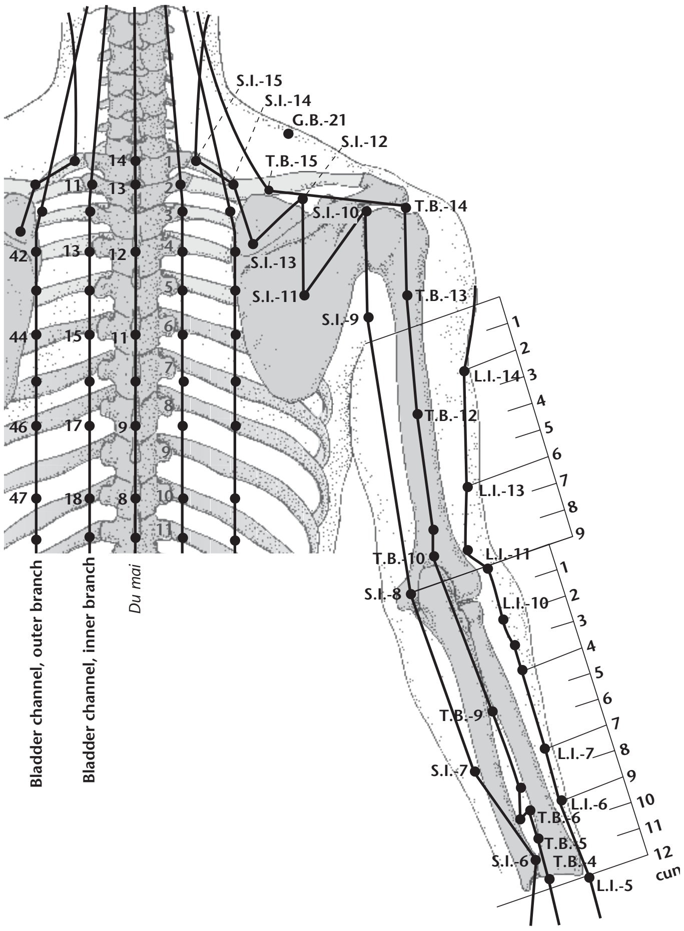
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
T.B.-14
S.I.-10
S.I.-13
S.I.-11
T.B.-13
S.I.-9
L.I.-14
T.B.-12
L.I.-13
L.I.-11
T.B.-10
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9 10 11 12
cun
14 13 12 11 9 8
1 2 3 4 5 6 7 8 9 10 11
11 13 15 17 18
42 44 46 47
Bladder channel, outer branch
Bladder channel, inner branch
Du mai

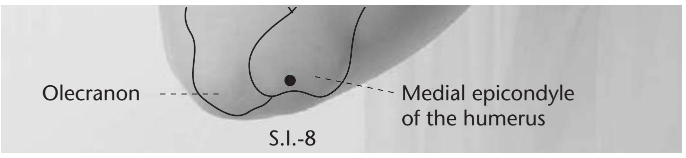
Olecranon
Medial epicondyle
of the humerus
S.I.-8

S.I.-8 (*xiaohai*) 'Small Sea' With the elbow flexed in the depression between the olecranon process of the ulna and the medial epicondyle of the humerus

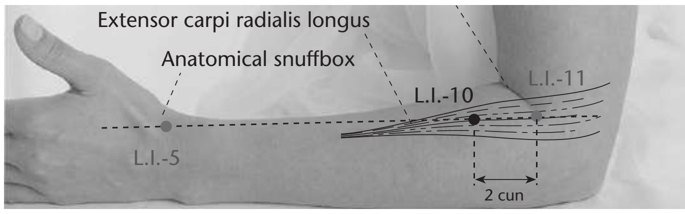
Extensor carpi radialis longus
Anatomical snuffbox
L.I.-11
L.I.-10
L.I.-5
2 cun

L.I.-10 (*shousanli*) 'Arm Three Miles' 2 cun distal to **L.I.-11**, on the line connecting **L.I.-5** and **L.I.-11**

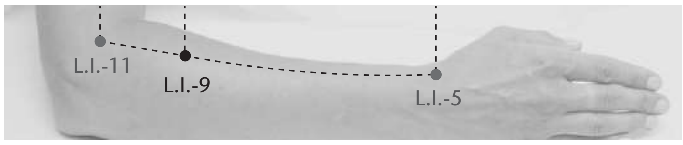
L.I.-11
L.I.-9
L.I.-5

L.I.-9 (*shanglian*) 'Upper Angle' 3 cun distal to **L.I.-11**, on the line connecting **L.I.-5** and **L.I.-11**

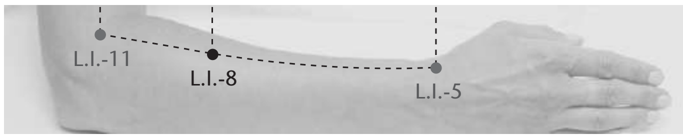
L.I.-11
L.I.-8
L.I.-5

L.I.-8 (*xialian*) 'Lower Angle' 4 cun distal L.I.-11, on the line connecting **L.I.-5** and **L.I.-11**

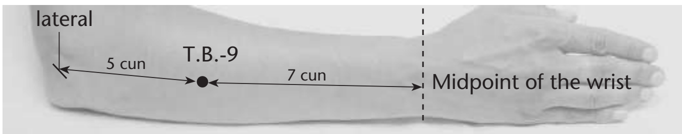
lateral
5 cun
T.B.-9
7 cun
Midpoint of the wrist

T.B.-9 (*sidu*) 'Four Rivers' 5 cun distal to the olecranon or 7 cun proximal to **T.B.-4**, between the radius and the ulna

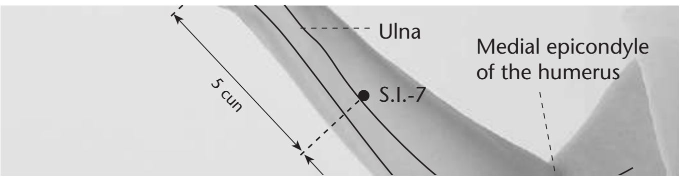
Ulna
Medial epicondyle
of the humerus
5 cun
S.I.-7

S.I.-7 (*zhizheng*) 'Branch of the Upright' 5 cun proximal to the wrist joint space, on the line connecting **S.I.-5** and **S.I.-8**

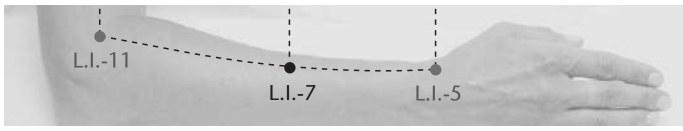
L.I.-11
L.I.-7
L.I.-5

L.I.-7 (*wenliu*) 'Warm Flow' 5 cun proximal to **L.I.-5** or midway between **L.I.-5** and **L.I.-11**

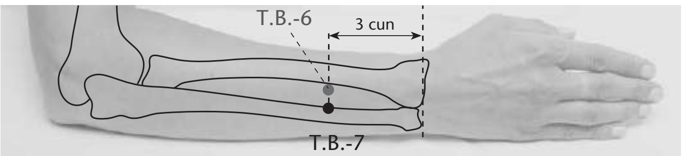
T.B.-6
3 cun
T.B.-7

T.B.-7 (*huizhong*) 'Ancestral Meeting' 3 cun proximal to **T.B.-4**, between radius and ulna, ulnar to extensor digitorum communis

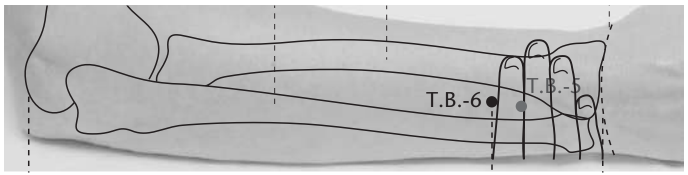

T.B.-6 (*zhigou*) 'Branch Ditch' 3 cun proximal to **T.B.-4**, between radius and ulna, radial to the tendon of extensor digitorum communis

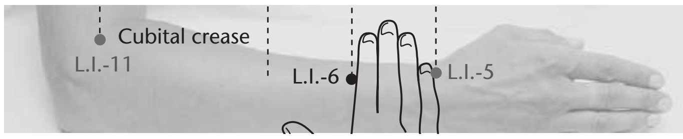
Cubital crease
L.I.-11
L.I.-6
L.I.-5

L.I.-6 (*pianli*) 'Veering Passage' 3 cun proximal to **L.I.-5**, on the line connecting **L.I.-5** and **L.I.-11**

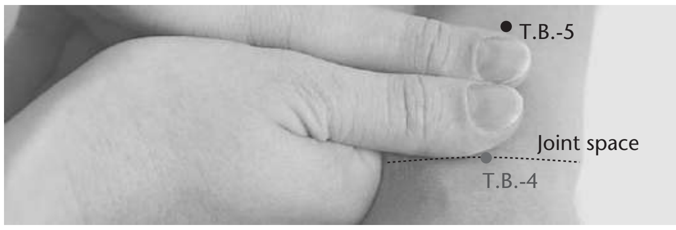
T.B.-5
Joint space
T.B.-4

T.B.-5 (*waiguan*) 'Outer Pass' 2 cun proximal to **T.B.-4** (on the dorsal wrist joint space, between radius and ulna.

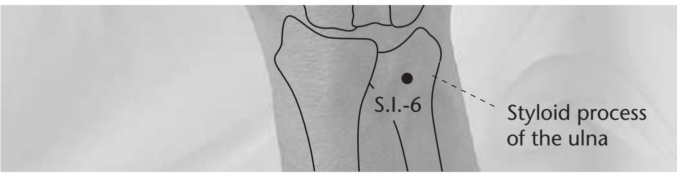
S.I.-6
Styloid process
of the ulna

S.I.-6 (*yanglao*) 'Support the Aged' In the depression radial and proximal to the styloid process of the ulna that forms when turning the hand from a pronated to a supinated position.

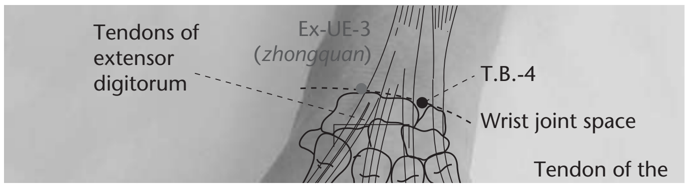
Tendons of
extensor
digitorum
Ex-UE-3
(zhongquan)
T.B.-4
Wrist joint space
Tendon of the

T.B.-4 (*yangchi*) 'Yang Pool' On the dorsum of the wrist, in the gap ulnar to the tendons of the extensor digitorum

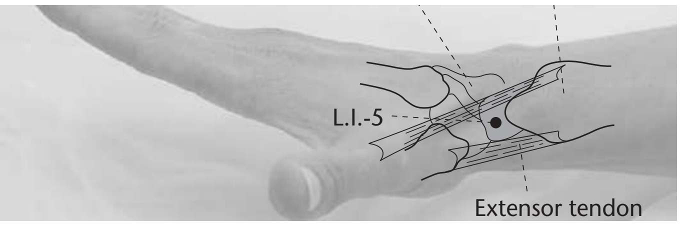
L.I.-5
Extensor tendon

L.I.-5 (*yangxi*) 'Yang Stream' With the thumb abducted in the depression between the tendons of the extensor pollicis longus and brevis muscles ('anatomical snuff box')

7.8 Anterior Shoulder and Arm

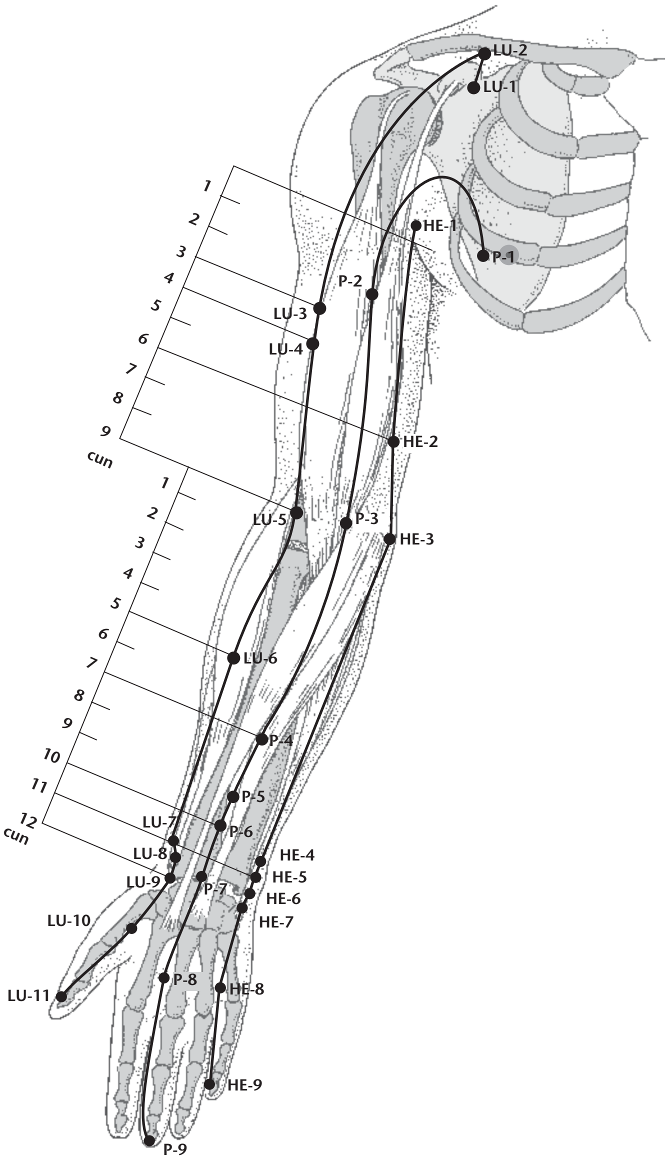
LU-2
LU-1
HE-1
P-1
P-2
LU-3
LU-4
HE-2
LU-5
P-3
HE-3
LU-6
P-4
P-5
P-6
LU-7
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9
1
2
3
4
5
6
7
8
9
cun
1
2
3
4
5
6
7
8
9
10
11
12
cun

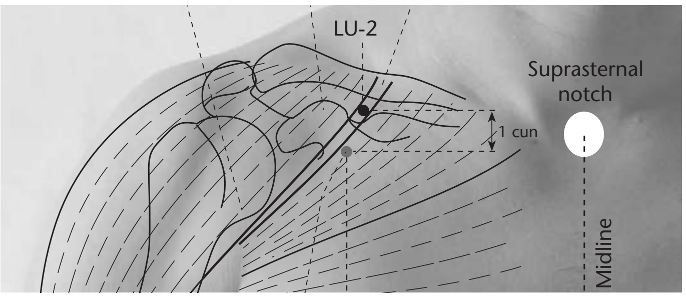
LU-2
Suprasternal notch
1 cun
Midline

LU-2 (*yunmen*) 'Cloud Gate' 6 cun lateral to the anterior midline, below the clavicle, in the centre of the deltopectoral triangle.

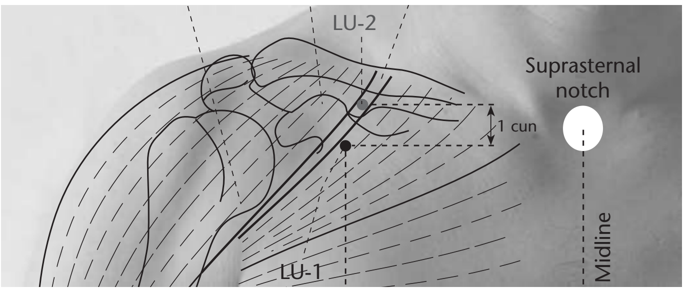
LU-2
Suprasternal notch
1 cun
Midline
LU-1

LU-1 (*zhongfu*) 'Middle Palace' 6 cun lateral to the anterior midline and approximately 1 cun inferior to **LU-2**, slightly medial to the lower border of the coracoid process

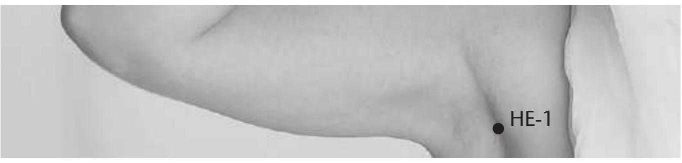

HE-1 (*jiquan*) 'Summit Spring' With the arm abducted, in the centre of the axilla, medial to the axillary artery

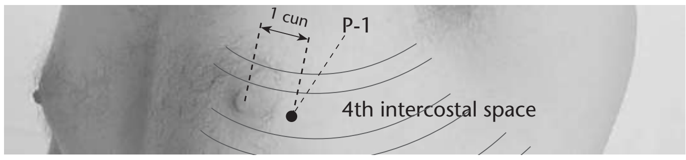
1 cun
P-1
4th intercostal space

P-1 (*tianchi*) 'Heavenly Pool' In the 4th ICS, 1 cun lateral to the nipple

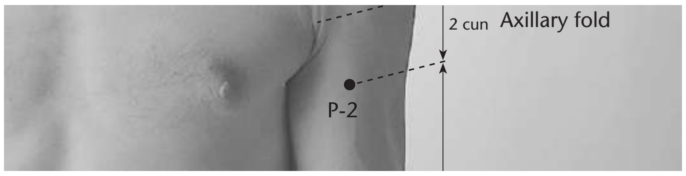
2 cun
Axillary fold
P-2

P-2 (*tianquan*) 'Heavenly Spring' Between the two heads of the biceps brachii muscle, 2 cun inferior to the axillary fold.

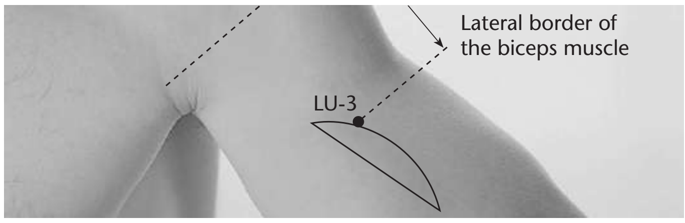
Lateral border of the biceps muscle
LU-3

LU-3 (*tianfu*) 'Palace of Heaven' 3 cun distal to the end of the anterior axillary fold, on the medial aspect of the upper arm, radial to the biceps brachialis

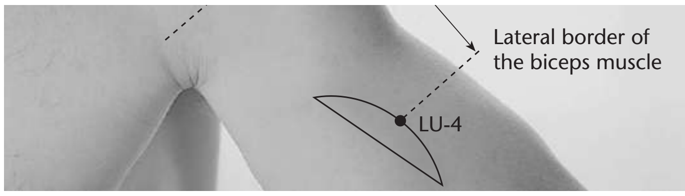
Lateral border of
the biceps muscle
LU-4

LU-4 (*xiabai*) ‘Clasping the White’ 4 cun distal to the end of the anterior axillary fold, on the medial aspect of the upper arm, radial to the biceps brachii

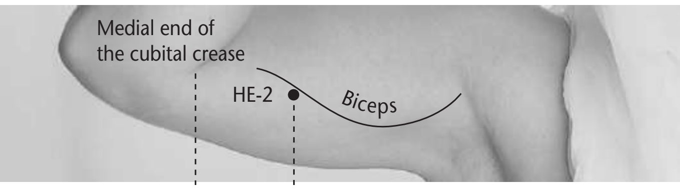
Medial end of
the cubital crease
HE-2
Biceps

HE-2 (*qingling*) ‘Green Spirit’ 3 cun proximal to the cubital crease, on the medial border of the biceps brachii

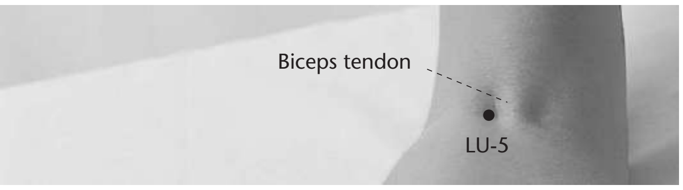
Biceps tendon
LU-5

LU-5 (*chize*) ‘Cubit Marsh’ In the cubital crease, radial to the tendon of the biceps brachii

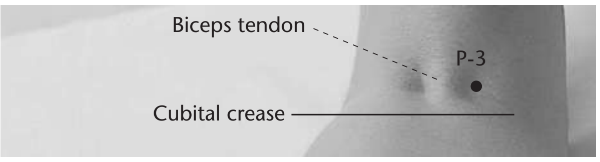
Biceps tendon
P-3
Cubital crease

P-3 (*quze*) ‘Marsh at the Crook’ In the cubital crease, ulnar to the biceps tendon

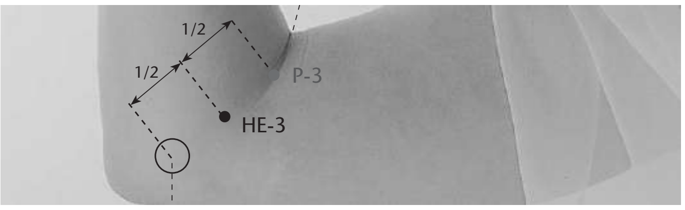

HE-3 (*shaohai*) ‘Lesser Sea’ With the elbow flexed, in the depression between the ulnar end of the cubital crease and the medial epicondyle of the humerus

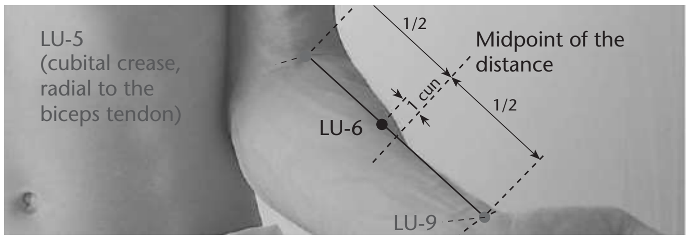
LU-5
(cubital crease,
radial to the
biceps tendon)
1/2
Midpoint of the
distance
1 cun
1/2
LU-6
LU-9

LU-6 (*kongzui*) ‘Maximum Opening’ 5 cun distal to **LU-5** or 7 cun proximal to the wrist joint space, on a line connecting **LU-5** and **LU-9**

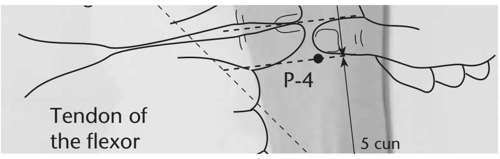
P-4
Tendon of
the flexor
5 cun

P-4 (*ximen*) ‘Xi-Cleft Gate’ 5 cun proximal to the palmar wrist joint space, between the tendons of palmaris longus and flexor carpi radialis, or 1 cun distal to the midpoint of the line connecting **P-3** and **P-7**

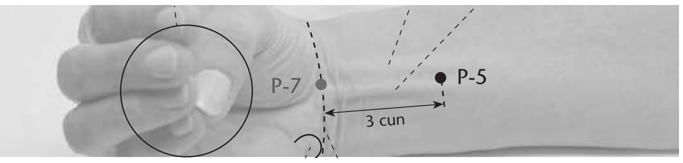

P-5 (*jianshi*) ‘Intermediate Messenger’ 3 cun proximal to the palmar wrist joint space, between the tendons of palmaris longus and flexor carpi radialis

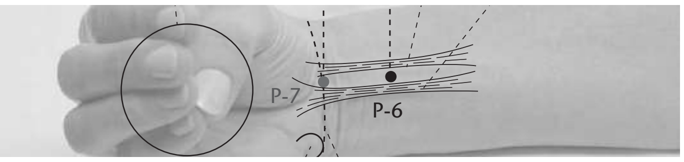

P-6 (*neiguan*) ‘Inner Pass’ 2 cun proximal to the palmar wrist joint space, between the tendons of palmaris longus and flexor carpi radialis

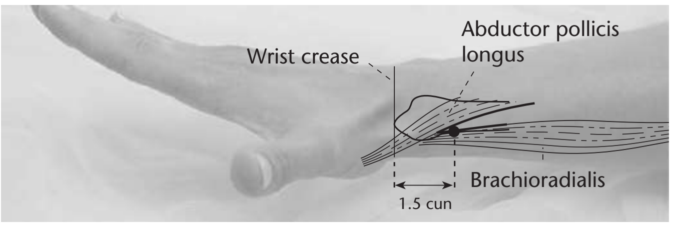
Wrist crease
Abductor pollicis
longus
Brachioradialis
1.5 cun

LU-7 (*lieque*) ‘Broken Sequence’ On the radial aspect of the forearm, immediately above the styloid process of the radius, approximately 1.5 cun proximal to the wrist joint space in a V-shaped groove

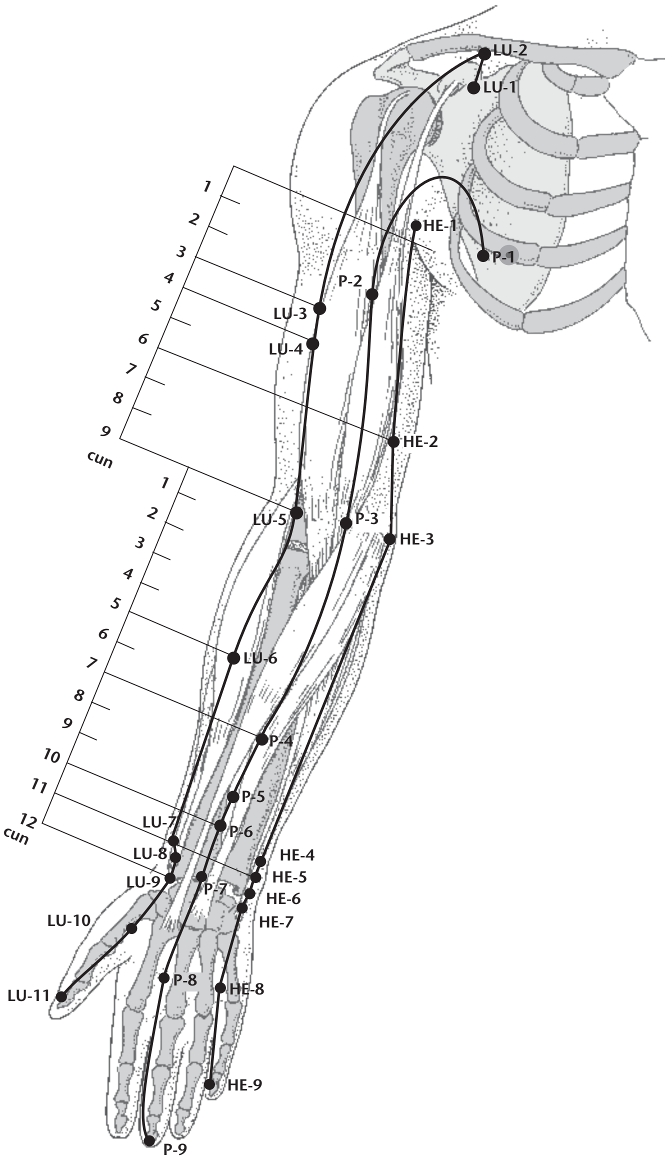
LU-2
LU-1
HE-1
P-1
1
2
3
4
5
6
7
8
9
cun
P-2
LU-3
LU-4
HE-2
1
2
3
4
5
6
7
8
9
10
11
12
cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
LU-8
HE-4
HE-5
LU-9
P-7
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9

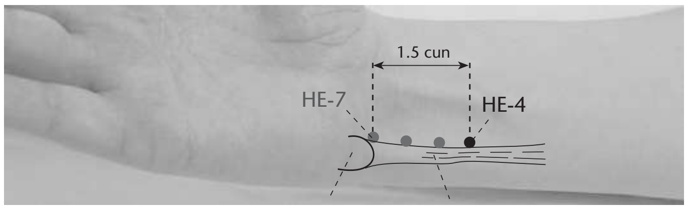
1.5 cun
HE-7
HE-4

HE-4 (*lingdao*) 'Spirit Path' 1.5 cun proximal to the palmar wrist joint space, radial to the flexor carpi ulnaris tendon

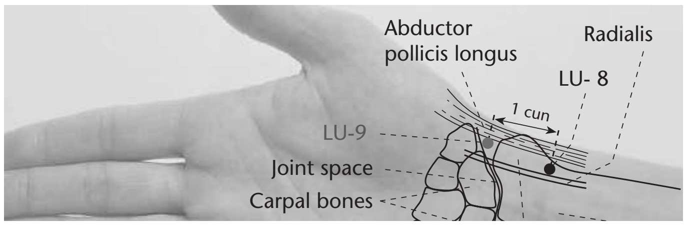
Abductor pollicis longus
Radialis
LU- 8
1 cun
LU-9
Joint space
Carpal bones

LU-8 (*jingqu*) 'Channel Gutter' 1 cun proximal to the palmar wrist joint space (**LU-9**), lateral to the radial artery

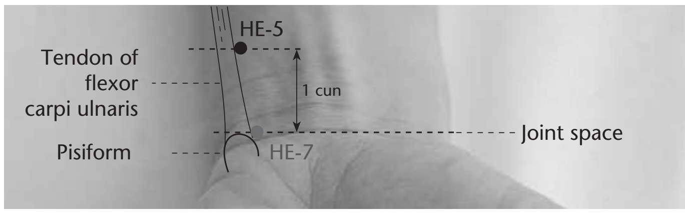
HE-5
Tendon of flexor carpi ulnaris
1 cun
Joint space
Pisiform
HE-7

HE-5 (*tongli*) 'Penetrating the Interior' 1 cun proximal to the palmar wrist joint space, on the radial side of the flexor carpi ulnaris tendon

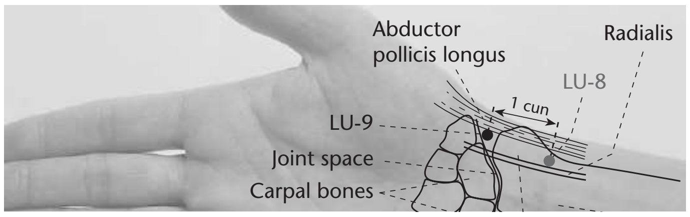
Abductor pollicis longus
Radialis
LU-8
1 cun
LU-9
Joint space
Carpal bones

LU-9 (*taiyuan*) 'Supreme Abyss' On the wrist joint space, lateral to the radial artery and ulnar to the tendon of abductor pollicis longus

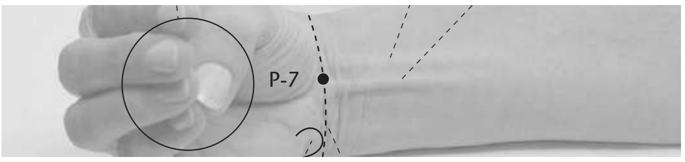

P-7 (*daling*) 'Great Mound' In the centre of the palmar aspect of the wrist joint space ('most distal wrist crease'), between the tendons of palmaris longus and flexor carpi radialis

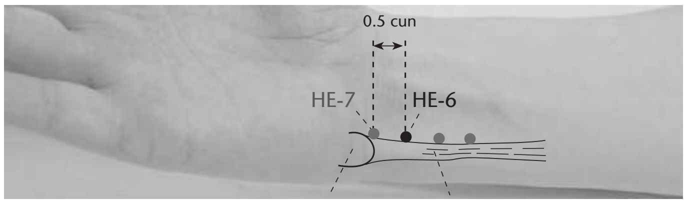
0.5 cun
HE-7
HE-6

HE-6 (*yinxi*) 'Yin Cleft' 0.5 cun proximal to the palmar wrist joint space, radial to the flexor carpi ulnaris tendon

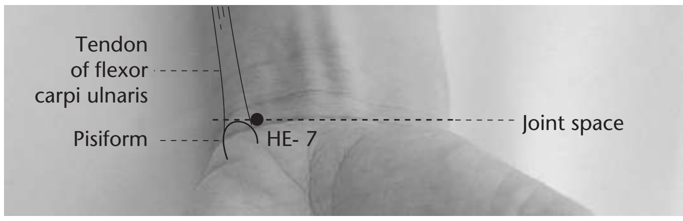
Tendon of flexor carpi ulnaris
Pisiform
HE- 7
Joint space

HE-7 (*shenmen*) 'Spirit Gate' On the palmar aspect of the wrist joint space ('most distal wrist crease'), in the depression radial to the insertion of the flexor carpi ulnaris tendon

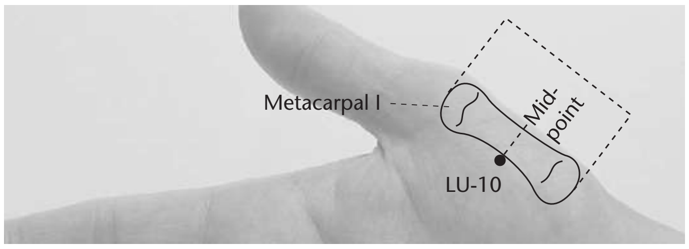
Metacarpal I
Mid-point
LU-10

LU-10 (*yuji*) 'Fish Border' At the midpoint of the palmar border of the 1st metacarpal bone

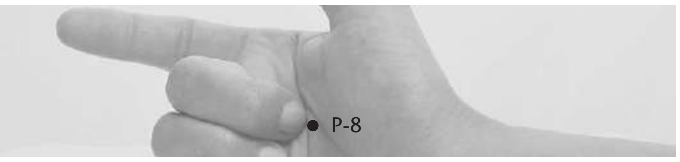

P-8 (*laogong*) 'Palace of Toil' In the centre of the palm, between the 2nd and 3rd metacarpal bones, slightly closer to the 3rd metacarpal bone. When a fist is made, the point is located beneath the tip of the middle finger

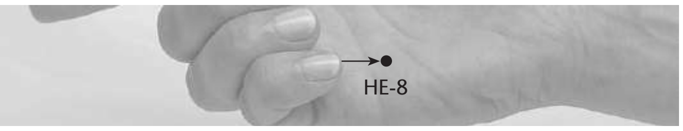

HE-8 (*shaofu*) 'Lesser Palace' When a fist is made, beneath the tip of the little finger, on the palm, between the 4th and 5th metacarpal bones

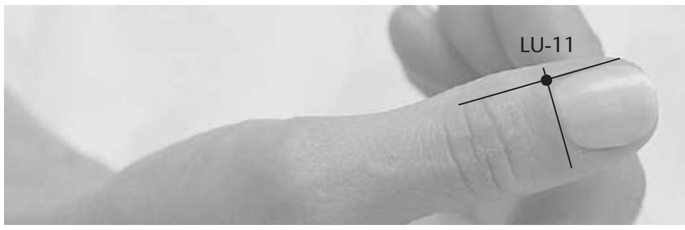

LU-11 (*shaoshang*) 'Lesser Shang' 0.1 cun from the radial corner of the thumb nail

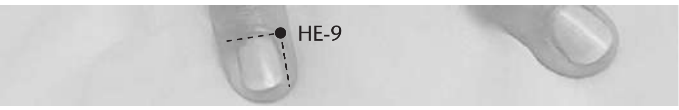

HE-9 (*shaochong*) 'Lesser Rushing' 0.1 cun lateral to the radial corner of the nail of the little finger

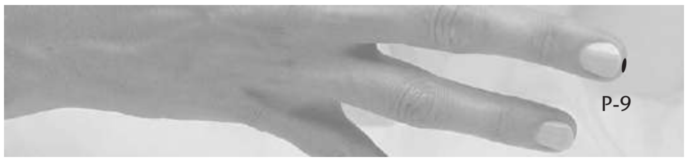

P-9 (*zhongchong*) 'Middle Rushing' On the most distal point of the middle finger

7.9 Dorsum of the Hand

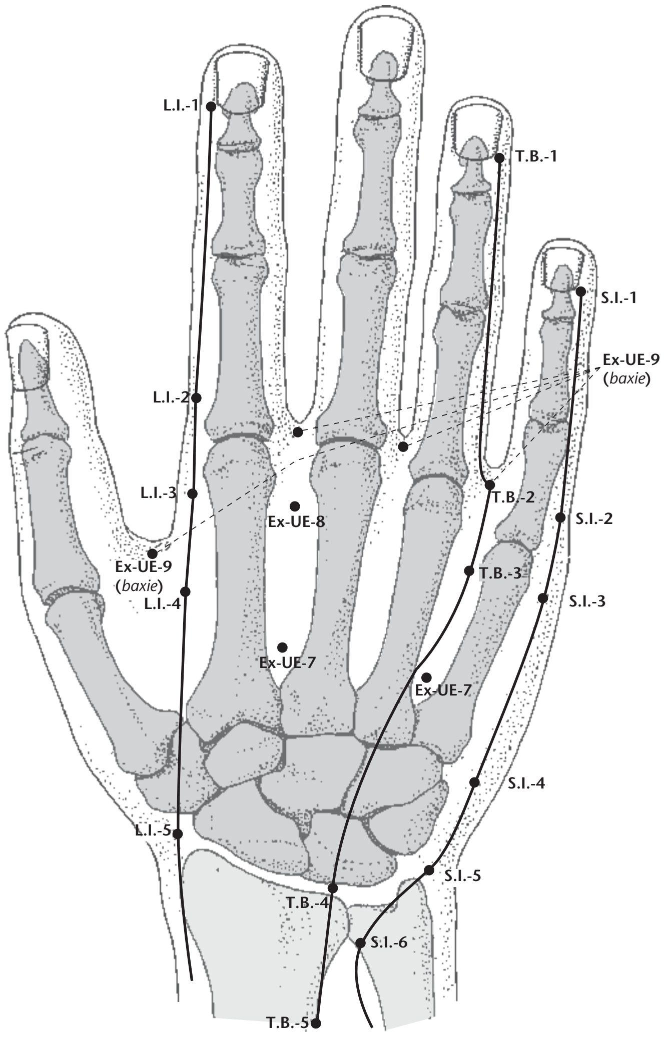
L.I.-1
T.B.-1
S.I.-1
Ex-UE-9
(*baxie*)
L.I.-2
L.I.-3
T.B.-2
S.I.-2
Ex-UE-8
Ex-UE-9
(*baxie*)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

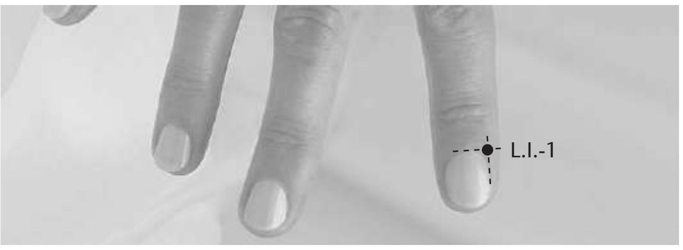

L.I.-1 (*shangyang*) 'Shang Yang' On the index finger, 0.1 cun from the radial corner of the nail

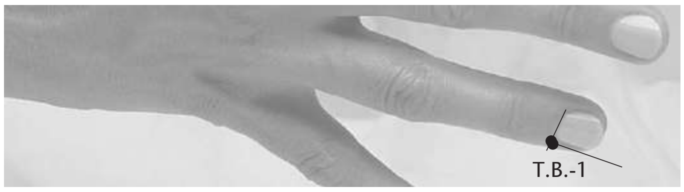

T.B.-1 (*guanchong*) 'Rushing Pass' On the ring finger (4th finger), 0.1 cun from the ulnar corner of the nail

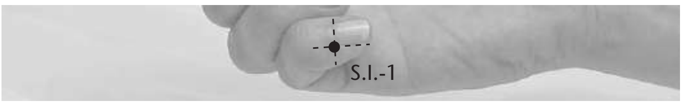

S.I.-1 (*shaoze*) 'Lesser Marsh' On the little finger, 0.1 cun from the ulnar corner of the nail

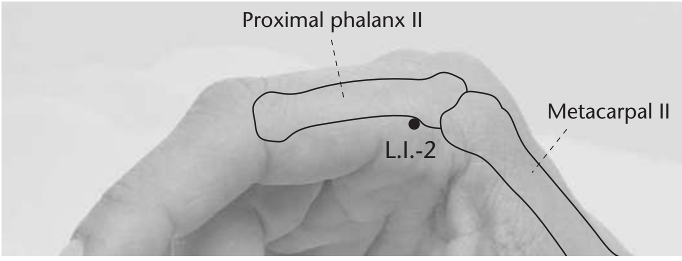
Proximal phalanx II
Metacarpal II
L.I.-2

L.I.-2 (*erjian*) 'Second Space' On the radial aspect of the index finger, distal to the metacarpophalangeal joint at the junction of the shaft and the basis of the proximal phalanx

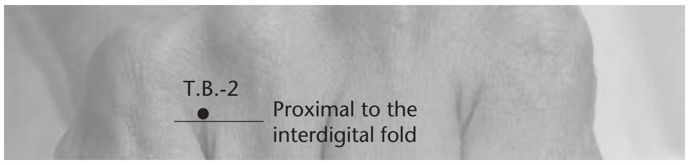
T.B.-2
Proximal to the
interdigital fold

T.B.-2 (*yemen*) 'Fluid Gate' Between the little finger and ring finger, proximal to the margin of the web; locate with the patient making a fist

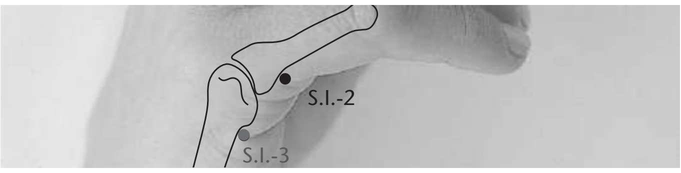

S.I.-2 (*qiangu*) 'Front Valley' On the ulnar aspect of the little finger, at the junction of the shaft and the base of the proximal phalanx

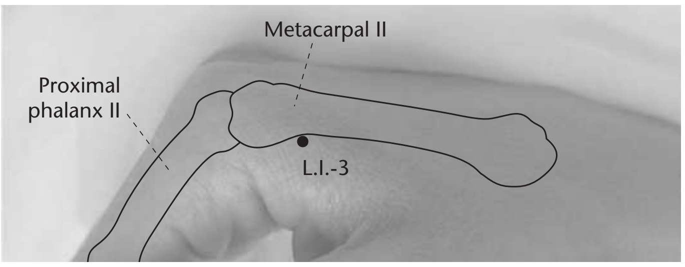
Metacarpal II
Proximal phalanx II
L.I.-3

L.I.-3 (*sanjian*) 'Third Space' On the radial aspect of the index finger, proximal to the metacarpophalangeal joint at the junction of the shaft and the head of the 2nd metacarpal bone

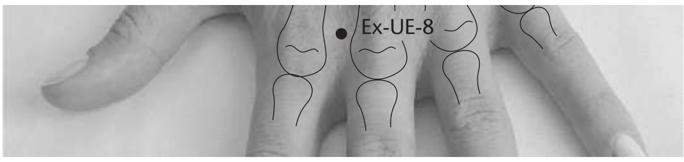

Ex-UE-8 (*wailaogong/luozhen/xianqiang*) 'Stiff Neck' On the dorsum of the hand, between the 2nd and 3rd metacarpal bones, approximately 0.5 cun proximal to the metacarpophalangeal joints

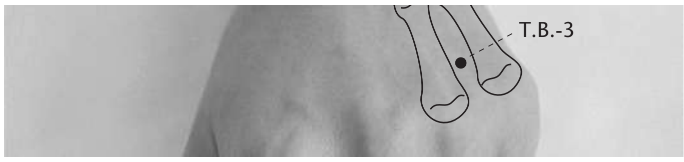

T.B.-3 (*zhongzhu*) 'Central Islet' On the dorsum of the hand, in the depression between the 4th and 5th metacarpal bones, proximal to the metacarpophalangeal joints. **T.B.-3** is located at the junction between the heads and shafts of the two metacarpals

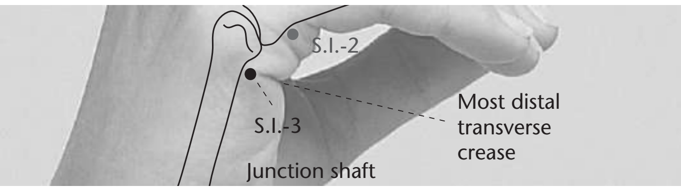
S.I.-2
S.I.-3
Most distal transverse crease
Junction shaft

S.I.-3 (*houxi*) 'Back Stream' On the ulnar border of the hand, in the depression proximal to the metacarpophalangeal joint, at the junction between the head and shaft of the 5th metacarpal

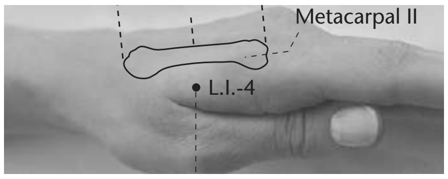
Metacarpal II
L.I.-4

L.I.-4 (*hegu*) 'Joining Valley' On the radial aspect of the hand, between the 1st and 2nd metacarpal bone, closer to the 2nd metacarpal bone and approximately at its midpoint

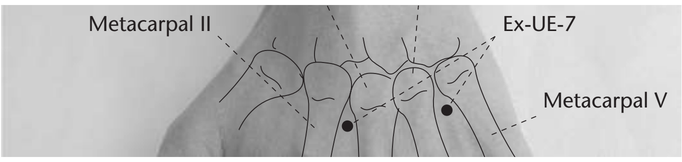
Metacarpal II
Ex-UE-7
Metacarpal V

Ex-UE-7 (*yaotongdian/yaotongxue*) 'Lumbar Pair Point' Two points on the dorsum of the hand, between the 2nd/3rd and 4th/5th metacarpal bones

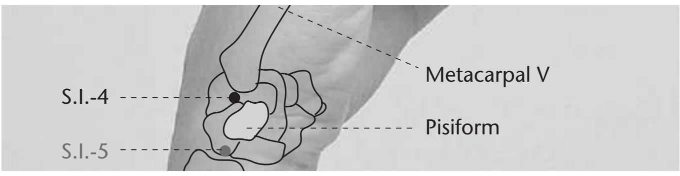
Metacarpal V
S.I.-4
Pisiform
S.I.-5

S.I.-4 (*wangu*) 'Wrist Bone' On the ulnar border of the hand, between the 5th metacarpal and the carpal bones, at the border of the red and white skin

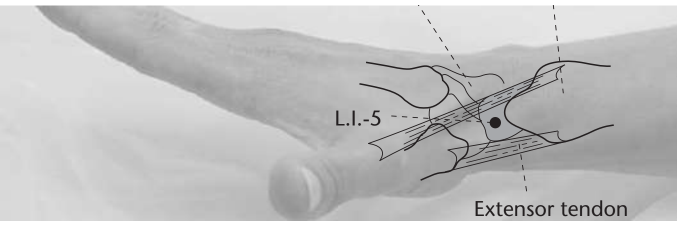
L.I.-5
Extensor tendon

L.I.-5 (*yangxi*) 'Yang Stream' With the thumb abducted in the depression between the tendons of the extensor pollicis longus and brevis muscles ('anatomical snuff box')

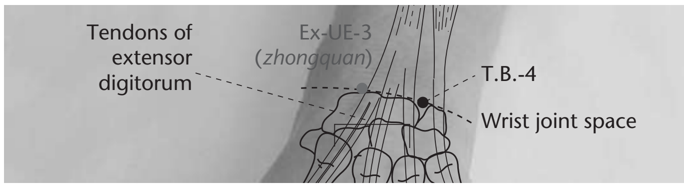
Tendons of extensor digitorum
Ex-UE-3 (zhongquan)
T.B.-4
Wrist joint space

T.B.-4 (*yangchi*) 'Yang Pool' On the dorsum of the wrist, in the gap ulnar to the tendons of extensor digitorum

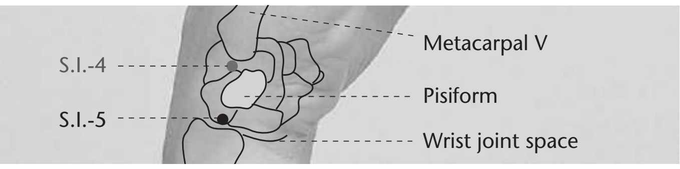
Metacarpal V
S.I.-4
Pisiform
S.I.-5
Wrist joint space

S.I.-5 (*yanggu*) 'Yang Valley' On the ulnar aspect of the wrist, at the level of the lateral joint space

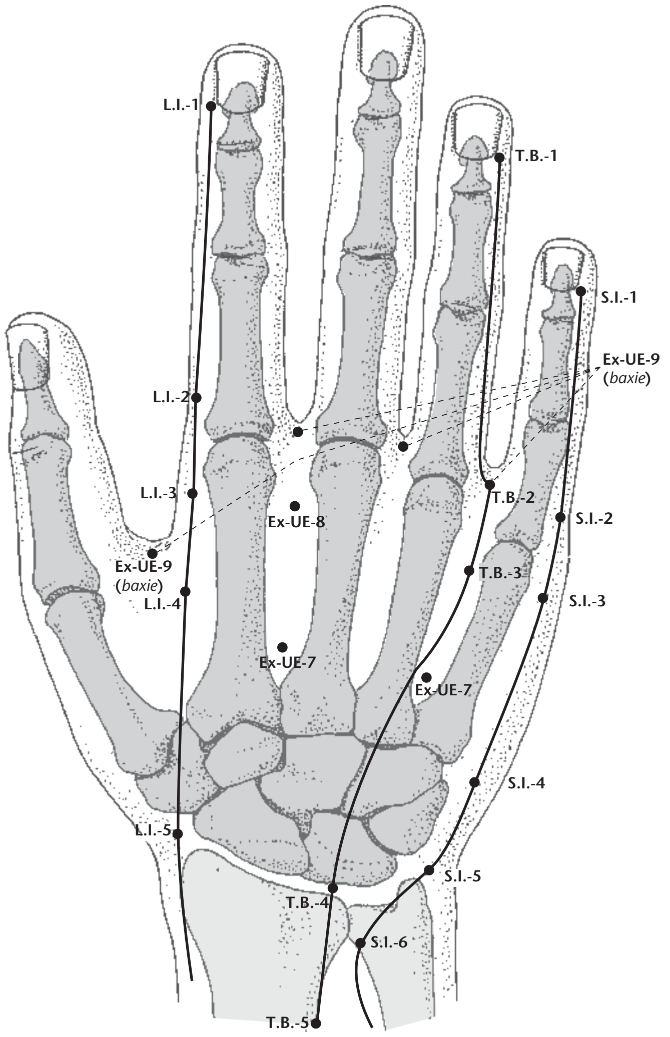
L.I.-1
T.B.-1
S.I.-1
Ex-UE-9
(*baxie*)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(*baxie*)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

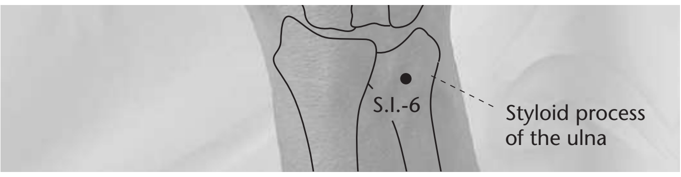
S.I.-6
Styloid process
of the ulna

S.I.-6 (*yanglao*) 'Nourishing the Old' In the depression radial and proximal to the styloid process of the ulna that forms when the hand is turned from a pronated to a supinated position

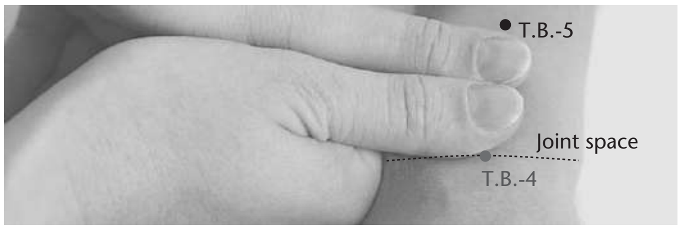
T.B.-5
Joint space
T.B.-4

T.B.-5 (*waiguan*) 'Outer Pass' 2 cun proximal to **T.B.-4** (on the dorsal wrist joint space, between radius and ulna)

7.10 Anterior Leg

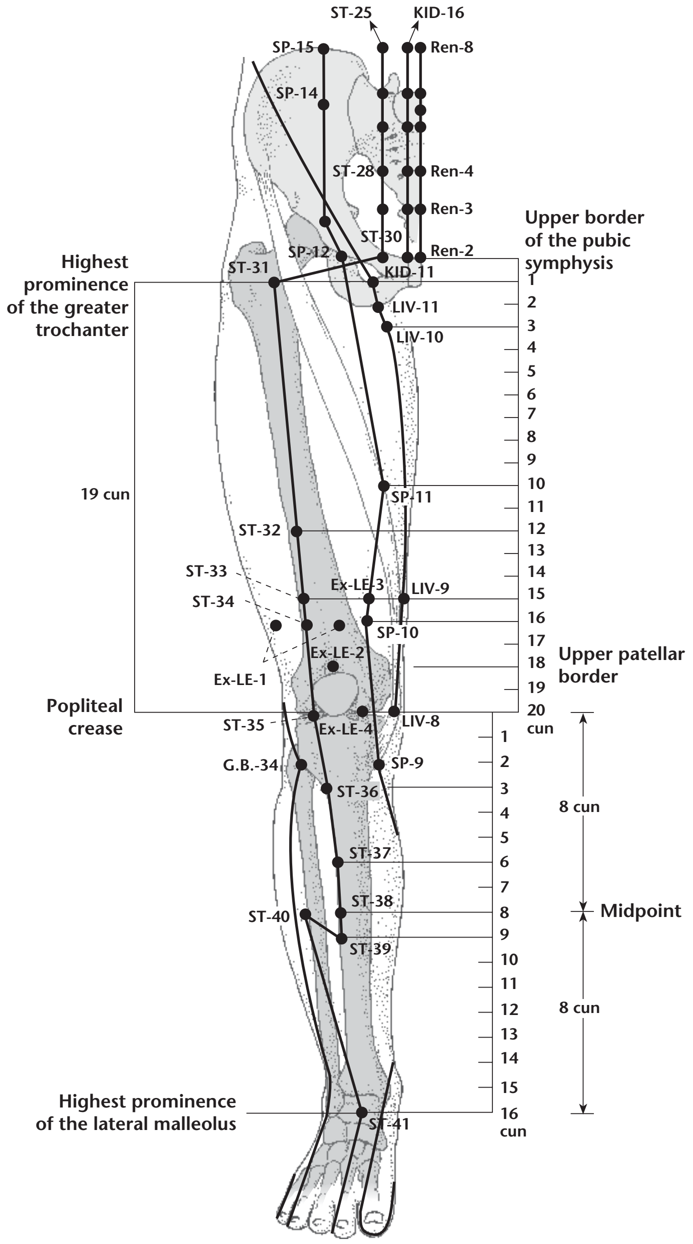
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
Upper border
of the pubic
symphysis
ST-30
SP-12
Ren-2
ST-31
KID-11
Highest
prominence
of the greater
trochanter
LIV-11
LIV-10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
cun
19 cun
SP-11
ST-32
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Upper patellar
border
Ex-LE-1
Popliteal
crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
8 cun
ST-37
ST-38
Midpoint
ST-40
ST-39
8 cun
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun
Highest prominence
of the lateral malleolus
ST-41

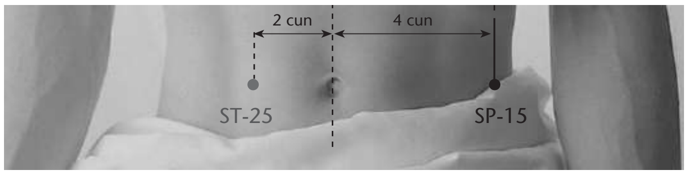
2 cun
4 cun
ST-25
SP-15

SP-15 (*daheng*) 'Great Horizontal' 4 cun lateral to the centre of the umbilicus, on the mamillary line

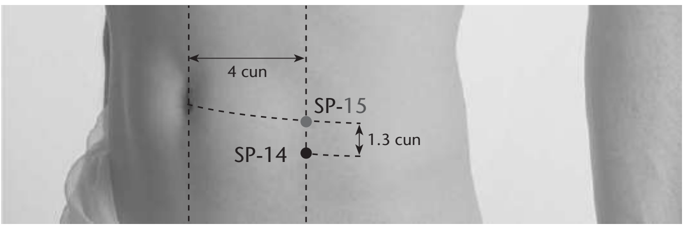
4 cun
SP-15
SP-14
1.3 cun

SP-14 (*fujie*) 'Abdomen Knot' 4 cun lateral to the anterior midline (mamillary line), 3 cun superior to **SP-13** or 1.3 cun inferior to **SP-15**

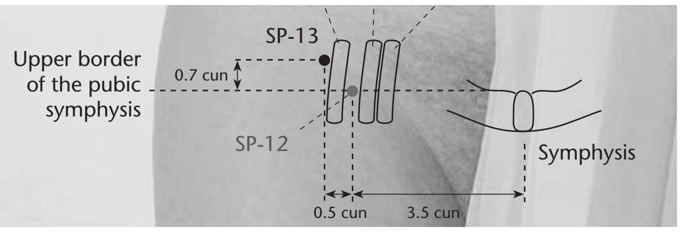
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

SP-13 (*fushe*) 'Meeting of the Fu Organs' 4 cun lateral to the anterior midline (mamillary line) and 0.7 cun superior to the upper border of the pubic symphysis

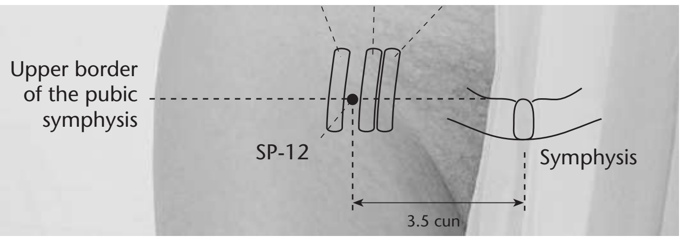
Upper border of the pubic symphysis
SP-12
Symphysis
3.5 cun

SP-12 (*chongmen*) 'Rushing Gate' 3.5 cun lateral to the midline, level with the upper border of the pubic symphysis, lateral to the femoral artery

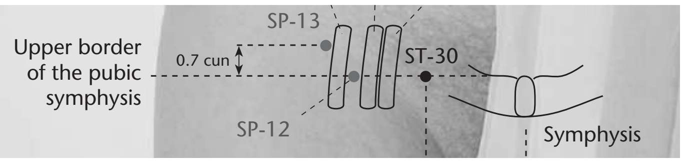
SP-13
Upper border of the pubic symphysis
0.7 cun
ST-30
SP-12
Symphysis

ST-30 (*qichong*) 'Rushing Qi' 2 cun lateral to the upper border of the pubic symphysis, medial to the femoral artery; at the level of **Ren-2**

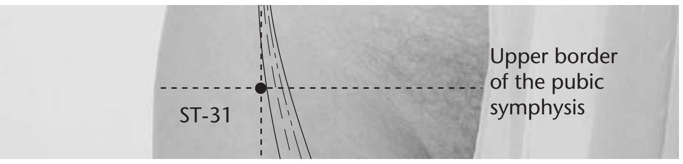
Upper border of the pubic symphysis
ST-31

ST-31 (*biguan*) 'Thigh Gate' With the hip joint flexed, in the depression below the anterior superior iliac spine, lateral to the sartorius

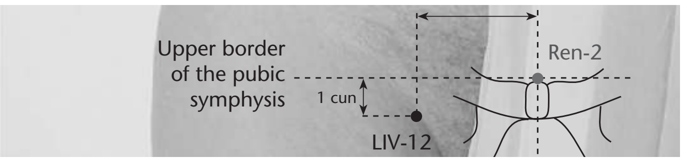
Upper border of the pubic symphysis
Ren-2
1 cun
LIV-12

LIV-12 (*jimai*) 'Urgent Pulse' In the inguinal groove, 2.5 cun lateral to the midline and 1 cun inferior to the upper border of the symphysis, inferior to **ST-30** and medial to the femoral artery

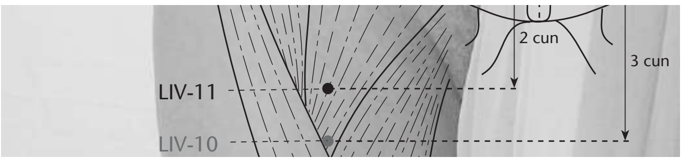
2 cun
3 cun
LIV-11
LIV-10

LIV-11 (*yinlian*) 'Yin Corner' 2 cun inferior to **ST-30** (level with the upper border of the pubic symphysis), on the lateral border of the adductor longus muscle or approximately 1 cun inferior to where the femoral artery passes the inguinal ligament

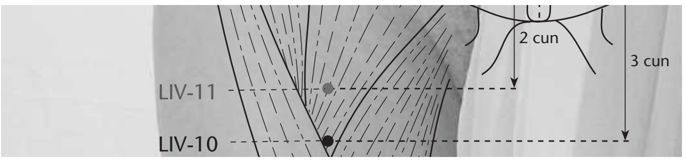
2 cun
3 cun
LIV-11
LIV-10

LIV-10 (*zuwuli*) 'Leg Five Miles' On the anterior aspect of the thigh, 3 cun inferior to the upper border of the symphysis, on the lateral border of adductor longus

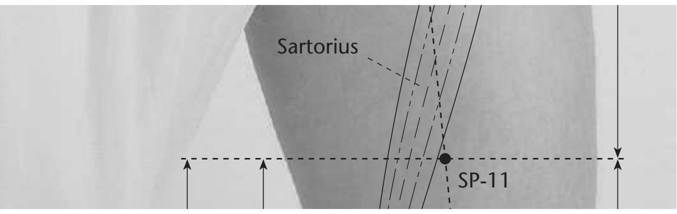
Sartorius
SP-11

SP-11 (*jimen*) 'Winnowing Gate' 6 cun proximal to **SP-10** at the midpoint of the femur, in the depression between the sartorius and vastus lateralis

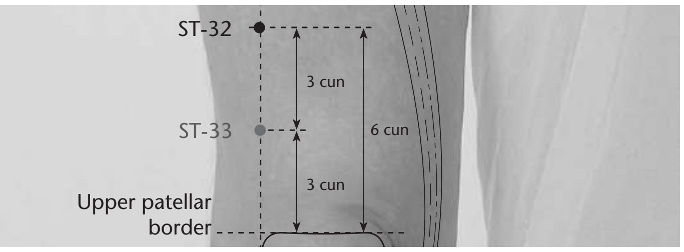
ST-32
3 cun
ST-33
6 cun
3 cun
Upper patellar border

ST-32 (*futu*) 'Crouching Rabbit' 6 cun proximal to the lateral upper border of the patella, on a line connecting with the anterior superior iliac spine

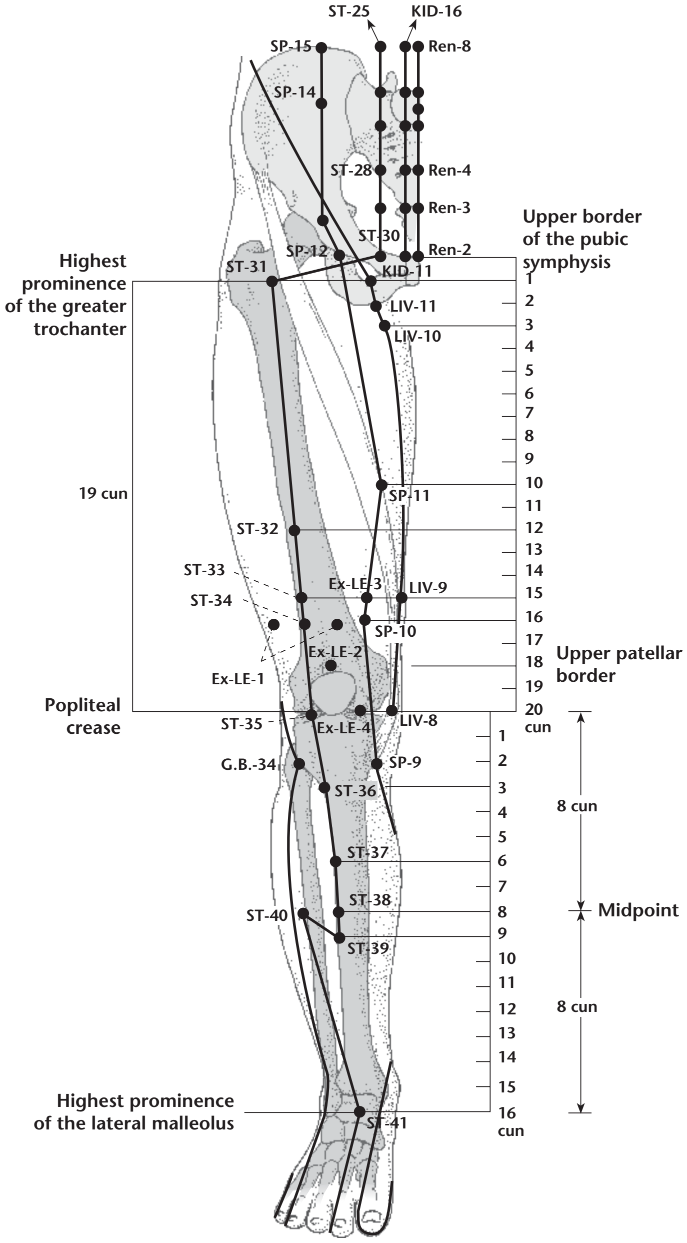
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
Upper border of the pubic symphysis
ST-30
SP-12
Ren-2
Highest prominence of the greater trochanter
ST-31
KID-11
LIV-11
LIV-10
19 cun
SP-11
ST-32
ST-33
Ex-LE-3
LIV-9
ST-34
SP-10
Ex-LE-2
Upper patellar border
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
8 cun
ST-37
ST-38
Midpoint
ST-40
ST-39
8 cun
Highest prominence of the lateral malleolus
ST-41
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 cun
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 cun

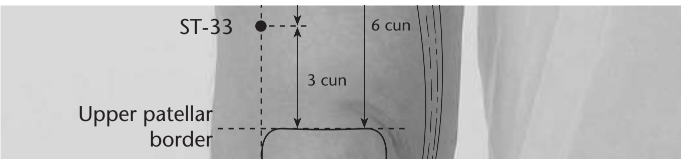
ST-33
6 cun
3 cun
Upper patellar border

ST-33 (*yinshi*) 'Yin Market' 3 cun superior to the upper lateral border of the patella

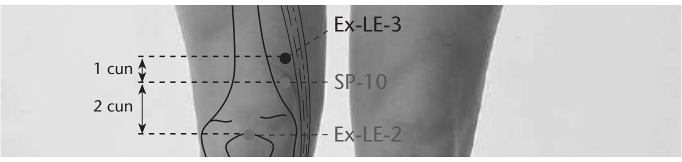
Ex-LE-3
1 cun
SP-10
2 cun
Ex-LE-2

Ex-LE-3 (*baichongwo*) 'Hundred Insect Burrow' 3 cun superior and 1 cun medial to the upper medial border of the patella, in a small depression on the vastus medialis muscle or 1 cun superior to **SP-10**

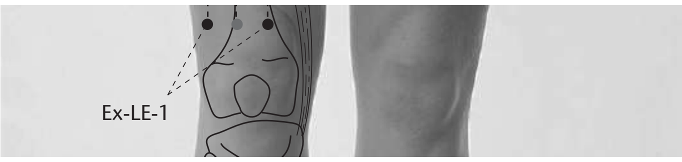

Ex-LE-1 (*kuangu*) 'Hip Bone' A pair of points 2 cun superior to the patella and 1.5 cun lateral and medial to **ST-34**

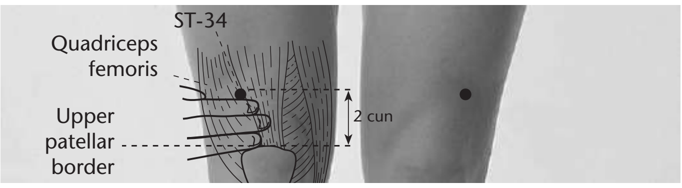
ST-34
Quadriceps femoris
Upper patellar border
2 cun

ST-34 (*liangqiu*) 'Ridge Mound' 2 cun proximal to the upper lateral border of the patella, in a depression on vastus lateralis

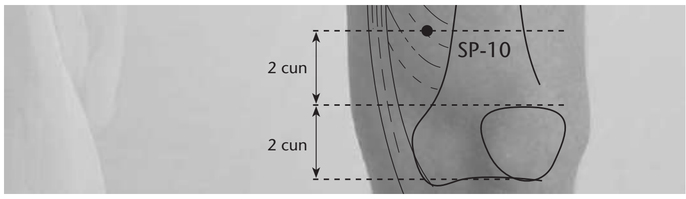
SP-10
2 cun
2 cun

SP-10 (*xuehai*) 'Sea of Blood' With the knee flexed, 2 cun proximal to the medial superior border of the patella. To find: place the right palm on the patient's left patella and locate the point at the tip of the thumb

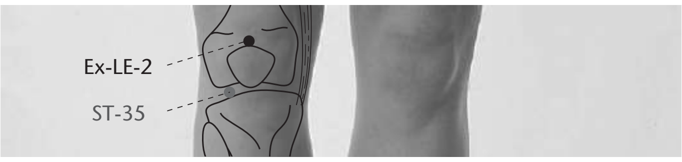

Ex-LE-2 (*heding/xiding*) 'Crane's Summit' With the knee flexed, at the midpoint of the upper patellar border

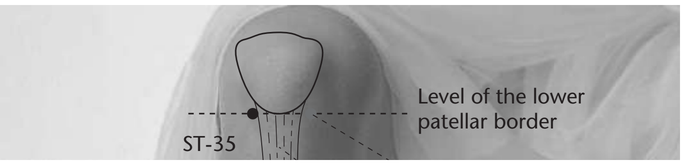
Level of the lower patellar border
ST-35

ST-35 (*dubi*) 'Calf's Nose' With the knee flexed, in the depression inferior to the patella and lateral to the patellar ligament; corresponds to the lateral eye of the knee; **Ex-LE-5** (*xiyan*) is the combination of this point and **Ex-LE-4**

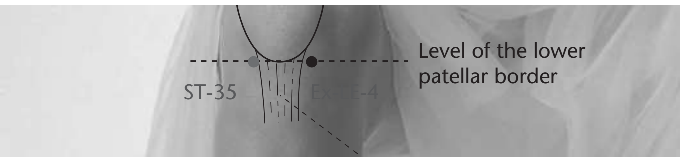
Level of the lower patellar border
ST-35

Ex-LE-4 (*neixiyan*) 'Inner Eye of the Knee' Inferior to the patella, in a depression medial to the patellar ligament; **Ex-LE-5** (*xiyan*) is the combination of this point and **ST-35**

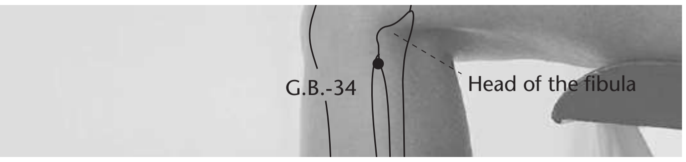
G.B.-34
Head of the fibula

G.B.-34 (*yanglingquan*) 'Yang Mound Spring' In the depression anterior and inferior to the head of the fibula

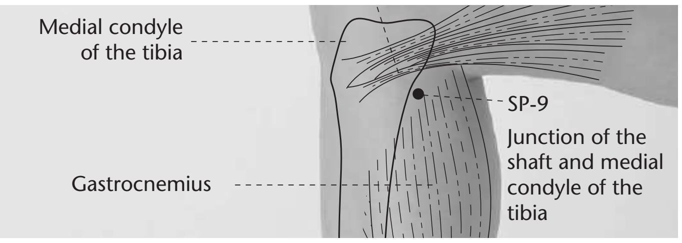
Medial condyle of the tibia
Gastrocnemius
SP-9
Junction of the shaft and medial condyle of the tibia

SP-9 (*yinlingquan*) 'Yin Mound Spring' With the knee flexed in a depression distal to the medial condyle of the tibia. To find: palpate from the medial posterior border of the tibia towards proximal, at roughly the same level as **G.B.-34**

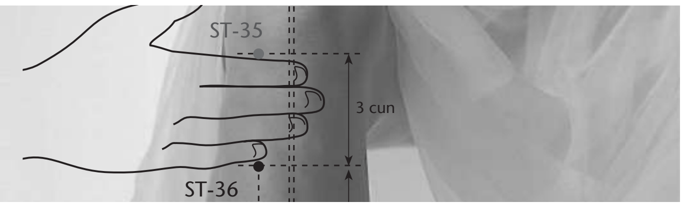
ST-35
3 cun
ST-36

ST-36 (*zusanli*) 'Leg Three Miles' one fingerbreadth (middle finger) lateral to the anterior crest of the tibia, 3 cun distal to **ST-35**

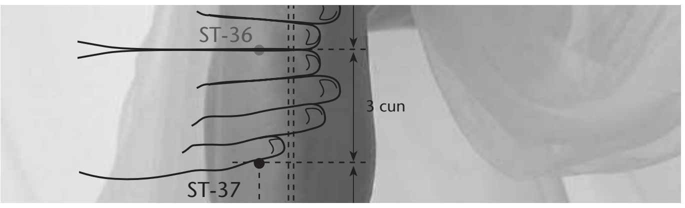
ST-36
3 cun
ST-37

ST-37 (*shangjuxu*) 'Upper Great Void' 1 fingerbreadth lateral to the anterior crest of the tibia, 6 cun distal to **ST-35** or 3 cun distal to **ST-36**

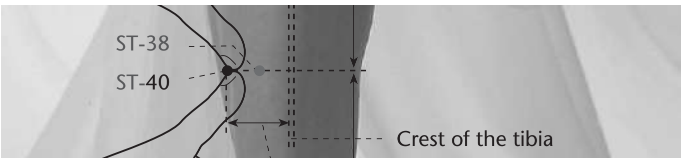
ST-38
ST-40
Crest of the tibia

ST-40 (*fenglong*) 'Abundant Bulge' 2 fingerbreadths lateral to the anterior crest of the tibia, at the midpoint of the line joining **ST-35** ('lateral eye of the knee', level with the knee joint space) and the highest prominence of the lateral malleolus

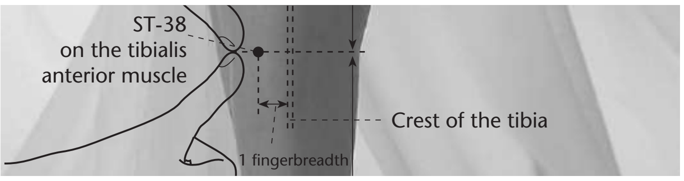
ST-38 on the tibialis anterior muscle
Crest of the tibia
1 fingerbreadth

ST-38 (*tiaokou*) 'Lines Opening' 1 fingerbreadth lateral to the anterior crest of the tibia, at the midpoint of the line joining **ST-35** ('lateral eye of the knee', level with the knee joint space) and the highest prominence of the lateral malleolus

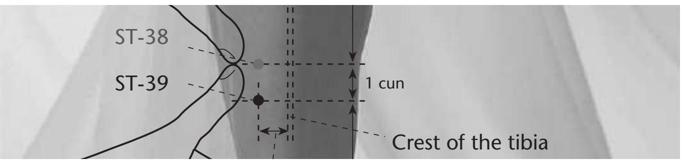
ST-38
ST-39
1 cun
Crest of the tibia

ST-39 (*xiajuxu*) 'Lower Great Void' 1 cun distal to **ST-38**

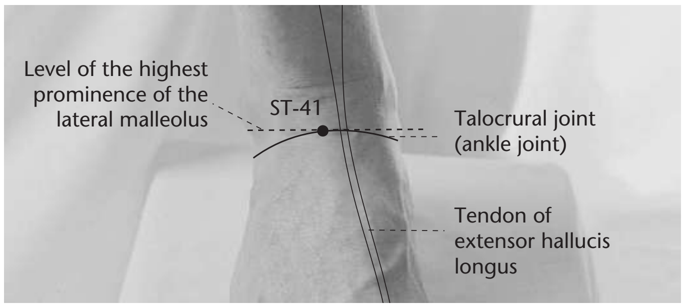
Level of the highest prominence of the lateral malleolus
ST-41
Talocrural joint (ankle joint)
Tendon of extensor hallucis longus

ST-41 (*jiexi*) 'Stream Divide' On the ankle, between the tendons of the extensor digitorum and the extensor hallucis longus, at the level of the highest prominence of the lateral malleolus

7.11 Medial Leg

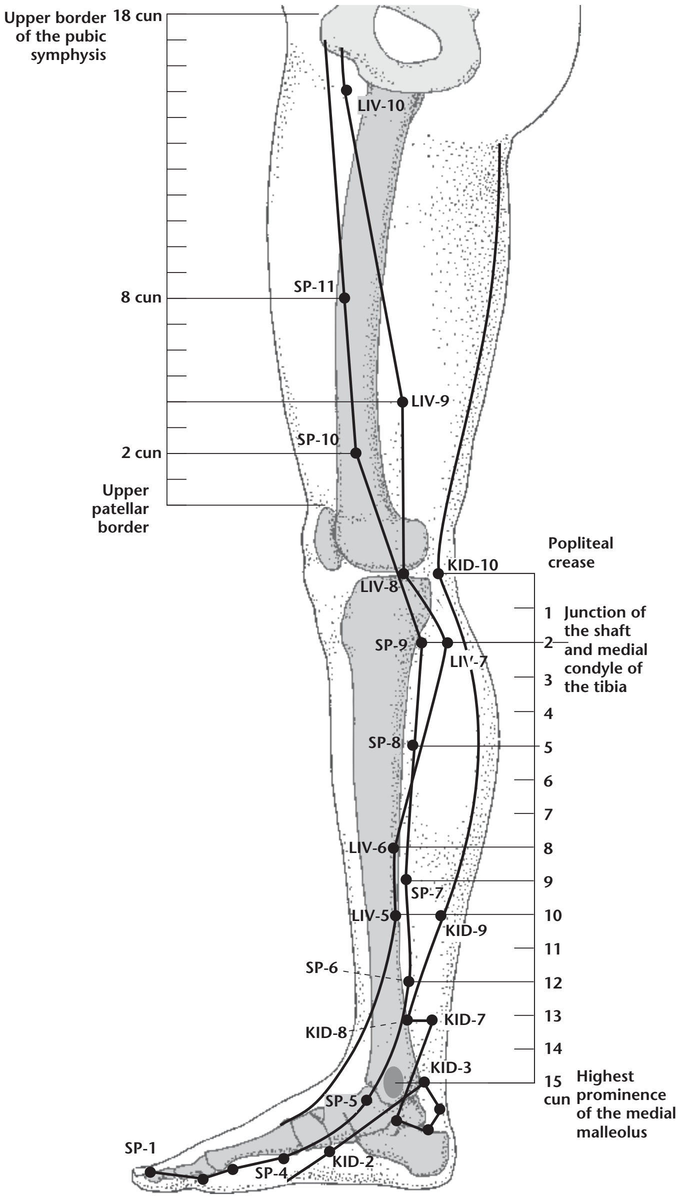
Upper border of the pubic symphysis
18 cun
8 cun
2 cun
Upper patellar border
LIV-10
SP-11
LIV-9
SP-10
Popliteal crease
KID-10
LIV-8
SP-9
LIV-7
1
Junction of the shaft and medial condyle of the tibia
2
3
4
5
6
7
8
9
10
11
12
13
14
15 cun
SP-8
LIV-6
SP-7
LIV-5
KID-9
SP-6
KID-8
KID-7
KID-3
SP-5
Highest prominence of the medial malleolus
SP-1
SP-4
KID-2

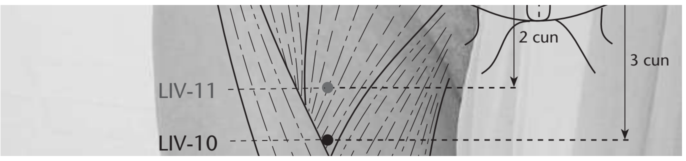
2 cun
3 cun
LIV-11
LIV-10

LIV-10 (*zuwuli*) 'Leg Five Miles' On the anterior aspect of the thigh, 3 cun inferior to the upper border of the symphysis, on the lateral border of the adductor longus

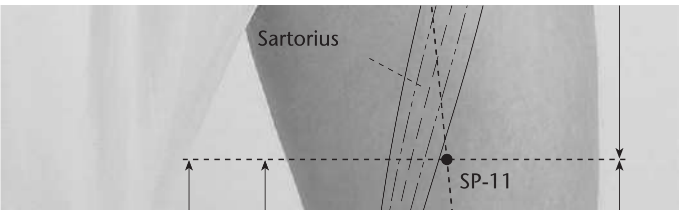
Sartorius
SP-11

SP-11 (*jimen*) 'Winnowing Gate' 6 cun proximal to **SP-10** at the midpoint of the femur, in the depression between the sartorius and vastus lateralis

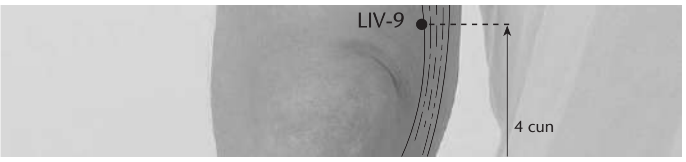

LIV-9 (*yinbao*) 'Yin Wrapping' 4 cun proximal to the medial condyle of the femur, between the sartorius and vastus medialis

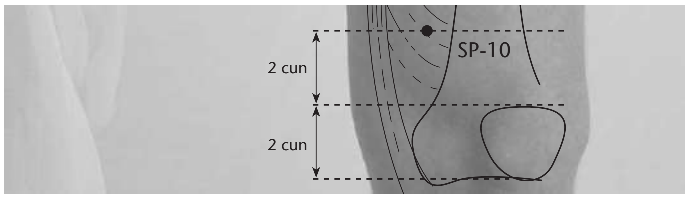
2 cun
SP-10
2 cun

SP-10 (*xuehai*) 'Sea of Blood' With the knee flexed, 2 cun proximal to the medial superior border of the patella. To find: place the right palm on the patient's left patella and locate the point at the tip of the thumb

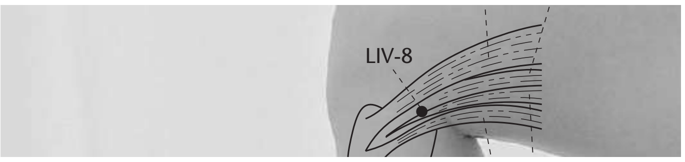

LIV-8 (*ququan*) 'Spring at the Crook' With the knee flexed, at the medial end of the popliteal crease, in the depression anterior to the tendons of semitendinosus and semimembranosus

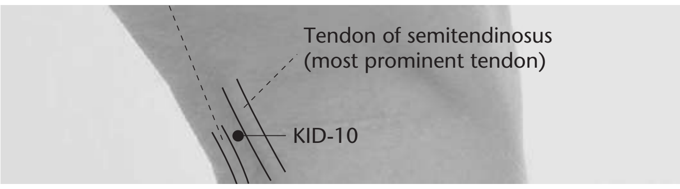
Tendon of semitendinosus (most prominent tendon)
KID-10

KID-10 (*yingu*) 'Yin Valley' At the medial end of the popliteal crease, between the tendons of semimembranosus and semitendinosus, at the level of the knee joint space

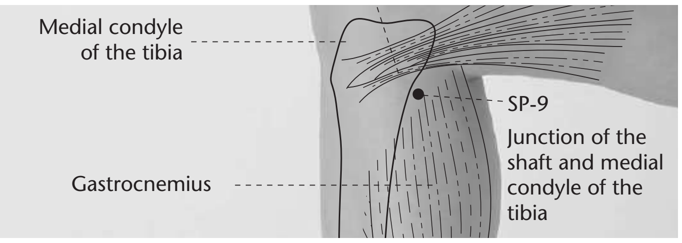
Medial condyle of the tibia
SP-9
Junction of the shaft and medial condyle of the tibia
Gastrocnemius

SP-9 (*yinlingquan*) 'Yin Mound Spring' With the knee flexed, in a depression distal to the medial condyle of the tibia. To find: palpate from the medial posterior border of the tibia towards proximal, at roughly the same level as **G.B.-34**

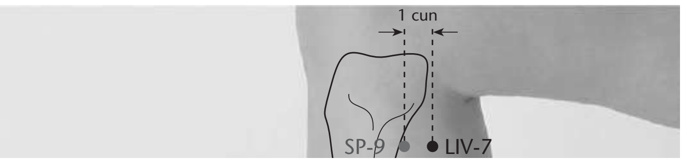
1 cun
SP-9
LIV-7

LIV-7 (*xiguan*) 'Knee Joint' At the junction of the shaft and the medial condyle of the tibia, 1 cun posterior to **SP-9**

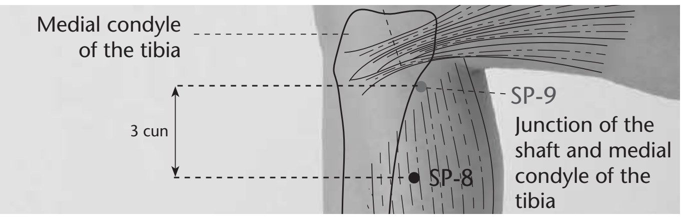
Medial condyle of the tibia
SP-9
Junction of the shaft and medial condyle of the tibia
3 cun
SP-8

SP-8 (*diji*) 'Earth Pivot' 3 cun distal to **SP-9** (at the junction of the shaft and the medial condyle of the tibia) on the posterior border of the tibia

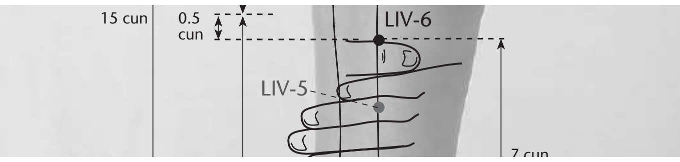
15 cun
0.5 cun
LIV-6
LIV-5
7 cun

LIV-6 (*zhongdu*) 'Central Capital' 7 cun proximal to the highest prominence of the medial malleolus, just posterior to the medial crest of the tibia

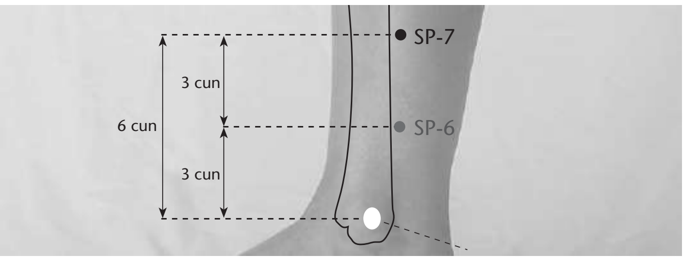
SP-7
3 cun
6 cun
SP-6
3 cun

SP-7 (*lougu*) 'Dripping Valley' 6 cun proximal to the highest prominence of the medial malleolus or 3 cun proximal to **SP-6**, on the medial posterior border of the tibia

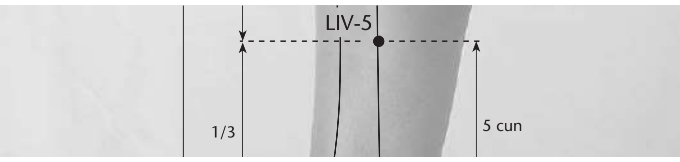
LIV-5
1/3
5 cun

LIV-5 (*ligou*) 'Woodworm Canal' 5 cun proximal to the highest prominence of the medial malleolus, just posterior to the medial crest of the tibia

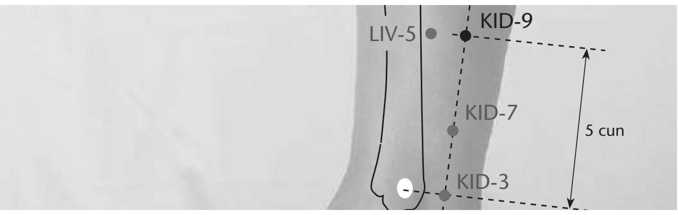
LIV-5
KID-9
KID-7
5 cun
KID-3

KID-9 (*zhubin*) 'Guest House' 5 cun proximal to KID-3 and 2 cun posterior to the medial border of the tibia

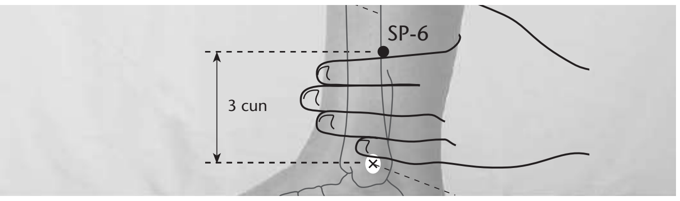

SP-6 (*sanyinjiao*) 'Three Yin Intersection' 3 cun proximal to the highest prominence of the medial malleolus, posterior border to the medial border of the tibia

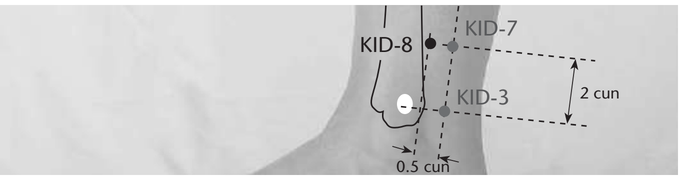
KID-8
KID-7
2 cun
KID-3
0.5 cun

KID-8 (*jiaoxin*) 'Exchange Belief' 2 cun proximal to the highest prominence of the medial malleolus, posterior to the medial border of the tibia

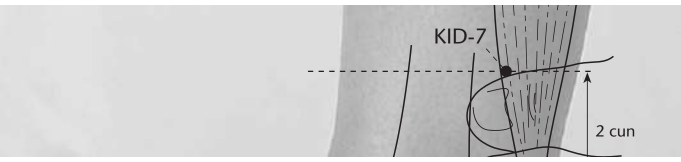

KID-7 (*fuliu*) 'Returning Current' 2 cun proximal to KID-3, in a depression on the anterior border of the Achilles tendon

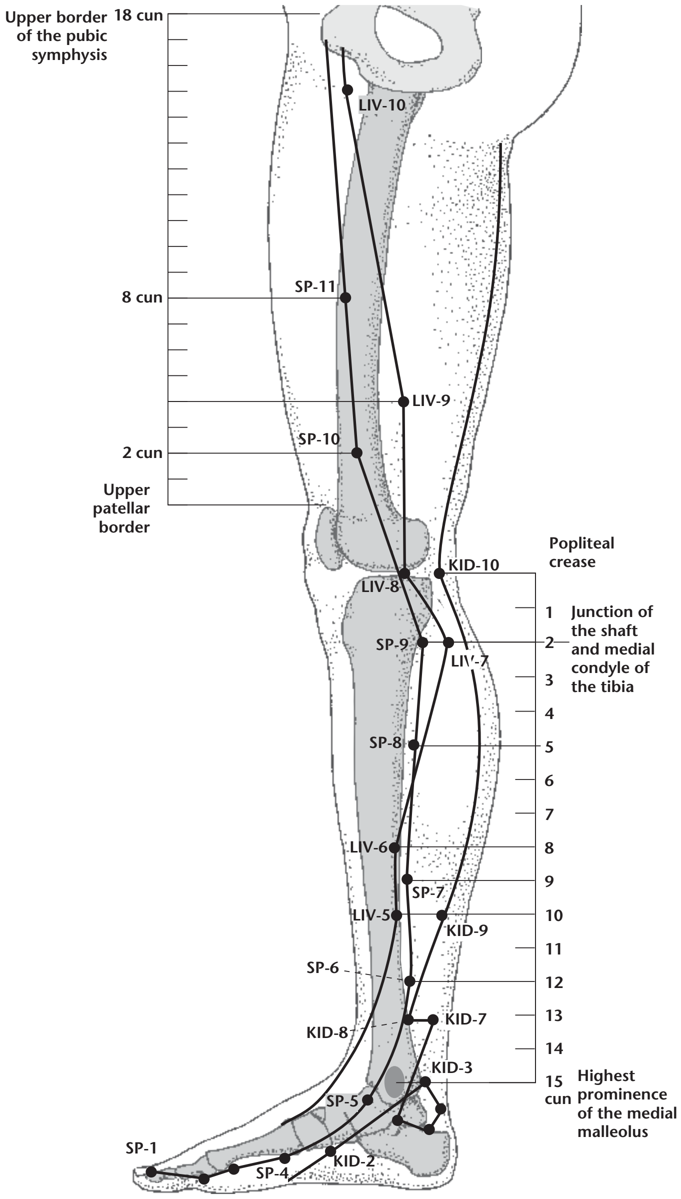
Upper border of the pubic symphysis
18 cun
LIV-10
8 cun
SP-11
LIV-9
SP-10
2 cun
Upper patellar border
Popliteal crease
KID-10
LIV-8
1
Junction of the shaft and medial condyle of the tibia
SP-9
2
LIV-7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
LIV-5
10
KID-9
11
SP-6
12
KID-8
KID-7
13
14
KID-3
15 cun
Highest prominence of the medial malleolus
SP-5
SP-1
KID-2
SP-4

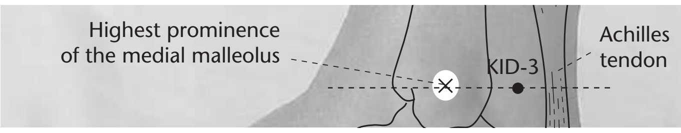
Highest prominence of the medial malleolus
KID-3
Achilles tendon

KID-3 (*taixi*) 'Supreme Stream' In the depression between the highest prominence of the medial malleolus and the Achilles tendon

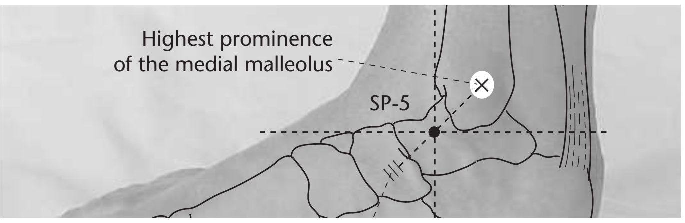
Highest prominence of the medial malleolus
SP-5

SP-5 (*shangqiu*) 'Shang Mound' In the depression at the junction of a vertical line along the anterior border and a horizontal line along the lower border of the medial malleolus. Or: in the depression halfway between the highest prominence of the medial malleolus and the navicular tuberosity

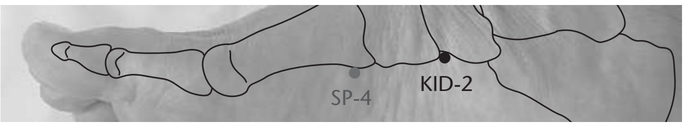

KID-2 (*rangu*) 'Blazing Valley' In a depression at the anterior inferior border of the navicular bone, at the border of the red and white skin

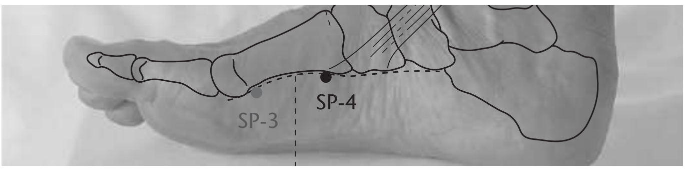

SP-4 (*gongsun*) 'Grandfather Grandson' In a depression distal and inferior to the base of the 1st metatarsal bone, at the border of the red and white skin

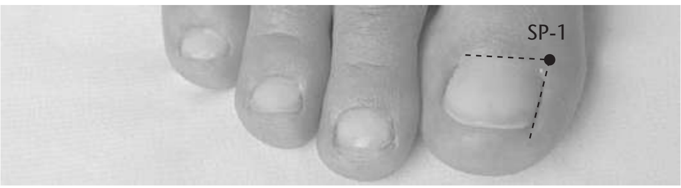

SP-1 (*yinbai*) 'Hidden White' On the big toe, 0.1 cun from the medial corner of the nail

7.12 Lateral Leg

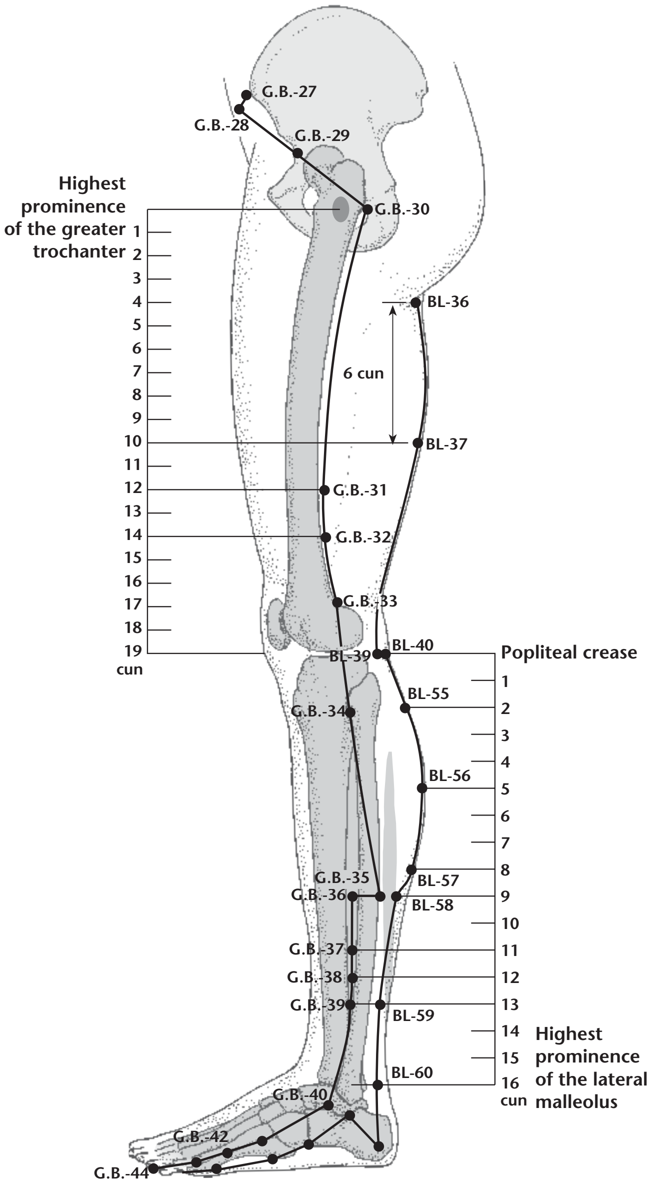
G.B.-27
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-40
BL-39
Popliteal crease
G.B.-34
BL-55
BL-56
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
Highest prominence of the lateral malleolus
G.B.-40
G.B.-42
G.B.-44
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 cun
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 cun

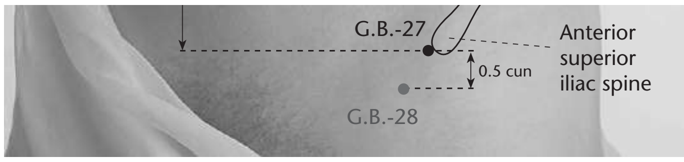
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

G.B.-27 (*wushu*) 'Five Pivots' In the depression medial to the anterior superior iliac spine, approximately level with **Ren-4**

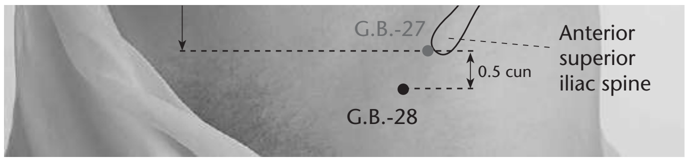
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

G.B.-28 (*weidao*) 'Linking Path' Anterior and inferior to the anterior superior iliac spine, 0.5 cun anterior and inferior to **G.B.-27**

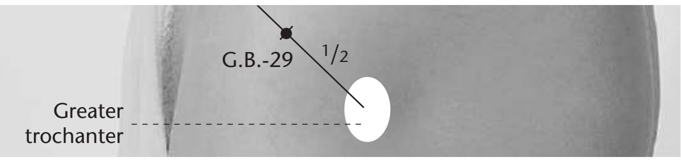
G.B.-29
1/2
Greater trochanter

G.B.-29 (*juliao*) 'Stationary Crevice' At the midpoint of a line connecting the anterior superior iliac spine and the greater trochanter, at the anterior border of the iliac crest

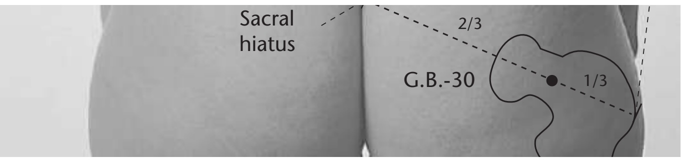
Sacral hiatus
2/3
G.B.-30
1/3

G.B.-30 (*huantiao*) 'Jumping Circle' With the patient lying on his/her side, at the junction between the medial two-thirds and the lateral third on a line connecting the greater trochanter and the sacral hiatus

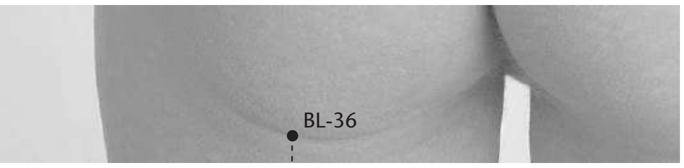

BL-36 (*chengfu*) 'Hold and Support' In the gluteal crease, superior to the midpoint of the popliteal crease

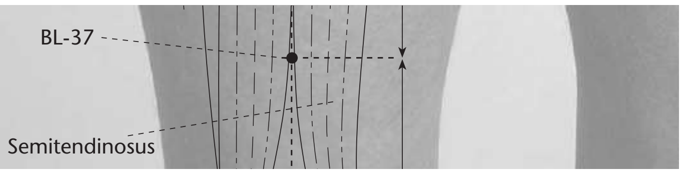
BL-37
Semitendinosus

BL-37 (*yinmen*) 'Gate of Abundance' 6 cun distal to **BL-36** (gluteal crease), on a line connecting BL-36 and BL-40 (centre of the popliteal crease), in a gap in the musculature

Ch07-F10028.qxd 2/22/08 10:09 PM Page 654

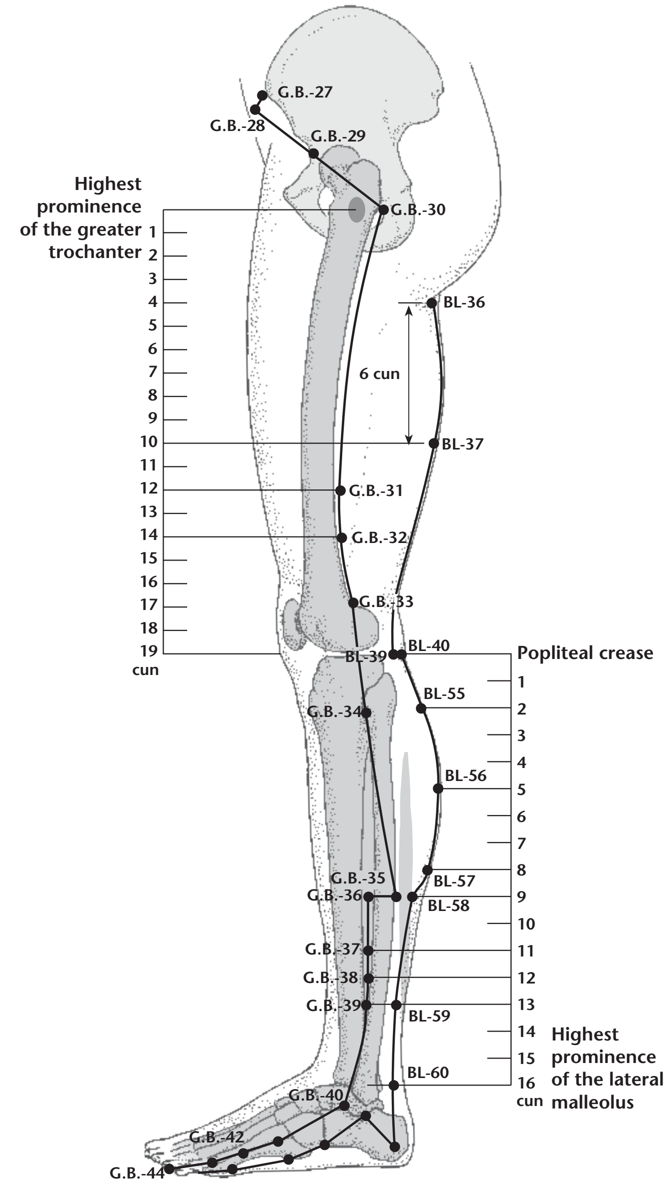
G.B.-27
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-40
BL-39
Popliteal crease
BL-55
G.B.-34
BL-56
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
G.B.-40
G.B.-42
G.B.-44
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus

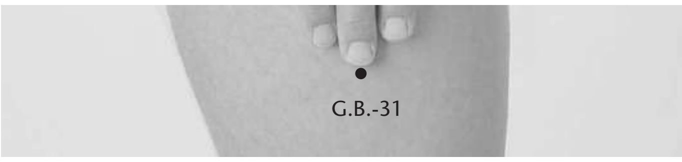

G.B.-31 (*fengshi*) 'Wind Market' On the lateral aspect of the thigh, approximately 7 cun proximal to the popliteal crease. With the patient standing upright, the point is where the patient's middle fingertip touches the thigh

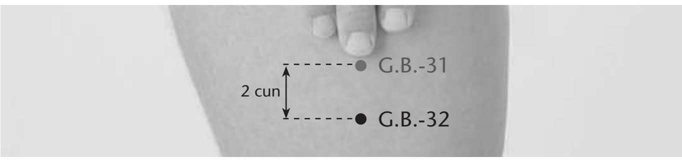
G.B.-31
2 cun
G.B.-32

G.B.-32 (*zhongdu*) 'Middle Ditch' On the lateral aspect of the thigh, 2 cun inferior to **G.B.-31** or 5 cun superior to the popliteal crease, between vastus lateralis and biceps femoris

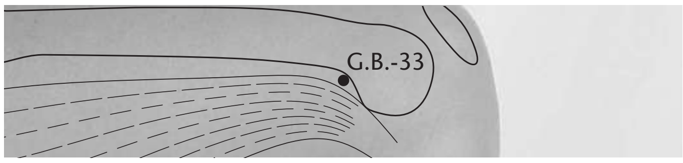

G.B.-33 (*xiyangguan*) 'Knee Yang Gate' On the lateral aspect of the knee, with the knee flexed, in the depression between the lateral epicondyle of the femur and the tendon of biceps femoris, approximately 3 cun proximal to **G.B.-34**

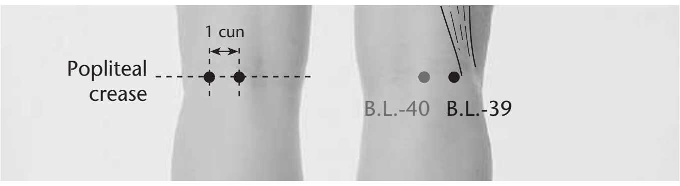
1 cun
Popliteal crease
B.L.-40
B.L.-39

BL-39 (*weiyang*) 'Outside of the Crook' At the lateral end of the popliteal crease, on the medial side of the tendon of the long head of the biceps femoris, 1 cun lateral to **BL-40** (in the centre of the popliteal crease).

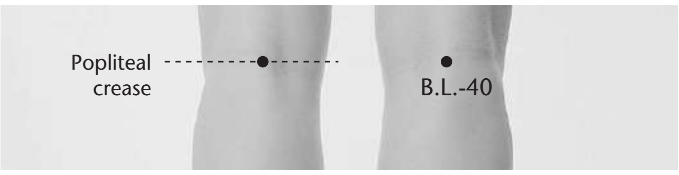
Popliteal crease
B.L.-40

BL-40 (*weizhong*) 'Middle of the Crook' In the centre of the popliteal crease, between the tendons of the biceps femoris and semitendinosus.

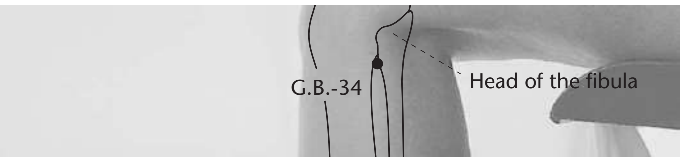
G.B.-34
Head of the fibula

G.B.-34 (*yanglingquan*) 'Yang Mound Spring' In the depression anterior and inferior to the head of the fibula

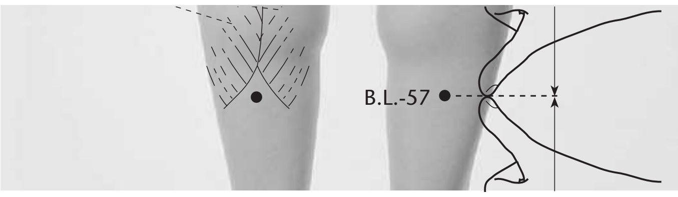

BL-57 (*chengshan*) 'Support of the Mountain' In the middle of the calf, between the two heads of the gastrocnemius, on a line connecting **BL-40** and **BL-60**, approximately 8 cun distal to **BL-40** (spreading hands technique → 2)

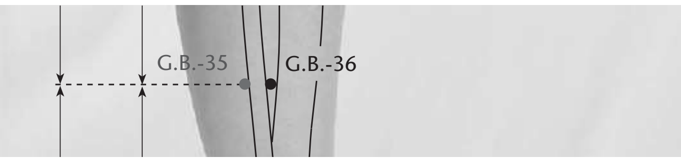

G.B.-36 (*waiqiu*) 'Outer Hill' 7 cun proximal to the highest prominence of the lateral malleolus, on the anterior border of the fibula

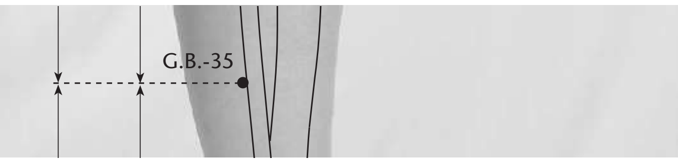

G.B.-35 (*yangjiao*) 'Yang Intersection' 7 cun proximal to the highest prominence of the lateral malleolus, on the posterior border of the fibula

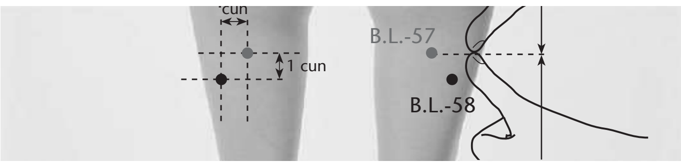
cun
1 cun
B.L.-57
B.L.-58

BL-58 (*feiyang*) 'Soaring Upwards' 1 cun distal and 1 cun lateral to **BL-57** or 7 cun proximal to **BL-60**, at the posterior border of the fibula and on the lower border of the gastrocnemius

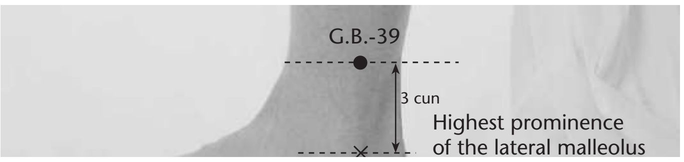
G.B.-39
3 cun
Highest prominence of the lateral malleolus

G.B.-39 (*xuanzhong*) 'Suspended Bell' 3 cun proximal to the highest prominence of the lateral malleolus, on the anterior border of the fibula

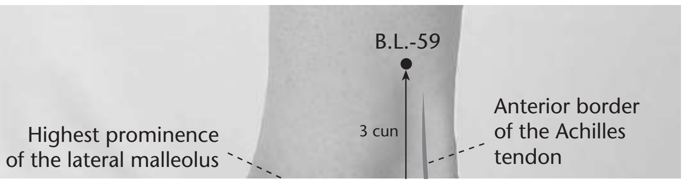
B.L.-59
3 cun
Highest prominence of the lateral malleolus
Anterior border of the Achilles tendon

BL-59 (*fuyang*) 'Instep Yang' 3 cun superior to **BL-60** (in the depression between the highest prominence of the lateral malleolus and the Achilles tendon)

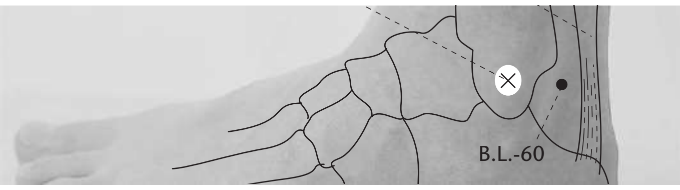

BL-60 (*kunlun*) 'Kunlun Mountains' In the depression between the Achilles tendon and the highest prominence of the lateral malleolus

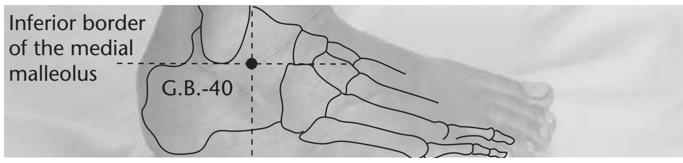
Inferior border of the medial malleolus
G.B.-40

G.B.-40 (*qiuxu*) 'Large Mound' In the depression anterior and inferior to the lateral malleolus, lateral to the tendons of the extensor digitorum longus muscle.

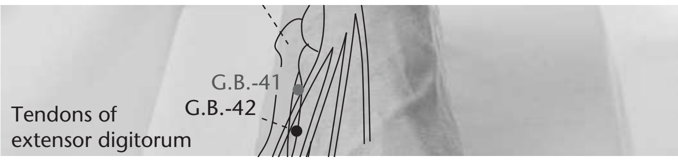
G.B.-41
G.B.-42
Tendons of extensor digitorum

G.B.-42 (*diwuhui*) 'Earth Five Meetings' Between the 4th and 5th metatarsal bones, proximal to the metatarsophalangeal joints and medial to the tendon of extensor digitorum longus

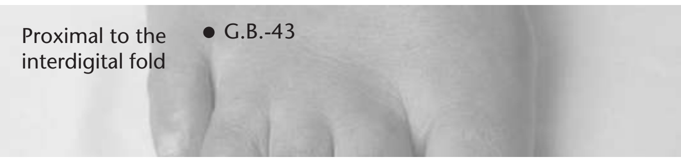
Proximal to the interdigital fold
G.B.-43

G.B.-43 (*jiaxi/xiaxi*) 'Clamped Stream' Between the 4th and the 5th toes, proximal to the margin of the interdigital web

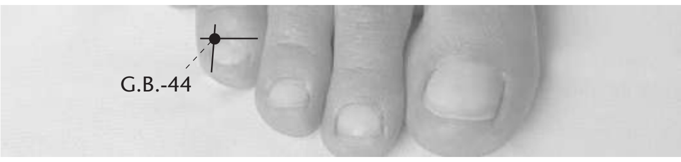

G.B.-44 (*zuqiaoyin*) 'Yin Portals of the Foot' On the 4th toe, 0.1 cun from the lateral corner of the nail

7.13 Dorsum of the Foot

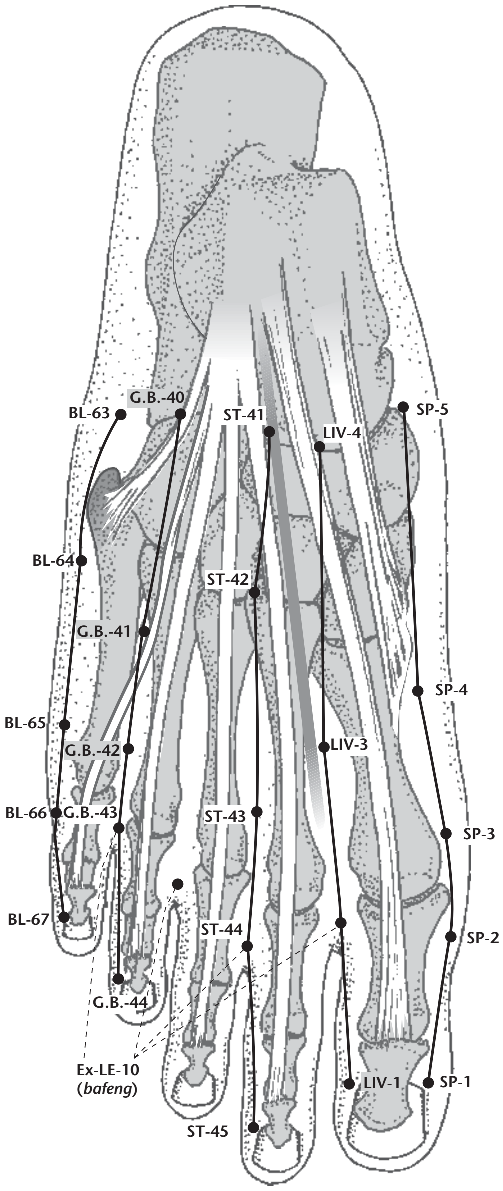
G.B.-40
BL-63
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
BL-66
G.B.-43
ST-43
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

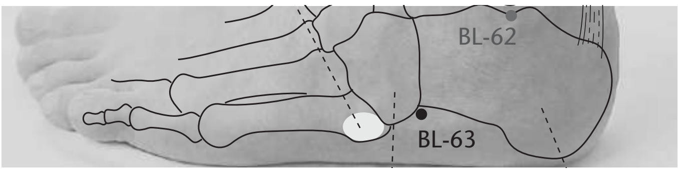

BL-63 (*jinmen*) 'Golden Gate' In a depression anterior and inferior to **BL-62** between the calcaneus and the cuboid bone, proximal to the tuberosity of the 5th metatarsal bone

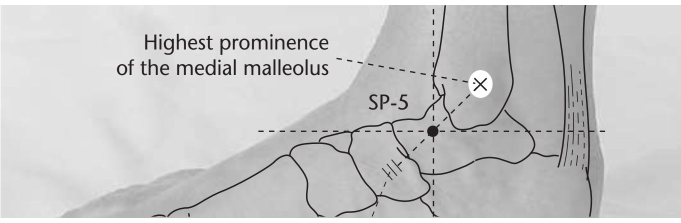
Highest prominence
of the medial malleolus
SP-5

SP-5 (*shangqiu*) 'Shang Mound' In the depression at the junction of a vertical line along the anterior border and a horizontal line along the lower border of the medial malleolus. Or: in the depression halfway between the highest prominence of the medial malleolus and the navicular tuberosity

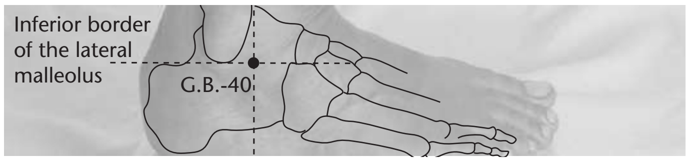
Inferior border
of the lateral
malleolus
G.B.-40

G.B.-40 (*qiuxu*) 'Large Mound' In the depression anterior and inferior to the lateral malleolus, lateral to the tendons of extensor digitorum longus muscle.

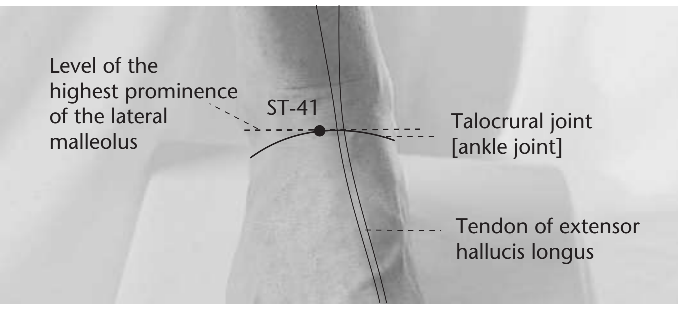
Level of the
highest prominence
of the lateral
malleolus
ST-41
Talocrural joint
[ankle joint]
Tendon of extensor
hallucis longus

ST-41 (*jiexi*) 'Stream Divide' On the ankle, between the tendons of extensor digitorum and extensor hallucis longus, at the level of the highest prominence of the lateral malleolus

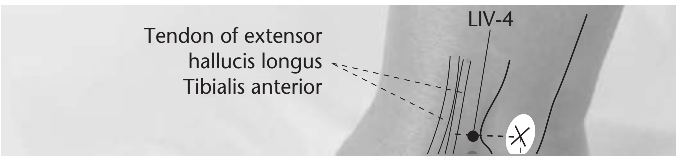
Tendon of extensor
hallucis longus
Tibialis anterior
LIV-4

LIV-4 (*zhongfeng*) 'Middle Seal' 1 cun anterior to the prominence of the medial malleolus, medial to the tendon of the tibialis anterior

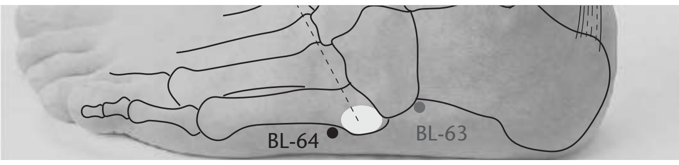

BL-64 (*jinggu*) 'Capital Bone' Distal to the tuberosity of the 5th metatarsal bone, at the border of the red and white skin

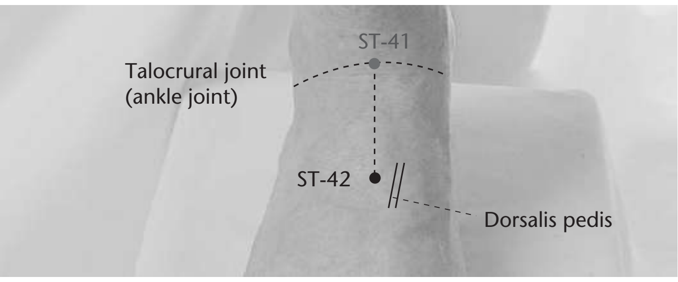
ST-41
Talocrural joint
(ankle joint)
ST-42
Dorsalis pedis

ST-42 (*chongyang*) 'Rushing Yang' On the highest point of the dorsum of the foot, between the tendons of extensor hallucis longus and extensor digitorum longus, directly lateral to where the dorsalis pedis artery may be palpated. The point is bordered proximally by the second and third metatarsal bones and distally by the second and third cuneiform bones. **Alternative location:** sometimes this point may be located lateral to the medial portion of the extensor digitorum longus tendon (joining the second toe)

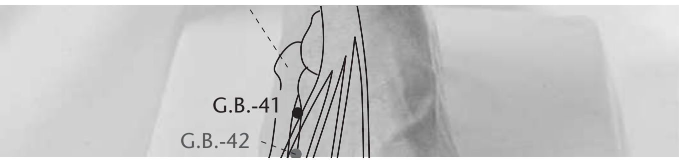

G.B.-41 (*zulinqi*) 'Foot Governor of Tears' In the depression at the junction of the shafts and the bases of the 4th and 5th metatarsal bones, lateral to the branch of the extensor digitorum longus tendon running to the little toe

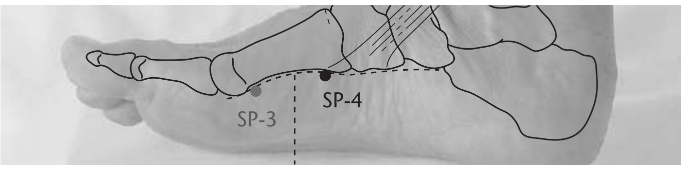

SP-4 (*gongsun*) 'Grandfather Grandson' In a depression distal and inferior to the base of the 1st metatarsal bone, at the border of the red and white skin

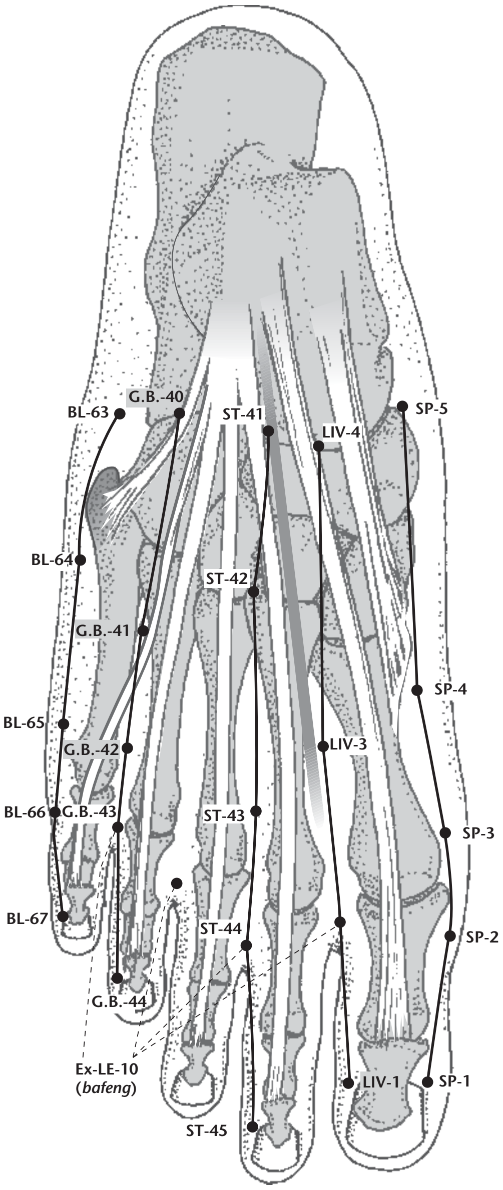
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
BL-66
G.B.-43
ST-43
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(bafeng)
LIV-1
SP-1
ST-45

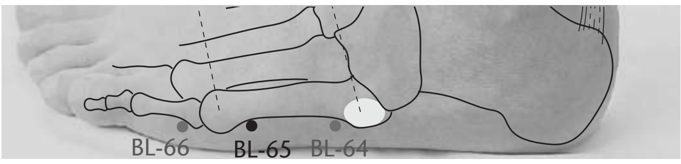
BL-66
BL-65
BL-64

BL-65 (*shugu*) 'Restraining Bone' In the depression proximal to the head of the 5th metatarsal bone, at the border of the red and white skin

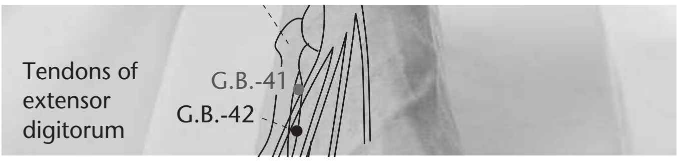
Tendons of extensor digitorum
G.B.-41
G.B.-42

G.B.-42 (*diwuhui*) 'Earth Five Meetings' Between the 4th and 5th metatarsal bones, proximal to the metatarsophalangeal joints and medial to the tendon of the extensor digitorum longus

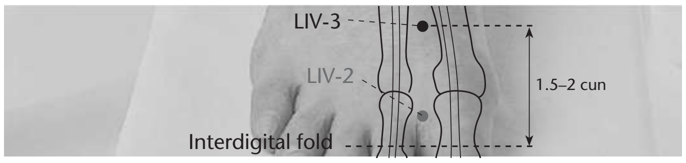
LIV-3
LIV-2
1.5–2 cun
Interdigital fold

LIV-3 (*taichong*) 'Great Rushing' On the dorsum of the foot, in the depression between the 1st and 2nd metatarsal bones, approximately 1.5–2 cun proximal to the interdigital web

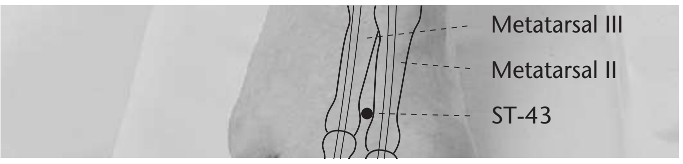
Metatarsal III
Metatarsal II
ST-43

ST-43 (*xiangu*) 'Sunken Valley' In the depression between the second and third metatarsal bones, at the junction of the shafts and the heads of these metatarsal bones

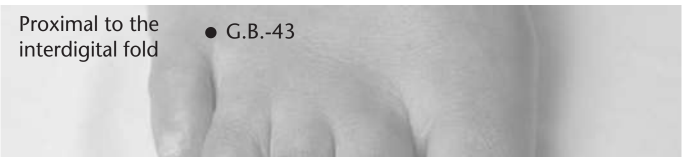
Proximal to the interdigital fold
G.B.-43

G.B.-43 (*xiaxi*) 'Clamped Stream' Between the 4th and the 5th toes, proximal to the margin of the interdigital web

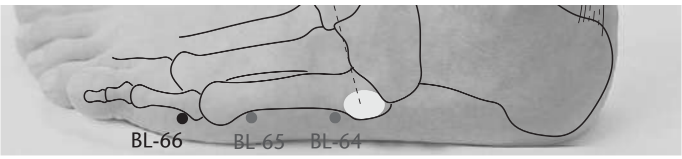
BL-66
BL-65
BL-64

BL-66 (*zutonggu*) 'Foot Connecting Valley' In a depression distal to the metatarsophalangeal joint of the little toe

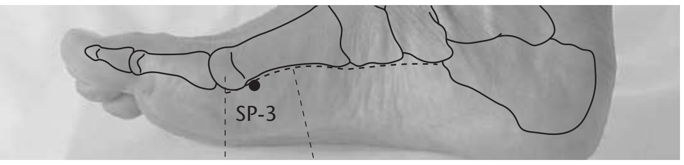

SP-3 (*taibai*) 'Supreme White' In the depression proximal to the head of the 1st metatarsal bone, at the border of the red and white skin

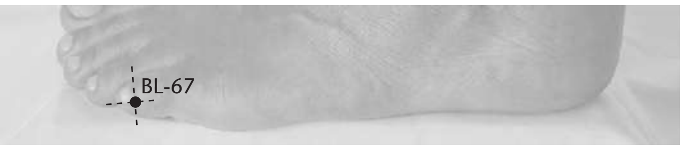

BL-67 (*zhiyin*) 'Reaching Yin' On the little toe, 0.1 cun from the lateral corner of the nail

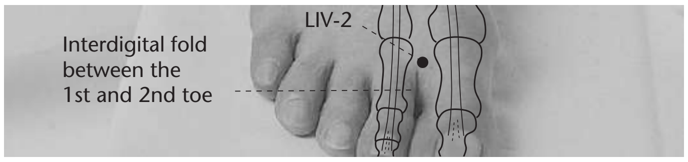

LIV-2 (*xingjian*) 'Moving Between' Between the 1st and 2nd toe, proximal to the margin of the interdigital web

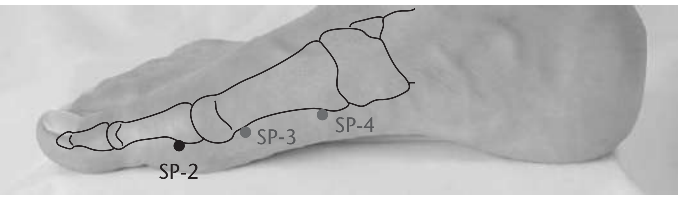

SP-2 (*dadu*) 'Great Metropolis' On the medial aspect of the big toe, in the depression distal to the first metatarsophalangeal joint, on the border of the red and white skin

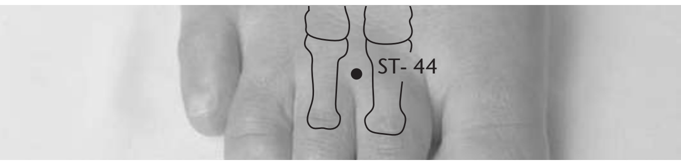

ST-44 (*neiting*) 'Inner Court' Between the 2nd and the 3rd toe, proximal to the interdigital fold

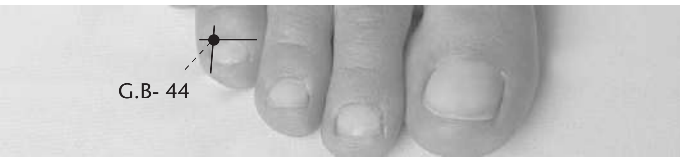

G.B.-44 (*zuqiaoyin*) 'Yin Portals of the Foot' On the 4th toe, 0.1 cun from the lateral corner of the nail

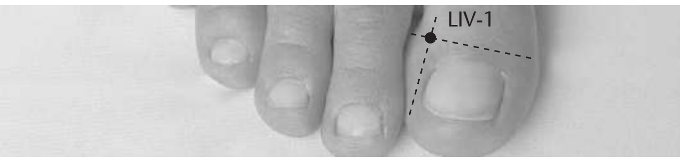

LIV-1 (*dadun*) 'Big Mound' On the lateral aspect of the big toe, 0.1 cun proximal and lateral to the corner of the nail

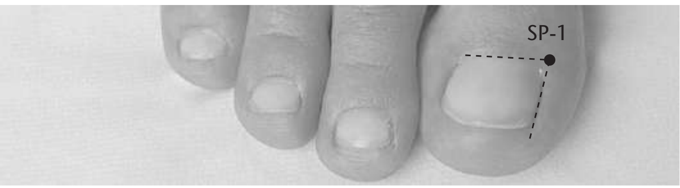

SP-1 (*yinbai*) 'Hidden White' On the big toe, 0.1 cun from the medial corner of the nail

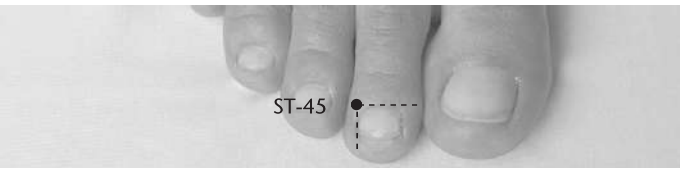

ST-45 (*lidui*) 'Strict Exchange' On the 2nd toe, 0.1 cun from the lateral corner of the nail

Ch07-F10028.qxd 2/22/08 10:10 PM Page 660

7.14 Medial Foot

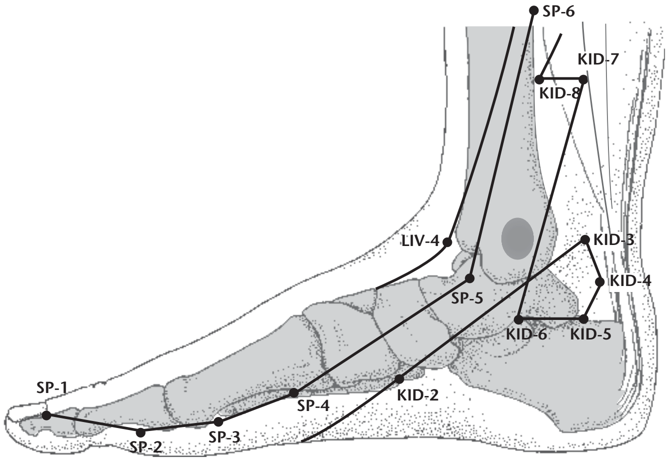
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

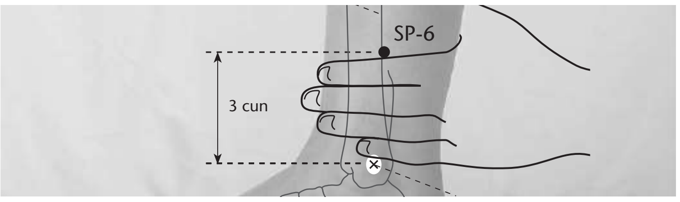

SP-6 (*sanyinjiao*) 'Three Yin Intersection' 3 cun proximal to the highest prominence of the medial malleolus, on the posterior border of the medial crest of the tibia

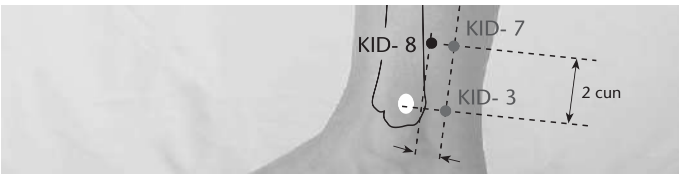
KID- 8
KID- 7
KID- 3
2 cun

KID-8 (*jiaoxin*) 'Exchange Belief' 2 cun proximal to the highest prominence of the medial malleolus, posterior to the medial border of the tibia.

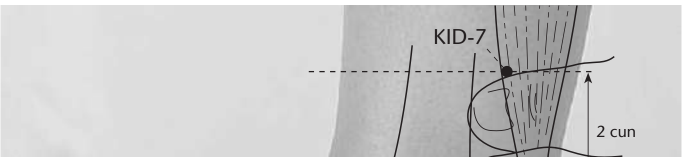

KID-7 (*fuliu*) 'Returning Current' 2 cun proximal to **KID-3**, in a depression on the anterior border of the Achilles tendon

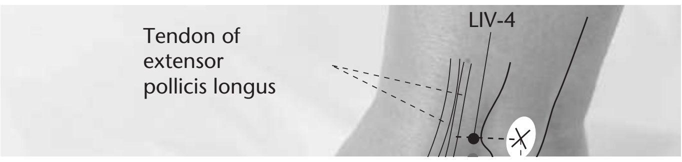
Tendon of
extensor
pollicis longus
LIV-4

LIV-4 (*zhongfeng*) 'Middle Seal' 1 cun anterior to the prominence of the medial malleolus, medial to the tendon of tibialis anterior

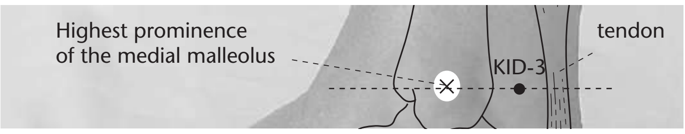
Highest prominence
of the medial malleolus
tendon
KID-3

KID-3 (*taixi*) 'Supreme Stream' In the depression between the highest prominence of the medial malleolus and the Achilles tendon

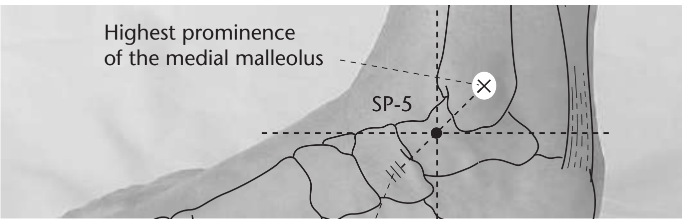
Highest prominence
of the medial malleolus
SP-5

SP-5 (*shangqiu*) 'Shang Mound' In the depression at the junction of a vertical line along the anterior border and a horizontal line along the lower border of the medial malleolus. Or: in the depression halfway between the highest prominence of the medial malleolus and the navicular tuberosity

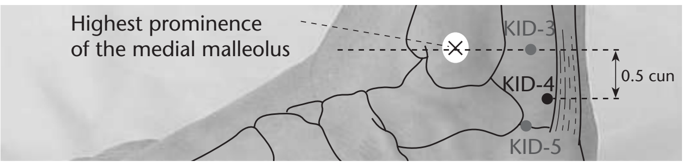
Highest prominence
of the medial malleolus
KID-3
KID-4
KID-5
0.5 cun

KID-4 (*dazhong*) 'Great Bell' Anterior to the medial border of the Achilles tendon, superior to its insertion at the calcaneus

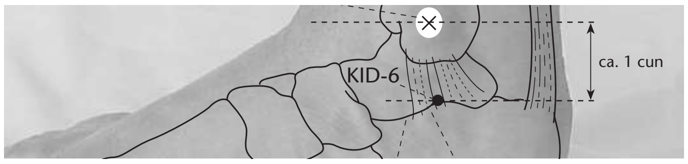
KID-6
ca. 1 cun

KID-6 (*zhaohai*) 'Shining Sea' In the depression inferior to the highest prominence of the medial malleolus, over the joint space between the talus and the calcaneus

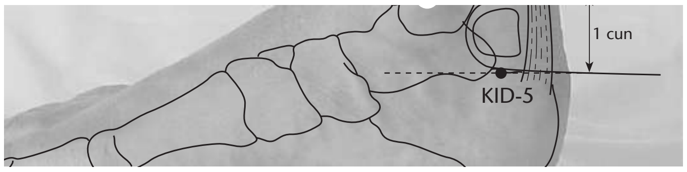

KID-5 (*shuiquan*) 'Water Spring' 1 cun distal to **KID-3**, in a depression over the joint space between the talus and calcaneus

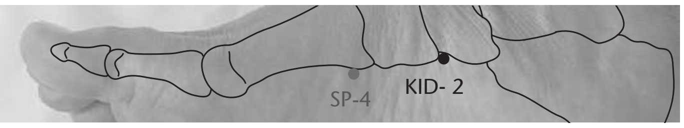

KID-2 (*rangu*) 'Blazing Valley' In a depression at the anterior inferior border of the navicular bone, at the border of the red and white skin

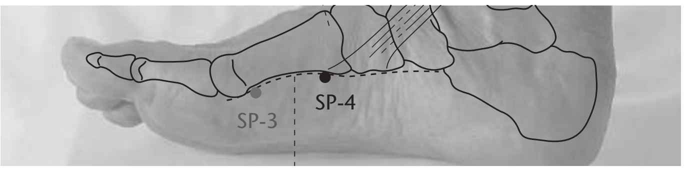

SP-4 (*gongsun*) 'Grandfather Grandson' In a depression distal and inferior to the base of the 1st metatarsal bone, at the border of the red and white skin

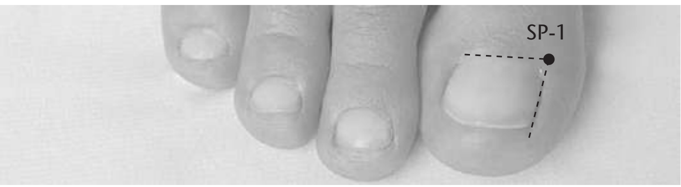

SP-1 (*yinbai*) 'Hidden White' On the big toe, 0.1 cun from the medial corner of the nail

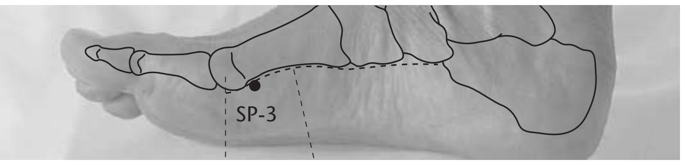

SP-3 (*taibai*) 'Supreme White' In the depression proximal to the head of the 1st metatarsal bone, at the border of the red and white skin

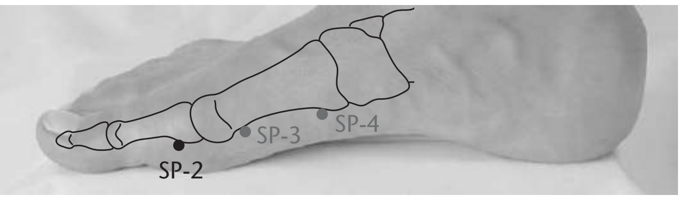

SP-2 (*dadu*) 'Great Metropolis' On the medial aspect of the big toe, in the depression distal to the first metatarsophalangeal joint, on the border of the red and white skin

7.15 Lateral Foot

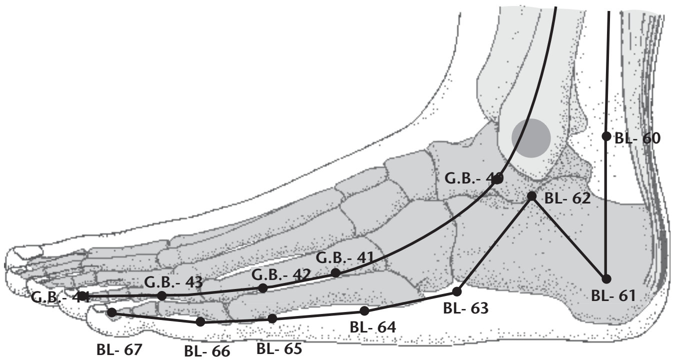
BL- 60
G.B.- 40
BL- 62
G.B.- 41
G.B.- 43
G.B.- 42
G.B.- 44
BL- 61
BL- 63
BL- 64
BL- 67
BL- 66
BL- 65

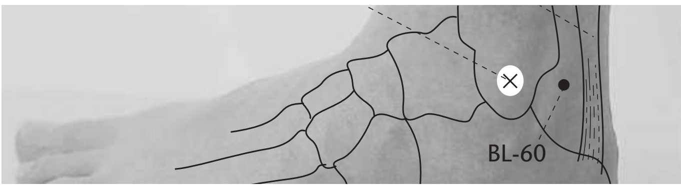

BL-60 (*kunlun*) 'Kunlun Mountains' In the depression between the Achilles tendon and the highest prominence of the lateral malleolus

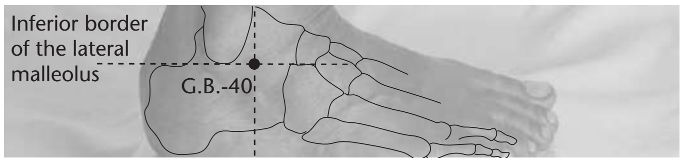
Inferior border of the lateral malleolus
G.B.-40

G.B.-40 (*qiuxu*) 'Large Mound' In the depression anterior and inferior to the medial malleolus, lateral to the tendons of extensor digitorum longus muscle.

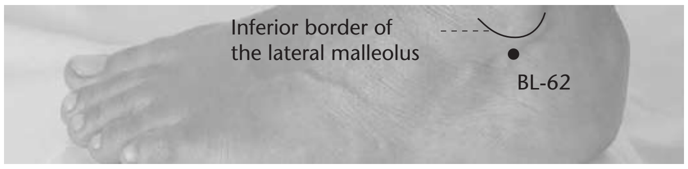
Inferior border of the lateral malleolus
BL-62

BL-62 (*shenmai*) 'Extending Vessel' In a depression directly inferior to the highest prominence of the lateral malleolus, over the joint space between the talus and calcaneus

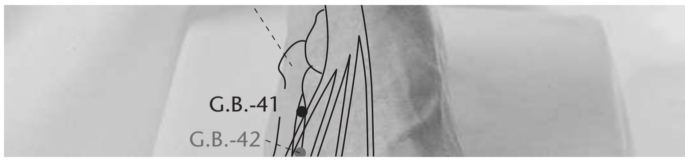

G.B.-41 (*zulinqi*) 'Foot Governor of Tears' In the depression at the junction of the shafts and the bases of the 4th and 5th metatarsal bones, lateral to the branch of the extensor digitorum longus tendon running to the little toe

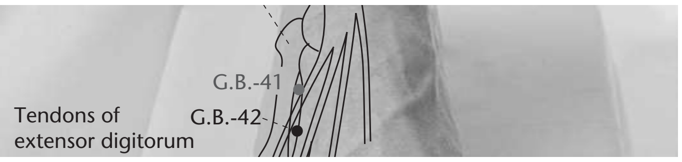
G.B.-41
Tendons of extensor digitorum
G.B.-42

G.B.-42 (*diwuhui*) 'Earth Five Meetings' Between the 4th and 5th metatarsal bones, proximal to the metatarsophalangeal joints and medial to the tendon of extensor digitorum longus

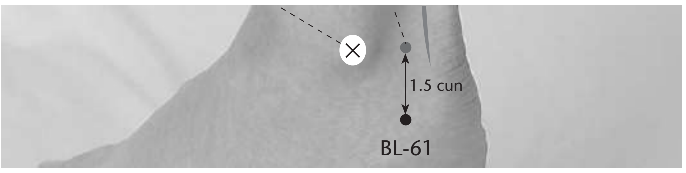

BL-61 (*pucan*) 'Servant's Respect' In a depression on the calcaneus, approximately 1.5 cun inferior to **BL-60**

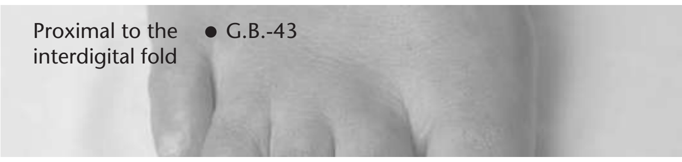
Proximal to the interdigital fold
G.B.-43

G.B.-43 (*xiaxi*) 'Clamped Stream' Between the 4th and the 5th toes, proximal to the margin of the interdigital web

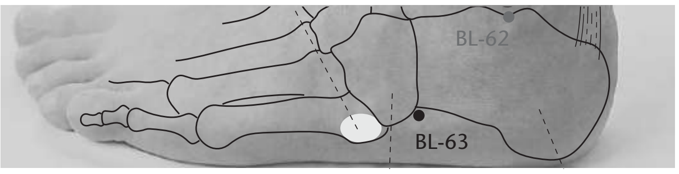

BL-63 (*jinmen*) 'Golden Gate' On the lateral aspect of the foot, proximal to the tuberosity of the 5th metatarsal bone, in a depression anterior and inferior to **BL-62** between the calcaneus and the cuboid bone

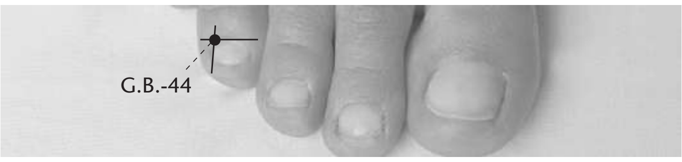

G.B.-44 (*zuqiaoyin*) 'Yin Portals of the Foot' On the 4th toe, 0.1 cun from the lateral corner of the nail

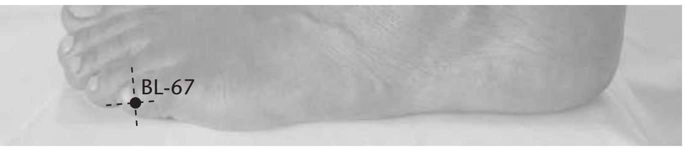

BL-67 (*zhiyin*) 'Reaching Yin' On the little toe, 0.1 cun from the lateral corner of the nail

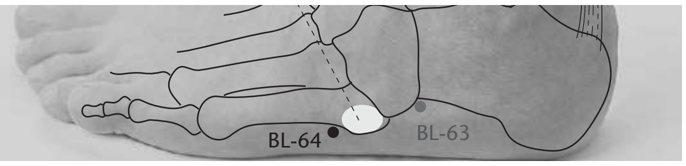

BL-64 (*jinggu*) 'Capital Bone' Distal to the tuberosity of the 5th metatarsal bone, at the border of the red and white skin

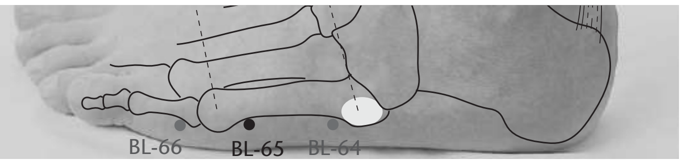

BL-65 (*shugu*) 'Restraining Bone' In the depression proximal to the head of the 5th metatarsal bone, at the border of the red and white skin

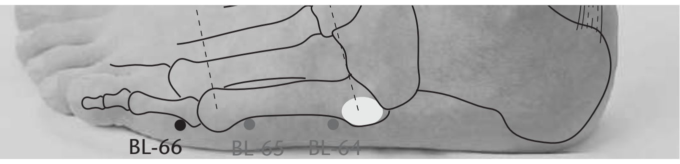

BL-66 (*zutonggu*) 'Foot Connecting Valley' In a depression distal to the metatarsophalangeal joint of the little toe

8 Point Categories and Point Selection

Claudia Focks

There are a total of 361 acupuncture points located on the channels (➞ Chapter 1). Translated literally from the Chinese language, the meaning of the word for 'point' is rather 'opening' or 'hole'. Stimulation of the points can be achieved through massage (in acupressure) or through the insertion of needles (in acupuncture). Besides the regular acupuncture points, there exist a wide variety of so-called extra points, located outside the pathways of the channels. The most commonly used of these extra points are described in Chapter 6, as classified by the WHO.

The acupuncture points are effective

- **locally** (for example, points on the elbow treat disorders of the elbow)
- **regionally** (the channel pathway – for example, points on the arm treat disorders of the upper extremity)
- some points have an effect on their **pertaining *zangfu***-Organ
- some points have **additional actions** such as calming the *shen* or eliminating Heat.

The following table provides an overview for the general indications of the channel points (modified after Ellis and Wiseman 1991)

Indications of the channel points

Channel	Indications		
Hand-Yin channels			
LU	Lungs, throat		Disorders of the thoracic region
P	Heart, Stomach	Spirit (*shen*) disorders	
HE	Heart		
Hand-Yang channels			
L.I.	Face, mouth, occiput, teeth, nose		Eyes, disorders of the throat, febrile diseases
T.B.	Head/lateral costal region	Disorders of the ears	
S.I.	Occipital region, shoulder, scapula, *shen* disorders		
Foot-Yin channels			
SP	Spleen, Stomach, intestines		Urogenital disorders, menstrual disorders, vaginal discharge, spirit (*shen*) disorders
LIV	Liver, genitalia		
KID	Kidneys, Lungs, throat		
Foot-Yang channels			
ST	Frontal aspect of the head, face, mouth, teeth, throat, Stomach, intestines		Psychoemotional disorders (*shen* disorders), febrile colds
G.B.	Lateral aspect of the head, ears, lateral costal region	Disorders of the eyes	
BL	Posterior aspect of the head, lower back		

8.1 Point Categories

Certain acupuncture points or groups of points are categorised according to their actions or location on their respective channels. (For an overview of these points and point groups ➞ inside back cover.)

8.1.1 *Yuan*-Source Points

At the *yuan*-source points, the Original *Qi* (*yuan qi* ➞ 1.1.4) of the respective Organ system emerges to the surface of the body.

Location: near the hand or ankle.

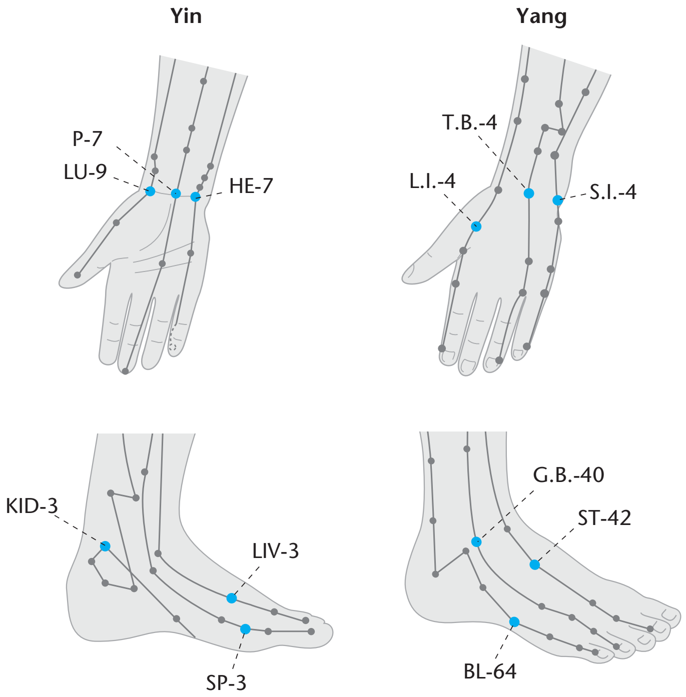

Points

Yuan-source points of the Yin channels

LU-9 (*taiyuan*)	**SP-3** (*taibai*)	**HE-7** (*shenmen*)
KID-3 (*taixi*)	**P-7** (*daling*)	**LIV-3** (*taichong*)

- On the Yin channels the *yuan*-source point is always the **3rd point** from the distal end of the channel
- They are identical with the *shu*-stream points and the Earth point of the channel

Function

- They tonify deficiency or weakness of their pertaining *zang*-Organ
- They regulate the Yin/Yang balance, restoring homeostasis
- As Earth points they have a 'grounding' effect on the body, the emotions and the mind.

Yuan-source points of the Yang channels

L.I.-4 (*hegu*)	**ST-42** (*chongyang*)	**S.I.-4** (*wangu*)
BL-64 (*jinggu*)	**T.B.-4** (*yangchi*)	**G.B.-40** (*qiuxu*)

- On the Yang channels the *yuan*-source point is always the **4th point** from the distal end of the channel, except for the G.B. channel (5th point)
- Regarding the energetics of the channel, they are all located between the *shu*-stream point and the *jing*-river point.

Function

- Their main function is to expel pathogenic factors in excess syndromes
- They tonify their pertaining *fu*-Organ.

Clinical application

Diagnosis

Disorders of a particular channel or its pertaining *zangfu*-Organ will often manifest themselves with tenderness, skin changes such as discolorations, swellings, redness, etc. in the area of the respective *yuan*-source point.

Therapy

- For deficiency of the relevant channel or Organ, use tonifying needling techniques; for excess, use reducing techniques
- Combine with the *luo*-connecting point of its Interiorly–Exteriorly paired channel (➞ 1.2.2) in order to increase the therapeutic effect and to balance Yin and Yang (➞ 8.3.3)
- Combine with the respective Back-*shu* point (➞ 8.1.4) for disorders of the *zangfu*-Organs

8.1.2 *Luo*-Connecting Points

The *luo*-connecting points are located in those regions of the body where the *luo*-network vessels (➞ 1.5) branch out, connecting the Interiorly–Exteriorly paired Yin and Yang channels (➞ 1.2.2).

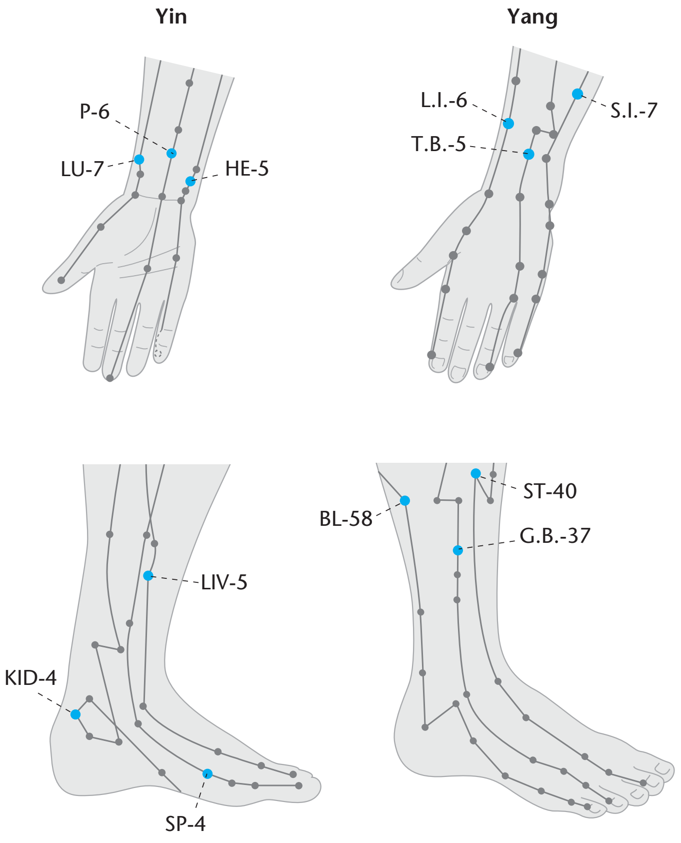

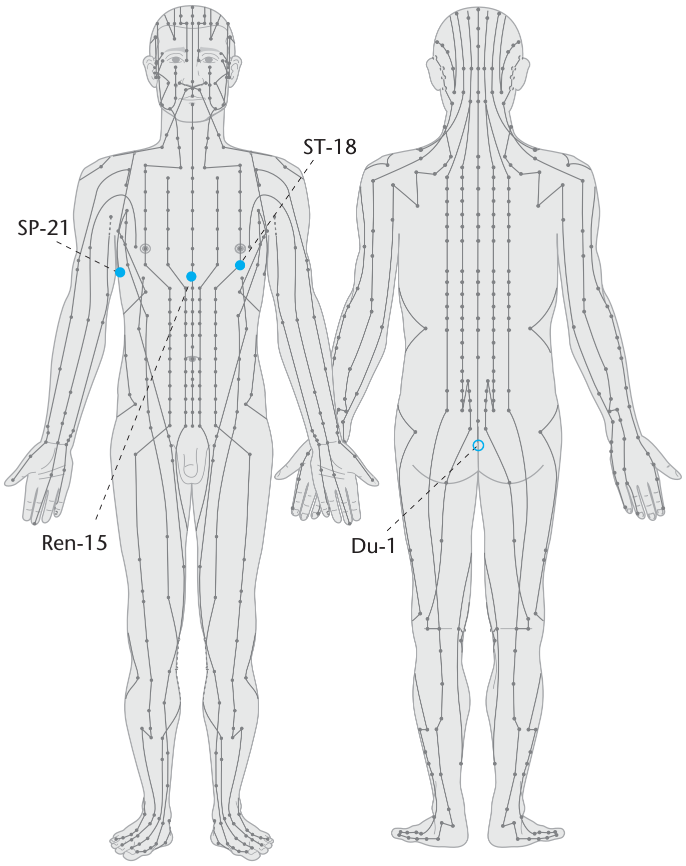

Points

The 12 *luo*-connecting points of the primary channels are located between the fingers and the elbow on the upper extremity and between the toes and the knee on the lower extremity.

- *Luo*-connecting points of the primary Yin channels
  - hand Yin channels: **LU-7** (*lieque*), **P-6** (*neiguan*), **HE-5** (*tongli*)
  - foot Yin channels: **SP-4** (*gongsun*), **LIV-5** (*ligou*), **KID-4** (*dazhong*)
- *Luo*-connecting points of the primary Yang channels
  - hand Yang channels: **L.I.-6** (*pianli*), **T.B.-5** (*waiguan*), **S.I.-7** (*zhizheng*)
  - foot Yang channels: **ST-40** (*fenglong*), **G.B.-37** (*guangming*), **BL-58** (*feiyang*)
- *Luo*-connecting points of the *ren mai:* **Ren-15** (*jiuwei*); action: regulates the *luo*-connecting vessels of the Yin channels; indication: abdominal diseases and disorders
- *Luo*-connecting points of the *du mai:* **Du-1** (*changqiang*); action: regulates the *luo*-connecting vessels of the Yang channels; indication: disorders of the back
- Great *luo*-connecting point of the Spleen: **SP-21** (*dabao*); action: connects with all *luo*-vessels and governs all *luo* and Blood vessels of the body; indication: disorders of the joint, general weakness and pain
- Great *luo*-connecting point of the Stomach: ***xu li****, near the left **ST-18**; moxibustion is contraindicated, only superficial needling is recommended. Indication: breathing and heart disorders such as asthma, dyspnoea, emphysema and bradycardia, also for shortness of breath due to anxiety attacks.

Clinical application

Diagnosis

- In excess syndromes the *luo* vessels often become visible near the *luo*-connecting points, for example skin discolorations and swellings may occur. With pain due to Cold the predominant colour may be bluish to greenish, with Heat syndromes reddish, with chronic stagnation and stasis dark to dark purple.
- In deficiency syndromes there are often no colour changes. However, in severe chronic cases (for example severe Qi deficiency) there will more likely be distinct muscular flaccidity, presenting as a dip or depression in the area of the point.

Therapy

- For disorders of the pertaining *zangfu*-Organ or channel
- For diseases of the Interiorly–Exteriorly paired Organ or channel
- For disorders due to the accumulation of pathogenic factors or substances, for example Qi stagnation, Blood stasis, retention of Dampness or Phlegm
- For psychoemotional problems, especially *luo*-connecting points of the Yin channels, for example classically **P-6** (*neiguan*) and **HE-5** (*tongli*)
- Bloodletting (three-edged needle, lancet or plum-blossom needle) on the *luo*-connecting point of the affected channel: this is particularly indicated for excess syndromes with Qi stagnation and Blood stasis. For more details on point selection ➞ 8.3.3

8.1.3 *Xi*-Cleft Points

Synonyms: Accumulation points (Maciocia)
The *xi*-cleft points are all located between the fingers and the elbow on the upper extremity and between the toes and the knee on the lower extremity. Only **ST-34** is located more proximally on the thigh.

---

* *Xu li* is only rarely mentioned in Western literature. The location and significance of this point is based on *Su Wen*, chapter 18: 'The geat *luo* vessel of the Stomach is called *xu li* and emerges at **ST-18** under the left breast, passes through the diaphragm and ascends to connect with the Lung. One can feel its pulsation with the hand. It is the place where the *zong qi* (➞ 1.1.4) of the channels resides.' (Ngyen Van Nghi, 1989). *Xu li* is the only *luo*-connecting vessel to directly emerge from a *fu*-Organ. It plays an important role in Korean acupuncture.

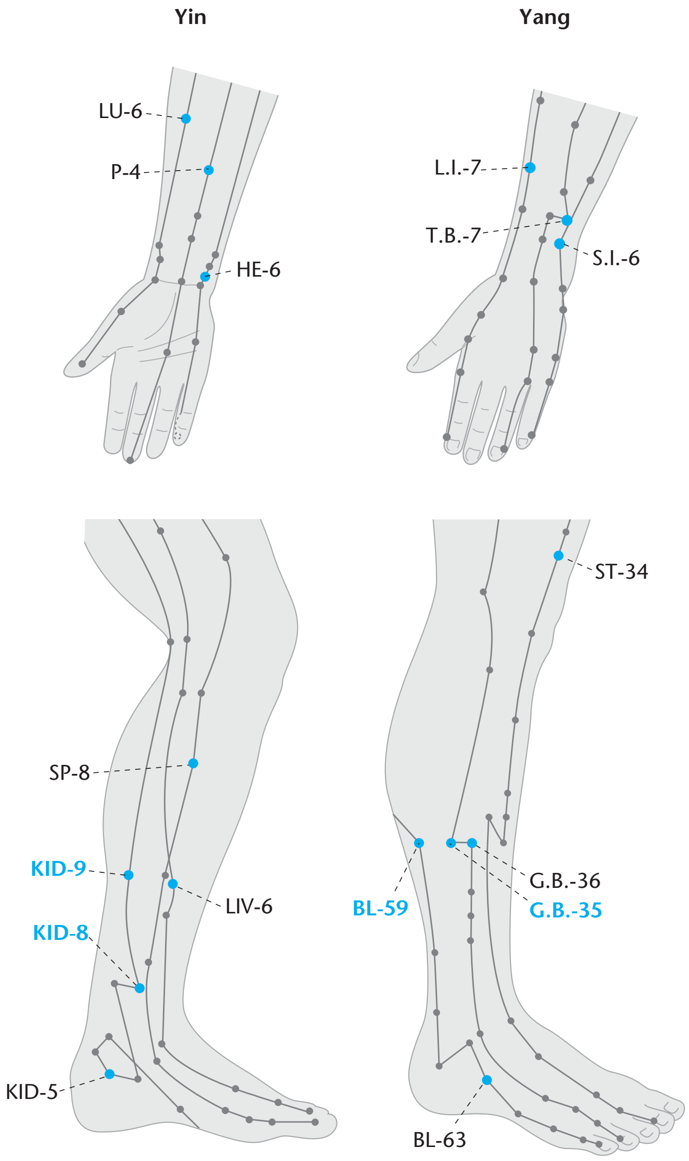
Yin
Yang
LU-6
P-4
HE-6
L.I.-7
T.B.-7
S.I.-6
ST-34
SP-8
KID-9
KID-8
LIV-6
KID-5
BL-59
G.B.-36
G.B.-35
BL-63

Points

Xi-cleft points of the primary Yin channels

Hand Yin channels

LU-6 (*kongzui*)
HE-6 (*yinxi*)
P-4 (*ximen*)

Foot Yin channels

SP-8 (*diji*)
KID-5 (*shuiquan*)
LIV-6 (*zhongdu*)

Xi-cleft points of the primary Yang channels

Hand Yang channels
L.I.-7 (*wenliu*)
S.I.-6 (*yanglao*)
T.B.-7 (*huizhong*)
Foot Yang channels
ST-34 (*liangqiu*)
BL-63 (*jinmen*)
G.B.-36 (*waiqiu*)

Xi-cleft points of the extraordinary vessels (point names printed in blue)

- *Xi*-cleft point of the *yang qiao mai*: **BL-59** (*fuyang*)
- *Xi*-cleft point of the *yin qiao mai*: **KID-8** (*jiaoxin*)
- *Xi*-cleft point of the *yang wei mai*: **G.B.-35** (*yangjiao*)
- *Xi*-cleft point of the *yin wei mai*: **KID-9** (*zhubin*)

Clinical application

Diagnosis

- They provide information about disorders affecting their pertaining channel or *zangfu*-Organ
  - Excess: acute, intense pain upon pressure or redness and swelling in the area of the point
  - Deficiency: dull, mild pain upon pressure or presence of a depression in the area of the point

Therapy

- Mainly used for acute and/or therapy-resistant disorders, especially for excess conditions and pain of the affected channel and/or pertaining *zangfu*-Organ. In patients with a robust constitution use reducing needling techniques.
- Add the *xi*-cleft points of the Yin channels for disorders of the Blood.
- In order to increase the therapeutic effect, the *xi*-cleft point may be combined with a relevant *hui*-meeting point (→ 8.1.7).

8.1.4 Back-*Shu* Points

Synonym: Back-transporting points
The Back-*shu* points are located on the inner branch of the BL channel at the level of their pertaining *zangfu*-Organ. They are mostly at the same level as the corresponding Front-*mu* point (→ 8.1.5).

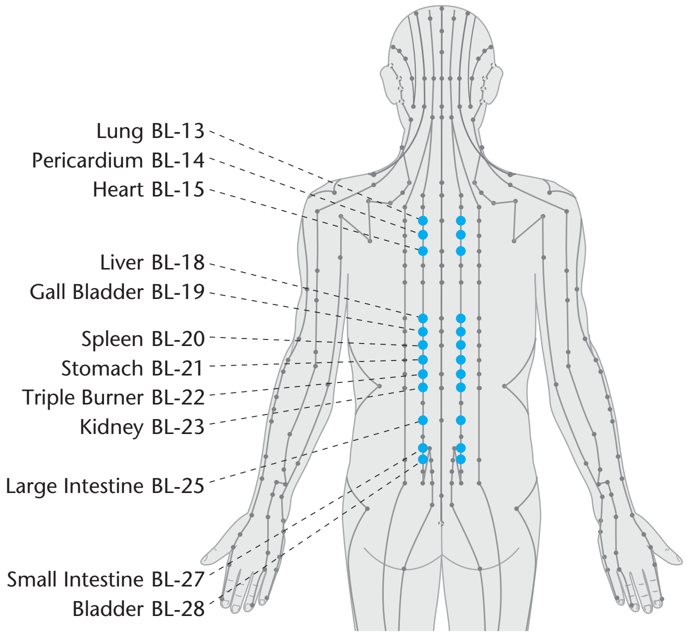
Lung BL-13
Pericardium BL-14
Heart BL-15
Liver BL-18
Gall Bladder BL-19
Spleen BL-20
Stomach BL-21
Triple Burner BL-22
Kidney BL-23
Large Intestine BL-25
Small Intestine BL-27
Bladder BL-28

Points

Organ	Back-*shu* point
Lung	**BL-13** (*feishu*)
Large Intestine	**BL-25** (*dachangshu*)
Stomach	**BL-21** (*weishu*)
Spleen	**BL-20** (*pishu*)
Heart	**BL-15** (*xinshu*)
Small Intestine	**BL-27** (*xiaochangshu*)
Bladder	**BL-28** (*pangguangshu*)
Kidney	**BL-23** (*shenshu*)
Pericardium	**BL-14** (*jueyinshu*)
Triple Burner	**BL-22** (*sanjiaoshu*)
Gall Bladder	**BL-19** (*danshu*)
Liver	**BL-18** (*ganshu*)

Clinical application

Diagnosis

- Often spontaneously tender and/or sensitive upon pressure in disorders of the pertaining *zangfu*-Organ

Therapy

- Mainly used for disorders of the corresponding *zangfu*-Organ. These points have a tonifying effect for chronic disorders and deficiency syndromes, but can also be used for acute disorders of their pertaining *zangfu*-Organ.
- For disorders of the sensory organs, use the Back-*shu* point of the pertaining *zangfu*-Organ.
  Example: for disorders of the eyes, needle the Back-*shu* point of the Liver **BL-18** (*ganshu*).

For excess syndromes, the Back-*shu* points can be needled with reducing needling techniques; for deficiency syndromes, tonifying needling techniques should be applied. When Cold signs are present, moxibustion can also be used. The needles should not be retained longer than 10 minutes – longer needle retention can have a sedating effect and lead to tiredness.

8.1.5 Front-*Mu* Points

Synonym: Alarm points
The Front-*mu* points are located on the thorax and abdomen in the area of their pertaining *zangfu*-Organs.

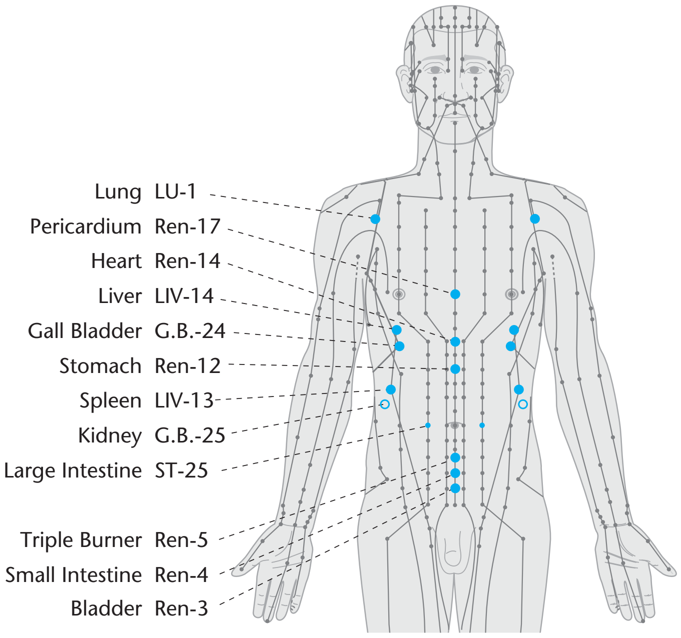

Points

Organ	Back-*shu* point
Lung	**LU-1** (*zhongfu*)
Large Intestine	**ST-25** (*tianshu*)
Stomach	**Ren-12** (*zhongwan*)
Spleen	**LIV-13** (*zhangmen*)
Heart	**Ren-14** (*juque*)
Small Intestine	**Ren-4** (*guanyuan*)
Bladder	**Ren-3** (*zhongji*)
Kidney	**G.B.-25** (*jingmen*)
Pericardium	**Ren-17** (*danzhong*)
Triple Burner	**Ren-5** (*shimen*)
Gall Bladder	**G.B.-24** (*riyue*)
Liver	**LIV-14** (*qimen*)

Clinical application

Diagnosis

- Often spontaneously tender or sensitive upon pressure in disorders of the pertaining *zangfu*-Organ

Therapy

- Used to regulate the pertaining *zangfu*-Organ. For excess syndromes use reducing needling techniques; for deficiency syndromes use tonifying methods.
- Commonly used point combinations:
  - *Shu–mu* method (➞ 8.3.2)
  - For disorders of the *fu*-Organs: simultaneous needling of the lower *he*-sea point (➞ 8.1.9) and the Front-*mu* point of the affected *fu*-Organ.

8.1.6 The Five *Shu*-Transporting Points

The five *shu-t*ransporting points are located **between the fingers and the elbow on the upper extremity, between the toes and the knee on the lower extremity**. The classification and application of these points is based on two theories: on the one hand they are applied according to the Qi flow in the channel from distal to proximal, on the other hand they are used according to the Five Phases (➞ 8.2.5).

While the location of the five *shu-t*ransporting points is identical to the Five Phase points, their dynamics and range of application based on the Qi flow in the channels differs from or is even contrary to their actions as Five Phase points. The theoretical foundations of the Qi flow in the channels as applied in the instance of the *shu*-transporting point are based on a rather earlier circulatory concept (➞ 1.1.1).

The therapeutic application of the Five Phase points will be explained in more detail in ➞ 8.2.5, while the clinical application according to channel flow is described below.

Theoretical background of the Qi circulation in the channels

According to the more original centripetal model (➞ 1.1.1), the Qi flows from the peripheral parts of the body to the elbows and knees, similar to a river coursing from its spring to its estuary and finally into the sea:

- The Qi flow begins at the anastomosis at the *jing*-well point. Here the water surfaces, it is dynamic and unstable (change of polarity from Yin to Yang)
- The water course then changes to a spring (*ying*) and a stream (*shu*) – both are at this point still quite dynamic features, the water flowing rapidly and superficially.
- The water then becomes increasingly deeper and slower, taking on the characteristics of a river (*jing*).
- Eventually it flows as a big river into the sea (*he*).

This image of a river system is often used to explain disorders caused by pathogenic factors (*xie qi*). Due to the superficial position of these points on the extremities, pathogens can here easily penetrate the body. By the same token, they can be removed again with relatively little effort. Barbara Kirschbaum (course material 1999) provides a useful visual image: she suggests imagining a ship carrying a cargo of pathogens that begins its journey at the tips of the extremities. The further a point is located peripherally in the body, the easier it is for the ship to get rid of its dangerous cargo (or to expel the pathogen according to the dynamics of the Qi flow). However, once the ship has reached the *he*-sea point, the ship and its cargo will disappear in the big, wide sea, having the potential to injure the internal *zangfu*-Organ. Thus it is the therapeutic goal to eliminate pathogens from the body as soon as possible.

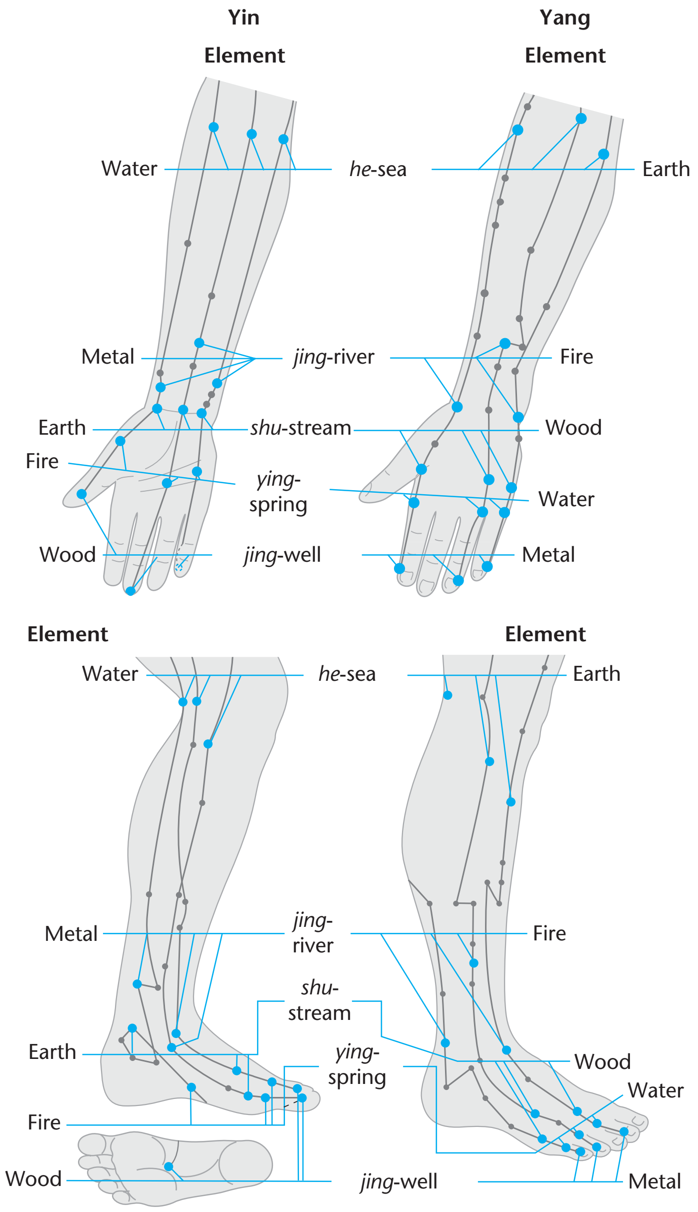

Overview: *shu*-transporting points

Shu-transporting points of the Yin channels					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

Shu-transporting points of the Yang channels					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41+	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

Clinical application of the *shu*-transporting points

Jing-well points (*jing xue*)

Shu-transporting Points of the Yin channels					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

Shu-transporting points of the Yang channels					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

According to the Five Phases (→ 8.2.5) the *jing*-well points of the **Yin channels** correspond to the **Wood points**, while those of the **Yang channels** correspond to the **Metal points**.

Location: as the **last or first point of each channel**, they are located on the **tips of the fingers or the toes** (exception: **KID-1** of the sole of the foot). The course of the channel is here at its **most superficial**. It is also here that a change of polarity is taking place in the energy: from Yin to Yang or vice versa. **Caution**: needling these points can be quite intense and painful, presenting a strong stimulus to the body.

Actions

- Clear Heat, restore consciousness, eliminate Heat and excess from the upper end of the channel: commonly pricked to bleed.
- Treat fullness and Cold below the Heart
- Treat disorders of the spirit (*shen*) for psychological disorders

Indications

- For acute conditions and emergencies: the *jing*-well points can quickly expel Exterior pathogenic factors (*xie qi*), especially Heat. They have a strong effect on the psyche, for example for coma, collapse and shock as well as for severe irritability and restlessness.
- The points on the Yin channels also have an effect on their pertaining *zang*-Organs and can eliminate internal Wind.

Examples

- **LU-11**: prick to bleed for acute sore throat
- **KID-1**: for spasms and unconsciousness
- **HE-9**: for collapse
- **SP-1**: for uterine bleeding

For psychological disorders:

- **KID-1**: for severe restlessness
- **P-9** and **HE-9**: for extreme restlessness and insomnia

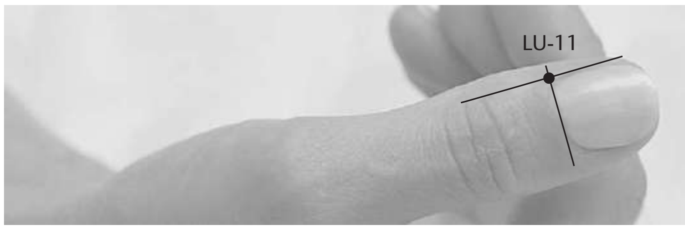

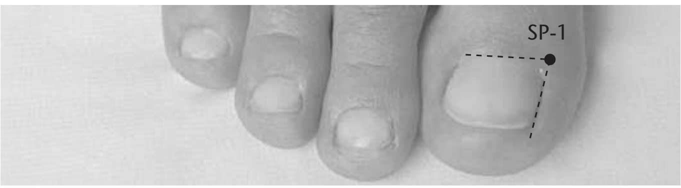

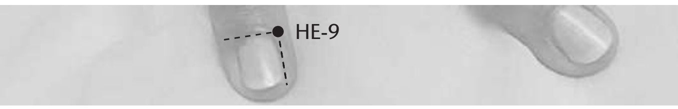

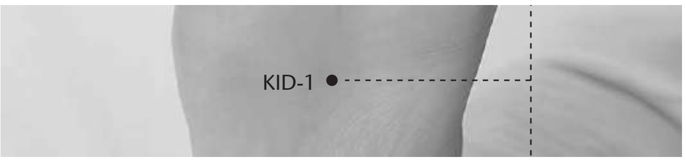

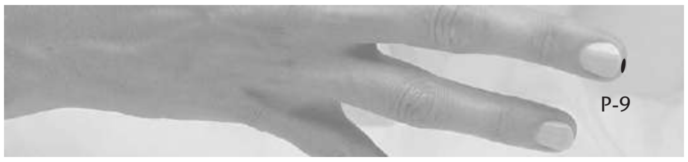

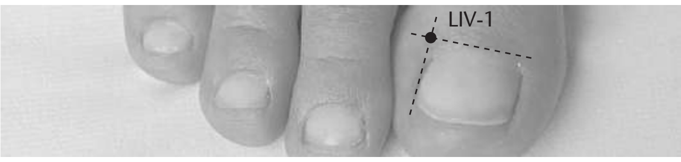

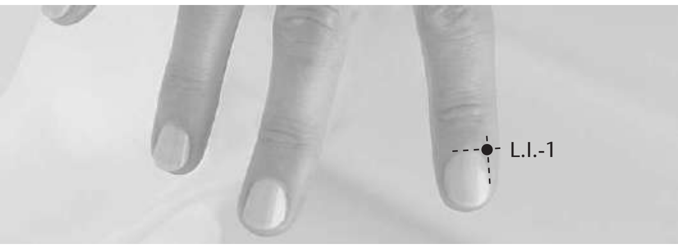

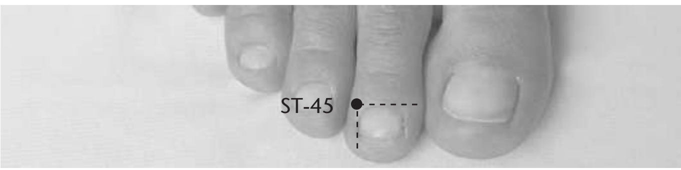

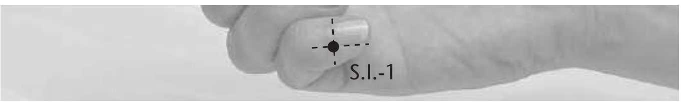

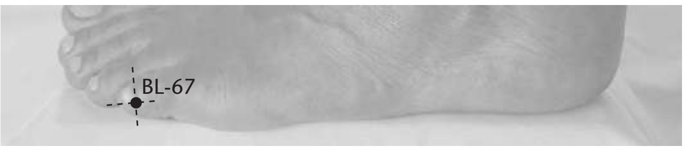

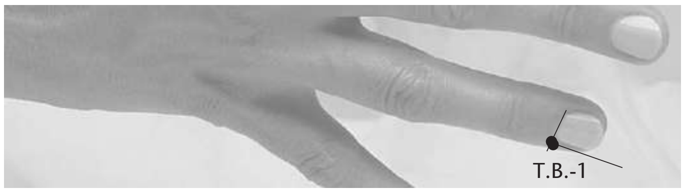

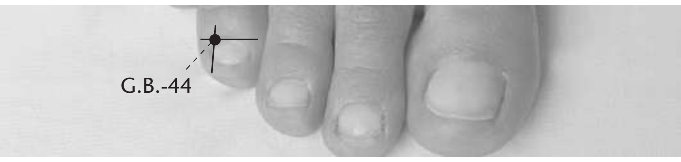

Ying-spring points (*ying xue*)

Shu-transporting points of the Yin channels					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point
***Shu*-transporting points of the Yang channels**					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43 +	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

According to the Five Phases (→ 8.2.5) the *ying*-spring points of the Yin channels correspond to the Fire points, while those on the Yang channels correspond to the Water points. **Caution**: needling some of these points can be painful, for example **LU-10, P-8, HE-8**.

Location: as the second most distal point of each channel, they are located on the upper extremity between the phalanges and metacarpal bones and on the lower extremity between the phalanges with the exception of KID-2.

Actions

- They are very dynamic points that eliminate both external and internal pathogenic factors.
- They clear Heat from their pertaining *zangfu*-Organ and channel, especially in the upper half of the body.

Indications

- For example for febrile diseases: Exterior Heat syndromes are mainly treated using points on the Yang channels while Interior Heat or deficient Heat syndromes are treated with points on the Yin channels.

Examples

- **LU-10:** for acute inflammations of the throat
- **ST-44:** for acute frontal headaches or acute inflammations affecting the face
- **LIV-2:** for severe temporal headaches, for example due to blazing Liver Fire

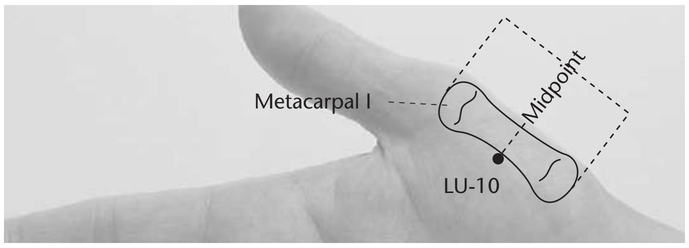

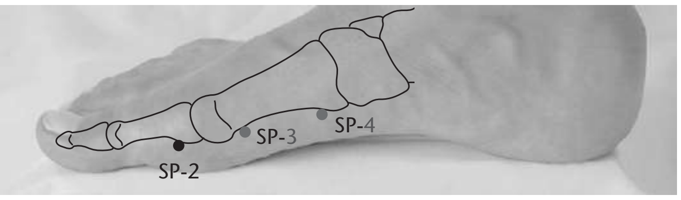

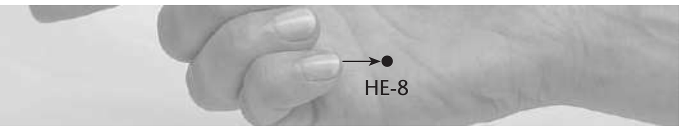

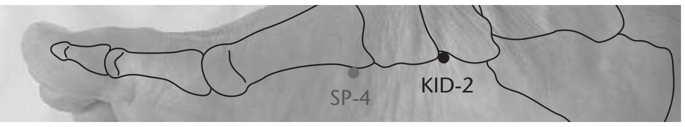

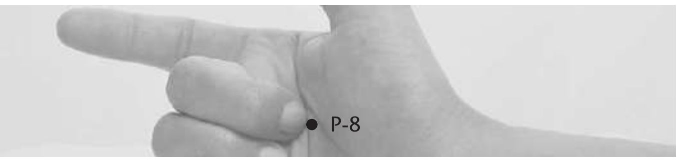

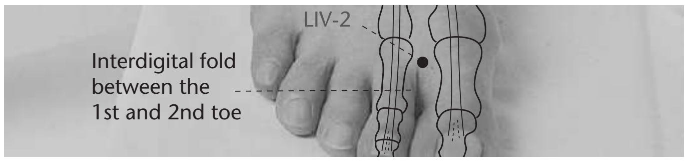

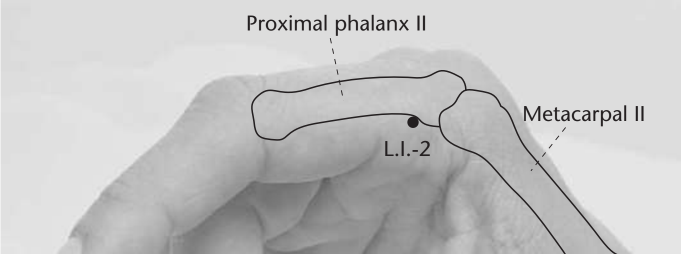

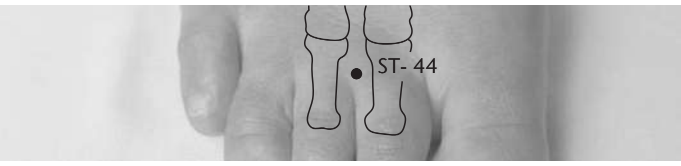

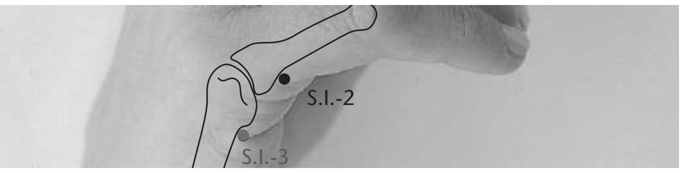

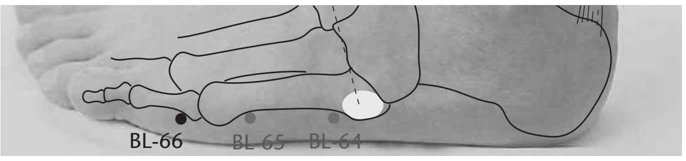
BL-66
BL-65
BL-64

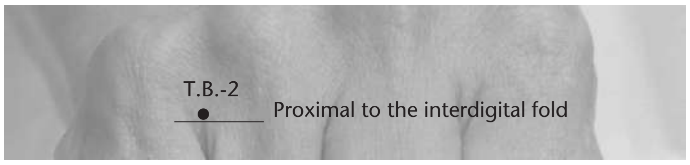
T.B.-2
Proximal to the interdigital fold

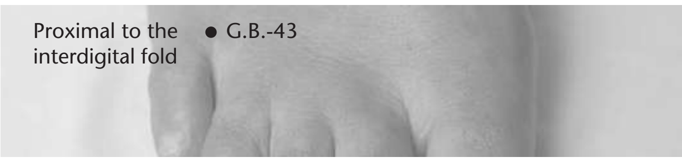
Proximal to the interdigital fold
G.B.-43

Shu-stream points (*shu xue*)

***Shu*-transporting points of the Yin channels**					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2−	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

***Shu*-transporting points of the Yang channels**					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11+
ST	ST-45 −	ST-44	ST-43	ST-41+	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38−	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; −, sedation point

According to the Five Phases (→ 8.2.5) the *shu*-stream points of the **Yin channels** correspond to the **Earth points** and ***yuan*-Source points** (→ 8.1.1), while those on the **Yang channels** correspond to the **Wood points**.

Location: counting from the periphery, they are the **third point of each channel**, the only exception being the G.B. channel (4th point on the channel). They are located proximal to the metacarpophalangeal or metatarsophalangeal joints, except for **KID-3** (posterior to the medial malleolus), **LIV-3** and **SP-3** (both located at the distal ends of the metatarsal bones) as well as **LU-9, HE-7** and **P-7** (all located on the wrist joint space).

Actions

They are considered to be the entry points for pathogenic factors; needling these points can strengthen the Defensive Qi (*wei qi*) (→ 1.1.4) and eliminate pathogenic factors from the respective channel. At the *shu*-stream points, the Qi flow is already somewhat deeper: the bed of the river is widening and deepening.

Indications

For *bi*-syndromes of the joints, especially due to Damp or pathogenic Wind. They are generally used for heaviness of the body as well as for *shaoyang* syndrome with intermittent symptoms.

- Points on the Yang channels are predominantly used for Exterior syndromes: they have a specific action on their respective channel.
- Points on the Yin channels are more commonly used for Interior syndromes; they strengthen and harmonise their respective *zang*-Organs (this is in part due to these points also being the *yuan*-source points).

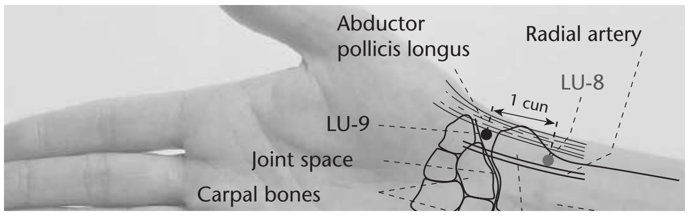
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones

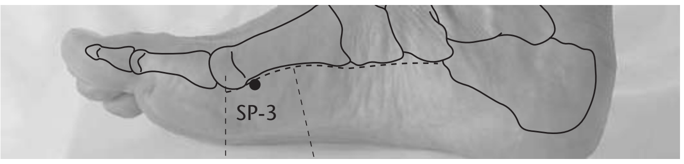

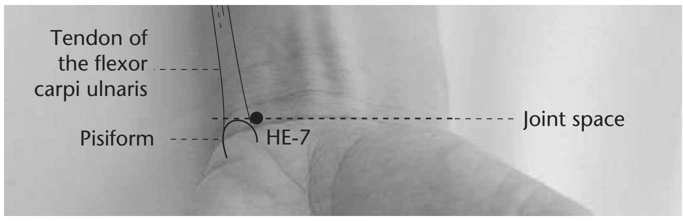
Tendon of the flexor carpi ulnaris
Pisiform
HE-7
Joint space

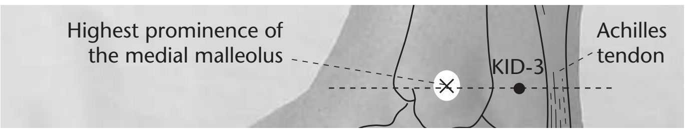
Highest prominence of the medial malleolus
KID-3
Achilles tendon

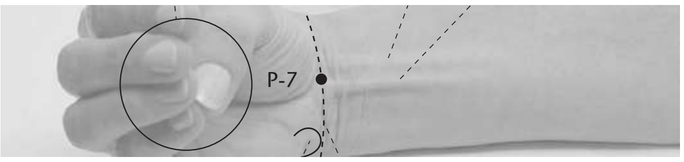

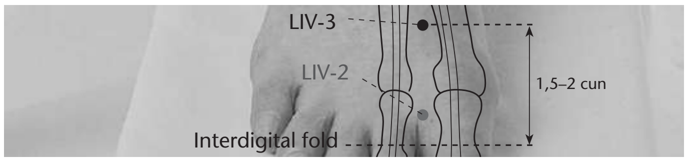
LIV-3
LIV-2
1,5–2 cun
Interdigital fold

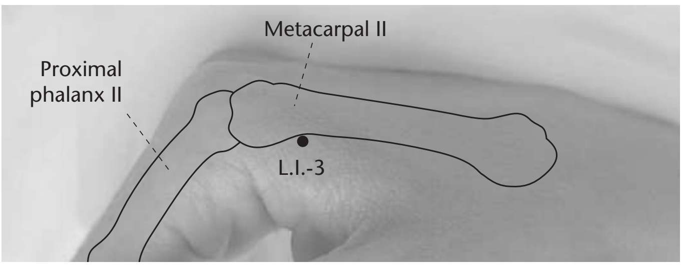

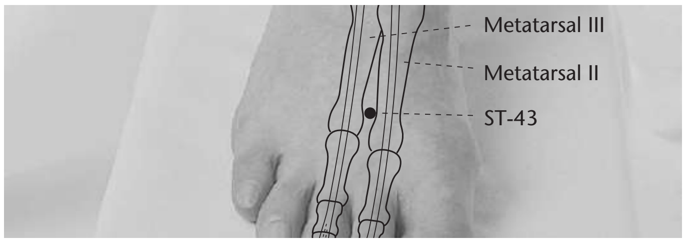

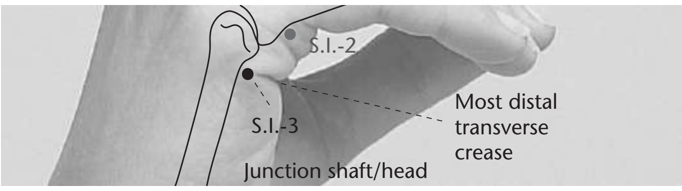

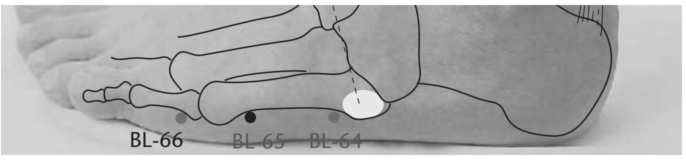

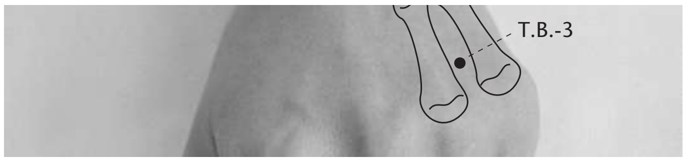

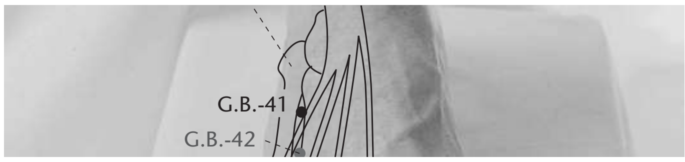

Jing-river points (*jing xue*)

***Shu*-transporting points of the Yin channels**					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7−	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point
***Shu*-transporting points of the Yang channels**					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

According to the Five Phases (➞ 8.2.5), the *jing*-river points of the **Yin channels** correspond to the **Metal points**, while those on the **Yang channels** correspond to the **Fire points**.

Location: they are located **between** the ***shu*-stream points** and the ***he*-sea points**. On the upper extremity, they are located either at, or distal or proximal to the wrist, while on the lower extremity they are located either at, or distal or proximal to the ankle (see figure on p. 676).

Actions

They divert external pathogenic factors away from joints, bones and tendons towards the Exterior. At the *jing*-river points the Qi flow becomes increasingly wider, slower, deeper and larger – thus the actions of these points are also less dynamic compared to the first three, more peripheral, *shu*-transporting points. At the *jing*-river points external pathogens have already penetrated deeper into the body, especially into the joints, bones and tendons.

Indications

- Dyspnoea, disorders of the throat such as inflammations, cough, disorders manifesting in the voice.
- Alternating fever and chills (especially points on the Yin channels)
- Also for *bi*-syndromes such as painful joint and tendon disorders.

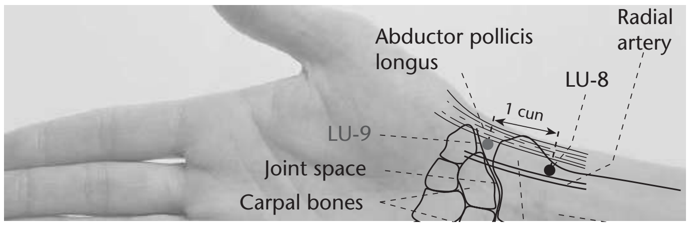
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones

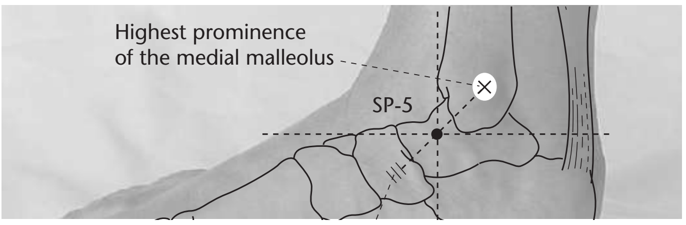
Highest prominence of the medial malleolus
SP-5

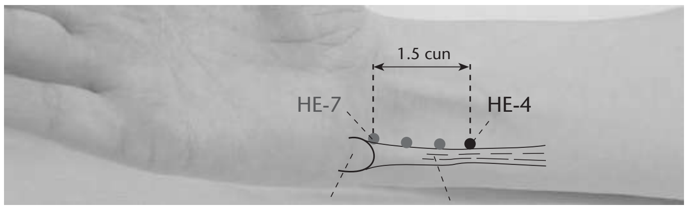
1.5 cun
HE-7
HE-4

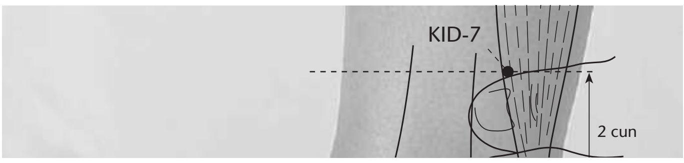

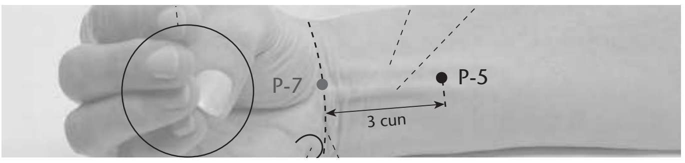

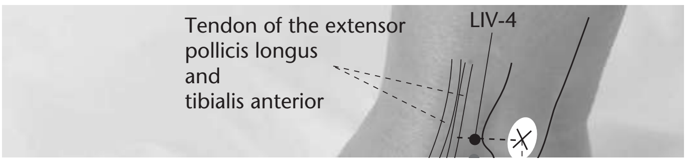
Tendon of the extensor pollicis longus and tibialis anterior
LIV-4

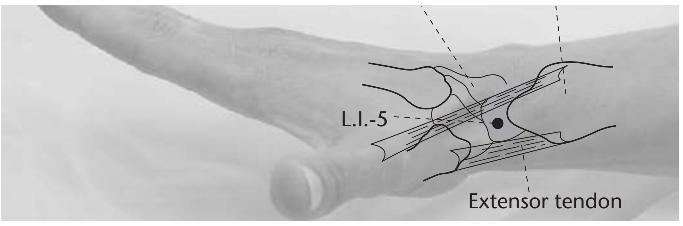
L.I.-5
Extensor tendon

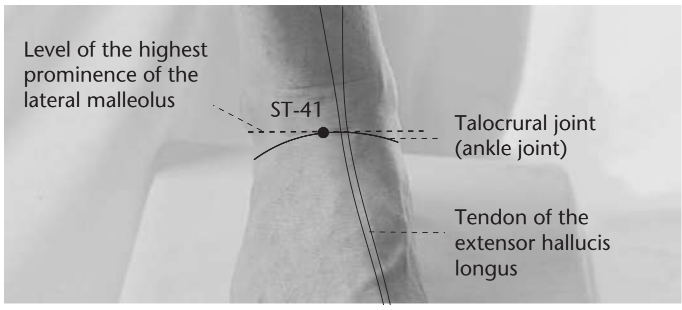
Level of the highest prominence of the lateral malleolus
ST-41
Talocrural joint (ankle joint)
Tendon of the extensor hallucis longus

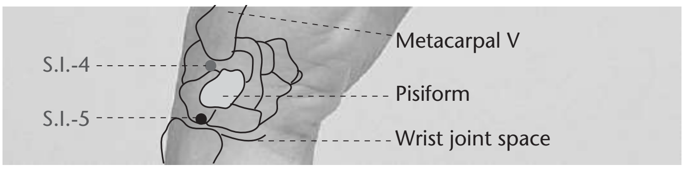
S.I.-4
S.I.-5
Metacarpal V
Pisiform
Wrist joint space

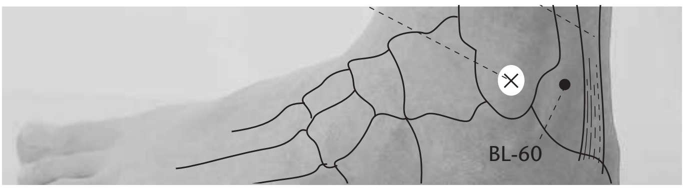

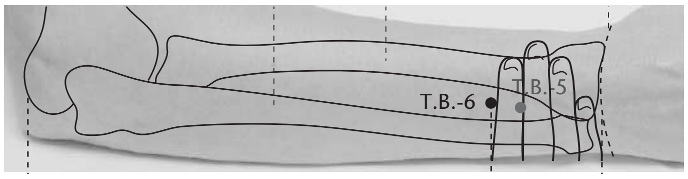

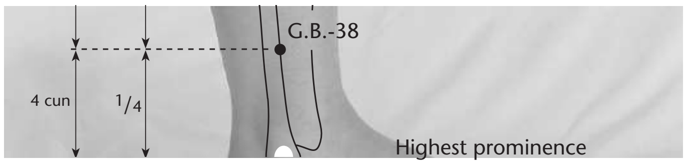
G.B.-38
4 cun
1/4
Highest prominence

He-sea points (*he xue*)

***Shu*-transporting points of the Yin channels**					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7−	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9+	P-8	P-7−	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

***Shu*-transporting points of the Yang channels**					
Phase/ Channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38−	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

According to the Five Phases (→ 8.2.5) the *he*-sea points of the **Yin channels** correspond to the **Water points**, while those on the **Yang channels** correspond to the **Earth points**.

Location: on the upper extremity these points are located near the **elbow**, on the lower extremity near the **knee**.

Actions

- They descend counterflow Qi and stop diarrhoea. At the *he*-sea points the Qi flow is very wide, deep and slow, 'terminating in the deep sea' where the 'toxic cargo' can 'disappear' into the body unless treated adequately. External pathogens (*xie qi*) which have not been eliminated from the channel beforehand can here penetrate the Interior of the body and injure the *zangfu*-Organs.
- Diseases of the *fu*-Organs (here especially points on the lower extremities are indicated).

Indications

- Disorders of the *fu*-Organs, especially gastro-intestinal disorders with symptoms such as vomiting and diarrhoea
- Some *he*-sea points are also indicated for skin diseases, for example **LI-11, BL-40**.

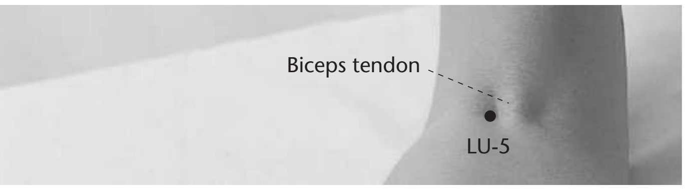
Biceps tendon
LU-5

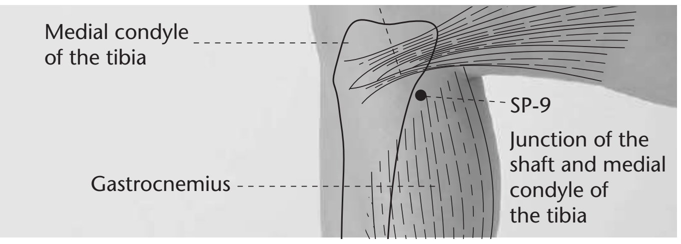
Medial condyle of the tibia
SP-9
Junction of the shaft and medial condyle of the tibia
Gastrocnemius

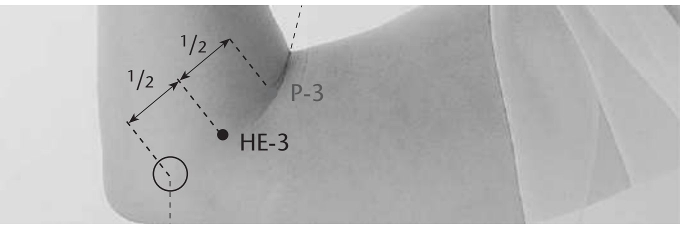

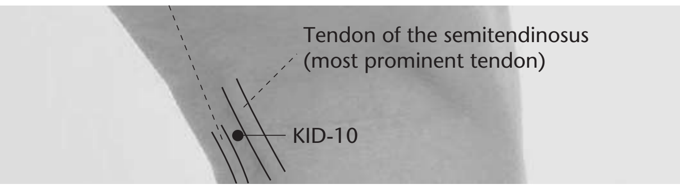
Tendon of the semitendinosus (most prominent tendon)
KID-10

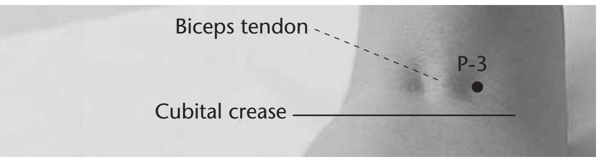
Biceps tendon
P-3
Cubital crease

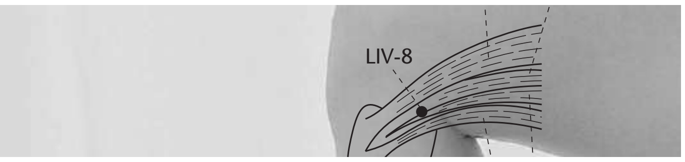

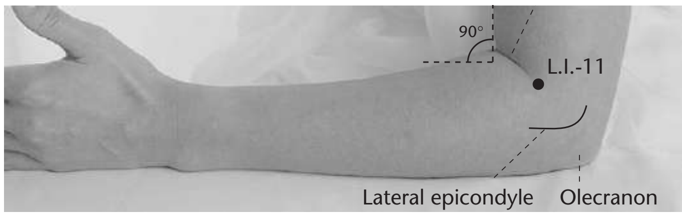
90°
L.I.-11
Lateral epicondyle
Olecranon

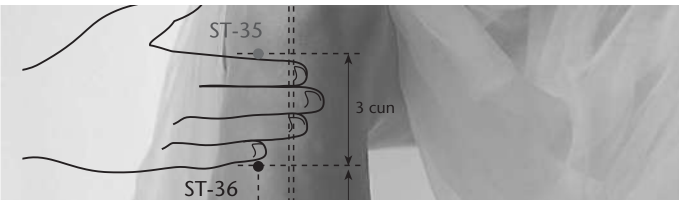
ST-35
3 cun
ST-36

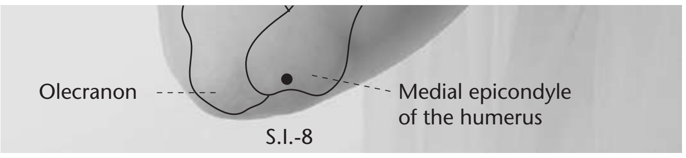
Olecranon
Medial epicondyle of the humerus
S.I.-8

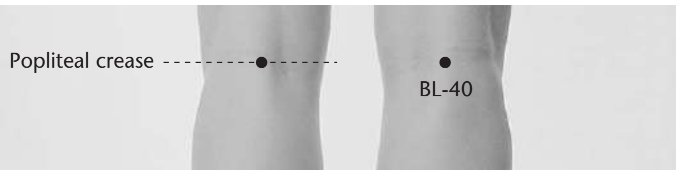
Popliteal crease
BL-40

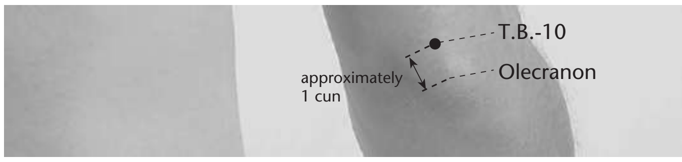
T.B.-10
approximately 1 cun
Olecranon

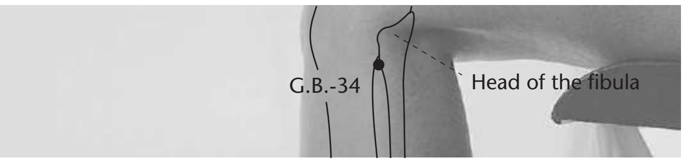
G.B.-34
Head of the fibula

8.1.7 The Eight *Hui*-Meeting Points

Synonym: Gathering Points (Maciocia)
The Qi of a particular tissue, Organ system or substance gathers at the *hui*-meeting points and can be influenced through these points.

Points and clinical application

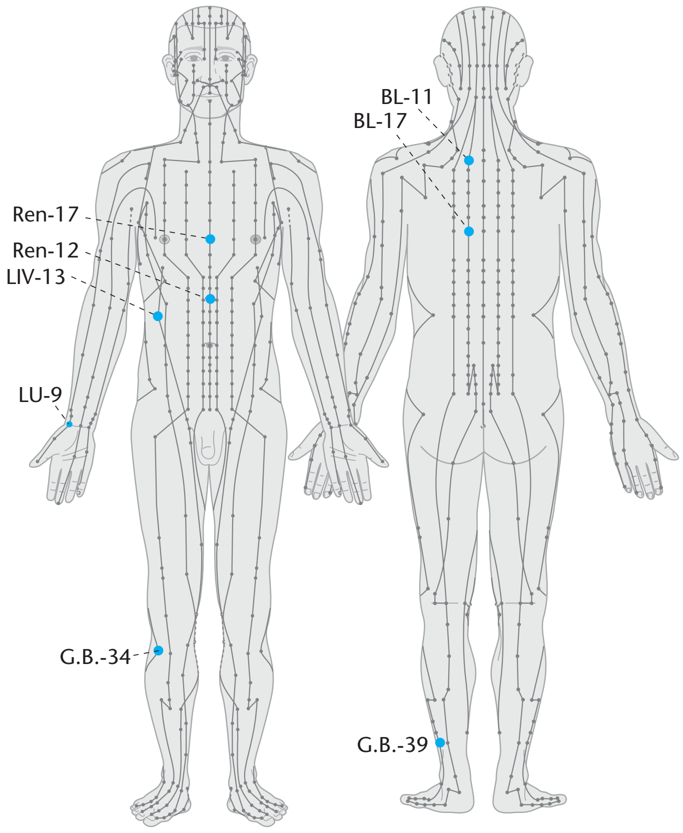
BL-11
BL-17
Ren-17
Ren-12
LIV-13
LU-9
G.B.-34
G.B.-39

The Eight *hui*-Meeting Points		
Point	Tissue/organ system/substance	Indications
LIV-13 (*zhangmen*)	*zang*-Organs	Disorders of the *zang*-Organs (tonifies the Spleen and indirectly all *zang*-Organs)
Ren-12 (*zhongwan*)	*fu*-Organs	Disorders of the *fu*-Organs, for example of the gastro-intestinal tract
Ren-17 (*danzhong or shanzhong*)	Qi, has a strong effect on the *zong qi*	Respiratory diseases and disorders of the Qi flow such as hiccups; controls the *zong qi* and therefore functioning of the Lungs and Heart
BL-17 (*geshu*)	Blood (*xue*)	Disorders of the Blood such as anaemia, Blood stasis, haemorrhagia and gynaecological disorders
G.B.-34 (*yangling-quan*)	Sinews	Disorders of the joints, sinews and muscles
LU-9 (*taiyuan*)	Blood vessels	Stimulates the peripheral Blood circulation for disorders of the blood vessels such as vasculitis and arteriosclerosis
BL-11 (*dazhu*)	Bones	Disorders of the Bones such as shoulder and spine problems, joint and bone pain
G.B.-39 (*xuanzhong*)	Marrow	Disorders of the Bones, the Marrow and Brain (nerves)

8.1.8 Opening Points of the Eight Extraordinary Vessels

Synonyms: Confluent points (Deadman et al 1998), Master points (Pirog 1996)
The table below shows the opening points of the eight extraordinary vessels. For more detail on their origin and pathways → 1.7, Chapter 5.

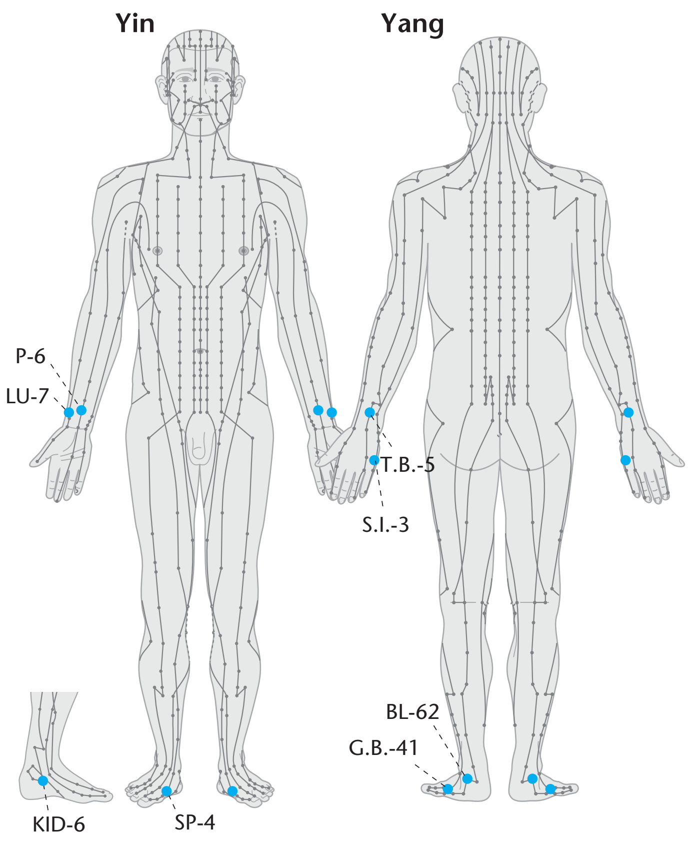
Yin
Yang
P-6
LU-7
T.B.-5
S.I.-3
BL-62
G.B.-41
KID-6
SP-4

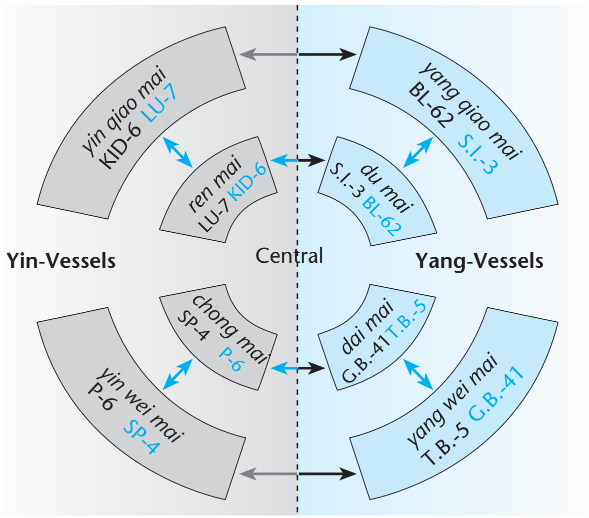
yin qiao mai
KID-6 LU-7
yang qiao mai
BL-62 S.I.-3
ren mai
LU-7 KID-6
du mai
S.I.-3 BL-62
Yin-Vessels
Central
Yang-Vessels
chong mai
SP-4 P-6
dai mai
G.B.-41 T.B.-5
yin wei mai
P-6 SP-4
yang wei mai
T.B.-5 G.B.-41

black	Opening point
blue	Coupled point

Central ↔ Peripheral (Yin–Yin, Yang–Yang)
Peripheral ↔ Peripheral (Yin–Yang)
Central ↔ Central (Yin–Yang)

Points

Pairs	Extraordinary Vessel	Opening point	Coupled point (opening point of the paired vessel)	Supported body regions
Pair 1	*chong mai*	**SP-4** (*gongsun*)	**P-6** (*neiguan*)	Heart, chest, Stomach
	yin wei mai	**P-6** (*neiguan*)	**SP-4** (*gongsun*)	
Pair 2	*du mai*	**S.I.-3** (*houxi*)	**BL-62** (*shenmai*)	Occiput, shoulder, back
	yang qiao mai	**BL-62** (*shenmai*)	**S.I.-3** (*houxi*)	
Pair 3	*dai mai*	**G.B.-41** (*zulinqi*)	**T.B.-5** (*waiguan*)	Temples, ears, lateral aspects of the body
	yang wei mai	**T.B.-5** (*waiguan*)	**G.B.-41** (*zulinqi*)	
Pair 4	*ren mai*	**LU-7** (*lieque*)	**KID-6** (*zhaohai*)	Face, throat, chest, Lungs, abdomen
	yin qiao mai	**KID-6** (*zhaohai*)	**LU-7** (*lieque*)	

Clinical application

Various approaches exist for needling the opening/coupled points of the extraordinary vessels. The following method **for 'opening' or regulating the extraordinary vessels** has been used with good therapeutic results in clinical practice:

- Needle the **opening point** of the relevant channel first, then needle the **contralateral coupled point** (opening point of the paired extraordinary vessel). This will mobilise and regulate the Qi of the extraordinary vessel, making it more dynamic.
- During a subsequent session, needle **either the *xi*-cleft point** of the previously treated extraordinary vessel (→ 8.1.3) or add **other channel points of the vessel** (see channel pathways, chapter 5) during the same or subsequent session. This will further support the dynamics of the Qi flow in the vessel.

In addition Maciocia (1989) and Kirschbaum (1995) suggest the following procedure for 'opening' the extraordinary vessels:

- **In women:** First needle the opening point on the *right* side with even method, then the coupled point on the *left* side with even method. Afterwards other points on the respective vessel may be added.
- **In men:** First needle the opening point on the *left* side with even method, then the coupled point on the *right* side with even method. Afterwards other points on the respective vessel may be added.
- **Needle retention:** the needles should be retained for 20–25 minutes and should be removed in the opposite order.

8.1.9 Lower *He*-Sea Points

The lower *he*-sea points are used as the main points for disorders of their pertaining *fu*-Organ.

Points and clinical application

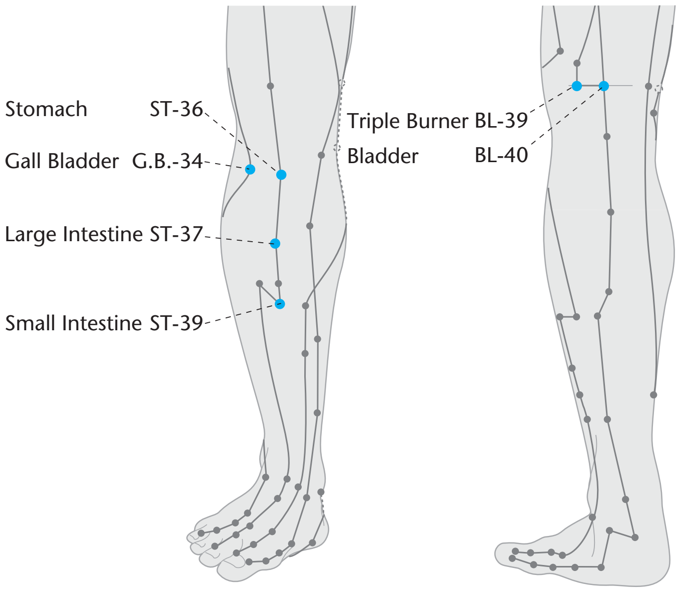

- **ST-36** (*zusanli*): Lower *he*-sea point of the Stomach; indication: loss of appetite, epigastric pain and fullness, acid reflux, abdominal pain, constipation, diarrhoea
- **ST-37** (*shangjuxu*): Lower *he*-sea point of the Large Intestine; indication: appendicitis, diarrhoea
- **ST-39** (*xiajuxu*): Lower *he*-sea point of the Small Intestine; indication: diarrhoea, abdominal pain
- **G.B.-34** (*yanglingquan*): Lower *he*-sea point of the Gall Bladder; indication: cholecystitis, vomiting
- **BL-40** (*weizhong*): Lower *he*-sea point of the Bladder; indication: urinary incontinence, urinary retention
- **BL-39** (*weiyang*): Lower *he*-sea point of the Triple Burner; indication: urinary incontinence, urinary retention

8.1.10 Meeting Points (*Jiaohui* Points)

Synonym: Intersecting points (Practical Dictionary)

Meeting points (*jiaohui* points) represent intersections between channels and vessels. Needling them allows more than one channel to be influenced and widens the range of actions of a given point. By the same token, the number of needles may be reduced while maintaining a good therapeutic effect.

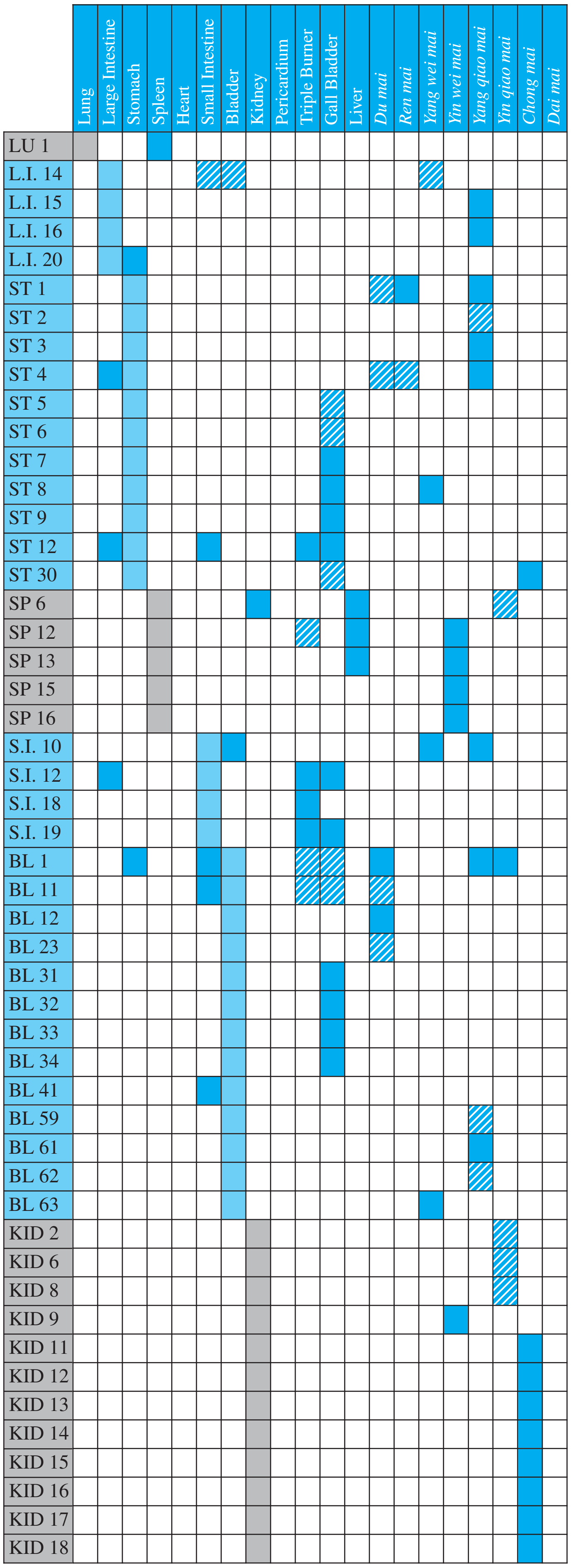

	Lung	Large Intestine	Stomach	Spleen	Heart	Small Intestine	Bladder	Kidney	Pericardium	Triple Burner	Gall Bladder	Liver	*Du mai*	*Ren mai*	*Yang wei mai*	*Yin wei mai*	*Yang qiao mai*	*Yin qiao mai*	*Chong mai*	*Dai mai*
LU 1	▒			■																
L.I. 14		░				▨	▨								▨					
L.I. 15		░															■			
L.I. 16		░															■			
L.I. 20		░	■																	
ST 1			░										▨	■			■			
ST 2			░														▨			
ST 3			░														■			
ST 4		■	░										▨	▨			■			
ST 5			░								▨									
ST 6			░								▨									
ST 7			░								■									
ST 8			░								■				■					
ST 9			░								■									
ST 12		■	░			■				■	■									
ST 30			░								▨								■	
SP 6				▒				■				■						▨		
SP 12				▒						▨		■				■				
SP 13				▒								■				■				
SP 15				▒												■				
SP 16				▒												■				
S.I. 10						░	■								■		■			
S.I. 12		■				░				■	■									
S.I. 18						░				■										
S.I. 19						░				■	■									
BL 1			■			■	░			▨	▨		■				■	■		
BL 11						■	░			▨	▨		▨							
BL 12							░						■							
BL 23							░						▨							
BL 31							░				■									
BL 32							░				■									
BL 33							░				■									
BL 34							░				■									
BL 41						■	░													
BL 59							░										▨			
BL 61							░										■			
BL 62							░										▨			
BL 63							░								■					
KID 2								▒										▨		
KID 6								▒										▨		
KID 8								▒										▨		
KID 9								▒								■				
KID 11								▒											■	
KID 12								▒											■	
KID 13								▒											■	
KID 14								▒											■	
KID 15								▒											■	
KID 16								▒											■	
KID 17								▒											■	
KID 18								▒											■	

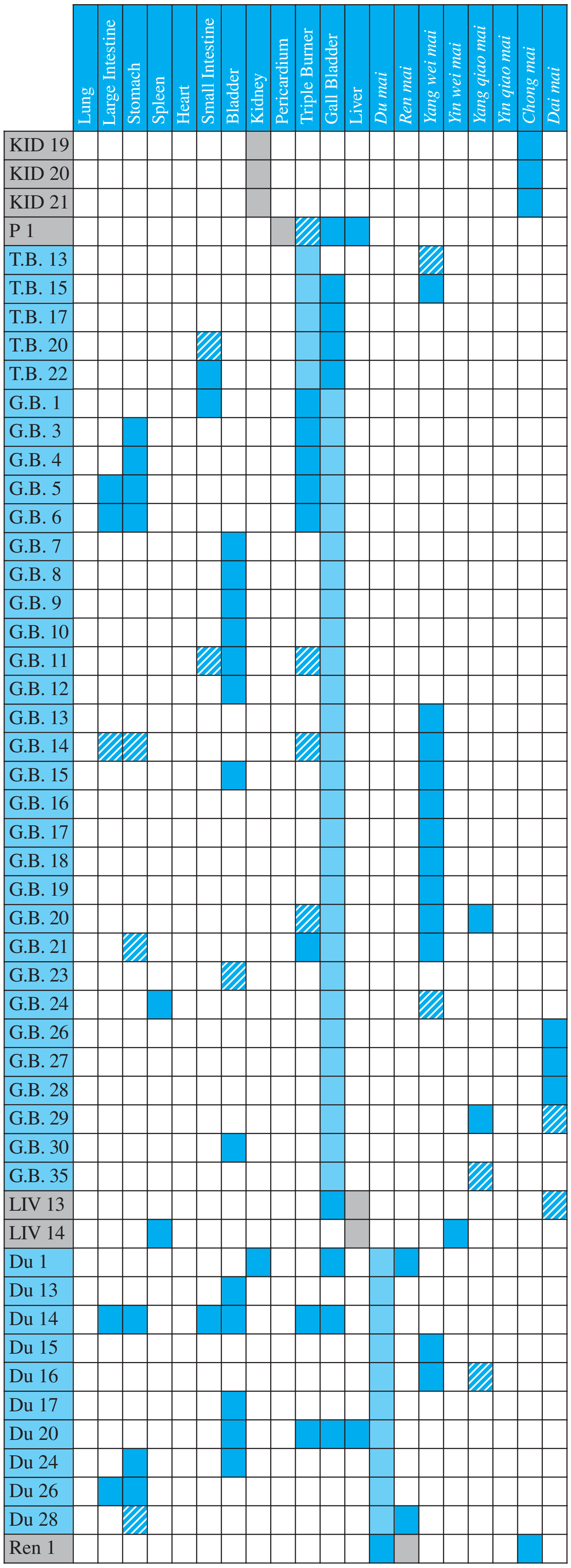

	Lung	Large Intestine	Stomach	Spleen	Heart	Small Intestine	Bladder	Kidney	Pericardium	Triple Burner	Gall Bladder	Liver	*Du mai*	*Ren mai*	*Yang wei mai*	*Yin wei mai*	*Yang qiao mai*	*Yin qiao mai*	*Chong mai*	*Dai mai*
KID 19								▒											■	
KID 20								▒											■	
KID 21								▒											■	
P 1									▒	▨	■	■								
T.B. 13										░					▨					
T.B. 15										░	■				■					
T.B. 17										░	■									
T.B. 20						▨				░	■									
T.B. 22						■				░	■									
G.B. 1						■				■	░									
G.B. 3			■							■	░									
G.B. 4			■							■	░									
G.B. 5		■	■							■	░									
G.B. 6		■	■							■	░									
G.B. 7							■				░									
G.B. 8							■				░									
G.B. 9							■				░									
G.B. 10							■				░									
G.B. 11						▨	■			▨	░									
G.B. 12							■				░									
G.B. 13											░				■					
G.B. 14		▨	▨							▨	░				■					
G.B. 15							■				░				■					
G.B. 16											░				■					
G.B. 17											░				■					
G.B. 18											░				■					
G.B. 19											░				■					
G.B. 20										▨	░				■		■			
G.B. 21			▨							■	░				■					
G.B. 23							▨				░									
G.B. 24				■							░				▨					
G.B. 26											░									■
G.B. 27											░									■
G.B. 28											░									■
G.B. 29											░						■			▨
G.B. 30							■				░									
G.B. 35											░						▨			
LIV 13											■	▒								▨
LIV 14				■								▒				■				
Du 1								■			■		░	■						
Du 13							■						░							
Du 14		■	■			■	■			■	■		░							
Du 15													░		■					
Du 16													░		■		▨			
Du 17							■						░							
Du 20							■			■	■	■	░							
Du 24			■				■						░							
Du 26		■	■										░							
Du 28			▨										░	■						
Ren 1													■	▒					■	

	Lung	Large Intestine	Stomach	Spleen	Heart	Small Intestine	Bladder	Kidney	Pericardium	Triple Burner	Gall Bladder	Liver	*Du mai*	*Ren mai*	*Yang wei mai*	*Yin wei mai*	*Yang qiao mai*	*Yin qiao mai*	*Chong mai*	*Dai mai*
Ren 2												■		░						
Ren 3				■				■				■		░						
Ren 4				■				■				■		░						
Ren 7								▨						░					■	
Ren 10			▨	■										░						
Ren 12			■			■				▨				░						
Ren 13			■			■								░						
Ren 17				■		■		■		■				░						
Ren 22														░		■				
Ren 23														░		■				
Ren 24		■	■										■	░						

▨ only mentioned by some authors

8.1.11 The Gao Wu Command Points

Synonym: none known
These points have the ability to affect certain regions of the body. By combining these points with local and other specific points they can increase the therapeutic effect in a particular area of the body.

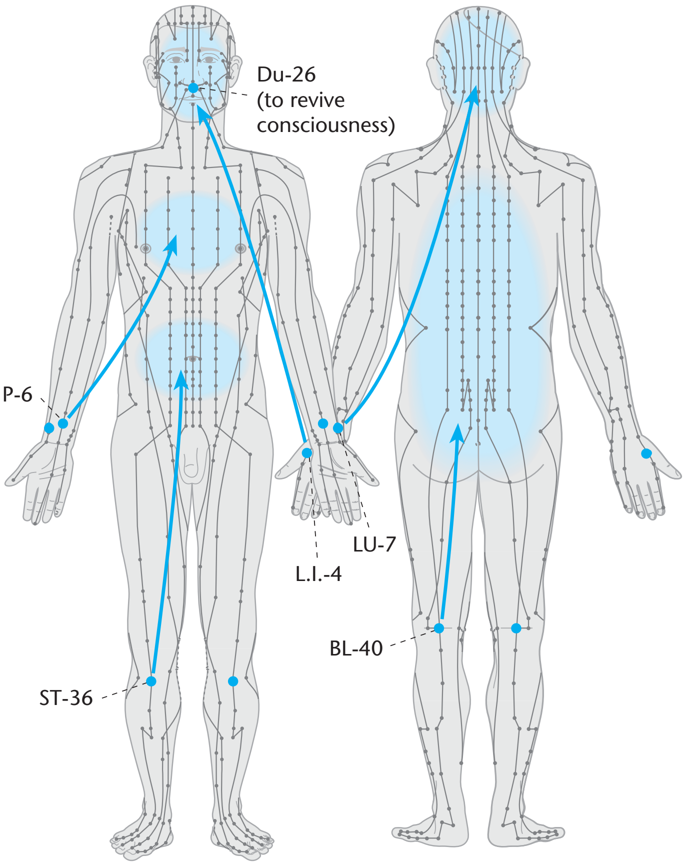

Points and clinical application

- **ST-36** (*zusanli*): for all disorders of the abdomen
- **L.I.-4** (*hegu*): for disorders of the face and mouth
- **LU-7** (*lieque*): for disorders of the occiput
- **BL-40** (*weizhong*): for disorders of the back and lumbar region
- **P-6** (*neiguan*): for disorders of the thorax
- **Du-26** (*renzhong*): for restoring consciousness (The last two points were added later, → Deadman et al 1998)

8.1.12 The Window of Heaven Points

In Chinese classical literature there are only a few references to these points. In modern times, both Ross (1995) and Deadman et al (1998) mention this group of points.

Points

LU-3 (*tianfu*)	**L.I.-18** (*futu*)	**P-1** (*tianchi*)
T.B.-16 (*tianyou*)	**S.I.-16** (*tianchuang*)	**S.I.-17** (*tianrong*)
ST-9 (*renying*)	**BL-10** (*tianzhu*)	**Ren-22** (*tiantu*)
Du-16 (*fengfu*)		

Location: Eight of the Window of Heaven points are located in the neck region (exceptions are **LU-3** on the upper arm and **P-1** on the thorax). This might provide a clue about the action of these points in regulating the Qi flow between the head and the body.
Comment: According to Deadman, the Ming dynasty physician Ma Shi remarked that **S.I.-17** should be replaced by **G.B.-9** (*tianchong*). If that was the case all six Yang channels running to the head would be represented by a Window of Heaven point.

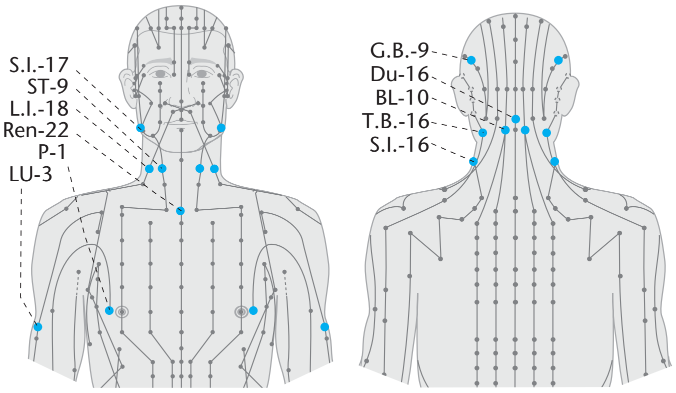

Clinical application

- For disorders of the Qi flow:
  - **LU-3, L.I.-18, Ren-22** for counterflow Qi causing cough, wheezing, etc.
  - **P-1** for counterflow Lung Qi with copious Phlegm
  - **ST-9** for counterflow Lung and Stomach Qi
  - **BL-10** treats dizziness, headaches and stiffness of the occiput, etc.
  - **Du-16** for internal Liver Wind

- For goitre, swellings, pain and Qi stagnation in the occiput and neck: especially indicated if these occur locally
- For disorders with an acute onset, for example
  - **L.I.-18** for acute aphonia
  - **ST-9** for acute diarrhoea
  - **S.I.-16** for acute aphonia, for example following a stroke
  - **BL-10** for epilepsy
  - **T.B.-16** for acute hearing loss
  - **Ren-22** for acute dyspnoea
  - **Du-16** for acute aphasia following a stroke
- For psychological disorders:
  - **LU-3** for somnolence, grief, disorientation, forgetfulness, insomnia
  - **S.I.-16** for manic agitation and bipolar disorder
  - **BL-10** for mania, hallucinations, epilepsy, childhood convulsions
  - **T.B.-16** for vivid dreaming
  - **Du-16** for mania, grief, anxiety with anxiety-induced palpitations
- For disorders of the sensory organs:
  - **LU-3** for nose bleeding, blurry vision, myopia
  - **ST-9** for blurry vision
  - **S.I.-16** for deafness, tinnitus, ear pain
  - **S.I.-17** for tinnitus and deafness
  - **BL-10** for eye pain, blurry vision, excessive tearing, speech disorders, blocked nasal passages, loss of sense of smell
  - **T.B.-16** for hearing disorders, visual disorders, eye pain, excessive tearing, loss of sense of smell, blocked nasal passages
  - **Du-16** for speech disorders (for example following a stroke), blurry vision, nosebleeds
  - **Ren-22** for speech disorders

8.1.13 Points of the Four Seas

These points have a particular, supportive effect on their respective 'Sea'.

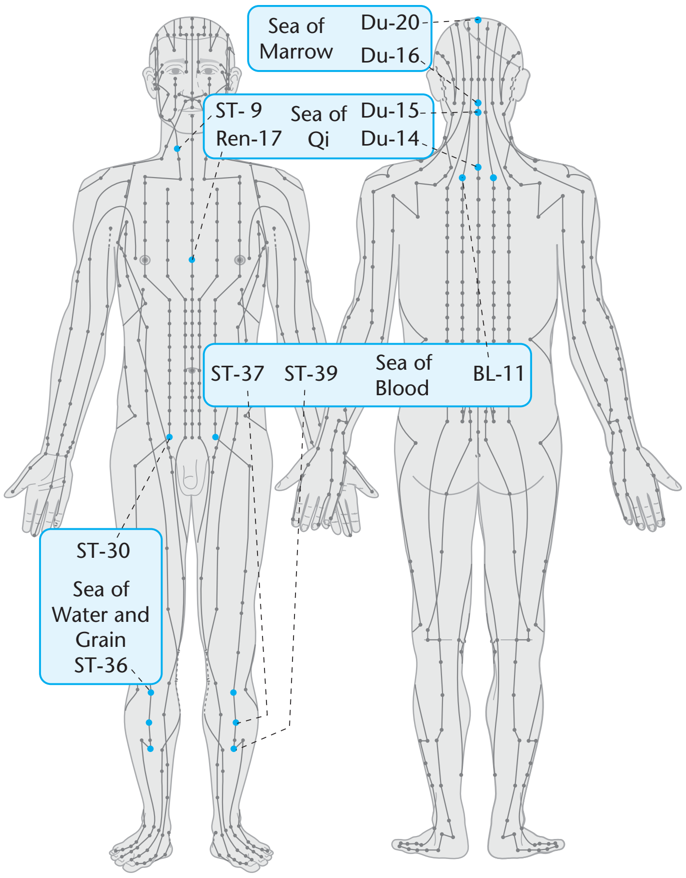

Sea of Qi	**ST-9** (*renying*), **Ren-17** (*danzhong or shanzhong*), **Du-15** (*yamen*), **Du-14** (*dazhui*)
*Ling Shu**: 'When the Sea of Qi is in excess there is fullness in the chest, urgent breathing and a red complexion. When the Sea of Qi is insufficient, there is scanty energy, insufficient for speech.'	
Sea of Blood	**BL-11** (*dazhu*), **ST-37** (*shangjuxu*), **ST-39** (*xiajuxu*)
*Ling Shu**: 'The Penetrating vessel is the sea of the twelve channels. In the upper it is conveyed to BL-11 (*dazhu*), and in the lower it emerges at both the ST-37 (*shangjuxu*) and ST-39 (*xiajuxu*) …When the Sea of Blood is in excess, then one has a sensation of the body being big; one feels disquiet, but does not know what the illness is; when the Sea of Blood is insufficient, one has the sensation of one's body being small; one feels reduced but does not know what the illness is.'	
Sea of Water and Grain	**ST-30** (*qichong*) as upper point, **ST-36** (*zusanli*) as lower point
*Ling Shu**: 'When the Sea of Water and Grain is in excess, there is abdominal fullness, and when it is deficient, there is hunger with inability to eat.'.	
Sea of Marrow	**Du-20** (*baihui*), **Du-16** (*fengfu*)
*Ling Shu**: 'When the Sea of Marrow is in excess, there is lightness of the body and much strength, and a person's feeling of 'self' exceeds the normal level; when the Sea of Marrow is insufficient there is a whirling sensation of the brain, dizziness, tinnitus, pain of the lower legs, impairment of vision, indolence and desire to sleep.'	

---

* Quotes from the *Ling Shu* according to Deadman et al. (1998).

8.1.14 The 12 Heavenly Star Points of Ma Dan Yang

Ma Dan Yang, a famous physician of the Jin dynasty, considered these initial 11 points as the most important acupuncture points. It was the physician Xu Feng who later added **LIV-3** as the 12th point.

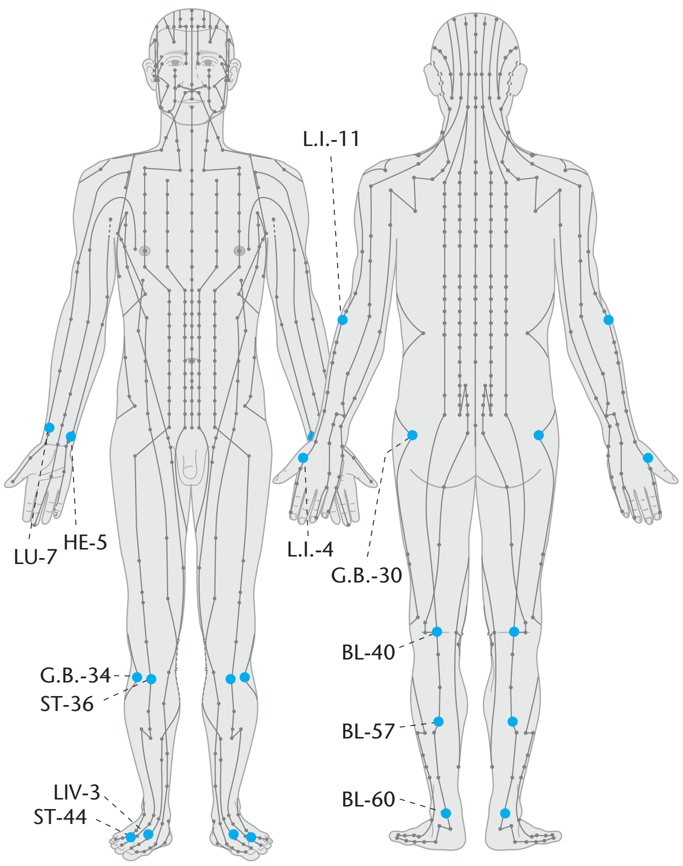

LU-7 (*lieque*)
ST-36 (*zusanli*)
BL-40 (*weizhong*)
G.B.-30 (*huantiao*)
L.I.-4 (*hegu*)
ST-44 (*neiting*)
BL-57 (*chengshan*)
G.B.-34 (*yanglingquan*)
L.I.-11 (*quchi*)
HE-5 (*tongli*)
BL-60 (*kunlun*)
LIV-3 (*taichong*)

8.1.15 The 13 Ghost Points of Sun Si Miao

Sun Si Miao, a famous physician of the Tang Dynasty, applied these 13 points (also called demon points) for treating disorders that today would be considered severe, manic, psychological disorders and/or epilepsy. Each *Sun Si Miao* Ghost point has an alternative Chinese name containing the component 'demon' or 'ghost' (see list according to Deadman et al. 2000). According to Deadman et al., some authors assumed that *guixin* is **LU-9** (*taiyuan*) rather than **P-7** (*daling*), while *guilu* is believed to be **P-5** (*jianshi*) or **P-8** (*laogong*) rather than **BL-62** (*shenmai*). There also exist different versions of this list of points; for example the physician Gao Wu omitted **BL-62, Du-23, Ren-1*** (in the original text of *Sun Si Miao: yumentou/yinxiafeng*) and **L.I.-11**, adding **Du-24, ST-17, LIV-2 and G.B.-34** instead.

	Name	Alternative name	Translation
Du-26	*renzhong*	*guigong*	Ghost Palace
Du-16	*fengfu*	*guizhen*	Ghost Pillow
Du-23	*shangxing*	*guitang*	Ghost Hall
Ren-24	*chengjiang*	*guishi*	Ghost Market
LU-11	*shaoshang*	*guixin*	Ghost Faith
L.I.-11	*quchi*	*guitui*	Ghost Leg
ST-6	*jiache*	*guichuang*	Ghost's Bed
SP-1	*yinbai*	*guilei*	Ghost Fortress
P-7	*daling*	*guixin*	Ghost Heart
P-8	*laogong*	*guicu*	Ghost Cave
BL-62	*shenmai*	*guilu*	Ghost Path
Ex-HN-11	*haiquan*	*guifeng*	Ghost Seal
Ren-1*	*huiyin*	*guicang*	Ghost Store

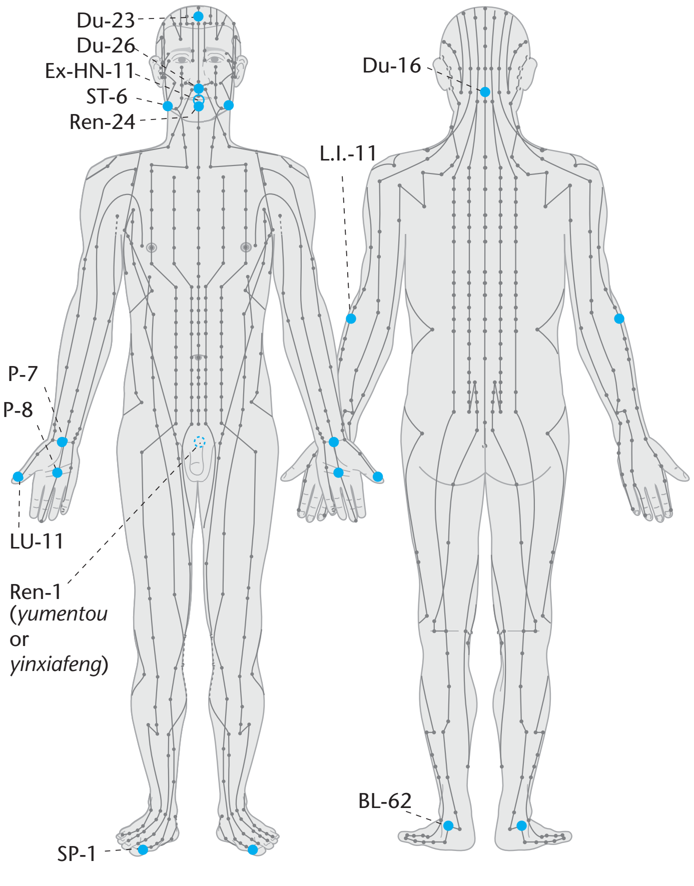

---

* *Yumentou* (in women) and *yinxiafeng* (in men): these two extra points approximately correspond to Ren-1 (*huiyin*/*guicang*) (Ghost Store).

8.1.16 Entry/Exit Points

While in modern Chinese literature entry/exit points are not mentioned as a separate point category, they are listed as such in some Western literature (for example Jarrett 2003, Hicks et al. 2004, Pirog 1996). However, Hicks et al. (→ 2004, p. 250) postulate that original Chinese sources exist for these points. The concept of entry/exit points is based on a continuous flow of Qi from one channel to the next, the sequence corresponding to the Organ clock (→ 1.1.4). Shunts are suggested in order to facilitate the flow of Qi from the end of one channel to the beginning of the next. At an entry point the Qi therefore flows from the preceding channel into the following one, while at the exit points it leaves the current channel in order to enter the next one.
However, it should be noted that the entry/exit points do not always correspond to the first or last point on a channel. In the following table these points are marked by grey shading.

Channel (sequence according to Organ clock)	**Time of maximum Qi flow**		**Entry/exit points in sequential order**		
LU (LU-1–LU-11)	3–5am		LU-1 (*zhongfu*)	➡	LU-7 (*lieque*)
					⬇
L.I. (L.I.-1–L.I.-20)	5–7am		L.I.-20 (*yingxiang*)	⬅	L.I.-4 (*hegxu*)
			⬇		
ST (ST-1–ST-45)	7–9am		ST-1 (*chengqi*)	➡	ST-42 (*chongyang*)
					⬇
SP (SP-1–SP-21)	9–11am		SP-21 (*dabao*)	⬅	SP-1 (*yinbai*)
			⬇		
HE (HE-1–HE-9)	11am–1pm		HE-1 (*jiquan*)	➡	HE-9 (*shaochong*)
					⬇
S.I. (S.I.-1–S.I.-19)	1–3pm		S.I.-19 (*tinggong*)	⬅	S.I.-1 (*shaoze*)
			⬇		
BL (BL-1–BL-67)	3–5pm		BL-1 (*jingming*)	➡	BL-67 (*zhiyin*)
					⬇
KID (KID-1–KID-27)	5–7pm		KID-22 (*bulang*)	⬅	KID-1 (*yongquan*)
			⬇		
P (P-1–P-9)	7–9pm		P-1 (*tianchi*)*	➡	P-8 (*laogong*)*
					⬇
T.B. (T.B.-1–T.B.-23)	9–11pm		T.B.-22 (*erheliao*)	⬅	T.B.-1 (*guanchong*)
			⬇		
G.B. (G.B.-1–G.B.-44)	11pm–1am		G.B.-1 (*tongziliao*)	➡	G.B.-41 (*zulinqi*)
					⬇
LIV (LIV-1–LIV-14)	1–3am		LIV-14 (*qimen*)	⬅	LIV-1 (*dadun*)

---

* In women P-2 is often substituted for P-1 due to its anatomical location close to the breast.

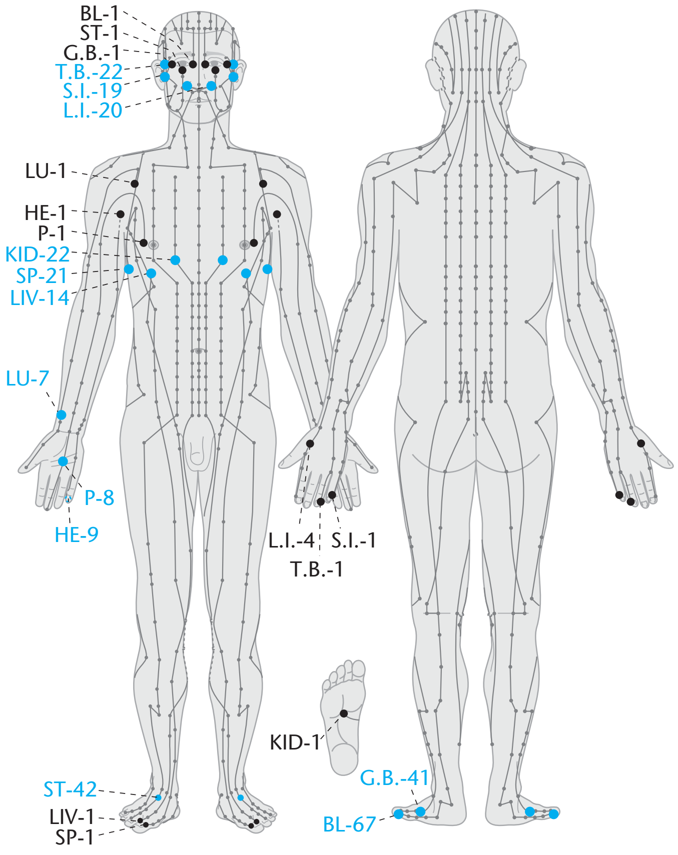
BL-1
ST-1
G.B.-1
T.B.-22
S.I.-19
L.I.-20
LU-1
HE-1
P-1
KID-22
SP-21
LIV-14
LU-7
P-8
HE-9
L.I.-4
S.I.-1
T.B.-1
KID-1
G.B.-41
ST-42
LIV-1
SP-1
BL-67

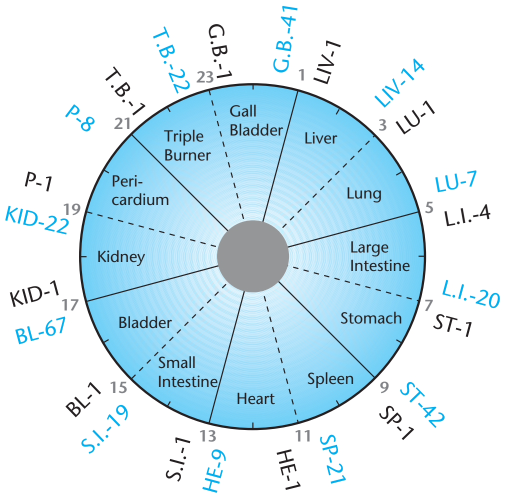
G.B.-41
LIV-1
1
LIV-14
LU-1
3
LU-7
L.I.-4
5
L.I.-20
ST-1
7
ST-42
SP-1
9
SP-21
HE-1
11
HE-9
S.I.-1
13
S.I.-19
BL-1
15
BL-67
KID-1
17
KID-22
P-1
19
P-8
T.B.-1
21
T.B.-22
G.B.-1
23
Gall Bladder
Liver
Lung
Large Intestine
Stomach
Spleen
Heart
Small Intestine
Bladder
Kidney
Peri-cardium
Triple Burner

Entry/exit block and therapy

Pulse diagnosis can reveal the partial or complete block between entry and exit points. An entry/exit block will inhibit the Qi flow from one channel to the next. In order to restore the blocked flow, the exit point of the blocked channel and the entry point of the following channel are needled at the same time. According to Hicks et al. (2004), further exit and entry points preceding or following the affected channel can be added in order to regulate the Qi flow over a greater distance. Needling should always be **bilateral**, even if the disorder occurs only on one side of the body. If it is not only the shunt between two channels, but the circulation in a channel as a whole that is affected, the flow in the whole channel can be strengthened and stimulated by simultaneous needling of both its entry point and its exit point.

Pirog (1996) compares the entry/exit points to input and output valves regulating the flow in a series of pipes (the channels) by 'opening' or 'closing' them.

- Tonifying needling techniques at the entry point (opening the input valve) will increase the flow into the channel. For example: tonifying **LU-1** opens the input valve. This will increase Qi flow in the LU channel for deficiency in this channel.
- Tonifying needling techniques at the exit point (opening the output valve) will decrease the flow in the channel. For example: Tonifying **LIV-14** opens the output valve: excess in the LIV channel will be drained.
- Reducing needling techniques at the entry point (closing the input valve) will reduce the flow into the channel. For example: Reducing **LU-1** closes the input valve and reduces the flow from the LIV channel into the LU channel, for example if there is pre-existing excess in the LU channel.
- Reducing needling techniques at the exit point (closing the output valve) will strengthen the flow in the channel. For example: Reducing **LIV-14** closes the output valve and prevents a further loss of Qi if there is a pre-existing deficiency in the LIV channel.

8.2 Strategies for Point Selection

8.2.1 Local, Adjacent and Distal Points

Local points

Local points are located in the immediate vicinity of the affected region. Each point that is tender with pressure (*ashi* point) can be considered to be a local point.

Adjacent points

Adjacent points are located near the affected or painful region.

Distal points

Distal points, despite being located at a distance from the affected area, still have a therapeutic effect, either directly or by being connected to the affected area by a channel (→ 1, also → 8.2.2, selection of distal points). The most effective and dynamic distal points are located distal to the knee (for the leg) or the elbow (for the arm). Generally the distal points of the foot channels tend to be more dynamic and effective than those on the arm channels.

Needling technique: In acute conditions or with severe pain the distal points should be needled with a reducing technique in order to activate the Qi flow more strongly in the blocked channels.

Stimulation of distal points: For a limited range of motion accompanied by pain the relevant distal point is strongly stimulated with a reducing technique while the patient moves the affected joint.

Enhancing the therapeutic effect: The therapeutic effect can be enhanced by combining distal points on the hand channels with those on the foot channels or by combining distal points with relevant local and adjacent points (→ 8.3.1).

8.2.2 Point Selection According to Affected Channel

Here the method of selecting points is based on the channel system (→ 1):

- Selecting points on the channel which **traverses the affected area**
- Selecting points on the channel that is **Interiorly–Exteriorly pertaining** (→ 1.2) to the affected channel
- Selecting points from the channel which is connected to the affected channel by **hand–foot pairing** (for example, *taiyang* axis, etc. → 1.2.3)

Differential selection of distal points

The following possibilities for selecting distal points are, among others, based on course material about the 'one point therapy' by R. Thambirajah (1990, 1991), R. Tan (2003) and my own clinical experience.

Selection of distal points according to corresponding areas

The distal points can be selected according to corresponding areas. The affected area is carefully examined, then the corresponding area is needled contralaterally (→ Fig.).

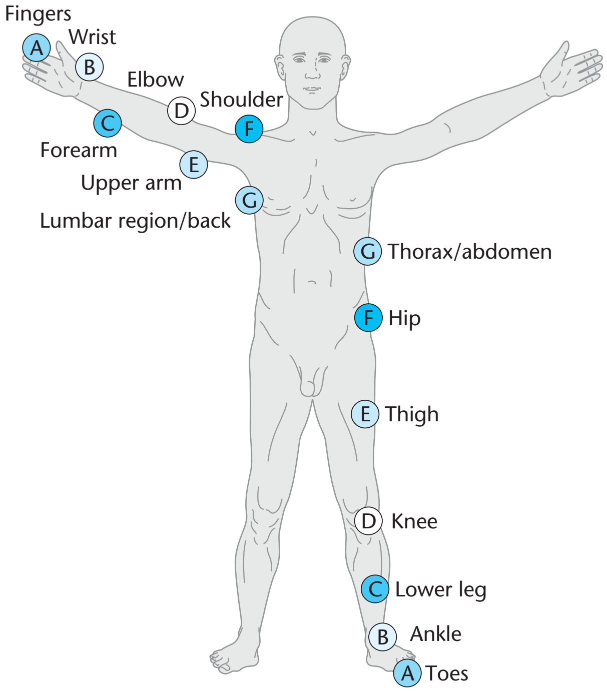

Selection of distal points according to Interiorly–Exteriorly connected channels

Here points are selected on the Interiorly–Exteriorly connected channel (Yin/Yang). The affected area is carefully located, then the corresponding point (mirror point) is needled with reducing technique on the contralateral side on the Interiorly–Exteriorly pertaining channel. Another specific example for this method of point selection is the *yuan–luo* combination (→ 8.3.3).

Example: For a disorder affecting the ST channel (foot *yangming*) select a point on the Interiorly–Exteriorly pertaining SP channel (foot *taiyin*). For example, for knee pain at **ST-35**, needle **SP-9** contralaterally (→ Fig.).

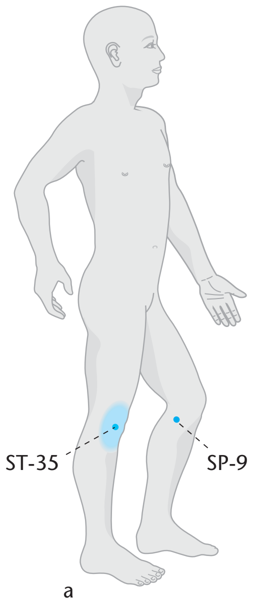

Selection of distal points according to the six great channels (*liu jing*, hand–foot pairing, ➞ 1.2.3)

The point selection with this method is based on the six great channels (hand–foot pairing: *taiyin, taiyang*, etc., ➞ 1.2.3). Disorders affecting the area of the hand channel are therefore treated with points on the corresponding foot channel and vice versa.

Example: Disorders in the area of the hand *taiyin* channel (LU channel) can be balanced by contralateral needling of points on the foot *taiyin* channel (SP channel). For example, shoulder pain with limited range of motion and maximum pain near **LU-1** (*zhongfu*) can be improved by contralateral needling of **SP-9** (*yinlingquan*) with reducing technique.

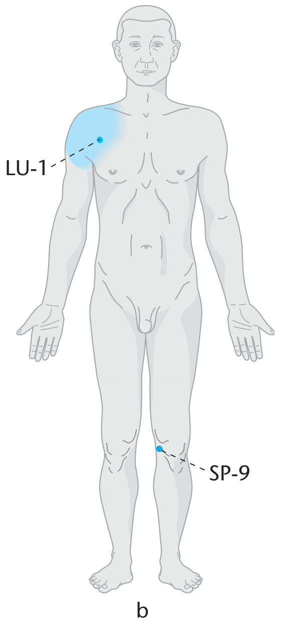

Selection of distal points according to the Organ clock (➞ 1.4)

Distal points are selected on the channel that lies chronologically opposite to the affected channel on the Organ clock (➞ 1.4). With this method, hand channels will treat disorders affecting the foot channels and vice versa.

Example: For a disorder affecting the G.B. channel (foot *shaoyang*), needle a point on the HE channel (hand *shaoyin*) located opposite the G.B. channel on the Organ clock. Needling can be contralateral or ipsilateral.

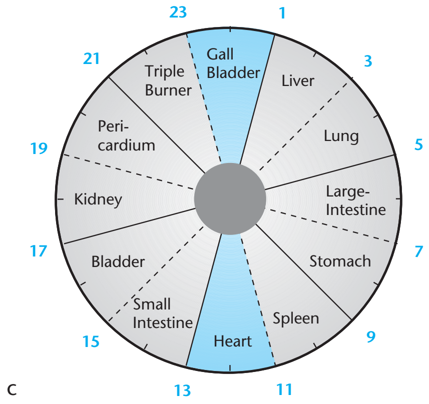

8.2.3 Point Selection Based on Symptoms

Points can also be selected based on traditional or empirical therapeutic experiences. While this method cannot replace a root treatment based on a differential diagnosis, it can be very helpful in situations where quick action is required.

Examples:

- **Du-14** or **L.I.-11** for fever
- **ST-36, Ren-12** for acute gastrointestinal disorders

8.2.4 Point Selection Based on the Qi Flow of the Organs

Descend counterflow Qi:

- Stomach: **Ren-10, Ren-13, ST-34, ST-44, ST-45, L.I.-4**
- Lung: **LU-1, LU-5, LU-7**
- Heart: **HE-5, HE-8, Ren-15**
- Kidney: **KID-1, KID-7, Ren-4**
- Liver: **LIV-2, LIV-3, LIV-1, LIV-14**

Raise Qi:

Spleen: **Ren-6, Ren-12, Du-20, BL-20**

8.2.5 Point Selection According to the Five Phases (Elements)

While the Five Phase points are identical to the *shu*-transporting points in location, their function and application differ from the latter (for more detail → 8.1.6). There exist various very specific methods regarding diagnosis, point selection and therapy based on the Five Phases, which are practised by several different schools of thought. Within the framework of this Atlas, only a simplified overview of the possible clinical applications is listed below.

***Shu*-transporting points of the Yin channels**					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point
***Shu*-transporting points of the Yang channels**					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43 +	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

Cycles

The Five Phases influence each other both physiologically (*sheng* and *ke* cycle) as well as pathologically (*cheng*, *wu* and to some extent also *sheng* cycle). If the balance between the Five Phases is disturbed, pathological syndromes will appear.

- *Sheng* cycle (generating cycle, mother–son cycle): a phase generates and nourishes the following phase. Each phase nourishes (as mother) and is nourished (as son) at the same time. Fire nourishes Earth, Earth nourishes Metal, Metal nourishes Water, Water nourishes Wood, and Wood nourishes Fire. In a pathological condition either the mother is too weak to nourish the son sufficiently or the son is too strong, draining his mother and weakening her.
- *Ke* cycle (controlling cycle): one phase controls another and is itself controlled by yet another phase.
- *Cheng* cycle (overacting cycle): The controlled phase is pathologically suppressed or weakened.
- *Wu* cycle (insulting cycle): A phase is pathologically stronger than its controlling phase.

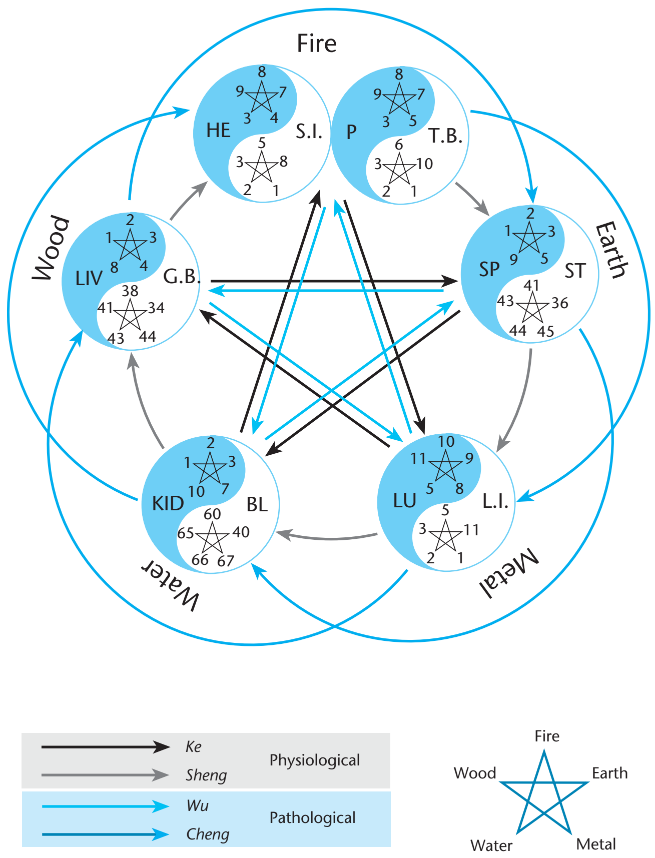

Simple point selection according to the *sheng* cycle

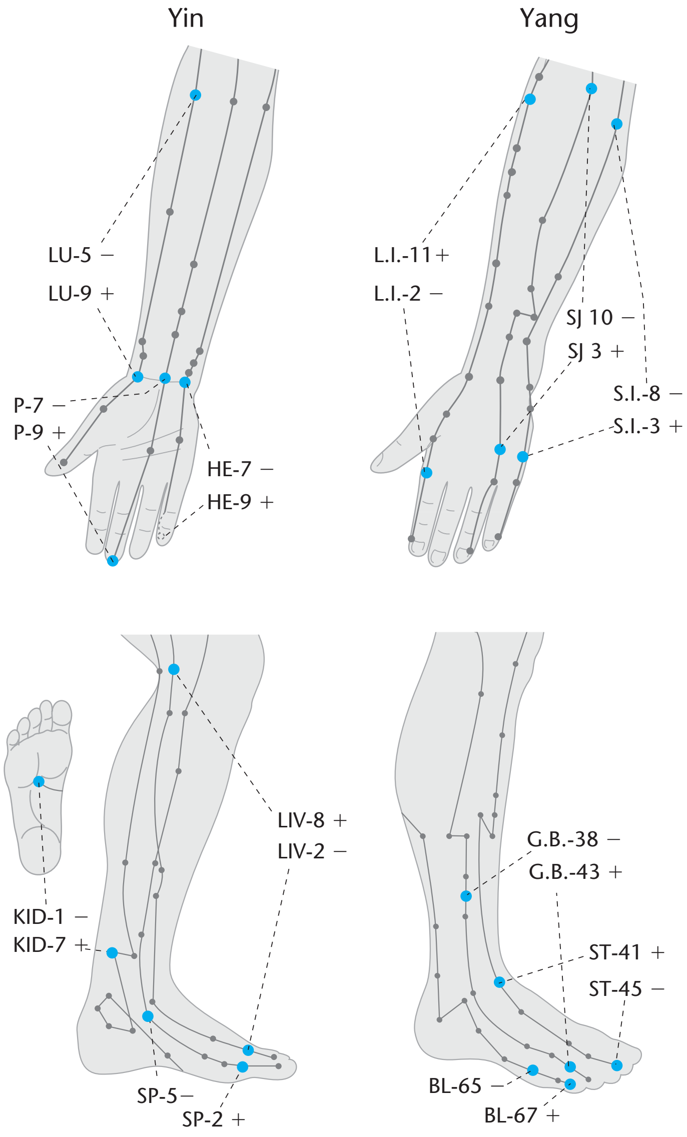
Yin
Yang
LU-5 –
LU-9 +
P-7 –
P-9 +
HE-7 –
HE-9 +
L.I.-11 +
L.I.-2 –
SJ 10 –
SJ 3 +
S.I.-8 –
S.I.-3 +
LIV-8 +
LIV-2 –
KID-1 –
KID-7 +
SP-5 –
SP-2 +
G.B.-38 –
G.B.-43 +
ST-41 +
ST-45 –
BL-65 –
BL-67 +

Treatment principle: For deficiency tonify the mother, for excess drain (sedate) the son. Mother–son points are identical to the tonification and sedation points of a channel. The correct needling technique is important in order to obtain the desired result:

- tonifying method for tonification points
- reducing method for sedation points

Example for the clinical application of the *sheng* cycle

For deficiency: In cases of deficiency of a particular channel/Organ, tonify that point on the affected channel that corresponds to the phase of the mother.

Example: Earth is the mother of Metal. If the Lung (Metal) is deficient, needle the Earth point (**LU-9**) on the Lung channel with tonifying technique.

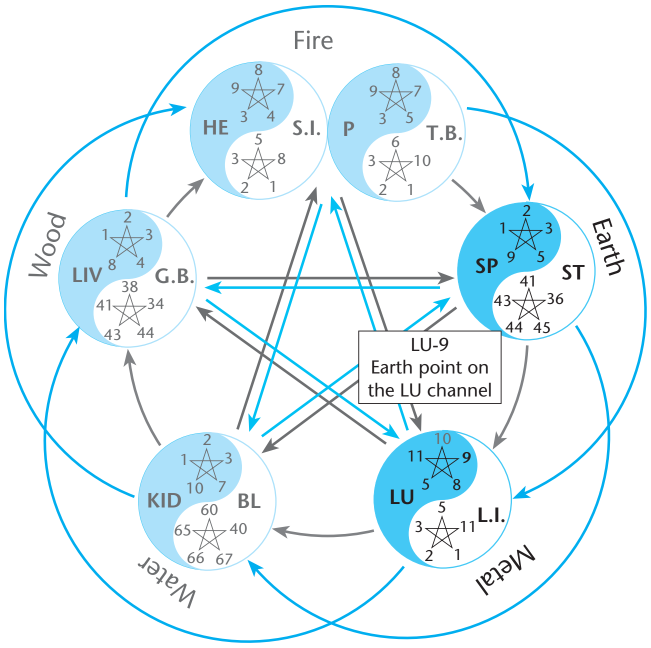
Fire
Earth
Metal
Water
Wood
HE
S.I.
P
T.B.
SP
ST
LU
L.I.
KID
BL
LIV
G.B.
LU-9
Earth point on
the LU channel

For excess: If there is excess in a particular channel/Organ, select that point on the channel which corresponds to the phase of the son and needle it with reducing technique.

Example: Water is the son of Metal. If the Lung (Metal) is in excess, needle the Water point (LU-5) on the Lung channel with reducing technique.

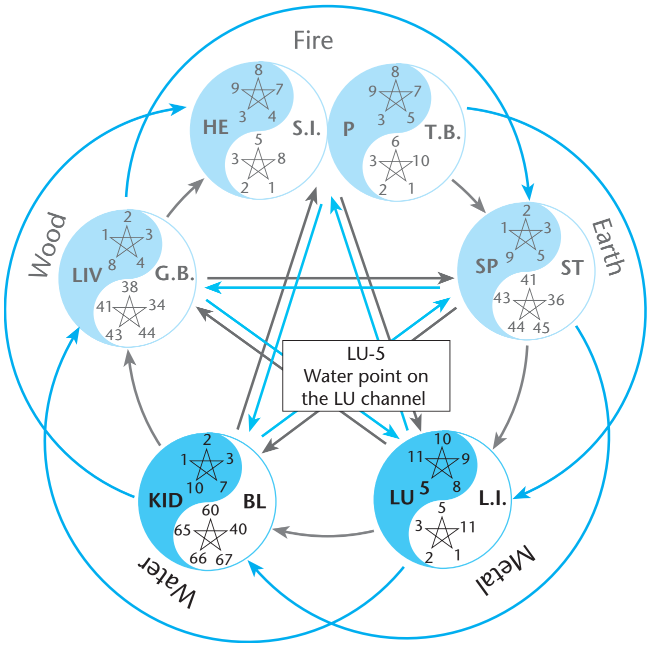
Fire
Earth
Metal
Water
Wood
HE
S.I.
P
T.B.
SP
ST
LU
L.I.
KID
BL
LIV
G.B.
LU-5
Water point on
the LU channel

Addendum: The terms tonification/sedation points are somewhat controversial, since the properties of a point can often be overshadowed by its other features. For example, according to the Five Phase theory, **P-9** and **HE-9** are considered tonification points, since they correspond to the phase of the mother. However, in clinical practice they are much more frequently used as *jing-we*ll points for acute situations – thus they are considered from the point of view of the Qi flow in the channels, especially in order to drain Heat.

Application of the Five Phase points for the elimination of external pathogenic factors

A connection exists between the Five Phases and pathogenic factors (Maciocia 1994; according to Ross 1998, however, these points should only be used for internal disorders):

- Wood corresponds to Wind
- Fire corresponds to Heat
- Earth corresponds to Dampness
- Metal corresponds to Dryness. According to Maciocia this correlation presents an exception: the Metal point should not be needled to treat Dryness; a moistening treatment is here the preferred choice.
- Water corresponds to Cold.

In accordance with these interrelations, Five Phase points can be applied to eliminate pathogenic factors.

Example: Acute sore throat with fever, redness and swelling of the throat due to Wind-Heat

Affected phases: Wood (Wind), Fire (Heat)

Affected Organ/channel: Lungs

Therapy: Needling of the Wood point on the LU channel (**LU-11**) and the Fire point on the LU channel (**LU-10**).

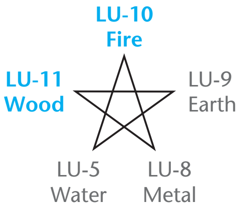

The application of the ***ke* cycle (controlling cycle)** allows a more differentiated approach in the therapy and elimination of pathogenic factors:

- On the Yin channel, needle the point corresponding to the pathogenic factor with reducing technique (for example, for Heat needle the Fire point)
- In addition, needle the point which corresponds to the controlling phase (according to the *ke* cycle/controlling cycle → 8.2.5) of the pathogenic factor on the paired Yang channel with tonifying technique.

Ben points

***Shu*-transporting points of the Yin channels**					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9+	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2−	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point
***Shu*-transporting points of the Yang channels**					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38−	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; –, sedation point

Ben points (root points, element points) represent the same phase as its pertaining channel. **Example:** The Spleen corresponds to Earth; thus the Earth point on the Spleen channel is its *ben* point.

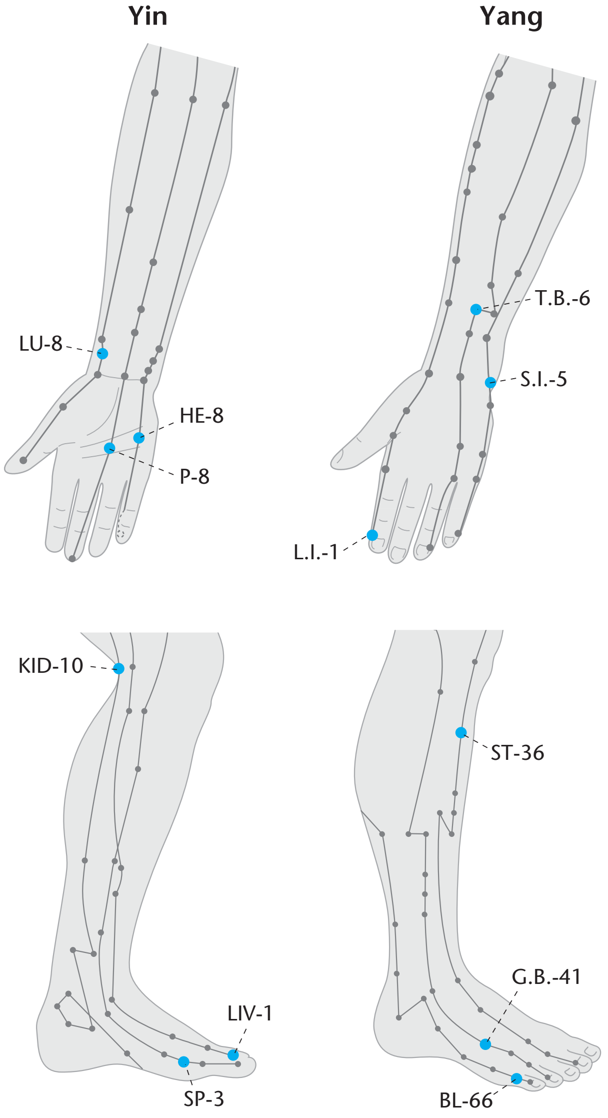

Clinical application

- **For deficiency conditions:** needling the *ben* point with tonifying technique will tonify its corresponding Organ or channel
- **For excess conditions:** needling the *ben* point with reducing technique will drain the excess from its corresponding Organ or channel
- **Spiritual aspect:** the *ben* points on the Yin channels also affect the spiritual aspect of their corresponding phases. Thus the LU *ben* point **LU-8** allows the *po* (corporeal soul) to unfold its potential. Similarly the HE *ben* point **HE-8** will empower the *shen* (spirit), the SP *ben* point **SP-3** the *yi* (thinking), the LIV *ben* point **LIV-1** the *hun* (ethereal soul) and the KID *ben* point **KID-10** the *zhi* (will).

8.3 Point Combinations

8.3.1 Combining Local and Distal Points

Point selection and combination can be carried out according to affected areas and channels (→ 8.2.1, → 8.2.2).

For disorders affecting the channel: For example, for *bi*-syndromes affecting the sinews and joints, distal points can be stimulated first. Then local and adjacent points can be selected based on tenderness. These are needled and/or subsequently cupped.

For disorders of the *zangfu*-Organs: Generally, only distal points are applied in acute conditions. Once the condition (for example, acute pain) has improved, local points can be selected. In chronic conditions the combination of local and distal points – and especially Back-*shu* and Front-*mu* points as adjacent points (→ 8.3.2) – is commonly used. Another possibility is the combination of the opening point of an extraordinary vessel (→ 8.1.8) with local points in the affected area.

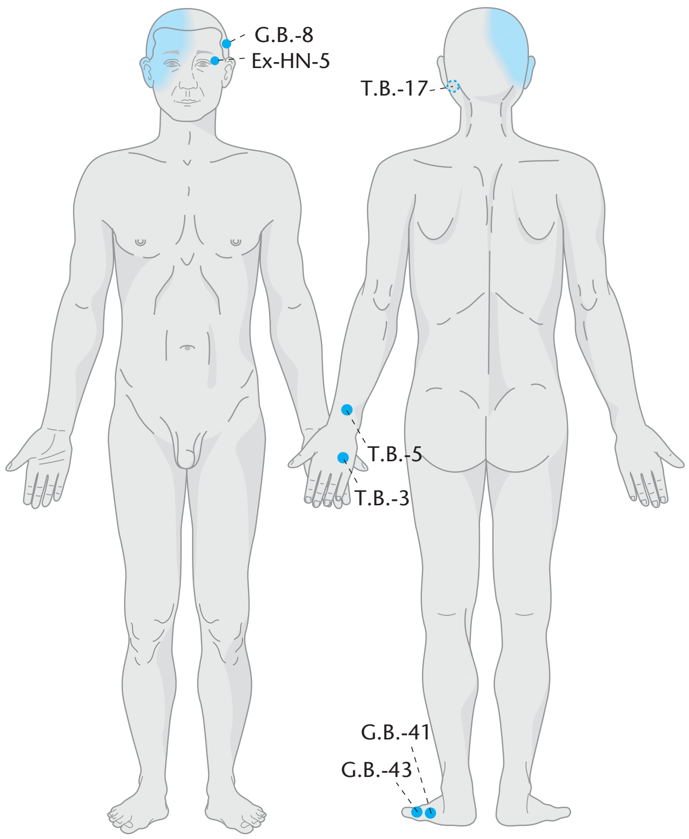

8.3.2 Combining Points on the Front and Back of the Body

Here, points on the anterior and posterior aspect of the body are combined:

- **Anterior points**: mainly located on the thorax and abdomen
- **Posterior points**: mainly located on the back and lumbar region

Back-*shu*/Front-*mu* method

The combination of the Back-*shu* point (→ 8.1.4) and the Front-*mu* point (→ 8.1.5) of the affected Organ is the most important way of combining points on the front and the back of the body. The Back-*shu*/Front-*mu* method enhances and extends the therapeutic effect, compared to using either of those points only. It has a strong Yin/Yang balancing effect and it is particularly beneficial for chronic disorders of the *zangfu*-Organs.

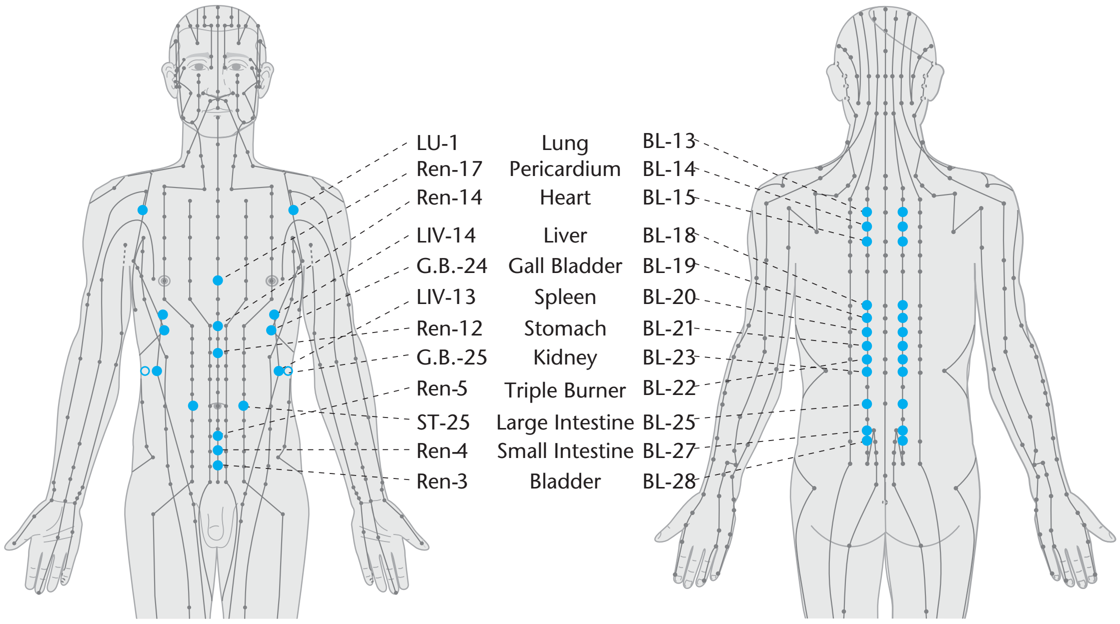
LU-1 Lung BL-13
Ren-17 Pericardium BL-14
Ren-14 Heart BL-15
LIV-14 Liver BL-18
G.B.-24 Gall Bladder BL-19
LIV-13 Spleen BL-20
Ren-12 Stomach BL-21
G.B.-25 Kidney BL-23
Ren-5 Triple Burner BL-22
ST-25 Large Intestine BL-25
Ren-4 Small Intestine BL-27
Ren-3 Bladder BL-28

Clinical application of the Back-*shu*/Front-*mu* method

- Needling of the Back-*shu* and Front-*mu* points during one session
- For repeated treatments within a short period of time, alternate needling of the Back-*shu* and Front-*mu* points (for example, during the first treatment select the relevant Back-*shu* point, in the next treatment select the corresponding Front-*mu* point, etc.)
- The Back-*shu*/Front-*mu* combination can also be applied to balance an incorrect treatment. For example, if the needles have been retained too long in the Back-*shu* points and the patient is tired, Front-mu points can be needled to restore balance.

Balancing the *du mai* and the *ren mai*

The points on the *ren mai* are located on the anterior aspect of the body, while those of the *du mai* are predominantly located on the posterior aspect. Combining points on the *ren mai* and *du mai* has a balancing affect on the Yin and Yang and can regulate the ascending and descending flow of Qi. This combination therefore not only balances the front and the back of the body, but also the Yin and the Yang. Furthermore, it has a strong effect on the psyche and, depending on the needling technique and point selection, will have a calming or stimulating effect (figure modified according to Ross 1998).

Ch08-F10028.qxd 2/23/08 4:41 PM Page 693

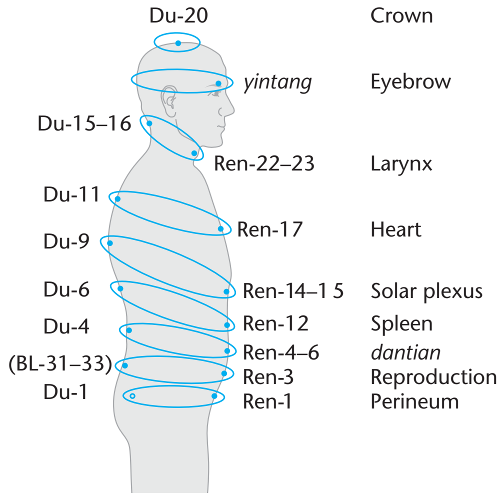

8.3.3 Yin–Yang Combination

Synonym: Interior/Exterior combination
The Yin–Yang combination balances the flow of Yin and Yang in the channels. It is based on the concepts of channel energetics (→ channel circuits, Fig. 1.12).

> Needling too many points on the Yang channels can make the patient nervous and restless. In this case, needling points on the Yin channels can restore balance and have a calming effect.
> Needling too many points on the Yin channels can lead to tiredness. In this case, needling points on the Yang channels has a balancing, stimulating effect.

Balancing Yin and Yang in paired channels

- Points on the Interiorly–Exteriorly paired channels are combined in order to balance Yin and Yang. For example the LU and L.I. channels are the Interiorly–Exteriorly paired channels of the hand, while the ST and SP channels are the Interiorly–Exteriorly paired channels of the foot of the first circuit.
- This method increases the therapeutic effect compared to using only Yin or only Yang points.

Yuan/luo combination

Most important Yin–Yang combination (Synonym: host–guest combination): The *yuan*-source point (host) of the primarily affected channel/Organ is combined with the *luo*-connecting point of the Interiorly–Exteriorly paired channel.

Important *yuan/luo* combinations

- **L.I.-4** (*hegu*) and **LU-7** (*lieque*) support the descending function of the Lung Qi, expel pathogenic factors, tonify the Defensive Qi (*wei qi*) and calm the *shen*
- **SP-3** (*taibai*) and **ST-40** (*fenglong*) tonify the Spleen and transform Phlegm
- **LIV-3** (*taichong*) and **G.B.-37** (*guangming*) benefit the eyes for Liver syndromes
- **T.B.-4** (*yangchi*) and **P-6** (*neiguan*) regulate the Triple Burner, move Liver Qi, calm the *shen*, relax the muscles in the occiput and shoulders.

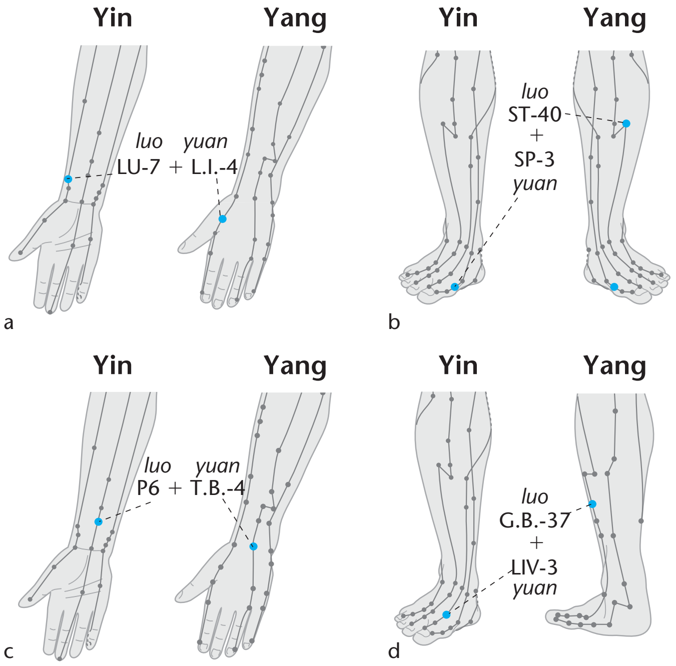

Further methods of combining points of Interiorly–Exteriorly paired channels and modified *yuan/luo* combinations

a Combining the *luo*-connecting points of Interiorly–Exteriorly paired channels tends to enhance the therapeutic effect; for example combine **ST-40** and **SP-4** for abdominal pain

b For chronic disorders, combine the *yuan*-source point and the *luo*-connecting point of the same Yin channel in order to increase the therapeutic effect; for example needle **LU-9** and **LU-7** for chronic cough.

c For disorders affecting a Yin channel, only needle the *yuan*-source point of the Interiorly–Exteriorly paired Yang channel. For example for an acute cold (a Lung syndrome), needle only **L.I.-4** (*yuan*-source point).

d For unilateral musculoskeletal disorders affecting a Yang channel: in addition to local and adjacent points on the affected channel, needle the contralateral *luo*-connecting point of the Interiorly–Exteriorly paired Yin channel. This method, which promotes balance between Yin and Yang, is mainly used for channel-pertaining disorders. For example: for unilateral arm or shoulder pain along the L.I. channel, needle points on the L.I. channel on the affected side as well as **LU-7** on the opposite side of the body.

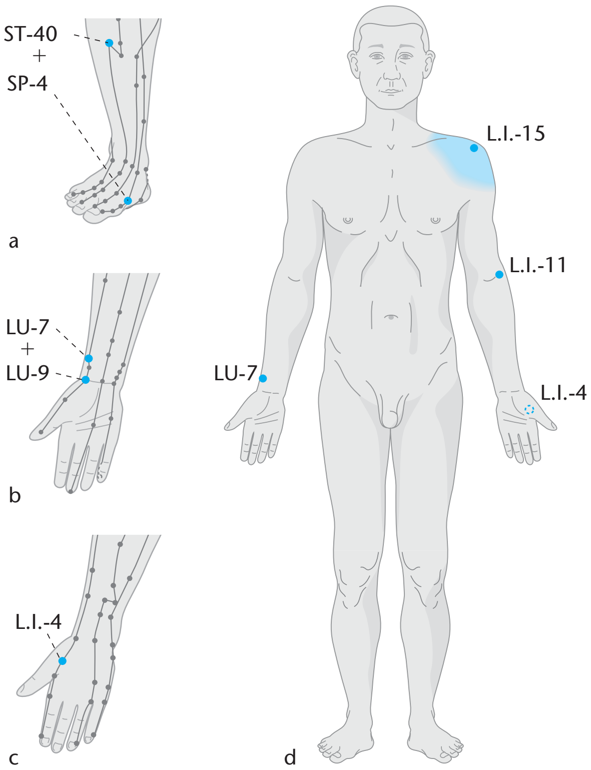

Yin–Yang balance for unpaired channels

Examples:

ST-36 + P-6	Harmonises the Middle Burner and descends counterflow Qi
L.I.-4 + LIV-3	Most important combination for regulating the Qi flow in the whole body

8.3.4 Combining Points from Above and Below

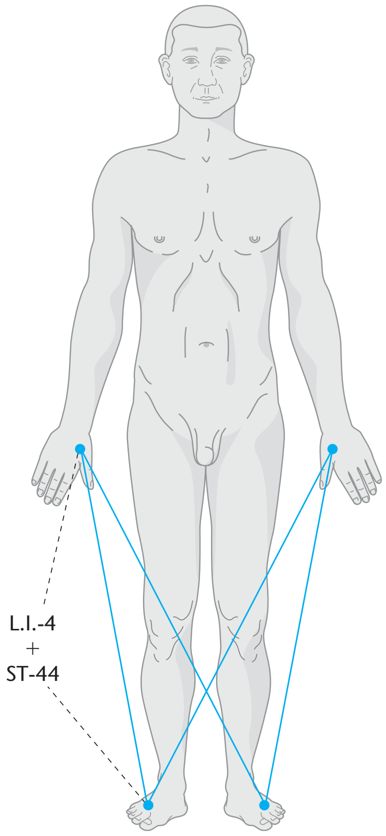

- A balance between the **upper and lower** half of the body is achieved by selecting points that are evenly distributed. For headaches, therefore, **L.I.-4** on the upper extremity is combined with **ST-44** on the lower extremity.
- Combining points above with points below **promotes the smooth flow of Qi**. For example: combining the opening point and coupled point of an extraordinary vessel (→ 8.1.8).

When points from above and below should not be combined

- for acute disorders of the joints and the back – here only distal points should be needled with the reducing method.
- if there is a pre-existing energetic disharmony between above and below. For example: for Heat sensations of the head (for example due to Kidney Yin deficiency with Empty Fire) needle **KID-1** in order to cause the Qi to descend. On the other hand, for a uterine prolapse use moxibustion **on Du-20** in order to raise the Qi.

8.3.5 Left–Right Combination

Often the balance between the right and left half of the body can be restored through bilateral needling of the relevant points. This will increase the therapeutic effect as opposed to unilateral needling.

Clinical application for Interiorly–Exteriorly paired channels (→ 1.2.3)

- For acute unilateral channel disorders caused by the invasion of pathogenic factors, combine **local points** on the affected side with the ***luo*-connecting point** on the contralateral side.
- For chronic unilateral channel disorders caused by the invasion of pathogenic factors, combine **local points** on the affected side with the *luo*-**connecting point** of the **Interiorly–Exteriorly paired channel** on the **contralateral** side (with tonifying technique).

Indications for unilateral and/or contralateral needling

- **For acute, painful disorders:** contralateral or diagonal needling (corresponding hand and foot channel, also see → 8.2.1 and → 8.2.2)
- **For chronic joint disorders:** ipsilateral needling of several points balanced by needling of contralateral points.
- **In paediatric acupuncture** (use fewer needles)
- **For ongoing treatment sessions:** occasional balancing contralateral needling of the healthy side of the body

8.3.6 Chain and Lock Point Association Method

Two, three or more points on the same channel are selected and needled one after the other (in a row). This method is predominantly applied for disorders of the musculoskeletal system or the nervous system.

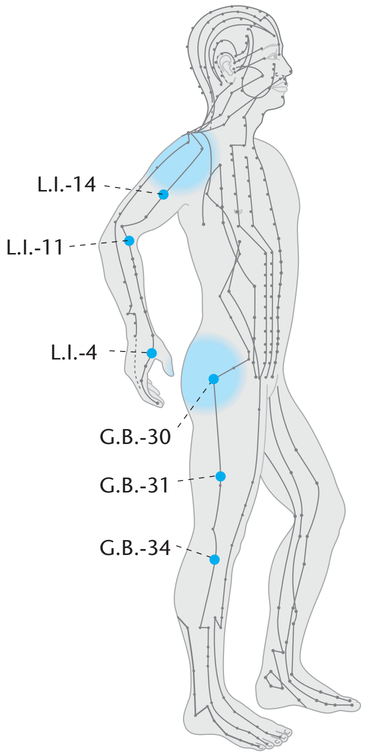

8.3.7 Point Selection According to the Organ Clock

Within a 24-hour cycle each channel experiences a two-hour period of maximum Qi flow (→ Fig. 1.17, Chinese Organ clock). Some modern (Western) schools extrapolate the circadian Qi flow to the Organs (internal channel pathways and divergent channels). If one Organ is at its maximum period, the Organ on the opposite side of the clock will be at its minimum period, in other words the Qi flow will at that time be at its lowest. Symptoms of a disorder occurring at a particular time may provide a hint about the Organ affected, which at that time will be during its **maximal period in excess syndromes** or during its **minimal period in deficiency syndromes**.

For example: Always waking up between 1 and 3 am can be a sign of a Liver syndrome. Thus the tonification point and/or *luo*-connecting point of the channel opposite to the Liver on the Organ clock could be selected, in this example **S.I.-3** (*houxi*, tonification point) or **S.I.-7** (*zhizheng*, *luo*-connecting point).

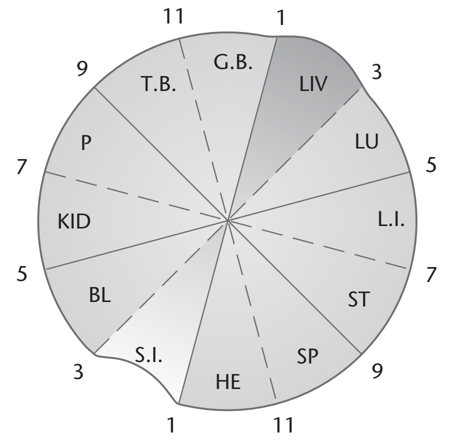

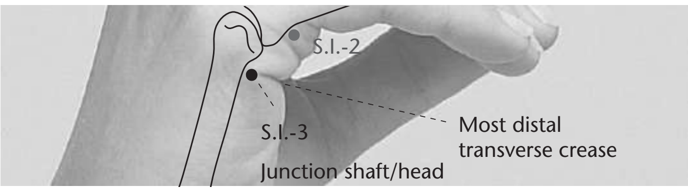

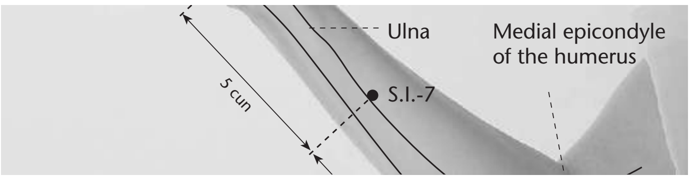

Addendum: There are some very specific methods for chrono-acupuncture, for exampel the *zi wu liu zhu* method and the *ling gui ba fa* method, where points are needled at specific times (→ Kubiena and Ramakers 2002).

8.3.8 Clinical Application

The following illustration shows the clinical application of several point selection methods (8.3.1 – 8.3.5) for the treatment of dysmenorrhoea.

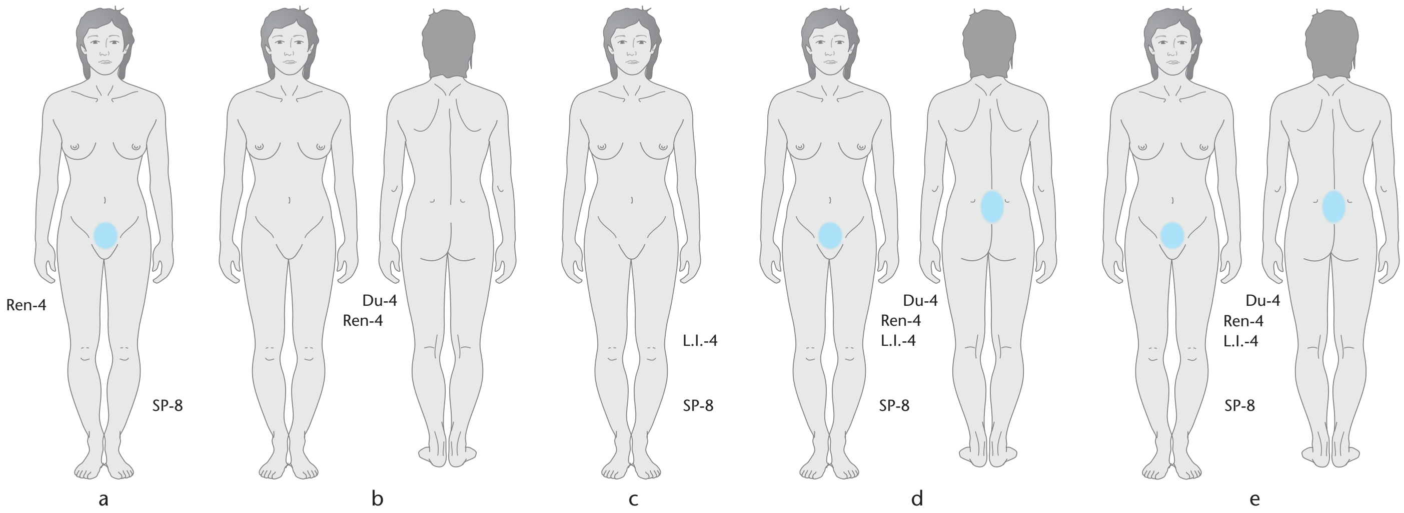

Possible point combinations for the treatment of dysmenorrhoea:

a. combining local and distal points (➞ 8.3.1)
b. combining points on the front and the back of the body (➞ 8.3.2) and Yin–Yang (➞ 8.3.3)
c. Yin–Yang combination (➞ 8.3.3) and above–below (➞ 8.3.4)
d. Combination of all of the above methods (➞ 8.3.1–8.3.4)
e. Additionally combining points on the left and right side of the body (➞ 8.3.5)

9 Scientific Research

Ingolf Hosbach

9.1 Introduction

The aim of this chapter is to give a comprehensive overview about the scientifically proven or disproved effectiveness of certain points. It is not the aim of this chapter to prove the effectiveness of acupuncture as such. Based on the latest research, and 1000 recent human studies with mostly positive results, it is assumed that acupuncture is effective, despite criticism. If applied properly it is a well-established, low-risk method with minimal side effects.

Accordingly, the scope of acupuncture research has developed from simple questions, such as whether acupuncture works, to more specific questions such as which acupuncture points can be used for which conditions and with which stimulation methods. Furthermore, studies are increasingly designed for precise diagnoses and treatments based on the principles of Traditional Chinese Medicine. By the same token, the quality of the biomathematical methods used for the design and analysis of the studies is increasing. The number of randomised, controlled studies (RCTs) is also on the rise (Fig. 1).

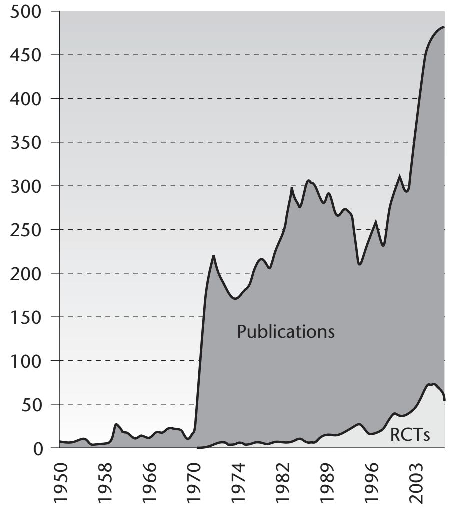

Fig. 1: Number of all publications and RCTs per year between 1954 and 2005. (Source: PubMed, US National Library of Medicine, Bethesda, Maryland)

The percentage of RCTs of all publications is nearly 7%, a figure that compares well with most conventional forms of therapy (for example: arthroscopy 5%, beta-blockers 7%, pain medication 7%, lowering cholesterol medication 8–11%).

Fundamental principles for scientific assessment

One of the prerequisites for including a study in this chapter was the mention of specific acupuncture points in the abstract. Out of 1092 studies listed in PubMed until May 2006 under the keyword 'acupuncture', 259 were therefore chosen for this chapter. In some exceptional cases, studies based on animal experiments were also considered. Regarding the presentation of the results, particular emphasis was placed on the role of biomathematical methods as a parameter for reliability. The terms which allow assessment of a study in terms of its biomathematical quality are explained below:

Case study: Usually a retrospective (=in hindsight) summary of results of a particular treatment method in patients with a particular disorder; no control group or control therapy. Most basic form of clinical trial.

Controlled study: Usually a prospective (=planned in advance) scientific comparison of a treatment method with a group of untreated patients or of patients treated with a placebo or conventional therapy.

Placebo: A form of 'mock' or sham therapy. This can take the form of a pill containing no active ingredient or a therapy without any authentic foundation. The administration of a placebo tries to eliminate the non-specific improvement of the researched disorder (such as the therapist's care of the patient and expectations of the patient) compared to the study group receiving authentic treatment. While the possibilities for placebo-acupuncture are plentiful, none is free of disadvantages. The effectiveness of acupuncture as a holistic therapy can be seen as the sum of many psychological as well as non-specific and specific physiological effects. Depending on the choice of the acupuncture placebo, certain effects can be excluded. Based on the concept of the sinew channels, however, there is an ongoing debate whether non-acupuncture points exist at all. The highest level of placebo acupuncture is the so-called 'sham acupuncture' where actual deep needling takes place at non-acupuncture points. The value of a study increases considerably if it is controlled by a placebo therapy.

Randomisation: The random assignment of a patient to a treatment group, without any influence of either the patient or the therapist. Randomisation significantly increases the value of a study.

Waitlist: A control group formed by patients receiving no treatment as a form of control. While a patient is waiting for his/her treatment, all his/her data relevant to the treatment are collected and later compared to those data obtained during treatment. Since there is neither randomisation nor a placebo treatment, the value of a waitlist group as a control group is very low.

Cross-over study: A particular design of controlled study. Each patient will participate in two different courses of treatment (for example treatment 1: medication; treatment 2: acupuncture). The order in which the treatments are administered can differ from patient to patient and be random (=randomised cross-over study). However, any long-term effect of the first form of treatment poses a problem for the final result. Therefore so-called washout phases during which patients receive no treatments at all are implemented to eliminate these effects and prevent a distortion of the data to be collected during the second phase. Since the effects of some acupuncture treatments have been shown to last for up to a year, cross-over studies are not very suitable for acupuncture research and particularly non-significant results should be questioned.

The ‘n of 1’ study is another form of research: a patient will receive various different treatments in order to find out which form of treatment led to the best improvement for the relevant disorder. However, the same limitations as discussed above will affect this type of study.

Significance: In simple terms, the significance of a study indicates if the difference of a particular parameter which has been observed between two groups is caused by the treatment (and is therefore a true difference) or whether the difference is the result of random deviation. For example, if in a large number of treated patients the observed data differ ‘significantly’ from a group of untreated patients, it is **highly probable** that this difference is not due to a random deviation. This probability can be expressed mathematically. If, for example, the significance is expressed as $p = 0.03$ this means that, considering a random error of 3%, the difference noted between the two groups is statistically significant. The convention is to accept a random error of up to 5% as acceptable. The significance of a study is based on the number of patients treated and the extent of the difference in the treatment result (for example by assuming the same random error, the difference in treatment results has to be much larger in a group of 10 patients ($n = 10$) than if the same study was carried out with 100 patients ($n = 100$)). Only differences classified as ‘significant’ are accepted for research studies.

Power of the study: This refers to the ability of a study to be statistically significant. The higher the number of participants, the higher the power of the study. By international standards, 80% is an acceptable level of power. Depending on the difference between the two study groups, the number of patients required can therefore easily be in the 100s or 1000s.

Experimental studies: Through standardising and controlling conditions of the experiment strictly, every attempt is made to avoid random results, often leading to just a small number of participants. Generally these tend to be healthy volunteers, who are subjected to a clearly defined stimulus. In this case acupuncture as a form of therapy will achieve a high level of validity. More questionable are studies of healthy subjects without a well-defined stimulus, since in that case acupuncture as a balancing form of treatment is not used according to its principles (for example changes in peristalsis caused by acupuncture in healthy subjects).

Blinding: The gold standard of clinical trials are the so-called double blind studies in which both the patient and the therapist remain ignorant as to which treatment (for example the tablet with or without the active ingredient) is administered to the patient. However, blinding of an acupuncturist is not possible. Even in a single blind study (only the patient is ‘blinded’), his/her behaviour will distort the result of the trial. This effect can be counterbalanced to some extent by blinding the assessor of the treatment results (for example changes in the range of motion) so that he/she does not know which treatment (authentic or sham acupuncture) was received by the patient.

Standardisation: In pharmaceutical trials the amount of the active ingredient per tablet can easily be stated (for example standardisation of a 10 mg tablet). However, acupuncture cannot be standardised. The acupuncturist’s point selection will largely be influenced by experience, prior knowledge, specialisation and cultural background. Strictly speaking, acupuncture trials therefore do not investigate acupuncture per se but acupuncture as performed by an acupuncturist or a group of acupuncturists.

The fact that acupuncture is to a large extent centred on the therapist has in the past led to a paradoxical situation: experienced Chinese acupuncturists performing the ‘best’ acupuncture produced biomathematically ‘bad’ trials due to their lack of training, while ‘bad’ Western acupuncturists produced ‘good’ trials from a biomathematical point of view. ‘Importing’ Chinese acupuncturists into Western trials did not solve the problem either, since the Chinese clinicians were now treating Europeans with diseases not common in China and which reacted differently to acupuncture. Only as experience grows on both sides – which would have to be evaluated individually for each trial – will the value of acupuncture as a researched form of therapy increase accordingly.

9.2 Scientific Research According to Channels

Lung Channel

LU-5 → Ren-23; L.I.-10

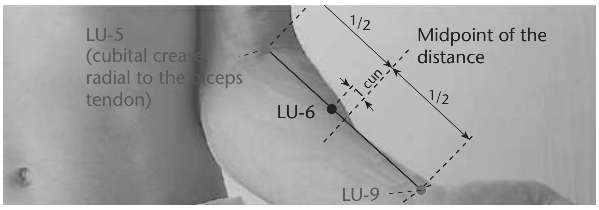

LU-6 192 patients with bronchial asthma were treated with **LU-6** and **LU-10**. 98.9% of the patients experienced an immediate improvement; the rate of clinical remission plus marked improvement was 76.5%. Best therapeutic results were achieved for allergic asthma according to Western medicine (asthma due to Cold according to TCM) and with longer needle retention (approximately 40 minutes). The lack of blinding and control groups reduce the value of this precise therapeutic case study. **Zang, J Tradit Chin Med 1990**

LU-7 → ST-8

LU-10 → LU-6

A Chinese prospective, randomised, controlled, non-blinded trial investigated the effect of bilateral bloodletting at **LU-10**, **P-8**, **HE-8**, **L.I.-2**, **T.B.-2** and **S.I.-3** on heart rate and consciousness in patients with **apoplexy** that had occurred within 3 days. Only patients with mild injuries showed improved consciousness while the heart rated increased independently of the size of injury. **Yi et al**, **J Tradit Chin Med 2005**

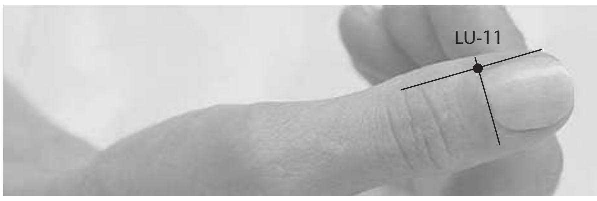

LU-11 Three Russian animal trials demonstrated the antipyretic action of **LU-11**. LU-11 can also strengthen the effect of hypothermic substances, suggesting a combination with antipyretic Chinese herbs. **Nezhenzev et al**, **Biull Eksp Bio Med 1992, and Fizion Zh SSSR Im I M Sechenova 1991**

A Taiwanese prospective, randomised, non-blinded, controlled trial ($n = 76$) investigated whether bloodletting at **LU-11** and **L.I.-1** can prevent or treat the occurrence of **laryngospasm** after tracheal extubation in children. The incidence of laryngospasm was significantly less in the bloodletting group (5% vs 24%). If laryngospasm developed, this could be relieved with the treatment of LU-11 or L.I.-1 within one minute in all patients. **Lee et al, Anaesthesia 1998**

Large Intestine Channel

L.I.-1 → LU-11

L.I.-2 → LU-10

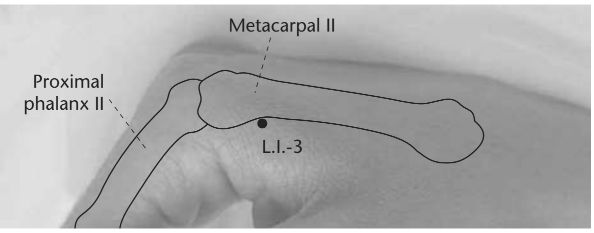

L.I.-3 A Chinese prospective, randomised, controlled study ($n = 210$) reports the significantly better effects of needling **L.I.-3** and ***lingxia*** in the treatment of **periarthritis of the humerus** compared to the control group treated with local points. Of interest is the application of *lingxia*, an unofficial extra point 2 cun inferior to G.B.-34. **Feng, J Tradit Chin Med 2003**

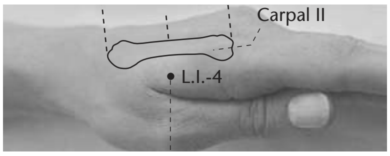

L.I.-4 → ST-36; P-5, P-6; ST-8; G.B.-43; Du-20; LIV-3; S.I.-18; SP-9; L.I.-11; ST-2

A Turkish prospective, controlled, non-blinded, group-comparison trial ($n = 55$) investigated the effect of electro-acupuncture (EA) at **L.I.-4**, **L.I.-11**, **ST-25**, **ST-36**, **ST-44** and **LIV-3** (combined with auricular points *shenmen* and *sanjiao*) on **obesity in women**. Treatments were given daily for 20 days. One control group received a 1425 kcal diet for 20 days while the other control group received no treatment. There was a 4.8% weight reduction in the EA group, whereas women on the restricted diet had a 2.5% weight loss. Both groups showed a significant reduction in total cholesterol and triglyceride levels. In the EA group only, LDL levels were also reduced significantly. **Cabioglu et Ergene**, **Am J Chin Med 2005**

A Chinese (Hong Kong) prospective, randomised, controlled, non-blinded study ($n = 29$) investigated the effect of **L.I.-4** and **P-6** on the heart rate variability (HRV) in normal subjects in **fatigue** and **non-fatigue states** respectively. Only fatigued subjects showed a significant change in HRV. The authors concluded that the effect of acupuncture depends not only on the points selected but also on the functional state of the subject, one of the principles of Chinese Medicine. **Li et al, Eur J Appl Physiol 2005**

An unusual veterinary case study investigated the effects of acupuncture on the **immune system**. According to the authors, immunostimulant points **include LI-4**, **LI-11**, **ST-36**, **GB-39**, **SP-6**, **GV-14**, **BL-11**, **BL-20**, **BL-23**, **BL-24**, **BL-25**, **BL-26**, **BL-27**, **BL-28,** and **Ren-12**. **BL-47** is considered to be immunosuppressive. Anti-febrile points include **ST-36** and **Du-14**. **Rogers et al, Probl Vet Med 1992**

A Chinese prospective, randomised, single-blinded, experimental trial ($n = 20$) compared the effect of transcutaneous electric nerve stimulation (TENS) at **L.I.-4** on the **heat pain threshold** and **vibration threshold** in humans with a placebo control point. In the study group, after TENS at **L.I.-4**, the pain threshold for heat was higher for up to half an hour compared to the control group. There was no effect on the threshold for feeling vibration. **Wang et al, Chin Med J (Engl) 2003**

A Korean prospective, randomised, placebo-controlled trial ($n = 35$) evaluated the efficacy of electro-acupuncture (EA) and moxibustion on **L.I.-4**, **L.I.-10**, **L.I.-11** and **T.B.-5** on **spasticity due to stroke**. The control group was treated with routine acupuncture. Only the EA group experienced a significant, immediate and lasting reduction in spasticity (using a modified Ashworth scale). **Moon et al, Am J Chin Med 2003**

A German prospective, randomised, controlled, non-blinded trial ($n = 29$) investigated if acupuncture at **L.I.-4**, **T.B.-5**, **ST-36**, **P-6**, **Du-20** and **Ex-UE-9** (*baxie*) had a vasospasmolytic effect on patients with **Raynaud's syndrome** during the winter season. There was no significant difference regarding the number and intensity of episodes, nor was there a significant effect on skin microcirculation between the acupuncture group and the placebo group before and after treatment. **Hahn et al, J Dtsch Dermatol Ges 2004**

A Japanese prospective, randomised, controlled study ($n = 21$) investigated the effect of **L.I.-4**, **ST-36 and SP-6** on objective and subjective parameters regarding **exhaustion** in elite female soccer players during a competition period. Only the subjects treated with acupuncture experienced a better rating for both subjective and objective parameters. **Akimoto et al, Med Sci Sports Exerc 2003**

A Chinese prospective, randomised, controlled trial ($n = 89$) researched the effect of **L.I.-4**, **ST-36**, **BL-18** and **BL-23** combined with CT positioning scalp circum-needling (SCN) and herbal therapy in patients with **poly-infarctional vascular dementia**. The control group received herbal treatment only. The treated group had a significantly better short-term result (96.5%) than the control group (75%) regarding changes of clinical symptoms, intelligence and haemorrheological characteristics. **Lun et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2003**

A Chinese prospective, randomised, controlled experimental trial ($n = 11$) investigated the changes in **brain activation patterns** evoked by stimulating **L.I.-4** manually versus stimulation with electro-acupuncture (EA). These were assessed by functional magnetic resonance imaging (fMRI). Results showed that EA mainly produced fMRI signal increases in precentral gyrus, postcentral gyrus/inferior parietal lobule and putamen/insula; in contrast, manual needle manipulation produced prominent decreases of fMRI signals in posterior cingulate, superior temporal gyrus, putamen/insula. These results suggest that different brain mechanisms may be recruited during manual and EA. **Kong et al, J Altern Complement Med 2002**

An American prospective, randomised, controlled, non-blinded experimental trial ($n = 13$) evaluated the effect of transcutaneous electric acupoint stimulation (TEAS) on **L.I.-4** and **P-6** on **tonic pain** induced with the cold-pressor test. While TEAS alone had a significant analgesic effect, this could be increased by combining TEAS analgesia with low-dose morphine. **Yuan et al, J Clin Pharmacol 2002**

A Taiwanese prospective, randomised, placebo-controlled experimental study ($n = 13$) investigated the effect of manual acupuncture as well as transcutaneous electrical nerve stimulation (TENS) with 2 Hz and 100 Hz respectively at **L.I.-4** on the spinal cord by measuring the **H-reflex**. While the application of TENS with both 2 Hz and 100 Hz changed the amplitude of the H-reflex, it was retained longer with the 100 Hz stimulation. These findings suggest that TENS enhances the excitability of the motoneuron pool in the spinal cord. **Chang et al, Acupunct Electrother Res 2001**

An Austrian prospective, randomised, controlled study ($n = 45$) evaluated whether acupuncture at **L.I.-4** and **SP-6** at term can influence cervical ripening, **induce labour** and thus reduce the need for postdate induction. Compared to the control group, the study group showed a significantly shorter time span (three days) between estimated date of confinement (EDC) and actual time of delivery. Medical induction of labour 10 days after EDC was also significantly lower in the acupuncture group (20%) than in the control group (35%). **Rabl et al, Wien Klin Wochenschr 2001**

An Israeli prospective, randomised, double-blinded controlled trial ($n = 25$) investigated the effect of acupuncture in the treatment of **irritable bowel syndrome**. True acupuncture was performed at **LI-4** and sham acupuncture at **BL-60** during a total of two sessions. There was no significant difference between the two groups. However, the choice of acupuncture points seems somewhat peculiar. While both points are Ma Dan Yang points (the 12 most important points), they have no specific action for intestinal disorders. Obviously the authors chose the points according to the channel (Large Intestine). Furthermore, only two treatments does not seem sufficient. **Fireman et al, Digestion 2000**

A British prospective experimental trial ($n = 14$) investigated the effect of acupuncture at **L.I.-4** on the **resting EEG** of healthy subjects. Acupuncture did not induce any changes in the resting EEG. **Rosted et al, Complement Ther Med 2001**

An Irish prospective, randomised, controlled trial ($n = 30$) applied transcutaneous electric nerve stimulation (TENS) either at **L.I.-4** or over the median nerve, in order to investigate **cutaneous blood flow**. The control group received no treatment. Only TENS over the median nerve showed a significant increase in cutaneous blood flow. **Cramp et al, Acupunct Electrother Res 2001**

A Japanese prospective, randomised, controlled, non-blinded trial ($n = 22$) investigated the effect of acupuncture at **L.I.-4**, **ST-6** and **ST-7** on pain after mandibular **wisdom tooth extraction**. Compared to the control group there was a significant reduction in pain in the treatment group. **Kitade et al, Acupunct Electrother Res 2000**

A Swedish prospective, randomised, controlled experimental trial ($n = 12$) investigated if acupuncture at **L.I.-4** induced **changes in the sympathetic and/or parasympathetic nervous system** in healthy subjects. For this purpose, regular and superficial needling of **L.I.-4** was compared. Changes were measured by using power spectral analysis and heart rate frequency analysis. Only deeper needling at L.I.-4 led to a significant activation of the sympathetic and parasympathetic nervous system during and after treatment as well as to a significant reduction of the heart rate frequency. **Haker et al, J Auton Nerv Syst 2000**

A US prospective, randomised, sham-controlled, single-blinded trial (*n* = 101) assessed the effect of varying intensities of transcutaneous electric acupoint stimulation (TEAS) at **L.I.-4** on the **postoperative patient-controlled analgesia** (PCA) requirement for **hydromorphine** (HM), the incidence of opioid-related side effects, and the recovery profile after lower abdominal surgery in women. In the sham-TEAS group HM-requirement was reduced by 23%, low-intensity TEAS produced a 34% decrease in the HM requirement while the high-intensity TEAS reduced requirement significantly by 65%, accompanied by a reduction in the occurrence of dizziness, nausea and pruritus. **Wang et al, Anesth Analg 1997**

A Chinese prospective, randomised, controlled trial (*n* = 110) investigated the effect of transcutaneous electric acupoint stimulation TEAS) at **L.I.-4**, **Ex-HN-4** ***(yuyao)*** and **G.B.-31** during enflurane anaesthesia in patients undergoing craniotomy. In comparison to the control group (anaesthesia maintained with enflurane only) the **minimum alveolar concentration** (MAC) of enflurane decreased in the TEAS-supplemented group by 38–47%. If anaesthesia was supplemented by TEAS plus scalp infiltration with procaine, there was an even higher reduction of 42–66%. The haemodynamic balance was also more stable during the operation, and postoperative recovery was faster. **Wang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1994**

A German prospective, waitlist-controlled trial (*n* = 6) investigated the effect of electro-acupuncture at **L.I.-4**, **ST-25**, **LIV-3** and **BL-25** on stool frequency and colonic transit time in **chronic constipation**. There were no significant differences in all the parameters between the control group and the waitlist group. **Klauser et al, Z Gastroenterol 1993**

A Japanese experimental trial researched the influence of **L.I.-4 on vibration induced finger flexion reflex**. Unilateral acupuncture at L.I.-4 suppressed the reflex in both hands. **Takakura et al, Am J Chin Med 1992**

A Chinese prospective, randomised, matched controlled trial (*n* = 12) investigated the analgesic effect of aqueous acupuncture (injection of glucose solution) at **L.I.-4** and **G.B.-34** in **postoperative pain control**. In comparison to the control group, the intensity of postoperative pain as well as the requirement for analgesic medication were significantly lower in the study group. **Chen et al, Gaoxiong Yi Xue Ke Xue Za Zhi 1991**

A German prospective, non-randomised, sham-controlled trial (*n* = 36) investigated the **analgesic effect** of acupuncture at **L.I.-4**, **P-6**, **ST-36** and **SP-4** prior to a **colonoscopy**. The pain experienced by the acupuncture patients was significantly lower compared to the group receiving sham acupuncture. The analgesics and sedatives required were also significantly less. **Li et al, Dtsch Med Wochenschr 1991**

A prospective, randomised, single-blinded experimental study (*n* = 39) investigated the effect of acupuncture and laser acupuncture at **L.I.-4** and **Ex-UE** ***jianqian*** on the **pain threshold**. Compared to laser acupuncture, acupuncture significantly increased the pain threshold. **Brockhaus et al, Pain 1990**

A British prospective, controlled experimental cross-over study investigated the effectiveness of electro-acupuncture at **L.I.-4** on experimental **tooth pain**. After 30 minutes of electric stimulation the pain threshold increased by 27%. This increase could be partially blocked by naloxone. **Ernst et al, Acupunct Electrother Res 1987**

A British prospective, randomised, placebo-controlled study (*n* = 51) investigated the effect of acupuncture at **L.I.-4** as an **analgesic for operative dentistry**. There were no significant differences between the treatment group and the placebo group. **Taub et al, Oral Surg Oral Med Oral Pathol 1979**

A prospective non-randomised, placebo-controlled experimental trial (*n* = 40) investigated the effect of acupuncture at **L.I.-4** on the **pain threshold of postoperative dental pain** compared to codeine. Both **L.I.-4** and codeine had a significantly better effect than the placebo. The combination of both had the strongest effect. **Sung et al, Anesth Analg 1977**

An Austrian prospective, sham-controlled, single-blinded experimental trial (*n* = 12) measured the effect of **L.I.-4** and **P-6** on **pain threshold and pain tolerance**. In contrast to sham acupuncture, true acupuncture significantly raised the pain threshold. **Stacher et al, Am J Chin Med 1975**

L.I.-5 → P-5
L.I.-6 → P-5
L.I.-7 → P-5

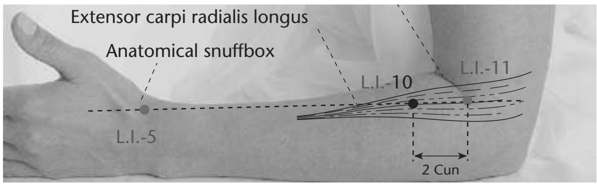

L.I.-10 → L.I.-4; Du-12; ST-36

A Swedish prospective, randomised, controlled, double-blinded trial (*n* = 49) investigated the effect of laser acupuncture on **L.I.-10**, **L.I.-11**, **L.I.-12**, **LU-5** and **T.B.-5** on **lateral humeral epicondalgia (tennis elbow)**. No significant difference occurred at a dose of 0.36 J/point. **Haker et al, Pain 1990**

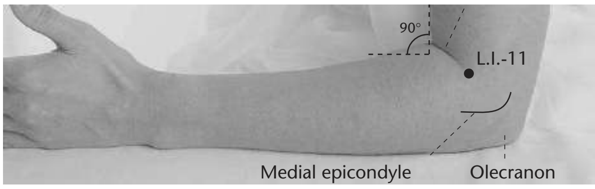

L.I.-11 → L.I.-4; ST-36; LIV-3; L.I.-10; L.I.-15

A Taiwanese prospective, randomised, placebo-controlled, single-blinded trial (*n* = 40) investigated the effect of acupuncture at **L.I.-11** on **uraemic pruritus**. The results of the pruritus score questionnaires demonstrated that only true acupuncture at **L.I.-11** resulted in a significantly reduced pruritus lasting for a period of three months. Uraemia levels remained unchanged. **Che-yi et al, Nephro Dial Transplant 2005**

A German prospective, randomised, placebo-controlled, single-blinded experimental trial (*n* = 22) investigated the effect of **L.I.-11** on the **phagocytic immune system**. Based on laboratory parameters, the **activity of neutrophils** was significantly higher in the study group. **Karst et al, Complement Ther Med 2003**

A Swedish prospective, non-randomised, placebo-controlled, single-blinded experimental longitudinal trial investigated the influence of electro-acupuncture at **L.I.-11** and **L.I.-4** on sympathetic nerve activity and the pain threshold. Only electro-acupuncture at L.I.-11 and L.I.-4 increased the pain threshold in combination with increased sympathetic nerve activity. **Knardahl et al, Pain 1998**

A Swedish prospective, randomised, placebo-controlled, single-blinded trial (*n* = 58) investigated the effect of laser acupuncture at **L.I.-11** and **L.I.-12** for **lateral humeral epicondalgia (tennis elbow)** compared to laser application on the painful area. There were no significant differences between the laser and placebo treatments after the treatment period. **Haker et al, Arch Phys Med Rehabil 1991**

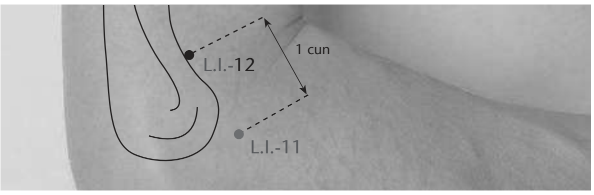

L.I.-12 → L.I.-11; L.I.-10

A Chinese prospective, randomised, controlled trial (*n* = 64) investigated the effect of electro-acupuncture at **L.I.-12**, **T.B.-5**, **G.B.-30** and **ST-36** accompanied by medication for **acute cerebral infarction** with regard to the plasma and CSF somatostatin levels. The control group received medication only. Both plasma and CSF somatostatin levels increased significantly in the acupuncture group. **Zhang et al, J Tradit Chin Med 1999**

A Chinese prospective, randomised, controlled trial (*n* = 64) investigated the effect of electro-acupuncture at **L.I.-12**, **T.B.-5**, **G.B.-30** and **ST-36** on the plasma and CSF levels of VIP (vasoactive intestinal peptide), somatostatin and pancreatic polypeptide in patients with **acute cerebral infarction**. The CSF VIP level dropped significantly while the plasma pancreatic polypeptide level increased significantly. In patients with a good response both plasma and CSF somatostatin levels increased significantly after acupuncture. **Zhang et al, Zhen Ci Yan Jiu 1996**

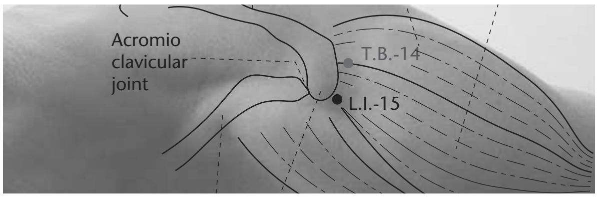

L.I.-15 A Taiwanese prospective, randomised, controlled, non-blinded trial (*n* = 150) investigated the effect of electro-acupuncture at **L.I.-15**, **G.B.-21** and **Ex-UE *jianneiling*** with or without regional nerve block (RNB) (stellate ganglion and suprascapular plexus) for **'frozen shoulder' (adhesive capsulitis)**. This was compared to RNB treatment alone. The combination of electro-acupuncture and RNB produced the most significant reduction of pain and duration of pain relief as well as improvement of range of motion. **Lin et al, Acta Anaesthesiol Sin 1994**

A Chinese prospective, randomised, controlled, non-blinded trial (*n* = 63) investigated the effect of acupuncture at **L.I.-15**, **L.I.-11**, **L.I.-4**, **G.B.-30**, **G.B.-34**, **G.B.-37**, **LIV-3**, **G.B.-20**, **ST-25**, **ST-40**, **ST-36**, **SP-6** and **KID-3** on symptoms of **cerebral infarction** compared to the treatment with the calcium channel blocker verapamil. The effectiveness in the acupuncture group was 94%, in the verapamil group 84%. This difference is statistically significant. However, the choice of a calcium channel blocker seems unusual for this indication. **Zou et al, Zhong Xi Yi Jie He Za Zhi 1990**

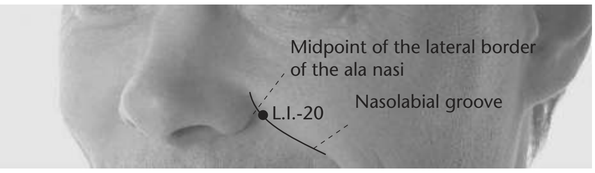

L.I.-20 → G.B.-14

A British prospective, randomised, controlled trial (*n* = 20) investigated the effect of a 30 second massage at **L.I.-20** on **nasal airflow** compared to an untreated control group. Symptoms improved significantly in the massage group. **Takeuchi et al, Am J Rhinol 1999**

Stomach channel

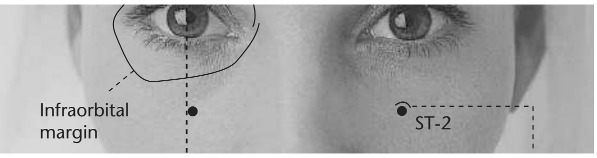

ST-2 → G.B.-14

A prospective, non-randomised experimental trial (*n* = 33) investigated the effect of acupuncture and electro-acupuncture at **ST-2**, **L.I.-4** and **ST-44** on experimentally induced **tooth pain**.

While both procedures increased the pain threshold, the increase was not sufficient to justify acupuncture as a means of pain control in conservative dentistry. **Bakke, Scand J Dent Res 1976**

ST-4 → G.B.-14
ST-6 → G.B.-14; L.I.-4
ST-7 → L.I.-4

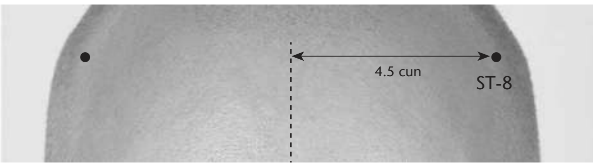

ST-8 A large prospective, randomised, multi-centre, group-comparison trial ($n = 120$) in patients suffering from **migraine without aura** demonstrated that acupuncture at **ST-8**, **G.B.-5**, **G.B.-20**, **Du-14** and **LU-7** was significantly more effective than conventional drug therapy, regarding both frequency and duration of the migraine attacks as well as number of sick days. The authors calculated that by treating migraine with acupuncture the Italian health system could save approximately 0.5 billion Euros per year. **Liguori et al., J Tradit Chin Med 2000**

A Bulgarian case study ($n = 35$) reports the successful treatment of headaches following caesarian sections with spinal anaesthesia with acupuncture at ST-8, L.I.-4, G.B.-11, G.B.-14, BL-10, Du-14 and Du-20 during 1–3 treatments. **Tsenov, Akush Ginekol (Sofia) 1996**

A Swedish prospective, placebo-controlled trial investigated the effect of intrasegmental (**ST-8**) and extrasegmental (**L.I.-4**) electro-acupuncture on sensory thresholds. Only the pain threshold could be increased, and this only with intrasegmental acupuncture at ST-8. **Lundeberg et al, Am J Chin Med 1989**

ST-18 → G.B.-14; Du-1
ST-20 → T.B.-8
ST-21 → Du-20

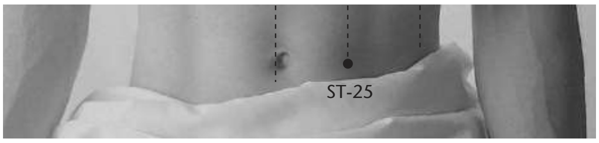

ST-25 → Du-1; Du-20; L.I.-4; L.I.-15

A Chinese prospective, randomised, controlled trial ($n = 62$) investigated the effect of acupuncture and moxibustion at ST-25 and Ren-4 for chronic colitis. This was compared to modern drug therapy. The acupuncture group showed the same therapeutic results but with fewer side effects. **Yang et al, J Tradit Chin Med 1999**

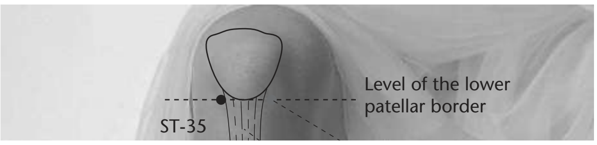

ST-32 → G.B.-31
ST-34 → ST-36; SP-9

ST-35 A Chinese prospective, randomised, single-blinded, controlled, long-term trial ($n = 24$) investigated the effect of electro-acupuncture (EA) and transcutaneous electrical nerve stimulation (TENS) at **ST-35** and **Ex-LE-4** (eight treatments over a period of two weeks) in older patients (average age: 85 years) suffering from **painful osteoarthritic knees**. This was compared to conventional therapy for this disorder. Both TENS and EA significantly reduced the knee pain. EA also had an effect on range of motion. **Ng et al, Altern Complement Med 2003**

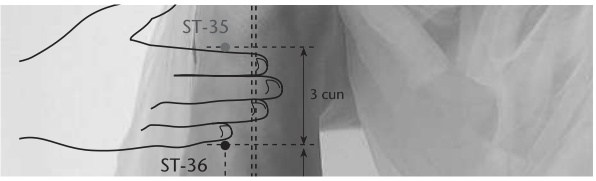

ST-36 → L.I.-4; BL-27; P-6; T.B.-8; G.B.-26; G.B.-31; Du-1, Ex-B-3; Du-4; Du-20; LIV-3; SP-9; BL-60; L.I.-12; L.I.-15; Ren-17

An Iranian prospective, randomised, placebo-controlled, single-blinded trial ($n = 51$) investigated the effect of transcutaneous electric nerve stimulation (TENS) at **ST-36** and **ST-38** on nitroglycerin-induced **hypotension under general anaesthesia with halothane**. With true acupuncture at the above points, the mean arterial blood pressure (MAP) was significantly lower (60 vs 66 mmHg). The effect of acupuncture was significantly greater with lower dosages of nitroglycerin compared with higher dosage. In addition, the time to reach steady state MAP was significantly shorter (10 vs 15.8 minutes) and the quality of operative ischaemia was excellent in 83.3% subjects compared to 25% in the sham group. **Saghaei M. et al, Acta Anaesthesiol Taiwan 2005**

A US prospective non-blinded experimental trial investigated the response to acupuncture at **ST-36** as evidenced by **fMRI**. The limbic and paralimbic structures of the cortical and subcortical regions in the telencephalon, diencephalon, brainstem and cerebellum were activated in response to acupuncture. **Hui et al, Neuroimage 2005**

A prospective, randomised placebo-controlled single-blinded trial ($n = 13$) investigated the effect of electro-acupuncture at **ST-36** on the **heart rate variability** (HRV) in healthy subjects. There was no acupuncture-specific effect on HRV. **Chang et al, Am J Chin Med 2005**

A German prospective randomised placebo-controlled single-blinded experimental crossover study ($n = 10$) investigated the effect of acupuncture at **ST-36**, **L.I.-11**, **SP-10 and Du-14** on

leukocyte circulation in healthy young subjects. There was a significant decrease in leukocyte and lymphocyte values in the acupuncture group while cortisol and norepinephrine plasma levels remained unchanged. **Kou et al, Brain Behav Immun 2005**

A US prospective, non-randomised, placebo-controlled, single-blinded trial (*n* = 7) investigated if electro-acupuncture at **ST-36, SP-6**, **ST-34** and **L.I.-4** before and during induction of general anaesthesia (desflurane) influenced the **anaesthetic requirement** compared to a control group. There was no reduction in pain in the treatment group. **Chernyak et al Anesth Analg 2005**

A Taiwanese prospective, randomised, placebo-controlled, single-blinded experimental cross-over study (*n* = 15) investigated if electro-acupuncture (EA) at **ST-36** can normalise **atropine-induced gastric dysrhythmia**. While there was a significant increase in the percentage of normal frequency, atropine-induced gastric dysrhythmia was not normalised by EA (no significant differences in the EGG). **Chang et al, Dig Dis Sci 2002**

A Taiwanese prospective, randomised, controlled experimental trial (*n* = 15) investigated the effect of **ST-36** on **gastric myo-electrical regularity**. There was a significant decrease in the tachygastric and bradygastric rhythm during electro-acupuncture on ST-36. **Chang et al, Digestion 2002**

A Taiwanese prospective, randomised, sham-controlled trial (*n* = 100) investigated the effect of electro-acupuncture at **ST-36** on **postoperative pain and opioid-related side effects**. The postoperative pain, the requirement for morphine (via PCA) and the incidence of nausea and dizziness were all significantly reduced. **Lin et al, Pain 2002**

A US prospective, randomised, double-blinded, sham-controlled experimental trial (*n* = 14) investigated if electro-acupuncture at **ST-36**, **G.B.-34** and **BL-60** could reduce the anaesthetic requirement (desflurane) for **experimental pain stimuli**. Electro-stimulation of these points did not reduce desflurane requirements compared to the control group. **Morioka et al, Anesth Analg 2002**

A Taiwanese prospective case study (*n* = 15) investigated the effect of bilateral electro-acupuncture **at ST-36** on gastric slow waves in **diabetic patients** with symptoms suggesting **gastric motor dysfunction**. Changes of gastric activity were measured by EGG. During and after acupuncture there was a significant increase in the percentages of normal frequency. In addition the percentage of tachygastric frequency was decreased significantly during and after acupuncture. **Chang et al, Digestion 2001**

A Chinese case study (*n* = 104) reports the successful injection-acupuncture at **ST-36** for **chemotherapy-induced leukopenia**. **Yin et al, J Tradit Chin Med 2001**

A Chinese prospective, randomised, controlled, non-blinded trial (*n* = 26) investigated the effect of acupuncture at **ST-36** plus auricular plaster therapy on **peristalsis** for postoperative recovery of intestinal function after abdominal surgery. 92% of patients in the treatment group showed recovery of normal peristalsis within 72 hours, while in the control group only 46% patients recovered within the same time-period. **Wan et al, J Tradit Chin Med, 2000**

A US prospective, randomised, controlled pilot study (*n* = 17) investigated the effect of acupressure at **ST-36** and **P-6** on the intensity and frequency of **nausea** in patients undergoing **chemotherapy due to breast cancer**. Both the frequency and intensity of the nausea was significantly lower in the treatment group than in the control group. **Dibble et al, Oncol Nurs Forum 2000**

An Italian prospective, randomised, controlled experimental trial (*n* = 120) investigated the effect of acupuncture at **ST-36** and **L.I.-4** regarding changes in the levels of beta-endorphins and other parameters (VIP, lymphocyte subsets, NK cells and monocyte phagocytosis) in patients suffering from various painful disorders. These changes were compared to an untreated control group. Only in the acupuncture group were the endorphin levels, the CD3 and CD4 values and monocyte phagocytosis increased for at least 24 hours after treatment. At the same time there was an increase of the CD8 values. **Petti et al, J Tradit Chin Med 1998**

A Chinese retrospective case study reports good results in treating **leukopenia** with acupuncture at **ST-36**. **Wei, J Tradit Chin Med 1998**

A Taiwanese prospective, randomised, sham-controlled experimental trial (*n* = 18) investigated the effect of acupuncture at **ST-36** and **L.I.-4** on central nervous system pathways with functional MR imaging of the brain. Besides a significant reduction of the heart rate frequency, acupuncture at both points resulted in activation of the hypothalamus and nucleus accumbens (structures of the descending antinociceptive pathway) and deactivation of the rostral part of the anterior cingulate cortex, amygdala formation, and hippocampal complex (limbic areas of pain recognition). Therefore this study provides explanations for the **pain-reducing effect** of acupuncture. **Wu et al, Radiology 1999**

A prospective, sham-controlled, single-blinded experimental longitudinal trial (*n* = 13) investigated the effect of acupuncture at ST-36 only, compared to acupuncture at **ST-36 plus L.I.-10** on **auditory endogenous potentials** (P300). The effect of the different forms of acupuncture was identical. In contrast to the sham group, it led to a decrease of P300 amplitude in the acupuncture group. **Hsieh et al, Am J Chin Med 1998**

A US prospective, randomised, sham-controlled, single-blinded trial (*n* = 100) investigated the effect of transcutaneous electric nerve stimulation (TENS) at **ST-36** on **opioid analgesic requirement** by PCA following hysterectomy or myomectomy compared to peri-incisional dermatomal stimulation by TENS. TENS applied at the dermatomal level of the incision is as effective as stimulation at ST-36, and both were more effective than stimulation at a sham location. **Chen et al, Anesth Analg 1998**

A Chinese prospective, randomised, controlled trial (*n* = 69) compared the so-called immunotherapy at **ST-36** to conventional desensitisation therapy in patients suffering from **anaphylactic asthma**. Results in the acupuncture group were significantly better, regarding both clinical and immunohistochemical parameters. **Chen et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1996**

A Mongolian prospective, randomised, controlled study ($n = 48$) compared the effect of acupuncture at **ST-36** and **P-6** combined with epidural anaesthesia to simple epidural anaesthesia for subtotal gastrectomies. The acupuncture group, while requiring a lower dosage of anaesthetic, experienced a stronger analgesic effect, the abdominal muscles were more relaxed and the haemodynamic disturbances were smaller. **Sun, Zhen Ci Yan Jiu 1996**

A US prospective, randomised, controlled, experimental trial ($n = 11$) researched the effect of transcutaneous electric nerve stimulation (TENS) at **ST-36** and **ST-37** on **peripheral blood circulation** and **haemodynamics**. Twenty minutes of TENS on healthy subjects did not produce a significant change in the measured parameters. **Balogun et al, Disabil Rehabil 1996**

A Taiwanese prospective, randomised, placebo-controlled, single-blinded trial investigated the influence of acupuncture at **ST-36** on the **pulse spectrum**. The results indicated that acupuncture at ST-36 has a specific effect on the Fourier components of the pulse: C2 and C4 decreased while C5, C6, C8 and C9 increased. This specific frequency effect was not found when acupuncture was applied on a non-acupuncture point. **Wang et al, Am J Chin Med 1995**

A Chinese prospective, randomised, controlled trial ($n = 45$) investigated the effect of **ST-36**, **L.I.-11** and **KID-6** on the immunoactivity of natural killer cells and the interleukin-2 level in **malignant tumour** patients. After receiving one treatment of 30 minutes daily for 10 days, parameters in the treatment group had increased significantly compared to the control group. **Wu et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1994**

A Chinese prospective, randomised, controlled trial ($n = 80$) compared **anaesthesia with electro-acupuncture** at **ST-36** and **L.I.-4** to epidural anaesthesia in **appendectomies**. The operations were equally successful in both groups, without any significant differences regarding the success rate. **Sun et al, Zhen Ci Yan Jiu 1992**

A Chinese prospective, randomised, controlled trial ($n = 39$) investigated the effect of acupuncture at **ST-36** and **SP-6** on **bowel movements following abdominal surgery**. In the acupuncture group the first postoperative bowel movement occurred after approximately 58 hours, but only after 86 hours in the untreated control group. **Liu et al, Zhong Xi Yi Jie He Za Zhi 1991**

ST-37 → ST-36
ST-38 → G.B.-34; ST-36
ST-40 → L.I.-15
ST-44 → L.I.-4; ST-2

Spleen channel

SP-1 A US prospective experimental trial ($n = 13$) investigated the effect of electro-acupuncture (EA) at **SP-1** and **LIV-1** on **thermal pain thresholds**. The pain threshold increased significantly 30 seconds after onset of EA. The authors postulate that EA at the above points has an inhibitory effect on the C-fibre afferents; the analgesic benefit observed is most probably A-delta afferent mediated. **Leung et al, J Altern Complement Med 2005**

SP-4 → Du-4; P-6; L.I.-4

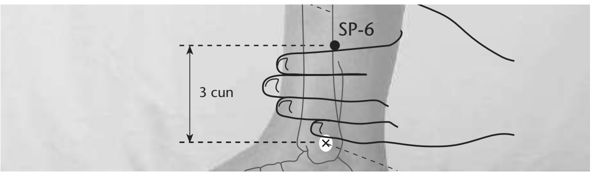

SP-6 → ST-36; L.I.-4; P-6; G.B.-26; Du-4; Du-20; LIV-3; L.I.-15

In a prospective, randomised, double-blinded group-comparison trial ($n = 56$) patients suffering from **interstitial cystitis** performed daily laser therapy on SP-6 at home for 30 seconds over a period of 12 weeks. There were no significant differences between the treatment and control cohorts. **O'Reilly et al, J Urol 2004**

An Australian prospective, randomised, placebo-controlled trial ($n = 20$) investigated the effect of transcutaneous electrical stimulation (TENS) at **SP-6** and **LIV-3** on **uterine contractions** in post-date pregnant women. A significant increase in frequency and strength of contractions was found in the TENS group compared with the placebo group. **Dunn et al, Obstet Gynecol 1989**

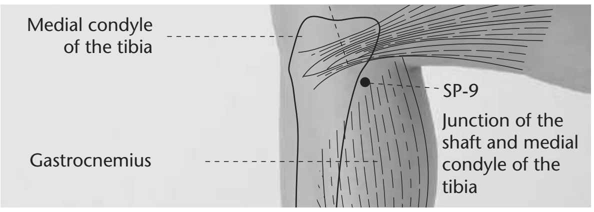
Medial condyle of the tibia
SP-9
Junction of the shaft and medial condyle of the tibia
Gastrocnemius

SP-9 A British prospective, randomised, controlled, assessor-blinded trial ($n = 44$) investigated the influence of **SP-9**, **SP-10**, **ST-34**, **ST-36** and **L.I.-4** on unilateral versus bilateral acupuncture in patients with **advanced osteoarthritis of the knee**. The symptoms improved significantly in both groups. This improvement was sustained for six months. There was no statistically significant difference between the groups. **Tillu et al, Acupunct Med 2001**

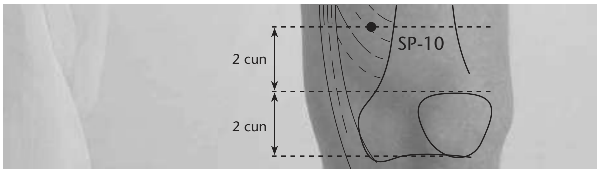

SP-10 → SP-9; ST-36; Ren-17

A Chinese prospective, randomised, controlled trial ($n = 62$) investigated the effect of Q-wave millimetre microwave application at **SP-10** and **BL-17** on **chemotherapy-induced leukopenia in gastrointestinal cancer patients**. One group received

irradiation before the start of chemotherapy, the other group only after the onset of the chemotherapy-induced leukopenia. Beginning the irradiation before starting chemotherapy produced a significantly better result (86% versus 73% with later application). **Wu et al, Zhongguo Zhong Xi Yi Jie He za Zhi 1997**
SP-12 → Du-4

Heart channel

HE-3 → P-6
HE-5 → P-6

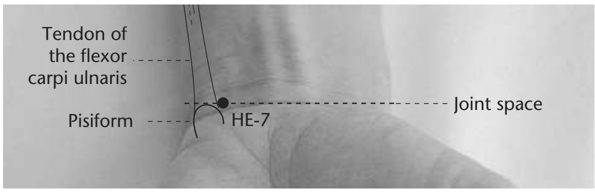

HE-7 → G.B.-26; Du-4; Du-20; S.I.-3; P-6
A German prospective, randomised, placebo-controlled, single-blinded trial ($n = 36$) investigated the effect of **H-7**, **P-6**, **Du-20**, **BL-62** and **Ex-HN-6** ***erjian*** on the cardiac autonomous nervous system in patients with **minor depression and anxiety disorders**. Only the acupuncture group demonstrated a relative increase of cardiovagal modulation of heart rate. **Agelink et al, Fortschr Neurol Psychiatr 2003**
A British pilot study ($n = 17$) investigated the effect of **HE-7** on certain **stress parameters**. There was an average decrease of 44% (based on the Edinburgh Postnatal Depression Scale (EPDS). **Chan et al, Acupunct Med 2002**
HE-8 → LU-10
HE-9 → Du-4

Small Intestine channel

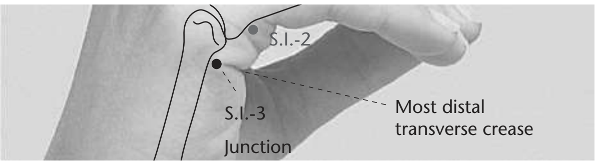

S.I.-3 → LU-10; Du-12
A larger-scale Taiwanese prospective, controlled, randomised trial ($n = 100$) demonstrated a significant improvement of neurological functioning in patients with **severe spinal cord injuries** (ASIA status A and B) after electro-acupuncture at **S.I.-3** and **BL-62** plus auricular acupuncture versus the control group (conventional and rehabilitation therapy only). **Wong et al, Am J Phys Med Rehabil 2003**
A non-controlled pilot study ($n = 18$) found that the treatment of **acute torticollis** with a single treatment of ipsilateral acupuncture at **S.I.-3** and **M-UE-8** resulted in a mean improvement of lateral head rotation of 53%. **Samuels N, Am J Chin Med 2003**

A Chinese prospective, randomised, controlled trial ($n = 62$) investigated the effect of acupuncture at **S.I.-3** and **HE-7** in patients with **cerebral traumatic dementia** compared to conventional physiotherapy. There were significant changes of MMSE scores and auditory P300 values in the acupuncture group only. **Zhang et al, Zhen Ci Yan Jiu 1996** and **Zhang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1995**

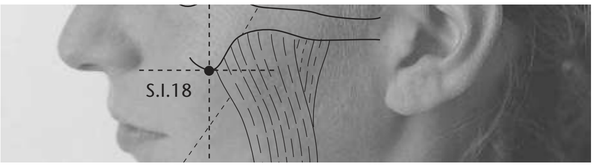

S.I.-18 A Chinese prospective, randomised, controlled, single-blinded trial ($n = 42$) investigated the effect of electro-acupuncture at **S.I.-18**, **Ex-HN-04**, **G.B.-20** and **L.I.-4** on controlled hypotension induced by isoflurane during a cerebral operation in order to avoid blood loss. The concentration of isoflurane necessary to achieve controlled hypotension was significantly reduced by 31–42% compared to the control group. **Wang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2000**

Bladder channel

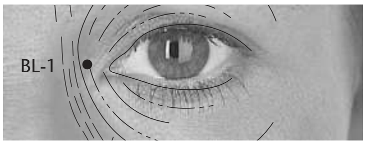

BL-1 → G.B.-14
A non-randomised group-comparison trial ($n = 34$) investigated the effect of warmed needle acupuncture at **BL-1** in patients suffering from **epiphora** due to dysfunction of the lacrimal duct. The effect in the treatment group was significantly better in comparison to the control group treated with lacrimal duct irrigation and norfloxacin eye drops (92% vs 54%). **Ni et al, J Tradit Chin Med 2002**
A Chinese case study reports the treatment of **epiphora** due to **insufficiency of lacrimal passage** with acupuncture at **BL-1**. Of the 68 eyes treated in 42 patients, 28 eyes were cured and 35 improved. 34 eyes improved after only one course of treatment. **Ni Y et al, J Tradit Chin Med 1999**

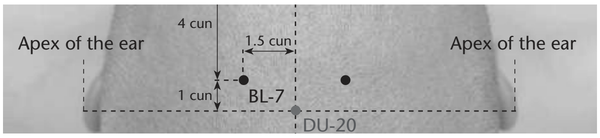

BL-7 A Chinese experimental trial carried out in 1988 was the first to discuss the possible effect of **BL-7**, **G.B.-6** and **Ex-HN-1** on **hemiplegia** through changes of the micro-circulation in the nail bed. **Shun et al, Zhen Ci Yan Jiu 1988**

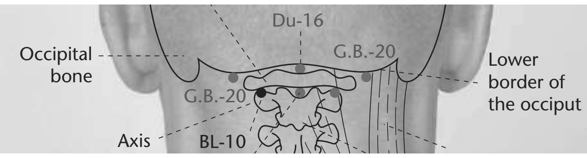

BL-10 → ST-8
A placebo-controlled, non-blinded group-comparison trial (*n* = 65) investigated the effect of treating children with an acuplaster at **BL-10**, **BL-11** and **G.B.-34** the night before undergoing **strabismus surgery,** for the prevention of **post-operative nausea and emesis**. Nausea and emesis were reduced by 50% during the early and later postoperative stage compared to the placebo-group. The well-researched point P-6 often failed to prevent nausea in this type of surgery. **Chu et al, Acta Anaesthesiol sin 1998**
BL-11 → Du-12; BL-10; L.I.-4; Ex-B-1
BL-12 → Ex-B-1
BL-13 → Ex-B-1; Du-14

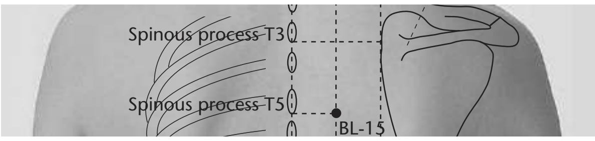

BL-15 → Ex-B-1; P-6
In an animal experiment, moxibustion at **BL-27** was able to reduce the renal excretion of Na^+ while at the same time decreasing urinary volume. Systolic blood pressure remained unchanged while plasma levels of aldosterone and ANP (atrial natriuretic peptide) decreased. In contrast, moxibustion at BL-15 increased urinary volume while reducing excretion of Na^+ and decreasing systolic blood pressure. Plasma levels of aldosterone and ANP decreased. These results suggest that BL-15 and BL-27 might be applied for certain forms of **hypertension**. **Lee et al, Am J Chin Med 1997**
BL-17 → SP-10

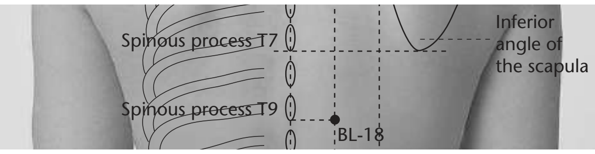

BL-18 → BL-23
A Japanese prospective, randomised, placebo-controlled, double-blinded trial (*n* = 189) investigated the effect of Japanese intradermal acupuncture at **BL-18**, **BL-19**, **BL-20**, **BL-21**, **BL-22**, **BL-23**, **BL-24**, **BL-25** and **BL-26** on **analgesic requirement, nausea and emesis** as well as stress markers after **abdominal surgery**. Starting from the recovery room and including the second postoperative day incisional pain at rest and during coughing was significantly lower in the verum group than in the control group. The requirement for epidural morphine was also significantly lower (only 50%). Nausea and emesis were also significantly lower (20–30%). Plasma cortisol and epinephrine concentrations were reduced by 30–50% in the acupuncture group during recovery and on the first postoperative day. This is one of the most valuable trials due to its excellent methodology and design. **Kotani et al, Anesthesiology 2001**
BL-19 → BL-18
BL-20 → L.I.-4; Du-1; BL-18; Du-14; P-6
BL-21 → BL-18
BL-22 → BL-18; Ex-B-2
BL-23 → L.I.-4; Du-4; Du-20; BL-18; Du-14
BL-24 → L.I.-4; BL-18
BL-25 → L.I.-4; BL-18; Ex-B-2
BL-26 → L.I.-4; BL-18

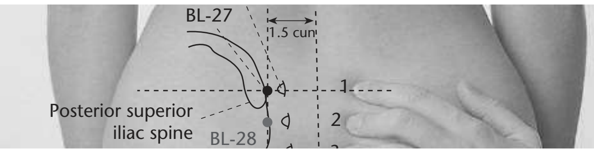

BL-27 → L.I.-4; BL-15
A German randomised, placebo-controlled experimental trial (*n* = 42) compared the effect of acupuncture at **BL-27** on pain reduction with a point of the Yamamoto New Scalp Acupuncture (YNSA). Experimental pain stimuli were set on the **upper calcaneus edge**. The difference in pain reduction between the acupuncture group and the scalp acupuncture group was highly significant. There are also highly significant differences concerning the verum and the placebo treatment. BL-27 could therefore be a useful point for treating calcaneal pain. **Ogal et al, Anasthesiol Intensivmed Notfallmed Schmerzther 2002**
An animal experiment in 15 healthy dogs investigated the effect of acupuncture on intestinal (duodenum) motility. Electro-acupuncture at BL-27 decreased the frequency of intestinal motility by 31% during stimulation, while showing an increase of 18% after stimulation. Electro-stimulation of ST-36 led to the opposite result: a 20% increase during stimulation followed by a decrease of 7% after stimulation. Based on this result, BL-27 could be applied for **paralytic ileus** or **constipation** while ST-36 could be used for treating **bloating, tenesmus and diarrhoea**. **Choi et al, J Vet Sci 2001**
BL-28 → L.I.-4; Du-4

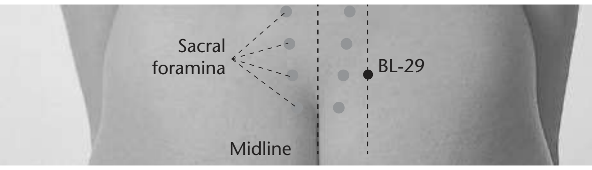

BL-29 There exists only a small case study describing **various disorders**. This is of educational value rather than anything else. **Chen Y., J Tradit Chin Med 2002**

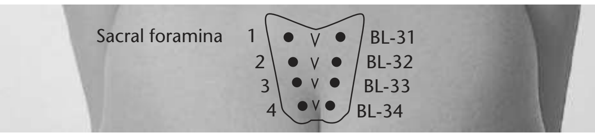
Sacral foramina
1
2
3
4
BL-31
BL-32
BL-33
BL-34

BL-32 → Du-4; Ren-3
A medium-scale Chinese case study ($n = 30$) reports the successful treatment of **urinary and faecal incontinence** due to nerve damage with electro-acupuncture at **BL-35** and **BL-32**. **Yang et al, J Tradit Chin Med 2003**

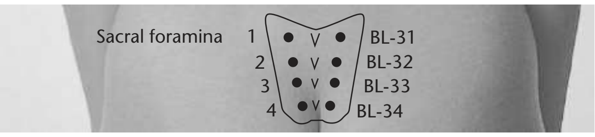
Sacral foramina
1
2
3
4
BL-31
BL-32
BL-33
BL-34

BL-33 A Japanese case study ($n = 10$) documents a significant reduction of pain for **chronic pelvic pain syndrome** with **intrapelvic venous congestion** by weekly treatments with bilateral acupuncture at **BL-33** (needle retention 10 minutes, manual rotation). **Honjo et al, Int J Urol 2004**
A Japanese prospective case study ($n = 13$) investigated the effect of bilateral acupuncture at **BL-33** on cystometry and **incontinence** in **spinal-cord-injured patients**. Incontinence completely disappeared in 15% of cases and decreased to 50% or less compared to baseline in a further 46% of patients. Maximum cystometric bladder capacity increased significantly from 75 ml to 148 ml. When cystometry was repeated in six patients one month after the last acupuncture treatment, bladder capacity showed a further average increase to 187 ml. **Honjo et al, Urol Int 2000**

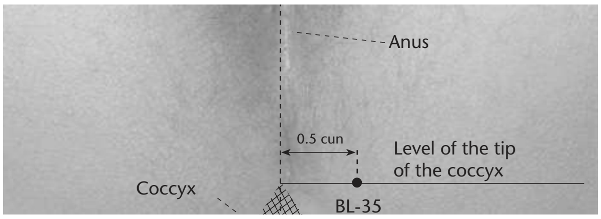
Anus
0.5 cun
Level of the tip
of the coccyx
Coccyx
BL-35

BL-35 → BL-29; BL-32
A larger-scale Chinese case study ($n = 103$) reports the successful treatment of **post-partum urinary retention** with classic acupuncture at **BL-35**. **Li et al, J Tradit Chin Med 1996**
BL-47 → L.I.-4

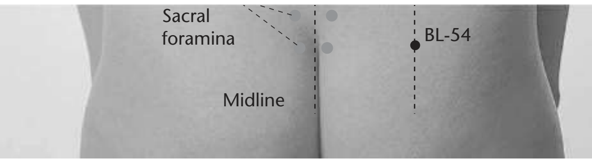
Sacral
foramina
BL-54
Midline

BL-54 → Du-3
A larger-scale Chinese case study ($n = 100$) reports the successful treatment of **lumbar pain** with acupuncture at **BL-54**. **Cui, J Tradit Chin Med 1992**

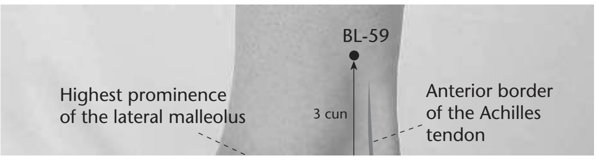
BL-59
Highest prominence
of the lateral malleolus
3 cun
Anterior border
of the Achilles
tendon

BL-59 A smaller-scale Chinese case study reports the treatment of **acute lumbar pain** with acupuncture at **BL-59**. **Hu, J Tradit Chin Med 1993**

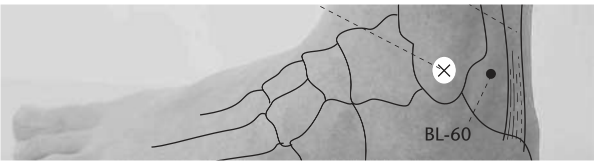

BL-60 → Du-3; ST-36; L.I.-4
A US prospective case study ($n = 7$) reports the positive subjective (questionnaire) and objective (tibial H-reflex) effect of non-invasive electro-acupuncture at **BL-60**, **ST-36**, **KID-1** and **LIV-3** on **HIV-related neuropathy**. Results showed a significant overall improvement in functional activities. **Galantino et al, J Altern Complement Med 1999**
BL-62 → S.I.-3; HE-7; Du-20

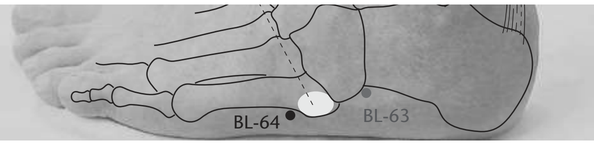

BL-64 An animal experiment demonstrated that electro-acupuncture at BL-64 and BL-65 increased the activity of nitric oxide synthase in the brain stem nuclei (nucleus gracilis). **Ma at al, Acupunct Electrother Res 2002**
BL-65 → BL-64

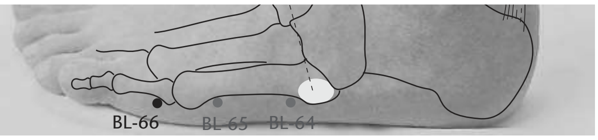

BL-66 An experimental study in rats investigated the antipyretic effect of manual acupuncture at **BL-66** on lipopolysaccharide-induced **fever**. Acupuncture at this point reduced the fever and also reduced the production of the pro-inflammatory cytokines interleukin-1 beta and interleukin-6 in the hypothalamus of the treated rats. **Son et al, Neurosci Lett 2002**

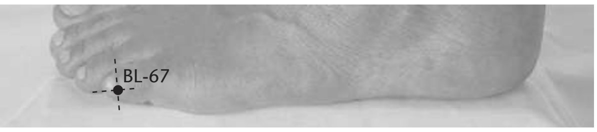

BL-67 An Italian prospective, randomised, controlled, single-blinded trial (*n* = 123) investigated the effect of moxibustion at **BL-67** on the conversion rate in **breech presentations**. The trial had to be terminated prematurely due to a high drop-out rate of participants. At that time, there were no significant differences between the moxibustion group and the control group. The authors conclude that there is too little acceptance for moxibustion treatment in non-Asian cultures. **Cardini et al, Br J Obstet Gynaecol 2005**

A randomised, controlled, prospective larger-scale trial (*n* = 67) demonstrated a higher rate of conversion of fetal **breech presentations** in a group treated with 30 minutes of manual acupuncture at **BL-67** (76.4%) compared to an untreated control group (45.4%). **Habek et al, Fetal Diagn Ther 2003**

A large-scale (*n* = 260), randomised, controlled, prospective trial demonstrated a higher rate of conversion of fetal **breech presentations** in women treated with moxibustion at **BL-67** for 7–14 days (75.4%) than the conventionally treated control group (47.7%). **Cardini et al, JAMA 1998**

A medium-scale case study (*n* = 48) with retrospective control groups demonstrated that electro-acupuncture at **BL-67** for the conversion of fetal **breech presentations** had a 81.3% success rate. **Li et al, J Tradit Chin Med 1996**

Acupuncture at **BL-67** resulted in a significant increase in the number of **c-Fos-positive cells in the primary visual cortex** of binocularly deprived rat pups to levels suggesting that BL-67 has an influence on the activity of the primary visual cortex. **Lee et al, Am J Chin Med 2002**

A recent Italian prospective, randomised trial investigated the effect of acupuncture and moxibustion at **BL-67** on the conversion of **breech presentations** in 240 pregnant women. At delivery, breech presentation was significantly lower in the active-treatment group (37%) than in the observation group (54%). Therefore the proportion of caesarean sections indicated for breech presentation was also significantly lower in the treatment group than in the observation group (52% vs 67%). **Neri et al, J Matern Neonatal Med 2004**

A Croatian prospective, randomised, controlled, non-blinded group-comparison trial (*n* = 67) assessed the value of acupuncture at **BL-67** for 30 minutes in the conversion of **breech presentation**. The success rate of the acupuncture correction of the breech presentation was 76% while the rate in the untreated control group was only 45%. This is a highly significant difference. **Habek et al, Fetal Diagn Ther 2003**

An Austrian prospective, randomised, placebo-controlled experimental trial (*n* = 10) investigated the effect of laser acupuncture at **BL-67** on the visual cortex using fMRI to monitor **neuroradiological changes**. Only verum acupuncture resulted in an activation in the cuneus (Brodmann Area BA 18) and the medial occipital gyrus (BA 19). These results confirm the traditional application of this point for ophthalmic disorders. **Siedentopf et al, Neurosci Lett 2002**

An Italian prospective, non-randomised, sham-controlled, single-blinded trial (*n* = 12 pregnant women with **breech presentation**) demonstrated that acupuncture at **BL-67** significantly reduced the fetal heart rate while increasing fetal movement and producing more accelerations. **Neri et al, J Soc Gynecol Invest 2002**

Kidney channel

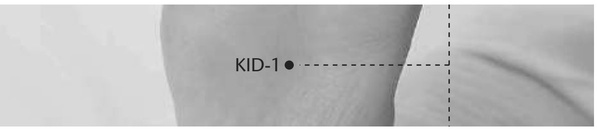

KID-1 → BL-60

A Chinese prospective, randomised, controlled, non-blinded trial (*n* = 150) investigated the effect of acupuncture at **KID-1** in the treatment of **children with bronchial asthma** compared to a control group treated with conventional drugs. The clinical effectiveness of acupuncture at KID-1 was significantly higher than treatment with drugs (89% vs 64%). In addition, in the acupuncture group the level of eosinophils and IgE was highly significant and significantly lower respectively than in the control group. **Gao et Zhu, Zhongguo Zhong Xi Yi Jie He Za Zhi 2005**

A Chinese retrospective case study reports the successful local stimulation of **KID-1** by applying a paste in the treatment of **essential hypertension**. **Gong et al, J Tradit Chin Med 1995**

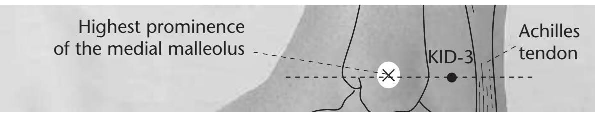

KID-3 → Du-4; L.I.-15

A Taiwanese prospective, randomised, controlled experimental trial (*n* = 157) compared the absorption of Tc-99 m (a radioactive technetium marker) at **KID-3** with non-acupuncture points. Absorption at KID-3 was significantly better than at non-acupuncture points. **Wu et al, Am J Chin Med 1994**

KID-4 → P-6
KID-5 → Du-4
KID-6 → ST-36

Pericardium channel

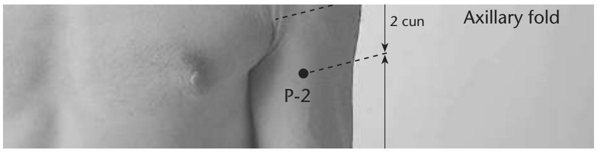

P-2 A Chinese experimental trial ($n = 100$) examined cardiac function before and during acupuncture at **P-2**, **P-3**, **P-4** and **P-6** as well as at control points (some arbitrary points and **G.B.-37**) in patients with **coronary heart disease**. In contrast to the control points the verum points had an effect – even if only a minor one – on the measured parameters (PEP, LVET, P/L, HI, SV, CO, ST segment and T wave of ECG). **You et al, Zhen Ci Yan Jiu 1993**

P-3 → P-2
P-4 → P-2

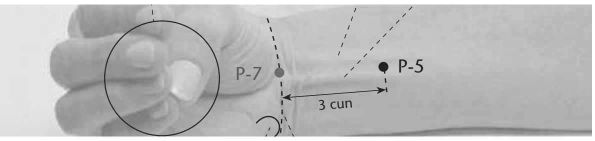

P-5 An earlier Chinese trial interpreted the synchronising effect of electro-acupuncture at **P-5** on EEG activity as having a **homeostatic effect on the whole body**. **Huang et al, Zhen Ci Yan Jiu 1990**

Electro-acupuncture at **P-5** improved the ST segment of ECG in rabbits with induced **acute myocardial infarction**. **Cao et al, Zhen Ci Yan Jiu 1990**

An earlier Chinese experimental trial ($n = 10$) found that electro-acupuncture at **P-5** and **P-6** shortened sino-atrial conduction in healthy subjects and increased frequency at the sinus node (**positive chronotropic and bathmotropic effect**). **Xi et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1993**

Previous work of the team led by Li and Tjen-A-Looi suggests that the inhibitory effect of electro-acupuncture (EA) on the pressor reflex induced by bradykinin (BK) applied to the gallbladder is in part due to the activation of opioid receptors most likely located in the rostral ventrolateral medulla (rVLM). This trial investigated the specific opioid receptor subtypes and neurotransmitters responsible for this inhibition. Therefore BK was applied to the gallbladder of anaesthetised cats to induce an increase of arterial blood pressure. It was found that applying EA at P-5 and P-6 activated the mu and delta opioid receptors located in the rVLM, thus preventing a rise of the arterial blood pressure for a longer period of time through the activating influence of the splanchnic nerve (for example due to stretching of the gallbladder or stomach). This experiment substantiated the positive effect of EA at P-5 and P-6, especially for myocardial ischaemia due to **coronary heart disease**. **Li et al, Auton Neurosci 2001; Li et al, Am J Physiol Regul Integr Comp Physiol 2002; Tjen-A-Looi et al, Auton Neurosci 2003**

In a US study 17 healthy subjects were subjected to weekly bicycle training for 3–4 weeks. Subjects were asked to perform the training with and without electro-acupuncture (EA) at **P-5**, **P-6**, **L.I.-4**, **L.I.-5**, **L.I.-6**, **L.I.-7**, **G.B.-37**, **G.B.-38** and **G.B.-39**. In 70% of subjects, EA at points on the Pericardium and Large Intestine channels led to an increase in maximal workload with decreased blood pressure. Points on the Gall Bladder channel had no effect. Further trials may investigate the application of these points in the treatment of hypertension induced by exercise stress. **Li et al, Clin Auton Res 2004**

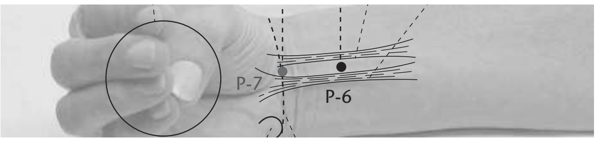

P-6 → P-2; P-5; Du-20; Du-14; HE-7; LIV-3; L.I.-4; ST-36

A Taiwanese prospective, randomised, placebo-controlled, double-blinded trial ($n = 110$) investigated the effect of acupressure bands at **P-6** on **nausea and vomiting during spinal anaesthesia for caesarean delivery**. The reduction and occurrence of nausea (64% vs 71%) and vomiting (22% vs 27%) was not significant. **Ho et al, Anesth Analg 2006**

A US prospective, randomised, placebo-controlled, single-blinded trial ($n = 94$) investigated the effect of acupressure bands at **P-6** on the occurrence of **nausea and vomiting during caesarean delivery under spinal anaesthesia**. There was no statistically significant difference between the active and sham control groups in the incidence of intraoperative nausea (30% vs 43%), postoperative nausea (23% vs 41%), intraoperative vomiting (13% vs 9%), postoperative vomiting (26% vs 34%), intraoperative antiemetic requirement (23% vs 18%), postoperative antiemetic requirement (34% vs 39%), complete intraoperative response (55% vs 57%) or complete postoperative response (51% vs 34%). These results did not reach the selected level of significance. There were also no significant differences between the groups in nausea scores, number of vomiting episodes and patient satisfaction. **Habib et al, Anesth. Analg 2006**

A Swedish prospective, randomised placebo-controlled experimental double-blinded trial ($n = 60$) investigated the effect of acupressure at **P-6** on **nausea induced by eccentric rotation**. Mean time to nausea was significantly the longest in the P-6 group (352 seconds), compared to the placebo-acupressure group (280 seconds) and the untreated control group (151 seconds). **Alkaissi et al, Can J Anaesth 2005.**

A Turkish prospective randomised controlled non-blinded trial ($n = 90$) investigated the effect of transcutaneous electrical acupoint stimulation (TEAS) at **P-6** and **Ren-13** compared to ondansetron on **postoperative nausea following paediatric**

tonsillectomy. There was the same incidence of nausea episodes in the TEAS group and in the ondansetron group, but significantly fewer episodes than in the untreated control group. Side-effects were significantly higher in the ondansetron group. **Kabalak et al, J Altern Complement Med 2005**

An Indian prospective, randomised, placebo-controlled, double-blinded trial (*n* = 120) investigated the effect of capsicum-plasters on **P-6** on **postoperative nausea and vomiting** (PONV) compared to conventional treatment with ondansetron. For the duration of the plaster application (6 hours after the operation) the incidence of PONV and the requirement for antiemetics were significantly lower in both groups. **Misra et al, Can J Anaesth 2005**

A Turkish prospective, randomised, placebo-controlled, single-blinded trial (*n* = 127) investigated the effect of transcutaneous electric nerve stimulation (TENS) at **P-6** on the incidence and severity of **nausea and gagging during gastroscopies**. There were no significant differences between the treatment group and placebo group. **Tarcin et al, Turk J Gastroenterol 2004**

A Chinese prospective, non-blinded, randomised, group-comparison trial (*n* = 30) investigated the effect of acupuncture at **P-6** plus mexiletin on **intermittent ventricular extrasystole**. The treatment group had a total effective rate of 90% compared with the control group, who was administered mexiletin only (80% total effective rate). Only the acupuncture group showed no increase in extrasystole. **Zhang et al, J Tradit Chin Med 2004.**

A Chinese group-comparison trial investigated the effect of acupuncture at **P-6** in the treatment of patients with angina pectoris and **acute myocardial infarct**. The effectiveness of acupuncture was significantly better (91%) than conventional treatment with isosorbide dinitrate and nifedipine. **Meng et al, J Tradit Chin Med 2004**

A US prospective, controlled, group-comparison trial (*n* = 77) investigated the effect of stimulation at **P-6** on the symptoms of **motion sickness** during exposure to optokinetic drum rotation. Despite the extreme test conditions, stimulating P-6 delayed the onset of symptoms considerably. **Miller et al, Aviat Space Environ Med 2004**

A Croatian prospective, single-blinded group-comparison trial (*n* = 36) investigated the efficiency of acupuncture at **P-6** in the treatment of **hyperemesis gravidarum**. Treatment results were measured by the requirement for anti-emetic medication. Acupuncture at P-6 had an efficiency rate of 90%, acupressure at P-6 64%, placebo acupuncture 12.5% and placebo acupressure 0%. **Habek et al, Forsch Komplementarmed Klass Naturheilkd 2004**

In a German prospective, randomised, placebo-controlled double-blinded group-comparison trial acupuncture at **P-6** significantly reduced the incidence of **emesis after gynaecological and breast surgery** (40% placebo, 25% P-6). Postoperative emesis was only reduced for gynaecological surgery, not for breast surgery. **Streitberger et al, Anaesthesia 2004**

A Chinese prospective, randomised, single-blinded, placebo-controlled group-comparison trial (*n* = 41) investigated the effect of twice daily acupressure at **P-6**, **ST-36** and **SP-6** on **gastrointestinal motility following trans-abdominal hysterectomy**. In comparison to the control group receiving acupressure on sham points, gastrointestinal motility (measured with a multifunctional stethoscope) increased significantly in the acupressure group. **Chen et al, Am J Chin Med 2003**

A German prospective, randomised, placebo-controlled, single-blinded trial (*n* = 80) investigated the anti-emetic effect of acupuncture at **P-6** in addition to administration of ondansetron in patients with **chemotherapy-related nausea**. There was no significant difference between the treatment and control group. **Streitberger et al, Clin Cancer Res 2003**

A US prospective, randomised, placebo-controlled, double-blinded trial (*n* = 230) investigated the effect of TENS at **P-6** for the relief of **nausea and vomiting in pregnancy**. Based on the Rhodes Index, results in the verum group were significantly better than in the control group. **Rosen et al, Obstet Gynecol 2003**

A British prospective, partially randomised, partially blinded, placebo-controlled trial (*n* = 301 patients with **acute myocardial infarction**, 125 of whom acted as a non-randomised control group) compared the effect of wristband-acupressure at **P-6** with placebo-acupressure. P-6 led to a significant reduction in **nausea and vomiting** during the last 20 hours of the 24-hour study phase (18%), compared with the placebo group (32%) and the control group (43%). **Dent et al, Complement Ther Med 2003**

A Swedish prospective case study (*n* = 39) investigated the effect of **P-6** combined with ondansetron for **nausea and vomiting** associated with **cyclophosphamide chemotherapy**. The authors stated that, compared with ondansetron treatment alone, the combined acupuncture-ondansetron treatment was significantly more effective, but they do not mention a control group. **Josefson et al, Rheumatology (Oxford) 2003**

In a US prospective, randomised, single-blinded trial (*n* = 53) the application of 'minute sphere' acupressure at **P-6**, **ST-36**, **SP-6** and **SP-4** showed no decrease of **postoperative pain and morphine requirement after abdominal surgery**. **Sakurai et al**, **Anesth Analg 2003**

A Swedish prospective, randomised, placebo-controlled, double-blinded, multi-centre trial (*n* = 410) investigated the effect of **P-6** for the treatment of **postoperative nausea and vomiting after gynaecological surgery**. The incidence of nausea and vomiting was significantly lower in the acupuncture group (33%) than in the control group (46%). **Alkaissi et al, Can J Anaesth 2002**

A US prospective, randomised, placebo- and sham-controlled, double-blinded trial (*n* = 120) investigated the effect of the prophylactic application of a 'ReliefBand' at **P-6** on the incidence of **nausea and vomiting after plastic surgery,** compared to the application of the ReliefBand in addition to 4 mg ondansetron. The occurrence of nausea and vomiting as well as the need for antiemetic 'rescue' medication was significantly the lowest in the P-6 plus ondansetron group. **White et al, Anesthesiology 2002**

A Korean prospective, randomised, placebo-controlled, double-blinded trial (*n* = 160) investigated among other parameters the effect of a capsicum plaster applied at **P-6** on **postoperative nausea and vomiting**. The incidence of nausea and vomiting was significantly less (P-6: 26% within 24 hours, placebo: 57%). **Kim et al, Anesth Analg 2002**

A Japanese case study investigated the effect of different stimulation techniques at **P-6** on the coronary arteries in patients with **coronary heart disease**. The mean coronary dilation with acupuncture was 69% of that caused by isosorbide dinitrate. **Kurono et al, Am J Chin Med 2002**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 187$) investigated the effect of **P-6** on the incidence of **postoperative nausea and vomiting** in children. P-6 produced significantly better results than the sham points and was equally effective as administration of droperidol, but without the sedative, hypotensive side effects. **Wang et al, Anesthesiology 2002**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 27$) investigated the effect of **P-6** on the incidence of **nausea and vomiting during chemotherapy**. Compared to the control group, the anti-emetic requirement was significantly lower in the P-6 group. **Roscoe et al, Altern Ther Health Med 2002**

An Austrian prospective, randomised, placebo-controlled, experimental cross-over study ($n = 51$) investigated the effect of acupuncture at **P-6** on the **skin blood perfusion** measured by laser Doppler perfusion imaging. This was compared to needling a sham point. Changes in the skin blood perfusion occurred significantly earlier in the acupuncture group than in the sham group, with a basically more pronounced reduction in skin blood perfusion at the verum point. **Litscher et al, Lasers Med Sci 2002**

A large-scale Australian prospective, randomised, sham-controlled trial ($n = 593$) investigated the effect of acupuncture at **P-6** on the incidence **of nausea and vomiting in early pregnancy**. Women receiving traditional acupuncture reported significantly less nausea after the second treatment and significantly less dry retching after the third week compared with women in the control group. Women treated with sham acupuncture showed significantly less nausea and retching after the third treatment compared with untreated women. Individualised acupuncture resulted in significantly less nausea and retching already after the first treatment. In this study none of the treatments had a significant influence on the frequency of vomiting. **Smith et al, Birth 2002**

A US prospective, randomised, sham-controlled trial ($n = 120$) investigated the influence of electro-acupuncture (EA) at **P-6** on postoperative **nausea and vomiting** in paediatric patients who had undergone **ENT surgery** and were awake. EA at P-6 significantly decreased the incidence of nausea (P-6: 60%, sham: 85%, control group: 93%). While vomiting had the lowest incidence in the P-6 group, the difference from the other groups was not significant. **Rusy et al, Anesthesiology 2002**

A Swedish prospective, randomised, placebo-controlled pilot study ($n = 60$) investigated the effect of tuina (acupressure) at **P-6** on **pregnancy-related nausea and vomiting**. The frequency of both nausea and vomiting was significantly lower in the acupressure group compared to the sham acupressure group and untreated control group. **Werntoft et al, J Reprod Med 2001**

A US prospective, randomised, controlled trial ($n = 25$) investigated the effect of an acupressure wrist band at **P-6** on **motion sickness**. Both the symptoms of motion sickness and abnormal gastric activity as recorded via EEG were significantly lower in the group treated with wristbands at P-6. **Stern et al, Altern Ther Health Med 2001**

A US prospective, randomised, placebo-controlled, single-blinded trial investigated the effect of ‘sea-bands’ (acupressure wristbands) at **P-6** on the incidence of **postoperative nausea and vomiting**. There were no significant differences. **Windle et al, J Perianesth Nurs 2001**

A Swedish prospective, randomised, sham-controlled, single-blinded cross-over study ($n = 33$) investigated the effect of acupuncture at **P-6** on pregnant women with **hyperemesis gravidarum** (vomiting in pregnancy). The verum group showed a significantly faster reduction of the nausea, and also a higher number of the patients experienced no vomiting. **Carlsson et al, J Pain Symptom Manage 2000**

A Chinese prospective, controlled, randomised, group-comparison study ($n = 181$) and a Chinese case study ($n = 33$) demonstrated that electro-acupuncture at **P-6** and **P-8** as well as at **L.I.-4**, **T.B.-5**, **ST-36** and **SP-6** with de-escalating treatment frequency from four times daily to once weekly for a period of 15 and 28 days had a significant effect on the **withdrawal symptoms of heroin addicts**. **Zhang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2000; Wu et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2000**

A Taiwanese prospective, non-randomised, placebo-controlled, single-blinded experimental cross-over study ($n = 44$) investigated the effect of acupuncture at **P-6** on the left ventricular ejection fraction (LVEF) in 22 healthy subjects and 22 patients with **coronary artery disease** (CAD). There was no change of LVEF in the healthy subjects while LVEF significantly increased in the CAD patients. **Ho et al, Am J Chin Med 1999**

A US prospective, randomised, placebo-controlled, single-blinded trial ($n = 100$) investigated the effect of pre-operative acupressure (verum group) and intra-operative acupuncture (verum and control group) at **P-6** on **postoperative nausea and vomiting** after **tonsillectomy** in children. There were no significant differences. **Shenkman et al, Anesthesiology 1999**

A Swedish prospective, randomised, placebo-controlled, double-blinded trial ($n = 60$) investigated the effect of acupressure at **P-6** on **nausea and vomiting** after **minor gynaecological surgery**. Only the group receiving acupressure at P-6 experienced no vomiting and also required no anti-emetic medication. **Alkaissi et al, Acta Anaesthesiol Scand 1999**

An Austrian prospective, randomised, placebo-controlled, double-blinded trial investigated the effect of laser acupuncture at **P-6** on **nausea and vomiting in children undergoing strabismus surgery**. The laser acupuncture was administered 15 minutes before induction of anaesthesia and 15 min after arriving in the recovery room. In the laser stimulation group, the incidence of vomiting was significantly lower (25%) than that in the placebo group (85%). **Schlager et al, Br J Anaesth 1998**

A Chinese prospective, randomised, placebo-controlled, double-blinded trial ($n = 163$) investigated the effect of acupoint stickers at **P-6**, **Ren-8** and **T.B.-18** on **motion sickness**. All tested acupuncture points showed a significantly better result than the active control points (scopolamine plaster) and the placebo control (lactose plaster). T.B.-18 was the most effective point (symptoms 100% reduced), followed by Ren-8 (81%) and P-6 (46%). **Pei et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1998**

A Chinese prospective, randomised, controlled trial ($n = 50$) compared two forms of anaesthesia for the **anterior approach cervical discectomy**: local anaesthesia plus intravenous anaesthesia (IVA) in the control group and electro-acupuncture at **P-6** and **L.I.-4** plus IVA. The effect of the anaesthesia was the same in both groups. The authors therefore recommend electro-acupuncture at P-6 and L.I.-4 as a less risky alternative. **Li et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1997**

A British prospective, randomised, placebo-controlled, double-blinded trial ($n = 81$) investigated the effect of intra-operative acupuncture at **P-6** in the prevention of **postoperative nausea and vomiting** in patients undergoing **gynaecological laparoscopic surgery**. The use of acupuncture significantly reduced the incidence of postoperative nausea and vomiting by 30–38% compared with placebo. **Al-Sadi et al, Anaesthesia 1997**

A US prospective, randomised, placebo- and sham-controlled, double-blinded trial ($n = 75$) investigated the effect of acupressure at **P-6** on **nausea and vomiting during caesarean section under spinal anaesthesia** compared with administration of 10 mg metoclopramide. Acupressure at P-6 is as effective as 10 mg metoclopramide. **Stein et al, Anesth Analg 1997**

An Australian prospective, randomised, placebo-controlled, double-blinded trial ($n = 84$) investigated the effect of TENS at **P-6** and **L.I.-4** on **postoperative nausea and vomiting** in **paediatric patients** undergoing **minor urological surgery**. The differences between the treatment and control group were not statistically significant. **Schwager et al, Anaesth Intensive Care 1996**

A Taiwanese prospective, randomised, sham-controlled, single-blinded experimental trial ($n = 48$) evaluated the relationship between electro-acupuncture at **P-6** and **ST-36** and cardiopulmonary function in healthy subjects. In the P-6/ST-36 group **resting heart rate**, **carbon dioxide production** and **oxygen requirement** were significantly decreased, indicating that acupuncture can lower the metabolic rate. **Lin et al, Chin Med J (Engl) 1996**

A Taiwanese prospective, randomised, placebo-controlled, double-blinded trial ($n = 60$) investigated the effect of acupressure wristbands at **P-6** on the incidence of **nausea and vomiting after epidural morphine for post-caesarean section pain relief**. Compared to the control group, the incidence of nausea significantly decreased from 43% to 3%, vomiting significantly decreased from 27% to 0%. **Ho et al, Acta Anaesthesiol Scand 1996**

A US prospective, randomised, placebo-controlled, single-blinded cross-over study ($n = 9$) investigated the effect of acupressure wristbands at **P-6** on nausea and vomiting for **seasickness**. Depending on the time of application (earlier or later during the trip) subjects were either symptom-free or complained about nausea and vomiting. The difference was highly significant. **Bertolucci et al, Aviat Space Environ Med 1995**

A US prospective, randomised, placebo- and sham-controlled, single-blinded experimental trial ($n = 9$) investigated the effect of acupressure at **P-6** on the incidence of visually-induced **motion sickness**. Acupressure at P-6 significantly reduced the incidence of nausea and abnormal gastric myoelectric activity. **Hu et al, Aviat Space Environ Med 1995**

A Chinese prospective, randomised, controlled trial ($n = 40$) investigated the effect of acupuncture on **P-6**, **L.I.-4**, **ST-36** and **KID-4** on the regulation of cellular immune function in patients with **malignant tumours**. There was a highly significant increase of the CD3+ and CD4+ levels, an increase in the CD4+/CD8+ ratio, a higher endorphin level and a decreased level of soluble interleukin-2 receptor (SIL-2R). **Wu, Zhen Ci Yan Jiu 1995**

A British prospective, randomised, controlled trial ($n = 46$) investigated the effect of acupressure at **P-6** on **nausea and vomiting** following **laparoscopy** for **gynaecological surgery** and the requirement for anti-emetic therapy. There was a significant reduction in the requests for anti-emetic therapy in the P-6-acupuncture group. **Allen et al, Anaesth Intensive Care 1994**

A Chinese prospective paired trial ($n = 40$) investigated the effect of electro-acupuncture (EA) at **P-6** during different times of the day (*chen*-time: 7–9am, *xu*-time: 7–9pm) on left ventricular function (LVF) in patients with **coronary heart disease**. EA performed at *chen*-time improved LVF, while EA administered under the same conditions during *xu*-time led to an impairment of LVF. **Li et al, J Tradit Chin Med 1994**

A British prospective case study ($n = 27$) reports that wearing a special wrist band decreased **pregnancy-related nausea and vomiting** more than 50%. **Stainton et al, Health Care Women Int 1994**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 60$) investigated the effect of acupressure at **P-6** on **pregnancy-related nausea and vomiting**. There was a significant decrease of nausea in the P-6 acupressure group. The incidence of vomiting did not differ from the sham control group. **Belluomini et al, Obstet Gynecol 1994**

A German prospective, randomised, placebo-controlled, single-blinded trial ($n = 60$) investigated the effect of acupressure at **P-6** on **nausea** in patients undergoing **gynaecological operations of longer duration** (6–8 h). Nausea was reduced from 53% in the placebo group to 23% in the acupressure group. **Gieron et al, Anaesthesist 1993**

A Chinese prospective, randomised, placebo-controlled, single-blinded trial ($n = 15$) investigated the effect of acupuncture at **P-6**, **HE-7** and **HE-3** on the frequency and duration of **angina pectoris** during exercise. The control groups received either sham acupuncture or no treatment at all. In the verum group, the anginal attack occurred significantly later and the duration of the attack after stopping the exercise was significantly shorter than in the control groups. **Zhou et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1993**

A Taiwanese prospective, randomised, controlled trial ($n = 120$) investigated the effect of **P-6** acupoint injection with 0.2 ml 50% glucose in water compared to intravenous injection of 20 micrograms/kg droperidol for the prevention of **vomiting after gynaecological laparoscopy**. A non-controlled comparison-group received no treatment at all. The incidence of vomiting significantly decreased in both the P-6 group and the droperidol group. **Yang et al, Acta Anaesthesiol Scand 1993**

An Italian prospective, randomised, placebo-controlled, double-blinded cross-over study ($n = 60$) investigated the effect of unilateral and bilateral acupressure at **P-6** for the treatment of **morning sickness** compared to acupressure at a sham point. Compared to the placebo group both unilateral and bilateral acupressure at P-6 significantly reduced the frequency of morning sickness from 60% to 30%. **De Aloysio et al, Obstet Gynecol 1992**

A US ‘n of 1’ trial investigated the effect of **P-6** wristbands on **nausea and vomiting in hospice patients** ($n = 6$). There were no significant differences between the verum group, the placebo group and the group without wristbands. **Brown et al, Am J Hosp Palliat Care 1992**

A Canadian prospective, randomised, controlled, double-blinded trial ($n = 90$) investigated the effect of acupuncture at **P-6**, administered after induction of anaesthesia, on the frequency of **nausea and vomiting after strabismus surgery in children**. The control group received droperidol. There was no significant difference between the two groups. **Yentis et al, Can J Anaesth 1992**

A Canadian prospective, randomised, controlled, single-blinded trial ($n = 45$) investigated the effect of acupuncture at **P-6** administered after induction of anaesthesia on the frequency of **nausea and vomiting after tonsillectomy in children**. There was no significant difference between the verum group and the untreated control group. **Yentis et al, Br J Anaesth 1991**

A US prospective, randomised, controlled, double-blinded trial ($n = 66$) investigated the effect of acupuncture at **P-6** administered after induction of anaesthesia on the frequency of **nausea and vomiting after strabismus surgery in children**. There was no significant difference between the treatment group and placebo group. **Lewis et al, Br J Anaesth 1991**

An Irish case study ($n = 100$) reports that TENS at **P-6** reduced **chemotherapy-induced nausea** by 75%. **Dundee et al, J R Soc Med 1991**

A Swedish prospective, randomised, controlled cross-over study ($n = 21$) investigated the effect of acupuncture at **P-6**, **HE-5**, **BL-15**, **BL-20** and **ST-36** on the frequency of **angina pectoris** (AP) and the performance before onset of pain during exercise in patients suffering from coronary heart disease. The frequency of AP attacks decreased significantly from 11 to 6 attacks per week. Accordingly, the performance before onset of pain during exercise increased significantly from 82 W to 94 W. **Richter et al, Eur Heart J 1991**

A Chinese prospective, randomised, non-blinded, controlled trial ($n = 64$) investigated the effect of a 3 ml normal saline injection into **P-6** on the incidence of **postoperative vomiting**. Compared to the control group, vomiting was significantly reduced (2 vs 10 patients) in the treatment group. **Shyr et al, Ma Zui Xue Za Zhi 1990**

A British prospective, randomised, controlled experimental trial ($n = 18$) compared the effectiveness of acupressure bands at **P-6** on **motion sickness** to a placebo and 0.6 mg scopolamine. Only the subjects taking scopolamine showed a significant increase in tolerance to a laboratory nauseogenic cross-coupled motion challenge. **Bruce et al, Aviat Space Environ Med 1990**

A Taiwanese prospective, randomised, controlled trial ($n = 100$) compared the effect of electro-acupuncture (EA) and TENS at **P-6** on **nausea and vomiting after laparoscopy**. Compared to the control group (44%), nausea was significantly reduced in the EA and perchlorperazine group (12% each), but not in the TENS group (36%). **Ho et al, Anaesthesia 1990**

A British prospective, randomised, controlled trial ($n = 162$) compared the effect of acupressure wristbands at **P-6** on **postoperative nausea and vomiting** in the recovery room with conventional anti-emetic treatment. Acupressure at P-6 significantly reduced the intensity of the nausea. The reduced anti-emetic requirement and the reduced incidence of vomiting were not statistically significant. **Barsoum et al, J R Soc Med 1990**

A Chinese prospective, randomised, controlled trial ($n = 40$) compared the effect of electro-acupuncture (EA) at **L.I.-4** and **P-6** on intra-operative **analgesia** in patients undergoing **thyroidectomy** with superficial cervical plexus block. There was no significant difference between the two groups. **Ouyang et al, Zhen Ci Yan Jiu 1990**

An Irish prospective, randomised, sham-controlled, single-blinded trial investigated the effect of pre-operative electro-acupuncture and acupuncture at **P-6** on the frequency of **nausea and vomiting after minor gynaecological surgery** compared to sham acupuncture. The frequency of nausea and vomiting was significantly reduced in the P-6-group only. **Dundee et al, Br J Anaesth 1989**

A US prospective, randomised, controlled cross-over study ($n = 16$) investigated the efficacy of acupressure wristbands at **P-6** on **morning sickness**. Use of the acupressure wristbands significantly relieved morning sickness and also significantly reduced anxiety, depression and behavioural dysfunction. **Hyde, J Nurse Midwifery 1989**

An Irish prospective, randomised, sham-controlled cross-over study ($n = 10$) investigated the effect of electro-acupuncture on **P-6** on **chemotherapy-induced nausea**. In the verum group sickness was either completely absent or reduced considerably in 97% of patients and no side effects were encountered. The same research team confirmed these results in a larger-scale case study ($n = 105$). **Dundee et al, J R Soc Med 1988**

An Irish prospective, randomised, sham-controlled trial ($n = 350$) compared the effect of acupressure at P-6 on **morning sickness** to sham acupressure and a control group. Only acupressure at P-6 significantly relieved the symptoms of morning sickness. **Dundee et al, J R Soc Med 1988**

A prospective, randomised, placebo-controlled trial ($n = 75$) investigated the effect of acupuncture at **P-6** on **nausea and vomiting after minor gynaecological surgery**. Acupuncture at P-6 significantly reduced both nausea and vomiting. **Dundee et al, Br Med J (Clin Res Ed) 1986**

P-8 → P-6; LU-10

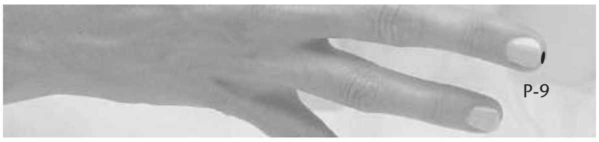

P-9 A larger-scale Chinese case study ($n = 100$) describes the positive effect of **P-9** on **infantile morbid night crying** in otherwise healthy children. **Zhao, J Tradit Chin Med 2002**

Triple Burner channel

T.B.-5 → P-6; L.I.-4; L.I.-12; L.I.-10

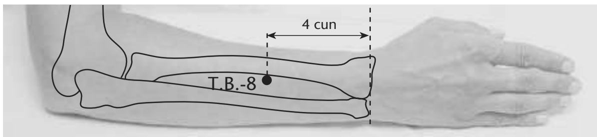

T.B.-8 Acupuncture at **T.B.-8** and **Du-15** activates areas important for **word generation** in the right inferior frontal gyrus, but not in the left inferior frontal gyrus. **Li et al, Hum Brain Mapp 2003**

A Russian case study reports the positive effects of acupuncture at **T.B.-8**, **T.B.-9**, **ST-20**, **ST-36**, **G.B.-14** and **Du-20** in 66 patients with **duodenal ulcer** accompanied by apparent shifts in psychovegetative correlations. **Kravtsova et al, Vopr Kurortol Fizioter Lech Fiz Kult 1994**

T.B.-9 → T.B.-8

T.B.-18 → P-6

Gall Bladder channel

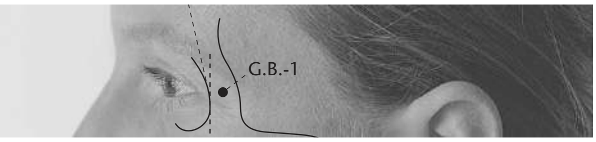

G.B.-1 An earlier Italian case study reports the successful application of **G.B.-1** for **heroin withdrawal** in nine male heroin addicts. After six months, 66% of these patients were classified as abstinent. Besides reservations regarding methodology of such a small-scale trial, the accompanying medication also poses problems in evaluating the effect of the acupuncture treatments. **Cocchi et al, Minerva Med 1979**

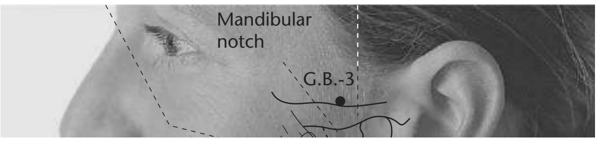

G.B.-3 A US prospective, randomised, double-blinded, experimental cross-over study ($n = 20$) investigated the effect of TENS at **G.B.-3** on the **anaesthetic requirement** (desflurane) to prevent purposeful movement of the extremities in response to noxious electrical stimulation. In comparison with the placebo group, bilateral application of TENS at G.B.-3 significantly reduced the desflurane requirement by 11%. **Greif et al, Anesthesiology 2002**

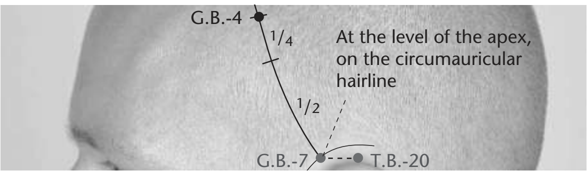

G.B.-4 Following the induction of an **ischaemic cerebral infarction** by occlusion of the middle cerebral artery in rats, animals treated with electro-acupuncture (EA) at **G.B.-4**, **G.B.-5**, **G.B.-6** and **G.B.-7** showed a significantly faster recovery than animals in the untreated control group or in the group treated with EA at **Du-20** and **Du-26**. Infarct volume was significantly reduced in the EA-treated groups. The authors discuss the acupuncture-induced production of vascular endothelial growth factor (VEGF) in astrocytes of the peri-infarct area as a causative factor. **Wang et al, Neurol Res 2003**

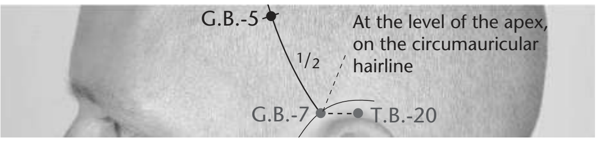

G.B.-5 → G.B.-4; ST-8

A Chinese case study recommends a particular needling technique (from **G.B.-5** or **Ex-HN-5** to **G.B.-8**) for the treatment of **migraine**. **Gan et al, J Tradit Chin Med 1986**

G.B.-6 → BL-7; G.B.-4

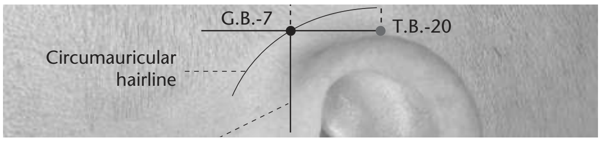

G.B.-7 → G.B.-4

A large-scale Chinese case study (*n* = 500) reports the successful application of acupuncture at **G.B.-7** in the treatment of **hemiplegia** after apoplexy. **Sun et al, J Tradit Chin Med 1985**

G.B.-8 → G.B.-5

G.B.-11 → ST-8

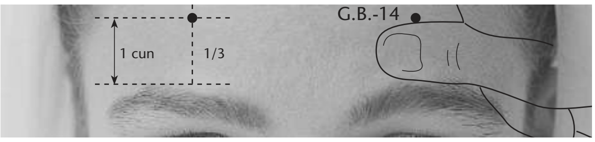

G.B.-14 → T.B.-8; ST-8; LIV-3

A case study reports that 63 patients with **facial paralysis** were treated with acupuncture at **G.B.-14**, **BL-1**, **ST-2**, **L.I.-20**, **ST-4**, **ST-6** and **ST-18** plus auxiliary points according to the underlying TCM pattern. Just over half of the patients were cured and all except one patient showed a marked improvement. **Liu et al, Zhen Ci Yan Jiu 1992**

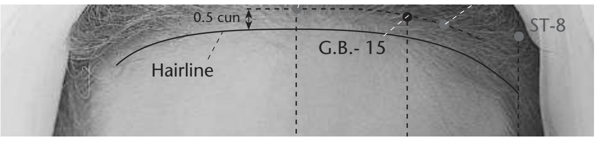

G.B.-15 An uncontrolled, non-randomised, group-comparison trial investigated the effect of acupuncture on **childhood diarrhoea**. Group I received scalp acupuncture (after Yamamoto) with points on line 3 starting at **G.B.-15**; Group II received body acupuncture; and Group III was treated with antibiotics. Group I showed the best results, followed by Group II and Group III, with results in Group I and II significantly better than in Group III. However, besides some methodological shortcomings, the authors did not seem to have a very firm grasp of their subject matter: while using the correct points for the correct indication, they refer to line 3 as line 2 and to G.B.-15 as BL-15! **Lin et al, J Tradit Chin Med 1993**

G.B.-20 → ST-8; LIV-3; S.I.-18; L.I.-15

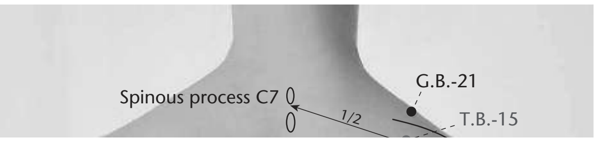

G.B.-21 → L.I.-15

A Chinese case study reports the successful treatment of **achalasia of the cardia** with acupuncture at **G.B.-21**. **Shi et al, J Tradit Chin Med 1994**

G.B.-24 → LIV-14

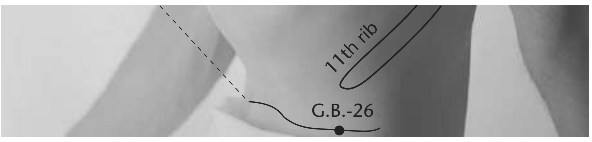

G.B.-26 An earlier prospective group-comparison trial investigated the **analgesic effect** of electro-acupuncture (EA) at **G.B.-26**, **ST-36**, **SP-6** and **HE-7** in patients who underwent hysterectomy by subumbilical midline incision. EA for 40 minutes had the same analgesic effect as 30 mg pentazocine. However, in contrast to pentazocine; EA also improved the vital capacity of the lung for 3–4 hours after the treatment. **Facco et al, Am J Chin Med 1981**

G.B.-30 → L.I.-12; L.I.-15

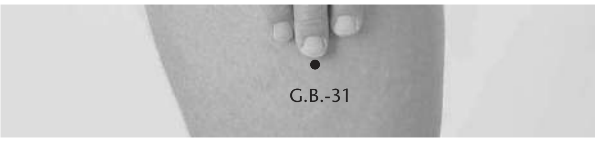

G.B.-31 → L.I.-4

Three publications by a Canadian research team led by Romita demonstrated that electro-acupuncture (EA) at **G.B.-31**, **ST-32** and **ST-36** resulted in a more pronounced inhibition of the **nociceptive tail withdrawal reflex** than classical acupuncture in slightly anaesthetised rats. **Romita et al, Brain Res 1997, Romita et al, Brain Res Bull 1997, Romita et al, Brain Res 1996**

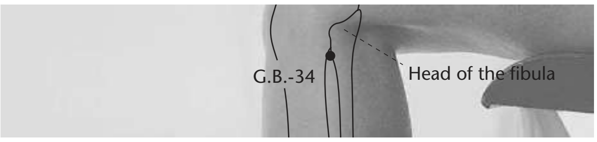

G.B.-34 → BL-10; Du-20; ST-36; L.I.-4; L.I.-15

A Korean prospective randomised placebo-controlled experimental trial (*n* = 10) investigated the effect of acupuncture at **G.B.-34** on motor cortex activities as evidenced by **fMRI**. Bilateral sensorimotor areas BA 3, 4, 6 and 7 were activated during acupuncture, providing a basis for future investigations regarding therapeutic interventions in **stroke** patients. **Jeun et al, Am J Chin Med 2005**

A Chinese prospective, randomised, controlled trial (*n* = 20) investigated the effectiveness of electro-acupuncture (EA) at **G.B.-34** and **ST-38** on **epicondylitis lateralis (tennis elbow)** compared to manual acupuncture at the same points. After six treatments over a period of two weeks, the EA group showed significantly better results in relation to pain relief (Pain visual analogue scale) and pain free hand grip strength (PFG). **Tsui et al, Acupunct Electrother 2002**

A British, prospective, randomised, sham-controlled, double-blinded experimental trial ($n = 240$) investigated the effect of transcutaneous electric nerve stimulation (TENS) at **G.B.-34** on mechanical **pain thresholds**. Low frequency, high intensity stimulation showed a significant analgesic effect that was sustained post-stimulation. **Chesterton et al, Pain 2002**

G.B.-37 → P-2; P-5; L.I.-15

G.B.-38 → P-5

G.B.-39 → LIV-3

G.B.-40 → LIV-3

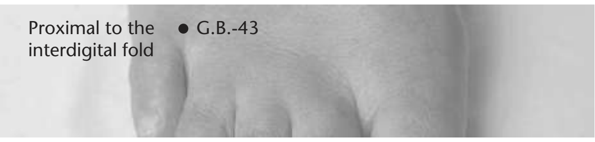

G.B.-43 An Austrian prospective, randomised, placebo-controlled, single-blinded trial investigated the effect of laser acupuncture at **G.B.-43** on cerebral cortical and subcortical activations as evidenced by **fMRI**. Only true laser acupuncture led to ipsilateral activations within the thalamus, nucleus subthalamicus, nucleus ruber, the brainstem, and the Brodmann areas 40 and 22. **Siedentopf et al, Laser Med Sci 2005**

A purely descriptive study investigated the effect of **G.B.-43** and **L.I.-4** on **somatosensory evoked potentials**. Particular attention was paid to the effect of the *deqi*-sensation. **Wu et al, Zhen Ci Yan Jiu 1993**

Liver channel

LIV-1 → SP-1

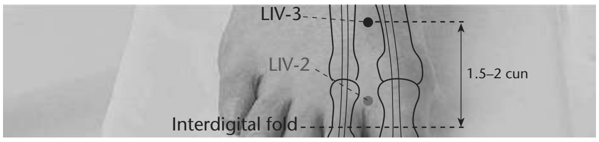

LIV-3 → Du-20; BL-60; L.I.-4; L.I.-15; SP-6

In a Chinese-German single-blinded experimental trial, acupuncture at **LIV-3** and **G.B.-40** resulted in the activation of the secondary somatosensory cortical areas, frontal areas, the right side of the thalamus and the left side of the cerebellum. Sham-points had no effect. **Fang et al, Neuroradiology 2004**

An Italian open, prospective, randomised group-comparison trial investigated the effect of acupuncture, laser acupuncture and TENS on **LIV-3**, **SP-6**, **L.I.-4**, **G.B.-20**, **Du-20** and **Ex-HN-5** on **transformed migraines**. The number of days with headache per month significantly decreased during treatment in all groups. Acupuncture showed the best effectiveness over time. **Allais et al, Neurol Sci 2003**

An Italian prospective, randomised, controlled, non-blinded trial ($n = 160$) investigated the effectiveness of acupuncture at **LIV-3**, **SP-6**, **ST-36**, **Ren-12**, **L.I.-4**, **P-6**, **G.B.-20**, **G.B.-14**, **Ex-HN-5** and **Du-20** on **migraine attacks** without aura compared to flunarizine. The number of attacks after two and four months of therapy was significantly lower in the acupuncture group than in the flunarizine group, and analgesic consumption was accordingly also significantly lower in the acupuncture group. Pain intensity was significantly reduced only by acupuncture treatment. **Allais et al, Headache 2002**

A British prospective, randomised, placebo-controlled cross-over study ($n = 56$) investigated the effect of **LIV-3** on **rheumatoid arthritis (RA)**. There were no significant changes between the treatment and placebo group. The treatment of only one point for a condition such as RA is unusual. **David et al, Rheumatology (Oxford) 1999**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 10$) investigated the effect of electro-acupuncture (EA) on **LIV-3**, **ST-36** and **L.I.-11** on lowering diastolic blood pressure in hypertensive subjects. Compared to the sham group, diastolic blood pressure showed a significant decrease in the EA-group. **Williams et al, Phys Ther 1991**

LIV-13 → Du-20

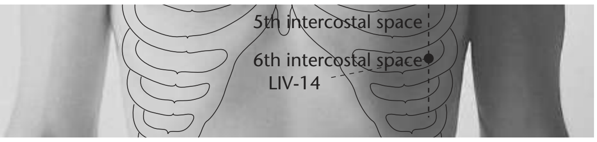

LIV-14 A Chinese prospective, non-randomised, controlled, non-blinded trial ($n = 56$ plus control group) investigated the effect of water injections at **LIV-14**, **G.B.-24** and **Ren-14** on biliary colic pain. The pain disappeared in 57% of cases and showed improvement in a further 39%. This result is significantly better compared to the control group receiving conventional treatment. **Jiang et al, J Tradit Chin Med 1995**

Du mai (governing vessel)

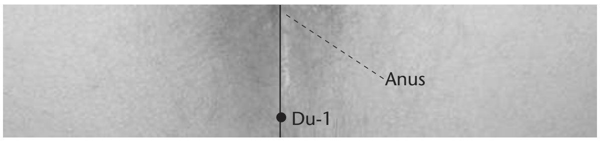

Du-1 A prospective, controlled, group-comparison trial in **hypogalactic postpartum sows** ($n = 42$) showed that apitherapy (acupuncture with bee venom) at **Du-1** and **ST-18** was superior to treatment with conventional drugs. **Choi et al, J Vet Sci 2001**

A prospective, controlled, group-comparison animal experiment showed that apitherapy at **Du-1** and **ST-25** in piglets with preweaning **diarrhoea** ($n = 91$) is as effective as conventional treatment with drugs. **Choi et al, Am J Chin Med 2003**

In an earlier prospective, controlled, group-comparison trial in piglets ($n = 44$) with induced **enteropathogenic *Escherichia coli*** diarrhoea, traditional acupuncture at **Du-1**, **ST-36**, **BL-20**, **Ren-12** and **ST-25** produced the best results in comparison to the groups treated with electro-acupuncture or antibiotics. There were too many groups given the number of cases. **Hwang et al, Am J Vet Res 1988**

Several animal experiments investigated the effect of acupuncture on signalling and modulating pain. Acupuncture at **Du-1** and **Du-2** inhibited the reaction on **experimental pain stimuli**. **Bing et al, Neuroscience 1991, Xu et al, Zhen Ci Yan Jiu 1989, Xu et al, Zhen Ci Yan Jiu 1989**

Du-2 → Du-1

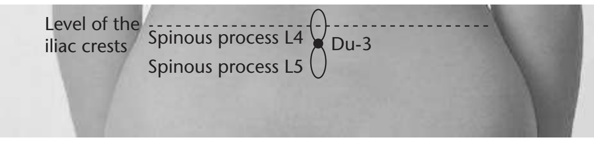
Level of the iliac crests
Spinous process L4
Spinous process L5
Du-3

Du-3 Rats treated with electro-acupuncture at **Du-3**, **BL-54** and **BL-60** 15 minutes after a standardised **spinal cord injury** at T8, showed a marked improvement (regarding both morphology and functioning) three days post-operatively compared to rats not treated with acupuncture. However, none of the beneficial effects occurred in rats given acupuncture treatment 24 hours after spinal cord injury. **Politis et al, Acupunct Electrother Res 1990**

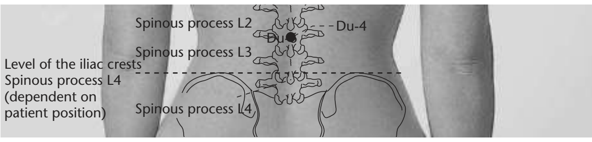
Spinous process L2
Du-4
Du-5
Spinous process L3
Level of the iliac crests
Spinous process L4
(dependent on
patient position)
Spinous process L4

Du-4 A Turkish prospective, randomised, non-blinded, controlled group-comparison trial ($n = 24$) investigated the efficacy of acupressure at **Du-4**, **Du-15**, **Du-20**, **BL-23**, **BL-28**, **BL-32**, **HE-7**, **HE-9**, **ST-36**, **SP-4**, **SP-6**, **SP-12**, **Ren-2**, **Ren-3**, **Ren-6**, **KID-3** and **KID-5** on **enuresis** compared to oxybutynin given to the control group. Acupressure was administered to 12 patients by the parents. While treatments in both groups were successful, complete recovery was significantly higher in the acupressure group (83.3%) compared to the control group (58.3%). **Yuksek et al, J Int Med Res 2003**

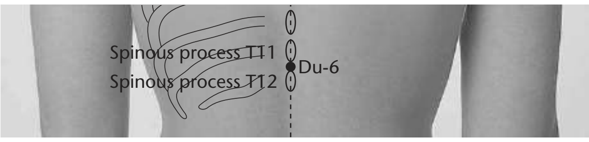
Spinous process T11
Spinous process T12
Du-6

Du-6 An earlier Chinese case study reports the application of **Du-6** for **acupuncture anaesthesia for hysterectomies. Authors' Collective, Chin Med J (Engl) 1978**

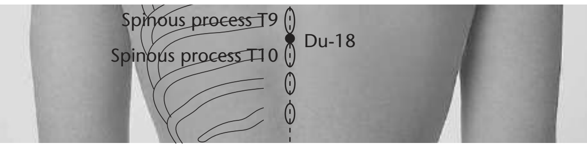
Spinous process T9
Spinous process T10
Du-18

Du-8 An experiment in rats supports the effect of **Du-8** on **seizures** (by releasing melatonin) and on **cerebral infarction** (reduction of the cerebral infarction volume due to release of taurine and decrease of aspartame). **Chao et al, Acupunct Electrother Res 2001, Zhao et al, Acupunct Electrother Res 1997, Yin et al, Zhen Ci Yan Jiu 1994**

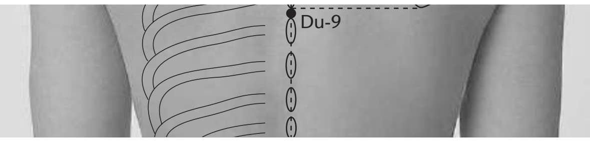

Du-9 The effect of electro-acupuncture at **Du-9** was investigated in 23 patients suffering from **coronary heart disease**. The results showed a slight dilation of the coronary arteries and a slight decrease of the heart rate frequency (monitored by coronary arteriography). **Yan et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1998**

Two earlier Chinese case studies report decreased symptoms in patients with coronary heart disease after acupuncture or tuina at Du-9. Wang et al, Zhong Xi Yi Jie He Za Zhi 1988, Wang et al, Zhong Xi Yi Jie He Za Zhi 1987

A Chinese case study reports the successful treatment of **angina pectoris** by implanting a micro-depressor at **Du-9**. **Wang et al, Zhong Xi Yi Jie He Za Zhi 1988**

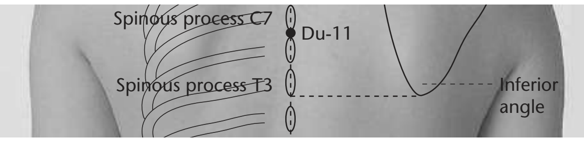
Spinous process C7
Du-11
Spinous process T3
Inferior
angle

Du-11 Electro-acupuncture at **Du-11** and **Du-16** prior to, during and after an experimentally induced **transient ischaemia** in gerbils suggests the protective influence of these points. They seem to suppress both glutamate release and reperfusion injury following the ischaemic insult. **Pang et al, Am J Chin Med 2003**

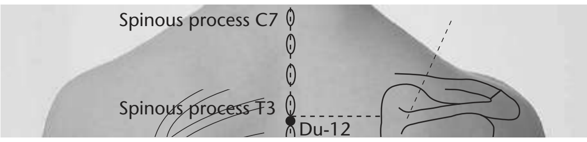
Spinous process C7
Spinous process T3
Du-12

Du-12 A Chinese case study reports that classic acupuncture at **Du-12**, **L.I.-10**, **BL-11** and **S.I.-3** with *deqi* sensation in the direction of the disorder in 55 patients with **cervical vertebra**

disease and **impaired cerebral blood flow capacity** led to a significant improvement of blood flow capacity. **Qie et al, Zhong Xi Yi Jie He Za Zhi 1991**

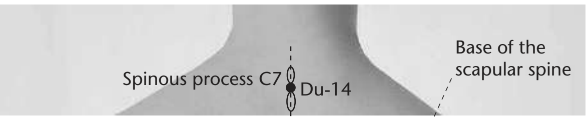

Du-14 → L.I.-4, ST-8; ST-36

A Chinese experimental trial (n = 40) investigated the effect of acupuncture at **Du-14** and **P-6** on driving performance during a three-hour driving simulator test. The acupuncture group showed a significantly better driving performance and a decrease in **driving fatigue**, compared to the untreated control group. **Li et al, Accid Anal Prev 2004**

A Chinese prospective, randomised, controlled, non-blinded trial (n = 40) investigated the effect of acupuncture at **Du-14 and P-6** on heart rate and driving performance during a three-hour driving simulator test. The acupuncture group showed a significantly decreased **activation of the sympathetic nervous system** and increased **parasympathetic activity** compared to the untreated control group. **Li et al, Eur J Appl Physiol 2003**

A Chinese retrospective case study (n = 25) investigated the effect of cupping at **Du-14**, **BL-13**, **BL-20** and **BL-23** on **bronchial asthma** while reducing orally administered cortisone. The author reports that 56% of patients showed a marked improvement after 30 treatments. **Hu, J Tradit Chin Med 1998**

A Chinese prospective case study (n = 50) reports the successful treatment of **perennial allergic rhinitis** by applying 10% *Cantharides* extract at **Du-14** and **P-6**. The results showed an effective rate of 88%. The allergic nasal mucosa provocative test of the treated group improved after the treatment, the number of eosinophils and basophils in nasal secretion decreased and the serum total IgE also reduced significantly. **Tang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1995**

Du-15 → T.B.-8; Du-4

Du-16 → Du-11

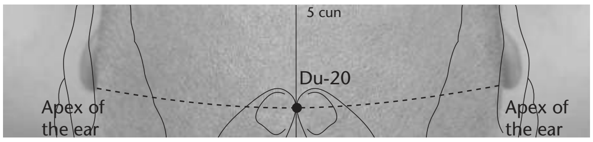

Du-20 → T.B.-8; G.B.-4, ST-8; Du-4; LIV-3; HE-7, L.I.-4

A Chinese prospective randomised controlled single-blinded group-comparison trial (n = 50) investigated the effect of a standardised acupuncture protocol, used either on its own or accompanied by needling of **Du-20**, **Du-26** and **HE-7** (either singly or in combination) on **vascular dementia** in hemiplegic patients. Clinical symptoms were observed and scales such as HDS-R, ADL and functional activities questionnaire (FAQ) used for assessment. The group additionally treated with **Du-20 and HE-7** showed marked improvement of memory, orientation and reaction. At the same time mental rigidity and trance decreased. **Du-20** on its own improved understanding, calculating and social adaptation, while **Du-26** was helpful in treating mental retardation, failure to perform daily activities, trance and poor memory. The combination of the three points showed overall the best results in improving intelligence and social adaptation. The particular design of this trial – combining a standard protocol used in both groups with the actually relevant points in the verum group – presents an interesting model for eliminating persistent problems regarding placebo methods and blinding. However, in this particular trial the number of participants (5 groups of 10 subjects) was too low to determine if the results were statistically significant. **Lai et Huang, Chin J Integr Med 2005**

A German prospective, non-blinded, pair-matched, controlled trial (n = 121 pregnant women between 30 and 39 weeks of gestation) investigated the effect of acupuncture at **Du-20** and **ST-36** on **fetal and maternal cardiotocographic parameters and maternal circulation**. The main difference was a transient increase of the Fisher score in the treatment group compared to the control group, with a persistent increase of the Fisher score. **Scharf et al, Z Geburtshilfe Neonatol 2003**

A Croatian prospective, randomised, placebo-controlled trial (n = 57) investigated the effectiveness of **Du-20**, **L.I.-4**, **Ren-3**, **Ren-4**, **Ren-6**, **G.B.-34**, **BL-23**, **SP-6** and **HE-7** on **primary dysmenorrhoea** compared to placebo-acupuncture. Subjects were observed for a period of two years. After one year, 93.3% of acupuncture patients were symptom-free and/or not taking medication compared to only 3.7% of the placebo patients. **Habek et al, Gynakol Geburtshilfliche Rundsch 2003**

A Polish prospective, randomised, controlled, non-blinded trial (n = 69) investigated the effect of acupuncture at **Du-20**, **Ren-12**, **ST-36**, **ST-21**, **ST-25**, **LIV-3**, **LIV-13**, **P-6** and **HE-7** in addition to a low calorie diet in the treatment of obesity (BMI approximately 33 kg/m^2). This diet consisted of a daily intake of about 1200–1500 kcal, which was about 1000 kcal below the daily energy requirement of the patients. The acupuncture group showed a significantly better result after only 12 treatments over a period of 6 weeks. The mean body weight decreased from 85.5 kg to 72.7 kg, while it only decreased to 77.1 kg in the diet group. **Wozniak et al, Ginekol Pol 2003**

A German prospective, randomised, placebo-controlled, double-blinded trial (n = 56) investigated the effect of acupuncture at **Du-20**, **Ex-HN-6**, **HE-7**, **P-6** and **BL-62** in the treatment of **depression** and **generalised anxiety**. Results were assessed using questionnaires. After 10 treatments, the verum acupuncture group showed a significantly larger clinical improvement compared to the placebo group (61% vs 21%). **Eich et al, Fortschr Neurol Psychiatr 2000**

Du-26 → G.B.-4; Du-20

Ren mai (conception vessel)

Ren-2 → Du-4

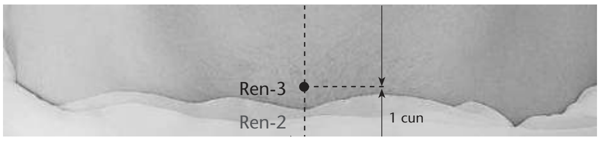

Ren-3 → Du-4; Du-20

A Taiwanese prospective, randomised, controlled trial ($n = 60$) investigated the influence of electro-acupuncture (EA) at **Ren-3**, **Ren-4** and **BL-32** on the **neurogenic bladder of spinal cord injured patients** compared to the conventional bladder training programme received by the control group. The acupuncture group took significantly less time to achieve balanced voiding than the control group (on average 28 days less). The best results were achieved in patients who received acupuncture within three weeks after injury (39 days less). **Cheng et al, Spinal Cord 1998**

Ren-4 → Du-20; ST-25; Ren-3

Ren-6 → Du-4; Du-20; Ren-17

Ren-8 → P-6

A Chinese prospective, randomised, controlled, non-blinded trial ($n = 56$) investigated the effect of moxibustion at **Ren-8** in addition to chemotherapy for **nasopharyngeal carcinoma** (stage III and IV). There were significantly fewer toxic and side effects in the moxibustion group compared to the control group. The five-year survival rate in the two groups were 50.0% and 35.7% respectively. **Chen et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2000**

A Chinese prospective, randomised, controlled trial ($n = 140$) investigated the effect of applying a topical ointment at **Ren-8** on **epigastric pain**. Results in the treatment group were significantly better than in the control group (94% vs 53%). **Ba et al, J Tradit Chin Med 1999**

Ren-12 → L.I.-4; Du-1; Du-20; LIV-3; Ren-17

Ren-13 → P-6

Ren-14 → LIV-14

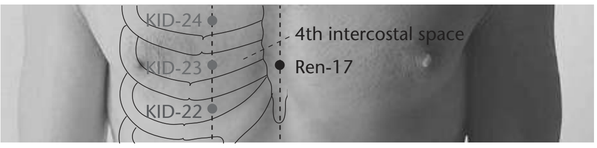

Ren-17 A Chinese prospective, randomised, controlled trial ($n = 60$) investigated the effect of **Ren-17**, **Ren-12**, **Ren-6**, **ST-36 and SP-10** on the symptoms of **vascular dementia** compared to conventional treatment. Both acupuncture and conventional treatment showed a highly significant increase of MMSE, HDS-R and ADL scores. However, the results of the acupuncture group were statistically more significant regarding MMSE, HDS-R and the total rate of effectiveness (80% vs 47%). **Yu et al, Neurol Res 2006**

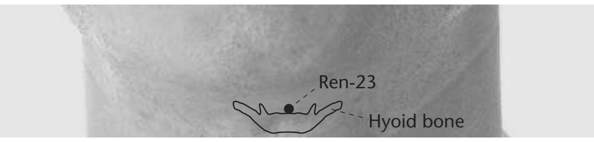

Ren-23 A Chinese retrospective case study ($n = 120$) reports the successful treatment of **pseudobulbar palsy** with acupuncture at **Ren-23** and **LU-5. Wang et al, J Tradit Chin Med 1998**

Ren-24 A German prospective, single-blinded, controlled trial ($n = 41$) reports that acupuncture at **Ren-24** significantly reduced the severity of the **gag reflex** during transoesophageal echocardiography compared to acupuncture at a sham point. **Rosler et al, J Altern Complement Med 2003**

Extra points

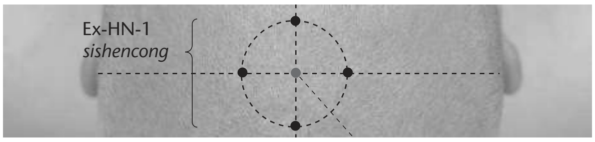

Ex-HN-1 → BL-7

A Taiwanese, prospective, randomised, placebo-controlled, single-blinded experimental trial ($n = 9$) investigated the effect of the *sishencong*-points (**Ex-HN-1**) on the variance of heart rate frequency. The authors observed an increase in the vagal activities but a suppression of the sympathetic activity. **Wang et al, Auton Neurosci 2002**

A Chinese case study ($n = 51$) reports the successful treatment of **hyperthyreosis** with acupuncture at Ex-HN-1. **Xie et al, J Tradit Chin Med 1994**

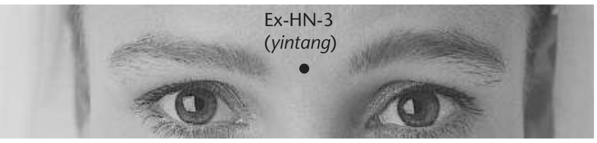

Ex-HN-3 An Austrian prospective, randomised, single-blinded experimental cross-over study ($n = 25$) showed that acupuncture (but not laser acupuncture) at *yintang* monitored by EEG had a sedating effect. Clinical application of this point, for example for **sleeping disorders**, requires further investigation. **Litscher et al, Eur J Anaesthesiol 2004**

Ex-HN-4 → S.I.-18; L.I.-4

Ex-HN-5 → G.B.-5; LIV-3

Ex-HN-6 → HE-7; Du-20

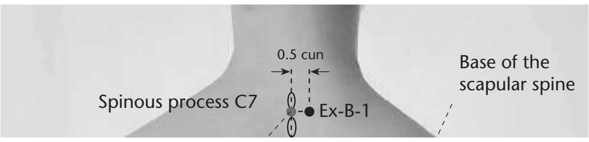

Ex-B-1 A Chinese prospective, randomised, controlled trial ($n = 84$) investigated the effect of a herbal plaster at **Ex-B-1**, **BL-11**, **BL-12**, **BL-13** and **BL-15** on **chronic bronchitis** compared to intramuscular injection therapy. In the plaster group the clinical total effective rate was 93.2% immediately after the treatment course and 91.5% after 18 months (control group: 80% and 80%). X-rays of the thorax showed the effectiveness of the plaster as 40.7% and 89.8% after 18 months (control group: 20% and 76%). **Huo et al, Zhongguo Zhong Xi Yi Jie He za Zhi 2001**

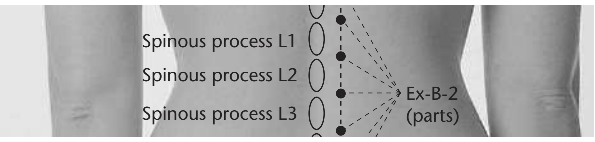

Ex-B-2 A Chinese case study ($n = 43$) reports the positive effects of treating **ulcerative colitis** with acupuncture at **Ex-B-2** as well as stimulation with a plum-blossom needle at **BL-22** and **BL-25. Yue and Zhenhui, J Tradit Chin Med 2005**

A Chinese retrospective case study ($n = 168$) reports the successful treatment of **sciatica** with acupuncture at the **Ex-B-2** points. **Pei et al, J Tradit Chin Med 1994**

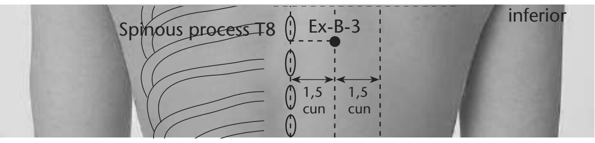

Ex-B-3 An experiment in diabetic rabbits demonstrated that electro-acupuncture at **Ex-B-3** decreased plasma glucose levels and inhibited the release of pancreatic glucagon. The effect was increased by adding **ST-36**. This latter point led to no significant changes when needled alone. **Zeng et al, J Tradit Chin Med 2002**

Ex-UE-9 → L.I.-4

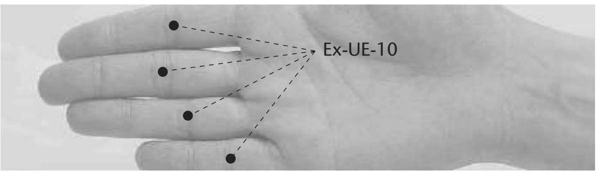

Ex-UE-10 A case study describes the successful treatment of **hiccups** with acupuncture at **Ex-UE-10**. **Qi, J Tradit Chin Med 1993**

Ex-UE (***JIANNEILING***) → L.I.-15

Ex-UE (***JIANQIAN***) → L.I.-4

Ex-LE-4 → ST-35

Ex-UE (***ZHONGPING***)

A Chinese retrospective case study ($n = 115$) reports the successful treatment of **shoulder and back pain** with acupuncture at ***zhongping***. The point *zhongping* is located midway between ST-36 and ST-37. **Wang et al, J Tradit Chin Med 1995**

9.3 Summary

The data presented clearly demonstrates that several actions of a number of acupuncture points are undoubtedly effective. These points (for example P-6) were tested according to the highest scientific standards, comparing standard acupuncture to deep acupuncture at non-acupuncture points (a methodology considered the most superior acupuncture placebo).

This proof is even more important since the results of the two largest acupuncture trials to date, ART (Acupuncture Randomized Trials) and GERAC (German Acupuncture Trial), may have led to the impression that, while acupuncture is indeed effective, its effect is independent of the location of the acupuncture points. The hitherto unsurpassed size of the trials and the undoubtedly intelligent and biomathematically impeccable design (especially of the GERAC trial) further contribute to this impression. Due to their flexible point prescriptions and their complexity, these studies have not been included in section 9.2.

But what are the basic results of these trials? (up to May 2006: **Witt et al, Lancet 2005; Linde et al, JAMA 2005; Melchart et al, Br Med J 2005; Brinkhaus, Arch Intern Med 2006**; **Diener et al, Lancet Neurol 2006**)

For both **gonarthrosis** and **lower back pain,** acupuncture had significantly better results than the untreated waitlist (ART) or conventional therapy (GERAC). For **migraine**, acupuncture is significantly better than the untreated waitlist (ART), and treatments with acupuncture for only six weeks had the same results as treatments with conventional medicine for six months (GERAC). For **tension headaches**, acupuncture provided significantly better results than the waitlist (ART). A comparison between acupuncture and conventional therapy in the treatment of tension headaches within the framework of GERAC was not possible, since all but two patients refused the stigmatised conventional treatment with amitriptyline due to the high incidence

Johannes W. Rohen
Chihiro Yokochi
Elke Lütjen-Drecoll

Color Atlas of Anatomy

FOURTH EDITION

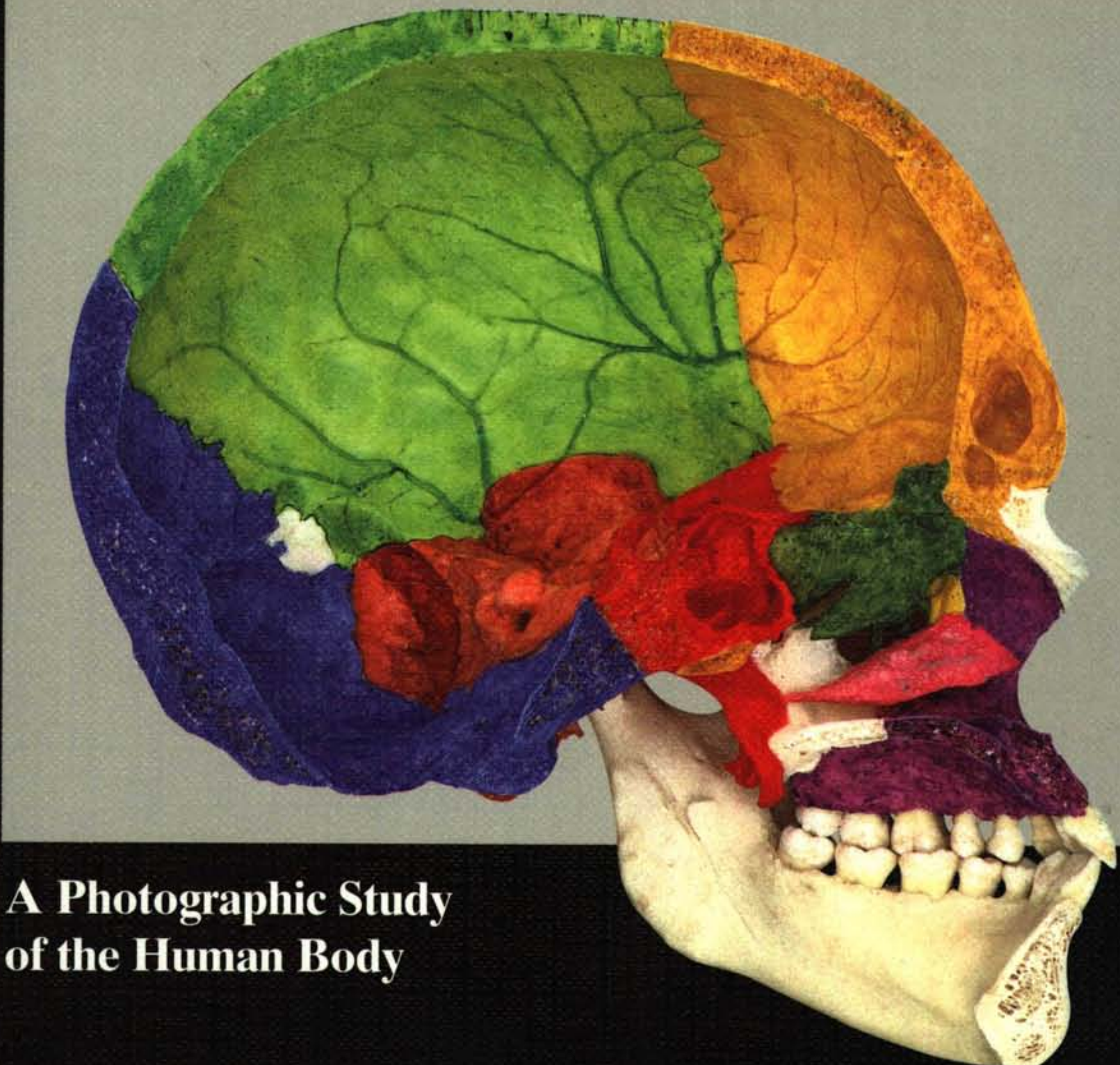

A Photographic Study
of the Human Body

In contrast to most other mammals the human body is adapted for bipedal locomotion. Three general principles in the architecture of the human organism are recognizable:

1. The principle of **segmentation,** which dominates in the trunk. The vertebral column and the thorax consist of relatively equal, segmentally arranged elements.
2. The principle of **bilateral symmetry.** Both sides of the body are separated by a midsagittal plane and resemble each other like image and mirror-image.
3. The principle of **polarity** between the head at one end of the body and the lower extremities at the other. As the center of the information system the head contains the main sensory organs and the brain. The head has a predominantly spherical form while the extremities consist of radially formed skeletal elements, the number of which increases distally.

A. The **skull** consists of two parts: 1. a **cranial part** containing mainly the brain and the sensory organs and 2. a **facial part** which contains the nasal and oral cavity and the chewing apparatus. The cranial cavity is continuous with the vertebral canal which contains the spinal cord.

B. The **thorax** contains the respiratory and circulatory organs (lung, heart, etc.) but also some of the abdominal organs which are located underneath the diaphragm.

C. The **abdominal cavity** contains the organs of metabolism such as the liver, the stomach and the intestinal tract as well as the excretory and genital organs (kidney, uterus, urinary bladder, etc.). The latter are located primarily in the **pelvic cavity** with the exception of the testes.

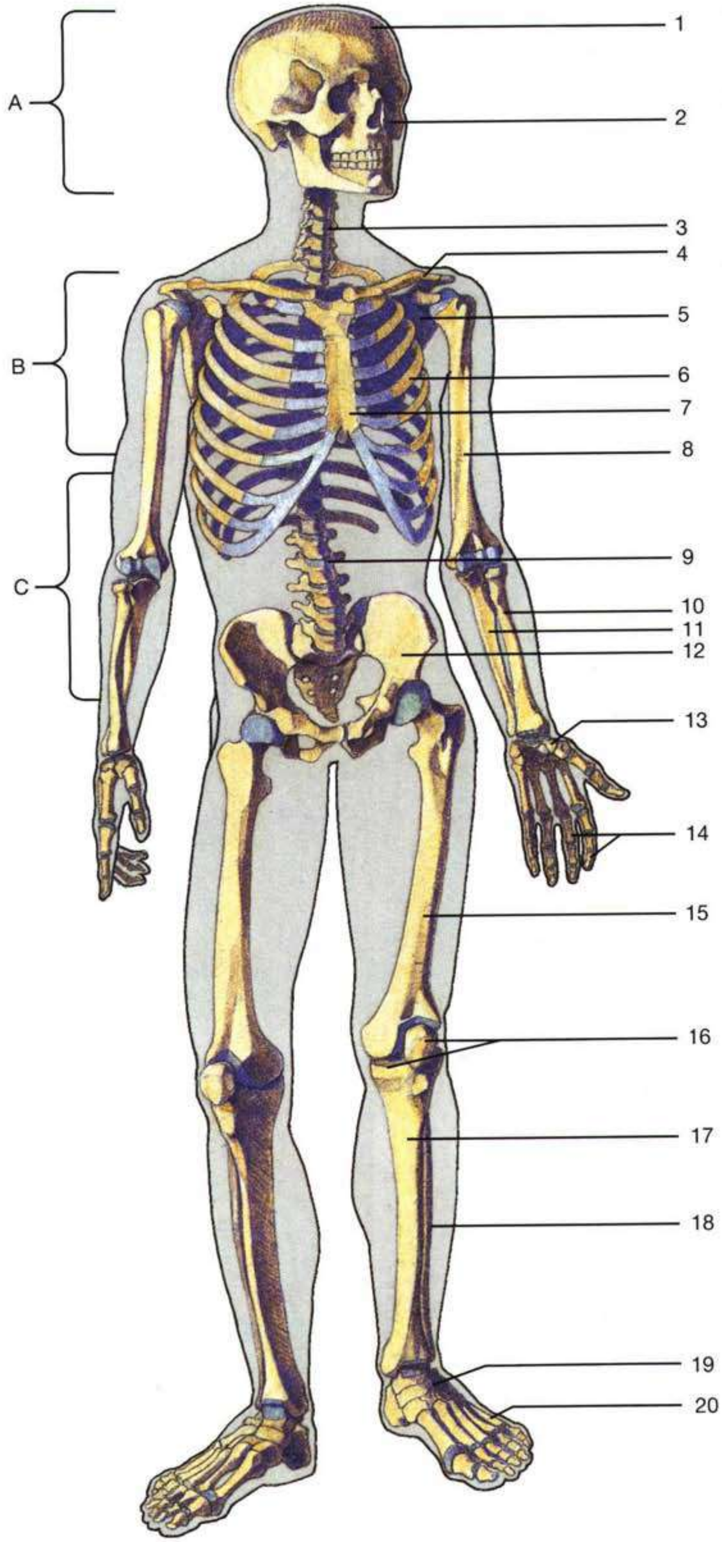

Structure of the human body and the skeleton. Blue = joints.
A Head (caput) B Thorax (thoracic cavity) C Abdominal and pelvic cavities

1	Cranial part	of the skull
2	Facial part	
3	Vertebral column (cervical part)	
4	Clavicle	
5	Scapula	
6	Ribs	
7	Sternum	
8	Arm (humerus)	
9	Vertebral column (lumbar part)	
10	Radius	forearm
11	Ulna	
12	Pelvis	
13	Wrist (carpals)	hand
14	Fingers (phalanges)	
15	Thigh (femur)	
16	Patella and knee joint	
17	Tibia	leg
18	Fibula	
19	Tarsals	foot
20	Metatarsals	

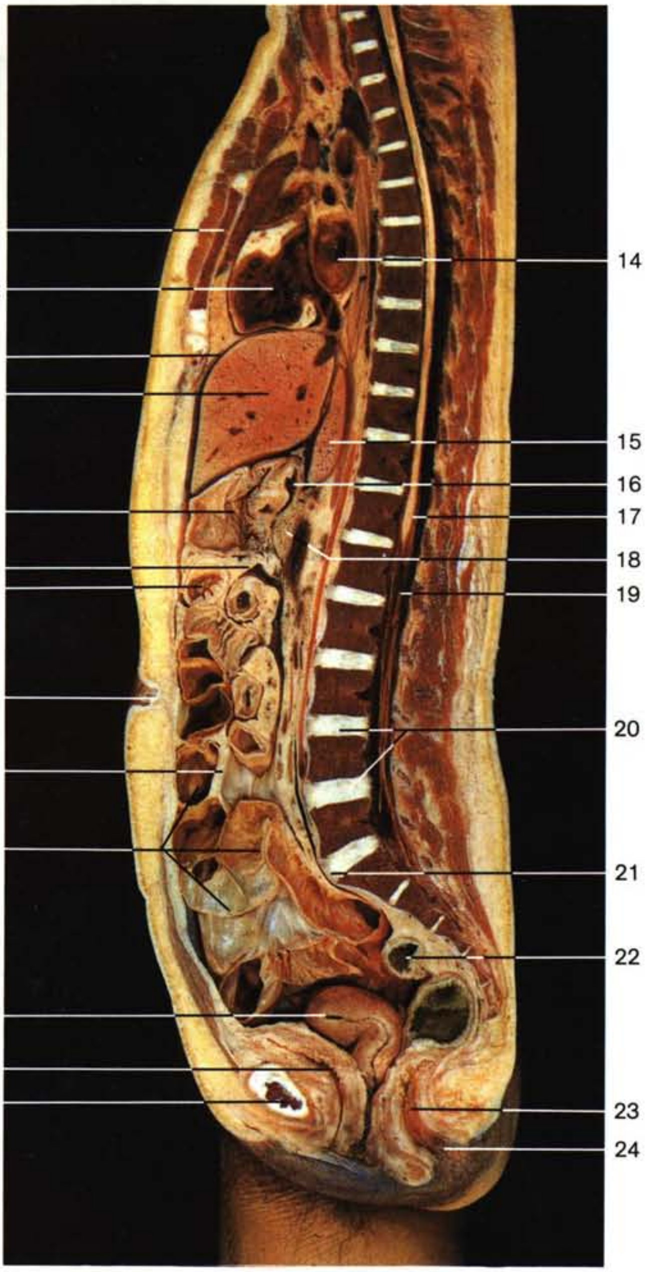
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Median section through the trunk (female).

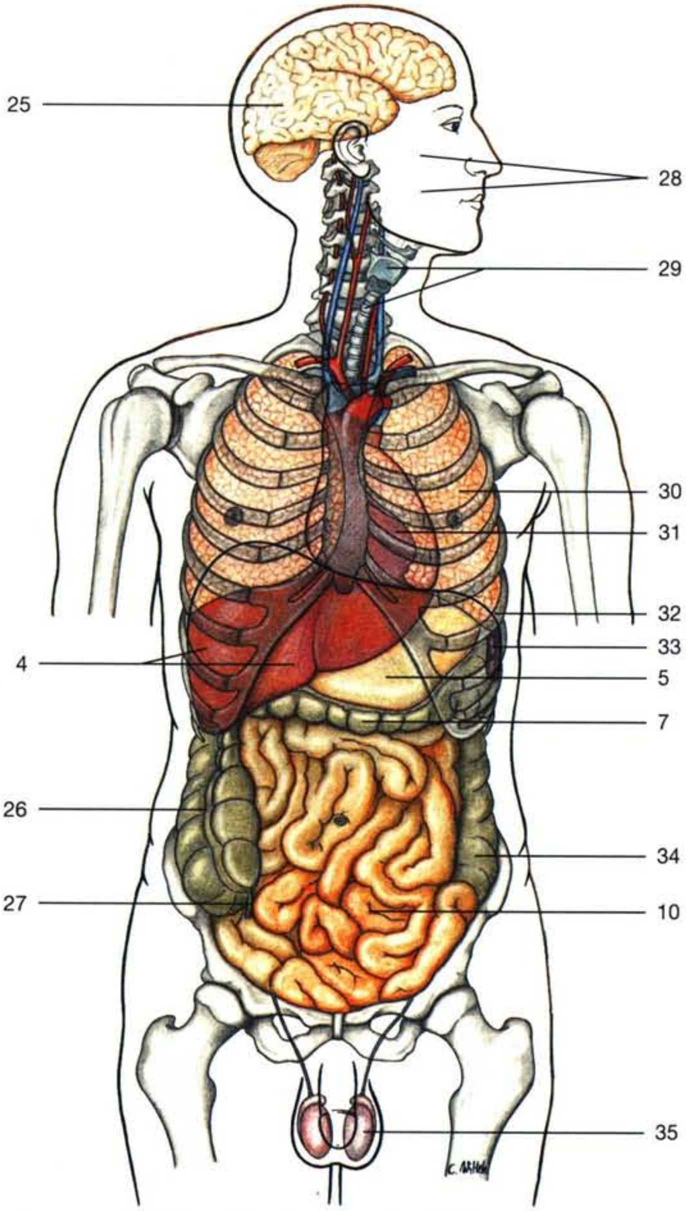
25
28
29
30
31
32
33
4
5
7
26
34
27
10
35

Positon of the inner organs of the human body (anterior aspect)
The main cavities of the body and their contents.

1 Sternum
2 Right ventricle of heart
3 Diaphragma
4 **Liver**
5 **Stomach**
6 Transverse mesocolon
7 Transverse colon
8 Umbilicus
9 Mesentery
10 **Small intestine**
11 Uterus
12 Urinary bladder
13 Pubic symphysis
14 Left atrium of heart
15 Caudate lobe of liver
16 Omental bursa or lesser sac
17 Conus medullaris
18 **Pancreas**
19 Cauda equina
20 Intervertebral discs (lumbar vertebral column)
21 Sacral promontory
22 Sigmoid colon
23 Anal canal
24 Anus
25 Head (neurocranium) with brain
26 Ascending colon
27 Appendix
28 Facial region (viscerocranium) with oral and nasal cavities
29 Trachea and larynx
30 Thorax with the lungs
31 **Heart**
32 Surface projection of the diaphragm
33 **Spleen**
34 Descending colon
35 Testis

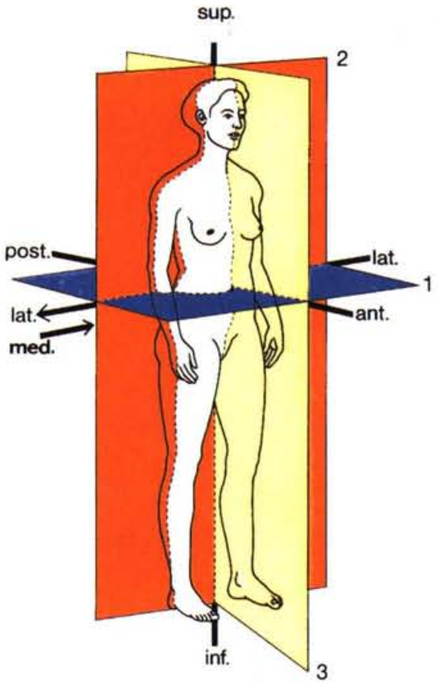
sup.
2
post.
lat.
1
lat.
ant.
med.
inf.
3

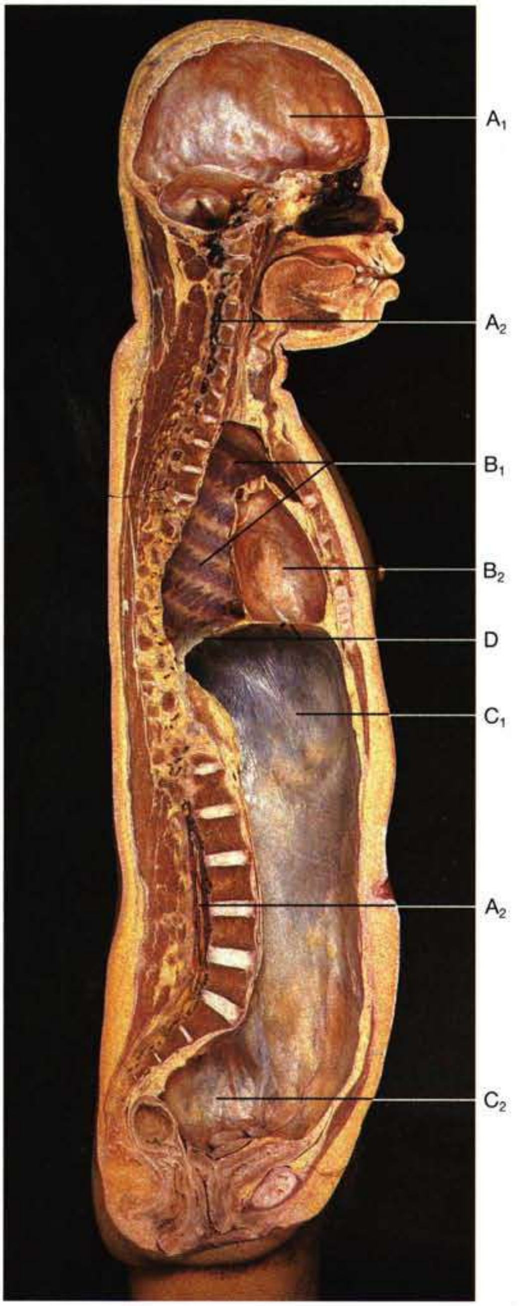

Sagittal section through the human body (female). Demonstration of the main cavities of the body. Internal organs are removed.

A_1 Cranial cavity
A_2 Vertebral canal
B_1 Thoracic cavity
B_2 Pericardial cavity
C_1 Abdominal cavity
C_2 Pelvic cavity
D Diaphragm

◁

Planes of the body

1 Transverse plane
2 Frontal plane
3 Sagittal plane (midsagittal)

Lines of direction

ant. = anterior
inf. = inferior
lat. = lateral
med. = medial
post. = posterior
sup. = superior

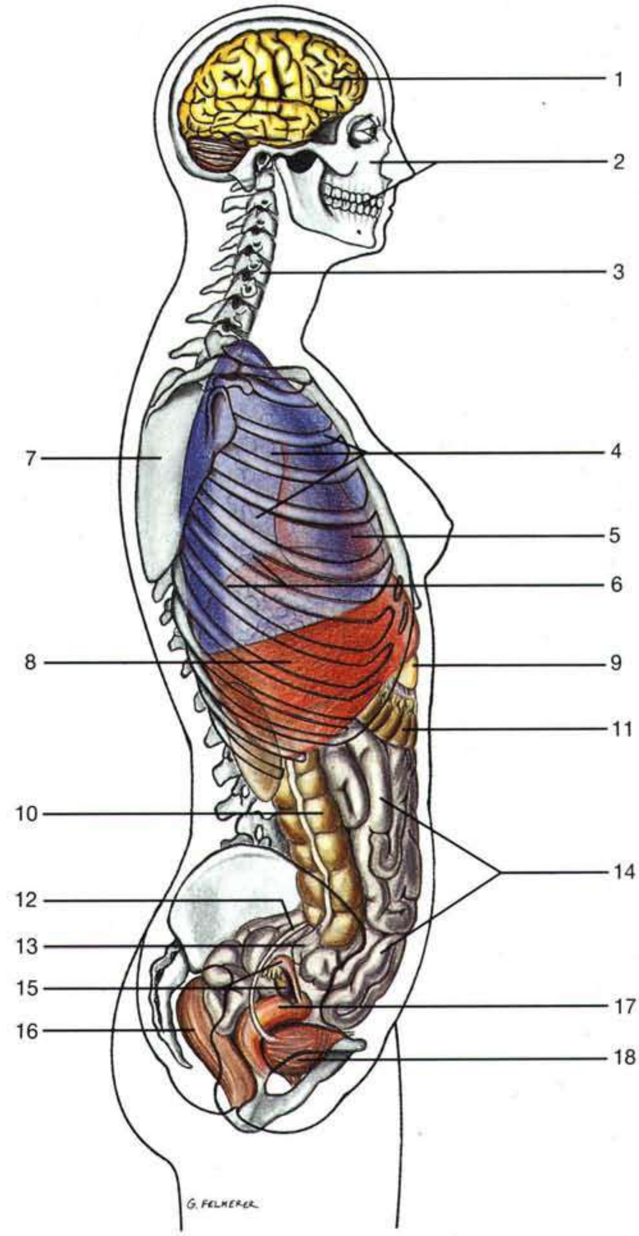

Position of the inner organs of the human body (lateral aspect).
The three main cavities of the body and their contents.

1 Head (neurocranium) with the brain
2 Facial bones with oral and nasal cavities
3 Vertebral column (cervical part)
4 Thorax with the lungs
5 Heart
6 Surface projection of the diaphragm
7 Scapula
8 Liver
9 Stomach
10 Ascending colon
11 Transverse colon
12 Ureter
13 Appendix
14 Small intestine
15 Ovary, uterine tube
16 Rectum
17 Uterus
18 Urinary bladder

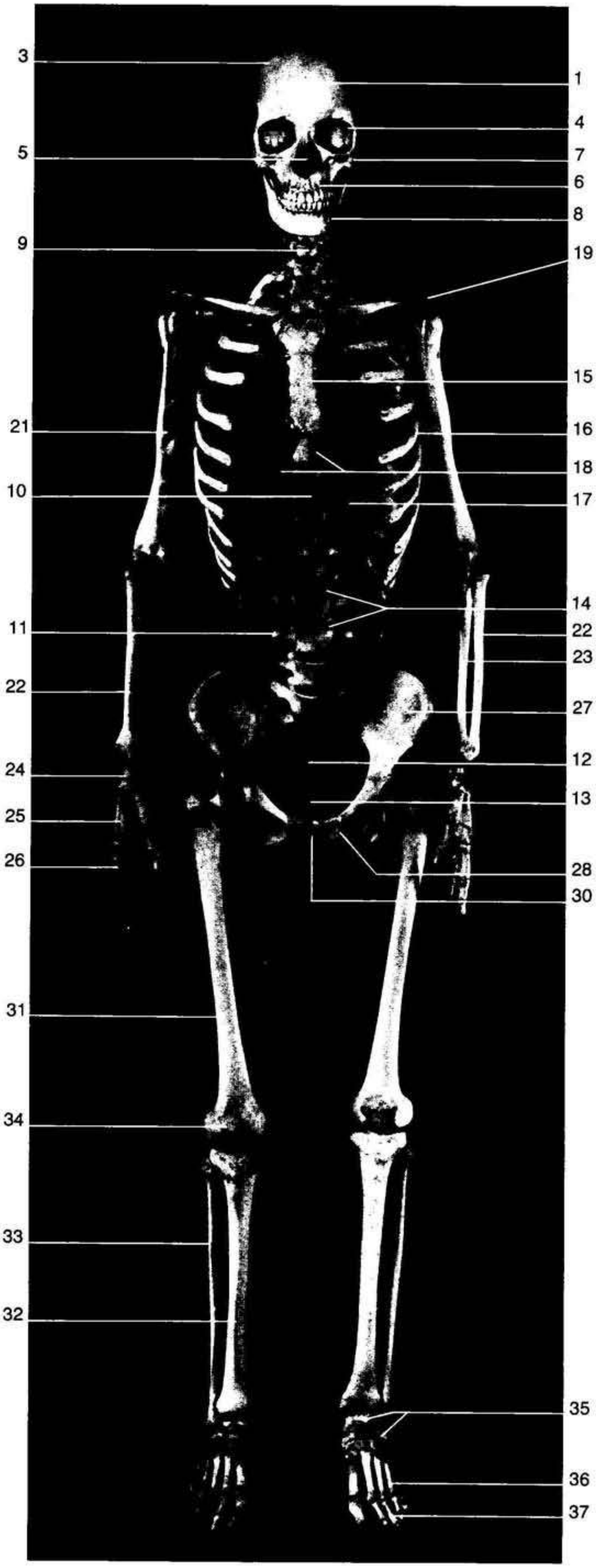
3
1
4
5
7
6
8
9
19
15
21
16
18
10
17
14
11
22
23
22
27
24
12
25
13
26
28
30
31
34
33
32
35
36
37

Skeleton of a female adult (anterior aspect).

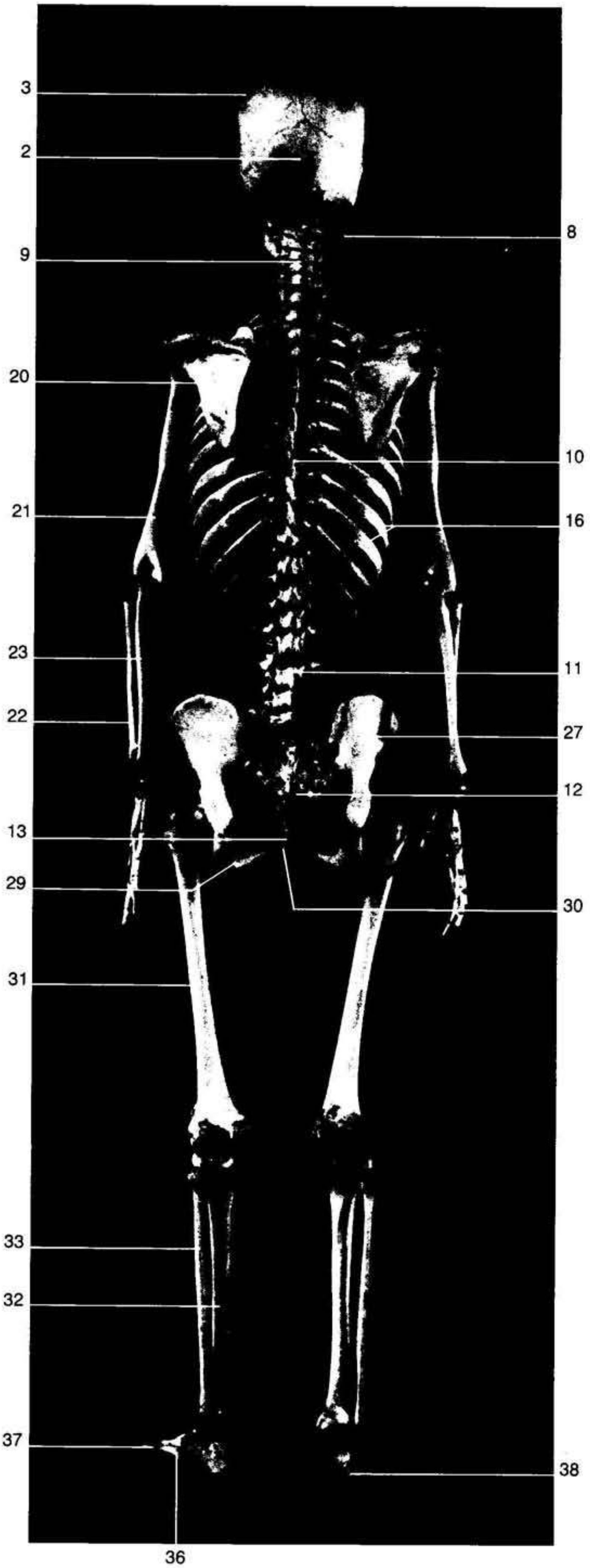
3
2
8
9
20
10
21
16
23
11
22
27
12
13
29
30
31
33
32
37
38
36

Skeleton of a female adult (posterior aspect).

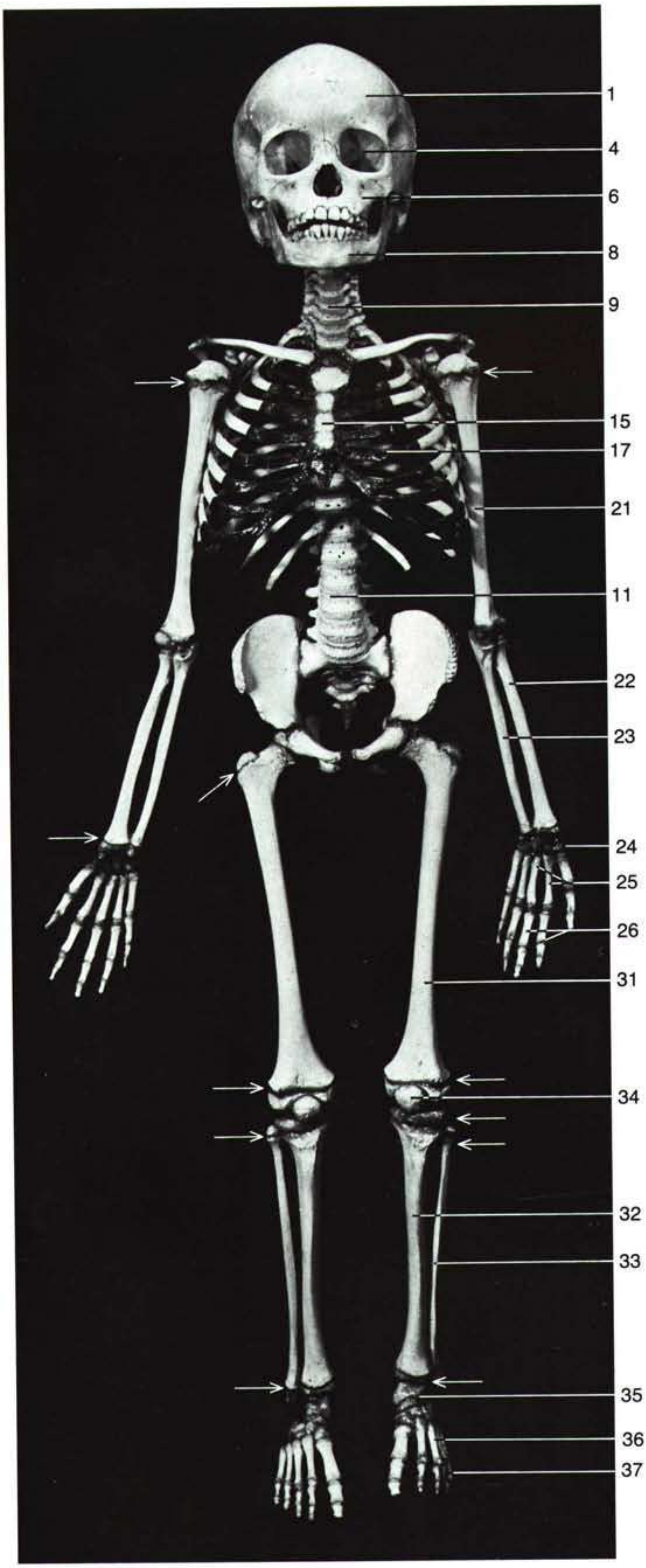

Skeleton of a 5-year-old child (anterior aspect).
The zones of the cartilaginous growth plates are seen (arrows). In contrast to the adult, the ribs show a predominantly horizontal position.

Axial skeleton

Head

1 Frontal bone
2 Occipital bone
3 Parietal bone
4 Orbit
5 Nasal cavity
6 Maxilla
7 Zygomatic bone
8 Mandible

Trunk and thorax

Vertebral column

9 Cervical vertebrae
10 Thoracic vertebrae
11 Lumbar vertebrae
12 Sacrum
13 Coccyx
14 Intervertebral discs

Thorax

15 Sternum
16 Ribs
17 Costal cartilage
18 Infrasternal angle

Appendicular skeleton

Upper limb and shoulder girdle

19 Clavicle
20 Scapula
21 Humerus
22 Radius
23 Ulna
24 Carpal bones
25 Metacarpal bones
26 Phalanges of the hand

Lower limb and pelvis

27 Ilium
28 Pubis
29 Ischium
30 Symphysis pubis
31 Femur
32 Tibia
33 Fibula
34 Patella
35 Tarsal bones
36 Metatarsal bones
37 Phalanges of the foot
38 Calcaneus

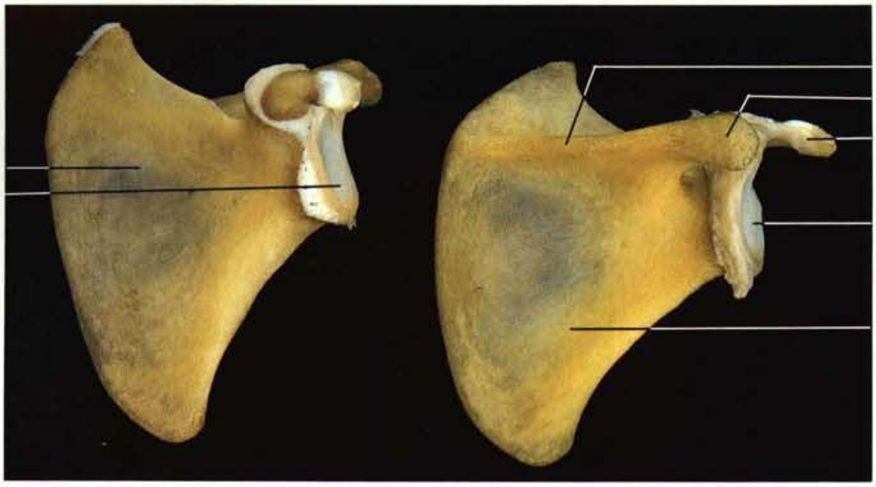

1 Subscapular fossa
2 Coracoid process
3 Glenoid fossa
4 Acromion
5 Spine of scapula
6 Infraspinous fossa

Ossification of the scapula
(left: anterior aspect, right: posterior aspect).

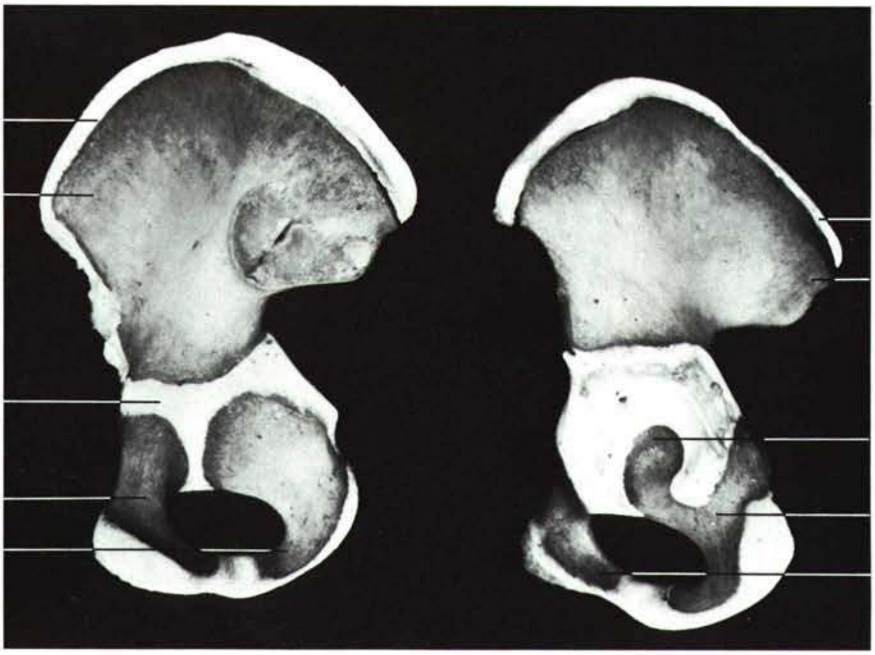

1 Cartilage of the iliac crest
2 Ilium
3 Cartilage
4 Pubis
5 Ischium
6 Acetabulum

Ossification of the hip bone
(left: medial aspect, right: lateral aspect).

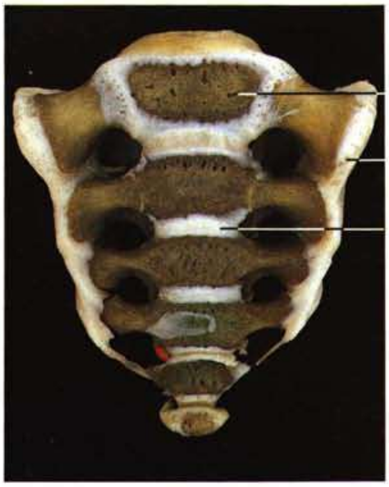

1 Bone tissue (vertebral body)
2 Cartilaginous tissue (lateral epiphysis)
3 Intervertebral discs

Ossification of the sacrum (anterior aspect).
Note the five vertebral bones, which are still separated from each other.

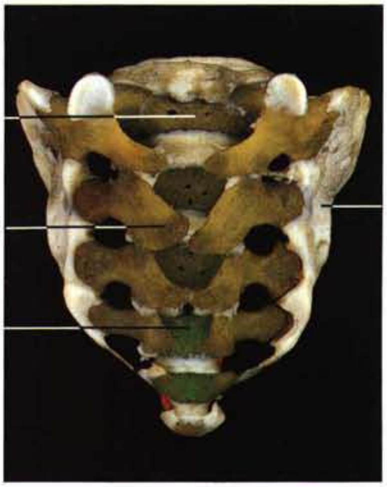

1 Bony tissue (center of ossification)
2 Vertebral arch (not completely united)
3 Cartilaginous tissue (lateral epiphysis)
4 Sacral canal

Ossification of the sacrum
(posterior aspect).

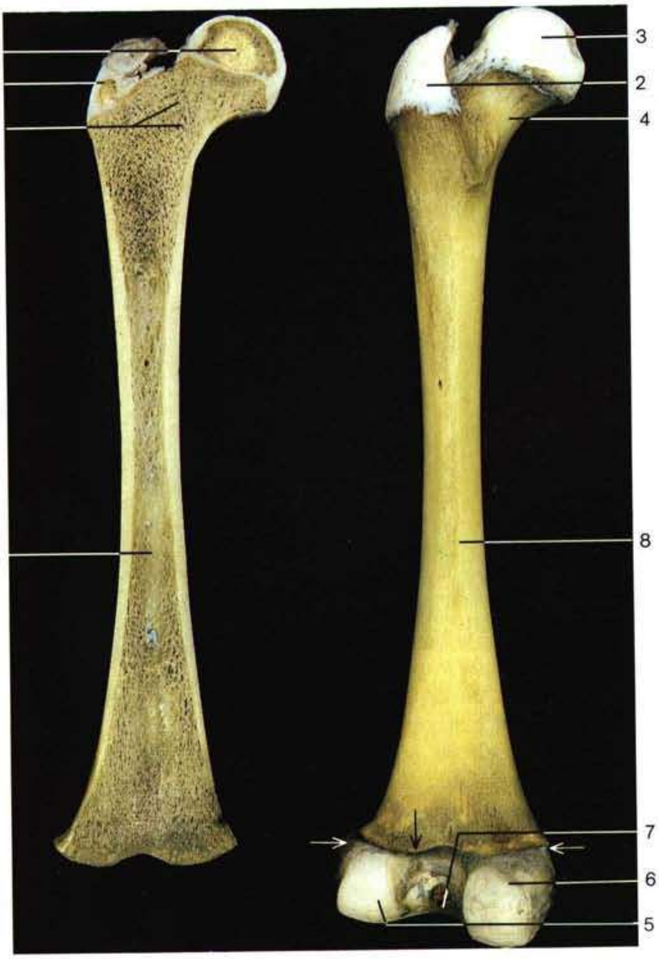

Ossification of the femur (left: coronal section, right: posterior view of the femur). Arrows: distal epiphysis.

1 Ossification center in the head of the femur
2 Greater trochanter
3 Head of the femur
4 Neck of the femur
5 Lateral condyle
6 Medial condyle
7 Intercondylar notch
8 Diaphysis

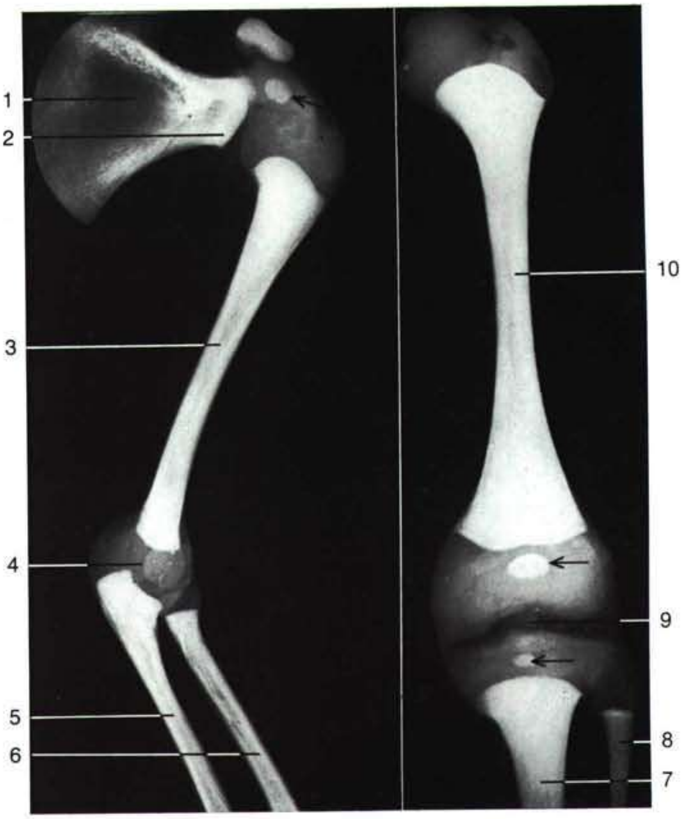

X-ray of the upper and lower limb of a newborn child.
Left: upper limb. Right: lower limb.
Arrows: ossification centers.

1 Scapula
2 Shoulder joint
3 Humerus
4 Elbow joint
5 Ulna
6 Radius
7 Tibia
8 Fibula
9 Knee joint
10 Femur

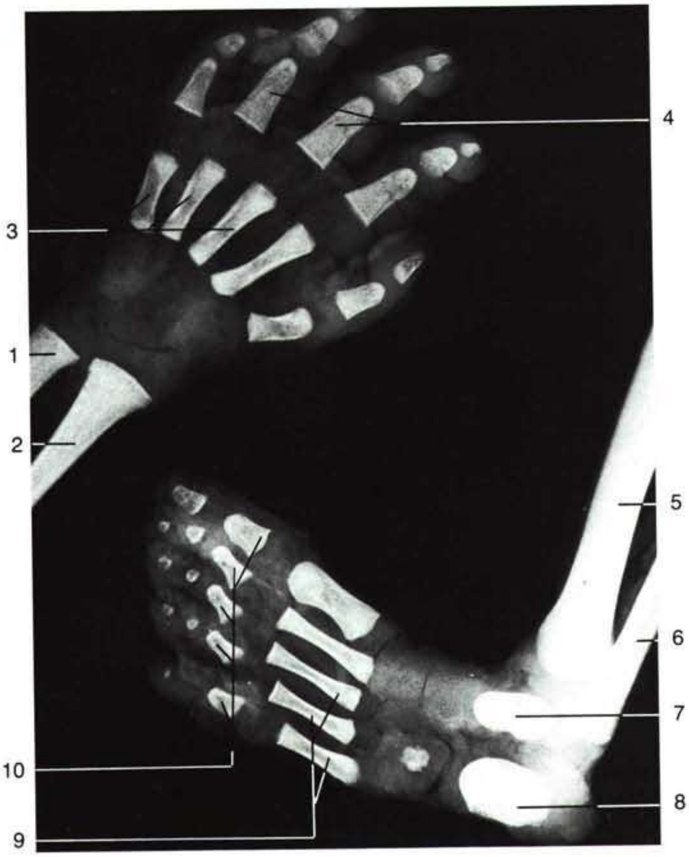

X-ray of hand and foot of a newborn.

1 Ulna
2 Radius
3 Metacarpals
4 Phalanges
5 Tibia
6 Fibula
7 Talus
8 Calcaneus
9 Metatarsals
10 Phalanges

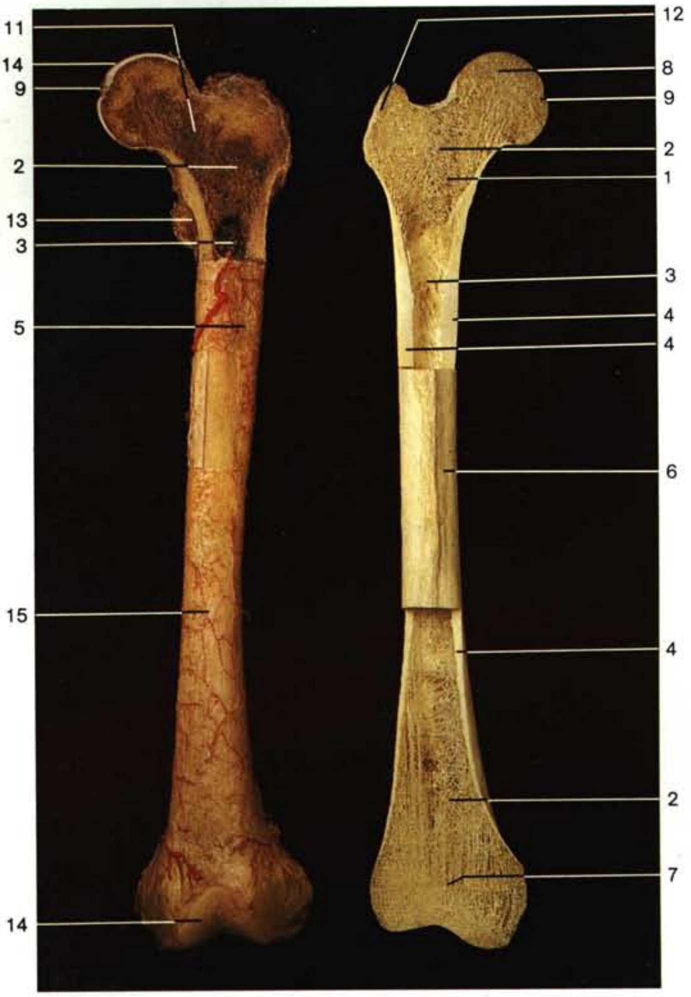

Femur of the adult. Left: the periosteum and the nutrient vessels are preserved. Right: coronal section of the proximal and distal epiphyses to display the spongy bone and the medullary cavity.

1 Metaphysis
2 Spongy bone
3 Medullary cavity in the diaphysis
4 Compact bone
5 Nutrient canal
6 Diaphysis
7 Epiphyseal line (remnants of the epiphyseal plate)
8 Epiphysis (head of the femur)
9 Fovea of head
10 Trabeculae of spongy bone
11 Neck of the femur
12 Greater trochanter
13 Lesser trochanter
14 Articular surface
15 Periosteum
16 Skin
17 Vastus medialis muscle
18 Sartorius muscle
19 Femoral artery and vein
20 Great saphenous vein
21 Gracilis muscle
22 Adductor longus muscle
23 Adductor magnus muscle
24 Semimembranosus muscle
25 Semitendinosus muscle
26 Rectus femoris muscle
27 Vastus lateralis muscle
28 Femur and medullary cavity
29 Vastus intermedius muscle
30 Sciatic nerve
31 Biceps femoris muscle
32 Spongy bone trabeculae containing bone marrow
33 Compact bone
34 Osteon with Haversian lamellae
35 Periosteum
36 Blood vessels and nerves for periosteum and bone

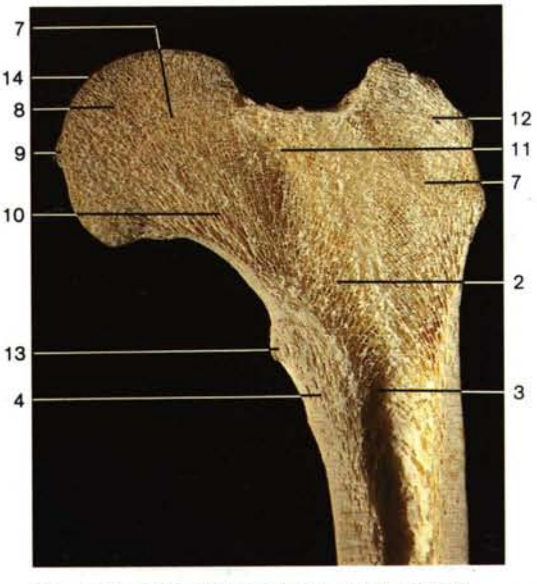

Coronal section through the proximal end of the adult femur, revealing the characteristic trajectorial structure of the spongy bone.

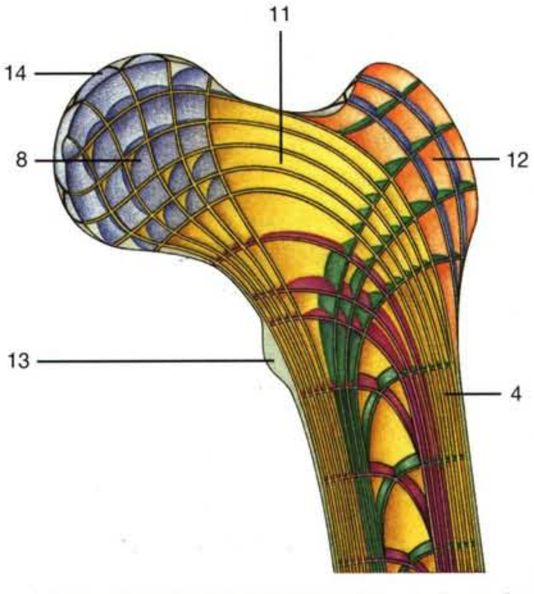

Three dimensional representation on the trajectorial lines of the femoral head (according to B. Kummer).

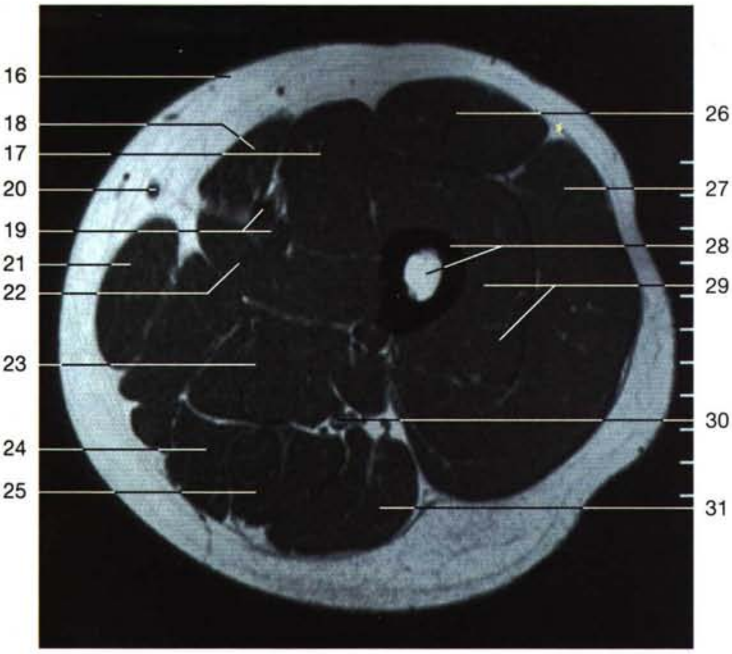

MR-Image of the thigh (axial section through the middle of the left thigh, the same level as the CT-Image).

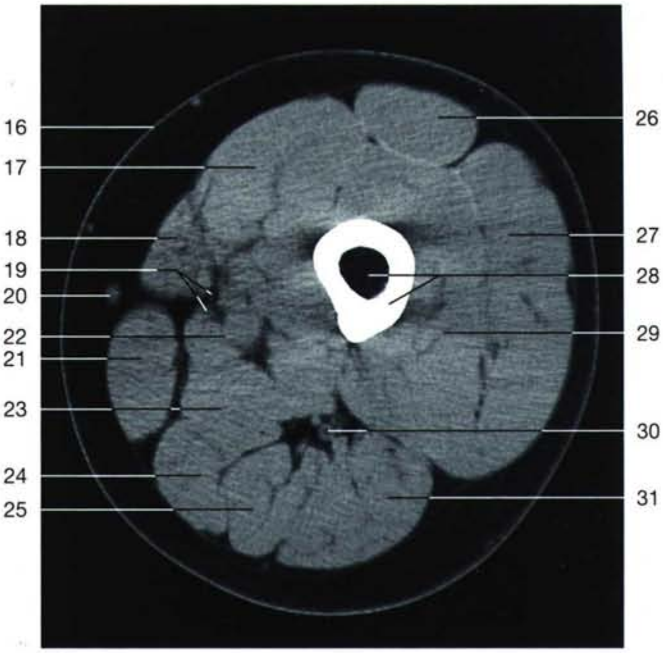

CT-Image of the section through the middle of the left thigh (axial section). Note the differences between the CT- and MR-Image (see text below).

The **bones** of the skeletal system consist of two different parts, the spongy and the compact bone. The spongy bone trabeculae are highly adapted to mechanical forces revealing a trajectorial structure. The intertrabecular spaces are filled with bone marrow, the site of blood formation. The appearance of bones, muscles and soft tisssues is quite different in **CT-** and **MR-Images**. The CT-images relate well to radiographs in that areas of great absorption such as bones are white, and those with little absorption such as fat appear black. In contrast, the intensity of signals in MR-Images, obtained without X-rays but by magnetic forces, is different so that dense areas of bones appear black and soft tissues such as bone marrow and fat appear white (for comparison see above figures).

A highly innervated **periosteum** is an essential structure for bone nutrition, blood supply, growth and bone repair.

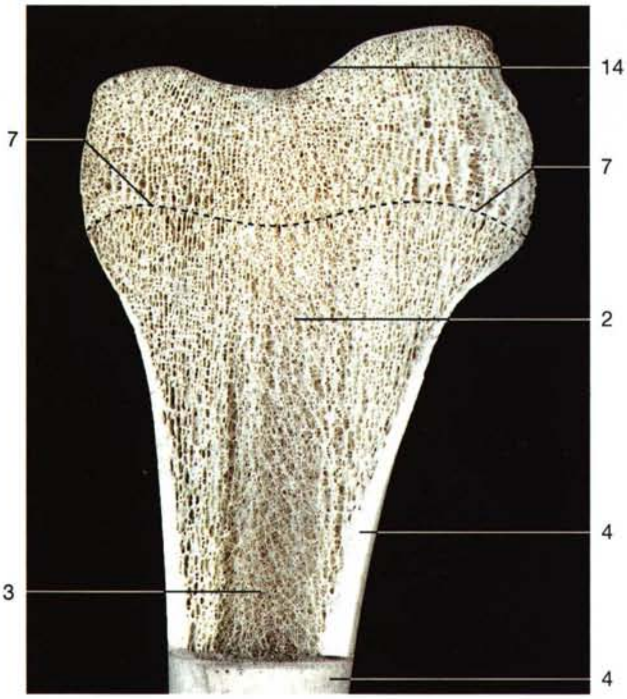

Coronal section through the proximal epiphysis of the adult tibia. Note the zone of dense bone at the site of the former epiphyseal plate (dotted line).

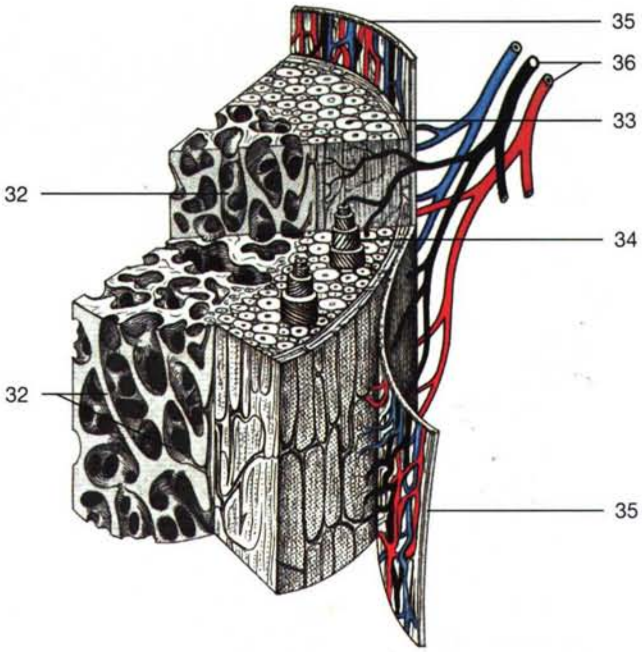

Structure of bones of the skeletal system (after Benninghoff). Note that the compact bone reveals a lamellar structure with Haversian lamellae and canals.

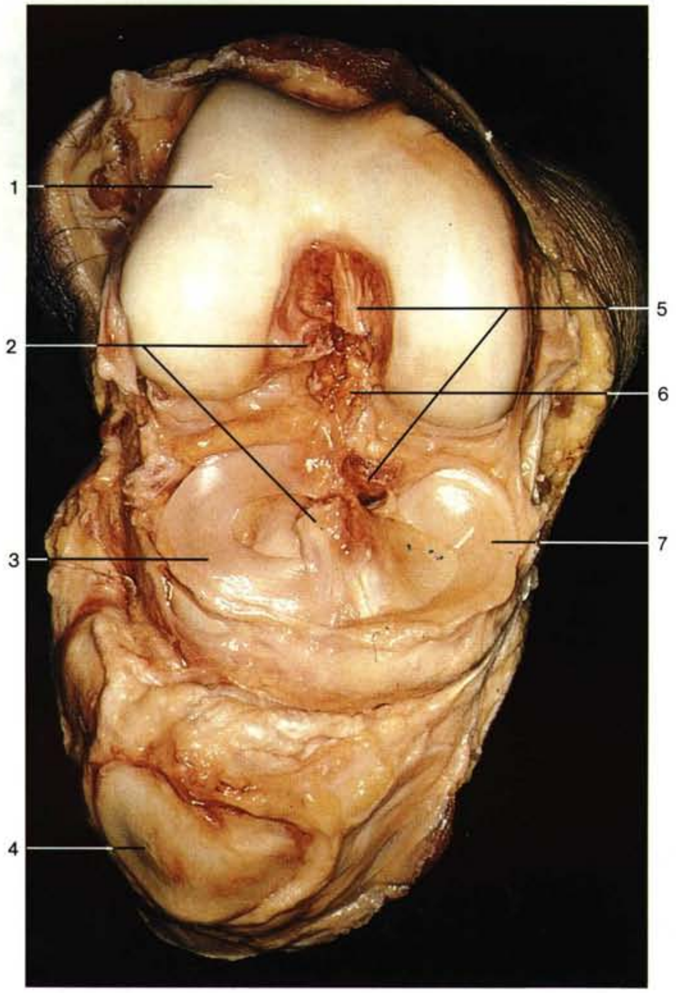

Knee joint. Anterior aspect, showing menisci and cruciate ligaments (cut). Quadriceps tendon cut and patella reflected distally.

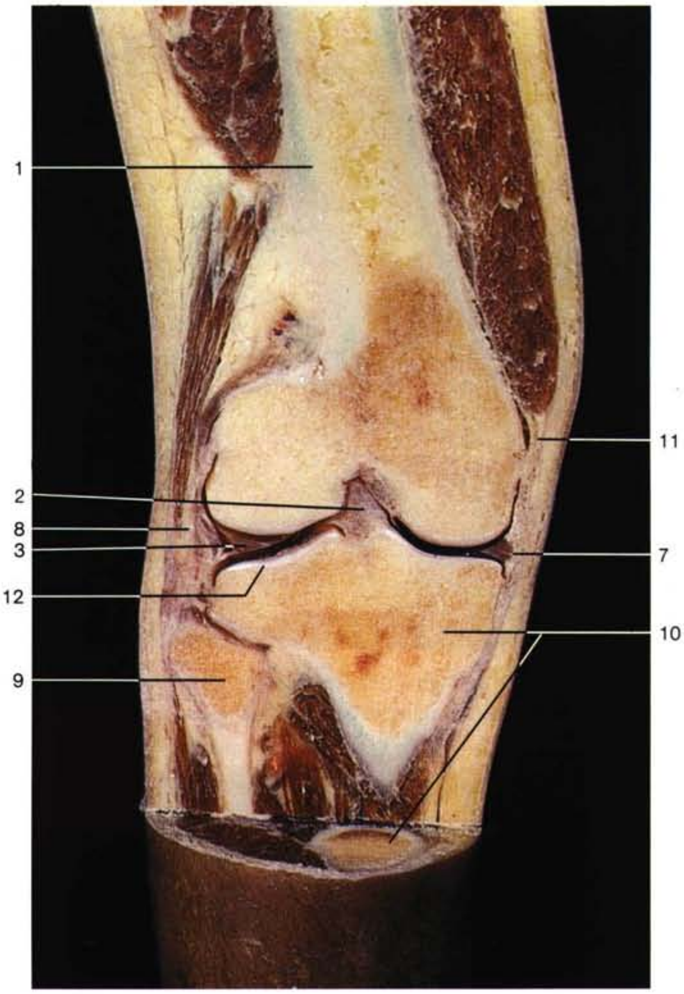

Coronal section through the knee joint. Anterior aspect of the right joint in extension.

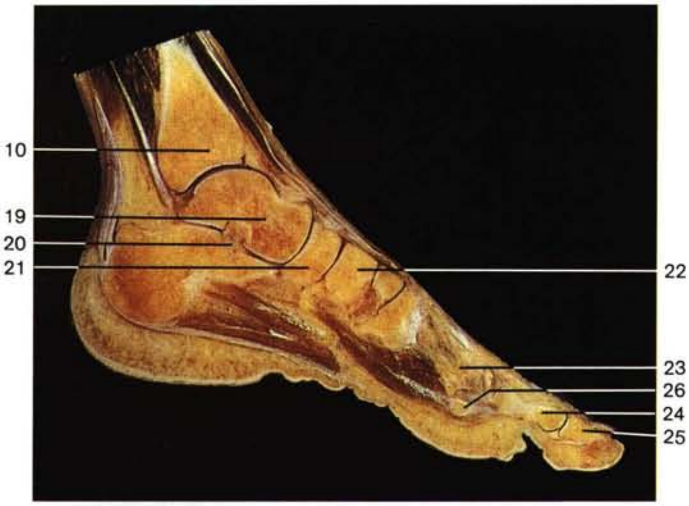

Sagittal section through the lower limb and the foot.

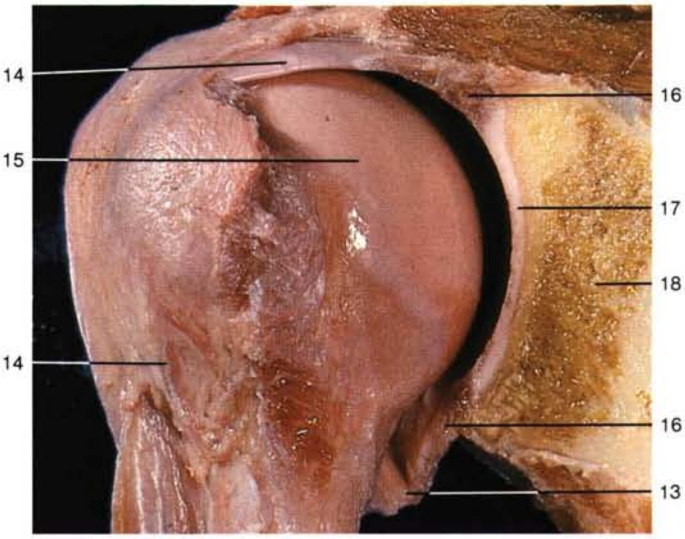

Shoulder joint (anterior view). The anterior part of the articular capsule has been removed.

1 Femur
2 Anterior cruciate ligament
3 Lateral meniscus
4 Patella
5 Posterior cruciate ligament
6 Posterior meniscofemoral ligament
7 Medial meniscus
8 Fibular collateral ligament
9 Fibula
10 Tibia
11 Tibial collateral ligament
12 Articular cartilage
13 Articular capsule
14 Tendon of long head of biceps brachii muscle
15 Head of humerus
16 Glenoid labrum
17 Articular cartilage of glenoid fossa
18 Scapula
19 Talus
20 Interosseous talocalcaneal ligament
21 Navicular bone
22 Medial cuneiform bone
23 First metatarsal bone
24 Proximal phalanx of the hallux (great toe)
25 Distal phalanx of the hallux
26 Sesamoid bone

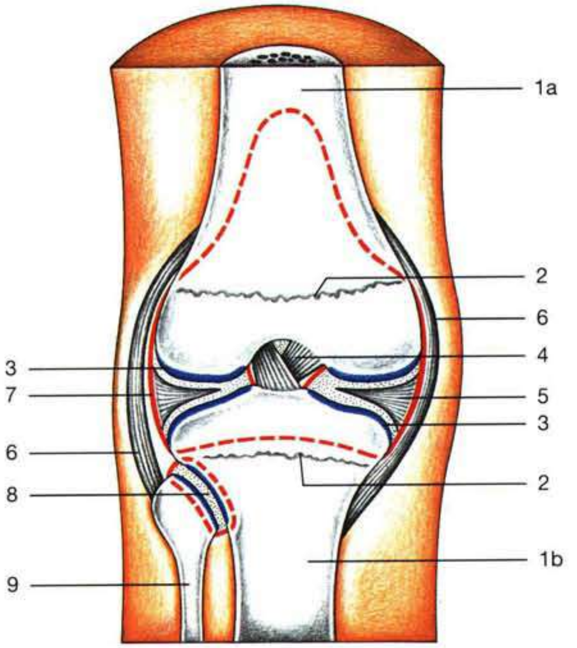

General architecture of a synovial joint with 2 articulating bones and a synovial cavity (right side, anterior view).
Coronal section throught the knee joint.
Red line = Articular capsule with synovial membrane.
Dotted red line = extension of articular capsule (suprapatellar bursa).

1 Articulating bones: a) Femur, b) Tibia
2 Epiphysial line
3 **Articular cartilage**
4 Intraarticular ligaments (e.g. cruciate ligaments)
5 Fibrocartilaginous disk (e.g. meniscus)
6 Collateral ligaments
7 Articular capsule with synovial membrane
8 **Tibiofibular articulation** (example of gliding synovial joint)
9 Fibula

	Movement	Examples
A Fibrous joints		
1 Sutures	No movements	Sutures of the skull
2 Syndesmoses	No movements	Distal tibiofibular joint
3 Gomphosis	No movements	Roots of teeth in alveolar process
B Cartilaginous joints		
1 Synchondroses	No movements	Epiphyseal plates
2 Symphyses	Slight movement	Symphysis pubis intervertebral discs
C Synovial joints		
1 Gliding	Monaxial	Intercarpal joint Intertarsal joint Sacroiliacal joint
2 Hinge	Monaxial	Interphalangeal joint Humeroulnar joint Talocrural joint
3 Pivot	Monaxial	Atlantoaxial joint Radioulnar joint
4 Ellipsoidal	Biaxial	Radiocarpal joint
5 Saddle	Biaxial	Carpometacarpal joint of the thumb
6 Ball-and-socket	Multiaxial	Shoulder and hip joint

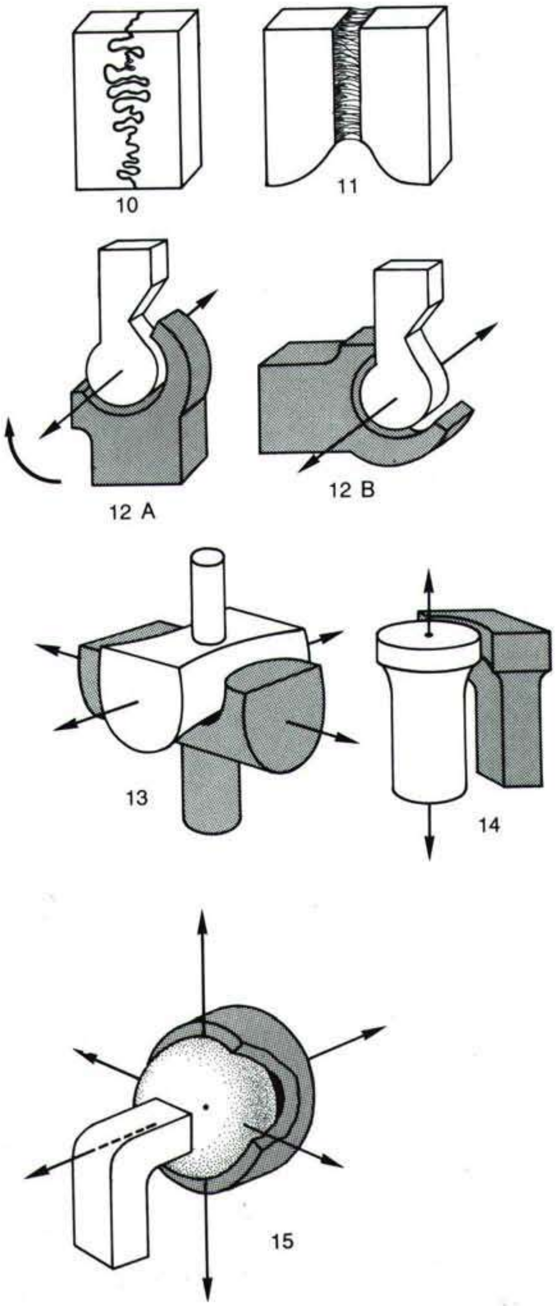

Main types of joints. Arrows: axes of movement.

Fibrous joints (synarthroses)
10 Serrate suture
11 Syndesmosis

Synovial joints (diarthroses)
12 Hinge joints (monaxial) ginglymus
  A Extension
  B Flexion
13 Saddle joint (biaxial)
14 Pivot joint (monaxial, rotation)
15 Ball-and-socket joint (multiaxial)

An articulation or joint is the functional connection between two or more bones. Joints can be divided into two categories depending upon whether the articulating surfaces of the bones are separated by a real cavity (joint cavity) so that they are movable against each other **(synovial joints)** or whether the bones are firmly connected by fibrous or cartilaginous tissue and practically immovable (**fibrous joints, cartilaginous joints,** symphysis, etc.). Synovial joints always possess a joint capsule (with a vascularized synovial membrane), articular cartilages, and a joint cavity. They are grouped according to the degree of movement they permit. A hinge joint (ginglymus) permits movement in only one plane about a single axis **(uniaxial** or **monaxial),** an ellipsoidal joint permits movements in two planes **(biaxial),** and ball-and-socket joints permit a range of movements around several axes **(multiaxial).** The following survey gives a few examples of these types of articulation.

1. Fibrous joints

A Sutures

a) Serrate or dentate suture

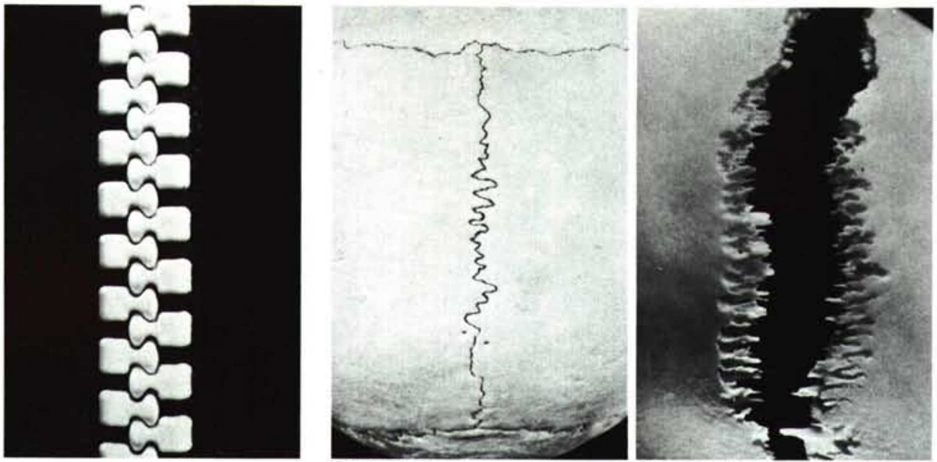

Sagittal suture

Separated

b) Squamous suture (arrows)

1 Parietal bone
2 Sphenoid bone
3 Squama of temporal bone

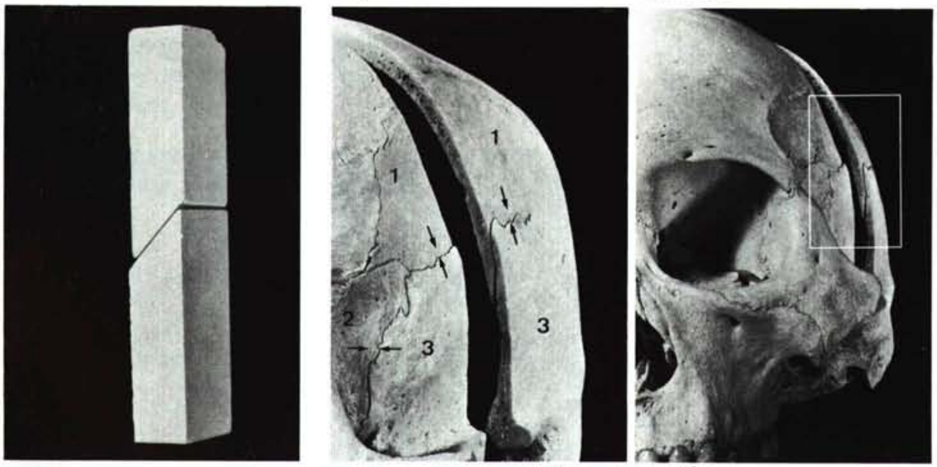

Squamous suture

c) Plane suture (arrow)

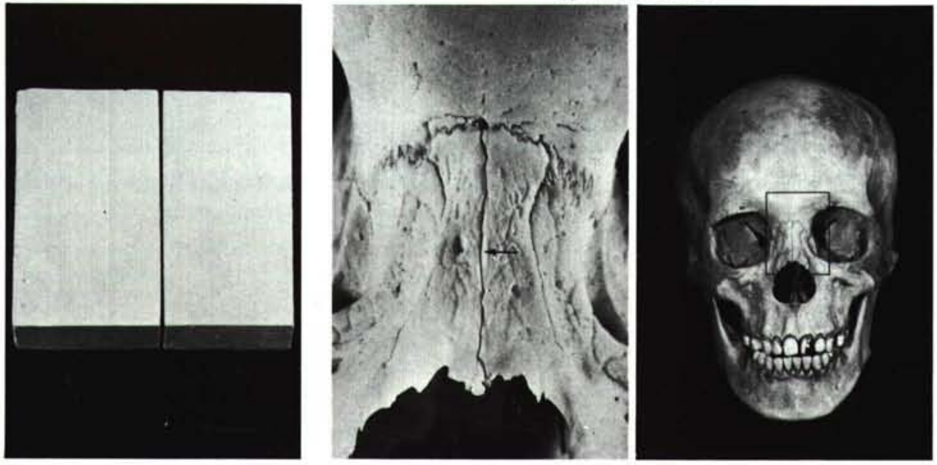

Nasal bones

B Peg suture
(Gomphosis)

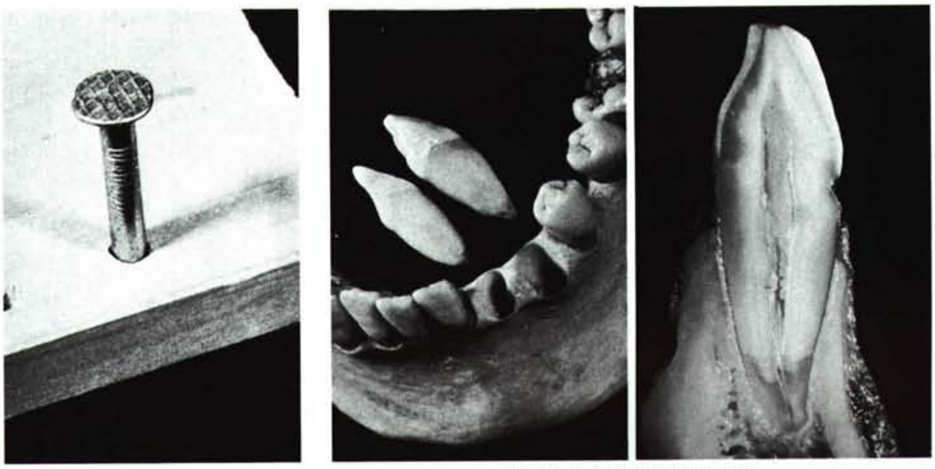

Teeth and alveolar process

2. Cartilaginous joints

a) Symphysis
(fibrocartilage)

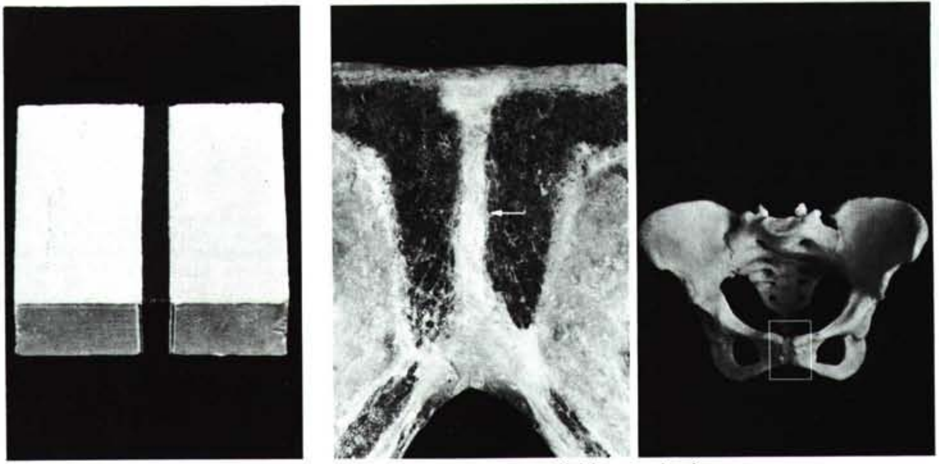

Pubic symphysis

b) Synchondrosis
(hyaline cartilage)

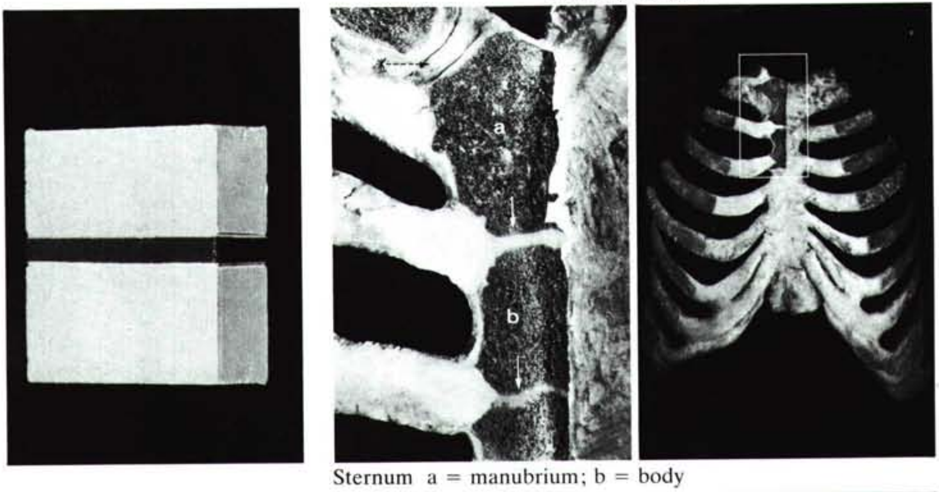

* Articular disc
(sternoclavicular joint)

Sternum a = manubrium; b = body

3. Osseous joints
(synostosis)

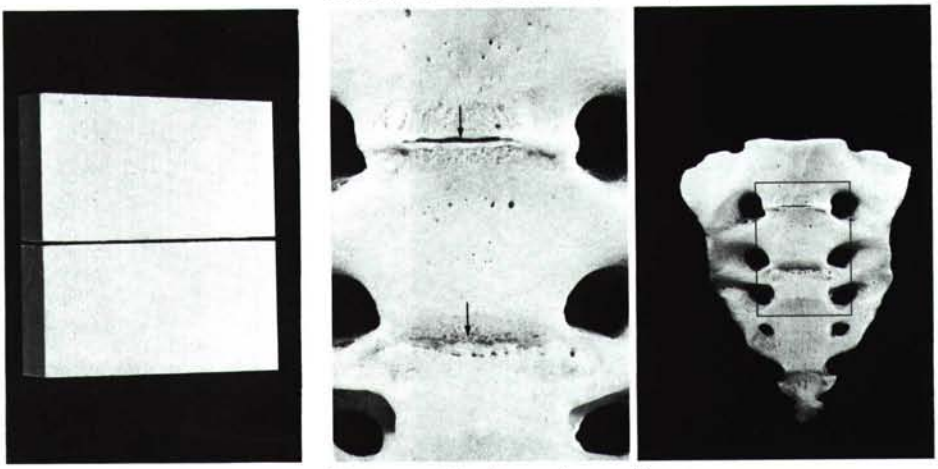

Transverse ridge (arrows)

Sacrum

1. Ball-and-socket joint

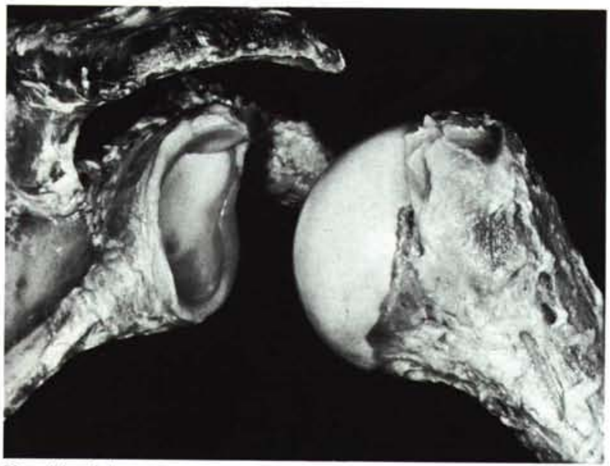

Shoulder joint

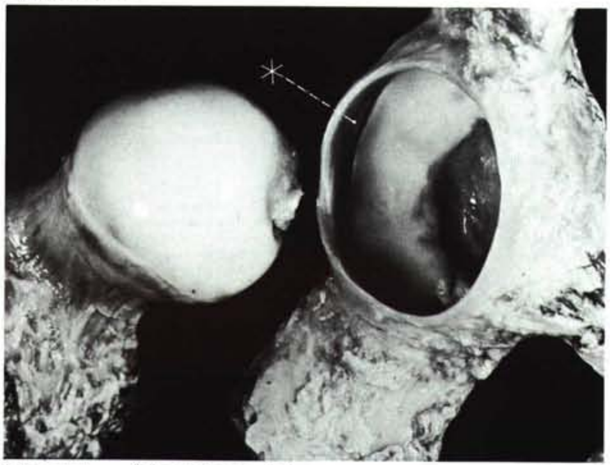

Hip joint *Acetabular labrum

2. Ellipsoid joint

OFFICIAL
Rugby Football
OFFICIAL
SIZE-WEIGHT

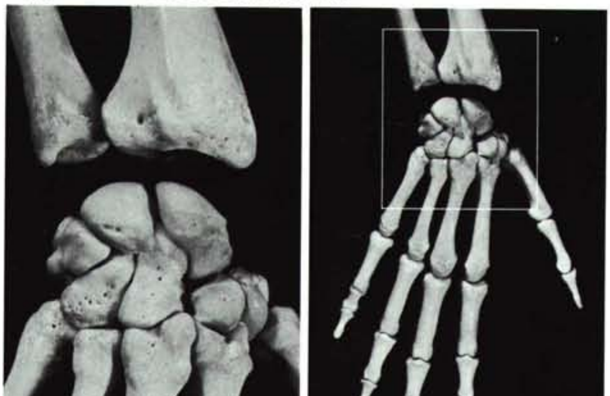

Radiocarpal joint

3. Hinge joint

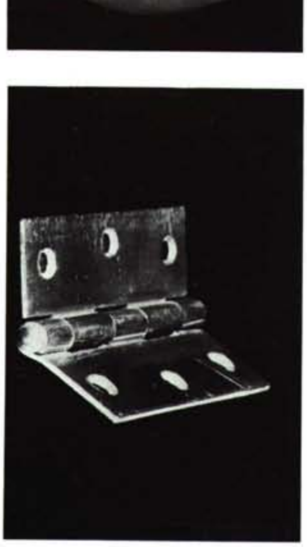

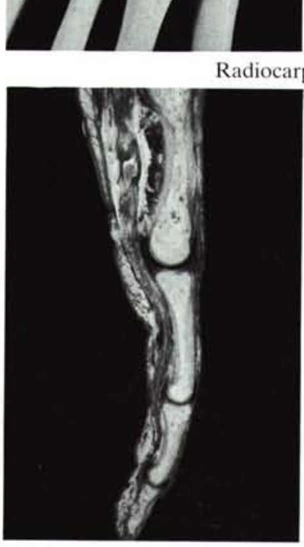

Interphalangeal joint

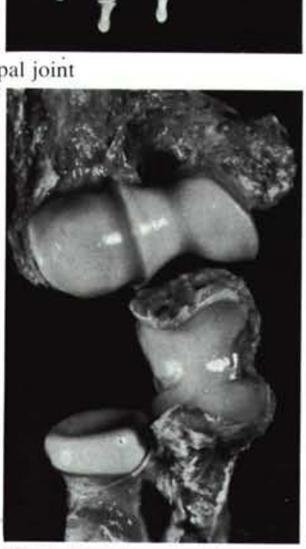

Elbow joint (humeroulnar joint)

4. Pivot joint

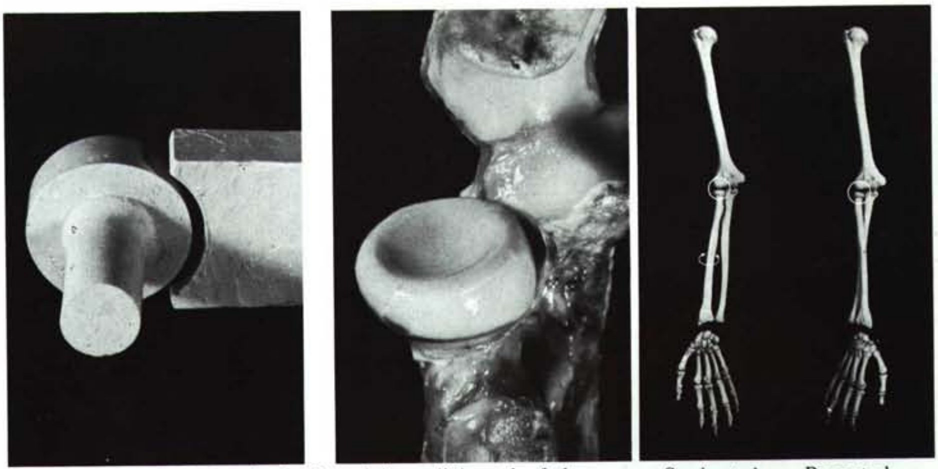

Radioulnar joint; left: head of radius, right: radial notch of ulna

Supinated Pronated

5. Hinge joint

(but a gradually tilting axis produces a slight spiral motion)

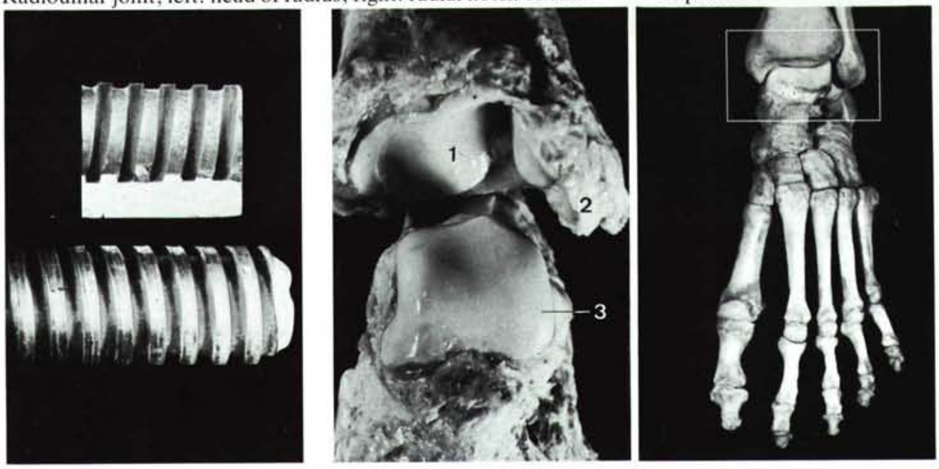

Ankle joint: 1 = tibia; 2 = fibula; 3 = talus

6. Saddle joint

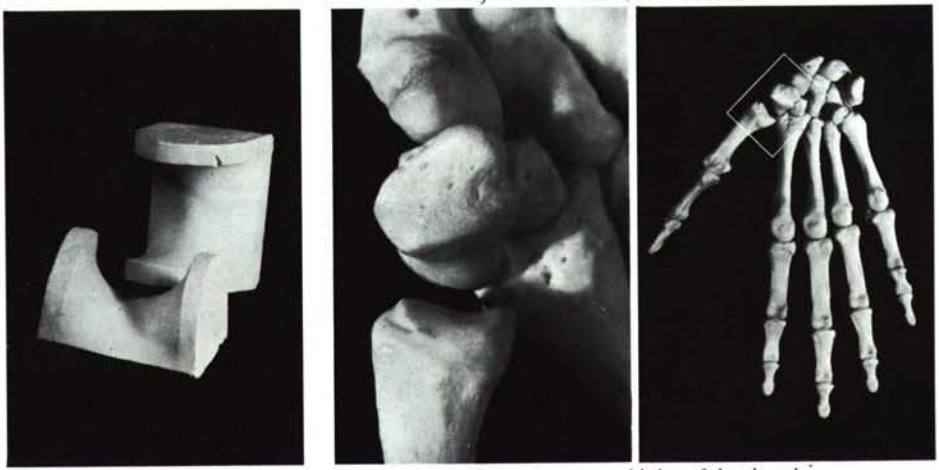

Carpometacarpal joint of the thumb

7. Plane joint

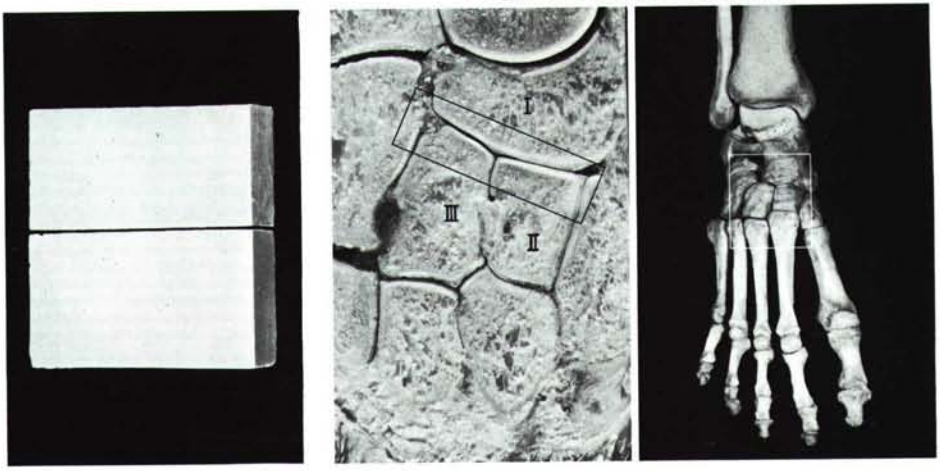

I = navicular; **II** = intermediate cuneiform; **III** = lateral cuneiform

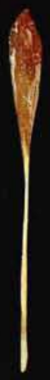

Fusiform
(palmaris longus muscle)

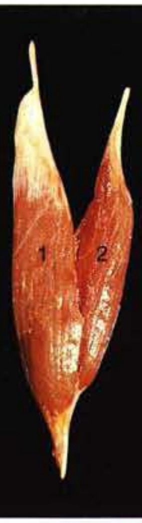

Bicipital
(biceps brachii muscle)

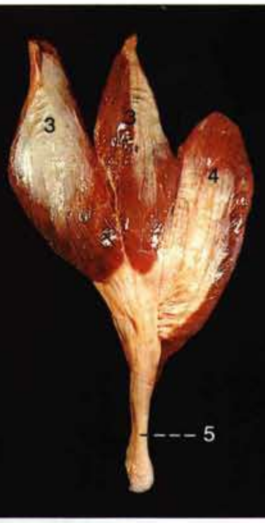

Tricipital
(triceps surae muscle)

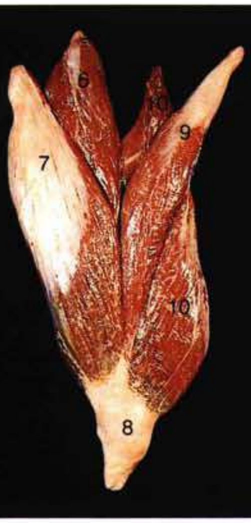

Quadricipital
(quadriceps femoris muscle)

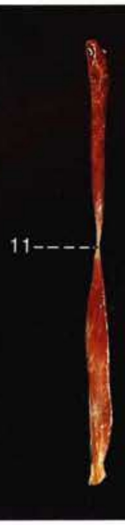

Digastric
(omohyoid muscle)

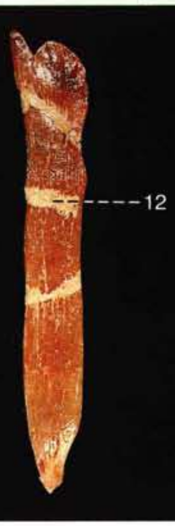

Multiventral
(rectus abdominis muscle)

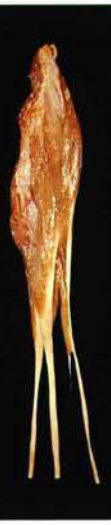

Multicaudal
(flexor prof. muscle)

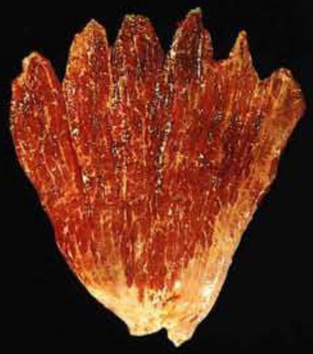

Serrated
(serratus anterior muscle)

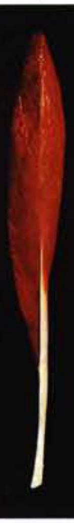

Bipennate
(tibialis anterior muscle)

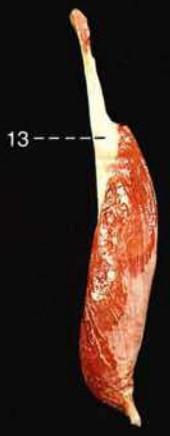

Unipennate
(semimembranous muscle)

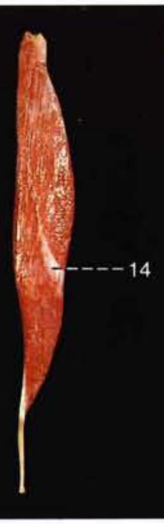

Semitendinous
(semitendinous muscle)

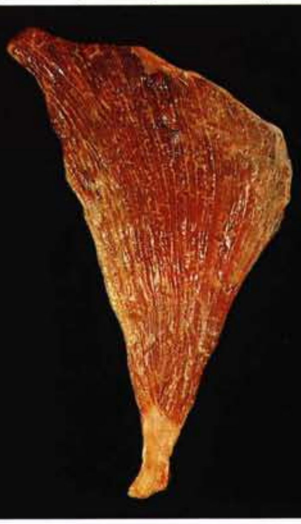

Broad, flat muscle
(latissimus dorsi muscle)

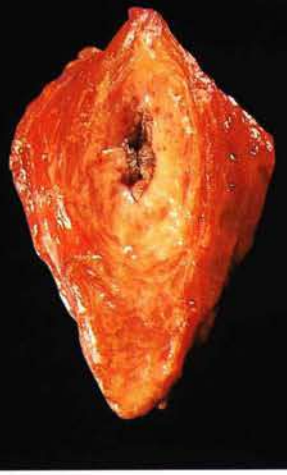

Ring-like
(external anal sphincter muscle)

1 Long head
2 Short head
3 Gastrocnemius muscle (medial head, lateral head)
4 Soleus muscle
5 Achilles tendon
6 Vastus intermedius muscle
7 Vastus lateralis muscle
8 Patella
9 Rectus femoris muscle
10 Vastus medialis muscle
11 Intermediate tendon
12 Tendinous intersection
13 Aponeurosis
14 Tendinous intersection

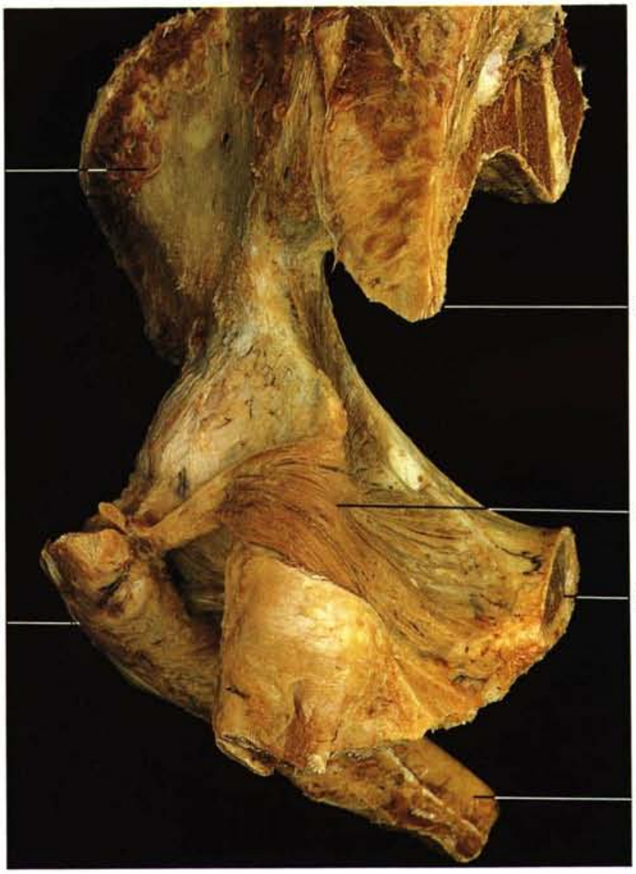

Left half of the pelvis (posterior aspect).
Obturator internus muscle as an example of a muscle, the tendon of which does not act in the direction of the main muscle fibers. Its fibers originate at the internal aspect of the obturator foramen, turn around the posterior rim of the ischium and insert at the greater trochanter of the femur. The ischium thereby serves as a pulley.

1 Ilium
2 Greater trochanter
3 Coccyx
4 Obturator internus muscle
5 Pubis
6 Femur

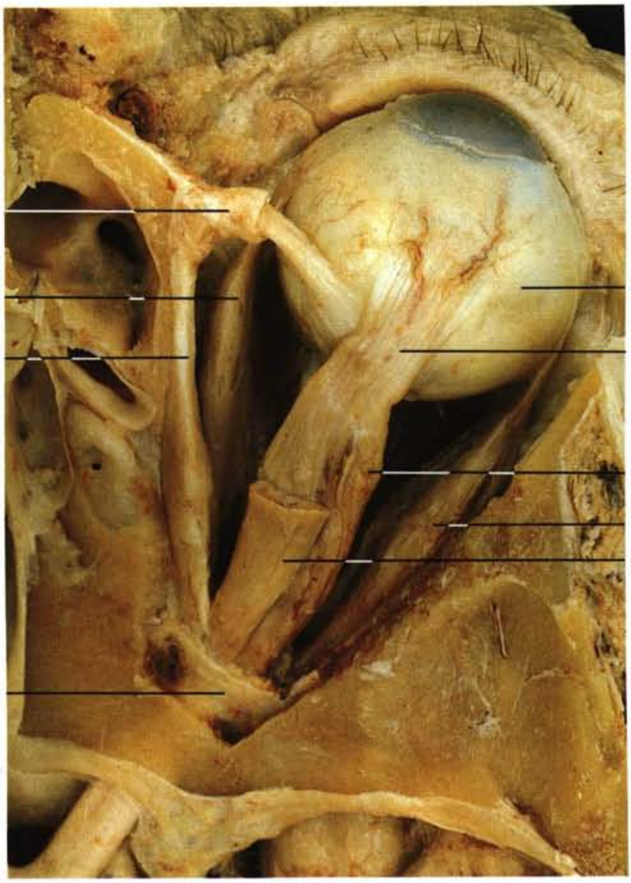

Superior oblique muscle of the eyeball, right eye (superior aspect). The tendon of this muscle bends over the trochlea changing its direction so that it becomes attached to the posterior lateral quadrant of the eyeball.

1 Trochlea
2 Medial rectus muscle
3 Superior oblique muscle
4 Common annular tendon
5 Eyeball
6 Superior rectus muscle
7 Lateral rectus muscle
8 Superior rectus muscle (tendon)
9 Levator palpebrae superioris muscle (divided)

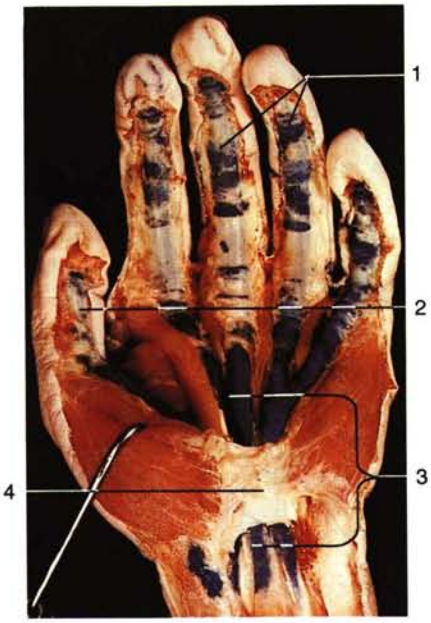

1 Digital synovial sheaths of the tendons of flexor digitorum superficialis and profundus muscles
2 Digital synovial sheaths of the tendon of long flexor pollicis longus muscle
3 Common flexor synovial sheaths of flexor digitorum superficialis and profundus muscles
4 Flexor retinaculum

The synovial sheaths of the tendons on the palmar aspect of the left wrist (colored fluid has been injected).

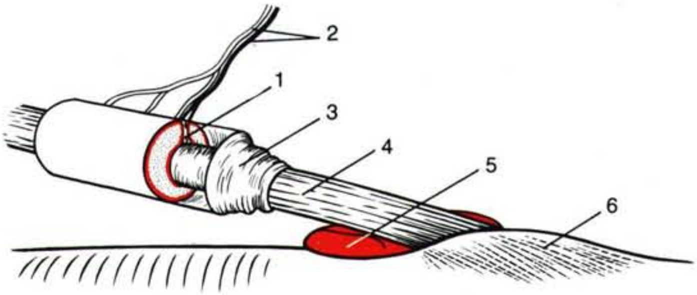

1 Mesotendon
2 Blood vessels
3 Synovial sheath
4 Tendon
5 Synovial bursa
6 Bone (tuberosity)

Structure of a tendon sheath. The synovial membrane which also forms the mesotendon is indicated in red. (Schematic drawing.)

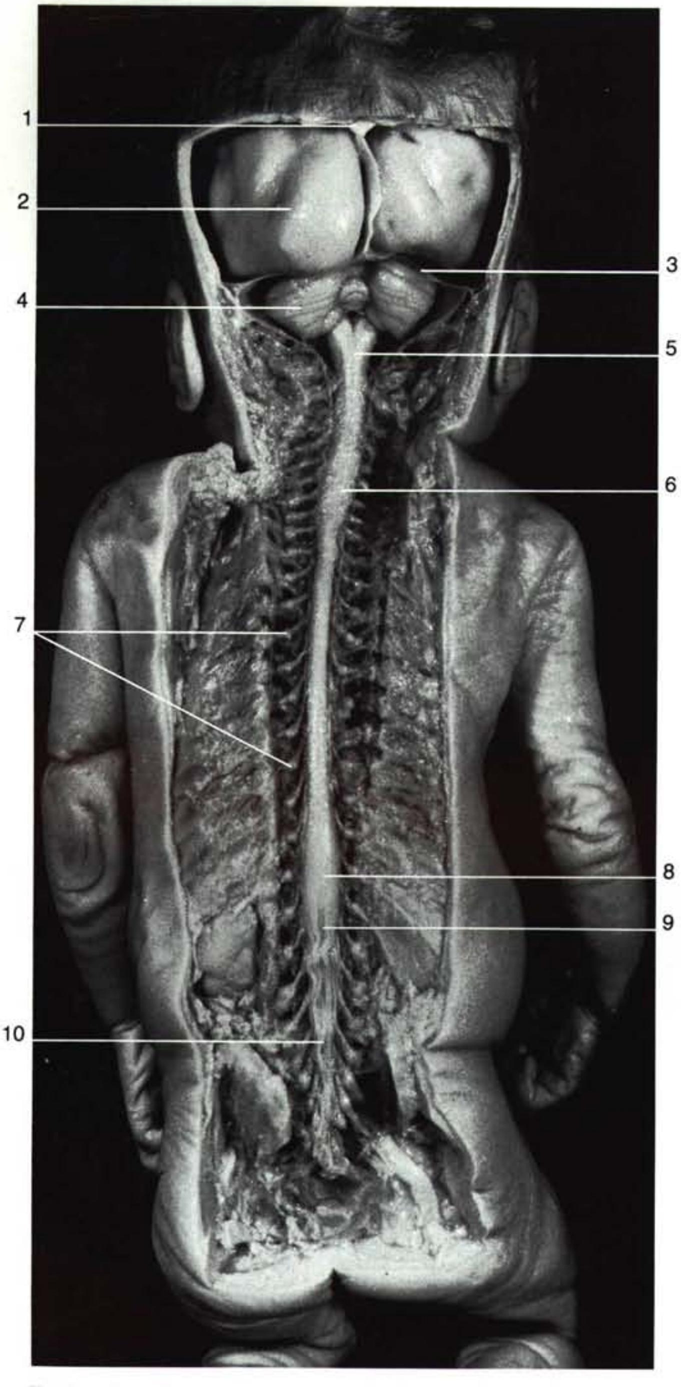

1 Falx cerebri
2 Cerebral hemispheres
3 Tentorium cerebelli
4 Cerebellum
5 Medulla oblongata
6 Spinal cord, cervical enlargement
7 Spinal ganglia
8 Spinal cord, lumbar enlargement
9 Conus medullaris
10 Cauda equina
11 Cervical plexus (formed from ventral rami of C_1–C_4)
12 Brachial plexus (formed from ventral rami of C_5–T_1)
13 Lumbosacral plexus (formed from ventral rami of L_1–S_4)
14 Sympathetic trunk

Brain, the **spinal cord** and the **spinal nerves** in the fetus (posterior aspect).

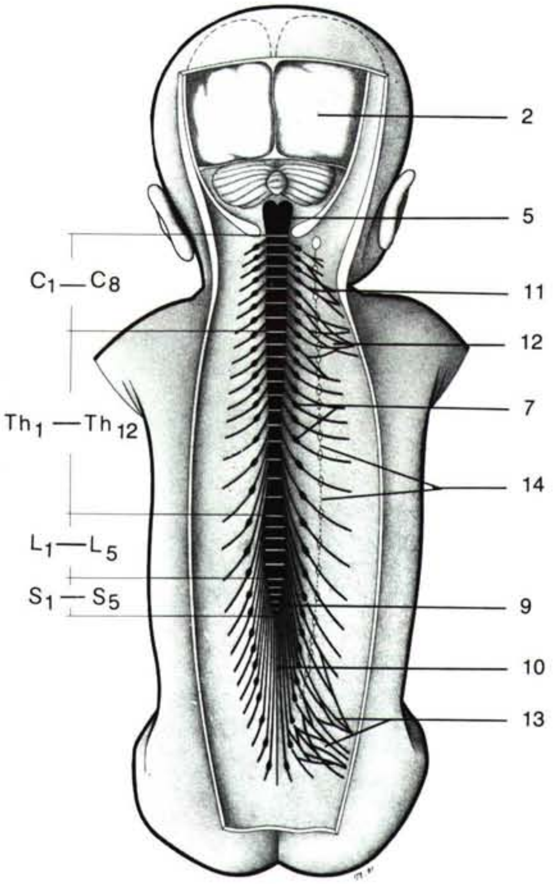

Schematic drawing to illustrate the three main parts of the nervous system in general.

The **nervous system** can be divided into three, functionally distinct parts: 1. the cranial part which comprises the great sensory organs and the brain, 2. the spinal cord which shows a segmental structure and serves predominantly as a reflex-organ, and 3. the autonomic nervous system which controls the unvoluntary functions (subconscious control) of organs and tissues. The autonomic part of the nervous system forms many delicate plexus within the organs. At certain places these plexus contain aggregations of nerve cells (prevertebral and intramural ganglia). The spinal nerves leave the spinal cord at regular intervals, forming the 8 cervical, 12 thoracic, 5 lumbar, 5 sacral and a varying number of coccygeal segments. The ventral rami of the first four cervical spinal nerves (C_1–C_4) form the cervical plexus (for innervation of the anterior neck), the ventral rami of the lower cervical spinal nerves (C_5–T_1) form the brachial plexus which innervates the upper leg, and the ventral rami of the lumbar and sacral spinal nerves form the lumbosacral plexus (L_1–S_4) which innervates the pelvic and genital organs and the lower extremity.

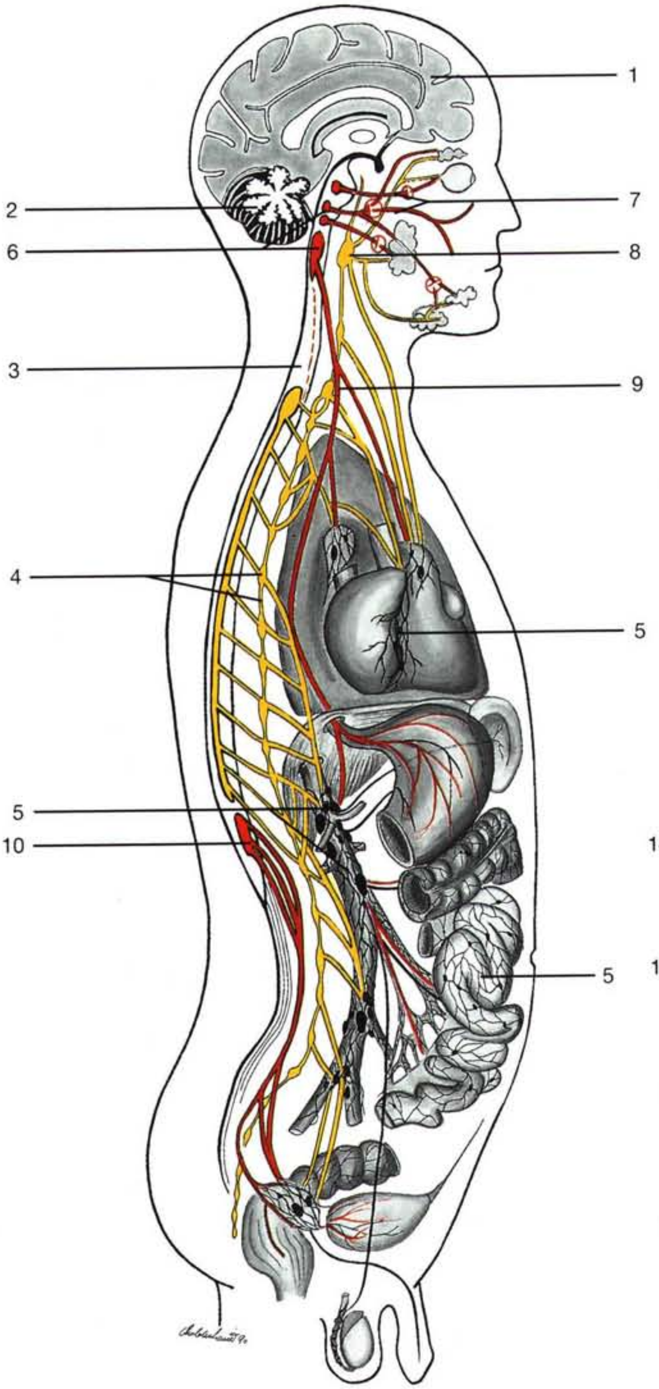

Diagram illustrating the localization of the **three functional portions** of the **nervous system** (brain, spinal cord and autonomic nervous system). Yellow = sympathetic system; red = parasympathetic system.

1 Cerebrum
2 Cerebellum
3 Spinal cord
4 Sympathetic trunk and ganglion
5 Plexus and ganglia of the autonomic nervous system
6 Cranial autonomic system
7 Cranial nerves (n. III and n. VII)
8 Superior cervical ganglion
9 Vagus nerve (n. X)
10 Sacral autonomic system

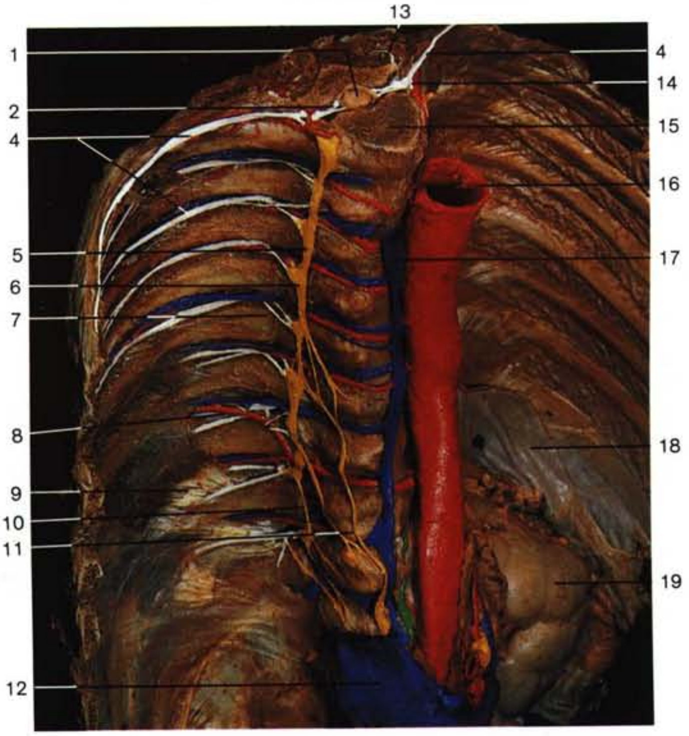

Posterior part of the thorax. Cross-section at the level of the 5th thoracic segment. Spinal nerves and their connections to the sympathetic trunk.

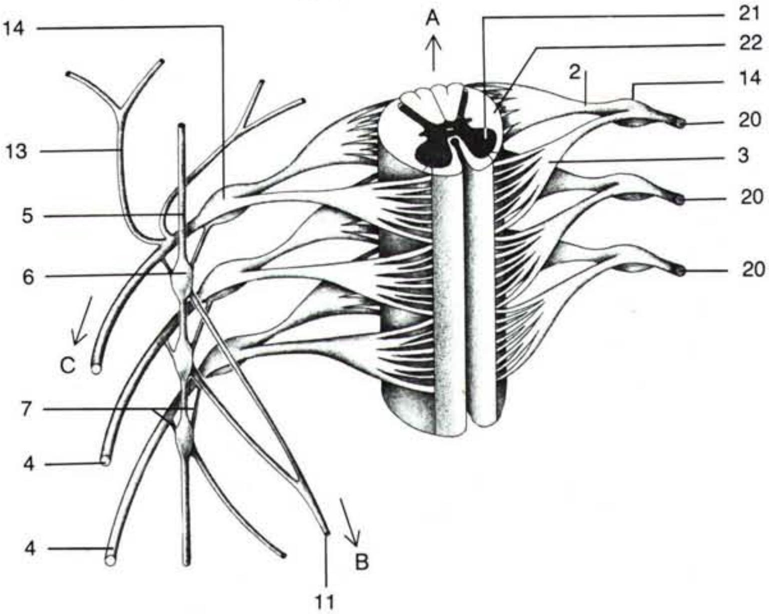

Organization of the spinal cord in structurally equal segments which form the paired spinal nerves. A = connections to the brain; B = connections to the autonomic nervous system; C = connections to the trunk and extremities (intercostal nerves and plexus). (Schematic drawing.)

1 Spinal cord
2 **Dorsal root**
3 **Ventral root**
4 Intercostal nerves
5 Sympathetic trunk
6 Ganglia of the sympathetic trunk
7 Rami communicantes
8 Intercostal artery and vein
9 Subcostalis muscle
10 Lesser splanchnic nerve
11 Greater splanchnic nerve
12 Inferior vena cava
13 Dorsal ramus of spinal nerve
14 **Spinal (dorsal root) ganglion**
15 Body of the vertebra
16 Aorta
17 Azygos vein
18 Diaphragm
19 Left kidney
20 **Spinal nerve**
21 Gray matter of the spinal cord
22 White matter of the spinal cord

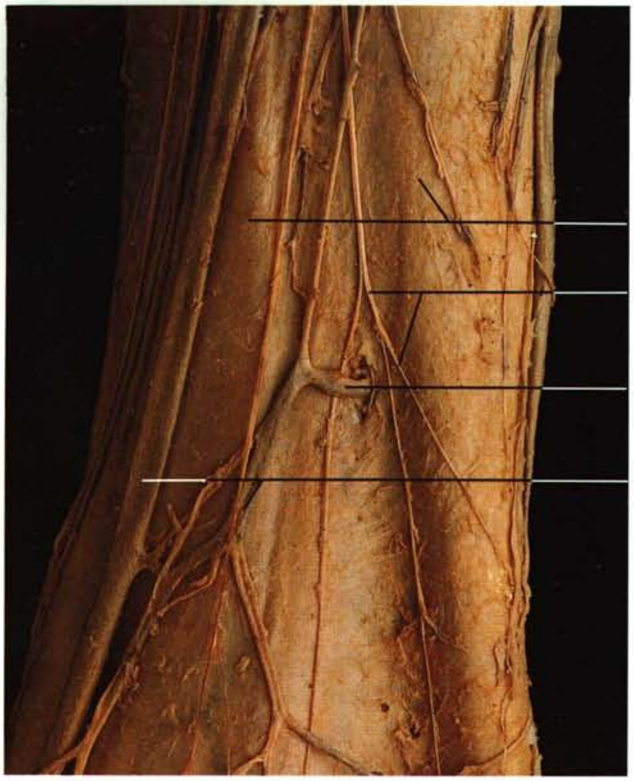

Superficial nerves and vessels of the lower leg, illustrating the structural differences between veins and nerves.

1 Crural fascia (fascia cruris)
2 Cutaneous nerves
3 Superficial cutaneous veins
4 Perforating vein

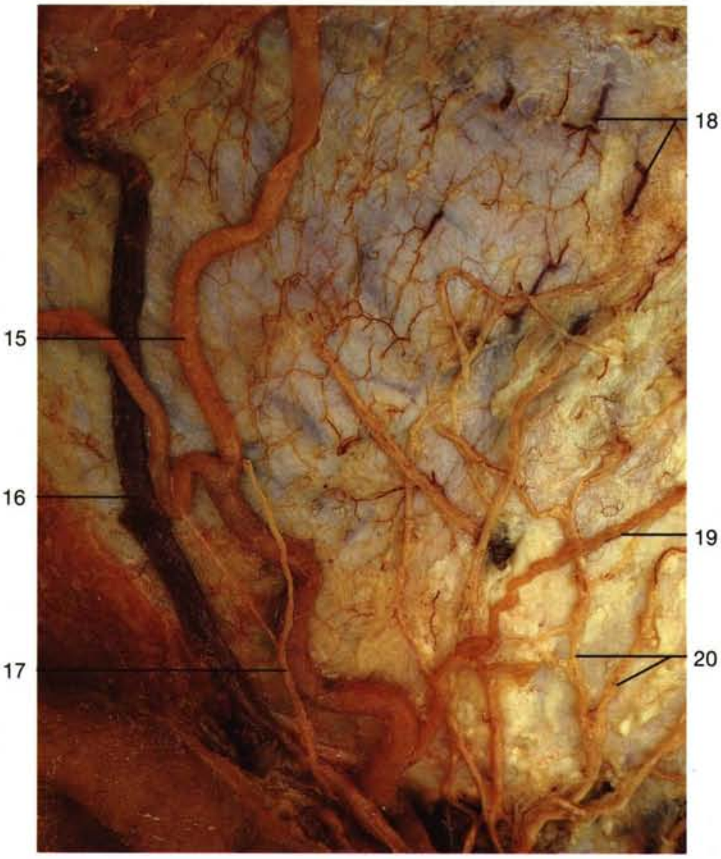

Superficial nerves and vessels. Temporal region. Note the differences between arteries, veins and nerves.

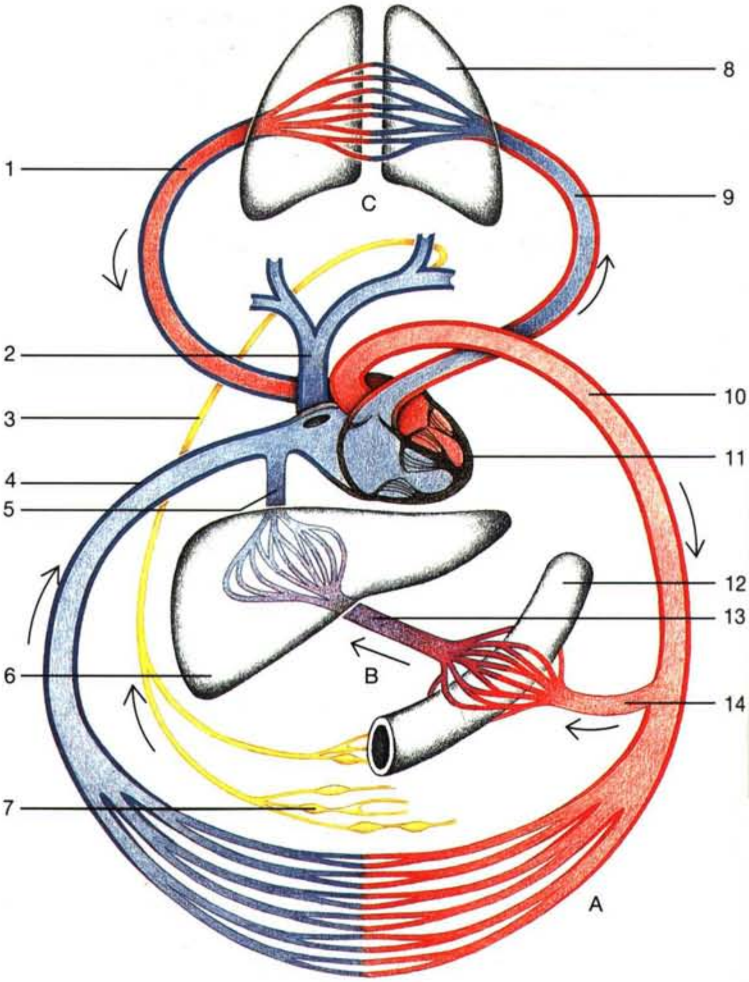

Organization of the circulatory system. Arrows: direction of the blood flow.

Vessel wall in red = Arteries
Vessel wall in blue = Veins
Yellow = Lymphatic vessels

A Systemic circulation
B Hepatic portal circulation
C Pulmonary circulation

1 Pulmonary vein
2 Superior vena cava
3 Thoracic duct
4 Inferior vena cava
5 Hepatic vein
6 Liver
7 Lymph nodes and lymphatic vessels
8 Lung
9 Pulmonary artery
10 Aorta
11 Heart
12 Small intestine with capillary network
13 Portal vein
14 Mesenteric artery
15 Superficial temporal artery
16 Superficial temporal vein
17 Auriculotemporal nerve
18 Perforating veins for subcutaneous fatty tissue
19 Small artery
20 Small nerves (branches of facial nerve)

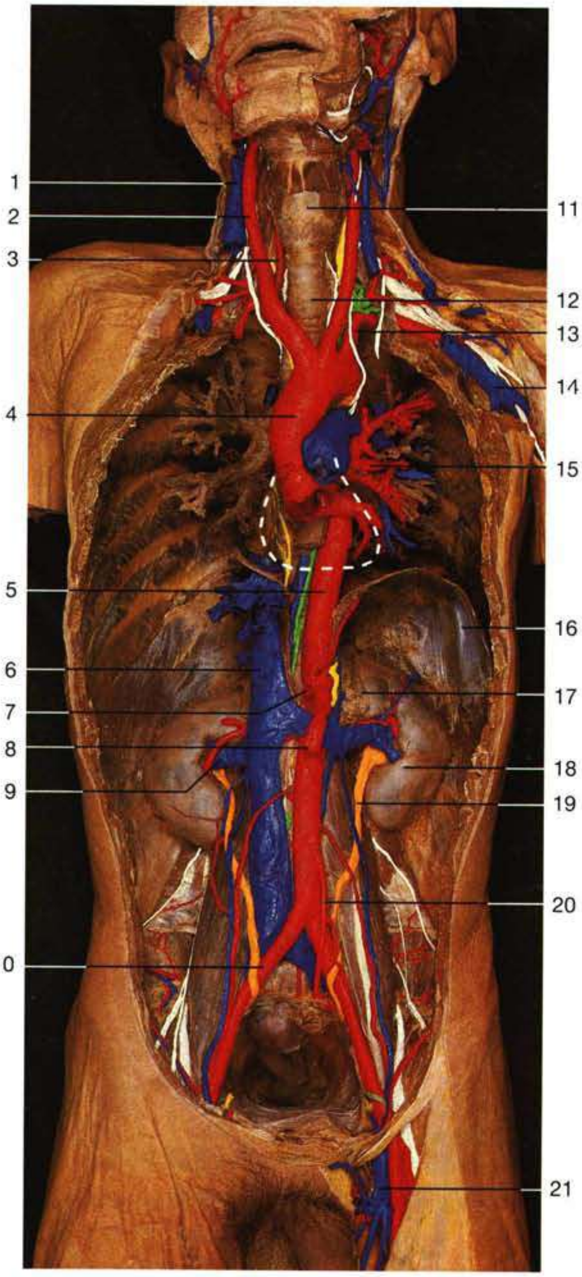

Major vessels of the trunk. The position of the heart is indicated by the dotted line.

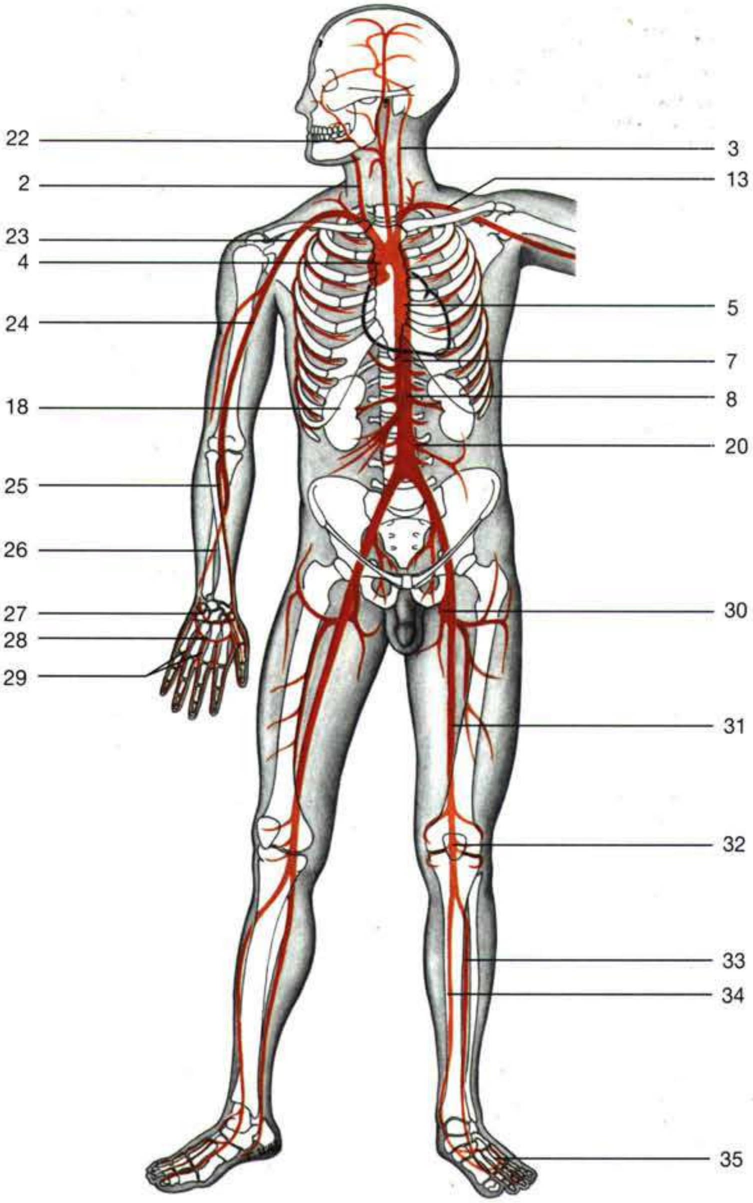

Major arteries of the human body. (Schematic drawing.)

1 Internal jugular vein
2 **Common carotid artery**
3 Vertebral artery
4 **Ascending aorta**
5 **Descending aorta**
6 **Inferior vena cava**
7 Celiac trunk
8 Superior mesenteric artery
9 Renal vein
10 **Common iliac artery**
11 Larynx
12 Trachea
13 Left subclavian artery
14 Left axillary vein
15 **Pulmonary veins**
16 Diaphragm
17 Suprarenal gland
18 Kidney
19 Ureter
20 Inferior mesenteric artery
21 **Femoral vein**
22 Facial artery
23 Axillary artery
24 Brachial artery
25 Radial artery
26 Ulnar artery
27 Deep palmar arch
28 Superficial palmar arch
29 Common palmar digital arteries
30 Profunda femoris artery
31 **Femoral artery**
32 Popliteal artery
33 Anterior tibial artery
34 Posterior tibial artery
35 Plantar arch

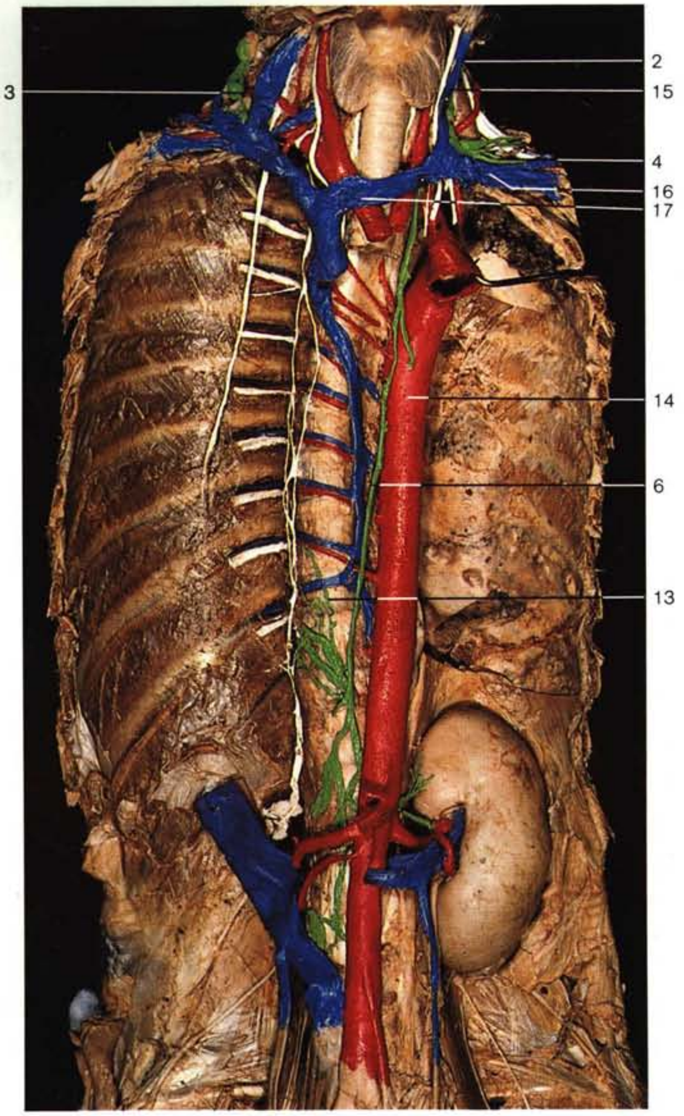

Major lymph vessels of the trunk.

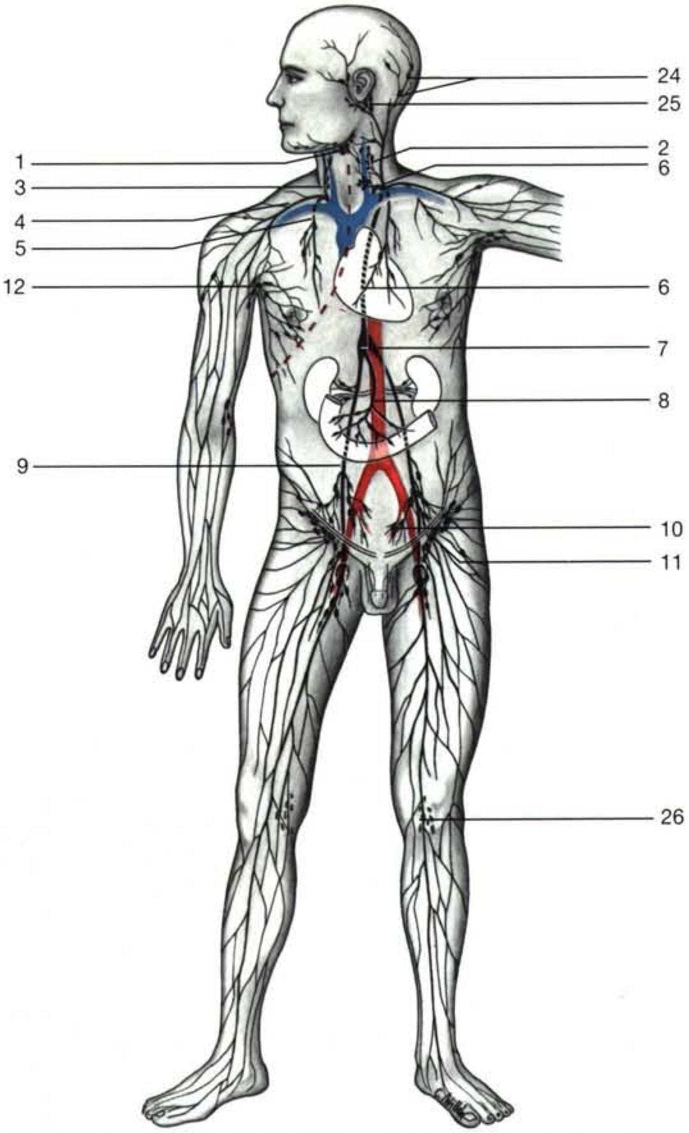

Lymphatic system. Course of the main lymphatic vessels and lymph nodes in the body. Dotted line = border between lymphatic vessels draining towards the right venous angle and towards the left.

1 Submandibular nodes
2 Deep cervical nodes
3 Right jugular trunk
4 Subclavian trunk
5 Right bronchomediastinal trunk
6 Thoracic duct
7 **Cisterna chyli**
8 **Intestinal trunk**
9 Right **lumbar trunk**
10 Internal iliac nodes
11 **Inguinal nodes**
12 **Axillary nodes**
13 Descending trunk
14 Descending aorta
15 Internal jugular vein
16 Subclavian vein
17 Left brachiocephalic vein
18 Mandible
19 Larynx
20 Internal jugular vein
21 **Deep cervical nodes**
22 Cervical nerve plexus
23 Superficial layer of deep fascia
24 Occipital nodes
25 Parotid nodes
26 Popliteal nodes

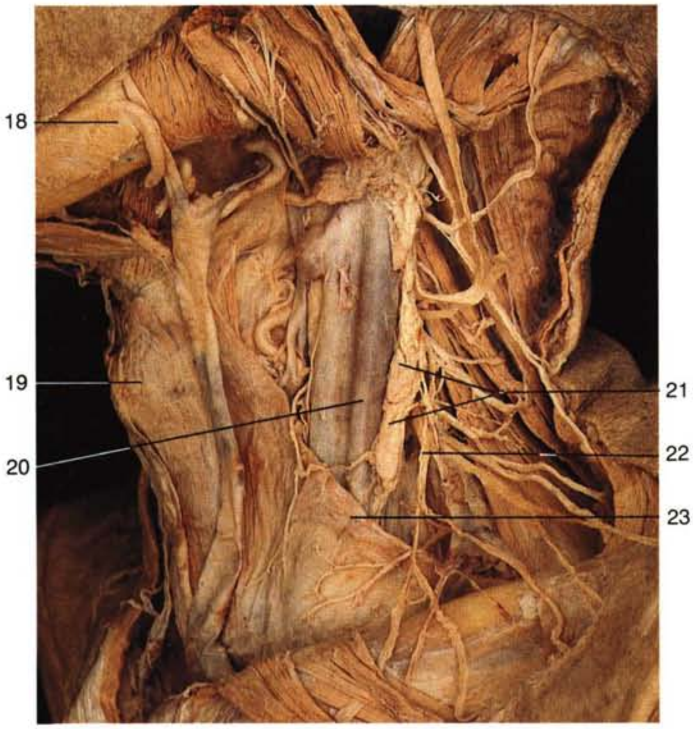

Deep cervical nodes, adjacent to the internal jugular vein.

Lymphatic vessels originate as blind-ending tubes in the tissue spaces (lymph capillaries) and unite to form larger vessels (lymphatics). These resemble veins but have a much thinner wall, more valves and are interrupted by lymph nodes at various intervals. Large groups of lymph nodes are located in the inguinal and axillary regions, deep to the mandible and sternocleidomastoid muscle and within the root of the mesentery of the intestine.

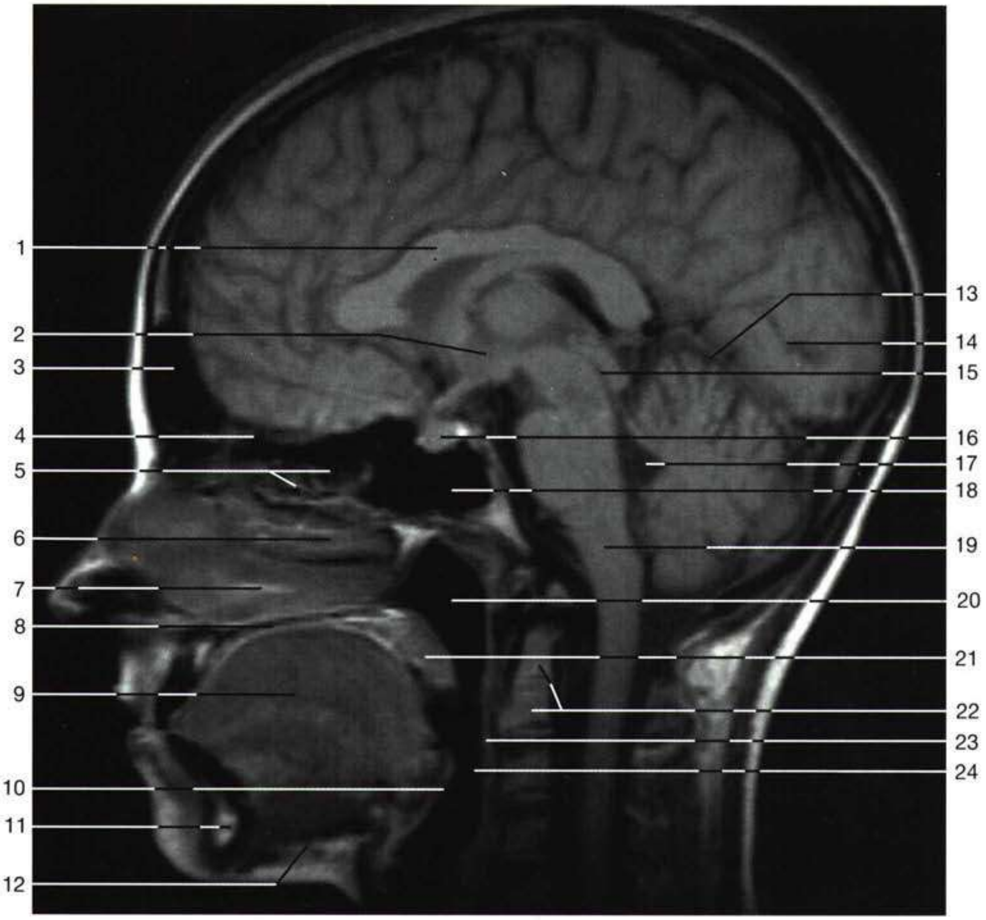

1 Corpus callosum
2 **Hypothalamus**
3 Frontal sinus
4 Cribriform plate
5 Ethmoidal air cells
6 Middle nasal concha
7 Inferior nasal concha
8 Hard palate
9 **Tongue**
10 Epiglottis
11 Mandible
12 Mylohyoid muscle
13 Tentorium of cerebellum
14 Calcarine fissure
15 Cerebral aqueduct
16 **Pituitary gland**
17 Fourth ventricle
18 **Sphenoidal sinus**
19 Medulla oblongata
20 Nasopharynx
21 Uvula
22 Dens of axis
23 Constrictor muscle of pharynx
24 Oral part of pharynx
25 **Cerebrum** (right hemisphere)
26 Calvaria
27 **Cerebellum**

Sagittal section through head and neck (MR-Scan. 23-year-old female, courtesy of PD Dr. A. Heuck, Munich).

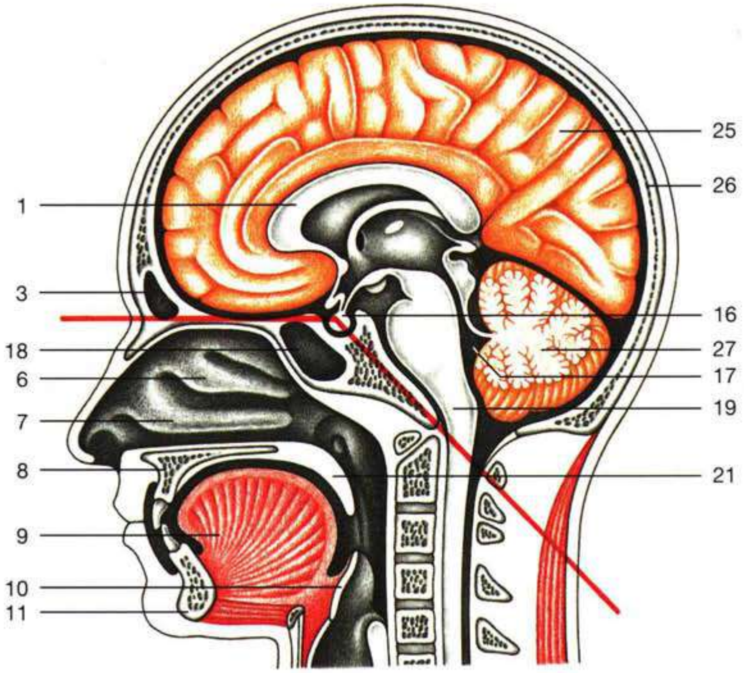

Sagittal section through the head
(schematic drawing). The red line represents the border between the neurocranium and viscerocranium forming the clivus angle. The neural cavity contains the brain; the viscerocranium comprises the orbit, the nasal cavity and the oral cavity arranged one beneath the other.

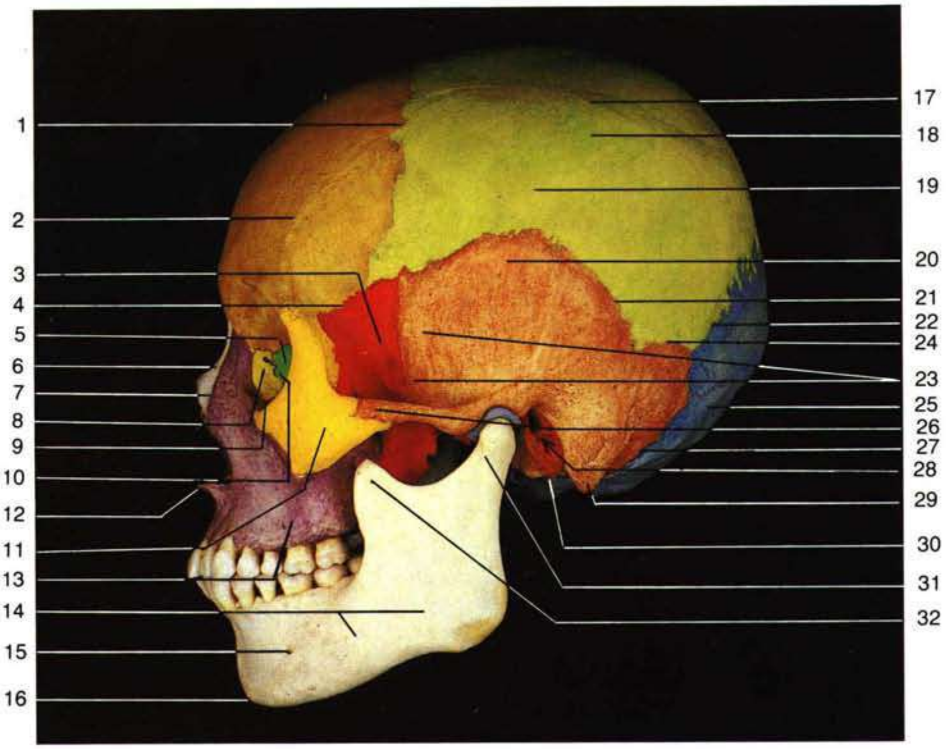

General architecture of the skull (lateral aspect). The different bones are indicated in color (numbers cf. table).

1	Coronal suture
2	**Frontal bone**
3	**Sphenoid bone**
4	Sphenofrontal suture
5	**Ethmoid bone**
6	**Nasal bone**
7	Nasomaxillary suture
8	**Lacrimal bone**
9	Lacrimomaxillary suture
10	Lacrimoethmoid suture
11	**Zygomatic bone**
12	Anterior nasal spine
13	**Maxilla**
14	**Mandible**
15	Mental foramen
16	Mental protuberance
17	Superior temporal line
18	Inferior temporal line
19	**Parietal bone**
20	**Temporal bone**
21	Squamous suture
22	Lambdoid suture
23	Temporal fossa
24	Parietomastoid suture
25	**Occipital bone**
26	Zygomatic arch
27	Occipitomastoid suture
28	External acoustic meatus
29	Mastoid process
30	Tympanic portion of temporal bone
31	Condylar process of mandible
32	Coronoid process of mandible

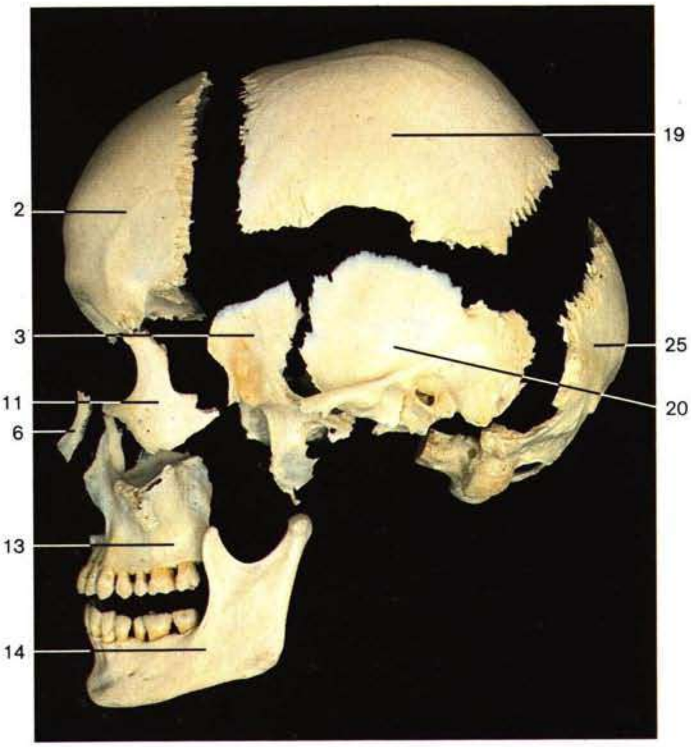

Lateral aspect of the disarticulated skull (palatine bone, lacrimal bone, ethmoid bone and vomer are not depicted).

2	Frontal bone (orange)		Cranial bones
19	Parietal bone (light green)		
3	Greater wing of sphenoid bone (red)		
25	Squama of occipital bone (blue)		
20	Squama of temporal bone (brown)		
5	Ethmoid bone (dark green)		Base of skull
3	Sphenoid bone (red)		
	Temporal bone excluding squama (brown)		
30	Tympanic portion of temporal bone (dark brown)		
	Occipital bone excluding squama (blue)		
6	Nasal bone (white)		Facial bones
8	Lacrimal bone (yellow)		
	Inferior nasal concha		
	Vomer		
11	Zygomatic bone (light yellow)		
	Palatine bone		
13	Maxilla (violet)		
14	Mandible (white)		
	Malleus	within petrous portion of temporal bone	Auditory ossicles
	Incus		
	Stapes		
	Hyoid		

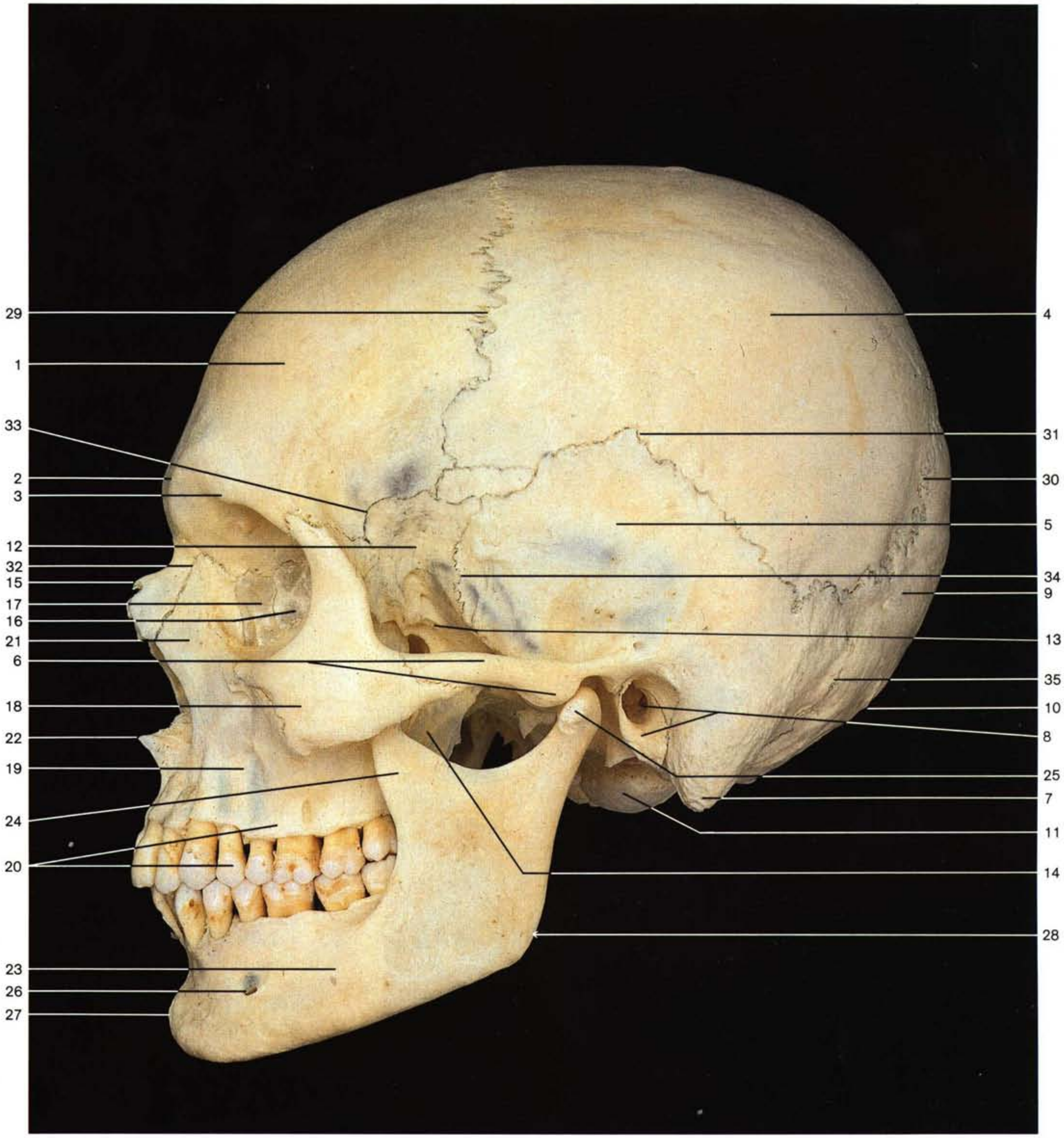

Lateral aspect of the skull.

1 **Frontal bone**
2 Glabella
3 Supraorbital margin
4 **Parietal bone**
5 **Temporal bone** (squamous part)
6 Zygomatic process (articular tubercle)
7 Mastoid process
8 Tympanic part (tympanic plate) and external acoustic meatus
9 **Occipital bone** (squamous part)
10 External occipital protuberance
11 Occipital condyle
12 **Sphenoid bone** (greater wing)
13 Infratemporal crest of sphenoid
14 Pterygoid process (lateral pterygoid plate)
15 **Nasal bone**
16 **Ethmoid bone** (orbital part)
17 **Lacrimal bone**
18 **Zygomatic bone**
19 **Maxilla** (body)
20 Alveolar process and teeth
21 Frontal process
22 Anterior nasal spine
23 **Mandible** (body)
24 Coronoid process
25 Condylar process
26 Mental foramen
27 Mental protuberance
28 Angle of the mandible

Sutures

29 Coronal suture
30 Lambdoid suture
31 Squamous suture
32 Nasomaxillary suture
33 Frontosphenoid suture
34 Sphenosquamosal suture
35 Occipitomastoid suture

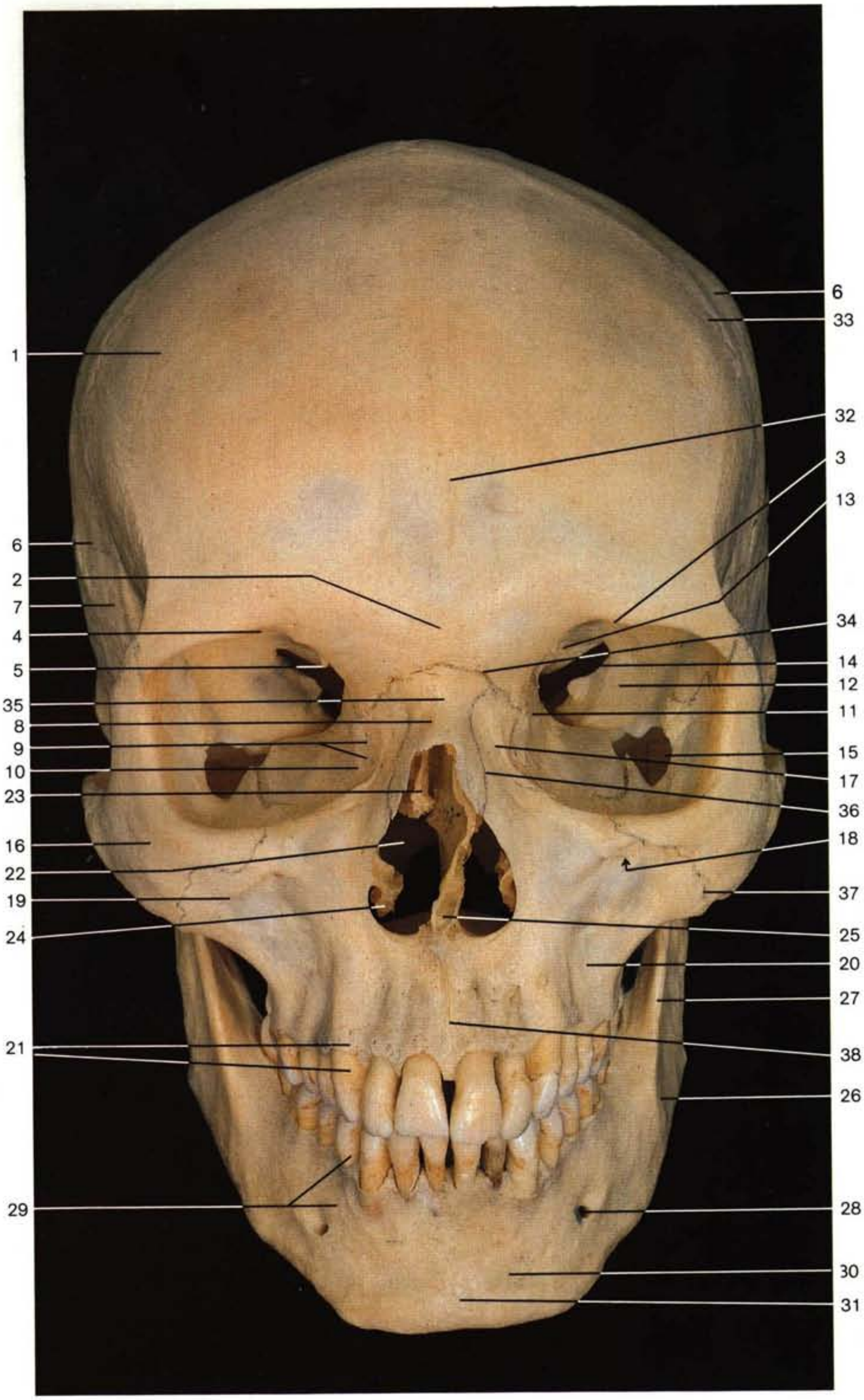

1 **Frontal bone**
2 Glabella
3 Supraorbital margin
4 Supraorbital notch
5 Trochlear spine
6 **Parietal bone**
7 **Temporal bone**
8 **Nasal bone**

Orbit
9 Lacrimal bone
10 Posterior lacrimal crest
11 Ethmoid bone

Sphenoid bone
12 Greater wing of sphenoid bone
13 Lesser wing of sphenoid bone
14 Superior orbital fissure
15 Inferior orbital fissure
16 **Zygomatic bone**

Maxilla
17 Frontal process
18 Infraorbital foramen
19 Zygomatic process
20 Body of maxilla
21 Alveolar process with teeth

Nasal cavity
22 Anterior nasal aperture
23 Middle nasal concha
24 Inferior nasal concha
25 Nasal septum, vomer

Mandible
26 Body of mandible
27 Ramus of mandible
28 Mental foramen
29 Alveolar part with teeth
30 Base of mandible
31 Mental protuberance

Sutures
32 Frontal suture
33 Coronal suture
34 Frontonasal suture
35 Internasal suture
36 Nasomaxillary suture
37 Zygomaticomaxillary suture
38 Intermaxillary suture

Anterior aspect of the skull.

The skull comprises a mosaic of numerous complicated bones which form the cranial cavity protecting the brain **(neurocranium)** and several cavities such as nasal and oral cavities in the facial region. The neurocranium consists of large bony plates which develop directly from the surrounding sheets of connective tissue **(desmocranium).** The bones of the skull base are formed out of cartilaginous tissue **(chondrocranium)** which ossifies secondarily. The **visceral skeleton** which, in fish, gives rise to the gills, has in higher vertebrates been transformed into the bones of the masticatory and auditory apparatus (maxilla, mandible, auditory ossicles and hyoid bone).

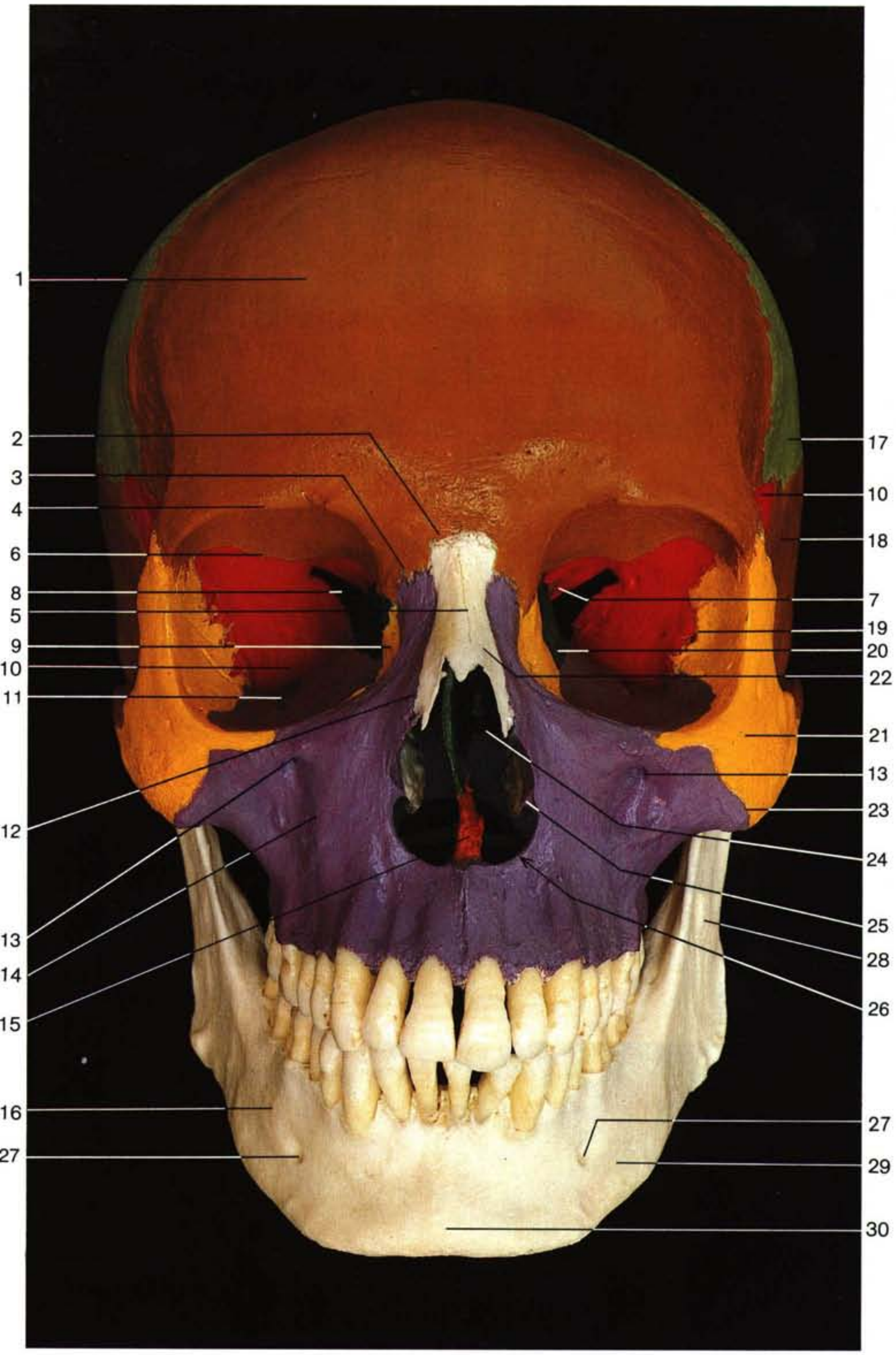

1 **Frontal bone**
2 Frontonasal suture
3 Frontomaxillary suture
4 Supraorbital margin
5 Internasal suture
6 Sphenofrontal suture
7 Optic canal in lesser wing of sphenoid bone
8 Superior orbital fissure
9 **Lacrimal bone**
10 **Sphenoid bone** (greater wing)
11 Inferior orbital fissure
12 Nasomaxillary suture
13 Infraorbital foramen
14 **Maxilla**
15 **Vomer**
16 Body of mandible
17 **Parietal bone**
18 **Temporal bone**
19 Sphenozygomatic suture
20 **Ethmoid bone**
21 **Zygomatic bone**
22 **Nasal bone**
23 Zygomaticomaxillary suture
24 Middle nasal concha
25 Inferior nasal concha
26 Anterior nasal aperture
27 Mental foramen
28 Ramus of mandible
29 Base of mandible
30 Mental protuberance

Bones
Frontal bone (brown)
Parietal bone (light green)
Temporal bone (dark brown)
Sphenoidal bone (red)
Zygomatic bone (yellow)
Ethmoid bone (dark green)
Lacrimal bone (yellow)
Vomer (orange)
Maxilla (violet)
Nasal bone (white)
Mandible (white)

Anterior aspect of the skull (individual bones indicated by color).

The following series of figures are arranged so that the mosaic-like pattern of the skull becomes understandable. It starts with the bones of the **skull base** (sphenoid and occipital) to which the other bones are added step by step. The facial skeleton is built up by the ethmoid bone to which the palatine bone and maxilla are attached laterally; the small nasal and lacrimal bones fill the remaining spaces. Cartilages remain only in the external part of the nose.

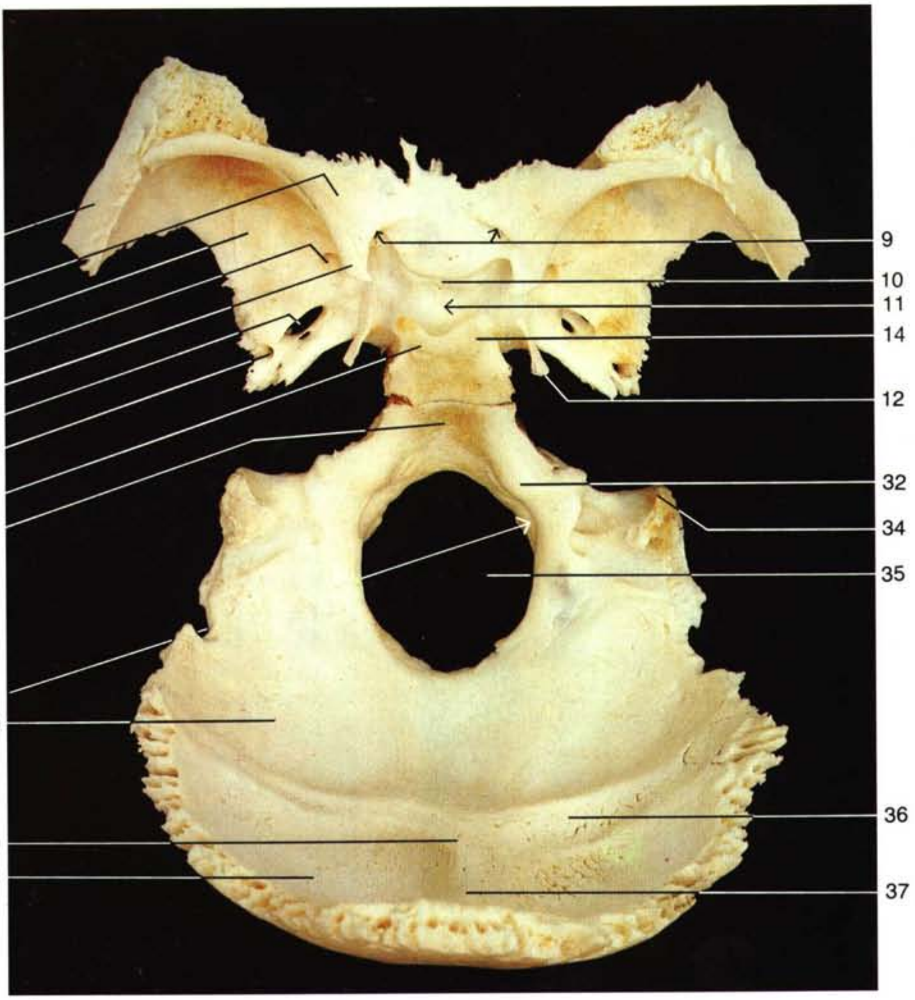
1
2
3
4
5
6
7
8
27
28
29
30
31
9
10
11
14
12
32
34
35
36
37

Sphenoid and occipital bone (from above).

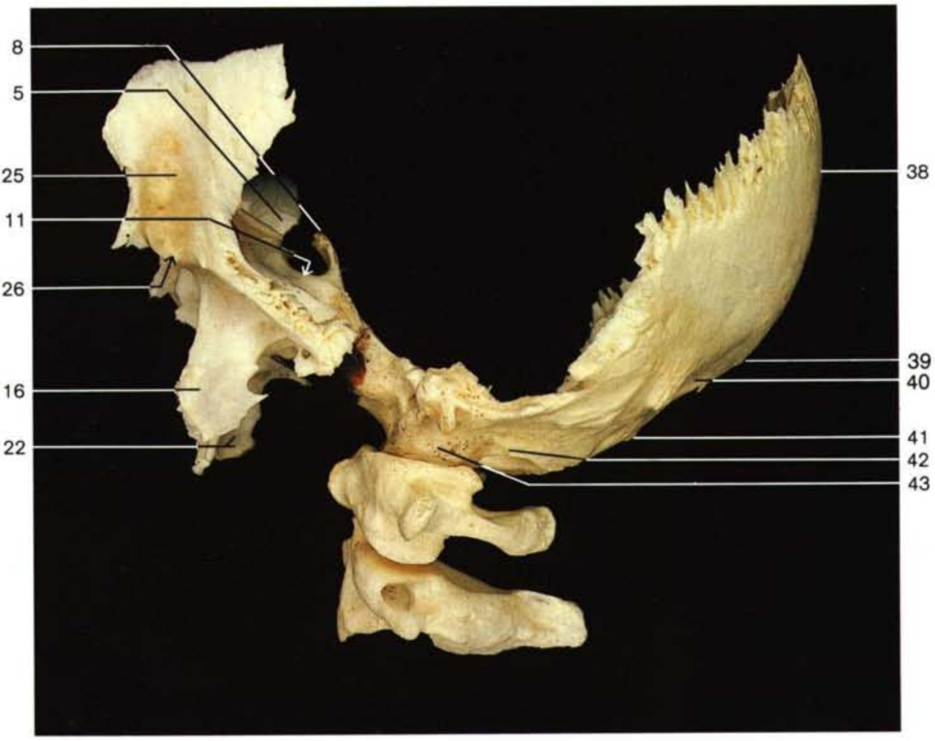
8
5
25
11
26
16
22
38
39
40
41
42
43

Sphenoid and occipital bone in connection with the atlas and axis (1st and 2nd cervical vertebrae) (left lateral view).

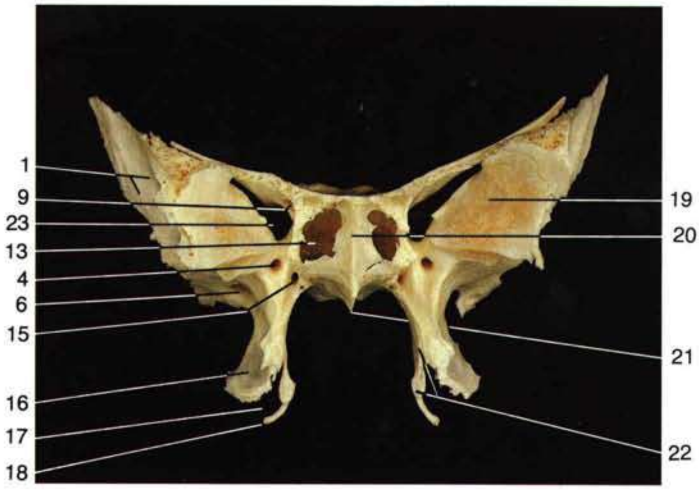

Sphenoid bone (anterior aspect).

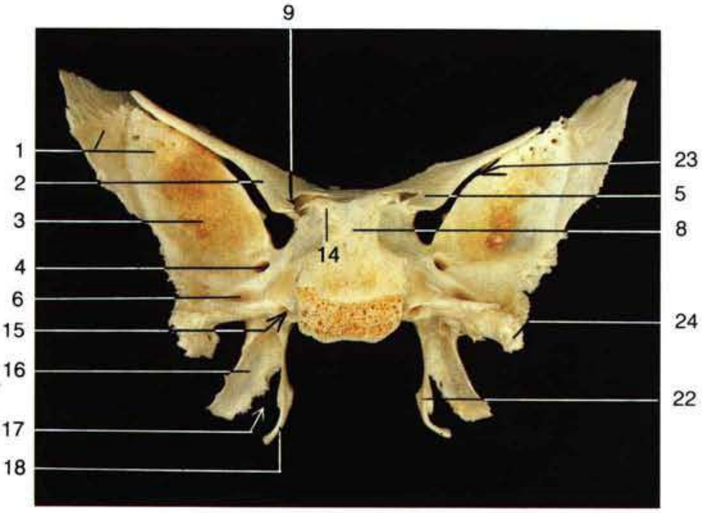

Sphenoid bone (posterior aspect).

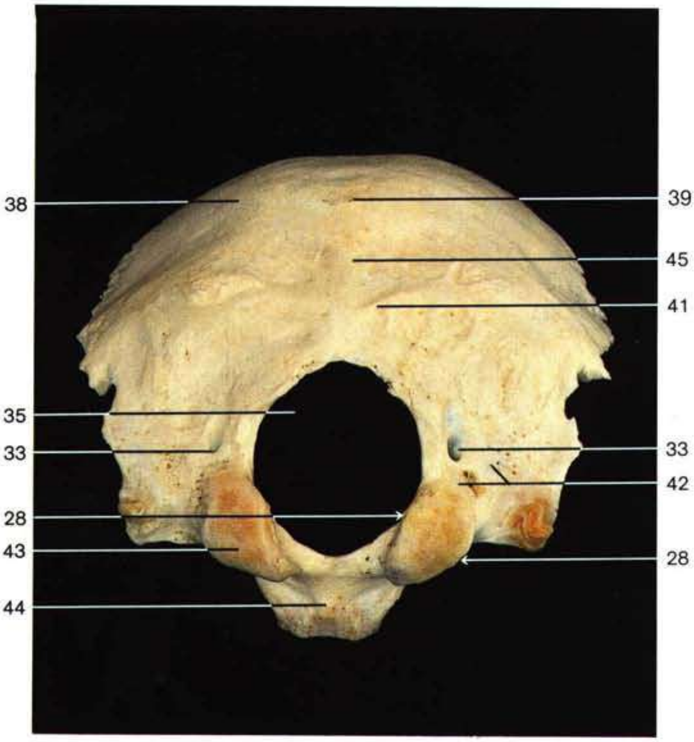

Occipital bone (from below).

Sphenoid bone

1 **Greater wing**
2 **Lesser wing**
3 Cerebral or superior surface of greater wing
4 **Foramen rotundum**
5 Anterior clinoid process
6 Foramen ovale
7 Foramen spinosum
8 Dorsum sellae
9 Optic canal
10 Chiasmatic groove (sulcus chiasmatis)
11 **Hypophysial fossa (sella turcica)**
12 Lingula
13 Opening of sphenoidal sinus
14 Posterior clinoid process
15 **Pterygoid canal**
16 Lateral pterygoid plate of **pterygoid process**
17 Pterygoid notch
18 Pterygoid hamulus
19 Orbital surface of greater wing
20 Sphenoid crest
21 Sphenoid rostrum
22 Medial pterygoid plate
23 Superior orbital fissure
24 Spine of sphenoid
25 Temporal surface of greater wing
26 Infratemporal crest

Occipital bone

27 **Clivus** with basilar part of occipital bone
28 **Hypoglossal canal**
29 Fossa for cerebellar hemisphere
30 Internal occipital protuberance
31 Fossa for cerebral hemisphere
32 Jugular tubercle
33 Condylar canal
34 Jugular process
35 **Foramen magnum**
36 Groove for transverse sinus
37 Groove for superior sagittal sinus
38 **Squamous part** of the occipital bone
39 External occipital protuberance
40 Superior nuchal line
41 Inferior nuchal line
42 Condylar fossa
43 **Condyle**
44 Pharyngeal tubercle
45 External occipital crest

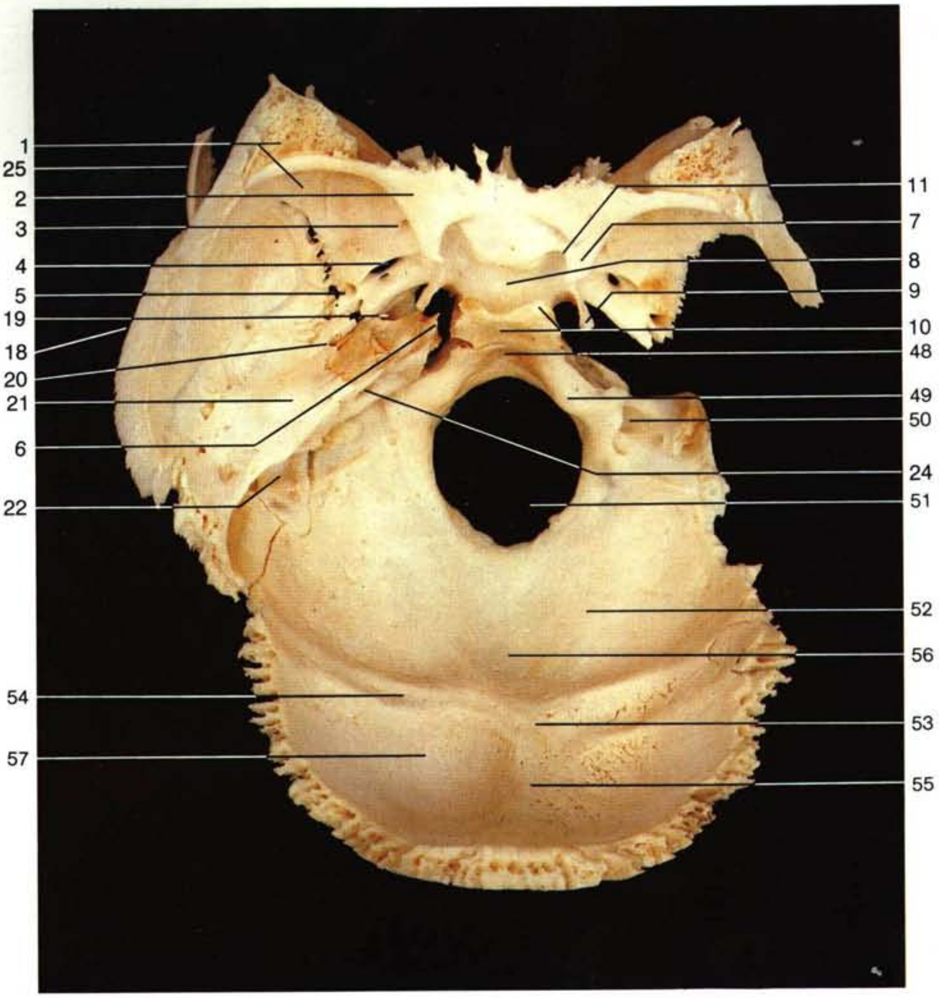

Sphenoid, occipital and left temporal bone (from above). Internal aspect of the base of the skull. The left temporal bone has been added to the preceeding figure.

Sphenoid bone

1 **Greater wing**
2 **Lesser wing**
3 Foramen rotundum
4 **Foramen ovale**
5 Foramen spinosum
6 Foramen lacerum
7 Anterior clinoid process
8 Hypophysial fossa (**sella turcica**)
9 Lingula
10 Dorsum sellae and posterior clinoid process
11 **Optic canal**
12 Sphenoid rostrum
13 Medial pterygoid plate
14 Lateral pterygoid plate
15 Pterygoid hamulus
16 Infratemporal crest
17 Body of the sphenoid

Temporal bone

18 **Squamous part**
19 **Carotid canal**
20 Hiatus of facial canal (for the greater petrosal nerve)
21 Arcuate eminence
22 Groove for the sigmoid sinus
23 Mastoid foramen
24 **Internal acoustic meatus**
25 Zygomatic process
26 Mandibular fossa
27 Petrotympanic fissure
28 Canalis musculotubarius (bony part of auditory tube)
29 **External acoustic meatus**
30 Styloid process (remnant only)
31 Stylomastoid foramen
32 Mastoid canaliculus
33 Jugular fossa
34 **Mastoid process**
35 Mastoid notch

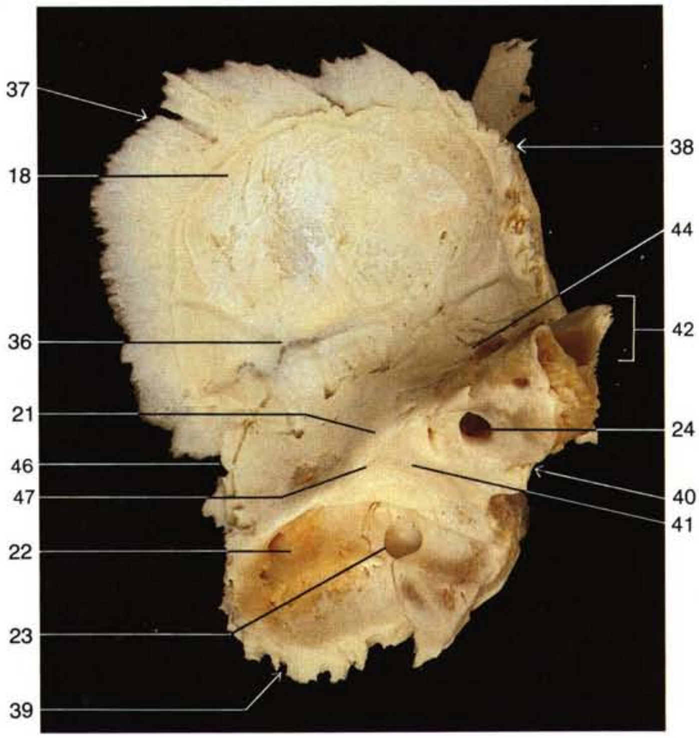

Left temporal bone (medial aspect).

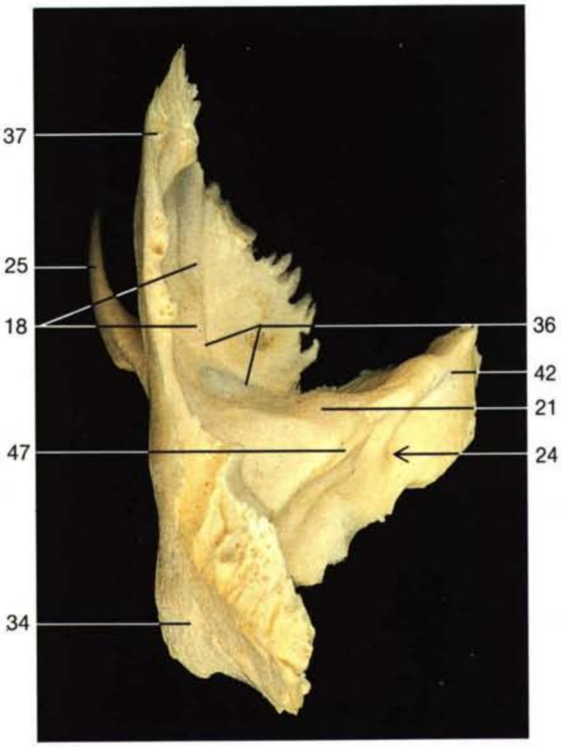

Left temporal bone (from above).

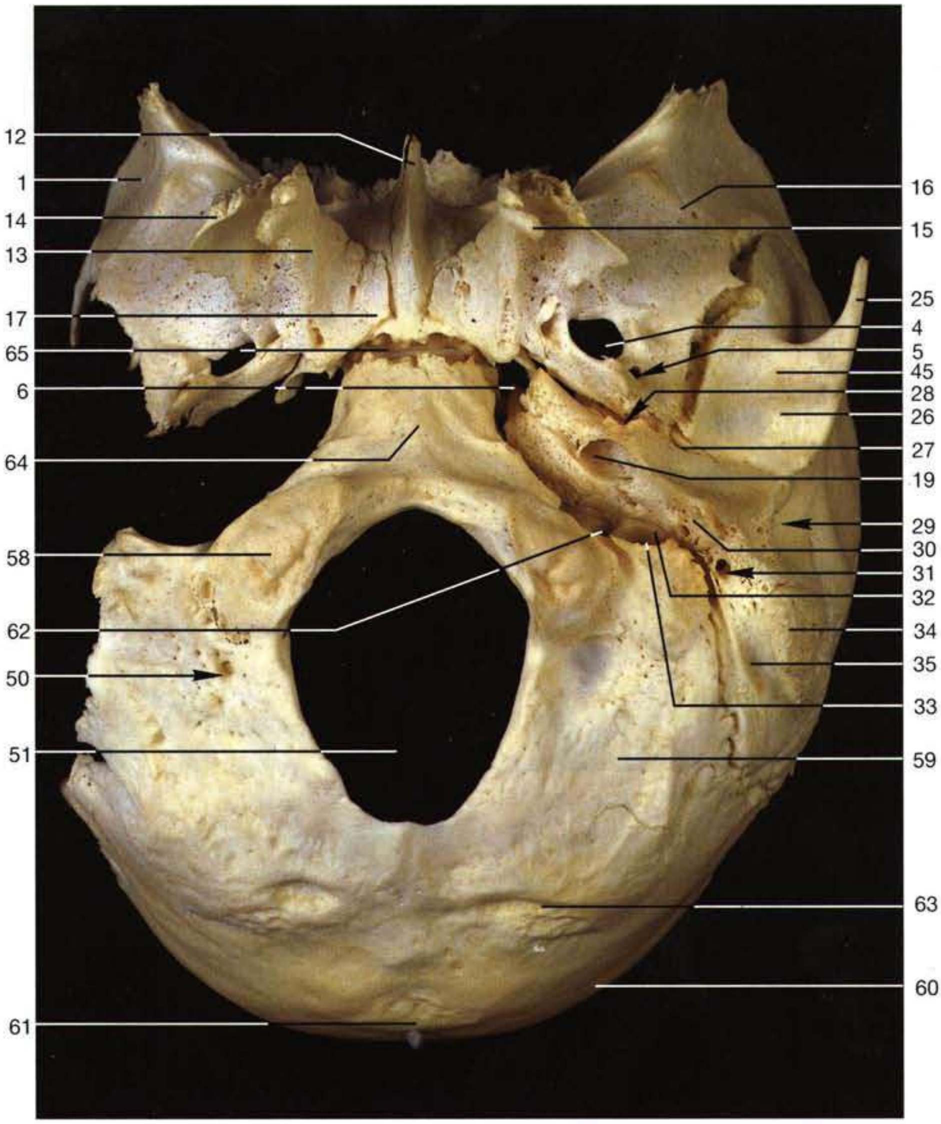
12
1
14
13
17
65
6
64
58
62
50
51
61
16
15
25
4
5
45
28
26
27
19
29
30
31
32
34
35
33
59
63
60

Sphenoid, occipital and left temporal bone.
Base of the skull (external aspect).

36 Groove for middle meningeal vessels
37 Parietal margin
38 Sphenoid margin
39 Occipital margin
40 Cochlear canaliculus
41 Aqueduct of the vestibule
42 Apex of the petrous part
43 **Tympanic part**
44 Trigeminal impression
45 **Articular tubercle**
46 Parietal notch
47 Groove for the superior petrosal sinus

Occipital bone
48 **Clivus**
49 Jugular tubercle
50 Condylar canal
51 **Foramen magnum**
52 Lower part of squamous occipital bone (cerebellar fossa)
53 Internal occipital protuberance
54 Groove for the transverse sinus
55 Groove for the superior sagittal sinus
56 Internal occipital crest
57 Upper part of squamous occipital bone (cerebral fossa)
58 **Condyle**
59 Nuchal plane
60 Superior nuchal line
61 **External occipital protuberance**
62 **Jugular foramen**
63 Inferior nuchal line
64 Pharyngeal tubercle
65 **Sphenooccipital synchondrosis**

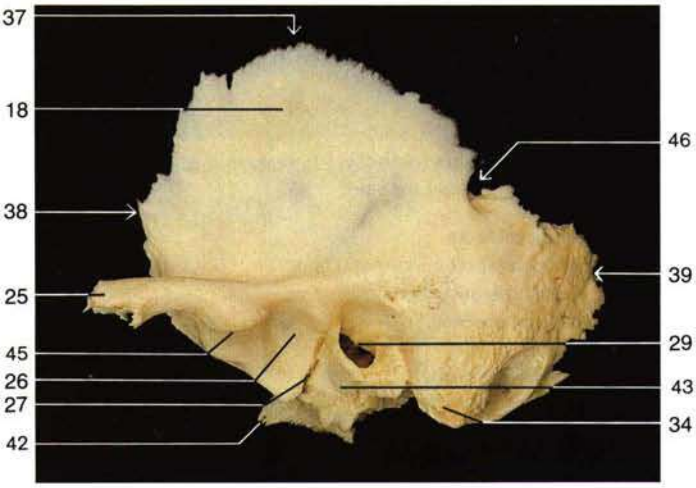
37
18
38
25
45
26
27
42
46
39
29
43
34

Left temporal bone (lateral aspect).

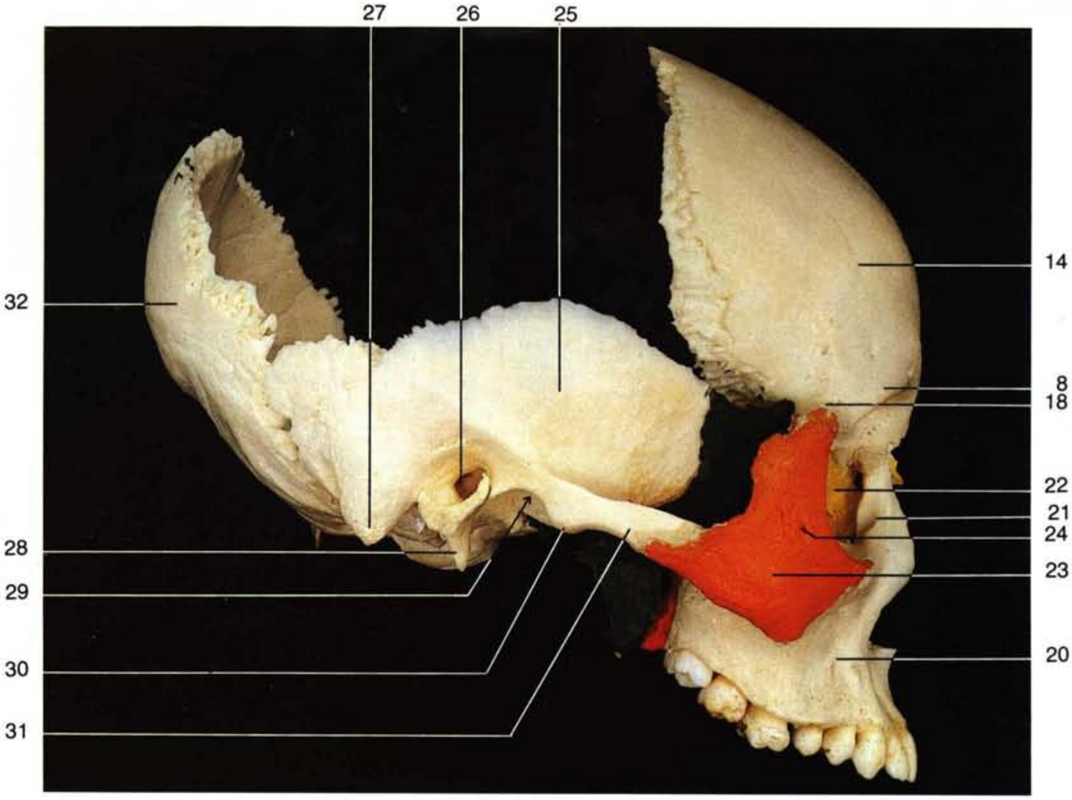

Part of a disarticulated skull (right lateral aspect). The **frontal bone** and the maxilla are connected with the temporal bone by the zygomatic bone (orange). Sphenoid bone (green), palatine bone (red), lacrimal bone (yellow).

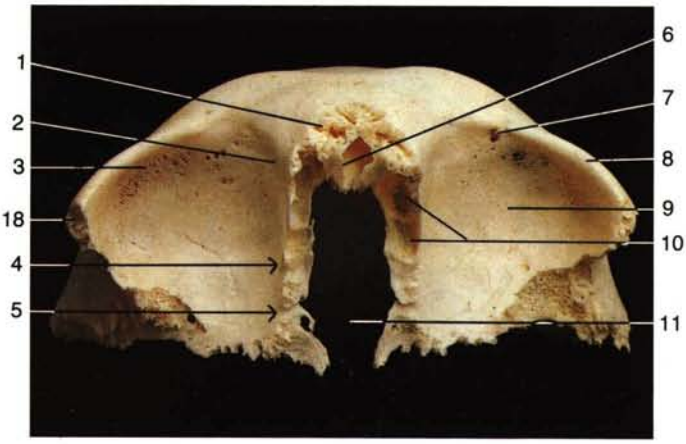

Frontal bone (inferior aspect). The ethmoidal foveolae cover the ethmoidal cavities of the ethmoid bone.

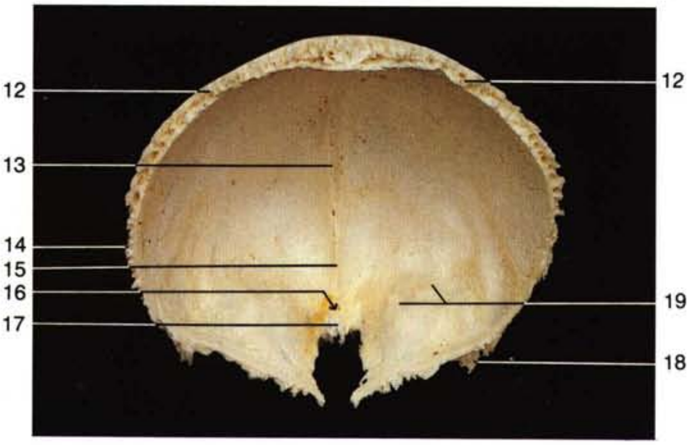

Frontal bone (posterior aspect).

Frontal bone

1 Nasal margin
2 Trochlear fossa
3 Fossa for lacrimal gland
4 Anterior ethmoidal foramen
5 Posterior ethmoidal foramen
6 Nasal spine
7 Supraorbital notch
8 Supraorbital margin
9 **Orbital plate**
10 Roofs of the ethmoidal air cells
11 Ethmoidal notch
12 Parietal margin
13 Groove for superior sagittal sinus
14 **Squamous part** of frontal bone
15 Frontal crest
16 Foramen cecum
17 Nasal spine
18 **Zygomatic process** of frontal bone
19 Juga cerebralia

20 **Maxilla**
21 Frontal process of maxilla
22 **Lacrimal bone** (yellow)
23 **Zygomatic bone** (orange)
24 Zygomaticofacial foramen

Temporal bone

25 Squamous part of temporal bone
26 External acoustic meatus
27 **Mastoid process**
28 **Styloid process**
29 **Mandibular fossa**
30 Articular tubercle
31 Zygomatic process

Occipital bone

32 Squamous part of occipital bone

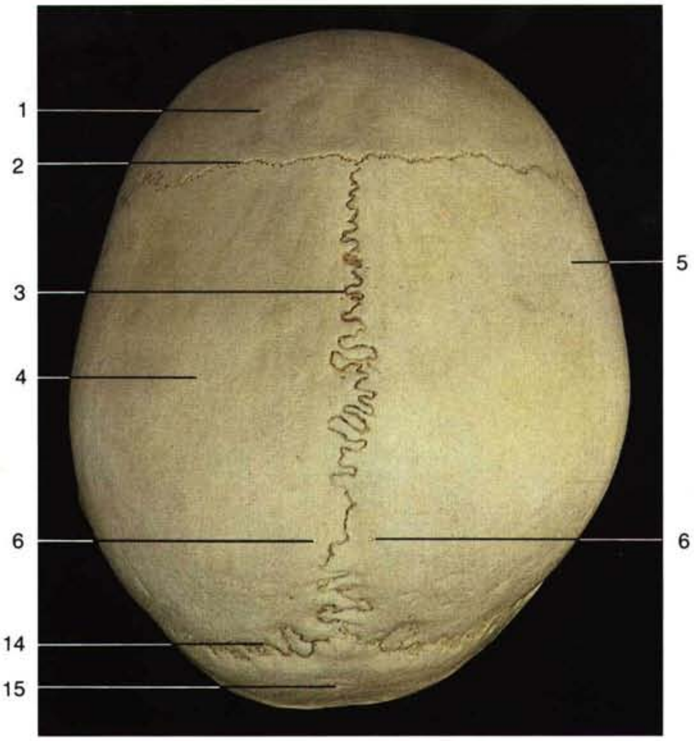

Calvaria (superior aspect).

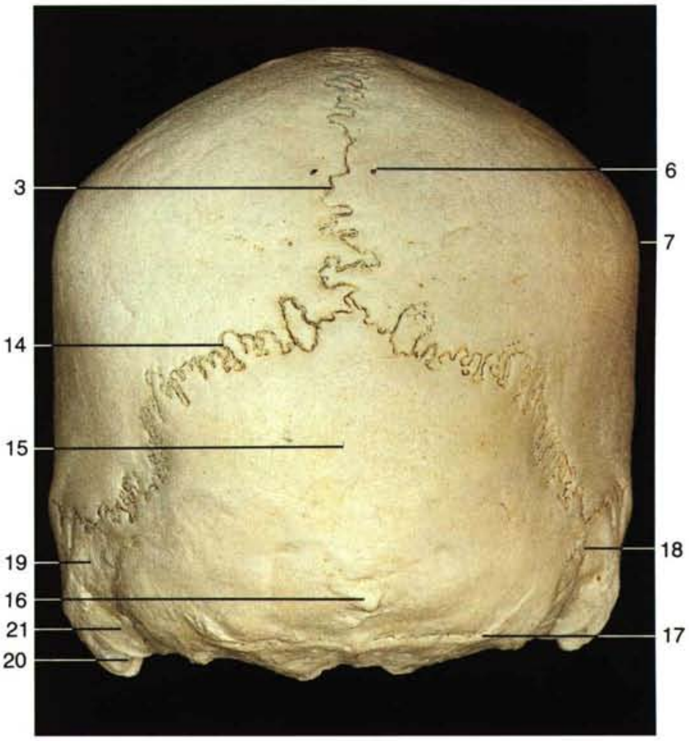

Calvaria (posterior aspect).

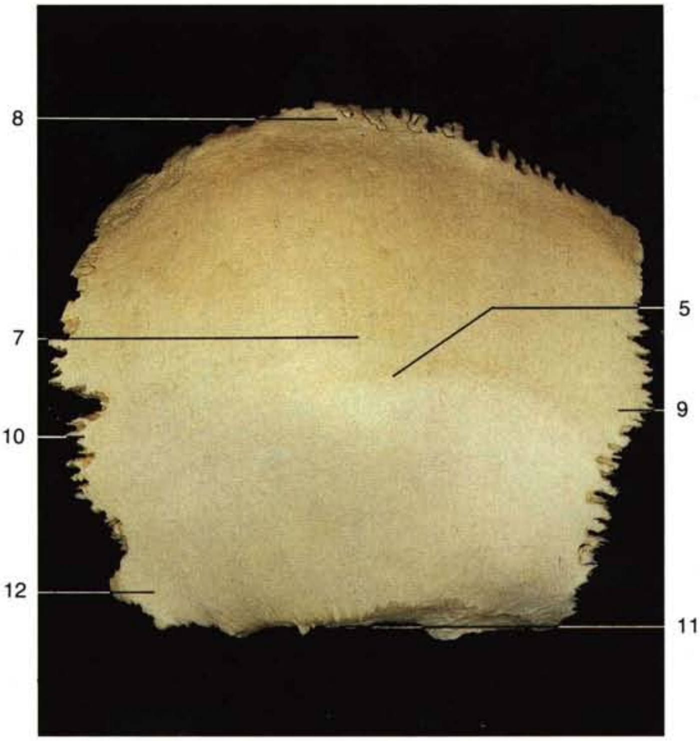

Left parietal bone (external aspect).

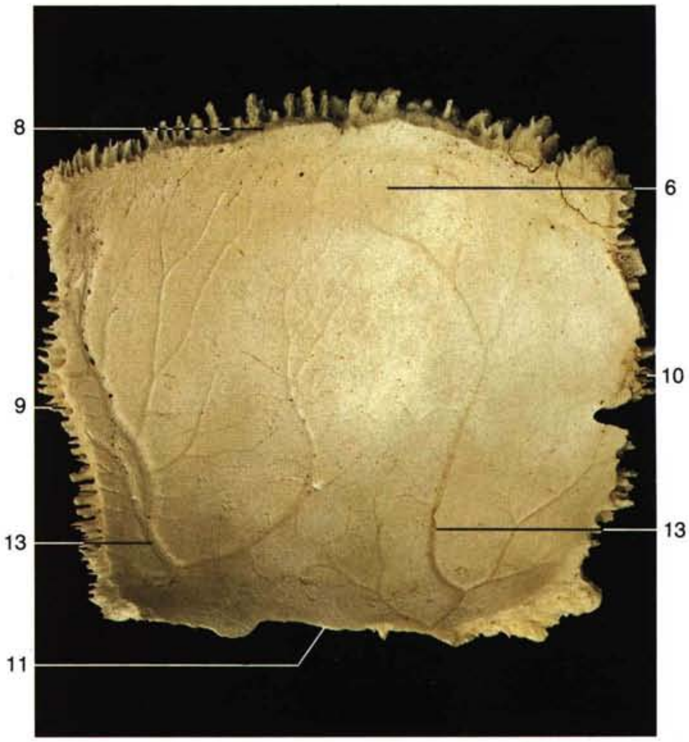

Left parietal bone (internal aspect).

1 **Frontal bone**
2 Coronal suture
3 Sagittal suture
4 **Parietal bone**
5 Superior temporal line
6 Parietal foramen
7 Parietal tuber or eminence
8 Sagittal margin
9 Occipital margin
10 Frontal margin
11 Squamous margin
12 Sphenoidal angle
13 Groove for middle meningeal artery
14 Lambdoid suture
15 **Occipital bone**
16 External occipital protuberance
17 Inferior nuchal line
18 Occipitomastoid suture
19 **Temporal bone**
20 Mastoid process
21 Mastoid notch

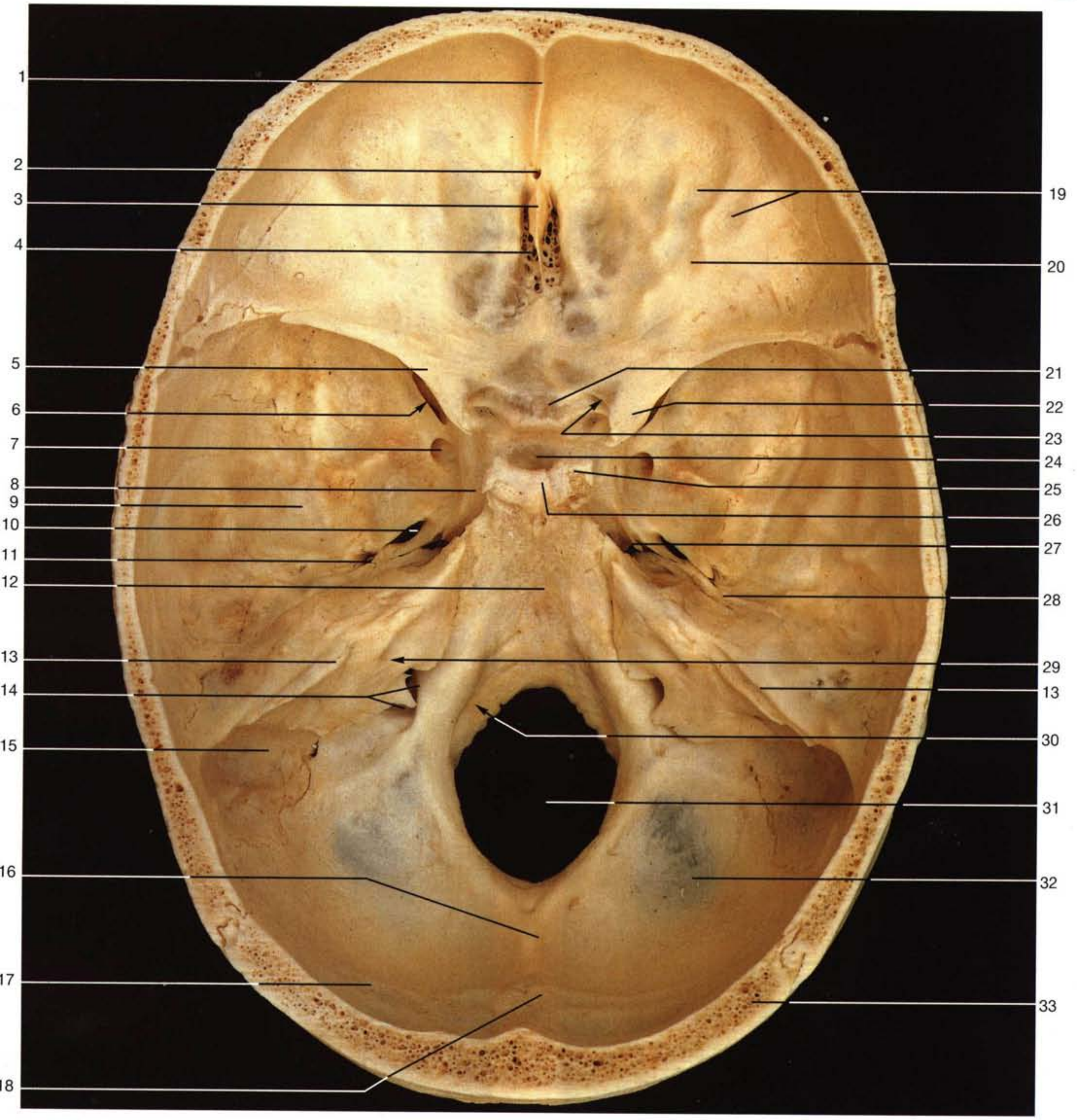

Base of the skull, calvaria removed (internal aspect).

1	Frontal crest
2	Foramen cecum
3	**Crista galli**
4	**Cribriform plate** of ethmoid bone
5	Lesser wing of sphenoid bone
6	Superior orbital fissure
7	**Foramen rotundum**
8	**Carotid sulcus**
9	Middle cranial fossa
10	**Foramen ovale**
11	Foramen spinosum
12	**Clivus**
13	Groove for superior petrosal sinus
14	**Jugular foramen**
15	Groove for sigmoid sinus
16	Internal occipital crest
17	Groove for transverse sinus
18	Internal occipital protuberance
19	Marking of brain convolutions
20	Anterior cranial fossa
21	Chiasmatic sulcus
22	Anterior clinoid process
23	**Optic canal**
24	**Sella turcica** (hypophysial fossa)
25	Posterior clinoid process
26	Dorsum sellae
27	**Foramen lacerum**
28	Groove for greater petrosal nerve
29	Internal acoustic meatus
30	**Hypoglossal canal**
31	**Foramen magnum**
32	Posterior cranial fossa
33	Diploe

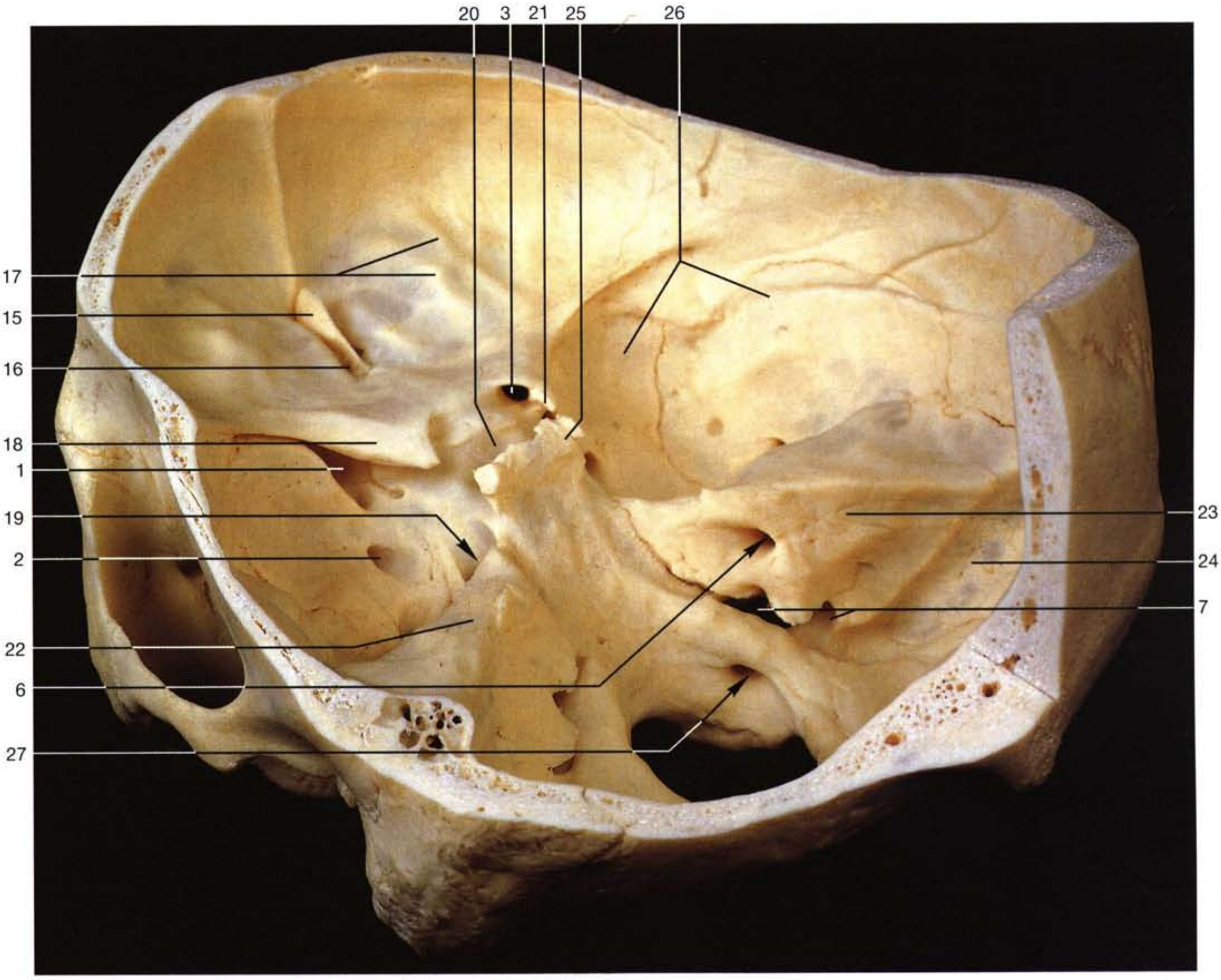

Base of the skull (internal aspect, oblique lateral view from left side).

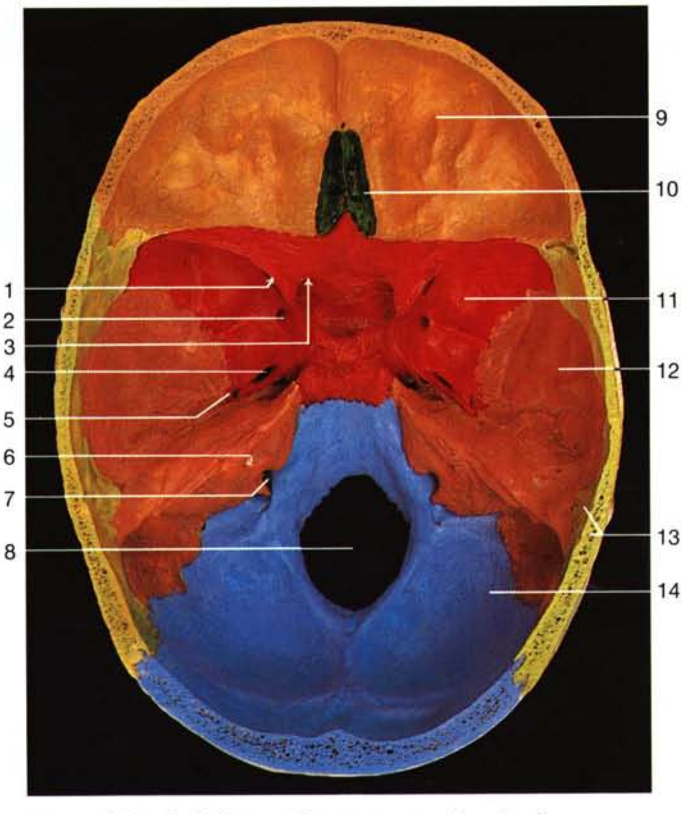

Base of the skull (internal aspect, superior view). Individual bones indicated by color.

Canals, fissures and foramina of the base of the skull

1 Superior orbital fissure
2 Foramen rotundum
3 Optic canal
4 Foramen ovale
5 Foramen spinosum
6 Internal acoustic meatus
7 Jugular foramen
8 Foramen magnum

Bones

9 Frontal bone (orange)
10 Ethmoid bone (dark green)
11 Sphenoid bone (red)
12 Temporal bone (brown)
13 Parietal bone (yellow-green)
14 Occipital bone (blue)

Details of bones

15 Crista galli
16 Cribriform plate
17 Marking of brain convolutions (frontal bone)
18 Lesser wing of sphenoid bone
19 Foramen lacerum
20 Hypophysial fossa (sella turcica)
21 Anterior clinoid process
22 Depression for trigeminal ganglion
23 Petrous part of temporal bone
24 Groove for sigmoid sinus
25 Dorsum sellae (posterior clinoid process)
26 Greater wing of sphenoid bone, groove for middle meningeal artery
27 Hypoglossal canal

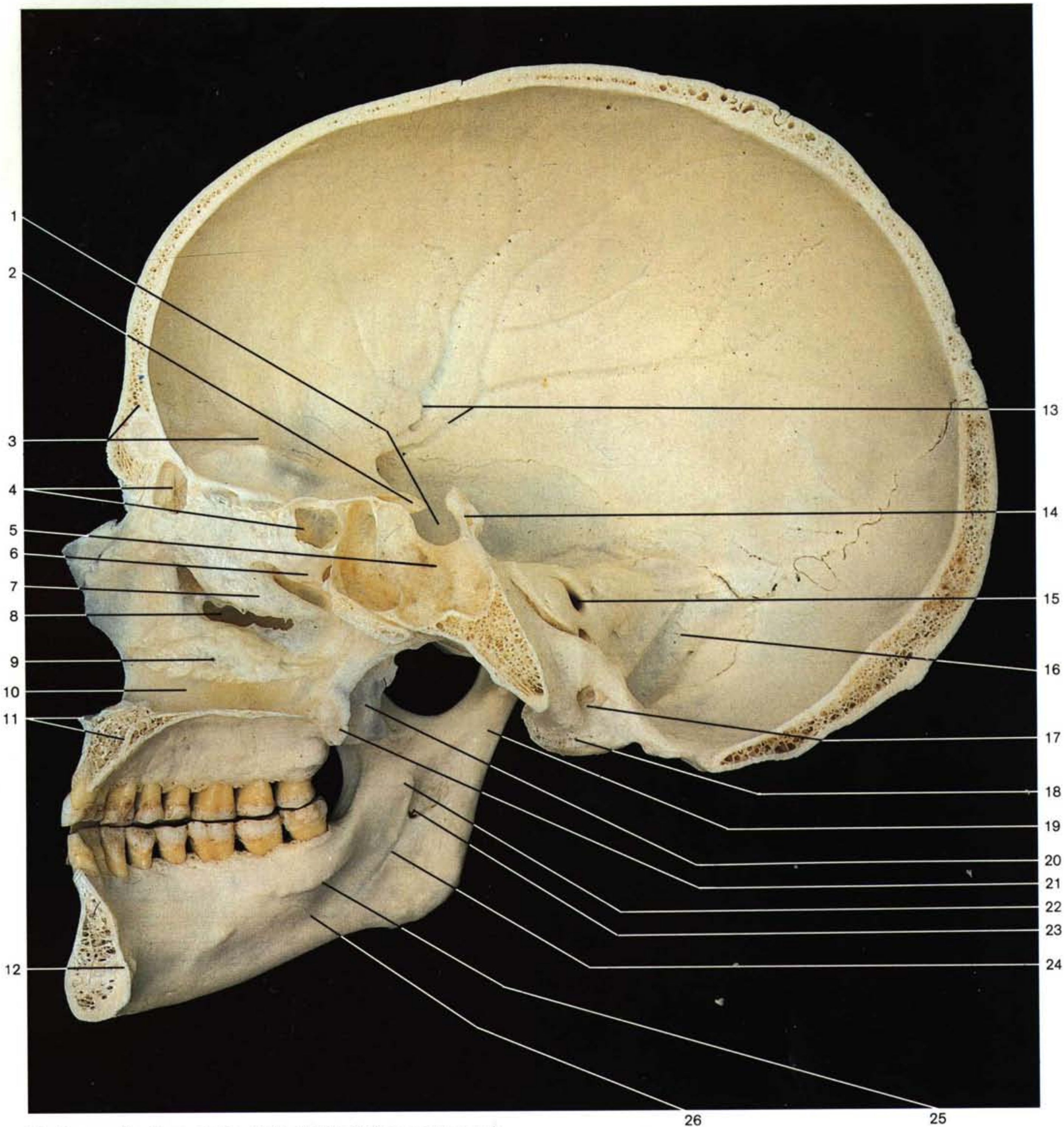

Median section through the skull, right half (internal aspect).

1	**Hypophysial fossa (sella turcica)**
2	Anterior clinoid process
3	**Frontal bone**
4	**Ethmoidal air cells**
5	**Sphenoidal sinus**
6	Superior concha
7	Middle concha
8	**Maxillary hiatus**
9	**Inferior concha**
10	Inferior meatus
11	Anterior nasal spine and maxilla
12	Mental spine or genial tubercle
13	Groove for middle meningeal artery
14	Dorsum sellae
15	Internal acoustic meatus
16	Groove for sigmoid sinus
17	**Hypoglossal canal**
18	**Occipital condyle**
19	Condylar process
20	Lateral pterygoid plate } of **pterygoid process**
21	Medial pterygoid plate } of **pterygoid process**
22	Lingula of mandible
23	**Mandibular foramen**
24	Mylohyoid groove
25	Mylohyoid line
26	Submandibular fovea

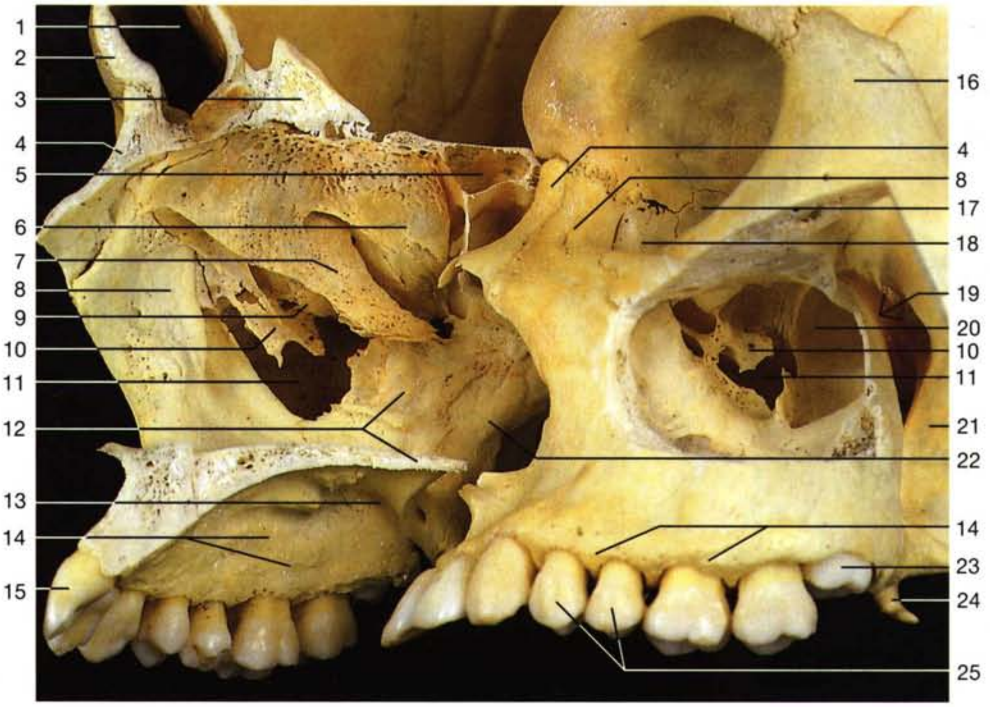

Facial part of the skull (viscerocranium), divided in two halves (lateral and medial aspect). Right inferior concha has been removed to show the maxillary hiatus. Left maxillary sinus opened.

1 Frontal sinus
2 **Frontal bone**
3 Crista galli
4 **Nasal bone**
5 Sphenoidal sinus
6 **Superior concha** } of **ethmoid**
7 **Middle concha** } **bone**
8 Frontal process of maxilla
9 Ethmoidal bulla
10 **Uncinate process**
11 Maxillary hiatus
12 **Palatine bone**
13 Greater palatine foramen
14 **Alveolar process of maxilla**
15 Central incisor
16 Zygomatic bone
17 Ethmoid bone
18 **Lacrimal bone**
19 Pterygopalatine fossa
20 **Maxillary sinus**
21 Lateral pterygoid plate
22 Medial pterygoid plate
23 Third molar tooth
24 Pterygoid hamulus
25 Two premolar teeth

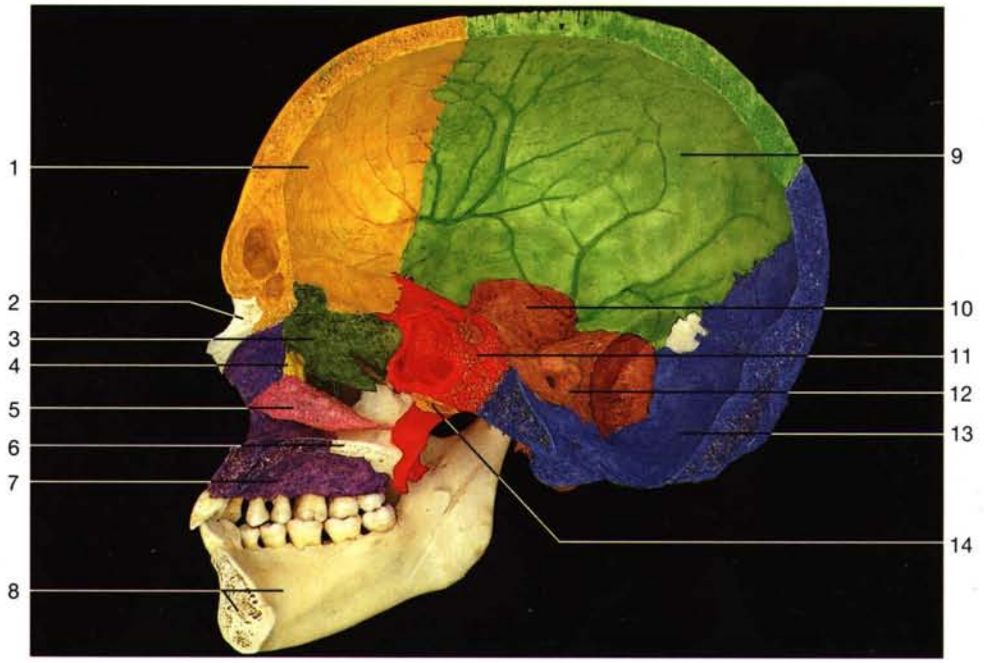

Median section through the skull. The nasal septum has been removed. Bones indicated by colors.

Bones (indicated by colors)
1 Frontal bone (yellow)
2 Nasal bone (white)
3 Ethmoid bone (dark green)
4 Lacrimal bone (yellow)
5 Inferior nasal concha (pink)
6 Palatine bone (white)
7 Maxilla (violet)
8 Mandible (white)
9 Parietal bone (light green)
10 Temporal bone (brown)
11 Sphenoid bone (red)
12 Petrous part of temporal bone (brown)
13 Occipital bone (blue)
14 Ala of vomer (light brown)

Because of the upright posture which the man developed in the course of evolution, the cranial cavity greatly increased in size whereas the facial skeleton decreased. As a result, the base of the skull developed an angulation of about 120° between the clivus and the cribriform plate. The hypophysial fossa containing the pituitary gland lies at the angle formed between these two planes.

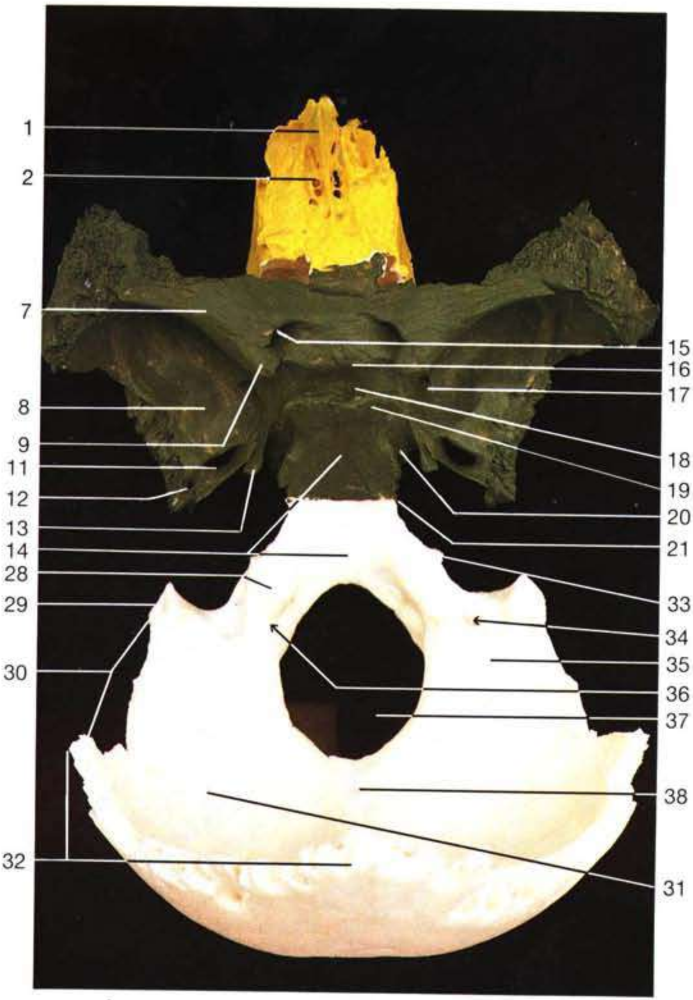

Part of the disarticulated base of the skull.
Ethmoid, sphenoid and occipital bones (from above).
Green = sphenoid bone; yellow = ethmoid bone.

Ethmoid bone
1 Crista galli
2 Cribriform plate
3 Ethmoidal air cells
4 Middle concha
5 Perpendicular plate (part of nasal septum)
6 Orbital plate

Sphenoid bone
7 Lesser wing
8 Greater wing
9 Anterior clinoid process
10 Posterior clinoid process
11 Foramen ovale
12 Foramen spinosum
13 Lingula of the sphenoid
14 Clivus
15 Optic canal
16 Tuberculum sellae
17 Foramen rotundum (right side)
18 Hypophysial fossa (sella turcica)
19 Dorsum sellae
20 Carotid sulcus
21 Sphenooccipital synchondrosis
22 Lateral pterygoid plate
23 Greater wing of sphenoid bone (orbital surface)
24 Greater wing of sphenoid bone (maxillary surface)
25 Foramen rotundum (left side)
26 Superior orbital fissure
27 Infratemporal crest of the greater wing

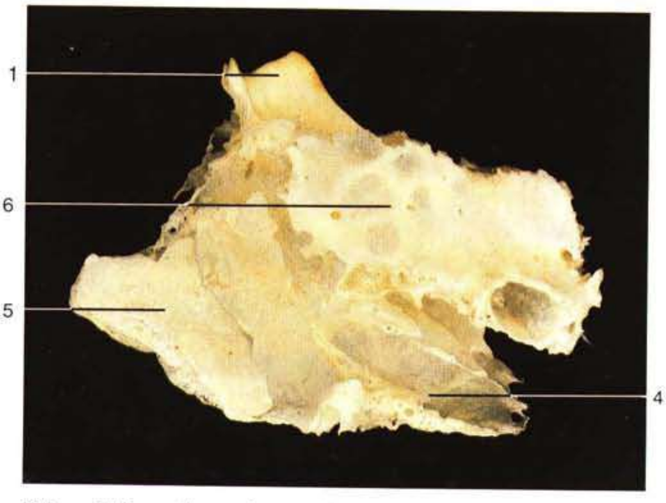

Ethmoid bone (lateral aspect), posterior portion to the right.

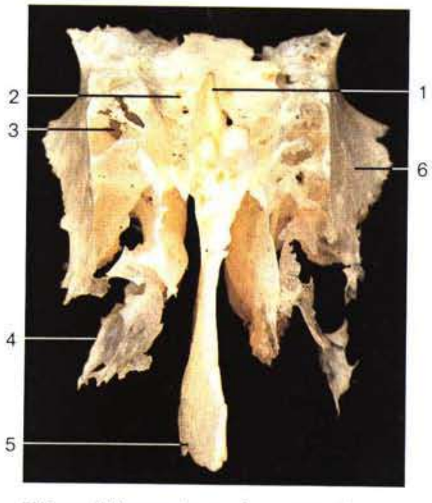

Ethmoid bone (anterior aspect).

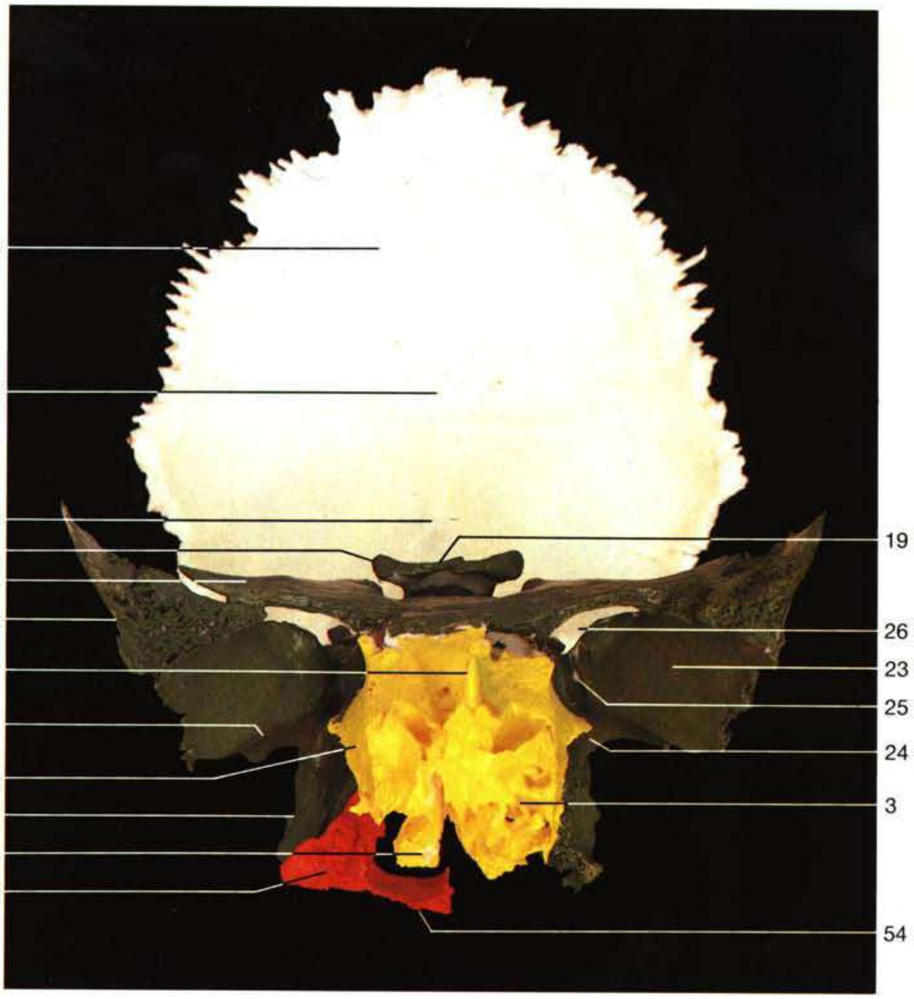

Disarticulated base of the skull (anterior aspect).
Green = sphenoid bone; yellow = ethmoid bone; red = palatine bone.

Occipital bone
28 Jugular tubercle
29 Jugular process
30 Mastoid margin
31 Posterior cranial fossa
32 Lambdoid margin
33 Intrajugular process
34 Condylar canal
35 Lateral part of occipital bone
36 Hypoglossal canal
37 Foramen magnum
38 Internal occipital crest
39 Squamous part of occipital bone
40 Internal occipital protuberance

Maxilla
41 Orbital surface
42 Infraorbital groove
43 Maxillary tuberosity with foramina
44 Frontal process
45 Nasolacrimal groove
46 Infraorbital margin
47 Anterior nasal spine
48 Zygomatic process
49 Alveolar process

Palatine bone
50 Orbital process
51 Sphenopalatine notch
52 Sphenoidal process
53 Perpendicular plate
54 Horizontal plate
55 Pyramidal process

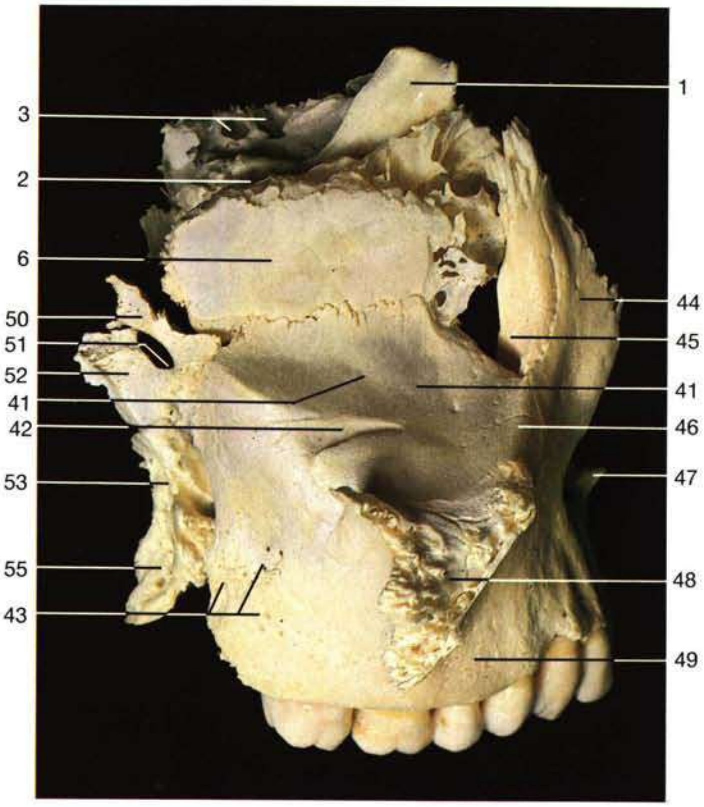

Right maxilla, ethmoid and palatine bone (lateral aspect).

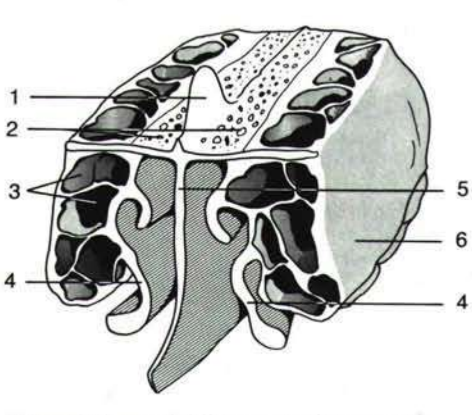

Ethmoid bone (oblique anterior aspect).
(Schematic drawing.)

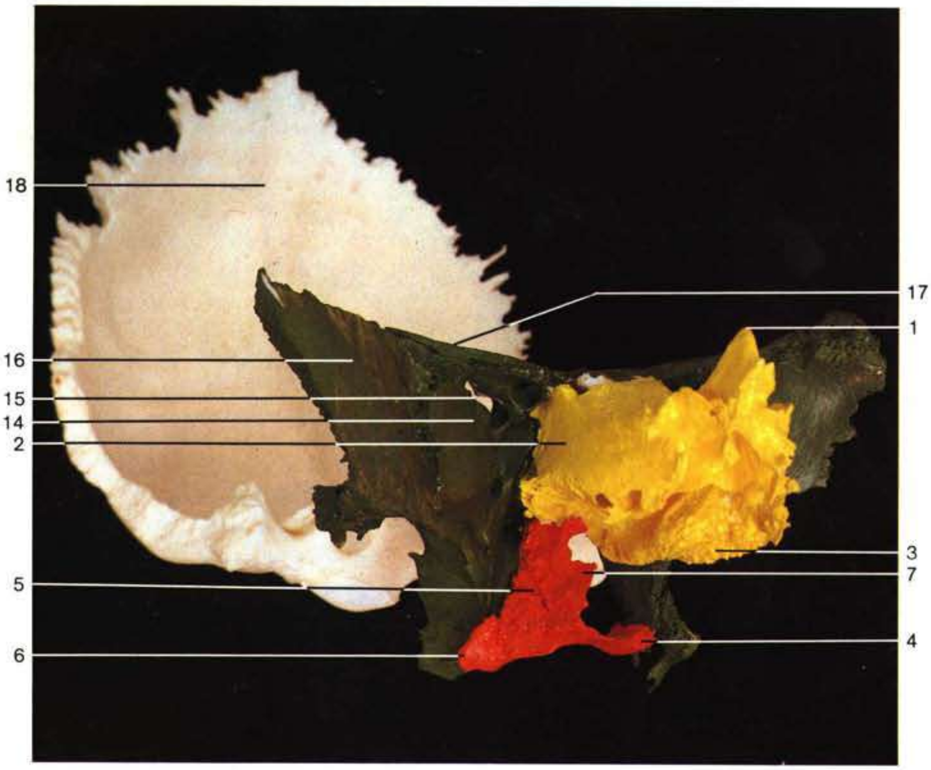

Part of a disarticulated skull base, similar to the foregoing figures, but with palatine bone. Green = sphenoid bone; yellow = ethmoid bone; red = palatine bone.

Ethmoid bone

1 Crista galli
2 Orbital plate
3 Middle concha

Palatine bone

4 Horizontal plate of palatine bone
5 Greater palatine canal
6 Pyramidal process
7 Maxillary process
8 Orbital process
9 Sphenopalatine notch
10 Perpendicular plate of palatine bone
11 Conchal crest
12 Nasal crest
13 Sphenoidal process

Sphenoid bone

14 Greater wing
15 Superior orbital fissure
16 Greater wing (orbital surface)
17 Lesser wing

Occipital bone

18 Squamous part of occipital bone

Maxilla

19 Maxillary tuberosity
20 Frontal process
21 Orbital surface
22 Infraorbital margin
23 Infraorbital groove
24 Zygomatic process
25 Alveolar process

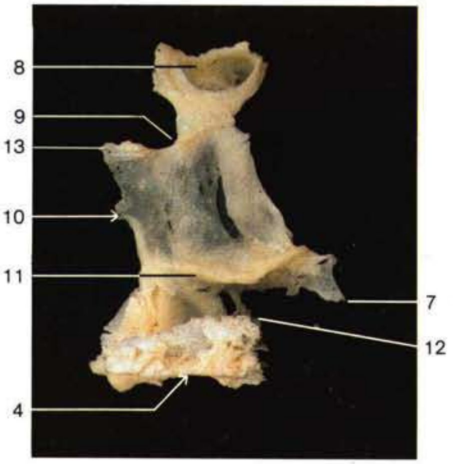

Left palatine bone (medial aspect, posterior aspect to the left).

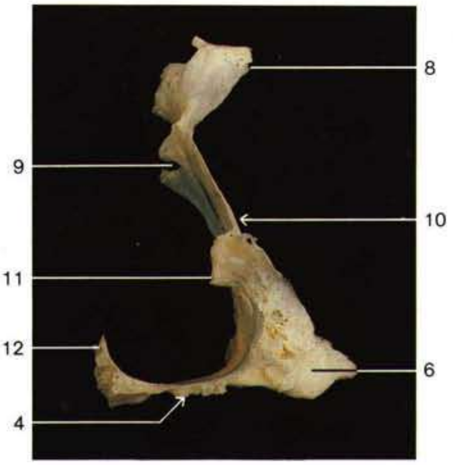

Left palatine bone (anterior aspect).

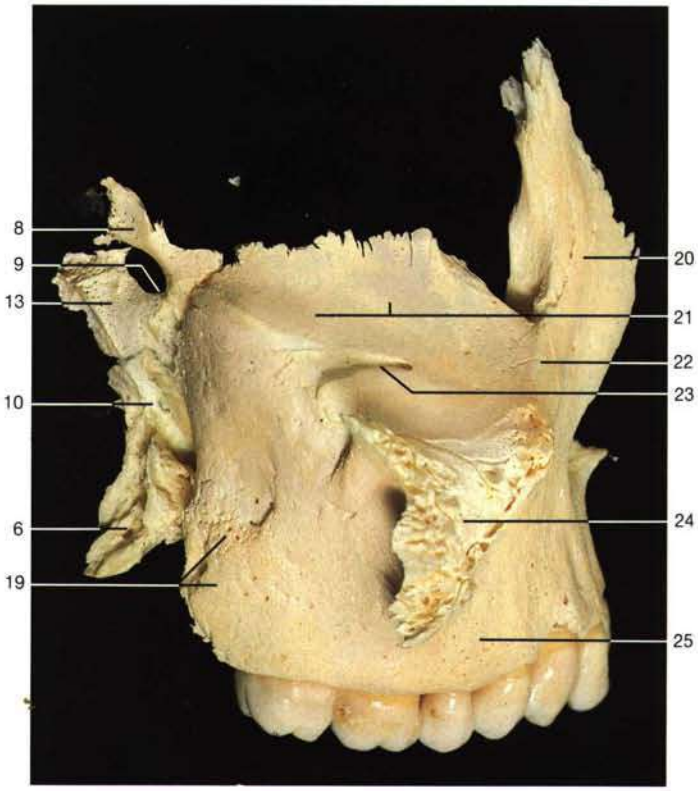

Right maxilla and right palatine bone (lateral aspect).

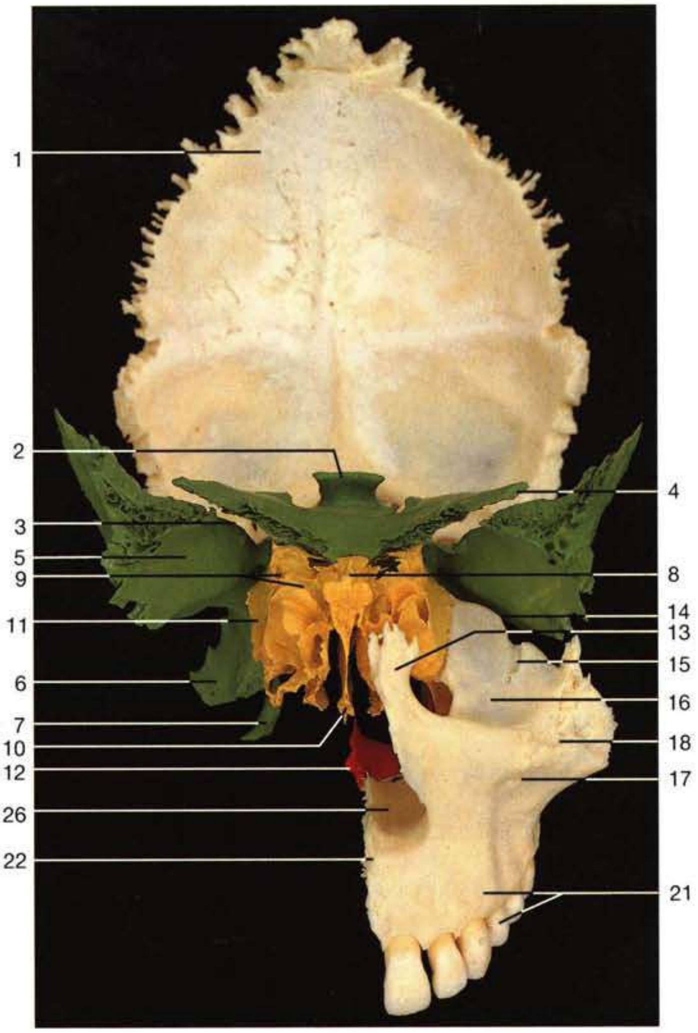

Part of a disarticulated skull.
The left **maxilla** is added to the preceding specimen.

Occipital bone
1 Squamous part

Sphenoid bone
2 Dorsum sellae
3 Superior orbital fissure
4 Lesser wing
5 Greater wing (orbital surface)
6 Lateral pterygoid plate
7 Medial pterygoid plate

Ethmoid bone
8 Crista galli
9 Ethmoidal air cells
10 Perpendicular plate
11 Orbital plate

Palatine bone
12 Horizontal plate (nasal crest)

Maxilla
13 Frontal process
14 Inferior orbital fissure
15 Infraorbital groove
16 Orbital surface
17 Infraorbital foramen
18 Zygomatic process
19 Anterior lacrimal crest
20 Canine fossa
21 Alveolar process with teeth
22 Anterior nasal spine
23 Juga alveolaria (elevations formed by roots of teeth)
24 Lacrimal groove
25 Maxillary tuberosity with alveolar foramina
26 Palatine process of maxilla

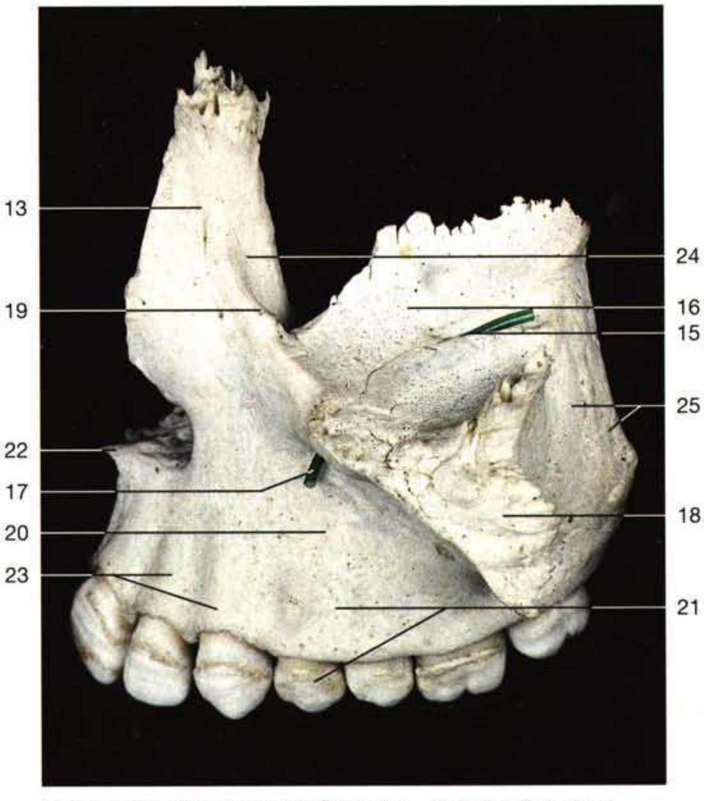

Left maxilla (lateral aspect). Probe = infraorbital canal.

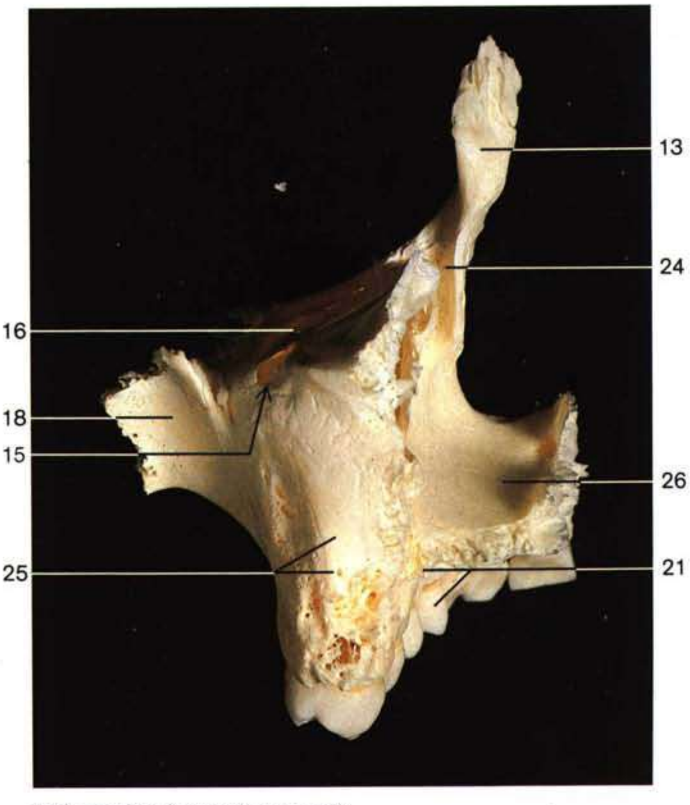

Left maxilla (posterior aspect).

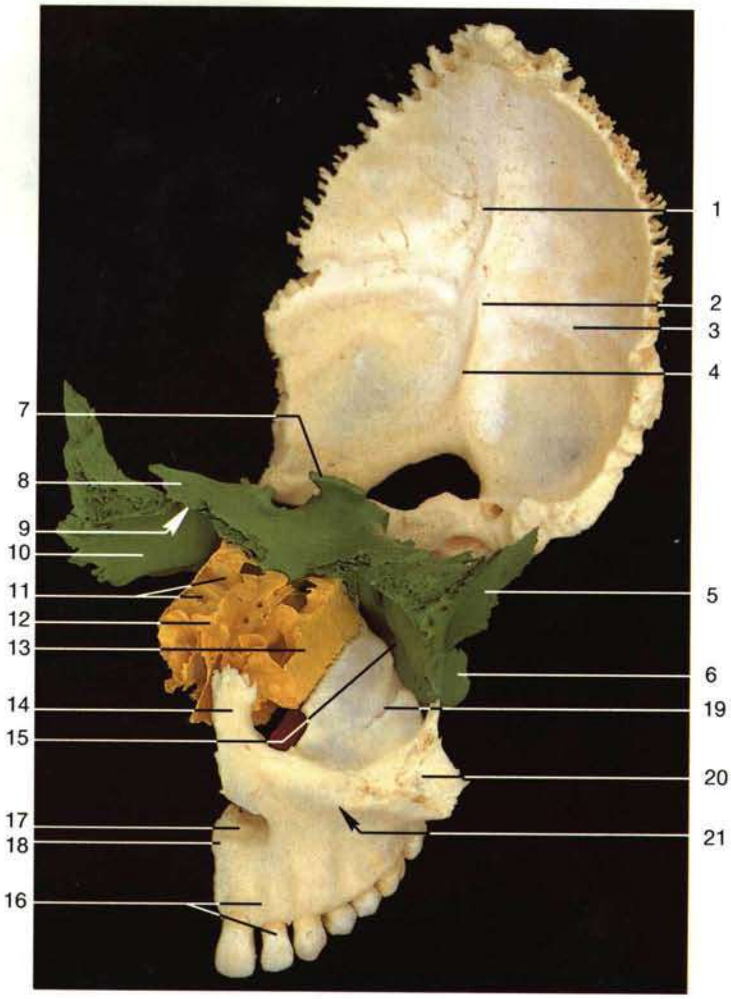

Part of a disarticulated base of skull. The mosaic of the facial bones [sphenoid bone (green), ethmoid bone (yellow), and palatine bone (red)] is seen from the anterior-lateral aspect.

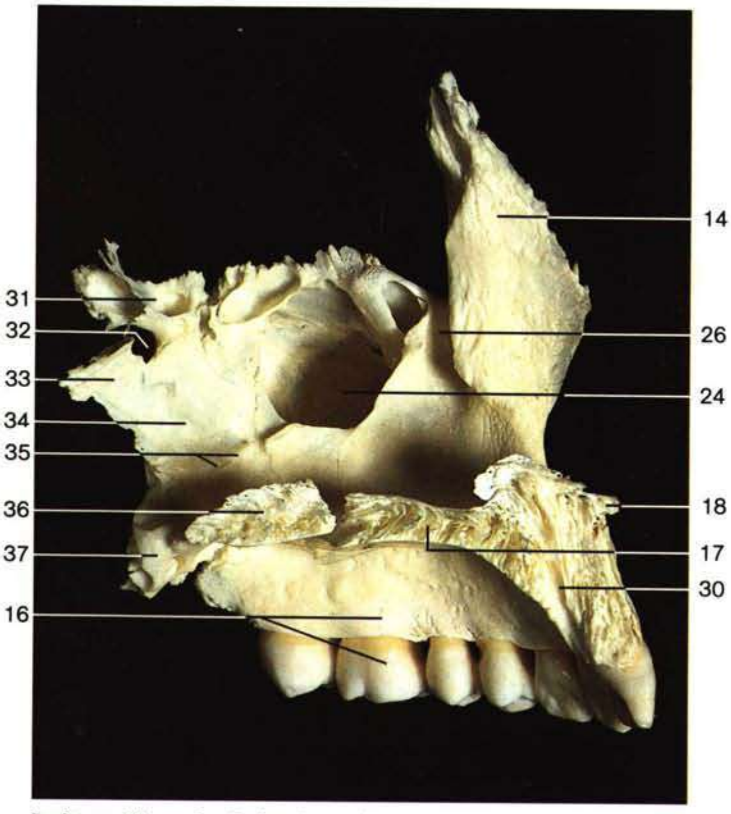

Left maxilla and palatine bone (medial aspect).

Occipital bone

1 Groove for superior sagittal sinus
2 Internal occipital protuberance
3 Groove for transverse sinus
4 Internal occipital crest

Sphenoid bone

5 Greater wing (temporal surface)
6 Lateral pterygoid plate
7 Dorsum sellae
8 Lesser wing
9 Superior orbital fissure
10 Greater wing (orbital surface)

Ethmoid bone

11 Ethmoidal air cells
12 Crista galli
13 Orbital plate

Maxilla

14 Frontal process
15 Inferior orbital fissure
16 Alveolar process with teeth
17 Palatine process
18 Anterior nasal spine
19 Infraorbital groove
20 Zygomatic process
21 Location of infraorbital foramen
22 Middle nasal meatus
23 Inferior nasal meatus
24 Maxillary hiatus (leading to maxillary sinus)
25 Third molar
26 Lacrimal groove
27 Conchal crest
28 Body of maxilla (nasal surface)
29 Nasal crest
30 Incisive canal

Palatine bone

31 Orbital process
32 Sphenopalatine notch
33 Sphenoidal process
34 Perpendicular plate
35 Conchal crest
36 Horizontal plate
37 Pyramidal process

Frontal bone

38 Squamous part
39 Supraorbital foramen
40 Frontal notch
41 Frontal spine

Inferior nasal concha

42 Inferior nasal concha with maxillary process

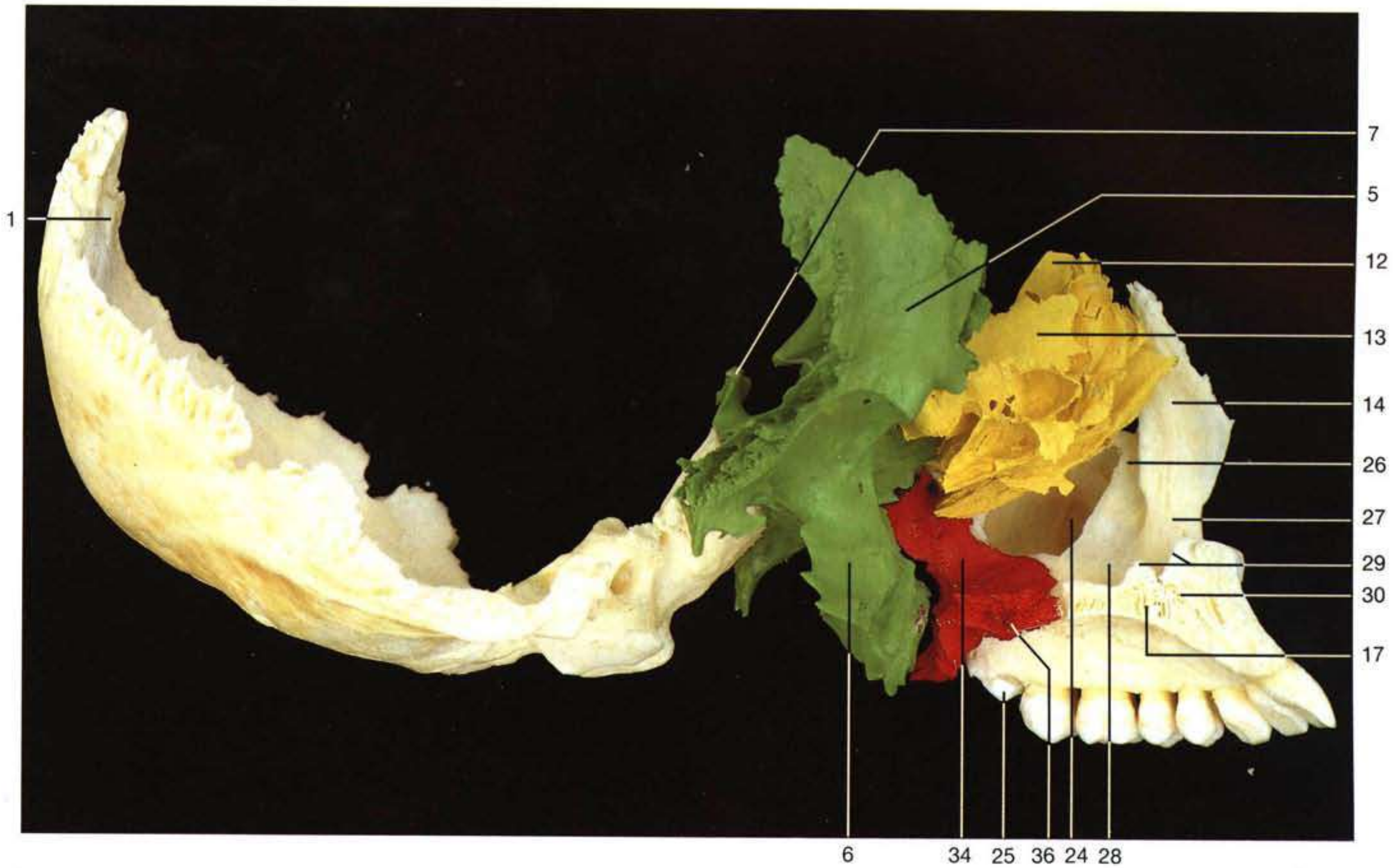
1
7
5
12
13
14
26
27
29
30
17
6
34
25
36
24
28

Part of a disarticulated base of skull (medial aspect). Green = sphenoid bone; yellow = ethmoid bone; red = palatine bone; natural colored = left maxilla.

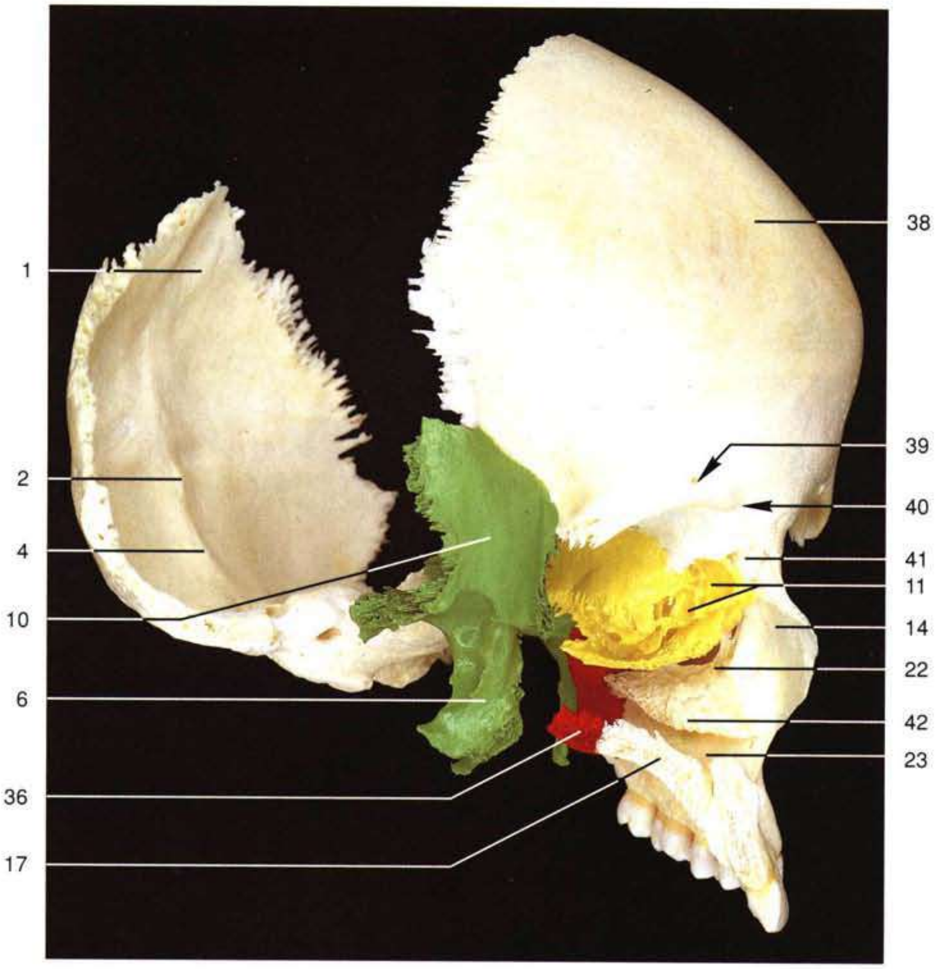
1
2
4
10
6
36
17
38
39
40
41
11
14
22
42
23

Part of a disarticulated base of skull. The same specimen as shown above but with frontal bone (oblique-lateral aspect).

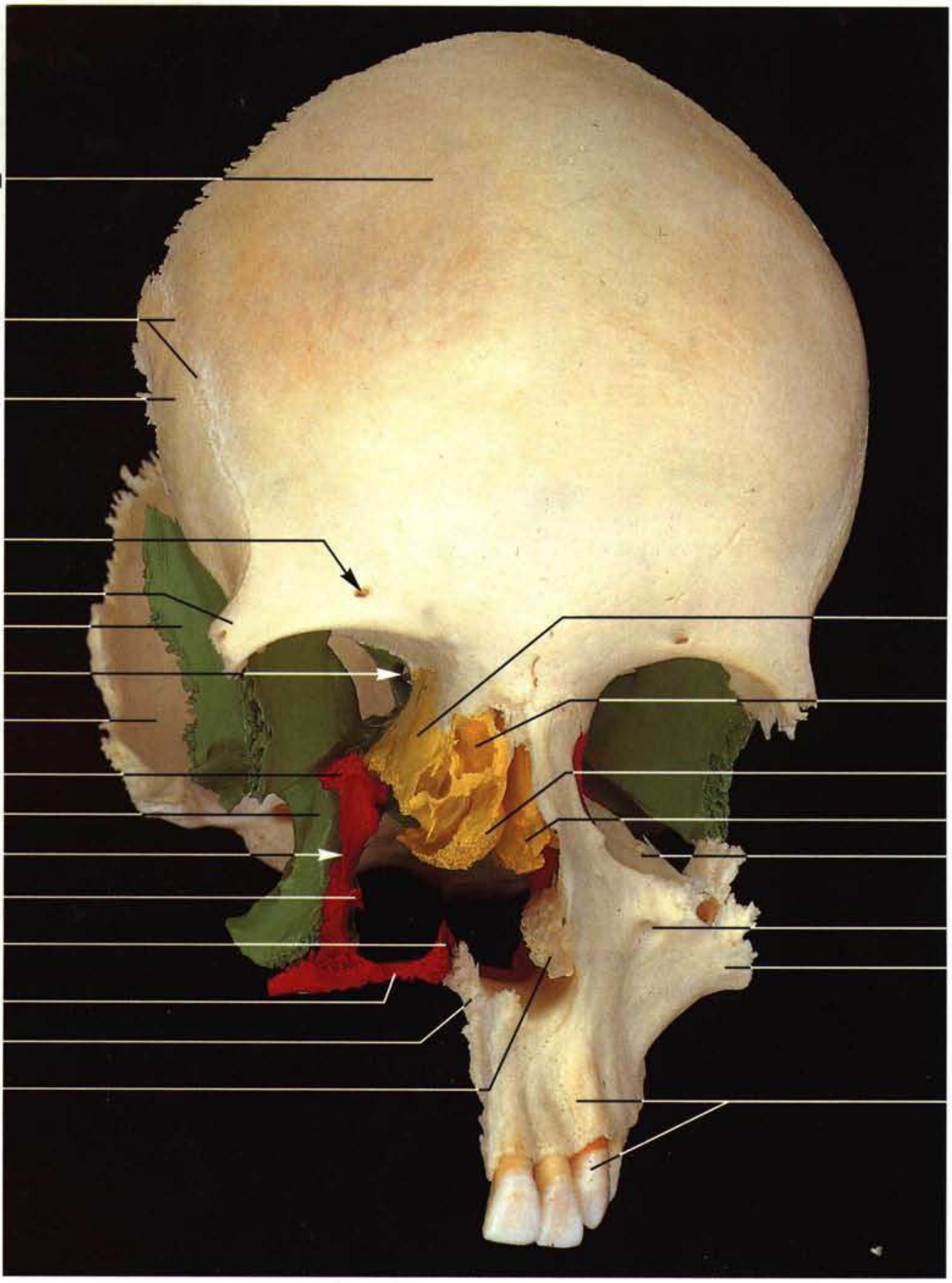

Part of a disarticulated skull, showing the connection of the palatine bone (red) and the maxilla with ethmoid bone (yellow) and sphenoid bone (light green) (anterior aspect).

Frontal bone

1 Squamous part
2 Inferior temporal line
3 Temporal surface
4 Supraorbital foramen
5 Zygomatic process

Occipital bone

6 Squamous part

Sphenoid bone

7 Greater wing (temporal surface)
8 Optic canal within the lesser wing
9 Lateral pterygoid plate

Palatine bone

10 Orbital process
11 Perpendicular plate
12 Conchal crest
13 Nasal crest
14 Horizontal plate

Ethmoid bone

15 Orbital plate
16 Ethmoidal air cell
17 Middle concha
18 Perpendicular plate (part of bony nasal septum)

Maxilla

19 Infraorbital groove
20 Infraorbital foramen
21 Zygomatic process
22 Alveolar process with teeth
23 Palatine process

Left inferior nasal concha

24 Anterior part of inferior concha

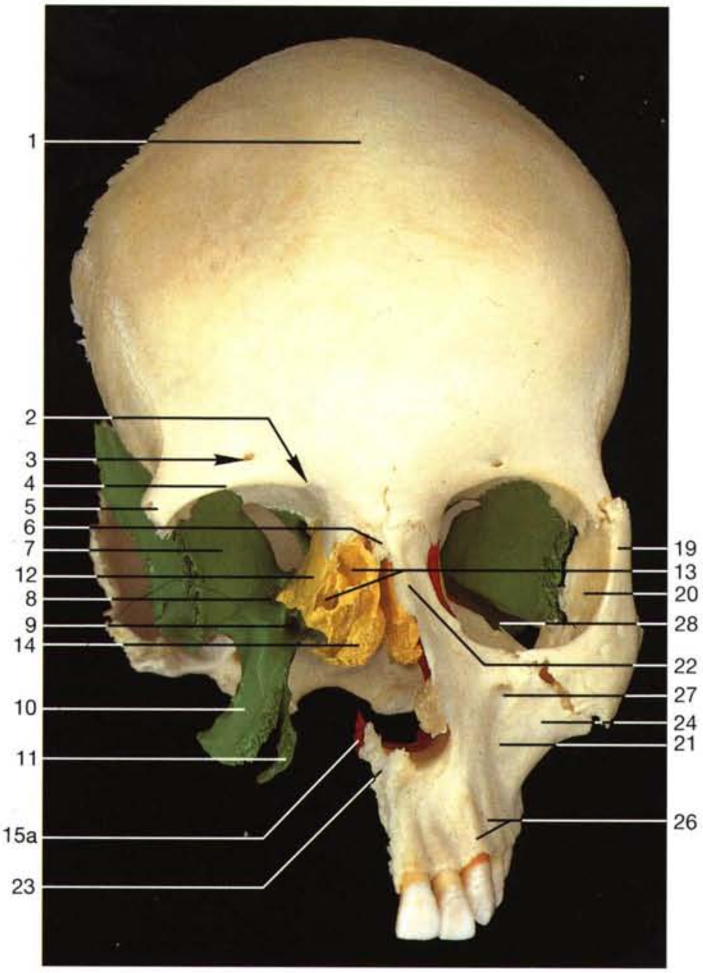

Anterior view of a disarticulated skull, showing the connection of the maxilla with the frontal and zygomatic bones. Yellow = ethmoid bone; red = palatine bone; green = sphenoid bone.

Frontal bone

1 Squamous part
2 Frontal notch
3 Supraorbital foramen
4 Supraorbital margin
5 Zygomatic process
6 Frontal spine

Sphenoid bone

7 Greater wing (orbital surface)
8 Foramen rotundum
9 Pterygoid or Vidian canal
10 Lateral pterygoid plate
11 Medial pterygoid plate

Ethmoid bone

12 Orbital plate
13 Ethmoidal air cells
14 Middle concha

Palatine bone

15 Horizontal plate
15a Nasal crest
16 Pyramidal process
17 Lesser palatine foramen
18 Greater palatine foramen

Zygomatic bone

19 Frontal process
20 Orbital surface

Maxilla

21 Canine fossa
22 Frontal process
23 Palatine process
24 Zygomatic process
25 Alveolar process and teeth
26 Juga alveolaria
27 Infraorbital foramen
28 Infraorbital groove
29 Anterior nasal aperture
30 Anterior nasal spine

Incisive bone

31 Central incisor and incisive bone or premaxilla
32 Incisive fossa

Vomer

33 Ala of the vomer

Sutures and choanae

34 Median palatine suture
35 Transverse palatine suture
36 Choanae

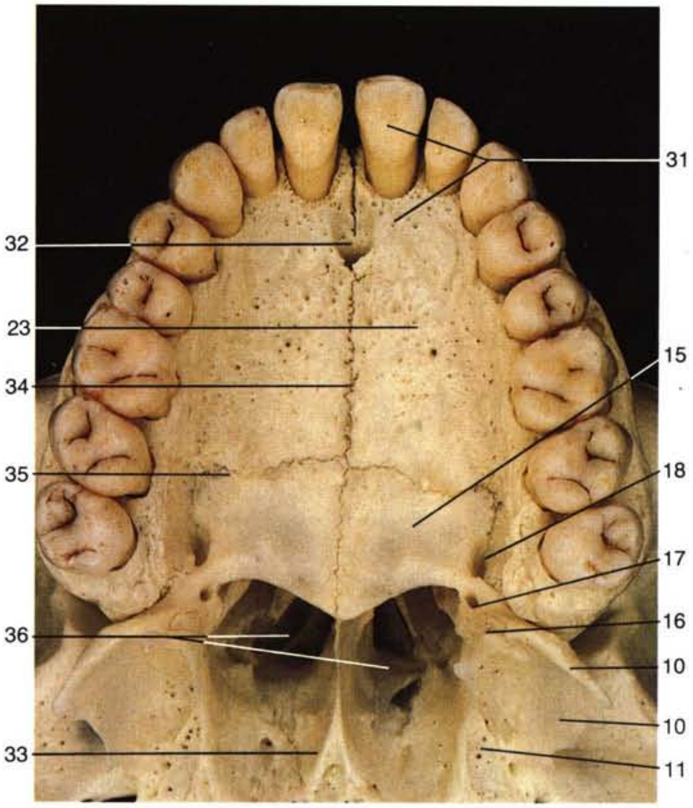

Bony palate and teeth of the maxillae (from below).

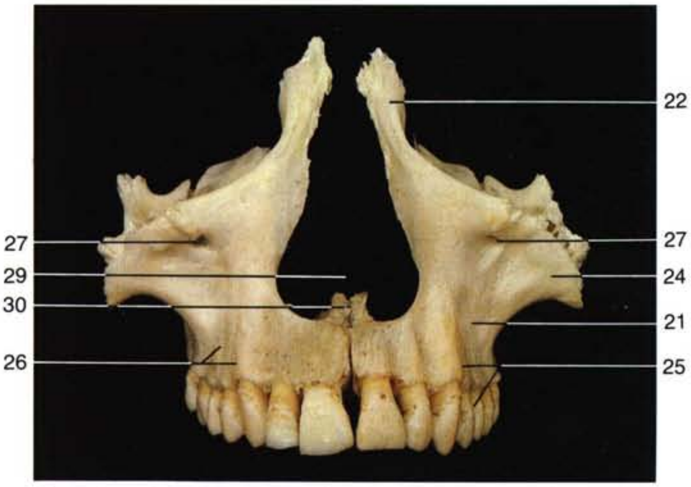

Anterior view of both maxillae, forming the anterior bony aperture of the nose.

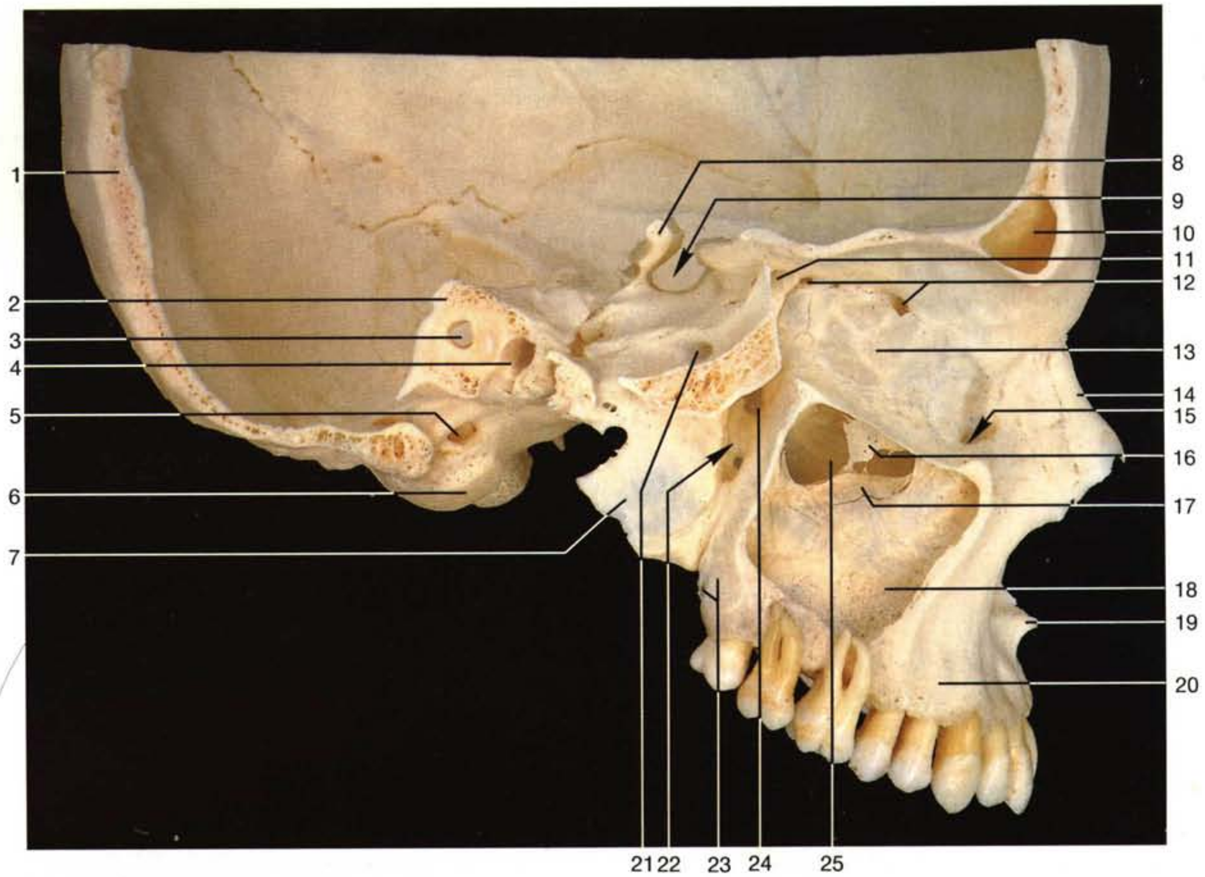

Paramedian section through the skull, right side (lateral aspect). Frontal and maxillary sinus are opened.

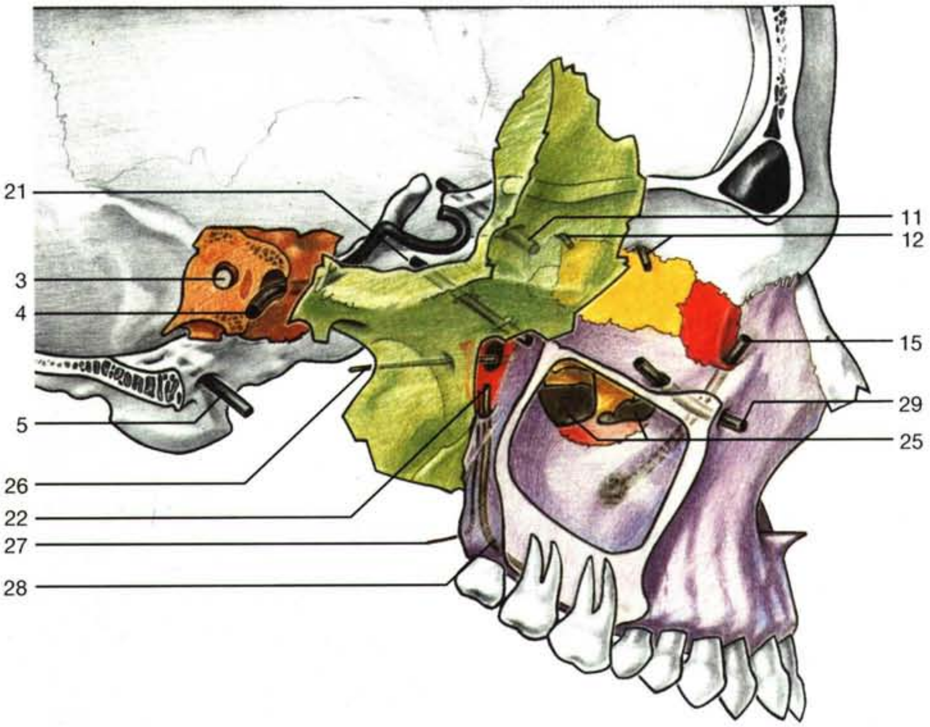

Illustration of canals and foramina connected with the right orbit and pterygopalatine fossa (compare the above figure). The greater wing of sphenoid bone (green) is shown as being transparent. Brown = temporal bone; yellow = ethmoid bone; red = laerimal bone; light red = inferior nasal concha; violet = maxilla; orange = palatine bone.

1 Occipital bone
2 Temporal bone (petrous part)
3 **Internal acoustic meatus**
4 **Carotid canal**
5 **Hypoglossal canal**
6 Occipital condyle
7 Lateral plate of pterygoid process
8 Dorsum of sella turcica
9 Sella turcica
10 Frontal sinus
11 **Optic canal**
12 **Posterior** and **anterior ethmoidal foramen**
13 Orbital plate of ethmoidal bone
14 Nasal bone
15 **Nasolacrimal canal**
16 Uncinate process
17 Inferior nasal concha (maxillary process)
18 Maxillary sinus
19 Anterior nasal spine
20 Alveolar process of maxilla
21 **Foramen rotundum**
22 **Pterygopalatine fossa**
23 Tuberosity of maxilla with **alveolar foramina**
24 **Sphenopalatine foramen**
25 Maxillary hiatus
26 **Pterygoid** or Vidian **canal**
27 **Lesser palatine canal**
28 **Greater palatine canal**
29 **Infraorbital canal**

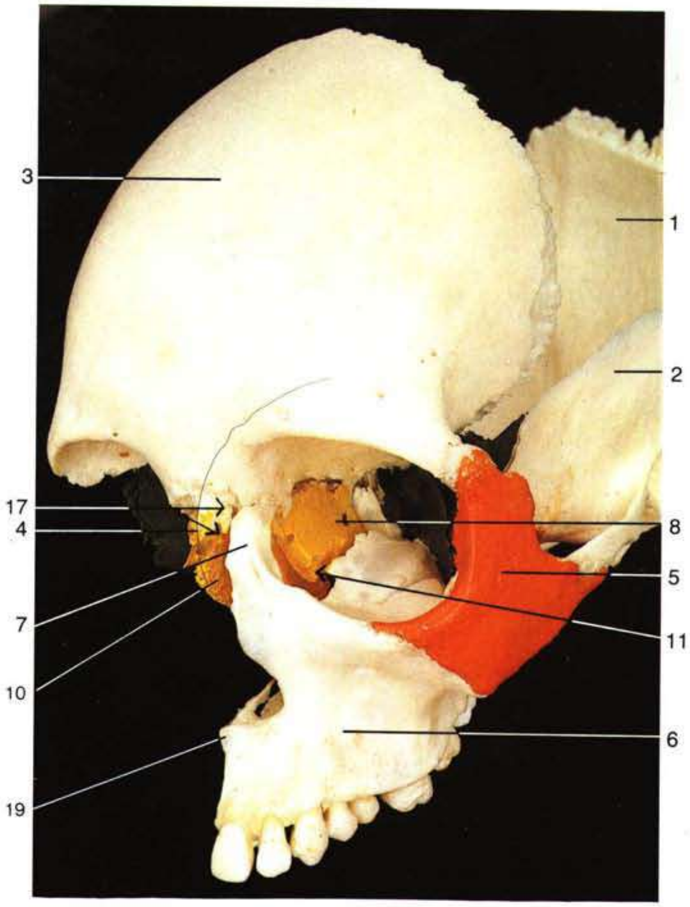

Anterior part of a disarticulated skull.
Orange = zygomatic bone; yellow = ethmoid bone; green = sphenoidal bone. The arrows indicate the locations of the lacrimal bone (11) and the nasal bone (17).

1 **Occipital bone**
2 **Temporal bone**
3 **Frontal bone**
4 Nasal spine of frontal bone
5 **Zygomatic bone**
6 **Maxilla**
7 Frontal process of maxilla
8 **Ethmoid bone**
9 Orbital plate of ethmoid bone
10 Perpendicular plate of ethmoid bone
11 Site of **lacrimal bone**
12 Lacrimal groove of lacrimal bone
13 Posterior lacrimal crest
14 Fossa for lacrimal sac
15 Lacrimal hamulus
16 Nasolacrimal canal
17 Site of **nasal bone**
18 Nasal foramina of nasal bone
19 Anterior nasal spine of maxilla
20 **Vomer**
21 Greater wing of sphenoid bone
22 Anterior and posterior ethmoidal foramina
23 Optic canal
24 Superior orbital fissure
25 Inferior orbital fissure
26 Infraorbital groove
27 Infraorbital foramen

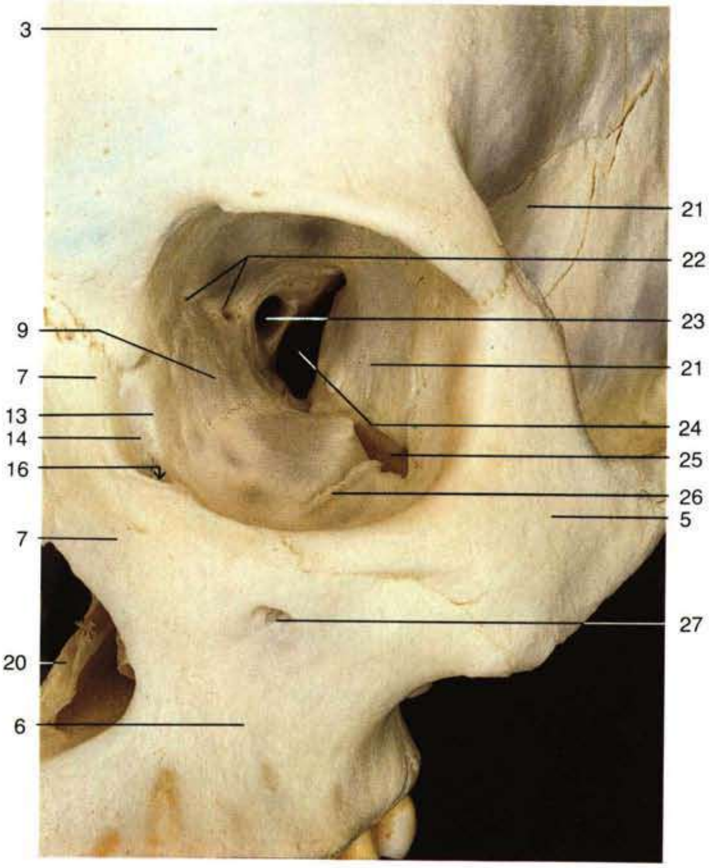

Left orbit (anterior aspect).

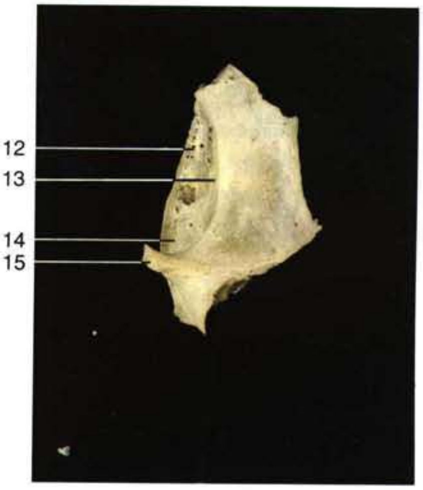

Left lacrimal bone (anterior aspect).

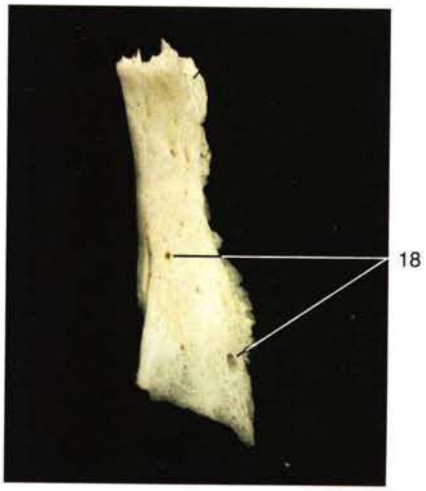

Left nasal bone (anterior aspect).

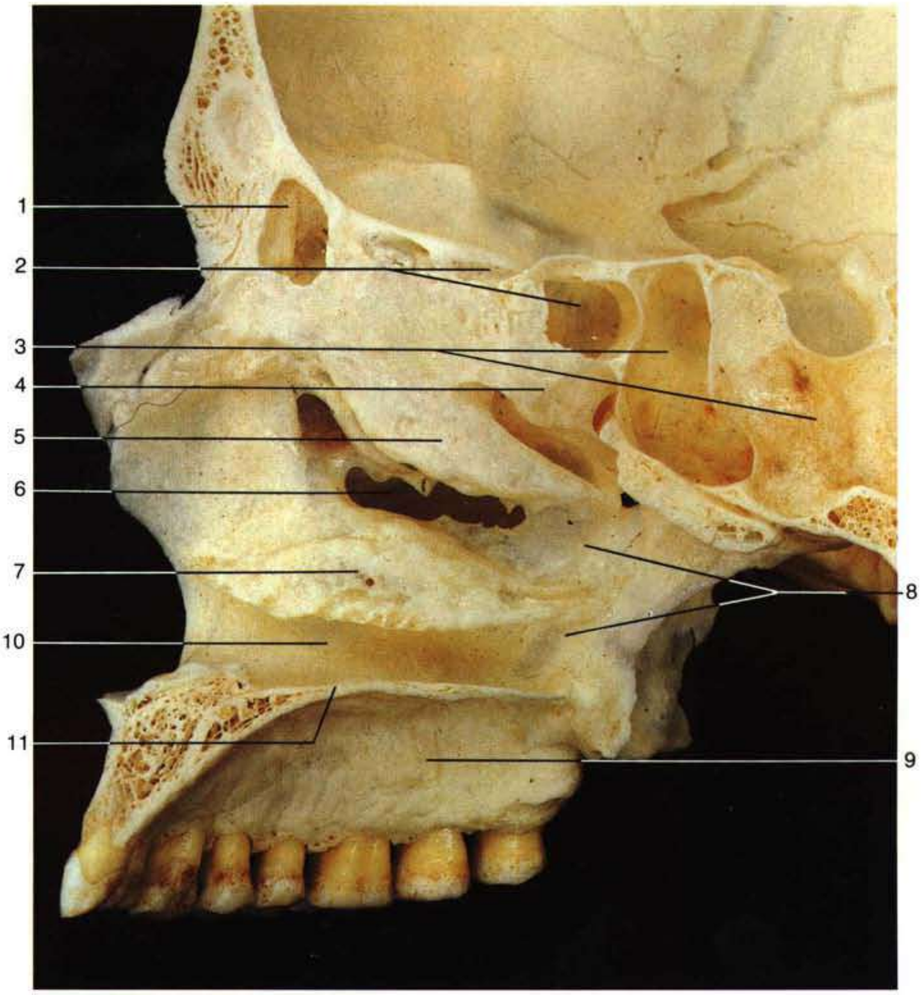

Lateral wall of the nasal cavity. Median section through the skull.

1 Frontal sinus
2 Ethmoidal air cells
3 Sphenoid sinus
4 Superior nasal concha
5 Middle nasal concha
6 Maxillary hiatus
7 Inferior nasal concha
8 Palatine bone
9 Maxilla
10 Inferior meatus
11 Palatine process of the maxilla

▷

To page 49:

Blue = Occipital bone
Light green = Parietal bone
Light brown = Frontal bone
Dark brown = Temporal bone
Red = Sphenoid bone
Dark green = Ethmoid bone
Light blue = Nasal bone
Pink = Inferior concha
Orange = Vomer
Violet = Maxilla
White = Palatine bone
White = Mandible

Inferior Concha and Vomer

1 Ethmoidal process
2 Anterior part of concha
3 Inferior border
4 Ala of vomer
5 Posterior border of nasal septum
6 Lacrimal process
7 Posterior part of concha
8 Maxillary process

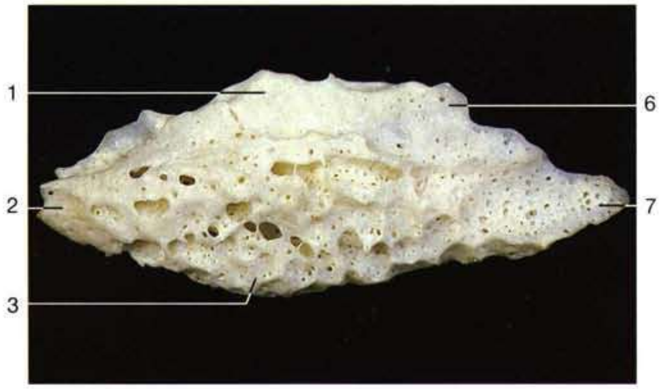

Right inferior nasal concha (medial aspect). Anterior part to the left.

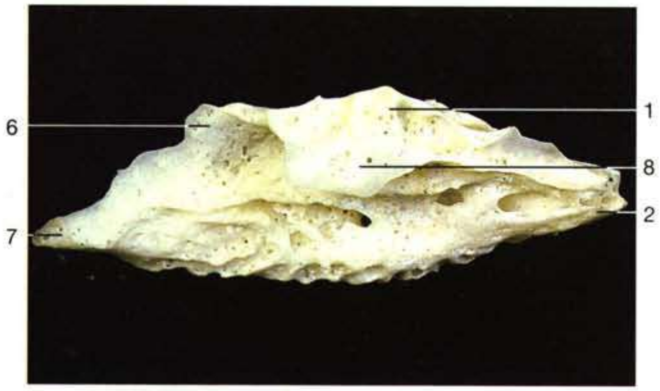

Right inferior nasal concha (lateral aspect). Anterior part to the right.

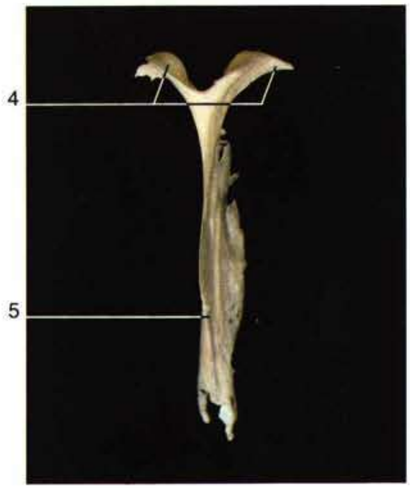

Vomer (posterior aspect).

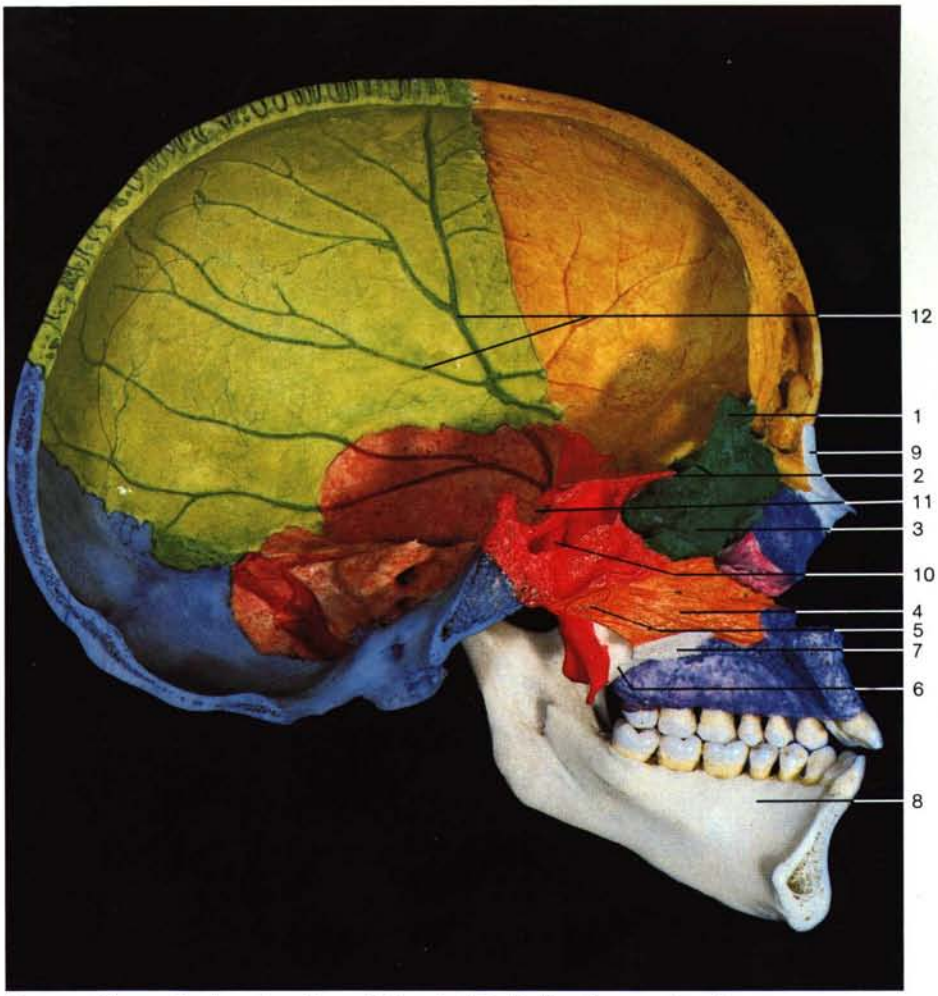

Paramedian sagittal section through the skull including the nasal septum.

1 Crista galli
2 Cribriform plate of ethmoid bone
3 Perpendicular plate of ethmoid bone
4 Vomer
5 Ala of the vomer
6 Palatine bone (perpendicular process)
7 Palatine bone (horizontal plate)
8 Mandible
9 Nasal bone
10 Sphenoidal sinus
11 Hypophysial fossa (sella turcica)
12 Grooves for the middle meningeal artery

Cartilages of the nose

13 Lateral nasal cartilages
14 Greater alar cartilage
15 Lesser alar cartilages
16 Septal cartilage

17 Location of nasal bone

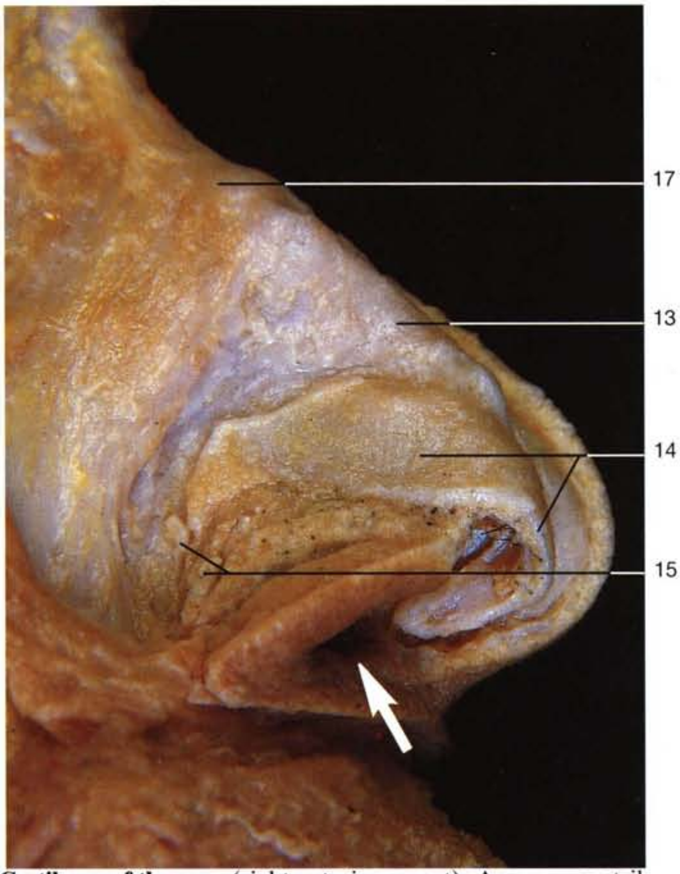

Cartilages of the nose (right anterior aspect). Arrow = nostril, framed by nasal wing.

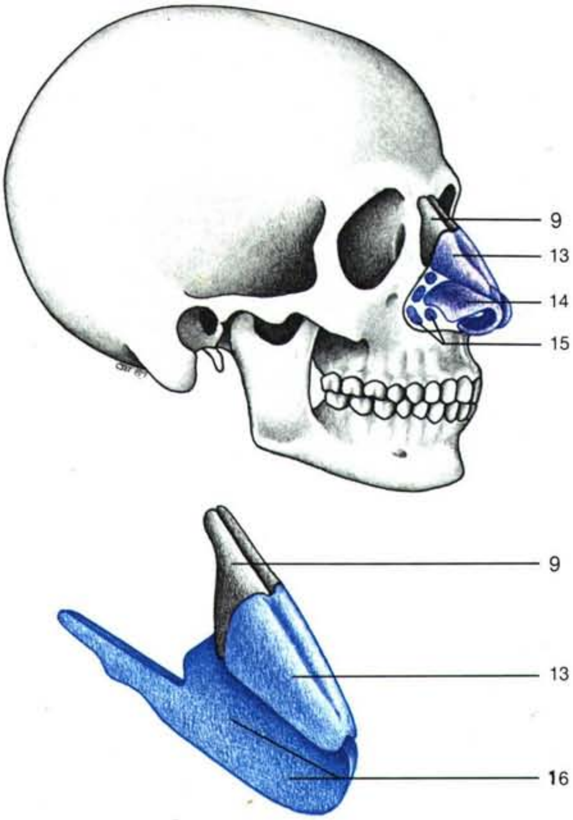

Cartilages of the nose.
Schematic diagram of the external nose.

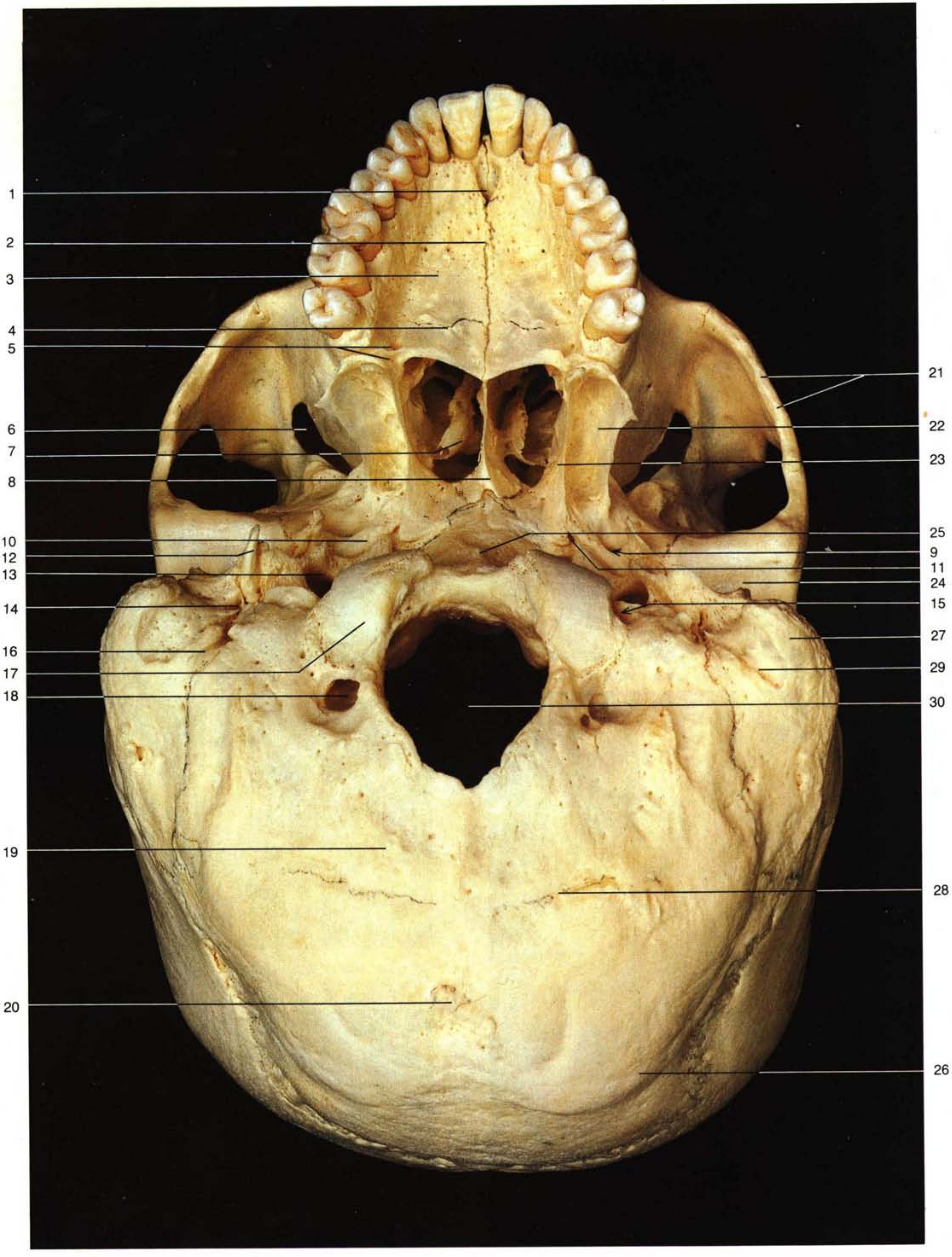
1
2
3
4
5
6
7
8
10
12
13
14
16
17
18
19
20
21
22
23
25
9
11
24
15
27
29
30
28
26

Base of the skull (inferior aspect).

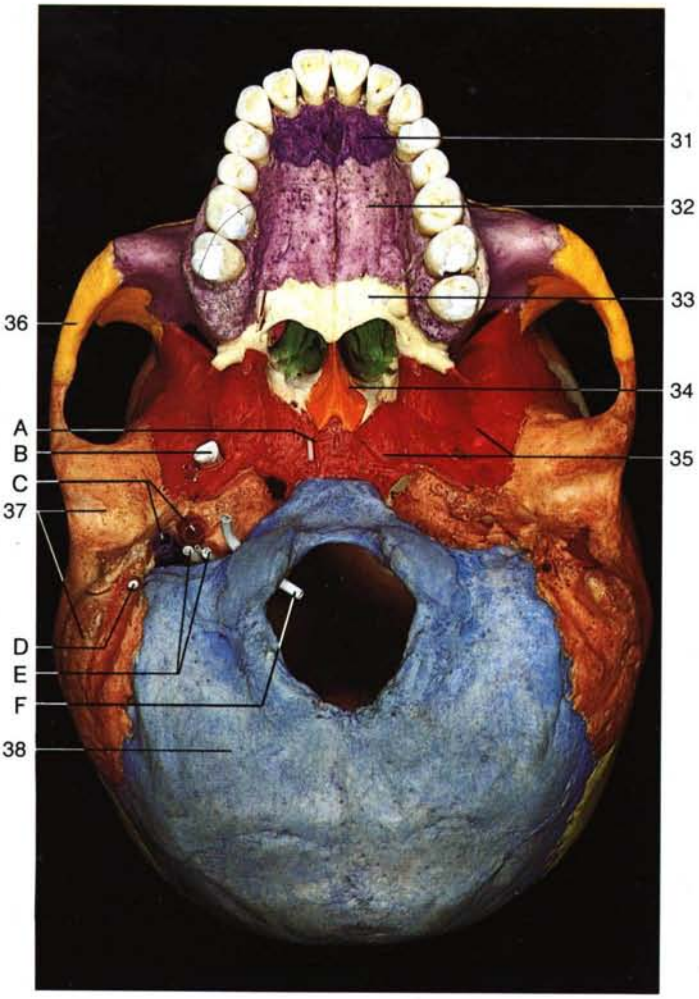

Base of the skull (from below). The individual bones are indicated by different colours.

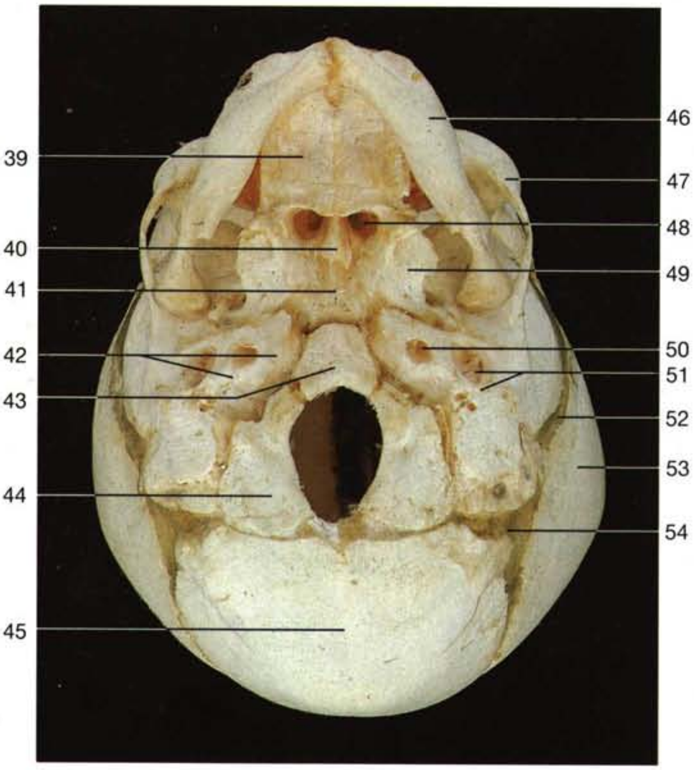

Skull of the newborn (inferior aspect).

A **Pterygoid canal**
B **Foramen ovale**
C Internal carotid artery within **carotid canal** and internal jugular vein within the venous part of jugular foramen
D **Stylomastoid foramen** (facial nerve)
E **Jugular foramen** (glossopharyngeal, vagus and accessory nerves)
F **Hypoglossal canal** (hypoglossal nerve)

1 Incisive canal
2 Median palatine suture
3 Palatine process of maxilla
4 Palatomaxillary suture
5 Greater and lesser palatine foramina
6 Inferior orbital fissure
7 Middle concha (process of ethmoid bone)
8 Vomer
9 Foramen ovale
10 Groove for auditory tube
11 Pterygoid canal
12 Styloid process
13 Carotid canal
14 Stylomastoid foramen
15 Jugular foramen
16 Groove for occipital artery
17 Occipital condyle
18 Condylar canal
19 Nuchal plane
20 External occipital protuberance
21 Zygomatic arch
22 Lateral pterygoid plate
23 Medial pterygoid plate
24 Mandibular fossa
25 Pharyngeal tubercle
26 Superior nuchal line
27 Mastoid process
28 Inferior nuchal line
29 Mastoid notch
30 Foramen magnum
31 **Incisive bone** or **premaxilla** (dark violet)
32 **Maxilla** (violet)
33 **Palatine bone** (white)
34 **Vomer** (orange)
35 **Sphenoid bone** (red)
36 **Zygomatic bone** (yellow)
37 **Temporal bone** (brown)
38 **Occipital bone** (blue)
39 Palatine process of maxilla
40 Vomer
41 Sphenoid bone
42 Petrous part of temporal bone
43 Basilar part
44 Lateral part } of occipital bone
45 Squamous part
46 Mandible
47 Zygomatic arch
48 Choana
49 Pterygoid process of sphenoid bone
50 Carotid canal
51 External acoustic meatus (tympanic annulus)
52 Sphenoidal fontanelle
53 Parietal bone
54 Mastoid fontanelle

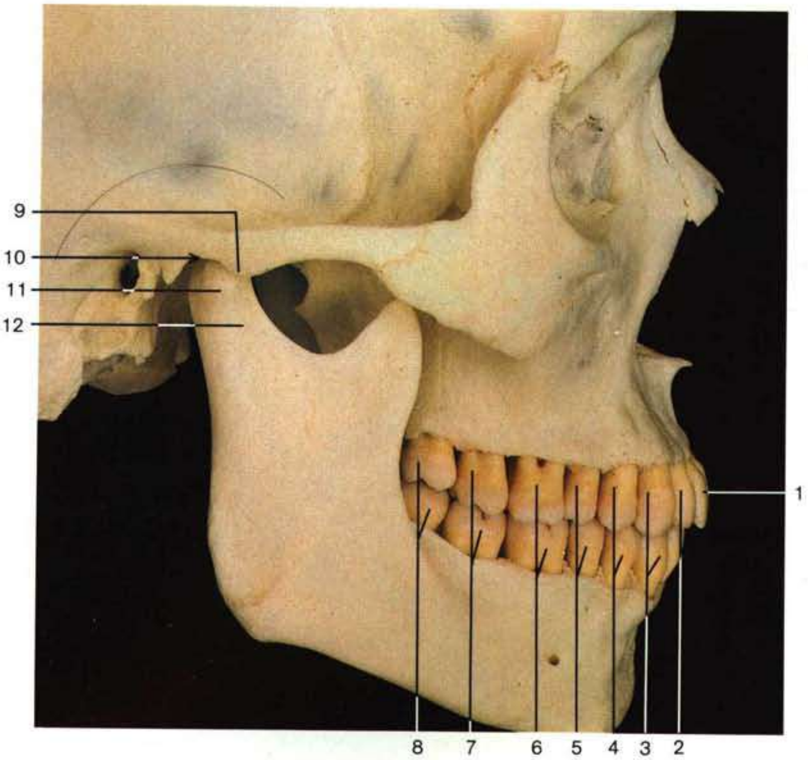

1 Central incisor
2 Lateral incisor
3 Canines
4 First premolars or bicuspids
5 Second premolars or bicuspids
6 First molars
7 Second molars
8 Third molars
9 Articular tubercle
10 Mandibular fossa
11 Head of mandible
12 Condylar process

Normal position of teeth. Dentition in centric occlusion (lateral view).

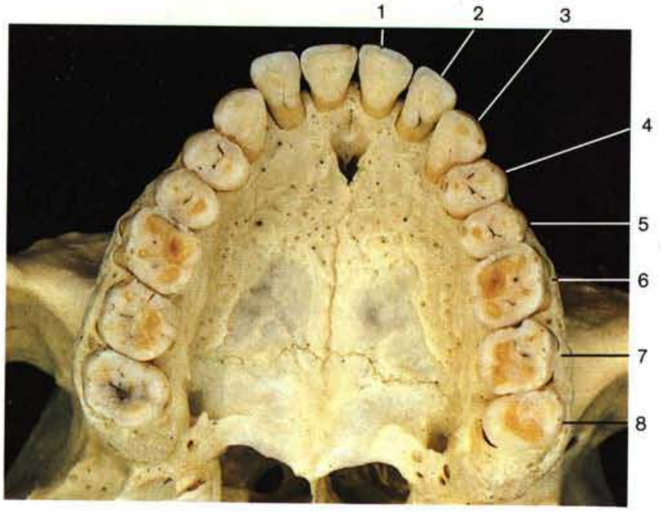

Upper teeth of the adult (inferior aspect).

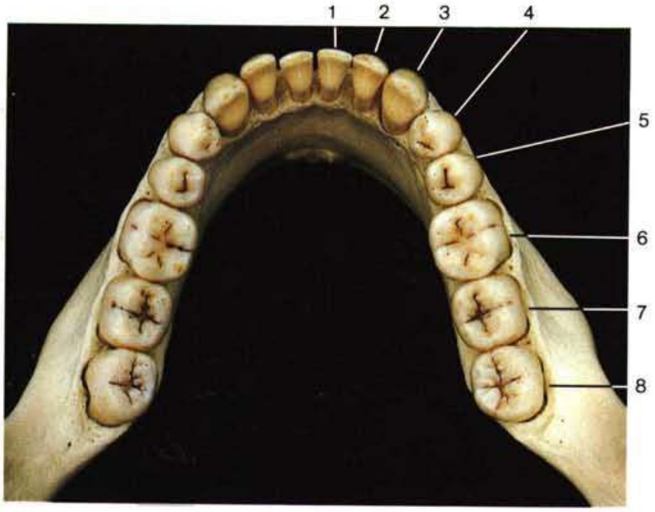

Lower teeth of the adult (superior aspect).

Table of dentition. Eruption of deciduous and permanent teeth (after C. Röse according to A. Kröncke).

Primary dentition (Deciduous teeth)	Maxilla months post partum	Mandible months
1. Central incisor	10.3	8.6
2. Lateral incisor	12.2	14.4
3. Cuspid incisor	19.5	20.1
4. First molar	15.5	16.5
5. Second molar	24.8	24.5

Secondary dentition (Permanent teeth)	years and months ♂	♀	years and months ♂	♀
1. Central incisor	7/8	7/5	6/10	6/7
2. Lateral incisor	8/11	8/6	7/11	7/7
3. Cuspid incisor	12/2	11/7	11/12	10/3
4. First premolar	10/5	10/1	11/3	10/8
5. Second premolar	11/4	11/1	12/0	11/7
6. First molar	6/7	6/6	6/5	6/3
7. Second molar	12/9	12/5	12/3	11/9

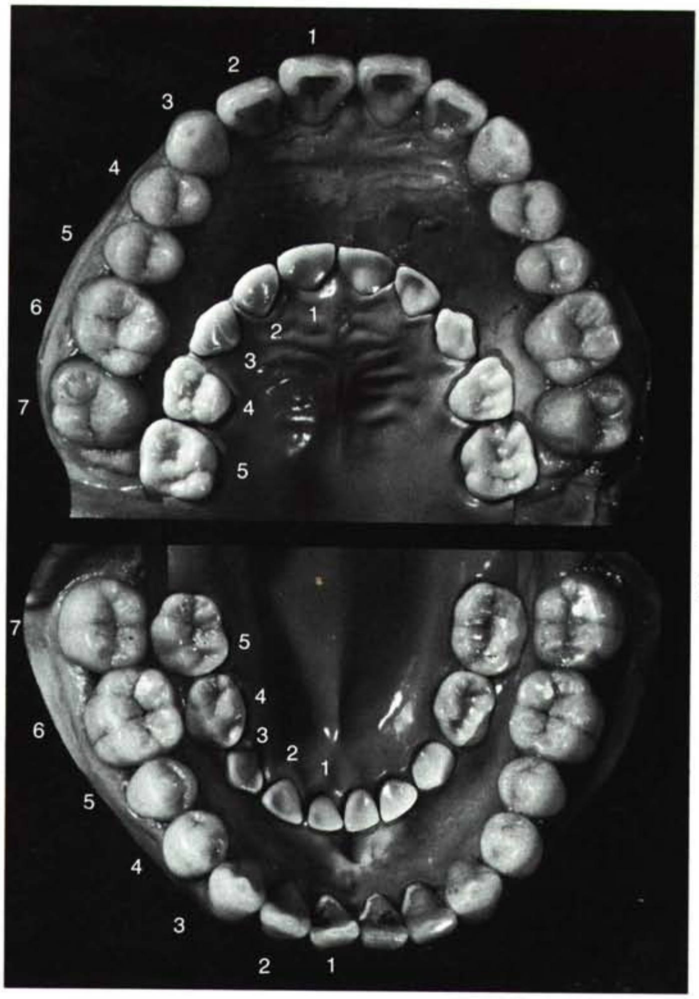

Comparison of the deciduous and permanent teeth. Notice that the breadth of the alveolar arch of the child's mandible and maxilla holding the deciduous teeth is nearly the same as the comparable portion in the jaws of the adult. Note the unerupted third molars.

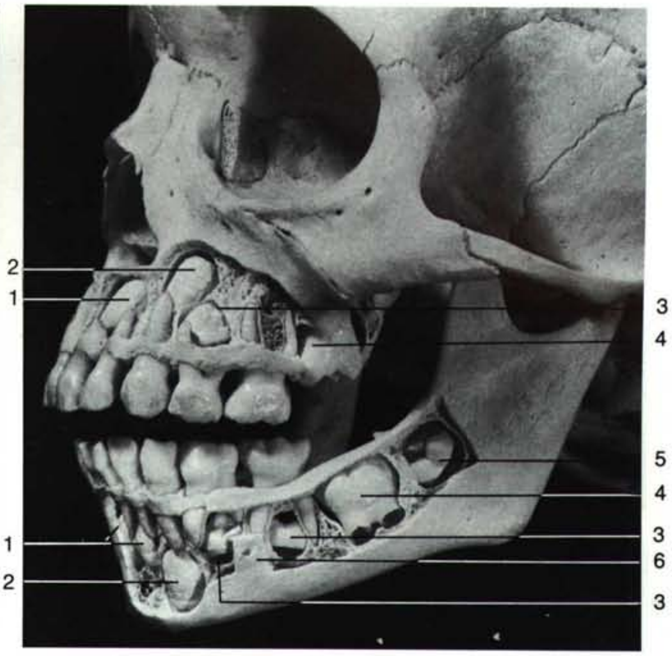

Deciduous teeth in child's skull. The developing crowns of the permanent teeth are displayed in their sockets in the maxilla and mandible.

△

1 Permanent incisors
2 Permanent cuspid (canine)
3 Premolars
4 First permanent molar
5 Second permanent molar
6 Mental foramen

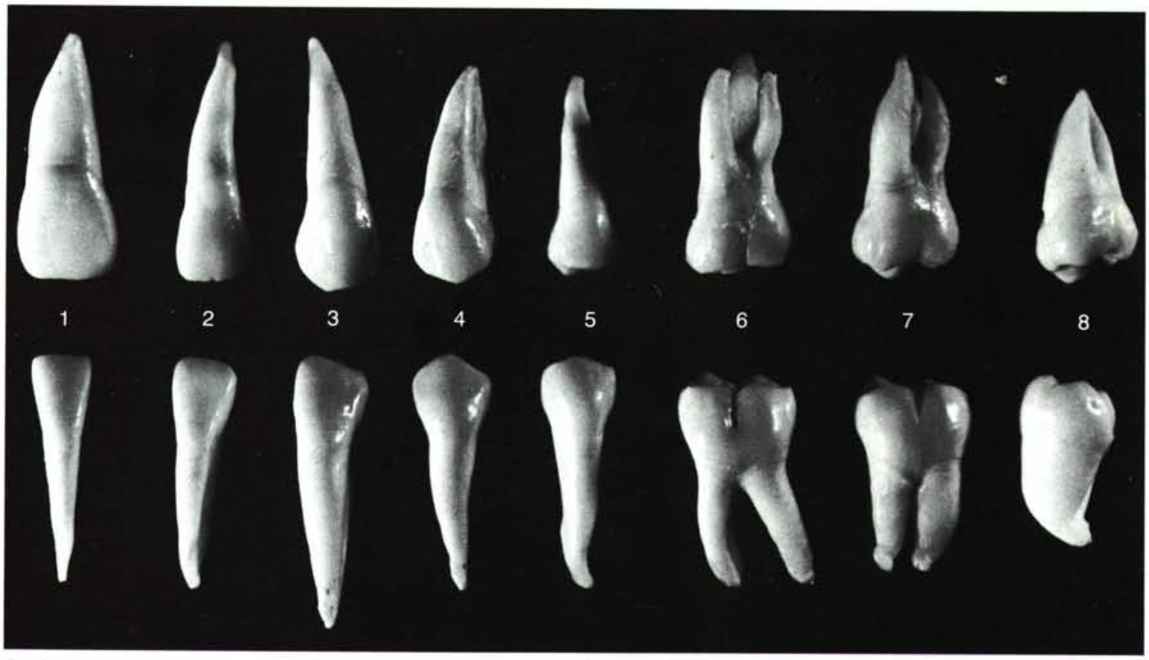

Isolated teeth of the alveolar part of the maxilla (top row) and the mandible (lower row), labial surface of the teeth.

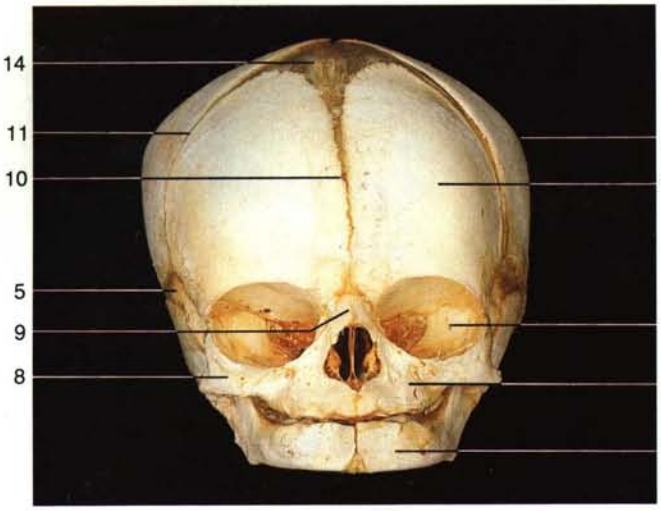

Skull of the newborn (anterior aspect).

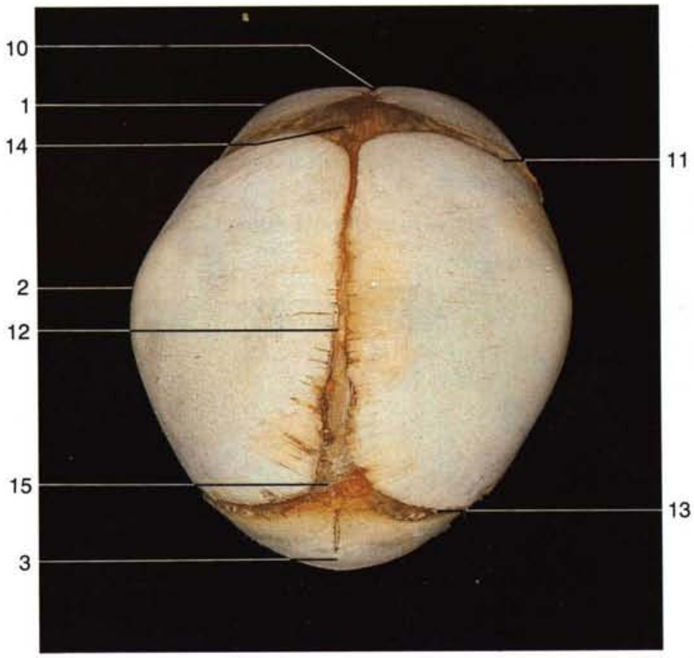

Skull of the newborn (superior aspect). Calvaria.

Cranial skeleton

1 Frontal tuber or eminence
2 Parietal tuber or eminence
3 Occipital tuber or eminence
4 Squamous part of temporal bone
5 Greater wing of sphenoid bone

Facial skeleton

6 Maxilla
7 Mandible
8 Zygomatic bone
9 Nasal bone

Sutures and fontanelles

10 Frontal suture
11 Coronal suture
12 Sagittal suture
13 Lambdoid suture
14 Anterior fontanelle
15 Posterior fontanelle
16 Sphenoidal fontanelle
17 Mastoid fontanelle

Base of the skull

18 Frontal bone
19 Ethmoid bone
20 Sphenoid bone
21 Hypophysial fossa (sella turcica)
22 Dorsum sellae
23 Temporal bone
24 Mastoid fontanelle
25 Occipital bone

In the newborn the facial skeleton, in contrast to the cranial skeleton, appears relatively small. There are no teeth presenting. The bones of the cranium are separated by wide fontanelles.

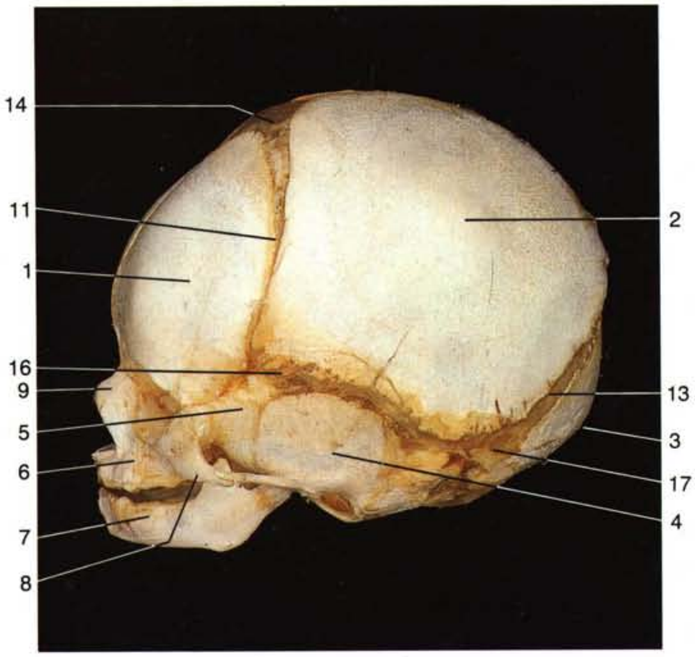

Skull of the newborn (lateral aspect).

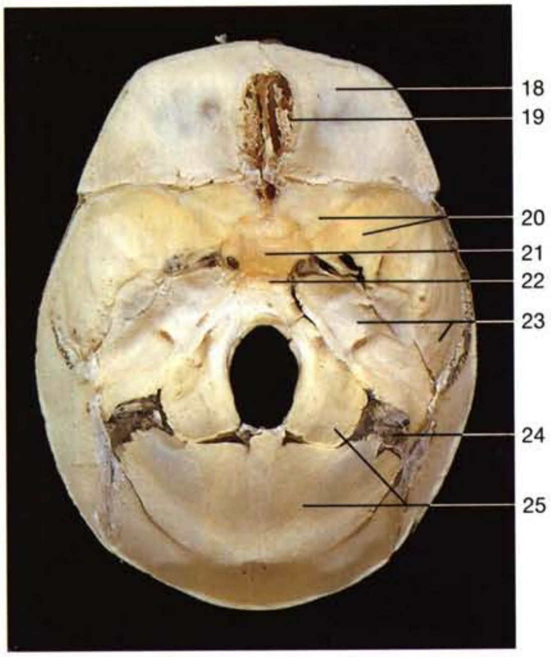

Base of the skull of the newborn (internal aspect).

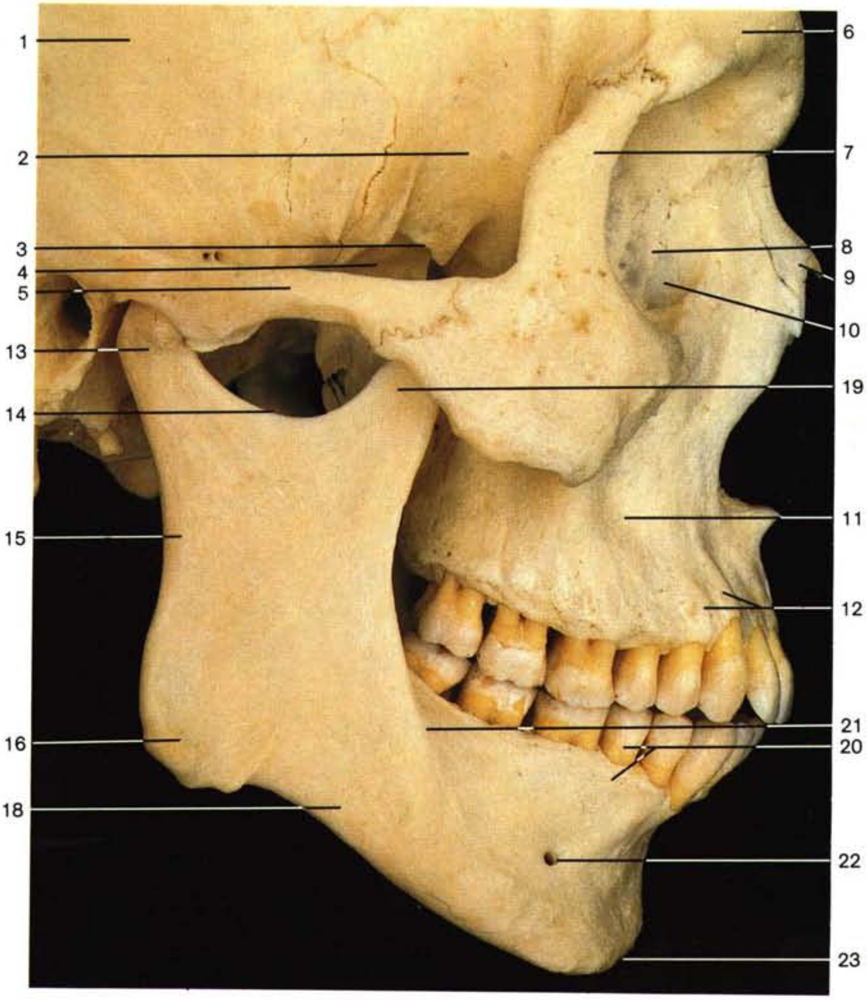

Lateral aspect of the facial bones. Mandible and teeth in the position of occlusion. Upper and lower jaw occluded.

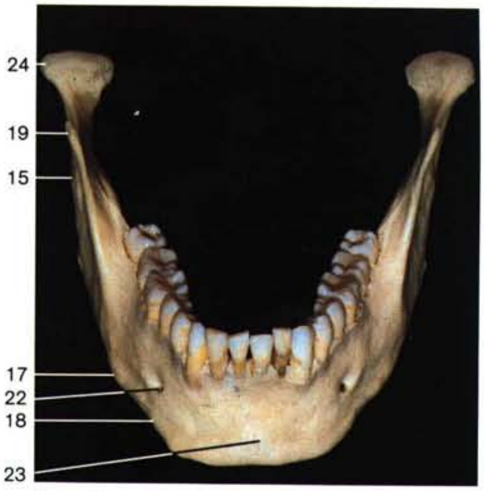

Mandible of the adult (anterior aspect).

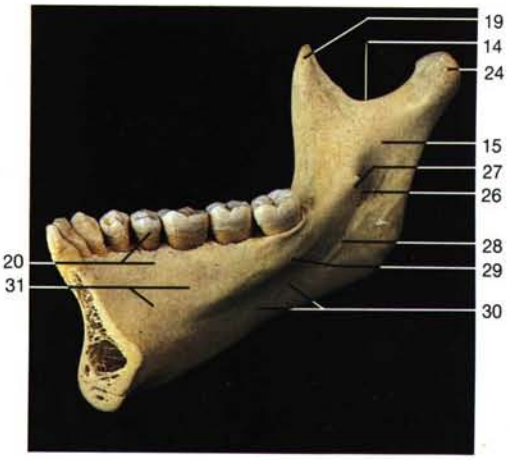

Right half of mandible (medial aspect).

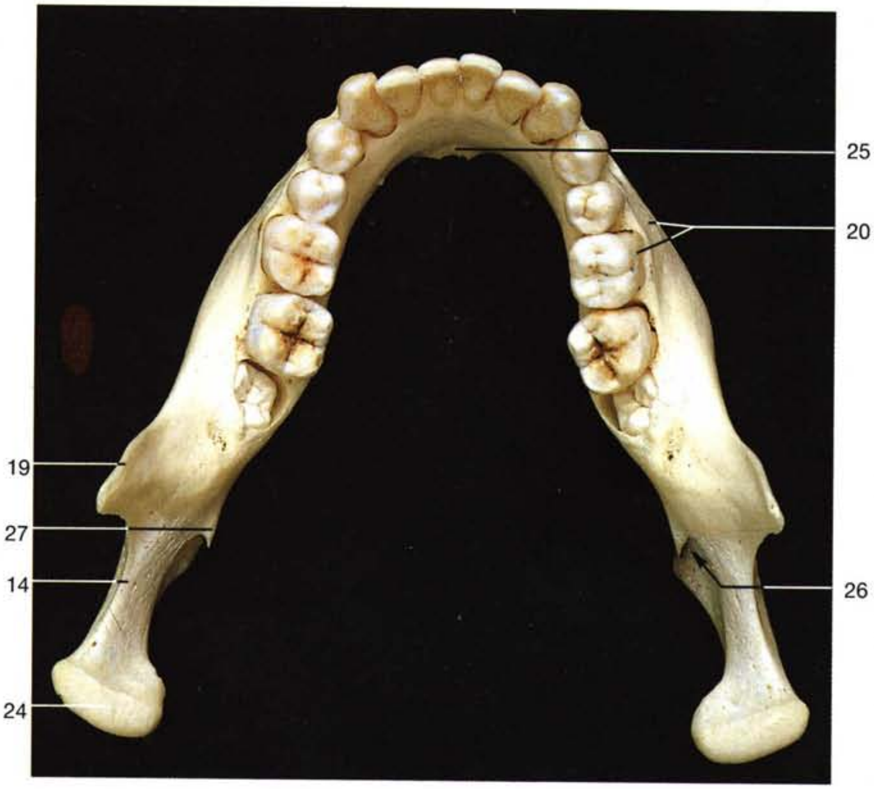

Mandible of the adult (superior aspect).

1 **Temporal bone**
2 Temporal fossa (greater wing of sphenoid bone)
3 Infratemporal crest
4 Infratemporal fossa
5 Zygomatic arch
6 **Frontal bone**
7 **Zygomatic bone** (frontal process)
8 **Lacrimal bone**
9 **Nasal bone**
10 Lacrimal groove
11 **Maxilla** (canine fossa)
12 Alveolar process of maxilla

Mandible
13 Condylar process
14 Mandibular notch
15 Ramus of the mandible
16 Masseteric tuberosity
17 Angle of the mandible
18 Body of the mandible
19 Coronoid process
20 Alveolar process including teeth
21 Oblique line
22 Mental foramen
23 Mental protuberance
24 Head of the mandible
25 Genial tubercle or mental spine
26 Mandibular foramen (entrance to mandibular canal)
27 Lingula
28 Mylohyoid sulcus
29 Mylohyoid line
30 Submandibular fossa
31 Sublingual fossa

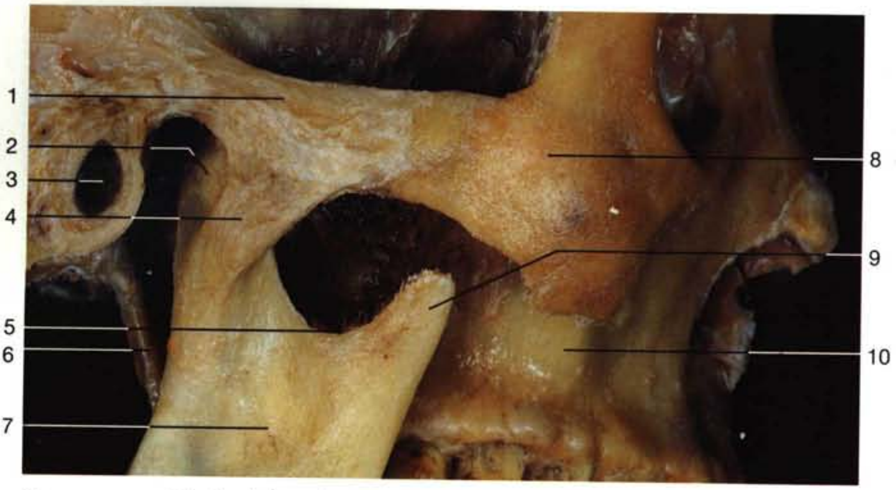

Temporomandibular joint with ligaments.

1. Zygomatic arch
2. Articular capsule
3. External acoustic meatus
4. **Lateral ligament**
5. Mandibular notch
6. **Stylomandibular ligament**
7. Ramus of the mandible
8. Zygomatic bone
9. Coronoid process
10. Maxilla
11. Articular cartilage of condylar process
12. Styloid process
13. Mandibular fossa
14. **Articular disc**
15. Articular tubercle
16. Lateral pterygoid muscle
17. Condylar process of mandible

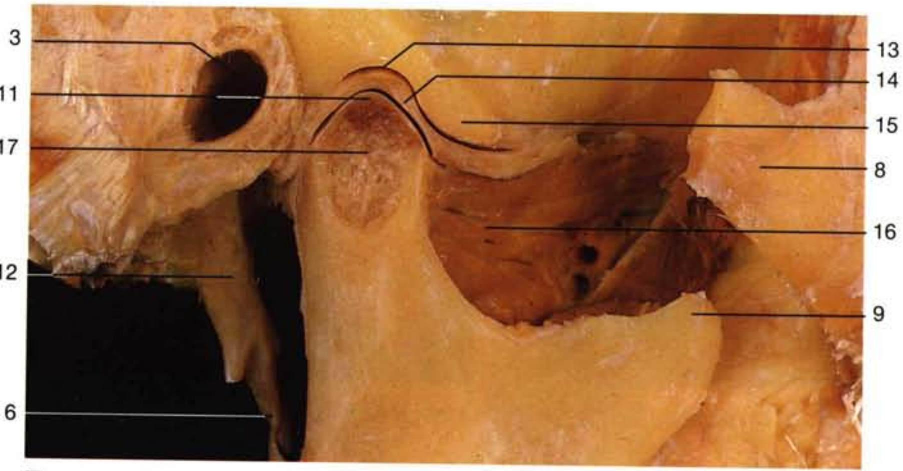

Temporomandibular joint, sagittal section.

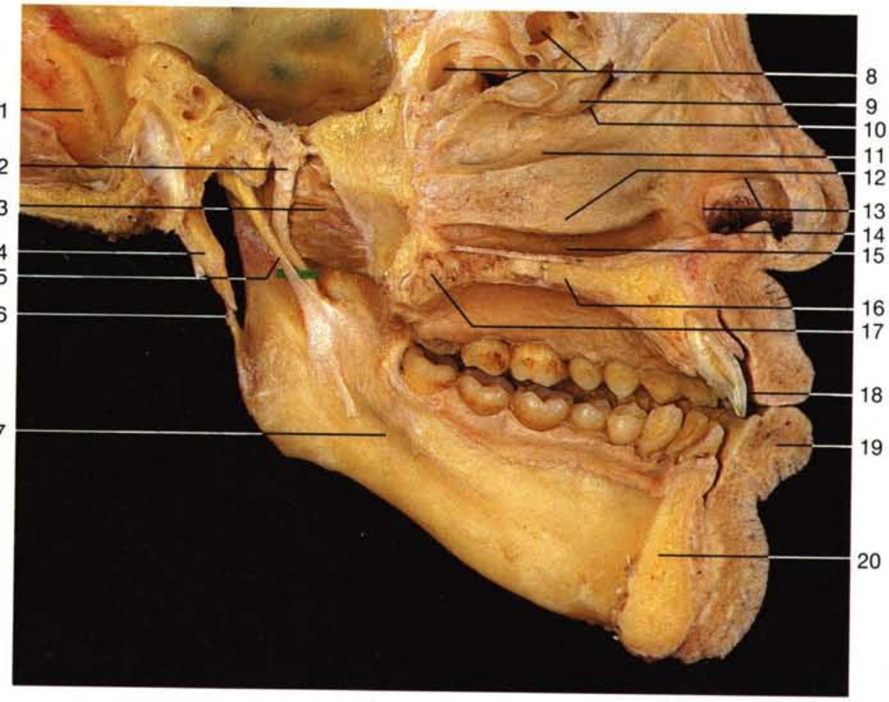

Ligaments of temporomandibular joint. Left half of the head (medial aspect).

1. Groove for sigmoid sinus
2. Mandibular nerve
3. Lateral pterygoid muscle
4. Styloid process
5. **Sphenomandibular ligament**
6. **Stylomandibular ligament**
7. Mylohyoid groove
8. **Ethmoidal air cells**
9. Ethmoidal bulla
10. **Hiatus semilunaris**
11. Middle meatus
12. **Inferior nasal concha**
13. Limen nasi
14. Vestibule with hairs
15. Inferior meatus
16. Hard palate
17. Soft palate
18. Vestibule of oral cavity
19. Lower lip
20. Mandible

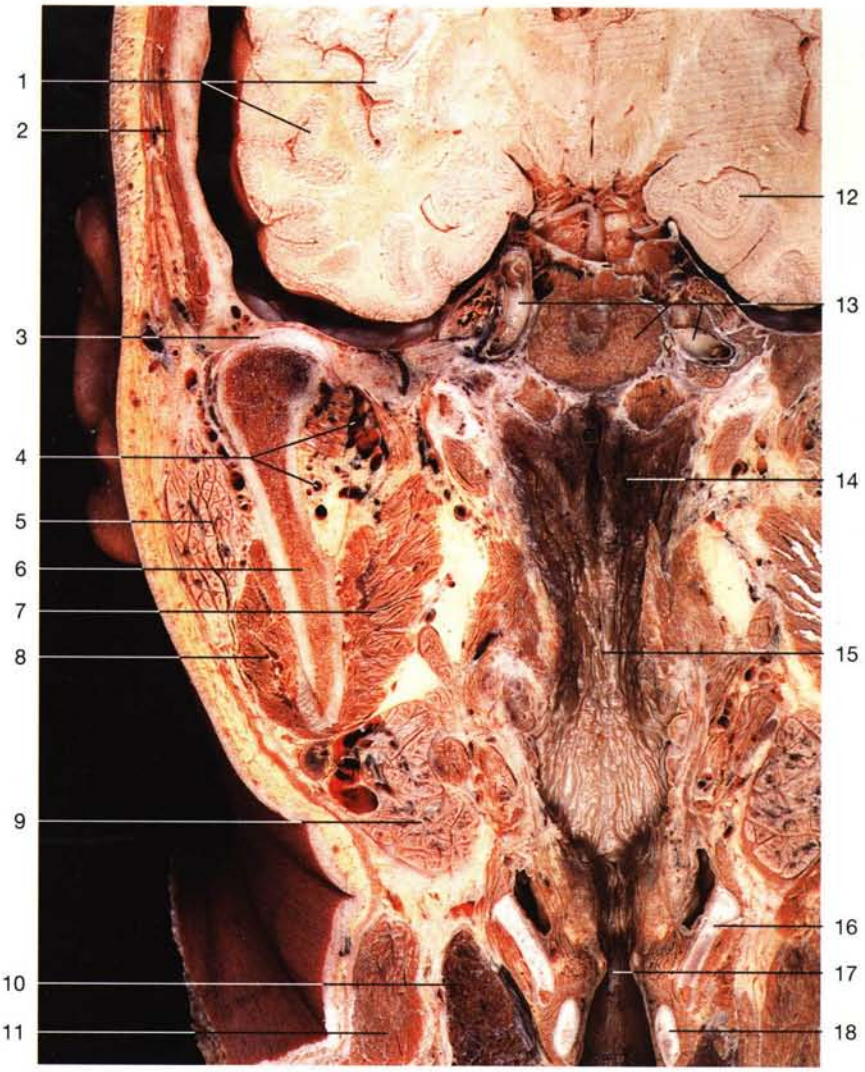

Coronal section through the head at the level of the temporomandibular joint (right side, anterior aspect).

1 Insular lobe and temporal lobe
2 Temporalis muscle
3 **Articular disc of temporomandibular joint**
4 Maxillary artery and pterygoid venous plexus
5 Parotid gland
6 **Mandible**
7 **Medial pterygoid muscle**
8 **Masseter muscle**
9 Submandibular gland
10 Thyroid gland
11 Sternocleidomastoid muscle
12 Hippocampus
13 Internal carotid artery and sphenoid bone
14 Pharyngeal tonsil
15 Pharynx
16 Thyroid cartilage
17 Rima glottidis
18 Cricoid cartilage
19 **Hard palate** and palatine glands
20 **Oral cavity**
21 Upper molar
22 **Oral vestibule**
23 Lower molar
24 Platysma muscle
25 Maxillary sinus
26 Superior longitudinal muscle of tongue
27 **Transverse muscle of tongue**
28 Buccinator muscle
29 Inferior longitudinal muscle of tongue
30 Sublingual gland
31 **Genioglossus muscle**
32 Mastoid process
33 Styloid process
34 **Stylomandibular ligament**
35 Articular capsule
36 **Lateral ligament**
37 Zygomatic arch
38 **Sphenomandibular ligament**
39 Mandibular foramen

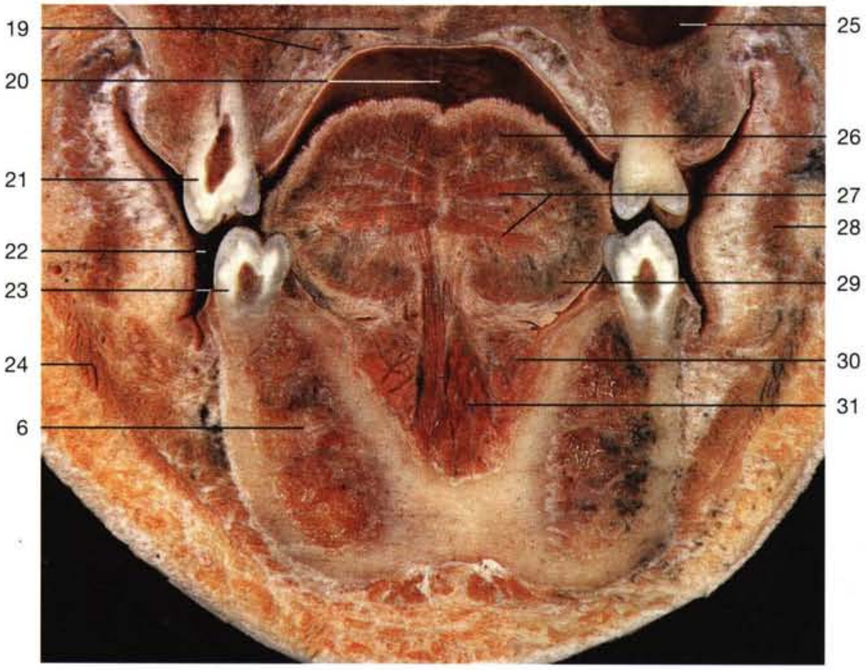

Coronal section through the oral cavity

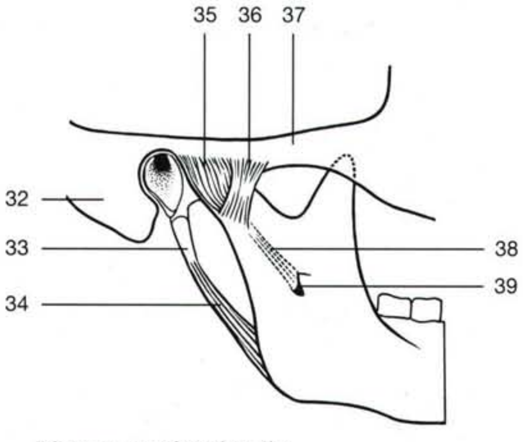

Ligaments related to the **temporomandibular joint**

Temporalis and masseter muscles. The temporal fascia has been removed, the temporomandibular joint severed and the zygomatic arch displayed.

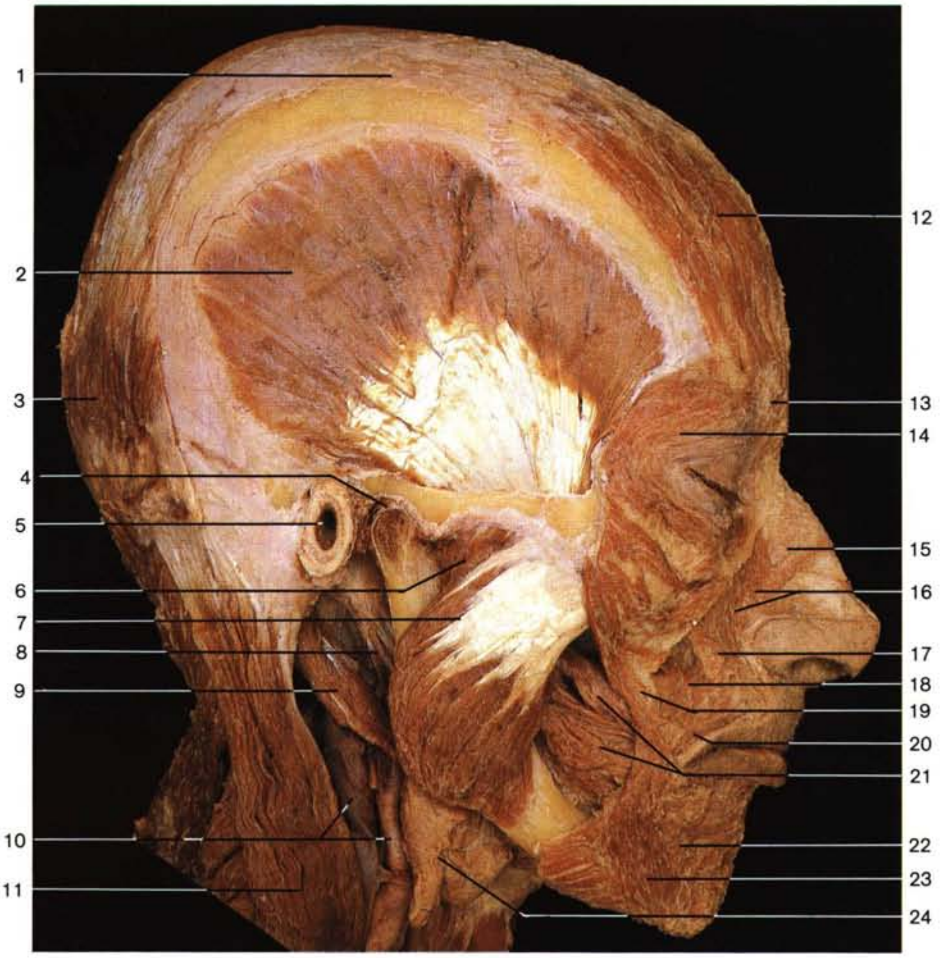

1 Galea aponeurotica
2 **Temporalis muscle**
3 Occipital belly of occipitofrontalis muscle
4 **Temporomandibular joint**
5 External acoustic meatus
6 Deep layer of **masseter muscle**
7 Superficial layer of masseter muscle
8 Stylohyoid muscle
9 Posterior belly of **digastric muscle**
10 Internal jugular vein and external carotid artery
11 **Sternocleidomastoid muscle**
12 Frontal belly of occipitofrontalis muscle
13 Depressor supercilii muscle
14 **Orbicularis oculi muscle**
15 Transverse part of nasalis muscle
16 Levator labii superioris alaeque nasi muscle
17 Levator labii superioris muscle
18 Levator anguli oris muscle
19 **Zygomaticus major muscle**
20 **Orbicularis oris muscle**
21 **Buccinator muscle**
22 Depressor labii inferioris muscle
23 Depressor anguli oris muscle
24 **Submandibular gland**
25 Lateral nasal cartilage
26 Greater alar cartilage (lateral part)
27 Lesser alar cartilages
28 Greater alar cartilage (medial part)
29 Infraorbital nerve
30 Anterior belly of digastric muscle
31 Hypoglossal nerve and hyoglossus muscle
32 Superior thyroid artery
33 **Zygomatic arch**
34 Internal carotid artery
35 Common carotid artery

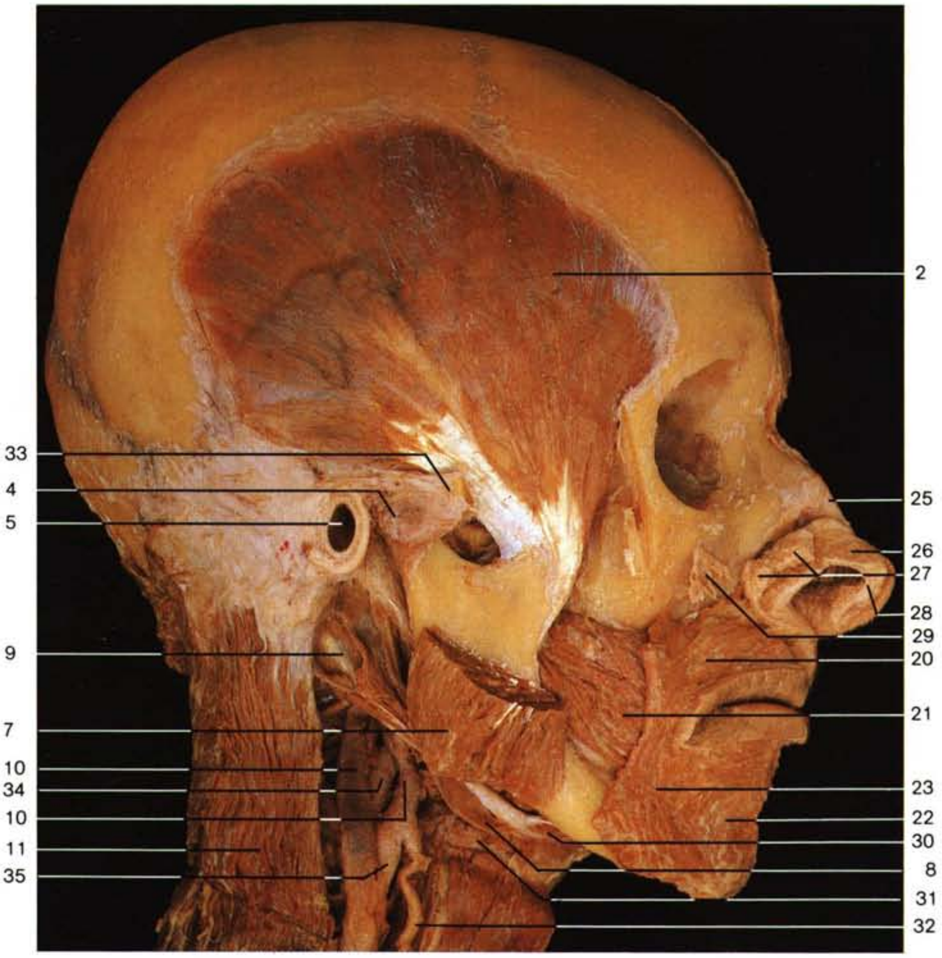

Temporalis muscle and temporomandibular joint. The zygomatic arch and the masseter muscle have been partially severed to display the insertion of the temporalis muscle.

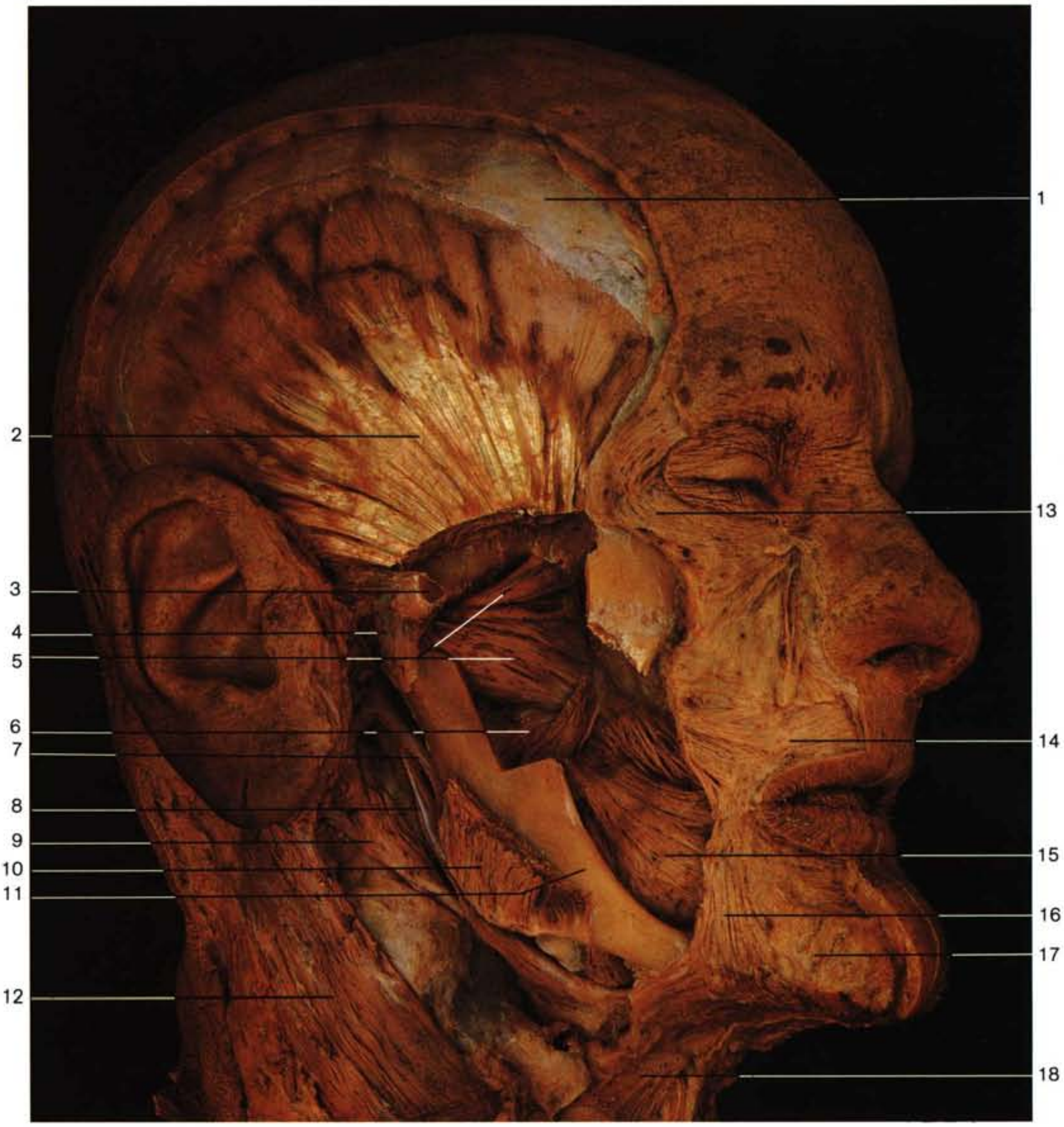

1 Periosteum
2 **Temporalis muscle**
3 Zygomatic arch
4 Articular capsule of temporomandibular joint
5 **Lateral pterygoid muscle** (upper and lower head)
6 **Medial pterygoid muscle**
7 Styloglossus muscle
8 Stylohyoid muscle
9 Posterior belly of digastric muscle
10 **Masseter muscle** (severed)
11 Mandible
12 Sternocleidomastoid muscle
13 Orbicularis oculi muscle
14 Orbicularis oris muscle
15 Buccinator muscle
16 Depressor anguli oris muscle
17 Depressor labii inferioris muscle
18 Platysma muscle
19 **Articular disc of temporomandibular joint**
20 Head of mandible
21 Anterior belly of digastric muscle
22 Mylohyoid muscle
23 Hyoid bone

Medial and lateral pterygoid muscles. A portion of the mandible and the zygomatic arch has been removed revealing the pterygoid region or infratemporal fossa.

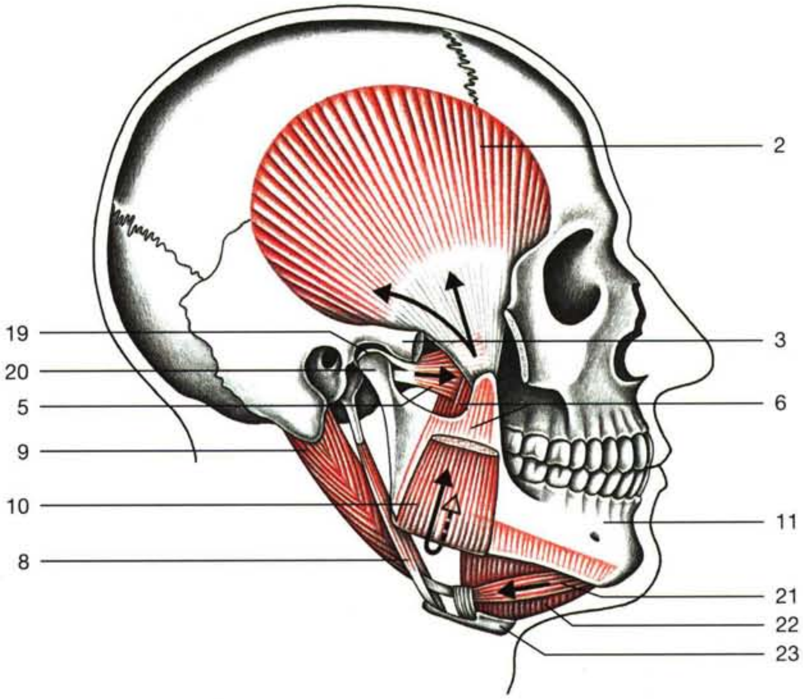

Effect of the muscles of mastication on the temporomandibular joint (arrows).

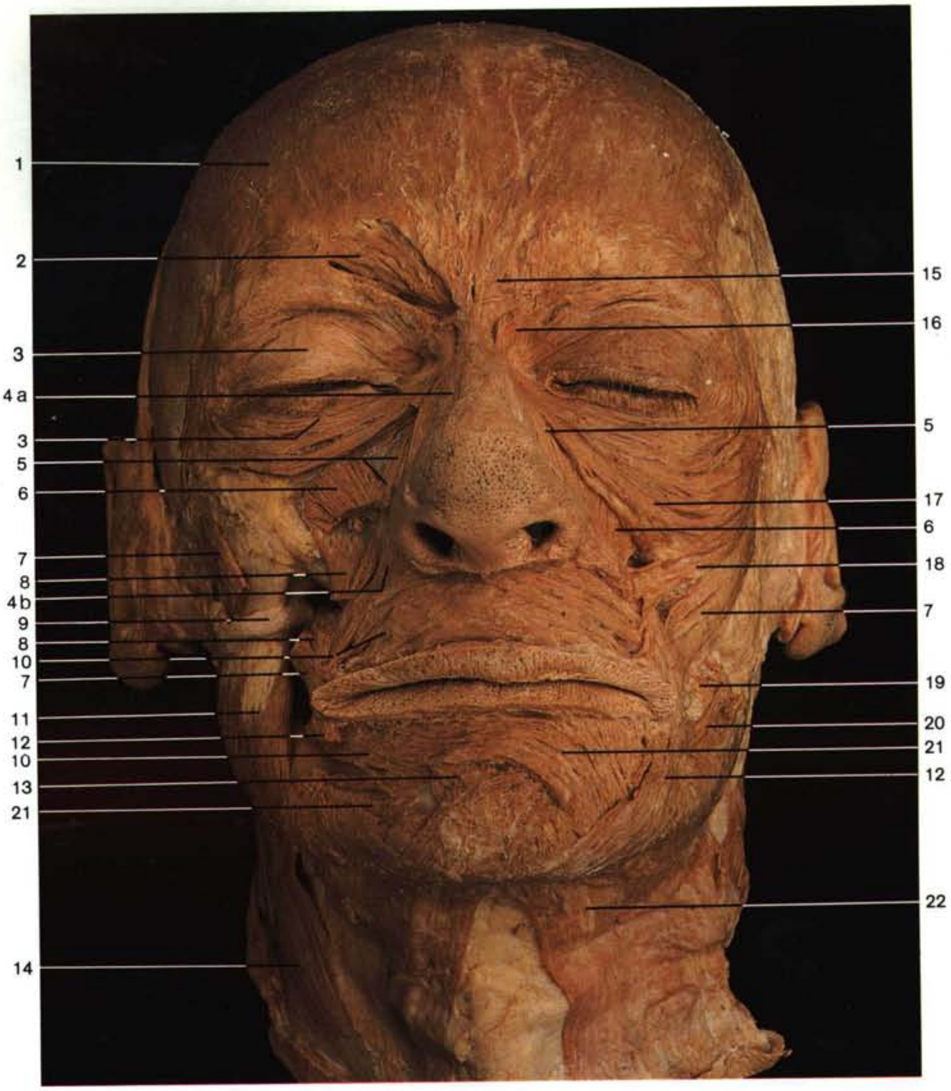

1 Frontal belly of **occipitofrontalis muscle**
2 **Corrugator supercilii muscle**
3 Palpebral part of orbicularis oculi muscle
4a Transverse part of nasalis muscle
4b Alar part of **nasalis muscle**
5 Levator labii superioris alaeque nasi muscle
6 **Levator labii superioris muscle**
7 **Zygomaticus major muscle**
8 Levator anguli oris muscle
9 Parotid duct
10 **Orbicularis oris muscle**
11 Masseter muscle
12 **Depressor anguli oris muscle**
13 **Mentalis muscle**
14 Sternocleidomastoid muscle
15 **Procerus muscle**
16 **Depressor supercilii muscle**
17 Orbital part of **orbicularis oculi muscle**
18 Zygomaticus minor muscle
19 **Buccinator muscle**
20 Risorius muscle
21 **Depressor labii inferioris muscle**
22 **Platysma muscle**
23 Galea aponeurotica
24 **Temporoparietalis muscle**
25 Occipital belly of occipitofrontalis muscle
26 Parotid gland with fascia
27 Temporal fascia
28 **Orbicularis oculi muscle**
29 Parotid duct, masseter muscle

Facial muscles (anterior aspect). Left side: superficial layer; right side: deeper layer.

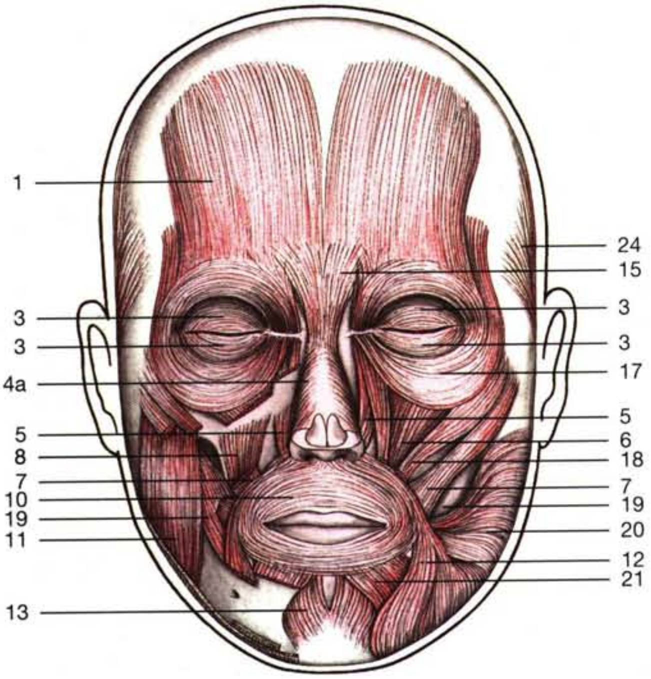

Facial muscles (schematic drawing). Left side: superficial layer; right side: deeper layer.

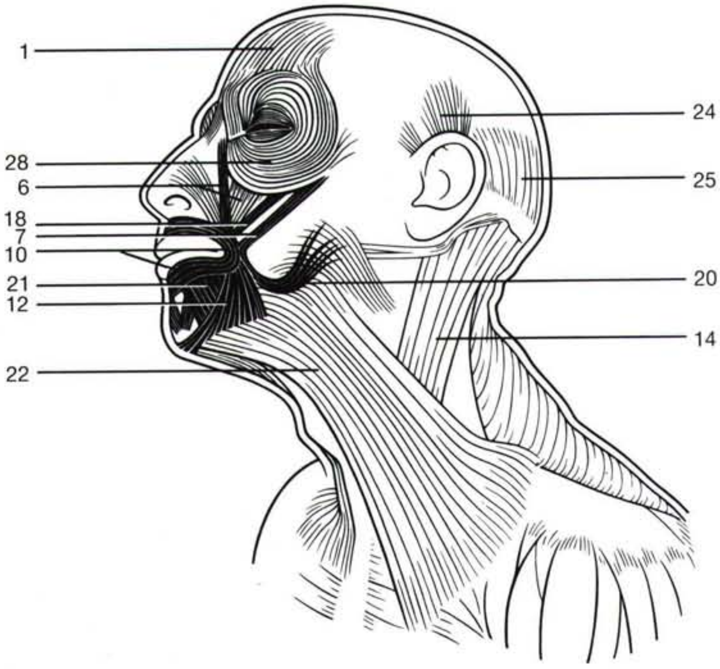

Facial muscles. Sphincter-like muscles surround the orifices of the head. Radially arranged muscles work as their antagonists.

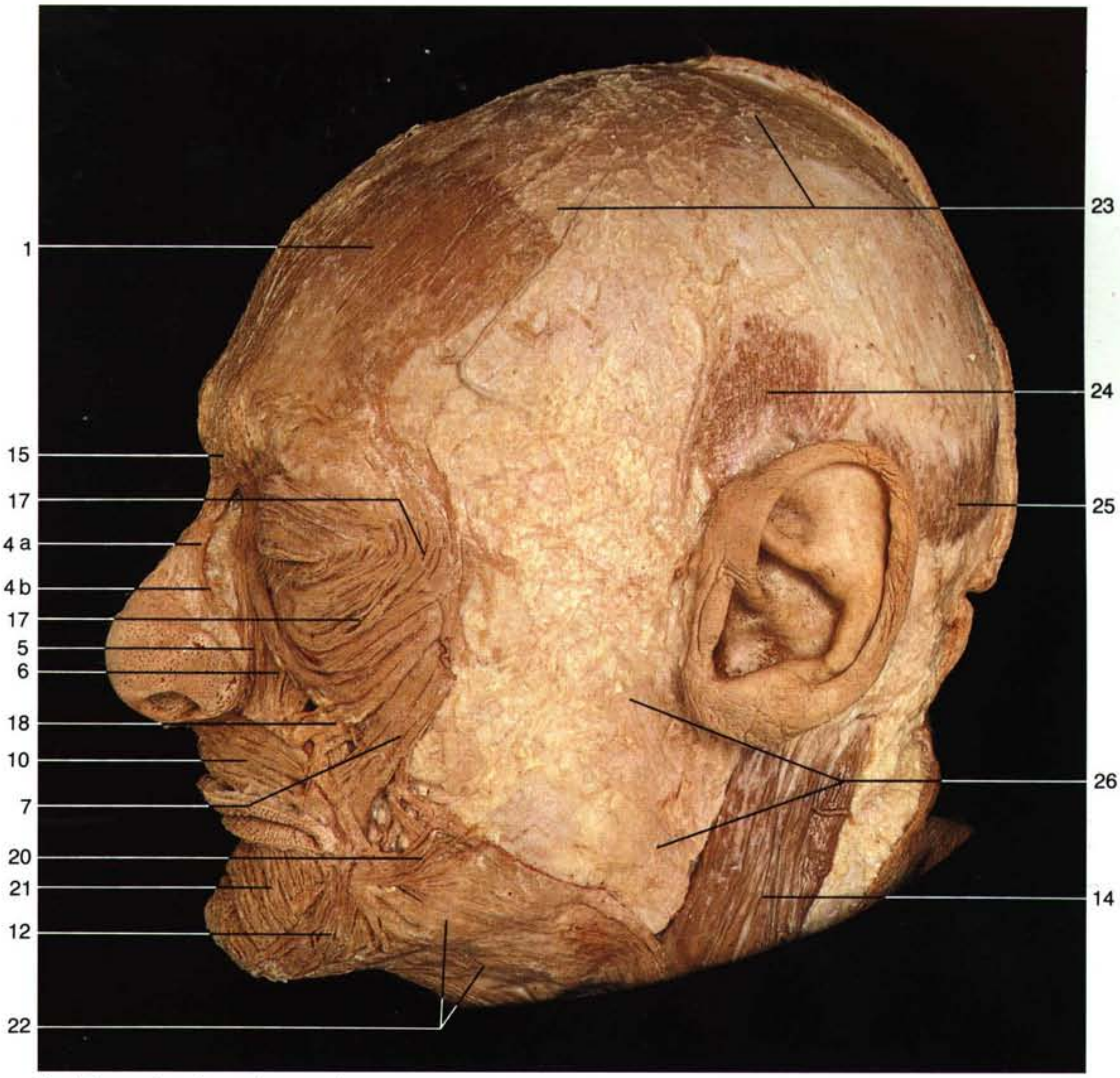
1
15
17
4 a
4 b
17
5
6
18
10
7
20
21
12
22
23
24
25
26
14

Facial muscles (lateral aspect).

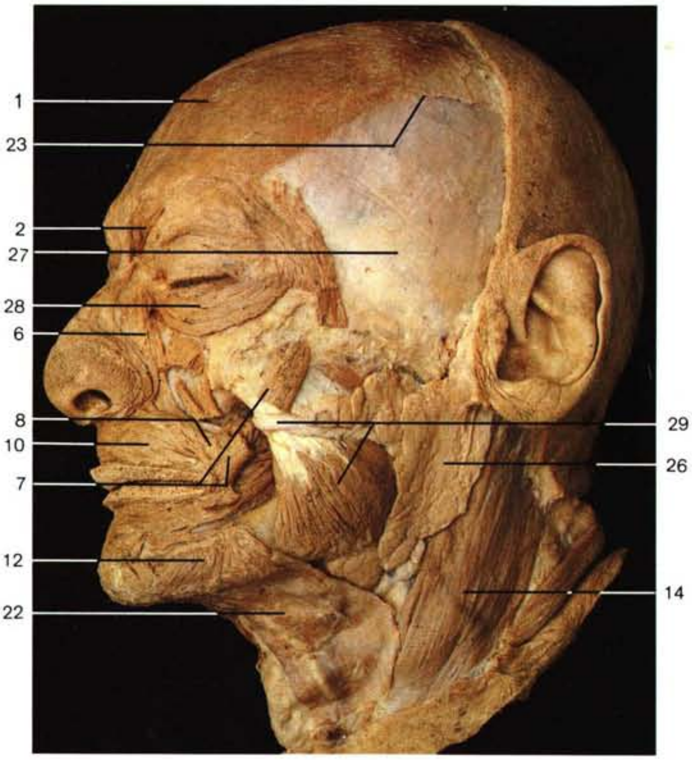
1
23
2
27
28
6
8
10
7
12
22
29
26
14

Facial muscles and parotid gland (lateral aspect).

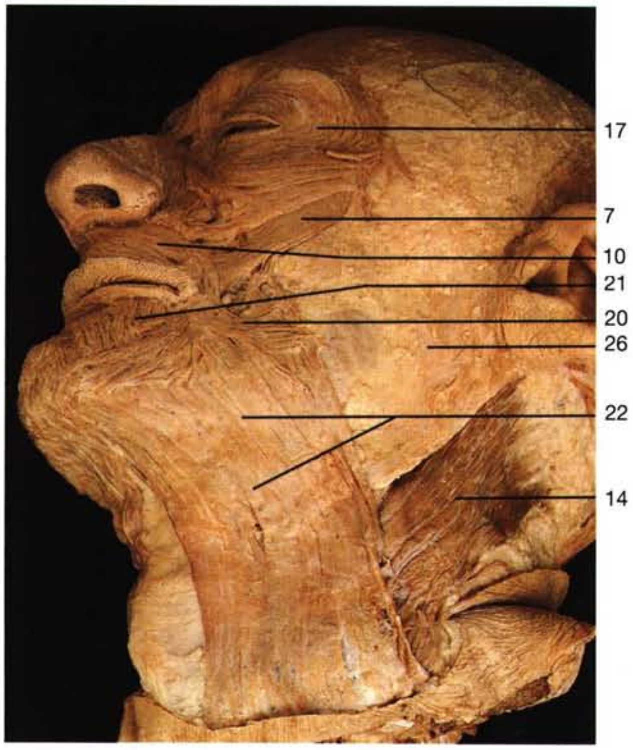
17
7
10
21
20
26
22
14

Platysma muscle (oblique lateral aspect). Superficial lamina of cervical fascia partly removed.

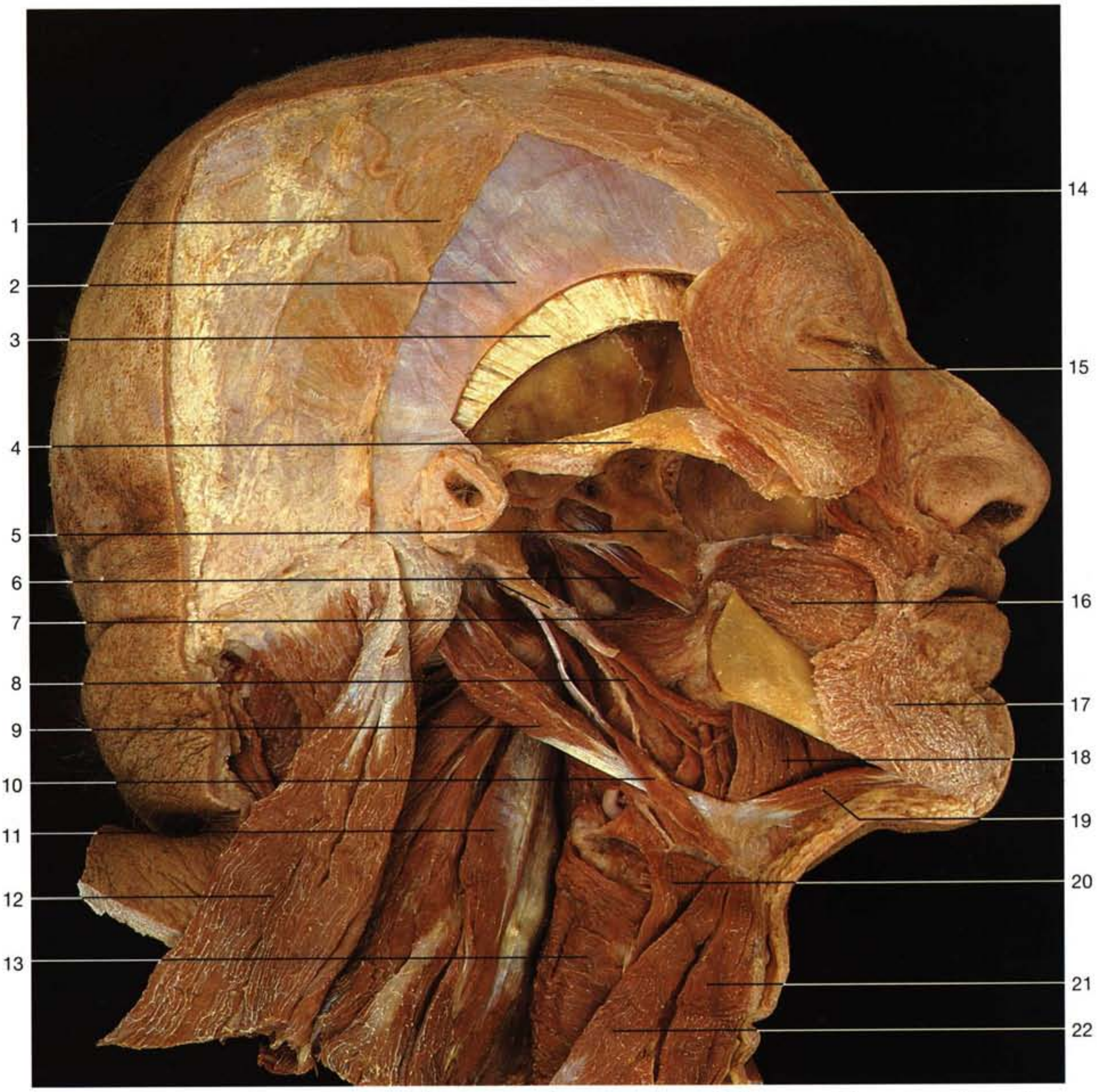

Supra- and infrahyoid muscles, pharynx I (lateral aspect). Ramus of mandible, pterygoid muscles and insertion of temporalis muscle removed.

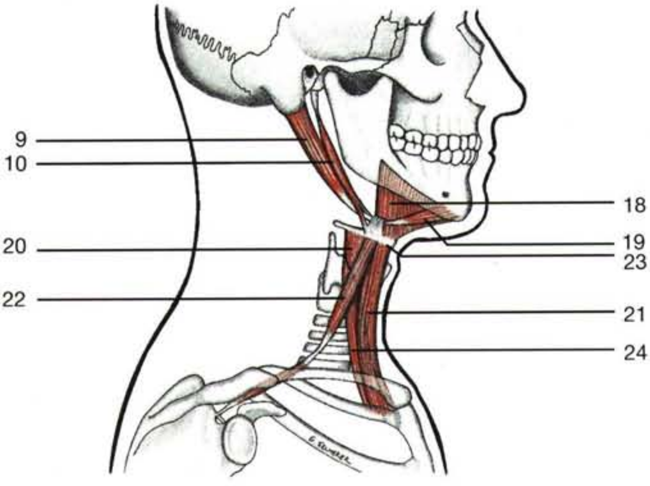

Supra- and infrahyoid muscles. (Schematic diagram.)

1 Galea aponeurotica
2 **Temporal fascia**
3 Tendon of temporalis muscle
4 Zygomatic arch
5 Lateral pterygoid plate
6 Tensor veli palatini muscle (styloid process)
7 Superior constrictor muscle of pharynx
8 **Styloglossus muscle**
9 Posterior belly of digastric muscle
10 **Stylohyoid muscle**
11 Longus capitis muscle
12 Sternocleidomastoid muscle (reflected)
13 Inferior constrictor of pharynx
14 Frontal belly of occipitofrontalis muscle
15 Orbital part of orbicularis oculi muscle
16 **Buccinator muscle**
17 Depressor anguli oris muscle
18 **Mylohyoid muscle**
19 Anterior belly of digastric muscle
20 **Thyrohyoid muscle**
21 **Sternohyoid muscle**
22 **Omohyoid muscle**
23 Hyoid bone
24 Sternothyroid muscle

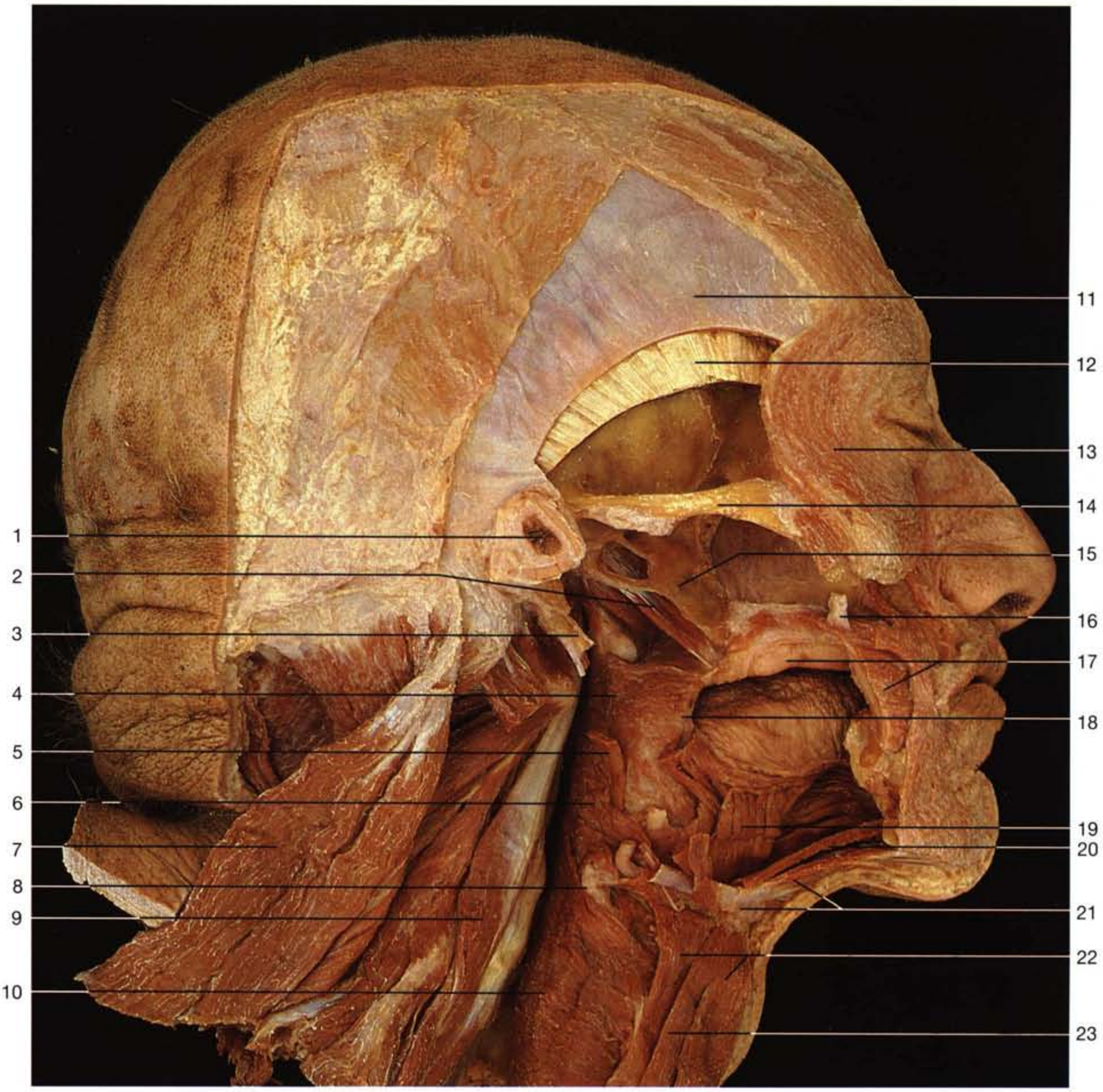

Supra- and infrahyoid muscles, pharynx II. Buccinator muscle removed; oral cavity opened.

1	External acoustic meatus	13	Orbicularis oculi muscle
2	Tensor veli palatini muscle	14	Zygomatic arch
3	Styloid process	15	Lateral pterygoid plate
4	**Superior constrictor muscle of pharynx**	16	Parotid duct
5	Stylopharyngeus muscle (divided)	17	Gingiva of upper jaw (without teeth), buccinator muscle (divided)
6	**Middle constrictor muscle of pharynx**	18	Pterygomandibular raphe
7	Sternocleidomastoid muscle	19	Hyoglossus muscle
8	Greater horn of hyoid bone	20	Mylohyoid muscle
9	Longus capitis	21	Anterior belly of digastric muscle (hyoid bone)
10	**Inferior constrictor muscle of pharynx**	22	Sternohyoid and thyrohyoid muscles
11	Temporal fascia	23	Omohyoid muscle
12	Tendon of temporalis muscle		

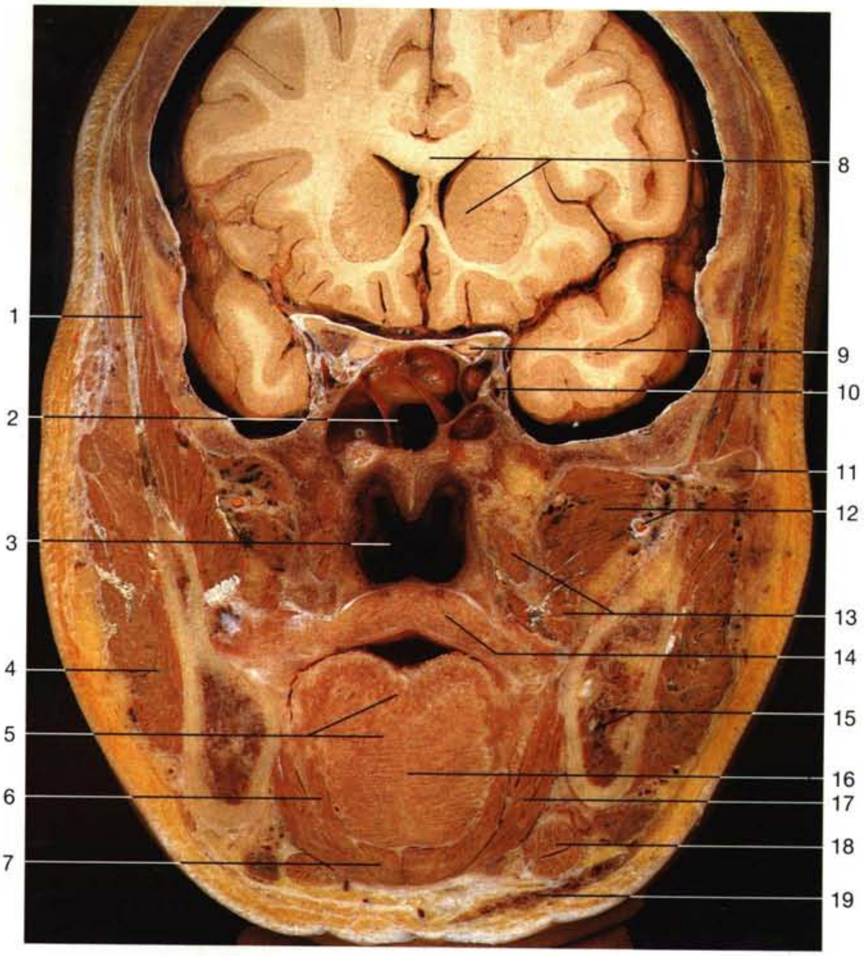

Coronal section through cranial, nasal and oral cavity at the level of sphenoidal sinus.

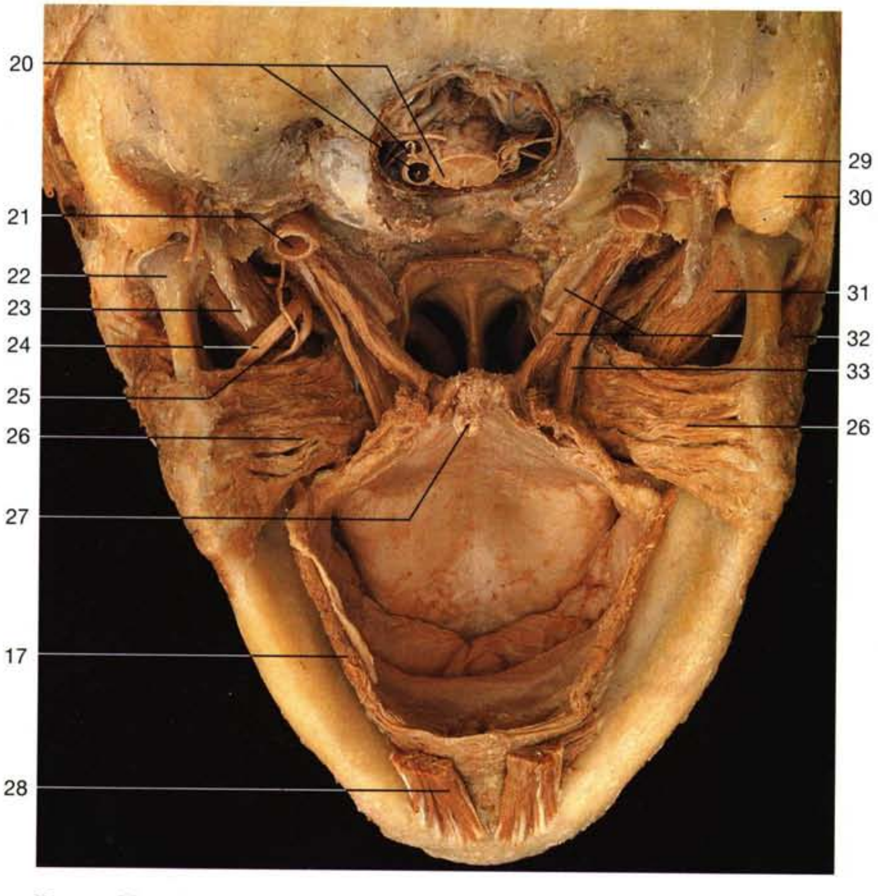

Pterygoid and palatine muscles (posterior aspect).

1 Temporalis muscle
2 **Sphenoidal sinus**
3 **Nasopharynx**
4 Masseter muscle
5 Superior longitudinal, transverse and vertical muscles of tongue
6 Hyoglossus muscle
7 Geniohyoid muscle
8 Corpus callosum (caudate nucleus)
9 Optic nerve
10 Cavernous sinus
11 Zygomatic arch
12 Cross section of lateral pterygoid muscle and maxillary artery
13 Section of medial pterygoid muscle
14 **Soft palate**
15 Mandible and inferior alveolar nerve
16 Septum of the tongue
17 **Mylohyoid muscle**
18 Submandibular gland
19 Platysma muscle
20 Foramen magnum, vertebral artery and spinal cord
21 Internal carotid artery
22 **Head of mandible**
23 Styloid process
24 Inferior alveolar nerve
25 Lingual nerve and chorda tympani nerve
26 **Medial pterygoid muscle**
27 Uvula
28 Anterior belly of digastric muscle (cut)
29 Condyle of occipital bone
30 Mastoid process
31 **Lateral pterygoid muscle**
32 Auditory tube and **levator veli palatini muscle**
33 **Tensor veli palatini muscle**

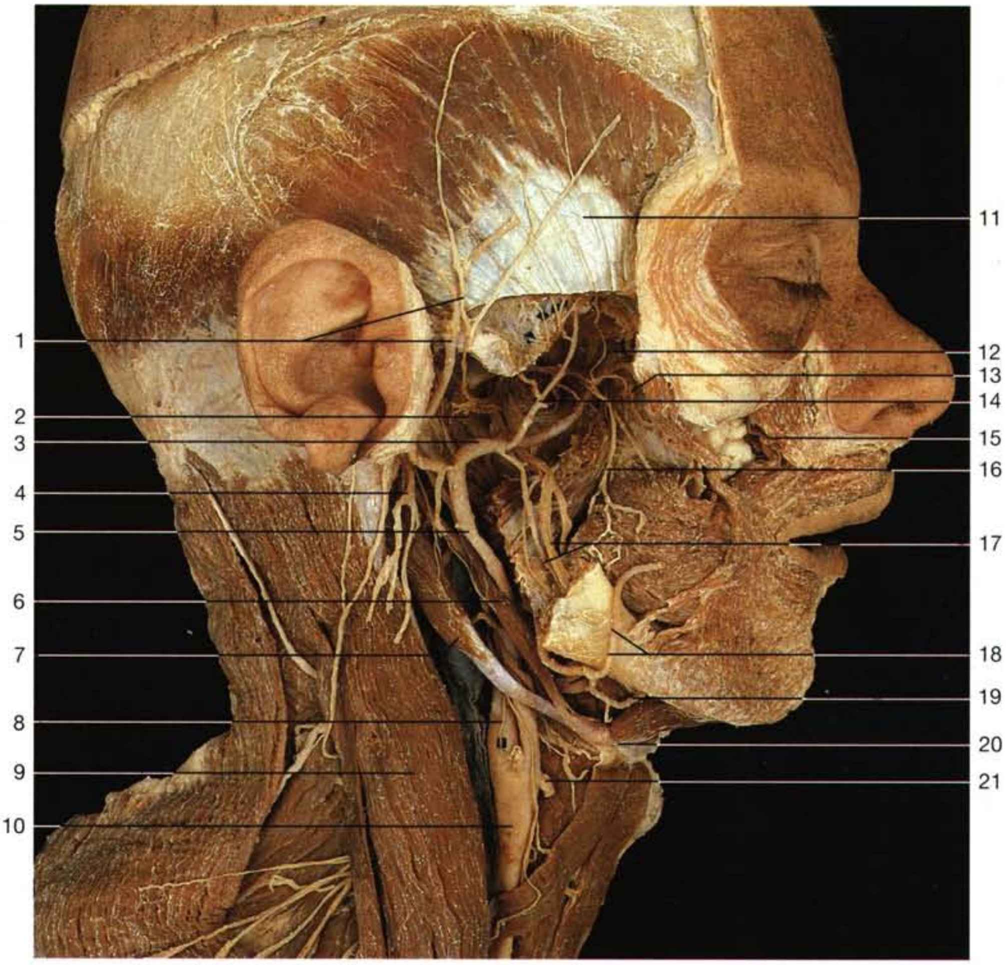

1 Superficial temporal artery and auriculotemporal nerve
2 Middle meningeal artery
3 **Maxillary artery**
4 Facial nerve (divided and reflected)
5 **External carotid artery**
6 Stylohyoid muscle
7 Digastric muscle (posterior belly)
8 Internal carotid artery and carotid sinus branch of glossopharyngeal nerve
9 Sternocleidomastoid muscle
10 Common carotid artery
11 Temporalis muscle
12 Pterygopalatine fossa
13 Posterior superior alveolar artery
14 Deep temporal artery
15 Infraorbital artery
16 Buccal nerve
17 Inferior alveolar artery and nerve
18 **Facial artery**
19 Submental artery
20 Hyoid bone
21 Superior thyroid artery (divided)

Dissection of maxillary artery. Mandible and lateral pterygoid muscle partly removed.

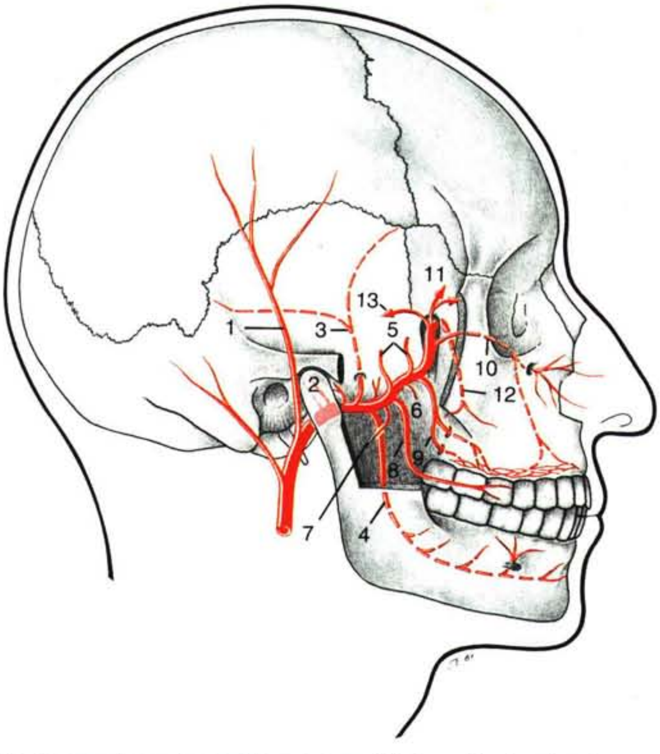

1 Superficial temporal artery

Branches of the first part
2 Deep auricular artery and anterior tympanic artery
3 Middle meningeal artery
4 Inferior alveolar artery

Branches of the second part
5 Deep temporal branches
6 Pterygoid branches
7 Masseteric artery
8 Buccal artery

Branches of the third part
9 Posterior superior alveolar artery
10 Infraorbital artery
11 Sphenopalatine artery and branches to the nasal cavity
12 Descending palatine artery
13 Artery of the pterygoid canal

Main branches of maxillary artery. (Schematic drawing.)

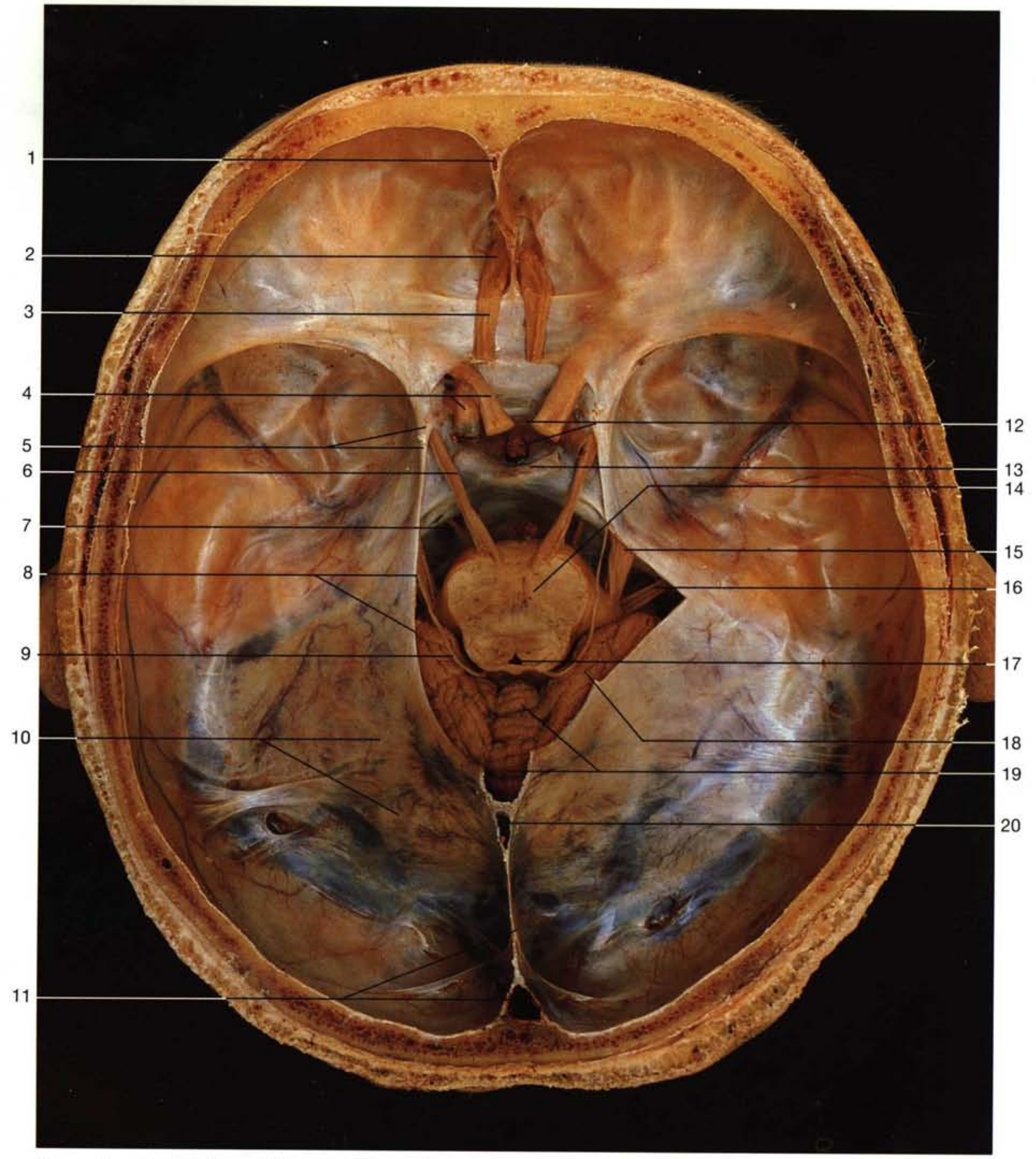

Base of the skull with cranial nerves (internal aspect). Both cerebral hemispheres and upper part of the brain stem removed. Incision on the right tentorium cerebelli to display the cranial nerves of the infratentorial space.

1 Superior sagittal sinus with falx cerebri
2 Olfactory bulb
3 **Olfactory tract**
4 **Optic nerve** and internal carotid artery
5 Anterior clinoid process and anterior attachment of tentorium cerebelli
6 **Oculomotor nerve** (n. III)
7 **Abducens nerve** (n. VI)
8 Tentorial notch (incisura tentorii)
9 **Trochlear nerve** (n. IV)
10 Tentorium cerebelli
11 Falx cerebri and confluence of sinuses
12 Hypophysial fossa, infundibulum, and diaphragma sellae
13 Dorsum sellae
14 Midbrain (divided)
15 **Trigeminal nerve** (n. V)
16 **Facial nerve** (n. VII), nervus intermedius, and **vestibulocochlear nerve** (n. VIII)
17 Cerebral aqueduct
18 Right hemisphere of cerebellum
19 Vermis of cerebellum
20 Straight sinus

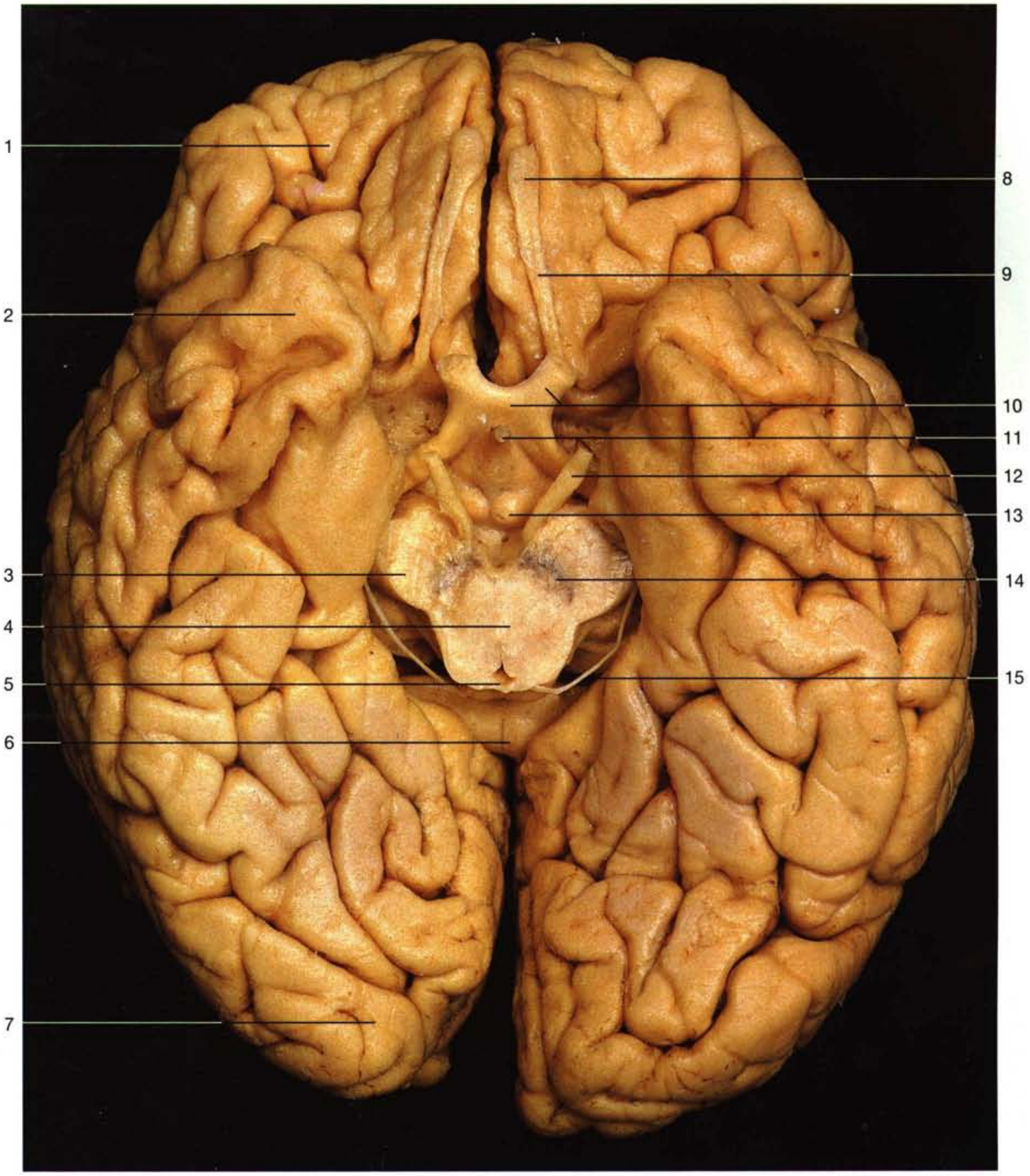

Inferior aspect of the brain with cranial nerves. Midbrain divided.

1 **Frontal lobe**
2 **Temporal lobe**
3 Pedunculus cerebri
4 **Midbrain** (divided)
5 Cerebral aqueduct
6 Splenium of corpus callosum
7 **Occipital lobe**
8 Olfactory bulb
9 Olfactory tract
10 **Optic nerve and optic chiasma**
11 Infundibulum
12 **Oculomotor nerve** (n. III)
13 Mamillary body
14 Substantia nigra
15 **Trochlear nerve** (n. IV)

Cranial nerves

I = Olfactory nerves
II = Optic nerve
III = Oculomotor nerve
IV = Trochlear nerve
V = Trigeminal nerve
VI = Abducens nerve
VII = Facial nerve
VIII = Vestibulocochlear nerve
IX = Glossopharyngeal nerve
X = Vagus nerve
XI = Accessory nerve
XII = Hypoglossal nerve

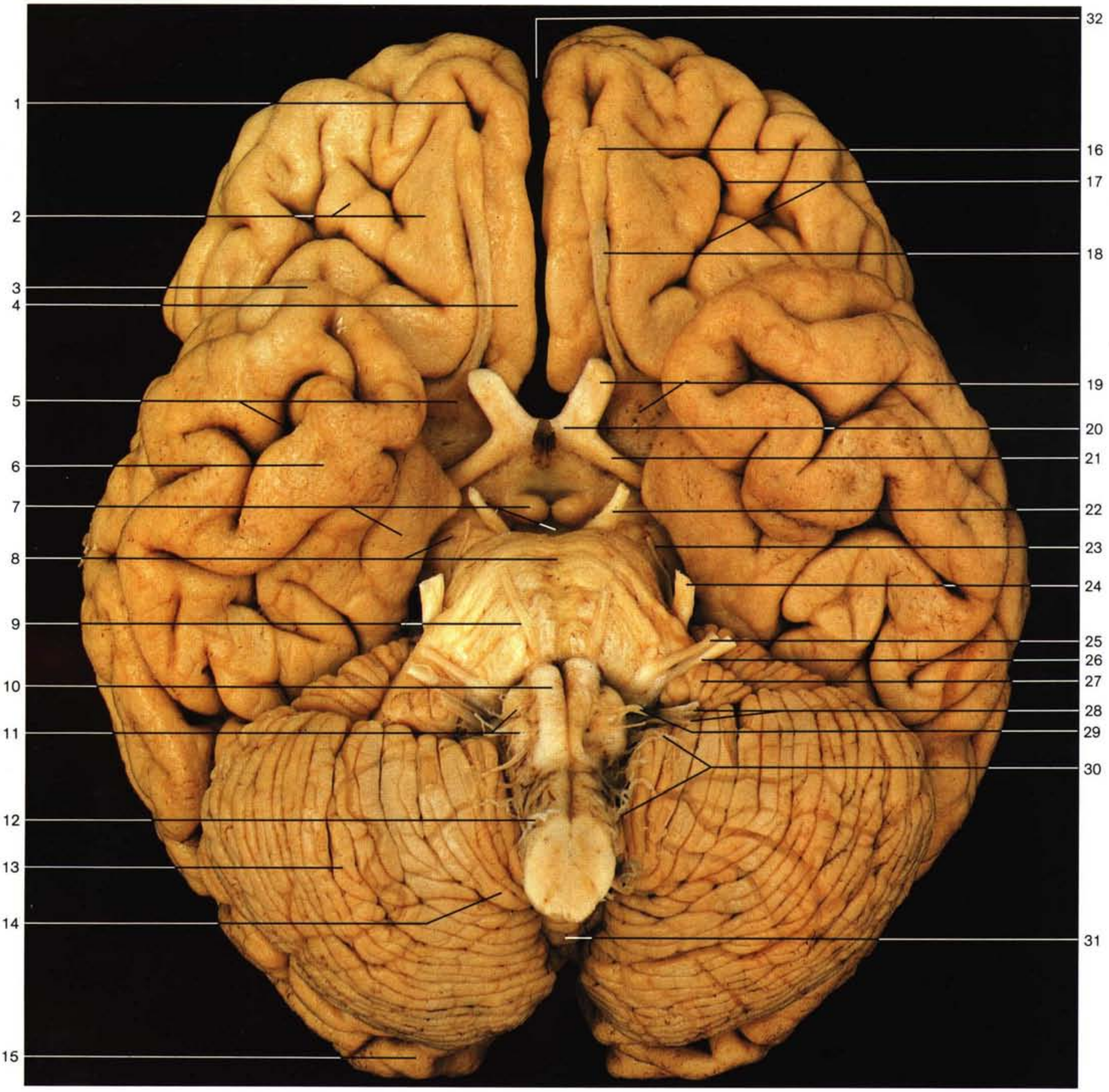

Cranial nerves. Brain (inferior aspect).

1 Olfactory sulcus (termination)
2 Orbital gyri
3 Temporal lobe
4 Straight gyrus
5 Olfactory trigone and inferior temporal sulcus
6 Medial occipitotemporal gyrus
7 Parahippocampal gyrus, mamillary body, and interpeduncular fossa
8 Pons and cerebral peduncle
9 **Abducens nerve** (n. VI)
10 Pyramid
11 Lower part of olive
12 Cervical spinal nerves
13 Cerebellum
14 Tonsil of cerebellum
15 Occipital lobe (posterior pole)
16 Olfactory bulb
17 Orbital sulci of frontal lobe
18 **Olfactory tract**
19 **Optic nerve** (n. II) and anterior perforated substance
20 Optic chiasma
21 Optic tract
22 **Oculomotor nerve** (n. III)
23 **Trochlear nerve** (n. IV)
24 **Trigeminal nerve** (n. V)
25 **Facial nerve** (n. VII)
26 **Vestibulocochlear nerve** (n. VIII)
27 Flocculus of cerebellum
28 **Glossopharyngeal** (n. IX) and **vagus nerve** (n. X)
29 **Hypoglossal nerve** (n. XII)
30 **Accessory nerve** (n. XI)
31 Vermis of cerebellum
32 Longitudinal fissure

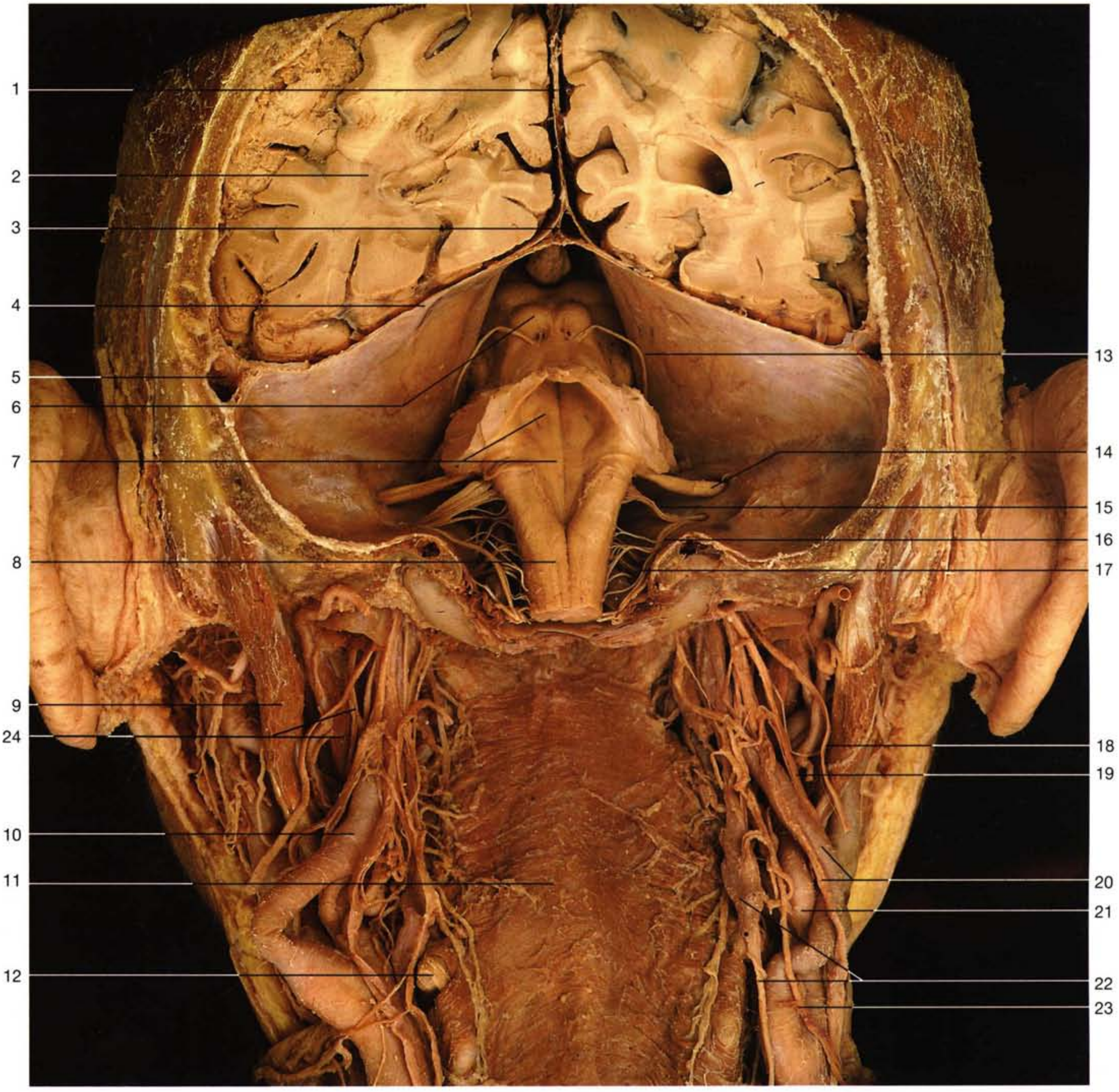

Brain stem and pharynx with cranial nerves (posterior aspect). Cranial cavity opened and cerebellum removed.

1 Falx cerebri
2 Occipital lobe
3 Straight sinus
4 Tentorium cerebelli
5 Transverse sinus
6 Inferior colliculus of midbrain
7 **Rhomboid fossa**
8 Medulla oblongata
9 Posterior belly of digastric muscle
10 Internal carotid artery
11 Pharynx (middle constrictor muscle)
12 Hyoid bone (greater horn)
13 **Trochlear nerve** (n. IV)
14 **Facial nerve** (n. VII), **vestibulocochlear nerve** (n. VIII)
15 **Glossopharyngeal nerve** (n. IX) and vagus nerve (n. X)
16 **Accessory nerve** (intracranial portion) (n. XI)
17 **Hypoglossal nerve** (intracranial portion) (n. XII)
18 **Accessory nerve** (n. XI)
19 **Hypoglossal nerve** (n. XII)
20 **Vagus nerve** (n. X) and internal carotid artery
21 External carotid artery
22 **Sympathetic trunk** and superior cervical ganglion
23 Ansa cervicalis (superior root of hypoglossal nerve)
24 **Glossopharyngeal nerve** (n. IX) and stylopharyngeus muscle

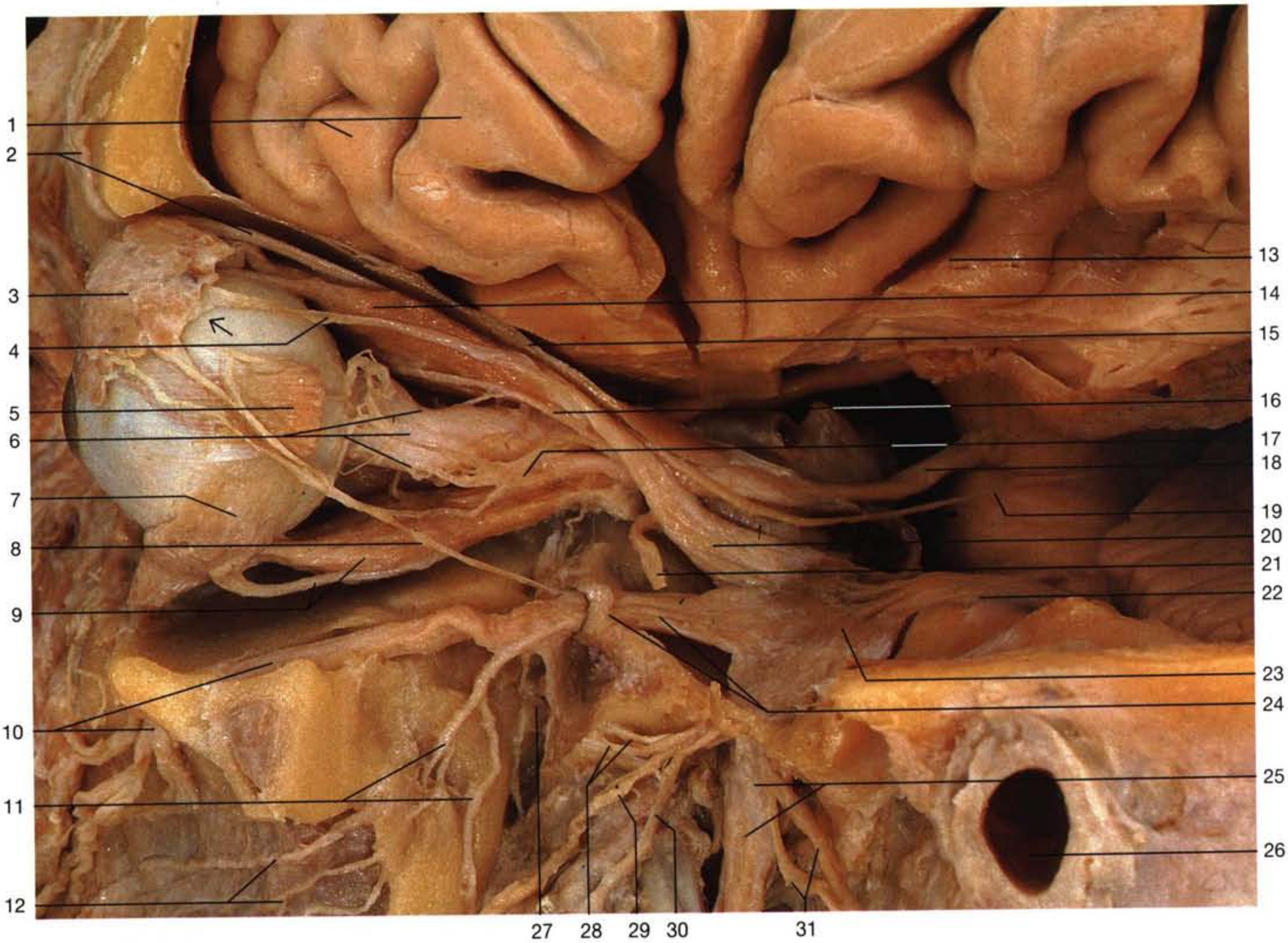

Cranial nerves of the orbit and pterygopalatine fossa. Left orbit (lateral aspect). Note the zygomaticolacrimal anastomosis (arrow).

1 Frontal lobe
2 Supraorbital nerve
3 Lacrimal gland
4 Lacrimal nerve
5 Lateral rectus muscle (divided)
6 **Optic nerve** and short ciliary nerves
7 Inferior oblique muscle
8 Zygomatic nerve
9 Inferior branch of oculomotor nerve and inferior rectus muscle
10 Infraorbital nerve
11 Posterior superior alveolar nerves
12 Branches of superior alveolar plexus adjacent to mucous membrane of maxillary sinus
13 Central sulcus of insula
14 Superior rectus muscle
15 Periorbita (roof of orbit)
16 Nasociliary nerve
17 Ciliary ganglion
18 **Oculomotor nerve** (n. III)
19 **Trochlear nerve** (n. IV)
20 **Ophthalmic nerve** (n. V_1)
21 **Abducens nerve** (n. VI) (divided)
22 **Trigeminal nerve** (n. V)
23 Trigeminal ganglion
24 **Maxillary nerve** (n. V_2) and foramen rotundum
25 **Mandibular nerve** (n. V_3)
26 External acoustic meatus
27 Pterygopalatine nerves
28 Deep temporal nerves
29 Buccal nerve
30 Masseteric nerve
31 Auriculotemporal nerve
32 Trochlea and superior oblique muscle

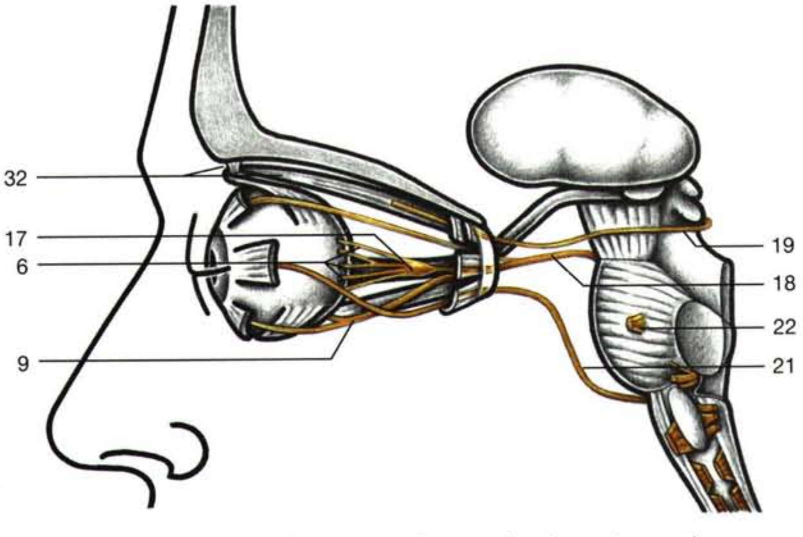

Cranial nerves innervating extraocular muscles (lateral aspect). (Schematic drawing.)

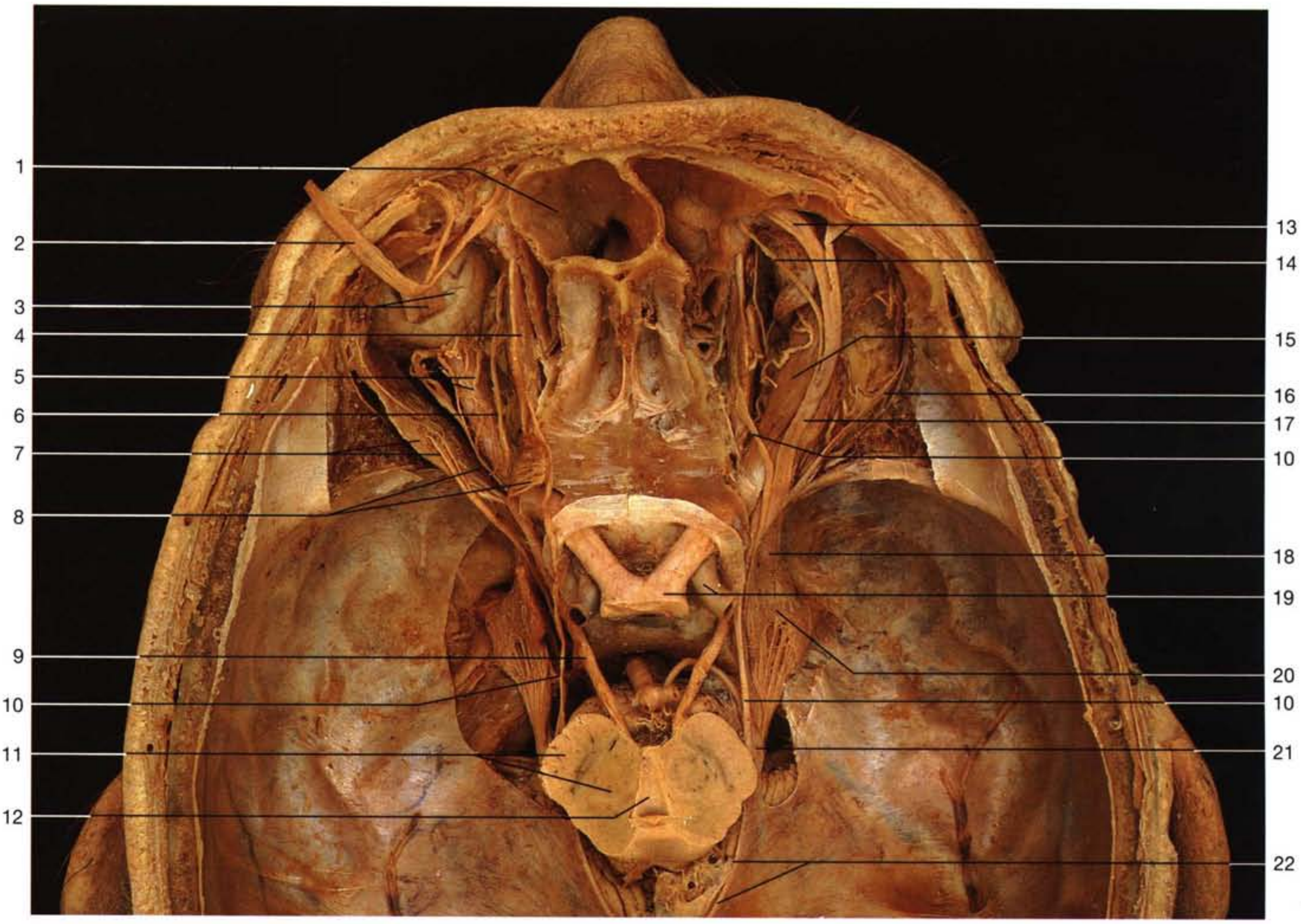

Cranial nerves of the orbit (superior aspect). Right side: superficial layer; left side: middle layer of the orbit (superior rectus muscle and frontal nerve divided and reflected). Tentorium and dura mater partly removed.

1 Frontal sinus (enlarged)
2 Frontal nerve (divided and reflected)
3 Superior rectus muscle (divided) and eyeball
4 Superior oblique muscle
5 Short ciliary nerves and optic nerve (n. II)
6 Nasociliary nerve
7 **Abducens nerve** (n. VI) and lateral rectus muscle
8 Ciliary ganglion and superior rectus muscle (reflected)
9 **Oculomotor nerve** (n. III)
10 **Trochlear nerve** (n. IV)
11 Crus cerebri and midbrain
12 Inferior wall of the third ventricle connected with cerebral aqueduct
13 Lateral and medial branch of supraorbital nerve
14 Supratrochlear nerve
15 Superior rectus muscle
16 Lacrimal nerve
17 Frontal nerve
18 **Ophthalmic nerve** (n. V1)
19 Optic chiasma and internal carotid artery
20 Trigeminal ganglion
21 **Trigeminal nerve** (n. V)
22 Tentorial notch
23 Falx cerebri
24 Cerebellum
25 Infundibulum
26 Olfactory tract

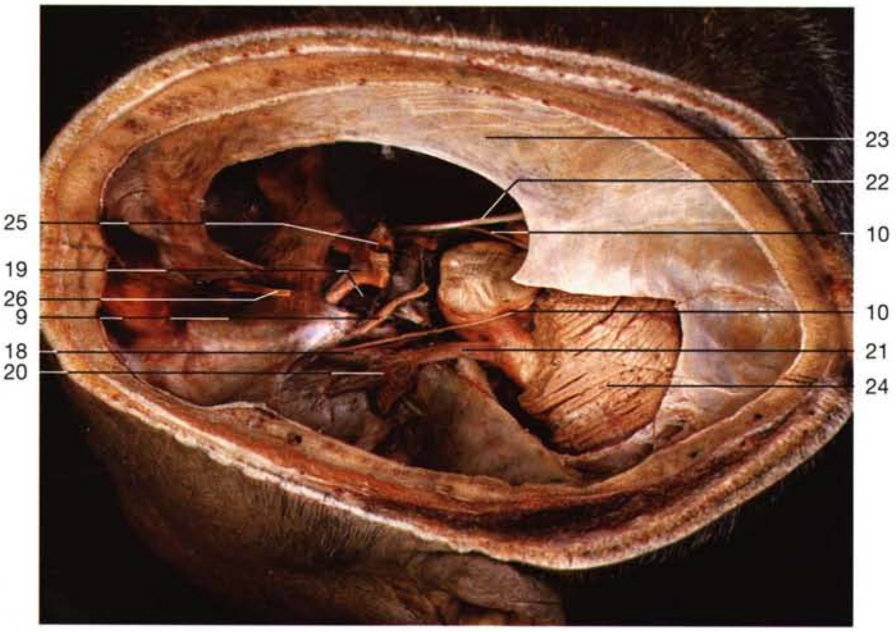

Cranial nerves at the skull base. The brain stem was divided and the tentorium fenestrated. Both hemispheres were removed.

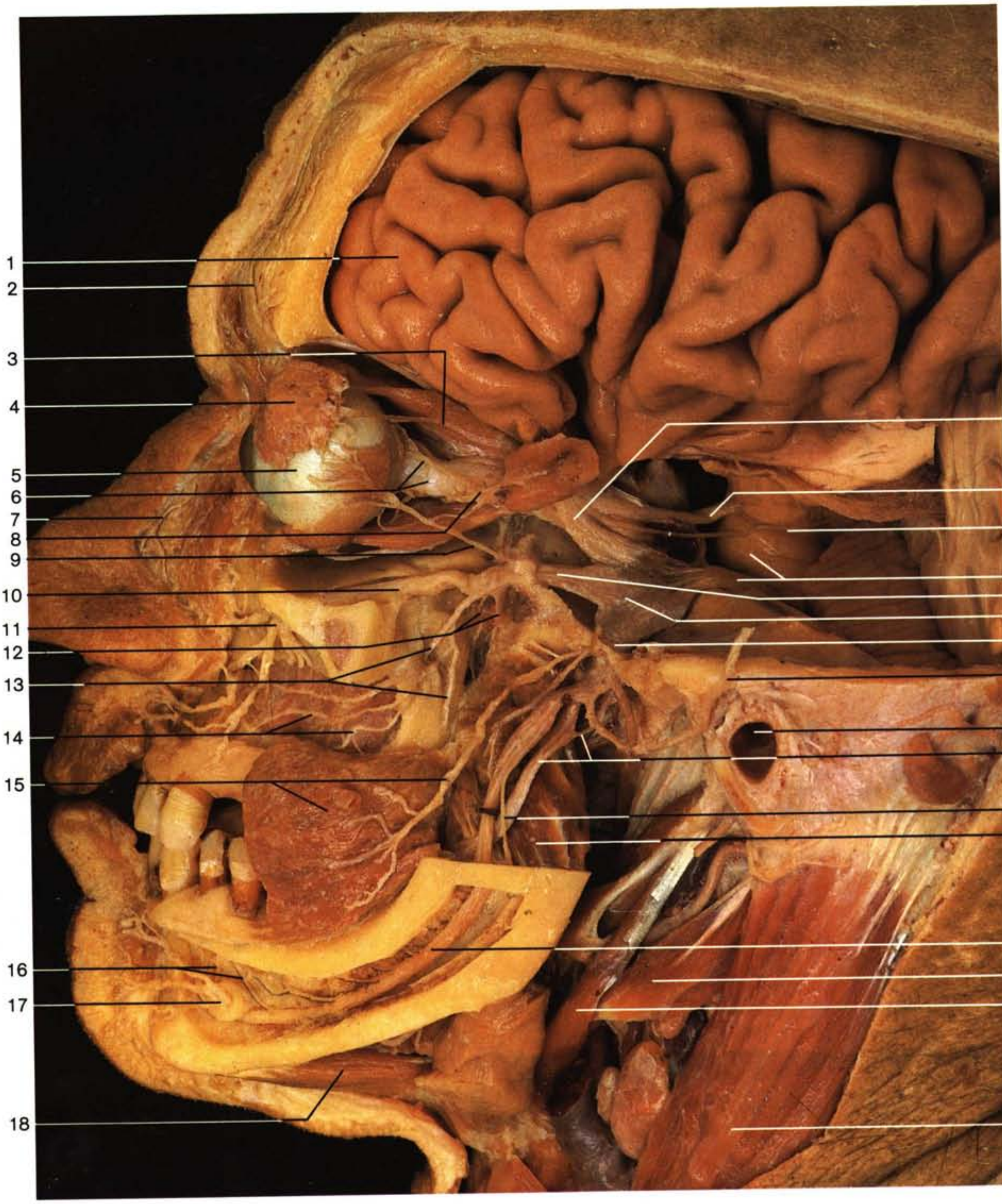

Dissection of the trigeminal nerve in its entirety. Lateral wall of cranial cavity, lateral wall of orbit, zygomatic arch and ramus of the mandible have been removed and the mandibular canal opened.

1 Frontal lobe of cerebrum
2 Supraorbital nerve
3 Lacrimal nerve
4 Lacrimal gland
5 Eyeball
6 **Optic nerve** and short ciliary nerves
7 External nasal branch of anterior ethmoidal nerve
8 Ciliary ganglion
9 Zygomatic nerve
10 Infraorbital nerve
11 Infraorbital foramen and terminal branches of infraorbital nerve
12 Pterygopalatine ganglion and pterygopalatine nerves
13 Posterior superior alveolar nerves
14 Superior dental plexus
15 Buccinator muscle and buccal nerve
16 Inferior dental plexus
17 Mental foramen and mental nerve
18 Anterior belly of digastric muscle
19 **Ophthalmic nerve** (n. V_1)
20 **Oculomotor nerve** (n. III)
21 **Trochlear nerve** (n. IV)
22 Trigeminal nerve and pons
23 **Maxillary nerve** (n. V_2)
24 Trigeminal ganglion
25 **Mandibular nerve** (n. V_3)
26 Auriculotemporal nerve
27 External acoustic meatus (divided)
28 Lingual nerve and chorda tympani
29 Mylohyoid nerve
30 Medial pterygoid muscle
31 Inferior alveolar nerve
32 Posterior belly of digastric muscle
33 Stylohyoid muscle
34 Sternocleidomastoid muscle

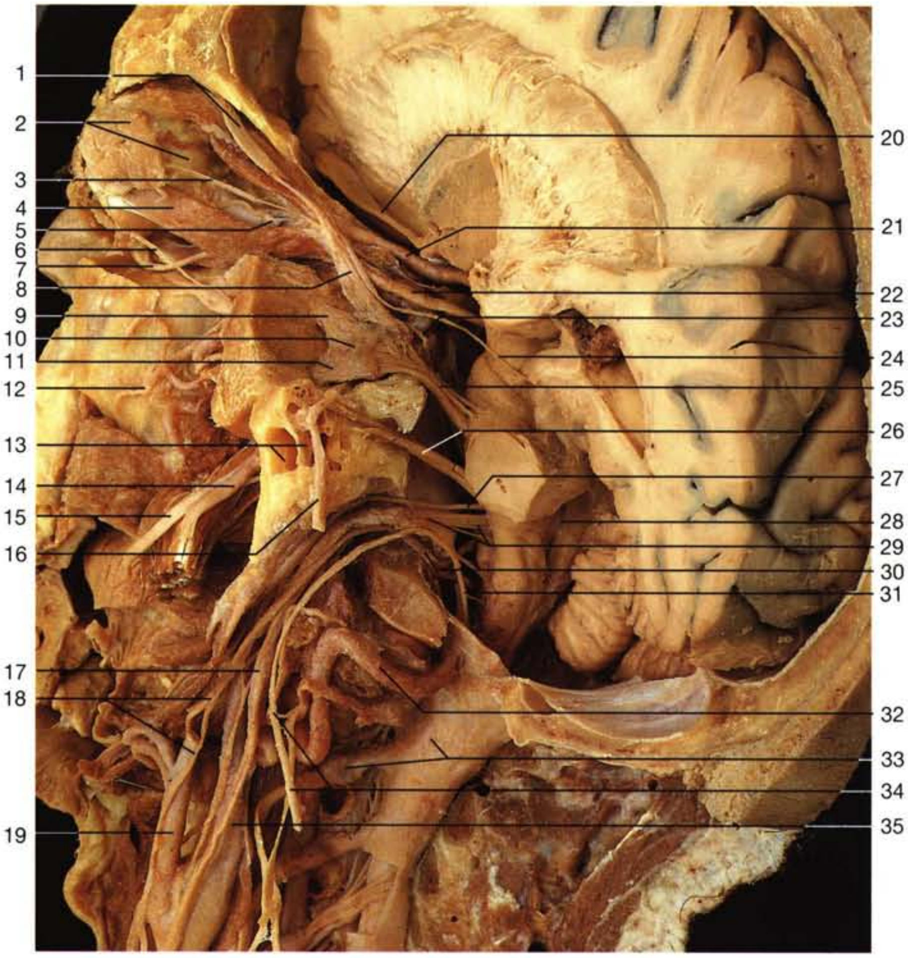

Cranial nerves in connection with the brain stem. Left side (lateral superior aspect). Left half of brain and head partly removed. Notice the location of trigeminal ganglion.

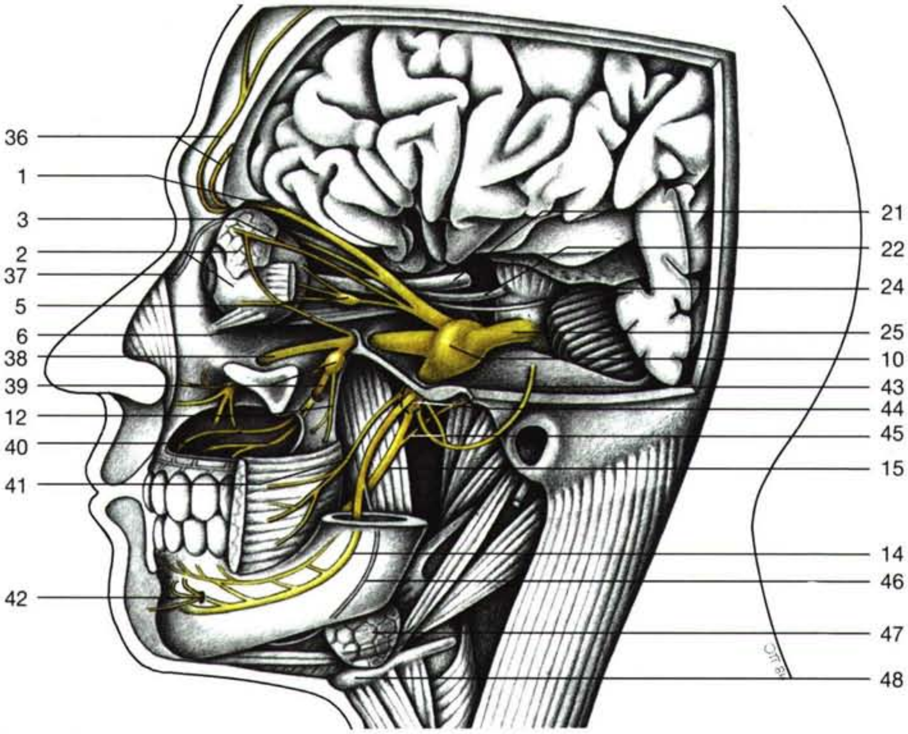

Main branches of trigeminal nerve. (Schematic drawing of figure on opposite page.)

1 Frontal nerve
2 Lacrimal gland and eyeball
3 Lacrimal nerve
4 Lateral rectus muscle
5 **Ciliary ganglion** lateral to optic nerve
6 Zygomatic nerve
7 Inferior branch of oculomotor nerve
8 **Ophthalmic nerve** (n. V_1)
9 **Maxillary nerve** (n. V_2)
10 **Trigeminal ganglion**
11 **Mandibular nerve** (n. V_3)
12 Posterior superior alveolar nerves
13 Tympanic cavity, external acoustic meatus, and tympanic membrane
14 Inferior alveolar nerve
15 Lingual nerve
16 **Facial nerve** (n. VII)
17 **Vagus nerve** (n. X)
18 **Hypoglossal nerve** (n. XII) and superior root of ansa cervicalis
19 External carotid artery
20 **Olfactory tract** (n. I)
21 **Optic nerve** (n. II) (intracranial part)
22 **Oculomotor nerve** (n. III)
23 **Abducens nerve** (n. VI)
24 **Trochlear nerve** (n. IV)
25 **Trigeminal nerve** (n. V)
26 **Vestibulocochlear nerve** (n. VIII) and **facial nerve** (VII)
27 **Glossopharyngeal nerve** (n. IX) (leaving brain stem)
28 Rhomboid fossa
29 **Vagus nerve** (n. X) (leaving brain stem)
30 **Hypoglossal nerve** (n. XII) (leaving medulla oblongata)
31 **Accessory nerve** (n. XI) (ascending from foramen magnum)
32 Vertebral artery
33 Spinal ganglion and dura mater of spinal cord
34 Accessory nerve (n. XI)
35 Internal carotid artery
36 Lateral and medial branch of supraorbital nerve
37 Infratrochlear nerve
38 Infraorbital nerve
39 **Pterygopalatine ganglion** and middle superior alveolar nerve
40 Middle superior alveolar nerves (entering superior dental plexus)
41 Buccal nerve
42 Mental nerve and mental foramen
43 Auriculotemporal nerve
44 Otic ganglion (dotted line)
45 Chorda tympani
46 Mylohyoid nerve
47 Submandibular gland
48 Hyoid bone

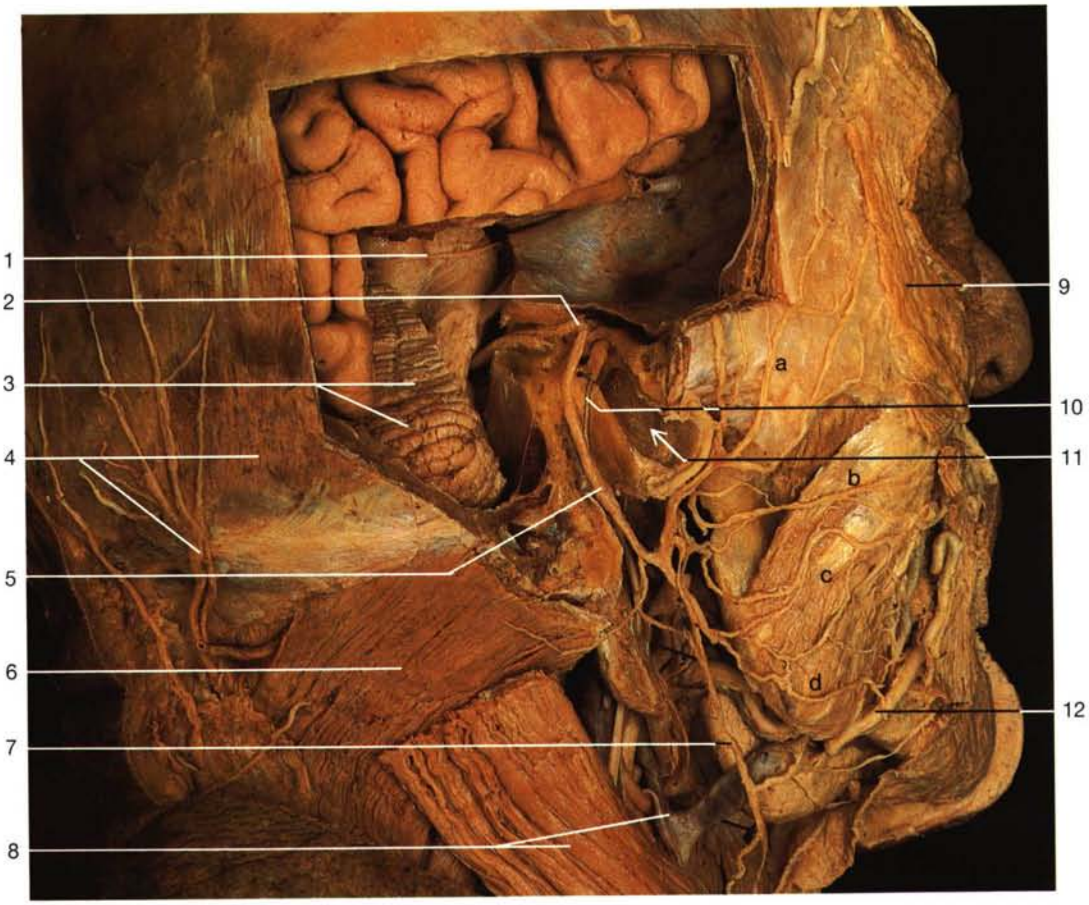

Dissection of facial nerve in its entirety. Cranial cavity fenestrated; temporal lobe partly removed. Facial canal and tympanic cavity opened, posterior wall of external acoustic meatus removed. **Branches of facial nerve:** a = temporal branch; b = zygomatic branches; c = buccal branches; d = marginal mandibular branch.

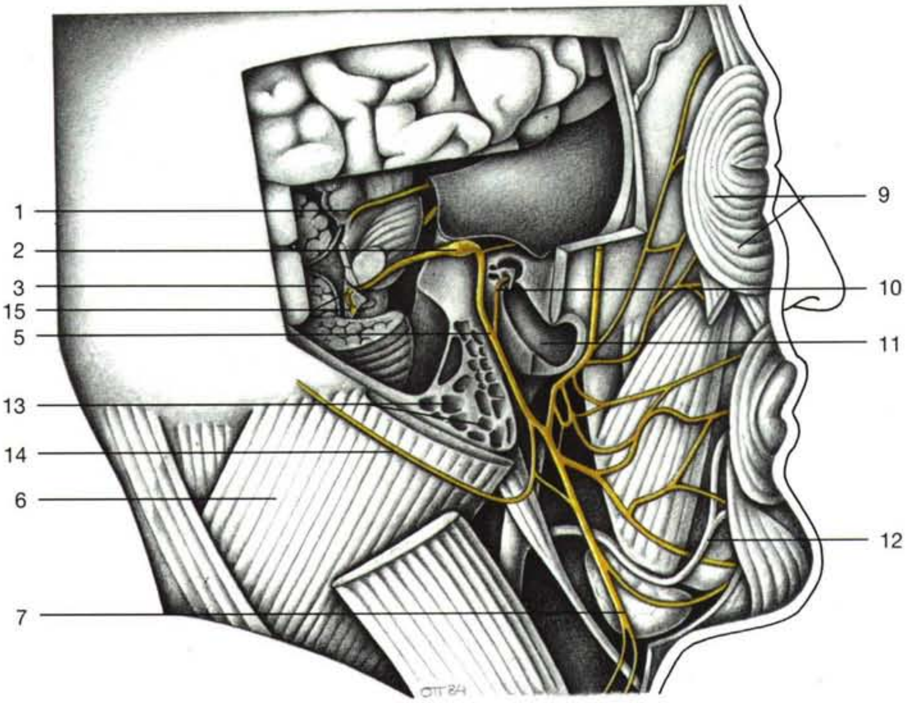

Facial nerve. (Schematic drawing of the dissection above.)

1 Trochlear nerve
2 Facial nerve with geniculate ganglion
3 Cerebellum (right hemisphere)
4 Occipital belly of occipitofrontalis and greater occipital nerve
5 Facial nerve at stylomastoid foramen
6 Splenius capitis muscle
7 Cervical branch of facial nerve
8 Sternocleidomastoid muscle and retromandibular vein
9 Orbicularis oculi muscle
10 Chorda tympani
11 External acoustic meatus
12 Facial artery
13 Mastoid air cells
14 Posterior auricular nerve
15 Nucleus and genu of facial nerve

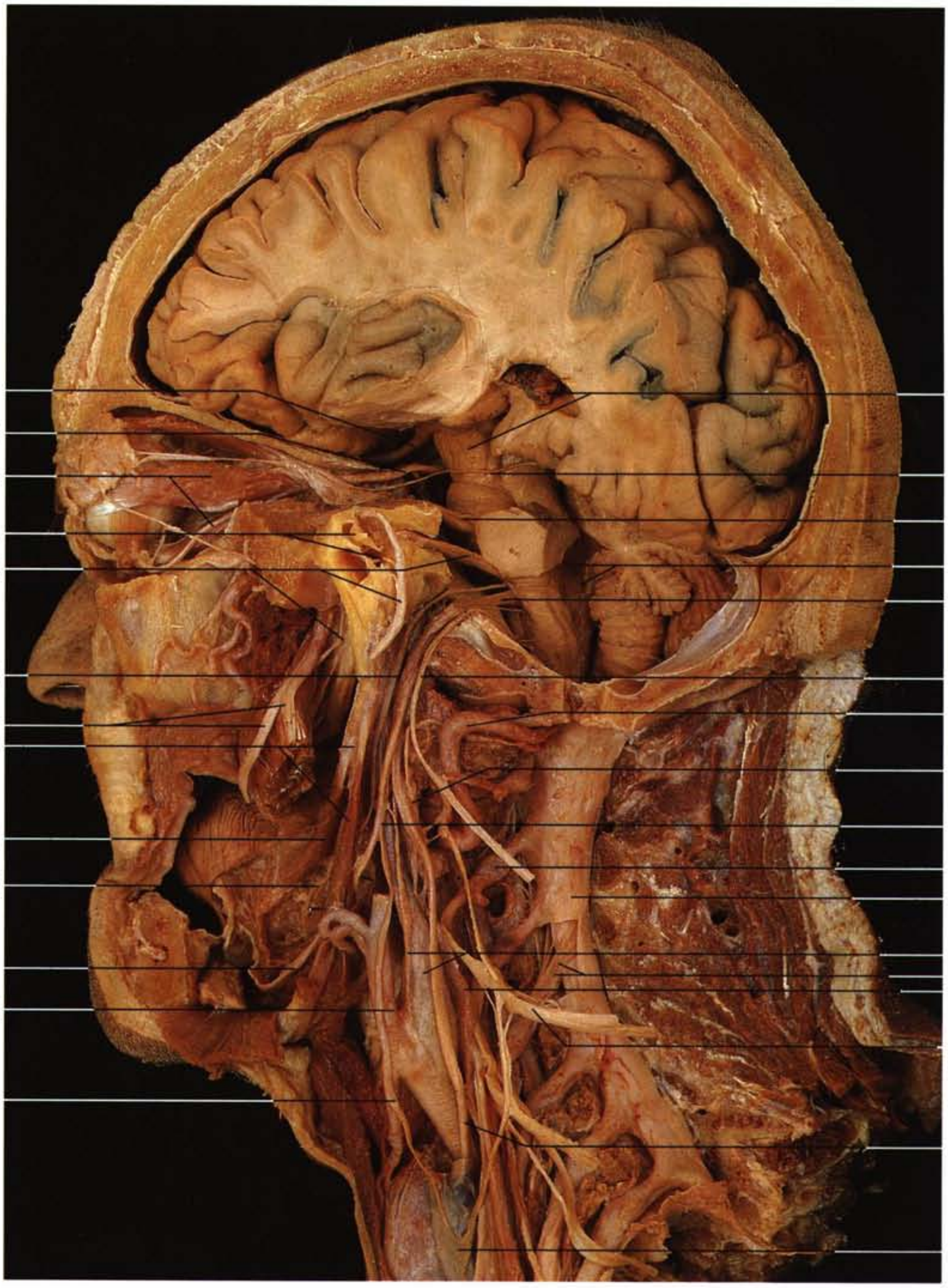

Cranial nerves in connection with the brain stem (oblique-lateral aspect). Lateral portion of the skull, brain, neck and facial structures, lateral wall of orbit and oral cavity have been removed. The tympanic cavity has been opened. The mandible has been divided and the muscles of mastication have been removed.

1 Optic tract
2 **Oculomotor nerve** (n. III)
3 Lateral rectus muscle and inferior branch of oculomotor nerve
4 Malleus and chorda tympani
5 Chorda tympani, **facial nerve** (n. VII), and **vestibulocochlear nerve** (n. VIII)
6 **Glossopharyngeal nerve** (n. XI)
7 Lingual nerve and inferior alveolar nerve
8 Styloid process and stylohyoid muscle
9 Styloglossus muscle
10 Lingual branches of glossopharyngeal nerve
11 Lingual branch of hypoglossal nerve
12 External carotid artery
13 Superior root of ansa cervicalis (branch of hypoglossal nerve)
14 Lateral ventricle with choroid plexus and cerebral peduncle
15 **Trochlear nerve** (n. IV)
16 **Trigeminal nerve** (n. V)
17 Fourth ventricle and rhomboid fossa
18 **Vagus nerve** (n. X)
19 **Accessory nerve** (n. XI)
20 Vertebral artery
21 Superior cervical ganglion
22 **Hypoglossal nerve** (n. XII)
23 Spinal ganglion with dural sheath
24 Dura mater of spinal cord
25 Internal carotid artery and carotid sinus branch of glossopharyngeal nerve
26 Dorsal roots of spinal nerve
27 Sympathetic trunk
28 Branch of cervical plexus (ventral primary ramus of third cervical spinal nerve)
29 Ansa cervicalis

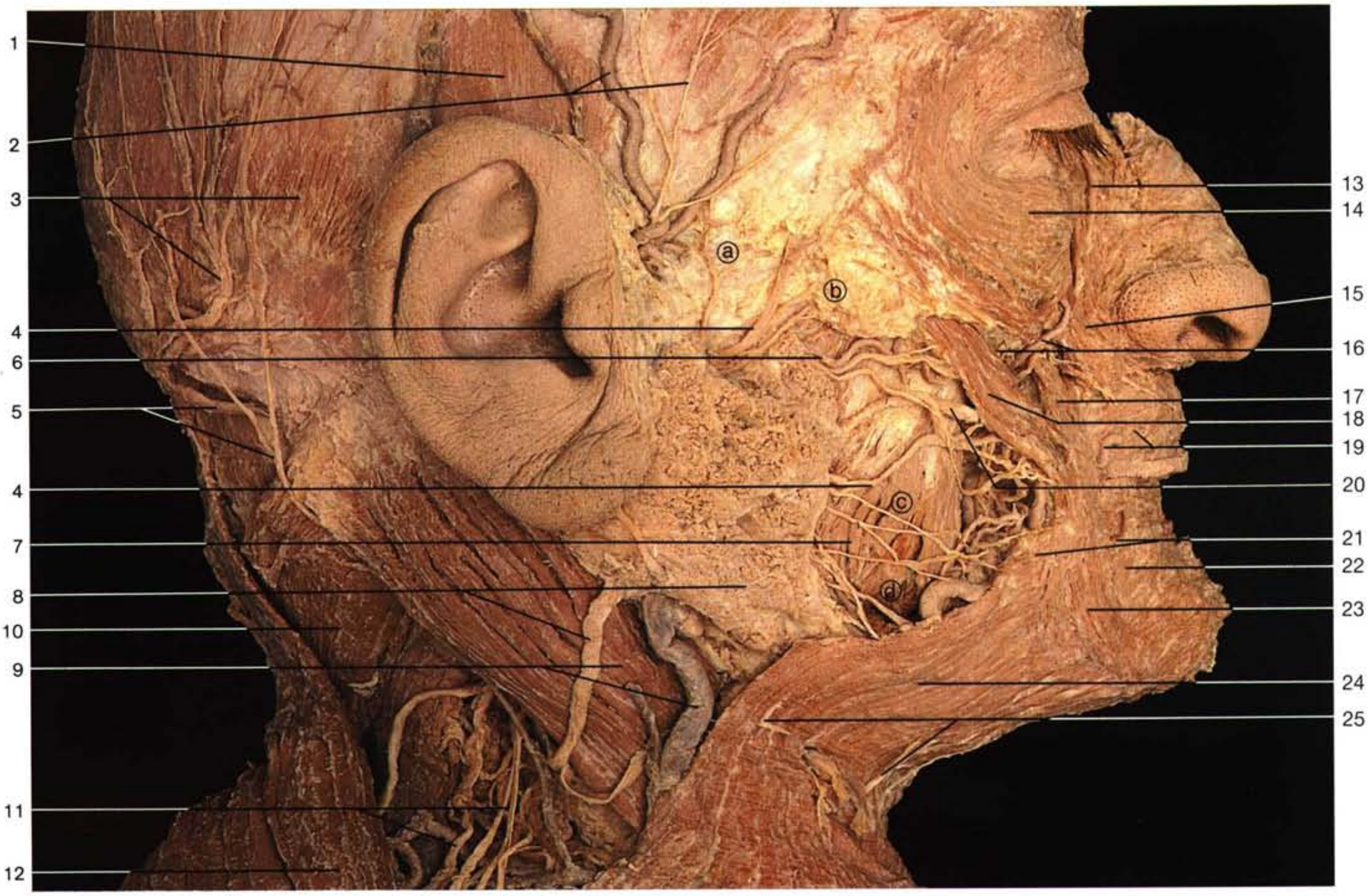

Lateral superficial aspect of the face. Peripheral distribution of facial nerve. a–d = branches of facial nerve: a = temporal branch; b = zygomatic branches; c = buccal branches; d = marginal mandibular branch.

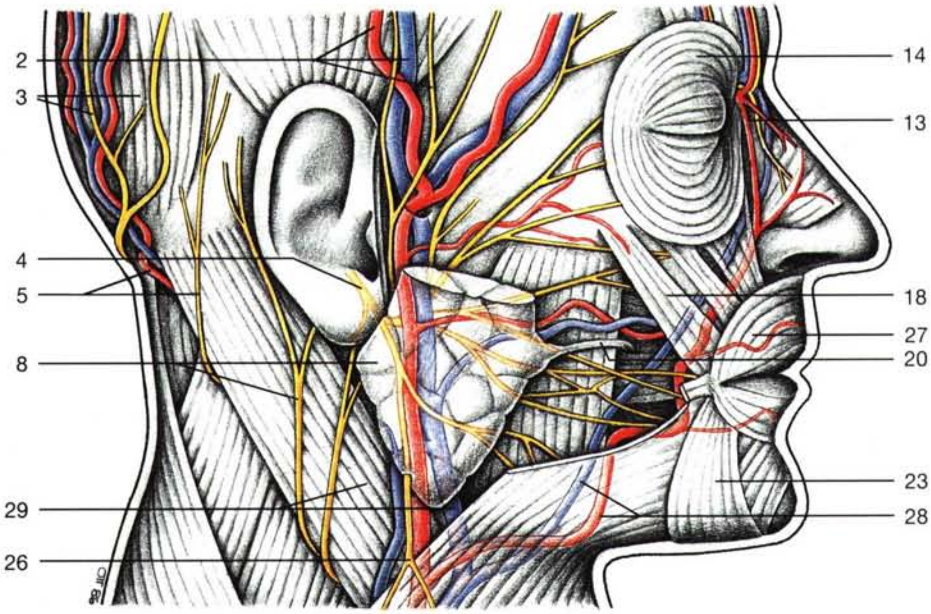

Superficial region of the face. Note the facial plexus within the parotid gland. (Semischematic drawing.)

1 Temporoparietalis muscle
2 Parietal branch of **superficial temporal artery** and **vein** and **auriculotemporal nerve**
3 Occipital belly of occipitofrontalis muscle and greater occipital nerve
4 **Facial nerve**
5 Lesser occipital nerve and **occipital artery**
6 Transverse facial artery
7 Masseter muscle
8 **Parotid gland** and great auricular nerve
9 Sternocleidomastoid muscle and external jugular vein
10 Splenius capitis muscle
11 Branches of cervical plexus
12 Trapezius muscle
13 Angular artery (terminal portion of facial artery)
14 Orbicularis oculi muscle
15 Levator labii superioris alaeque nasi muscle
16 **Facial artery** and zygomaticus minor muscle
17 Levator anguli oris muscle
18 Zygomaticus major muscle
19 Orbicularis oris muscle and superior labial artery
20 Parotid duct
21 Risorius muscle and inferior labial artery

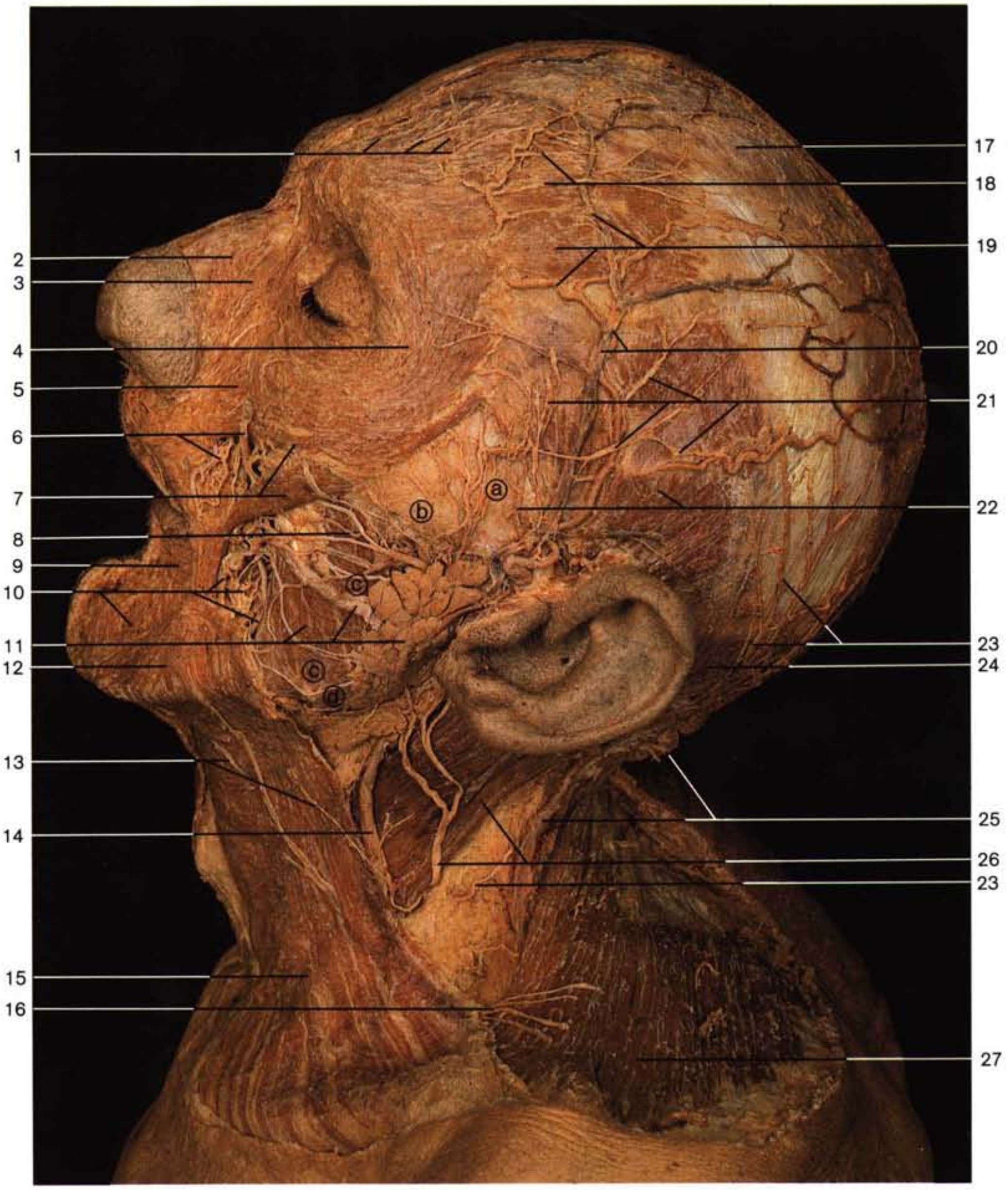

Superficial layer of the head and neck. Parotid fascia removed. Branches of **facial nerve:** a = temporal branch; b = zygomatic branches; c = buccal branches; d = marginal mandibular branch.

1 Medial branch of supraorbital nerve
2 Nasalis muscle
3 Levator labii superioris alaeque nasi muscle
4 **Orbicularis oculi muscle**
5 Levator labii superioris muscle
6 **Facial artery and vein**
7 Zygomaticus minor and major muscle
8 Transverse facial artery
9 **Orbicularis oris muscle**
10 Buccal nerves, depressor labii inferioris muscle, **facial artery** and **vein**
11 **Parotid gland** and duct and masseter muscle
12 Depressor anguli oris muscle
13 Transverse cervical nerve
14 External jugular vein
15 **Platysma muscle**
16 Supraclavicular nerve
17 Galea aponeurotica
18 Lateral branches of the supraorbital nerve
19 Frontal belly of occipitofrontalis muscle and branches of superficial temporal vein and artery
20 Superficial temporal artery and vein
21 **Auriculotemporal nerve**
22 Zygomatico-orbital artery and temporoparietalis muscle
23 Lesser occipital nerve
24 Occipital belly of occipitofrontalis muscle
25 Occipital vein and occipital artery
26 **Great auricular nerve** and sternocleidomastoid muscle
27 Trapezius muscle
28 Greater petrosal nerve
29 Geniculate ganglion
30 Chorda tympani
31 Posterior auricular nerve
32 Stylomastoid foramen

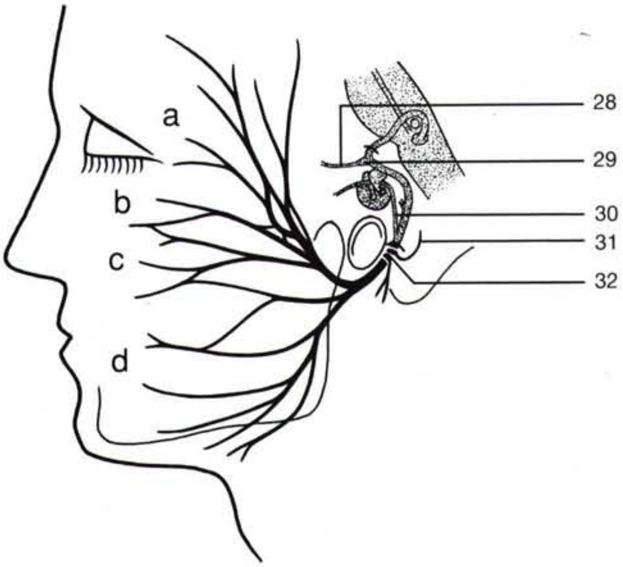

Facial nerve. Main branches. (Schematic diagram.)

◁ **to page 76**
22 Depressor labii inferioris muscle
23 Depressor anguli oris muscle
24 Platysma muscle
25 Terminal branches of transverse cervical nerve
26 Cervical branch of facial nerve
27 Orbicularis oris muscle
28 **Facial artery** and **vein**
29 Sternocleidomastoid muscle and retromandibular vein

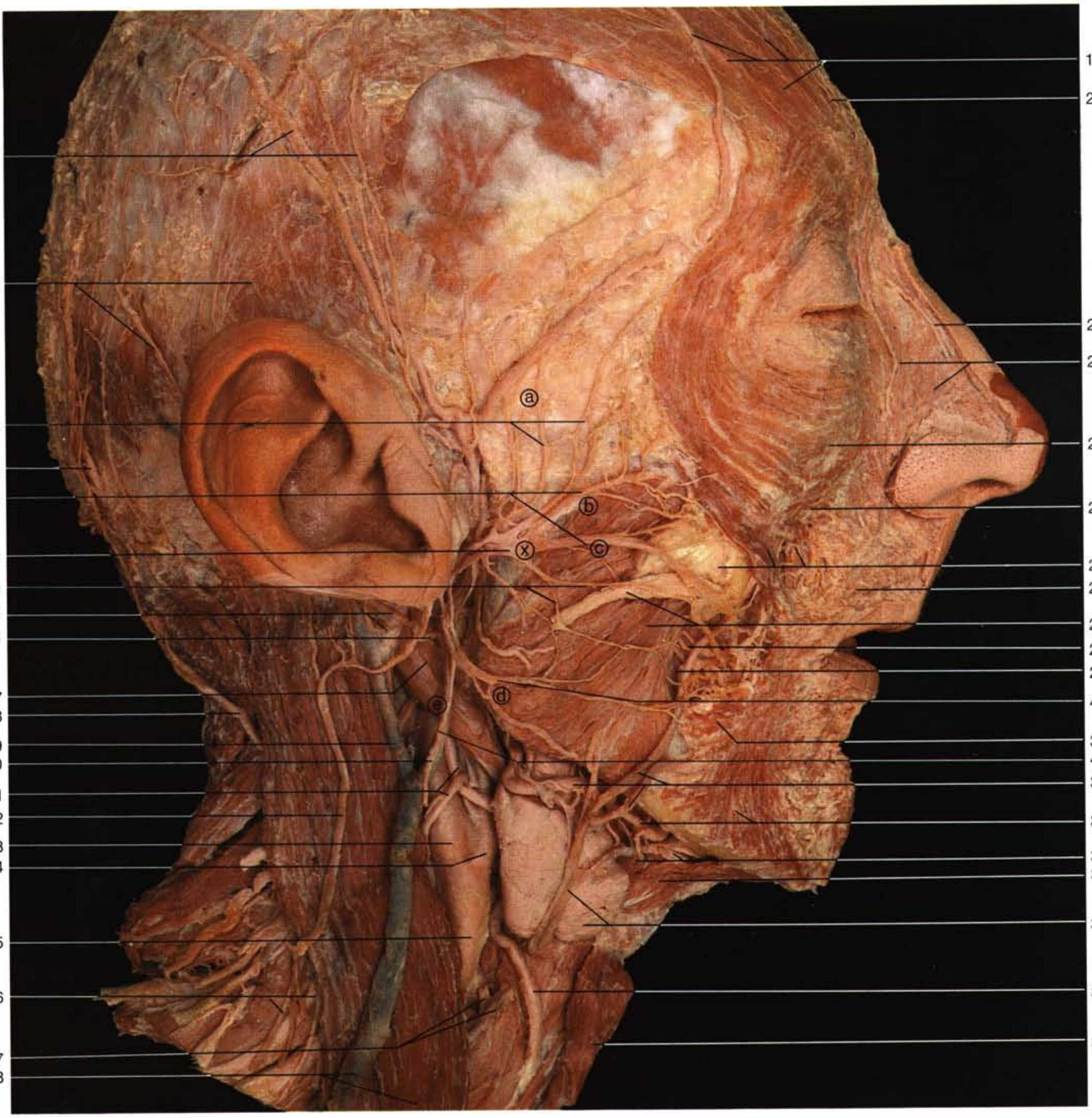

Lateral superficial aspect of the face. The parotid gland has been removed to display the parotid plexus of the facial nerve. a–e = branches of facial nerve: a = temporal branch; b = zygomatic branches; c = buccal branches; d = marginal mandibular branch; e = cervical branch.

1 Superficial temporal artery and auriculotemporal nerve
2 Posterior auricular artery and nerve and temporoparietalis muscle
3 Occipital artery
4 **Facial nerve** (n. VII) (**parotid plexus** [×])
  a Temporal branches
  b Zygomatic branches
  c Buccal branches
  d Marginal mandibular branch
  e Cervical branch
5 Posterior auricular nerve
6 Posterior auricular artery
7 Digastric muscle (posterior belly)
8 **Lesser occipital nerve**
9 Posterior auricular vein
10 Retromandibular vein
11 **Hypoglossal nerve** and sternocleidomastoid artery
12 **Great auricular nerve**
13 Internal carotid artery
14 External carotid artery
15 Common carotid artery
16 Branches of cervical plexus
17 Superior laryngeal artery and vein
18 External jugular vein and sternocleidomastoid muscle
19 Frontal branch of superficial temporal artery, lateral branch of supraorbital nerve, and frontal belly of occipitofrontalis muscle
20 Medial branch of supraorbital nerve
21 Dorsal nasal artery
22 Angular artery and nasalis muscle
23 **Orbicularis oculi muscle**
24 Zygomaticus minor muscle
25 Buccal fat pad, zygomaticus major muscle, and infraorbital nerve
26 **Orbicularis oris muscle**
27 **Parotid duct** and **masseter muscle**
28 Buccal artery and nerve
29 Buccinator muscle
30 Risorius muscle
31 **Facial artery** and **vein**
32 Submental artery and depressor anguli oris muscle
33 Mylohyoid nerve and mylohyoid muscle
34 Digastric muscle (anterior belly)
35 Facial vein and submandibular gland
36 **Superior thyroid artery**
37 Sternohyoid muscle

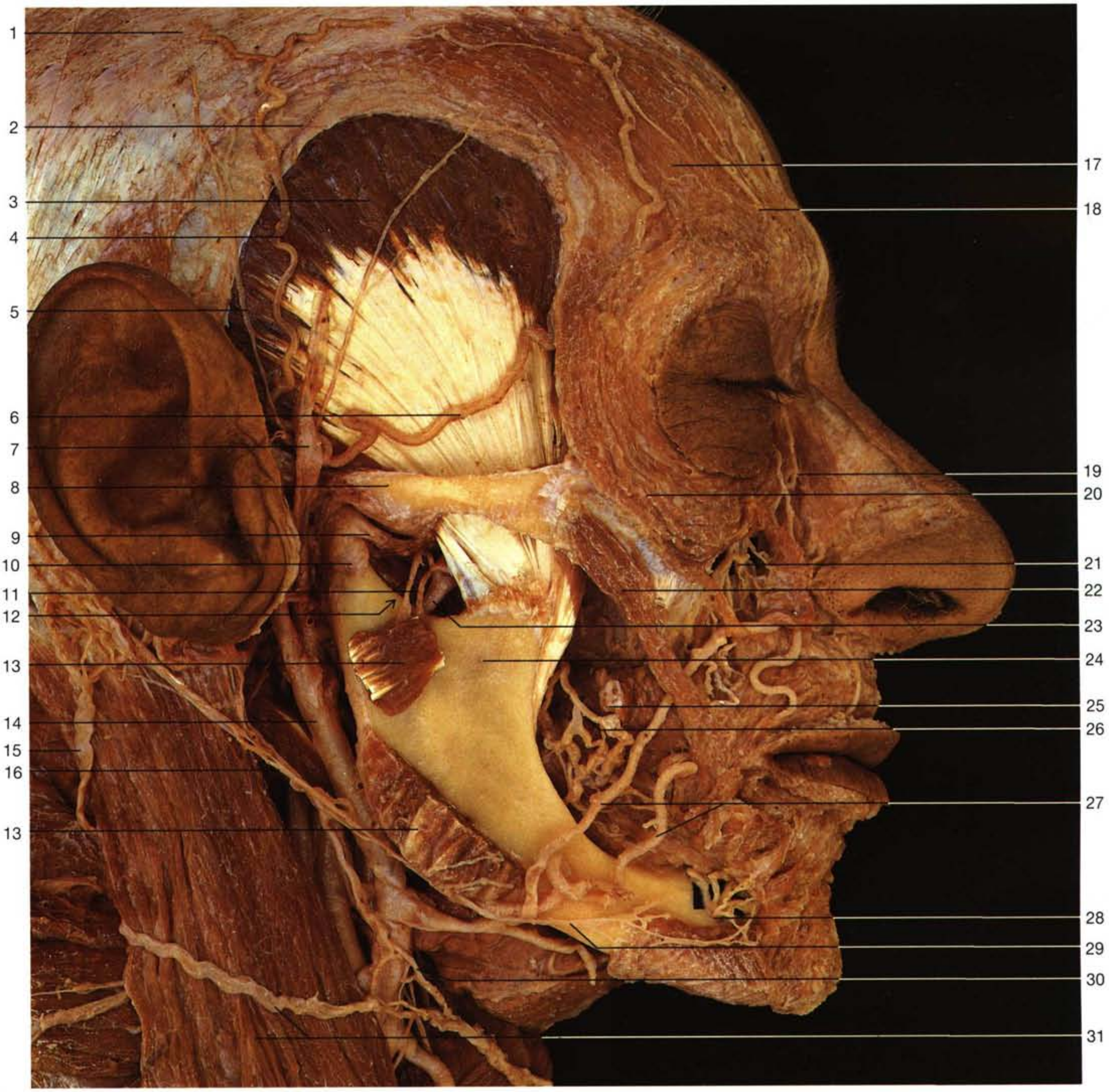

Lateral superficial aspect of the face. Masseter muscle and temporal fascia have been partly removed to display the masseteric artery and nerve.

1 Galea aponeurotica
2 Temporal fascia
3 **Temporalis muscle**
4 Parietal branch of **superficial temporal artery**
5 **Auriculotemporal nerve**
6 Frontal branch of superficial temporal artery
7 Superficial temporal vein
8 **Zygomatic arch**
9 **Articular disc of temporomandibular joint**
10 **Head of mandible**
11 Masseteric artery and nerve
12 Mandibular notch
13 Masseter muscle (divided)
14 **External carotid artery**
15 Great auricular nerve
16 Facial nerve (reflected)
17 Frontal belly of occipitofrontalis muscle
18 Medial branch of supraorbital nerve
19 **Angular artery**
20 Orbicularis oculi muscle
21 **Infraorbital nerve**
22 Zygomaticus major muscle
23 **Maxillar artery**
24 **Coronoid process**
25 Parotid duct (divided)
26 Buccal nerve
27 **Facial artery** and **vein**
28 **Mental nerve**
29 Mandibular branch of facial nerve
30 Cervical branch of facial nerve
31 Transverse cervical nerve (communicating branch with facial nerve) and sternocleidomastoid muscle

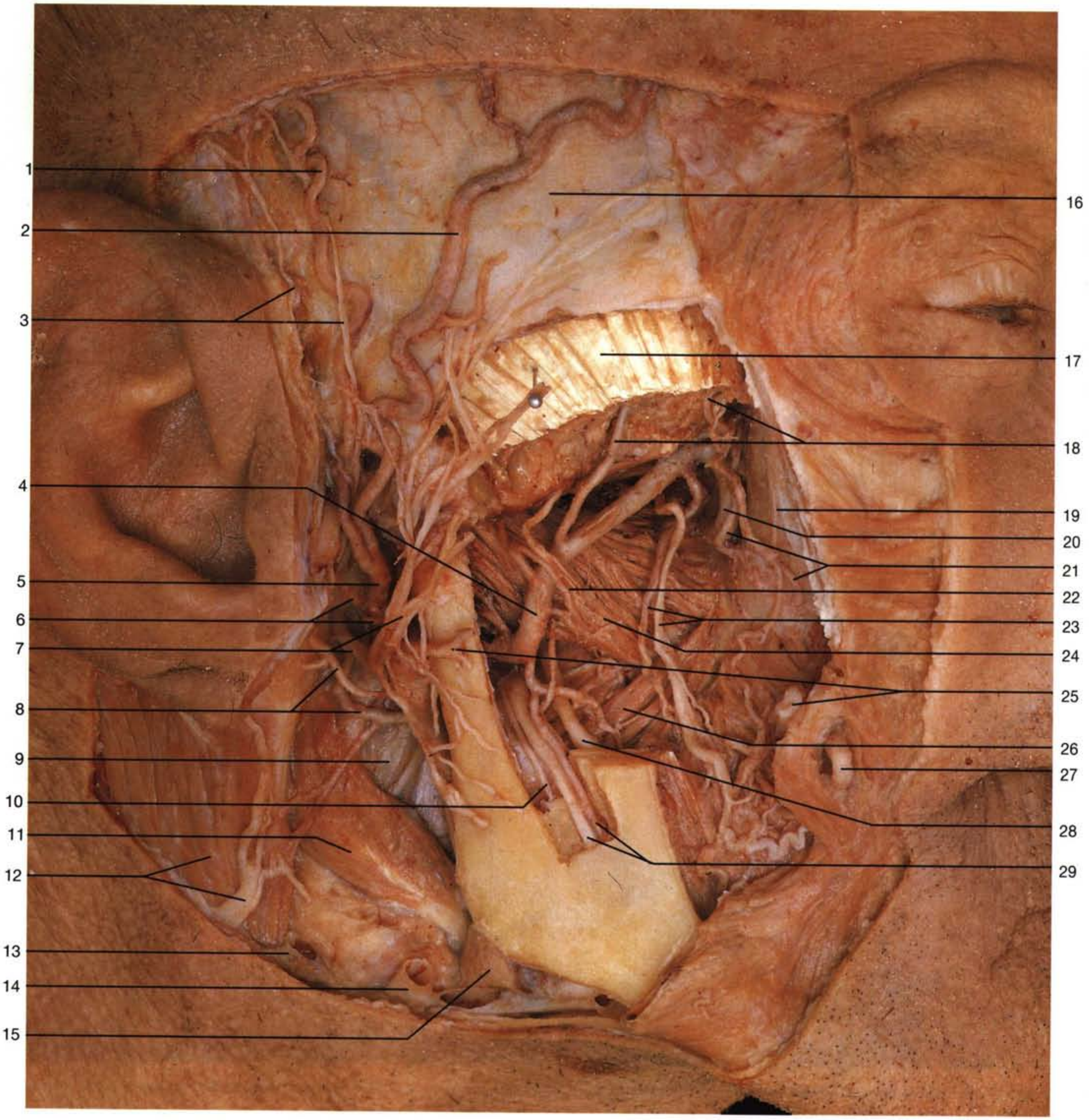

Deep dissection of facial and retromandibular regions. The coronoid process together with the insertions of temporalis muscle have been removed to display the maxillary artery. The upper part of the mandibular canal has been opened.

1 Parietal branch of the superficial temporal artery
2 Frontal branch of the superficial temporal artery
3 Auriculotemporal nerve
4 **Maxillary artery**
5 **Superficial temporal artery**
6 Communicating branches between facial and auriculotemporal nerves
7 **Facial nerve**
8 Posterior auricular artery and anterior auricular branch of superficial temporal artery
9 Internal jugular vein
10 Mylohyoid nerve
11 Posterior belly of digastric muscle
12 Great auricular nerve and sternocleidomastoid
13 External jugular vein
14 Retromandibular vein
15 Submandibular gland
16 Temporal fascia
17 **Temporalis tendon**
18 Deep temporal arteries
19 Posterior superior alveolar nerve
20 Sphenopalatine artery
21 Posterior superior alveolar arteries
22 Masseteric artery and nerve
23 **Buccal nerve and artery**
24 **Lateral pterygoid**
25 Transverse facial artery and parotid duct (divided)
26 Medial pterygoid
27 **Facial artery**
28 **Lingual nerve**
29 **Inferior alveolar artery and nerve** (mandibular canal opened)

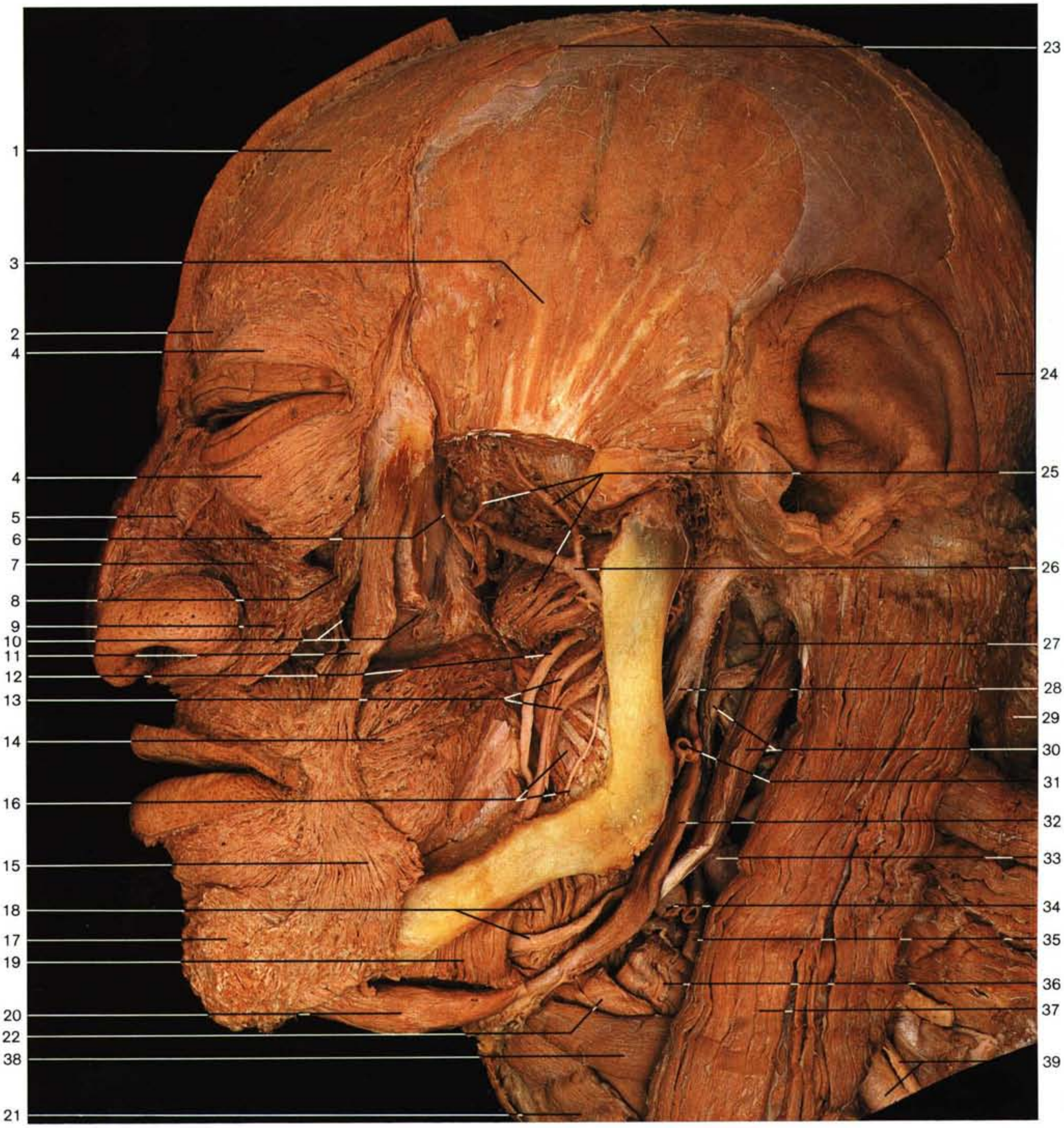

Retromandibular region with maxillary artery and branches of trigeminal nerve (n. V).

1 Frontal belly of occipitofrontalis muscle
2 Depressor supercilii muscle
3 Temporalis muscle
4 **Orbicularis oculi muscle**
5 Nasalis muscle
6 Infraorbital artery
7 Levator labii superioris alaeque nasi muscle
8 Zygomaticus minor muscle
9 Levator labii superioris muscle
10 Infraorbital artery and nerve and posterior superior alveolar artery
11 **Zygomaticus major muscle**
12 **Lingual nerve**
13 **Inferior alveolar artery and nerve**
14 Buccinator muscle
15 Depressor anguli oris muscle
16 **Medial pterygoid muscle** and **mylohyoid nerve**
17 Depressor labii inferioris muscle
18 **Hypoglossal nerve** and **hyoglossus muscle**
19 Mylohyoid muscle
20 Anterior belly of diagastric muscle
21 Sternohyoid muscle
22 Thyrohyoid muscle
23 Galea aponeurotica
24 Occipital belly of occipitofrontalis muscle
25 **Lateral pterygoid muscle** and deep temporal artery
26 **Maxillary artery**
27 Internal jugular vein
28 Styloglossus muscle
29 Splenius capitis muscle
30 Posterior belly of diagastric muscle and occipital artery
31 Superficial temporal artery
32 **Stylohyoid muscle**
33 External carotid artery
34 Retromandibular vein
35 Superior thyroid artery
36 Inferior constrictor muscle of the pharynx
37 Sternocleidomastoid muscle
38 Omohyoid muscle
39 Common carotid artery and vagus nerve

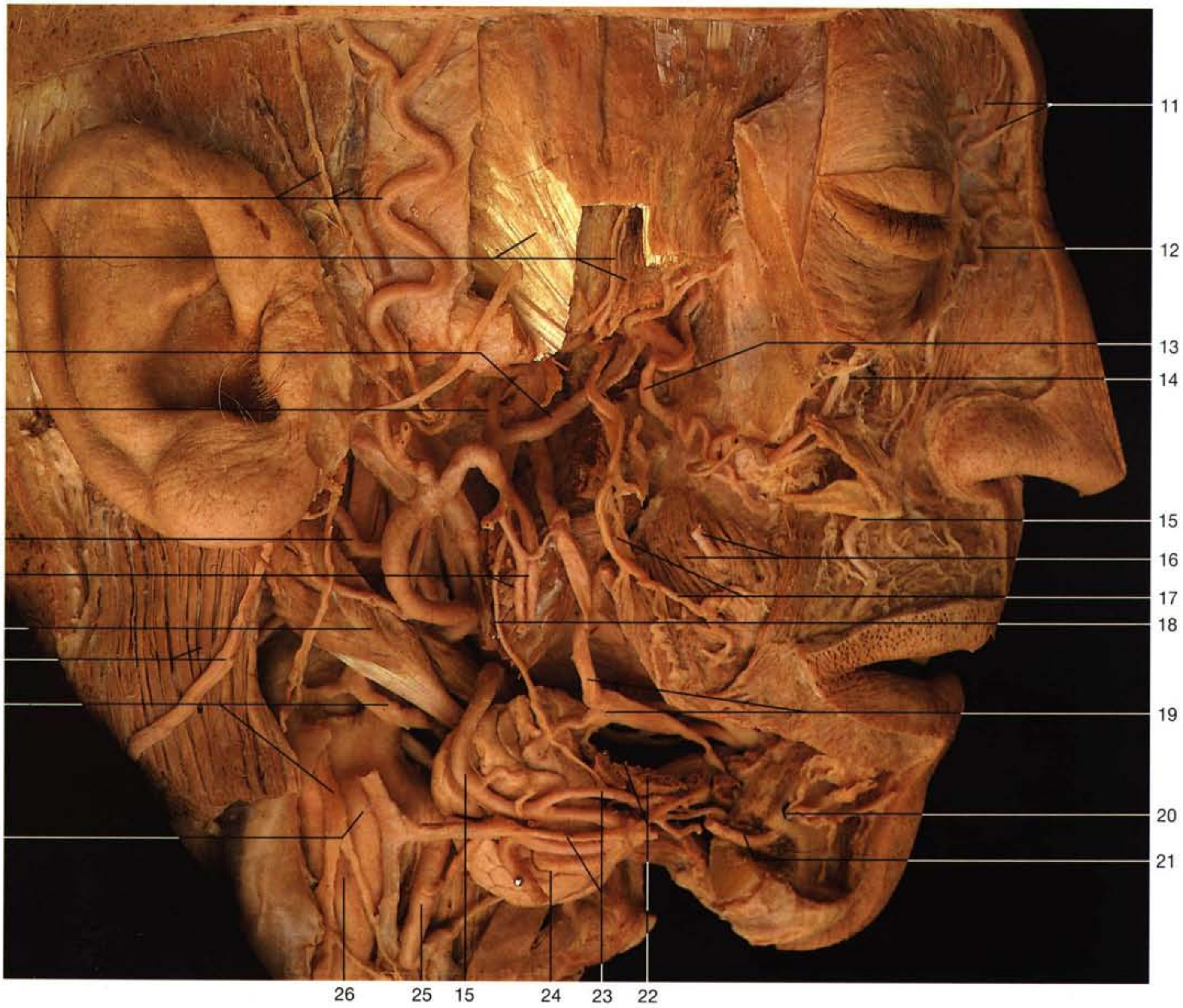

Dissection of deep facial and retromandibular regions after removal of mandible. Pterygoid muscles removed, temporalis muscle fenestrated.

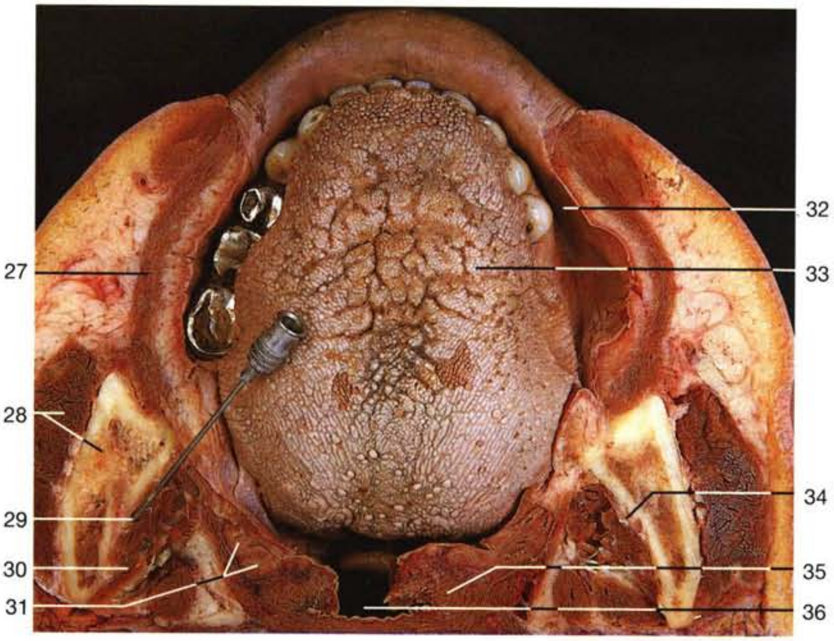

Transverse section through oral cavity and pharynx. The location of inferior alveolar nerve and artery is indicated by a needle.

1 Superficial temporal artery and vein and auriculotemporal nerve
2 Temporalis tendon, deep temporal nerves and artery
3 **Maxillary artery**
4 Middle meningeal artery
5 Occipital artery
6 Inferior alveolar artery and nerve
7 Posterior belly of digastric muscle
8 Great auricular nerve and sternocleidomastoid muscle
9 Hypoglossal nerve and superior root of ansa cervicalis
10 **External carotid artery**
11 Supratrochlear nerve and medial branch of supraorbital artery
12 Angular artery
13 Posterior superior alveolar artery
14 Infraorbital nerve
15 **Facial artery**
16 Parotid duct (divided) and buccinator muscle
17 Buccal artery and nerve

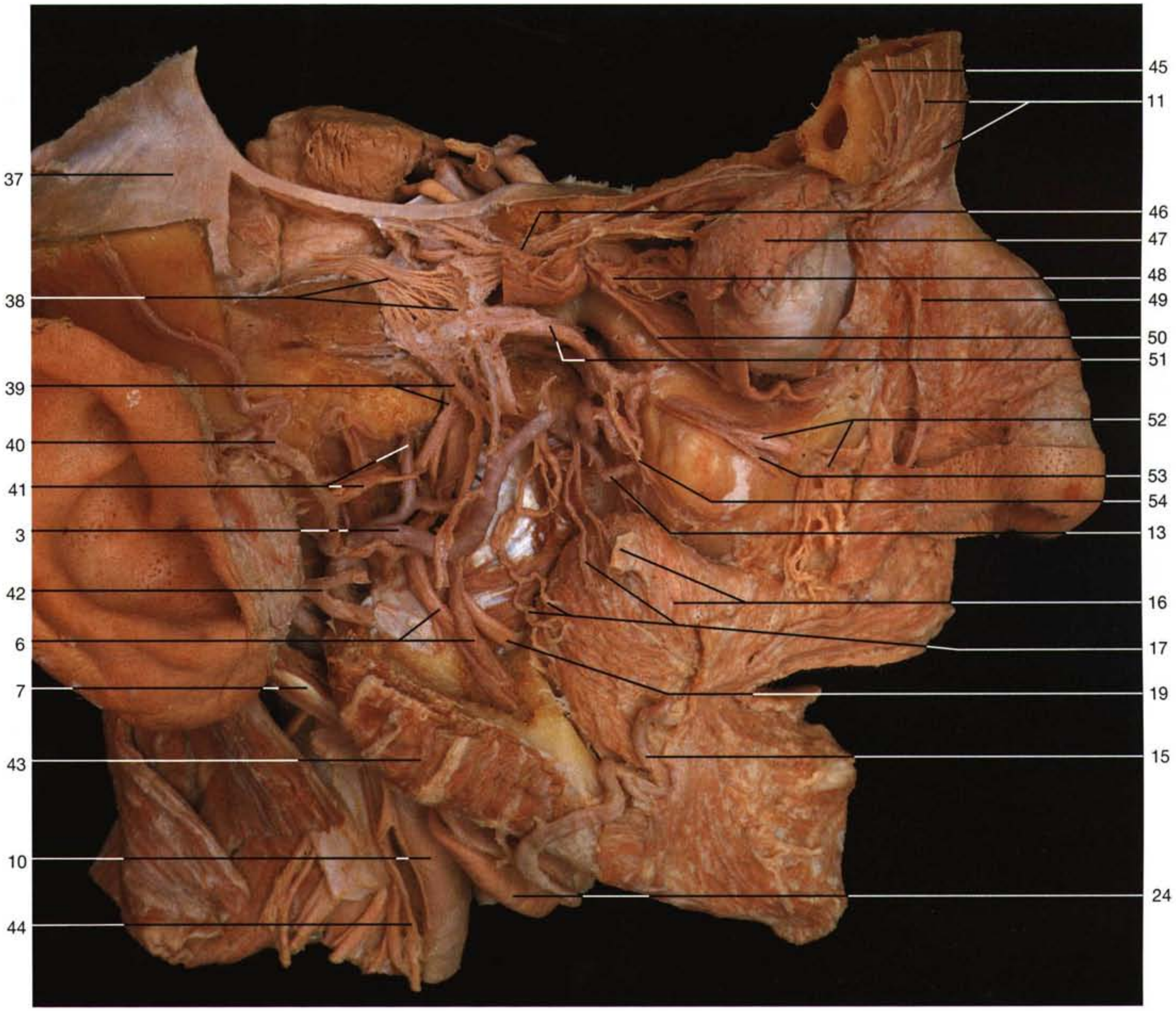

Para- and retropharyngeal regions. The mandible and the lateral wall of the orbit have been removed. The main branches of the trigeminal nerve and its ganglion are displayed.

18	Mylohyoid nerve
19	**Lingual nerve** and **submandibular ganglion**
20	Mental nerve and mental foramen
21	Inferior alveolar nerve
22	Mylohyoid muscle (divided) and hypoglossal nerve
23	Submental artery and vein
24	Submandibular gland
25	Superior thyroid artery
26	Common carotid artery
27	Buccinator muscle
28	Masseter muscle and mandible
29	Entrance of mandibular canal
30	Medial pterygoid muscle
31	Palatine tonsil
32	Oral vestibule
33	Tongue
34	Inferior alveolar nerve, artery and vein
35	Paryngeal constrictor muscle
36	Pharynx
37	Tentorium of cerebellum
38	**Trigeminal nerve** and **ganglion**
39	**Mandibular nerve**
40	Superficial temporal artery
41	Auriculotemporal nerve and middle meningeal artery
42	Facial nerve (divided)
43	Masseter muscle
44	Superior root of ansa cervicalis
45	Lateral branch of supraorbital nerve
46	**Ophthalmic nerve**
47	Lacrimal gland
48	Ciliary ganglion and short ciliary nerves
49	Angular artery
50	Inferior branch of oculomotor nerve
51	**Maxillary nerve**
52	Infraorbital nerve
53	Anterior superior alveolar nerve
54	Posterior superior alveolar nerve

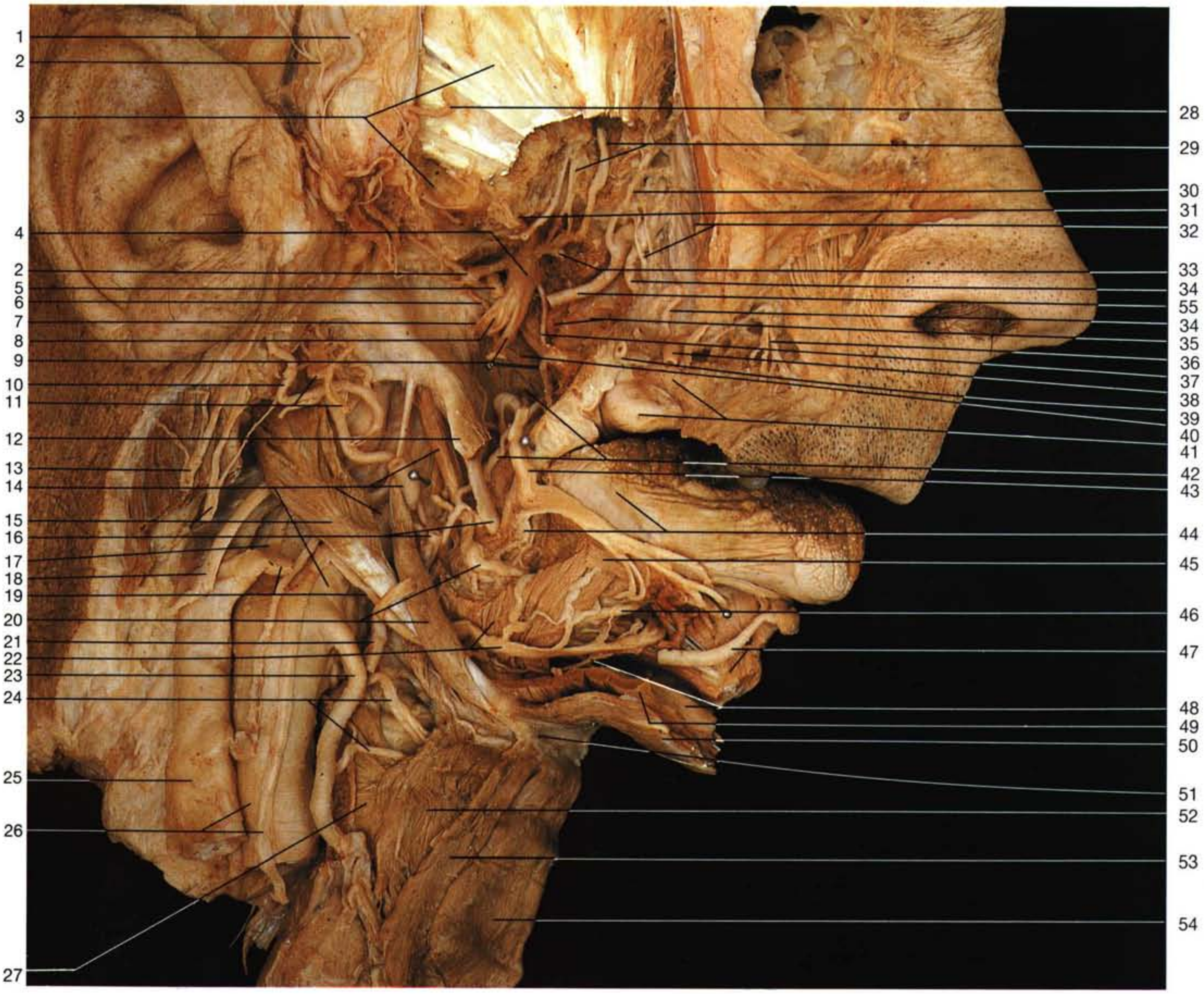

Retropharyngeal and sublingual regions. The mandible has been completely removed.

1 Parietal branch of superficial temporal artery
2 Auriculotemporal nerve
3 Temporalis muscle tendon and zygomatic arch (cut)
4 Mandibular nerve (n. V_3)
5 Middle meningeal artery
6 Chorda tympani
7 Mylohyoid nerve
8 Inferior alveolar nerve
9 **Lingual nerve**
10 Posterior auricular artery
11 Facial nerve
12 Styloglossus muscle (cut)
13 Great auricular nerve
14 Maxillary artery, external carotid artery, and stylopharyngeus muscle
15 Posterior belly of digastric muscle
16 Styloid process and facial artery
17 Vagus nerve (n. X)
18 **Accessory nerve** (n. XI)
19 Hypoglossal nerve (n. XII) and retromandibular vein (cut)
20 Stylohyoid muscle and glossopharyngeal nerve (n. IX)
21 Facial vein
22 Hypoglossal nerve (n. XII) and hyoglossus muscle
23 Superior thyroid artery
24 Superior laryngeal artery and internal laryngeal nerve
25 Internal jugular vein
26 Common carotid artery and superior root of ansa cervicalis
27 Inferior pharyngeal constrictor muscle
28 Middle temporal artery
29 Deep temporal nerves and posterior deep temporal artery
30 Anterior deep temporal artery
31 Masseteric nerve
32 Posterior superior alveolar branches of maxillary artery and nerve
33 Pterygoid branches of mandibular nerve
34 Posterior superior alveolar artery
35 Lateral pterygoid plate and medial pterygoid muscle
36 Infraorbital nerve and artery
37 Buccal nerve
38 Facial artery
39 Parotid duct

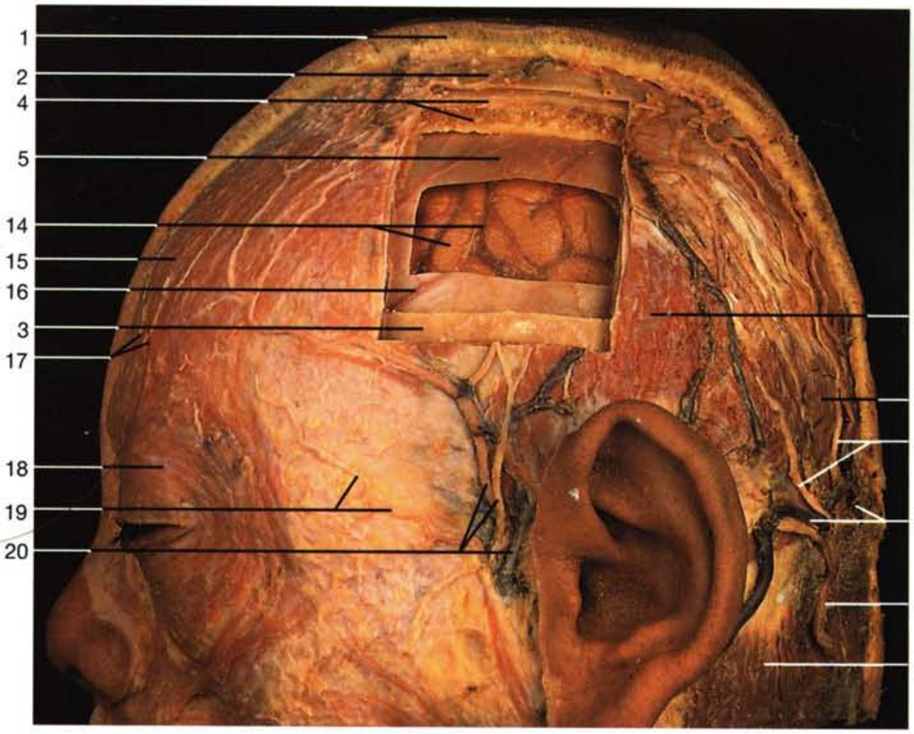

Nerves and blood vessels of the scalp. Scalp and meninges are demonstrated by a series of window-like openings.

1 Skin
2 **Galea aponeurotica**
3 Pericranium (periosteum)
4 Skull with diploe
5 **Dura mater**
6 Subdural space
7 **Arachnoid mater**
8 **Subarachnoid space**
9 Arachnoid granulations
10 Superior sagittal sinus
11 **Pia mater** with cerebral vessels
12 Falx cerebri
13 Cerebral cortex
14 Arachnoid and pia mater with cortical vessels
15 Frontal belly of occipitofrontal muscle
16 Branch of middle meningeal artery
17 Lateral and medial branch of supraorbital nerve
18 Orbicularis oculi muscle
19 Zygomatico-orbital artery
20 **Auriculotemporal nerve, superficial temporal artery** and **vein**
21 Superior auricular muscle
22 Occipital belly of occipitofrontalis muscle
23 Branches of **greater occipital nerve**
24 **Occipital artery** and **vein**
25 Greater occipital nerve
26 Sternocleidomastoid muscle

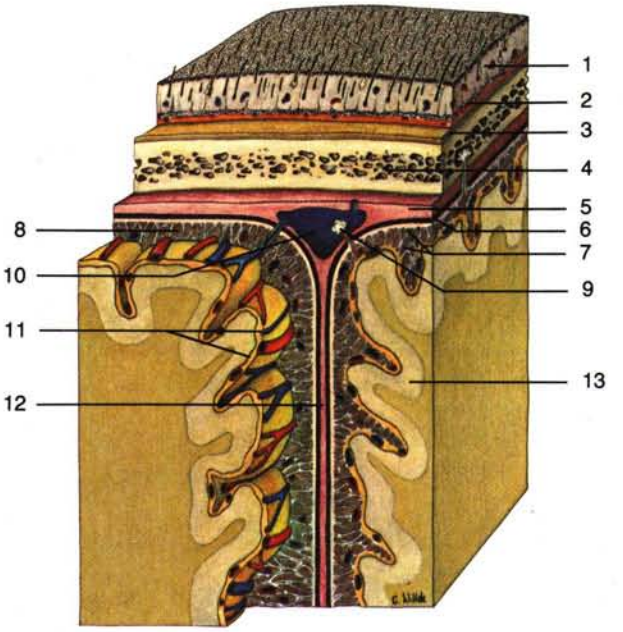

A coronal **section through the vertex of the skull** showing the arrangement of the meninges and vessels of the brain. Together the arachnoid mater and the pia mater form the leptomeninx.

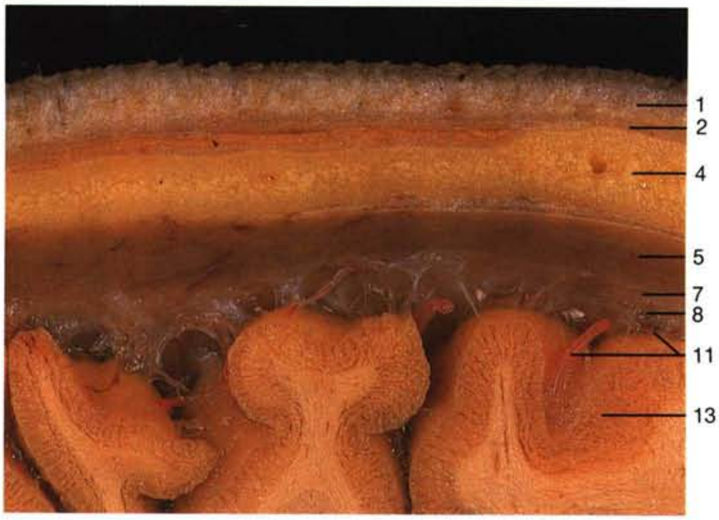

Cross-section of the scalp and the meninges. The subarachnoid space is shown.

40 Levator veli palatini muscle
41 Gingiva and buccinator muscle
42 Ascending pharyngeal artery and superior pharyngeal constrictor muscle (pterygopharyngeal part)
43 Lingual nerve
44 **Submandibular ganglion** and tongue
45 Palatoglossus muscle
46 Deep lingual artery
47 Submandibular duct and genioglossus muscle
48 Geniohyoid nerve and muscle
49 Mylohyoid muscle
50 Anterior belly of digastric muscle
51 Hyoid bone
52 Thyrohyoid muscle
53 Omohyoid muscle (superior belly)
54 Sternohyoid muscle
55 **Maxillary artery**

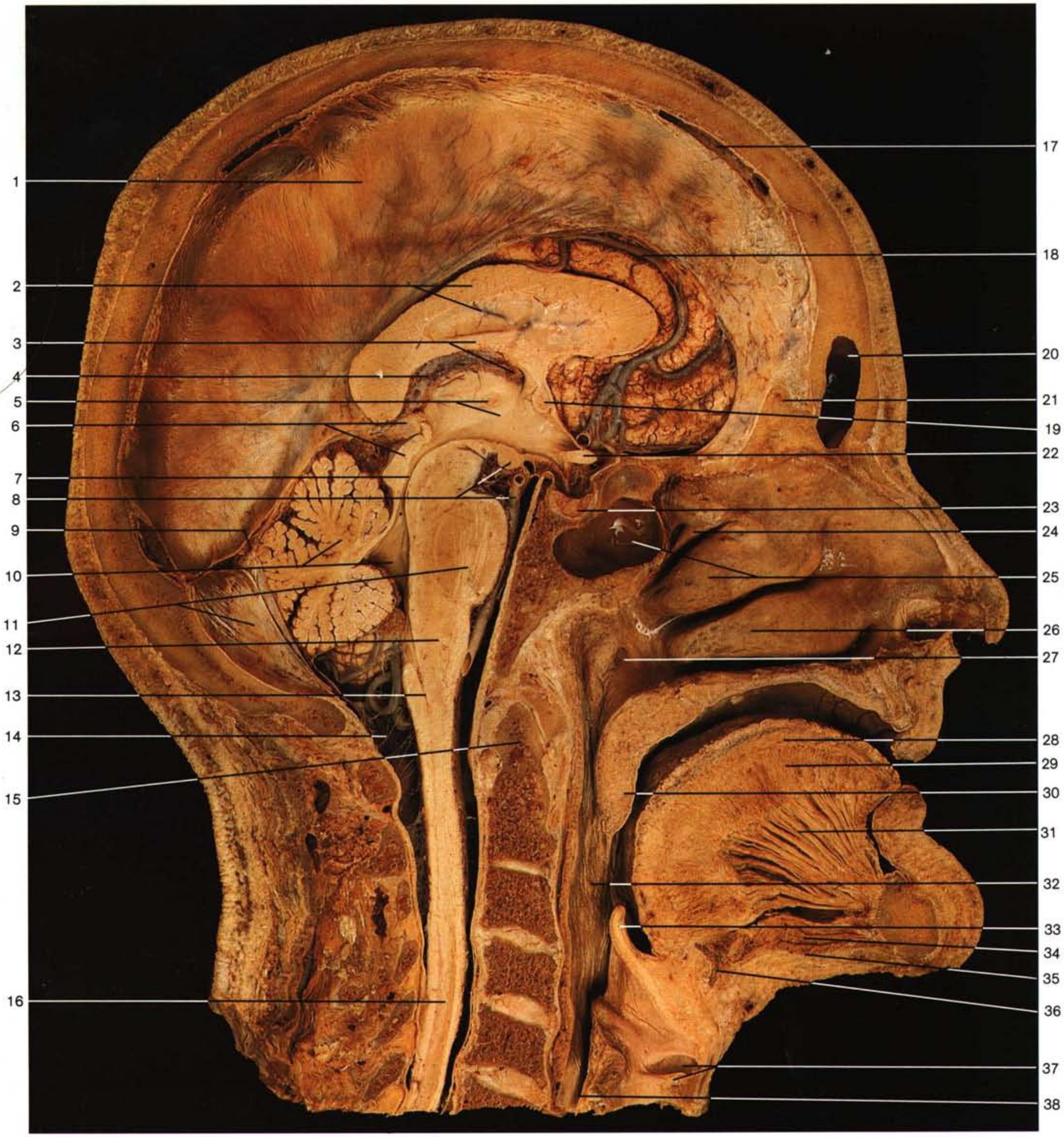

Median sagittal section through the head and neck.

1 Falx cerebri
2 **Corpus callosum** and **septum pellucidum**
3 Interventricular foramen and fornix
4 Choroid plexus of third ventricle and internal cerebral vein
5 **Third ventricle** and interthalamic adhesion
6 **Pineal body** and **colliculi of the midbrain**
7 Cerebral aqueduct
8 Mamillary body and basilar artery
9 Straight sinus
10 **Fourth ventricle** and **cerebellum**
11 Pons and falx cerebelli
12 **Medulla oblongata**
13 Central canal
14 **Cerebellomedullary cistern**
15 Dens of the axis (odontoid process)
16 Spinal cord
17 **Superior sagittal sinus**
18 **Anterior cerebral artery**
19 Anterior commissure
20 **Frontal sinus**
21 Crista galli
22 Optic chiasma
23 Pituitary gland (hypophysis)
24 Superior nasal concha
25 **Middle nasal concha** and **sphenoid sinus**
26 **Inferior nasal concha**
27 **Pharyngeal opening of auditory tube**
28 Superior longitudinal muscle of tongue
29 Vertical muscle of the tongue
30 **Uvula**
31 Genioglossus muscle
32 Pharynx
33 **Epiglottis**
34 Geniohyoid muscle
35 Mylohyoid muscle
36 Hyoid bone
37 Vocal fold and sinus of larynx
38 Esophagus

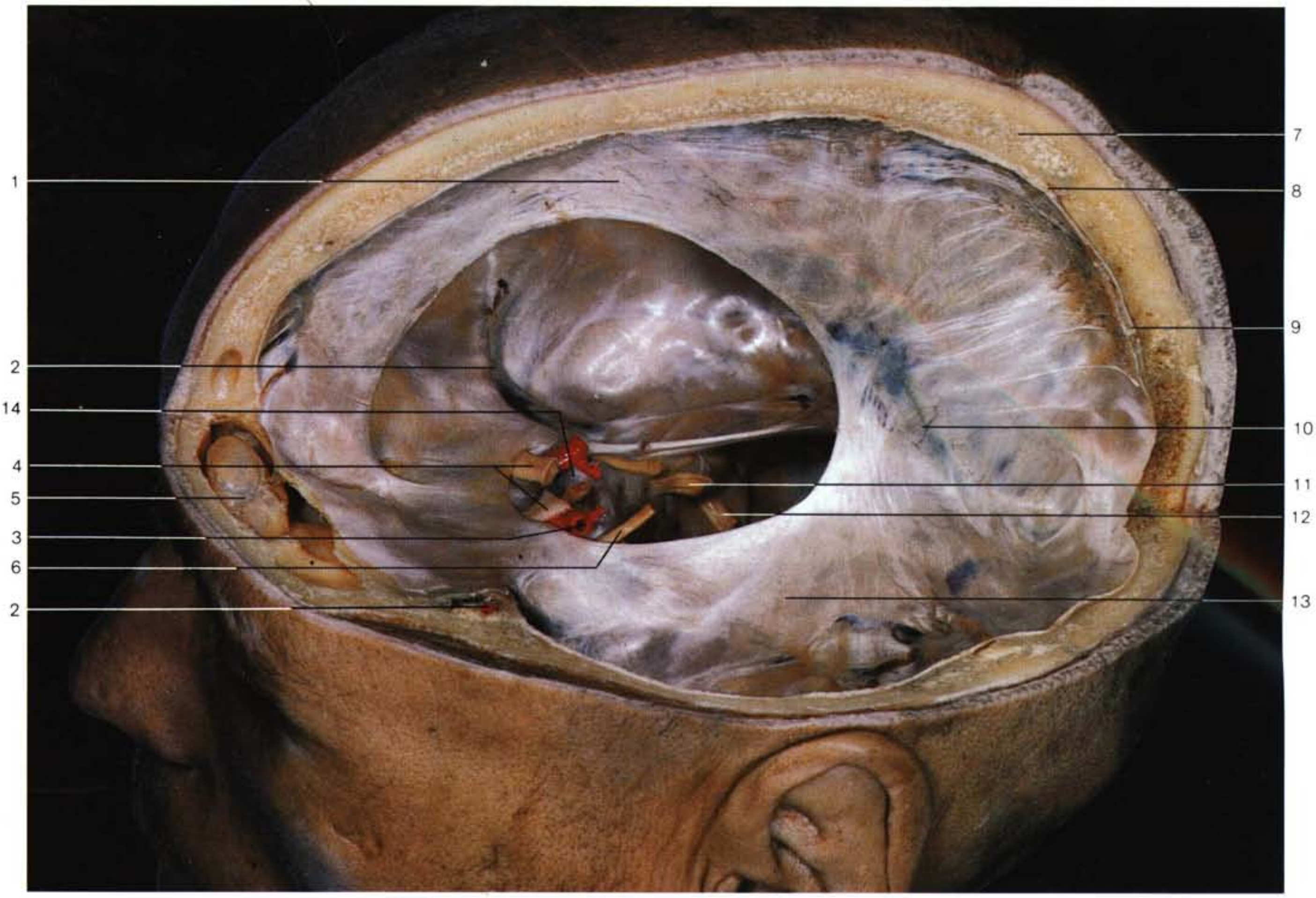

Dura mater and venous sinuses of the dura mater. The brain has been removed (oblique lateral aspect).

1 Falx cerebri
2 Position of middle meningeal artery and vein
3 Internal carotid artery
4 **Optic nerves** (n. II)
5 Frontal sinus
6 **Oculomotor nerve** (n. III)
7 Diploe
8 **Dura mater**
9 Superior sagittal sinus
10 Straight sinus
11 **Trigeminal nerve** (n. V)
12 Facial and vestibulocochlear nerve (n. VII and n. VIII)
13 **Tentorium cerebelli**
14 **Pituitary gland (hypophysis)**
15 Inferior sagittal sinus
16 Sigmoid sinus
17 Confluence of sinuses
18 Inferior petrosal sinus
19 Transverse sinus
20 Superior petrosal sinus
21 **Cavernous** and **intercavernous sinuses**

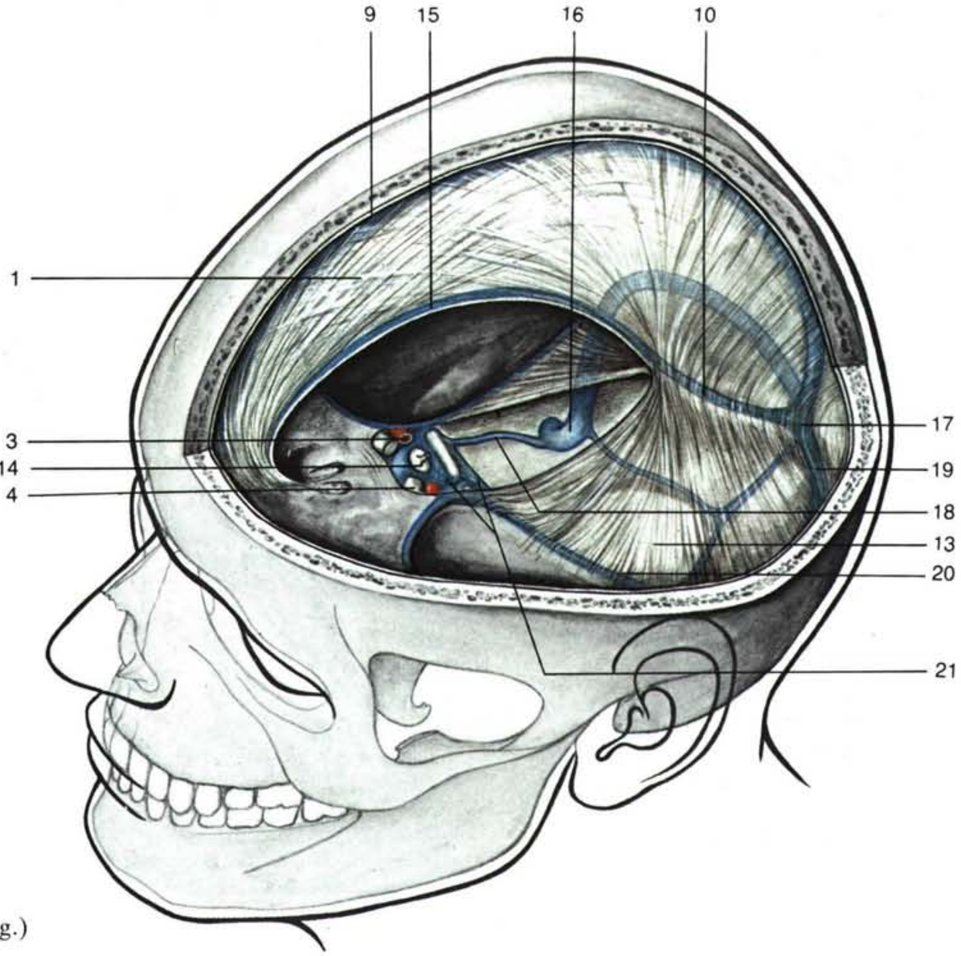

Dura mater and **venous sinuses** (left lateral aspect). (Schematic drawing.)

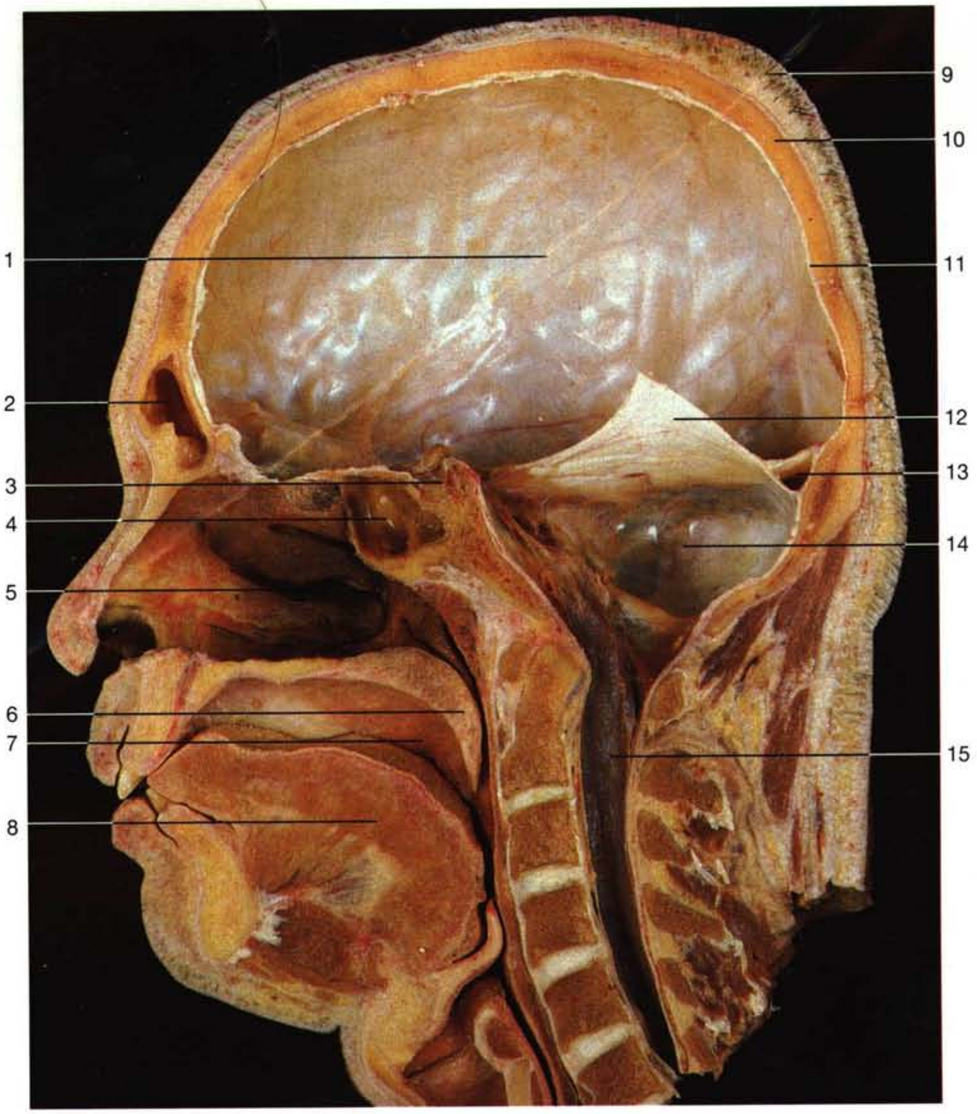

Median section through the head. Demonstration of dura mater covering the cranial cavity. Brain and spinal cord are removed (right half of the head, as seen from medial).

1 **Cranial cavity** with dura mater (the right cerebral hemisphere has been removed)
2 Frontal sinus
3 Hypophysial fossa with pituitary gland
4 Sphenoidal sinus
5 **Nasal cavity**
6 Soft palate (uvula)
7 **Oral cavity**
8 Tongue
9 Skin
10 Calvaria
11 Dura mater
12 Tentorium cerebelli
13 Confluence of sinuses
14 **Infratentorial space** (cerebellum and part of the brain stem have been removed)
15 Vertebral canal
16 Frontal branch of middle meningeal artery and veins
17 Middle meningeal artery
18 **Diploe**
19 Parietal branch of middle meningeal artery and vein
20 Occipital pole of left hemisphere covered with dura mater

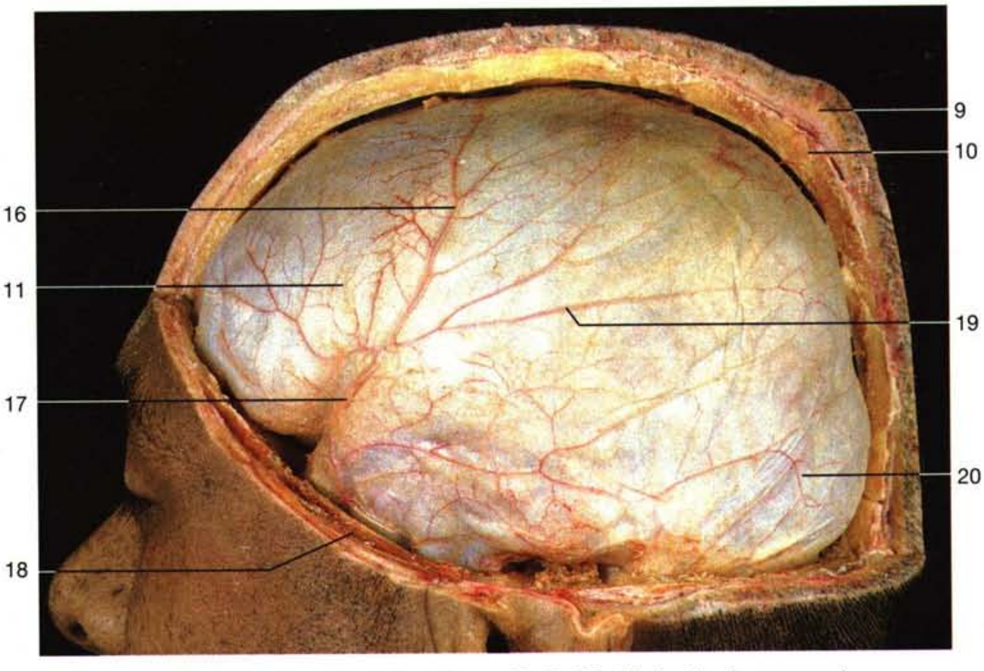

Dissection of dura mater and meningeal vessels. Left half of calvaria removed.

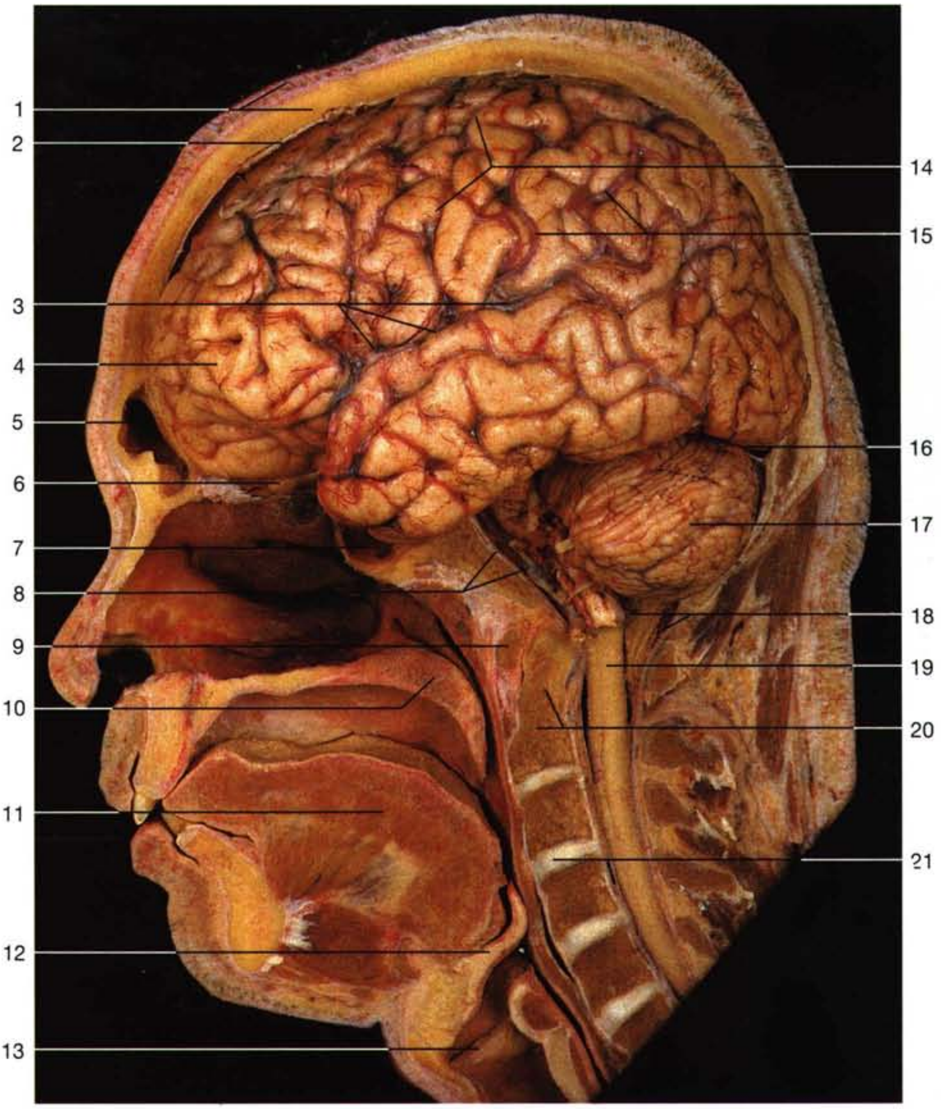

Dissection of the brain with pia mater and arachnoid in situ. The head is cut in half except for the brain, which is shown in its entirety.

1 Calvaria and skin of the scalp
2 Dura mater (divided)
3 Position of lateral sulcus
4 **Frontal lobe** covered by **arachnoid** and **pia mater**
5 Frontal sinus
6 Olfactory bulb
7 **Sphenoidal sinus**
8 Dura mater on clivus and basilar artery
9 Atlas (anterior arch, divided)
10 **Soft palate**
11 Tongue
12 Epiglottis
13 Vocal fold
14 Position of central sulcus
15 Superior cerebral veins
16 Tentorium (divided)
17 **Cerebellum**
18 Cerebellomedullary cistern
19 Position of foramen magnum and **spinal cord**
20 Dens of axis
21 Intervertebral disc

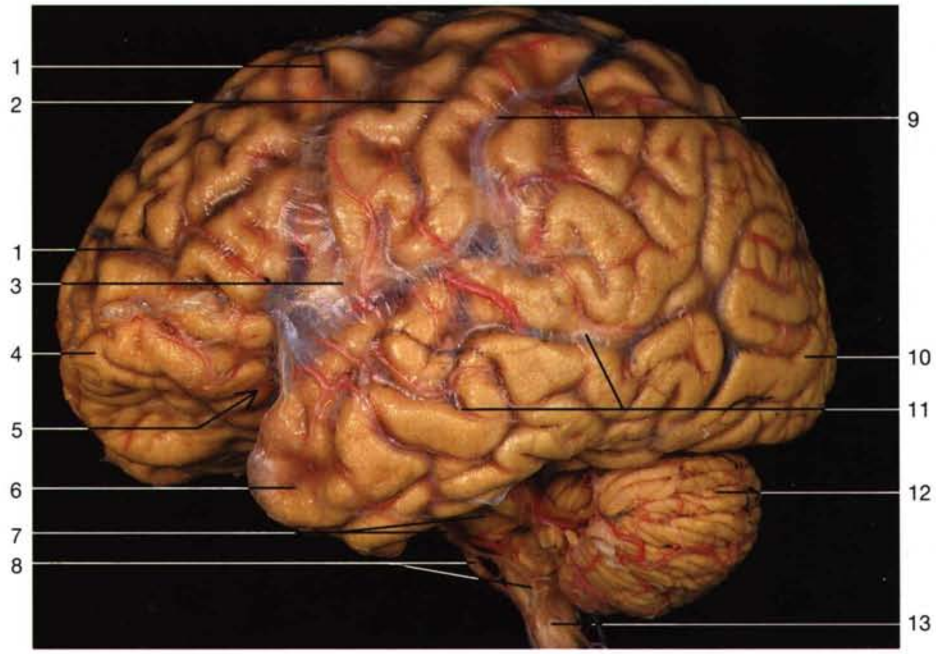

Brain with pia mater and arachnoid. Frontal pole to the left (lateral aspect).

1 Superior cerebral veins
2 Position of central sulcus
3 Position of lateral sulcus and **cistern of lateral cerebral fossa**
4 **Frontal pole**
5 Lateral sulcus (arrow)
6 **Temporal pole**
7 Pons and basilar artery
8 Vertebral arteries
9 Superior anastomotic vein
10 Occipital pole
11 Inferior cerebral veins
12 Hemisphere of **cerebellum**
13 Medulla oblongata

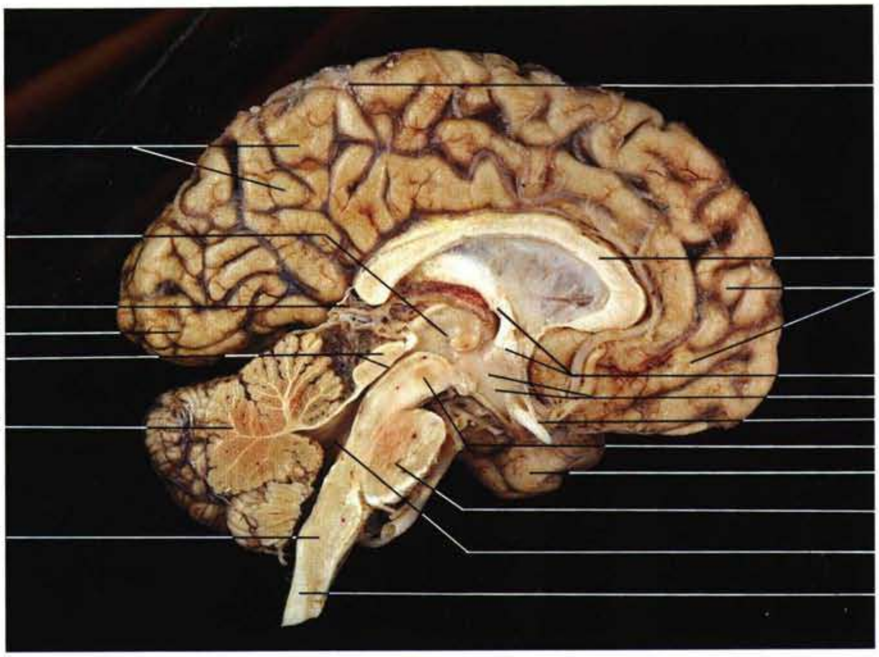
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

Brain and brain stem, median section. Frontal pole to the right.

1 Parietal lobe
2 Thalamus, **third ventricle** and intermediate mass
3 Great cerebral vein
4 **Occipital lobe**
5 Colliculi of the **midbrain** and **cerebral aqueduct**
6 Cerebellum
7 **Medulla oblongata**
8 Central sulcus
9 **Corpus callosum**
10 **Frontal lobe**
11 Fornix and anterior commissure
12 **Hypothalamus**
13 Optic chiasma
14 Midbrain
15 Temporal lobe
16 **Pons**
17 **Fourth ventricle**
18 Spinal cord
19 Inferior concha and nasal cavity
20 Alveolar process of maxilla
21 Tongue
22 Dens of axis
23 Oral part of pharynx
24 Alveolar process of mandible
25 Epiglottis

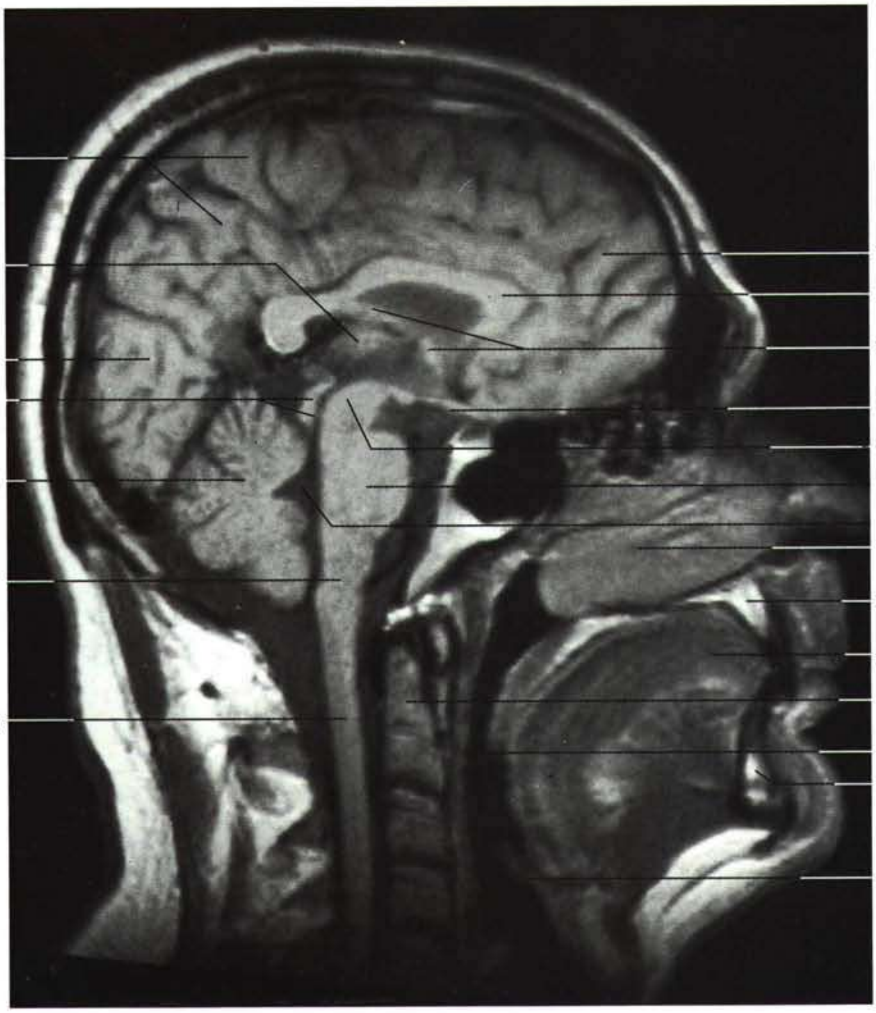
1
2
4
5
6
7
18
10
9
11
13
14
16
17
19
20
21
22
23
24
25

Median section through the head. (MR-Scan, cf. section on opposite page.)

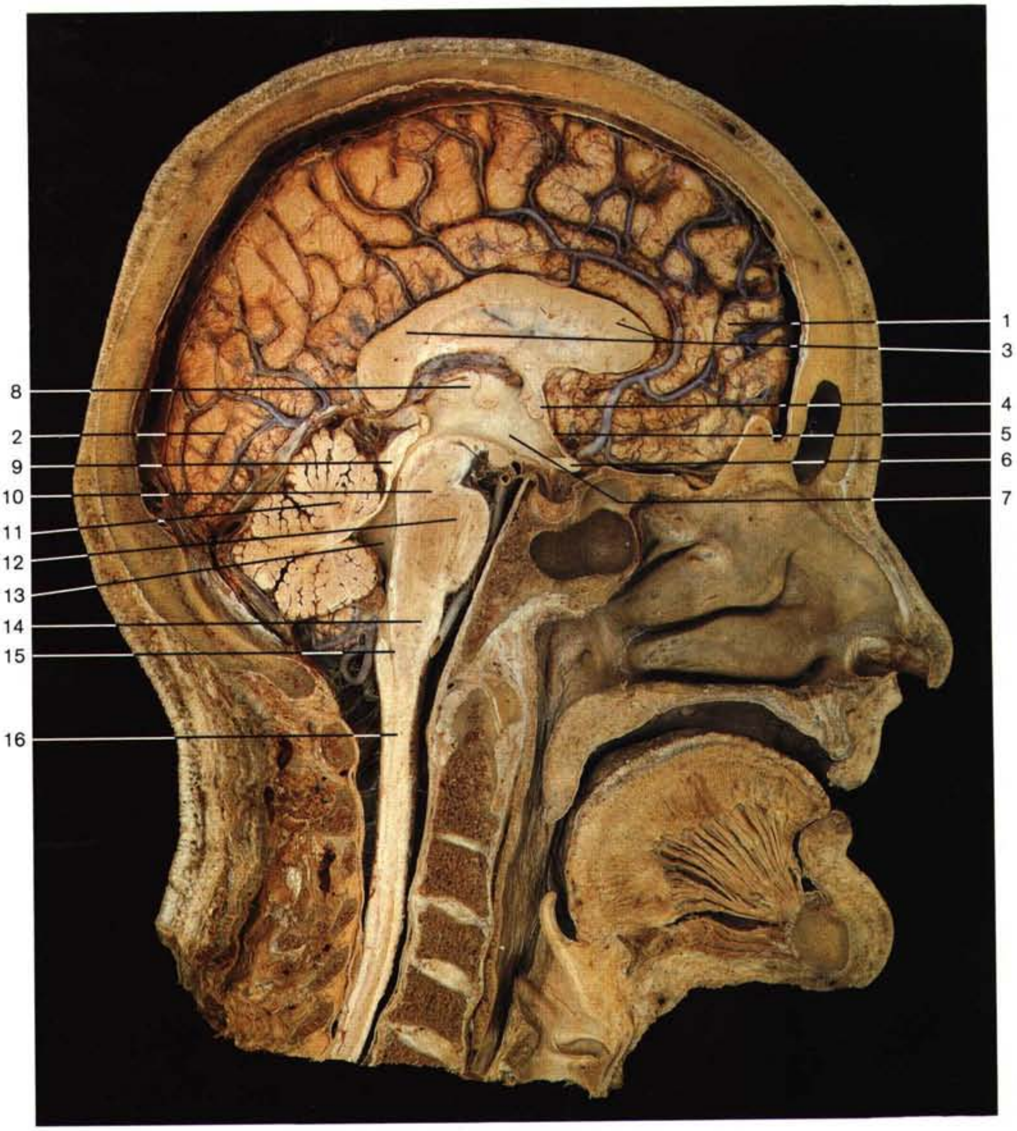

1 Frontal lobe of cerebrum
2 Occipital lobe of cerebrum
3 **Corpus callosum**
4 Anterior commissure
5 **Lamina terminalis**
6 Optic chiasma
7 **Hypothalamus**
8 Thalamus and third ventricle
9 **Colliculi of the midbrain**
10 Midbrain (inferior portion)
11 **Cerebellum**
12 Pons
13 **Fourth ventricle**
14 Medulla oblongata
15 Central canal
16 **Spinal cord**

Median section through the head. Regions of the brain. Falx cerebri removed.

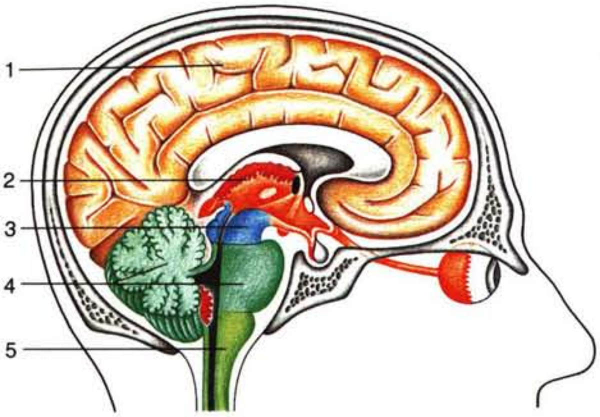

Scheme of brain divisions (cf. table).
Red = choroidal plexus. (Schematic drawing.)
1 Telencephalon (yellow) with lateral ventricles
2 Diencephalon (orange) with III ventricle, optic nerve and retina
3 Mesencephalon (blue) with cerebral aqueduct
4 Metencephalon (green) with IV ventricle
5 Myelencephalon (yellow-green)

I. Prosencephalon (forebrain)	1. Telencephalon (cerebral hemispheres, striatum, etc.) 2. Diencephalon (thalamus, metathalamus, hypothalamus, etc.)
II. Mesencephalon (midbrain)	3. Mesencephalon (colliculi, cerebral peduncles, tegmentum)
III. Rhombencephalon (hindbrain)	4. Metencephalon (pons, cerebellum) 5. Myelencephalon (medulla oblongata)

Main divisions of the brain
I-III = primary brain vesicles; 1-5 = secondary brain vesicles

Midbrain, pons and medulla oblongata are collectively termed the **brain stem.**

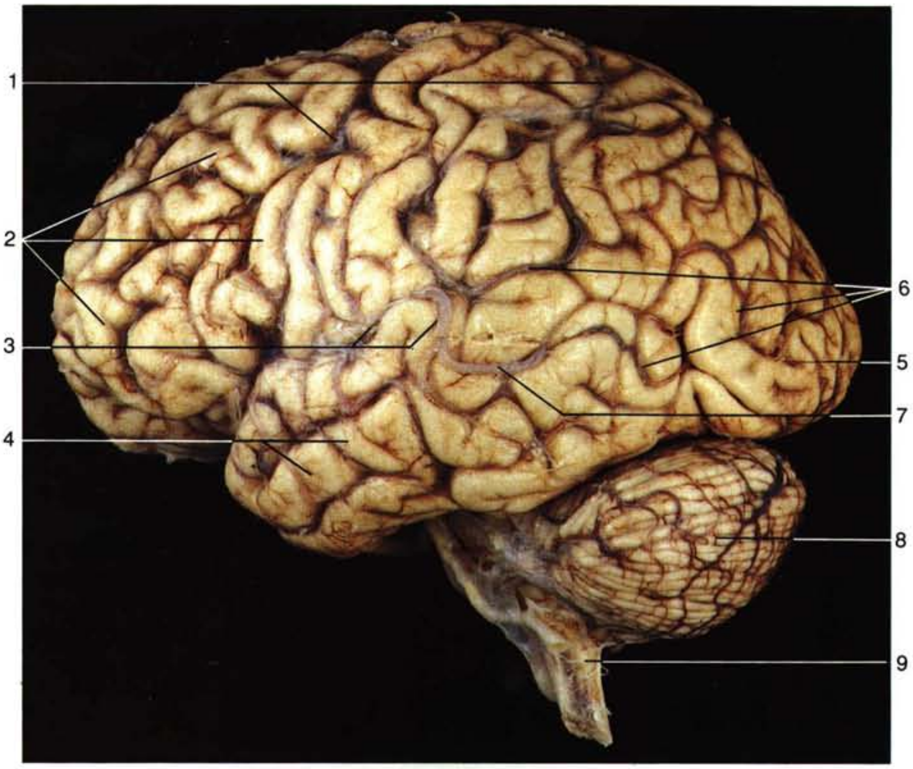

1 Superior cerebral veins and parietal lobe
2 Frontal lobe
3 Superficial middle cerebral vein and **cistern of lateral cerebral fossa**
4 Temporal lobe
5 Occipital lobe
6 Inferior cerebral veins and transverse occipital sulcus
7 Inferior anastomotic vein
8 Cerebellum
9 Medulla oblongata

Brain with pia mater. Cerebral veins (bluish). In the lateral sulcus the cistern of the lateral fossa is recognizable. Frontal lobe to the left.

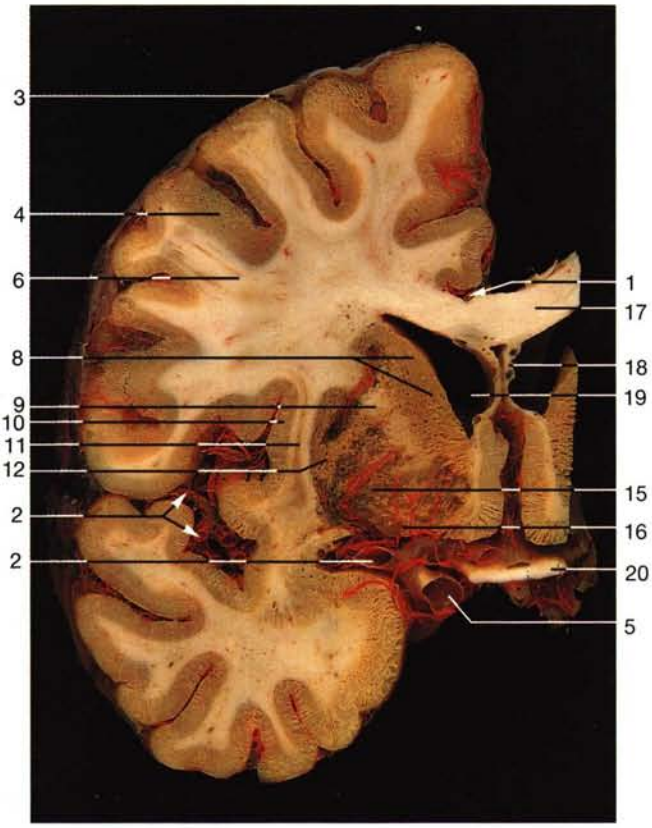

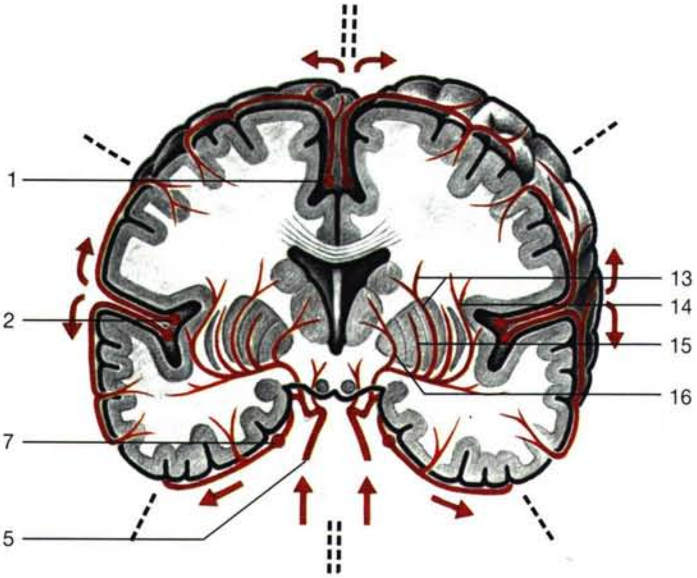

Arteries of the brain. Coronal section. Areas supplied by cortical and central arteries. Dotted lines indicate boundaries of arterial supply areas; arrows: direction of blood flow.

◁

Coronal section through the right hemisphere showing arachnoid, pia mater, and the arterial blood supply (anterior aspect).

1 **Anterior cerebral artery**
2 **Middle cerebral arteries**
3 Arachnoid
4 Cortex
5 **Internal carotid artery**
6 Frontal lobe (white matter)
7 **Posterior cerebral artery**
8 Caudate nucleus
9 Internal capsule
10 Insular lobe
11 Claustrum
12 Putamen
13 Posterior striate branch
14 Insular artery
15 Pallidostriate artery
16 Thalamic artery
17 Corpus callosum
18 Pellucid septum
19 Lateral ventricle
20 Optic chiasm

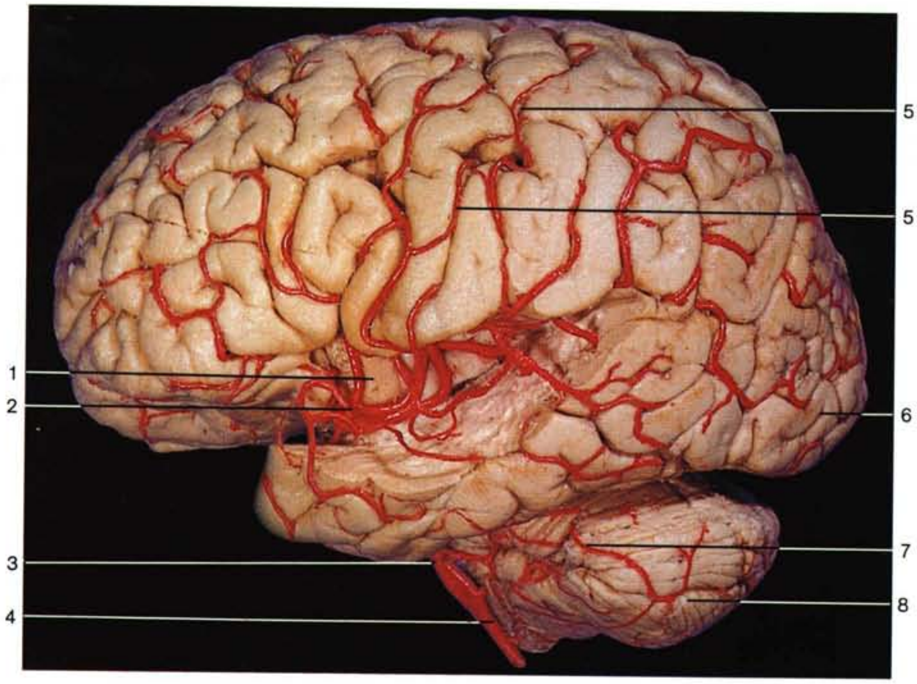

1 Insula
2 **Middle cerebral artery** [2 branches:
a) Parietal branches,
b) Temporal branches]
3 Basilar artery
4 **Vertebral artery**
5 Central sulcus
6 Occipital lobe
7 Superior cerebellar artery
8 Cerebellum
9 **Anterior cerebral artery**
10 Ethmoidal arteries
11 Ophthalmic artery
12 **Internal carotid artery**
13 Posterior communicating artery
14 **Posterior cerebral artery**
15 Anterior inferior cerebellar artery
16 Posterior inferior cerebellar artery

Cerebral arteries. Lateral aspect of the left hemisphere. The upper part of the temporal lobe has been removed to display the insula and cerebral arteries.

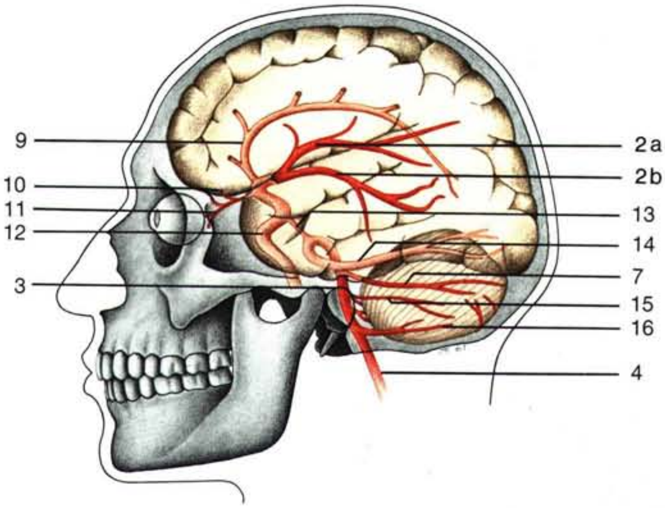

Arteries of the brain.

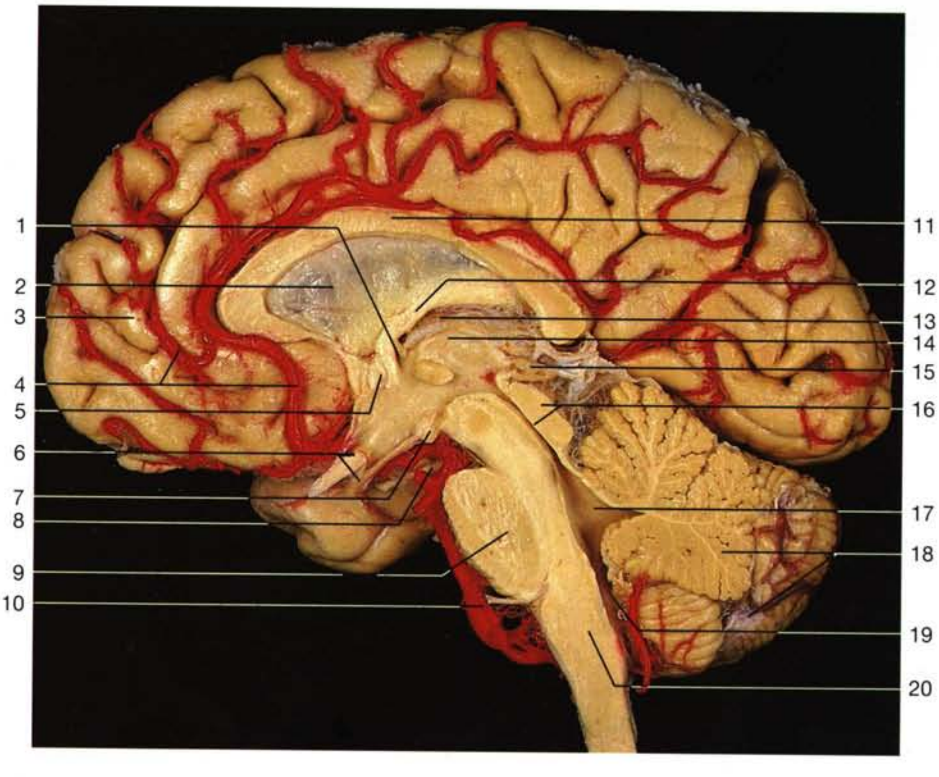

1 Interventricular foramen
2 Septum pellucidum
3 Frontal lobe
4 **Anterior cerebral artery**
5 Anterior commissure
6 Optic chiasma and infundibulum
7 Mamillary body
8 Oculomotor nerve (n. III)
9 Pons
10 **Basilar artery**
11 Corpus callosum
12 **Fornix**
13 **Choroid plexus**
14 **Third ventricle**
15 Pineal body
16 **Tectum** and **cerebral aqueduct**
17 **Fourth ventricle**
18 Cerebellum (arbor vitae, vermis)
19 Median aperture of Magendie
20 Medulla oblongata

Median section through the brain and brain stem. Cerebral arteries injected with red resin.

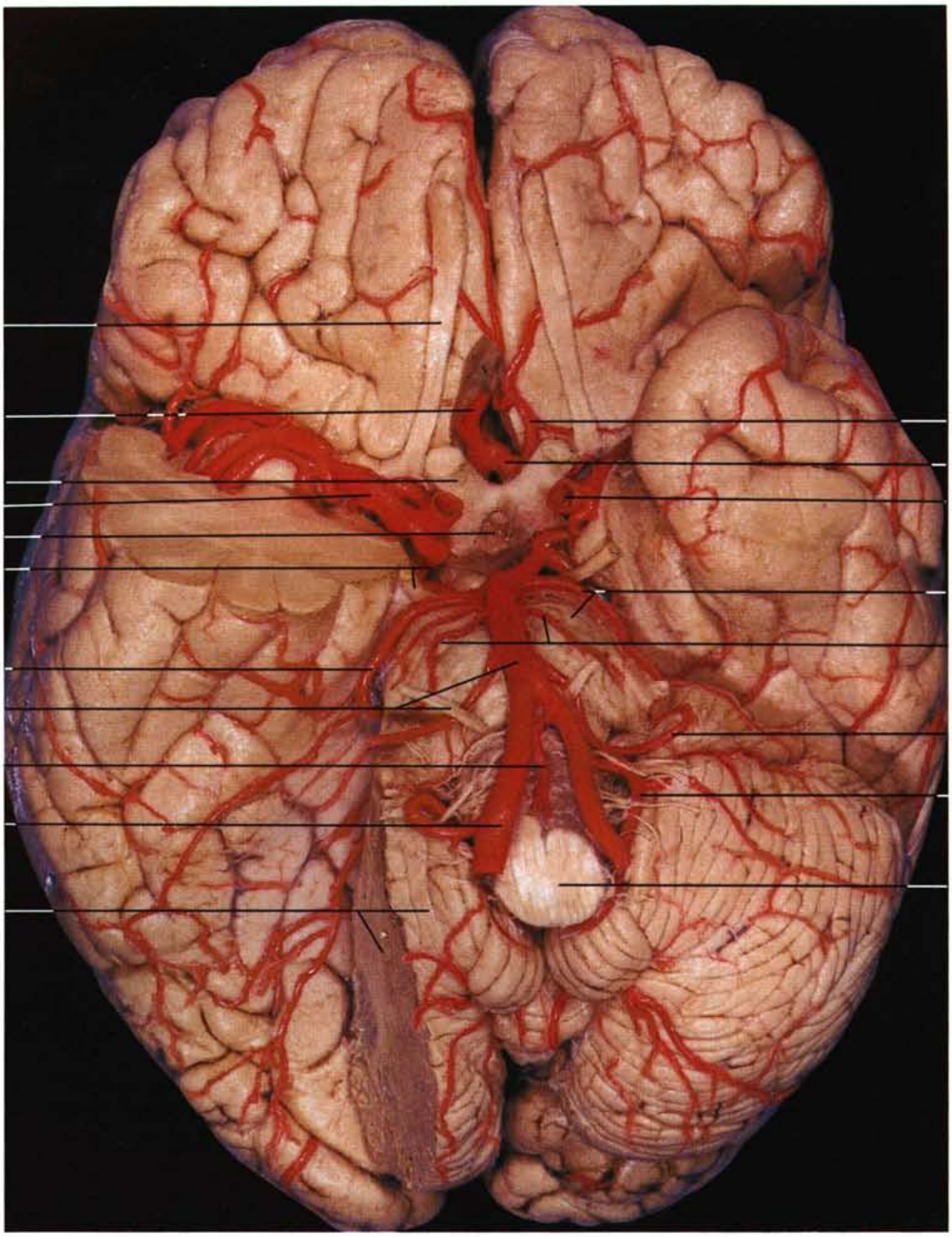

Arteries of the brain (inferior aspect). Frontal pole above; right temporal lobe and cerebellum partly removed.

1 Anterior communicating artery
2 Left **anterior cerebral artery**
3 **Internal carotid artery**
4 Pons and left **superior cerebellar artery**
5 **Anterior inferior cerebellar artery**
6 **Posterior inferior cerebellar artery**
7 Medulla oblongata
8 **Right anterior cerebral artery**
9 Olfactory tract
10 Optic nerve
11 **Middle cerebral artery**
12 Infundibulum
13 Oculomotor nerve and **posterior communicating artery**
14 **Posterior cerebral artery**
15 **Basilar artery** and abducens nerve (n. VI)
16 Anterior spinal artery
17 **Vertebral artery**
18 Cerebellum
19 Labyrinthine arteries
20 Posterior spinal artery
21 Right **superior cerebellar artery**
22 Olfactory bulb

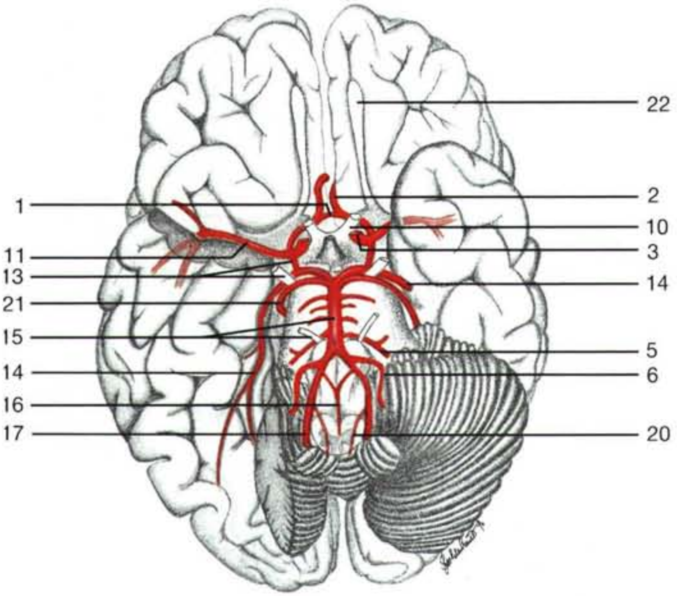

Arteries of the base of the brain, arterial circle of Willis. (Schematic drawing.)

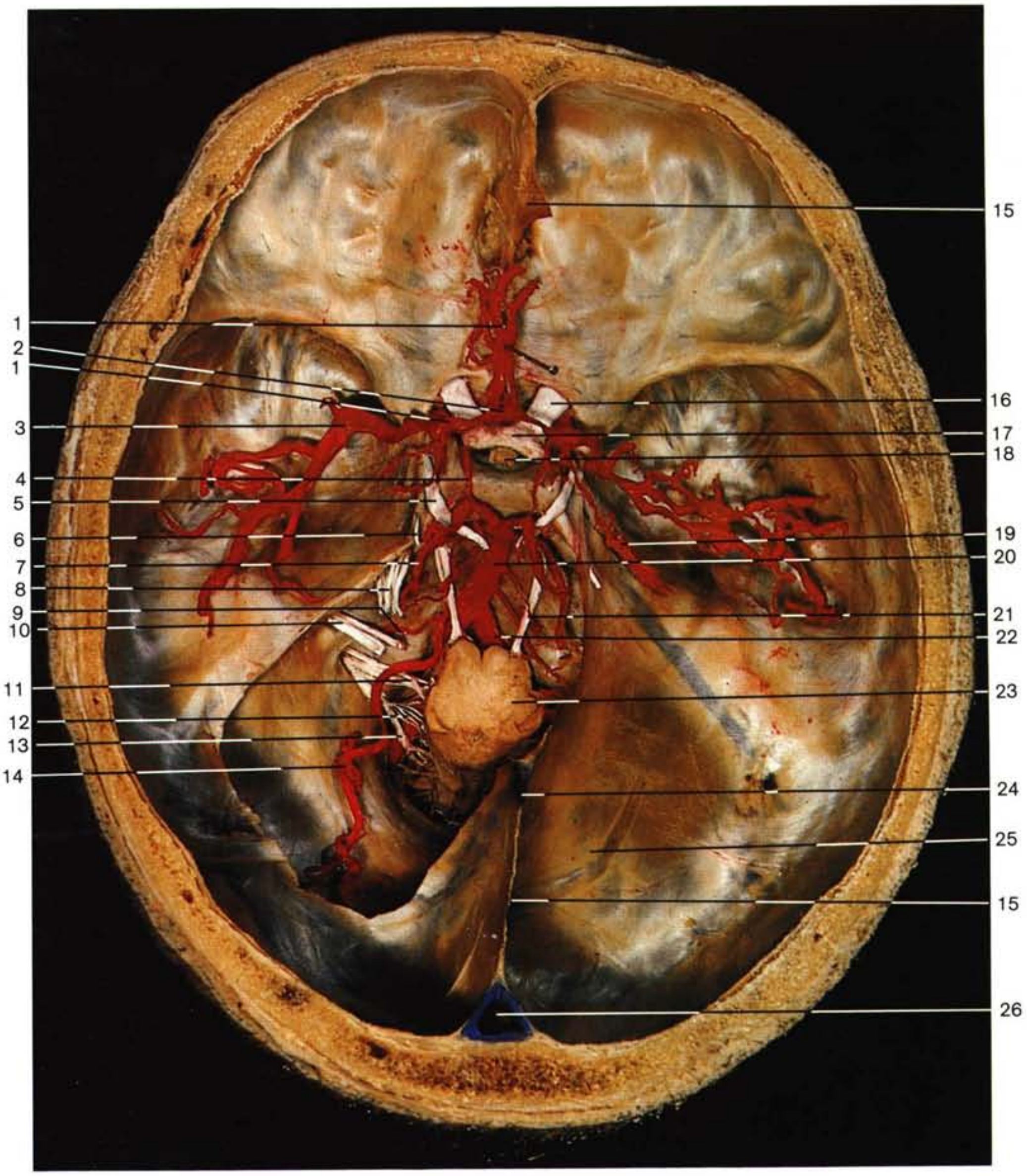

Base of the cranial cavity (internal aspect). Left tentorium has been cut.

1	Anterior cerebral artery
2	Anterior communicating artery
3	**Middle cerebral artery**
4	Posterior communicating artery
5	**Oculomotor nerve**
6	**Trochlear nerve**
7	Posterior cerebral artery
8	**Trigeminal nerve**
9	Internal auditory artery
10	**Facial nerve** and **vestibulocochlear nerve**
11	**Glossopharyngeal nerve** and **vagus nerve**
12	**Hypoglossal nerve**
13	**Accessory nerve**
14	Anterior inferior cerebellar artery
15	Falx cerebri
16	**Optic nerve**
17	Optic chiasma
18	Infundibulum and pituitary gland
19	Anterior choroidal artery and choroid plexus
20	**Basilar artery**
21	**Abducens nerve**
22	Vertebral arteries
23	Medulla oblongata
24	Inferior sagittal sinus
25	Tentorium cerebelli
26	Superior sagittal sinus and confluence of sinuses
27	Anterior spinal artery
28	Internal carotid artery
29	Superior cerebellar artery
30	Anterior inferior cerebellar artery
31	Posterior inferior cerebellar artery
32	Posterior spinal artery
33	Ophthalmic artery

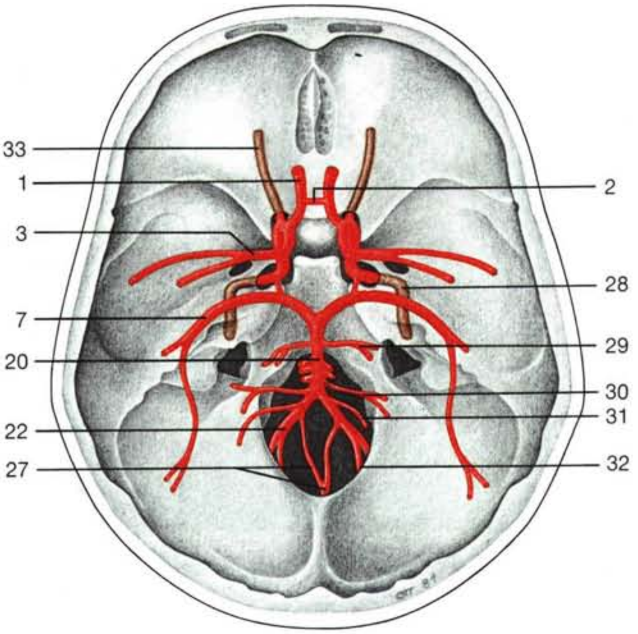

Arterial circle of Willis (superior aspect). (Schematic drawing.)

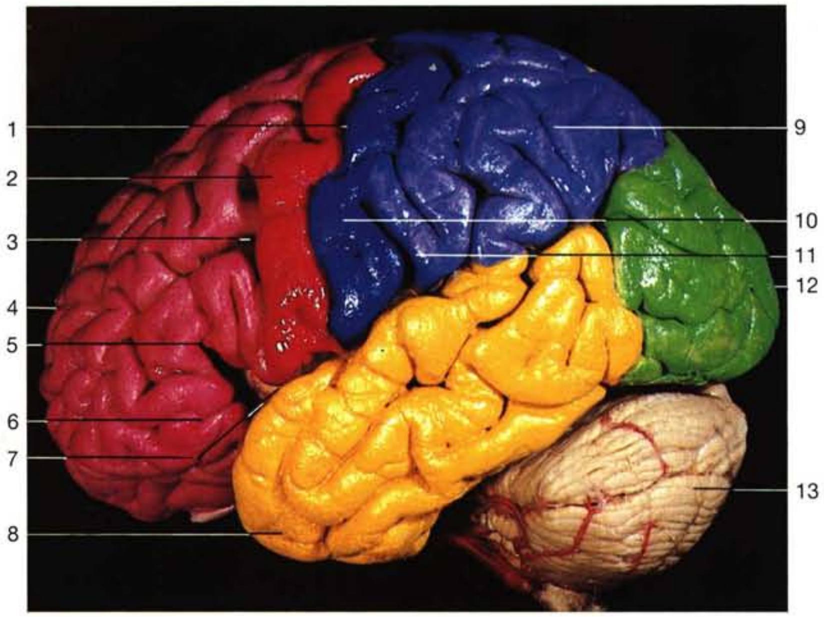

1 Central sulcus
2 Precentral gyrus
3 Precentral sulcus
4 **Frontal lobe**
5 Anterior ascending ramus of lateral sulcus
6 Anterior horizontal ramus of lateral sulcus
7 Lateral sulcus
8 **Temporal lobe**
9 **Parietal lobe**
10 Postcentral gyrus
11 Postcentral sulcus
12 **Occipital lobe**
13 Cerebellum
14 Superior frontal sulcus
15 Middle frontal gyrus
16 Lunate sulcus
17 Longitudinal fissure
18 Arachnoid granulations

Brain, left hemisphere (lateral aspect). Frontal pole to the left.

Pink = Frontal lobe
Blue = Parietal lobe
Green = Occipital lobe
Yellow = Temporal lobe
Dark red = Precentral gyrus
Dark blue = Postcentral gyrus

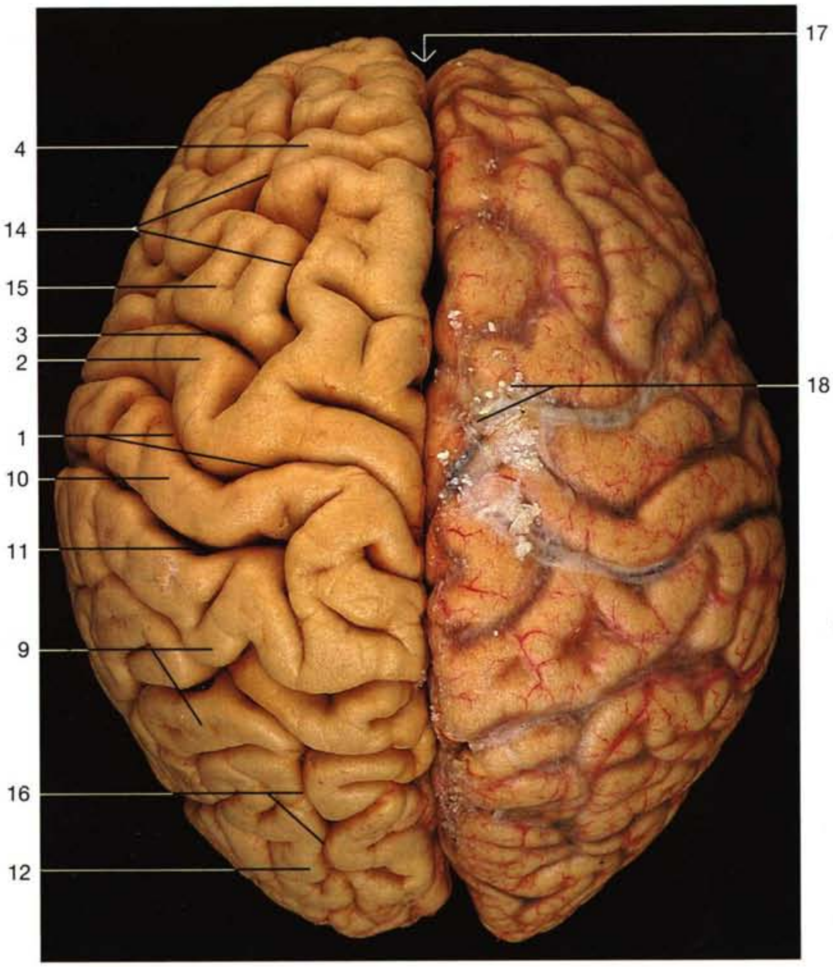

Brain (superior aspect). Right hemisphere with arachnoid and pia mater.

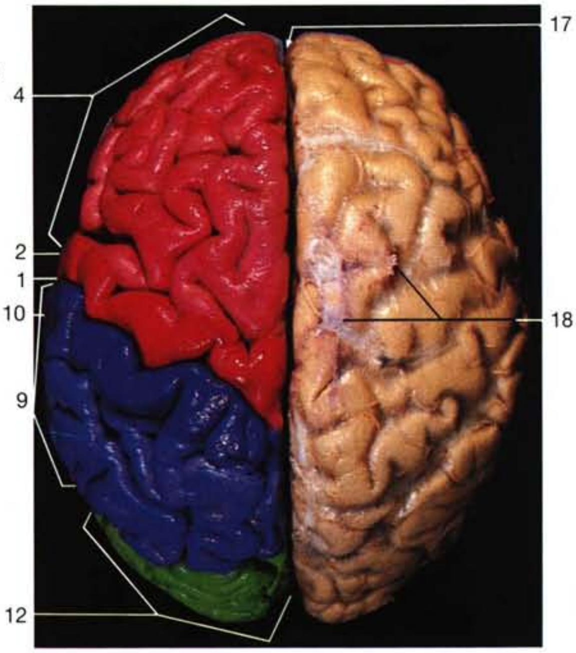

Brain (superior aspect). Lobes of the left hemisphere indicated by color; right hemisphere is covered with arachnoid and pia mater.

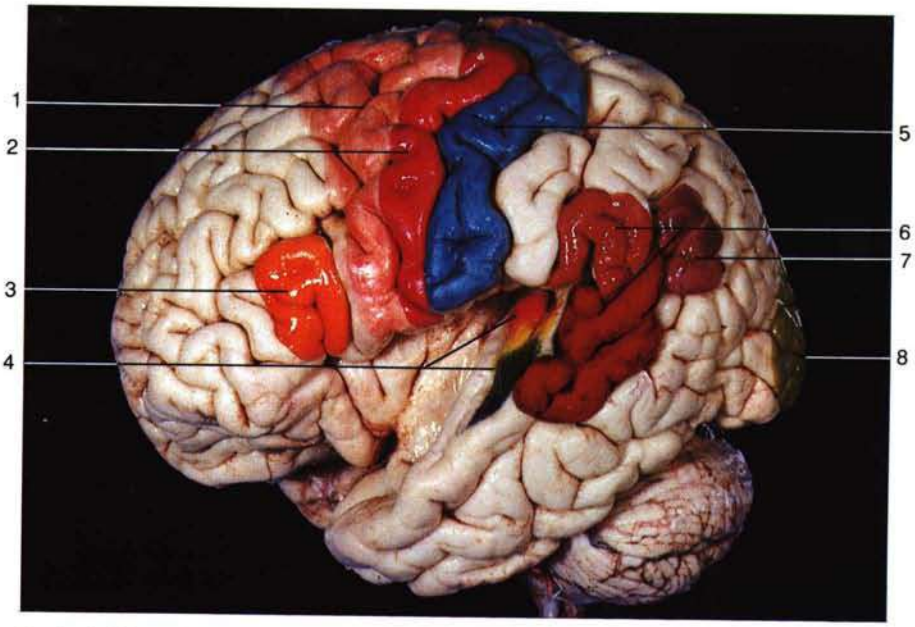

1 Premotor area
2 Somatomotor area
3 Motor speech area of Broca
4 Acoustic area
(red: high tone; blue: low tone)
5 Somatosensory area
6 Sensory speech area of Wernicke
7 Reading comprehension area
8 Visuosensory area

Brain, left hemisphere (lateral aspect). **Main cortical areas** are colored. The lateral sulcus has been opened to display the insula and the inner surface of the temporal lobe.

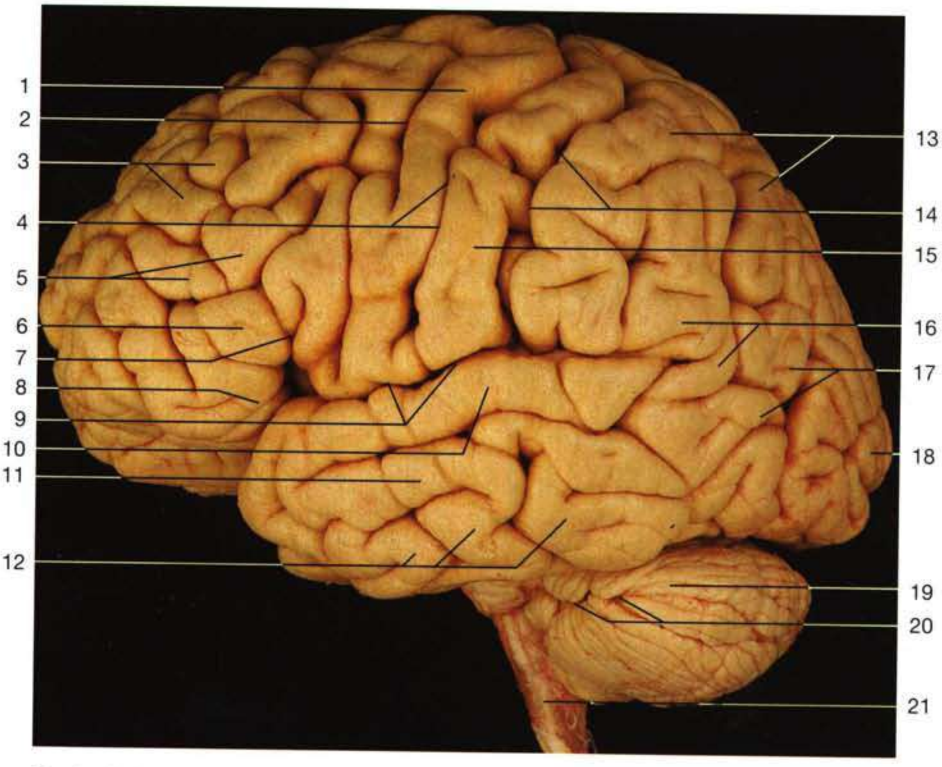

1	Precentral gyrus	
2	Precentral sulcus	
3	Superior frontal gyrus	
4	**Central sulcus**	
5	Middle frontal gyrus	
6	Inferior frontal gyrus	
7	Ascending ramus	of lateral sulcus
8	Horizontal ramus	
9	Posterior ramus	
10	Superior temporal gyrus	
11	Middle temporal gyrus	
12	Inferior temporal gyrus	
13	Parietal lobule	
14	Postcentral sulcus	
15	Postcentral gyrus	
16	Supramarginal gyrus	
17	Angular gyrus	
18	**Occipital lobe**	
19	**Cerebellum**	
20	Horizontal fissure of cerebellum	
21	Medulla oblongata	

Brain, left hemisphere (lateral aspect). Frontal pole to the left.

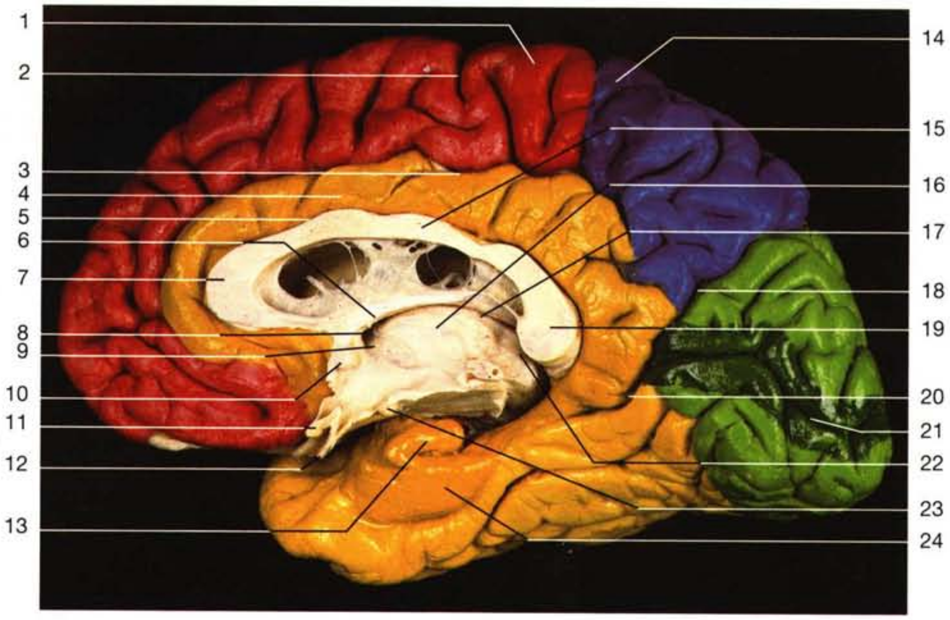

Brain, right hemisphere (medial aspect). Frontal pole to the left (midbrain divided, cerebellum and inferior part of brainstem removed).

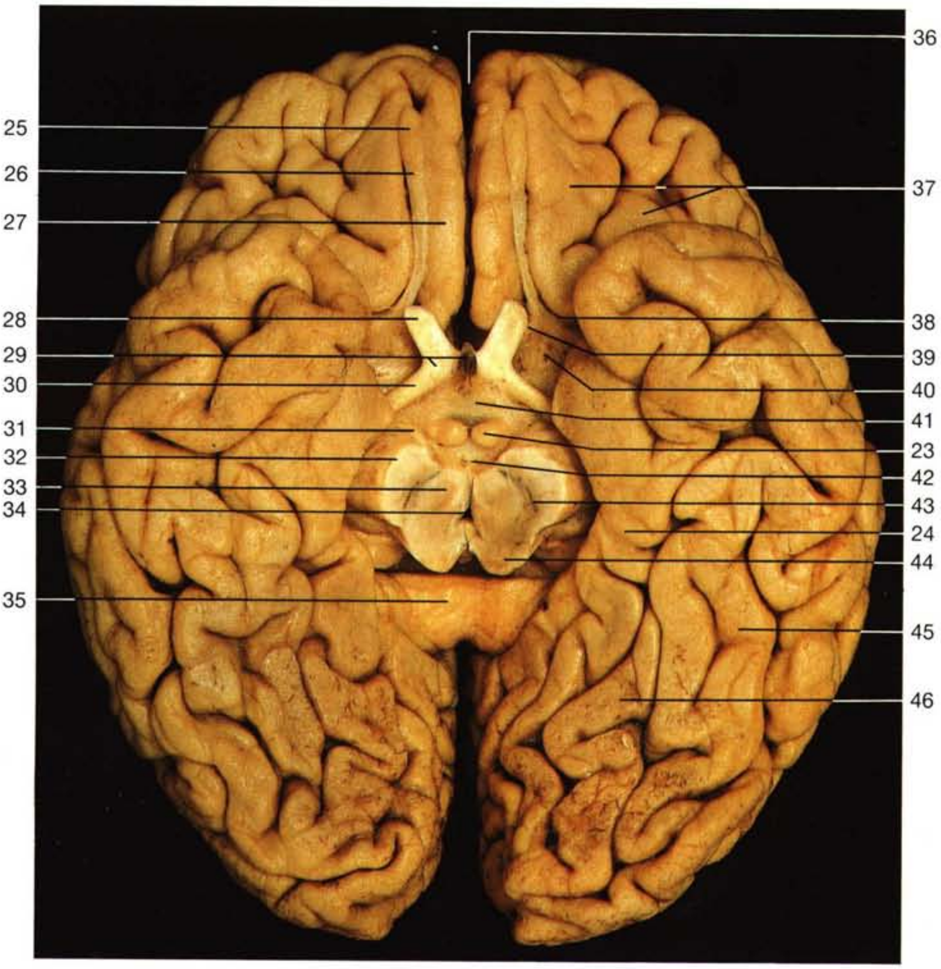

Brain (inferior aspect). Midbrain divided. Cerebellum and inferior part of brain stem removed. Frontal pole at the top.

1 **Precentral gyrus**
2 Precentral sulcus
3 Cingulate sulcus
4 **Cingulate gyrus**
5 Sulcus of corpus callosum
6 **Fornix**
7 Genu of **corpus callosum**
8 Interventricular foramen
9 Intermediate mass
10 Anterior commissure
11 **Optic chiasma**
12 **Infundibulum**
13 Uncus hippocampi
14 **Postcentral gyrus**
15 Body of corpus callosum
16 Third ventricle and thalamus
17 Stria medullaris
18 **Parietooccipital sulcus**
19 Splenium of corpus callosum
20 Communication of calcarine and parietooccipital sulcus
21 Calcarine sulcus
22 **Pineal body**
23 **Mamillary body**
24 **Parahippocampal gyrus**
25 **Olfactory bulb**
26 Olfactory tract
27 Gyrus rectus
28 **Optic nerve**
29 Infundibulum and optic chiasma
30 **Optic tract**
31 Oculomotor nerve
32 Pedunculus cerebri
33 **Red nucleus**
34 **Cerebral aqueduct**
35 Corpus callosum
36 Longitudinal fissure
37 Orbital gyri
38 Lateral root of olfactory tract
39 Medial root of olfactory tract
40 Olfactory tubercle and anterior perforated substance
41 Tuber cinereum
42 Interpeduncular fossa
43 **Substantia nigra**
44 Colliculi of the **midbrain**
45 Lateral occipitotemporal gyrus
46 Medial occipitotemporal gyrus

Pink = Frontal lobe
Blue = Parietal lobe
Green = Occipital lobe
Yellow = Temporal lobe
Dark red = Precentral lobe
Dark blue = Postcentral lobe
Orange = Limbic cortex (cingulate and parahippocampal gyri)

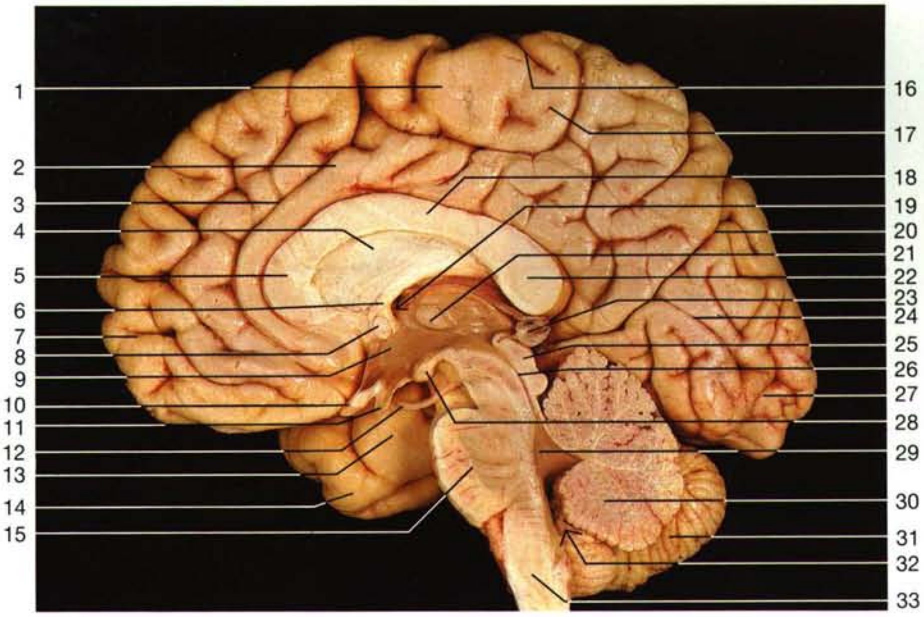

Brain (sagittal section). Frontal pole to the left.

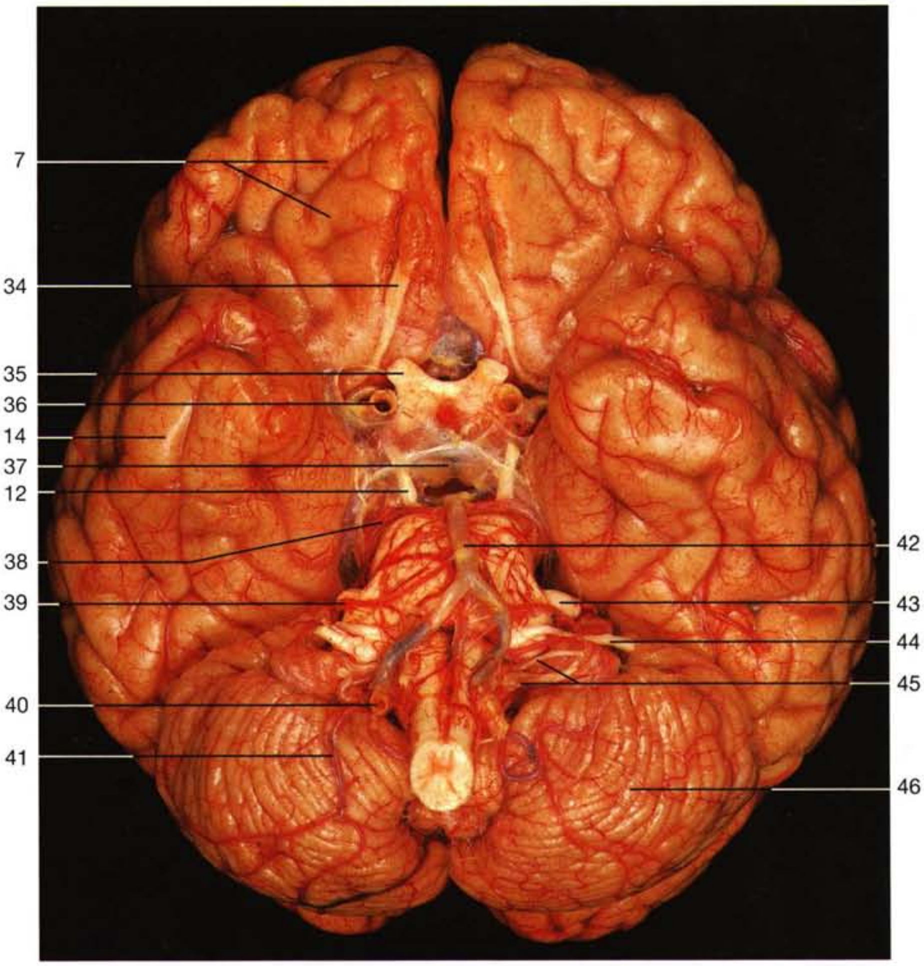

Brain, with pia mater and blood vessels (inferior aspect).

1	Precentral gyrus
2	Cingulate gyrus
3	Cingulate sulcus
4	Septum pellucidum
5	Genu of corpus callosum
6	Fornix
7	**Frontal lobe**
8	Anterior commissure
9	**Hypothalamus**
10	Optic chiasma
11	Infundibulum
12	Oculomotor nerve
13	Uncus
14	**Temporal lobe**
15	Pons
16	Central sulcus
17	Postcentral gyrus
18	Body of corpus callosum
19	Interventricular foramen (arrow)
20	Parietooccipital sulcus
21	Intermediate mass
22	Splenium of corpus callosum
23	Pineal body
24	Calcarine sulcus
25	Colliculi of midbrain
26	Cerebral aqueduct
27	**Occipital lobe**
28	Mamillary body
29	Fourth ventricle
30	Vermis of cerebellum
31	Right hemisphere of cerebellum
32	Median aperture of Magendie (arrow)
33	Medulla oblongata
34	Olfactory tract
35	Optic nerve
36	**Internal carotid artery**
37	Interpeduncular cistern
38	Superior cerebellar artery
39	Anterior inferior cerebellar artery
40	**Vertebral artery**
41	Posterior inferior cerebellar artery
42	**Basilar artery**
43	Trigeminal nerve (n. V)
44	Facial nerve (n. VII)
45	Accessory nerve (n. XI), hypoglossal nerve (n. XII)
46	Cerebellum

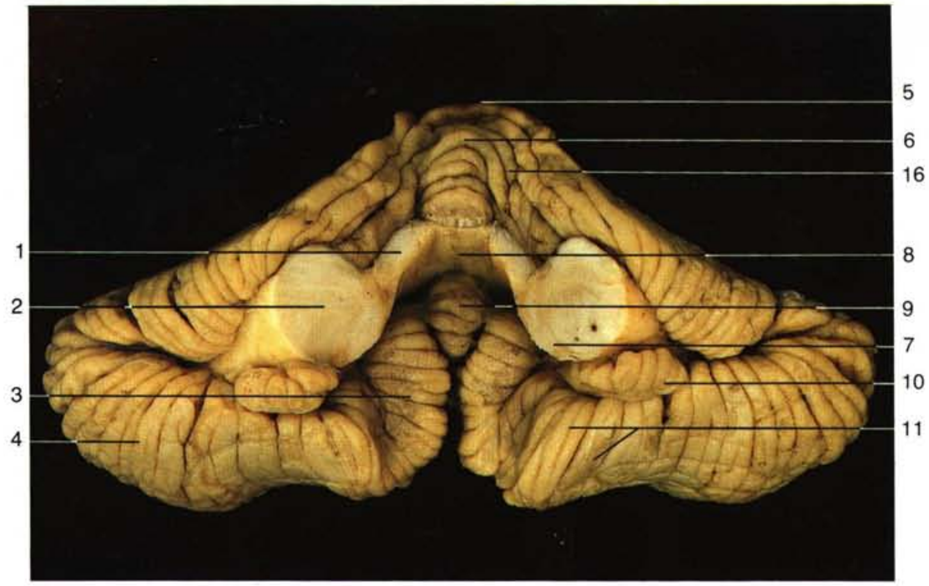

Cerebellum (inferior anterior aspect). The cerebellar peduncles have been severed.

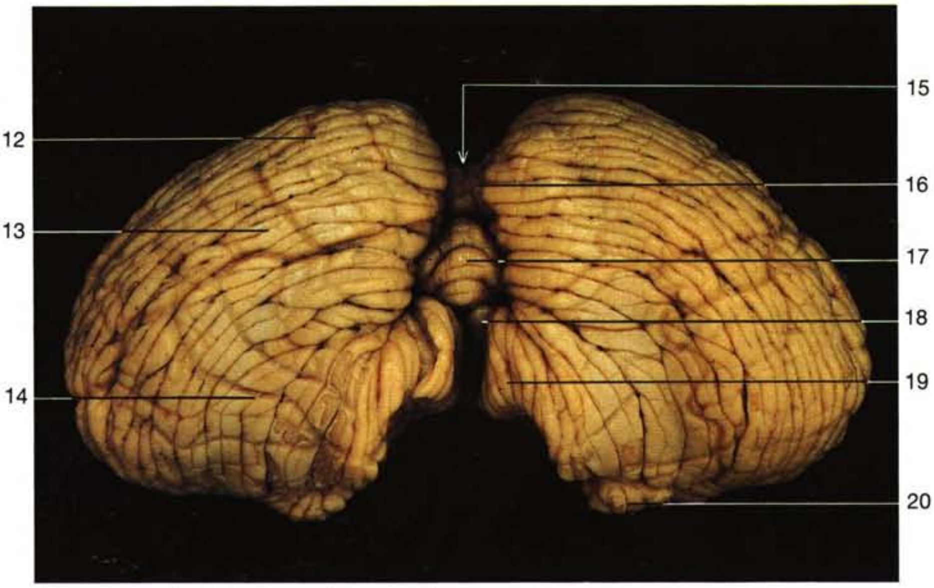

Cerebellum (inferior posterior aspect).

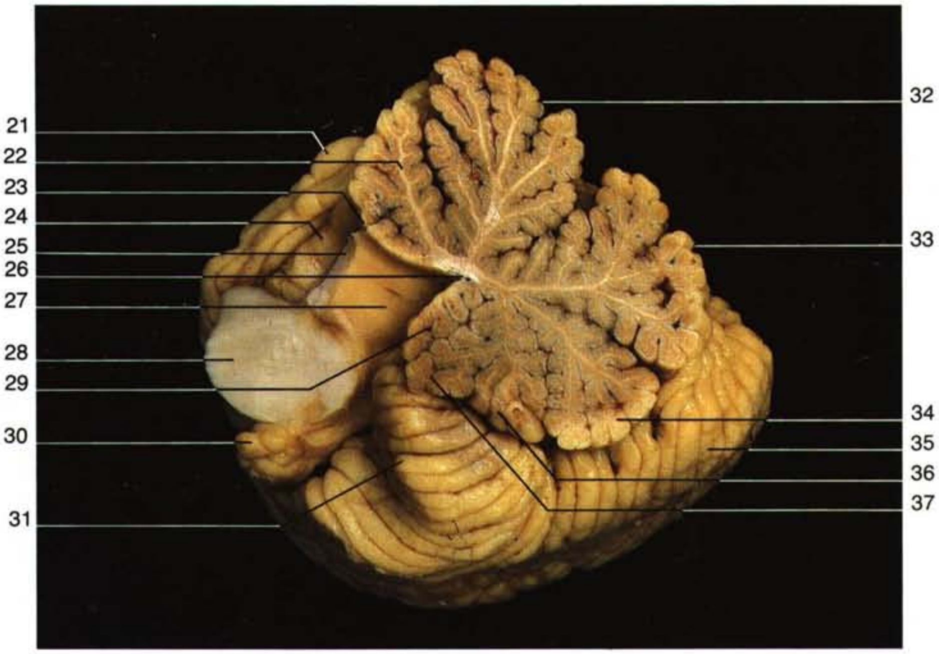

Median section through the cerebellum. Right cerebellar hemisphere and right half of vermis.

1 Superior cerebellar peduncle
2 Middle cerebellar peduncle
3 Cerebellar tonsil
4 Inferior semilunar lobule
5 **Vermis**
6 Central lobule of vermis
7 Inferior cerebellar peduncle
8 Superior medullary velum
9 **Nodule of vermis**
10 **Flocculus**
11 Biventral lobule
12 Left cerebellar hemisphere
13 Inferior semilunar lobule
14 Biventral lobule
15 Vermis of cerebellum
16 Tuber of vermis
17 Pyramid of vermis
18 Uvula of vermis
19 Tonsil of cerebellum
20 Floccule of cerebellum
21 Right cerebellar hemisphere
22 Vermis (central lobule)
23 Cerebellar lingula
24 Ala of central lobule
25 **Superior cerebellar peduncle**
26 Fastigium
27 **Fourth ventricle**
28 **Middle cerebellar peduncle**
29 Nodule of vermis
30 Flocculus of cerebellum
31 Cerebellar tonsil
32 Culmen of vermis
33 Declive of vermis
34 Tuber of vermis
35 Inferior semilunar lobule
36 Pyramis of vermis
37 Uvula of vermis

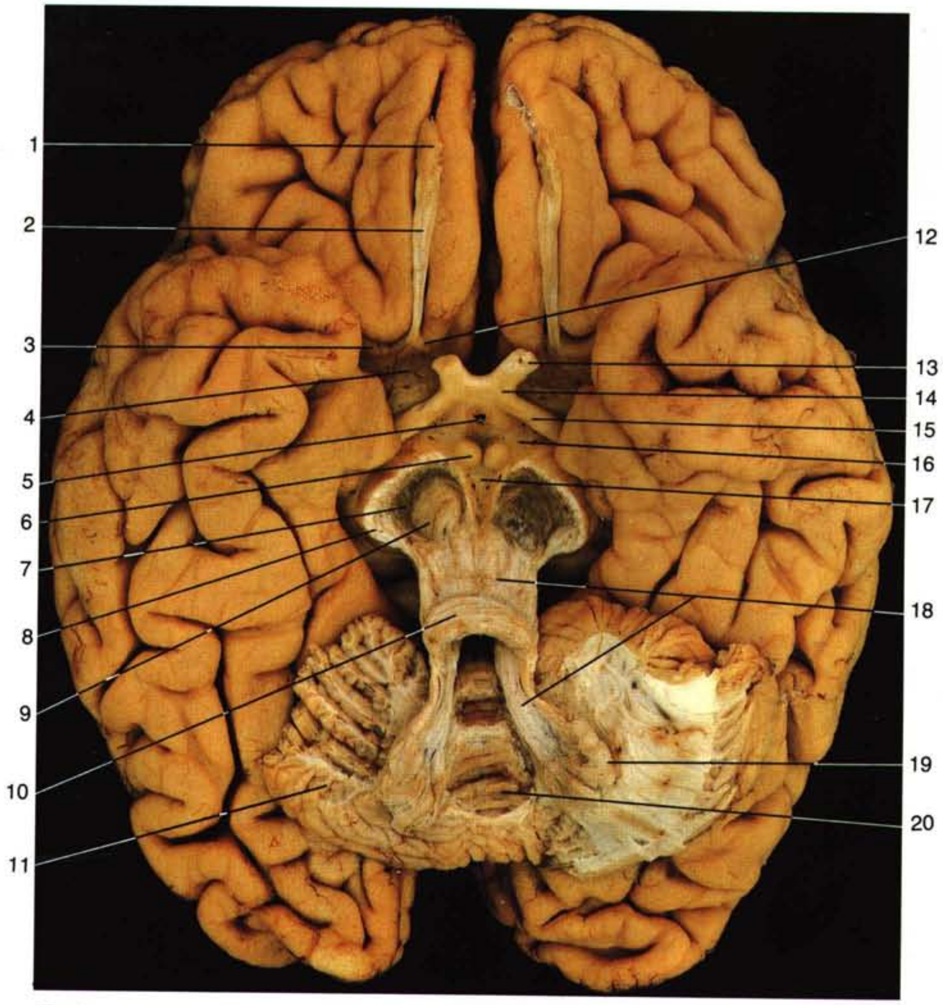

Brain and cerebellum (inferior aspect). Parts of the cerebellum have been removed to display the dentate nucleus and the main pathway to the midbrain (cerebellorubral tract).

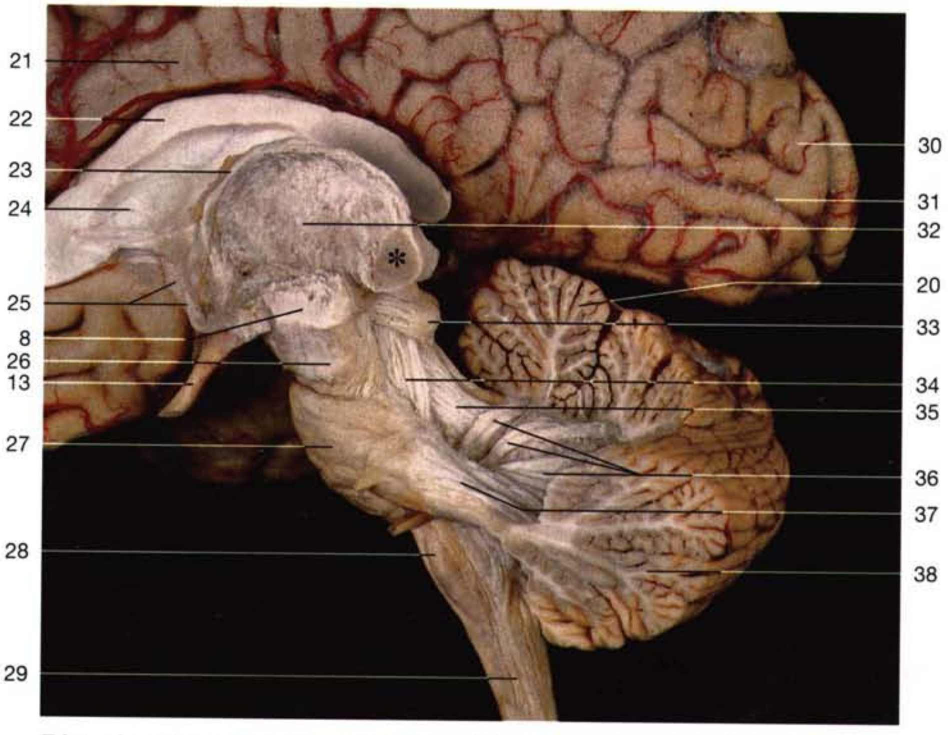

Dissection of the cerebellar peduncles and their connection with midbrain and diencephalon. A small part of pulvinar thalami (*) has been cut to show inferior brachium.

1 Olfactory bulb
2 **Olfactory tract**
3 Lateral olfactory stria
4 Anterior perforated substance
5 Infundibulum (divided)
6 **Mamillary body**
7 Substantia nigra
8 Pedunculus cerebri
9 **Red nucleus**
10 Decussation of superior cerebellar peduncle
11 **Cerebellar hemisphere**
12 Medial olfactory stria
13 **Optic nerve**
14 Optic chiasma
15 Optic tract
16 Posterior perforated substance
17 Interpeduncular fossa
18 **Superior cerebellar peduncle** and cerebellorubral tract
19 **Dentate nucleus**
20 **Vermis of cerebellum**
21 Cingulate gyrus
22 **Corpus callosum**
23 Stria terminalis
24 Septum pellucidum
25 **Columna fornicis**
26 Cerebral peduncle at midbrain level
27 Pons
28 Inferior olive
29 Medulla oblongata with lateral pyramidal tract
30 Occipital lobe
31 Calcarine sulcus
32 **Thalamus**
33 Inferior colliculus with brachium
34 **Medial lemniscus**
35 **Superior cerebellar peduncle**
36 **Inferior cerebellar peduncle**
37 **Middle cerebellar peduncle**
38 Cerebellar hemisphere

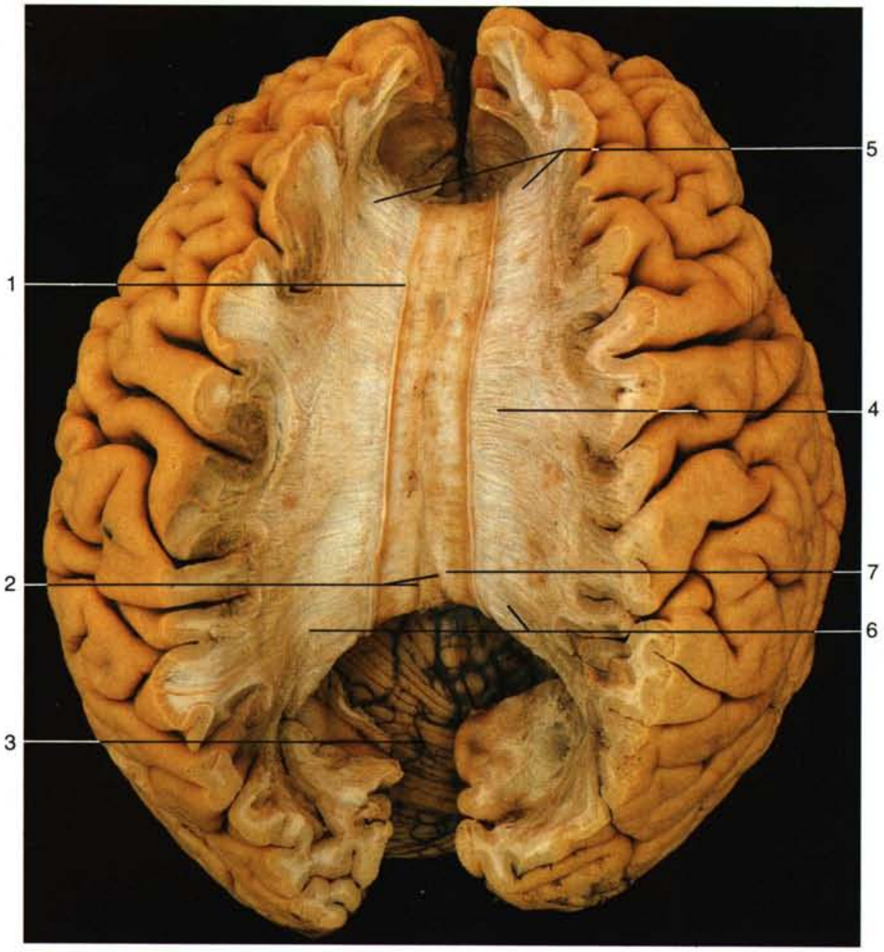

1 Lateral longitudinal stria of indusium griseum
2 Medial longitudinal stria of indusium griseum
3 Cerebellum
4 Radiating fibers of the corpus callosum
5 Forceps minor of corpus callosum
6 Forceps major of corpus callosum
7 Splenium of corpus callosum

Dissection of the brain I. The fiber system of the corpus callosum has been displayed by removing the cortex lying above it. Frontal pole at top.

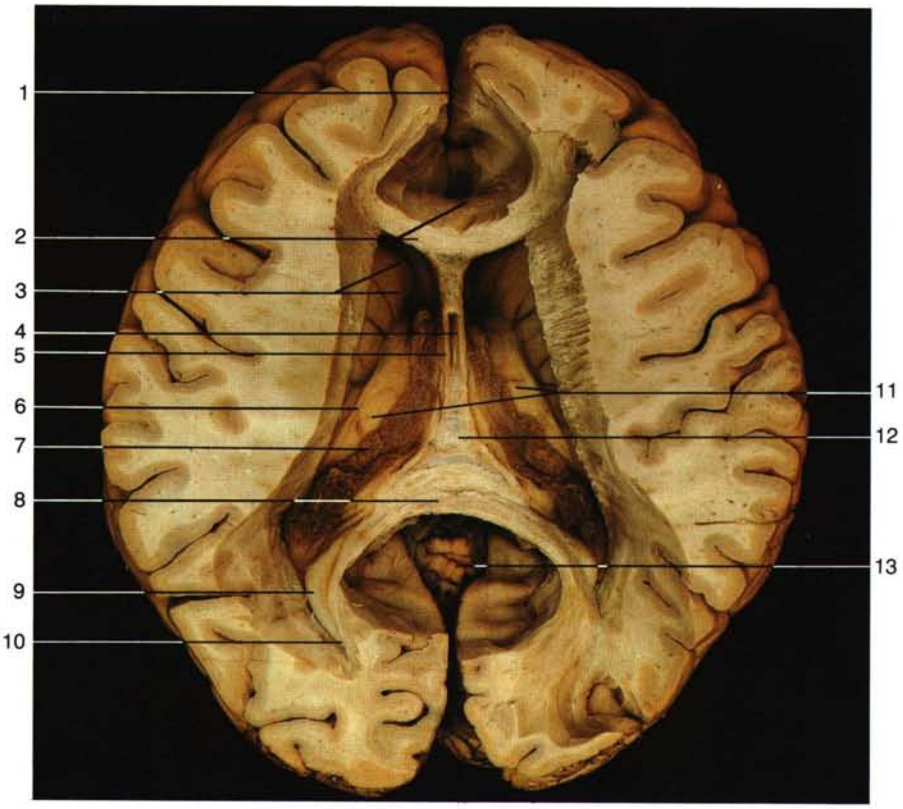

1 Longitudinal cerebral fissure
2 Genu of corpus callosum
3 Head of **caudate nucleus** and anterior horn of lateral ventricle
4 Cavum of septum pellucidum
5 **Septum pellucidum**
6 Stria terminalis
7 **Choroid plexus** of **lateral ventricle**
8 Splenium of **corpus callosum**
9 Calcar avis
10 Posterior horn of lateral ventricle
11 **Thalamus** (lamina affixa)
12 Commissure of fornix
13 Vermis of cerebellum

Dissection of the brain II. The lateral ventricles and subcortical nuclei of the brain are dissected. The corpus callosum has been partly removed. Frontal pole at the top.

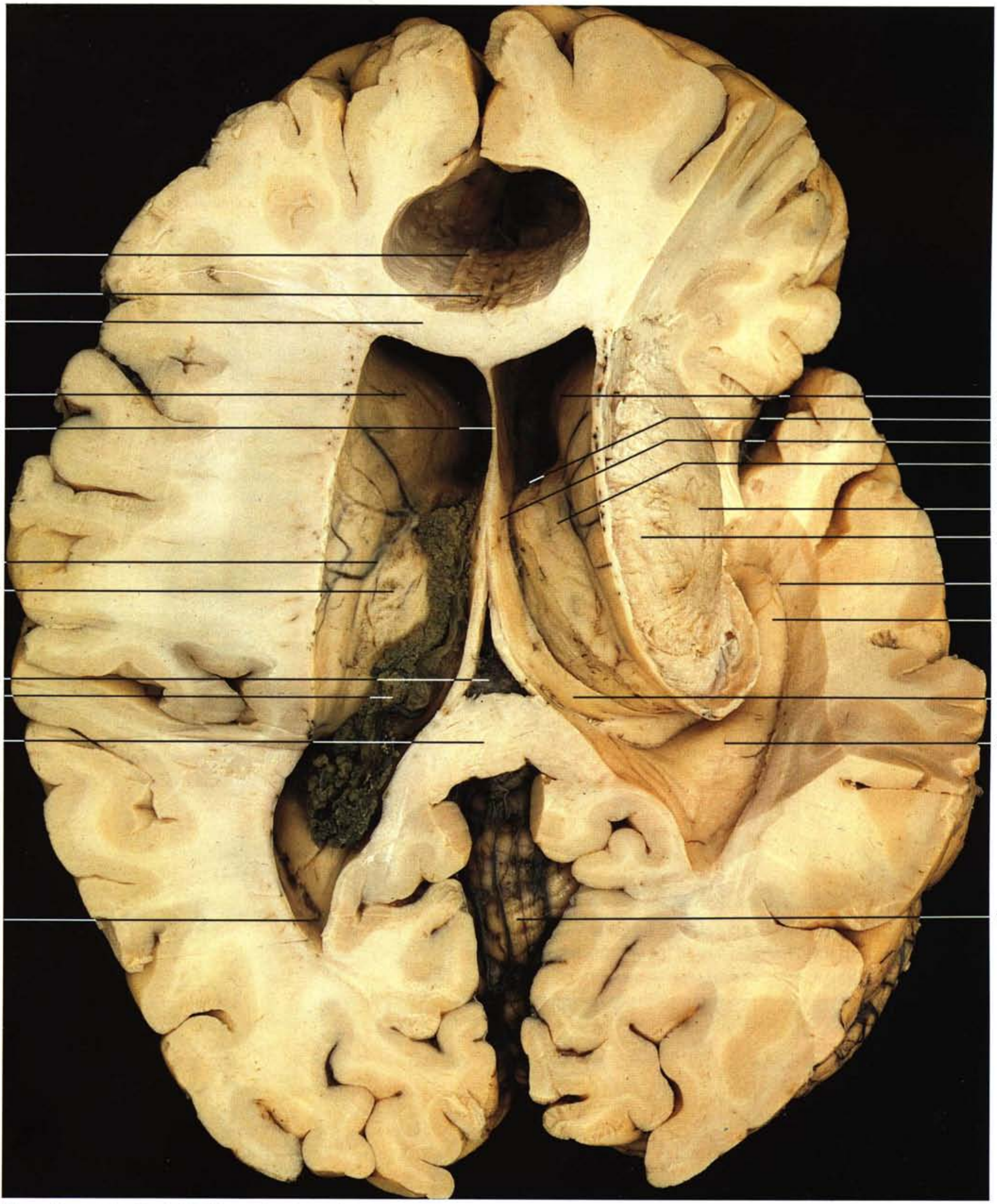

Dissection of the brain III (superior view of lateral ventricle and subcortical nuclei of the brain). Corpus callosum partly removed. At right, the entire lateral ventricle has been opened, the insula with claustrum, extreme and external capsules have been removed, exposing the lentiform nucleus and the internal capsule.

1	Lateral longitudinal stria
2	Medial longitudinal stria
3	Genu of corpus callosum
4	Head of **caudate nucleus**
5	**Septum pellucidum**
6	Stria terminalis
7	**Thalamus** (lamina affixa)
8	Choroid plexus of third ventricle
9	**Choroid plexus** of **lateral ventricle**
10	Splenium of corpus callosum
11	Posterior horn of lateral ventricle
12	Anterior horn of lateral ventricle (head of caudate nucleus)
13	**Putamen** of **lentiform nucleus**
14	**Internal capsule**
15	Inferior horn of lateral ventricle
16	**Pes hippocampi**
17	**Crus of fornix**
18	Vermis of cerebellum with arachnoid and pia mater
19	Interventricular foramen
20	Right column of fornix
21	Collateral eminence

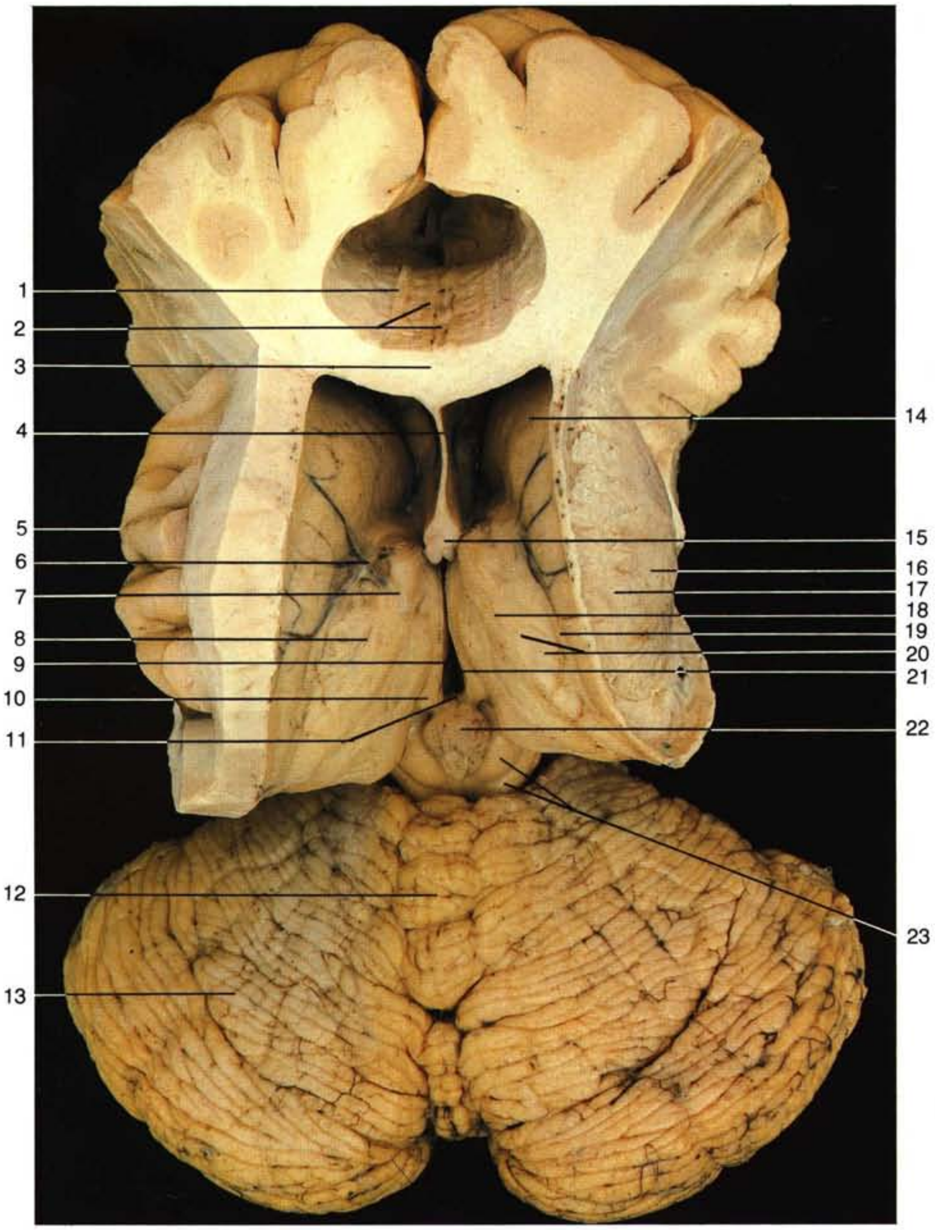

1 Lateral longitudinal stria
2 Medial longitudinal stria
3 **Corpus callosum**
4 Septum pellucidum
5 Insular gyri
6 Thalamostriate vein
7 Anterior tubercle of thalamus
8 **Thalamus**
9 Medullary stria of thalamus
10 Habenular trigone
11 Habenular commissure
12 **Vermis of the cerebellum**
13 Left **hemisphere of cerebellum**
14 **Head of caudate nucleus**
15 **Columns of fornix**
16 Putamen of lentiform nucleus
17 **Internal capsule**
18 Taenia of choroid plexus
19 Stria terminalis and thalamostriate vein
20 Lamina affixa
21 **Third ventricle**
22 **Pineal body**
23 Superior and inferior colliculus of midbrain

Dissection of the brain IVa. Temporal lobe, fornix and the posterior corpus callosum have been removed (this part of the specimen is depicted below). Frontal pole at top (superior aspect).

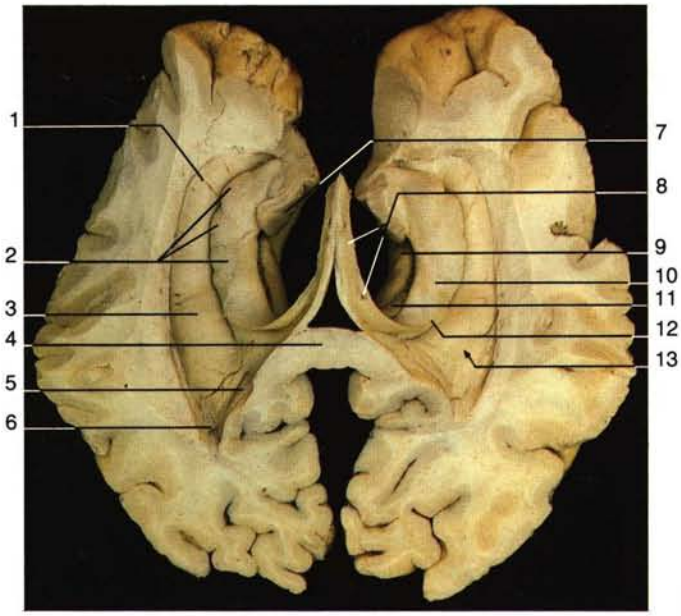

1 Inferior horn of lateral ventricle
2 **Hippocampal digitations**
3 Collateral eminence
4 Splenium of corpus callosum
5 Calcar avis
6 Posterior horn of lateral ventricle
7 **Uncus** of parahippocampal gyrus
8 Body and crus of **fornix**
9 **Parahippocampal gyrus**
10 **Pes hippocampi**
11 Dentate gyrus
12 Hippocampal fimbria
13 Lateral ventricle

Dissection of the brain IVb. Depicted is the portion of the brain removed from the specimen above. **Temporal lobe and limbic system** (superior aspect). Columns of fornix are served.

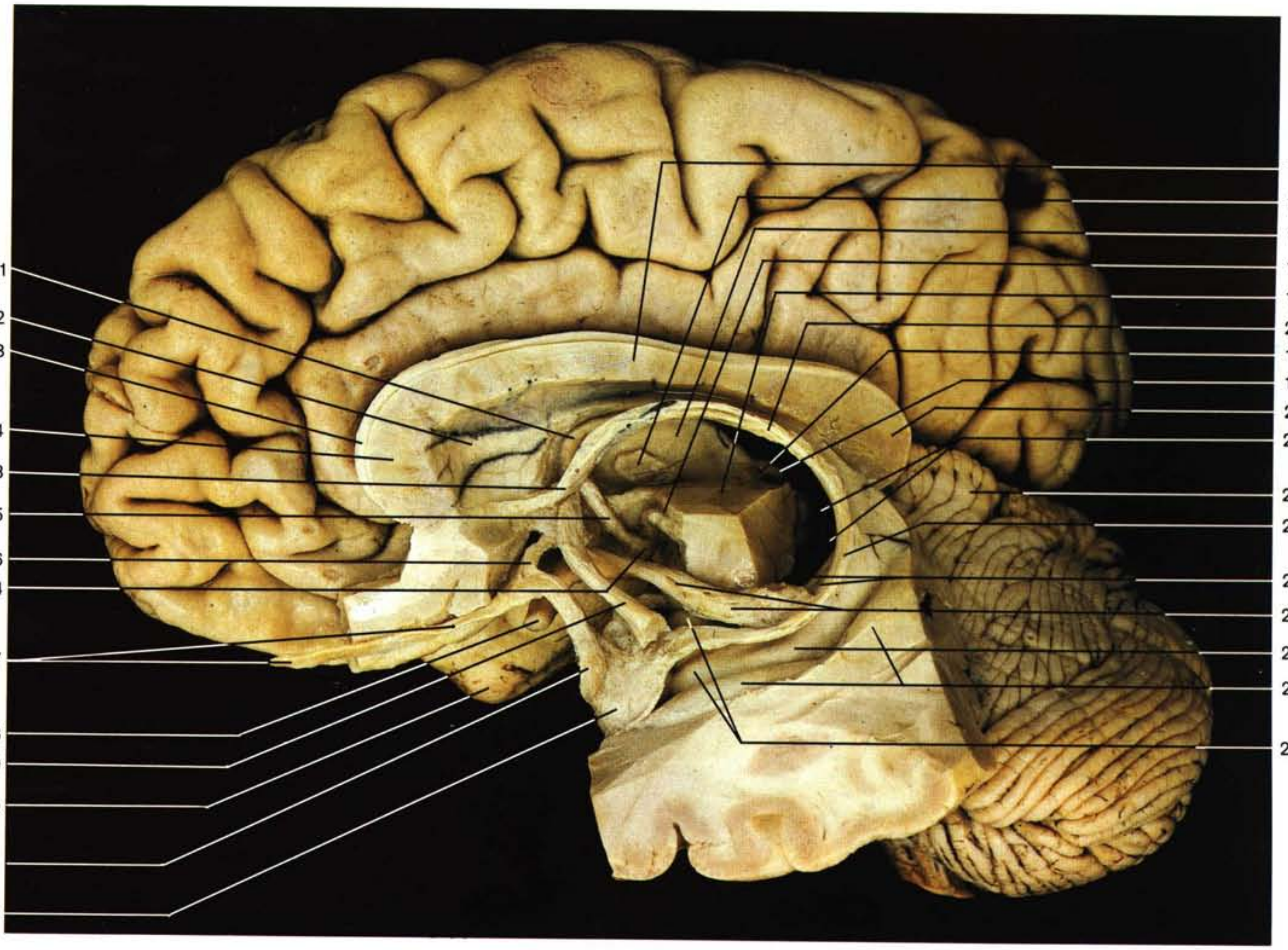

Dissection of the limbic system. Left side, lateral aspect. Corpus callosum has been cut in the median plane. The left thalamus and the left hemisphere have been partly removed.

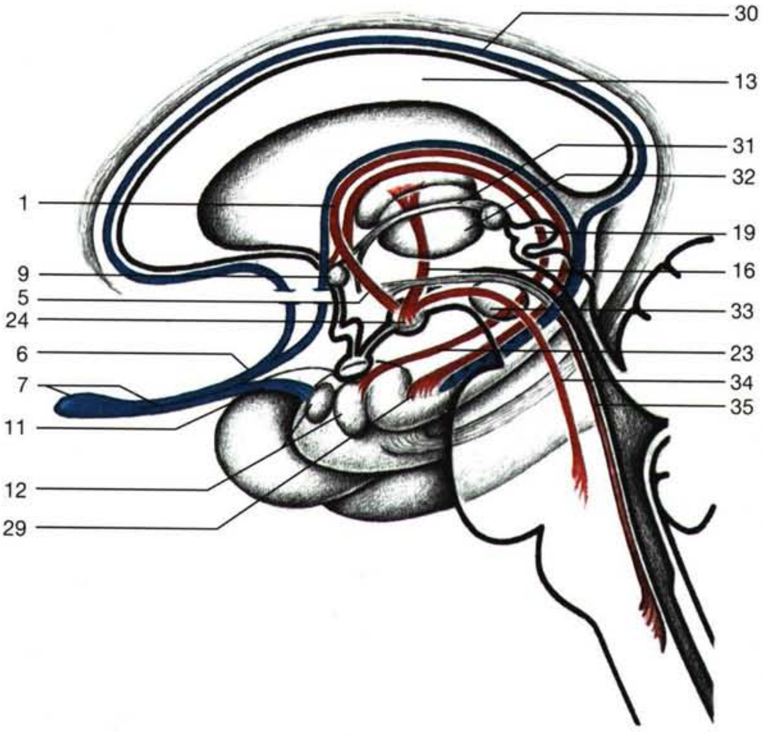

Main pathways of **limbic and olfactory system.** (Schematic drawing.) Blue = afferent pathways; red = efferent pathways.

1 **Body of fornix**
2 Septum pellucidum
3 **Lateral longitudinal stria**
4 Genu of corpus callosum
5 **Column of fornix**
6 Medial olfactory stria
7 **Olfactory bulb** and **olfactory tract**
8 Optic nerve
9 Anterior commissure (left half)
10 Right temporal lobe
11 **Lateral olfactory stria**
12 **Amygdala**
13 Body of **corpus callosum**
14 Interthalamic adhesion
15 Third ventricle and right thalamus
16 **Mamillothalamic tract**
17 Part of the thalamus
18 Habenular commissure
19 Pineal body
20 Splenium of corpus callosum
21 Colliculi of midbrain
22 Vermis of cerebellum
23 **Stria terminalis**
24 **Mamillary body**
25 Fimbria of hippocampus and pes hippocampi
26 Left optic tract and lateral geniculate body
27 Lateral ventricle and **parahippocampal gyrus**
28 Collateral eminence
29 Hippocampal digitations
30 **Supracallosal gyrus** (longitudinal stria)
31 Stria medullaris thalami
32 Thalamus
33 Red nucleus
34 Mamillotegmental tract
35 Dorsal longitudinal fasciculus (Schütz)

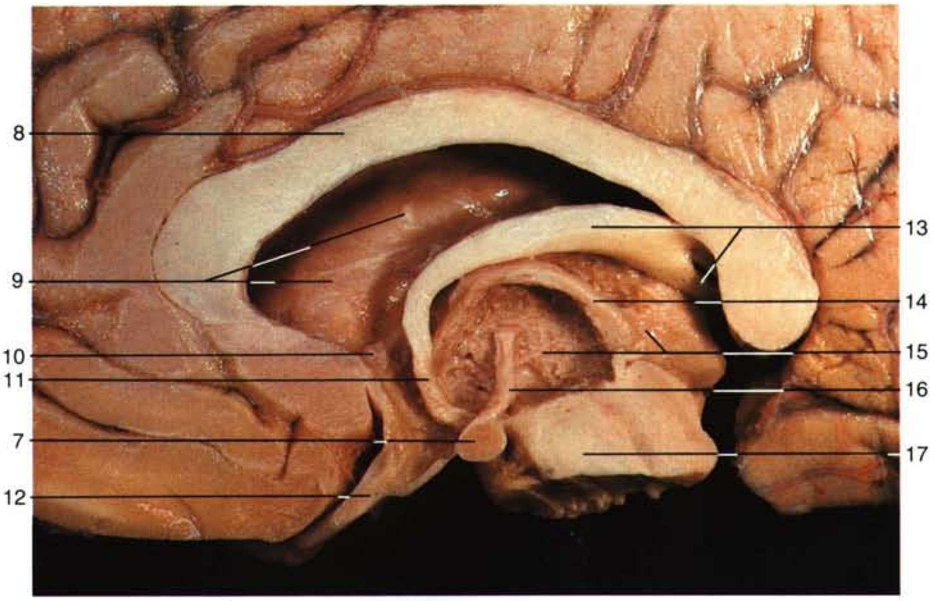

Median section through the diencephalon. Medial part of the thalamus and septum pellucidum have been removed to show the fornix and mamillothalamic tract.

1 Paraventricular nucleus
2 Preoptic nucleus
3 Ventromedial nucleus
4 Supraoptic nucleus
5 Posterior nucleus
6 Dorsomedial nucleus
} **Hypothalamic nuclei**
7 **Mamillary body**
8 **Corpus callosum**
9 Lateral ventricle (showing caudate nucleus)
10 Anterior commissure
11 **Column of fornix**
12 Optic chiasma
13 **Crus of fornix**
14 Medullary stria of thalamus
15 Thalamus and interthalamic adhesion
16 **Mamillothalamic tract** of Vicq d'Azyr
17 Cerebral peduncle
18 **Pineal body**
19 Tectum of midbrain
20 Lamina terminalis

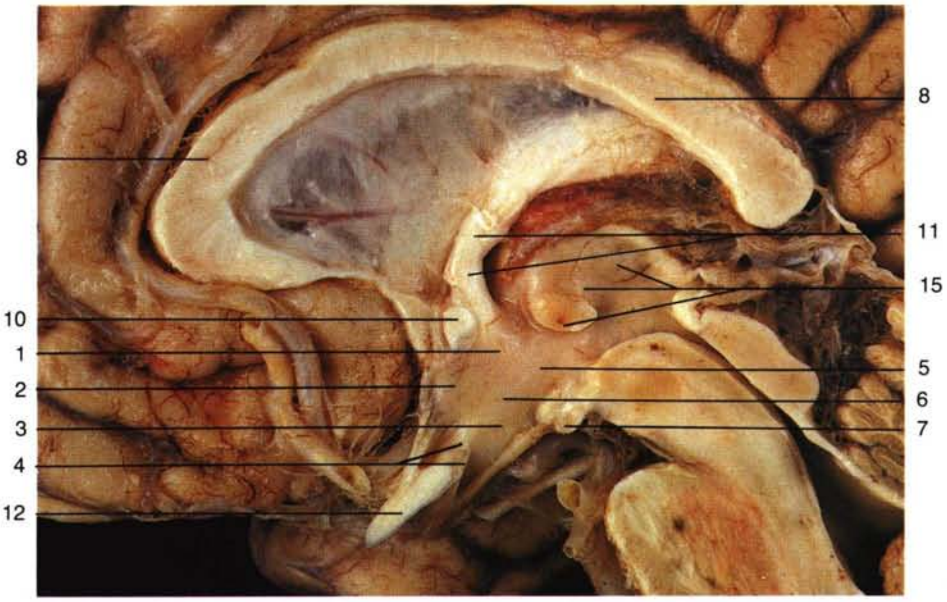

Median section through the diencephalon and midbrain; location of hypothalamic nuclei.

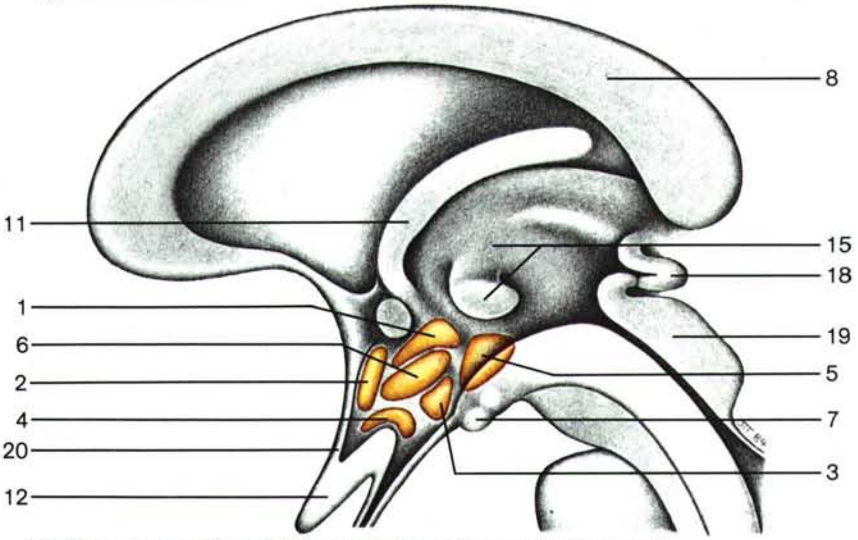

Position of main hypothalamic nuclei. (Schematic diagram.)

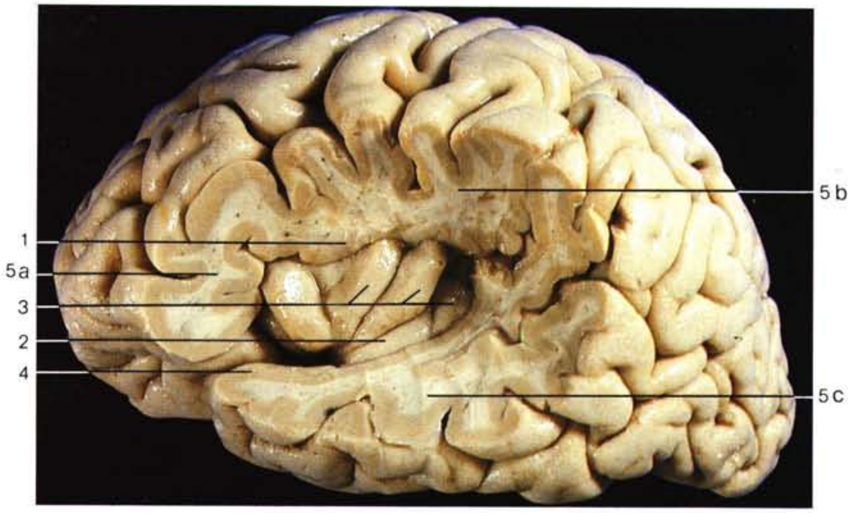

Insula (Reili). The opercula of the frontal, parietal and temporal lobes have been removed to display the insular gyri. Left hemisphere.

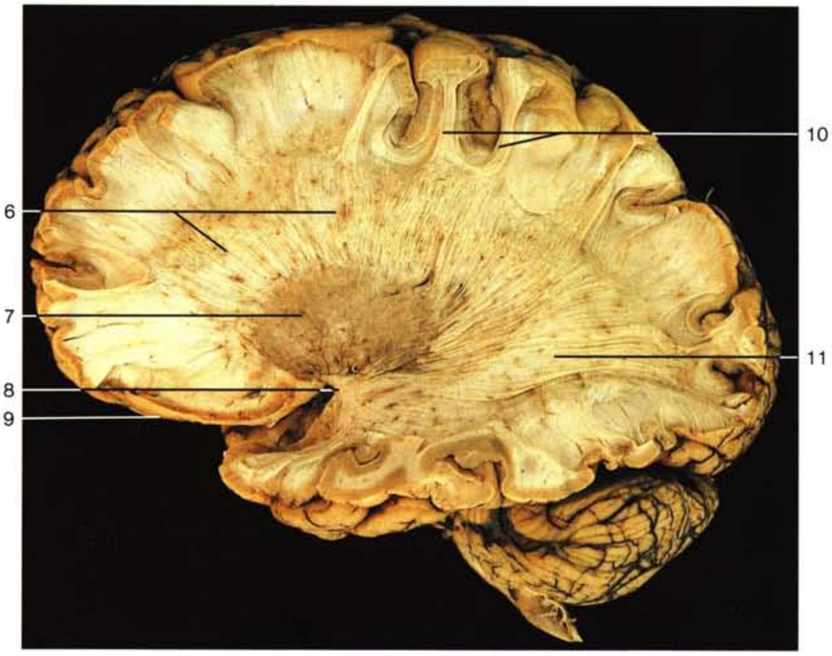

Dissection of the corona radiata, left hemisphere. Frontal pole on the left.

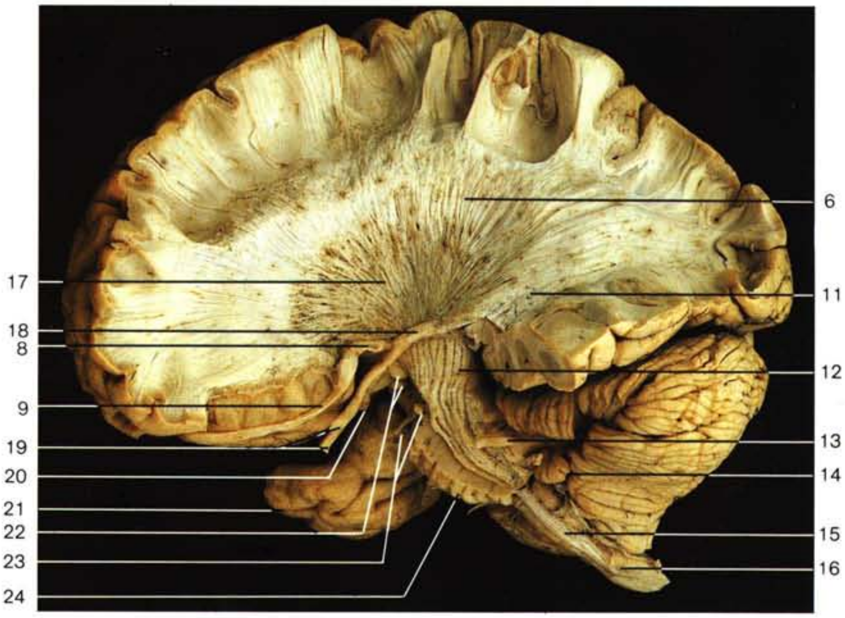

◁ **Corona radiata and internal capsule,** left hemisphere. Lentiform nucleus removed (frontal pole to the left).

1 Circular sulcus of insula
2 Long gyrus of insula
3 Short gyri of insula
4 Limen insulae
5 Opercula (cut)
  a Frontal operculum
  b Frontoparietal operculum
  c Temporal operculum
6 **Corona radiata**
7 **Lentiform nucleus**
8 **Anterior commissure**
9 Olfactory tract
10 Cerebral arcuate fibers
11 Optic radiation
12 Cerebral peduncle
13 Trigeminal nerve (n. V)
14 Flocculus of cerebellum
15 **Pyramidal tract**
16 Decussation of pyramidal tract
17 Internal capsule
18 Optic tract
19 Optic nerve (n. II)
20 Infundibulum
21 Temporal lobe (right side)
22 Mamillary bodies
23 Oculomotor nerve (n. III)
24 Transverse fibers of pons

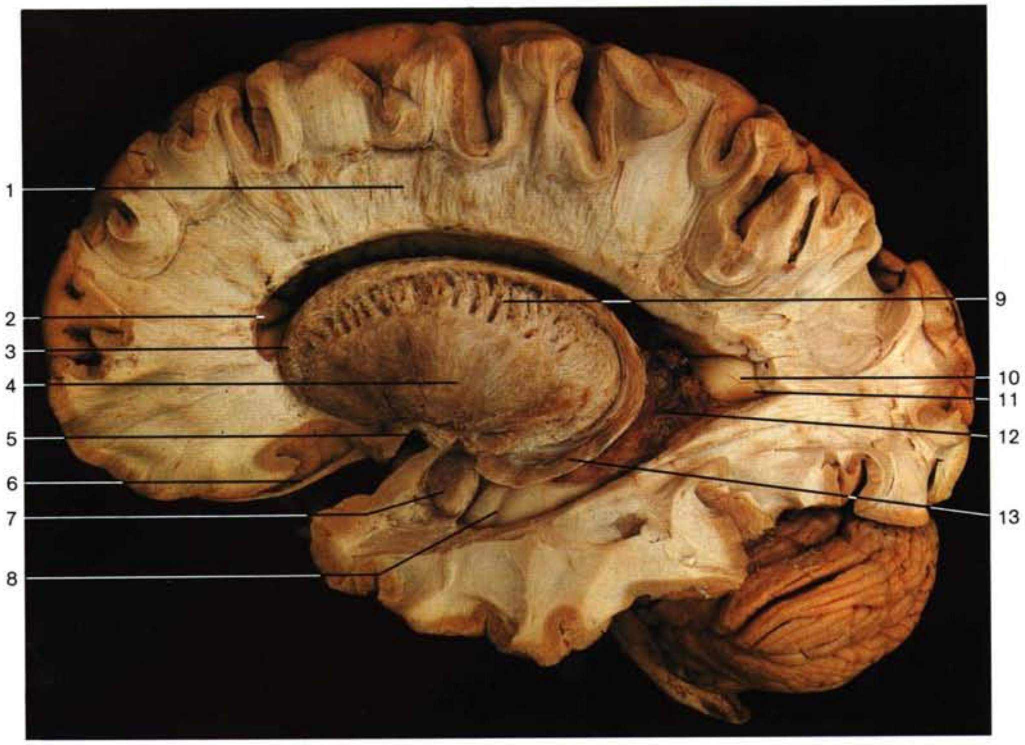

1 Corona radiata
2 Anterior horn of lateral ventricle
3 Head of **caudate nucleus**
4 **Putamen**
5 Anterior commissure
6 Olfactory tract
7 **Amygdala**
8 Hippocampal digitations
9 **Internal capsule**
10 Calcar avis
11 Posterior horn of lateral ventricle
12 Choroid plexus of lateral ventricle
13 Caudal extremity of caudate nucleus
14 **Thalamus**
15 Cerebral arcuate fibers
16 Globus pallidus (remnants)

Dissection of the subcortical nuclei and internal capsule, left hemisphere (lateral aspect). Frontal pole to the left. The lateral ventricle has been opened, and the insular gyri and claustrum have been removed, revealing the lentiform nucleus and the internal capsule.

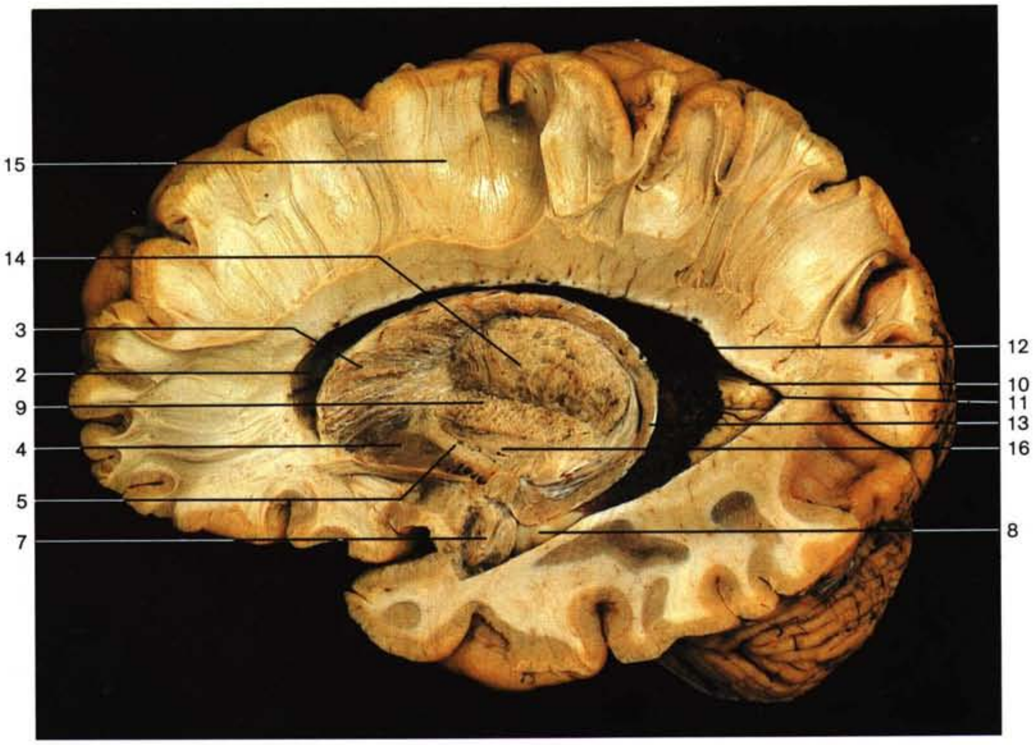

Dissection of the subcortical nuclei (lateral aspect). Lentiform nucleus removed, frontal pole to the left.

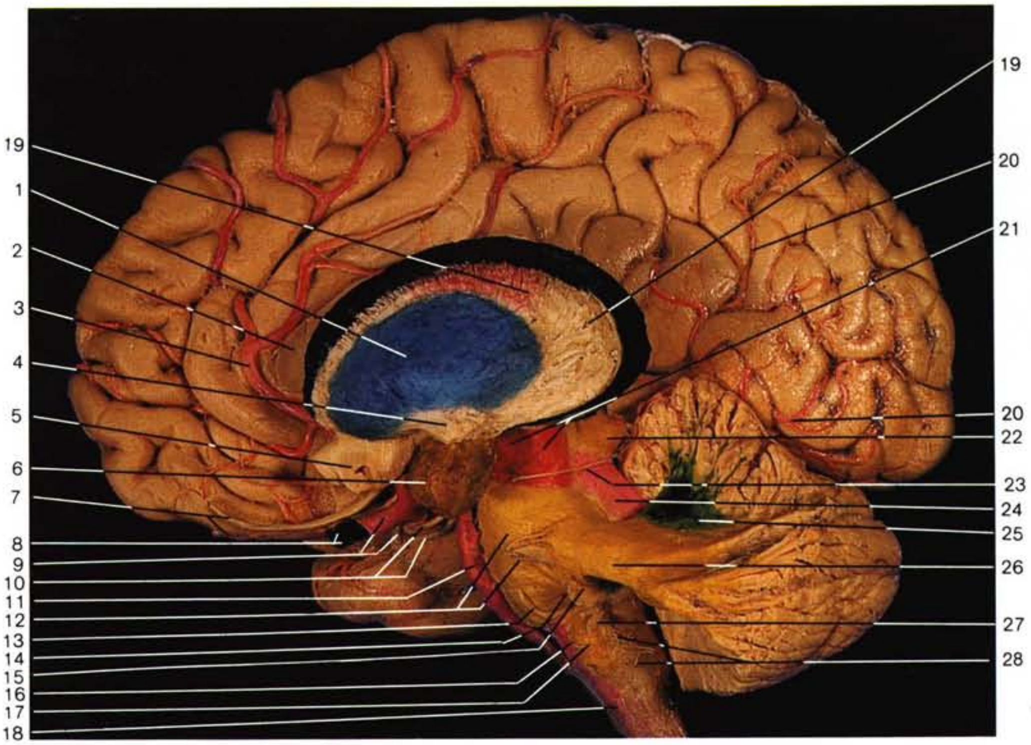

Brain stem and the connections with the cerebellum. Internal capsule (lateral aspect). Red = pyramidal tract; yellow = middle cerebellar peduncle; green = inferior cerebellar peduncle; pink = superior cerebellar peduncle.

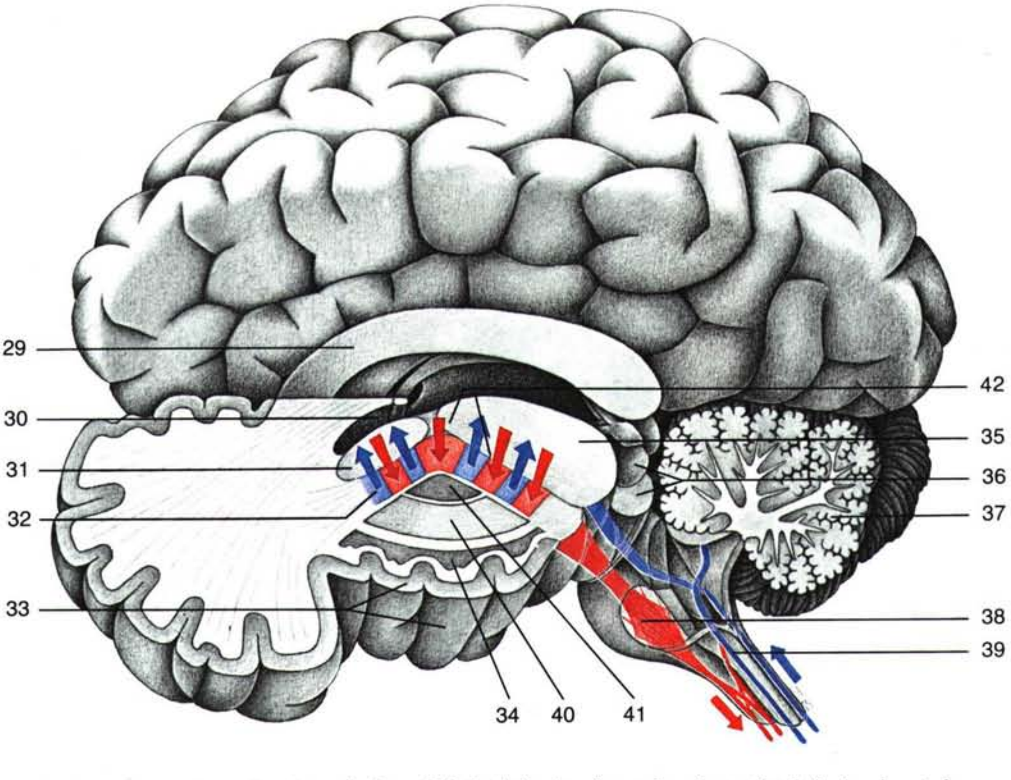

Internal capsule and subcortical nuclei. Left brain, frontal pole to the left, horizontal section. (Semischematic drawing.) Blue = afferent tracts; red = efferent tracts.

1 **Putamen**
2 Genu of corpus callosum
3 Anterior cerebral artery
4 Anterior commissure
5 Subcallosal area
6 Amygdala
7 Olfactory tract
8 Optic nerve (n. II)
9 Internal carotid artery and infundibulum
10 Oculomotor nerve (right and left nerve, n. III)
11 **Basilar artery**
12 Pons and trigeminal nerve (n. V)
13 Abducens nerve (n. VI)
14 Facial nerve (n. VII)
15 Vestibulocochlear nerve (n. VIII)
16 Hypoglossal nerve (n. XII)
17 Olive
18 Pyramidal tract
19 **Internal capsule**
20 Posterior cerebral artery
21 Cerebral peduncle
22 **Colliculi of midbrain**
23 Trochlear nerve (n. IV)
24 Superior cerebellar peduncle
25 Inferior cerebellar peduncle
26 **Middle cerebellar peduncle**
27 Glossopharyngeal nerve (n. IX)
28 Vagus (n. X) and accessory nerves (n. XI)
29 **Corpus callosum**
30 Lateral ventricle (anterior horn)
31 Caudate nucleus
32 Internal capsule (anterior limb)
33 Insula
34 **Claustrum**
35 **Thalamus**
36 Superior and inferior colliculus of midbrain
37 Cerebellum
38 Middle peduncle of cerebellum (efferent tracts)
39 Medulla oblongata (afferent tracts)
40 Putamen of lentiform nucleus
41 **Globus pallidus** of lentiform nucleus
42 Genu and posterior limb of internal capsule

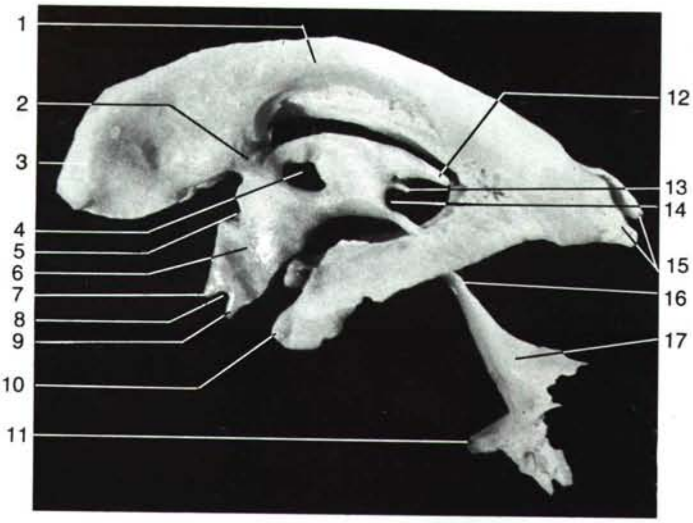

Cast of ventricular cavities of the brain (lateral aspect), frontal pole to the left.

1 Central part of the **lateral ventricle**
2 **Interventricular foramen** of Monro
3 Anterior horn of the lateral ventricle
4 Site of interthalamic adhesion
5 Notch for anterior commissure
6 **Third ventricle**
7 Optic recess
8 Notch for optic chiasma
9 Infundibular recess
10 Inferior horn of lateral ventricle with indentation of amygdaloid body
11 Lateral recess and lateral aperture of Luschka
12 Suprapineal recess
13 Pineal recess
14 Notch for posterior commissure
15 Posterior horn of lateral ventricle
16 **Cerebral aqueduct**
17 **Fourth ventricle**
18 Median aperture of Magendie
19 **Cerebellomedullary cistern**
20 Superior sagittal sinus
21 Inferior sagittal sinus
22 Intervaginal space of optic nerve
23 Arachnoid granulations of Pacchioni
24 Straight sinus

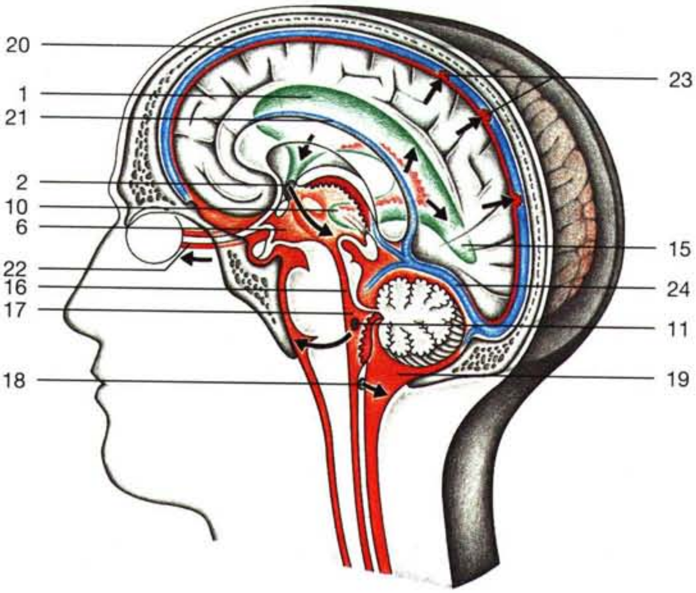

◁ **Position of ventricular cavities.** (Schematic drawing.) The direction of flow of cerebrospinal fluid is indicated by arrows.
Green = Right lateral ventricle
Red = Choroidal plexus

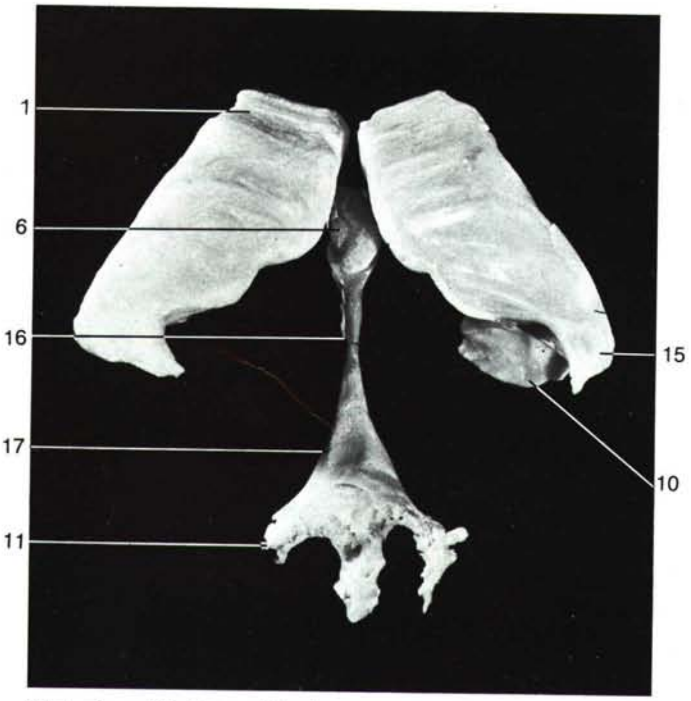

Cast of ventricular cavities (posterior aspect).

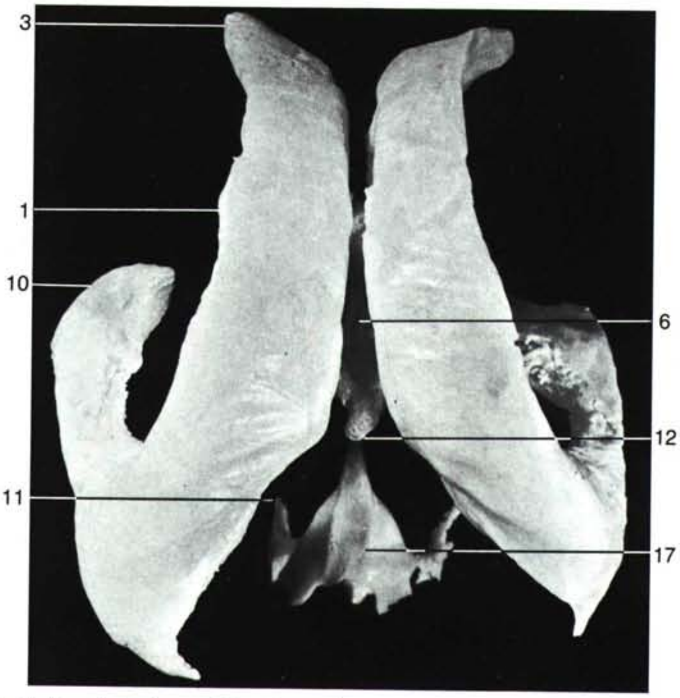

Cast of ventricular cavities (superior aspect). Frontal pole at top.

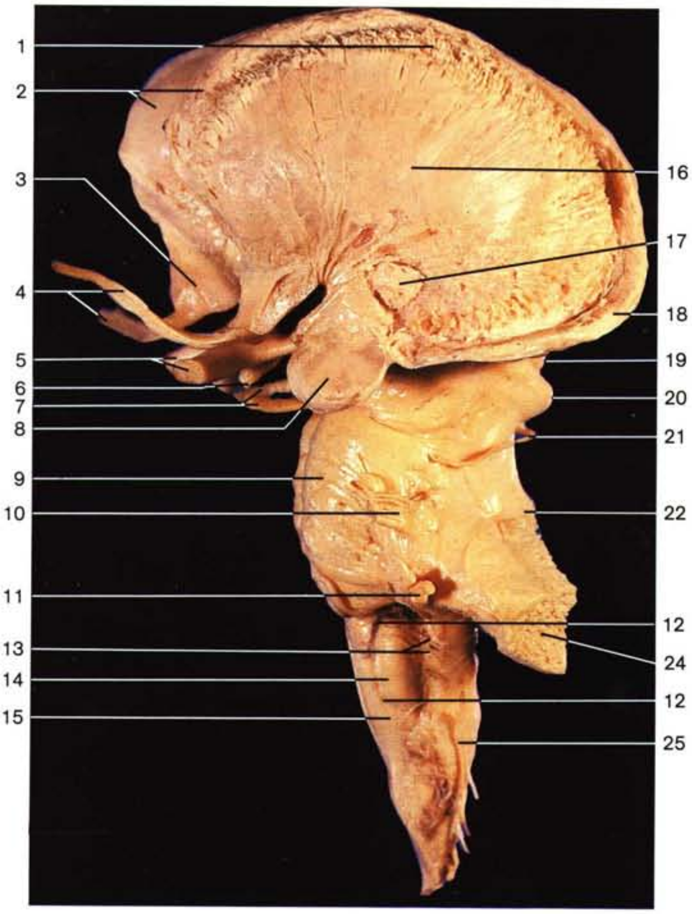

Brain stem (left lateral aspect). Cerebellar peduncles have been severed, cerebellum and cerebral cortex have been removed.

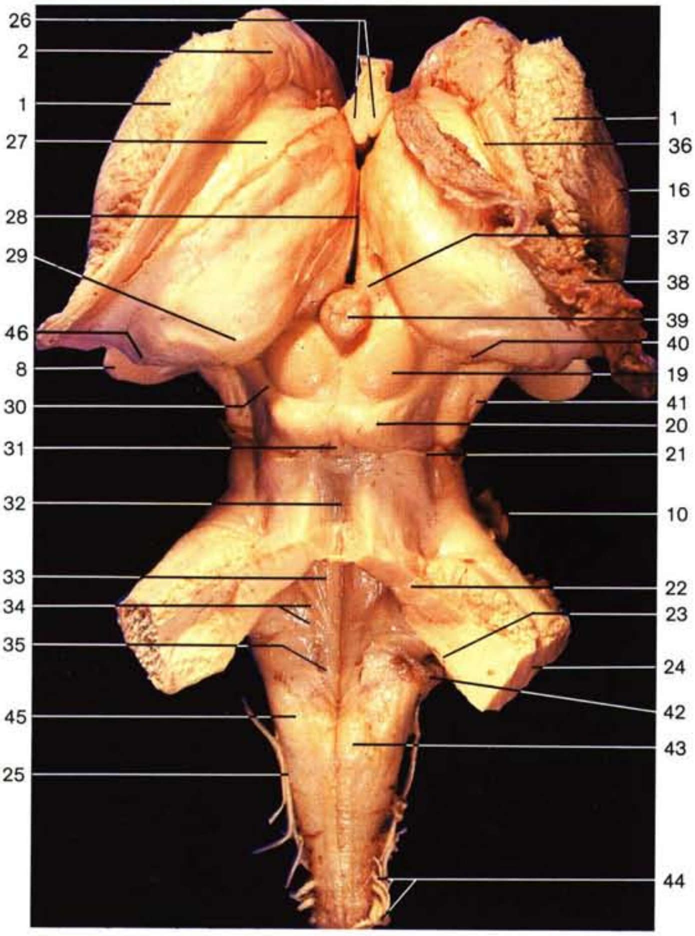

Brain stem (dorsal aspect). Cerebellum removed.

1	**Internal capsule**
2	Head of the caudate nucleus
3	Olfactory trigone
4	Olfactory tracts
5	Optic nerves
6	Infundibulum
7	Oculomotor nerve
8	**Amygdaloid body**
9	**Pons**
10	Trigeminal nerve
11	Facial and vestibulocochlear nerves
12	Hypoglossal nerve
13	Glossopharyngeal and vagus nerves
14	**Olive**
15	Medulla oblongata
16	**Lentiform nucleus**
17	**Anterior commissure**
18	Tail of caudate nucleus
19	Superior colliculus
20	Inferior colliculus
21	Trochlear nerve
22	Superior cerebellar peduncle
23	Inferior cerebellar peduncle
24	**Middle cerebellar peduncle**
25	Accessory nerve (n. XI)
26	Columns of fornix (divided)
27	Lamina affixa
28	**Third ventricle**
29	Pulvinar of thalamus
30	Inferior brachium
31	Frenulum veli
32	Superior medullary velum
33	Facial colliculus
34	Striae medullares and rhomboid fossa
35	Hypoglossal triangle
36	Stria terminalis and thalamostriate vein
37	Habenular trigone
38	Choroid plexus of lateral ventricle
39	**Pineal body**
40	Medial geniculate body
41	Cerebral peduncle
42	Choroid plexus of fourth ventricle
43	Clava
44	Dorsal root of cervical nerve
45	Cuneate tubercle
46	Lateral geniculate body

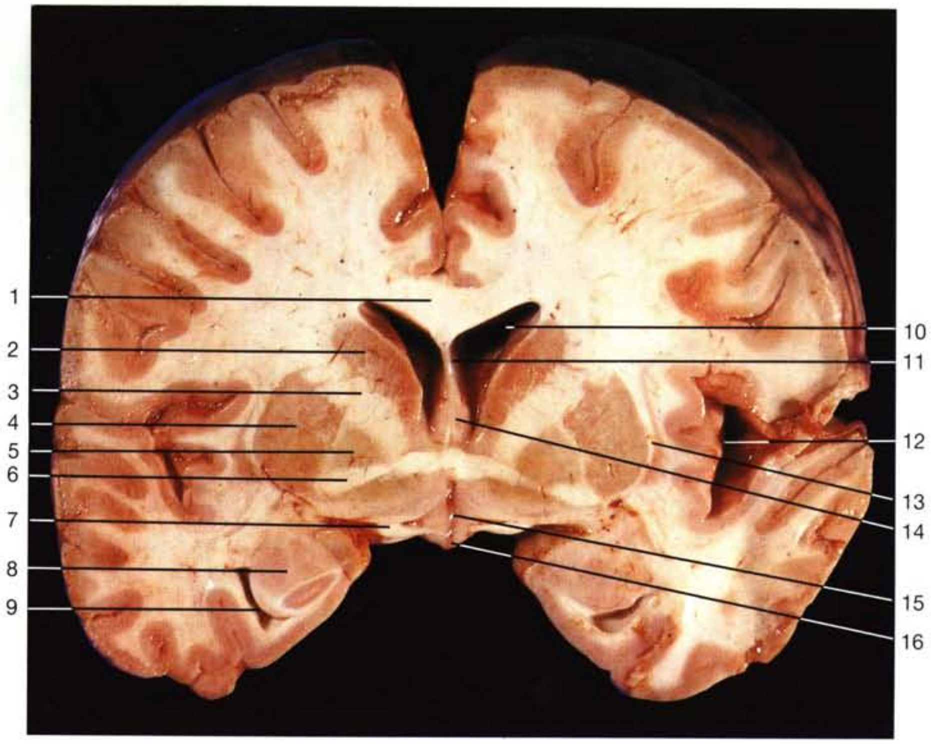

Coronal section through the brain at the level of the anterior commissure. Section 1.

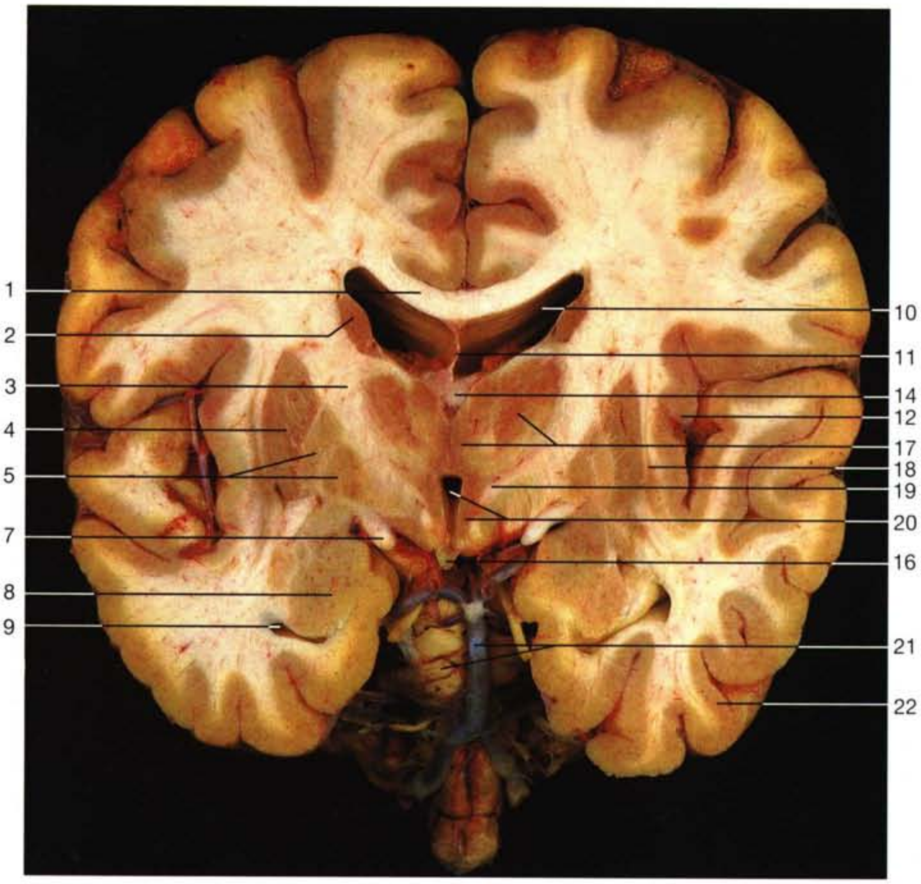

Coronal section through the brain at the level of the third ventricle and the interthalamic adhesion. Section 2.

1 Corpus callosum
2 Head of **caudate nucleus**
3 **Internal capsule**
4 **Putamen**
5 **Globus pallidus**
6 Anterior commissure
7 Optic tract
8 Amygdaloid body
9 Inferior horn of lateral ventricle
10 **Lateral ventricle**
11 Septum pellucidum
12 Lobus insularis (insula)
13 External capsule
14 Column of fornix
15 Optic recess
16 Infundibulum
17 **Thalamus**
18 Claustrum
19 Lenticular ansa
20 **Third ventricle** and **hypothalamus**
21 Basilar artery and pons
22 Cortex of temporal lobe
23 Inferior colliculus
24 Superior colliculus
25 Cerebral aqueduct
26 Red nucleus
27 Substantia nigra
28 Cerebral peduncle
29 Trochlear nerve (n. IV)
30 Gray matter
31 Nucleus of oculomotor nerve (nucleus of Edinger-Westphal)
32 Fibers of oculomotor nerve (n. III)
33 Vermis of cerebellum
34 Fourth ventricle
35 Reticular formation
36 Pons and transverse pontine fibers
37 Emboliform nucleus
38 Dentate nucleus
39 Middle cerebellar peduncle
40 Choroid plexus
41 Hypoglossal nucleus at rhomboid fossa
42 Medial longitudinal fasciculus
43 Trigeminal nerve (n. V.)
44 Inferior olivary nucleus
45 Corticospinal fibers and arcuate fibers
46 Fourth ventricle with choroid plexus
47 Vestibular nuclei
48 Nucleus and tractus solitarius
49 Inferior cerebellar peduncle (restiform body)
50 Reticular formation
51 Medial lemniscus
52 Cuneate nucleus of Burdach
53 Central canal
54 **Pyramidal tract**
55 Flocculus of cerebellum
56 Cerebellar hemisphere with pia mater
57 "Arbor vitae" of cerebellum
58 Nucleus gracilis of Goll
59 Lateral recess of choroid plexus of IVth ventricle
60 Posterior inferior cerebellar artery
61 Choroid plexus of lateral ventricle

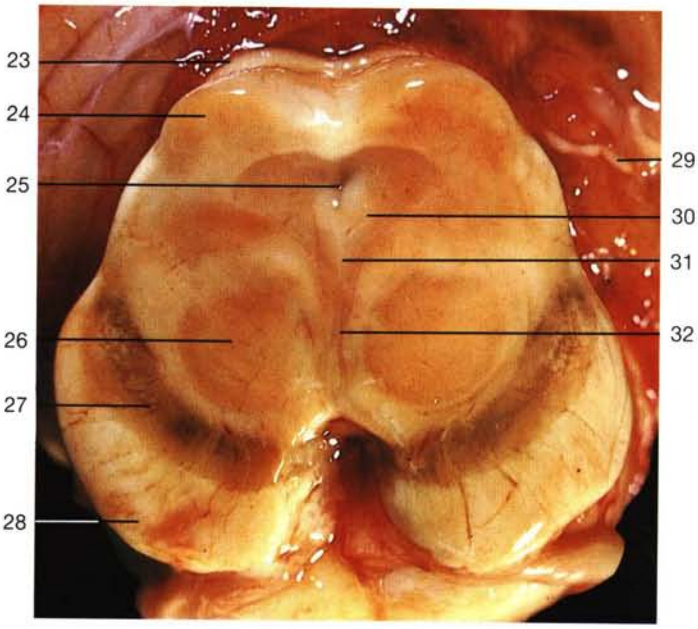
23
24
25
26
27
28
29
30
31
32

Cross-section of the midbrain (mesencephalon) at the level of the superior colliculus (superior aspect). Section 4.

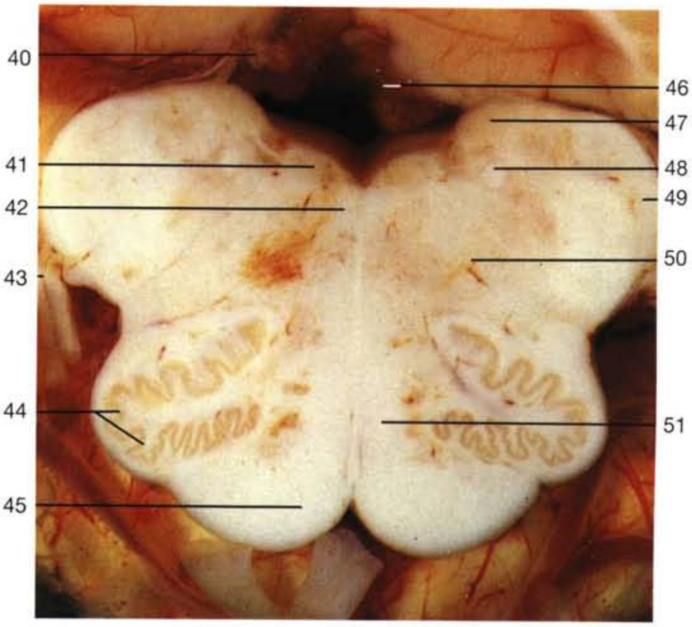
40
41
42
43
44
45
46
47
48
49
50
51

Cross-section of the rhombencephalon at the level of the olive (inferior aspect). Section 6.

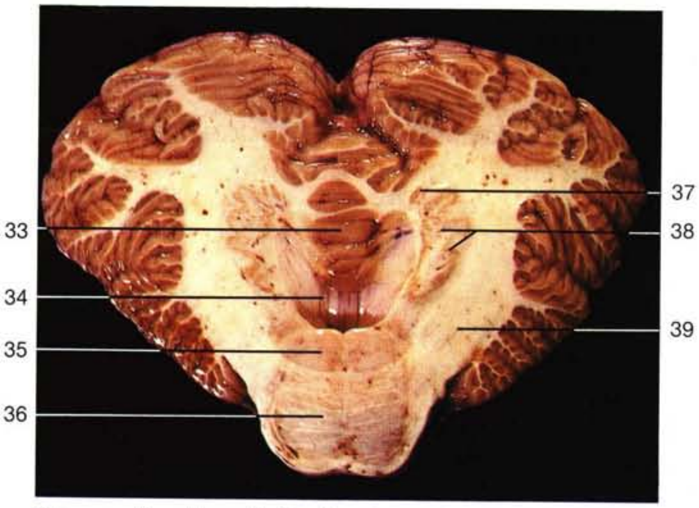
33
34
35
36
37
38
39

Cross-section through the rhombencephalon at the level of pons (inferior aspect). Section 5.

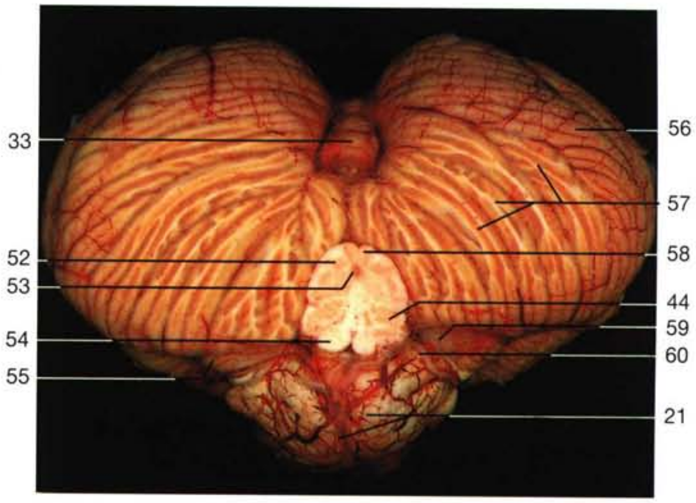
33
52
53
54
55
56
57
58
44
59
60
21

Cross-section through medulla oblongata and cerebellum (inferior aspect). Section 7.

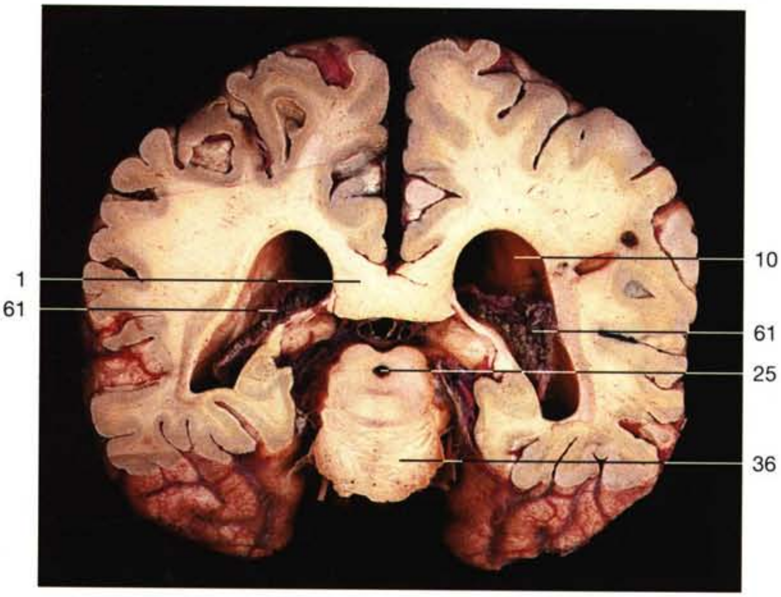
1
61
10
61
25
36

Coronal section at the level of **inferior colliculus** (posterior aspect). Section 3.

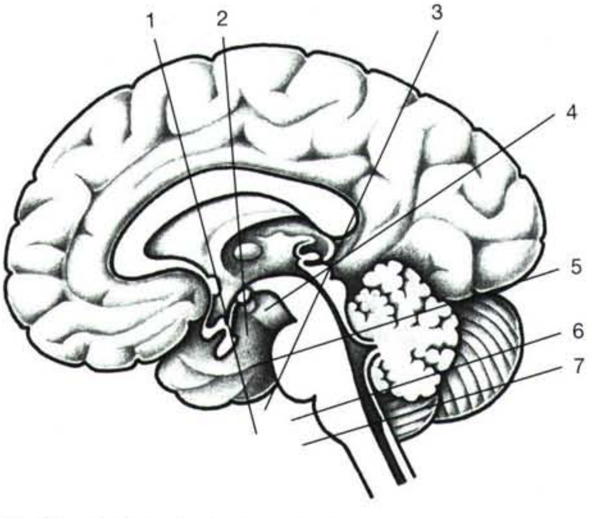

Right half of the brain. Levels of the sections are indicated.

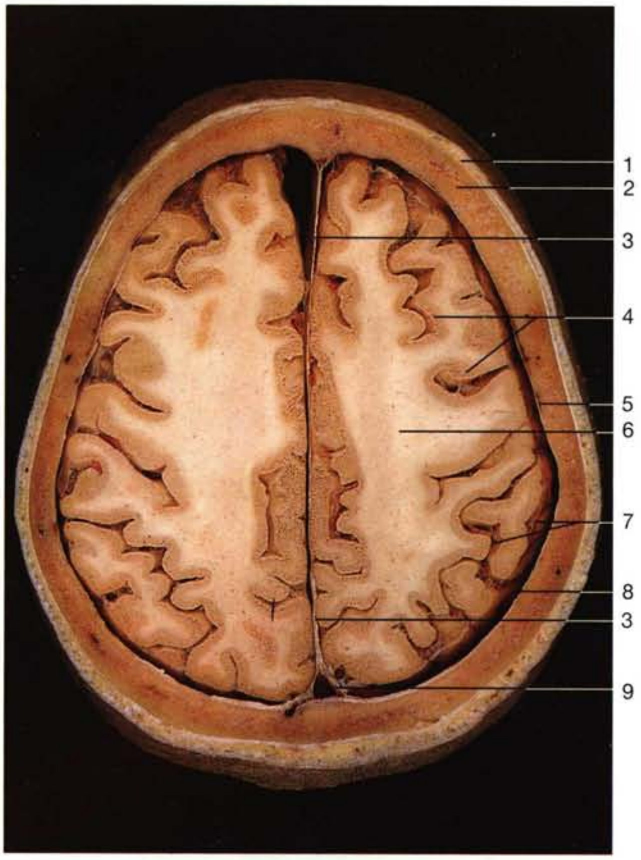

Horizontal section through the head.
Section 1.

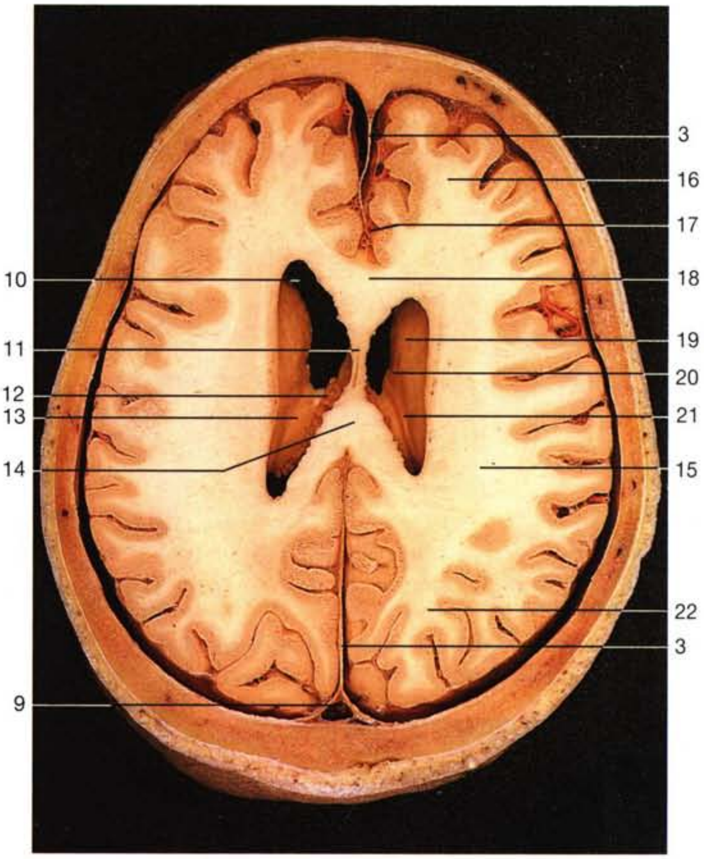

Horizontal section through the head.
Section 2.

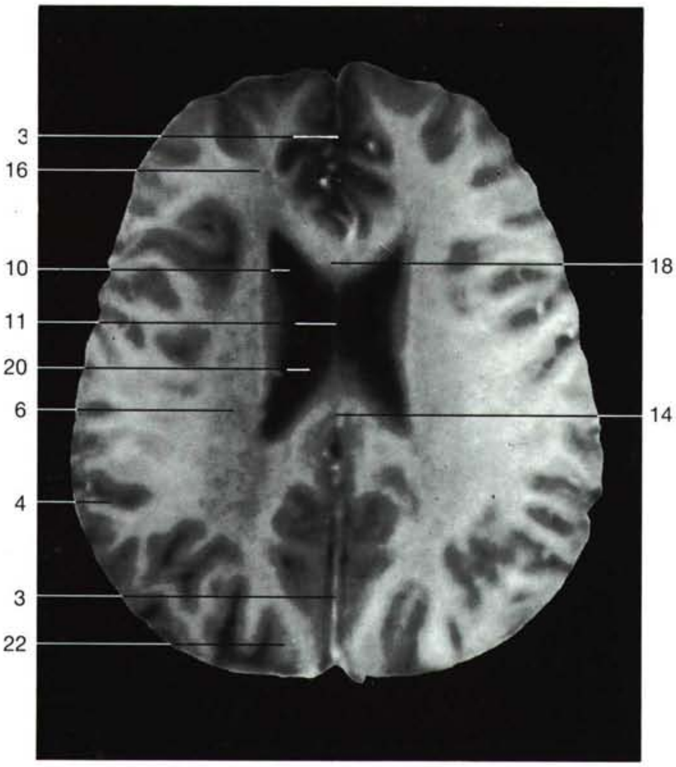

MR-Scan of the human head at the level of section 2.

1 Skin of scalp
2 Calvaria (diploe of the skull)
3 **Falx cerebri**
4 Gray matter of brain (cortex)
5 Dura mater
6 **White matter of brain**
7 Arachnoid and pia mater with vessels
8 Subdural space (slightly expanded due to shrinkage of the brain)
9 Superior sagittal sinus
10 Anterior horn of lateral ventricle
11 **Septum pellucidum**
12 Choroid plexus
13 Thalamus
14 Splenium of corpus callosum
15 Parietal lobe
16 Frontal lobe
17 Anterior cerebral artery
18 Genu of corpus callosum
19 Caudate nucleus
20 **Central part of lateral ventricle**
21 Stria terminalis
22 Occipital lobe

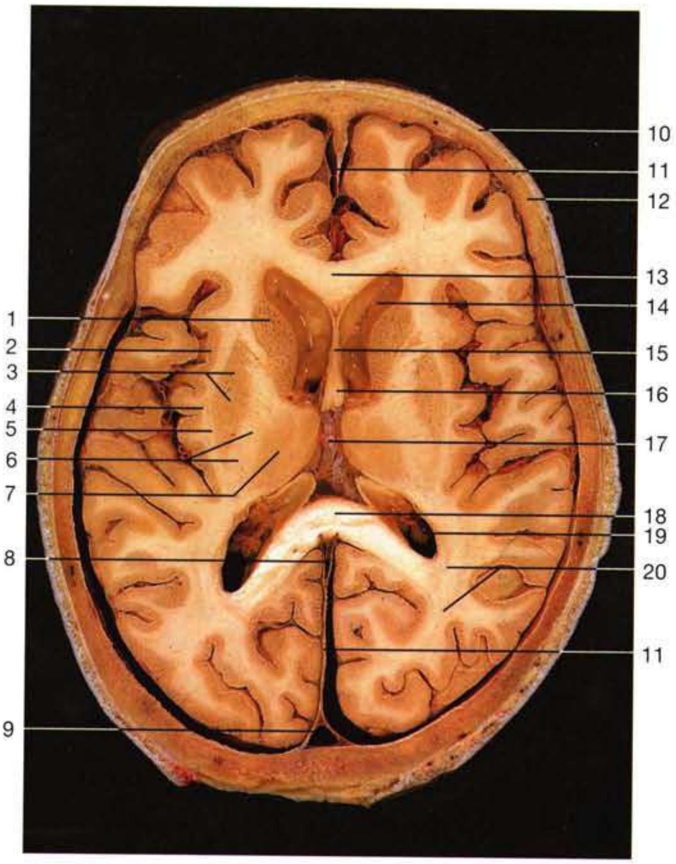

Horizontal section through the head at the level of third ventricle of internal capsule and neighbouring nuclei. Section 3.

1 **Caudate nucleus**
2 Lobus insularis (insula)
3 **Lentiform nucleus**
4 Claustrum
5 External capsule
6 Internal capsule
7 **Thalamus**
8 Straight sinus (sinus rectus)
9 Superior sagittal sinus
10 Skin of scalp
11 Falx cerebri
12 Calvaria (diploe of skull)
13 Genu of corpus callosum
14 Anterior horn of lateral ventricle
15 **Septum pellucidum**
16 Column of fornix
17 Choroid plexus of third ventricle
18 Splenium of corpus callosum
19 Entrance to inferior horn of lateral ventricle with choroid plexus
20 Optic radiation
21 **Third ventricle**

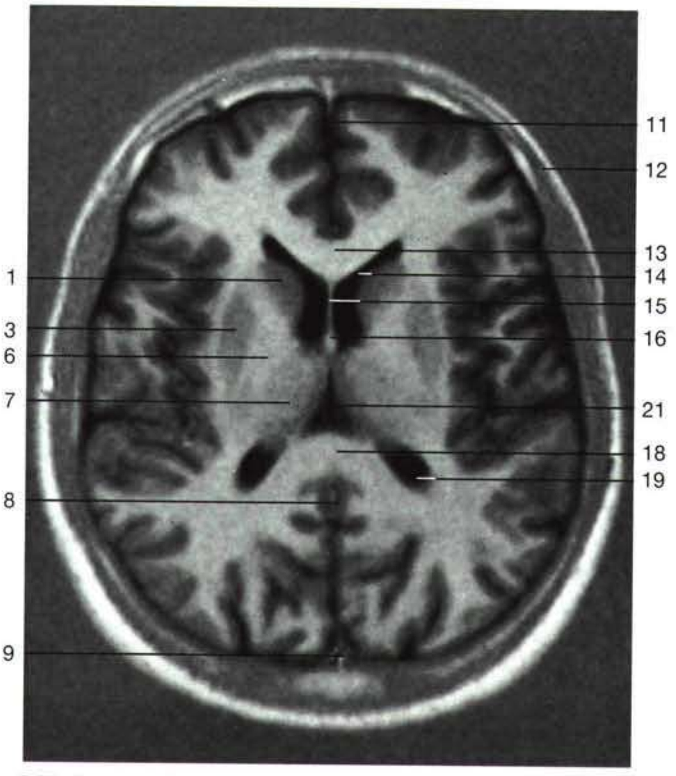

MR-Scan at the corresponding level to the above figure. Section 3.

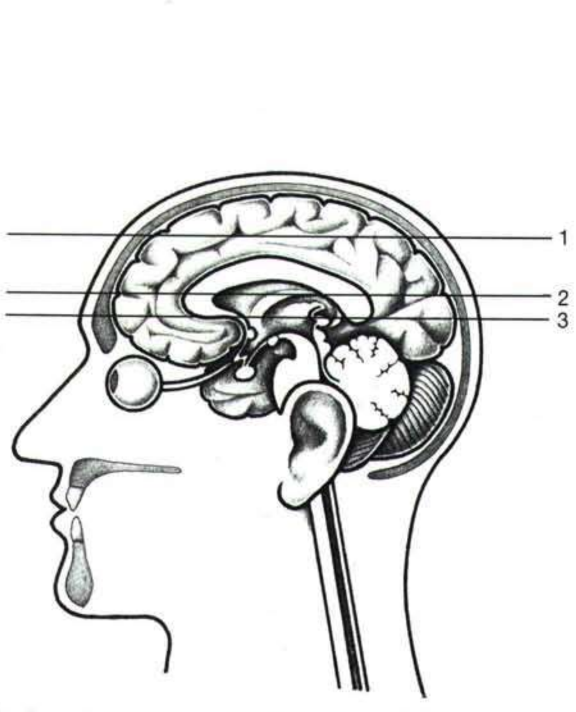

Sagittal section through the head. Levels of the horizontal sections are indicated.

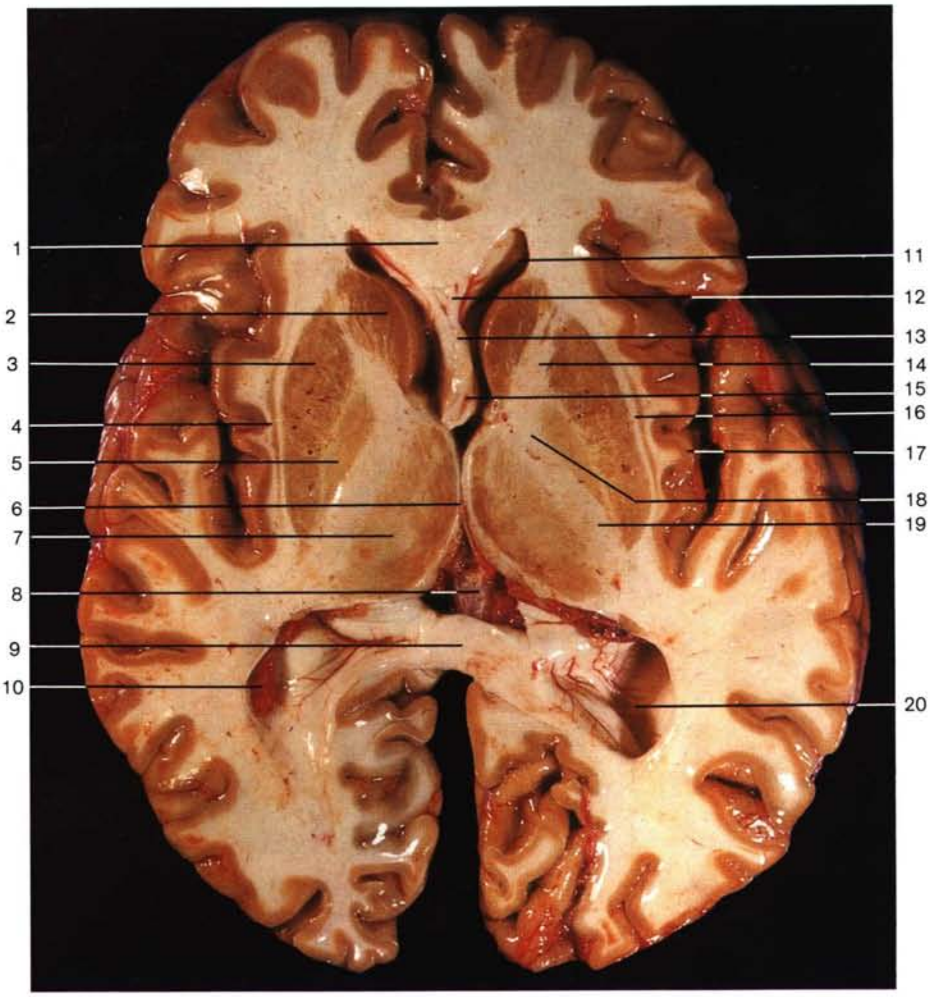

Horizontal section through the brain, showing the subcortical nuclei and internal capsule. Section 1.

1 Genu of corpus callosum
2 **Head of caudate nucleus**
3 **Putamen**
4 **Claustrum**
5 **Globus pallidus**
6 Third ventricle
7 Thalamus
8 **Pineal body**
9 Splenium of corpus callosum
10 Choroid plexus of the lateral ventricle
11 Anterior horn of lateral ventricle
12 Cavity of septum pellucidum
13 **Septum pellucidum**
14 Anterior limb of internal capsule
15 Column of fornix
16 External capsule
17 **Lobus insularis** (insula)
18 Genu of **internal capsule**
19 Posterior limb of internal capsule
20 Posterior horn of lateral ventricle
21 Anterior commissure
22 Optic radiation
23 Falx cerebri
24 **Maxillary sinus**
25 Position of auditory tube
26 Tympanic cavity
27 External acoustic meatus
28 **Medulla oblongata**
29 Fourth ventricle
30 Cerebellum (left hemisphere)
31 Temporomandibular joint
32 Tympanic membrane
33 Base of cochlea
34 Mastoid air cells
35 Sigmoid sinus
36 Vermis of cerebellum
37 Intermediate mass

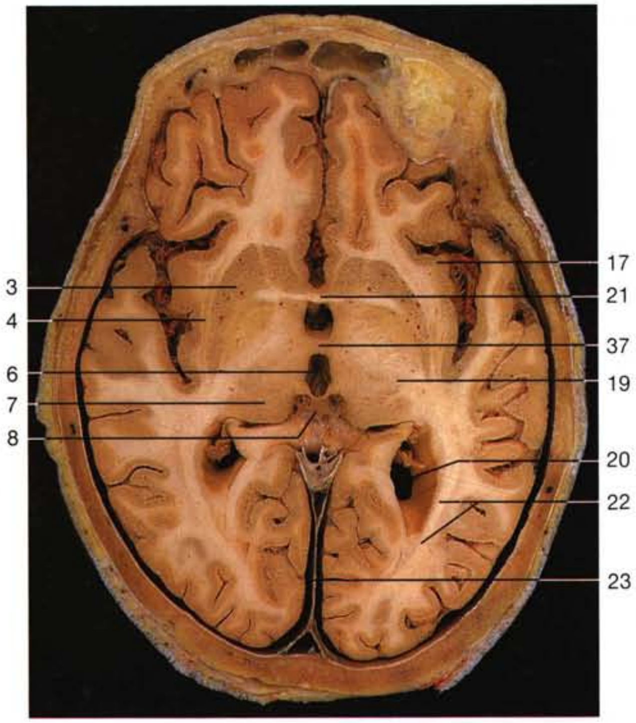

Horizontal section through the head. Section 2.

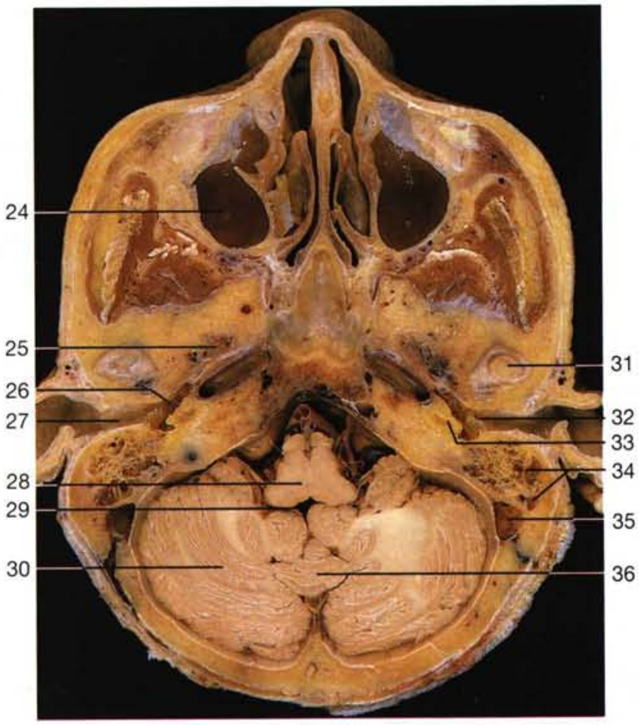

Horizontal section through the head. Section 4.

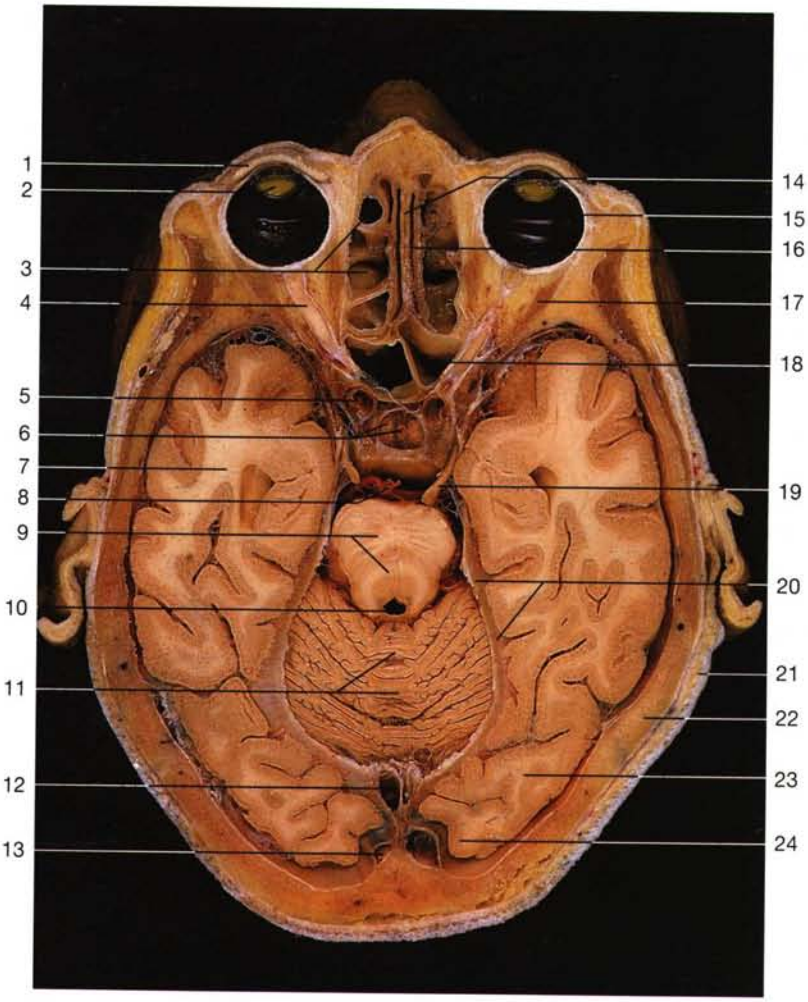

Horizontal section through the head. Section 3.

1 Upper lid (tarsal plate)
2 Lens
3 Ethmoidal sinus
4 **Optic nerve** (n. II)
5 Internal carotid artery
6 Infundibulum and **pituitary gland**
7 Temporal lobe
8 Basilar artery
9 Pons (cross section of brain stem)
10 **Cerebral aqueduct** (beginning of fourth ventricle)
11 Vermis of cerebellum
12 Straight sinus
13 Transverse sinus
14 Nasal septum
15 **Eyeball** (sclera)
16 Nasal cavity
17 Lateral rectus muscle
18 Sphenoidal sinus
19 Oculomotor nerve (n. III)
20 **Tentorium of cerebellum**
21 Skin of scalp
22 Calvaria
23 Occipital lobe
24 Striate cortex (visual cortex)

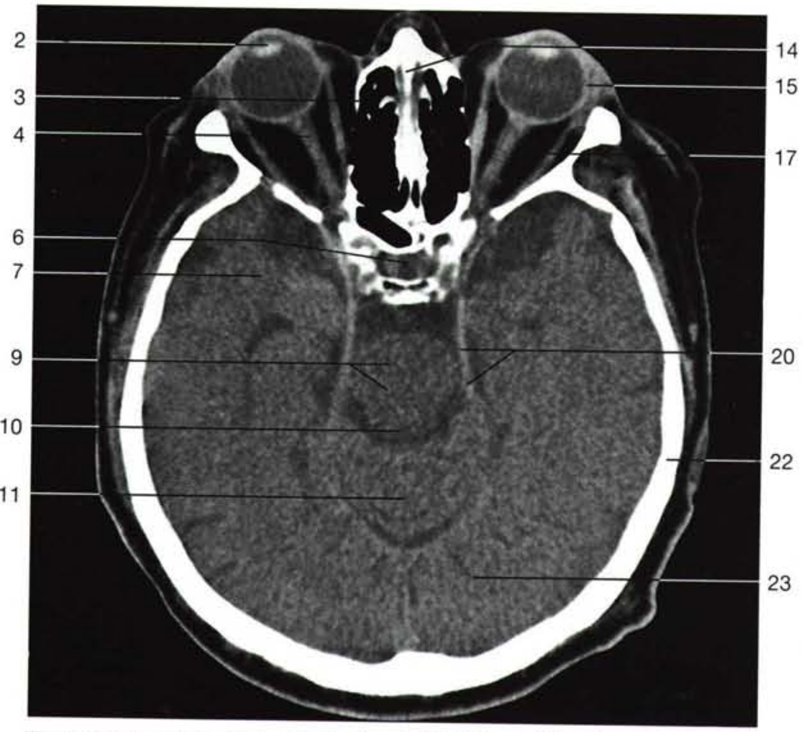

Horizontal section through the head. (CT-Scan.) Section 3.

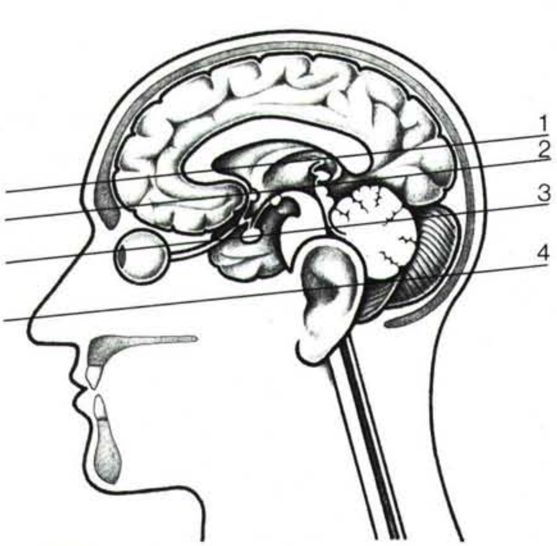

Sagittal section through the head.
Levels of the horizontal sections are indicated.

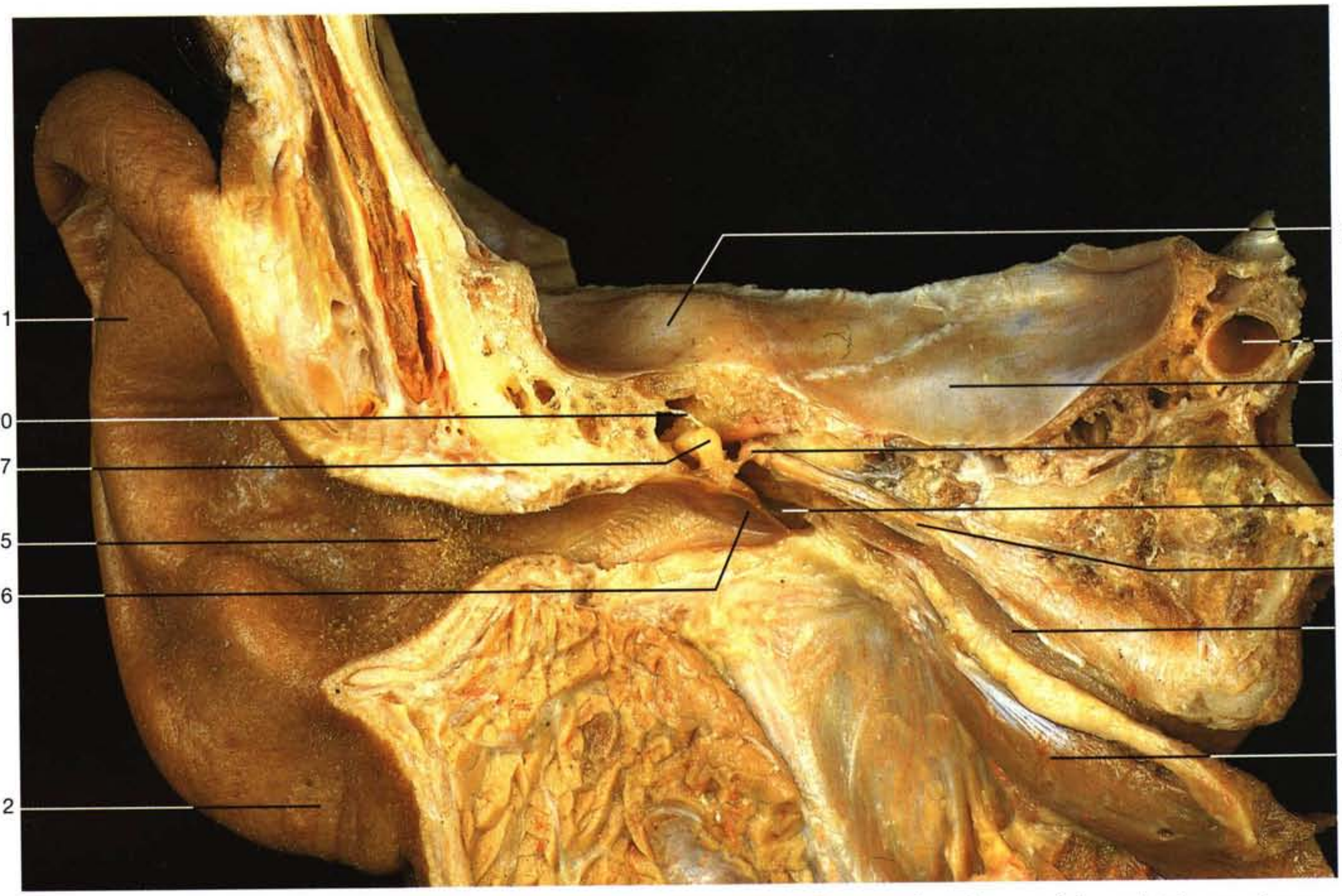

Longitudinal section through the right temporal bone I. The outer and middle ear and auditory ossicles and tube are shown (anterior aspect).

Outer ear

1 Auricle
2 Lobule of auricle
3 Helix
4 Tragus
5 External acoustic meatus

Middle ear

6 Tympanic membrane
7 Malleus
8 Incus
9 Stapes
10 Tympanic cavity
11 Mastoid process
12 Auditory tube
13 Tensor tympani muscle

Inner ear

14 Anterior semicircular duct
15 Posterior semicircular duct
16 Lateral semicircular duct
17 Cochlea
18 Vestibulocochlear nerve
19 Petrous part of the temporal bone

Additional structures

20 Superior ligament of malleus
21 Arcuate eminence
22 Internal carotid artery
23 Anterior surface of pyramid with dura mater
24 Stapes
25 Levator veli palatini muscle

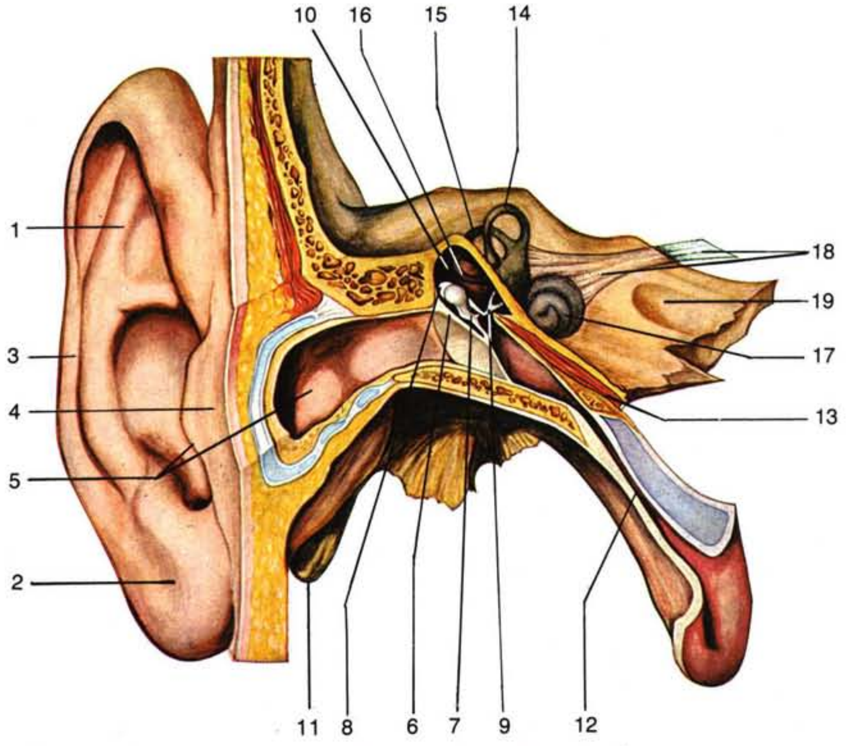

Right **auditory and vestibular apparatus** (anterior aspect). (Schematic drawing.)

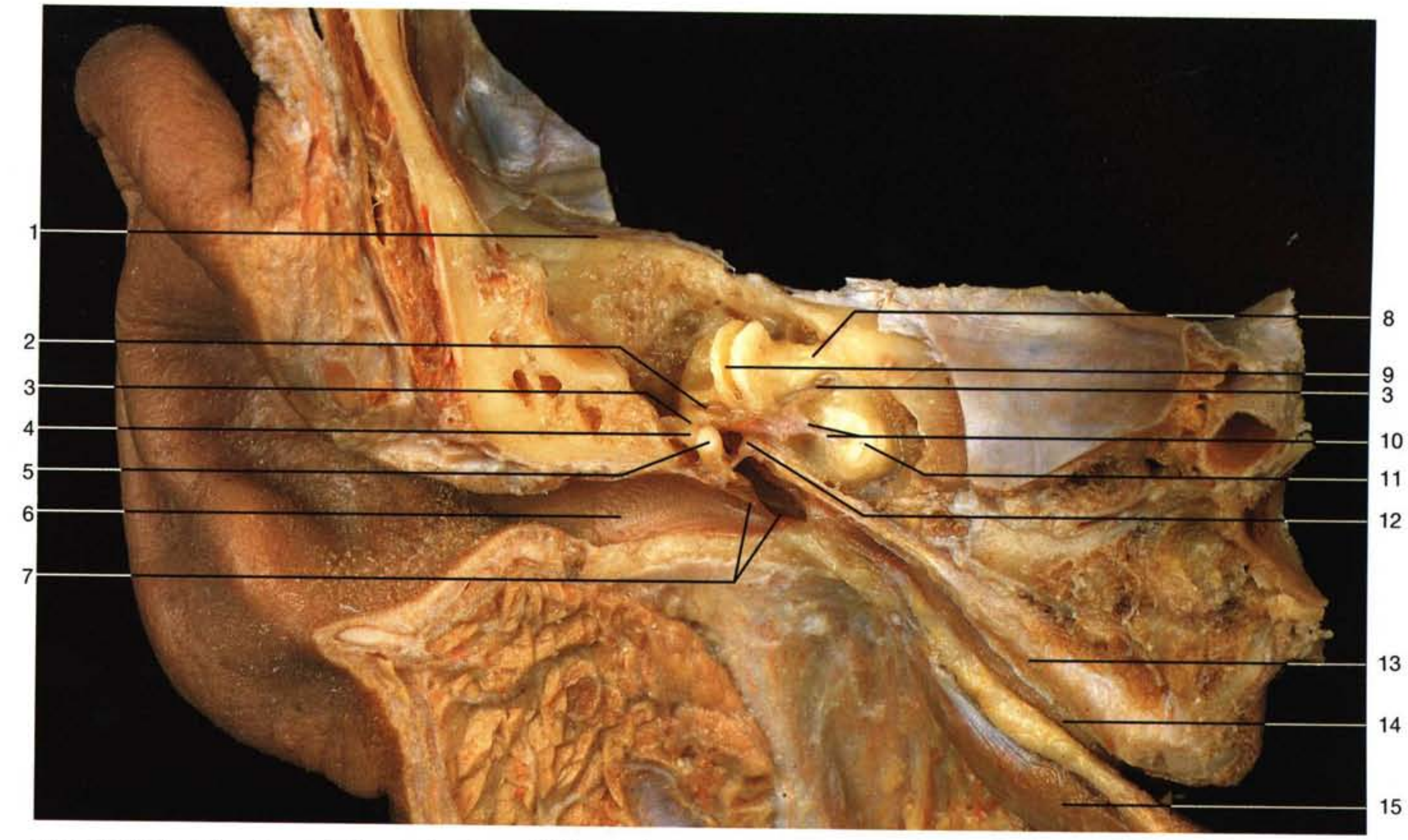

Longitudinal section through the right outer, middle and inner ear II. The cochlea and semicircular canals have been further dissected (anterior aspect).

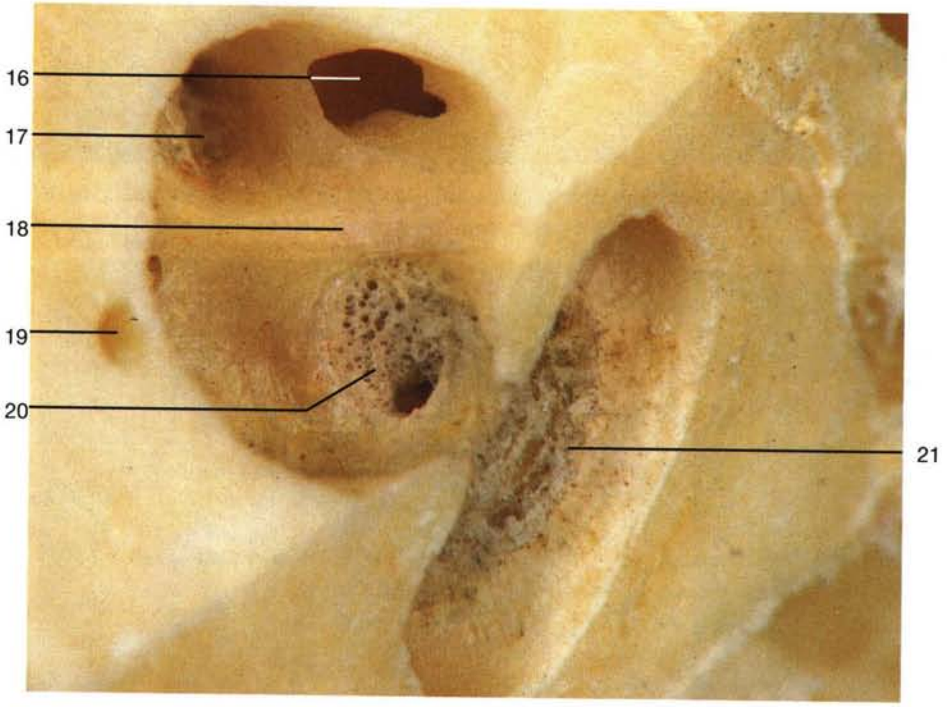

Internal acoustic meatus, left side. The bone was partly removed to show the bottom of the meatus.

1	Roof of tympanic cavity
2	**Lateral osseous semicircular canal**
3	**Facial nerve**
4	Incus
5	**Malleus**
6	External acoustic meatus
7	**Tympanic cavity** and **tympanic membrane**
8	Vestibulochochlear nerve
9	Anterior osseous semicircular canal
10	Geniculate ganglion and greater petrosal nerve
11	**Cochlea**
12	**Stapes**
13	Tensor tympani muscle
14	**Auditory tube**
15	Levator veli palatini muscle
16	Area of facial nerve
17	Superior vestibular area
18	Transverse crest
19	Foramen singulare
20	Foraminous spiral tract (outlet of cochlear part of vestibulocochlear nerve)
21	Base of cochlea

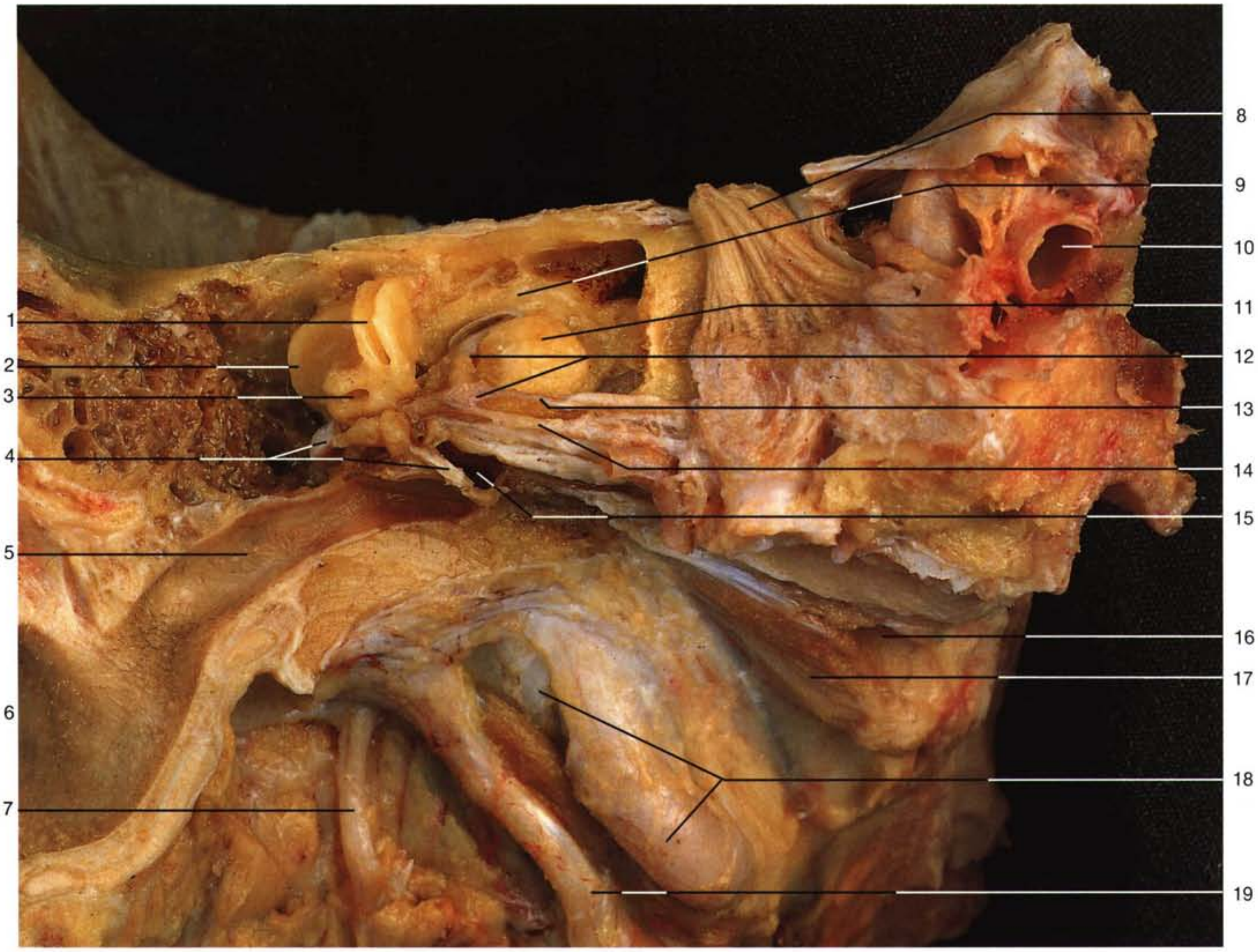

Longitudinal section through the outer, middle and inner ear III. Deeper dissection to display facial nerve and lesser and greater petrosal nerves (anterior aspect).

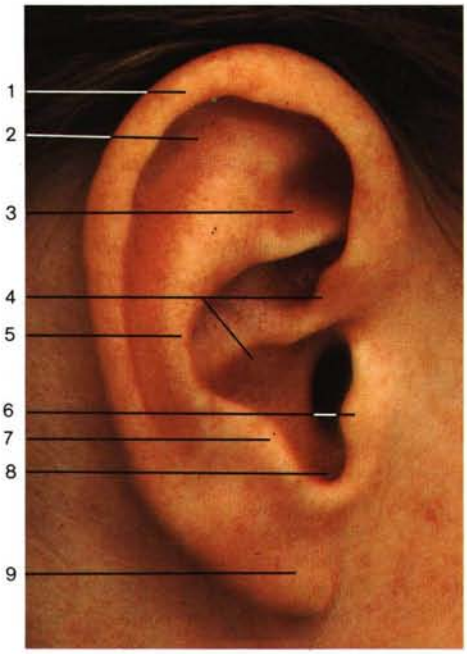

Right auricle (lateral aspect).

1 Helix
2 Scaphoid fossa
3 Triangular fossa
4 Concha
5 Antihelix
6 Tragus
7 Antitragus
8 Intertragic notch
9 Lobule

1 **Anterior osseous semicircular canal** (opened)
2 **Posterior osseous semicircular canal**
3 **Lateral osseous semicircular canal** (opened)
4 Facial nerve and chorda tympani
5 External acoustic meatus
6 Auricle
7 **Facial nerve**
8 Trigeminal nerve
9 Bony base of internal acoustic meatus
10 Internal carotid artery within cavernous sinus
11 **Cochlea**
12 **Facial nerve** with **geniculate ganglion**
13 Greater petrosal nerve
14 Lesser petrosal nerve
15 **Tympanic cavity**
16 Auditory tube
17 Levator veli palatini muscle
18 Internal carotid artery, internal jugular vein
19 Styloid process

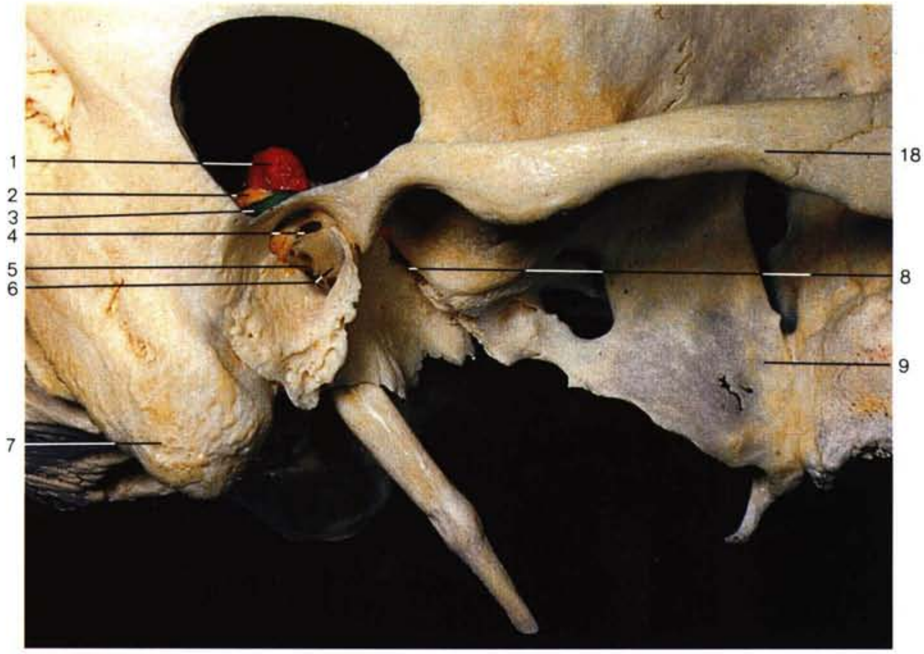

Right temporal bone (lateral aspect). Petrosquamous portion has been partly removed to display the semicircular canals.

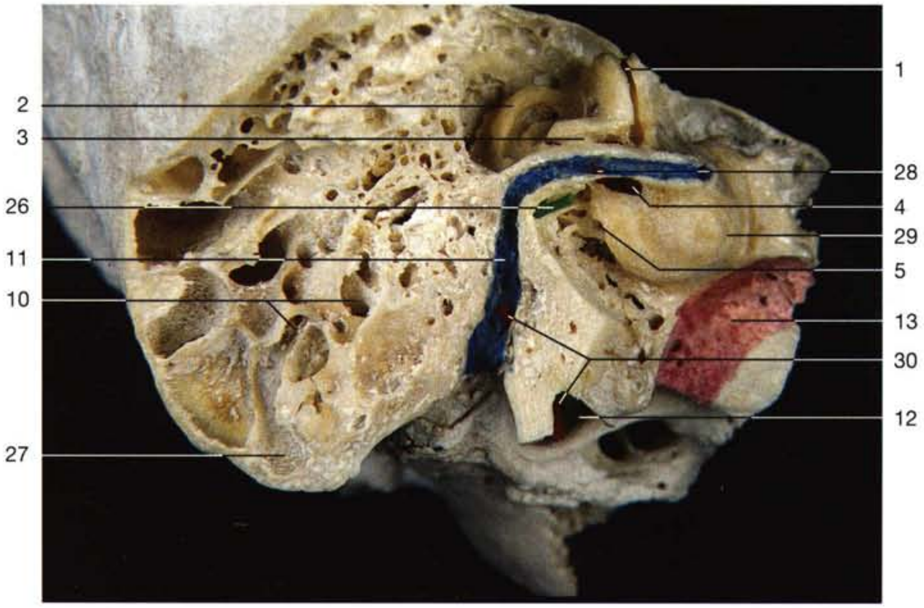

Right temporal bone (lateral aspect). Mastoid air cells and facial canal had been opened. The 3 semicircular canals were dissected.

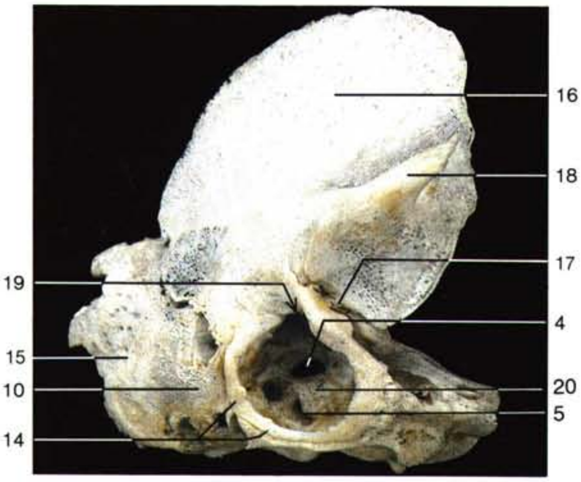

Right temporal bone of the newborn (lateral aspect).

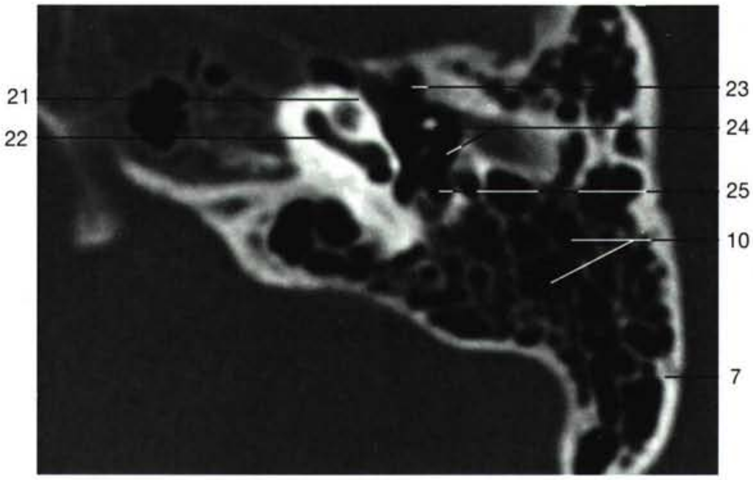

Frontal section through petrous part. (CT-Scan.)

1 Anterior semicircular canal (red)
2 Posterior semicircular canal (yellow)
3 Lateral or horizontal semicircular canal (green)
4 **Fenestra vestibuli**
5 **Fenestra cochleae**
6 Tympanic cavity
7 Mastoid process
8 **Petrotympanic fissure** (red probe: **chorda tympani**)
9 Lateral pterygoid plate
10 Mastoid air cells
11 **Facial canal** (blue)
12 Foramen ovale
13 **Carotid canal** (red)
14 **Tympanic ring**
15 Petromastoid part of temporal bone
16 Squamous part of temporal bone
17 Squamomastoid suture
18 Zygomatic process of temporal bone
19 Incisure of tympanic ring
20 **Promontory**
21 Apex of cochlea (cupula)
22 Spiral canal of cochlea at base of cochlea
23 Epitympanic recess
24 Auditory ossicles and tympanic cavity
25 Hypotympanic recess
26 Canaliculus chordae tympani (green probe)
27 Mastoid process
28 Canaliculus for stapedius nerve (red)
29 **Cochlea**
30 Canaliculus mastoideus (red probe)

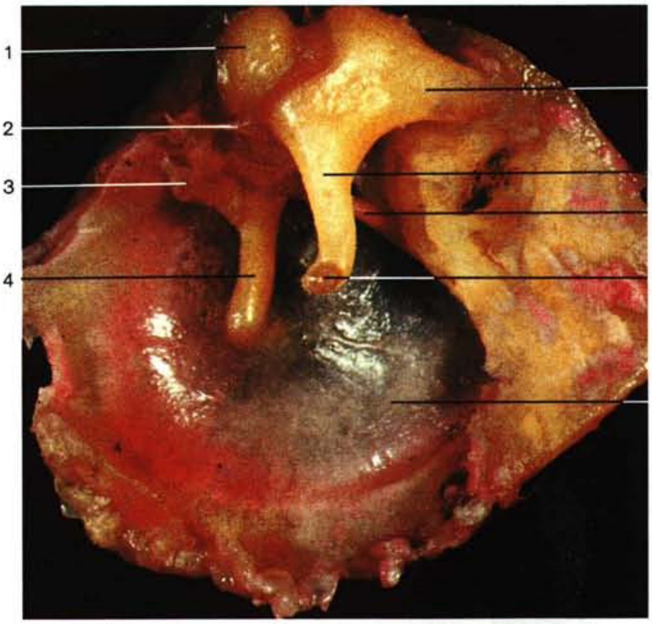

1 Head of malleus
2 Anterior ligament of malleus
3 Tendon of tensor tympani muscle
4 Handle of **malleus**
5 Short crus of incus
6 Long crus of **incus**
7 Chorda tympani
8 Lenticular process
9 **Tympanic membrane**

Tympanic membrane with malleus and incus (internal aspect; right side).

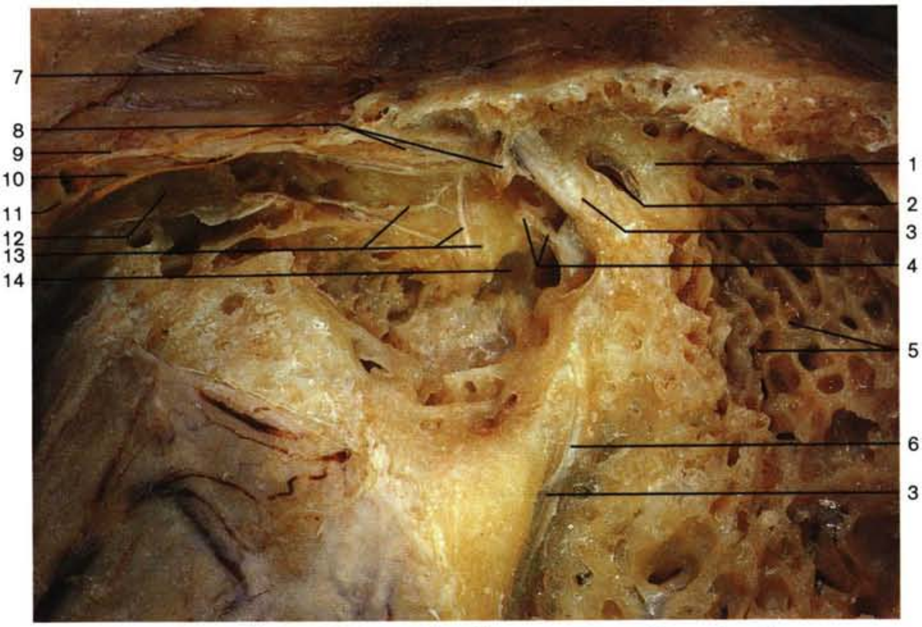

1 Tympanic antrum
2 Lateral semicircular canal (opened)
3 **Facial canal**
4 Stapes with tendon of stapedius
5 Mastoid air cells
6 Chorda tympani (intracranial part)
7 Greater petrosal nerve
8 Tensor tympani muscle (processus cochleariformis)
9 Lesser petrosal nerve
10 Anterior tympanic artery
11 Middle meningeal artery
12 **Auditory tube**
13 **Promontory with tympanic plexus**
14 Fenestra cochleae

Tympanic cavity, medial wall. External auditory meatus and lateral wall of tympanic cavity together with incus. Malleus and tympanic membrane have been removed; mastoid air cells are opened (left side).

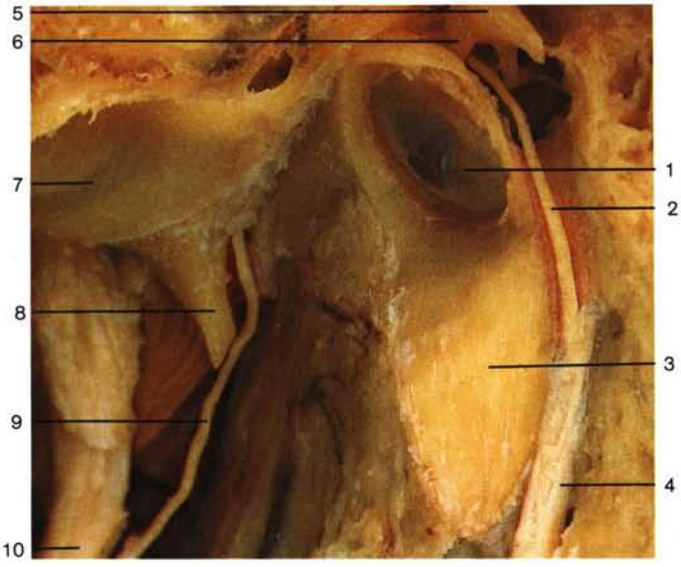

1 **Tympanic membrane**
2 Chorda tympani (intracranial part)
3 Floor of the external acoustic meatus
4 **Facial nerve** and facial canal
5 Incus
6 Head of malleus
7 Mandibular fossa
8 Spine of sphenoid
9 **Chorda tympani** (extracranial part)
10 Styloid process

Tympanic membrane (lateral aspect). External acoustic meatus and facial canal have been opened to expose the chorda tympani (magn. ~1.5×) (left side).

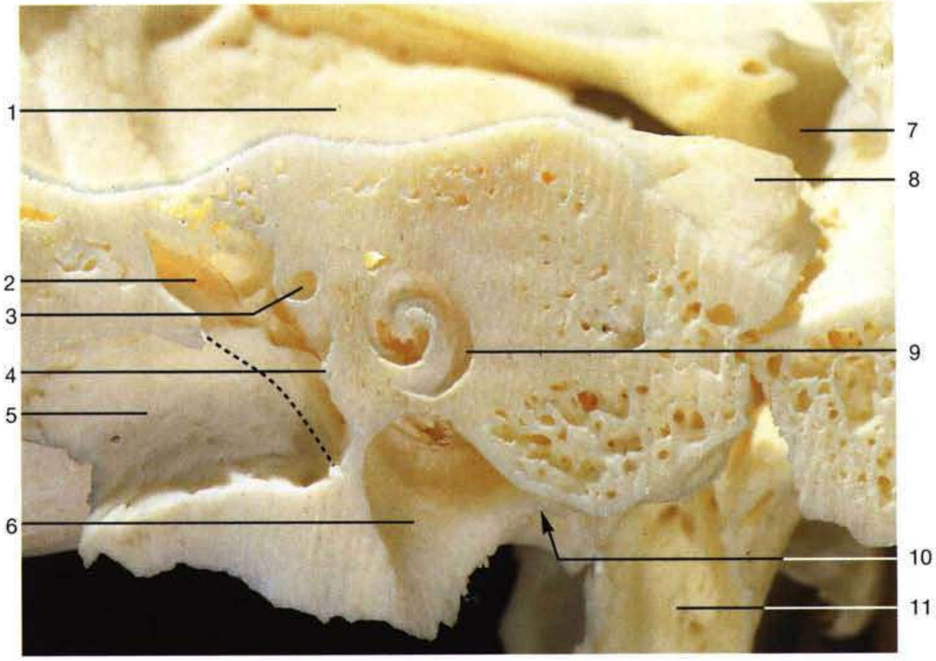

Frontal section through the petrous part of the left temporal bone at the level of the cochlea (posterior aspect). Position of tympanic membrane indicated by dotted line.

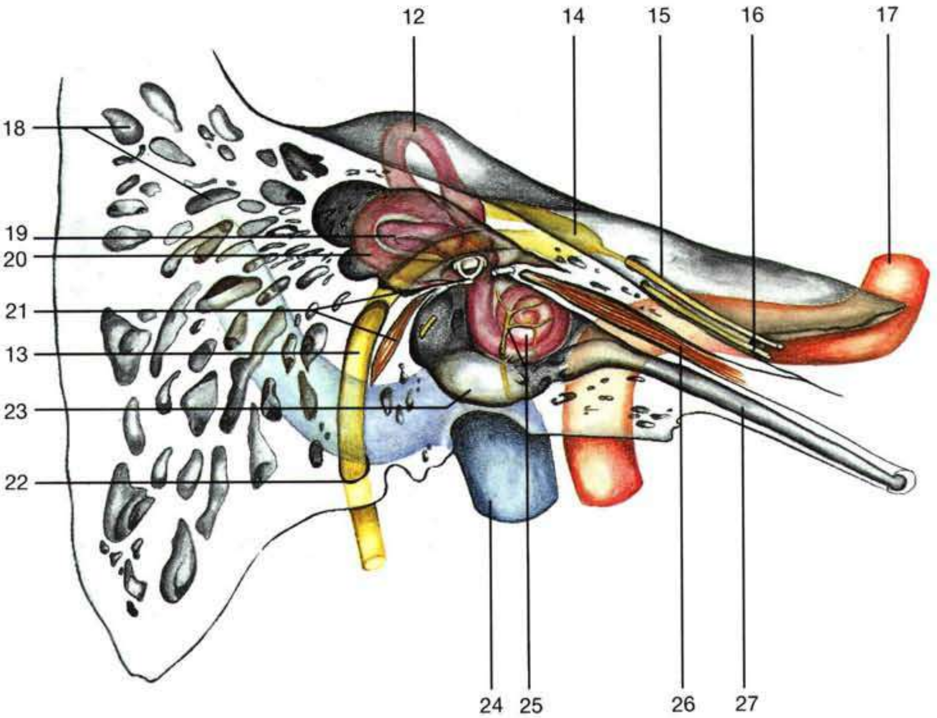

Medial wall of tympanic cavity and its relation to neighboring structures of the inner ear, facial nerve, and blood vessels. (Schematic drawing.) Frontal section through the right temporal bone (anterior aspect).

1 Anterior surface of the pyramid
2 Mastoid antrum
3 Lateral semicircular canal
4 Cochleariform process
5 External acoustic meatus
6 **Jugular fossa**
7 Foramen lacerum
8 Apex of petrous part
9 Position of **cochlea** (modiolus with crista spiralis ossea)
10 Carotid canal
11 Pterygoid process
12 Anterior semicircular duct
13 **Facial nerve**
14 Geniculate ganglion
15 Greater petrosal nerve
16 Lesser petrosal nerve
17 **Internal carotid artery**
18 Mastoid air cells
19 Lateral semicircular duct
20 Posterior semicircular duct
21 Stapes with stapedius muscle
22 Stylomastoid foramen
23 Inferior recess of tympanic cavity (hypotympanon)
24 **Internal jugular vein**
25 Promontory with tympanic plexus (position of cochlea)
26 Tensor muscle of tympanum
27 **Auditory tube**

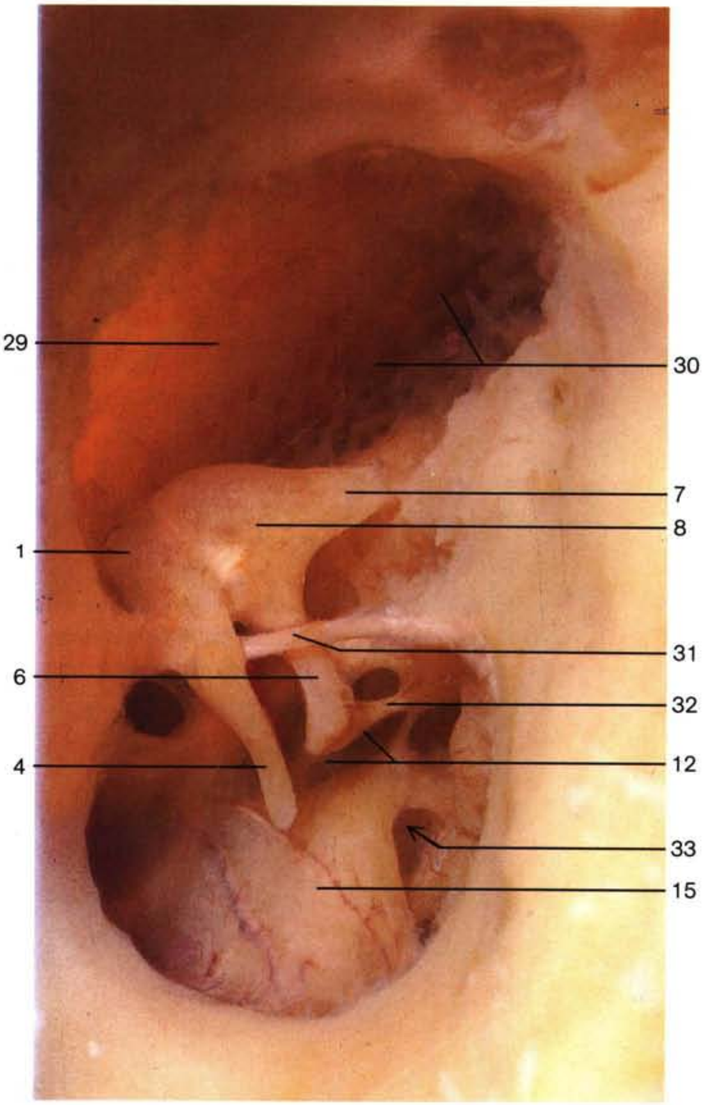

Tympanic cavity with malleus, incus and stapes, left side (lateral aspect). Tympanic membrane removed, mastoid antrum opened.

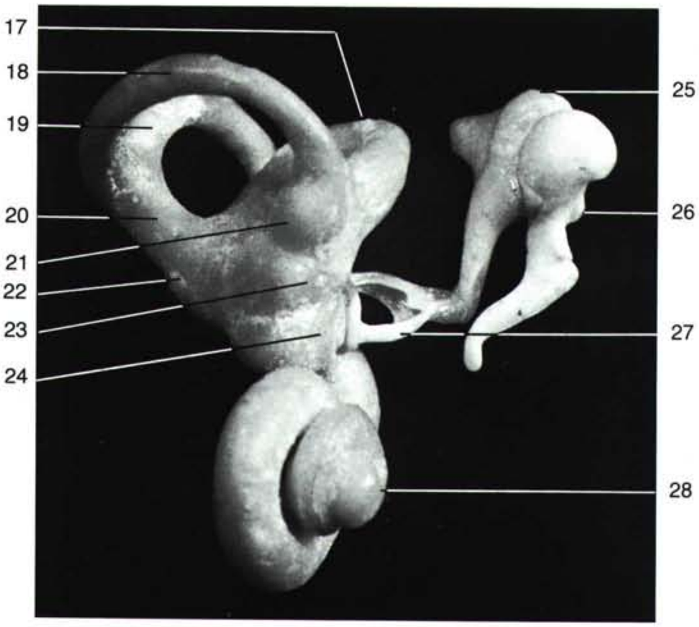

Chain of auditory ossicles in connection with the inner ear, left side (anterior-lateral aspect).

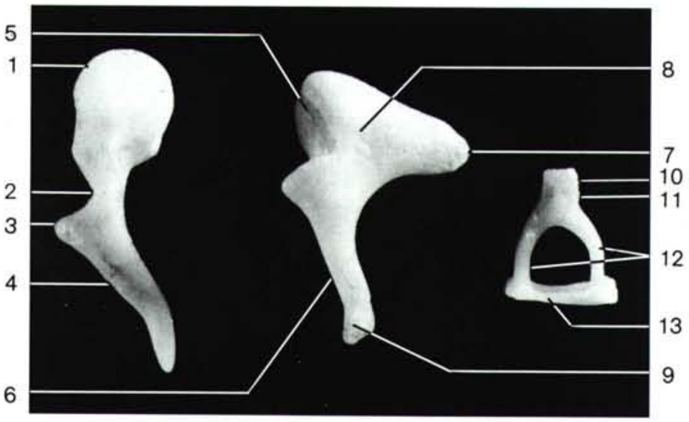

Auditory ossicles (isolated).

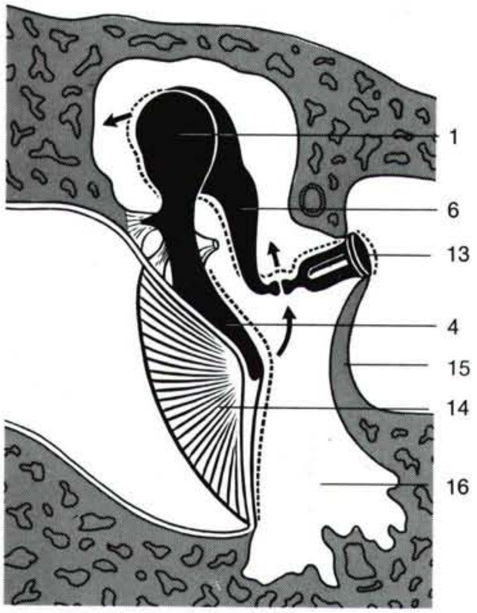

Position and movements of the auditory ossicles. (Schematic diagram.)

Malleus

1 Head
2 Neck
3 Lateral process
4 Handle

Incus

5 Articular facet for malleus
6 Long crus
7 Short crus
8 Body
9 Lenticular process

Stapes

10 Head
11 Neck
12 Anterior and posterior crura
13 Base

Walls of tympanic cavity

14 Tympanic membrane
15 Promontory
16 Hypotympanic recess of tympanic cavity

Internal ear (labyrinth)

17 Lateral semicircular duct
18 Anterior semicircular duct
19 Posterior semicircular duct
20 Common crus
21 Ampulla
22 Beginning of endolymphatic duct
23 Utricular prominence
24 Saccular prominence
25 Incus
26 Malleus
27 Stapes
28 Cochlea

Tympanic cavity

29 Epitympanic recess
30 Mastoid antrum
31 Chorda tympani
32 Tendon of stapedius muscle
33 Round window (fenestra cochleae)

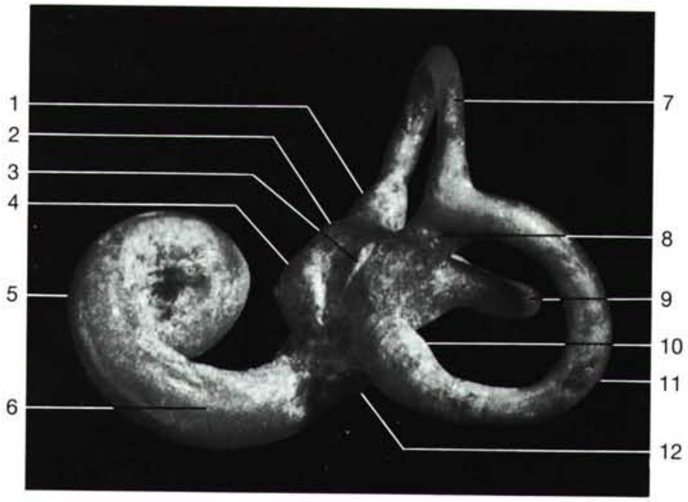

Cast of the right labyrinth (postero-medial aspect).

1 Ampulla (anterior semicircular canal)
2 Elliptical recess
3 Aqueduct of the vestibule
4 Spherical recess
5 **Cochlea**
6 Base of cochlea
7 **Anterior semicircular canal**
8 Crus commune or common limb
9 **Lateral semicircular canal**
10 Posterior bony ampulla
11 **Posterior semicircular canal** (posterior canal)
12 **Fenestra cochleae**
13 Bony ampulla
14 **Fenestra vestibuli**
15 Cupula of cochlea
16 External acoustic meatus
17 Mastoid air cells
18 Tympanic cavity and fenestra cochleae (probe)
19 External acoustic meatus
20 Facial canal
21 Base of cochlea and canalis musculotubarius
22 Malleus and incus
23 Stapes
24 Tympanic membrane
25 Tympanic cavity
26 Aqueduct of cochlea
27 Saccus endolymphaticus
28 Ductus endolymphaticus
29 Macula of utricle
30 Macula of saccule

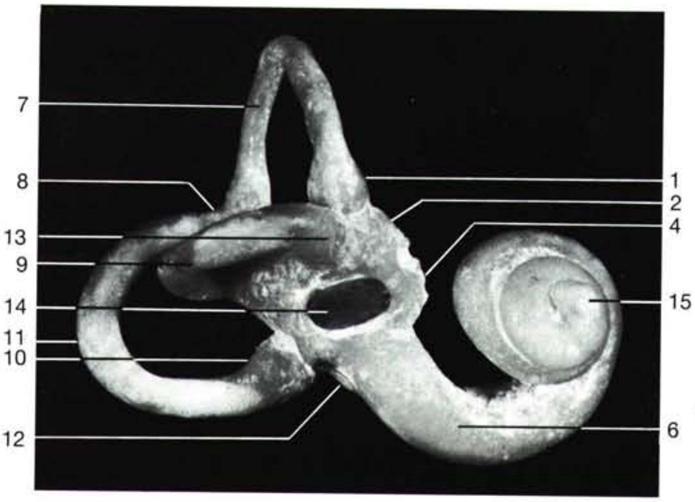

Cast of the right labyrinth (lateral aspect).

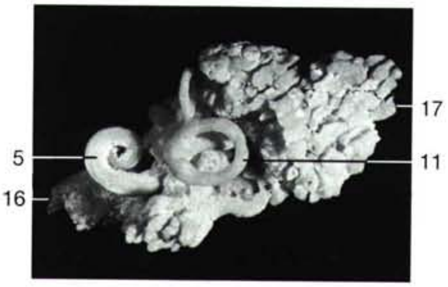

Cast of the labyrinth and mastoid cells. Life size (posterior aspect).

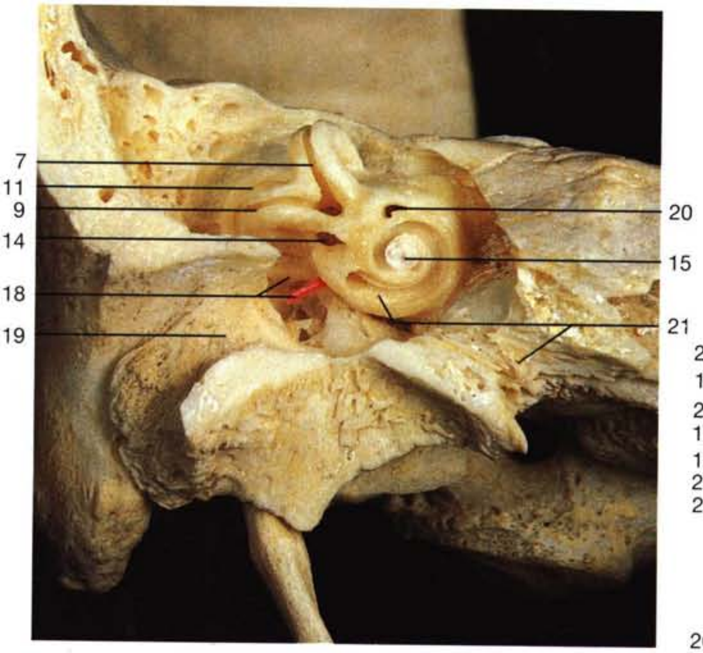

Dissection of bony labyrinth in situ. Semicircular canals and cochlear duct opened.

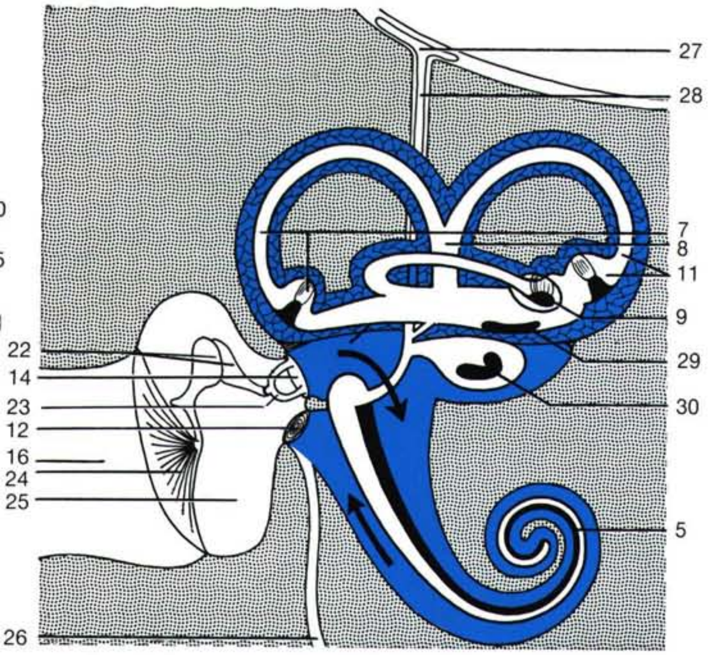

Auditory and vestibular apparatus. Arrows: direction of sound waves; blue = perilymphatic ducts. (Schematic diagram.)

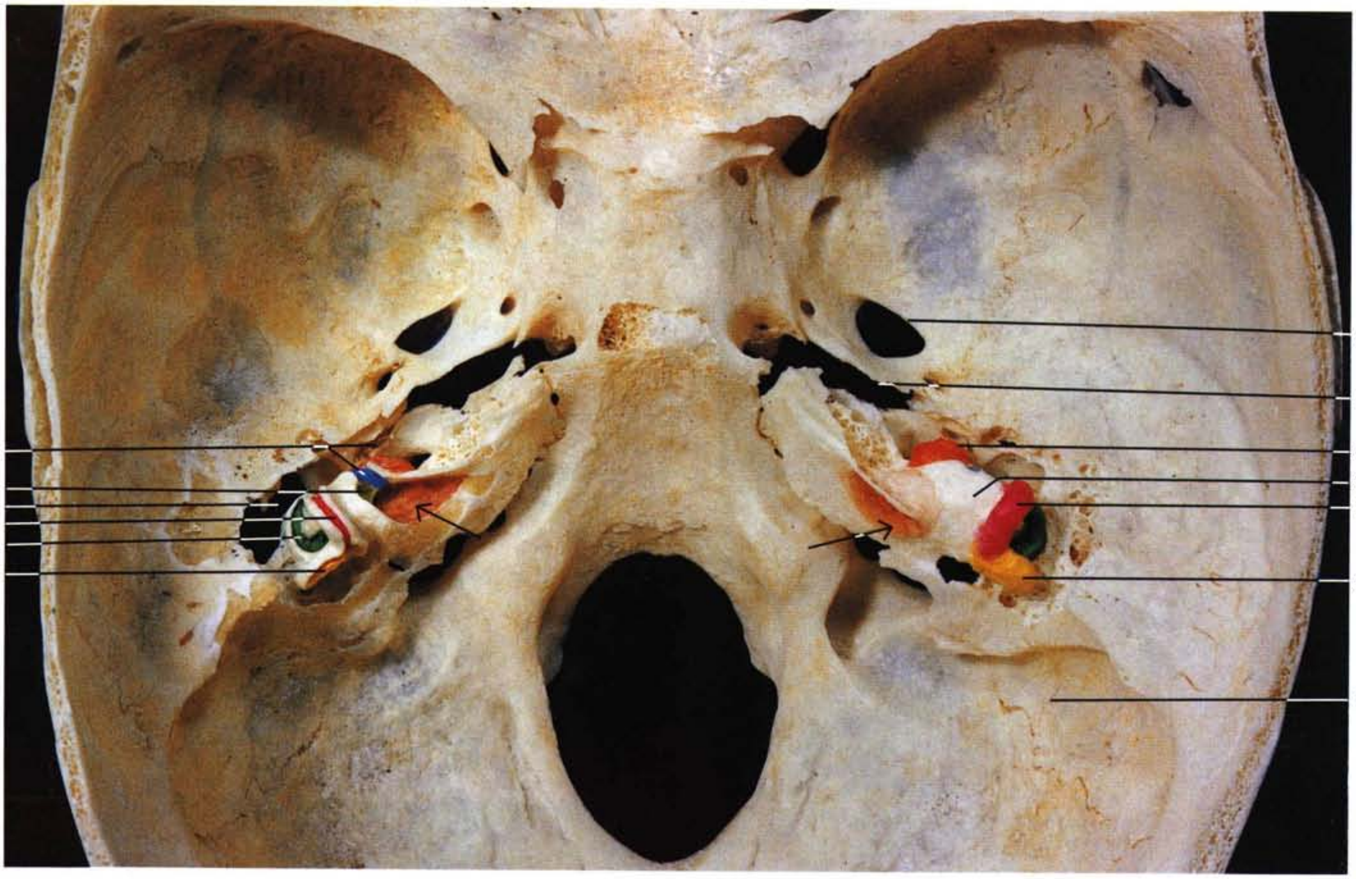

Bony labyrinth, petrous part of the temporal bone (from above). At left: semicircular canals opened, at right closed. Arrows: internal acoustic meatus.

1 Facial canal and semicanal of auditory tube
2 Superior vestibular area
3 Foramen ovale
4 Foramen lacerum
5 Cochlea
6 Vestibule
7 Anterior semicircular canal
8 Lateral semicircular canal
9 Posterior semicircular canal
10 Groove for sigmoid sinus
11 Sigmoid sinus
12 Tympanic cavity
13 Auditory tube
14 Mastoid air cells
15 Facial and vestibulocochlear nerves
16 Temporal fossa
17 Fenestra vestibuli
18 Promontory
19 Zygomatic process
20 Fenestra cochleae
21 Mastoid process

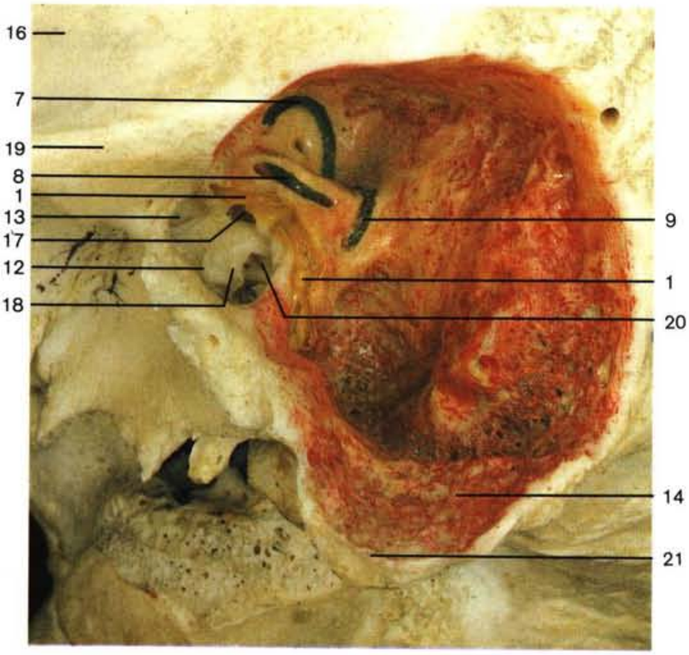

Bony labyrinth (left lateral aspect). Temporal and tympanic bone partly removed, semicircular canals opened.

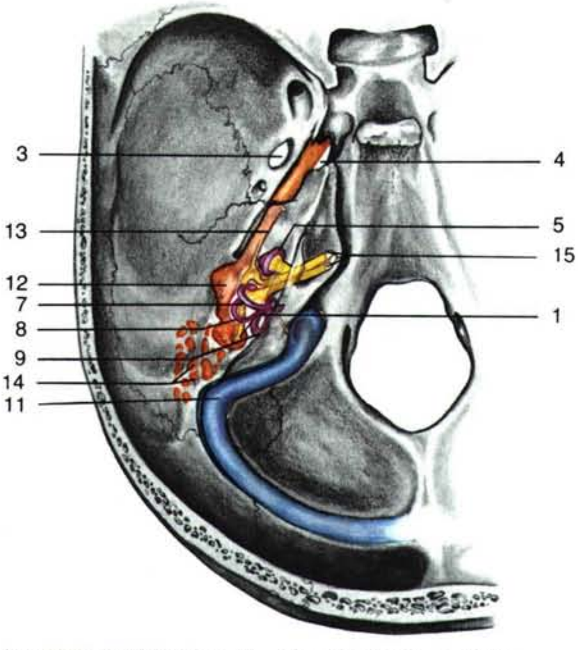

Internal ear. Diagram showing the position of the membranous labyrinth and the tympanic cavity.

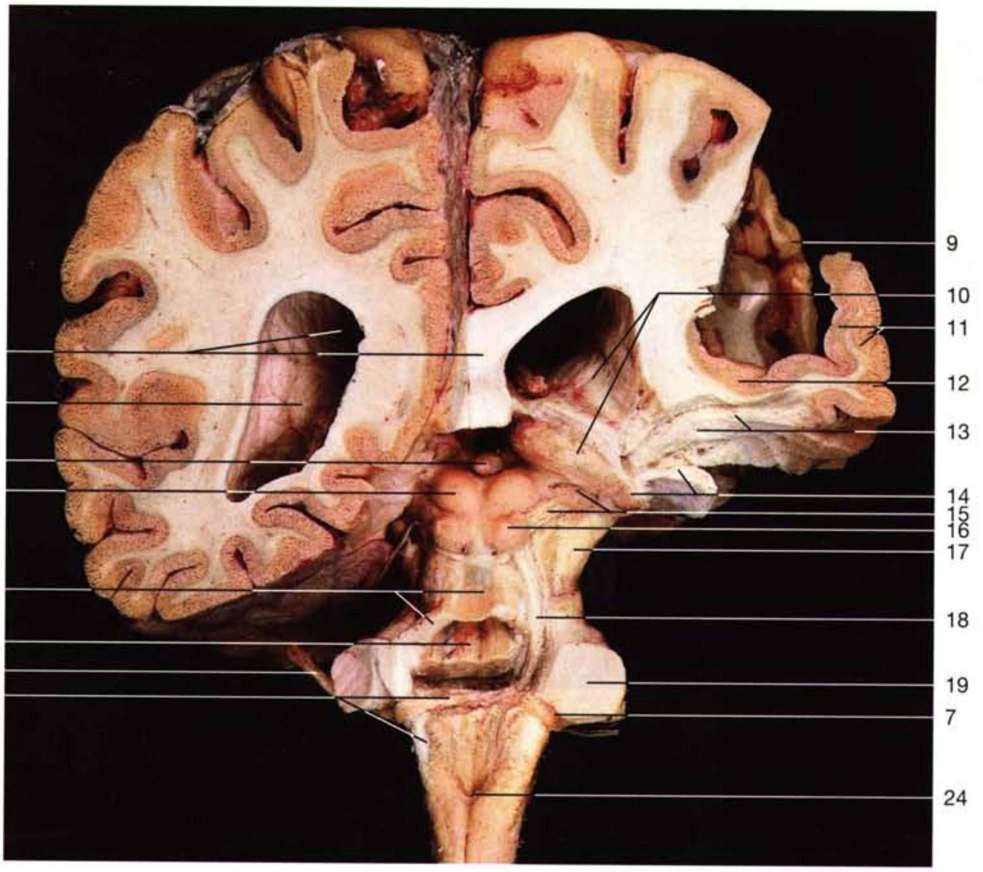

Dissection of the brain stem showing the acoustic pathway. Cerebellum and posterior part of the two hemispheres have been removed (dorsal aspect).

1 Left lateral ventricle and corpus callosum
2 Thalamus
3 Pineal gland (epiphysis)
4 Superior colliculus
5 Anterior cerebellar velum and superior cerebellar peduncle
6 Rhomboid fossa
7 **Vestibulocochlear nerve** (n. VIII)
8 **Dorsal acoustic striae** and inferior cerebellar peduncle
9 Insular lobe
10 Caudate nucleus and thalamus
11 Temporal lobe (superior temporal gyrus) (area of **acoustic centers**)
12 Transverse temporal gyri of Heschl (area of **primary acoustic centers**)
13 Acoustic radiation of internal capsule
14 Lateral geniculate body and optic radiation (cut)
15 **Medial geniculate body** and brachium of inferior colliculus
16 **Inferior colliculus**
17 Cerebral peduncle
18 **Lateral lemniscus**
19 Middle cerebellar peduncle
20 Dorsal (posterior) cochlear nucleus
21 Ventral (anterior) cochlear nucleus
22 Inferior olive with tractus olivocochlearis of Rasmussen (red)
23 Ganglion spirale
24 Obex
25 Frontal lobe
26 Temporal lobe
27 Middle temporal gyrus (area of **tertiary acoustic centers)**
28 Trapezoid body

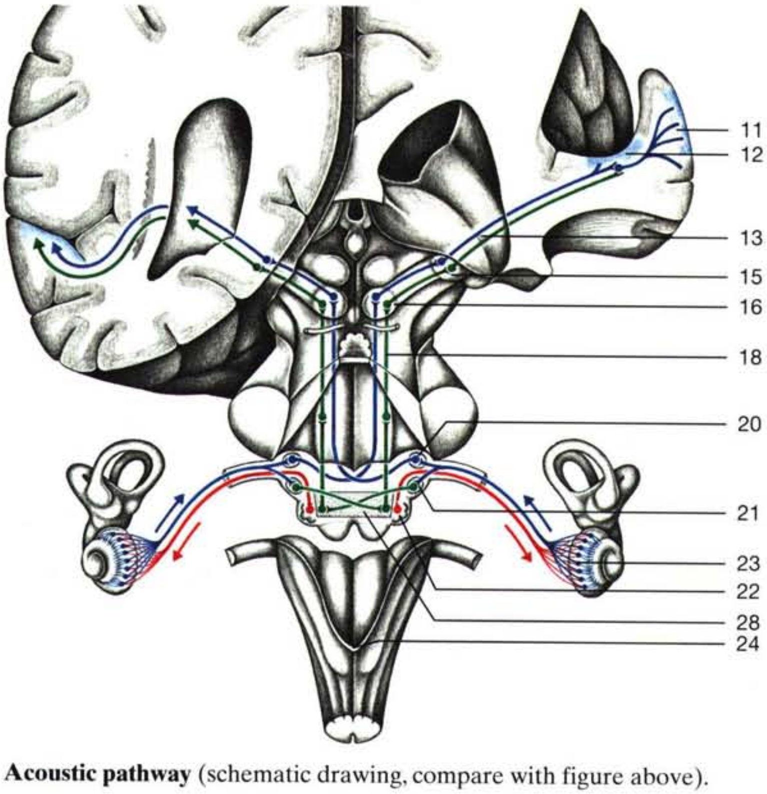

Acoustic pathway (schematic drawing, compare with figure above).

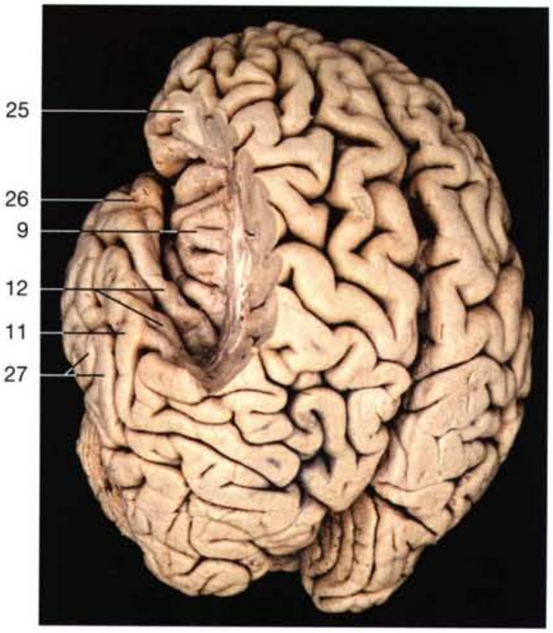

Acoustic areas in the left hemisphere (superior lateral aspect). Parts of the frontal and parietal lobe have been removed.

1 Frontal bone
2 Nasal bone
3 Lacrimal bone
4 Maxilla (frontal process)
5 Ethmoidal foramina
6 Lesser wing of sphenoid bone and optic canal
7 Superior orbital fissure
8 Greater wing of sphenoid bone
9 Orbital process of palatine bone
10 Orbital plate of ethmoid bone
11 Inferior orbital fissure
12 Infraorbital sulcus
13 Nasolacrimal canal
14 Zygomatic bone
15 Frontal sinus
16 **Superior rectus muscle**
17 Orbital fatty tissue
18 **Optic nerve**
19 **Sclera**
20 **Inferior rectus muscle**
21 Periorbita and maxilla
22 Maxillary sinus
23 Levator palpebrae superioris muscle
24 **Superior conjunctival fornix**
25 Superior tarsal plate
26 Inferior tarsal plate
27 **Inferior conjunctival fornix**
28 **Inferior oblique muscle**
29 **Lateral rectus muscle**
30 **Medial rectus muscle**
31 **Superior oblique muscle**
32 Nasal septum
33 Middle nasal concha
34 Inferior nasal concha
35 Tenon's space
36 Ophthalmic artery
37 **Cornea**
38 **Lens**

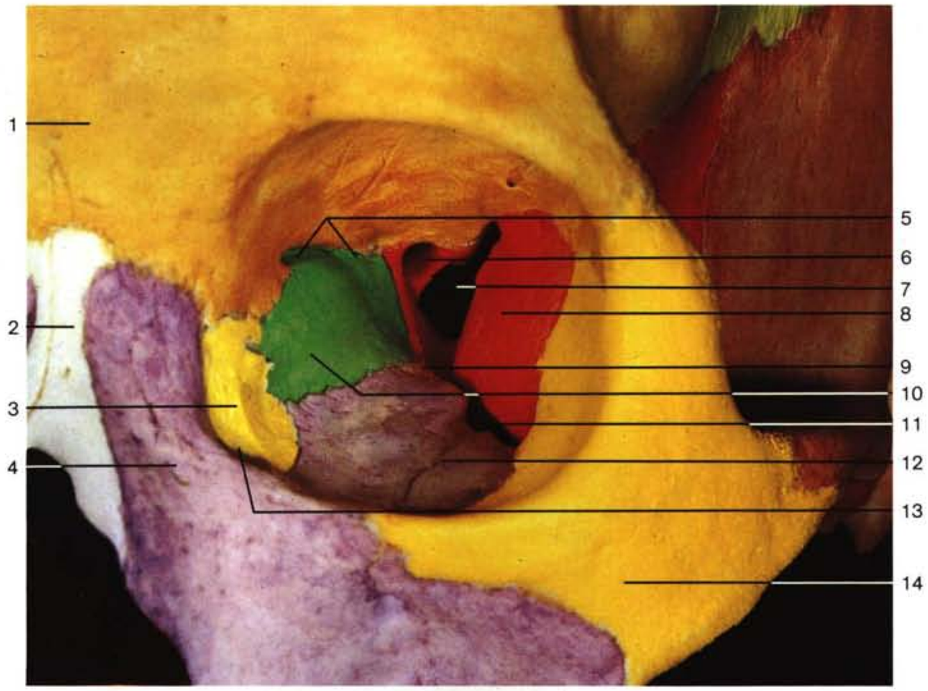

Bones of the left orbit (indicated by different colors).

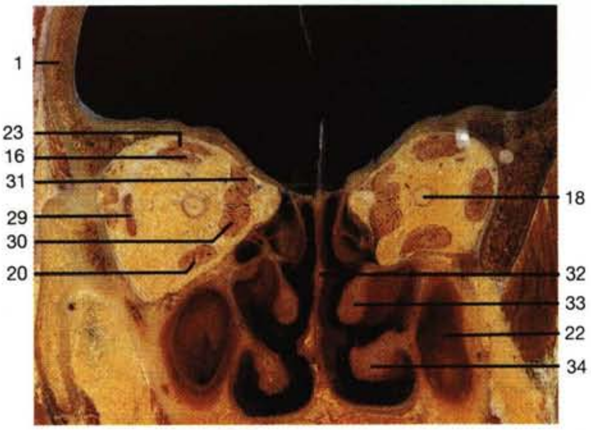

Frontal section through the posterior part of the orbit.

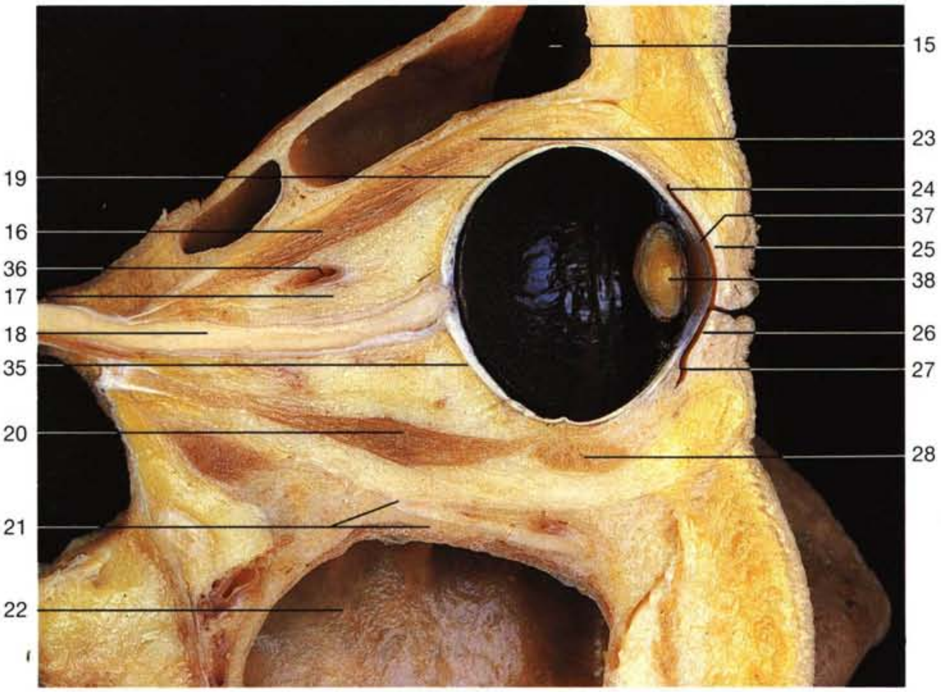

Sagittal section through orbit and eyeball.

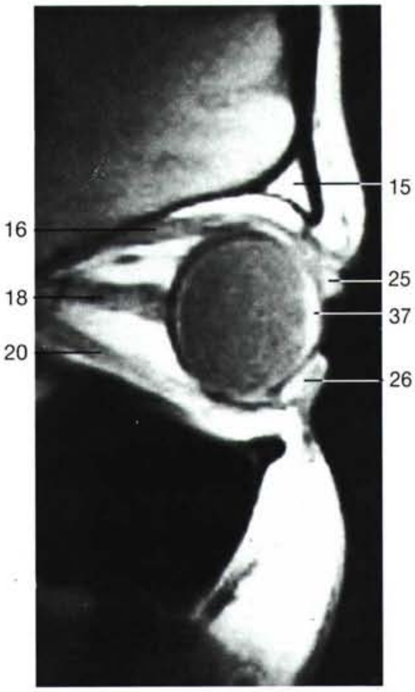

Sagittal section through orbit and eyeball. (MR-Scan.)

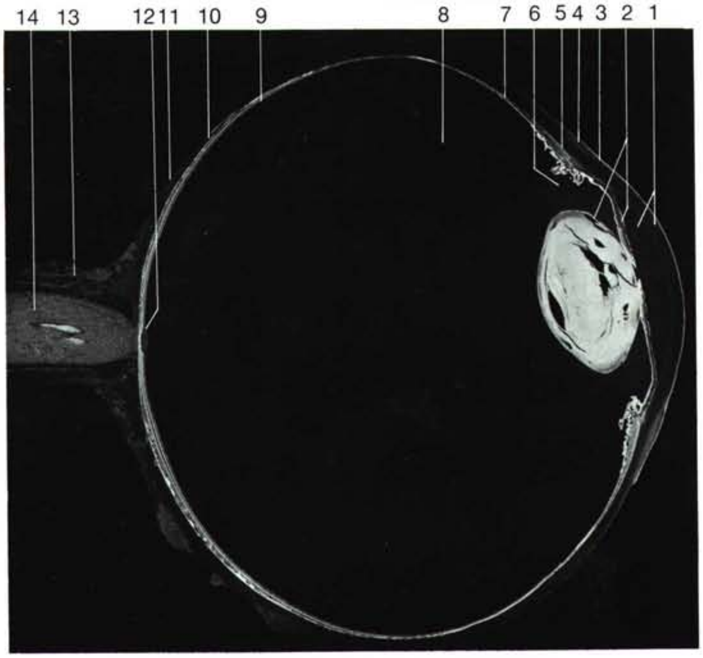

Horizontal section through the human eye (2×).

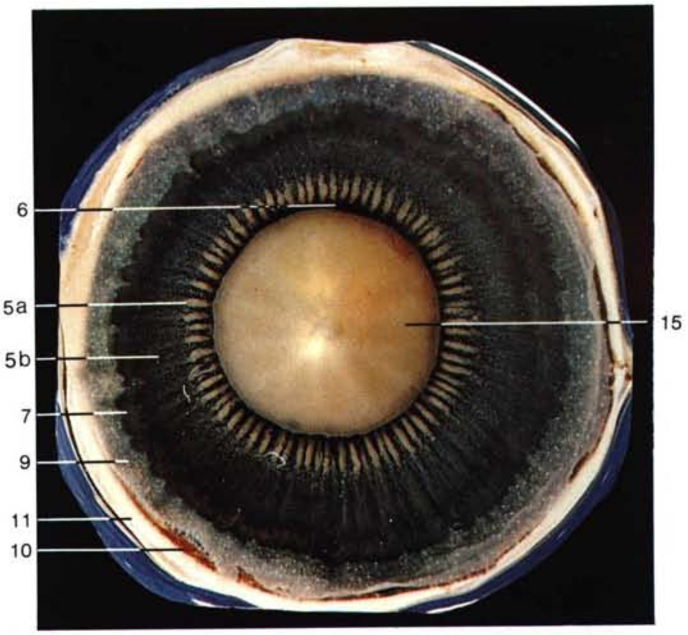

Anterior segment of the eyeball (posterior aspect). The opacity of the lens is an artifact.

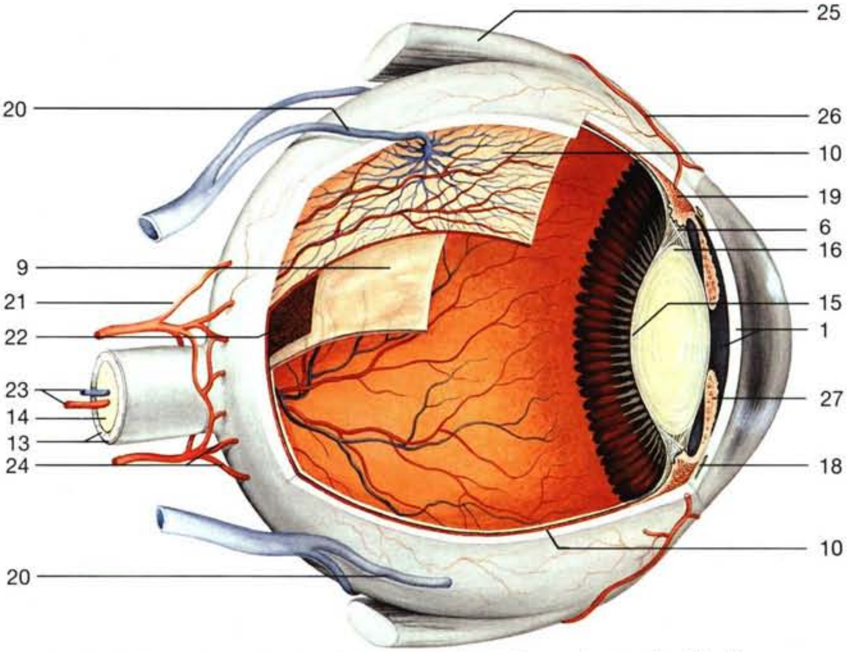

Organization of the eyeball. Demonstration of vascular tunic of bulb. (Schematic drawing.)

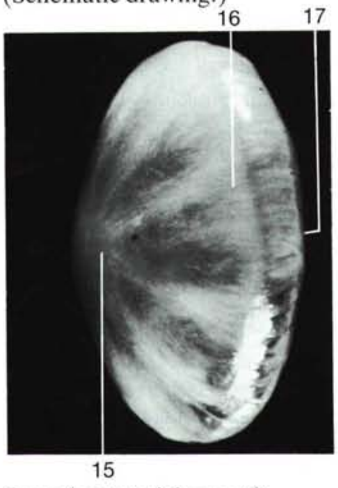

Lens (equatorial aspect), anterior pole to the right.

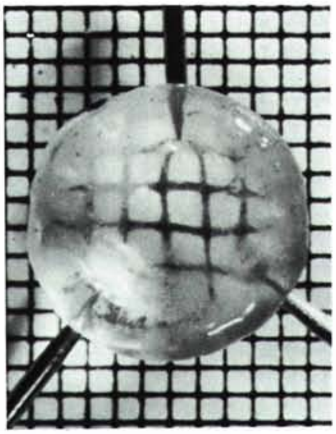

Lens (frontal aspect). Note the magnification effect.

1 **Cornea** and anterior chamber
2 Iris and lens
3 Transitional zone between corneal and conjunctival epithelium
4 Conjunctiva of the eyeball
5 **Ciliary body**
  a Ciliary processes (pars plicata)
  b Ciliary ring (pars plana)
6 Zonular fibers
7 **Ora serrata**
8 Vitreous body
9 **Retina**
10 **Choroid**
11 **Sclera**
12 Optic disc
13 Dura mater and subarachnoid space
14 **Optic nerve** (n. II)
15 Lens (posterior pole)
16 Equator of lens
17 **Lens** (anterior pole)
18 Canal of Schlemm
19 Ciliary muscle
20 Vena vorticosa
21 Long posterior ciliary artery
22 Retinal pigmented epithelium
23 Central retinal artery and vein
24 Short posterior ciliary arteries
25 External ocular muscle
26 Anterior ciliary artery
27 **Iris**

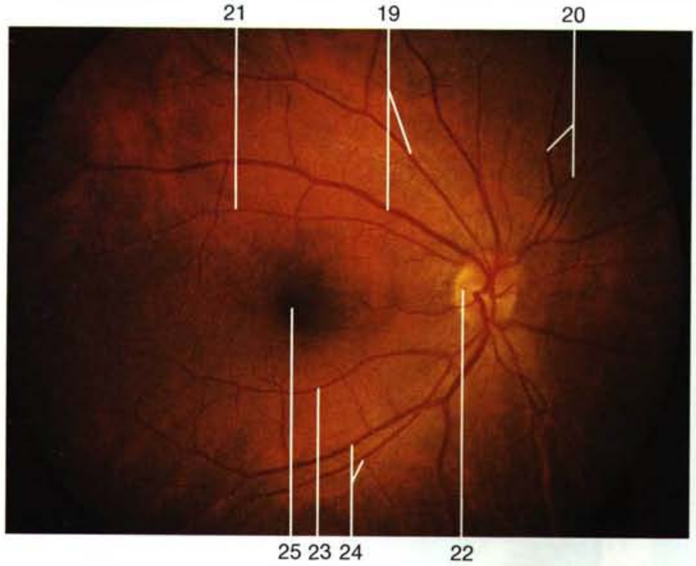

Fundus of a normal right eye (courtesy of Prof. Dr. R. Okamura, Univ. Eye Dept., Kumamoto/Japan). Notice, the arteries are smaller and lighter than the veins.

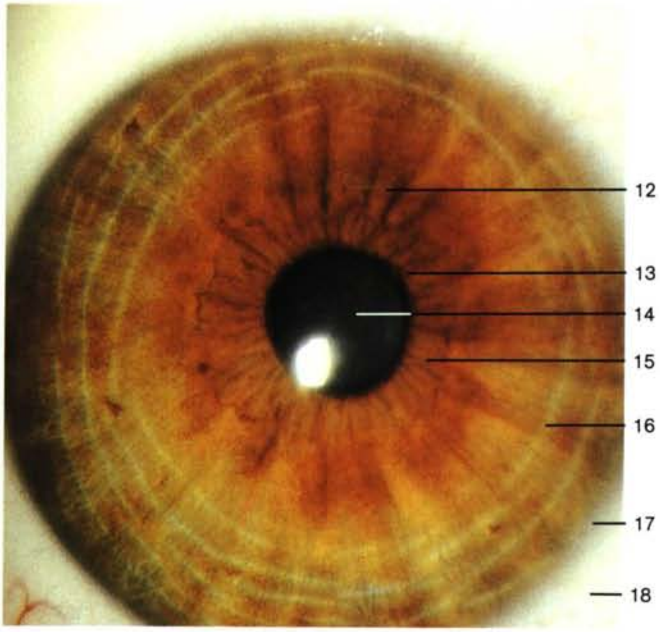

Anterior segment of the human eye (courtesy of Prof. Dr. G. O. H. Naumann, Eye Dept., University of Erlangen, FRG). Note the colored iris and the anterior lens surface.

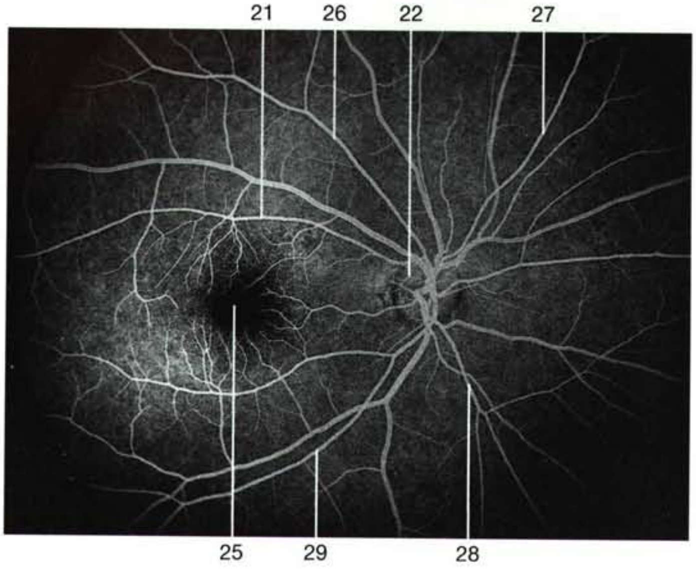

Fluorescent angiography of the right eye; retinal vessels. The same eye as above. (Courtesy of Prof. Dr. R. Okamura, Univ. Eye Dept., Kumamoto/Japan.)

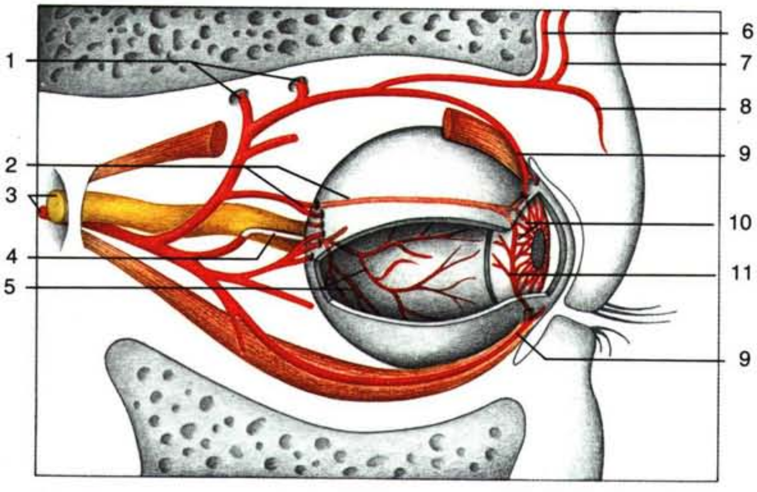

Diagram of the ophthalmic artery and its branches.

1 Posterior and anterior ethmoidal arteries
2 Long and short **posterior ciliary arteries**
3 Optic nerve and **ophthalmic artery**
4 Central retinal artery
5 Retinal arteries
6 Supratrochlear artery
7 Supraorbital artery
8 Dorsal nasal artery
9 **Anterior ciliary artery**
10 Iridial arteries
11 Circulus arteriosus major of iris
12 Iridial fold
13 Pupillary margin of iris
14 Anterior pole of lens
15 Lesser circle of **iris**
16 Greater circle of iris
17 Margin of cornea or limbus
18 Sclera
19 Superior temporal artery and vein of retina
20 Superior nasal artery and vein of retina
21 Superior macular artery
22 **Optic disc**
23 Inferior macular artery
24 Inferior temporal artery and vein
25 **Fovea centralis** and **macula lutea**
26 Superior temporal artery
27 Superior nasal artery
28 Inferior nasal artery
29 Inferior temporal artery
of retina

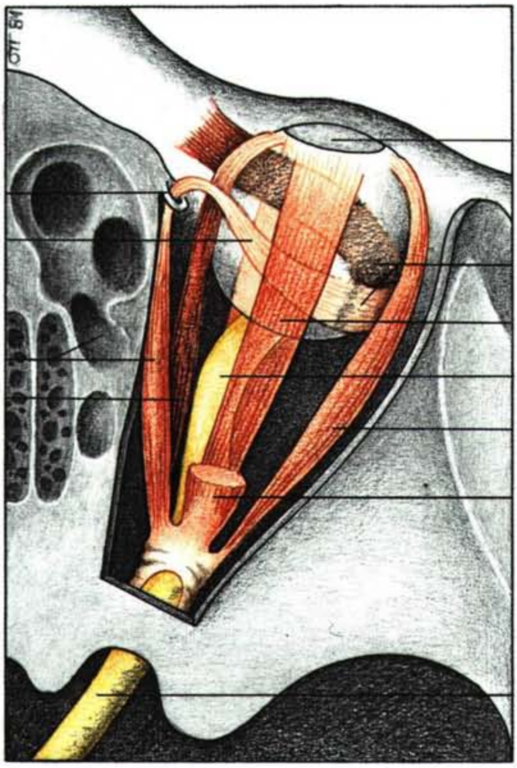

Schematic diagram of the extraocular muscles. Right orbit (from above). Levator palpebrae superioris muscle has been severed.

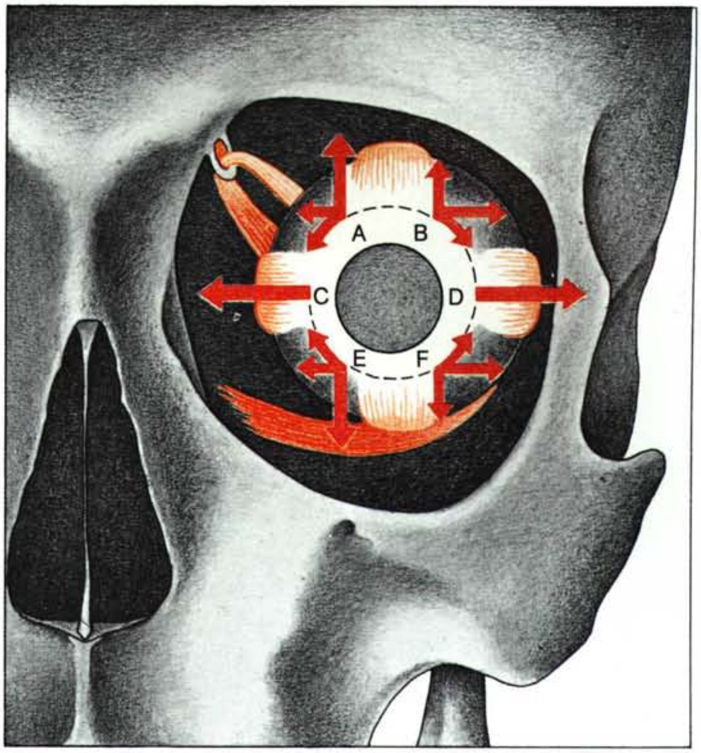

The action of the extraocular muscles. Left orbit (anterior aspect).

A Superior rectus muscle
B Inferior oblique muscle
C Medial rectus muscle
D Lateral rectus muscle
E Inferior rectus muscle
F Superior oblique muscle

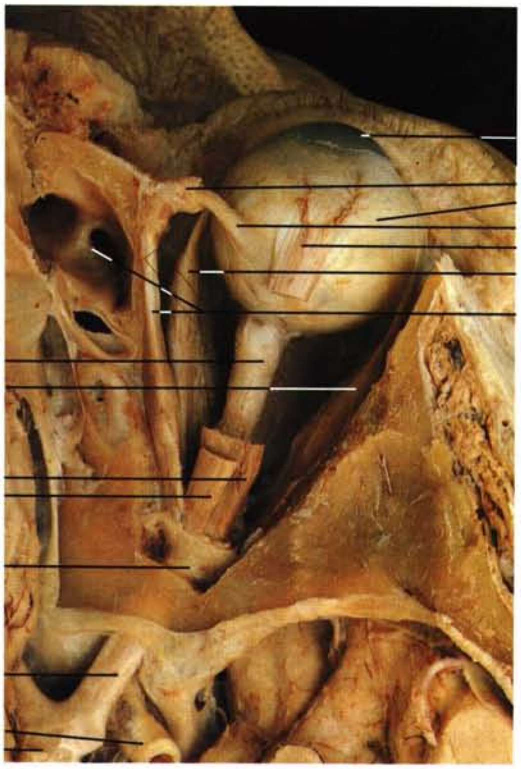

Right orbit with eyeball and extraocular muscles (from above). The roof of the orbit has been removed, the superior rectus muscle and the levator palpebrae superioris muscle have been severed.

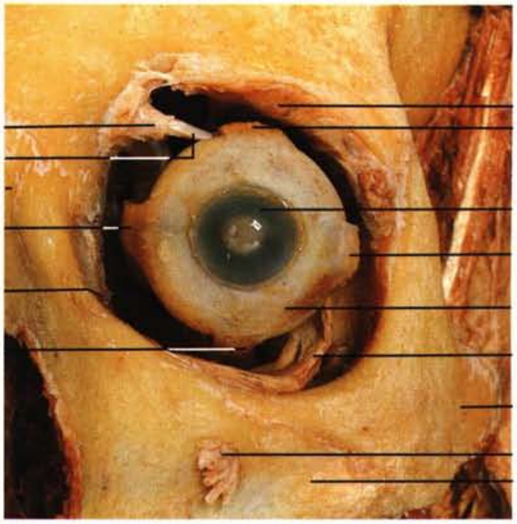

Left orbit with eyeball and extraocular muscles (anterior aspect). Lids, conjunctiva and lacrimal apparatus have been removed.

1 **Superior oblique muscle** and ethmoid air cells
2 **Medial rectus muscle**
3 Trochlea
4 Tendon of superior oblique muscle
5 **Superior rectus muscle**
6 Cornea
7 Eyeball
8 Optic chiasma
9 Optic nerve (intracranial part)
10 Internal carotid artery
11 **Common annular tendon**
12 Levator palpebrae superioris muscle
13 **Superior rectus muscle**
14 Optic nerve (extracranial part)
15 **Lateral rectus muscle**
16 Nasolacrimal duct
17 **Inferior oblique muscle**
18 Nasal bone
19 Maxilla
20 Infraorbital foramen and nerves
21 Zygomatic bone
22 **Inferior rectus muscle**

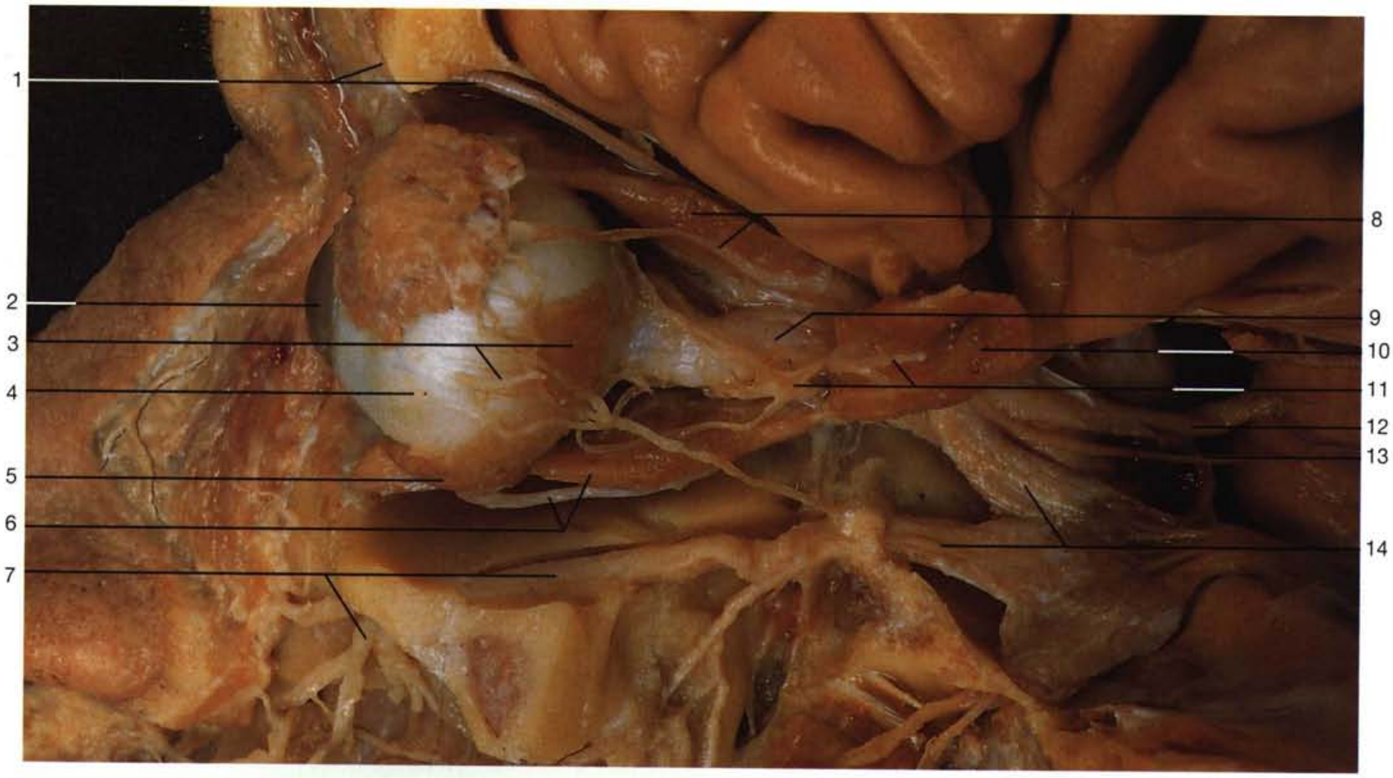

Extraocular muscles and their nerves (lateral aspect of left eye). Lateral rectus divided and reflected.

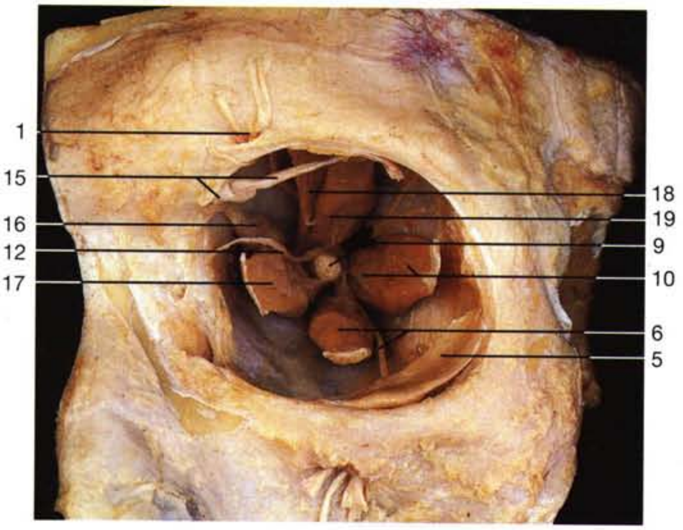

Left orbit with extraocular muscles (anterior aspect). Eyeball removed.

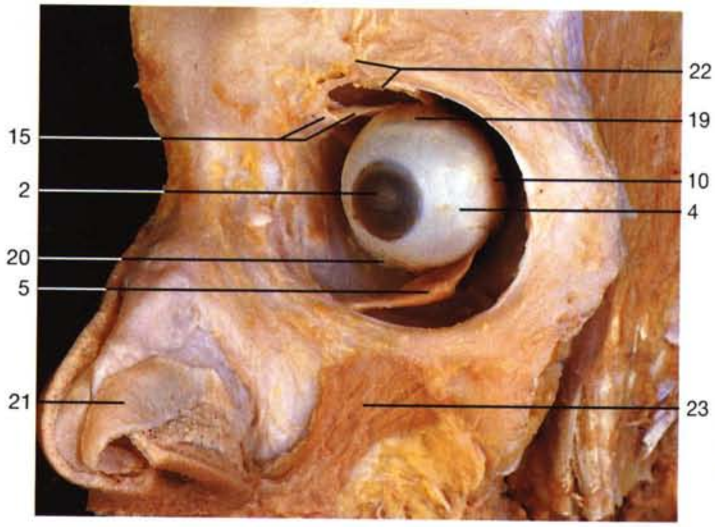

Extraocular eye muscles (anterior-lateral aspect).

1 Supraorbital nerve
2 Cornea
3 Insertion of lateral rectus muscle
4 Eyeball (sclera)
5 **Inferior oblique muscle**
6 **Inferior rectus muscle** and inferior branch of oculomotor nerve
7 Infraorbital nerve
8 Superior rectus muscle and lacrimal nerve
9 Optic nerve
10 **Lateral rectus muscle**
11 Ciliary ganglion and abducens nerve (n. VI)
12 Oculomotor nerve (n. III)
13 Trochlear nerve (n. IV)
14 Ophthalmic nerve (n. V_1) and maxillary nerve (n. V_2)
15 Trochlea and tendon of superior oblique muscle
16 **Superior oblique muscle**
17 **Medial rectus muscle**
18 Levator palpebrae superioris muscle
19 **Superior rectus muscle**
20 **Inferior rectus muscle**
21 Greater alar cartilage
22 Supraorbital nerve and levator palpebrae superioris muscle
23 Levator labii superioris muscle

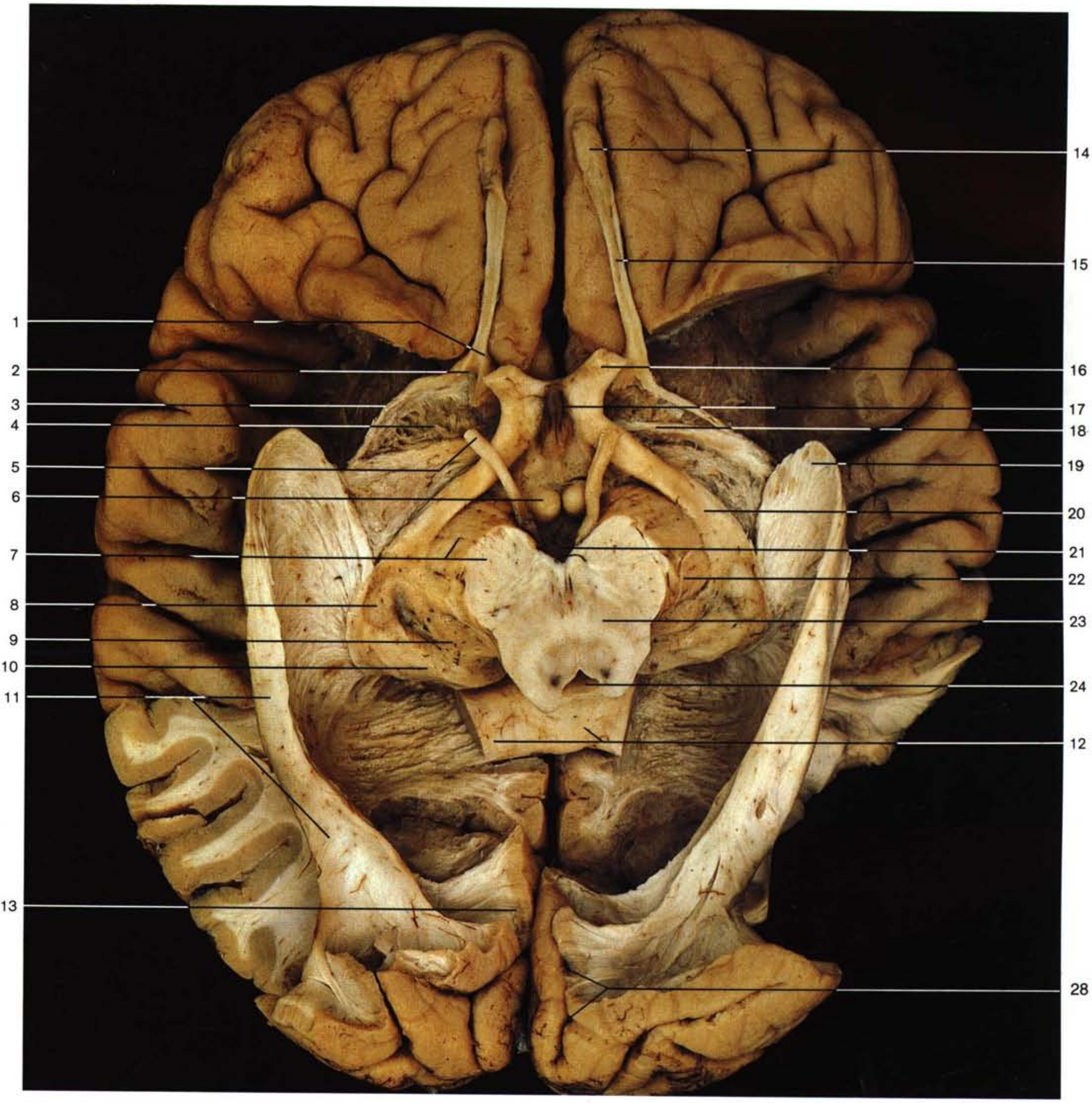

Dissection of the visual pathway (inferior aspect). Frontal pole at top, midbrain divided.

1	Medial olfactory stria
2	Olfactory trigone
3	Lateral olfactory stria
4	Anterior perforated substance
5	Oculomotor nerve (n. III)
6	Mamillary body
7	Cerebral peduncle
8	**Lateral geniculate body**
9	Medial geniculate body
10	**Pulvinar of thalamus**
11	Optic radiation
12	Splenium of the corpus callosum (commissural fibers)
13	Cuneus
14	Olfactory bulb
15	Olfactory tract
16	**Optic nerve** (n. II)
17	Infundibulum
18	Anterior commissure
19	**Genu of optic radiation**
20	**Optic tract**
21	Interpeduncular fossa and posterior perforated substance
22	Trochlear nerve (n. IV)
23	Substantia nigra
24	Cerebral aqueduct
25	**Visual cortex**
26	Line of Gennari
27	Gyrus of striate cortex
28	**Calcarine sulcus**

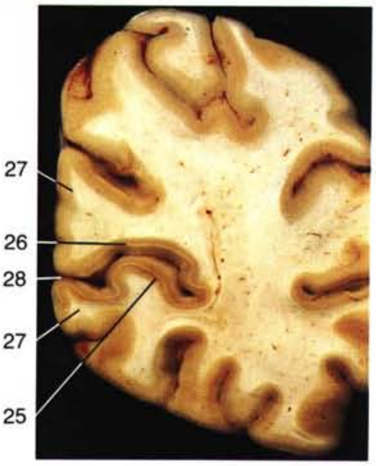

Frontal section of the striate cortex at the level of the striate area in the occipital lobe.

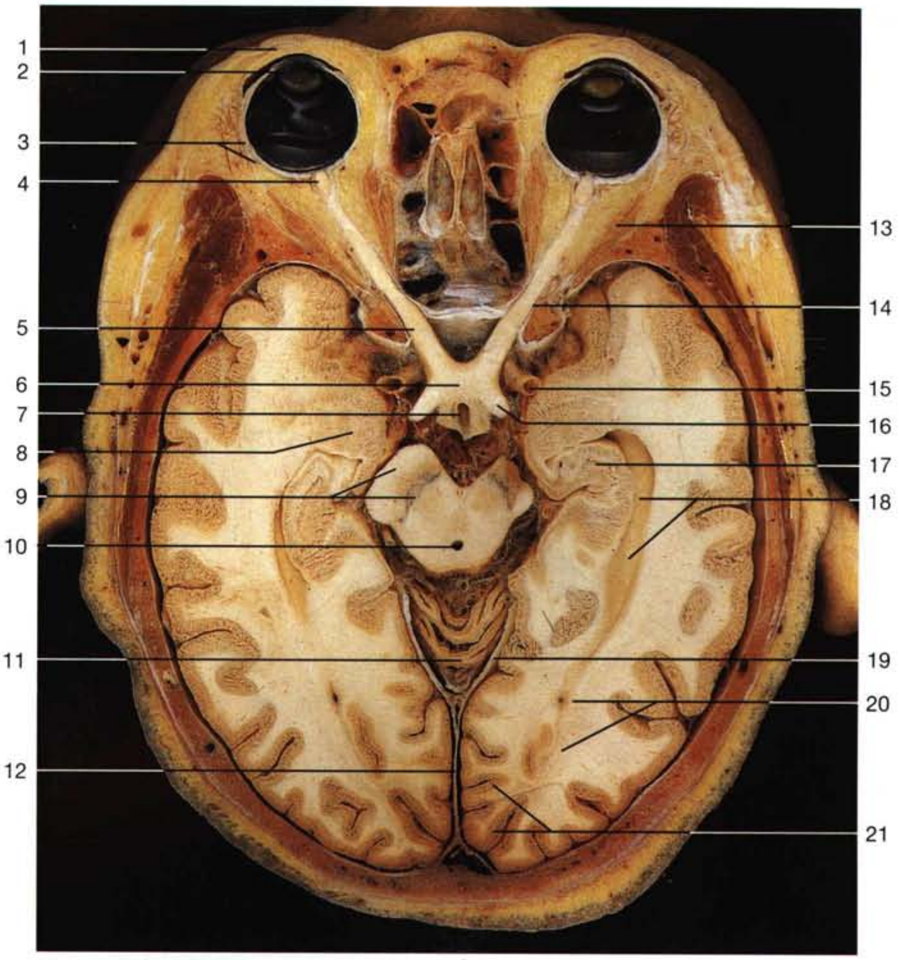

Horizontal section through the head at the level of optic chiasma and striate cortex (superior aspect). Note the relationship of hypothalamic infundibulum to optic chiasma.

1 Upper lid
2 Cornea
3 **Eyeball** (sclera, retina)
4 Head of optic nerve
5 Optic nerve
6 **Optic chiasma**
7 Infundibular recess of hypothalamus
8 Amygdaloid body
9 Substantia nigra and crus cerebri
10 Cerebral aqueduct
11 Vermis of cerebellum
12 Falx cerebri
13 Lateral rectus muscle
14 Optic canal
15 Internal carotid artery
16 **Optic tract**
17 Hippocampus
18 Inferior horn of lateral ventricle
19 Tentorium cerebelli
20 **Optic radiation** of Gratiolet
21 **Visual cortex (area calcarina, striate cortex)**
22 Lens
23 **Eyeball**
24 Ethmoidal cells
25 Optic nerve with dura sheath
26 Cerebral peduncle
27 Aqueduct of mesencephalon
28 Vermis of cerebellum

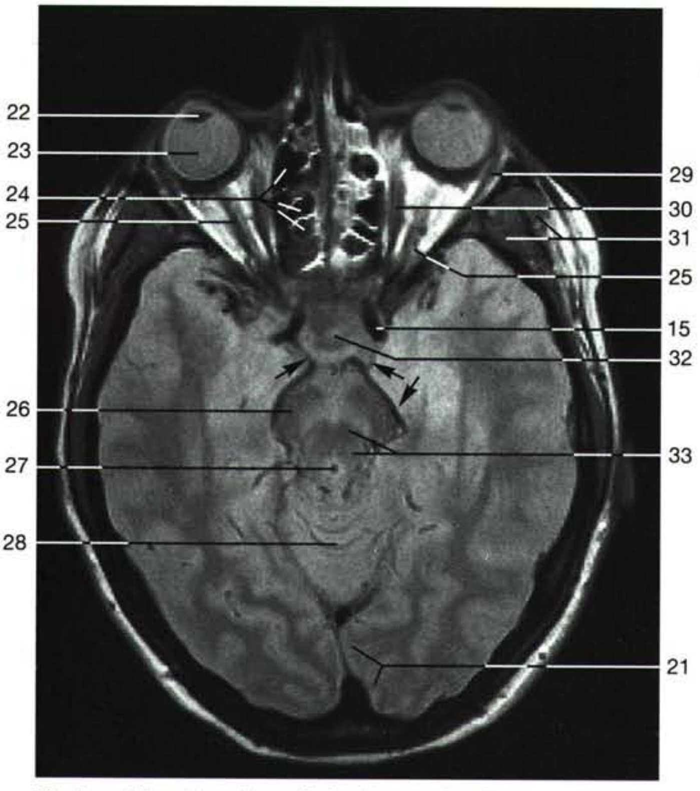

Horizontal section through the human head (MR-Scan, courtesy of Prof. W. J. Huk, Erlangen, FRG). Arrows = branches of arterial circle of Willis.

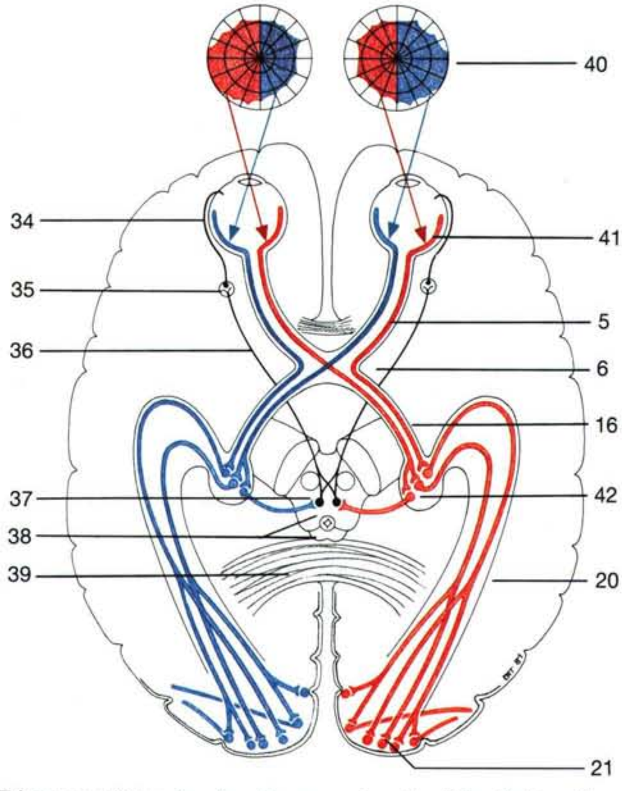

Diagram of the visual pathway and path of the light reflex.

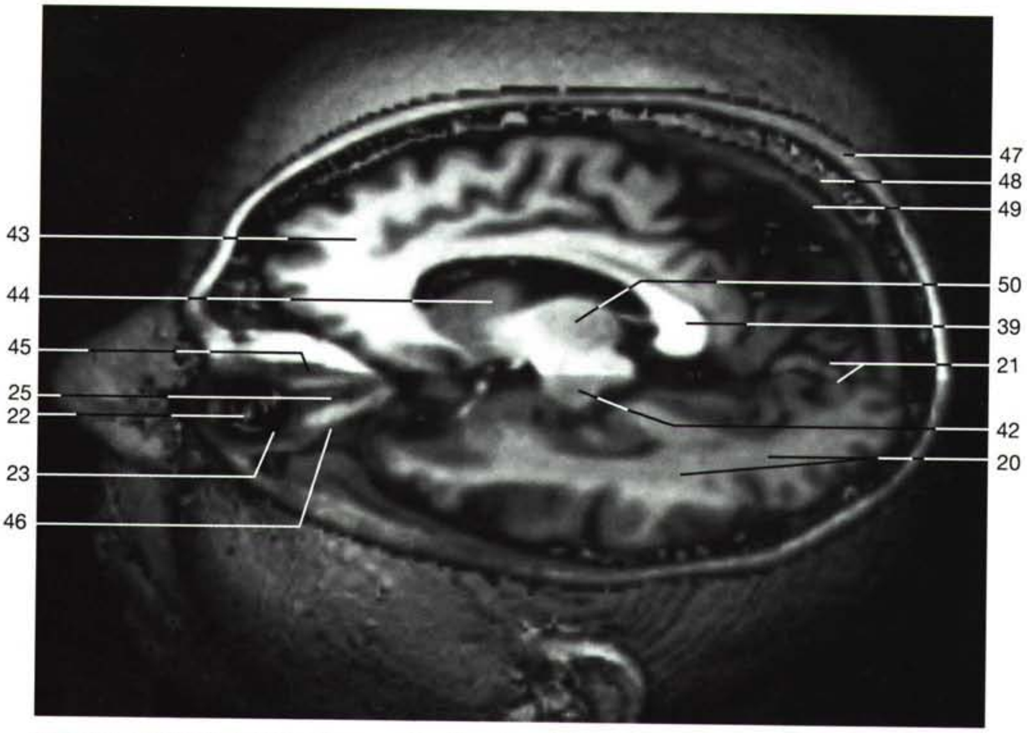

3-D-Reconstruction of the human visual system (MR-Image flash 40°, courtesy of Prof. W. J. Huk, University of Erlangen, FRG).

29	Lateral rectus muscle
30	Medial rectus muscle
31	Temporalis muscle
32	Hypophysis (pituitary gland)
33	Midbrain
34	Ciliary nerves (long and short)
35	Ciliary ganglion
36	Oculomotor nerve
37	Accessory oculomotor nucleus
38	Colliculi of midbrain
39	Corpus callosum
40	Visual field
41	Retina
42	**Lateral geniculate body**
43	Frontal lobe
44	**Caudate nucleus**
45	Medial rectus muscle
46	Lateral rectus muscle
47	Skin
48	Diploe (skull)
49	Dura mater
50	**Thalamus**
51	**Anterior cerebral artery**
52	Nucleus caudatus
53	Frontal sinus
54	**Internal capsule**
55	Lentiform nucleus (putamen)
56	**Hippocampus**
57	Temporal lobe of left hemisphere

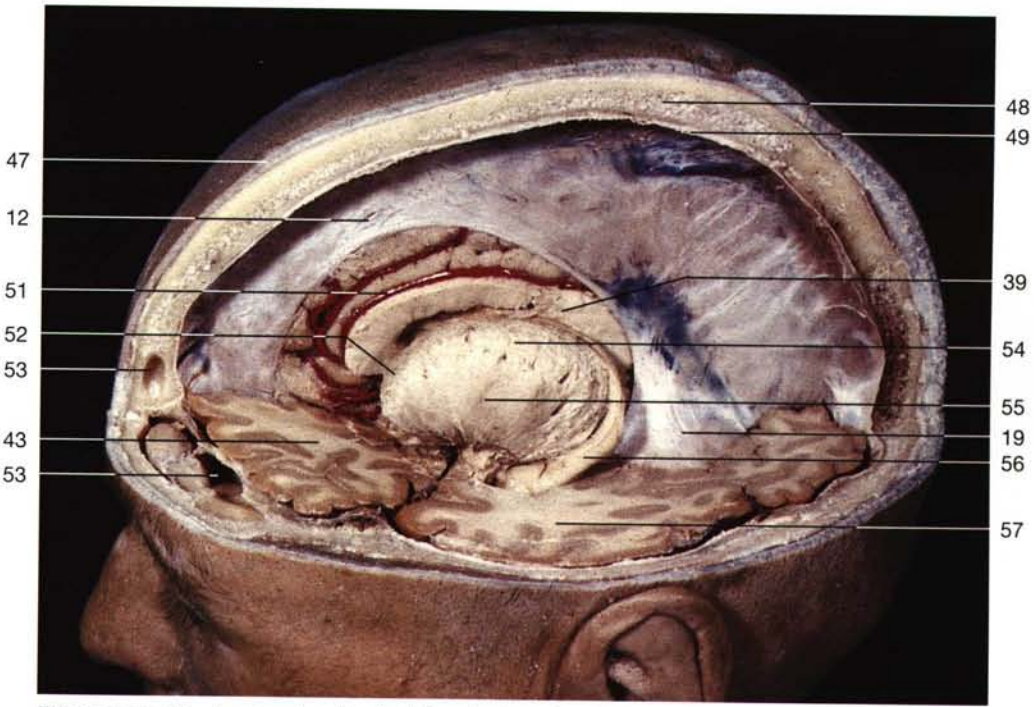

Dissection of brain stem in situ. Left hemisphere has been partly removed (compare with MR-Image above).

In **binocular vision** the visual field (40) is projected upon portions of both retinae (blue and red in the drawing). In the chiasma the fibers from the two retinal portions are combined to form the left optic tract. The fibers of the two eyes remain separated from each other throughout the entire visual pathway up to their final termination in the calcarine cortex (21). **Injuries** on the optic pathway produce visual defects whose nature depends on the location of the injury. Destruction of one optic nerve produces blindness in the corresponding eye with loss of pupillary light reflex. If lesions of the **chiasma** destroy the crossing fibers of the nasal portions of the retina, both temporal fields of vision are lost (bitemporal hemianopsia). If both lateral angles of the chiasma are compressed, the nondecussating fibers from the temporal retinae are affected resulting in loss of nasal visual fields (binasal hemianopsia). Lesions posterior to the chiasma (i. e., optic tract, lateral geniculate body, optic radiation or visual cortex) result in a loss of the entire opposite field of vision (homonymous hemianopsia).

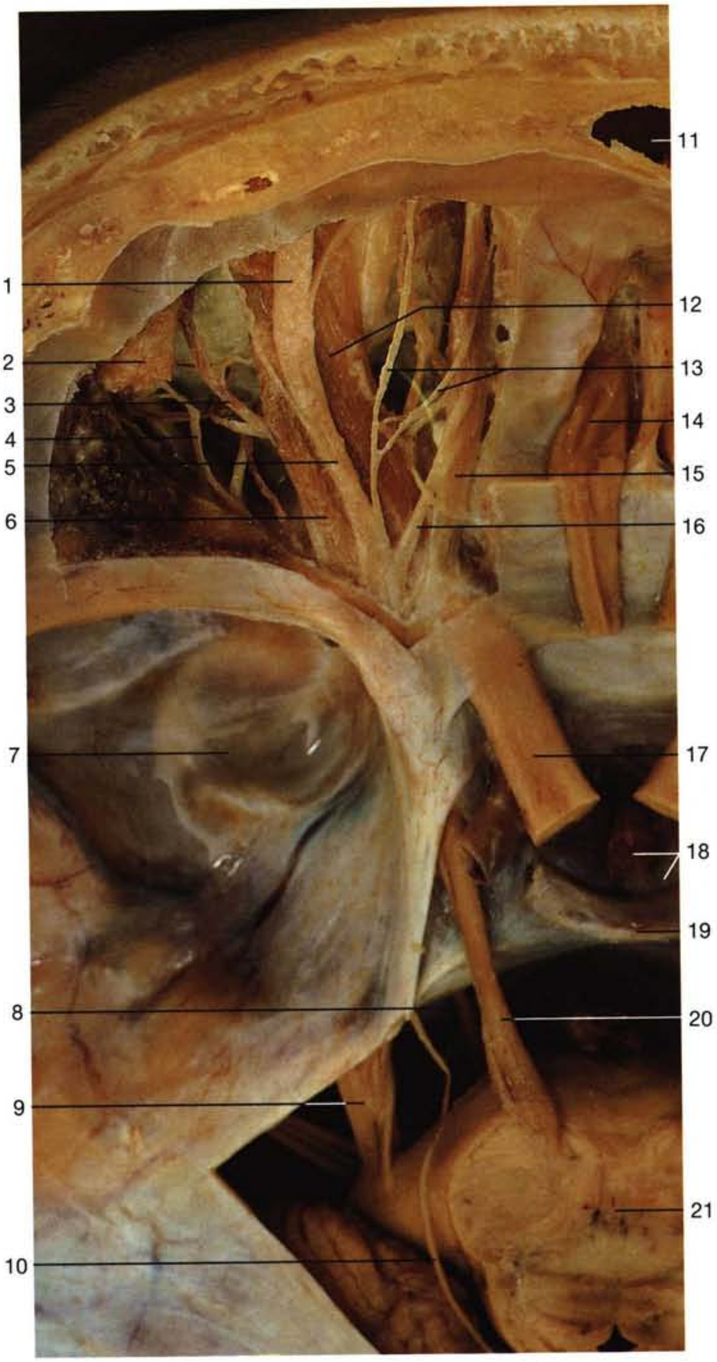

Superficial layer of the left orbit (superior aspect). The roof of the orbit and a portion of the left tentorium have been removed.

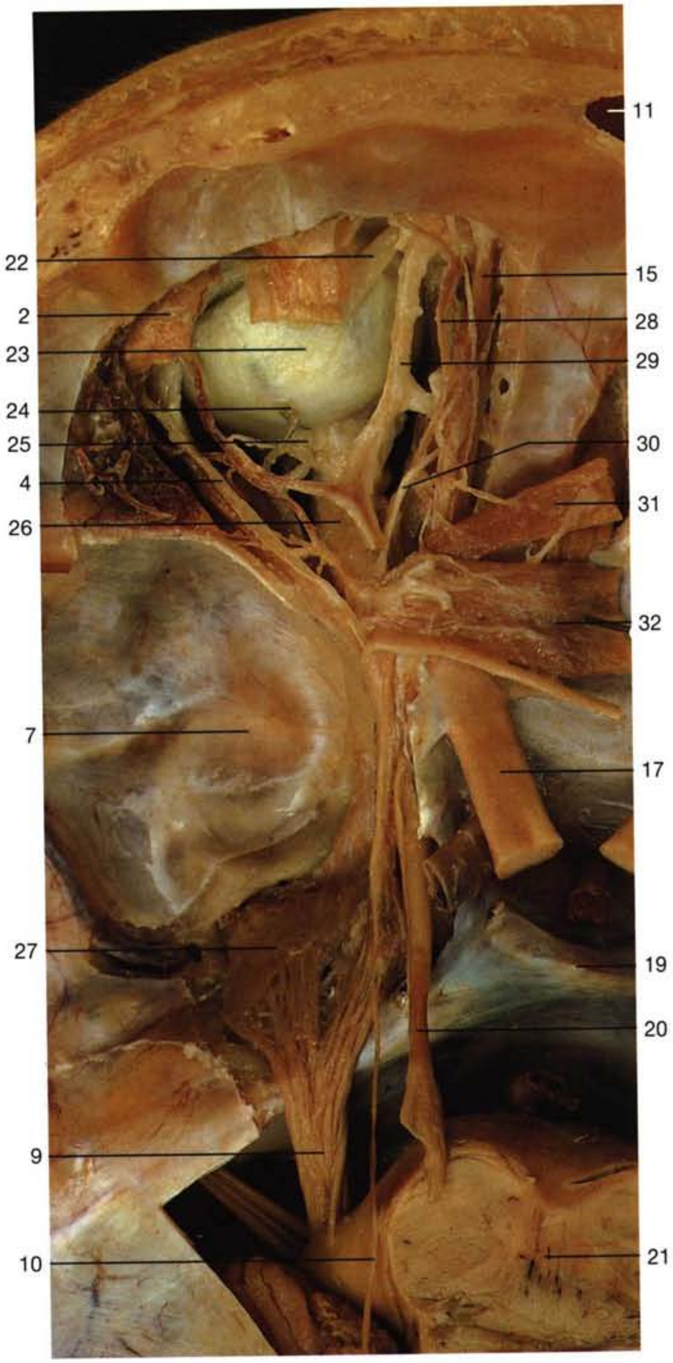

Middle layer of the left orbit (superior aspect). The roof of the orbit has been removed and the superior extraocular muscles have been divided and reflected.

1 Lateral branch of frontal nerve
2 Lacrimal gland
3 Lacrimal vein
4 Lacrimal nerve
5 **Frontal nerve**
6 Superior rectus
7 Middle cranial fossa
8 **Abducens nerve** (n. VI)
9 Trigeminal nerve (n. V)
10 **Trochlear nerve** (intracranial part) (n. IV)
11 Frontal sinus
12 Levator palpebrae superioris muscle
13 Branches of supratrochlear nerve
14 Olfactory bulb
15 Superior oblique muscle
16 **Trochlear nerve** (intraorbital part) (n. IV)
17 **Optic nerve** (intracranial part)
18 Pituitary gland and infundibulum
19 Dorsum sellae
20 **Oculomotor nerve** (n. III)
21 Midbrain
22 Tendon of superior oblique muscle
23 Eyeball
24 Vena vorticosa
25 Short ciliary nerves
26 Optic nerve (extracranial part)
27 **Trigeminal ganglion**

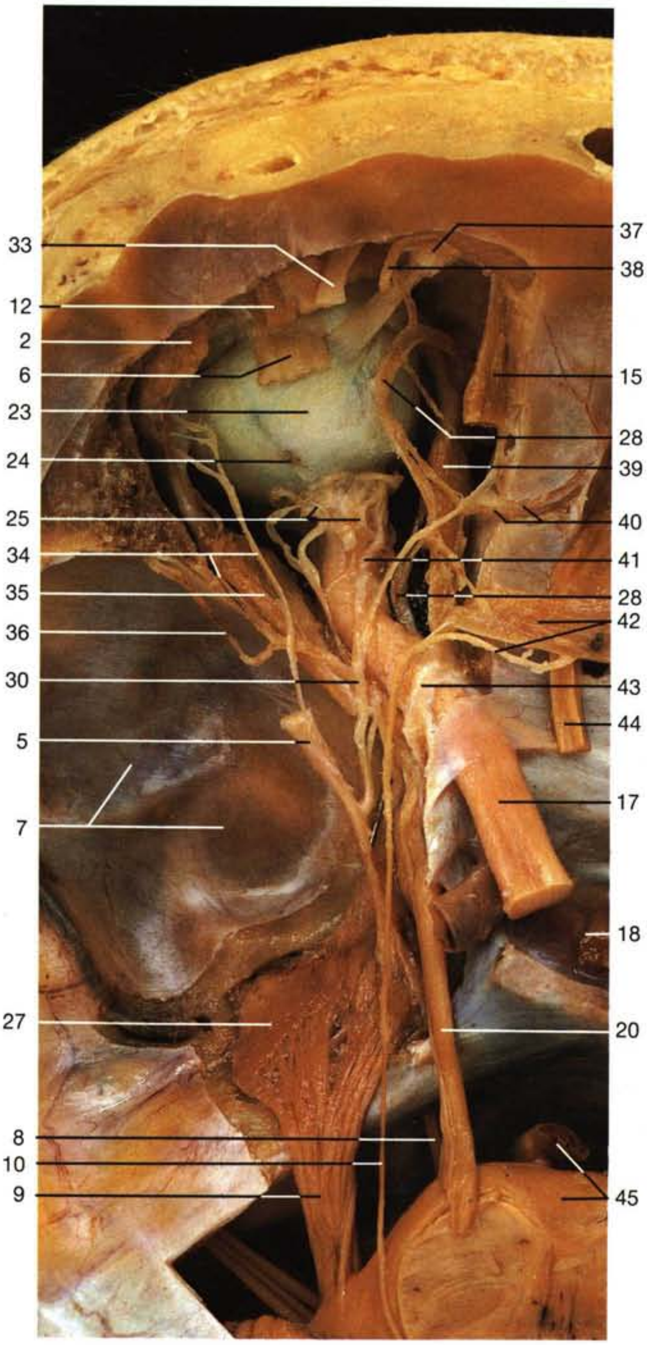
33
12
2
6
23
24
25
34
35
36
30
5
7
27
8
10
9
37
38
15
28
39
40
41
28
42
43
44
17
18
20
45

Middle layer of the left orbit (superior aspect). The roof of the orbit and the superior extraocular muscles have been removed.

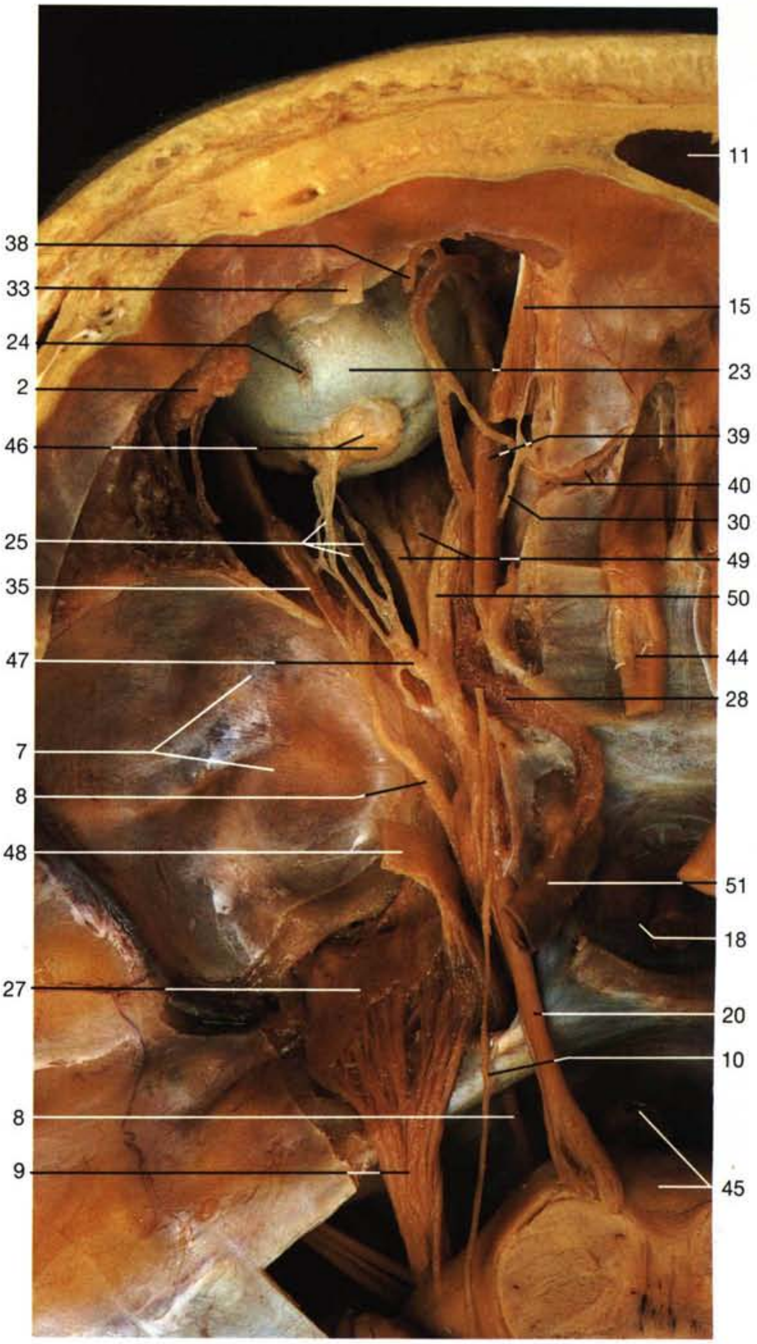
38
33
24
2
46
25
35
47
7
8
48
27
8
9
11
15
23
39
40
30
49
50
44
28
51
18
20
10
45

Deeper layer of the left orbit (superior aspect). The optic nerve has now been removed.

28 **Ophthalmic artery**
29 Superior ophthalmic vein
30 **Nasociliary nerve**
31 Levator palpebrae superioris muscle (reflected)
32 Superior rectus muscle (reflected)
33 Lateral branch of supraorbital nerve
34 Lacrimal nerve and artery
35 Lateral rectus muscle
36 Meningolacrimal artery (anastomosing with middle meningeal artery)
37 Trochlea
38 Medial branch of supraorbital nerve
39 Medial rectus muscle
40 Anterior ethmoidal artery and nerve
41 Long ciliary nerve
42 Superior oblique muscle and trochlear nerve
43 Common tendinous ring
44 Olfactory tract
45 Basilar artery and pons
46 Optic nerve (external sheath of optic nerve, divided)
47 **Ciliary ganglion**
48 Ophthalmic nerve (divided, reflected)
49 Inferior branch of oculomotor nerve and inferior rectus muscle
50 Superior branch of oculomotor nerve
51 Internal carotid artery

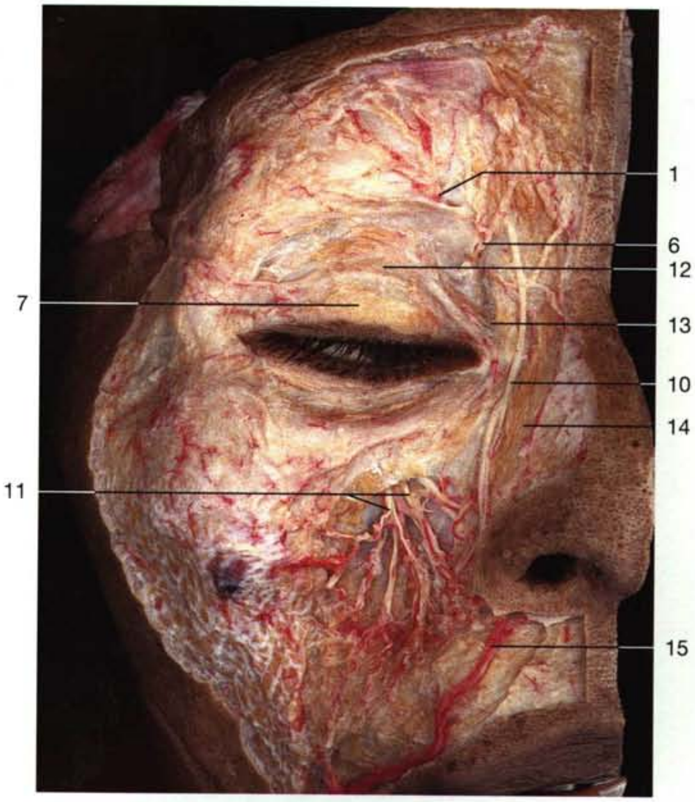

Eyelids (superficial layer, right side).

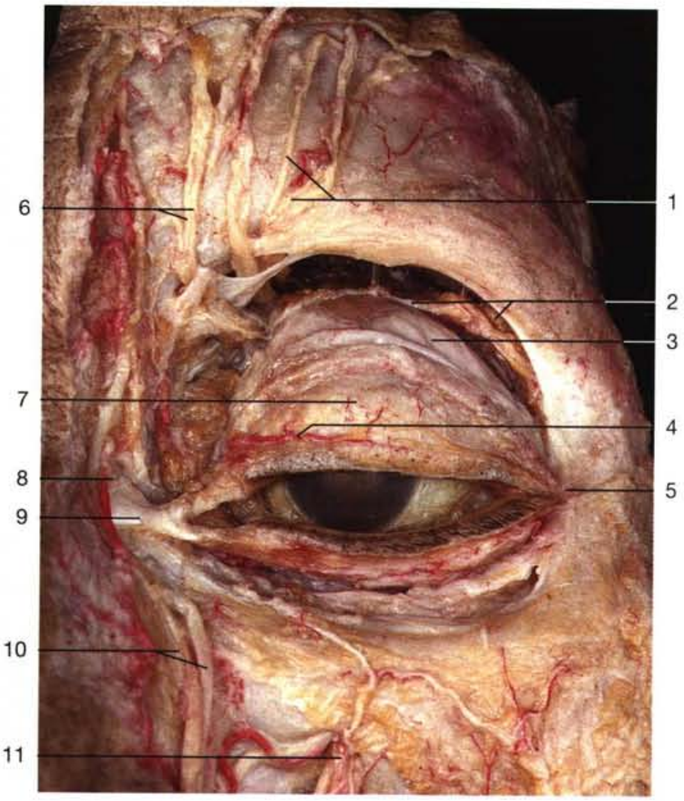

Eyelids (deeper layer, left side).

1 Supraorbital artery, supraorbital nerve, lateral branch
2 **Lacrimal gland**
3 Aponeurosis of levator palpebrae superioris muscle
4 Arterial arch of upper eyelid
5 Lateral palpebral ligament
6 Supratrochlear artery, supratrochlear nerve
7 Upper eyelid, tarsal plate
8 Lacrimal sac
9 **Medial palpebral ligament**
10 Angular artery and vein
11 Infraorbital artery, vein and nerve
12 Orbicularis oculi muscle
13 Infratrochlear nerve
14 Levator labii superioris alaeque nasi muscle
15 Facial artery and vein
16 **Superior lacrimal canaliculus**
17 **Inferior lacrimal canaliculus**
18 Lacrimal papilla and punctum
19 Lacrimal bone
20 **Lacrimal duct**
21 Mucous membrane of nasal cavity
22 Palpebral conjunctiva of lower eyelid
23 Lacrimal sac and superior lacrimal canaliculus
24 Lateral fixation of levator aponeurosis
25 Infraorbital foramen

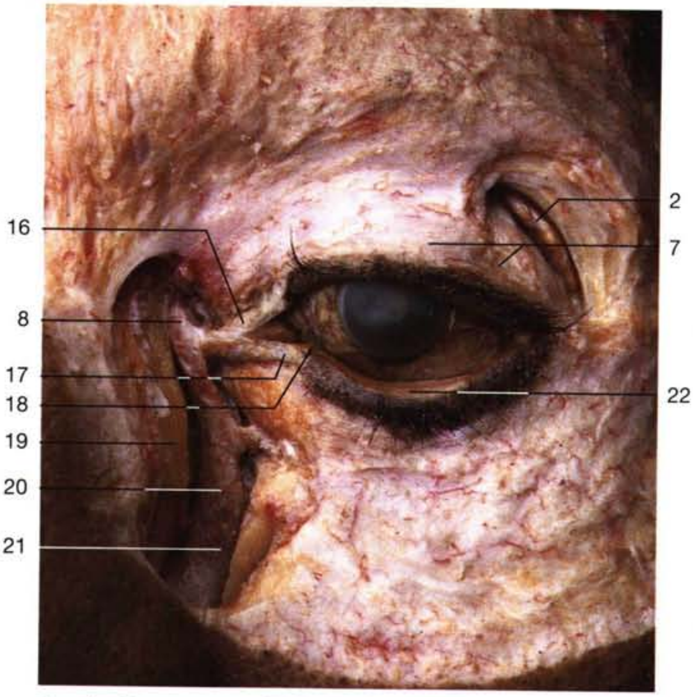

Lacrimal apparatus of left eye (anterior aspect).

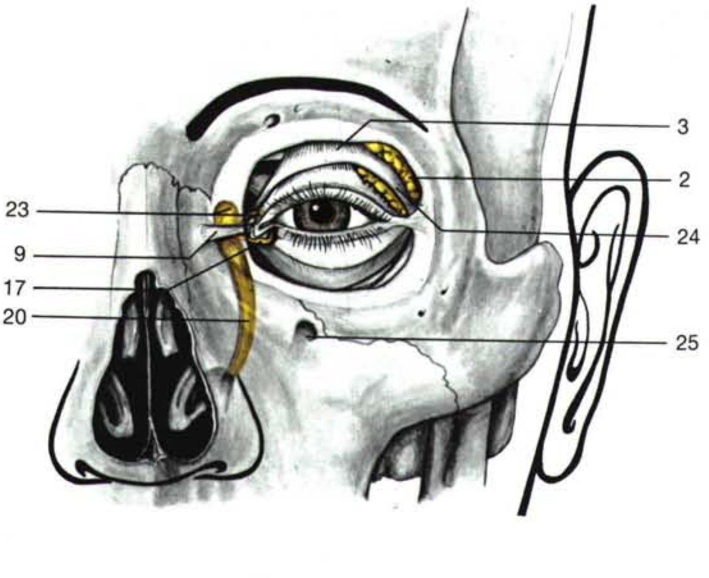

Lacrimal apparatus of left eye (anterior aspect).

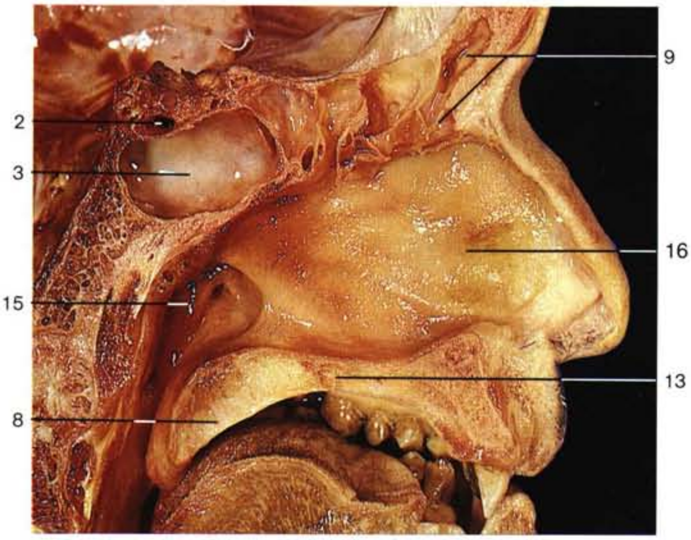

Nasal septum, covered by a mucous membrane.

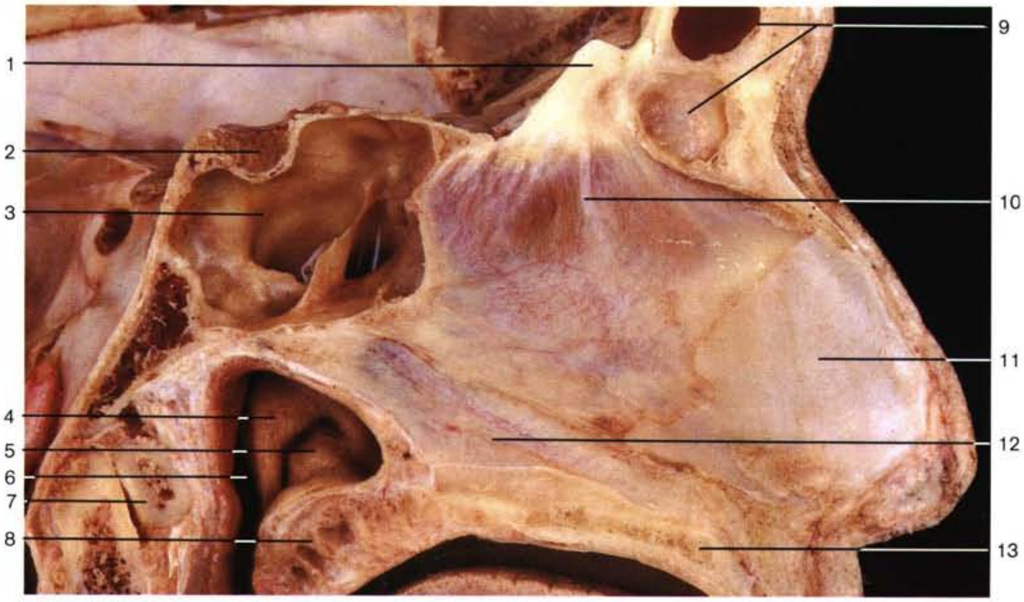

Nasal septum. Mucous membrane removed.

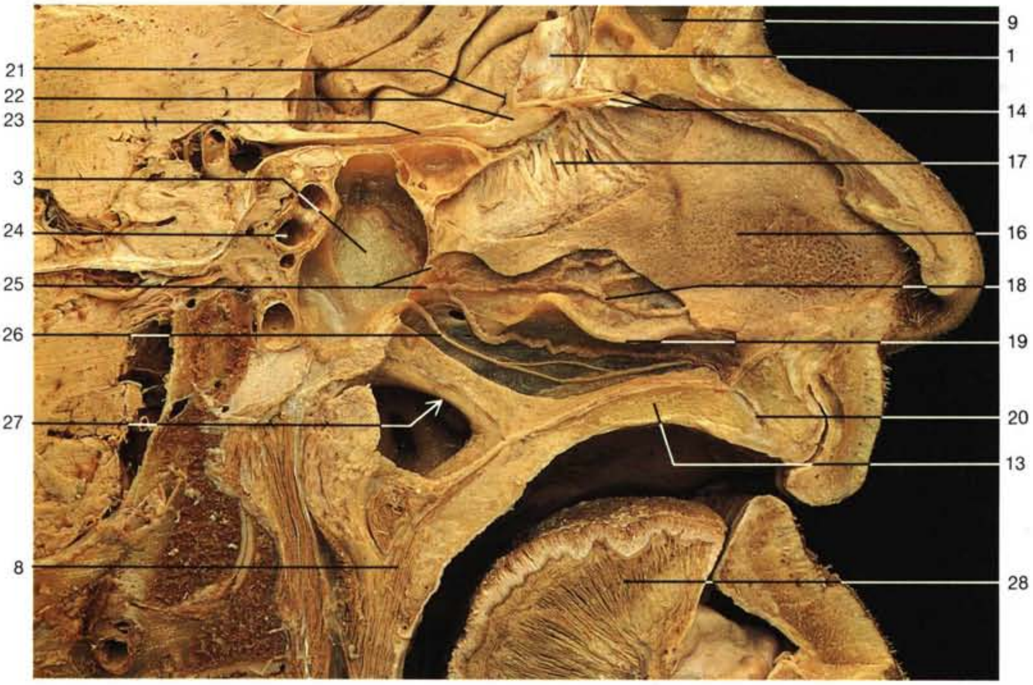

Nasal septum. Dissection of nerves and vessels.

1	Crista galli
2	Pituitary gland and sella turcica
3	**Sphenoidal sinus** (relatively large)
4	Tubal elevation
5	Pharyngeal opening of the **auditory tube**
6	Pharyngeal recess
7	Atlas (anterior arch)
8	Soft palate
9	Frontal sinus
10	Perpendicular plate of ethmoid
11	**Cartilage of nasal septum**
12	Vomer
13	Hard palate
14	Nasal branch of anterior ethmoidal artery and anterior ethmoidal nerve
15	Nasopharynx
16	Nasal septum
17	**Olfactory nerves**
18	Septal artery
19	Crest of nasal septum
20	Incisive canal
21	Anterior ethmoidal artery
22	Olfactory bulb
23	Olfactory tract
24	Internal carotid artery
25	Posterior nasal and septal arteries
26	Nasopalatine nerve
27	**Choana** (arrow)
28	Tongue

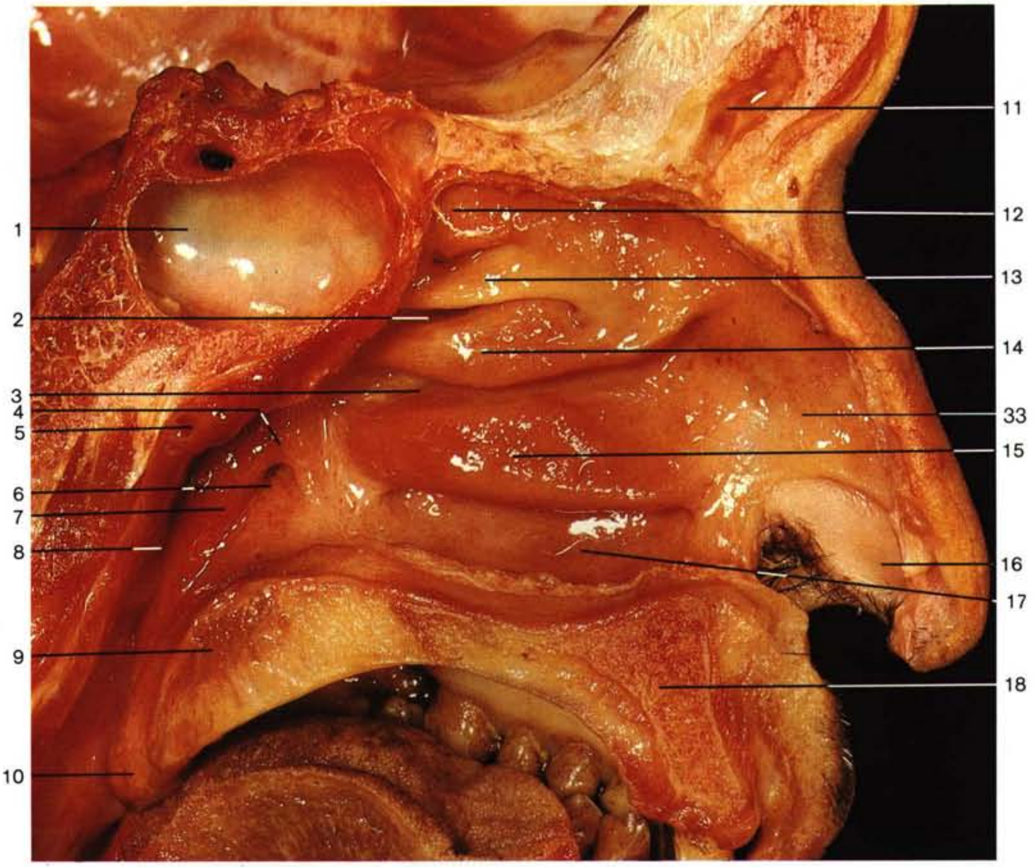

Lateral wall of the nasal cavity, septum removed.

1 Sphenoidal sinus
2 Superior meatus
3 Middle meatus
4 Tubal elevation
5 Pharyngeal tonsil
6 Pharyngeal orifice of **auditory tube**
7 Salpingopharyngeal fold
8 Pharyngeal recess
9 **Soft palate**
10 **Uvula**
11 Frontal sinus
12 Sphenoethmoidal recess
13 **Superior nasal concha**
14 **Middle nasal concha**
15 **Inferior nasal concha**
16 Vestibule
17 Inferior meatus
18 **Hard palate**
19 Grooves for the middle meningeal artery and parietal bone (light green)
20 **Maxillary hiatus**
21 Perpendicular process of palatine bone
22 Openings of ethmoidal air cells
23 Opening of frontal sinus
24 Medial pterygoid plate (red)
25 Horizontal plate of palatine process
26 **Ethmoidal air cells**
27 **Maxillary sinus**
28 Nasal septum
29 Pterygoid hamulus
30 Nasal bone (white)
31 Frontal process of maxilla (violet)
32 Palatine process of maxilla (violet)
33 Nasal atrium

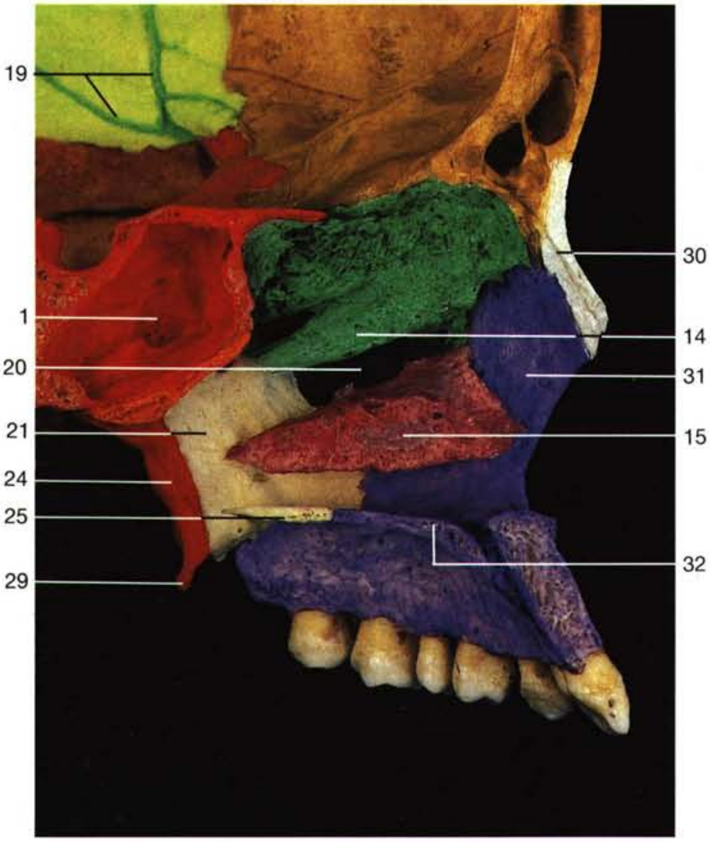

Bones of left nasal cavity, medial aspect.

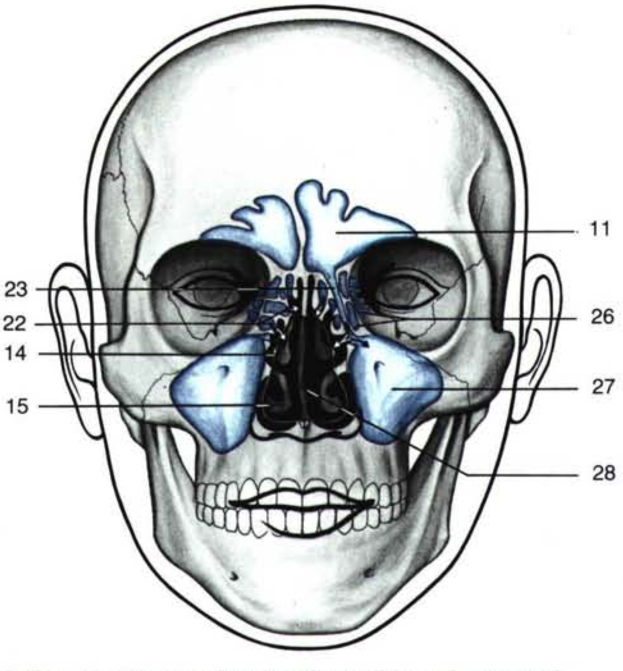

Schematic diagram showing the position of paranasal sinuses, openings indicated by arrows.

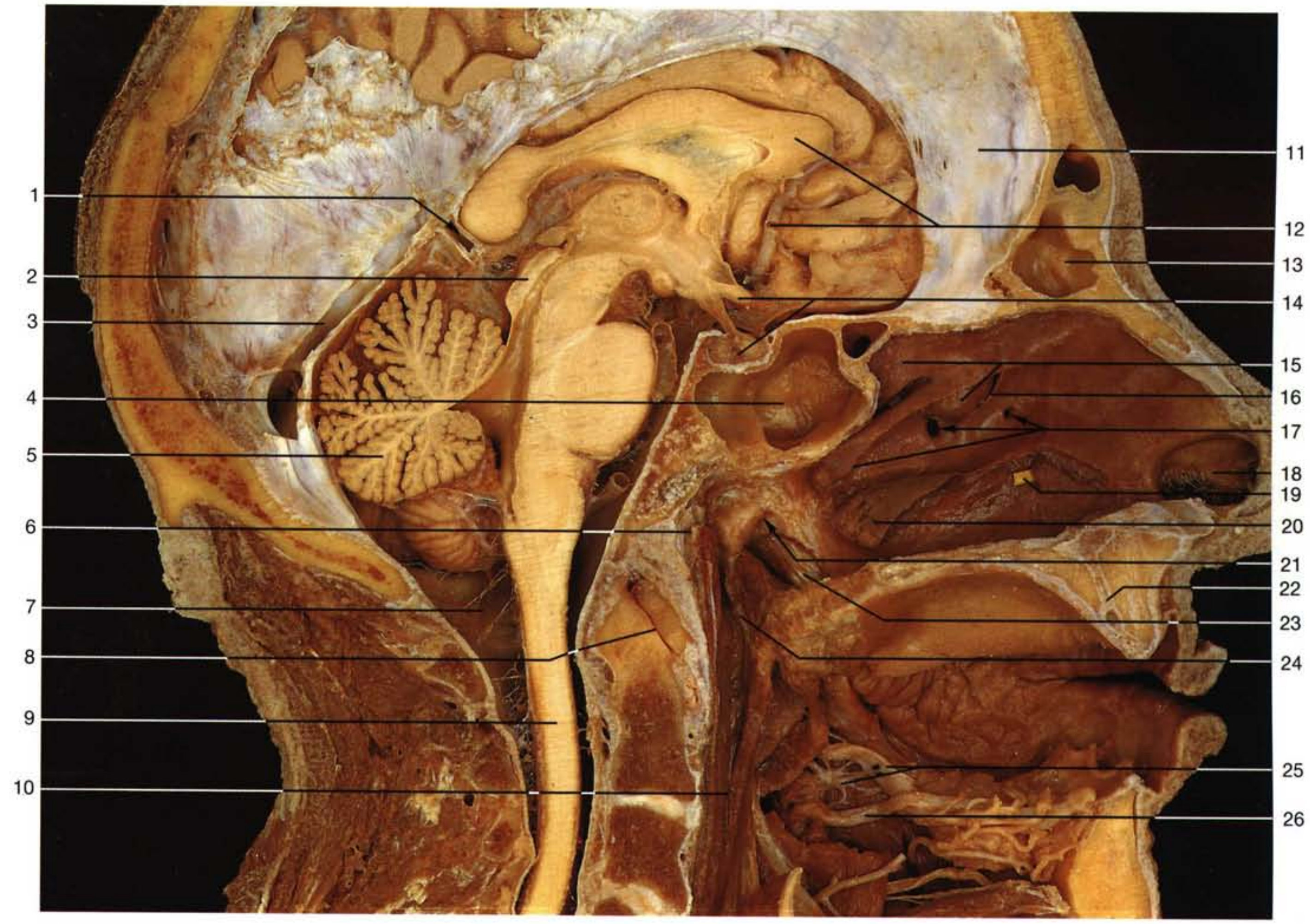

Median section through the head with nasal and oral cavity. The middle and inferior nasal conchae have been partly removed to show the openings of paranasal sinuses.

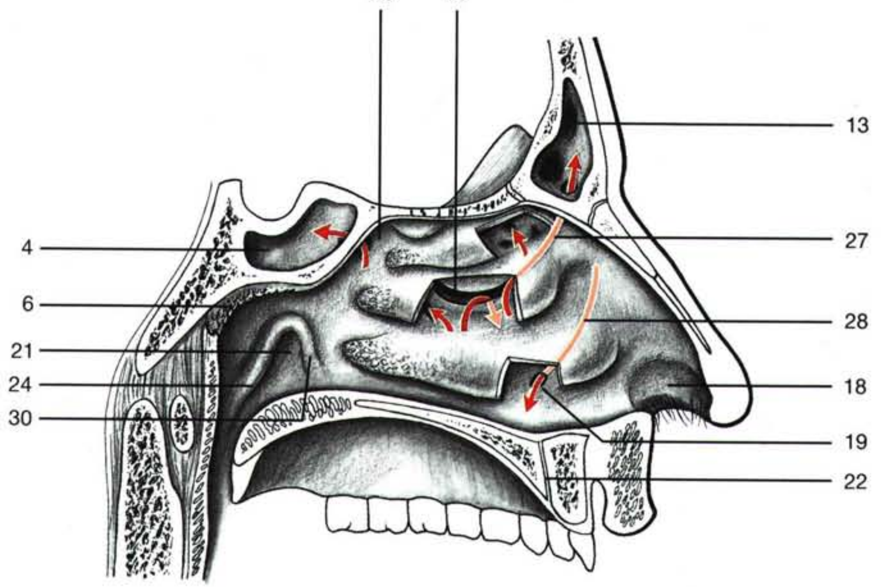

Lateral wall of nasal cavity. Openings indicated by red arrows. (Schematic drawing.)

1 Great cerebral vein (Galen's vein)
2 **Tectum of midbrain**
3 Straight sinus
4 **Sphenoidal sinus**
5 Cerebellum
6 **Pharyngeal tonsil**
7 **Cerebellomedullary cistern**
8 Median atlantoaxial joint
9 Spinal cord
10 Oral part of pharynx
11 Falx cerebri
12 Corpus callosum and anterior cerebral artery
13 **Frontal sinus**
14 Optic chiasm and pituitary gland
15 Superior nasal concha and ethmoidal bulla
16 **Semilunar hiatus**
17 Accessory **openings to maxillary sinus** and cut edge of middle nasal concha
18 Vestibule
19 **Opening of nasolacrimal duct**
20 Inferior nasal concha (cut)
21 **Opening of auditory tube**
22 **Incisive canal**
23 Levator veli palatini muscle
24 Salpingopharyngeal fold
25 Lingual nerve and submandibular ganglion
26 Submandibular duct
27 **Nasofrontal duct**
28 **Nasolacrimal duct**
29 Sphenoethmoidal recess (of Rosenmüller)
30 Salpingopalatine fold

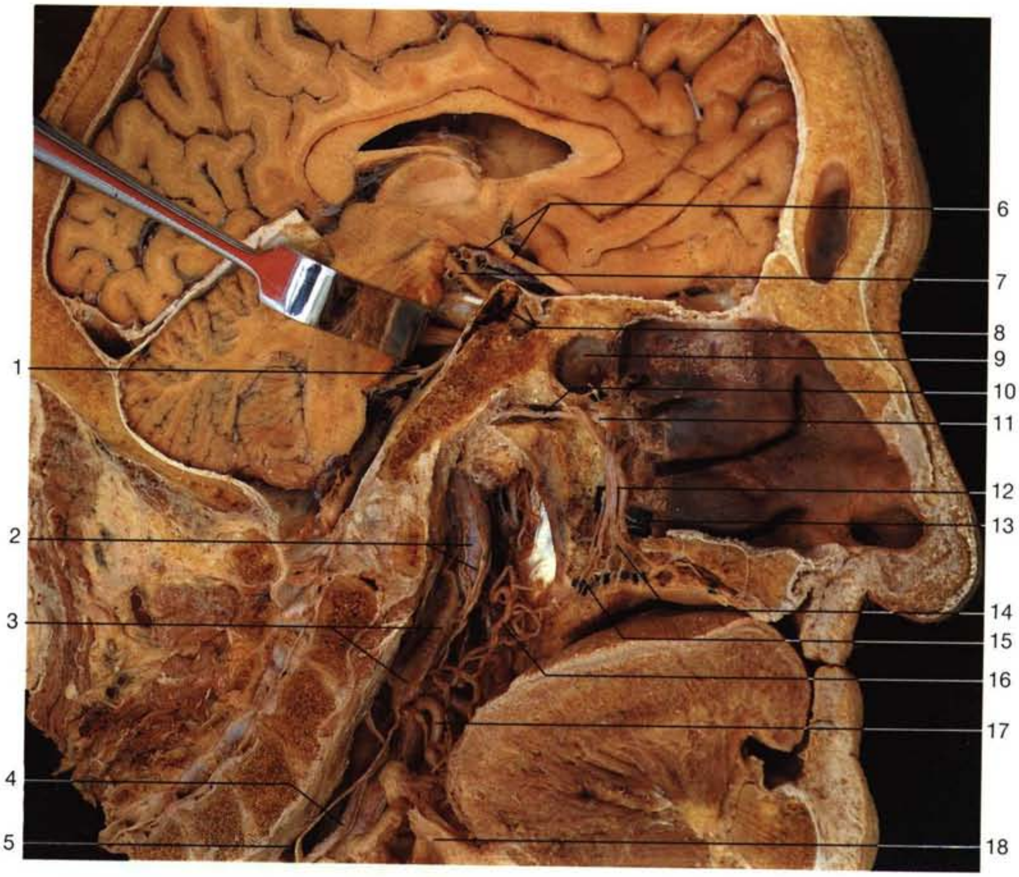

Nerves of the lateral wall of nasal cavity I. Sagittal section through the head. Mucous membranes partly removed, pterygoid canal opened.

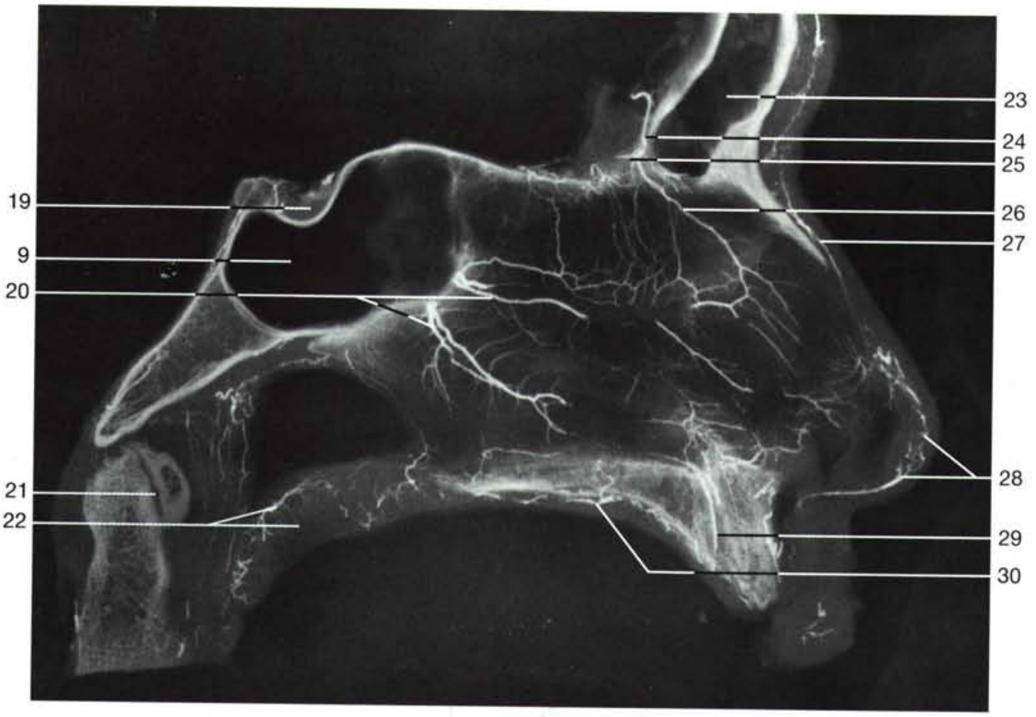

Arteriogram of the nasal septum, left side (lateral aspect).

1 Facial nerve
2 **Internal carotid artery** and **internal carotid plexus**
3 **Superior cervical ganglion**
4 Vagus nerve
5 Sympathetic trunk
6 Optic nerve and ophthalmic artery
7 Oculomotor nerve
8 Internal carotid artery and cavernous sinus
9 Sphenoidal sinus
10 Nerve of the pterygoid canal
11 **Pterygopalatine ganglion**
12 **Descending palatine artery**
13 Lateral inferior posterior nasal branches and lateral posterior nasal and septal arteries
14 Greater palatine nerves and artery
15 Lesser palatine nerves and arteries
16 Branches of ascending pharyngeal artery
17 Lingual artery
18 Epiglottis
19 **Sella turcica**
20 **Posterior lateral nasal and septal arteries** (branches of sphenopalatine artery)
21 Median atlanto-axial joint
22 Soft palate and lesser palatine arteries (branches of descending palatine artery)
23 Frontal sinus
24 Anterior meningeal artery
25 Anterior ethmoidal artery (branch of ophthalmic artery)
26 **Septal branch of anterior ethmoidal artery**
27 Dorsal nasal artery
28 Nasal branches of anterior ethmoidal artery
29 Incisive canal with nasopalatine artery
30 Hard palate and greater palatine artery (branch of descending palatine artery)
31 Tentorium cerebelli
32 Trochlear nerve
33 **Trigeminal nerve** with motor root
34 **Internal carotid plexus**
35 **Lingual nerve with chorda tympani**
36 Medial pterygoid muscle and medial pterygoid plate
37 **Inferior alveolar nerve**
38 **Sympathetic trunk**
39 Oculomotor nerve
40 **Palatine nerves**
41 Tongue
42 **Tringeminal ganglion**
43 Trigeminal nerve (n. V)
44 **Facial nerve** (n. VII)
45 Geniculate ganglion
46 Stylomastoid foramen
47 Medial pterygoid muscle

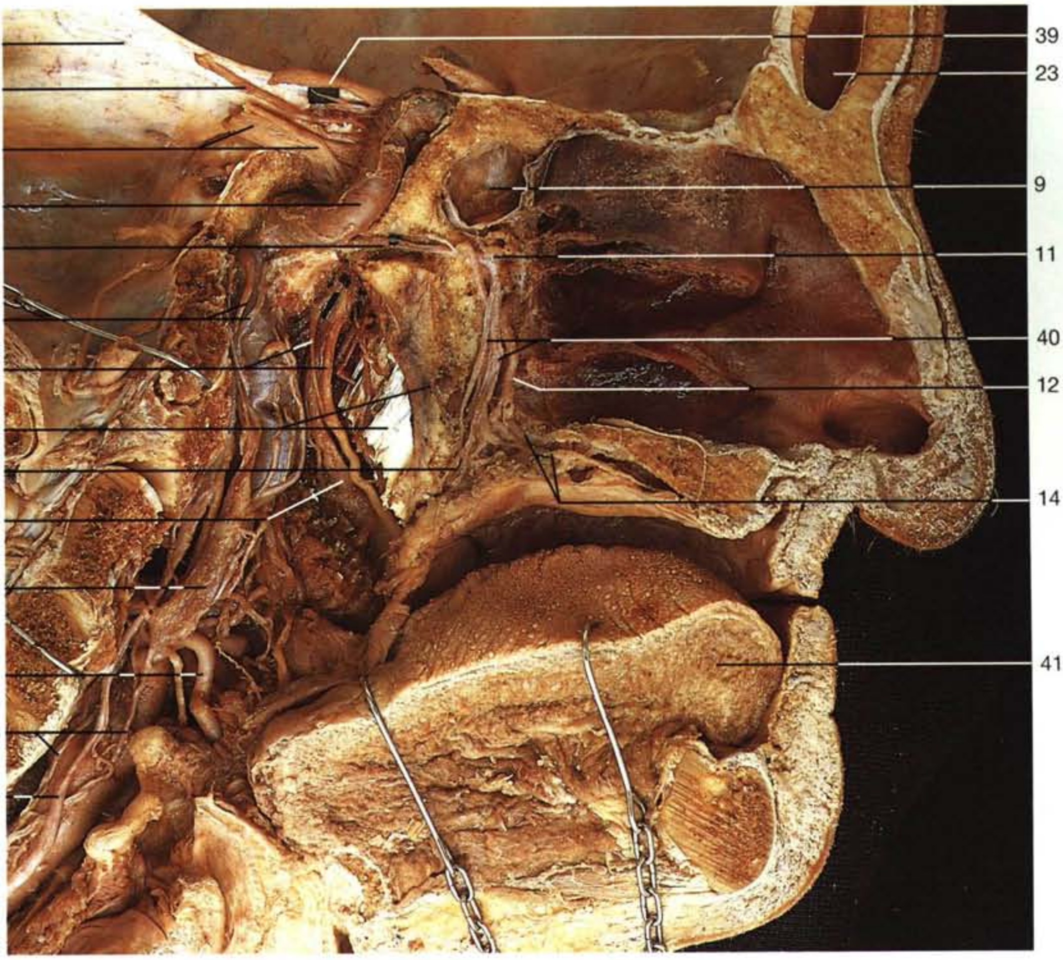
31
32
33
2
10
34
35
36
15
37
3
17
38
4
39
23
9
11
40
12
14
41

Nerves of the lateral wall of nasal cavity II. Carotid canal opened, mucous membranes of pharynx and nasal cavity partly removed.

48 **Greater petrosal nerve**
49 Maxillary nerve
50 Olfactory bulb
51 **Olfactory nerves**
52 Internal nasal branches of anterior ethmoidal nerve
53 Lateral superior posterior nasal branches
54 Lateral inferior posterior nasal branches
55 Incisive canal with nasopalatine nerve
56 **Greater palatine nerve**
57 **Deep petrosal nerve**
58 Mandibular nerve
59 Nasal cavity and inferior nasal concha
60 Opening of auditory tube
61 Tensor veli palatini muscle
62 Levator veli palatini muscle
63 Pharyngeal recess in the nasopharynx
64 Uvula
65 Palatoglossal arch
66 Tonsillar branch of ascending palatine artery
67 **Palatine tonsil**
68 Palatopharyngeal arch

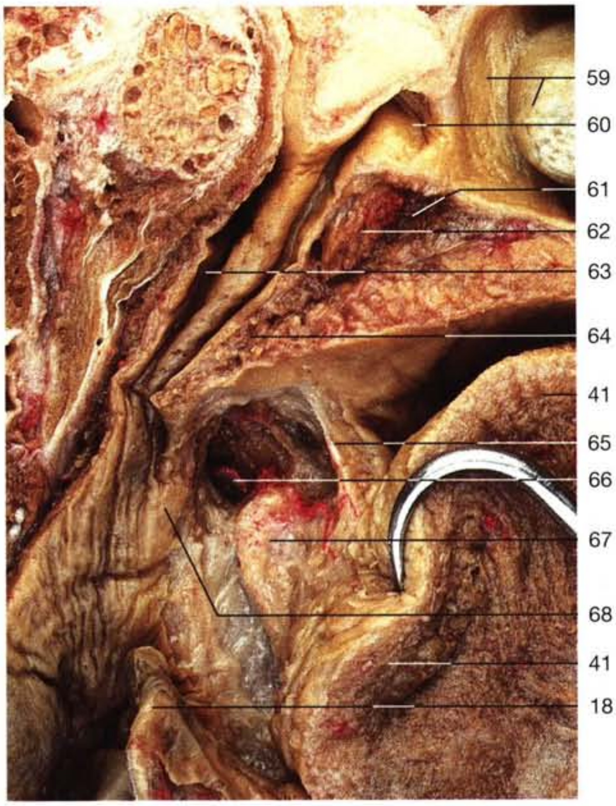
59
60
61
62
63
64
41
65
66
67
68
41
18

Dissection of palatine tonsil located in the lateral wall of the nasopharynx (left side). Root of tongue reflected.

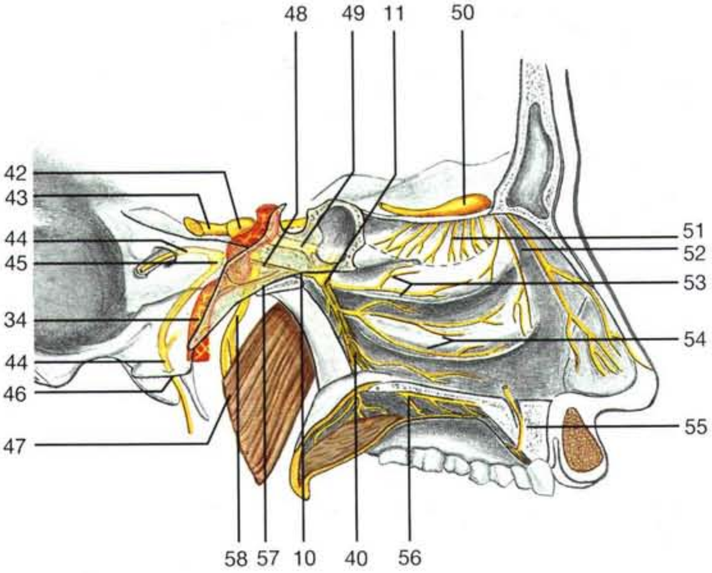
48
49
11
50
42
43
44
45
34
44
46
47
51
52
53
54
55
58
57
10
40
56

Nerves of the lateral wall of nassal cavity. Body of sphenoid bone appears transparent (schematic drawing).

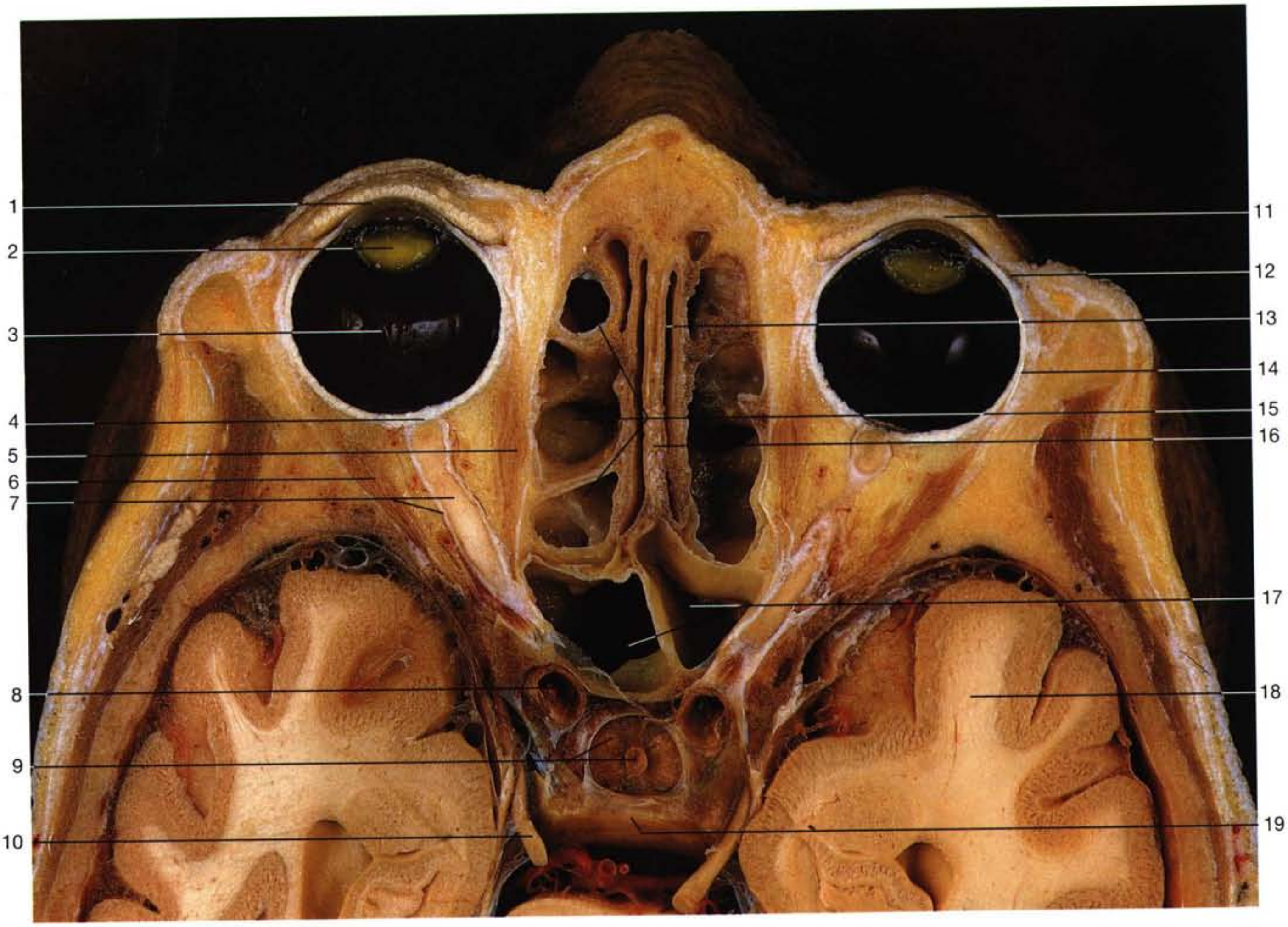

Horizontal section through the nasal cavity, the orbits and temporal lobes of the brain at the level of pituitary gland.

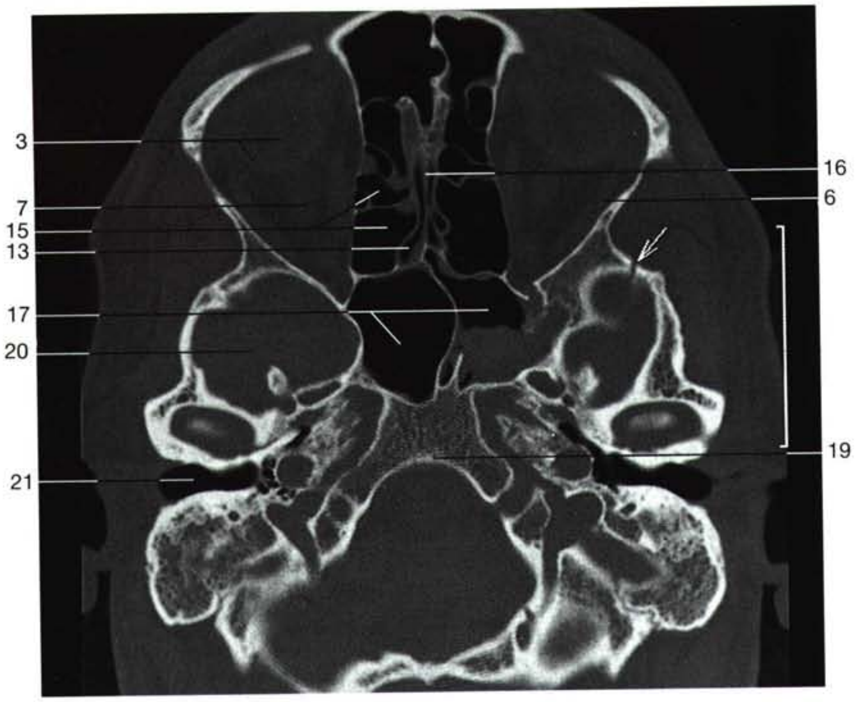

Horizontal section through the head. CT-Scan. Bar = 2 cm.
Arrow: fracture.

1 Cornea
2 Lens
3 Vitreous body (**eyeball**)
4 Head of optic nerve
5 Medial rectus muscle
6 Lateral rectus muscle
7 **Optic nerve** with dural sheath
8 Internal carotid artery
9 **Pituitary gland** and **infundibulum**
10 **Oculomotor nerve**
11 Superior tarsal plate of eye lid
12 Fornix of conjunctiva
13 **Nasal cavity**
14 Sclera
15 Ethmoidal sinus
16 **Nasal septum**
17 **Sphenoidal sinus**
18 Temporal lobe
19 Clivus
20 Middle cranial fossa
21 External acoustic meatus
22 Superior sagittal sinus
23 Falx cerebri
24 Superior rectus and levator palpebrae superioris muscles
25 Eyeball and lacrimal gland
26 Inferior rectus muscle
27 Zygomatic bone
28 **Maxillary sinus**
29 Inferior nasal concha
30 Hard palate
31 Superior longitudinal muscle of tongue
32 Median fibrosus septum of tongue

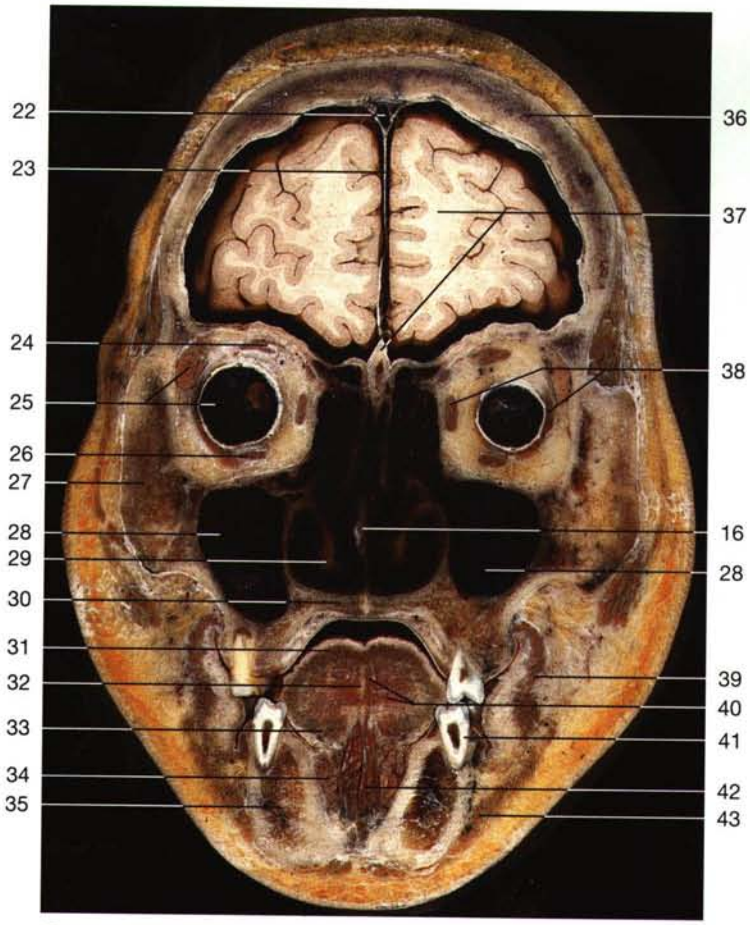
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
16
28
39
40
41
42
43

Coronal section through the head at the level of the second premolar of the mandible.

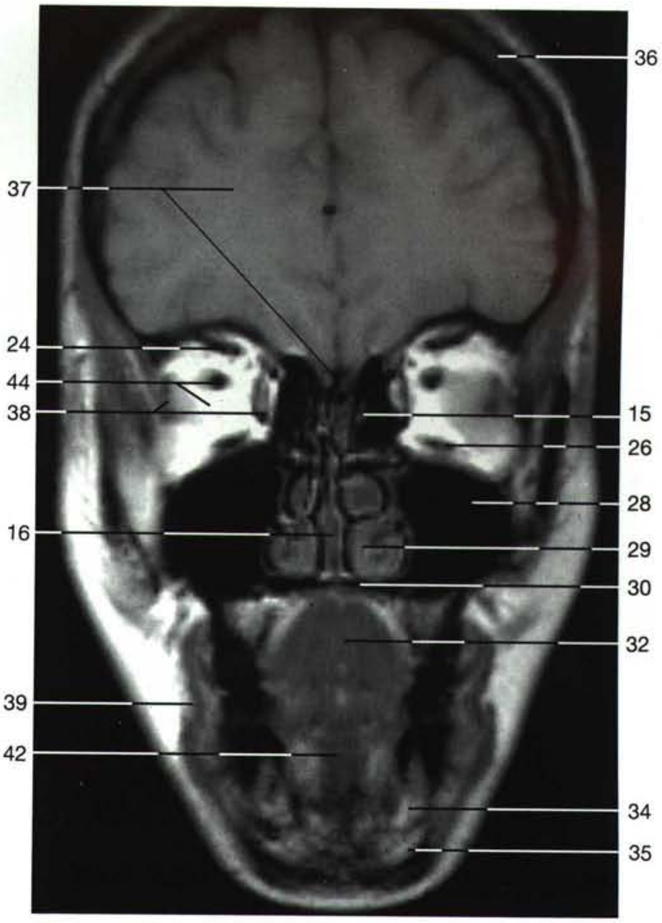
36
37
24
44
38
15
26
28
16
29
30
32
39
42
34
35

Coronal section through the head (MR-Scan, courtesy of Dr. A. Heuck, Munich). Note the situation of the head cavities.

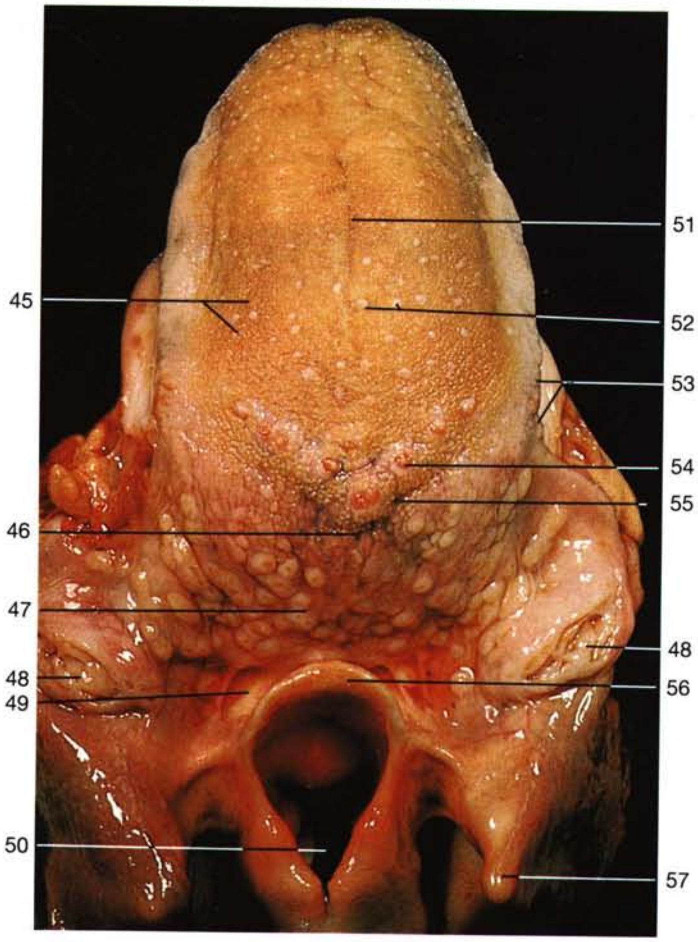
51
45
52
53
54
55
46
47
48
48
49
56
50
57

33 Inferior longitudinal muscle of tongue
34 **Sublingual gland**
35 Mandible
36 Calvaria
37 Frontal lobe of brain and crista galli
38 Lateral and medial eye rectus muscles
39 Buccinator muscle
40 Vertical and transverse muscles of tongue
41 Second premolar of mandible
42 Genioglossus muscle
43 Platysma muscle
44 **Orbit** and **optic nerve**
45 Filiform papillae
46 **Foramen cecum**
47 **Root of tongue** (lingual tonsil)
48 **Palatine tonsil**
49 Vallecula of epiglottis
50 Vestibule of larynx
51 Median sulcus of tongue
52 Fungiform papillae
53 Foliate papillae
54 **Circumvallate papilla**
55 Sulcus terminalis
56 Epiglottis
57 Greater cornu of hyoid bone

Dorsal surface of the tongue and laryngeal inlet.

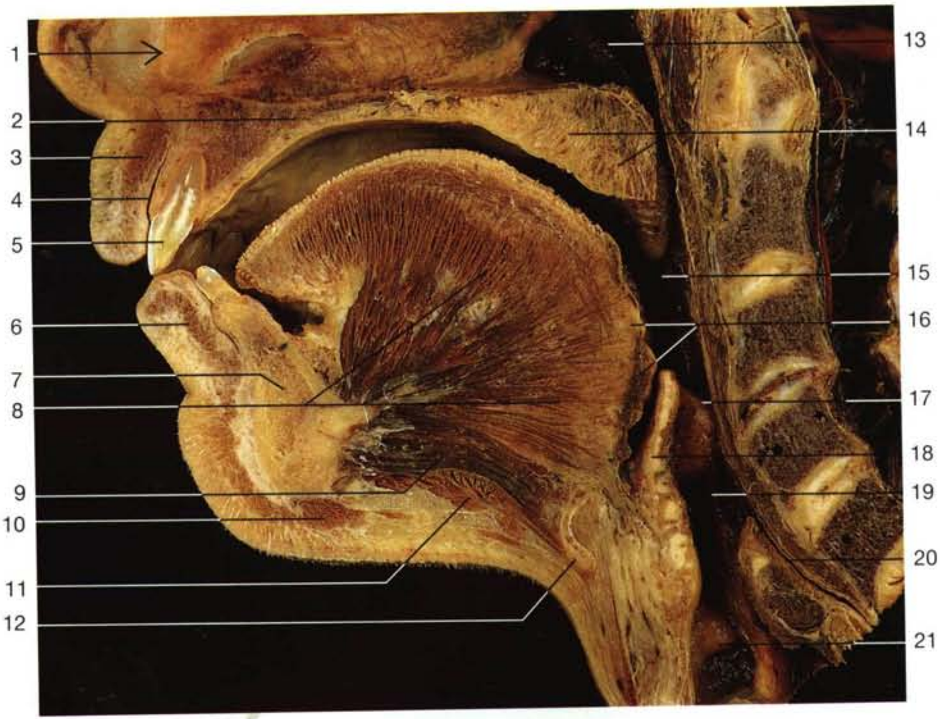

Median sagittal section through the oral cavity and pharynx.

1 Nasal cavity
2 **Hard palate**
3 Upper lip and orbicularis oris muscle
4 Vestibule of oral cavity
5 First incisor
6 Lower lip and orbicularis oris muscle
7 Mandible
8 Genioglossus muscle
9 Geniohyoid muscle
10 Anterior belly of diagastric muscle
11 Mylohyoid muscle
12 **Hyoid bone**
13 **Nasopharynx**
14 **Soft palate** and **uvula**
15 **Oropharynx**
16 Root of tongue and lingual tonsil
17 **Laryngopharynx**
18 Epiglottis
19 Aryepiglottic fold
20 Laryngopharynx continuous with esophagus
21 Larynx

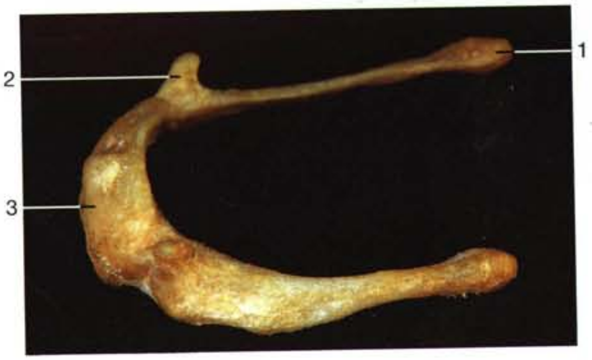

Hyoid bone (oblique lateral aspect).

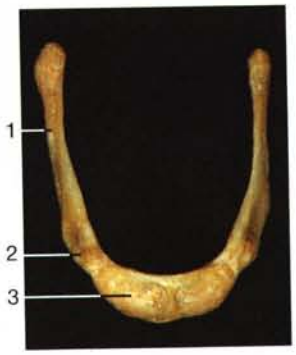

Hyoid bone (anterior aspect).

1 Greater cornu }
2 Lesser cornu } of hyoid bone
3 Body }

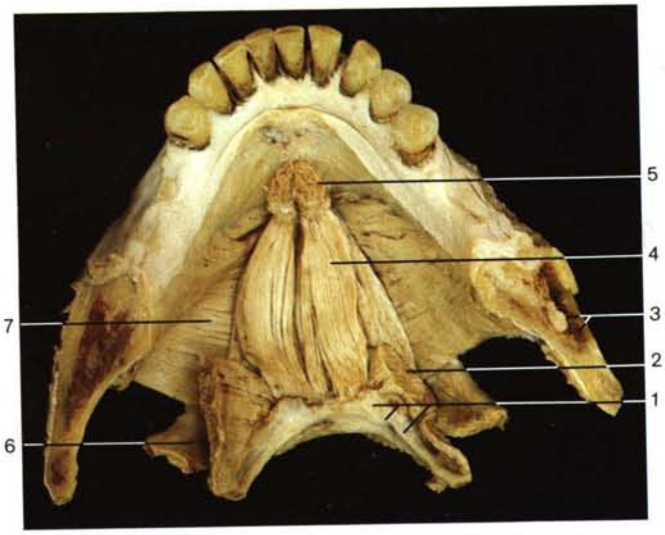

Muscles of the floor of the oral cavity (superior aspect).

1 Lesser cornu and body of hyoid bone
2 Hyoglossus muscle (divided)
3 Ramus of mandible and inferior alveolar nerve
4 **Geniohyoid muscle**
5 Genioglossus muscle (divided)
6 Stylohyoid muscle (divided)

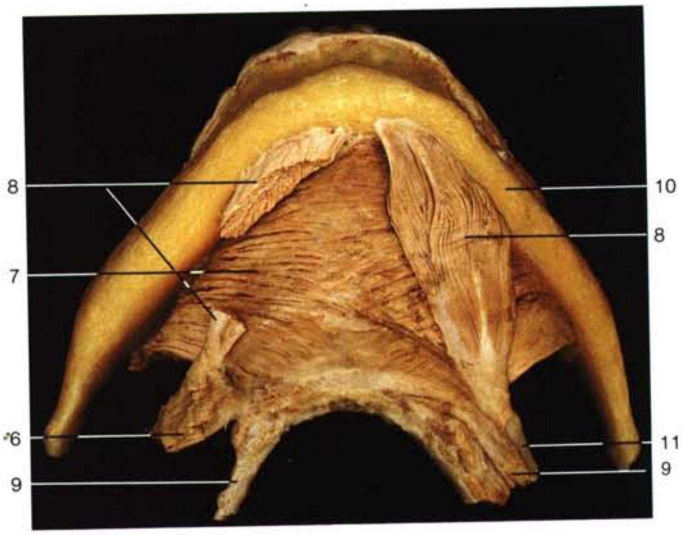

Oral diaphragm, muscles (inferior aspect). Cut on the base.

7 **Mylohyoid muscle**
8 Anterior belly of **digastric muscle**
9 Hyoid bone
10 Mandible
11 Intermediate tendon of digastric muscle

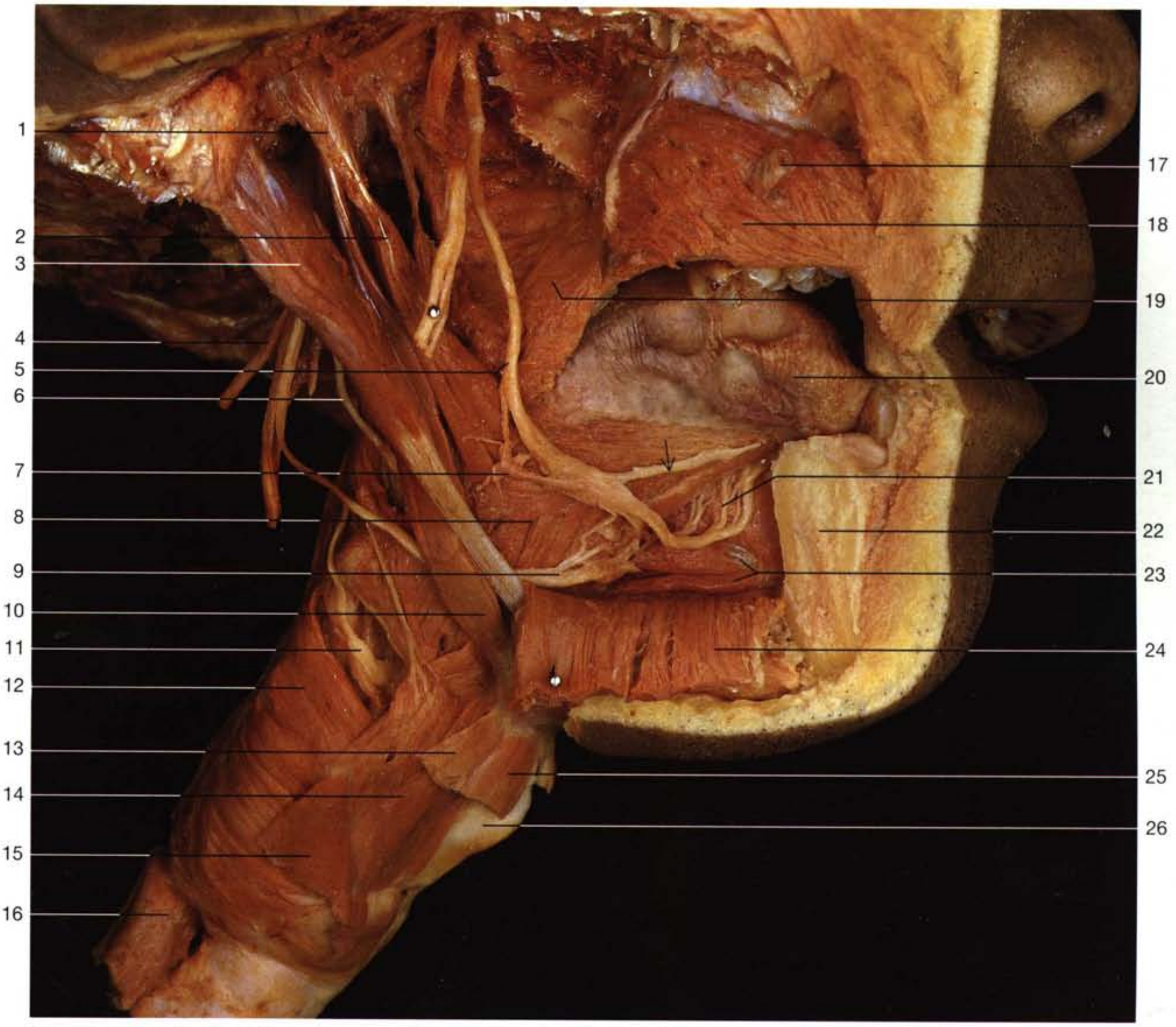

Parapharyngeal and sublingual regions. Innervation of the tongue. Lateral part of face and mandible removed, oral cavity opened. Arrow: submandibular duct.

1 Styloid process
2 Styloglossus muscle
3 Digastric muscle (posterior belly)

4 Vagus nerve (n. X)
5 **Lingual nerve** (n. V_3)
6 **Glossopharyngeal nerve** (n. IX)

7 Submandibular ganglion
8 Hyoglossus muscle
9 **Hypoglossal nerve** (n. XII)
10 Stylohyoid muscle
11 Internal branch of superior laryngeal nerve (branch of vagus nerve, not visible)
12 Middle constrictor muscle of pharynx
13 Omohyoid muscle (divided)
14 Thyrohyoid muscle
15 Sternothyroid muscle
16 Esophagus
17 Parotid duct (divided)
18 Buccinator
19 Superior constrictor muscle of pharynx
20 Tongue
21 Terminal branches of lingual nerve
22 Mandible (divided)
23 Genioglossus and geniohyoid muscles
24 Mylohyoid muscle (divided and reflected)
25 Sternohyoid muscle (divided)
26 Thyroid cartilage
27 Anterior belly of digastric muscle
28 Hyoid bone

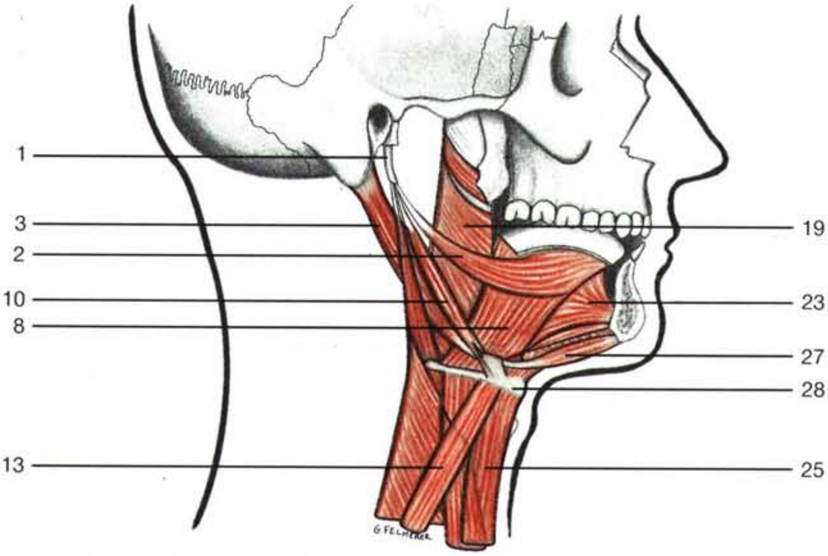

Supra- and infrahyoid muscles and pharynx (schematic drawing).

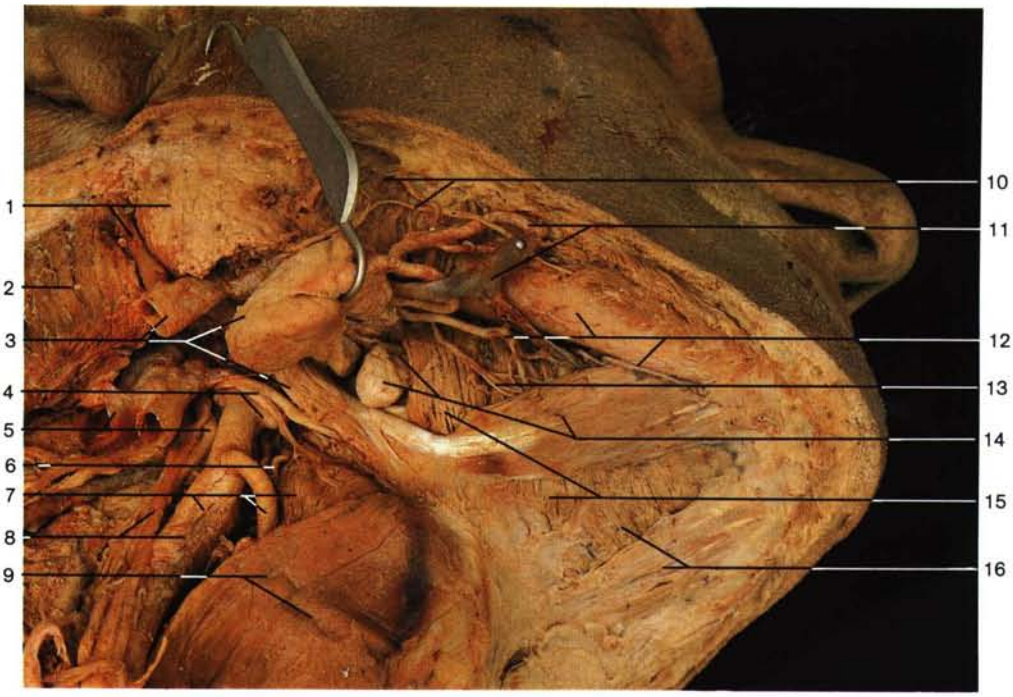

Submandibular triangle, superficial dissection. Right side (inferior aspect). Submandibular gland has been reflected.

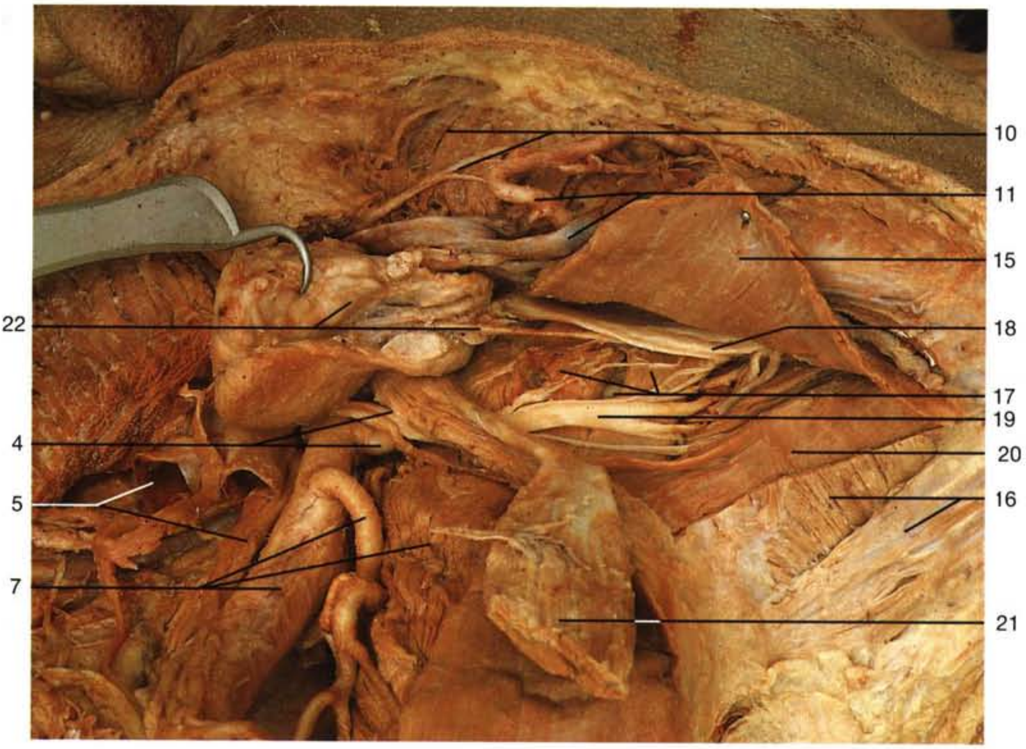

Submandibular triangle, deep dissection. Right side. Mylohyoid muscle has been severed and reflected to display the lingual and hypoglossal nerves.

1 Parotid gland and retromandibular vein
2 Sternocleidomastoid
3 Retromandibular vein, submandibular gland and stylohyoid muscle
4 Hypoglossal nerve and lingual artery
5 Vagus nerve and internal jugular vein
6 Superior laryngeal artery
7 **External carotid artery,** thyrohyoid muscle, and superior thyroid artery
8 Common carotid artery and superior root of ansa cervicalis
9 Omohyoid and sternohyoid
10 Masseter and marginal mandibular branch of facial nerve
11 **Facial artery and vein**
12 Mandible and submental artery and vein
13 Mylohyoid nerve
14 Submandibular duct, sublingual gland, and anterior belly of digastric muscle
15 Mylohyoid (right side)
16 Left mylohyoid and anterior belly of left digastric muscle
17 Hyoglossus muscle and lingual artery
18 **Lingual nerve**
19 **Hypoglossal nerve**
20 Geniohyoid muscle
21 Anterior belly of right digastric muscle
22 Submandibular gland and duct

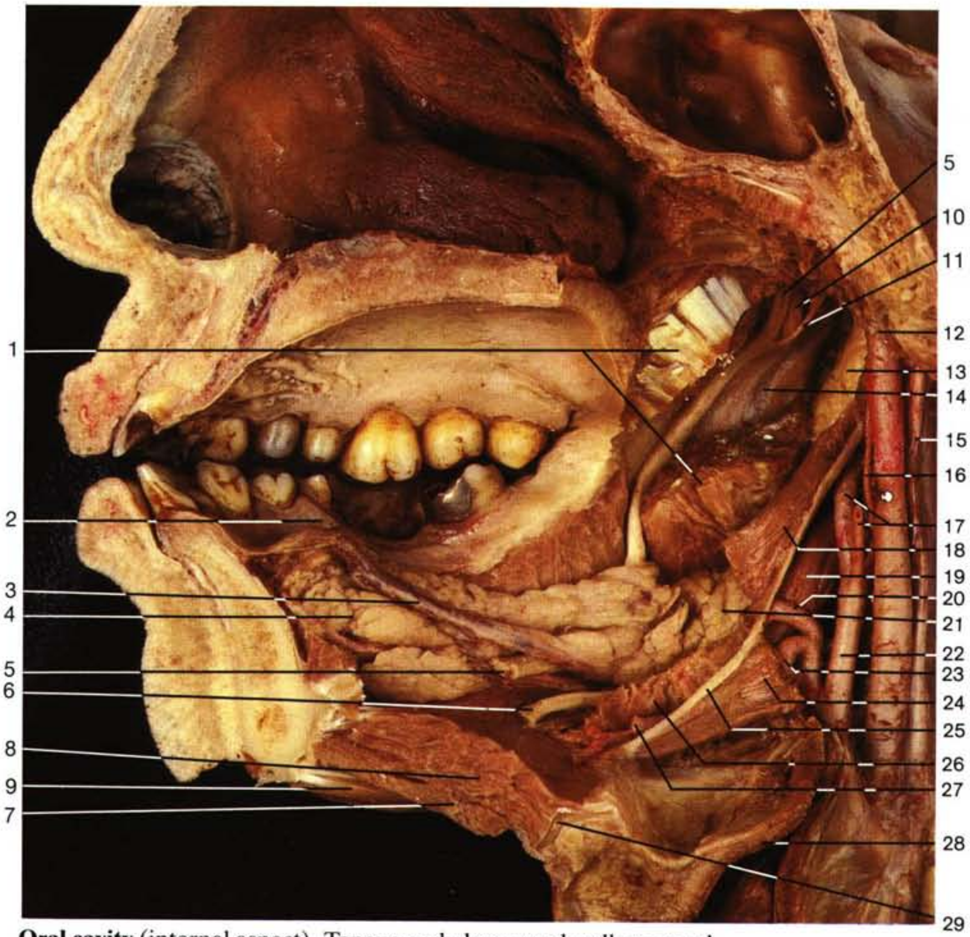

Oral cavity (internal aspect). Tongue and pharyngeal wall removed.

1	Medial pterygoid muscle
2	Sublingual papilla
3	**Submandibular duct**
4	**Sublingual gland**
5	**Lingual nerve**
6	Hypoglossal nerve
7	Mylohyoid muscle
8	Geniohyoid muscle
9	Anterior belly of digastric muscle
10	Inferior alveolar nerve
11	**Chorda tympani**
12	Internal carotid artery
13	Parotid gland
14	Sphenomandibular ligament
15	Vagus nerve
16	Glossopharyngeal nerve
17	Superficial temporal artery and ascending pharyngeal artery
18	Styloglossus muscle
19	Posterior belly of digastric muscle
20	**Facial artery**
21	Submandibular gland
22	**External carotid artery**
23	Lingual artery
24	Middle pharyngeal constrictor muscle
25	**Stylohyoid ligament**
26	Hyoglossus muscle
27	Deep lingual artery
28	Epiglottis
29	Hyoid bone
30	Buccinator muscle
31	Tongue
32	Mandible (divided)
33	**Parotid duct**
34	Masseter muscle
35	Right and left sublingual papilla

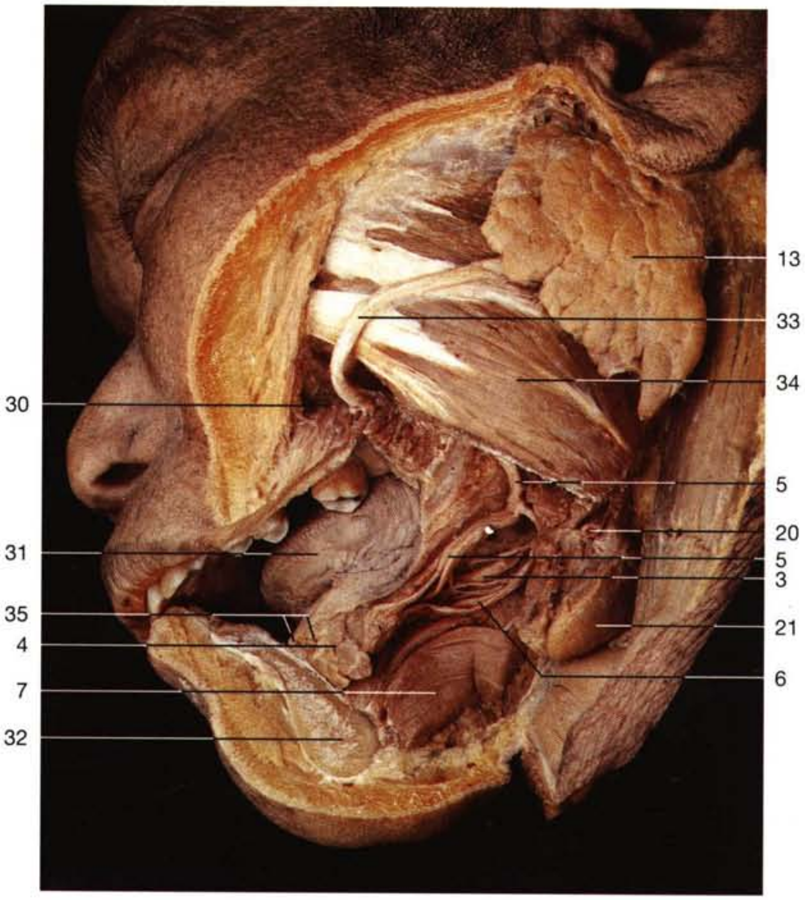

Dissection of major salivary glands. Left mandible and buccinator muscle partly removed to view the oral cavity (inferior lateral aspect).

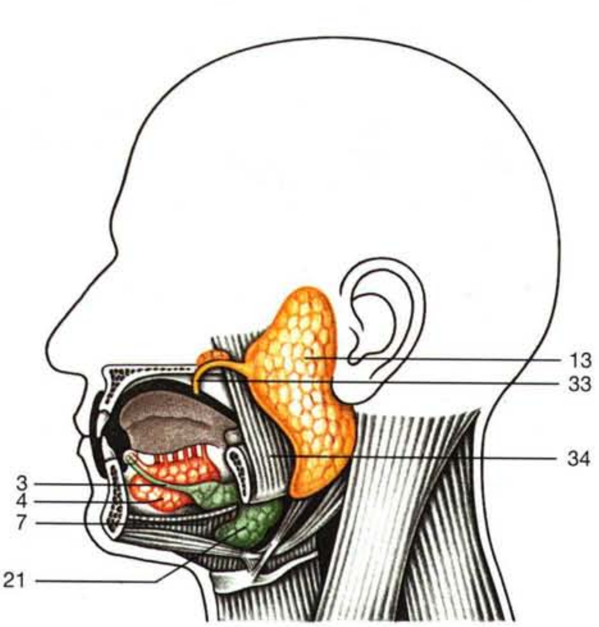

Location of the major salivary glands in relation to the oral cavity.

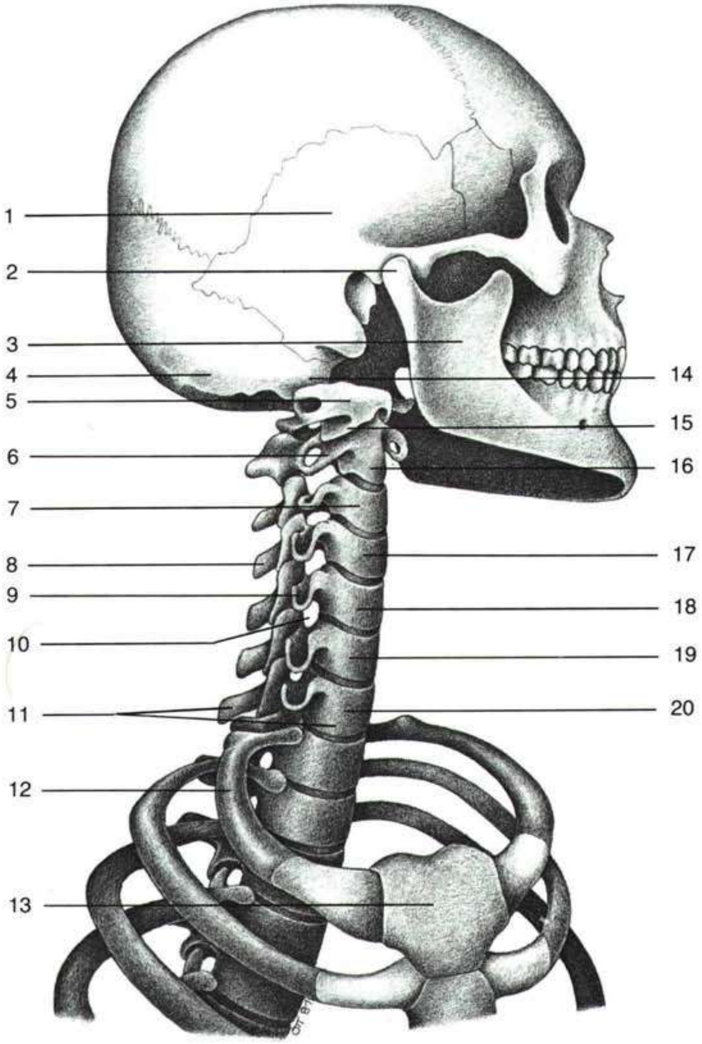

Cervical spine (oblique lateral aspect).

1 Temporal bone
2 Head of mandible
3 Mandible
4 Occipital bone
5 Atlas
6 Axis
7 Third cervical vertebra
8 Spinous process of fourth cervical vertrebra
9 Transverse process of fifth cervical vertebra with groove for spinal nerve
10 Intervertebral foramen
11 Vertebra prominens (C_7)
12 First rib
13 Sternum (manubrium sterni)
14 Atlantooccipital joint
15 Lateral atlantoaxial joint
16 Body of axis
17 Body of fourth cervical vertebra
18 Body of fifth cervical vertebra
19 Body of sixth cervical vertebra
20 Body of seventh cervical vertebra

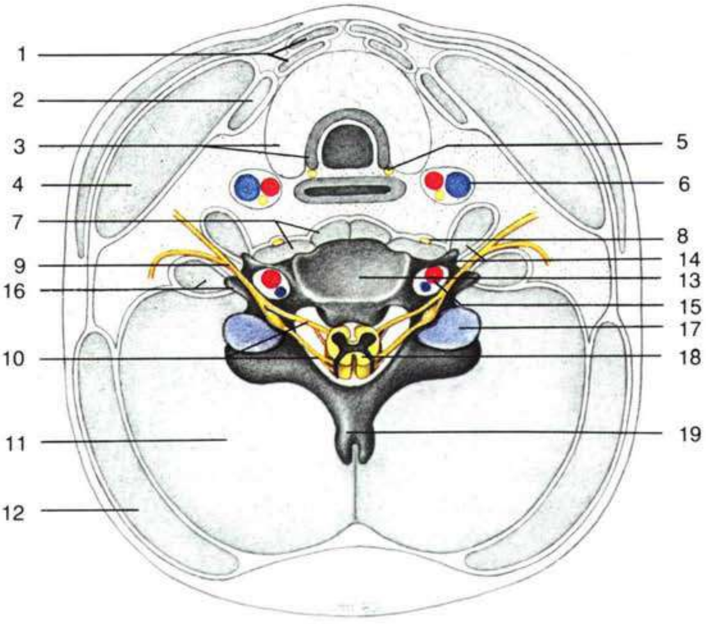

Cervical vertebra and the **organization of the neck.** (Schematic drawing.)

1 Sternohyoid and sternothyroid muscles
2 Omohyoid muscle
3 Thyroid gland and trachea
4 Sternocleidomastoid muscle
5 Recurrent laryngeal nerve
6 Internal jugular vein, common carotid artery and vagus
7 Longus colli and longus capitis muscles
8 Sympathetic trunk
9 Spinal nerve
10 Ventral and dorsal root of spinal nerve
11 True muscles of the neck
12 Trapezius muscle
13 Body of cervical vertebra
14 Anterior tubercle of transverse process and origin of scalenus anterior muscle
15 Vertebral artery and foramen transversarium
16 Posterior tubercle of transverse process and origin of scalenus medius and posterior muscles
17 Superior facet of articular process
18 Spinal cord
19 Spinous process

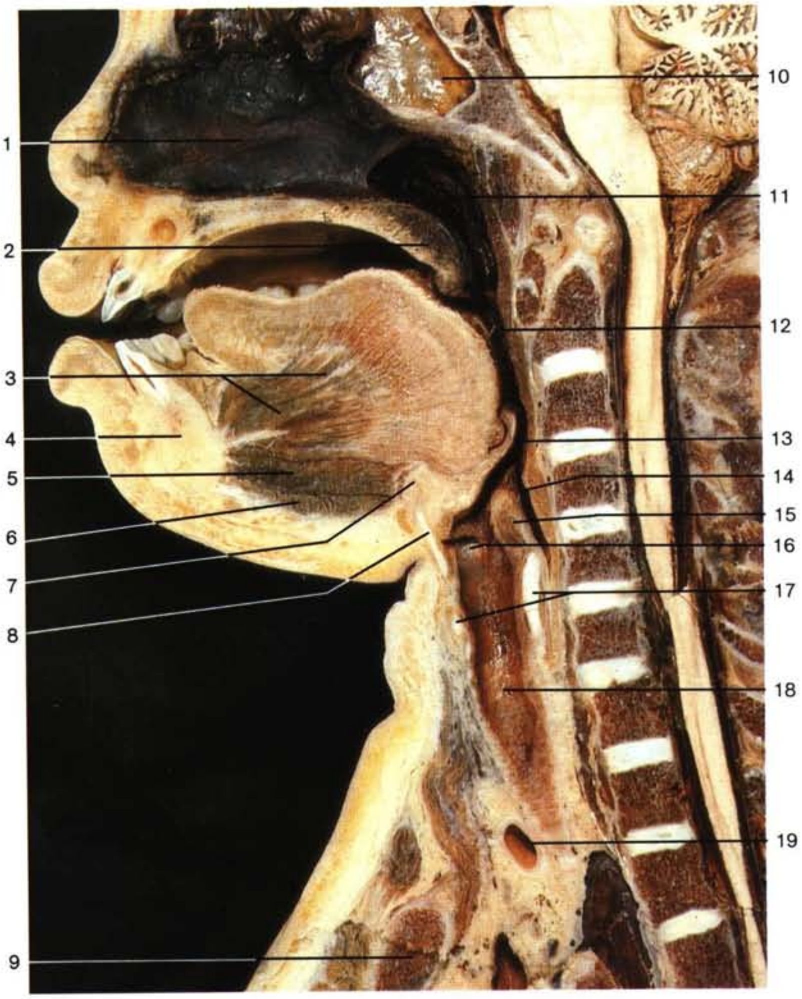

Median section through adult head and neck. Note the low position of the adult larynx when compared with that of the neonate (cf. with figure below).

1 Nasal septum
2 Uvula
3 Genioglossus muscle
4 Mandible
5 Geniohyoid muscle
6 Mylohyoid muscle
7 **Hyoid bone**
8 Thyroid cartilage
9 Manubrium sterni
10 Sphenoidal sinus
11 **Nasopharynx**
12 **Oropharynx**
13 **Epiglottis**
14 **Laryngopharynx**
15 Arytenoid muscle
16 **Vocal fold**
17 Cricoid cartilage
18 **Trachea**
19 Left brachiocephalic vein
20 Thymus
21 Esophagus
22 Occipital lobe
23 Cerebellum and 4th ventricle
24 Medulla oblongata
25 Dens of axis
26 Intervertebral discs of cervical vertebral column

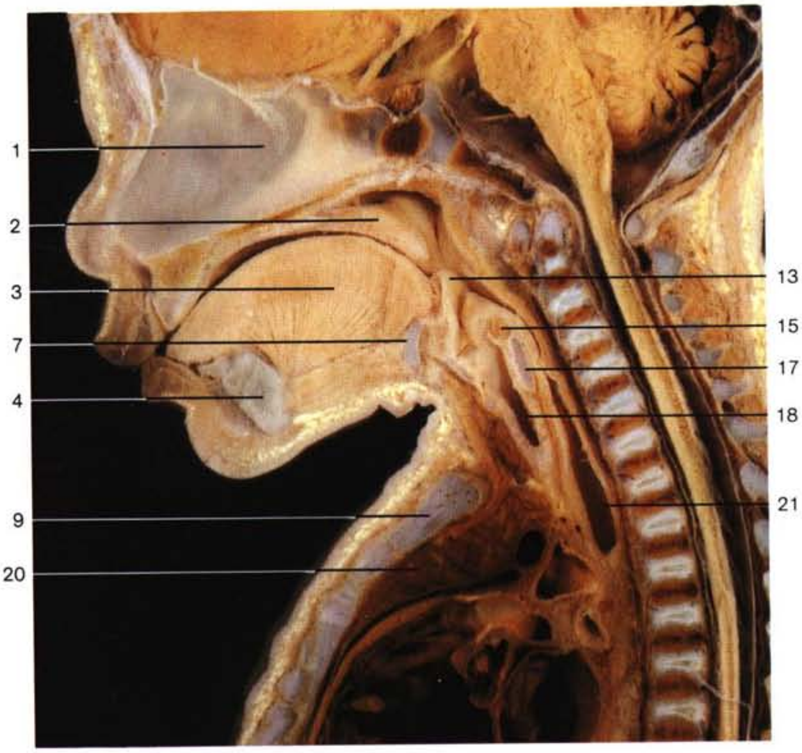

Median section through neonate head and neck. Note the high position of the larynx permitting the epiglottis nearly to reach the uvula (cf. with the figure above).

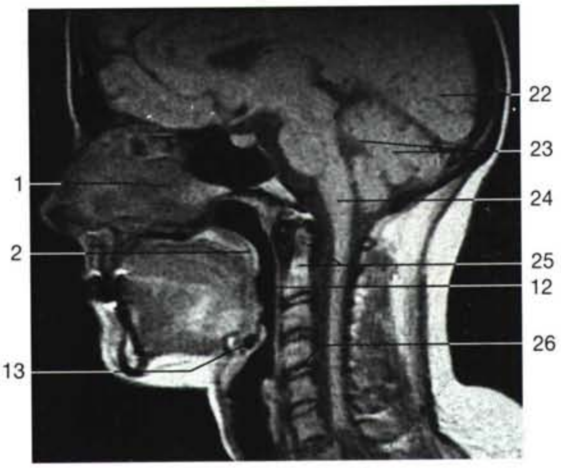

Sagittal section through the head. (MR-Scan.)

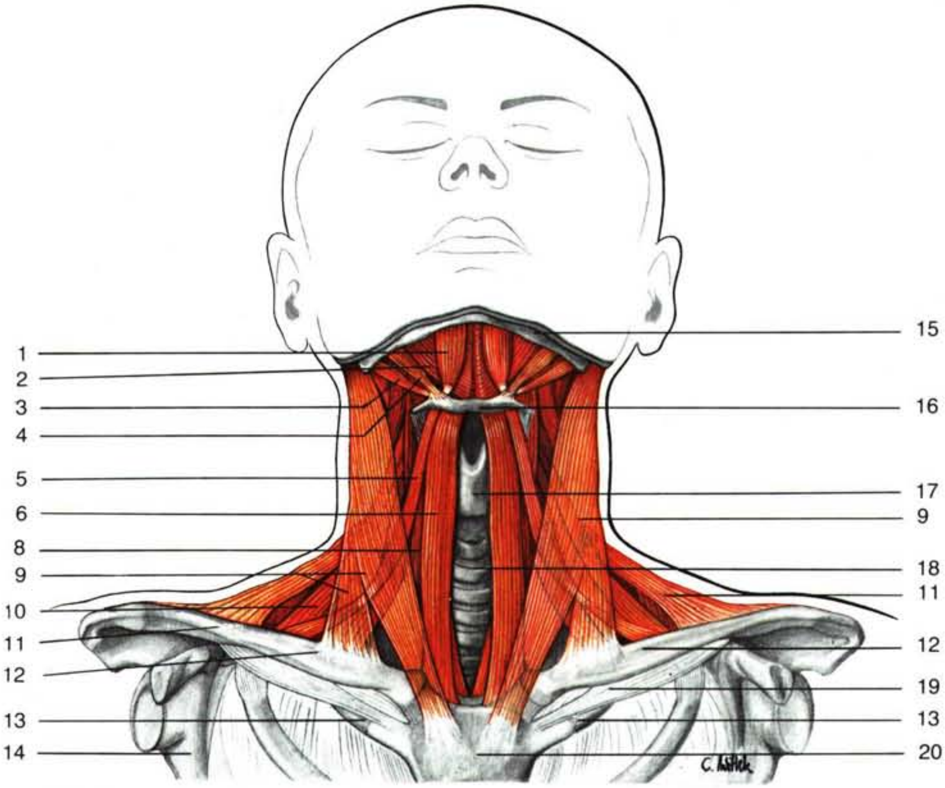

Muscles of the neck (anterior aspect).

Suprahyoid muscles

1 Anterior belly of digastric muscle
2 Mylohyoid muscle
3 Posterior belly of digastric muscle
4 Stylohyoid muscle

Infrahyoid muscles

5 Omohyoid muscle
6 Sternohyoid muscle
7 Thyrohyoid muscle
8 Sternothyroid muscle

Other structures

9 Sternocleidomastoid muscle
10 Scalenus muscles
11 Trapezius muscle
12 Clavicle
13 First rib
14 Scapula
15 Mandible
16 Hyoid bone
17 Larynx (thyroid cartilage)
18 Trachea
19 Subclavius muscle
20 Manubrium sterni
21 Mucous membrane of larynx (conus elasticus)
22 Cricoid cartilage
23 Inferior horn of thyroid cartilage
24 Esophagus
25 Body of cervical vertebra
26 Posterior root ganglion
27 Spinal cord
28 Spinous process
29 Internal jugular vein, common carotid artery, and vagus nerve
30 True muscles of the neck (semispinalis cervicis and capitis muscles)

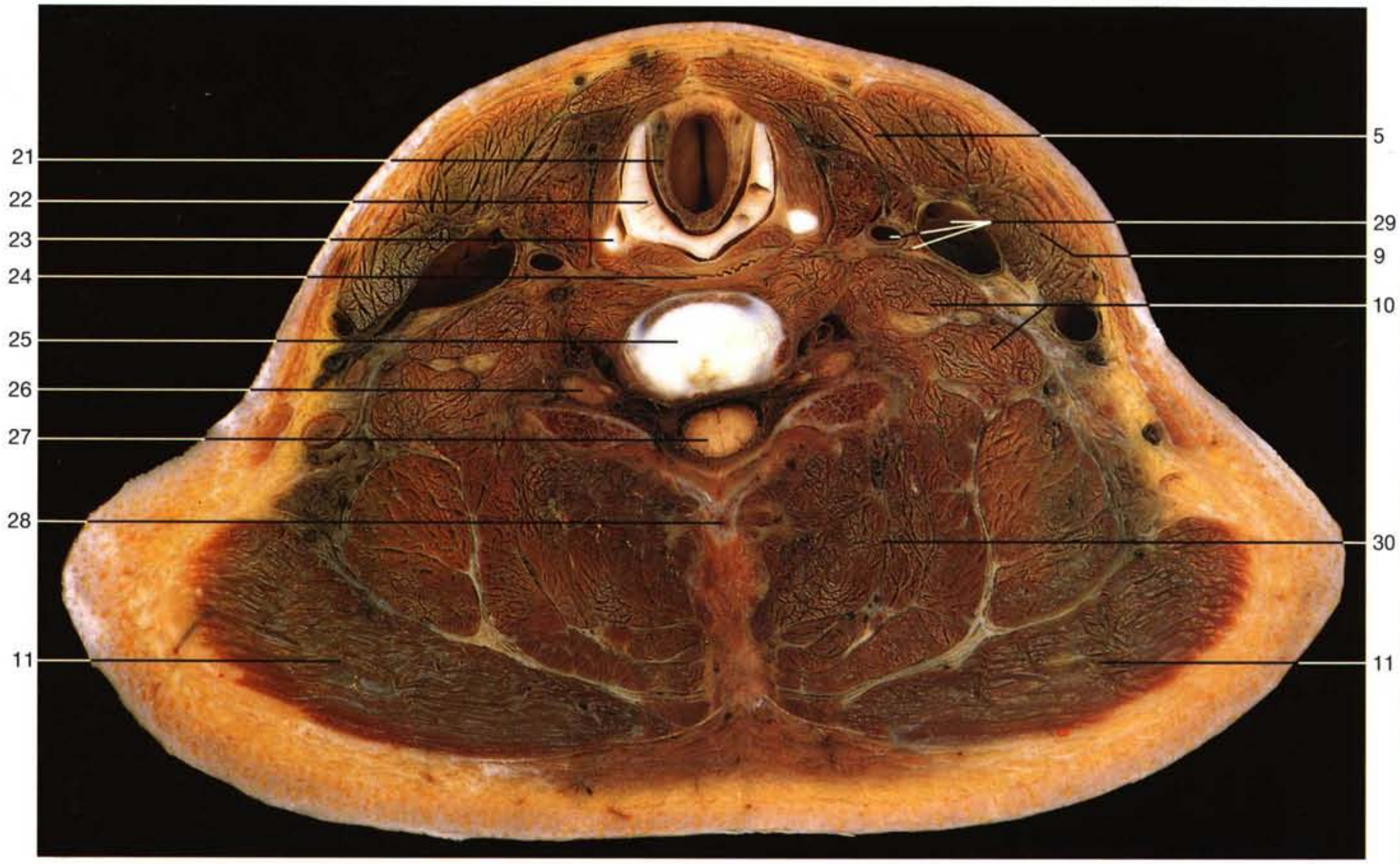

Cross-section of the neck at the level of the intervertebral disc between the 5th and 6th cervical vertebra (inferior aspect).

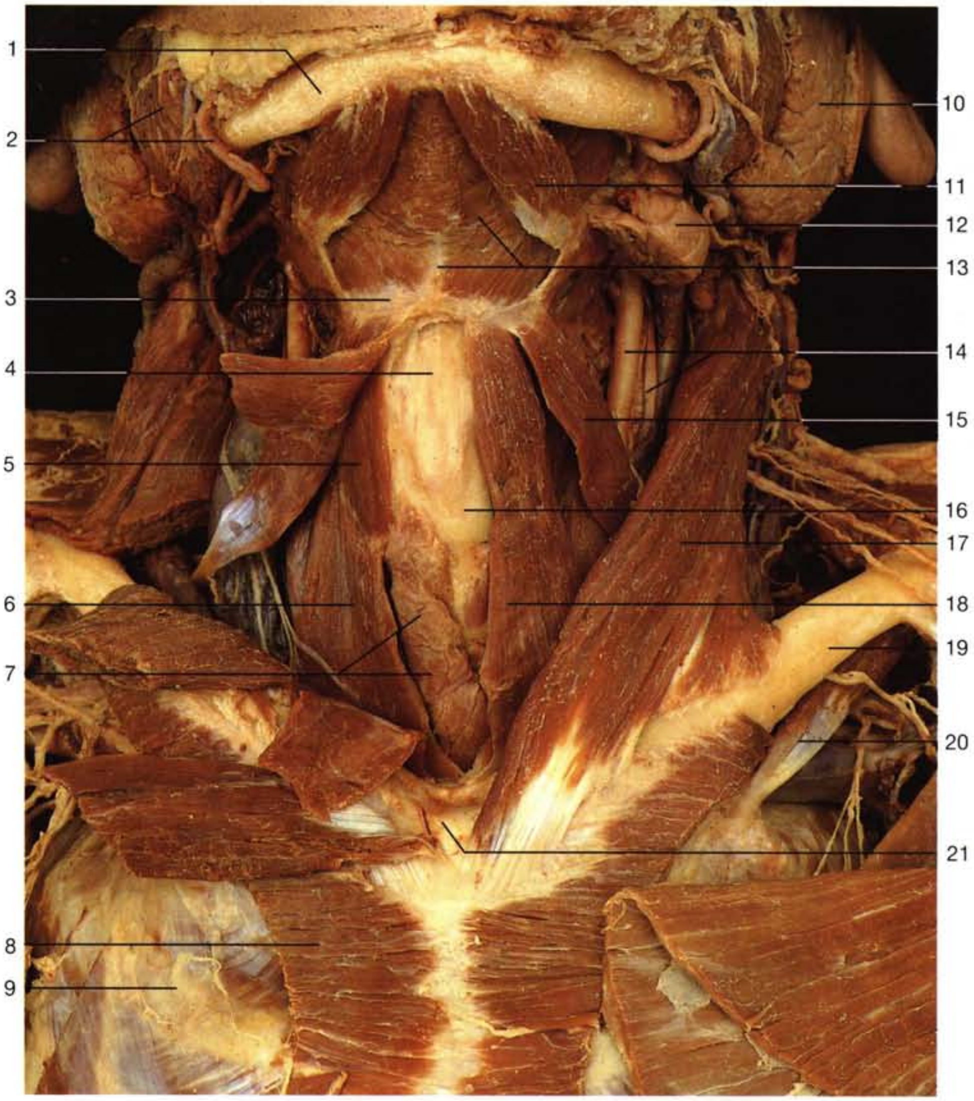

1 Mandible
2 Masseter muscle and facial artery
3 **Hyoid bone**
4 Median thyrohyoid ligament
5 **Thyrohyoid muscle**
6 **Sternothyroid muscle**
7 Thyroid gland (pyramidal lobe)
8 Pectoralis major muscle
9 Second rib
10 Parotid gland
11 Anterior belly of **digastric muscle**
12 Submandibular gland (divided)
13 **Mylohyoid muscle** and mylohyoid raphe
14 External carotid artery and vagus nerve
15 **Omohyoid muscle**
16 Thyroid cartilage
17 Sternocleidomastoid muscle
18 **Sternohyoid muscle**
19 Clavicle
20 Subclavius
21 Jugular fossa or suprasternal notch

Muscles of the neck (anterior aspect). Sternocleidomastoid and sternohyoid muscles on the right have been divided and reflected.

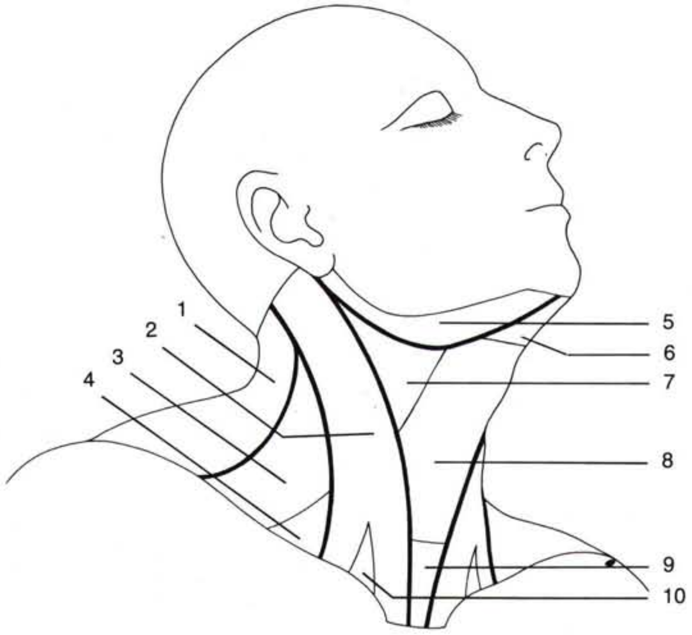

Regions and triangles of the neck.

1	Trapezius	
2	Sternocleidomastoid	
3	Occipital triangle	**Posterior triangle**
4	Supraclavicular triangle	
5	Submandibular triangle	**Anterior triangle**
6	Submental triangle	
7	Carotid triangle	
8	Muscular triangle	
9	Jugular fossa	
10	Lesser supraclavicular fossa	

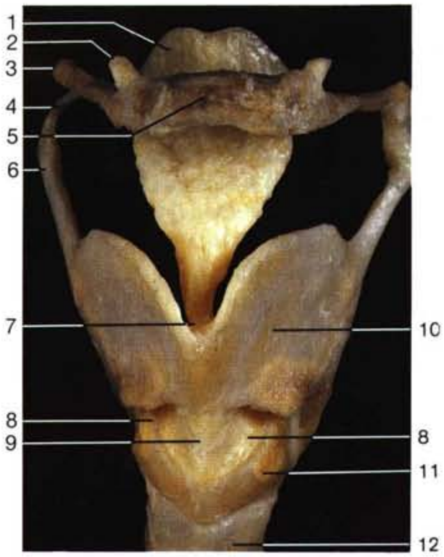

Cartilages of the larynx and the hyoid bone (anterior aspect).

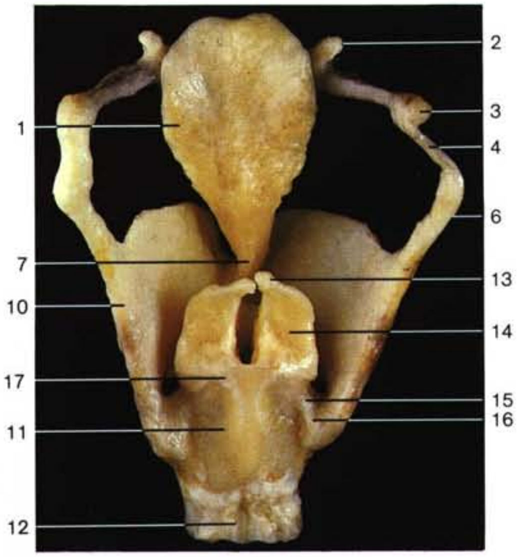

Cartilages of the larynx and the hyoid bone (posterior aspect).

1 **Epiglottis**
2 Lesser cornu of hyoid bone
3 Greater cornu of hyoid bone
4 Lateral thyrohyoid ligament
5 Body of **hyoid bone**
6 Superior cornu of thyroid cartilage
7 Thyroepiglottic ligament
8 Conus elasticus
9 Cricothyroid ligament
10 **Thyroid cartilage**
11 **Cricoid cartilage**
12 Trachea
13 Corniculate cartilage
14 **Arytenoid cartilage**
15 Posterior cricoarytenoid ligament
16 **Cricothyroid joint**
17 **Cricoarytenoid joint**

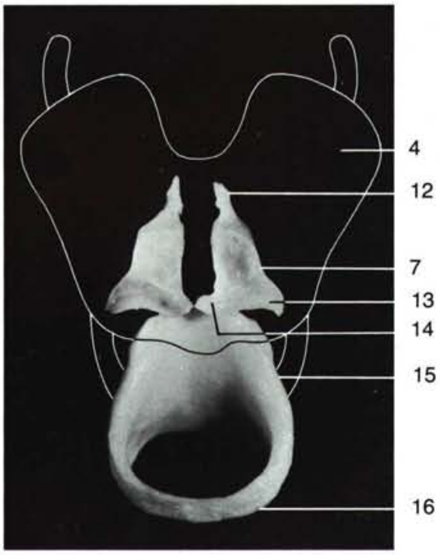

Cartilages of the larynx (anterior aspect). Thyroid cartilage is indicated by the outline.

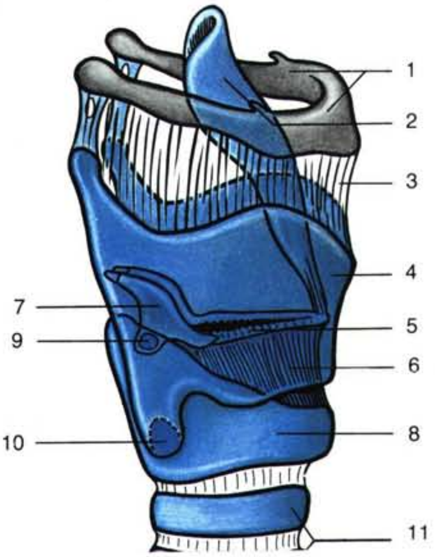

Cartilages and ligaments of the larynx (lateral aspect). (Schematic drawing.)

1 Hyoid bone
2 Epiglottis
3 Thyrohyoid membrane
4 **Thyroid cartilage**
5 **Vocal ligament**
6 Conus elasticus
7 Arytenoid cartilage
8 **Cricoid cartilage**
9 **Cricoarytenoid joint**
10 **Cricothyroid joint**
11 Tracheal cartilages
12 Corniculate cartilage
13 Muscular process of arytenoid cartilage
14 Vocal process of arytenoid cartilage
15 Lamina of cricoid cartilage
16 Arch of cricoid cartilage

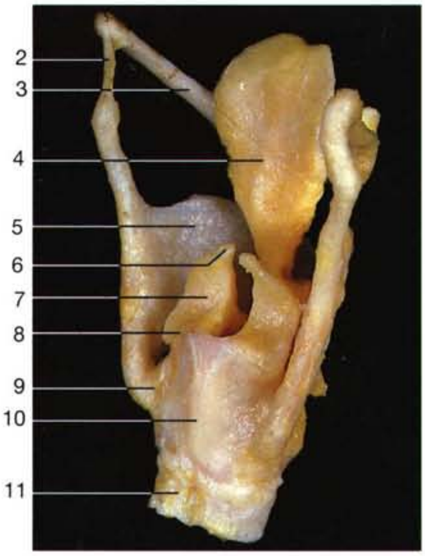

Cartilages of the larynx (oblique-posterior aspect).

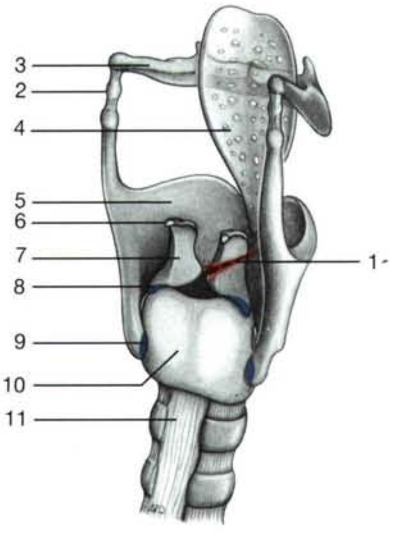

Cartilages of the larynx (oblique-posterior aspect).

1 **Vocal ligament** (red)
2 Lateral thyrohyoid ligament
3 Greater cornu of hyoid bone
4 **Epiglottis**
5 **Thyroid cartilage**
6 Corniculate cartilage
7 **Arytenoid cartilage**
8 **Cricoarytenoid joint**
9 **Cricothyroid joint**
10 **Cricoid cartilage**
11 Trachea

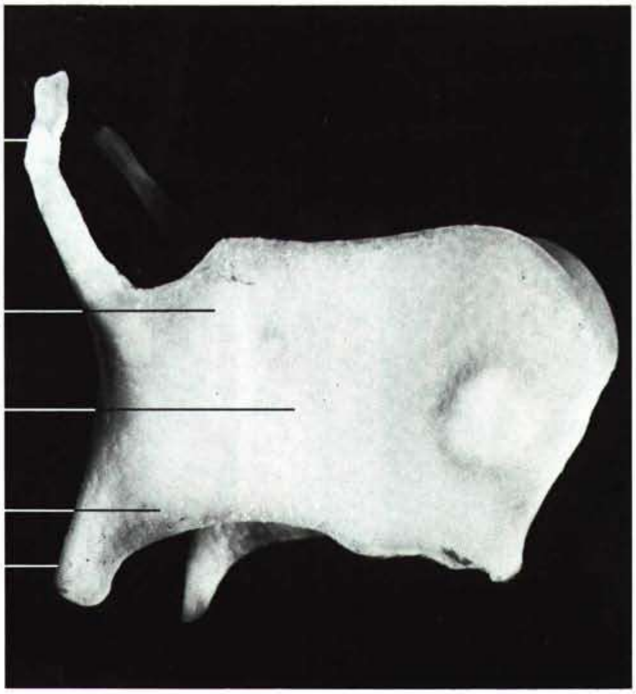

Thyroid cartilage (lateral aspect).

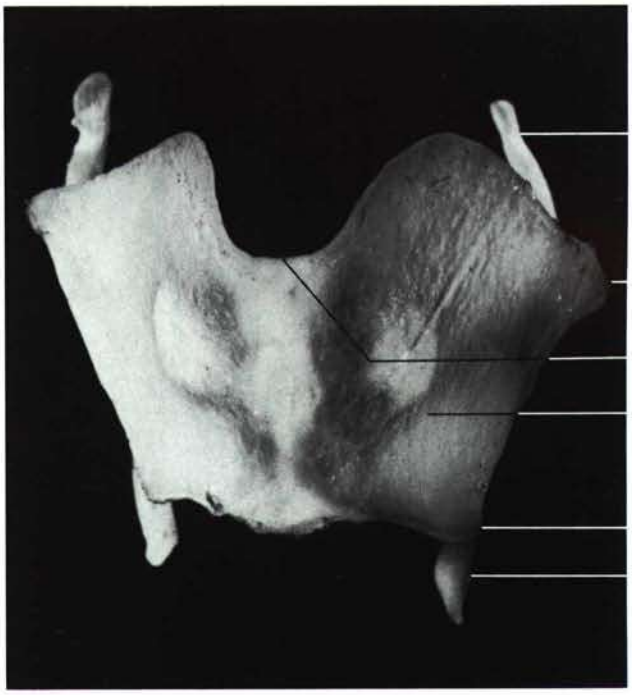

Thyroid cartilage (anterior aspect).

1 Superior cornu
2 Superior thyroid tubercle
3 Lamina of thyroid cartilage
4 Inferior thyroid tubercle
5 Inferior cornu
6 Superior thyroid notch

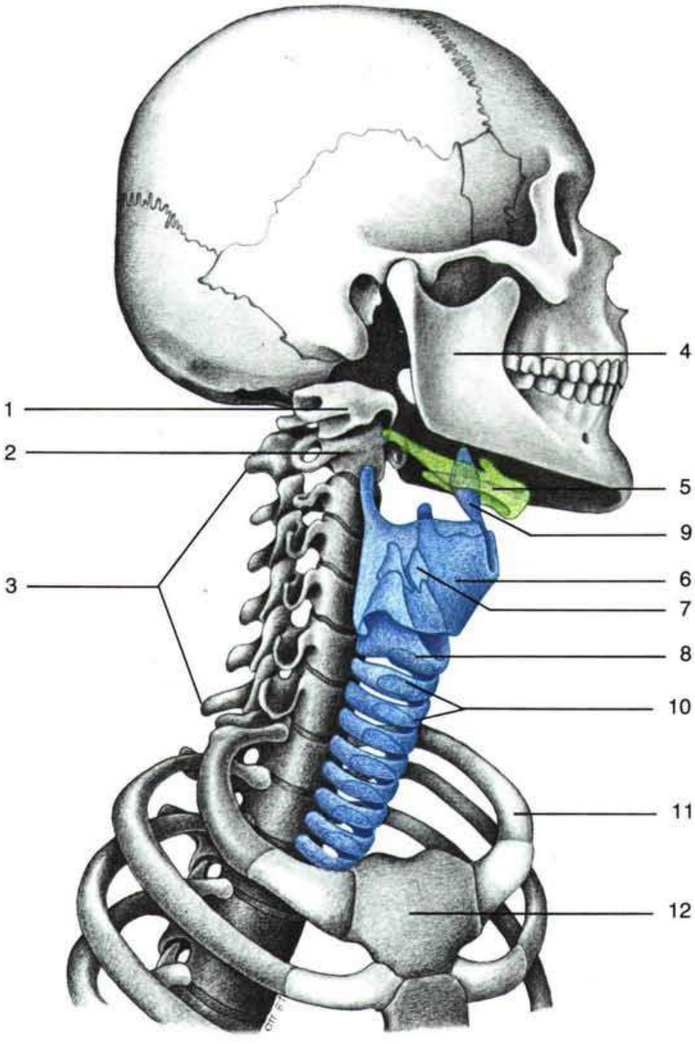

1 Atlas
2 Axis
3 Cervical vertebrae (C_2–C_7)
4 Mandible
5 **Hyoid bone**
6 **Thyroid cartilage**
7 Arytenoid cartilage
8 **Cricoid cartilage**
9 Epiglottis
10 **Tracheal cartilages**
11 First rib
12 Manubrium sterni

Position of the larynx in the neck (oblique lateral aspect). (Schematic drawing.)

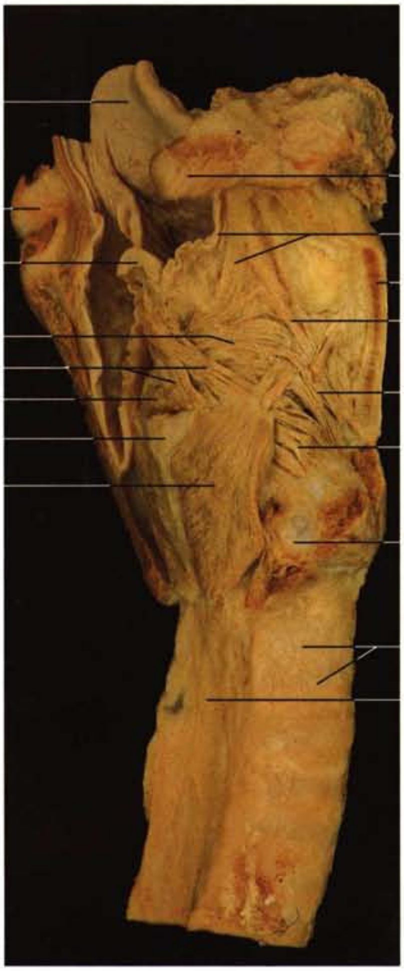

Laryngeal muscles I (lateral aspect). Thyroid cartilage and thyroarytenoid muscle have been partly removed.

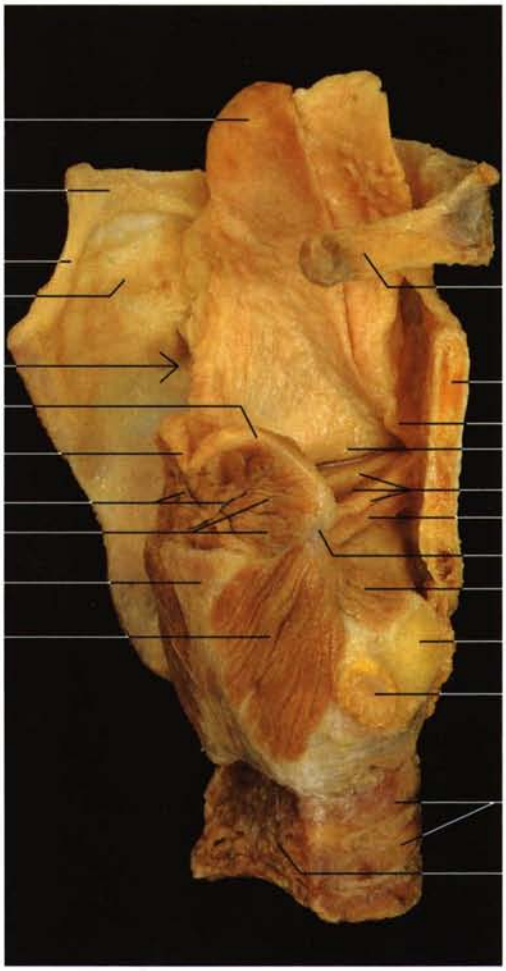

Laryngeal muscles II (lateral aspect). Half of the right side of the thyroid cartilage has been removed.

1 Epiglottis
2 Greater horn of hyoid bone
3 Corniculate cartilage
4 Transverse arytenoid muscle
5 Oblique arytenoid muscle
6 Lamina of cricoid cartilage
7 **Posterior cricoarytenoid muscle**
8 Aryepiglottic fold and muscle
9 Lamina of thyroid cartilage
10 Thyroepiglottic muscle
11 Thyroarytenoid muscle
12 **Lateral cricoarytenoid muscle**
13 Articular facet of thyroid cartilage
14 Cartilages of trachea
15 Membranous part of trachea
16 Lateral thyrohyoid ligament
17 Thyrohyoid membrane
18 Position of piriform recess
19 Arytenoid cartilage
20 Stem of epiglottis
21 Vestibular fold
22 Vocal folds and fissure of glottis
23 **Vocalis muscle**
24 Cricoarytenoid articulation
25 Arch of cricoid cartilage
26 Cricothyroid muscle
27 Cricoid cartilage
28 Piriform recess
29 Superior horn of thyroid cartilage
30 Vocal ligament

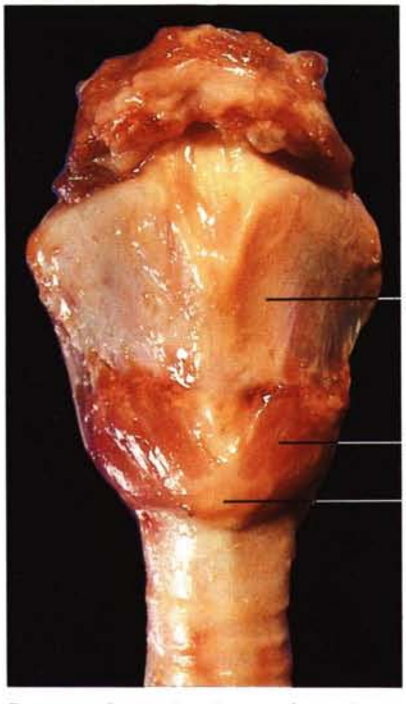

Laryngeal muscles, larynx (anterior aspect).

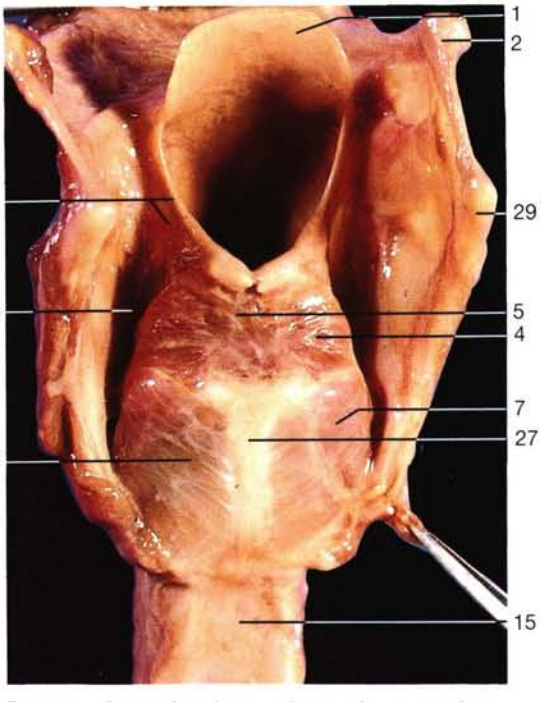

Laryngeal muscles, larynx (posterior aspect).

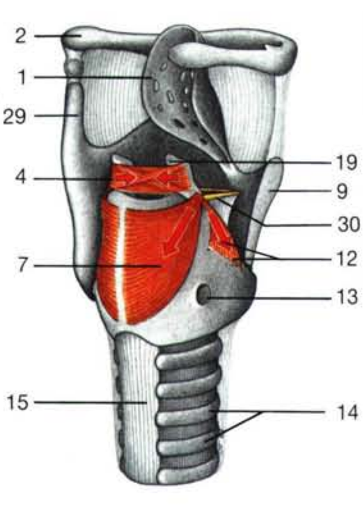

Action of internal muscles of the larynx. (Schematic drawing.)

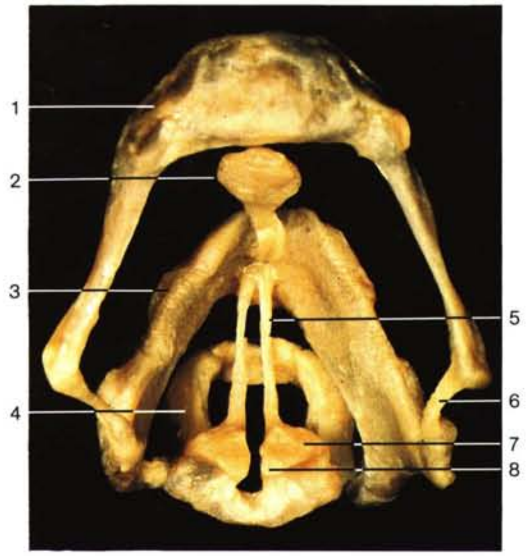

Laryngeal cartilages (superior aspect).

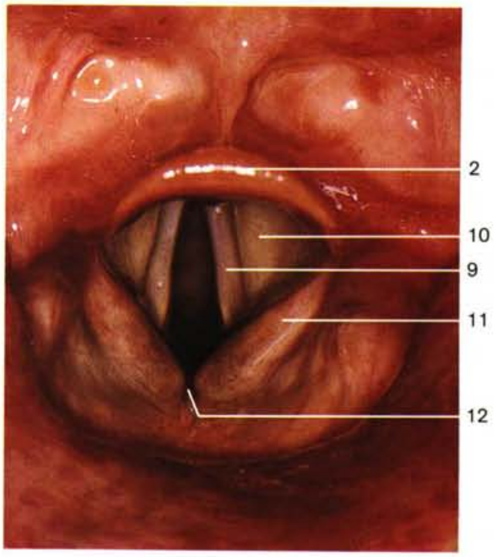

Glottis in vivo (superior aspect).

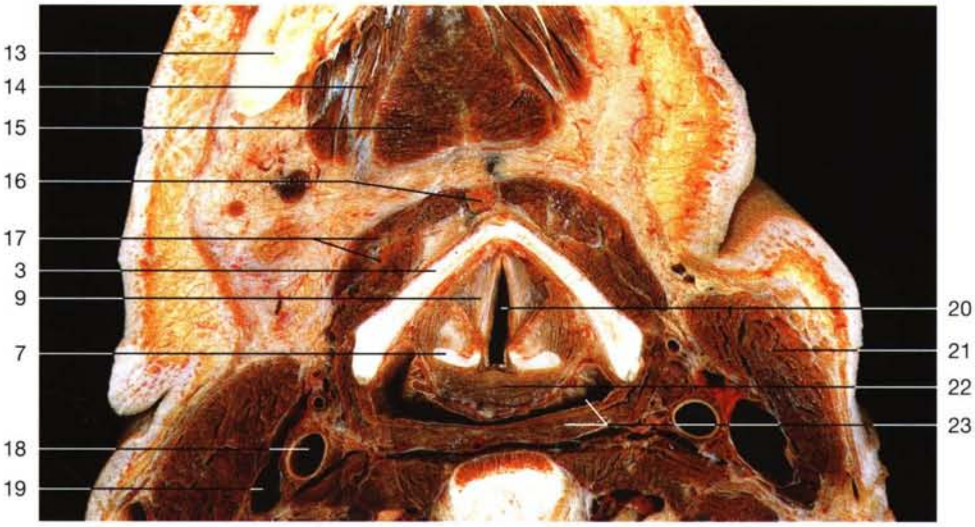

Horizontal section through the larynx at the level of the vocal folds (superior aspect).

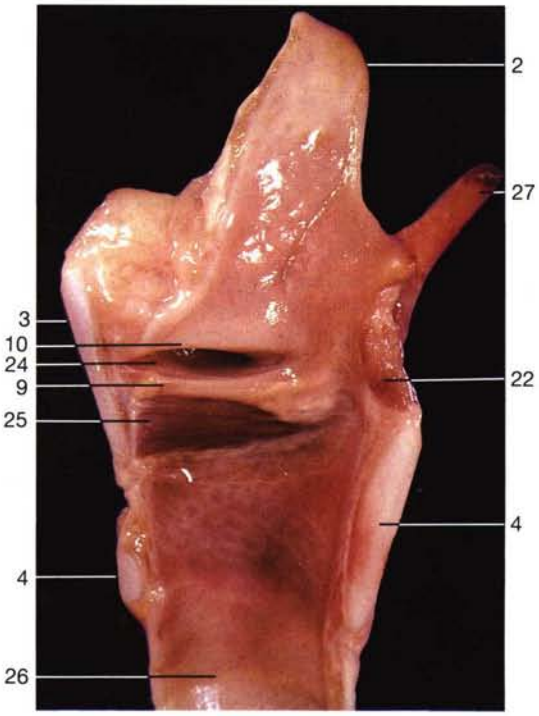

Sagittal section through the larynx.

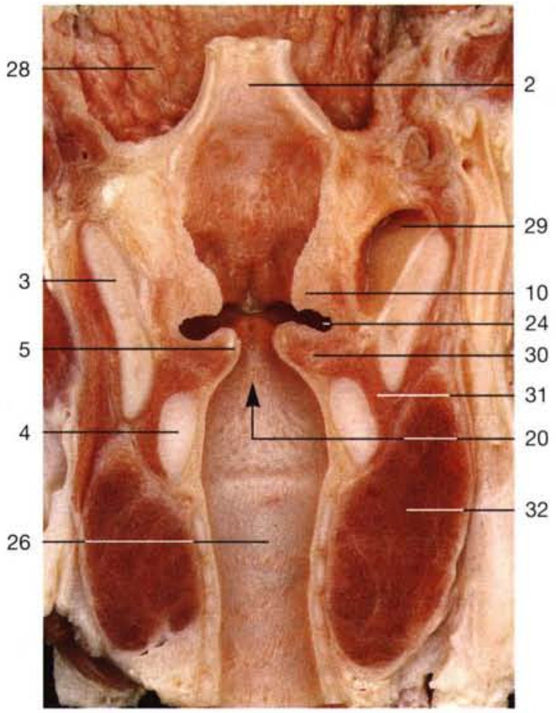

Coronal section through larynx and trachea.

1 **Hyoid bone**
2 Epiglottis
3 **Thyroid cartilage**
4 **Cricoid cartilage**
5 **Vocal ligament**
6 Thyrohyoid ligament
7 **Arytenoid cartilage**
8 Corniculate cartilage
9 **Vocal fold**
10 **Vestibular fold**
11 Aryepiglottic fold
12 Interarytenoid notch
13 Mandible
14 Anterior belly of digastric muscle
15 Mylohyoid muscle
16 Pyramidal process of thyroid gland
17 Sternohyoid and sternothyroid muscles
18 Common carotid artery
19 Internal jugular vein
20 **Rima glottidis**
21 Sternocleidomastoid muscle
22 Transverse arytenoid muscle
23 Pharynx and inferior constrictor muscle
24 Ventricle of larynx
25 **Vocalis muscle**
26 Trachea
27 Superior cornu of thyroid cartilage
28 Root of tongue (lingual tonsil)
29 Piriform recess
30 Vocalis muscle
31 Lateral cricoarytenoid muscle
32 **Thyroid gland**

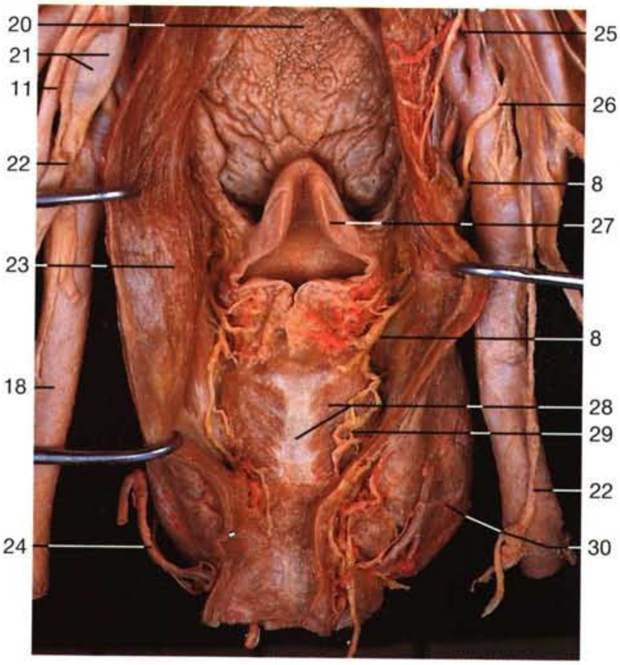

Larynx and its innervation (posterior aspect). Dissection of superior and inferior laryngeal nerves. Pharynx has been opened.

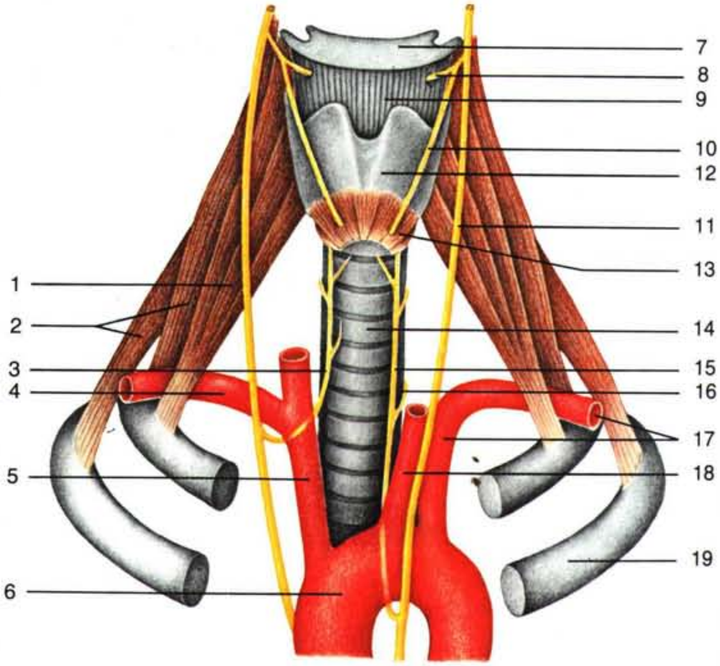

Innervation of the larynx. (Schematic diagram.)

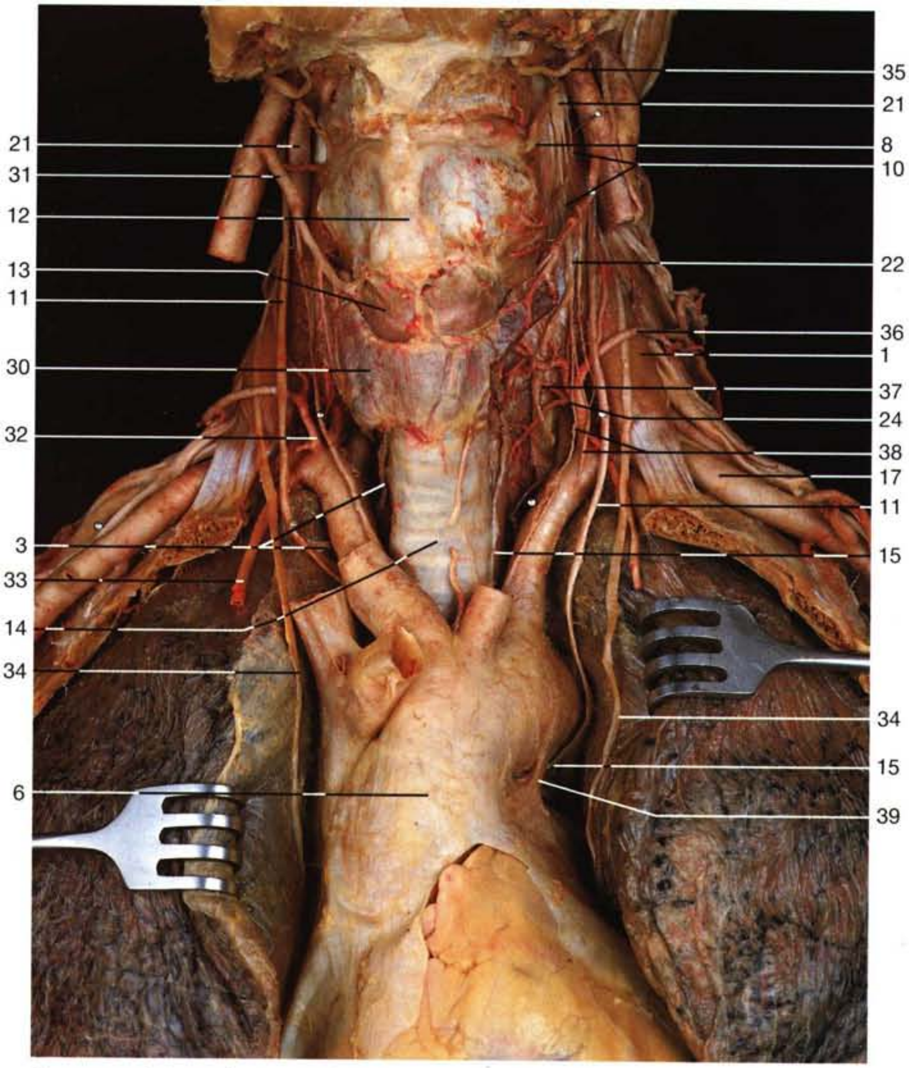

Larynx and thoracic organs (anterior aspect). Dissection of vagus and recurrent laryngeal nerves.

1 Scalenus anterior muscle
2 Scalenus medius and posterior muscles
3 Right **recurrent laryngeal nerve**
4 Right subclavian artery
5 Brachiocephalic trunk
6 Aortic arch
7 Hyoid bone
8 **Internal branch of superior laryngeal nerve**
9 Thyrohyoid membrane
10 **External branch of superior laryngeal nerve**
11 **Vagus nerve**
12 Thyroid cartilage
13 Cricothyroid muscle
14 Trachea
15 **Left recurrent larnygeal nerve**
16 Esophagus
17 Left subclavian artery
18 Left common carotid artery
19 Second rib
20 Tongue
21 Superior cervical ganglion
22 Sympathetic trunk
23 Inferior constrictor muscle of pharynx
24 **Inferior thyroid artery**
25 Glossopharyngeal nerve
26 Superior laryngeal nerve
27 Epiglottis
28 Posterior cricoarytenoid muscle and cricoid cartilage
29 Inferior laryngeal branch of recurrent laryngeal nerve
30 **Thyroid gland**
31 Superior thyroid artery
32 **Thyrocervical trunk**
33 Internal thoracic artery
34 **Phrenic nerve**
35 Hypoglossal nerve
36 Transverse cervical artery
37 Middle cervical ganglion
38 Middle cervical cardiac nerves (branches of sympathetic trunk)
39 Ligamentum arteriosum

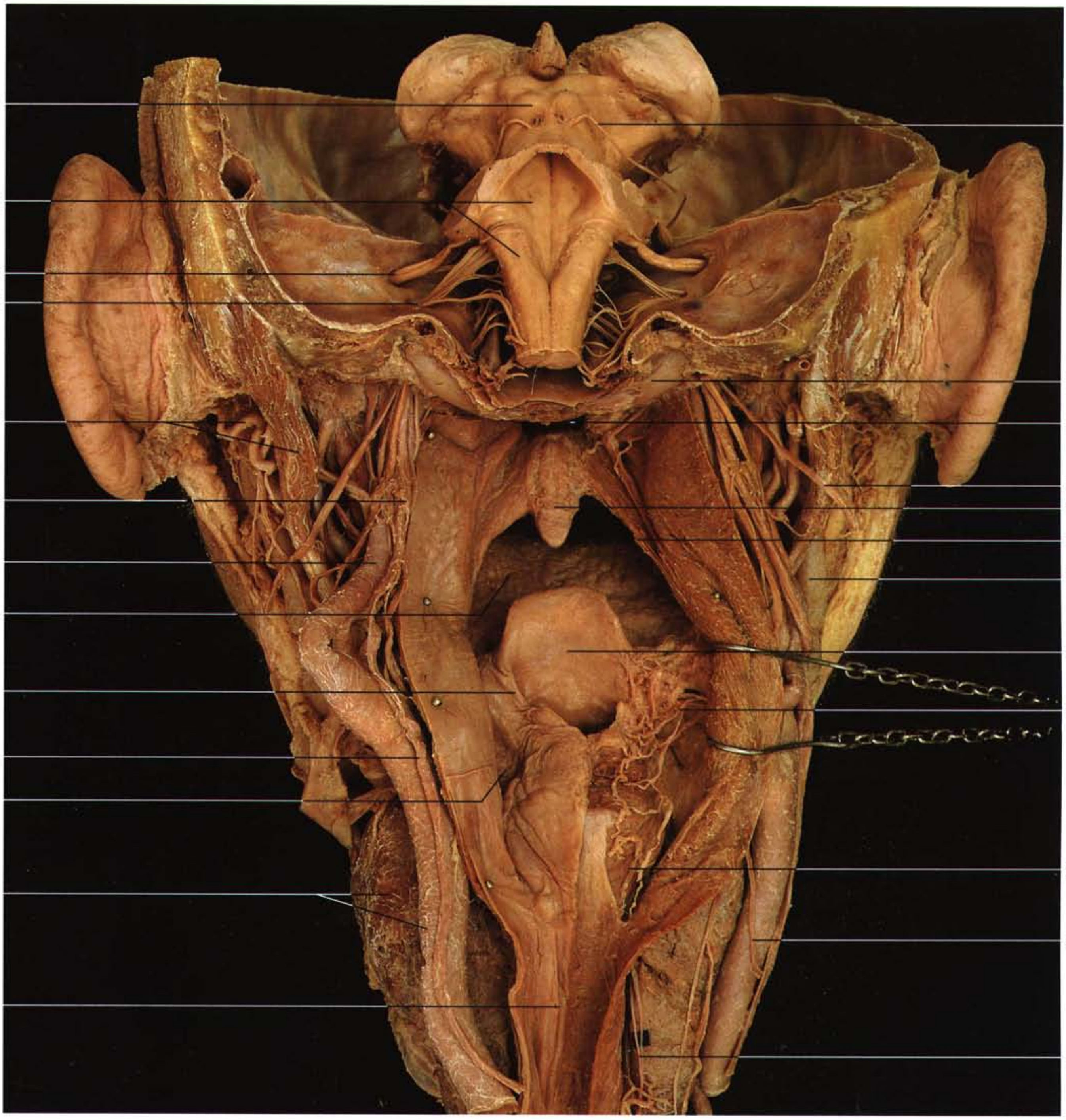

Larynx and oral cavity (posterior aspect). Mucous membrane on the right half of pharynx has been removed.

1 Midbrain (inferior colliculus)
2 Rhomboid fossa and medulla oblongata
3 Vestibulocochlear and facial nerve
4 Glossopharyngeal, vagus and accessory nerve
5 Occipital artery and posterior belly of digastric muscle
6 Superior cervical ganglion
7 **Internal carotid artery**
8 **Oral cavity** (tongue)
9 Aryepiglottic fold
10 Vagus nerve
11 **Piriform recess**
12 Thyroid gland and common carotid artery
13 Esophagus
14 Trochlear nerve
15 Occipital condyle
16 **Nasal cavity** (choana)
17 Accessory nerve
18 **Uvula** and soft palate
19 Palatopharyngeus muscle
20 External carotid artery
21 Epiglottis
22 **Internal branch of superior laryngeal nerve**
23 **Inferior laryngeal nerve**
24 Ansa cervicalis

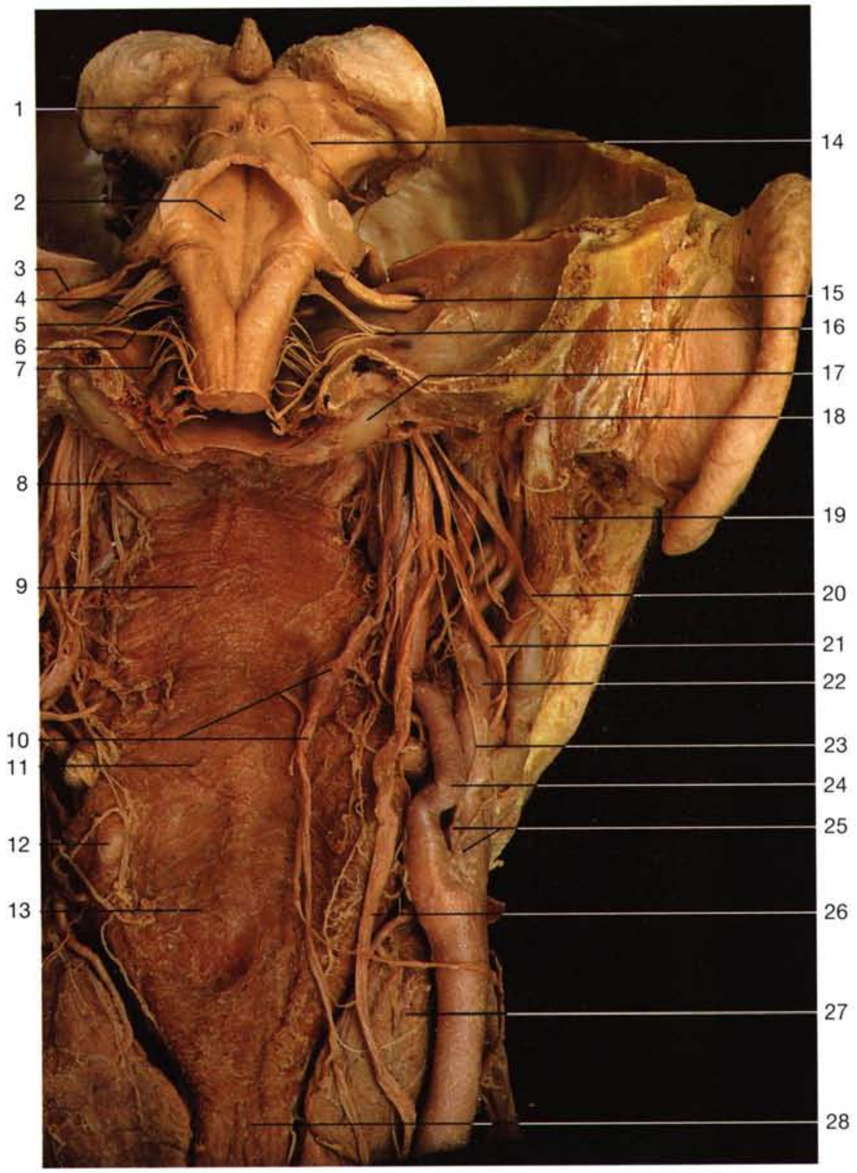

Pharynx and parapharyngeal nerves in connection with brain stem (posterior aspect).

1 Inferior colliculus of midbrain
2 Facial colliculus in **floor of rhomboid fossa**
3 Vestibulocochlear and facial nerves
4 Glossopharyngeal nerve
5 Vagus nerve
6 Accessory nerve
7 Hypoglossal nerve
8 Pharyngobasilar fascia
9 Superior constrictor muscle of pharynx
10 **Sympathetic trunk** and **superior cervical ganglion** (medially displaced)
11 Middle constrictor muscle of pharynx
12 Greater cornu of hyoid bone
13 Inferior constrictor muscle of pharynx
14 Trochlear nerve
15 **Internal acoustic meatus** with facial and vestibulocochlear nerves
16 **Jugular foramen** with glossopharyngeal, vagus and assessory nerves
17 Occipital condyle
18 Occipital artery
19 Posterior belly of digastric muscle
20 **Accessory nerve** (extracranial part)
21 **Hypoglossal nerve** (extracranial part)
22 External carotid artery
23 Carotid sinus nerve
24 **Internal carotid artery**
25 **Carotid sinus and carotid body**
26 Vagus nerve
27 Thyroid gland
28 Esophagus
29 Choanae
30 Medial pterygoid plate
31 Foramen lacerum
32 **Pharyngeal tubercle**
33 Hard palate
34 Greater and lesser palatine foramen
35 Pterygoid hamulus
36 Lateral pterygoid plate
37 Pterygoid canal
38 Foramen ovale
39 Mandibular fossa
40 Carotid canal
41 Styloid process and stylomastoid foramen

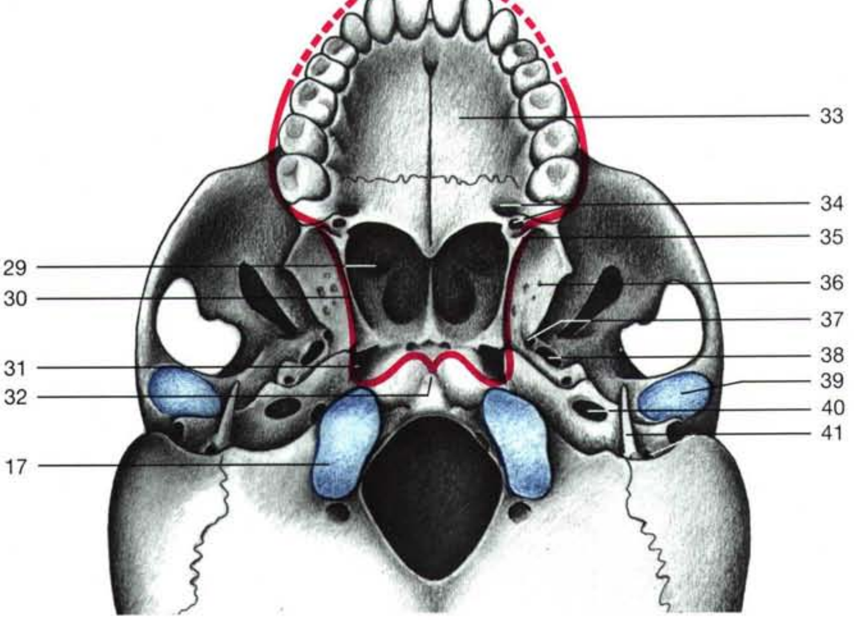

Inferior aspect of the skull. Red line = outline of superior constrictor muscle in continuation with buccinator muscle and orbicularis oris muscle. (Semischematic drawing.)

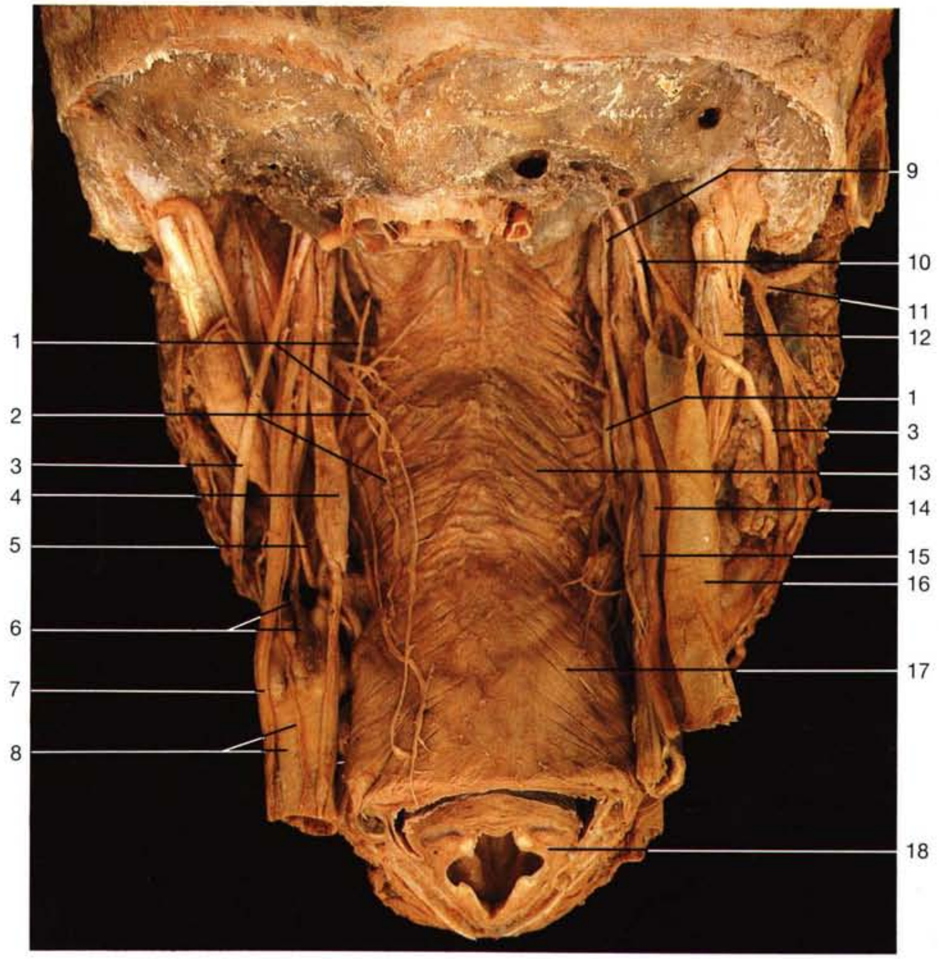

Parapharyngeal nerves and vessels. Dorsal aspect of the pharynx.

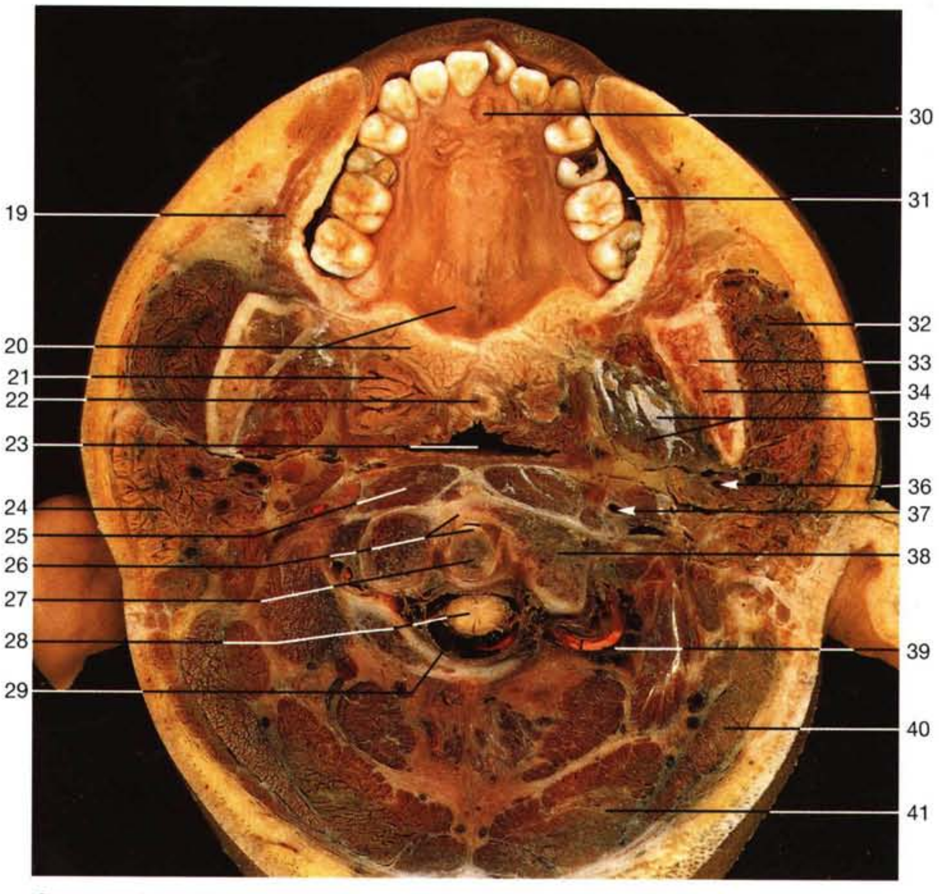

Cross-section of head and neck at the level of the atlas (inferior aspect).

1 Ascending pharyngeal artery
2 Pharyngeal plexus
3 **Accessory nerve**
4 Superior cervical ganglion of sympathetic trunk
5 Superior laryngeal nerve
6 Carotid body and carotid sinus nerve
7 Left **vagus nerve**
8 Common carotid artery and cardiac branch of vagus nerve
9 **Glossopharyngeal nerve**
10 **Hypoglossal nerve**
11 **Facial nerve**
12 Posterior belly of digastric muscle
13 Middle constrictor muscle of pharynx
14 Right vagus nerve
15 **Sympathetic trunk**
16 **Internal jugular vein**
17 Inferior constrictor muscle of pharynx
18 Larynx
19 **Buccinator muscle**
20 Soft palate and palatine glands
21 Palatine tonsil
22 Uvula of palate
23 **Pharynx** (oral part)
24 Parotid gland
25 Longus capitis muscle
26 Medioatlantoaxial joint and anterior arch of atlas
27 **Dens of axis**
28 **Spinal cord**
29 Dura mater
30 Incisive papilla
31 **Oral vestibule**
32 Masseter muscle
33 Mandible
34 Mandibular canal with vessels and nerve
35 Medial pterygoid muscle
36 External carotid artery
37 Internal carotid artery
38 **Atlas**
39 **Vertebral artery**
40 Splenius capitis muscle
41 Semispinalis capitis muscle

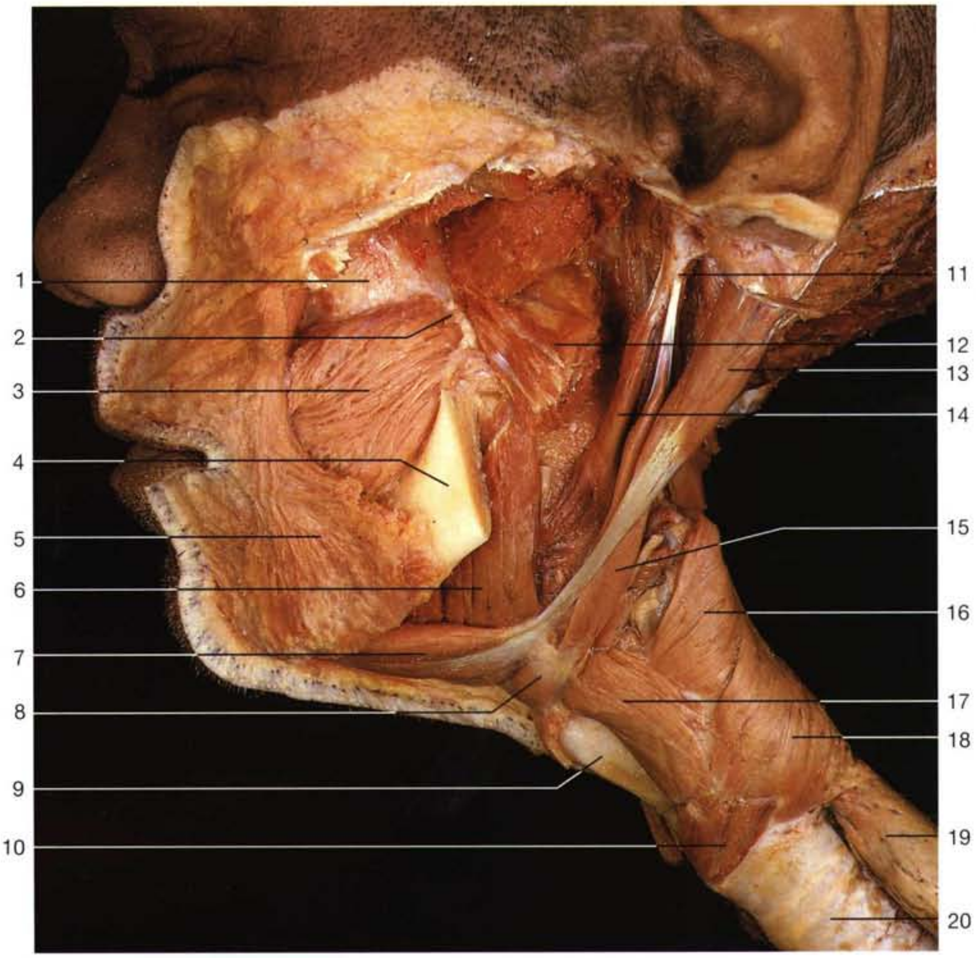

Dissection of pharynx, supra- and infrahyoid muscles I. Mandible partly removed (lateral aspect).

1 Maxilla
2 Pterygomandibular raphe
3 Buccinator muscle
4 Mandible (divided)
5 Depressor anguli oris muscle
6 Mylohyoid muscle
7 Anterior belly of digastric muscle
8 Hyoid bone
9 Thyroid cartilage
10 Cricothyroid muscle
11 Styloid process
12 Medial pterygoid muscle (divided)
13 Posterior belly of digastric muscle
14 Styloglossus muscle
15 Stylohyoid muscle
16 **Thyropharyngeal part of inferior constrictor muscle of pharynx**
17 Thyrohyoid muscle
18 **Cricopharyngeal part of inferior constrictor muscle of pharynx**
19 Esophagus
20 Trachea
21 First molar of maxilla
22 Tongue
23 Inferior longitudinal muscle of tongue
24 Genioglossus muscle
25 **Superior constrictor muscle of pharynx**
26 Hypoglossal nerve
27 Hyoglossus muscle
28 Superior laryngeal nerve and superior laryngeal artery

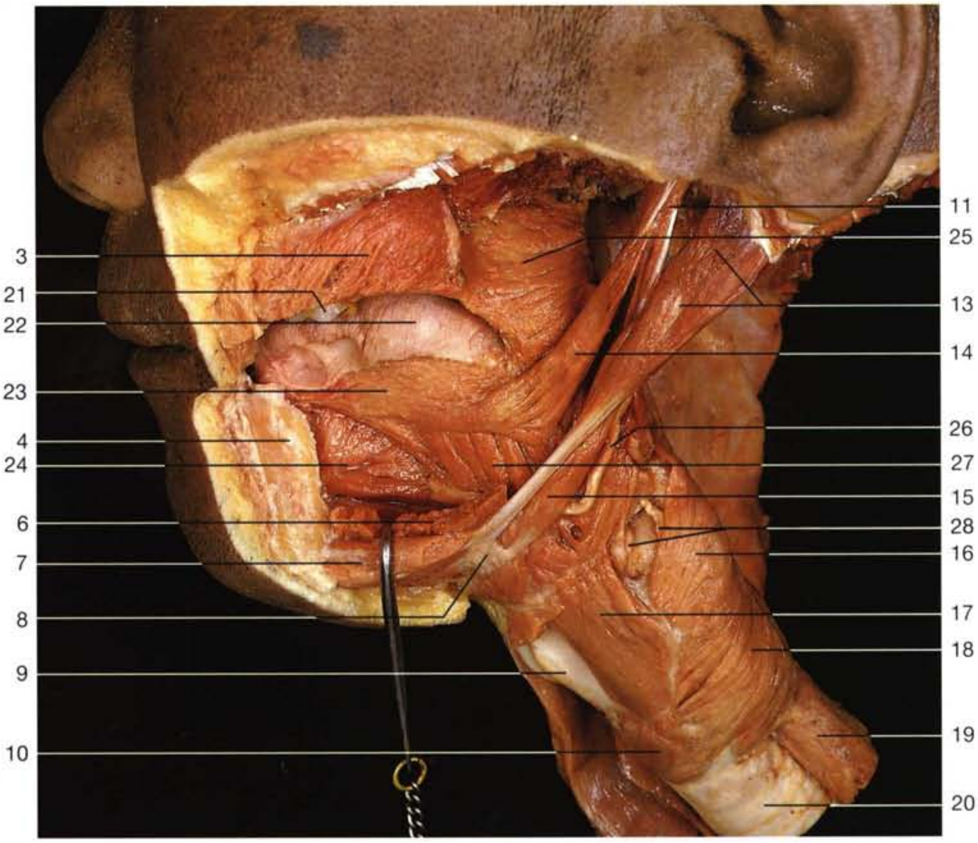

Dissection of pharynx, supra- and infrahyoid muscles II. Oral cavity opened (lateral aspect).

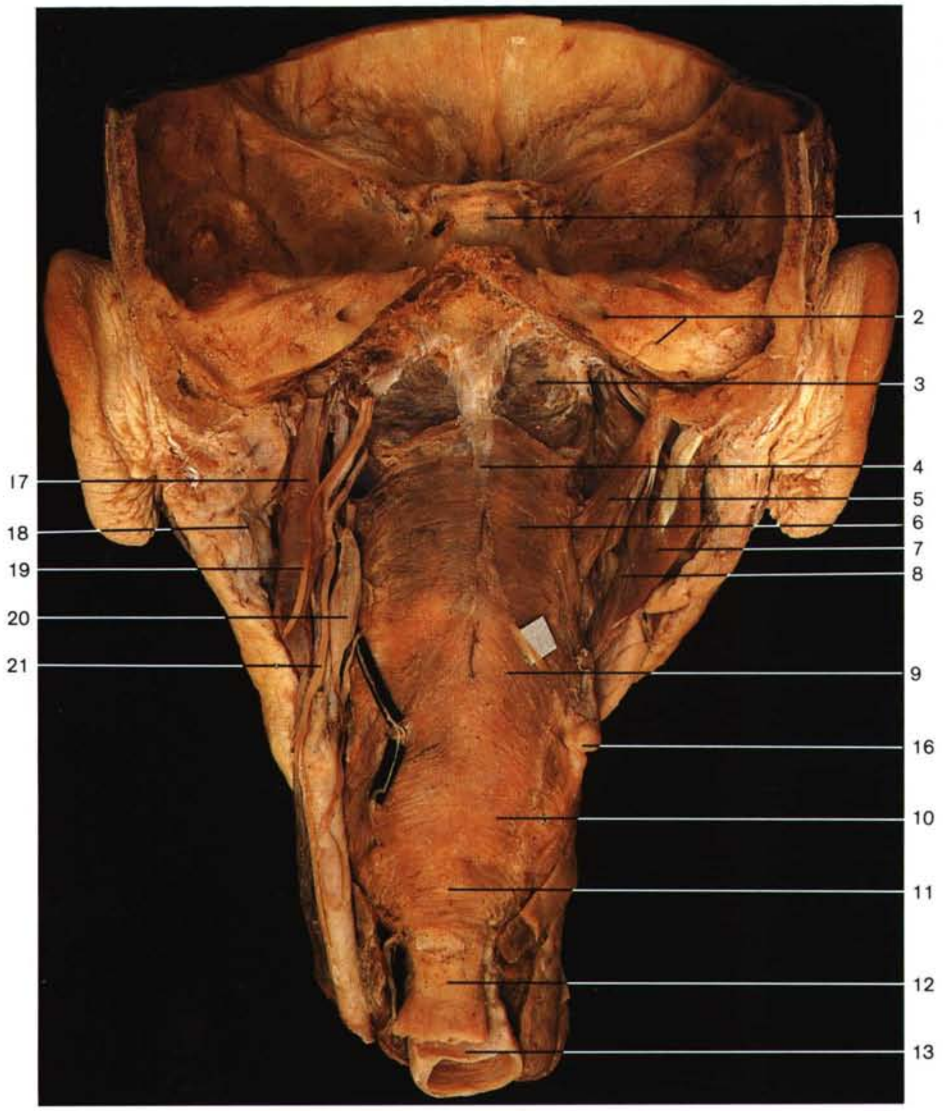

Muscles of the pharynx (posterior aspect).

1 Sella turcica
2 Internal acoustic meatus and petrous part of temporal bone
3 Pharyngobasilar fascia
4 Fibrous raphe of pharynx
5 **Stylopharyngeus muscle**
6 **Superior constrictor muscle of pharynx**
7 Posterior belly of digastric muscle
8 Stylohyoid muscle
9 **Middle constrictor muscle of pharynx**
10 **Inferior constrictor muscle of pharynx**
11 Muscle-free area (Killian's triangle)
12 Esophagus
13 Trachea
14 Thyroid and parathyroid glands
15 Medial pterygoid muscle
16 Greater horn of hyoid bone
17 Internal jugular vein
18 Parotid gland
19 Accessory nerve
20 **Superior cervical ganglion** of sympathetic trunk
21 Vagus nerve
22 Laimer's triangle (area prone to developing diverticula)
23 Orbicularis oculi muscle
24 Nasalis muscle
25 Levator labii superioris and levator labii alaeque nasi muscles
26 Levator anguli oris muscle
27 Orbicularis oris muscle
28 **Buccinator muscle**
29 Depressor labii inferioris muscle
30 **Hyoglossus muscle**
31 Thyrohyoid muscle
32 Thyroid cartilage
33 Cricothyroid muscle
34 Pterygomandibular raphe
35 Tensor veli palatini muscle
36 Levator veli palatini muscle
37 Depressor anguli oris muscle
38 Mentalis muscle
39 Styloglossus muscle

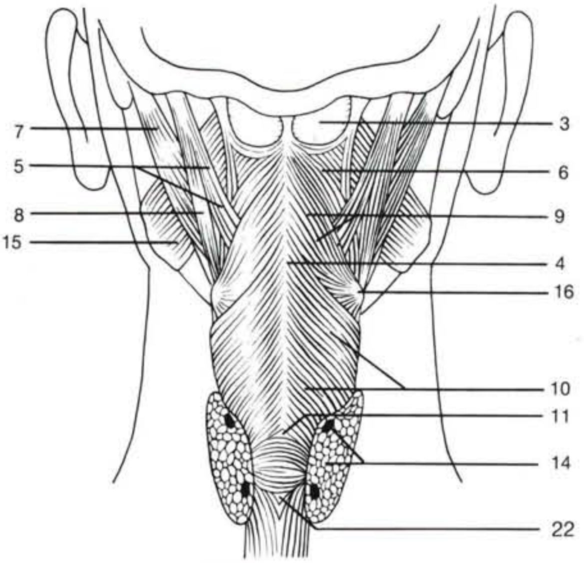

Muscles of the pharynx. (Schematic drawing.)

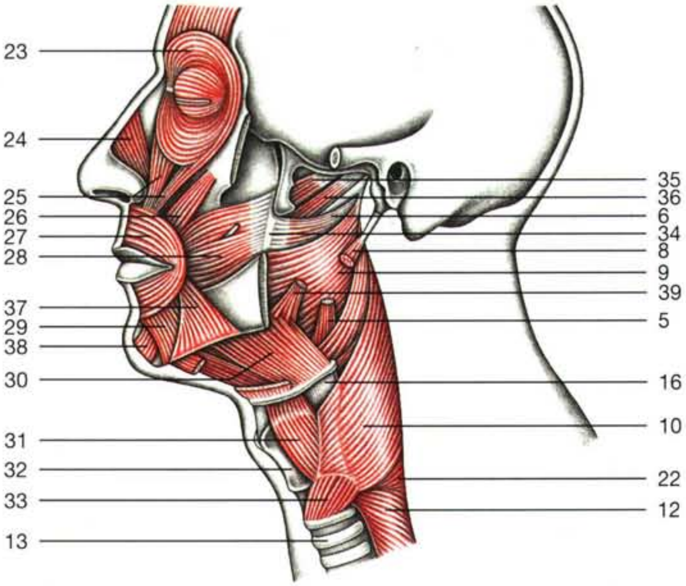

Muscles of pharynx (lateral aspect). (Schematic drawing.)

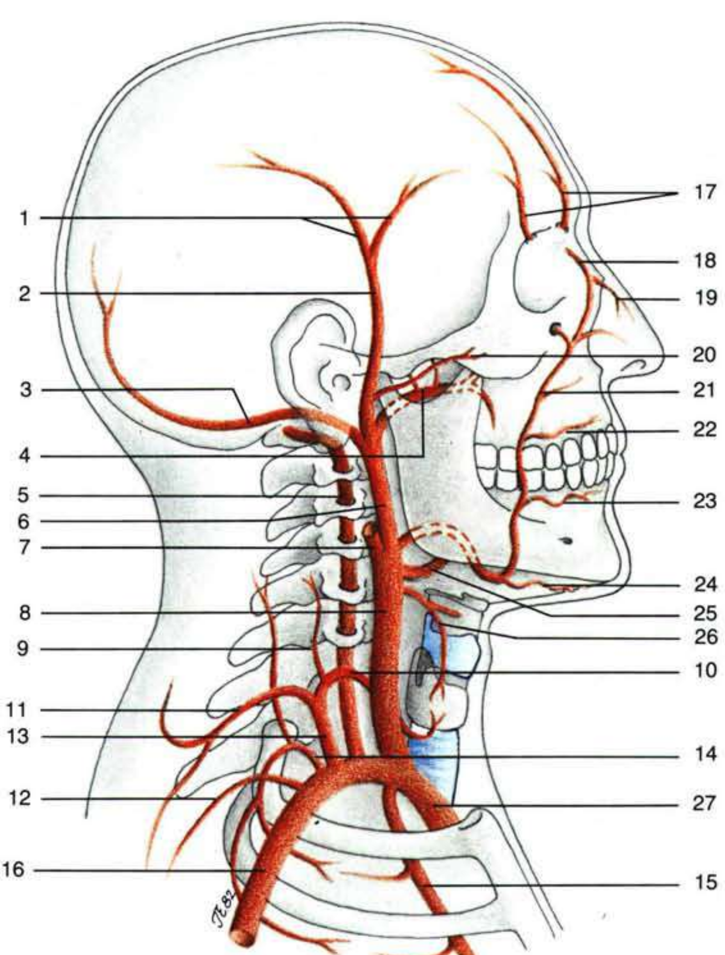

Arteries of head and neck. Diagram of the main branches of external carotid and subclavian artery.

1 Frontal and parietal branches of superficial temporal artery
2 Superficial temporal artery
3 **Occipital artery**
4 **Maxillary artery**
5 **Vertebral artery**
6 **External carotid artery**
7 Internal carotid artery
8 Common carotid artery (divided)
9 Ascending cervical artery
10 Inferior thyroid artery
11 Transverse cervical artery with 2 branches (superficial cervical artery and descending scapular artery)
12 Suprascapular artery
13 **Thyrocervical trunk**
14 Costocervical trunk with 2 branches (deep cervical artery and superior intercostal artery)
15 Internal thoracic artery
16 **Axillary artery**
17 Supraorbital and supratrochlear arteries
18 Angular artery
19 Dorsal nasal artery
20 Transverse facial artery
21 **Facial artery**
22 Superior labial artery
23 Inferior labial artery
24 Submental artery
25 Lingual artery
26 Superior thyroid artery
27 Brachiocephalic trunk

to page 165 ▷

1 Galea aponeurotica
2 Frontal branch } of superficial
3 Parietal branch } temporal artery
4 Superior auricular muscle
5 **Superficial temporal artery and vein**
6 Middle temporal artery
7 Auriculotemporal nerve
8 Branches of facial nerve
9 Facial nerve
10 **External carotid artery** within the retromandibular fossa
11 Posterior belly of digastric muscle
12 Sternocleidomastoid artery
13 Sympathetic trunk and superior cervical ganglion
14 Sternocleidomastoid muscle (divided and reflected)
15 Clavicle (divided)
16 Transverse cervical artery
17 Ascending cervical artery and phrenic nerve
18 Scalenus anterior muscle
19 Suprascapular artery
20 Dorsal scapular artery
21 Brachial plexus and **axillary artery**
22 Thoracoacromial artery
23 Lateral thoracic artery
24 Median nerve (displaced) and pectoralis minor muscle (reflected)
25 Frontal belly of occipitofrontalis muscle
26 Orbital part of orbicularis oculi muscle
27 **Angular artery** and vein
28 **Facial artery**
29 Superior labial artery
30 Zygomaticus major muscle
31 Inferior labial artery
32 Parotid duct
33 Buccal fat pad
34 **Maxillary artery**
35 Masseter muscle
36 Facial artery and mandible
37 Submental artery
38 Anterior belly of digastric muscle

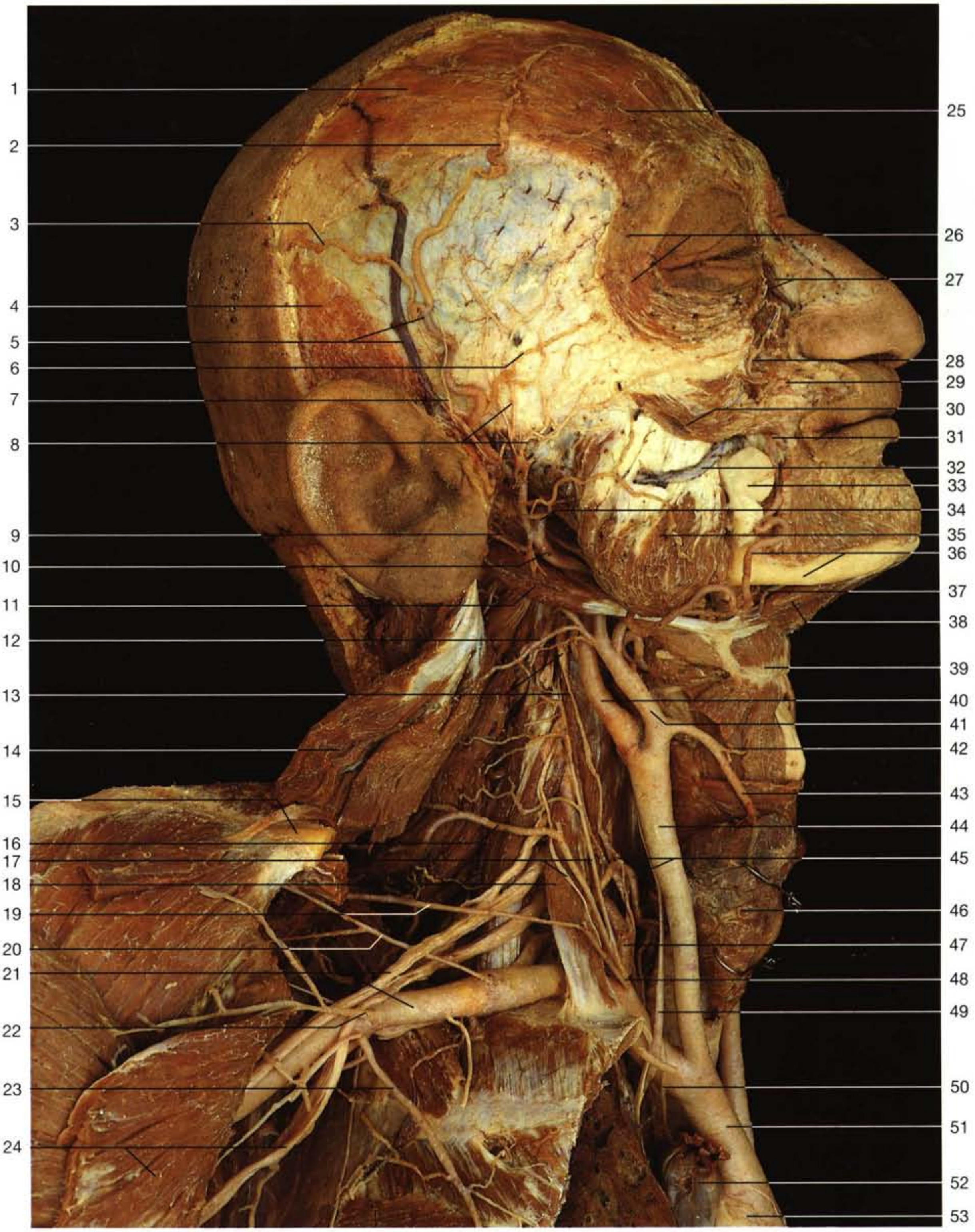
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

Main branches of head and neck arteries (lateral aspect). Anterior thoracic wall and clavicle partly removed; pectoralis muscles have been reflected to display the subclavian and axillary arteries.

39 Hyoid bone
40 **Internal carotid artery**
41 **External carotid artery**
42 Superior laryngeal artery
43 Superior thyroid artery
44 **Common carotid artery**
45 Thyroid ansa of sympathetic trunk and **inferior thyroid artery**

46 Thyroid gland (right lobe)
47 **Vertebral artery**
48 **Thyrocervical trunk**
49 Vagus nerve
50 Ansa subclavia of sympathetic trunk
51 **Brachiocephalic trunk**
52 Superior vena cava (divided)
53 Aortic arch

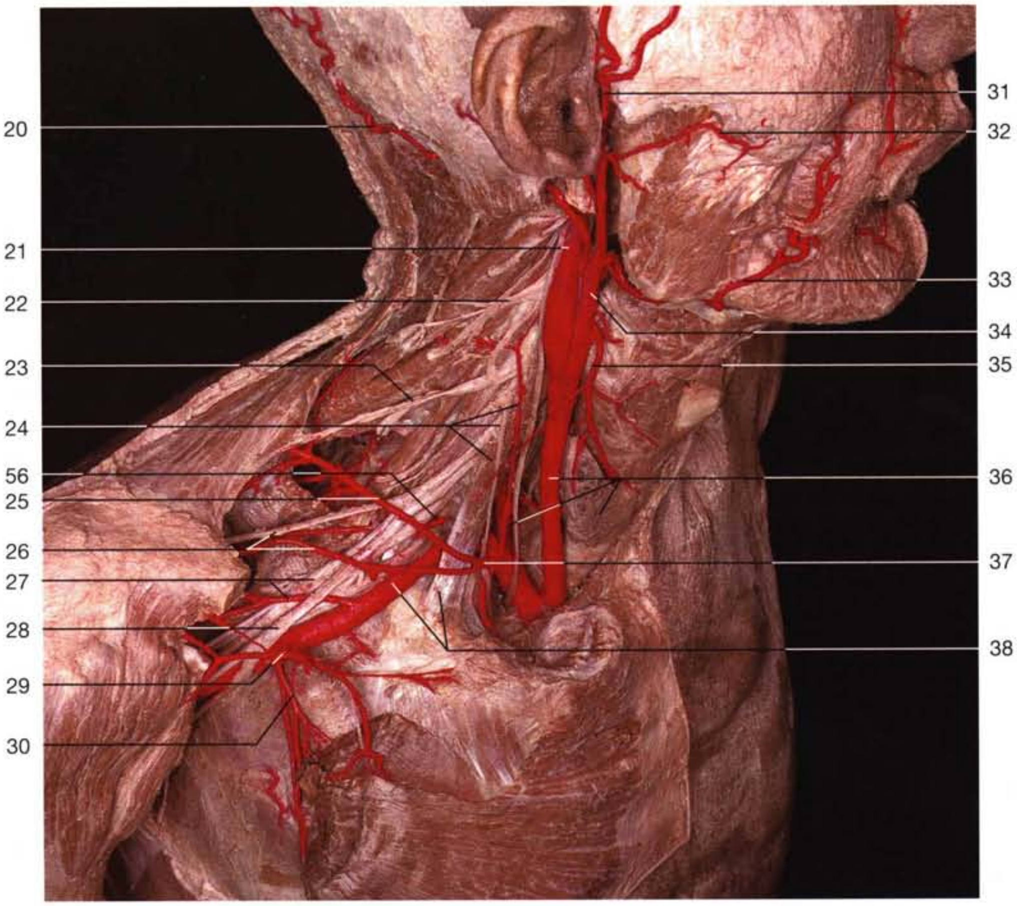

Arteries of head and neck (anterior-lateral aspect). Clavicle, sternocleidomastoid and veins have been partly removed, the arteries were colored.

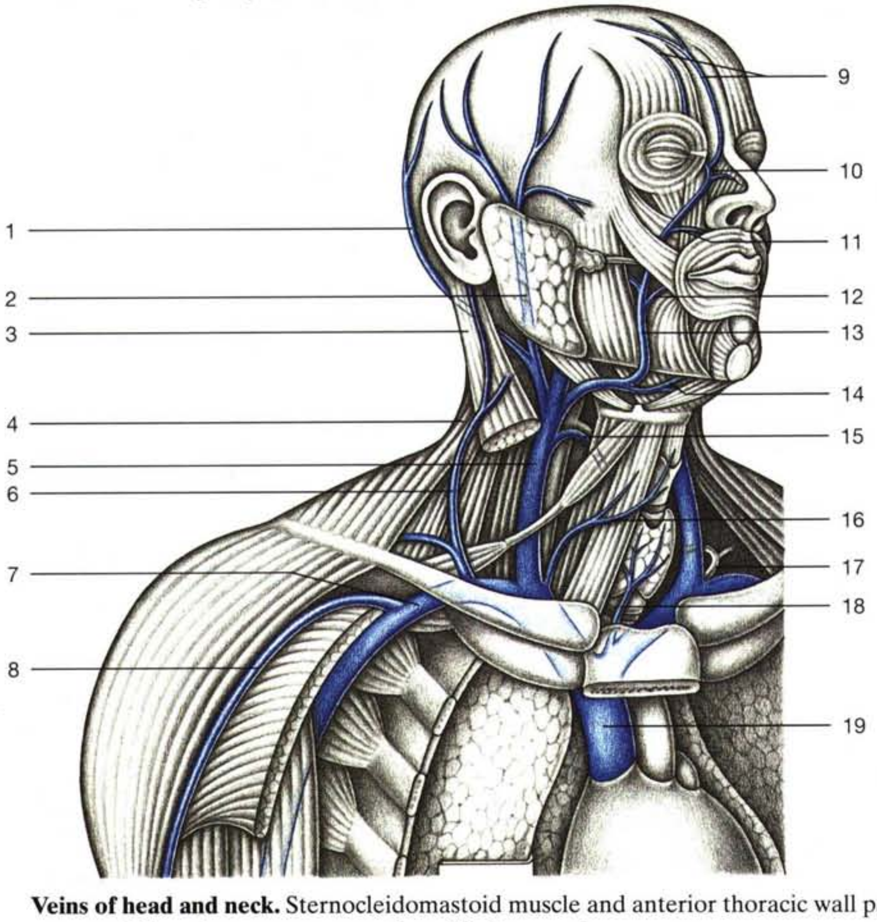

Veins of head and neck. Sternocleidomastoid muscle and anterior thoracic wall partly removed. Note the venous connection with the superior vena cava.

1 Occipital vein
2 Superficial temporal vein
3 Sternocleidomastoid muscle
4 Trapezius muscle
5 **Internal jugular vein**
6 External jugular vein
7 **Subclavian vein**
8 Cephalic vein
9 Supraorbital veins
10 Angular vein
11 Superior labial vein
12 Inferior labial vein
13 **Facial vein**
14 Submental vein
15 Superior thyroid vein
16 Anterior jugular vein
17 Thoracic duct
18 Inferior thyroid vein
19 Superior vena cava
20 Occipital artery
21 **Internal carotid artery**
22 Cervical plexus
23 Supraclavicular nerve
24 Phrenic nerve and ascending cervical artery on scalenus anterior muscle
25 Superficial cervical artery
26 Suprascapular artery and nerve
27 Brachial plexus and anterior circumflex humeral artery
28 Lateral cord of brachial plexus
29 **Thoracoacromial artery**
30 Lateral thoracic artery
31 Superficial temporal artery
32 Transverse facial artery
33 Facial artery
34 External carotid artery
35 Superior thyroid artery
36 **Common carotid artery, vagus nerve** and thyroid gland
37 **Thyrocervical trunk**
38 **Subclavian artery** and scalenus anterior muscle
39 Parotid gland and facial nerve
40 Great auricular nerve
41 External jugular vein
42 Brachial plexus
43 Cephalic vein in deltopectoral groove
44 Axillary vein and artery
45 Right brachiocephalic vein
46 **Superior vena cava**
47 Right lung (reflected)
48 Superficial temporal artery and vein
49 Facial artery and vein
50 Cervical branch of facial nerve and submandibular gland
51 **Internal jugular vein,** common carotid artery and omohyoid muscle
52 Anterior jugular vein and thyroid gland
53 Jugular venous arch
54 Left brachiocephalic vein
55 Pericardium of heart (location of right atrium)
56 Transverse cervical artery

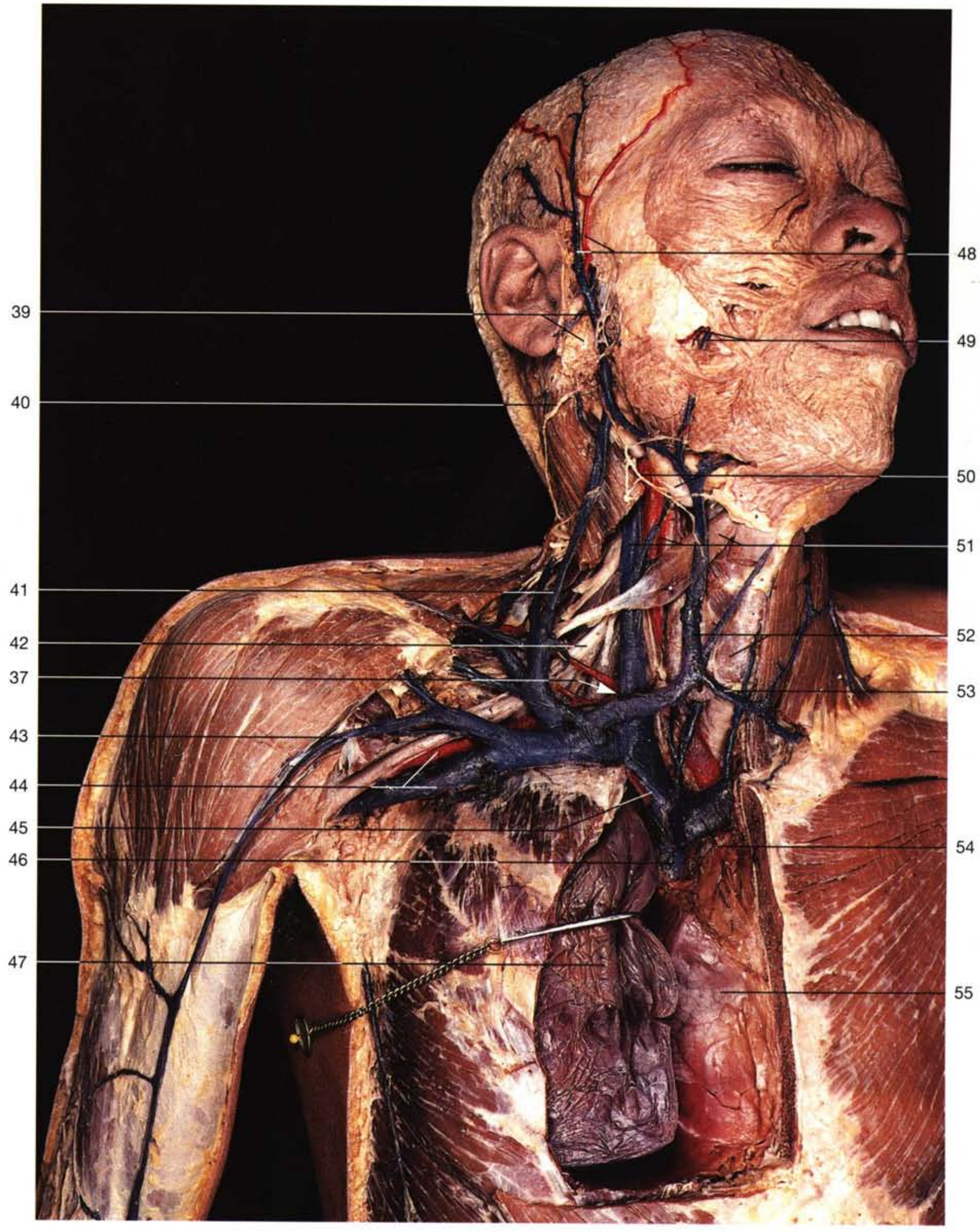

Veins of head and neck (anterior aspect). Part of the thoracic wall, clavicle and sternocleidomastoid muscle have been removed. Veins were colored blue, arteries red.

The **internal jugular vein** is the continuation of the sigmoid sinus which drains most of the venous blood from the brain together with the external cerebrospinal fluid. By joining the subclavian vein it forms the right brachiocephalic vein which continues on the right side directly into the superior vena cava. The common way to introduce the lead from a pacemaker device into the heart is by way of the cephalic vein. On the left side the thoracic duct joins the internal jugular vein at that point where the subclavian vein and the internal jugular vein form the left brachiocephalic vein. Note that the subclavian vein lies in front of the scalenus anterior muscle whereas the subclavian artery together with the plexus brachialis lies posterior to that muscle. The cephalic vein joins the axillary vein by passing into the deltopectoral triangle. The subclavian vein is strongly fixed to the first rib so that it can be punctured with a needle at that point (underneath the sternal end of clavicle) to introduce a catheter (subclavian line).

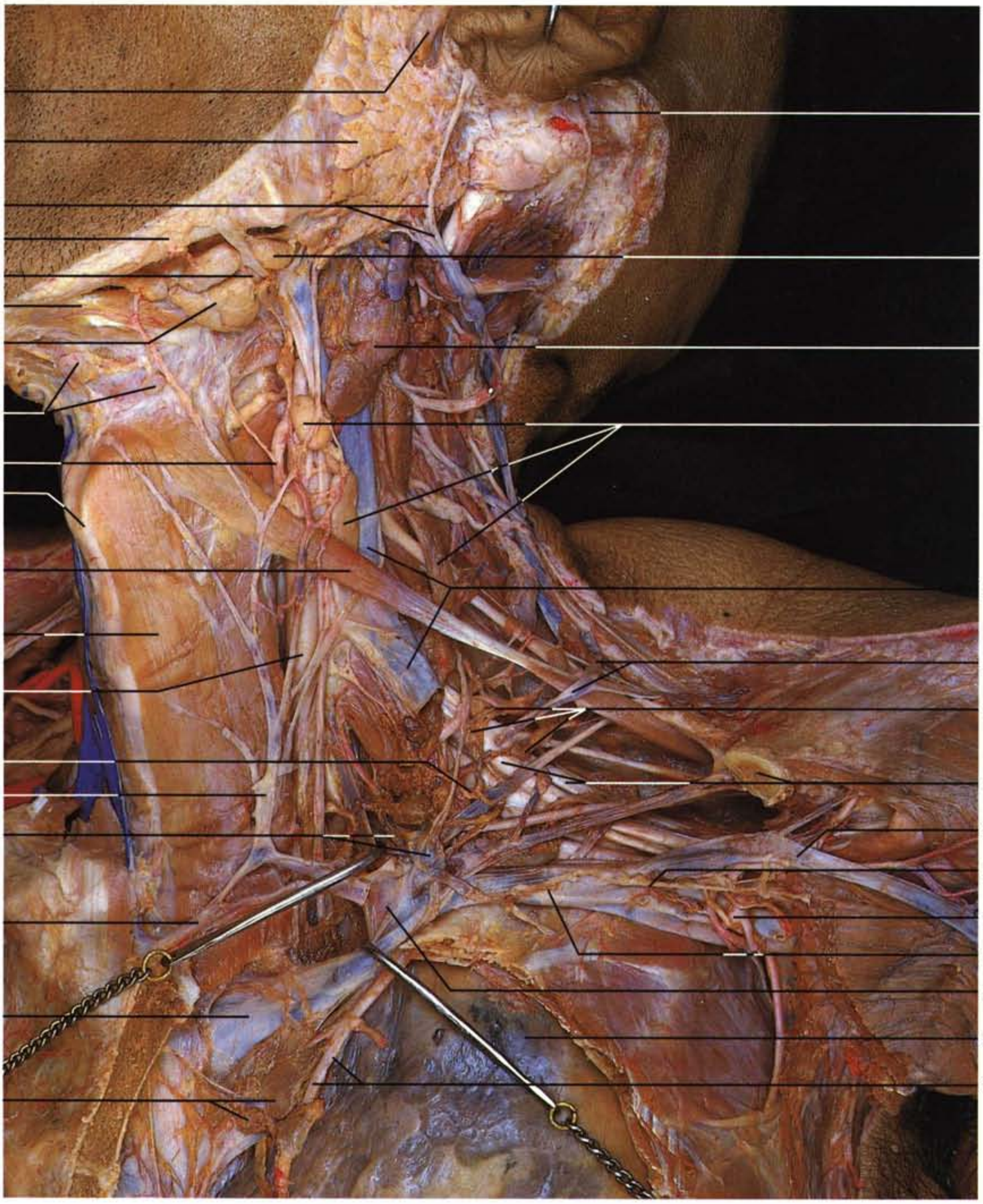

Lymph nodes and lymph vessels of the neck, left side oblique (lateral aspect). The sternocleidomastoid muscle and the left half of the thoracic wall have been removed. Lower part of the internal jugular vein has been cut and laterally displaced to show the thoracic duct.

1	**Superficial parotid lymph node**
2	Parotid gland
3	Great auricular nerve
4	Mandible
5	Facial vein
6	Anterior belly of digastric muscle
7	Submandibular gland
8	**Submental lymph nodes**
9	Superior thyroid artery
10	Thyroid cartilage
11	Omohyoid muscle
12	Sternohyoid muscle
13	Common carotid artery
14	**Supraclavicular lymph nodes**
15	Anterior jugular vein
16	Thoracic duct
17	Jugular venous arch
18	Left brachiocephalic vein
19	**Superior mediastinal lymph nodes**
20	Retroauricular lymph nodes
21	**Submandibular nodes**
22	**Superficial cervical lymph nodes**
23	**Jugulodigastric lymph nodes and jugular trunk**
24	**Internal jugular vein**
25	External jugular vein
26	**Jugulo-omohyoid lymph nodes**
27	Brachial plexus
28	Cephalic vein
29	Subclavian trunk
30	**Infraclavicular lymph nodes**
31	Subclavian vein
32	Lung
33	Internal thoracic artery and vein

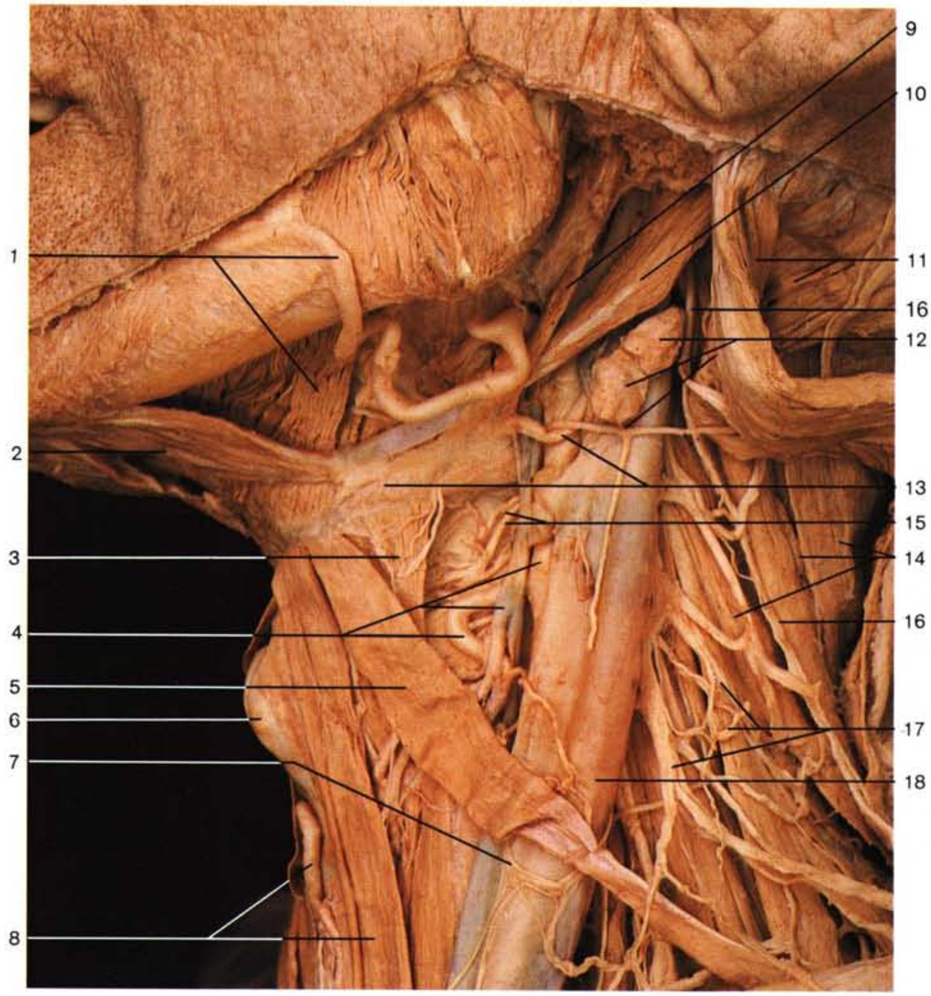

Carotid triangle, left side (lateral aspect). Sternocleidomastoid muscle reflected.

1 Mylohyoid muscle and facial artery
2 Anterior belly of digastric muscle
3 Thyrohyoid
4 **External carotid artery,** superior thyroid artery and vein
5 Omohyoid muscle
6 Thyroid cartilage
7 **Ansa cervicalis**
8 Sternohyoid muscle and superior thyroid artery
9 Stylohyoid muscle
10 Posterior belly of digastric muscle
11 Sternocleidomastoid muscle (reflected)
12 **Superior cervical lymph nodes** and sternocleidomastoid artery
13 Hyoid bone and hypoglossal nerve (n. XII)
14 Splenius capitis and levator scapulae muscles
15 Superior laryngeal artery and internal branch of superior laryngeal nerve
16 **Accessory nerve**
17 Cervical plexus
18 **Internal jugular vein**
19 Facial vein
20 Submental nodes
21 **Thoracic duct**
22 Retroauricular nodes
23 Parotid nodes
24 Occipital nodes
25 **Submandibular nodes**
26 Jugulodigastric nodes } **deep cervical**
27 Jugulo-omohyoid nodes } **nodes**
28 Jugular trunk
29 Subclavian trunk
30 Infraclavicular nodes
31 External jugular vein

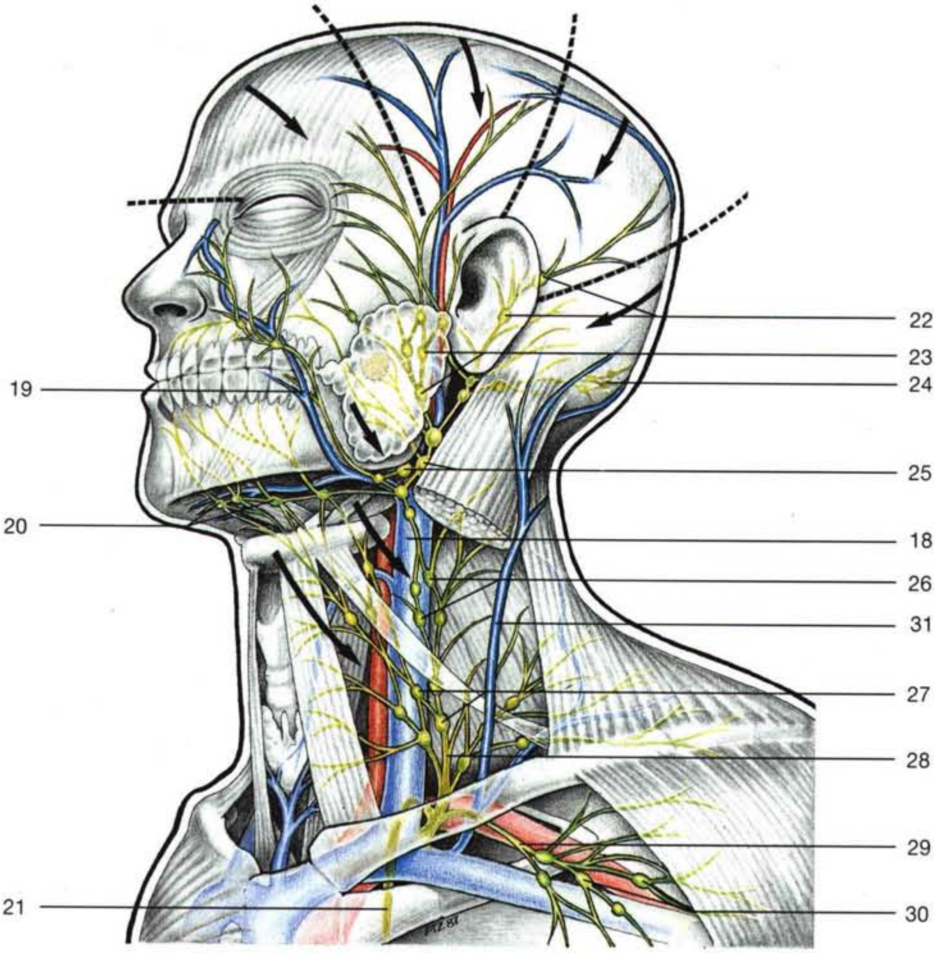

Lymph nodes and veins of head and neck. Dotted lines = border between irrigation areas; arrows: direction of lymph flow.

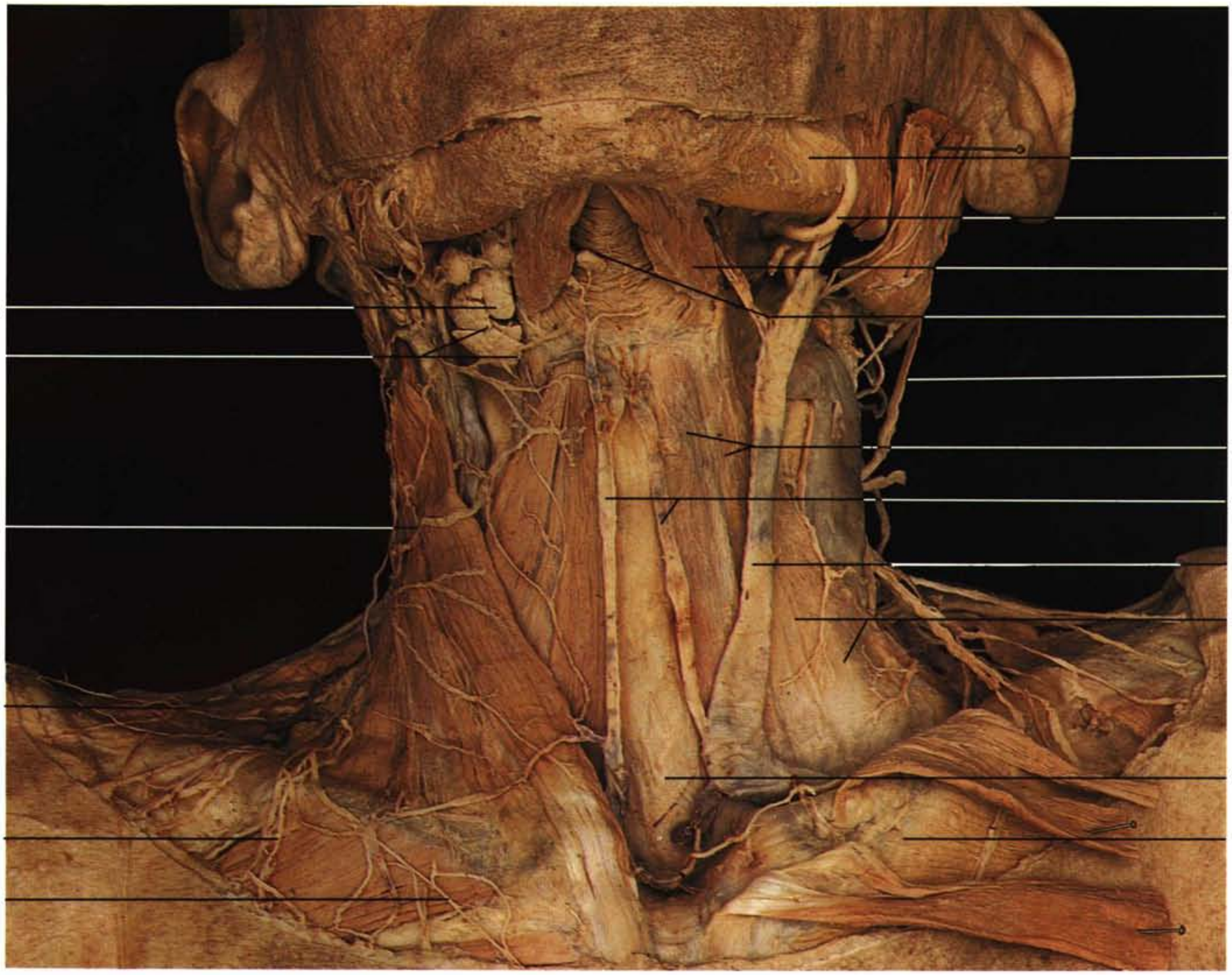

Neck (anterior aspect).
The superficial fascia has been removed.

1 Mandible
2 **Facial artery and vein**
3 Anterior belly of digastric muscle
4 Mylohyoid muscle
5 Infrahyoid muscles (sternohyoid, sternothyroid and omohyoid)
6 **Anterior jugular veins**
7 **External jugular vein**
8 Sternocleidomastoid muscle
9 **Thyroid gland**
10 **Submandibular gland**
11 Cervical branch of facial nerve
12 Great auricular nerve
13 Transverse cervical nerves
14 Lateral supraclavicular nerves
15 Middle supraclavicular nerves
16 Medial supraclavicular nerves
(12–16) Cutaneous branches of cervical plexus
17 Clavicle
18 Platysma muscle
19 Prevertebral lamina of cervical fascia, covering longus colli muscle
20 Vertebral artery and vein
21 Scalenus muscles
22 Trapezius muscle
23 **Superficial lamina of cervical fascia**
24 **Pretracheal lamina of cervical fascia**
25 **Prevertebral lamina of cervical fascia** with sympathetic trunk
26 **Carotid sheath** with common carotid artery, internal jugular vein and vagus nerve
27 Cervical part of sympathetic trunk
28 Carotid sheath

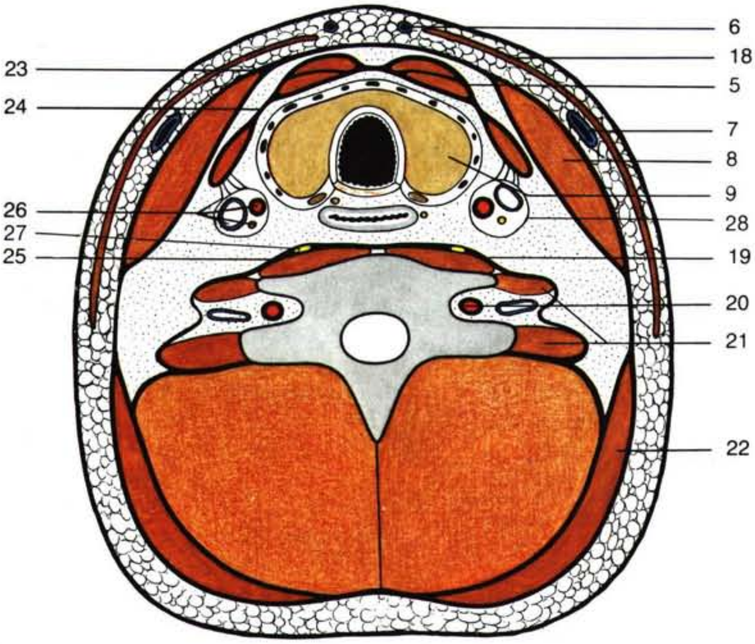

Cross-section of the neck at the level of the thyroid gland.
Notice the postion of the three laminae of cervical fascia (23, 24, 25).

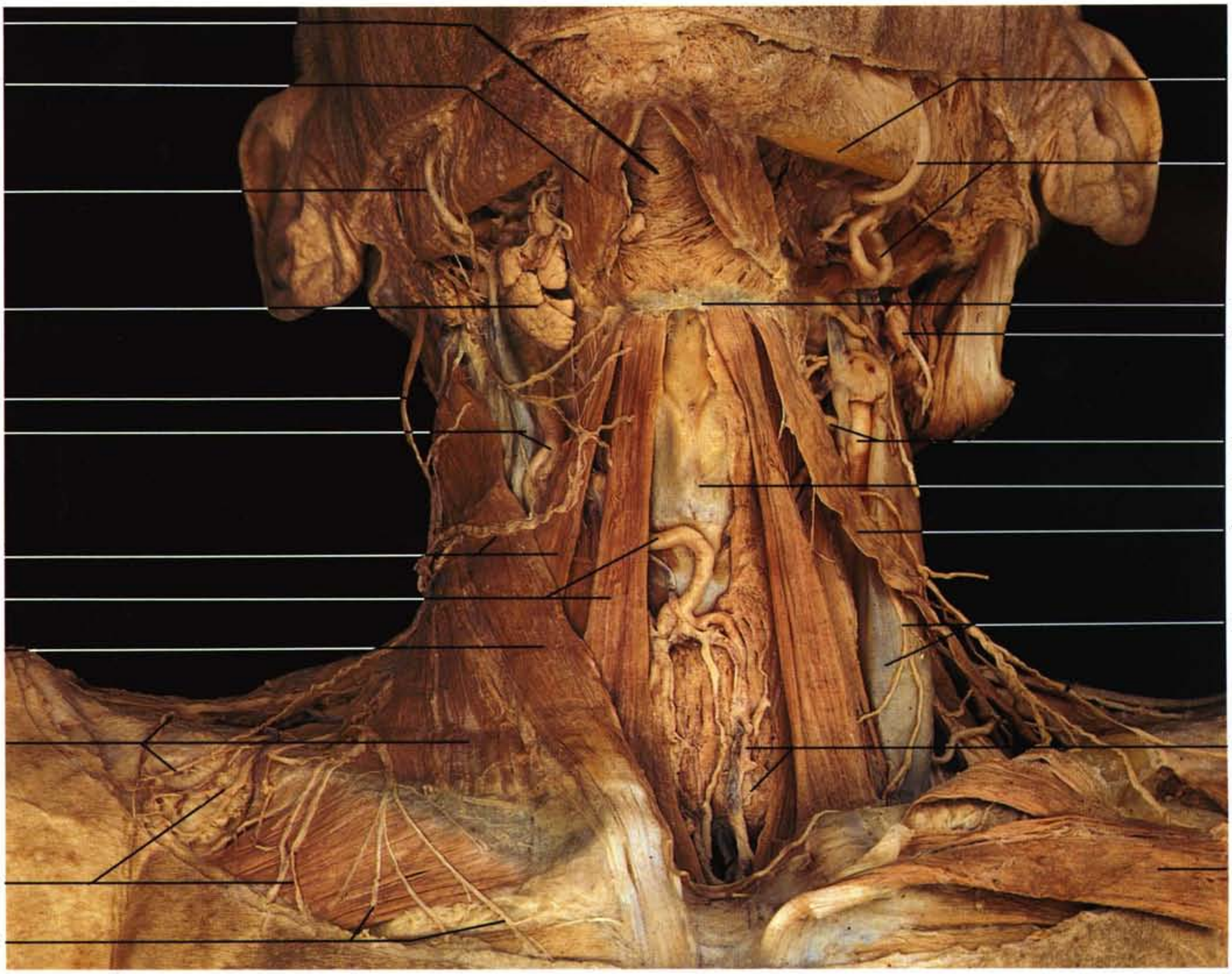

Anterior triangle (anterior aspect). The pretracheal lamina of cervical fascia and left sternocleidomastoid muscle have been removed.

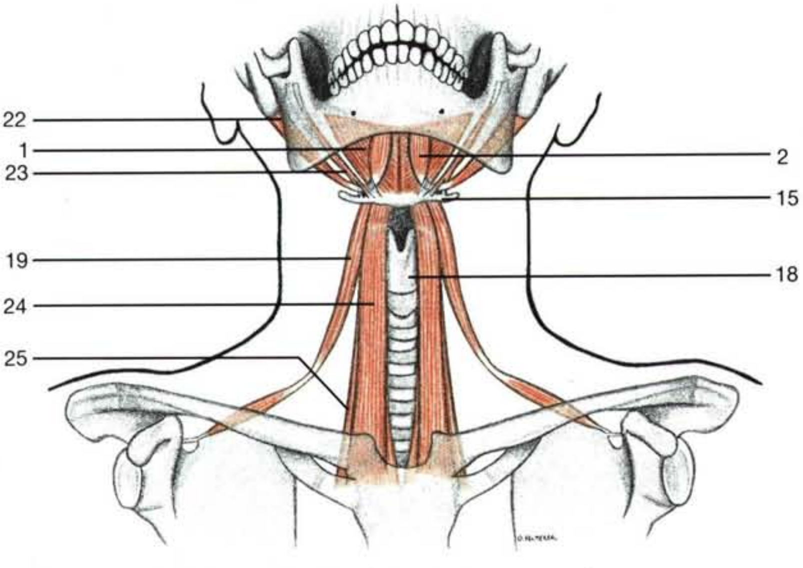

Supra- and infrahyoid muscles. (Schematic drawing.)

1 Mylohyoid muscle
2 Anterior belly of digastric muscle
3 **Facial artery**
4 **Submandibular gland**
5 Great auricular nerve
6 Internal jugular vein and common carotid artery
7 Transverse cervical nerve and omohyoid muscle
8 **Sternohyoid muscle** and **superior thyroid artery**
9 Sternocleidomastoid muscle (sternal head)
10 Left sternocleidomastoid muscle (reflected)
11 Sternocleidomastoid muscle (clavicular head) and lateral supraclavicular nerves
12 Middle **supraclavicular nerves**
13 Medial supraclavicular nerves
14 Mandible
15 **Hyoid bone**
16 Superficial cervical lymph nodes
17 Left superior thyroid artery and external carotid artery
18 **Thyroid cartilage**
19 **Omohyoid muscle** (superior belly)
20 **Internal jugular vein** and branches of ansa cervicalis
21 Thyroid gland and unpaired inferior thyroid vein
22 Posterior belly of digastric muscle
23 **Stylohyoid muscle**
24 **Sternohyoid muscle**
25 **Sternothyroid muscle**

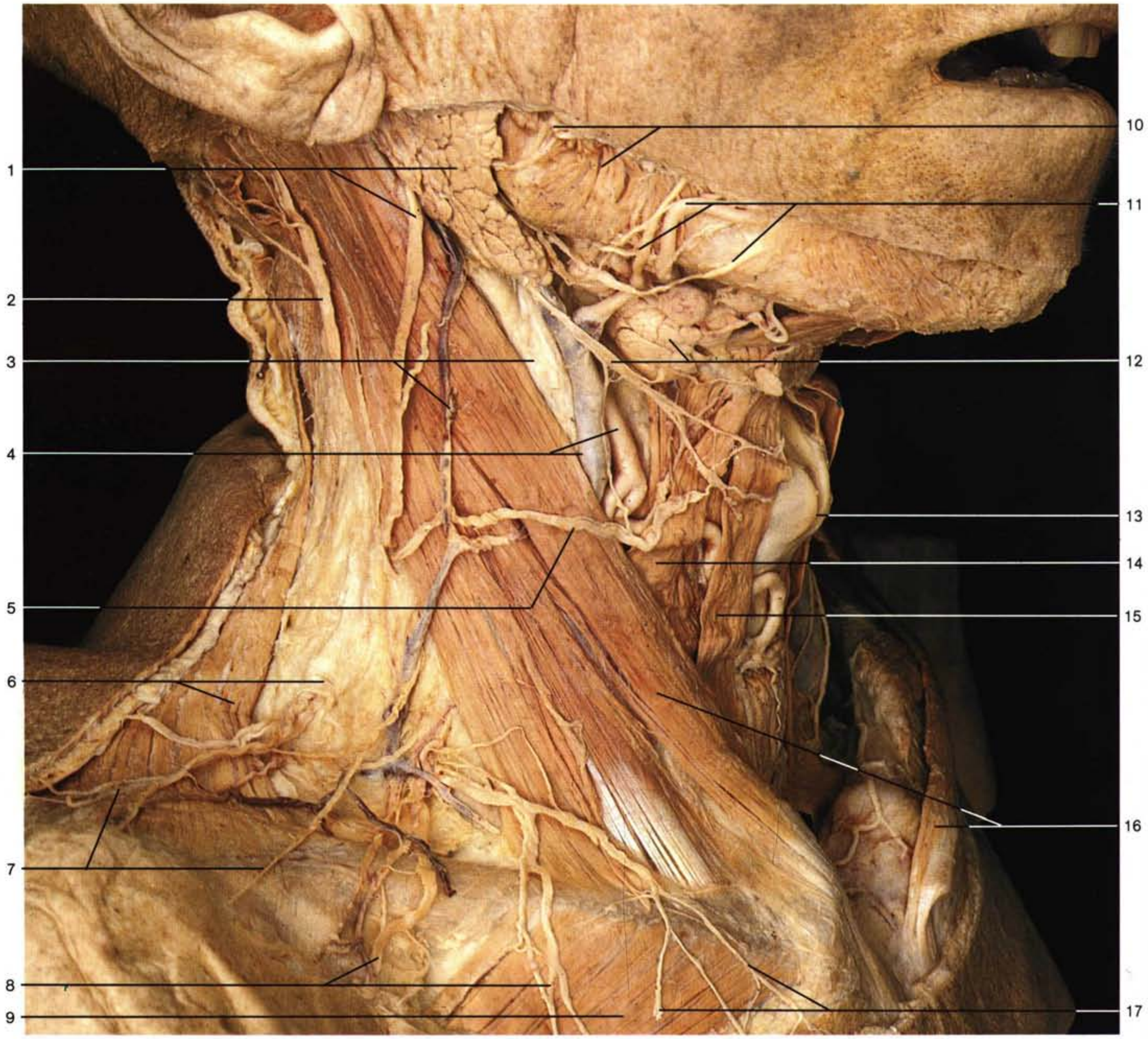

Posterior and carotid triangles (lateral aspect).
Superficial dissection.

1 **Parotid gland** and **great auricular nerve**
2 **Lesser occipital nerve**
3 Internal and external jugular veins
4 Retromandibular vein and external carotid artery
5 **Transverse cervical nerve** with communicating branch to cervical branch of facial nerve
6 Trapezius muscle and superficial lamina of cervical fascia
7 Lateral supraclavicular nerves
8 Middle **supraclavicular nerves**
9 Pectoralis major muscle
10 Buccal branch of facial nerve and masseter
11 **Facial artery and vein** and mandibular branch of facial nerve
12 Cervical branch of facial nerve and submandibular gland
13 Thyroid cartilage
14 Omohyoid muscle
15 Sternohyoid muscle
16 Sternocleidomastoid muscle
17 Medial supraclavicular nerves
18 Mandibular branch of facial nerve
19 Cervical branch of facial nerve with communicating branch to transverse cervical nerve

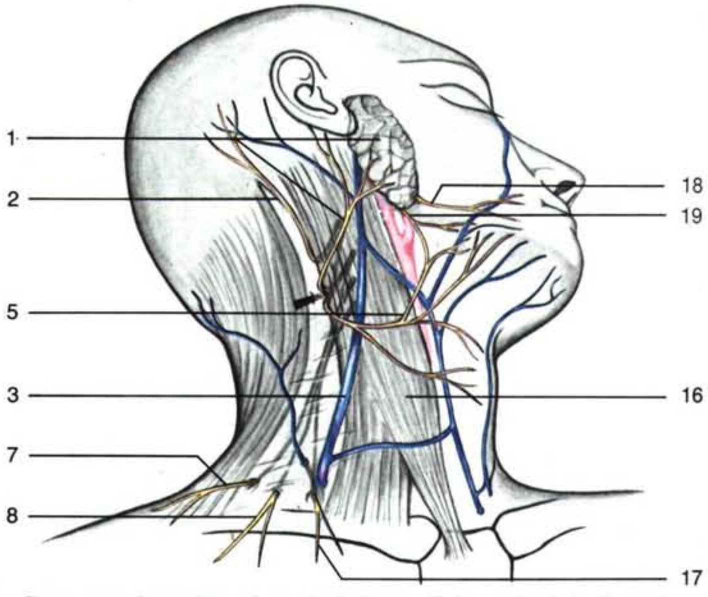

Cutaneous branches of cervical plexus. Erb's point is indicated by an arrowhead. (Schematic diagram.)

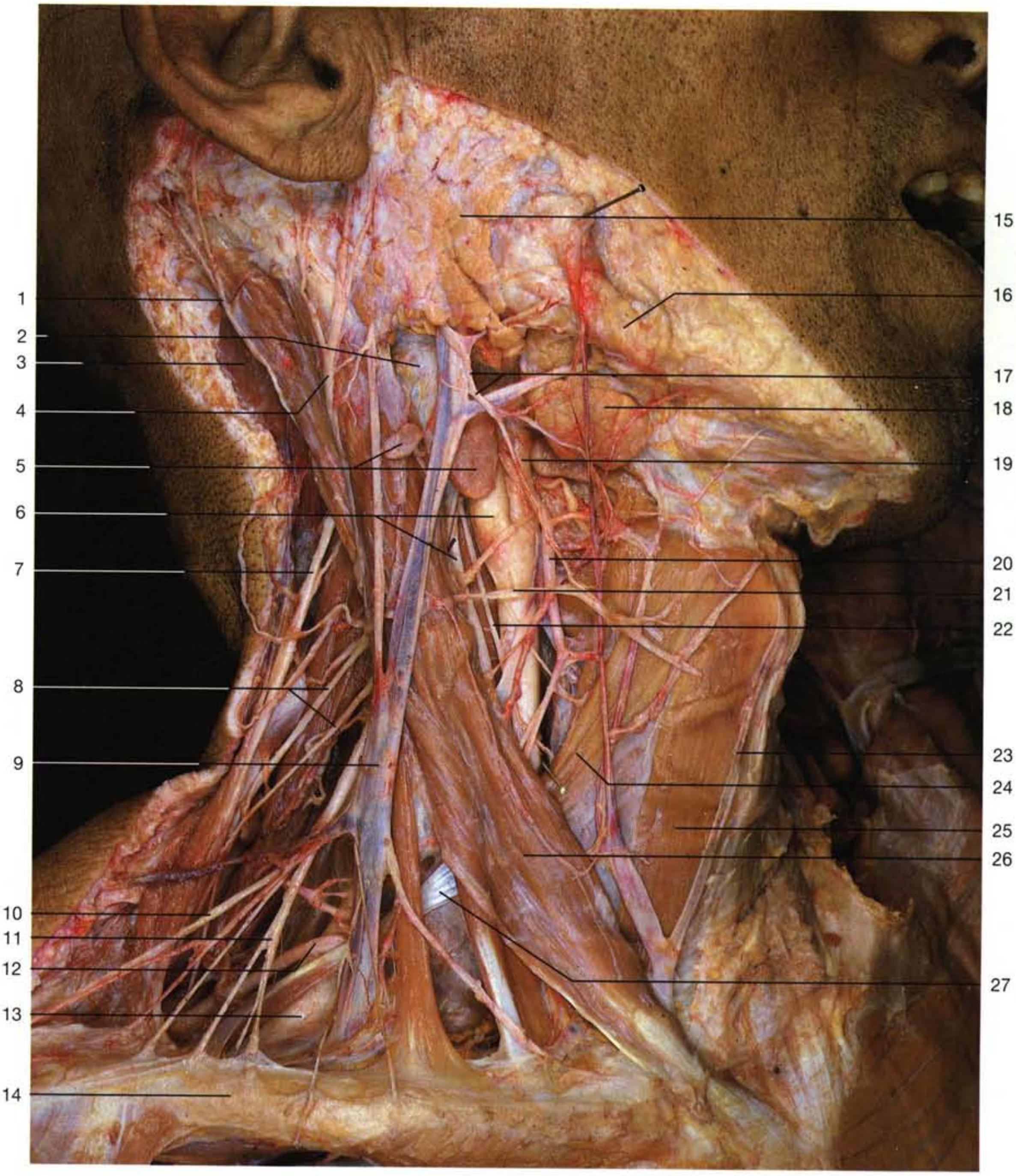

Posterior and carotid triangles (lateral aspect). Superficial dissection. The superficial lamina of cervical fascia has been removed to display the cutaneous branches of the cervical plexus and subcutaneous veins.

1	**Lesser occipital nerve**	11	**Middle supraclavicular nerves**	21	Transverse cervical nerve
2	**Internal jugular vein**	12	Suprascapular artery	22	Superior root of ansa cervicalis
3	Splenius capitis muscle	13	Pretracheal lamina of fascia of neck	23	Anterior jugular vein
4	**Great auricular nerve**	14	Clavicle	24	Omohyoid muscle
5	Submandibular nodes	15	Parotid gland	25	Sternohyoid muscle
6	Internal carotid artery and vagus nerve	16	Mandible	26	Sternocleidomastoid muscle
7	**Accessory nerve**	17	Cervical branch of facial nerve	27	Intermediate tendon of omohyoid muscle
8	Muscular branches of cervical plexus	18	Submandibular gland		
9	External jugular vein	19	External carotid artery		
10	**Posterior supraclavicular nerves**	20	Superior thyroid artery		

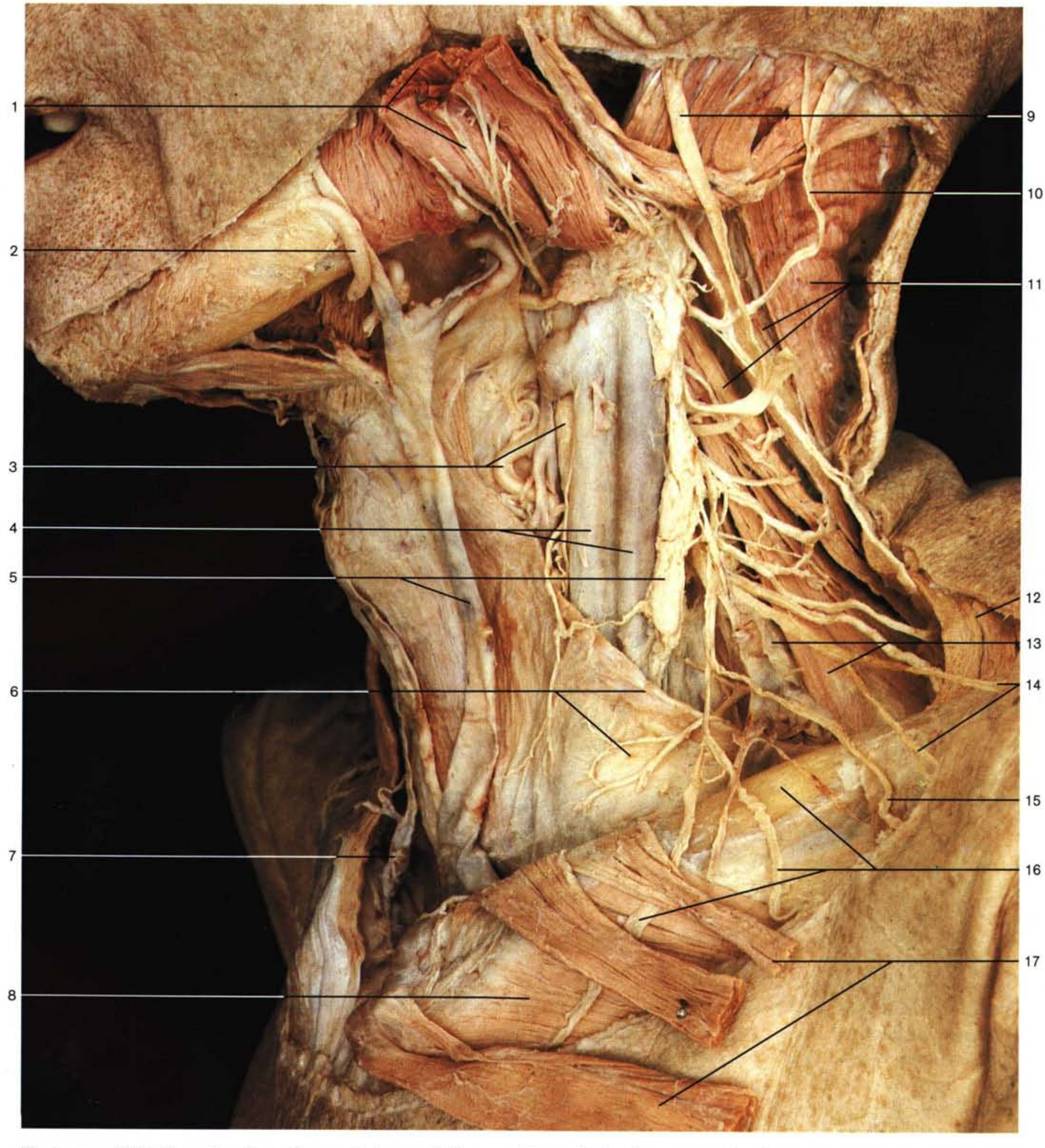

Neck, superficial dissection (lateral aspect). Sternocleidomastoid muscle has been cut and reflected to display the pretracheal lamina of the cervical fascia.

1 Sternocleidomastoid muscle (reflected) and branch of accessory nerve
2 Facial artery
3 External carotid artery and superior thyroid artery
4 **Internal jugular vein**
5 **Deep cervical lymph nodes** and external jugular vein
6 Omohyoid muscle and **pretracheal lamina of cervical fascia**
7 Anterior jugular vein
8 Pectoralis major muscle
9 **Great auricular nerve**

10 **Lesser occipital nerve**
11 Splenius capitis and levator scapulae muscles
12 Trapezius muscle
13 Scalenus medius muscle and brachial plexus
14 Posterior **supraclavicular nerves**
15 Middle supraclavicular nerve
16 Clavicle and anterior supraclavicular nerves
17 Sternocleidomastoid muscle (reflected)

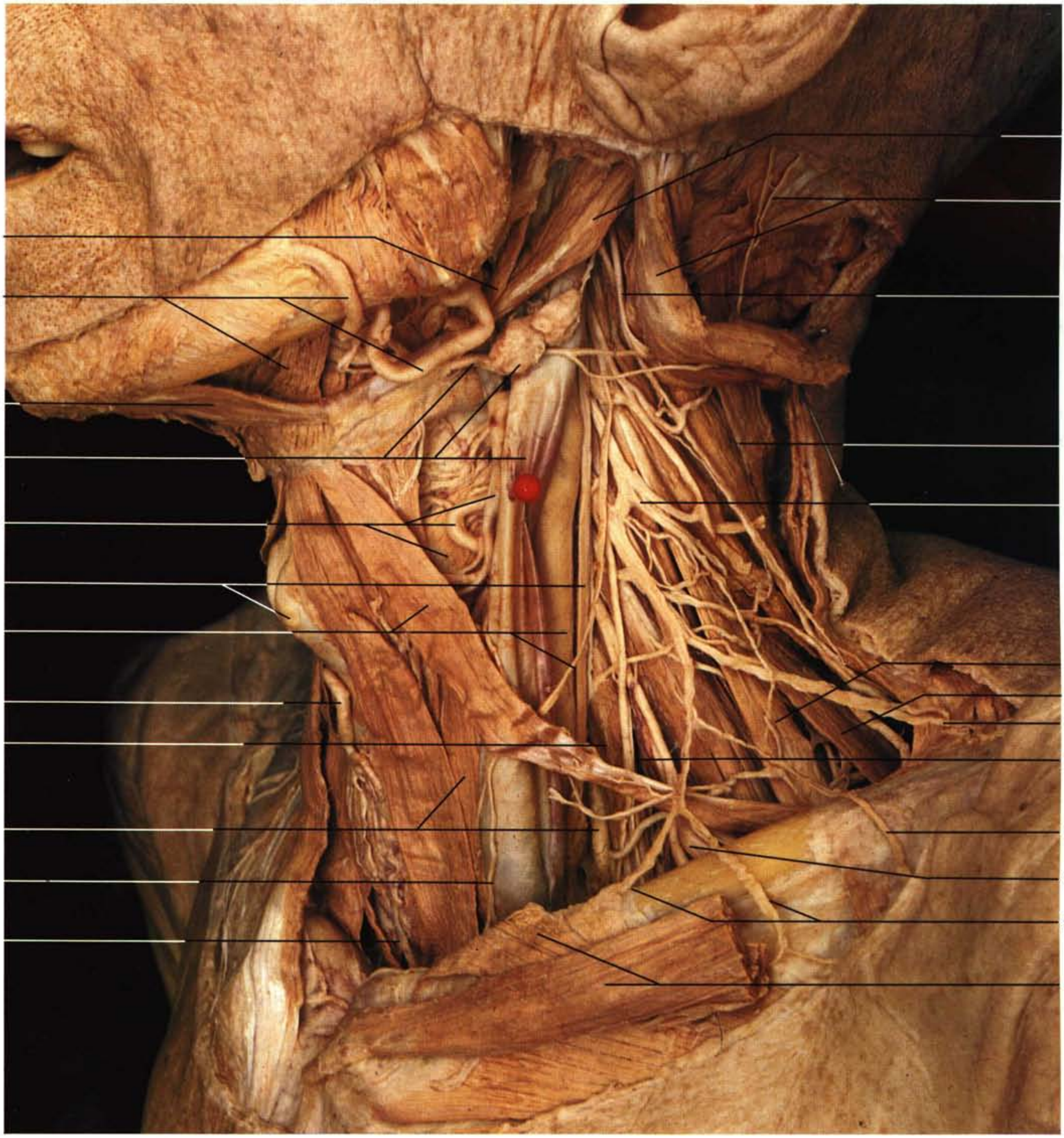

Neck, deep dissection (lateral aspect). The internal jugular vein has been reflected to expose the carotid artery and vagus nerve.

1 Stylohyoid muscle
2 Facial artery and mylohyoid muscle
3 Anterior belly of digastric muscle
4 **Internal jugular vein,** hypoglossal nerve, and superficial cervical lymph nodes
5 **Superior thyroid artery and vein** and inferior pharyngeal constrictor muscle
6 Thyroid cartilage and vagus nerve
7 **Ansa cervicalis,** omohyoid muscle, and common carotid artery
8 Right superior thyroid artery
9 Scalenus anterior muscle
10 Sternothyroid muscle and inferior thyroid artery
11 Muscular branches of ansa cervicalis to the infrahyoid muscles
12 Inferior thyroid vein

13 Posterior belly of digastric muscle
14 Sternocleidomastoid muscle and lesser occipital nerve
15 **Accessory nerve**
16 Splenius capitis muscle
17 **Cervical plexus**
18 Scalenus posterior muscle
19 Levator scapulac muscle
20 Posterior **supraclavicular nerves**
21 Phrenic nerve
22 Middle supraclavicular nerve
23 **Brachial plexus**
24 Anterior supraclavicular nerves
25 Sternocleidomastoid muscle

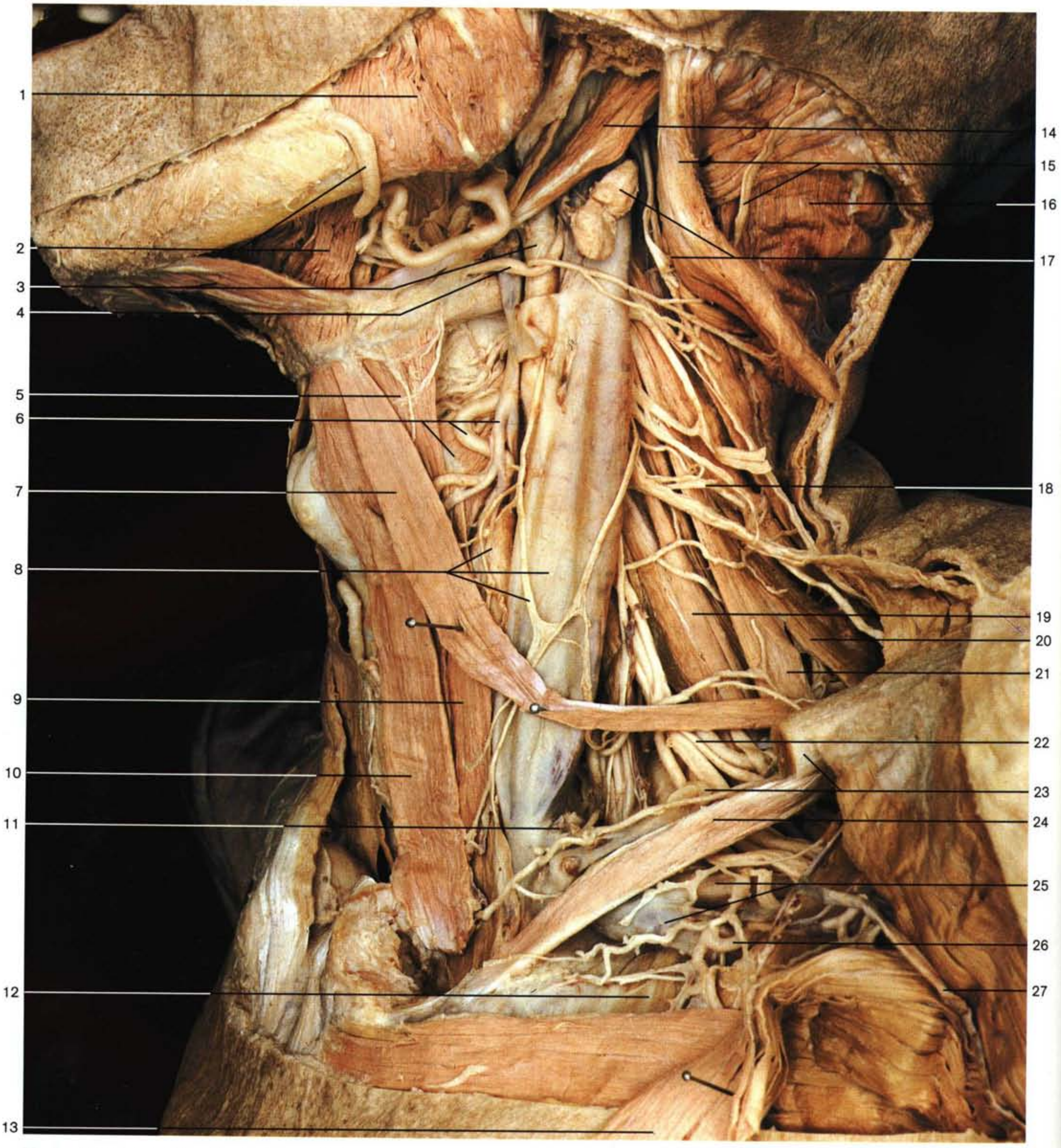

Neck, deeper dissection (lateral aspect). **Ansa cervicalis.** The cervical fascia and the clavicle are partly removed. Ansa cervicalis and infrahyoid muscles are displayed.

1 Masseter muscle
2 Mylohyoid muscle and facial artery
3 **External carotid artery** and anterior belly of digastric muscle
4 **Hypoglossal nerve**
5 Thyrohyoid muscle
6 Superior thyroid artery and vein and inferior pharyngeal constrictor muscle
7 Omohyoid muscle (superior belly)
8 **Ansa cervicalis,** thyroid gland and **internal jugular vein**
9 Sternothyroid muscle
10 Sternohyoid muscle
11 **Thoracic duct**
12 Pectoralis minor muscle
13 Pectoralis major muscle
14 Posterior belly of digastric muscle
15 Sternocleidomastoid muscle and lesser occipital nerve
16 Splenius capitis muscle
17 Superficial cervical lymph nodes and accessory nerve
18 **Cervical plexus**
19 Scalenus medius muscle
20 Levator scapulae muscle
21 Scalenus posterior muscle
22 Brachial plexus
23 Transverse cervical artery and clavicle
24 Subclavius muscle
25 **Subclavian artery and vein**
26 Thoracoacromial artery
27 Cephalic vein

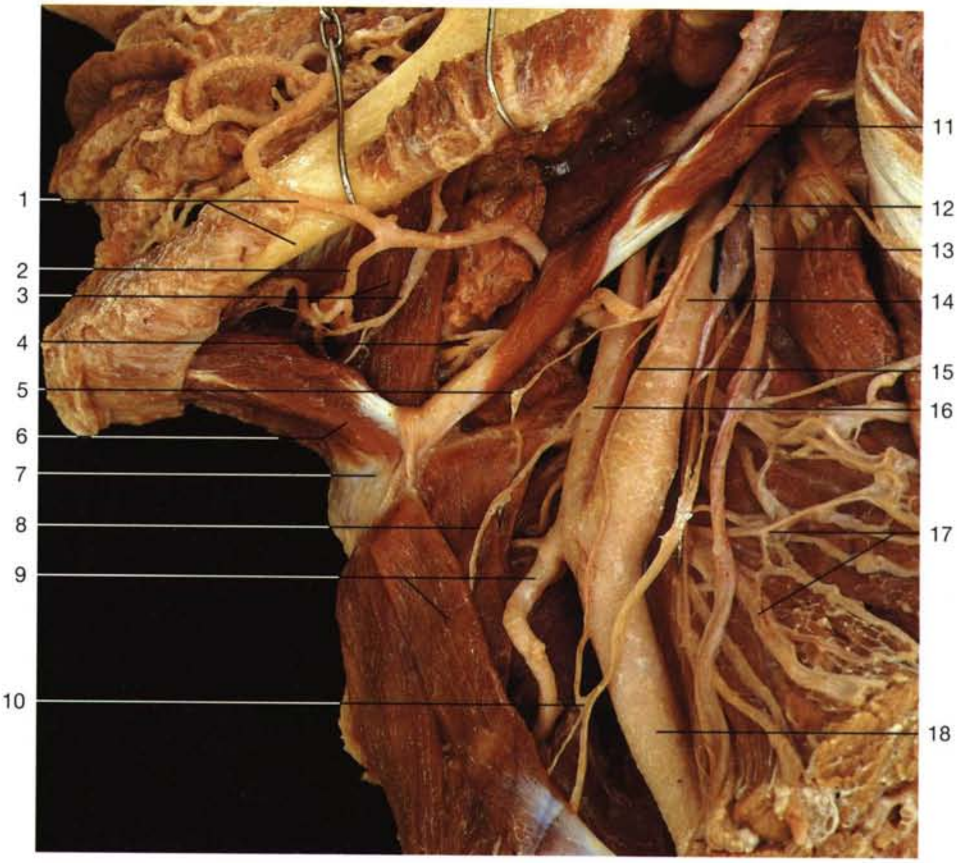

1 **Facial artery** and mandible
2 Submental artery
3 Mylohyoid muscle and nerve
4 **Hypoglossal nerve** (lingual branches)
5 Thyrohyoid branch of hypoglossal nerve (n. XII)
6 Anterior belly of digastric muscle
7 Hyoid bone
8 Omohyoid branch of hypoglossal nerve (n. XII)
9 Omohyoid muscle and superior thyroid artery
10 **Ansa cervicalis**
11 Posterior belly of digastric muscle
12 **Hypoglossal nerve** (n. XII)
13 **Vagus nerve** (n. X)
14 Internal carotid artery
15 Superior root of ansa cervicalis
16 **External carotid artery**
17 Cervical plexus
18 Common carotid artery

Neck, submandibular region (lateral aspect). **Hypoglossal nerve** (n. XII). Mandible slightly elevated.

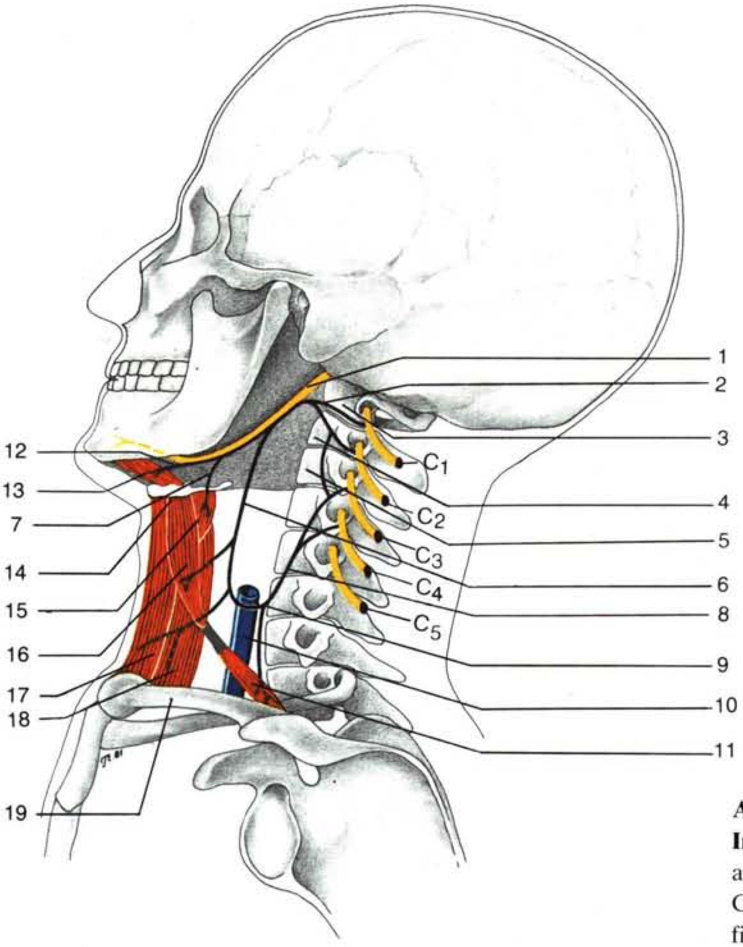

1 **Hypoglossal nerve** (n. XII)
2 Communication from the ventral ramus of the first cervical spinal nerve
3 Atlas
4 Axis
5 Third cervical vertebra
6 Superior root of ansa cervicalis
7 Thyrohyoid branch of hypoglossal nerve
8 Inferior root of ansa cervicalis
9 **Ansa cervicalis**
10 Internal jugular vein
11 Inferior belly of omohyoid muscle
12 Geniohyoid branch of hypoglossal nerve
13 Geniohyoid muscle
14 Hyoid bone
15 Thyrohyoid muscle
16 Superior belly of omohyoid muscle
17 Sternohyoid muscle
18 Sternothyroid muscle
19 Clavicle

Ansa cervicalis.
Innervation of infrahyoid muscles. Cervical plexus and its communication with the hypoglossal nerve. C_1–C_4 = ventral rami of cervical spinal nerves of the first four segments.

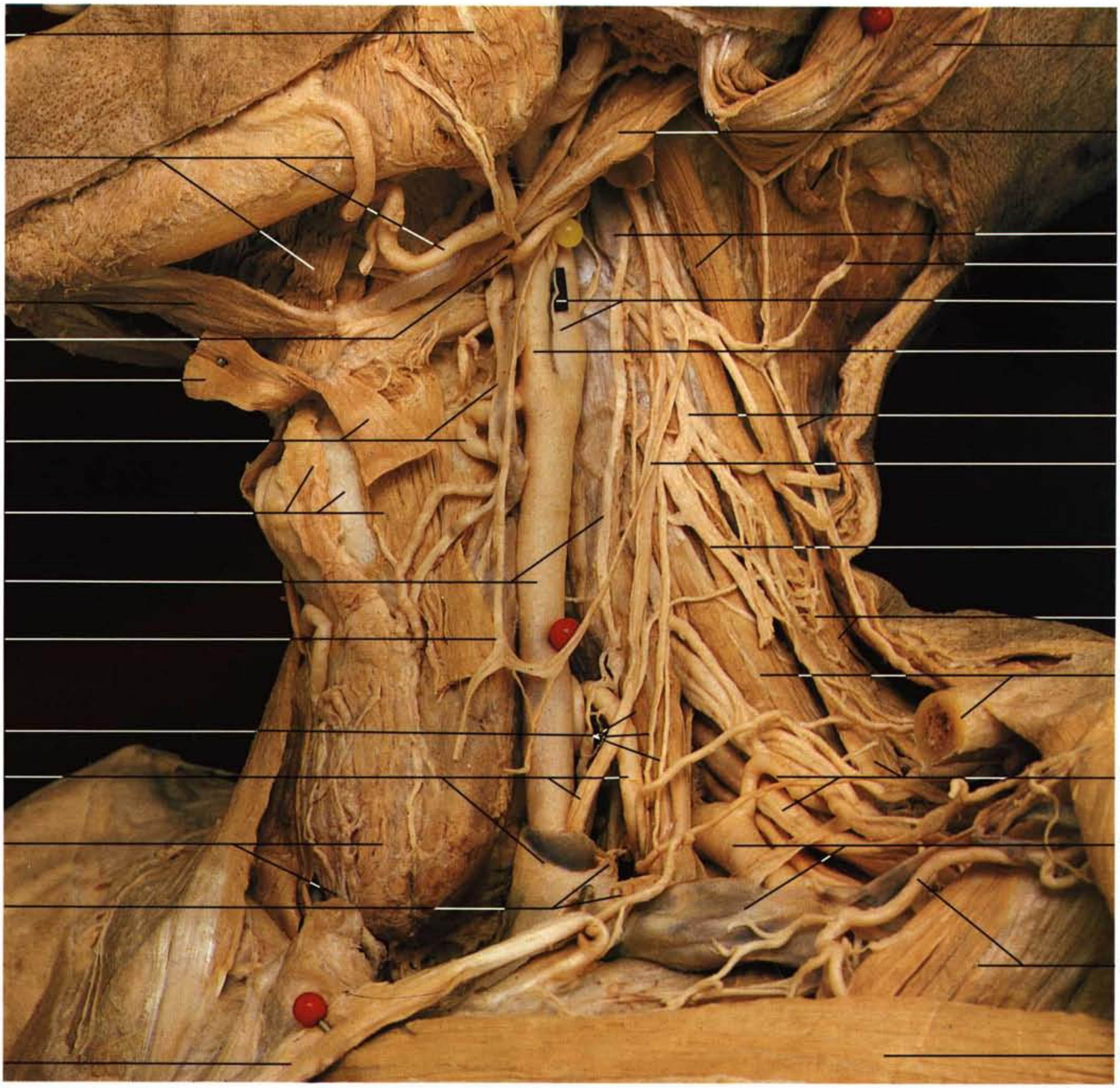

Neck, deep dissection (lateral aspect). Clavicle partly removed to show the slit between the scalenus muscles. Internal jugular vein removed.

1 Masseter muscle
2 Mylohyoid muscle and **facial artery**
3 Anterior belly of digastric muscle
4 **Hypoglossal nerve**
5 Sternohyoid muscle
6 Omohyoid muscle, superior thyroid artery and vein
7 Sternothyroid muscle, thyroid cartilage, and pyramidal lobe of thyroid gland
8 Common carotid artery and sympathetic trunk
9 **Ansa cervicalis**
10 **Phrenic nerve,** ascending cervical artery and anterior scalenus muscle
11 Inferior thyroid artery, vagus nerve and internal jugular vein (cut)
12 **Thyroid gland** and unpaired inferior thyroid venous plexus
13 Thoracic duct and left subclavian trunk
14 Subclavius muscle (reflected)
15 Sternocleidomastoid muscle (reflected)
16 Posterior belly of digastric muscle
17 **Superior cervical ganglion** and splenius muscle
18 **Lesser occipital nerve**
19 Internal carotid artery and branch of the glossopharyngeal nerve to the carotid body
20 **External carotid artery**
21 **Cervical plexus** and **accessory nerve**
22 Inferior root of ansa cervicalis
23 Supraclavicular nerve
24 Levator scapulae muscle
25 Scalenus medius muscle and clavicle
26 Transverse cervical artery, brachial plexus, and scalenus posterior muscle
27 **Subclavian artery and vein**
28 Thoracoacromial artery and pectoralis minor muscle
29 Pectoralis major muscle

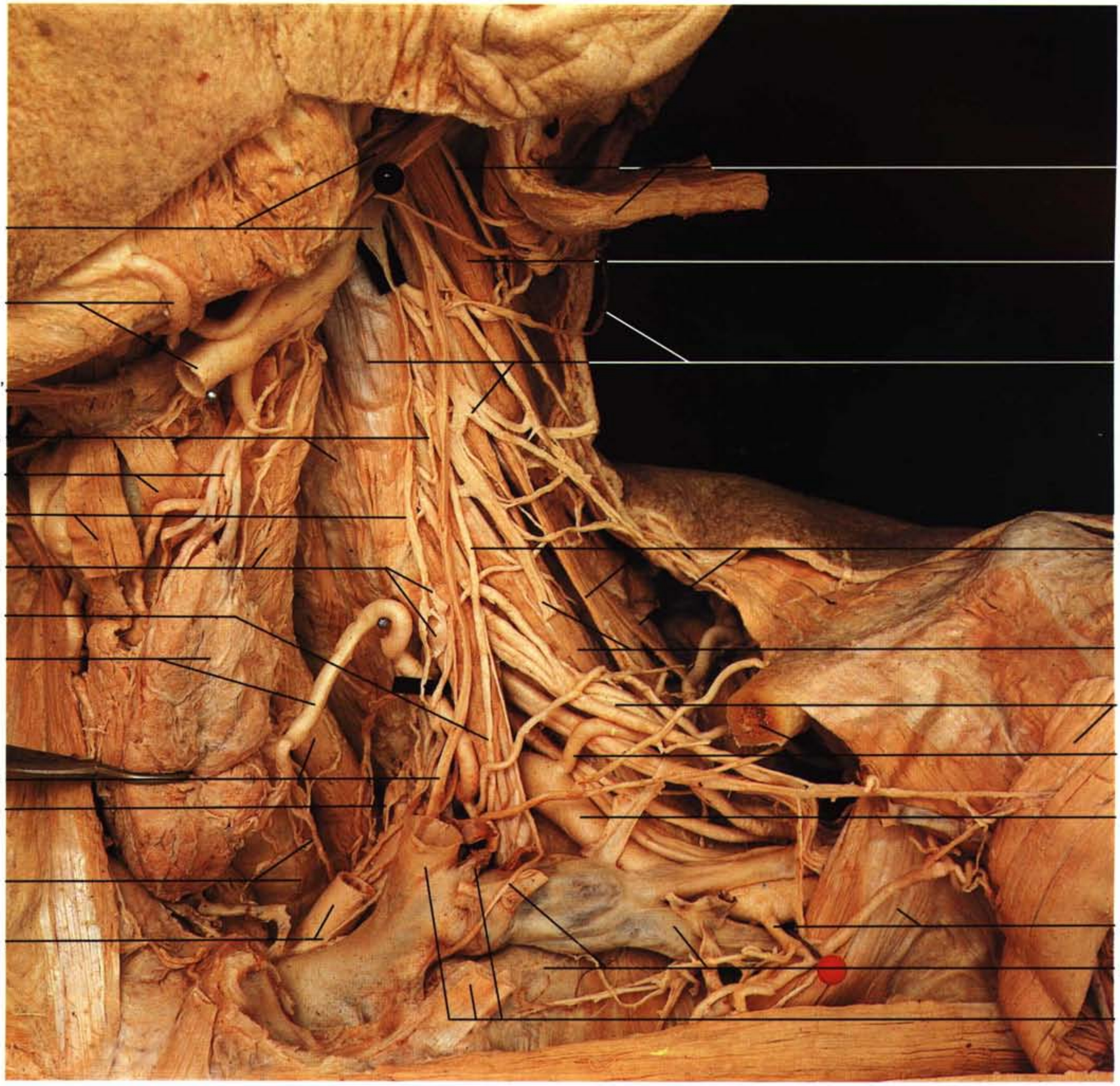

Neck, deepest dissection (anterolateral aspect). Thyroid gland reflected to expose the esophagus and the recurrent laryngeal nerve.

1 **Superior cervical ganglion** of sympathetic trunk and posterior belly of digastric muscle
1' Anterior belly of digastric muscle
2 **Facial artery** and common carotid artery (reflected anteriorly)
3 Ascending cervical artery and longus colli muscle
4 Omohyoid muscle and superior thyroid artery
5 **Sympathetic trunk** and sternohyoid muscle
6 Middle cervical ganglion and inferior pharyngeal constrictor muscle
7 Scalenus anterior muscle and phrenic nerve
8 **Thyroid gland** and **inferior thyroid artery**
9 Vagus nerve and esophagus
10 **Stellate ganglion**
11 **Recurrent laryngeal nerve** and trachea
12 Common carotid artery and cervical cardiac branch of vagus nerve
13 Sternocleidomastoid muscle and accessory nerve
14 Splenius capitis muscle
15 Lesser occipital nerve, longus capitis muscle and cervical plexus
16 **Phrenic nerve,** scalenus posterior muscle and levator scapulae muscle
17 Supraclavicular nerves and scalenus medius muscle
18 Brachial plexus and pectoralis major muscle (clavicular head)
19 Transverse cervical artery and clavicle
20 **Subclavian artery**
21 **Thoracoacromial artery** and pectoralis minor muscle
22 First rib, accessory phrenic nerve and subclavian vein
23 Internal jugular vein, thoracic duct and subclavius muscle

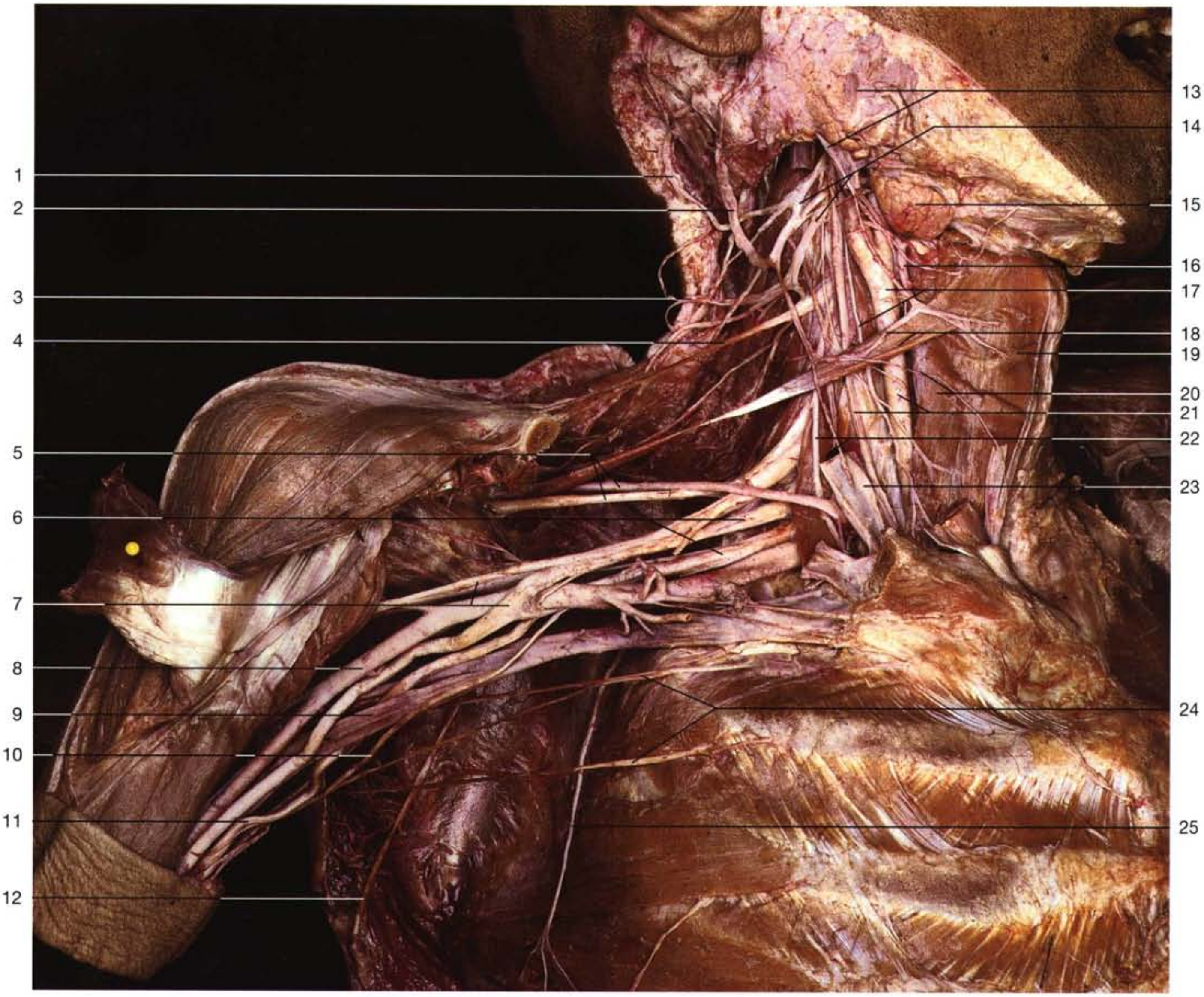

Neck and arm, deepest dissection (anterior-lateral aspect). Cervical and brachial plexus and their relation to the blood vessels are shown. Note the location and content of scalene triangle. Sternocleidomastoid muscle and clavicle have been removed; the internal jugular vein was divided to display the roots of cervical and brachial plexus.

1 Lesser occipital nerve
2 Great auricular nerve
3 Cutaneous branches of cervical plexus
4 Supraclavicular nerve
5 Suprascapular nerve and artery
6 **Brachial plexus**
7 **Median nerve** (with two roots) and musculocutaneous nerve
8 Axillary artery
9 Axillary vein
10 Medial brachial cutaneous nerve
11 **Ulnar nerve**
12 Thoracodorsal nerve
13 Parotid gland and facial nerve (cervical branch)
14 **Cervical plexus**
15 Submandibular gland
16 Superior thyroid artery
17 Common carotid artery dividing in internal and external carotid artery and superior root of ansa cervicalis
18 Omohyoid muscle and cervical branch of facial nerve joining the transverse cervical nerve (C_2, C_3)
19 Sternohyoid muscle
20 Transverse cervical nerve and sternothyroid muscle
21 **Common carotid artery** and **vagus nerve**
22 **Phrenic nerve** and scalenus anterior muscle
23 Internal jugular vein
24 Intercostobrachial nerves
25 Long thoracic nerve

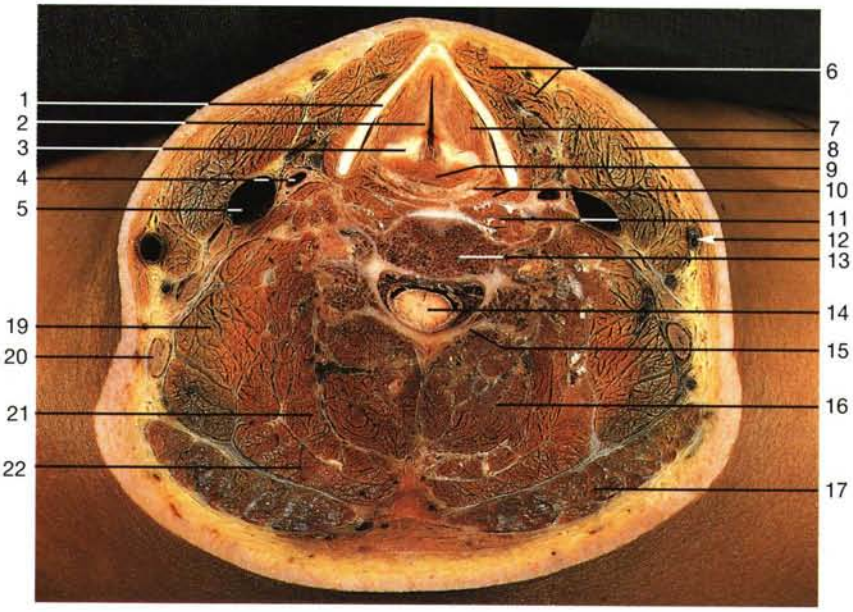

◁ **Horizontal section through the neck** at the level of the fissure of glottis, viewed from above.

1 **Thyroid cartilage**
2 **Vocal fold** and **glottis** (rima glottidis)
3 Arytenoid cartilage
4 Common carotid artery
5 **Internal jugular vein**
6 Infrahyoid muscles
7 Lateral thyroarytenoid muscle
8 Sternocleidomastoid muscle
9 Transverse arytenoid muscle
10 Laryngopharynx and inferior constricor muscle of pharynx
11 Longus colli muscle
12 External jugular vein
13 Body of cervical vertebra (C_5)
14 **Spinal cord**
15 Vertebra arch
16 Deep muscles of neck (semispinalis cervicis muscle)
17 Trapezius muscle
18 Rima glottidis
19 Levator scapulae muscle
20 Lymph node
21 Semispinalis capitis muscle
22 Splenius capitis muscle

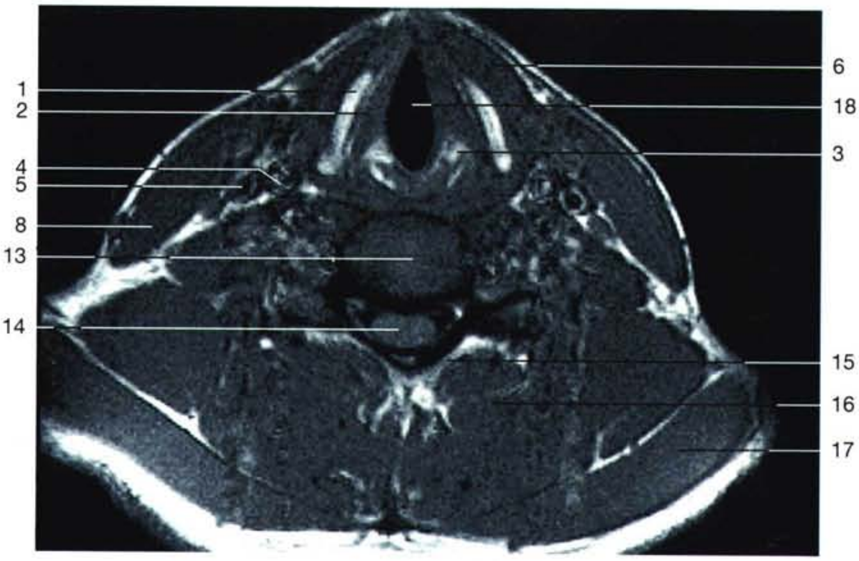

◁ **Section through the neck at the level of larynx.** (MR-Scan.)

1 Hyoid bone
2 Thyroid cartilage
3 **Cervical plexus** (C_1-C_4)
4 **Phrenic nerve** (C_4)
5 Sealenus anterior muscle
6 **Brachial plexus** (C_5-T_1)
7 Scalenus medius and posterior muscles
8 Subclavian artery
9 Subclavian vein
10 Superior vena cava
11 Cricoid cartilage
12 Thyroid gland
13 **Internal jugular vein**
14 **Common carotid artery**
15 Inferior thyroid vein
16 Ascending aorta
17 Descending aorta
18 Second rib

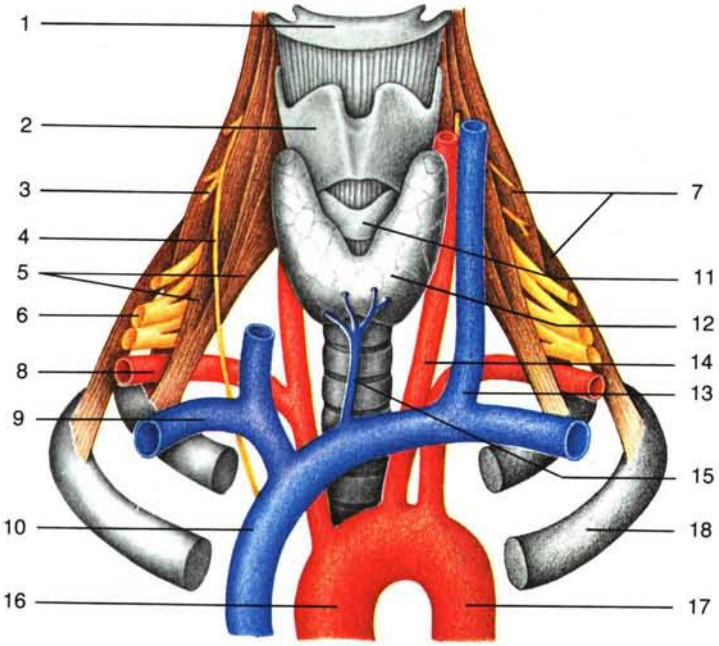

◁ **Scalene triangle, arrangements of blood vessels, and brachial plexus at the lower part of the neck.** (Schematic diagram.)

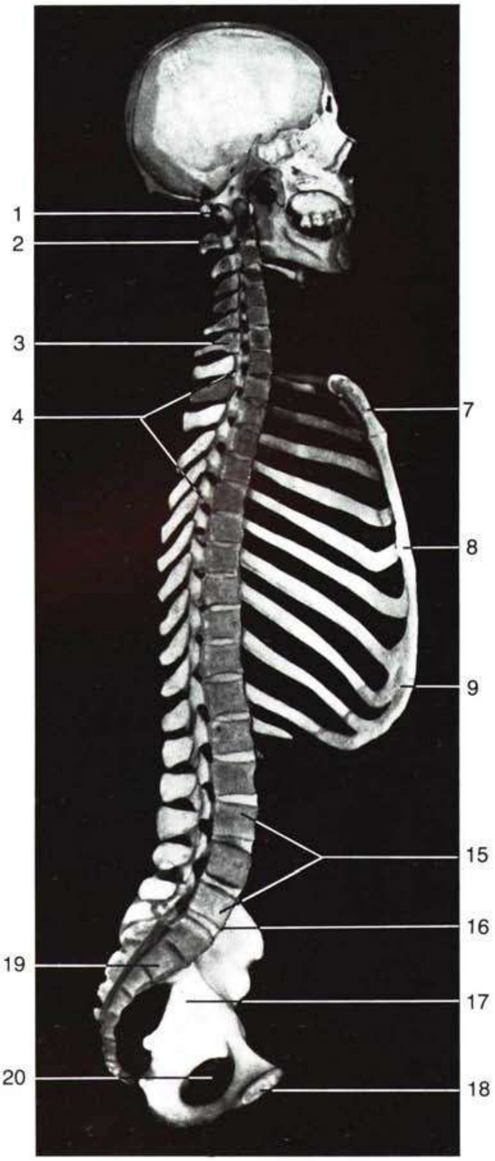

Median sagittal section through the vertebral column, head, and thorax of the adult.

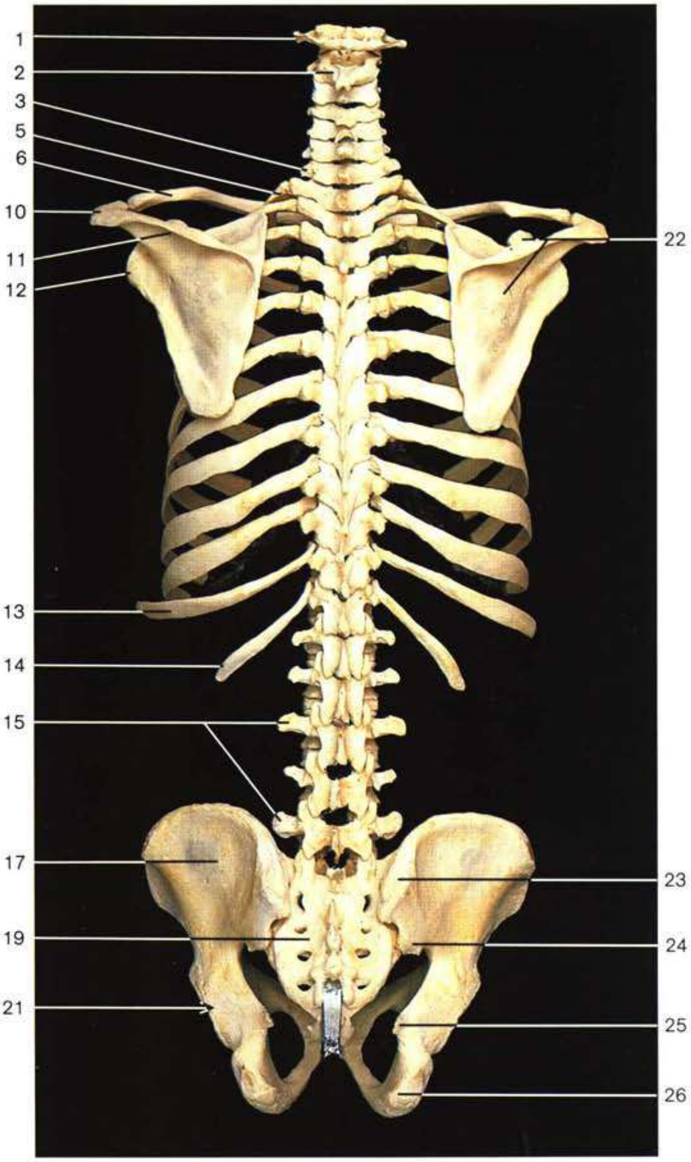

Skeleton of the trunk, vertebral column, thorax, and pelvis (posterior aspect).

1 Atlas
2 Axis
3 Seventh cervical vertebra (vertebra prominens)
4 Vertebral canal
5 First rib
6 Clavicle
7 Manubrium sterni
8 Body of sternum
9 Costal arch
10 Acromion
11 Spine of scapula
12 Glenoid cavity (lateral angle of scapula)
13 Eleventh rib
14 Twelfth rib
15 Lumbar vertebrae
16 Sacral promontory
17 Hip bone
18 Symphysis pubis
19 Sacrum
20 Obturator foramen
21 Acetabulum
22 Scapula with coracoid process
23 Posterior superior iliac spine
24 Posterior inferior iliac spine
25 Ischial spine
26 Ischial tuberosity

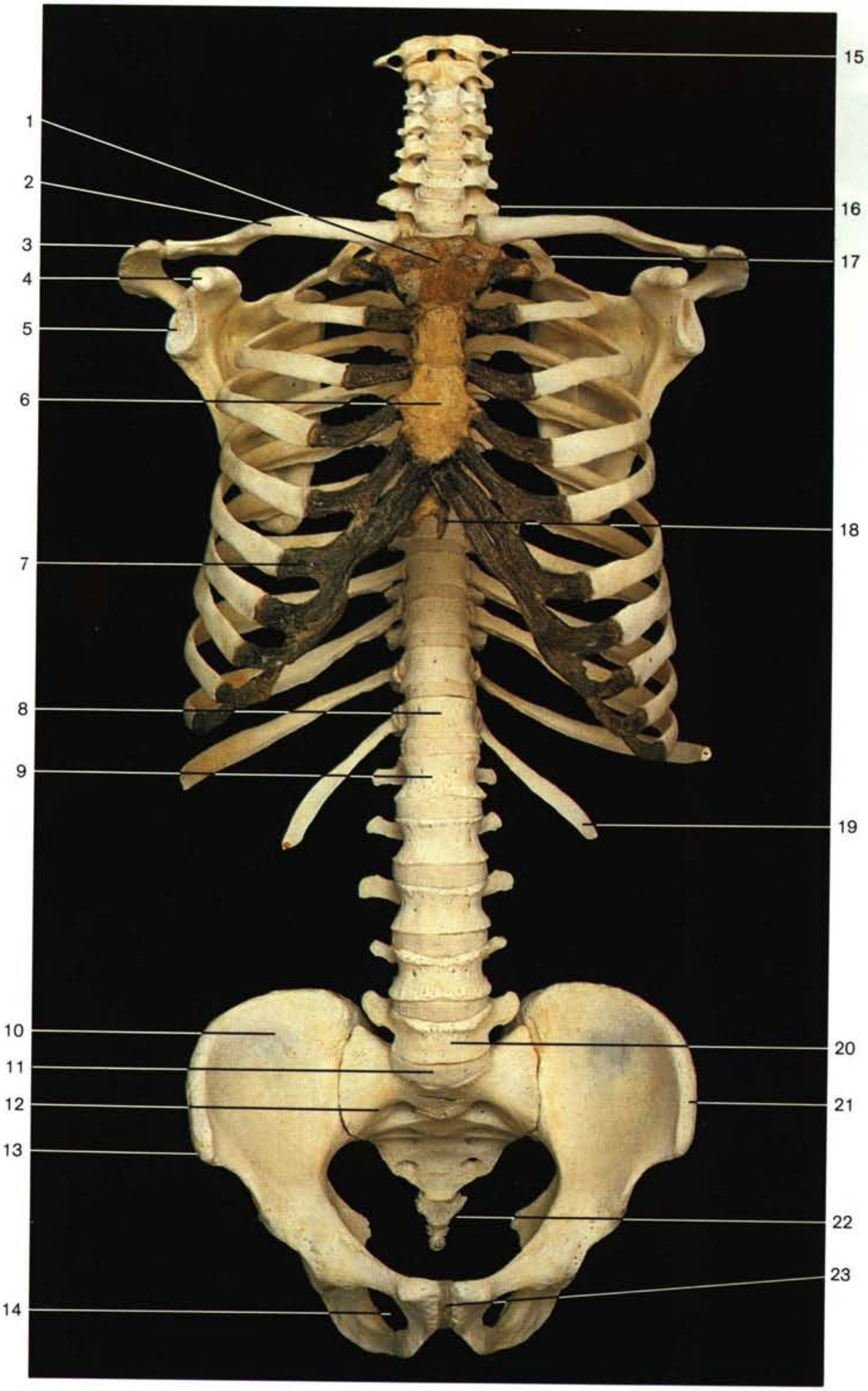

1	Manubrium sterni
2	**Clavicle**
3	Acromion
4	Coracoid process
5	Glenoid cavity
6	Body of **sternum**
7	Costal cartilage
8	Body of the twelfth **thoracic vertebra**
9	Body of the first **lumbar vertebra**
10	Hip bone
11	Sacral promontory
12	**Sacrum**
13	Anterior superior iliac spine
14	Obturator foramen
15	Atlas
16	Seventh **cervical vertebra**
17	First rib
18	Xiphoid process
19	Twelfth rib
20	Body of the fifth lumbar vertebra
21	Iliac crest
22	**Coccyx**
23	Symphysis pubis

Skeleton of the trunk, vertebral column, pelvis, thorax, and shoulder girdle (anterior aspect).

The trunk is divided into segments best visible in the thoracic region, where each segment consists of a pair of ribs connected anteriorly by the sternum and posteriorly by a thoracic vertebra. In the lumbar part of the vertebral column only vestiges of ribs are present which form what appear to be the transverse processes. In cervical vertebrae, remnants of ribs are part of the transverse processes. Each segment also comprises muscles (e.g. intercostal muscles), nerves, and vessels. However, in the cervical and lumbar region the muscular segments fuse with each other forming large muscle plates, for example, the oblique muscles of the abdomen, while vessels and nerves still retain their segmental pattern.

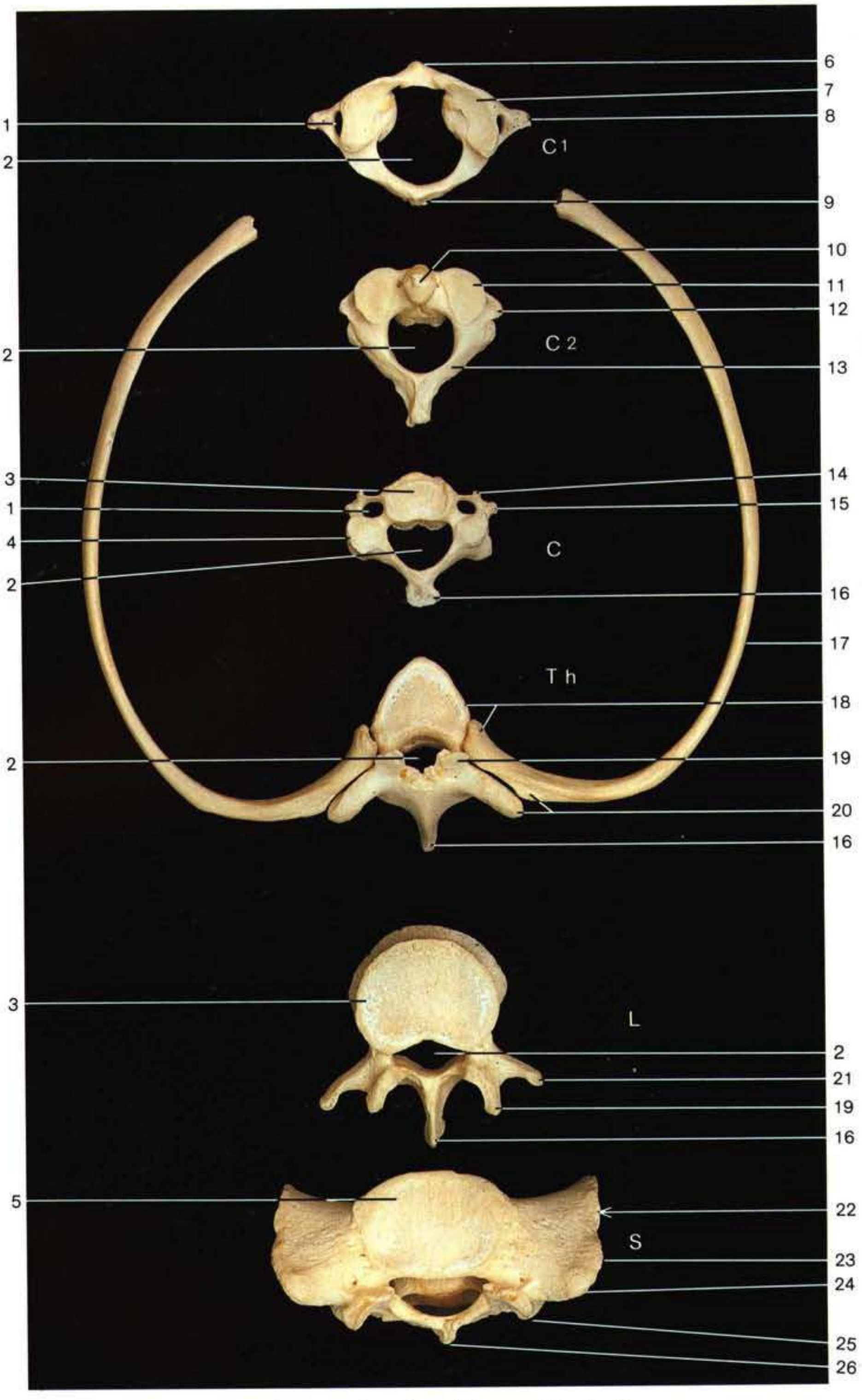
1
2
2
2
3
1
4
2
2
3
5
6
7
8
C1
9
10
11
12
C2
13
14
15
C
16
17
Th
18
19
20
16
L
2
21
19
16
S
22
23
24
25
26

Representative vertebrae from each region of the vertebral column (superior aspect). From top to bottom: atlas (C_1), axis (C_2), cervical vertebra (C), thoracic vertebra (Th), lumbar vertebra (L), and sacrum (S).

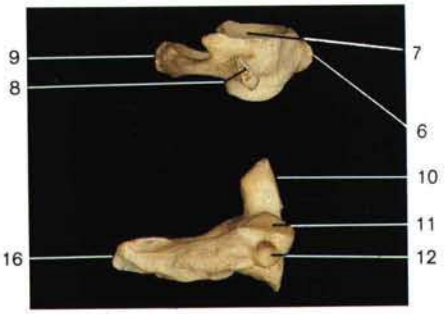
9
8
7
6
10
11
16
12

Atlas and axis

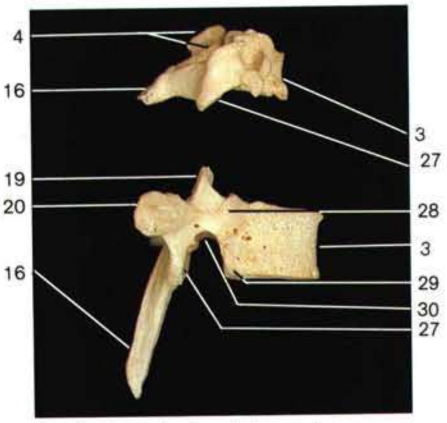
4
16
3
27
19
20
28
3
16
29
30
27

Typical cervical and thoracic vertebrae

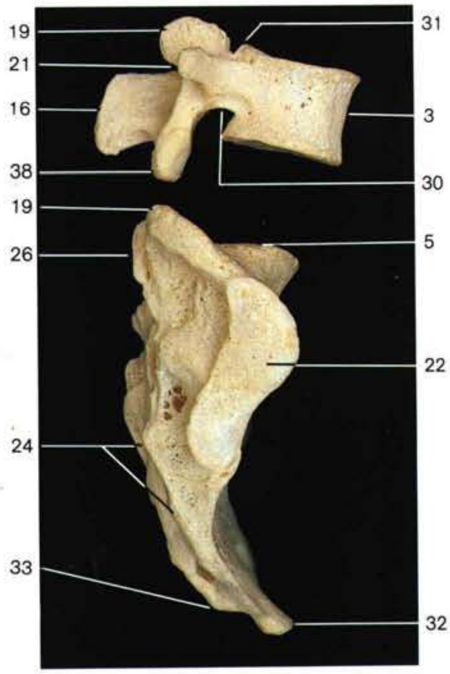
19
21
16
38
31
3
30
19
26
5
22
24
33
32

Typical lumbar vertebra and sacrum

Representative vertebrae from each region of the vertebral column (lateral aspect, ventral surface on the right).

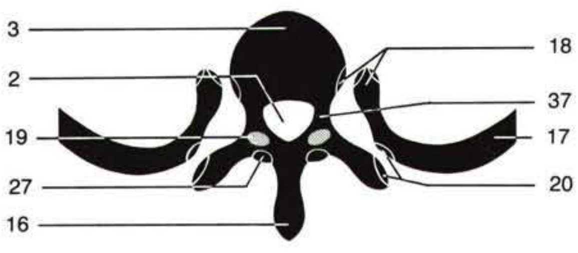
3
2
19
27
16
18
37
17
20

General organization of ribs and vertebrae. (Schematic diagram.)

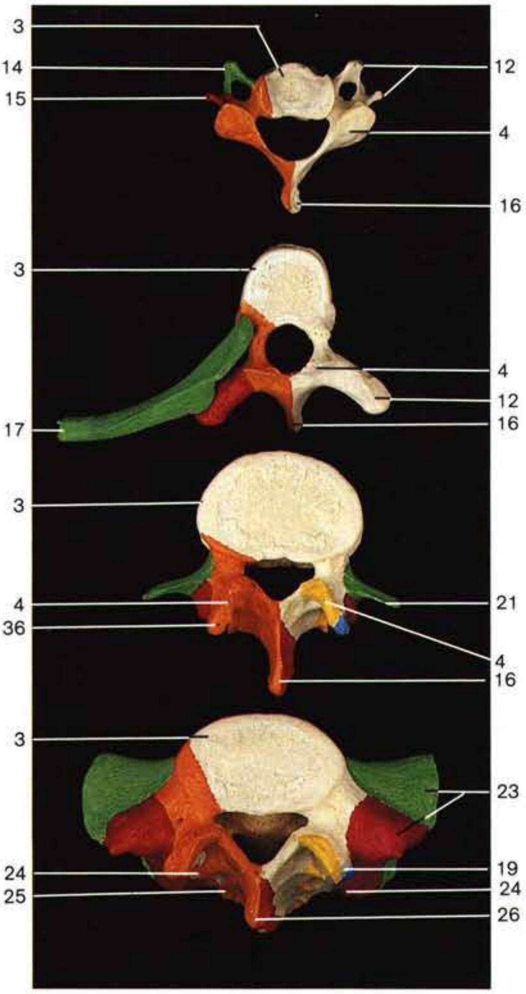

General characteristics of the vertebrae. Typical cervical, thoracic, lumbar vertebrae and sacrum.

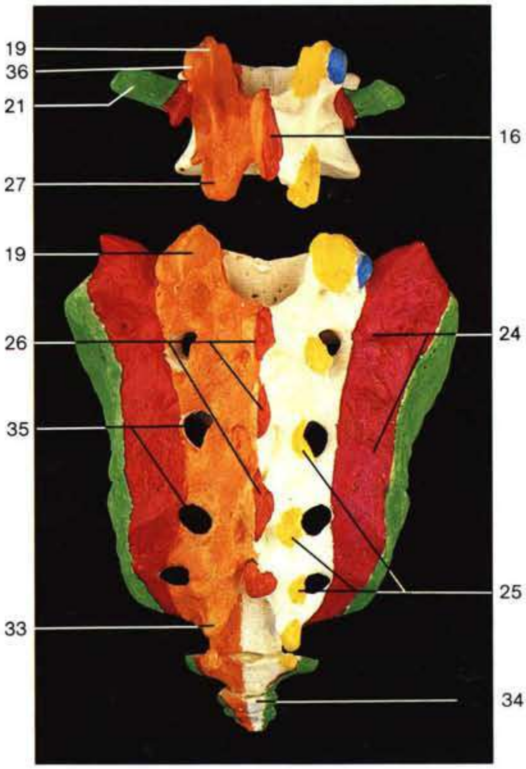

General characteristics of lumbar vertebrae and sacrum (posterior aspect).

1 Foramen transversarium
2 Vertebral foramen
3 Body of vertebra
4 Superior articular facet
5 Base of sacrum
6 Anterior tubercle of atlas
7 Superior articular facet of atlas
8 Transverse process
9 Posterior tubercle of atlas
10 Dens of axis
11 Superior articular surface
12 Transverse process
13 Arch of vertebra
14 Anterior tubercle of transverse process
15 Posterior tubercle of transverse process
16 Spinous process
17 Shaft of rib
18 Body of vertebra and head of rib articulating with each other (costovertebral joint)
19 Superior articular process
20 Transverse process and tubercle of rib articulating with each other (costotransverse joint)
21 Transverse process
22 Auricular surface
23 Lateral part of sacrum
24 Lateral sacral crest
25 Intermediate sacral crest
26 Median sacral crest
27 Inferior articular facet
28 Superior demifacet for head of rib
29 Inferior demifacet for head of rib
30 Inferior vertebral notch
31 Superior vertebral notch
32 Apex of the sacrum
33 Sacral cornu
34 Coccyx
35 Dorsal sacral foramina
36 Mamillary process
37 Pedicle
38 Inferior articular process

Green = Ribs or homologous processes
Red = Muscular processes (transverse and spinous processes)
Orange = Laminae and articular processes
Yellow = Articular facets

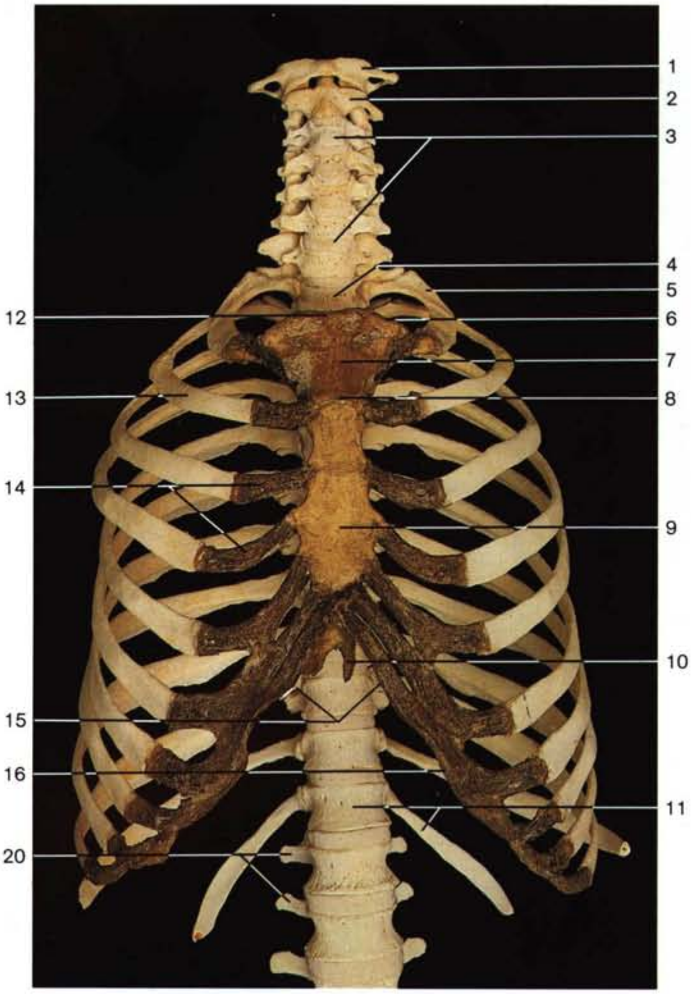

Skeleton of the thorax (anterior aspect).

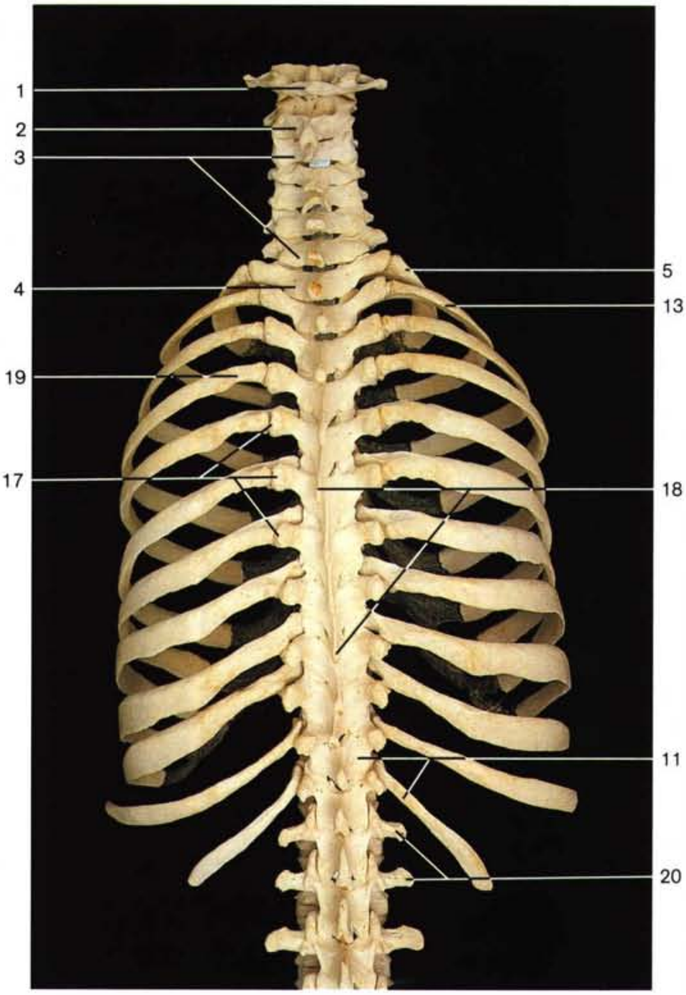

Skeleton of the thorax (posterior aspect).

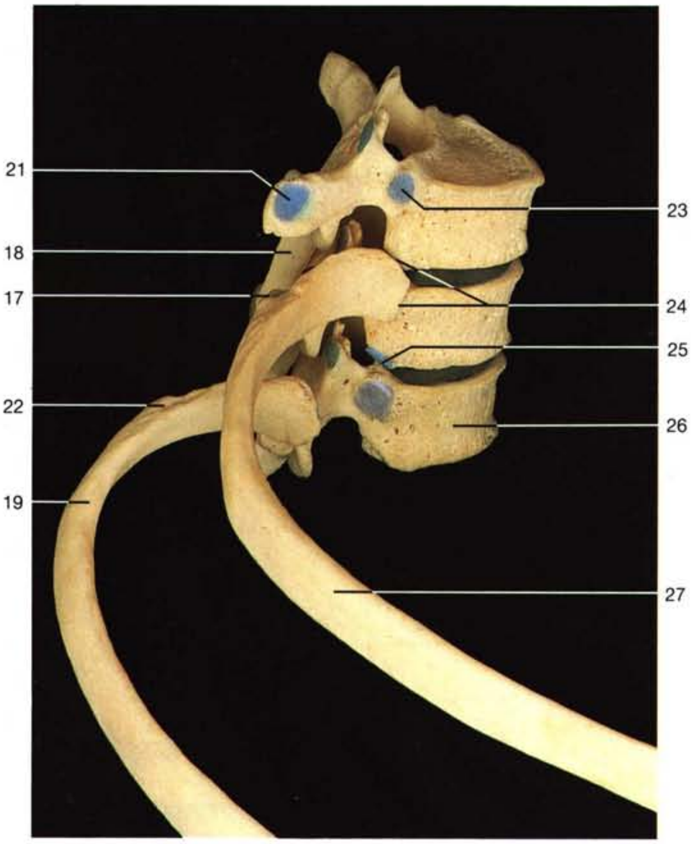

Costovertebral articulation (right lateral aspect).

1 Atlas
2 Axis
3 **Cervical vertebrae**
4 First thoracic vertebra
5 First rib
6 Facet for clavicle and clavicular notch
7 **Manubrium sterni**
8 Sternal angle
9 Body of sternum
10 Xiphoid process
11 Twelfth thoracic vertebra and rib
12 Jugular notch
13 Second rib
14 **Costal cartilages**
15 Infrasternal angle
16 **Costal arch**
17 Costotransverse joints between the transverse processes of thoracic vertebra and the tubercles of the ribs
18 Spinous processes
19 Costal angle
20 Transverse processes of lumbar vertebrae
21 Facet for articulation with rib
22 Tubercle of rib
23 Superior facet for articulation with head of rib
24 Articulation of head of rib with two vertebrae
25 Inferior facet for articulation with head of rib
26 Body of thoracic vertebra
27 Body or shaft of rib

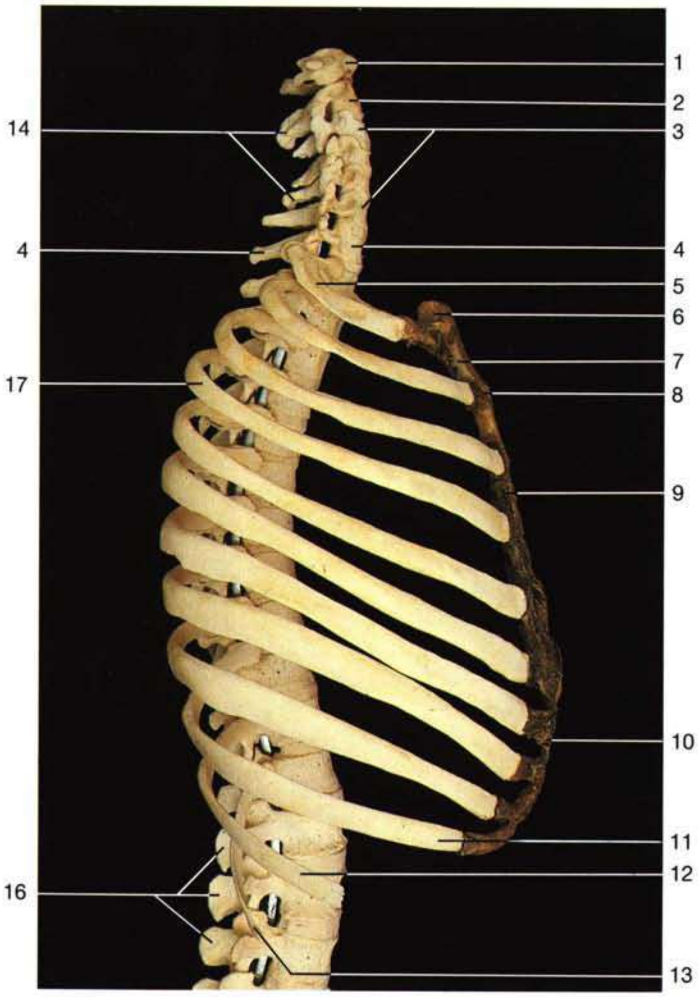
1
2
14
3
4
4
5
6
7
17
8
9
10
11
16
12
13

Skeleton of the thorax (right lateral aspect).

1 Atlas
2 Axis
3 **Cervical vertebrae**
4 Seventh cervical vertebra (vertebra prominens)
5 First rib
6 Facet for clavicle
7 **Manubrium sterni**
8 Sternal angle
9 Body of sternum
10 Costal arch
11 Tenth rib
12 Eleventh rib
13 Twelfth rib
14 Spinous processes of cervical vertebrae
15 Spinous processes of thoracic vertebrae
16 Spinous processes of lumbar vertebrae
17 Costal angle
18 Intervertebral foramina
19 Intervertebral discs
20 **Cervical curvature**
21 **Thoracic curvature**
22 **Lumbar curvature**
23 Sacrum
24 Coccyx

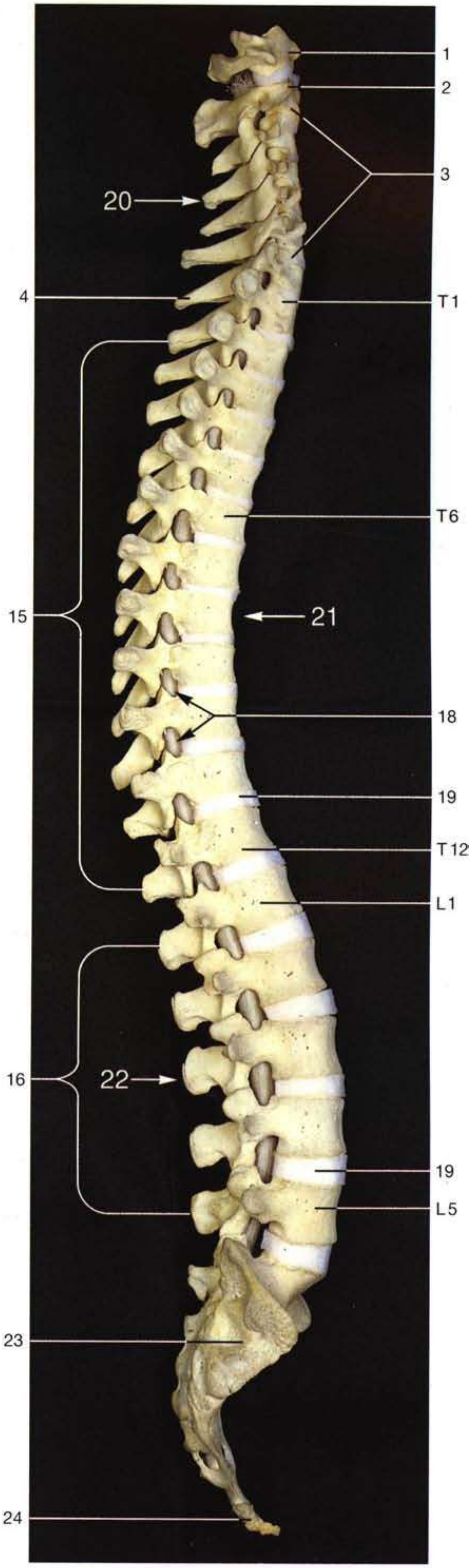
1
2
3
20
4
T1
T6
15
21
18
19
T12
L1
16
22
19
L5
23
24

Vertebral column (right lateral aspect).

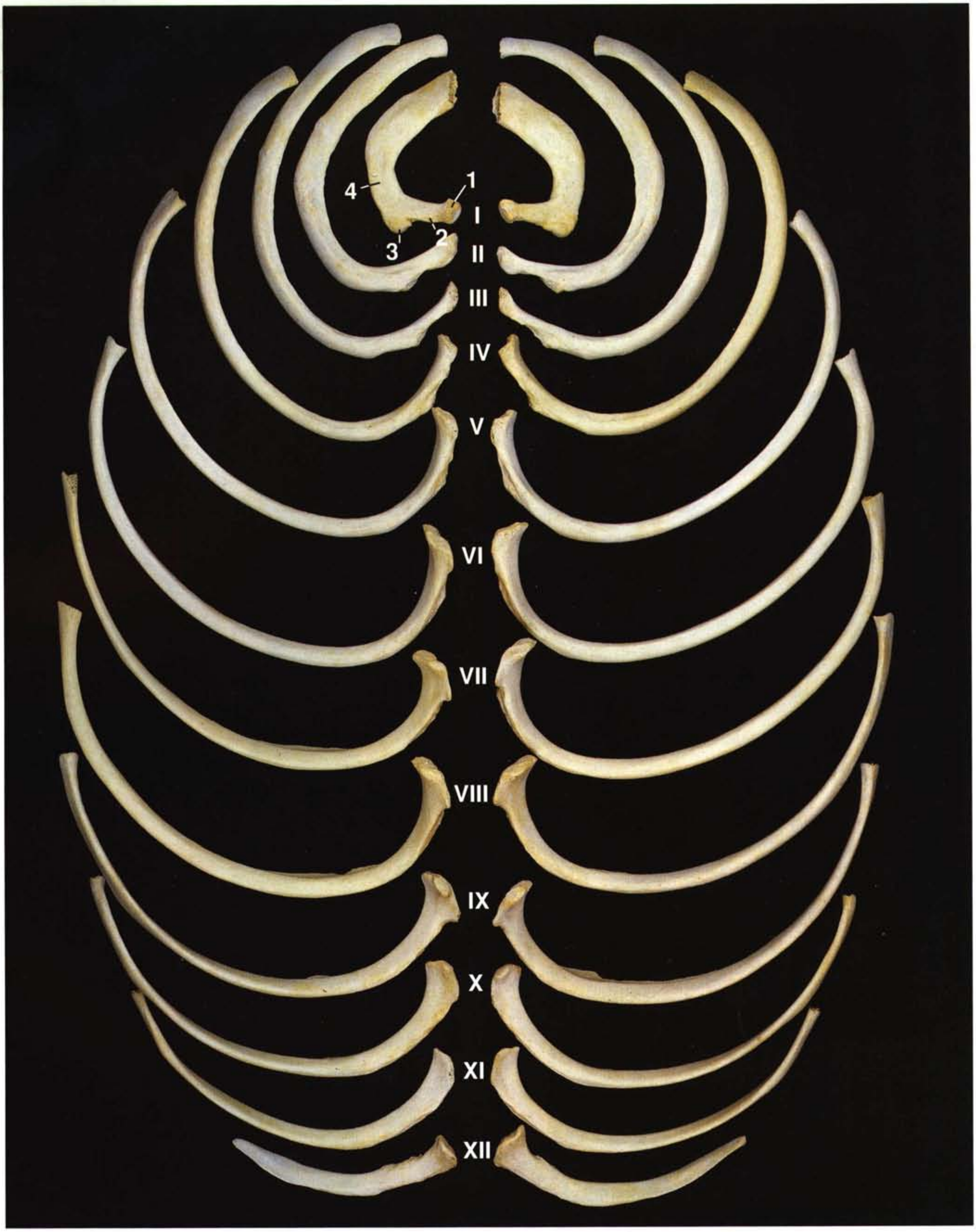

Disarticulated ribs. The twelve ribs arranged in a craniocaudal direction, first rib at top. I–XII = pairs of ribs; 1 = head of rib; 2 = neck of rib; 3 = tubercle of rib; 4 = body of rib.

The head of a typical rib articulates with the bodies of adjacent vertebrae and the intervening intervertebral disc. The first seven ribs are directly connected through the costal cartilages (not depicted) to the sternum (true ribs) while the remaining five ribs are only indirectly connected to the sternum (false ribs, ribs VIII to X) or end freely in the lateral wall of the trunk (floating ribs, ribs XI and XII).

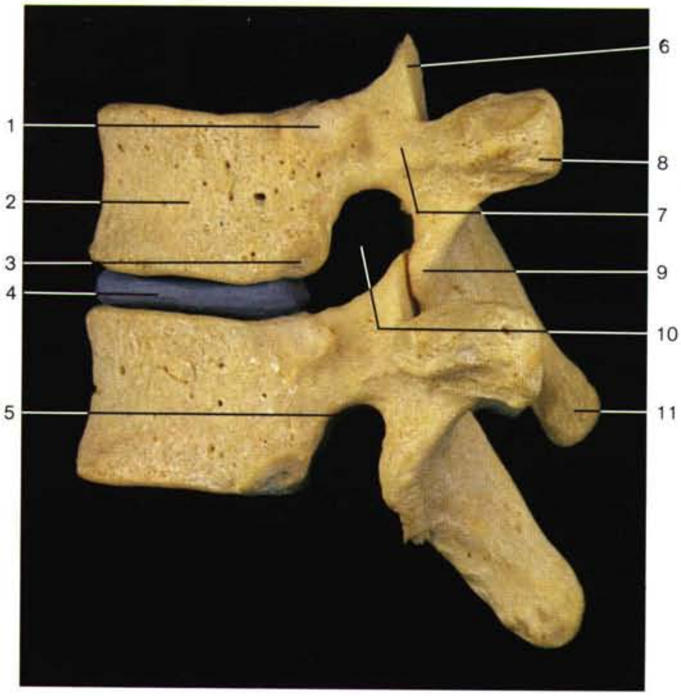

Two **thoracic vertebrae** (left lateral aspect).

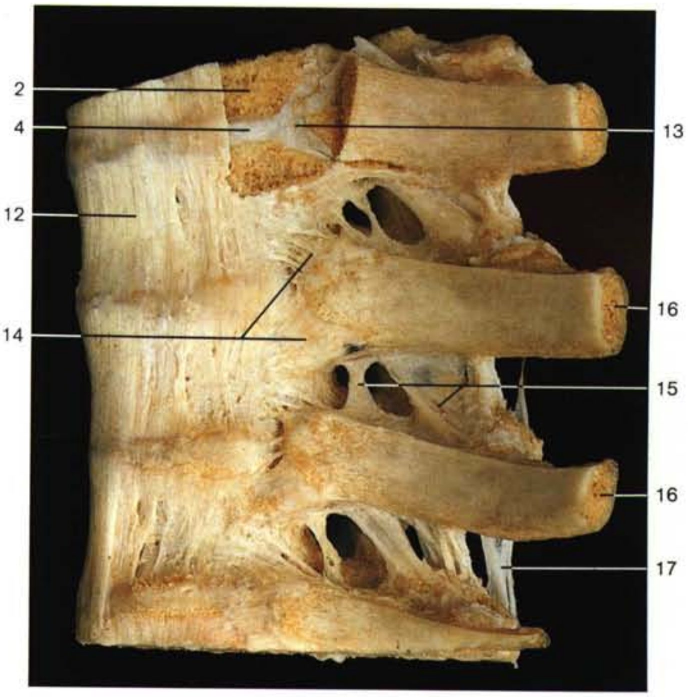

Ligaments of thoracic vertebrae and costovertebral joints (left anterolateral aspect).
In the upper joint, most of the radiate ligament and the anterior part of the head of the rib have been removed to expose the two joint cavities and the interposed intra-articular ligament.

1 Superior demifacet for head of rib
2 Body of vertebra
3 Inferior demifacet for head of rib
4 Intervertebral disc
5 Inferior vertebral notch
6 Superior articular facet and superior articular process
7 Pedicle
8 Transverse process and facet for tubercle of rib
9 Inferior articular process
10 Intervertebral foramen
11 Spinous process
12 Anterior longitudinal ligament
13 Intraarticular ligament
14 Radiate ligament
15 Superior costotransverse ligament
16 Body of rib
17 Intertransverse ligament

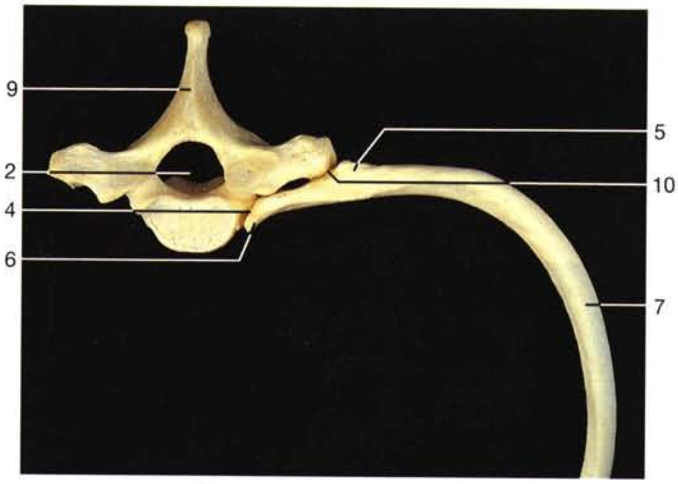

Location of costovertebral joints (superior aspect).

Costovertebral joints. Two thoracic vertebrae with an articulating rib (separated). Axis of movement indicated by dotted line. Blue = articular facets. (Schematic diagram.)

1 Superior articular process
2 Vertebral canal
3 Body of thoracic vertebra
4 Costovertebral joint (articular facets)
5 Tubercle of rib
6 Head of rib
7 Shaft or body of rib
8 Transverse process with articular facet
9 Spinous process
10 Costotransverse joint (articular facets)

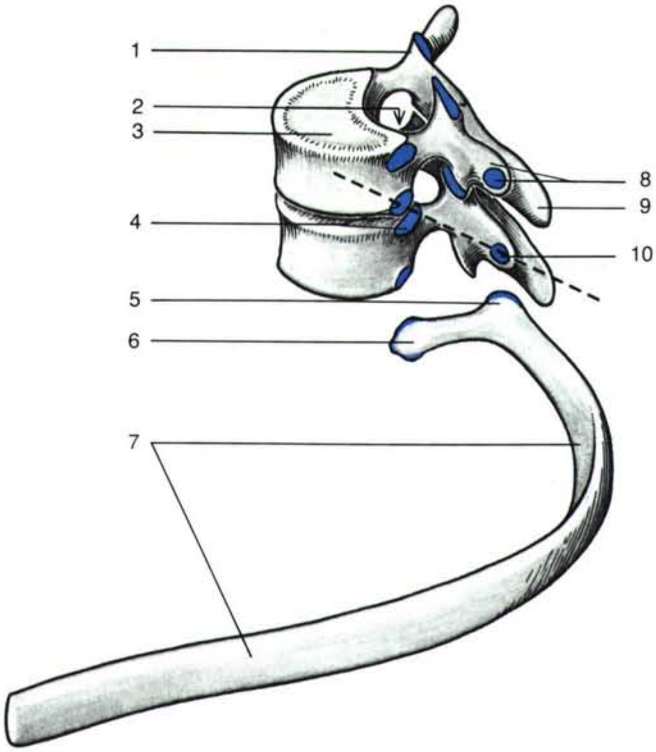

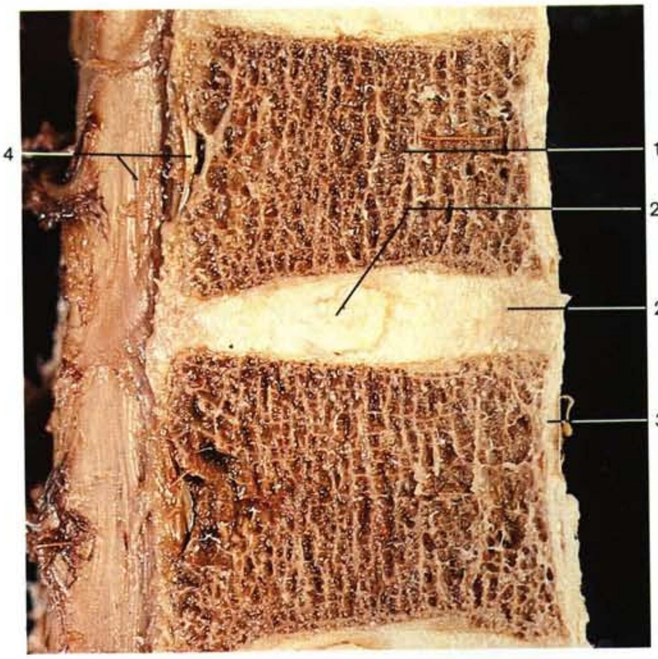

Median-sagittal section of the bodies of the vertebrae showing the **intervertebral discs,** each of which consists of an outer laminated portion and an inner core.

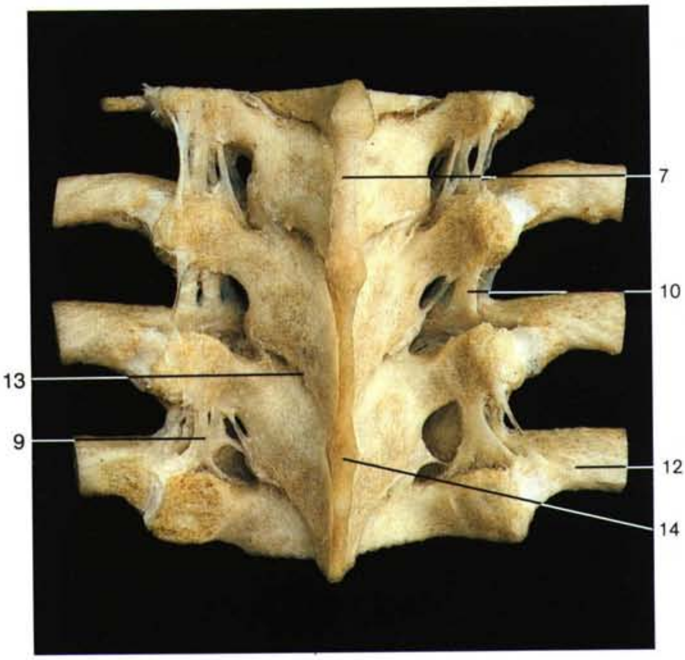

Ligaments of the vertebral column (dorsal aspect).

1 Body of vertebra
2 **Intervertebral disc**
  a Outer portion (anulus fibrosus)
  b Inner core **(nucleus pulposus)**
3 Anterior longitudinal ligament
4 Posterior longitudinal ligament and spinal dura mater
5 Transverse process of lumbar vertebra
6 Sacrum
7 **Supraspinous ligament**
8 Interspinous ligament
9 **Intertransverse ligament**
10 Superior costotransverse ligament
11 Transverse process of thoracic vertebra
12 Rib
13 **Ligamentum flavum**
14 Spinous process
15 Intervertebral foramen

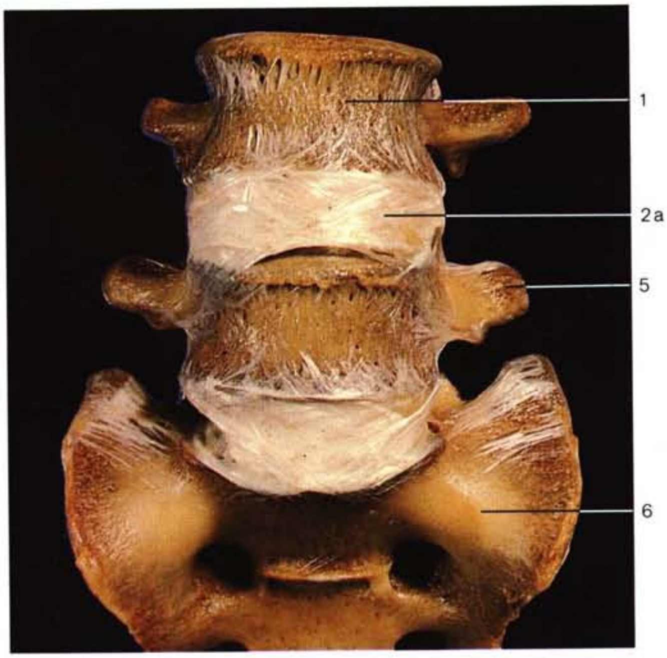

The two caudal lumbar vertebrae and the sacrum with their intervertebral discs (anterior aspect). Anterior longitudinal ligament removed.

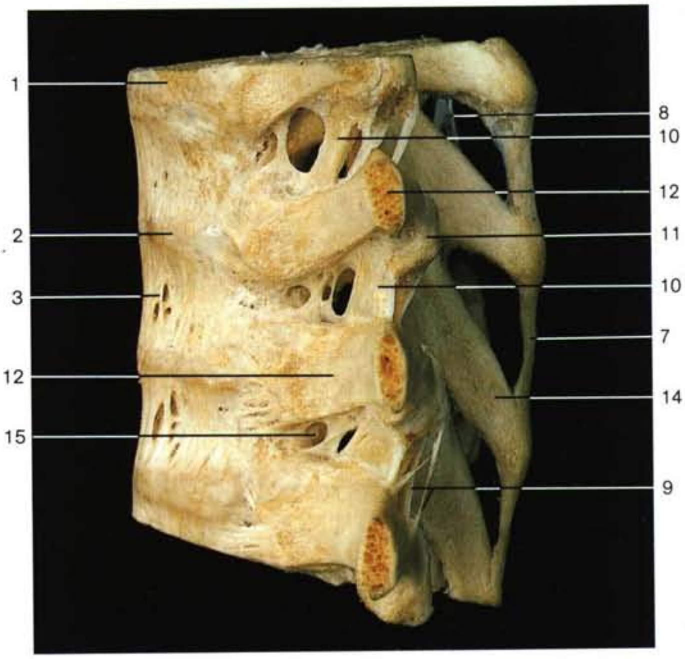

Ligaments of the vertebral column, thoracic part (left lateral aspect).

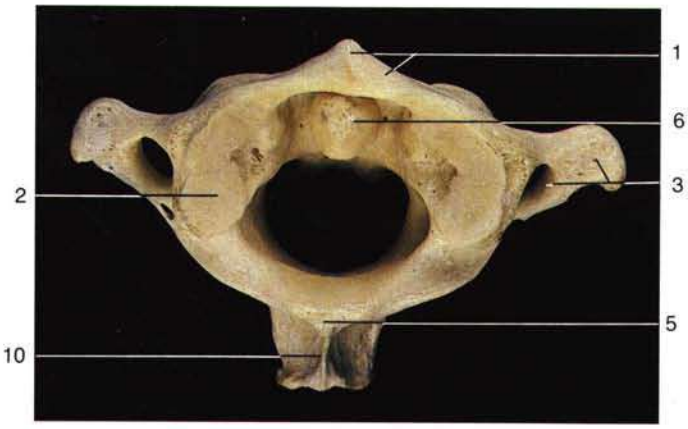

Atlas and axis (from above).

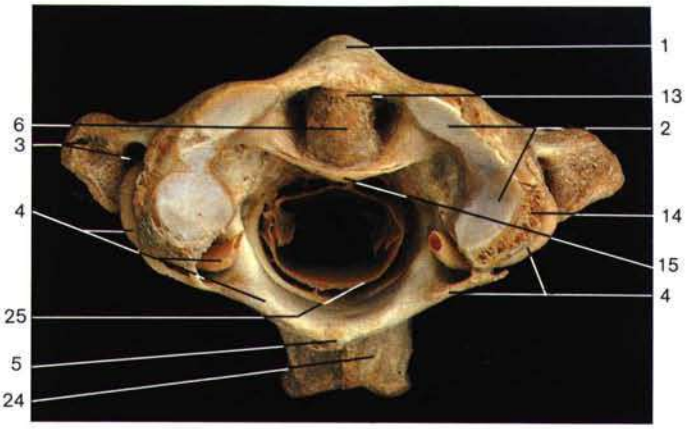

Median atlantoaxial joint and transverse ligament of atlas (from above). Dens of axis partly severed.

1 Anterior arch of **atlas** with anterior tubercle
2 Superior articular facet of atlas
3 **Foramen transversarium** and transverse process
4 Posterior arch of atlas and vertebral artery
5 Posterior tubercle of atlas
6 **Dens of axis**
7 Superior articular surface of axis
8 Body of axis
9 Pedicle and lamina of axis
10 Spinous process
11 Inferior articular process
12 Transverse process and vertebrarterial foramen (foramen transversarium)
13 **Median atlantoaxial joint** (anterior part)
14 Articular capsule of **atlantooccipital joint**
15 Transverse ligament of atlas
16 Occipital bone
17 **Atlantooccipital joint**
18 **Lateral atlantoaxial joint**
19 Third cervical vertebra
20 Superior longitudinal band of cruciform ligament
21 Alar ligaments
22 Transverse ligament of atlas
23 Inferior longitudinal band of cruciform ligament
24 Spinous process of axis
25 Dura mater

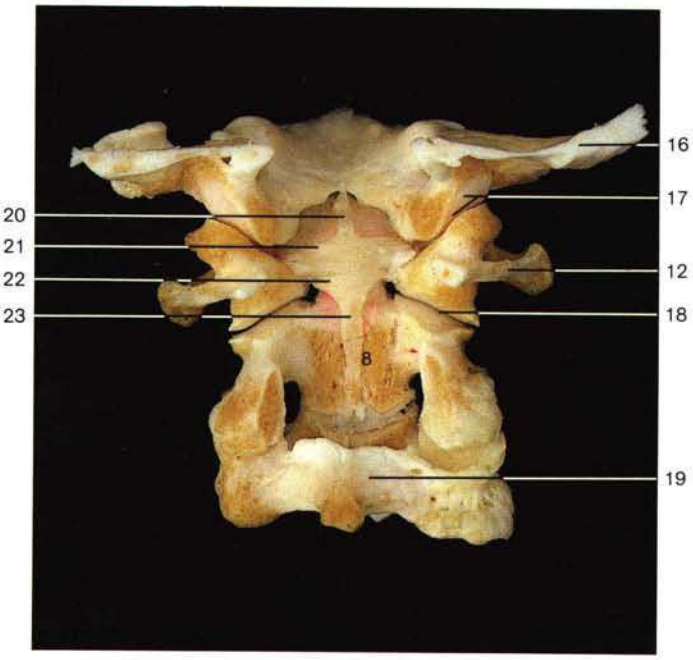

Atlantooccipital and atlantoaxial joints (posterior aspect). Posterior part of occipital bone, posterior arch of atlas and axis have been removed to show the cruciform ligament.

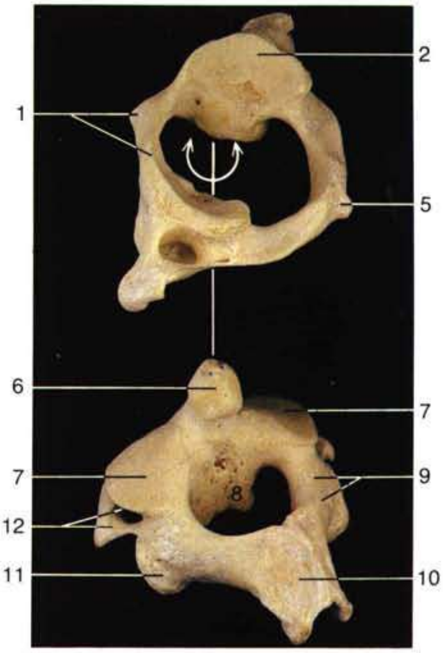

Atlas and axis. Left oblique posterolateral aspect, demonstrating the articulation of the dens of axis with atlas (cf. arrows).

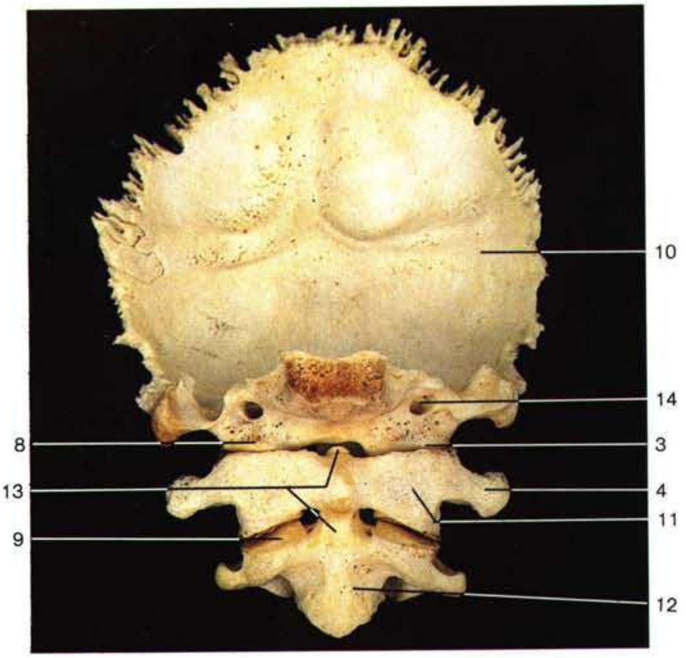

Occipital bone, atlas and axis (anterior aspect).

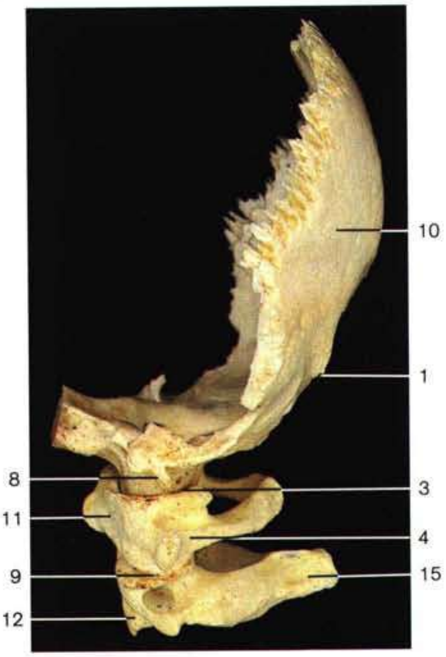

Occipital bone, atlas and axis (left lateral aspect).

1 External occipital protuberance
2 Foramen magnum
3 Atlantooccipital joint
4 Transverse process of atlas
5 Membrana tectoria
6 Posterior longitudinal ligament
7 Spinous process of third cervical vertebra
8 Occipital condyle
9 Lateral atlantoaxial joint
10 Occipital bone
11 Atlas
12 Axis
13 Dens of axis
14 Hypoglossal canal
15 Spinous processes of axis
16 Anterior longitudinal ligament

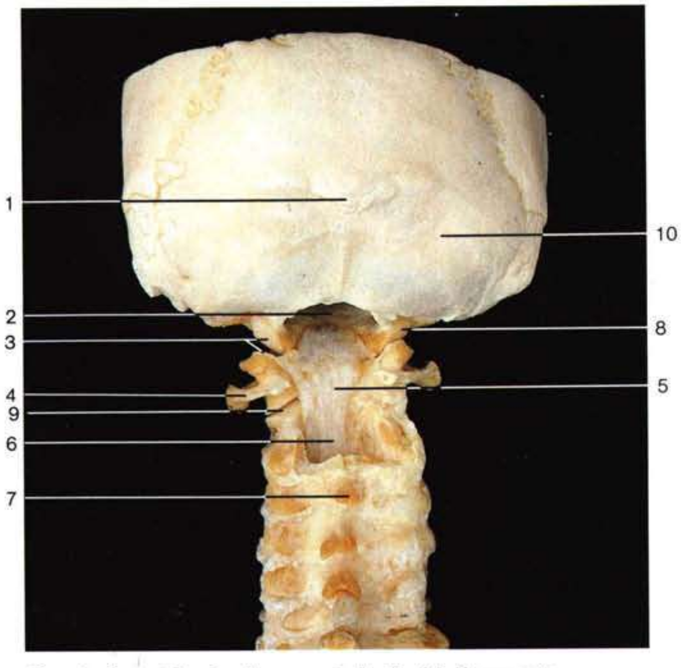

Cervical vertebral column and skull with ligaments (posterior aspect). Posterior arch of atlas and axis removed to show the membrana tectoria.

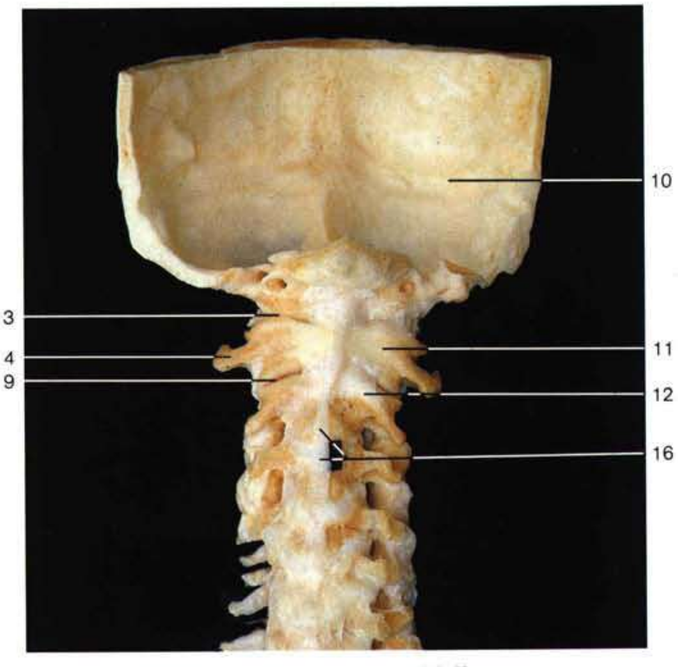

Cervical vertebral column and skull with ligaments (anterior aspect). Anterior part of occipital bone removed.

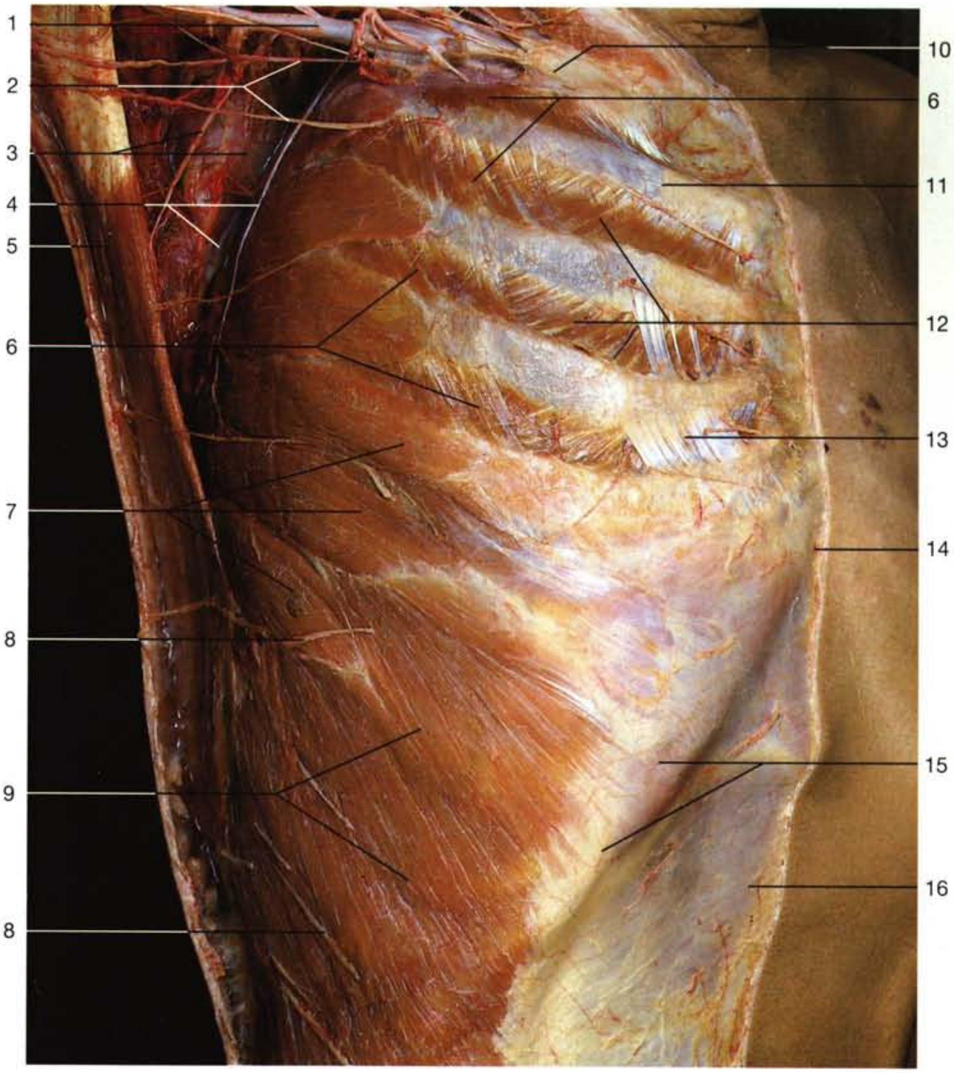

1 Axillary vein
2 Intercostobrachial nerves
3 Subscapularis muscle and thoracodorsal nerve
4 Long thoracic nerve, lateral thoracic artery and vein
5 Latissimus dorsi muscle
6 **External intercostal** muscles
7 **Serratus anterior** muscle
8 Lateral cutaneous branches of intercostal nerves
9 External abdominal oblique muscle
10 Clavicle (divided)
11 Second rib (costo-chondral junction)
12 **Internal intercostal muscles**
13 External intercostal membrane
14 Position of xiphoid process
15 Costal arch or margin
16 Anterior layer of rectus sheath

Muscles of the thorax, superficial layer (lateral aspect). Upper limb elevated. Pectoralis major and minor muscles have been removed.

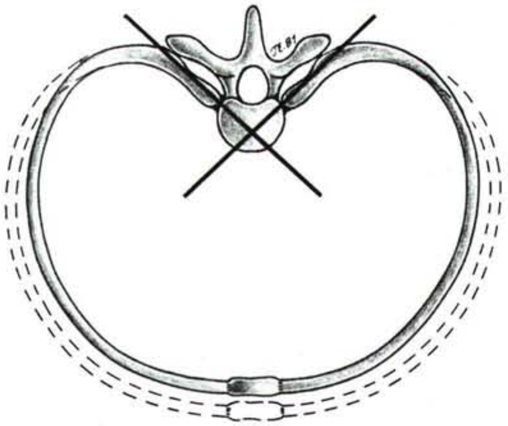

Effect of intercostal muscles on the costovertebral and costotransverse joints. Axes of movement indicated by lines; direction of movements indicated by arrows.

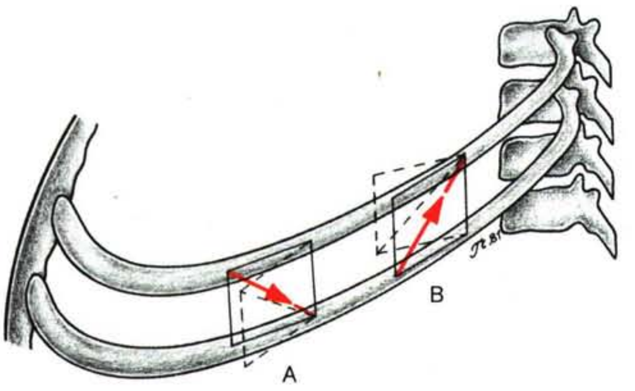

A Action of internal intercostal muscles (expiration)
B Action of external intercostal muscles (inspiration)

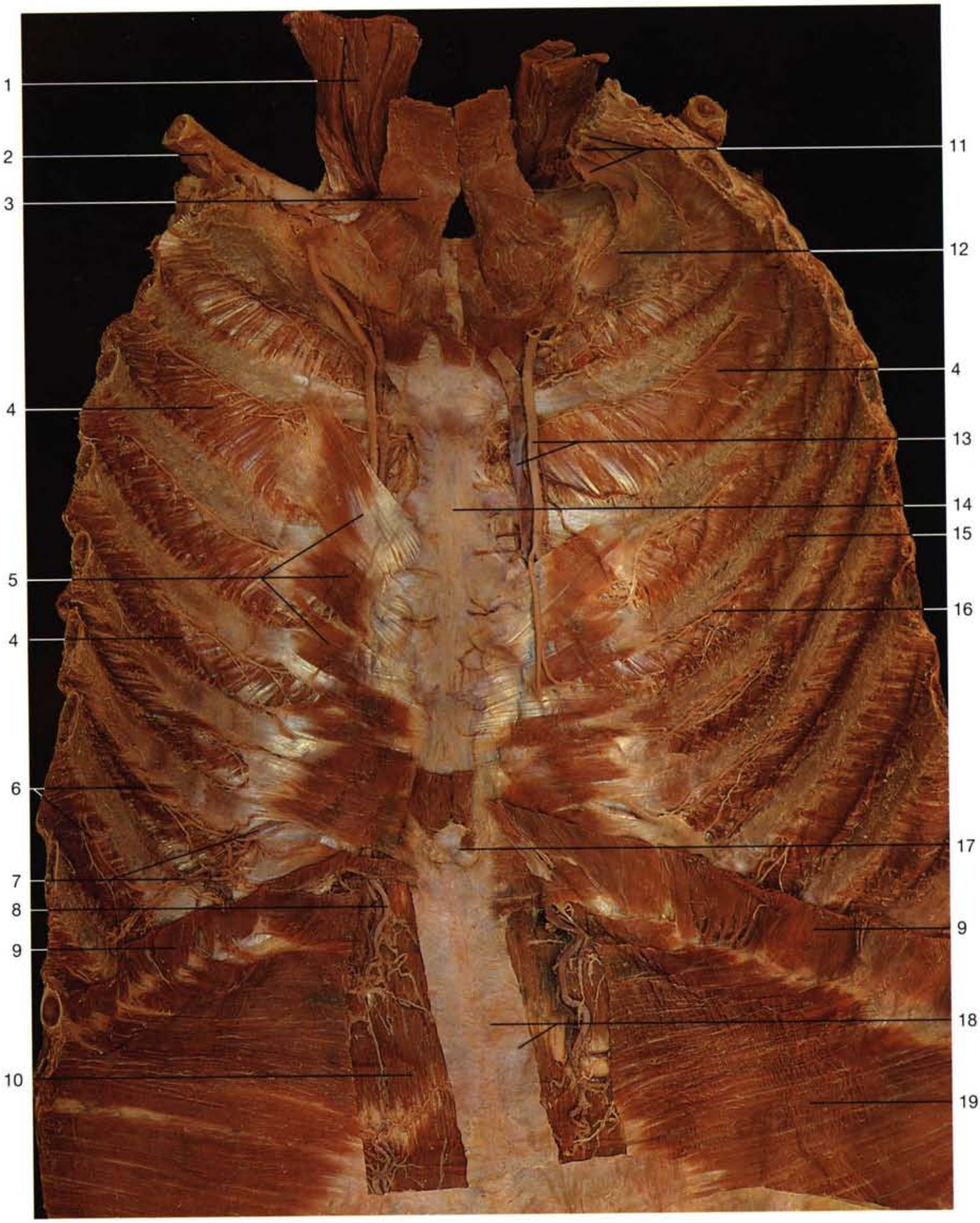

Anterior thoracic wall (posterior aspect). Diaphragm partly removed, posterior layer of rectus sheath fenestrated on both sides.

1 Sternocleidomastoid muscle (divided)
2 Clavicle
3 Sternothyroid muscle
4 **Internal intercostal muscle**
5 **Transversus thoracis muscle**
6 Intercostal arteries and nerves
7 Musculophrenic artery
8 Superior epigastric artery and vein
9 Diaphragm (divided)
10 Rectus abdominis muscle
11 Subclavian artery and brachial plexus
12 First rib
13 Internal thoracic artery and vein
14 Sternum
15 Innermost intercostal muscle
16 Intercostal artery and vein
17 Xiphoid process
18 Linea alba and posterior layer of rectus sheath
19 **Transversus abdominis muscle**

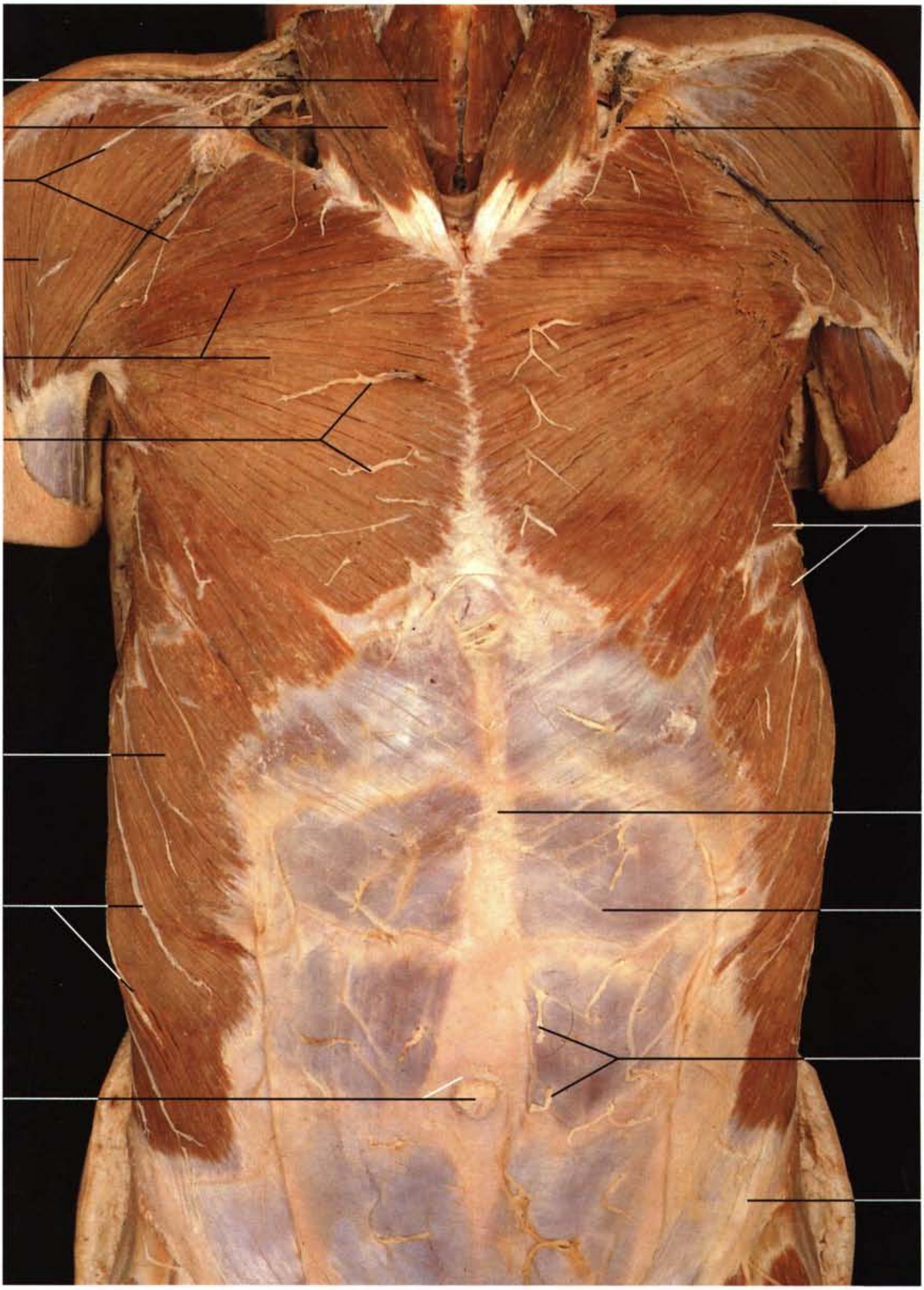

Superficial muscles of the anterior thoracic and abdominal wall. The fascia of pectoralis major muscle and the abdominal wall have been removed; the anterior layer of the sheath of the rectus abdominis muscle is displayed.

1 Sternohyoid muscle
1 Sternocleidomastoid muscle
3 Supraclavicular nerves (branches of cervical plexus)
4 Deltoid muscle
5 **Pectoralis major muscle**
6 **Anterior cutaneous branches of intercostal nerves**
7 External abdominal oblique muscle
8 **Lateral cutaneous branches of intercostal nerves**

9 Umbilicus and umbilical ring
10 Clavicle
11 Cephalic vein
12 Serratus anterior muscle
13 Linea alba
14 **Sheath of rectus abdominis** muscle (anterior layer)
15 Inguinal ligament

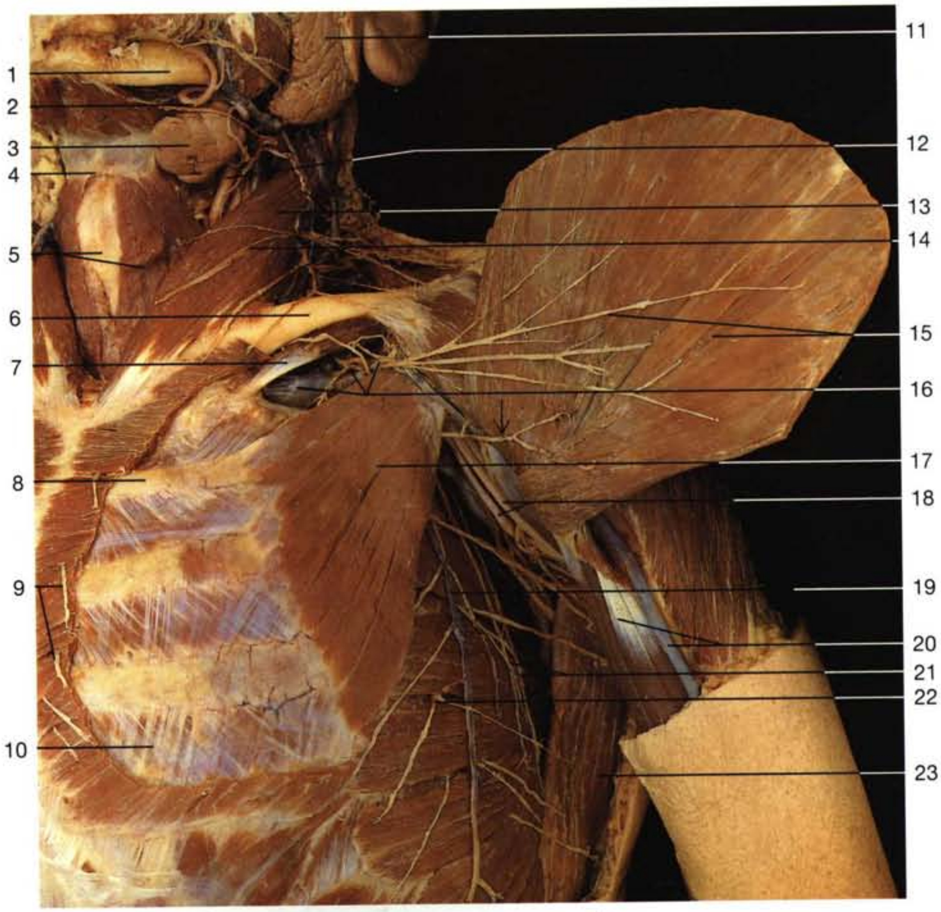

Thoracic wall I (anterior aspect). Left pectoralis major muscle has been divided and reflected. Note the connection of the cephalic vein with the subclavian vein. Arrow: medial pectoral nerve.

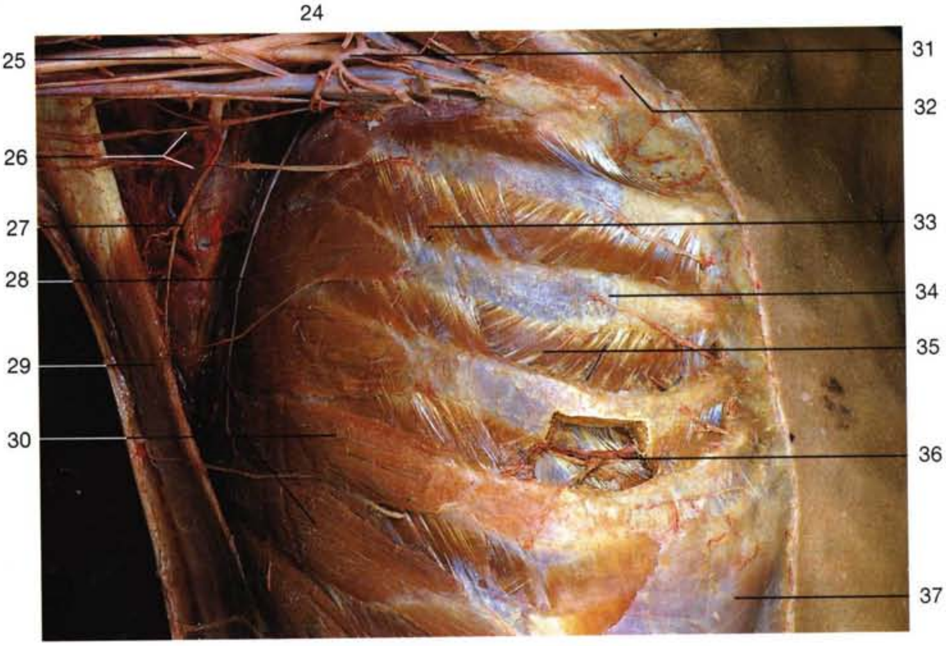

Thoracic wall (lateral aspect). Pectoralis major and minor muscles have been removed. A section of the 4th rib has been cut and removed to display the intercostal vessels and nerve.

1. Mandible
2. Facial artery
3. Submandibular gland
4. Hyoid bone
5. Thyroid cartilage and sternohyoid muscle
6. Clavicle
7. Subclavius muscle
8. Second rib
9. Anterior cutaneous branches of intercostal nerves
10. External intercostal membrane
11. Parotid gland
12. External carotid artery
13. Sternocleidomastoid muscle and cutaneous branches of cervical plexus
14. Supraclavicular nerves
15. Pectoralis major muscle and lateral pectoral nerves
16. Thoracoacromial artery and subclavian vein
17. **Pectoralis minor muscle**
18. **Median and ulnar nerve**
19. Thoracoepigastric vein
20. Cephalic vein and long head of biceps brachii muscle
21. **Lateral thoracic artery and long thoracic nerve**
22. Lateral cutaneous branches of intercostal nerve
23. Latissimus dorsi muscle
24. Median nerve
25. Axillary artery
26. Intercostobrachial nerves
27. **Thoracodorsal nerve**
28. Long thoracic nerve
29. Latissimus dorsi muscle
30. **Serratus anterior muscle**
31. Thoracoacromial artery
32. Clavicle
33. External intercostal muscle
34. Third rib
35. **Internal intercostal muscle**
36. Anterior intercostal artery and vein
37. Costal arch or margin

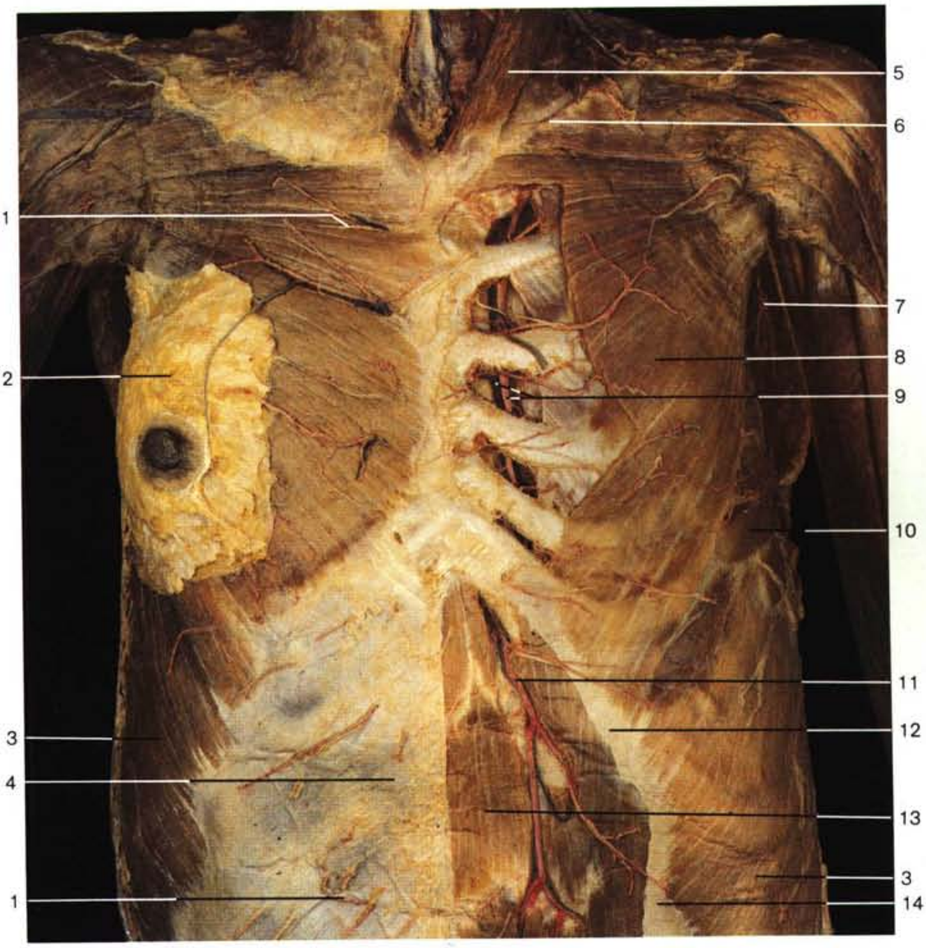

Thoracic wall II (anterior aspect). Dissection of the **internal thoracic artery and vein.** Left pectoralis major muscle partly removed. Anterior lamina of the rectus sheath on the left side has been removed.

1 Anterior perforating branches of intercostal nerve
2 Mammary gland
3 External abdominal oblique muscle
4 Rectus sheath (anterior layer)
5 Sternocleidomastoid muscle
6 Clavicle
7 Lateral thoracic artery and vein
8 Pectoralis major muscle
9 **Internal thoracic artery and vein**
10 Serratus anterior muscle
11 **Superior epigastric artery and vein**
12 Costal margin
13 Rectus abdominis muscle
14 Cut edge of the anterior layer of the rectus sheath
15 Subclavian artery
16 Highest intercostal artery
17 Internal thoracic artery
18 Musculophrenic artery
19 Superficial epigastric artery
20 Deep circumflex iliac artery
21 Superior epigastric artery
22 Inferior epigastric artery
23 Superficial circumflex iliac artery

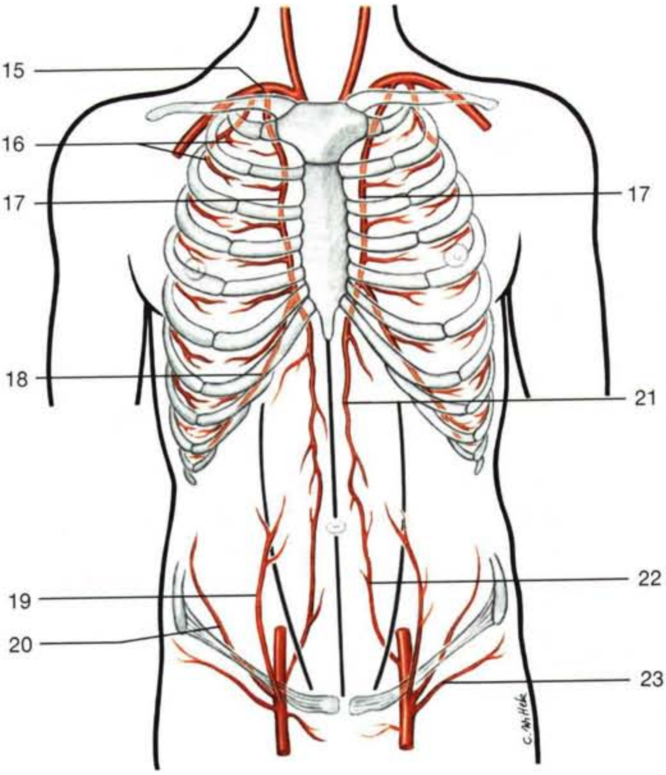

Main arteries of thoracic and abdominal wall.

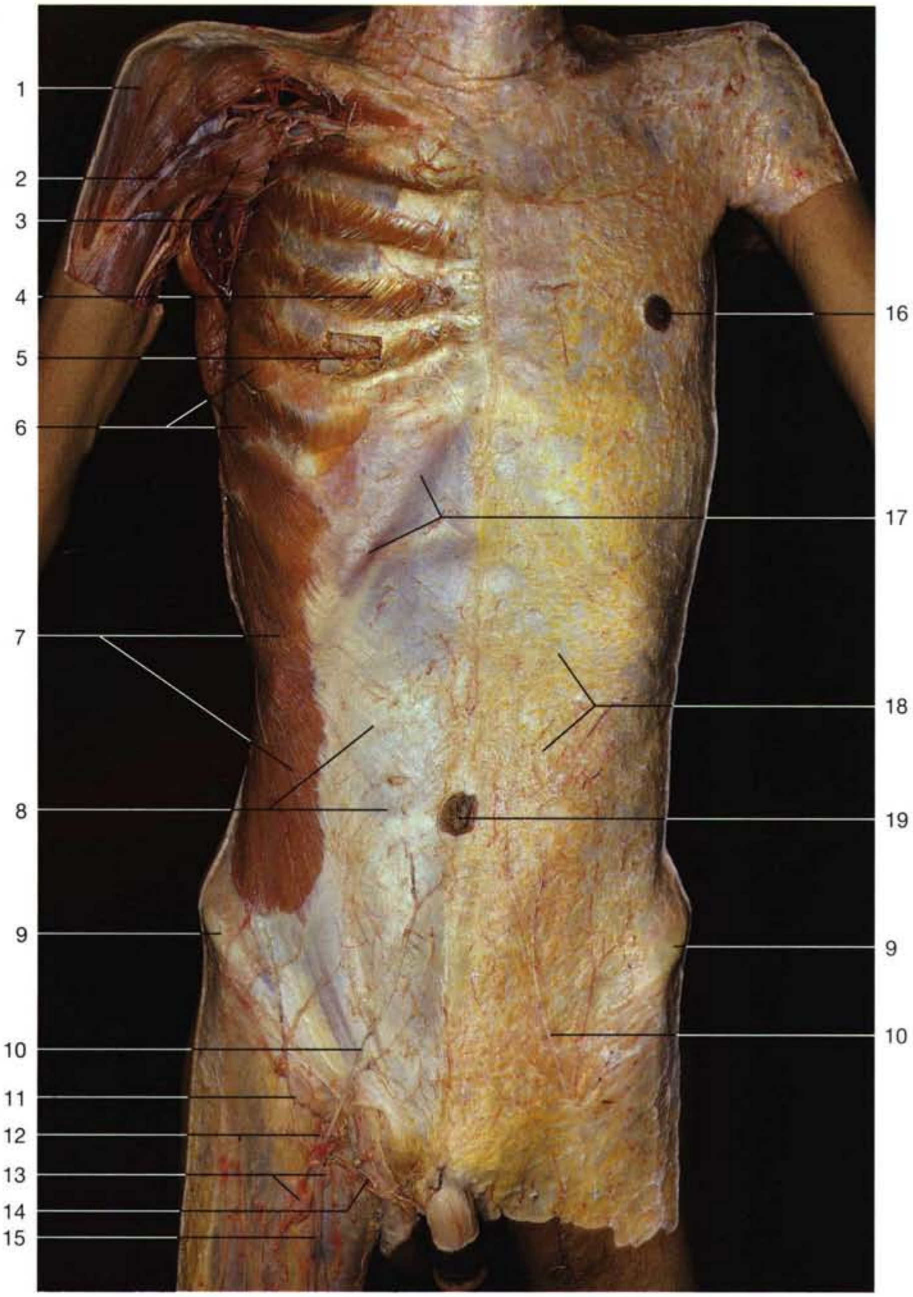

1 Deltoid muscle
2 Cephalic vein
3 Pectoralis major muscle (divided)
4 **Internal intercostal muscle**
5 **Intercostal artery and vein** (intercostal space, fenestrated)
6 Serratus anterior muscle
7 **External abdominal oblique muscle**
8 **Anterior layer of rectus sheath**
9 Iliac crest
10 Superficial epigastric vein
11 Superficial circumflex iliac vein
12 Saphenous opening
13 Superficial inguinal lymph nodes
14 Superficial external pudendal veins
15 Great saphenous vein
16 Nipple
17 Costal margin
18 Subcutaneous fatty tissue
19 Umbilicus
20 Anterior layer of rectus sheath
21 Rectus abdominis muscle
22 Posterior layer of rectus sheath
23 **Internal abdominal oblique muscle**
24 **External abdominal oblique muscle**
25 **Transversus abdominis muscle**
26 Transversal fascia and peritoneum
27 Psoas major muscle
28 Body of lumbar vertebra (L_4)
29 Quadratus lumborum muscle
30 Medial tract of erector spinae muscle
31 Lateral tract of erector spinae muscle (longissimus and iliocostalis muscles)
32 Small intestine
33 Left ureter
34 **Abdominal aorta**
35 **Inferior vena cava**
36 Descending colon
37 Spinous process

Thoracic and abdominal wall I. Right pectoralis major and minor muscles are divided. Muscles of thoracic and abdominal wall on right side are displayed.

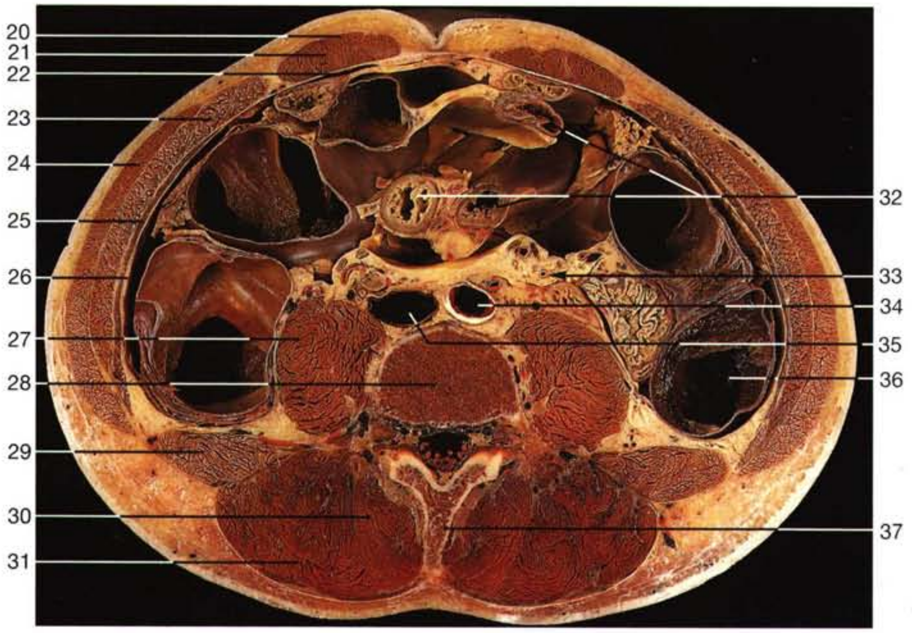

Horizontal section of the trunk at the level of the umbilicus, superior to arcuate line (inferior aspect).

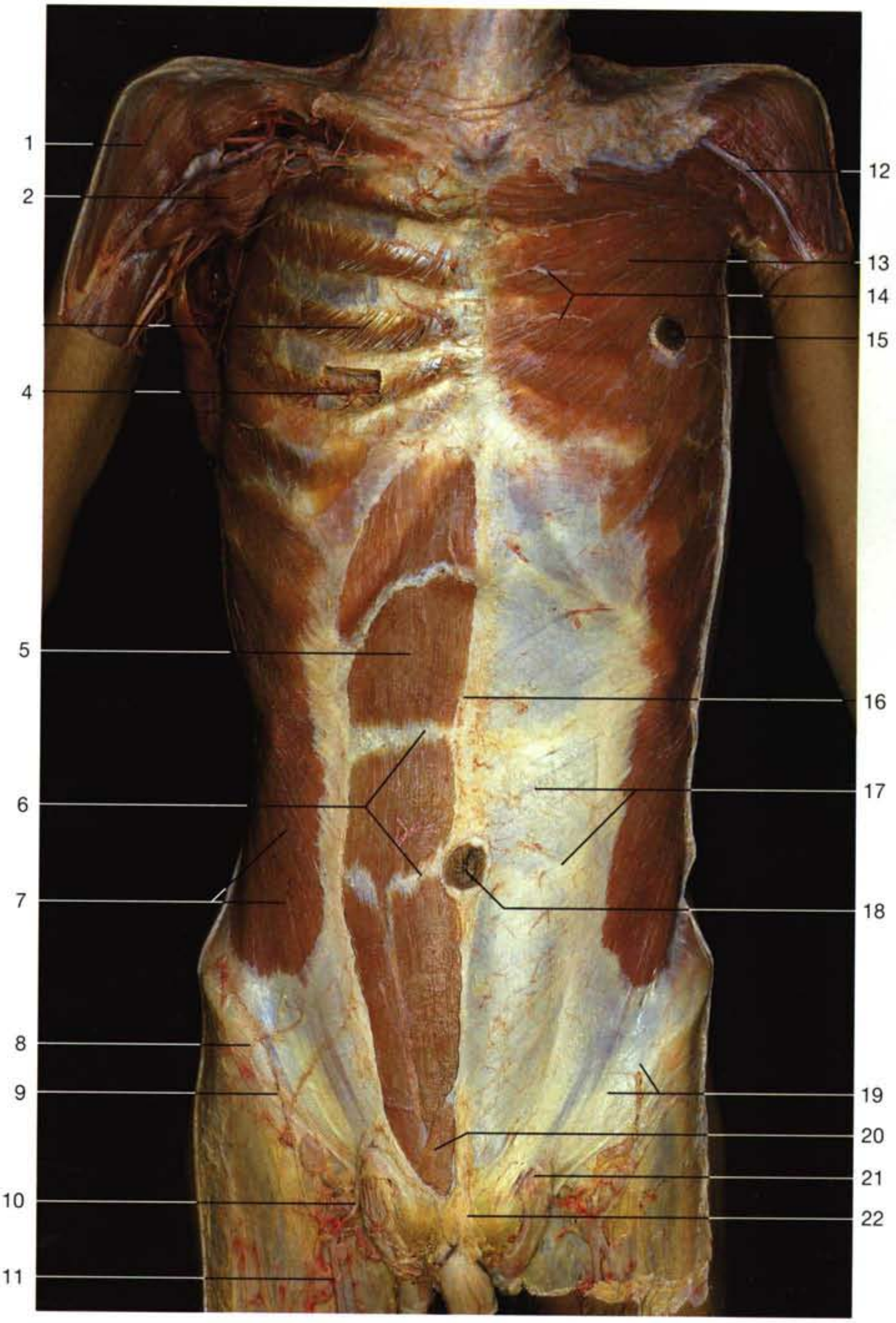

Thoracic and abdominal wall II. Right pectoralis major and minor muscles and anterior layer of rectus sheath have been removed on the right side.

1 Deltoid muscle
2 Pectoralis major muscle (divided)
3 **Internal intercostal muscle**
4 Intercostal artery and vein
5 **Rectus abdominis muscle**
6 Tendinous intersections
7 **External abdominal oblique muscle**
8 Anterior superior iliac spine
9 Superficial circumflex iliac vein
10 Superficial epigastric vein
11 Great saphenous vein
12 Cephalic vein
13 **Pectoralis major muscle**
14 Anterior cutaneous branches of intercostal nerves
15 Nipple
16 Linea alba
17 **Anterior layer of rectus sheath**
18 Umbilicus
19 **Inguinal ligament**
20 Pyramidalis muscle
21 Superficial inguinal ring and spermatic cord
22 Suspensory ligament of penis
23 Longissimus and iliocostalis muscles
24 Multifidus muscle
25 Quadratus lumborum muscle
26 Latissimus dorsi muscle
27 Psoas major muscle
28 Spinous process
29 Body of 1st lumbar vertebra
30 **Transversus abdominis muscle**
31 **Internal abdominal oblique muscle**

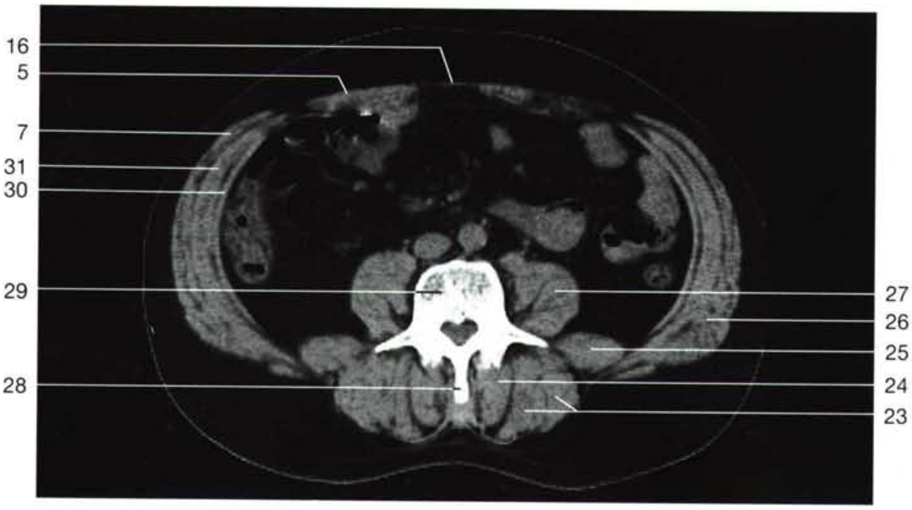

Horizontal section through the body at the level of 4th lumbar vertebra; seen from below. (CT-Scan.)

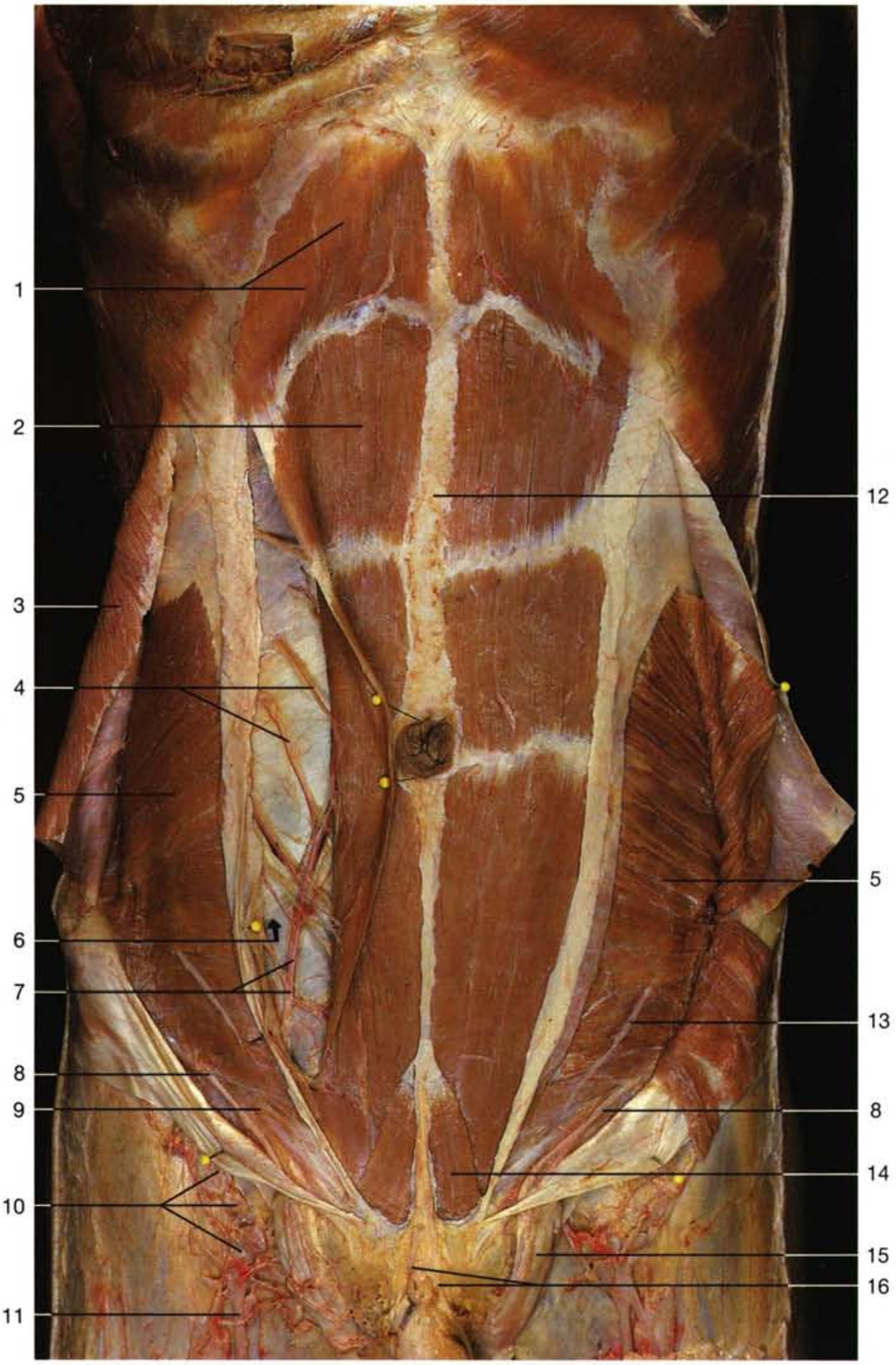

Thoracic and abdominal wall III. External abdominal oblique muscle has been divided and reflected on both sides. The right rectus muscle has been reflected medially to display the posterior layer of rectus sheath. Arrow: location of arcuate line.

1 Costal margin
2 **Rectus abdominis muscle**
3 External abdominal oblique muscle (reflected)
4 Thoracoabdominal (intercostal) nerves with accompanying vessels
5 Internal abdominal oblique muscle
6 **Arcuate line** (arrow)
7 **Inferior epigastric artery and vein**
8 Ilioinguinal nerve
9 Position of deep inguinal ring
10 Superficial inguinal lymph nodes
11 Great saphenous vein
12 Linea alba
13 Iliohypogastric nerve
14 **Pyramidalis** muscle
15 Spermatic cord
16 Fundiform ligament of penis

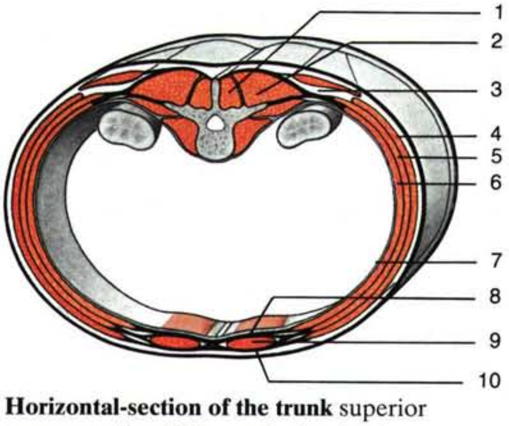

Horizontal-section of the trunk superior to arcuate line. (Schematic drawing.)

1 Medial column of intrinsic muscles of the back
2 Lateral column of erector spinae muscle
3 Thoracolumbar fascia with superficial and deep layer
4 External oblique muscle
5 Internal oblique muscle
6 Transversus abdominis muscle
7 Transversalis fascia
8 Posterior layer of rectus sheath
9 Rectus abdominis muscle
10 Anterior layer of rectus sheath

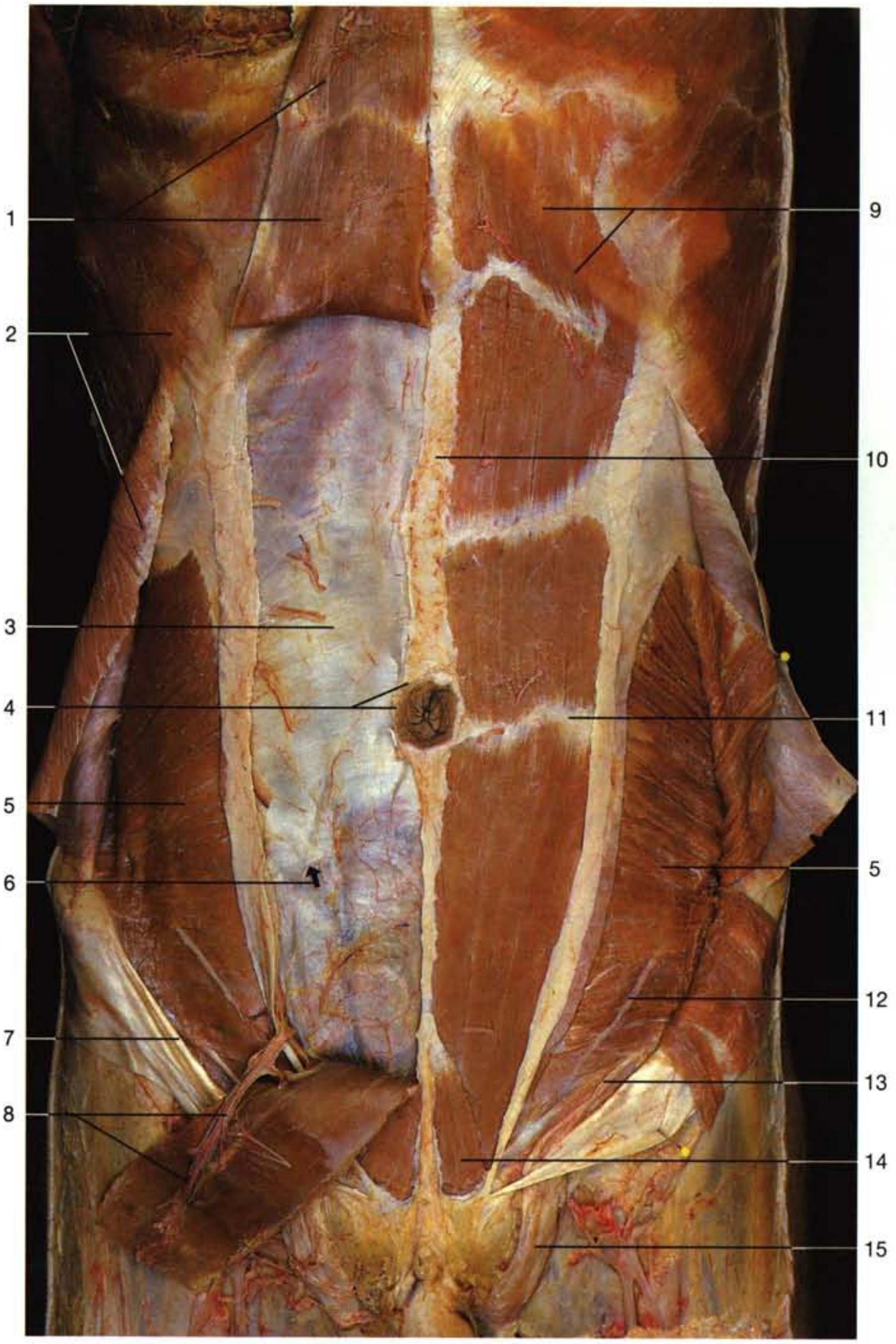

1 **Rectus abdominis muscle** (reflected)
2 External abdominal oblique muscle (divided)
3 Posterior layer of rectus sheath
4 Umbilical ring
5 **Internal abdominal oblique muscle**
6 Arcuate line (arrow)
7 Inguinal ligament
8 **Inferior epigastric artery and vein** and rectus abdominis muscle (divided and reflected)
9 Costal margin
10 Linea alba
11 Tendinous intersection
12 Iliohypogastric nerve
13 **Ilioinguinal nerve**
14 **Pyramidalis** muscle
15 Spermatic cord

Thoracic and abdominal wall IV. External abdominal oblique muscle has been divided and reflected on both sides. The right rectus muscle has been cut and reflected to display the posterior layer of rectus sheath. Arrow: location of arcuate line.

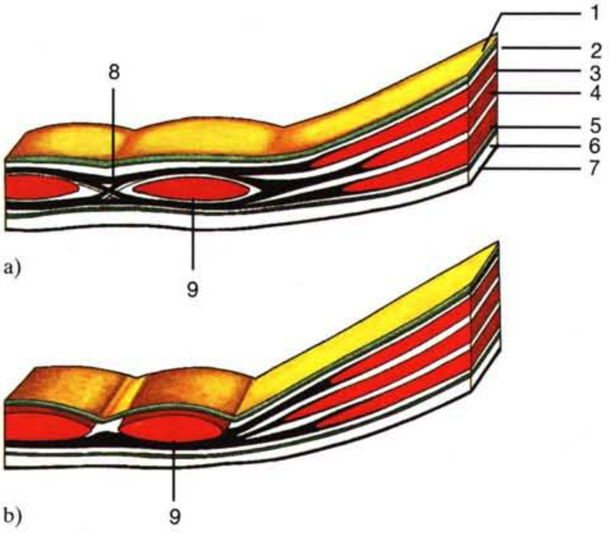

1 Peritoneum
2 Transversalis fascia (green)
3 Transversus abdominis muscle
4 Internal abdominal oblique muscle
5 External abdominal oblique muscle
6 Fascia of external abdominal oblique muscle (green)
7 Skin
8 Linea alba
9 Rectus abdominis muscle

Transverse sections through the abdominal wall superior (a) and inferior (b) to arcuate line.

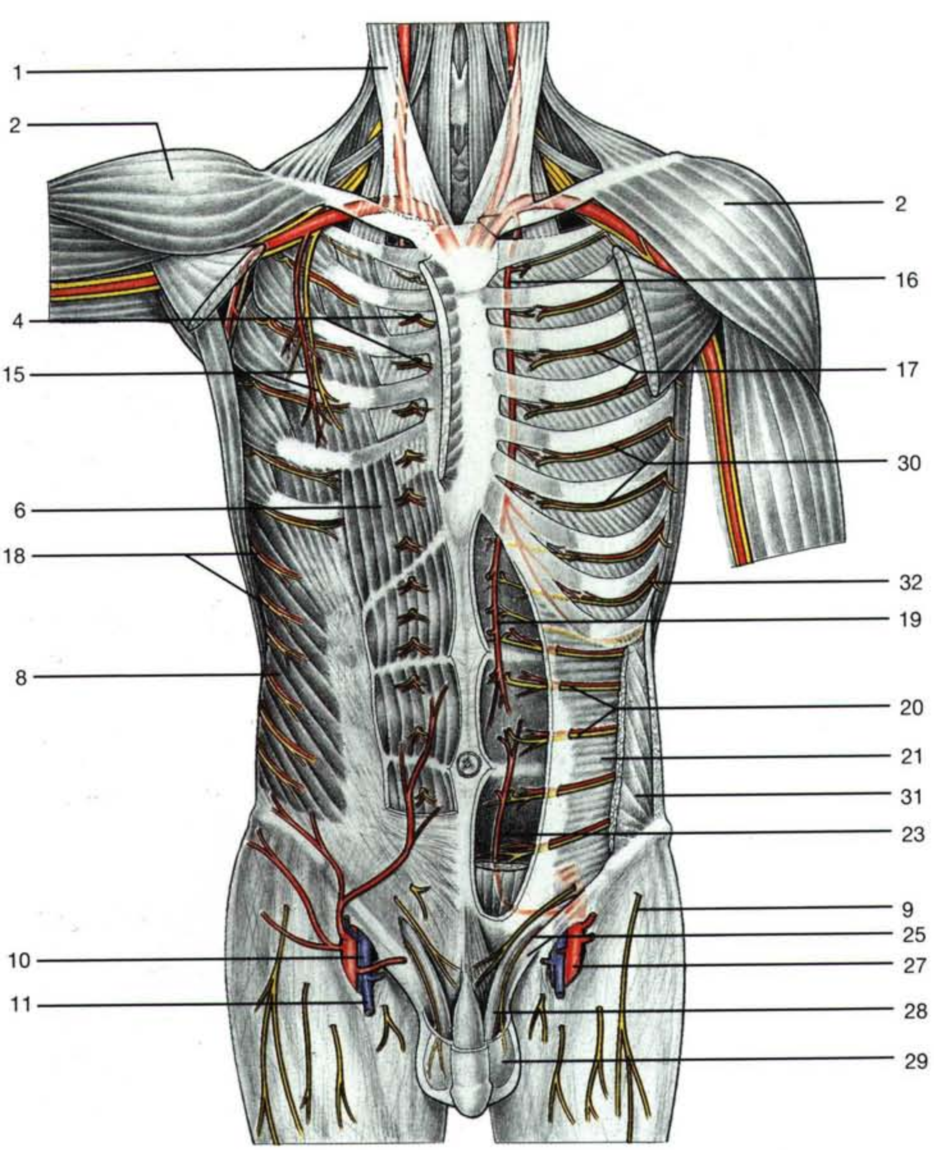

Thoracic and abdominal wall (schematic drawing). Note the segmental organization of the blood vessels and nerves. Right side: superficial layers; left side: deeper layers.

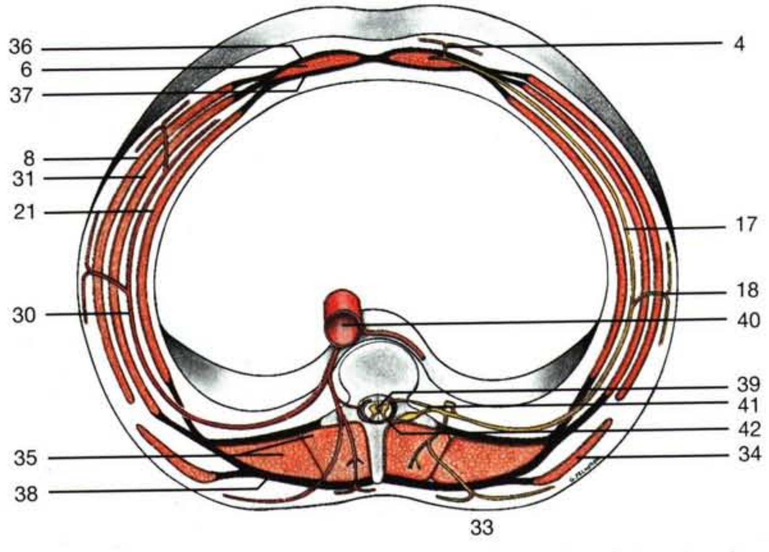

Horizontal section of the abdominal wall (from above) showing the location of the intercostal arteries (left side) and nerves (right side).

1 Sternocleidomastoid muscle
2 Deltoid muscle
3 Pectoralis major muscle
4 **Anterior cutaneous branches of intercostal nerves**
5 Cut edge of anterior layer of rectus sheath
6 **Rectus abdominis muscle**
7 Tendinous intersection
8 External abdominal oblique muscle
9 Lateral femoral cutaneous nerve
10 Femoral vein
11 Great saphenous vein
12 Medial supraclavicular nerves
13 Pectoralis minor muscle (reflected) and medial pectoral nerves
14 Axillary vein
15 Long thoracic nerve and lateral thoracic artery
16 Internal thoracic artery
17 **Intercostal nerves**
18 **Lateral cutaneous branches of intercostal nerves**
19 Superior epigastric artery
20 Thoracoabdominal (intercostal) nerves
21 Transversus abdominis muscle
22 Posterior layer of rectus sheath
23 **Inferior epigastric artery**
24 Lateral femoral cutaneous nerve
25 Inguinal ligament and ilioinguinal nerve
26 Femoral nerve
27 Femoral artery
28 Spermatic cord
29 Testis
30 **Posterior intercostal arteries**
31 Internal abdominal oblique muscle
32 Lateral cutaneous branch of intercostal nerve
33 Dorsal branch of spinal nerve
34 Latissimus dorsi muscle
35 Deep muscles of the back (medial and lateral tract)
36 Anterior layer of rectus sheath
37 Posterior layer of rectus sheath
38 Thoracolumbar fascia
39 Spinal cord
40 Aorta
41 Ventral root } of spinal
42 Dorsal root } nerve

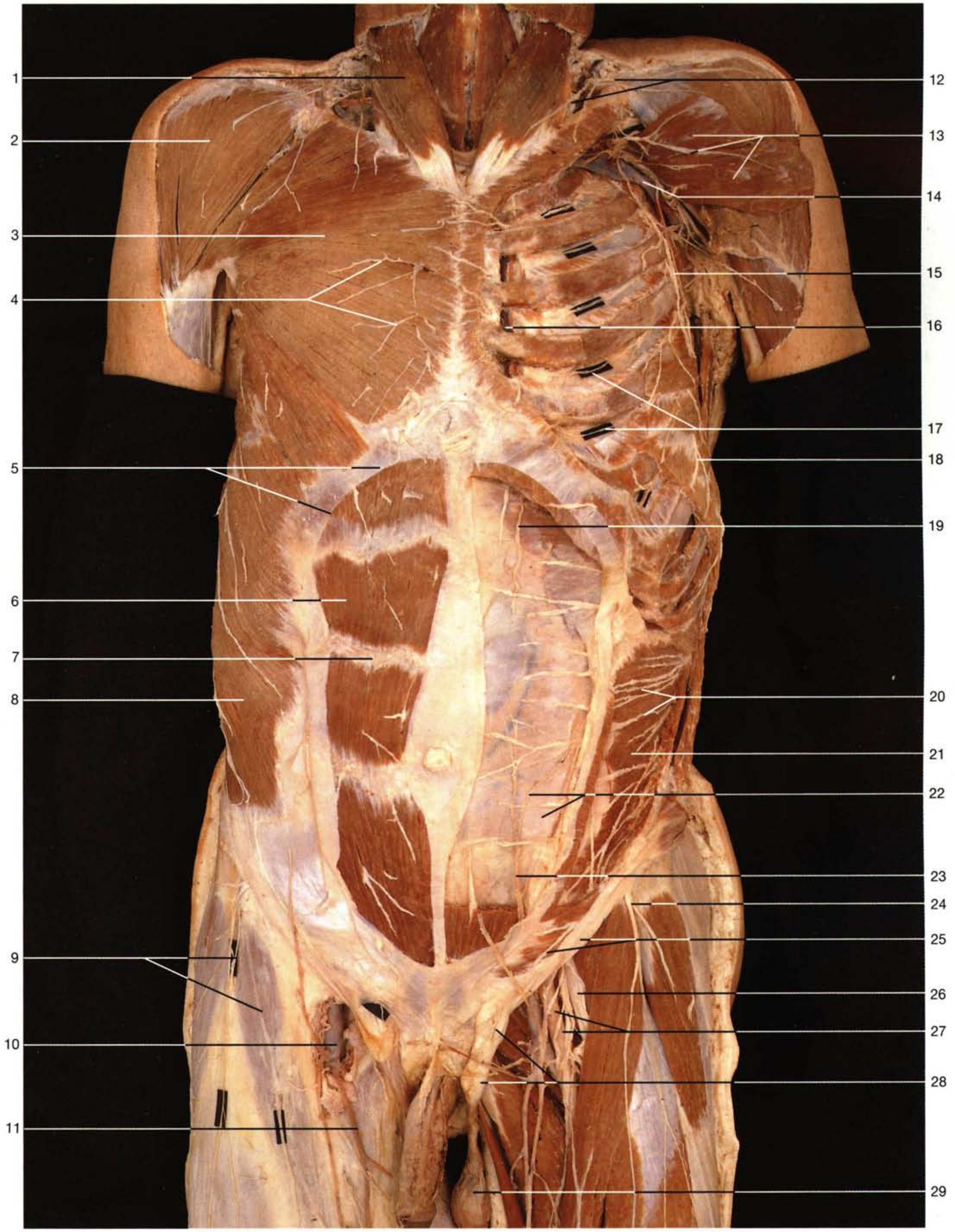
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

Thoracic wall and abdominal wall V. Right side: superficial layers; left side: deeper layers (anterior aspect). Pectoralis major and minor muscles, the external and internal intercostal muscles on the left side have been removed to display the intercostal nerves. The anterior layer of rectus sheath, the left rectus abdominis muscle and the external and internal abdominal oblique muscle have been removed to show the thoracoabdominal nerves within the abdominal wall.

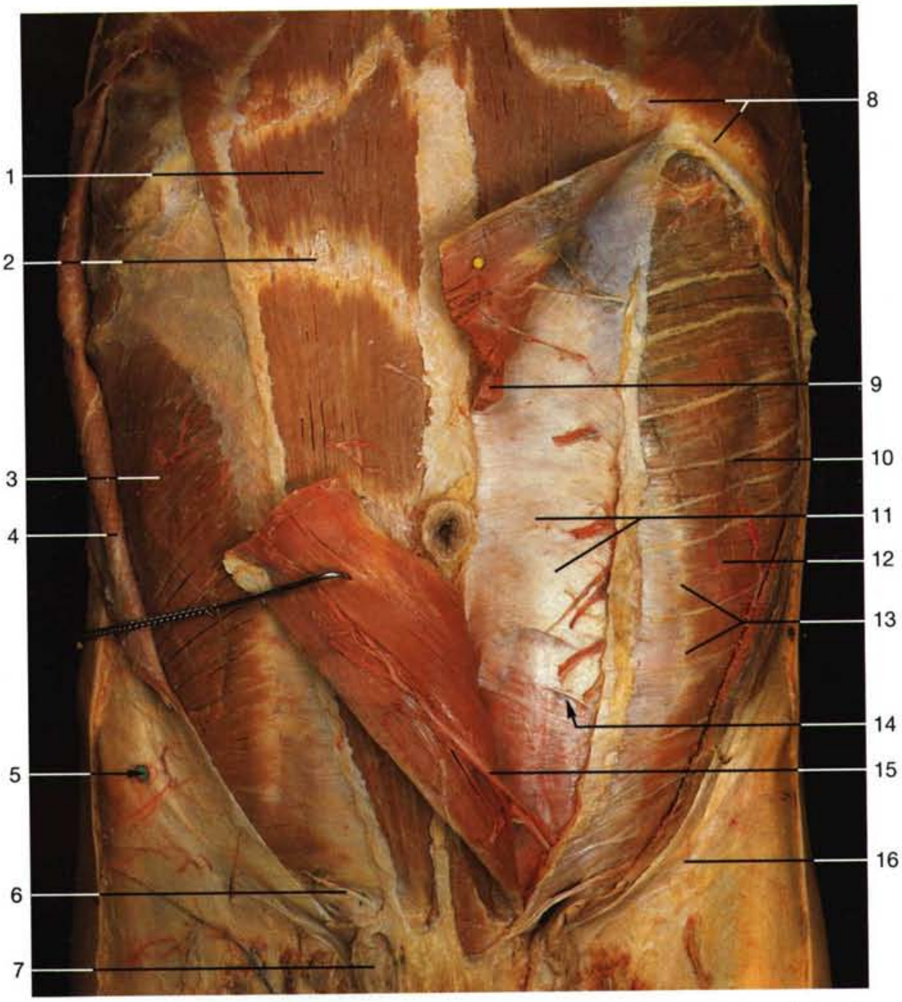

1 **Rectus abdominis muscle**
2 Tendinous intersection
3 **Internal abdominal oblique muscle**
4 External abdominal oblique muscle (reflected)
5 Anterior superior iliac spine
6 Ilioinguinal nerve
7 Spermatic cord
8 Costal margin
9 Superior epigastric artery
10 **Thoracoabdominal** (intercostal) **nerves**
11 Posterior layer of rectus sheath
12 **Transversus abdominis muscle**
13 Semilunar line
14 Arcuate line
15 **Inferior epigastric artery**
16 Inguinal ligament

Abdominal wall with vessels and nerves. The left rectus abdominis muscle has been divided and reflected to display the inferior epigastric vessels. The left internal abdominal oblique muscle has been removed to show the thoracoabdominal nerves.

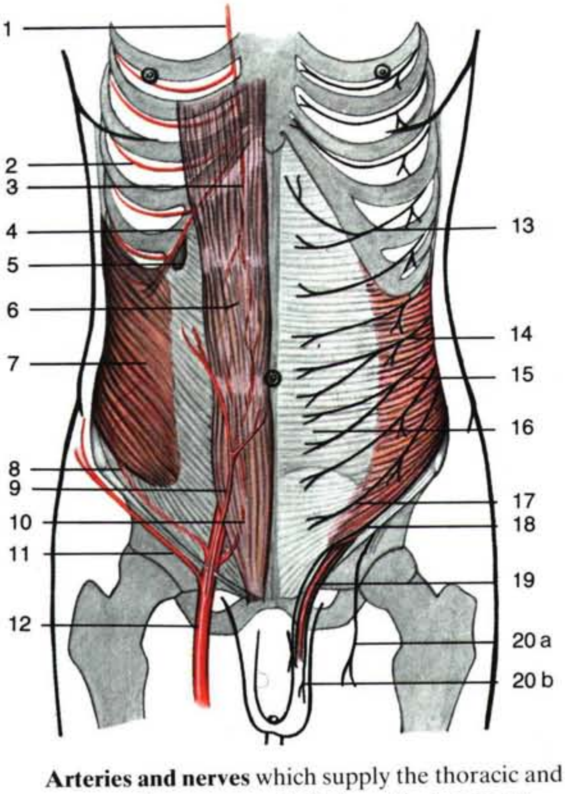

1 Internal thoracic artery
2 Intercostal artery
3 Superior epigastric artery
4 Musculophrenic artery
5 Gallbladder
6 Rectus abdominis muscle
7 External abdominal oblique muscle
8 Deep circumflex iliac artery
9 Superficial epigastric artery
10 Inferior epigastric artery
11 Superficial circumflex iliac artery
12 Femoral artery
13 Intercostal nerve
14 Thoracoabdominal nerve (T_{10})
15 Transversus abdominis muscle
16 Posterior layer of the rectus sheath
17 Iliohypogastric nerve (L_1)
18 Ilioinguinal nerve (L_1)
19 Spermatic cord
20 Genitofemoral nerve (L_1, L_2)
  a Femoral branch
  b Genital branch

Arteries and nerves which supply the thoracic and abdominal wall. Note their segmental arrangement. (Schematic drawing.)

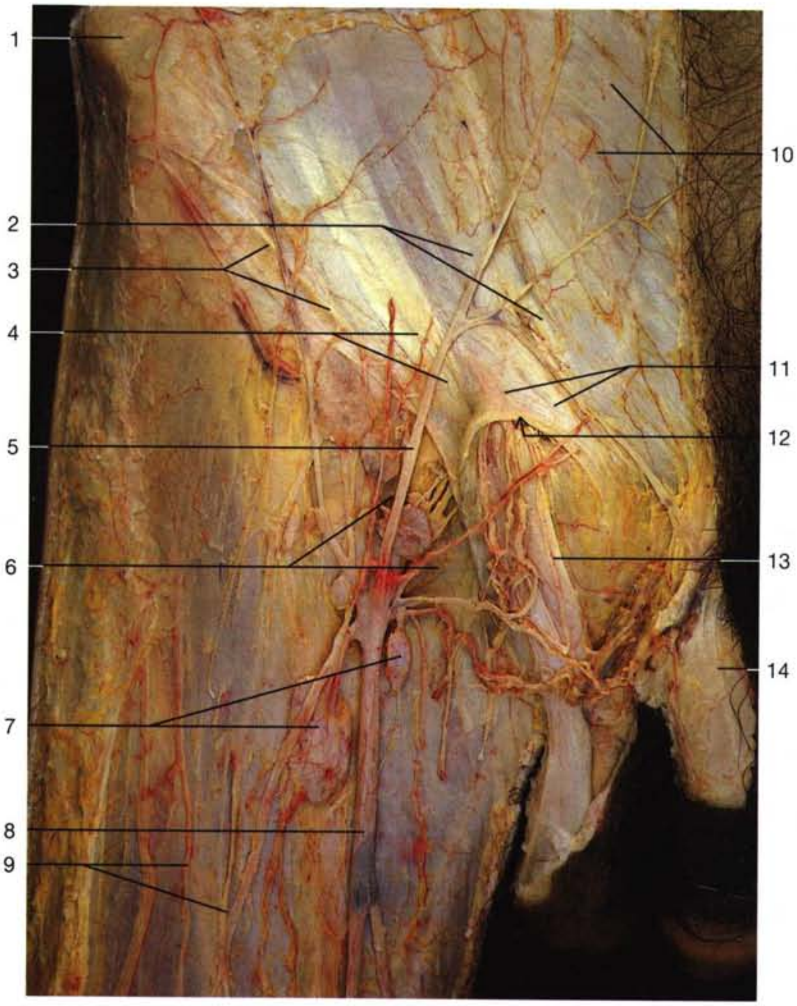

Inguinal canal in the male I. Dissection of superficial layer, right side (anterior aspect).

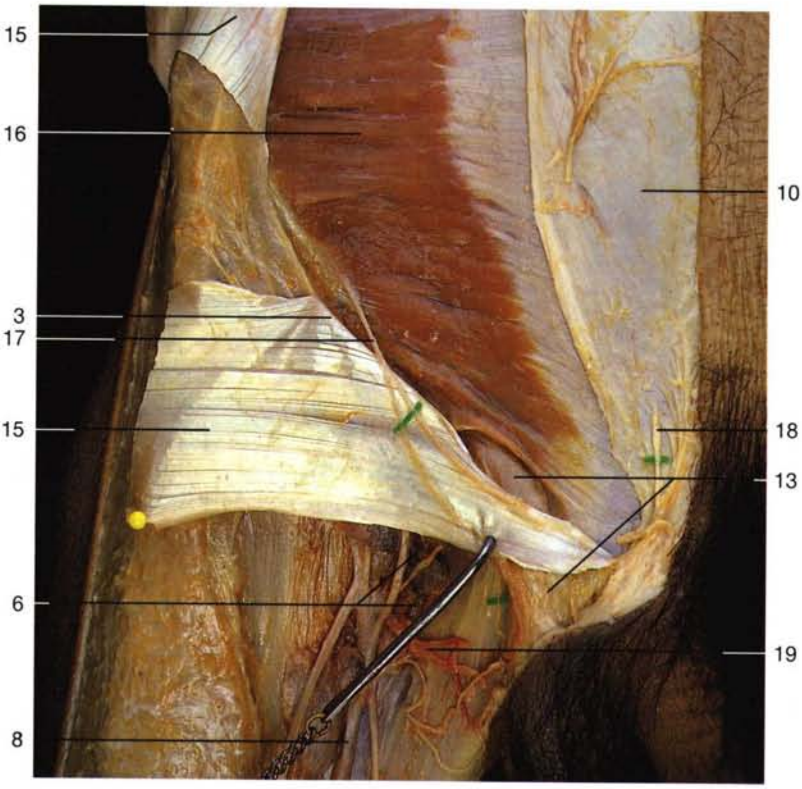

Inguinal canal in the male II, right side (anterior aspect). The external abdominal oblique muscle has been divided to display the inguinal canal.

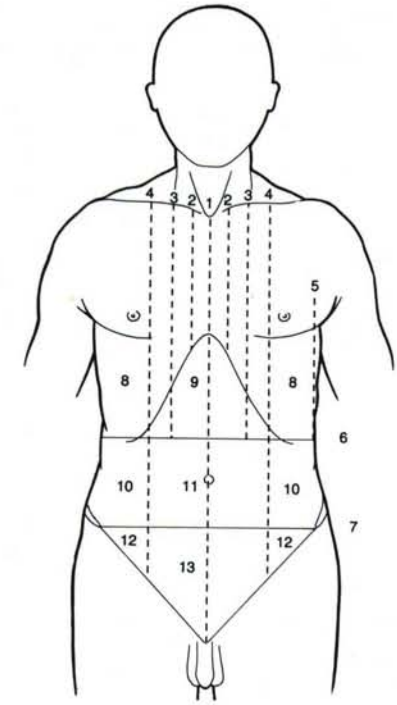

Regions and reference lines
for delineating surface projections.

Reference lines and regions

1 Median line
2 Lateral sternal line
3 Parasternal line
4 Left lateral or midclavicular line
5 Midaxillary line
6 Transpyloric plane
7 Transtubercular plane
8 Hypochondriac region
9 Epigastric region
10 Lumbar region
11 Umbilical region
12 Iliac region
13 Hypogastric region

1 Anterior superior iliac spine
2 Medial crus of inguinal ring
3 Inguinal ligament
4 Lateral crus of inguinal ring
5 Superficial epigastric vein
6 Saphenous opening
7 **Superficial inguinal lymph nodes**
8 **Great saphenous vein**
9 Anterior cutaneous branches of femoral nerve
10 **Anterior layer of rectus sheath**
11 Intercrural fibers
12 **Superficial inguinal ring**
13 Spermatic cord and genital branch of genitofemoral nerve
14 Penis
15 Aponeurosis of external abdominal oblique muscle (divided and reflected)
16 **Internal abdominal oblique muscle**
17 Ilioinguinal nerve
18 Anterior cutaneous branches of iliohypogastric nerve
19 Superficial external pudendal veins

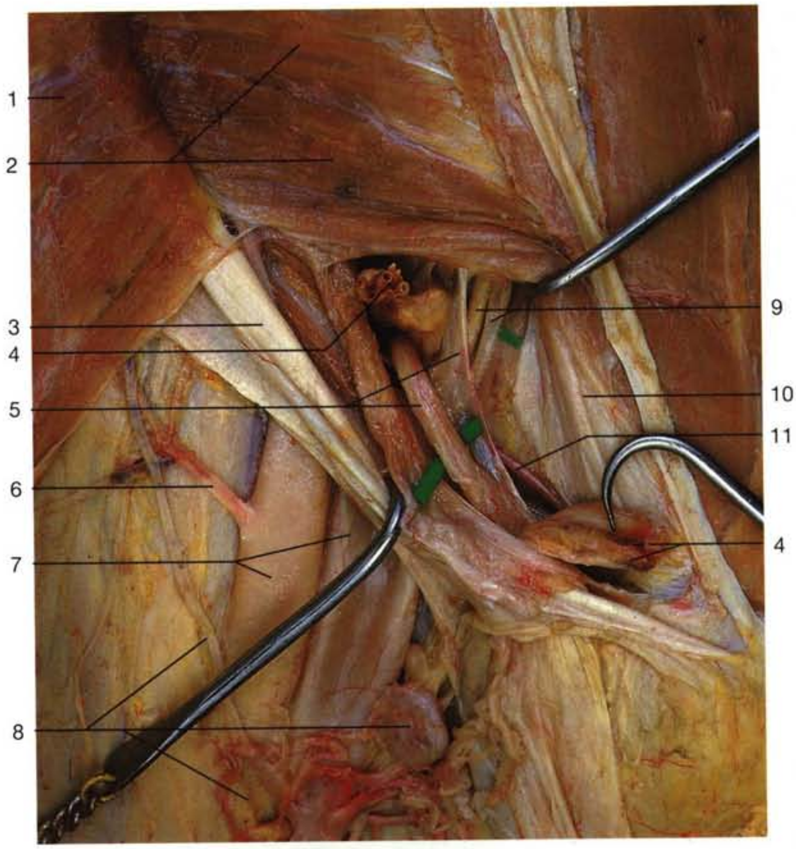

Inguinal canal in the male III. Deep dissection (anterior aspect, right side). Spermatic cord with exception of ductus deferens (probe) has been divided and reflected.

1 Internal abdominal oblique muscle (reflected)
2 Transversus abdominis muscle
3 Inguinal ligament
4 **Spermatic cord** with the exception of the ductus deferens (divided and reflected)
5 **Ductus deferens** and interfoveolar ligament
6 Superficial circumflex iliac artery
7 **Femoral artery and vein**
8 Superficial inguinal lymph nodes and inguinal lymph vessel
9 **Inferior epigastric artery and vein**
10 Falx inguinalis or conjoint tendon (cut)
11 Pubic branch of inferior epigastric artery
12 Superficial inguinal ring
13 Penis
14 External abdominal oblique muscle
15 Anterior superior iliac spine
16 Intercrural fibers
17 Fascia lata and sartorius muscle
18 Saphenous opening and great saphenous vein
19 Deep inguinal ring
20 Skin of scrotum and dartos muscle
21 Cremaster muscle
22 Internal spermatic fascia
23 Ductus deferens
24 Epididymis
25 Peritoneum (blue)
26 Remnant of processus vaginalis
27 Tunica vaginalis testis
28 Rectus abdominis muscle
29 Spermatic cord with ductus deferens covered by external spermatic fascia
30 Anterior layer of rectus sheath
31 Suspensory ligament of penis
32 **Testis and epididymis**
33 Ductus deferens
34 Pampiniform venous plexus and testicular artery
35 Inferior epigastric artery
36 Lateral femoral cutaneous nerve
37 Ilioinguinal nerve
38 **Femoral nerve**
39 Sartorius muscle
40 Deep dorsal vein of penis

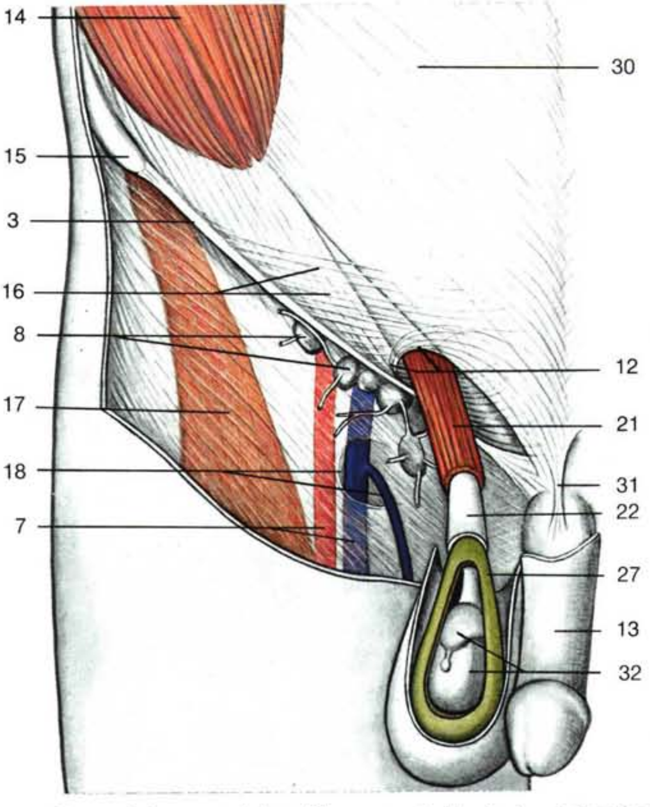

General characteristics of lower part of anterior abdominal wall and inguinal canal. (Schematic drawing.)

Inguinal hernias may either pass through the inguinal canal lateral to the inferior epigastric artery (indirect or lateral inguinal hernias, A and C) or directly penetrate the abdominal wall through the inguinal triangle located medial to the inferior epigastric artery (direct or medial inguinal hernias, B). The lateral hernias can be congenital if the vaginal process remains open (C) or acquired (A) if the hernia develops independently of a patent processus vaginalis.

Femoral hernias generally protrude through the femoral ring below the inguinal ligament. Proper assessment of the site of herniation requires the identification of both the inguinal ligament and the epigastric artery.

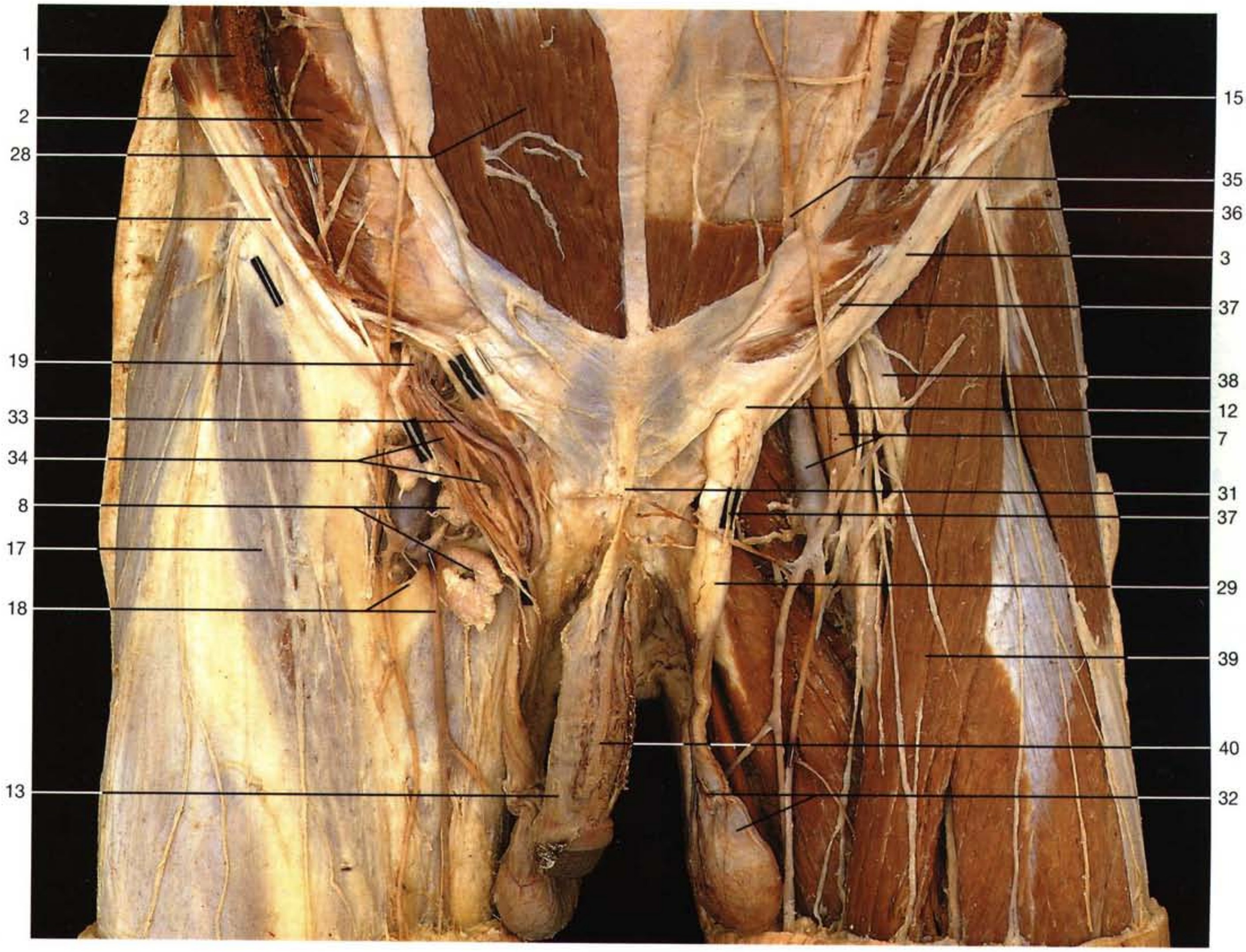
1
2
28
3
19
33
34
8
17
18
13
15
35
36
3
37
38
12
7
31
37
29
39
40
32

Inguinal and femoral region in the male (anterior aspect). On the right, the spermatic cord was dissected to display the ductus deferens and the accompanying vessels and nerves. The fascia lata on the left side has been removed.

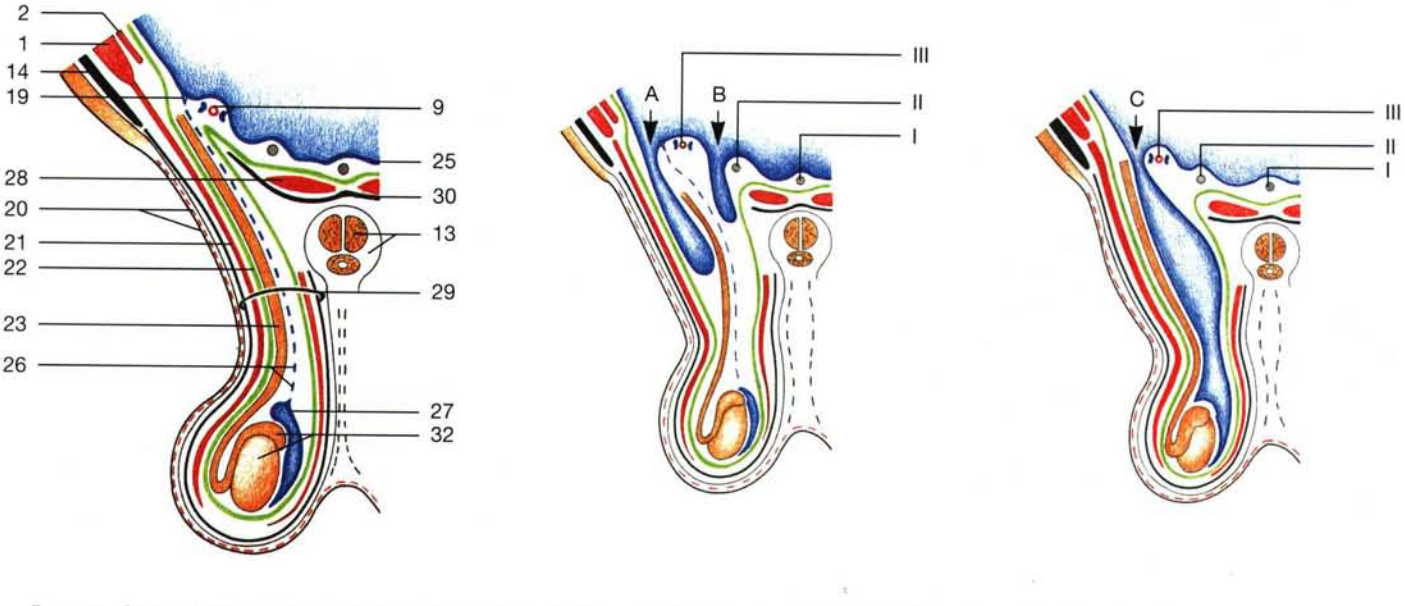
2
1
14
19
28
20
21
22
23
26
9
25
30
13
29
27
32
A
B
III
II
I
C
III
II
I

Layers of spermatic cord and types of hernias. Left: Normal situation; middle: Location of acquired inguinal hernias; A = indirect; B = direct inguinal hernia. Right: congenital indirect inguinal hernia (C); the vaginal process remained open.

I = Median umbilical fold, containing urachus chord.

II = Medial umbilical fold with remnants of umbilical artery and vein.

III = Lateral umbilical fold with inferior epigastric artery and vein.

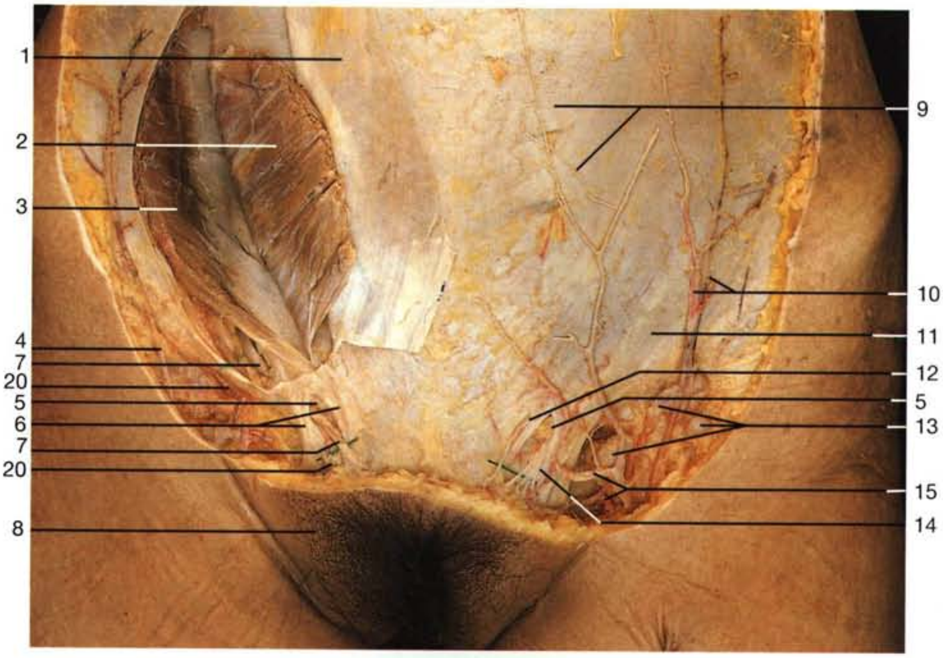

Inguinal region in the female (anterior aspect). Left side: superficial layer; right side: external and internal abdominal oblique muscle divided and reflected.

1 Aponeurosis of external abdominal oblique muscle
2 Internal abdominal oblique muscle (divided and reflected)
3 Transversus abdominis muscle
4 Superficial circumflex iliac artery and vein
5 Superficial inguinal ring with fat pad
6 Medial and lateral crural fibers
7 **Round ligament** (ligamentum teres uteri)
8 Labium majus pudendi
9 Anterior layer of rectus sheath
10 Superficial epigastric artery and vein
11 **Inguinal ligament**
12 Cutaneous branch of ilioinguinal nerve
13 Superficial inguinal lymph nodes
14 Entrance of round ligament into the labium majus
15 External pudendal artery and vein
16 Position of deep inguinal ring
17 **Ilioinguinal nerve**
18 Internal abdominal oblique muscle
19 Pubic branch of inferior epigastric artery
20 Genital branch of genitofemoral nerve
21 Fat pad of inguinal canal
22 Ilioinguinal nerve
23 Sheath of round ligament (inguinal canal)
24 Transversalis fascia

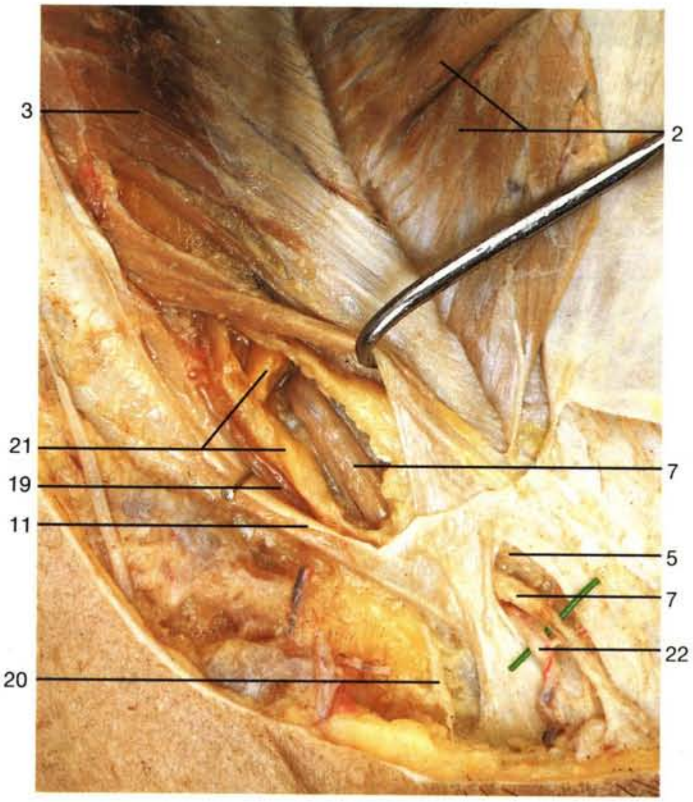

Inguinal canal of the female I (anterior aspect, right side). The external abdominal oblique muscle has been divided and reflected, to display the ilioinguinal nerve and the round ligament.

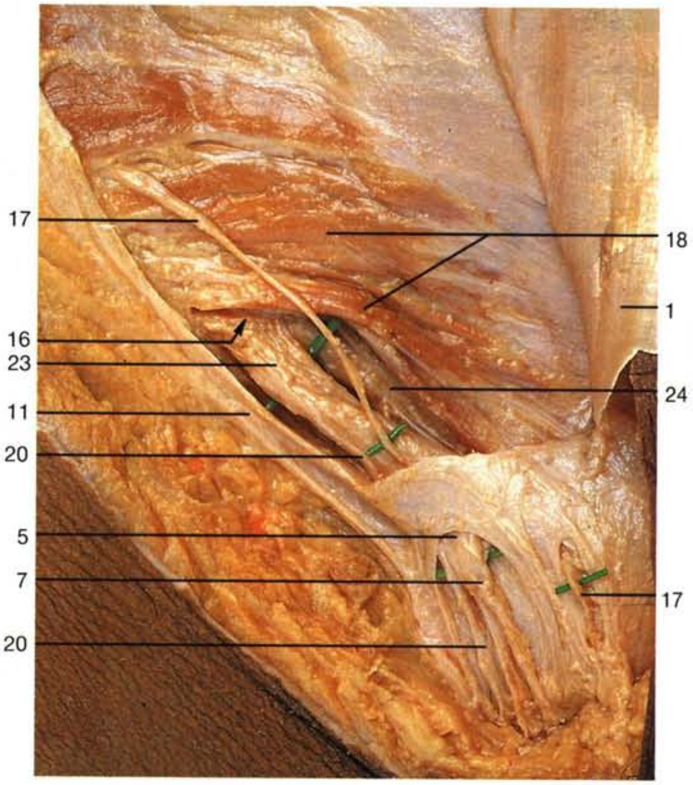

Inguinal canal of the female II (anterior aspect, right side). The external and internal abdominal oblique muscle have been divided and reflected to show the content of the inguinal canal.

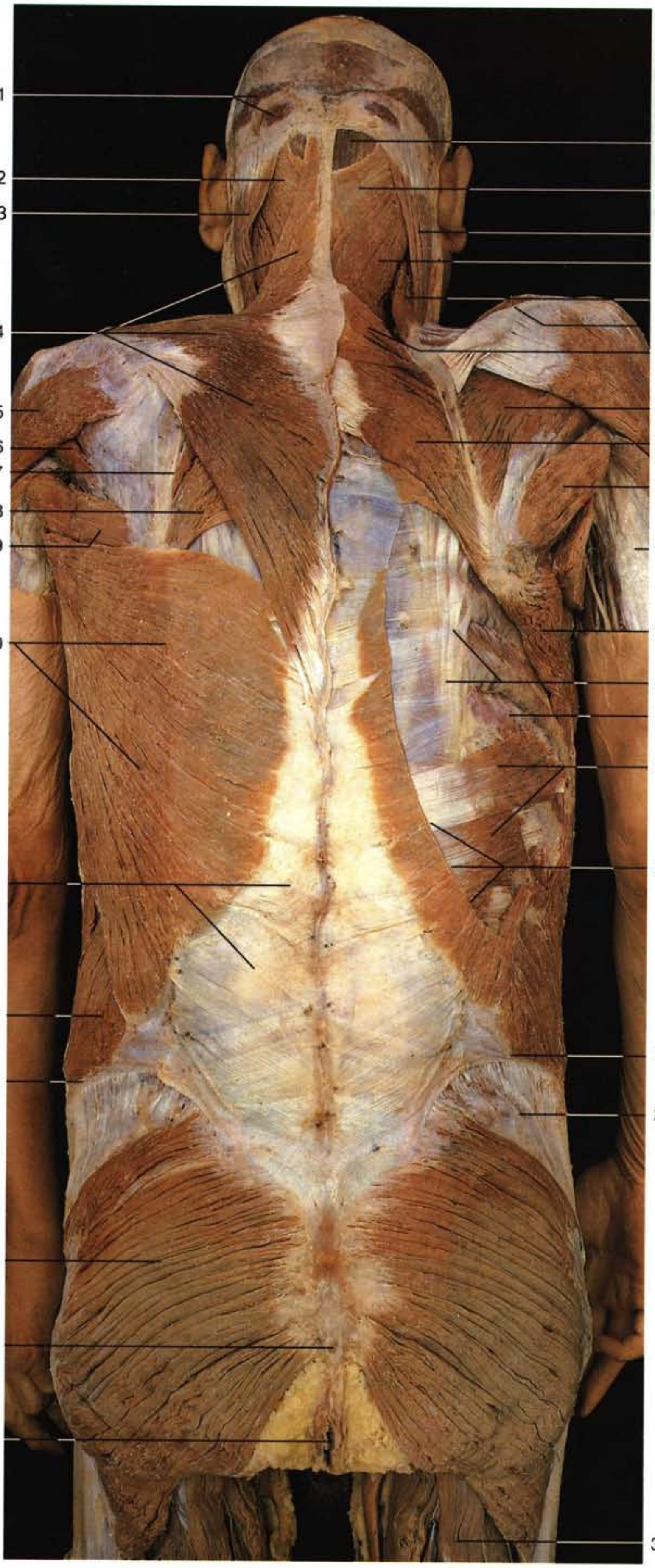

1 Occipital belly of occipitofrontalis muscle
2 Splenius capitis muscle
3 Sternocleidomastoid muscle
4 **Trapezius muscle**
5 Deltoid muscle
6 Teres minor muscle
7 Medial margin of scapula
8 **Rhomboid major muscle**
9 Teres major muscle
10 **Latissimus dorsi muscle**
11 Thoracolumbar fascia
12 External abdominal oblique muscle
13 Iliac crest
14 **Gluteus maximus muscle**
15 Position of last coccygeal vertebra
16 Anus
17 Semispinalis capitis muscle
18 Splenius cervicis muscle
19 Levator scapulae muscle
20 Spine of scapula
21 Rhomboideus minor muscle
22 Infraspinatus muscle
23 Triceps brachii muscle
24 Serratus anterior muscle
25 **Iliocostalis muscle**
26 External intercostal muscle
27 **Serratus posterior inferior muscle**
28 Latissimus dorsi muscle (cut edge)
29 Fascia over gluteus medius muscle
30 Long head of biceps femoris muscle

Muscles of the back I. Superficial layer on the left, deeper layer on the right. Right latissimus dorsi and trapezius muscles are removed.

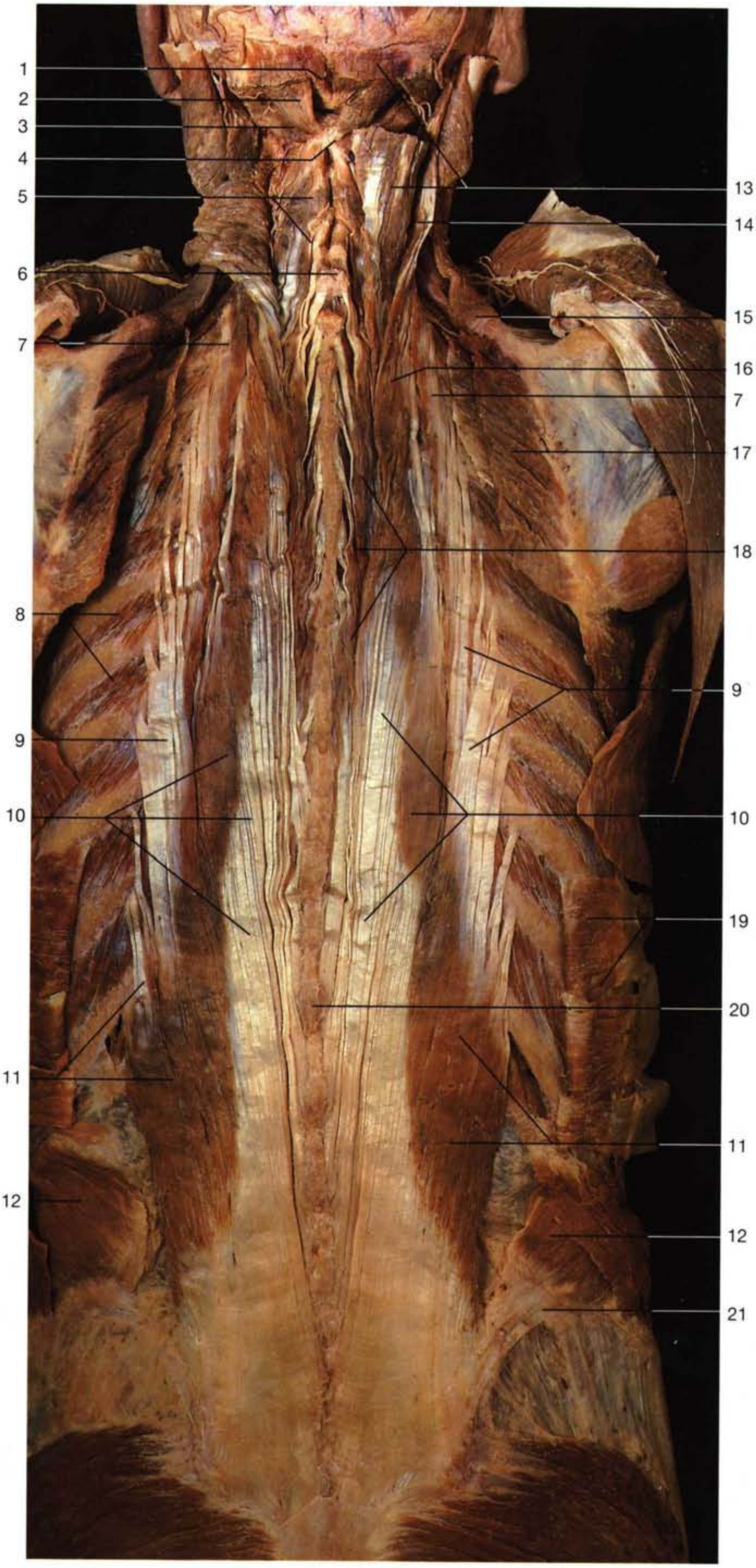

Muscles of the back II. Dissection of deep muscles. Erector spinae muscle.

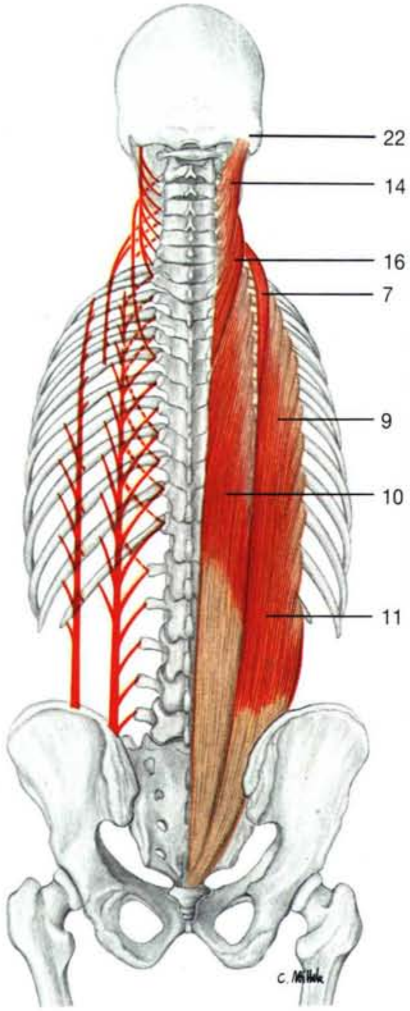

Origin and insertion of iliocostalis and longissimus muscles. (Schematic drawing.)

1 Rectus capitis posterior minor muscle
2 Rectus capitis posterior major muscle
3 Obliquus capitis inferior muscle
4 Spinous process of axis
5 Semispinalis cervicis muscle
6 Spinous process of seventh vertebra
7 **Iliocostalis cervicis muscle**
8 External intercostal muscles
9 **Iliocostalis thoracis muscle**
10 **Longissimus thoracis muscle**
11 **Iliocostalis lumborum muscle**
12 Internal abdominal oblique muscle
13 Semispinalis capitis muscle (divided)
14 **Longissimus capitis muscle**
15 Levator scapulae muscle
16 **Longissimus cervicis muscle**
17 Rhomboid major muscle
18 **Spinalis thoracis muscle**
19 Serratus posterior inferior muscle (reflected)
20 Spinous process of second lumbar vertebra
21 Iliac crest
22 Mastoid process

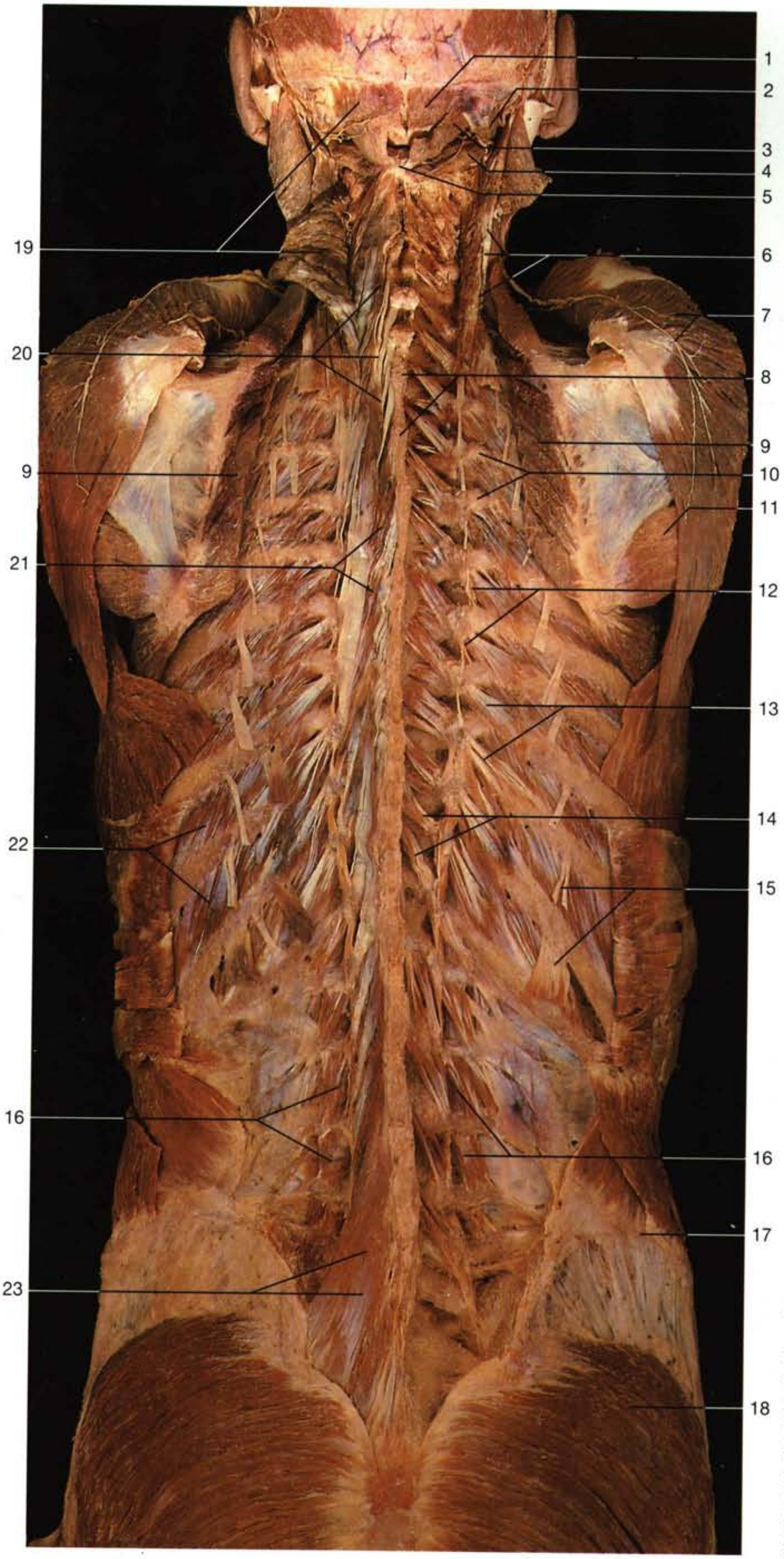

Muscles of the back III. Transversospinal muscles, deepest layer on the right, where all parts of semispinalis and multifidus muscles have been removed.

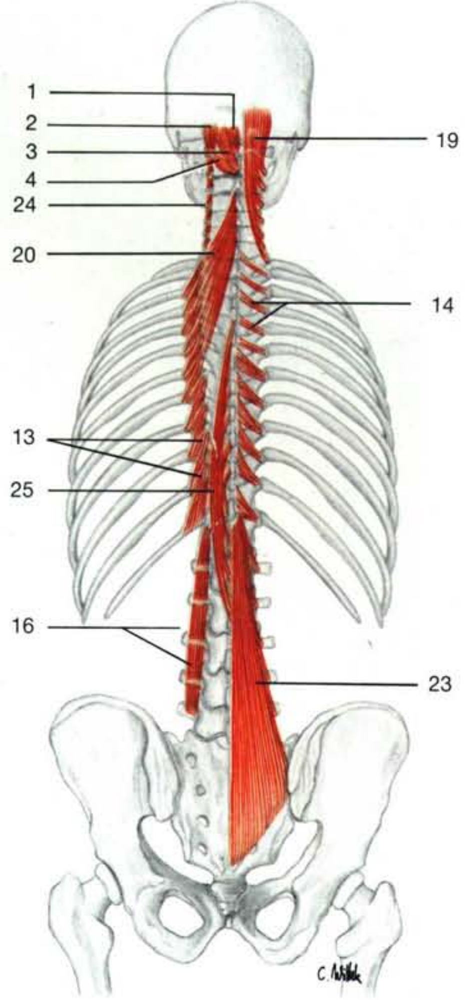

Medial column of intrinsic muscles of the back. Transversospinal and intertransversal system. (Schematic drawing.)

1 Rectus capitis posterior minor muscle
2 Obliquus capitis superior muscle
3 Rectus capitis posterior major muscle
4 Obliquus capitis inferior muscle
5 Spinous process of axis
6 **Longissimus capitis muscle**
7 Trapezius muscle (reflected) and accessory nerve (n. XI)
8 Spinous processes
9 Rhomboid major muscle
10 Transverse processes of thoracic vertebrae
11 Teres major muscle
12 Intertransverse ligaments
13 **Levatores costarum muscles**
14 **Rotatores muscles**
15 Tendons of iliocostalis muscle
16 Intertransversarii lumborum muscle (lateral)
17 Iliac crest
18 Gluteus maximus muscle
19 Semispinalis capitis muscle
20 **Semispinalis cervicis muscle**
21 Semispinalis thoracis muscle
22 **External intercostal muscles**
23 **Multifidus muscle**
24 Intertransversarii posterior cervicis muscle
25 Spinalis thoracis muscle

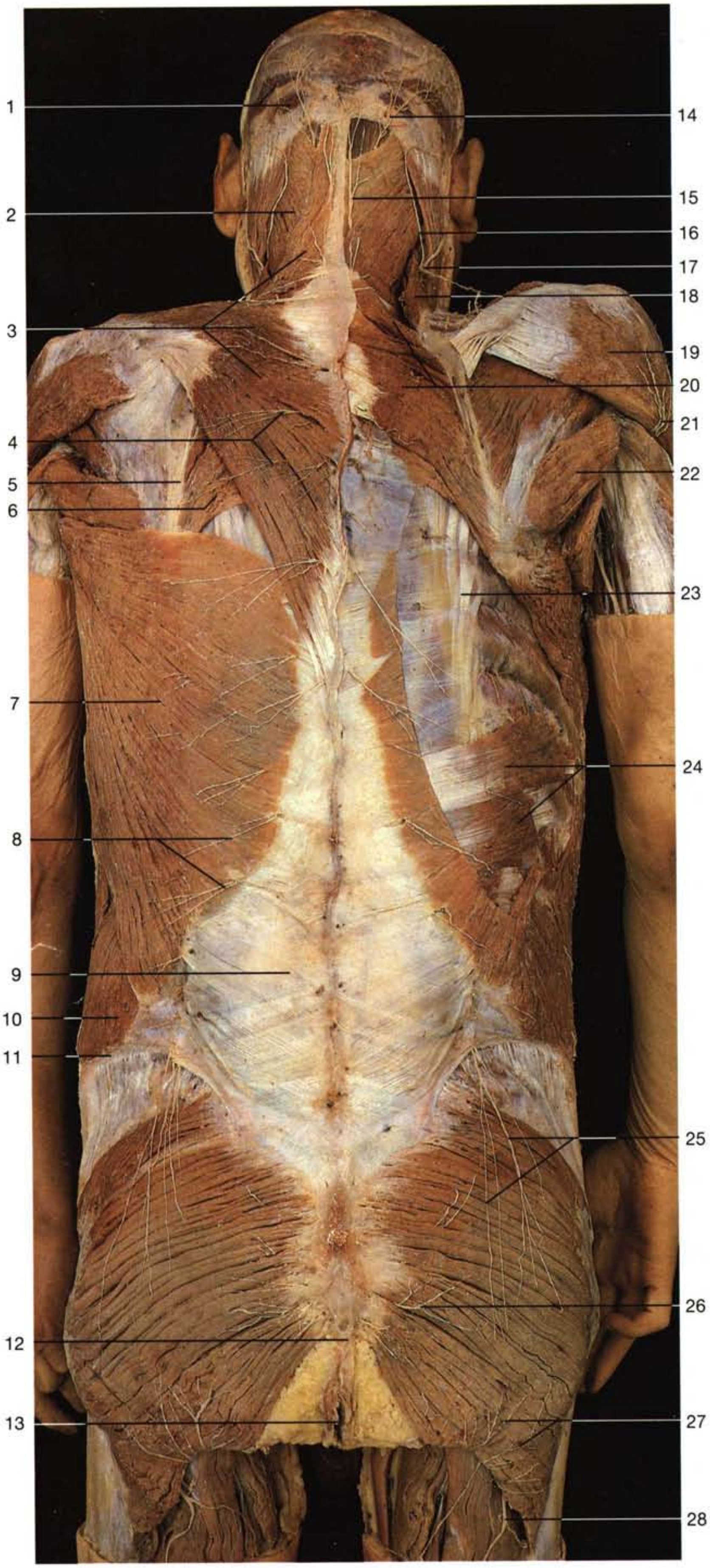

Innervation of the back I. Superficial (left) and deeper (right) layers. Right trapezius and latissimus dorsi muscles removed.

1 Occipital belly of occipitofrontalis muscle
2 Splenius capitis muscle
3 **Trapezius muscle**
4 **Medial cutaneous branches of dorsal rami of spinal nerves**
5 Medial margin of scapula
6 Rhomboid major muscle
7 **Latissimus dorsi muscle**
8 **Lateral cutaneous branches of dorsal rami of spinal nerves**
9 Thoracolumbar fascia
10 External abdominal oblique muscle
11 Iliac crest
12 Last coccygeal vertebra
13 Anus
14 Greater occipital nerve
15 Third occipital nerve
16 Lesser occipital nerve
17 Cutaneous branches of cervical plexus
18 Levator scapulae muscle
19 Deltoid muscle
20 **Rhomboid major and minor muscles**
21 Upper lateral cutaneous nerve of arm (branch of axillary nerve)
22 **Teres major muscle**
23 Iliocostalis thoracis muscle
24 Serratus posterior inferior muscle
25 Superior cluneal nerves
26 Middle cluneal nerves
27 Inferior cluneal nerves
28 Posterior femoral cutaneous nerve

▷

To page 213:

1 **Trapezius muscle**
2 Infraspinatus muscle
3 Left **latissimus dorsi muscle**
4 Thoracolumbar fascia
5 Splenius cervicis muscle
6 Serratus posterior superior muscle
7 **Medial branches of dorsal rami of thoracic spinal nerves**
8 **Lateral branches of dorsal rami of thoracic spinal nerves**
9 Iliocostalis muscle
10 Serratus posterior inferior muscle
11 Latissimus dorsi muscle (reflected)

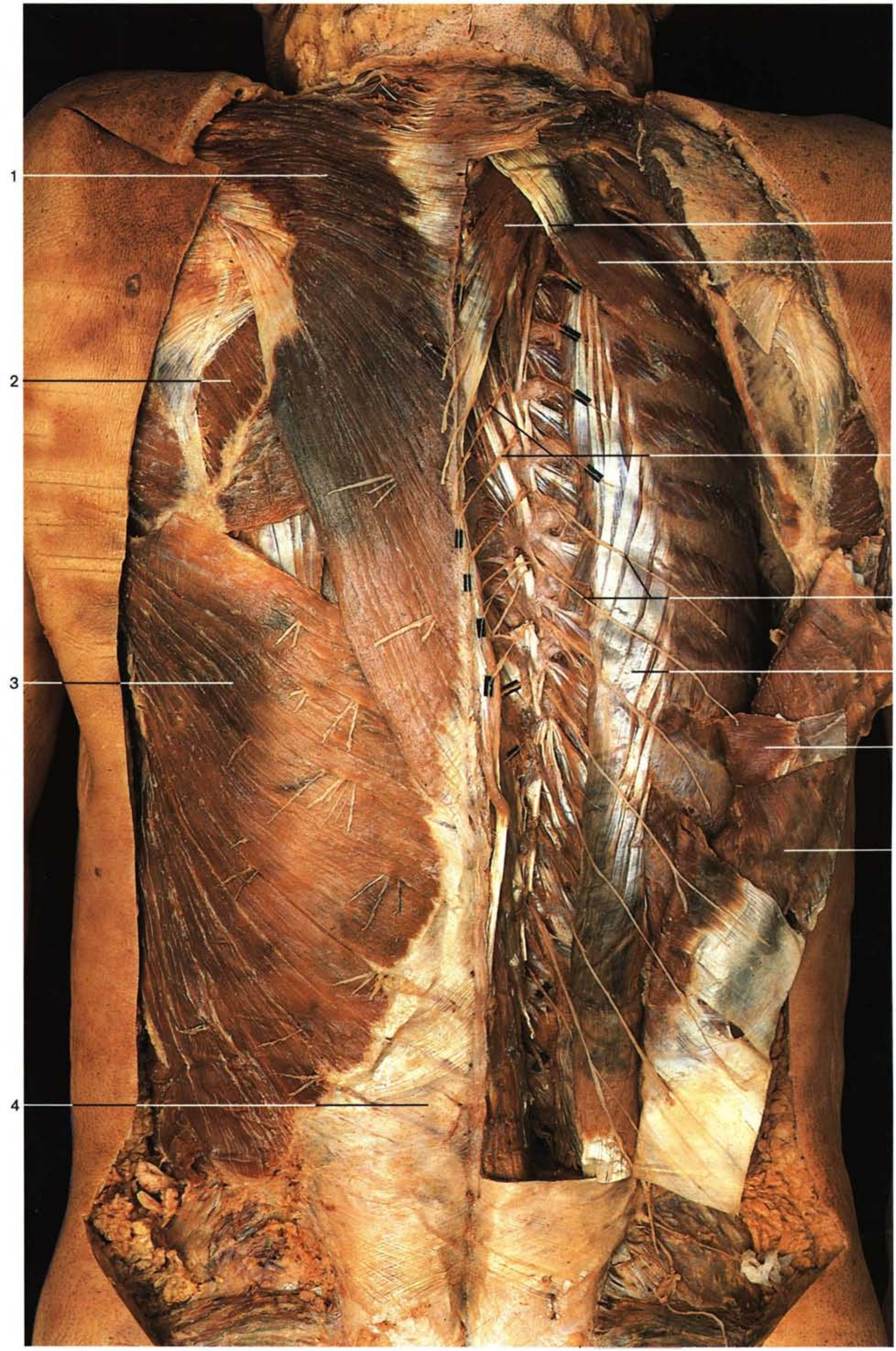

Innervation of the back II. Dissection of the dorsal branches of spinal nerves. On the right, longissimus thoracis muscle has been removed and iliocostalis muscle laterally reflected.

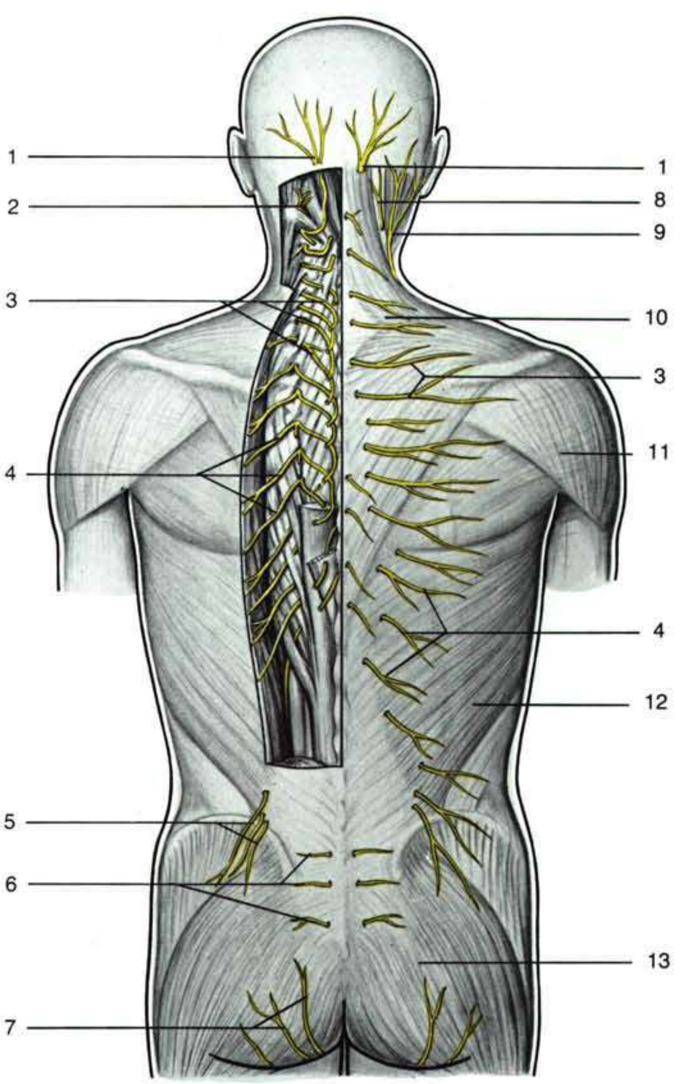

General characteristics of the innervation of the back. Distribution of dorsal branches of spinal nerves. Note the segmental arrangement of the innervation of the dorsal part of the trunk. (Schematic drawing.)

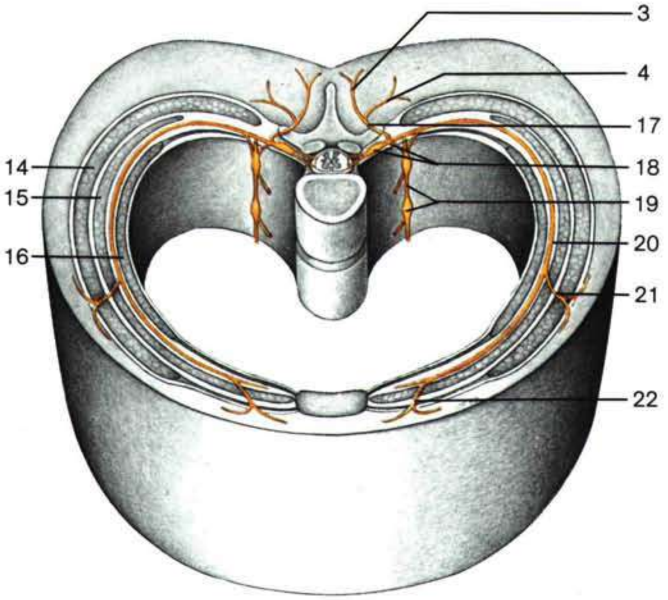

Position and branches of spinal nerves in one segment of thoracic wall. (Schematic drawing.)

1 Greater occipital nerve (C_2)
2 Suboccipital nerve (C_1)
3 **Medial branches of dorsal rami of spinal nerves**
4 **Lateral branches of dorsal rami of spinal nerves**
5 Superior cluneal nerves (L_1–L_3)
6 Middle cluneal nerves (S_1–S_3)
7 Inferior cluneal nerves (derived from branches of the sacral plexus, ventral rami)
8 Lesser occipital nerve
9 Great auricular nerve
10 Trapezius muscle
11 Deltoid muscle
12 Latissimus dorsi muscle
13 Gluteus maximus muscle
14 External intercostal muscle
15 Internal intercostal muscle
16 Innermost intercostal muscle
17 Dorsal ramus of spinal nerve
18 **Spinal nerve** and **spinal ganglion**
19 Sympathetic trunk with ganglion
20 Intercostal nerve
21 Lateral cutaneous branch } of **intercostal nerve**
22 Anterior cutaneous branch }

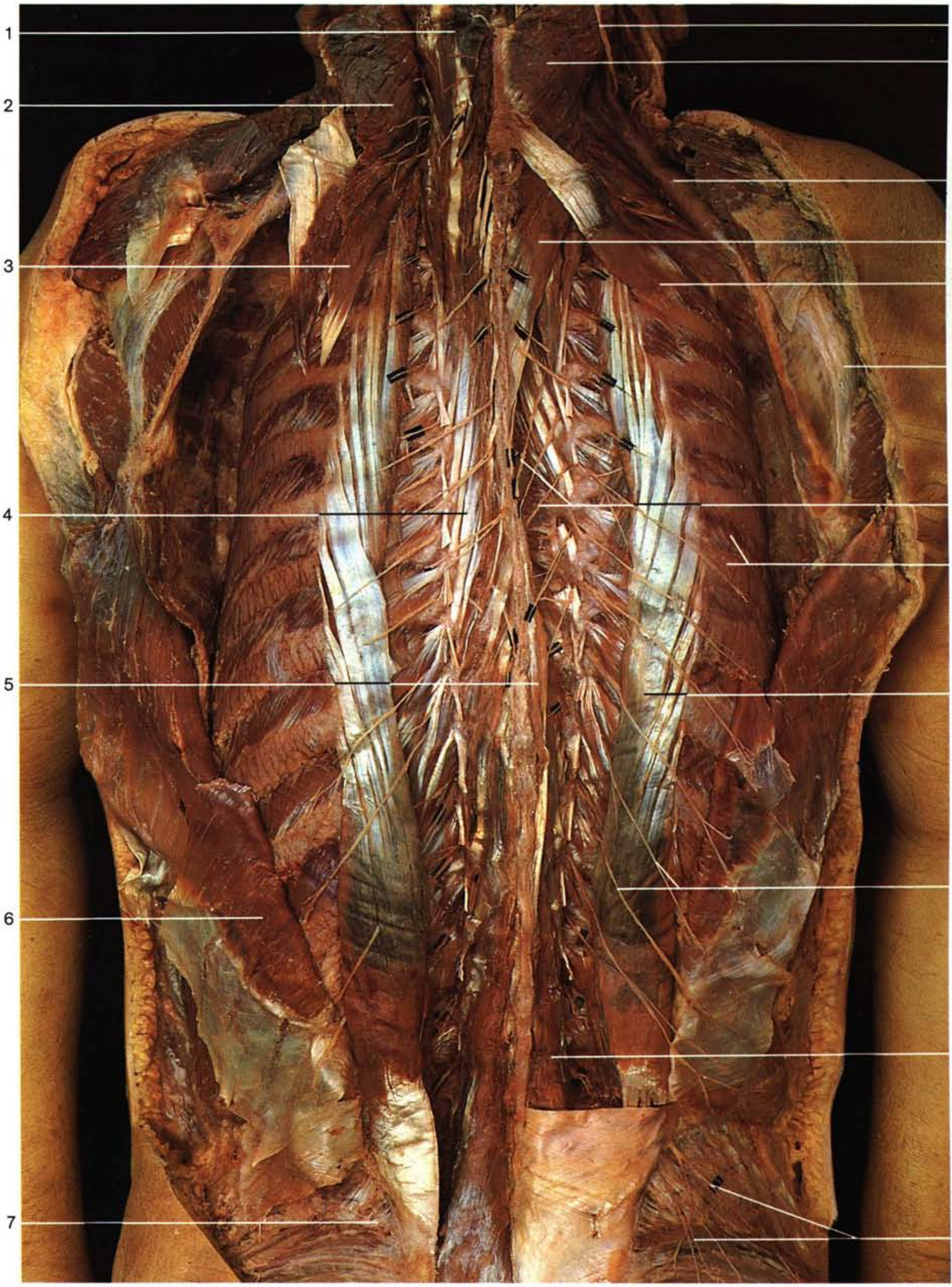

Innervation of the back III. Deeper layer (dorsal aspect).

1	Semispinalis capitis muscle	7	Iliac crest	14	**Medial branches of dorsal rami of spinal nerves**
2	Left splenius capitis muscle (cut and reflected)	8	Lesser occipital nerve	15	Rib and external intercostal muscle
3	Left splenius cervicis muscle (cut and reflected)	9	Splenius capitis muscle	16	Iliocostalis thoracis muscle
4	Semispinalis thoracis muscle	10	Levator scapulae muscle	17	**Lateral branches of dorsal rami of spinal nerves**
5	**Spinalis thoracis muscle**	11	**Splenius cervicis muscle**	18	Multifidus muscle
6	Latissimus dorsi muscle (reflected)	12	Serratus posterior superior muscle	19	Superior cluneal nerves
		13	Scapula		

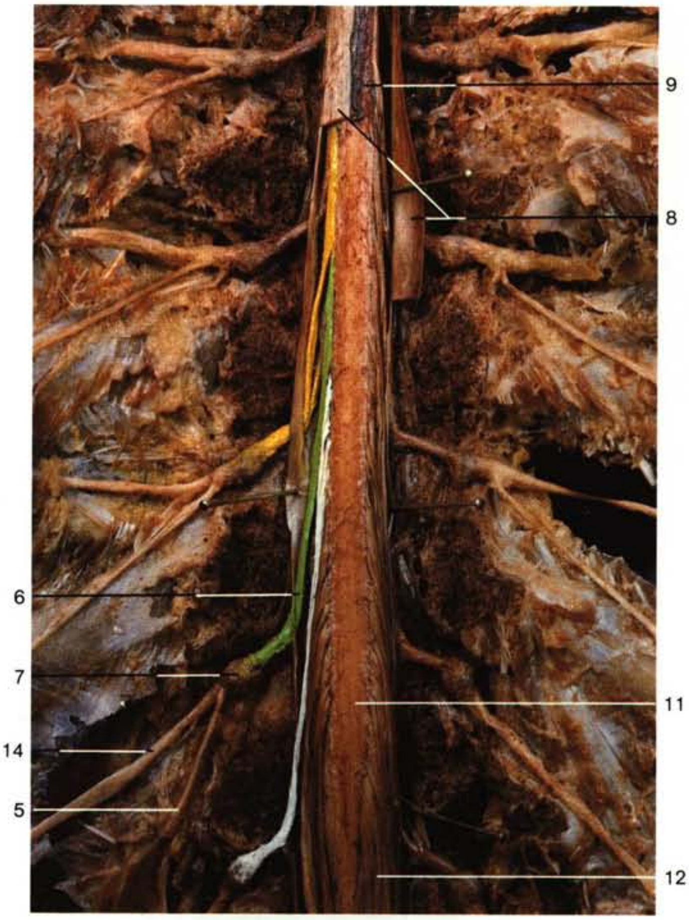

Lumbar portion of spinal cord. Note the relation between the nervous and muscular segments.

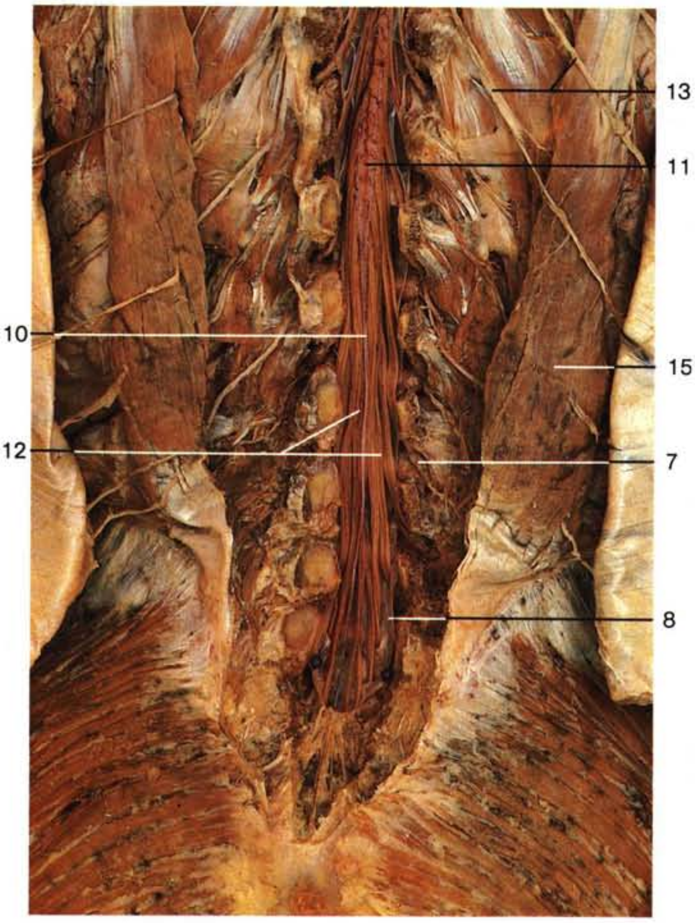

Terminal part of spinal cord. Dura removed.

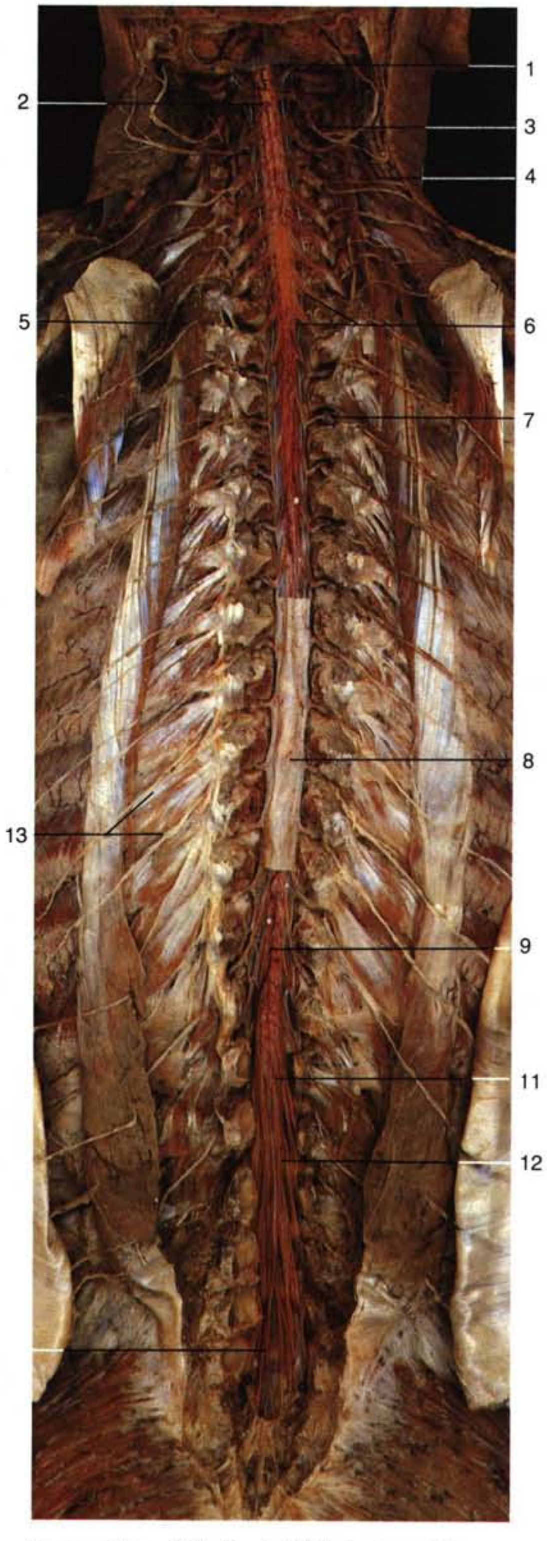

Innervation of the back IV. Spinal cord in the vertebral canal (opened). Longissimus dorsi muscle has been removed and iliocostal muscle reflected.

1 Cerebellomedullary cistern and cerebellum
2 Medulla oblongata
3 Greater occipital nerve (C_2)
4 Third cervical nerve (C_3)
5 Dorsal primary ramus
6 Dorsal roots
7 Spinal ganglion
8 Spinal dura mater
9 Spinal arachnoid mater
10 Filum terminale
11 Conus medullaris
12 Cauda equina
13 Lateral branches of dorsal rami of spinal nerves
14 Ventral ramus of spinal nerve (intercostal nerve)
15 Iliocostalis muscle

1 Arch of vertebra (divided)
2 Spinal nerve with meningeal coverings
3 Dorsal roots of thoracic spinal nerves
4 Spinal cord (thoracic portion)
5 Spinal ganglia with meningeal coverings
6 Pia mater with blood vessels
7 Dura mater (opened)
8 Denticulate ligament
9 Lateral branch of dorsal ramus
10 Dorsal ramus of spinal nerve (dividing into a medial and lateral branch)
11 Medial branch of dorsal ramus of spinal nerve
12 Spinal dura mater
13 Spinal nerves of sacral segments
14 Filum terminale

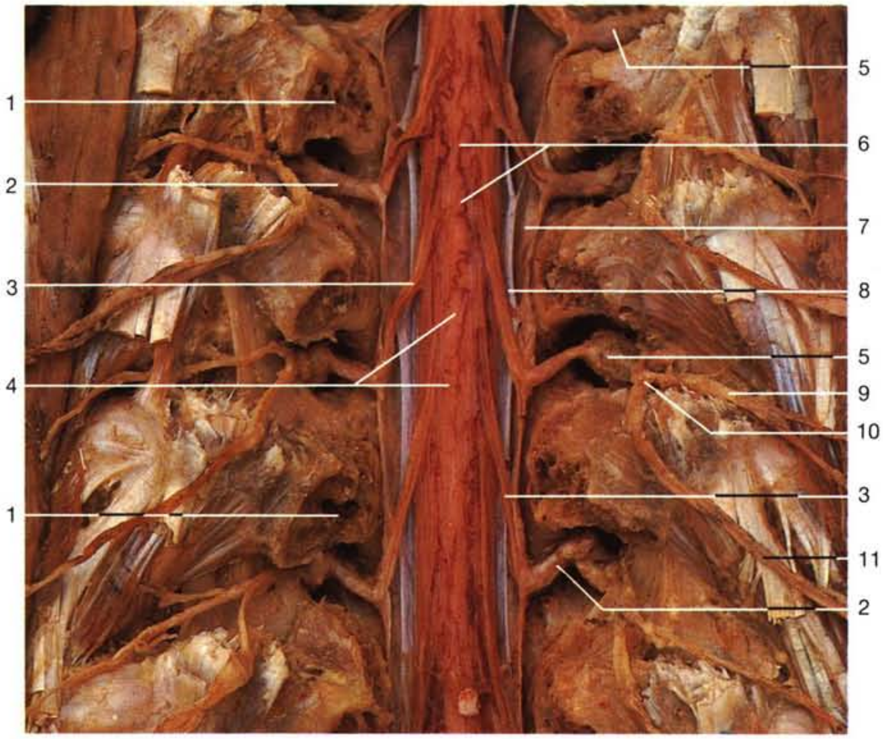

Thoracic portion of spinal cord (dorsal aspect). Vertebral canal and dura mater opened.

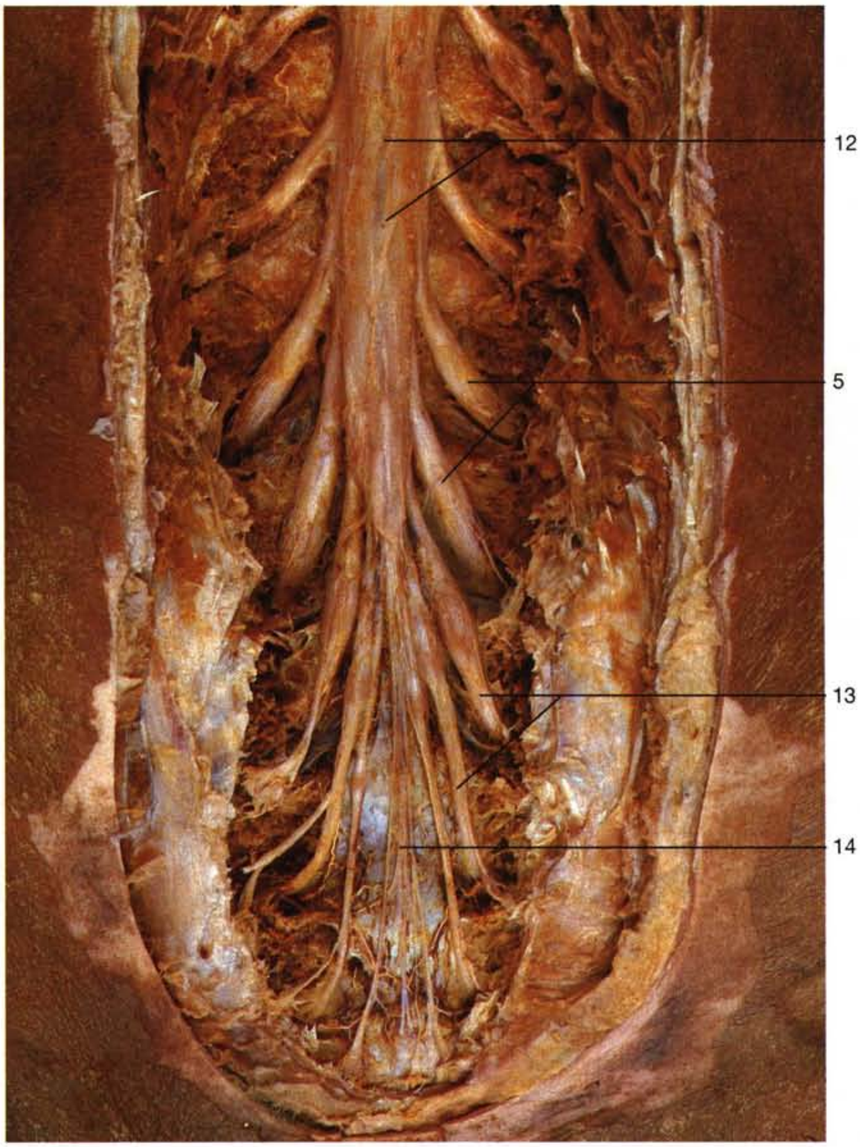

Terminal part of spinal cord with dura mater (dorsal aspect). Dorsal part of sacrum removed.

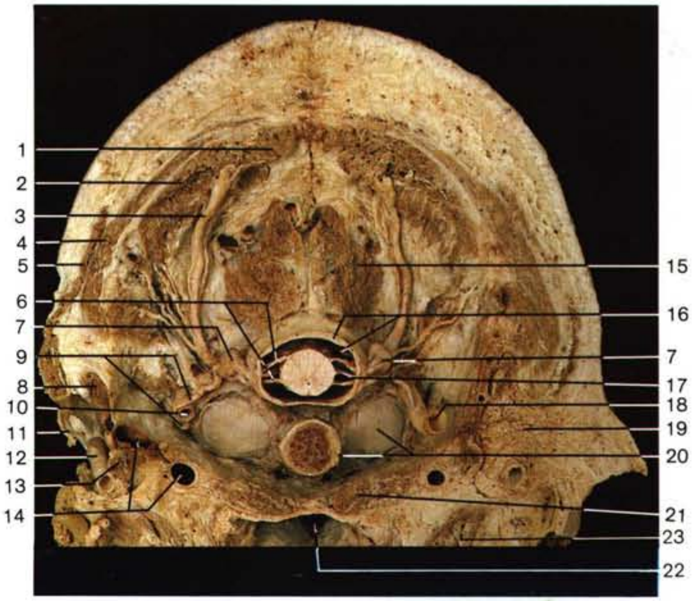

Horizontal section of the neck. Dissection of the second cervical spinal nerve. Posterior surface at top of figure.

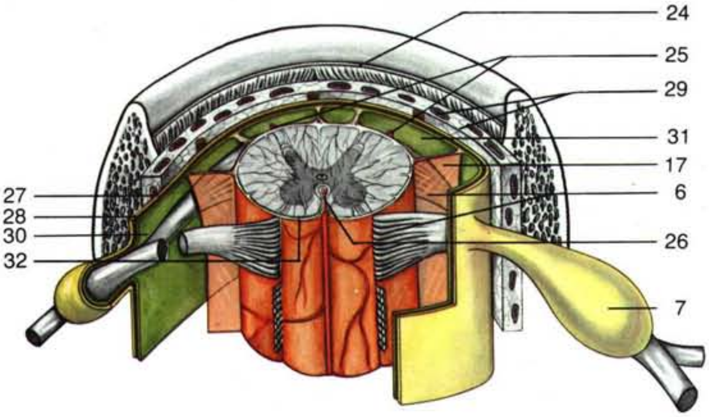

Meningeal coverings of the spinal cord (anterior aspect). (Schematic drawing.)

1 Trapezius muscle
2 Semispinalis capitis muscle
3 **Dorsal ramus of spinal nerve**
4 Sternocleidomastoid muscle
5 Platysma muscle
6 **Dorsal and ventral roots of spinal nerves**
7 **Spinal ganglion**
8 Posterior belly of digastric muscle
9 **Ventral ramus of spinal nerve**
10 Vertebral artery
11 Great auricular nerve
12 Superficial temporal artery
13 Styloid process
14 Internal jugular vein and internal carotid artery
15 Rectus capitis posterior major muscle
16 Dura mater and subarachnoid space
17 Denticulate ligament
18 Vertebral artery
19 Parotid gland
20 Dens of axis (divided) and inferior articular facet of atlas
21 Longus capitis muscle
22 Pharyngeal cavity
23 Medial pterygoid muscle
24 Periosteum of vertebral canal
25 Posterior spinal arteries
26 Anterior spinal artery

Meningeal coverings

27 Dura mater
28 Subdural space
29 Extradural or epidural space with venous plexus and fatty tissue
30 Arachnoid (green)
31 Subarachnoid space
32 Pia mater (pink)
33 Nucleus pulposus
34 Crus of diaphragm
35 Intervertebral disc
36 Body of first lumbar vertebra
37 Spinal cord
38 Conus medullaris
39 Cauda equina
40 Filum terminale
41 Spinous process

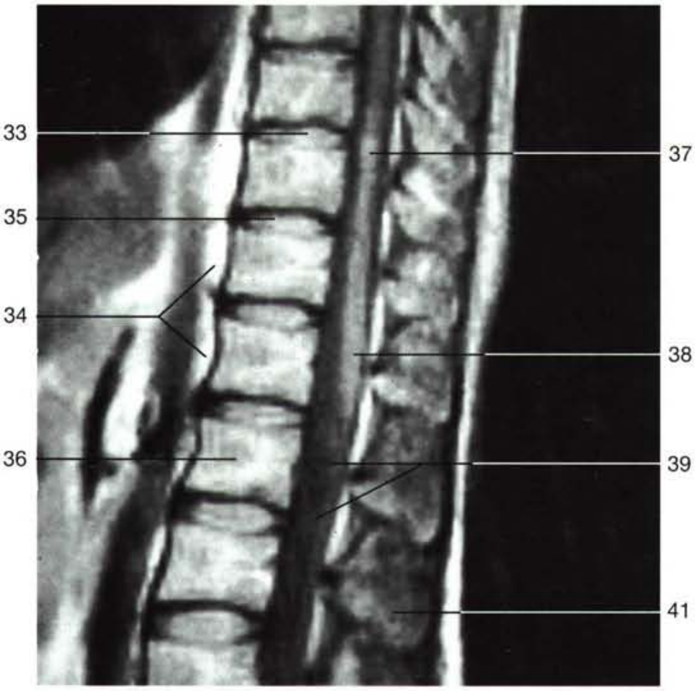

Sagittal section through the vertebral canal, T_9–L_2. (MR-Scan.)

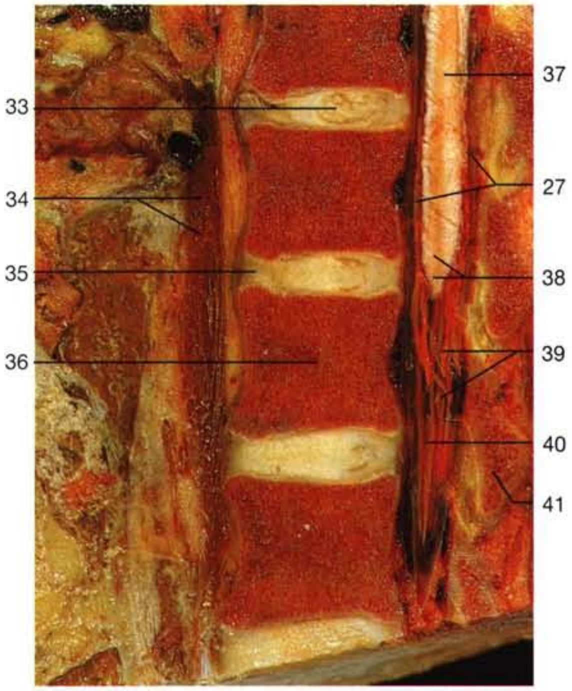

Sagittal section through vertebral canal, T_{12}–L_2. Notice red bone marrow (unfixed).

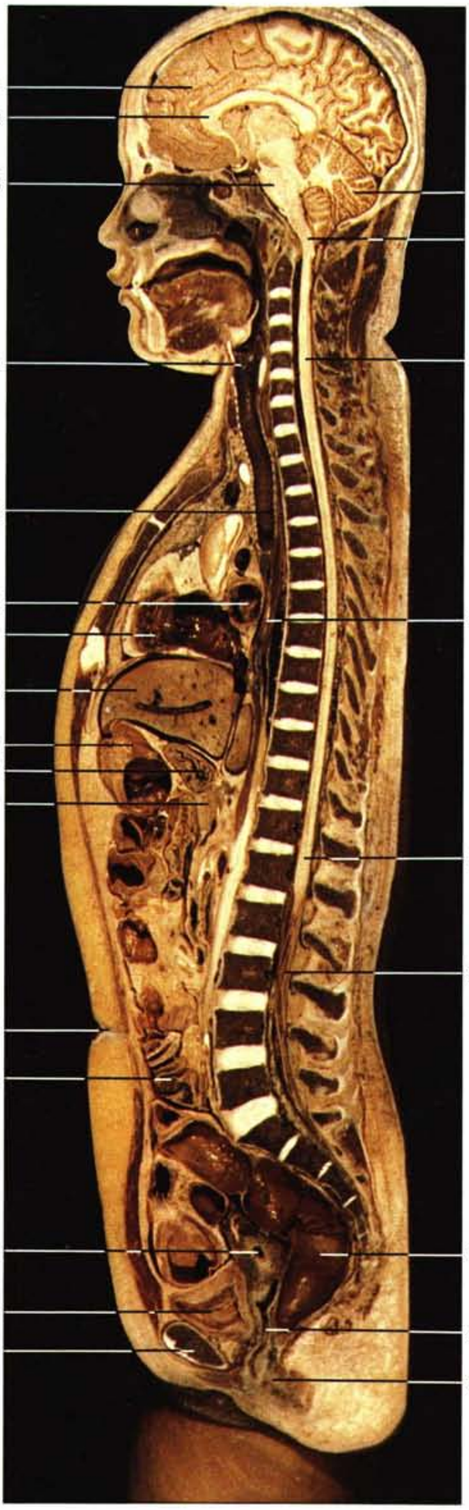
1
2
3
18
19
4
20
5
6
8
7
9
12
10
11
21
22
13
14
15
23
16
24
17
25

Median section of the head and trunk in the adult (female). The conus medullaris of the spinal cord is located at the level of L$_1$.

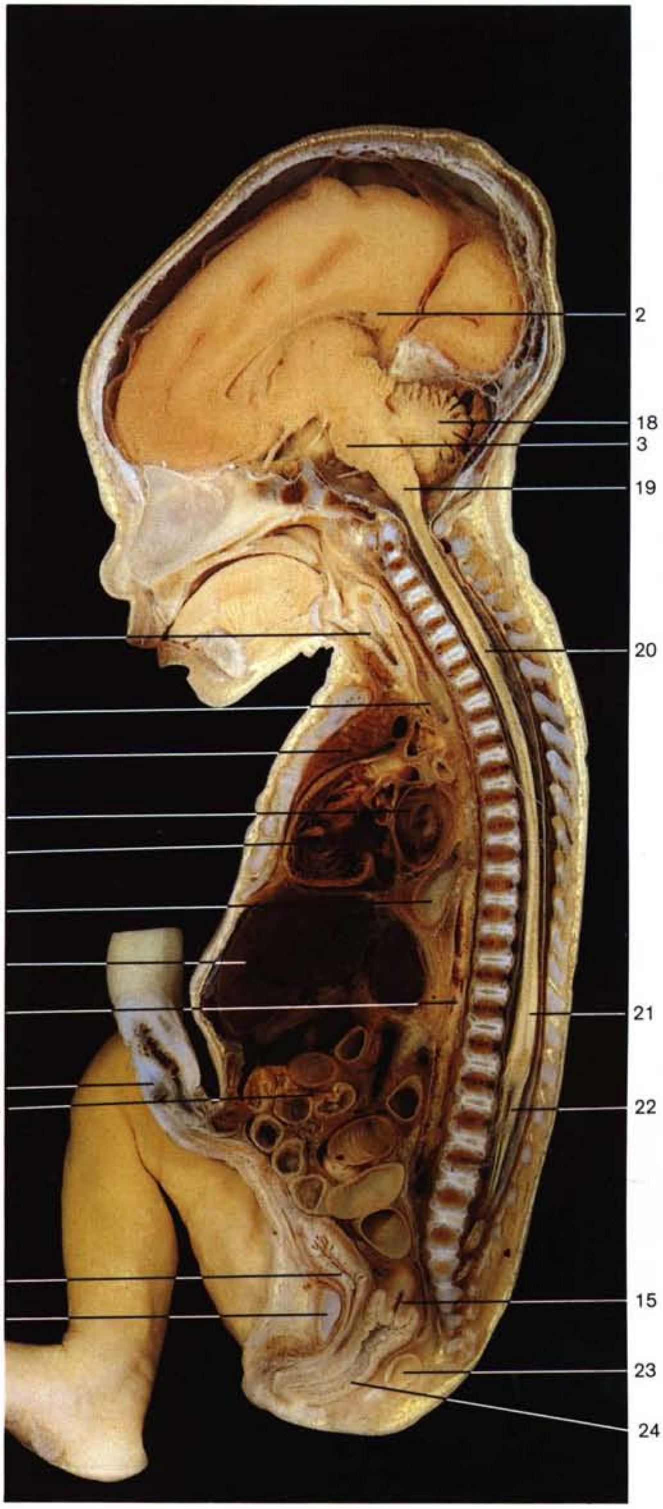
2
18
3
19
4
20
8
29
6
7
26
9
27
21
28
14
22
16
15
17
23
24

Median section of the head and trunk in the neonate. Note that in the neonate the conus medullaris of the spinal cord extends far more caudally than in the adult.

1 Cerebrum
2 Corpus callosum
3 Pons
4 Larynx
5 Trachea
6 Left atrium
7 **Right ventricle**
8 Esophagus
9 **Liver**
10 Stomach
11 Pancreas
12 Transverse colon
13 Umbilicus
14 Small intestine
15 Uterus
16 Urinary bladder
17 Pubic symphysis
18 Cerebellum
19 Medulla oblongata
20 **Spinal cord**
21 **Conus medullaris**
22 **Cauda equina**
23 Rectum
24 Vagina
25 Anus
26 Inferior vena cava
27 Aorta
28 Umbilical cord
29 Thymus

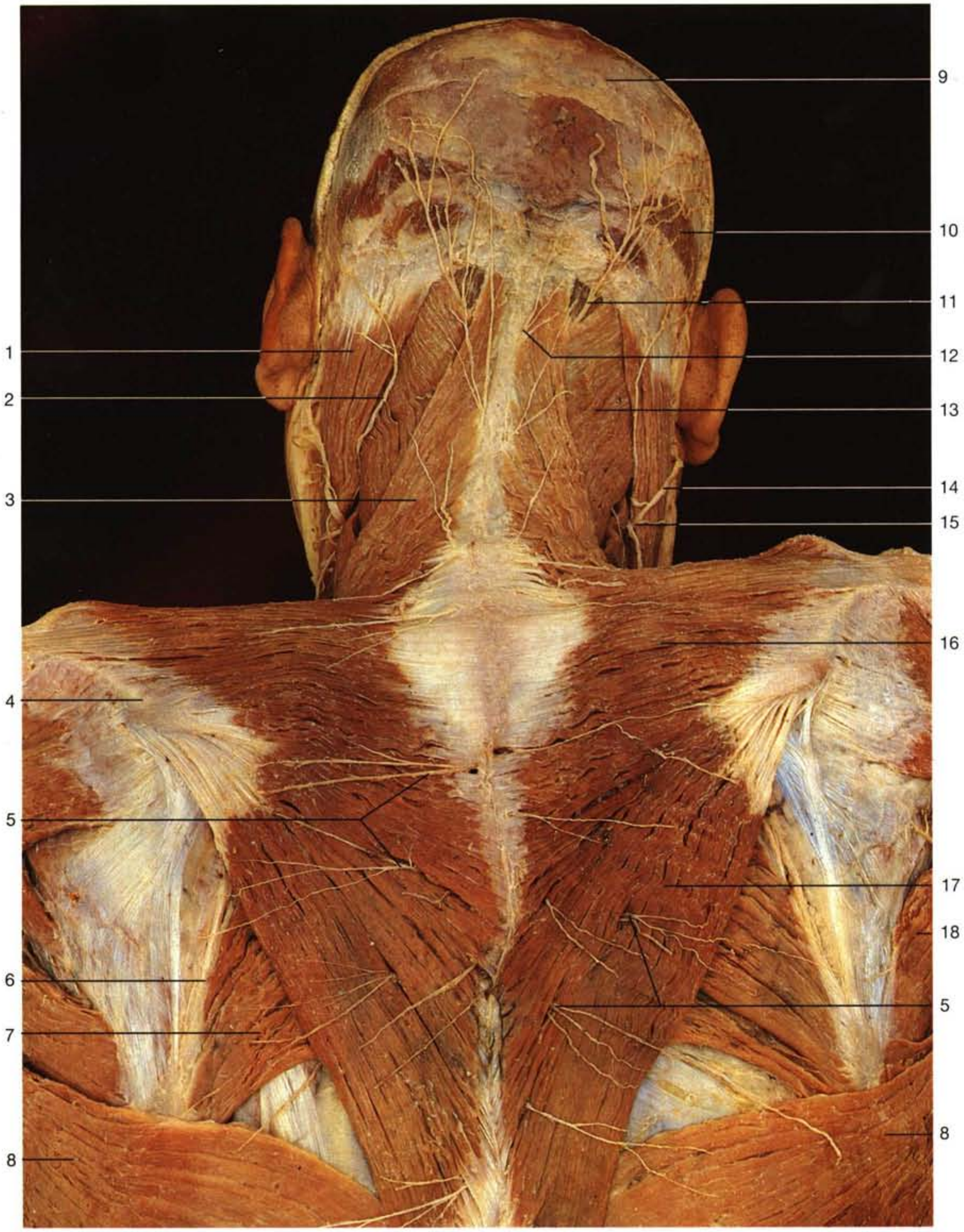

Dorsal aspect of the neck I. Superficial layer. Nuchal region and shoulder.

1 Sternocleidomastoid muscle
2 **Lesser occipital nerve**
3 Descending fibers of trapezius muscle
4 Spine of scapula
5 **Medial cutaneous branches of dorsal rami of spinal nerves**
6 Medial margin of scapula
7 Rhomboid major muscle
8 Latissimus dorsi muscle
9 Galea aponeurotica

10 Occipital belly of occipitofrontalis muscle
11 **Greater occipital nerve**
12 Third occipital nerve
13 Splenius capitis muscle
14 **Great auricular nerve**
15 Cutaneous nerves of cervical plexus
16 Transverse fibers of trapezius muscle
17 Ascending fibers of trapezius muscle
18 Teres major muscle

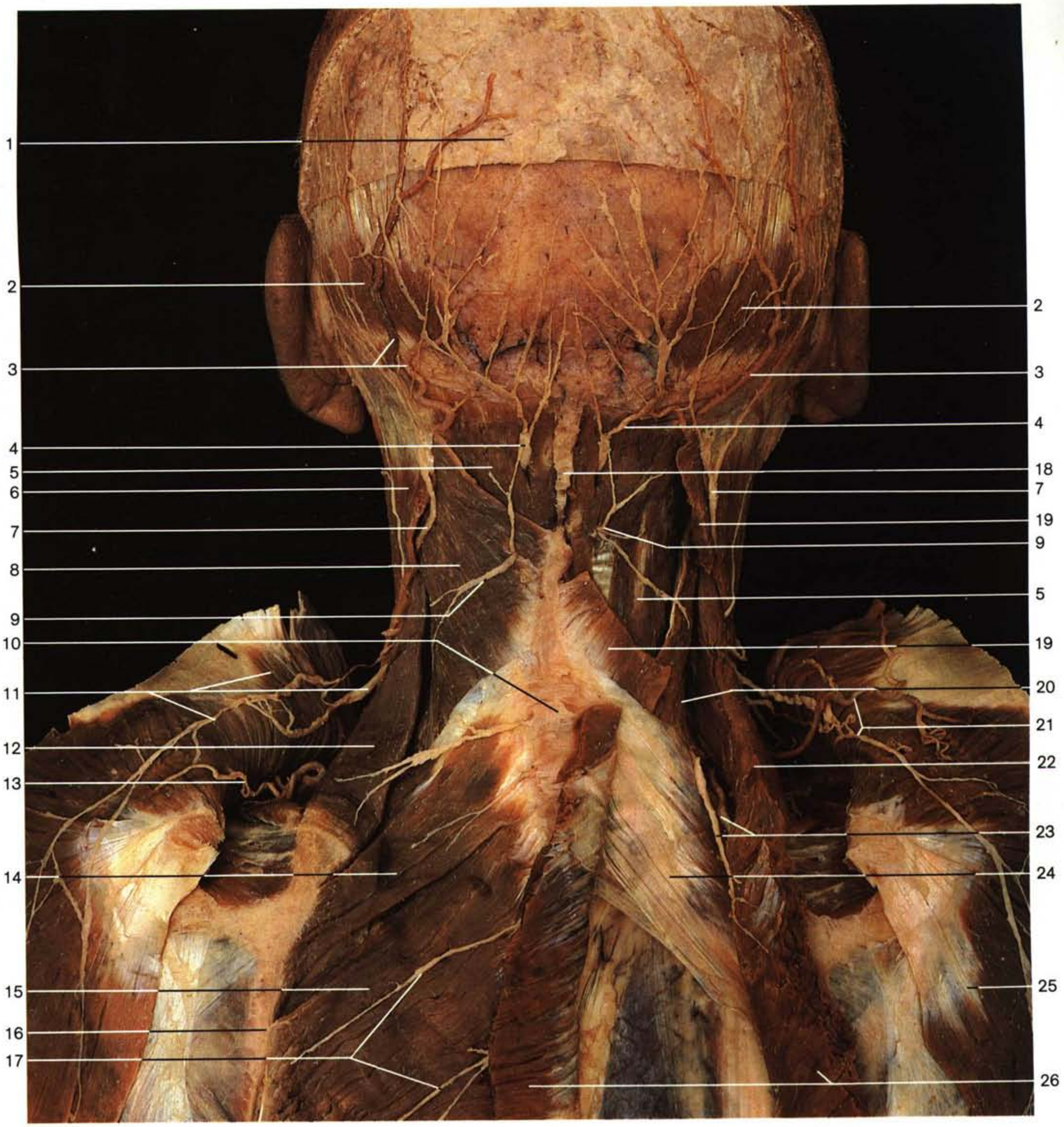

Dorsal aspect of neck II. Deeper layer. The left trapezius muscle has been divided and reflected. On the right, trapezius, rhomboid and splenius muscles have been divided. Right levator scapulae muscle has been slightly reflected.

1 Galea aponeurotica
2 Occipital belly of occipitofrontalis muscle
3 Occipital artery
4 **Greater occipital nerve** (C_2)
5 Semispinalis capitis muscle
6 Sternocleidomastoid muscle
7 Lesser occipital nerve
8 Left splenius capitis muscle
9 Third occipital nerve (C_3)
10 Spinous process of vertebra prominens (C_7)
11 Left trapezius muscle and **accessory nerve**
12 Levator scapulae muscle
13 Superficial branch of transverse cervical artery
14 Thomboid minor muscle
15 Rhomboid major muscle
16 Medial margin of scapula
17 **Medial branches of dorsal rami of spinal nerves**
18 Ligamentum nuchae
19 Splenius capitis muscle (divided)
20 Splenius cervicis muscle
21 Right **accessory nerve** and superficial branch of transverse cervical artery
22 Right levator scapulae muscle
23 Dorsal scapular nerve and deep branch of transverse cervical artery
24 Serratus posterior superior muscle
25 Right trapezius muscle (divided and reflected)
26 Right rhomboid major muscle (divided and reflected)

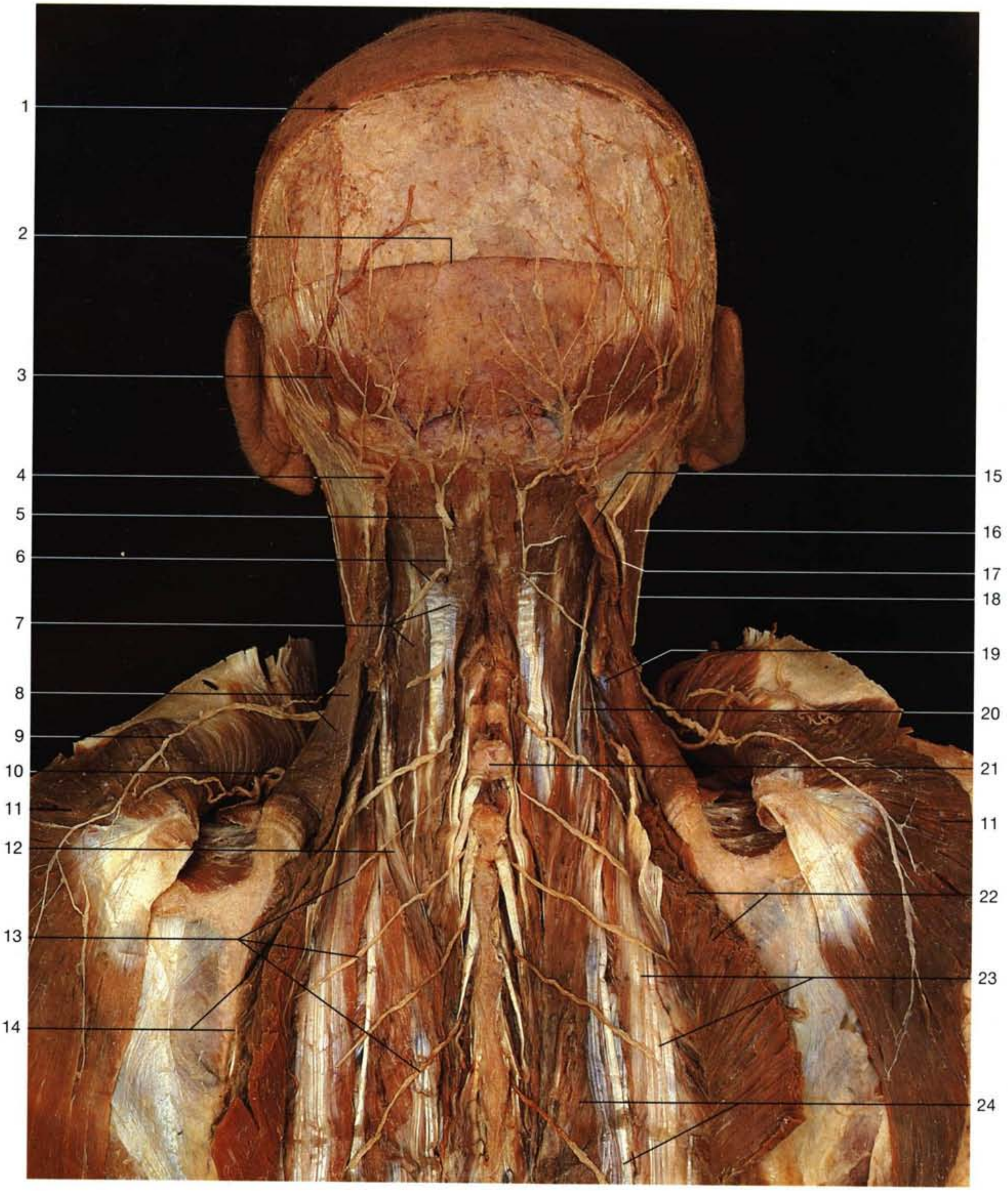

Dorsal aspect of neck III. Deepest layer. Nuchal region. Trapezius muscle and splenius capitis and cervicis muscles have been divided and partly removed or reflected.

1 Skin of scalp
2 Galea aponeurotica
3 Occipital belly of occipitofrontalis muscle
4 **Occipital artery**
5 **Greater occipital nerve**
6 Third occipital nerve
7 Semispinalis capitis muscle
8 Levator scapulae muscle
9 **Accessory nerve** (n. XI)
10 Superficial cervical artery
11 Trapezius muscle (reflected)
12 Longissimus cervicis muscle
13 **Medial cutaneous branches of dorsal rami of spinal nerves**
14 Medial margin of scapula
15 Splenius capitis muscle (divided)
16 Sternocleidomastoid muscle
17 **Lesser occipital nerve**
18 Great auricular nerve
19 Splenius cervicis muscle
20 Longissimus cervicis muscle
21 Spinous process of seventh cervical vertebra (vertebra prominens)
22 Rhomboid muscles (divided)
23 Iliocostalis thoracis muscle
24 Longissimus thoracis muscle

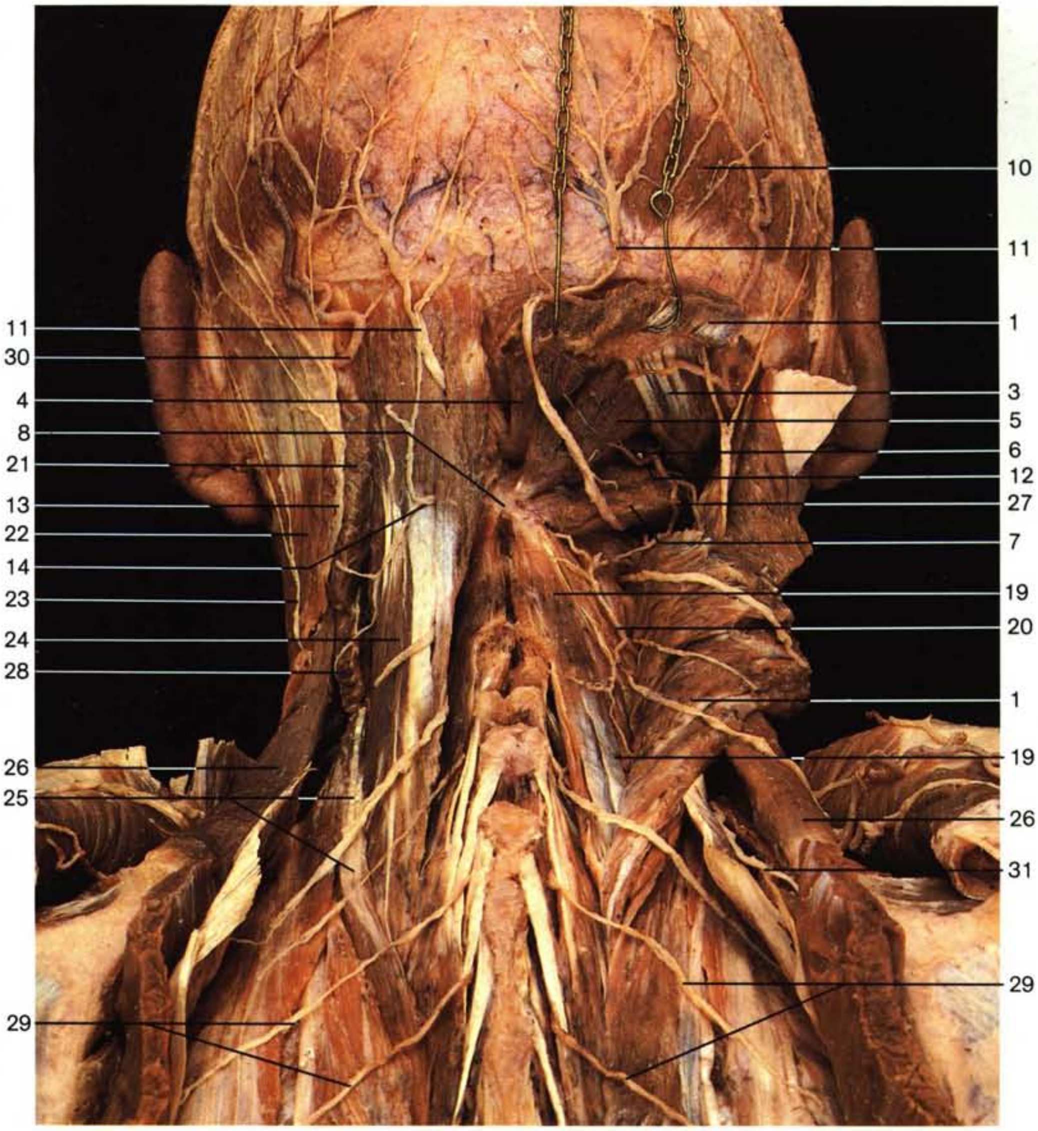

Dorsal aspect of neck IV. Deepest layer, suboccipital triangle; right semispinalis capitis muscle divited and reflected.

1	Semispinalis capitis muscle (divided)
2	External occipital protuberance
3	Obliquus capitis superior muscle
4	Rectus capitis posterior minor muscle
5	Rectus capitis posterior major muscle
6	Vertebral artery
7	Obliquus capitis inferior muscle
8	**Spinous process of axis**
9	Third cervical vertebra
10	Occipital belly of occipitofrontalis muscle
11	**Greater occipital nerve**
12	Suboccipital nerve (C_1)
13	Lesser occipital nerve
14	Third occipital nerve (C_3)
15	Mastoid process and splenius capitis muscle
16	Atlas
17	Axis
18	Spinous process of third cervical vertebra
19	Right semispinalis cervicis muscle
20	Deep cervical artery
21	Left splenius capitis muscle (divided)
22	Left sternocleidomastoid muscle
23	Great auricular nerve
24	Left **semispinalis capitis muscle**
25	Left longissimus cervicis muscle
26	Levator scapulae muscle
27	Muscular branch of vertebral artery
28	Left semispinalis cervicis muscle (divided)
29	**Medial branches of dorsal rami of spinal nerves**
30	**Occipital artery**
31	Dorsal scapular nerve

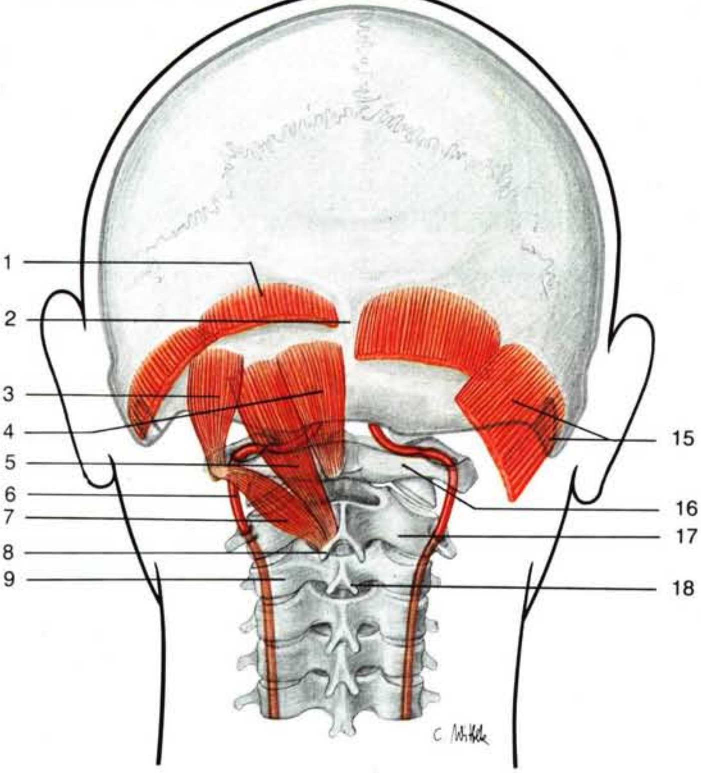

Suboccipital triangle and position of the vertebral artery. (Schematic drawing.)

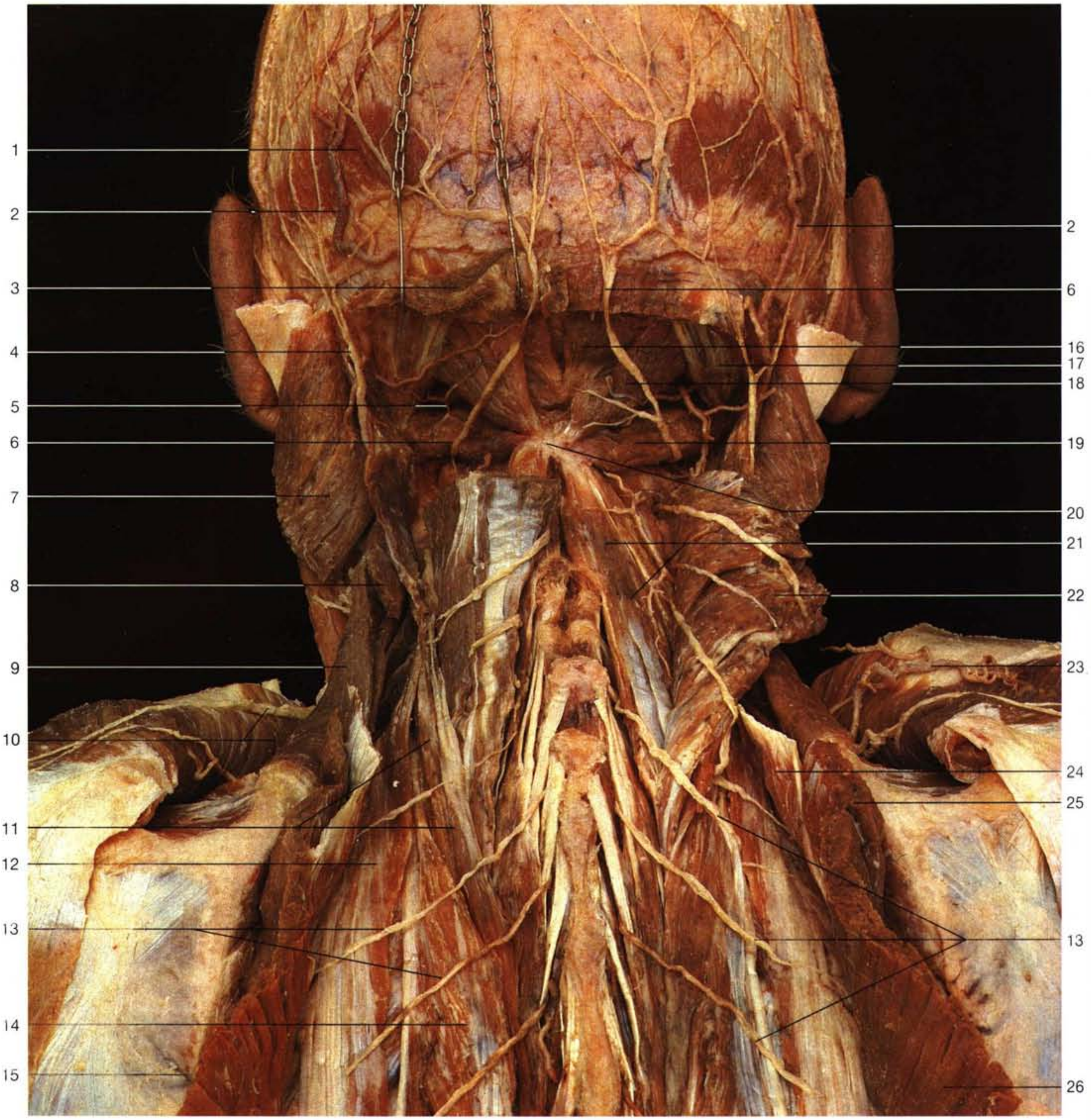

Dorsal aspect of neck V. Nuchal region. Deepest layer. Dissection of suboccipital triangle on both sides.

1 Occipital belly of occipitofrontalis muscle
2 **Occipital artery**
3 Insertion of semispinalis capitis muscle (divided)
4 Lesser occipital nerve (from cervical plexus)
5 Suboccipital nerve (C_1)
6 **Greater occipital nerve** (C_2)
7 Splenius capitis muscle (reflected)
8 Splenius cervicis muscle
9 Levator scapulae muscle
10 Accessory nerve (n. XI), trapezius muscle
11 Longissimus cervicis muscle
12 Iliocostalis cervicis muscle
13 **Medial cutaneus branches of dorsal rami of spinal nerves** (C_7, C_8)
14 Longissimus thoracis muscle
15 Medial margin of scapula
16 **Rectus capitis posterior minor muscle**
17 **Obliquus capitis superior muscle**
18 **Rectus capitis posterior major muscle**
19 **Obliquus capitis inferior muscle**
20 Spinous process of axis
21 Semispinalis cervicis muscle
22 Semispinalis capitis muscle (divided and reflected)
23 Transverse cervical artery (superficial branch)
24 Serratus posterior superior muscle (divided and reflected)
25 Rhomboid minor muscle (divided and reflected)
26 Rhomboid major muscle (divided and reflected)

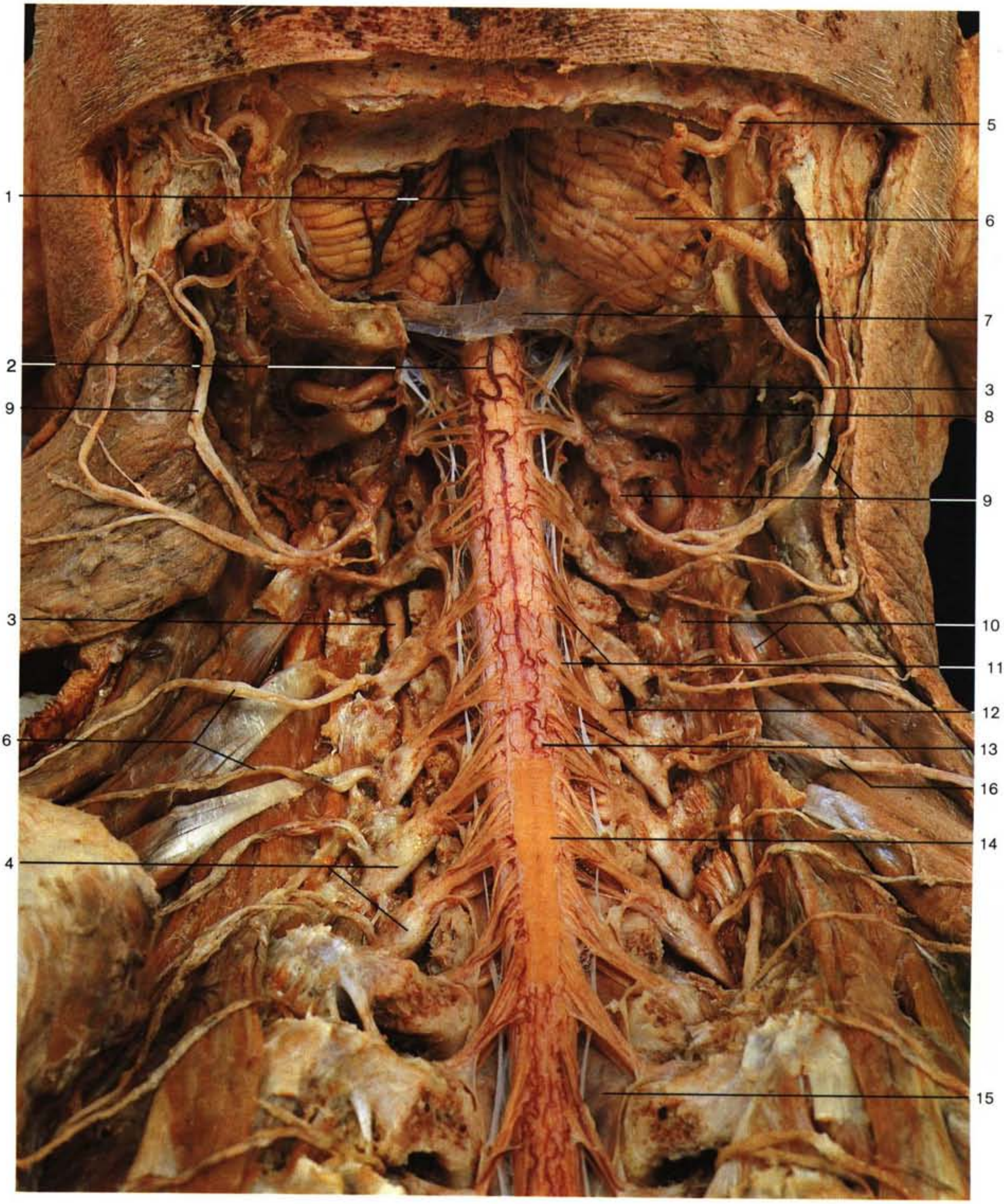

Neck, deepest layer. Spinal cord and medulla oblongata (dorsal aspect). Cranial cavity opened.

1 Vermis of the cerebellum
2 Medulla oblongata and posterior spinal artery
3 Vertebral artery
4 **Spinal ganglion**
5 Occipital artery
6 Cerebellum
7 **Cerebellomedullary cistern**
8 Atlas

9 Greater occipital nerve (C_2)
10 Levator scapulae muscle and intertransverse ligament
11 **Dorsal roots of spinal nerves**
12 Vertebral arch
13 **Denticulate ligament and arachnoid mater**
14 Area where pia mater has been removed
15 Dura mater
16 **Dorsal rami of spinal nerves**

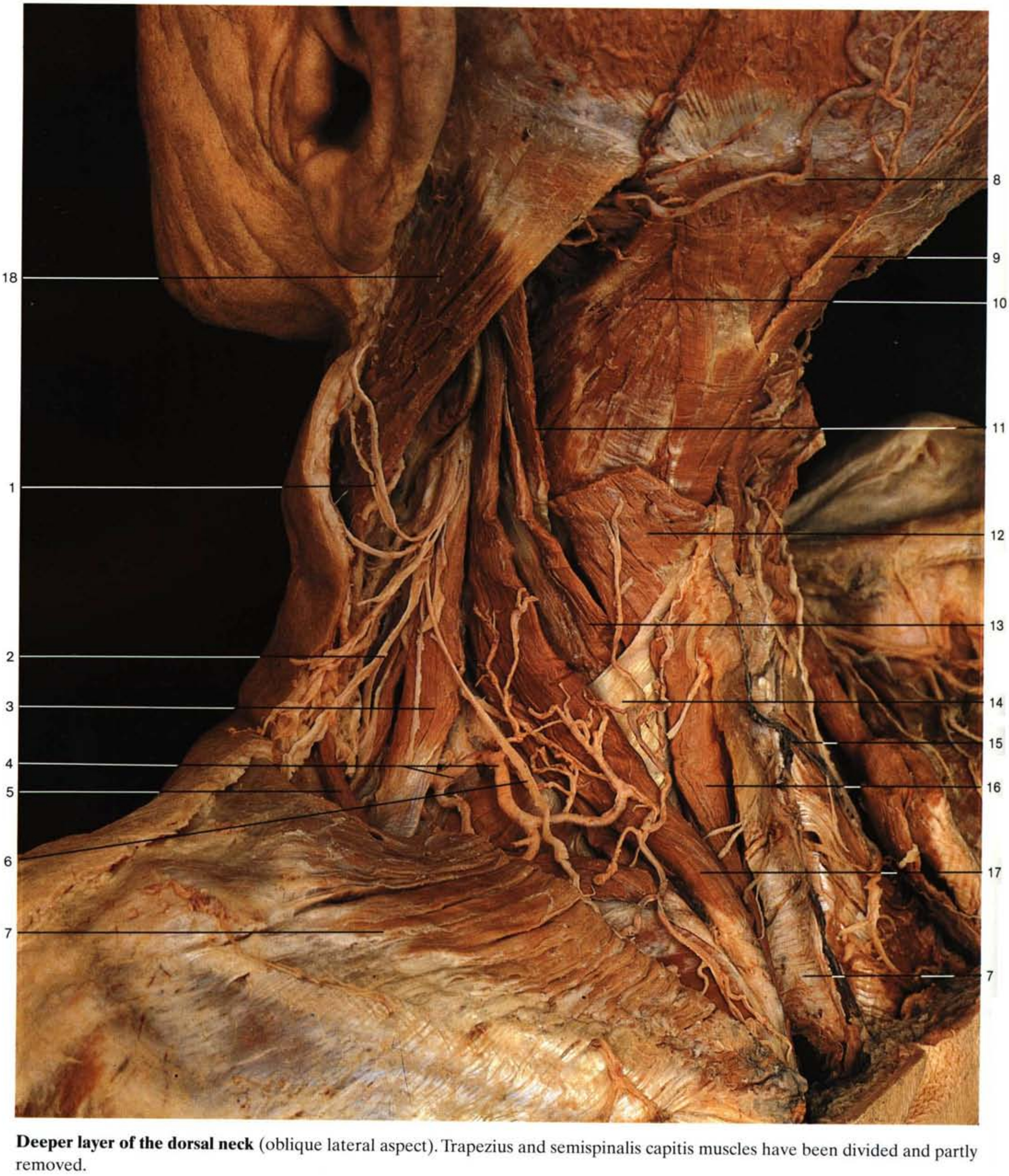

Deeper layer of the dorsal neck (oblique lateral aspect). Trapezius and semispinalis capitis muscles have been divided and partly removed.

1 Great auricular nerve
2 Accessory nerve (n. XI)
3 Scalenus posterior muscle
4 Transverse cervical artery
5 Omohyoid muscle
6 Muscular branch of cervical plexus

7 Trapezius muscle
8 Occipital artery
9 Greater occipital nerve
10 Semispinalis capitis muscle
11 Longissimus capitis muscle
12 Splenius capitis muscle

13 Splenius cervicis muscle
14 Serratus posterior superior muscle
15 Deep cervical vein (variant)
16 Rhomboid minor muscle
17 Levator scapulae muscle
18 Sternocleidomastoid muscle

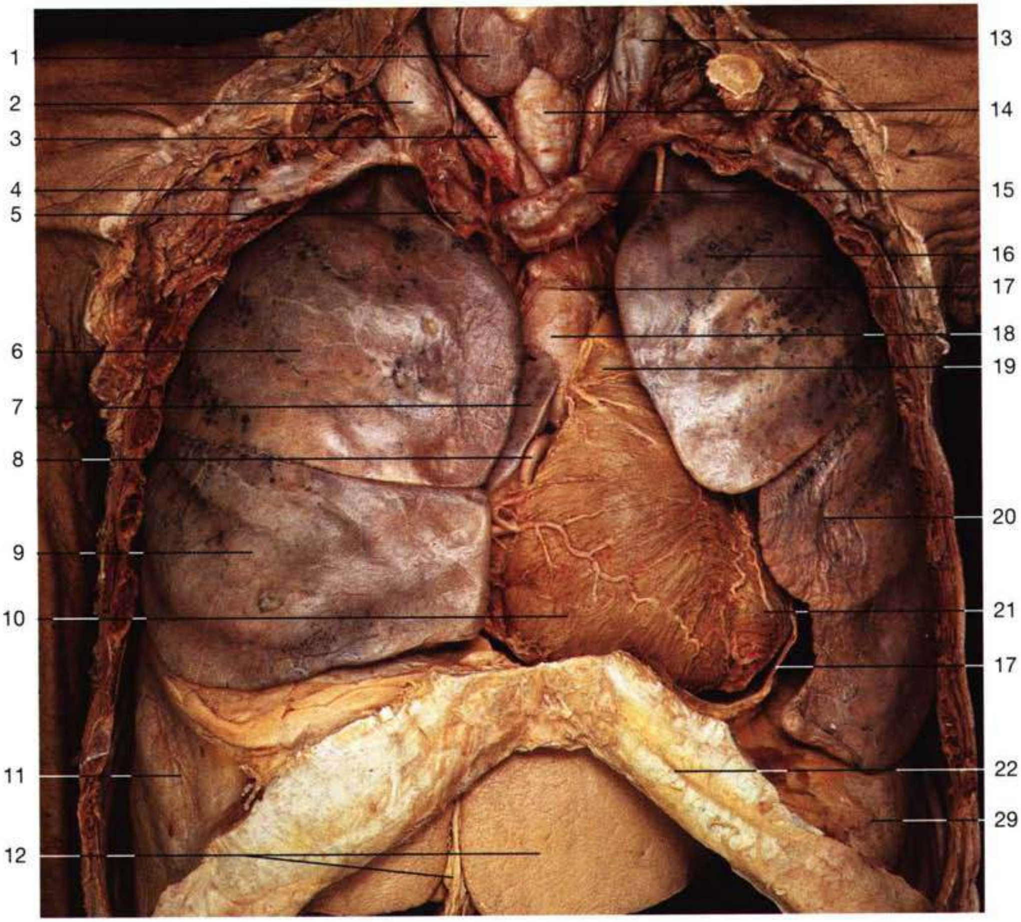

1 Thyroid gland
2 Internal jugular vein
3 Right common carotid artery
4 Right axillary vein
5 Right brachiocephalic vein
6 Superior lobe of right lung
7 Right atrium
8 Right coronary artery
9 Middle lobe of right lung
10 **Right ventricle**
11 Diaphragm
12 Liver (left lobe) and falciform ligament
13 Left internal jugular vein
14 **Trachea**
15 Left brachiocephalic vein
16 Superior lobe of left lung
17 Cut edge of pericardium
18 **Ascending aorta**
19 **Pulmonary trunk**
20 Inferior lobe of left lung
21 **Left ventricle**
22 Costal margin
23 Right pulmonary artery
24 Esophagus
25 Descending aorta
26 Pericardium
27 Aortic valve
28 Thymus
29 Diaphragm

Thoracic organs, heart and lungs in situ (ventral aspect). Anterior thoracic wall, parietal pleura and pericardium had been removed.

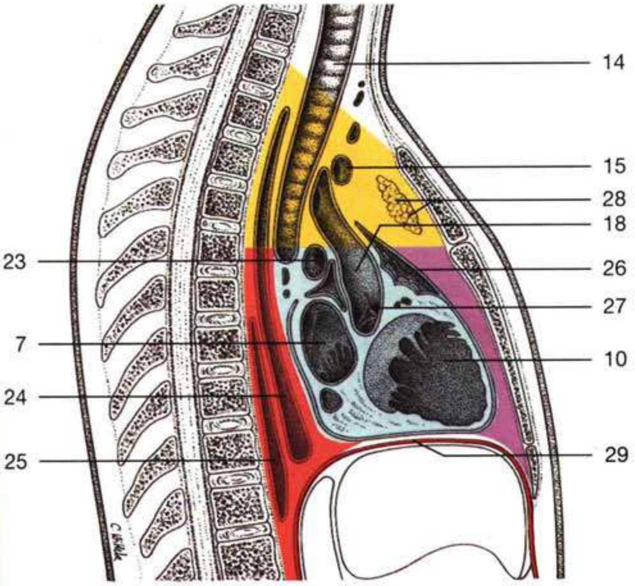

Sagittal section through thoracic cavity. The divisions of the mediastinum are indicated by colors.

Divisions of mediastinum	**Main content**
Superior mediastinum (yellow)	Trachea, brachiocephalic veins, thymus, aortic arch, esophagus, thoracic duct
Middle mediastinum (light blue)	Heart, ascending aorta, pulmonary trunk, pulmonary veins, phrenic nerves
Posterior mediastinum (red)	Esophagus with vagus nerves, descending aorta, thoracic duct, sympathetic trunks
Anterior mediastinum (pink)	Small vessels, connective and fatty tissue, and thymus in a child

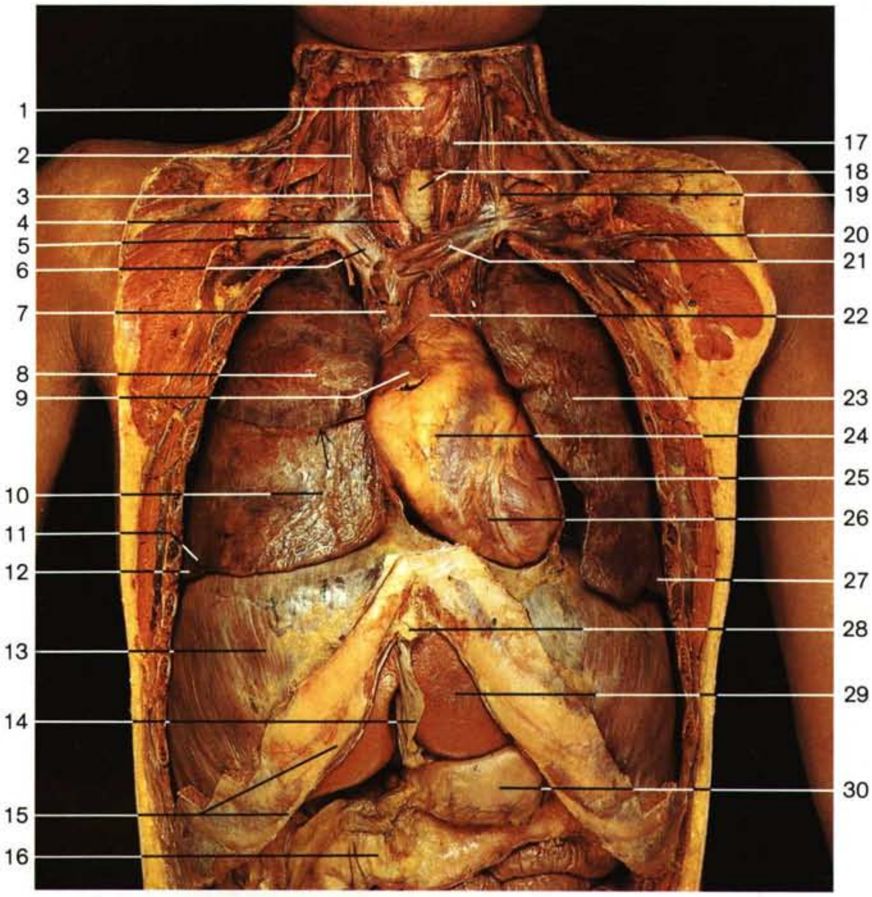
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

1 Cricothyroid muscle
2 Right internal jugular vein
3 Vagus nerve
4 Right common carotid artery
5 Right subclavian vein
6 Right brachiocephalic vein
7 Superior vena cava
8 Upper lobe of right lung
9 Right auricle
10 Middle lobe of right **lung**
11 Oblique fissure of right lung
12 Lower lobe of right lung
13 **Diaphragm**
14 Falciform ligament
15 Costal margin
16 Transverse colon
17 Thyroid gland
18 Trachea
19 Left internal jugular vein
20 Left cephalic vein
21 Left brachiocephalic vein
22 Pericardium (cut edge)
23 Upper lobe of left **lung**
24 **Right ventricle**
25 **Left ventricle**
26 Anterior interventricular sulcus
27 Lower lobe of left lung
28 Xiphoid process
29 **Liver**
30 **Stomach**
31 Pectoralis major muscle

Positions of thoracic organs. The anterior thoracic wall has been removed. Arrow: horizontal fissure of right lung.

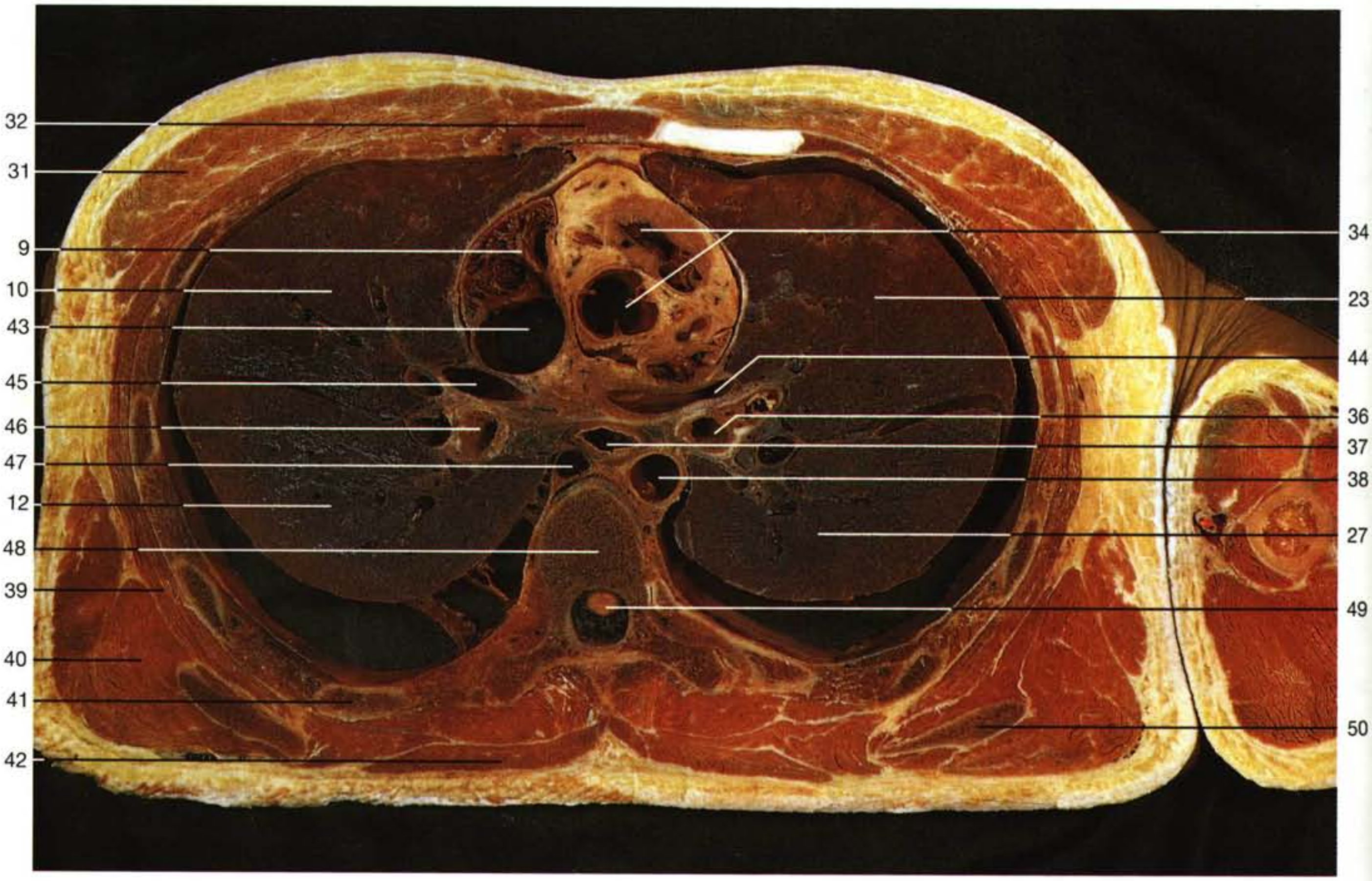
32
31
9
10
43
45
46
47
12
48
39
40
41
42
34
23
44
36
37
38
27
49
50

Horizontal section through the thorax at the level of the 7th thoracic vertebra (from below).

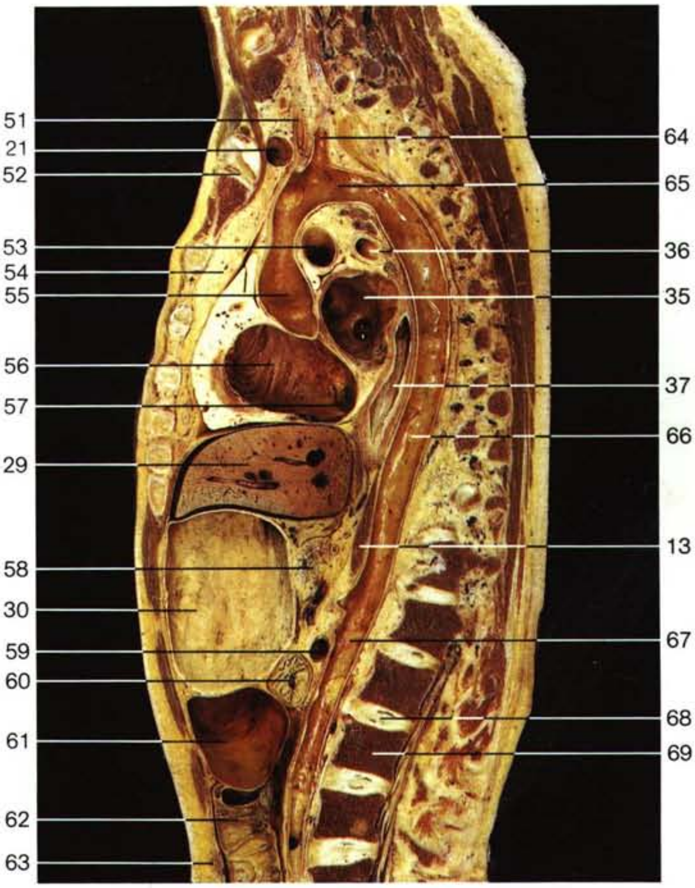
51
21
52
53
54
55
56
57
29
58
30
59
60
61
62
63
64
65
36
35
37
66
13
67
68
69

Sagittal section through the left thorax, 2 cm lateral to the median plane.

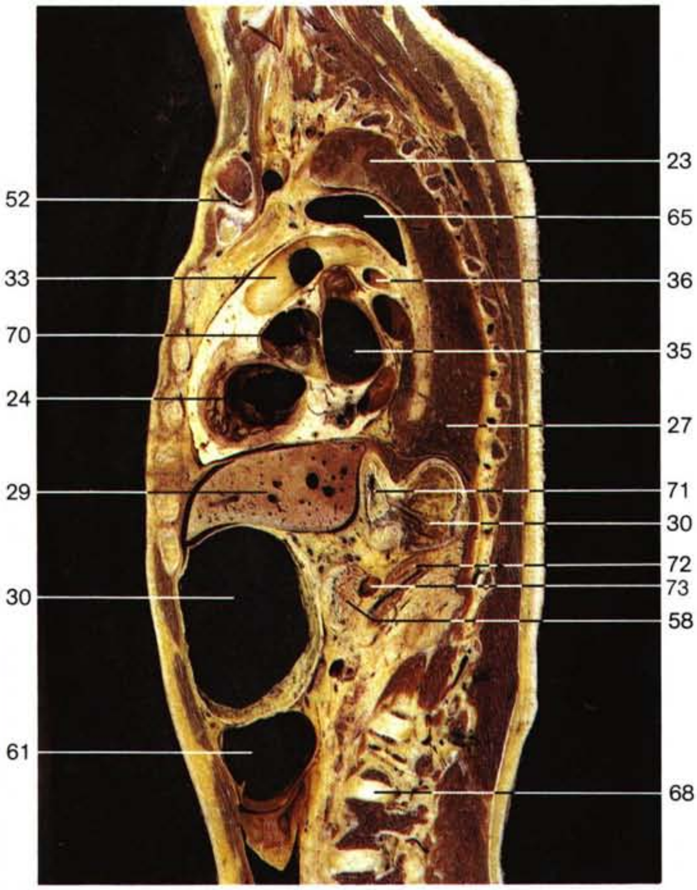
52
33
70
24
29
30
61
23
65
36
35
27
71
30
72
73
58
68

Sagittal section through the left thorax, 3.5 cm lateral to the median plane.

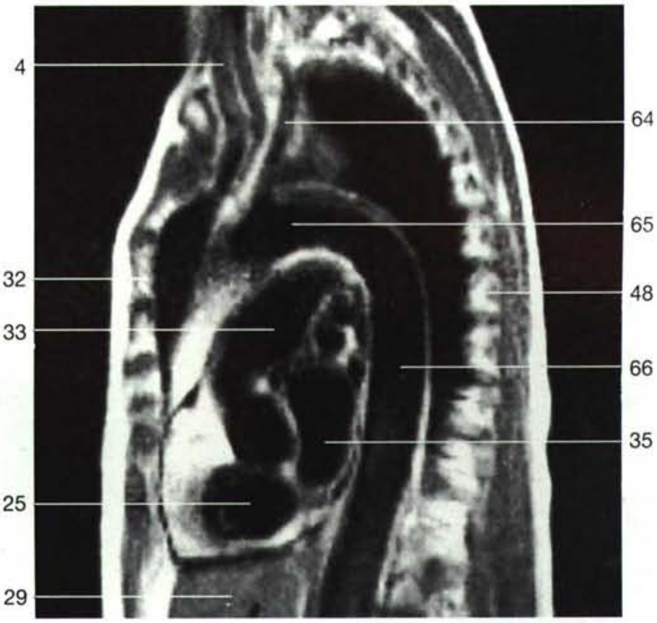
4
32
33
25
29
64
65
48
66
35

Sagittal section through the thorax. MR-Scan.

32 Sternum
33 Pulmonary trunk
34 Left ventricle and bulb of aorta
35 Left atrium
36 Left main bronchus
37 **Esophagus**
38 Descending aorta
39 Serratus anterior muscle
40 Teres major muscle
41 Rib
42 Trapezius muscle
43 Right atrium
44 Left pulmonary vein
45 Right pulmonary vein
46 Right main bronchus
47 Azygos vein
48 Body of vertebra
49 Spinal cord
50 Scapula
51 Left common carotid artery
52 Sternoclavicular articulation with disc
53 Right pulmonary artery
54 Remnants of thymus
55 Aortic bulb
56 Right atrium
57 Entrance of inferior vena cava in right atrium
58 **Pancreas**
59 Left renal vein
60 **Duodenum**
61 Transverse colon (dilated)
62 Small intestine
63 Umbilicus
64 Left subclavian artery
65 Aortic arch
66 **Thoracic aorta**
67 Abdominal aorta
68 Intervertebral disc
69 Body of lumbar vertebra
70 Aortic valve
71 Cardia of stomach
72 Suprarenal gland
73 Splenic vein

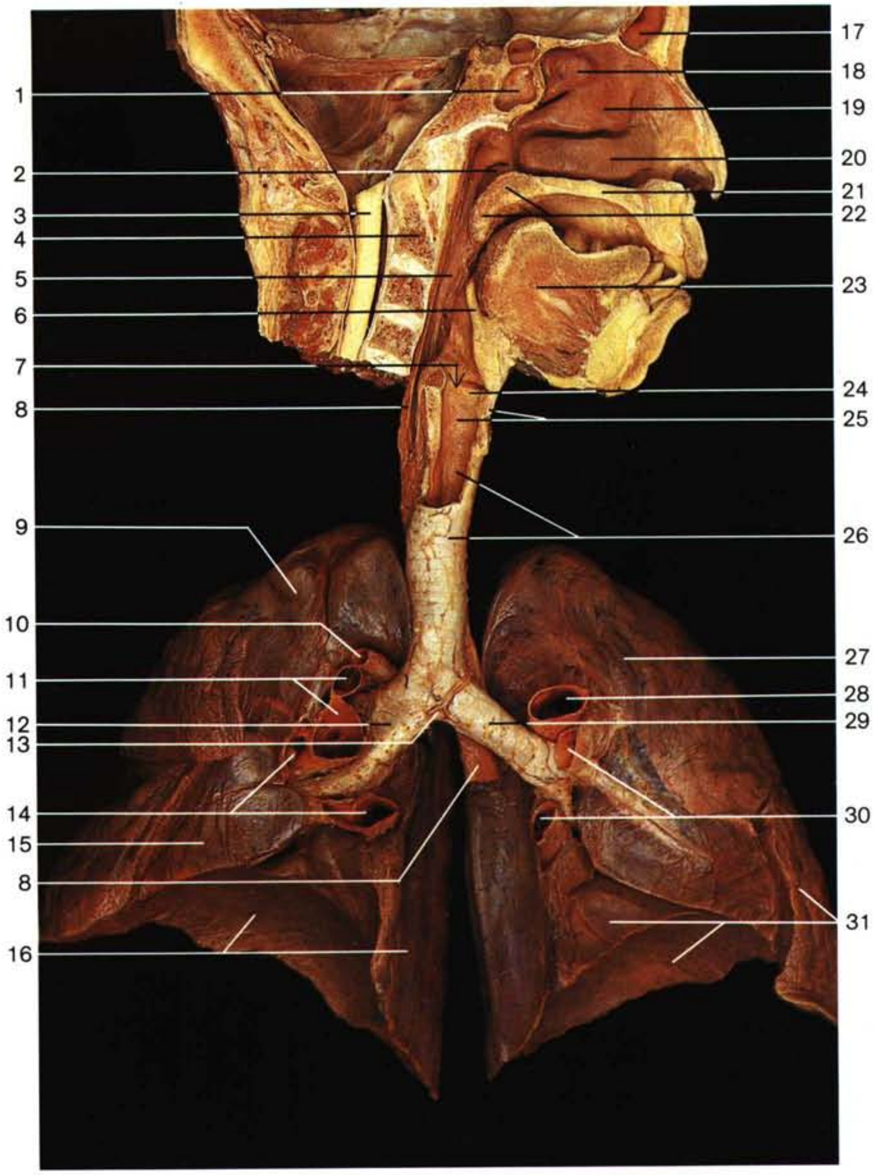

Respiratory system. The lungs have been fixed in expiration and turned laterally. Head bisected and turned laterally.

1 Sphenoid sinus
2 Pharyngeal opening of auditory tube
3 Spinal cord
4 Dens of axis
5 **Oropharynx** (oropharyngeal isthmus)
6 Epiglottis
7 Entrance of larynx
8 Esophagus
9 Upper lobe of right lung
10 Azygos vein
11 Branches of pulmonary artery
12 **Right main bronchus**
13 **Bifurcation of trachea**
14 Tributaries of right pulmonary veins
15 Middle lobe of right lung
16 Lower lobe of right lung
17 Frontal sinus
18 Superior nasal concha
19 Middle nasal concha
20 Inferior nasal concha
21 Hard palate
22 Soft palate with uvula
23 Tongue
24 Vocal fold
25 Larynx
26 Trachea
27 Upper lobe of left lung
28 Left pulmonary artery
29 **Left main bronchus**
30 Left pulmonary veins
31 Lower lobe of left lung

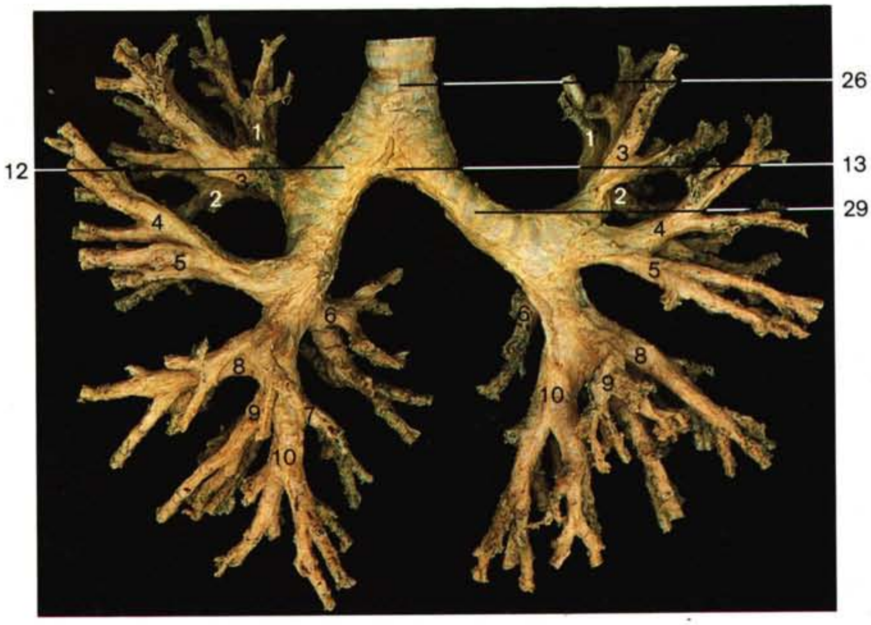

Bronchial tree (ventral aspect). The lung tissue has been removed. The bronchopulmonary segments are numbered 1–10.

▷

To page 231:
1 Nasal cavity
2 Pharynx
3 Larynx (thyroid cartilage)
4 Trachea
5 Upper lobe of right lung
6 Bifurcation of trachea
7 Right main bronchus
8 Horizontal fissure of right lung
9 Middle lobe of right lung
10 Oblique fissures of lungs
11 Lower lobe of right lung
12 Clavicle
13 Upper lobe of left lung
14 Left main bronchus
15 Bronchi supplying bronchopulmonary segments
16 Lower lobe of left lung
17 Costal margin
18 Hyoid bone
19 Right superior lobe bronchus
20 Right middle lobe bronchus
21 Right inferior lobe bronchus
22 Left superior lobe bronchus
23 Left inferior lobe bronchus
24 Segmental bronchi
25 Branches of pulmonary arteries
26 Branches of pulmonary veins

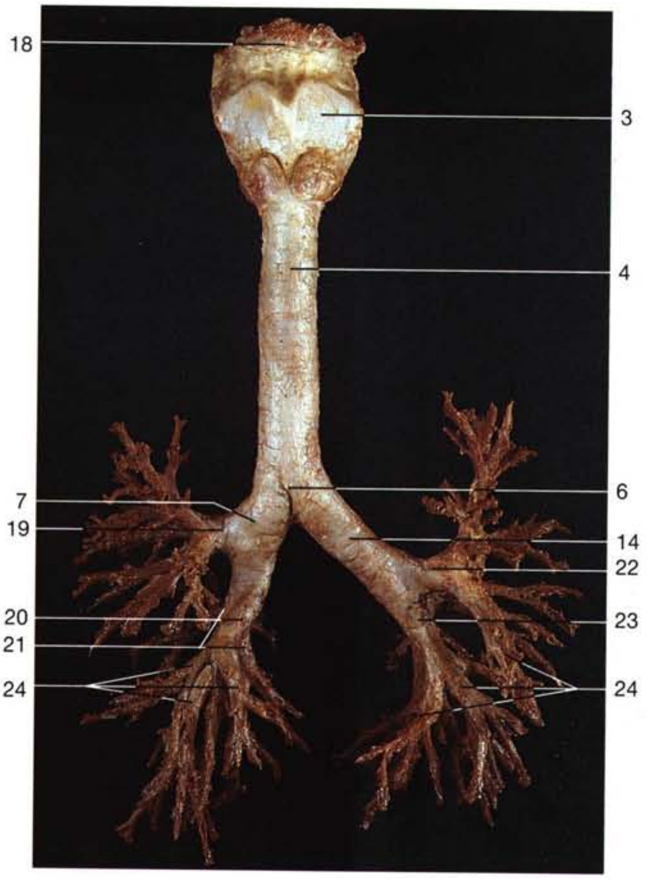
18
3
4
7
19
6
14
22
20
23
21
24
24

Larynx, trachea and bronchial tree (anterior aspect).

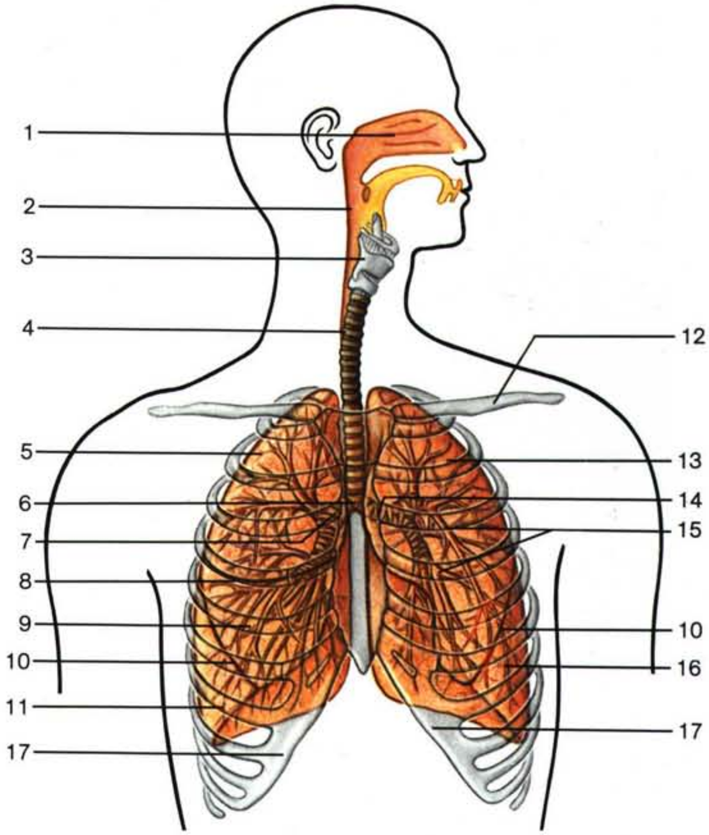
1
2
3
4
12
5
13
6
14
7
15
8
9
10
10
16
11
17
17

Organization and positions of respiratory organs. (Schematic drawing.)

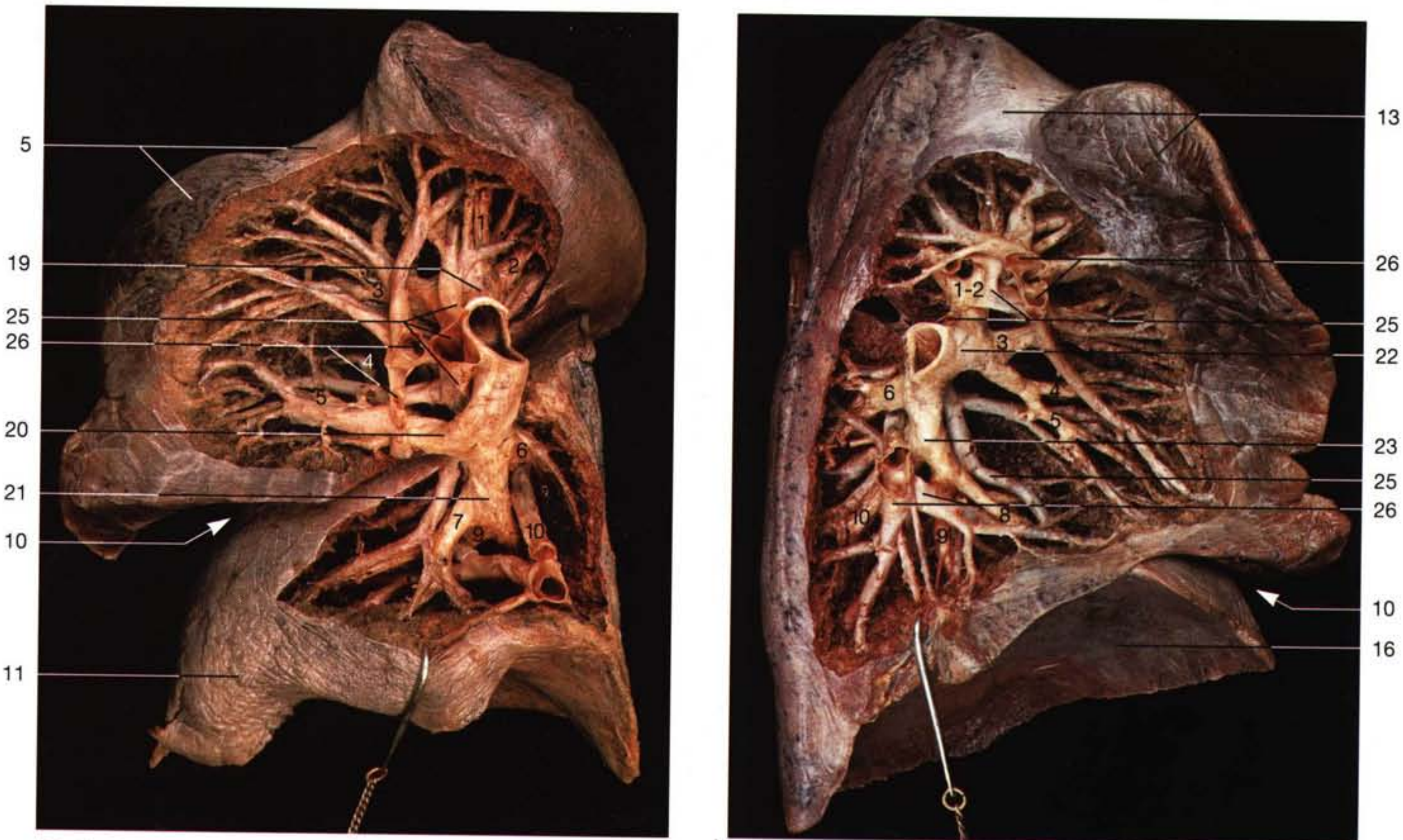
5
19
25
26
20
21
10
11
1
2
3
4
5
6
7
9
10
13
26
25
22
23
25
26
10
16
1-2
3
4
5
6
8
9
10

Mediastinal dissection of the bronchial tree, pulmonary veins, and pulmonary arteries of right lung (left) and left lung (right) (medial aspect). Segmental bronchi are numbered 1–10.

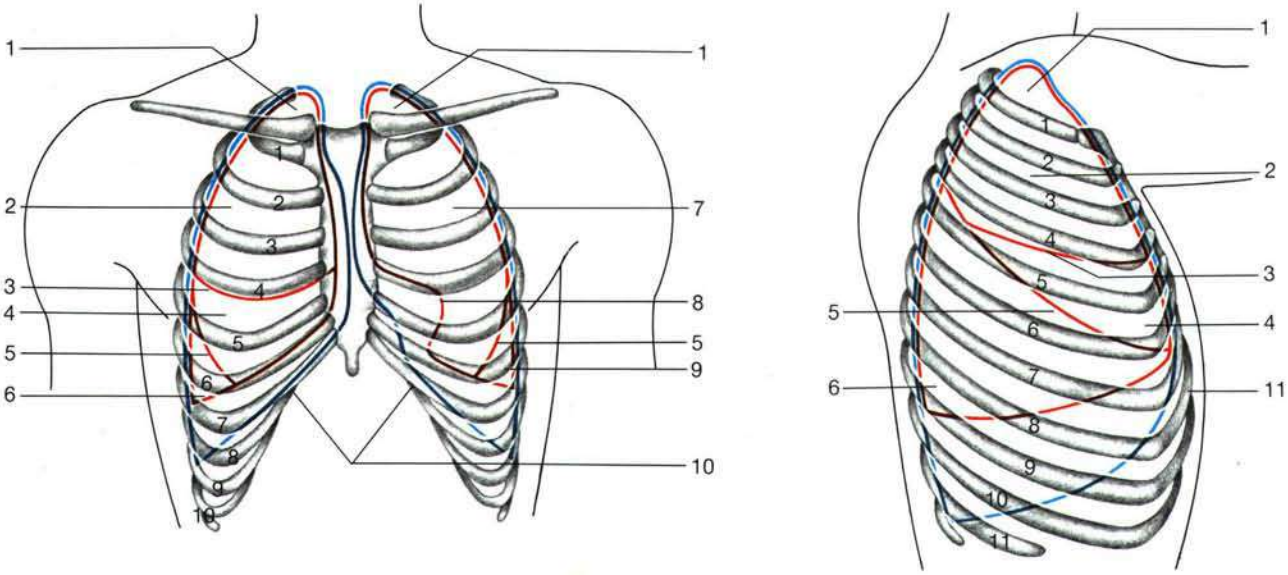
1
1
1
2
3
4
5
6
7
8
9
10
2
7
3
4
5
6
8
5
9
10
1
1
2
3
4
5
6
7
8
9
10
11
2
3
5
4
6
11

Surface projections of lungs and pleura on the thoracic wall. Left: anterior aspect; right: right-lateral aspect. Red = margins of the lung; blue = margins of pleura. The numbers indicate ribs.

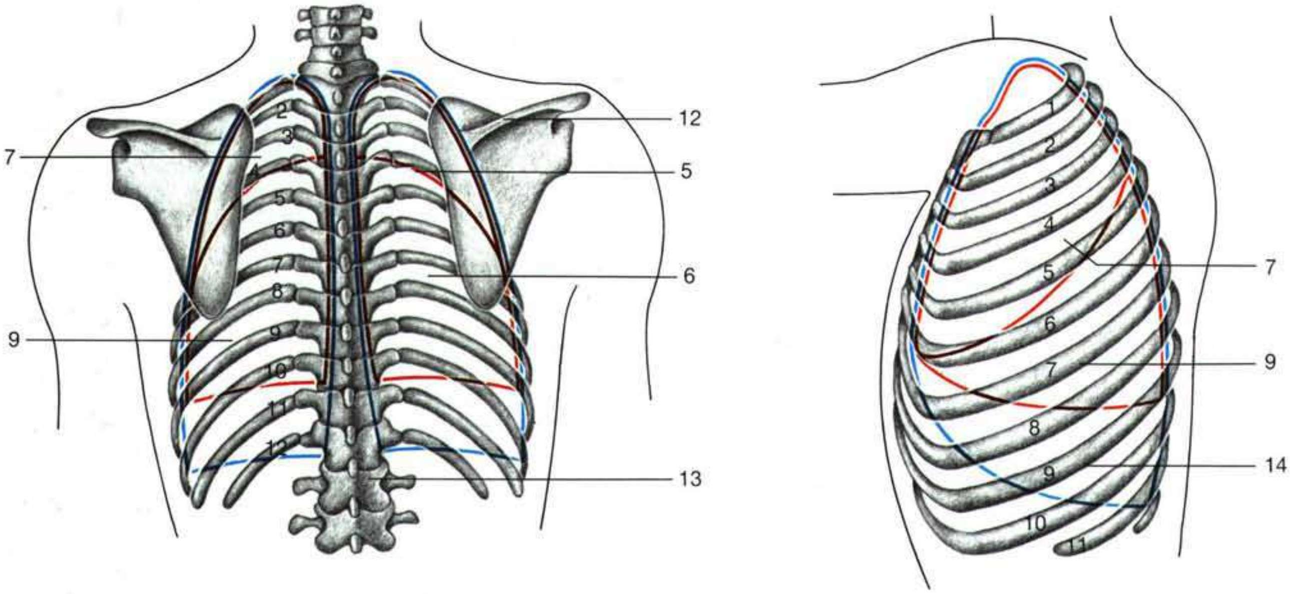
2
3
7
4
5
6
7
8
9
10
11
12
12
5
6
9
13
1
2
3
4
5
6
7
8
9
10
11
7
9
14

Surface projections of lungs and pleura on thoracic wall. Left: posterior aspect; right: left-lateral aspect. Red = margins of lung; blue = margins of pleura. The number indicate ribs.

1 Apex of lung
2 Upper lobe of right lung
3 Horizontal fissure of right lung
4 Middle lobe of right lung
5 Oblique fissures of lungs

6 Lower lobe of right lung
7 Upper lobe of left lung
8 Cardiac notch of left lung
9 Lower lobe of left lung
10 Infrasternal angle

11 Costal margin
12 Spine of scapula
13 1st lumbar vertebra
14 Space between border of lung and pleura (costodiaphragmatic recess)

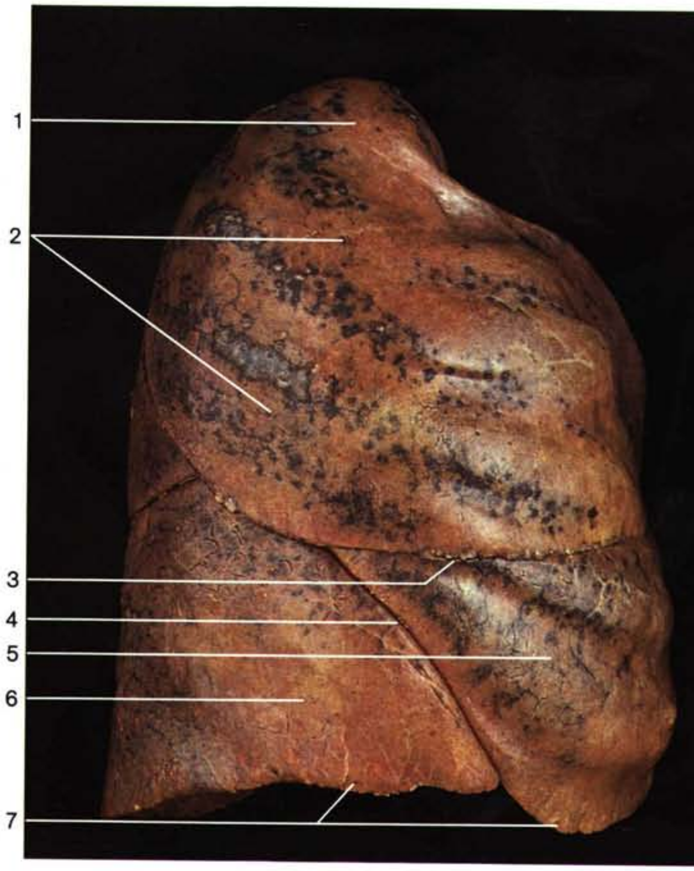

Right lung (lateral aspect).

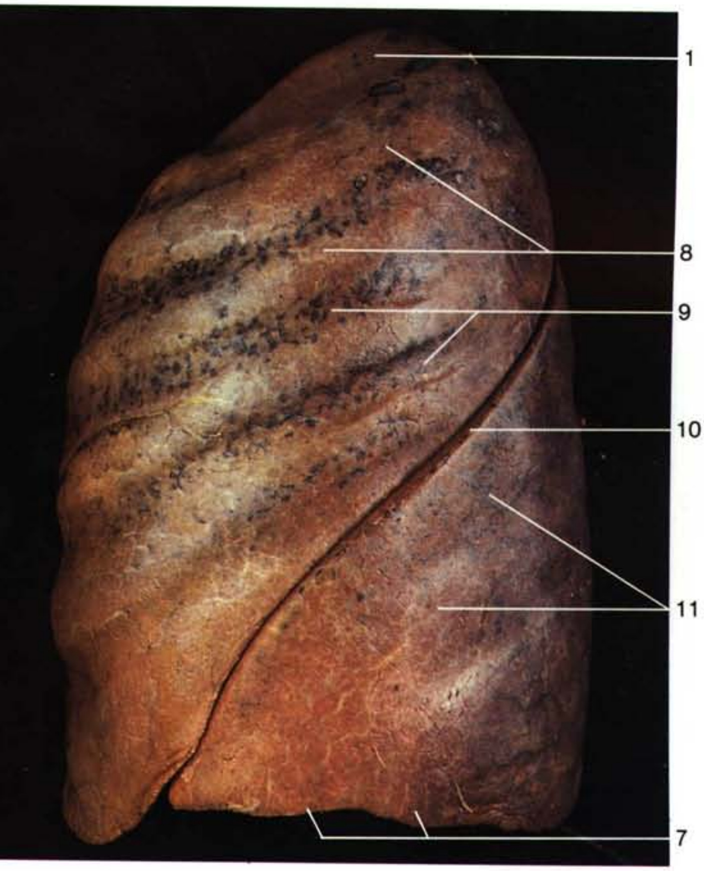

Left lung (lateral aspect).

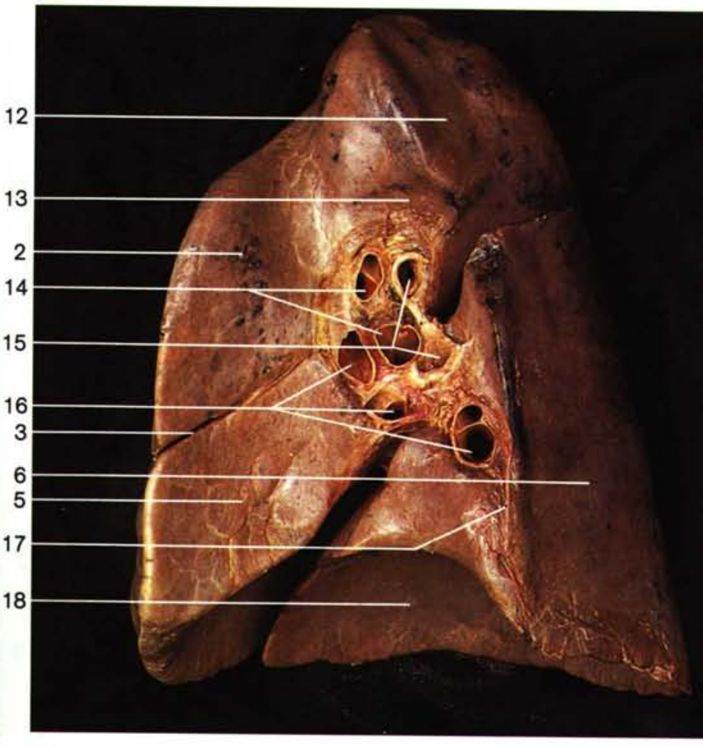

Right lung (medial aspect).

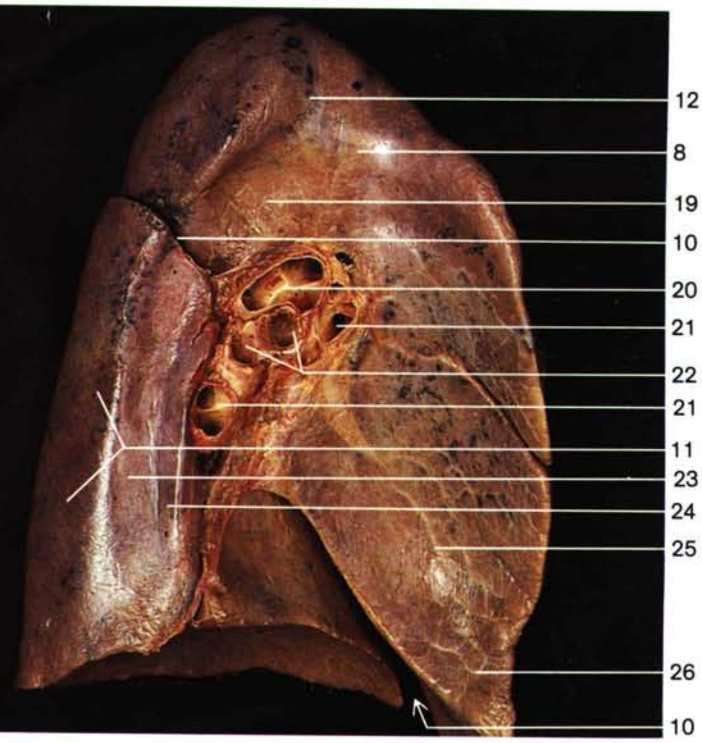

Left lung (medial aspect).

1 Apex of lung
2 Upper lobe of right lung
3 Horizontal fissure of right lung
4 Oblique fissure of right lung
5 Middle lobe of right lung
6 Lower lobe of right lung
7 Inferior border
8 Upper lobe of left lung
9 Impressions of ribs
10 Oblique fissure of left lung
11 Lower lobe of left lung
12 Groove of subclavian artery
13 Groove of azygos arch
14 Branches of right pulmonary artery
15 Bronchi
16 Right pulmonary veins
17 Pulmonary ligament
18 Diaphragmatic surface
19 Groove of aortic arch
20 Left pulmonary artery
21 Branches of left pulmonary veins
22 Left secondary bronchi
23 Groove of thoracic aorta
24 Groove of esophagus
25 Cardiac impression
26 Lingula

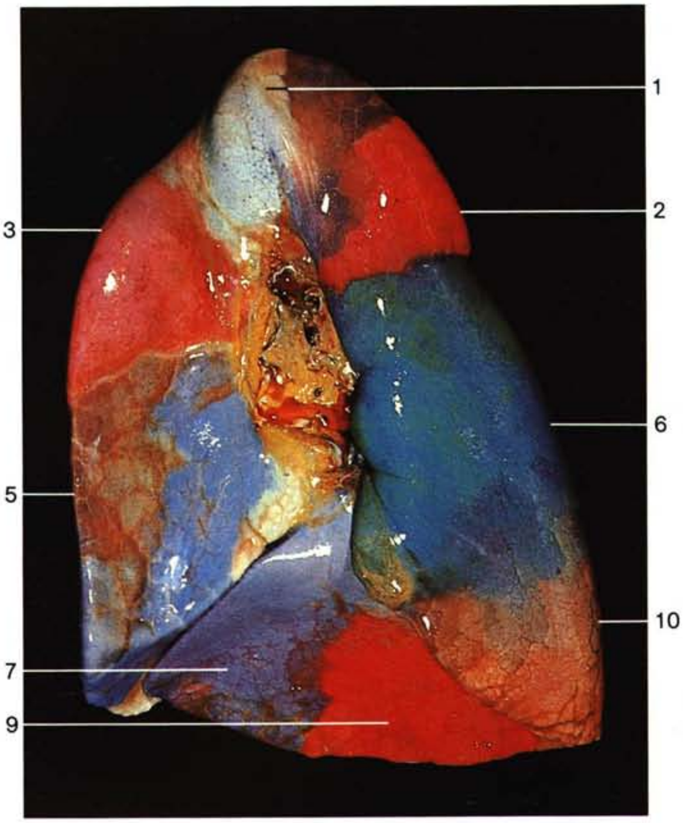

Right lung (medial aspect).

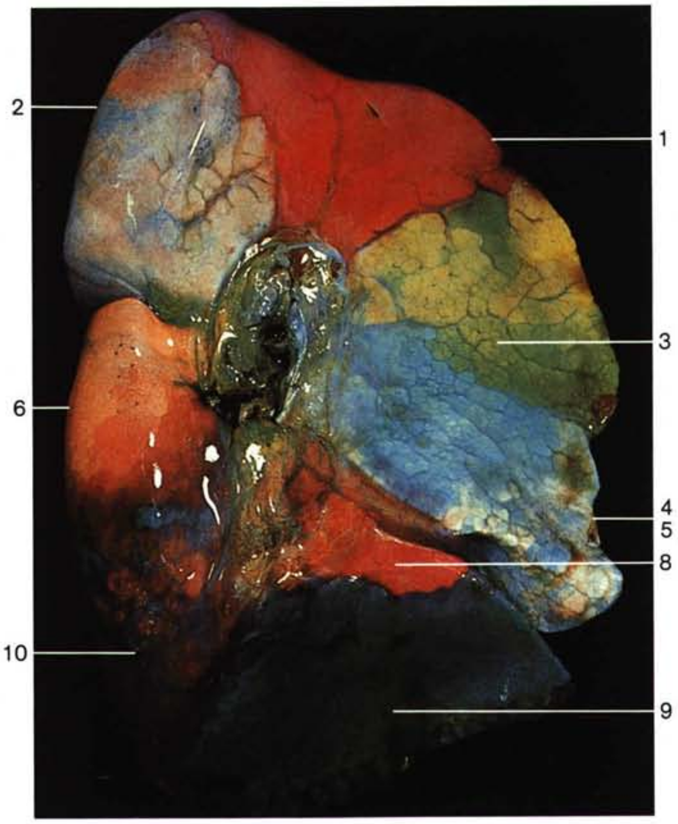
2
1
3
6
4
5
8
10
9

Left lung (medial aspect).

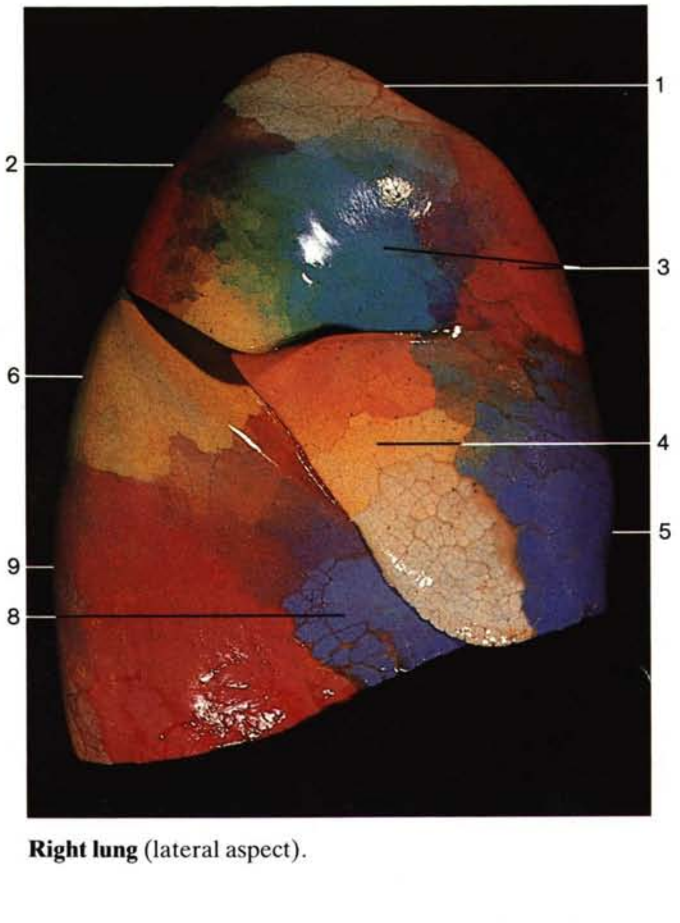

Right lung (lateral aspect).

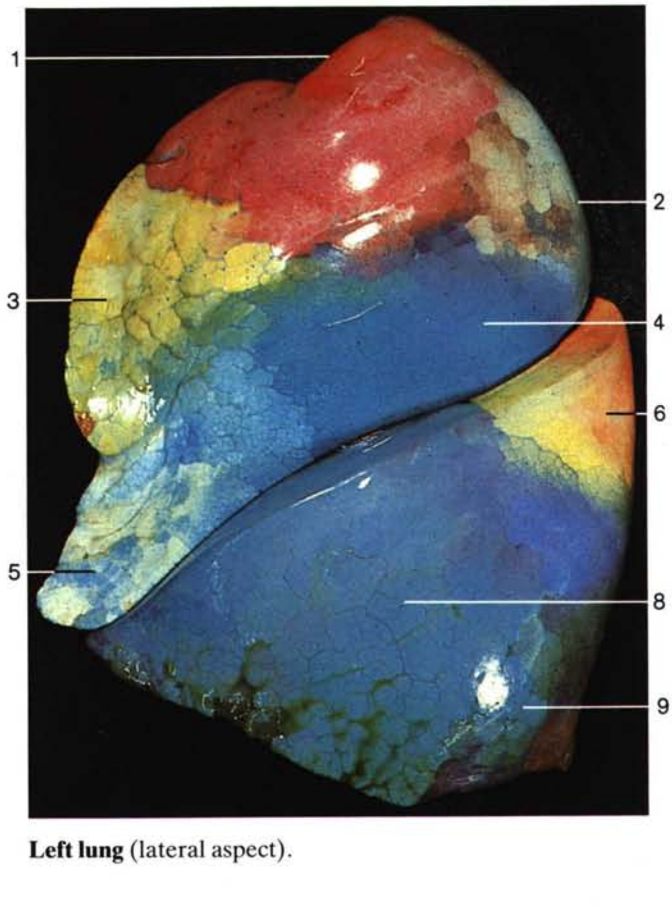

Left lung (lateral aspect).

The bronchopulmonary segments of the lungs are differentiated by the various colors. Notice that there is no segment in the left lung that corresponds to the 7th segment of the right lung. Compare with the schematic drawing on the facing page.

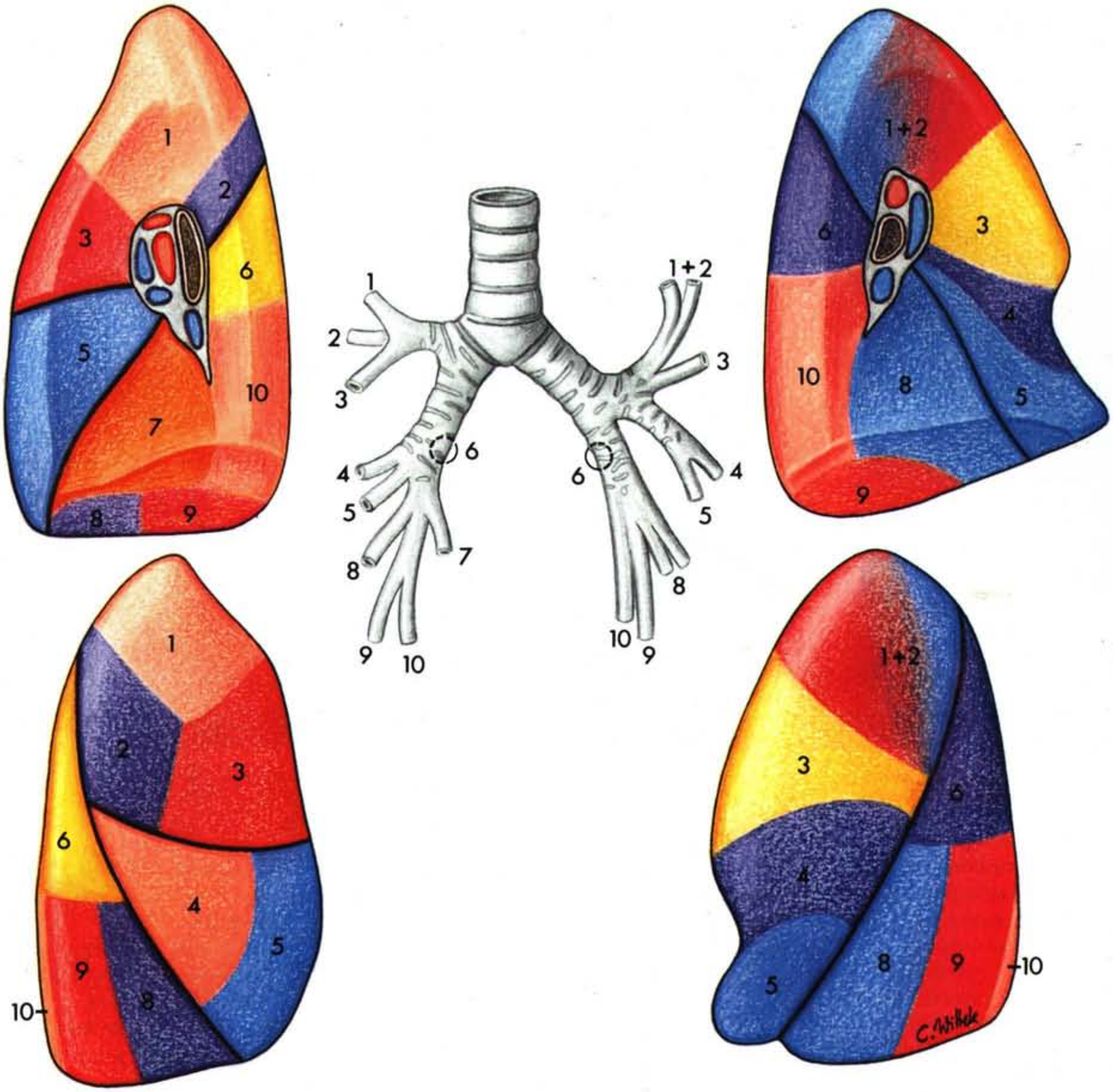

Distribution of bronchopulmonary segments of the lungs and their relation to the bronchial tree (after J. F. Huber).
The bronchopulmonary segments are morphologically and functionally separate independent respiratory units of the lung tissue. Each segment is surrounded by connective tissue which is continuous with the visceral pleura. The segmental bronchi in a segment are central, closely accompanied by branches of the pulmonary arteries whereas the tributaries of the pulmonary veins run **between** the segments. Thus, the veins serve two adjacent segments which drain for the most part into more than one vein. A bronchopulmonary segment is therefore not a complete vascular unit, but segmentation is the result of a specific architecture of the lung vasculature.

	Right lung	
1	Apical segment	Upper lobe bronchus
2	Posterior segment	
3	Anterior segment	
4	Lateral segment	Middle lobe bronchus
5	Medial segment	
6	Superior (apical) segment	Lower lobe bronchus
7	Medial basal segment	
8	Anterior basal segment	
9	Lateral basal segment	
10	Posterior basal segment	

	Left lung		
1+2	Apico-posterior segment	Superior division	Upper lobe bronchus
3	Anterior segment		
4	Superior lingular segment	Inferior division	
5	Inferior lingular segment		
6	Superior (apical) segment	Lower lobe bronchus	
7	Absent		
8	Anteromedial basal segment		
9	Lateral basal segment		
10	Posterior basal segment		

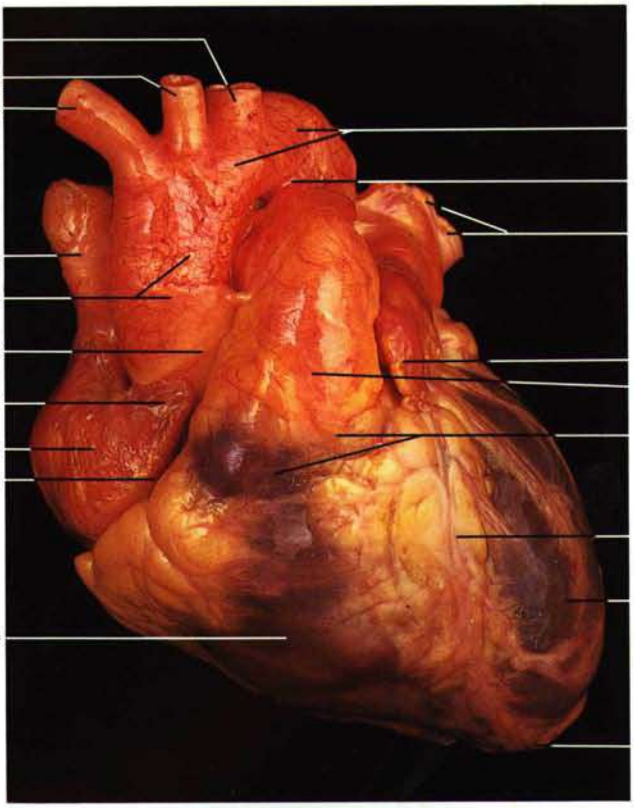

Heart of 30-year-old woman (anterior aspect).

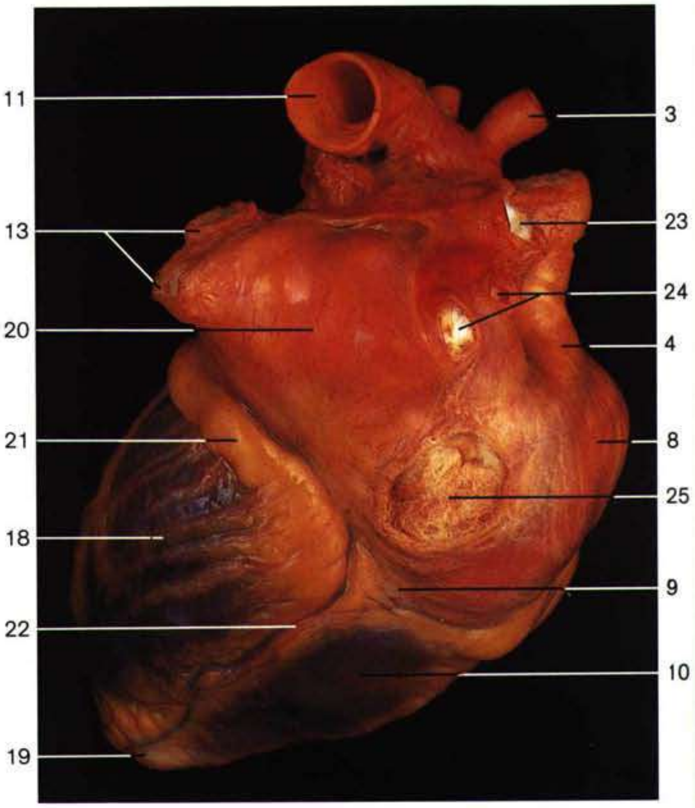

Heart of 30-year-old woman (oblique posterior view).

1 Left subclavian artery
2 Left common carotid artery
3 Brachiocephalic trunk
4 Superior vena cava
5 Ascending aorta
6 Bulb of the aorta
7 Right auricle
8 **Right atrium**
9 Coronary sulcus
10 **Right ventricle**
11 Aortic arch
12 Ligamentum arteriosum
13 Left pulmonary veins
14 Left auricle
15 **Pulmonary trunk**
16 Sinus of pulmonary trunk
17 Anterior interventricular sulcus
18 **Left ventricle**
19 Apex of the heart
20 **Left atrium**
21 Epicardial fat overlying coronary sinus
22 Posterior interventricular sulcus
23 Right pulmonary artery
24 Right pulmonary veins
25 Inferior vena cava

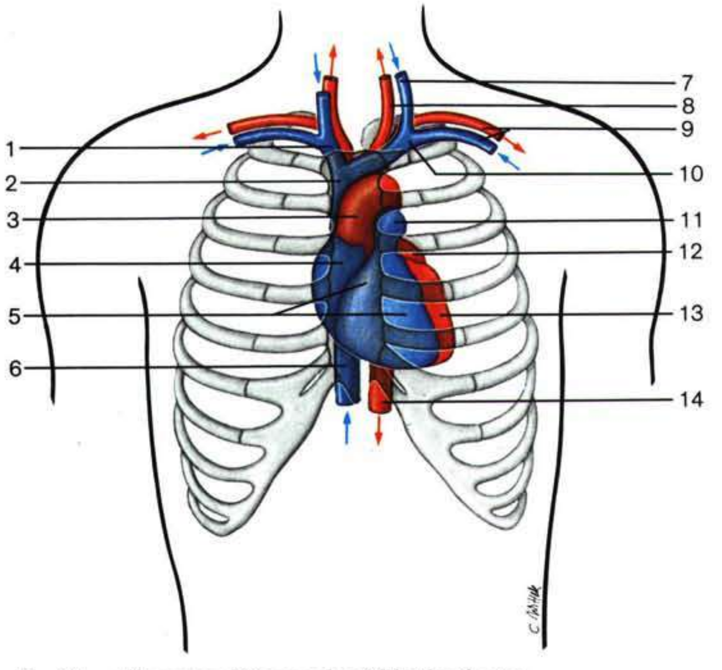

Position of heart and its vessels within the thorax.
(Schematic drawing.)

1 Right brachiocephalic vein
2 Superior vena cava
3 Ascending aorta
4 Right atrium
5 Right ventricle
6 Inferior vena cava
7 Left internal jugular vein
8 Left common carotid artery
9 Left axillary artery and vein
10 Left brachiocephalic vein
11 Pulmonary trunk
12 Left auricle
13 Left ventricle
14 Descending aorta

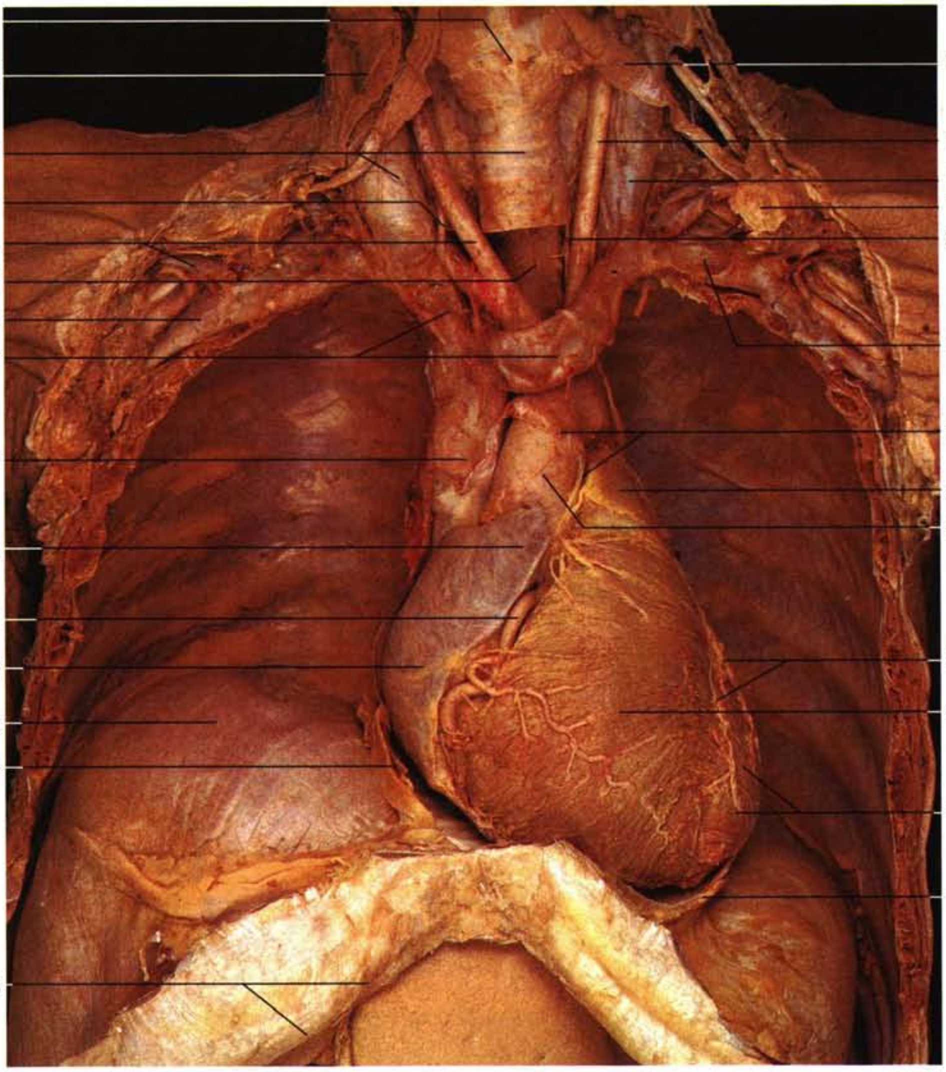

Heart and related vessels in situ (anterior aspect). Anterior thoracic wall, pericardium and epicardium have been removed; trachea divided.

1 Larynx (thyroid cartilage)
2 Sternocleidomastoid muscle (divided)
3 Trachea (divided) and right internal jugular vein
4 Vagus nerve
5 Right common carotid artery and cephalic vein
6 Esophagus
7 Right axillary vein
8 Right and left brachiocephalic veins
9 **Superior vena cava**
10 Right auricle
11 Right **coronary artery**
12 **Right atrium**
13 Diaphragm
14 Pericardium (cut edges)
15 Costal margin
16 Omohyoid muscle
17 Left **common carotid artery**
18 Left internal jugular vein
19 Clavicle (divided)
20 Left recurrent laryngeal nerve
21 **Subclavian vein**
22 Pericardial reflection
23 **Pulmonary trunk**
24 **Ascending aorta**
25 Anterior interventricular sulcus and anterior interventricular branch of left coronary artery
26 **Right ventricle**
27 **Left ventricle**
28 Aortic valve
29 Tricuspid or right atrioventricular valve
30 Inferior vena cava
31 Pulmonary veins
32 Pulmonary valve
33 Left atrioventricular (bicuspid or mitral) valve

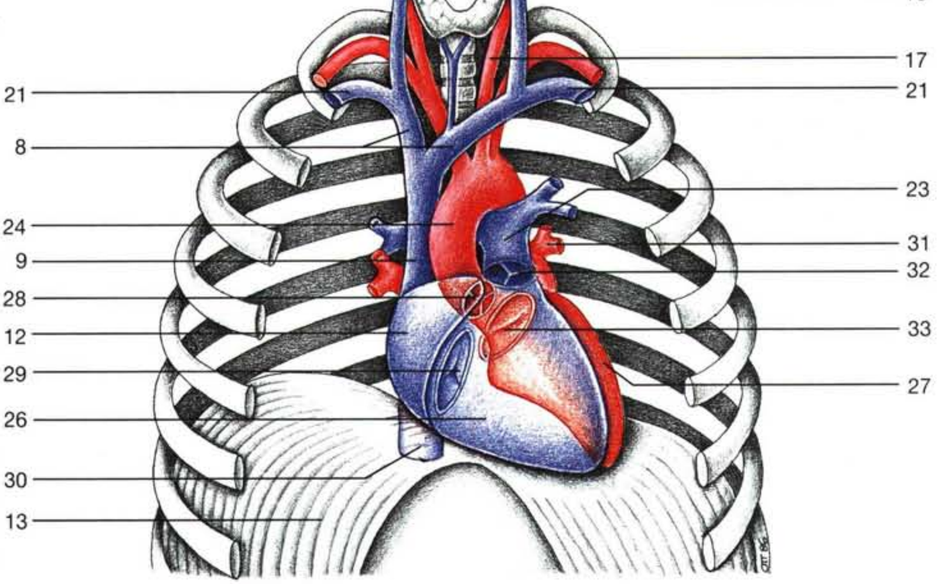

Heart in situ. Position of valves (anterior aspect). (Schematic drawing.)

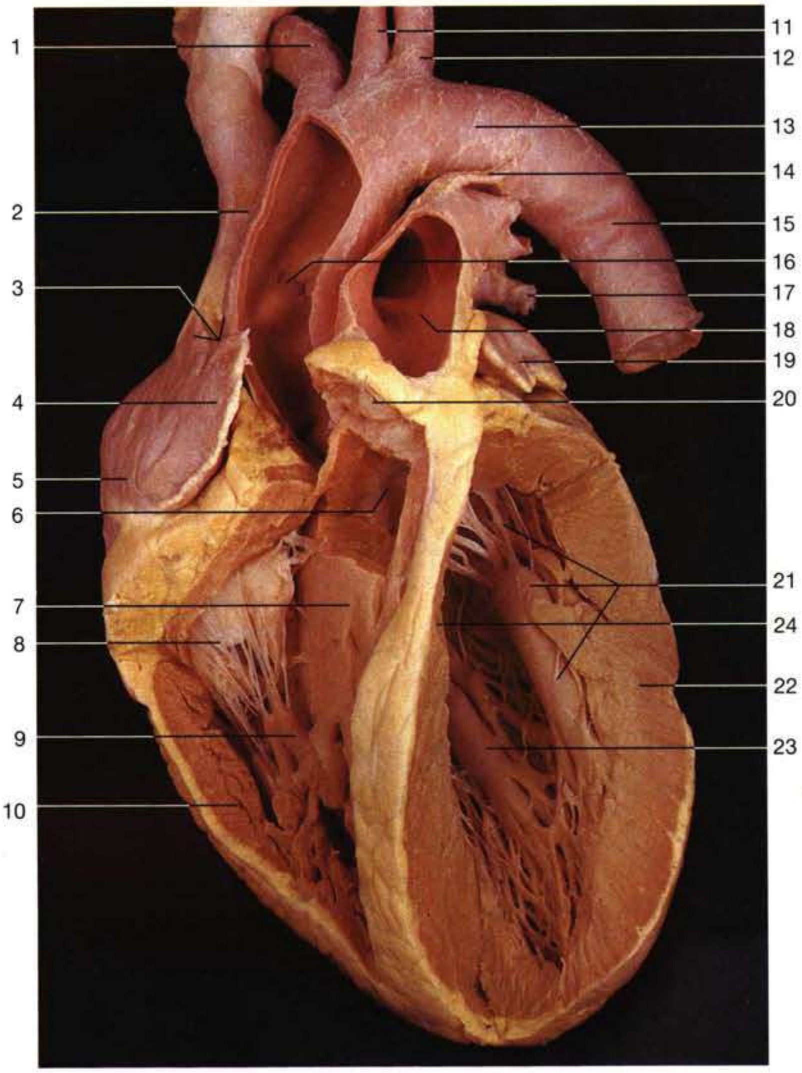

Anterior aspect of the heart. The anterior walls of the ventricles and of the aorta and pulmonary trunk have been fenestrated to show the aortic valve.

1 Brachiocephalic trunk
2 Superior vena cava
3 Sulcus terminalis
4 Right auricle
5 Right atrium
6 **Aortic valve**
7 Conus arteriosus (interventricular septum)
8 **Right atrioventricular (tricuspid) valve**
9 **Anterior papillary muscle**
10 Myocardium of right ventricle
11 Left common carotid artery
12 Left subclavian artery
13 **Aortic arch**
14 Ligamentum arteriosum (remnant of ductus arteriosus)
15 Thoracic aorta (descending aorta)
16 **Ascending aorta**
17 Left pulmonary vein
18 **Pulmonary trunk**
19 Left auricle
20 Pulmonic valve
21 **Anterior papillary muscle** with chordae tendineae
22 Myocardium of left ventricle
23 Posterior papillary muscle
24 Interventricular septum
25 Right and left brachiocephalic veins
26 **Chordae tendineae**
27 Papillary muscles of right ventricle
28 Left atrium
29 Infundibulum
30 Anterior papillary muscle of left ventricle
31 Left atrioventricular (bicuspid or mitral) valve and chordae tendineae
32 Apex of heart

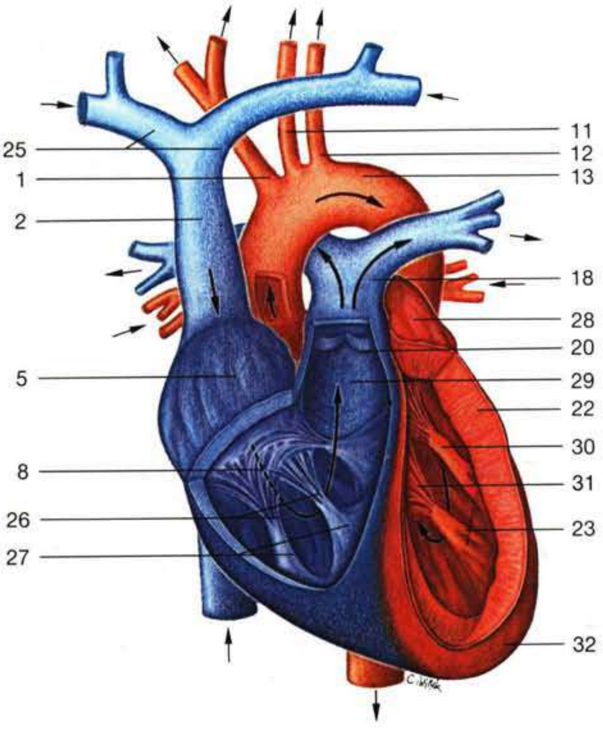

Circulation within the heart. (Schematic drawing.) The arrows indicate the direction of the blood flow through the heart. Blue = right heart; red = left heart.

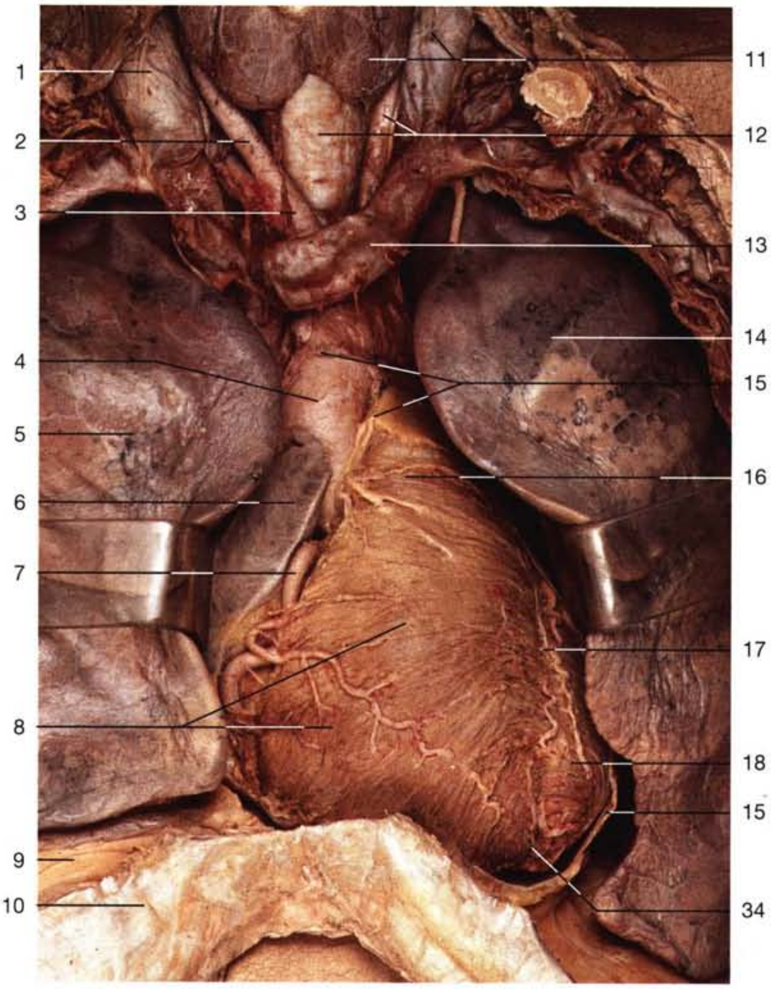

Heart in situ. Myocardium and coronary arteries (anterior aspect).

1 Internal jugular vein
2 Common carotid artery
3 Brachiocephalic trunk
4 **Ascending aorta**
5 Right lung
6 Right auricle
7 **Right coronary artery**
8 **Myocardium of right ventricle**
9 Diaphragm
10 Costal margin
11 Thyroid gland and internal jugular vein
12 Trachea and left common carotid artery
13 Left brachiocephalic vein
14 Left lung
15 **Pericardium** (cut edge)
16 Pulmonary trunk
17 Anterior interventricular artery
18 **Myocardium of left ventricle**
19 **Muscular vortex** (right ventricle)
20 Posterior interventricular sulcus
21 Anterior interventricular sulcus
22 **Muscular vortex** (left ventricle)
23 Aortic arch
24 Left atrium
25 Coronary sinus
26 Superior vena cava
27 Right pulmonary vein
28 Right atrium
29 Inferior vena cava
30 Coronary sulcus
31 Myocardium of left ventricle
32 Left pulmonary artery
33 Left pulmonary vein
34 Apex of heart

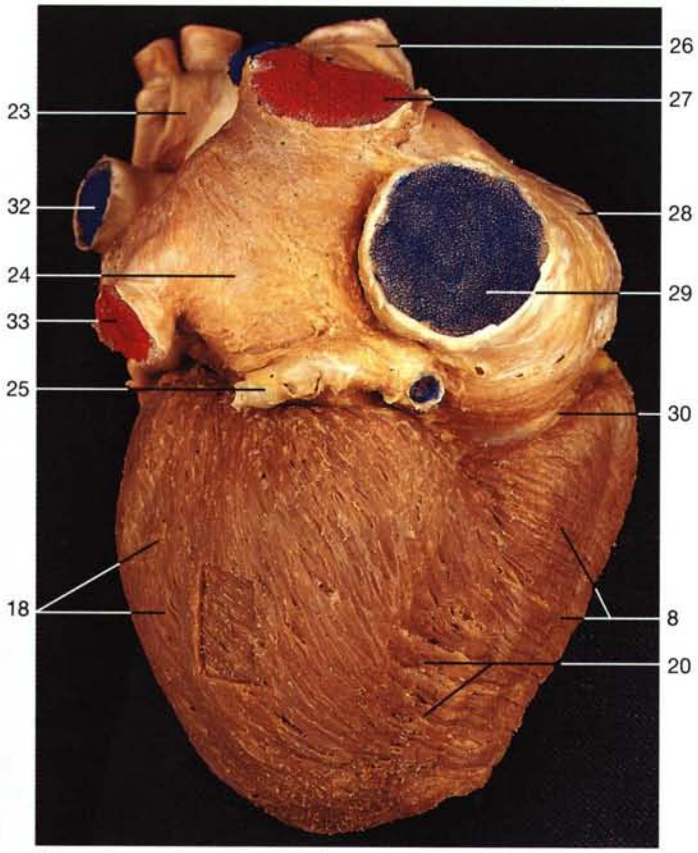

Heart (posterior aspect). The myocardium of the left ventricle has been fenestrated to show the muscle fiber bundles of the deeper layer with their more circular course.

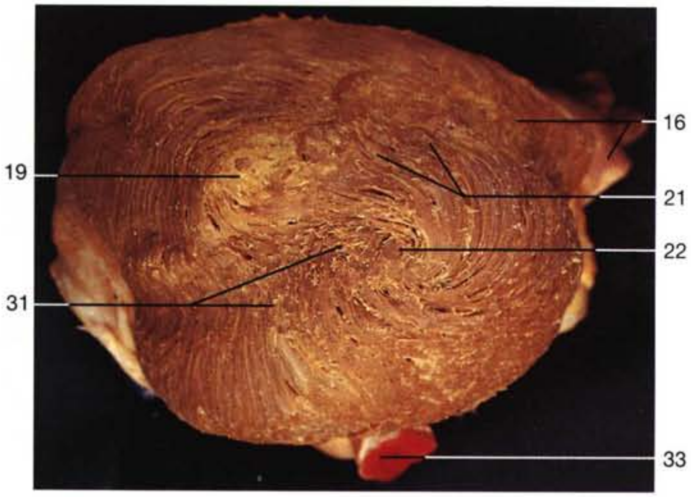

Vortex of cardiac muscle fibers (from below).

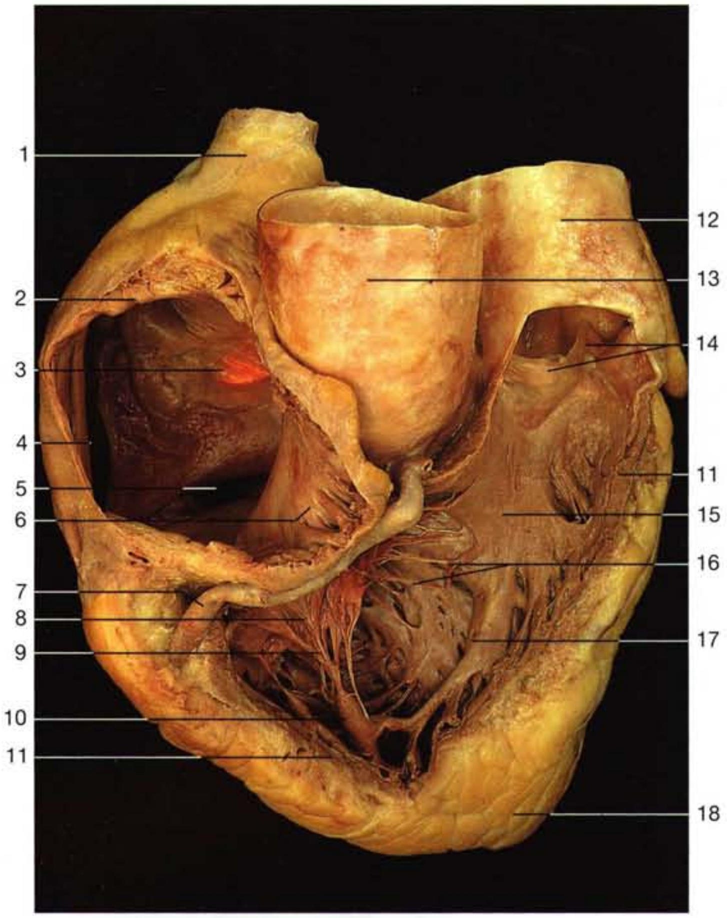

Right heart (anterior aspect). Anterior wall of right atrium and ventricle removed.

1	Superior vena cava
2	Crista terminalis
3	Fossa ovalis
4	Opening of inferior vena cava
5	Opening of coronary sinus
6	Right auricle
7	Right coronary artery and coronary sulcus
8	**Anterior cusp** of **tricuspid valve**
9	Chordae tendineae
10	**Anterior papillary muscle**
11	Myocardium
12	Pulmonary trunk
13	Ascending aorta
14	**Pulmonic valve**
15	Conus arteriosus (interventricular septum)
16	Septal papillary muscles
17	Septomarginal or moderator band
18	Apex of heart
19	Left auricle
20	**Aortic valve**
21	**Left ventricle**
22	Pulmonary veins
23	Position of fossa ovalis
24	**Left atrium**
25	**Left atrioventricular** (bicuspid or **mitral**) **valve**
26	Coronary sinus
27	Left coronary artery
28	Posterior papillary muscle
29	Left subclavian artery
30	Descending aorta
31	Left pulmonary artery

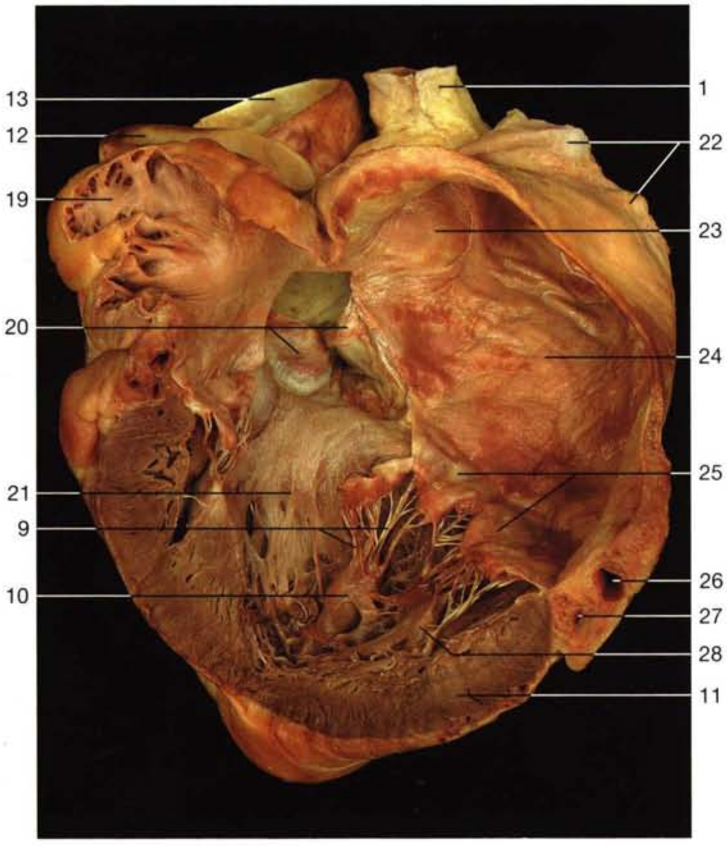

Left heart. Left atrium and ventricle opened. Aortic and mitral valve cut open.

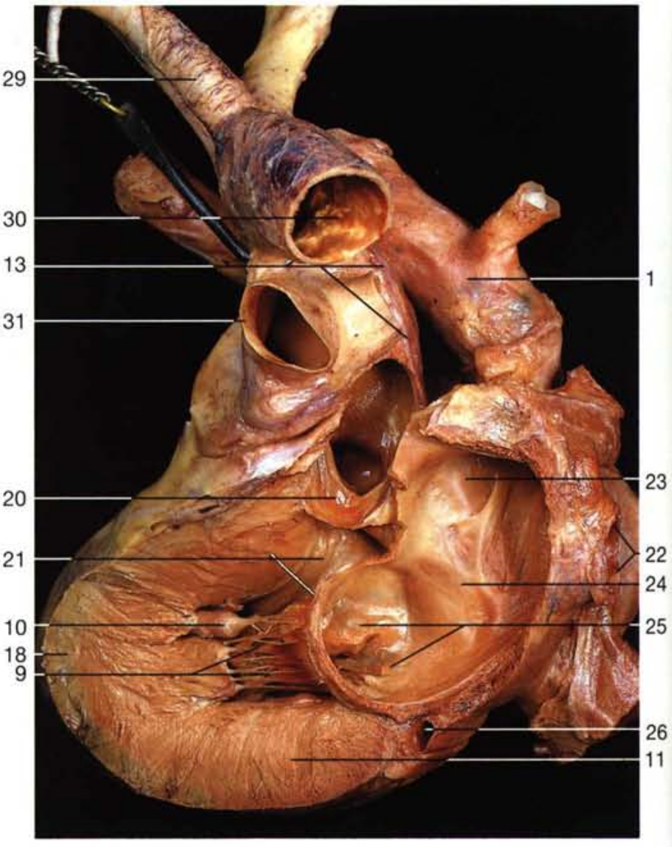

Left heart. Aortic valve cut open. Left atrium and ventricle opened. The atrial wall was cut near the anulus fibrosus of the mitral valve.

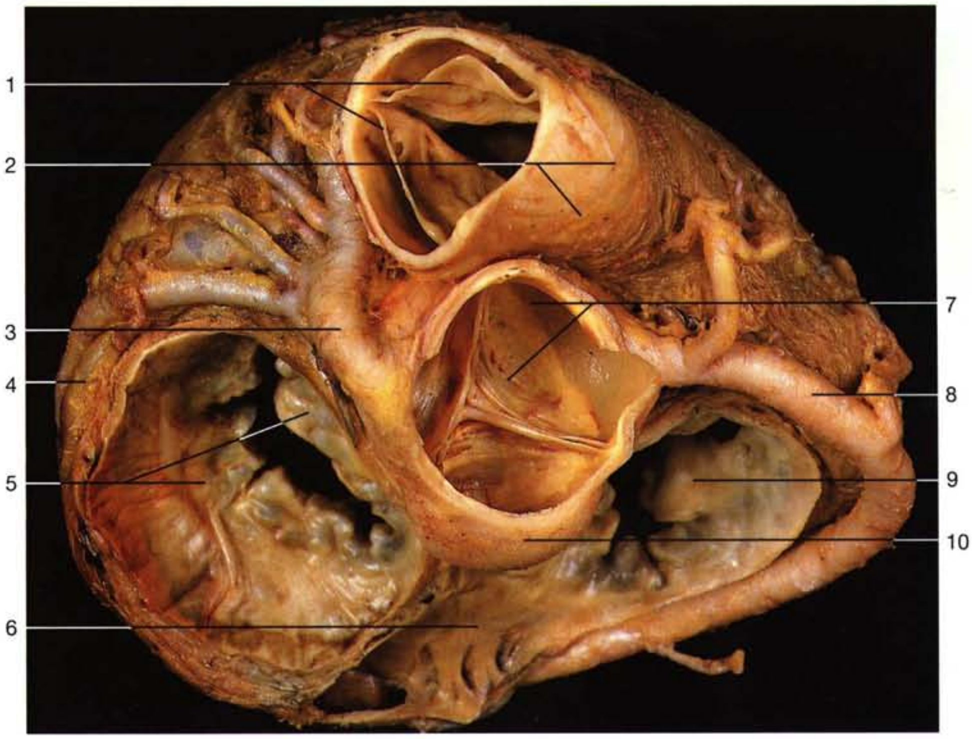

Valves of heart (superior aspect). Left and right atrium removed. Dissection of coronary arteries. Above: anterior wall of the heart.

1 **Pulmonic valve**
2 Sinus of pulmonary trunk
3 **Left coronary artery**
4 Great cardiac vein
5 **Left atrioventricular (mitral) valve**
6 **Coronary sinus**
7 **Aortic valve**
8 **Right coronary artery**
9 **Right atrioventricular (tricuspid) valve**
10 Bulb of aorta
11 Anterior semilunar cusp of pulmonic valve
12 Left semilunar cusp of pulmonic valve
13 Right semilunar cusp of pulmonic valve
14 Left semilunar cusp of aortic valve
15 Right semilunar cusp of aortic valve
16 Posterior semilunar cusp of aortic valve
17 Right atrium
18 Anterior cusp of tricuspid valve
19 **Chordae tendineae**
20 Trabeculae carneae
21 Interventricular septum
22 Septal cusp of tricuspid valve
23 **Anterior papillary muscle**
24 Myocardium of right ventricle

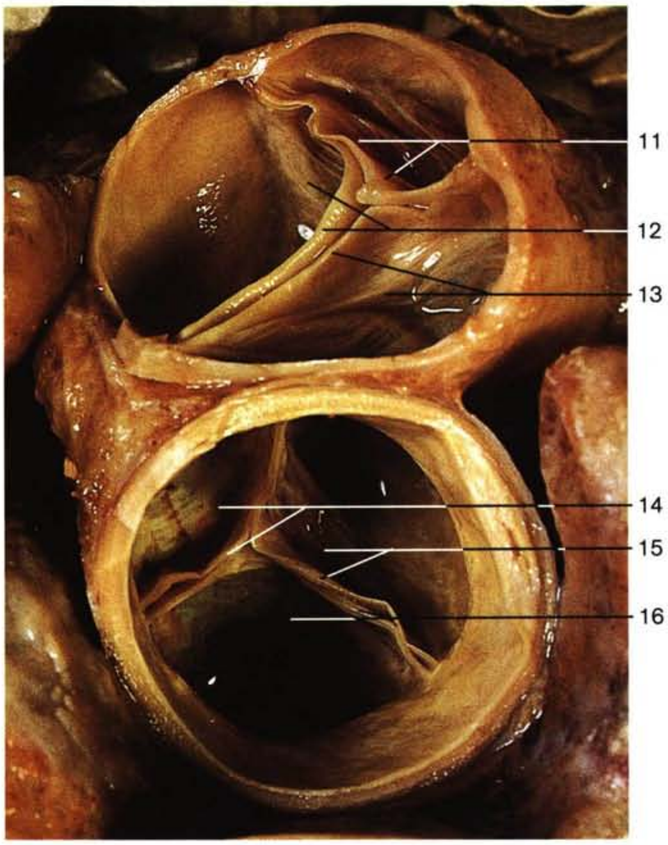

Pulmonic and aortic valves (from above). Anterior wall of the heart at the top. Both valves are closed.

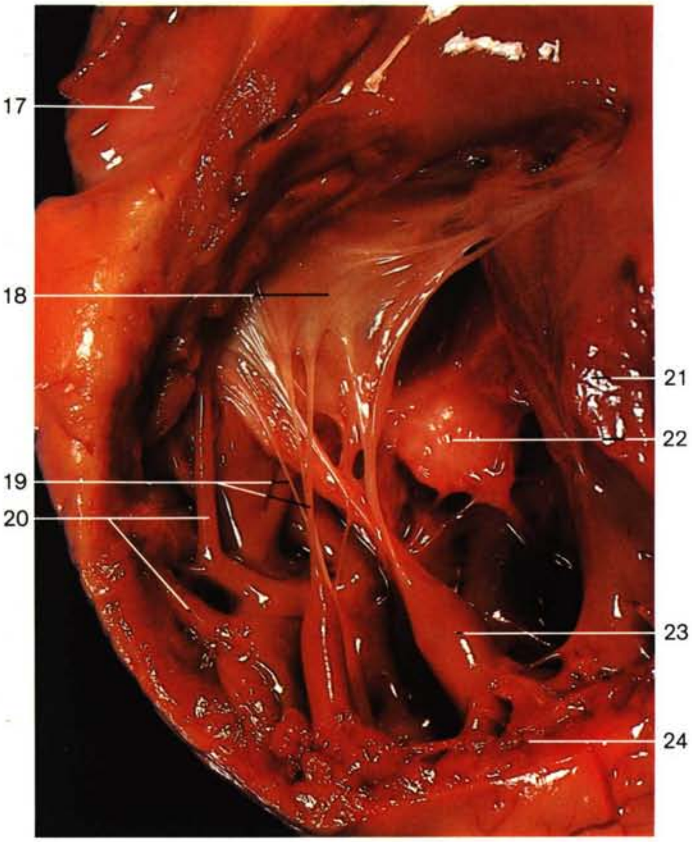

Right atrioventricular (tricuspid) valve (anterior aspect after removal of the anterior wall of the right ventricle).

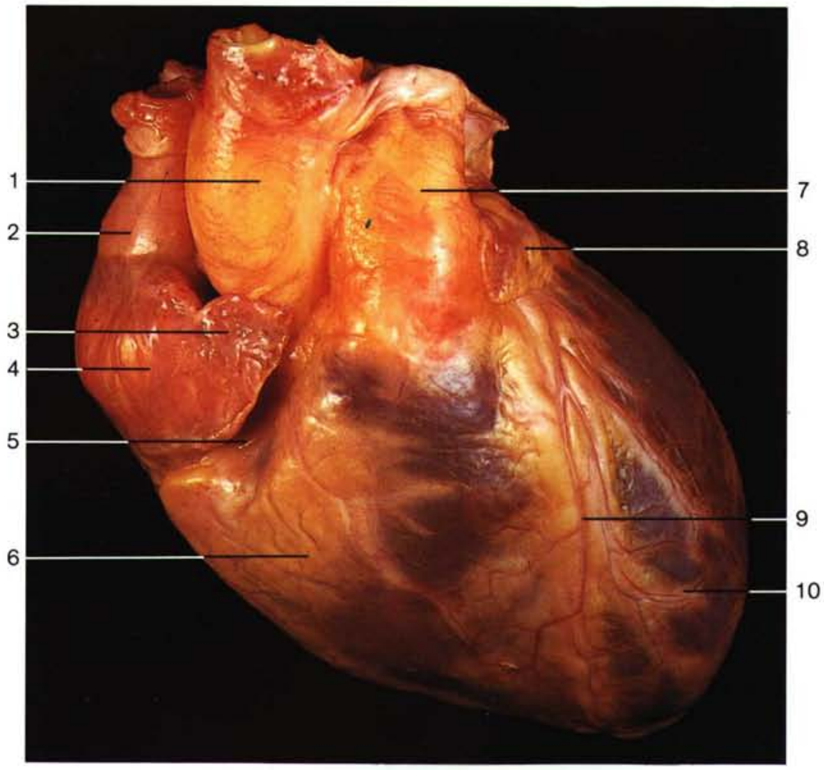

Heart, fixed in **diastole** (anterior aspect). The ventricles are relaxed, atria contracted.

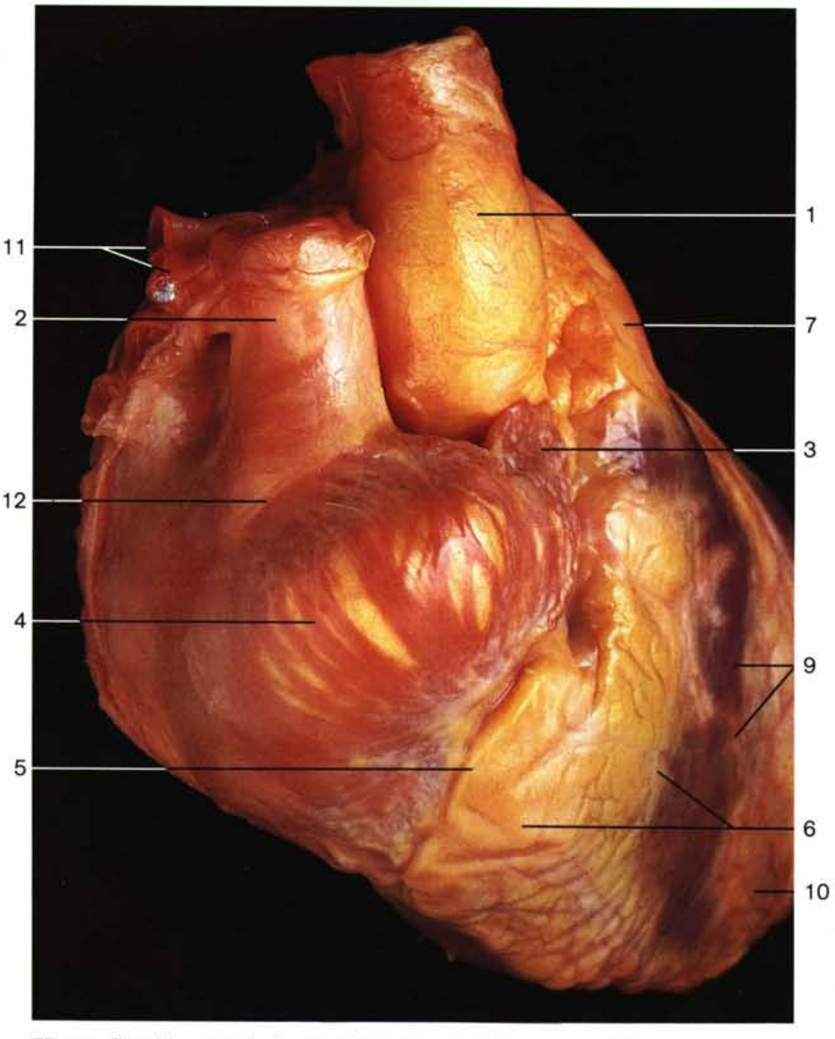

Heart, fixed in **systole** (anterolateral aspect). The ventricles are contracted, atria dilated.

1 Ascending aorta
2 Superior vena cava
3 Right auricle
4 **Right atrium**
5 Coronary sulcus
6 **Right ventricle**
7 Pulmonary trunk
8 Left auricle
9 Anterior interventricular sulcus
10 **Left ventricle**
11 Right pulmonary artery
12 Sulcus terminalis with sinuatrial node
13 Line indicating plane of position of valves
14 Myocardium of right atrium
15 Inferior vena cava
16 Valve of pulmonary trunk
17 Tricuspid valve
18 Myocardium of right ventricle

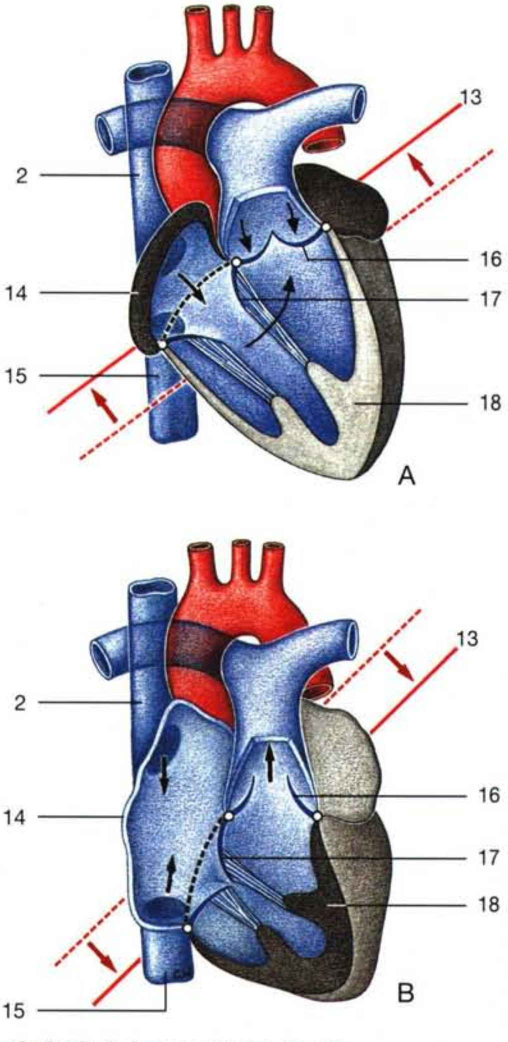

Morphological changes during heart movements. Note the changes in positon of the valves (arrows). Contracted portions of heart are indicated in red.

A. **Diastole,** muscles of the ventricles relaxed; atrioventricular valves open, semilunar valves closed.

B. **Systole,** muscles of ventricles contracted; atrioventricular valves closed, semilunar valves open.

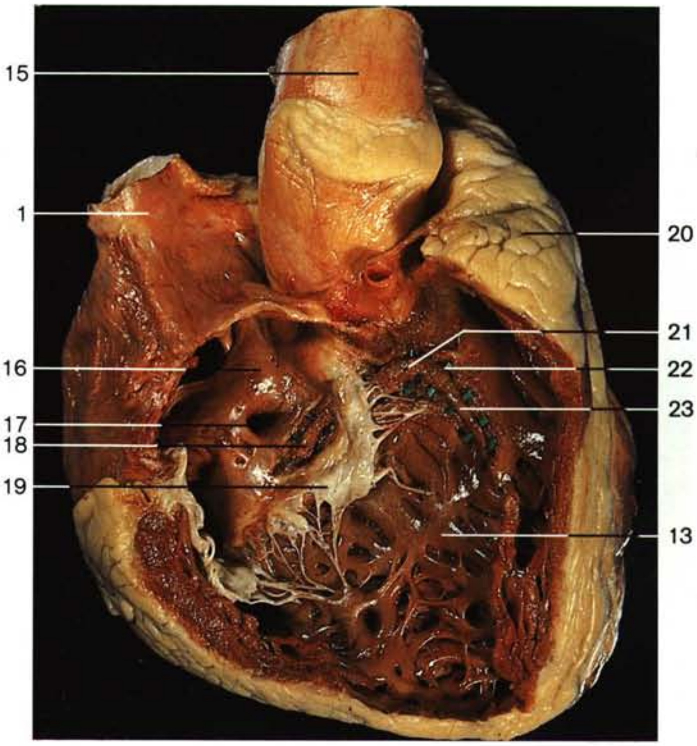

Right ventricle, dissection of **atrioventricular node, atrioventricular bundle (bundle of His)** and right limb or bundle branch (probes).

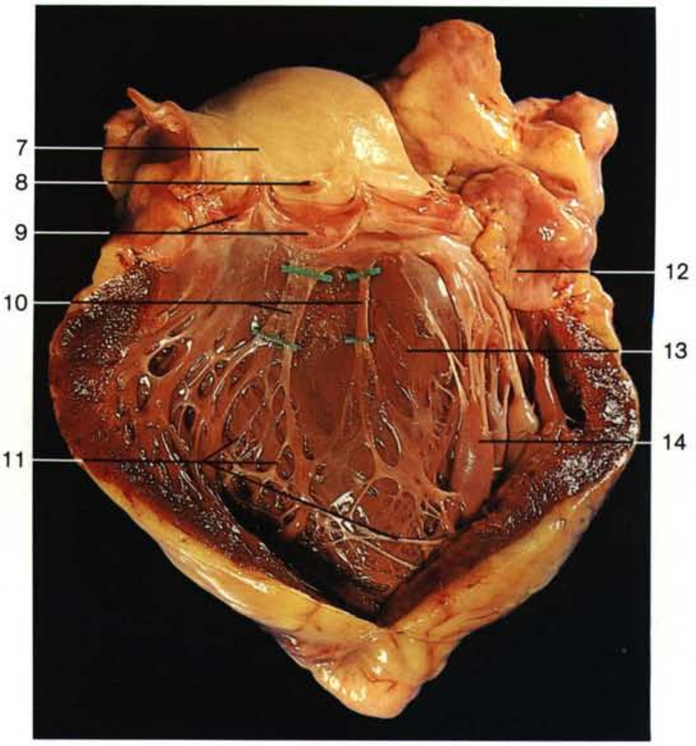

Left ventricle, dissection of the left limb or bundle branch of conducting system (probes).

1 Superior vena cava
2 Sulcus terminalis
3 Bulb of aorta
4 Sinuatrial node (arrows)
5 Muscle fiber bundles of right atrium
6 Coronary sulcus (with right coronary artery)
7 Aortic sinus
8 Entrance to left coronary artery
9 Aortic valve
10 Branches of **left bundle branch**
11 Purkinje fibers
12 Left auricle
13 Interventricular septum
14 Papillary muscles
15 Ascending aorta
16 Right atrium
17 Opening of coronary sinus
18 **Atrioventricular node**
19 Septal cusp of tricuspid valve
20 Pulmonary trunk
21 **Atrioventricular bundle** (bundle of His)
22 Bifurcation of atrioventricular bundle
23 **Right bundle branch**
24 Inferior vena cava
25 Left atrium
26 Left bundle branch
27 Papillary muscles with Purkinje fibers

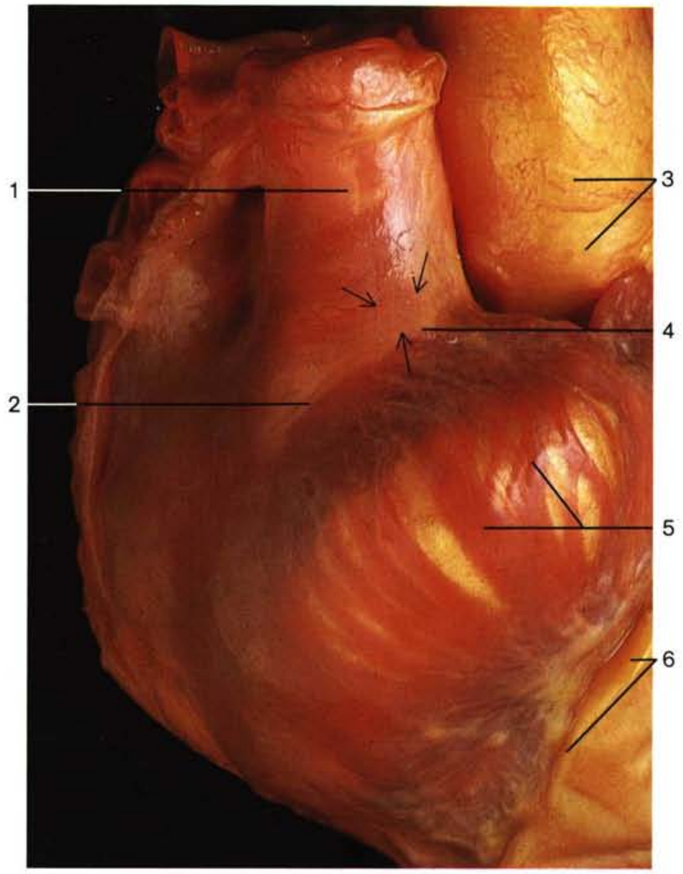

Right atrium, anterior wall, showing the location of the **sinuatrial node** (arrows).

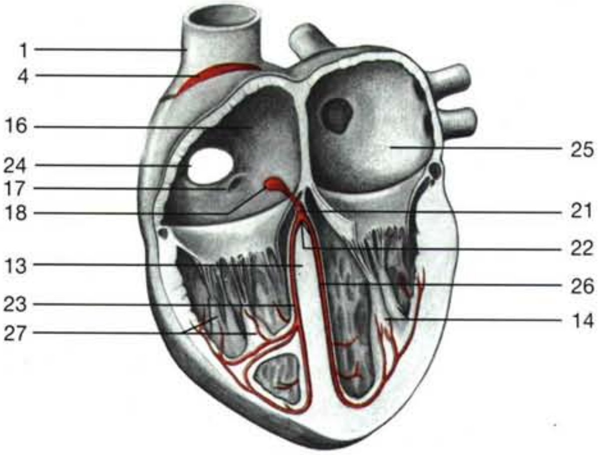

Conducting system of the heart. (Schematic drawing.)

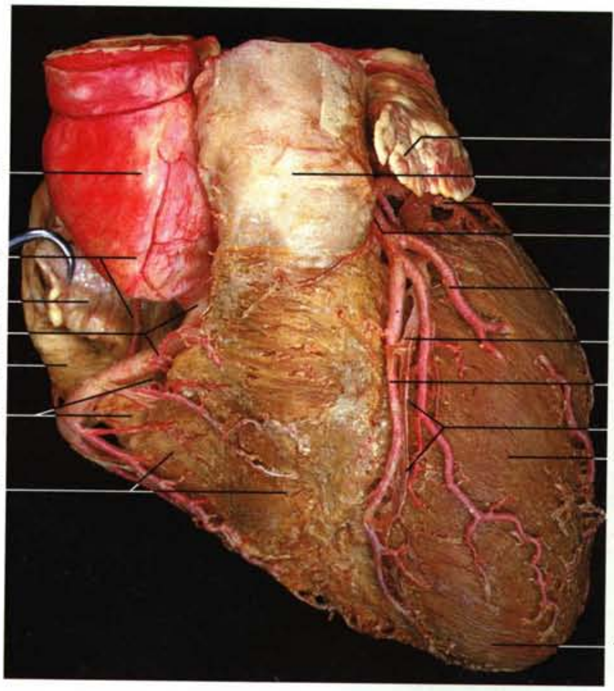

Coronary arteries (anterior aspect). The epicardium and subepicardial fatty tissue have been removed. The arteries have been injected with red resin from the aorta.

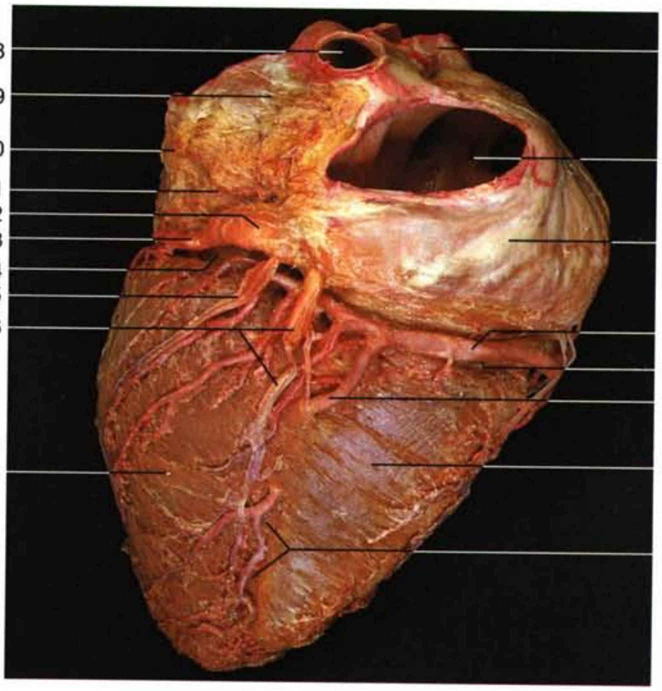

Right coronary artery and veins of the heart (dorsal aspect). The epicardium and subepicardial fatty tissue have been removed.

1 Ascending aorta
2 Aortic bulb and (in the above specimen) sinuatrial branch of right coronary artery
3 Right auricle
4 **Right coronary artery**
5 Right atrium
6 Coronary sulcus
7 Right ventricle
8 Left auricle
9 Pulmonary trunk
10 Circumflex branch of left coronary artery
11 **Left coronary artery**
12 Diagonal branch of left artery
13 Great cardiac vein
14 Anterior interventricular artery
15 Anterior interventricular sulcus
16 Left ventricle
17 Apex of heart
18 Right pulmonary vein
19 Left atrium
20 Left pulmonary veins
21 Oblique vein of left atrium (Marshall's vein)
22 **Coronary sinus**
23 **Great cardiac vein**
24 Coronary sulcus (posterior portion)
25 Posterior vein of left ventricle
26 **Middle cardiac vein**
27 Left pulmonary artery
28 Inferior vena cava
29 Right atrium
30 Posterior interventricular branch of right coronary artery
31 Posterior interventricular sulcus
32 Superior vena cava
33 Right marginal branch
34 Branch of sinuatrial node
35 Minimal cardiac veins
36 **Small cardiac vein**

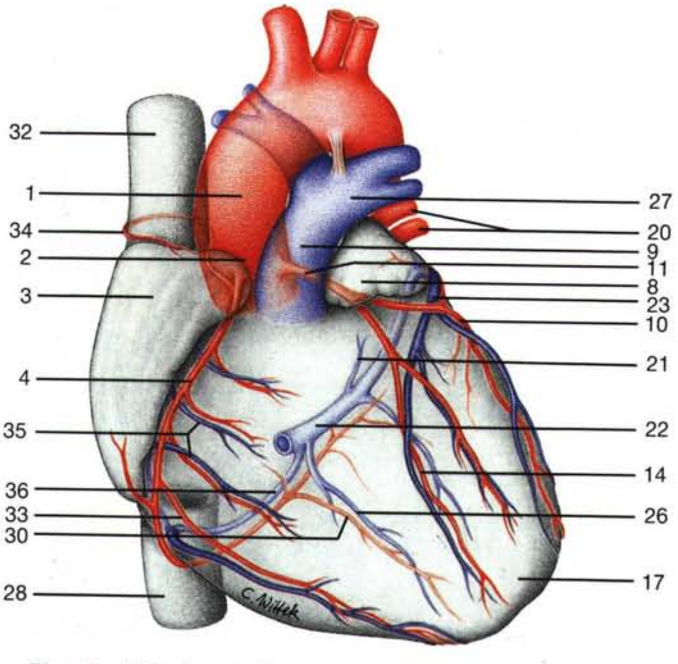

Vessels of the heart. Coronary arteries (red) and veins (blue) of the heart (anterior aspect).

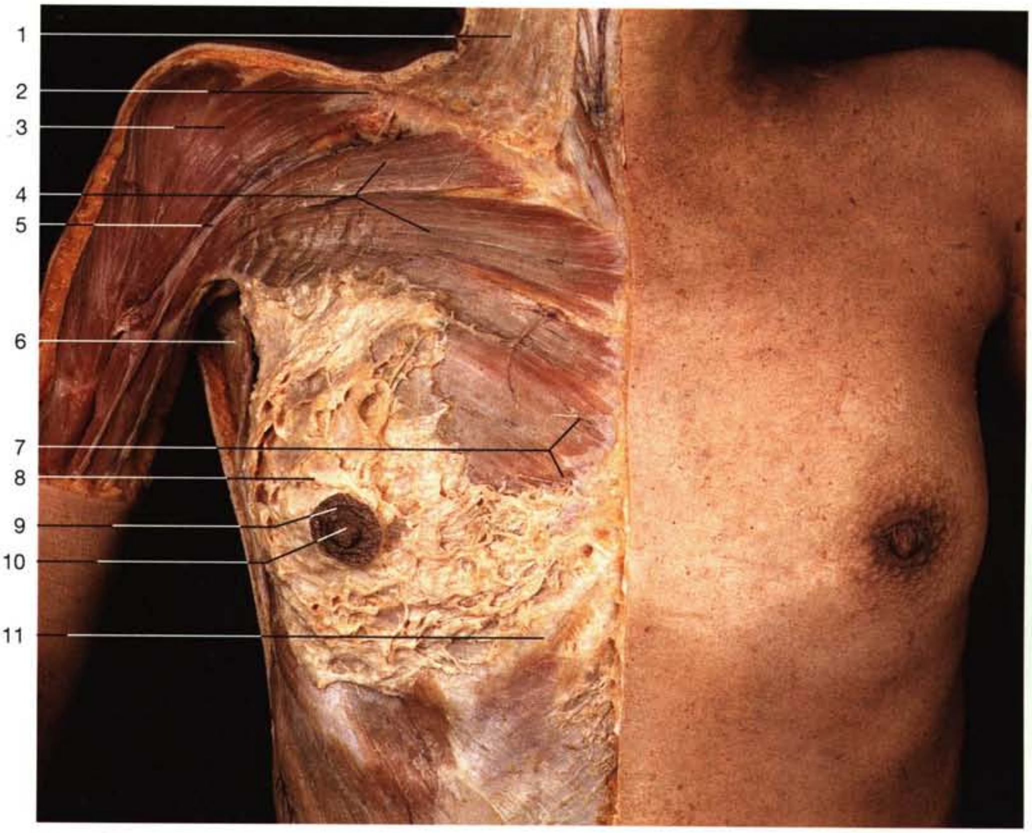

Dissection of mammary gland (anterior aspect).

1 Platysma muscle
2 Clavicle
3 Deltoid muscle
4 Pectoralis major muscle
5 Deltopectoral groove and cephalic vein
6 Latissimus dorsi muscle
7 Medial mamarian branches of intercostal nerves
8 Breast tissue
9 Areola
10 Nipple (papilla)
11 Costal margin
12 Pectoral fascia
13 Mammary gland
14 Serratus anterior muscle
15 Lactiferous sinus
16 Lactiferous ducts

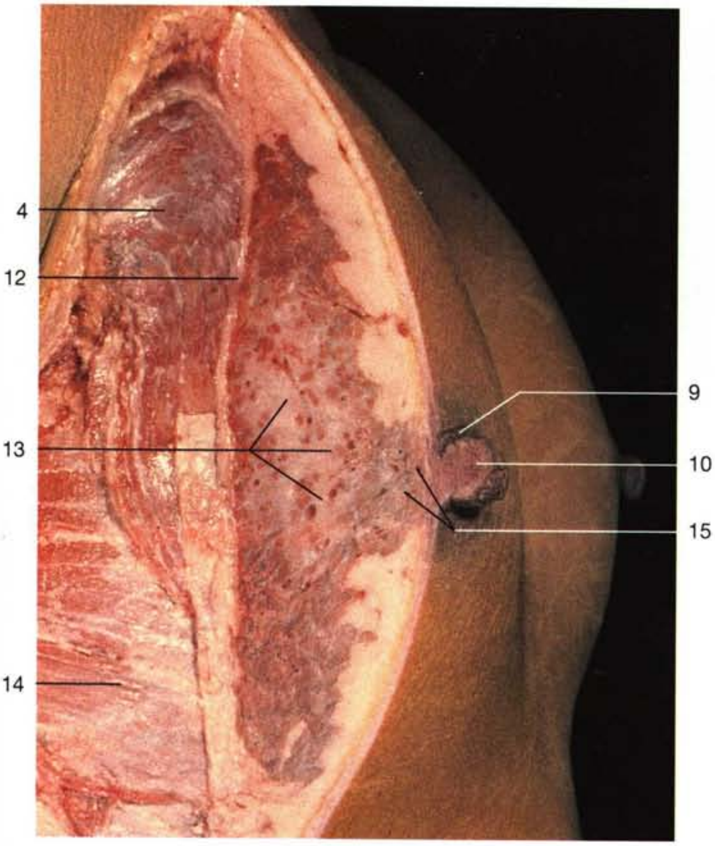

Mammary gland (sagittal section; pregnant female).

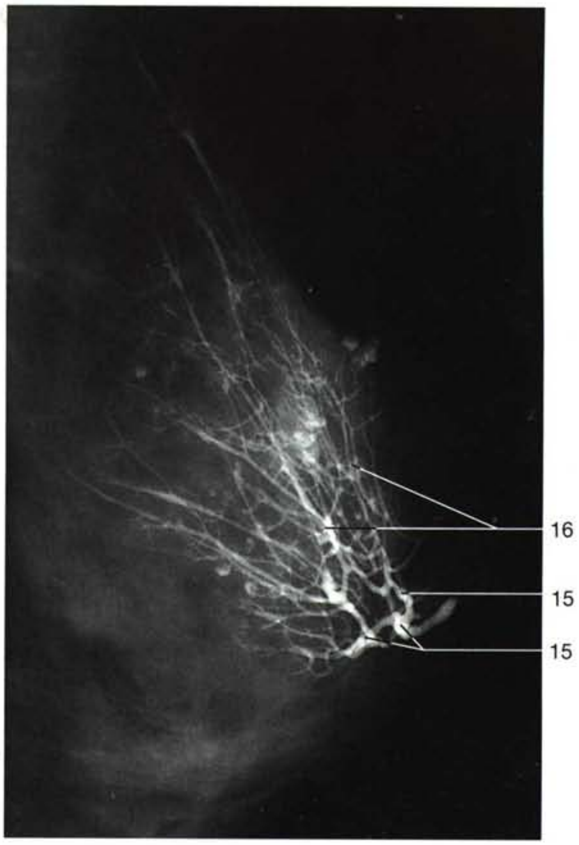

Galactogram of mammary gland (adult female). Notice multiple duct cysts.

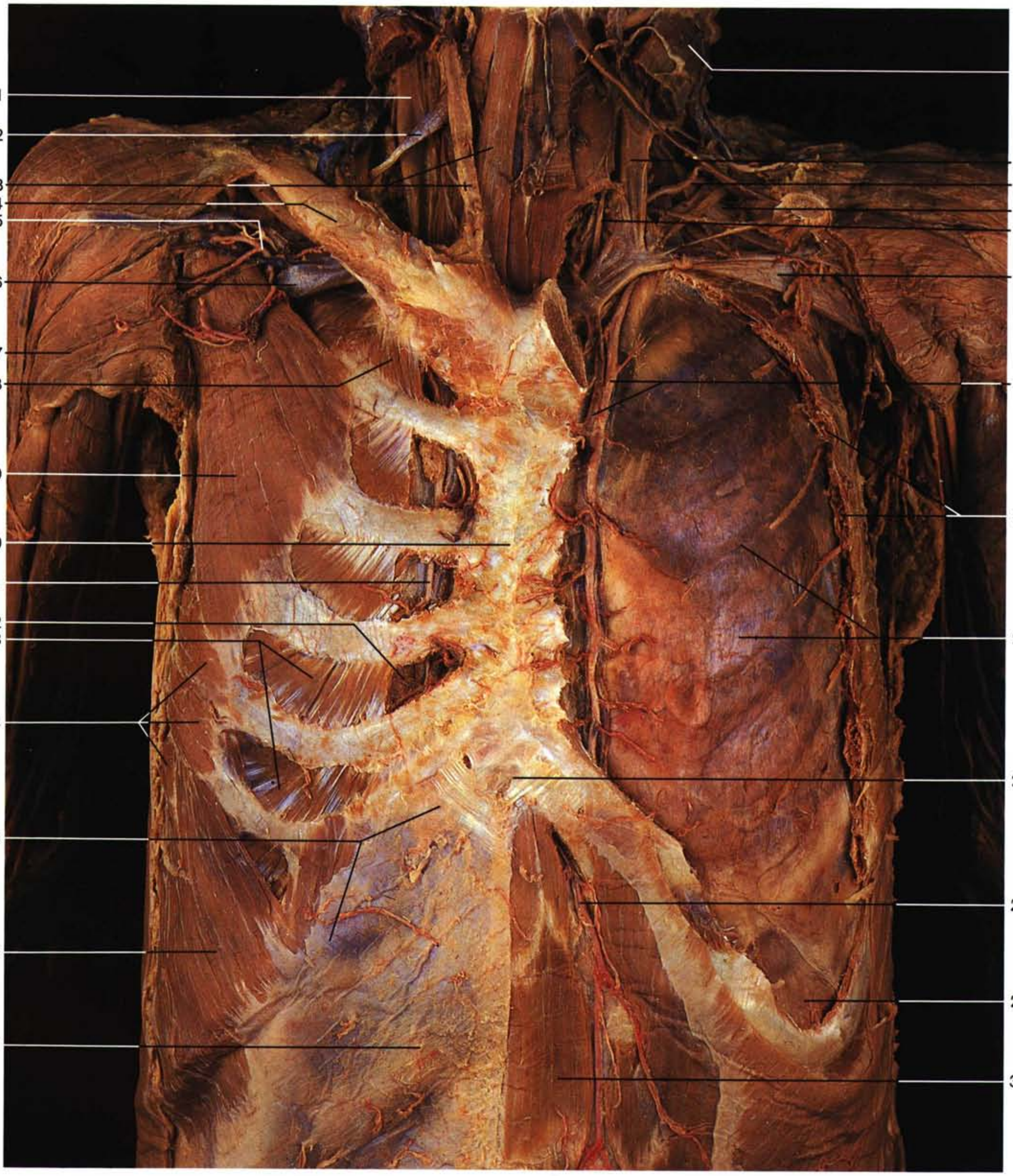

Thoracic wall and organs (ventral aspect). The left clavicle and ribs have been partially removed, and the right intercostal spaces have been opened to show the internal thoracic vein and artery.

1 Right internal jugular vein
2 Omohyoid muscle
3 Sternohyoid muscle and external jugular vein
4 Clavicle
5 Thoracoacromial artery
6 Right **subclavian vein**
7 Pectoralis major muscle
8 External intercostal muscle
9 Pectoralis minor muscle
10 Body of sternum

11 Right internal thoracic artery and vein
12 Fascicles of transversus thoracis muscle
13 **Internal intercostal muscles**
14 Serratus anterior muscle
15 Costal margin
16 External abdominal oblique muscle
17 Anterior sheath of rectus abdominis muscle
18 Sternocleidomastoid muscle
19 Left internal jugular vein
20 Transverse cervical artery

21 Brachial plexus
22 Vagus nerve
23 Left axillary vein
24 Left **internal thoracic artery and vein**
25 Ribs and thoracic wall (cut)
26 Costal pleura
27 Xiphoid process
28 **Superior epigastric artery**
29 Diaphragm
30 Rectus abdominis muscle

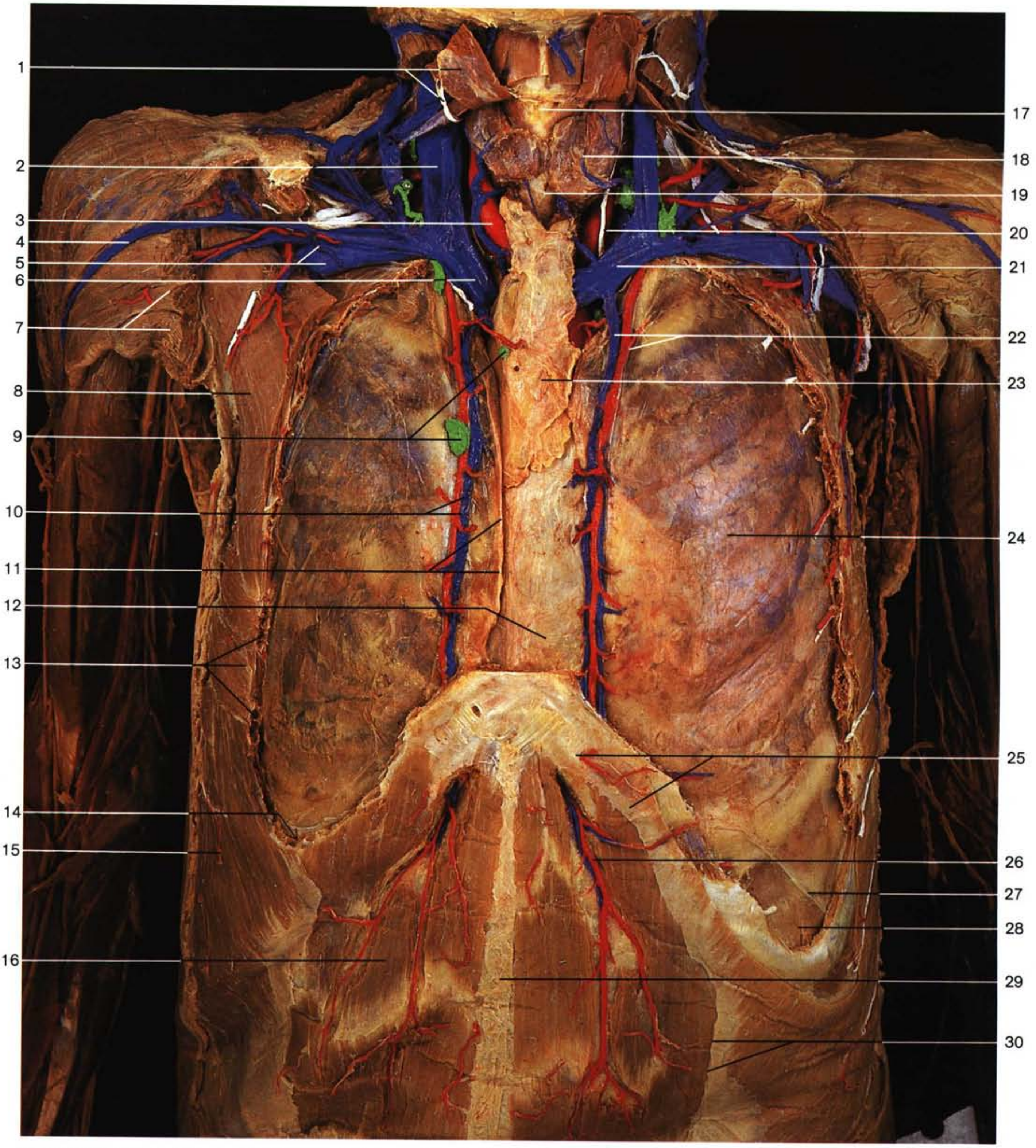

Thoracic organs, anterior mediastinum and pleura. Ribs, clavicle and sternum have been partly removed. Red = arteries; blue = veins; green = lymph vessels and nodes.

1 Sternothyroid muscle and its nerve (a branch of the ansa cervicalis)
2 Right **internal jugular vein**
3 Right common carotid artery
4 Cephalic vein
5 Right **subclavian vein**
6 **Right brachiocephalic vein**
7 Pectoralis major muscle (divided)
8 Pectoralis minor muscle (divided)
9 Parasternal lymph nodes
10 Internal thoracic artery and vein
11 Anterior margin of costal pleura
12 **Pericardium**
13 5th and 6th ribs (divided) and serratus anterior muscle
14 **Costodiaphragmatic recess**
15 External abdominal oblique muscle
16 Rectus abdominis muscle
17 Larynx (thyroid cartilage)
18 Thyroid gland
19 Trachea
20 Left vagus nerve
21 Left brachiocephalic vein
22 Left **internal thoracic artery and vein**
23 Thymus
24 **Costal pleura**
25 Costal margin
26 **Superior epigastric artery**
27 Margin of costal pleura
28 **Diaphragm**
29 **Linea alba**
30 Cut edge of anterior sheath of rectus abdominis muscle

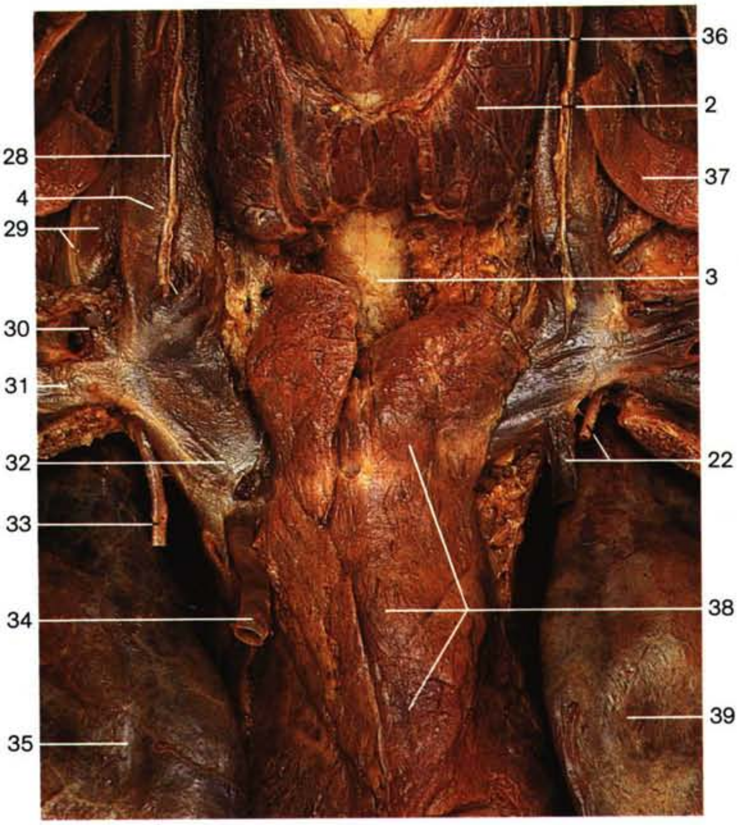

The **thymus** above the heart, showing its position and size.

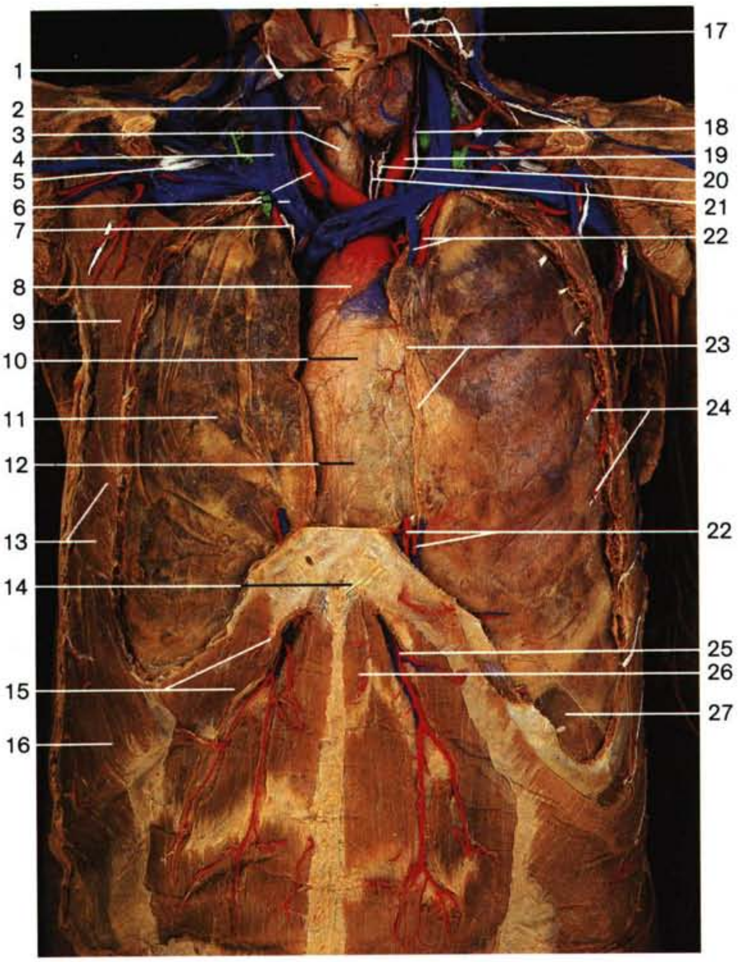

Thoracic organs (ventral aspect). The internal thoracic vessels have been removed, and the anterior margins of the pleura and lungs have been slightly reflected to display the **anterior** and **middle mediastinum,** including the **heart** and **great vessels.**

1 Larynx (thyroid cartilage)
2 Thyroid gland
3 **Trachea**
4 Internal jugular vein
5 Brachial plexus
6 Right brachiocephalic vein and common carotid artery
7 Right phrenic nerve
8 **Ascending aorta**
9 Pectoralis minor muscle (divided)
10 Pulmonary trunk (covered by pericardium)
11 **Costal pleura**
12 Pericardium and **heart**
13 Serratus anterior muscle
14 Xiphoid process
15 Costal margin
16 External abdominal oblique muslce
17 Sternothyroid muscle (divided and reflected)
18 Vagus nerve
19 Left common carotid artery
20 Left sympathetic trunk
21 Left recurrent laryngeal nerve
22 Left internal thoracic artery and vein (divided)
23 Margin of costal pleura
24 Intercostal nerves and vessels
25 Superior epigastric artery
26 Rectus abdominis muscle
27 **Diaphragm**
28 Ansa cervicalis
29 Phrenic nerve and scalenus anterior muscle
30 External jugular vein (divided)
31 Right subclavian vein
32 Right brachiocephalic vein
33 Internal thoracic artery (divided)
34 Internal thoracic vein (divided)
35 Right lung
36 Cricothyroid muscle
37 Omohyoid muscle
38 **Thymus**
39 Left lung

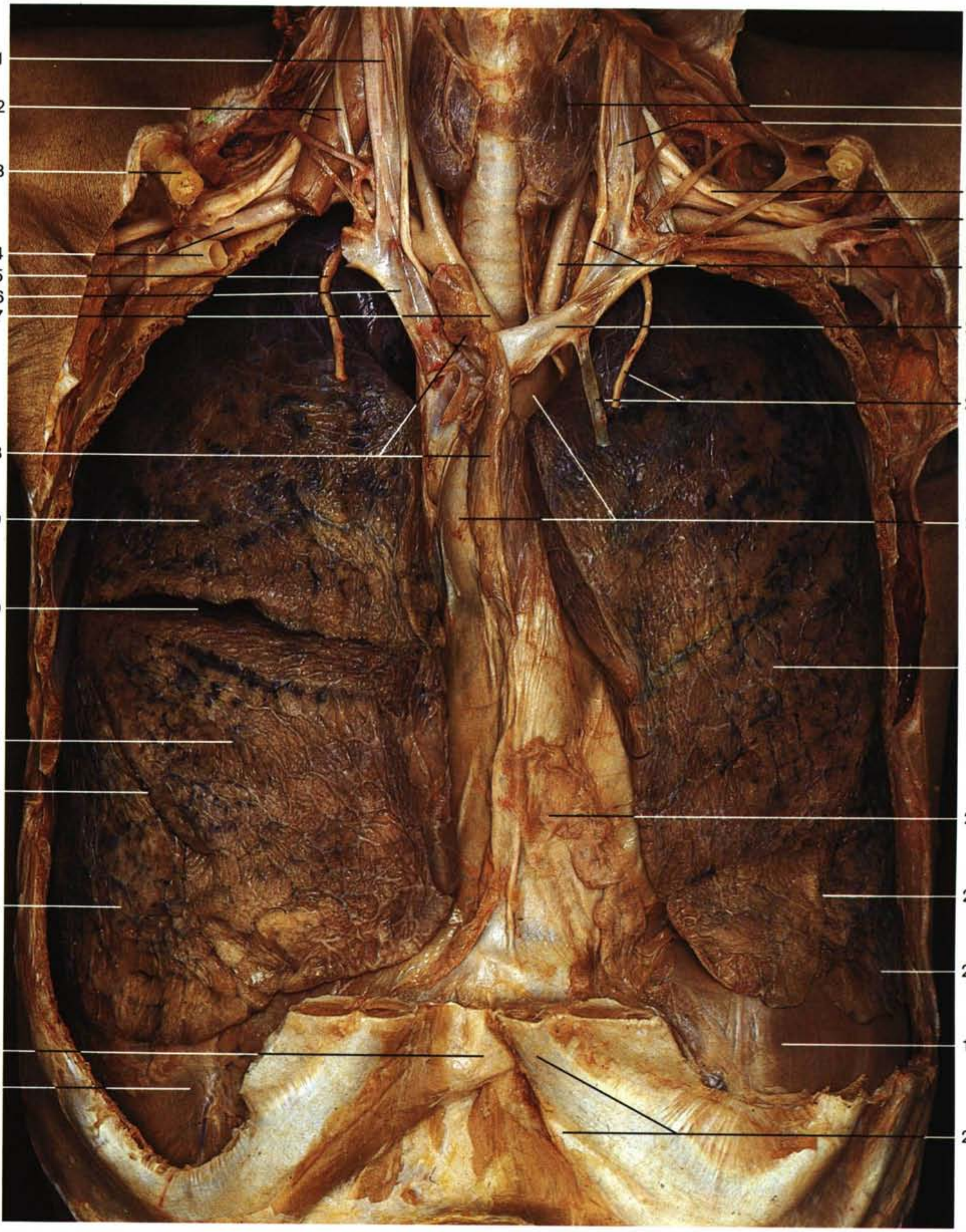

Thoracic organs (ventral aspect). The pleura has been opened and the lungs exposed. Remnants of the thymus and pericardium are seen.

1 Right internal jugular vein
2 Phrenic nerve and scalenus anterior muscle
3 Clavicle (divided)
4 Right subclavian artery and vein
5 Internal thoracic artery
6 **Right brachiocephalic vein**
7 Brachiocephalic trunk
8 Thymus (atrophic)
9 Upper lobe of right **lung**
10 Horizontal fissure of right lung (incomplete)
11 Middle lobe of right lung
12 Oblique fissure of right **lung**
13 Lower lobe of right lung
14 Xiphoid process
15 **Diaphragm**
16 **Thyroid gland**
17 Left internal jugular vein
18 Brachial plexus
19 Left cephalic vein
20 Left common carotid artery and vagus nerve
21 **Left brachiocephalic vein**
22 Internal thoracic artery and vein (divided)
23 **Ascending aorta** and aortic arch
24 Upper lobe of left **lung**
25 **Pericardium**
26 Oblique fissure of left lung
27 Lower lobe of left lung
28 Costal margin

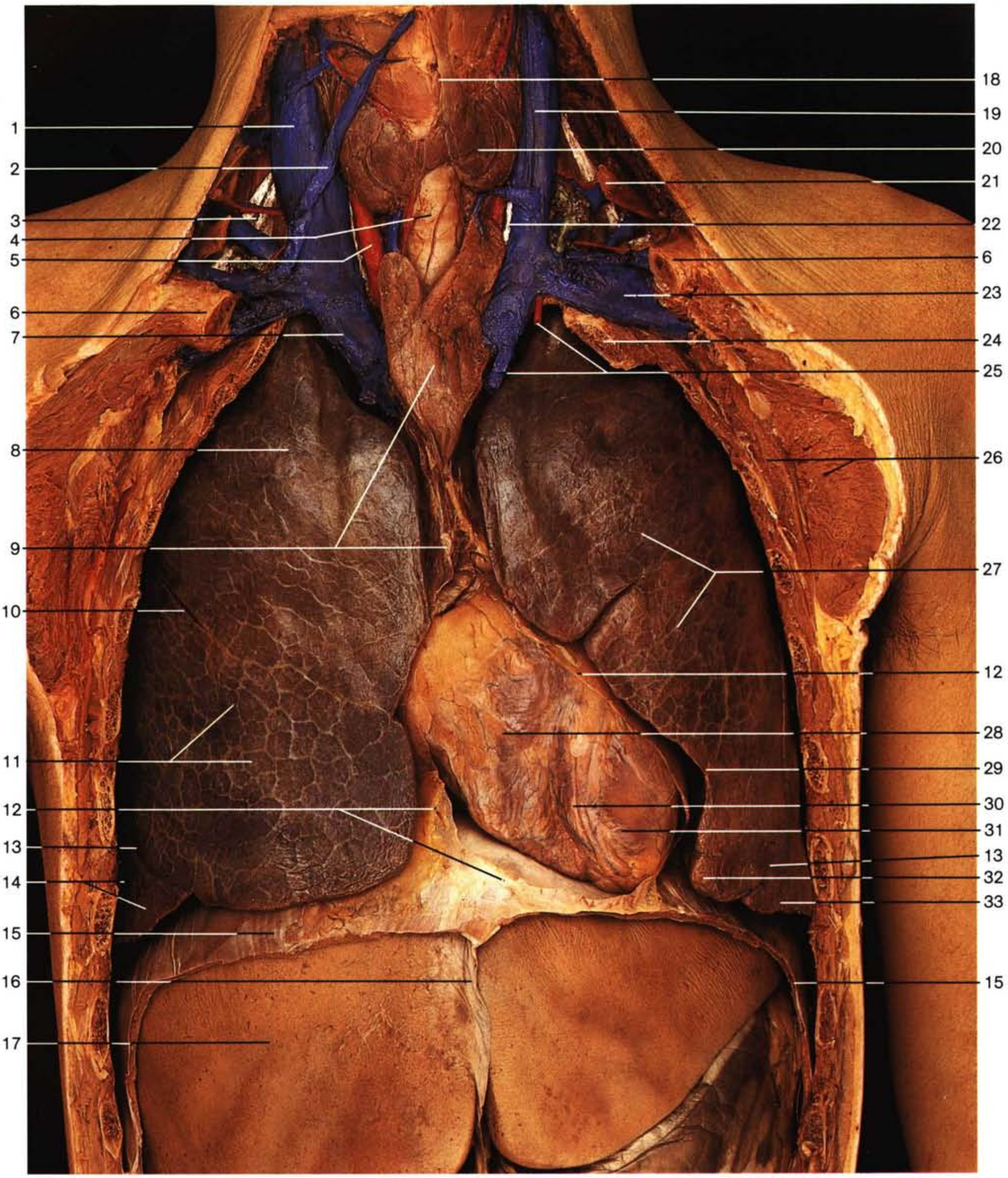

Thoracic organs (ventral aspect). The thoracic wall, costal pleura, pericardium and diaphragm have been partly removed.

1 Internal jugular vein
2 External jugular vein (displaced medially)
3 Brachial plexus
4 Trachea
5 Right common carotid artery
6 Clavicle (divided)
7 Right brachiocephalic vein
8 Upper lobe of right lung
9 **Thymus** (atrophic)
10 Horizontal fissure of right lung
11 Middle lobe of right lung

12 Pericardium (cut edges)
13 Oblique fissure of lung
14 Lower lobe of right lung
15 **Diaphragm**
16 Falciform ligament
17 **Liver**
18 Location of larynx
19 Left internal jugular vein
20 Thyroid gland
21 Omohyoid muscle (divided)
22 Vagus nerve

23 Left subclavian vein
24 1st rib (divided)
25 Internal thoracic artery and vein
26 Pectoralis major and pectoralis minor muscles (cut edges)
27 **Upper lobe of left lung**
28 **Right ventricle**
29 Cardiac notch of left lung
30 Interventricular sulcus of heart
31 **Left ventricle**
32 Lingula
33 Lower lobe of left lung

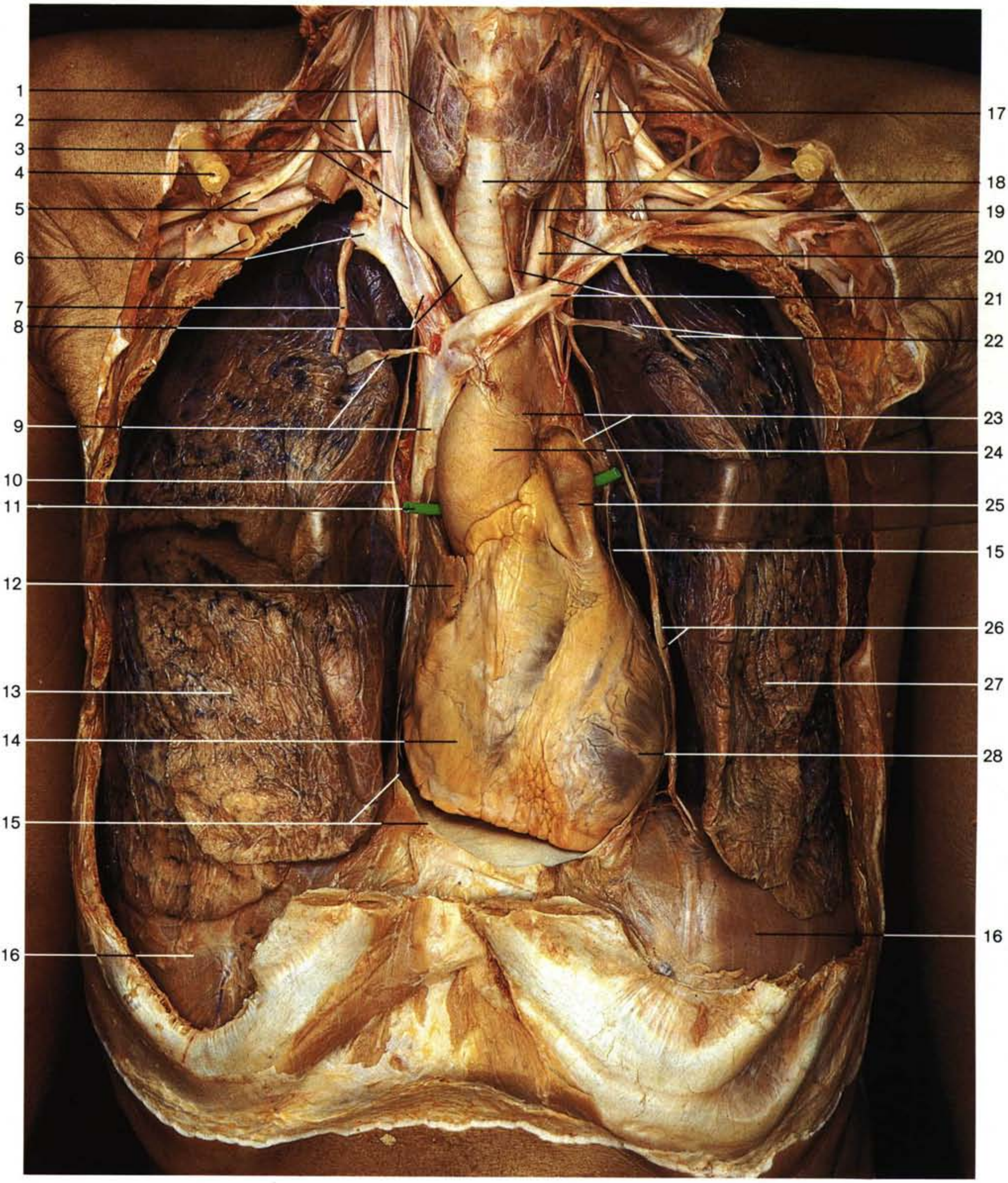

Thoracic organs, position of the heart, middle mediastinum (ventral aspect). The anterior wall of the thorax, the costal pleura and the pericardium have been removed and the lungs slightly reflected.

1 Thyroid gland
2 Phrenic nerve and scalenus anterior muscle
3 Vagus nerve and internal jugular vein
4 Clavicle (divided)
5 Brachial plexus and subclavian artery
6 Subclavian vein
7 Internal thoracic artery
8 Brachiocephalic trunk and right brachiocephalic vein
9 **Superior vena cava** and thymic vein
10 **Right phrenic nerve**
11 Transverse pericardial sinus (probe)
12 Right auricle
13 Middle lobe of right lung
14 **Right ventricle**
15 Cut edge of pericardium
16 Diaphragm
17 Internal jugular vein
18 Trachea
19 Left recurrent laryngeal nerve
20 Left common carotid artery and vagus nerve
21 Left brachiocephalic vein and inferior thyroid vein
22 Left internal thoracic artery and vein (divided)
23 Upper margin of pericardial sac
24 **Ascending aorta**
25 Pulmonary trunk
26 **Left phrenic nerve** and left pericardiacophrenic artery and vein
27 Upper lobe of left lung
28 Left ventricle

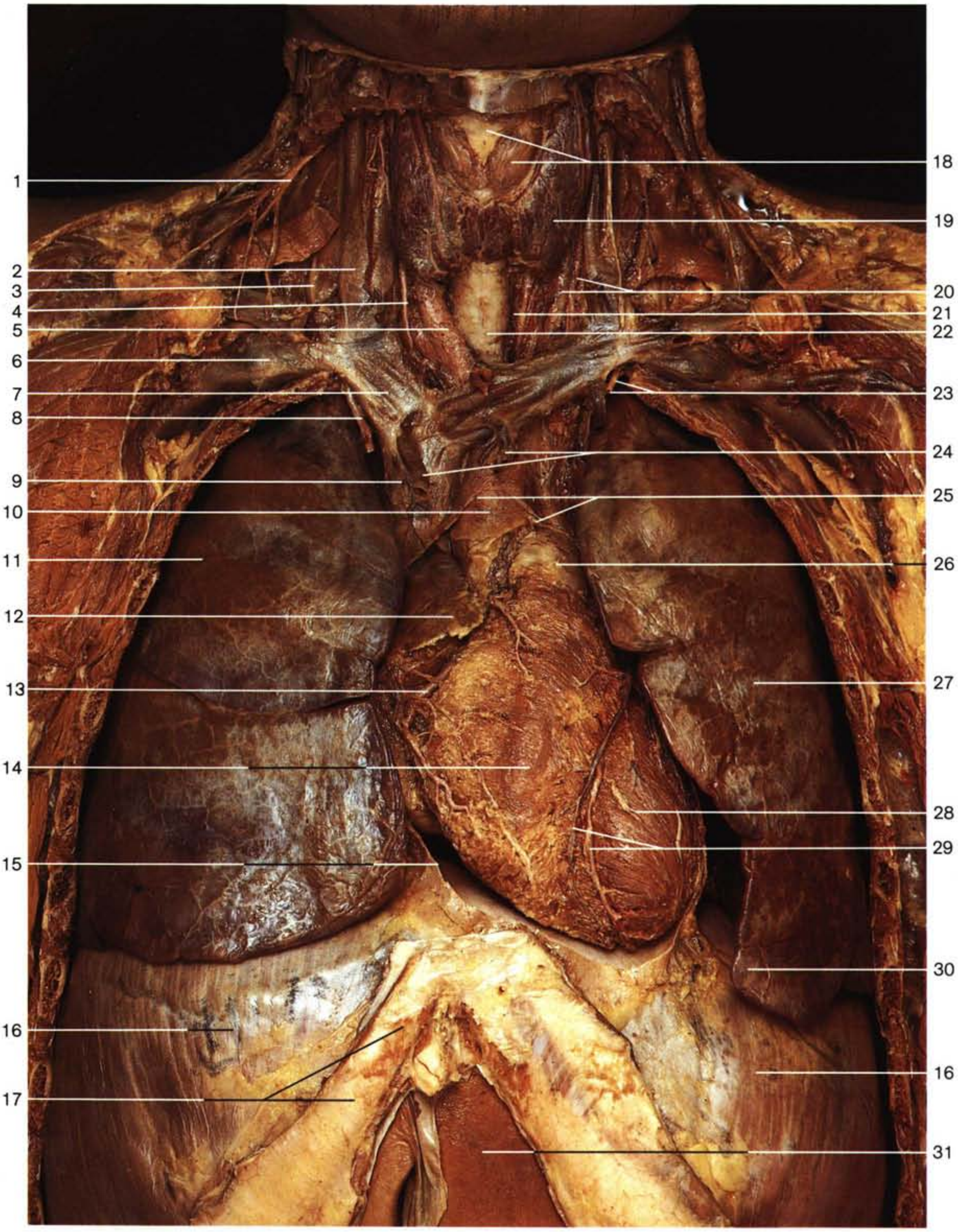

Thoracic organs, position of heart, dissection of coronary vessels in situ (ventral aspect). The anterior wall of thorax, costal pleura and pericardium have been removed.

1 Intermediate supraclavicular nerve
2 Internal jugular vein
3 Right phrenic nerve
4 Right vagus nerve
5 Right common carotid artery
6 Right subclavian vein
7 Right brachiocephalic vein
8 Right internal thoracic artery
9 **Superior vena cava**
10 Ascending aorta
11 Right lung
12 Right atrium
13 **Right coronary artery** and small cardiac vein
14 **Right ventricle**
15 Cut edge of pericardium
17 Costal margin
18 Larynx (cricothyroid muscle and thyroid cartilage)
19 Thyroid gland
20 Left common carotid artery and left vagus nerve
21 Left recurrent laryngeal nerve
22 Trachea
23 Left internal thoracic artery and vein (divided)
24 Thymic veins
25 Margin of pericardial sac
26 Pulmonary trunk
27 Left lung
28 Left ventricle
29 Anterior interventricular artery and vein
30 Lingula
31 Liver

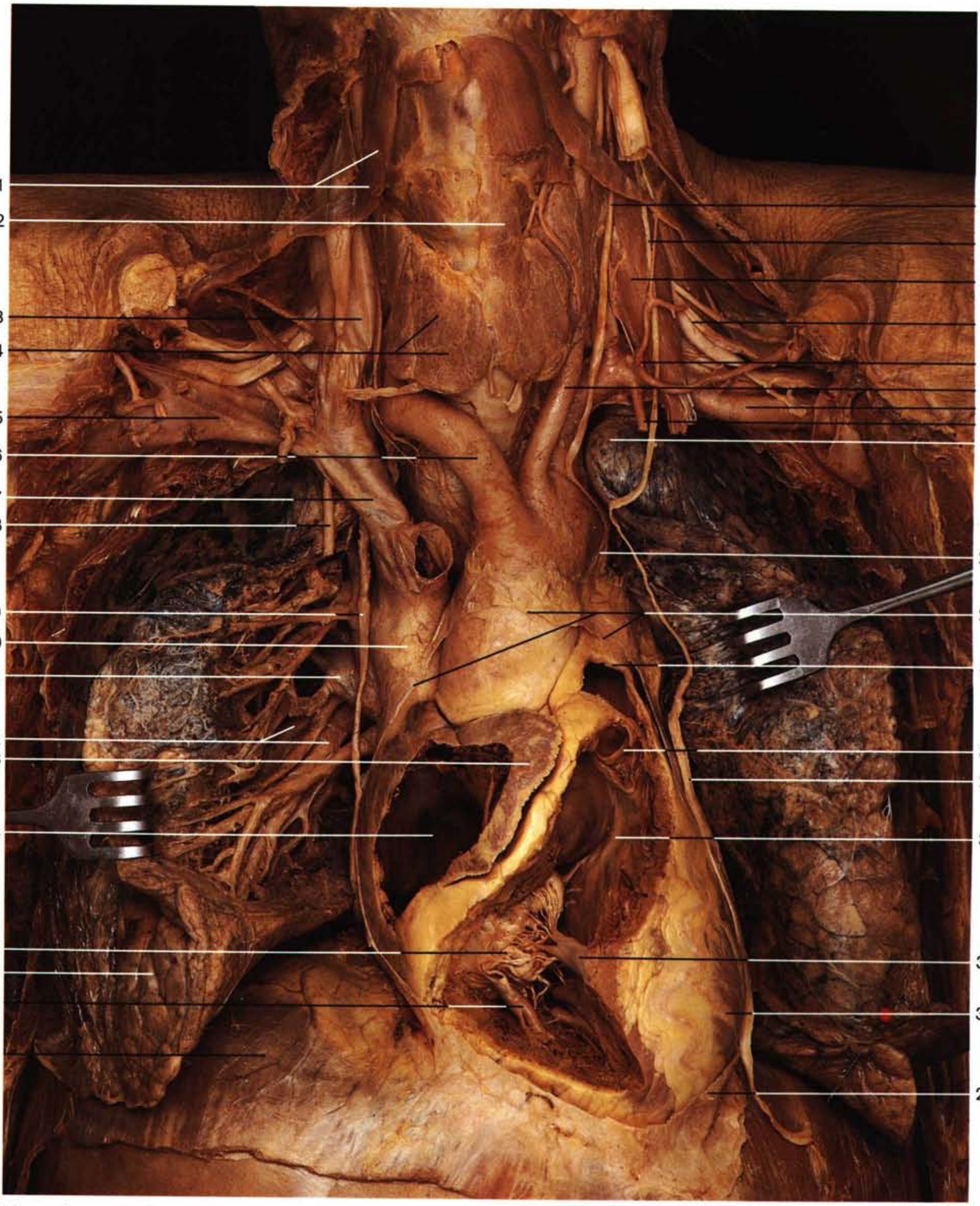

Thoracic organs, heart with valves in situ (ventral aspect). Anterior wall of thorax, pleura and anterior portion of pericardium have been removed. The right atrium and ventricle have been opened to show the right atrioventricular and pulmonary valves.

1 Omohyoid muscle
2 Pyramidal lobe of thyroid gland
3 Internal jugular vein
4 Thyroid gland
5 Right subclavian vein
6 Brachiocephalic trunk
7 **Right brachiocephalic vein**
8 Right internal thoracic artery
9 Right phrenic nerve
10 **Superior vena cava**
11 Pulmonary vein
12 Branches of pulmonary artery
13 Right auricle
14 **Right atrium**
15 **Right atrioventricular (tricuspid) valve**
16 Right lung
17 Posterior papillary muscle
18 Diaphragm
19 Left **vagus nerve**
20 Left **phrenic nerve**
21 Scalenus anterior muscle
22 Brachial plexus
23 Thyrocervical trunk
24 Left common carotid artery
25 Left subclavian artery
26 Left internal thoracic artery
27 Apex of left lung
28 Left recurrent laryngeal nerve
29 Cut edge of **pericardium**
30 Pulmonary trunk (fenestrated)
31 Pulmonic valve
32 Supraventricular crest
33 **Anterior papillary muscle**
34 **Left ventricle**

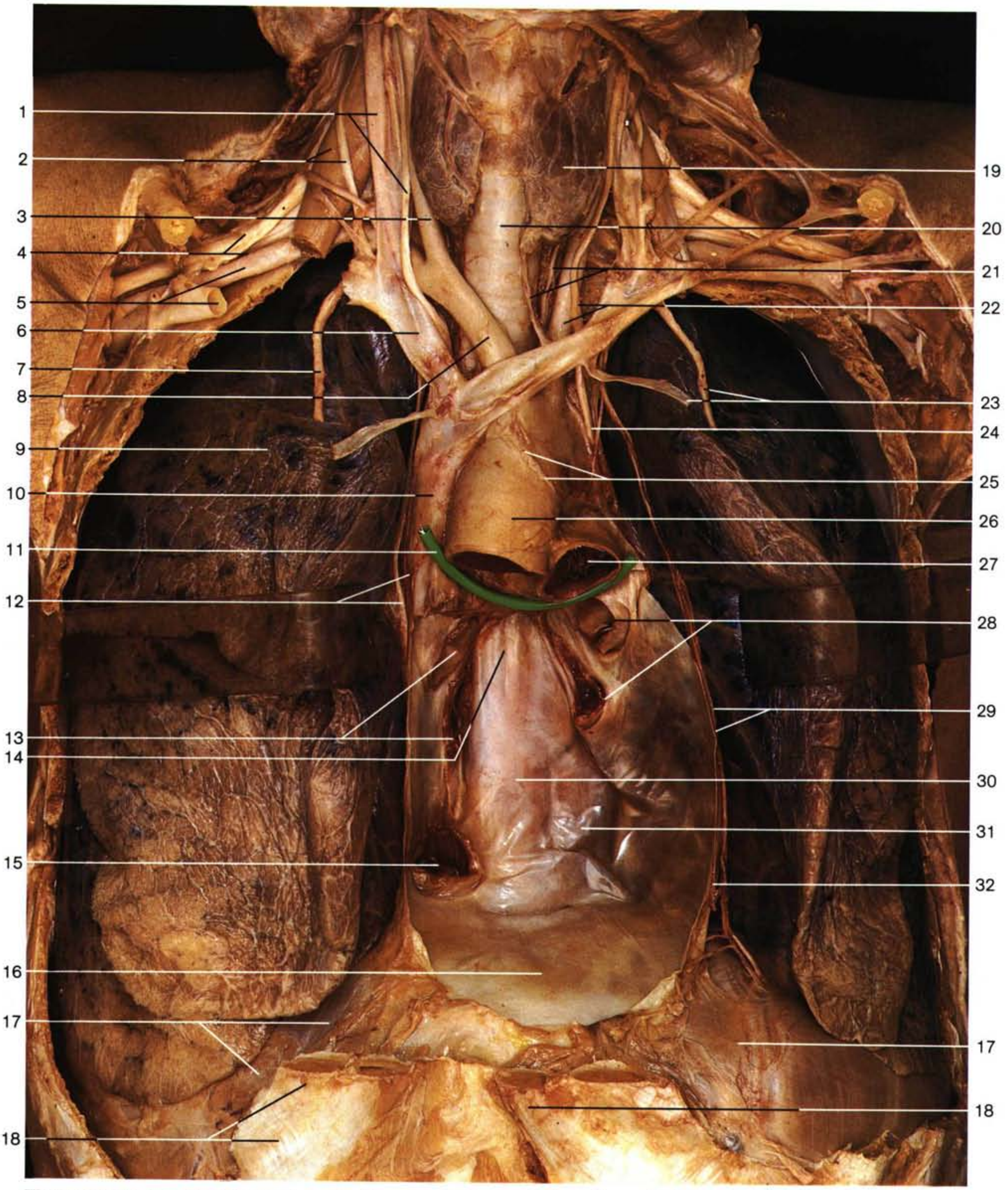

Thoracic organs, pericardium and **mediastinum** (ventral aspect). Anterior wall of thorax and heart have been removed and the lungs slightly reflected. Note probe within transverse pericardial sinus.

1 Right internal jugular vein and right vagus nerve
2 Right phrenic nerve and scalenus anterior muscle
3 Right common carotid artery
4 Brachial plexus
5 Right subclavian artery and vein
6 Right brachiocephalic vein
7 Right internal thoracic artery (divided)
8 **Brachiocephalic trunk**
9 Upper lobe of right lung
10 **Superior vena cava**
11 Transverse pericardial sinus (probe)
12 Right **phrenic nerve** and right pericardiacophrenic artery and vein
13 Right pulmonary veins
14 Oblique sinus of pericardium
15 **Inferior vena cava**
16 Diaphragmatic part of pericardium
17 Diaphragm
18 Costal margin
19 Thyroid gland
20 **Trachea**
21 Left recurrent laryngeal nerve and inferior thyroid vein
22 Left common carotid artery and left vagus nerve
23 Left internal thoracic artery and vein (divided)
24 **Vagus nerve** at aortic arch
25 Cut edge of pericardium
26 **Ascending aorta**
27 Pulmonary trunk (divided)
28 Left pulmonary veins
29 Left phrenic nerve and left pericardiacophrenic artery and vein
30 Contour of esophagus beneath pericardium
31 Contour of aorta beneath pericardium
32 **Pericardium** (cut edge)

1 Internal thoracic vein
2 Superior vena cava
3 Oblique sinus of pericardium
4 Right pulmonary veins
5 Esophagus
6 Branches of right vagus nerve
7 Mesocardium
8 Inferior vena cava
9 Middle lobe of right lung
10 Diaphragm
11 Upper lobe of left lung
12 Ascending aorta
13 **Pulmonary trunk**
14 Transverse pericardial sinus
15 Left pulmonary veins
16 Descending aorta and left vagus nerve
17 Left lung (adjacent to pericardium)
18 **Pericardium**
19 Left subclavian artery
20 Vagus nerve
21 Left recurrent laryngeal nerve
22 Descending aorta
23 Pulmonary artery
24 **Left atrium**
25 **Left ventricle**
26 Coronary sinus
27 Left common carotid artery
28 Brachiocephalic trunk
29 Azygos arch
30 **Right atrium**
31 **Right ventricle**
32 Aortic arch

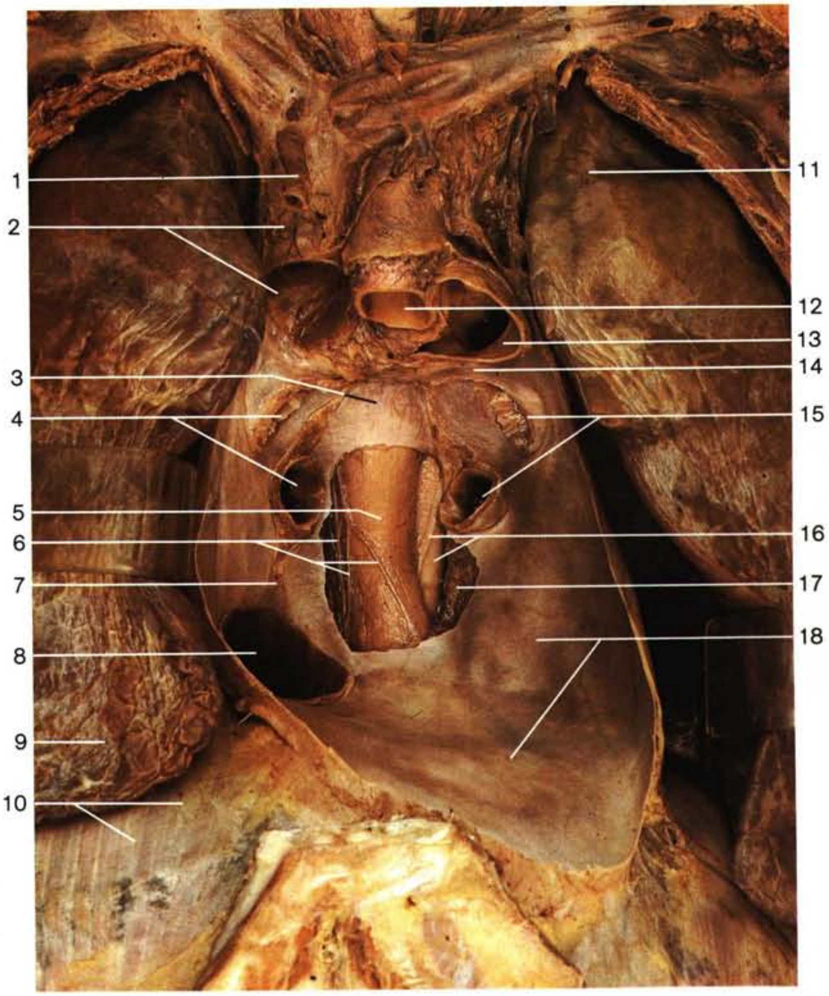
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

Pericardial sac (ventral aspect). The heart has been removed, and the posterior wall of the pericardium has been opened to show the adjacent esophagus and aorta.

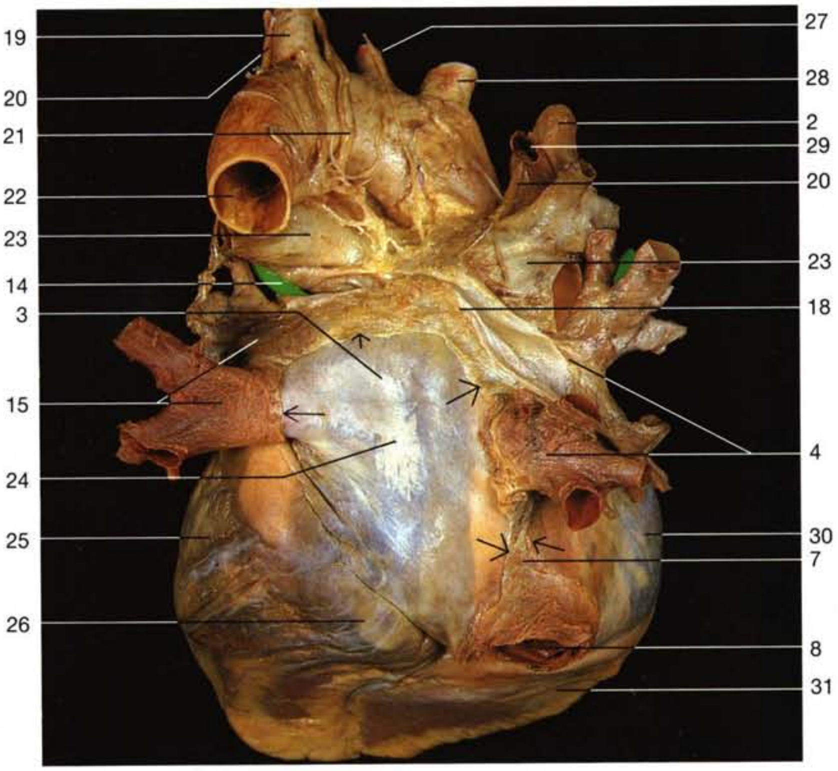
19
20
21
22
23
14
3
15
24
25
26
27
28
2
29
20
23
18
4
30
7
8
31

Heart with epicardium (posterior aspect). Arrows: oblique sinus.

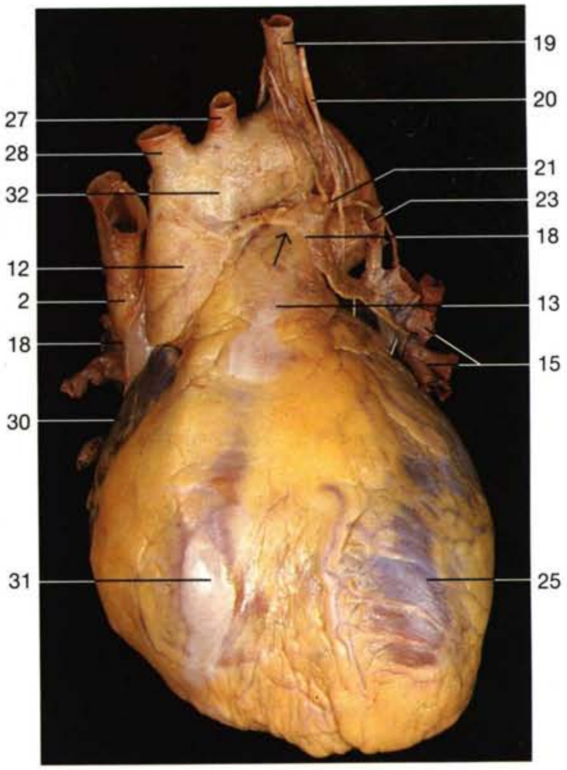
27
28
32
12
2
18
30
31
19
20
21
23
18
13
15
25

Heart with epicardium (anterior aspect). Arrow: pericardial reflection.

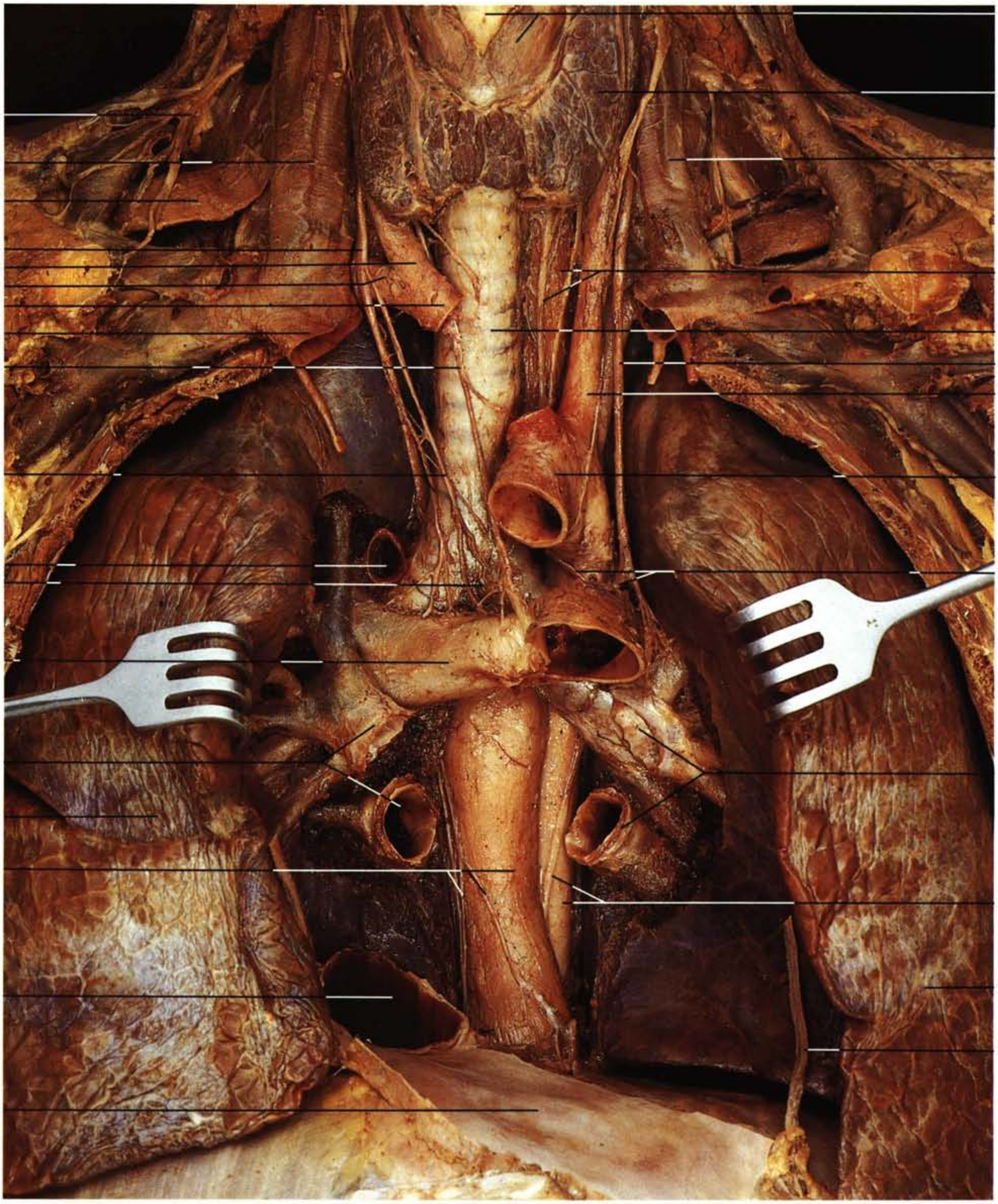

Mediastinal organs after removal of heart and pericardium (ventral aspect). Both lungs have been slightly reflected.

1 Supraclavicular nerves
2 Internal jugular vein
3 Omohyoid muscle
4 **Right vagus nerve**
5 Right common carotid artery
6 Right subclavian artery
7 Brachiocephalic trunk
8 Right brachiocephalic vein
9 Superior cervical cardiac branch of vagus nerve
10 Inferior cervical cardiac branches of vagus nerve
11 Azygos arch (divided)
12 Bifurcation of trachea
13 Right **pulmonary artery**
14 Right **pumonary veins**
15 Right lung
16 **Esophagus** and branches of right vagus nerve
17 Inferior vena cava
18 Pericardium
19 Larynx (thyroid cartilage, cricothyroid muscle)
20 Thyroid gland
21 Internal jugular vein
22 Esophagus and left recurrent laryngeal nerve
23 **Trachea**
24 **Left vagus nerve**
25 Left common carotid artery
26 Aortic arch
27 Left recurrent laryngeal nerve branching off from vagus nerve
28 Left pulmonary veins
29 **Thoracic aorta** and left vagus nerve
30 Left lung
31 Left phrenic nerve (divided)

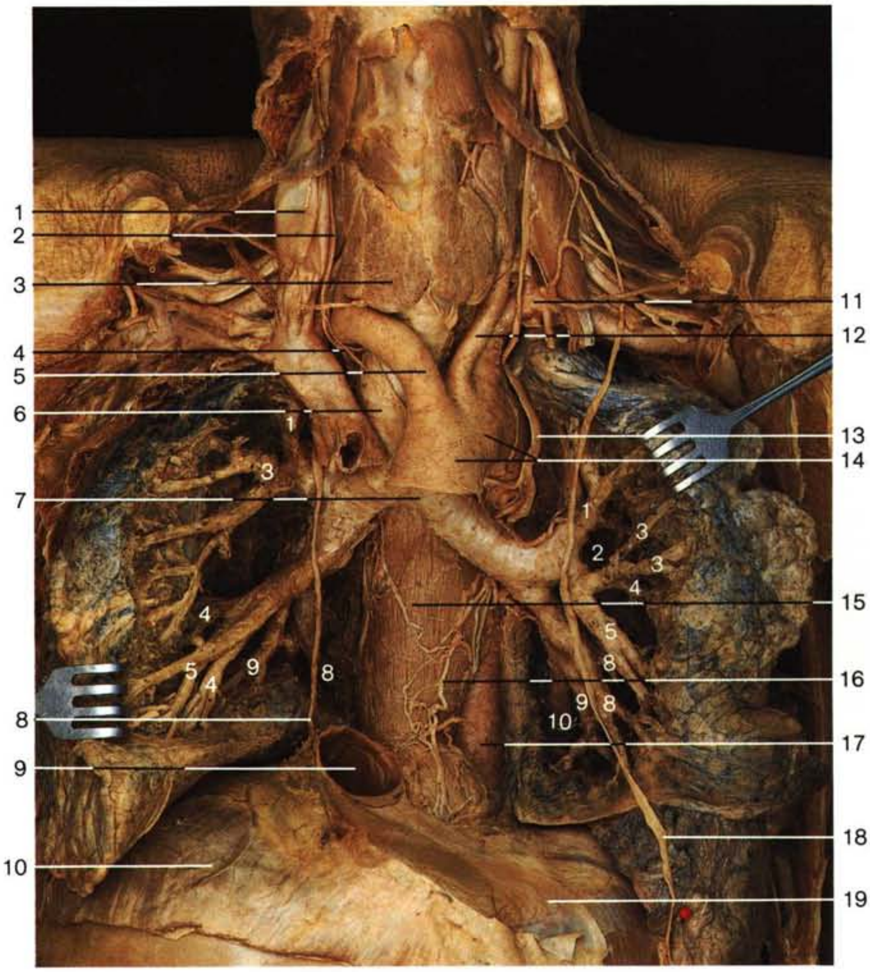
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
1
3
4
5
4
9
8
1
2
3
3
4
5
8
9
8
10

Bronchial tree in situ (ventral aspect). Heart and pericardium have been removed; the bronchi of the bronchopulmonary segments are dissected. 1–10 = numbers of segments.

1	Internal jugular vein
2	Right vagus nerve
3	**Thyroid gland**
4	Right recurrent laryngeal nerve
5	Brachiocephalic trunk
6	Trachea
7	**Bifurcation of trachea**
8	**Right phrenic nerve**
9	Inferior vena cava
10	Diaphragm
11	Left subclavian artery
12	Left common carotid artery
13	Left **vagus nerve**
14	Aortic arch
15	**Esophagus**
16	Esophageal plexus
17	Thoracic aorta
18	**Left phrenic nerve**
19	Pericardium at the central tendon of diaphragm
20	Right pulmonary artery
21	Left pulmonary artery
22	Tracheal lymph nodes
23	Superior tracheobronchial lymph nodes
24	Bronchopulmonary lymph nodes

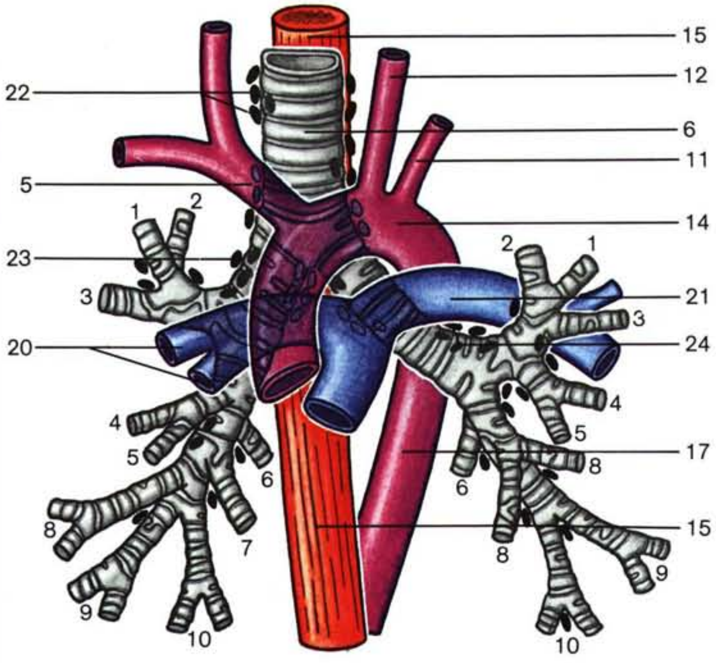
15
12
22
6
11
5
14
1
2
2
1
23
21
3
3
20
24
4
4
5
5
6
17
6
8
8
7
8
15
9
9
10
10

Relation of aorta, pulmonary trunk and esophagus to trachea and bronchial tree. (Schematic drawing.)
1–10 = number of segments (cf. p. 230 and 235).

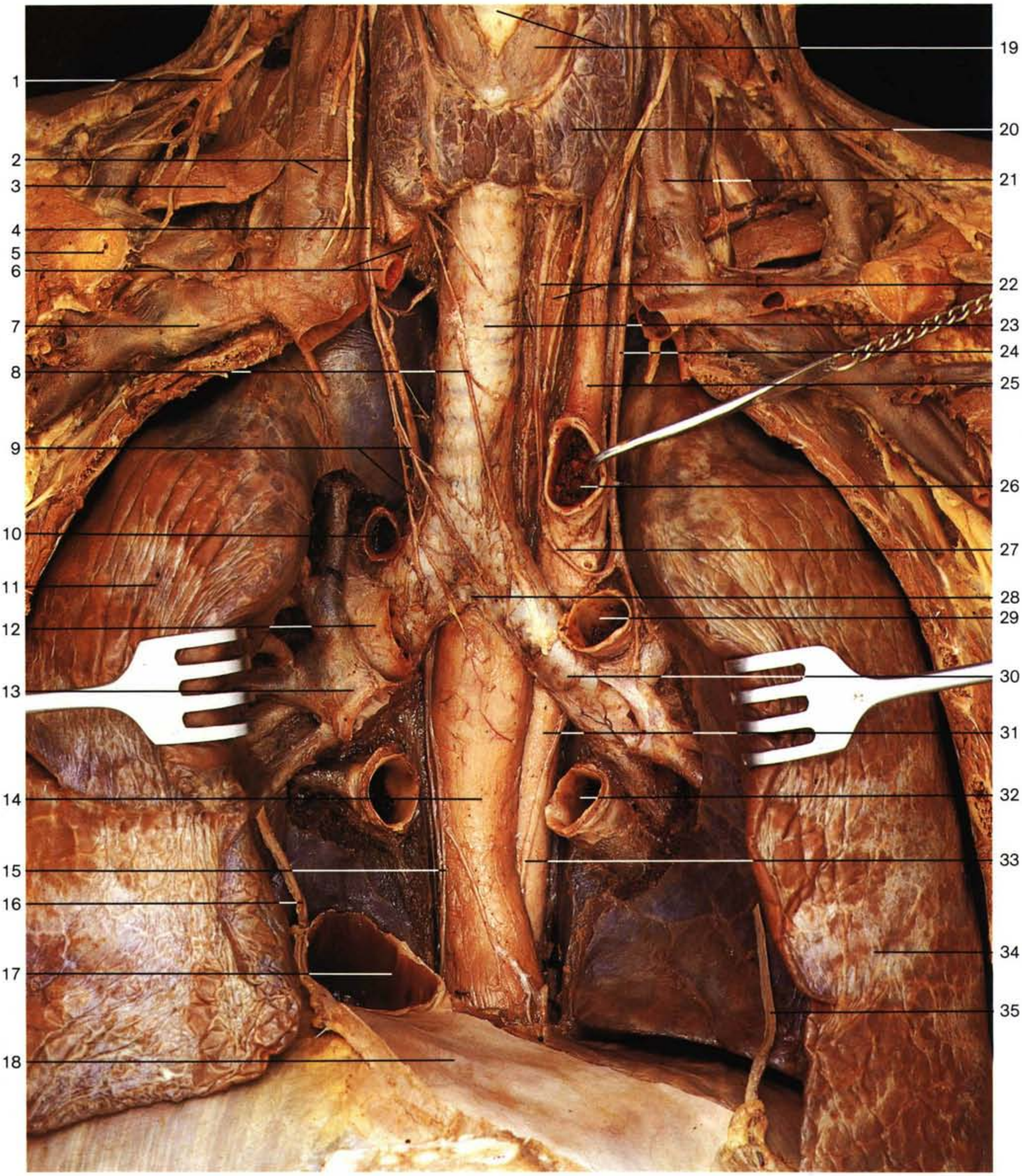

Organs of posterior mediastinum (ventral aspect). The heart with the pericardium has been removed, and the lungs and aortic arch have been slightly reflected to show the vagus nerves and their branches.

1 Supraclavicular nerves
2 Right internal jugular vein with ansa cervicalis
3 Omohyoid muscle
4 **Right vagus nerve**
5 Clavicle
6 Right subclavian artery and recurrent laryngeal nerve
7 Right subclavian vein
8 Superior cervical cardiac branch of vagus nerve
9 Inferior cervical cardiac branch of vagus nerve
10 Azygos arch (divided)
11 **Right lung**
12 **Right pulmonary artery**
13 **Right pulmonary veins**
14 **Esophagus**
15 Esophageal plexus
16 Right phrenic nerve (divided)
17 Inferior vena cava
18 Pericardium covering the diaphragm
19 **Larynx (thyroid cartilage and cricothyroid muscle)**
20 Thyroid gland
21 Left internal jugular vein
22 Esophagus and left recurrent laryngeal nerve
23 Trachea
24 **Left vagus nerve**
25 Left common carotid artery
26 Aortic arch
27 Left recurrent laryngeal nerve
28 **Bifurcation of trachea**
29 **Left pulmonary artery**
30 Left primary bronchus
31 Descending aorta
32 Left pulmonary veins
33 Branch of left vagus nerve
34 **Left lung**
35 Left phrenic nerve (divided)

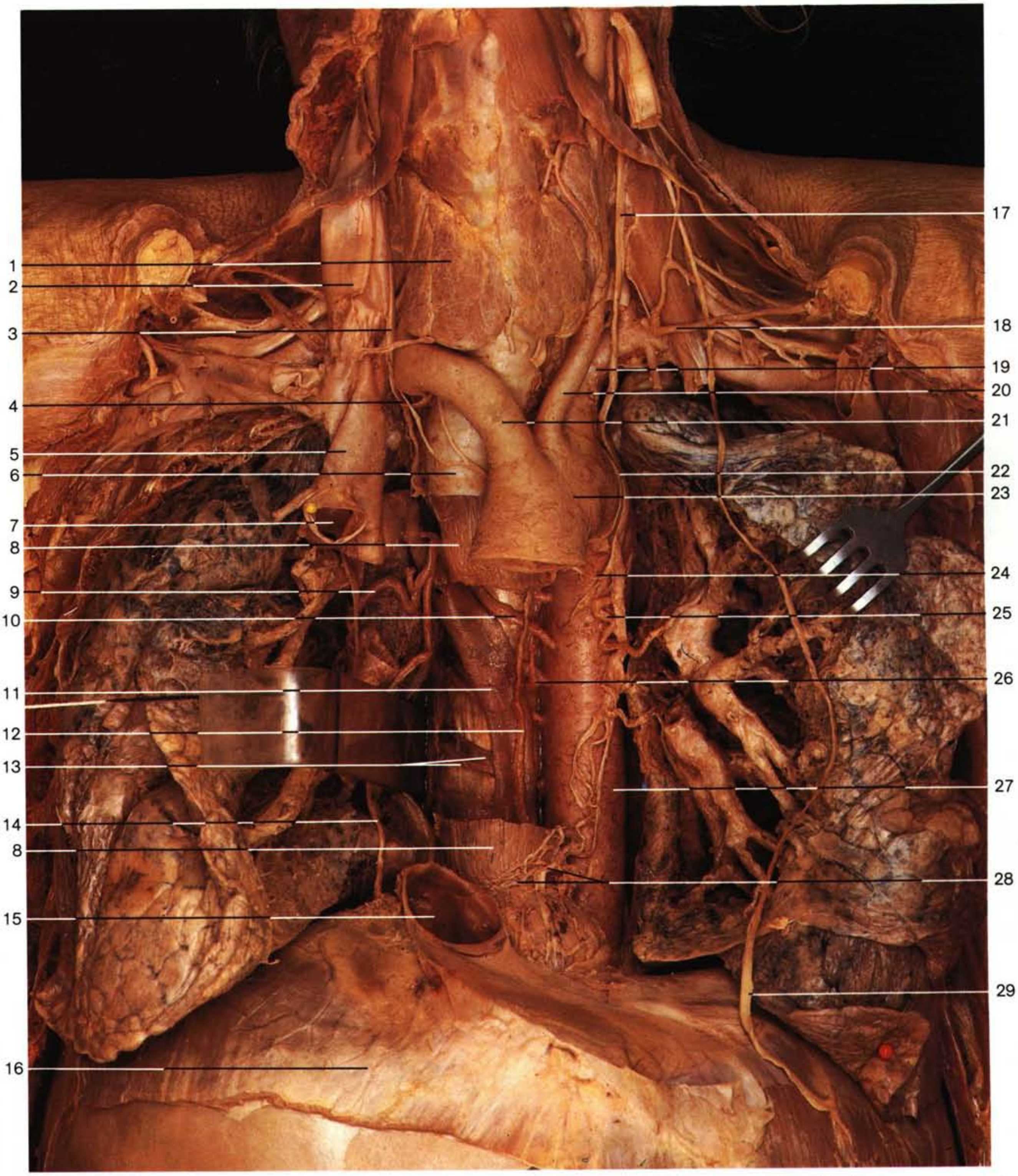

Mediastinal organs (ventral aspect). Heart and distal part of esophagus have been removed to display the vessels and nerves of the posterior mediastinum.

1 Thyroid gland
2 Right internal jugular vein
3 Right vagus nerve
4 Point, where right recurrent laryngeal nerve is branching off the vagus nerve
5 Right brachiocephalic vein
6 Trachea
7 Left brachiocephalic vein (reflected)
8 Esophagus
9 Right bronchial artery
10 Posterior intercostal artery
11 **Azygos vein**
12 **Thoracic duct**
13 Posterior intercostal artery and vein (in front of the vertebral column)
14 Right phrenic nerve
15 Inferior vena cava
16 Diaphragm
17 **Left vagus nerve**
18 Thyrocervical trunk
19 Left subclavian artery
20 Left common carotid artery
21 Brachiocephalic trunk
22 Left vagus nerve
23 **Aortic arch**
24 Left recurrent laryngeal nerve
25 Left bronchial artery
26 Lymph node
27 Thoracic aorta
28 Esophageal plexus
29 Left phrenic nerve

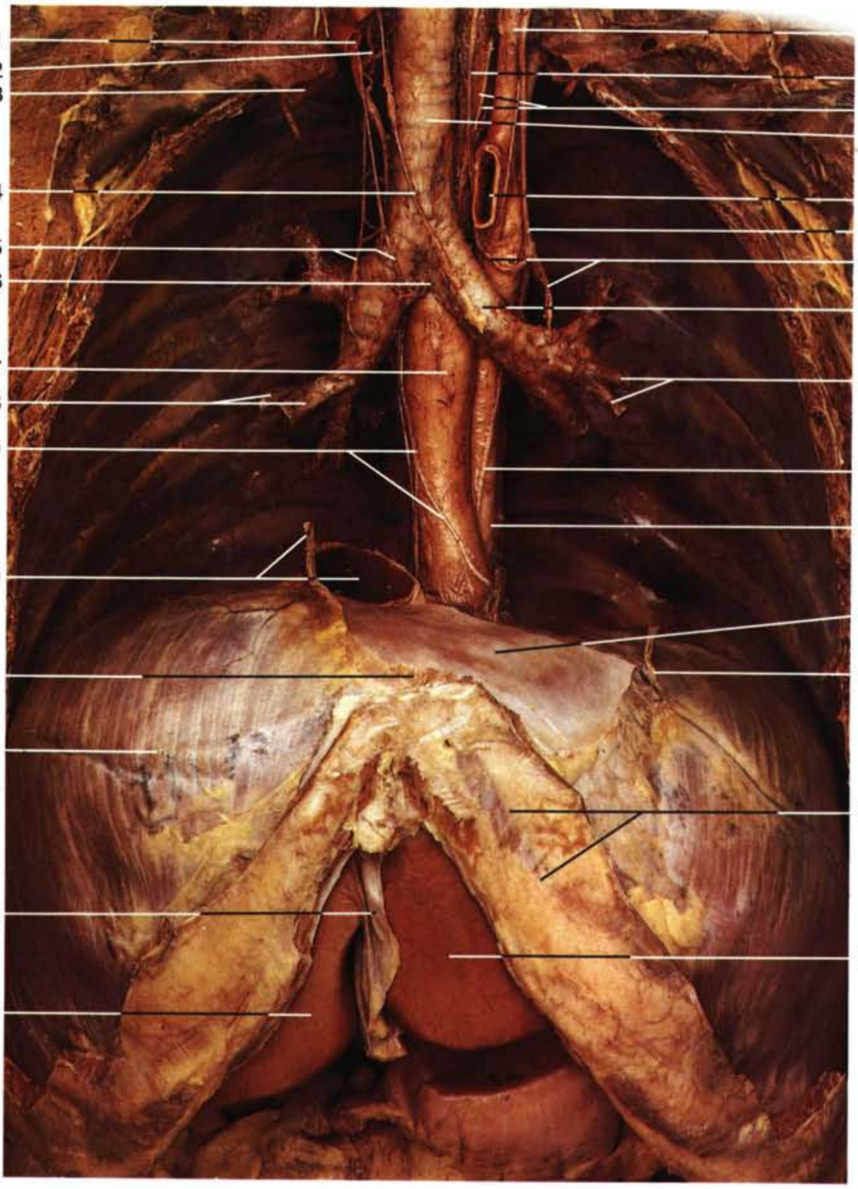

Diaphragm and organs of mediastinum (anterior aspect). Heart and lungs have been removed; the costal margin remains in place. Note the different courses of left and right vagus.

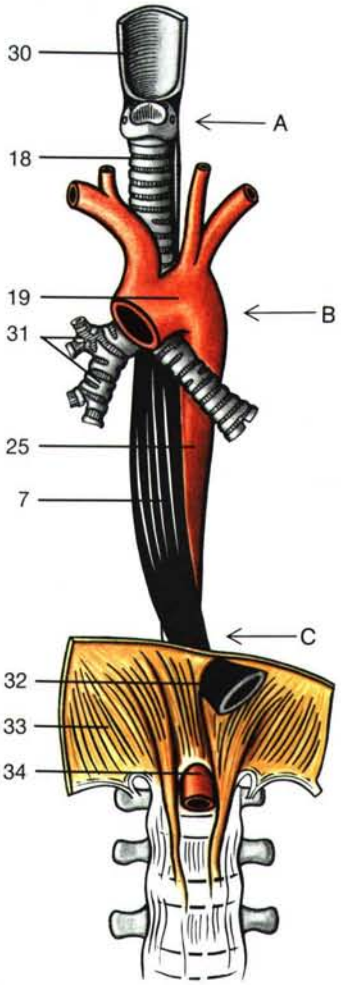

Organs of posterior mediastinum (ventral aspect). (Schematic drawing.) Three regions in which the esophagus is narrowed are shown:

A at the level of the cricoid cartilage;

B at the level of the aortic arch;

C at the level of the diaphragm.

1 Right subclavian artery
2 Right recurrent laryngeal nerve
3 Right brachiocephalic vein
4 Superior cervical cardiac nerve
5 Inferior cervical cardiac nerves and pulmonary branches
6 **Bifurcation of trachea**
7 **Esophagus** (thoracic part)
8 Bronchi of lateral and medial segments of middle lobe
9 Esophageal plexus and branches of right vagus nerve
10 Inferior vena cava and right phrenic nerve (cut)
11 Sternal part of diaphragm
12 Costal part of **diaphragm**
13 Falciform ligament of liver
14 Liver (quadrate lobe)
15 Left common carotid artery
16 Left recurrent laryngeal nerve
17 Esophageal branches of left vagus nerve and esophagus
18 **Trachea**
19 Aortic arch
20 Left vagus nerve
21 Left recurrent laryngeal nerve with inferior cardiac nerve
22 Left primary bronchus
23 Superior and inferior lingular bronchi
24 Esophageal plexus of left vagus nerve
25 **Descending aorta**
26 Central tendon of diaphragm covered with pericardium
27 Left phrenic nerve (divided)
28 Costal margin
29 **Liver,** left lobe
30 Pharynx
31 Secondary bronchi
32 Esophagus (abdominal part)
33 Diaphragm
34 Abdominal aorta

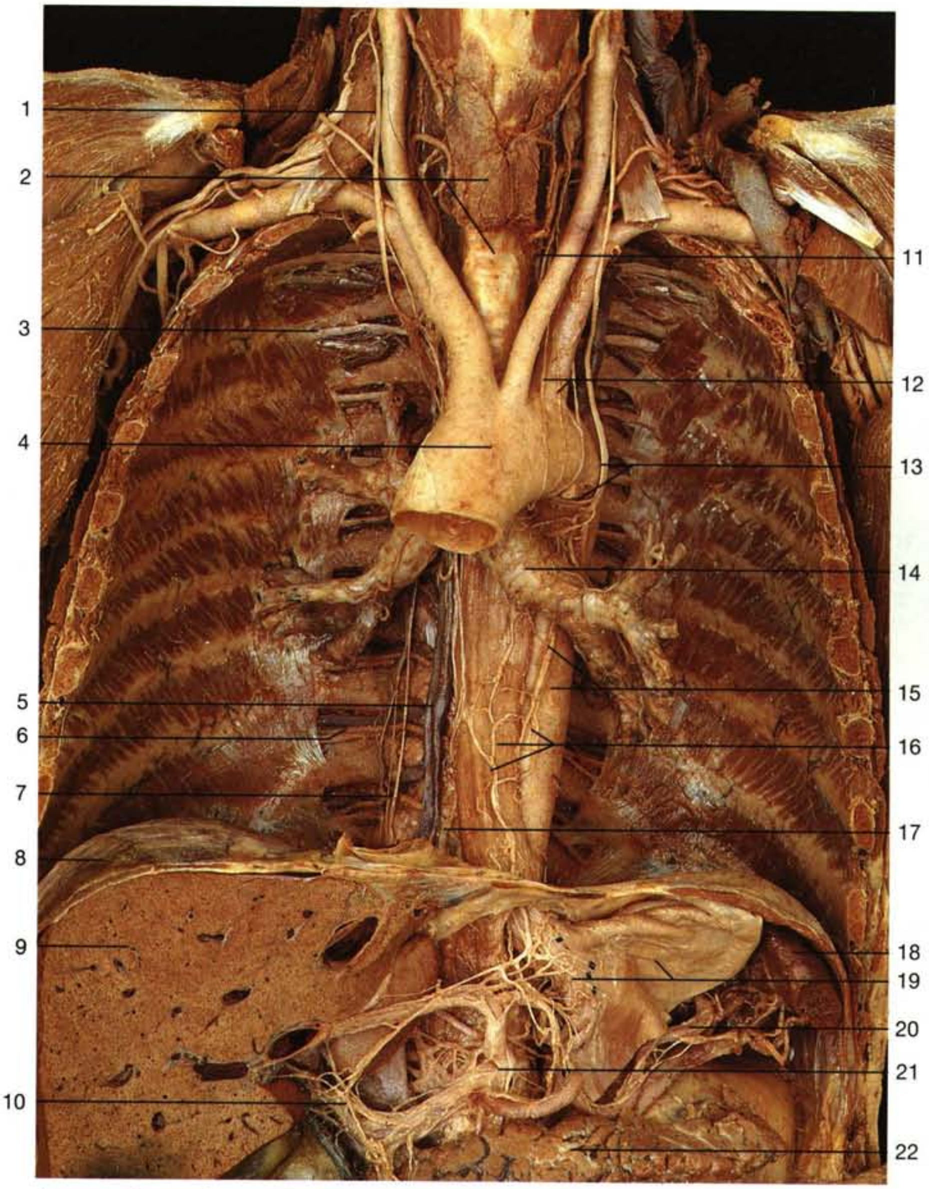

Organs of posterior mediastinum (anterior aspect).

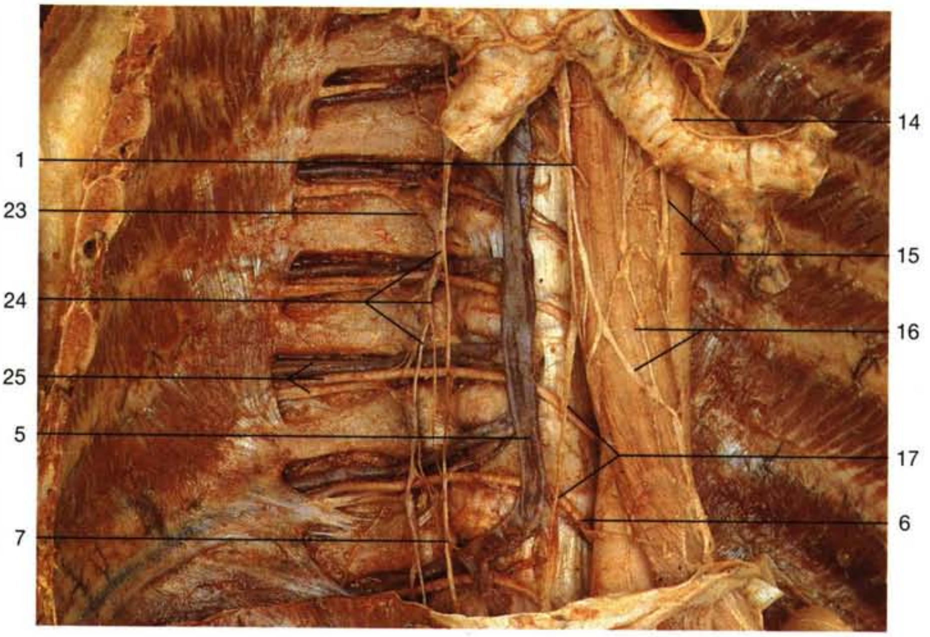

Inferior segment of posterior mediastinum (anterior aspect).

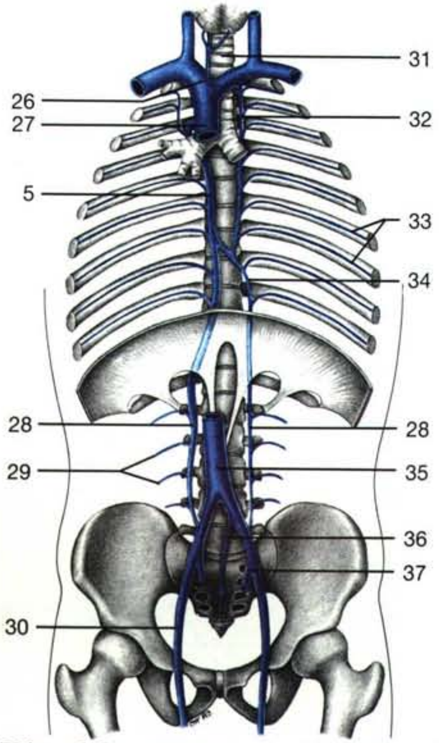

Veins of the posterior wall of thoracic and abdominal cavity. (Schematic drawing.)

1 Right vagus nerve
2 Thyroid gland and trachea
3 Intercostal nerve
4 Aortic arch
5 **Azygos vein**
6 Posterior intercostal artery
7 **Greater splanchnic nerve**
8 Diaphragm
9 Liver
10 Proper hepatic artery and hepatic plexus
11 Left recurrent laryngeal nerve
12 Inferior cervical cardiac nerves
13 Left vagus nerve and left recurrent laryngeal nerve
14 Left primary bronchus
15 Thoracic aorta and left vagus nerve
16 Esophagus and esophageal plexus
17 **Thoracic duct**
18 Spleen
19 Anterior gastric plexus and stomach (divided)
20 Splenic artery and splenic plexus
21 **Celiac trunk** and **celiac plexus**
22 Pancreas
23 Ramus communicans
24 **Sympathetic trunk** and **sympathetic ganglion**
25 **Posterior intercostal vein** and **artery** and **intercostal nerve**
26 Right brachiocephalic vein
27 Superior vena cava
28 Ascending lumbar vein
29 Lumbar veins
30 Right external iliac vein
31 Trachea
32 Accessory hemiazygos vein
33 Posterior intercostal veins
34 Hemiazygos vein
35 Inferior vena cava
36 Median sacral vein
37 Internal iliac vein

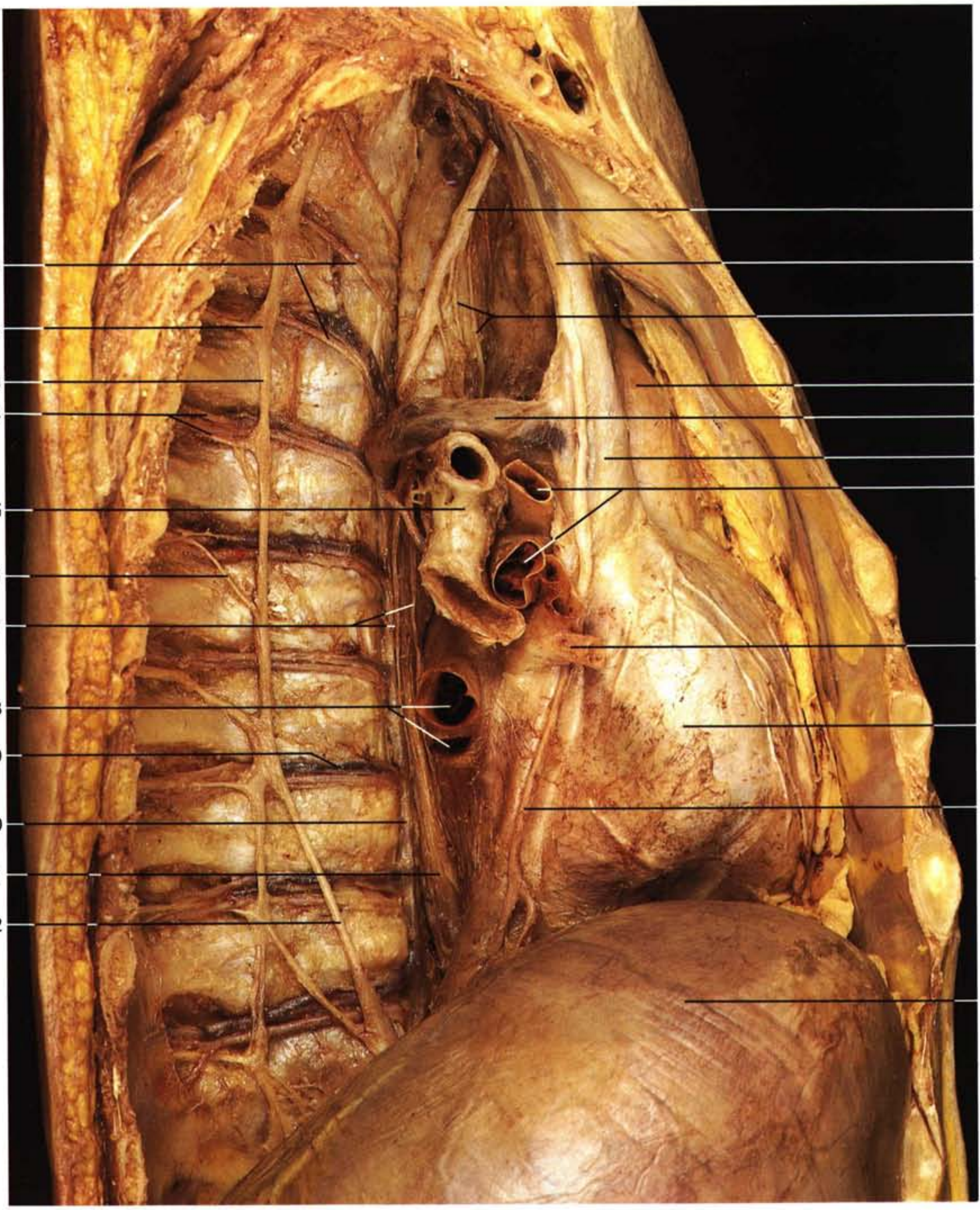

Mediastinal organs (right lateral aspect). Right lung and pleura of right half of the thorax have been removed.

1 Posterior intercostal arteries
2 Ganglion of sympathetic trunk
3 **Sympathetic trunk**
4 Vessels and nerves of the intercostal space (from above: Posterior intercostal vein and artery and intercostal nerve)
5 Right primary bronchus
6 Ramus communicans of sympathetic trunk
7 Esophageal plexus (branches of right vagus nerve)
8 Pulmonary veins
9 Posterior intercostal vein
10 **Azygos vein**
11 Esophagus
12 Greater splanchnic nerve
13 **Right vagus nerve**
14 **Right phrenic nerve**
15 Inferior cervical cardiac branches of vagus nerve
16 **Aortic arch**
17 Superior vena cava
18 Right pulmonary artery
19 **Heart with pericardium**
20 Diaphragm

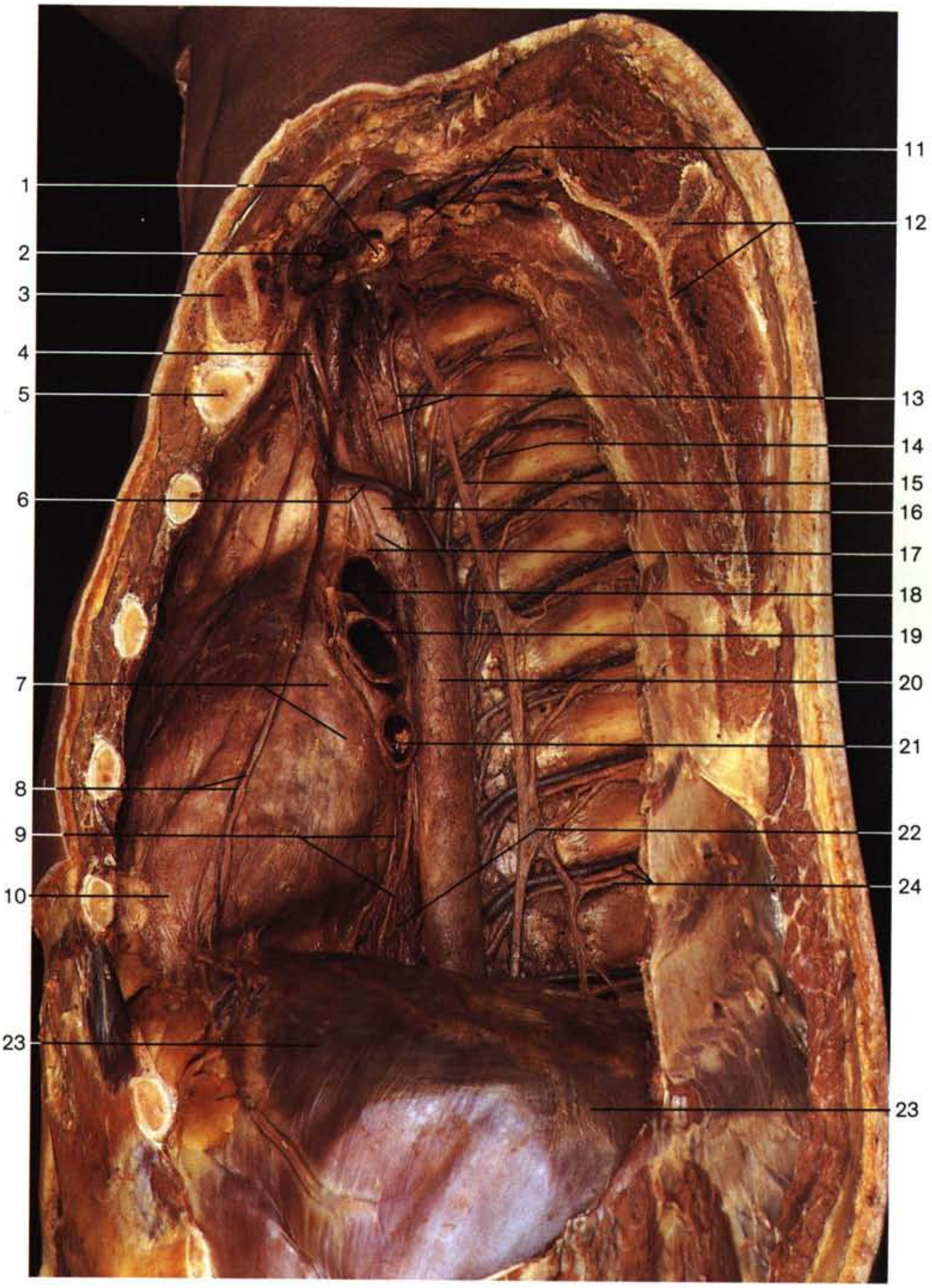

Organs of posterior and superior mediastinum (left lateral aspect).

1 Subclavian artery
2 Subclavian vein
3 Clavicle (divided)
4 **Left vagus nerve**
5 First rib (divided)
6 Left superior intercostal vein
7 **Left atrium with pericardium**
8 **Left phrenic nerve** and pericardiacophrenic artery and vein
9 Esophageal plexus (branches derived from left vagus nerve)
10 **Apex of heart with pericardium**
11 Brachial plexus
12 Scapula (divided)
13 Posterior intercostal arteries
14 Ramus communicans of sympathetic trunk
15 **Sympathetic trunk**
16 Aortic arch
17 Left vagus nerve and left recurrent laryngeal nerve
18 Left pulmonary artery
19 Left primary bronchus
20 **Thoracic aorta**
21 Pulmonary vein
22 Esophagus (thoracic part)
23 **Diaphragm**
24 Posterior intercostal artery and vein and intercostal nerve

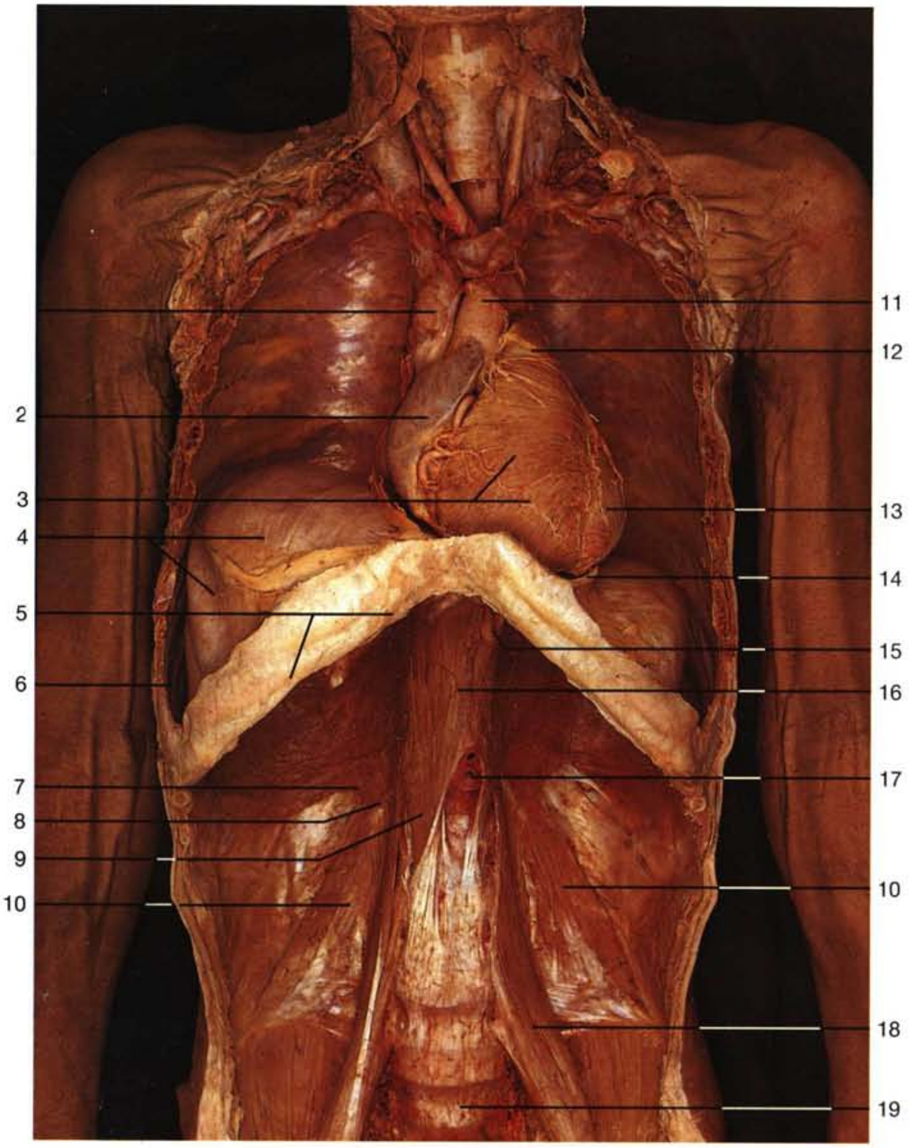

1 Superior vena cava
2 Right atrium
3 Right ventricle
4 **Costal part of diaphragm**
5 Costal margin
6 Position of costodiaphragmatic recess
7 **Lateral arcuate ligament**
8 **Medial arcuate ligament**
9 Right crus of lumbar part of diaphragm
10 Quadratus lumborum muscle
11 Ascending aorta
12 Pulmonary trunk
13 Left ventricle
14 Pericardium, diaphragm
15 Esophageal hiatus and abdominal part of esophagus (cut)
16 **Lumbar part of diaphragm**
17 Aortic hiatus
18 Psoas major muscle
19 Lumbar vertebra

Diaphragm in situ (anterior aspect). Anterior walls of thoracic and abdominal cavities have been removed. Natural position of the heart above the central tendon on the diaphragm is shown.

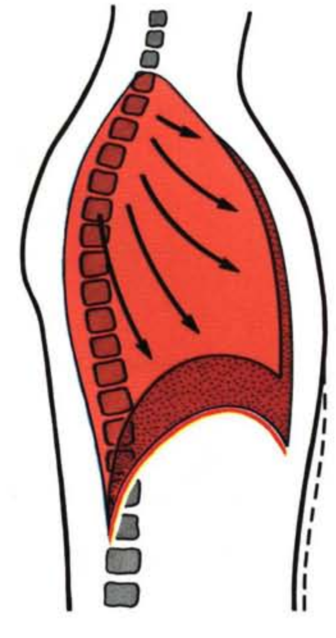

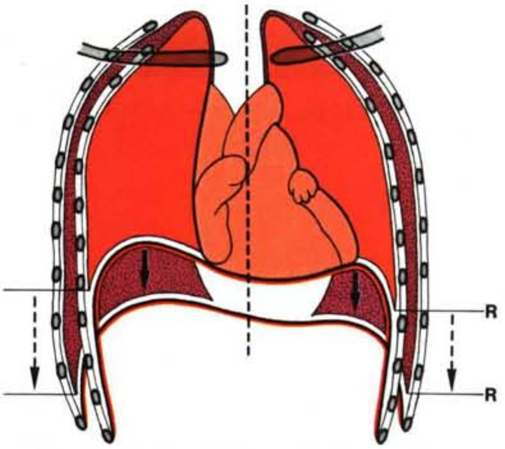

Changes in the position of the diaphragm and thoracic cage during respiration. Left: lateral aspect; right: anterior aspect. During inspiration the diaphragm moves downwards and the lower part of the thoracic cage expands forward and laterally, causing the costodiaphragmatic recess (R) to enlarge (cf. dotted arrows).

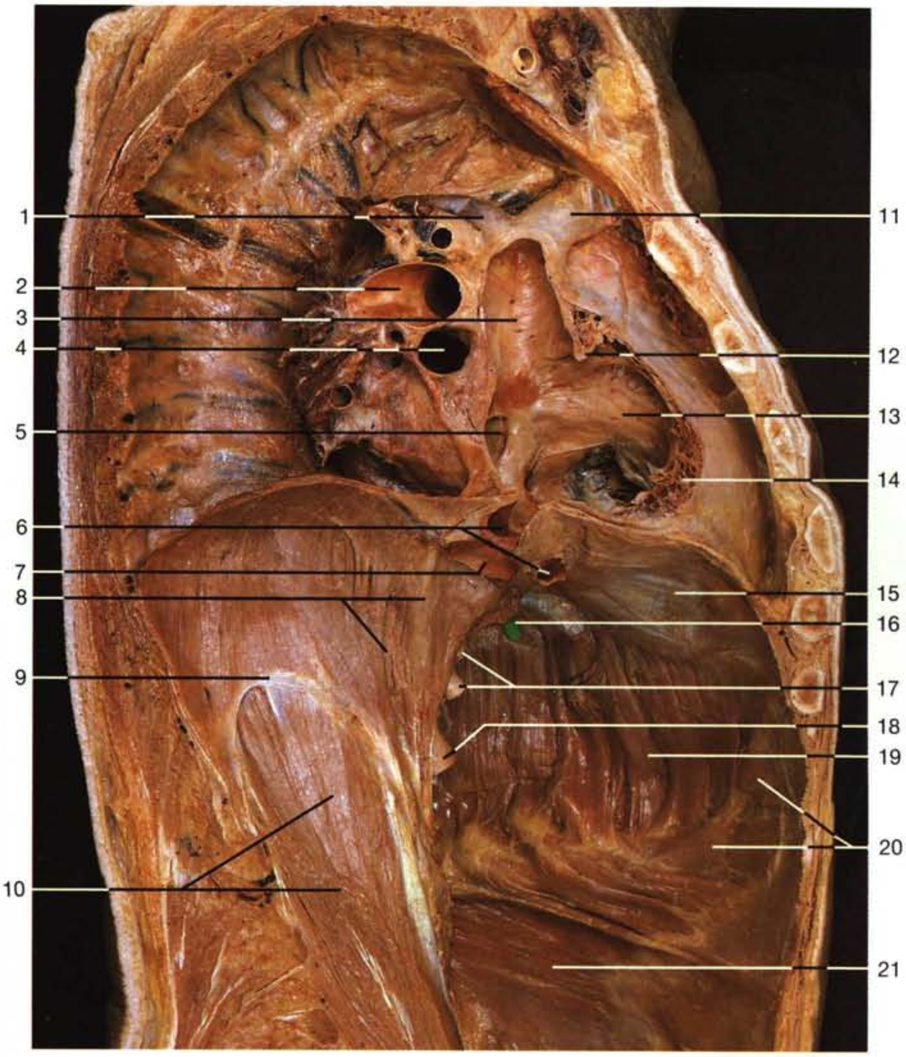

1 Azygos venous arch
2 Right pulmonary artery
3 **Superior vena cava**
4 Right pulmonary vein
5 Fossa ovalis
6 Hepatic veins
7 Inferior vena cava
8 Right crus of lumbar part of **diaphragm**
9 Medial arcuate ligament
10 Psoas major muscle
11 Left brachiocephalic vein
12 Terminal crista
13 **Right atrium**
14 Right auricle
15 Central tendon of diaphragm
16 Esophagus
17 Celiac trunk and superior mesenteric artery
18 Aorta
19 Costal part of diaphragm
20 Costal margin
21 Transversus abdominis muscle

Diaphragm. Paramedian section to the right of the median plane through thoracic and upper abdominal cavity. The plane passes through the superior and inferior vena cava just to the right of the vertebral bodies. Most of the heart remains in situ to the left of this plane (specimen is viewed from the right side).

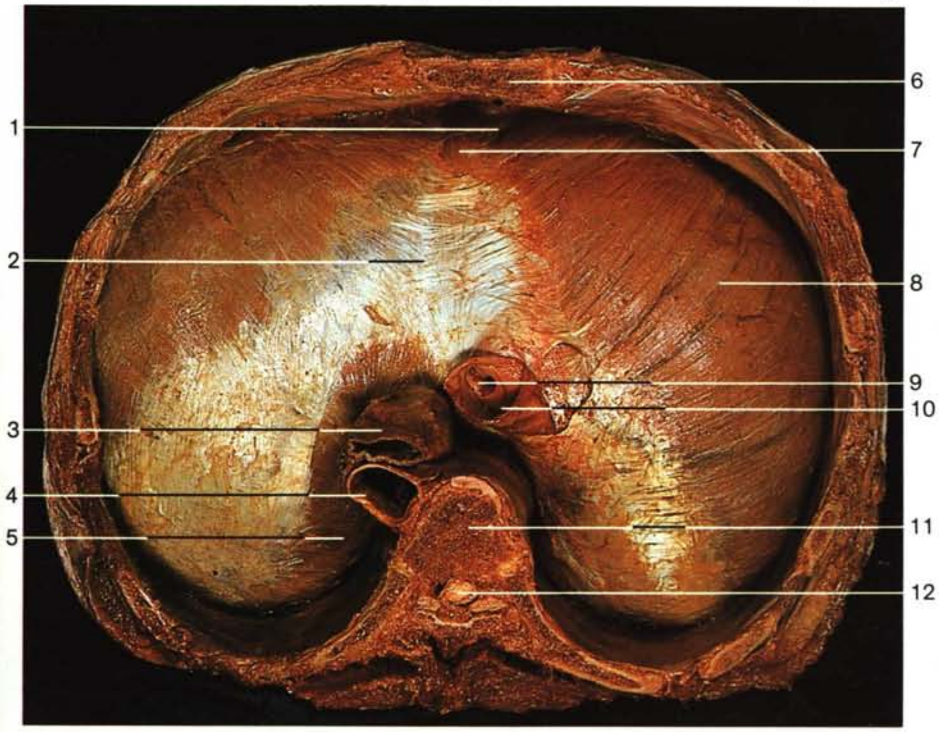

1 Sternocostal triangle
2 Central tendon (from above)
3 Esophagus
4 Aorta
5 Lumbar part of diaphragm
6 Sternum
7 Sternal part of diaphragm
8 Costal part of diaphragm
9 Entrance of hepatic veins
10 Inferior vena cava
11 Body of 9th thoracic vertebra
12 Spinal cord

Diaphragm (superior aspect). The pleura and thoracic wall have been removed.

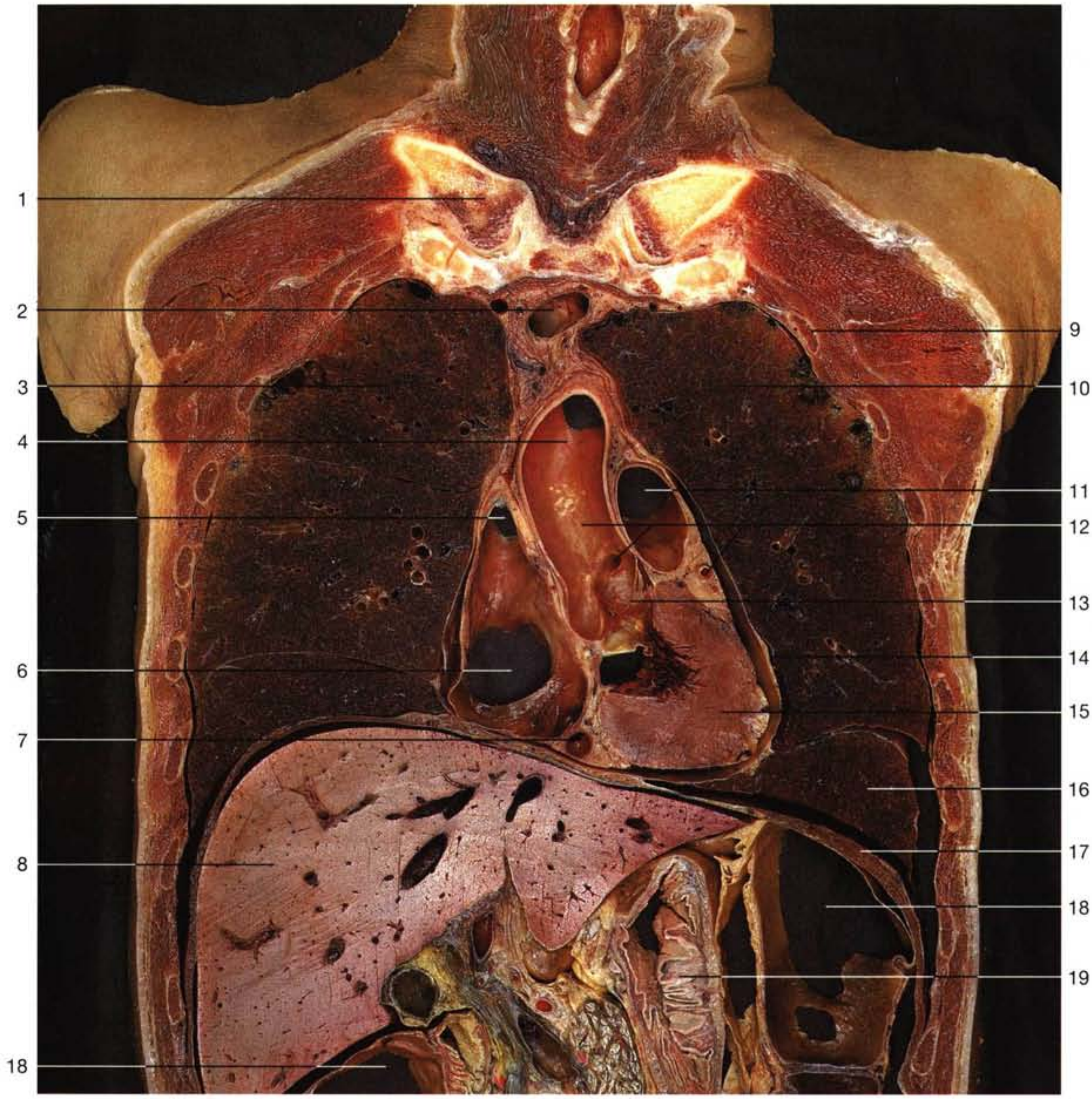

Coronal section through the thorax at the level of ascending aorta (anterior aspect).

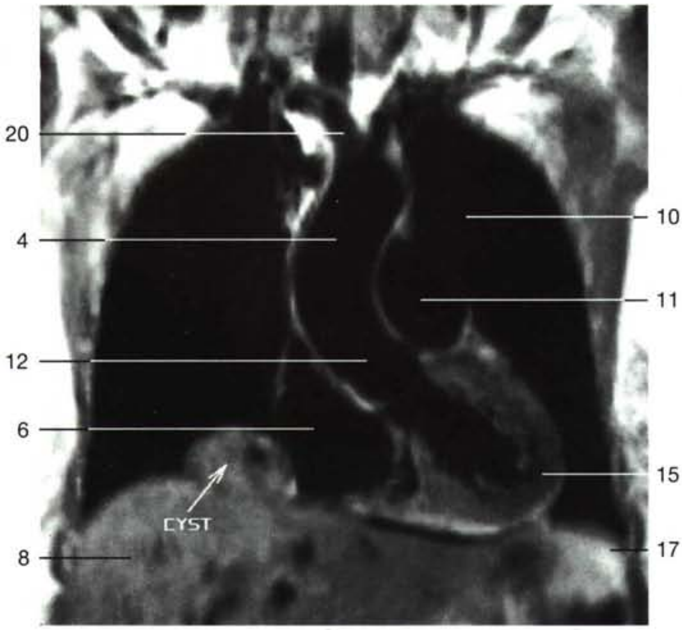

Coronal section through the thorax at the level of ascending aorta. (MR-Scan.)

1	Clavicle
2	Left brachiocephalic vein
3	Superior lobe of right lung
4	Aortic arch
5	Superior vena cava
6	Right atrium (entrance of inferior vena cava)
7	Coronary sinus
8	Liver
9	Second rib
10	Superior lobe of left lung
11	Pulmonary trunk
12	Ascending aorta and left coronary artery
13	Aortic valve
14	Pericardium
15	Myocardium of left ventricle
16	Lower lobe of left lung
17	Diaphragm
18	Colic flexures
19	Stomach
20	Brachiocephalic trunk

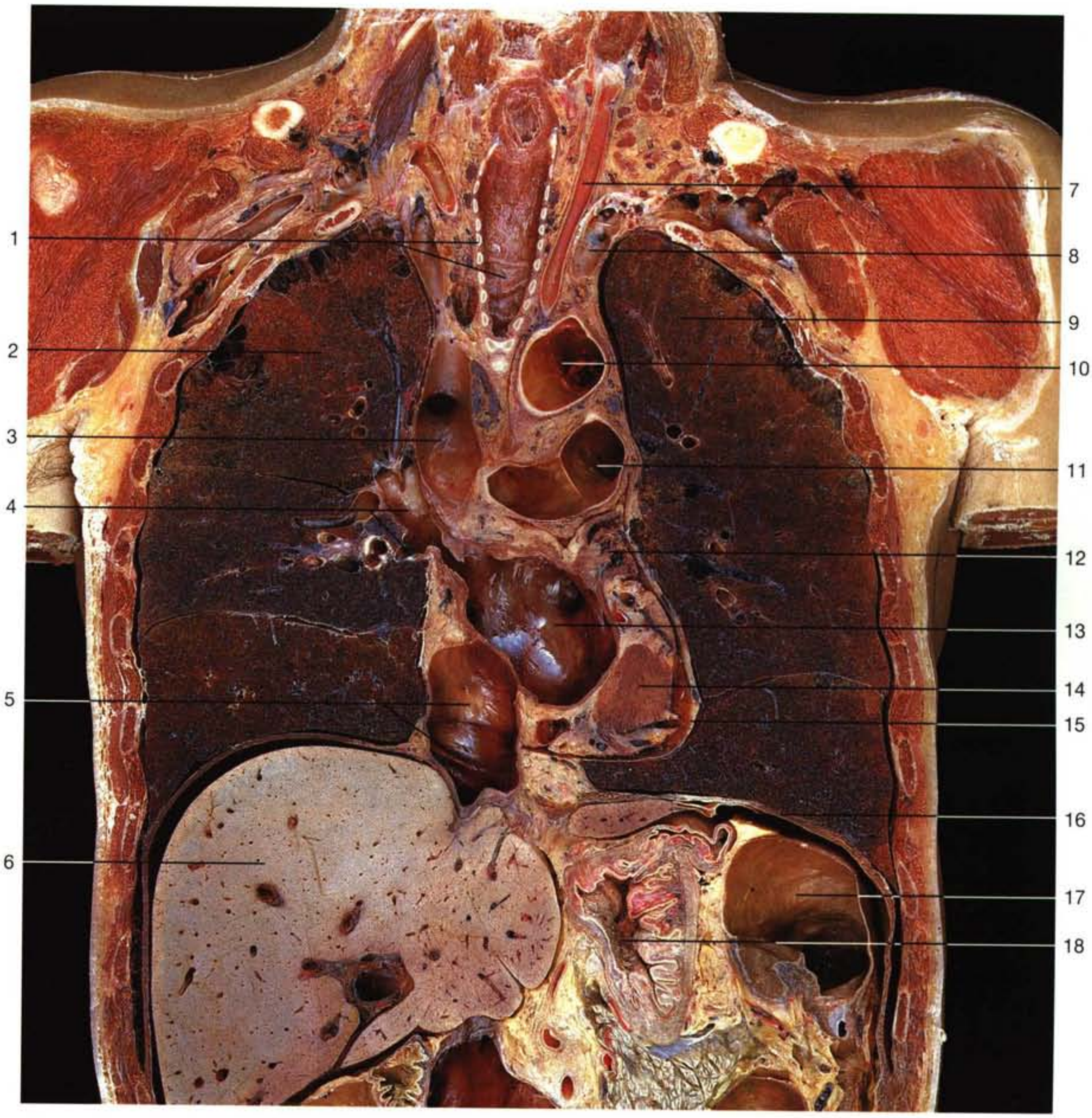

Coronal section through the thorax at the level of superior and inferior vena cava (anterior aspect).

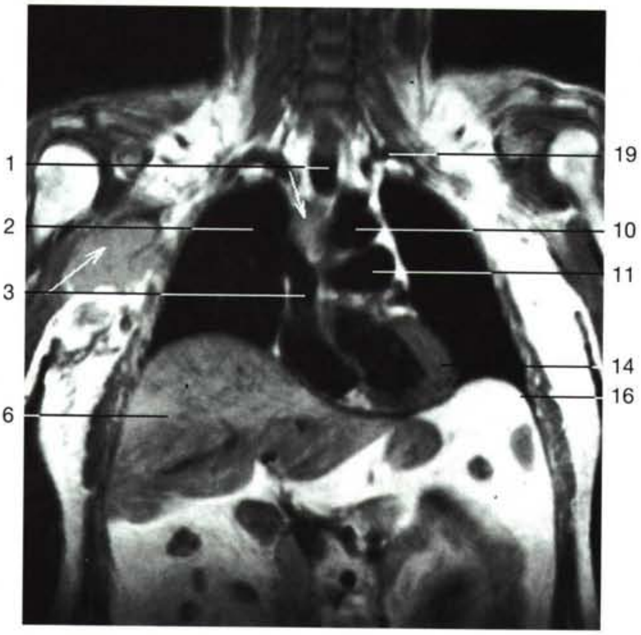

Coronal section through the thorax at the level of superior vena cava. (MR-Scan.)

1 Trachea
2 Upper lobe of right lung
3 **Superior vena cava**
4 Right pulmonary veins
5 Inferior vena cava and right atrium
6 Liver
7 Left common carotid artery
8 Left subclavian vein
9 Upper lobe of left lung
10 Aortic arch
11 Left pulmonary artery
12 Left auricle
13 **Left atrium** with orifices of pulmonary veins
14 Left ventricle (myocardium)
15 Pericardium
16 Diaphragm
17 Left colic flexure
18 Stomach
19 Left subclavian artery

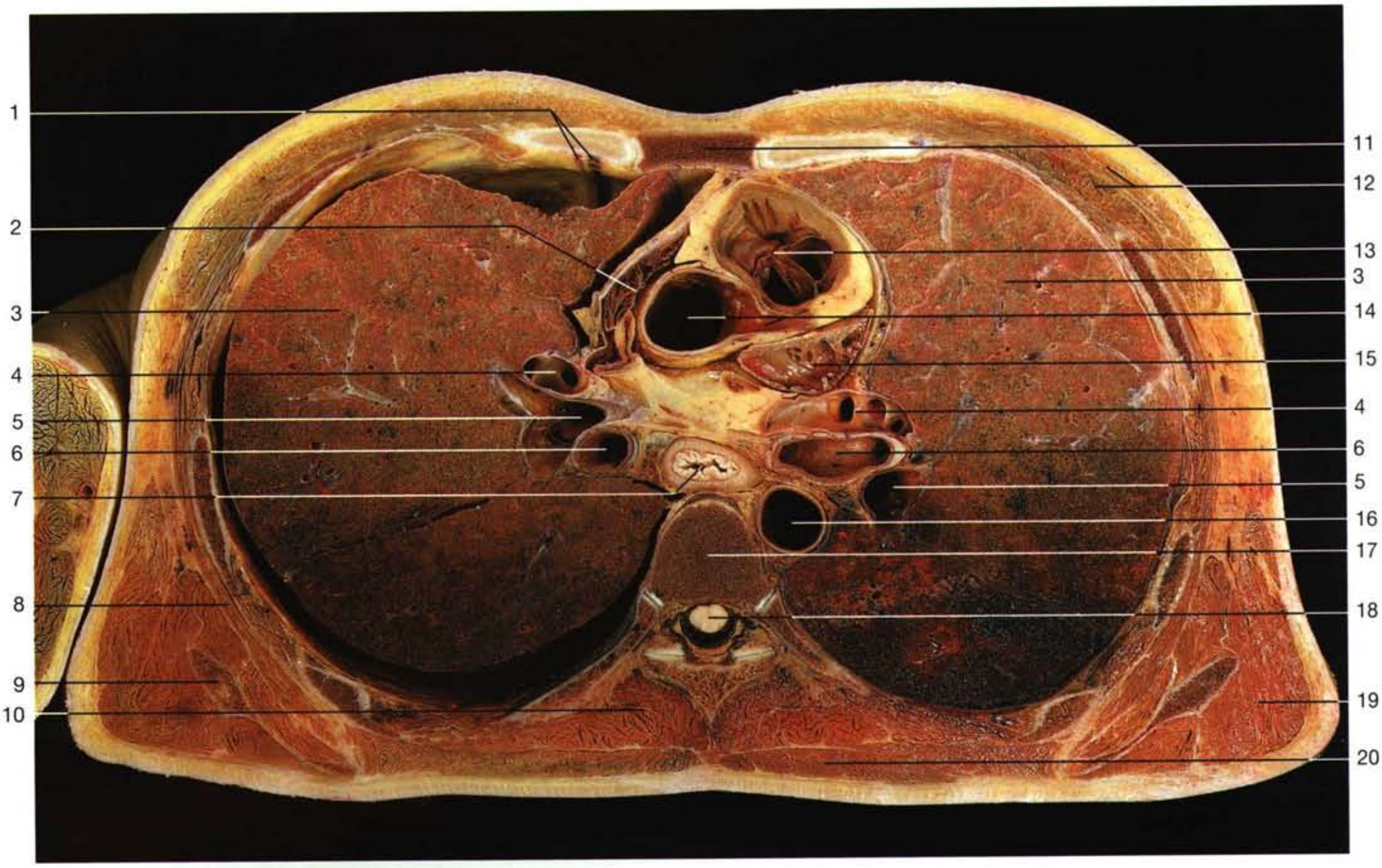

Horizontal section through the thorax at level 1 (from below).

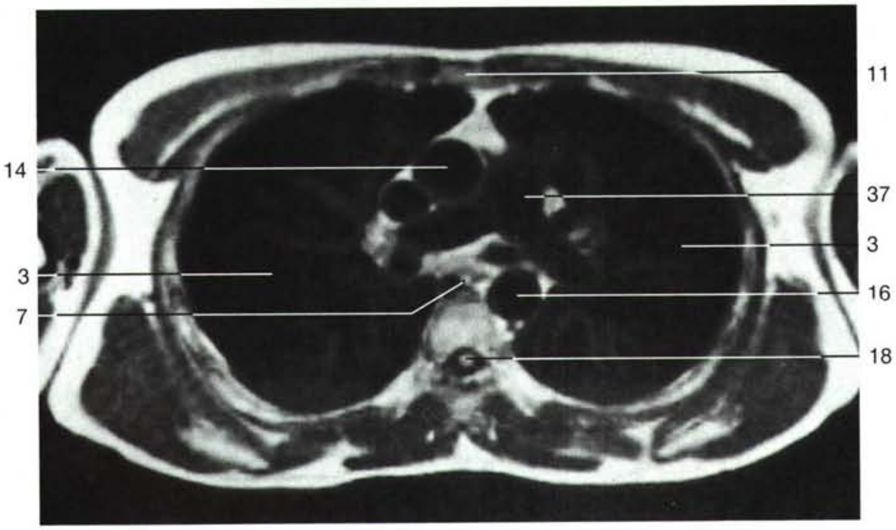

Horizontal section through the thorax at level 1 (from below). (MR-Scan.)

1 Internal thoracic artery and vein
2 Right atrium
3 Lung
4 Pulmonary artery
5 Pulmonary vein
6 Primary bronchus
7 Esophagus
8 Serratus anterior muscle
9 Scapula
10 Longissimus thoracis muscle
11 Sternum
12 Pectoralis major and minor muscles
13 Conus arteriosus (right ventricle), pulmonic valve
14 Ascending aorta and left coronary artery (only in upper figure)
15 **Left atrium**
16 **Descending aorta**
17 Thoracic vertebra
18 Spinal cord
19 Latissimus dorsi muscle
20 Trapezius muscle

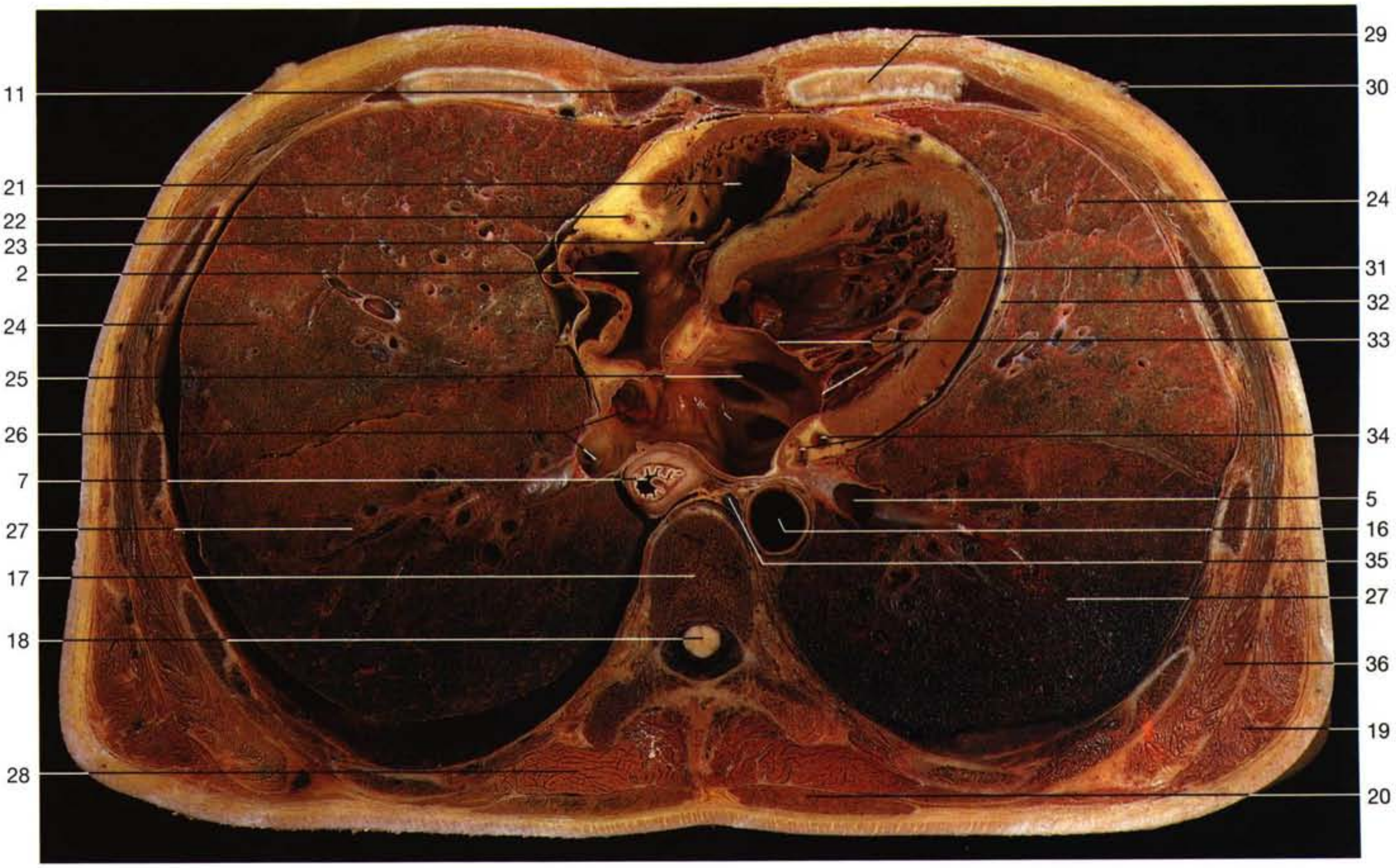
11
21
22
23
2
24
25
26
7
27
17
18
28
29
30
24
31
32
33
34
5
16
35
27
36
19
20

Horizontal section through the thorax at level 2 (from below).

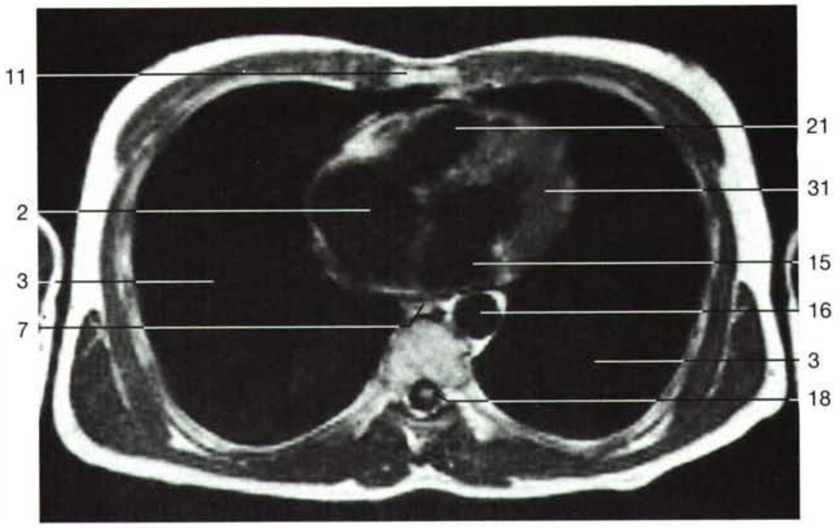
11
2
3
7
21
31
15
16
3
18

Horizontal section through the thorax at level 2 (from below). (MR-Scan.)

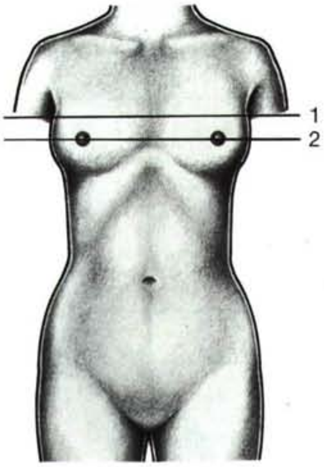

Levels of sections.

21 **Right ventricle**
22 Right coronary artery
23 Right atrioventricular valve
24 Lung (upper lobe)
25 **Left atrium**
26 Pulmonary veins
27 Lung (lower lobe)
28 Erector muscle of spine
29 Third costal cartilage
30 Nipple

31 **Left ventricle**
32 Pericardium
33 Left atrioventricular valve
34 Left coronary artery and coronary sinus
35 Accessory hemiazygos vein
36 Serratus anterior muscle
37 Pulmonary trunk

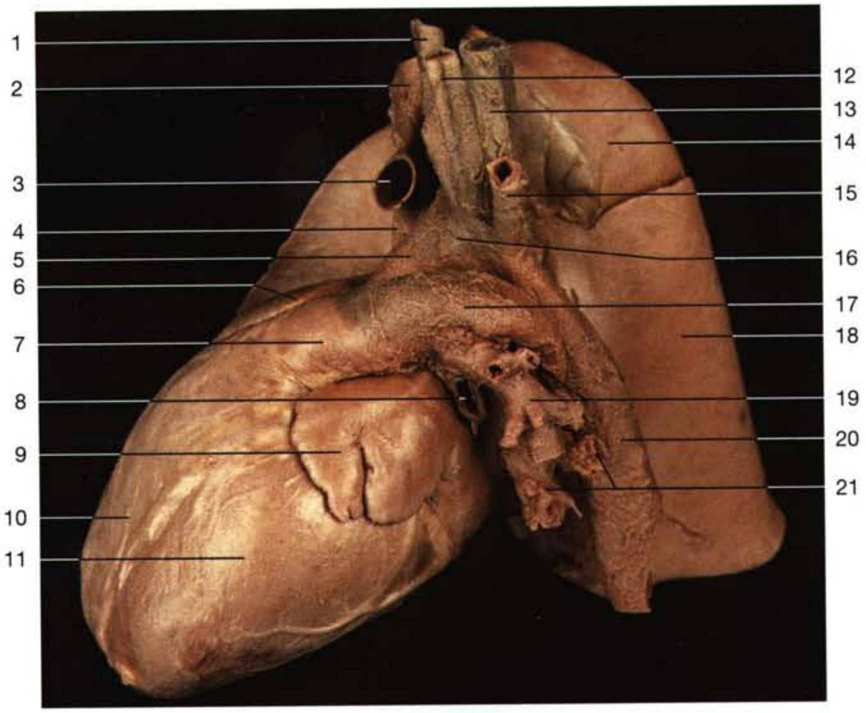

Heart and right lung of the fetus (viewed from left side). The left lung has been removed. Note the ductus arteriosus (Botalli).

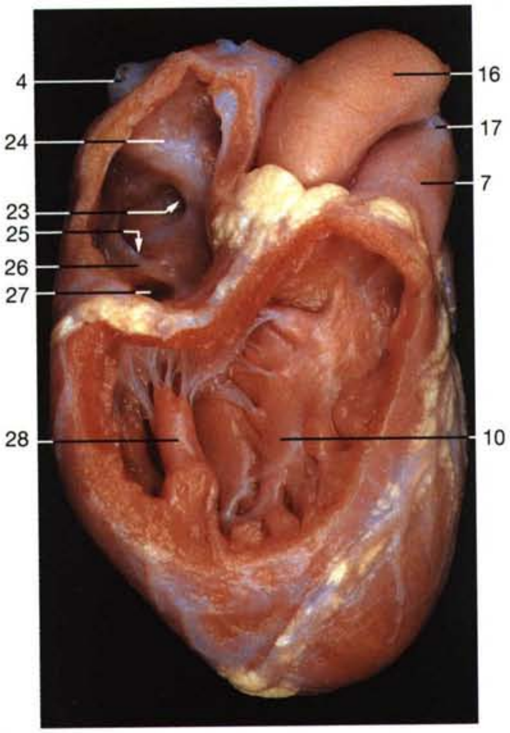

Heart of the fetus (anterior aspect). Right atrium and ventricle opened.

Shunts in the fetal circulation system		
1. Ductus venosus (of Arantius)	between umbilical vein and inferior vena cava	bypass of liver circulation
2. Foramen ovale	between right and left atrium	bypass of pulmonary circulation
3. Ductus arteriosus (Botalli)	between pulmonary trunk and aorta	

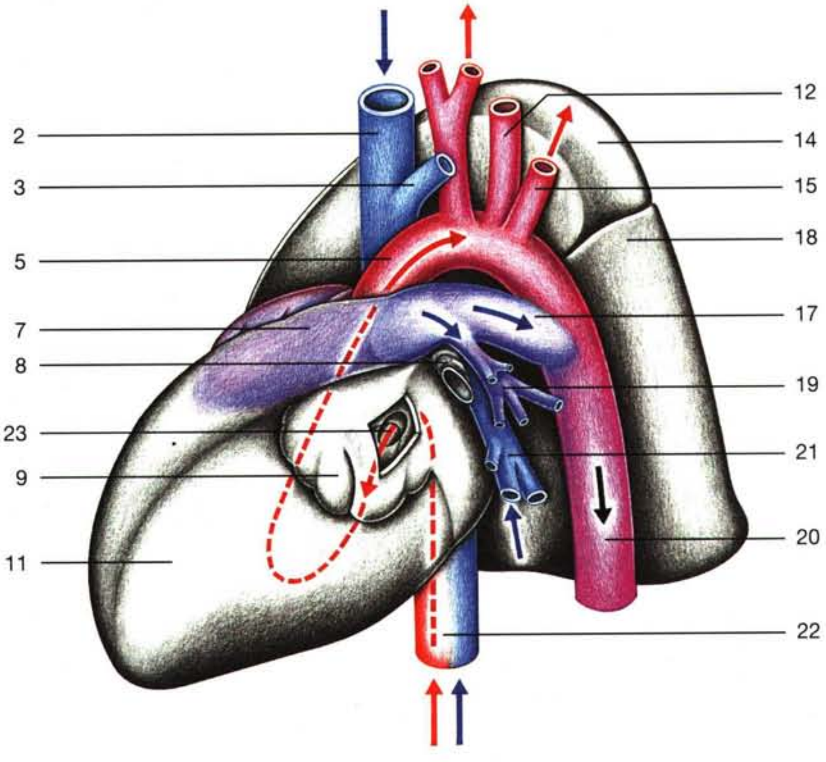

◁ **Heart of the fetus** (schematic drawing). Direction of blood flow indicated by arrows. Note the change in oxygenation of blood after ductus arteriosus entry into aorta.

1 Right common carotid artery
2 Right brachiocephalic vein
3 Left brachiocephalic vein
4 Superior vena cava
5 **Ascending aorta**
6 Right auricle
7 **Pulmonary trunk**
8 Left primary bronchus
9 Left auricle
10 Right ventricle
11 Left ventricle
12 Left common carotid artery
13 Trachea
14 Superior lobe of right lung
15 Left subclavian artery
16 Aortic arch
17 **Ductus arteriosus (Botalli)**
18 Inferior lobe of right lung
19 Left pulmonary artery with branches to the left lung
20 **Descending aorta**
21 Left pulmonary veins
22 Inferior vena cava
23 **Foramen ovale**
24 Right atrium
25 Opening of inferior vena cava
26 Valve of inferior vena cava (Eustachian valve)
27 Opening of coronary sinus
28 Anterior papillary muscle of right ventricle

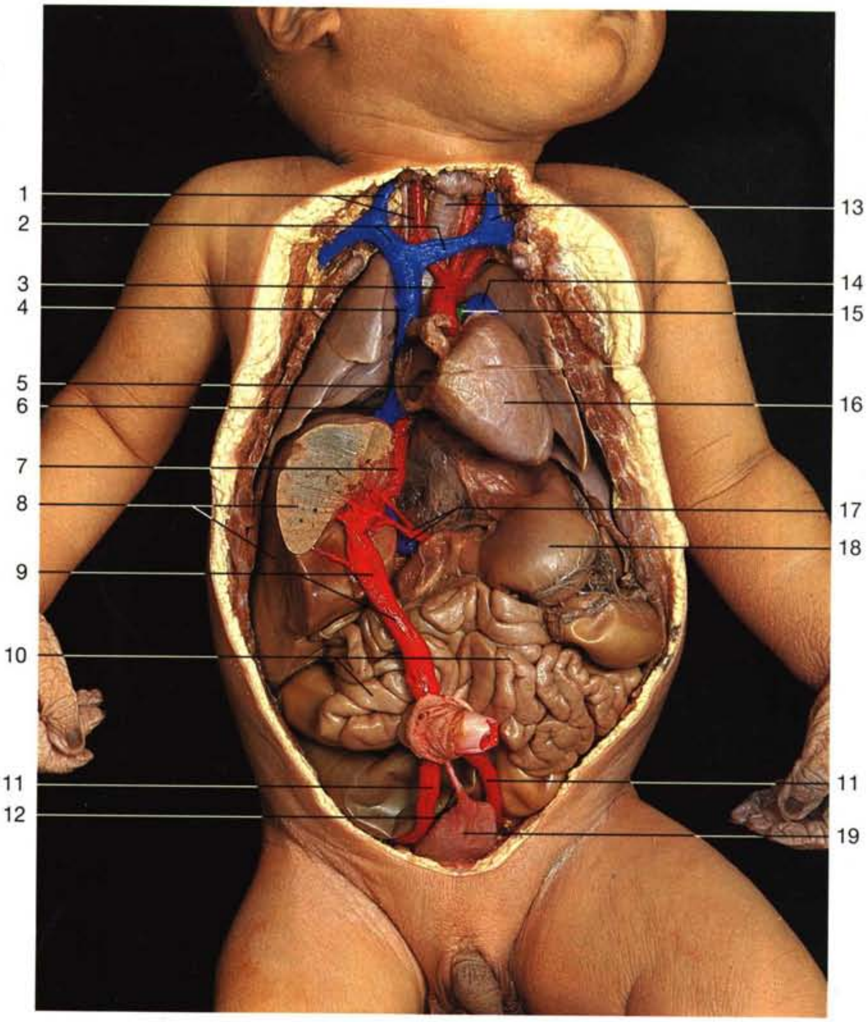

Thoracic and abdominal organs in the newborn (anterior aspect). The right atrium has been opened to show the foramen ovale. The left lobe of the liver has been removed.

1 Internal jugular vein and right common carotid artery
2 Right and left brachiocephalic vein
3 Aortic arch
4 Superior vena cava
5 **Foramen ovale**
6 Inferior vena cava
7 **Ductus venosus**
8 Liver
9 **Umbilical vein**
10 Small intestine
11 **Umbilical artery**
12 Urachus
13 Trachea and left internal jugular vein
14 Left pulmonary artery
15 **Ductus arteriosus** (Botalli)
16 Right ventricle
17 Hepatic arteries (red) and portal vein (blue)
18 Stomach
19 Urinary bladder
20 Portal vein
21 Pulmonary veins
22 Descending aorta
23 Placenta

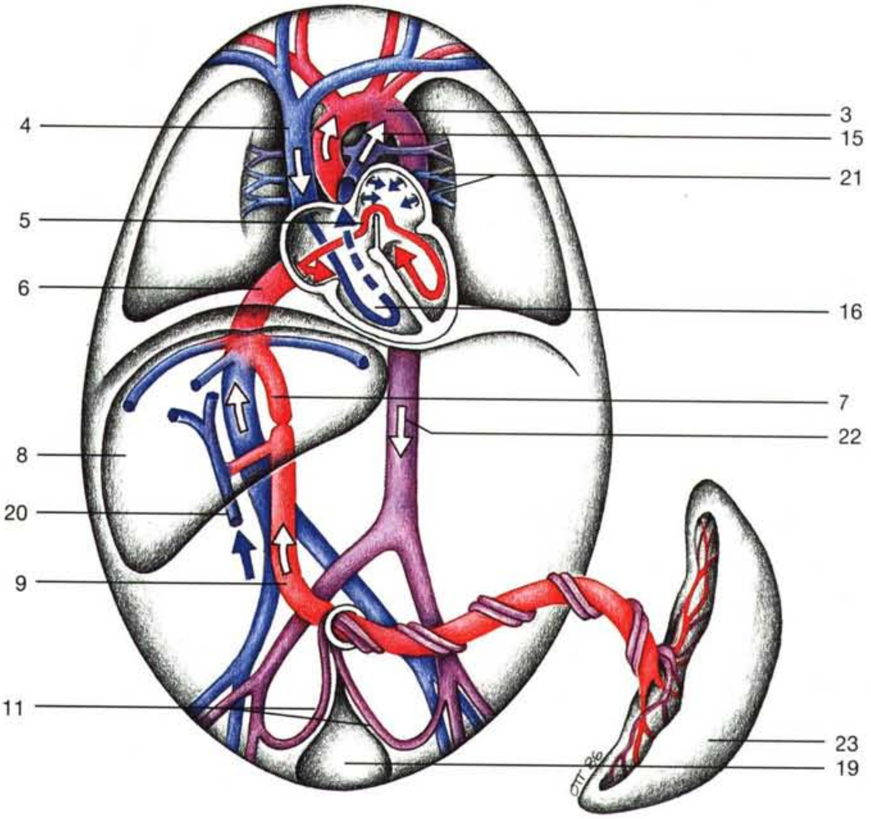

◁ **Fetal circulatory system.** (Schematic drawing.) The oxygen gradient is indicated by color.

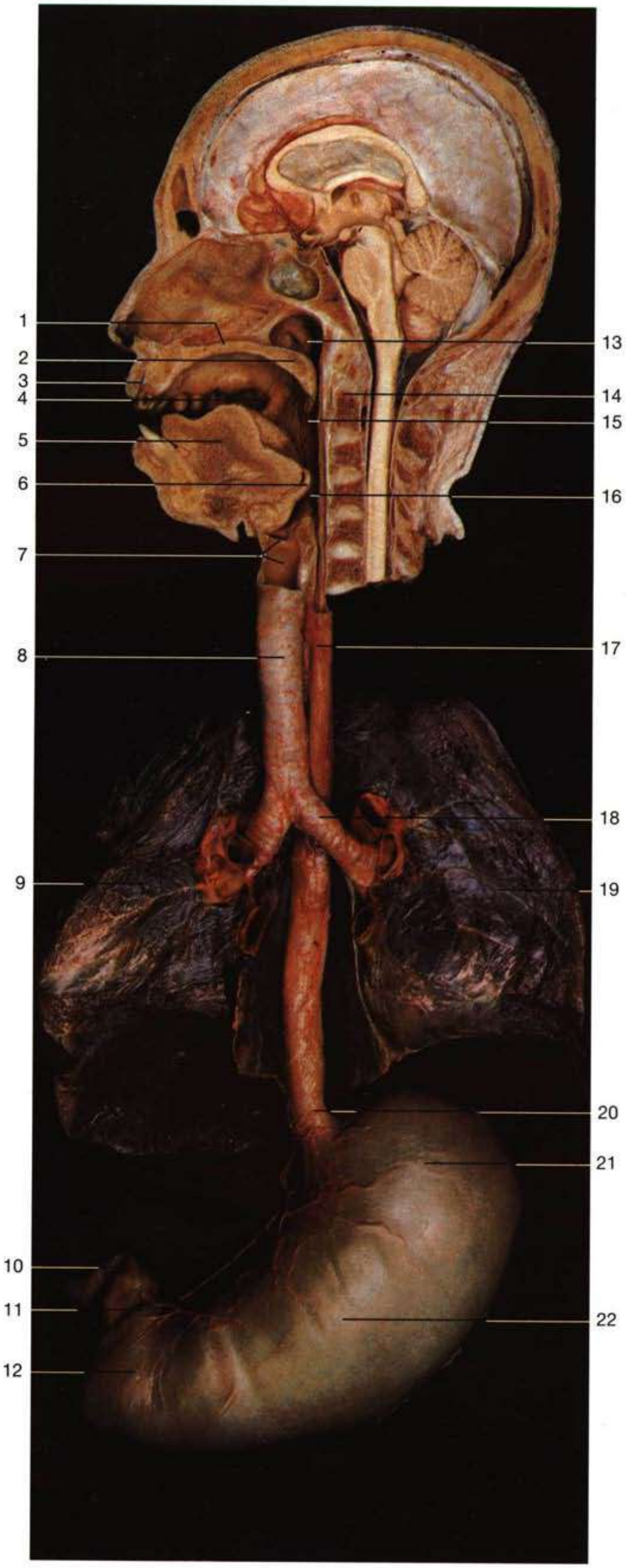

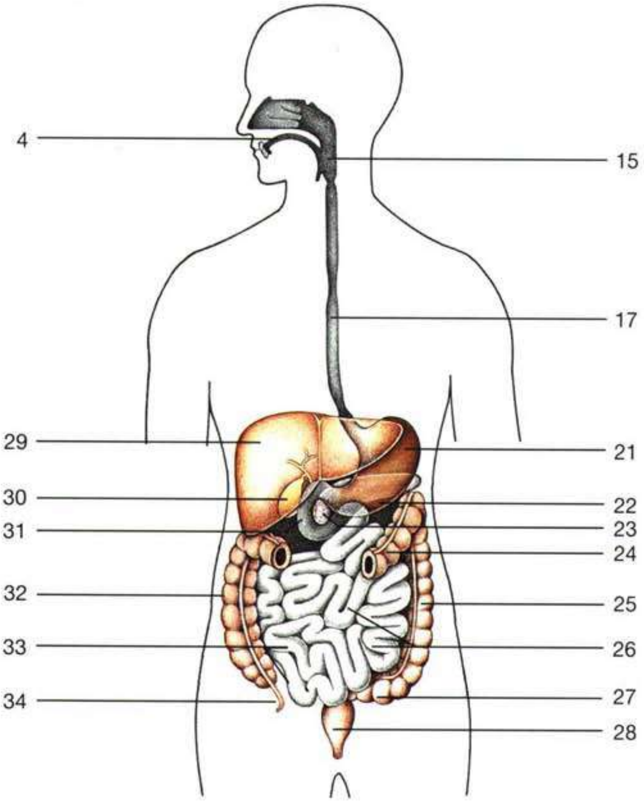

Organization of digestive system. Position of abdominal organs.

1 Hard palate
2 Soft palate with uvula
3 Vestibule of the mouth
4 Oral cavity proper
5 Tongue
6 Epiglottis
7 Vocal ligament and larynx
8 Trachea
9 Right lung
10 Superior part of duodenum
11 Pylorus
12 Pyloric antrum
13 Nasopharynx
14 Dens of axis
15 Oropharynx
16 Laryngopharynx
17 Esophagus (thoracic part)
18 Left primary bronchus
19 Left lung
20 Abdominal part of esophagus and cardia
21 Fundus of stomach
22 Body of stomach
23 Pancreas
24 Transverse colon (divided)
25 Descending colon
26 Jejunum
27 Sigmoid colon
28 Rectum
29 Liver
30 Gall bladder
31 Duodenum
32 Ascending colon
33 Ileum
34 Vermiform appendix

Survey of upper portion of digestive system. Oral cavity, pharynx, esophagus, and stomach.

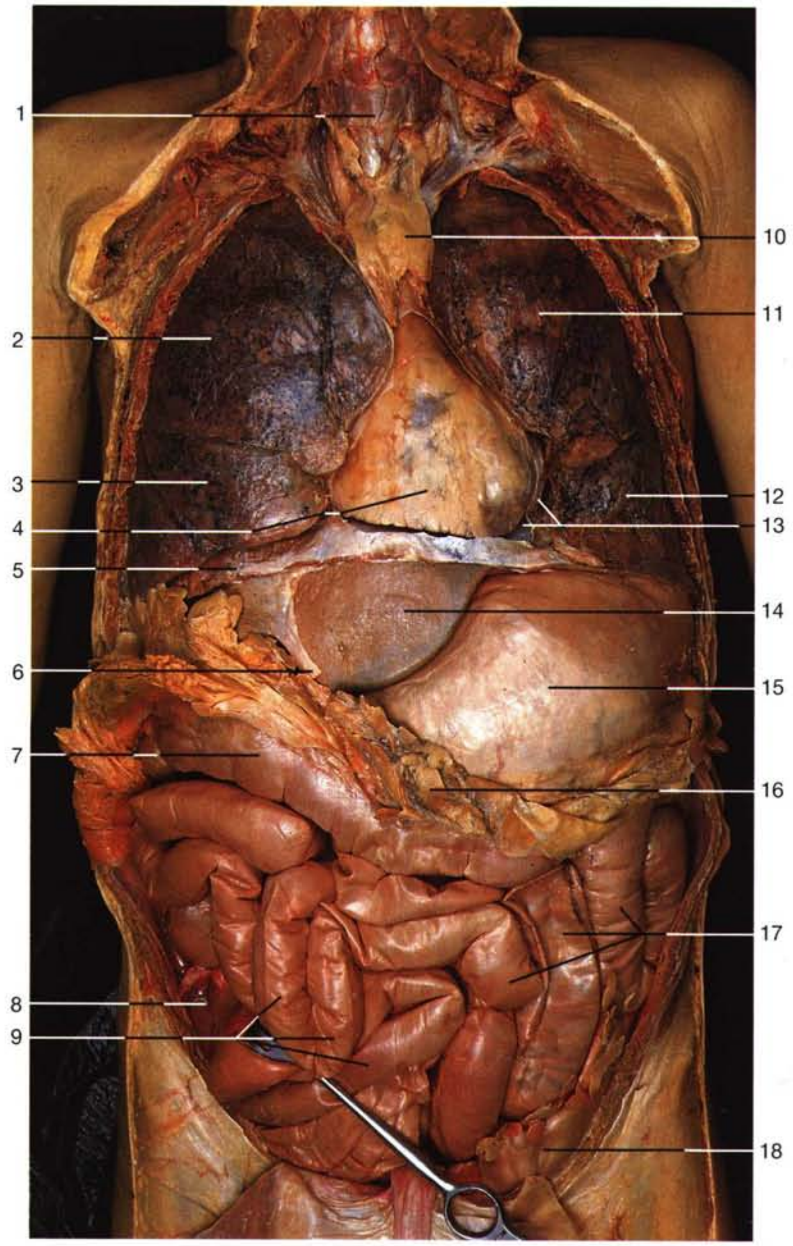

Abdominal organs in situ. The greater omentum has been partly removed or reflected.

1 Thyroid gland
2 Upper lobe of right lung
3 Middle lobe of right lung
4 Heart
5 Diaphragm
6 Round ligament of liver (ligamentum teres)
7 **Transverse colon**
8 Cecum
9 Small intestine (ileum)
10 Thymus
11 Upper lobe of left lung
12 Lower lobe of left lung
13 Pericardium (cut edge)
14 **Liver** (left lobe)
15 **Stomach**
16 Greater omentum
17 **Small intestine** (jejunum)
18 Sigmoid colon
19 Rectus abdominis muscle
20 Small intestine (section)
21 Rib
22 Common bile duct, duodenum and pancreas
23 Inferior vena cava
24 Liver
25 Body of second lumbar vertebra
26 Right kidney
27 Cauda equina and dura mater
28 Linea alba
29 Stomach and pylorus
30 Superior mesenteric artery and vein
31 Abdominal aorta
32 Left renal artery and vein
33 Left kidney
34 Psoas major muscle
35 Deep muscles of the back
36 Pancreas adjacent to lesser sac (bursa omentalis)
37 Falciform ligament with ligamentum teres

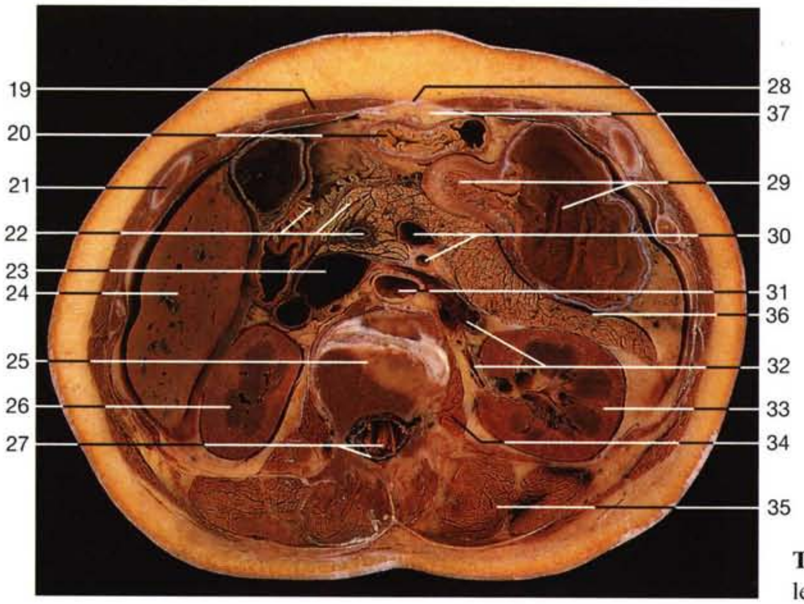

Transverse section through the abdominal cavity at the level of the second lumbar vertebra (from below).

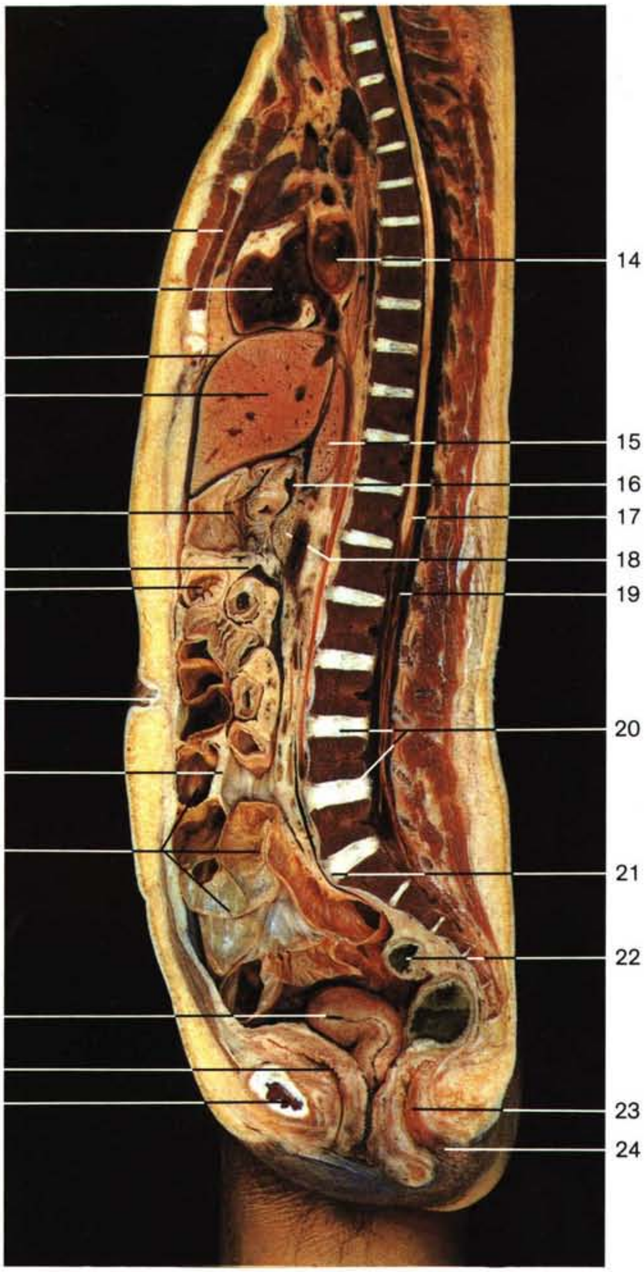
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Midsagittal section through the trunk (female).

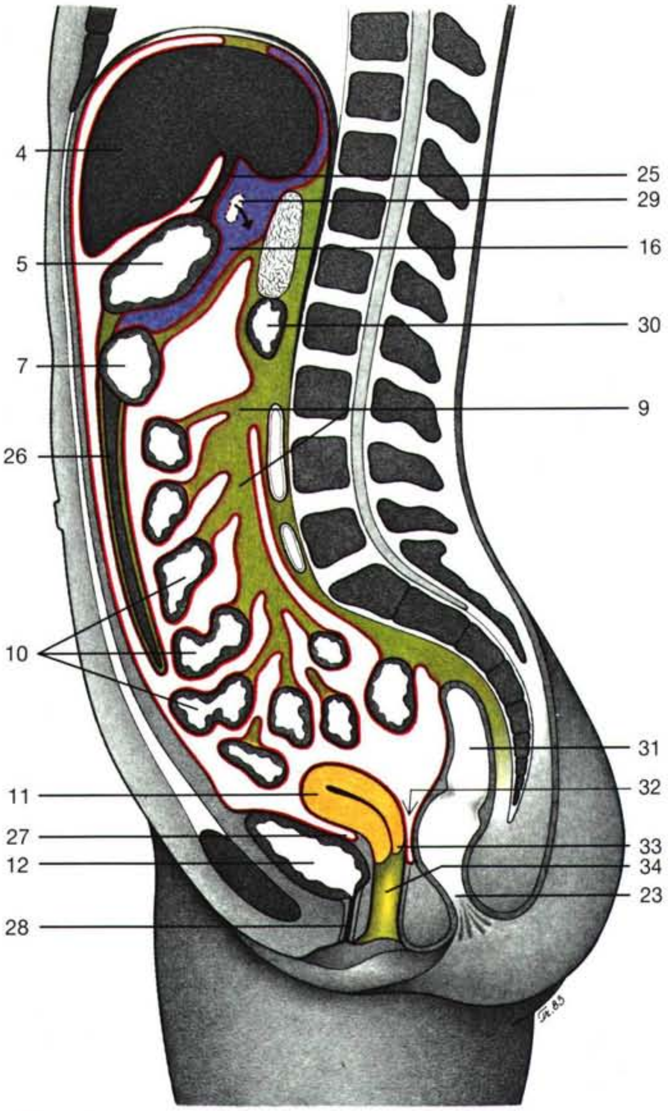
4
5
7
26
10
11
27
12
28
25
29
16
30
9
31
32
33
34
23

Midsagittal section through the tunk (female).
(Schematic drawing.)
Blue = omental bursa; red = peritoneum.

1 Sternum
2 Right ventricle of heart
3 Diaphragm
4 Liver
5 Stomach
6 Transverse mesocolon
7 Transverse colon
8 Umbilicus
9 Mesentery
10 Small intestine
11 Uterus
12 Urinary bladder
13 Pubic symphysis
14 Left atrium of heart
15 Caudate lobe of liver
16 Omental bursa or lesser sac
17 Conus medullaris
18 Pancreas
19 Cauda equina
20 Intervertebral discs (lumbar vertebral column)
21 Sacral promontory
22 Sigmoid colon
23 Anal canal
24 Anus
25 Lesser omentum
26 Greater omentum
27 Vesicouterine pouch
28 Urethra
29 Epiploic (omental) foramen
30 Duodenum
31 Rectum
32 Rectouterine pouch
33 Vaginal part of cervix of uterus
34 Vagina

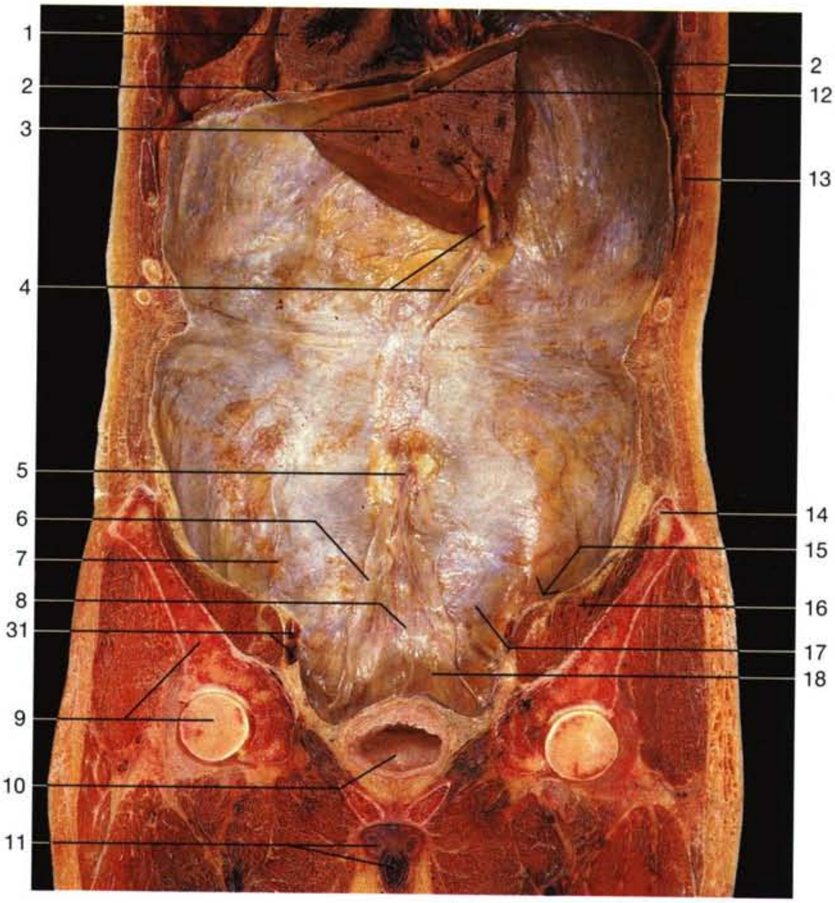

Anterior abdominal wall with pelvic cavity and thigh (frontal section, male) (internal aspect).

1 Left ventricle with pericardium
2 Diaphragm
3 Remnant of liver
4 Ligamentum teres (free margin of falciform ligament)
5 Site of umbilicus
6 **Medial umbilical fold** (containing the obliterated umbilical artery)
7 **Lateral umbilical fold** (containing inferior epigastric artery and vein)
8 **Median umbilical fold** (containing remnant of urachus)
9 Head of femur and pelvic bone
10 Urinary bladder
11 Root of penis
12 Falciform ligament of liver
13 Rib (divided)
14 Iliac crest (divided)
15 Site of **deep inguinal ring** and lateral inguinal fossa
16 Iliopsoas muscle (divided)
17 Medial inguinal fossa
18 Supravesical fossa
19 Posterior layer of rectus sheath
20 Transversus abdominis muscle
21 Umbilicus and arcuate line
22 **Inferior epigastric artery**
23 Femoral nerve
24 Iliopsoas muscle
25 Remnant of umbilical artery
26 Femoral artery and vein
27 Tendinous intersection of rectus abdominis muscle
28 Rectus abdominis muscle
29 Interfoveolar ligament
30 Pubic symphysis (divided)
31 External iliac artery and vein

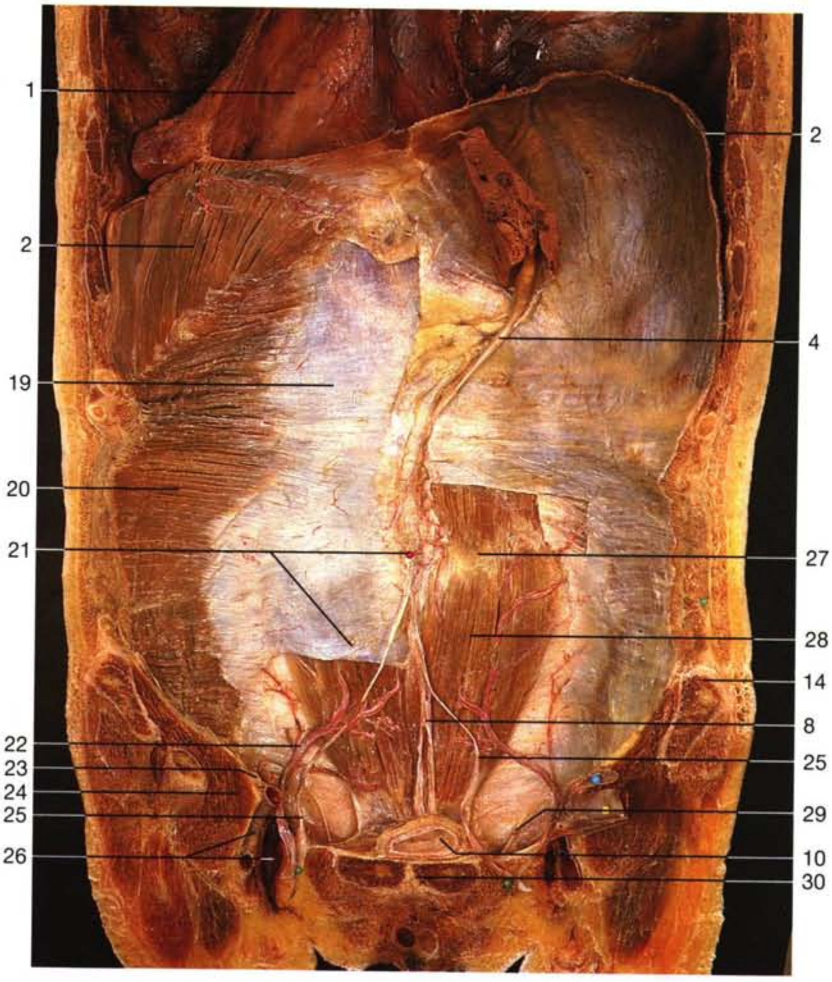

Anterior abdominal wall (male) (internal aspect). The peritoneum and parts of the posterior layer of rectus sheath have been removed. Dissection of inferior epigastric arteries and veins.

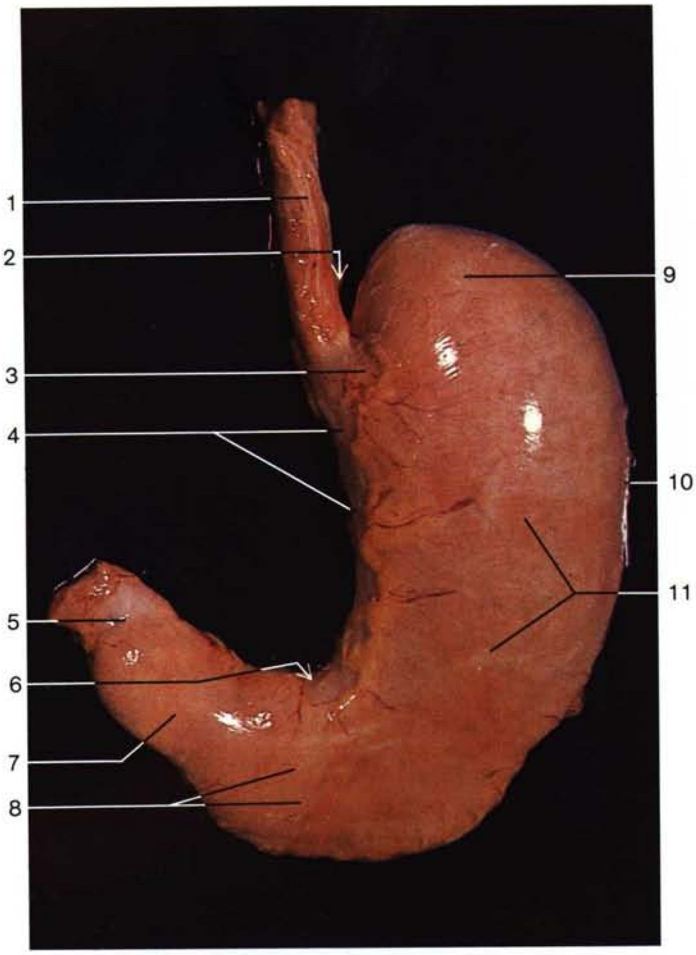

Stomach (ventral aspect).

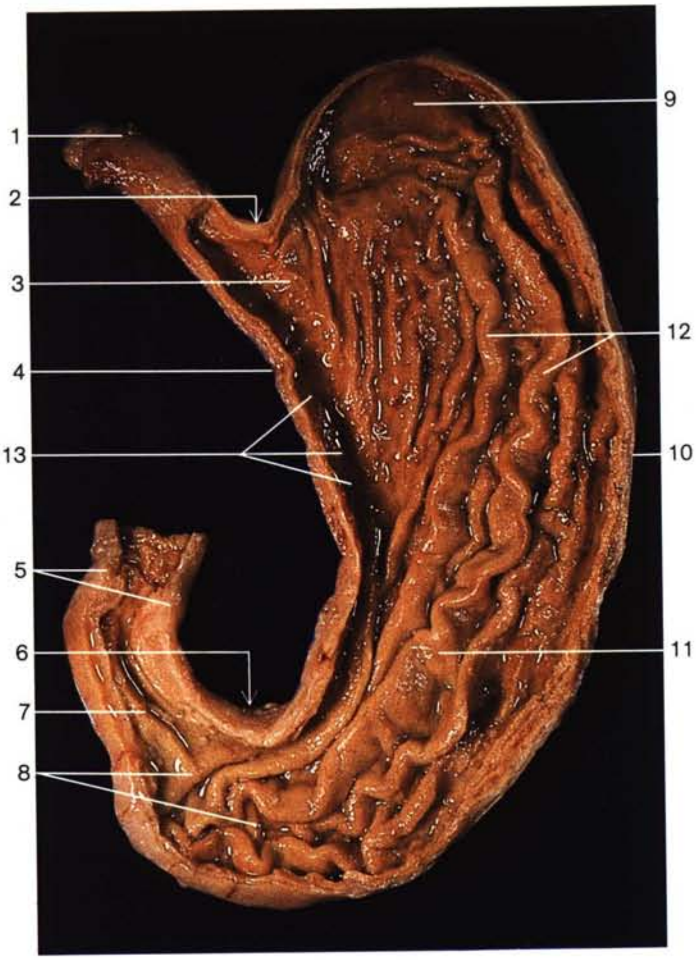

Mucosa of posterior wall of stomach (ventral aspect).

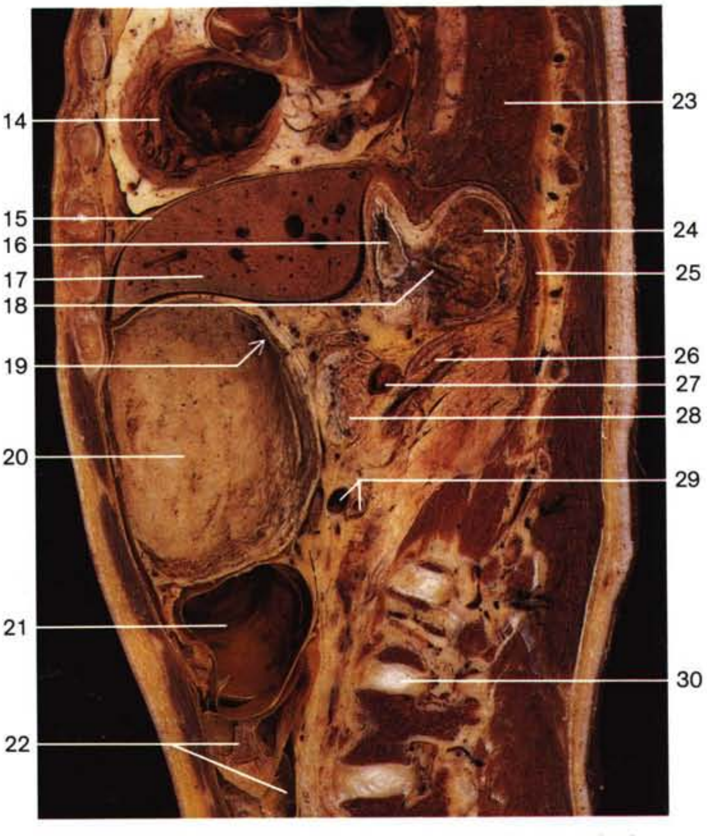

Parasagittal section through upper part of left abdominal cavity 3.5 cm lateral to median plane.

1 Esophagus
2 Cardiac notch
3 **Cardiac part of stomach**
4 Lesser curvature of stomach
5 **Pyloric sphincter**
6 Angular notch (incisura angularis)
7 **Pyloric canal**
8 Pyloric antrum
9 Fundus of stomach
10 Greater curvature of stomach
11 Body of stomach
12 Folds of mucous membrane (gastric rugae)
13 **Gastric canal**
14 Right ventricle of heart
15 Diaphragm (cut edge)
16 Abdominal portion of esophagus
17 **Liver**
18 Cardiac part of stomach (cut edge)
19 Position of pyloric canal
20 Body of stomach
21 Transverse colon
22 Small intestine
23 Lung (cut edge)
24 Fundus of stomach (section)
25 Lumbar portion of diaphragm (cut edge)
26 Suprarenal gland
27 Splenic vein
28 Pancreas
29 Superior mesenteric artery and vein
30 Intervertebral disc

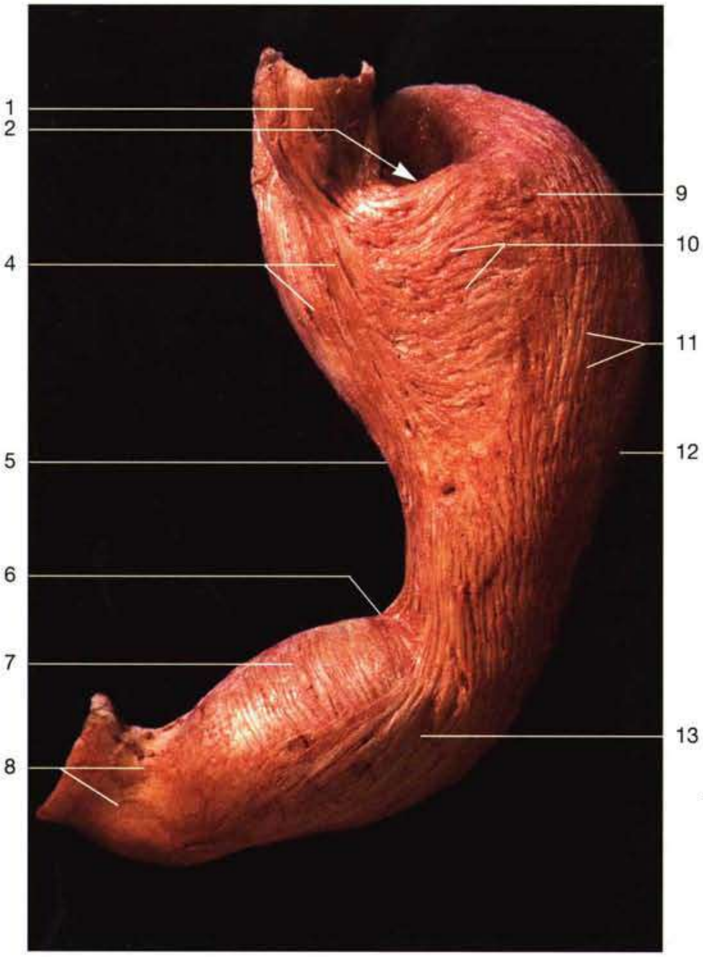

Muscular coat of stomach, outer layer (ventral aspect).

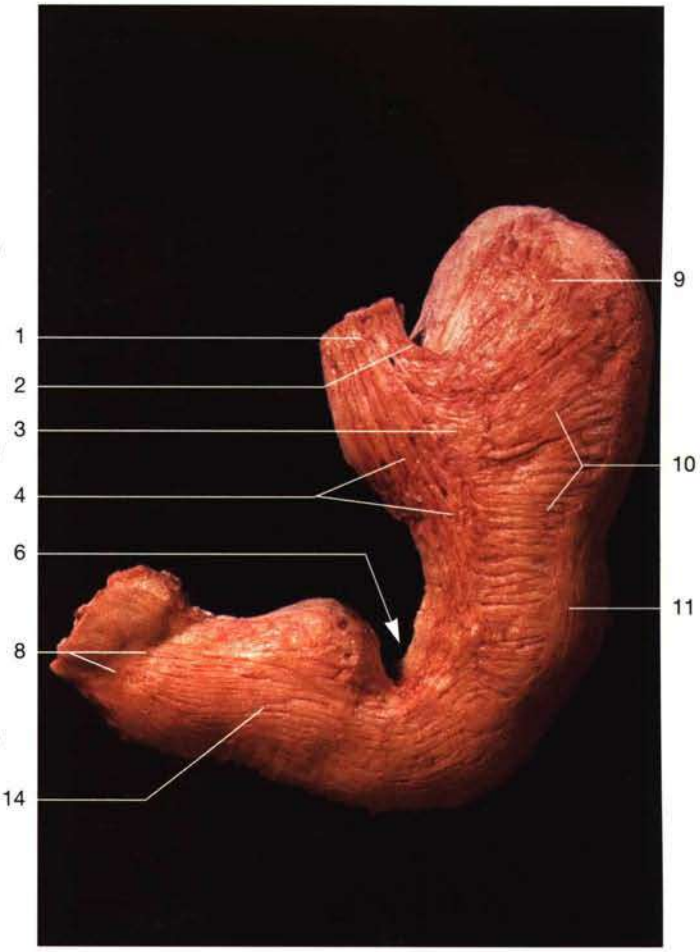

Muscular coat of stomach, middle layer (ventral aspect).

1 Esophagus (abdominal part)
2 Cardiac notch
3 Cardiac part of stomach
4 Longitudinal muscle layer at lesser curvature of stomach
5 Lesser curvature
6 Incisura angularis
7 Circular muscle layer of pyloric part of stomach
8 Pyloric sphincter muscle
9 Fundus of stomach
10 Circular muscle layer of fundus of stomach
11 Longitudinal muscle layer of greater curvature of stomach
12 Greater curvature of stomach
13 Longitudinal muscle layer (transition from body to pyloric part of stomach)
14 Pyloric part of stomach
15 Oblique muscle fibers

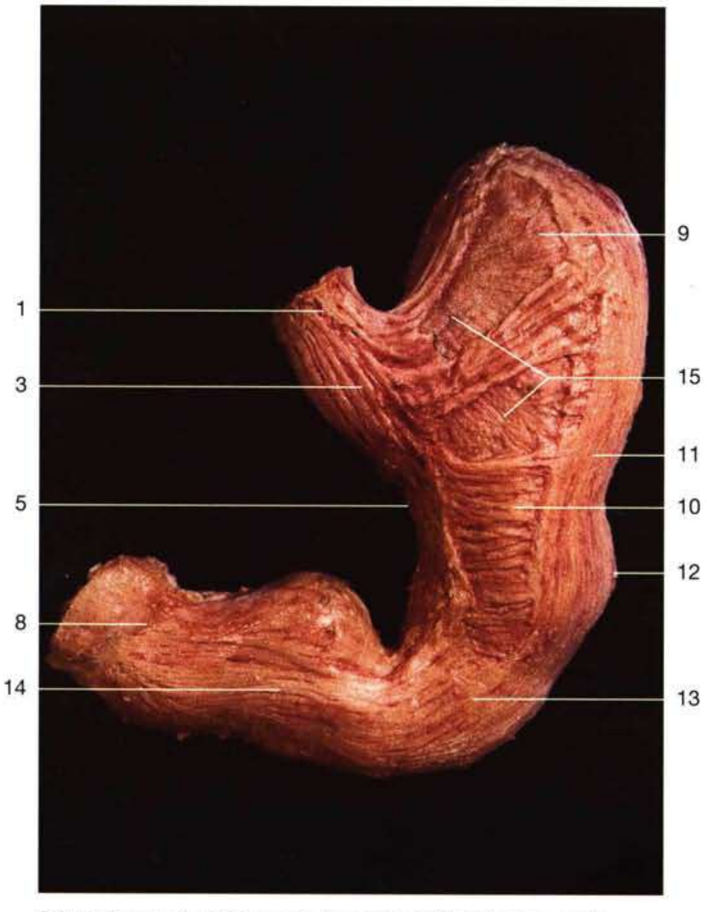

Muscular coat of stomach, inner layer (ventral aspect).

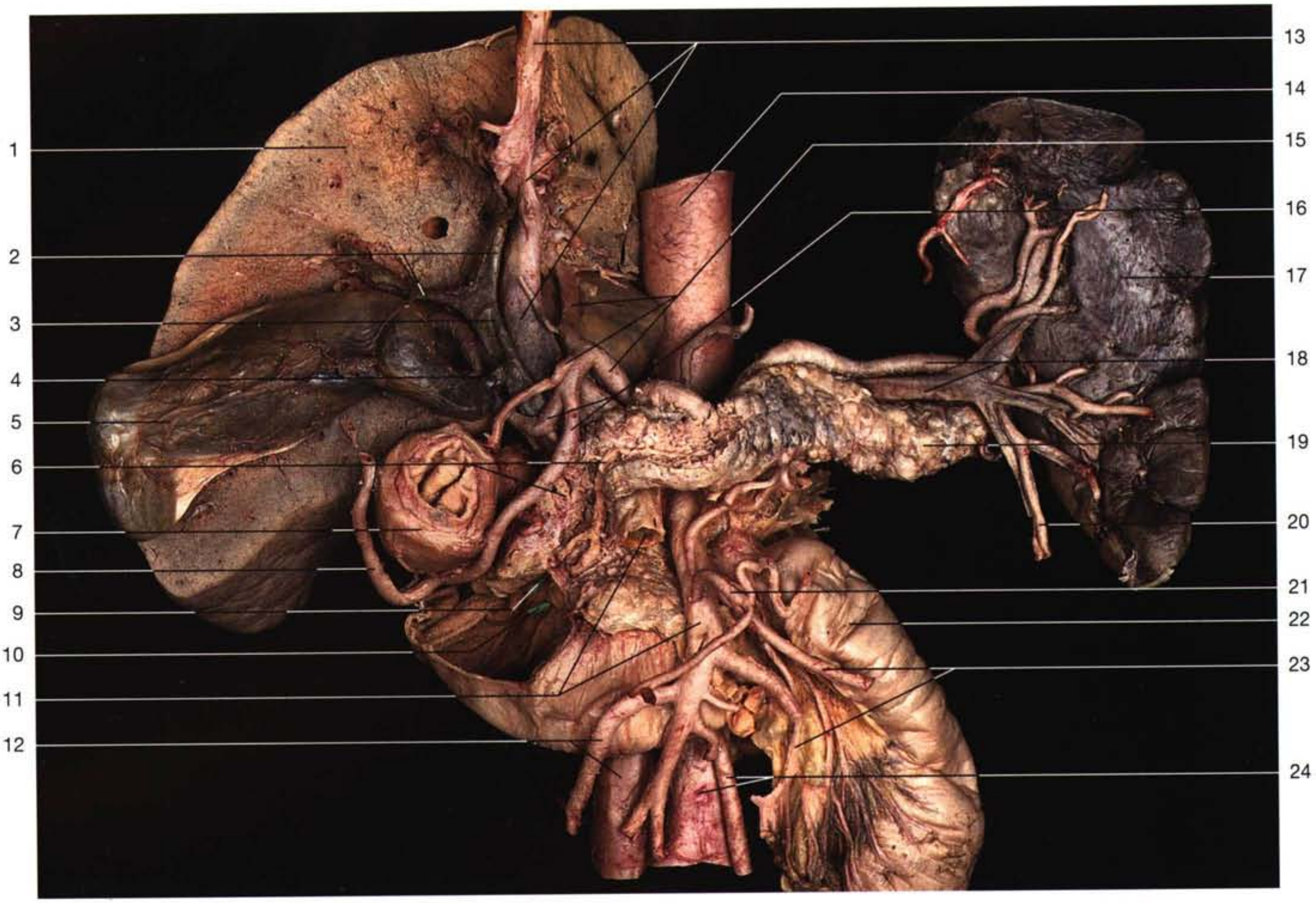

Pancreas with adjacent duodenum, spleen, liver and related arteries. The pancreatic ducts have been partly dissected. The duodenal papillae are indicated by probes. Left half of liver has been removed.

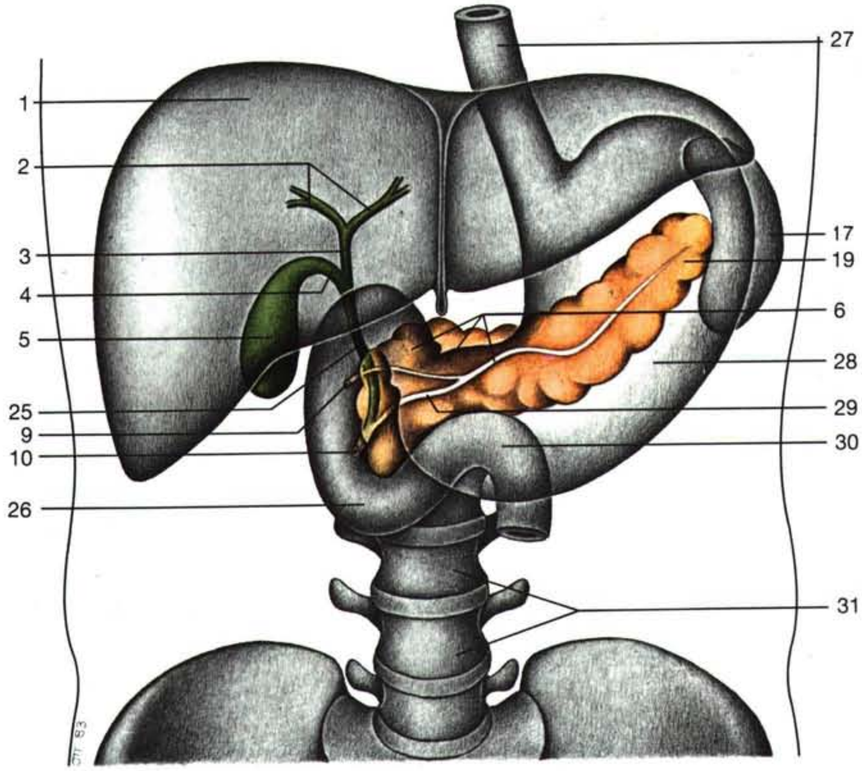

Pancreas, duodenum and extrahepatic bile ducts. (Schematic drawing.)

1 Liver
2 Left and right hepatic ducts
3 Common hepatic duct
4 **Cystic duct**
5 **Gall bladder**
6 **Pancreatic duct** and **head of pancreas**
7 Pylorus (cut)
8 Right gastro-omental (gastroepiploic) artery
9 Minor duodenal papilla (probe)
10 Major duodenal papilla (probe)
11 **Superior mesenteric artery and vein**
12 Inferior vena cava and iliocolic artery
13 **Portal vein,** lig. teres of liver and lig. venosum
14 Aorta
15 Common hepatic artery, left branch of hepatic artery proper, and right gastric artery
16 **Gastroduodenal artery** and right branch of hepatic artery proper
17 Spleen
18 Splenic artery and vein
19 **Tail of pancreas**
20 Left gastro-omental artery
21 Middle colic artery
22 Jejunum
23 Jejunal arteries
24 Aorta and inferior mesenteric artery
25 **Common bile duct**
26 Horizontal part of duodenum
27 Esophagus
28 Stomach
29 **Pancreatic duct**
30 Duodenojejunal flexure
31 Lumbal vertebrae

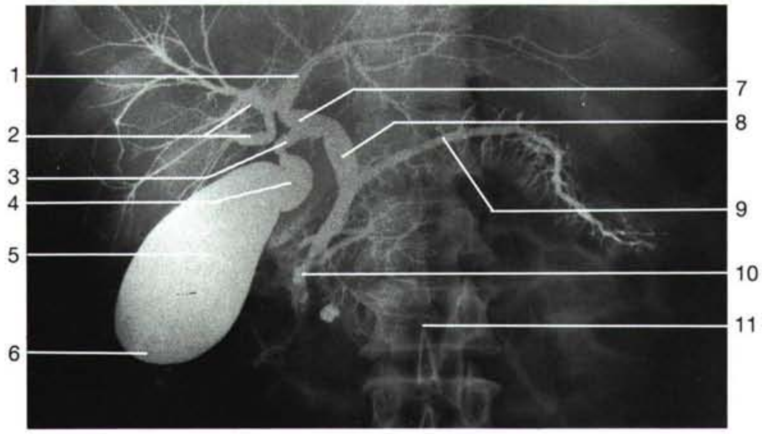

Radiograph of biliary ducts, gallbladder and pancreatic duct (anterior-posterior view).

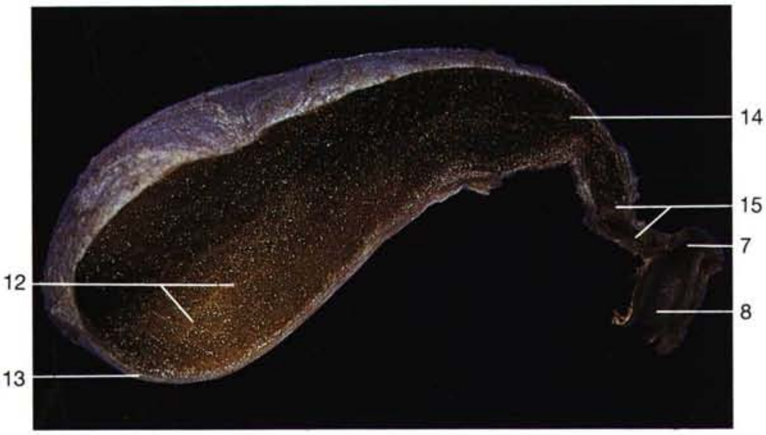

Isolated gallbladder and cystic duct (anterior aspect).
The gallbladder has been opened to display the mucous membrane.

1 Left hepatic duct
2 Right hepatic duct
3 Cystic duct
4 Neck of gallbladder
5 Body of gallbladder
6 Fundus of gallbladder
7 Common hepatic duct
8 **Common bile duct**
9 **Pancreatic duct**
10 **Greater duodenal papilla**
11 Second lumbar vertebra
12 Folds of mucous membrane of gallbladder
13 Muscular coat of gallbladder
14 Neck of gallbladder (opened)
15 Cystic duct with spiral fold
16 **Lesser duodenal papilla**
17 **Accessory pancreatic duct**
18 Uncinate process
19 Plica circularis of duodenum (Kerckrings's fold)
20 Head of pancreas
21 Body of pancreas
22 Tail of pancreas
23 Descending part of duodenum
24 Incisure of pancreas

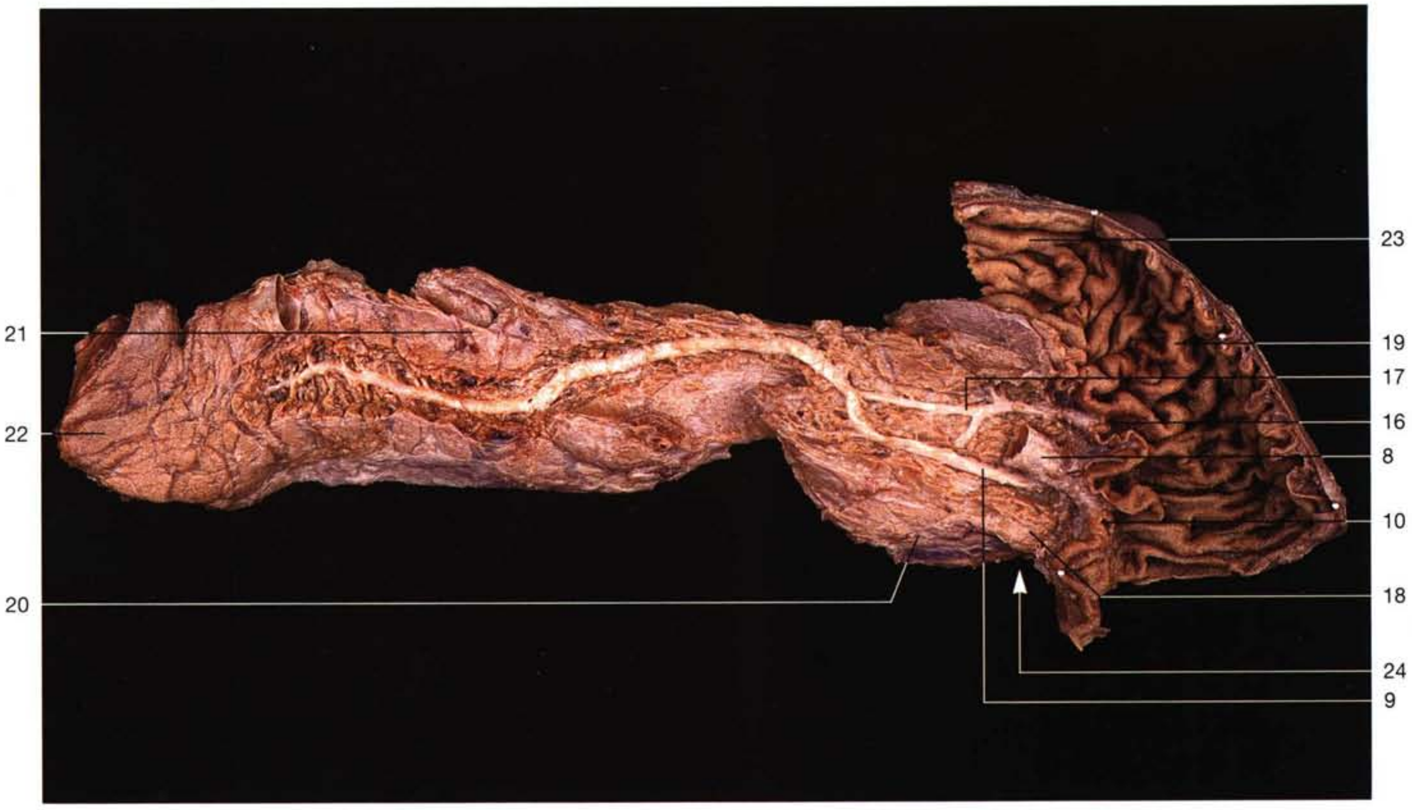

Pancreas with descending part of duodenum (posterior aspect). The duodenum was opened to display the duodenal papillae. Pancreatic duct has been dissected, the common bile duct has been divided. The sphincter of Oddi is shown.

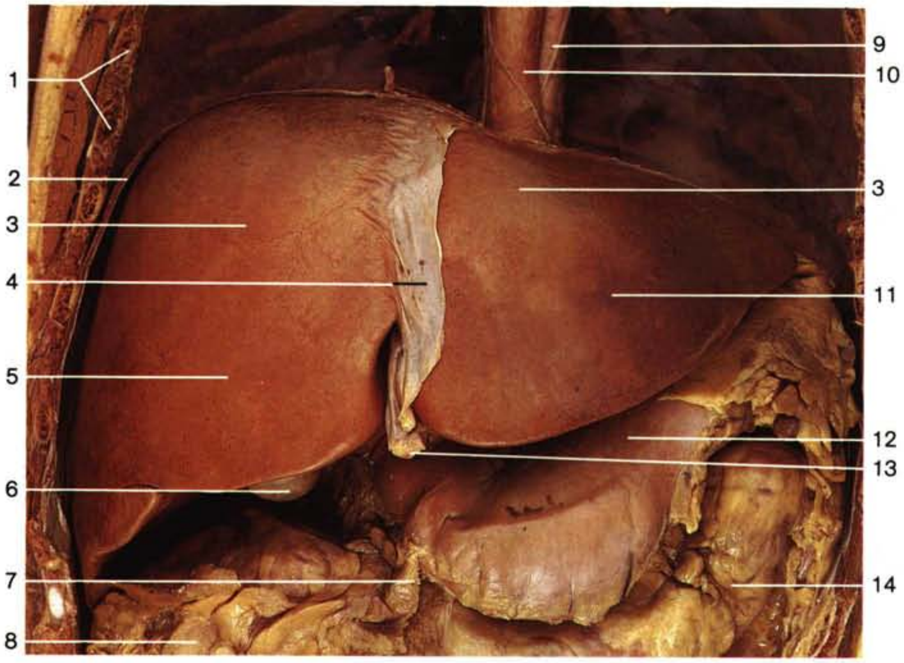

Liver in situ (ventral aspect). Part of the diaphragm has been removed.

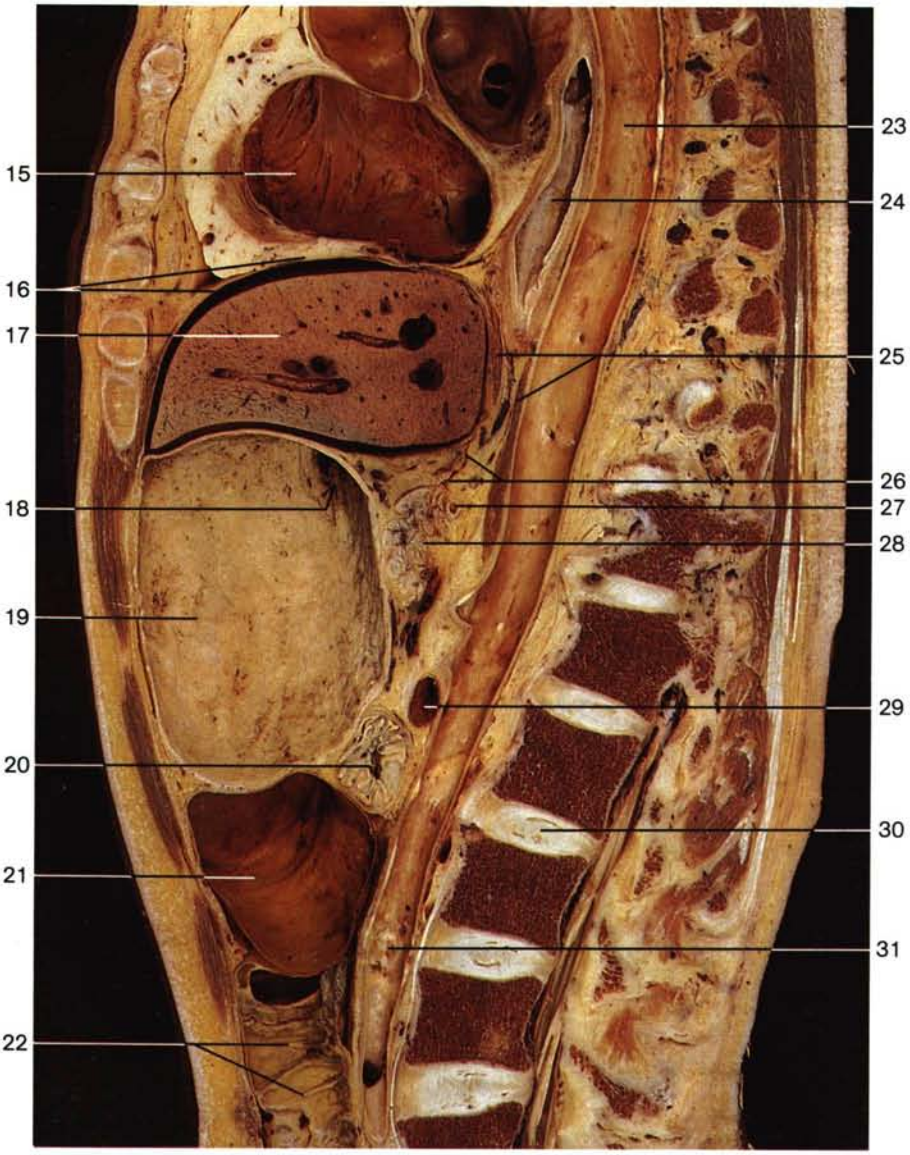

Liver in situ. Parasagittal section through the left side of the abdomen 2 cm lateral to median plane.

1 Ribs (cut edges)
2 Diaphragm
3 Diaphragmatic surface of liver
4 **Falciform ligament of liver**
5 Right lobe of liver
6 Fundus of gallbladder
7 Gastrocolic ligament
8 Greater omentum
9 Aorta
10 Esophagus
11 Left lobe of liver
12 Stomach
13 **Ligamentum teres**
14 Transverse colon
15 Right atrium of heart
16 Central tendon and sternal portion of diaphragm
17 Liver (cut edge)
18 Entrance to duodenum (pylorus)
19 Stomach
20 Duodenum
21 Transverse colon (divided, dilated)
22 Small intestine
23 Thoracic aorta (longitudinally divided)
24 Esophagus (longitudinally divided)
25 Esophageal hiatus of diaphragm
26 **Omental bursa** (lesser sac)
27 Splenic artery
28 Pancreas
29 Left renal vein
30 Intervertebral disc
31 Abdominal aorta (longitudinally divided)

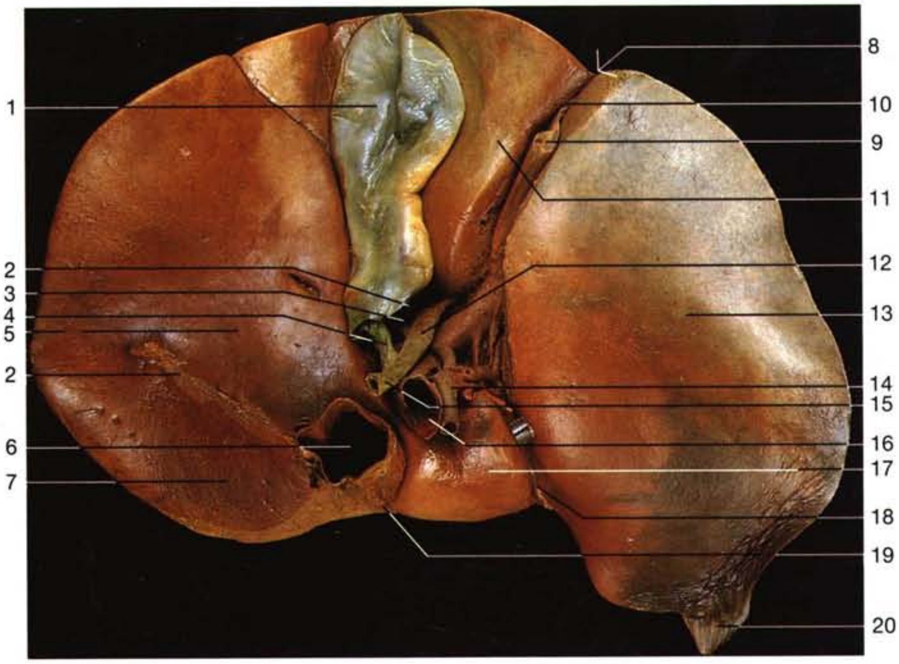

Liver (inferior aspect). Dissection of porta hepatis. Gallbladder partly collapsed. Ventral margin of liver above.

1 Fundus of **gallbladder**
2 Peritoneum (cut edges)
3 Cystic artery
4 **Cystic duct**
5 **Right lobe of liver**
6 Inferior vena cava
7 Bare area of liver
8 Notch for ligamentum teres and falciform ligament
9 Ligamentum teres
10 Falciform ligament of liver
11 **Quadrate lobe of liver**
12 Common hepatic duct
13 **Left lobe of liver**
14 Hepatic artery proper
15 **Common bile duct**
16 **Portal vein**
} Portal triad
17 **Caudate lobe of liver**
18 Ligamentum venosum
19 Ligament of inferior vena cava
20 Appendix fibrosa (left triangular ligament)
21 Coronary ligament of liver
22 Hepatic veins
23 Porta hepatis

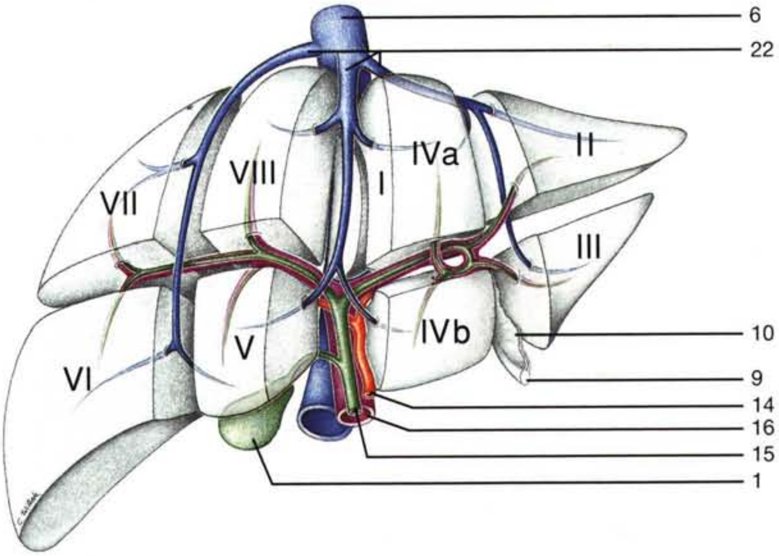

Segmentation of the liver (anterior aspect). Liver segments indicated by roman numerals.

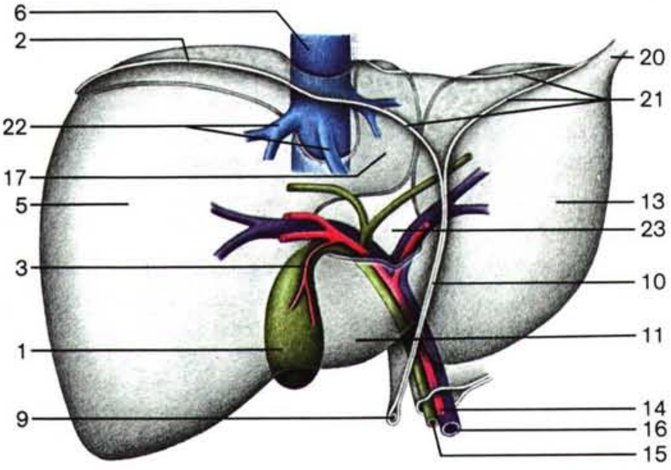

Liver (ventral aspect) (transparent drawing illustrating margins of peritoneal folds).

It should be noted that the anatomical left and right lobes of the liver do not reflect the internal distribution of the hepatic artery, portal vein, and biliary ducts. Using these structures as criteria the left lobe includes both the caudate and quadrate lobes and thus the line dividing the liver into left and right functional lobes passes through the gallbladder and inferior vena cava. The three main hepatic veins drain segments of the liver which have no visible external markings.

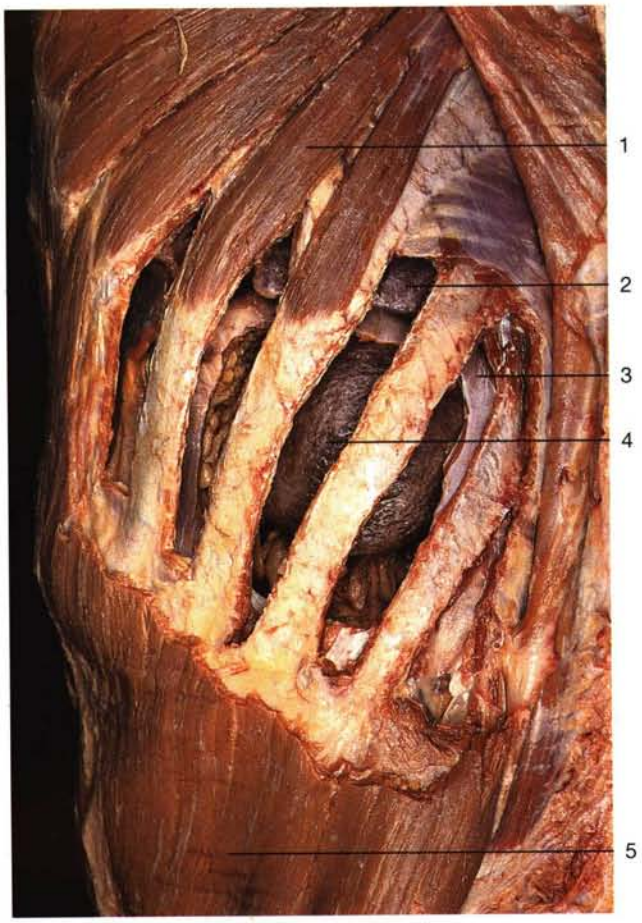

Location of the spleen in situ (left-lateral aspect). Intercostal spaces and diaphragm have been fenestrated.

1 Serratus anterior muscle
2 Left lung
3 Diaphragm
4 Spleen
5 External abdominal oblique muscle
6 Gastrosplenic ligament
7 **Splenic artery**
8 Pancreas tail
9 Superior margin of spleen
10 Anterior border of spleen
11 Inferior vena cava
12 **Hepatic veins**
13 Liver
14 **Portal vein**
15 Paraumbilical veins within falciform ligament
16 **Superior mesenteric vein**
17 Middle colic vein
18 Right colic vein
19 Ascending colon
20 Ileocolic vein
21 Vermiform appendix and appendicular vein
22 Spleen
23 Gastric and esophageal veins
24 **Splenic vein**
25 **Inferior mesenteric vein**
26 Right gastroepiploic vein
27 Ileal veins
28 Descending colon
29 Ileum
30 Sigmoid veins
31 Superior rectal veins

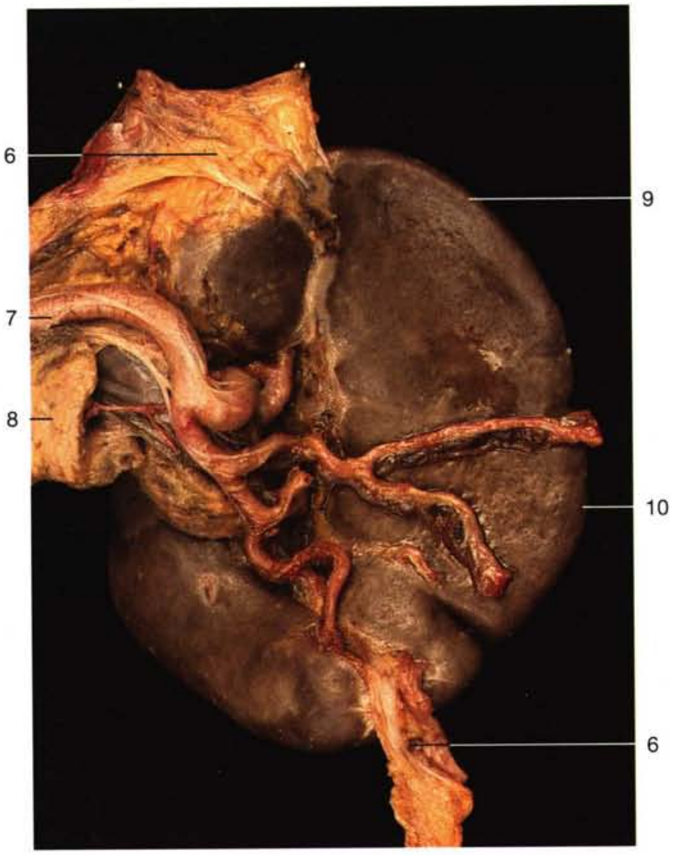

Spleen (visceral surface), hilum of spleen with vessels, nerves and ligaments.

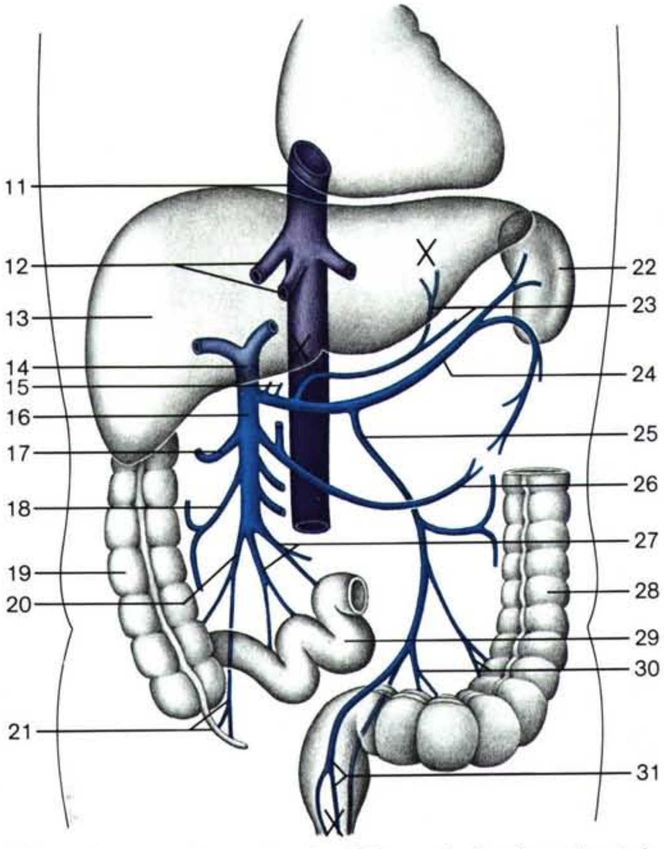

Main tributaries of portal vein. (Schematic drawing of portal circulation.) X = sites of portocaval anastomoses.

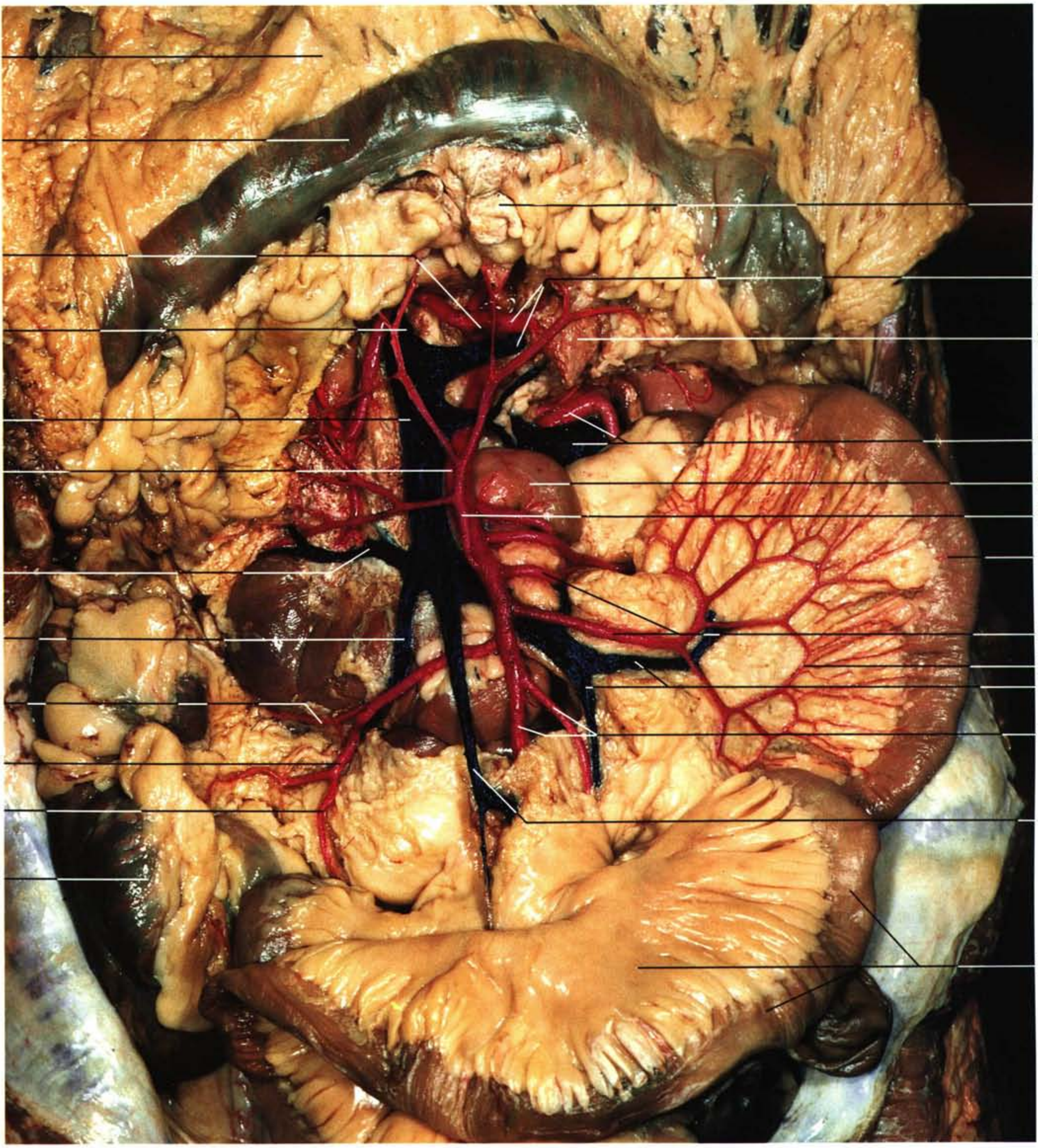

Tributaries of portal vein and branches of superior mesenteric artery (injected with colored solutions). Blue = veins; red = arteries. One layer of peritoneum has been removed to display the vascular arcades of the intestine. Part of the head of the pancreas and the mesocolon have also been removed to show the deeper vessels.

1 Greater omentum (reflected)
2 Transverse colon (raised cranially)
3 Celiac trunk
4 **Portal vein**
5 Superior mesenteric vein
6 **Superior mesenteric artery**
7 Right colic vein
8 Iliocolic vein
9 Right colic artery
10 Iliocolic artery
11 Appendicular artery
12 Cecum
13 Transverse mesocolon
14 **Splenic artery and vein**
15 Pancreas (divided)
16 Renal artery and vein
17 Duodenojejunal flexure
18 Middle colic artery
19 Jejunum
20 Jejunal arteries
21 Arterial arcades to intestine
22 Jejunal veins
23 Ileal arteries
24 Ileal vein
25 Ileum with mesentery

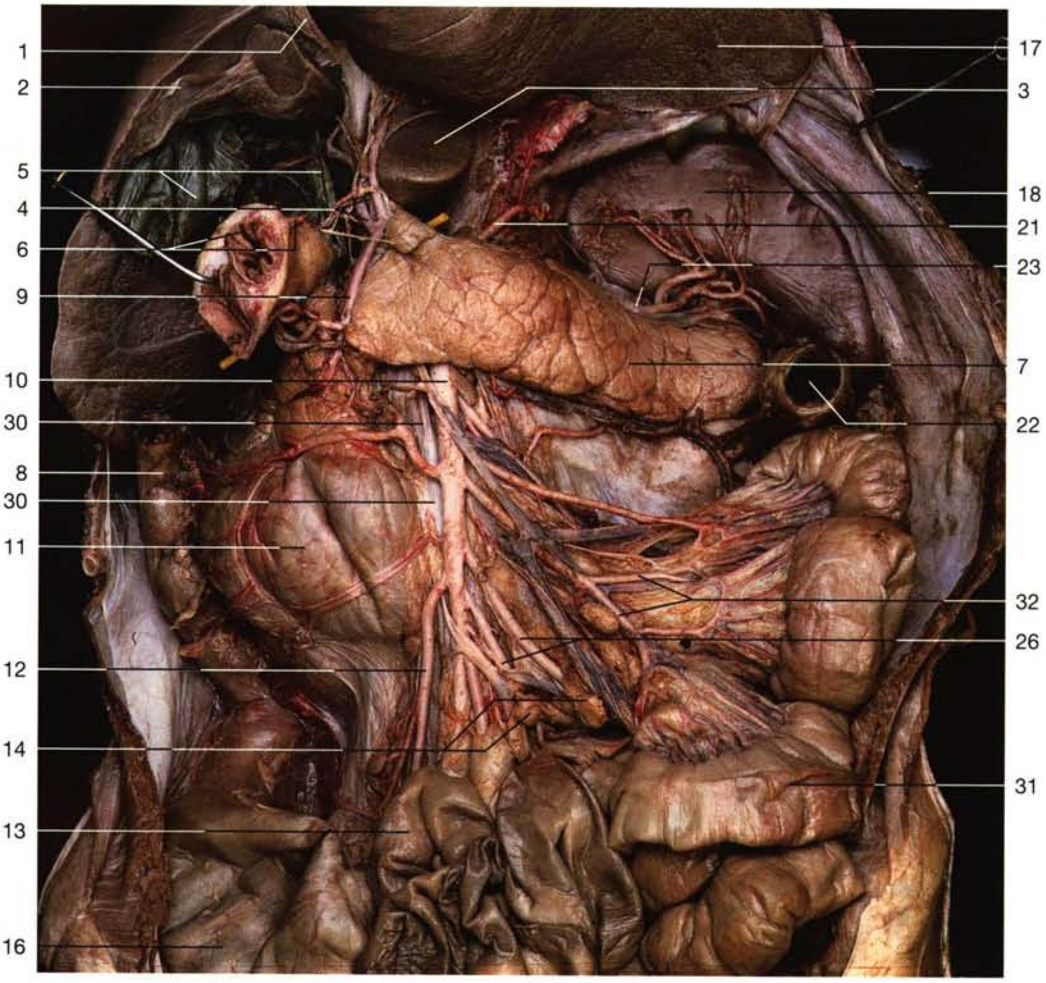

Superior mesenteric artery in relation to pancreas and duodenum. Stomach and transverse colon has been removed and the liver elevated. Note the location of the spleen. A yellow probe is inserted through the omental foramen.

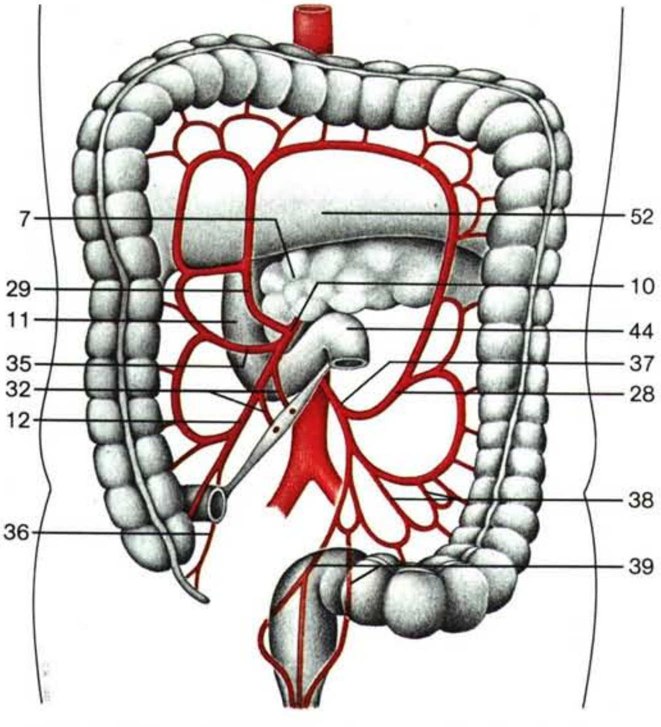

Main branches of superior and inferior mesenteric arteries. (Schematic drawing.)

1 Ligamentum teres
2 Liver
3 Caudate lobe of liver
4 Proper hepatic artery and portal vein
5 Gallbladder and common bile duct
6 Right gastric artery and pylorus
7 **Pancreas**
8 Ascending colon
9 Gastroduodenal artery
10 **Superior mesenteric artery**
11 Duodenum
12 **Iliocolic artery**
13 Terminal part of ileum
14 Mesenteric lymph nodes
15 Right common iliac artery
16 Cecum
17 Left lobe of liver
18 **Spleen**
19 Left gastro-omental (gastroepiploic) artery
20 Stomach
21 Left gastric artery
22 Left colic flexure (cut)
23 **Splenic artery**
24 Right gastro-omental (gastroepiploic) artery
25 Renal artery
26 **Ileal arteries**
27 Left kidney
28 Left colic artery
29 Middle colic artery
30 **Superior mesenteric vein**
31 Jejunum
32 **Jejunal arteries**
33 Inferior vena cava
34 Sigmoid colon
35 Right colic artery
36 Appendicular artery
37 **Inferior mesenteric artery**
38 Sigmoid arteries
39 Superior rectal artery
40 Fundus of gallbladder
41 Common bile duct
42 Portal vein
43 Beginning of jejunum
44 Duodenojejunal flexure
45 Ureter
46 Splenic artery
47 **Celiac trunk**
48 Renal vein
49 Abdominal aorta
50 Descending colon
51 Left common iliac artery
52 Transverse mesocolon

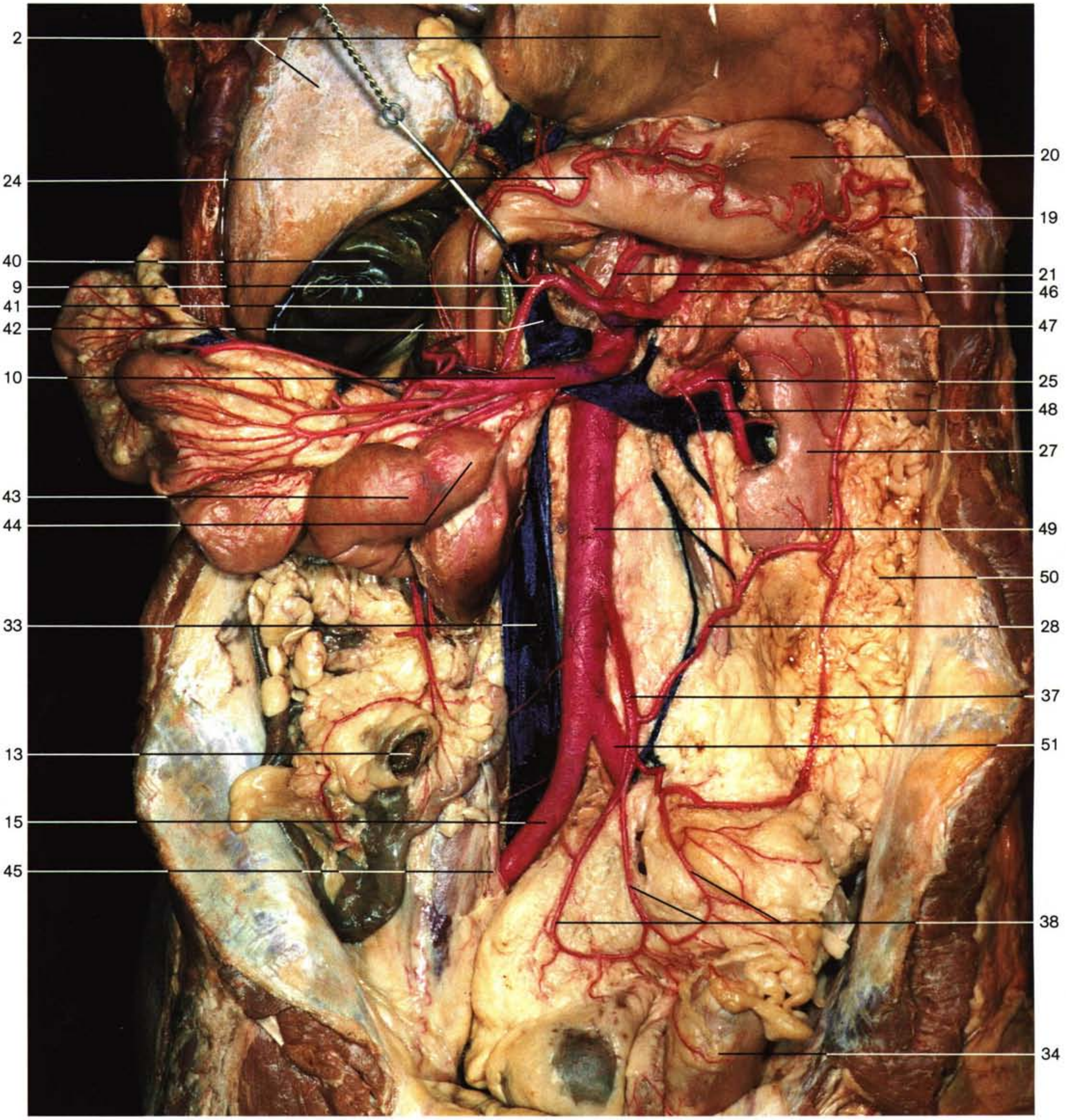

Vessels of abdominal organs; dissection of inferior mesenteric artery and celiac trunk (injected with colored solutions). Blue = veins; red = arteries. The small intestine including the duodenojejunal flexure has been reflected laterally and the stomach and liver have been raised. The peritoneum of the posterior abdominal wall has been removed to display the inferior mesenteric artery and its branches to the colon.

The **superior mesenteric artery** arises at the level of L_1 vertebra branching of the abdominal aorta. It supplies jejunum and ileum as the ascending and transverse colon. At the left colic flexure branches of the middle colic artery (from the superior mesenteric artery) and of the left colic artery (from the inferior mesenteric artery) anastomose (so-called **Riolani's anastomosis**). The arteries unite to form loops or arches called arterial arcades from which straight vessels (vasa recta) arise. The vasa recta do not anastomose within the mesentery. Thus occlusion of these vessels may lead to necrosis of the segment of bowel concerned and ileus. Within the wall of the intestine there are many anastomoses of blood vessels so that blockage of a single vessel is usually not dangerous. The arteries are paralleled by veins which convey blood to the portal vein. Closely related to the venous drainage are the lymph vessels which pass the mesenteric lymph nodes and finally enter the cisterna chyli.

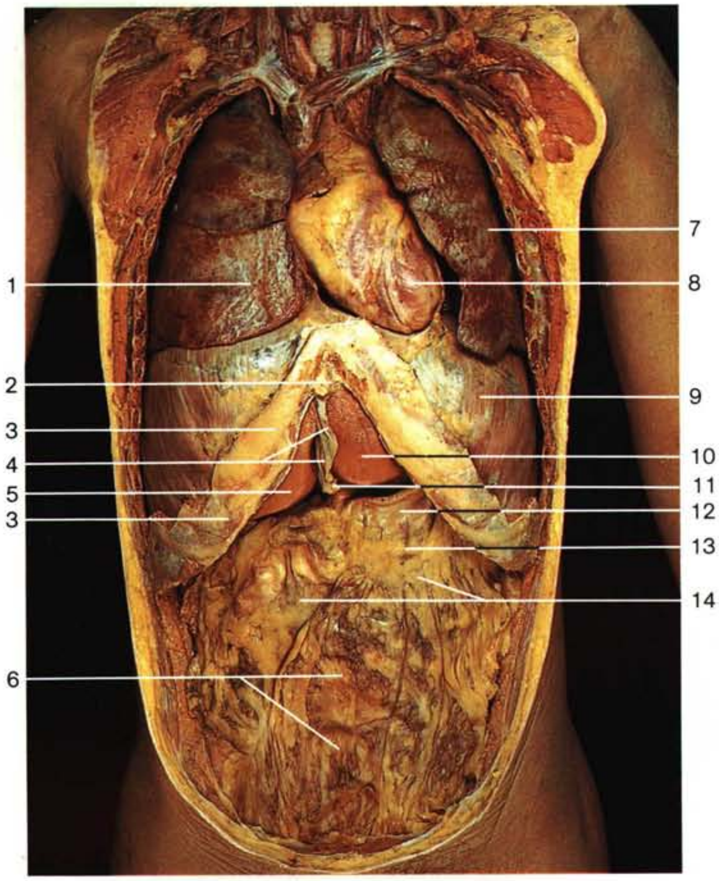

1	Middle lobe of right lung
2	Xiphoid process
3	Costal margin
4	Falciform ligament of liver
5	Quadrate lobe of liver
6	Greater omentum
7	Upper lobe of left lung
8	Heart
9	Diaphragm
10	Left lobe of liver
11	Ligamentum teres
12	Stomach
13	Gastrocolic ligament
14	Transverse colon
15	Taenia coli
16	Appendices epiploicae
17	Cecum
18	Taenia coli
19	Ileum
20	Transverse mesocolon
21	Jejunum
22	Sigmoid colon
23	Position of root of mesentery
24	Vermiform appendix
25	Duodenojejunal flexure
26	Mesentery

Abdominal organs. The anterior thoracic and abdominal walls have been removed.

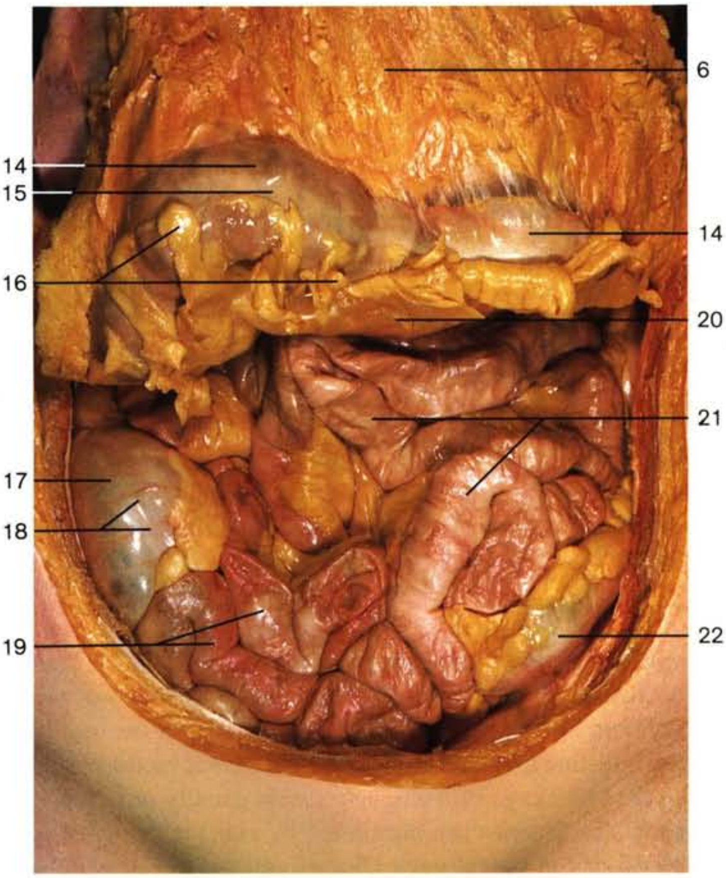

Abdominal organs. The greater omentum which is fixed to the transverse colon has been raised.

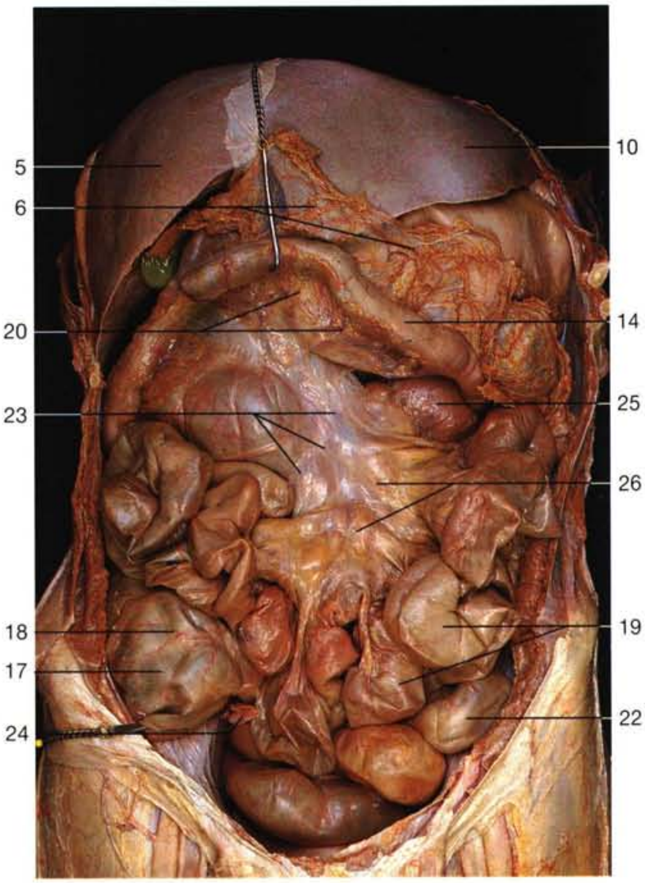

Abdominal organs (anterior aspect). The transverse colon has been reflected.

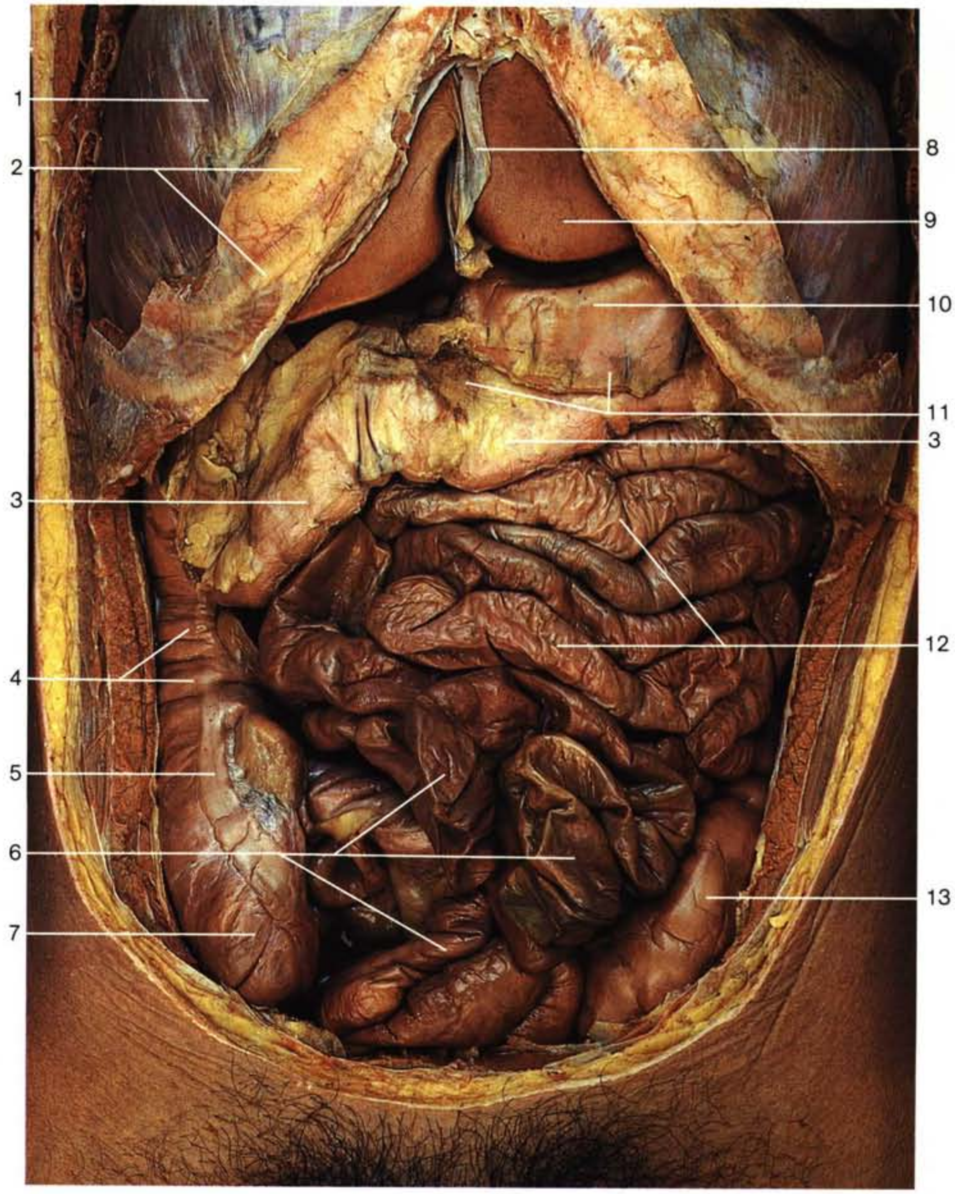

1 Diaphragm
2 Costal margin
3 Transverse colon
4 Ascending colon with haustra
5 Free taenia of cecum
6 **Ileum**
7 **Cecum**
8 **Falciform ligament of liver**
9 Liver
10 Stomach
11 Gastrocolic ligament
12 Jejunum
13 Sigmoid colon
14 **Vermiform appendix**
15 Terminal ileum
16 **Mesoappendix**
17 Mesentery

Abdominal organs in situ. The greater omentum has been removed.

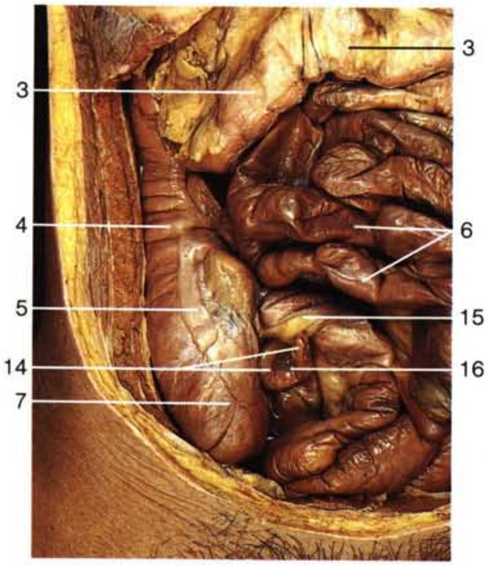

Ascending colon, cecum and vermiform appendix (detail of the preceding figure).

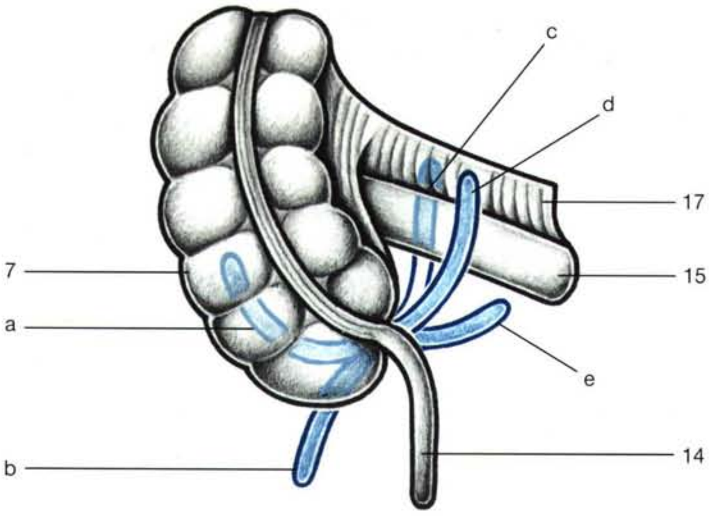

Variations in the position of the vermiform appendix. a = retrocecal; b = paracolic; c = retroileal; d = preileal; e = subcecal.

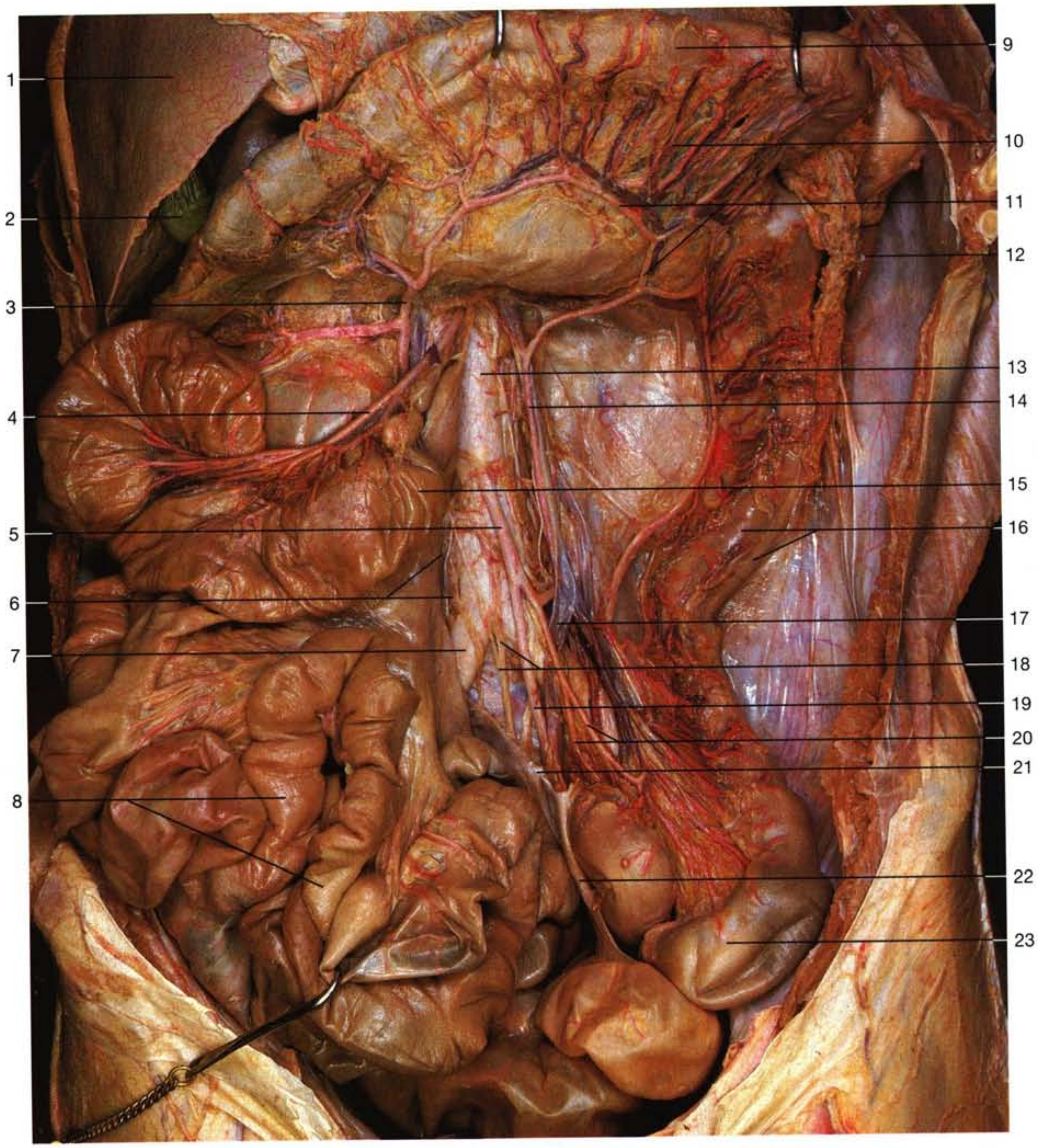

Abdominal organs. Dissection of inferior mesenteric artery and autonomic plexus. The transverse colon with mesocolon has been raised and the small intestine reflected.

1 Liver
2 Gallbladder
3 **Middle colic artery**
4 Jejunal artery
5 **Inferior mesenteric artery**
6 Sympathetic nerves and ganglia
7 Right common iliac artery
8 Small intestine (ileum)
9 Transverse colon (reflected)
10 Transverse mesocolon
11 Anastomosis between middle and left colic artery
12 Spleen
13 **Abdominal aorta**
14 **Left colic artery**
15 Duodenojejunal flexure
16 Descending colon (free taenia of colon)
17 Inferior mesenteric vein
18 Superior hypogastric plexus
19 **Superior rectal artery**
20 Sigmoid arteries
21 Peritoneum (cut edge)
22 Sigmoid mesocolon
23 Sigmoid colon

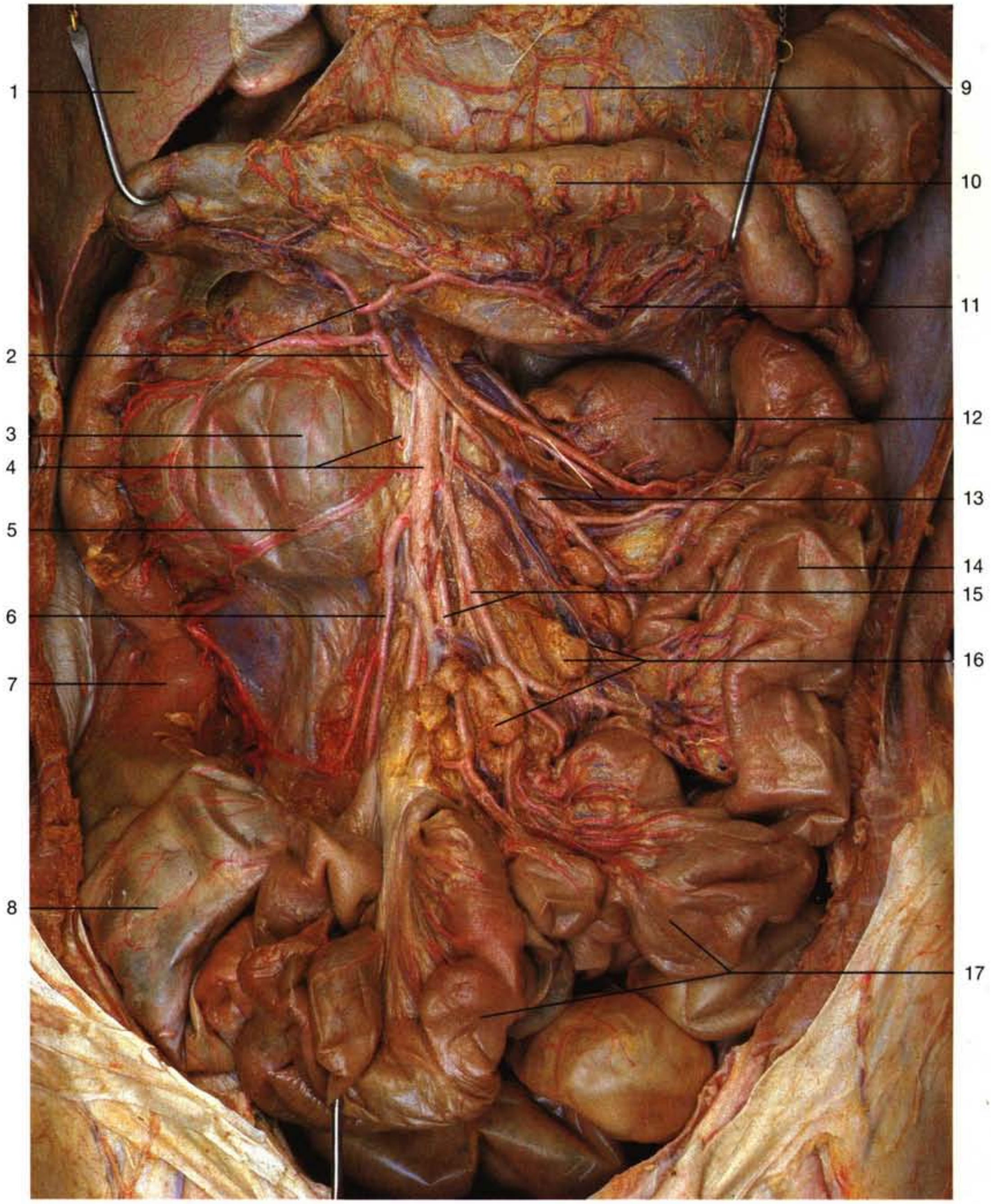

1. Liver
2. Middle colic artery
3. Horizontal part of duodenum (extended)
4. **Superior mesenteric artery and vein**
5. **Right colic artery**
6. **Iliocolic artery**
7. Ascending colon
8. Cecum
9. Greater omentum (reflected)
10. Transverse colon
11. Transverse mesocolon
12. Duodenojejunal flexure
13. **Jejunal arteries**
14. Jejunum
15. **Ileal arteries**
16. **Mesenteric lymph nodes and lymph vessels**
17. Ileum
18. Abdominal aorta
19. Inferior vena cava
20. Stomach
21. Spleen
22. Splenic artery
23. Head of pancreas
24. Superior mesenteric artery

Abdominal organs. Superior mesenteric artery. Mesenteric lymph nodes. Transverse colon reflected.

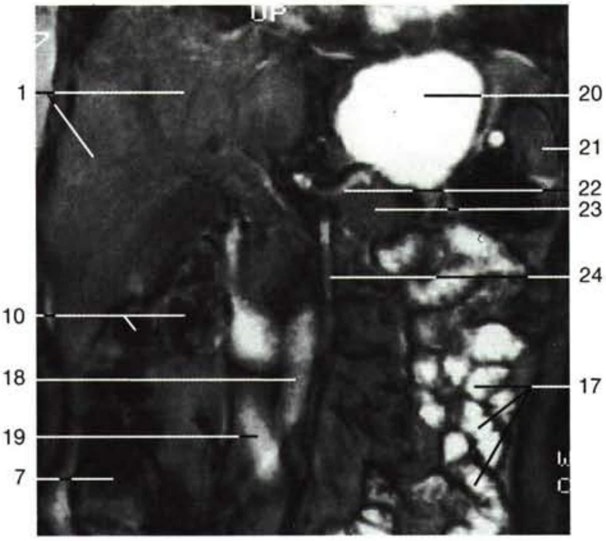

Frontal section through the abdominal cavity. (MR-Scan: the intestinal tract and vessels are filled with a paramagnetic substance [Gadolinium]; courtesy of Dr. W. Rödl, Erlangen.)

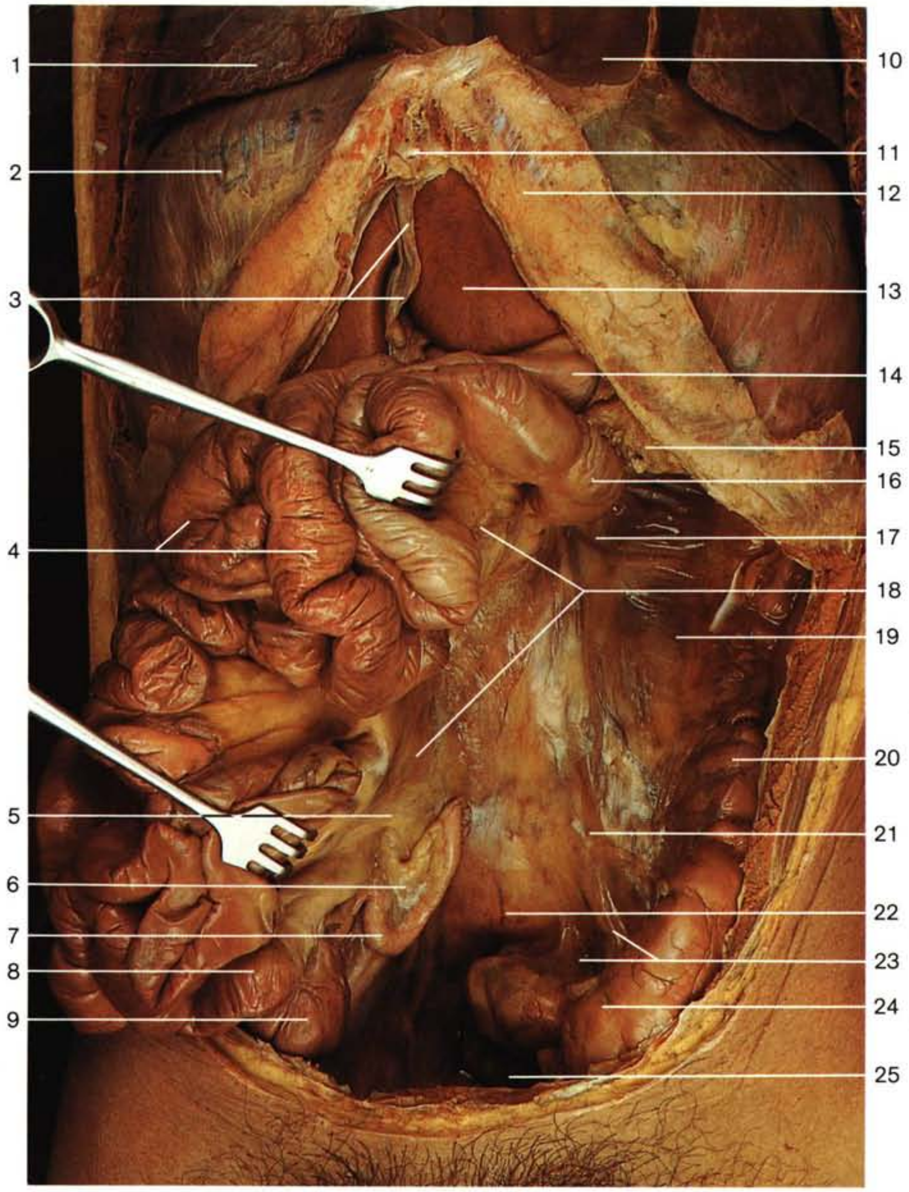

1 Lung
2 Diaphragm
3 Falciform ligament of liver
4 Jejunum
5 Ileocecal fold
6 Mesoappendix
7 Vermiform appendix
8 **Ileocecal junction**
9 **Cecum**
10 Pericardial sac
11 Xiphoid process
12 Costal margin
13 Liver
14 Stomach
15 Transverse colon
16 **Duodenojejunal flexure**
17 Inferior duodenal fold
18 Mesentery
19 Position of left kidney
20 Descending colon
21 Position of left common iliac artery
22 Sacral promontory
23 Sigmoid mesocolon
24 Sigmoid colon
25 Rectum
26 Beginning of jejunum
27 Peritoneum of posterior abdominal wall
28 Transverse mesocolon
29 Superior duodenal fold
30 Superior duodenal recess
31 Retroduodenal recess
32 Free taenia of ascending colon
33 **Ileocecal valve**
34 Frenulum of ileocecal valve
35 Orifice of vermiform appendix (probe)
36 **Ileocolic artery**
37 Vermiform appendix with appendicular artery
38 Ascending colon

Abdominal cavity. Mesenteries. The small intestine has been reflected laterally to demonstrate the mesentery.

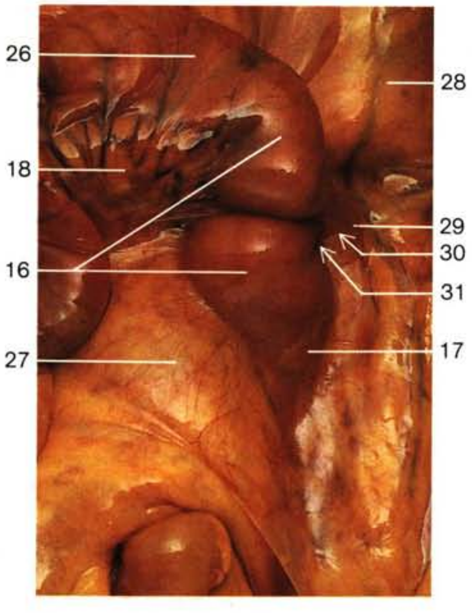

Duodenojejunal flexure
(enlargement of preceding figure).

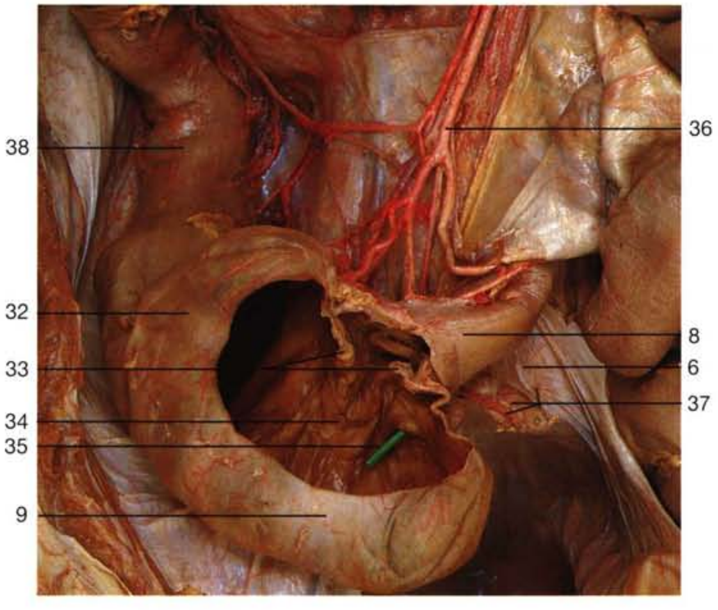

Ileocecal valve (ventral aspect). The cecum and terminal part of the ileum have been opened.

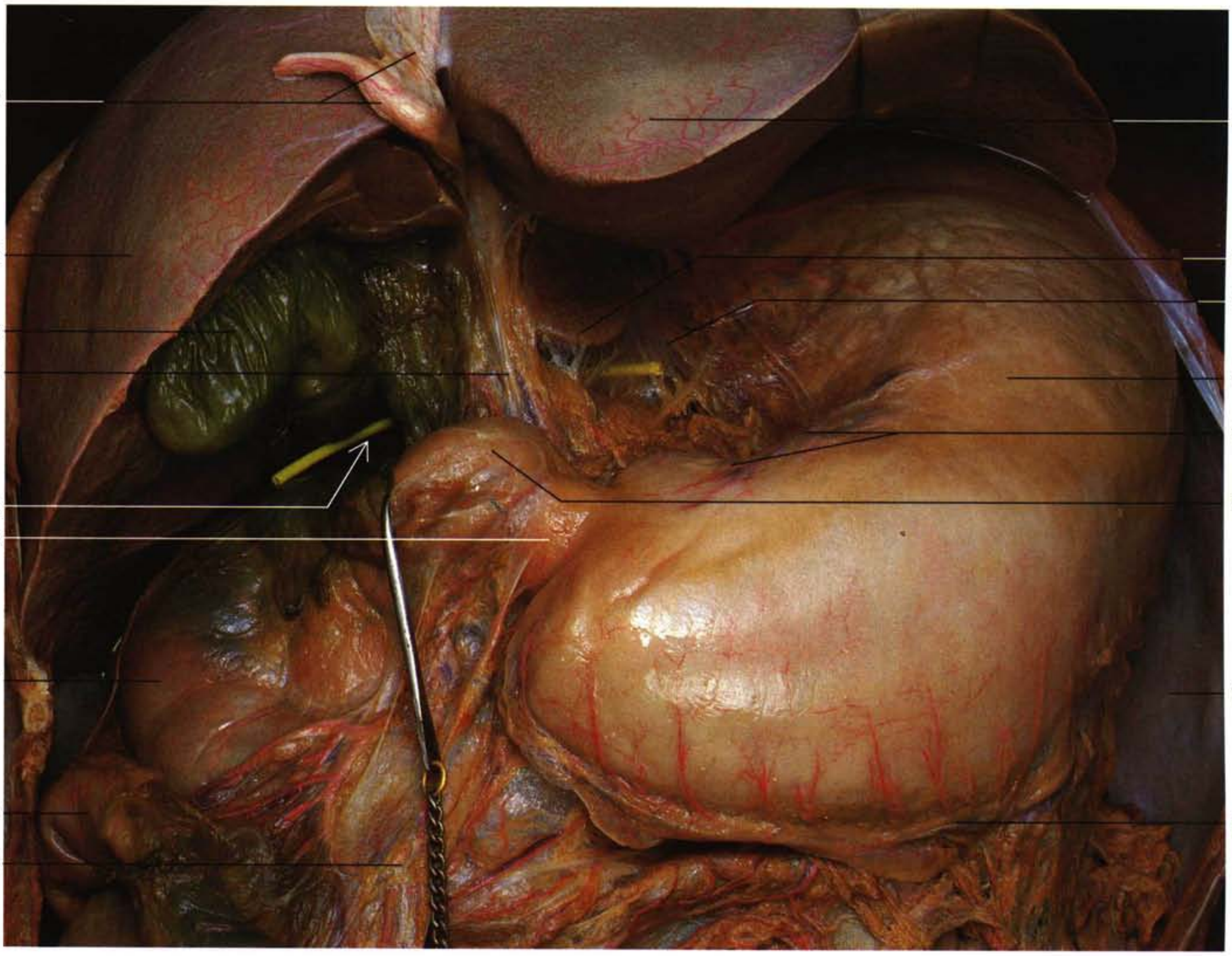

Upper abdominal organs. Thorax and anterior part of diaphragm have been removed and the liver raised to display the lesser omentum. A probe has been inserted into the epiploic foramen and lesser sac.

1 Falciform ligament and ligamentum teres
2 Liver
3 Gallbladder (fundus)
4 Hepatoduodenal ligament
5 **Epiploic foramen** (probe)
6 Pylorus
7 Descending part of duodenum
8 Right colic flexure
9 **Gastrocolic ligament**
10 Caudate lobe of liver (behind lesser omentum)
11 **Lesser omentum**
12 Stomach
13 Lesser curvature of stomach
14 Superior part of duodenum
15 Diaphragm
16 Greater curvature of stomach with gastroepiploic vessels
17 Twelfth thoracic vertebra
18 Right kidney
19 Right suprarenal gland
20 Inferior vena cava
21 Falciform ligament of liver
22 Abdominal aorta
23 Spleen
24 Lienorenal ligament
25 Gastrosplenic ligament
26 Pancreas
27 Lesser sac

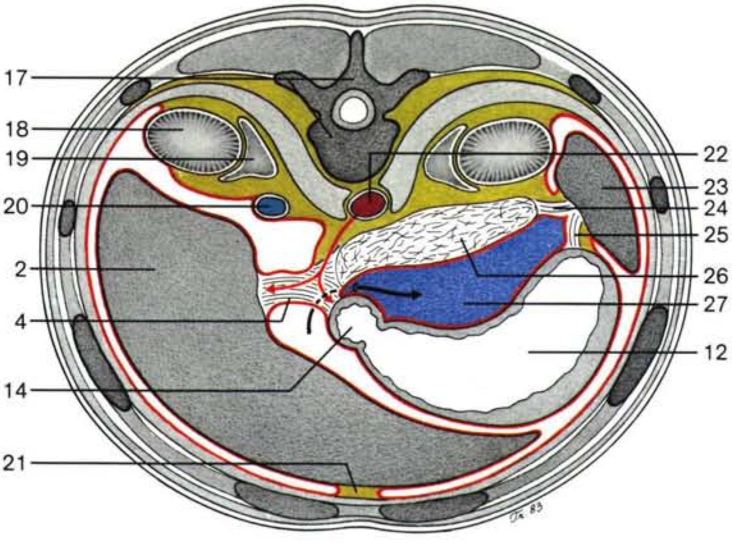

Horizontal section through omental bursa above the level of epiploic foramen (black arrow). Viewed from above. Red arrows: routes of the arterial branches of celiac trunk to liver, stomach, duodenum, and pancreas. (Schematic drawing.)

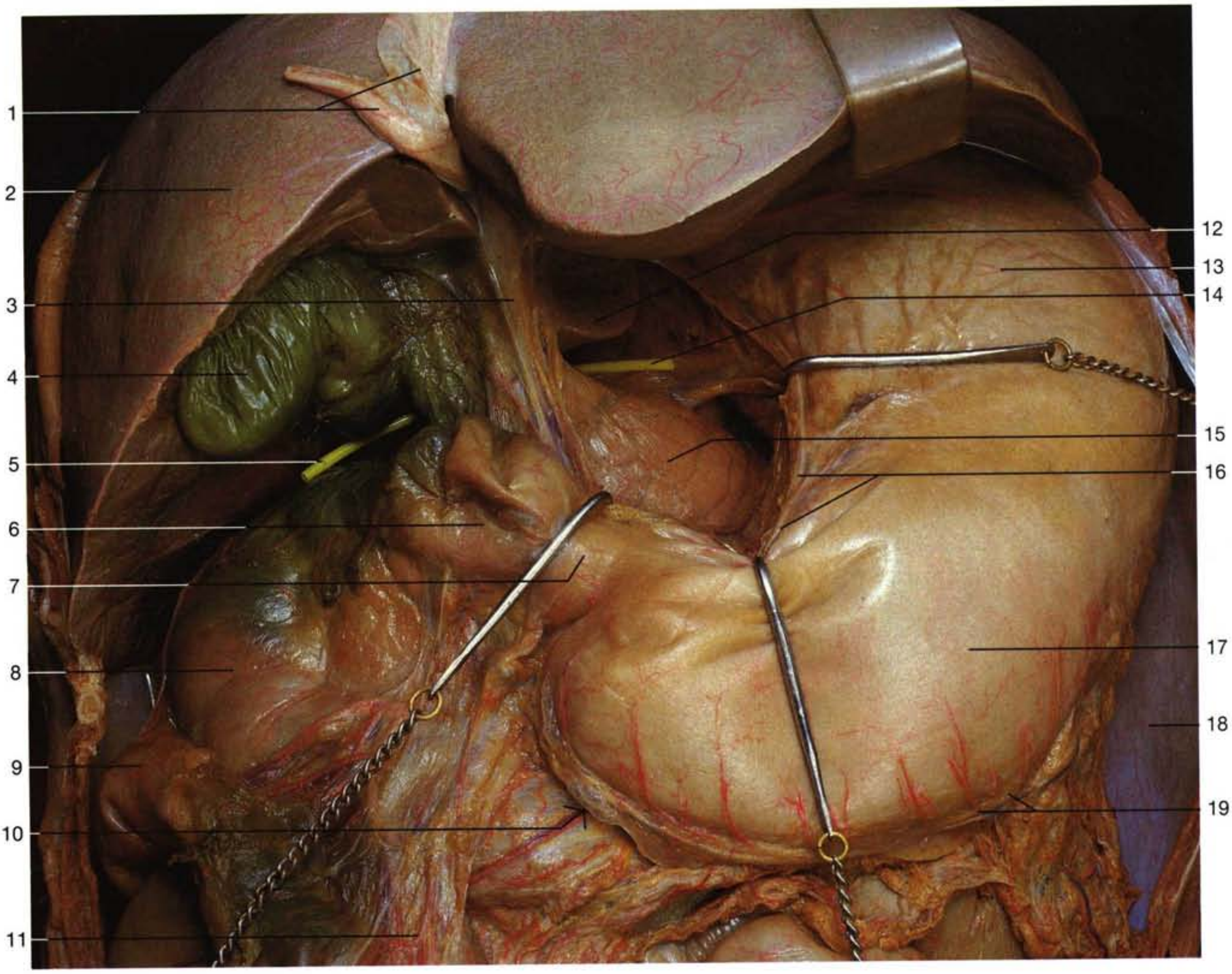

Upper abdominal organs, lesser sac, omental bursa (anterior aspect). Lesser omentum partly removed, liver and stomach slightly reflected.

1 Falciform ligament and ligamentum teres
2 Liver
3 **Hepatoduodenal ligament**
4 Gallbladder
5 Probe within the epiploic foramen
6 Superior part of duodenum
7 Pylorus
8 Descending part of duodenum
9 Right colic flexure
10 Gastrocolic ligament
11 **Greater omentum**
12 Caudate lobe of liver
13 Fundus of stomach
14 Probe at the level of the vestibule of lesser sac
15 **Head of pancreas**
16 Lesser curvature of stomach
17 **Body of stomach**
18 Diaphragm
19 Greater curvature with gastroepiploic vessels
20 Head of pancreas and gastropancreatic fold
21 **Spleen**
22 **Tail of pancreas**
23 Left colic flexure
24 Root of transverse mesocolon
25 Transverse mesocolon
26 Gastrocolic ligament (cut edge)
27 **Transverse colon**
28 Umbilicus
29 Small intestine
30 Lesser omentum
31 Lesser sac (omental bursa)
32 Duodenum
33 Mesentery
34 Sigmoid colon

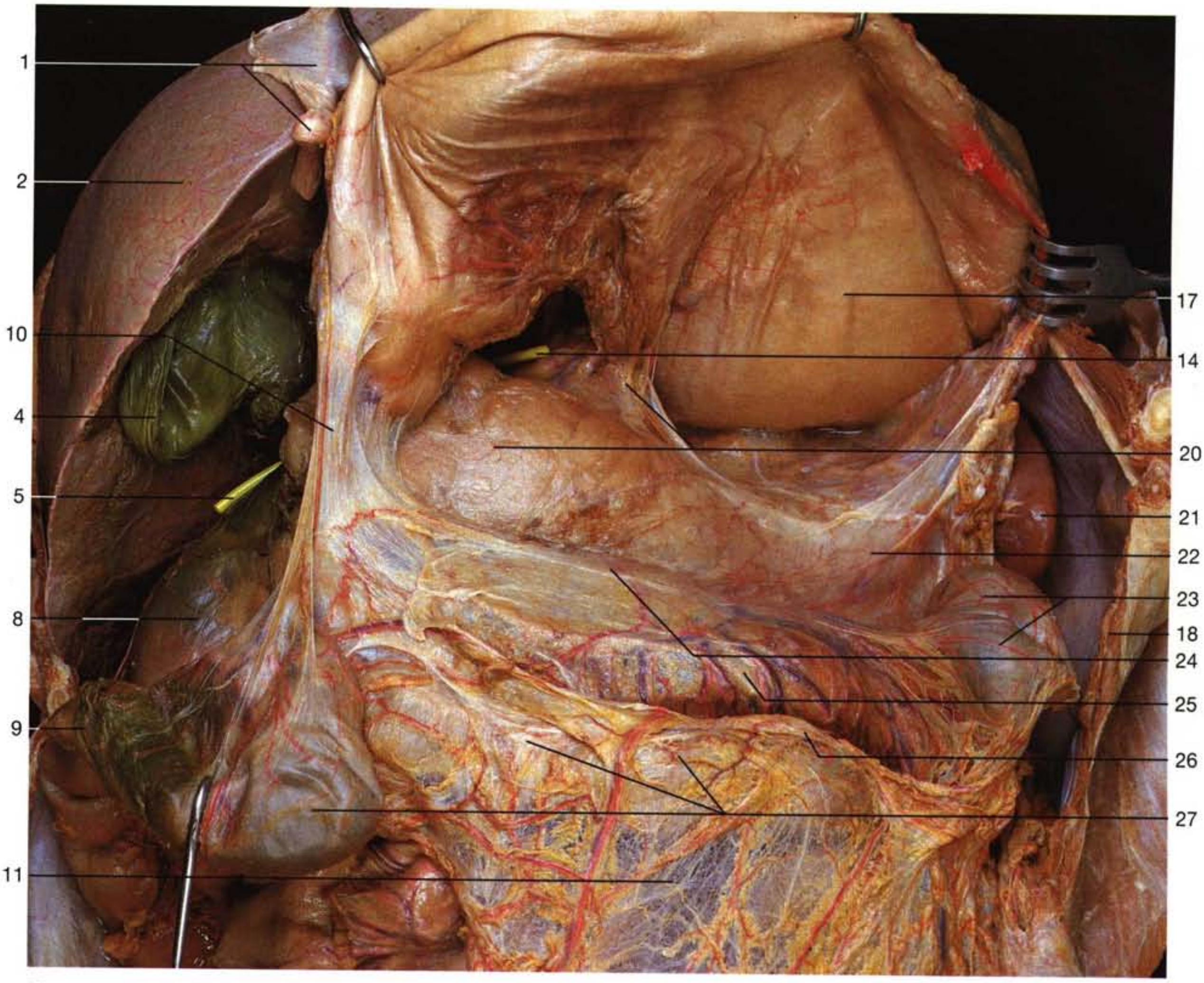
1
2
10
4
5
8
9
11
17
14
20
21
22
23
18
24
25
26
27

Upper abdominal organs, lesser sac, omental bursa (anterior aspect). The gastrocolic ligament has been divided and the whole stomach raised to display the posterior wall of the lesser sac.

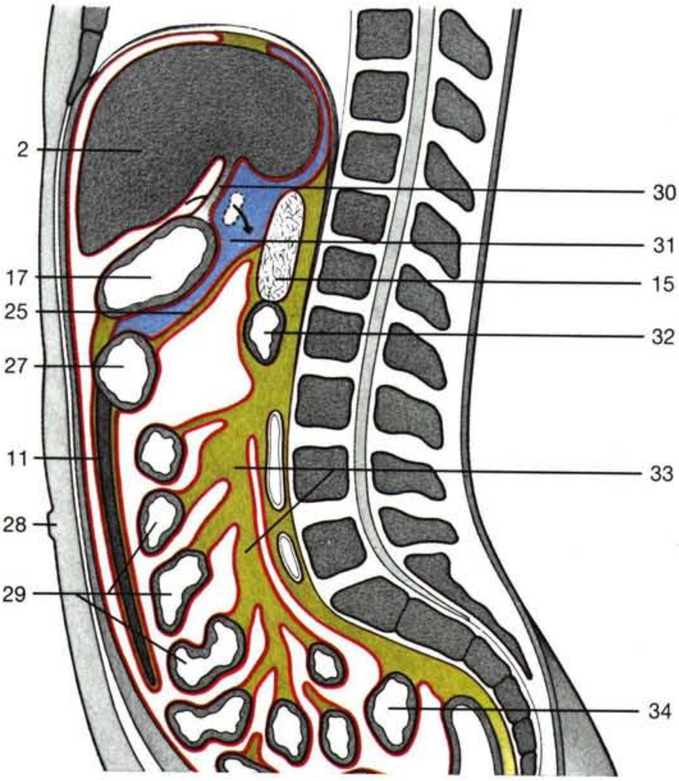
2
17
25
27
11
28
29
30
31
15
32
33
34

Midsagittal section through abdominal cavity demonstrating the site of lesser sac (blue). (Schematic drawing.) The epiploic foramen, entrance to the lesser sac, is indicated by an arrow. Red = peritoneum.

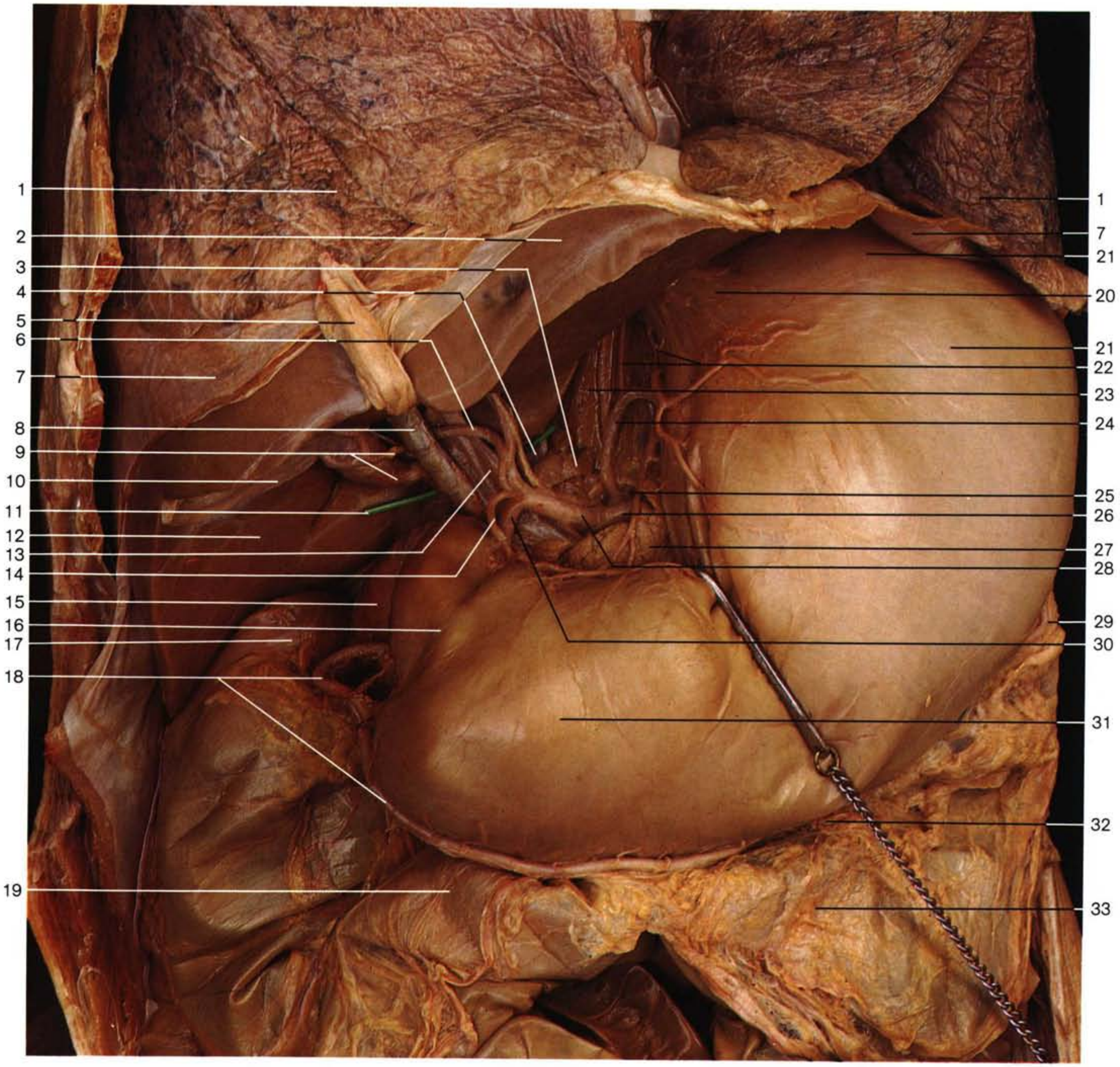
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
1
7
21
20
21
22
23
24
25
26
27
28
29
30
31
32
33

Arteries of upper abdominal organs; dissection of celiac trunk. The lesser omentum has been removed and the lesser curvature of the stomach reflected to display the branches of the celiac trunk.

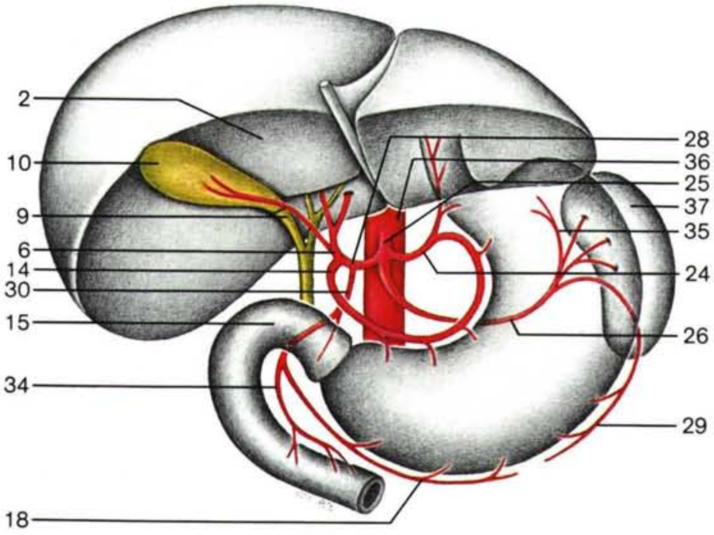
2
10
9
6
14
30
15
34
18
28
36
25
37
35
24
26
29

Branches of celiac trunk.
(Schematic drawing.)

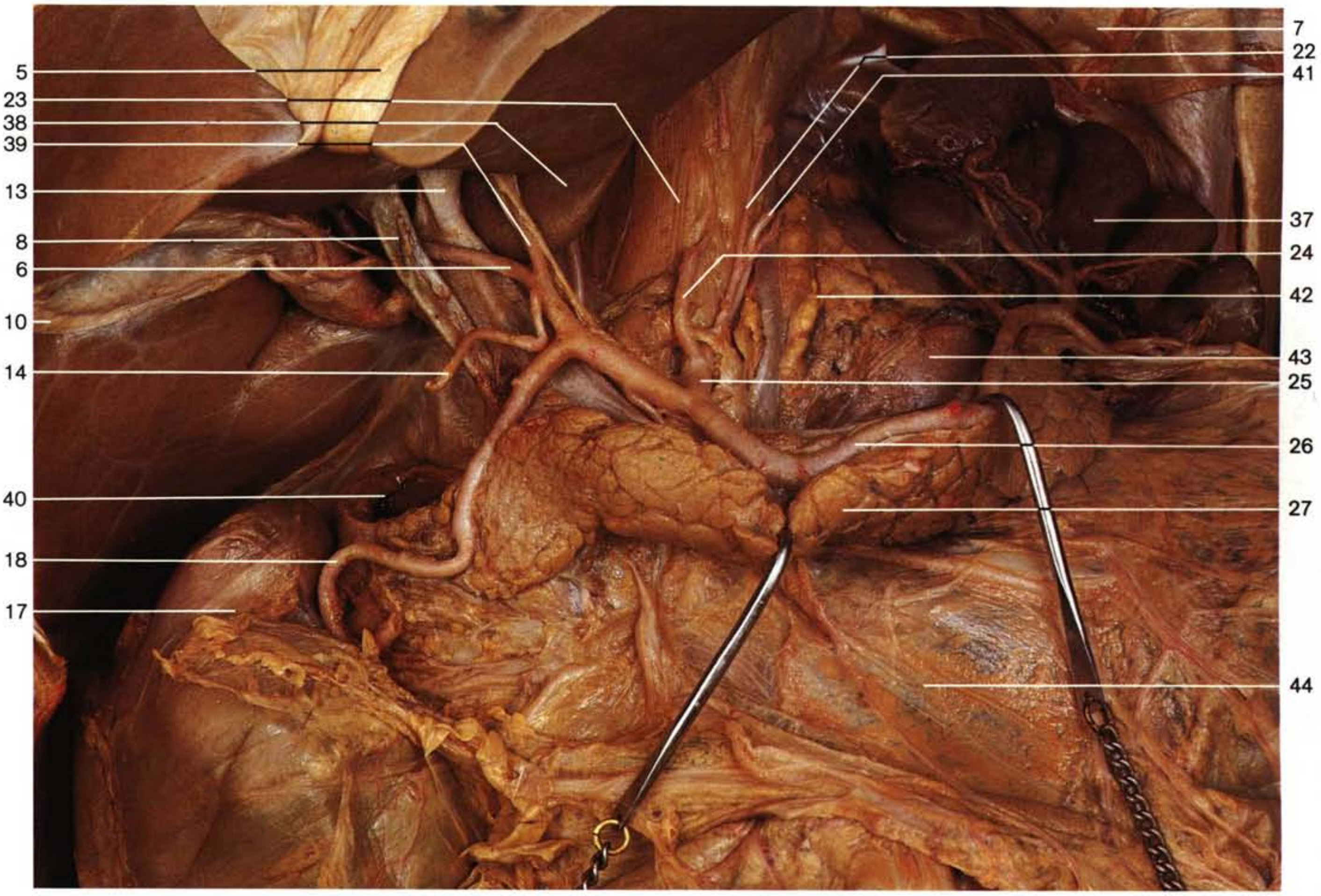

Branches of celiac trunk; blood supply of liver, pancreas and spleen. The stomach, superior part of duodenum and celiac ganglion have been removed to reveal the anterior aspect of the posterior wall of the lesser sac (omental bursa) and the vessels and ducts of the hepatoduodenal ligament. The pancreas has been slightly reflected anteriorly.

1 Lung
2 Liver (visceral surface)
3 Lymph node
4 Inferior vena cava
5 Ligamentum teres (reflected)
6 Right branch of hepatic artery proper
7 Diaphragm
8 **Common hepatic duct** (dilated)
9 Cystic duct and artery
10 Gallbladder
11 Probe in epiploic foramen
12 Right lobe of liver
13 **Portal vein**
14 **Right gastric artery**
15 Duodenum
16 Pylorus
17 Right colic flexure
18 **Right gastro-omental** (gastroepiploic) **artery**
19 Transverse colon
20 Abdominal part of esophagus (cardiac part of stomach)
21 Fundus of stomach
22 Esophageal branches of left gastric artery
23 Lumbar part of diaphragm
24 Left gastric artery
25 **Celiac trunk**
26 **Splenic artery**
27 Pancreas
28 **Common hepatic artery**
29 Left gastro-omental (gastroepiploic) artery
30 **Gastroduodenal artery**
31 Pyloric part of stomach
32 Greater curvature of stomach
33 Gastrocolic ligament
34 Superior pancreaticoduodenal artery
35 Short gastric arteries
36 Aorta
37 Spleen
38 Caudate lobe of liver
39 Left branch of hepatic artery proper
40 Descending part of duodenum (cut)
41 Left inferior phrenic artery
42 Suprarenal gland
43 Kidney
44 Transverse mesocolon

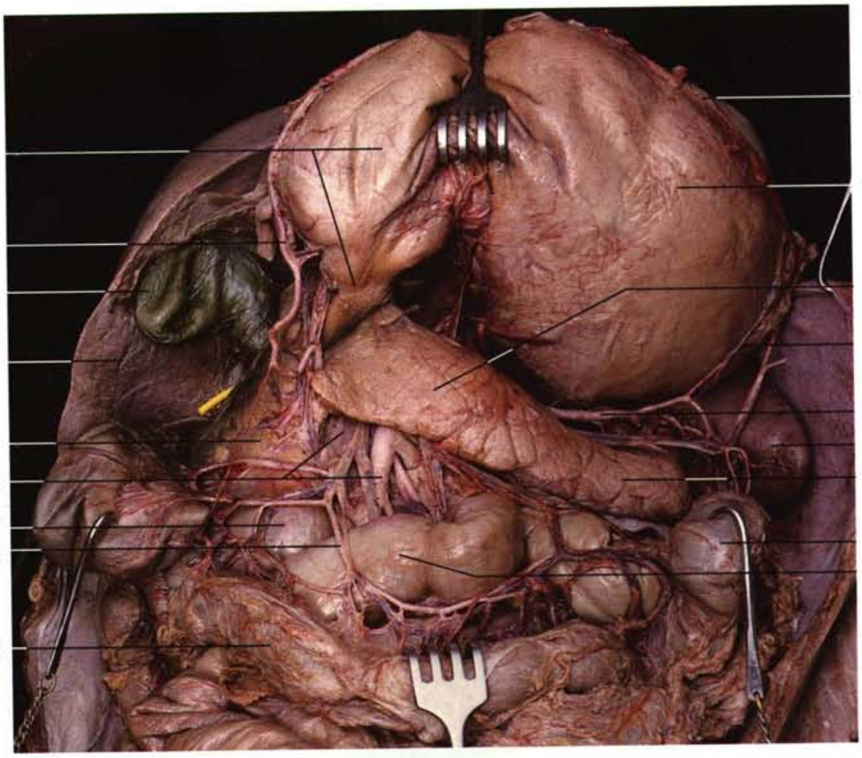

Pancreas and extrahepatic bile ducts in situ (anterior aspect). The gastrocolic ligament has been divided, the transverse colon and the stomach were replaced to display pancreas and superior mesenteric vessels.

1 Stomach (pyloric part) and pylorus
2 Right gastro-omental (gastroepiploic) artery
3 **Fundus of gall bladder**
4 Liver, right lobe
5 Head of pancreas
6 **Superior mesenteric artery** and **vein**
7 Duodenum
8 Middle colic artery
9 Transverse colon
10 Greater curvature of stomach (remnants of gastrocolic ligament)
11 **Body of stomach**
12 **Body of pancreas**
13 Left gastro-omental (gastroepiploic) artery
14 **Splenic artery**
15 **Spleen**
16 Tail of pancreas
17 Left colic flexure
18 Jejunum
19 Lung (inferior lobe of right lung)
20 **Common hepatic duct** and caudate lobe of liver
21 Portal vein and hepatic artery proper
22 Cystic duct and right gastric artery
23 **Common bile duct** and gastroduodenal artery
24 Pylorus (cut)
25 Major and minor duodenal papillas (probes)
26 Duodenum (inferior part)
27 Right kidney, ureter and ovarian artery
28 Iliocolic artery and inferior vena cava
29 Right and left vagal trunk
30 **Esophagus**
31 Diaphragm
32 Cut edge of esophagus, left gastric artery with esophageal branch
33 Celiac trunk
34 Common hepatic artery, main pancreatic duct and **pancreas**
35 Renal vein
36 Duodenojejunal flexure
37 Left kidney
38 Jejunal arteries
39 Abdominal aorta and left ureter
40 Cystic artery
41 Hepatic proper artery
42 **Celiac trunk**
43 Right gastric artery
44 **Common hepatic artery**
45 Gastroduodenal artery
46 **Superior mesenteric artery**
47 Superior posterior pancreaticoduodenal artery
48 Superior anterior pancreaticoduodenal artery
49 Short gastric arteries
50 Left gastric artery
51 Posterior pancreatic branch of splenic artery
52 Inferior pancreaticoduodenal artery

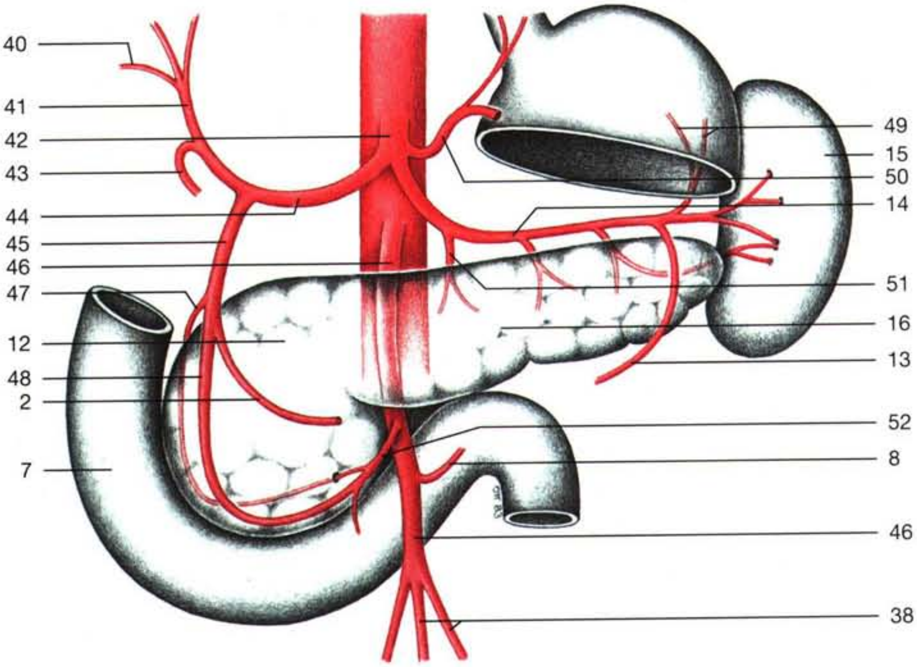

Blood supply of upper abdominal organs (branches of the celiac trunk and superior mesenteric artery). (Schematic drawing.)

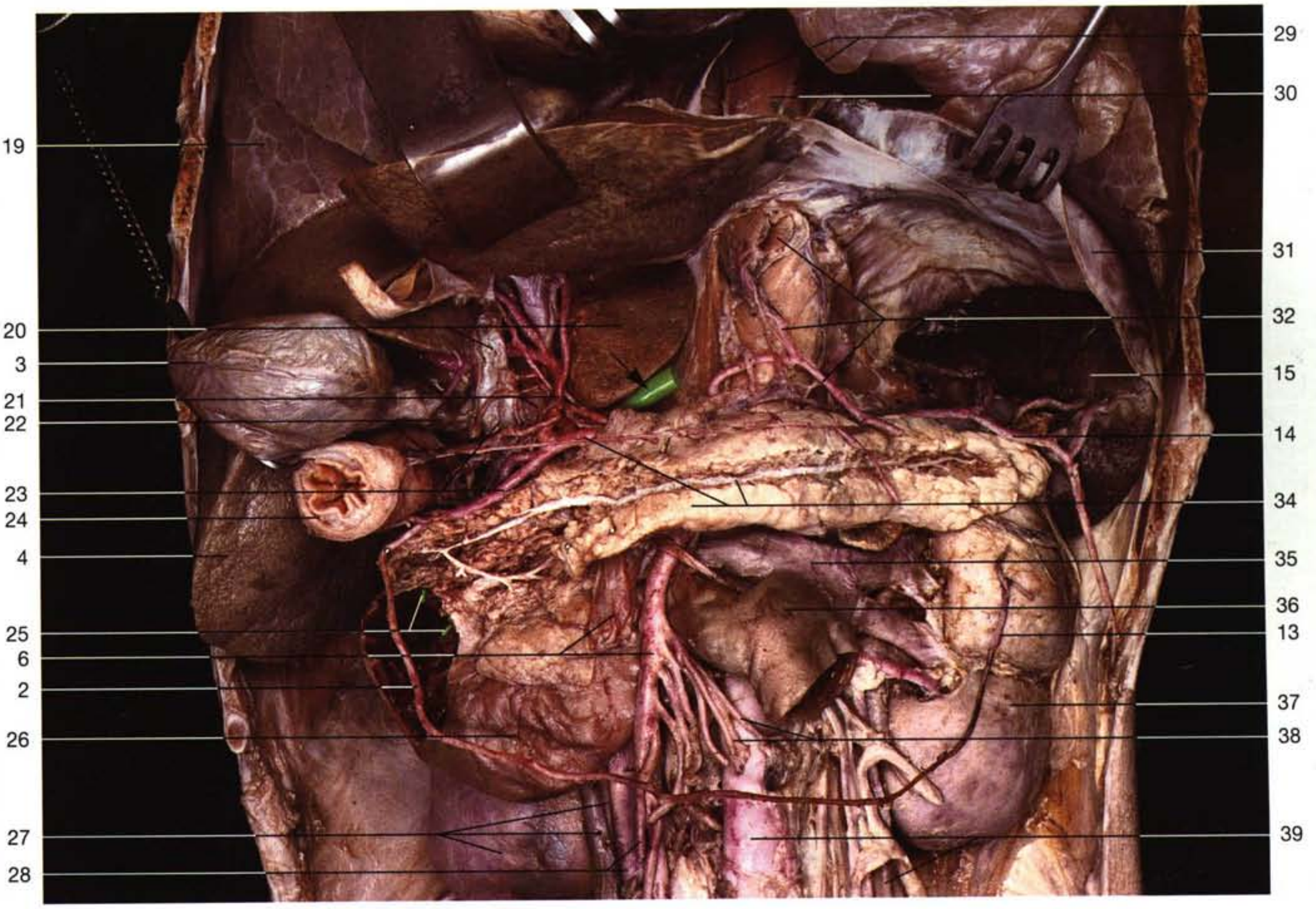
29
30
19
31
32
20
3
15
21
22
14
23
34
24
4
35
36
25
13
6
2
37
26
38
39
27
28

Pancreas, extrahepatic bile ducts, spleen and liver with their vessels in situ. The stomach has been removed, the liver elevated, and the descending part of duodenum opened to display the openings of pancreatic ducts. The pancreatic ducts were dissected. Note the location of superior mesenteric artery and vein between duodenum and pancreas. Arrow: probe within the epiploic foramen.

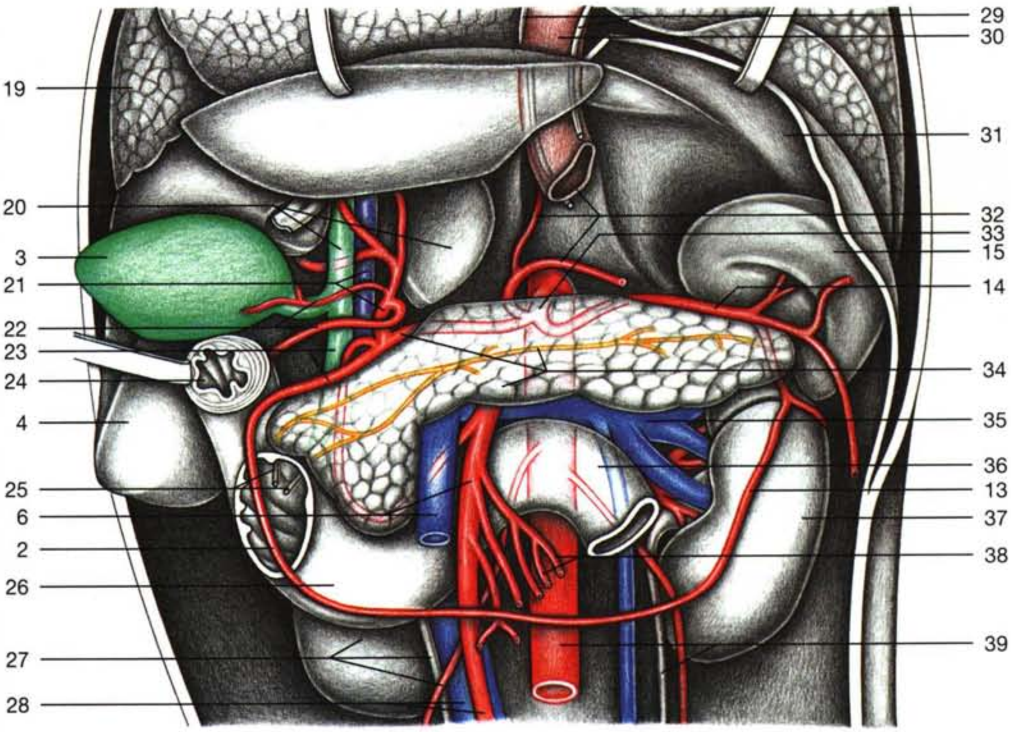
29
30
19
31
20
32
33
15
3
21
14
22
23
34
24
4
35
36
25
13
6
37
2
38
26
27
39
28

Schematic drawing of the specimen depicted above. Note the branching of the celiac trunk (partially covered by the body of the pancreas), and the main arterial supply of liver, spleen, pancreas, stomach, and duodenum.

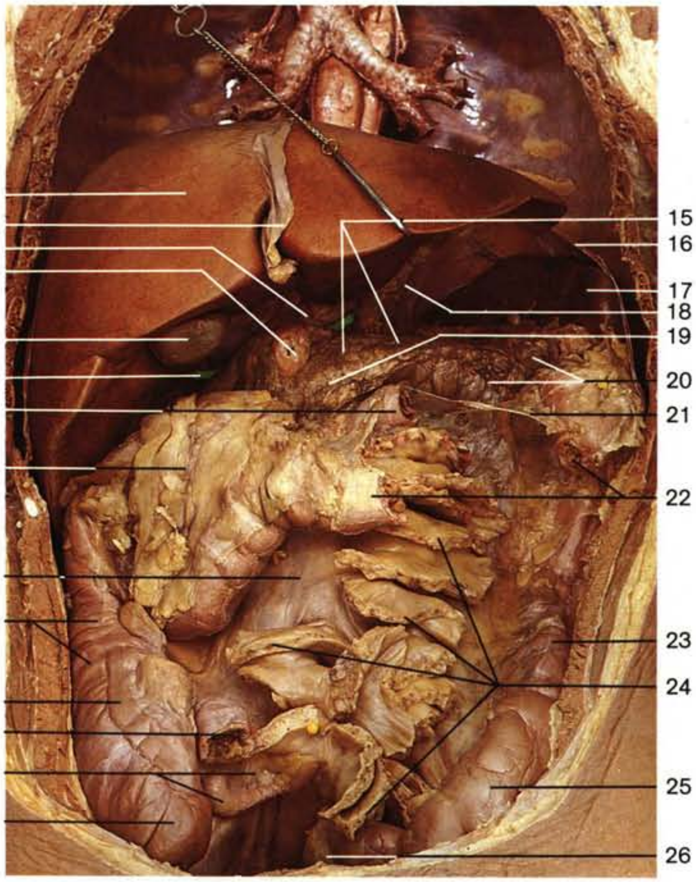

Abdominal cavity after removal of stomach, jejunum, ileum and part of the transverse colon. Liver has been slightly raised.

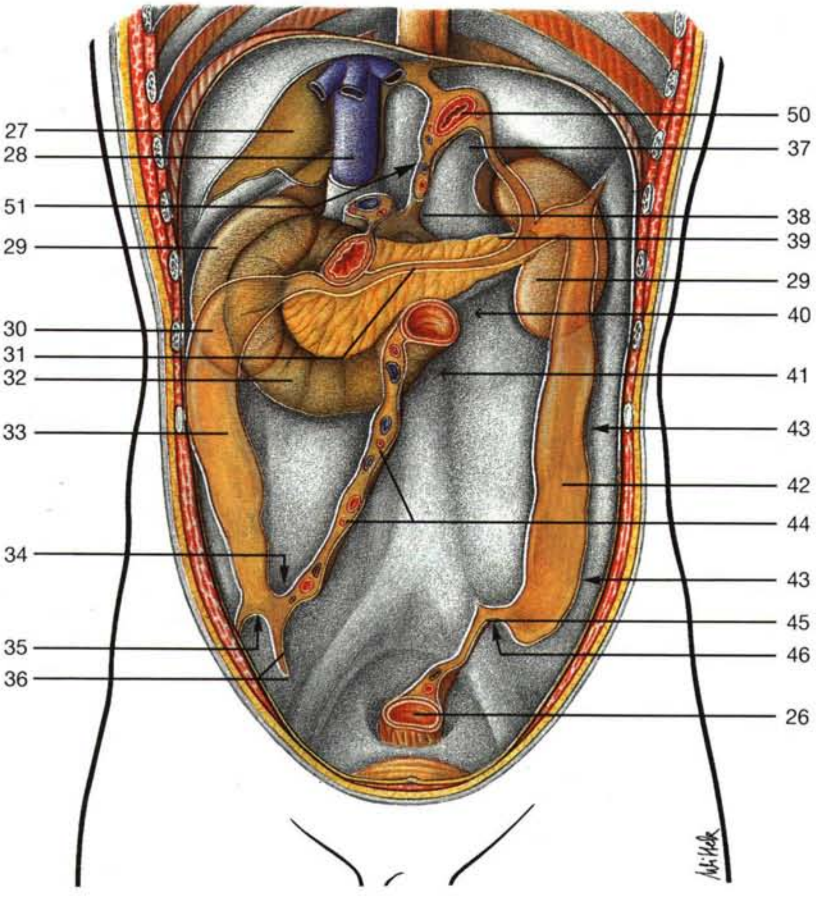

Peritoneal reflections from organs and the position of root of mesentry and peritoneal recesses on the posterior abdominal wall. (Schematic drawing.)

1 Liver
2 Falciform ligament
3 Hepatoduodenal ligament
4 Pylorus (divided)
5 Gallbladder
6 Probe within the epiploic foramen
7 Duodenojejunal flexure (divided)
8 **Greater omentum**
9 Root of mesentery
10 Ascending colon
11 Free colic taenia
12 End of ileum (divided)
13 Vermiform appendix with mesoappendix
14 Cecum
15 Pancreas and site of lesser sac
16 Diaphragm
17 Spleen
18 Cardia (part of stomach, divided)
19 Head of pancreas
20 Body and tail of pancreas
21 **Transverse mesocolon**
22 Transverse colon (divided)
23 Descending colon
24 Cut edge of mesentery
25 Sigmoid colon
26 Rectum
27 Attachment of bare area of liver
28 Inferior vena cava
29 Kidney
30 Attachment of right colic flexure
31 Root of transverse mesocolon
32 Junction between descending and horizontal parts of duodenum
33 Bare surface for ascending colon
34 **Ileocecal recess**
35 **Retrocecal recess**
36 Root of mesoappendix
37 Superior recess } of **lesser sac (omental bursa)**
38 Isthmus (opening) } of **lesser sac (omental bursa)**
39 Splenic recess } of **lesser sac (omental bursa)**
40 **Superior duodenal recess**
41 **Inferior duodenal recess**
42 Bare surface for descending colon
43 **Paracolic recesses**
44 **Root of mesentery**
45 Root of mesosigmoid
46 **Intersigmoid recess**
47 Hepatic veins
48 Duodenojejunal flexure
49 Attachment of left colic flexure
50 Esophagus
51 Entrance to lesser sac through the epiploic foramen

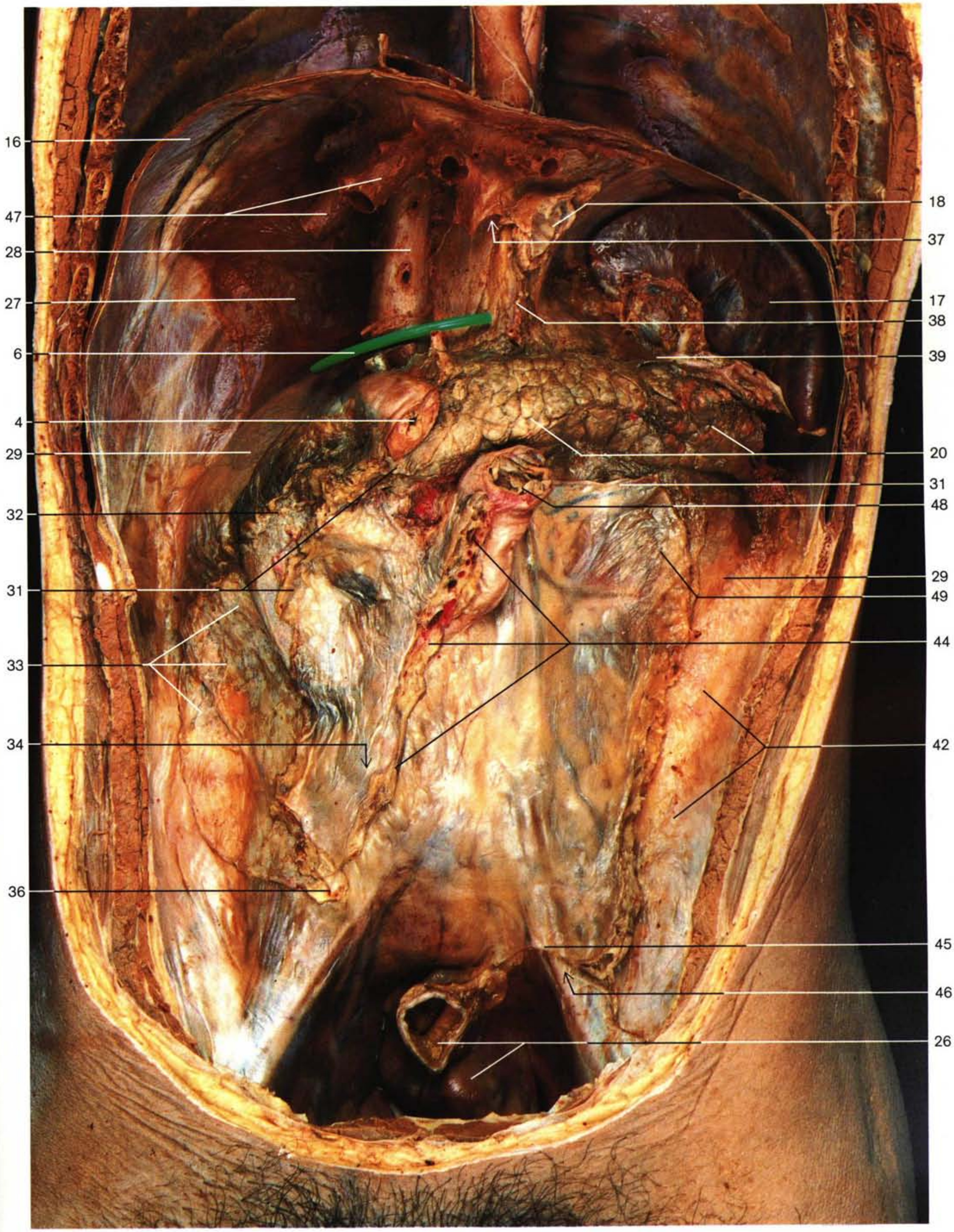
16
47
28
27
6
4
29
32
31
33
34
36
18
37
17
38
39
20
31
48
29
49
44
42
45
46
26

Peritoneal recesses on the posterior abdominal wall. The liver, stomach, jejunum, ileum, and colon have been removed. The duodenum, pancreas, and spleen have been left in place.

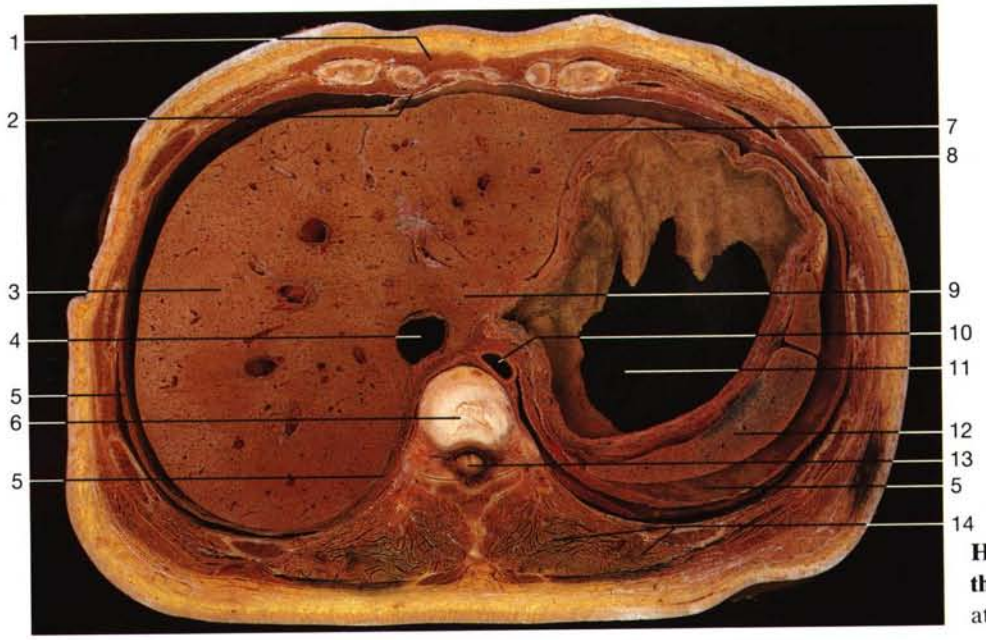

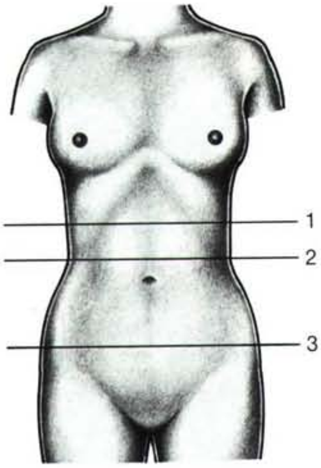

Horizontal section through the abdominal cavity at level 1 (from below).

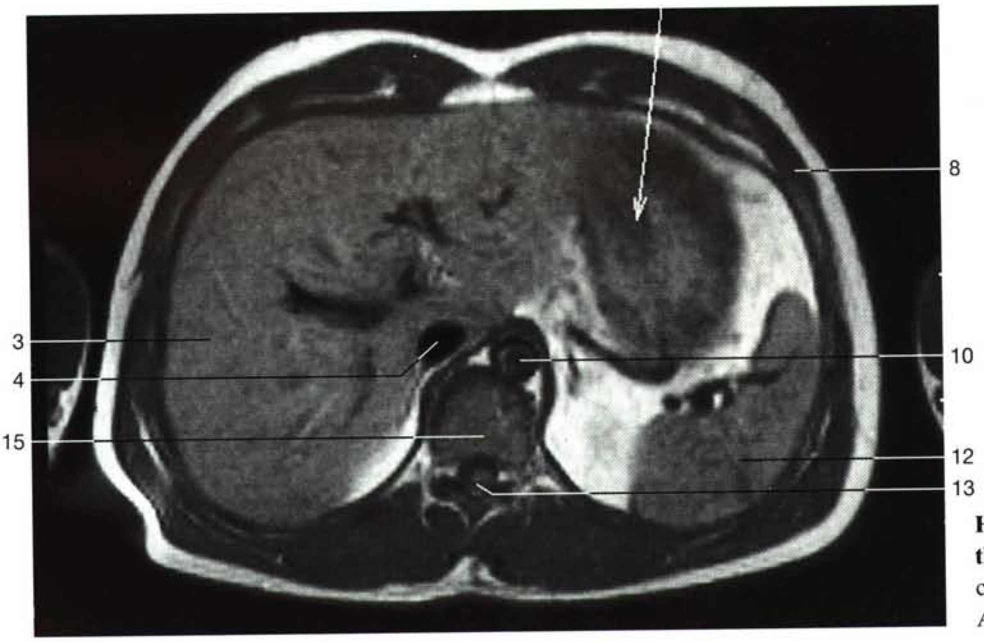

Horizontal section through the body. MR-Scan, corresponding to level 1. Arrow: stomach.

1 Rectus abdominis muscle
2 Falciform ligament
3 **Liver** (right lobe)
4 Inferior vena cava
5 Diaphragm
6 Intervertebral disc
7 Liver (left lobe)
8 Rib
9 Liver (caudate lobe)
10 Abdominal (descending) aorta
11 **Stomach**
12 **Spleen**
13 Spinal cord
14 Longissimus and iliocostalis muscles
15 Body of vertebra
16 Abdominal rectus muscle
17 External abdominal oblique muscle
18 Transverse colon
19 **Head of pancreas**
20 **Greater duodenal papilla**
21 Duodenum
22 Suprarenal gland and ureter
23 **Kidney**
24 Round ligament of liver
25 Superior mesenteric artery and vein
26 Psoas major muscle
27 Descending colon
28 Quadratus lumborum muscle
29 Cauda equina
30 Right renal vein
31 **Small intestine**
32 Iliacus muscle
33 Ilium
34 **Ileocecal valve**
35 Cecum
36 Common iliac artery and vein
37 Gluteus medius muscle
38 Vertebral canal and dura mater

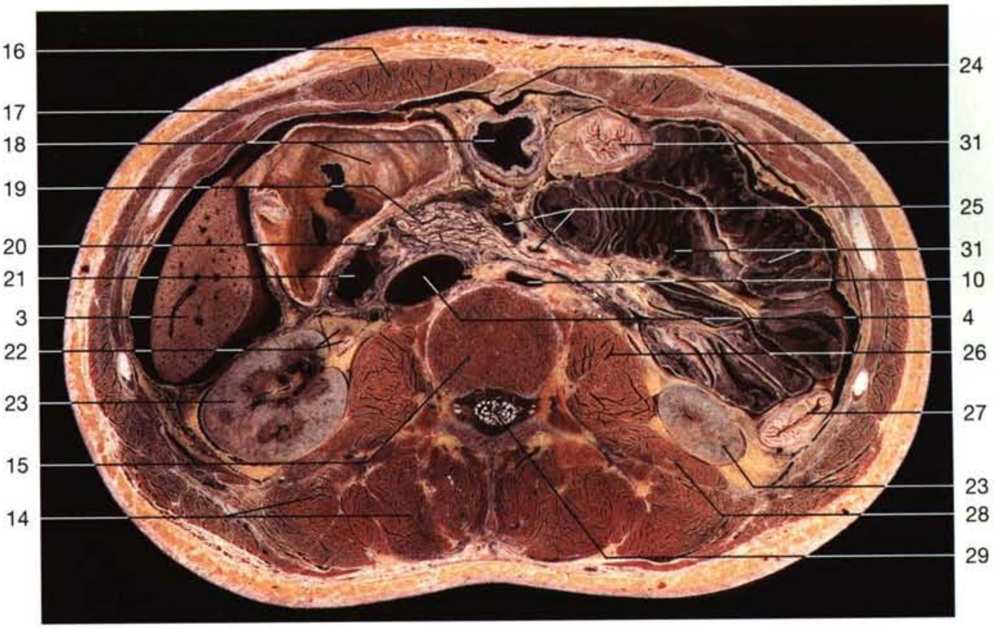
16
17
18
19
20
21
3
22
23
15
14
24
31
25
31
10
4
26
27
23
28
29

Horizontal section through the abdominal cavity at the level of greater duodenal papilla (from below).

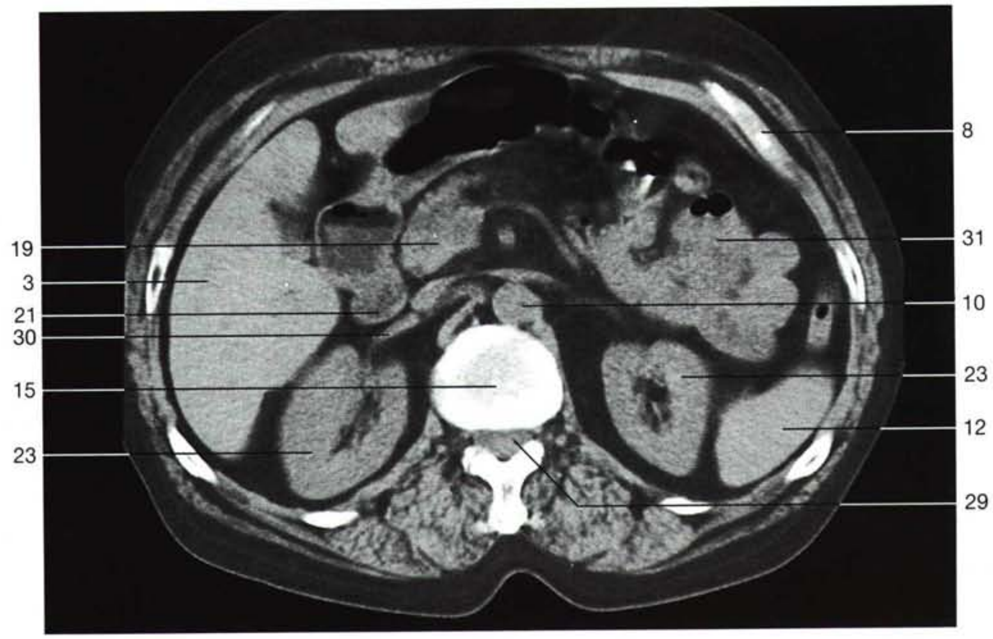
19
3
21
30
15
23
8
31
10
23
12
29

Horizontal section through the body. CT-Scan, corresponding to level 2.

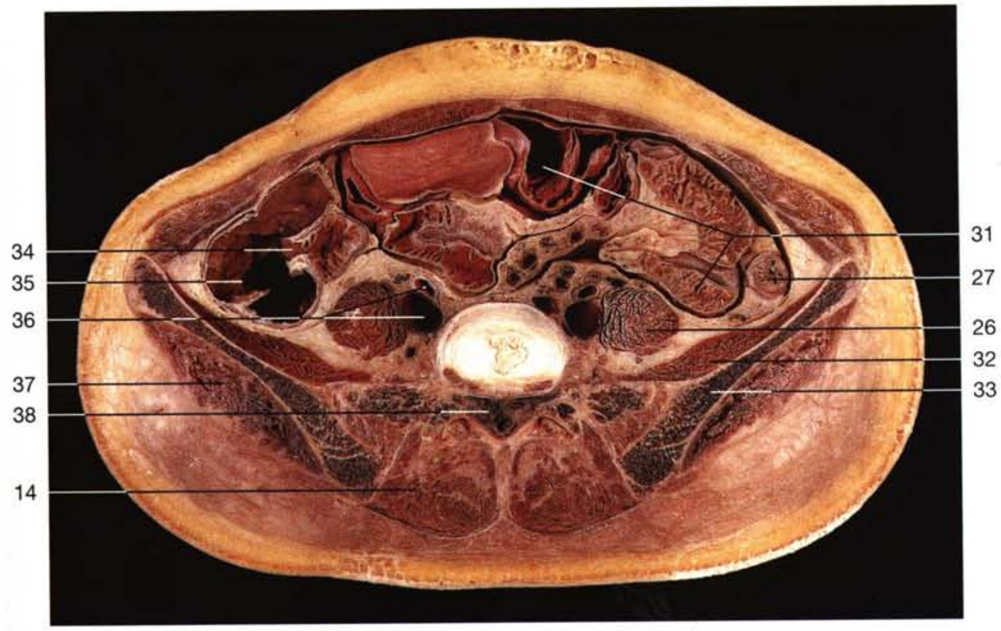
34
35
36
37
38
14
31
27
26
32
33

Horizontal section through the abdominal cavity at level 3 (from below).

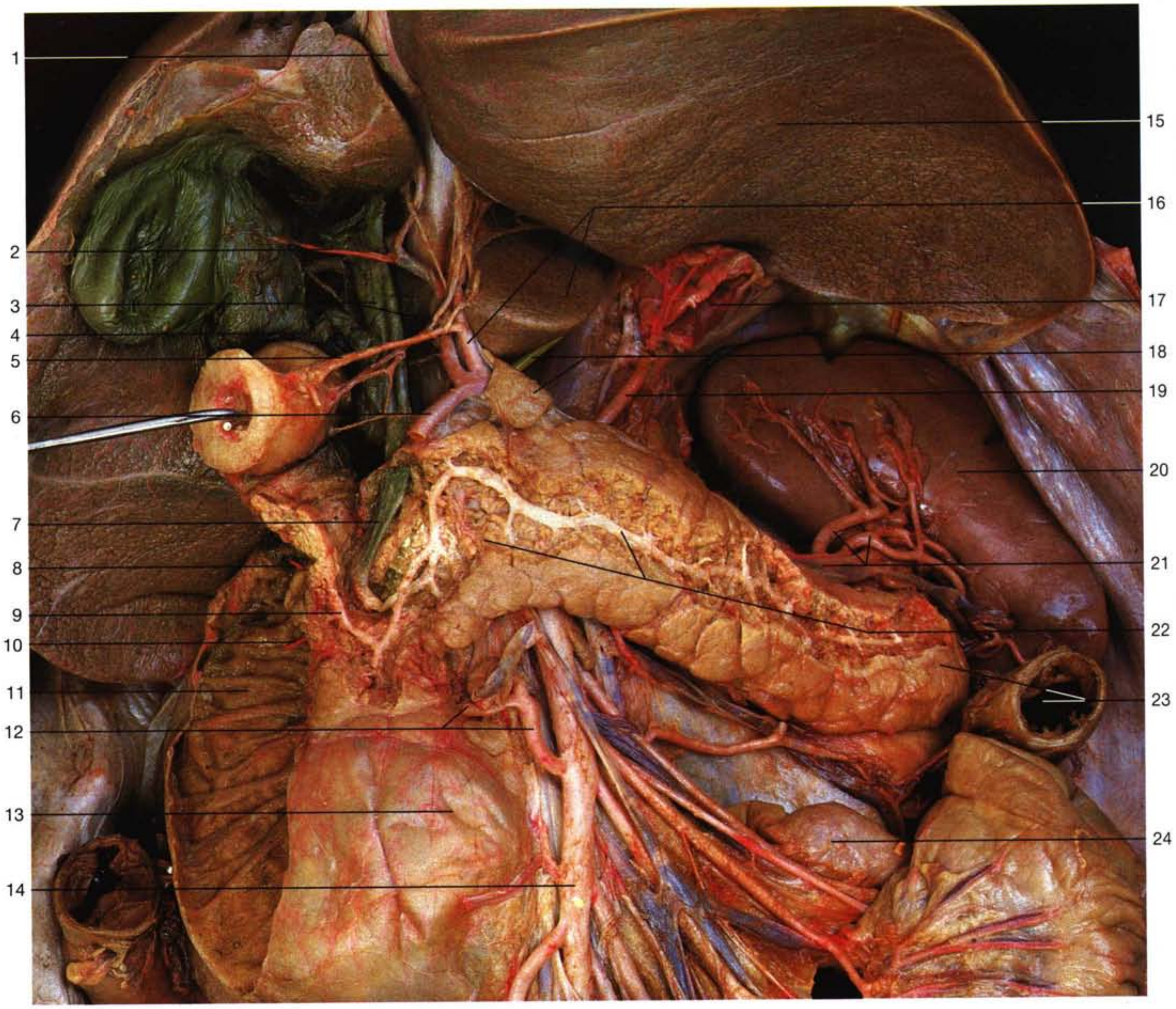

Posterior abdominal wall with duodenum, pancreas and spleen (anterior aspect). Dissection of pancreatic and common bile duct. The stomach has been removed, the liver raised and the duodenum anteriorly opened.

1	Ligamentum teres	13	Horizontal part of duodenum (distended)
2	Gallbladder and cystic artery	14	**Superior mesenteric artery**
3	Common hepatic duct and portal vein	15	Liver (left lobe)
4	Cystic duct	16	Caudate lobe of liver and hepatic artery proper
5	Right gastric artery (pylorus with superior part of duodenum, cut and reflected)	17	Abdominal part of esophagus (cut)
6	Gastroduodenal artery	18	Probe in epiploic foramen and lymph node
7	**Common bile duct**	19	Left gastric artery
8	Probe within the minor duodenal papilla	20	**Spleen**
9	**Accessory pancreatic duct**	21	Splenic vein and branches of splenic artery
10	Probe within the major duodenal papilla	22	**Pancreatic duct and head of pancreas**
11	Descending part of duodenum (opened)	23	Left colic flexure and tail of pancreas
12	Middle colic artery and inferior pancreaticoduodenal artery	24	Duodenojejunal flexure

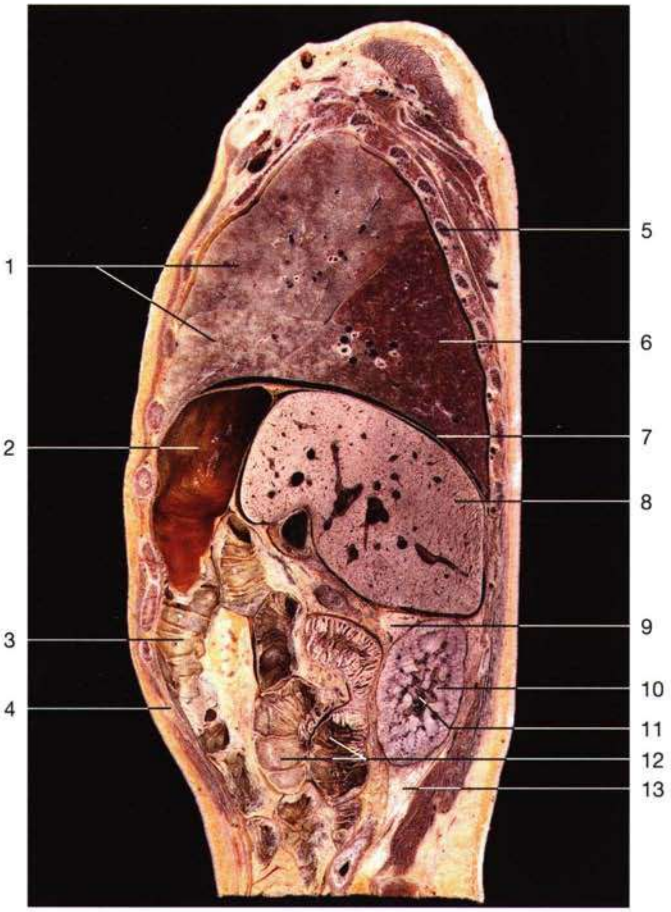

Parasagittal section through the thoracic and abdominal cavities (right side, medial aspect).

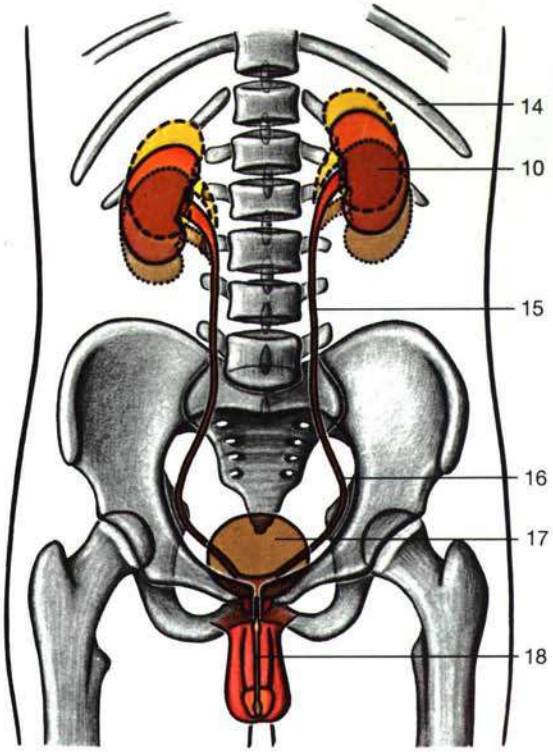

Position of kidneys and urinary system (anterior view). The excursions of the kidneys with the respiratory movements of the diaphragm are indicated. (Schematic drawing.)

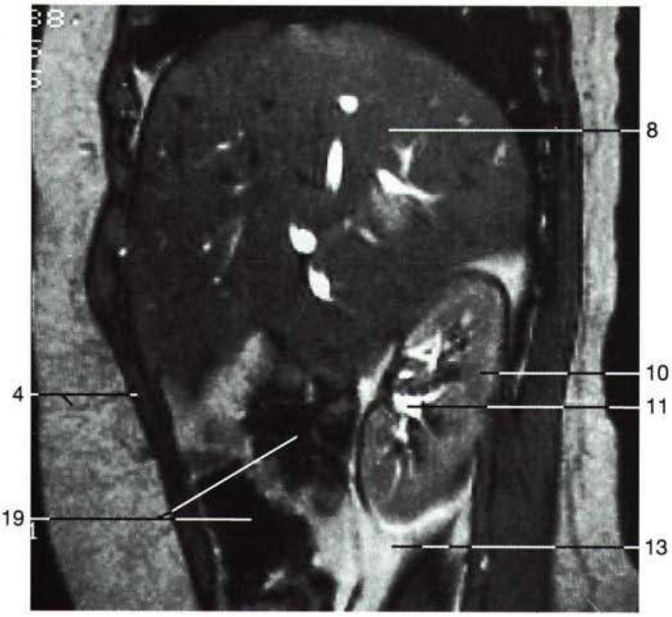

Parasagittal section through the abdominal cavity. (MR-Scan, courtesy of Prof. W. Rödl, Erlangen).

1 Right lung (superior and middle lobes)
2 Transverse colon
3 Jejunum
4 Abdominal wall
5 Fourth rib
6 Right lung (inferior lobe)
7 Diaphragm
8 **Liver**
9 Suprarenal gland
10 **Kidney**
11 Renal pelvis
12 Small intestine
13 Perirenal fat pad
14 Eleventh rib
15 Ureter (abdominal part)
16 Ureter (pelvic part)
17 **Urinary bladder**
18 Urethra
19 Colon

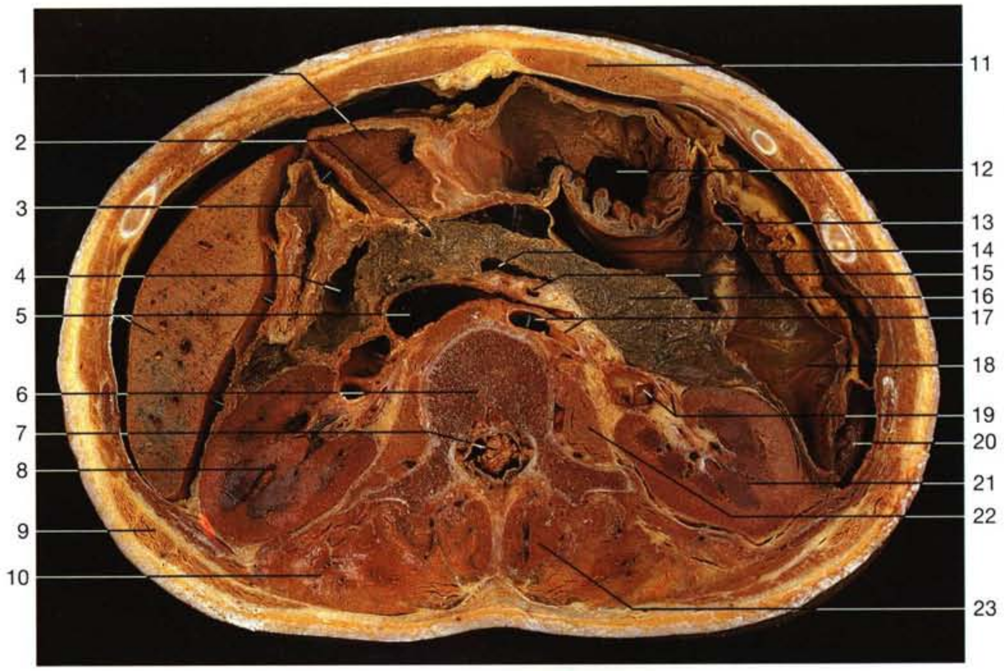

Horizontal section through the abdominal cavity at the level of the first lumbar vertebra (from below).

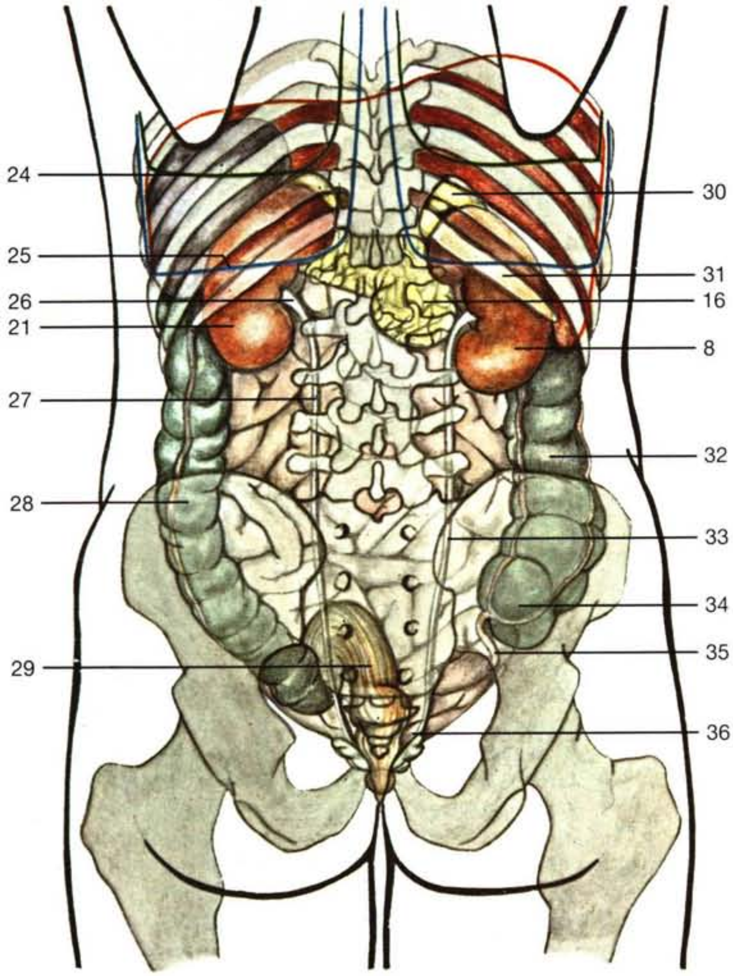

Positions of urinary organs (posterior view). Notice that the upper part of the kidney reaches the level of the margin of pleura and lung.

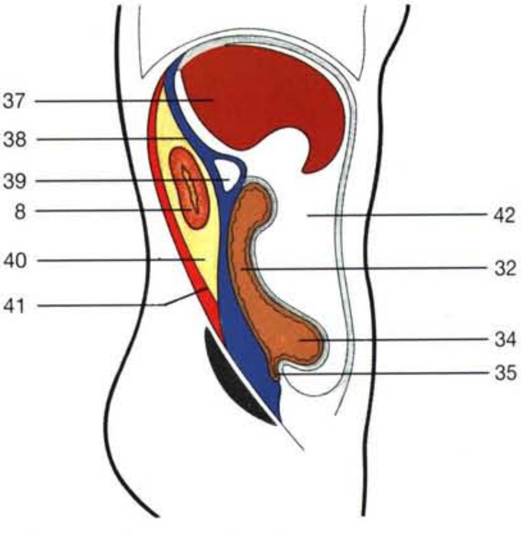

Retroperitoneal tissue, location of the right kidney. (Schematic drawing.) Yellow = adipose capsule of kidney.

1	Pyloric antrum
2	Gastroduodenal artery
3	Descending part of duodenum
4	Vestibule of lesser sac
5	Inferior vena cava and liver
6	Body of first lumbar vertebra
7	Cauda equina
8	**Right kidney**
9	Latissimus dorsi muscle
10	Iliocostalis muscle
11	Rectus abdominis muscle
12	Stomach
13	Lesser sac
14	Splenic vein
15	Superior mesenteric artery
16	Pancreas
17	Aorta and left renal artery
18	Transverse colon
19	Renal artery and vein
20	Spleen
21	**Left kidney**
22	Psoas major muscle
23	Multifidus muscle
24	Margin of lung
25	Margin of pleura
26	Renal pelvis
27	**Left ureter**
28	Descending colon
29	Rectum
30	Right suprarenal gland
31	12th rib
32	Ascending colon
33	**Right ureter**
34	Cecum
35	Vermiform appendix
36	Urinary bladder
37	Liver
38	Anterior layer of renal fascia
39	Duodenum
40	Perirenal fatty tissue
41	Posterior layer of renal fascia
42	Abdominal cavity

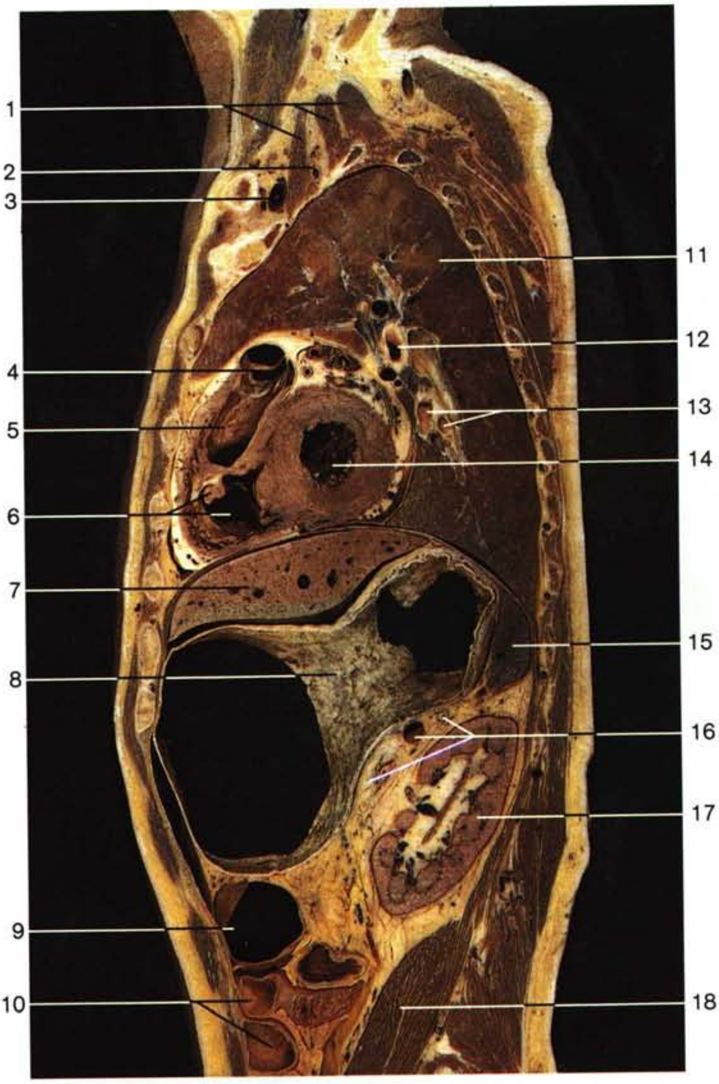

Parasagittal section through the thoracic and abdominal cavities at the level of the left kidney
(5.5 cm left of median plane).

1 Scalenus anterior, medius and posterior muscles
2 Left subclavian artery
3 Left subclavian vein
4 Pulmonic valve
5 Arterial cone
6 Right ventricle of heart
7 Liver
8 Stomach
9 Transverse colon
10 Small intestine
11 Left lung
12 Left main bronchus
13 Branches of pulmonary vein
14 Left ventricle of heart
15 Spleen
16 Splenic artery and vein and pancreas
17 **Left kidney**
18 Psoas major muscle
19 Inferior vena cava
20 **Renal vein**
21 Body of 12th thoracic vertebra and vertebral canal
22 **Right kidney**
23 Superior mesenteric artery
24 Superior mesenteric vein
25 Pancreas
26 Abdominal aorta
27 Left psoas major and quadratus lumborum muscles
28 Anterior layer of renal fascia } of Gerota
29 Posterior layer of renal fascia
30 Perirenal fatty tissue
31 Abdominal cavity
32 Descending and sigmoid colon

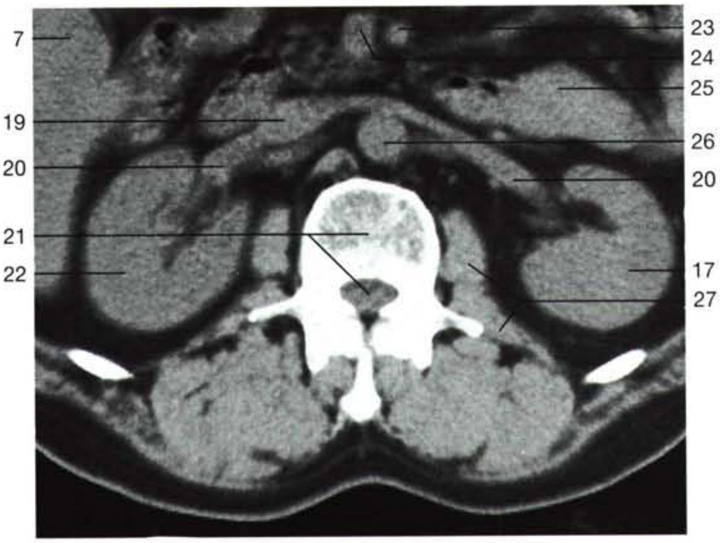

Horizontal section through the retroperitoneal region at level of 12th thoracic vertebra. (CT-Scan, from below.)

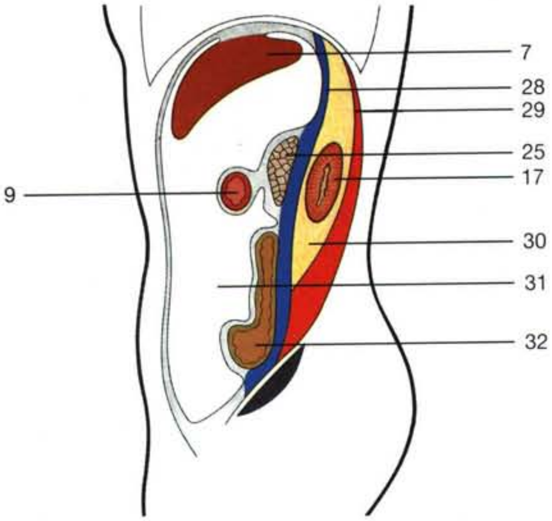

Retroperitoneal tissue, position of left kidney. (Schematic drawing.)

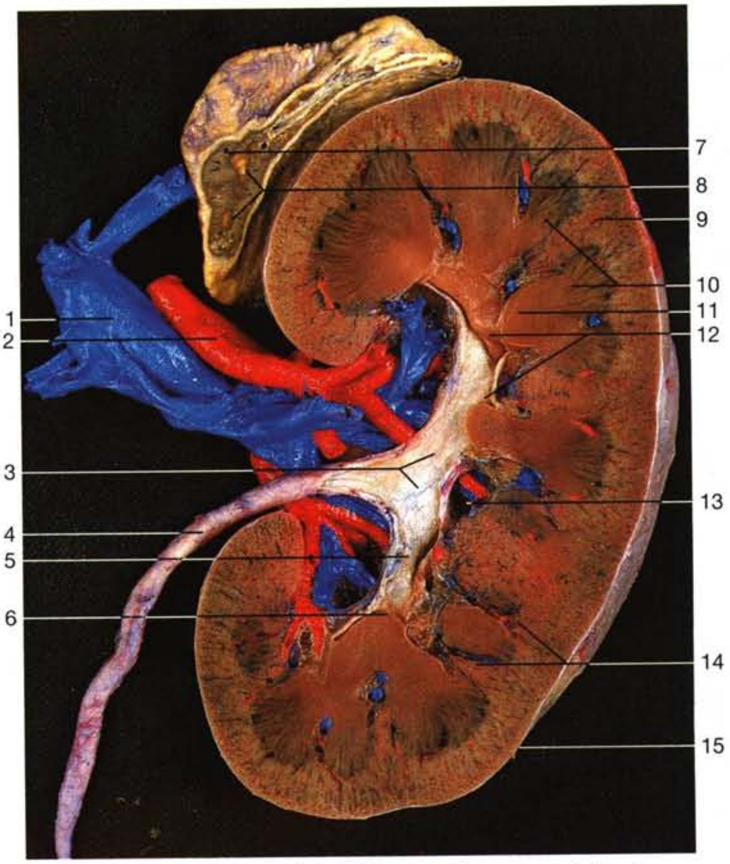

Coronal section through right kidney and suprarenal gland (posterior view). The renal pelvis has been opened and the fatty tissue removed to display the renal vessels.

1 Renal vein
2 Renal artery
3 **Renal pelvis**
4 Abdominal part of ureter
5 **Major renal calyx**
6 Cribriform area of renal papilla
7 Cortex of suprarenal gland
8 Medulla of suprarenal gland
9 **Cortex of kidney**
10 **Medulla of kidney**
11 Renal papilla
12 **Minor renal calyx**
13 Renal sinus
14 Renal columns
15 **Fibrous capsule** of kidney

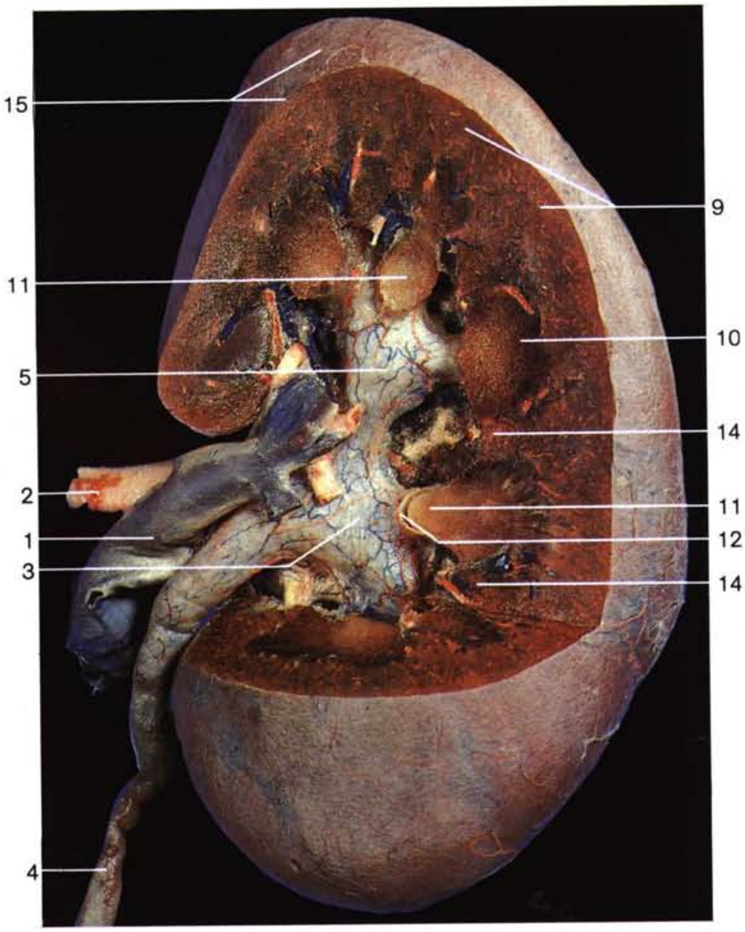

Right kidney (posterior view). Partial coronal section to expose internal aspect of the kidney.

Each kidney can be divided in five segments supplied by individual interlobar arteries considered as end arteries. Thus, obstruction leads to infarcts marking the trace of segment borders. The anterior kidney surface reveals four segments, the posterior only three (No. 1, 4 and 5).

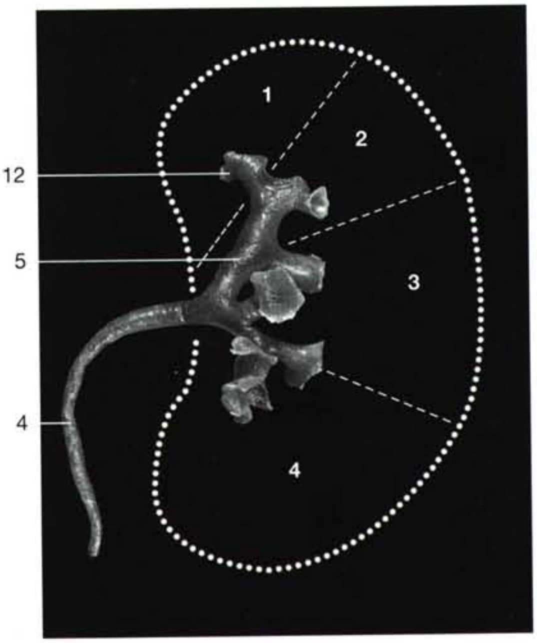

Cast of renal pelvis and calices.
1–4 = Renal segments on anterior surface.

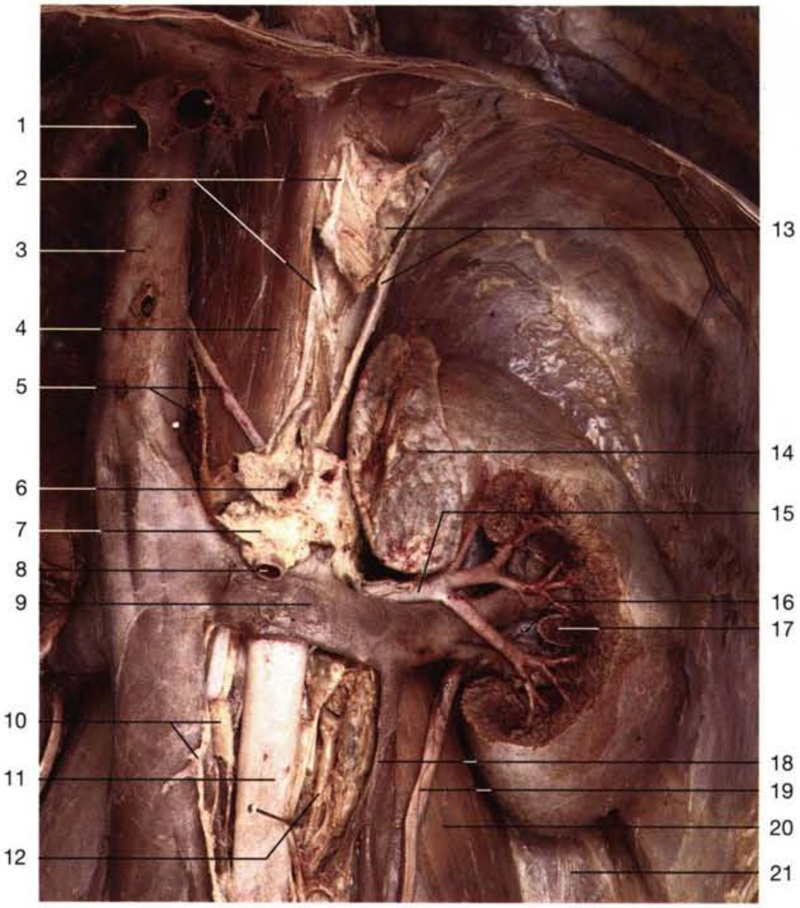

Left kidney and suprarenal gland in situ. The anterior cortical layer of the kidney has been removed to display the renal pelvis and papillae.

1 Hepatic vein
2 Anterior and posterior vagal trunk
3 Inferior vena cava
4 Lumbar part of diaphragm
5 Right greater and lesser splanchnic nerves
6 Celiac trunk
7 Celiac ganglion and plexus
8 Superior mesenteric artery
9 Left renal vein
10 Right sympathetic trunk and ganglion
11 Abdominal aorta
12 Left sympathetic trunk
13 Esophagus (cut), left greater splanchnic nerve
14 **Left suprarenal gland**
15 Left renal artery
16 Renal pelvis
17 Renal papilla with minor calyx
18 Left testicular vein
19 Left ureter
20 Psoas major muscle
21 Quadratus lumborum muscle
22 **Glomerulus**
23 Afferent arteriole of glomerulus
24 **Glomeruli**
25 Radiating cortical artery
26 Subcortical or arcuate artey
27 Subcortical or arcuate vein
28 Interlobular vein
29 Interlobular artery
30 **Interlobar artery and vein**
31 Vessels of renal capsule
32 Efferent arteriole of glomerulus
33 **Vasa recta** of renal medulla
34 Spiral arteries of renal pelvis

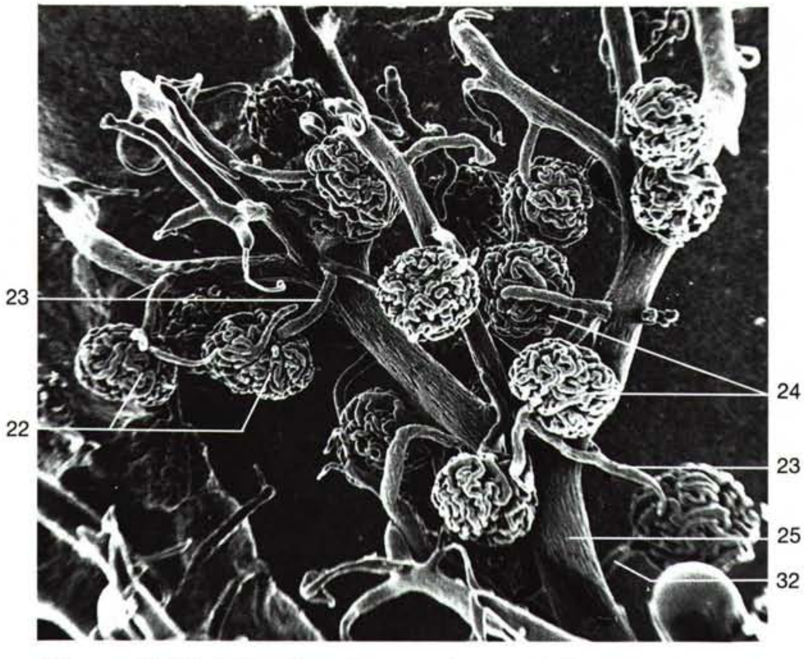

Glomeruli (210×). Scanning electron micrograph showing glomeruli and associated arteries.

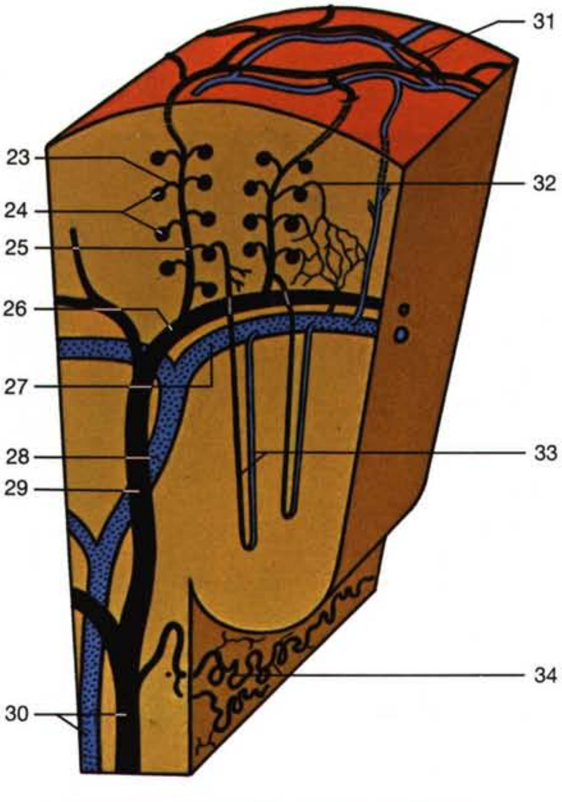

Architecture of vascular system of kidney. (Schematic drawing.)

1 Arteriolae rectae of renal medulla
2 Interlobar artery
3 Interlobular arteries
4 Cortical glomeruli
5 Juxtamedullary glomeruli
6 Body of first lumbar vertebra
7 Left renal artery
8 Abdominal aorta with catheter
9 Upper pole of kidney
10 Anterior branch } of renal artery
11 Posterior branch }
12 Anterior inferior segmental artery
13 Lower pole of kidney
14 Celiac trunk
15 Superior mesenteric artery
16 Middle colic artery
17 Splenic artery

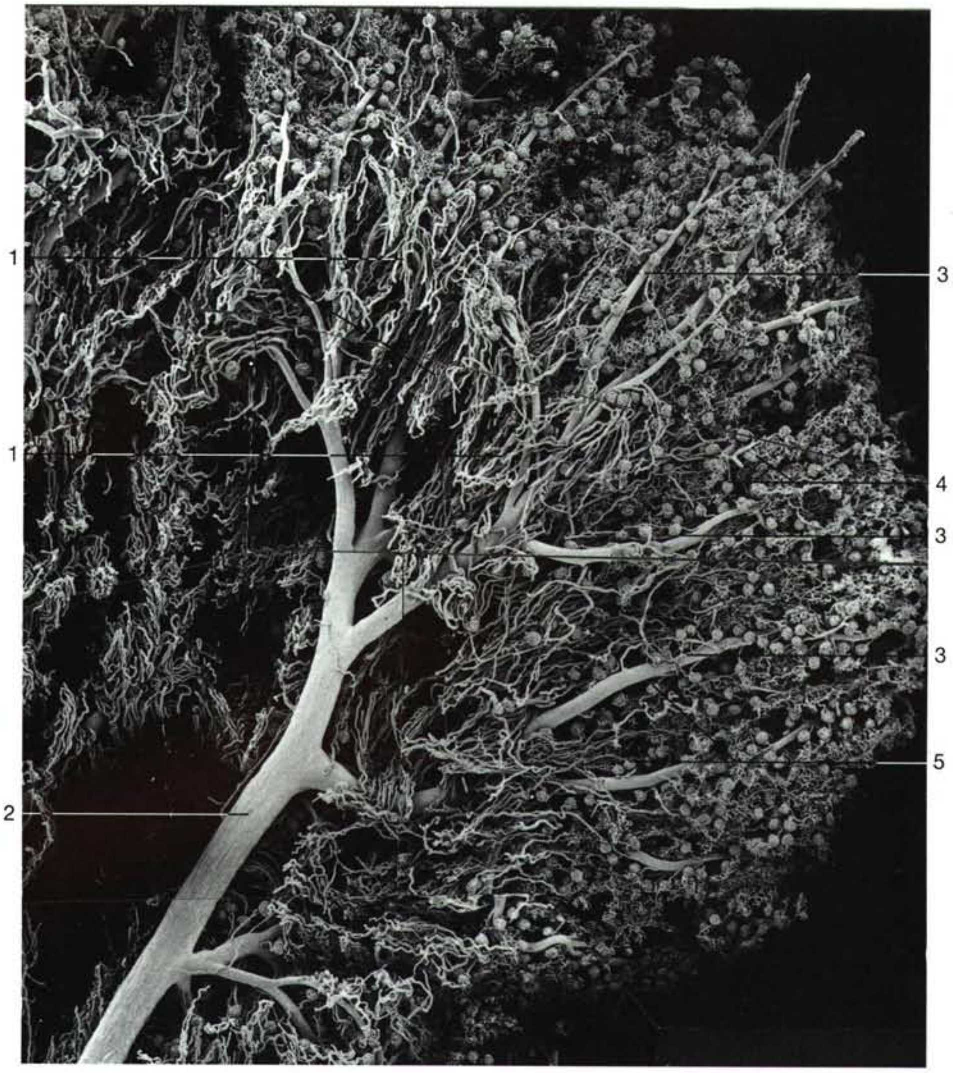

Resin cast of kidney arteries. (Scanning electron micrograph.)

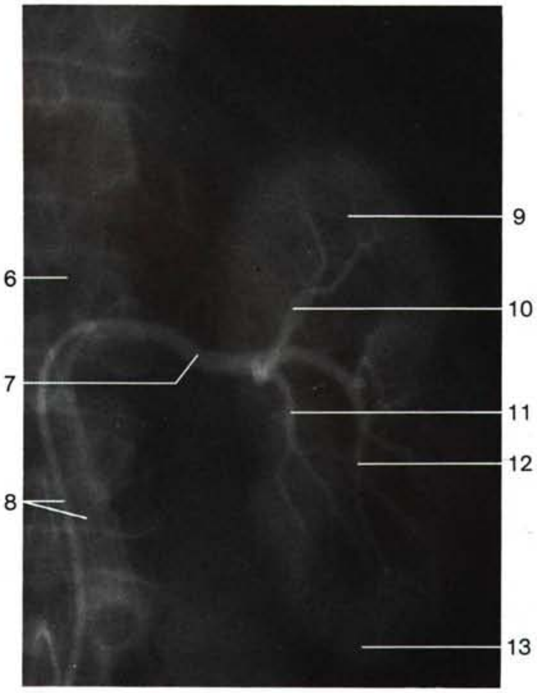

Left kidney. (Arteriogram.)

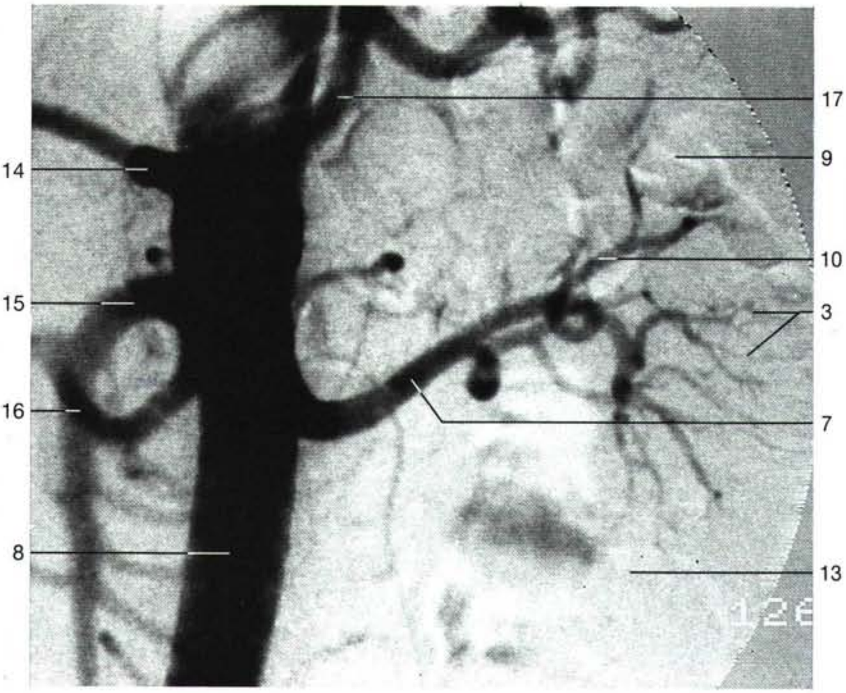

Abdominal aorta. (Subtraction angiograph.)

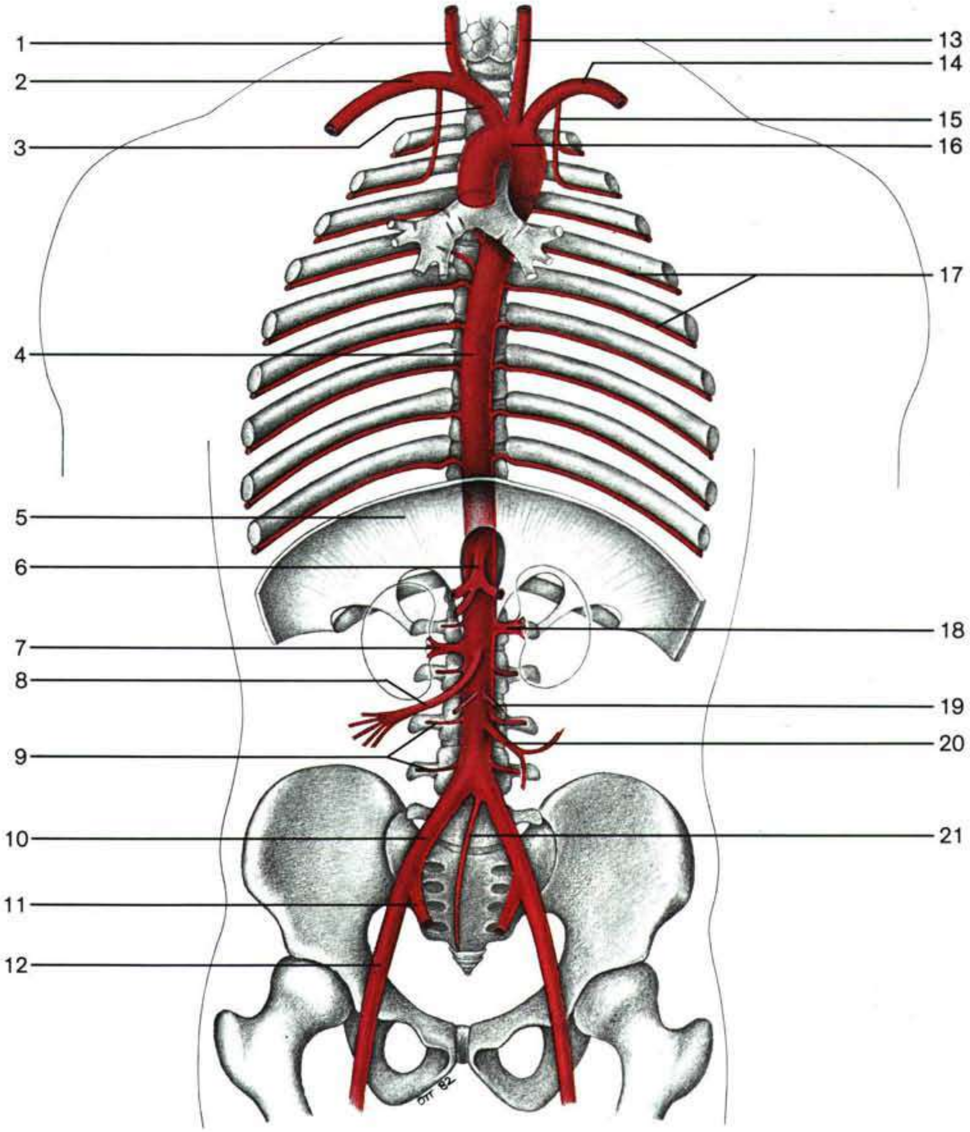

1 Right common carotid artery
2 Right subclavian artery
3 **Brachiocephalic trunk**
4 Thoracic aorta
5 Diaphragm
6 **Celiac trunk**
7 Right **renal artery**
8 **Superior mesenteric artery**
9 Lumbar arteries
10 Right **common iliac artery**
11 Internal iliac artery
12 **External iliac artery**
13 Left common carotid artery
14 Left subclavian artery
15 Highest intercostal artery
16 Aortic arch
17 **Posterior intercostal arteries**
18 Left renal artery
19 Left testicular (or ovarian) artery
20 Inferior mesenteric artery
21 Median sacral artery
22 **Superior suprarenal artery**
23 Upper capsular artery
24 Anterior branch of renal artery
25 Perforating artery
26 Lower capsular artery
27 Ureter
28 Right inferior phrenic artery
29 Left inferior phrenic artery
30 **Middle suprarenal artery**
31 **Inferior suprarenal artery**
32 Posterior branch of renal artery

Main branches of descending aorta. (Schematic drawing.)

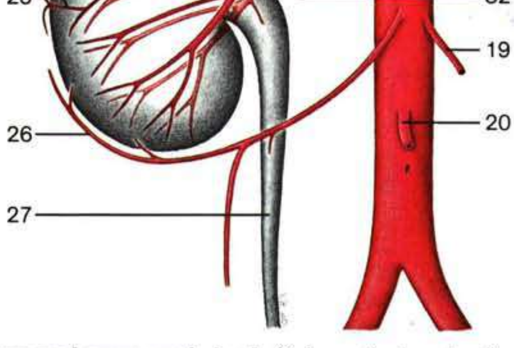

Arteries of kidney and suprarenal gland. (Schematic drawing.)

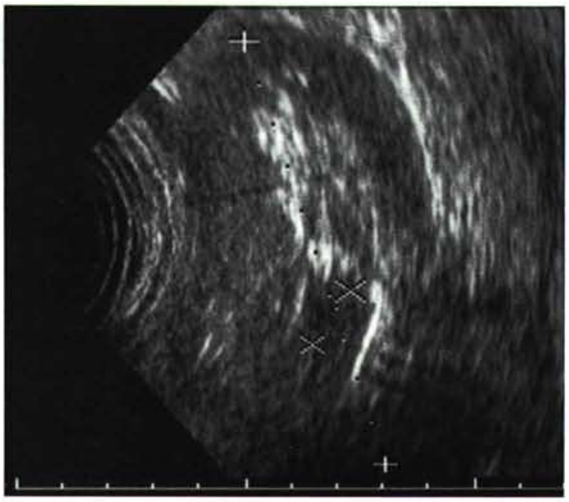

Ultrasound image of the right kidney (upper and lower border of the kidney marked by crosses.
× = small cortical cyst).

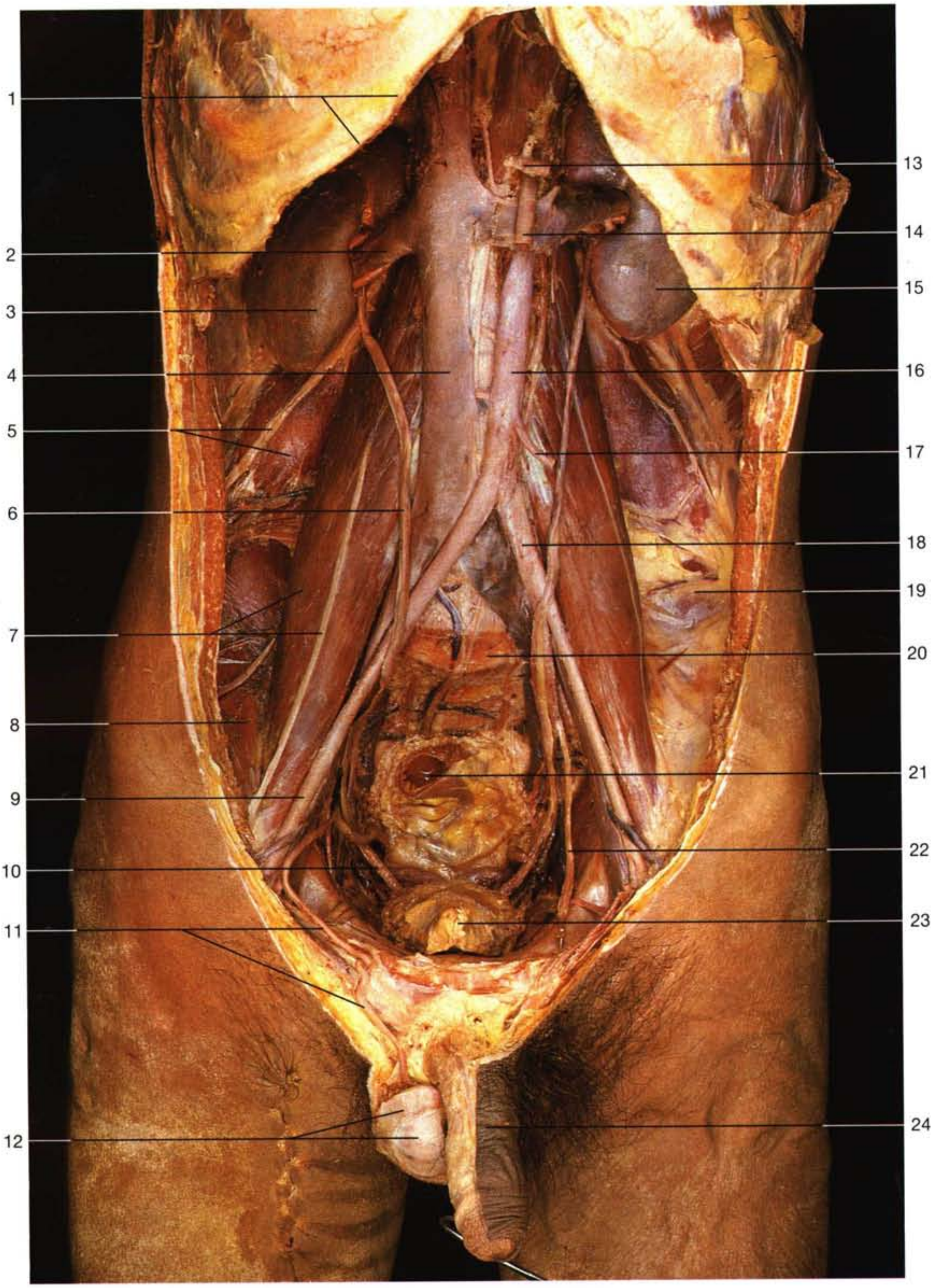

Retroperitoneal organs, urinary system in the male (anterior view). The peritoneum has been removed.

1	Costal arch
2	Right renal vein
3	**Right kidney**
4	Inferior vena cava
5	Iliohypogastric nerve and quadratus lumborum
6	**Ureter** (abdominal part)
7	Psoas major muscle and genitofemoral nerve
8	Iliacus muscle
9	External iliac artery
10	**Ureter** (pelvic part)
11	Ductus deferens
12	Testis and epididymis
13	Celiac trunk
14	Superior mesenteric artery
15	**Left kidney**
16	Abdominal aorta
17	Inferior mesenteric artery
18	Common iliac artery
19	Iliac crest
20	Sacral promontory
21	Rectum (cut)
22	Medial umbilical ligament
23	**Urinary bladder**
24	Penis

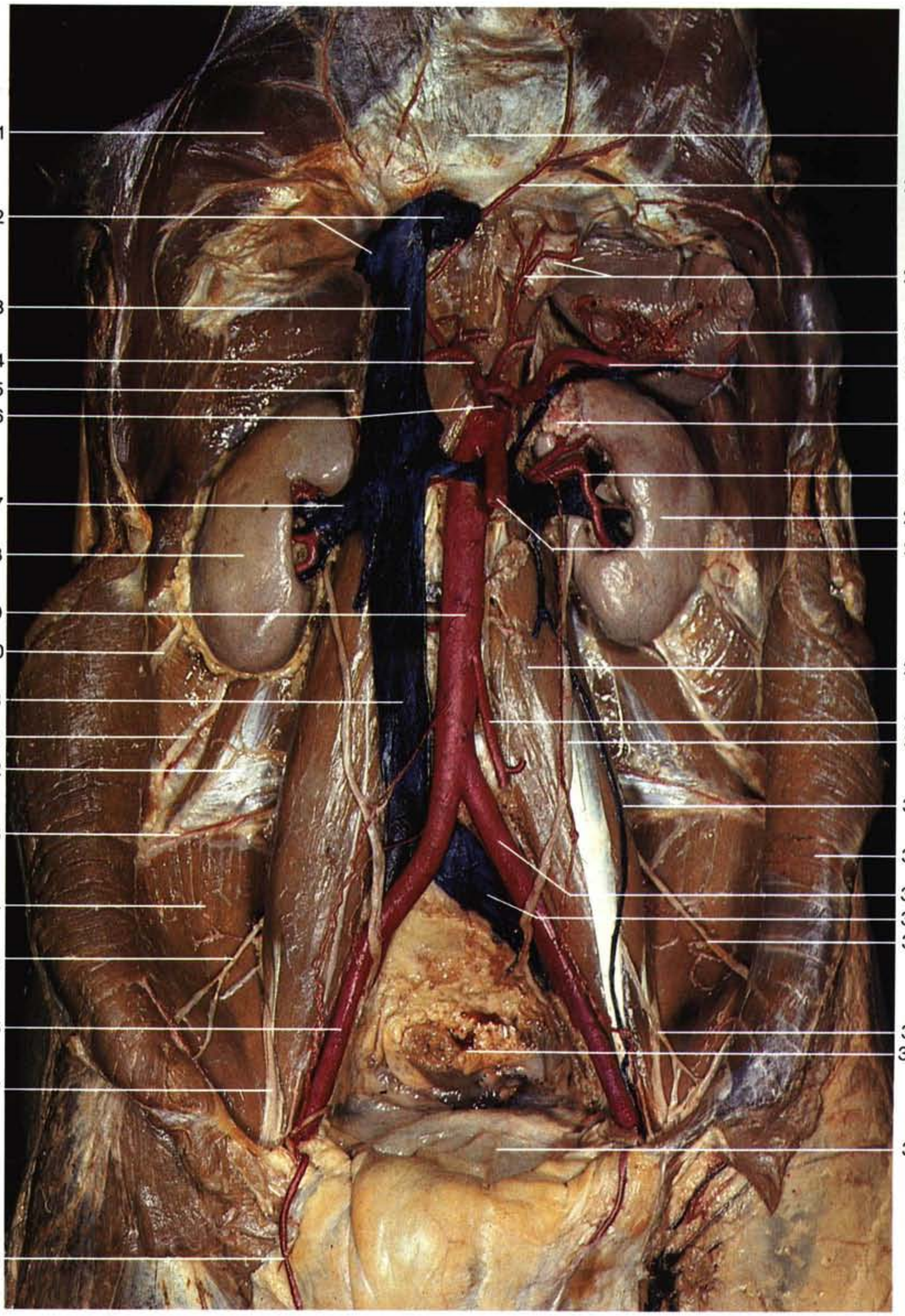

Retroperitoneal organs, urinary system in situ (anterior view). The peritoneum has been removed. Red = arteries; blue = veins.

1 Diaphragm
2 Hepatic veins
3 **Inferior vena cava**
4 Common hepatic artery
5 Right suprarenal gland
6 Celiac trunk
7 **Right renal vein**
8 Right kidney
9 **Abdominal aorta**
10 Subcostal nerve
11 Iliohypogastric nerve
12 Quadratus lumborum muscle
13 Iliac crest
14 Iliacus muscle
15 Right lateral femoral cutaneous nerve
16 **External iliac artery**
17 Femoral nerve
18 Right inferior epigastric artery
19 Central tendon of diaphragm
20 Inferior phrenic artery
21 Cardiac part of stomach and esophageal branches of left gastric artery
22 Spleen
23 Splenic artery
24 Left suprarenal gland
25 **Left renal artery**
26 Left **kidney**
27 Superior mesenteric artery
28 Psoas major muscle
29 Inferior mesenteric artery
30 **Ureter**
31 Testicular artery and vein
32 Transversus abdominis muscle
33 Left common iliac artery
34 Left common iliac vein
35 Lateral femoral cutaneous nerve
36 Genitofemoral nerve
37 Rectum (cut)
38 Urinary bladder

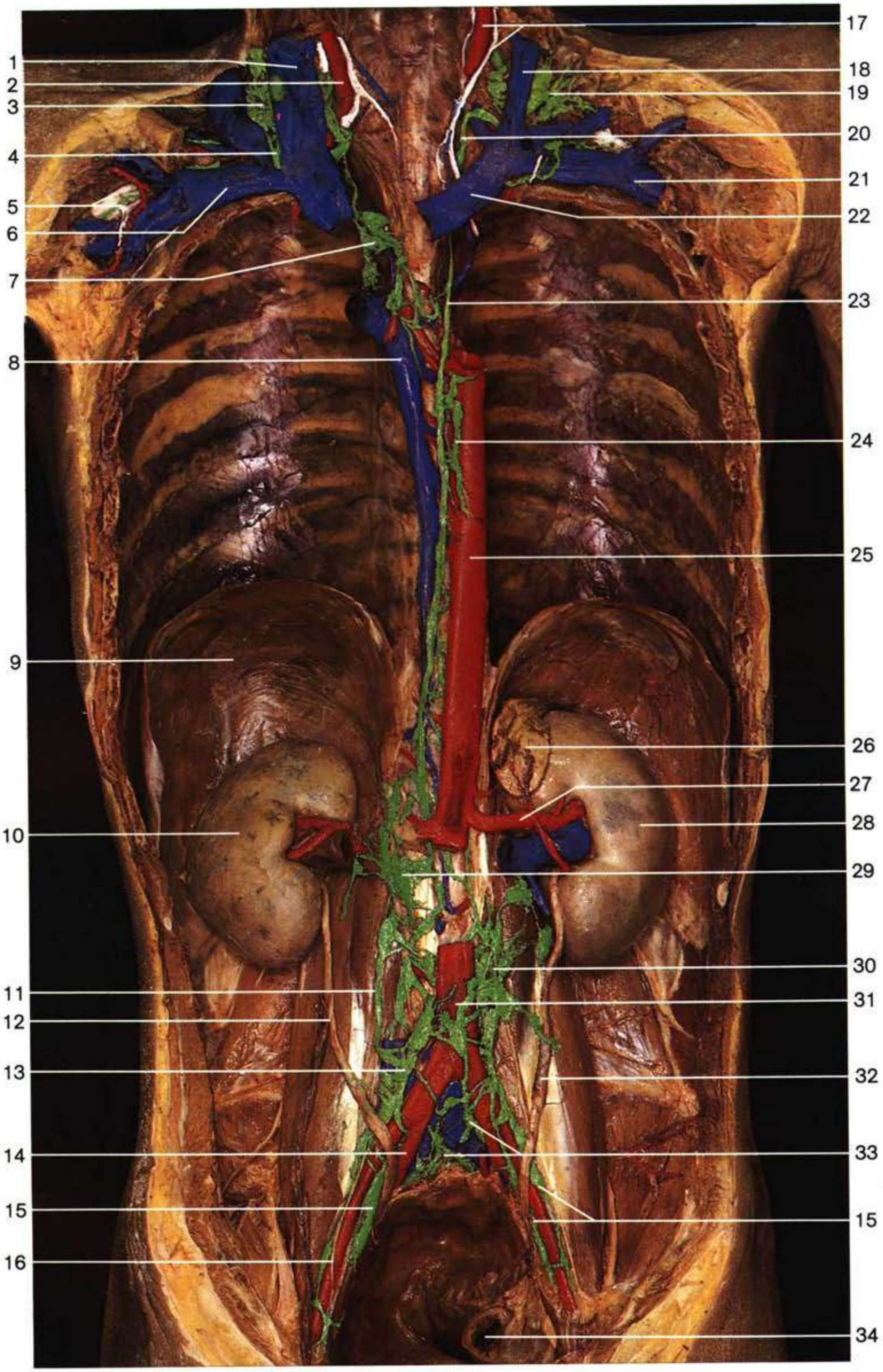

1 Internal jugular vein
2 Right common carotid artery and right vagus nerve
3 Jugulo-omohyoid lymph node
4 Right lymphatic duct
5 Subclavian trunk
6 Right subclavian vein
7 Bronchomediastinal trunk
8 Azygos vein
9 Diaphragm
10 Right kidney
11 Right lumbar trunk
12 Right ureter
13 Common iliac lymph nodes
14 Right internal iliac artery
15 External iliac lymph nodes
16 Right external iliac artery
17 Left common carotid artery and left vagus nerve
18 Internal jugular vein
19 Deep cervical lymph nodes
20 Thoracic duct entering left jugular angle
21 Left subclavian vein
22 Left brachiocephalic vein
23 **Thoracic duct**
24 Mediastinal lymph nodes
25 Thoracic aorta
26 Left suprarenal gland
27 Left renal artery
28 Left kidney
29 Cisterna chyli
30 Lumbar lymph nodes
31 Abdominal aorta
32 Left ureter
33 Sacral lymph nodes
34 Rectum (cut edge)

Lymph vessels and lymph nodes of the posterior wall of thoracic and abdominal cavities (anterior aspect). Green = lymph vessels and nodes; blue = veins; red = arteries; white = nerves.

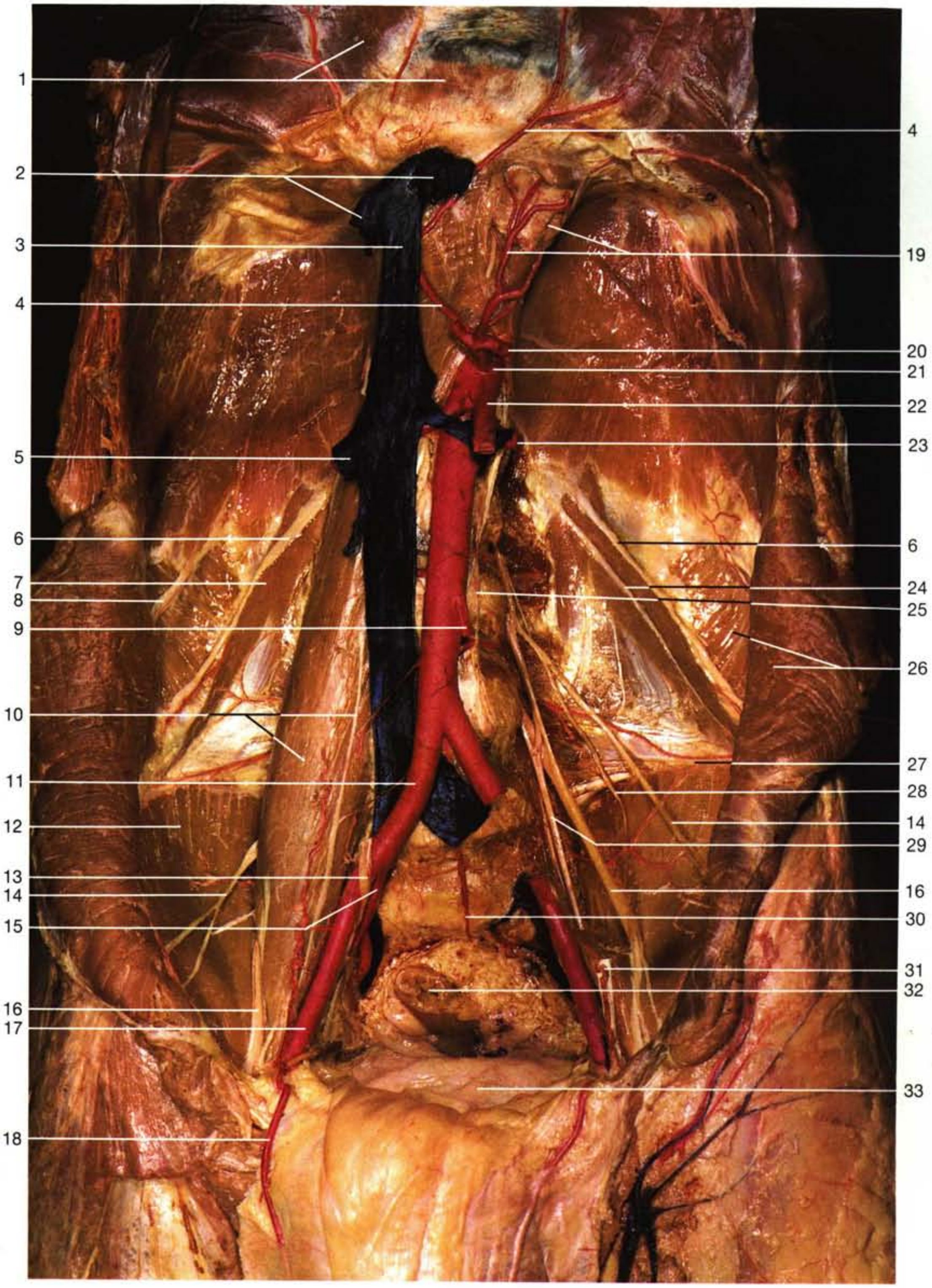

Vessels and nerves of posterior abdominal wall (anterior aspect). Part of the left psoas major muscle has been removed to display the lumbar plexus. Red = arteries, blue = veins.

1 Diaphragm
2 Hepatic veins
3 **Inferior vena cava**
4 Inferior phrenic artery
5 Right renal vein
6 **Iliohypogastric nerve**
7 Quadratus lumborum muscle
8 Subcostal nerve
9 Inferior mesenteric artery
10 Right genitofemoral nerve and psoas major muscle

11 Common iliac artery
12 Iliacus muscle
13 Right ureter (divided)
14 Lateral femoral cutaneous nerve
15 Internal iliac artery
16 **Femoral nerve**
17 **External iliac artery**
18 Inferior epigastric artery
19 Cardiac part of stomach and oesophageal branches of left gastric artery

20 Splenic artery
21 Celiac trunk
22 **Superior mesenteric artery**
23 Left renal artery
24 **Ilioinguinal nerve**
25 Sympathetic trunk
26 Transversus abdominis muscle
27 Iliac crest
28 Left **genitofemoral nerve**
29 Left **obturator nerve**
30 Median sacral artery

31 Psoas major muscle (divided) with supplying artery
32 Rectum (cut)
33 Urinary bladder

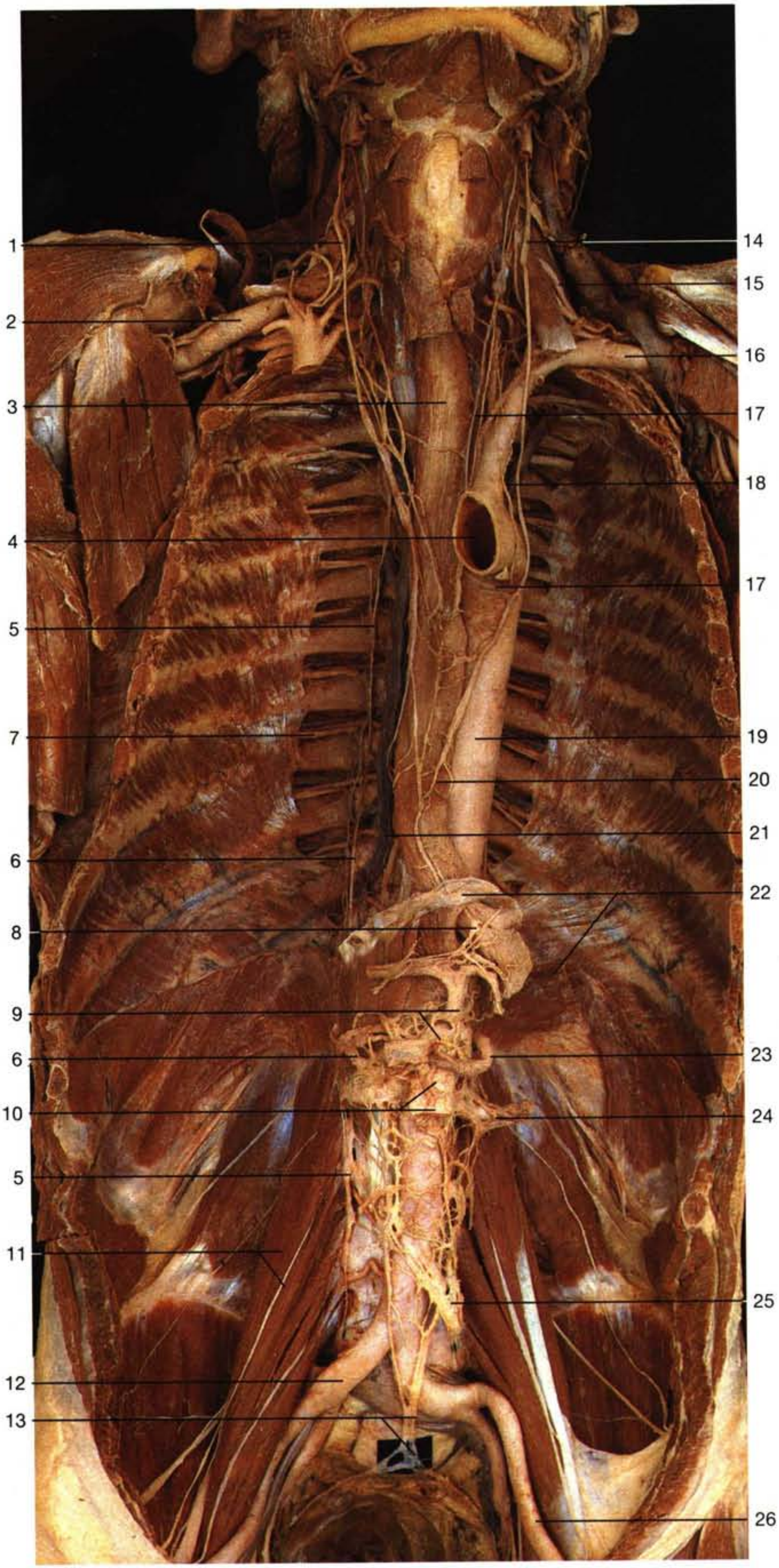
1
2
3
4
5
7
6
8
9
6
10
5
11
12
13
14
15
16
17
18
17
19
20
21
22
23
24
25
26

Posterior wall of thoracic and abdominal cavities with sympathetic trunk, vagus nerve and autonomic ganglia (anterior view). Thoracic and abdominal organs removed, except esophagus and aorta.

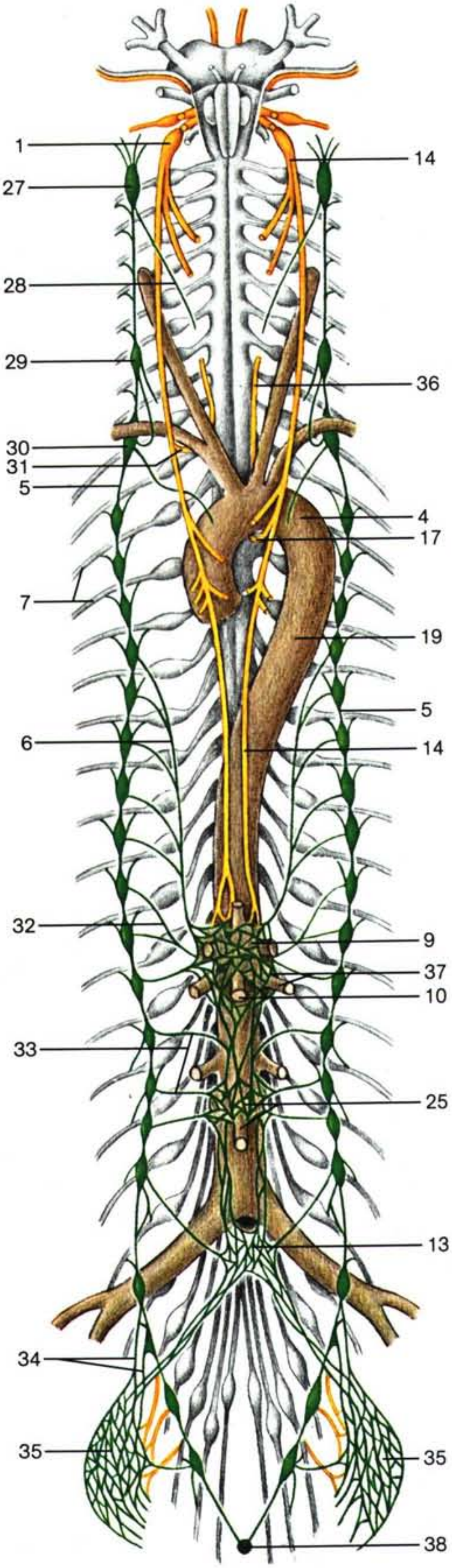
1
27
28
29
30
31
5
7
6
32
33
34
35
14
36
4
17
19
5
14
9
37
10
25
13
35
38

Organization of autonomic nervous system (after Mattuschka). (Schematic drawing.) Yellow = parasympathetic nerves; green = sympathetic nerves.

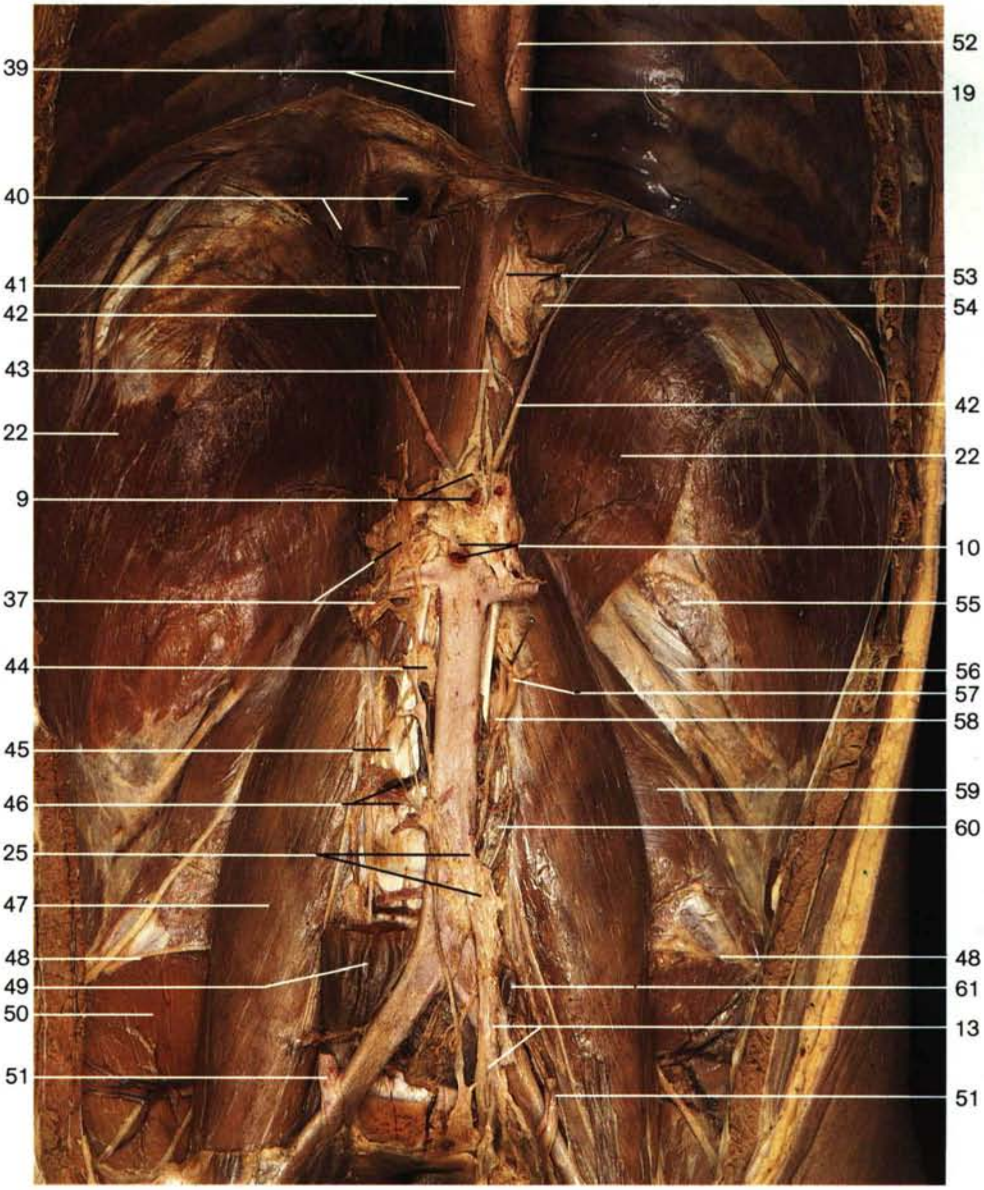

Ganglia and plexus of the autonomic nervous system within the retroperitoneal space (anterior view). The kidneys and the inferior vena cava with its tributaries have been removed (compare pp. 264 and 265).

1. **Right vagus nerve**
2. Right subclavian artery
3. Esophagus
4. Aortic arch
5. **Sympathetic trunk**
6. Greater splanchnic nerve
7. Intercostal nerve
8. Abdominal part of esophagus and vagal trunk
9. **Celiac trunk with celiac ganglion**
10. **Superior mesenteric artery and ganglion**
11. Psoas major muscle and genitofemoral nerve
12. Common iliac artery
13. **Superior hypogastric plexus and ganglion**
14. **Left vagus nerve**
15. Brachial plexus
16. Left subclavian artery
17. Left recurrent laryngeal nerve
18. Inferior cervical cardiac nerve
19. Thoracic aorta
20. Esophageal plexus
21. Azygos vein
22. Diaphragm
23. Splenic artery
24. Left renal artery and plexus
25. **Inferior mesenteric ganglion and artery**
26. Left external iliac artery
27. **Superior cervical ganglion** of sympathetic trunk
28. Superior cardiac branch of sympathetic trunk
29. Middle cervical ganglion of sympathetic trunk
30. **Inferior cervical ganglion** of sympathetic trunk
31. Right recurrent laryngeal nerve
32. Lesser splanchnic nerve
33. Lumbar splanchnic nerves
34. Sacral splanchnic nerves
35. **Inferior hypogastric ganglion and plexus**
36. Left recurrent laryngeal nerve
37. **Aorticorenal plexus** and renal artery
38. Ganglion impar
39. Esophagus with branches of vagus nerve
40. Hepatic veins
41. Right crus of diaphragm
42. Inferior phrenic artery
43. Right vagus nerve entering the celiac ganglion
44. Right lumbar lymph trunk
45. Lumbar part of right sympathetic trunk
46. Lumbar artery and vein
47. Psoas major muscle
48. Iliac crest
49. Inferior vena cava
50. Iliacus muscle
51. Ureter
52. Left vagus nerve forming the esophageal plexus
53. Left vagus nerve forming the gastric plexus
54. Esophagus continuing into the cardiac part of stomach
55. Lumbocostal triangle
56. Position of 12th rib
57. Left lumbar lymph trunk
58. Ganglion of sympathetic trunk
59. Quadratus lumborum muscle
60. Lumbar part of left sympathetic trunk
61. Iliac lymph vessels

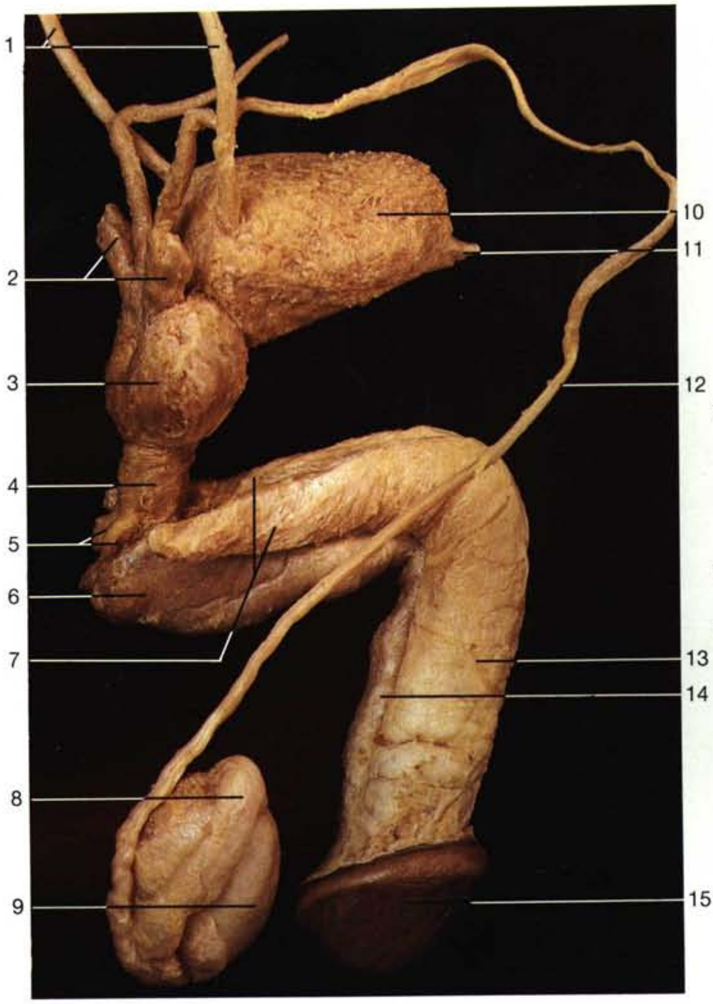
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Male genital organs isolated (right lateral aspect).

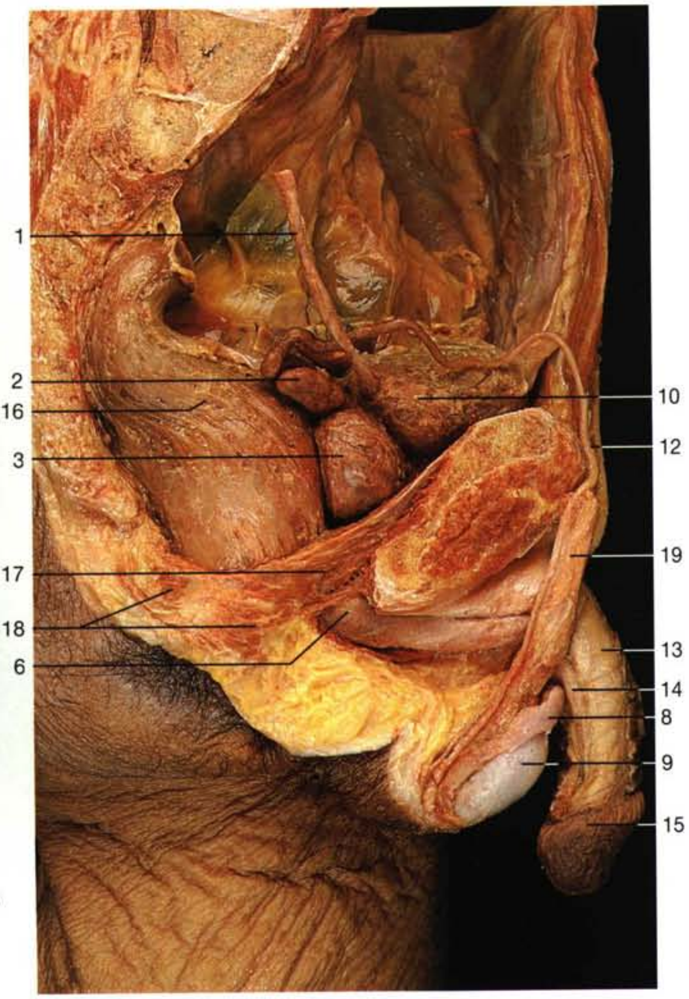
1
2
16
3
17
18
6
10
12
19
13
14
8
9
15

Male genital organs in situ (right lateral aspect).

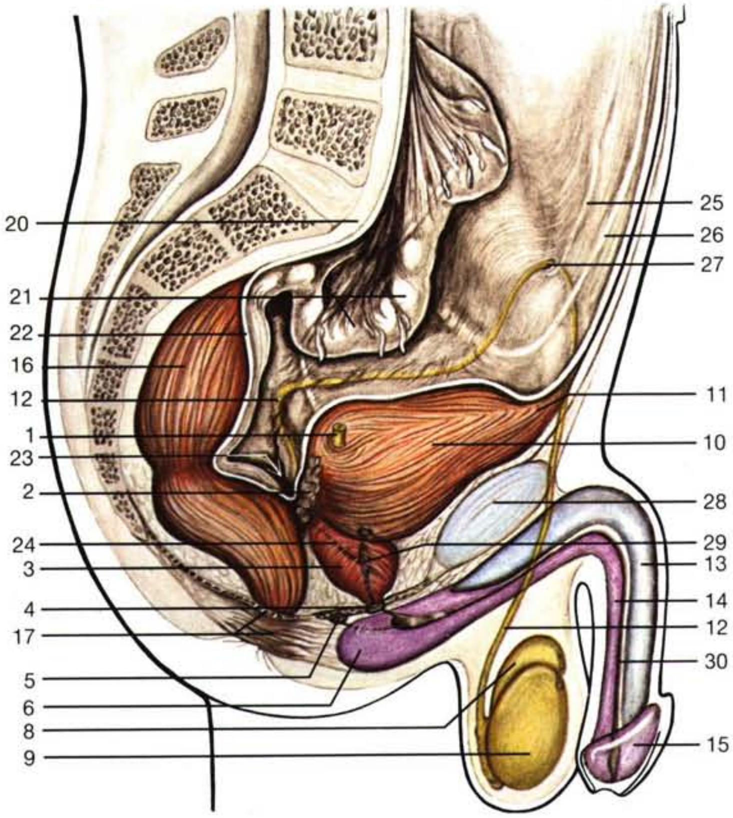
20
21
22
16
12
1
23
2
24
3
4
17
5
6
8
9
25
26
27
11
10
28
29
13
14
12
30
15

Positions of male genital organs (right lateral aspect).
(Schematic drawing.)

1 Ureter
2 **Seminal vesicle**
3 **Prostate gland**
4 Urogenital diaphragm and membranous part of urethra
5 Bulbourethral or Cowper's gland
6 Bulb of penis
7 Left and right crus penis
8 **Epididymis**
9 **Testis**
10 **Urinary bladder**
11 Apex of urinary bladder
12 **Ductus deferens**
13 **Corpus cavernosum of penis**
14 **Corpus spongiosum of penis**
15 Glans penis
16 Ampulla of rectum
17 Levator ani muscle
18 Anal canal and external anal sphincter muscle
19 Spermatic cord (cut)
20 Sacral promontory
21 Sigmoid colon
22 Peritoneum (cut edge)
23 **Rectovesical pouch**
24 Ejaculatory duct
25 Lateral umbilical fold
26 Medial umbilical fold
27 Deep inguinal ring and ductus deferens
28 Pubic symphysis
29 Prostatic part of urethra
30 Spongy urethra

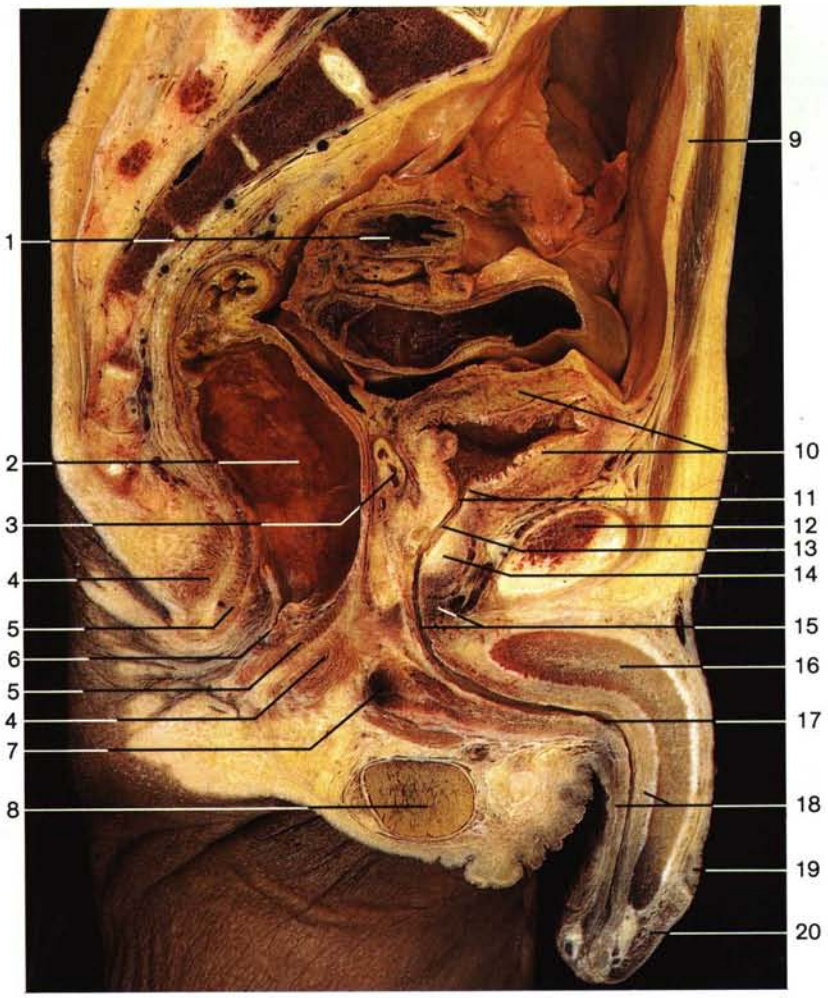

Midsagittal section through the male pelvis.

1 Sigmoid colon
2 **Ampulla of rectum**
3 Ampulla of ductus deferens
4 External anal sphincter muscle
5 Internal anal sphincter muscle
6 **Anal canal**
7 Bulb of penis
8 **Testis** (cut surface)
9 Median umbilical ligament
10 **Urinary bladder**
11 Internal urethral orifice and sphincter muscle (sphincter vesicae)
12 Pubic symphysis
13 Prostatic part of urethra
14 **Prostate gland**
15 Membranous urethra and external urethral sphincter muscle
16 **Corpus cavernosum of penis**
17 **Spongy urethra**
18 Corpus spongiosum of penis
19 Foreskin or prepuce
20 Glans penis
21 Kidney
22 Renal pelvis
23 Abdominal part of ureter
24 Pelvic part of ureter
25 Seminal vesicle
26 Ejaculatory duct
27 Bulbourethral or Cowper's gland
28 **Ductus deferens**
29 **Epididymis**
30 Umbilicus
31 Trigone of bladder and ureteric orifice
32 Navicular fossa of urethra
33 External urethral orifice
34 **Testis**

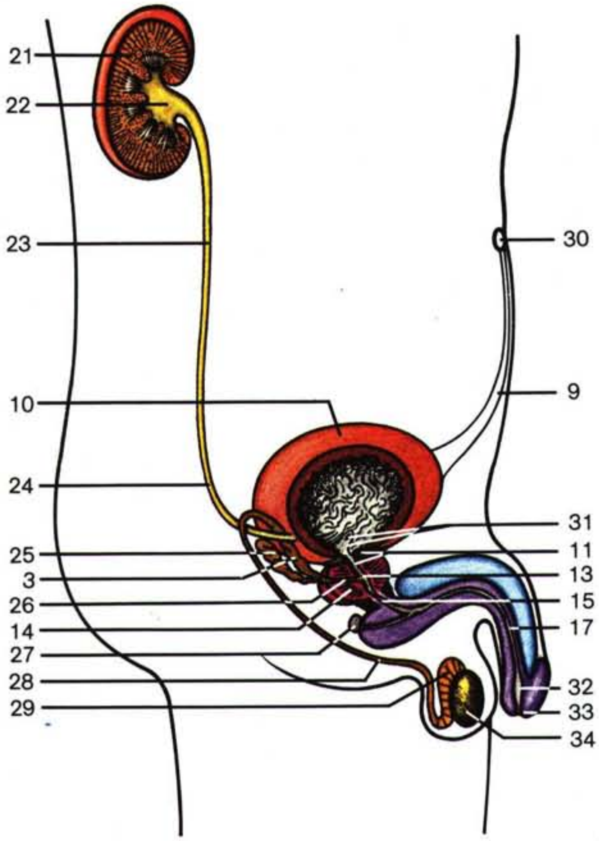

Male urogenital system. (Schematic drawing.)

The **prostate** is located between bladder and urogenital diaphragm. The penis includes the **urethra** and thus serves for both ejaculation and micturition. The internal (involuntary) and external (voluntary) urethral sphincters are widely separated. The **ureter** having crossed the ductus deferens enters the urinary bladder at its base. The peritoneum is reflected off the posterior surface of the bladder onto the rectum thus forming the rectovesical pouch.

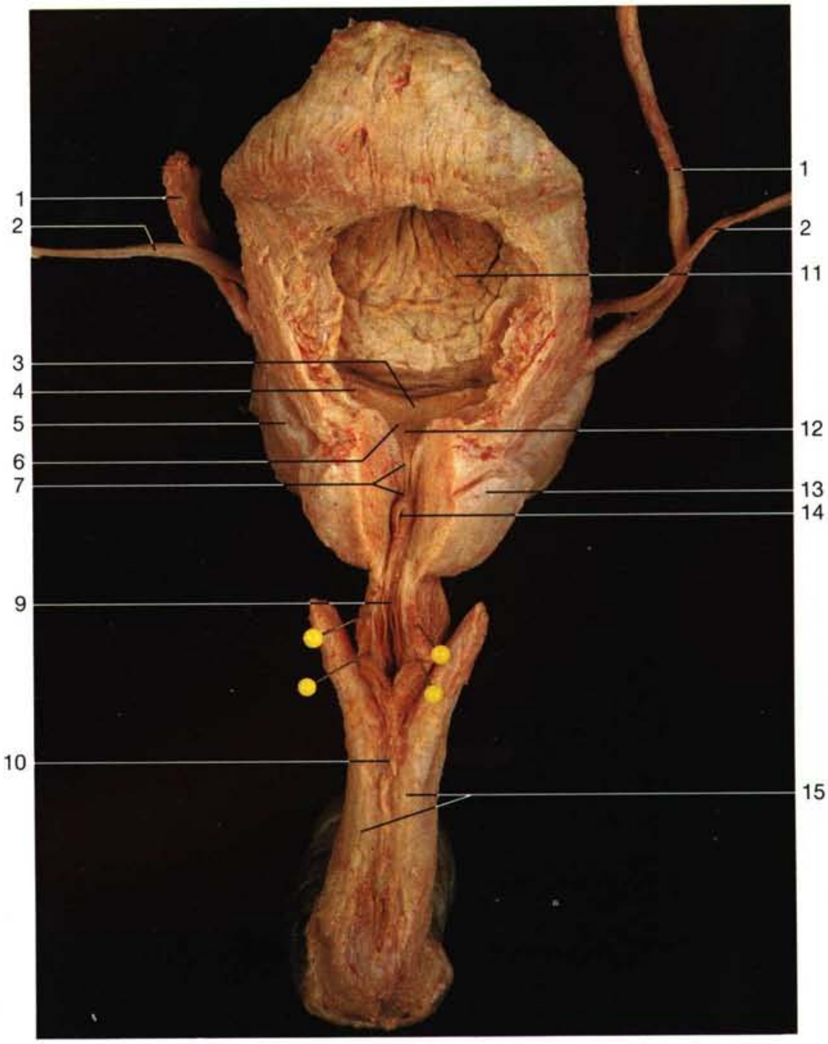

Male urogenital organs; isolated (anterior view). Urinary bladder, prostate and urethra have been opened. The urinary bladder is contracted.

1 Ureter
2 Ductus deferens
3 Interureteric fold
4 Ureteric orifice
5 Seminal vesicle
6 Trigone of bladder
7 Prostatic urethra with seminal colliculus and urethral crest
8 Deep transverse perineal muscle
9 Membranous urethra
10 **Spongy urethra**
11 Mucous membrane of **urinary bladder**
12 Internal urethral orifice and uvula of bladder
13 **Prostate**
14 Prostatic utricle
15 Right and left **corpus cavernosum of penis**
16 Ejaculatory duct
17 **Sphincter urethrae muscle**
18 **Median umbilical fold** with remnant of urachus
19 Medial umbilical fold with remnant of umbilical artery
20 **Urinary bladder**
21 Rectovesical pouch
22 Rectum
23 Sacrum
24 Deep iliac circumflex artery
25 Deep inguinal ring and ductus deferens
26 External iliac artery and vein
27 **Femoral nerve**
28 Obturator nerve
29 Ilium and sacrum
30 Inferior epigastric artery
31 Iliopsoas muscle

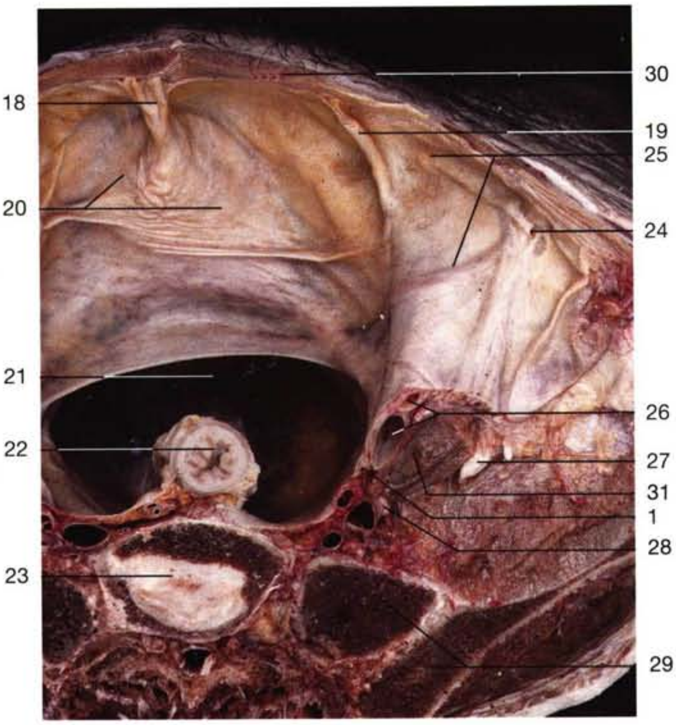

Pelvic cavity in the male (viewed from above).

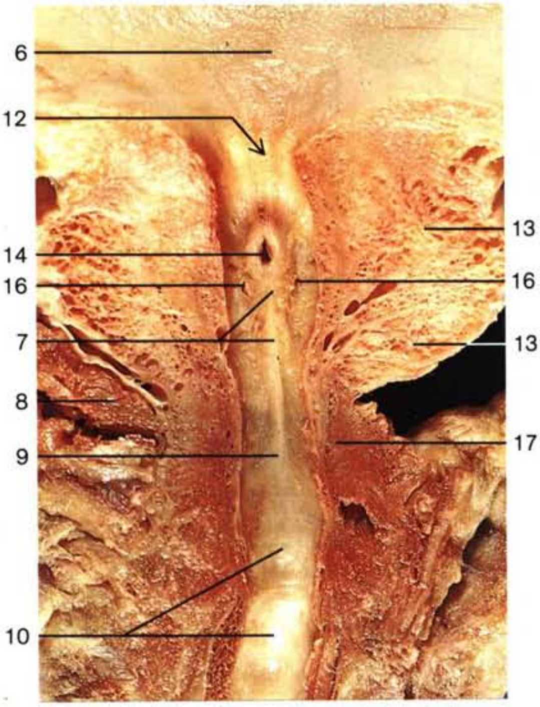

Posterior half of male urethra and prostate in continuity with neck of bladder (anterior view).

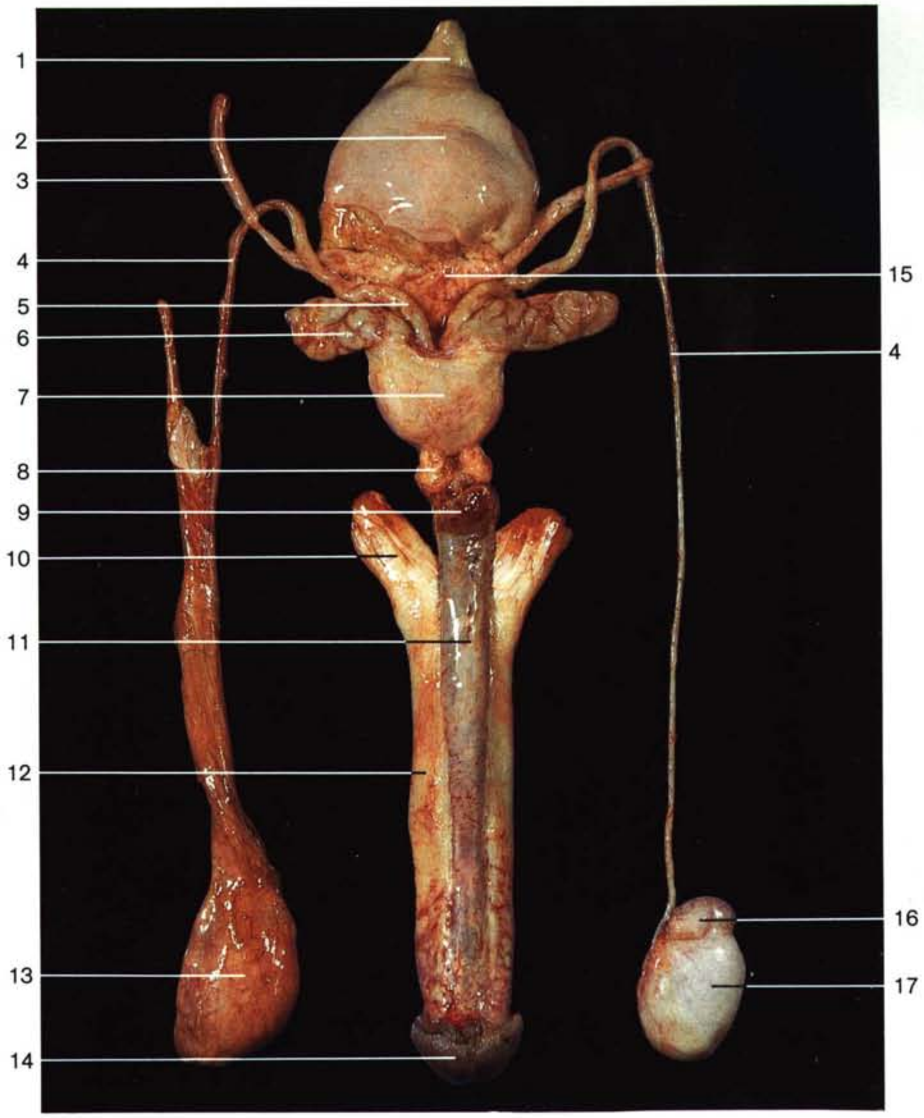

1 Apex of urinary bladder with urachus
2 **Urinary bladder**
3 Ureter
4 **Ductus deferens**
5 Ampulla of ductus deferens
6 **Seminal vesicle**
7 **Prostate**
8 **Bulbourethral or Cowper's gland**
9 **Bulb of penis**
10 Crus penis
11 **Corpus spongiosum of penis**
12 **Corpus cavernosum of penis**
13 Testis and epididymis with coverings
14 Glans penis
15 Fundus of bladder
16 **Head of epididymis**
17 **Testis**
18 Mucous membrane of bladder
19 Trigone of bladder
20 Ureteric orifice
21 Internal urethral orifice
22 Seminal colliculus
23 Prostate
24 Prostatic urethra
25 Membranous urethra
26 Spongy (penile) urethra
27 Skin of penis
28 Deep dorsal vein of penis (unpaired)
29 Dorsal artery of penis (paired)
30 **Tunica albuginea of corpora cavernosa**
31 Septum of penis
32 Deep artery of penis
33 Tunica albuginea of corpus spongiosum
34 Deep fascia of penis

Male genital organs, isolated (posterior view).

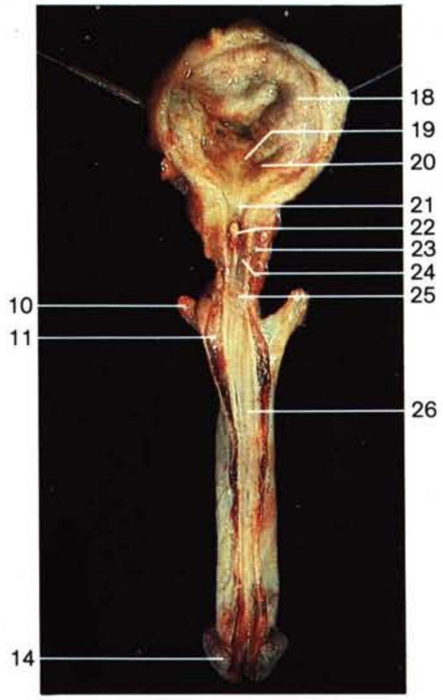

Urinary bladder, urethra and penis (anterior view, open longitudinally).

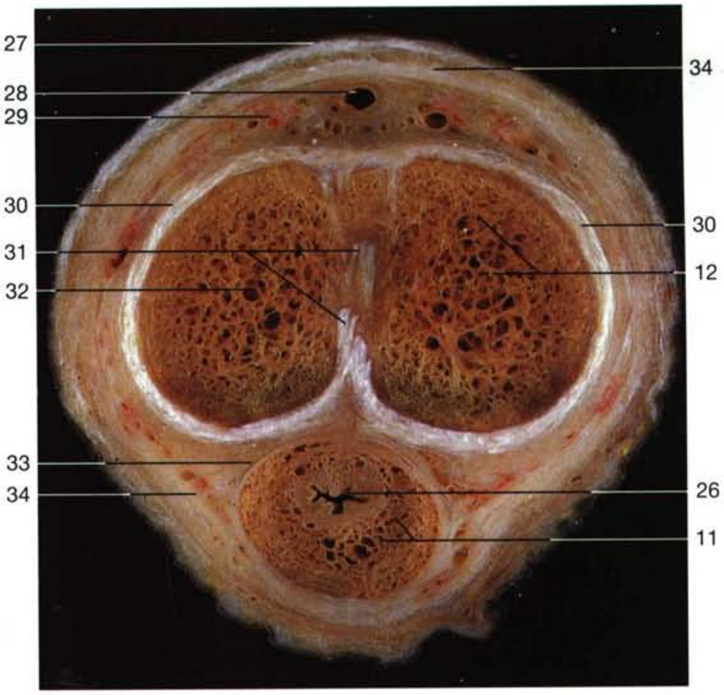

Cross-section of penis (anterior view).

1 **Corpus cavernosum of penis**
2 **Corpus spongiosum of penis**
3 Corona of glans penis
4 **Glans penis**
5 Suspensory ligament of penis
6 Inferior pubic ramus
7 Crus penis
8 **Bulb of penis**
9 Deep dorsal vein of penis
10 Septum pectiniforme
11 Dorsal artery of penis
12 Bulbourethral or Cowper's gland
13 Urinary bladder
14 **Seminal vesicle**
15 Ampulla of ductus deferens
16 Ductus deferens
17 Membranous urethra
18 **Prostate**
19 Ureter

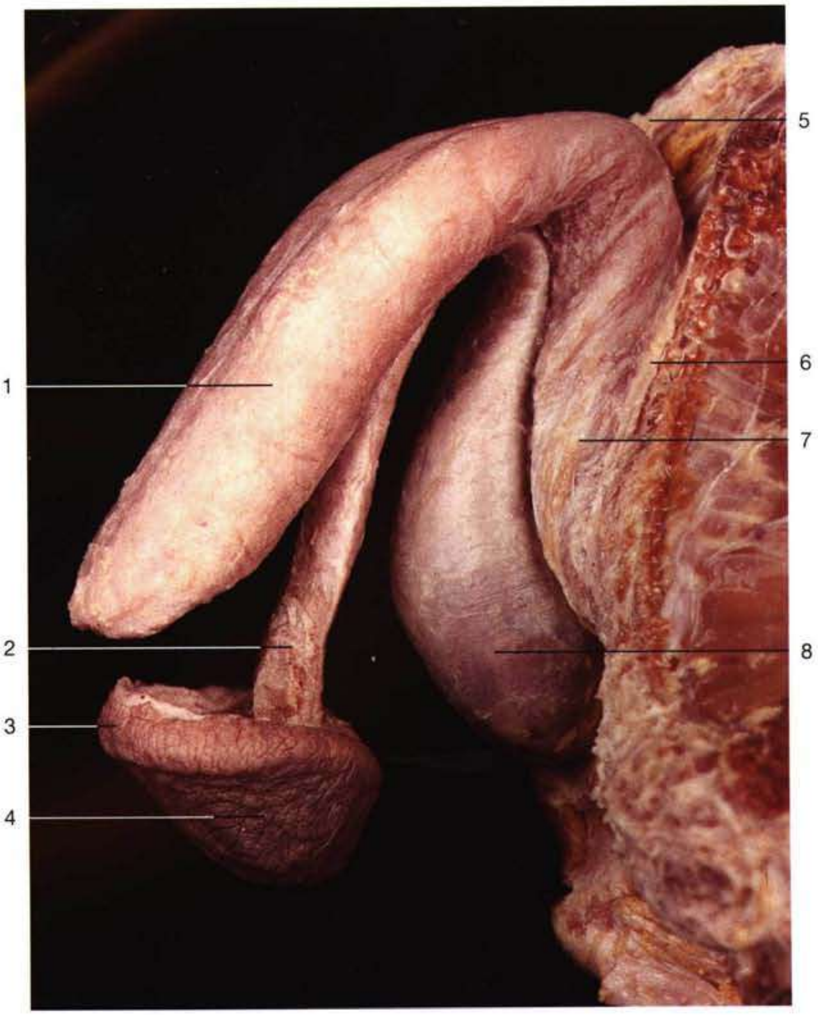

Male external genital organs (lateral view). The corpus spongiosum of the penis with the glans penis has been isolated and reflected.

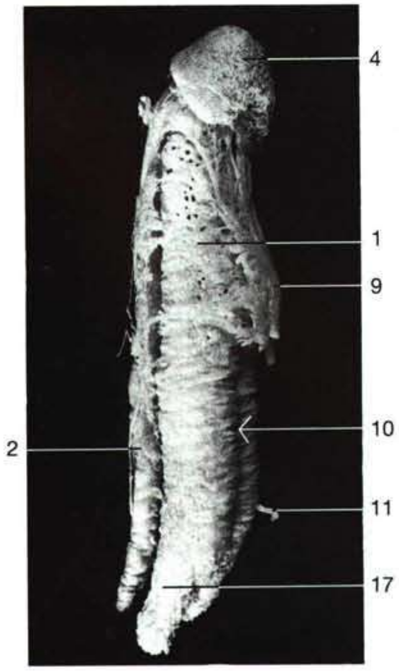

Resin cast of erected penis.

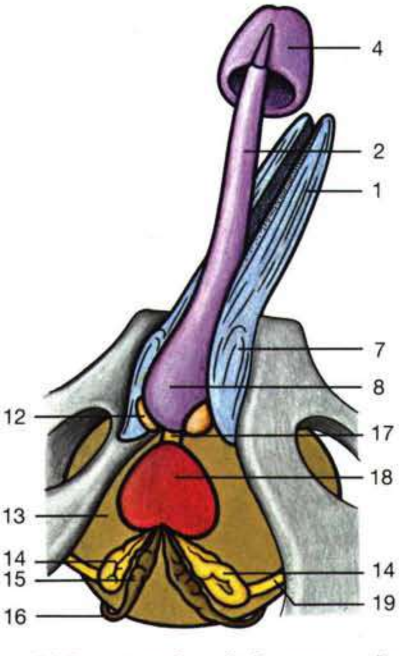

Male external genital organs and accessory glands. (Schematic drawing.)

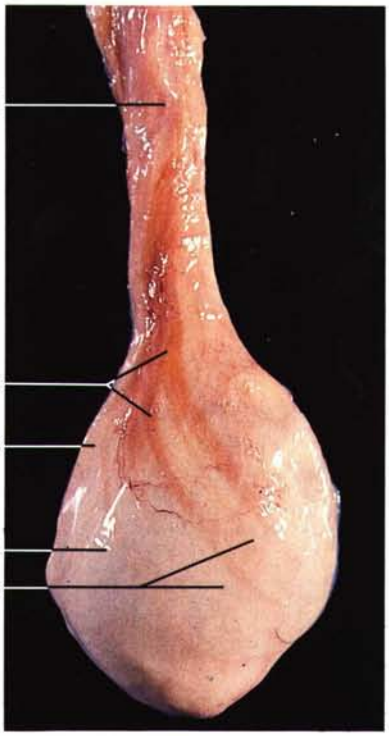

Testis and epididymis with investing layers (lateral view).

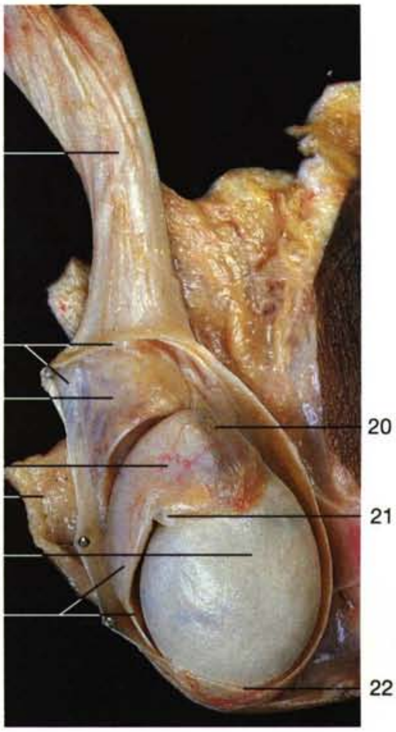

Testis and epididymis (lateral view). The tunica vaginalis has been opened.

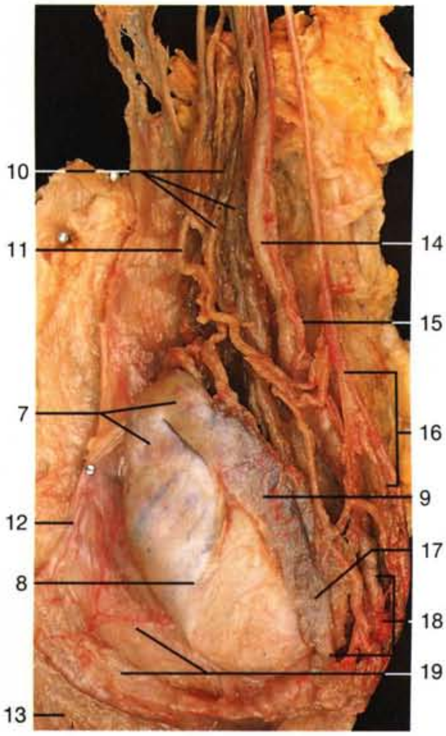

Testis, epididymis and spermatic cord. Dissection of spermatic cord and ductus deferens (left side, posterolateral aspect).

1. Spermatic cord covered with cremasteric fascia
2. **Cremaster muscle**
3. Position of epididymis
4. Internal spermatic fascia
5. Position of testis
6. Internal spermatic fascia with adjacent investing layers of testis (cut surface)
7. **Head of epididymis**
8. **Testis with tunica vaginalis** (visceral layer)
9. Body of epididymis
10. Pampiniform venous plexus (anterior veins)
11. **Testicular artery**
12. Tunica vaginalis (parietal layer, cut edge)
13. Skin and dartos muscle (reflected)
14. **Ductus deferens**
15. Artery of ductus deferens
16. Posterior veins of pampiniform plexus
17. Tail of epididymis
18. Transition of epididymal duct to ductus deferens and venous plexus
19. Parietal layer of tunica vaginalis
20. Appendix of epididymis
21. Appendix of testis
22. Gubernaculum testis

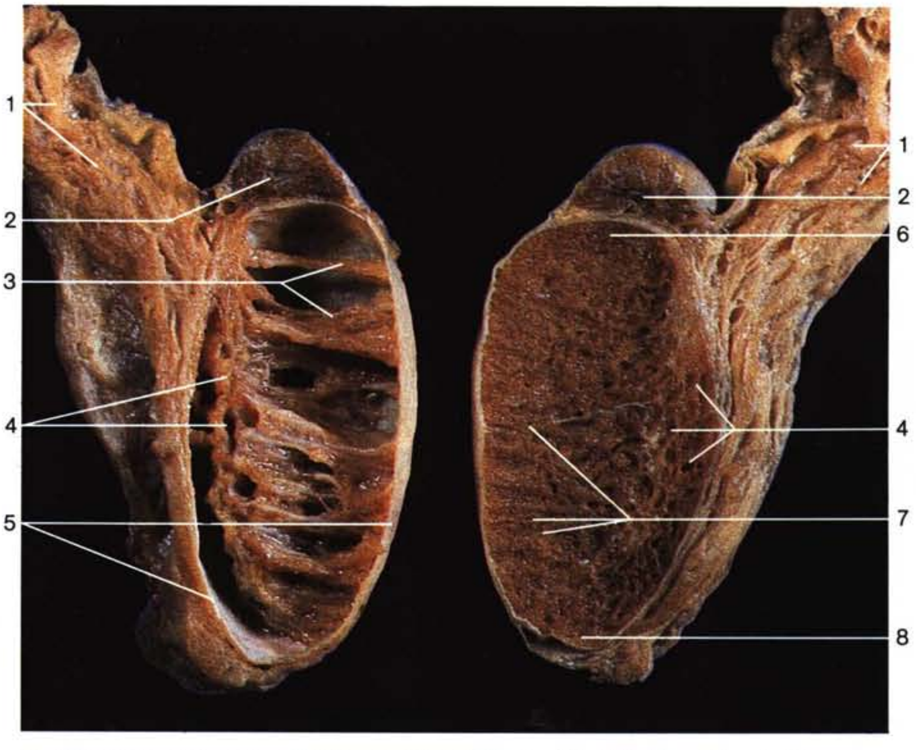

Longitudinal section through testis and epididymis. The left figure shows the testicular septa after removal of the seminiferous tubules.

1. Spermatic cord (cut surface)
2. Head of epididymis (cut surface)
3. Septa of testis
4. Mediastinum testis
5. Tunica albuginea
6. Superior pole of testis
7. Convoluted seminiferous tubules
8. Inferior pole of testis

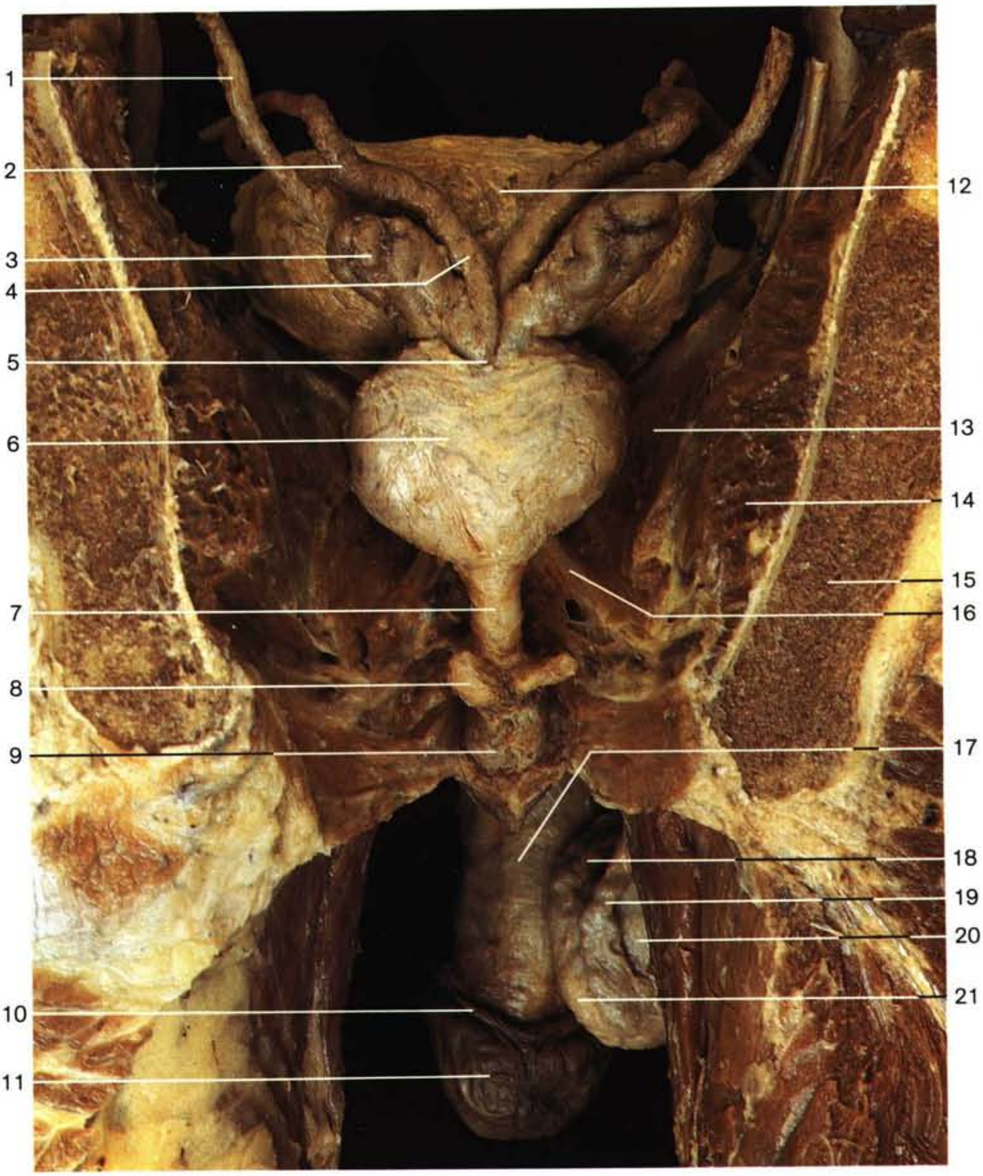

Accessory glands of male genital organs in situ. Coronal section through the pelvic cavity. Posterior aspect of urinary bladder, prostate and seminal vesicles.

1 Ureter
2 **Ductus deferens**
3 **Seminal vesicle**
4 **Ampulla of ductus deferens**
5 Ejaculatory duct (proximal portion)
6 **Prostate**
7 Membranous urethra
8 **Bulbourethral or Cowper's gland**
9 Bulb of penis
10 Penis
11 Glans penis
12 **Urinary bladder**
13 Levator ani muscle
14 Obturator internus muscle
15 Pelvic bone (cut edge)
16 Puboprostatic ligament
17 **Corpus spongiosum of penis**
18 Head of epididymis
19 Beginning of ductus deferens
20 Testis
21 Tail of epididymis
22 Corpus cavernosum of penis
23 Spermatic cord
24 Pectineus and adductor muscles
25 Pubic bone
26 **Prostatic part of urethra** (seminal colliculus)
27 Rectum
28 Sciatic nerve
29 Great saphenous vein
30 Sartorius muscle
31 Femoral artery and vein
32 Rectus femoris muscle
33 Tensor fasciae latae muscle
34 Pectineus muscle
35 Iliopsoas muscle
36 Vastus lateralis muscle
37 Obturator externus muscle
38 Femur
39 Ischial tuberosity
40 Gluteus maximus muscle

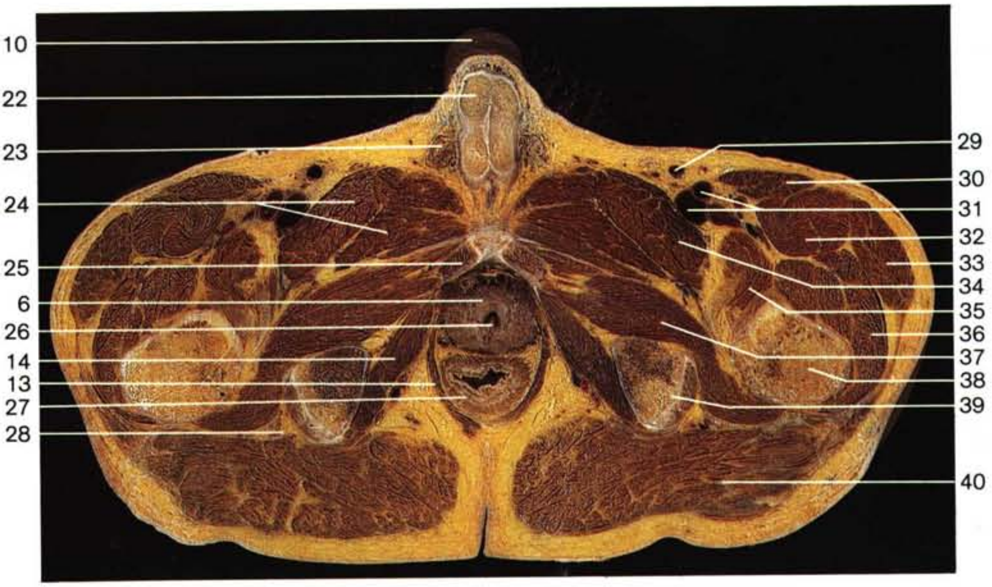

Horizontal section through pelvic cavity at level of prostate.

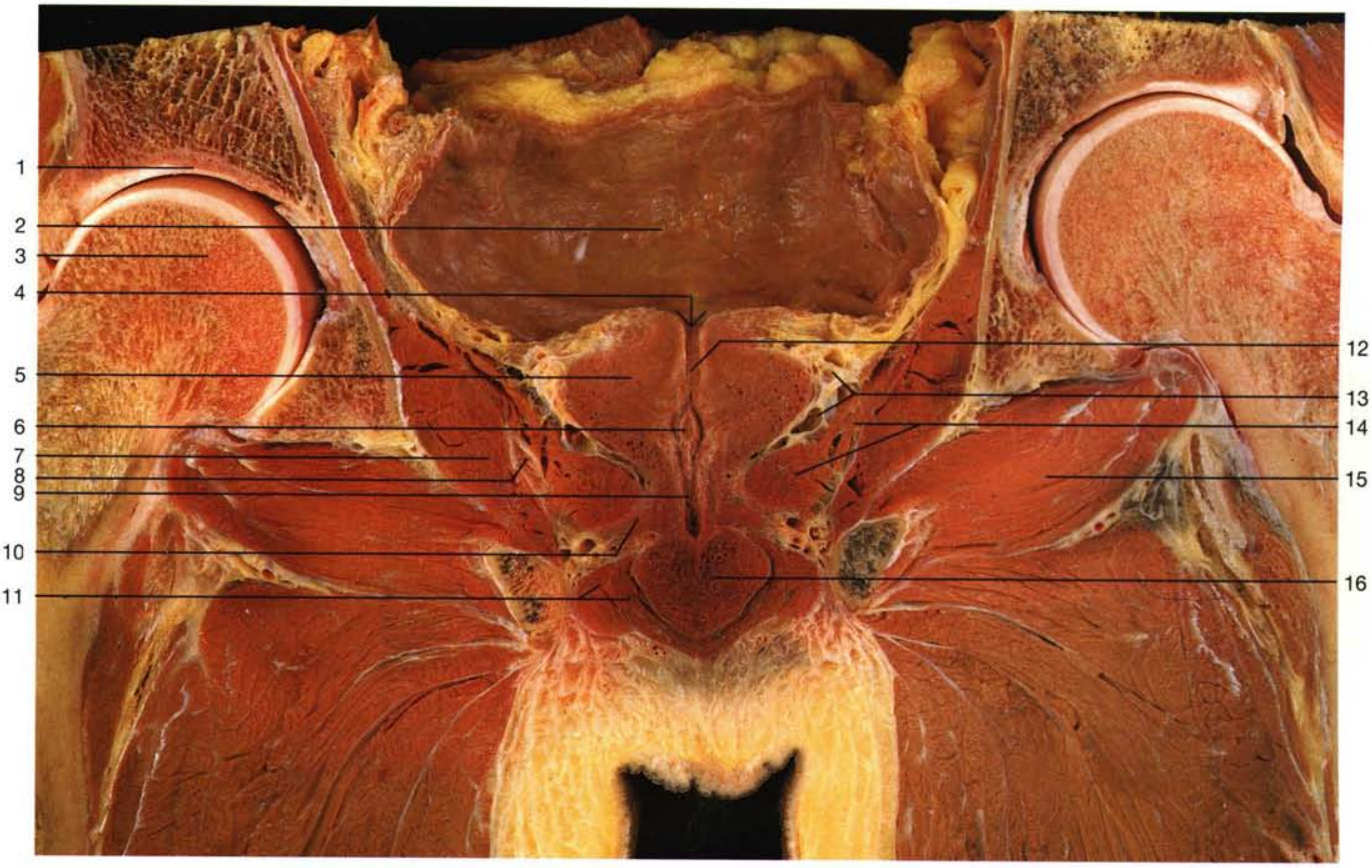

Coronal section through pelvic cavity at the level of prostate and hip joint (anterior aspect).

1 Acetabulum of hip joint
2 **Urinary bladder**
3 Head of femur
4 Internal urethral orifice
5 **Prostate**
6 **Seminal colliculus**
7 Obturator internus muscle
8 Ischiorectal fossa
9 Membranous urethra
10 Deep transverse perineus muscle
11 **Crus penis** and **ischiocavernosus muscle**
12 Prostatic part of urethra
13 Prostatic plexus
14 Levator ani muscle
15 Obturator externus muscle
16 **Bulb of penis**
17 Ampulla of rectum
18 **Anal canal**
19 Seminal vesicle
20 Internal anal sphincter muscle
21 External anal sphincter muscle
22 Anus
23 Psoas major muscle
24 Intervertebral disc
25 Ilium
26 Ligament of the head of the femur
27 Sacral promontory

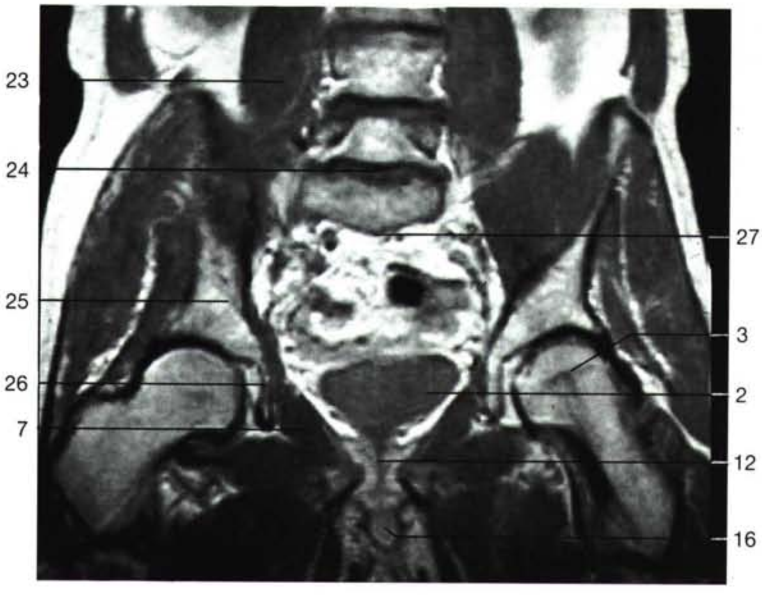

Coronal section through pelvic cavity. (MR-Scan.)

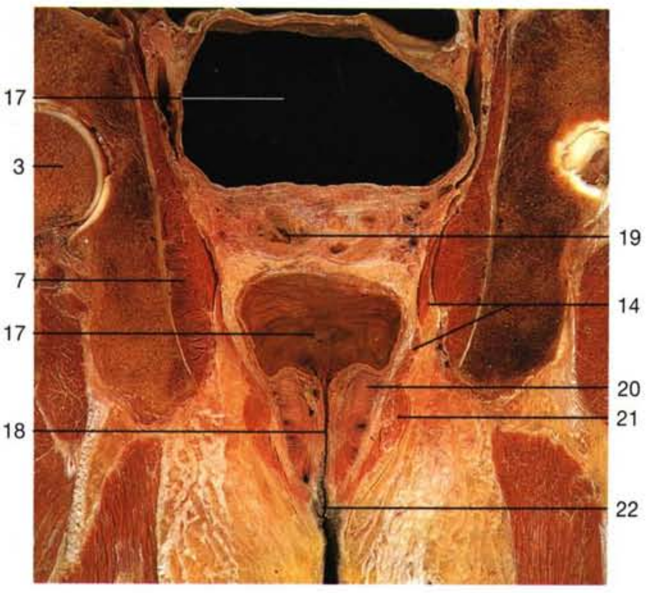

Coronal section through anal canal.

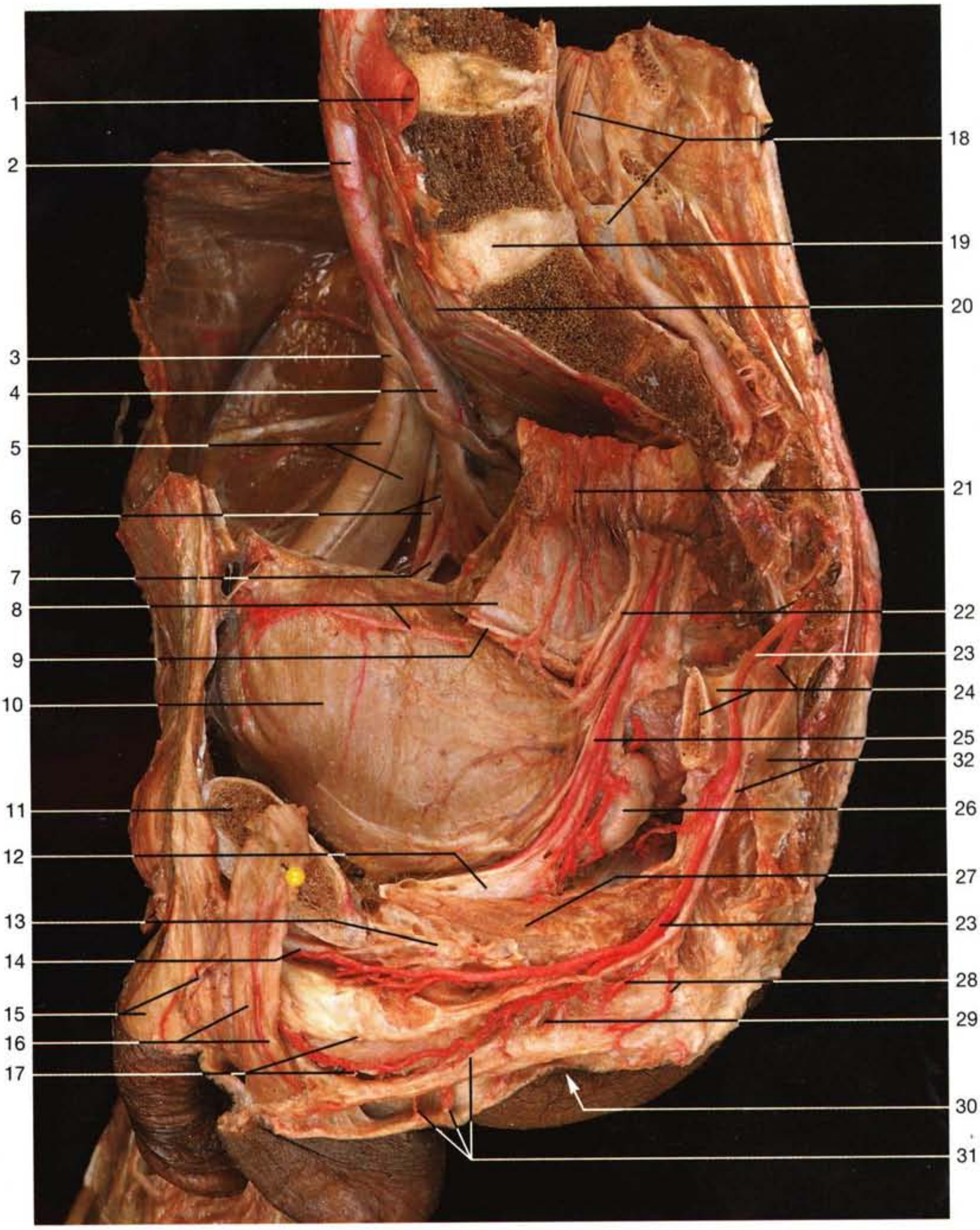

Pelvic cavity in the male (right half of parasagittal section). The arteries have been injected with red resin. The parietal layer of peritoneum has been removed. The urinary bladder is filled to a great extent.

1 Left common iliac artery
2 Right common iliac artery
3 Right **ureter**
4 Right **internal iliac artery**
5 Right external iliac artery and vein
6 Right **obturator artery and nerve**
7 Umbilical artery
8 Sigmoid and superior vesical artery
9 Left ductus deferens
10 **Urinary bladder**
11 Pubic bone (cut)
12 **Prostate**
13 Vesicoprostatic venous plexus
14 Deep dorsal vein of penis and dorsal artery of penis
15 Penis and superficial dorsal vein
16 **Spermatic cord and testicular artery**
17 **Bulb of penis** and deep artery of penis
18 Cauda equina and dura mater (divided)
19 Intervertebral disc between fifth lumbar vertebra and sacrum
20 **Sacral promontory**
21 Mesosigmoid
22 Left ureter
23 Left internal pudendal artery
24 Ischial spine (cut), sacrospinal ligament, inferior gluteal artery
25 Left **inferior vesical artery**
26 **Seminal vesicle**
27 Levator ani muscle
28 Branches of inferior rectal artery
29 Perineal artery
30 Anus
31 Posterior scrotal branches
32 **Pudendal nerve** and sacrotuberal ligament

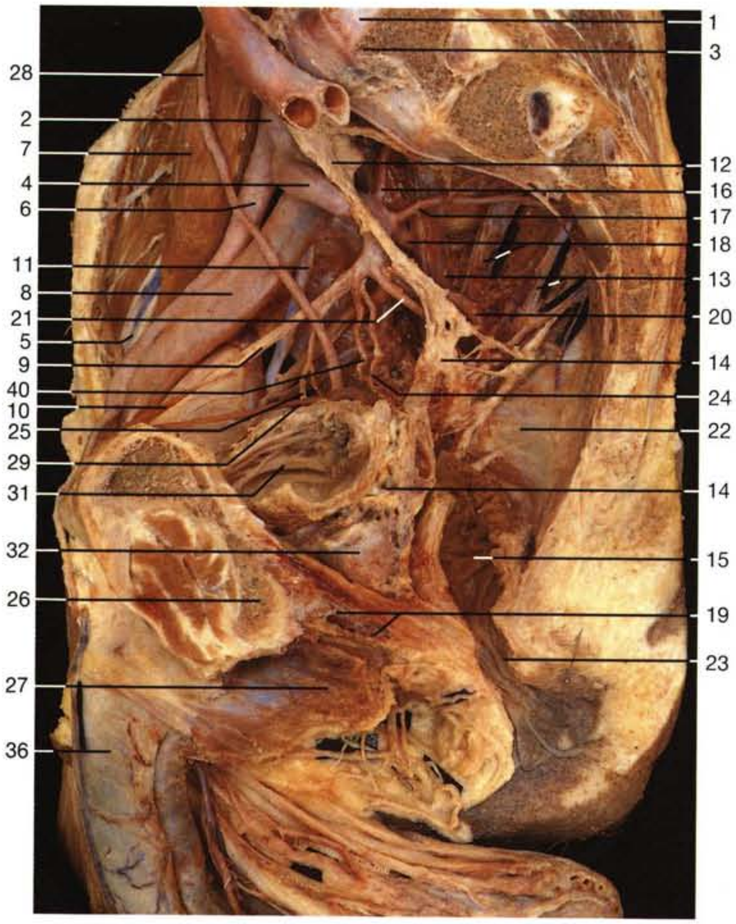

Vessels of pelvic cavity in the male (midsagittal section, right side, medial aspect). Urinary bladder and rectum have been partly removed; the hypogastric plexus is displayed.

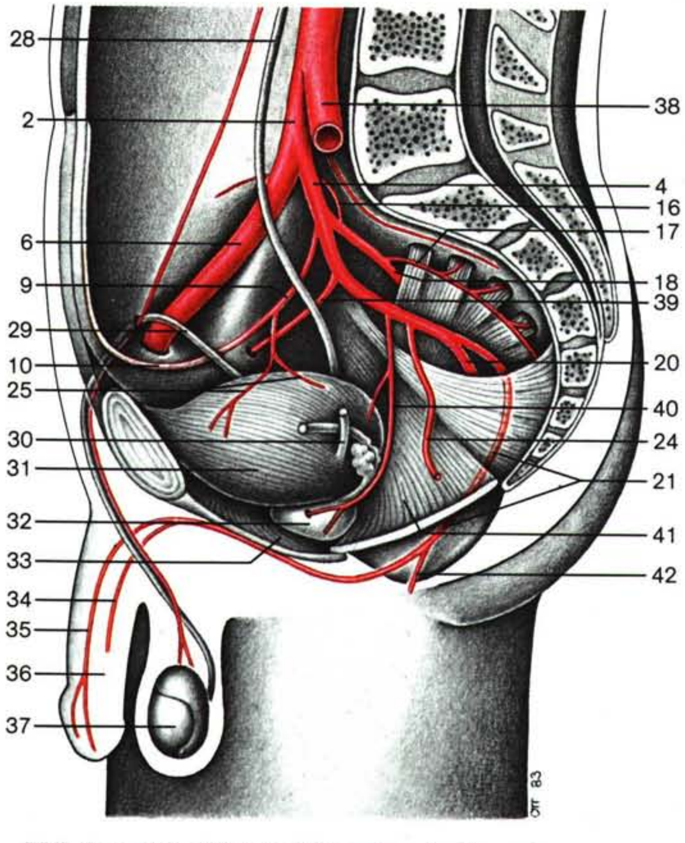

Main branches of internal iliac artery in the male. (Schematic drawing.)

1 Intervertebral disc between 5th lumbar vertebra and sacrum
2 Right common iliac artery
3 Sacral promontory
4 Right **internal iliac artery**
5 Femoral nerve
6 Right external iliac artery
7 Psoas major muscle
8 Right external iliac vein
9 **Umbilical artery**
10 Medial umbilical ligament with obliterated umbilical artery
11 Obturator nerve
12 Superior hypogastric plexus
13 Sacral spinal nerve (of sacral plexus)
14 **Inferior hypogastric plexus**
15 Rectum
16 Iliolumbar artery
17 Lateral sacral artery
18 Superior gluteal artery and sacral spinal nerve
19 Pudendal and coccygeal plexus
20 Inferior gluteal artery
21 **Internal pudendal artery**
22 Coccygeus muscle overlying sacrospinous ligament
23 **Anal canal**
24 Middle rectal artery
25 **Superior vesical artery** and branch to the ductus deferens
26 Pubic bone (cut)
27 Bulbospongiosus muscle
28 Right ureter
29 Right ductus deferens
30 Left ureter
31 **Urinary bladder**
32 **Prostate**
33 Urogenital diaphragm
34 Deep artery of penis
35 Dorsal artery of penis
36 Penis
37 Testis
38 Left common iliac artery
39 **Obturator artery**
40 **Inferior vesical artery**
41 Levator ani muscle
42 Inferior rectal artery

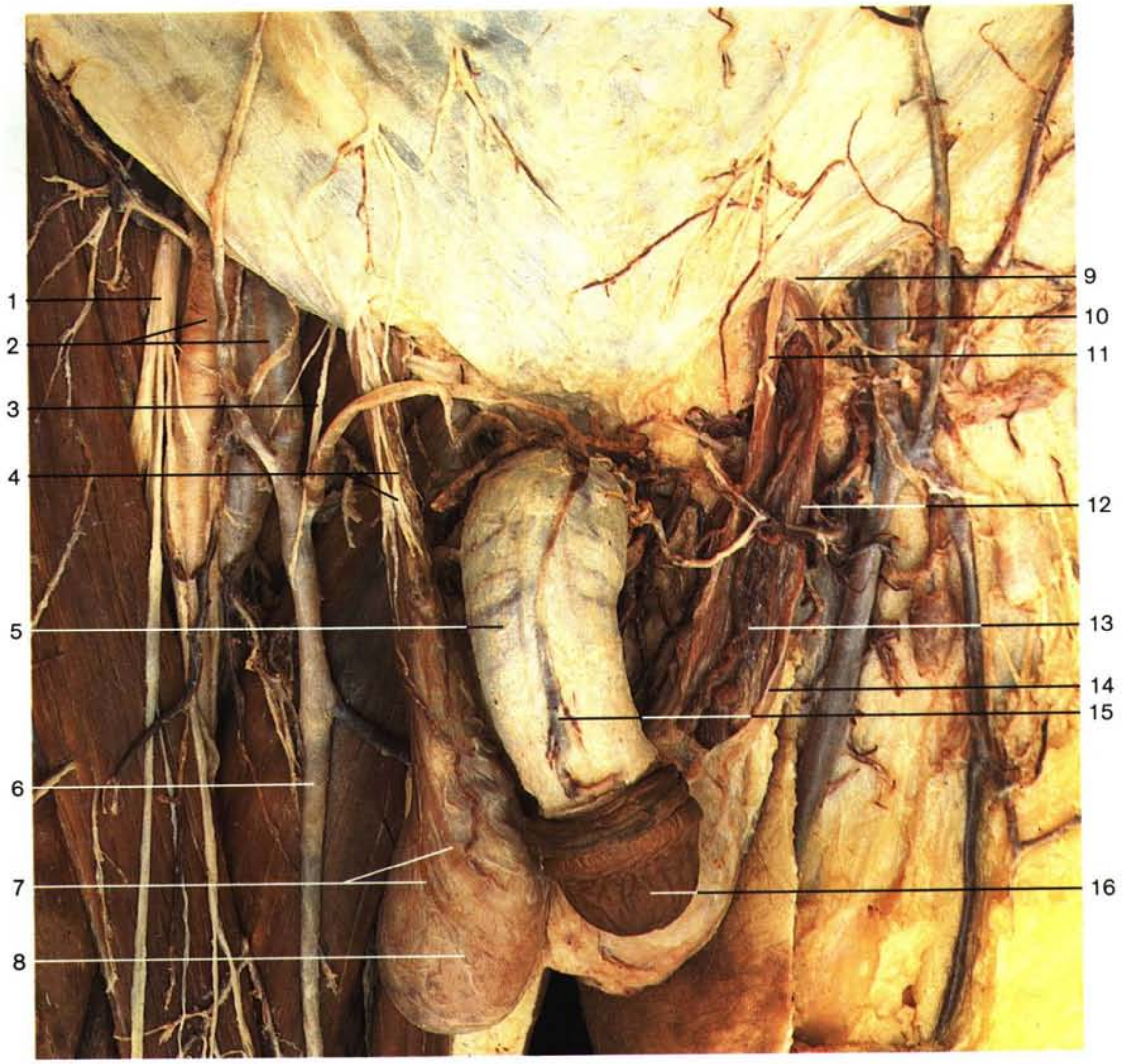
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Male external genital organs with penis, testis, and spermatic cord, superficial layers (anterior view).

1 Femoral nerve
2 Femoral artery and vein
3 Femoral branch of genitofemoral nerve
4 **Spermatic cord** with genital branch of genitofemoral nerve
5 **Penis** with deep fascia
6 Great saphenous vein
7 Cremaster muscle
8 **Testis with cremaster muscle**
9 Superficial inguinal ring
10 Internal spermatic fascia (cut edge)
11 Ilioinguinal nerve
12 Left spermatic cord
13 Pampiniform venous plexus
14 External spermatic fascia
15 Superficial dorsal vein of penis
16 **Glans penis**

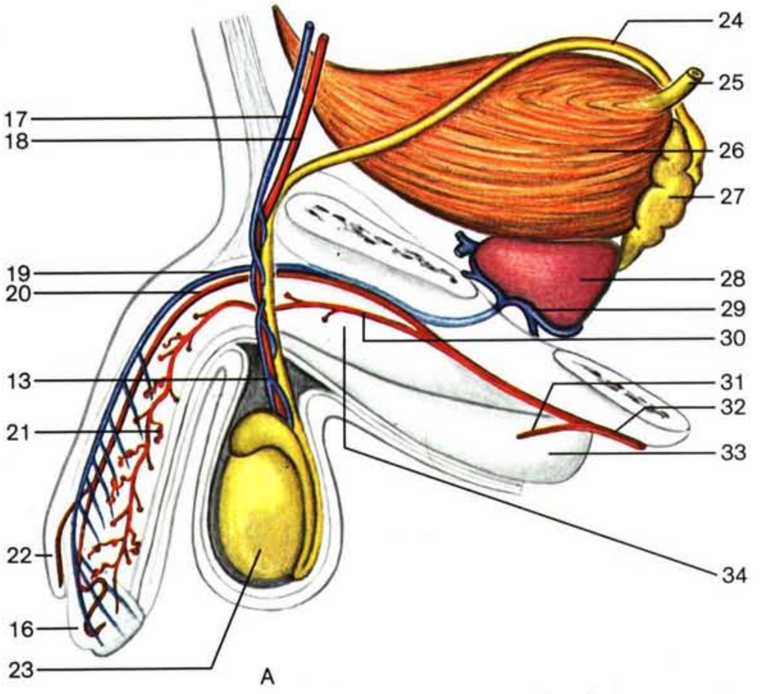
24
25
17
18
26
27
19
20
28
29
30
13
31
32
21
33
22
34
16
23
A

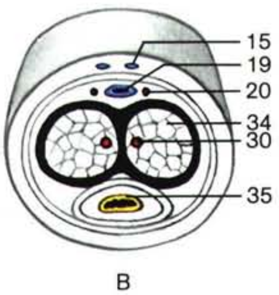
15
19
20
34
30
35
B

Vessels of male genital organs. (Schematic drawing.)
A = lateral aspect; B = cross-section of penis.

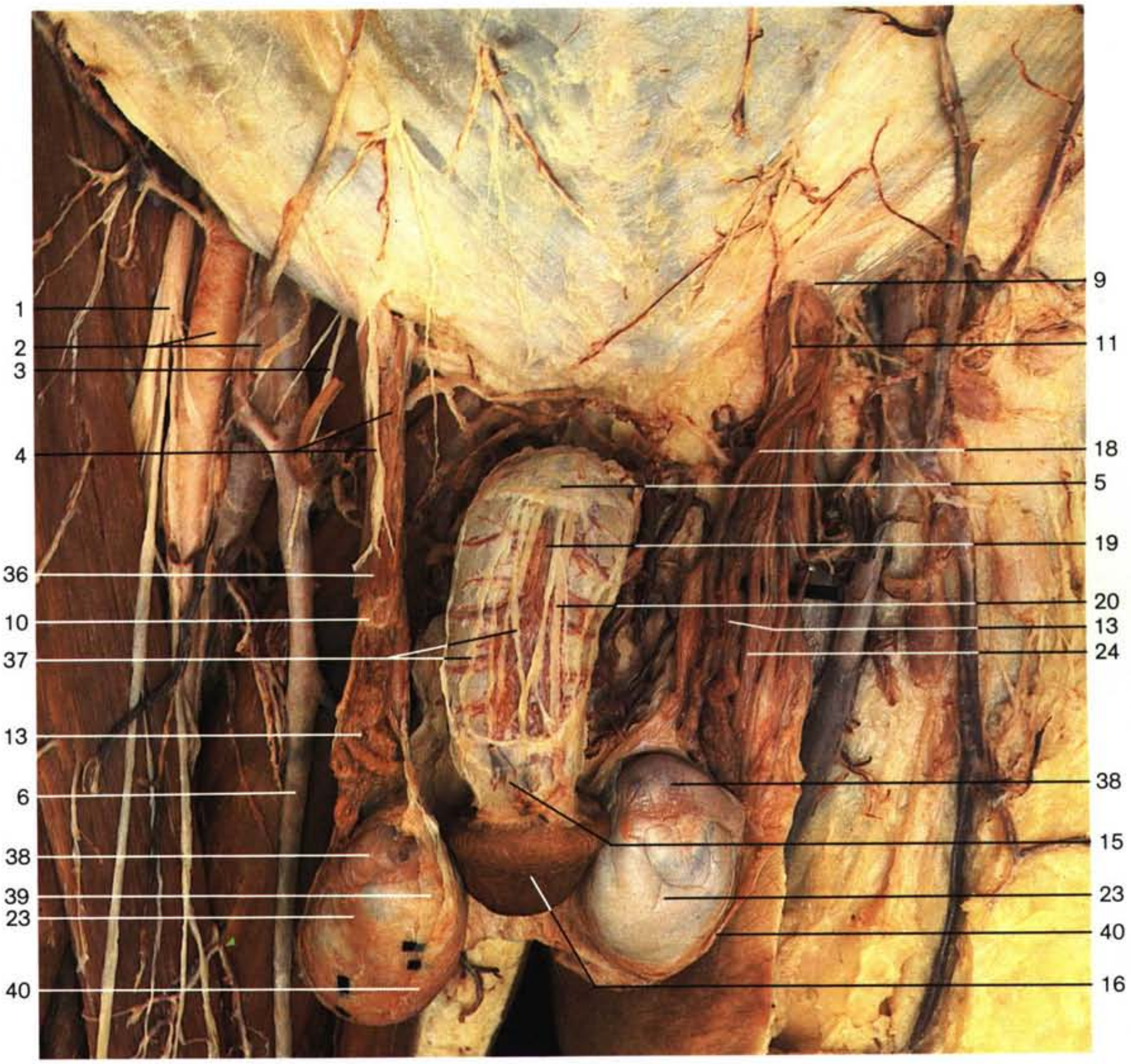

Male external genital organs with penis, testis and spermatic cord, deeper layers (ventral aspect). The deep fascia of the penis has been opened to display the dorsal nerves and vessels.

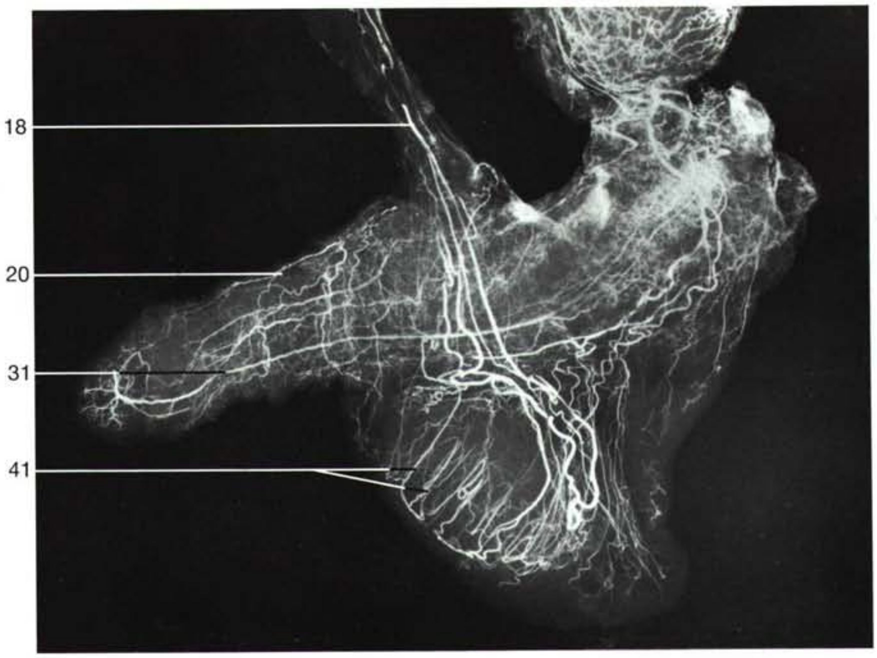

Male genital organs (lateral aspect). (Arteriogram.)

17 Testicular vein
18 Testicular artery
19 Deep dorsal vein of penis
20 **Dorsal artery of penis**
21 **Helicine arteries**
22 Prepuce
23 **Testis** with **tunica albuginea**
24 Ductus deferens
25 Ureter
26 Urinary bladder
27 **Seminal vesicle**
28 **Prostate**
29 Vesicoprostatic venous plexus
30 **Deep artery of penis**
31 Artery of bulb of penis
32 Internal pudendal artery
33 Corpus spongiosum of penis
34 Corpus cavernosum of penis
35 Urethra
36 Cremasteric fascia with cremaster muscle
37 Dorsal nerve of penis
38 **Epididymis**
39 **Tunica vaginalis** (visceral layer)
40 Tunica vaginalis (parietal layer)
41 Testis with vascular loops

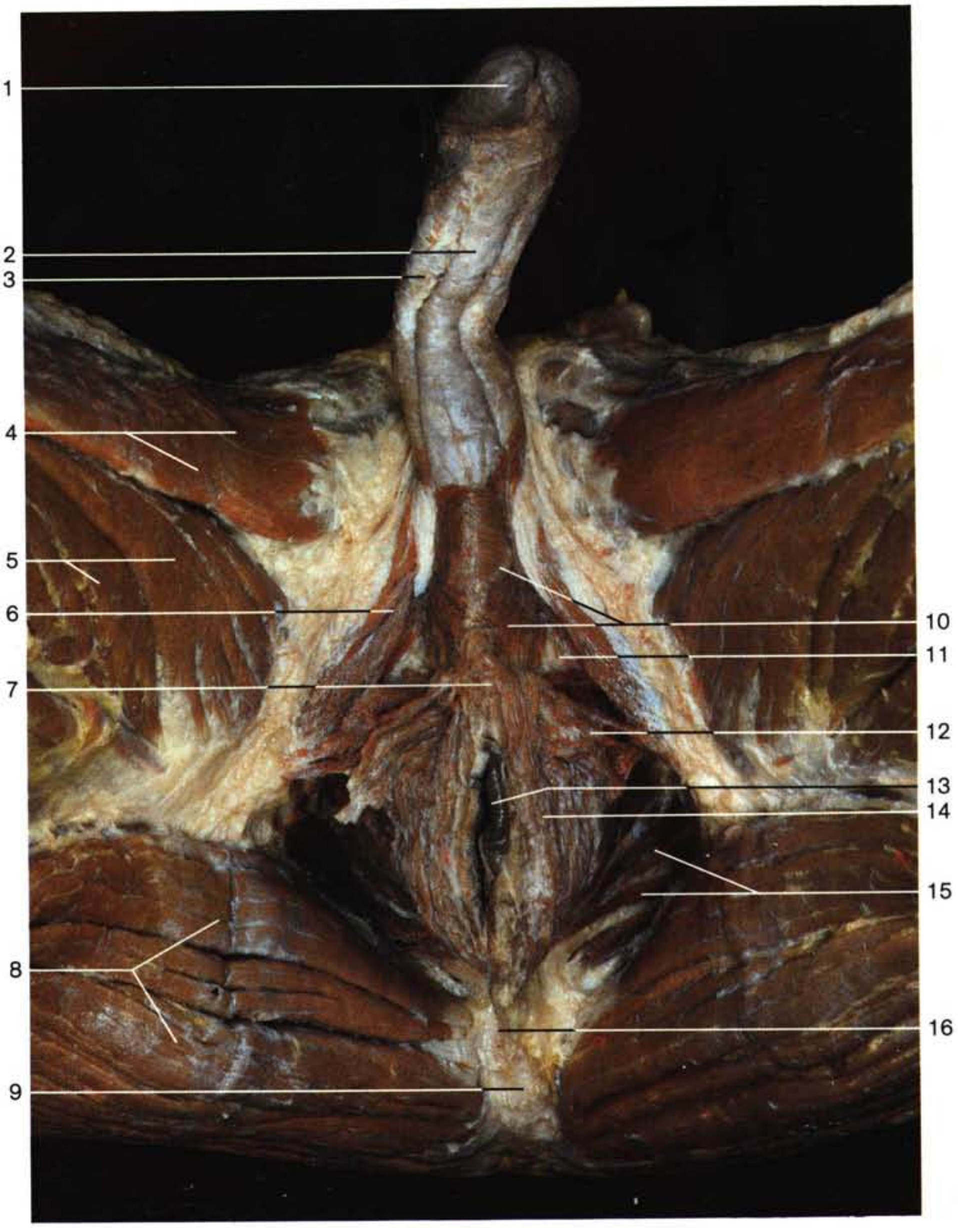

1. Glans penis
2. **Corpus spongiosum of penis**
3. **Corpus cavernosum of penis**
4. Gracilis muscle
5. Adductor muscles
6. Ischiocavernosus muscle overlying crus of penis
7. Perineal body
8. Gluteus maximus muscle
9. Coccyx
10. **Bulbospongiosus** muscle
11. Deep transverse perineus muscle covered by inferior fascia of urogenital diaphragm
12. Superficial transverse perineus muscle
13. Anus
14. **External anal sphincter muscle**
15. **Levator ani muscle**
16. Anococcygeal ligament
17. Obturator internus muscle
18. Urethra
19. **Deep transverse perineus muscle**

Muscles of urogenital and pelvic diaphragms in the male (from below).

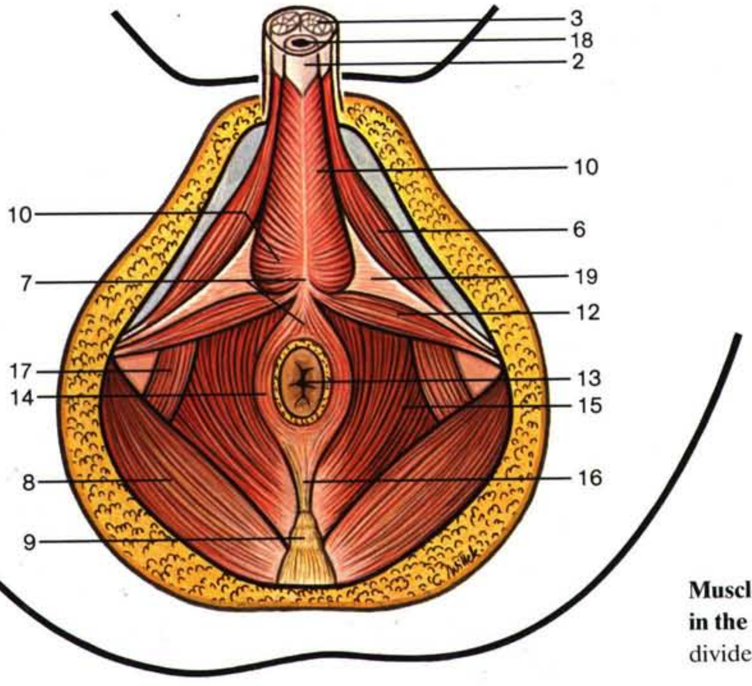

Muscles of urogenital and pelvic diaphragms in the male (from below). The penis has been divided. (Schematic drawing.)

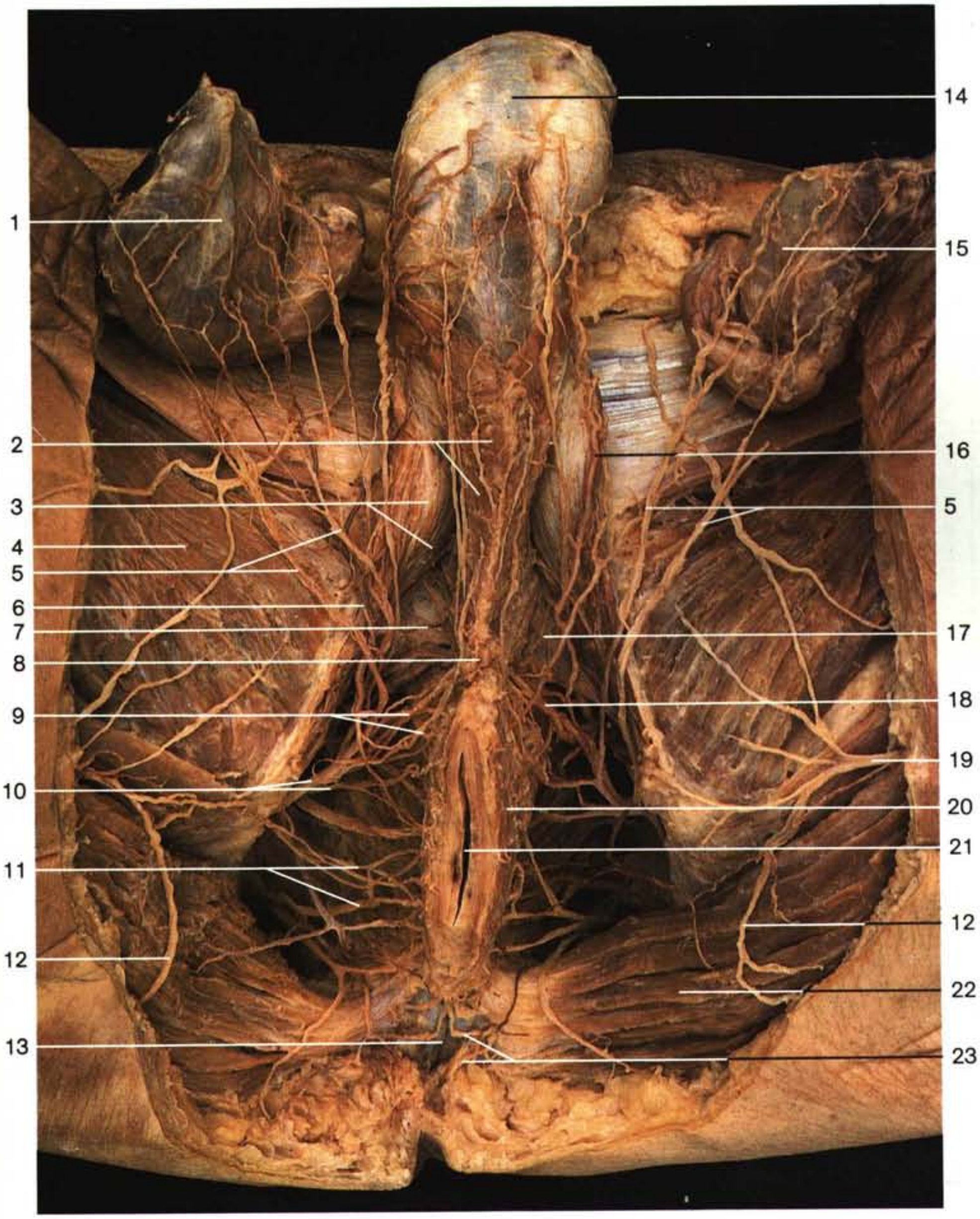

Urogenital diaphragm and external genital organs in the male with vessels and nerves (from below). The testes have been reflected laterally.

1	Right testis (reflected laterally and upward)
2	Bulbospongiosus muscle
3	Ischiocavernosus muscle
4	Adductor magnus muscle
5	**Posterior scrotal nerves** and superficial perineal arteries
6	Posterior scrotal artery and vein
7	Right artery of bulb of penis
8	Perineal body
9	Perineal branches of pudendal nerve
10	**Pudendal nerve** and **internal pudendal artery**
11	**Inferior rectal arteries** and **nerves**
12	Inferior cluneal nerve
13	Coccyx (location)
14	Penis
15	Left testis (reflected laterally)
16	Left posterior scrotal artery
17	**Deep transverse perineal muscle**
18	Left artery of bulb of penis
19	**Posterior femoral cutaneous nerve**
20	External anal sphincter muscle
21	**Anus**
22	Gluteus maximus muscle
23	Anococcygeal nerves
24	Acetabulum (femur removed)
25	Ligament of femoral head
26	Body of ischium (cut)
27	Sciatic nerve
28	Coccygeus muscle
29	**Levator ani muscle**
a	iliococcygeus muscle
b	pubococcygeus muscle
c	puborectalis muscle
30	Prostatic venous plexus
31	Body of pubis
32	Testis

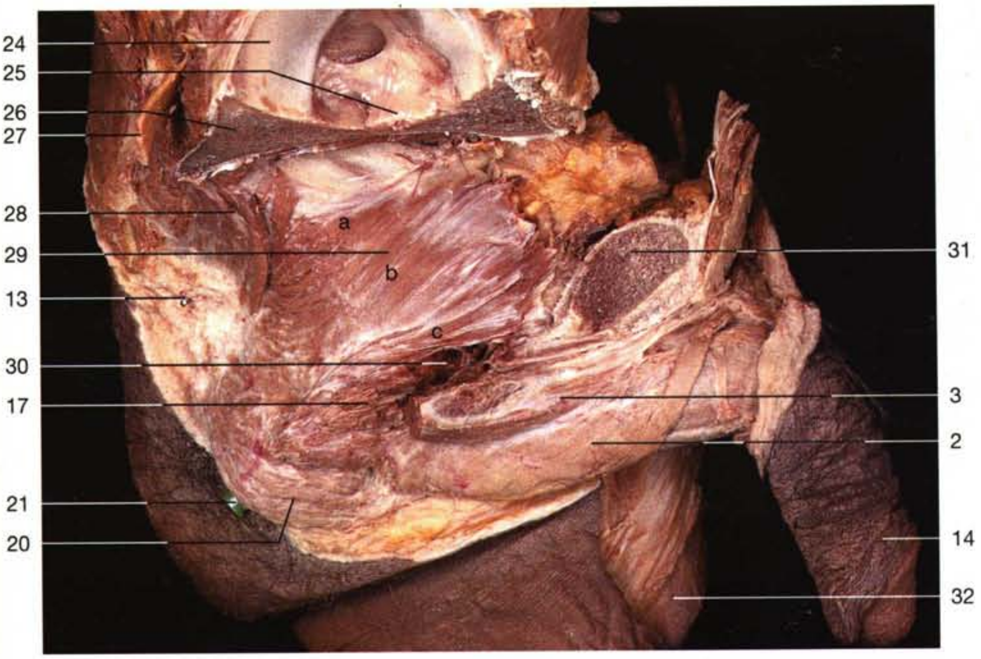

Pelvic diaphragm and external genital organs in the male. The right half of the pelvis including obturator internus muscle and femur had been removed to display the right half of levator ani muscle.

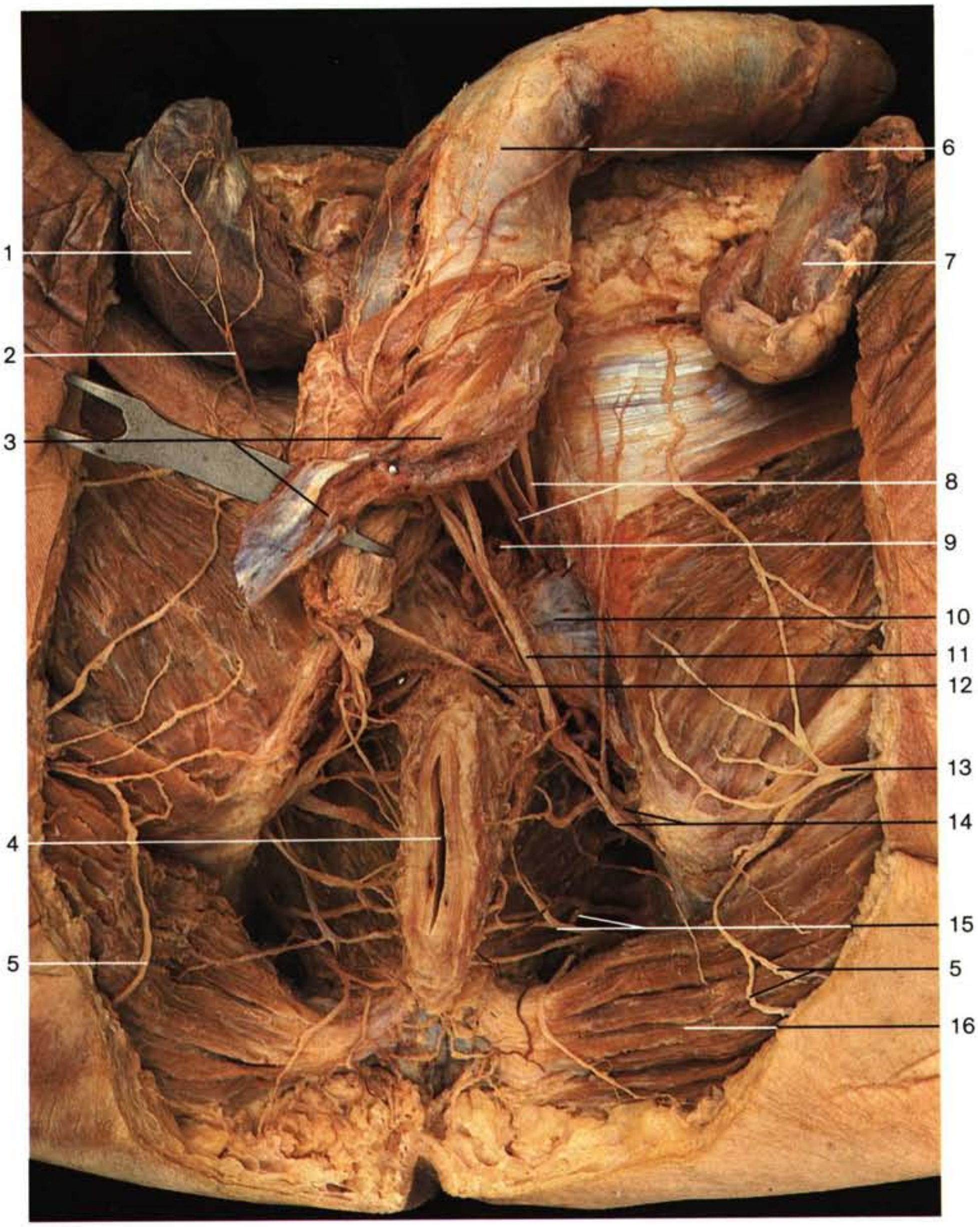

1 Right testis (reflected)
2 Posterior scrotal nerves
3 Left crus penis with ischiocavernosus muscle
4 **Anus**
5 Inferior cluneal nerves
6 **Penis**
7 Left testis (reflected)
8 Dorsal artery and nerve of penis
9 Urethra
10 **Deep transverse perineus muscle**
11 **Perineal branch of pudendal nerve**
12 Artery of bulb of penis (reflected)
13 Branch of posterior femoral cutaneous nerve
14 **Internal pudendal artery** and **pudendal nerve**
15 Inferior rectal arteries and nerves
16 Gluteus maximus muscle
17 Dorsal nerve of penis
18 Posterior femoral cutaneous nerve
19 Perineal branches of pudendal nerve
20 Inferior rectal nerves
21 **Bulbocavernosus muscle** (inside: dorsal artery of penis)
22 Perineal artery
23 **External anal sphincter muscle**
24 Inferior rectal artery and veins

Urogenital diaphragm and external genital organs in the male (from below).
The left crus penis has been isolated and reflected laterally together with the bulb of the penis. The urethra has been cut.

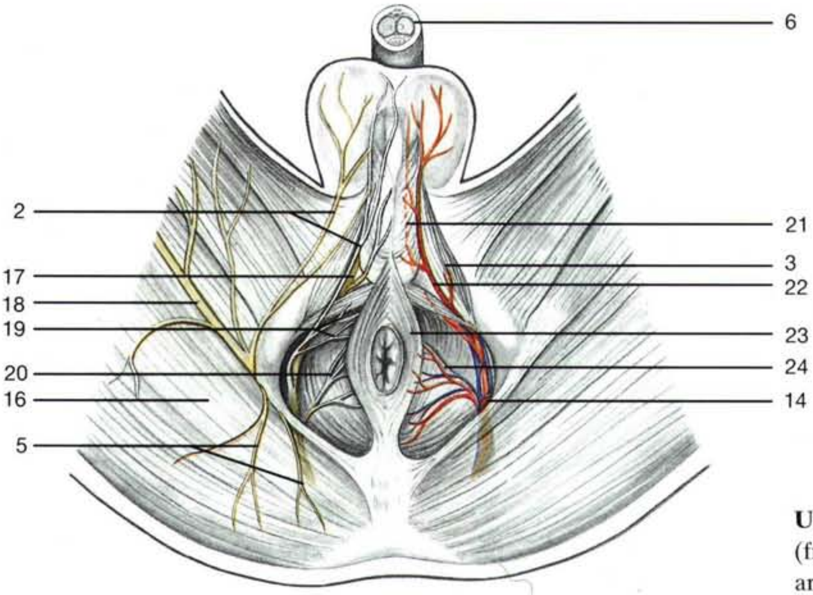

Urogenital and anal region in the male (from below). Right side: nerves; left side: arteries and veins.

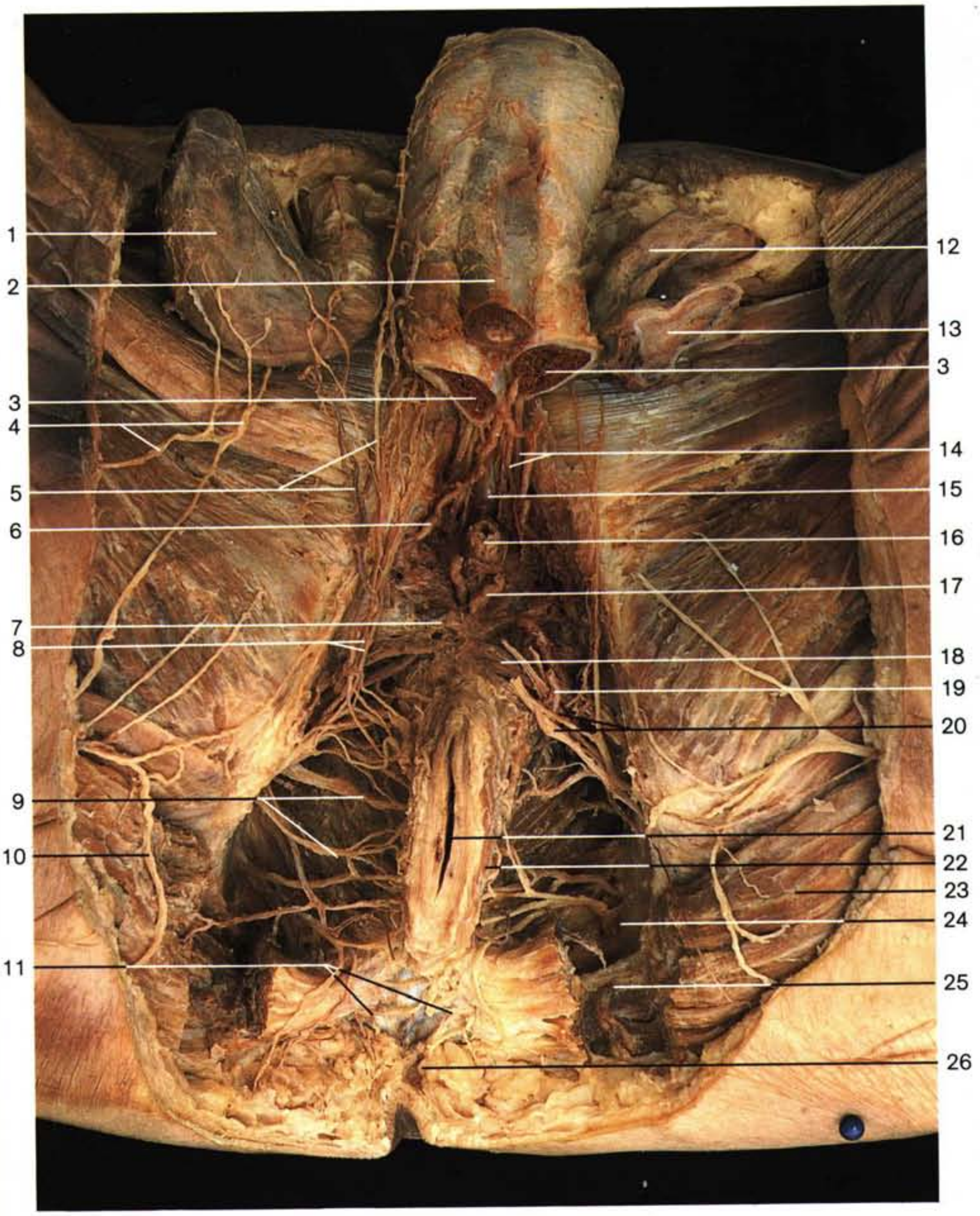

1 Right testis (reflected)
2 **Corpus spongiosum of penis**
3 **Corpus cavernosum of penis**
4 Perineal branch of posterior femoral cutaneous nerve
5 Posterior scrotal arteries and nerves
6 Deep artery of penis
7 Deep transverse perineal muscle
8 Right **perineal nerves**
9 **Inferior rectal nerves**
10 Inferior cluneal nerve
11 Anococcygeal nerves
12 Left spermatic cord
13 Left testis (cut surface)
14 Dorsal artery and nerve of penis
15 Deep dorsal vein of penis
16 **Urethra** (cut)
17 Artery of bulb of penis
18 Superficial transverse perineus muscle
19 Left artery of bulb of penis
20 Perineal branch of pudendal nerve
21 **Anus**
22 **External anal sphincter muscle**
23 Gluteus maximus muscle
24 Internal pudendal artery and pudendal nerve
25 Sacrotuberous ligament
26 Coccyx
27 Urogenital diaphragm (deep transverse perineus muscle)
28 Tendinous center of perineum (perineal body)
29 **Levator ani muscle**
30 Anococcygeal ligament
31 Obturator internus muscle
32 Dorsal artery of penis

Anal and urogenital region in the male (from below). The root of penis has been cut. Dissection of the urogenital diaphragm.

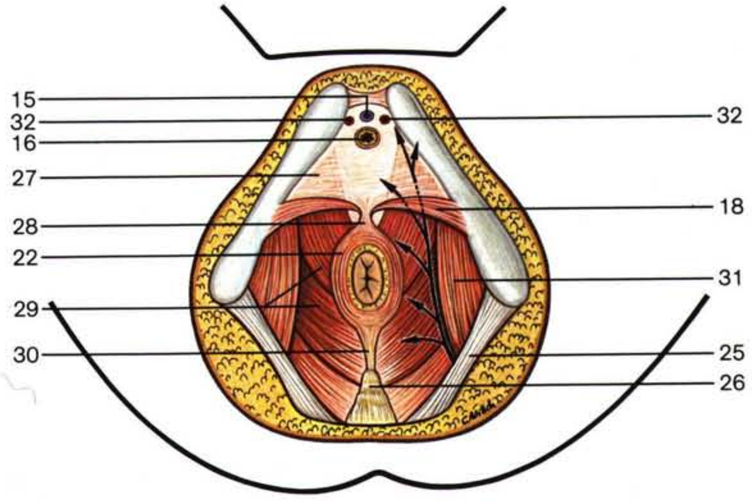

Urogenital and pelvic diaphragms in the male (from below). The penis has been removed. The arrows indicate the course of vessels and nerves. (Schematic drawing.)

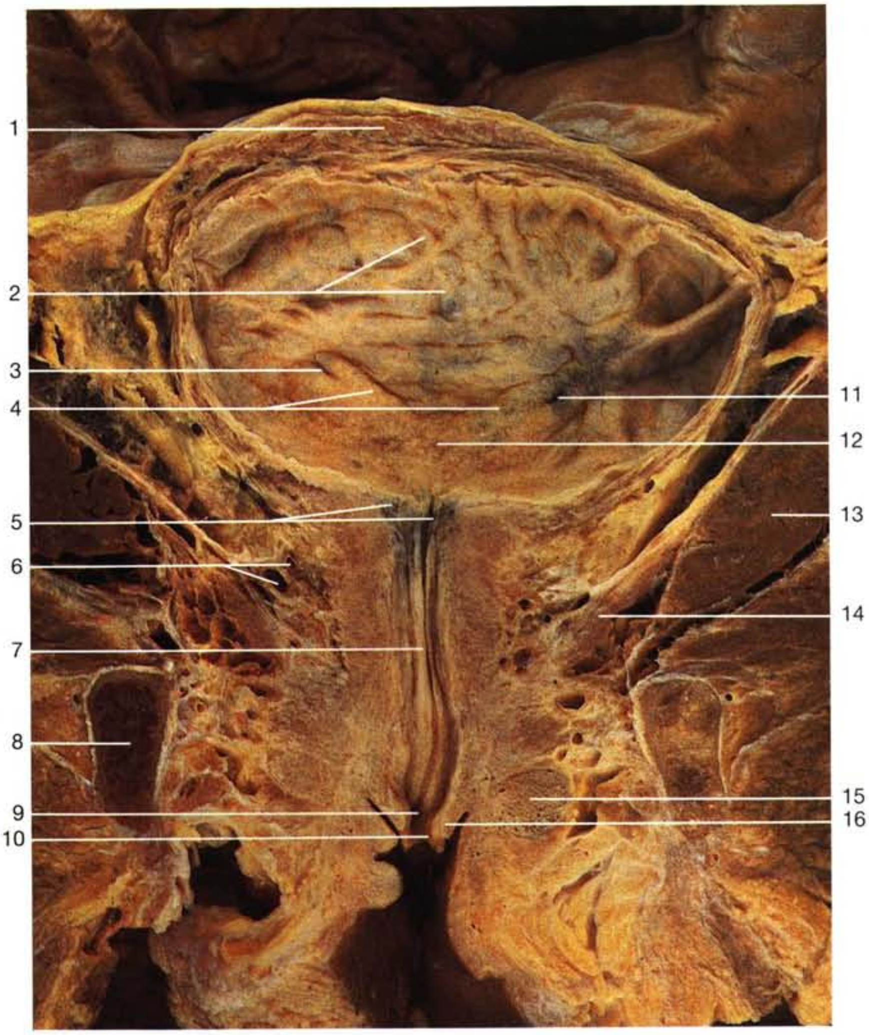

Coronal section through the female urinary bladder and urethra (anterior view).

1 Muscular coat of **urinary bladder**
2 Folds of mucous membrane of urinary bladder
3 Right ureteric orifice
4 Interureteric fold
5 Internal urethral orifice
6 Vesicouterine venous plexus
7 **Urethra**
8 Pubic bone (cut edge)
9 External urethral orifice
10 Vestibule of vagina
11 Left ureteric orifice
12 Trigone of bladder
13 Obturator internus muscle
14 Levator ani muscle
15 Bulb of the vestibule
16 Left labium minus
17 **Uterine tube**
18 Mesosalpinx
19 **Ovary**
20 Sigmoid colon
21 Saphenous opening
22 Urinary bladder
23 Vesicouterine pouch
24 **Fundus of uterus**
25 Rectouterine pouch (of Douglas)
26 Ampulla of rectum
27 Kidney
28 Abdominal part of ureter
29 Pelvic part of ureter
30 Anal canal
31 Perineum (perineal body)
32 Umbilicus
33 Infundibulum of uterine tube
34 Vaginal portion of cervix of uterus
35 Vagina
36 Pubic symphysis
37 Clitoris
38 Deep transverse perineus muscle

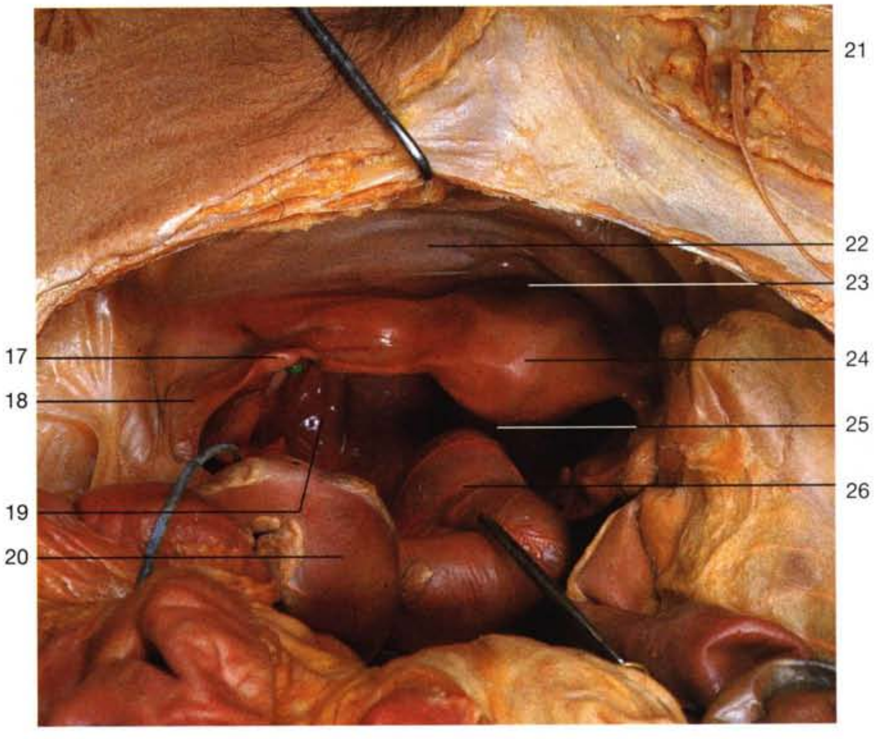

Femal internal genital organs. Pelvic cavity (from above).

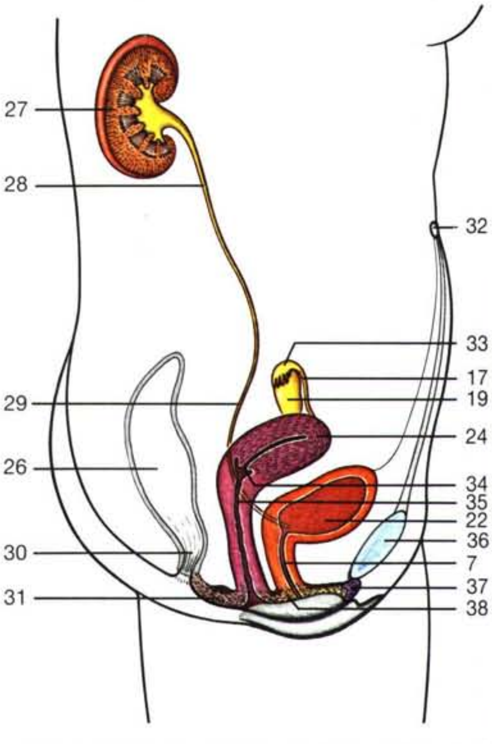

Female urogenital system (midsagittal section). (Schematic drawing.)

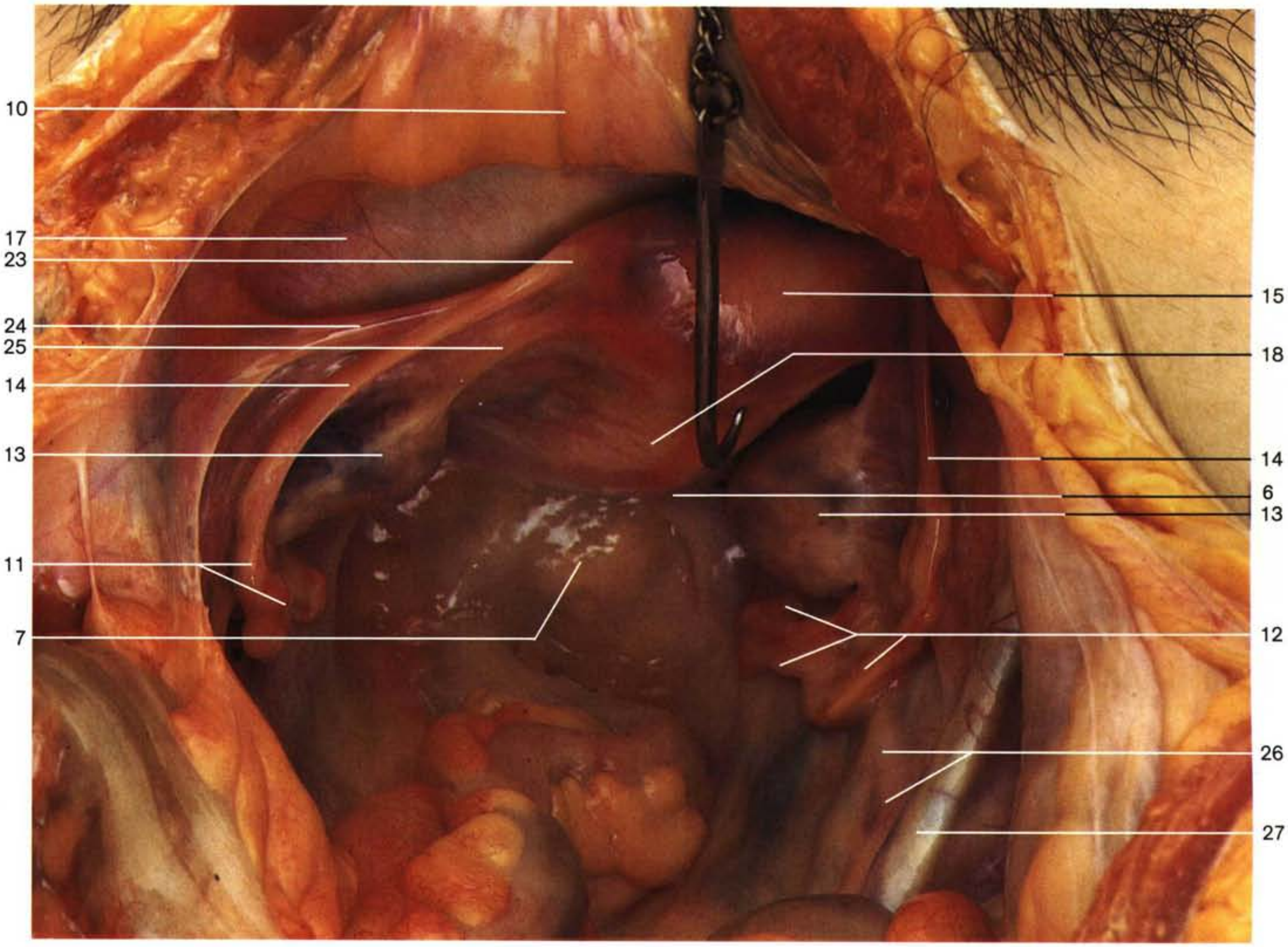

Female internal genital organs. Pelvic cavity, seen from above. The uterus has been reflected to the right.

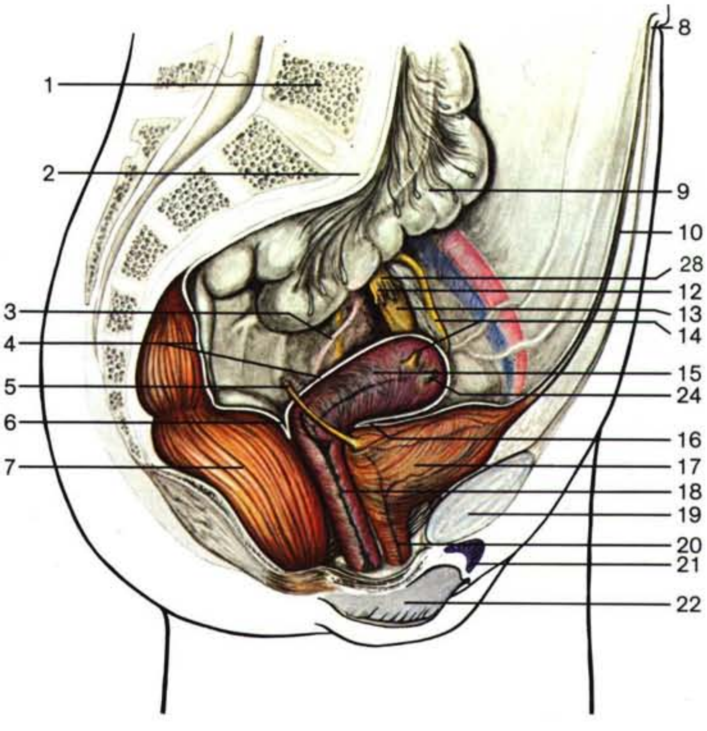

Regional relations of female internal genital organs (medial aspect). (Schematic drawing.)

1 Body of 5th lumbar vertebra
2 Sacral promontory
3 Left ureter
4 Peritoneum (cut edge)
5 Right ureter (divided)
6 **Recto-uterine pouch** (of Douglas)
7 Rectum
8 Umbilicus
9 Sigmoid colon
10 Median umbilical fold with urachus
11 Ampulla of uterine tube
12 Fimbriae of uterine tube
13 **Ovary**
14 **Uterine tube** (isthmus)
15 **Uterus**
16 **Vesico-uterine pouch**
17 Urinary bladder
18 Vagina
19 Pubic symphysis
20 Urethra
21 Clitoris
22 Labium minus
23 Insertion of uterine tube at fundus of uterus
24 Round ligament of uterus
25 Ligament of the ovary
26 Suspensory ligament of ovary
27 Right common iliac artery (covered by peritoneum)
28 Infundibulum of uterine tube

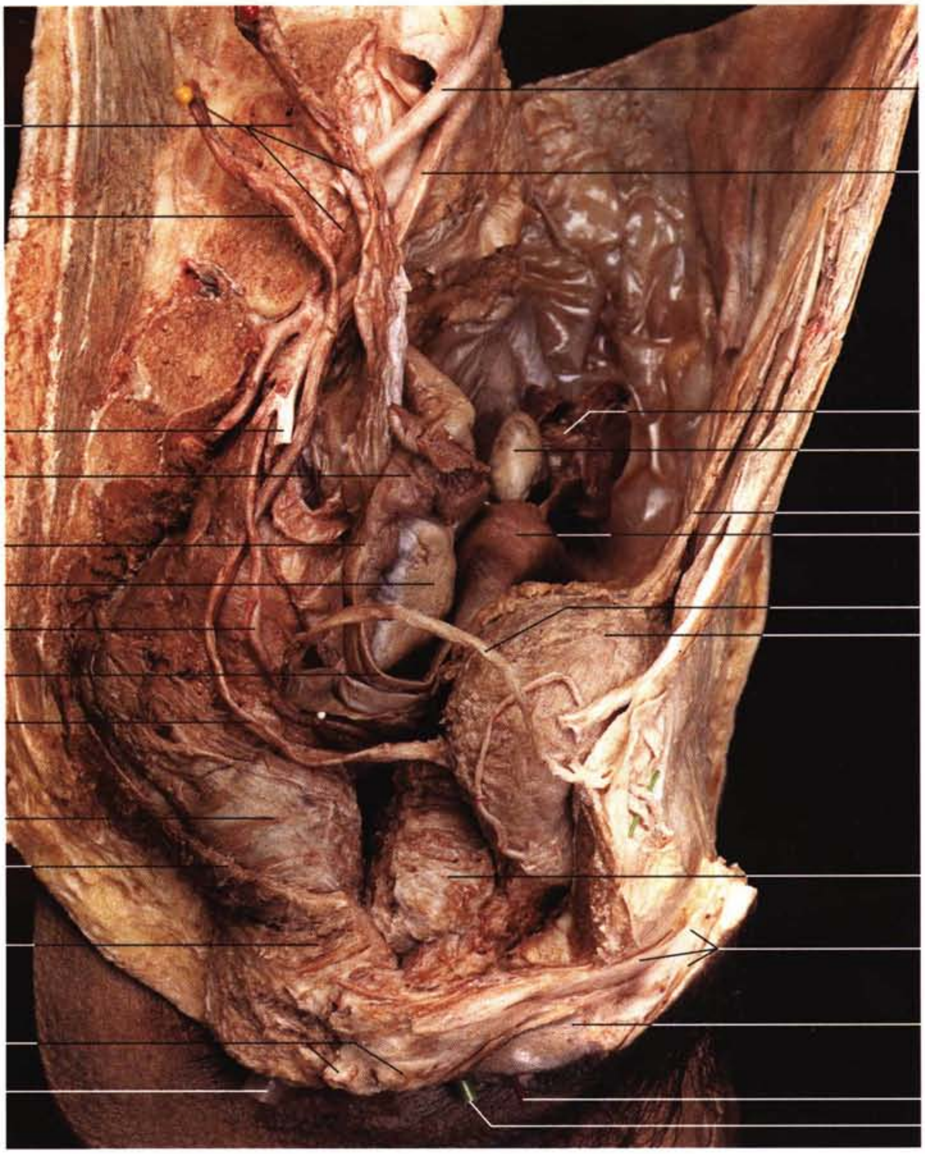

Female internal genital organs in situ. Right half of the pelvis and sacrum have been removed.

1 Body of 5th lumbar vertebra, **suspensory ligament of ovary** and sacral promontory
2 Ureter
3 Medial umbilical ligament (remnant of umbilical artery) (cut)
4 Infundibulum of uterine tube
5 Ampulla of uterine tube
6 **Ovary**
7 **Uterine artery**
8 **Uterine tube**
9 Rectum
10 Levator ani muscle (pelvic diaphragm – cut edge)
11 External anal sphincter muscle
12 Anus (probe)
13 Internal iliac artery
14 Remnant of urachus (median umbilical ligament)
15 **Uterus**
16 Round ligament of uterus
17 **Urinary bladder**
18 **Vagina**
19 **Clitoris**
20 Labium minus
21 External orifice of urethra (red probe)
22 Vaginal orifice (green probe)
23 Lateral umbilical ligament
24 Inferior epigastric artery
25 Obturator artery, vein and nerve
26 External iliac artery
27 **Recto-uterine pouch** (of Douglas)
28 Recto-uterine fold
29 Vesico-uterine pouch
30 Suspensory ligament of ovary
31 Greater vestibular gland and bulb of the vestibule

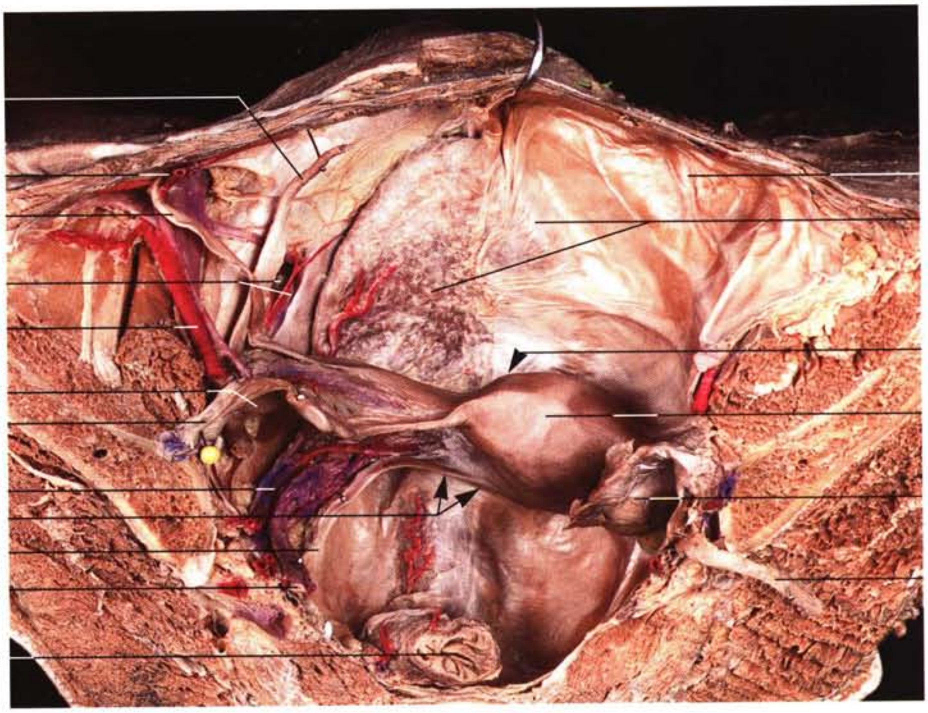

Female internal genital organs in situ (seen from above). The peritoneum at the left half of pelvic cavity has been removed to display uterine tube, vessels, and nerves.

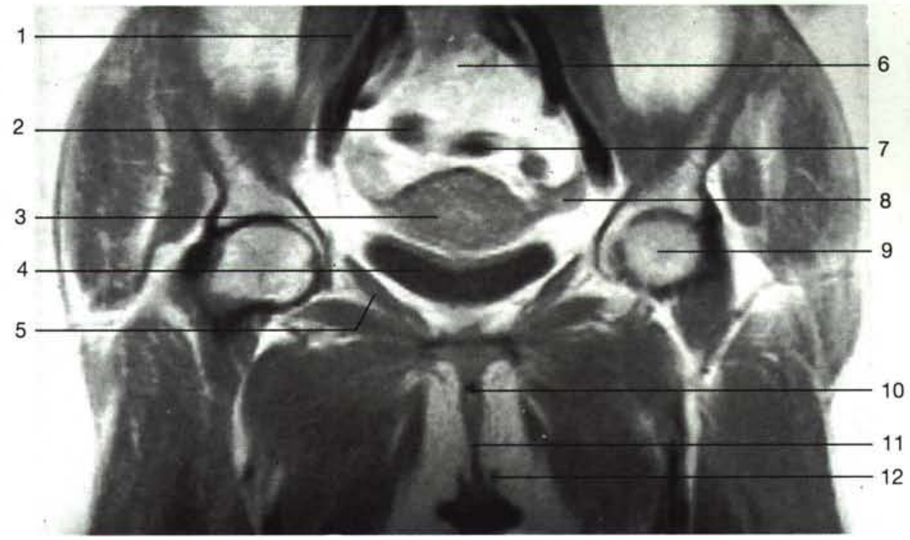

Coronal section through the pelvic cavity of the female. (MR-Scan.)

1	Psoas major muscle
2	Ampulla of rectum
3	**Uterus**
4	**Urinary bladder**
5	Obturator internus muscle
6	Promontory
7	Sigmoid colon
8	**Uterine tube**
9	Head of femur
10	Urethra
11	Vagina
12	Labium minus
13	Umbilicus
14	Duodenum
15	Ascending part of duodenum
16	Root of mesentery
17	Mesentery
18	Vesico-uterine pouch
19	**Urinary bladder** (collapsed)
20	Pubic symphysis
21	Anterior fornix of vagina
22	Clitoris
23	Labium minus
24	Labium majus
25	Vertebral canal with cauda equina
26	Intervertebral disc
27	Body of 5th lumbar vertebra
28	Sacral promontory
29	Mesosigmoid
30	**Recto-uterine pouch** (of Douglas)
31	Posterior fornix of vagina
32	Cervix of uterus
33	External anal sphincter muscle
34	Anal canal
35	Internal anal sphincter muscle
36	Anus
37	Hymen
38	Small intestine
39	Rectus abdominis muscle

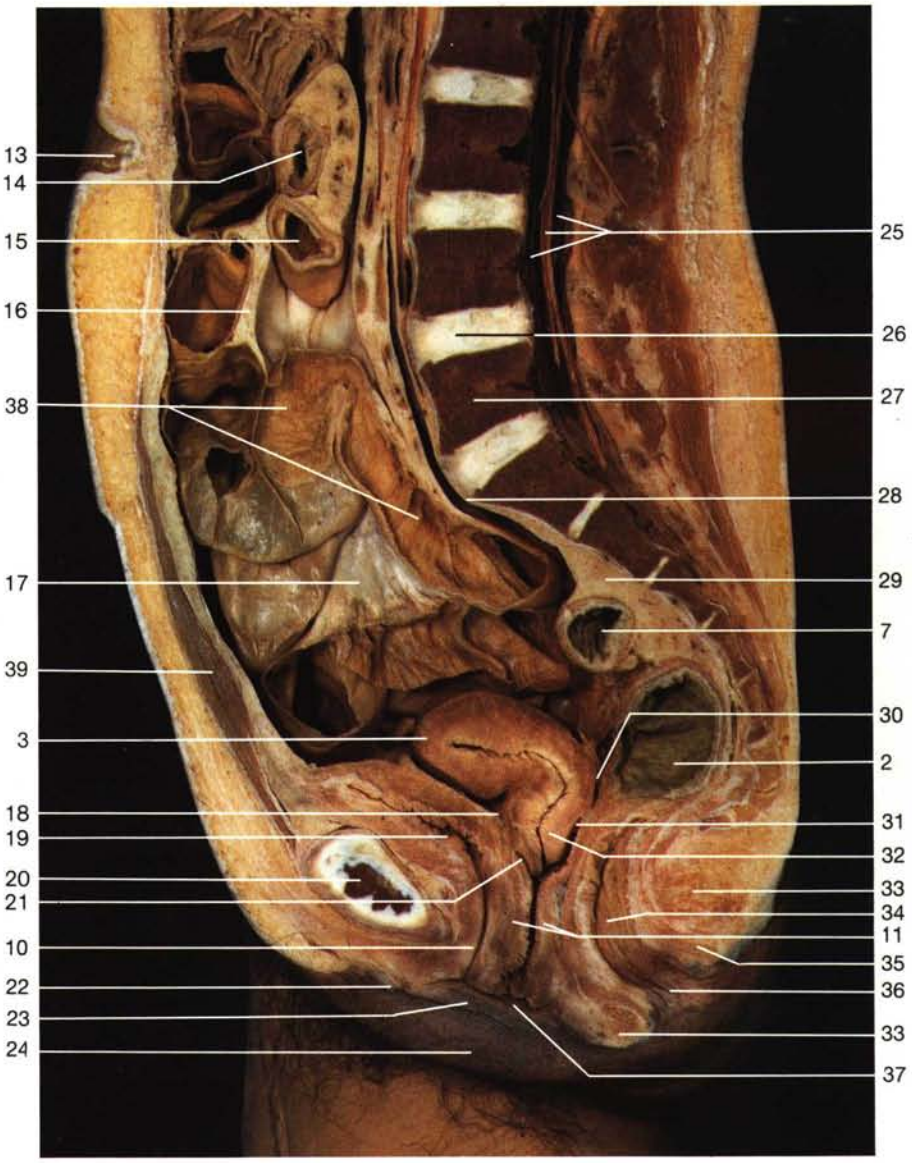

Midsagittal section through the female trunk. The urinary bladder is empty, the position and shape of the uterus are normal.

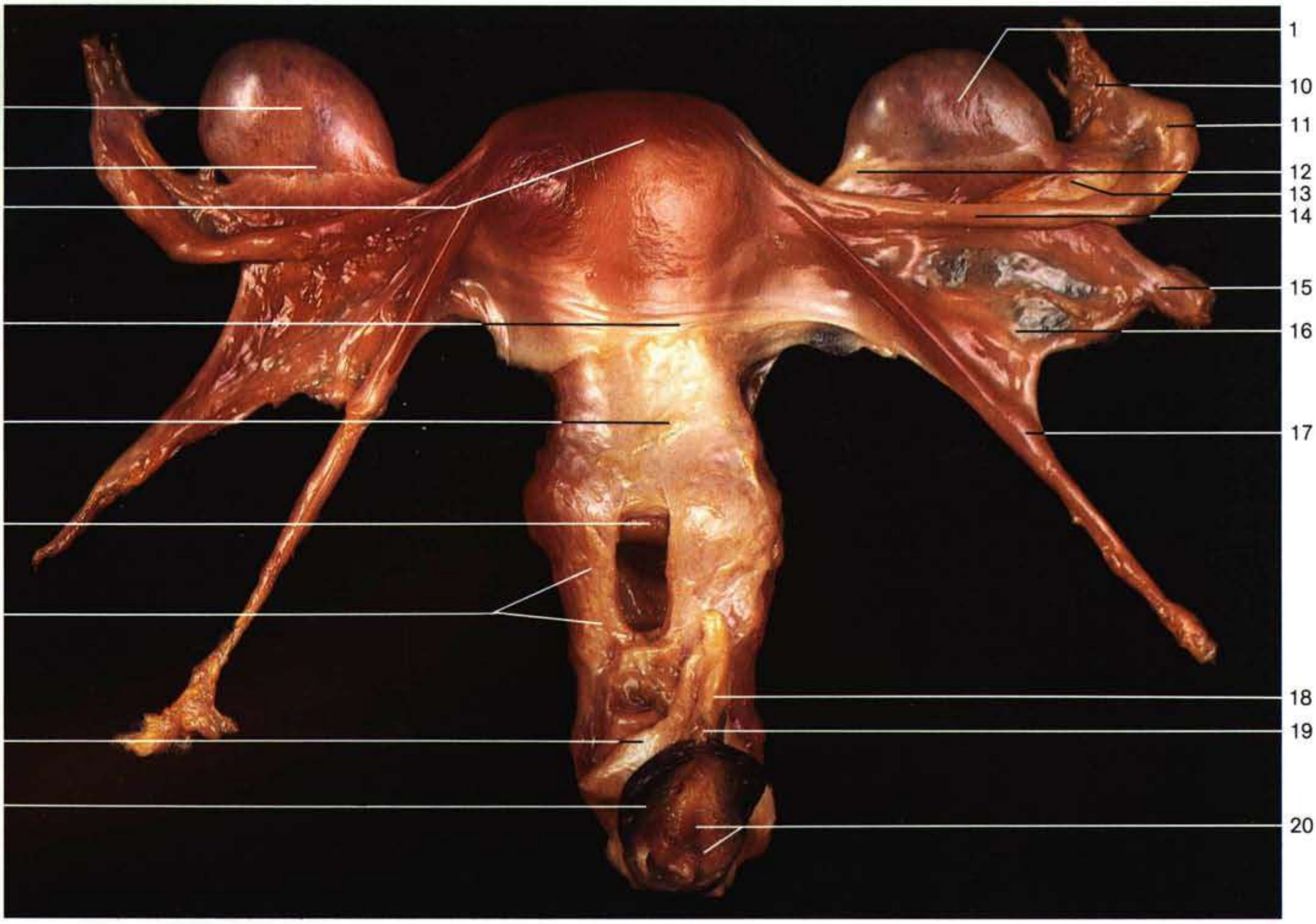

Female genital organs, isolated (anterior view). The anterior wall of the vagina has been opened to display the vaginal portion of the cervix.

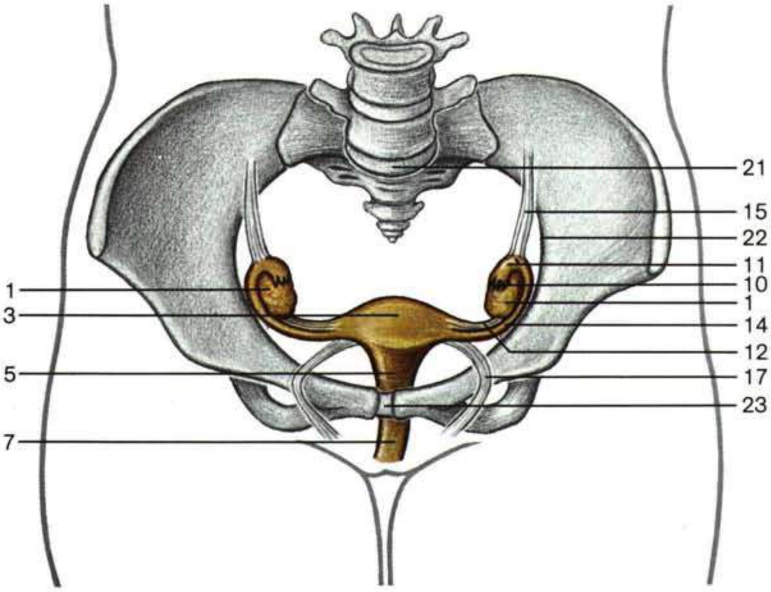

Female internal genital organs. (Schematic drawing.)

1 **Ovary**
2 Mesovarium
3 **Fundus of uterus**
4 Vesico-uterine pouch
5 **Cervix of uterus**
6 Vaginal portion of cervix
7 **Vagina**
8 Crus of clitoris
9 Labium minus
10 Fimbriae of uterine tube
11 Infundibulum of uterine tube
12 Ligament of the ovary
13 Mesosalpinx
14 Uterine tube
15 Suspensory ligament of ovary (caudally displaced)
16 Broad ligament of uterus
17 Round ligament of uterus
18 Corpus cavernosum of clitoris
19 Glans of clitoris
20 **Hymen, vaginal orifice**
21 Promontory
22 Linea terminalis of pelvis
23 Pubic symphysis

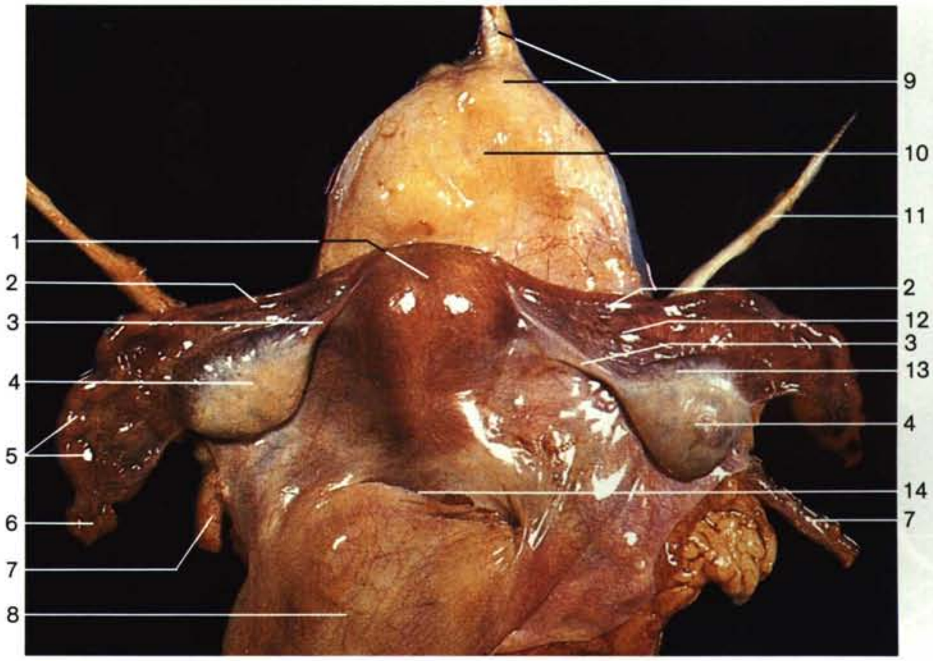

Female internal genital organs, isolated (superior-posterior view).

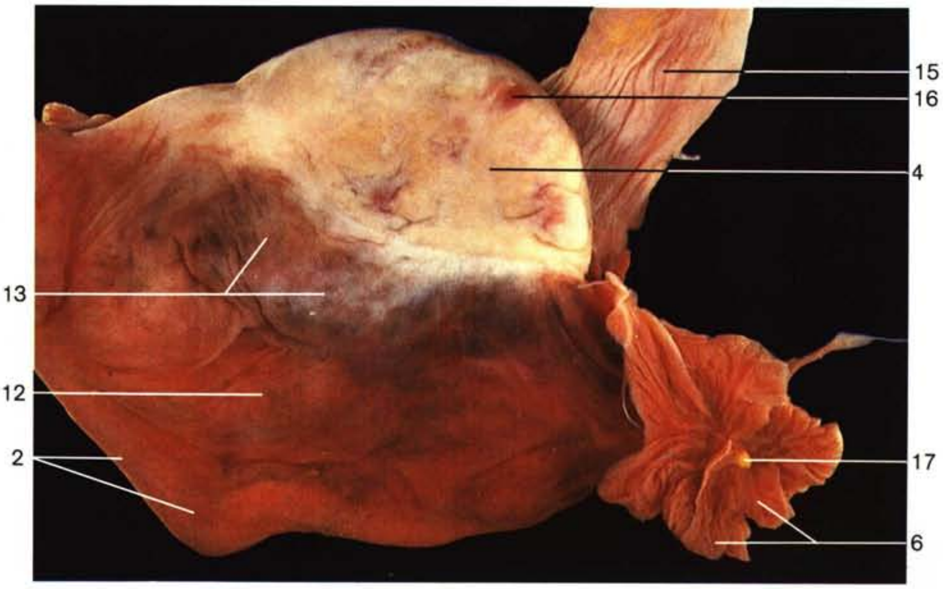

Right ovary and uterine tube, isolated (superior-posterior view). The fimbriae of the uterine tube have been reflected to show the abdominal ostium.

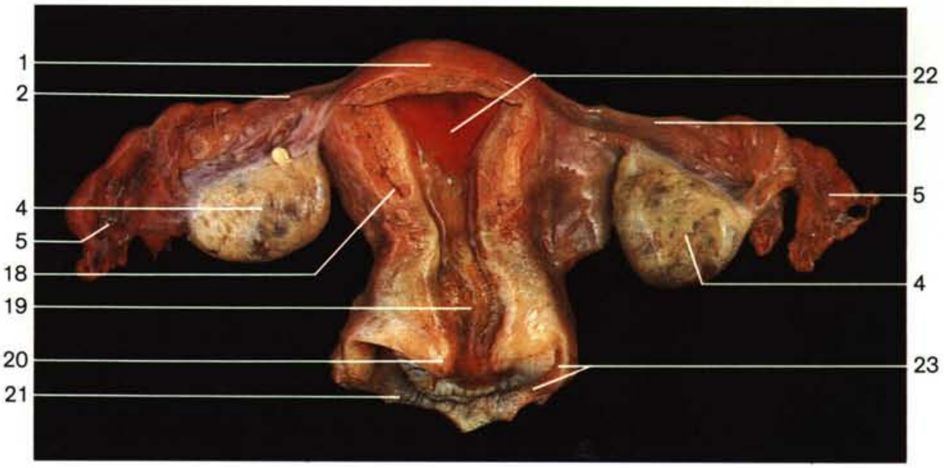

Uterus and related organs (posterior view). The posterior wall of the uterus has been opened.

1 **Fundus of uterus**
2 Uterine tube
3 **Ligament of the ovary**
4 **Ovary**
5 Infundibulum of uterine tube
6 Fimbriae of uterine tube
7 **Ureter**
8 Rectum
9 Apex of urinary bladder and median umbilical ligament
10 Urinary bladder
11 Round ligament of uterus
12 **Mesosalpinx**
13 **Mesovarium**
14 **Rectouterine pouch** (of Douglas)
15 Suspensory ligament of ovary
16 Scarring of ovary (following ovulation)
17 Abdominal opening of uterine tube
18 Body of uterus
19 **Cervical canal**
20 Vaginal portion of cervix of uterus (congestion)
21 Vagina
22 Mucous membrane of uterus
23 Anterior fornix of vagina

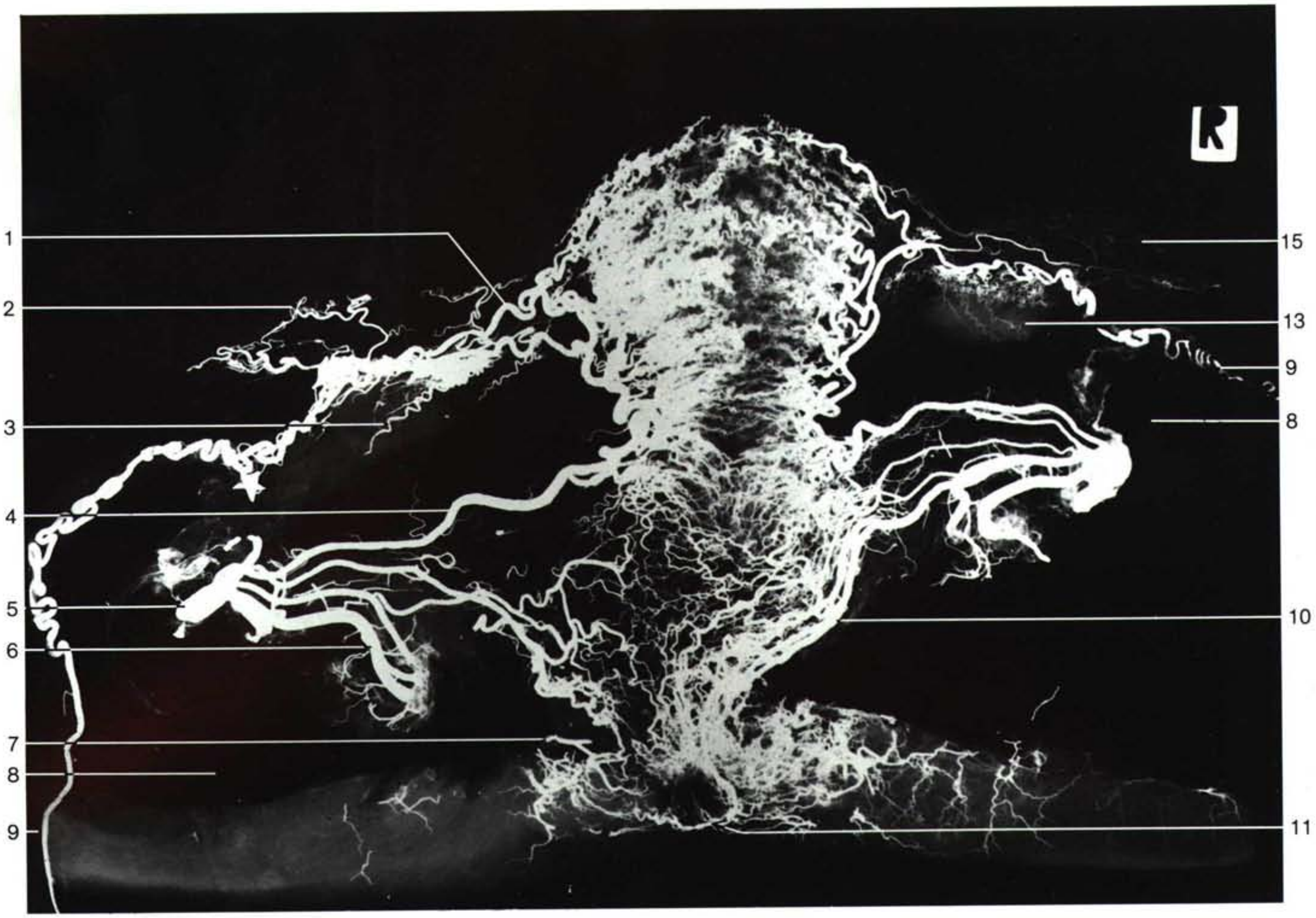

Arteriogram of female genital organs (anterior-posterior view). Notice the helicine arteries supplying uterus and ovary.

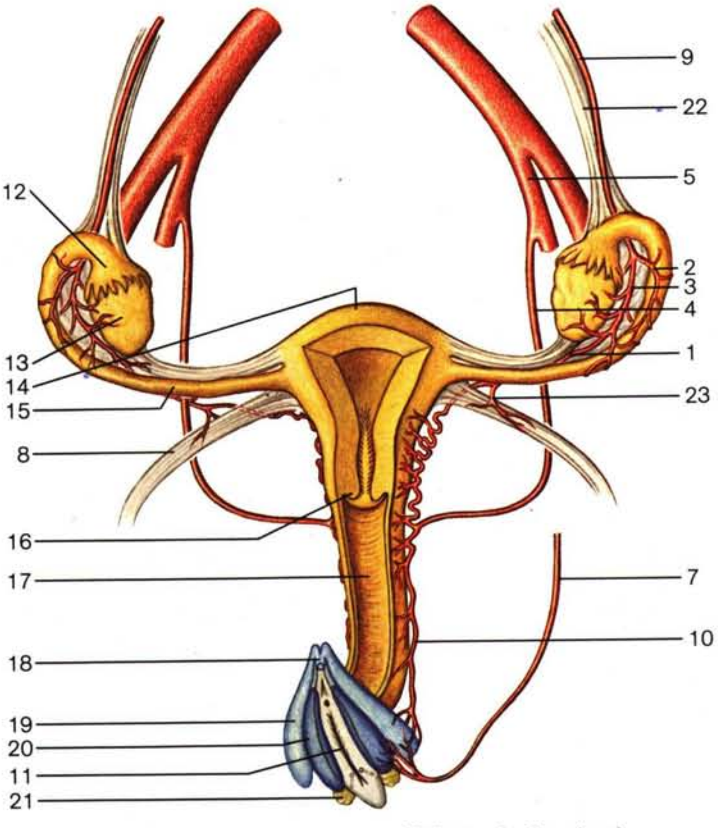

Arteries of female genital organs. (Schematic drawing.)

1 Ovarian branch of uterine artery (anastomoses with ovarian artery)
2 Tubal branch of ovarian artery
3 Ovarian branch of ovarian artery
4 **Uterine artery**
5 Internal iliac artery
6 Inferior gluteal artery
7 **Internal pudendal artery**
8 Round ligament of uterus
9 **Ovarian artery**
10 Vaginal artery
11 Vaginal orifice
12 Infundibulum of uterine tube
13 Ovary
14 Fundus of uterus
15 Uterine tube
16 Vaginal portion of cervix of uterus
17 Vagina
18 Clitoris
19 Corpus cavernosum of clitoris
20 Bulb of vestibule
21 Greater vestibular gland
22 Suspensory ligament of ovary
23 Artery of round ligament

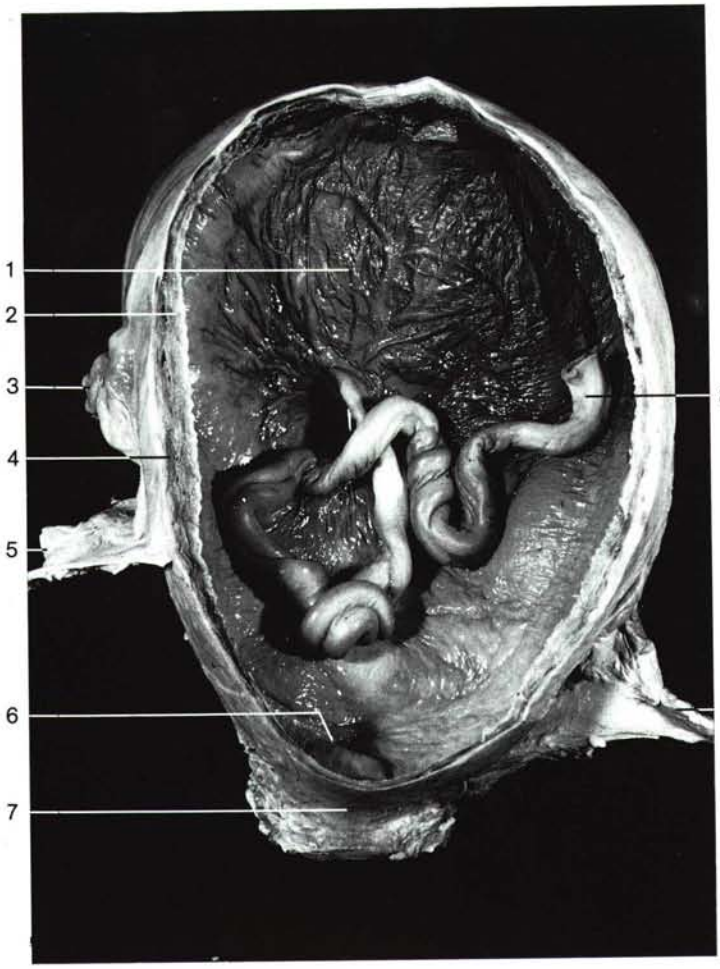

Full term uterus with placenta (anterior view). The anterior wall of the uterus has been removed to show the location of the placenta.

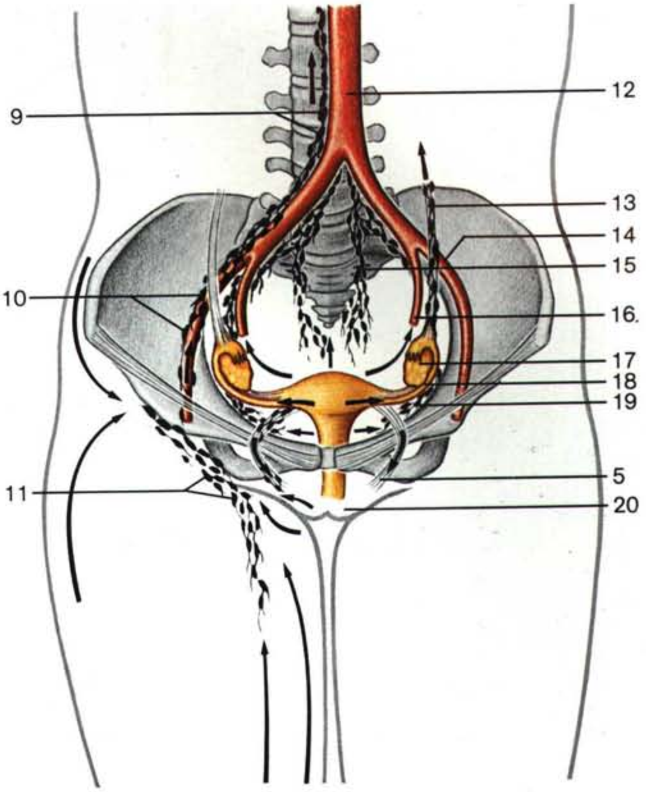

Main drainage routes of lymph vessels of uterus and its adnexa (indicated by arrows). (Schematic drawing.)
Red = arteries; black = lymph vessels and nodes; yellow = internal genital organs.

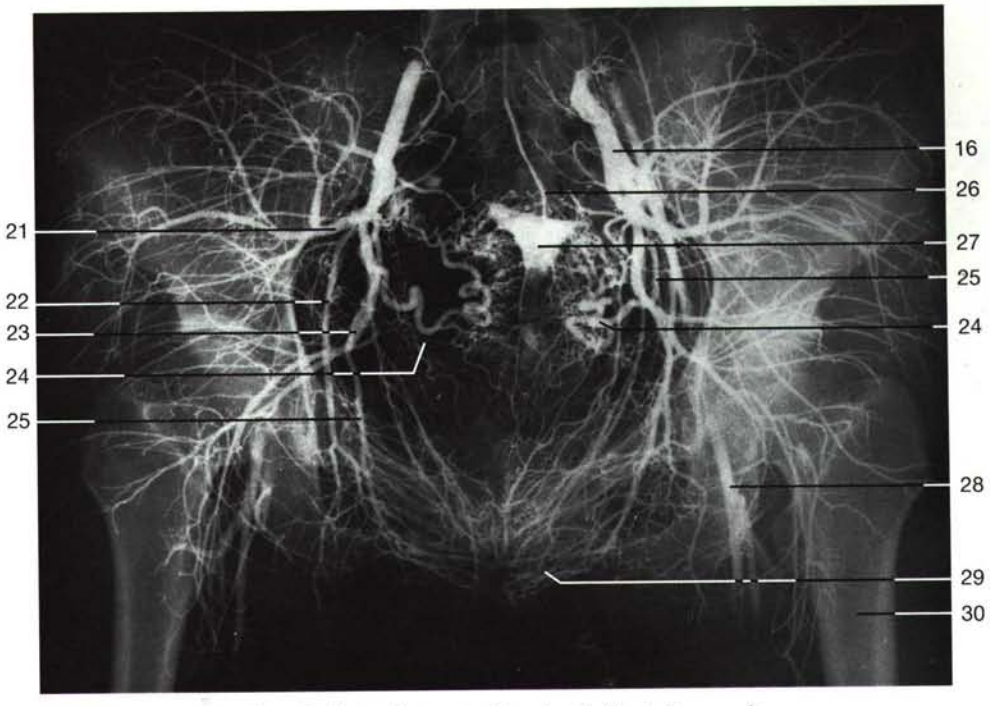

Pelvic vessels in the female (anterior-posterior view). (Arteriogram.)

1 Placenta
2 Amnion and chorion
3 Adnexa of uterus (uterine tube and ovaries)
4 Myometrium
5 Round ligament of uterus
6 Internal orifice of uterus
7 Cervix of uterus
8 **Umbilical cord**
9 Lumbar lymph nodes
10 External iliac lymph nodes
11 Inguinal lymph nodes
12 Abdominal aorta
13 Suspensory ligament of ovary
14 External iliac artery
15 Sacral lymph nodes
16 Internal iliac artery
17 Ovary
18 Uterine tube
19 Internal iliac lymph nodes
20 External genital organs
21 Superior gluteal artery
22 **Obturator artery**
23 Inferior gluteal artery
24 **Uterine artery**
25 **Internal pudendal artery**
26 Middle sacral artery
27 **Uterine cavity**
28 Femoral artery
29 Vessels of labium majus
30 Femur

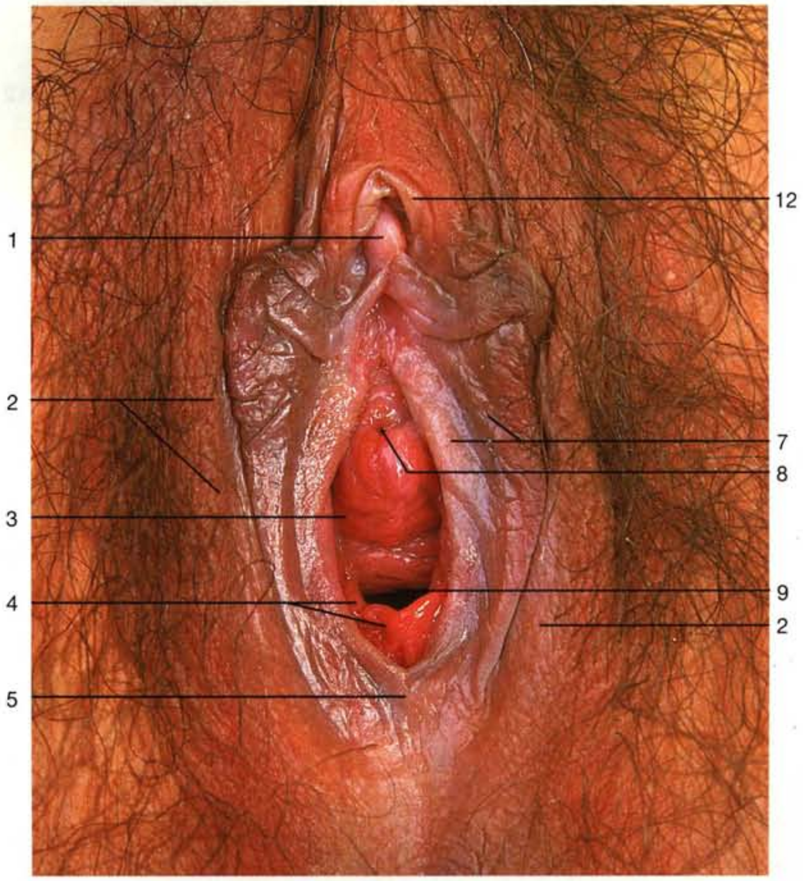

External genital organs in the female (anterior aspect). Labia reflected.

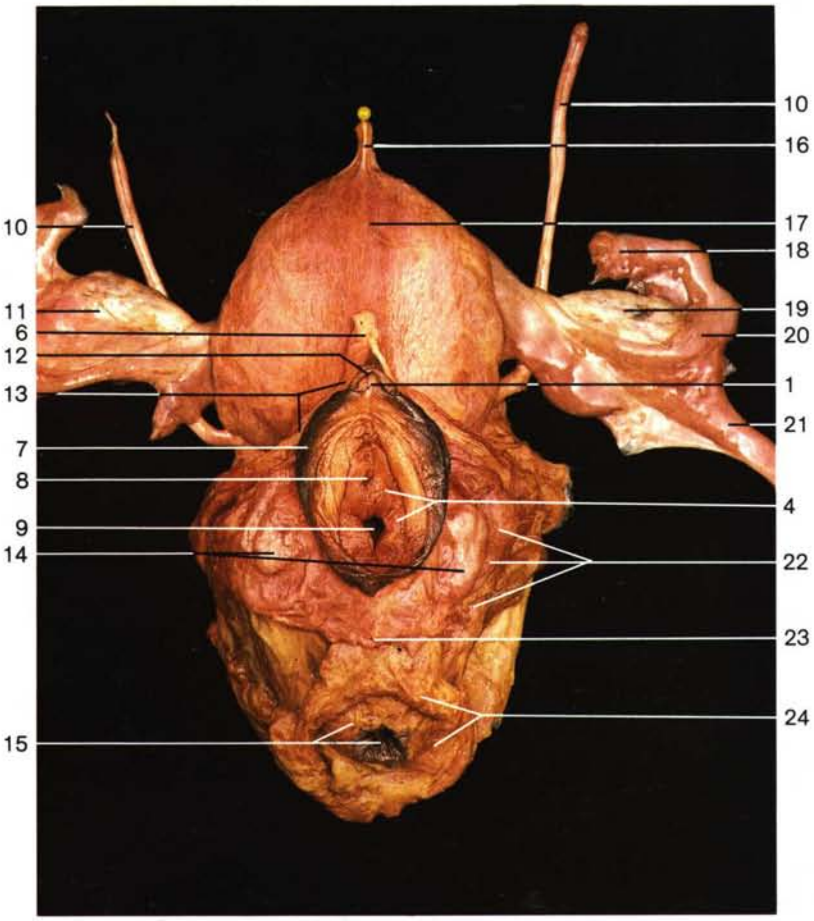

Female external genital organs in relation to internal genital organs and urinary system (isolated, anterior aspect).

1 **Glans clitoris**
2 Labium majus
3 Vestibule of vagina
4 **Hymen**
5 Posterior labial commissure
6 **Body of clitoris**
7 Labium minus
8 **External orifice of urethra**
9 **Vaginal orifice**
10 Ureter
11 Adnexa of uterus
12 Prepuce of clitoris
13 Crus of clitoris
14 Greater vestibular glands
15 Anus and internal anal sphincter muscle
16 Median umbilical ligament containing urachus
17 **Urinary bladder**
18 Infundibulum of uterine tube
19 Ovary
20 Ampulla of uterine tube
21 Suspensory ligament of the ovary
22 Bulbospongiosus muscle and bulb of vestibule
23 Central tendon of perineum (perineal body)
24 External anal sphincter muscle

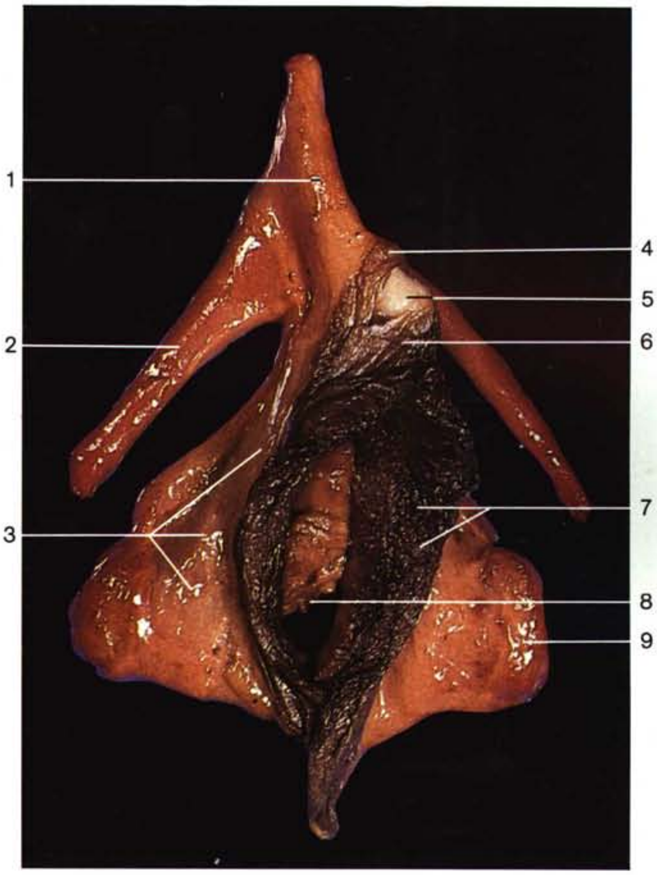

Cavernous tissue of female external genital organs, isolated (anterior aspect).

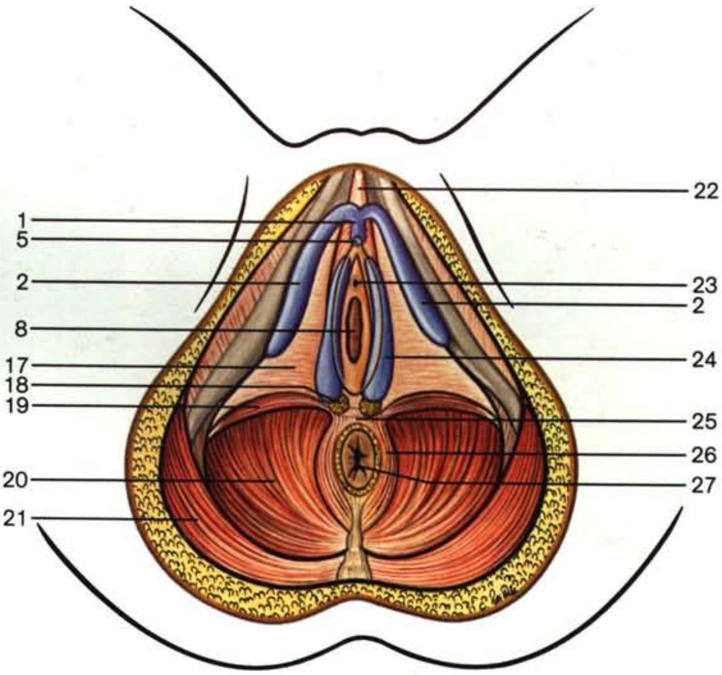

Urogenital and pelvic diaphragms (anterior aspect). (Schematic drawing.)
Blue = cavernous tissue of clitoris and bulb of vestibule.

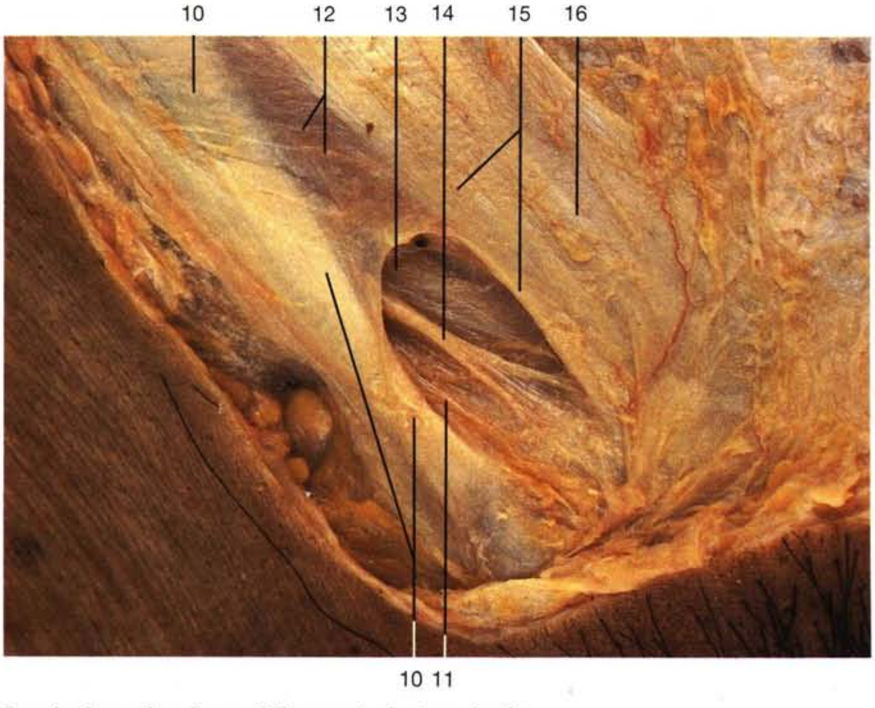

Inguinal canal and round ligament of uterus in situ (right side, ventral aspect).

1	Body of clitoris
2	Crus of clitoris
3	Bulb of vestibule
4	Prepuce of clitoris
5	**Glans of clitoris**
6	Frenulum of clitoris
7	**Labium minus**
8	**Vaginal orifice**
9	**Greater vestibular gland**
10	Lateral crus of superficial inguinal ring
11	Ilioinguinal nerve
12	Intercrural fibers
13	Superficial inguinal ring
14	**Round ligament of uterus**
15	Medial crus of superficial inguinal ring
16	Aponeurosis of external abdominal oblique muscle
17	Deep transverse perineus muscle with fascia
18	Greater vestibular gland
19	Superficial transverse perineus muscle
20	Levator ani muscle
21	Gluteus maximus muscle
22	Suspensory ligament of clitoris
23	External orifice of urethra
24	Bulb of vestibule
25	Perineal body
26	External anal sphincter muscle
27	Anus

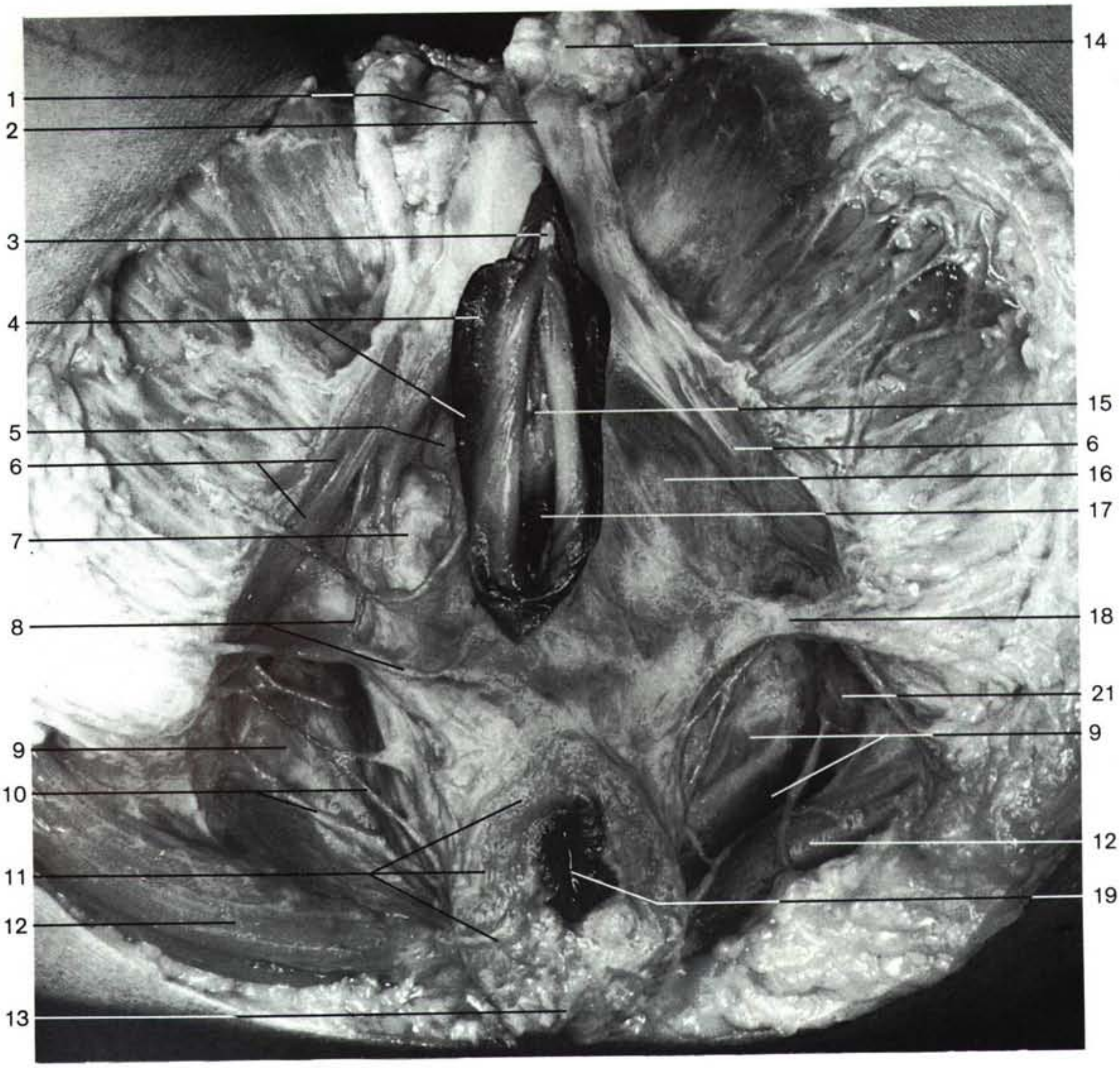

Female urogenital diaphragm and external genital organs, superficial layer (from below).

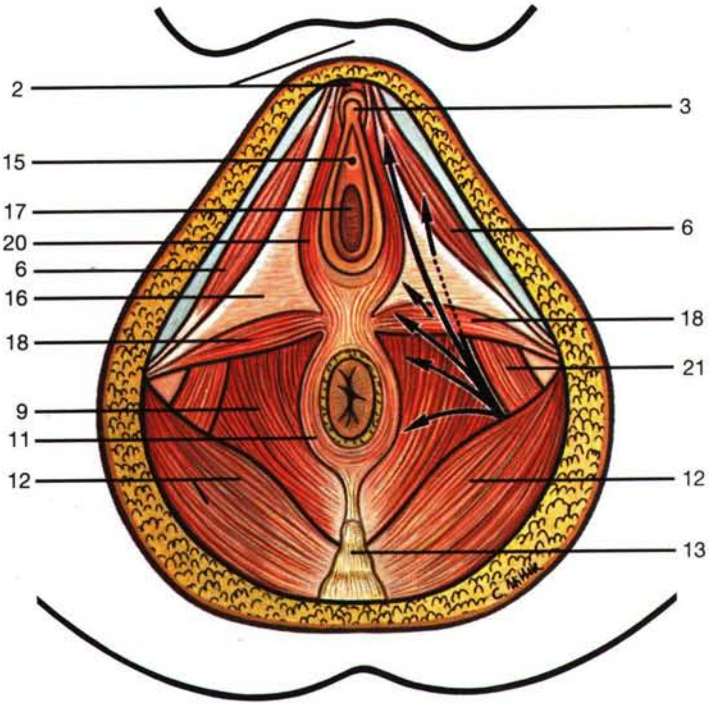

Muscles of pelvic and urogenital diaphragms (from below).
(Schematic drawing.)

1	Fatty tissue encasing round ligament
2	Position of pubic symphysis
3	Clitoris
4	Labium minus
5	Bulb of vestibule
6	Ischiocavernosus muscle
7	**Greater vestibular gland**
8	Perineal branches of pudendal nerve
9	Levator ani muscle
10	Inferior rectal nerves
11	External anal sphincter muscle
12	Gluteus maximus muscle
13	Coccyx
14	Fatty tissue of mons pubis
15	**External orifice of urethra**
16	Urogenital diaphragm with fascia of deep transverse perineus muscle
17	**Vaginal orifice**
18	Superficial transverse perineal muscle
19	**Anus**
20	Bulbospongiosus muscle
21	Obturator internus muscle

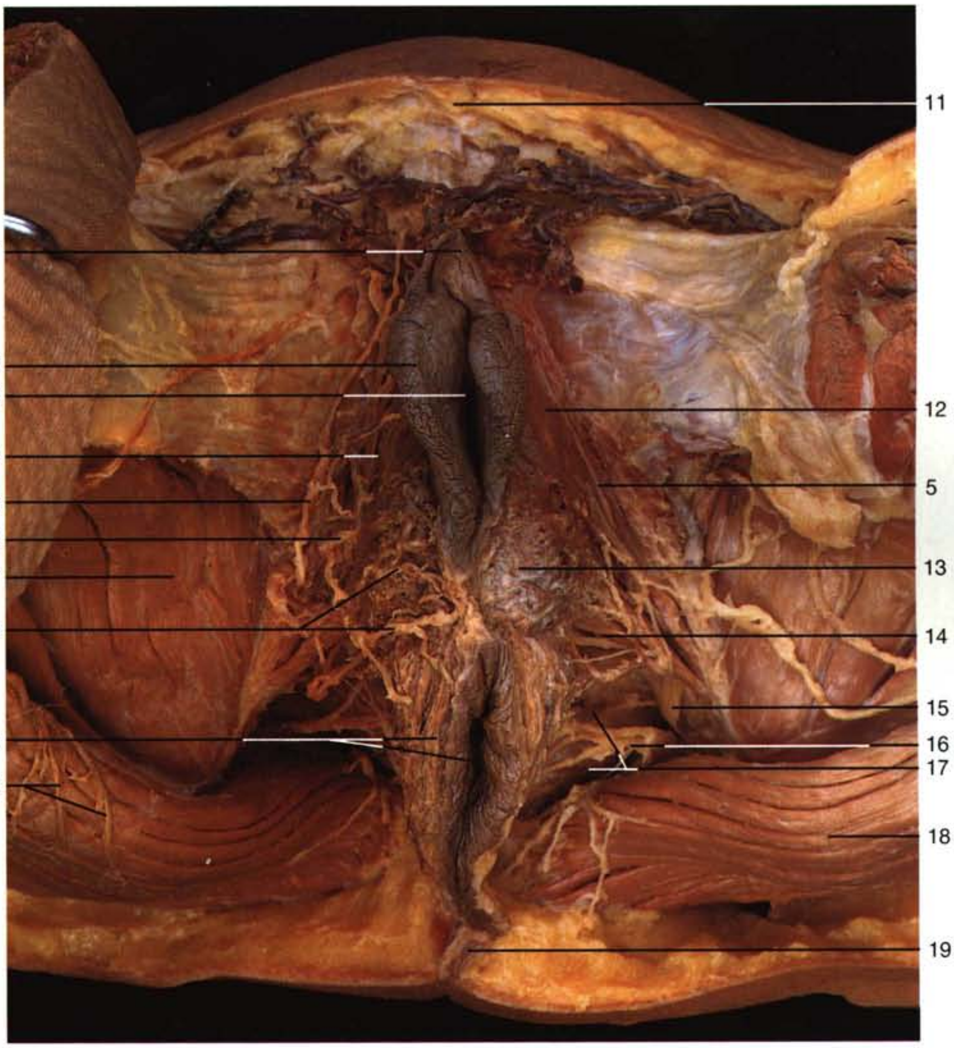

Urogenital diaphragm and external genital organs in the female, superficial layer (from below). On the right side the bulb of vestibule has been removed.

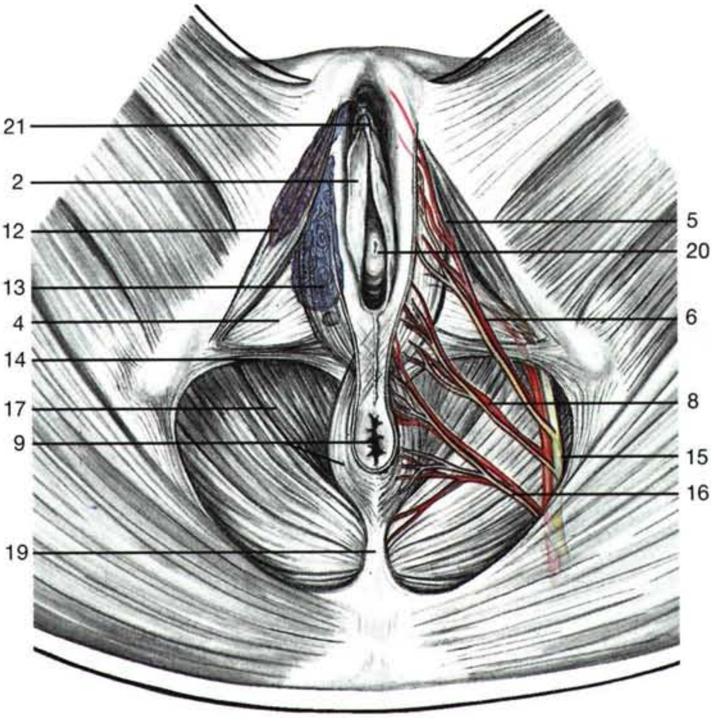

External female genital organs. Position of arteries and nerves; bulb of vestibule in blue. (Schematic drawing.)

1	**Prepuce of clitoris**
2	Labium minus
3	**Vaginal orifice**
4	Deep transverse perineus muscle
5	Dorsal nerve of clitoris
6	Posterior labial nerves
7	Great adductor muscle
8	Perineal branches of pudendal nerve
9	**Anus** and **external anal sphincter muscle**
10	Inferior cluneal nerves
11	Mons pubis
12	Crus of clitoris with ischiocavernosus muscle
13	Bulb of vestibule
14	Superficial transverse perineus muscle
15	**Pudendal nerve** and **internal pudendal artery**
16	Inferior rectal nerves
17	Levator ani muscle
18	Gluteus maximus muscle
19	Anococcygeal ligament
20	External urethral orifice
21	Glans of clitoris

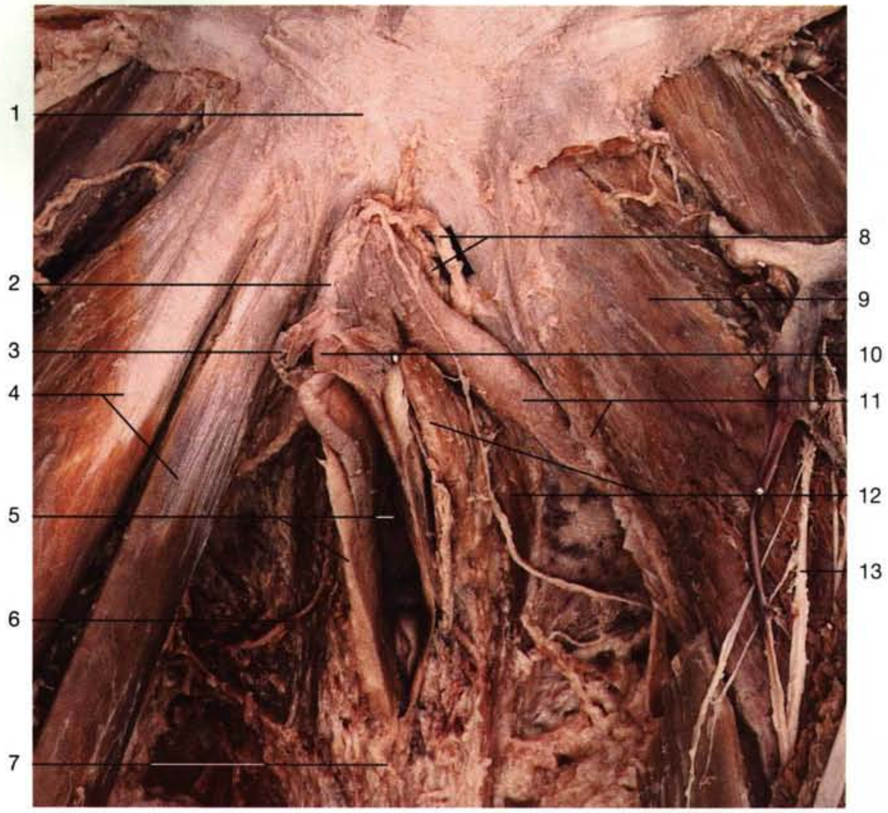

External female genital organs (inferior aspect). The clitoris has been dissected and slightly reflected to the right. The prepuce of clitoris has been divided to display the glans.

1 Position of pubic symphysis
2 Body of clitoris
3 Prepuce of clitoris
4 Adductor longus and gracilis muscles
5 External orifice of vagina and labium minus
6 Posterior labial nerve
7 Perineal body
8 Deep artery of clitoris and dorsal nerve of clitoris
9 Adductor brevis muscle
10 **Glans of clitoris**
11 Crus of clitoris and ischiocavernosus muscle
12 **Bulb of vestibule** and bulbospongiosus muscle
13 Anterior branch of obturator nerve
14 Labium minus
15 **Vaginal orifice**
16 Posterior labial nerves
17 Branches of pudendal nerve
18 External sphincter of anus
19 **Anus**
20 Bulb of vestibule (divided)
21 Dorsal artery of clitoris
22 Superficial transverse perineus muscle
23 Perineal branch of posterior femoral cutaneous nerve
24 Levator ani muscle
25 **Pudendal nerve** and **internal pudendal artery**
26 Inferior rectal nerves
27 Gluteus maximus muscle
28 Anococcygeal ligament

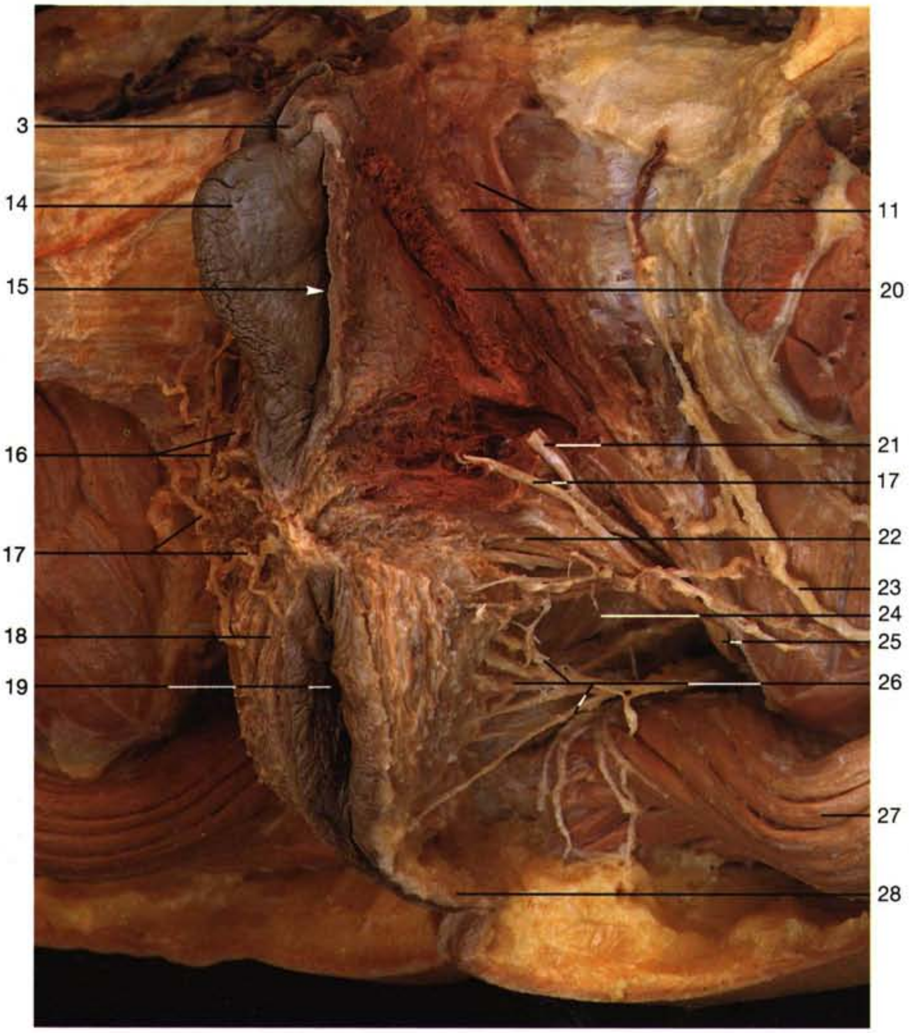

Urogenital diaphragm and external genital organs in the female (lateral inferior view). The bulb of vestibule has partly been removed, the left labium minus was cut away.

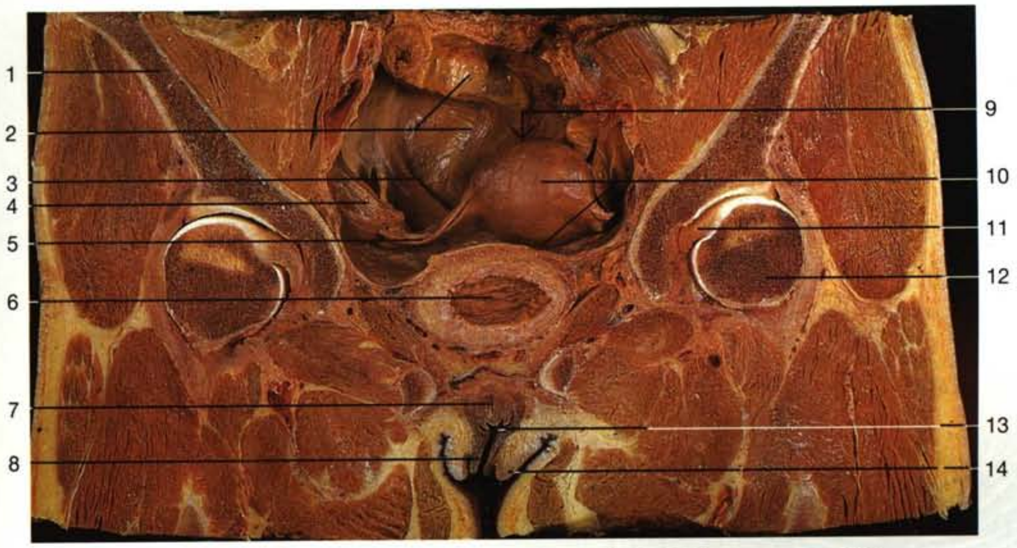

Coronal section through the pelvic cavity of the female (cf. MR-Scan on page 335).

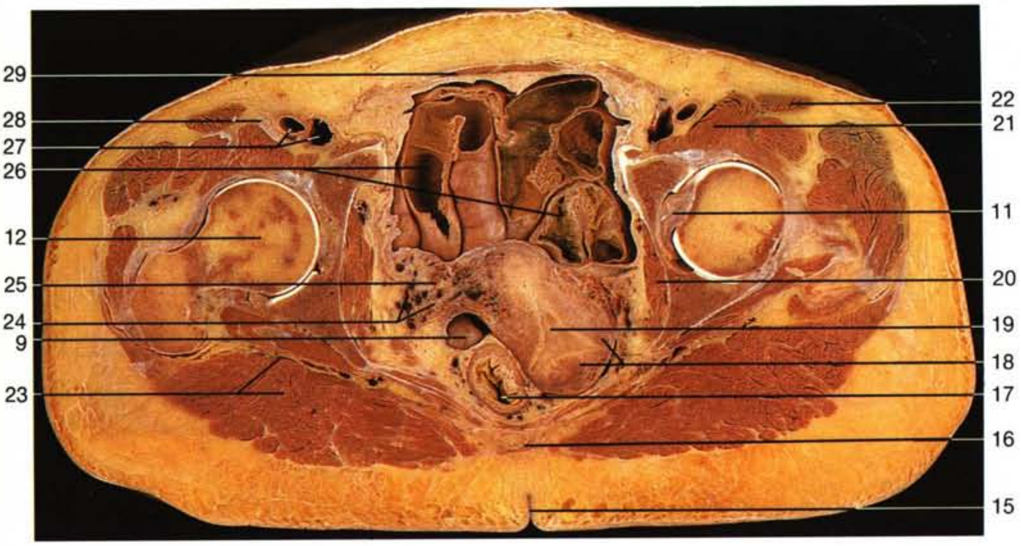

Horizontal section through pelvic cavity at level of uterus (from below). The uterus is retroverted to the left.

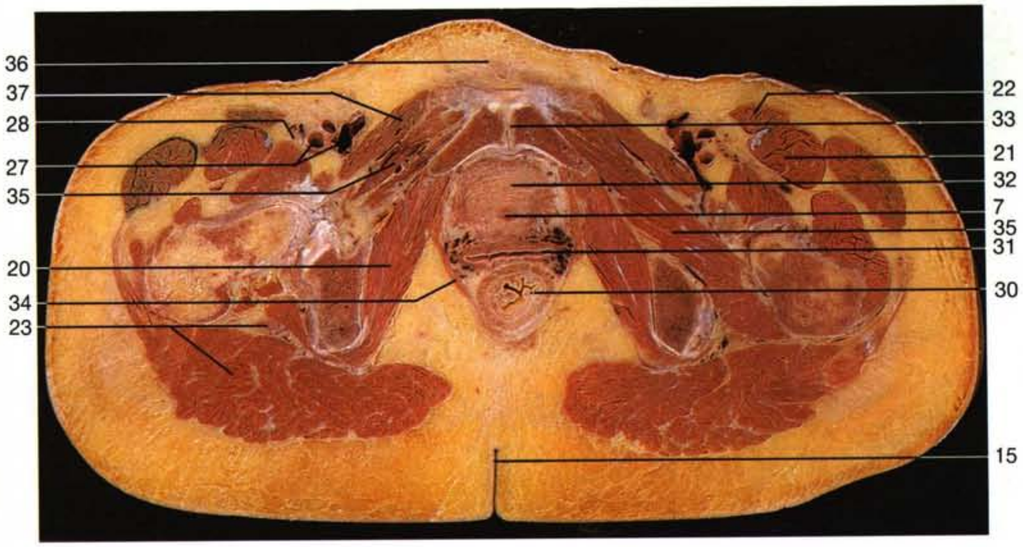

Horizontal section through the pelvic cavity at level of the urethral sphincter and vagina (from below).

1 Ilium
2 Rectum
3 Rectouterine fold
4 **Ovary**
5 **Uterine tube**
6 **Urinary bladder**
7 **Urethra**
8 Labium minus
9 Rectouterine pouch of Douglas
10 **Uterus** (uterovesical pouch)
11 Ligament of the head of the femur
12 Head of femur
13 Vestibule of vagina
14 Labium majus
15 Anal cleft
16 Coccyx
17 Rectum
18 Myometrium of uterus
19 Uterine cavity
20 Obturator internus muscle
21 Iliopsoas muscle
22 Sartorius muscle
23 Sciatic nerve and gluteus maximus muscle
24 Uterine venous plexus
25 Broad ligament
26 Small intestine
27 Femoral artery and vein
28 Femoral nerve
29 Pyramidalis muscle
30 Rectum (anal canal)
31 Vagina
32 Urethral sphincter muscle (base of urinary bladder)
33 Pubic symphysis
34 Levator ani muscle
35 Obturator externus muscle
36 Mons pubis
37 Pectineus muscle

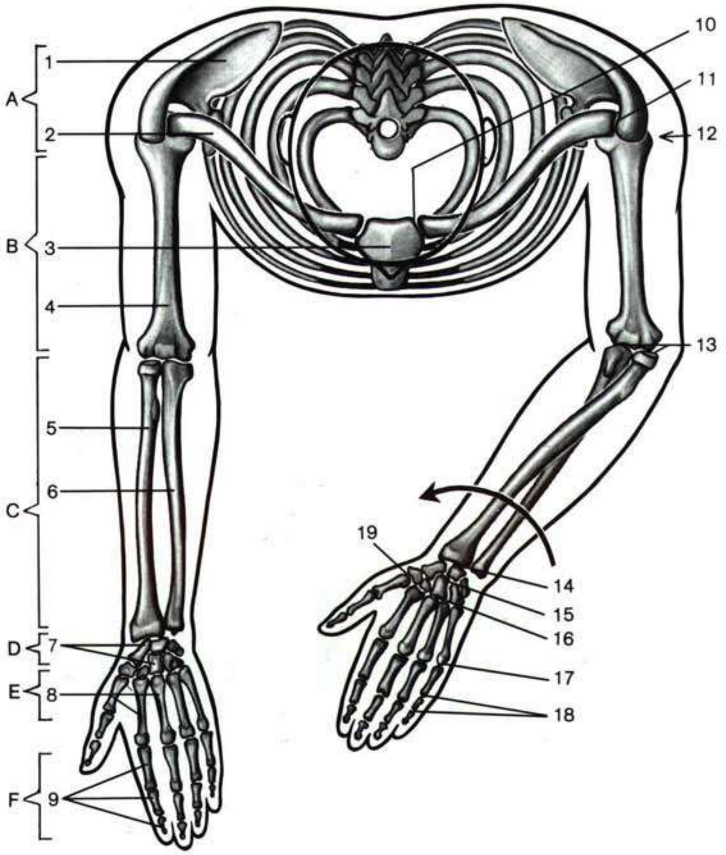

A Pectoral girdle
B Arm
C Forearm
D Wrist
E Palm of hand
F Finger

Bones

1 Scapula
2 Clavicle
3 Sternum
4 Humerus
5 Radius
6 Ulna
7 Carpal bones
8 Metacarpal bones
9 Phalanges

Joints

10 Sternoclavicular joint
11 Acromioclavicular joint
12 Shoulder joint
13 Elbow joint
14 Radiocarpal joint
15 Midcarpal joint
16 Carpometacarpal joint
17 Metacarpophalangeal joint
18 Interphalangeal joints of fingers
19 Carpometacarpal joint of thumb

Organization of shoulder girdle and upper limb (superior aspect). The two positions of the forearm essential to manual skills in the human, supination (right arm) and pronation (left arm) are shown.

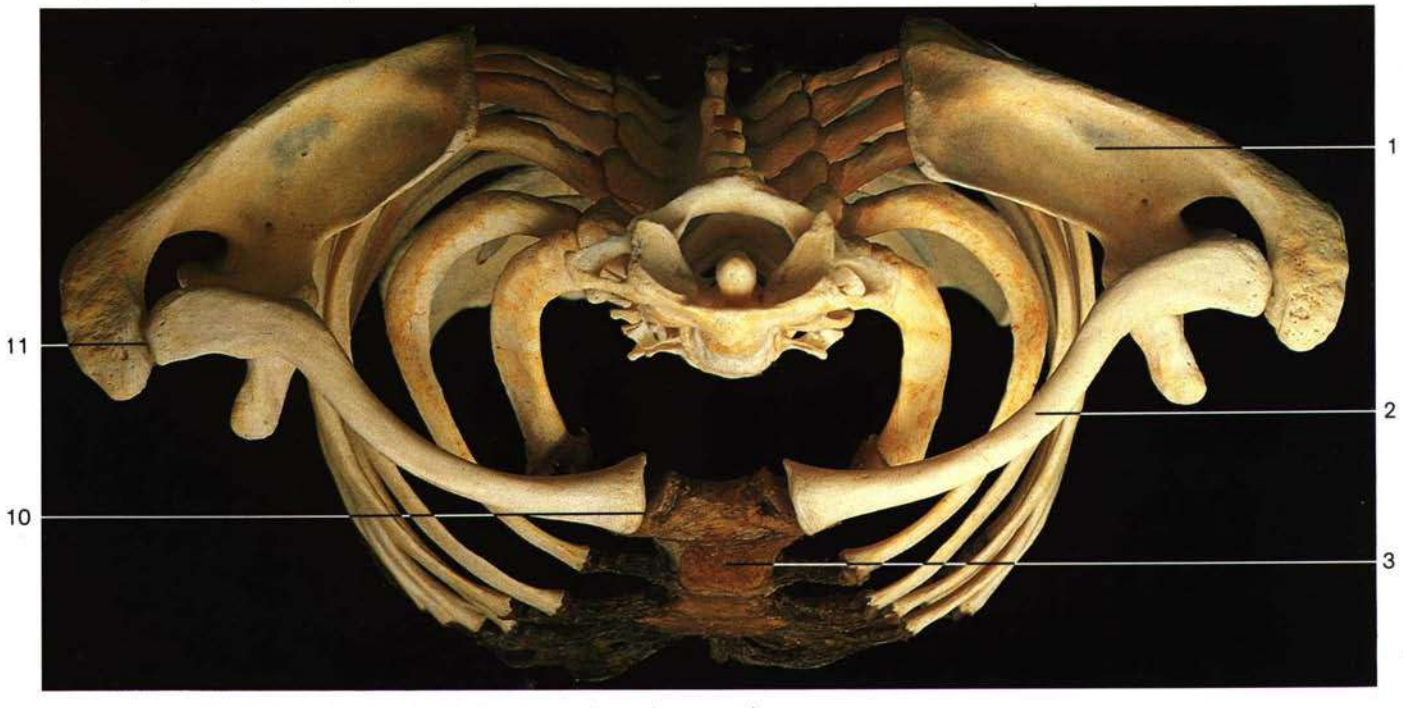

Bones of shoulder girdle articulated with the thorax (superior aspect).

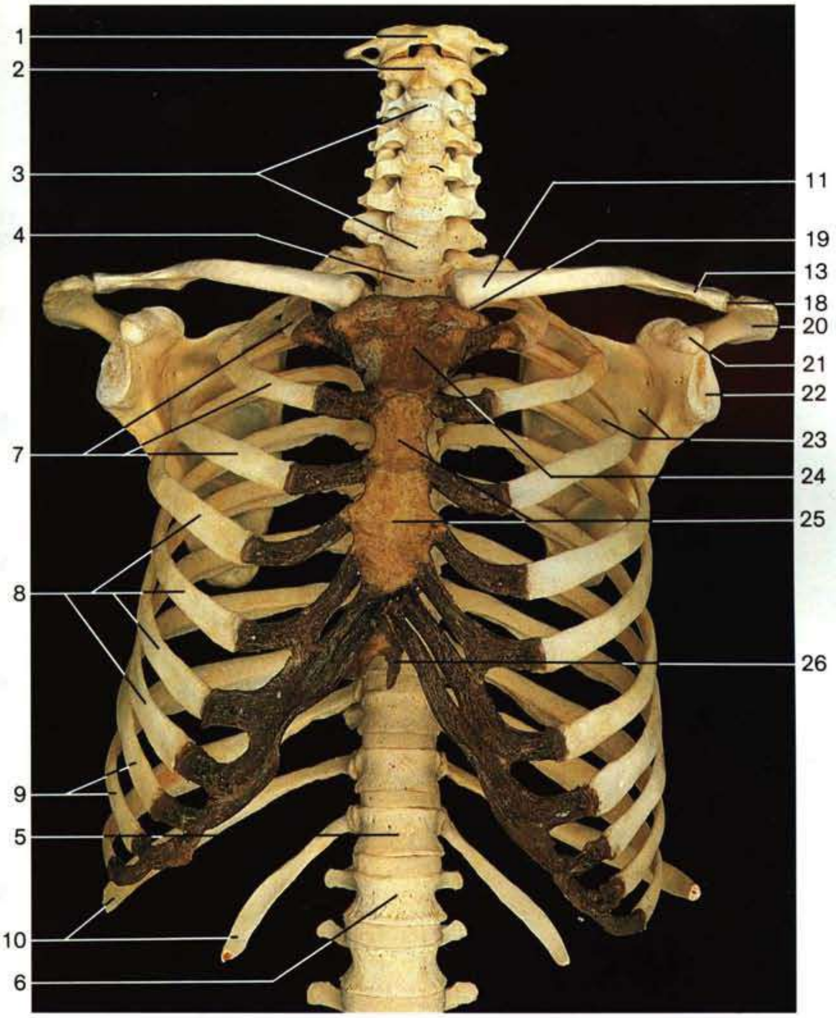

Skeleton of shoulder girdle and thorax (anterior aspect). The cartilaginous parts of the ribs appear dark brown.

Vertebral column

1 Atlas
2 Axis
3 3rd-7th cervical vertebrae
4 1st thoracic vertebra
5 12th thoracic vertebra
6 1st lumbar vertebra

Ribs

7 1st–3rd ribs } True ribs
8 4th–7th ribs
9 8th–10th ribs } False ribs
10 11th and 12th ribs (floating ribs)

Clavicle

11 Sternal end
12 Articular facet for sternum
13 Acromial end
14 Articular facet for acromion
15 Impression for costoclavicular ligament
16 Conoid tubercle
17 Trapezoid line
18 Site of acromioclavicular joint
19 Site of sternoclavicular joint

Scapula

20 Acromion
21 Coracoid process
22 Glenoid cavity
23 Costal surface

Sternum

24 Manubrium
25 Body
26 Xiphoid process

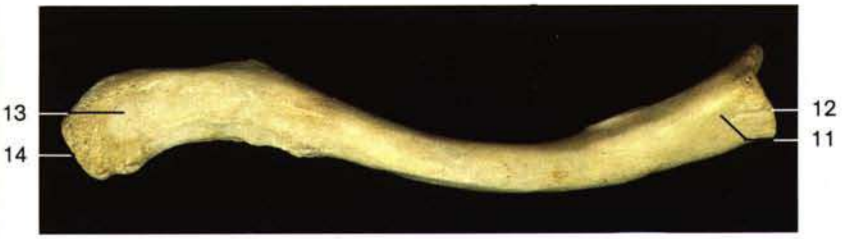

Right clavicle (superior aspect).

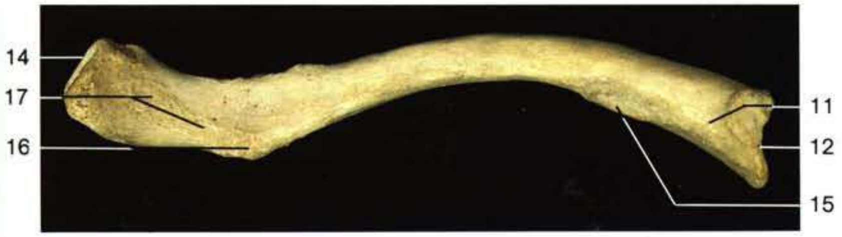

Right clavicle (inferior aspect).

Because of his upright posture, man's upper limb has developed a high degree of mobility. The shoulder girdle is to a great extent movable in the thorax and is connected with the trunk only by the sternoclavicular joint. A further characteristic of man's forearm is the capacity for rotation (i.e., pronation and supination).

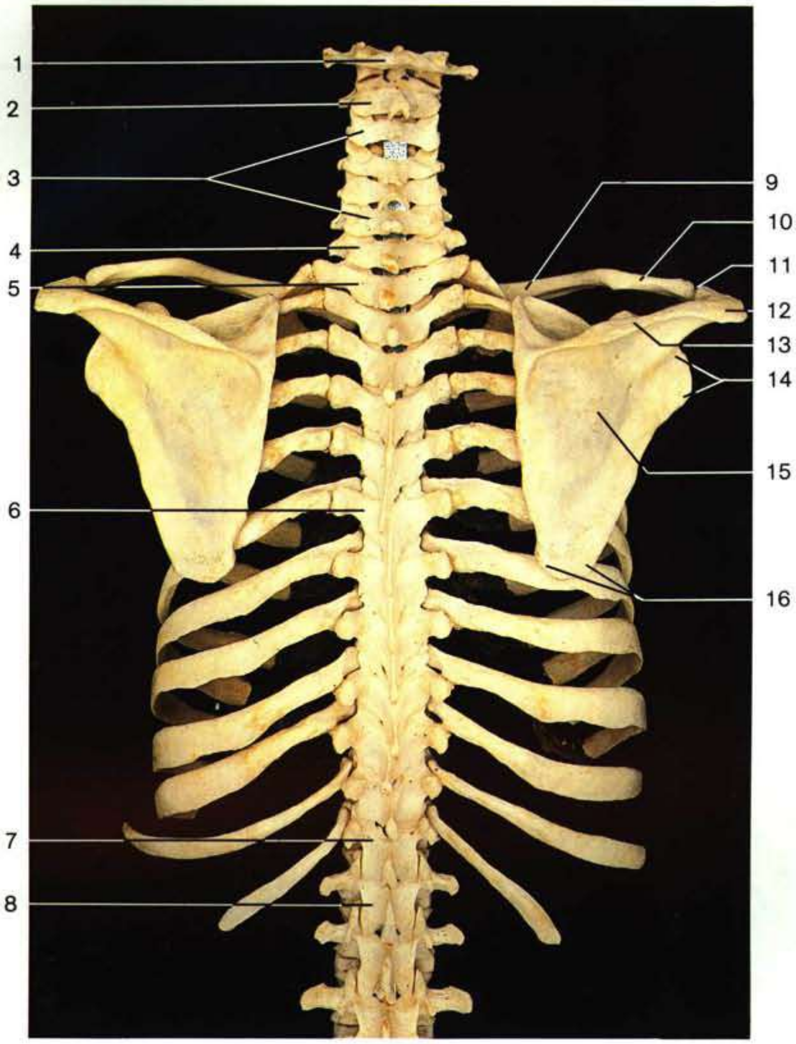

Skeleton of shoulder girdle and thorax (posterior view).

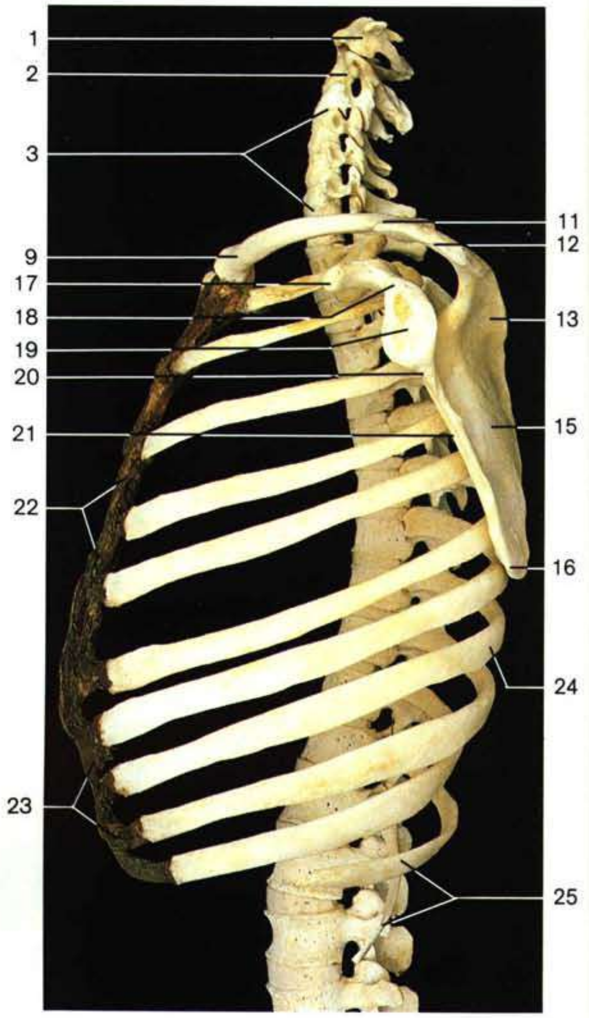

Skeleton of shoulder girdle and thorax (lateral view).

Vertebral column

1 Atlas
2 Axis
3 3rd–6th cervical vertebrae
4 7th vertebra (vertebra prominens)
5 1st thoracic vertebra
6 6th thoracic vertebra
7 12th thoracic vertebra
8 1st lumbar vertebra

Clavicle

9 Sternal end
10 Acromial end
11 Site of acromioclavicular joint

Scapula

12 Acromion
13 Spine of scapula
14 Lateral angle
15 Posterior surface
16 Inferior angle
17 Coracoid process
18 Supraglenoid tubercle
19 Glenoid cavity
20 Infraglenoid tubercle
21 Lateral margin

Thorax

22 Body of sternum
23 Costal arch
24 Angle of ribs
25 Free ribs

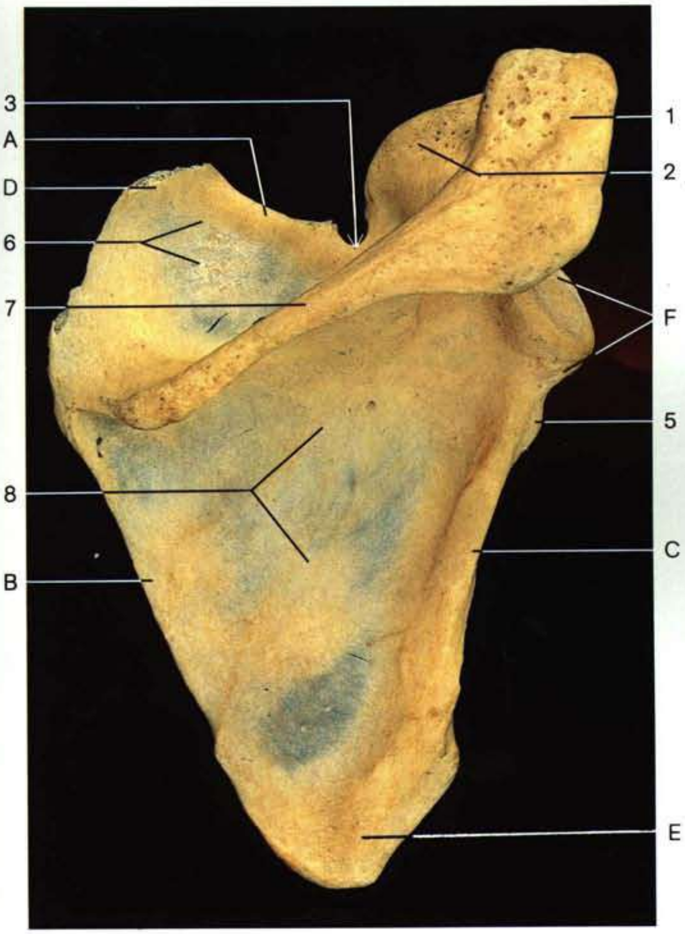

Right scapula (posterior aspect).

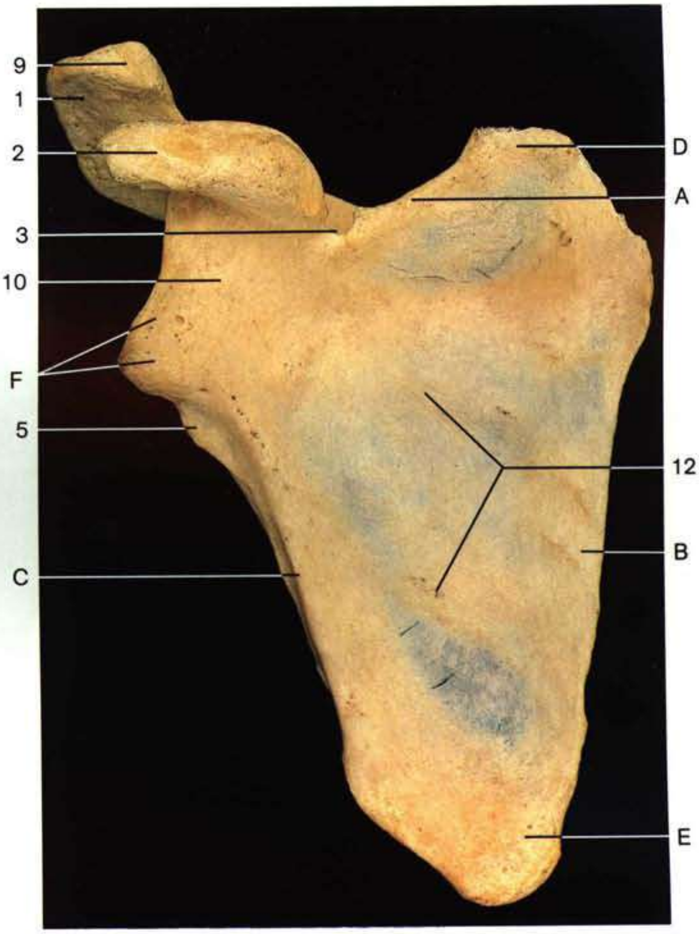

Right scapula (anterior aspect costal surface).

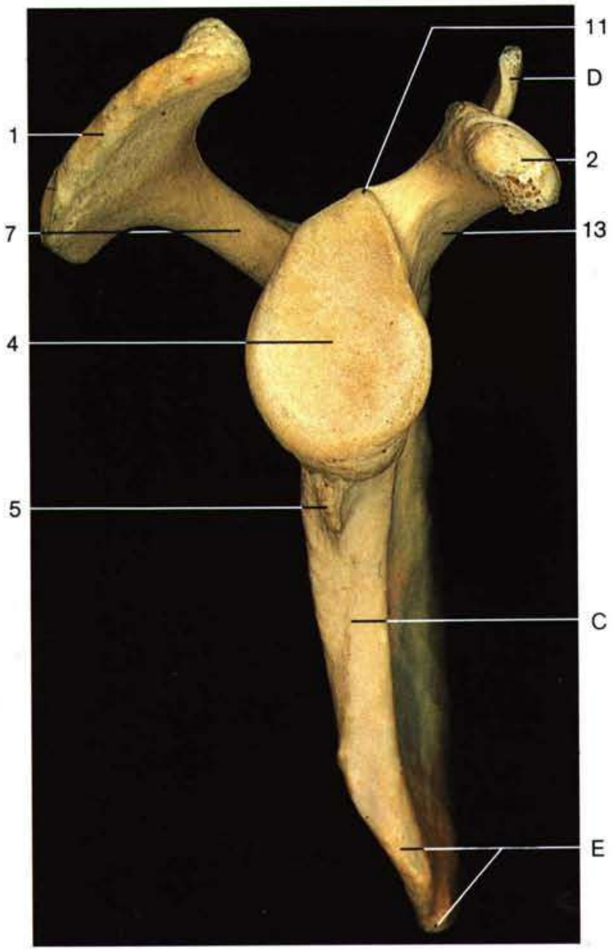

Right scapula (lateral view).

Scapula

A	Superior border
B	Medial border
C	Lateral border
D	Superior angle
E	Inferior angle
F	Lateral angle

1	Acromion
2	Coracoid process
3	Scapular notch
4	Glenoid cavity
5	Infraglenoid tubercle
6	Supraspinous fossa
7	Spine
8	Infraspinous fossa
9	Articular facet for acromion
10	Neck
11	Supraglenoid tubercle
12	Costal (anterior) surface
13	Base of coracoid process

1 1st rib
2 Position of costotransverse joints
3 4th–7th ribs
4 **Clavicle**
5 Position of acromioclavicular joint
6 Acromion
7 Scapular notch
8 Spine of scapula
9 Head of humerus
10 Glenoid cavity
11 Surgical neck of humerus
12 Posterior surface of **scapula**
13 Coracoid process
14 Infraglenoid tubercle
15 Greater tubercle of **humerus**
16 Anatomical neck of humerus

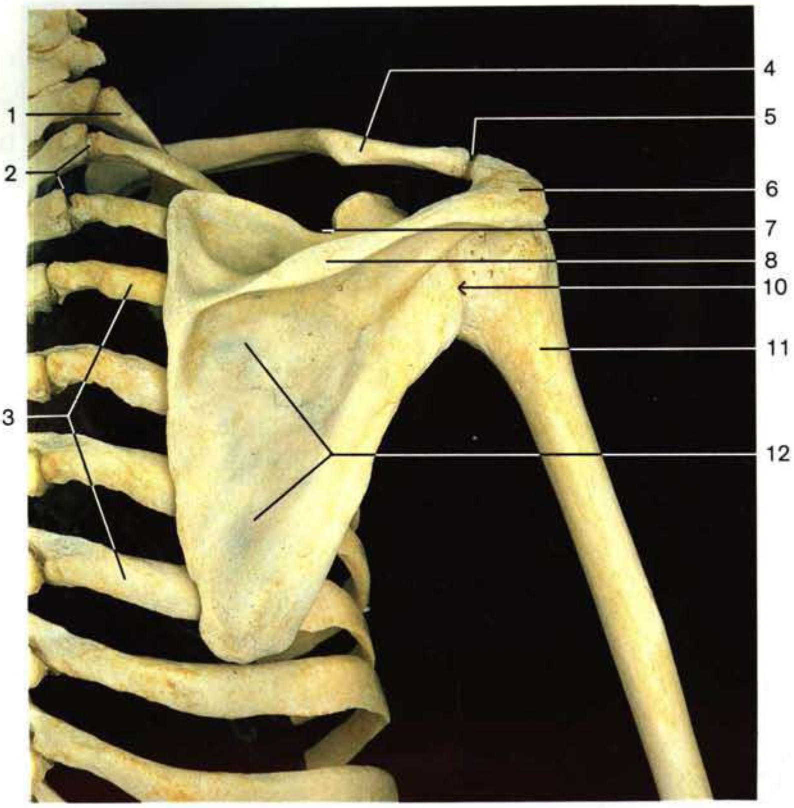

Bones of shoulder joint (posterior aspect).

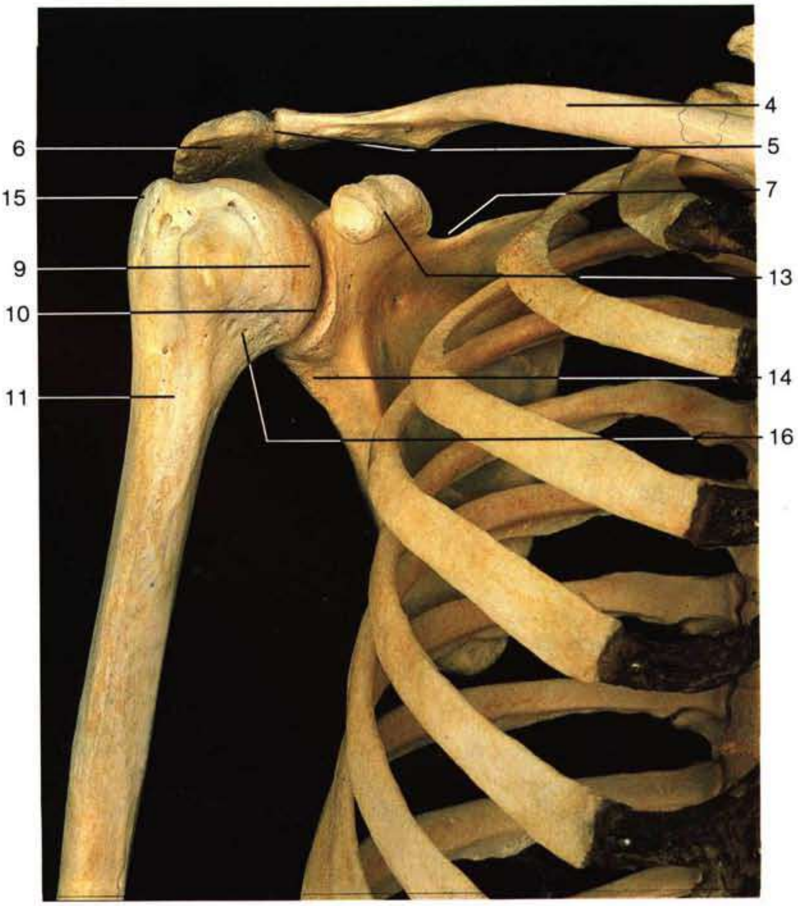

Bones of shoulder joint (anterior aspect).

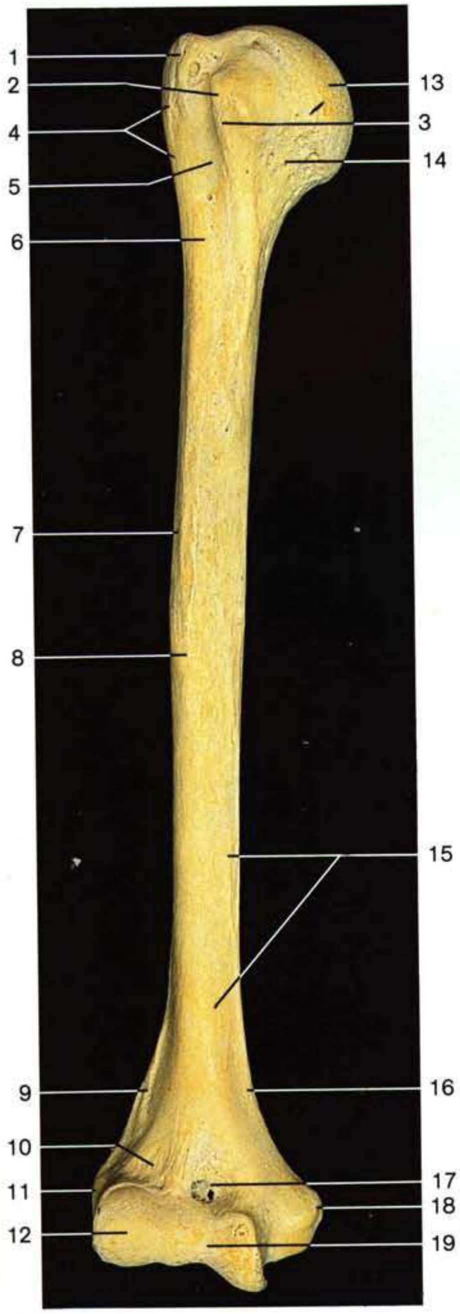
1
2
4
5
6
7
8
9
10
11
12
13
3
14
15
16
17
18
19

Right humerus (anterior aspect).

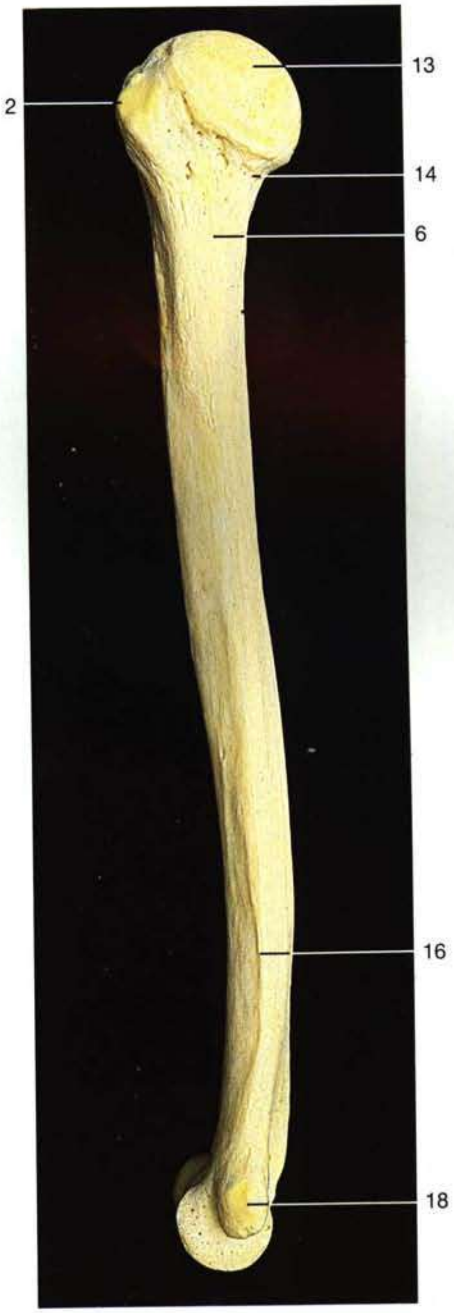
2
13
14
6
16
18

Right humerus (medial aspect).

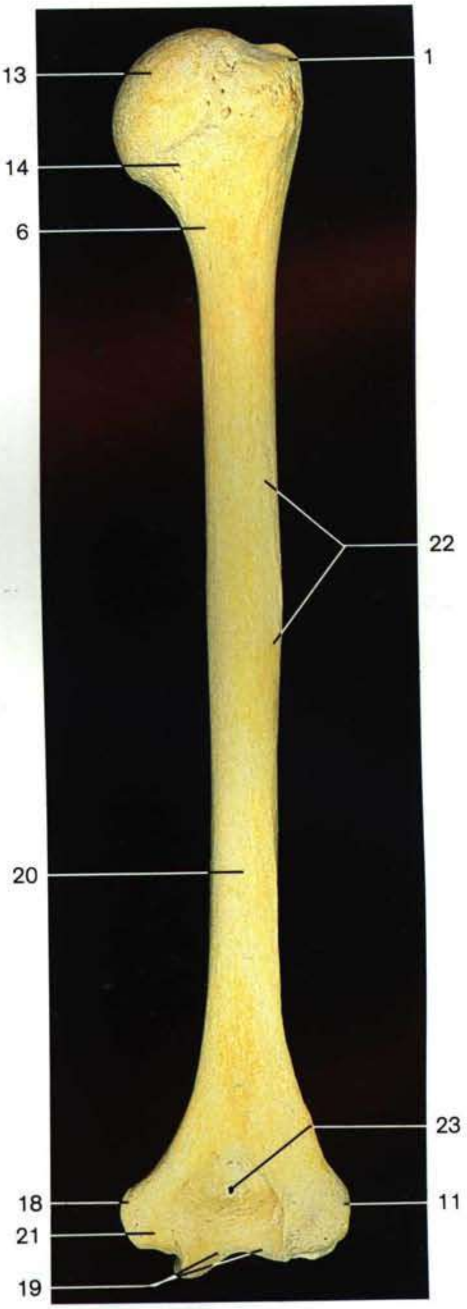
13
14
6
20
18
21
19
1
22
23
11

Right humerus (posterior aspect).

Humerus

1	Greater tubercle
2	Lesser tubercle
3	Crest of lesser tubercle
4	Crest of greater tubercle
5	Intertubercular sulcus
6	Surgical neck
7	Deltoid tuberosity
8	Anterolateral surface
9	Lateral supracondylar ridge
10	Radial fossa
11	Lateral epicondyle
12	Capitulum
13	Head
14	Anatomical neck
15	Anteromedial surface
16	Medial supracondylar ridge
17	Coronoid fossa
18	Medial epicondyle
19	Trochlea
20	Posterior surface
21	Groove for ulnar nerve
22	Groove for radial nerve
23	Olecranon fossa

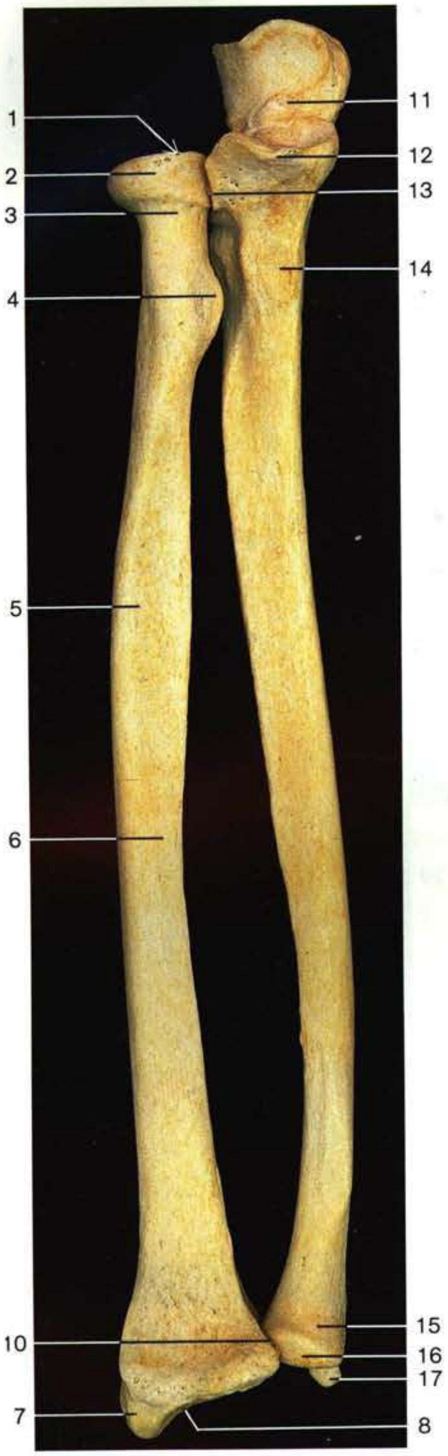

Bones of right forearm, radius and ulna (anterior aspect).

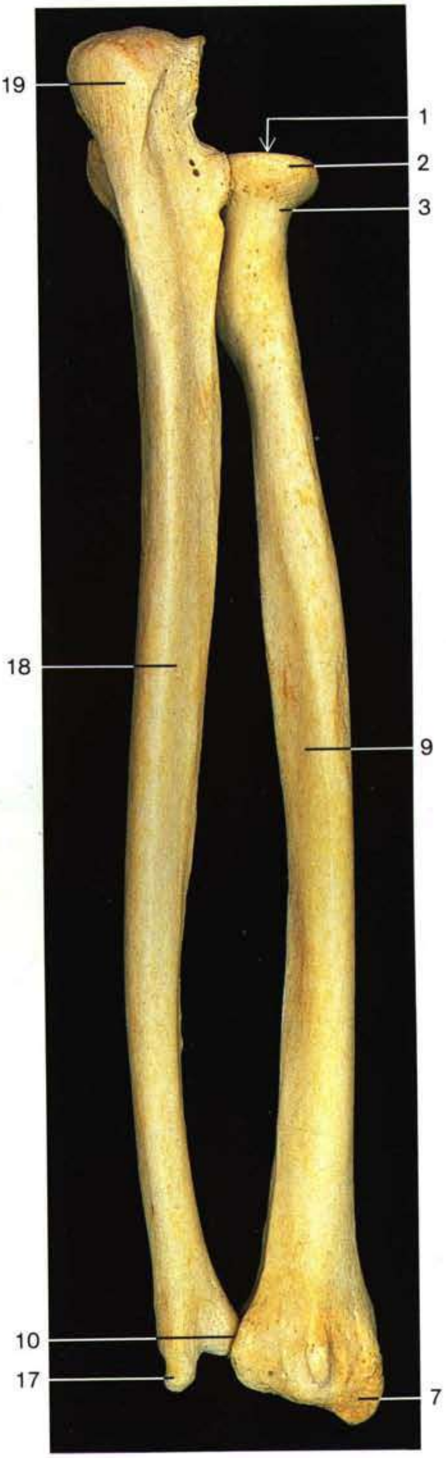

Bones of right forearm, radius and ulna (posterior aspect).

Radius

1	Head
2	Articular circumference
3	Neck
4	Radial tuberosity
5	Shaft
6	Anterior surface
7	Styloid process
8	Articular surface
9	Posterior surface
10	Ulnar notch

Ulna

11	Trochlear notch
12	Coronoid process
13	Radial notch
14	Ulnar tuberosity
15	Head
16	Articular circumference
17	Styloid process
18	Posterior surface
19	Olecranon

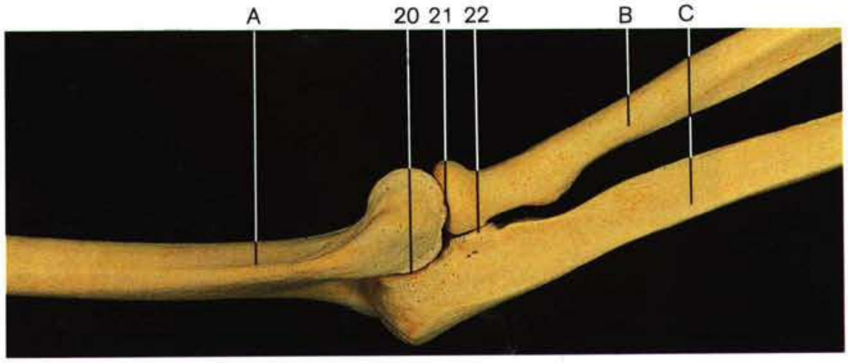

Bones of right elbow joint (lateral aspect).

Articulations at the right elbow

20	Site of humeroulnar joint
21	Site of humeroradial joint
22	Site of proximal radioulnar joint

A	Humerus
B	Radius
C	Ulna

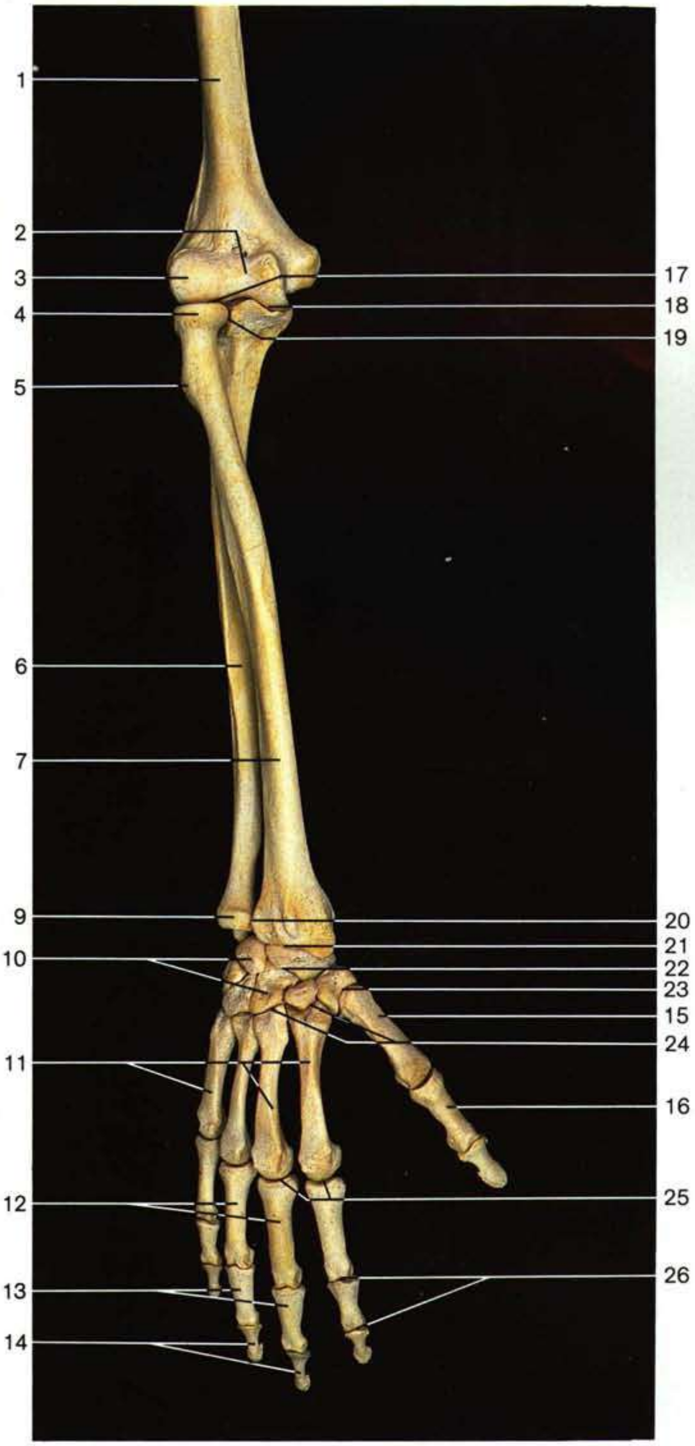
1
2
3
17
4
18
19
5
6
7
9
20
10
21
22
23
15
24
11
16
12
25
26
13
14

Skeleton of right forearm and hand in pronation.

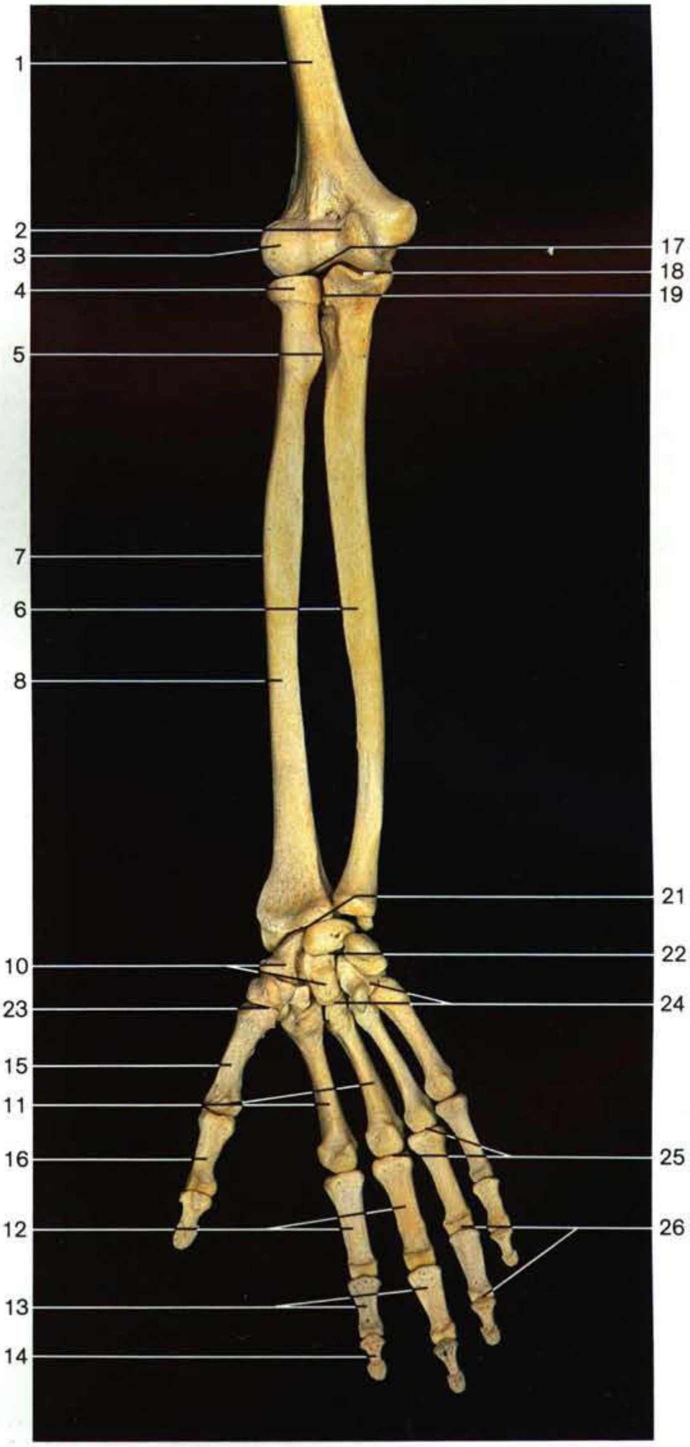
1
2
17
3
18
4
19
5
7
6
8
21
10
22
23
24
15
11
16
25
12
26
13
14

Skeleton of right forearm and hand in supination.

1 Humerus
2 Trochlea of humerus
3 Capitulum of humerus
4 Articular circumference of radius
5 Radial tuberosity
6 Anterior surface of **ulna**
7 Posterior surface of **radius**
8 Anterior surface of radius
9 Articular circumference of ulna
10 **Carpal bones**

11 **Metacarpal bones**
12 Proximal phalanges
13 Middle phalanges
14 Distal phalanges
15 Metacarpal bone of thumb
16 Proximal phalanx of thumb

Sites of joints

17 Humeroradial joint
18 Humeroulnar joint
19 Proximal radioulnar joint
20 Distal radioulnar joint
21 Radiocarpal joint
22 Midcarpal joint
23 Carpometacarpal joint of thumb
24 Carpometacarpal joints
25 Metacarpophalangeal joints
26 Interphalangeal joints of fingers

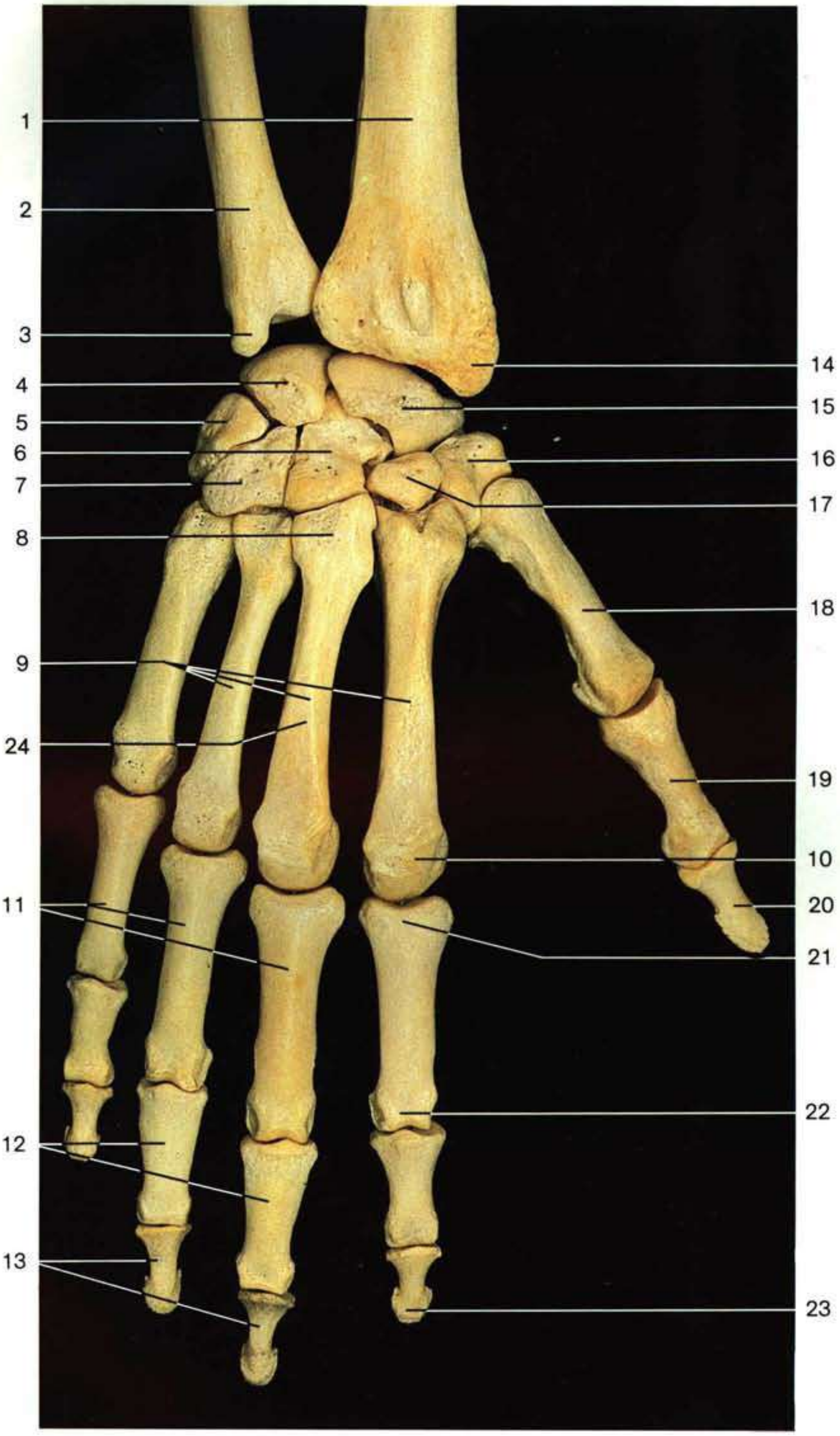
1
2
3
4
5
6
7
8
9
24
11
12
13
14
15
16
17
18
19
10
20
21
22
23

Skeleton of right wrist and hand (dorsal aspect).

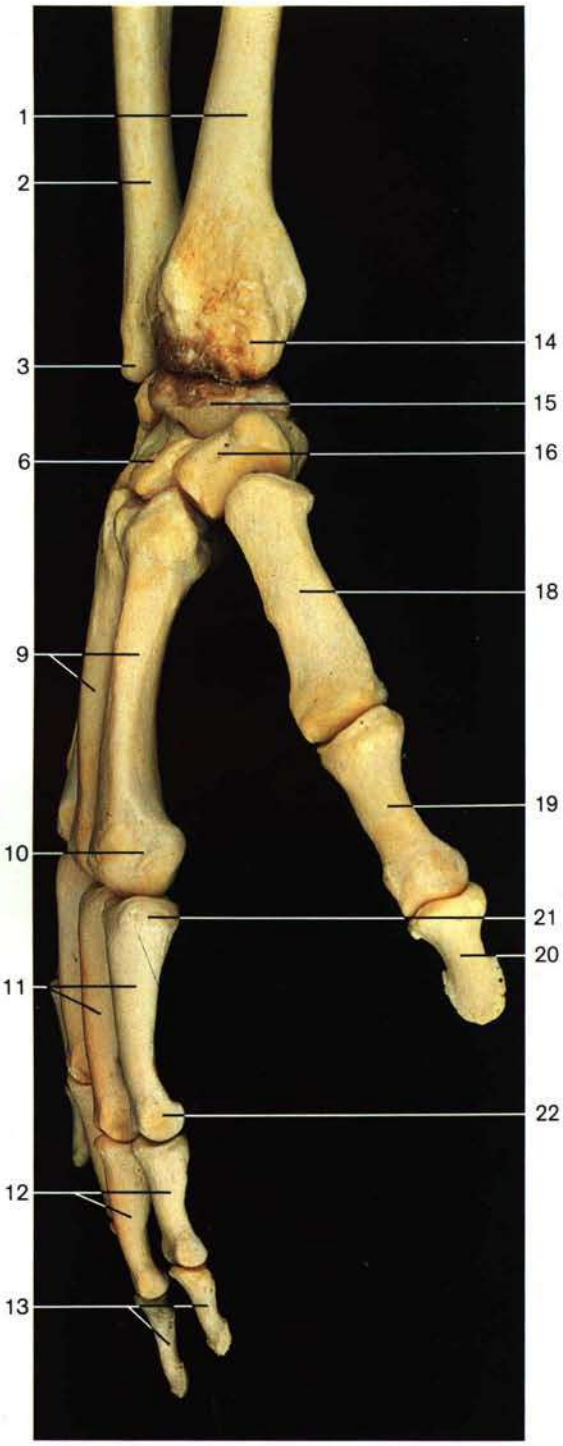
1
2
3
6
9
10
11
12
13
14
15
16
18
19
21
20
22

Skeleton of right wrist and hand (medial aspect).

1 **Radius**
2 **Ulna**
3 Styloid process of ulna
4 Lunate bone
5 Triquetral bone
6 Capitate bone
7 Hamate bone
Carpal bones
8 Base of third metacarpal bone
9 **Metacarpal bones**
10 Head of metacarpal bone
11 Proximal phalanges of hand
12 Middle phalanges of hand
13 Distal phalanges of hand
14 Styloid process of radius
15 Scaphoid bone
16 The trapezium
17 Trapezoid bone
Carpal bones
18 Metacarpal bone of thumb
19 Proximal phalanx of thumb
20 Distal phalanx of thumb
21 Base of second proximal phalanx
22 Head of second proximal phalanx
23 Tuberosity of distal phalanx
24 Body of third metacarpal bone

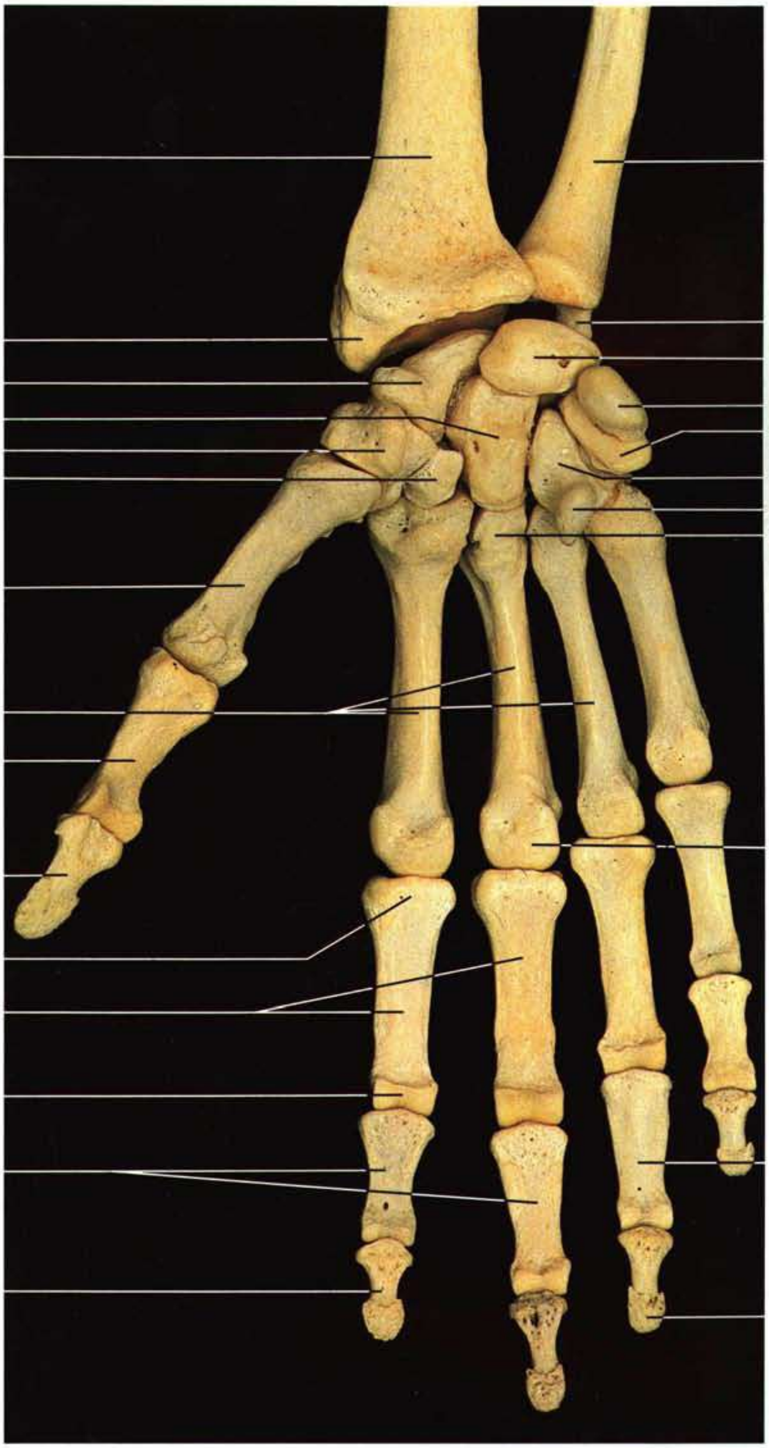

1	Radius	
2	Styloid process of radius	
3	Scaphoid bone	**Carpal bones**
4	Capitate bone	
5	Trapezium	
6	Trapezoid bone	
7	First metacarpal bone	
8	Second to fourth metacarpal bones	
9	Proximal phalanx of thumb	
10	Distal phalanx of thumb	
11	Base of second proximal phalanx	
12	Proximal phalanges	
13	Head of second proximal phalanx	
14	Middle phalanges	
15	Distal phalanx	
16	Ulna	
17	Styloid process of ulna	
18	Lunate bone	**Carpal bones**
19	Pisiform bone	
20	Triquetral bone	
21	Hamate bone	
22	Hamulus or hook of hamate bone	
23	Base of third metacarpal bone	
24	Head of metacarpal bone	
25	Tuberosity of distal phalanx	

Skeleton of right wrist and hand (palmar aspect).

The human hand is one of the most admirable structures of the human body. The carpometacarpal joint of the thumb, a saddle joint, enjoys wide mobility, so that the thumb can get in contact with all other fingers, thus enabling the hand to become an instrument for grasping and psychologic expression. During evolution these newly developed functions appeared after the erect posture of the human body was achieved. An inevitable prerequisite for the development of human cultures is not only the differentiation of the brain but also the development of an organ capable of realizing its ideas: the human hand.

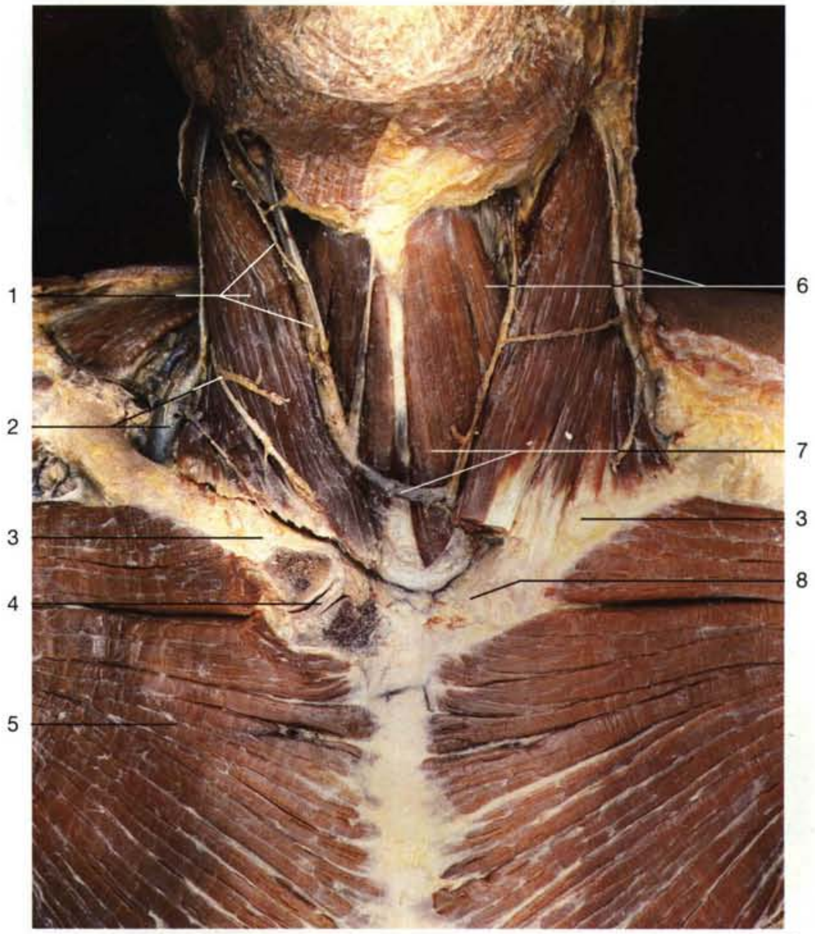

Right sternoclavicular joint (anterior aspect). On the right side the joint has been opened by a coronal section. Note the articular disk.

1 Sternocleidomastoid muscle, cervical branch of facial nerve and anterior jugular vein
2 External jugular vein and transverse cervical nerve (inferior branch)
3 Clavicle
4 **Sternoclavicular joint** (opened) with articular disk
5 Pectoralis major muscle
6 Omohyoid muscle and external jugular vein
7 Jugular venous arch and sternohyoid muscle
8 Sternoclavicular joint (not opened)
9 Acromial end of clavicle
10 **Acromioclavicular joint**
11 Acromion
12 Tendon of supraspinatus muscle (attached to the articular capsule)
13 Coraco-acromial ligament
14 Tendon of long head of biceps brachii muscle
15 Tendon of subscapularis muscle (attached to the articular capsule)
16 Intertubercular sulcus
17 Articular capsule of shoulder joint
18 Humerus
19 Trapezoid ligament
20 Coracoid process
21 Glenoid labrum
22 **Shoulder joint** (joint cavity)
23 Scapula
24 Supraspinatus muscle
25 Cartilage of glenoid cavity
26 Tendon of long head of triceps brachii muscle
27 Head of humerus (articular cartilage)
28 Epiphyseal line

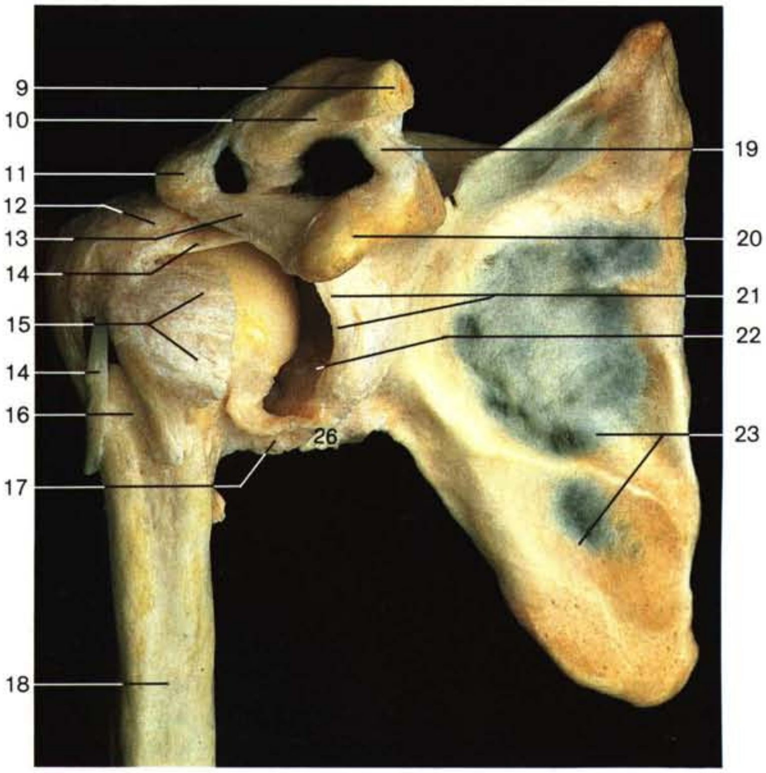

Right shoulder joint. The anterior part of the articular capsule has been removed and the head of the humerus has been slightly rotated outward to show the cavity of the joint.

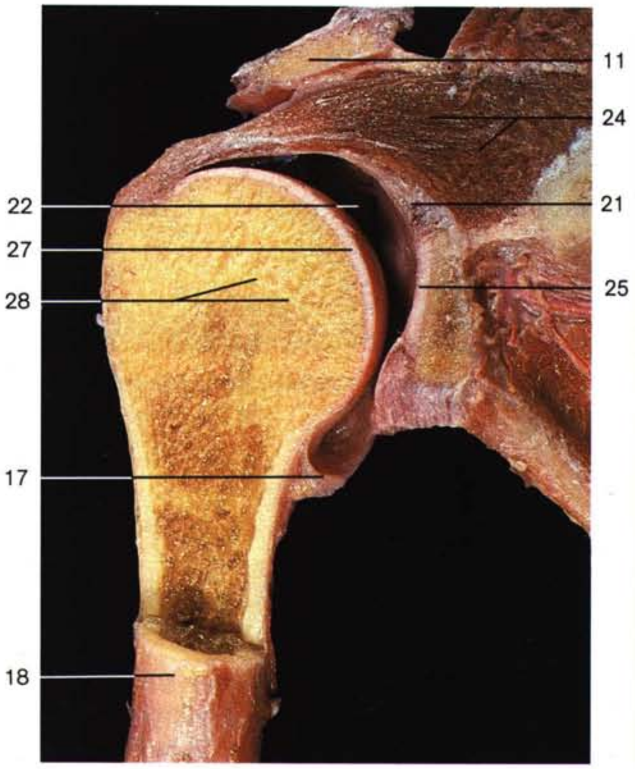

Coronal section of the right shoulder joint (anterior aspect).

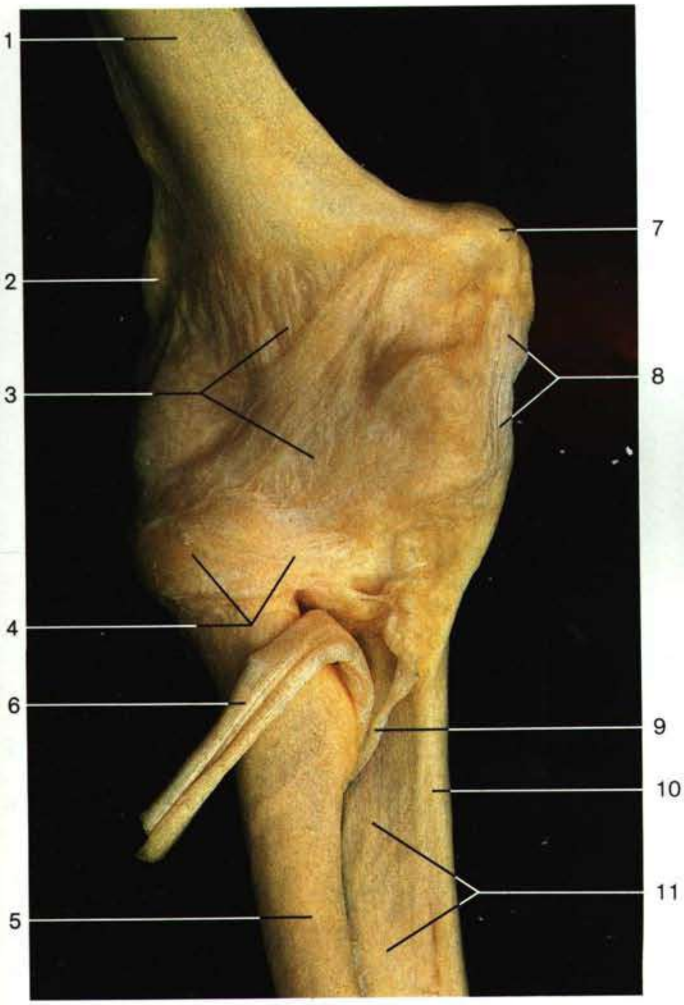

Ligaments of elbow joint (anterior aspect).

1 Humerus
2 Lateral epicondyle of humerus
3 Articular capsule
4 Annular ligament of proximal radioulnar joint
5 Radius
6 Tendon of biceps brachii muscle
7 Medial epicondyle of humerus
8 Ulnar collateral ligament
9 Oblique cord
10 Ulna
11 Interosseous membrane
12 Radial fossa
13 Capitulum of humerus
14 Head of radius
15 Radial collateral ligament
16 Coronoid fossa
17 Trochlea of humerus
18 Coronoid process of ulna
19 Olecranon
20 Radial tuberosity

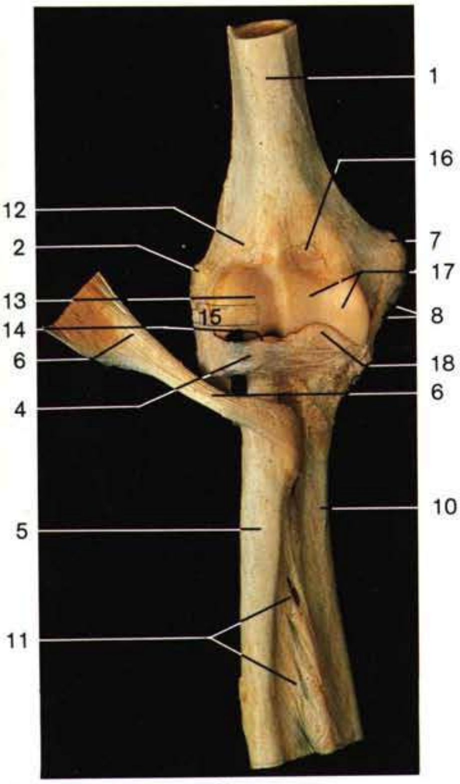

Elbow joint with ligaments (anterior aspect). Articular capsule has been removed to show the annular ligament.

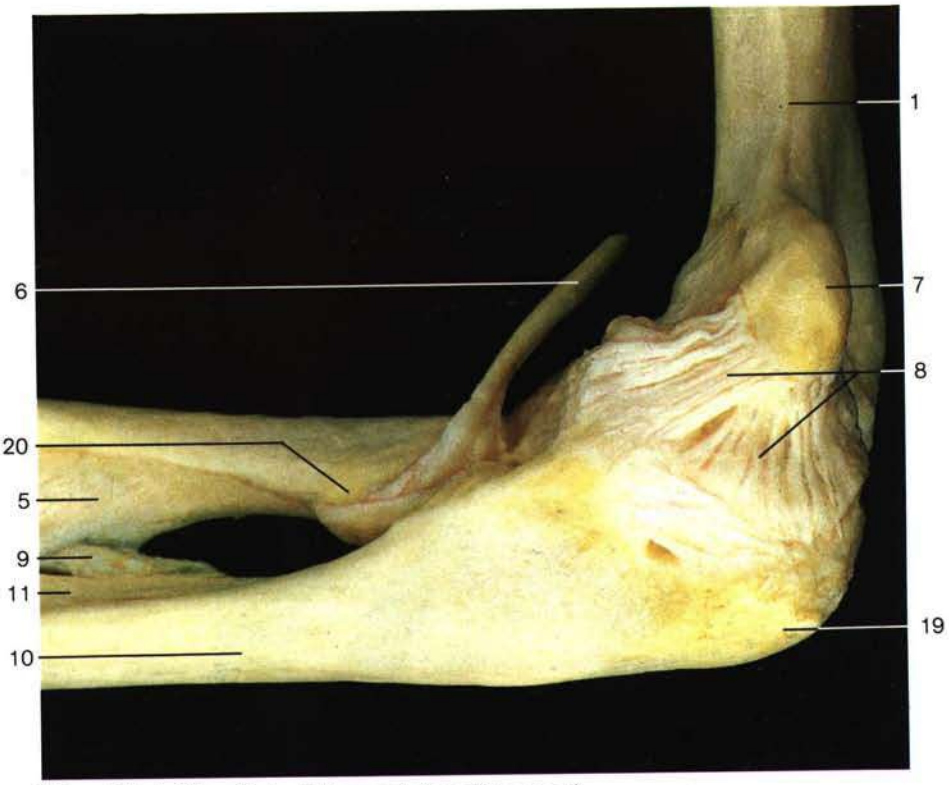

Elbow joint with collateral ligaments (medial aspect).

1 Ulna
2 Exostosis (pathological)
3 Head of ulna
4 **Ulnar carpal collateral ligament**
5 Deep intercarpal ligaments
6 Dorsal carpometacarpal ligaments
7 Dorsal metacarpal ligaments
8 Interosseous membrane
9 Radius
10 Styloid process of radius
11 **Dorsal radiocarpal ligament**
12 Radial collateral ligament
13 Articular capsule and dorsal intercarpal ligaments
14 **Palmar radiocarpal ligament**
15 Tendon of flexor carpi radialis muscle (cut)
16 Radiating carpal ligament
17 Palmar carpometacarpal ligaments
18 1st metacarpal bone
19 Palmar ulnocarpal ligament
20 Tendon of flexor carpi ulnaris muscle (cut)
21 **Pisohamate ligament**
22 **Pisometacarpal ligament**
23 Palmar metacarpal ligaments
24 5th metacarpal bone

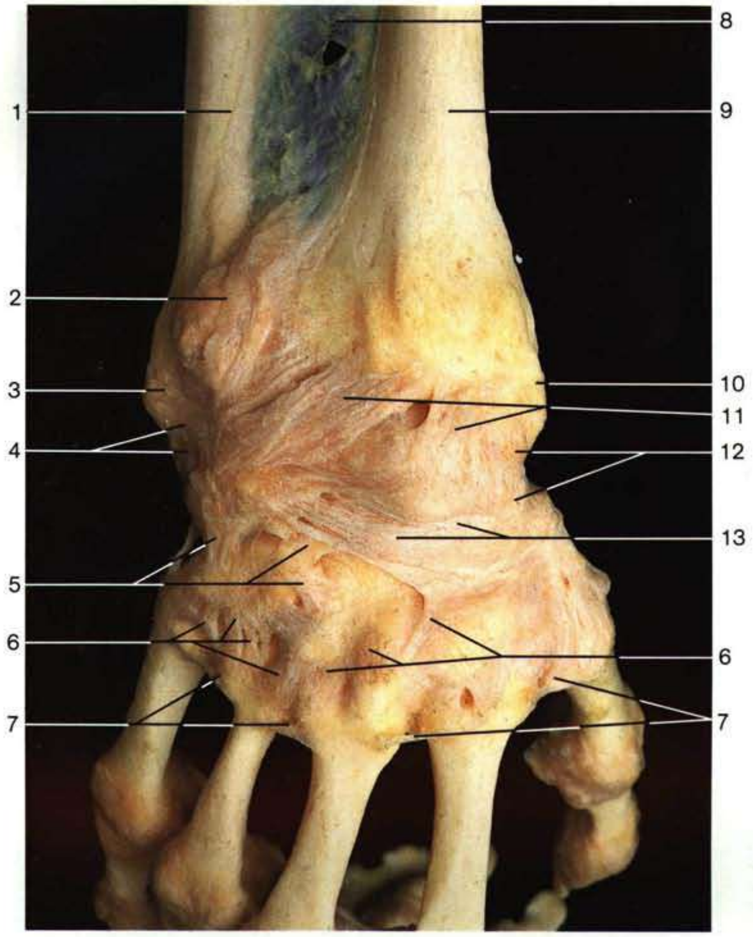

Ligaments of hand and wrist (dorsal aspect).

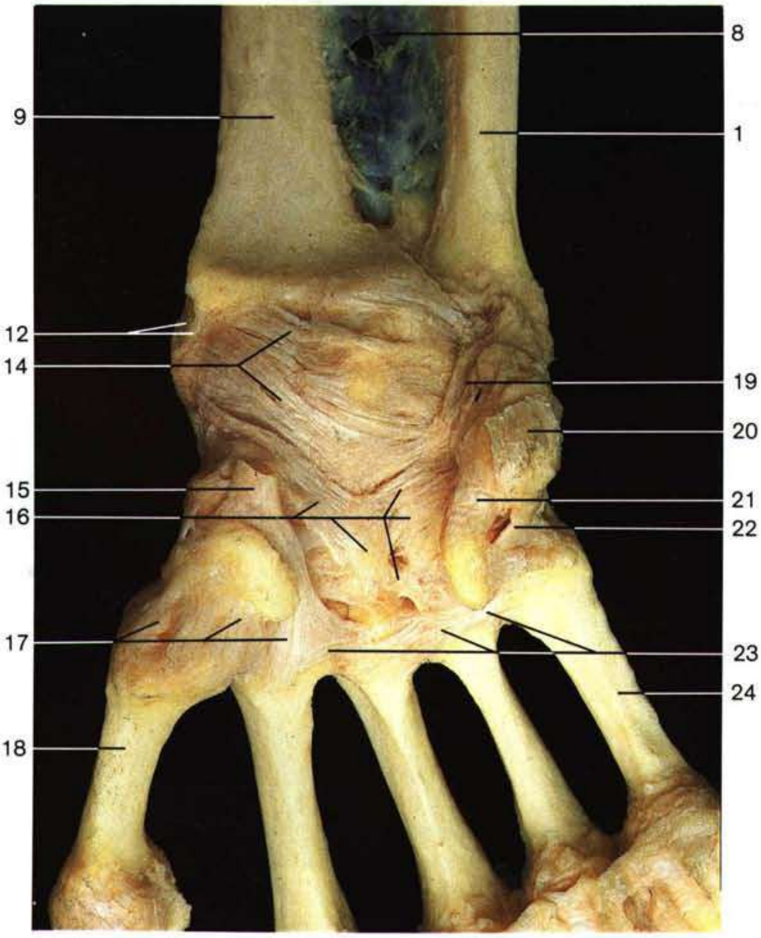

Ligaments of hand and wrist (palmar aspect).

1 Radius
2 Styloid process of radius
3 Palmar radiocarpal ligament
4 Tendon of flexor carpi radialis muscle (cut)
5 Radiating carpal ligament
6 Articular capsule of **carpometacarpal joint of thumb**
7 Articular capsule of **metacarpophalangeal joint of thumb**
8 Palmar ligaments and articular capsule of metacarpophalangeal joints
9 Palmar ligaments and articular capsule of interphalangeal joints
10 Articular capsule
11 Interosseous membrane
12 Ulna
13 **Distal radioulnar joint**
14 Styloid process of ulna
15 Palmar ulnocarpal ligament
16 **Pisiform bone with tendon of flexor carpi ulnaris muscle**
17 Pisometacarpal ligament
18 Pisohamate ligament
19 Metacarpal bone
20 Deep transverse metacarpal ligament
21 Tendons of extensor muscles and articular capsule
22 Collateral ligament of **interphalangeal joint**
23 Collateral ligaments of **metacarpophalangeal joints**
24 Second metacarpal bone

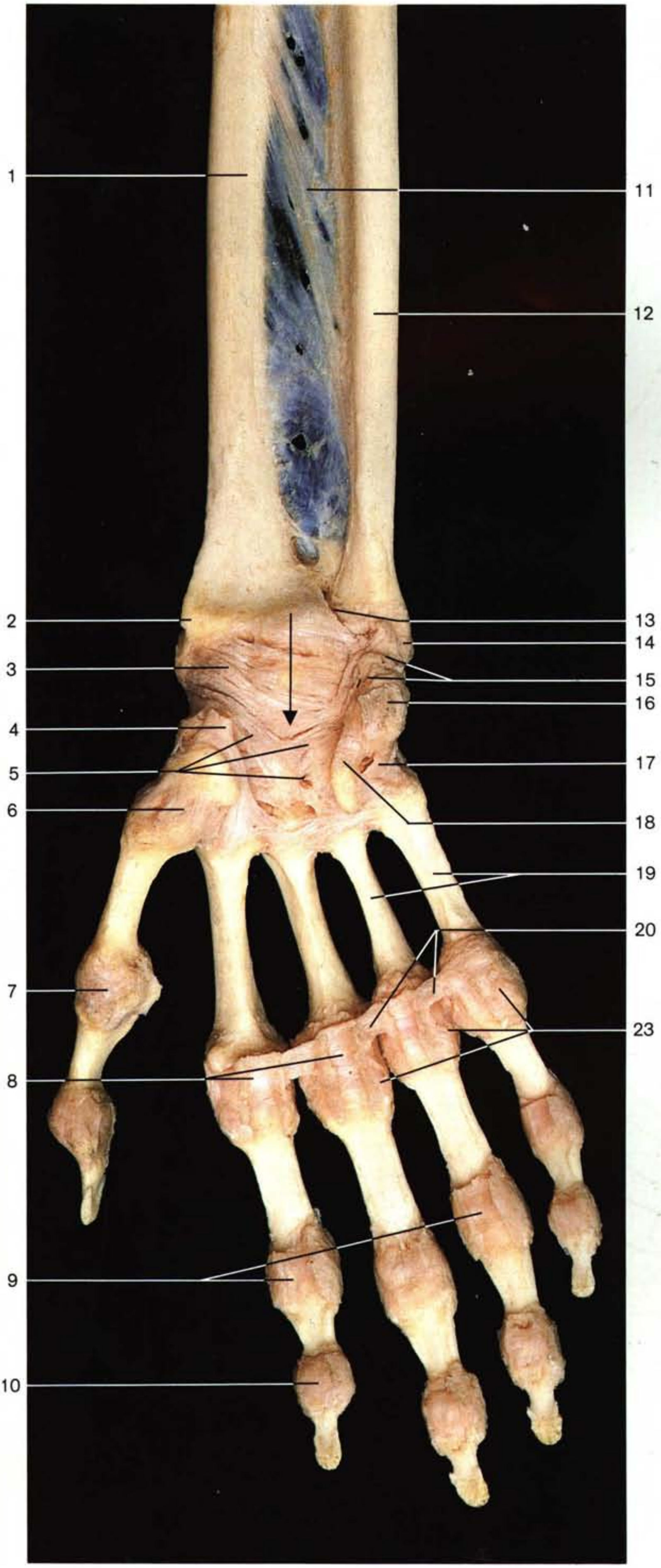

Ligaments of right forearm, hand and fingers (palmar aspect). The arrow indicates the location of the carpal tunnel.

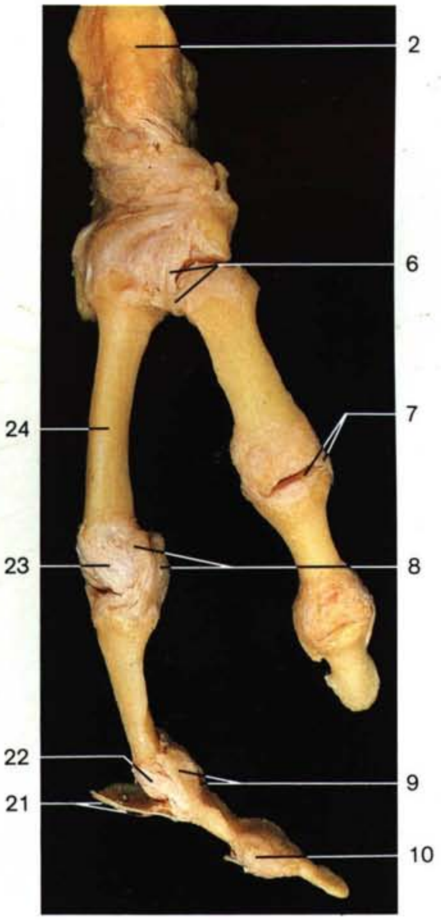

Ligaments of fingers (lateral aspect).

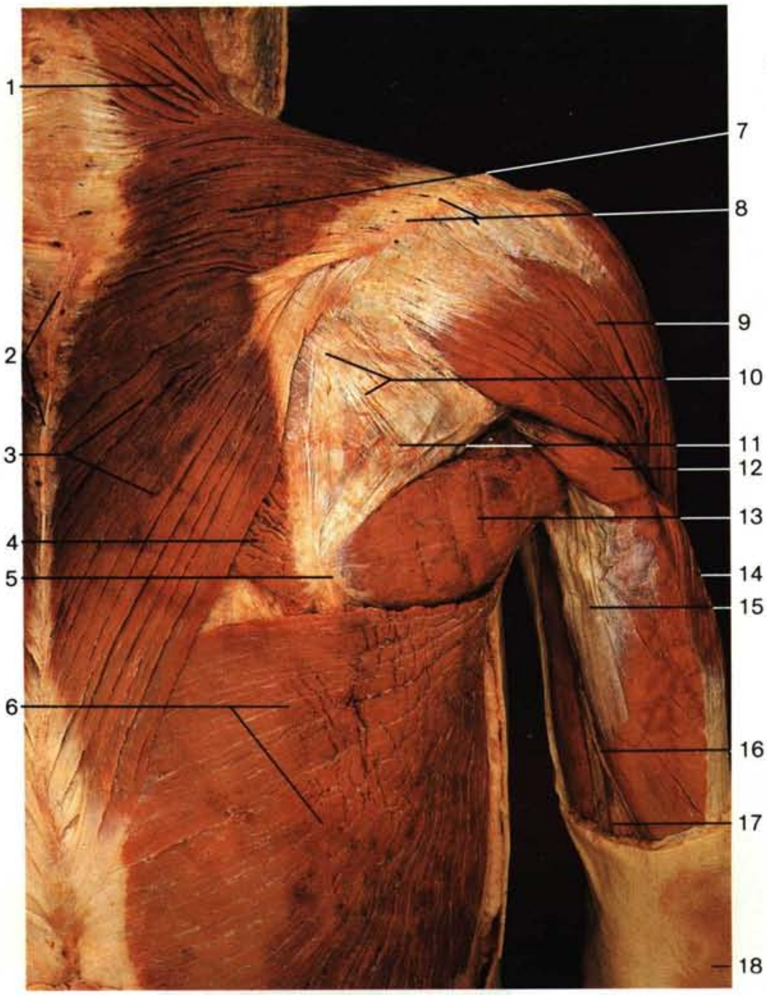

1 Descending fibers of **trapezius muscle**
2 Spinous processes of thoracic vertebrae
3 Ascending fibers of trapezius muscle
4 Rhomboid major muscle
5 Inferior angle of scapula
6 **Latissimus dorsi muscle**
7 Transverse fibers of trapezius muscle
8 Spine of scapula
9 Posterior fibers of **deltoid muscle**
10 Infraspinatus muscle and infraspinous fascia
11 Teres minor muscle and fascia
12 Long head of triceps brachii muscle
13 **Teres major muscle**
14 Lateral head of triceps brachii muscle
15 Medial head of triceps brachii muscle
16 Medial intermuscular septum
17 Ulnar nerve
18 Olecranon

Muscles of shoulder and arm, superficial layer (dorsal aspect).

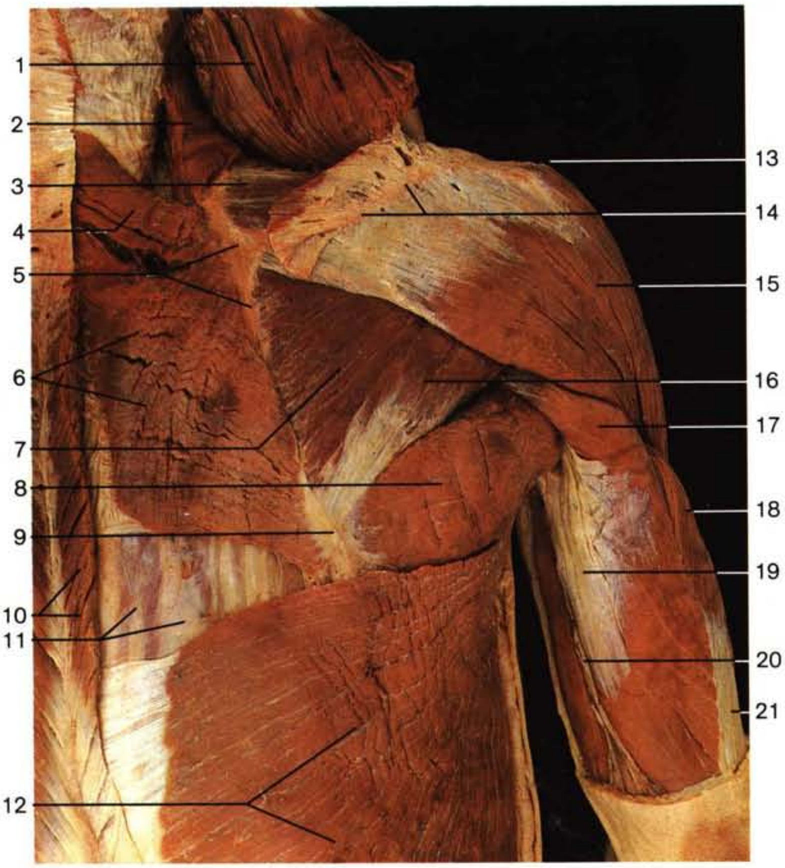

1 Trapezius muscle (reflected)
2 Levator scapulae muscle
3 Supraspinatus muscle
4 Rhomboid minor muscle
5 Medial border of scapula
6 **Rhomboid major muscle**
7 **Infraspinatus muscle**
8 **Teres major muscle**
9 Inferior angle of scapula
10 Cut edge of trapezius muscle
11 Intrinsic muscles of back with fascia
12 **Latissimus dorsi muscle**
13 Acromion
14 Spine of scapula
15 **Deltoid muscle**
16 Teres minor
17 Long head of triceps brachii muscle
18 Lateral head of triceps brachii muscle
19 Medial head of triceps brachii muscle
20 Medial intermuscular septum
21 Tendon of triceps brachii muscle

Muscles of shoulder and arm, deeper layer (right side, dorsal aspect). The trapezius has been cut near its origin at the vertebral column and reflected upward.

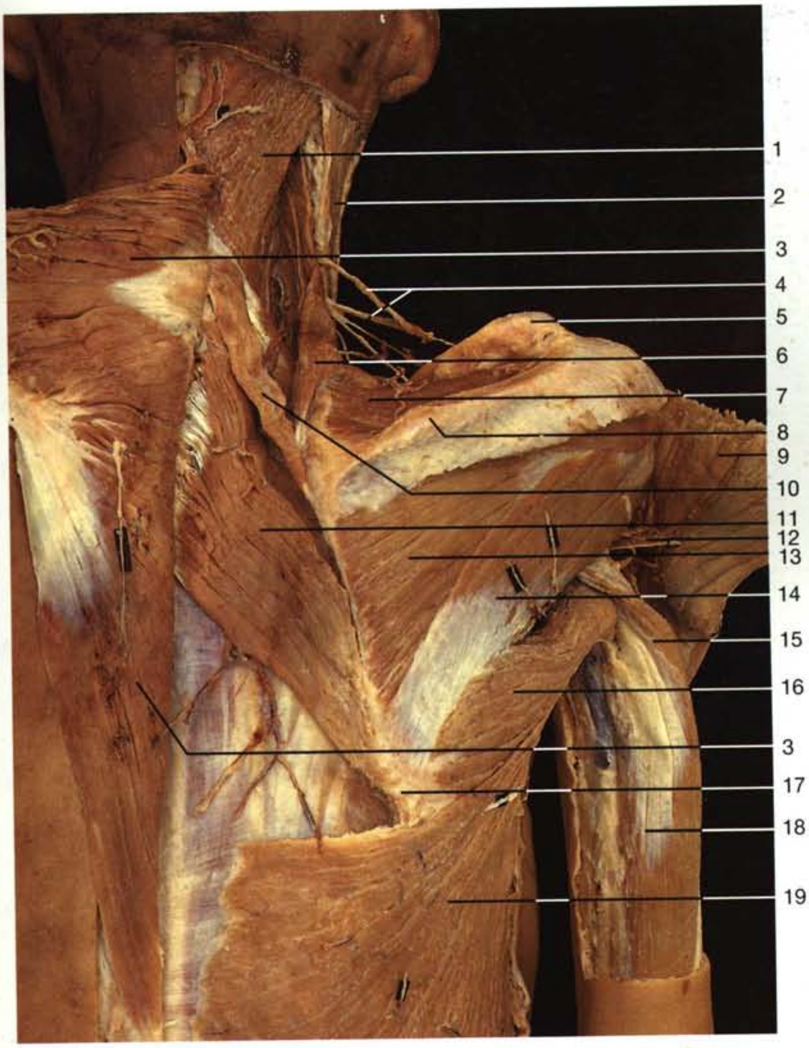

1 **Splenius capitis muscle**
2 Sternocleidomastoideus muscle
3 Trapezius muscle (reflected)
4 **Lateral supraclavicular nerves**
5 Clavicle
6 Levator scapulae muscle
7 Supraspinatus muscle
8 Spine of scapula
9 Deltoid muscle (reflected)
10 Rhomboid minor muscle
11 **Rhomboid major muscle**
12 **Axillary nerve and posterior circumflex humeral artery**
13 **Infraspinatus muscle**
14 **Teres minor muscle**
15 Long head of triceps brachii muscle
16 **Teres major muscle**
17 Inferior angle of scapula
18 Triceps brachii muscle
19 **Latissimus dorsi muscle**

Muscles of shoulder and arm, deeper layer, right side (dorsal aspect). The trapezius and deltoid muscles have been divided and reflected.

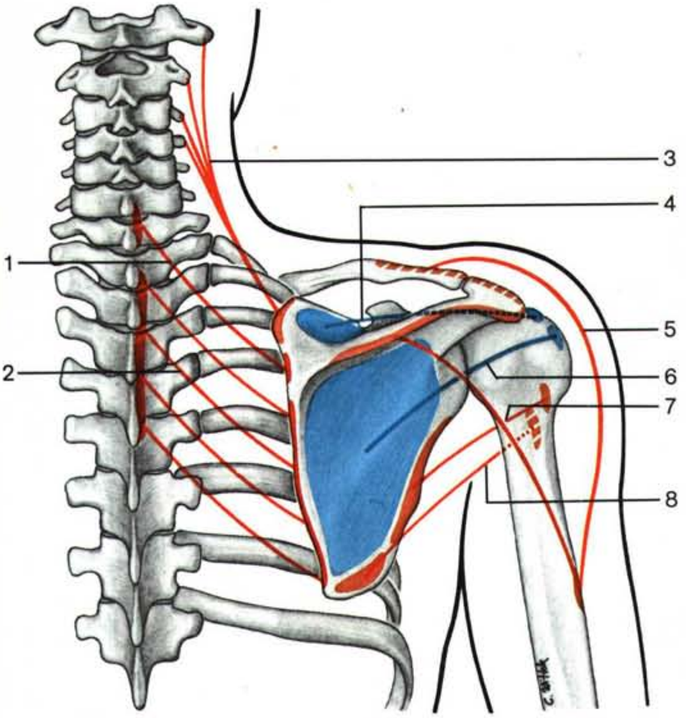

Shoulder muscles, schematic diagram illustrating the course of the main muscles of the dorsal aspect of the shoulder.

1 Rhomboid minor muscle (red)
2 Rhomboid major muscle (red)
3 Levator scapulae muscle (red)
4 Supraspinatus muscle (blue)
5 Deltoid muscle (red)
6 Infraspinatus muscle (blue)
7 Teres minor muscle (red)
8 Teres major muscle (red)

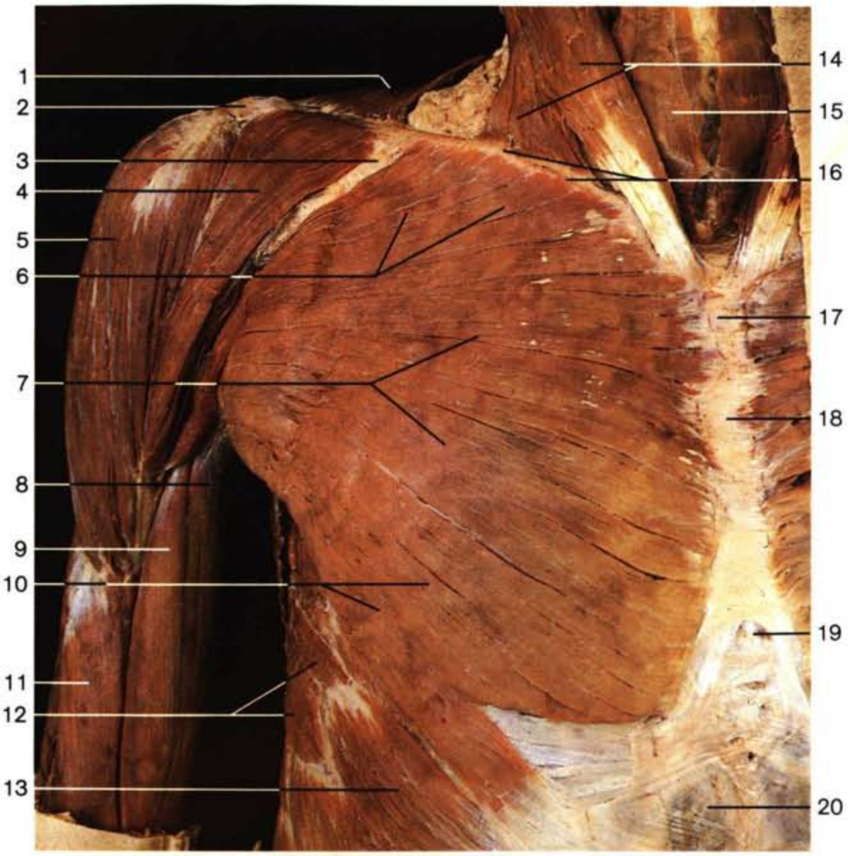

1 Trapezius muscle
2 Acromion
3 Deltopectoral triangle
4 Clavicular part of deltoid muscle (anterior fibers)
5 Acromial part of deltoid muscle (central fibers)
6 Clavicular part of pectoralis major muscle
7 Sternocostal part of pectoralis major muscle
8 Short head of biceps brachii muscle
9 Long head of biceps brachii muscle
10 Abdominal part of pectoralis major muscle
11 Brachialis muscle
12 Serratus anterior muscle
13 External abdominal oblique muscle
14 Sternocleidomastoid muscle
15 Infrahyoid muscles
16 Clavicle
17 Manubrium sterni
18 Body of sternum
19 Xiphoid process
20 Anterior layer of sheath of rectus abdominis muscle

Muscles of shoulder and arm, superficial layer (ventral aspect).

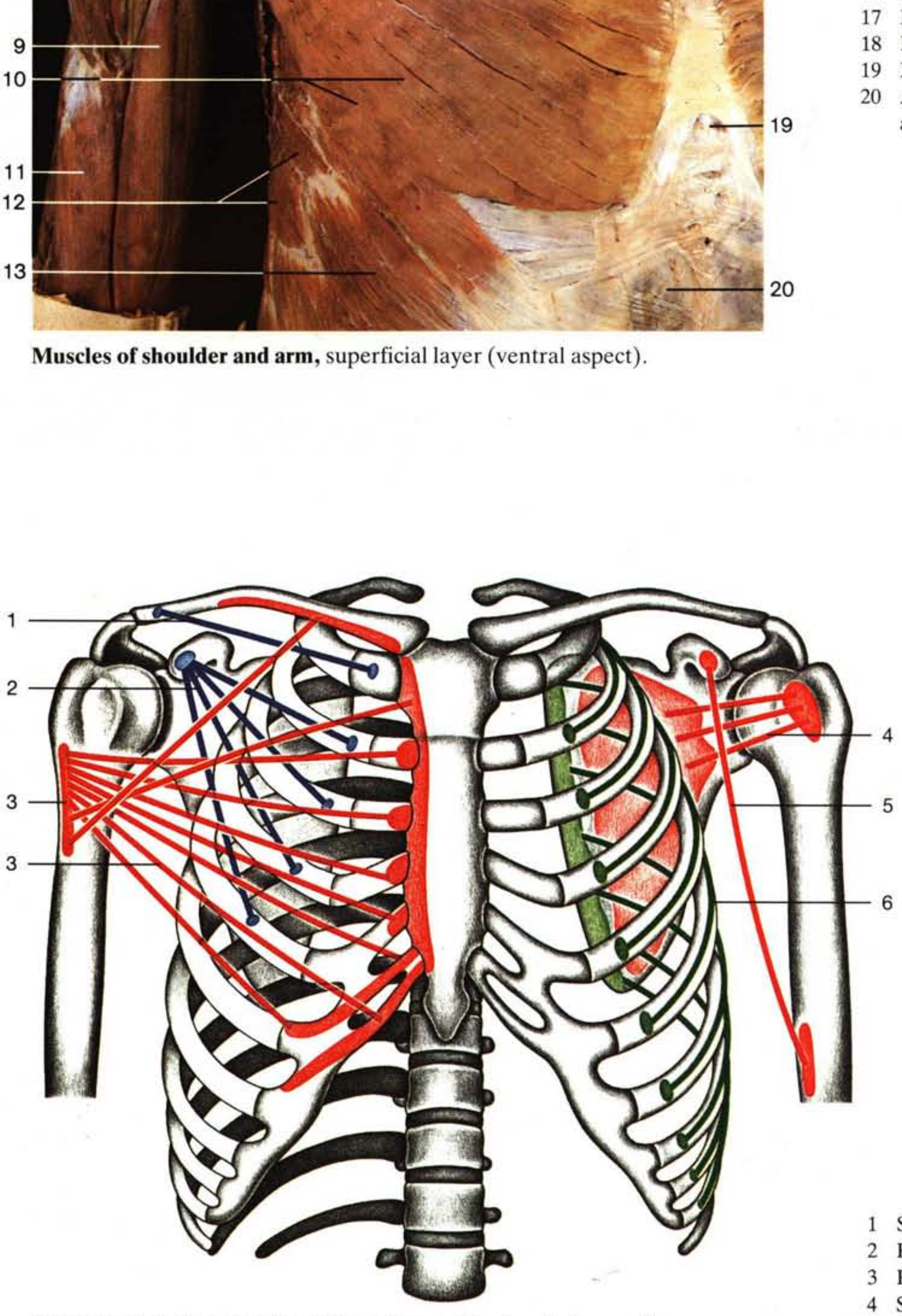

1 Subclavius muscle (blue)
2 Pectoralis minor muscle (blue)
3 Pectoralis major muscle (red)
4 Subscapularis muscle (red)
5 Coracobrachialis muscle (light-blue)
6 Serratus anterior muscle (green)

Arrangement of pectoral and shoulder muscles (ventral aspect).
(Schematic drawing.)

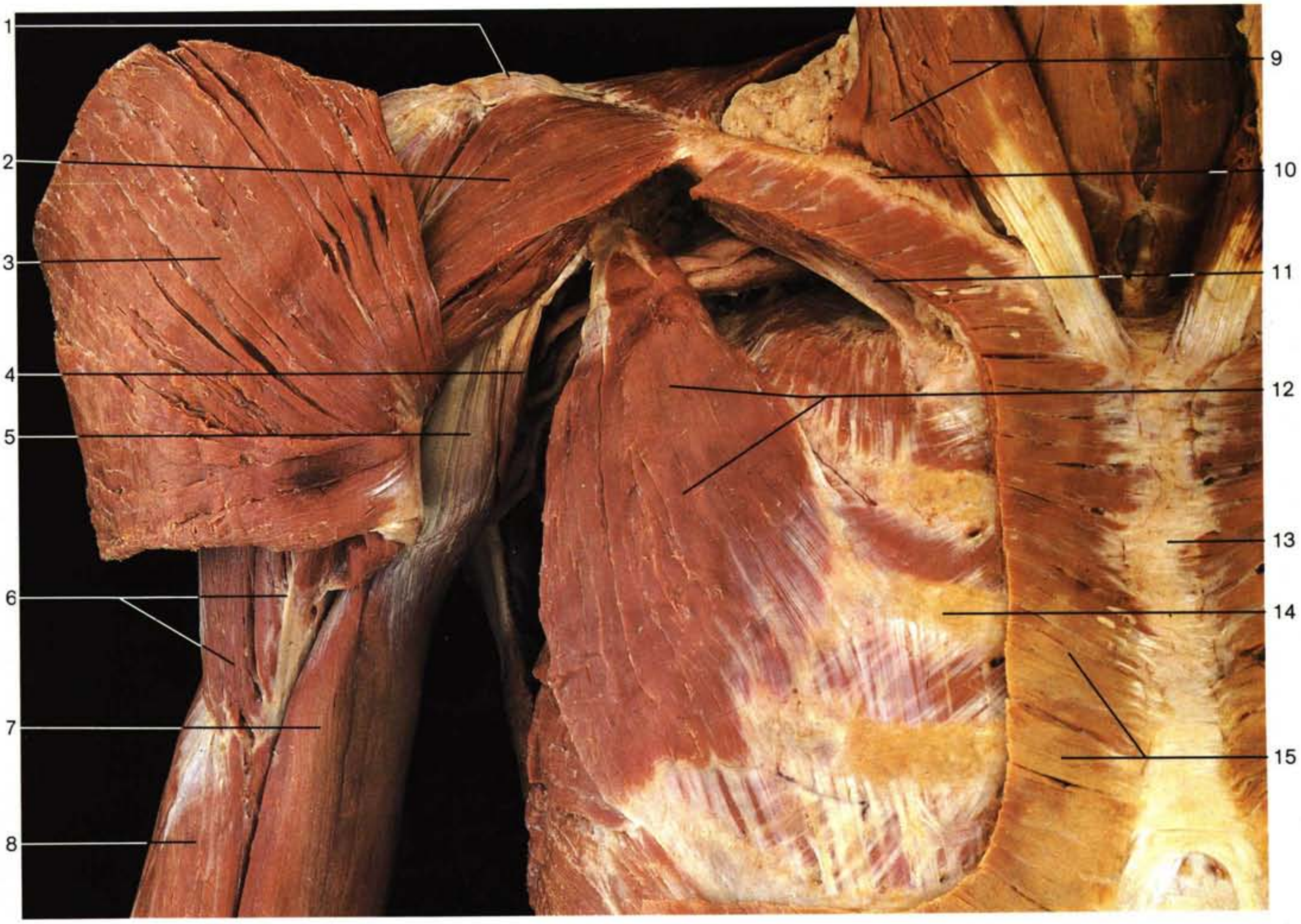

Muscles of shoulder and arm, deep layer (ventral aspect).

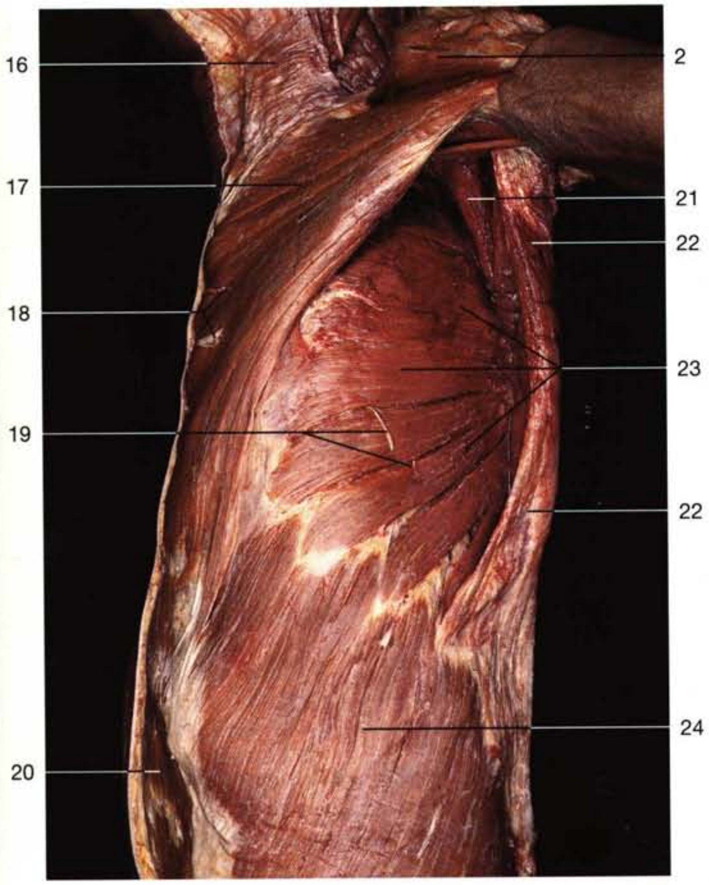

1 Acromion
2 Clavicular part of deltoid muscle
3 Pectoralis major muscle (reflected)
4 Coracobrachialis muscle
5 Short head of biceps brachii muscle
6 Deltoid muscle (insertion on humerus)
7 Long head of biceps brachii muscle
8 Brachialis muscle
9 Sternocleidomastoid muscle
10 Clavicle
11 Subclavius muscle
12 Pectoralis minor muscle
13 Sternum
14 3rd rib
15 Pectoralis major muscle
16 Platysma muscle
17 Pectoralis major muscle forming the anterior axillary fold
18 Anterior cutaneous branches of intercostal nerves
19 Lateral cutaneous branches of intercostal nerves
20 Rectus abdominis muscle
21 Subscapularis muscle
22 **Latissimus dorsi muscle** forming the posterior axillary fold
23 **Serratus anterior muscle** forming the medial wall of the axilla
24 External abdominal oblique muscle

Axillary fossa and serratus anterior muscle (left side, lateral aspect).

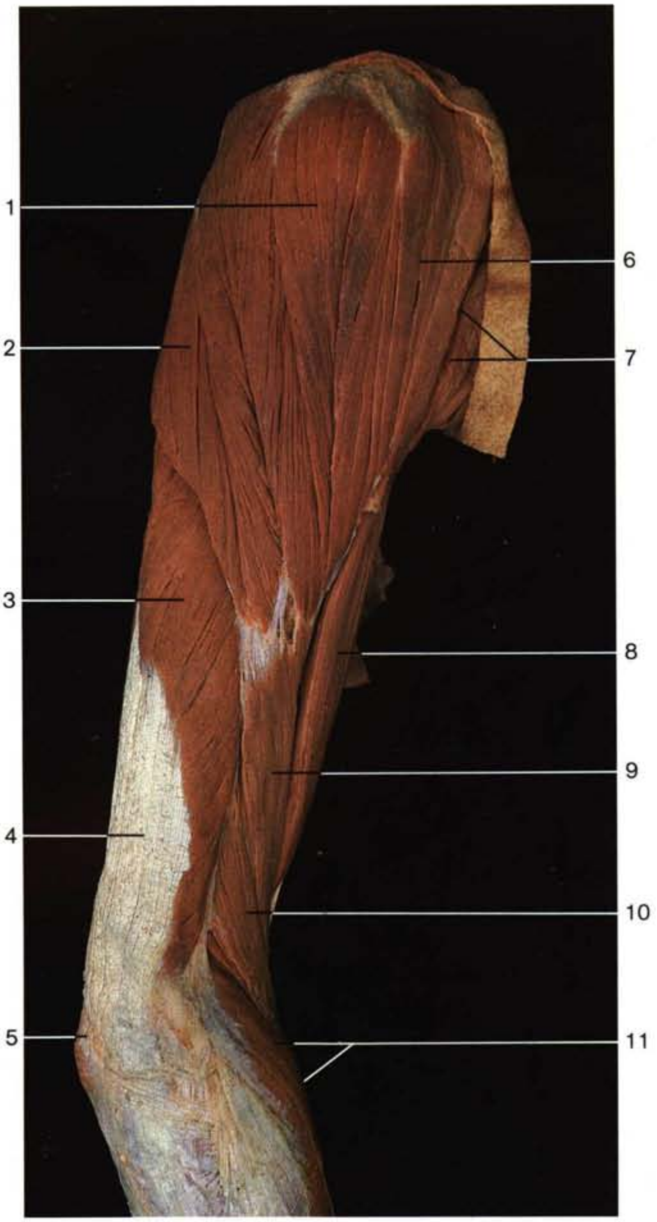
1
2
3
4
5
6
7
8
9
10
11

Muscles of right arm (lateral aspect).

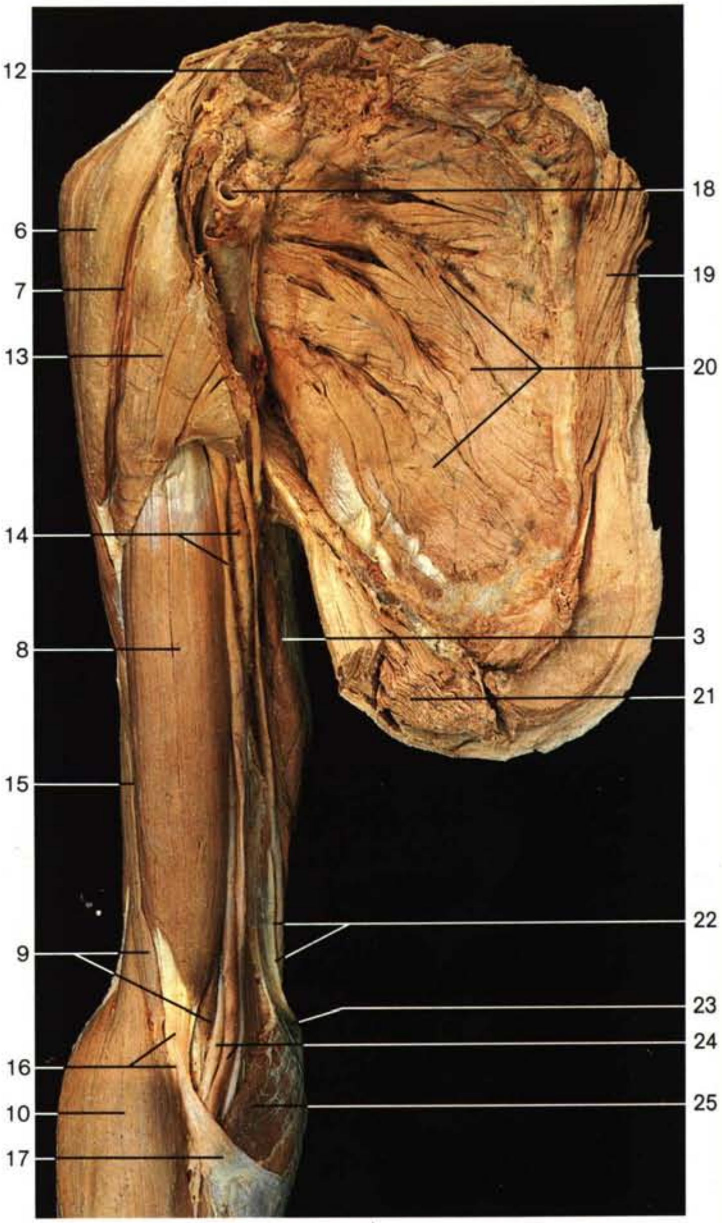
12
6
7
13
14
8
15
9
16
10
17
18
19
20
3
21
22
23
24
25

Muscles of right arm (ventral aspect). The arm with the scapula and attached muscles has been removed from the trunk.

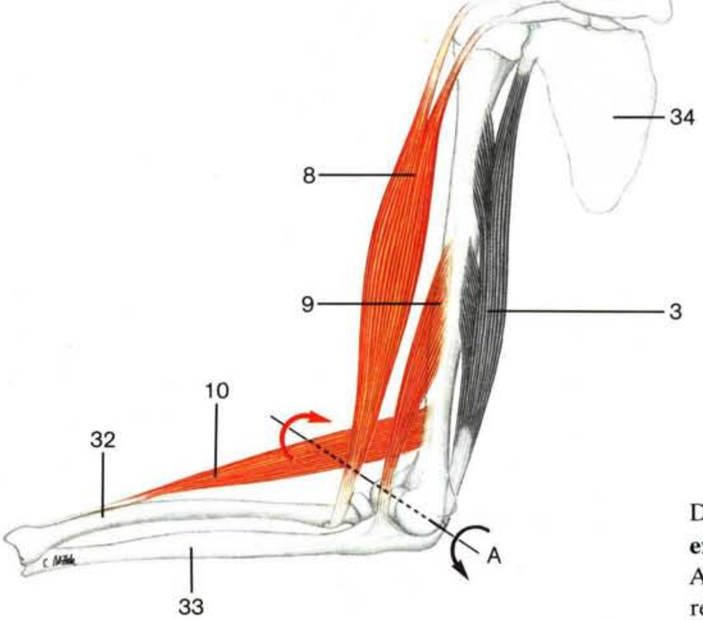
34
8
9
3
10
32
33
A

Diagram illustrating the position of the **flexor and extensor muscles of the arm** and their effect on the elbow joint. A = axis; arrows = direction of movements; red = flexion; black = extension.

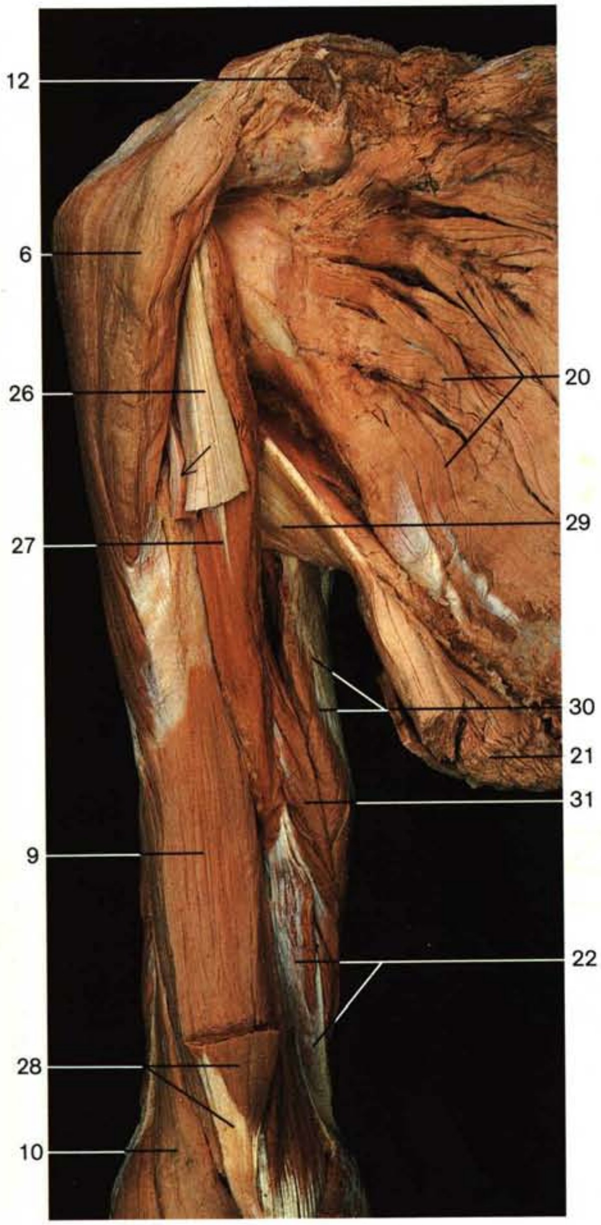

Flexor muscles of right arm (ventral aspect). Part of the biceps brachii muscle has been removed. Arrow: tendon of long head of biceps brachii muscle.

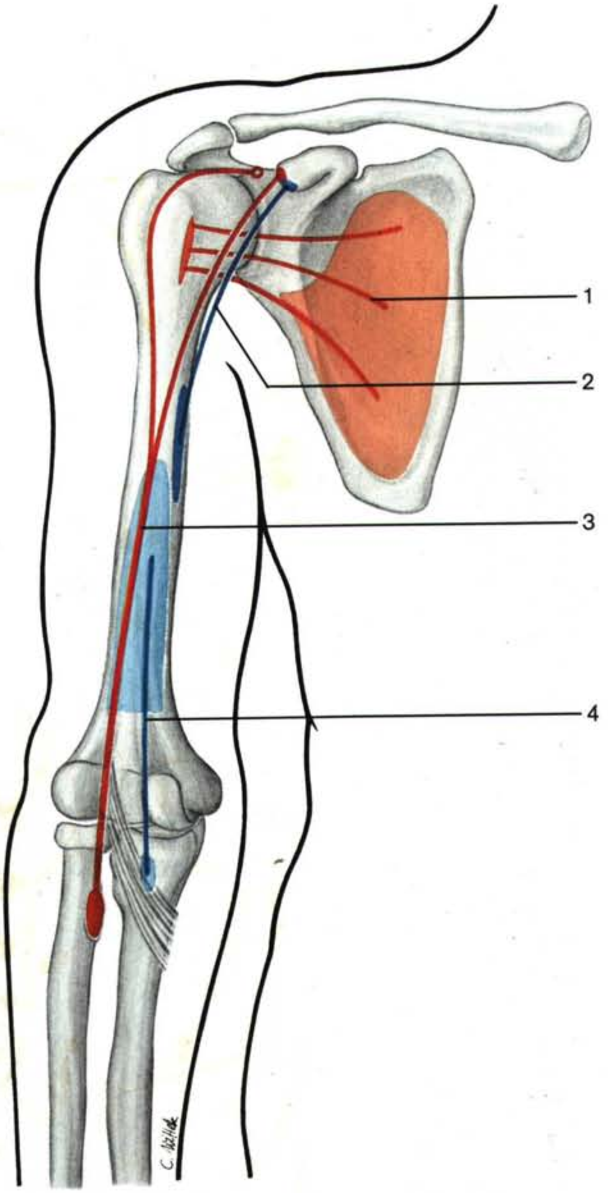

Position and course of flexors of arm. (Schematic drawing.)

1 Subscapularis muscle (red)
2 Coracobrachialis muscle (blue)
3 Biceps brachii muscle (red)
4 Brachialis muscle (blue)

1 Acromial part of deltoid muscle (central fibers)
2 Scapular part of deltoid muscle (posterior fibers)
3 Triceps brachii muscle
4 Tendon of triceps brachii muscle
5 Olecranon
6 Clavicular part of deltoid muscle (anterior fibers)
7 Deltopectoral groove
8 **Biceps brachii muscle**
9 **Brachialis muscle**
10 **Brachioradialis muscle**
11 Extensor carpi radialis longus muscle
12 Clavicle (divided)
13 Pectoralis major muscle
14 Medial intermuscular septum with vessels and nerves
15 Lateral intermuscular septum
16 Tendon of biceps brachii muscle
17 Bicipital aponeurosis
18 Axillary artery
19 Rhomboid major muscle
20 **Subscapularis muscle**
21 Latissimus dorsi muscle (divided)
22 Medial intermuscular septum
23 Medial epicondyle of humerus
24 Brachial artery and median nerve
25 Pronator teres muscle
26 Tendon of short head of biceps brachii muscle
27 Coracobrachialis muscle
28 Distal part of biceps brachii muscle
29 Teres major muscle
30 Long head of triceps brachii muscle
31 Medial head of triceps brachii muscle
32 Radius
33 Ulna
34 Scapula

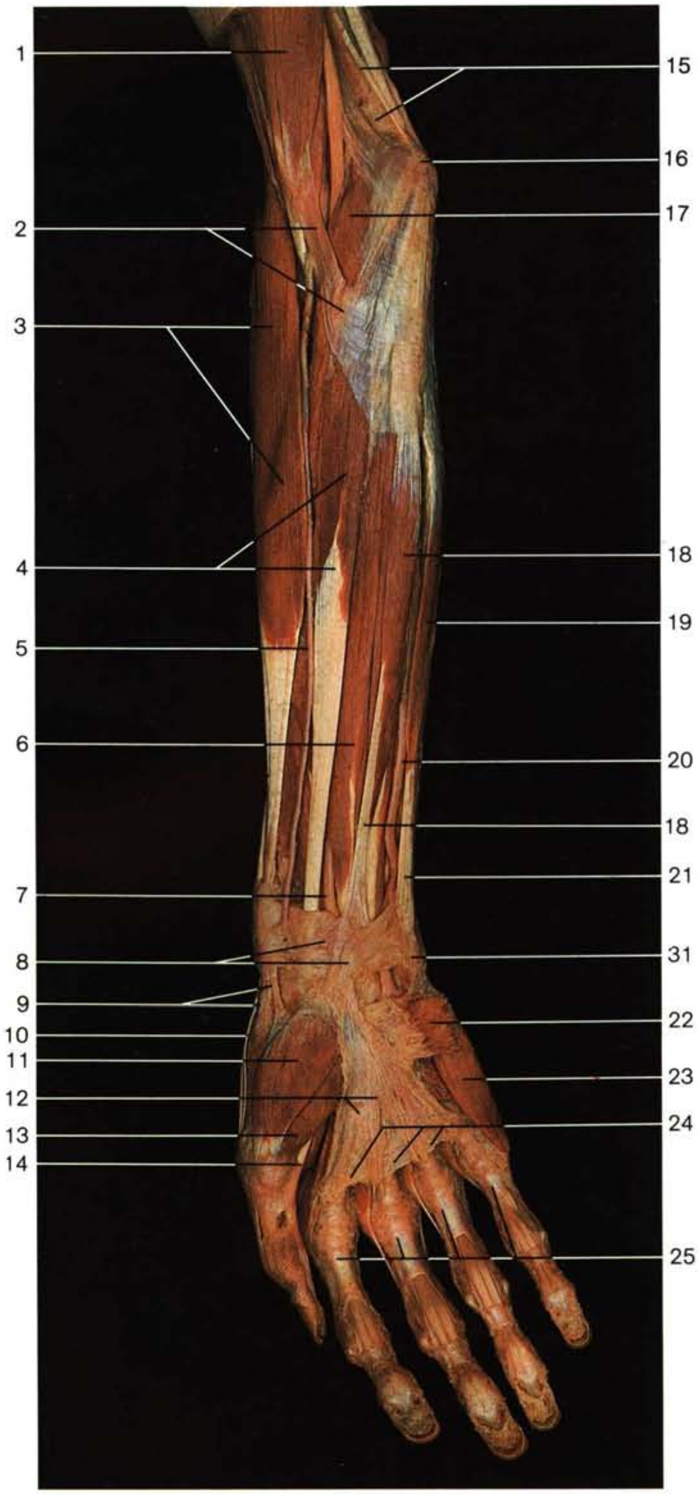
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
18
21
31
22
23
24
25

Forearm muscles, superficial layer (ventral aspect).

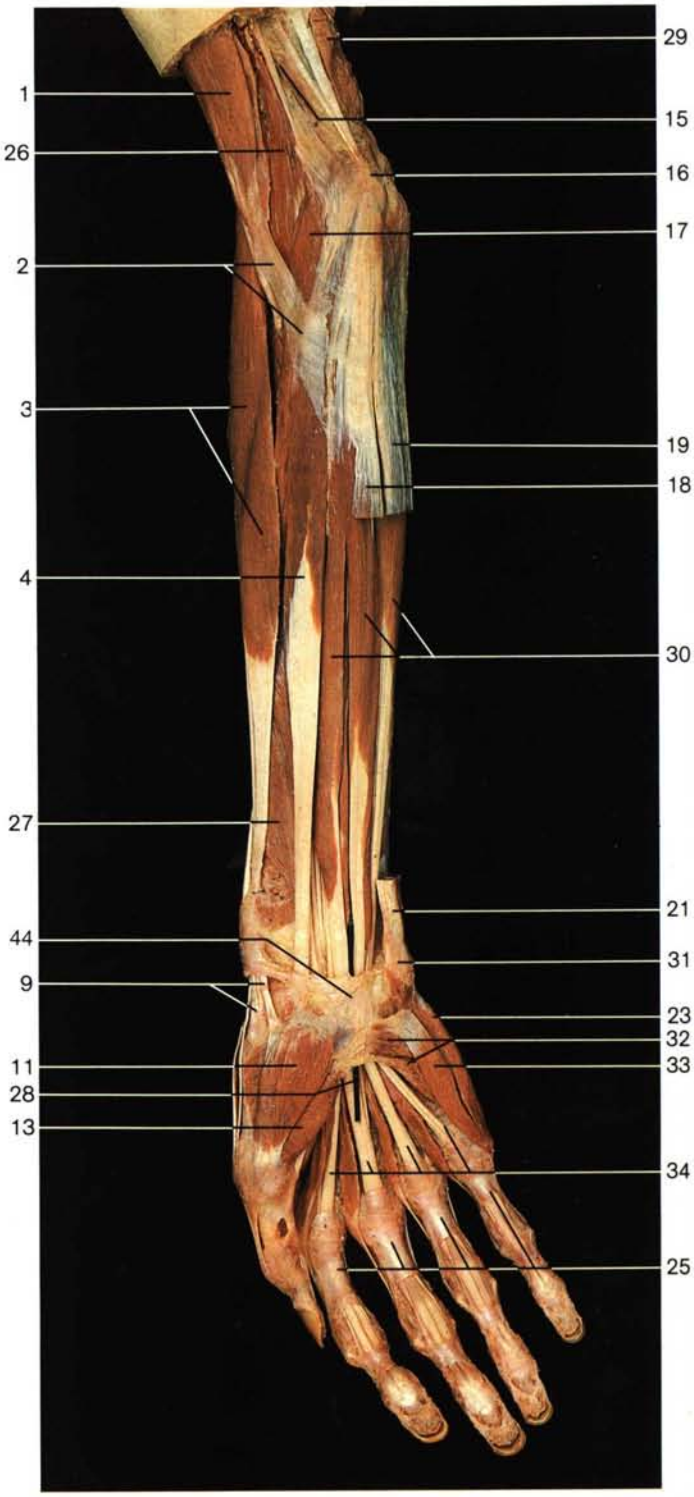
1
26
2
3
4
27
44
9
11
28
13
29
15
16
17
19
18
30
21
31
23
32
33
34
25

Forearm muscles, superficial layer (ventral aspect). The palmaris longus and flexor carpi ulnaris muscles have been removed.

1 Biceps brachii muscle
2 Bicipital aponeurosis
3 **Brachioradialis muscle**
4 **Flexor carpi radialis muscle**
5 Radial artery
6 **Flexor digitorum superficialis muscle**
7 Median nerve
8 Antebrachial fascia and tendon of palmaris longus muscle
9 Tendon of abductor pollicis longus muscle
10 Tendon of extensor pollicis brevis muscle
11 Abductor pollicis brevis muscle
12 **Palmar aponeurosis**
13 Superficial head of flexor pollicis brevis muscle
14 Tendon of flexor pollicis longus muscle
15 Medial intermuscular septum
16 Medial epicondyle of humerus
17 Humeral head of pronator teres muscle
18 **Palmaris longus muscle**
19 **Flexor carpi ulnaris muscle**
20 **Ulnar artery**
21 Tendon of flexor carpi ulnaris muscle
22 Palmaris brevis muscle

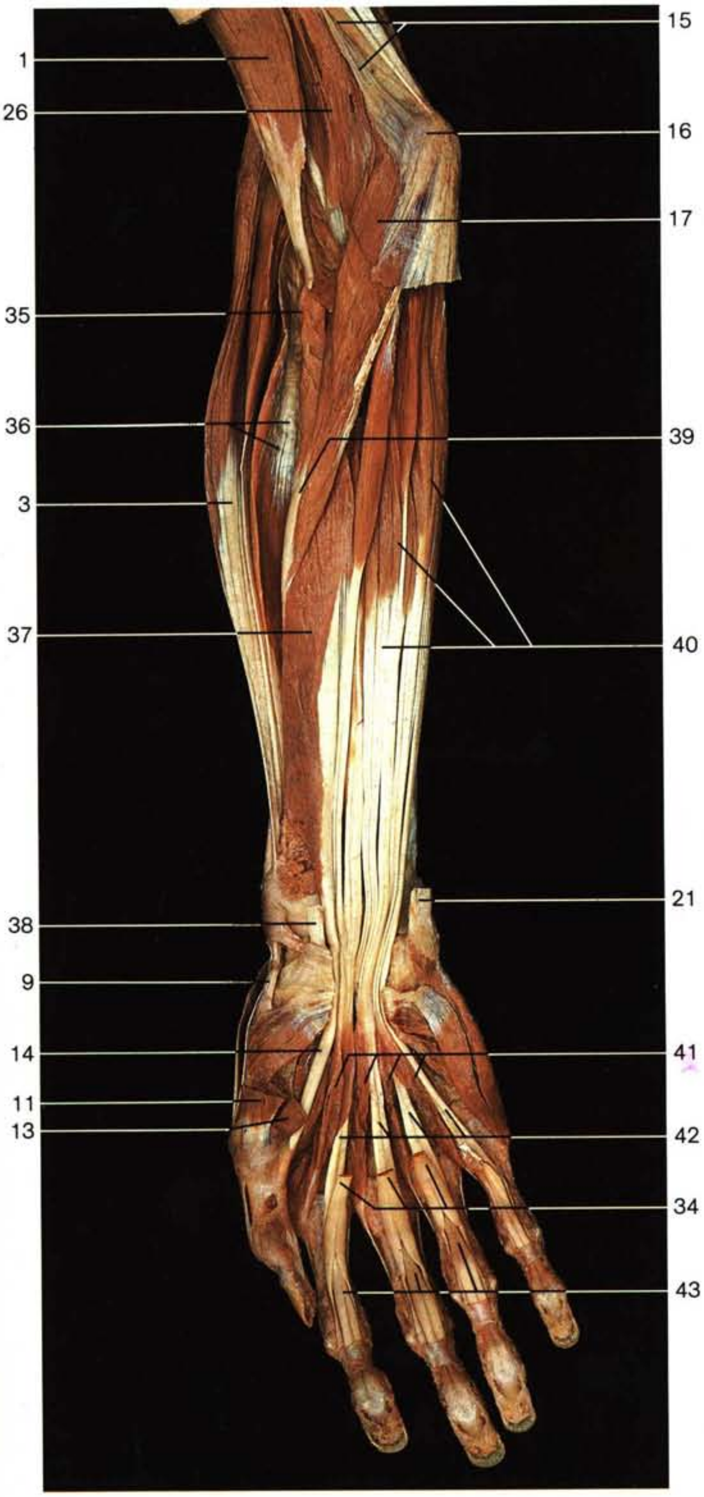

Forearm muscles, middle layer (ventral aspect). The palmaris longus, flexor carpi radialis and ulnaris muscles have been partly removed. The flexor retinaculum has been divided.

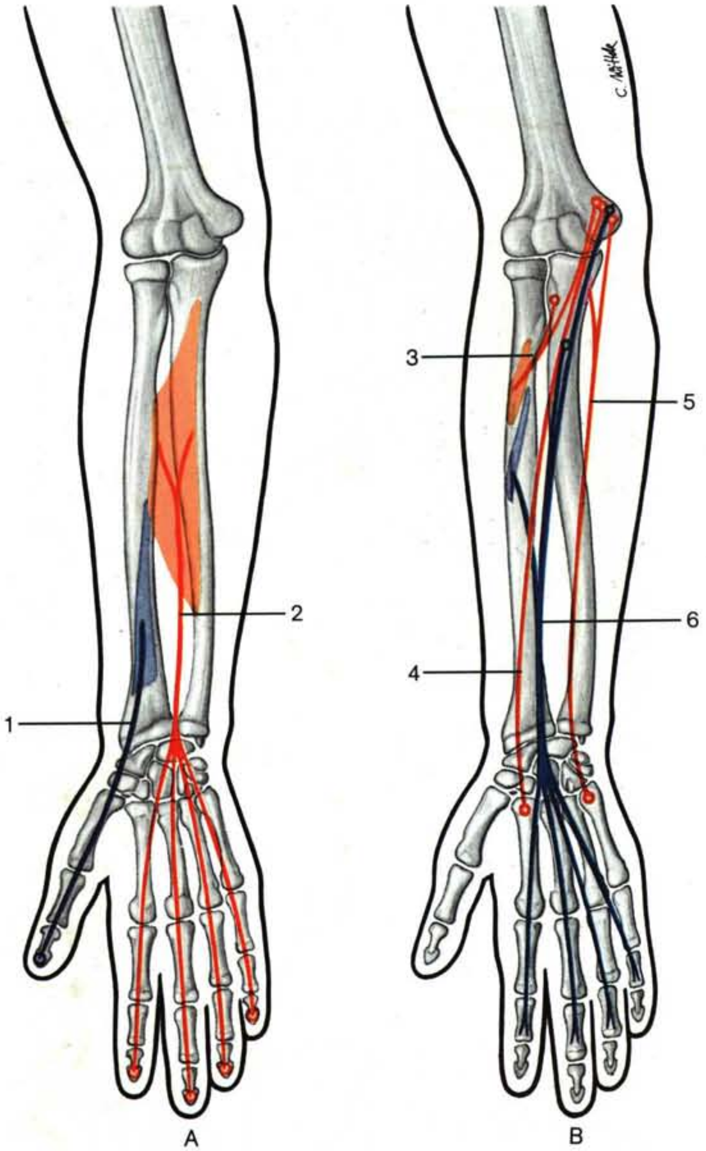

Position of flexors of fingers and hand. (Schematic drawing.)

A Deep layer
B Superficial layer
△
1 Flexor pollicis longus muscle (blue)
2 Flexor digitorum profundus muscle (red)
3 Pronator teres muscle (red)
4 Flexor carpi radialis muscle (red)
5 Flexor carpi ulnaris muscle (red)
6 Flexor digitorum superficialis muscle (blue)

23 Abductor digiti minimi muscle
24 Transverse fasciculi of palmar aponeurosis
25 Digital fibrous sheaths of tendons of flexor digitorum muscle
26 Brachialis muscle
27 Flexor pollicis longus muscle
28 Carpal tunnel (canalis carpi, probe)
29 Triceps brachii muscle
30 Flexor digitorum superficialis muscle
31 Pisiform bone
32 Opponens digiti minimi muscle
33 Flexor digiti minimi brevis muscle
34 Tendons of flexor digitorum superficialis muscle
35 Supinator muscle
36 Extensor carpi radialis brevis muscle
37 **Flexor pollicis longus muscle**
38 Tendon of flexor carpi radialis muscle
39 **Pronator teres muscle** (insertion of radius)
40 **Flexor digitorum profundus muscle**
41 **Lumbrical muscles**
42 Tendons of flexor digitorum profundus muscle
43 Tendons of flexor digitorum profundus muscle having passed through the divided tendons of the flexor digitorum superficialis muscle
44 Flexor retinaculum

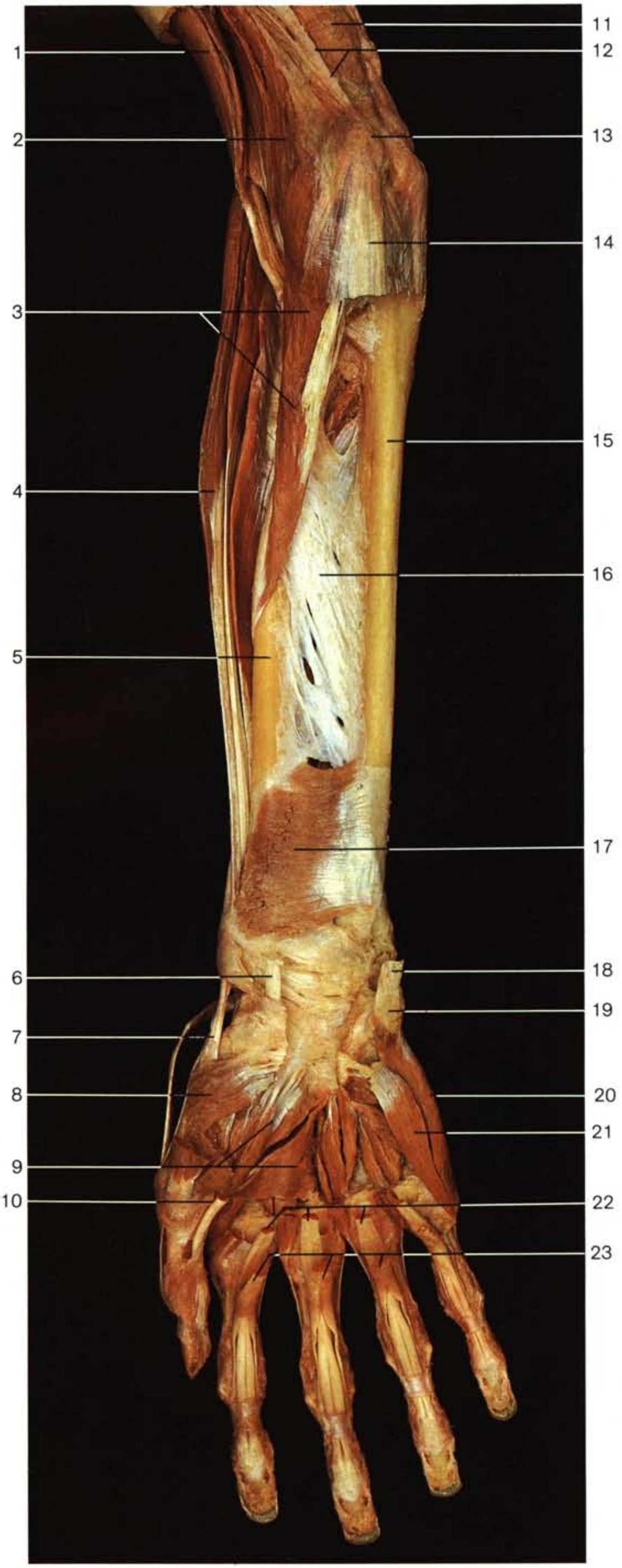

Muscles of the forearm, deep layer (ventral aspect). All flexors have been removed to display the pronator quadratus and pronator teres muscles together with the interosseous membrane. Forearm in supination.

1 Biceps brachii muscle
2 Brachialis muscle
3 **Pronator teres muscle**
4 Brachioradialis muscle
5 **Radius**
6 Tendon of flexor carpi radialis muscle
7 Tendon of abductor pollicis longus muscle
8 Opponens pollicis muscle
9 Adductor pollicis muscle
10 Tendon of flexor pollicis longus muscle
11 Triceps brachii muscle
12 Medial intermuscular septum
13 Medial epicondyle of humerus
14 Common flexor mass (divided)
15 **Ulna**
16 **Interosseous membrane**
17 **Pronator quadratus muscle**
18 Tendon of flexor carpi ulnaris muscle
19 Pisiform bone
20 Abductor digiti minimi muscle
21 Flexor digiti minimi brevis muscle
22 Tendons of flexor digitorum profundus muscle
23 Tendons of flexor digitorum superficialis muscle
24 Flexor retinaculum
25 Hypothenar muscles
26 Thenar muscles
27 Common synovial sheath of flexor tendons
28 Synovial sheath of tendon of flexor pollicis longus muscle
29 Digital synovial sheaths of flexor tendons

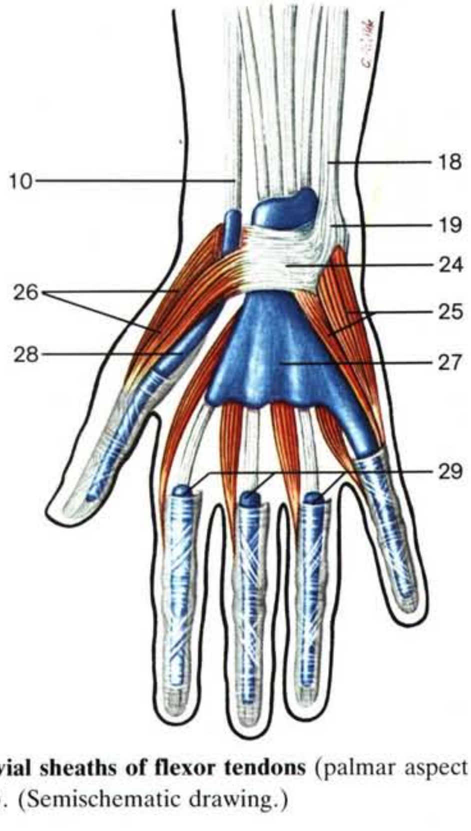

Synovial sheaths of flexor tendons (palmar aspect of right hand). (Semischematic drawing.)

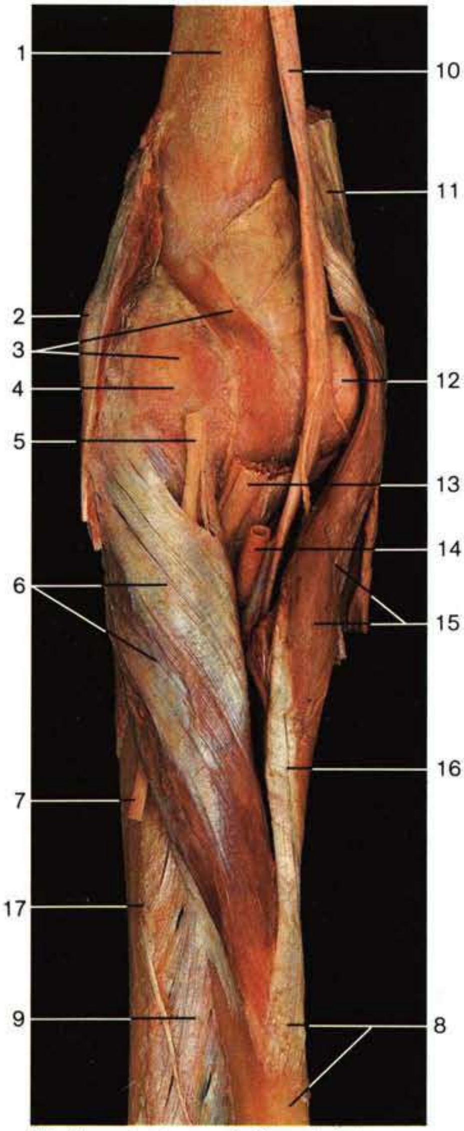

Right supinator and elbow joint
(anterior aspect). Forearm in pronation.

1	Humerus
2	Lateral epicondyle of humerus
3	Articular capsule
4	Position of capitulum of humerus
5	Deep branch of radial nerve
6	**Supinator muscle**
7	Entrance of deep branch of radial nerve to extensor muscles
8	Radius and insertion of pronator teres muscle
9	Interosseous membrane
10	Median nerve
11	Triceps brachii muscle
12	Trochlea of humerus
13	Tendon of biceps brachii muscle
14	Brachial artery
15	**Pronator teres muscle**
16	Tendon of pronator teres muscle
17	Ulna
18	**Pronator quadratus muscle**
19	Tendon of flexor carpi radialis muscle
20	Thenar muscles
21	Synovial sheath of tendon of flexor pollicis longus muscle
22	Fibrous sheath of flexor tendons
23	**Digital synovial sheath of flexor tendons**
24	Flexor digitorum superficialis muscle
25	Tendon of flexor carpi ulnaris muscle
26	**Common synovial sheath of flexor tendons**
27	Position of pisiform bone
28	Flexor retinaculum
29	Hypothenar muscles

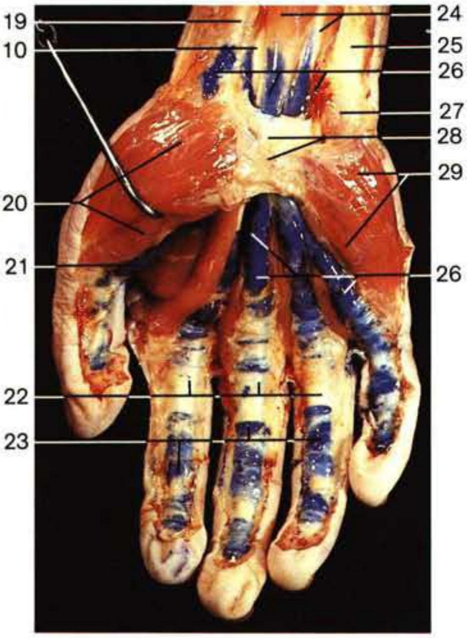

Synovial sheaths of flexor muscles
(palmar aspect of right hand).
Blue PVA-solution has been injected into the sheaths.

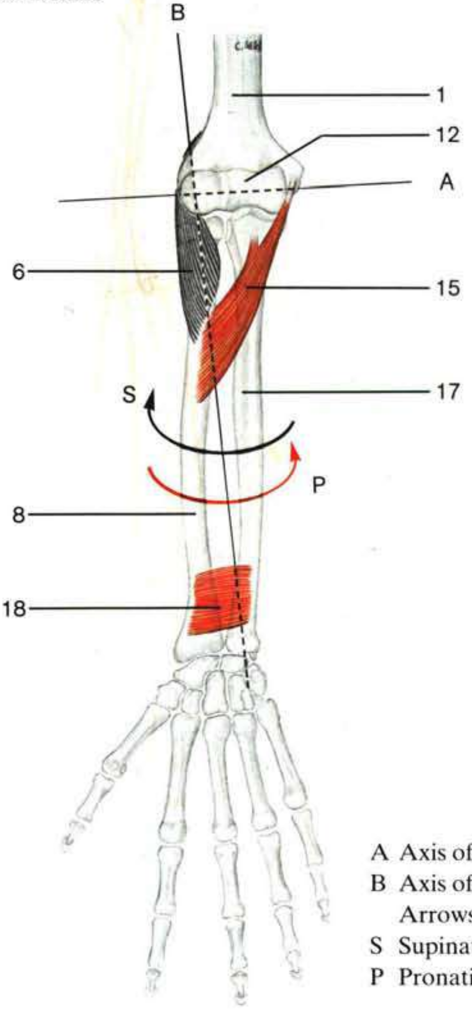

A Axis of flexion and extension
B Axis of rotation
Arrows:
S Supination
P Pronation

Diagram illustrating the two **axes of the elbow joint.**

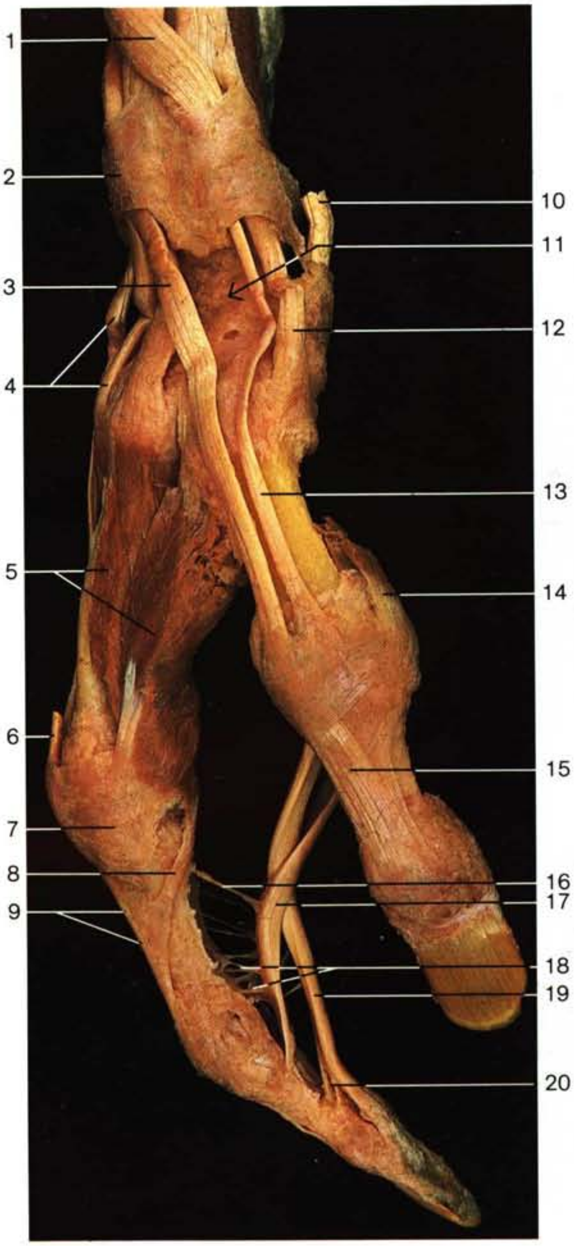

Muscles of thumb and index finger (medial aspect). The tendons of the extensor muscles of the thumb and the insertion of the flexor tendons of the index finger are displayed.

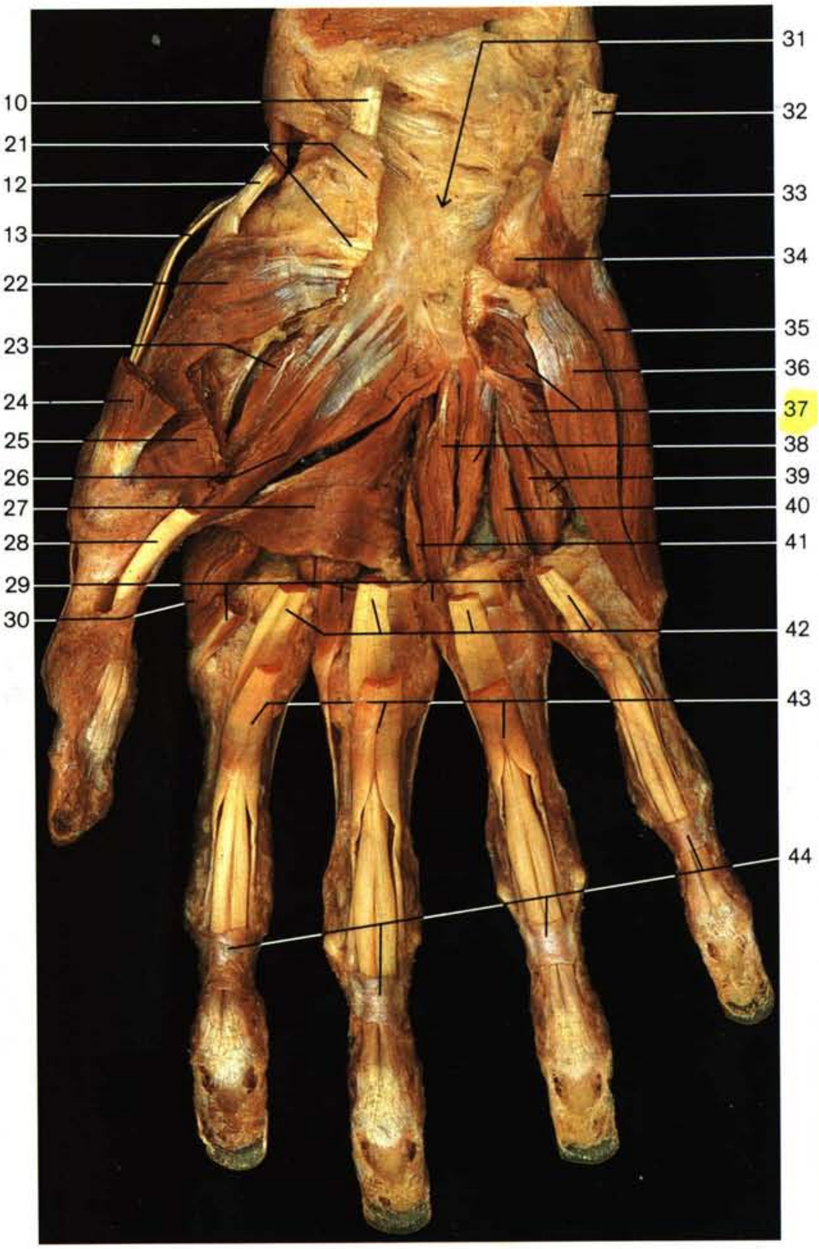

Muscles of hand (palmar aspect). The tendons of the flexor muscles and parts of the thumb muscles have been removed. The carpal tunnel has been opened.

1 Tendons of extensor pollicis brevis and abductor pollicis longus muscle
2 Extensor retinaculum
3 Tendon of extensor pollicis longus muscle
4 Tendons of extensor carpi radialis longus and brevis muscles
5 First dorsal interosseous muscle
6 Tendon of extensor digitorum muscle for index finger
7 Location of metacarpophalangeal joint
8 Tendon of lumbrical muscle
9 Extensor expansion of index finger
10 Tendon of flexor carpi radialis muscle (cut)
11 **Anatomical snuffbox**
12 Tendon of abductor pollicis longus muscle
13 Tendon of extensor pollicis brevis muscle
14 Tendon of abductor pollicis brevis muscle
15 Extensor expansion of extensor of thumb
16 Vinculum longum
17 Tendons of flexor digitorum superficialis muscle dividing to allow passage of deep tendons
18 Vincula of flexor tendons
19 Tendon of flexor digitorum profundus muscle
20 Vinculum breve
21 Radial carpal eminence (cut edge of flexor retinaculum)
22 Opponens pollicis muscle
23 Deep head of flexor pollicis brevis muscle
24 Abductor pollicis brevis muscle (cut)
25 Superficial head of flexor pollicis brevis muscle (cut)
26 Oblique head of adductor pollicis muscle
27 Transverse head of adductor pollicis muscle
28 Tendon of flexor pollicis longus muscle (cut)
29 Lumbrical muscles (cut)
30 First dorsal interosseous muscle
31 **Position of carpal tunnel**
32 Tendon of flexor carpi ulnaris muscle
33 Location of pisiform bone
34 Hook of hamate bone
35 Abductor digiti minimi muscle
36 Flexor digiti minimi brevis muscle
37 Opponens digiti minimi muscle
38 2nd palmar interosseous muscle
39 3rd palmar interosseous muscle
40 4th dorsal interosseous muscle
41 3rd dorsal interosseous muscle
42 Tendon of flexor digitorum profundus muscle (cut)
43 Tendons of flexor digitorum superficialis muscle (cut)
44 Fibrous flexor sheaths

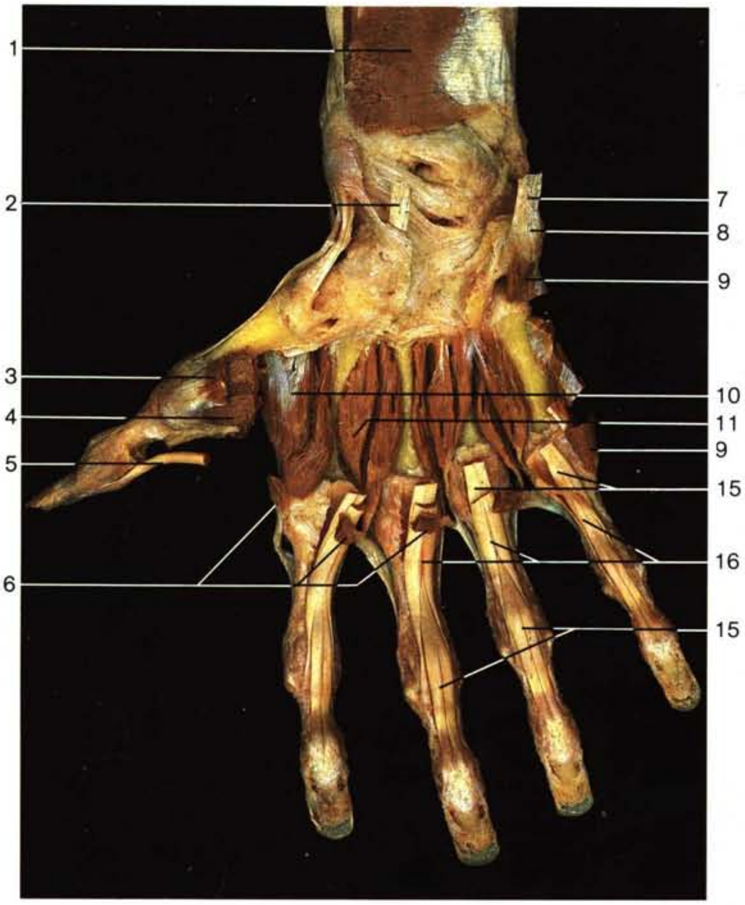

Muscles of right hand, deep layer (palmar aspect). The thenar and hypothenar muscles have been removed to display the interosseous muscles.

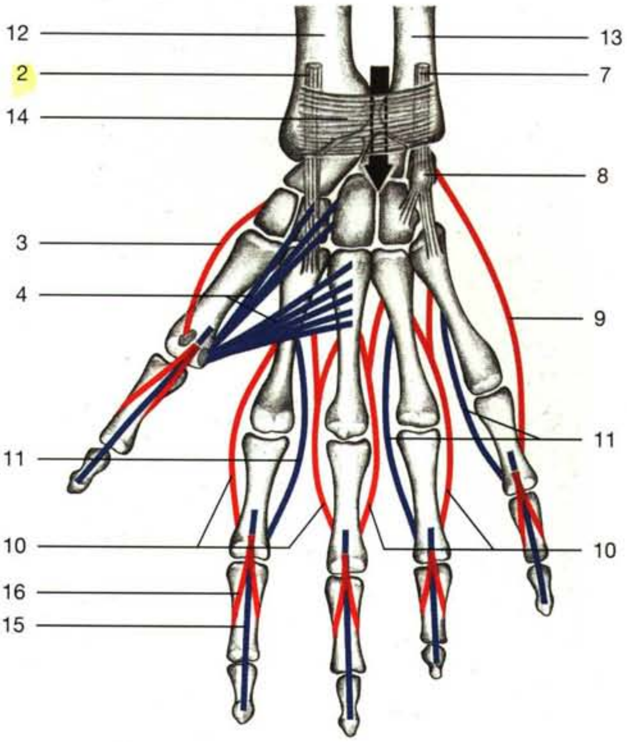

Actions of interosseous muscles in abduction and adduction of fingers (palmar aspect, schematic drawing). Arrow: carpal tunnel.
Red = **abduction**
(dorsal interosseous muscles, abductor digiti minimi and abductor pollicis brevis muscles).
Black = **adduction**
(palmar interosseous muscles, adductor pollicis muscle).

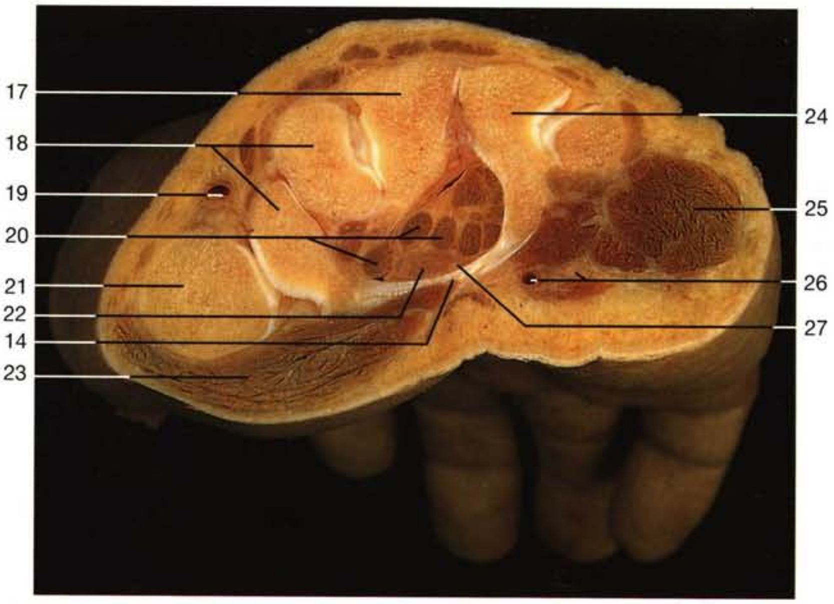

Transverse section through the right hand, showing the carpal tunnel (canalis carpi).

1	Pronator quadratus muscle
2	Tendon of flexor carpi radialis muscle
3	Abductor pollicis brevis muscle (divided)
4	Adductor pollicis muscle (divided)
5	Tendon of flexor pollicis longus muscle
6	Lumbrical muscles (cut)
7	Tendon of flexor carpi ulnaris muscle
8	Pisiform bone
9	Abductor digiti minimi muscle (divided)
10	**Dorsal interosseous muscles**
11	**Palmar interosseous muscles**
12	Radius
13	Ulna
14	**Flexor retinaculum**
15	Tendons of flexor digitorum profundus muscle
16	Tendons of flexor digitorum superficialis muscle
17	Capitate bone
18	Trapezium bone and trapezoid bone
19	Radial artery
20	Tendon of flexor muscles
21	First metacarpal bone
22	**Median nerve**
23	Thenar muscles
24	Hamate bone
25	Hypothenar muscles
26	Ulnar artery and nerve
27	**Carpal tunnel** (canalis carpi)

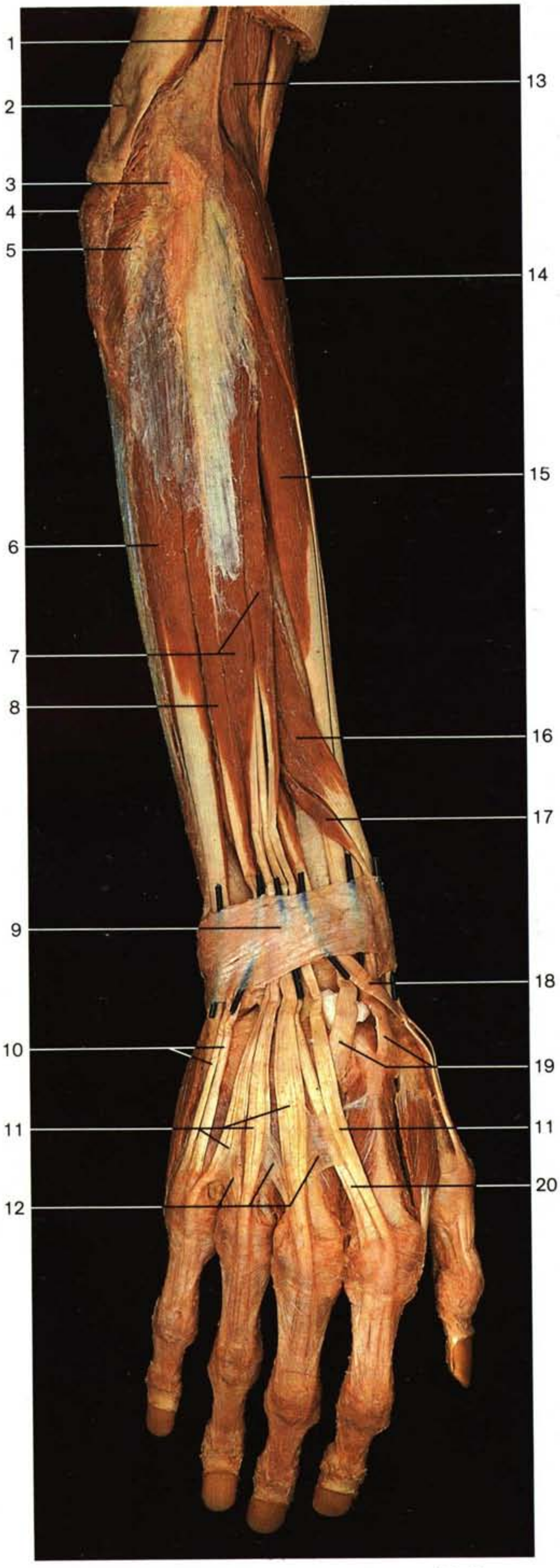

Extensor muscles of forearm and hand, superficial layer (dorsal aspect). Tunnels for extensor tendons indicated by probes.

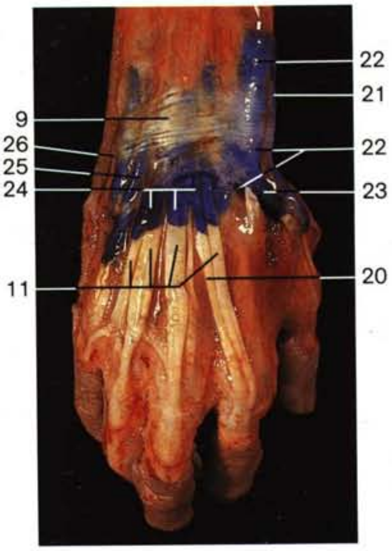

Synovial sheaths of extensor tendons. The sheaths have been injected with blue gelatin.

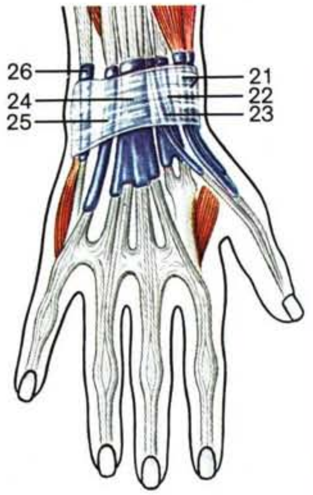

Synovial sheaths of extensor tendons on the back of right wrist (indicated in blue). Notice the six tunnels for the passage of the extensor tendons beneath the extensor retinaculum. (Semischematic drawing.)

1 Lateral intermuscular septum
2 Tendon of triceps brachii muscle
3 Lateral epicondyle of humerus
4 Olecranon
5 Anconeus
6 Extensor carpi ulnaris muscle
7 Extensor digitorum muscle
8 Extensor digiti minimi muscle
9 Extensor retinaculum
10 Tendons of extensor digiti minimi muscle
11 Tendons of extensor digitorum muscle
12 Intertendinous connections
13 Brachioradialis muscle
14 Extensor carpi radialis longus muscle
15 Extensor carpi radialis brevis muscle
16 Abductor pollicis longus muscle
17 Extensor pollicis brevis muscle
18 Tendon of extensor pollicis longus muscle
19 Tendons of both extensor carpi radialis longus and extensor carpi radialis brevis muscles
20 Tendon of extensor indicis muscle
21 1st tunnel: Abductor pollicis longus muscle
Extensor pollicis brevis muscle
22 2nd tunnel: Extensor carpi radialis longus and brevis muscles
23 3rd tunnel: Extensor pollicis longus muscle
24 4th tunnel: Extensor digitorum muscle
Extensor indicis muscle
25 5th tunnel: Extensor digiti minimi muscle
26 6th tunnel: Extensor carpi ulnaris muscle

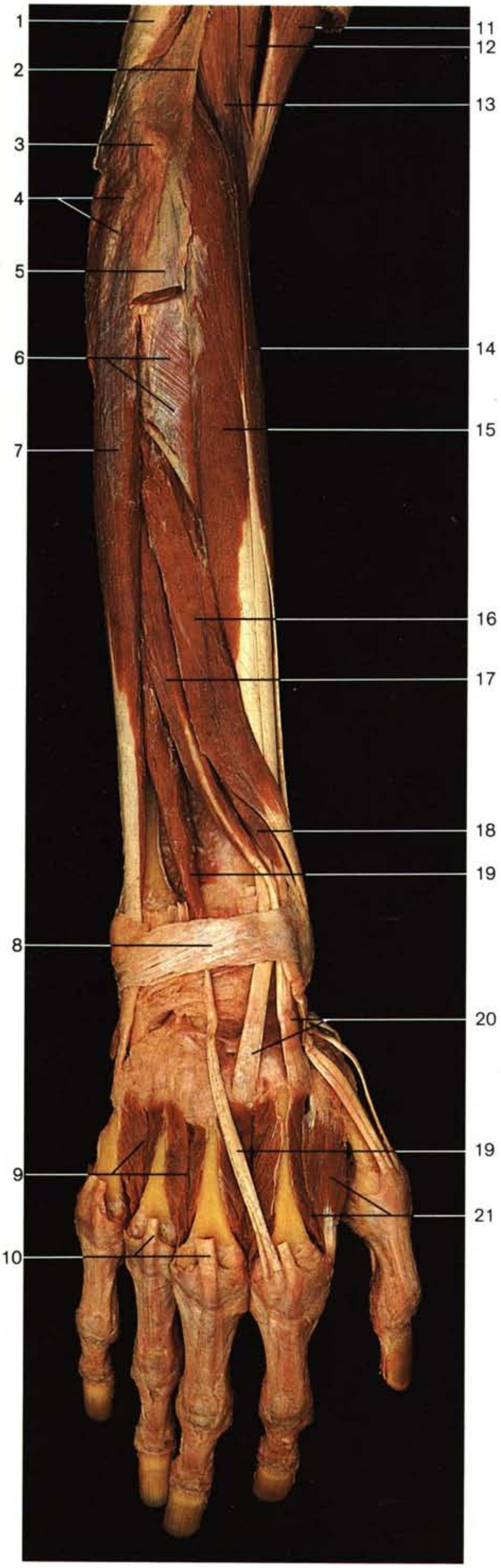

Extensor muscles of forearm and hand, deep layer (dorsal aspect).

1 Triceps brachii muscle
2 Lateral intermuscular septum
3 Lateral epicondyle of humerus
4 Anconeus muscle
5 Extensor digitorum and extensor digiti minimi muscles (cut)
6 Supinator muscle
7 Extensor carpi ulnaris muscle
8 Extensor retinaculum
9 3rd and 4th dorsal interosseous muscles
10 Tendons of extensor digitorum muscle (cut)
11 Biceps brachii muscle
12 Brachialis muscle
13 Brachioradialis muscle
14 Extensor carpi radialis longus muscle
15 Extensor carpi radialis brevis muscle
16 Abductor pollicis longus muscle
17 Extensor pollicis longus muscle
18 Extensor pollicis brevis muscle
19 Extensor indicis muscle
20 Tendons of the extensor carpi radialis longus and extensor carpi radialis brevis muscles
21 1st dorsal interosseous muscle

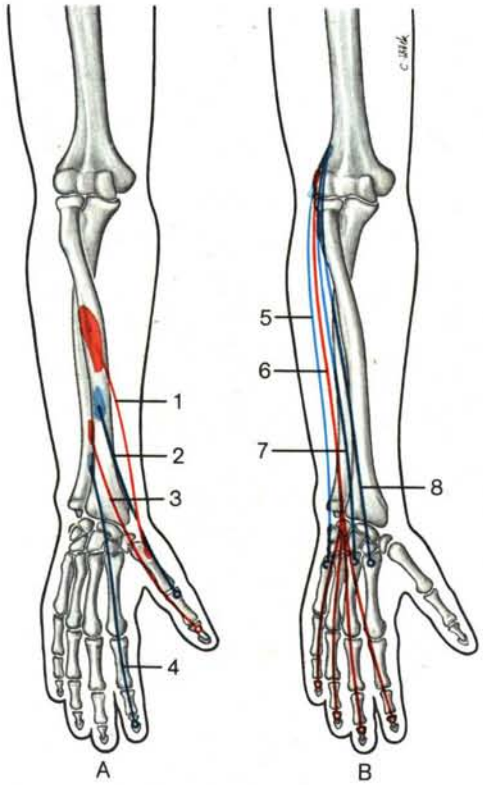

Position of extensor muscles of forearm and hand. (Semischematic drawing.)

A Extensors of thumb	B Extensors of fingers and hand
1 Abductor pollicis longus muscle (red)	5 Extensor carpi ulnaris muscle (blue)
2 Extensor pollicis brevis muscle (blue)	6 Extensor digitorum muscle (red)
3 Extensor pollicis longus muscle (red)	7 Extensor carpi radialis brevis muscle (blue)
4 Extensor indicis muscle (blue)	8 Extensor carpi radialis longus muscle (blue)

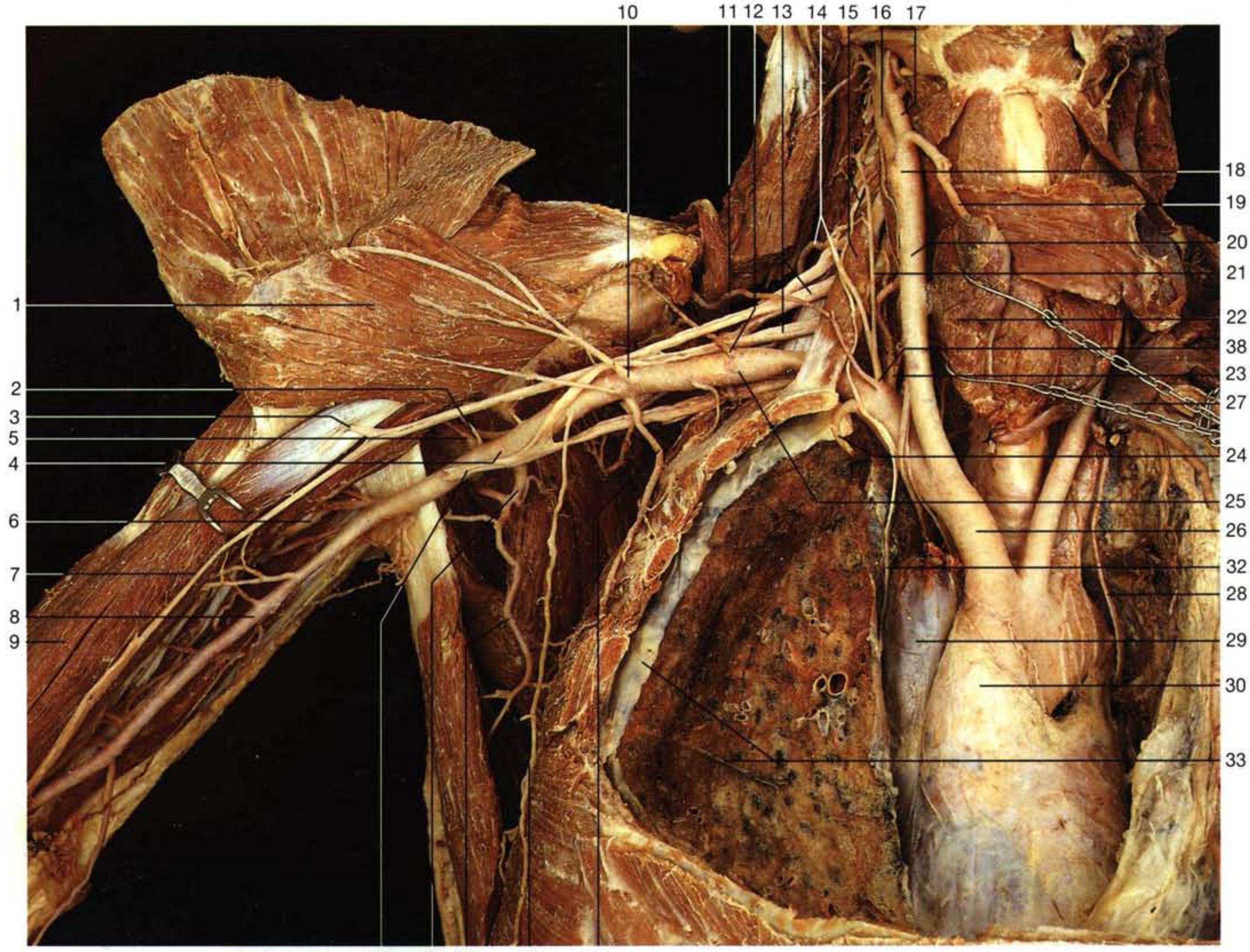

Main branches of right subclavian and axillary arteries (anterior aspect). Pectoralis muscles have been reflected, clavicle and anterior wall of thorax removed and right lung divided. Left lung with pleura and thyroid gland have been reflected laterally to display aortic arch and common carotid artery with their branches.

1 Pectoralis minor muscle (reflected)
2 Anterior circumflex humeral artery
3 Musculocutaneous nerve (divided)
4 **Axillary artery**
5 Posterior circumflex humeral artery
6 **Profunda brachii artery**
7 Median nerve (var.)
8 **Brachial artery**
9 Biceps brachii muscle
10 Thoracoacromial artery
11 Suprascapular artery
12 Descending scapular artery
13 Brachial plexus (middle trunk)
14 Transverse cervical artery
15 Scalenus anterior muscle and phrenic nerve
16 Right internal carotid artery
17 Right external carotid artery
18 Carotid sinus
19 Superior thyroid artery
20 Right common carotid artery
21 Ascending cervical artery
22 Thyroid gland
23 Inferior thyroid artery
24 Internal thoracic artery
25 Right subclavian artery
26 Brachiocephalic trunk
27 Left brachiocephalic vein (divided)
28 Left vagus nerve
29 Superior vena cava (divided)
30 **Ascending aorta**
31 Median nerve (divided)
32 Phrenic nerve
33 Right lung (divided) and pulmonary pleura
34 Thoracodorsal artery
35 Subscapular artery
36 Lateral mammary branches (variant)
37 Lateral thoracic artery
38 **Thyrocervical trunk**
39 Superior thoracic artery
40 Superior ulnar collateral artery
41 Inferior ulnar collateral artery
42 Middle collateral artery
43 Radial collateral artery
44 Radial recurrent artery
45 **Radial artery**
46 Anterior and posterior interosseous arteries
47 Princeps pollicis artery
48 **Deep palmar arch**
49 Common palmar digital arteries
50 Ulnar recurrent artery
51 Recurrent interosseous artery
52 Common interosseous artery
53 **Ulnar artery**
54 **Superficial palmar arch**
55 Median nerve and brachial artery
56 Biceps brachii muscle
57 Ulnar nerve
58 Flexor pollicis longus muscle
59 Palmar digital arteries
60 Anterior interosseous artery
61 Flexor carpi ulnaris muscle
62 Superficial palmar branch of radial artery

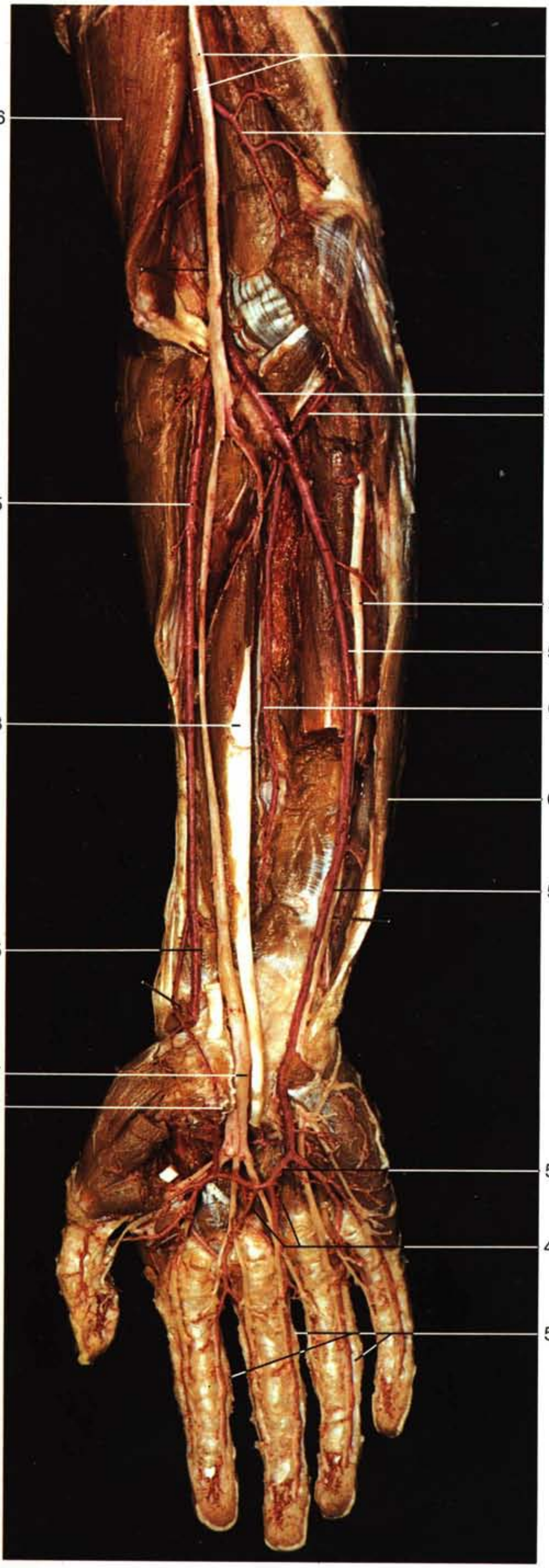
55
56
41
53
50
45
57
53
60
58
61
53
45
7
62
54
49
59

Dissection of the arteries of forearm and hand. The superficial flexor muscles have been removed, the carpal tunnel opened and the flexor retinaculum cut. The arteries have been filled with colored resin.

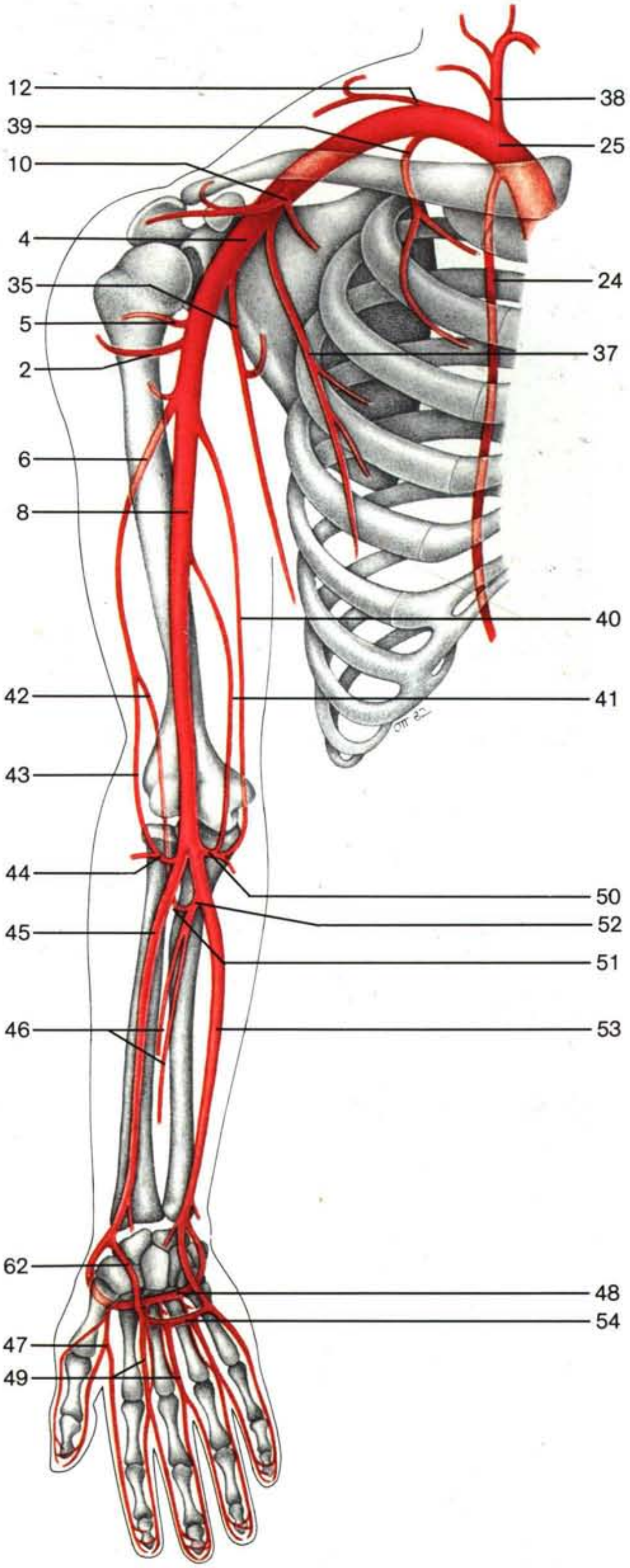
12
38
39
25
10
4
35
24
5
37
2
6
8
40
42
41
43
44
50
45
52
51
46
53
62
48
54
47
49

Arteries of the upper limb. (Schematic drawing.)

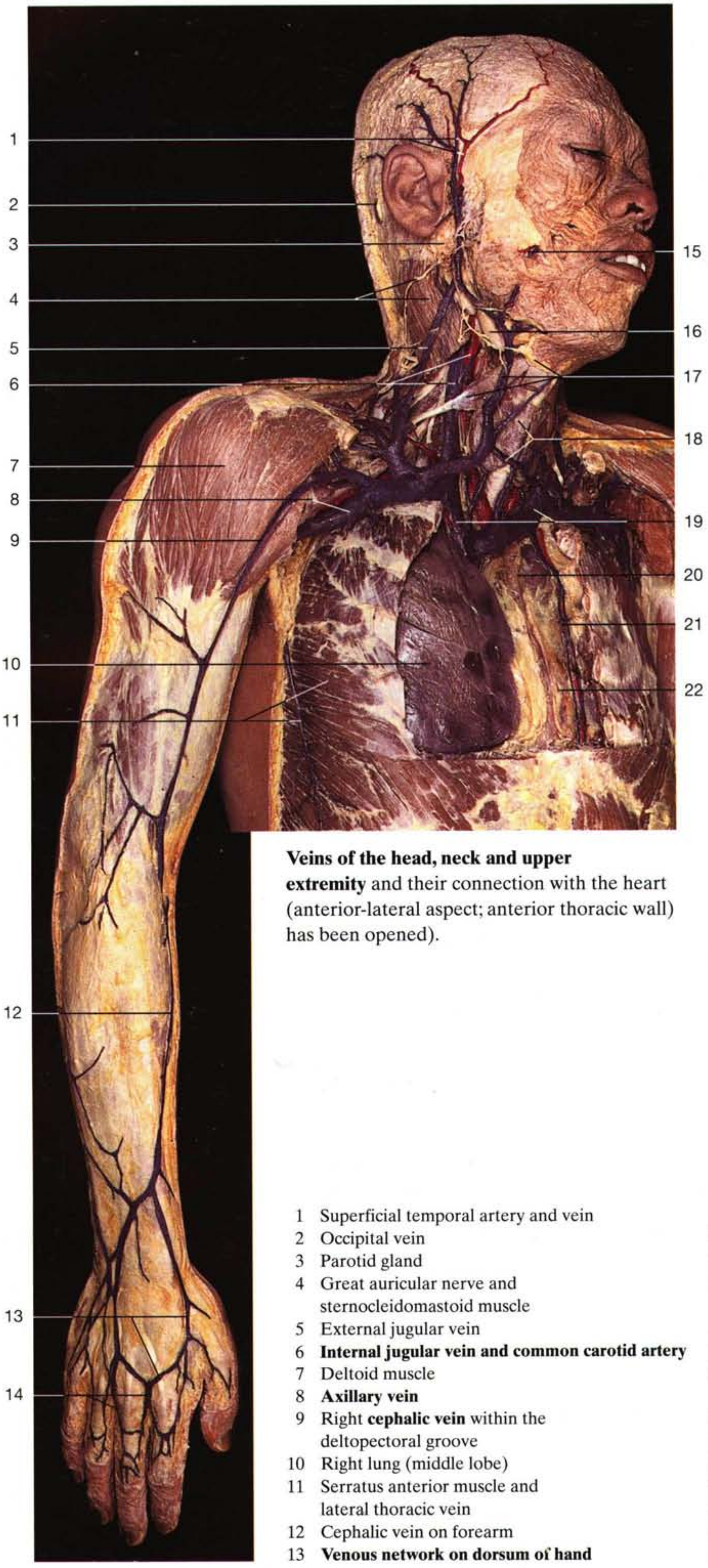

Veins of the head, neck and upper extremity and their connection with the heart (anterior-lateral aspect; anterior thoracic wall) has been opened).

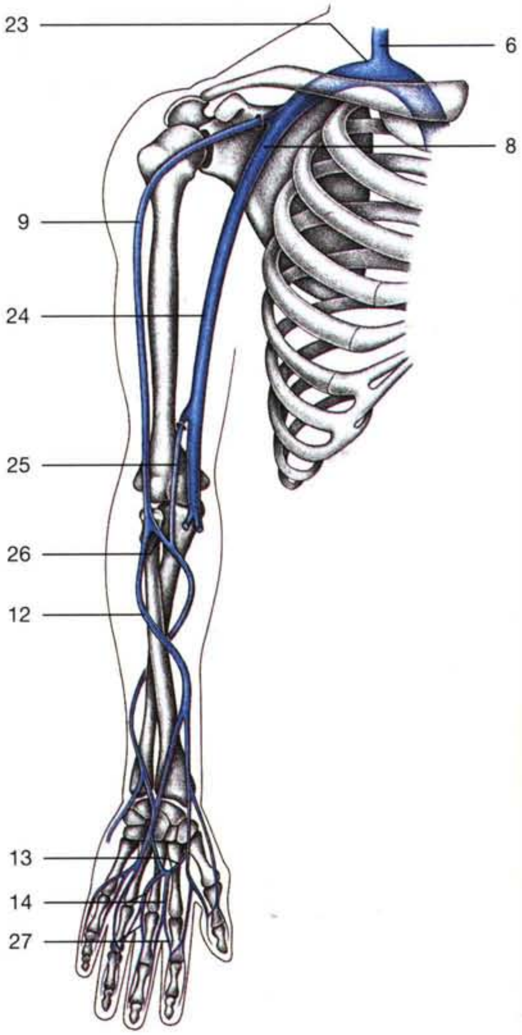

Superficial veins of upper limb. (Semischematic drawing.)

1 Superficial temporal artery and vein
2 Occipital vein
3 Parotid gland
4 Great auricular nerve and sternocleidomastoid muscle
5 External jugular vein
6 **Internal jugular vein and common carotid artery**
7 Deltoid muscle
8 **Axillary vein**
9 Right **cephalic vein** within the deltopectoral groove
10 Right lung (middle lobe)
11 Serratus anterior muscle and lateral thoracic vein
12 Cephalic vein on forearm
13 **Venous network on dorsum of hand**
14 Dorsal metacarpal veins
15 **Facial artery and vein**
16 Submandibular gland
17 Anterior jugular vein, hyoid bone and omohyoid muscle
18 Jugular venous arch and thyroid gland
19 Right and left brachiocephalic veins
20 Retrosternal body (remnant of thymus gland)
21 Internal thoracic artery and vein
22 Heart with pericardium
23 Right venous angle
24 **Brachial vein**
25 Basilic vein
26 Median cubital vein
27 Digital veins

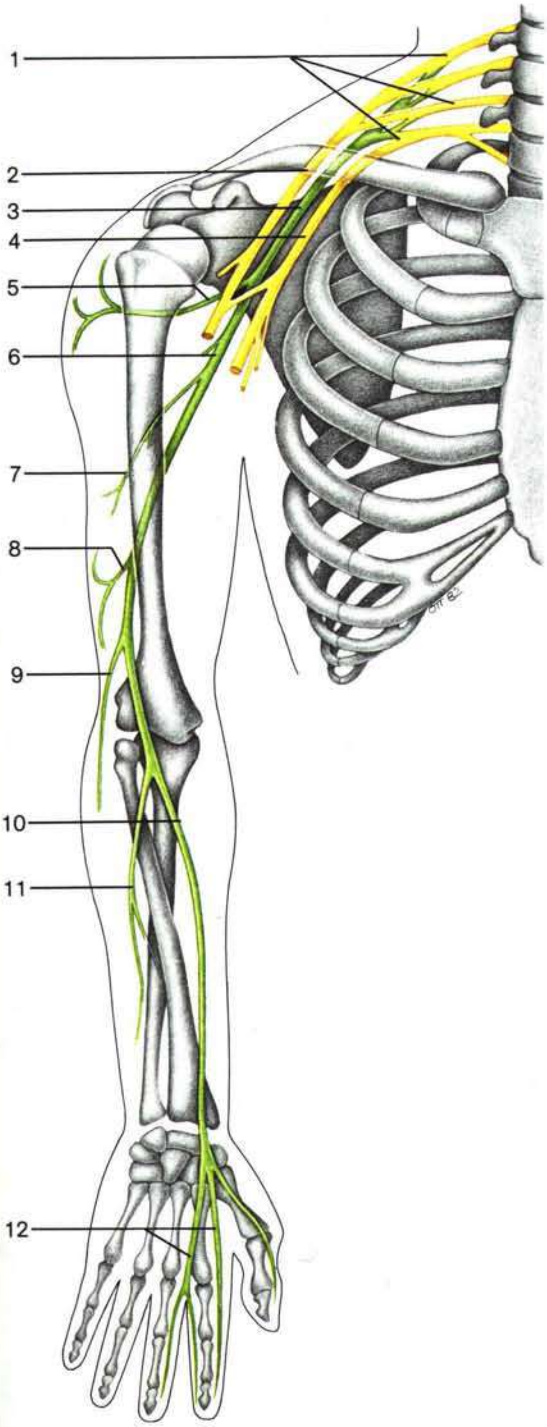

Main branches of radial nerve. (Schematic drawing.) Posterior divisions of trunks and posterior cord and its branches are indicated in green.

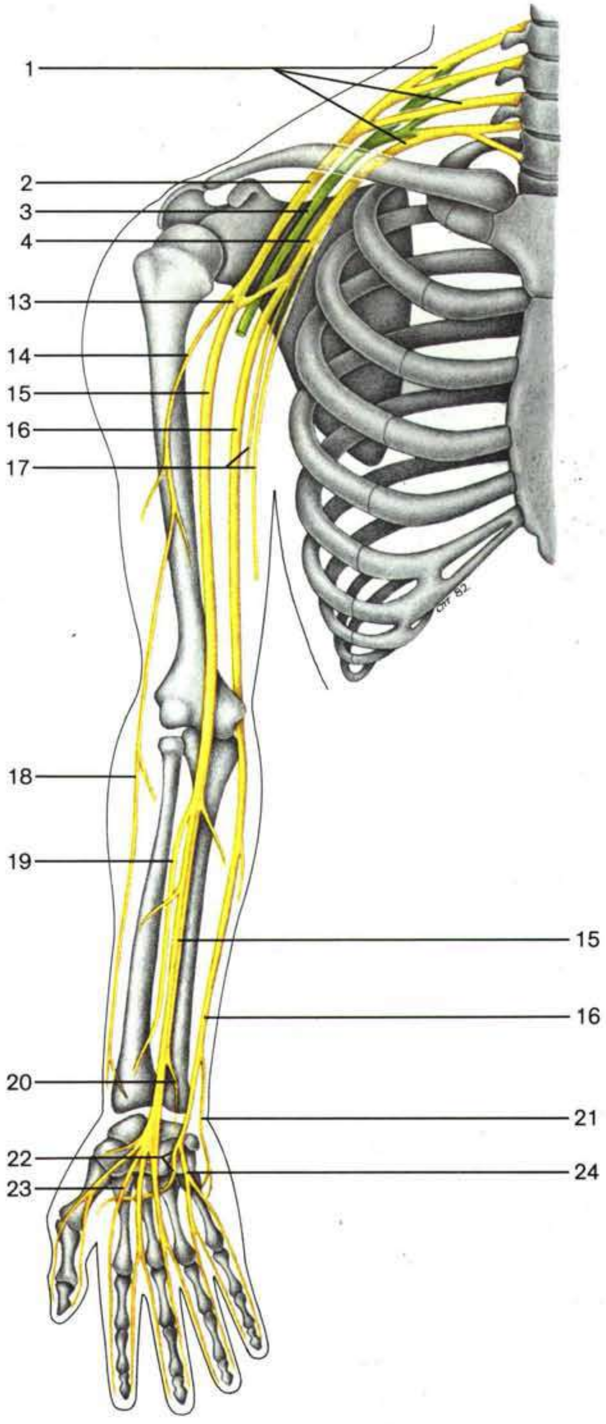

Main branches of musculocutaneous, median and ulnar nerves. (Schematic drawing.)

Anterior divisions of the trunks and all the components arising from them are indicated in yellow.

1 Brachial plexus
2 Lateral cord of brachial plexus
3 Posterior cord of brachial plexus
4 Medial cord of brachial plexus
5 **Axillary nerve**
6 **Radial nerve**
7 Posterior cutaneous nerve of arm
8 Lower lateral cutaneous nerve of arm
9 Posterior cutaneous nerve of forearm
10 Superficial branch of radial nerve
11 Deep branch of radial nerve
12 Dorsal digital nerves

13 Roots of median nerve
14 Musculocutaneous nerve
15 **Median nerve**
16 **Ulnar nerve**
17 Medial cutaneous nerves of arm and forearm
18 Lateral cutaneous nerve of forearm
19 Anterior interosseous nerve
20 Palmar branch of median nerve
21 Dorsal branch of ulnar nerve
22 Deep branch of ulnar nerve
23 Common palmar digital nerves of median nerve
24 Superficial branch of ulnar nerve

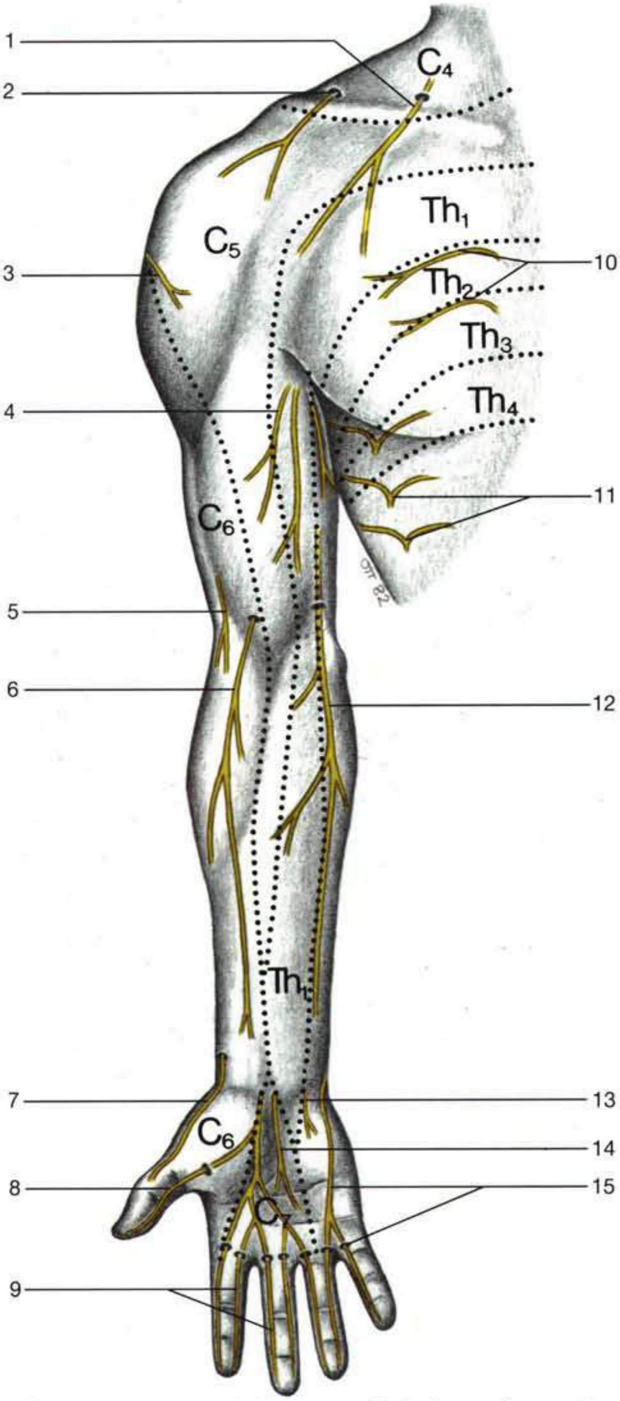

Cutaneous nerves of right upper limb (ventral aspect).
(Semischematic drawing.)

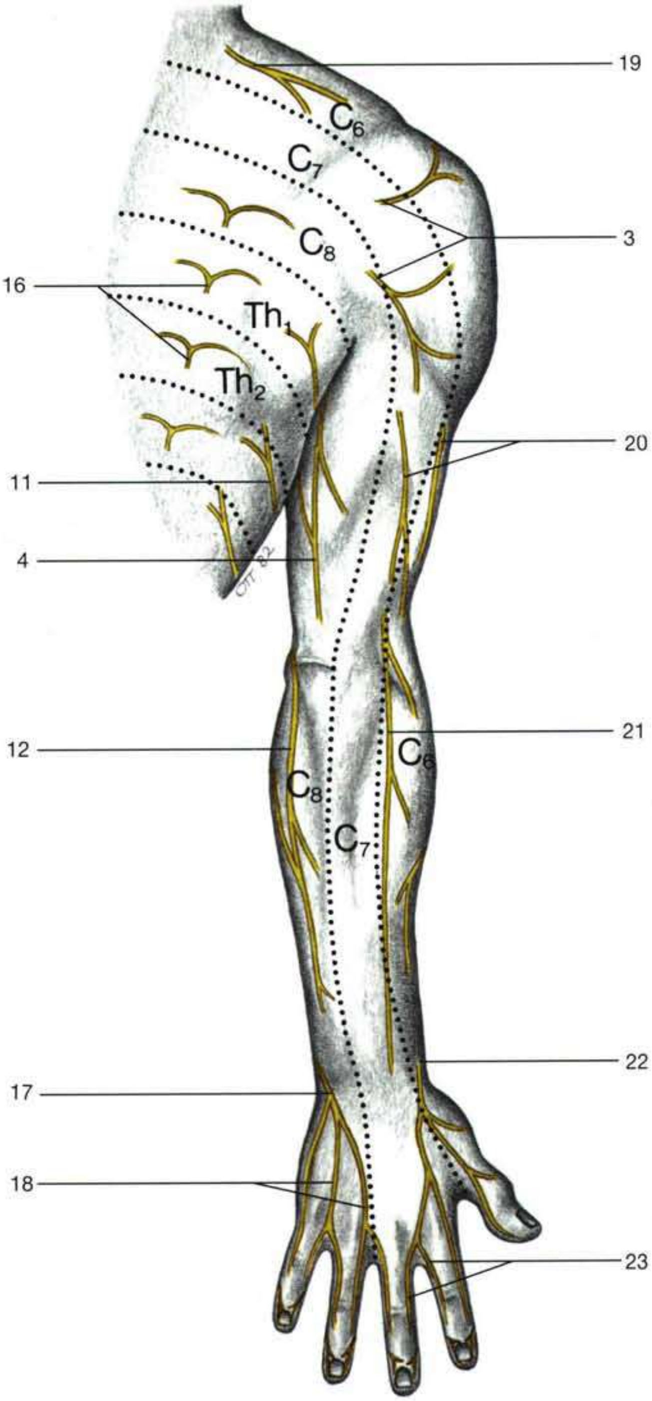

Cutaneous nerves of the upper limb (dorsal aspect).
(Semischematic drawing.)

1 Medial supraclavicular nerve
2 Intermediate supraclavicular nerve
3 Upper lateral cutaneous nerve of arm
4 Terminal branches of intercostobrachial nerves
5 Lower lateral cutaneous nerve of arm
6 Lateral cutaneous nerve of forearm
7 Terminal branch of superficial branch of radial nerve
8 Palmar digital nerve of thumb (branch of median nerve)
9 Palmar digital branches of median nerve
10 Anterior cutaneous branches of intercostal nerves
11 Lateral cutaneous branches of intercostal nerves
12 Medial cutaneous nerve of forearm

13 Palmar cutaneous branch of ulnar nerve
14 Palmar branch of median nerve
15 Palmar digital branches of ulnar nerve
16 Cutaneous branches of dorsal rami of spinal nerves
17 Dorsal branch of ulnar nerve
18 Dorsal digital nerves
19 Posterior supraclavicular nerve
20 Posterior cutaneous nerve of arm
21 Posterior cutaneous nerve of forearm
22 Superficial branch
23 Dorsal digital branches

} from radial nerve

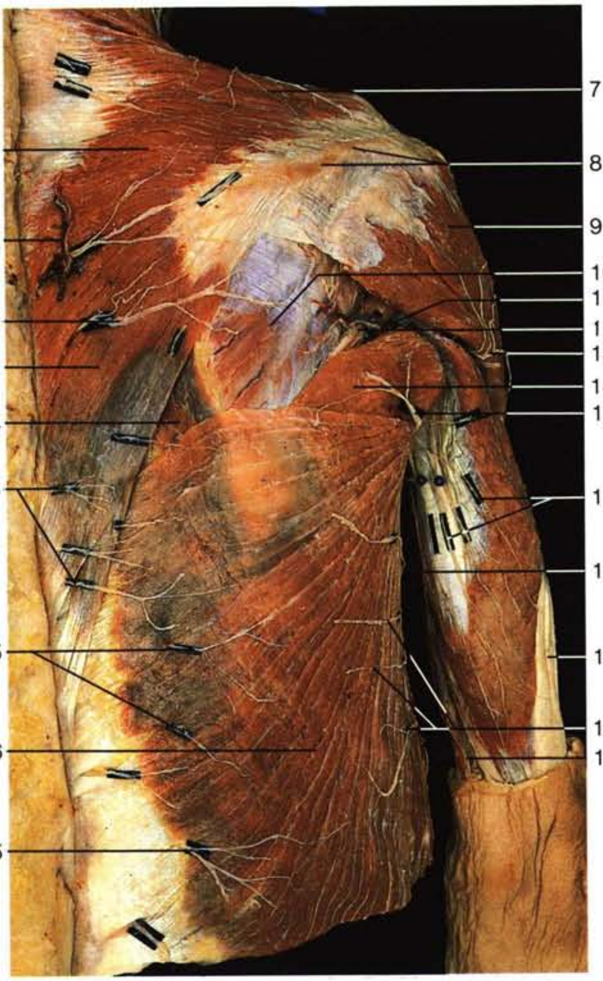

Dorsal region of shoulder, superficial layer. Note the segmental arrangement of the cutaneous nerves of the back.

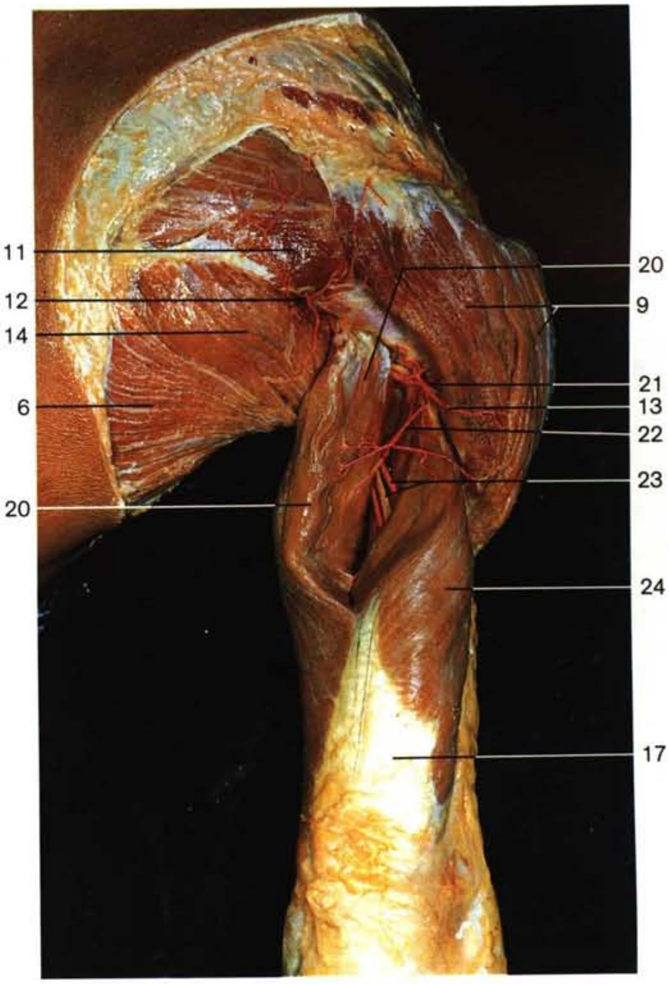

Shoulder and arm (dorsal aspect). Dissection of the quadrangular and triangular spaces of the axillary region.

1 Trapezius muscle
2 Dorsal branches of posterior intercostal artery and vein (medial cutaneous branches)
3 Medial branches of dorsal rami of spinal nerves
4 Rhomboid major muscle
5 Lateral branches of dorsal rami of spinal nerves
6 Latissimus dorsi muscle
7 Posterior supraclavicular nerves
8 Spine of scapula
9 Deltoid muscle
10 Infraspinatus muscle
11 Teres minor muscle
12 **Triangular space** with circumflex scapular artery and vein
13 Upper lateral cutaneous nerve of arm with artery
14 Teres major muscle
15 Terminal branches of intercostobrachial nerve
16 Medial cutaneous nerve of arm
17 Tendon of triceps brachii muscle
18 Lateral cutaneous branches of intercostal nerves
19 Medial cutaneous nerve of forearm
20 Long head of triceps brachii muscle
21 **Quadrangular space** with axillary nerve and posterior humeral circumflex artery
22 Anastomosis between profunda brachii artery and posterior humeral circumflex artery
23 Course of radial nerve and profunda brachii artery
24 Lateral head of triceps brachii muscle
25 Course of descending scapular artery and dorsal scapular nerve
26 Course of suprascapular nerve, artery and vein

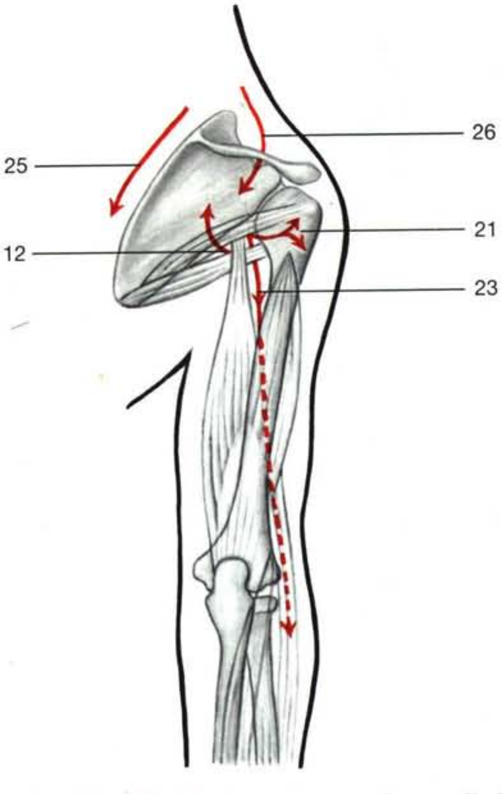

Course of vessels and nerves to shoulder and upper limb. (Schematic drawing.)

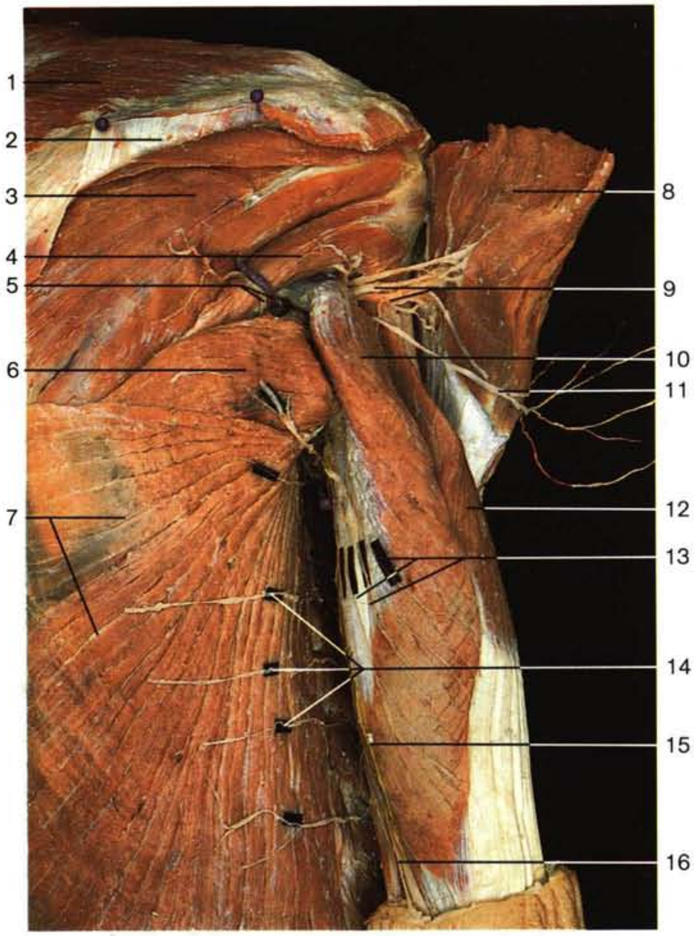

Scapular region, arm and shoulder, deep layer (dorsal aspect). Part of deltoid muscle has been cut and reflected to display the quadrangular and triangular spaces of the axillary region.

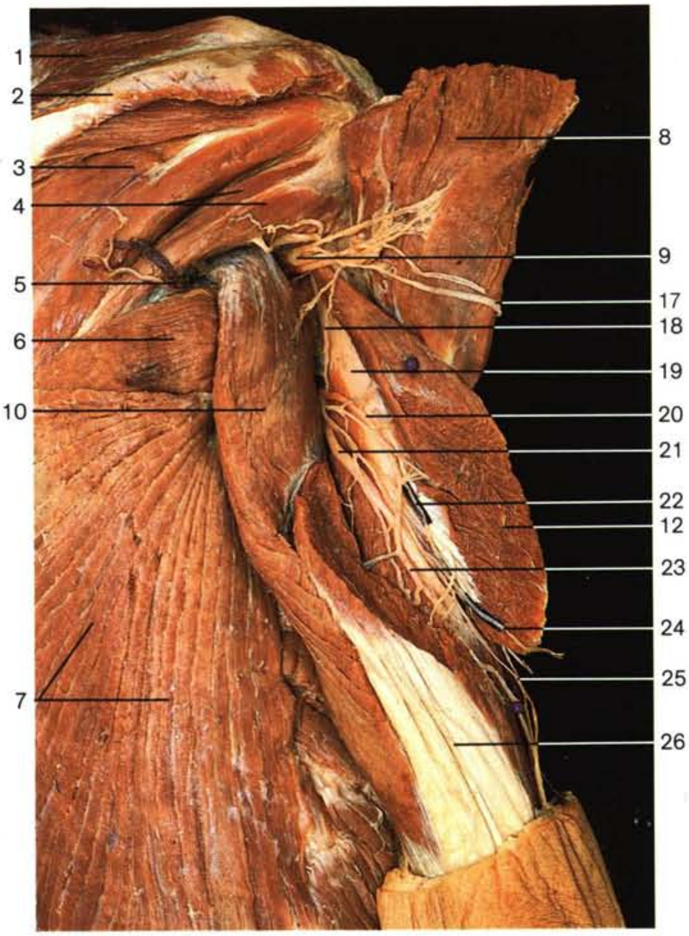

Scapular region and posterior brachial region; arm and shoulder, deep layer (dorsal aspect). The lateral head of the triceps brachii muscle has been cut to display the radial nerve and accompanying vessels.

1 Trapezius muscle
2 Spine of scapula
3 Infraspinatus muscle
4 Teres minor muscle
5 **Triangular space** containing circumflex scapular artery and vein
6 Teres major muscle
7 Latissimus dorsi muscle
8 Deltoid muscle (cut and reflected)
9 **Quadrangular space** containing axillary nerve and posterior circumflex humeral artery and vein
10 Long head of triceps brachii muscle
11 Cutaneous branch of axillary nerve
12 Lateral head of triceps brachii muscle
13 Terminal branches of intercostobrachial nerve
14 Lateral cutaneous branches of intercostal nerves
15 Medial cutaneous nerve of arm
16 Medial cutaneous nerve of forearm
17 Upper lateral cutaneous nerve of arm
18 Anastomosis between profunda brachii artery and posterior humeral circumflex artery
19 Humerus
20 **Profunda brachii artery**
21 **Radial nerve**
22 Radial collateral artery
23 Middle collateral artery
24 Lower lateral cutaneous nerve of arm
25 Posterior cutaneous nerve of forearm
26 Tendon of triceps brachii muscle

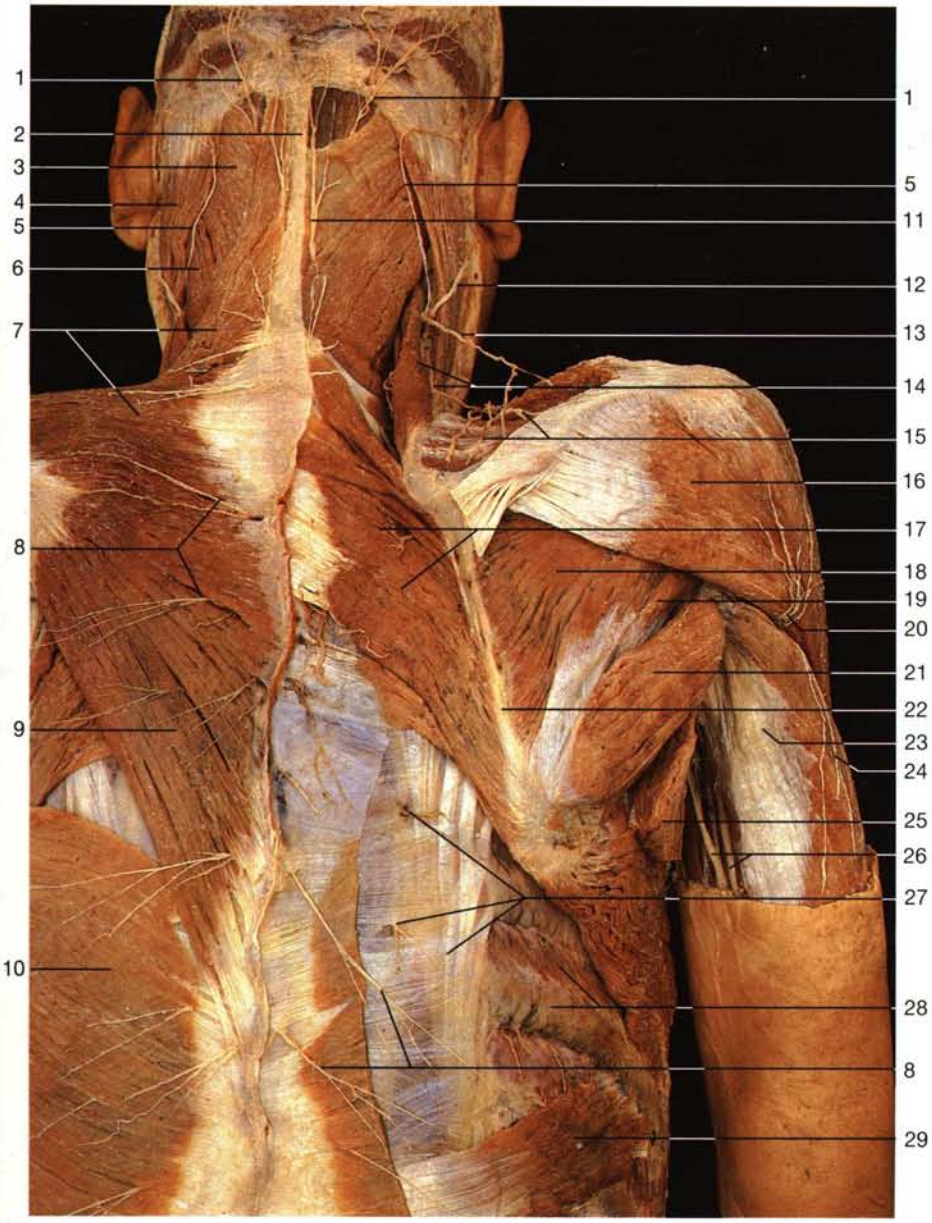

Dorsal regions of neck and shoulder (dorsal aspect). Left side: superficial layer. Right side: trapezius and latissimus dorsi muscles have been removed. Dissection of dorsal branches of spinal nerves.

1 **Greater occipital nerve**
2 Ligamentum nuchae
3 Splenius capitis muscle
4 Sternocleidomastoid muscle
5 Lesser occipital nerve
6 Splenius cervicis muscle
7 Descending and transverse fibers of trapezius muscle
8 **Medial cutaneous branches of dorsal rami of spinal nerves**
9 Ascending fibers of trapezius muscle
10 Latissimus dorsi muscle
11 Cutaneous branch of third occipital nerve
12 Great auricular nerve
13 Accessory nerve (n. XI)
14 Posterior supraclavicular nerve and levator scapulae muscle
15 Branches of suprascapular artery
16 Deltoid muscle
17 Rhomboid major muscle
18 Infraspinatus muscle
19 Teres minor muscle
20 Upper lateral cutaneous nerve of arm (branch of axillary nerve)
21 Teres major muscle
22 Medial margin of scapula
23 Long head of triceps muscle
24 Posterior cutaneous nerve of arm (branch of radial nerve)
25 Latissimus dorsi muscle (divided)
26 Ulnar nerve and brachial artery
27 **Lateral cutaneous branches of dorsal rami of spinal nerves** and iliocostalis thoracis muscle
28 External intercostal muscle and seventh rib
29 Serratus posterior inferior muscle

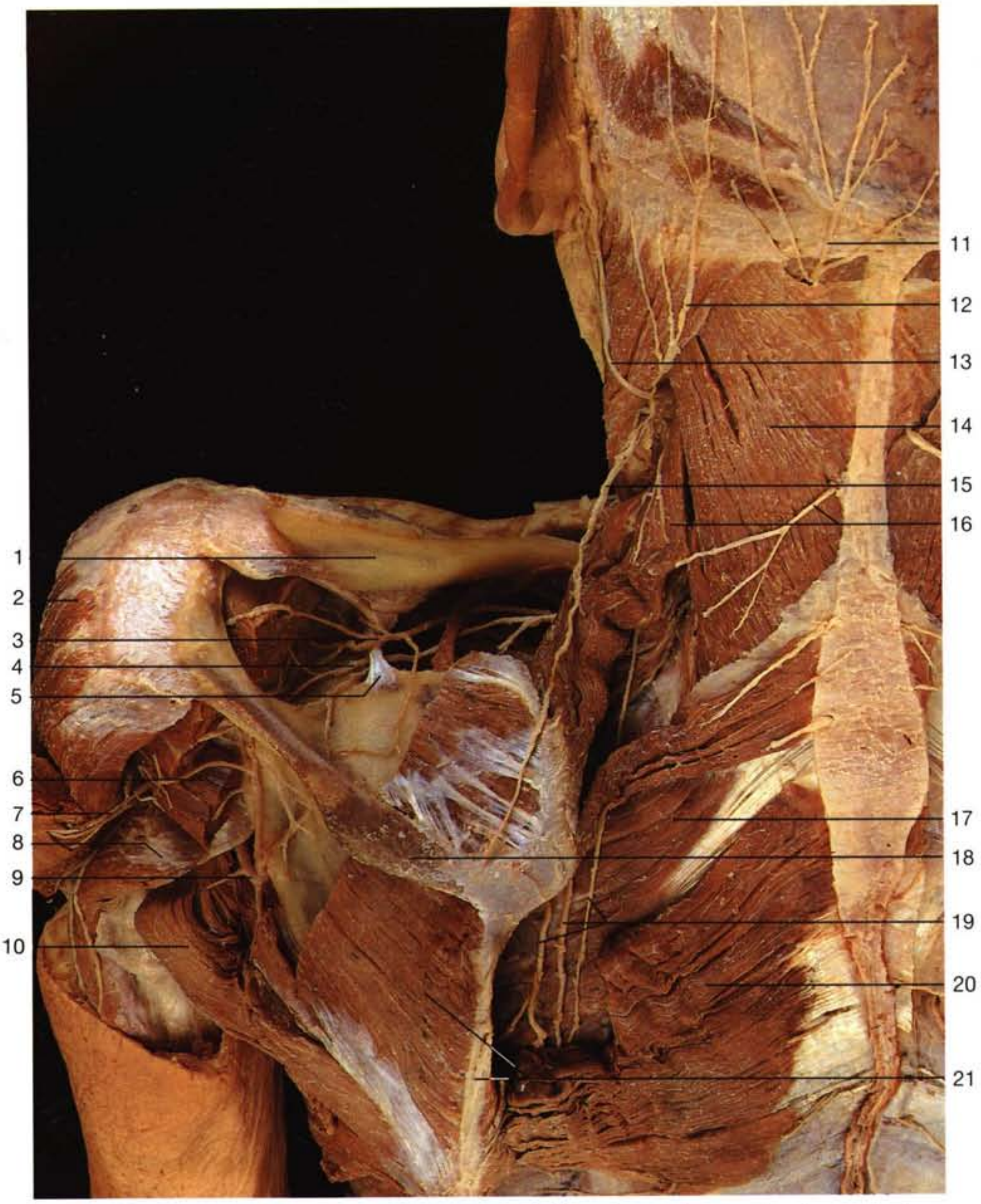

Dorsal region of shoulder; deepest layer. Rhomboid and scapular muscles fenestrated; posterior part of deltoid muscle reflected.

1 Clavicle
2 Deltoid muscle
3 **Suprascapular artery**
4 Suprascapular nerve
5 Superior transverse scapular ligament
6 Teres minor muscle
7 Axillary nerve and posterior circumflex humeral artery
8 Long head of triceps muscle
9 **Circumflex scapular artery**
10 Teres major muscle
11 Greater occipital nerve
12 Lesser occipital nerve
13 Great auricular nerve
14 Splenius capitis muscle
15 Accessory nerve (n. XI)
16 Third occipital nerve and levator scapulae muscle
17 Serratus posterior superior muscle
18 Spine of scapula
19 Descending scapular artery and dorsal scapular nerve
20 Rhomboid major muscle
21 Infraspinatus muscle and medial margin of scapula
22 Radial nerve and profunda brachii artery
23 Thoracodorsal artery
24 Thyrocervical trunk
25 Roots of brachial plexus

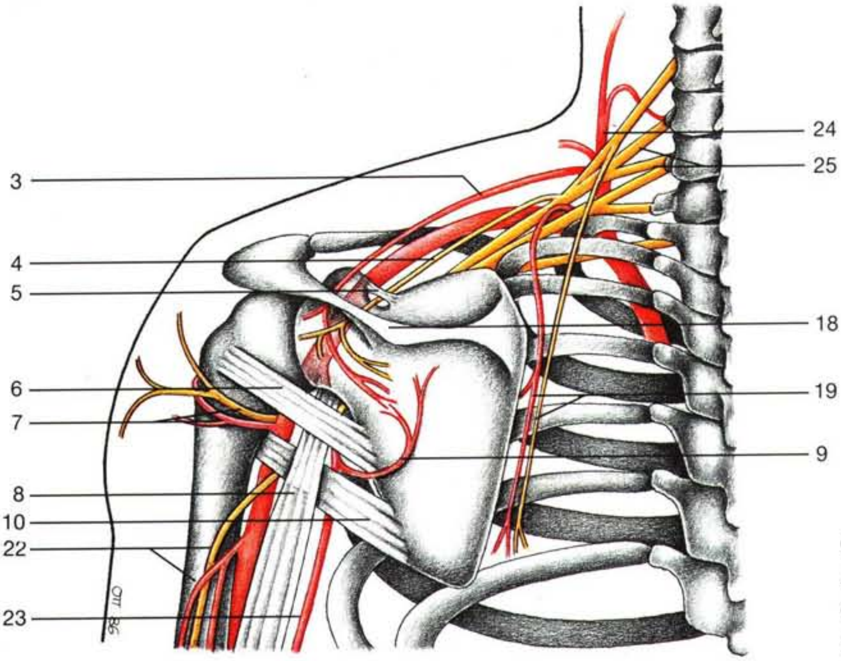

Collateral circulation of shoulder. Anastomosis of suprascapular and circumflex scapular arteries. (Semischematic drawing.)

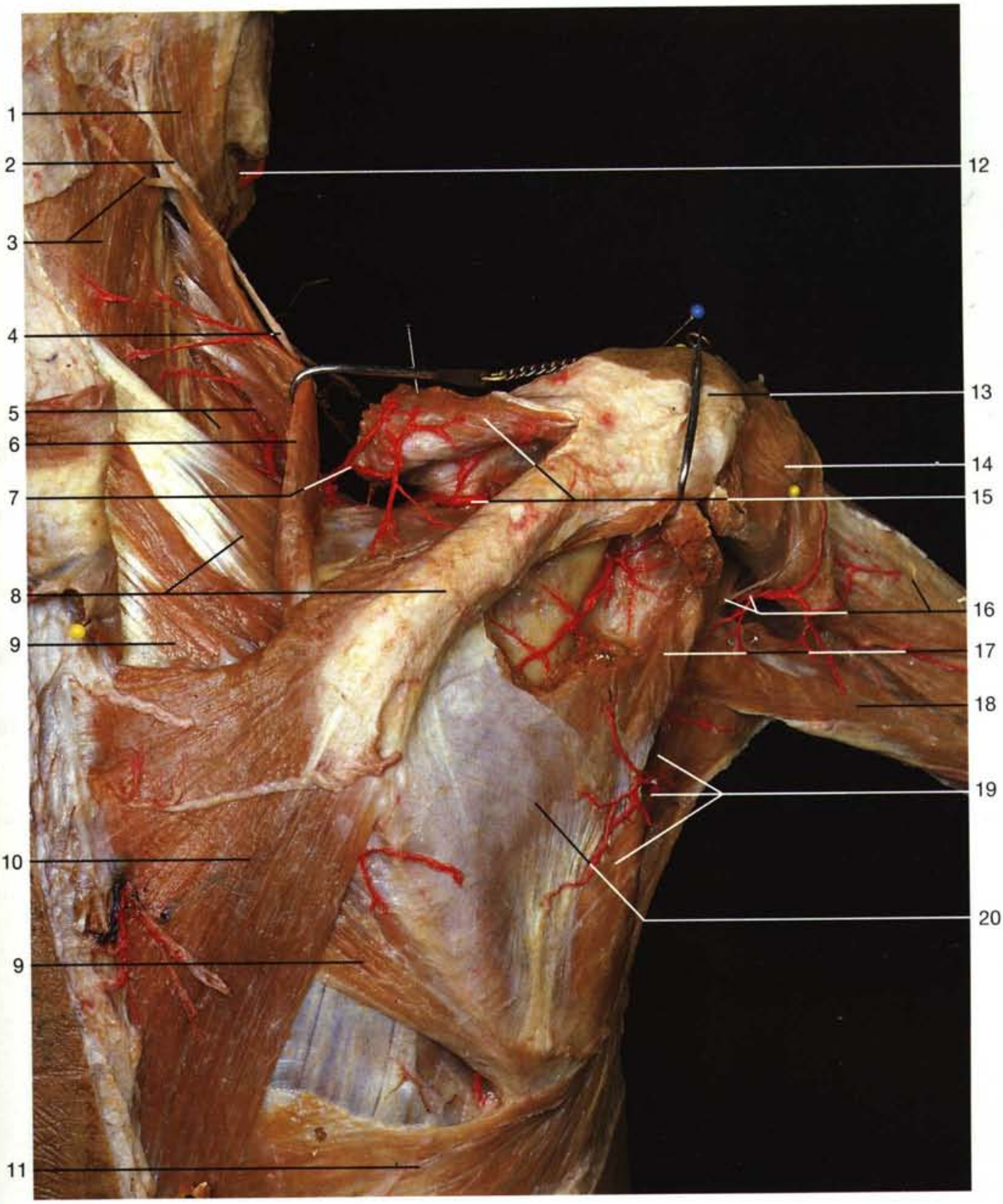

Scapular region, arm and shoulder (dorsal aspect). Arteries of scapular region are injected. Trapezius, deltoid and infraspinatus muscles are partially removed or reflected.

1	Sternocleidomastoid muscle	11	Latissimus dorsi muscle
2	Lesser occipital nerve	12	Facial artery
3	Splenius capitis muscle and third occipital nerve	13	Acromion
4	**Accessory nerve** (N. XI)	14	Deltoid muscle
5	Splenius cervicis muscle and transverse cervical artery (deep branch)	15	**Suprascapular artery** and supraspinatus muscle (reflected)
6	Levator of scapula muscle	16	**Axillary nerve,** posterior circumflex humeral artery and lateral head of triceps, brachii muscle
7	**Transverse cervical artery** (superficial branch)	17	Teres minor muscle
8	Spine of scapula and serratus posterior superior muscle	18	Long head of triceps brachii muscle
9	Rhomboid major muscle	19	**Circumflex scapular artery** and teres major
10	Trapezius muscle	20	Infraspinatus muscle

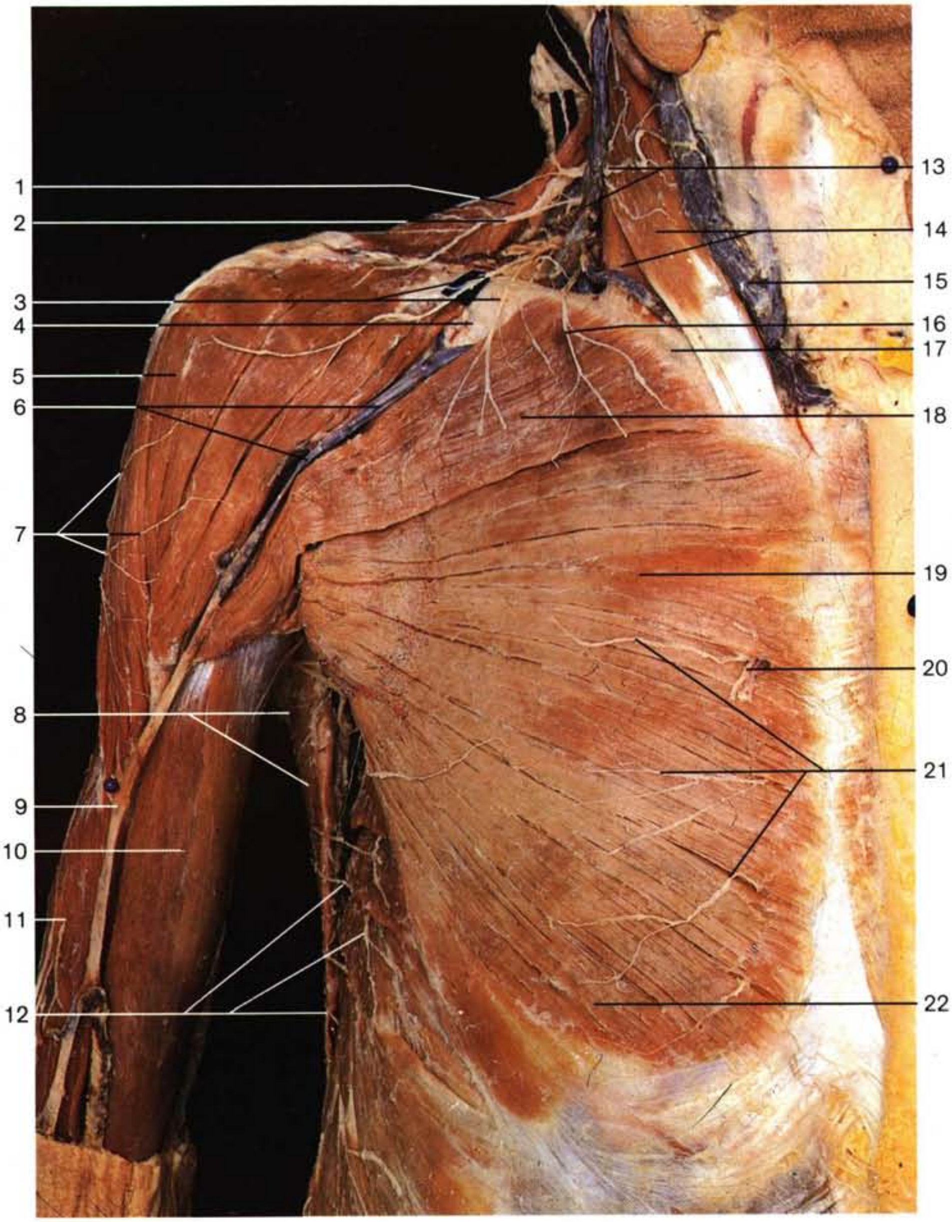

Right shoulder and thoracic wall, superficial layer (anterior aspect). Dissection of the cutaneous nerves and veins.

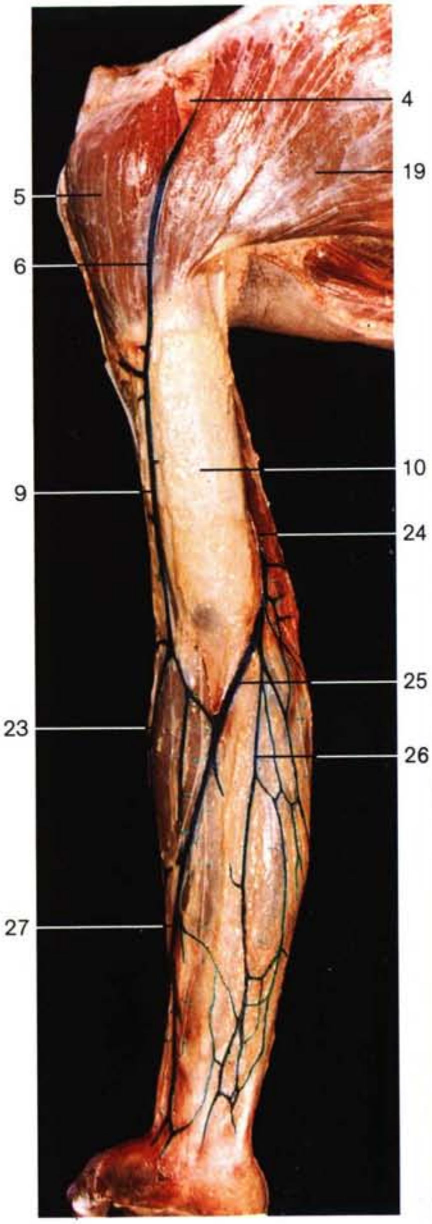

Superficial veins of right arm have been injected with blue gelatin.

1	Trapezius muscle
2	Posterior supraclavicular nerve
3	Middle supraclavicular nerve
4	Deltopectoral triangle
5	Deltoid muscle
6	Cephalic vein within the deltopectoral groove
7	Upper lateral cutaneous nerve of arm (branch of axillary nerve)
8	Latissimus dorsi muscle
9	**Cephalic vein**
10	Biceps brachii muscle
11	Triceps brachii muscle
12	Lateral cutaneous branches of intercostal nerves
13	Transverse cervical nerve and external jugular vein
14	Sternocleidomastoid muscle
15	**Anterior jugular vein**
16	Anterior supraclavicular nerve
17	Clavicle
18	Clavicular part of pectoralis major muscle
19	Sternocostal part of pectoralis major muscle
20	Perforating branch of internal thoracic artery
21	**Anterior cutaneous branches of intercostal nerves**
22	Abdominal part of pectoralis major muscle
23	Accessory cephalic vein
24	**Basilic vein**
25	**Median cubital vein**
26	Median vein of forearm
27	Cephalic vein in the forearm

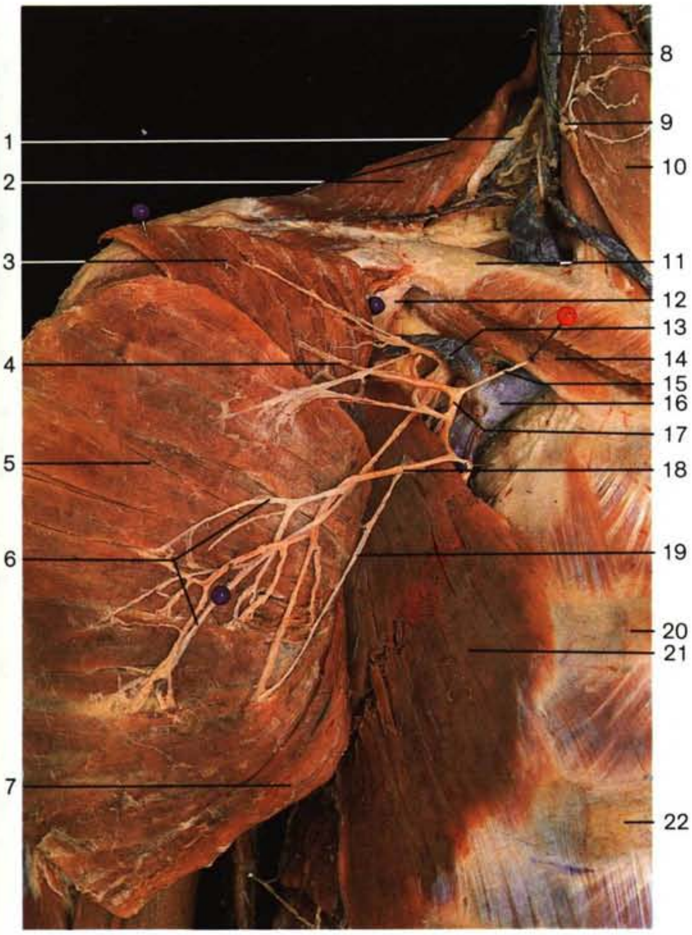

Right deltopectoral triangle, infraclavicular region (anterior aspect). The pectoralis major muscle has been cut and reflected.

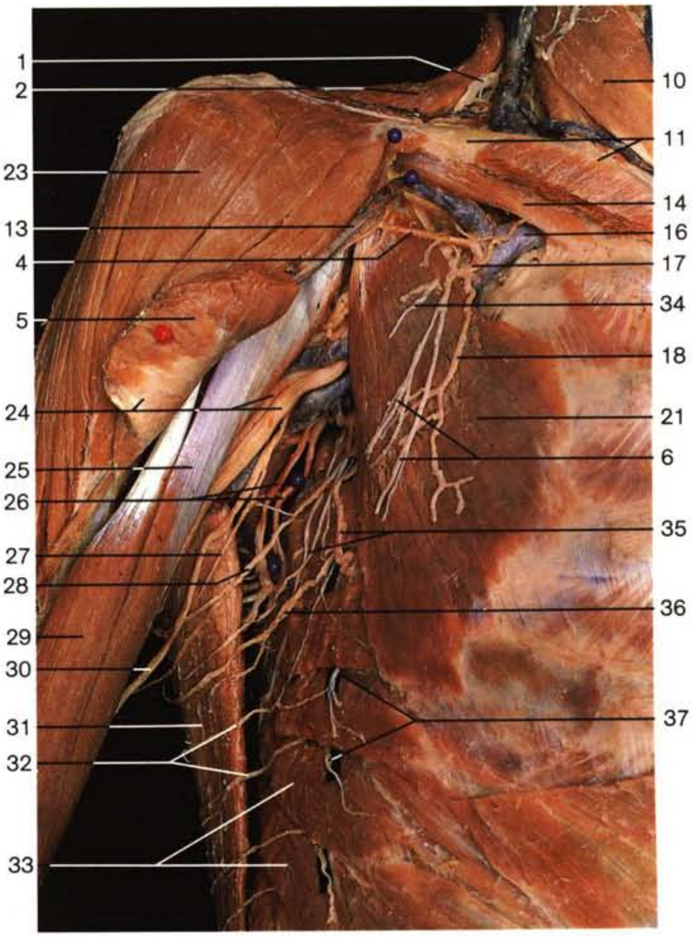

Thoracic wall and shoulder, deep layer. **Right axillary region** (anterior aspect). The pectoralis major muscle has been cut and partly removed.

1 Accessory nerve
2 Trapezius muscle
3 Pectoralis major muscle (clavicular part)
4 Acromial branch of thoracoacromial artery
5 Pectoralis major muscle
6 **Lateral pectoral nerves**
7 Abdominal part of pectoralis major muscle
8 External jugular vein
9 Cutaneous branches of cervical plexus
10 Sternocleidomastoid muscle
11 Clavicle
12 Clavipectoral fascia
13 **Cephalic vein**
14 Subclavius muscle
15 Clavicular branch of thoracoacromial artery
16 **Subclavian vein**
17 **Thoracoacromial artery**
18 Pectoral branch of thoracoacromial artery
19 Medial pectoral nerve
20 2nd rib
21 Pectoralis minor muscle
22 3rd rib
23 Deltoid muscle
24 Pectoralis major muscle (reflected), **brachial artery and median nerve**
25 Short head of biceps brachii muscle
26 Thoracodorsal artery and nerve
27 Medial cutaneous nerve of arm
28 Intercostobrachial nerve (T_2)
29 Long head of biceps brachii muscle
30 Medial cutaneous nerve of forearm
31 Latissimus dorsi muscle
32 Lateral cutaneous branches of intercostal nerves (posterior branches)
33 Serratus anterior muscle
34 Medial pectoral nerve
35 **Long thoracic nerve and lateral thoracic artery**
36 Intercostobrachial nerve (T_3)
37 **Lateral cutaneous branches of intercostal nerves** (anterior branches)

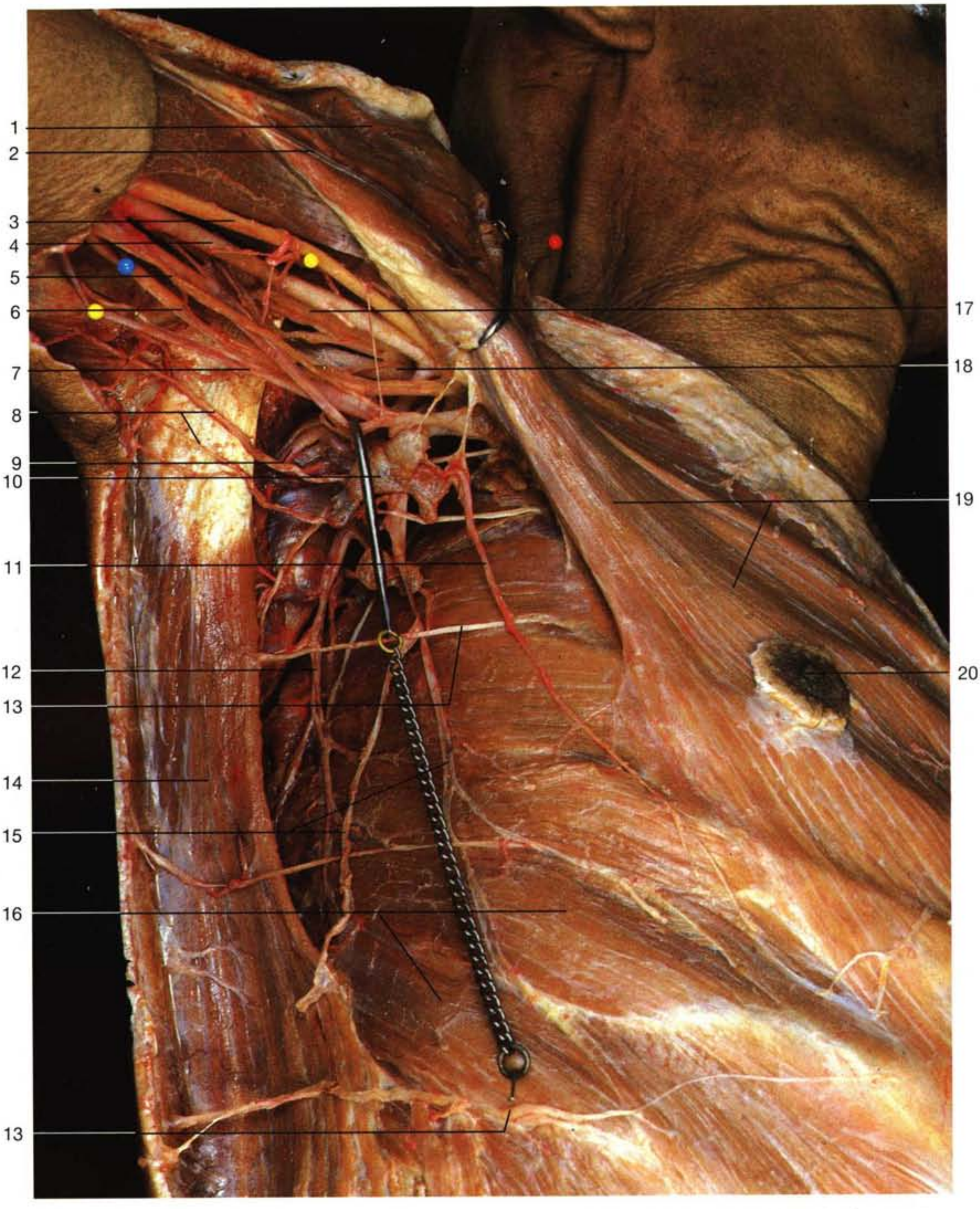

Right axillary region (inferior aspect). **Dissection of superficial axillary nodes and lymphatic vessels.** The pectoralis major muscle has been slightly elevated.

1	Deltoid muscle	12	Thoracodorsal artery
2	Cephalic vein	13	Lateral cutaneous branch of intercostal nerve
3	Median nerve	14	Latissimus dorsi muscle
4	Brachial artery	15	Thoracoepigastric vein
5	Medial cutaneous nerves of arm and forearm	16	Serratus anterior muscle
6	Ulnar nerve	17	Musculocutaneous nerve
7	Basilic Vein	18	Radial nerve
8	Intercostobrachial nerves	19	Pectoralis major muscle
9	Circumflex scapular artery	20	Nipple
10	**Superficial axillary nodes**		
11	Lateral thoracic artery		

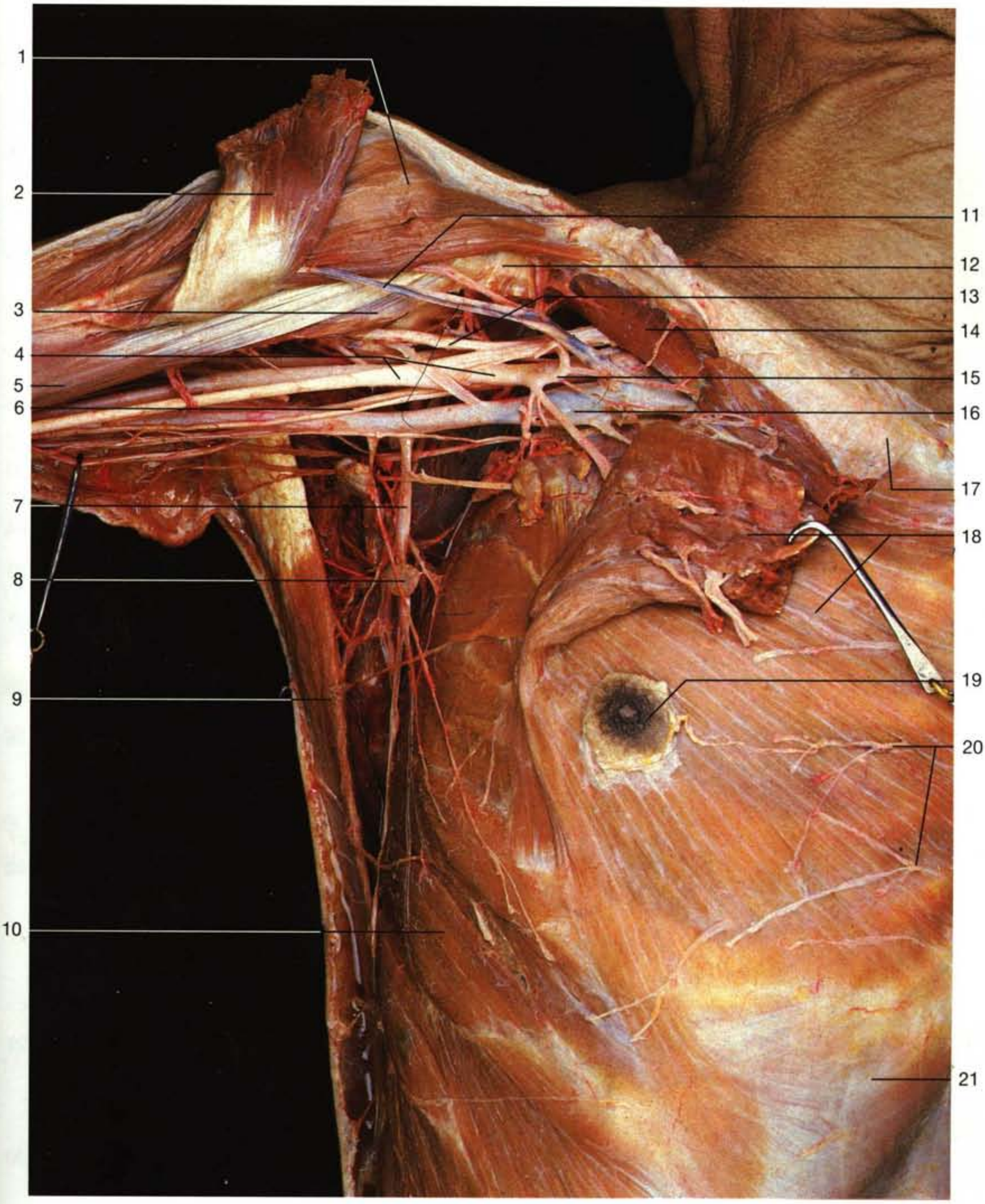

Axillary region (anterior aspect). Dissection of deep axillary nodes. Pectoralis major and minor divided and reflected. Shoulder girdle and arm elevated and reflected.

1	Deltoid muscle	13	**Musculocutaneous nerve**
2	Insertion of pectoralis major muscle	14	Subclavius muscle
3	Coracobrachialis muscle	15	Thoracoacromial artery
4	Roots of **median nerve, axillary artery**	16	**Axillary vein**
5	Short head of biceps brachii muscle	17	Clavicle
6	**Ulnar nerve** and medial cutaneous nerve of forearm	18	Pectoralis major and minor muscles (reflected)
7	Thoracoepigastric vein	19	Nipple
8	Deep axillary node	20	Anterior cutaneous branches of intercostal nerves
9	Latissimus dorsi muscle	21	Anterior layer of rectus sheaths
10	Serratus anterior muscle		
11	Cephalic vein		
12	Insertion of pectoralis minor muscle (coracoid process)		

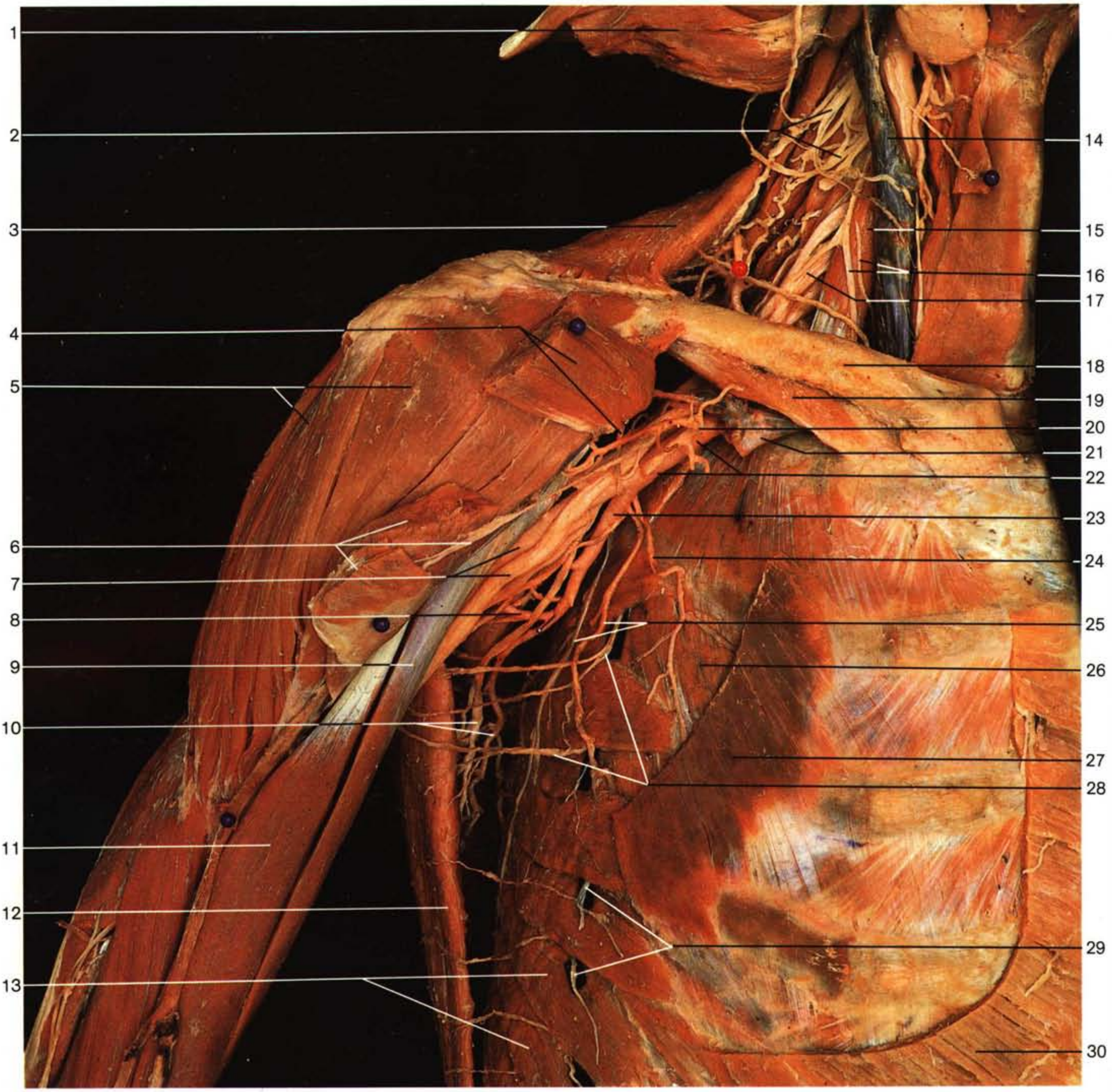

Right axillary region (anterior aspect). The pectoralis major and minor muscles have been cut and reflected to display the vessels and nerves of the axilla.

1 Sternocleidomastoid muscle (cut and reflected)
2 Cervical plexus
3 Trapezius muscle
4 Pectoralis minor muscle and medial pectoral nerve
5 Deltoid muscle
6 Pectoralis major muscle and lateral pectoral nerve
7 **Median nerve and brachial artery**
8 Circumflex scapular artery
9 Short head of biceps brachii muscle
10 **Thoracodorsal artery and nerve**
11 Long head of biceps brachii muscle
12 Latissimus dorsi muscle
13 Serratus anterior muscle
14 Internal jugular vein
15 **Scalenus anterior muscle**
16 Phrenic nerve and ascending cervical artery
17 Brachial plexus (at the levels of the trunks)
18 Clavicle
19 Subclavius muscle
20 **Thoracoacromial artery**
21 Subclavian vein (cut)
22 **Axillary artery**
23 Subscapular artery
24 Superior thoracic artery
25 **Lateral thoracic artery and long thoracic nerve**
26 External intercostal muscle
27 Insertion of pectoralis minor muscle
28 Intercostobrachial nerves
29 Lateral cutaneous branches of intercostal nerves
30 Insertion of pectoralis major muscle

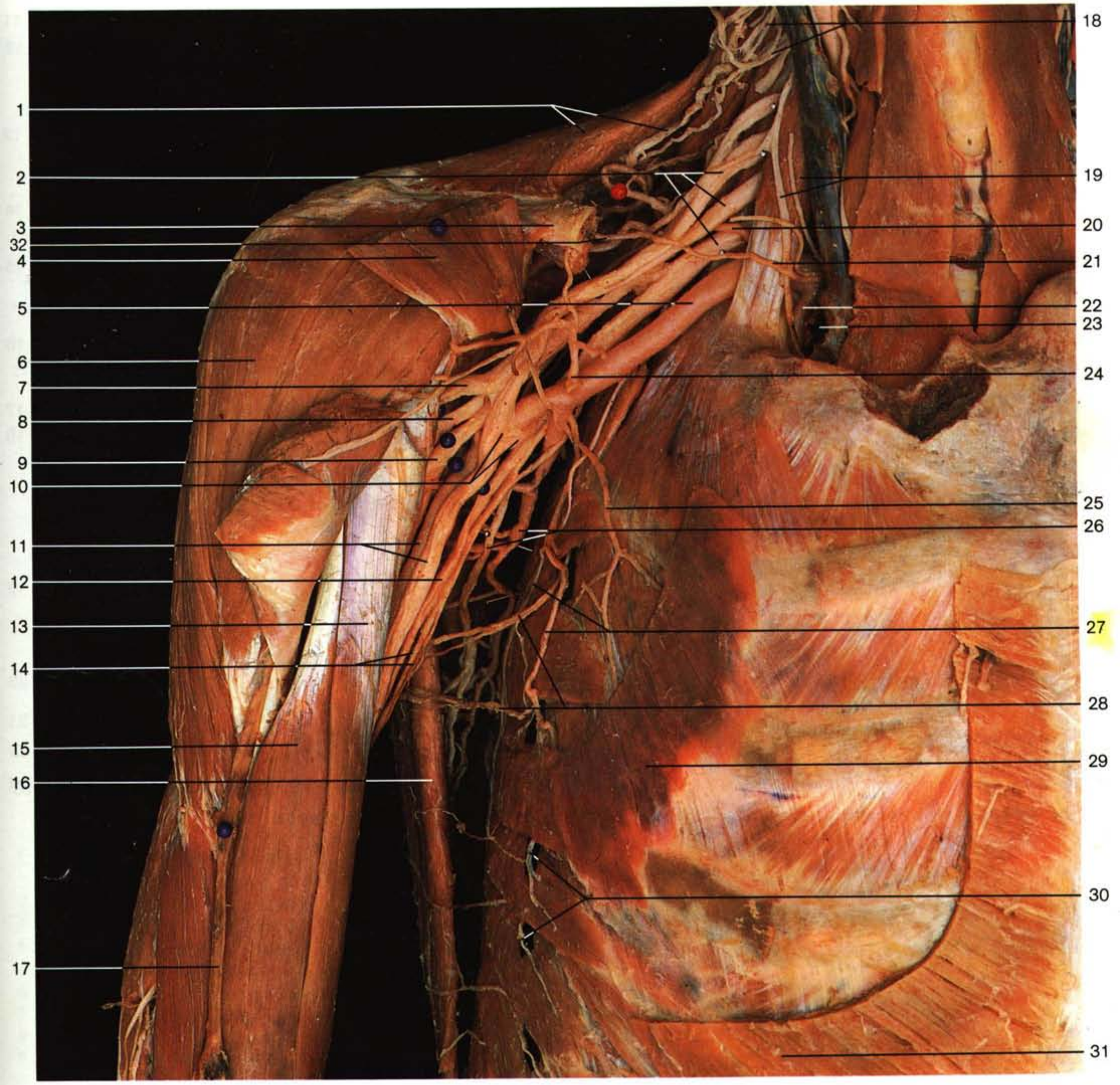

Shoulder and arm. Axillary region with brachial plexus (anterior aspect). The pectoralis major and minor muscles have been partly removed. The clavicle has been cut.

1 Trapezius muscle and accessory nerve
2 Brachial plexus
3 Clavicle (cut)
4 Pectoralis minor muscle
5 **Axillary (subclavian) artery**
6 Deltoid muscle
7 **Musculocutaneous nerve**
8 **Axillary nerve**
9 **Radial nerve**
10 Medial and lateral roots of the median nerve
11 **Median nerve** and **brachial artery**
12 **Ulnar nerve**
13 Short head of biceps brachii muscle
14 Medial cutaneous nerves of arm and forearm
15 Long head of biceps brachii muscle
16 Latissimus dorsi muscle
17 Cephalic vein
18 Cervical plexus
19 **Phrenic nerve and scalenus anterior muscle**
20 Descending scapular artery
21 Suprascapular artery
22 Internal thoracic artery
23 Subclavian vein
24 Thoracoacromial artery
25 Superior thoracic artery
26 **Thoracodorsal artery and nerve**
27 **Lateral thoracic artery and long thoracic nerve**
28 Intercostobrachial nerves
29 Insertion of the pectoralis minor muscle into the 2nd–5th ribs
30 Lateral cutaneous branches of intercostal nerves
31 Abdominal part of pectoralis major muscle (cut)
32 Suprascapular nerve and artery

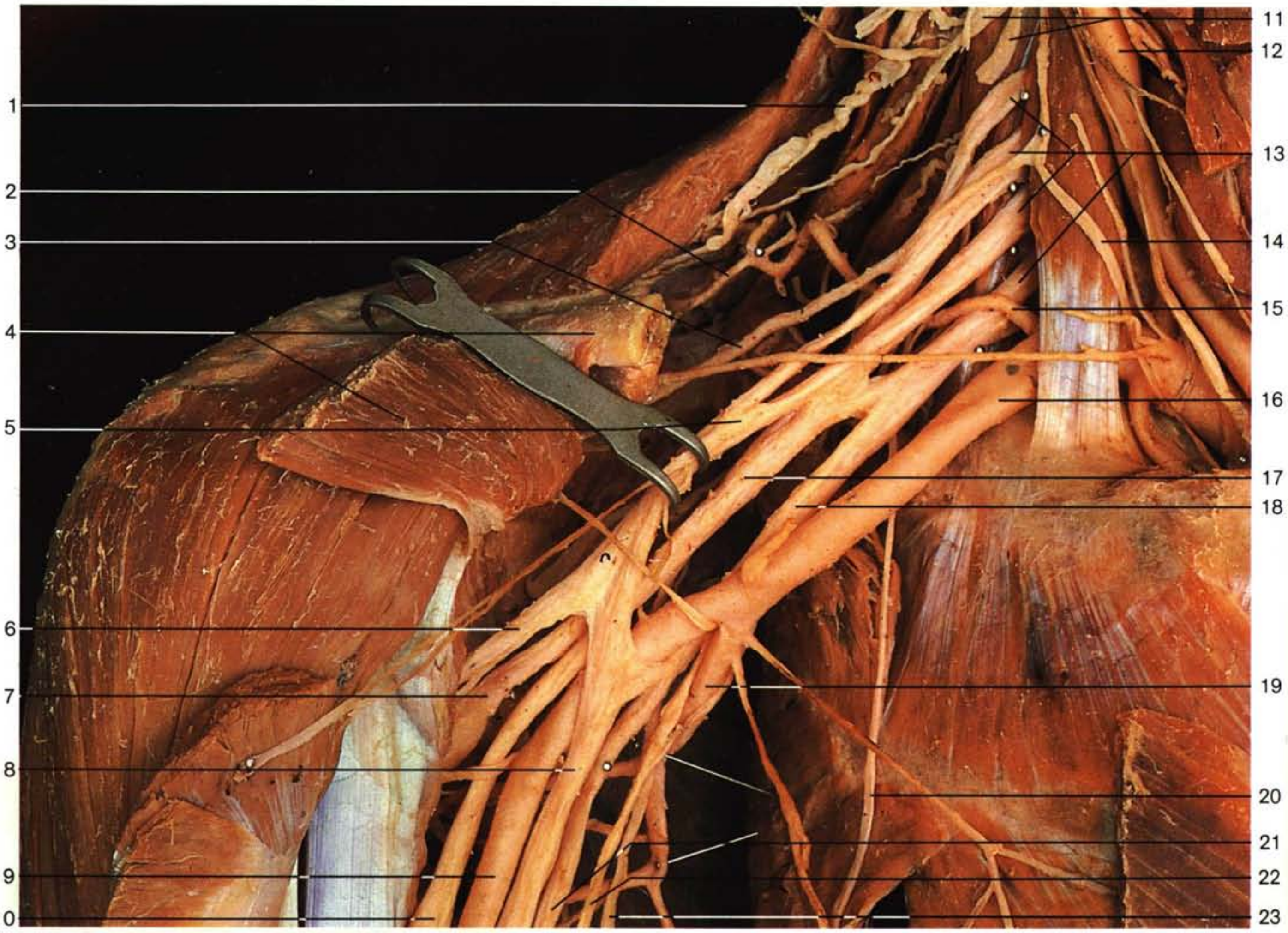

Brachial plexus (anterior aspect). Clavicle and the two pectoralis muscles have been partly removed.

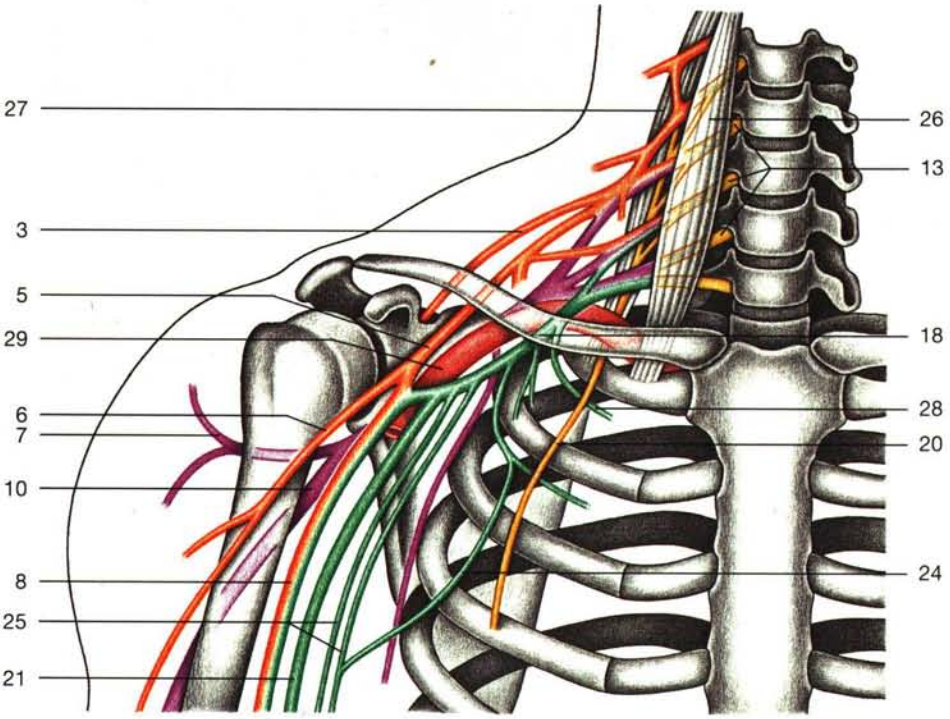

Main branches of brachial plexus. Posterior cord in red, lateral cord in yellow and medial cord in purple. (Schematic drawing.)

1 Accessory nerve
2 Dorsal scapular artery
3 Suprascapular nerve
4 Clavicle and pectoralis minor muscle
5 **Lateral cord** of brachial plexus
6 Musculocutaneous nerve
7 Axillary nerve
8 Median nerve
9 Brachial artery
10 Radial nerve
11 Cervical plexus
12 Common carotid artery
13 Roots of brachial plexus (C_5–T_1)
14 Phrenic nerve
15 Dorsal scapular artery
16 Subclavian artery
17 **Posterior cord** of brachial plexus
18 **Medial cord** of brachial plexus
19 Subscapular artery
20 Long thoracic nerve
21 Ulnar nerve
22 Medial cutaneous nerve of forearm
23 Thoracodorsal nerve
24 Intercostobrachial nerve
25 Medial cutaneous nerves of arm and forearm
26 Scalenus anterior muscle
27 Scalenus medius muscle
28 Intercostal nerve (T_1)
29 Axillary artery

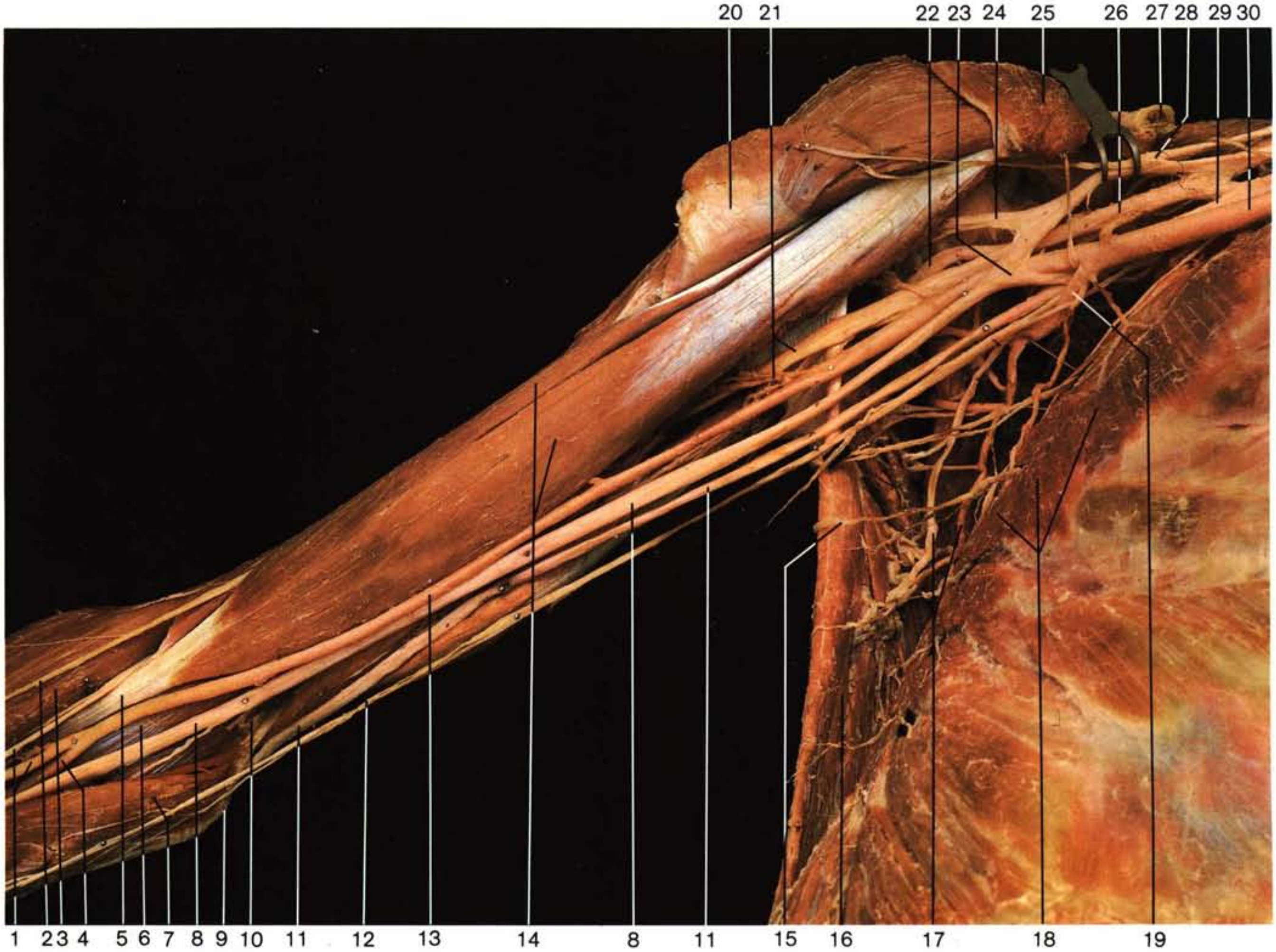

Right arm. Dissection of vessels and nerves (medial aspect). Shoulder girdle has been reflected slightly.

1 Radial artery and superficial branch of radial nerve
2 Lateral cutaneous nerve of forearm
3 Brachioradialis muscle
4 Ulnar artery
5 Tendon of biceps brachii muscle
6 Brachialis muscle
7 Pronator teres muscle
8 **Median nerve**
9 Medial epicondyle of humerus
10 Inferior ulnar collateral artery
11 **Ulnar nerve**
12 Medial cutaneous nerve of forearm
13 **Brachial artery**
14 Biceps brachii muscle
15 Intercostobrachial nerve (T_3)
16 Latissimus dorsi muscle
17 Thoracodorsal nerve and artery
18 Serratus anterior muscle
19 Subscapular artery
20 Pectoralis major muscle (reflected) and lateral pectoral nerve
21 Radial nerve and profunda brachii artery
22 Axillary nerve
23 Roots of the median nerve with axillary artery
24 **Musculocutaneous nerve**
25 Pectoralis minor muscle (reflected) and medial pectoral nerve
26 Posterior cord of brachial plexus
27 Clavicle (cut)
28 Lateral cord of brachial plexus
29 Medial cord of brachial plexus
30 Subclavian artery
31 Brachial vein

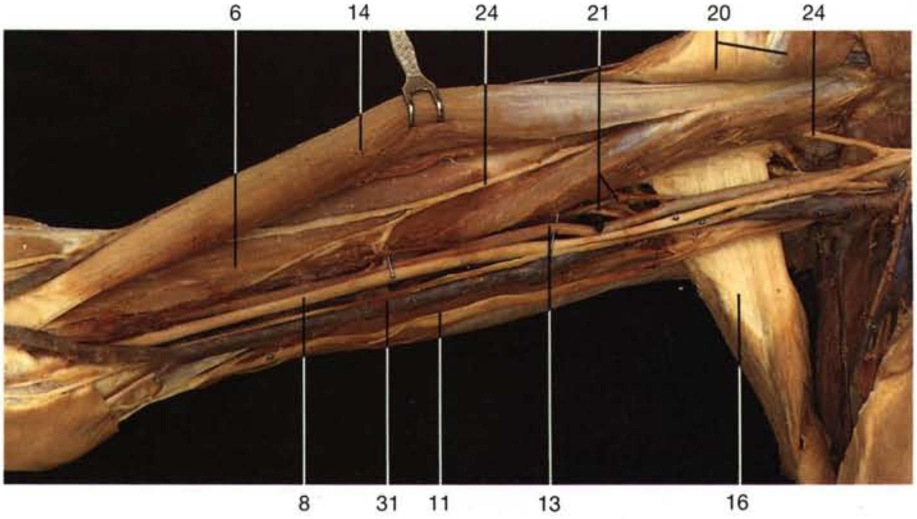

Right arm. Dissection of vessels and nerves, deeper layer. Biceps muscle has been reflected.

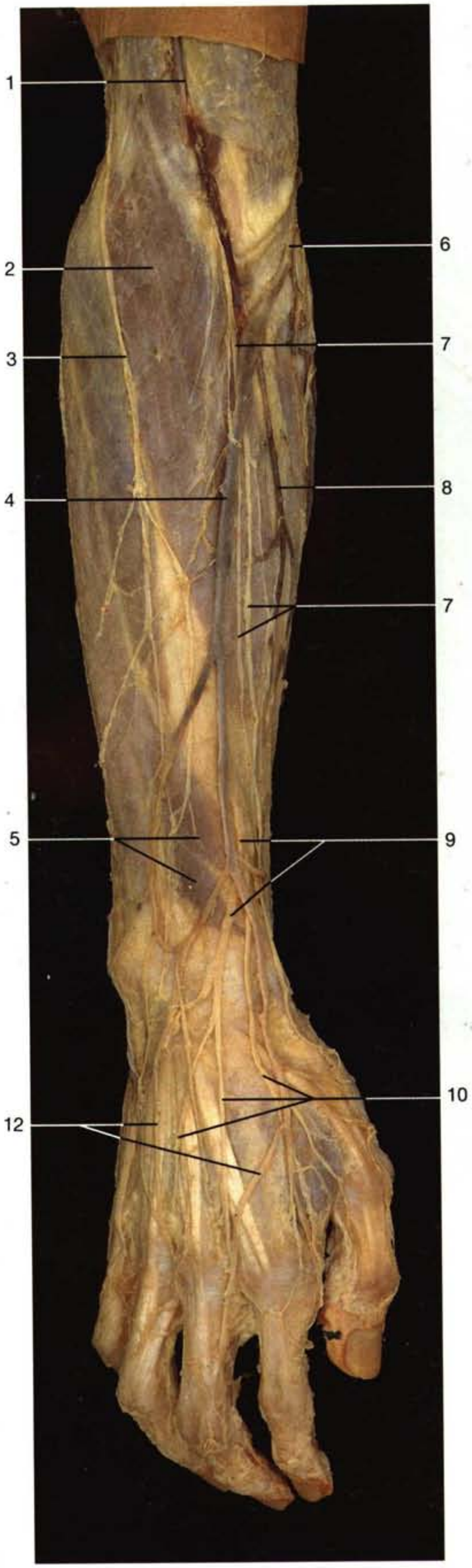

Superficial veins and cutaneous nerves of forearm and hand (anteromedial aspect).

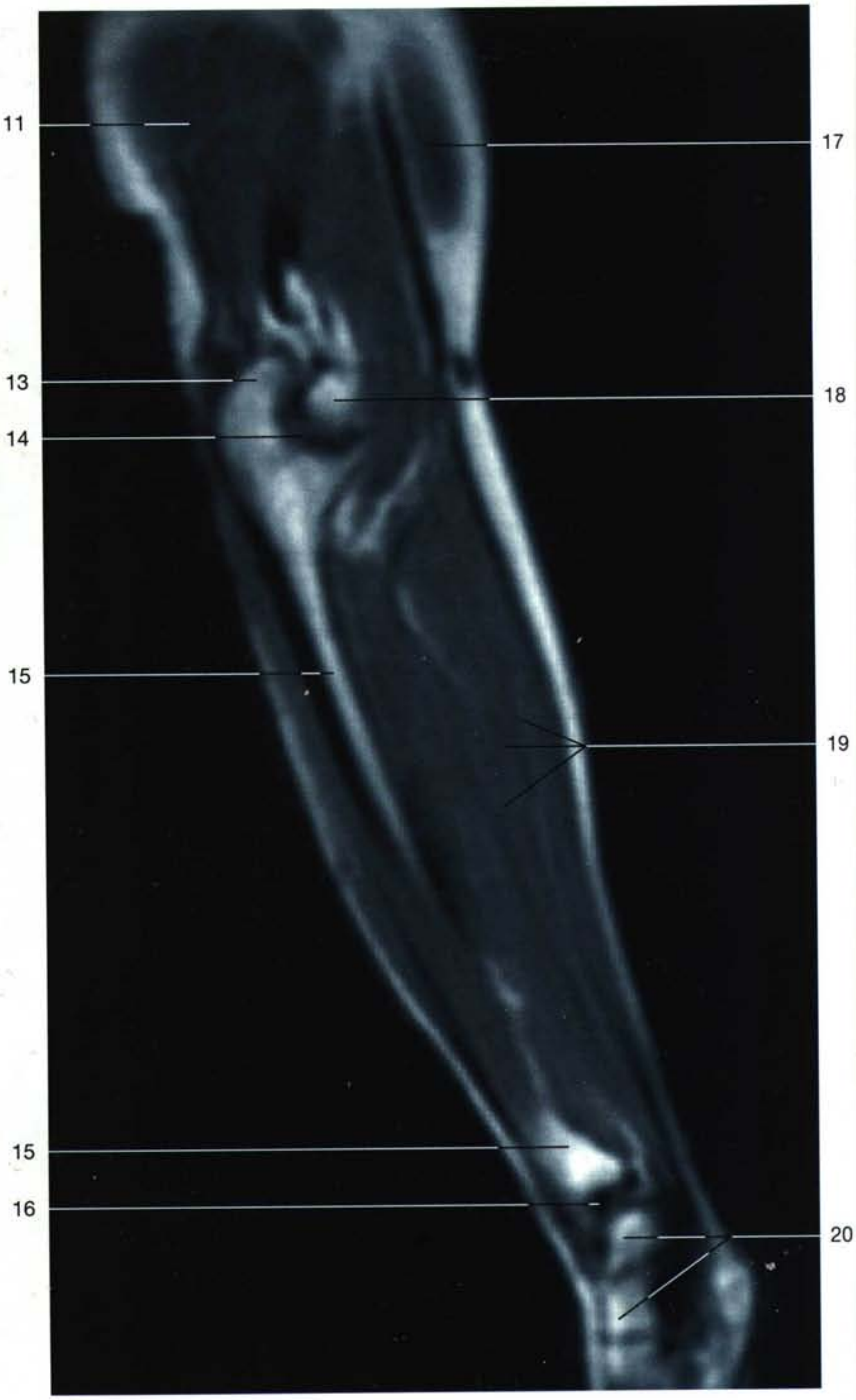

Sagittal section through the forearm at the level of the humeroulnar articulation (MR-Scan, after A. Heuck, G. Luttke, and J. W. Rohen, 1994).

1 **Cephalic vein**
2 Brachioradialis muscle covered by its fascia
3 **Posterior cutaneous nerve of forearm** (branch of radialis nerve)
4 Cephalic vein of forearm
5 Extensor pollicis longus and brevis muscles covered by their fascia
6 **Median cubital vein**
7 Lateral cutaneous nerves of forearm (branch of musculocutaneous nerve)
8 Intermedian vein of forearm
9 Superficial branch of radial nerve
10 **Dorsal digital branches of radial nerve**
11 Triceps brachii muscle
12 Venous network on the dorsum of the hand
13 Olecranon
14 **Humeroulnar articulation**
15 Ulna
16 Radiocarpal articulation
17 Biceps brachii muscle
18 Trochlea of humerus
19 Flexor muscles of forearm
20 Carpal bones

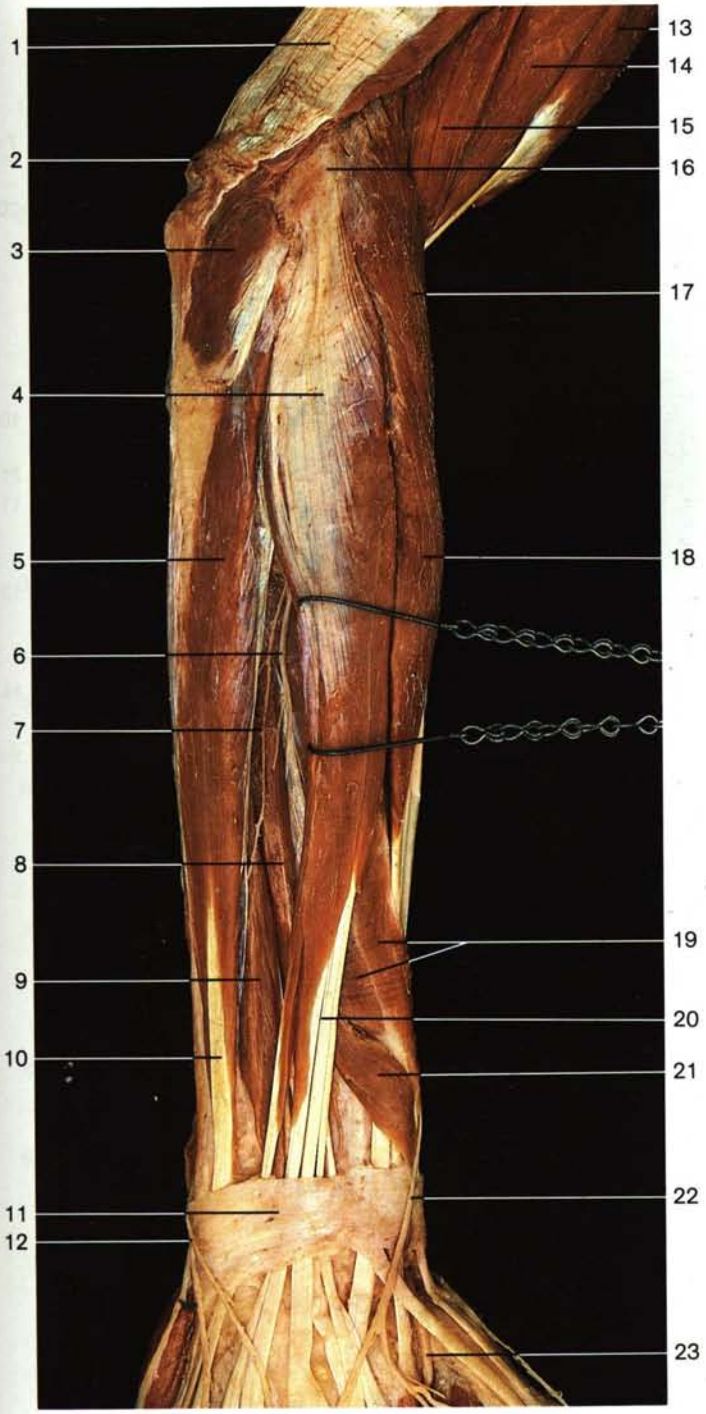

Vessels and nerves of right forearm, superficial layer (dorsal aspect).

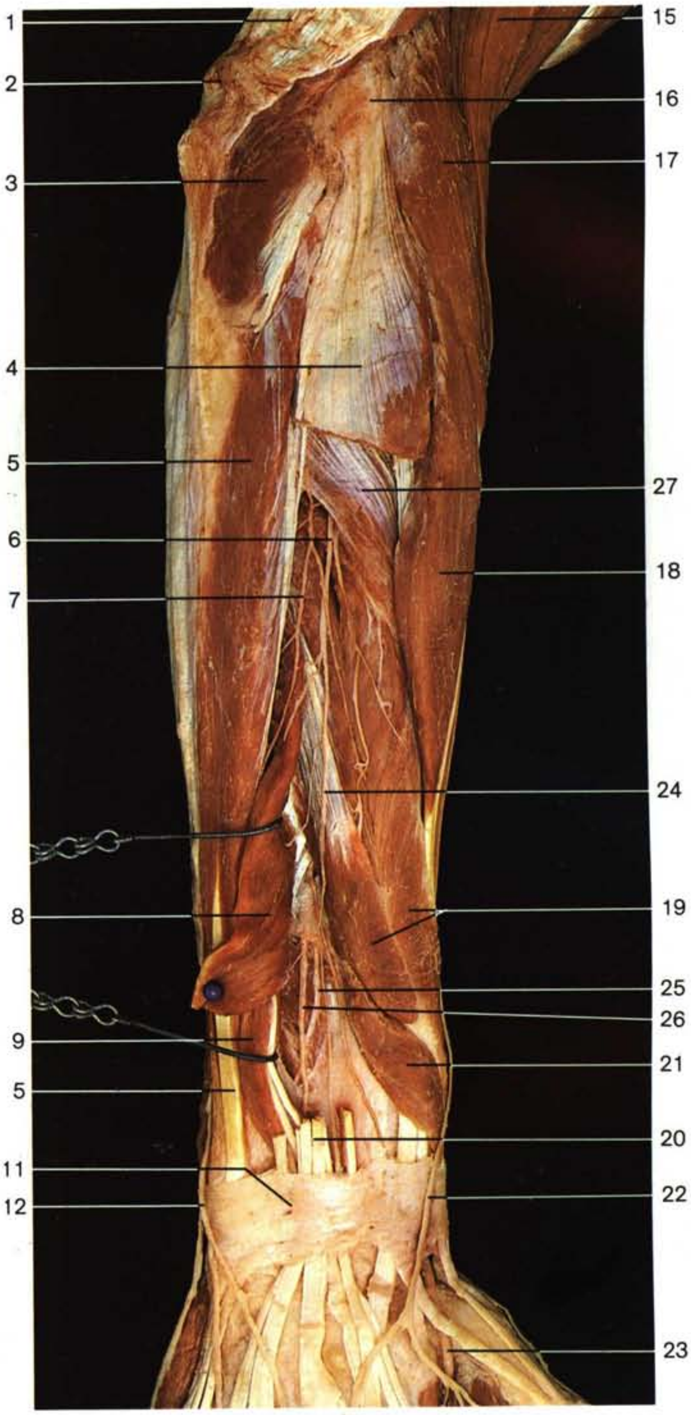

Vessels and nerves of right forearm, deep layer (dorsal aspect).

1 Tendon of triceps brachii muscle
2 Olecranon
3 Anconeus muscle
4 Extensor digitorum muscle
5 Extensor carpi ulnaris muscle
6 Deep branch of radial nerve
7 Posterior interosseous artery
8 Extensor pollicis longus muscle
9 Extensor indicis muscle
10 Tendon of extensor carpi ulnaris muscle
11 Extensor retinaculum
12 Dorsal branch of ulnar nerve
13 Biceps brachii muscle
14 Brachialis muscle
15 Brachioradialis muscle
16 Lateral epicondyle of humerus
17 Extensor carpi radialis longus muscle
18 Extensor carpi radialis brevis muscle
19 Abductor pollicis longus muscle
20 Tendons of extensor digitorum muscle
21 Extensor pollicis brevis muscle
22 Superficial branch of radial nerve
23 Radial artery
24 Posterior interosseous nerve
25 Posterior interosseous branch of radial nerve
26 Posterior branch of anterior interosseous artery
27 Supinator muscle

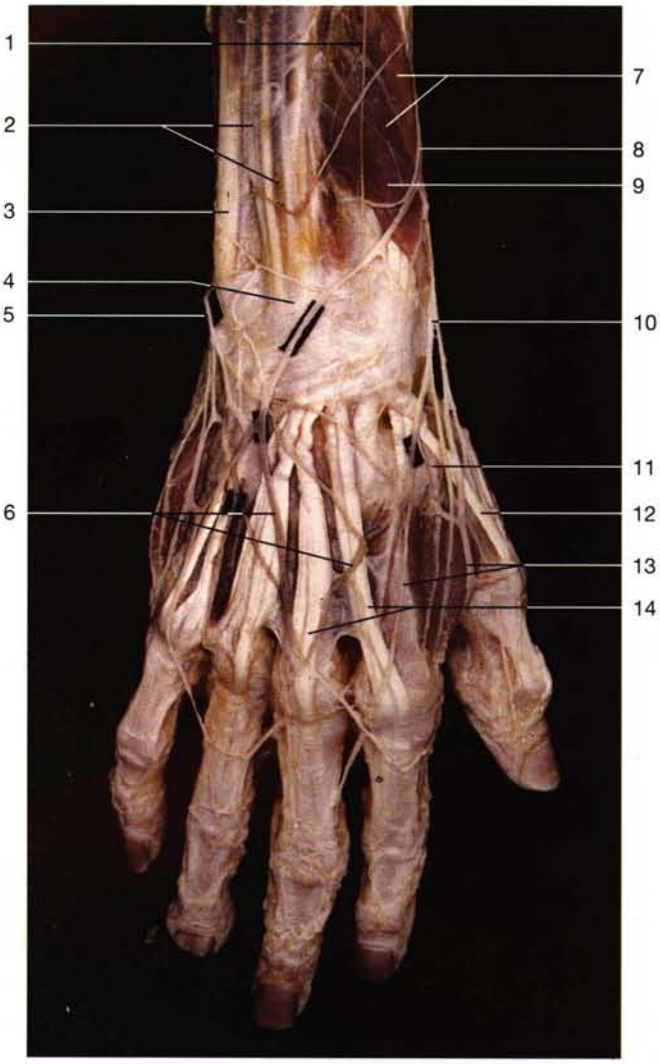

Cutaneous nerves and veins of forearm and hand (superficial layer, dorsal aspect).

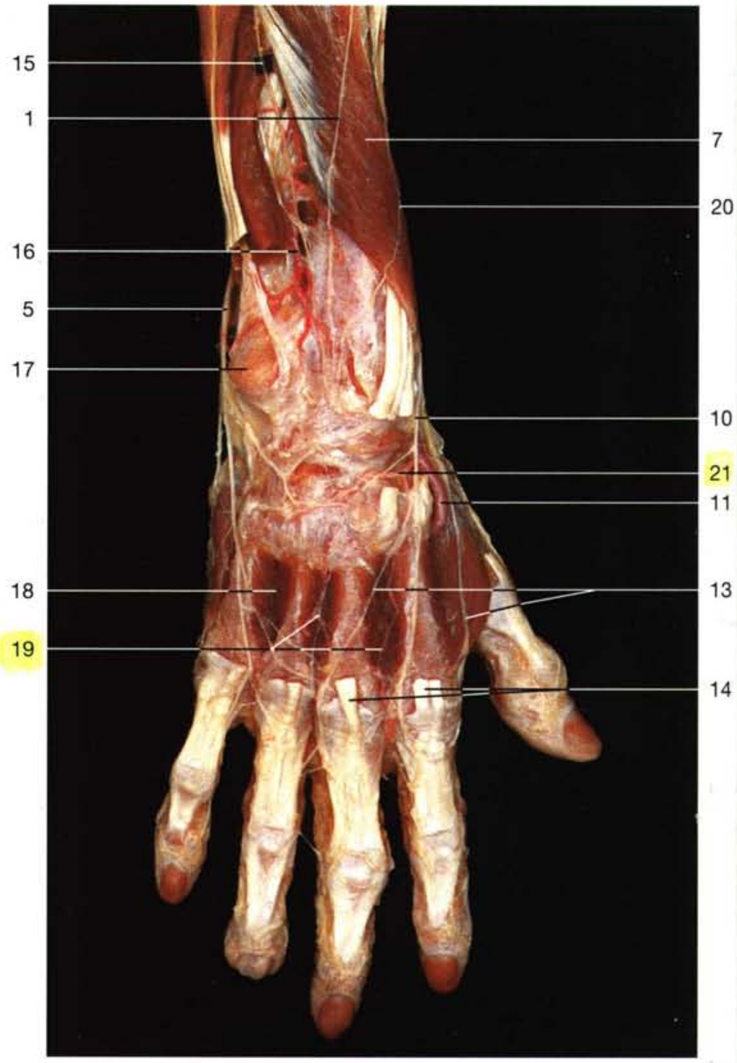

Dorsal region of forearm and hand (deeper layer). Extensor digitorum muscle has been partly removed.

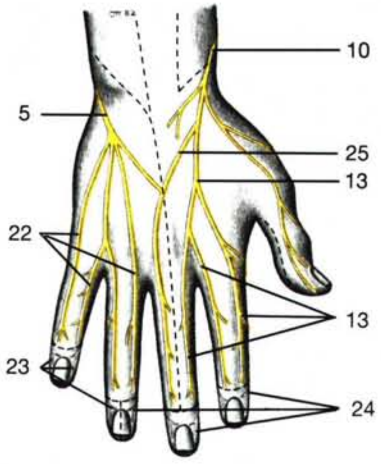

Innervation pattern of dorsal surfaces of hand
$2^1/_2$ digits by radial nerve, $2^1/_2$ digits by ulnar nerve. Note that the terminal branches to the dorsal surfaces of the distal phalanges are derived from the palmar digital nerves. The cutaneous distribution varies; often $3^1/_2$ digits are innervated by the radial and $1^1/_2$ digits by the ulnar nerve.

1 **Posterior cutaneous nerve of forearm** (branch of radial nerve)
2 Extensor digitorum muscle
3 Tendon of extensor carpi ulnaris muscle
4 Extensor retinaculum
5 **Ulnar nerve**
6 **Venous network of dorsum of hand**
7 Abductor pollicis longus muscle
8 Cephalic vein
9 Extensor pollicis brevis muscle
10 **Radial nerve,** superficial branch
11 **Radial artery**
12 Tendon of extensor pollicis longus muscle
13 Dorsal digital branches of radial nerve
14 Tendons of extensor digitorum muscle with intertendinous connections
15 Posterior interosseus nerve (branch of the deep radial nerve)
16 Posterior interosseous artery
17 Styloid process of ulna
18 Dorsal interosseus muscle IV
19 Dorsal carpal branch of radial artery
20 **Lateral cutaneous nerve of forearm** (branch of musculocutaneous nerve)
21 Dorsal metacarpal artery
22 Proper dorsal digital branches of ulnar nerve
23 Regions supplied by palmar digital nerves (ulnar nerve)
24 Regions supplied by palmar digital nerves (median nerve)
25 Communicating branch with ulnar nerve

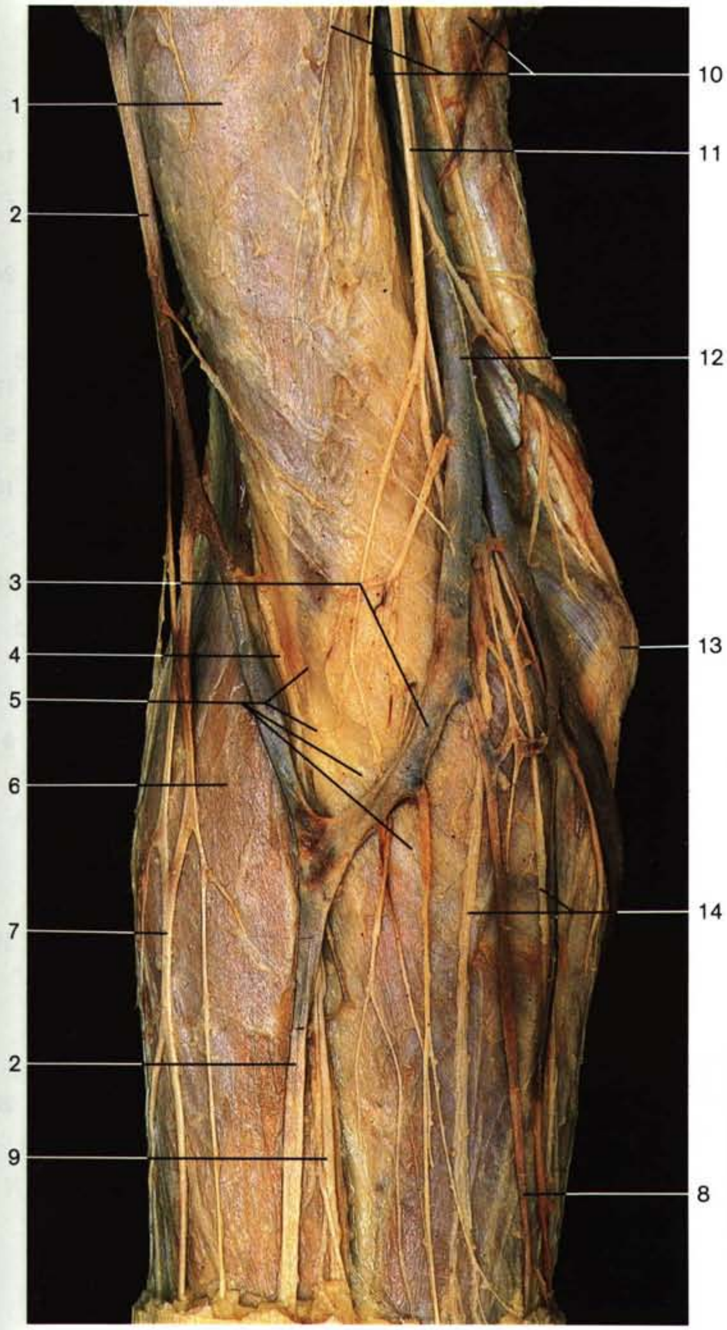

Cubital region (anterior aspect), dissection of cutaneous nerves and veins.

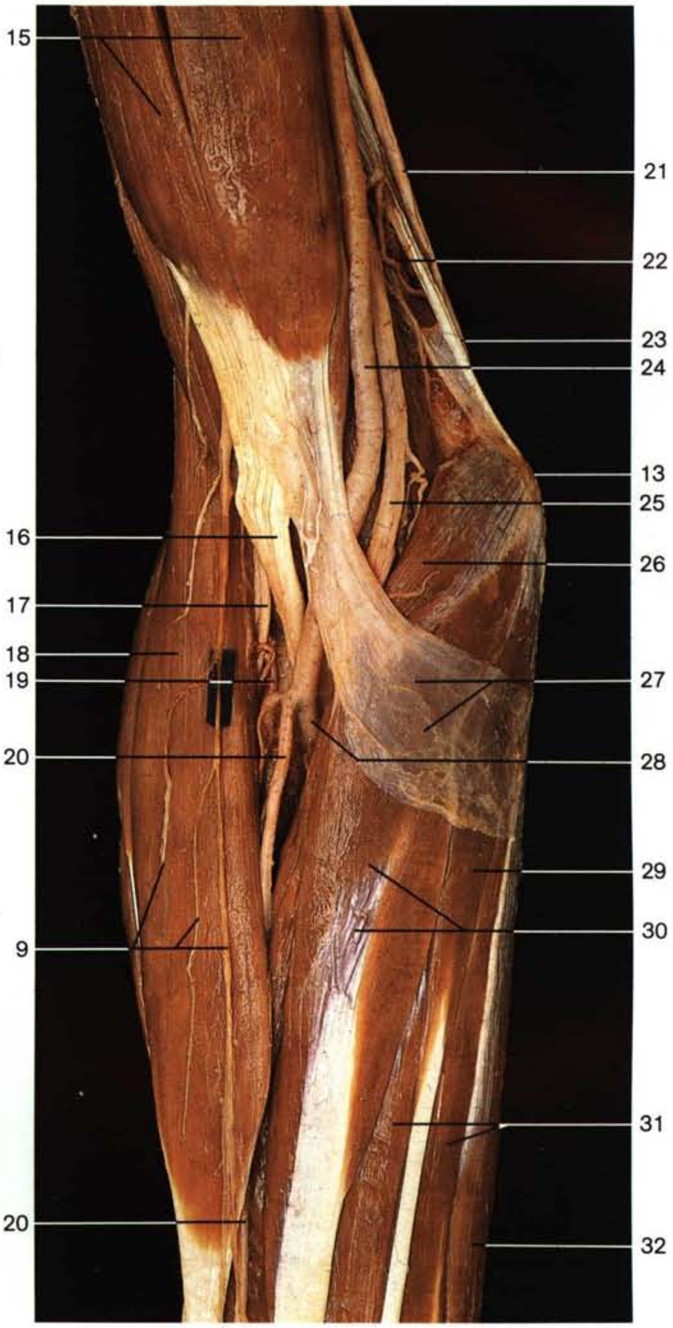

Cubital region, superficial layer (anterior aspect). The fasciae of the muscles have been removed.

1 Biceps brachii muscle with fascia
2 **Cephalic vein**
3 **Median cubital vein**
4 Lateral cutaneous nerve of forearm
5 Tendon and aponeurosis of biceps brachii muscle (covered by the antebrachial fascia)
6 Brachioradialis muscle with fascia
7 Accessory cephalic vein
8 Median vein of forearm
9 Branches of lateral cutaneous nerve of forearm
10 Terminal branches of medial cutaneous nerve of arm
11 **Medial cutaneous nerve of forearm**
12 **Basilic vein**
13 Medial epicondyle of humerus
14 Terminal branches of medial cutaneous nerve of forearm
15 Biceps brachii muscle

16 Tendon of biceps brachii muscle
17 Radial nerve
18 Brachioradiaiis muscle
19 Radial recurrent artery
20 **Radial artery**
21 **Ulnar nerve**
22 Superior ulnar collateral artery
23 Medial intermuscular septum
24 **Brachial artery**
25 **Median nerve**
26 Pronator teres muscle
27 Bicipital aponeurosis
28 Ulnar artery
29 Palmaris longus muscle
30 Flexor carpi radialis muscle
31 Flexor digitorum superficialis muscle
32 Flexor carpi ulnaris muscle

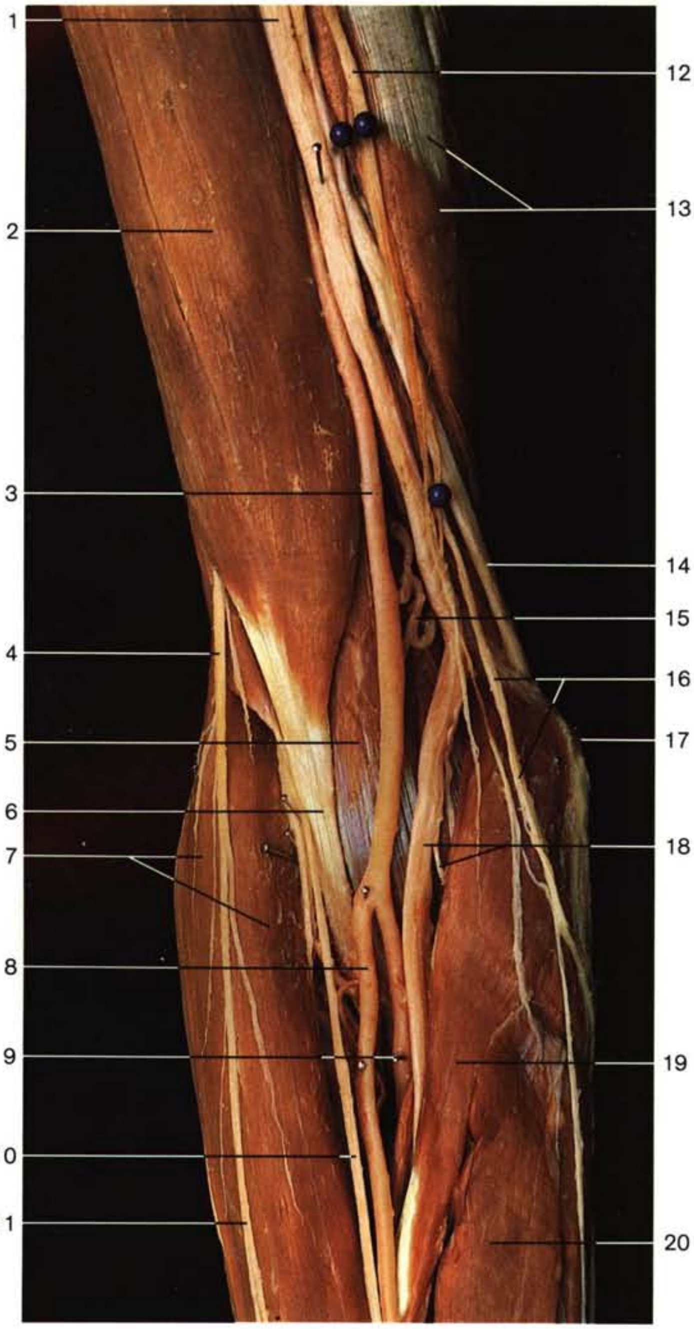

Cubital region, middle layer (anterior aspect). The bicipital aponeurosis has been removed.

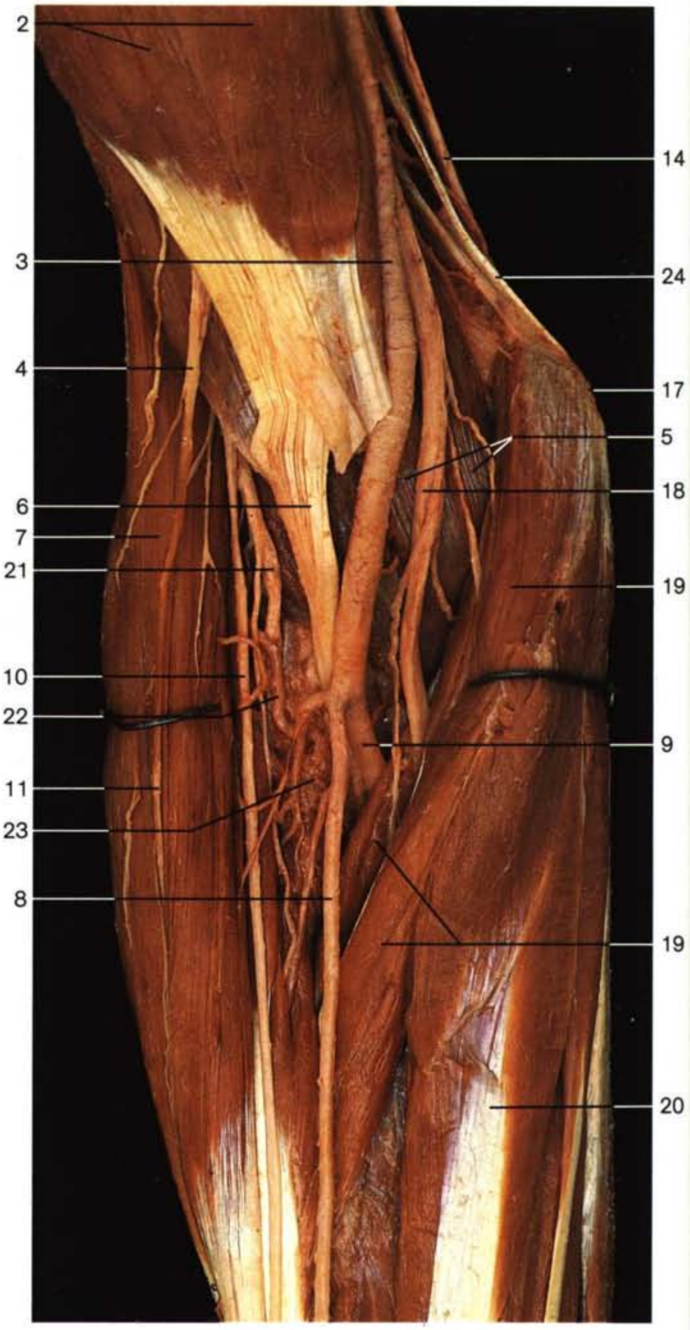

Cubital region, middle layer (anterior aspect). The pronator teres and brachioradialis muscles have been slightly reflected.

1 Median nerve
2 Biceps brachii muscle
3 Brachial artery
4 Lateral cutaneous nerve of forearm (terminal branch of **musculocutaneous nerve**)
5 Brachialis muscle
6 Tendon of biceps brachii muscle
7 Brachioradialis muscle
8 **Radial artery**
9 **Ulnar artery**
10 Superficial branch of radial nerve
11 Lateral cutaneous nerve of forearm
12 Medial cutaneous nerve of forearm
13 Triceps brachii muscle
14 **Ulnar nerve**
15 Inferior ulnar collateral artery
16 Anterior branch of medial cutaneous nerve of forearm
17 Medial epicondyle of humerus
18 **Median nerve** with branches to pronator teres muscle
19 Pronator teres muscle
20 Flexor carpi radialis muscle
21 Deep branch of radial nerve
22 Radial recurrent artery
23 Supinator muscle
24 Medial intermuscular septum of arm

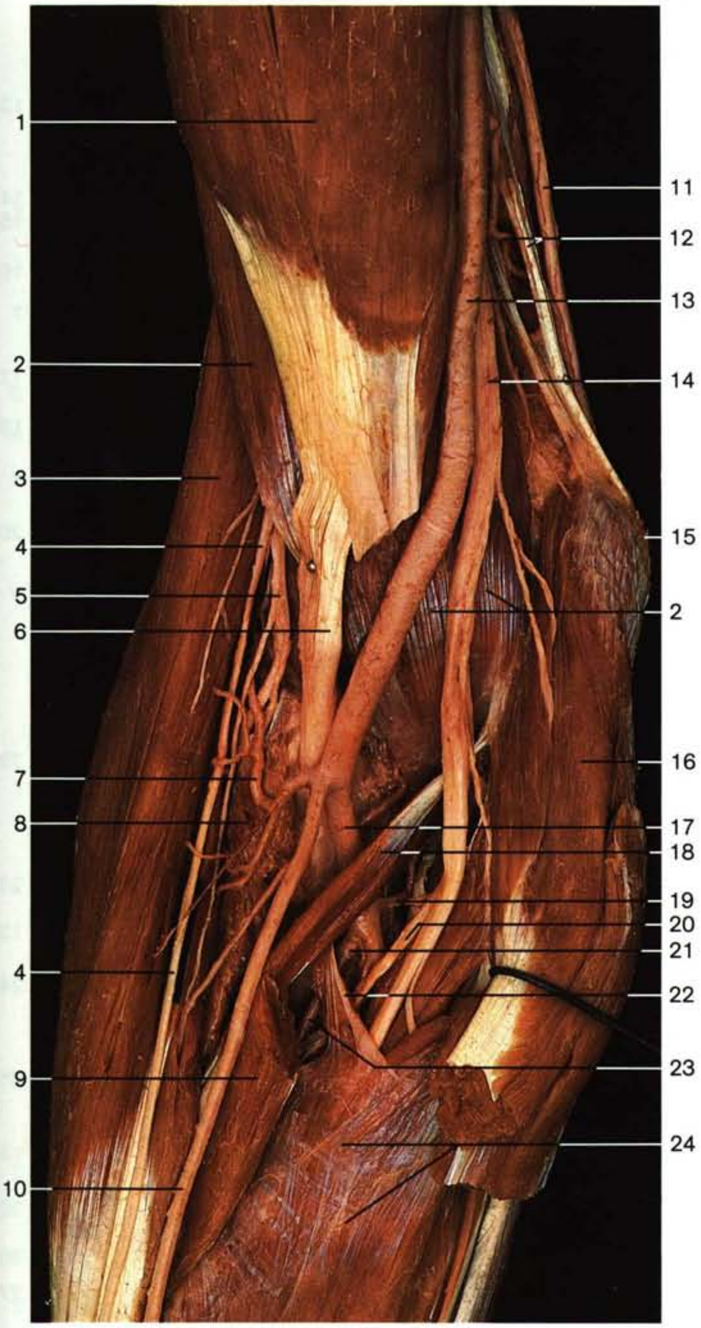

Cubital region, deep layer (anterior aspect). The pronator teres and flexor carpi ulnaris muscles have been cut and reflected.

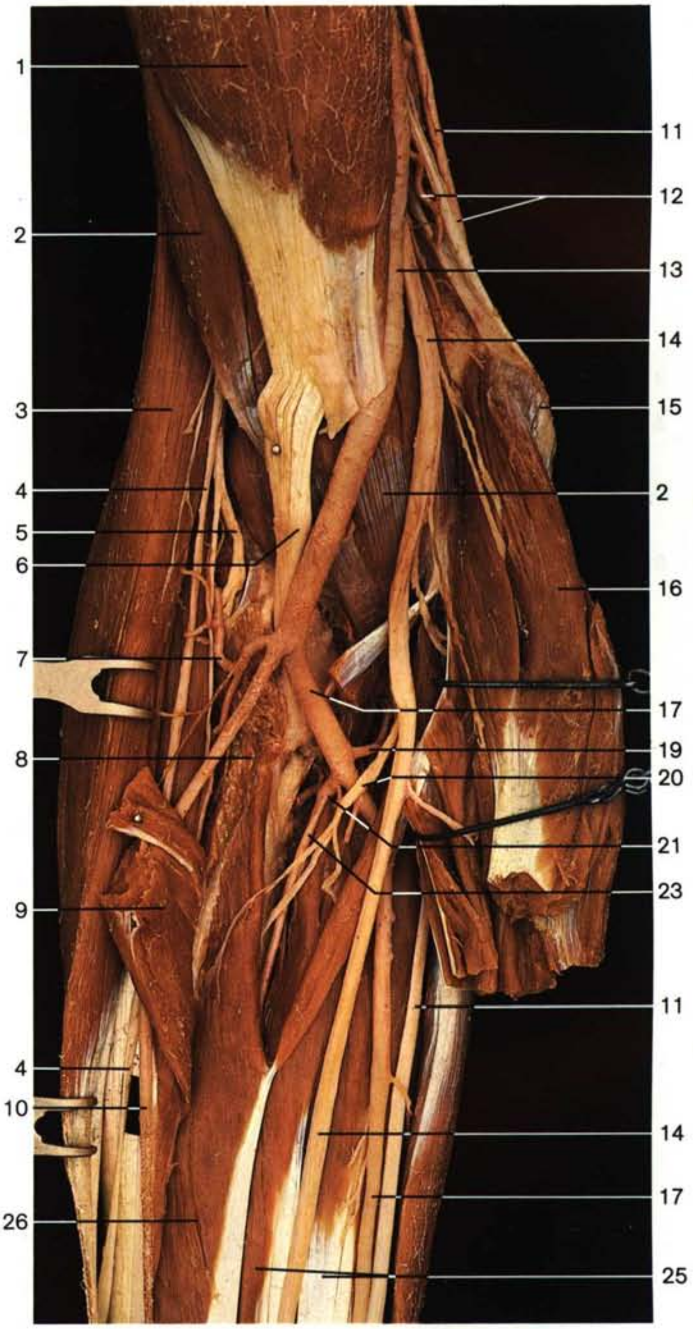

Cubital region, deepest layer (anterior aspect). The flexor digitorum superficialis and the ulnar head of the pronator teres have been cut and reflected.

1 Biceps brachii muscle
2 Brachialis muscle
3 Brachioradialis muscle
4 Superficial branch of radial nerve
5 Deep branch of radial nerve
6 Tendon of biceps brachii muscle
7 **Radial recurrent artery**
8 Supinator muscle
9 Insertion of pronator teres muscle
10 **Radial artery**
11 **Ulnar nerve**
12 Medial intermuscular septum of arm and superior ulnar collateral artery
13 **Brachial artery**
14 **Median nerve**
15 Medial epicondyle of humerus
16 Humeral head of pronator teres muscle
17 **Ulnar artery**
18 Ulnar head of pronator teres muscle
19 **Ulnar recurrent artery**
20 Anterior interosseous nerve
21 Common interosseous artery
22 Tendinous arch of flexor digitorum superficialis muscle
23 Anterior interosseous artery
24 Flexor digitorum superficialis muscle
25 Flexor digitorum profundus muscle
26 Flexor pollicis longus muscle

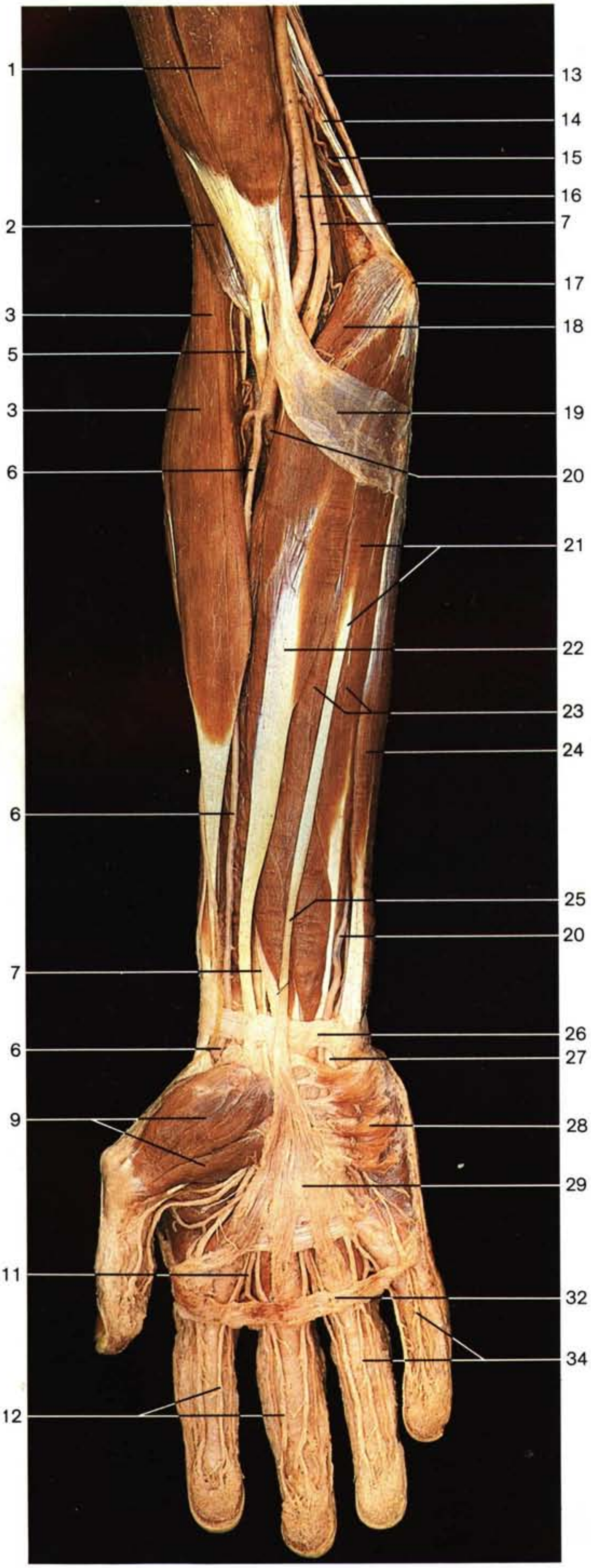
1
2
3
5
3
6
6
7
6
9
11
12
13
14
15
16
7
17
18
19
20
21
22
23
24
25
20
26
27
28
29
32
34

Vessels and nerves of right forearm and hand, superficial layer (palmar aspect).

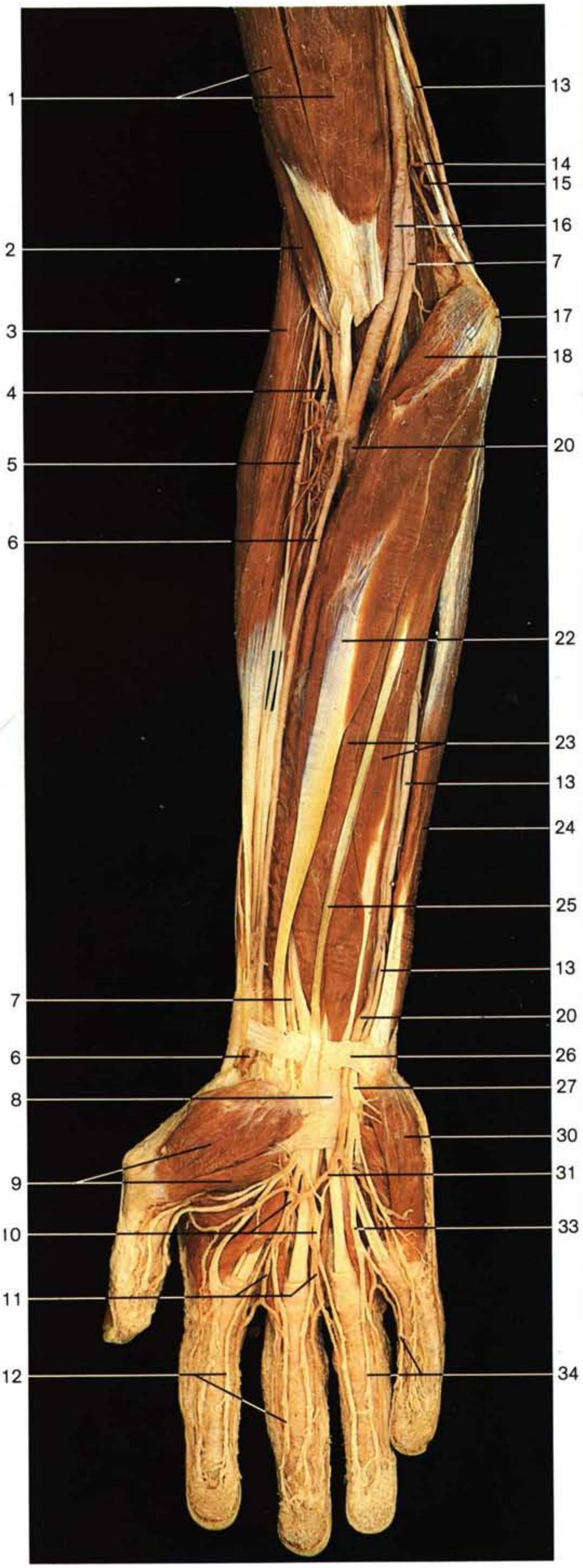
1
2
3
4
5
6
7
6
8
9
10
11
12
13
14
15
16
7
17
18
20
22
23
13
24
25
13
20
26
27
30
31
33
34

Vessels and nerves of right forearm and hand, superficial layer (palmar aspect). The palmar aponeurosis of the hand and the bicipital aponeurosis have been removed.

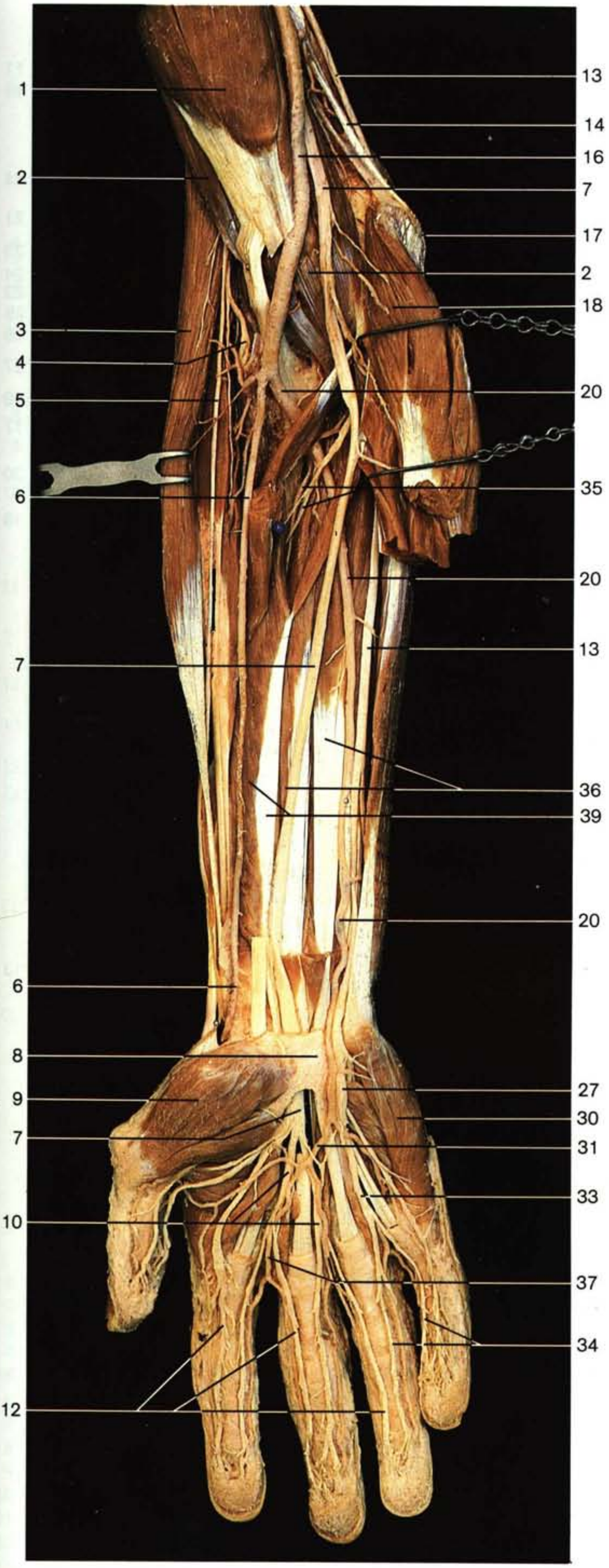

Vessels and nerves of forearm and hand, deep layer (anterior aspect). The superficial layer of the flexor muscles has been removed.

1 Biceps brachii muscle
2 Brachialis muscle
3 Brachioradialis muscle
4 **Deep branch of radial nerve**
5 **Superficial branch of radial nerve**
6 **Radial artery**
7 **Median nerve**
8 Flexor retinaculum
9 Thenar muscles
10 Common palmar digital branches of median nerve
11 Common palmar digital arteries
12 Proper palmar digital nerves (median nerve)
13 **Ulnar nerve**
14 Medial intermuscular septum of arm
15 Superior ulnar collateral artery
16 Brachial artery
17 Medial epicondyle of humerus
18 Pronator teres muscle
19 Bicipital aponeurosis
20 **Ulnar artery**
21 Palmaris longus muscle
22 Flexor carpi radialis muscle
23 Flexor digitorum superficialis muscle
24 Flexor carpi ulnaris muscle
25 Tendon of palmaris longus muscle
26 Remnant of antebrachial fascia
27 Superficial branch of ulnar nerve
28 Palmaris brevis muscle
29 Palmar aponeurosis
30 Hypothenar muscles
31 **Superficial palmar arch**
32 Superficial transverse metacarpal ligament
33 Common palmar digital branch of ulnar nerve
34 Proper palmar digital branches of ulnar nerve
35 Anterior interosseous artery and nerve
36 Flexor digitorum profundus muscle
37 Common palmar digital arteries
38 Palmar branch of median nerve
39 Flexor pollicis longus muscle
40 Palmar branch of ulnar nerve

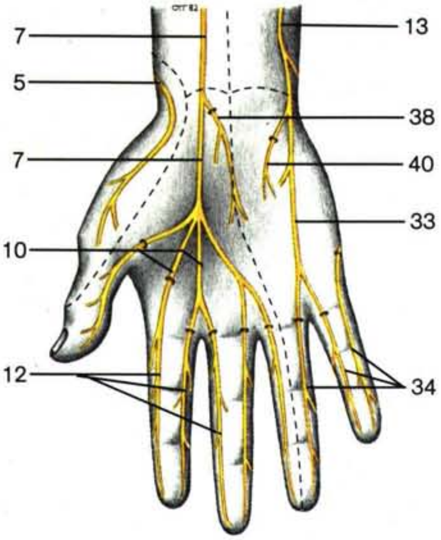

Cutaneous innervation of hand (palmar aspect). (Schematic drawing.)

Cutaneous innervation of palmar surface:
3½ digits by median nerve,
1½ digits by ulnar nerve.

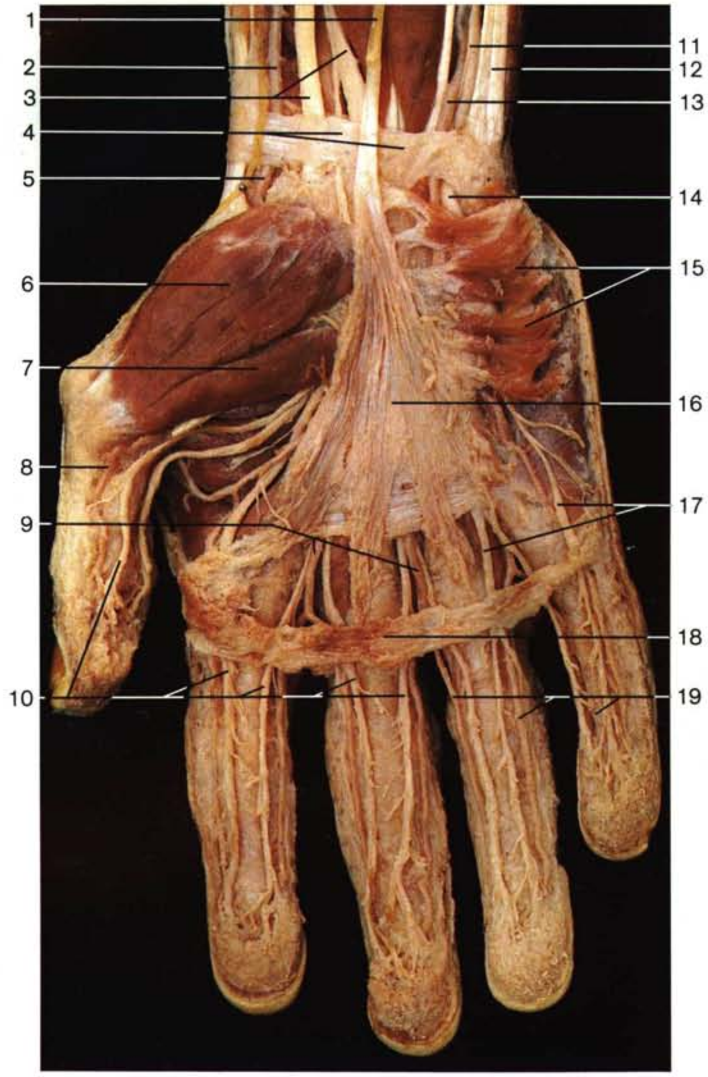

Right hand, superficial layer, dissection of vessels and nerves (palmar aspect).

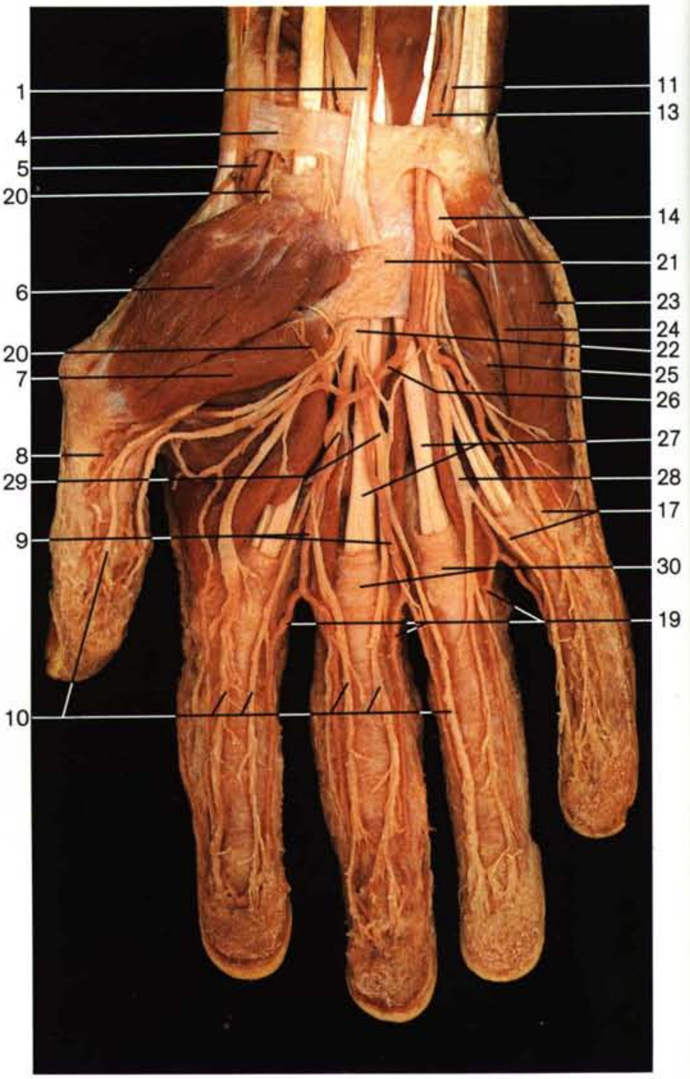

Right hand, superficial layer, dissection of vessels and nerves (palmar aspect). The palmar aponeurosis has been removed to display the superficial palmar arch.

1 Tendon of palmaris longus muscle
2 **Radial artery**
3 Tendon of flexor carpi radialis muscle and median nerve
4 Distal part of antebrachial fascia
5 Radial artery passing into the anatomical snuffbox
6 Abductor pollicis brevis muscle
7 Superficial head of flexor pollicis brevis muscle
8 Palmar digital artery of thumb
9 Common palmar digital arteries
10 Proper palmar digital nerves (median nerve)
11 **Ulmar nerve**
12 Tendon of flexor carpi ulnaris muscle
13 **Ulnar artery**
14 Superficial branch of ulnar nerve
15 Palmaris brevis muscle
16 Palmar aponeurosis
17 Palmar digital nerves (ulnar nerve)
18 Superficial transverse metacarpal ligament
19 Proper palmar digital arteries
20 Superficial palmar branch of radial artery (contributing to the superficial palmar arch)
21 Flexor retinaculum
22 **Median nerve**
23 Abductor digiti minimi muscle
24 Flexor digiti minimi brevis muscle
25 Opponens digiti minimi muscle
26 **Superficial palmar arch**
27 Tendons of flexor digitorum superficialis muscle
28 Common palmar digital branch of ulnar nerve
29 Common palmar digital branch of median nerve
30 Fibrous sheath of flexor tendons

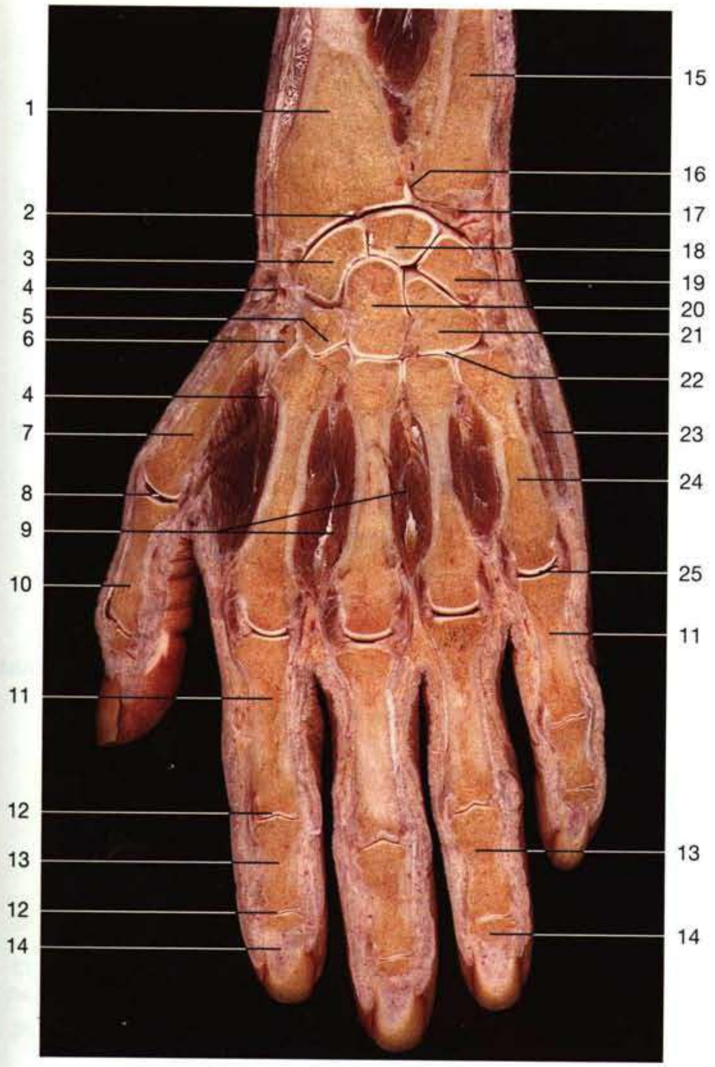
1
15
16
2
17
3
18
4
19
5
20
6
21
4
22
7
23
8
24
9
10
25
11
11
12
13
13
12
14
14

Coronal section through the right hand (palmar aspect).

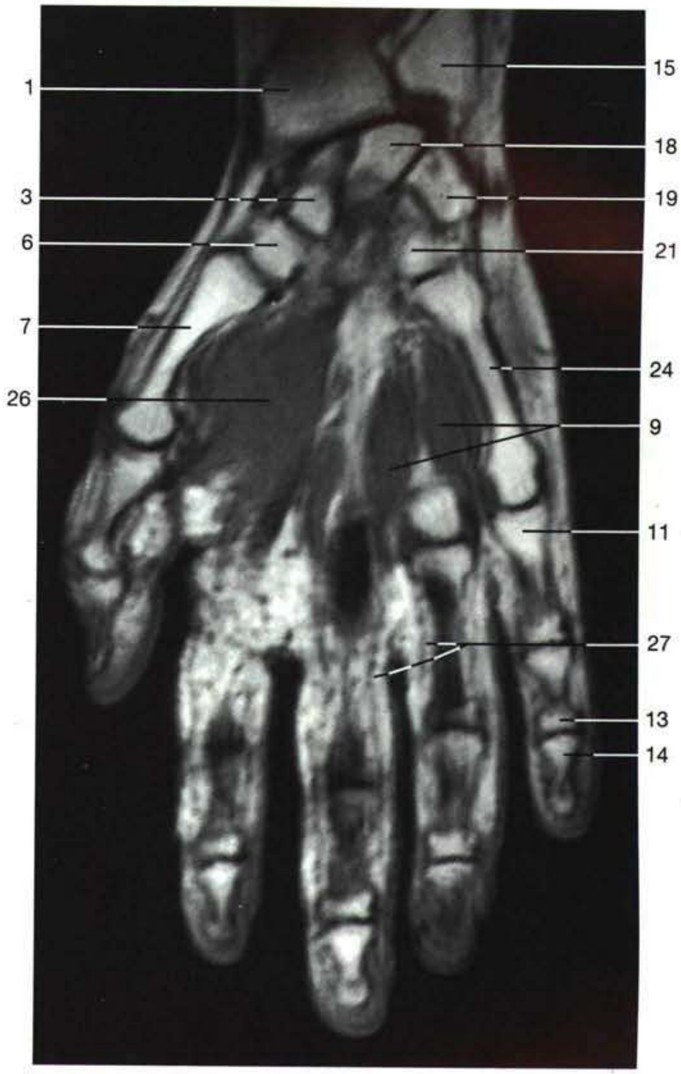
1
15
18
3
19
6
21
7
24
26
9
11
27
13
14

Coronal section through the right hand (palmar aspect) (MR-Scan, courtesy of Dr. A. Heuck, Munich).

1 Radius
2 **Radiocarpal articulation**
3 **Scaphoid** (navicular) **bone**
4 **Radial artery**
5 Trapezoid bone
6 **Trapezium bone**
7 First metacarpal bone
8 Metacarpophalangeal articulation of thumb
9 Interosseous muscles
10 Proximal phalanx of thumb
11 Proximal phalanx of fingers
12 Interphalangeal articulations
13 Middle phalanx
14 Distal phalanx
15 **Ulna**
16 Distal radioulnar articulation
17 Articular disk
18 **Lunate bone**
19 **Triangular bone**
20 **Capitate bone**
21 **Hamate bone**
22 Carpometacarpal articulations
23 Abductor digiti minimi muscle
24 Fifth metacarpal bone
25 Metacarpophalangeal articulation
26 Adductor pollicis muscle
27 Proper palmar digital arteries

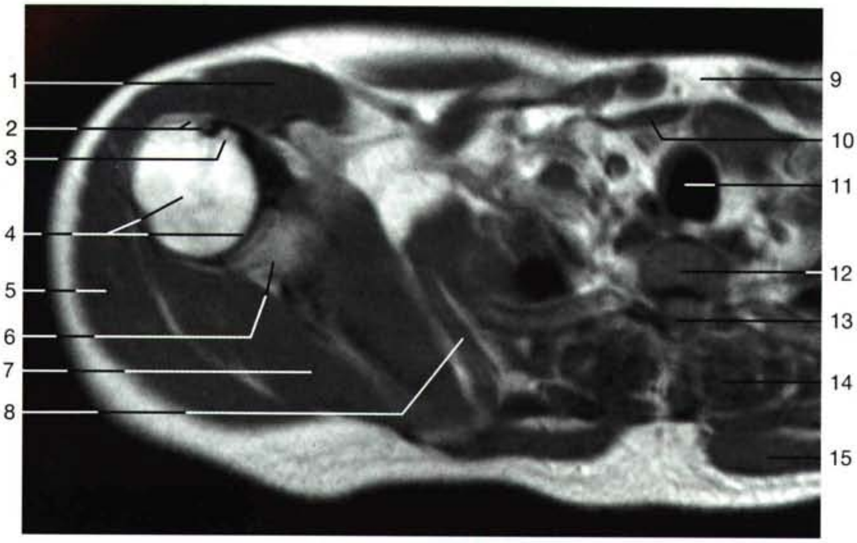
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Horizontal section through shoulder joint (section 1; MR-Image; inferior aspect).

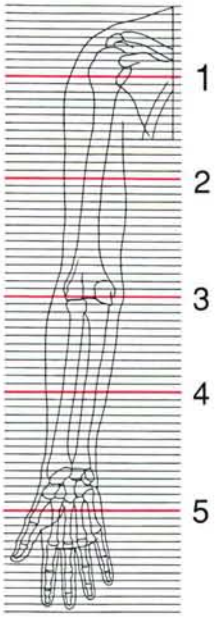

Upper extremity, location of sections 1–5 (MR-Scans, courtesy of Dr. A. Heuck, Munich).

1 Pectoralis major muscle
2 Greater tubercle and tendon of biceps muscle
3 Lesser tubercle
4 Head of humerus and articular cavity of shoulder joint
5 Deltoid muscle
6 Scapula
7 Infraspinatus muscle
8 Serratus anterior muscle
9 Sternum
10 Infrahyoid muscles
11 Trachea
12 Body of thoracic vertebra
13 Vertebral canal and spinal cord
14 Deep muscles of the back
15 Trapezius muscle
16 Brachialis muscle
17 Radial nerve and profunda brachii vessels

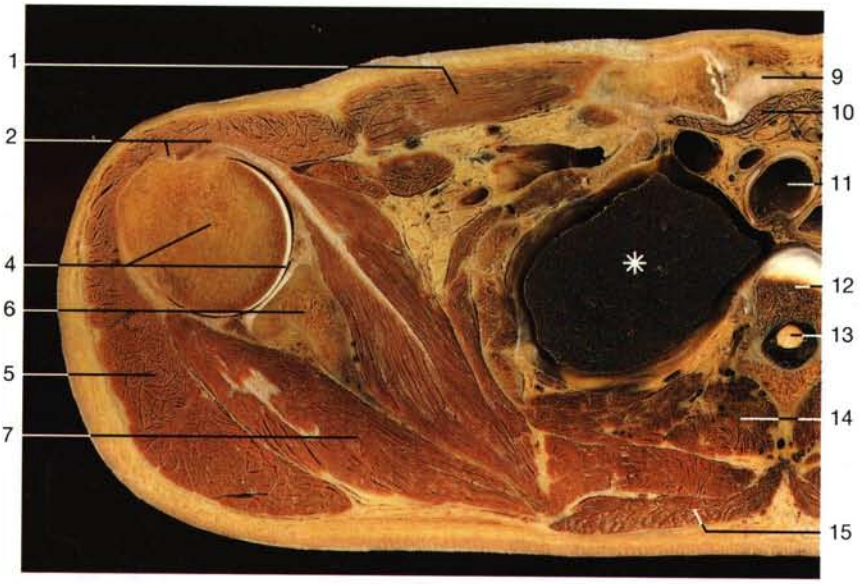
1
2
4
6
5
7
9
10
11
*
12
13
14
15

Horizontal section through the right shoulder at the level of T_1 (section 1; inferior aspect). * = Upper lobe of lung.

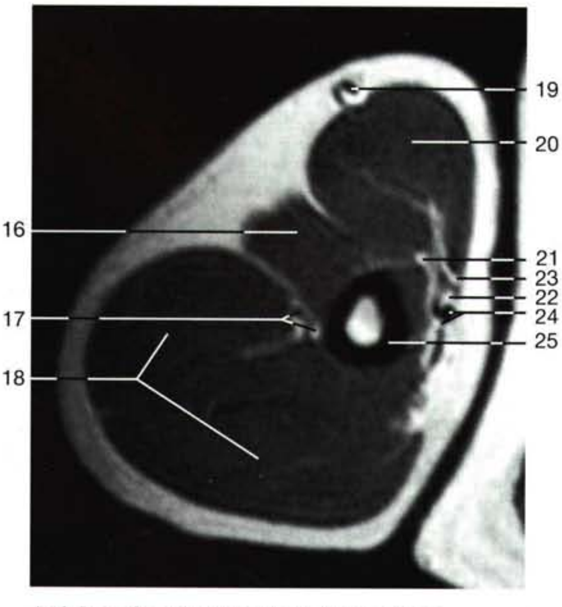
19
20
16
21
23
22
24
17
25
18

Axial section through the middle of the right arm (section 2; MR-Image, inferior aspect).

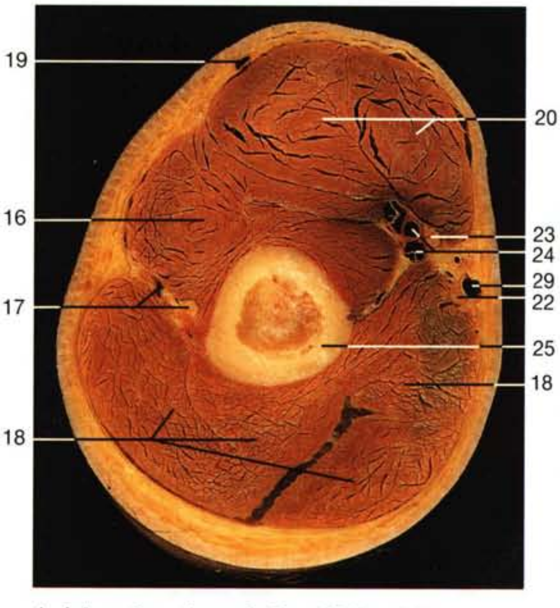
19
20
16
23
24
29
22
17
25
18
18

Axial section through the middle of the right arm (section 2; inferior aspect).

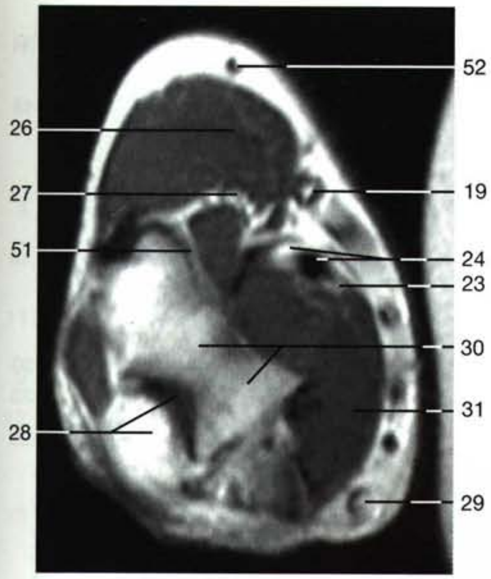

Axial section through the elbow joint (section 3; MR-Image; inferior aspect).

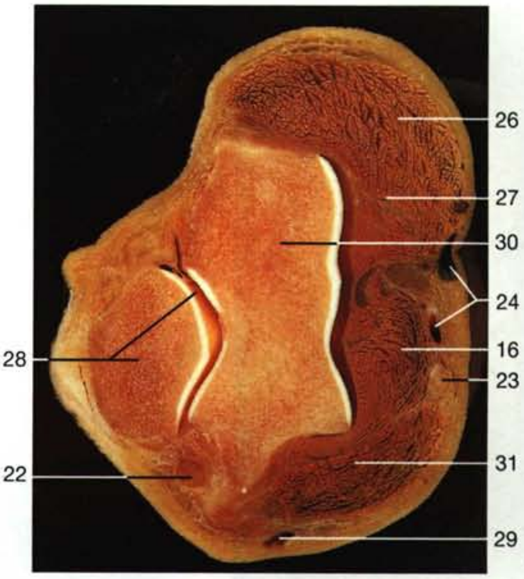

Axial section through the right elbow joint (section 3; inferior aspect).

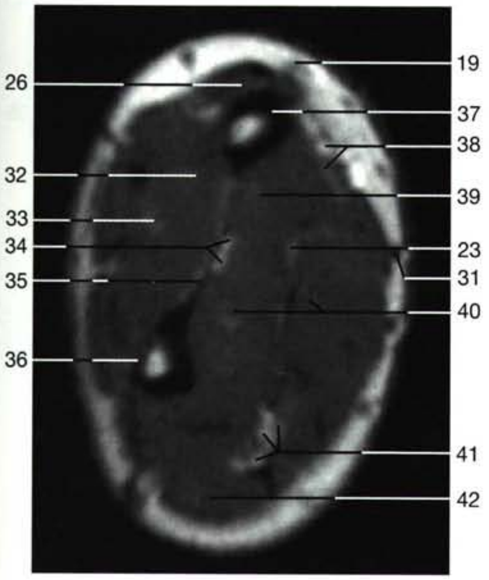

Axial section through the middle of the forearm (section 4; MR-Image; inferior aspect).

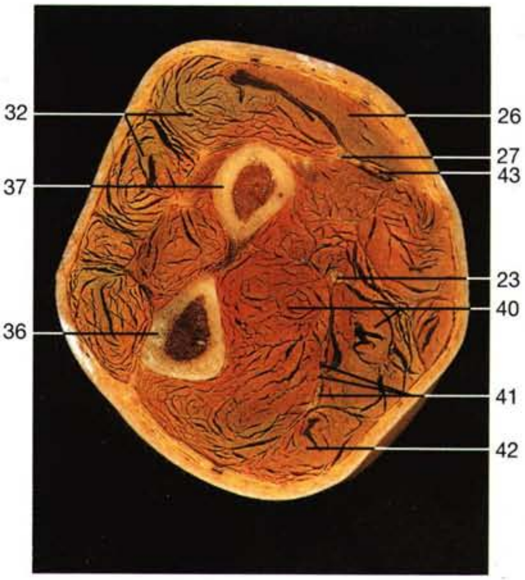

Axial section through the right forearm (section 4; inferior aspect).

18	Triceps brachii muscle
19	Cephalic vein
20	Biceps brachii muscle
21	Musculocutaneous nerve
22	**Ulnar nerve**
23	**Medianus nerve**
24	Brachial artery and vein
25	Shaft of humerus
26	Brachioradialis muscle
27	**Radial nerve**
28	Olecranon and articular cavity of elbow joint
29	Basilic vein
30	Humerus
31	Pronator teres muscle
32	Extensor muscles of forearm
33	Ramus profundus of radialis nerve
34	Anterior interosseus vessels and nerve
35	Interosseous membrane
36	Ulna
37	Radius
38	Radial artery and superficial branch of radial nerve
39	Flexor pollicis longus muscle
40	Flexor digitorum superficialis and profundus muscles
41	Ulnar nerve, ulnar artery and vein
42	Flexor carpi ulnaris muscle
43	Radial artery
44	Metacarpal bones III and IV
45	**Carpal canal** with tendons of flexor digitorum muscles
46	Hypothenar muscle
47	Median nerve
48	Interosseus muscles
49	First metacarpal bone
50	Thenar muscles
51	Articular cavity of humeroradial joint
52	Median cubital vein

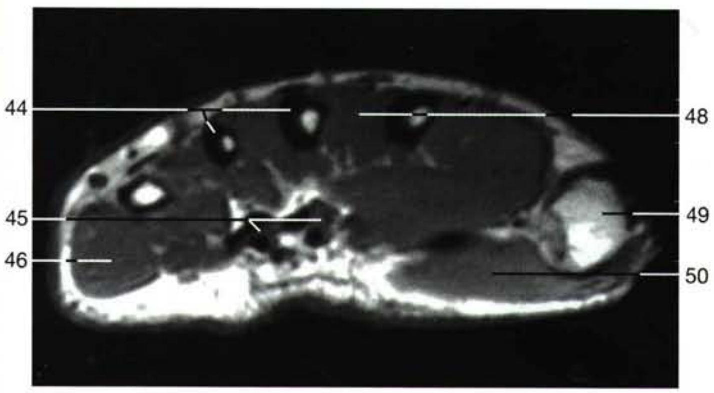

Axial section through the right hand (metacarpus; MR-Image; section 5; inferior aspect).

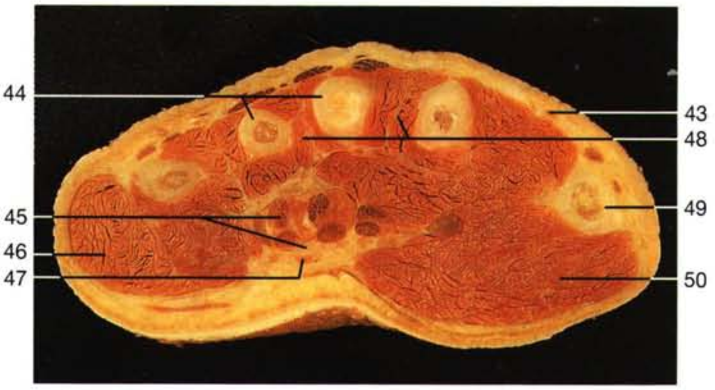

Axial section through the right hand at the level of the metacarpus (section 5; inferior aspect).

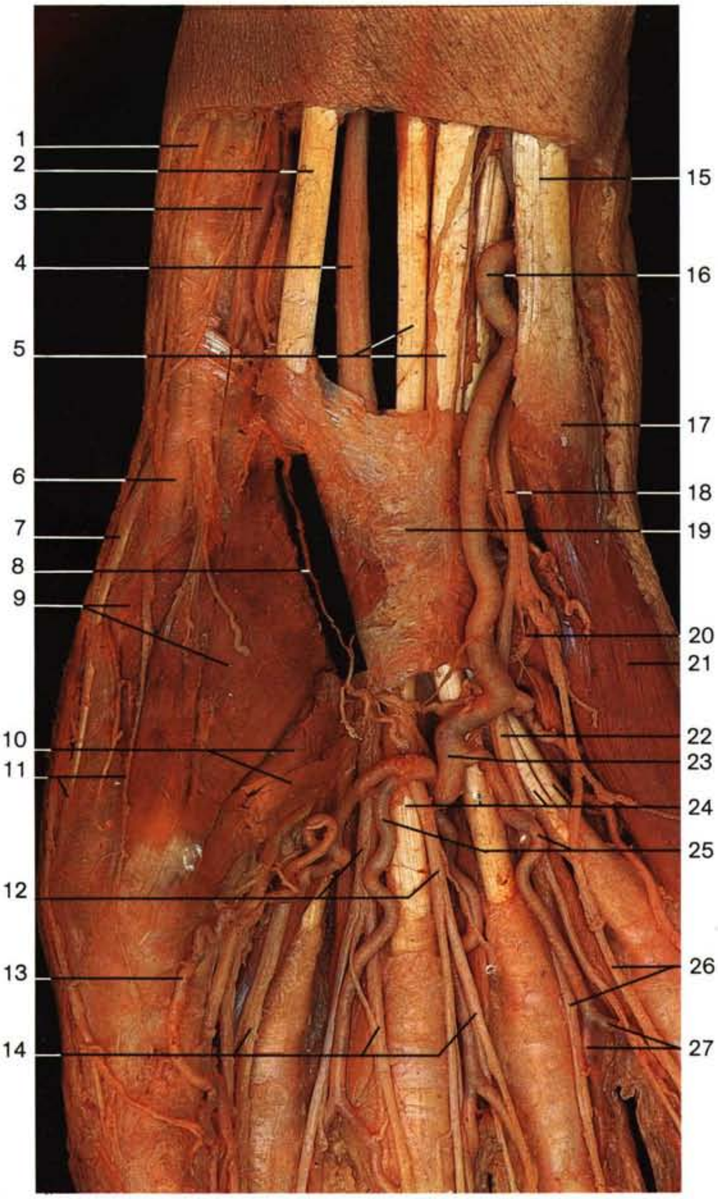
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

◁
Right hand, superficial layer (palmar aspect). Dissection of the superficial palmar arch.

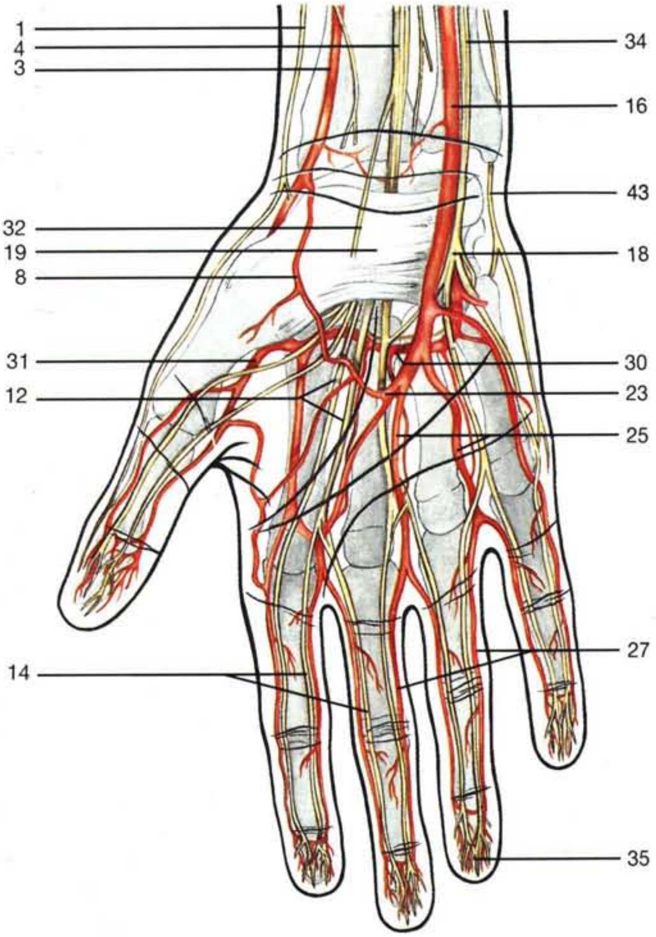
1
4
3
32
19
8
31
12
14
34
16
43
18
30
23
25
27
35

Arteries and nerves of the right hand (palmar aspect). (Schematic drawing.)

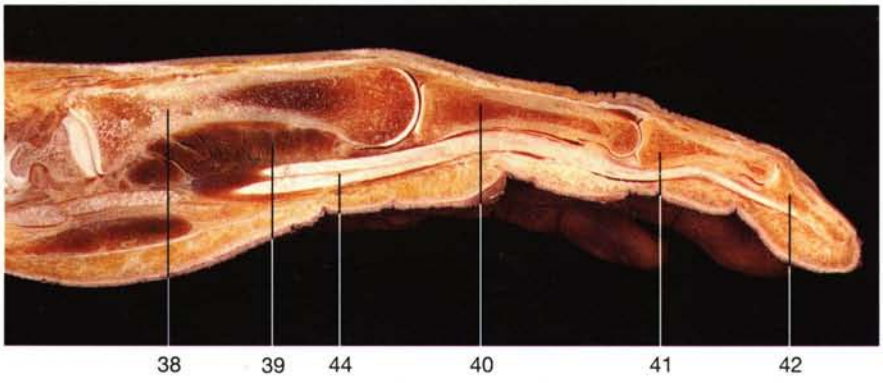
38
39
44
40
41
42

Longitudinal section through the hand at the level of the third finger

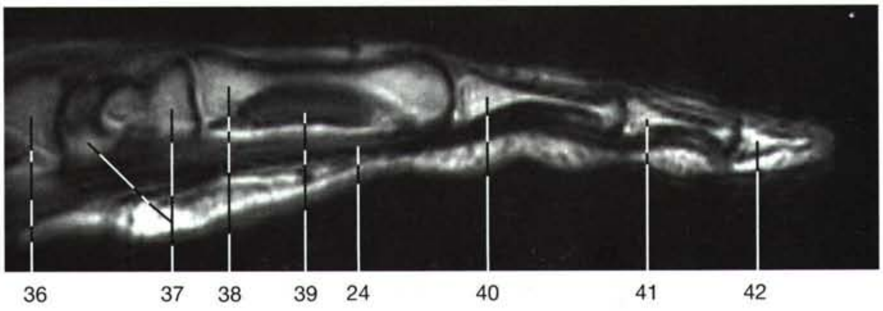
36
37
38
39
24
40
41
42

Longitudinal section through the hand at the level of the third finger (MR-Scan, courtesy of Dr. A. Heuck, Munich).

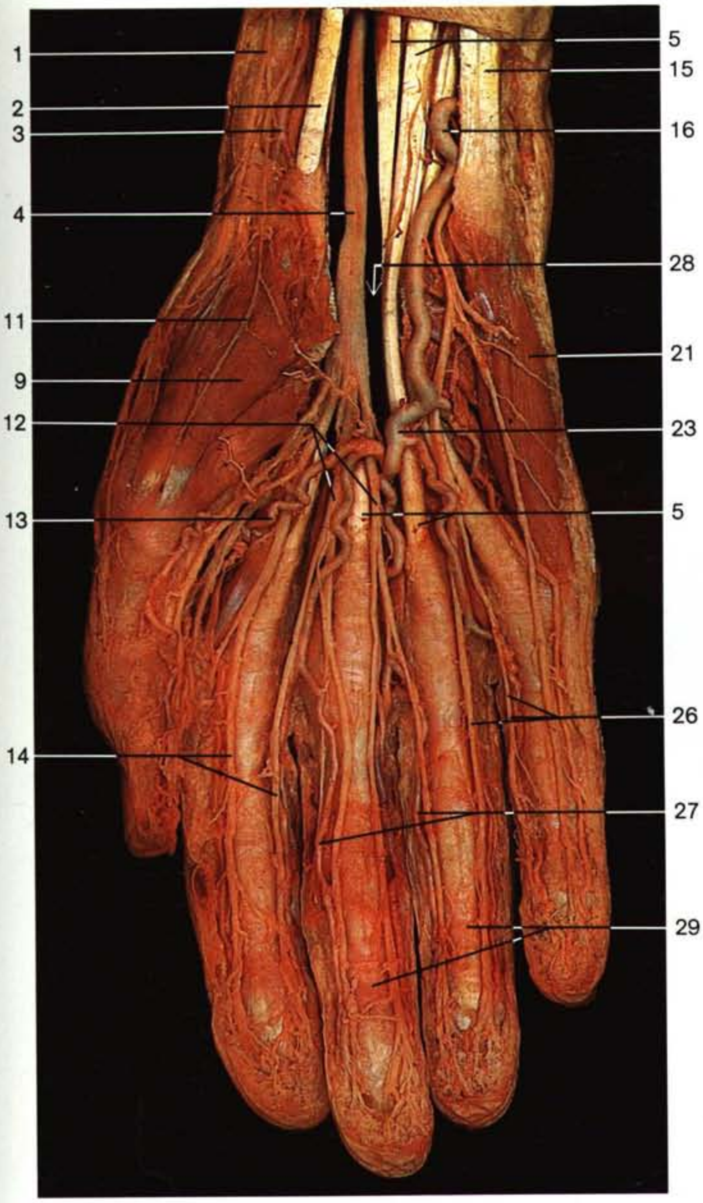

Right hand, middle layer (palmar aspect). The flexor retinaculum has been removed.

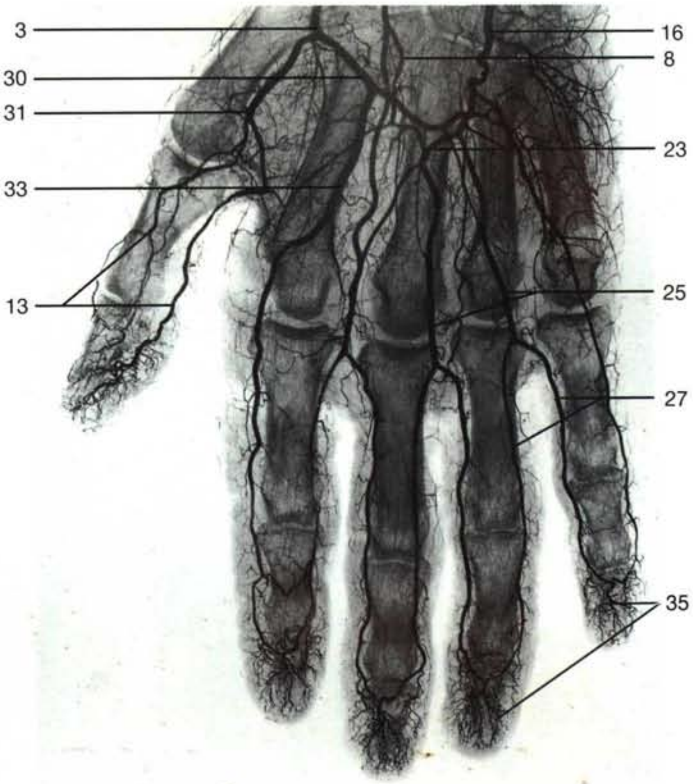

Arteriogram of the right hand (palmar aspect).

1 Superficial branch of radial nerve
2 Tendon of flexor carpi radialis muscle
3 **Radial artery**
4 **Median nerve**
5 Tendon of flexor digitorum superficialis muscle
6 Tendon of abductor pollicis longus muscle
7 Tendon of extensor pollicis brevis muscle
8 Superficial palmar branch of radial artery
9 Abductor pollicis brevis muscle
10 Superficial head of flexor pollicis brevis muscle
11 Terminal branches of superficial branch of radial nerve
12 Common palmar digital nerves (median nerve)
13 Proper palmar digital arteries of thumb
14 Proper palmar digital nerves (median nerve)
15 Tendon of flexor carpi ulnaris muscle
16 **Ulnar artery**
17 Position of pisiform bone
18 Superficial branch of ulnar nerve
19 Flexor retinaculum
20 Deep branch of ulnar nerve
21 Abductor digiti minimi muscle
22 Common palmar digital nerves (ulnar nerve)
23 **Superficial palmar arch**
24 Tendons of flexor digitorum muscles
25 Common palmar digital arteries
26 **Palmar digital nerves (ulnar nerve)**
27 **Proper palmar digital arteries**
28 **Carpal tunnel**
29 Fibrous sheaths for the tendons of flexor digitorum muscles
30 **Deep palmar arch**
31 Princeps pollicis artery
32 Palmar branch of median nerve
33 Common digital palmar artery
34 Ulnar nerve
35 Capillary network of finger
36 Radius
37 Carpal bones
38 Metacarpal bone
39 Interosseus muscles
40 Proximal phalanx
41 Middle phalanx
42 Distal phalanx
43 Dorsal branch of ulnar nerve
44 Tendons of flexor digitorum profundus (upper) and superficialis (lower) muscles

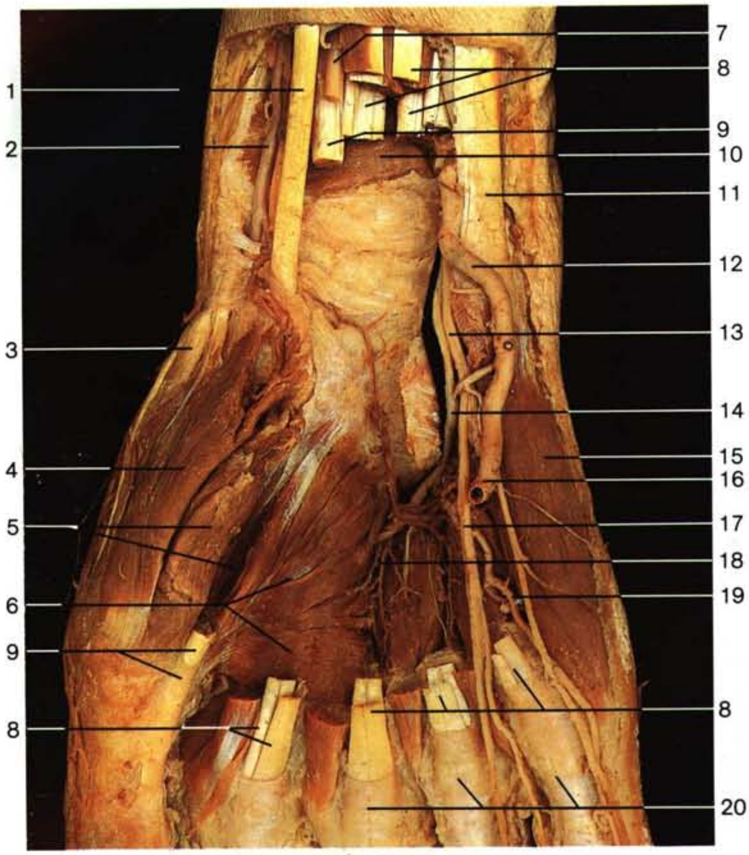

Right hand, deep layer (palmar aspect). The carpal tunnel has been opened, the tendons of the flexor muscles have been removed and the superficial palmar arch has been cut.

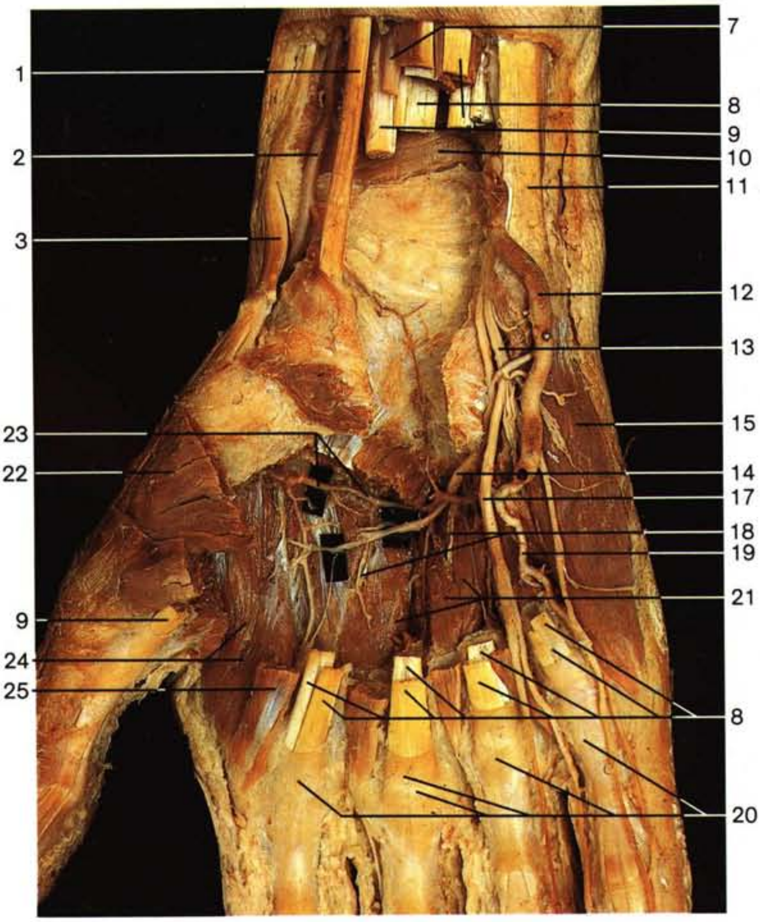

Right hand, deep layer (palmar aspect). Dissection of the deep palmar arch.

1	Tendon of flexor carpi radialis muscle
2	Radial artery
3	Tendon of abductor pollicis longus muscle
4	Abductor pollicis brevis muscle
5	Superficial and deep heads of flexor pollicis brevis muscle
6	Oblique and transverse heads of adductor pollicis muscle
7	Median nerve
8	Tendons of flexor digitorum superficialis and profundus muscles
9	Tendon of flexor pollicis longus muscle
10	Pronator quadratus muscle
11	Tendon of flexor carpi ulnaris muscle
12	Ulnar artery
13	Superficial branch of ulnar nerve
14	Deep branch of ulnar nerve
15	Abductor digiti minimi muscle
16	**Superficial palmar arch** (cut end)
17	Common palmar digital nerves (ulnar nerve)
18	Palmar metacarpal arteries of deep palmar arch
19	Palmar digital artery of the 5th finger
20	Fibrous sheaths of tendons of flexor muscles
21	Palmar interosseous muscles
22	Opponens pollicis muscle (cut)
23	**Deep palmar arch**
24	1st dorsal interosseous muscle
25	1st lumbrical muscle

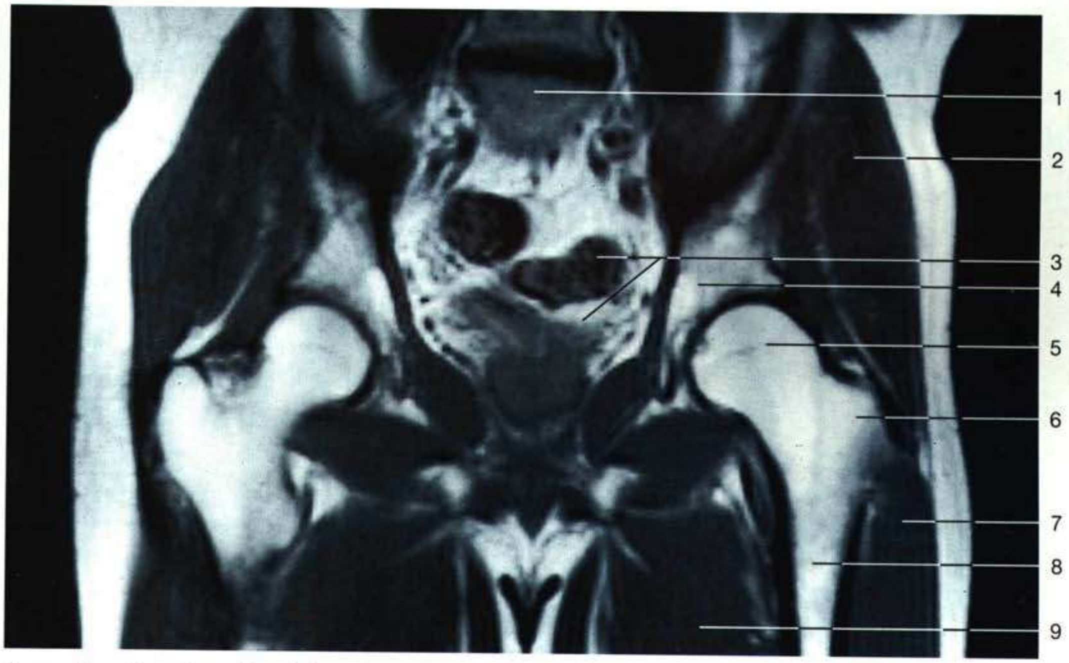

Coronal section through pelvis and thighs (MR-Scan of the male, courtesy of Dr. A. Heuck, Munich).

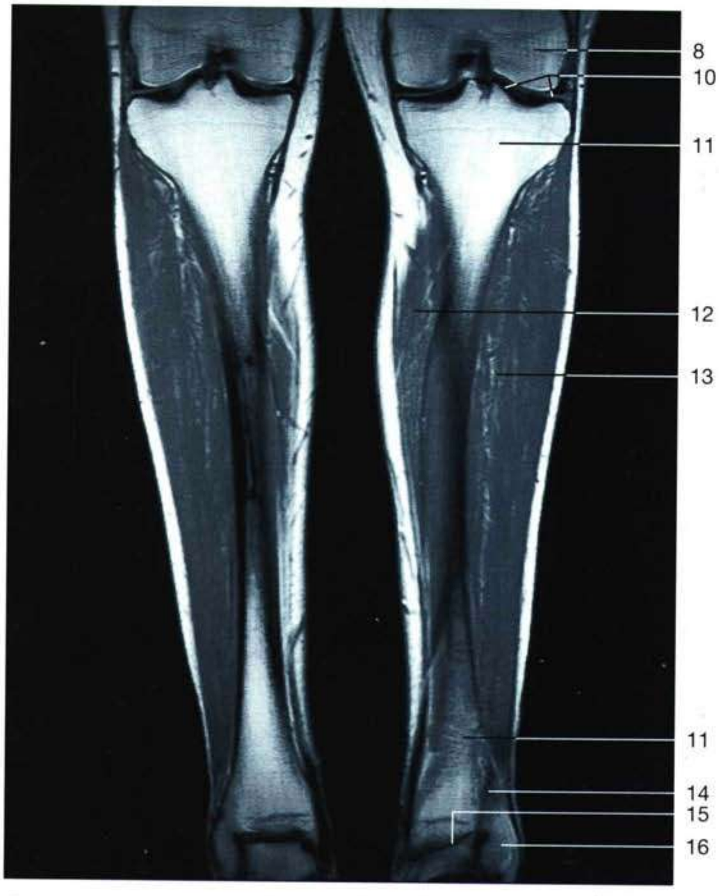

Coronal section through the legs (MR-Scan, courtesy of Dr. A. Heuck, Munich).

1 Sacral promontory
2 Gluteus medius muscle
3 Small intestine and urinary bladder
4 Acetabulum
5 **Head of femur**
6 Greater trochanter of femur
7 Vastus lateralis muscle
8 **Femur**
9 Adductor muscles
10 **Knee joint** with menisci
11 **Tibia**
12 Soleus muscle
13 Tibialis anterior muscle
14 Distal tibiofibular articulation
15 **Talocrural articulation**
16 Fibula (lateral malleolus)

The lower limb (extremity) is specialized for support of the upright posture, locomotion and maintaining balance. In contrast to the upper limb, the lower limb is more restricted in its movements, and the joints are more tight and fixed by strong ligaments. The hip joint is a ball and socket type of synovial joint between head of femur and acetabulum. The knee joint is a hinge type of synovial joint that permits only limited rotation. The talocrural joint is a hinge joint between talus, fibula, and tibia only allowing movements of flexion and extension.

The long axis of the foot is at right angle to that of the leg, thus, forming an effective arch for the upright stance of man.

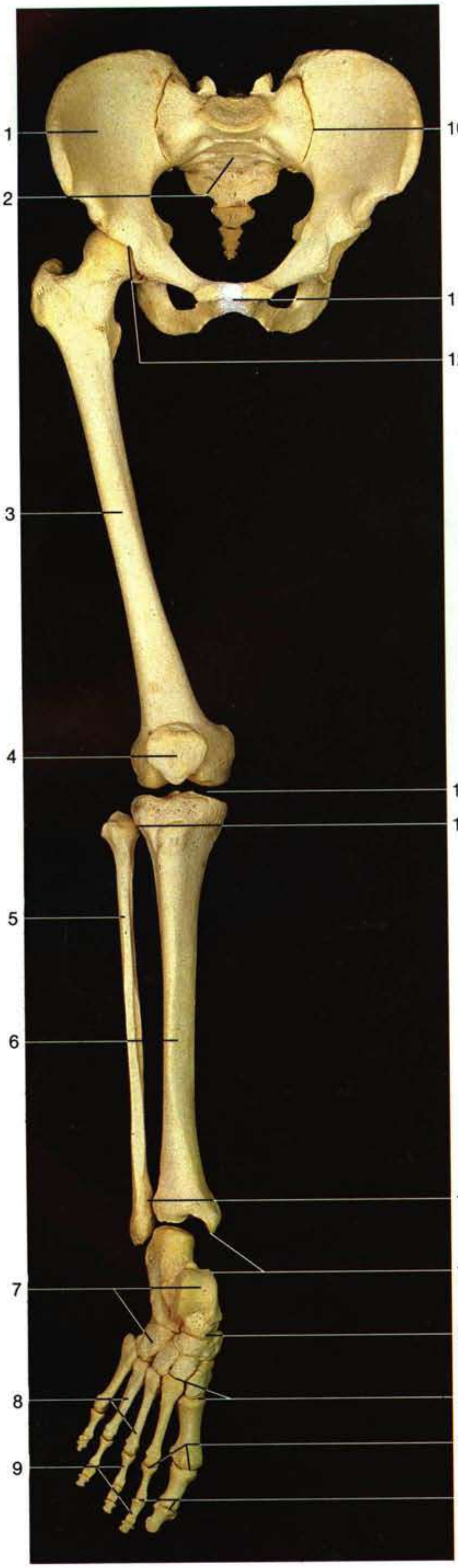

Skeleton of pelvic girdle and lower limb (anterior aspect). The talocrural joint has been dislocated.

A	Pelvic girdle
B	Thigh
C	Leg
D	Foot
1	Right hip bone
2	Sacrum
3	Femur
4	Patella
5	Fibula
6	Tibia
7	Tarsal bones
8	Metatarsal bones
9	Phalanges
10	Sacroiliac joint
11	Pubic symphysis
12	Hip joint
13	Knee joint
14	Proximal tibiofibular joint
15	Distal tibiofibular joint
16	Talocrural joint
17	Talocalcaneonavicular joint
18	Tarsometatarsal joints
19	Metatarsophalangeal joints
20	Interphalangeal joints

The pelvic girdle is firmly connected to the vertebral column at the sacroiliac joint. Therefore the body can be kept upright more easily even if only one limb is used for support (as in walking). The mobility of the lower limb is more limited than that of the upper limb.

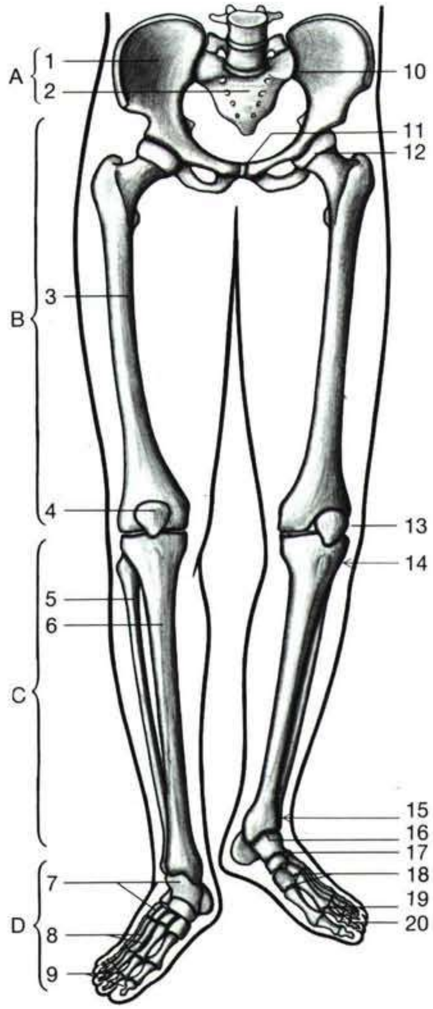

Organization of pelvic girdle and lower limb.

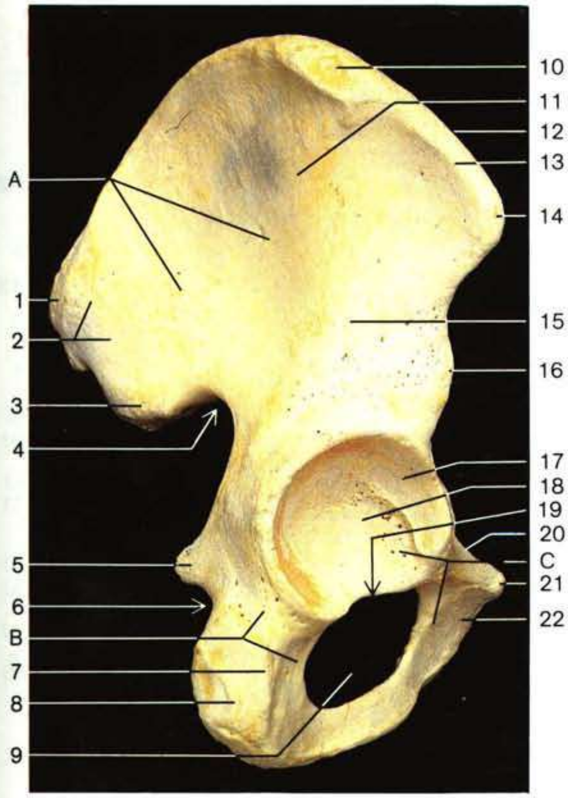

Right hip bone (lateral aspect).

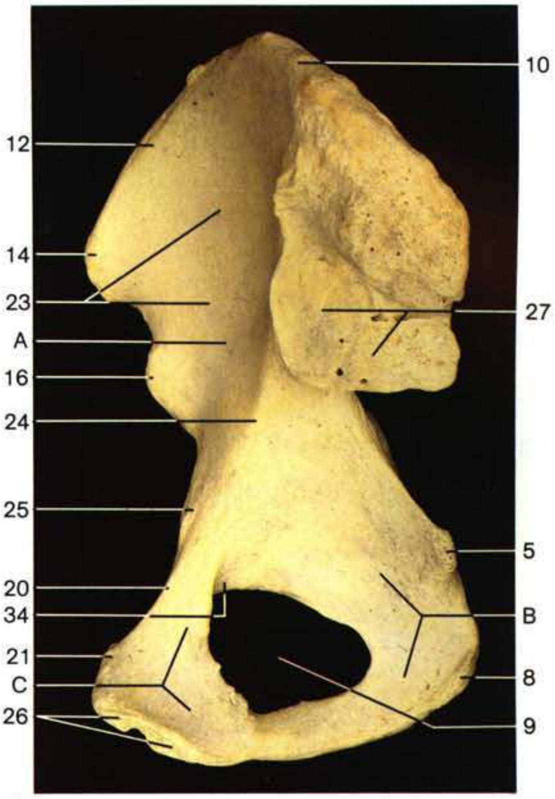

Right hip bone (medial aspect).

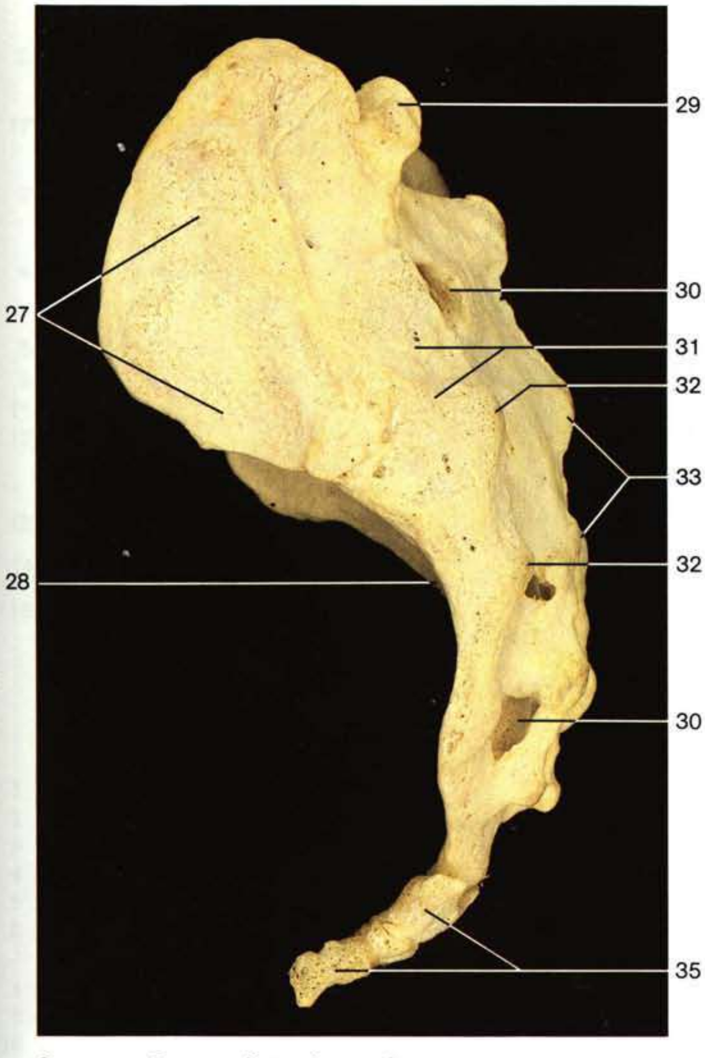

Sacrum and coccyx (lateral aspect).

A	Ilium
B	Ischium
C	Pubis
1	Posterior superior iliac spine
2	Posterior gluteal line
3	Posterior inferior iliac spine
4	Greater sciatic notch
5	Ischial spine
6	Lesser sciatic notch
7	Body of ischium
8	Ischial tuberosity
9	Obturator foramen
10	Iliac crest
11	Anterior gluteal line
12	Internal lip of iliac crest
13	External lip of iliac crest
14	Anterior superior iliac spine
15	Inferior gluteal line
16	Anterior inferior iliac spine
17	Lunate surface of acetabulum
18	Acetabular fossa
19	Acetabular notch
20	Pecten pubis
21	Pubic tubercle
22	Body of pubis
23	Iliac fossa
24	Arcuate line
25	Iliopubic eminence
26	Articular surface of pubis
27	Auricular surface of sacrum
28	Pelvic surface of sacrum
29	Superior articular process of sacrum
30	Dorsal sacral foramina
31	Sacral tuberosity
32	Lateral sacral crest
33	Median sacral crest
34	Obturator groove
35	Coccyx

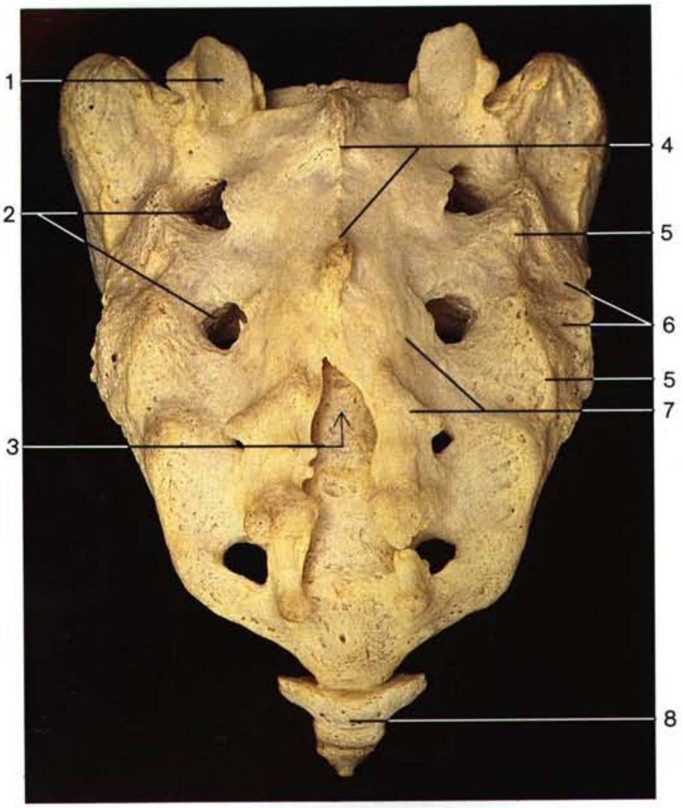

Sacrum (posterior aspect).

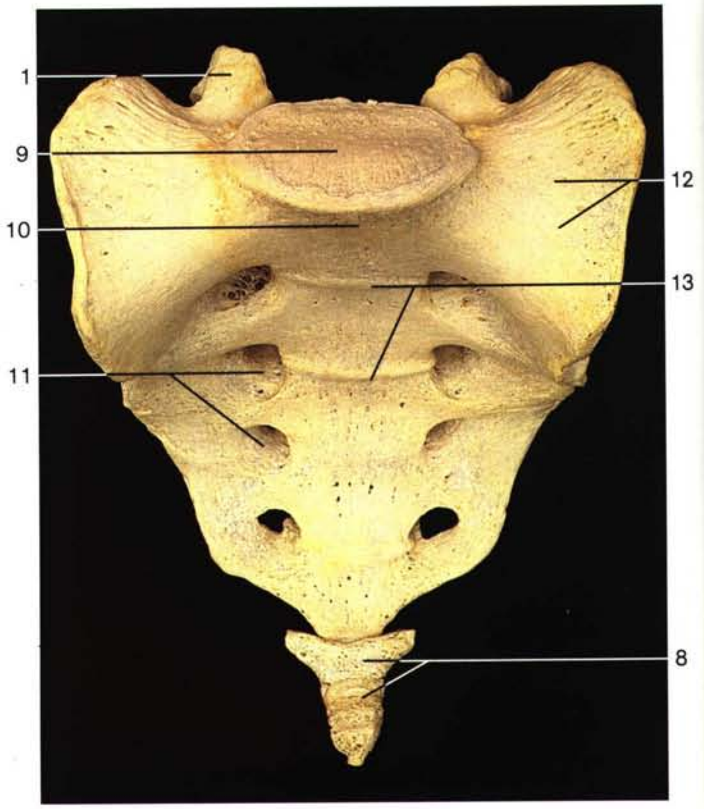

Sacrum (anterior aspect).

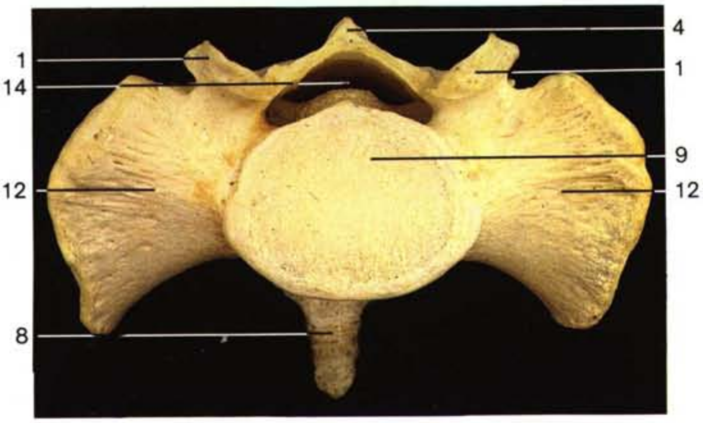

Sacrum (superior aspect).

1 Superior articular process of sacrum
2 Dorsal sacral foramina
3 Sacral hiatus
4 Median sacral crest
5 Lateral sacral crest
6 Sacral tuberosity
7 Intermediate sacral crest
8 Coccyx
9 Base of sacrum
10 Sacral promontory
11 Anterior sacral foramina
12 Lateral part of sacrum (ala)
13 Transverse line of sacrum
14 Sacral canal
15 Linea terminalis
16 True conjugate
17 Diagonal conjugate
18 Transverse diameter
19 Oblique diameter
20 Inferior pelvic aperture or outlet

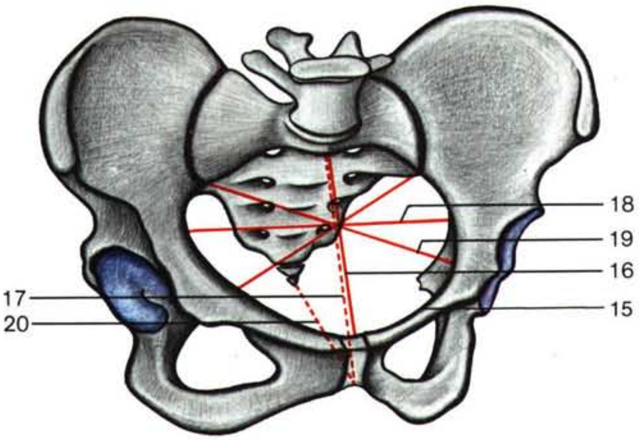

Diameters of pelvis (oblique superior aspect). (Schematic drawing.)

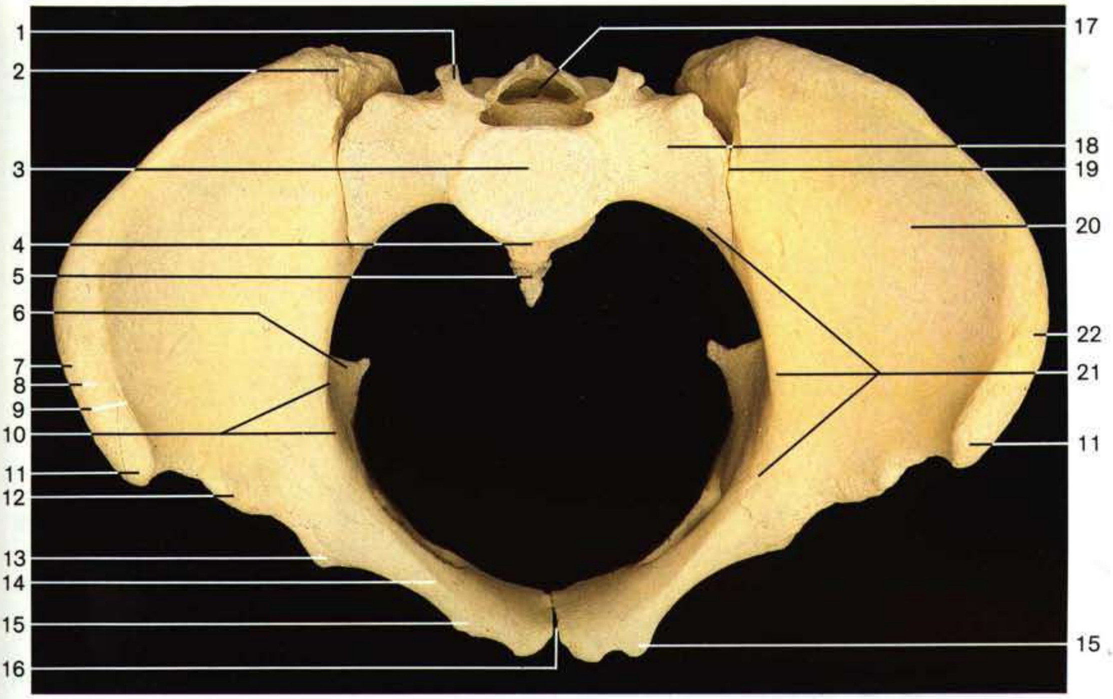
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
22
21
11
15

Female pelvis (superior aspect). Note the differences between the male and the female pelvis, predominantly in the form and dimensions of the sacrum, the superior and inferior apertures and the alae of the ilium.

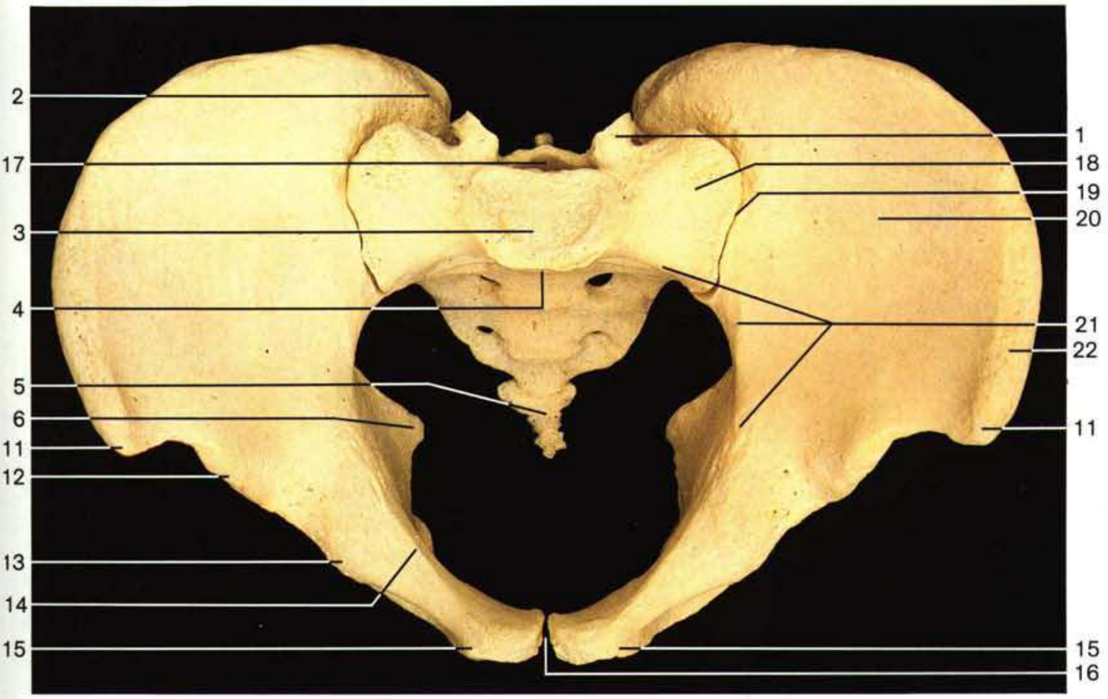
2
17
3
4
5
6
11
12
13
14
15
1
18
19
20
21
22
11
15
16

Male pelvis (superior aspect). Compare with the female pelvis (depicted above).

1 Superior articular process of sacrum
2 Posterior superior iliac spine
3 Base of sacrum
4 Sacral promontory
5 Coccyx
6 Ischial spine
7 External lip
8 Intermediate line
9 Internal lip
} of iliac crest
10 Arcuate line
11 Anterior superior iliac spine

12 Anterior inferior iliac spine
13 Iliopubic eminence
14 Pecten pubis
15 Pubic tubercle
16 Pubic symphysis
17 Sacral canal
18 Ala of sacrum
19 Position of sacroiliac joint
20 Iliac fossa
21 Linea terminalis
22 Iliac crest

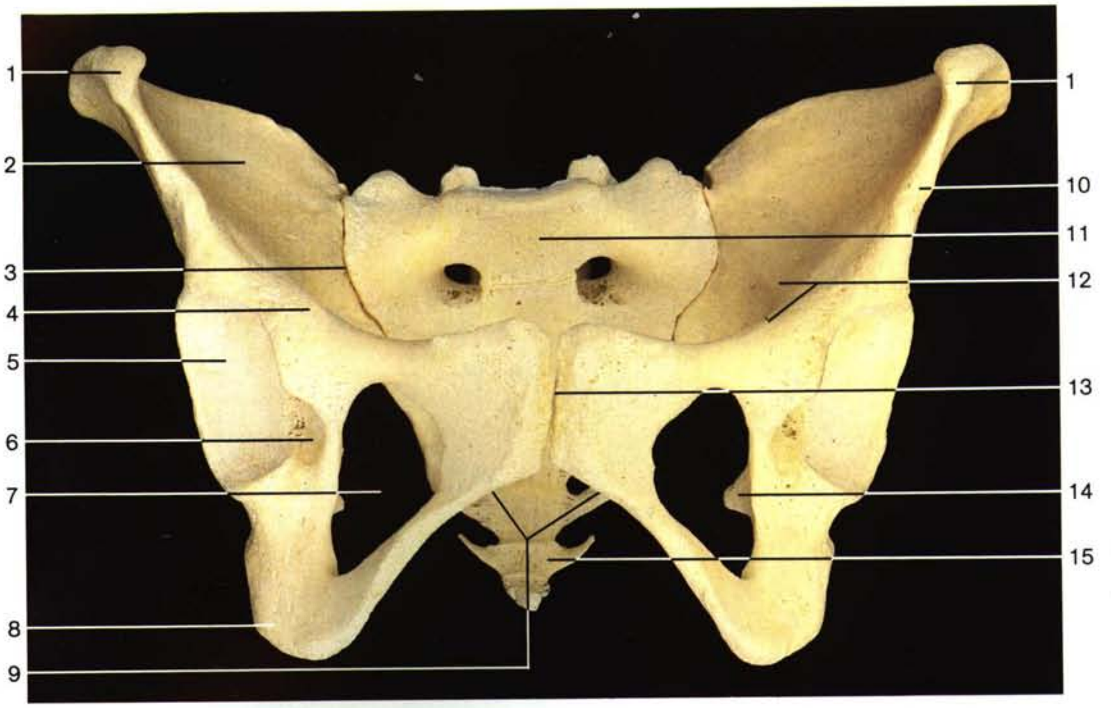

Female pelvis (anterior aspect). Note the differences between the form and dimensions of the male and the female pelvis. The female pubic arch is wider than the male. The obturator foramen in the female pelvis is triangular, while that in the male pelvis is ovoid.

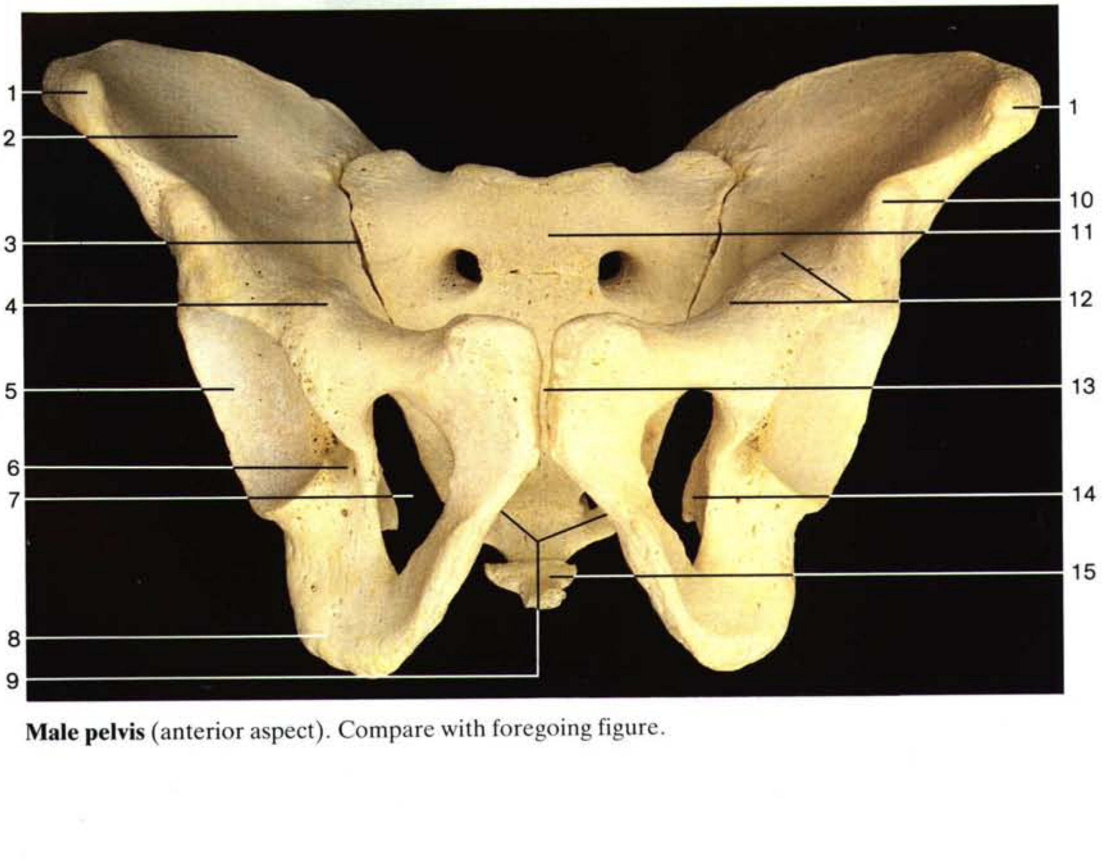

Male pelvis (anterior aspect). Compare with foregoing figure.

1 Anterior superior iliac spine
2 Iliac fossa
3 Position of sacroiliac joint
4 Iliopubic eminence
5 Lunate surface of acetabulum
6 Acetabular notch
7 Obturator foramen
8 Ischial tuberosity

9 Pubic arch
10 Anterior inferior iliac spine
11 Sacrum
12 Linea terminalis (at margin of superior aperture)
13 Pubic symphysis
14 Ischial spine
15 Coccyx

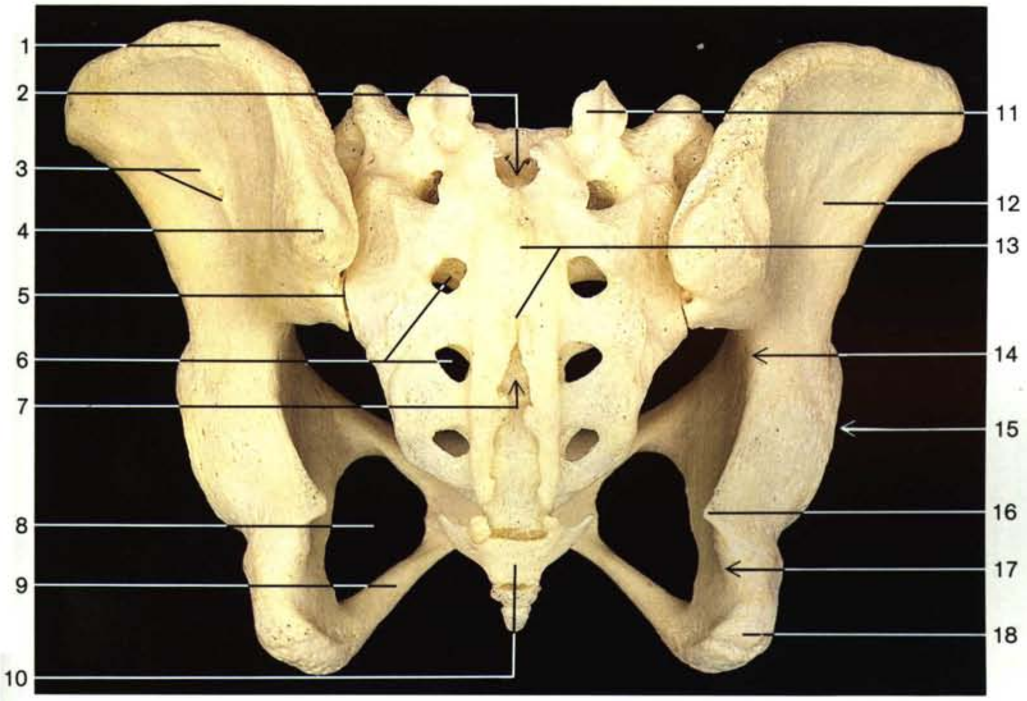

Female pelvis (posterior inferior aspect). Note the differences between the female and male pelvis, especially with respect to the inferior aperture, the shape of the sacrum, the two sciatic notches and the pubic arch.

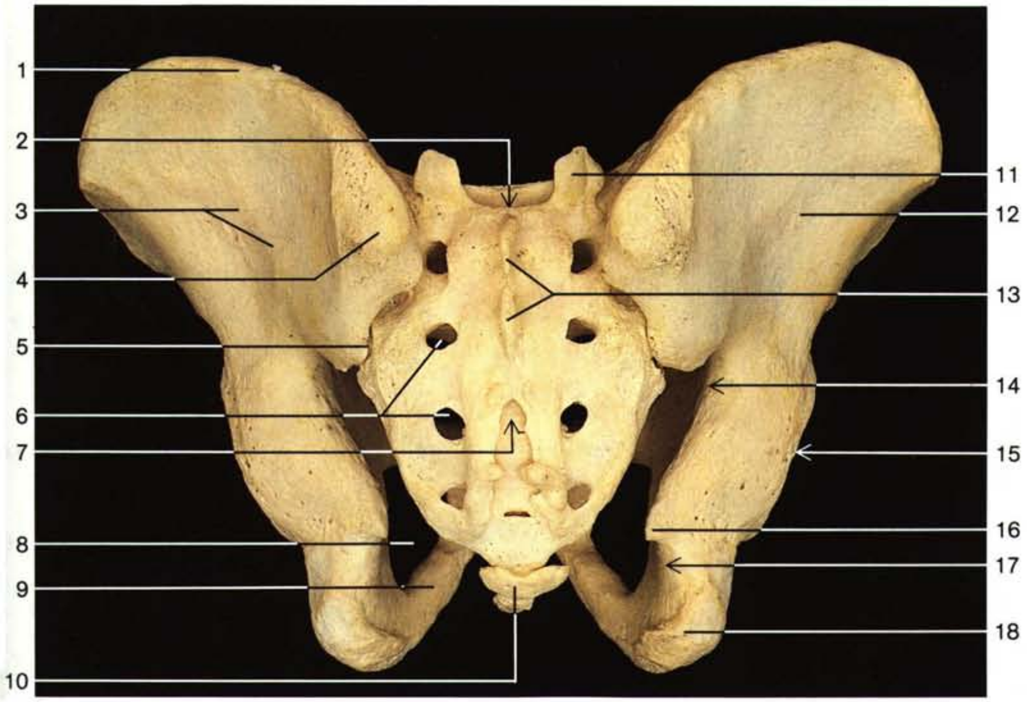

Male pelvis (posterior inferior aspect). Compare with the female pelvis (depicted above).

1 Iliac crest
2 Sacral canal
3 Posterior gluteal line
4 Posterior superior iliac spine
5 Position of sacroiliac joint
6 Dorsal sacral foramina
7 Sacral hiatus
8 Obturator foramen
9 Ramus of ischium
10 Coccyx
11 Superior articular process of sacrum
12 Gluteal surface of ilium
13 Median sacral crest
14 Greater sciatic notch
15 Position of acetabulum
16 Ischial spine
17 Lesser sciatic notch
18 Ischial tuberosity

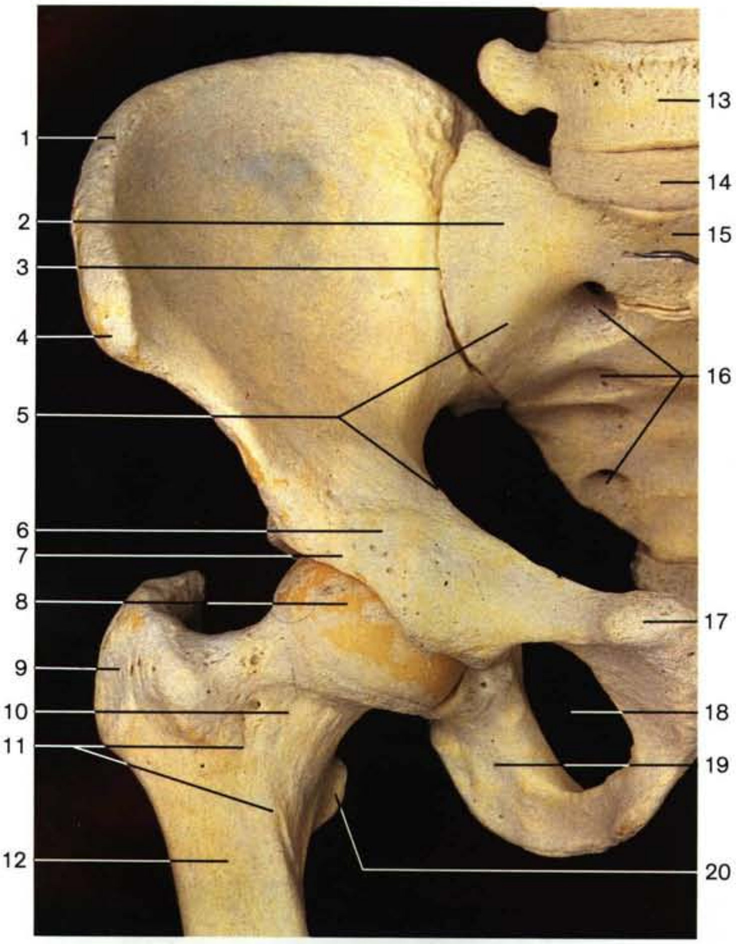

Bones of right hip joint (anterior aspect).

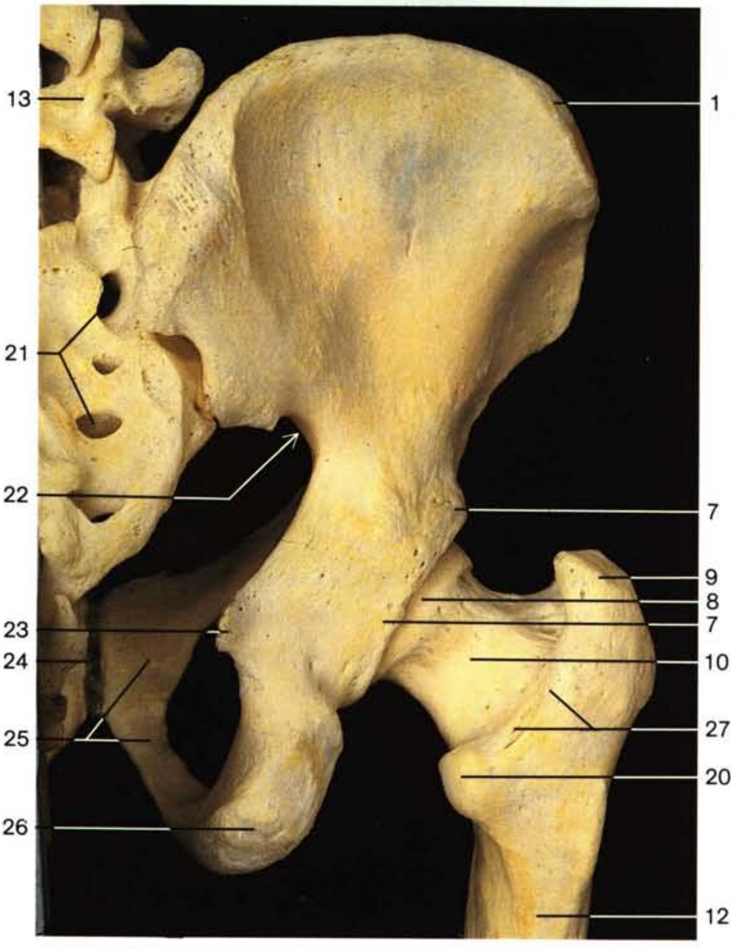

Bones of right hip joint (posterior aspect).

1 Iliac crest
2 Lateral part of sacrum (ala)
3 Position of sacroiliac joint
4 Anterior superior iliac spine
5 Linea terminalis
6 Iliopubic eminence
7 Bony margin of acetabulum
8 Head of femur
9 Greater trochanter
10 Neck of femur
11 Intertrochanteric line
12 Shaft of femur
13 5th lumbar vertebra
14 Imitation intervertebral disc between 5th lumbar vertebra and sacrum
15 Sacral promontory
16 Anterior sacral foramina
17 Pubic tubercle
18 Obturator foramen
19 Ramus of ischium
20 Lesser trochanter
21 Dorsal sacral foramina
22 Greater sciatic notch
23 Ischial spine
24 Pubic symphysis
25 Pubis
26 Ischial tuberosity
27 Intertrochanteric crest
28 Symphysial surface

Diameters of the pelvis

A True conjugate (11–11.5 cm) (Conjugata vera)
B Diagonal conjugate (12.5–13 cm)
C Largest diameter of pelvis
D Inferior pelvic aperture
E Pelvic inclination (60°)

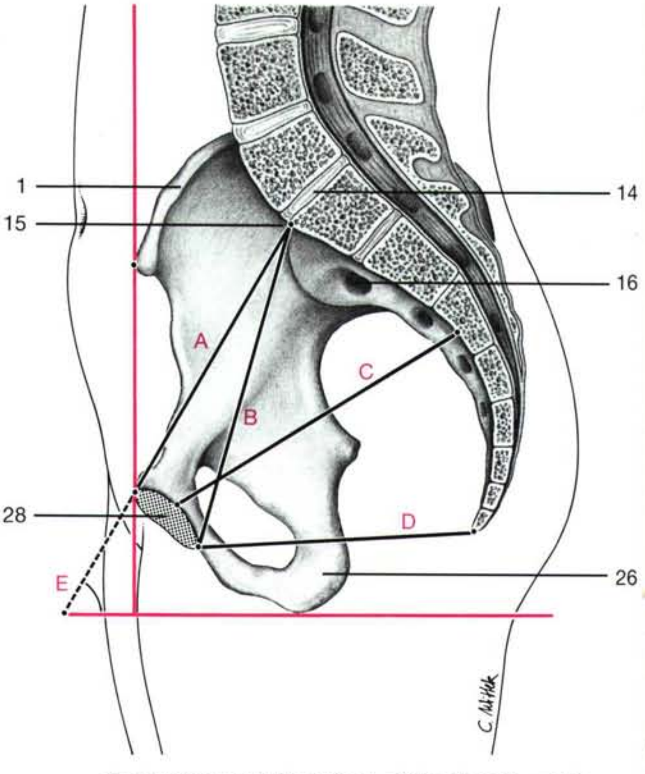

Inclination and diameters of the female pelvis, right half (medial aspect).

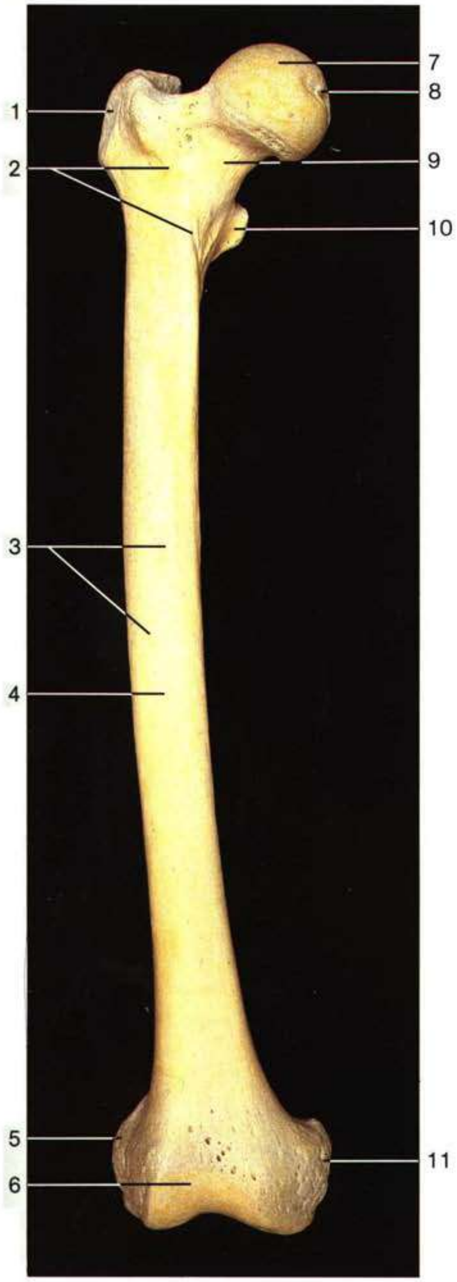
7
1
8
2
9
10
3
4
5
11
6

Right femur (anterior aspect).

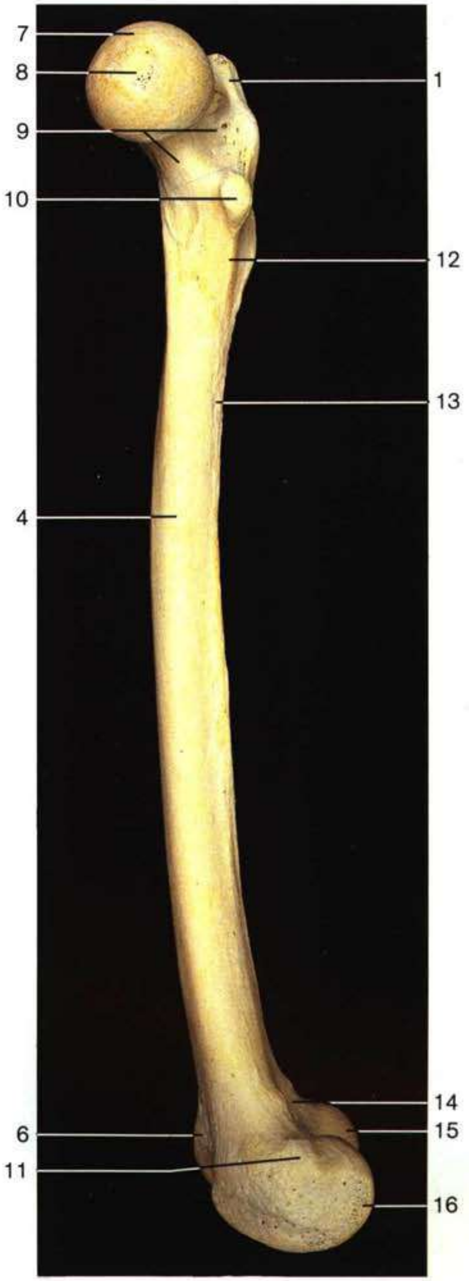
7
8
1
9
10
12
13
4
14
6
15
11
16

Right femur (medial aspect).

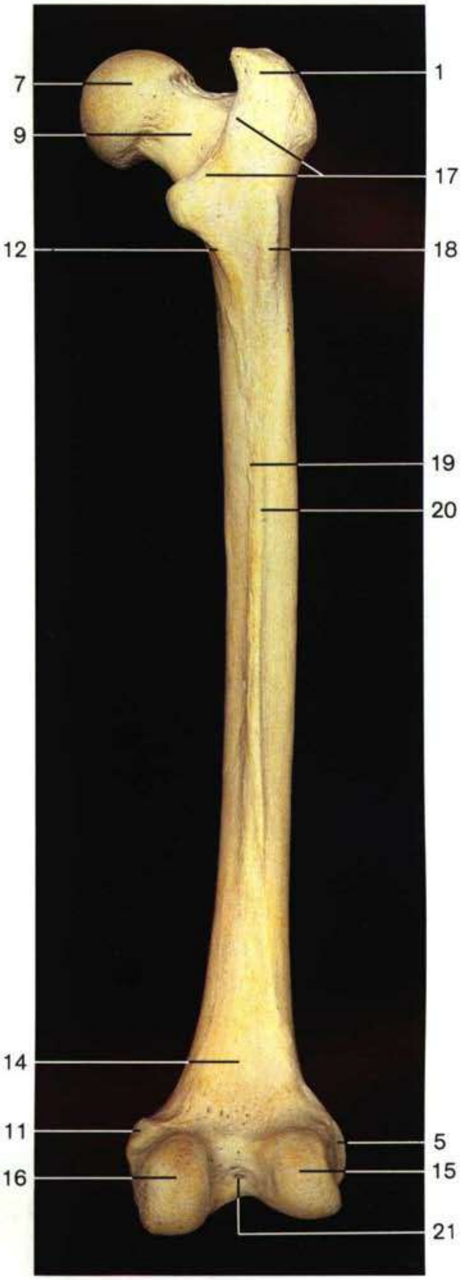
7
1
9
17
12
18
19
20
14
11
5
16
15
21

Right femur (posterior aspect).

1	Greater trochanter
2	Intertrochanteric line
3	Nutrient foramina
4	Shaft of femur (diaphysis)
5	Lateral epicondyle
6	Patellar surface
7	Head
8	Fovea of head
9	Neck
10	Lesser trochanter
11	Medial epicondyle
12	Pectineal line
13	Linea aspera
14	Popliteal surface
15	Lateral condyle
16	Medial condyle
17	Intertrochanteric crest
18	Third trochanter
19	Medial lip of linea aspera
20	Lateral lip of linea aspera
21	Intercondylar fossa

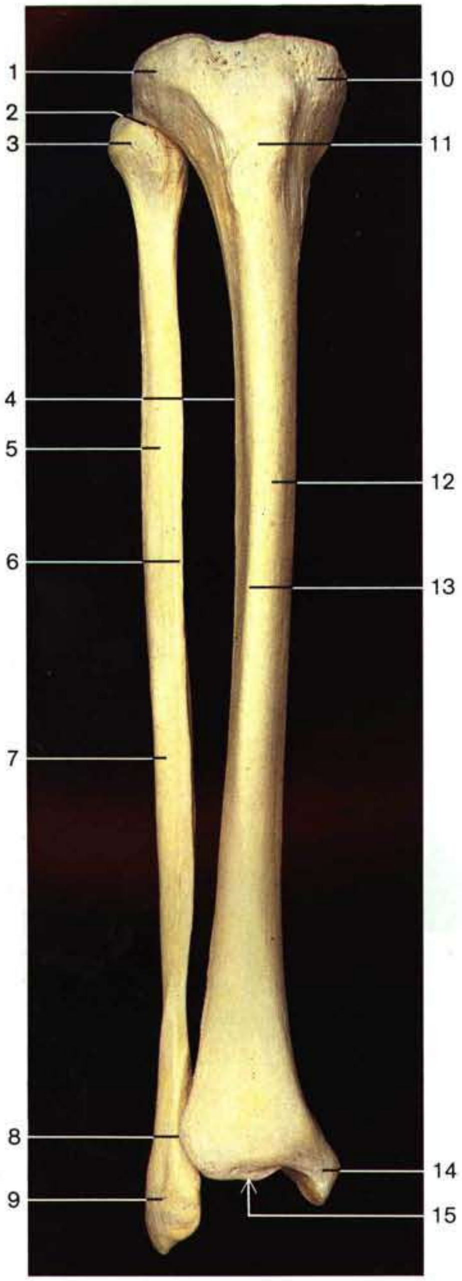

Bones of leg, right tibia and fibula (anterior aspect).

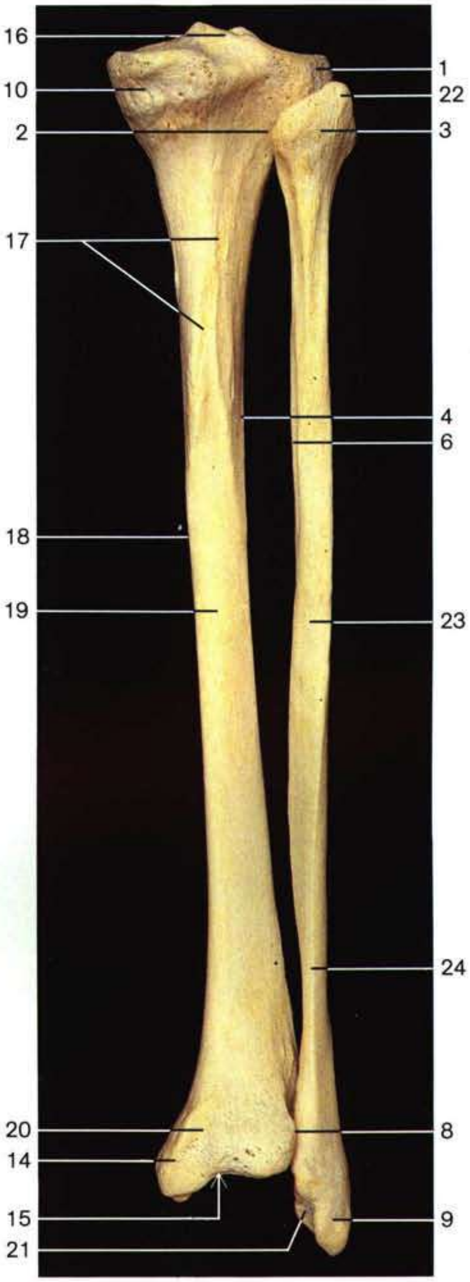

Bones of leg, right tibia and fibula (posterior aspect).

1 Lateral condyle of tibia
2 Position of tibiofibular joint
3 Head of fibula
4 Interosseous border of tibia
5 Shaft of fibula
6 Interosseous border of fibula
7 Lateral surface of fibula
8 Position of tibiofibular joint
9 Lateral malleolus
10 Medial condyle of tibia
11 Tuberosity of tibia
12 Shaft of tibia (diaphysis)
13 Anterior margin of tibia
14 Medial malleolus
15 Inferior articular surface of tibia
16 Intercondylar eminence
17 Soleal line
18 Medial border of tibia
19 Posterior surface of tibia
20 Malleolar sulcus of tibia
21 Malleolar articular surface of fibula
22 Apex of head of fibula
23 Posterior surface of fibula
24 Posterior border of fibula
25 Medial intercondylar tubercle
26 Posterior intercondylar area
27 Anterior intercondylar area
28 Lateral intercondylar tubercle

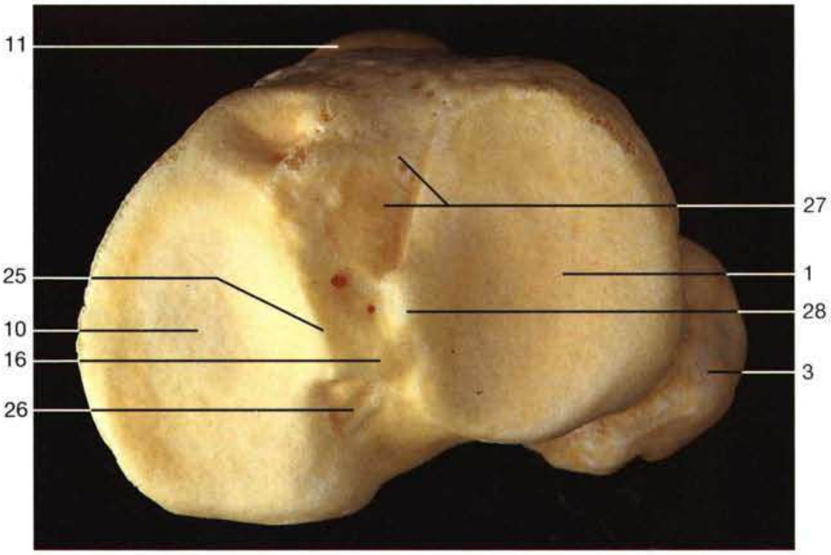

Upper end of right tibia with fibula (from above), anterior margin of tibia above. Superior articular surface of tibia.

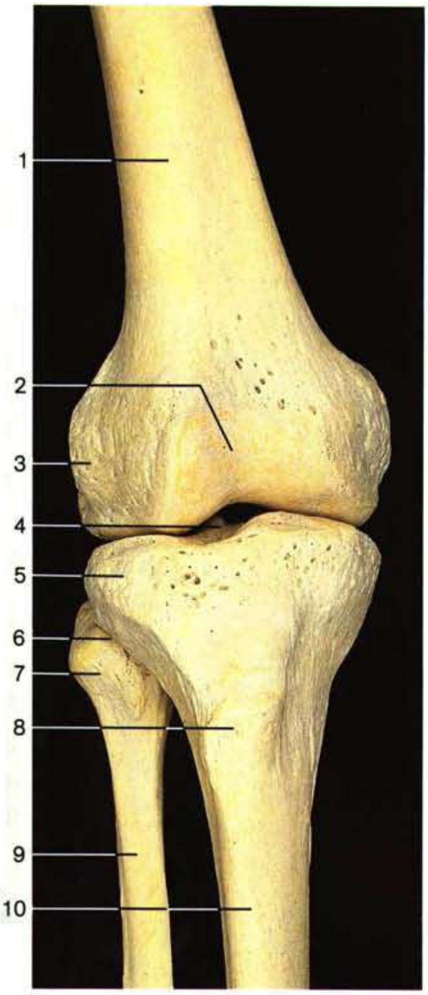

Bones of right knee joint
(anterior aspect).

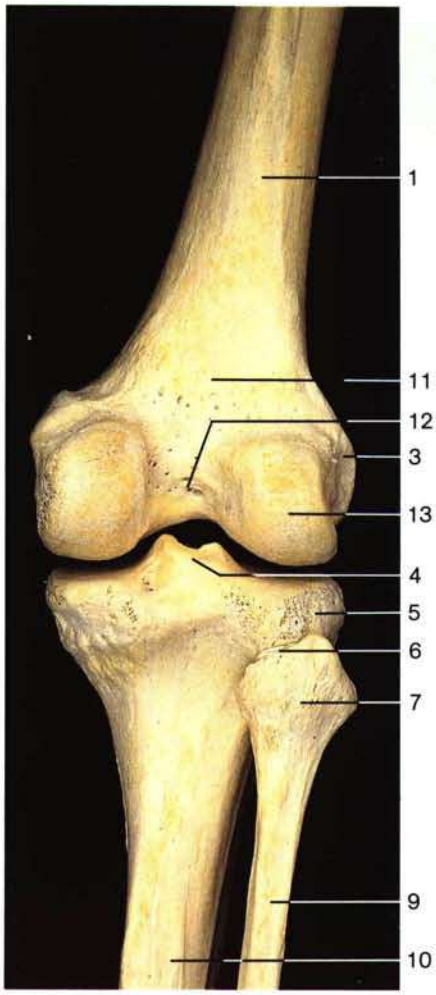

Bones of right knee joint
(posterior aspect).

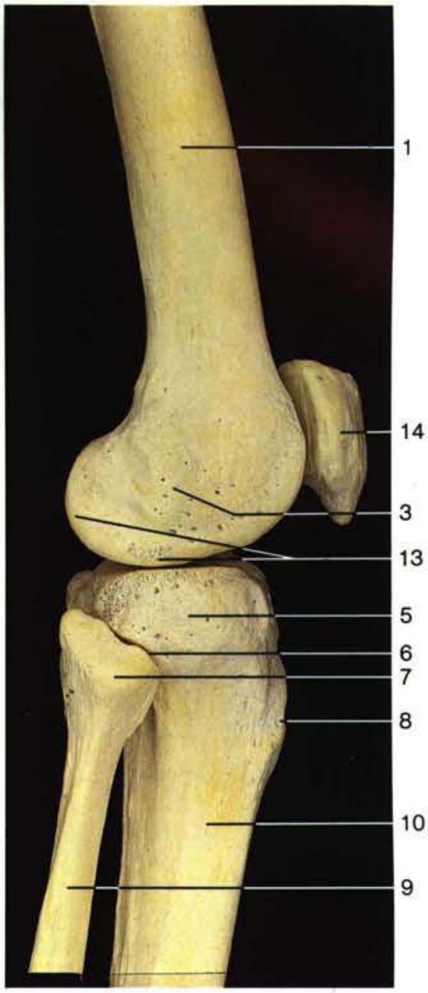

Bones of right knee joint
(lateral aspect).

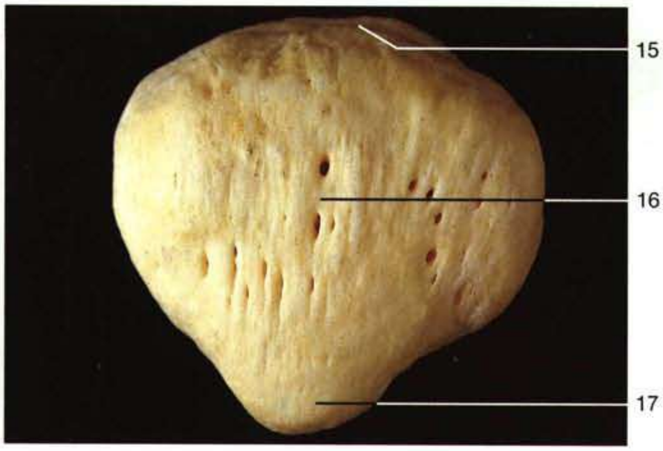

Right patella (anterior aspect).

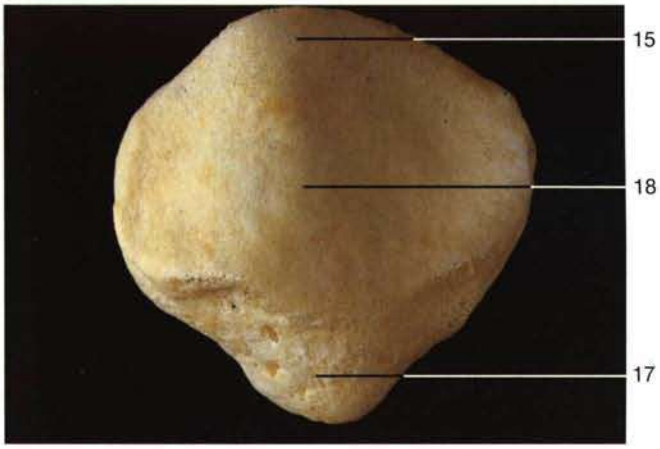

Right patella (posterior aspect).

1 **Femur**
2 Patellar surface of femur
3 Lateral epicondyle of femur
4 Intercondylar eminence of tibia
5 Lateral condyle of tibia
6 Position of tibiofibular joint
7 Head of fibula
8 Tuberosity of **tibia**
9 **Fibula**

10 Shaft of tibia
11 Popliteal surface of femur
12 Intercondylar fossa of femur
13 Lateral condyle of femur
14 **Patella**
15 Base of patella
16 Anterior surface of patella
17 Apex of patella
18 Articular surface of patella

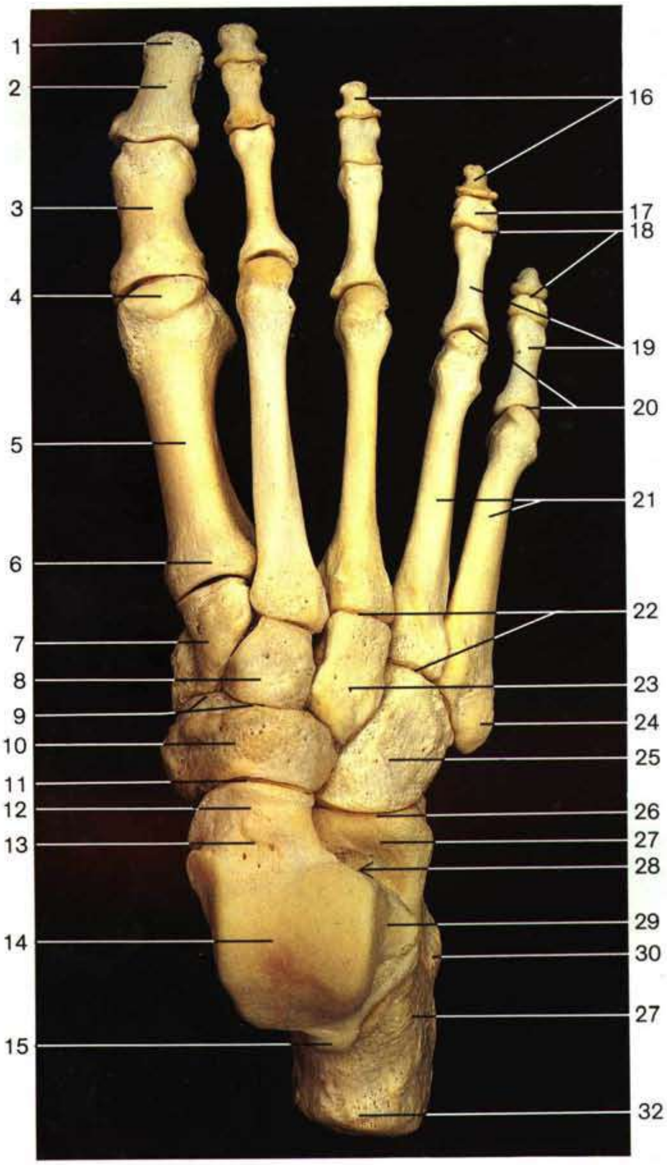
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
27
32

Bones of right foot (dorsal aspect).

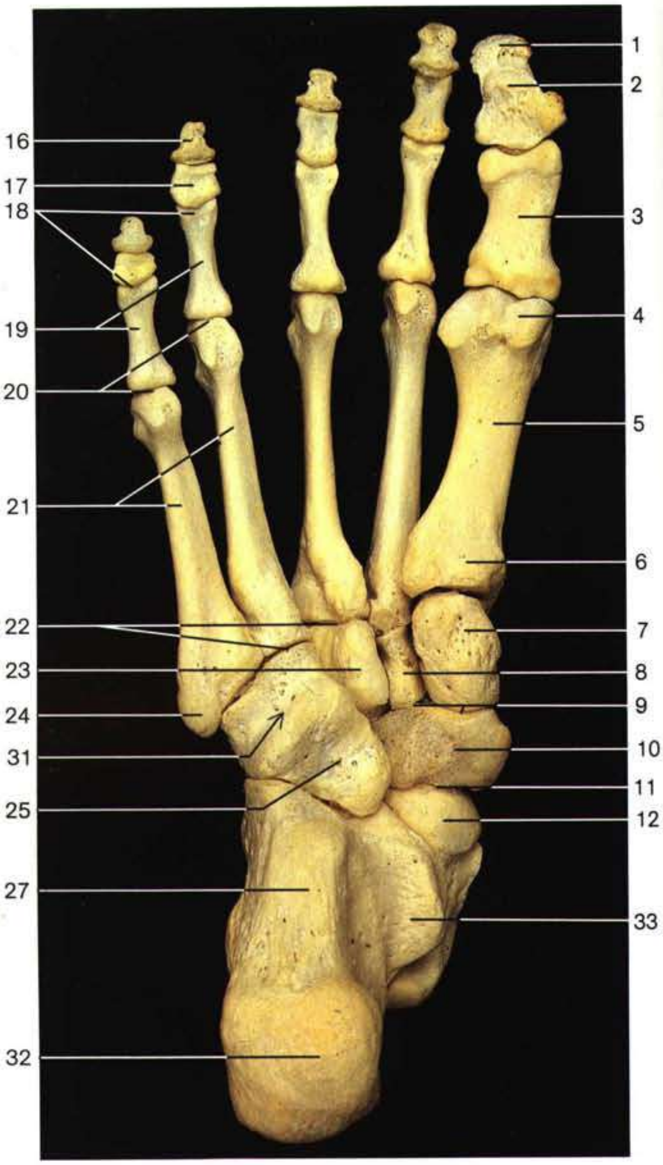
16
17
18
19
20
21
22
23
24
31
25
27
32
1
2
3
4
5
6
7
8
9
10
11
12
33

Bones of right foot (plantar aspect).

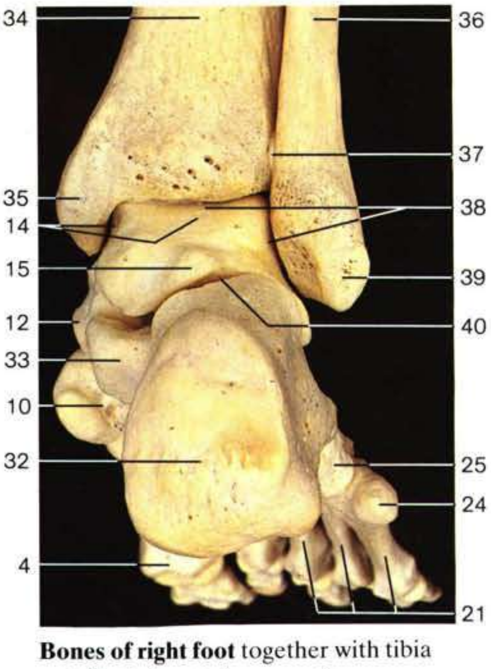
34
35
14
15
12
33
10
32
4
36
37
38
39
40
25
24
21

Bones of right foot together with tibia and fibula (posterior aspect).

1 Tuberosity of distal phalanx of great toe
2 Distal phalanx of great toe
3 Proximal phalanx of great toe
4 Head of first metatarsal bone
5 First metatarsal bone
6 Base of first metatarsal bone
7 Medial cuneiform bone
8 Intermediate cuneiform bone
9 Position of **cuneonavicular joint**
10 Navicular bone

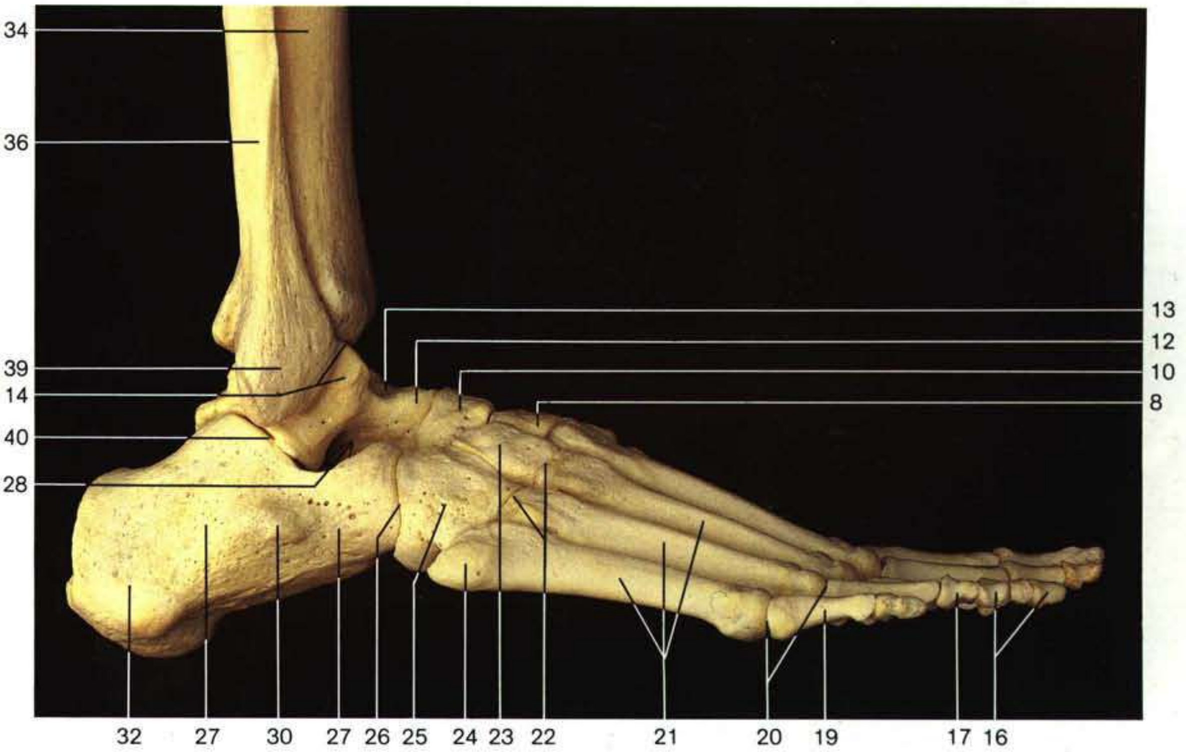

Bones of right foot, tibia and fibula (lateral aspect).

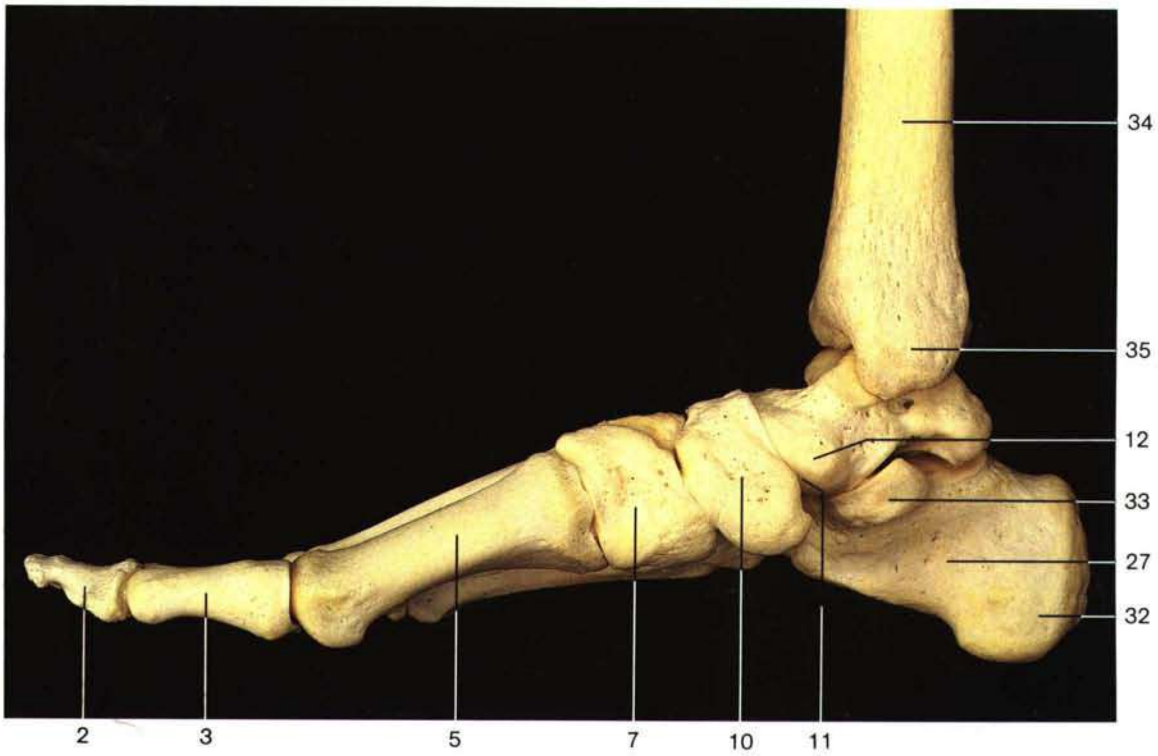

Bones of right foot, tibia and fibula (medial aspect).

11 Position of **talocalcaneonavicular joint**
12 Head of talus
13 Neck of talus
14 Trochlea of talus
15 Posterior talar process
16 Distal phalanges
17 Middle phalanges
18 Position of **interphalangeal joints**
19 Proximal phalanges
20 Position of **metatarsophalangeal joints**

21 Metatarsal bones
22 Position of **tarsometatarsal joints**
23 Lateral cuneiform bone
24 Tuberosity of 5th metatarsal bone
25 Cuboid bone
26 Position of **calcaneocuboid joint**
27 Calcaneus
28 Tarsal sinus
29 Lateral malleolar surface of talus
30 Peroneal trochlea of calcaneus

31 Groove for tendon of peroneus longus
32 Calcaneal tuberosity
33 Sustentaculum tali
34 Tibia
35 Medial malleolus
36 Fibula
37 Position of tibiofibular syndesmosis
38 Position of **talocrural joint**
39 Lateral malleolus
40 Position of **subtalar joint**

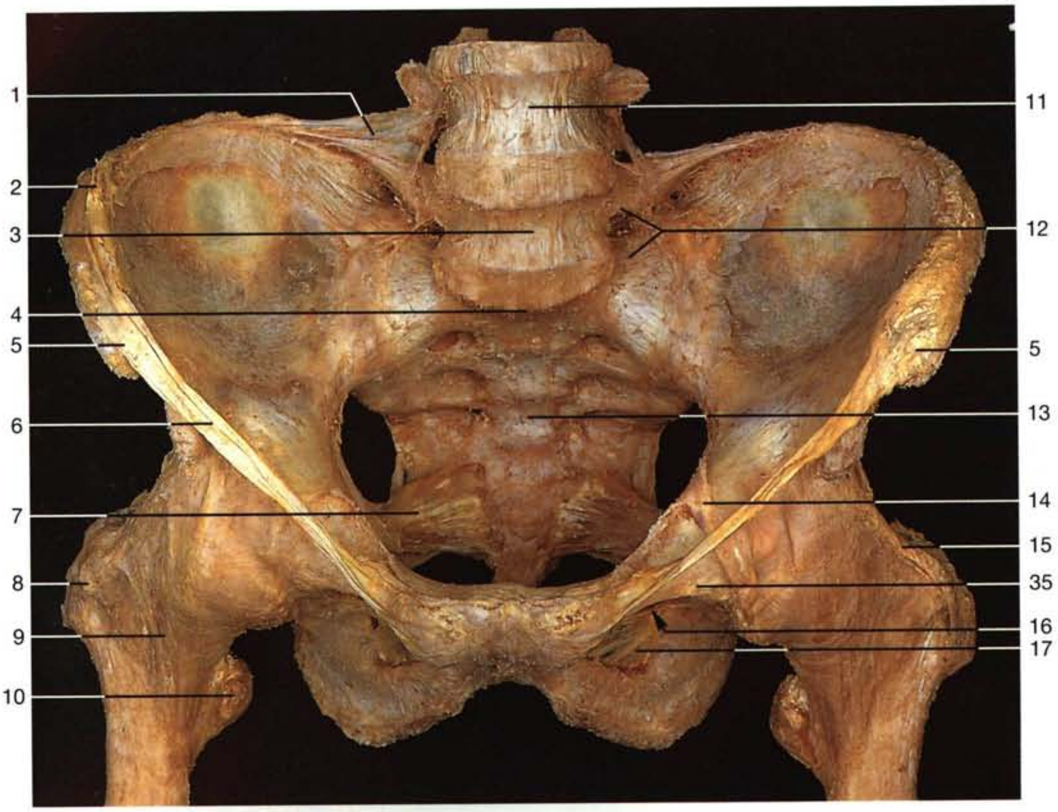

Ligaments of pelvis and hip joint (anterior aspect).

1 **Iliolumbar ligament**
2 Iliac crest
3 Fifth lumbar vertebra
4 Sacral promontory
5 Anterior superior iliac spine
6 Inguinal ligament
7 Sacrospinous ligament
8 Greater trochanter
9 **Iliofemoral ligament** (vertical band)
10 Lesser trochanter
11 Forth lumbar vertebra
12 Iliolumbar and ventral sacroiliac ligaments

13 Sacrum
14 Iliopectineal arch
15 **Iliofemoral ligament** (horizontal band)
16 Obturator canal
17 Obturator membrane
18 **Greater sciatic foramen**
19 Sacrospinous ligament
20 Sacrotuberous ligament
21 **Lesser sciatic foramen**
22 Ischial tuberosity
23 **Ischiofemoral ligament**
24 Intertrochanteric crest

25 Femur
26 Articular capsule of hip joint
27 Dorsal sacroiliac ligaments
28 Coccyx with superficial dorsal sacrococcygeal ligament
29 Head of femur
30 Articular cartilage of head of femur
31 Articular cavity of hip joint
32 Acetabular lip
33 Spongy bone
34 Ligament of head of femur
35 Pubofemoral ligament
36 Zona orbicularis

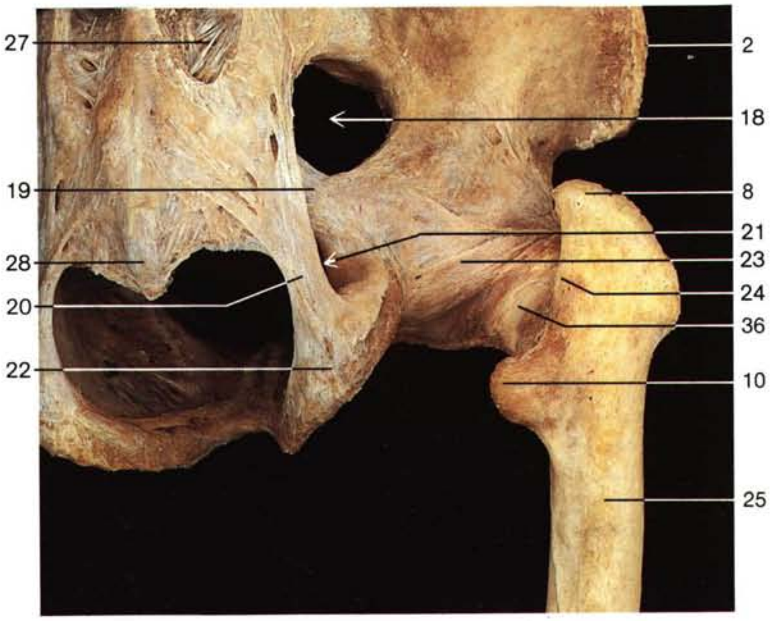

Ligaments of pelvis and hip joint (left anterior aspect).

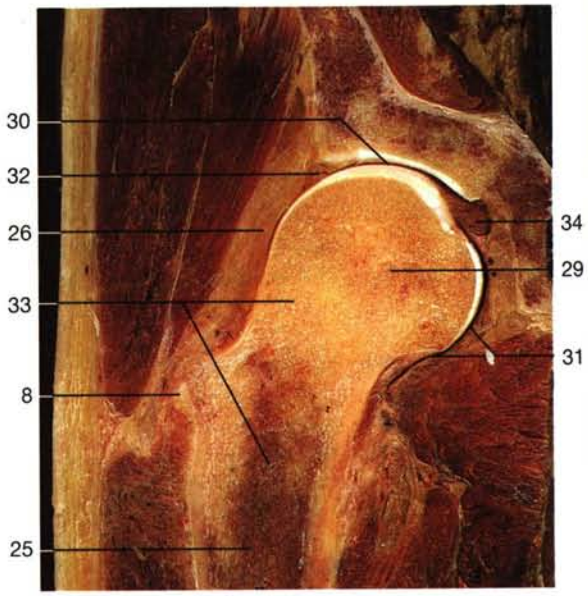

Coronal section of right hip joint (anterior view).

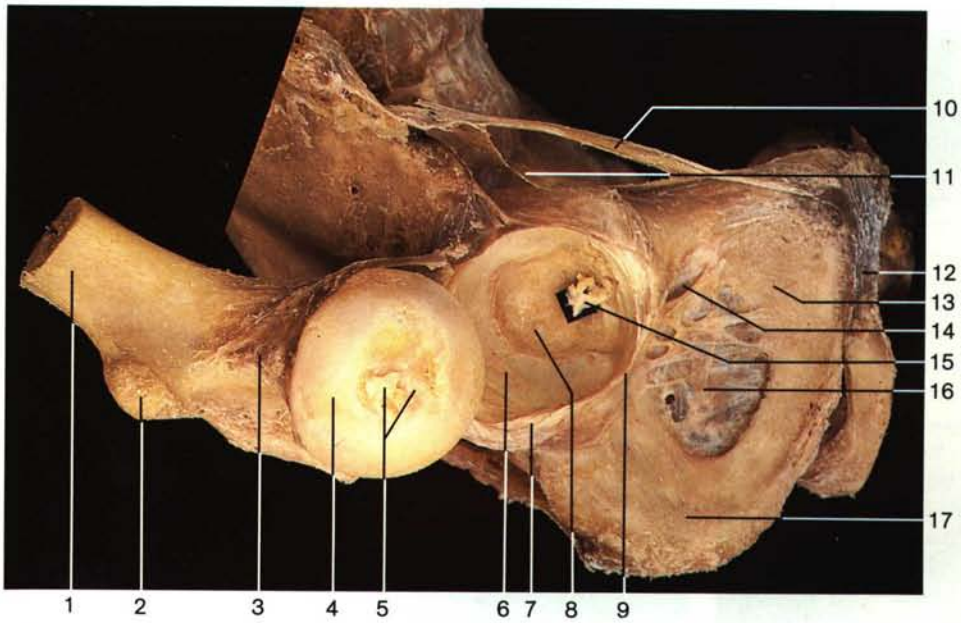

Right hip joint, opened (lateral anterior aspect). The ligament of the head of the femur has been divided and the femur has been posteriorly reflected.

1 Femur
2 Lesser trochanter
3 Neck of femur
4 Head of femur
5 Fovea of head with remnants of ligament of head
6 Lunate surface of acetabulum
7 Acetabular lip
8 Acetabular fossa
9 Transverse acetabular ligament
10 Inguinal ligament
11 Iliopectineal arch
12 Pubic symphysis
13 Pubic bone
14 Obturator canal
15 Ligament of head of femur
16 Obturator membrane
17 Ischium
18 Anterior longitudinal ligament (level of fifth lumbar vertebra)
19 Sacral promontory
20 Iliolumbar ligament
21 Iliac crest
22 Anterior superior iliac spine
23 Iliofemoral ligament (horizontal band)
24 Iliofemoral ligament (vertical band)
25 Greater trochanter
26 Pubofemoral ligament
27 Anterior inferior iliac spine
28 Ventral sacroiliac ligaments
29 **Sacrospinous ligament**
30 **Sacrotuberous ligament**
31 Intertrochanteric line
32 **Ischiofemoral ligament**
33 Zona orbicularis

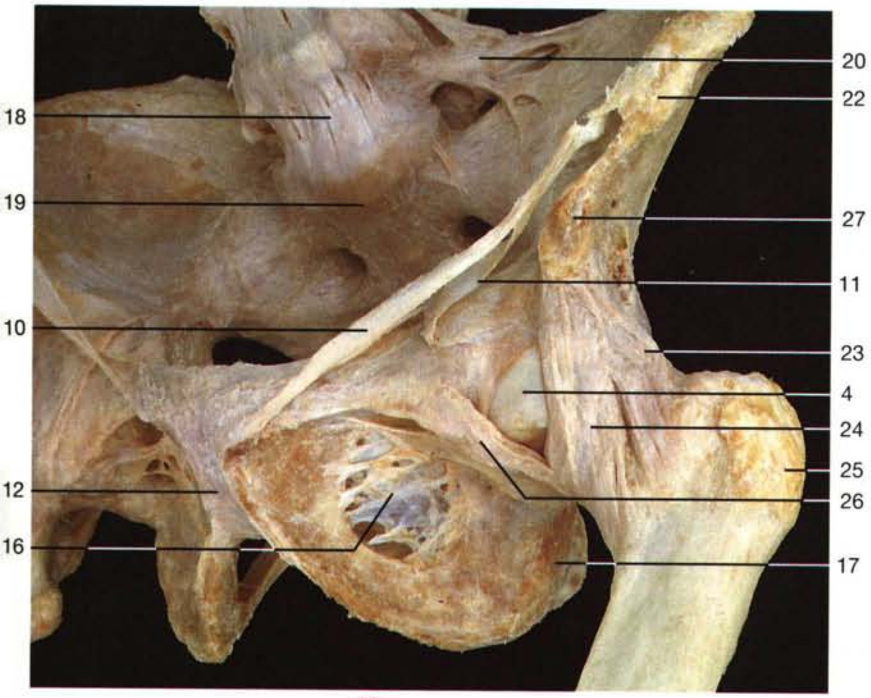

Ligaments of the pelvis and hip joint (antero-lateral aspect).

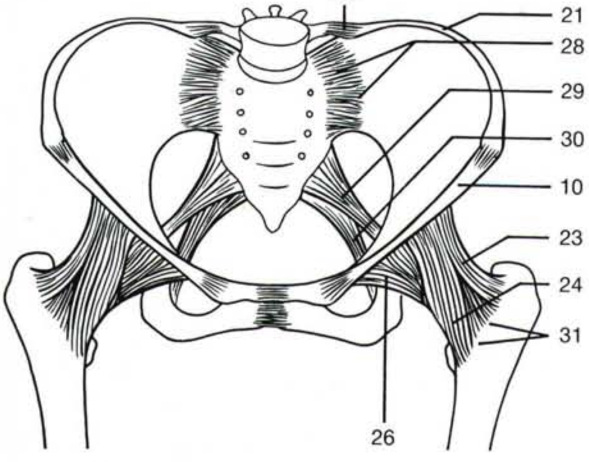

Ligaments of hip joint (anterior aspect). (Schematic drawing.)

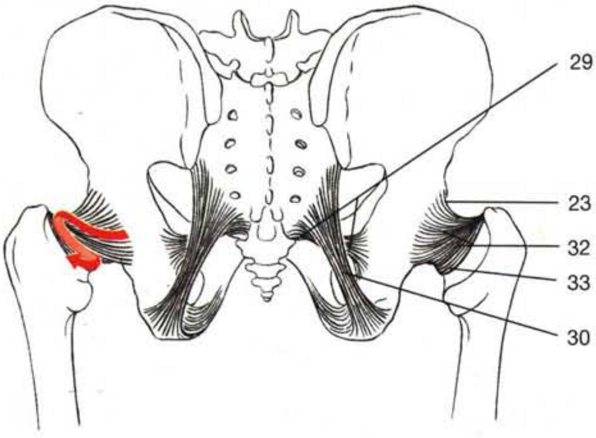

Ligaments of hip joints (posterior aspect). (Schematic drawing.)

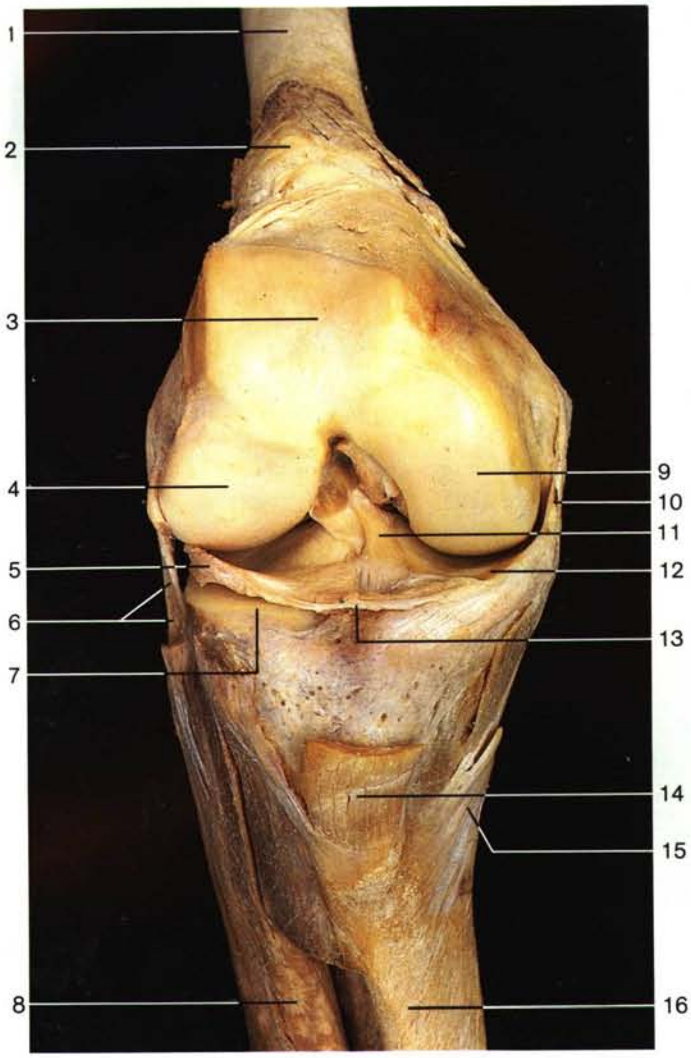

Right knee joint (opened) **with ligaments** (anterior aspect). The patella and articular capsule have been removed and the femur slightly flexed.

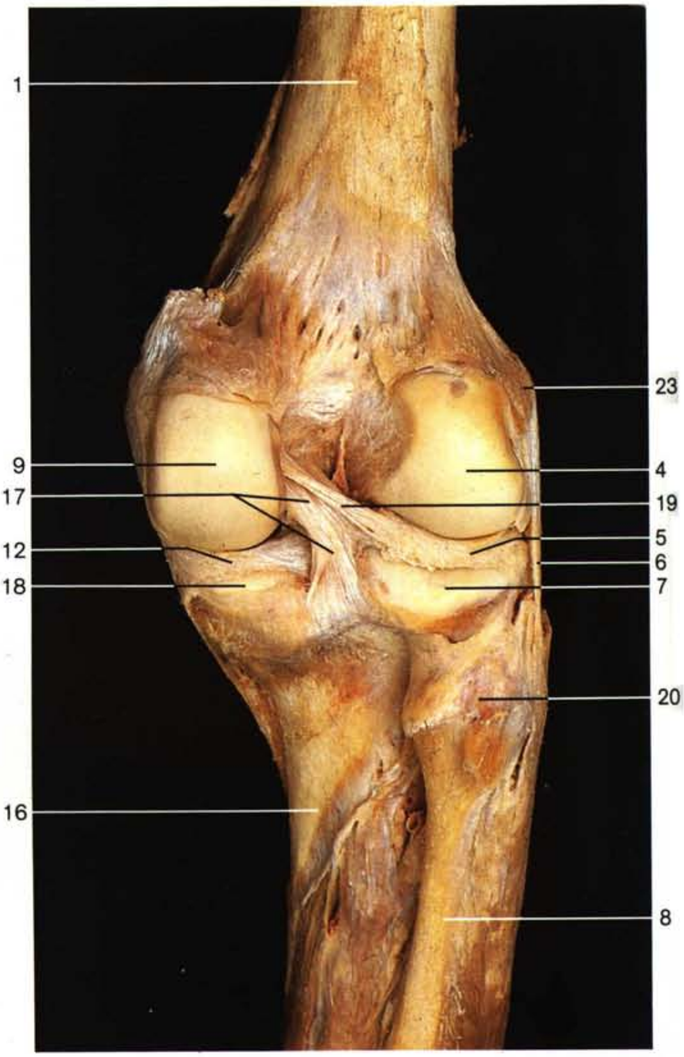

Right knee joint with ligaments (posterior aspect). The joint is extended and the articular capsule has been removed.

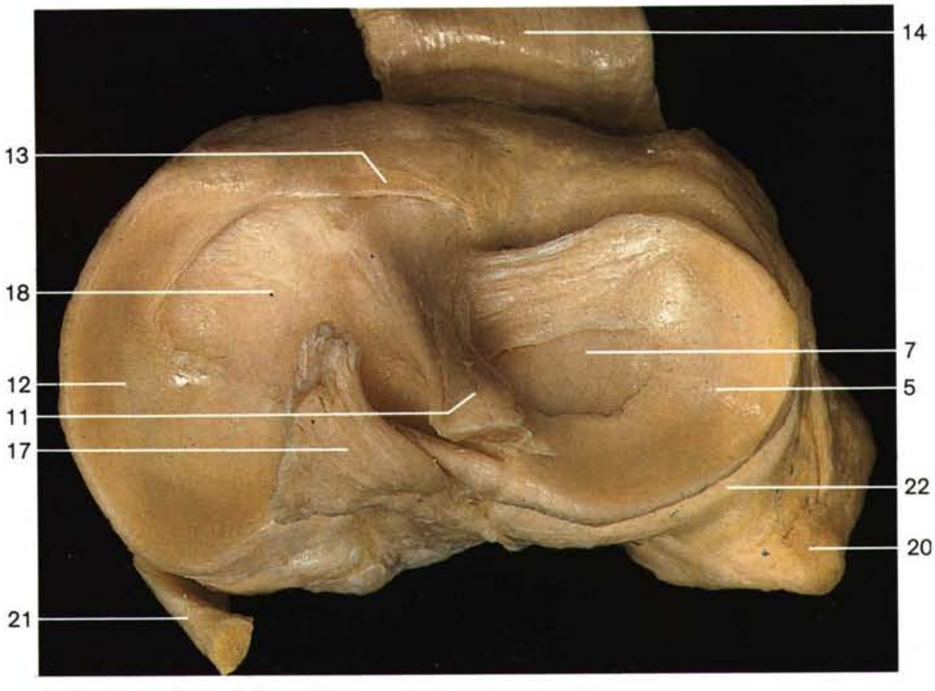

Articular surface of right tibia, menisci, and cruciate ligaments (superior aspect). Anterior margin of tibia above.

1 Femur
2 Articular capsule with **suprapatellar bursa**
3 Patellar surface
4 Lateral condyle of femur
5 **Lateral meniscus of knee joint**
6 Fibular collateral ligament
7 Lateral condyle of tibia (superior articular surface)
8 Fibula
9 Medial condyle of femur
10 Tibial collateral ligament
11 **Anterior cruciate ligament**
12 **Medial meniscus of knee joint**
13 Transverse ligament of knee
14 Patellar ligament
15 Common tendon of sartorius, semitendinosus and gracilis muscles
16 Tibia
17 **Posterior cruciate ligament**
18 Medial condyle of tibia (superior articular surface)
19 Posterior meniscofemoral ligament
20 Head of fibula
21 Tendon of semimembranosus muscle
22 Posterior attachment of articular capsule of knee joint
23 Lateral epicondyle of femur

1 Iliotibial tract
2 Articular muscle of knee
3 Patellar surface
4 Lateral condyle of femur
5 Articular capsule
6 Infrapatellar fat pad
7 **Patella** (articular surface)
8 **Suprapatellar bursa**
9 Quadriceps muscle of thigh (divided)
10 **Anterior cruciate ligament**
11 Medial condyle of femur
12 Fibular collateral ligament
13 **Posterior cruciate ligament**
14 Medial epicondyle of femur
15 Intercondylar fossa of femur
16 Tibial collateral ligament
17 **Medial meniscus** of knee joint
18 Medial intercondylar tubercle
19 Femur
20 Lateral epicondyle of femur
21 **Lateral meniscus** of knee joint
22 Epiphyseal line of tibia
23 Tibia

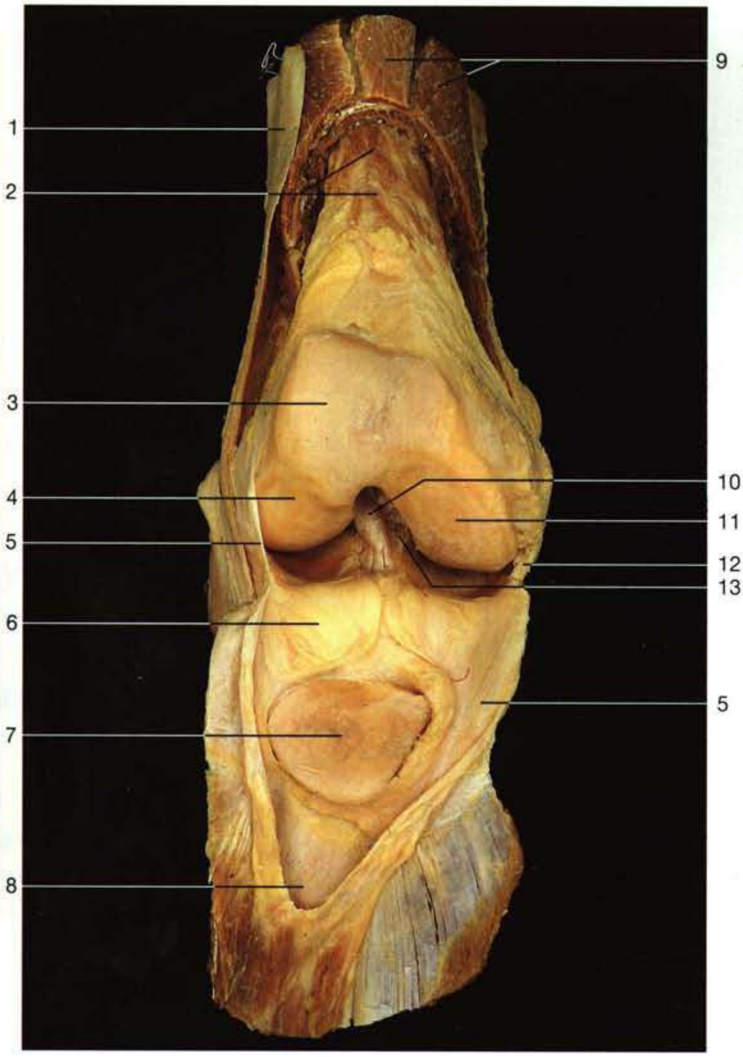

Right knee joint, opened (anterior aspect). Patellar ligament with patella reflected.

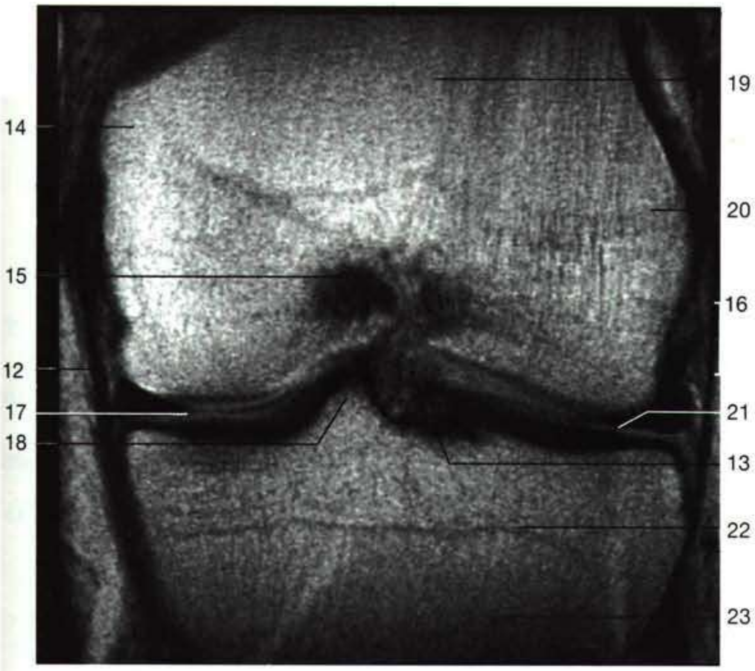

Right knee joint. Frontal section through the central part of the joint (posterior aspect, MR-Scan). (See also p. 10.)

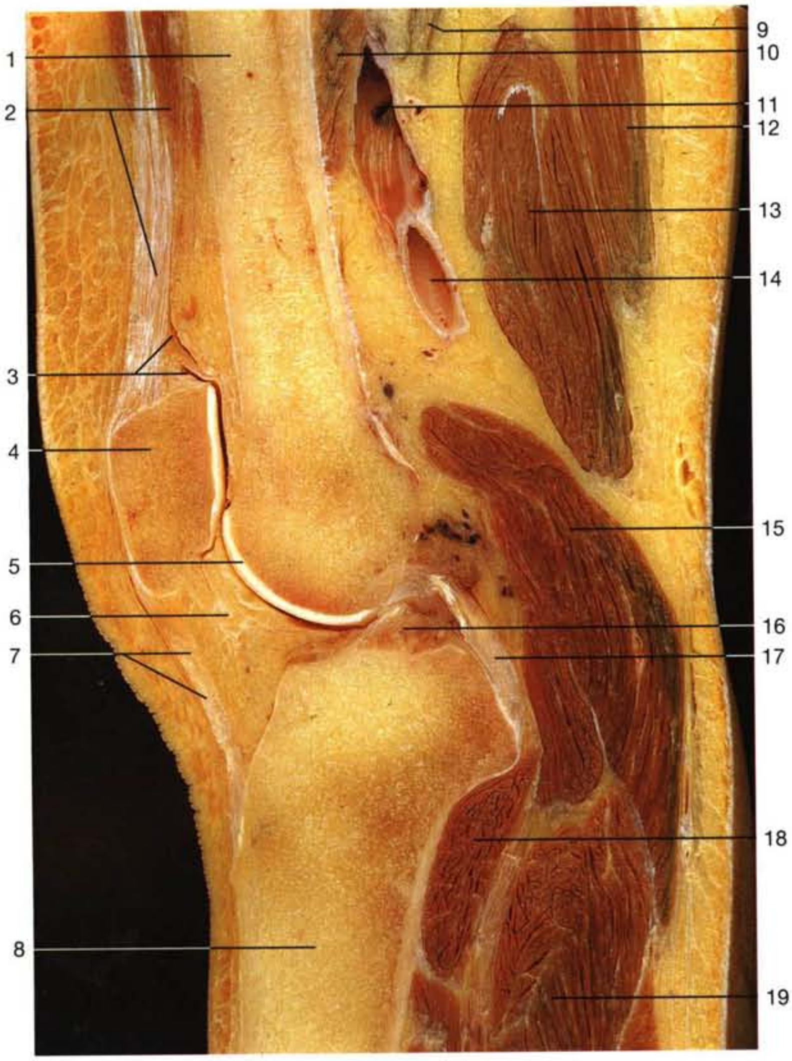

Sagittal section through the knee joint. Anterior surface to the left.

1 **Femur**
2 Quadriceps femoris muscle
3 Suprapatellar bursa and articular cavity
4 **Patella**
5 Patellar surface (articular cartilage)
6 Infrapatellar fat pad
7 Patellar ligament
8 **Tibia**
9 Tibial nerve
10 Adductor magnus muscle
11 Popliteal vein
12 Semitendinosus muscle
13 Semimembranosus muscle
14 **Popliteal artery**
15 Gastrocnemius muscle
16 Anterior cruciate ligament
17 Posterior cruciate ligament
18 Popliteus muscle
19 Soleus muscle
20 Deep flexor muscles of leg
21 Calcaneal tendon
22 Epiphyseal line of tibia
23 Calcaneus
24 **Talocrural joint**
25 Talus
26 Lateral meniscus
27 Epiphysial line

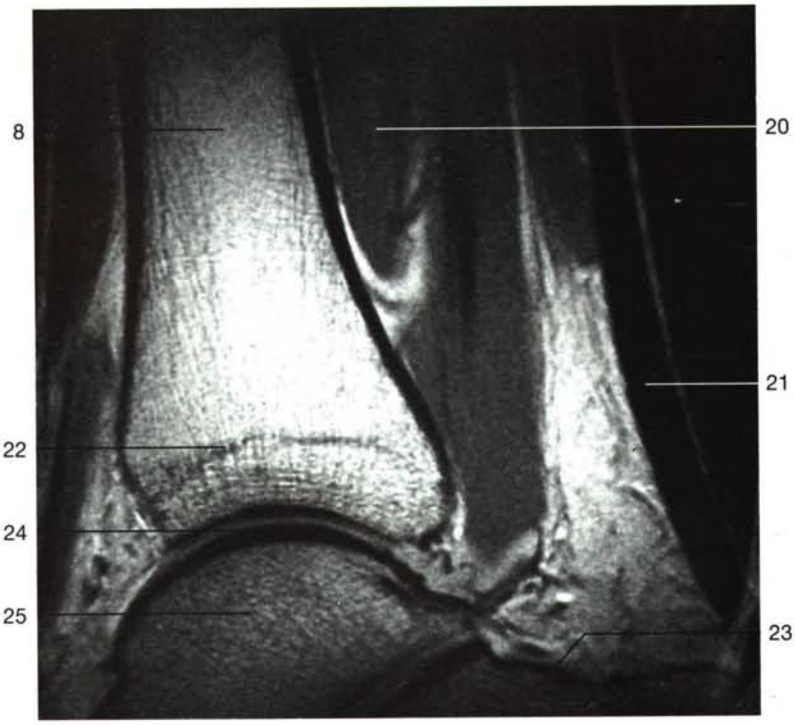

Talocrural joint. Sagittal section; anterior part to the left. (MR-Scan.)

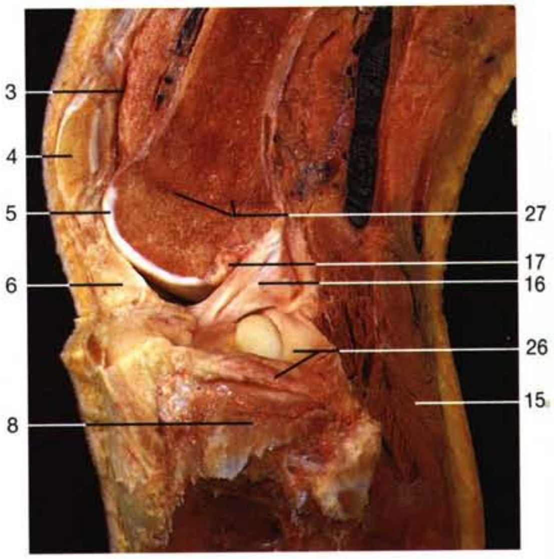

Left knee joint. Anterior cruciate ligament (lateral aspect).

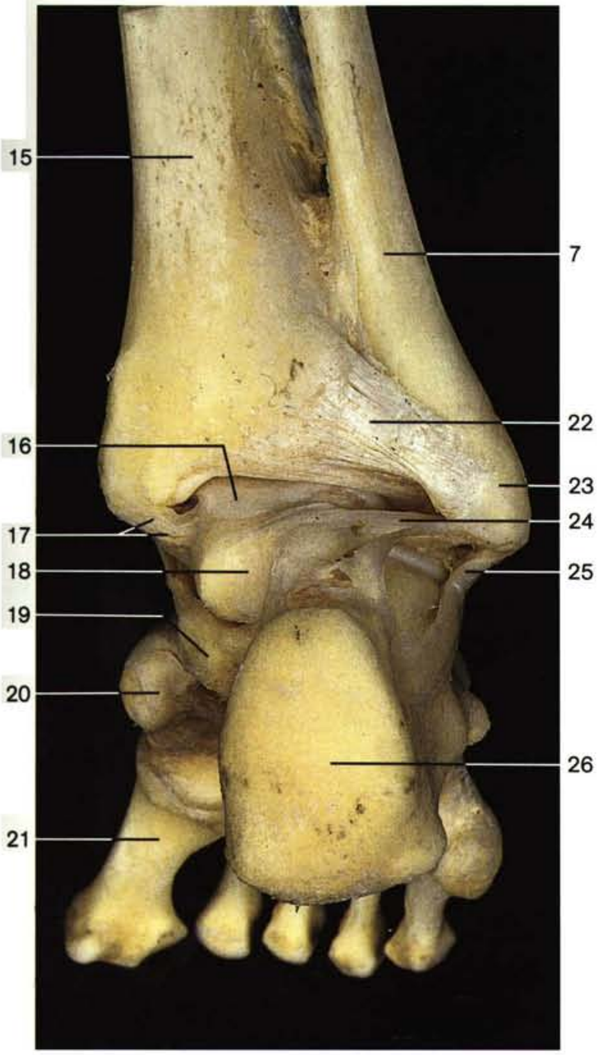

Ligaments of talocrural joint, right leg (posterior aspect).

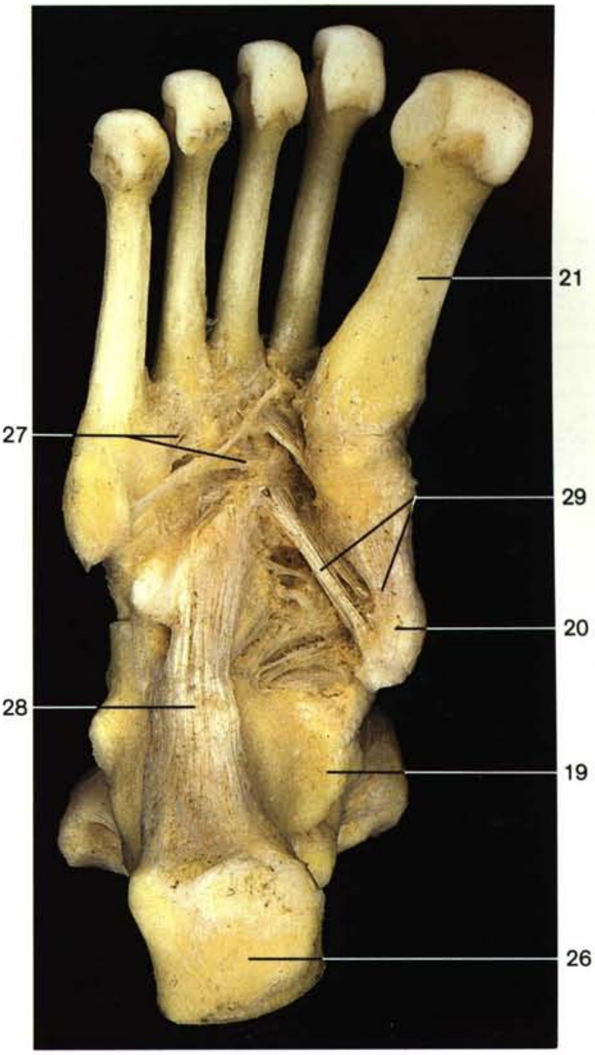

Deep ligaments of the foot, right foot (plantar aspect). The toes have been removed.

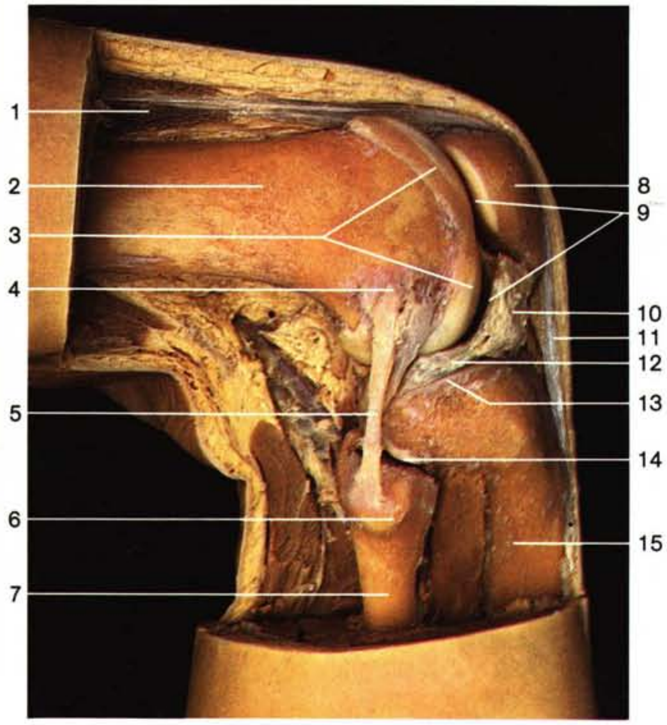

Right knee joint and tibiofibular joint with ligaments. Note the position of the lateral meniscus.

1 Quadriceps femoris muscle
2 Femur
3 Patellar surface
4 Lateral epicondyle of femur
5 **Fibular collateral ligament**
6 Head of fibula
7 **Fibula**
8 **Patella**
9 Articular cavity of knee joint
10 Infrapatellar fat pad
11 **Patellar ligament**
12 **Lateral meniscus of knee joint**
13 Lateral condyle of tibia (superior articular facet)
14 **Tibiofibular joint**
15 Tibia
16 Trochlea of talus (superior surface)
17 Deltoid ligament of ankle (posterior tibiotalar part)
18 Talus
19 **Sustentaculum tali**
20 Navicular bone
21 1st metatarsal bone
22 **Posterior tibiofibular ligament**
23 Lateral malleolus
24 Posterior talofibular ligament
25 Calcaneofibular ligament
26 Calcaneal tuberosity
27 Plantar tarsometatarsal ligaments
28 **Long plantar ligament**
29 Plantar cuneonavicular ligaments

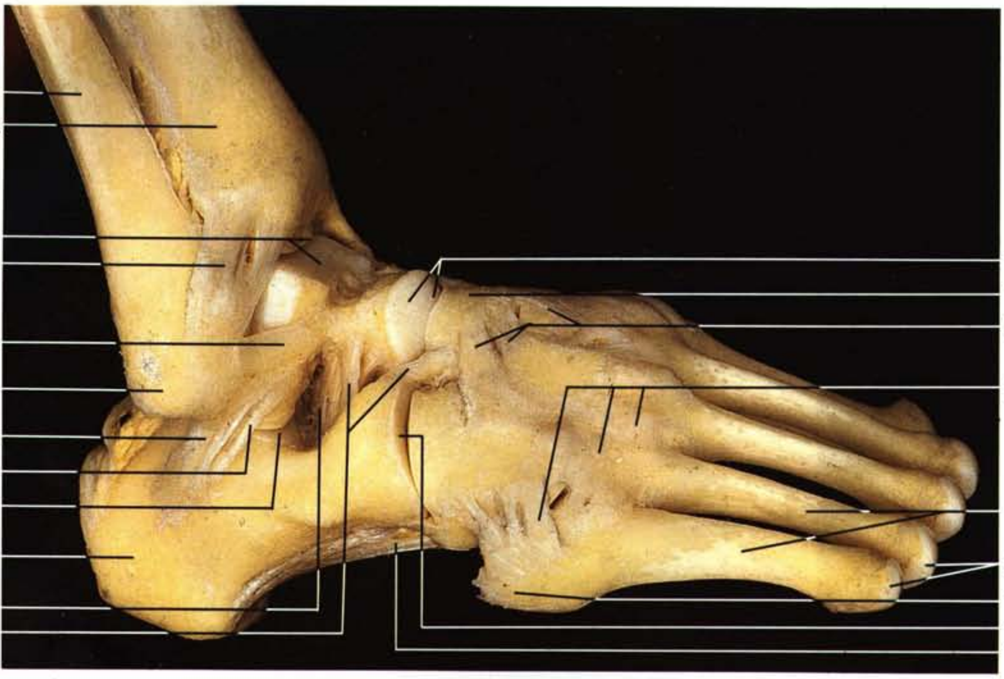

Ligaments of right foot (lateral aspect).

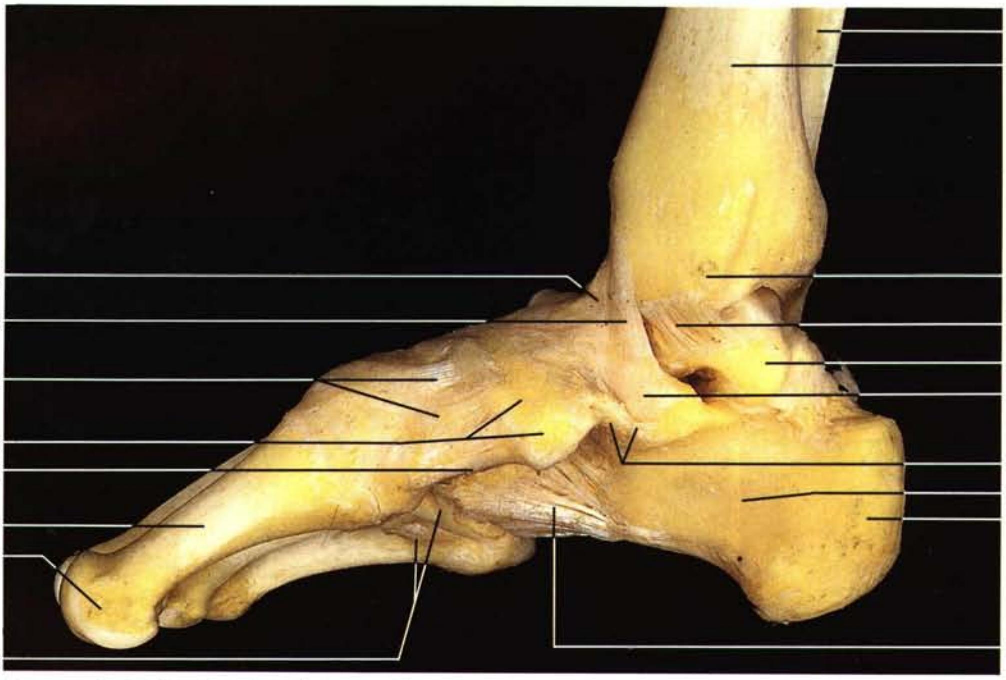

Ligaments of right foot (medial aspect).

1 Fibula
2 Tibia
3 Trochlea of talus and **talocrural joint**
4 Anterior tibiofibular ligament
5 Anterior talofibular ligament
6 Lateral malleolus
7 Calcaneofibular ligament
8 Lateral talocalcaneal ligament
9 **Subtalar joint**
10 Tuber calcanei
11 Interosseous talocalcaneal ligament
12 Bifurcate ligament
13 Long plantar ligament
14 **Calcaneocuboid joint**
15 Tuberosity of 5th metatarsal bone
16 Dorsal tarsometatarsal ligaments
17 Metatarsal bones
18 Head of talus and **talocalcaneonavicular joint**
19 Navicular bone
20 Dorsal cuneonavicular ligaments
21 Heads of metatarsal bones
22 Medial or deltoid ligament of ankle (tibionavicular part)
23 Medial or deltoid ligament of ankle (tibiocalcaneal part)
24 Dorsal cuneonavicular ligaments
25 Navicular bone
26 Plantar cuneonavicular ligament
27 1st metatarsal bone
28 Head of 1st metatarsal bone
29 Plantar tarsometatarsal ligaments
30 Plantar calcaneonavicular ligament
31 Sustentaculum tali
32 Calcaneus
33 Medial malleolus
34 Medial or deltoid ligament of ankle (posterior part)
35 Talus

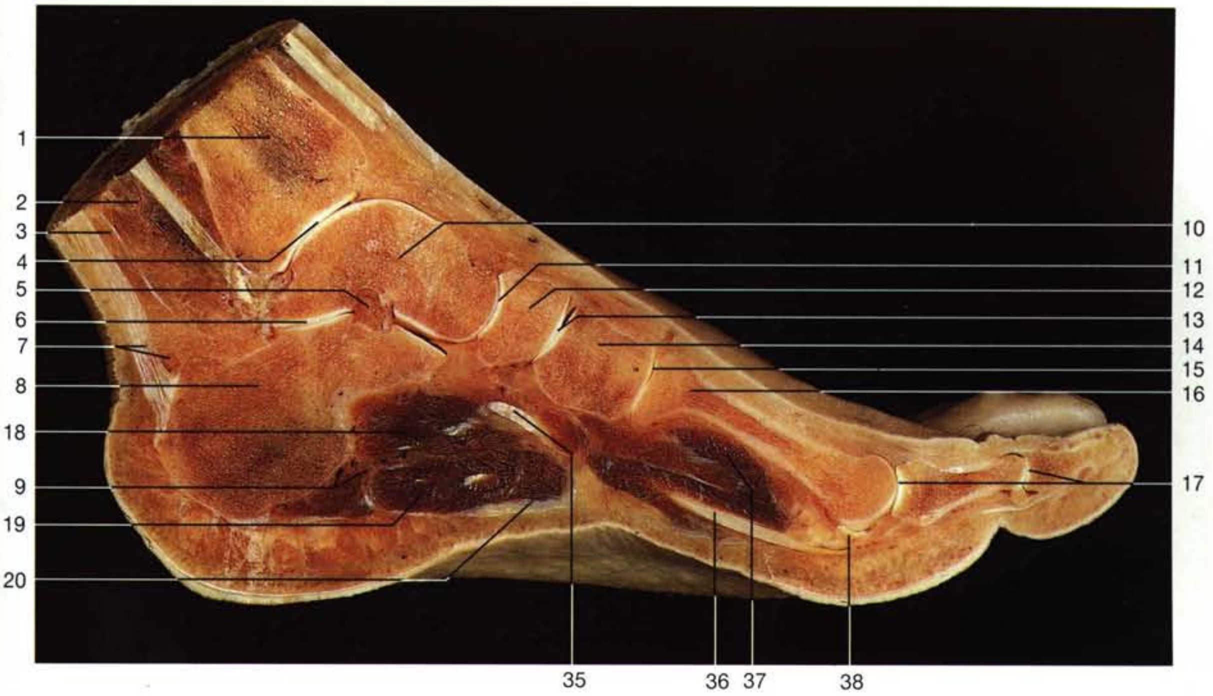

Longitudinal section through the foot at the level of first phalanx.

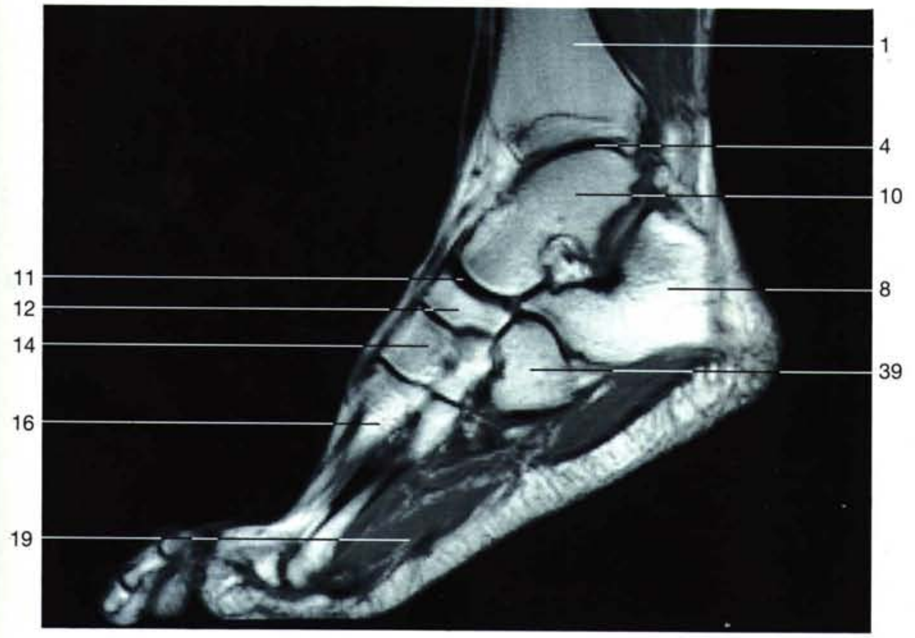

Sagittal section through the foot
(MR-Scan after A. Heuck, G. Luttke and J. W. Rohen, 1994).

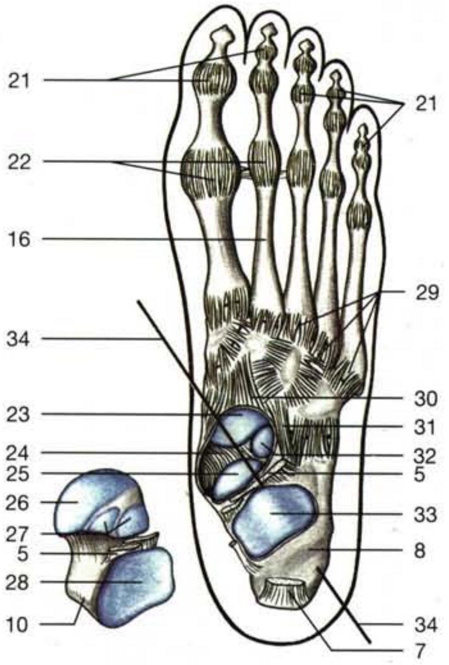

Talocalcaneonavicular joint.
The talus has been rotated to show the articular surfaces of the joint.

1 Tibia
2 Deep flexor muscles
3 Superficial flexor muscles
4 **Talocrural joint**
5 Interosseous talocalcaneal ligament
6 **Subtalar joint**
7 Calcaneal or Achilles tendon and bursa
8 Calcaneus
9 Vessels and nerves of foot
10 Talus
11 **Talocalcaneonavicular joint**
12 Navicular bone
13 **Cuneonavicular joint**
14 Intermediate cuneiform bone
15 **Tarsometatarsal joints**
16 Metatasal bones
17 **Metatarsophalangeal** and **interphalangeal joints**
18 Quadratus plantae muscle with flexor tendons
19 Flexor digitorum brevis muslce
20 Plantar aponeurosis
21 Articular capsules of interphalangeal joints
22 Articular capsules of metatarsophalangeal joints
23 Articular surface of navicular bone
24 Plantar calcaneonavicular ligament
25 Middle talar articular surface of calcaneus
26 Navicular articular surface of talus
27 Anterior and middle calcaneal surfaces of talus
28 Posterior calcaneal surface of talus
29 Dorsal tarsometatarsal ligaments
30 Talonavicular ligament
31 Bifurcate ligament
32 Anterior talar articular surface of calcaneus
33 Posterior talar articular surface of calcaneus
34 Axis for inversion and eversion
35 Tendon of tibialis posterior muscle
36 Tendon of flexor hallucis longus muscle
37 Flexor hallucis brevis muscle
38 Sesamoid bone
39 Cuboid bone

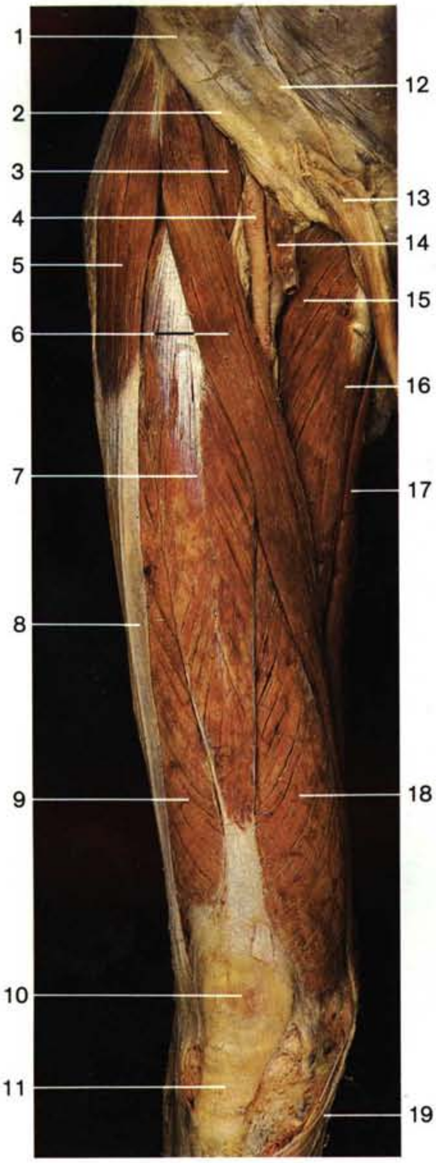

Extensor and adductor muscles of thigh, right thigh (anterior aspect).

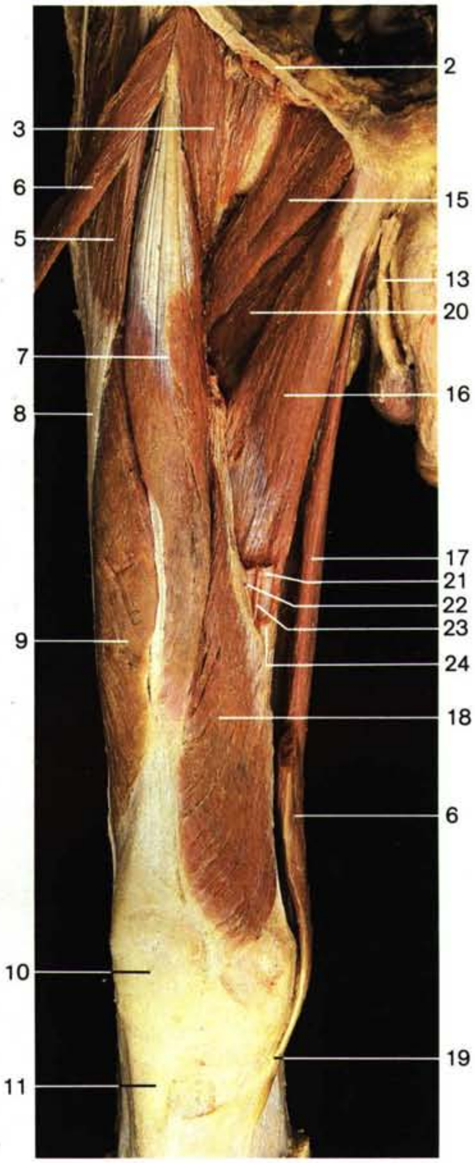

Quadriceps muscle and superficial layer of adductor muscles, right thigh (anterior aspect). The sartorius muscle has been divided.

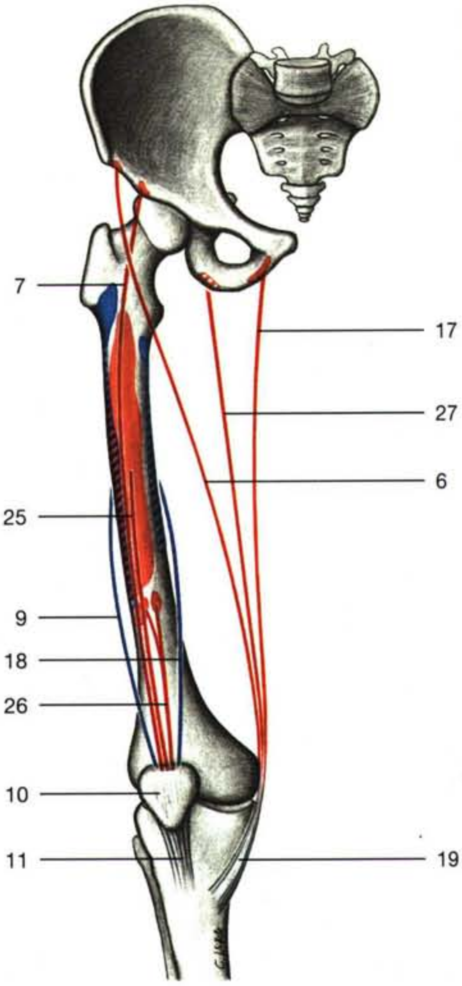

Course of extensor muscles of thigh and muscles inserting with common tendon on tibia.

1 Anterior superior iliac spine
2 Inguinal ligament
3 Iliopsoas muscle
4 Femoral artery
5 Tensor fasciae latae muscle
6 Sartorius muscle
7 **Rectus femoris muscle**
8 Iliotibial tract
9 Vastus lateralis muscle
10 Patella
11 Patellar ligament
12 Aponeurosis of external abdominal oblique muscle
13 Spermatic cord
14 Femoral vein

15 **Pectineus muscle**
16 **Adductor longus muscle**
17 Gracilis muscle
18 Vastus medialis muscle
19 Common tendon of sartorius, gracilis, and semitendinosus muscles (pes anserinus)
20 **Adductor brevis muscle**
21 Femoral artery
22 Femoral vein } entering the adductor canal
23 Saphenous nerve
24 Fascia of adductor canal
25 Vastus intermedius muscle
26 Articularis genus muscle
27 Semitendinosus muscle

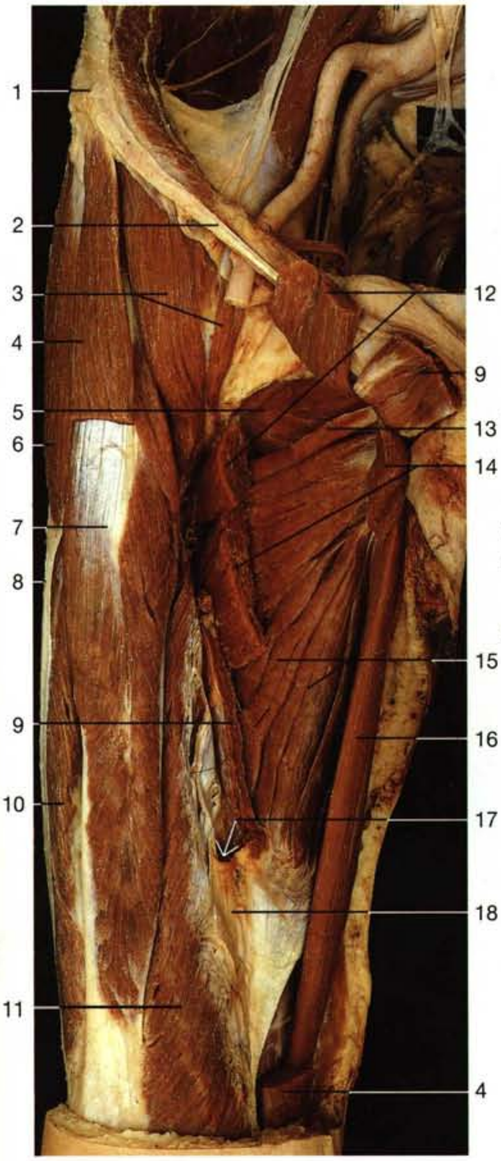

Deep layer of adductor muscles. Adductor magnus muscle (anterior aspect). Pectineus, adductor longus and brevis muscles have been divided.

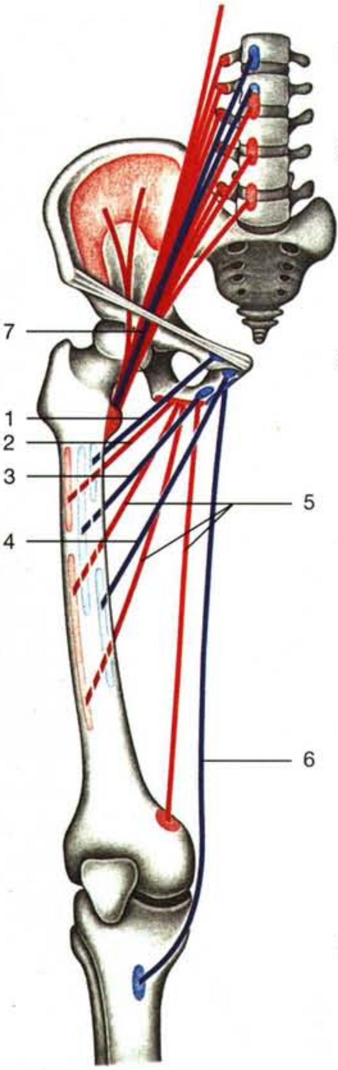

Course of adductor muscles. (Schematic drawing.)

1 Pectineus muscle (blue)
2 Adductor minimus muscle (red)
3 Adductor brevis muscle (blue)
4 Adductor longus muscle (blue)
5 Adductor magnus muscle (red)
6 Gracilis muscle (blue)
7 Iliopsoas muscle (red/blue)

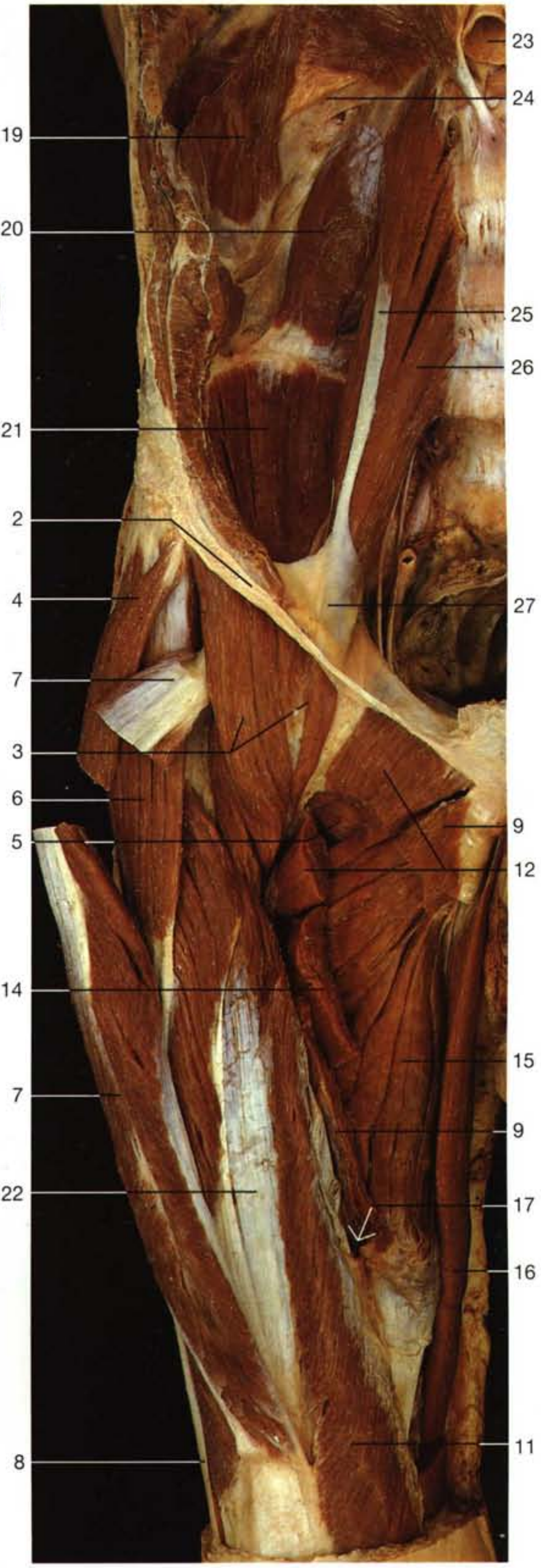

Iliopsoas and adductor muscles; deepest layer (anterior aspect). Pectineus, adductor longus and brevis, rectus femoris muscles have been divided.

1 Anterior superior iliac spine
2 Inguinal ligament
3 **Iliopsoas muscle**
4 Sartorius muscle
5 Obturator externus muscle
6 Tensor fasciae latae muscle
7 Rectus femoris muscle
8 Iliotibial tract
9 Adductor longus muscle (divided)
10 Vastus lateralis muscle
11 Vastus medialis muscle
12 Pectineus muscle (divided)
13 Adductor minimus muscle
14 Adductor brevis muscle (cut)
15 **Adductor magnus**
16 Gracilis muscle
17 Adductor hiatus
18 Vastoadductor membrane
19 Diaphragm
20 Quadratus lumborum muscle
21 **Iliacus muscle**
22 Vastus intermedius muscle
23 Aorta in aortic hiatus
24 Twelfth rib
25 **Psoas minor muscle**
26 **Psoas major muscle**
27 Iliopectineal arch

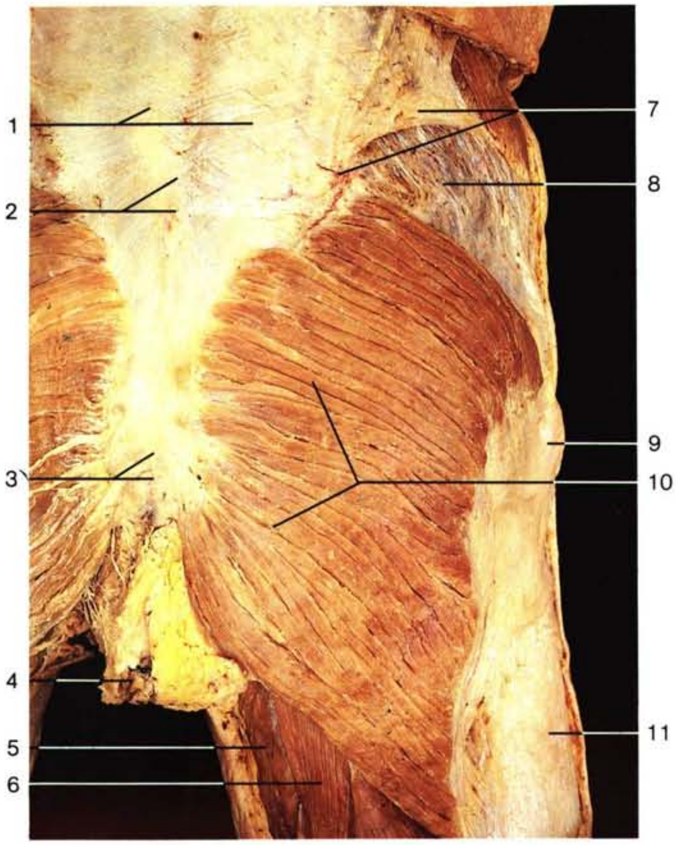
1
2
3
4
5
6
7
8
9
10
11

Gluteal muscles, superficial layer (posterior aspect).

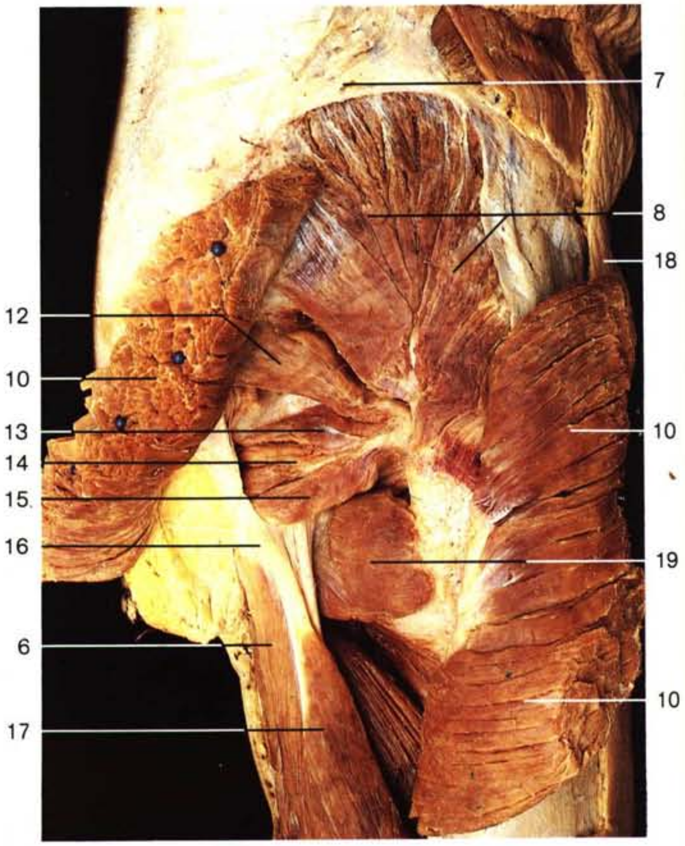
7
8
18
12
10
13
14
15
16
6
17
10
19
10

Gluteal muscles, deeper layer (posterior aspect).

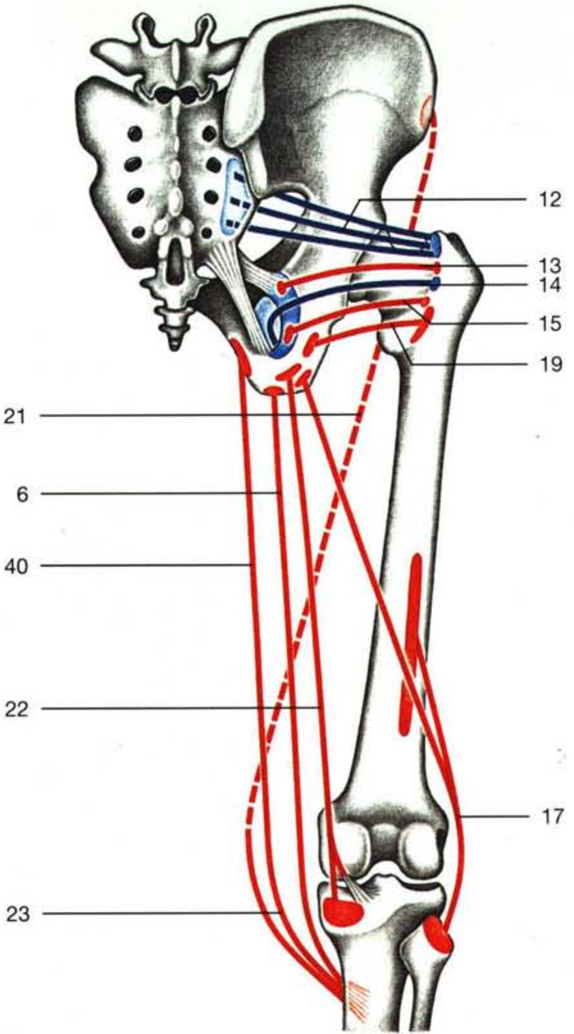
12
13
14
15
19
21
6
40
22
17
23

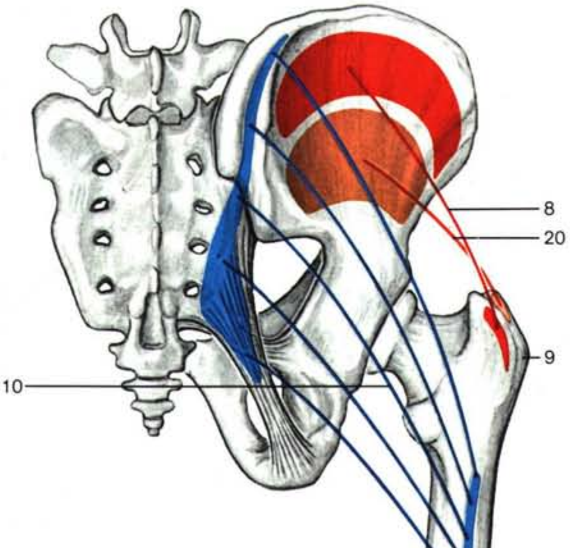

Course of gluteal muscles
(posterior aspect; schematic drawing).

◁

Course of gluteal muscles (deeper layer) **and of ischiocrural muscles** (posterior aspect).
Sartorius muscle is indicated by dotted line (schematic drawing).

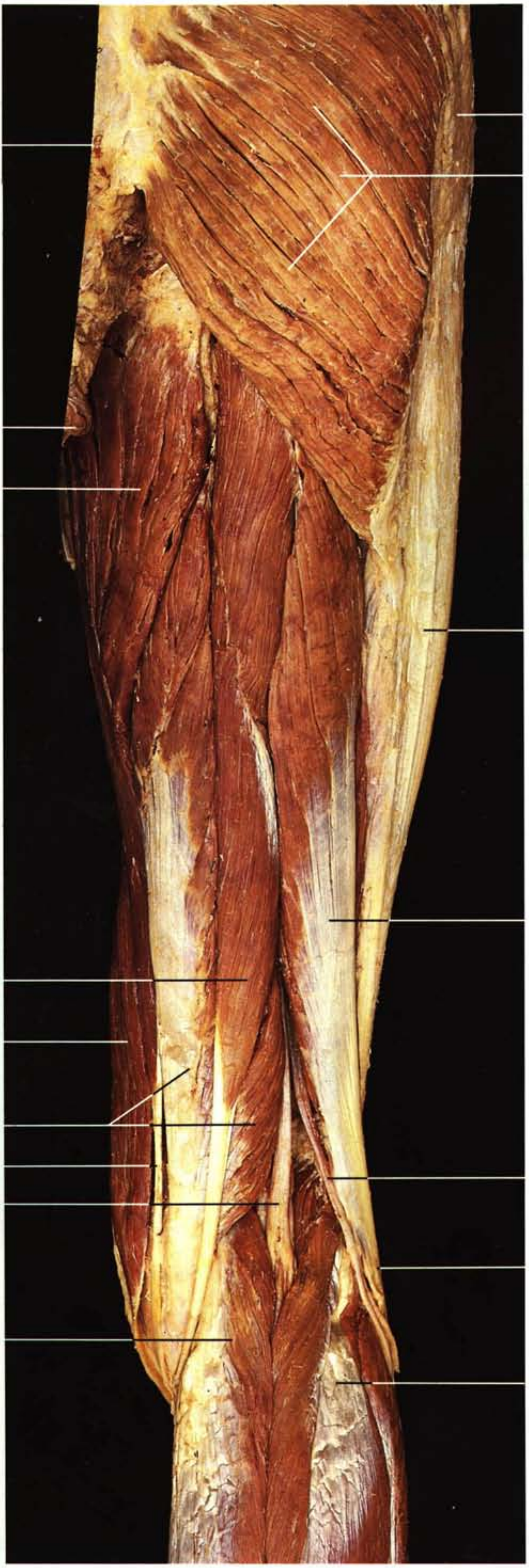

Flexors of the right thigh, superficial layer (dorsal aspect).

1 Thoracolumbar fascia
2 Spinous processes of lumbar vertebrae
3 Coccyx
4 Anus
5 Adductor magnus muscle
6 Semitendinosus muscle
7 Iliac crest
8 **Gluteus medius muscle**
9 Greater trochanter
10 **Gluteus maximus muscle**
11 Iliotibial tract
12 Piriformis muscle
13 Superior gemellus muscle
14 Obturator internus muscle
15 Inferior gemellus muscle
16 Ischial tuberosity
17 **Biceps femoris muscle**
18 Tensor fasciae latae muscle
19 Quadratus femoris muscle
20 **Gluteus minimus muscle**
21 Sartorius muscle
22 Semimembranosus muscle
23 Tendon of gracilis muscle
24 Tibial nerve
25 Medial head of gastrocnemius muscle
26 Common peroneal nerve
27 Tendon of biceps femoris muscle
28 Lateral head of gastrocnemius muscle
29 Rectus femoris muscle
30 Vastus medialis muscle
31 Vastus intermedius muscle
32 Vastus lateralis muscle
33 Sciatic nerve
34 Gluteus maximus muscle (insertion)
35 Great saphenous vein
36 Femoral artery
37 Femoral vein
38 Adductor longus muscle
39 Femur
40 Gracilis muscle
41 Septum between semitendinosus
and semimembranosus muscles

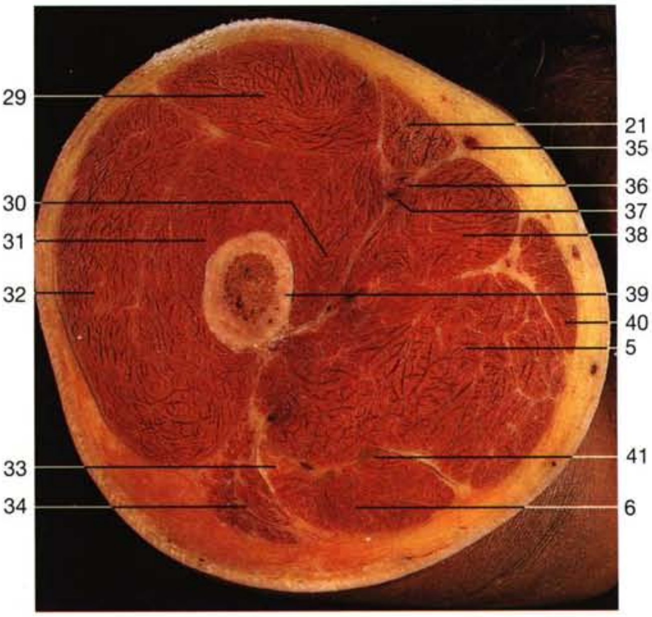

Cross section of right thigh (inferior aspect). Anterior side on top.

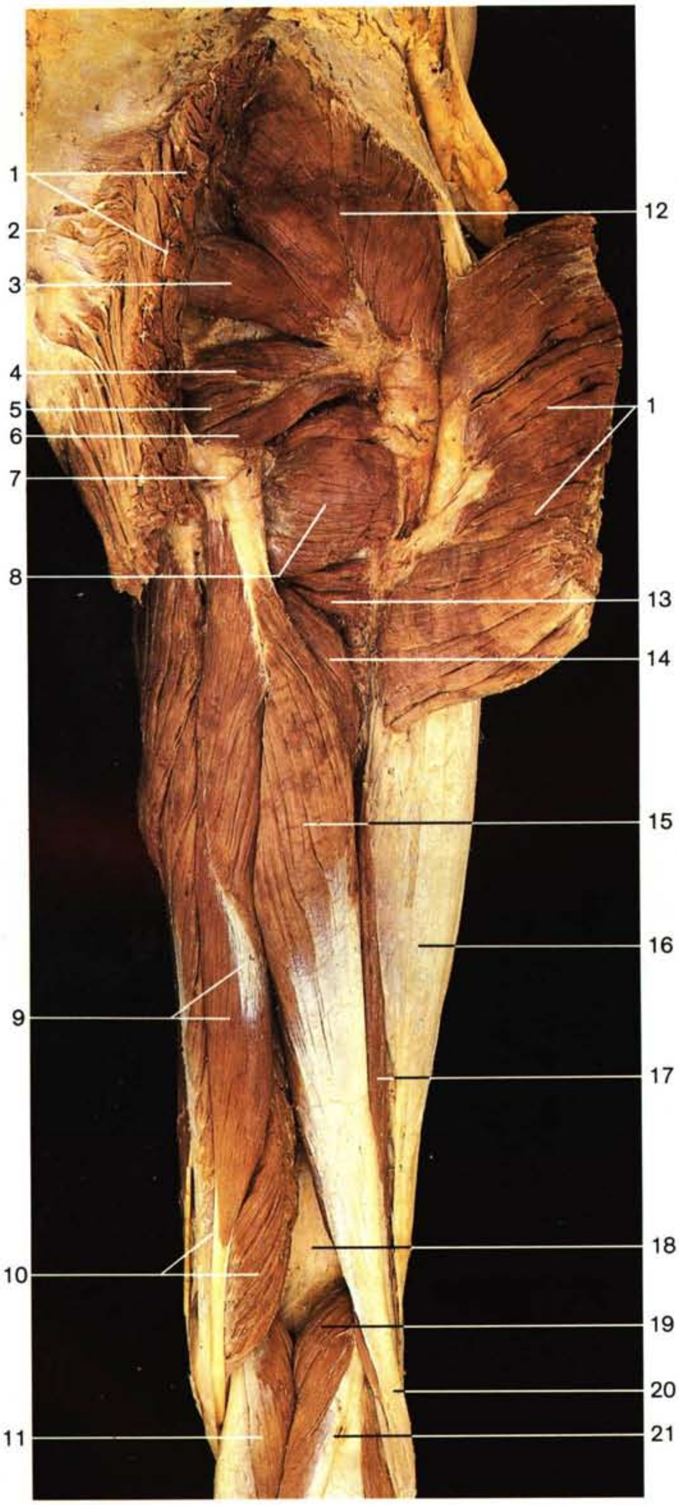

Dorsal muscles of right thigh (posterior aspect).
The gluteus maximus muscle has been cut and reflected.

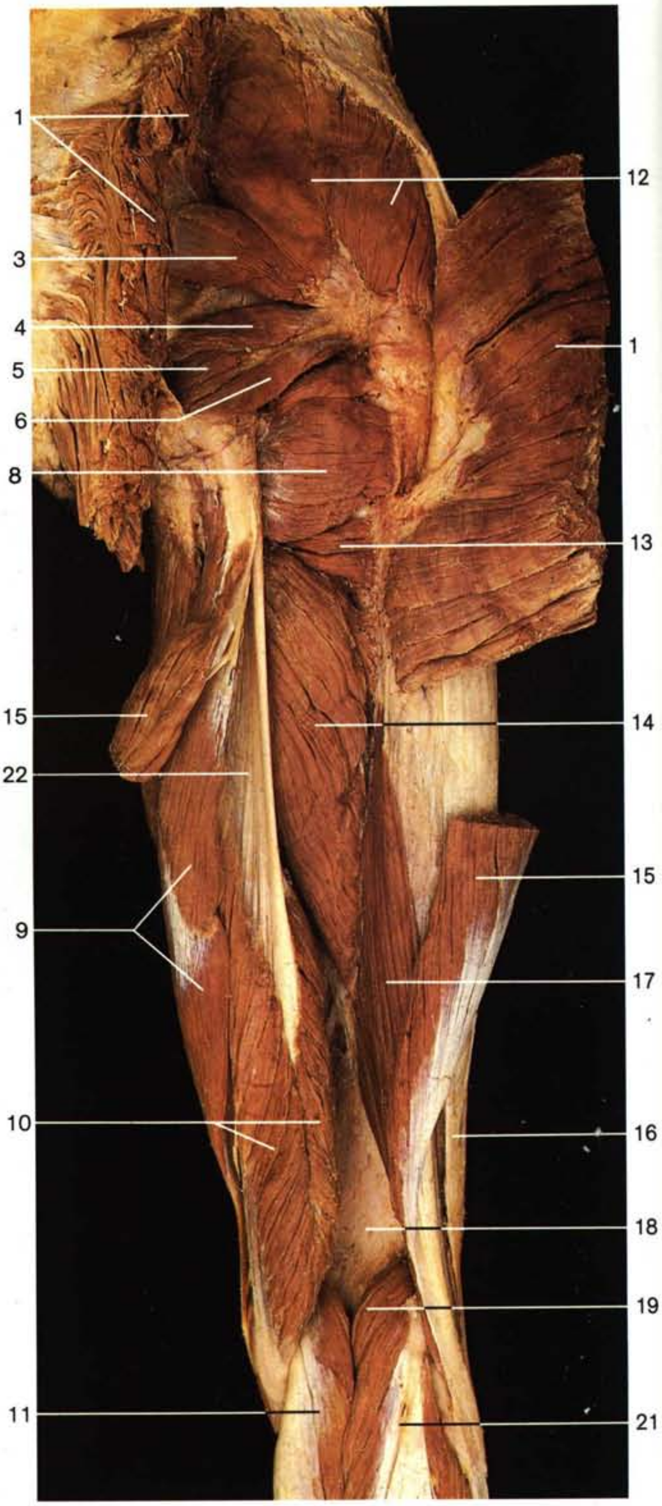

Dorsal muscles of right thigh (posterior aspect).
The gluteus maximus muscle and the long head of biceps femoris muscle have been divided and displaced.

1 Gluteus maximus muscle (divided)
2 Position of coccyx
3 Piriformis muscle
4 Superior gemellus muscle
5 Obturator internus muscle
6 Inferior gemellus muscle
7 Ischial tuberosity
8 Quadratus femoris muscle
9 Semitendinosus muscle with intermediate tendon
10 Semimembranosus muscle
11 Medial head of gastrocnemius muscle
12 Gluteus medius muscle
13 Adductor minimus muscle
14 Adductor magnus muscle
15 Long head of biceps femoris muscle
16 Iliotibial tract
17 Short head of biceps femoris muscle
18 Popliteal surface of femur
19 Plantaris muscle
20 Tendon of biceps femoris muscle
21 Lateral head of gastrocnemius muscle
22 Membranous part of semimembranosus muscle

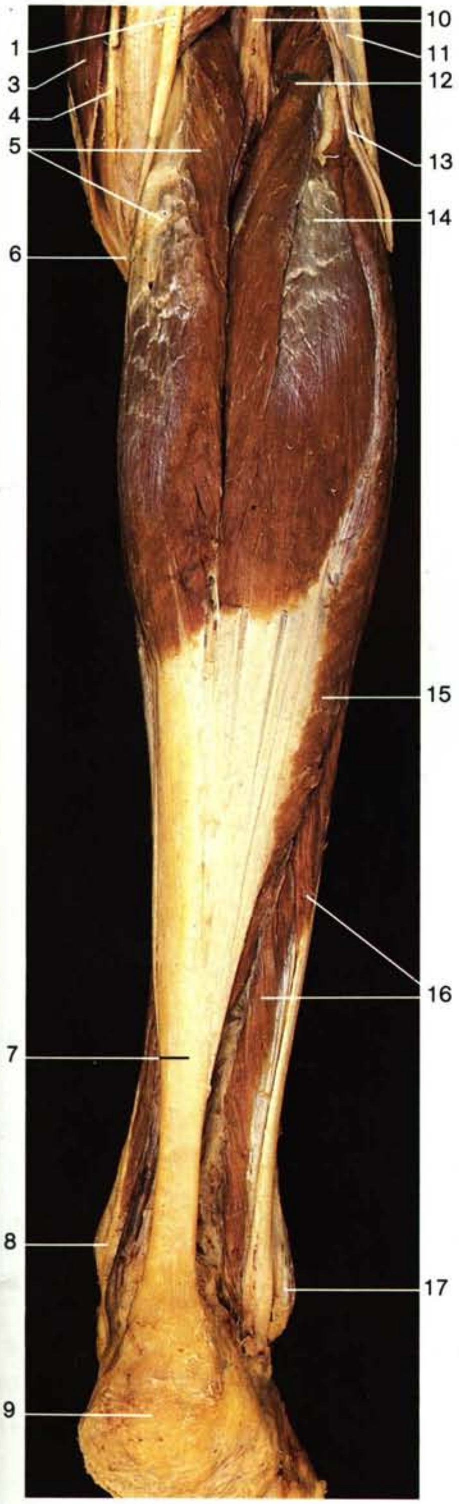

Flexor muscles of right leg (posterior aspect).

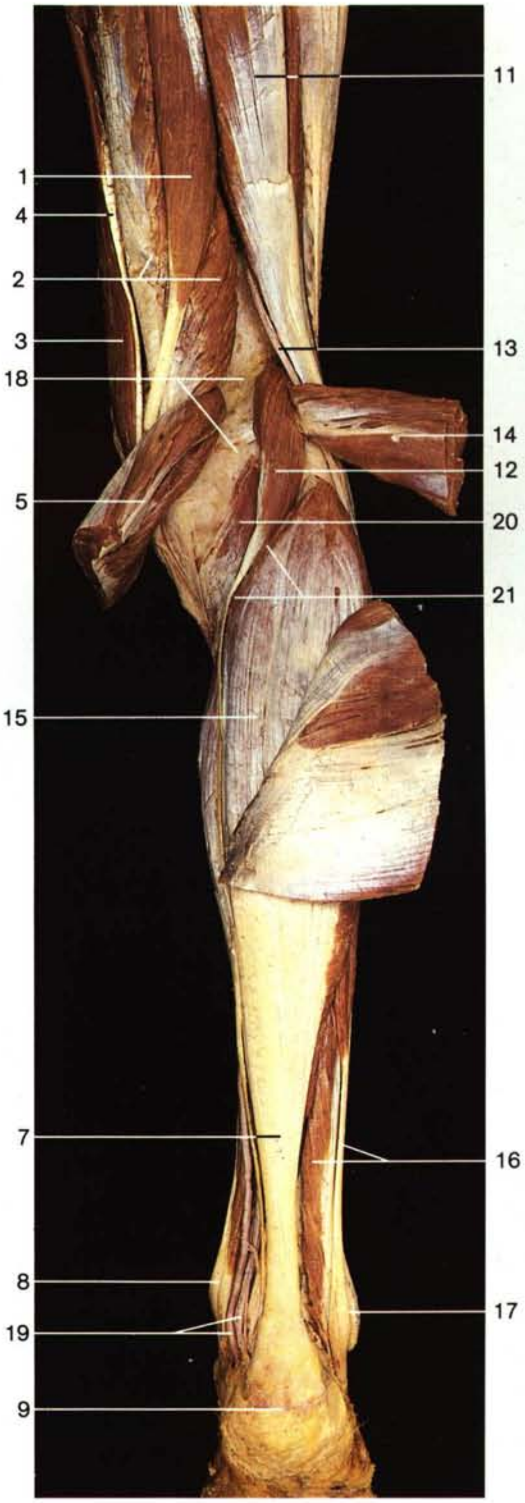

Flexor muscles of right leg (posterior aspect). Both heads of the gastrocnemius muscle have been cut and reflected.

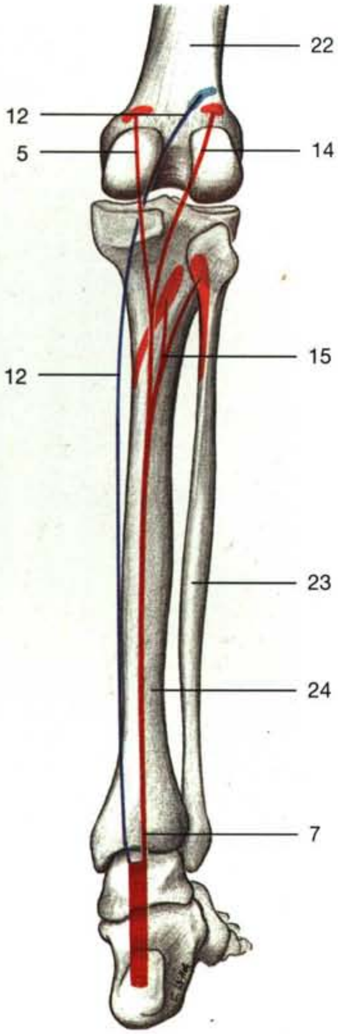

Flexor muscles of the leg (right side).

1 Semitendinosus muscle
2 Semimembranosus muscle
3 Sartorius muscle
4 Tendon of gracilis muscle
5 **Medial head of gastrocnemius muscle**
6 Common tendon of gracilis, sartorius, and semitendinosus muscles
7 Calcaneal or Achilles tendon
8 Medial malleolus
9 Calcaneal tuberosity
10 Tibial nerve
11 Biceps femoris muscle
12 Plantaris muscle
13 Common peroneal nerve
14 **Lateral head of gastrocnemius muscle**
15 **Soleus muscle**
16 Peroneus longus and brevis muscles
17 Lateral malleolus
18 Popliteal fossa
19 Tibial nerve and posterior tibial artery
20 Popliteus muscle
21 Tendinous arch of soleus muscle
22 Femur
23 Fibula
24 Tibia

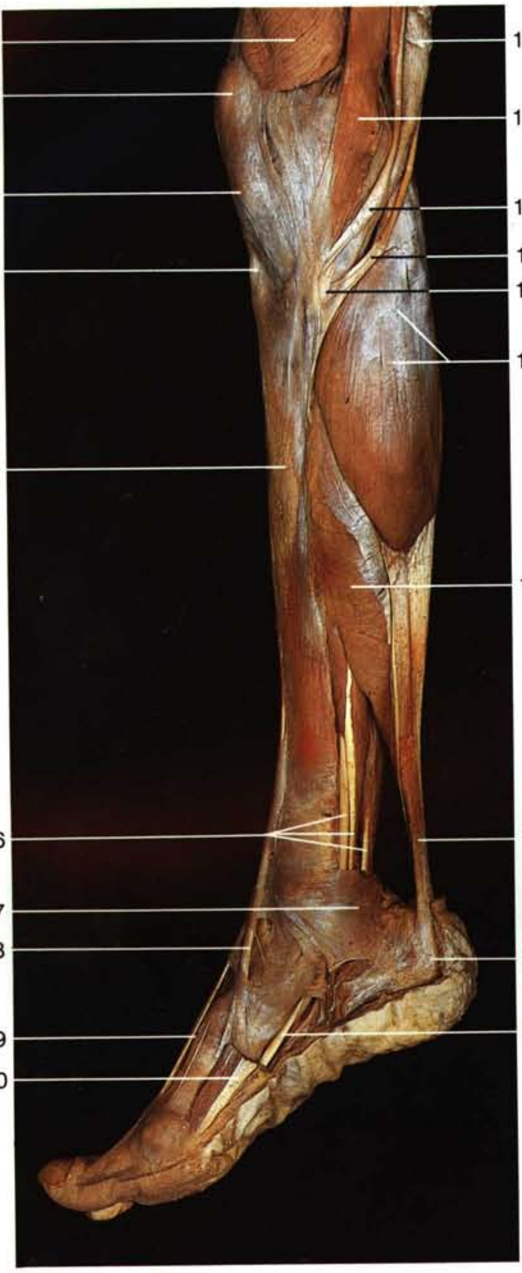

Muscles of right leg and foot (medial aspect).

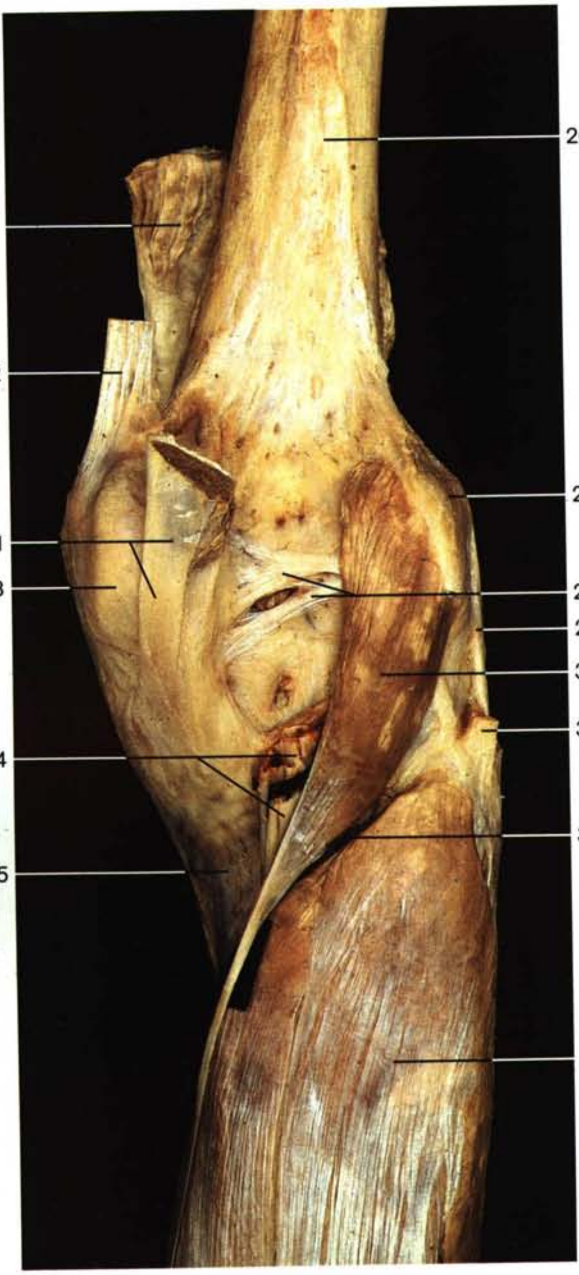

Popliteal region with plantaris and soleus, right side (dorsal aspect). Notice the insertion of the tendon of semimembranosus.

1 Vastus medialis muscle
2 Patella
3 Patellar ligament
4 Tibial tuberosity
5 Tibia
6 Tendons of deep flexor muscles (from anterior to posterior: 1. tibialis posterior; 2. flexor digitorum longus; 3. flexor hallucis longus muscles)
7 Flexor retinaculum
8 Tendon of tibialis anterior muscle
9 Tendon of extensor hallucis longus muscle
10 Abductor hallucis muscle
11 Semimembranosus muscle
12 Sartorius muscle
13 Tendon gracilis muscle
14 Tendon of semitendinosus muscle
15 Common tendon of gracilis, semitendinosus and sartorius muscles
16 **Medial head of gastrocnemius muscle**
17 **Soleus muscle**
18 Calcaneal or Achilles tendon
19 Calcaneus muscle
20 Tendon of flexor hallucis longus muscle
21 Quadriceps femoris muscle (divided)
22 Tendon of adductor magnus muscle (divided)
23 Medial condyle of femur
24 Popliteal artery and vein, tibial nerve
25 Tibia
26 Femur
27 Lateral epicondyle of femur
28 Oblique popliteal ligament
29 Lateral (fibular) collateral ligament
30 **Plantaris muscle**
31 Tendon of biceps femoris muscle (divided)
32 Tendinous arch of soleus muscle

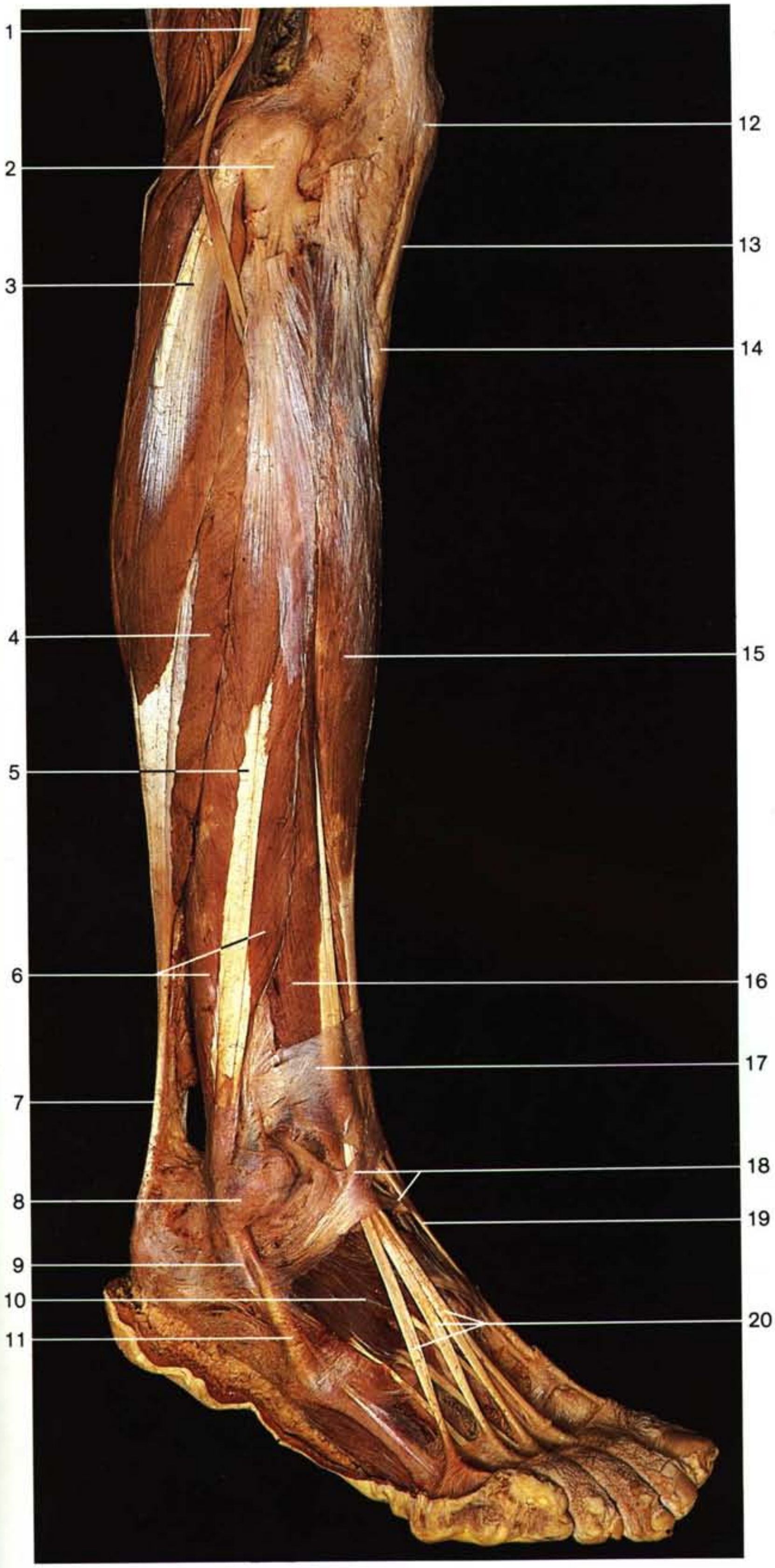

Muscles of right leg and foot (lateral aspect).

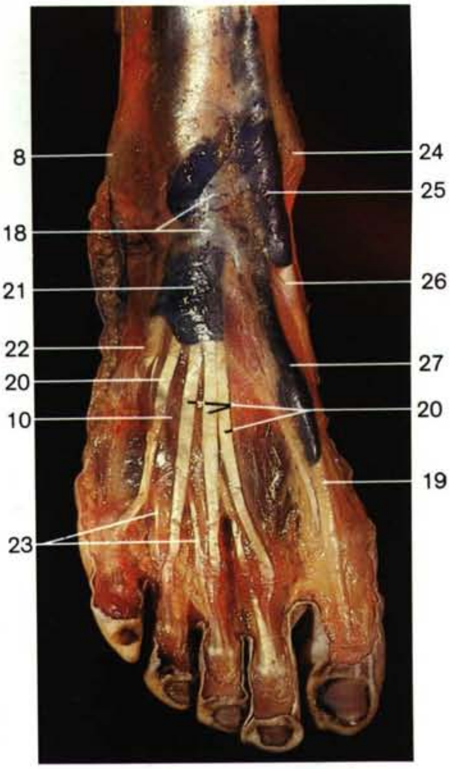

Right foot with synovial sheaths of extensor muscles (dorsal aspect). The synovial sheaths have been injected with blue solution.

1 Common peroneal nerve
2 Head of fibula
3 Lateral head of gastrocnemius muscle
4 Soleus muscle
5 Peroneus longus muscle
6 Peroneus brevis muscle
7 Calcaneal or Achilles tendon
8 Lateral malleolus muscle
9 Tendon of peroneus longus muscle
10 Extensor digitorum brevis muscle
11 Tendon of peroneus brevis muscle
12 Patella
13 Patellar ligament
14 Tuberosity of tibia
15 Tibialis anterior muscle
16 Extensor digitorum longus muscle
17 Superior extensor retinaculum
18 Inferior extensor retinaculum
19 Tendon of extensor hallucis longus muscle
20 Tendons of extensor digitorum longus muscle
21 Common synovial sheath of extensor digitorum longus muscle
22 Tendon of peroneus tertius muscle to the lateral margin of foot
23 Tendons of extensor digitorum brevis muscle
24 Medial malleolus
25 Synovial sheath of tendon of tibialis anterior muscle
26 Tendon of tibialis anterior muscle
27 Synovial sheath of tendon of extensor hallucis longus muscle

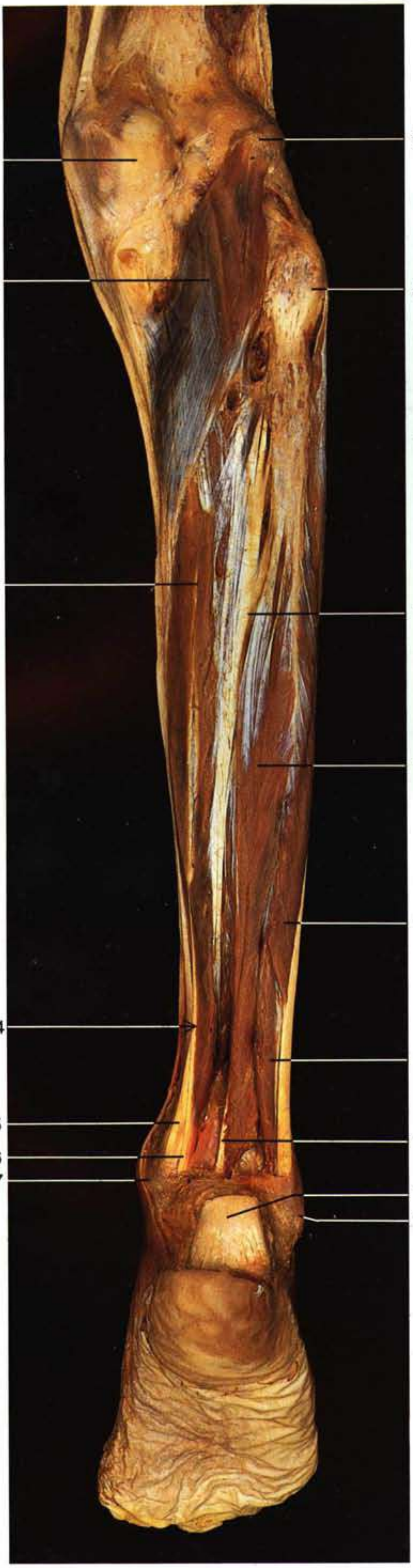

1 Medial condyle of femur
2 Popliteus muscle
3 Flexor digitorum longus muscle
4 Crosssing of tendons in leg
5 Tendon of tibialis posterior muscle
6 Tendon of flexor digitorum longus muscle
7 Medial malleolus
8 Lateral condyle of femur
9 Head of fibula
10 Tibialis posterior muscle
11 Flexor hallucis longus muscle
12 Peroneus longus muscle
13 Peroneus brevis muscle
14 Tendon of flexor hallucis longus muscle
15 Calcaneal tendon (divided)
16 Lateral malleolus

Deep flexor muscles of right leg (posterior aspect).

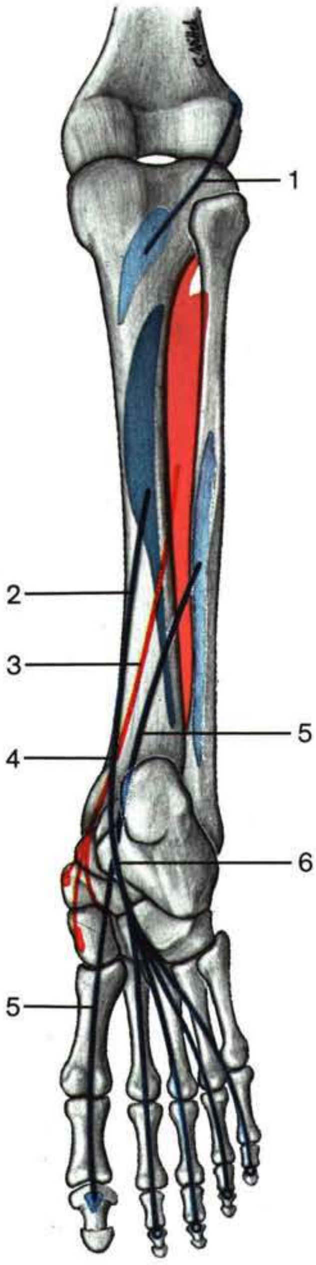

1 Popliteus muscle (blue)
2 Flexor digitorum longus muscle (blue)
3 Tibialis posterior muscle (red)
4 Crossing of tendons in leg
5 Flexor hallucis longus muscle (blue)
6 Crossing of tendons in sole

Course of deep flexor muscle of leg. (Schematic drawing.)

1 Medial condyle of femur
2 Tibia
3 Flexor digitorum longus muscle
4 Crossing of tendons in leg
5 Tendon of tibialis posterior muscle
6 Abductor hallucis muscle
7 Tendon of flexor hallucis longus muscle
8 Lateral condyle of femur
9 Head of fibula
10 Tibialis posterior muscle
11 Tendon of flexor digitorum longus muscle
12 Flexor retinaculum
13 Calcaneal tendon
14 Calcaneal tuberosity
15 Crossing of tendons in sole
16 Quadratus plantae muscle
17 Tendons of flexor digitorum longus muscle
18 Tendon of tibialis anterior muscle
19 Area of insertion of tibialis posterior muscle
20 Lumbrical muscles
21 Flexor hallucis longus muscle

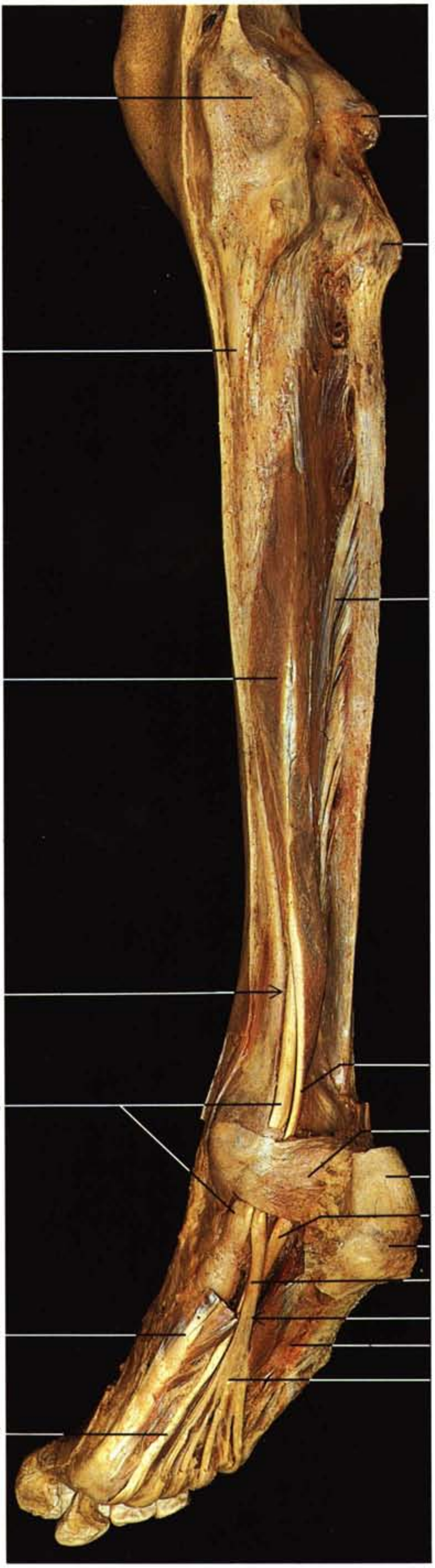

Deep flexor muscles of right leg and foot (posterior oblique medial aspect). Felxor digitorum brevis and flexor hallucis longus muscles have been removed.

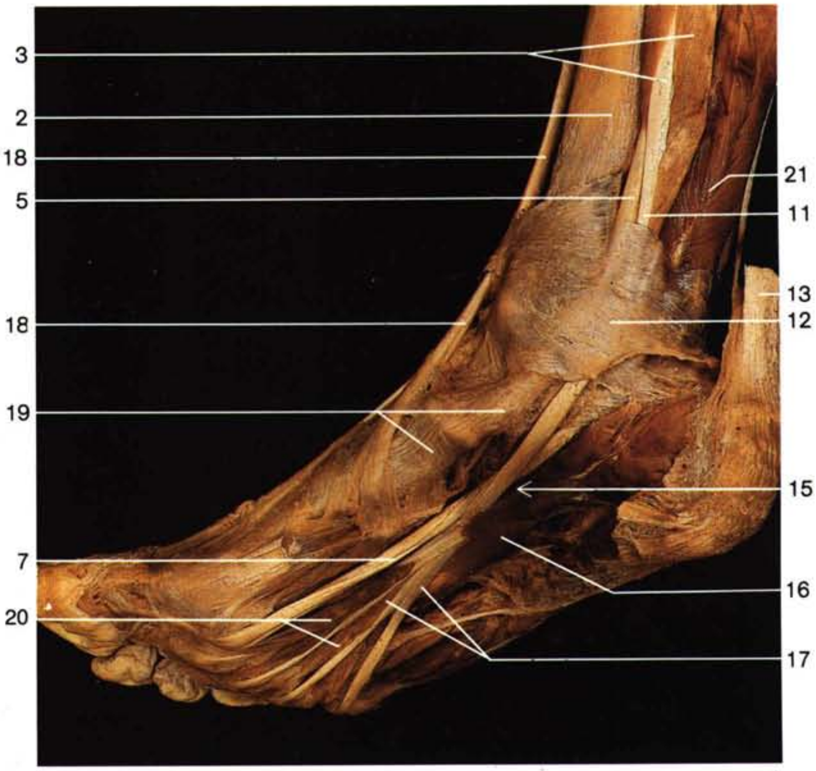

Sole of foot; tendons of long flexor muscles (oblique medial and inferior aspect).

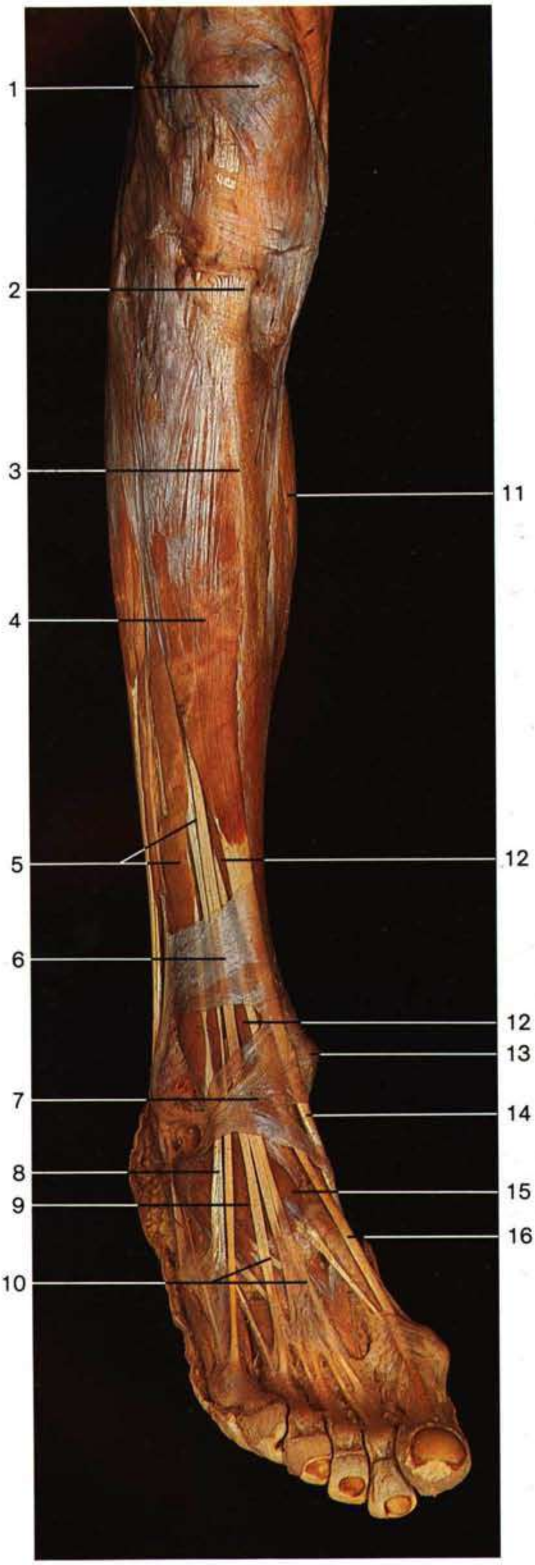

Extensor muscles of right leg and foot (oblique anterolateral aspect).

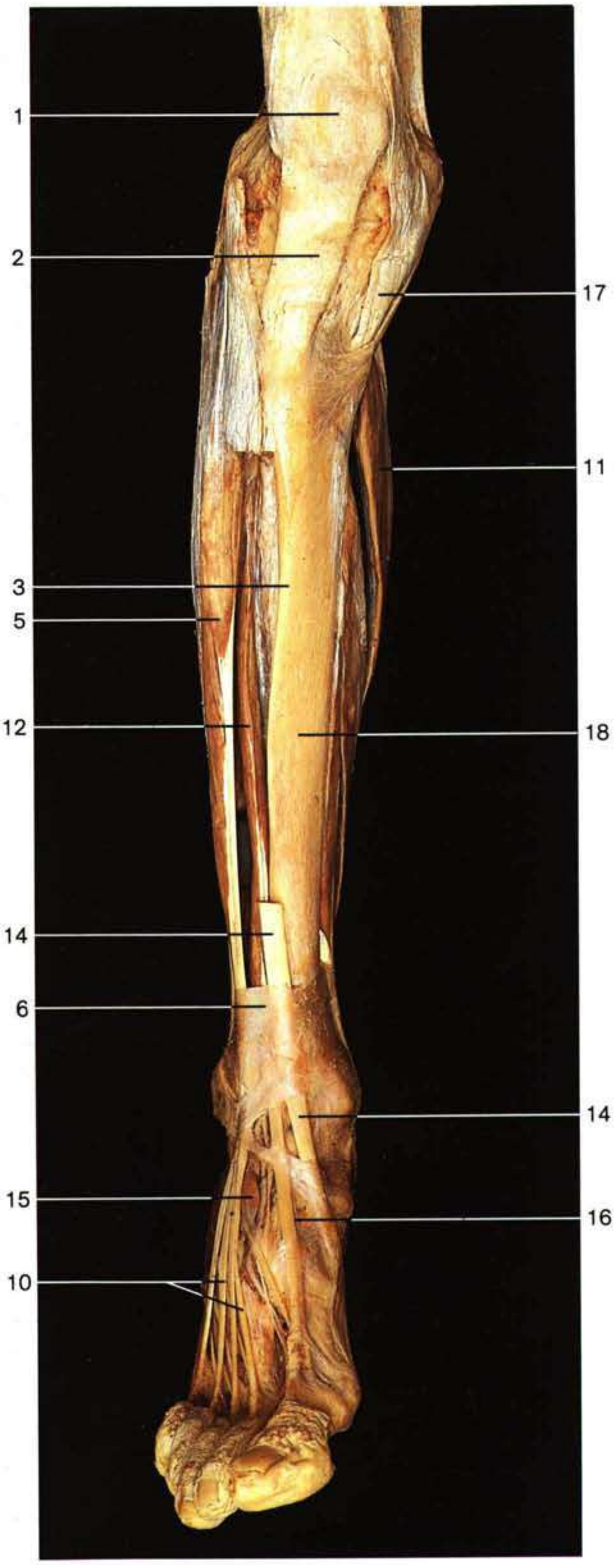

Extensors of the right leg and foot (anterior aspect). Part of the tibialis anterior muscle has been removed.

1 Patella
2 Patellar ligament
3 Anterior margin of tibia
4 Tibialis anterior muscle
5 Extensor digitorum longus muscle
6 Superior extensor retinaculum
7 Inferior extensor retinaculum
8 Tendon of peroneus tertius muscle
9 Extensor digitorum brevis muscle
10 Tendons of extensor digitorum longus muscle
11 Soleus
12 Extensor hallucis longus muscle
13 Medial malleolus
14 Tendon of tibialis anterior muscle
15 Extensor hallucis brevis muscle
16 Tendon of extensor hallucis longus muscle
17 Common tendon of gracilis, semitendinosus and sartoris muscles
18 Tibia

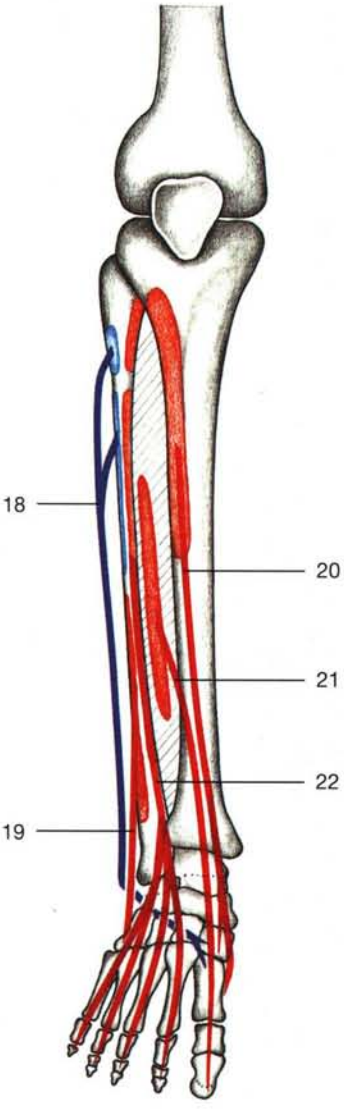

Extensor muscles of the leg (right side).

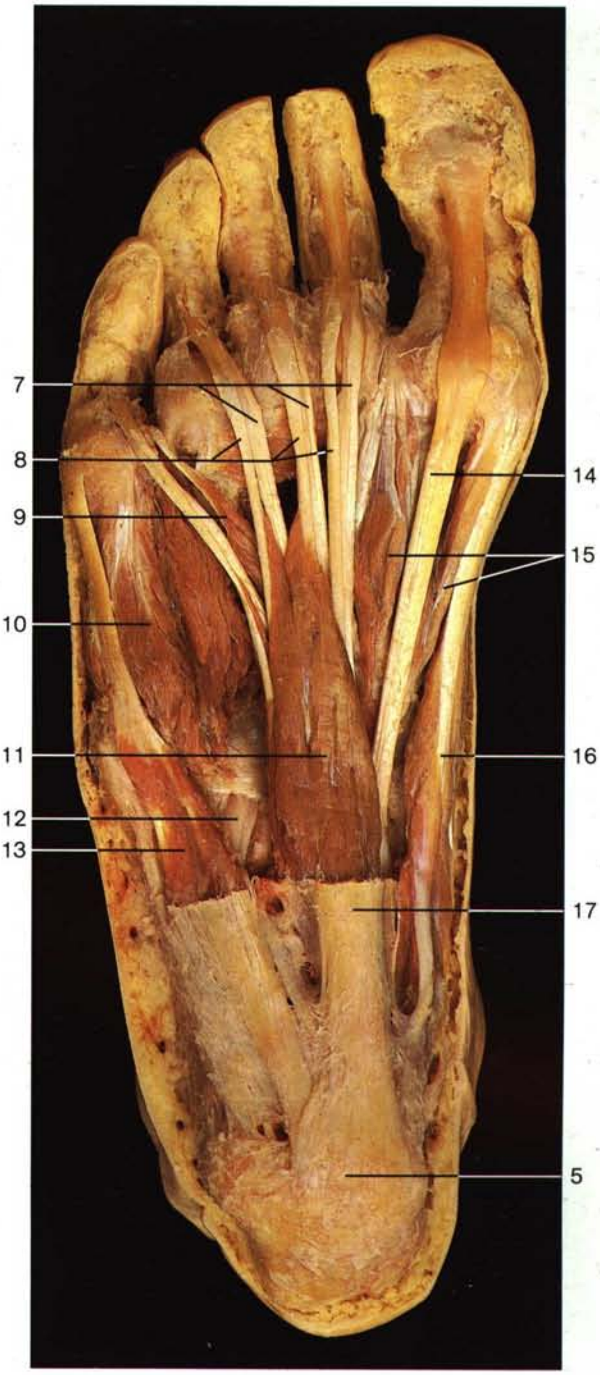

Sole of foot, first layer of muscles (from below). The plantar aponeurosis and the fasciae of the superficial muscles have been removed.

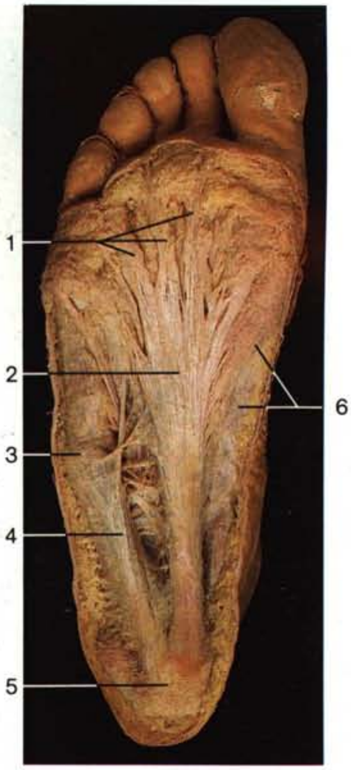

Sole of foot, plantar aponeurosis (from below).

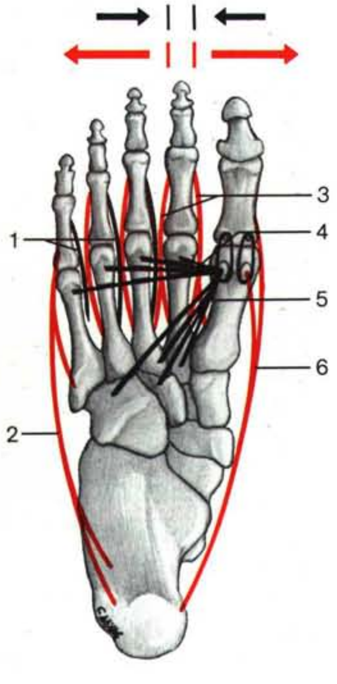

Course of abductor and adductor muscles of foot. (Schematic drawing.)
Red arrows = abduction.
Black arrows = adduction.

1 Longitudinal bands of plantar aponeurosis
2 Plantar aponeurosis
3 Position of tuberosity of 5th metatarsal bone
4 Muscles of 5th toe with fascia
5 Calcaneal tuberosity
6 Muscles of great toe with fascia
7 Tendons of flexor digitorum longus muscle
8 Tendons of flexor digitorum brevis muscle
9 Lumbrical muscle
10 Flexor digiti minimi brevis muscle
11 Flexor digitorum brevis muscle
12 Tendon of peroneus longus muscle
13 Abductor digiti minimi muscle
14 Tendon of flexor hallucis longus muscle
15 Flexor hallucis brevis muscle
16 Abductor hallucis muscle
17 Plantar aponeurosis (cut)
18 Peroneus longus muscle
19 Peroneus brevis muscle
20 Tibialis anterior muscle
21 Extensor hallucis longus muscle
22 Extensor digitorum longus muscle

1 Plantar interossei muscles (black)
2 Abductor digiti minimi muscles (red)
3 Dorsal interossei muscle (red)
4 Transverse head of adductor muscle (black)
5 Oblique head of adductor muscle (black)
6 Abductor hallucis muscle (red)

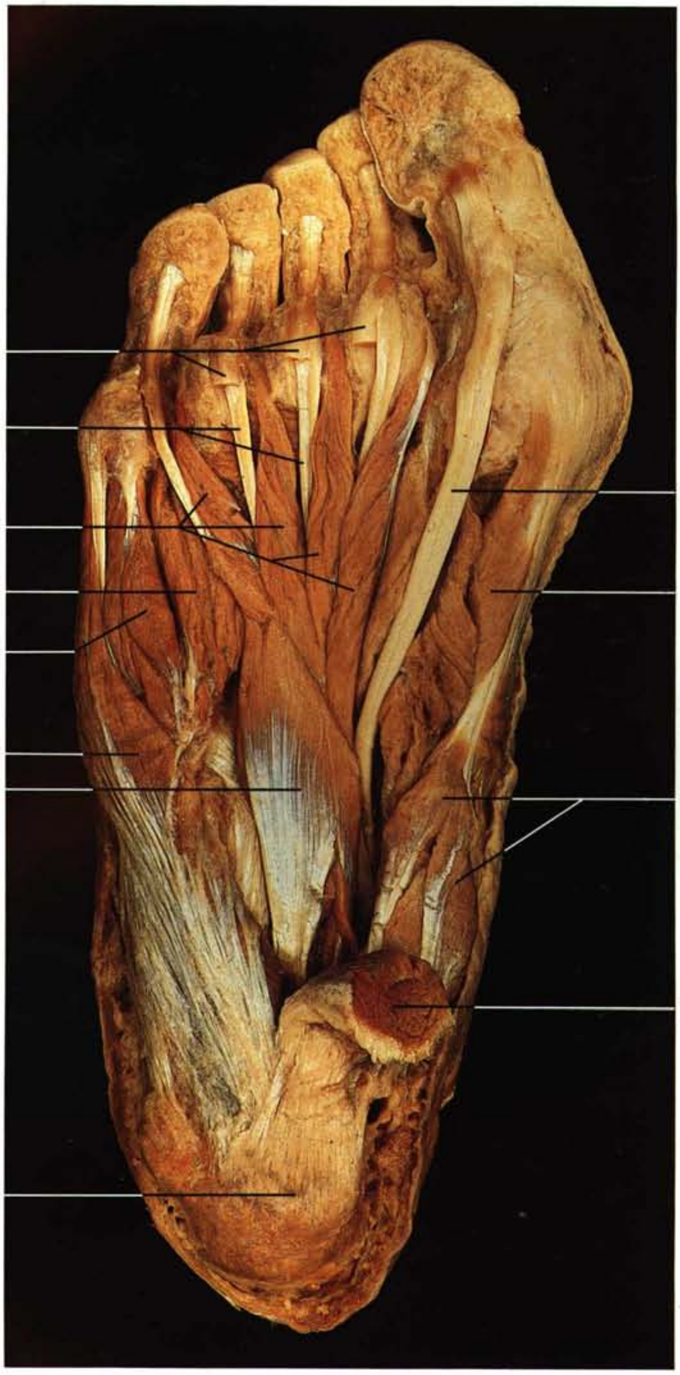

Muscles of sole of foot, second layer (from below). The flexor digitorum brevis muscle has been divided.

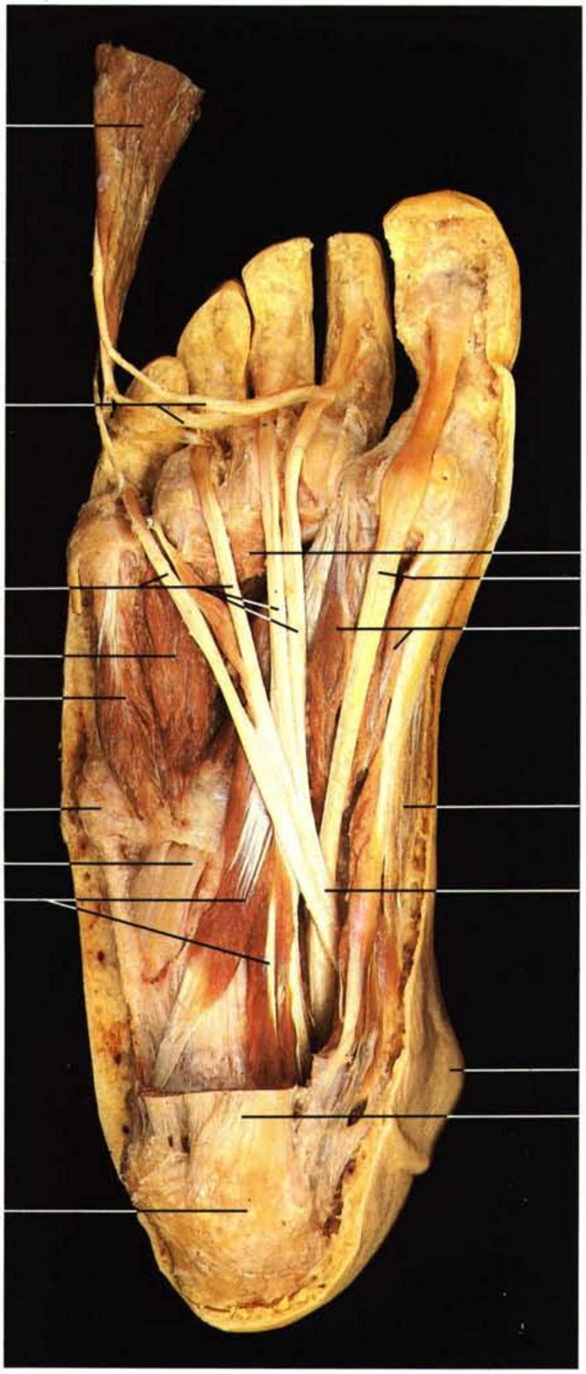

Muscles of sole of foot, second layer (from below). The tendons of the flexor muscles and the crossing of tendons are displayed. The flexor digitorum brevis muscle has been divided and reflected.

1 Tendons of flexor digitorum brevis muscle
2 Tendons of flexor digitorum longus muscle
3 Lumbrical muscles
4 Interossei muscles
5 Flexor digiti minimi brevis muscle
6 Abductor digiti minimi muscle
7 Quadratus plantae muscle
8 Calcaneal tuberosity
9 Tendon of flexor hallucis longus muscle
10 Flexor hallucis brevis muscle
11 Abductor hallucis muscle
12 Flexor digitorum brevis muscle (divided)
13 Tuberosity of 5th metatarsal bone
14 Tendon of peroneus longus muscle
15 Transverse head of adductor hallucis muscle
16 Crossing of tendons in sole of foot
17 Medial malleolus
18 Plantar aponeurosis (divided)

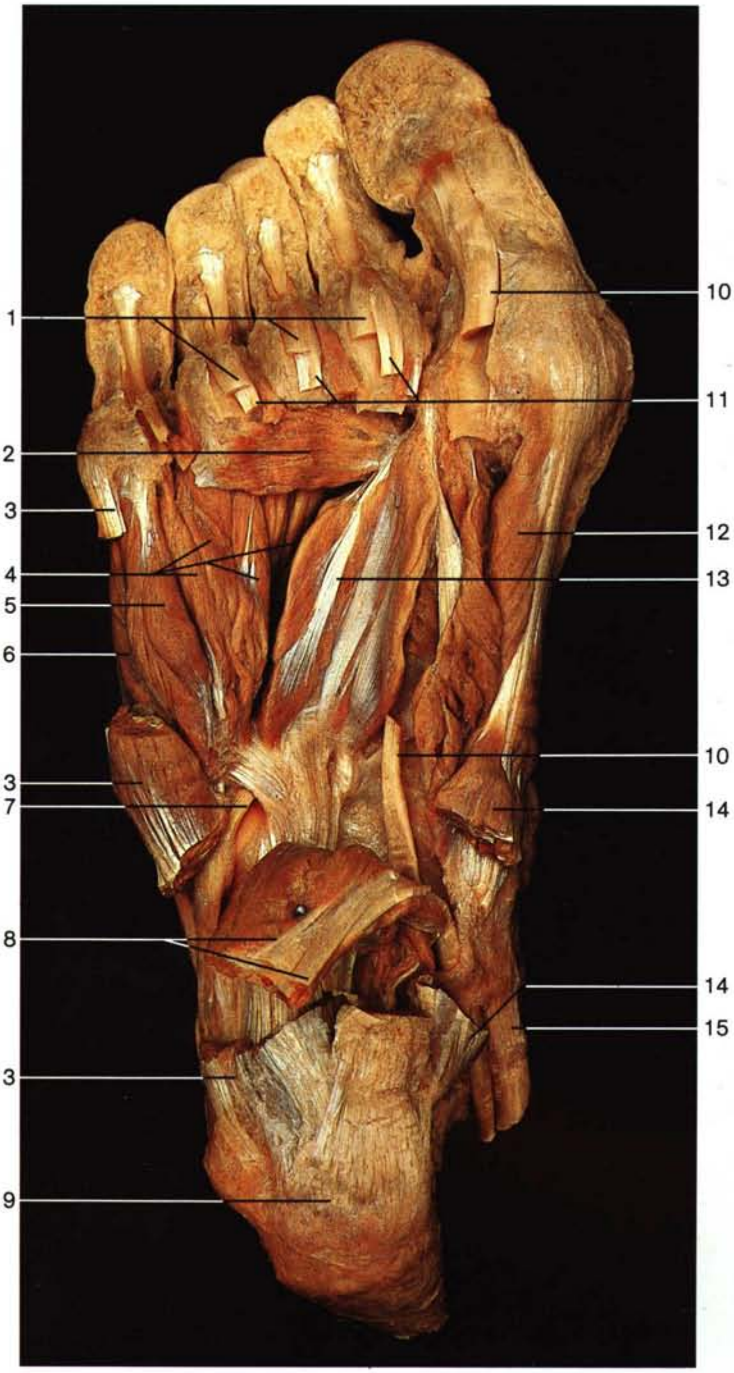

Muscles of sole of foot, third layer (from below). The flexor digitorum brevis muscle has been removed, and the quadratus plantae muscle and the abductor hallucis and digiti minimi muscles have been divided.

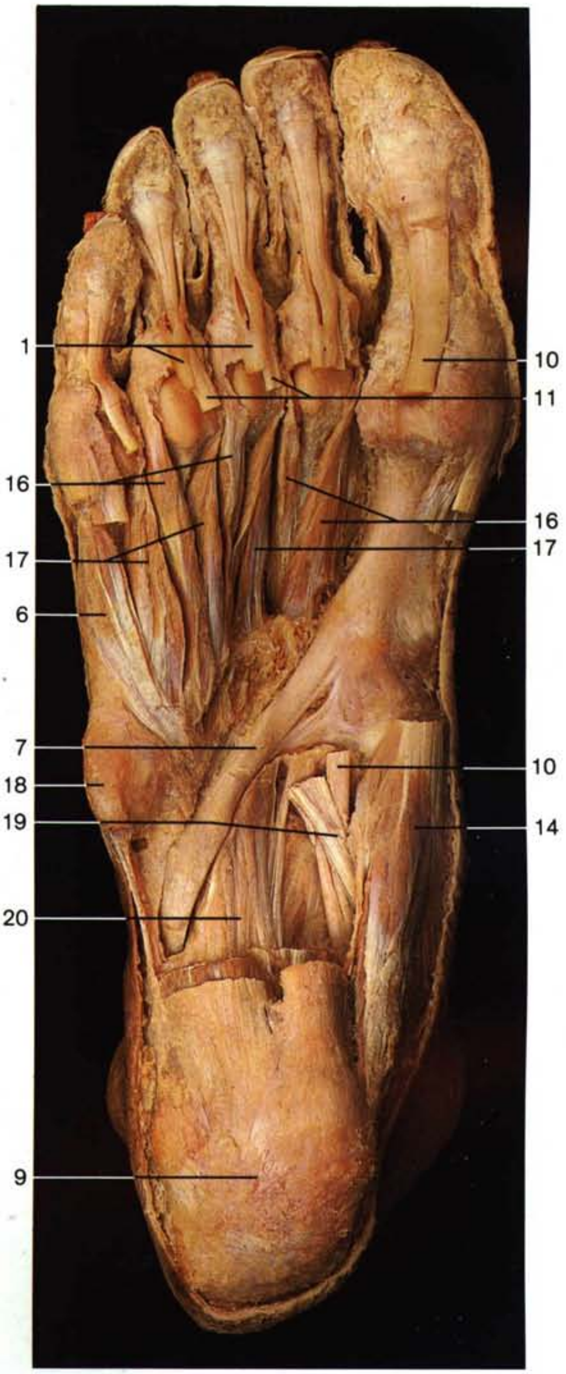

Muscles of sole of foot, fourth layer (from below). The interosseous muscles and the canal for the tendon of peroneus longus muscle are shown.

1 Tendons of flexor digitorum brevis muscle
2 Transverse head of adductor hallucis muscle
3 Abductor digiti minimi muscle
4 Interosseous muscles
5 Flexor digiti minimi brevis muscle
6 Opponens digiti minimi muscle
7 Tendon of peroneus longus muscle
8 Quadratus plantae muscle with tendon of flexor digitorum longus muscle
9 Calcaneal tuberosity
10 Tendons of flexor hallucis longus muscle (divided)
11 Tendon of flexor digitorum longus muscle
12 Flexor hallucis brevis muscle
13 Oblique head of adductor hallucis muscle
14 Abductor hallucis muscle (cut)
15 Tendon of tibialis posterior muscle
16 Dorsal interossei muscles
17 Plantar interossei muscles
18 Tuberosity of 5th metatarsal bone
19 Tendon of flexor digitorum longus muscle (crossing of plantar tendons)
20 Long plantar ligament

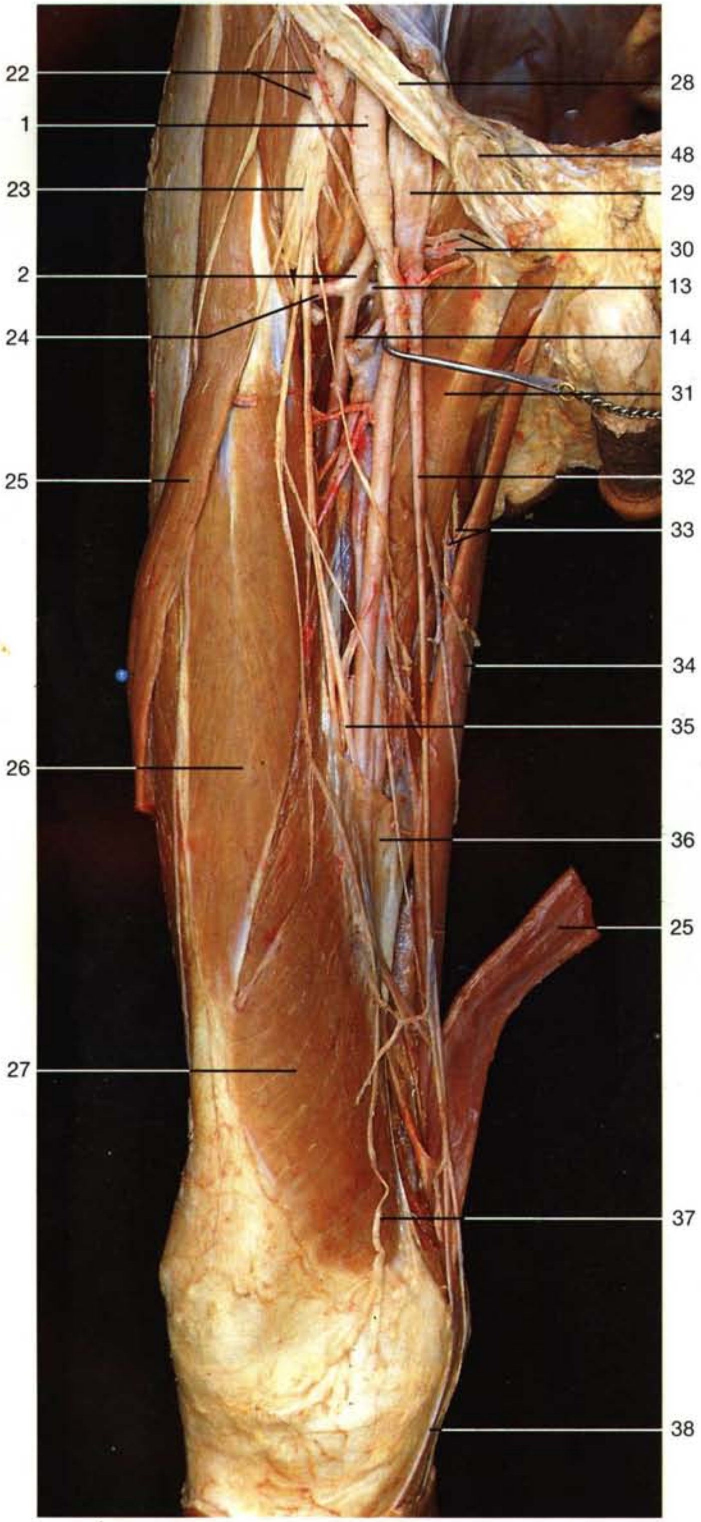
22
1
23
2
24
25
26
27
28
48
29
30
13
14
31
32
33
34
35
36
25
37
38

Main arteries and nerves of right thigh (anterior aspect). Sartorius muscle has been divided and reflected. The femoral vein has been partly removed to show the deep femoral artery. Notice: the vessels enter the adductor canal to reach the popliteal fossa.

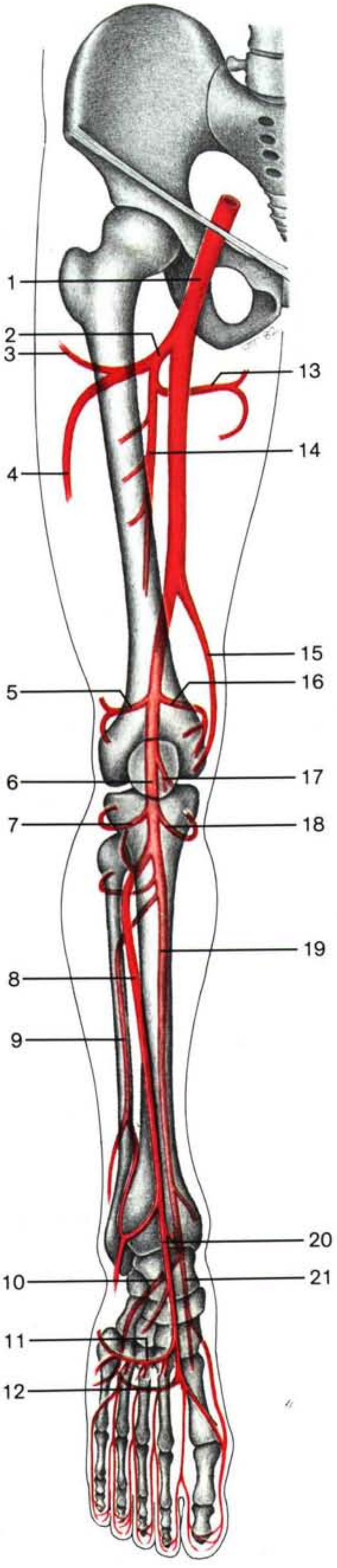
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

Main arteries of lower extremity, right side (ventral aspect). (Schematic drawing.)

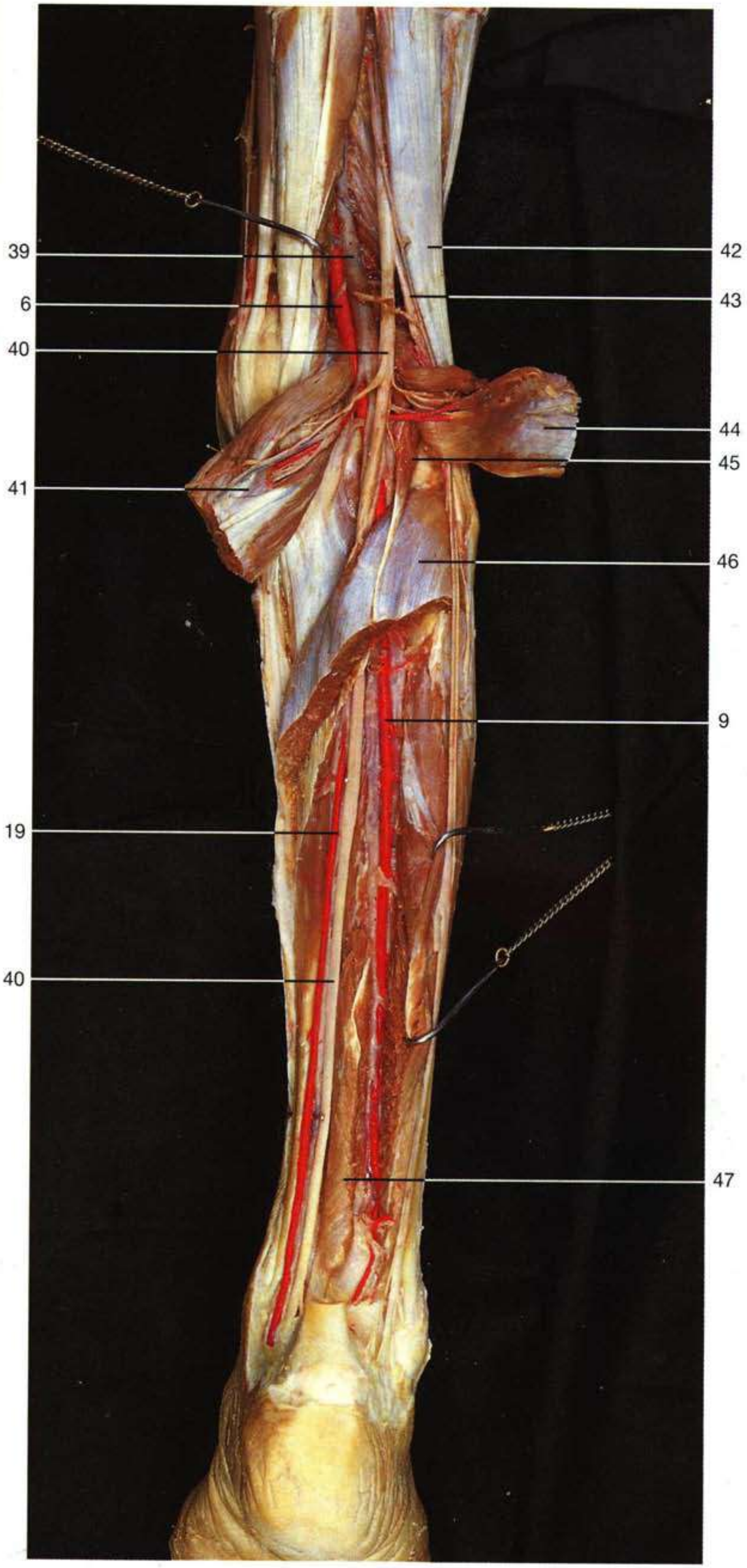

1 **Femoral artery**
2 **Profunda femoris artery**
3 Ascending branch of lateral circumflex femoral artery
4 Descending branch of lateral circumflex femoral artery
5 Lateral superior genicular artery
6 Popliteal artery
7 Lateral inferior genicular artery
8 **Anterior tibial artery**
9 **Peroneal artery**
10 **Lateral plantar artery**
11 Arcuate artery with dorsal metatarsal arteries
12 **Plantar arch** with plantar metatarsal arteries
13 Medial circumflex femoral artery
14 Profunda femoris artery with perforating arteries
15 Descending genicular artery
16 Medial superior genicular artery
17 Middle genicular artery
18 Medial inferior genicular artery
19 **Posterior tibial artery**
20 **Dorsalis pedis artery**
21 **Medial plantar artery**
22 Superficial and deep circumflex iliac arteries
23 **Femoral nerve**
24 Lateral circumflex femoral artery
25 Sartorius muscle (cut and reflected)
26 Rectus femoris muscle
27 Vastus medialis muscle
28 Inguinal ligament
29 **Femoral vein** (cut)
30 External pudendal artery and vein
31 Adductor longus muscle
32 Great saphenous vein
33 Obturator artery and nerve
34 Gracilis muscle
35 Saphenous nerve
36 Tendinous wall of adductor canal
37 Anterior cutaneous branch of femoral nerve
38 Infrapatellar branch of saphenous nerve
39 Popliteal vein
40 **Tibial nerve**
41 Medial head of gastrocnemius muscle
42 Biceps femoris muscle
43 Common peroneal nerve
44 Lateral head of gastrocnemius muscle
45 Plantaris muscle
43 Soleus muscle
47 Flexor hallucis longus muscle
48 Spermatic cord

Arteries of the right leg (posterior aspect).

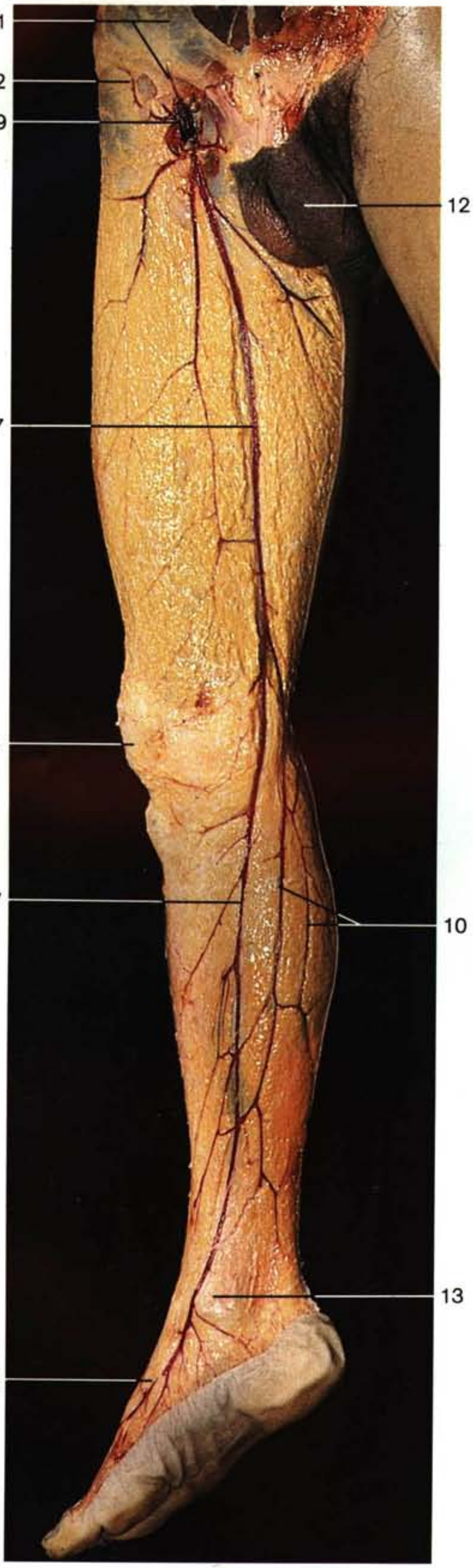

Superficial veins of lower limb, right side (medial anterior aspect). The veins have been injected with red solution.

Main veins of lower limb, right side (anterior aspect). (Schematic drawing.)

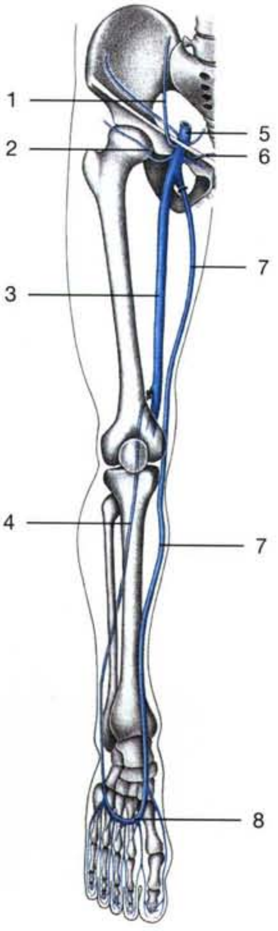

1 Superficial epigastric vein
2 Superficial circumflex iliac vein
3 Femoral vein
4 **Small saphenous vein**
5 External iliac vein
6 External pudendal vein
7 **Great saphenous vein**
8 Dorsal venous arch
9 Saphenous opening with femoral vein
10 Venous anastomoses of small saphenous vei
with great saphenous vein
11 Patella
12 Penis
13 Medial malleolus
14 Popliteal fossa
15 Perforating veins
16 Lateral malleolus
17 Dorsal digital veins of foot
18 Dorsal venous arch of foot
19 Dorsal metatarsal veins of foot
20 Anterior tibial artery and veins
21 Tibia
22 **Posterior tibial artery and veins**
23 Fibula
24 Peroneal artery and vein
25 Deep layer of crural fascia
26 Superficial layer of crural fascia
27 Perforating veins I-III (of Cockett)
28 **Tibial nerve**
29 **Arcuate vein**
30 **Saphenous nerve**
31 Medial dorsal cutaneous nerve
(branch of superficial peroneal nerve)
32 Posterior tibial vein

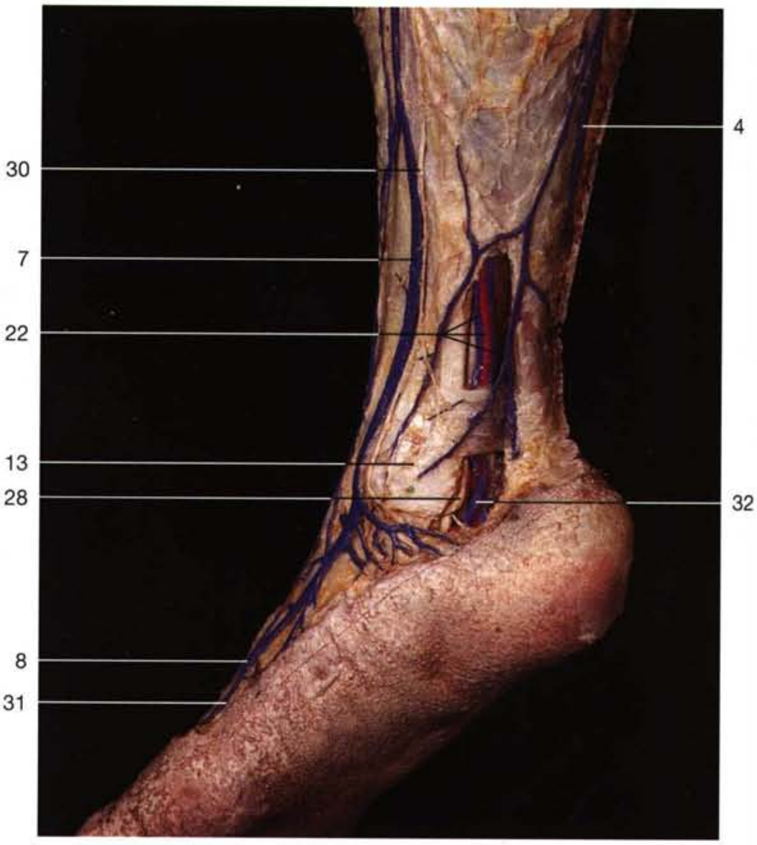

▷

Medial malleolar region. Dissection of tibial nerve, posterior tibial vessels, and great saphenous vein (veins injected with blue resin).

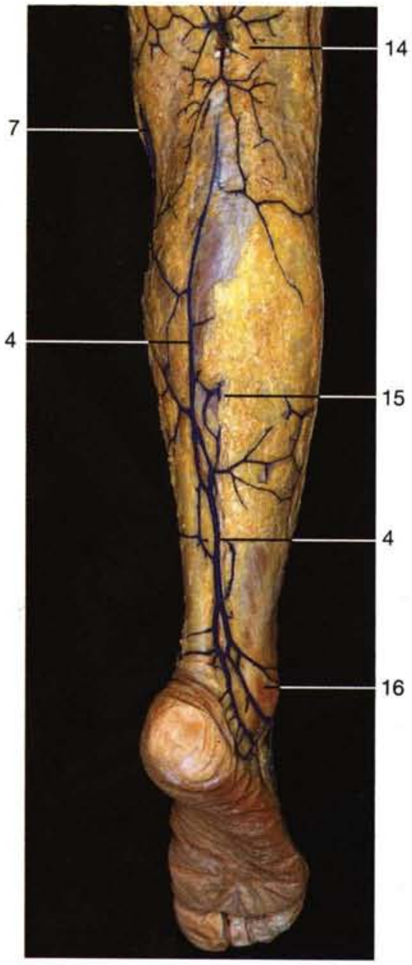

Superficial veins of leg (posterior aspect; injected with blue resin).

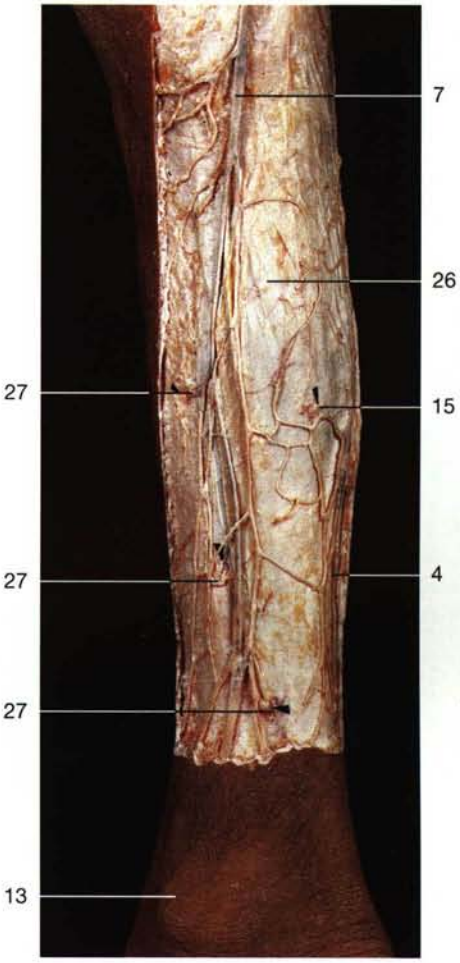
7
26
27
15
27
4
27
13

Superficial veins of leg. The perforating veins of Cockett have been dissected.

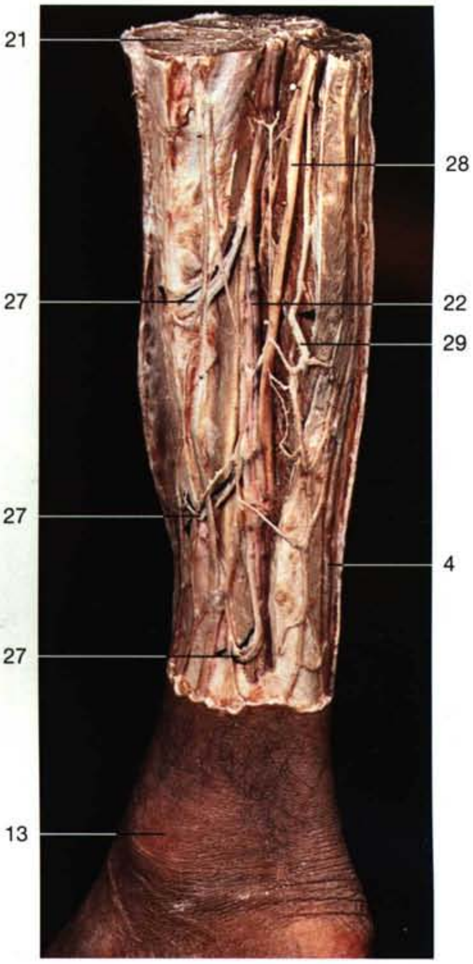
21
28
27
22
29
27
4
27
13

Veins of leg. The anastomoses between superficial and deeper veins are dissected.

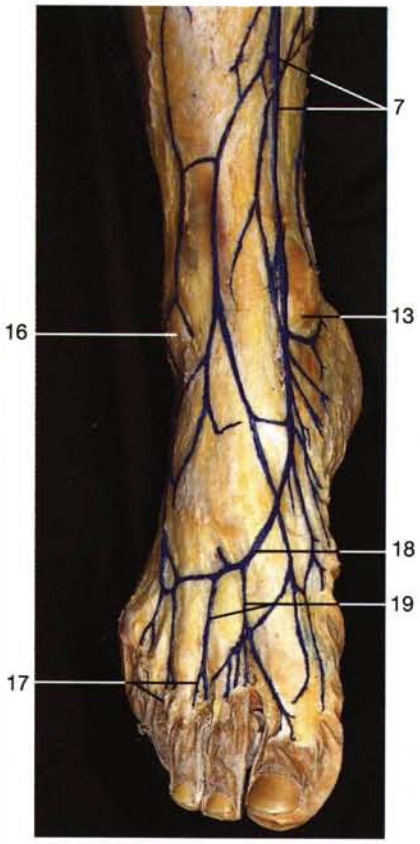
7
16
13
18
19
17

◁

Superficial veins on dorsum of foot (injected with blue resin).

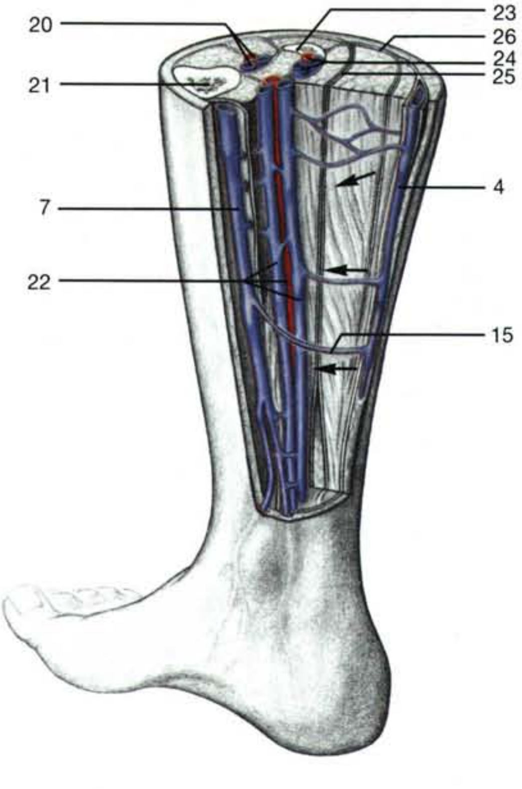
20
23
26
24
21
25
7
4
22
15

Anastomoses between superficial and deep veins of the leg (after Aigner). (Schematic drawing.) Arrows: directions of blood flow. ▷

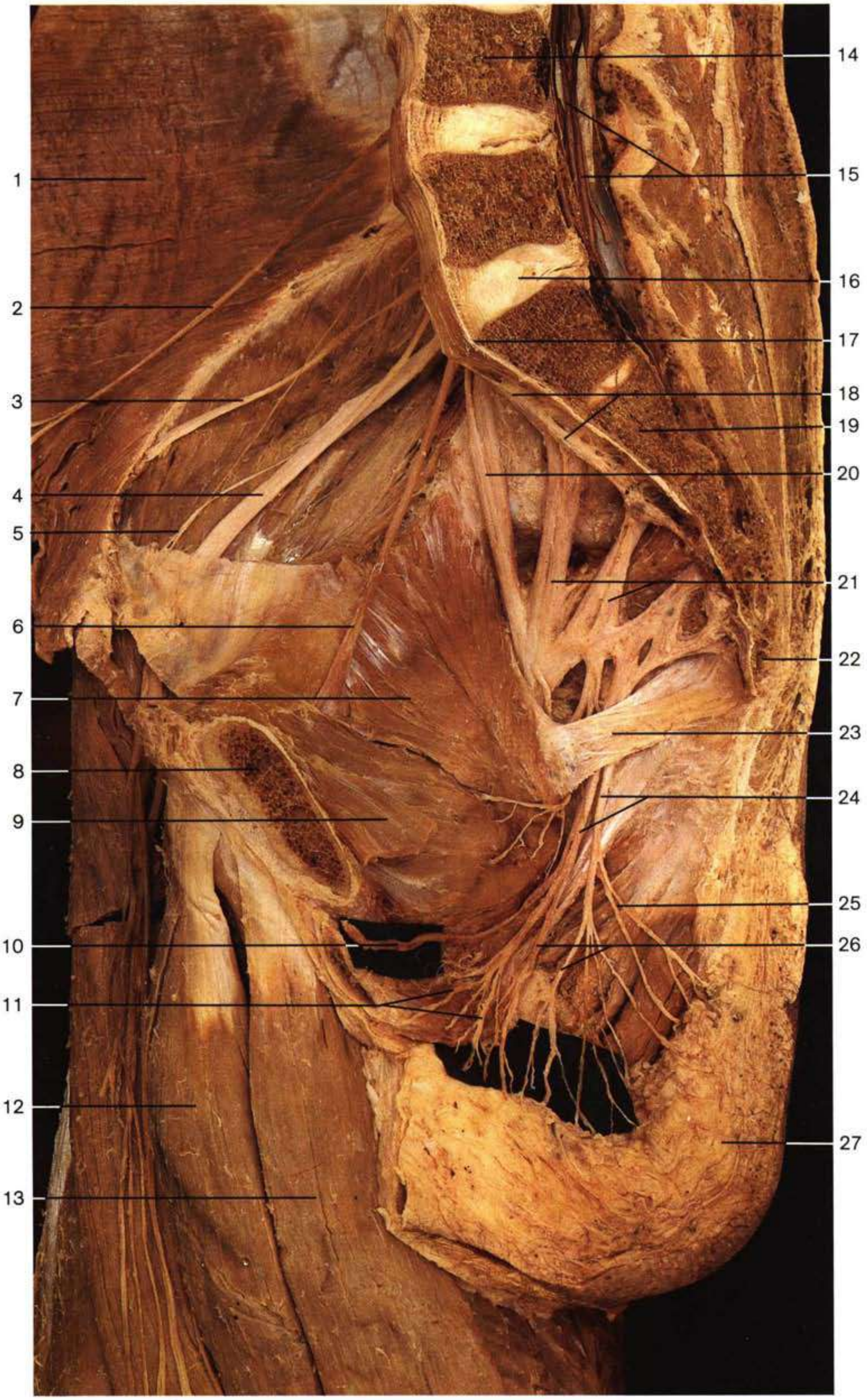

1 Transversus abdominis muscle
2 Iliohypogastric nerve
3 Ilioinguinal nerve
4 **Femoral nerve**
5 Lateral femoral cutaneous nerve
6 Obturator nerve
7 Obturator internus muscle
8 Pubic bone (cut edge)
9 Levator ani muscle (remnant)
10 Dorsal nerve of penis
11 Posterior scrotal nerves
12 Adductor longus muscle
13 Gracilis muscle
14 Body of 4th lumbar vertebra
15 Cauda equina
16 Intervertebral disc
17 Sacral promontory
18 Sympathetic trunk
19 Sacrum
20 **Lumbosacral trunk**
21 **Sciatic plexus**
22 Coccyx
23 Sacrospinous ligament
24 **Pudendal nerve**
25 Inferior rectal nerves
26 Perineal nerves
27 Subcutaneous fat tissue of gluteal region

Lumbosacral plexus in situ, right side (medial aspect).
Pelvic organs with peritoneum and part of the levator ani muscle have been removed.

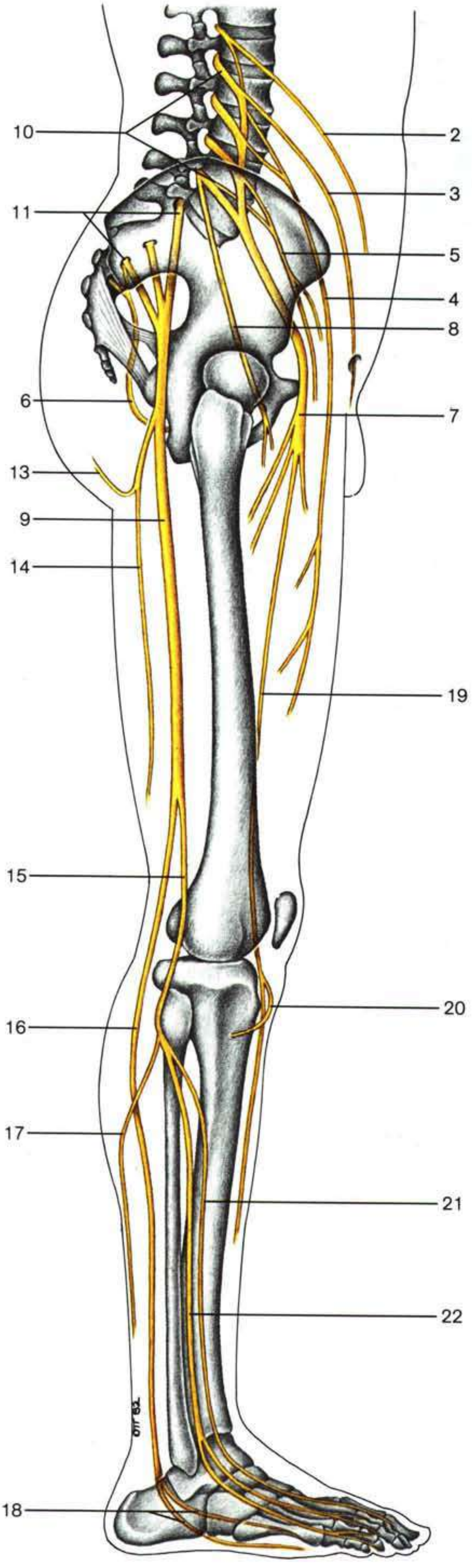

Nerves of lower limb, right side (lateral aspect). (Schematic drawing.)

1 Subcostal nerve
2 Iliohypogastric nerve
3 Ilioinguinal nerve
4 Lateral femoral cutaneous nerve
5 Genitofemoral nerve
6 **Pudendal nerve**
7 **Femoral nerve**
8 Obturator nerve
9 **Sciatic nerve**
10 Lumbar plexus (L_1–L_4) } lumbosacral plexus
11 Sacral plexus (L_4–S_4) } lumbosacral plexus
12 "Pudendal" plexus (S_2–S_4) } lumbosacral plexus
13 Inferior cluneal nerve
14 Posterior femoral cutaneous nerve
15 Common peroneal nerve
16 **Tibial nerve**
17 Lateral sural cutaneous nerve
18 Medial and lateral plantar nerves
19 Saphenous nerve
20 Infrapatellar branch of saphenous nerve
21 **Deep peroneal nerve**
22 **Superficial peroneal nerve**

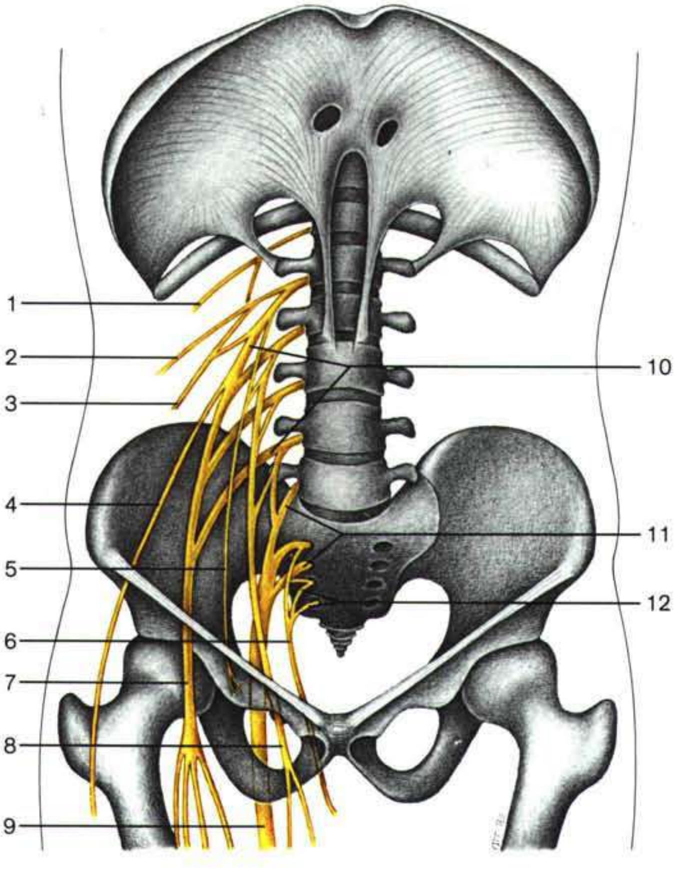

Main branches of lumbosacral plexus (ventral aspect). (Schematic drawing.)

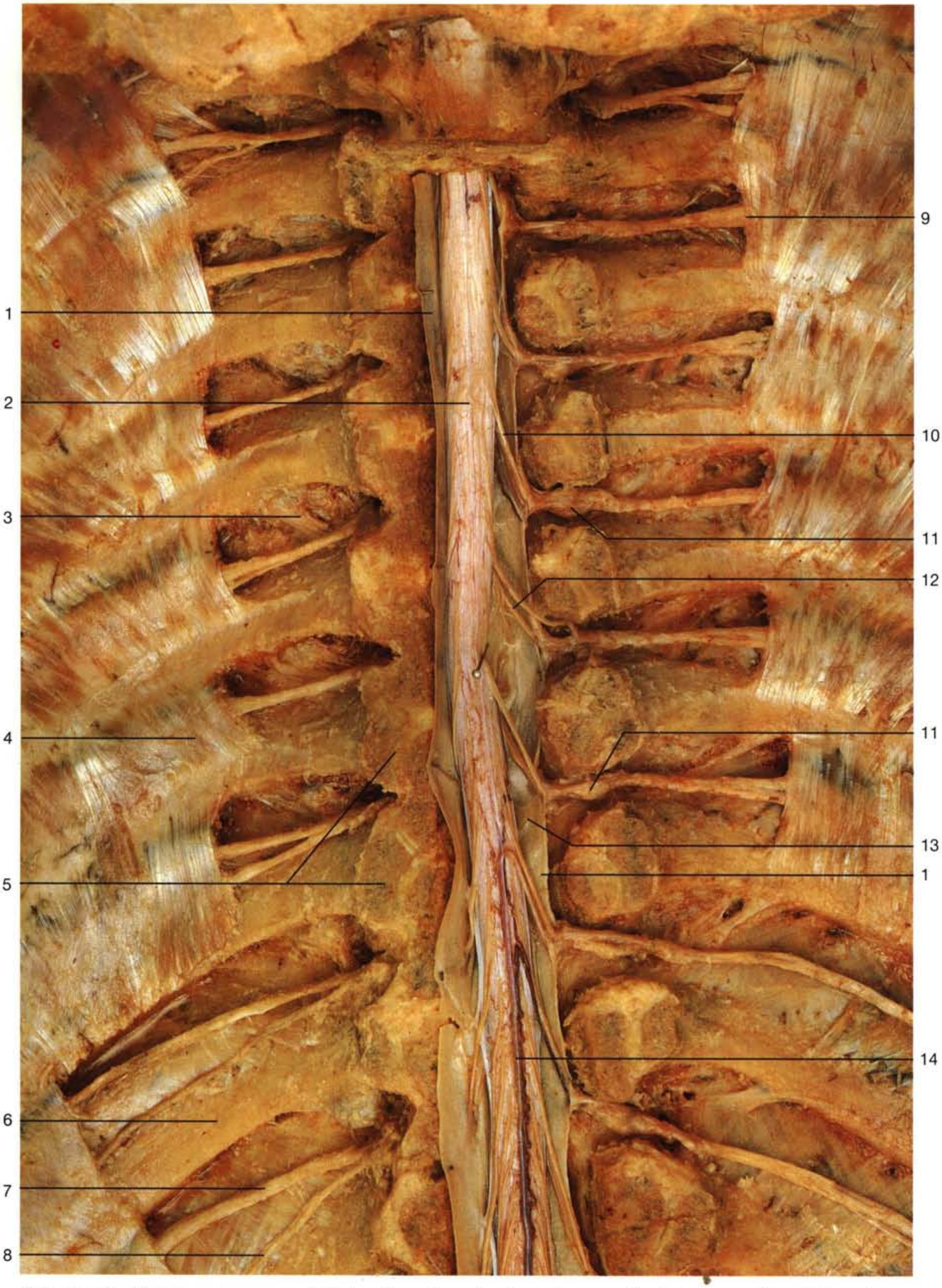

Spinal cord with intercostal nerves. Inferior thoracic region (anterior aspect). Anterior portion of thoracic vertebrae removed, dural sheath opened and spinal cord slightly reflected to the right to display the dorsal and ventral roots.

1	Dura mater	6	Eleventh rib	10	Anterior root filaments
2	**Spinal cord**	7	**Intercostal nerve**	11	Spinal (dorsal root) ganglion
3	Costotransverse ligament	8	Collateral branch of intercostal nerve	12	Posterior root filaments
4	Innermost intercostal muscle	9	Intercostal nerve (entering the intermuscular interval)	13	Arachnoid mater and denticulate ligament
5	Vertebral arches (cut surfaces)			14	Anterior spinal artery

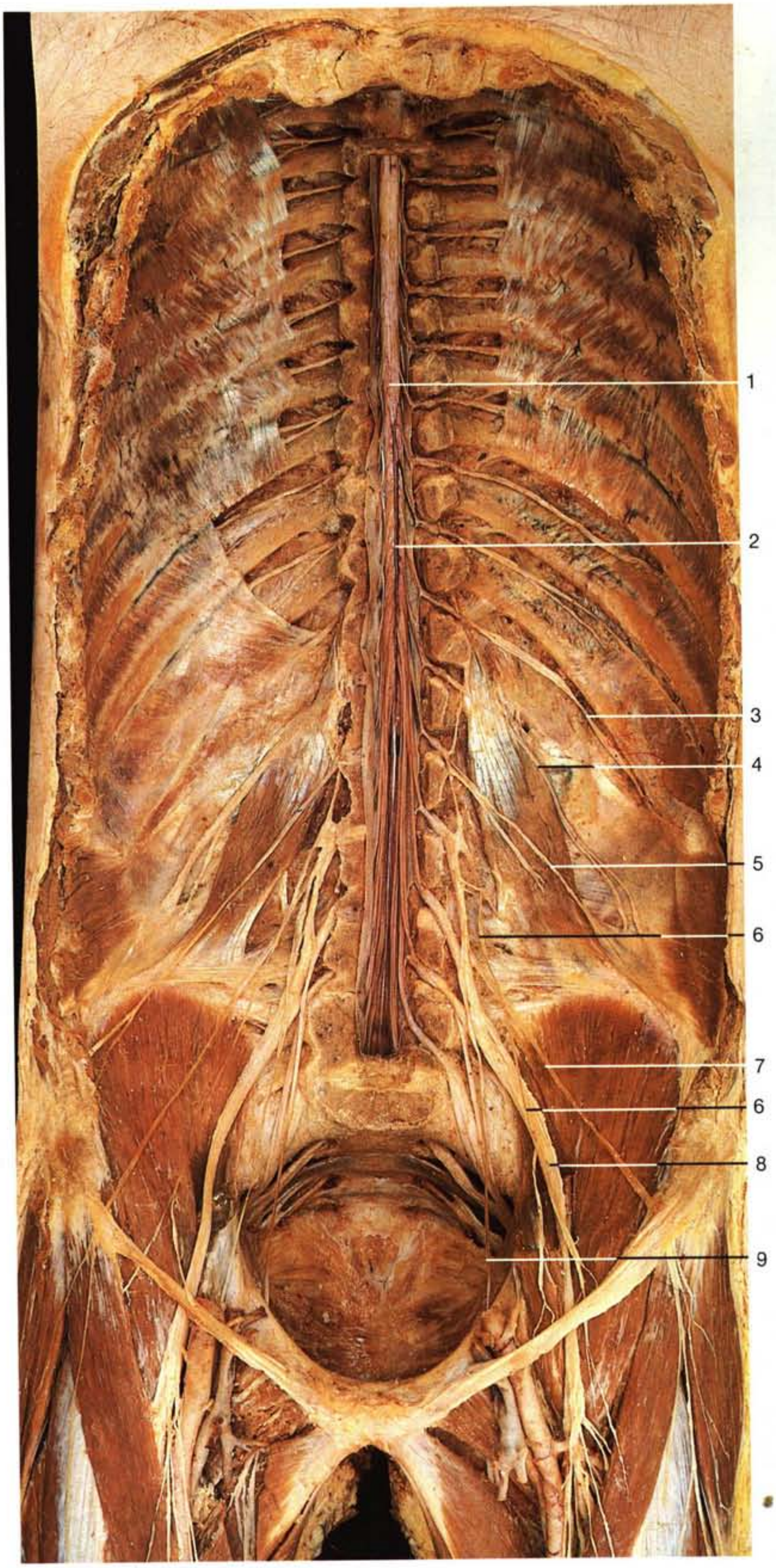
1
2
3
4
5
6
7
6
8
9

Spinal cord and lumbar plexus in situ (anterior aspect).

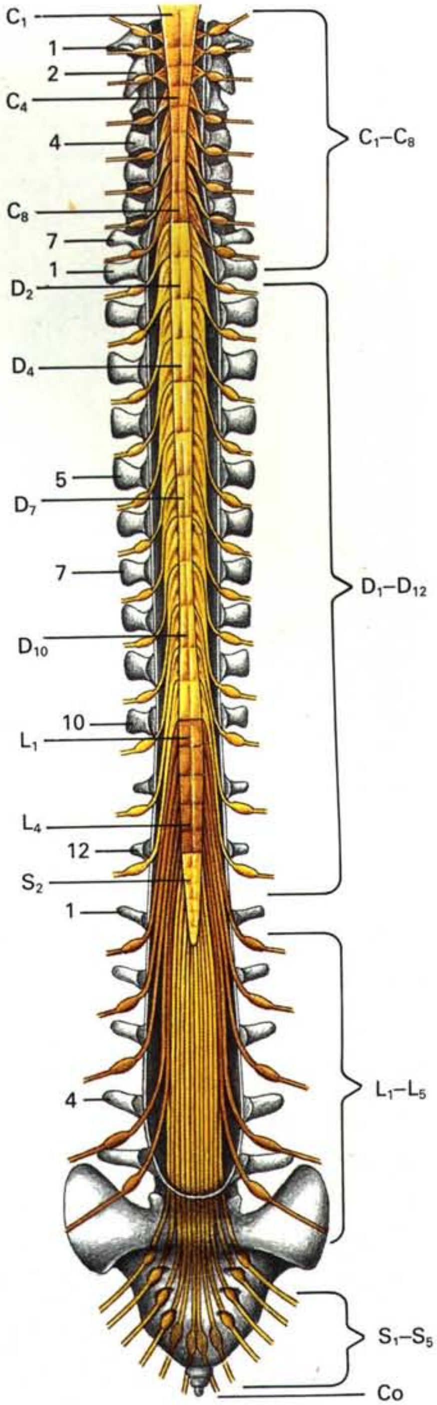
C_1
1
2
C_4
4
C_8
7
1
D_2
D_4
5
D_7
7
D_{10}
10
L_1
L_4
12
S_2
1
4
C_1–C_8
D_1–D_{12}
L_1–L_5
S_1–S_5
Co

Organization of spinal cord segments in relation to the vertebral column (anterior aspect). C = cervical; D = thoracic; L = lumbar; S = sacral segments; Co = coccygeal bone.

1 Conus medullaris
2 Filum terminale
3 Subcostal nerve
4 Iliohypogastric nerve
5 Ilioinguinal nerve
6 Genitofemoral nerve
7 Lateral femoral cutaneous nerve
8 Femoral nerve
9 Obturator nerve

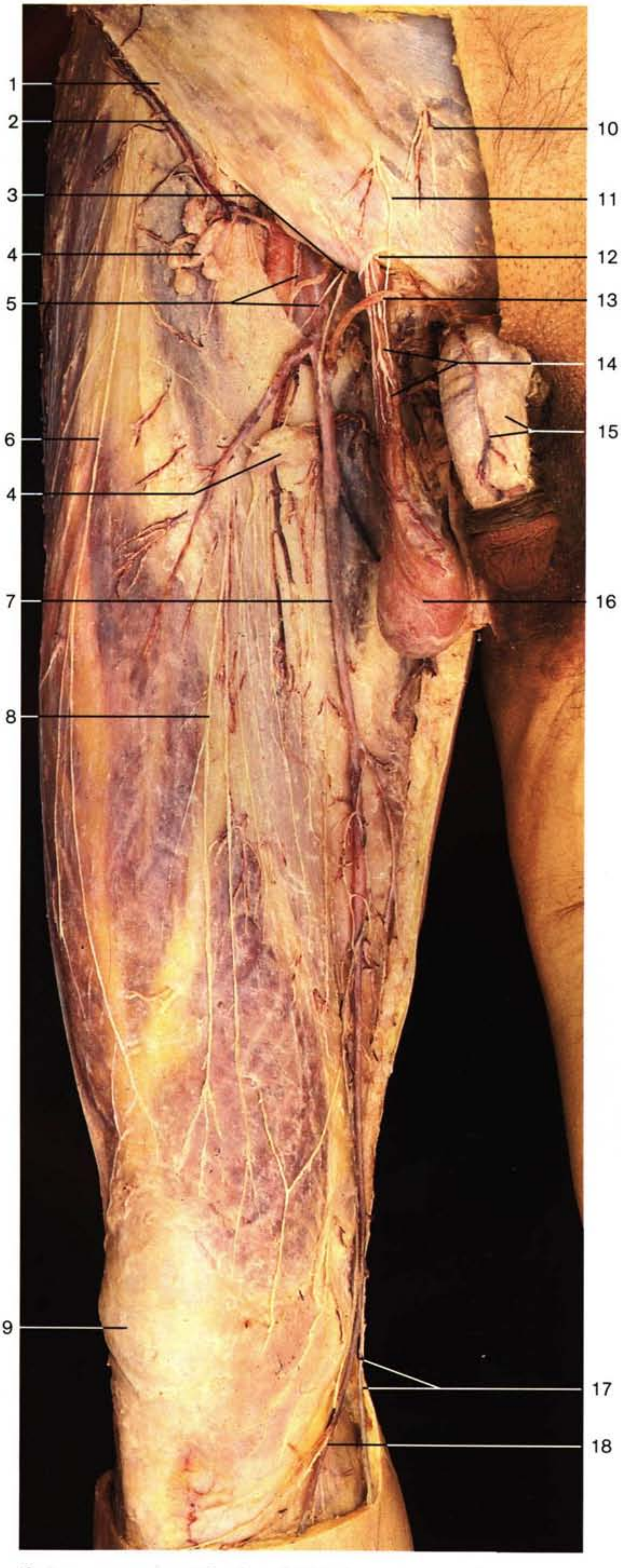
1
2
3
4
5
6
4
7
8
9
10
11
12
13
14
15
16
17
18

Cutaneous nerves and veins of thigh (anterior aspect).

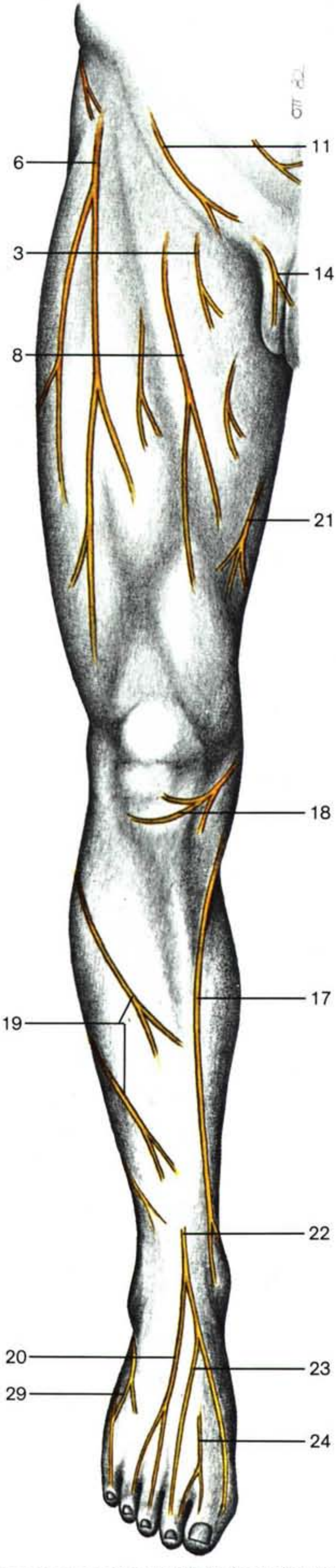
6
11
3
14
8
21
18
17
19
22
20
23
29
24

Cutaneous nerves of lower limb (anterior aspect). (Schematic drawing.)

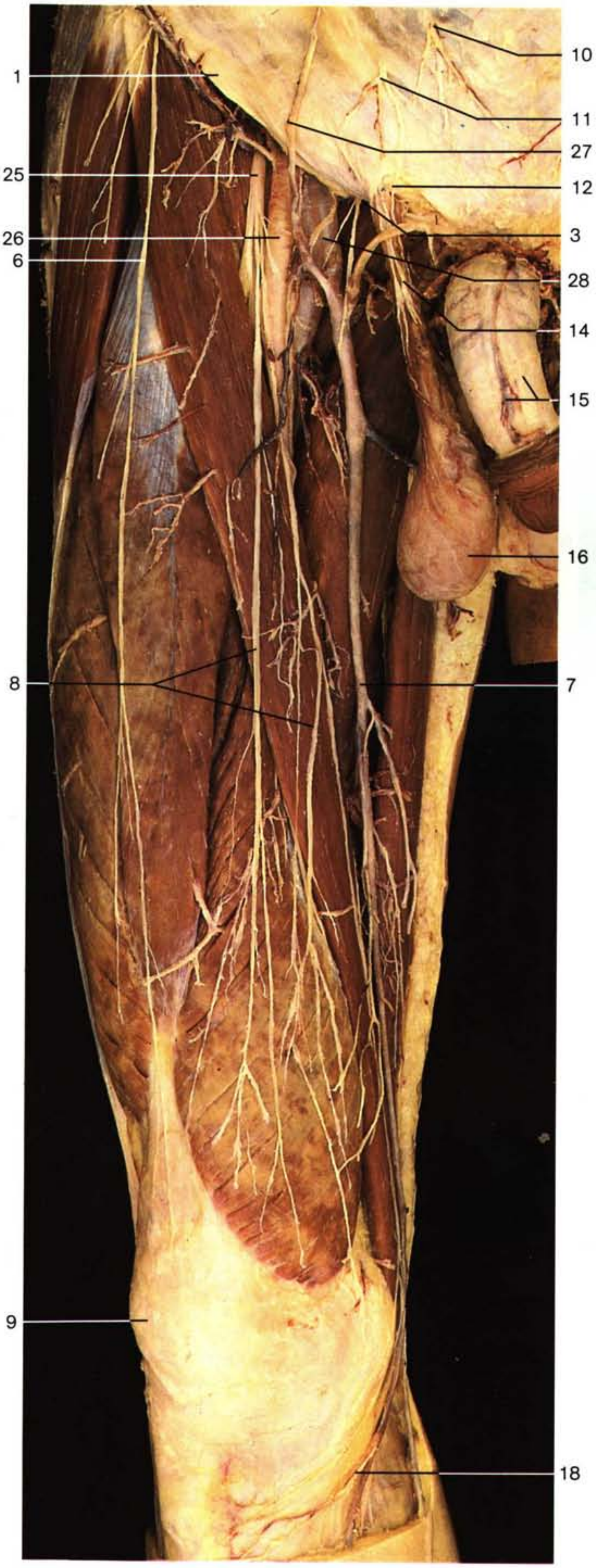

Cutaneous nerves and veins of thigh (anterior aspect). The fascia lata and fasciae of the thigh muscles have been removed.

1 Inguinal ligament
2 Superficial circumflex iliac vein
3 Femoral branch of genitofemoral nerve
4 Superficial inguinal lymph nodes
5 Saphenous opening with femoral artery and vein
6 Lateral femoral cutaneous nerve
7 Great saphenous vein
8 Anterior cutaneous branches of femoral nerve
9 Patella
10 Terminal branches of subcostal nerve
11 Terminal branches of iliohypogastric nerve
12 Superficial inguinal ring
13 External pudendal vein
14 Spermatic cord with genital branch of genitofemoral nerve
15 Penis with superficial dorsal vein of penis
16 Testis and its coverings
17 Saphenous nerve
18 Infrapatellar branch of saphenous nerve
19 Lateral sural cutaneous nerves
20 Intermediate dorsal cutaneous branch of superficial peroneal nerve
21 Cutaneous branch of obturator nerve
22 Superficial peroneal nerve
23 Medial dorsal cutaneous branch of superficial peroneal nerve
24 Deep peroneal nerve
25 **Femoral nerve**
26 **Femoral artery**
27 Superficial epigastric vein
28 **Femoral vein**
29 Lateral dorsal cutaneous branch of sural nerve
30 **Inguinal nodes** (enlarged)
31 Lymphatic vessels
32 Sartorius muscle

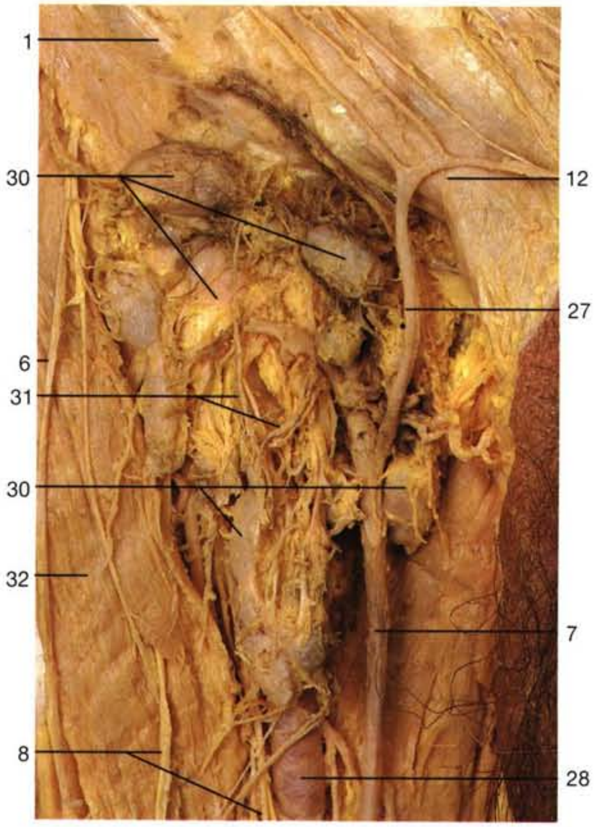

Inguinal nodes with lymphatic vessels (anterior aspect).

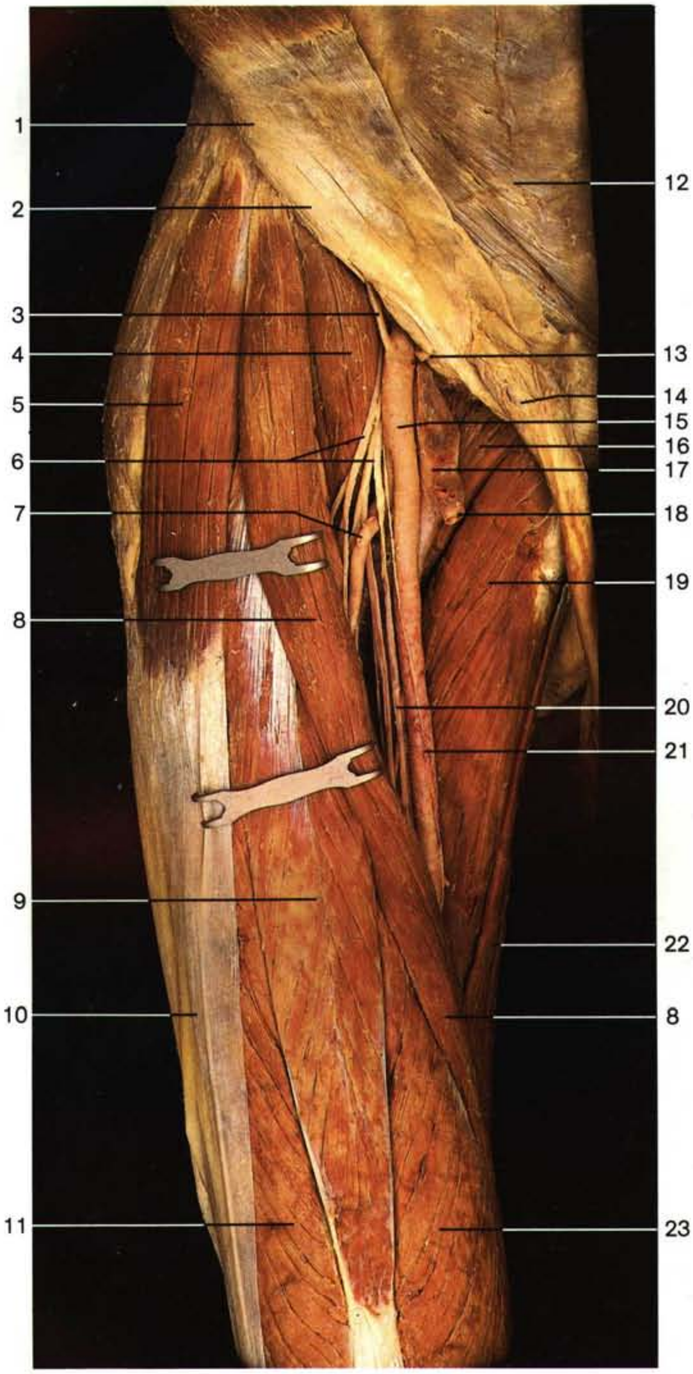

Anterior region of right thigh (anterior aspect).
The fascia lata has been removed, and the sartorius muscle has been slightly reflected.

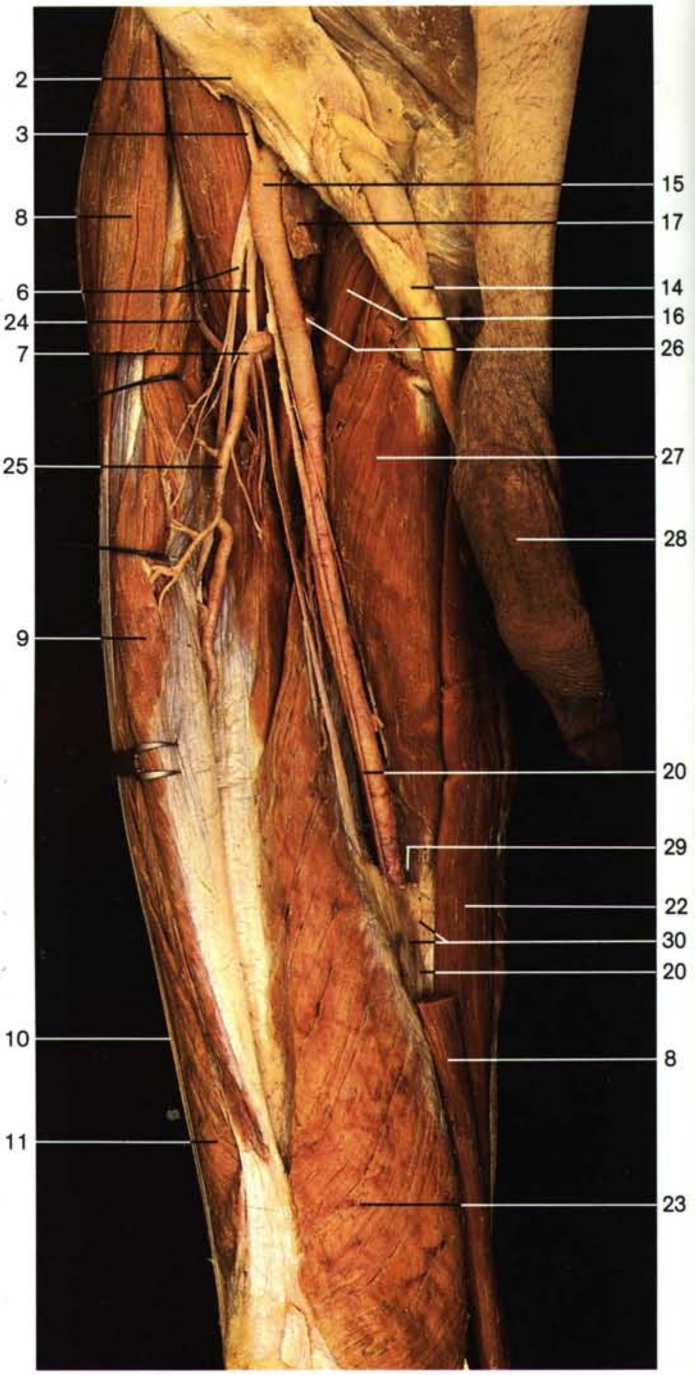

Anterior region of right thigh (anterior aspect).
The fascia lata has been removed, and the sartorius muscle has been divided.

1 Anterior superior iliac spine
2 Inguinal ligament
3 **Deep circumflex iliac artery**
4 Iliopsoas muscle
5 Tensor fasciae latae muscle
6 **Femoral nerve**
7 **Lateral circumflex femoral artery**
8 Sartorius muscle
9 Rectus femoris muscle
10 Iliotibial tract
11 Vastus lateralis muscle
12 Anterior sheath of rectus abdominis muscle
13 Inferior epigastric artery
14 Spermatic cord
15 **Femoral artery**
16 Pectineus muscle
17 **Femoral vein**
18 Great saphenous vein (divided)
19 Adductor longus muscle
20 **Saphenous nerve**
21 Muscular branch of femoral nerve
22 Gracilis muscle
23 Vastus medialis muscle
24 Ascending branch of lateral circumflex femoral artery
25 Descending branch of lateral circumflex femoral artery
26 Medial circumflex femoral artery
27 Adductor longus muscle
28 Penis
29 Entrance to adductor canal
30 Vastoadductory lamina of fascia beneath sartorius muscle

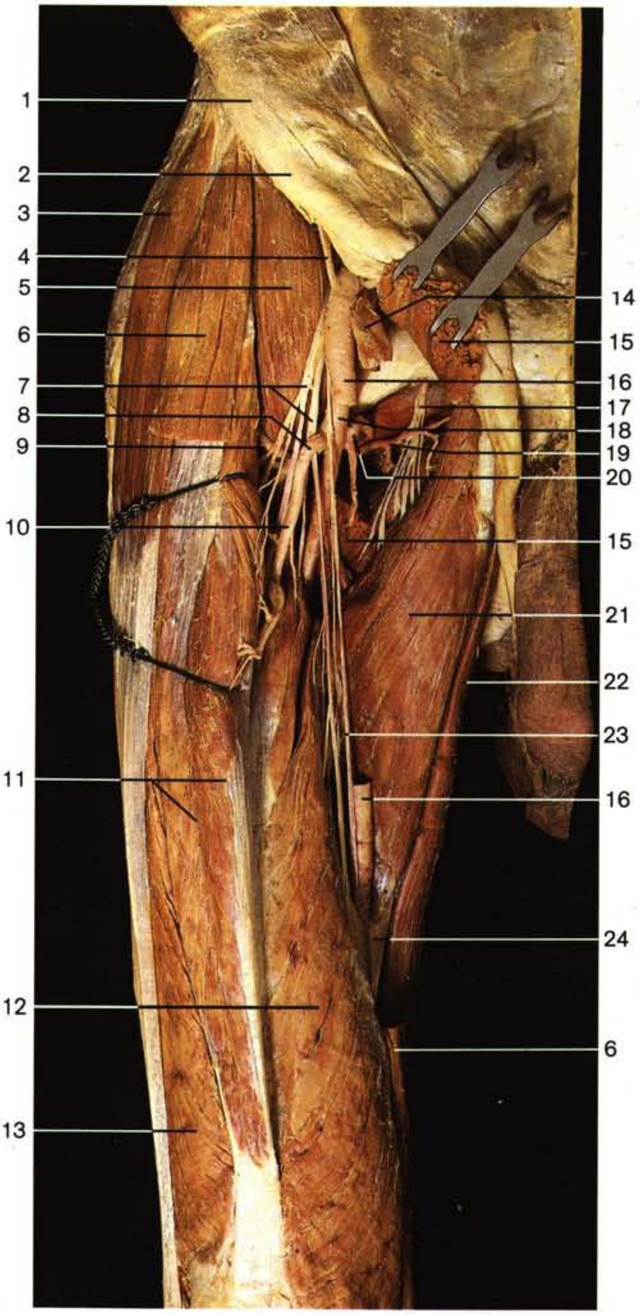

Anterior region of right thigh (anterior aspect). The fascia lata has been removed. Sartorius muscle, pectineus muscle and femoral artery have been cut to display the deep femoral artery with its branches. The rectus femoris muscle has been slightly reflected.

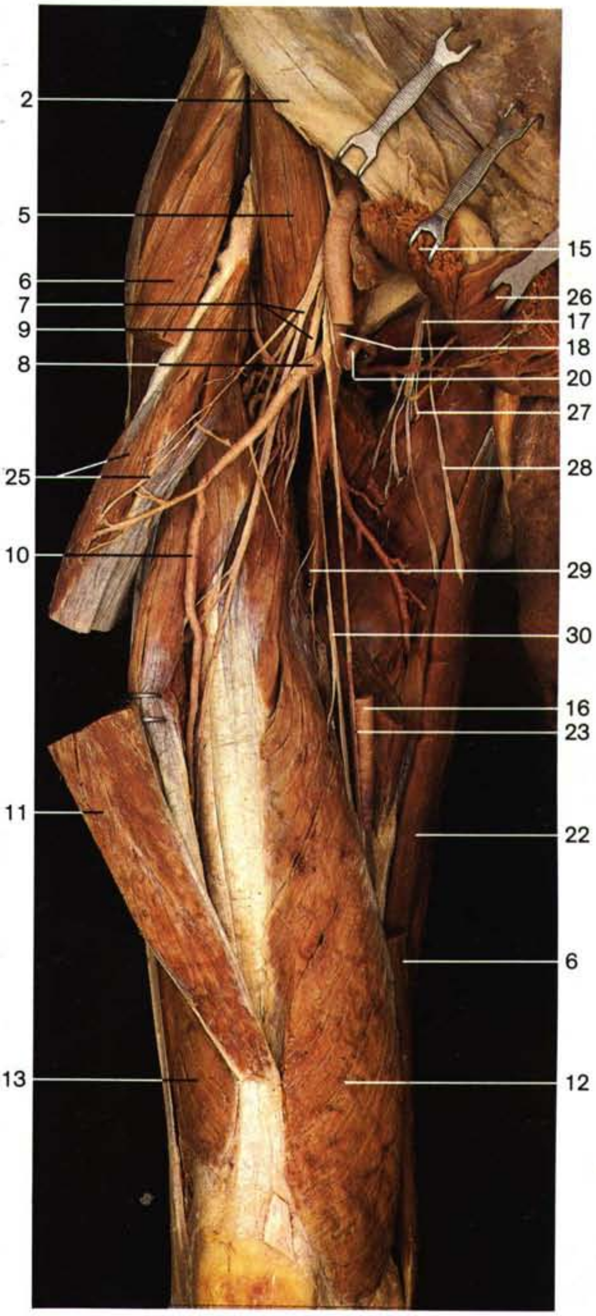

Anterior region of right thigh (anterior aspect). The sartorius, pectineus, adductor longus and rectus femoris muscles have been divided and reflected. The greater part of the femoral artery has been removed.

1 Anterior superior iliac spine
2 Inguinal ligament
3 Tensor fasciae latae muscle
4 Deep circumflex iliac artery
5 Iliopsoas muscle
6 Sartorius muscle (cut)
7 **Femoral nerve**
8 **Lateral circumflex femoral artery**
9 Ascending branch of lateral circumflex femoral artery
10 Descending branch of lateral circumflex femoral artery
11 Rectus femoris muscle
12 Vastus medialis muscle
13 Vastus lateralis muscle
14 Femoral vein
15 Pectineus muscle (cut)
16 Femoral artery (cut)
17 **Obturator nerve**
18 **Profunda femoris artery**
19 Ascending branch of medial circumflex femoral artery
20 **Medial circumflex femoral artery**
21 Adductor longus muscle
22 Gracilis muscle
23 Saphenous nerve
24 Distal part of vastoadductory lamina
25 Rectus femoris muscle with muscular branch of femoral nerve
26 Adductor longus muscle (divided)
27 Posterior branch of obturator nerve
28 Anterior branch of obturator nerve
29 Point at which perforating artery branches off from profunda femoris artery
30 Muscular branch to vastus medialis muscle

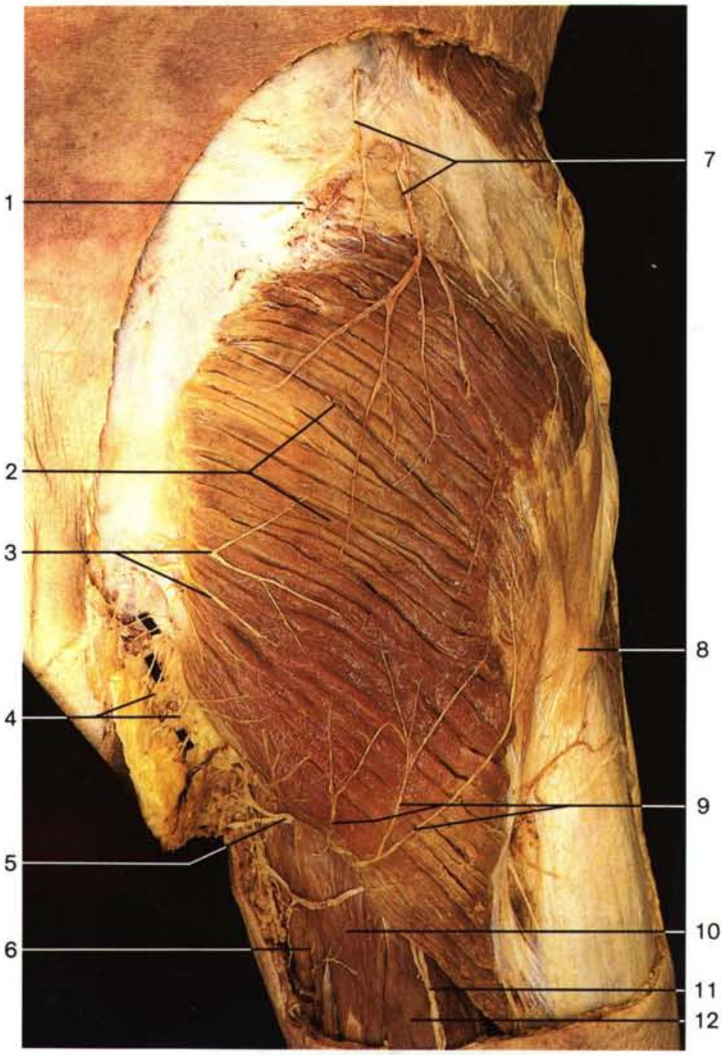

Gluteal region, right side (posterior aspect).

1 Iliac crest
2 Gluteus maximus muscle
3 **Middle cluneal nerves**
4 Anococcygeal nerves
5 Perineal branch of posterior femoral cutaneous nerve
6 Adductor magnus muscle
7 **Superior cluneal nerves**
8 Position of greater trochanter
9 **Inferior cluneal nerves**
10 Semitendinosus muscle
11 Posterior femoral cutaneous nerve
12 Long head of biceps femoris muscle

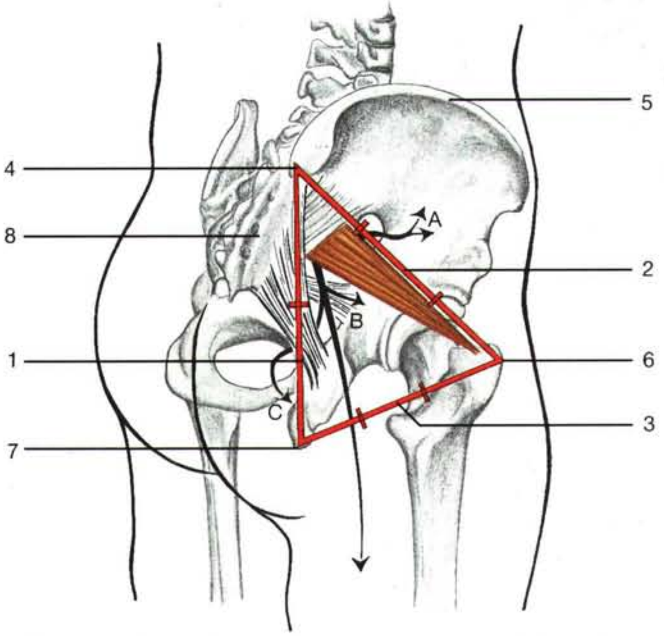

Gluteal region, right side (posterolateral aspect). Location of sciatic foramina in relation to the bones.
(Schematic drawing.)

A **Suprapiriform foramen**
of greater sciatic foramen

Superior gluteal artery, vein and nerve

B **Infrapiriform foramen**
of greater sciatic foramen

Sciatic nerve
Inferior gluteal artery, vein and nerve
Posterior femoral cutaneous nerve
Internal pudendal artery and vein
Pudendal nerve

C **Lesser sciatic foramen**

Pudendal nerve
Internal pudendal artery and vein

Red lines

1 **Spine-tuber line.**
In the middle of this line the infrapiriform foramen is situated
2 **Spine-trochanter line.**
In the upper third the suprapiriform foramen is located
3 **Tuber-trochanter line.**
Between the middle and posterior third, the ischiadic nerve can be found

Other structures

4 Posterior superior iliac spine
5 Iliac crest
6 Greater trochanter
7 Ischial tuberosity
8 Sacrum

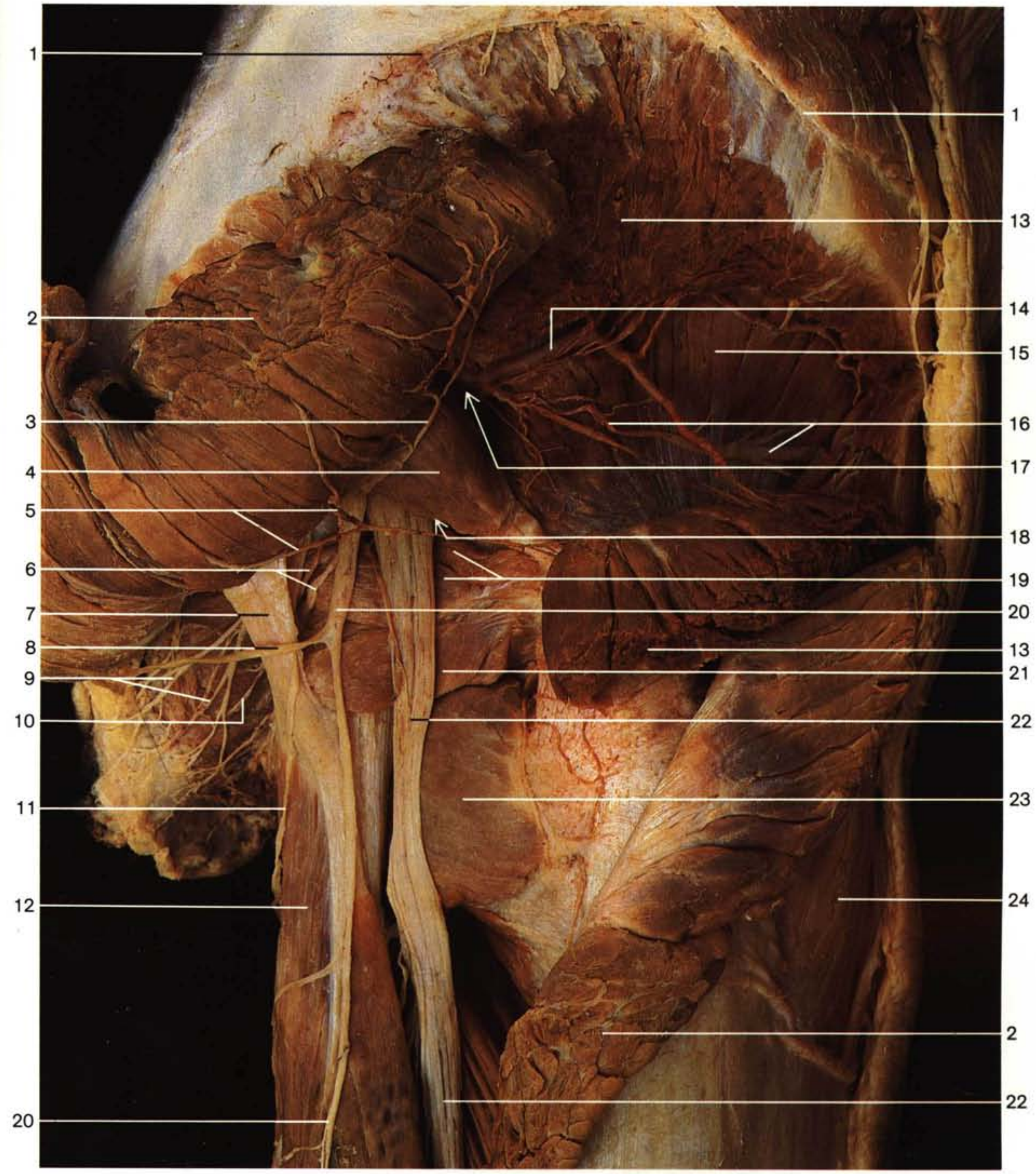

Gluteal region, right side (dorsal aspect). The gluteus maximus and gluteus medius muscles have been divided and reflected. Notice the position of the foramina above and below piriformis muscles and the lesser sciatic foramen.

1 Iliac crest
2 Gluteus maximus muscle (cut)
3 Inferior gluteal nerve
4 Piriformis muscle
5 Muscular branches of inferior gluteal artery
6 Pudendal nerve and internal pudendal artery within the **lesser sciatic foramen** (entrance to the pudendal canal)
7 Sacrotuberous ligament
8 Inferior cluneal nerve
9 Inferior rectal nerves
10 Inferior rectal arteries
11 Perforating cutaneous nerve
12 Long head of biceps femoris muscle

13 Gluteus medius muscle (cut)
14 Deep branch of superior gluteal artery
15 Gluteus minimus muscle
16 Superior gluteal nerve
17 **Suprapiriform foramen** } greater sciatic foramen
18 **Infrapiriform foramen** }
19 Tendon of obturator internus and superior gemellus muscles
20 Posterior femoral cutaneous nerve
21 Inferior gemellus muscle
22 **Sciatic nerve**
23 Quadratus femoris muscle
24 Tensor fasciae latae muscle

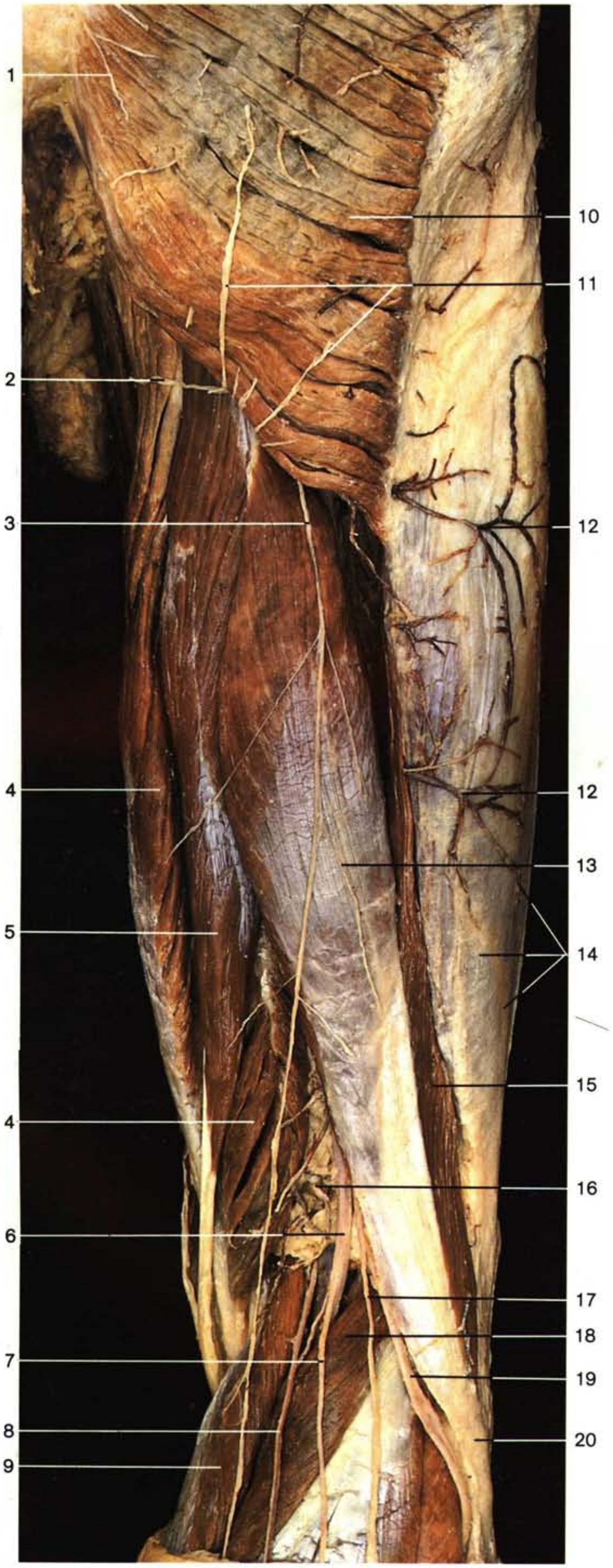

Cutaneous nerves of thigh (posterior aspect).
The fascia lata and the fasciae of muscles have been removed.

1 Middle cluneal nerves
2 Perineal branch of posterior femoral cutaneous nerve
3 **Posterior femoral cutaneous nerve**
4 Semimembranosus muscle
5 Semitendinosus muscle
6 **Tibial nerve**
7 Medial sural cutaneous nerve
8 **Small saphenous vein**
9 Medial head of gastrocnemius muscle
10 Gluteus maximus muscle
11 **Inferior cluneal nerve**
12 Cutaneous veins
13 Long head of biceps femoris muscle
14 Iliotibial tract
15 Short head of biceps femoris muscle
16 Popliteal fossa
17 Lateral sural cutaneous nerve
18 Lateral head of gastrocnemius muscle
19 **Common peroneal nerve**
20 Tendon of biceps femoris muscle
21 Inferior gluteal nerve
22 Sacrotuberous ligament
23 Inferior rectal branches of pudendal nerve
24 Anus
25 Gluteus medius muscle
26 Piriformis muscle
27 **Sciatic nerve**
28 Inferior gluteal artery
29 Gluteus maximus muscle (cut)
30 Quadratus femoris muscle
31 Sciatic nerve dividing into its two branches: the common peroneal nerve and the tibial nerve
32 Muscular branches of sciatic nerve to hamstring muscles
33 **Popliteal artery**
34 **Popliteal vein**
35 Small saphenous vein (cut)
36 Long head of biceps femoris muscle (cut)
37 **Superficial peroneal nerve**

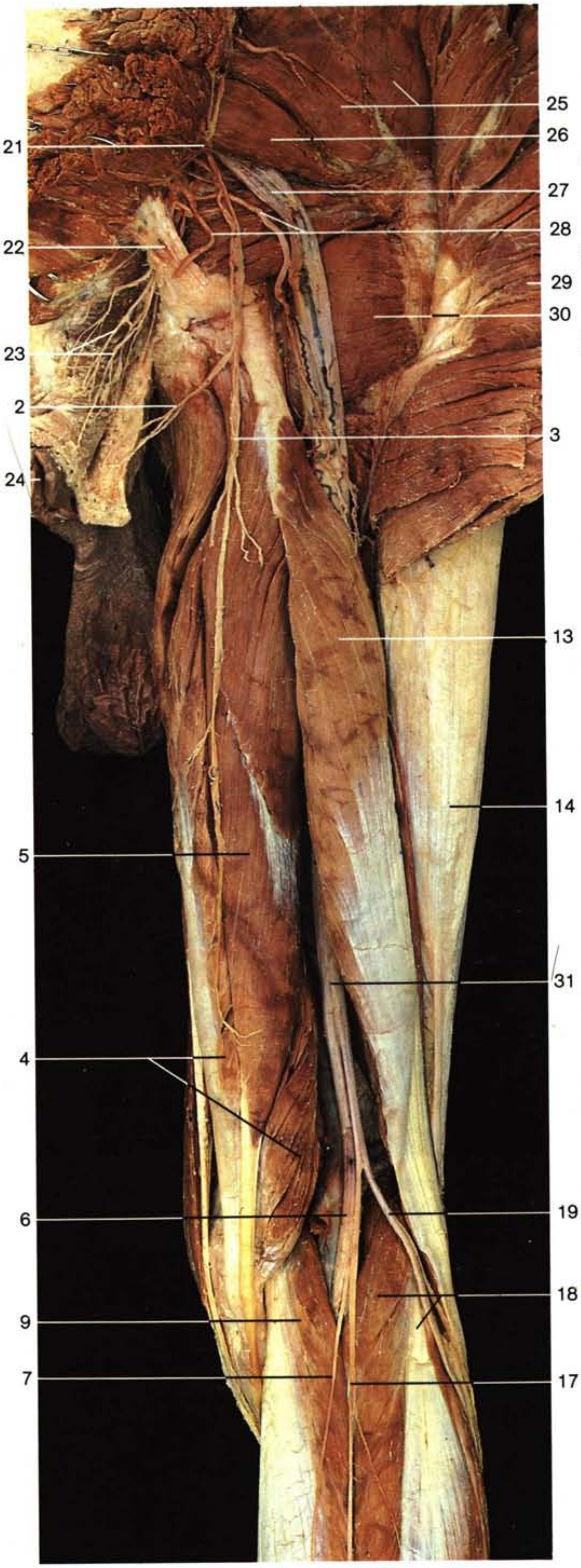
25
21
26
27
28
22
29
30
23
2
3
24
13
14
5
31
4
6
19
18
9
7
17

Posterior femoral region and gluteal region, right side (posterior aspect). The gluteus maximus muscle has been divided and reflected.

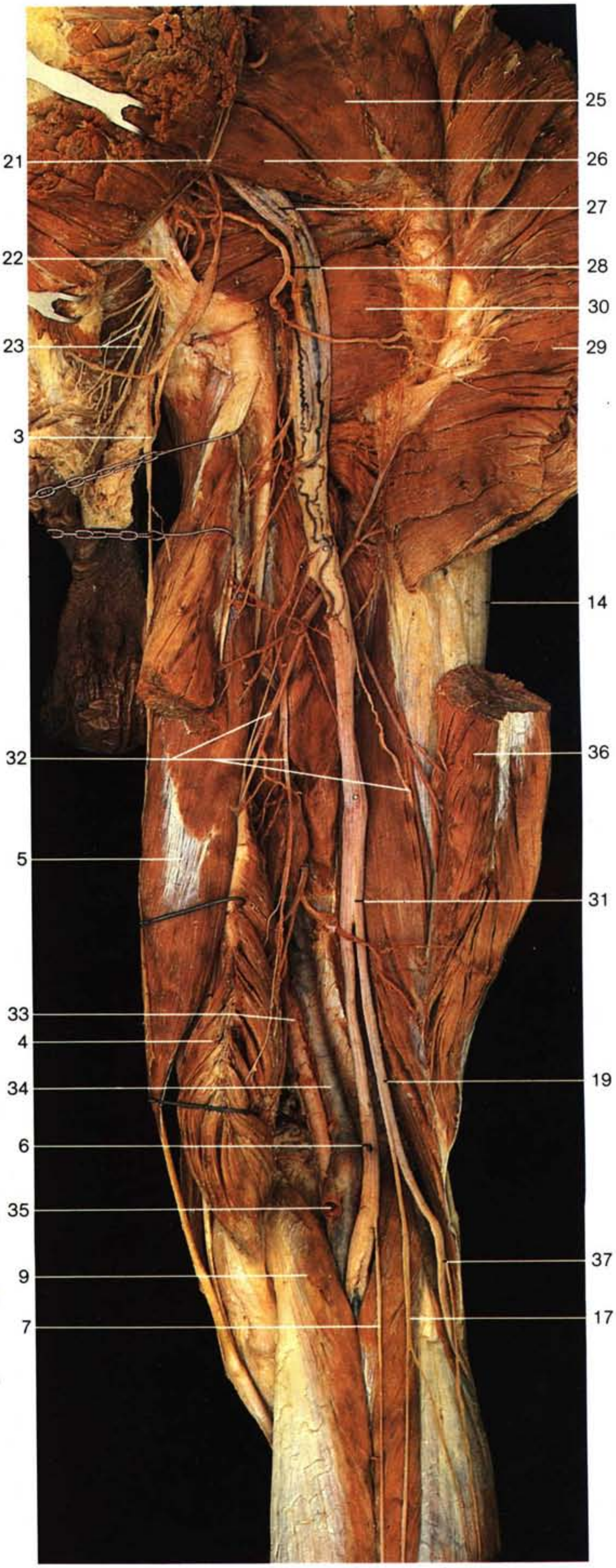
25
21
26
27
22
28
30
23
29
3
14
32
36
5
31
33
4
34
19
6
35
37
9
17
7

Posterior femoral region and gluteal region, right side (posterior aspect). The gluteus maximus muscle and the long head of biceps femoris muscle have been divided and reflected.

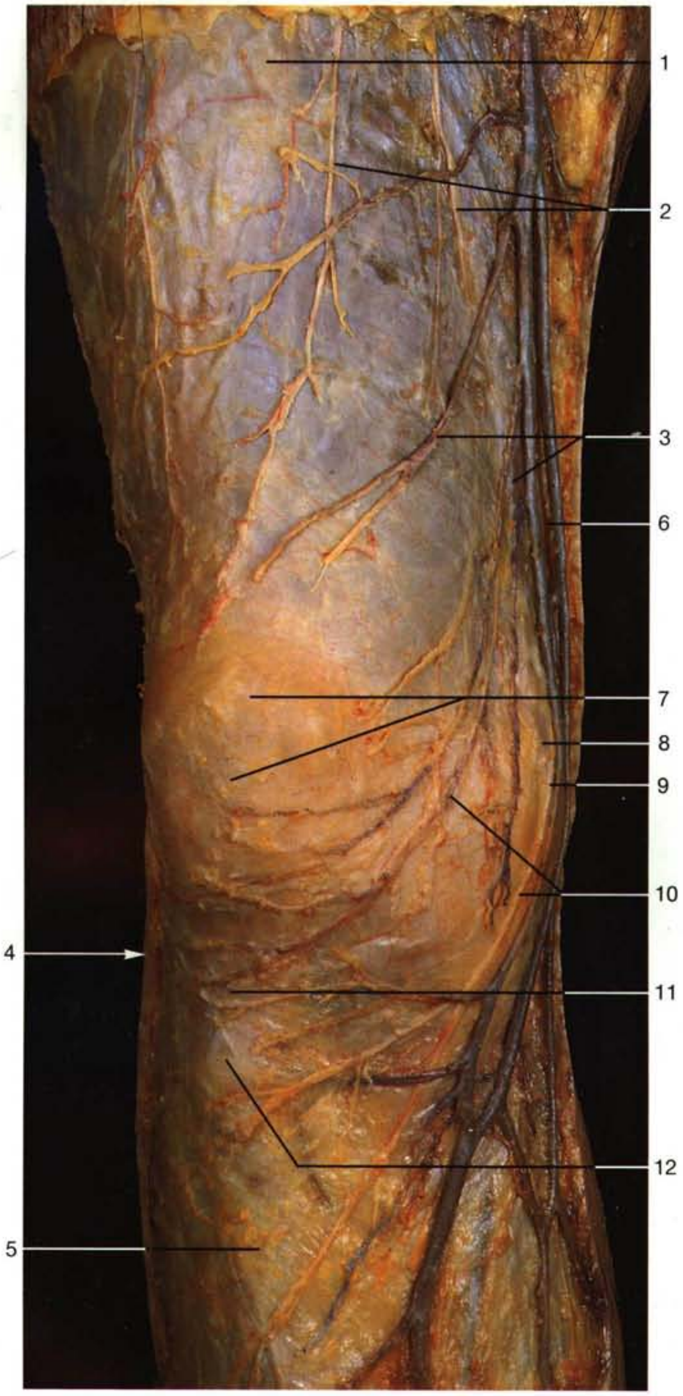

Anterior region of right knee; cutaneous nerves and veins (anterior aspect).

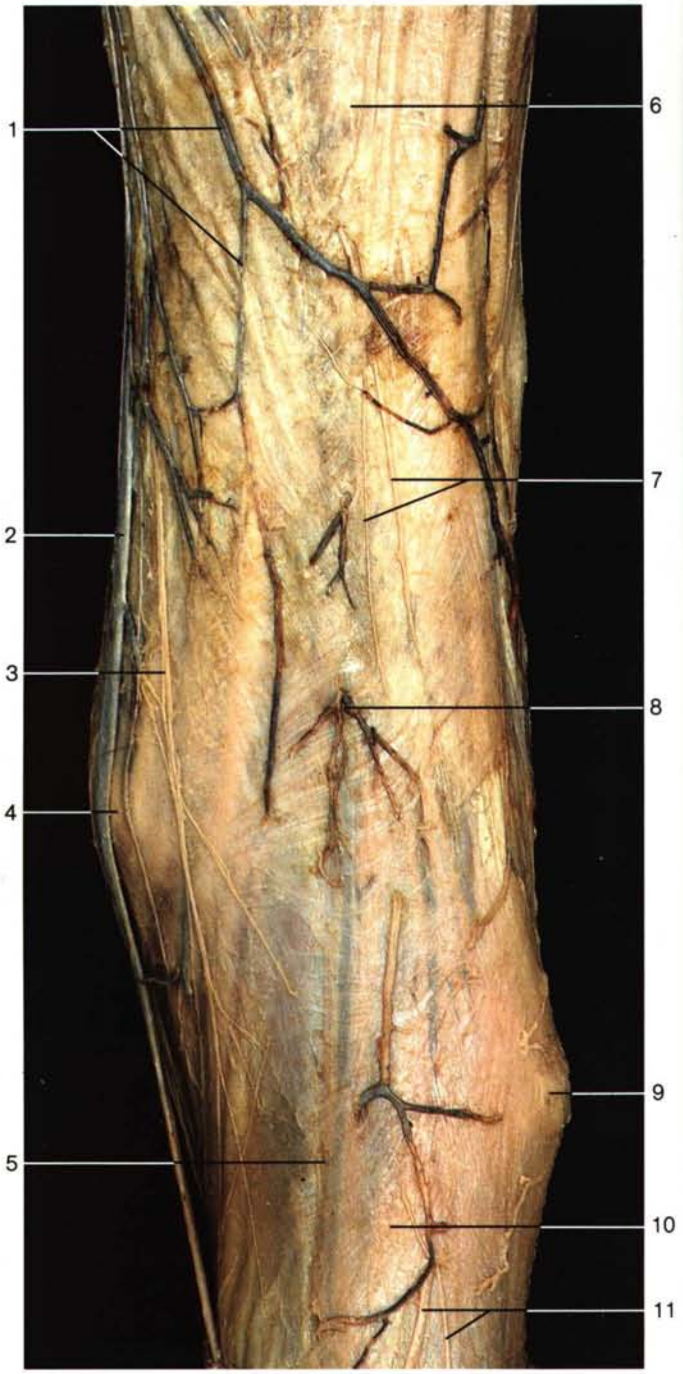

Posterior region of right kneee; cutaneous nerves and veins (posterior aspect).

1 Fascia lata
2 Terminal branches of anterior cutaneous branches of femoral nerve
3 Venous network around knee
4 Position of head of fibula
5 Superficial crural fascia
6 **Great saphenous vein**
7 Patella
8 Position of medial epicondyle of femur
9 Saphenous nerve
10 **Infrapatellar branches of saphenous nerve**
11 Patellar ligament
12 Position of tuberosity of tibia

1 Cutaneous veins (tributaries of great saphenous vein)
2 **Great saphenous vein**
3 Cutaneous branch of femoral nerve
4 Position of medial epicondyle of femur
5 Position of small saphenous vein
6 Fascia lata
7 Terminal branches of **posterior femoral cutaneous nerve**
8 Cutaneous veins of popliteal fossa
9 Position of head of fibula
10 Superficial layer of fascia cruris
11 Lateral sural cutaneous nerve

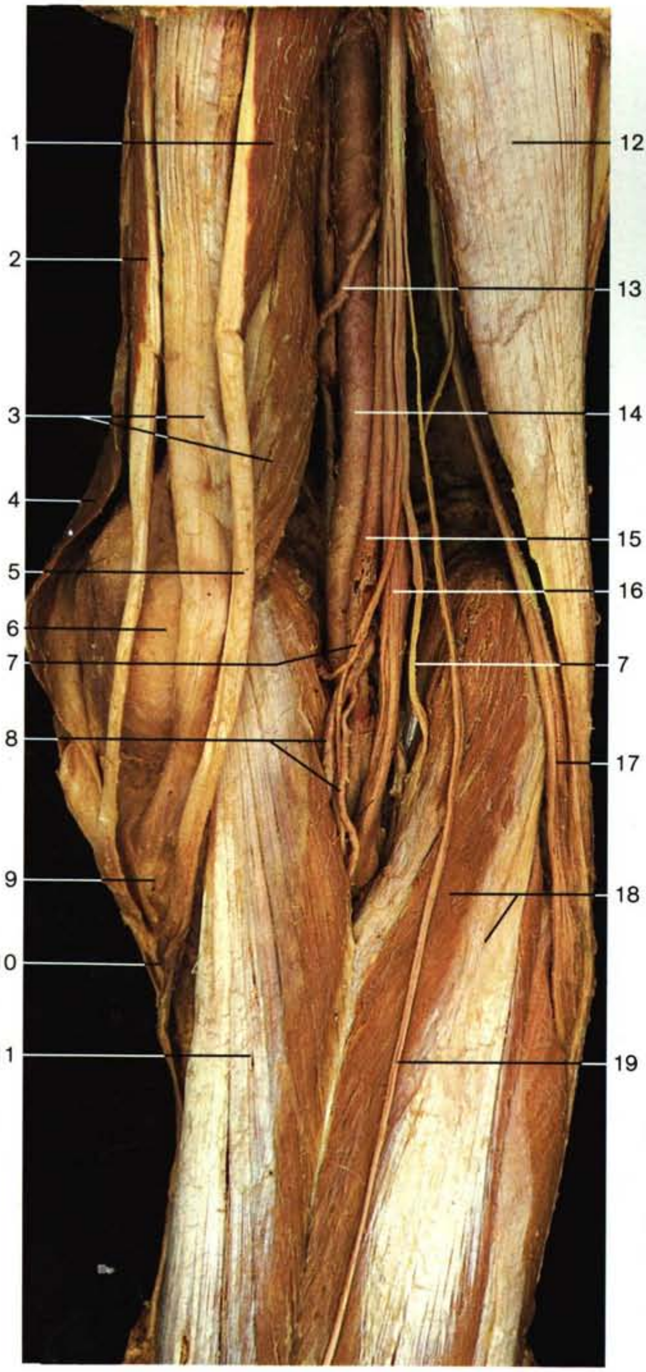

Right leg, posterior crural region (posterior aspect). The gastrocnemius muscle has been divided and reflected.

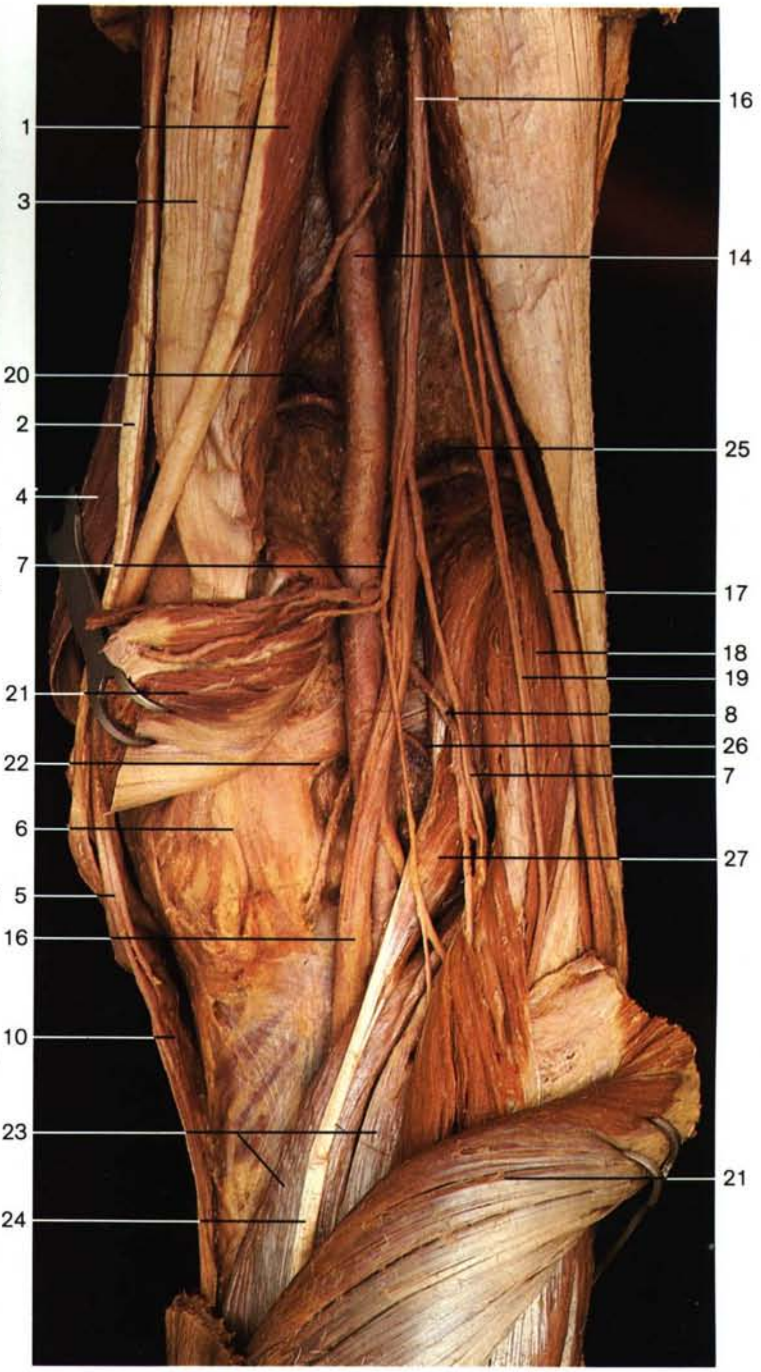

Right leg, posterior crural region, deep layer (posterior aspect). The gastrocnemius and the soleus muscles have been divided and reflected.

1 Semitendinosus muscle
2 Gracilis muscle
3 Semimembranosus muscle
4 Sartorius muscle
5 Tendon of semitendinosus muscle
6 Position of medial condyle of femur
7 Muscular branches of tibial nerve
8 Sural arteries and veins
9 Tendon of semimembranosus muscle
10 Common tendon of gracilis, semitendinosus and sartorius muscles
11 Medial head of gastrocnemius muscle
12 Biceps femoris muscle
13 Muscular branch of popliteal artery
14 **Popliteal artery**
15 **Popliteal vein**
16 **Tibial nerve**
17 **Common peroneal nerve**
18 Lateral head of gastrocnemius muscle
19 Medial sural cutaneous nerve
20 Medial superior genicular artery
21 Medial head of gastrocnemius muscle (cut and reflected)
22 Medial inferior genicular artery
23 Soleus muscle
24 Tendon of plantaris muscle
25 Lateral superior genicular artery
26 Lateral inferior genicular artery
27 Plantaris muscle

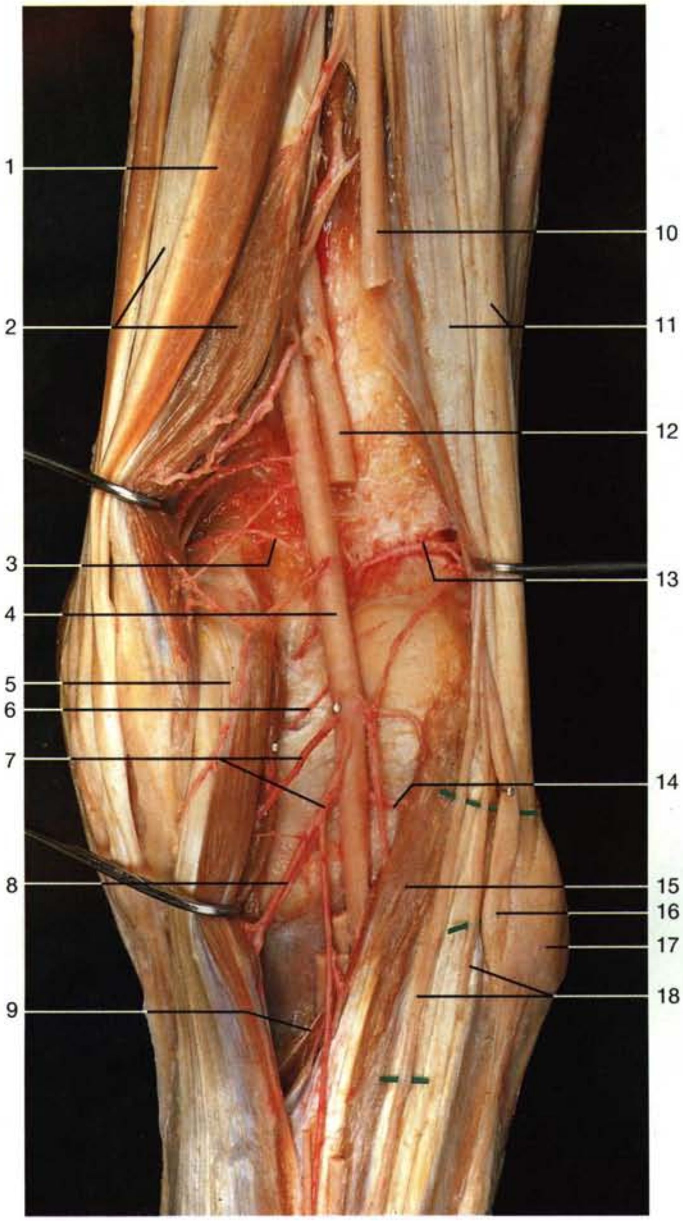

Right leg, popliteal fossa, deep layer (posterior aspect). The muscles have been reflected to display the genicular arteries.

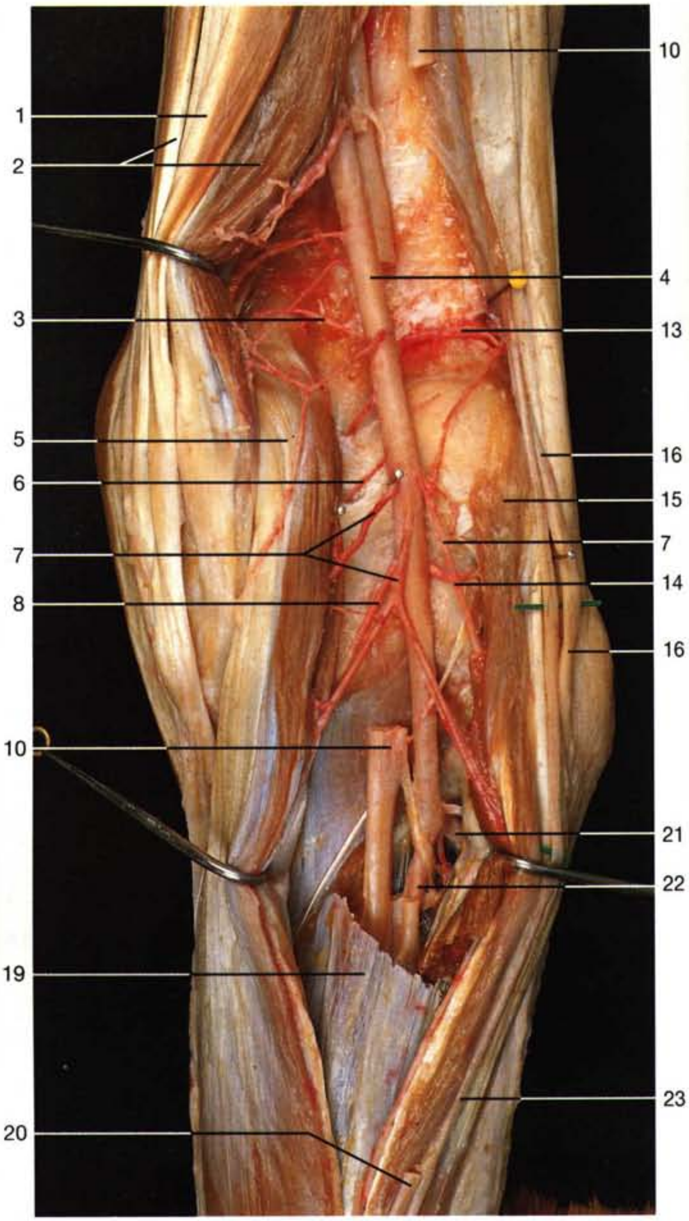

Right leg, popliteal fossa, deepest layer (posterior aspect). Tibial nerve and popliteal vein have been partly removed and a portion of the soleus muscle was cut away to display the anterior tibial artery.

1 Semitendinosus muscle
2 Semimembranosus muscle
3 Medial superior genicular artery
4 **Popliteal artery**
5 Medial head of gastrocnemius muscle
6 Middle genicular artery
7 Muscular branches
8 Medial inferior genicular artery
9 Tendon of plantaris muscle
10 **Tibial nerve** (cut)
11 Biceps femoris muscle
12 Popliteal vein (cut)
13 Lateral superior genicular artery
14 Lateral inferior genicular artery
15 Lateral head of gastrocnemius muscle
16 **Common peroneal nerve**
17 Head of fibula
18 Lateral sural cutaneous nerves
19 Soleus muscle
20 Medial sural cutaneous nerve
21 **Anterior tibial artery**
22 **Posterior tibial artery**
23 Lateral sural cutaneous nerve

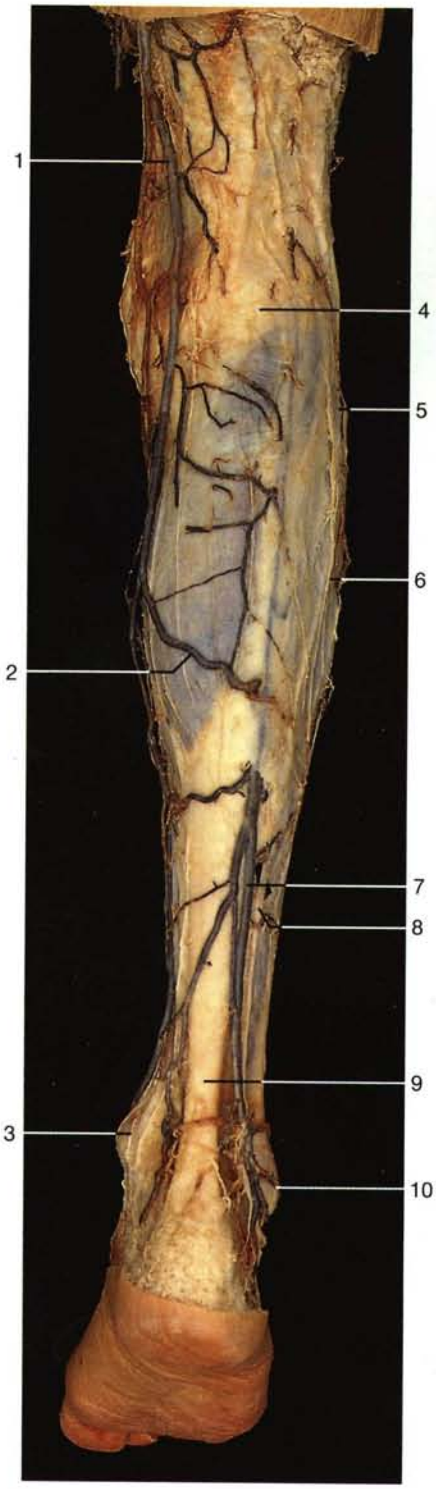

Right leg. Cutaneous veins and nerves (posterior aspect).

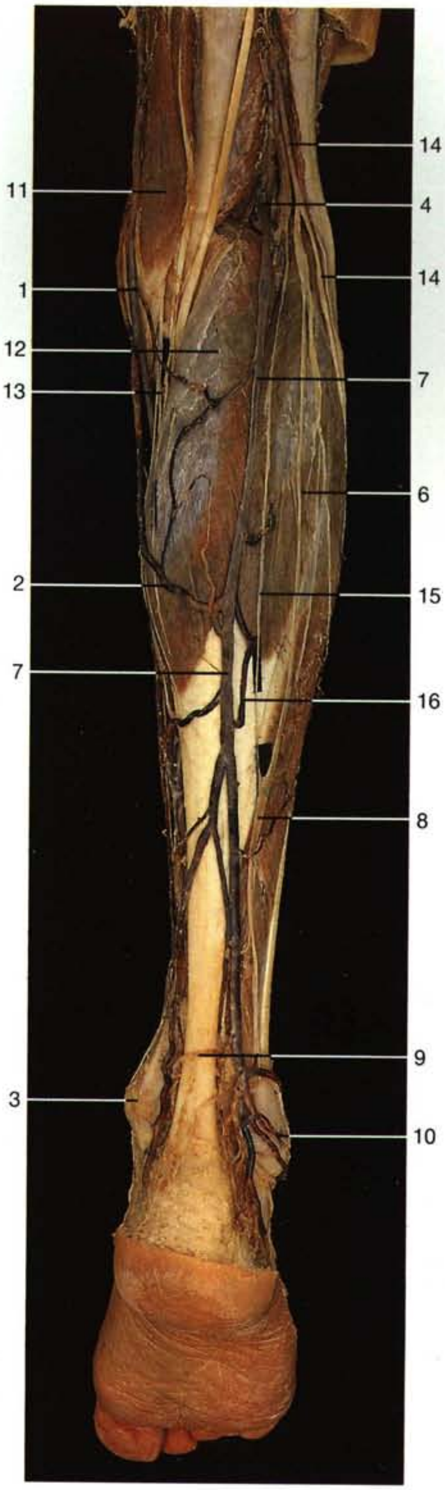

Right leg. Cutaneous nerves and veins (posterior aspect). The superficial layer of the crural fascia has been removed.

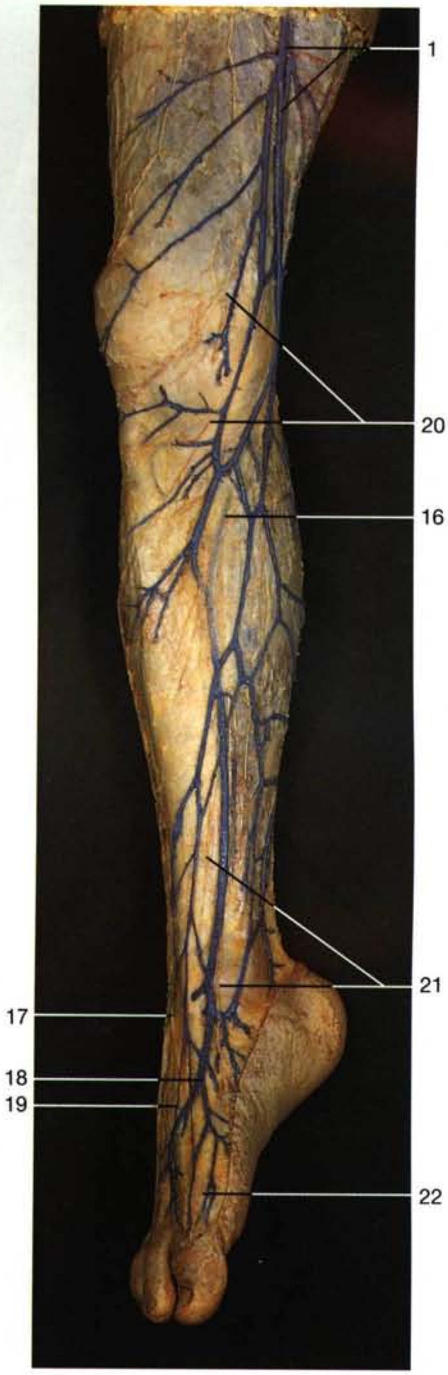

Cutaneous veins and nerves of the right leg (anterior-medial aspect; veins are colored).

1 **Great saphenous vein**
2 Venous anastomosis between small and great saphenous veins
3 Medial malleolus
4 Popliteal fossa
5 Position of head of fibula
6 Lateral sural cutaneous nerve
7 **Small saphenous vein**
8 **Sural nerve**
9 Calcaneal tendon
10 Lateral malleolus
11 Semitendinosus muscle
12 Medial head of gastrocnemius muscle
13 Saphenous nerve
14 **Common peroneal nerve**
15 Medial sural cutaneous nerve
16 Perforating veins
17 Superficial peroneal nerve
18 Dorsal venous arch
19 Intermediate dorsal cutaneous nerve
20 **Infrapatellar branches of saphenous nerve**
21 Terminal branches of saphenous nerve
22 Medial dorsal cutaneous nerve

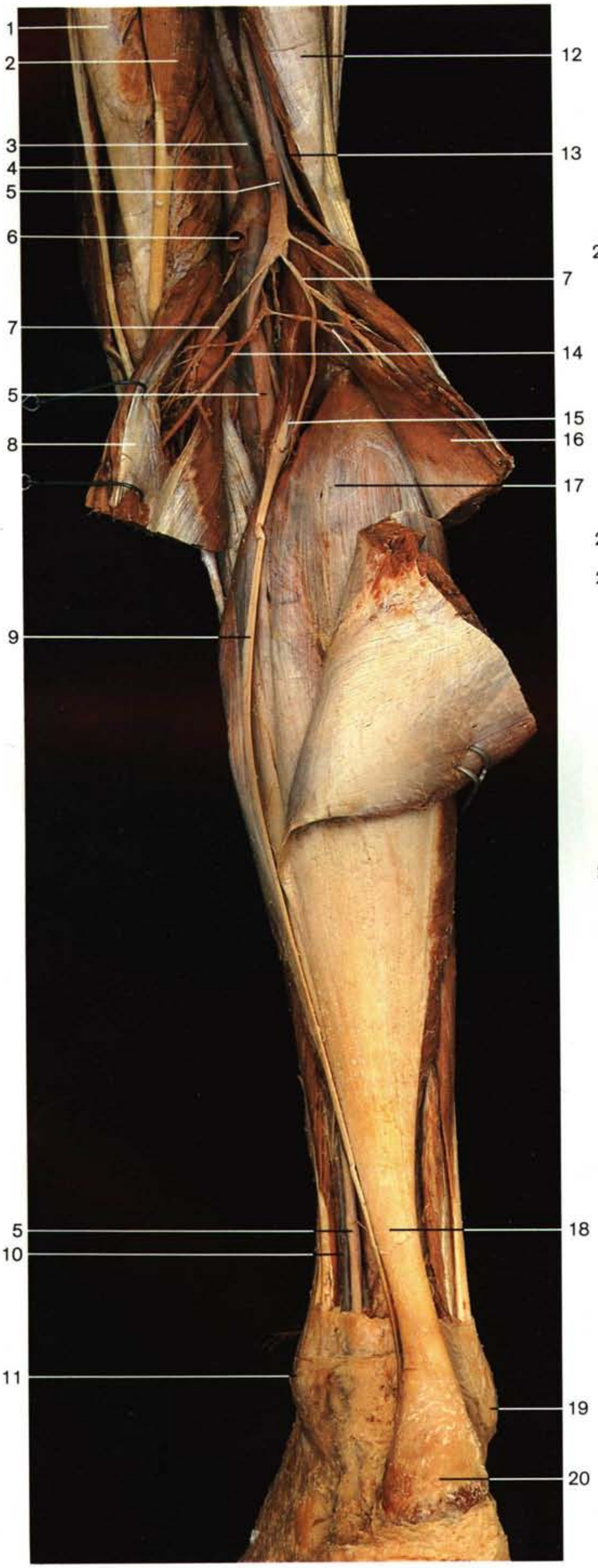
1
12
2
3
13
4
5
6
7
7
14
5
15
8
16
17
9
5
18
10
11
19
20

Right popliteal fossa, middle layer (posterior aspect).
The cutaneous veins and nerves have been removed.

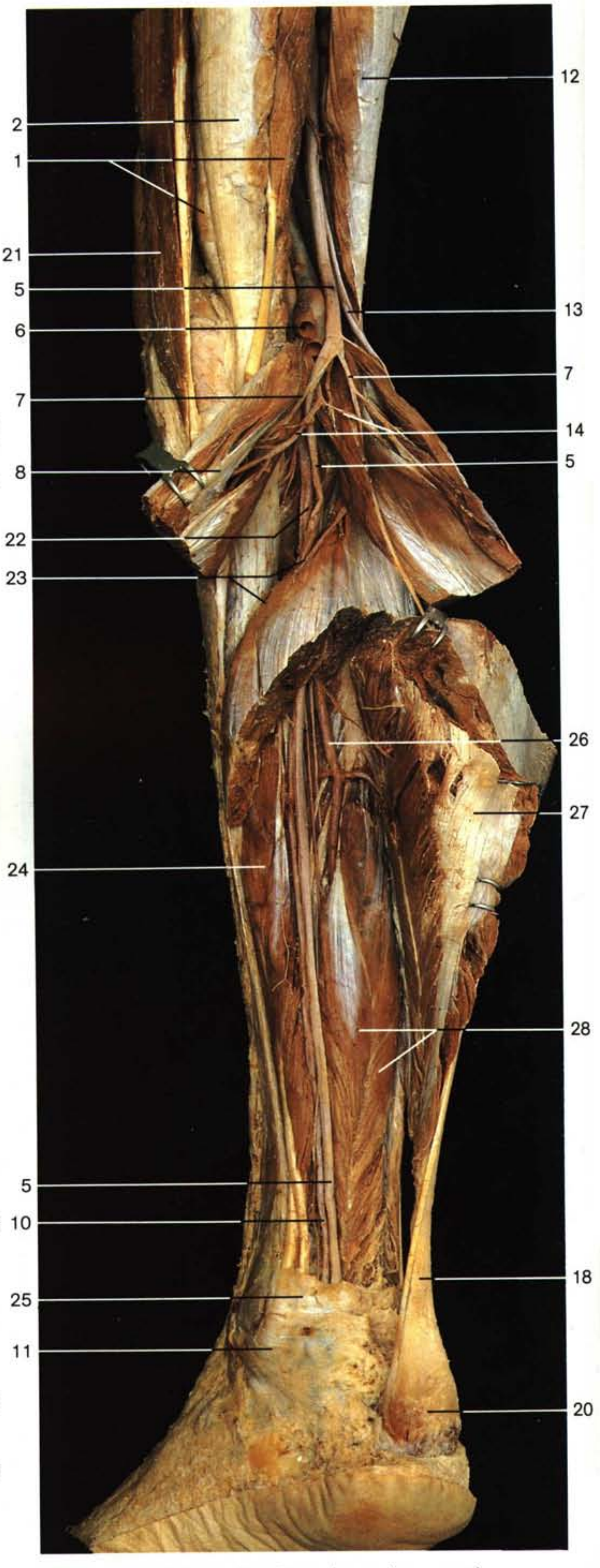
12
2
1
21
5
13
6
7
7
14
8
5
22
23
26
27
24
28
5
10
18
25
11
20

Right popliteal fossa, deep layer (posterior aspect).
The medial head of gastrocnemius muscle has been divided and reflected.

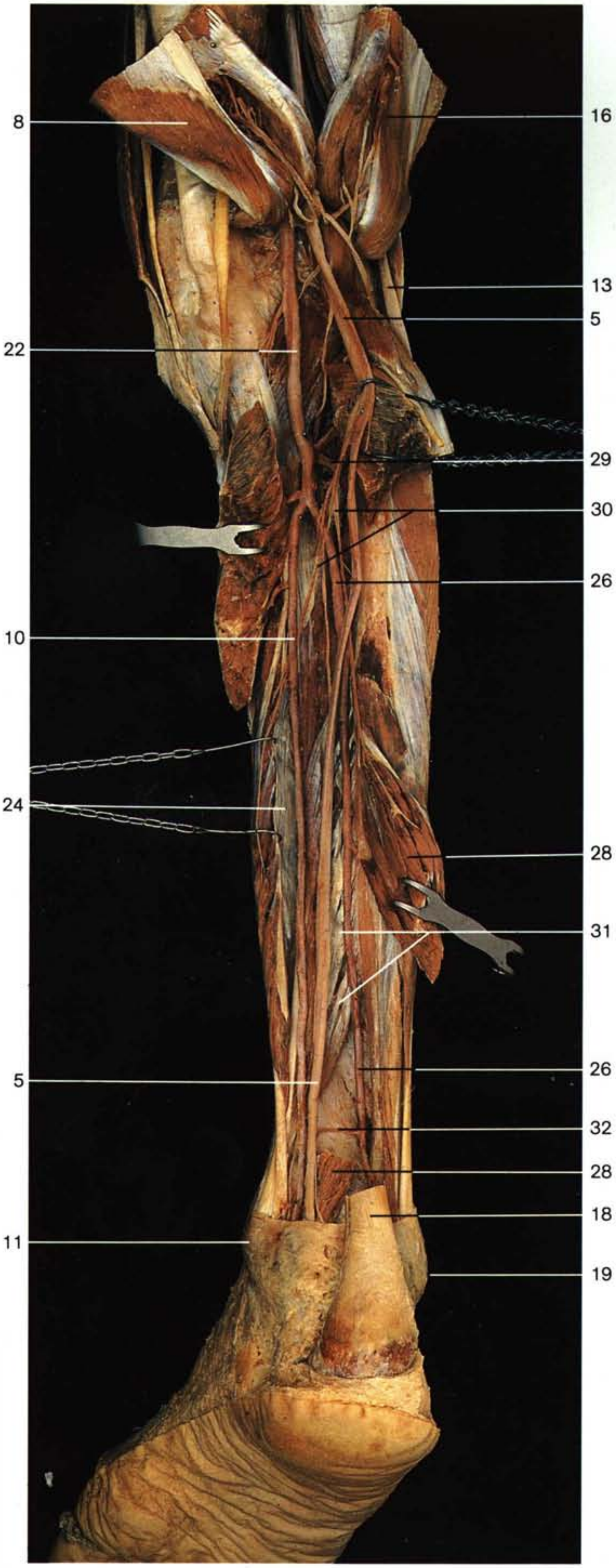

Right leg, posterior crural region, deepest layer (posterior aspect). Triceps surae (gastrocnemius and soleus) and flexor hallucis longus muscles have been cut and reflected.

1 Semimembranosus muscle
2 Semitendinosus muscle
3 **Popliteal vein**
4 Popliteal artery
5 **Tibial nerve**
6 Small saphenous vein (cut)
7 Muscular branch of tibial nerve
8 Medial head of gastrocnemius muscle
9 Tendon of plantaris muscle
10 **Posterior tibial artery**
11 Medial malleolus
12 Biceps femoris muscle
13 Common peroneal nerve
14 **Sural arteries**
15 Plantaris muscle
16 Lateral head of gastrocnemius muscle
17 Soleus muscle
18 Calcaneal tendon
19 Lateral malleolus
20 Calcaneal tuberosity
21 Sartorius muscle
22 **Popliteal artery**
23 Tendinous arch of soleus muscle
24 Flexor digitorum longus muscle
25 Flexor retinaculum
26 **Peroneal artery**
27 Soleus muscle
28 Flexor hallucis longus muscle
29 **Anterior tibial artery**
30 Muscular branches of tibial nerve
31 Tibialis posterior muscle
32 Communicating branch of peroneal artery
33 Tendon of tibialis anterior muscle
34 Tibia
35 Tendon of extensor hallucis longus muscle
36 Tendons of extensor digitorum longus muscle
37 Anterior tibialis artery
38 Fibula
39 Tendons of peroneus longus and brevis muscles

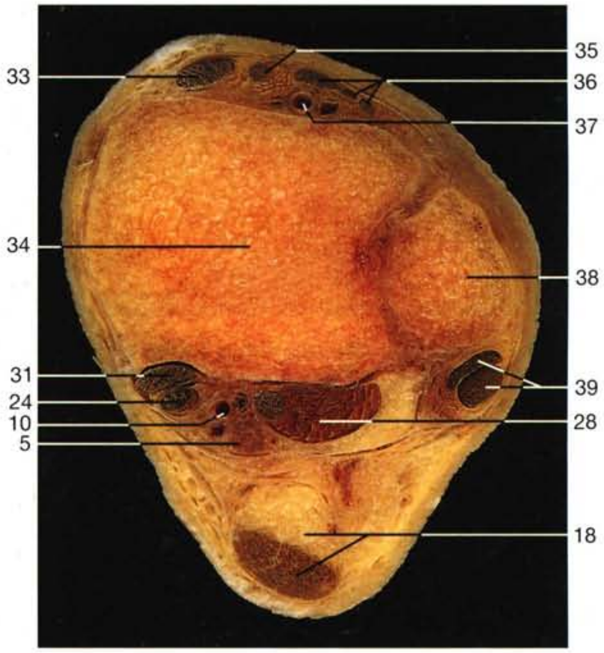

Cross-section of the leg, superior to the malleoli (from below).

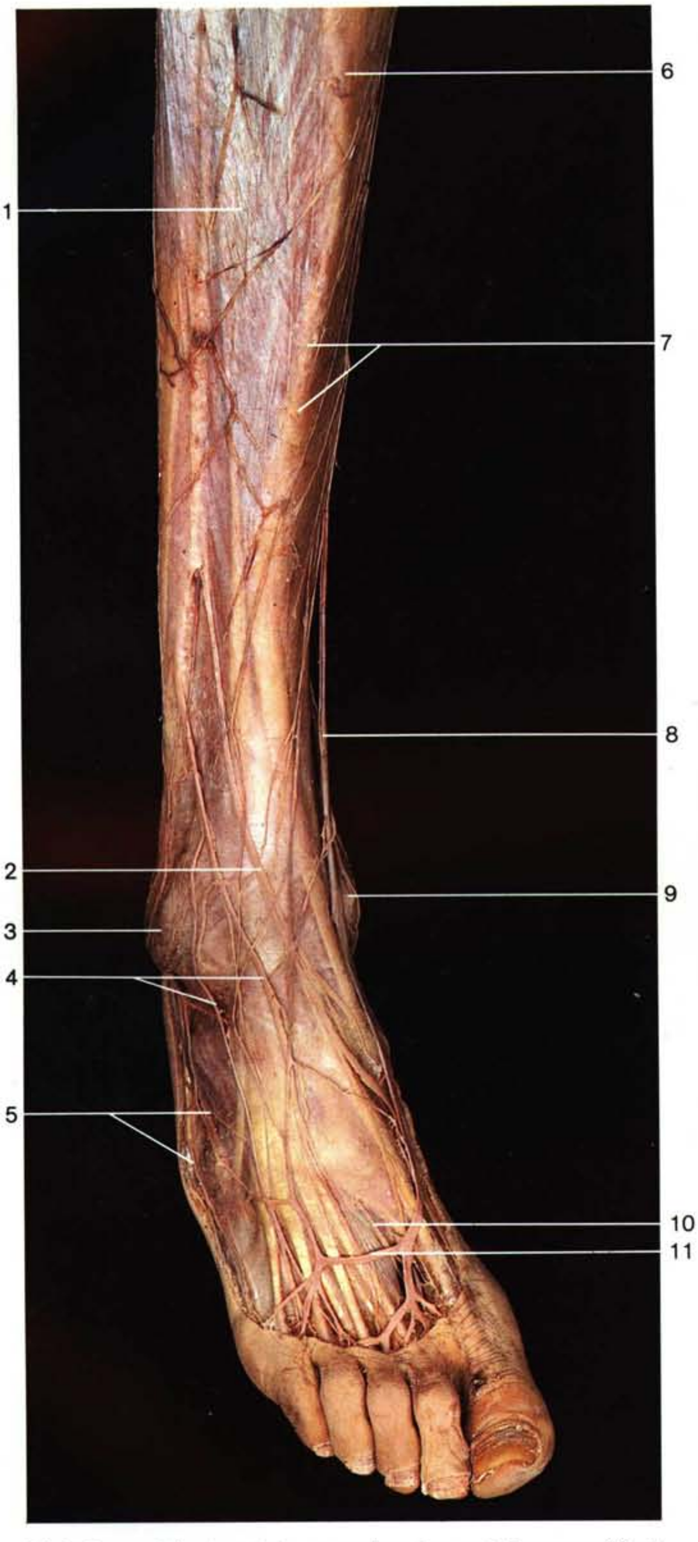

Right leg and foot, anterior crural region and dorsum of foot; cutaneous nerves and veins (anterior aspect).

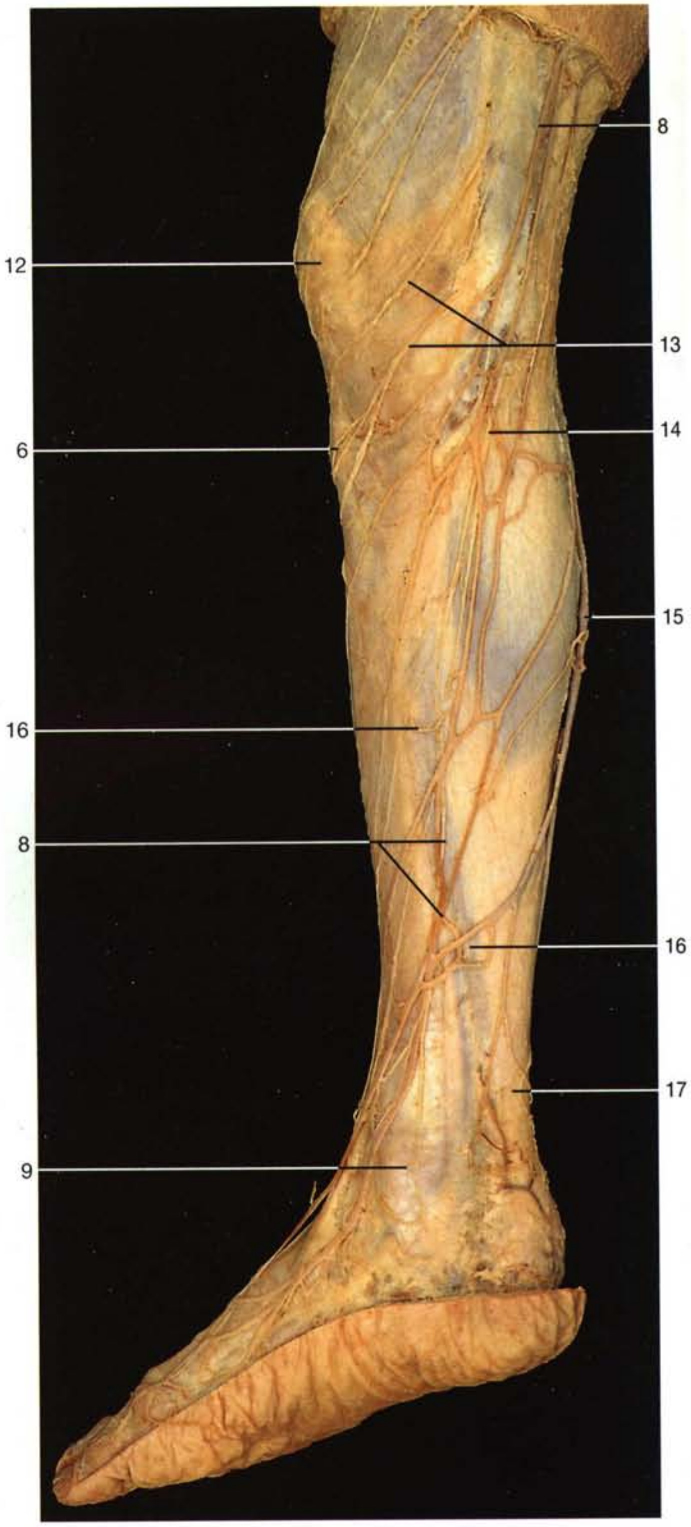

Right leg and foot; cutaneous nerves and veins (medial aspect).

1 Superficial crural fascia
2 Medial cutaneous branch of superficial peroneal nerve
3 Lateral malleolus
4 Lateral cutaneous branch of superficial peroneal nerve
5 Cutaneous branch of sural nerve
6 Position of tuberosity of tibia
7 Anterior margin of tibia
8 **Great saphenous vein**
9 Medial malleolus

10 Deep peroneal nerve
11 **Venous arch of dorsum of foot**
12 Position of patella
13 **Infrapatellar branches of saphenous nerve**
14 Saphenous nerve
15 Small saphenous vein
16 Perforating vein
17 Calcaneal tendon

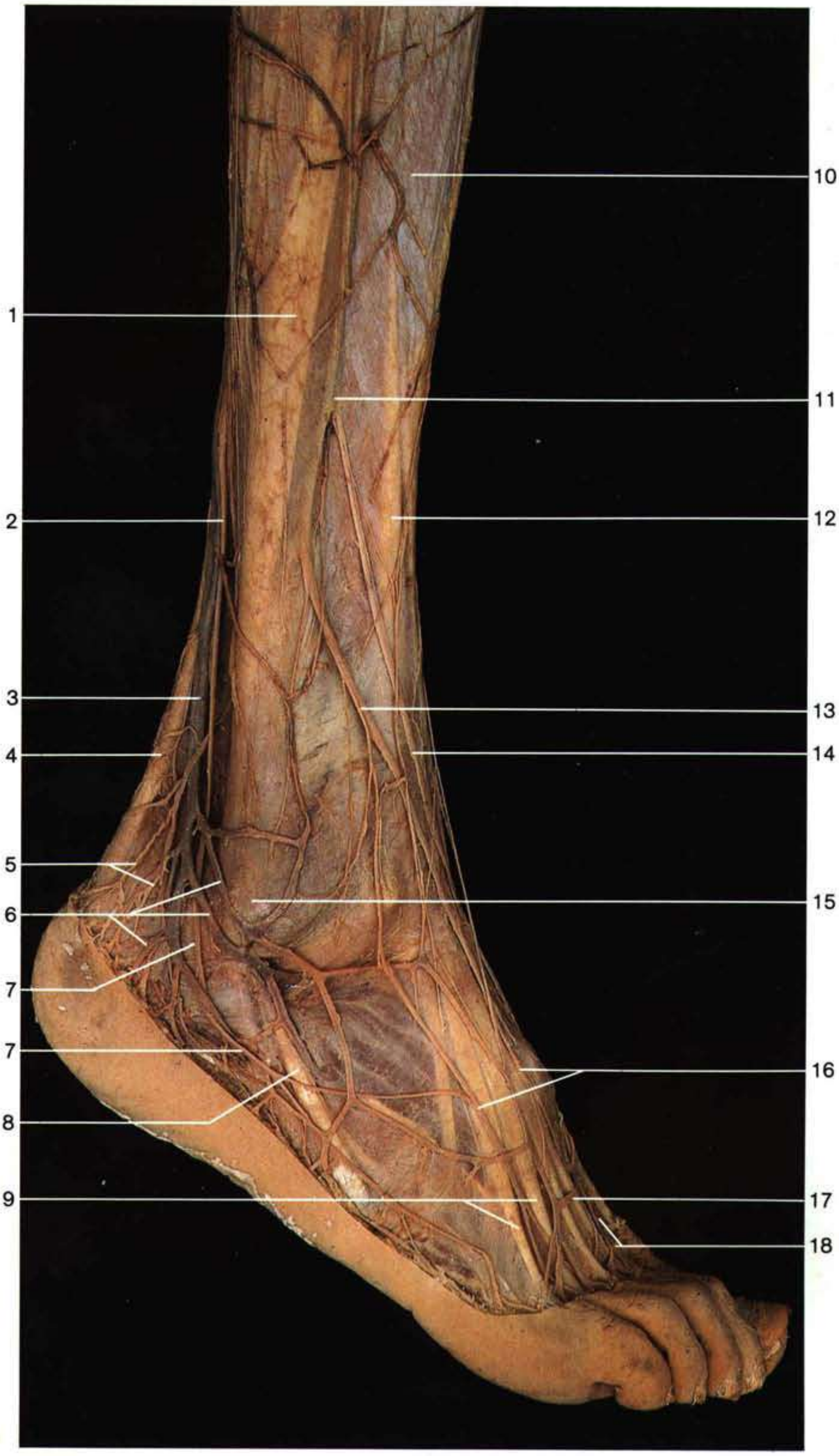

Right leg and dorsum of foot; cutaneous nerves and veins (lateral aspect).

1	Position of fibula	10	Fascia cruris
2	**Sural nerve**	11	Superficial peroneal nerve
3	**Small saphenous vein**	12	Position of tibia
4	Calcaneal tendon	13	Lateral cutaneous branch } of **superficial peroneal nerve**
5	Lateral calcaneal branches of sural nerve	14	Medial cutaneous branch } of **superficial peroneal nerve**
6	Venous network at lateral malleolus	15	Lateral malleolus
7	Cutaneous branch of sural nerve	16	Dorsal digital nerves
8	Tendon of peroneus brevis muscle	17	Dorsal venous arch
9	Tendons of extensor digitorum longus muscle	18	Deep peroneal nerve

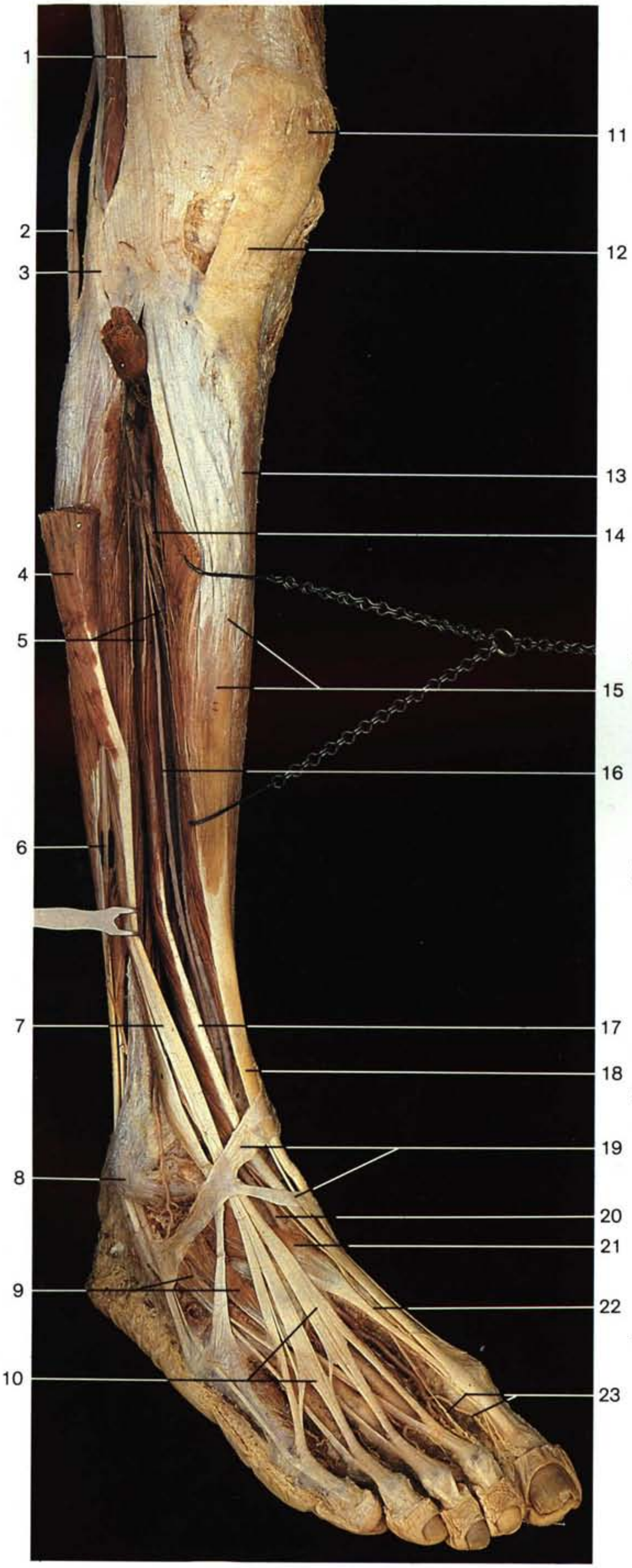
1
11
2
12
3
13
14
4
5
15
16
6
7
17
18
8
19
20
21
9
22
10
23

Right leg and dorsum of foot, middle layer (anterior lateral aspect). The extensor digitorum longus muscle has been divided and reflected laterally.

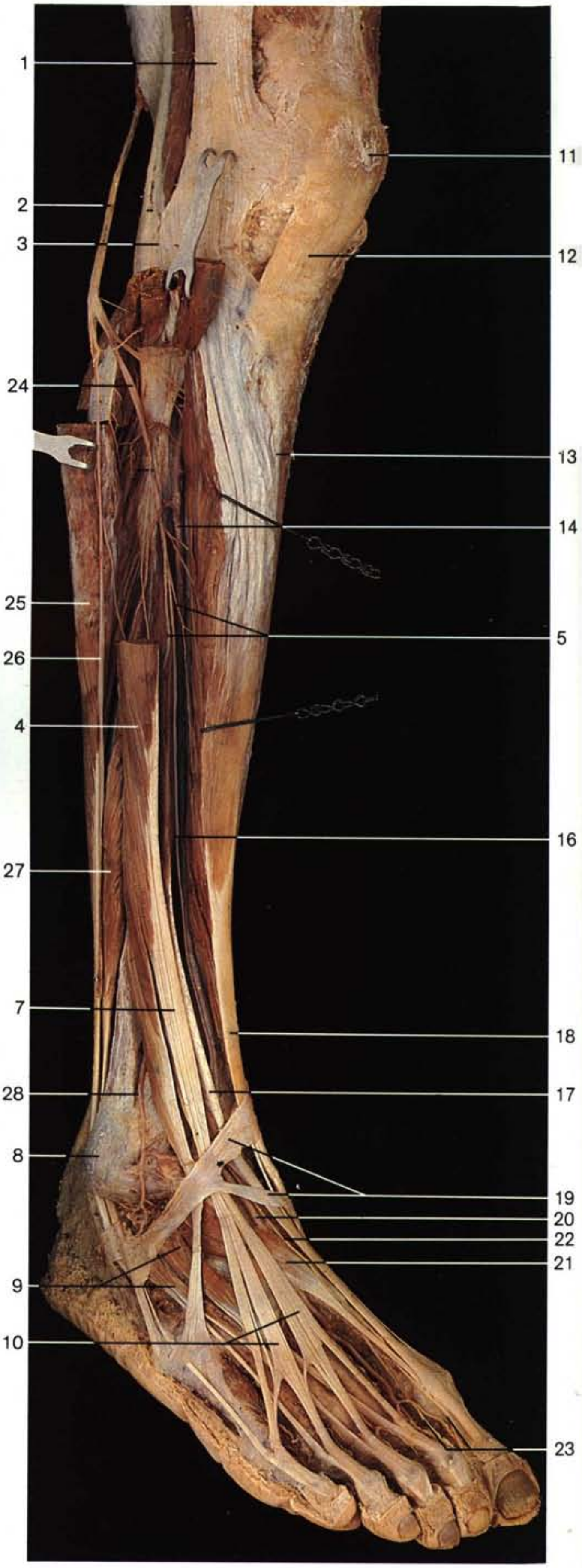
1
11
2
3
12
24
13
14
25
5
26
4
16
27
7
18
28
17
8
19
20
22
21
9
10
23

Right leg, deep layer (anterior lateral aspect). The extensor digitorum longus and the peroneus longus muscle have been divided or removed. The common peroneal nerve has been elevated to show its course around the head of fibula.

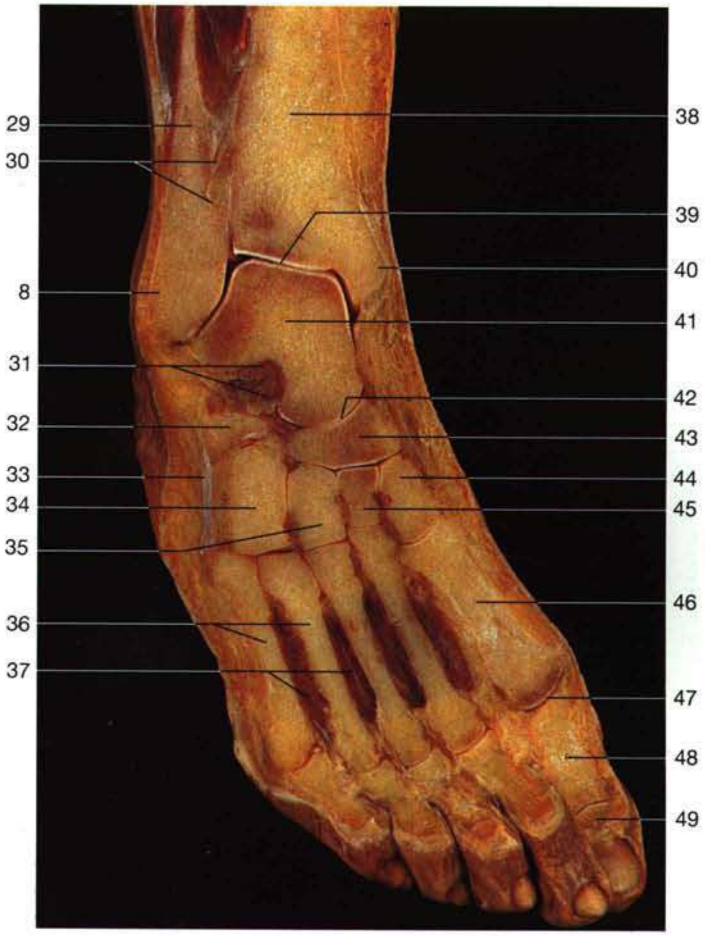

Coronal section through the foot and talocrural joint (anterior aspect).

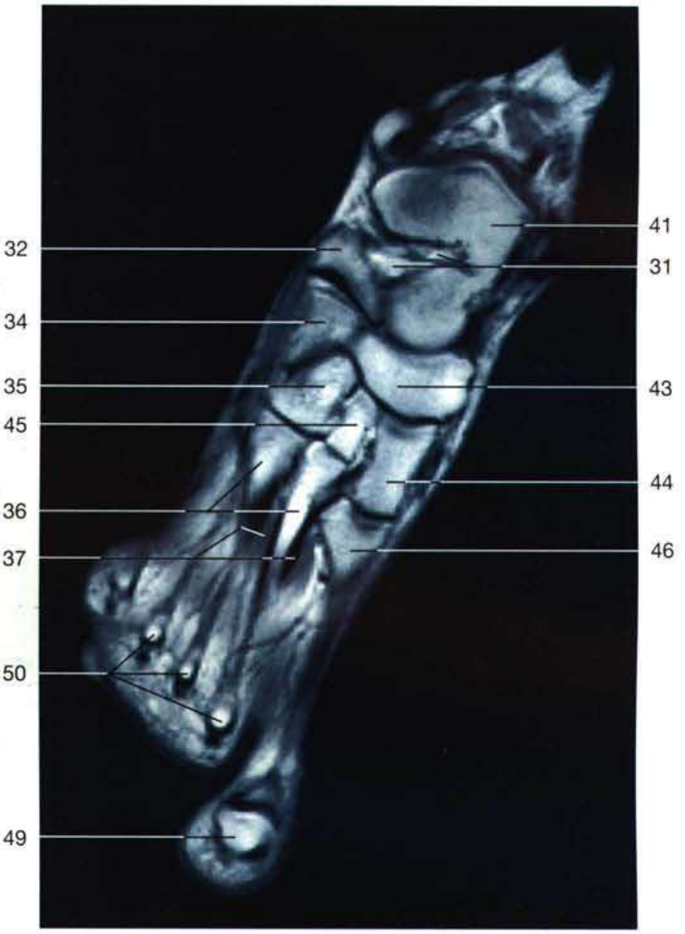

Coronal section through the foot (MR-Scan after A. Heuck, G. Luttke and J. W. Rohen, 1994).

1 Iliotibial tract
2 **Common peroneal nerve**
3 Position of head of fibula
4 Extensor digitorum longus muscle
5 Muscular branches of deep peroneal nerve
6 **Superficial peroneal nerve**
7 Tendon of extensor digitorum longus muscle
8 Lateral malleolus
9 Extensor digitorum brevis muscle
10 Tendons of extensor digitorum longus muscle
11 Patella
12 Patellar ligament
13 Anterior margin of tibia
14 **Anterior tibial artery**
15 Tibialis anterior muscle
16 Deep peroneal nerve
17 Extensor hallucis longus muscle
18 Tendon of tibialis anterior muscle
19 Extensor retinaculum
20 Dorsalis pedis artery
21 Extensor hallucis brevis muscle
22 Deep peroneal nerve (on dorsum of foot)
23 Dorsal digital nerves (terminal branches of deep peroneal nerve)
24 **Deep peroneal nerve**
25 Peroneus longus muscle (cut)
26 Superficial peroneal nerve (with peroneal muscles laterally reflected)
27 Peroneus brevis muscle
28 Lateral anterior malleolar artery
29 Fibula
30 Distal tibiofibular joint (syndesmosis)
31 Interosseous talocalcaneal ligament
32 Calcaneus
33 Tendon of peroneus brevis muscle
34 Cuboid bone
35 Lateral cuneiform bone
36 Metatarsal bones
37 Dorsal interosseous muscles
38 Tibia
39 **Talocrural joint**
40 Medial malleolus
41 Talus
42 **Talocalcaneonavicular joint**
43 Navicular bone
44 Medial cuneiform bone
45 Intermediate cuneiform bone
46 First metatarsal bone
47 Metatarsophalangeal joint of great toe
48 Proximal phalanx of great toe
49 Distal phalanx of great toe
50 Heads of metatarsal bones II-IV

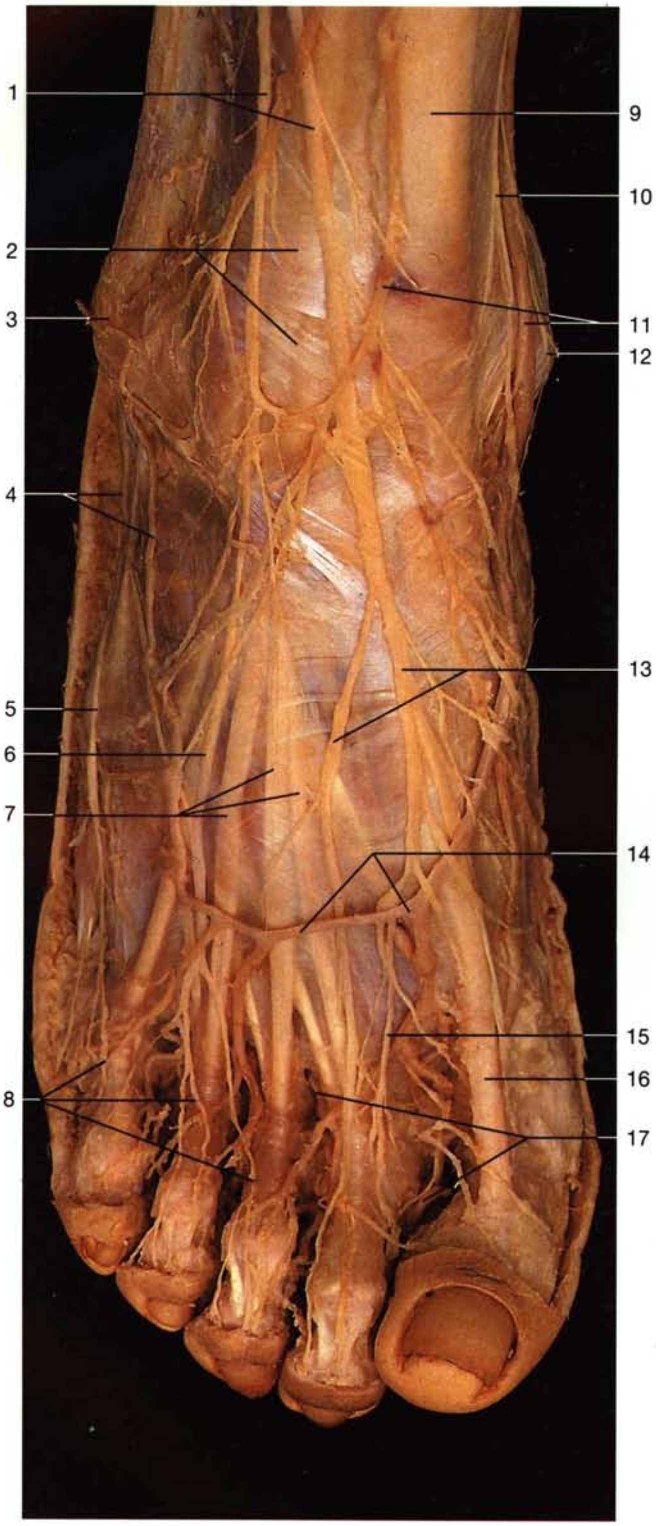

Dorsum of the right foot, superficial layer (anterior aspect).

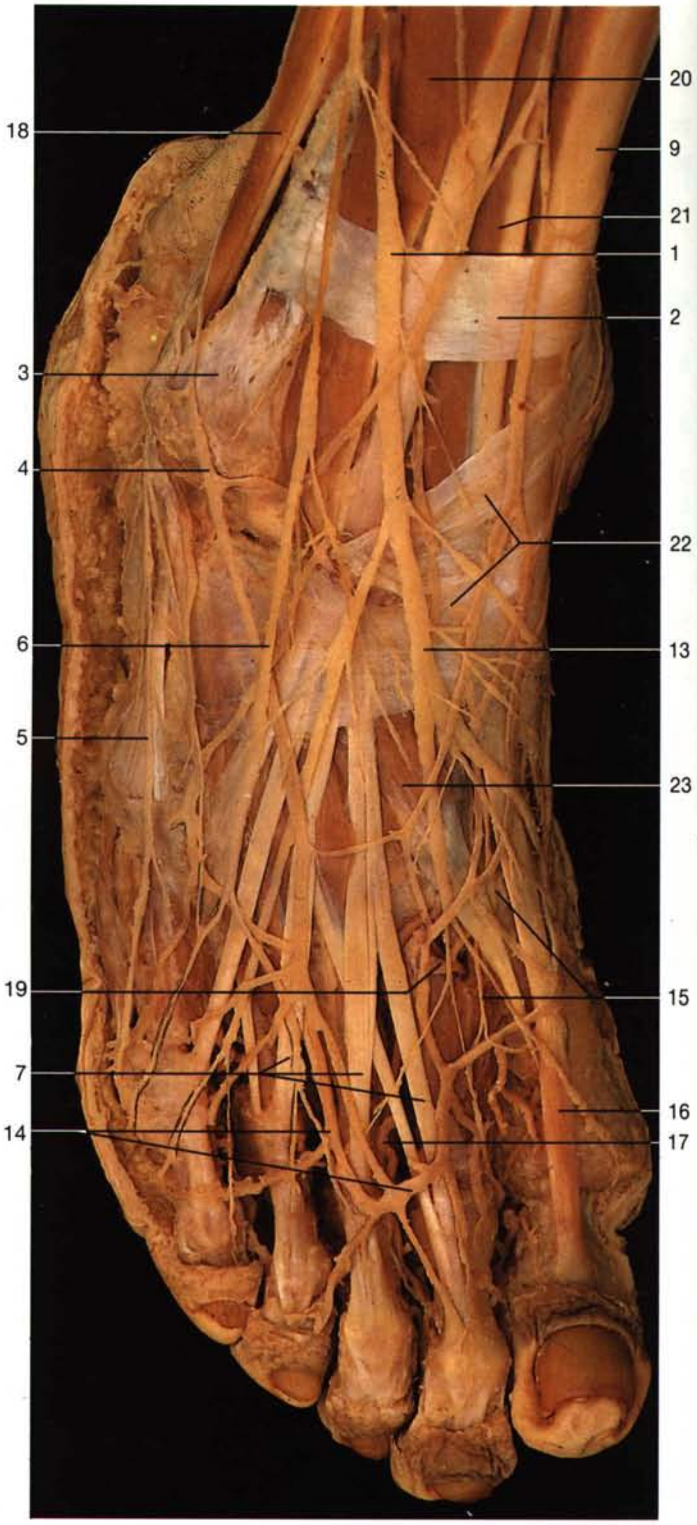

Dorsum of the right foot, superficial layer. The fascia of the dorsum has been removed.

1	**Superficial peroneal nerve**
2	Superior extensor retinaculum
3	Lateral malleolus
4	Venous network of lateral malleolus and tributaries of small saphenous vein
5	Lateral dorsal cutaneous nerve (branch of sural nerve)
6	Intermediate dorsal cutaneous nerve
7	Tendons of extensor digitorum longus muscle
8	Dorsal digital nerves
9	Tendon of tibialis anterior muscle
10	**Saphenous nerve**
11	Venous network of medial malleolus and tributaries of great saphenous vein
12	Medial malleolus
13	Medial dorsal cutaneous nerves
14	**Dorsal venous arch**
15	Dorsal digital nerve (of deep peroneal nerve)
16	Tendon of extensor hallucis longus muscle
17	Dorsal digital arteries
18	Peroneal muscles
19	Deep plantar branch of dorsalis pedis artery anastomosing with plantar arch
20	Extensor digitorum longus muscle
21	Extensor hallucis longus muscle
22	Inferior extensor retinaculum
23	Extensor hallucis brevis muscle

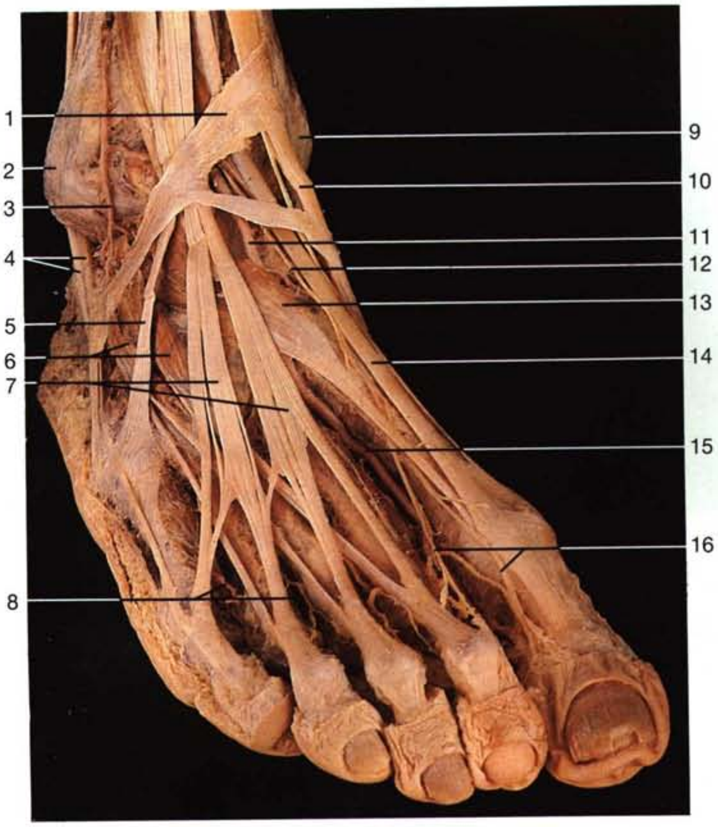

Dorsum of right foot, middle layer (anterior lateral aspect). The cutaneous nerves have been removed.

1 Extensor retinaculum
2 Lateral malleolus
3 Lateral anterior malleolar artery
4 Tendons of peroneal muscles
5 Tendon of peroneus tertius muscle
6 Extensor digitorum brevis muscle
7 Tendons of extensor digitorum longus muscle
8 Dorsal metatarsal arteries
9 Medial malleolus
10 Tendon of tibialis anterior muscle
11 **Dorsalis pedis artery**
12 Deep peroneal nerve (on dorsum of foot)
13 Extensor hallucis brevis muscle
14 Tendon of extensor hallucis longus muscle
15 Dorsalis pedis artery with deep plantar branch to the plantar arch
16 Dorsal digital nerves (terminal branches of deep peroneal nerve)
17 Lateral tarsal artery
18 Extensor digitorum brevis muscle (divided)
19 Arcuate artery
20 Dorsal interosseous muscles
21 **Deep peroneal nerve**
22 Medial cuneiform and 1st metatarsal bone
23 Tendon of peroneus longus muscle
24 Abductor hallucis and flexor hallucis brevis muscles
25 Medial plantar artery, vein and nerve
26 Fourth and fifth metatarsal bone
27 Adductor hallucis (oblique head)
28 Tendons of flexor digitorum longus muscle
29 Lateral plantar artery, vein and nerve
30 Flexor digitorum brevis muscle
31 Plantar aponeurosis

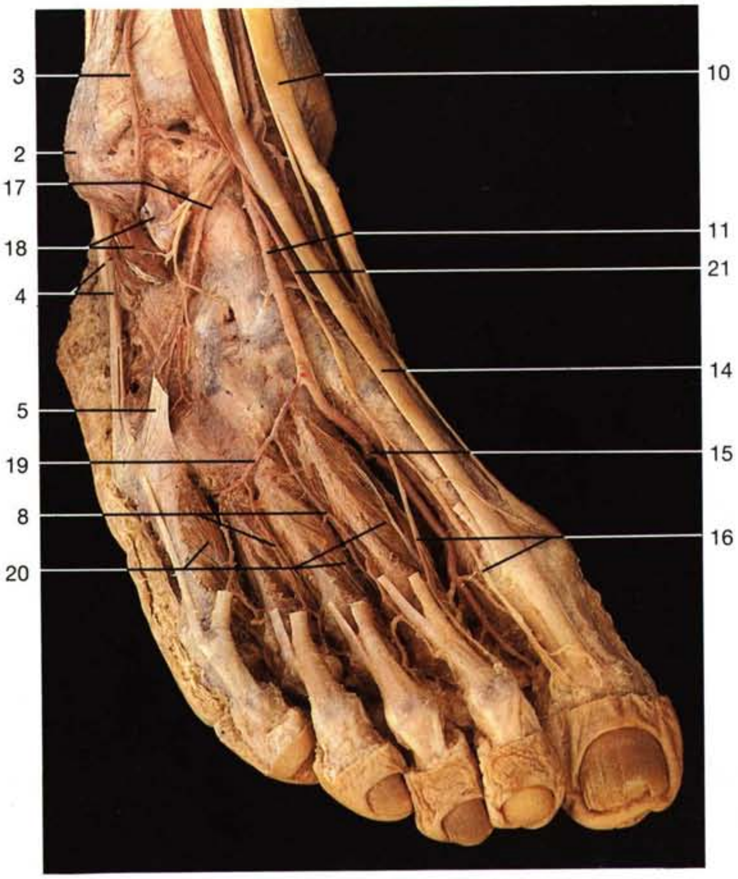

Dorsum of right foot, deep layer (anterior lateral aspect). The extensor digitorum and hallucis breves muscles have been removed.

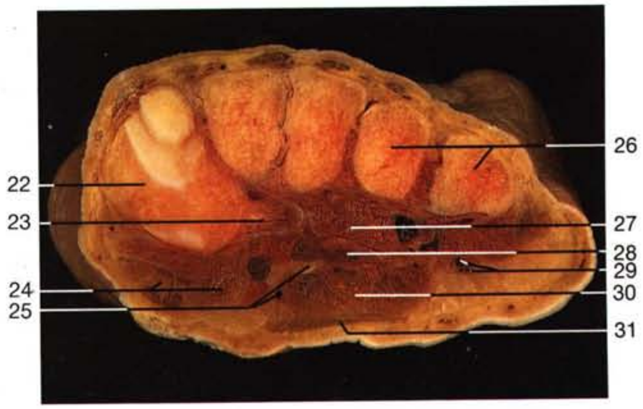

Cross-section of the rigt foot at the level of the metatarsal bones (posterior aspect).

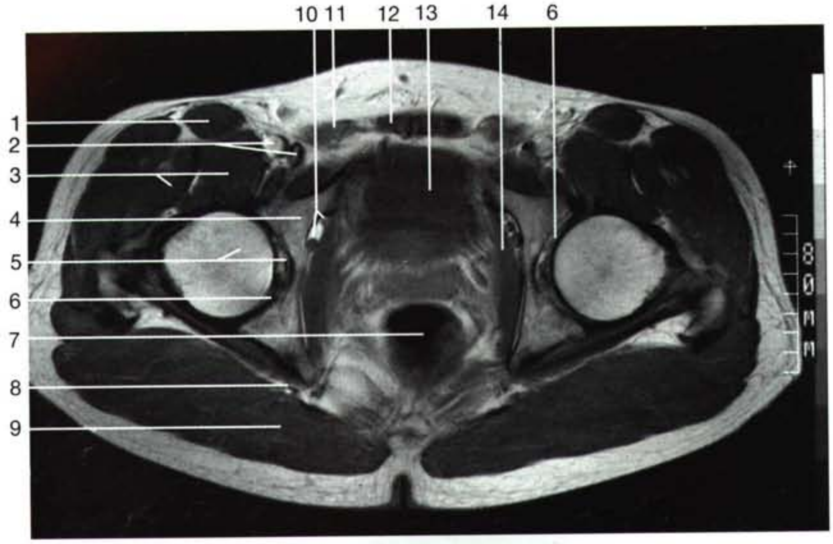
10
11
12
13
14
6
1
2
3
4
5
6
7
8
9

Axial section through the pelvis and the hip joints (section 1; MR-Image; inferior aspect).

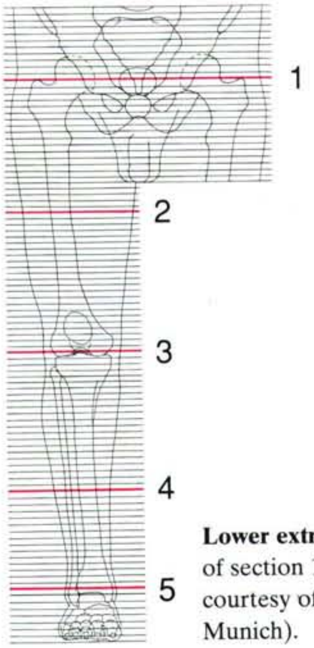

Lower extremity, location of section 1–5. (MR-Images, courtesy of Dr. A. Heuck, Munich).

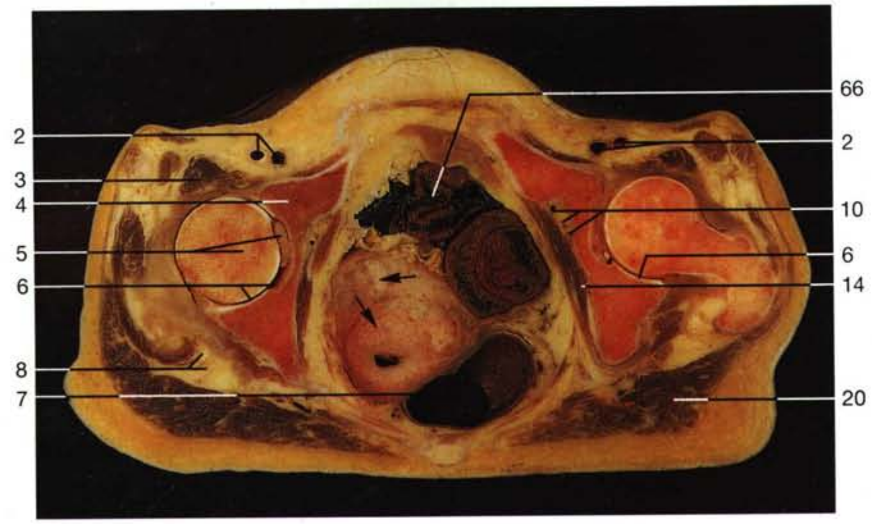
66
2
2
3
4
10
5
6
6
14
8
7
20

Axial section through the pelvis and hip joints in the female (section 1, inferior aspect). (Arrows: uterus, myometrium with myoma).

1 Sartorius muscle
2 **Femoral artery and vein**
3 Iliopsoas muscle
4 Pelvic bone
5 **Femoral head** with ligament of femoral head
6 Articular cavity
7 Rectum
8 **Sciatic nerve** and accompanying artery
9 Gluteus maximus muscle
10 Obturator vessels and obturator nerve
11 Rectus abdominis muscle
12 Pyramidalis muscle
13 Urinary bladder
14 Obturator internus muscle
15 Rectus femoris muscle
16 Vastus intermedius and vastus lateralis of quadriceps femoris muscle

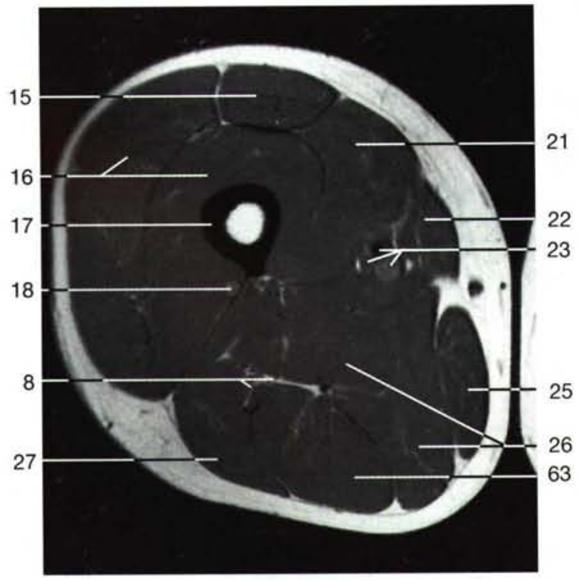
15
21
16
22
17
23
18
8
25
26
27
63

Axial section through the middle of the thigh (section 2; MR-Image, inferior aspect).

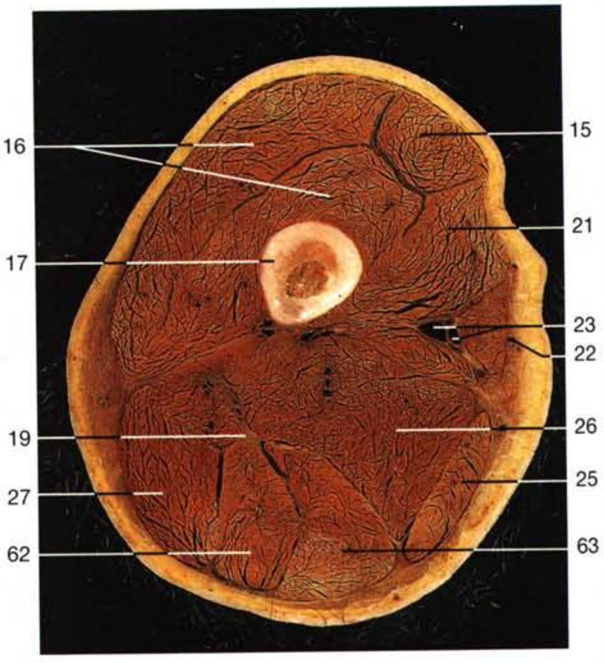
16
15
21
17
23
22
26
19
25
27
62
63

Axial section through the middle of the right thigh (section 2, inferior aspect).

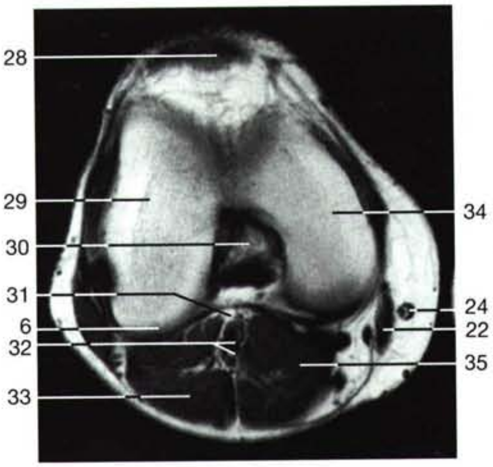

Axial section through the knee joint (section 3; MR-Image; inferior aspect).

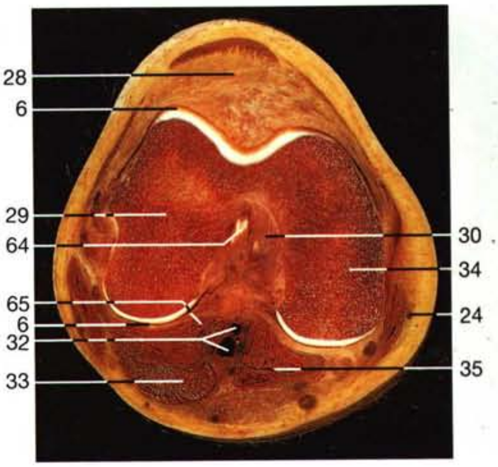

Axial section through the right knee joint (section 3, inferior aspect).

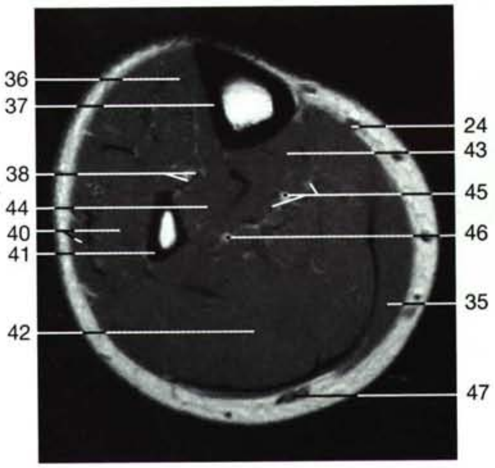

Axial section through the middle of the leg (section 4, MR-Image, inferior aspect).

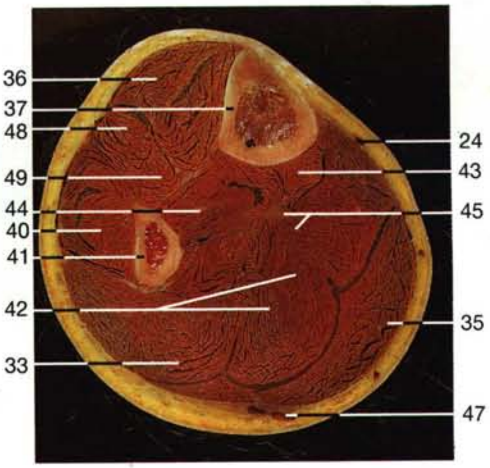

Axial section through the middle of the right leg (section 4, inferior aspect).

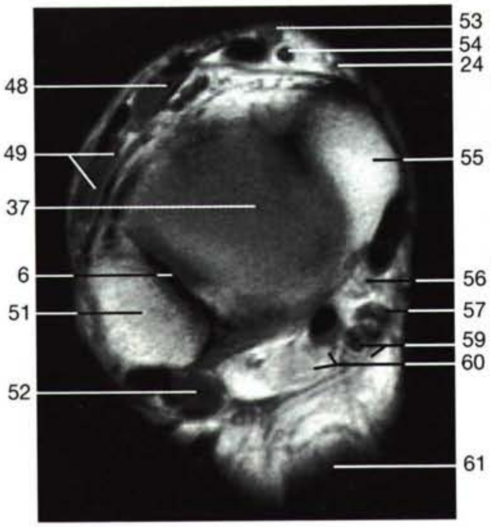

Axial section through the end of the right leg (section 5; MR-Image; inferior aspect).

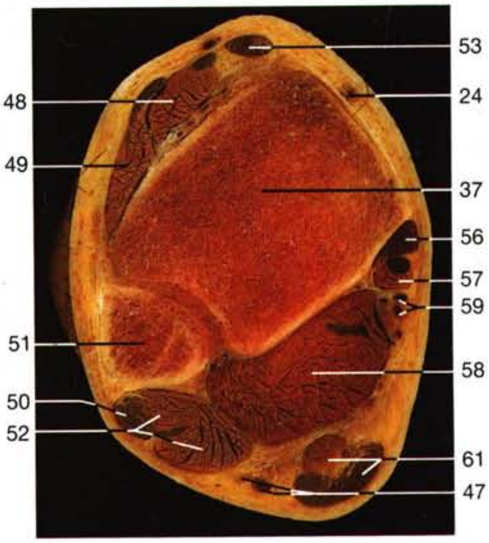

Axial section through the end of the right leg (section 5; inferior aspect).

17	**Femur**
18	Perforating artery
19	Sciatic nerve
20	Gluteus maximus muscle (insertion)
21	Vastus medialis muscle
22	Sartorius muscle
23	Femoral artery and vein
24	Great saphenous vein
25	Gracilis muscle
26	Adductor muscles
27	Biceps femoris muscle
28	Patellar ligament
29	Lateral condyle of femur
30	Posterior cruciate ligament
31	**Tibialis nerve**
32	Popliteal artery and vein
33	Lateral head of gastrocnemius muscle
34	Medial condyle of femur
35	Medial head of gastrocnemius muscle
36	Tibialis anterior muscle
37	**Tibia**
38	Deep peroneal nerve, anterior tibial artery and vein
40	Peroneus longus and brevis muscles
41	**Fibula**
42	Soleus muscle
43	Flexor digitorum longus muscle
44	Tibialis posterior muscle
45	Posterior tibial artery and vein and tibial nerve
46	Peroneal artery
47	Small saphenous vein and sural nerve
48	Extensor hallucis longus muscle
49	Extensor digitorum longus muscle
50	Tendon of peroneus longus muscle
51	Lateral malleolus (fibula)
52	Peroneus brevis muscle
53	Tibialis anterior muscle (tendon)
54	Dorsalis pedis artery
55	Medial malleolus (tibia)
56	Tibialis posterior muscle (tendon)
57	Flexor digitorum longus muscle (tendon with synovial sheath)
58	Flexor hallucis longus muscle
59	Posterior tibial artery and vein
60	Lateral and medial plantar nerves
61	Calcaneal tendon
62	Semitendinosus muscle
63	Semimembranosus muscle
64	Anterior cruciate ligament
65	Plantaris muscle
66	Small intestine

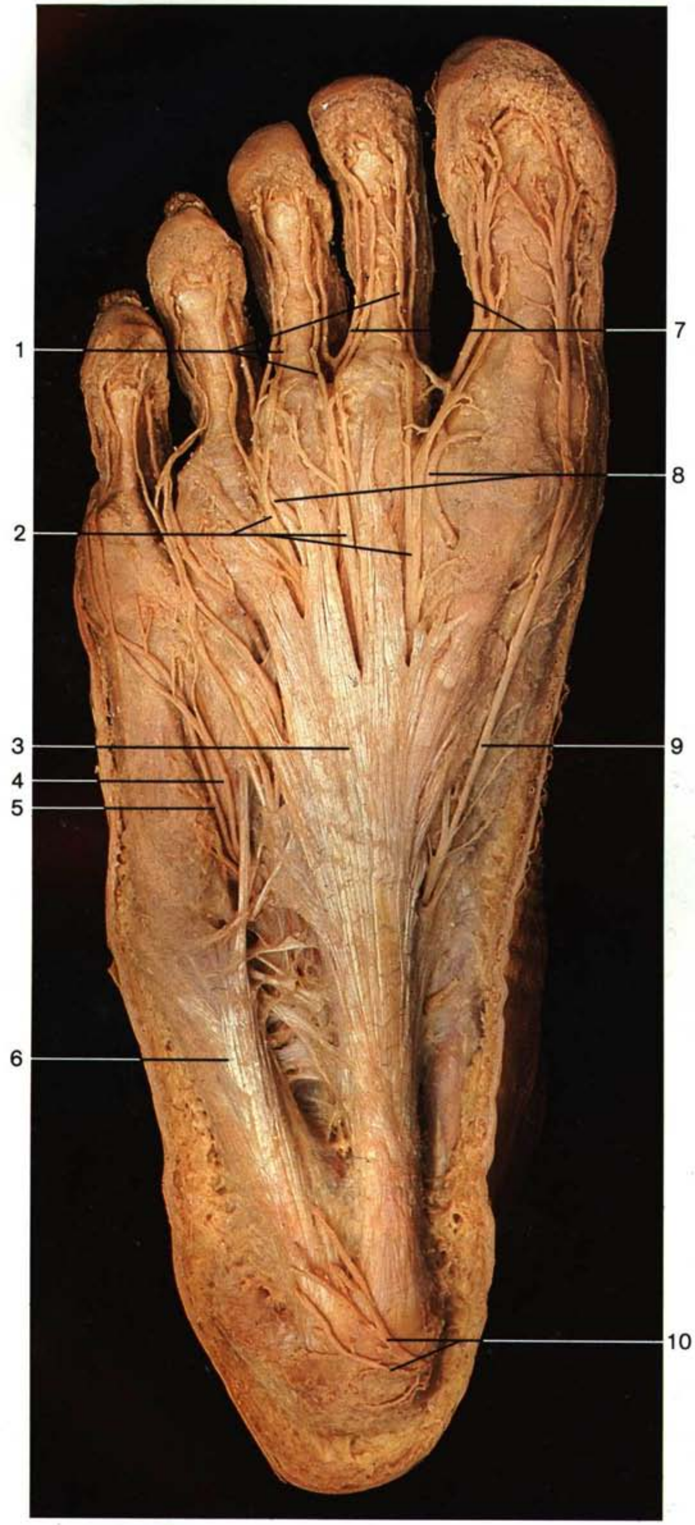
1
2
3
4
5
6
7
8
9
10

Sole of the right foot, superficial layer (from below): dissection of cutaneous nerves and vessels.

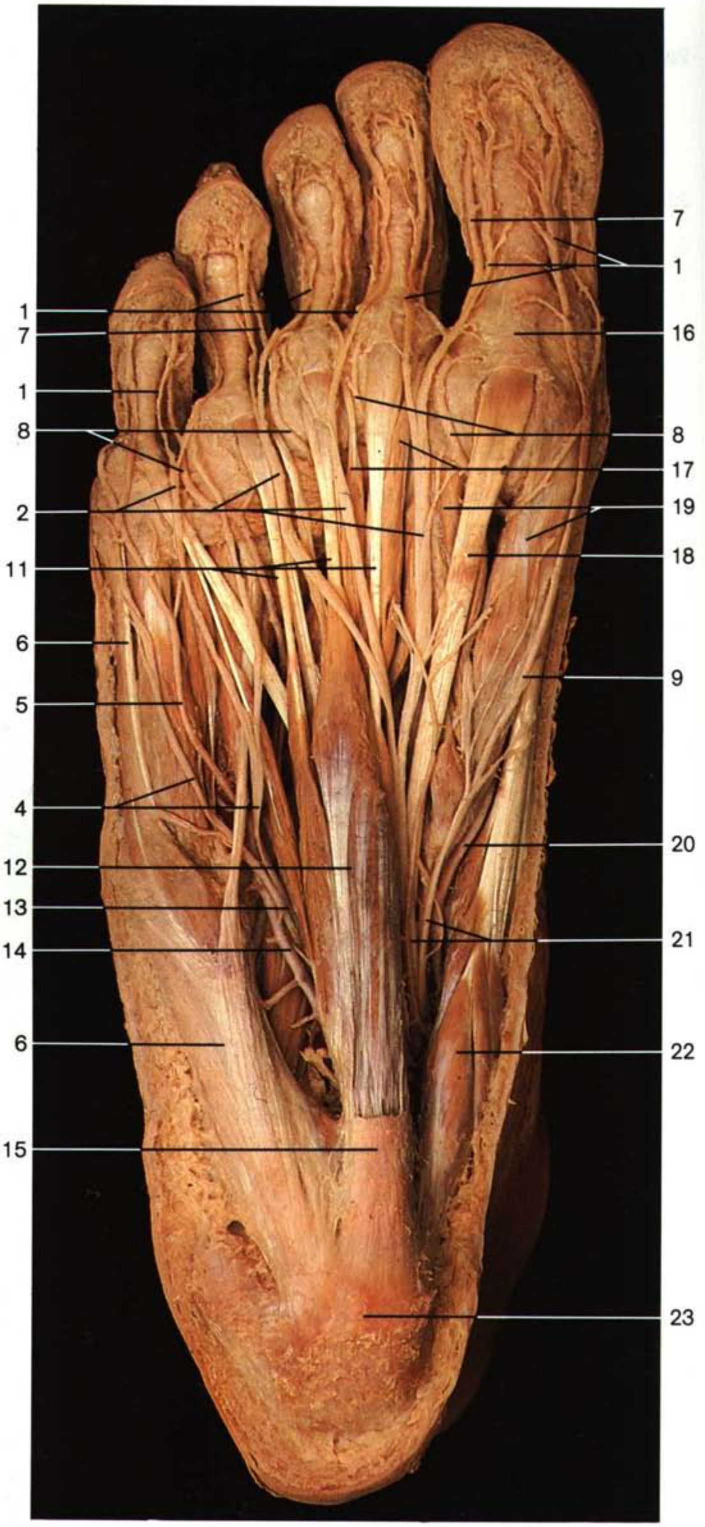
1
7
1
8
2
11
6
5
4
12
13
14
6
15
7
1
16
8
17
19
18
9
20
21
22
23

Sole of the right foot, middle layer (from below). The plantar aponeurosis has been removed.

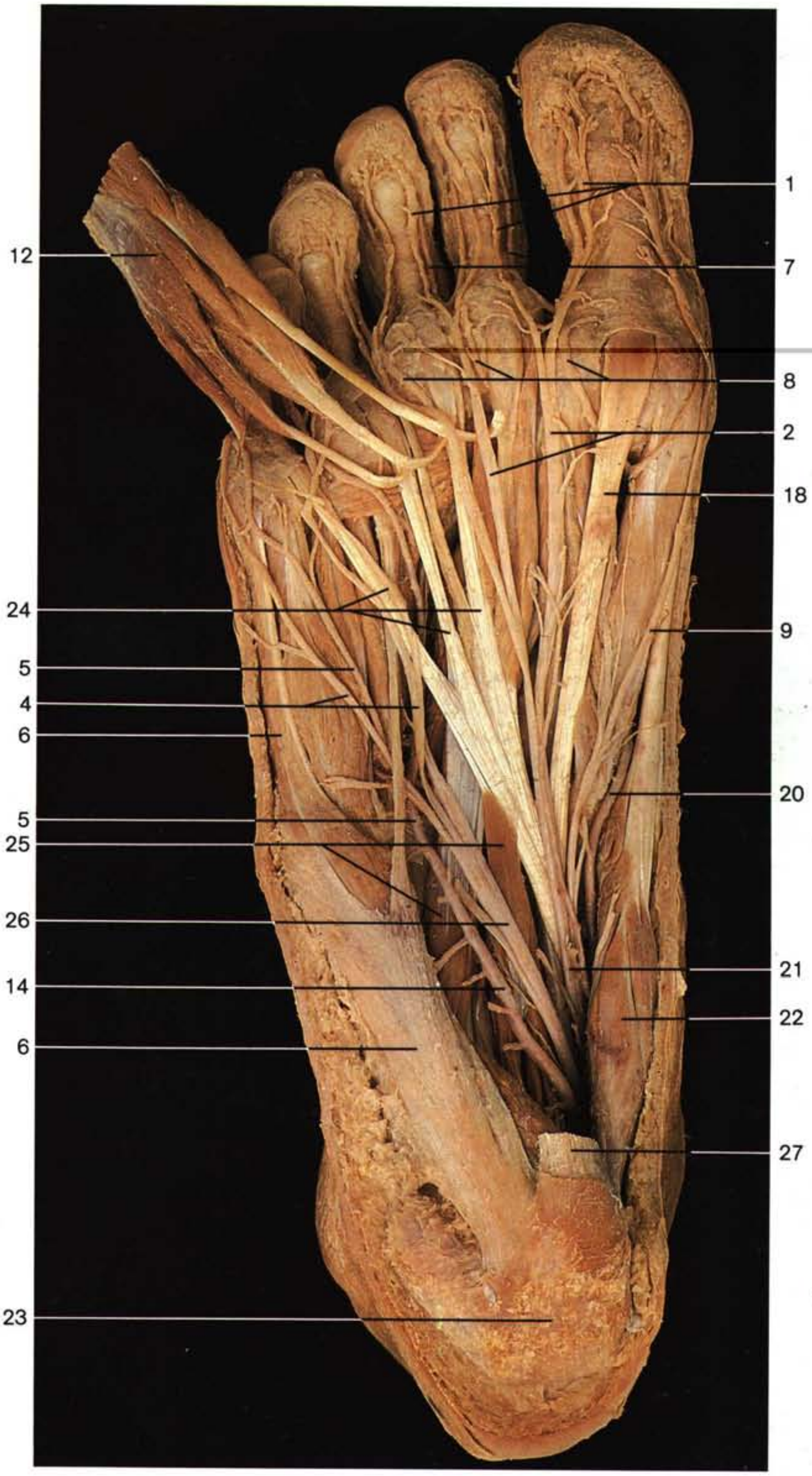

Sole of the right foot, middle layer (from below); dissection of vessels and nerves. The flexor digitorum brevis muscle has been divided and anteriorly reflected.

1	Proper plantar digital nerves
2	Common plantar digital nerves
3	Plantar aponeurosis
4	Superficial branch of lateral plantar nerve
5	Superficial branch of lateral plantar artery
6	Abductor digiti minimi
7	Proper plantar digital arteries
8	Common plantar digital arteries
9	Digital branch of medial plantar nerve to great toe
10	Medial calcaneal branches
11	Tendons of flexor digitorum brevis muscle
12	Flexor digitorum brevis muscle
13	**Superficial branch of lateral plantar nerve**
14	**Lateral plantar artery**
15	Plantar aponeurosis (remnant)
16	Digital synovial sheath
17	Lumbrical muscles
18	Tendon of flexor hallucis longus muscle
19	Flexor hallucis brevis muscle
20	**Medial plantar artery**
21	**Medial plantar nerve**
22	Abductor hallucis muscle
23	Calcaneal tuberosity
24	Tendons of flexor digitorum longus muscle
25	Quadratus plantae muscle
26	**Lateral plantar nerve**
27	Flexor digitorum brevis muscle (cut)
28	Synorial sheaths
29	Plantar arch

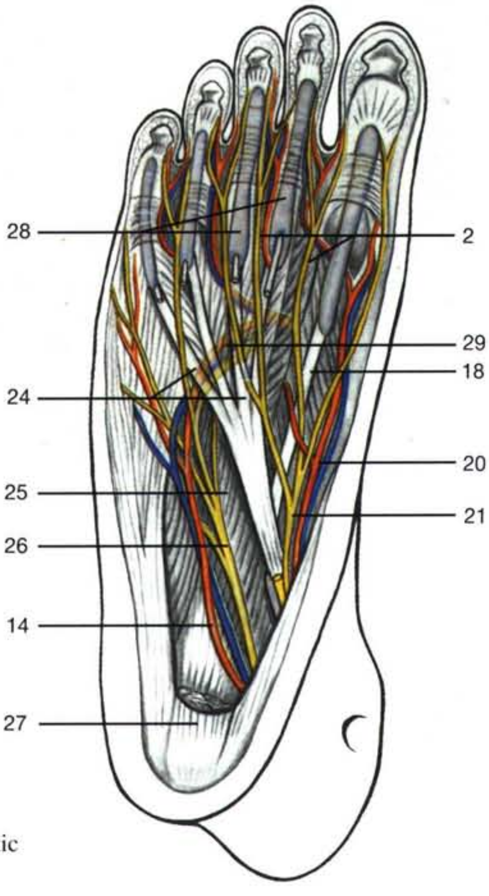

Sole of the right foot. Synovial sheaths of flexor tendons indicated in light blue. (Schematic drawing.)

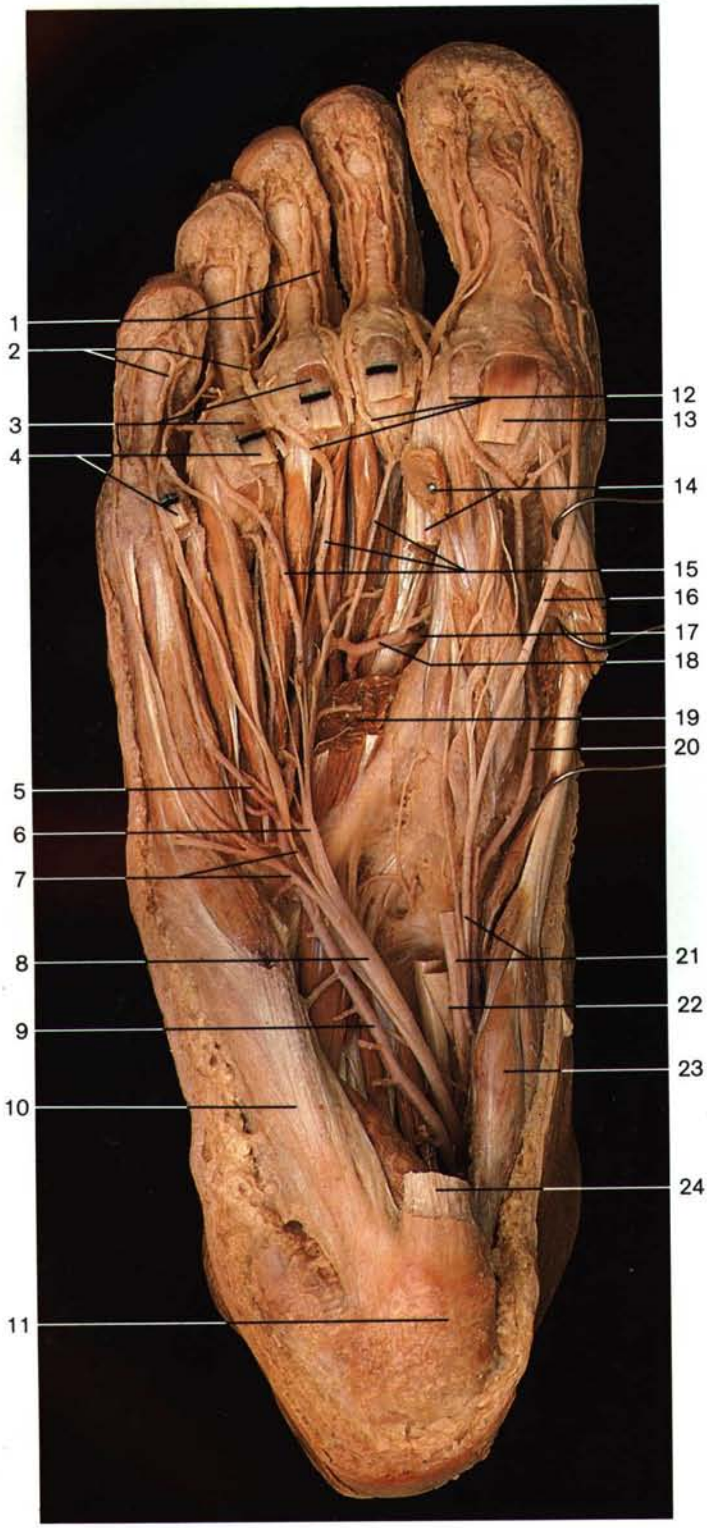

1 Proper plantar digital arteries
2 Proper plantar digital nerves
3 Tendons of flexor digitorum brevis muscle
4 Tendons of flexor digitorum longus muscle
5 Superficial branch of lateral plantar artery
6 Deep branch of lateral plantar nerve
7 Superficial branch of lateral plantar nerve
8 **Lateral plantar nerve**
9 **Lateral plantar artery**
10 Abductor digiti minimi muscle
11 Calcaneal tuberosity
12 Common plantar digital arteries
13 Tendon of flexor hallucis longus muscle
14 Insertion of both heads of adductor hallucis muscle
15 Plantar metatarsal arteries
16 Medial plantar nerve of great toe
17 Deep plantar branch of dorsalis pedis artery (perforating branch)
18 **Plantar arch**
19 Oblique head of adductor hallucis muscle (cut)
20 **Medial plantar artery**
21 **Medial plantar nerve**
22 Crossing of tendons in sole of foot (flexor hallucis longus and flexor digitorum longus muscles)
23 Abductor hallucis muscle
24 Origin of flexor digitorum brevis muscle

Sole of the right foot, deep layer (from below); dissection of vessels and nerves. The flexor digitorum brevis muscle, the quadratus plantae muscle with the tendons of the flexor digitorum longus muscle and some branches of the medial plantar nerve have been removed. The flexor hallucis brevis and adductor hallucis muscles have been cut and portions removed to show the somewhat atypical course of the medial plantar artery and deep muscles of the foot.

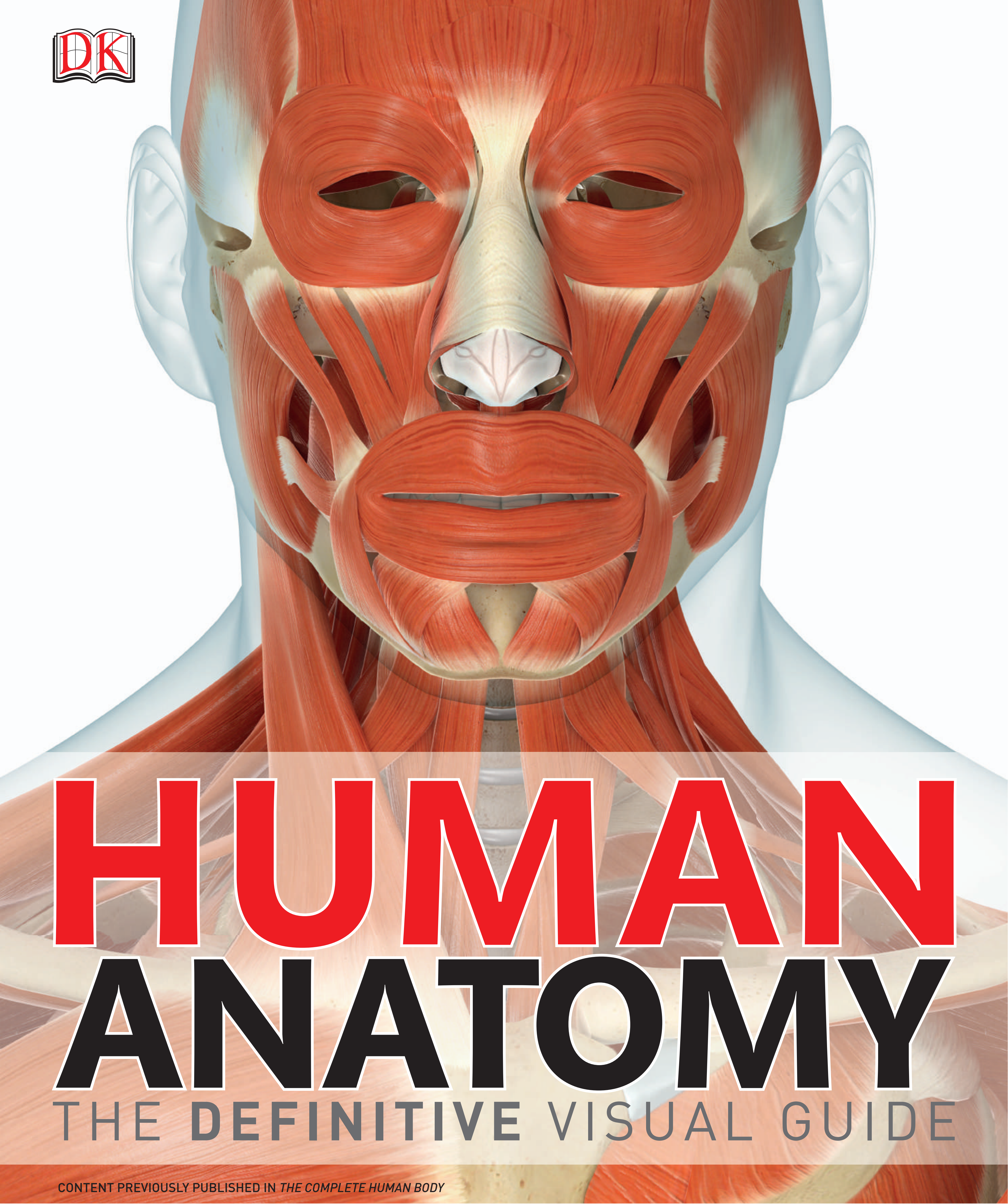

HUMAN ANATOMY

THE **DEFINITIVE** VISUAL GUIDE

CONTENT PREVIOUSLY PUBLISHED IN *THE COMPLETE HUMAN BODY*

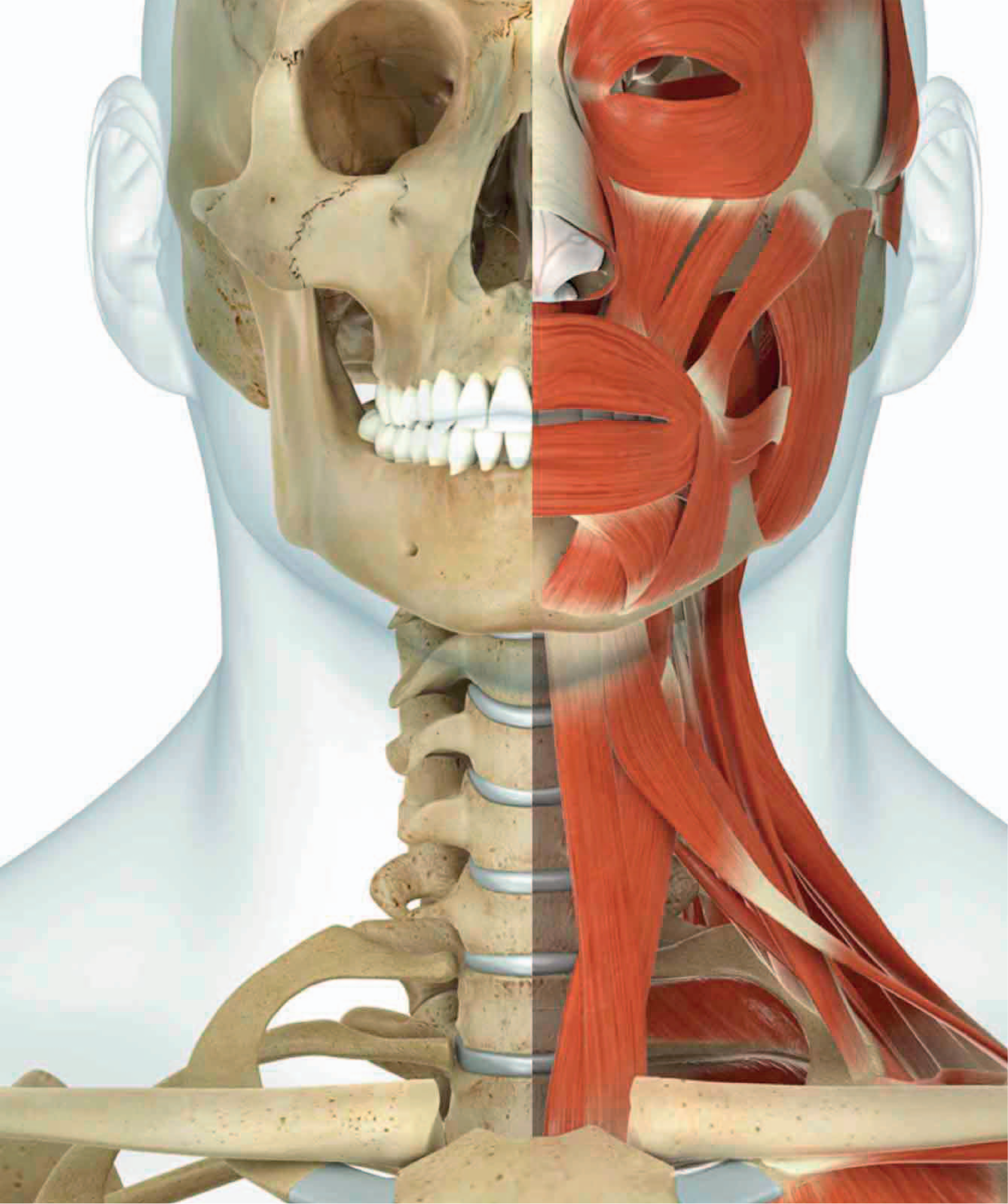

LONDON, NEW YORK, MELBOURNE,
MUNICH, AND DELHI

DK LONDON

Project Art Editor
Duncan Turner
Jacket Designer
Duncan Turner
Pre-Production Producer
Vikki Nousiainen
Managing Art Editor
Michelle Baxter
Production Editor
Rachel Ng
Art Director
Philip Ormerod
Associate Publishing Director
Liz Wheeler

Project Editor
Ruth O'Rourke-Jones
US Editor
Cheryl Ehrlich
Editor
Lili Bryant
Jacket Editor
Manisha Majithia
Managing Editor
Angeles Gavira
Publisher
Sarah Larter
Publishing Director
Jonathan Metcalf

DK DELHI

Senior Editor
Anita Kakkar
Art Editors
Suhita Dharamjit,
Amit Malhotra
Deputy Managing Art Editor
Sudakshina Basu
DTP Designer
Vishal Bhatia
Production Manager
Pankaj Sharma

Editor
Pallavi Singh
Design Assistant
Anjali Sachar
Managing Editor
Rohan Sinha
Pre-Production Manager
Balwant Singh
Picture Researcher
Sumedha Chopra

Illustrators
Medi-Mation (Creative Director: Rajeev Doshi)
Antbits Ltd. (Richard Tibbitts)
Dotnamestudios (Andrew Kerr)
Deborah Maizels

Editor-in-Chief Professor Alice Roberts

Authors

THE INTEGRATED BODY
Linda Geddes

BODY SYSTEMS,
IMAGING THE BODY
Professor Alice Roberts

Consultants

THE INTEGRATED BODY
Professor Mark Hanson,
University of Southampton, UK

BODY SYSTEMS,
IMAGING THE BODY
Professor Harold Ellis, King's College, London
Professor Susan Standring, King's College, London

Content previously published in *The Complete Human Body*
First American Edition, 2010
This American Edition, 2014

Published in the United States by
DK Publishing
4th floor, 345 Hudson Street
New York, New York 10014

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

256502—05/14

Copyright © 2010, 2014 Dorling Kindersley Limited
Foreword copyright © Alice Roberts
All rights reserved

Without limiting the rights under copyright reserved above, no part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted, in any form, or by any means (electronic, mechanical, photocopying, recording, or otherwise), without the prior written permission of both the copyright owner and the above publisher of this book.
Published in Great Britain by Dorling Kindersley Limited.

A catalog record for this book is available from the Library of Congress.

ISBN 978-1-4654-1954-5

DK books are available at special discounts when purchased in bulk for sales promotions, premiums, fund-raising, or educational use. For details, contact: DK Publishing Special Markets, 345 Hudson Street, New York, New York 10014 or SpecialSales@dk.com.

Printed and bound in China by South China

Discover more at
www.dk.com

CONTENTS

006 Foreword

008
THE INTEGRATED BODY

010 Human genetic formula
012 Cell
014 Body composition
016 Body systems
018 Terminology and planes

020
BODY SYSTEMS

022 SKIN, HAIR, AND NAIL

024 SKELETAL SYSTEM OVERVIEW
026 Head and neck
036 Thorax
040 Spine
042 Abdomen and pelvis
046 Pelvis

048 Shoulder and upper arm
054 Lower arm and hand
056 Hand and wrist joints
058 Hip and thigh
062 Hip and knee
064 Lower leg and foot
066 Foot and ankle

068 MUSCULAR SYSTEM OVERVIEW
070 Head and neck
076 Thorax
082 Abdomen and pelvis
086 Shoulder and upper arm
094 Lower arm and hand
098 Hip and thigh
106 Lower leg and foot

110 NERVOUS SYSTEM OVERVIEW
112 Brain
118 Head and neck
120 Brain (transverse and coronal sections)
122 Head and neck (cranial nerves)
124 Eye
126 Ear
128 Neck
130 Thorax
132 Abdomen and pelvis
134 Shoulder and upper arm
138 Lower arm and hand
140 Hip and thigh
144 Lower leg and foot

146 RESPIRATORY SYSTEM OVERVIEW
148 Head and neck
150 Thorax
152 Lungs

154 CARDIOVASCULAR SYSTEM OVERVIEW
156 Head and neck
160 Thorax
162 Heart
166 Abdomen and pelvis
168 Shoulder and upper arm
172 Lower arm and hand
174 Hip and thigh
178 Lower leg and foot

180 LYMPHATIC AND IMMUNE SYSTEM OVERVIEW
182 Head and neck
184 Thorax
186 Abdomen and pelvis
188 Shoulder and upper arm
190 Hip and thigh

192 DIGESTIVE SYSTEM OVERVIEW
194 Head and neck
196 Thorax
198 Abdomen and pelvis
200 Stomach and intestines
202 Liver, pancreas, and gallbladder

204 URINARY SYSTEM OVERVIEW
206 Abdomen and pelvis

208 REPRODUCTIVE SYSTEM OVERVIEW
210 Thorax
212 Abdomen and pelvis

216 ENDOCRINE SYSTEM OVERVIEW
218 Head and neck

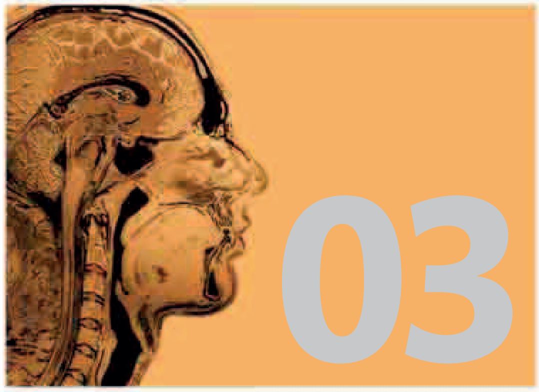

220
IMAGING THE BODY

222 Imaging techniques
224 Head and neck
226 Thorax
228 Abdomen and pelvis
230 Lower arm and hand
232 Lower limb and foot

234 GLOSSARY
241 INDEX
255 ACKNOWLEDGMENTS

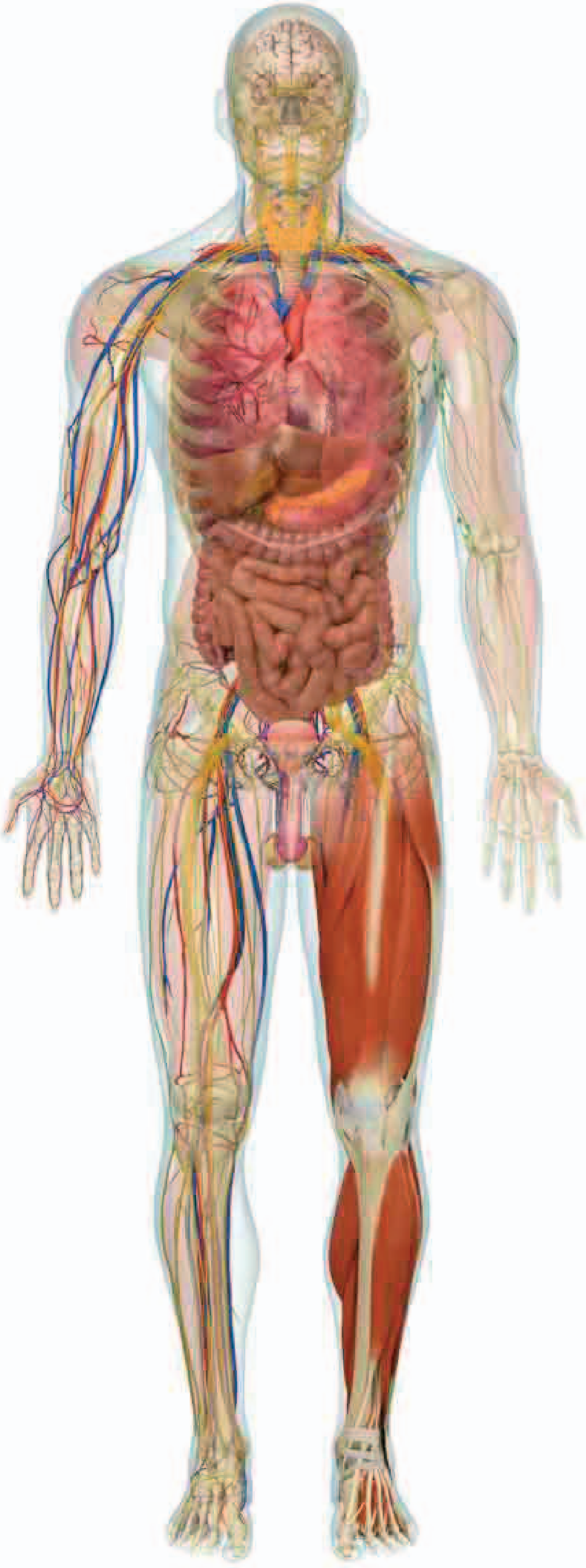

01 The Integrated Body

The human body comprises trillions of cells, each one a complex unit with intricate workings in itself. Cells are the building blocks of tissues, organs, and eventually, the integrated body systems that all interact—allowing us to function and survive.

010 Human genetic formula
012 Cell
014 Body composition
016 Body systems
018 Terminology and planes

HUMAN GENETIC FORMULA

DNA (deoxyribonucleic acid) is the blueprint for all life, from the humblest yeast to human beings. It provides a set of instructions for how to assemble the many thousands of different proteins that make us who we are. It also tightly regulates this assembly, ensuring that the components of the assemby do not run out of control.

THE MOLECULE OF LIFE

Although we all look different, the basic structure of our DNA is identical. It consists of chemical building blocks called bases, or nucleotides. What varies between individuals is the precise order in which these bases connect into pairs. When base pairs are strung together they can form functional units called genes, which "spell out" the instructions for making a protein. Each gene encodes a single protein, although some complex proteins are encoded by more than one gene. Proteins have a wide range of vital functions in the body. They form structures such as skin or hair, carry signals around the body, and fight off infectious agents such as bacteria. Proteins also make up cells, the basic units of the body, and carry out the thousands of basic biochemical processes needed to sustain life. However, only about 1.5 per cent of our DNA encodes genes. The rest consists of regulatory sequences, structural DNA, or has no obvious purpose – so-called "junk DNA".

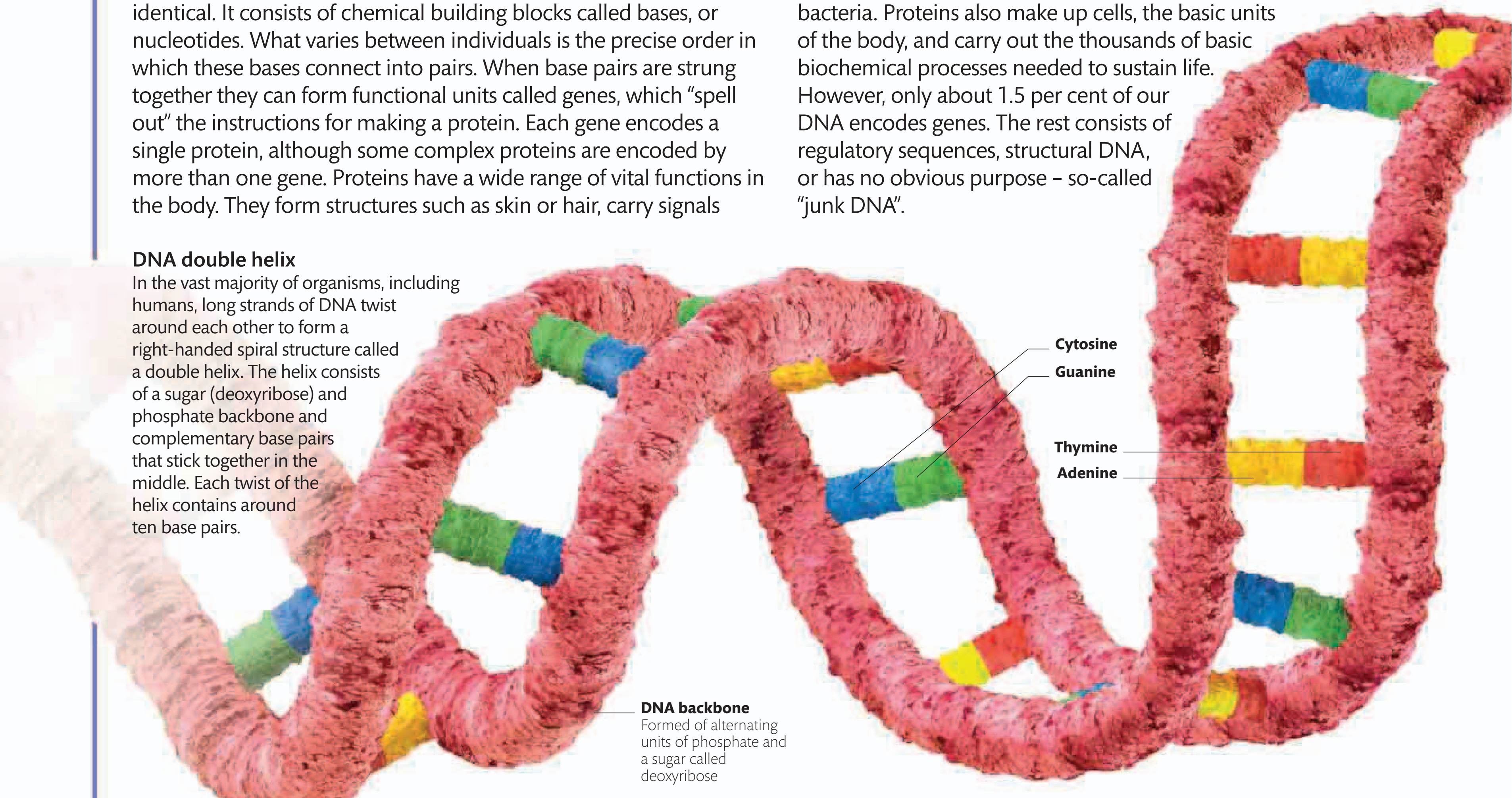

DNA double helix
In the vast majority of organisms, including humans, long strands of DNA twist around each other to form a right-handed spiral structure called a double helix. The helix consists of a sugar (deoxyribose) and phosphate backbone and complementary base pairs that stick together in the middle. Each twist of the helix contains around ten base pairs.

DNA backbone
Formed of alternating units of phosphate and a sugar called deoxyribose

PACKAGING DNA

The human genome is composed of approximately 3 billion bases of DNA—about 6½ft (2m) of DNA in every cell if it were stretched from end to end. So our DNA must be packaged in order to fit inside each cell. DNA is concentrated into dense structures called chromosomes. Each cell has 23 pairs of chromosomes (46 in total)—one set from each parent. To package DNA, the double helix must first be coiled around histone proteins, forming a structure that looks like a string of beads. These histone "beads" wind around and lock together into densely coiled "chromatin", which, when a cell prepares to divide, further winds back on itself into tightly coiled chromosomes.

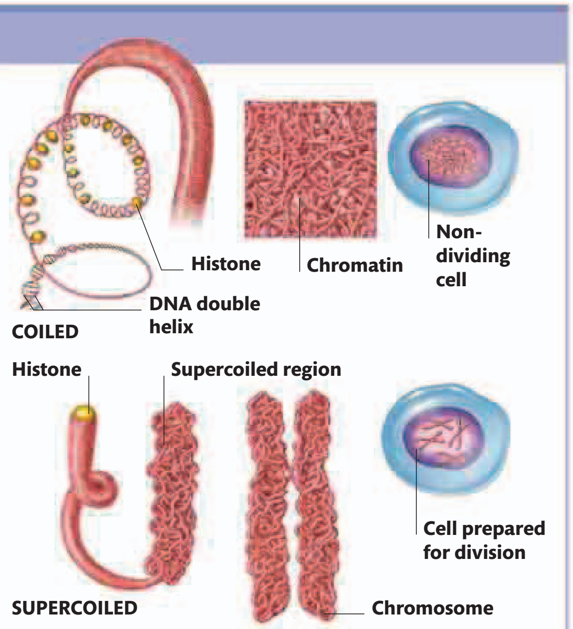

MAKING PROTEINS

Proteins consist of building blocks called amino acids, strung together in chains and folded. Every three base pairs of DNA codes for one amino acid. The body makes 20 different amino acids—others are obtained from the diet. Protein synthesis occurs in two steps: transcription and translation. In transcription, the DNA double helix unwinds, exposing single-stranded DNA. Complementary sequences of a related molecule called RNA (ribonucleic acid) then create a copy of the DNA sequence that locks into the exposed DNA bases to be translated into protein. This "messenger RNA" travels to ribosomes, where it is translated into strings of amino acids. These are then folded into the 3D structure of a particular protein.

BASE PAIRS

DNA consists of building blocks called bases. There are four types: adenine (A), thymine (T), cytosine (C), and guanine (G). Each base is attached to a phosphate group and a deoxyribose sugar ring to form a nucleotide. In humans, bases pair up to form a double-stranded helix in which adenine pairs with thymine, and cytosine with guanine. The two strands are "complementary" to each other. Even if they are unwound and unzipped, they can realign and rejoin.

GENES

A gene is a unit of DNA needed to make a protein. Genes range in size from just a few hundred to millions of base pairs. They control our development, but are also switched on and off in response to environmental factors. For example, when an immune cell encounters a bacterium, genes are switched on that produce antibodies to destroy it. Gene expression is regulated by proteins that bind to regulatory sequences within each gene. Genes contain regions that are translated into protein (exons) and non-coding regions (introns).

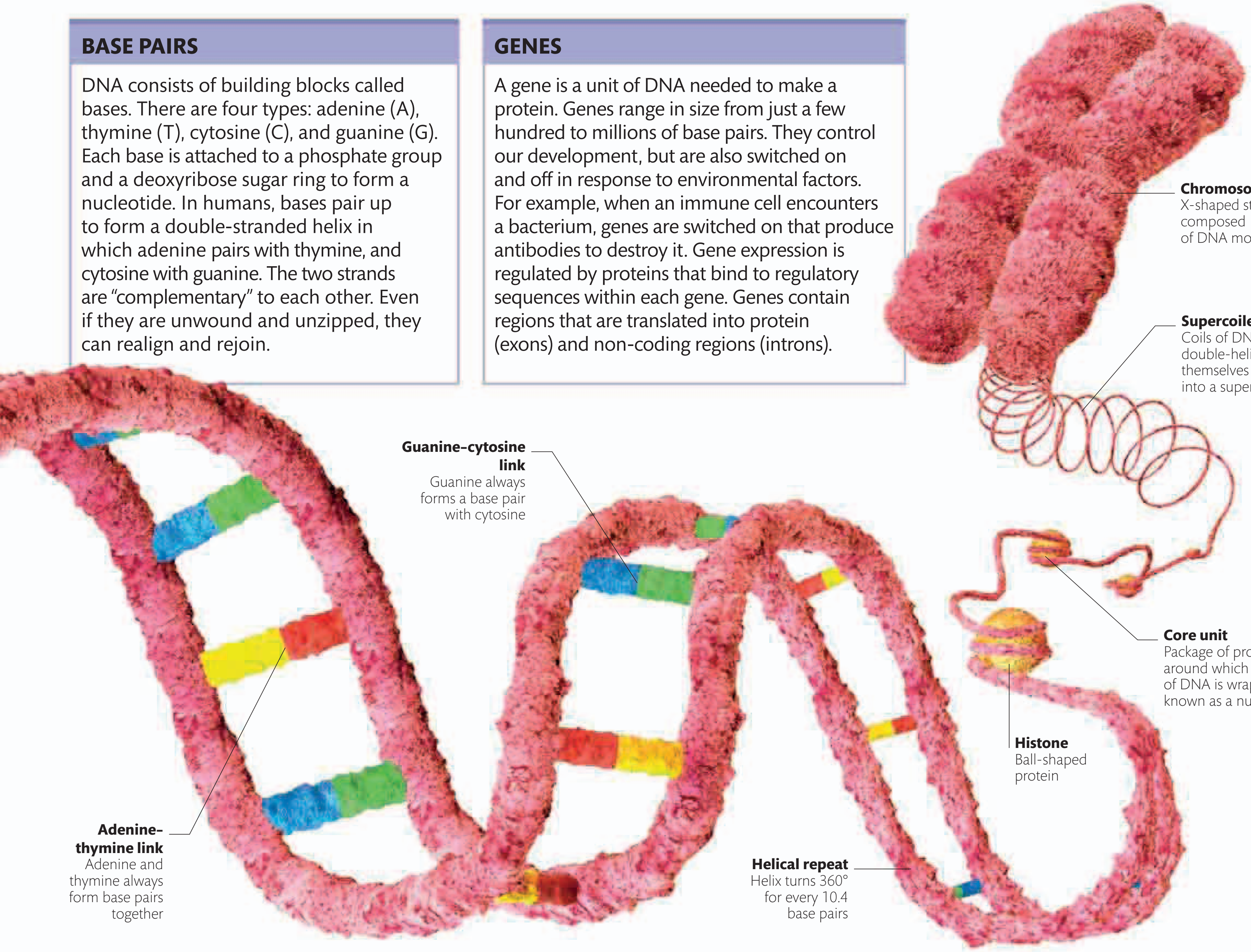

Chromosome X-shaped structure composed of DNA molecules

Supercoiled DNA Coils of DNA double-helix are themselves twisted into a supercoil

Guanine-cytosine link Guanine always forms a base pair with cytosine

Core unit Package of proteins around which 2–5 turns of DNA is wrapped; also known as a nucleosome

Histone Ball-shaped protein

Adenine-thymine link Adenine and thymine always form base pairs together

Helical repeat Helix turns 360° for every 10.4 base pairs

THE HUMAN GENOME

Different organisms contain different genes, but a surprisingly large proportion of genes are shared between organisms. For example, roughly half of the genes found in humans are also found in bananas. However, it would not be possible to substitute the banana version of a gene for a human one because variations in the order of the base pairs within each gene also distinguish us. Humans possess more or less the same genes, but many of the differences between individuals can be explained by subtle variations within each gene. In humans, DNA differs by only about 0.2 per cent, while human DNA differs from chimpanzee DNA by around 5 per cent. Human genes are divided unevenly between 23 pairs of chromosomes, and each chromosome consists of gene-rich and gene-poor sections. When chromosomes are stained, differences in these regions show up as light and dark bands, giving chromosomes a striped appearance. We still don't know the exact number of protein-coding genes in the human genome, but researchers currently estimate between 20,000 and 25,000.

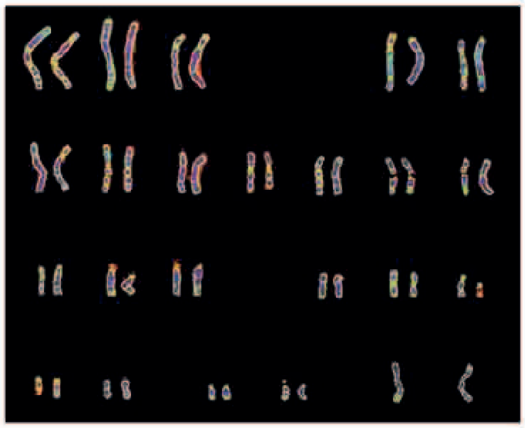

Karyotype
This is an organized profile of the chromosomes in someone's cells, arranged by size. Studying someone's karyotype enables doctors to determine whether any chromosomes are missing or abnormal.

GENETIC ENGINEERING

This form of gene manipulation enables us to substitute a defective gene with a functional one, or introduce new genes. Glow-in-the-dark mice were created by introducing a jellyfish gene that encodes a fluorescent protein into the mouse genome. Finding safe ways of delivering replacement genes to the correct cells in humans could lead to cures for many types of inherited diseases—so-called gene therapy.

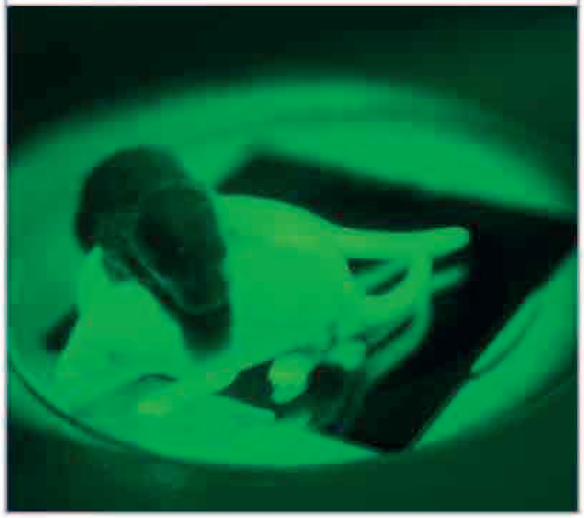

CELL

It is hard to comprehend what 75 trillion cells looks like, but observing yourself in a mirror would be a good start. That is how many cells exist in the average human body – and we replace millions of these cells every single day.

CELL ANATOMY

The cell is the basic functional unit of the human body. Cells are extremely small, typically only about 0.01mm across – even our largest cells are no bigger than the width of a human hair. They are also immensely versatile: some can form sheets like those in your skin or lining your mouth, while others can store or generate energy, such as fat and muscle cells. Despite their amazing diversity, there are certain features that all cells have in common, including an outer membrane, a control center called a nucleus, and tiny powerhouses called mitochondria.

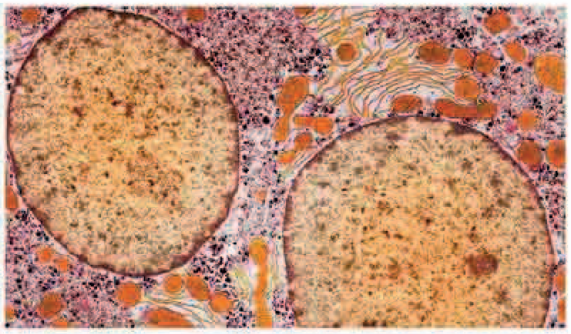

Liver cell
These cells make protein, cholesterol, and bile, and detoxify and modify substances from the blood. This requires lots of energy, so liver cells are packed with mitochondria (orange).

CELL METABOLISM

When a cell breaks down nutrients to generate energy for building new proteins or nucleic acids, it is known as cell metabolism. Cells use a variety of fuels to generate energy, but the most common one is glucose, which is transformed into adenosine triphosphate (ATP). This takes place in structures called mitochondria through a process called cellular respiration: enzymes within the mitochondria react with oxygen and glucose to produce ATP, carbon dioxide, and water. Energy is released when ATP is converted into adenoside diphosphate (ADP) via the loss of a phosphate group.

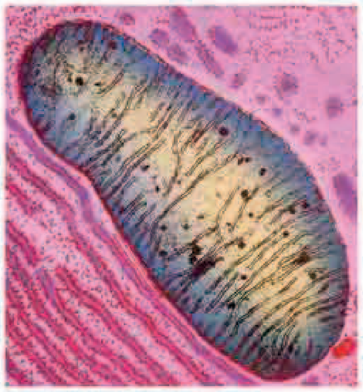

Mitochondrion
While the number of mitochondria varies between different cells, all have the same basic structure: an outer membrane and a highly folded inner membrane, where the production of energy actually takes place.

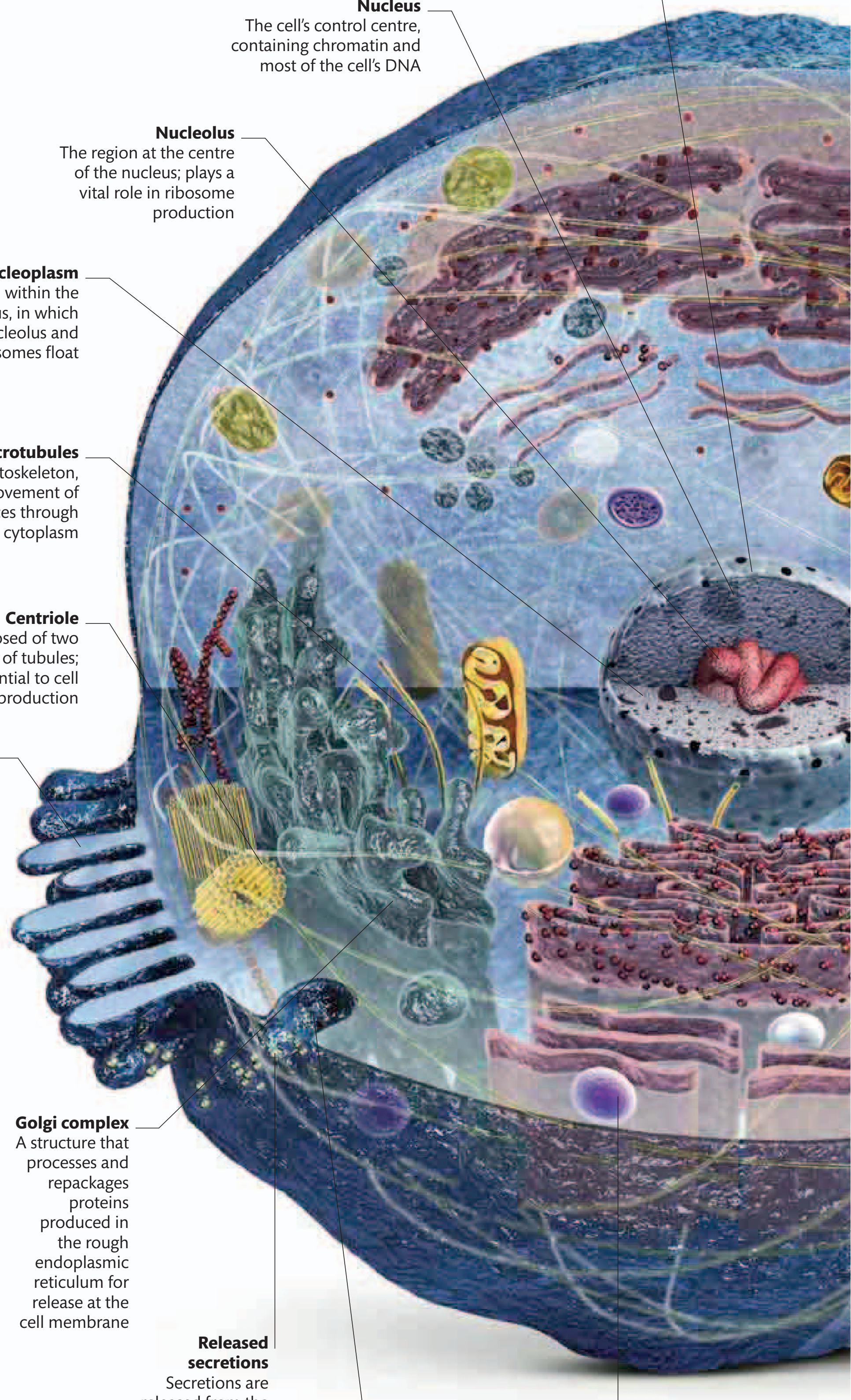

Nuclear membrane
A two-layered membrane with pores for substances to enter and leave the nucleus

Nucleus
The cell's control centre, containing chromatin and most of the cell's DNA

Nucleolus
The region at the centre of the nucleus; plays a vital role in ribosome production

Nucleoplasm
Fluid within the nucleus, in which nucleolus and chromosomes float

Microtubules
Part of cell's cytoskeleton, these aid movement of substances through the watery cytoplasm

Centriole
Composed of two cylinders of tubules; essential to cell reproduction

Microvilli
These projections increase the cell's surface area, aiding absorption of nutrients

Golgi complex
A structure that processes and repackages proteins produced in the rough endoplasmic reticulum for release at the cell membrane

Released secretions
Secretions are released from the cell by exocytosis, in which a vesicle merges with the cell membrane and releases its contents

Secretory vesicle
Sac containing various substances, such as enzymes, that are produced by the cell and secreted at the cell membrane

Lysosome
Produces powerful enzymes that aid in digestion and excretion of substances and worn-out organelles

Generic cell

At a cell's heart is the nucleus, where the genetic material is stored and the first stages of protein synthesis occur. Cells also contain other structures for assembling proteins, including ribosomes, the endoplasmic reticulum, and the Golgi apparatus. The mitochondria provide the cell with energy.

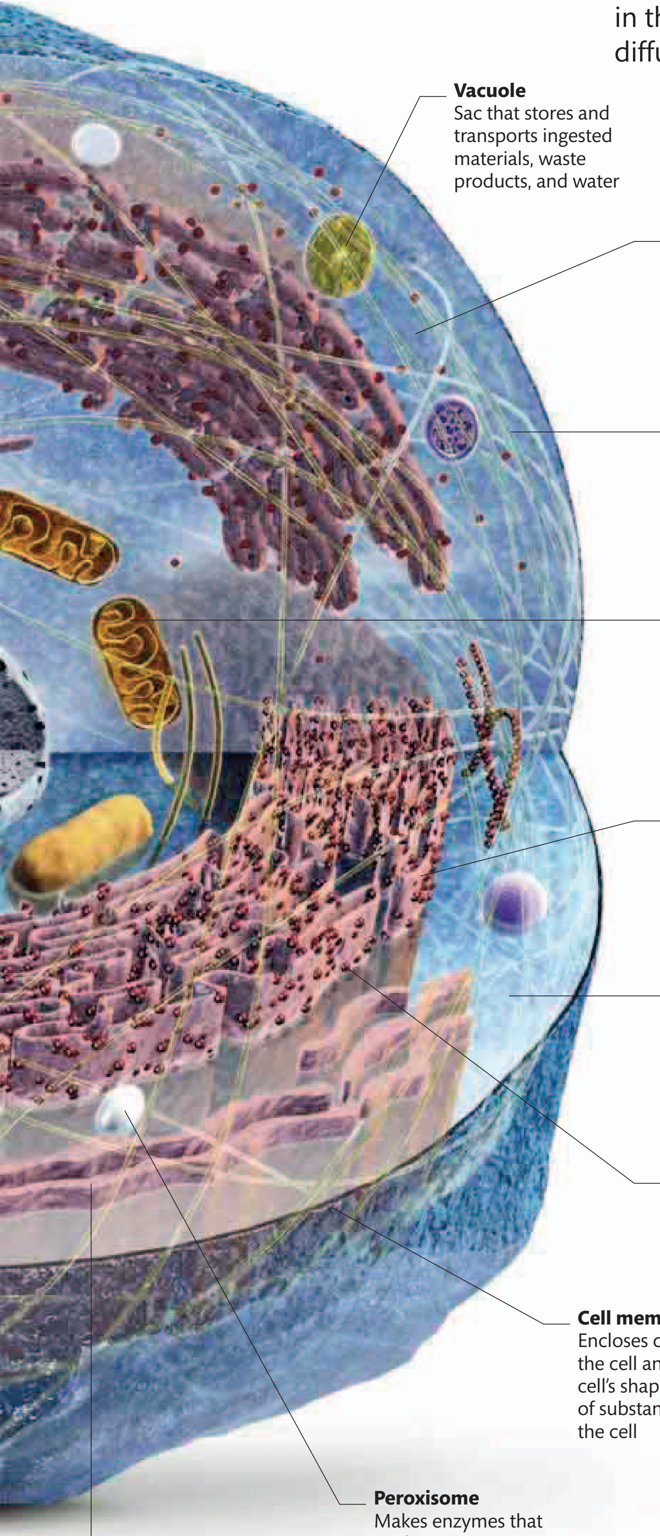

Vacuole
Sac that stores and transports ingested materials, waste products, and water

Cytoskeleton
Internal framework of the cell, made up of microfilaments and hollow microtubules

Microfilament
Provides support for the cell; sometimes linked to the cell's outer membrane

Mitochondrion
Site of fat and sugar digestion in the cell; produces energy

Rough endoplasmic reticulum
Consists of folded membranes, studded with ribosomes, that extend throughout the cell

Cytoplasm
Jellylike fluid in which organelles float; primarily water, but also contains enzymes and amino acids

Ribosome
Tiny structure that assists with protein assembly (see p.10)

Cell membrane
Encloses contents of the cell and maintains the cell's shape; regulates flow of substances in and out of the cell

Peroxisome
Makes enzymes that oxidize some toxic chemicals

Smooth endoplasmic reticulum
Network of tubes and flat, curved sacs that helps to transport materials through the cell; site of calcium storage; main location of fat metabolism

CELL TRANSPORT

Materials are constantly being transported in and out of the cell via the cell membrane. Such materials include fuel for generating energy, or building blocks for protein assembly. Some cells secrete signalling molecules to communicate with the rest of the body. The cell membrane is studded with proteins that help transport, allow cells to communicate, and identify a cell to other cells. The membrane is permeable to some molecules, but others need active transport through special channels in the membrane. Cells have three methods of transport: diffusion, facilitated diffusion, and active transport.

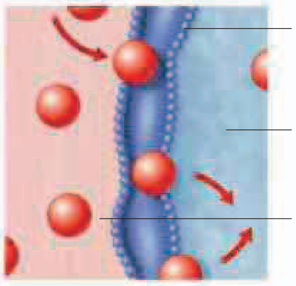

Diffusion

Molecules passively cross the membrane from areas of high to low concentration. Water and oxygen both cross by diffusion.

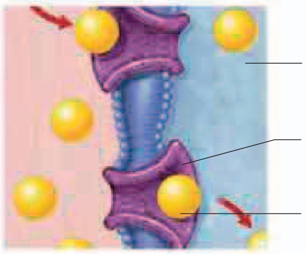

Facilitated diffusion

A carrier protein, or protein pore, binds with a molecule outside the cell, then changes shape and ejects the molecule into the cell.

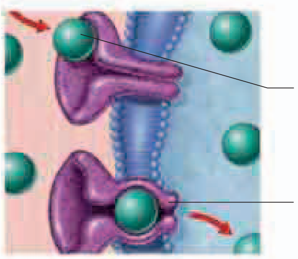

Active transport

Molecules bind to a receptor site on the cell membrane, triggering a protein, which changes into a channel that molecules travel through.

MAKING NEW BODY CELLS

While the cells lining the mouth are replaced every couple of days, some of the nerve cells in the brain have been there since before birth. Stem cells are specialized cells that constantly divide and give rise to new cells, such as blood cells. Cell division requires that a cell's DNA is accurately copied and then shared equally between two "daughter" cells, by a process called mitosis. The chromosomes are first replicated before being pulled to opposite ends of the cell. The cell then divides to produce two daughter cells, with the cytoplasm and organelles being shared between the two cells.

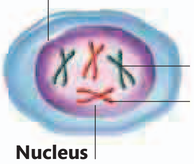

1 Preparation

The cell produces proteins and new organelles, and duplicates its DNA. The DNA condenses into X-shaped chromosomes.

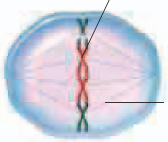

2 Alignment

The chromosomes line up along a network of filaments – spindle – linked to a larger network, called the cytoskeleton.

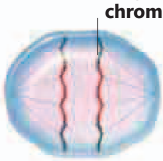

3 Separation

The chromosomes are pulled apart and move to opposite ends of the cell. Each end has an identical set of chromosomes.

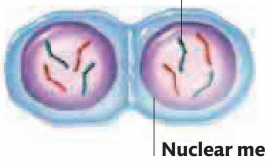

4 Splitting

The cell now splits into two, with the cytoplasm, cell membrane, and remaining organelles being shared roughly equally between the two daughter cells.

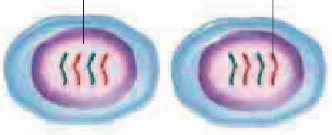

5 Offspring

Each daughter cell contains a complete copy of the DNA from the parent cell; this enables it to continue growing, and eventually divide itself.

BODY COMPOSITION

Cells are building blocks from which the human body is made. Some cells work alone—such as red blood cells, which carry oxygen—but many are organized into tissues. These tissues form organs, which in turn form specific body systems, where cells with various functions join forces to accomplish one or more tasks.

CELL TYPES

There are more than 200 types of cells in the body, each type specially adapted to its own particular function. Every cell contains the same genetic information, but not all of the genes are "switched on" in every cell. It is this pattern of gene expression that dictates the cell's appearance, its behavior, and its role in the body. A cell's fate is largely determined before birth, influenced by its position in the body and the cocktail of chemical messengers that it is exposed to in that environment. Early during development, stem cells begin to differentiate into three layers of specialized cells called the ectoderm, endoderm, and mesoderm. Cells of the ectoderm will form the skin and nails, the epithelial lining of the nose, mouth, and anus, the eyes, and the brain and spinal cord. Cells of the endoderm will become the inner linings of the digestive tract, the respiratory linings, and glandular organs. Mesoderm cells will develop into muscles, and the circulatory and excretory systems.

STEM CELLS

A few days after fertilization, an embryo consists of a ball of "embryonic stem cells" (ESCs). These cells have the potential to develop into any type of cell in the body. Scientists are trying to harness this property to grow replacement body parts. As the embryo grows, the stem cells become increasingly restricted in their potential and most are fully differentiated by the time we are born. Only a small number of stem cells remain in parts of the adult body, including in the bone marrow. Scientists believe that these cells could also be used to help cure diseases.

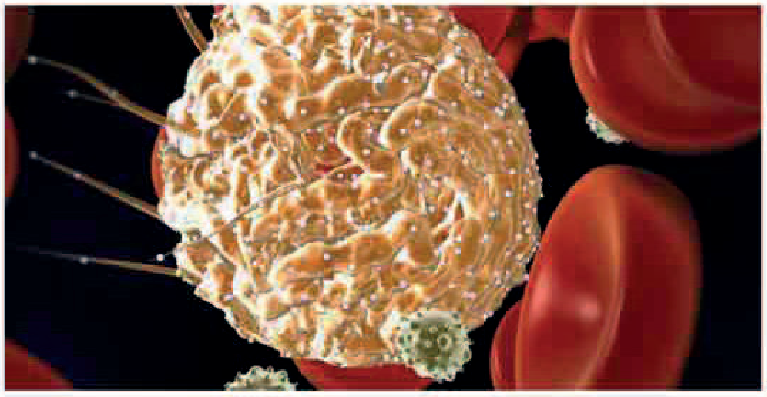

Adult stem cells
Adult stem cells, such as the large white cell in this image, are present in bone marrow, where they multiply and produce millions of blood cells, including red blood cells, also seen here.

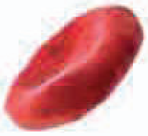

Red blood cells
Unlike other cells, red blood cells lack a nucleus and organelles. Instead, they have an oxygen-carrying protein (hemoglobin), which gives blood its red color.

Epithelial cells
The skin cells and the cells lining the lungs and reproductive tracts are among the barrier cells, called epithelial cells, which line the cavities and surfaces of the body.

Adipose (fat) cells
These cells are highly adapted for storing fat – the bulk of their interior is taken up by a large droplet of semi-liquid fat. When we gain weight, they fill up with more fat.

Nerve cells
These electrically excitable cells transmit electrical signals down an extended stem called an axon. Found throughout the body, they enable us to feel sensations.

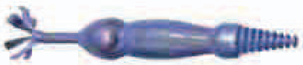

Photoreceptor cells
Located in the eye, these areof two types—cone and rod (left). Both have a light-sensitive pigment and generate electrical signals when struck by light, helping us see.

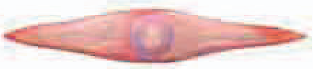

Smooth muscle cells
One of three types of muscle cell, smooth muscle cells are spindle-shaped cells found in the arteries and the digestive tract that produce contractions.

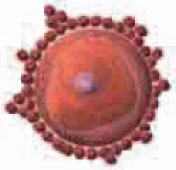

Ovum (egg) cells
The largest cells in the female human body, eggs are female reproductive cells. Like sperm, they have just 23 chromosomes.

Sperm cells
Sperm are male reproductive cells, with tails that enable them to swim up the female reproductive tract and fertilize an egg.

LEVELS OF ORGANIZATION

The overall organization of the human body can be visualized as a hierarchy of levels. At its lowest are the body's basic chemical constituents, forming organic molecules, such as DNA, the key to life. As the hierarchy ascends, the number of components in each of its levels—cells, tissues, organs, and systems—decreases, culminating in a single being at its apex. Cells are the smallest living units, with each adapted to carry out a specific role, but not in isolation. Groups of similar cells form tissues, which in turn form organs with a specific role. Organs with a common purpose are linked within a system, such as the cardiovascular system, shown right. These interdependent systems combine to produce a human body (see pp.16–17).

TISSUE TYPES

Cells of the same kind often group together to form tissues that carry out a specific function. However, not all cells within a tissue are necessarily identical. The four main types of tissue in the human body are muscle, connective tissue, nervous tissue, and epithelial tissue. Within these groups, different forms of these tissues can have very different appearances and functions. For example, blood, bone, and cartilage are all types of connective tissue, but so are fat layers, tendons, ligaments, and the fibrous tissue that holds organs and epithelial layers in place. Organs such as the heart and lungs are composed of several different kinds of tissue.

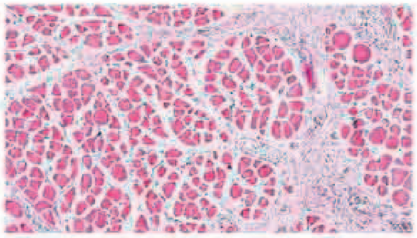

Skeletal muscle

This tissue enables voluntary limb movements. Its cells are arranged into bundles of fibers that connect to bones via tendons. They are packed with filaments that slide over one another to produce contractions.

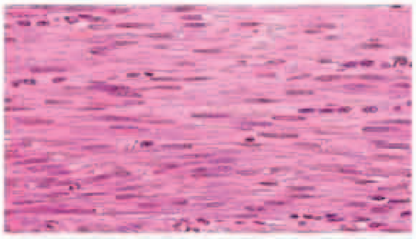

Smooth muscle

Able to contract involuntarily in long, wavelike motions, smooth muscle is found in sheets on the walls of specific organs. It is vital for maintaining blood pressure and for pushing food through the system.

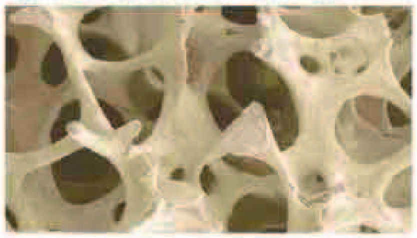

Spongy bone

Spongy bone is found in the center of bones (see p.24) and is softer and weaker than compact bone. The latticelike spaces in spongy bone are filled with bone marrow or connective tissue.

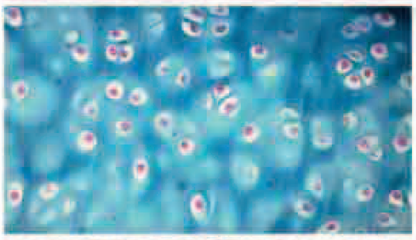

Cartilage

Its high water-content makes this tissue rubbery yet stiff. It is composed of cells, called chondrocytes, set in a matrix of gel-like materials secreted by the cells. Cartilage is found in the bone joints and in the ear and nose.

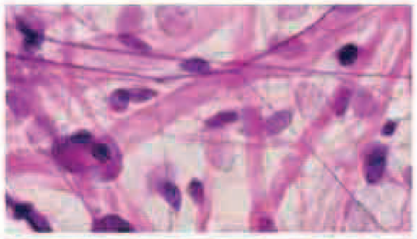

Loose connective tissue

This tissue type also contains cells called fibroblasts, which secrete loosely-organized fibers that make the tissue pliable. Loose connective tissue holds organs in place and provides support.

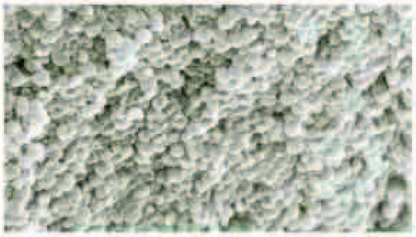

Dense connective tissue

This contains fibroblast cells, which secrete the fibrous protein called type 1 collagen. The fibers are organized into a regular parallel pattern, making the tissue very strong. This tissue type occurs in the base layer of skin.

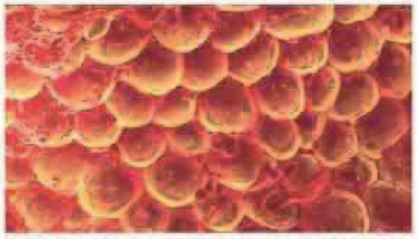

Adipose tissue

A type of connective tissue, adipose tissue is composed of fat cells called adipocytes, as well as some immune cells, fibroblast cells, and blood vessels. Its main task is to store energy, and to protect and insulate the body.

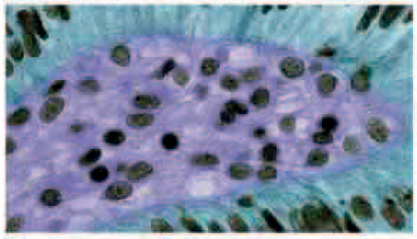

Epithelial tissue

This tissue forms a covering or lining for internal and external body surfaces. Some epithelial tissues can secrete substances such as digestive enzymes; others can absorb substances like food or water.

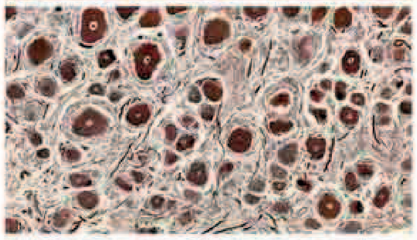

Nervous tissue

This forms the brain, spinal cord, and the nerves that control movement, transmit sensation, and regulate many body functions. It is mainly made up of networks of nerve cells (see opposite).

---

1. CHEMICALS

Key among the chemicals inside all cells is DNA (see pp.10–11). Its long molecules provide the instructions for making proteins. These, in turn, perform many roles, such as building cells.

2. CELLS

While cells may differ in size and shape, all have the same basic features: an outer membrane; organelles floating within jellylike cytoplasm; and a nucleus containing DNA (see pp.12–13).

3. HEART TISSUE

One of the three types of muscle tissue, cardiac muscle is found only in the walls of the heart. Its cells contract together, as a network, to make the heart squeeze and pump blood.

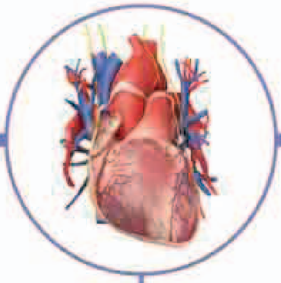

4. HEART

Like other organs, the heart is made of several types of tissue, including cardiac muscle tissue. Among the others are connective and epithelial tissues, found in the chambers and valves.

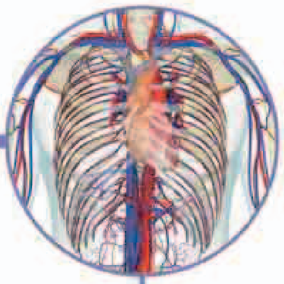

5. CARDIOVASCULAR SYSTEM

The heart, blood, and blood vessels form the cardiovascular system. Its main tasks are to pump blood, deliver nutrients, and remove waste from the tissue cells.

BODY SYSTEMS

The human body can do many different things. It can digest food, think, move, even reproduce and create new life. Each of these tasks is performed by a different body system—a group of organs and tissues working together to complete that task. However, good health and body efficiency rely on the different body systems working together in harmony.

SYSTEM INTERACTION

Think about what your body is doing right now. You are breathing, your heart is beating, and your blood pressure is under control. You are also conscious and alert. If you were to start running, specialized cells called chemoreceptors would detect a change in your body's metabolic requirements and signal to the brain to release adrenaline. This would in turn signal the heart to beat faster, boosting blood circulation and providing more oxygen to the muscles. After a while, cells in the hypothalamus might detect an increase in temperature and send a signal to the skin to produce sweat, which would evaporate and cool you down.

The individual body systems are linked together by a vast network of positive and negative feedback loops. These use signalling molecules such as hormones and electrical impulses from nerves to maintain equilibrium. Here, the basic components and functions of each system are described, and examples of system interactions are examined.

BREATHING IN AND OUT

The mechanics of breathing rely upon an interaction between the respiratory and muscular systems. Together with three accessory muscles, the intercostal muscles and the diaphragm contract to increase the volume of the chest cavity. This draws air down into the lungs. A different set of muscles is used during forced exhalation. These rapidly compress the chest cavity, forcing air out of the lungs.

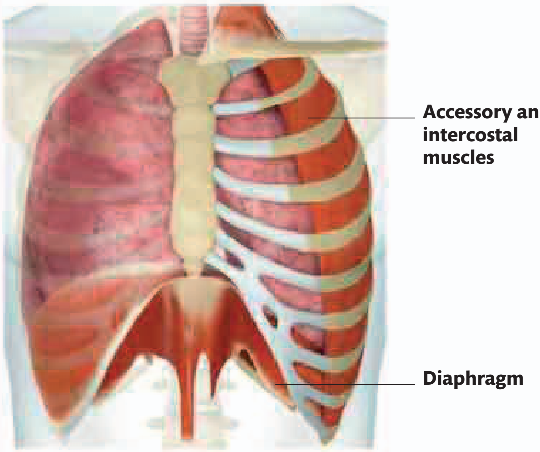

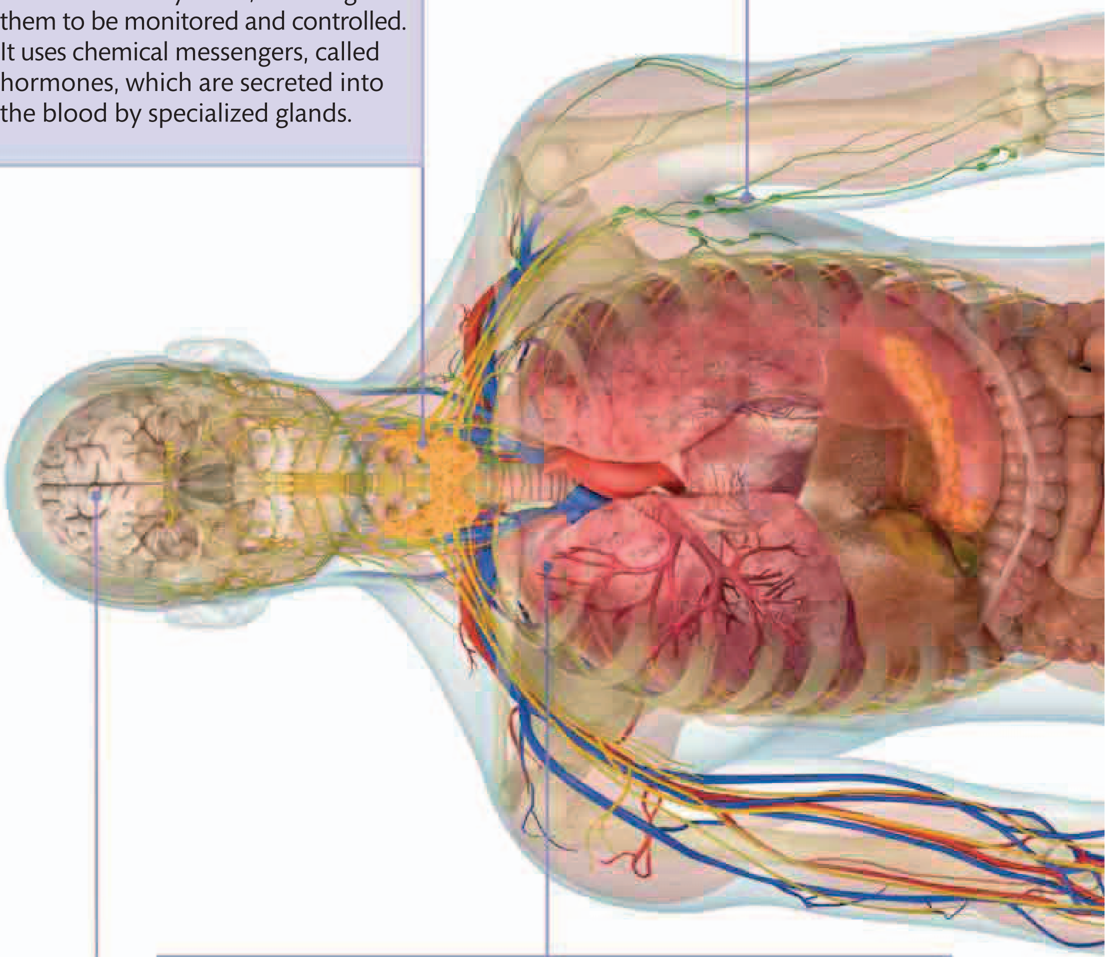

LYMPHATIC AND IMMUNE SYSTEM

The lymphatic system includes a network of vessels and nodes, which drain tissue fluid and return it to the veins. Its main functions are to maintain fluid balance within the cardiovascular system and to distribute immune cells around the body. Movement of lymphatic fluid relies on the muscles within the muscular system.

ENDOCRINE SYSTEM

The endocrine system communicates with the other systems, enabling them to be monitored and controlled. It uses chemical messengers, called hormones, which are secreted into the blood by specialized glands.

RESPIRATORY SYSTEM

Every cell in the body needs oxygen and must dispel carbon dioxide in order to function. The respiratory system ensures this by breathing air into the lungs, where the exchange of these molecules occurs between the air and blood. The cardiovascular system transports oxygen and carbon dioxide between the cells and the lungs.

NERVOUS SYSTEM

The brain, spinal cord, and nerves collect, process, and disseminate information from the body's internal and external environments. The nervous system communicates through networks of nerve cells, which connect with other systems. The brain controls and monitors all the other systems to ensure they are performing normally.

DIGESTIVE SYSTEM

As well as oxygen, every cell needs energy in order to function. The digestive system processes and breaks down the food we eat so that a variety of nutrients can be absorbed from the intestines into the circulatory system. These are then delivered to the cells of every body system in order to provide them with energy.

MUSCULAR SYSTEM

The muscular system is made up of three types of muscle: skeletal, smooth, and cardiac. It is responsible for generating movement – both in the limbs and within the other body systems. For example, smooth muscle aids the digestive system by helping to propel food down the esophagus and through the stomach, intestines, and rectum. The respiratory system needs the thoracic muscles to contract to fill the lungs with air (see opposite).

SKELETAL SYSTEM

This system uses bones, cartilage, ligaments, and tendons to provide the body with structural support and protection. It encases much of the nervous system within a skull and vertebrae, and the vital organs of the respiratory and circulatory systems within the ribcage. The skeletal system also supports our immune and the circulatory systems by manufacturing red and white blood cells.

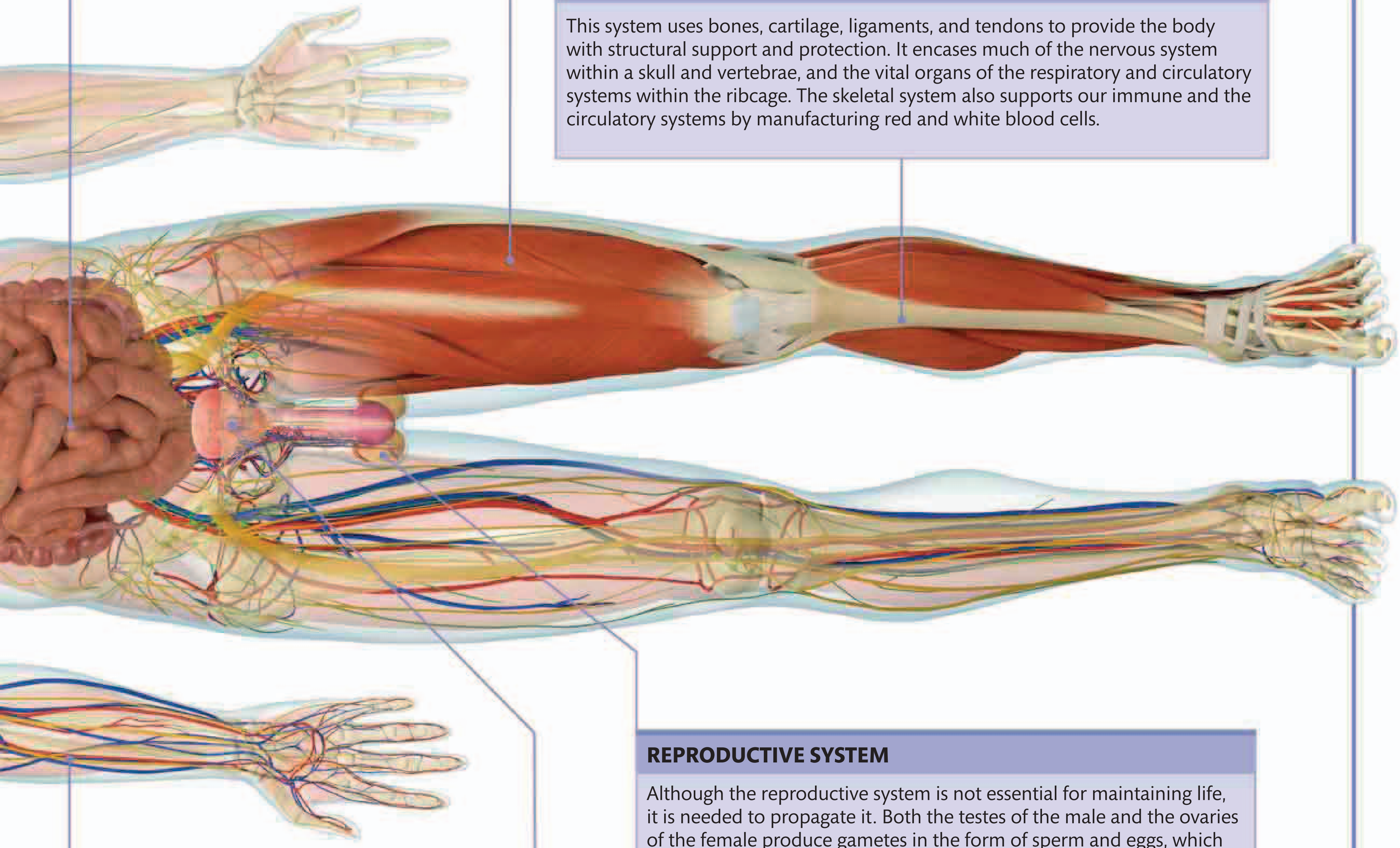

REPRODUCTIVE SYSTEM

Although the reproductive system is not essential for maintaining life, it is needed to propagate it. Both the testes of the male and the ovaries of the female produce gametes in the form of sperm and eggs, which fuse to create an embryo. The testes and ovaries also produce hormones including estrogen and testosterone, thus forming part of the endocrine system.

CARDIOVASCULAR SYSTEM

The cardiovascular system uses blood to carry oxygen from the respiratory system and nutrients from the digestive system to cells of all the body's systems. It also removes waste products from these cells. At the center of the cardiovascular system lies the muscular heart, which pumps blood through the blood vessels.

URINARY SYSTEM

The urinary system filters and removes many of the waste products generated by cells of the body. It does this by filtering blood through the kidneys and producing urine, which is collected in the bladder and then excreted through the urethra. The kidneys also help regulate blood pressure within the cardiovascular system by ensuring that the correct amount of water is reabsorbed by the blood.

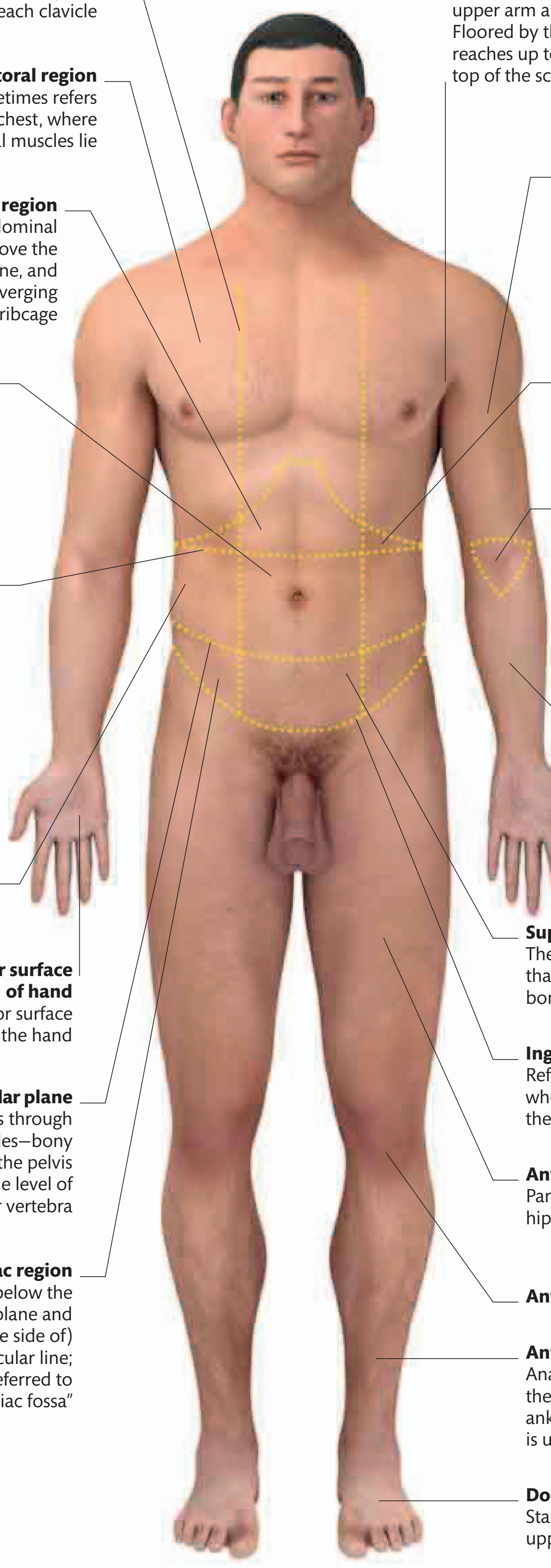

Midclavicular line
A vertical line running down from the midpoint of each clavicle

Pectoral region
The chest; sometimes refers to just the upper chest, where the pectoral muscles lie

Epigastric region
Area of the abdominal wall above the transpyloric plane, and framed by the diverging margins of the ribcage

Umbilical region
Central region of the abdomen, around the umbilicus (navel)

Transpyloric plane
Horizontal plane joining tips of the ninth costal cartilages, at the margins of the ribcage; level with the first lumbar vertebra and pylorus of the stomach

Lumbar region
The side of the abdominal wall

Palmar surface of hand
Anterior surface of the hand

Intertubercular plane
This plane passes through the iliac tubercles—bony landmarks on the pelvis—and lies at the level of the fifth lumbar vertebra

Iliac region
The area below the intertubercular plane and lateral to (to the side of) the midclavicular line; may also be referred to as the "iliac fossa"

Axilla
The armpit; more precisely, the pyramid-shaped part of the body between the upper arm and the side of the thorax. Floored by the skin of the armpit, it reaches up to the level of the clavicle, top of the scapula, and first rib

Anterior surface of arm
"Anterior" means front, and refers to the body when it is in this "anatomical position". "Arm" relates to the part of the upper limb between the shoulder and the elbow

Hypochondrial region
The abdominal region under the ribs on each side

Cubital fossa
Triangular area anterior to the elbow, bounded above by a line between the epicondyles of the humerus on each side, and framed below by the pronator teres and brachioradialis muscles

Anterior surface of forearm
The part of the body between the elbow and the wrist

Suprapubic region
The part of the abdomen that lies just above the pubic bones of the pelvis

Inguinal region
Refers to the groin area, where the thigh meets the trunk

Anterior surface of thigh
Part of the body between the hip and the knee

Anterior surface of knee

Anterior surface of leg
Anatomically, "leg" just refers to the part between the knee and ankle, and the term "lower limb" is used for the whole limb

Dorsum of foot
Standing upright, this is the upper surface of the foot

Anterior surface regions

The anterior surface of the body is divided into general anatomical areas by imaginary lines drawn on the body. The location of many of these lines is defined by reference to underlying features such as muscles or bony prominences; for example, the cubital fossa is defined by reference to epicondyles of the humerus, and the pronator teres and brachioradialis muscles. Many of the regions may be divided into smaller areas. For instance, the upper part of the anterior thigh contains the femoral triangle.

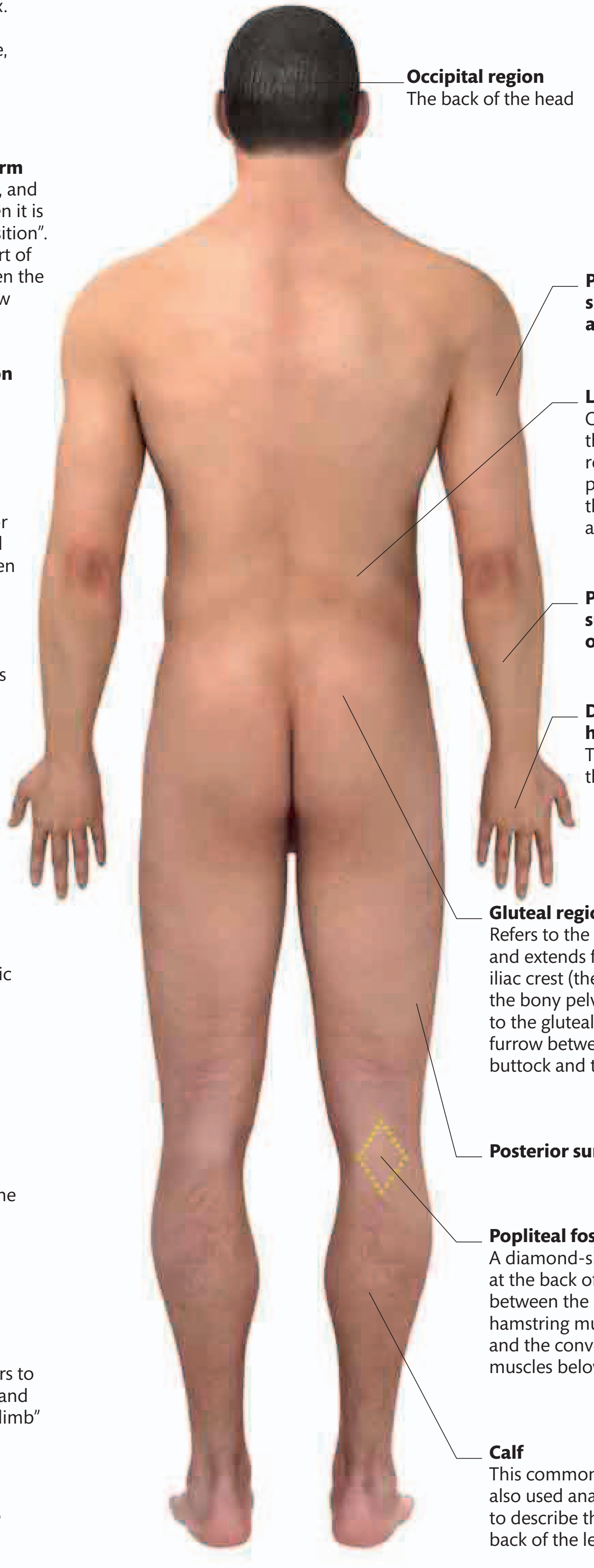

Occipital region
The back of the head

Posterior surface of arm

Lumbar region
On the back of the body it refers to the part between the thorax and the pelvis

Posterior surface of forearm

Dorsum of hand
The back of the hand

Gluteal region
Refers to the buttock, and extends from the iliac crest (the top of the bony pelvis) above, to the gluteal fold (the furrow between the buttock and thigh) below

Posterior surface of thigh

Popliteal fossa
A diamond-shaped cavity at the back of the knee, between the diverging hamstring muscles above and the converging calf muscles below

Calf
This common term is also used anatomically, to describe the fleshy back of the leg

Posterior surface regions

As with the anterior surface, the posterior surface can also be divided into anatomical regions. The anterior surface of the abdomen is divided by planes and mapped into nine regions—allowing doctors to describe precisely where areas of tenderness or lumps are felt on abdominal examination. The back is not divided into as many regions. This illustration shows some of the terms used for the broader regions of back of the body.

TERMINOLOGY AND PLANES

Anatomical language allows us to describe the structure of the body accurately and unambiguously. The illustrations here show the main regions of the anterior (front) and posterior (back) surfaces of the body. Sometimes it is easier to understand anatomy by dividing the body into two dimensional slices. The orientation of these planes through the body also have specific anatomical names. There are also terms to describe the relative position of structures within the body.

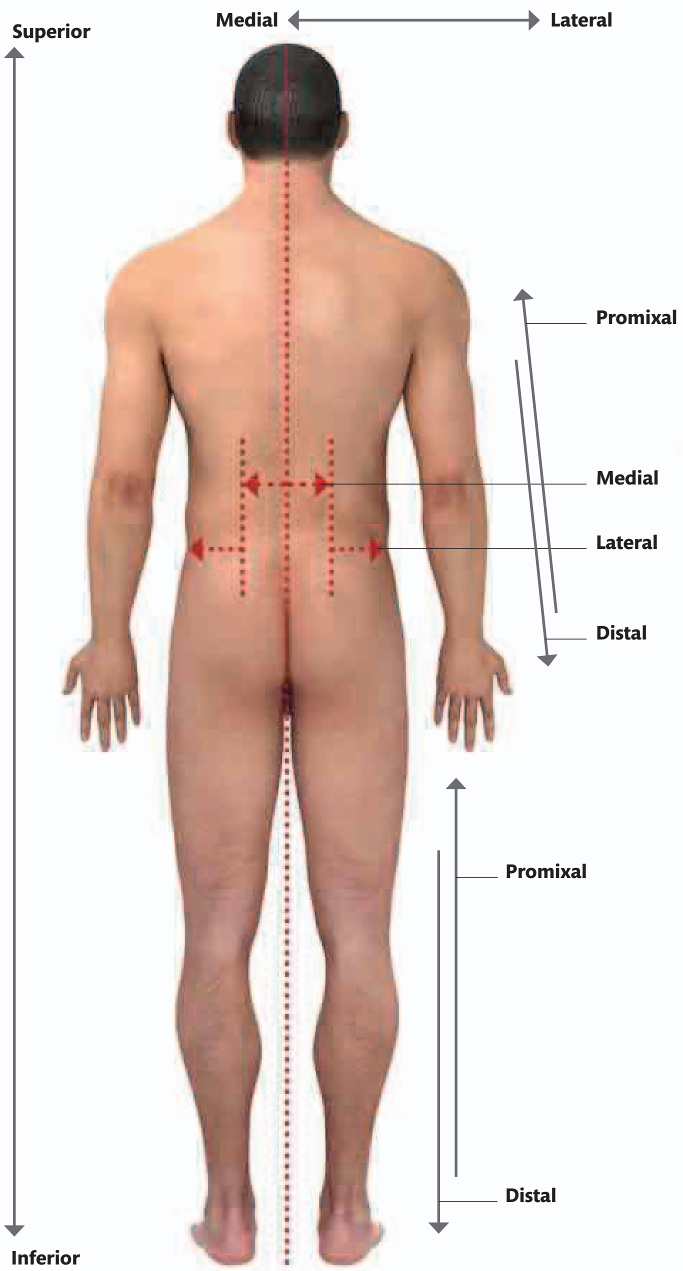

Directions and relative positions

As well as defining parts of the body, anatomical terminology also allows us to precisely and concisely describe the relative positions of various structures. These terms always refer back to relative positions of structures when the body is in the "anatomical position" (shown above). Medial and lateral describe positions of structures toward the midline, or toward the side of the body, respectively. Superior and inferior refer to vertical position—toward the top or bottom of the body. Proximal and distal are useful terms, describing a relative position toward the center or periphery of the body.

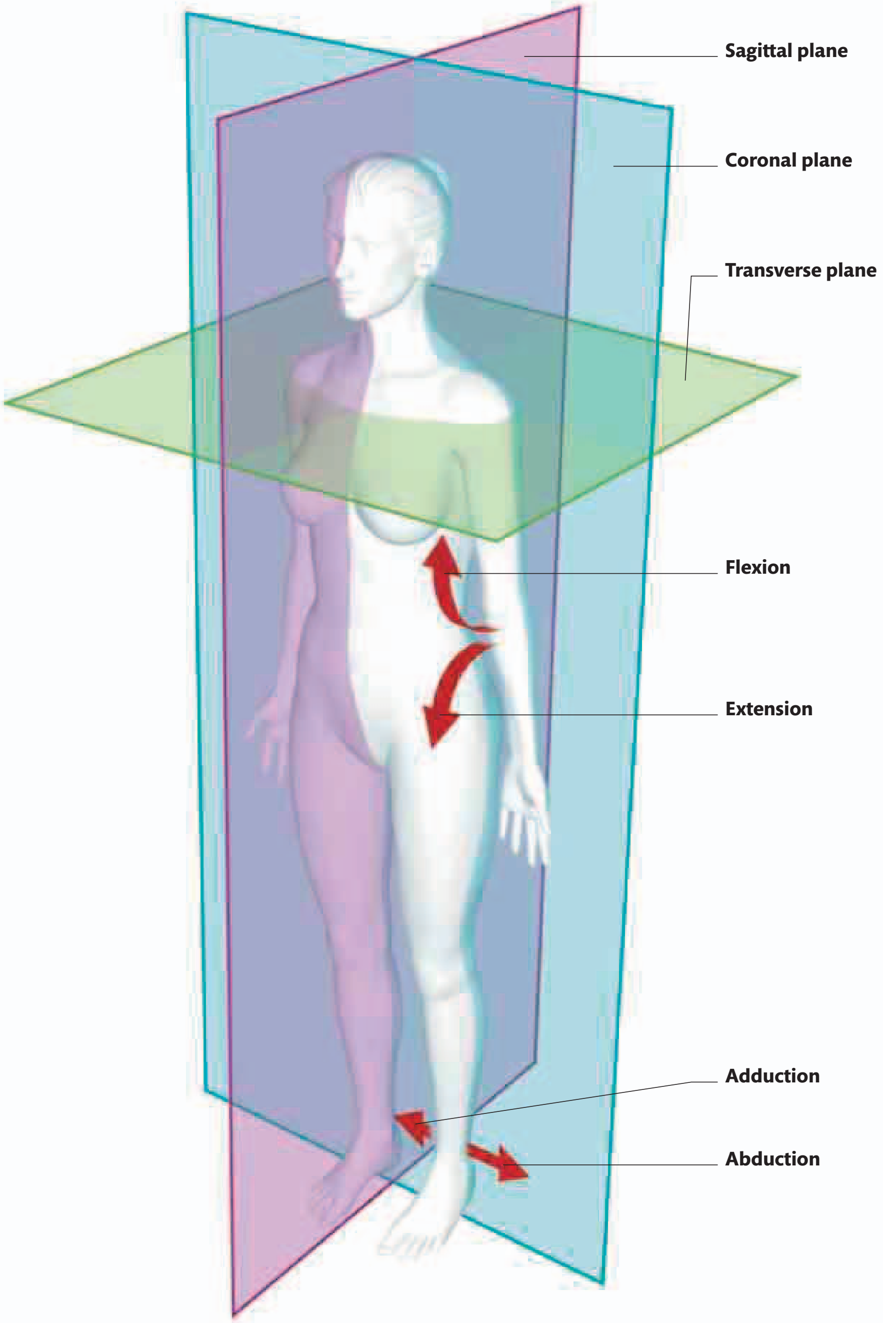

Anatomical terms for planes and movement

The diagram above shows the three planes—sagittal, coronal, and transverse—cutting through a body. It also illustrates some medical terms that are used to describe certain movements of body parts: flexion decreases the angle of a joint, such as the elbow, while extension increases it; adduction draws a limb closer to the sagittal plane, while abduction moves it further away from that plane.

02

Body Systems

The human body is made up of eleven functional systems. No one system works in isolation: for example the endocrine and nervous systems work closely to keep the body regulated, while the respiratory and cardiovascular systems combine to deliver vital oxygen to cells. To build the clearest picture of how the body is put together it is, however, helpful to strip back our anatomy and consider it system by system. This chapter gives an overview of the basic structure of each system before looking at each region in detail.

022 Skin, hair, and nail
024 Skeletal system
068 Muscular system
110 Nervous system
146 Respiratory system
154 Cardiovascular system
180 Lymphatic and immune system
192 Digestive system
204 Urinary system
208 Reproductive system
216 Endocrine system

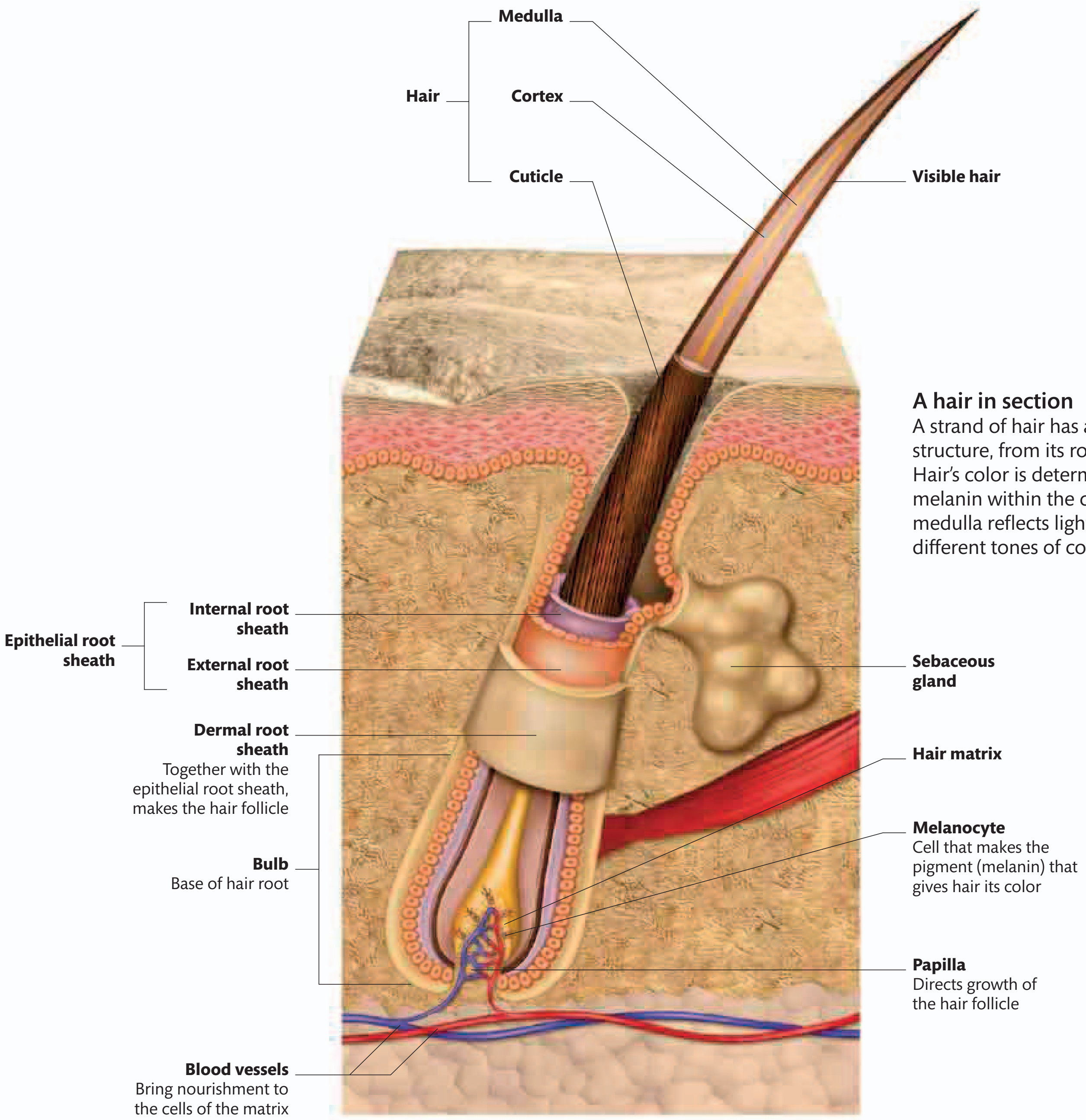
Medulla
Hair
Cortex
Cuticle
Visible hair
Epithelial root sheath
Internal root sheath
External root sheath
Sebaceous gland
Dermal root sheath
Together with the epithelial root sheath, makes the hair follicle
Hair matrix
Melanocyte
Cell that makes the pigment (melanin) that gives hair its color
Bulb
Base of hair root
Papilla
Directs growth of the hair follicle
Blood vessels
Bring nourishment to the cells of the matrix

SECTION THROUGH A HAIR

A hair in section

A strand of hair has a multilayered structure, from its root to the tip. Hair's color is determined by melanin within the cortex; the medulla reflects light so the different tones of color are seen.

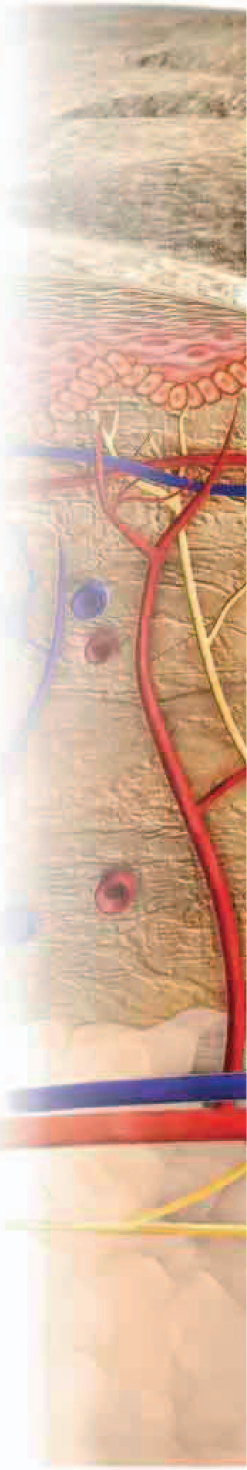

SKIN, HAIR, AND NAIL

The skin is our largest organ, weighing about 9lb (4kg) and covering an area of about 21 square feet (2 square meters). It forms a tough, waterproof layer, which protects us from the elements. However, it offers much more than protection: the skin lets us appreciate the texture and temperature of our environment; it regulates body temperature; it allows excretion in sweat; communication through blushing; gripping thanks to ridges on our fingertips, and vitamin D production in sunlight. Thick head hairs and fine body hairs help to keep us warm and dry. All visible hair is in fact dead; hairs are only alive at their root. Constantly growing and self-repairing, nails protect fingers and toes but also enhance their sensitivity.

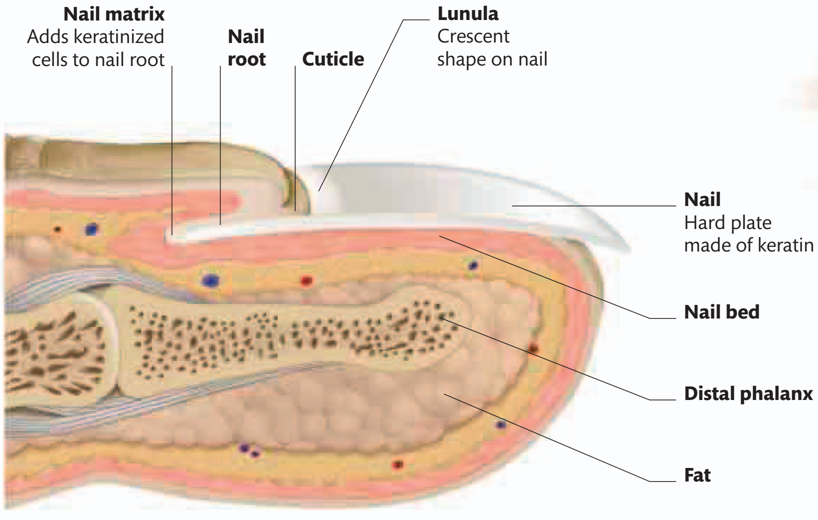
Nail matrix
Adds keratinized cells to nail root
Nail root
Cuticle
Lunula
Crescent shape on nail
Nail
Hard plate made of keratin
Nail bed
Distal phalanx
Fat

SECTION THROUGH A NAIL

SECTION THROUGH SKIN

Skin in section

In just one square centimeter (⅙in^2) of the skin, there are, on average, 21½in (55cm) of nerve fibers, 27½in (70cm) of blood vessels, 15 sebaceous glands, 100 sweat glands, and over 200 sensory receptors.

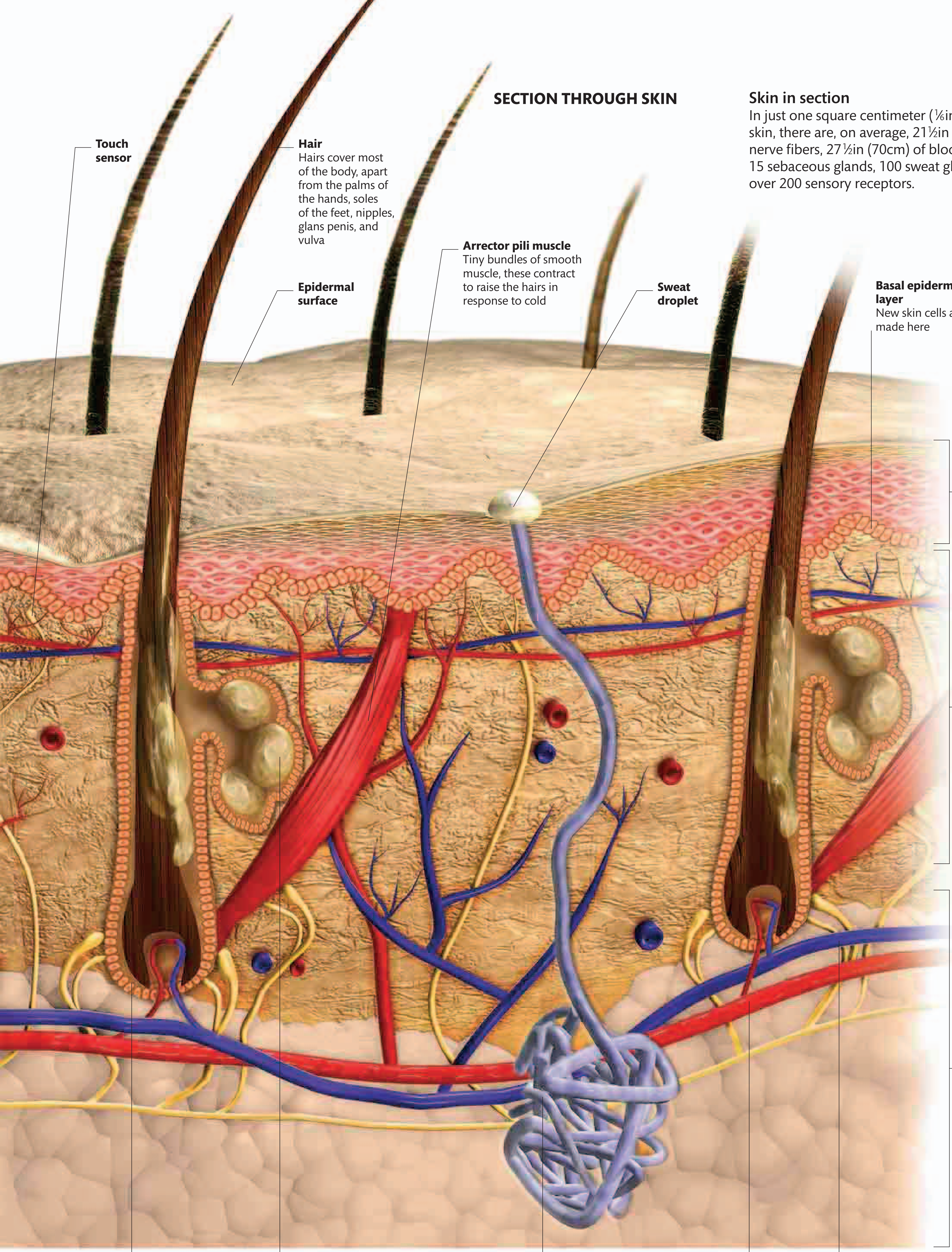

Touch sensor

Hair Hairs cover most of the body, apart from the palms of the hands, soles of the feet, nipples, glans penis, and vulva

Epidermal surface

Arrector pili muscle Tiny bundles of smooth muscle, these contract to raise the hairs in response to cold

Sweat droplet

Basal epidermal layer New skin cells are made here

Epidermis Outermost layer of the skin, comprising constantly renewing layers of cells called keratinocytes

Dermis Inner layer, composed of dense connective tissue, containing the nerves and blood vessels that supply the skin

Hypodermis Layer of loose connective tissue under the skin; also known as superficial fascia

Hair follicle Cuplike structure in the dermis or hypodermis forms a socket for a hair

Sebaceous gland Secretes sebum into the hair follicle; this oily secretion helps to waterproof the skin and keep it supple, and also has an antibacterial effect

Sweat gland Coiled tube extend upward from the dermis to open at a pore on the surface of the epidermis

Arteriole

Venule

SKELETAL SYSTEM OVERVIEW

The human skeleton gives the body its shape, supports the weight of all our other tissues, provides attachment for muscles, and forms a system of linked levers that the muscles can move. It also protects delicate organs and tissues, such as the brain within the skull, the spinal cord within the arches of the vertebrae, and the heart and lungs within the ribcage. The skeletal system differs between the sexes. This is most obvious in the pelvis, which is usually wider in a woman than in a man. The skull also varies, with men having a larger brow and more prominent areas for muscle attachment on the back of the head. The entire skeleton tends to be larger and more robust in a man.

BONE STRUCTURE

Most of the human skeleton develops first as cartilage, which is later replaced by bone throughout fetal development and childhood. Both bone and cartilage are connective tissues. Bone tissue consists of cells that are embedded in a mineralized matrix, making it extremely hard and strong. Bone is full of blood vessels and repairs easily.

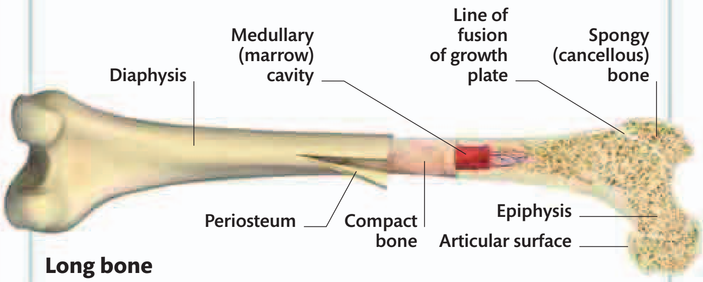
Diaphysis
Medullary (marrow) cavity
Line of fusion of growth plate
Spongy (cancellous) bone
Periosteum
Compact bone
Epiphysis
Articular surface

Long bone
Long bones are found in the limbs, and include the femur (shown above), humerus, radius, ulna, tibia, fibula, metatarsals, metacarpals, and phalanges. A long bone has flared ends (epiphyses), which narrow to form a neck (metaphysis), tapering down into a cylindrical shaft (diaphysis). Cartilage growth plates near the ends of bones allow rapid growth in childhood, but disappear by adulthood.

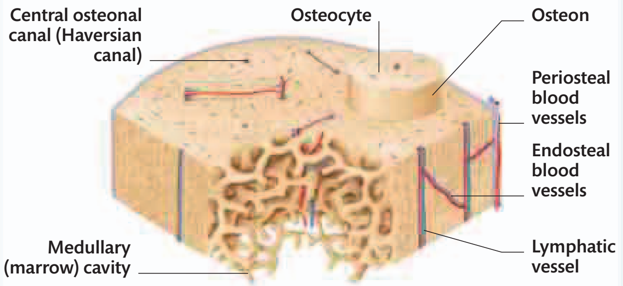
Central osteonal canal (Haversian canal)
Osteocyte
Osteon
Periosteal blood vessels
Endosteal blood vessels
Medullary (marrow) cavity
Lymphatic vessel

Compact bone
Also called cortical bone, compact bone is made up of osteons: concentric cylinders of bone tissue, each around 0.1–0.4mm in diameter, with a central vascular canal. Bone is full of blood vessels: those in the osteons connect to blood vessels within the medullary cavity of the bone as well as to vessels in the periosteum on the outside.

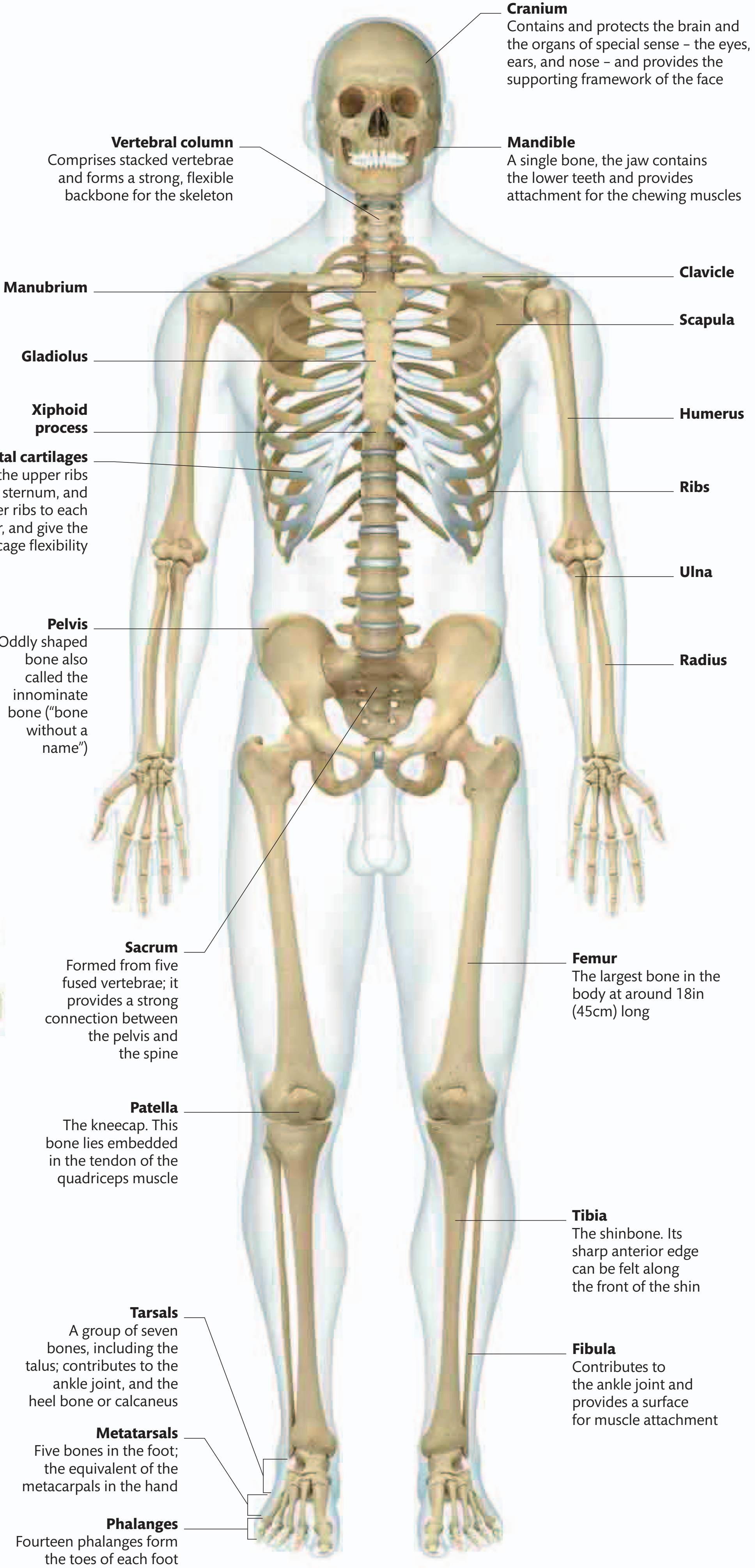
Sternum
Manubrium
Gladiolus
Xiphoid process
Clavicle
Scapula
Humerus
Ribs
Ulna
Radius

Cranium
Contains and protects the brain and the organs of special sense – the eyes, ears, and nose – and provides the supporting framework of the face

Vertebral column
Comprises stacked vertebrae and forms a strong, flexible backbone for the skeleton

Mandible
A single bone, the jaw contains the lower teeth and provides attachment for the chewing muscles

Costal cartilages
Attach the upper ribs to the sternum, and lower ribs to each other, and give the ribcage flexibility

Pelvis
Oddly shaped bone also called the innominate bone ("bone without a name")

Sacrum
Formed from five fused vertebrae; it provides a strong connection between the pelvis and the spine

Femur
The largest bone in the body at around 18in (45cm) long

Patella
The kneecap. This bone lies embedded in the tendon of the quadriceps muscle

Tibia
The shinbone. Its sharp anterior edge can be felt along the front of the shin

Tarsals
A group of seven bones, including the talus; contributes to the ankle joint, and the heel bone or calcaneus

Fibula
Contributes to the ankle joint and provides a surface for muscle attachment

Metatarsals
Five bones in the foot; the equivalent of the metacarpals in the hand

Phalanges
Fourteen phalanges form the toes of each foot

ANTERIOR (FRONT)

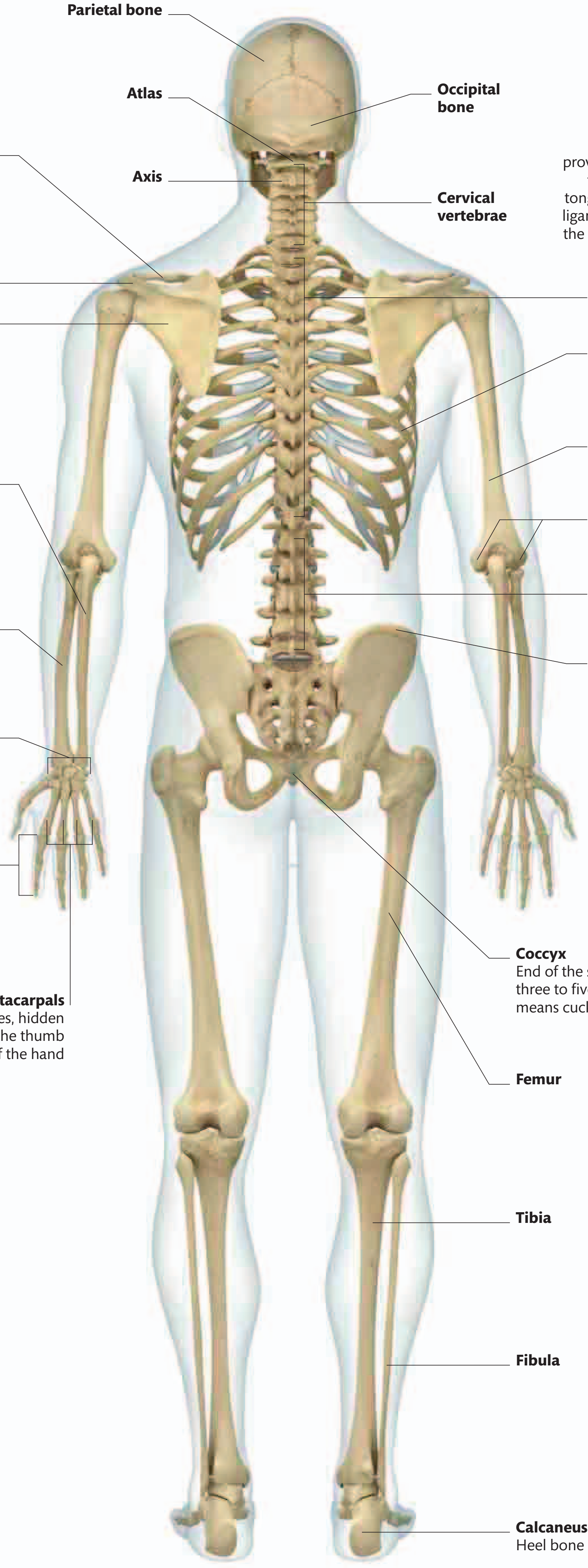
Parietal bone
Atlas
Occipital bone
Axis
Cervical vertebrae
Acromion
Thoracic vertebrae
Humerus
Epicondyles
Lumbar vertebrae
Femur
Tibia
Fibula

Clavicle Traces a sinuous curve at the base of the neck; it acts as a strut supporting the shoulder

Scapula Connects the arm to the trunk, and provides a secure but mobile anchor for the arm, allowing the shoulders to be retracted backward, protracted forward, and elevated

Ribs Twelve pairs of curving bones form the ribcage

Ulna Wide at its proximal end, where it articulates with the humerus at the elbow, this bone tapers down to a pointed styloid process near the wrist

Radius Forearm bone; it can rotate around the ulna to alter the orientation of the hand

Ilium Named after the Latin for hip

Carpals Eight small bones in the base of the hand. Two articulate with the radius to form the wrist joint

Phalanges Fourteen bones in each hand: two form the thumb, with three (proximal, middle, and distal) in each finger

Coccyx End of the spine made up of three to five tiny vertebrae; means cuckoo in Greek

Metacarpals Five slender bones, hidden in the base of the thumb and the palm of the hand

Calcaneus Heel bone

POSTERIOR (BACK)

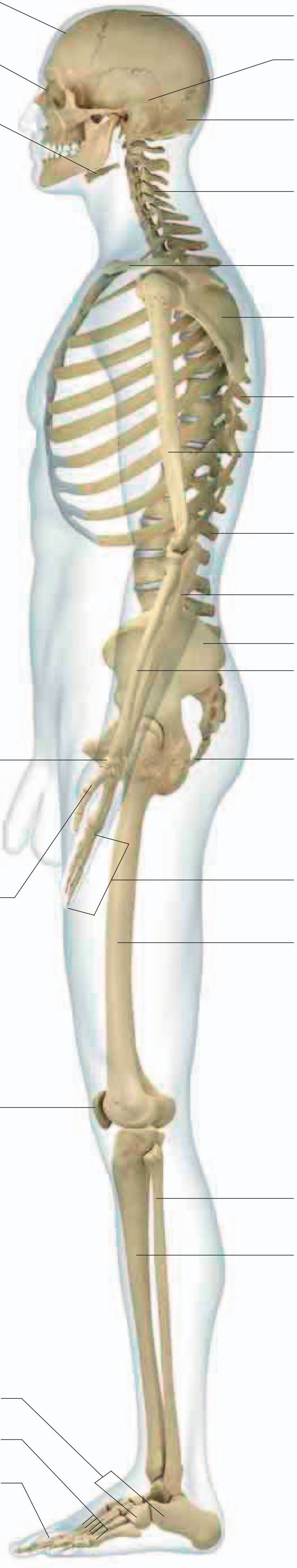
Frontal bone
Nasal bone
Cranium
Temporal bone
Occipital bone
Cervical vertebra
Clavicle
Scapula
Thoracic vertebra
Lumbar vertebra
Ulna
Pelvis
Radius
Carpal
Coccyx
Metacarpal
Phalanges
Femur
Patella
Fibula
Tibia
Tarsals
Metatarsals
Phalanx

Hyoid bone U-shaped bone; provides attachments for muscles of the tongue, as well as the ligaments suspending the larynx in the neck

Humerus Latin for shoulder

SIDE

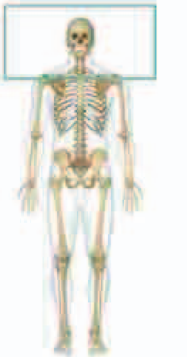

HEAD AND NECK

The skull comprises the cranium and mandible. It houses and protects the brain and the eyes, ears, nose, and mouth. It encloses the first parts of the airway and of the alimentary canal, and provides attachment for the muscles of the head and neck. The cranium itself comprises more than 20 bones that meet each other at fibrous joints called sutures. In addition to the main bones labeled on these pages, there are sometimes extra bones along the sutures. In a young adult skull, the sutures are visible as tortuous lines between the cranial bones; they gradually fuse with age. The mandible of a newborn baby is in two halves, with a fibrous joint in the middle. The joint fuses during early infancy, so that the mandible becomes a single bone.

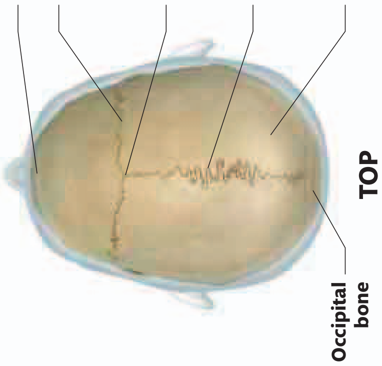

Frontal bone
Coronal suture Where the frontal and parietal bones meet; crosses the skull's highest part (the crown)
Bregma Where the sagittal and coronal sutures meet
Sagittal suture Joint on the midline (sagittal plane) where parietal bones meet
Parietal bone From the Latin for wall
Occipital bone
TOP

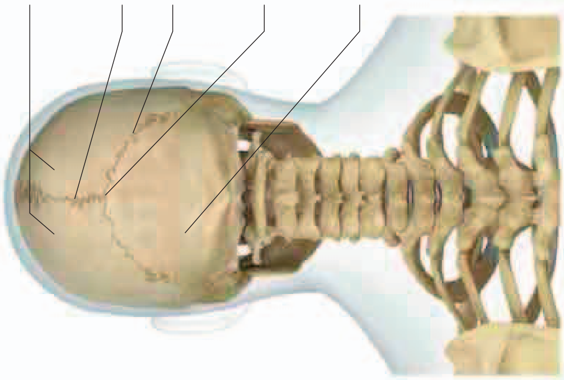

Parietal bones Paired bones forming most of the roof and sides of skull
Sagittal suture
Lambdoid suture Joint between occipital and parietal bones
Lambda Point where the sagittal suture meets the lambdoid suture
Occipital bone Forms lower part of back of skull, and back of cranial base
BACK

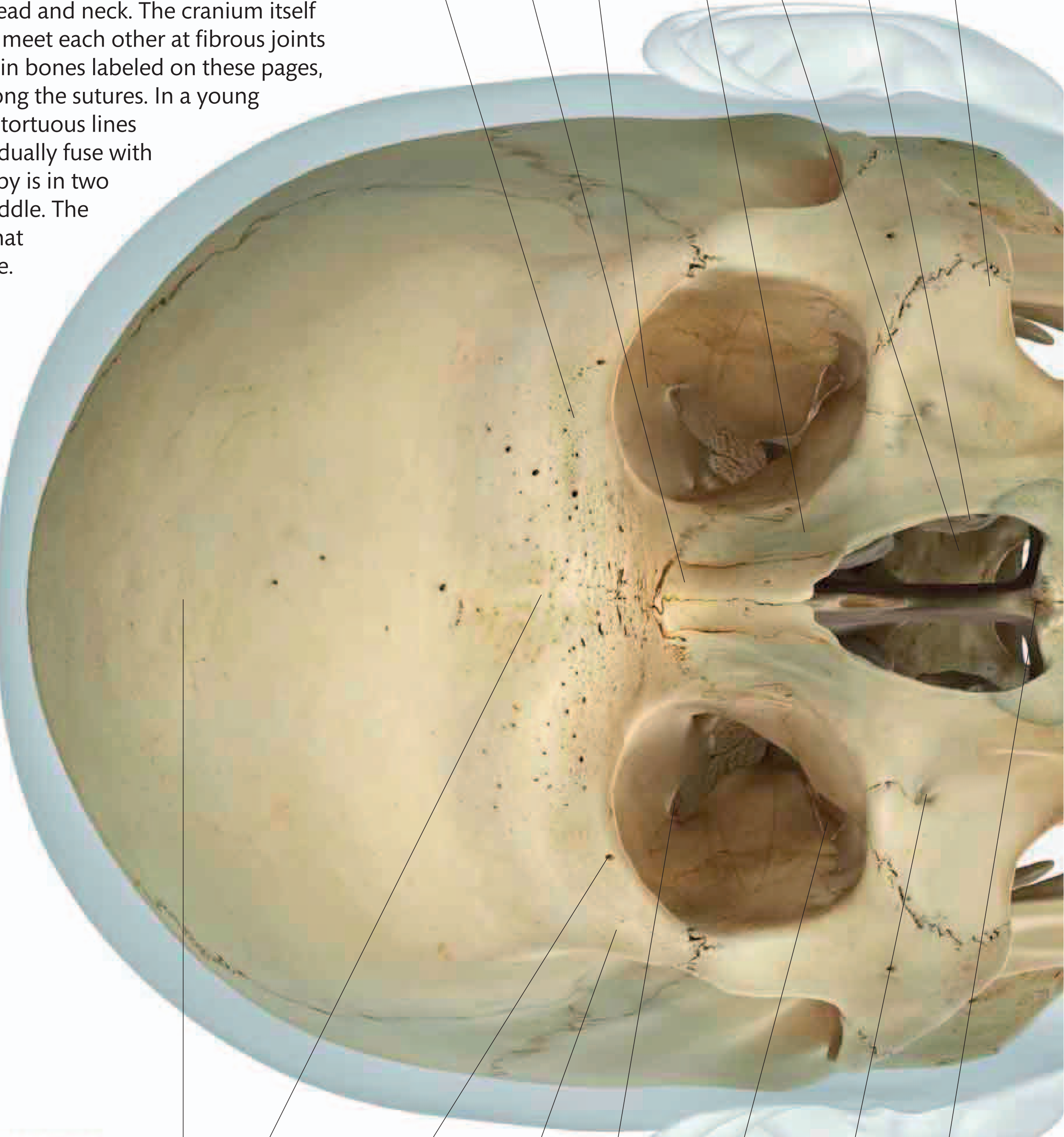

Superciliary arch Also called the supraorbital ridge, or brow ridge; from the Latin for eyebrow
Nasal bone Two small bones form the bony bridge of the nose
Orbit Technical term for the eye-socket, from the Latin for wheel track
Frontal process of maxilla Rises up on the medial (inner) side of the orbit
Piriform aperture Pear-shaped (piriform) opening; also called the anterior nasal aperture
Inferior nasal concha Lowest of the three curled protrusions on the lateral wall of the nasal cavity
Zygomatic process of maxilla Part of the maxilla that projects laterally (to the side)
Frontal bone
Glabella Area between the two superciliary arches; glabella comes from the Latin for smooth, and refers to the bare area between the eyebrows
Supraorbital foramen The supraorbital nerve passes through this hole to supply sensation to the forehead
Zygomatic process of frontal bone Runs down to join the frontal process of the zygomatic bone
Superior orbital fissure Gap between the sphenoid bone's greater and lesser wings, opening into the orbit
Inferior orbital fissure Gap between the maxilla and the greater wing of the sphenoid bone, opening into the back of the orbit
Infraorbital foramen Hole for infraorbital branch of maxillary nerve to supply sensation to the cheek
Nasal crest Where the two maxillae meet; the vomer (part of the septum) sits on the crest

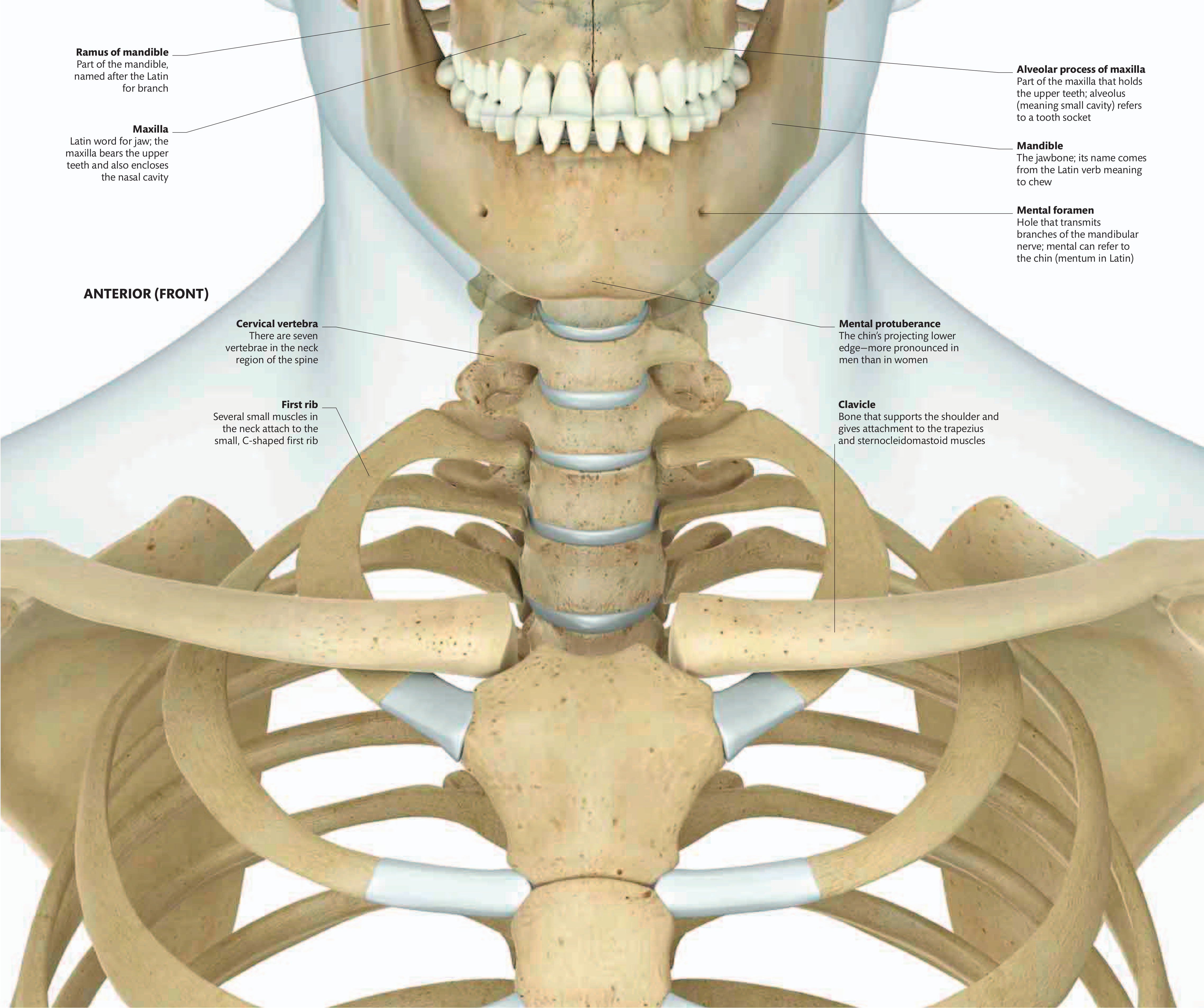

ANTERIOR (FRONT)

Ramus of mandible
Part of the mandible, named after the Latin for branch

Maxilla
Latin word for jaw; the maxilla bears the upper teeth and also encloses the nasal cavity

Cervical vertebra
There are seven vertebrae in the neck region of the spine

First rib
Several small muscles in the neck attach to the small, C-shaped first rib

Alveolar process of maxilla
Part of the maxilla that holds the upper teeth; alveolus (meaning small cavity) refers to a tooth socket

Mandible
The jawbone; its name comes from the Latin verb meaning to chew

Mental foramen
Hole that transmits branches of the mandibular nerve; mental can refer to the chin (mentum in Latin)

Mental protuberance
The chin's projecting lower edge—more pronounced in men than in women

Clavicle
Bone that supports the shoulder and gives attachment to the trapezius and sternocleidomastoid muscles

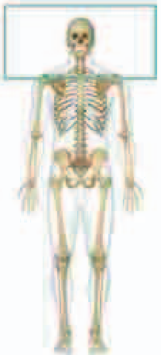

HEAD AND NECK

The cervical spine includes seven vertebrae, the top two of which have specific names. The first vertebra, which supports the skull, is called the atlas, after the Greek god who carried the sky on his shoulders. Nodding movements of the head occur at the joint between the atlas and the skull. The second cervical vertebra is the axis, from the Greek word for axle, so-called because when you shake your head from side to side, the atlas rotates on the axis. In this side view, we can also see more of the bones that make up the cranium, as well as the temporomandibular (jaw) joint between the mandible and the skull. The hyoid bone is also visible. This small bone is a very important anchor for the muscles that form the tongue and the floor of the mouth, as well as muscles that attach to the larynx and pharynx.

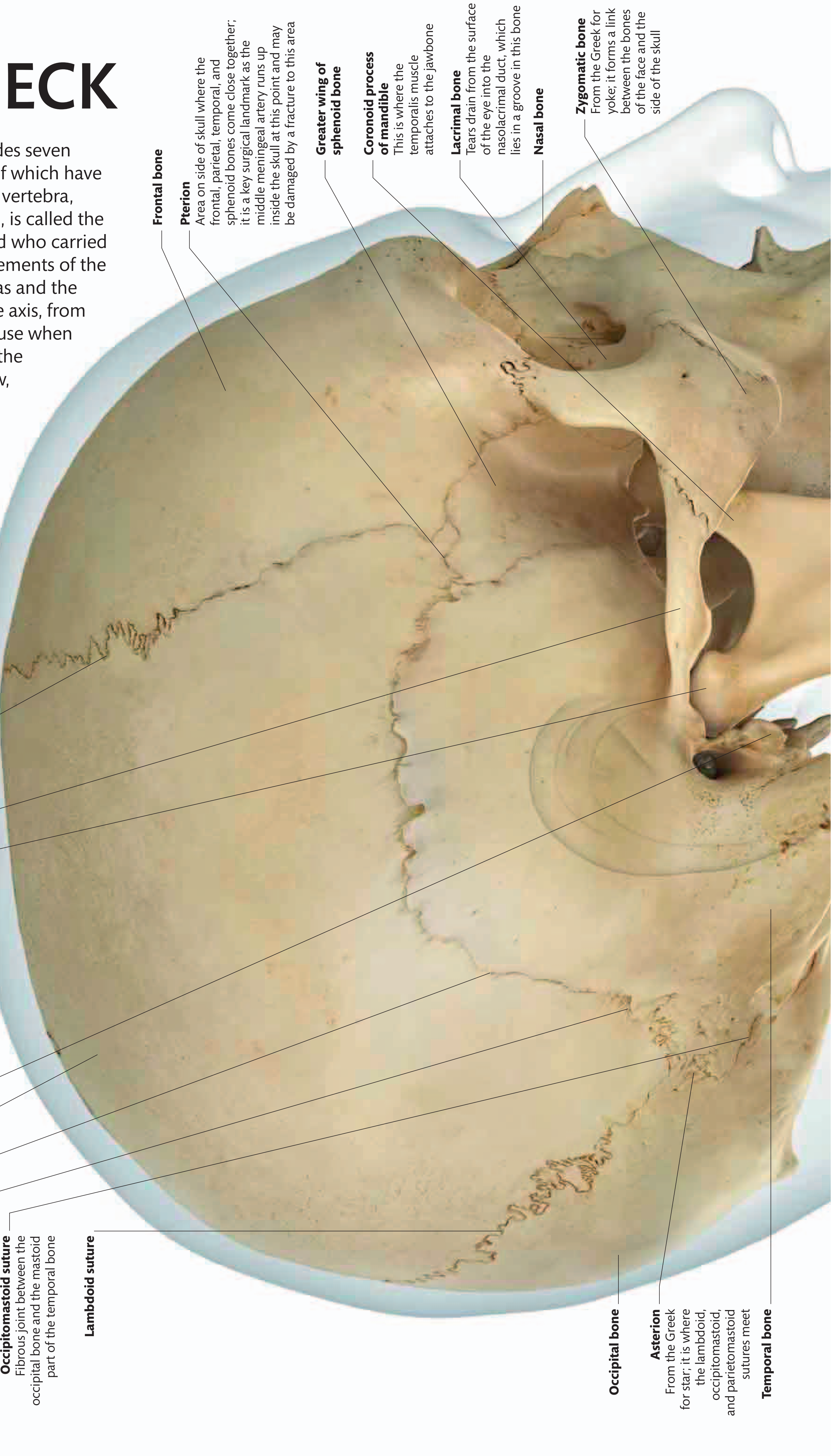

Frontal bone
Pterion
Area on side of skull where the frontal, parietal, temporal, and sphenoid bones come close together; it is a key surgical landmark as the middle meningeal artery runs up inside the skull at this point and may be damaged by a fracture to this area
Greater wing of sphenoid bone
Coronoid process of mandible
This is where the temporalis muscle attaches to the jawbone
Lacrimal bone
Tears drain from the surface of the eye into the nasolacrimal duct, which lies in a groove in this bone
Nasal bone
Zygomatic bone
From the Greek for yoke; it forms a link between the bones of the face and the side of the skull
Coronal suture
Zygomatic arch
Formed by the zygomatic process of the temporal bone projecting forward to meet the temporal process of the zygomatic bone
Condyle
Condylar process projects upwards to end as the condyle, or head of the mandible, which articulates with the cranium at the temporomandibular (jaw) joint
Tympanic part of temporal bone
Forms floor of the external acoustic meatus, at the inner end of which lies the tympanic membrane, or eardrum
Parietal bone
Squamosal suture
The articulation between squamous part of temporal bone and parietal bone
Parietomastoid suture
Here the parietal bone meets the posterior, mastoid part of the temporal bone
Occipitomastoid suture
Fibrous joint between the occipital bone and the mastoid part of the temporal bone
Lambdoid suture
Occipital bone
Asterion
From the Greek for star; it is where the lambdoid, occipitomastoid, and parietomastoid sutures meet
Temporal bone

SIDE

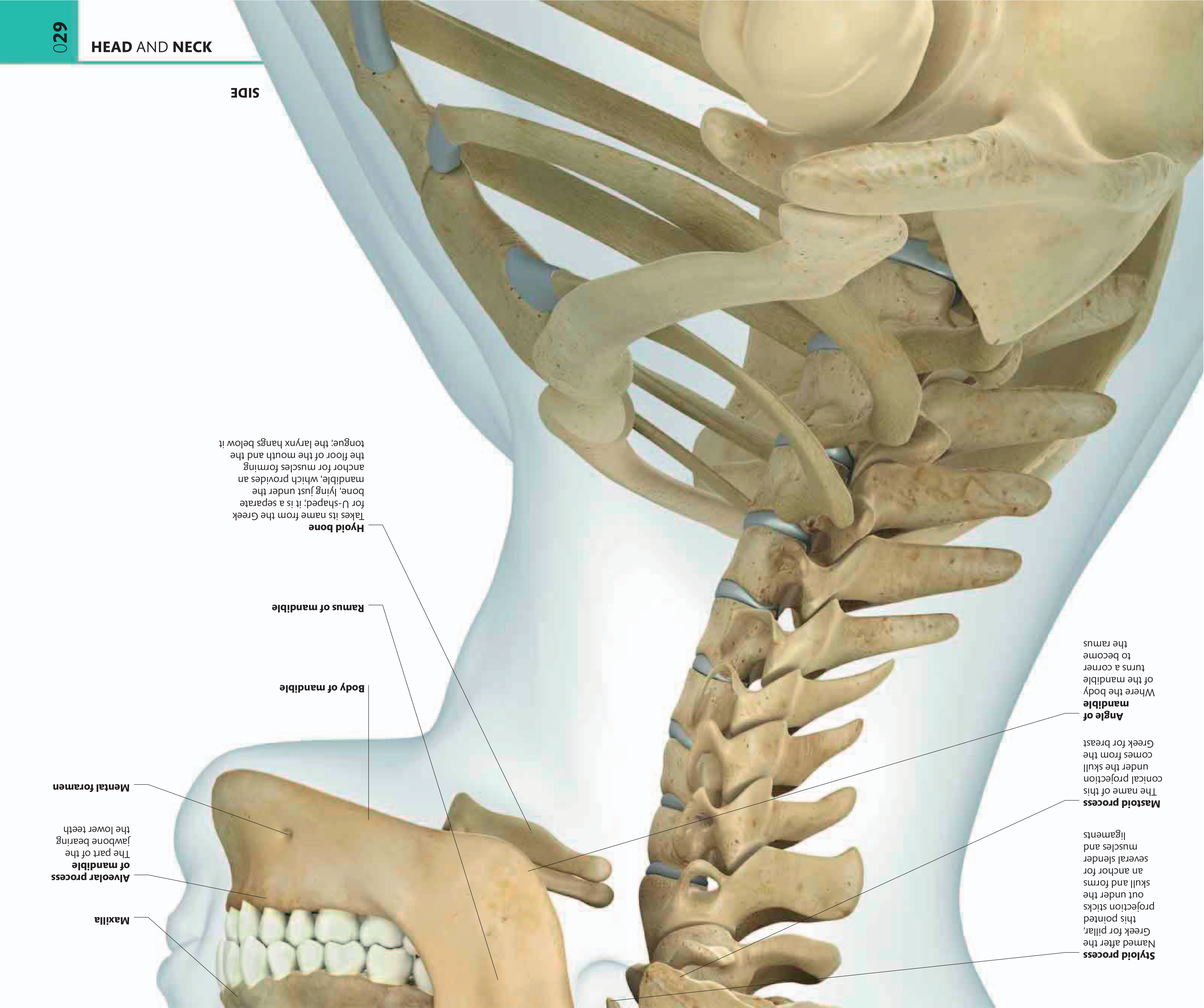
Maxilla
Alveolar process of mandible The part of the jawbone bearing the lower teeth
Mental foramen
Body of mandible
Ramus of mandible
Hyoid bone Takes its name from the Greek for U-shaped; it is a separate bone, lying just under the mandible, which provides an anchor for muscles forming the floor of the mouth and the tongue; the larynx hangs below it
Styloid process Named after the Greek for pillar, this pointed projection sticks out under the skull and forms an anchor for several slender muscles and ligaments
Mastoid process The name of this conical projection under the skull comes from the Greek for breast
Angle of mandible Where the body of the mandible turns a corner to become the ramus

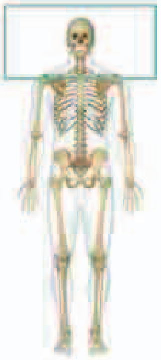

HEAD AND NECK

The most striking features of the skull viewed from these angles are the holes in it. In the middle, there is one large hole—the foramen magnum—through which the brain stem emerges to become the spinal cord. But there are also many smaller holes, most of them paired. Through these holes, the cranial nerves from the brain escape to supply the muscles, skin, and mucosa, and the glands of the head and neck. Blood vessels also pass through some holes, on their way to and from the brain. At the front, we can also see the upper teeth sitting in their sockets in the maxillae, and the bony, hard palate.

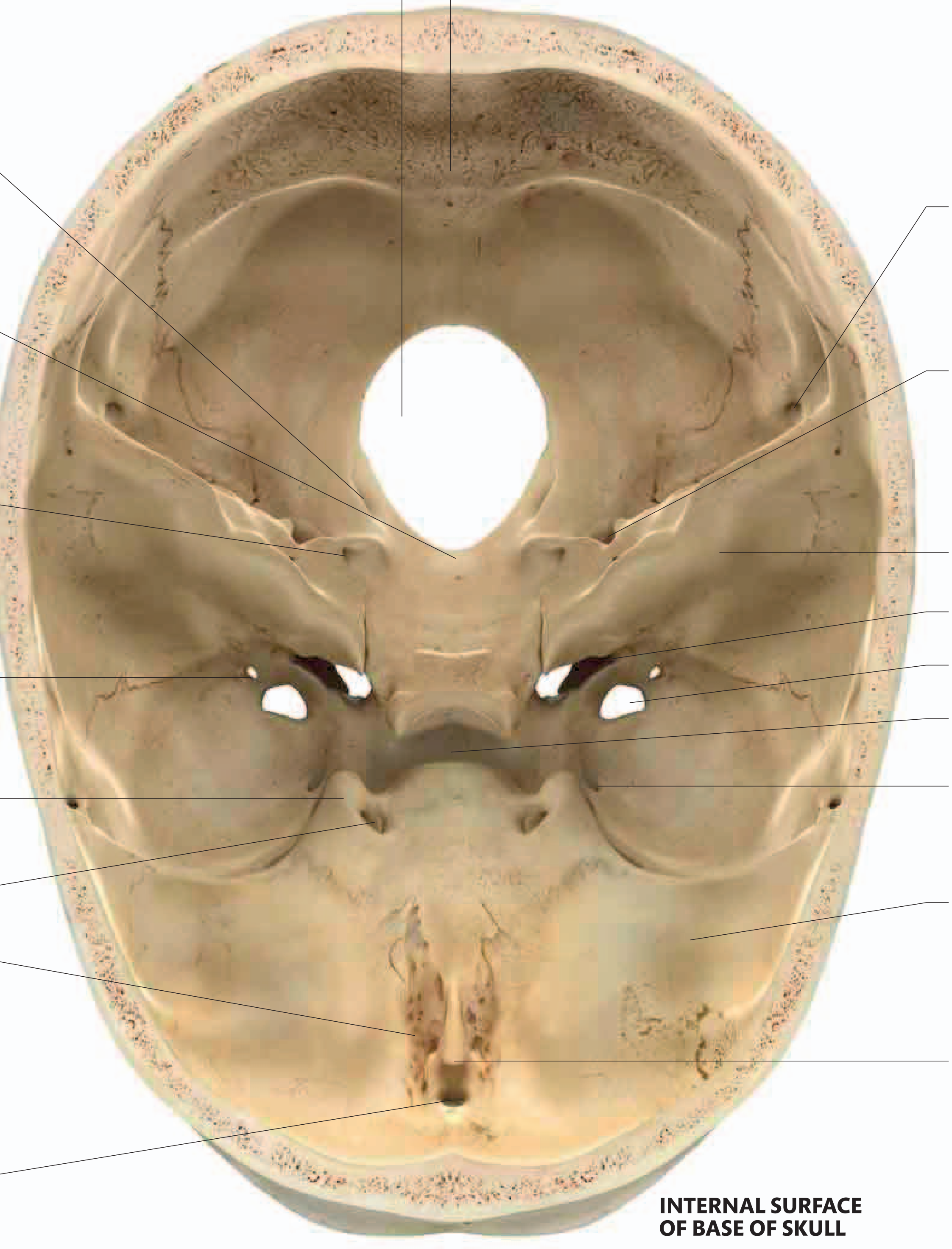

Foramen magnum
Latin for large hole; the brain stem emerges here

Internal occipital protuberance
Located near the confluence of the sinuses, where the superior sagittal, transverse, and straight sinuses (the large veins in the dura mater) meet

Hypoglossal canal
The hypoglossal nerve, supplying the tongue muscles, exits here

Basiocciput
Part of the occipital bone, in front of the foramen magnum, which fuses with the body of the sphenoid bone

Internal acoustic meatus
The facial and vestibulocochlear nerves pass through this hole

Foramen spinosum
Entry point of the middle meningeal artery, which supplies the dura mater and the bones of the skull

Lesser wing of sphenoid bone

Optic canal

Cribriform plate of ethmoid
Area of the ethmoid bone pierced by holes, through which the olfactory nerves pass. Cribriform is Latin for sievelike; ethmoid, taken from Greek, also means sievelike

Foramen caecum
Named after the Latin for blind, this is a blind-ended pit

Mastoid foramen
An emissary (valveless) vein passes out through this hole

Jugular foramen
The internal jugular vein and the glossopharyngeal, vagus, and accessory nerves emerge from this hole

Petrous part of temporal bone

Foramen lacerum

Foramen ovale

Pituitary fossa

Foramen rotundum
The maxillary division of the trigeminal nerve passes through this round hole

Orbital part of frontal bone
Part of the frontal bone that forms the roof of the orbit, and also the floor at the front of the cranial cavity

Crista galli
Vertical crest on the ethmoid bone that takes its name from the Latin for cock's comb; it provides attachment for the falx cerebri—the membrane between the two cerebral hemispheres

INTERNAL SURFACE OF BASE OF SKULL

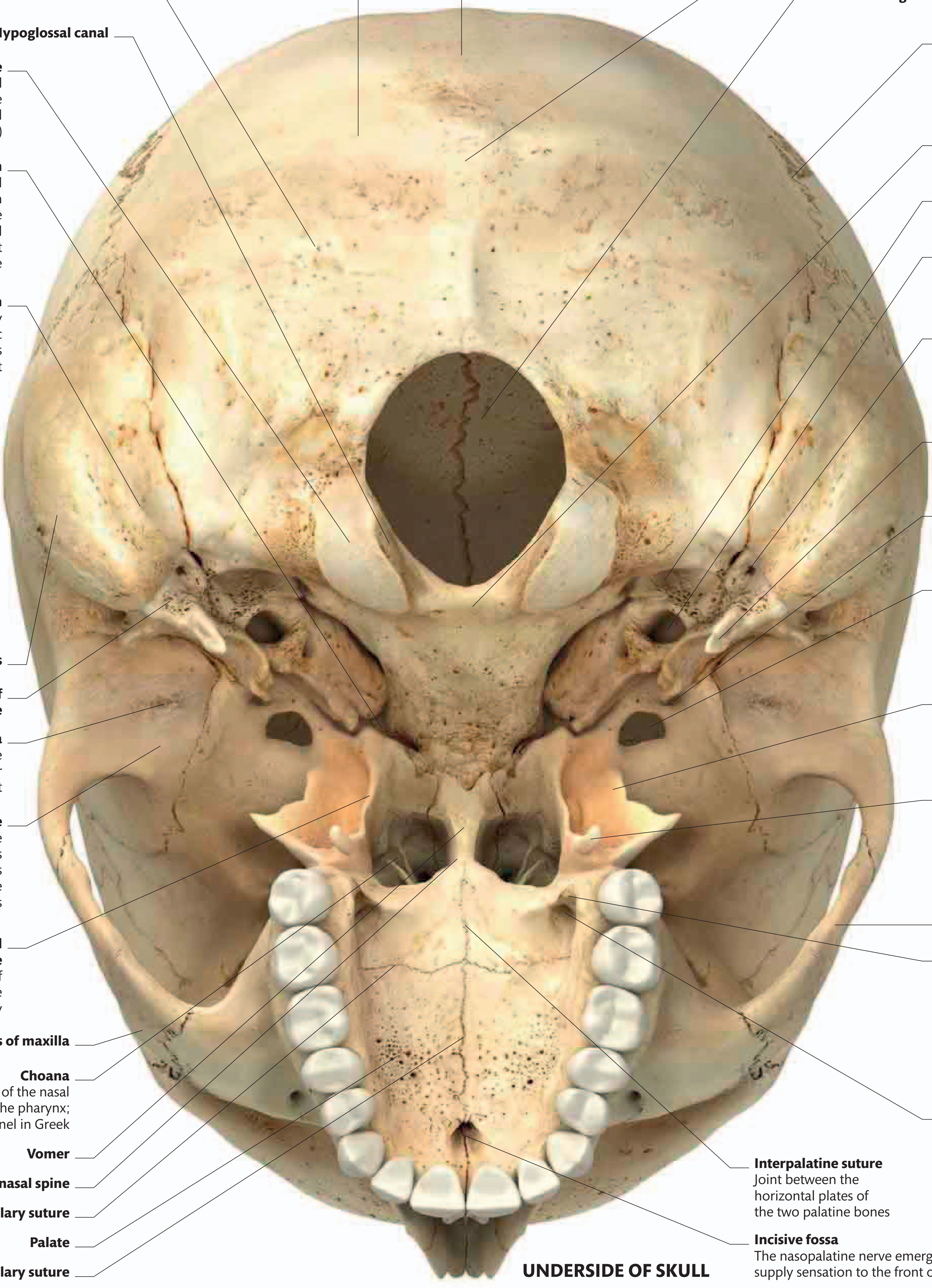

Superior nuchal line
The trapezius and sternocleidomastoid muscles attach to this ridge

Occipital bone

Inferior nuchal line
Slight ridge lying between the attachments of some of the deeper neck muscles

Hypoglossal canal

Occipital condyle
Where the skull articulates with the atlas (first cervical vertebra)

Foramen lacerum
Fibrocartilage-filled hole between the body of the phenoid bone and the petrous part of the temporal bone

Digastric notch
The posterior belly of the digastric muscle attaches to this pit

Mastoid process

Tympanic part of temporal bone

Mandibular fossa
Socket for the temporomandibular (jaw) joint

Articular eminence
The condyle of the mandible moves forward onto this area as the jaw opens

Medial pterygoid plate
Forms the back of the side wall of the nasal cavity

Zygomatic process of maxilla

Choana
Opening of the nasal cavity into the pharynx; from funnel in Greek

Vomer

Posterior nasal spine

Palatomaxillary suture

Palate

Intermaxillary suture

External occipital protuberance

Foramen magnum

Lambdoid suture

Pharyngeal tubercle

Jugular foramen

Carotid canal
The internal carotid artery enters here

Stylomastoid foramen
The facial nerve emerges through this hole

Styloid process

Foramen spinosum

Foramen ovale
The mandibular division of the trigeminal nerve goes through this hole

Lateral pterygoid plate
An anchor point for jaw muscles

Pterygoid hamulus
The word hamulus means small hook in Latin

Zygomatic arch

Lesser palatine foramina
The lesser palatine arteries and nerves pass through this hole and run backward to supply the soft palate

Greater palatine foramen
This hole transmits the greater palatine artery and nerve, which supply the hard palate

Interpalatine suture
Joint between the horizontal plates of the two palatine bones

Incisive fossa
The nasopalatine nerve emerges here to supply sensation to the front of the palate

UNDERSIDE OF SKULL

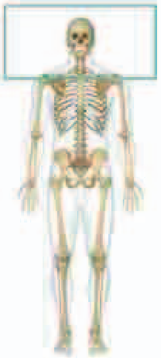

HEAD AND NECK

This section—right through the middle of the skull—lets us in on some secrets. We can clearly appreciate the size of the cranial cavity, which is almost completely filled by the brain, with just a small gap for membranes, fluid, and blood vessels. Some of those blood vessels leave deep grooves on the inner surface of the skull: we can trace the course of the large venous sinuses and the branches of the middle meningeal artery. We can also see that the skull bones are not solid, but contain trabecular bone (or diploe), which itself contains red marrow. Some skull bones also contain air spaces, like the sphenoidal sinus visible here. We can also appreciate the large size of the nasal cavity, hidden away inside the skull.

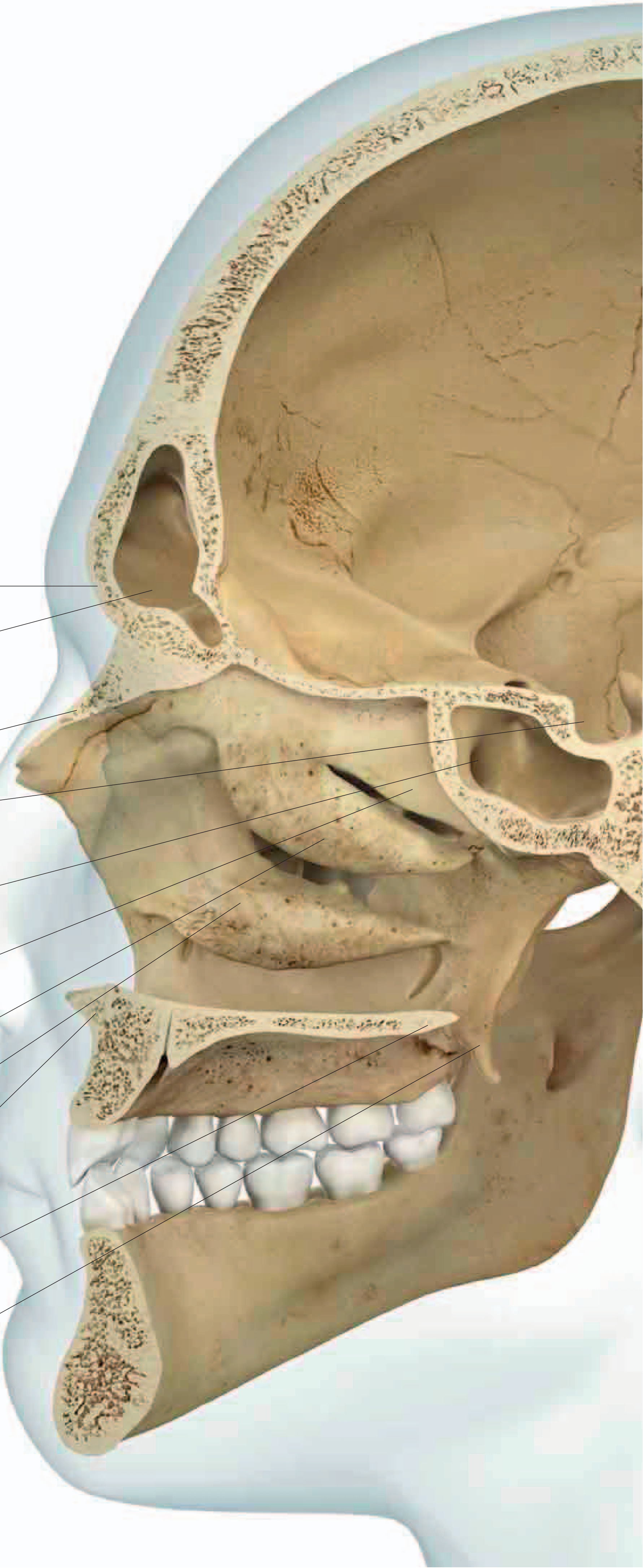

Frontal bone
Forms the anterior cranial fossa, where the frontal lobes of the brain lie, inside the skull

Frontal sinus
One of the paranasal air sinuses that drain into the nasal cavity, this is an air space within the frontal bone

Nasal bone

Pituitary fossa
Fossa is the Latin word for ditch; the pituitary gland occupies this small cavity on the upper surface of the sphenoid bone

Sphenoidal sinus
Another paranasal air sinus; it lies within the body of the sphenoid bone

Superior nasal concha
Part of the ethmoid bone, which forms the roof and upper side walls of the nasal cavity

Middle nasal concha
Like the superior nasal concha, this is also part of the ethmoid bone

Inferior nasal concha
A separate bone, attached to the inner surface of the maxilla; the conchae increase the surface area of the nasal cavity

Anterior nasal crest

Palatine bone
Joins to the maxilla and forms the back of the hard palate

Pterygoid process
Sticking down from the greater wing of the sphenoid bone, this process flanks the back of the nasal cavity and provides attachment for muscles of the palate and jaw

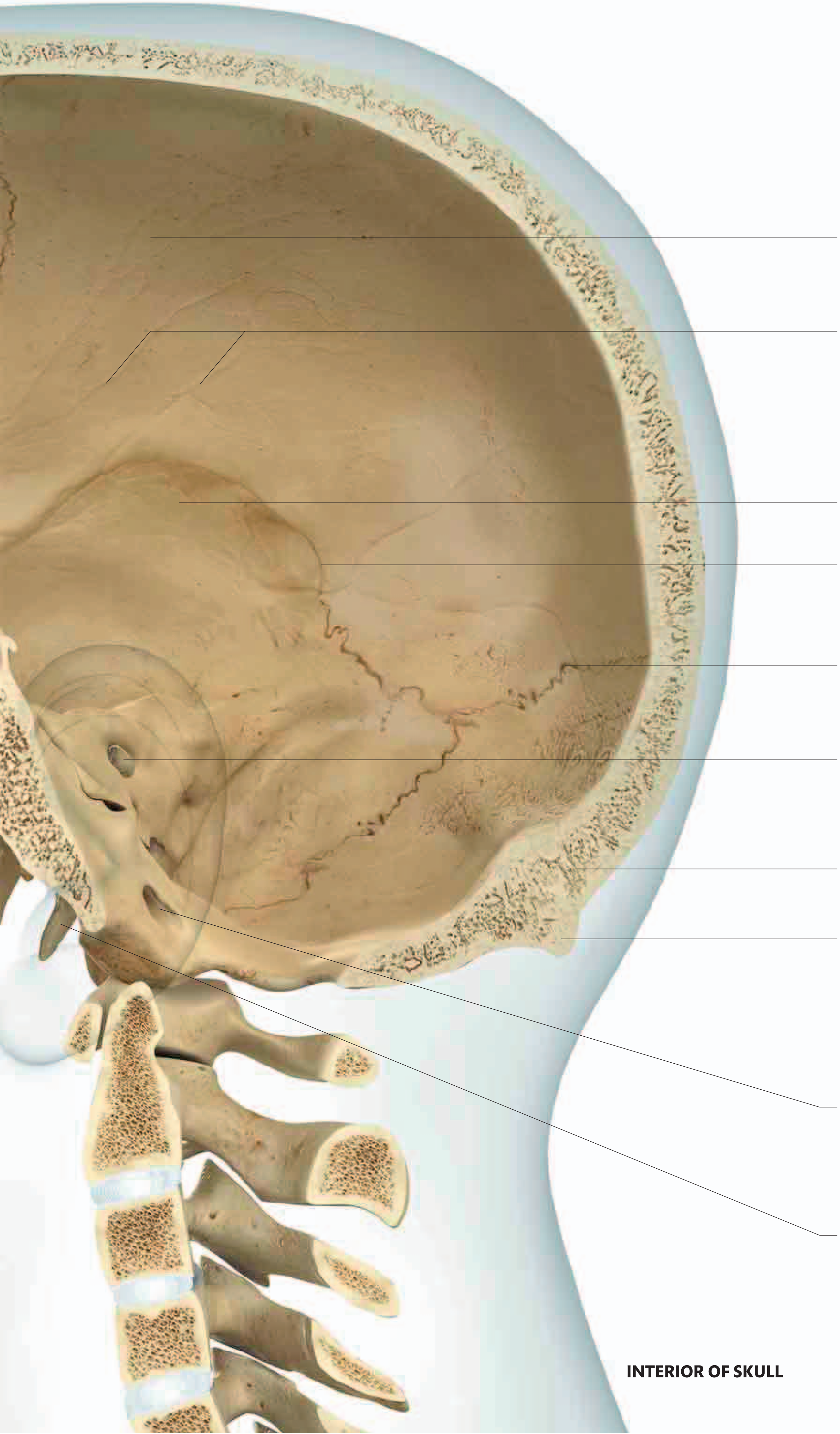
Parietal bone

Grooves for arteries
Meningeal arteries branch on the inside of the skull and leave grooves on the bones

Squamous part of the temporal bone

Squamosal suture

Lambdoid suture

Internal acoustic meatus
Hole in petrous part of the temporal bone that transmits both the facial and vestibulocochlear nerves

Occipital bone

External occipital protuberance
Projection from occipital bone that gives attachment to the nuchal ligament of the neck; much more pronounced in men than in women

Hypoglossal canal
Hole through occipital bone, in the cranial base, which transmits the hypoglossal nerve supplying the tongue muscles

Styloid process

INTERIOR OF SKULL

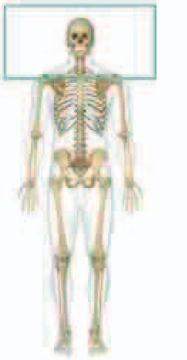

HEAD AND NECK

In this view of the skull, we can clearly see that it is not one single bone, and we can also see how the various cranial bones fit together to produce the shape we are more familiar with. The butterfly-shaped sphenoid bone is right in the middle of the action—it forms part of the skull base, the orbits, and the side walls of the skull, and it articulates with many of the other bones of the skull. The temporal bones also form part of the skull's base and side walls. The extremely dense petrous parts of the temporal bones contain and protect the delicate workings of the ear, including the tiny ossicles (malleus, incus, and stapes) that transmit vibrations from the eardrum to the inner ear.

FIBROUS JOINTS

In places, the connective tissue between developing bones solidifies to create fibrous joints. Linked by microscopic fibers of collagen, these fixed joints anchor the edges of adjacent bones, or bone and tooth, so that they are locked together. Such joints include the sutures of the skull, the teeth sockets (gomphoses), and the lower joint between the tibia and fibula.

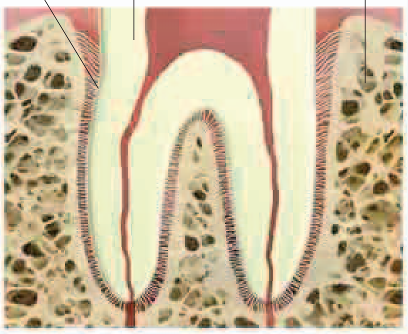

Periodontal ligament Dense connective tissue anchoring the tooth in the socket

Cement Covers the roots of the tooth

Alveolar bone Bone of the maxilla or mandible forming the tooth socket (alveolus)

TOOTH

Gomphosis
This name comes from the Greek word for bolted together. The fibrous tissue of the periodontal ligament connects the cement of the tooth to the bone of the socket.

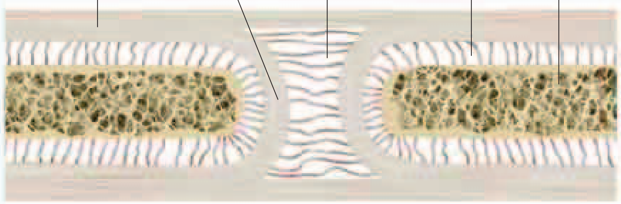

SKULL

Suture
These joints exist between flat bones of the skull. They are flexible in the skull of a newborn baby, and allow growth of the skull throughout childhood. The sutures in the adult skull are interlocking, practically immovable joints, and eventually fuse completely in later adulthood.

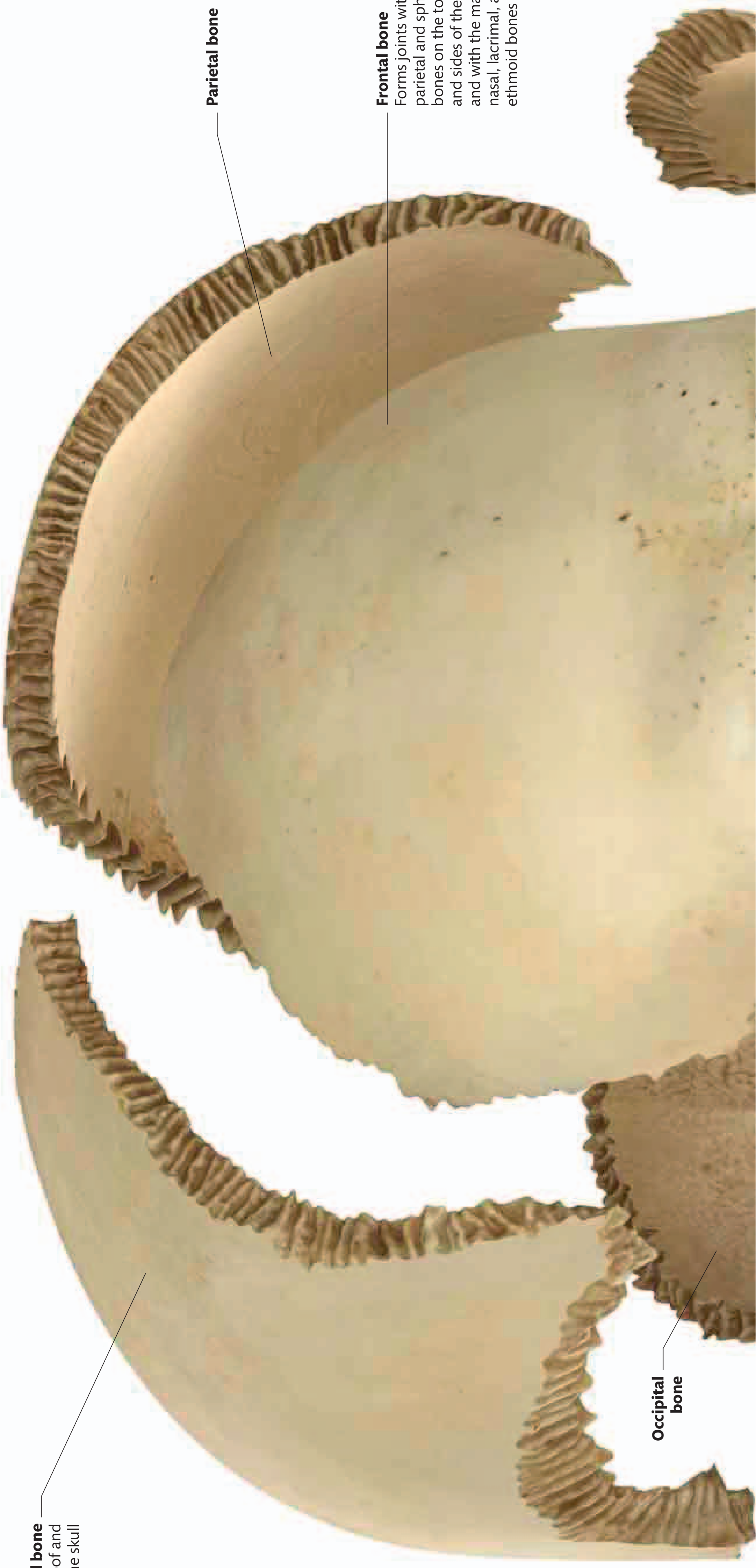

Parietal bone

Frontal bone Forms joints with the parietal and sphenoid bones on the top and sides of the skull, and with the maxilla, nasal, lacrimal, and ethmoid bones below

Occipital bone

Parietal bone Forms the roof and side of the skull

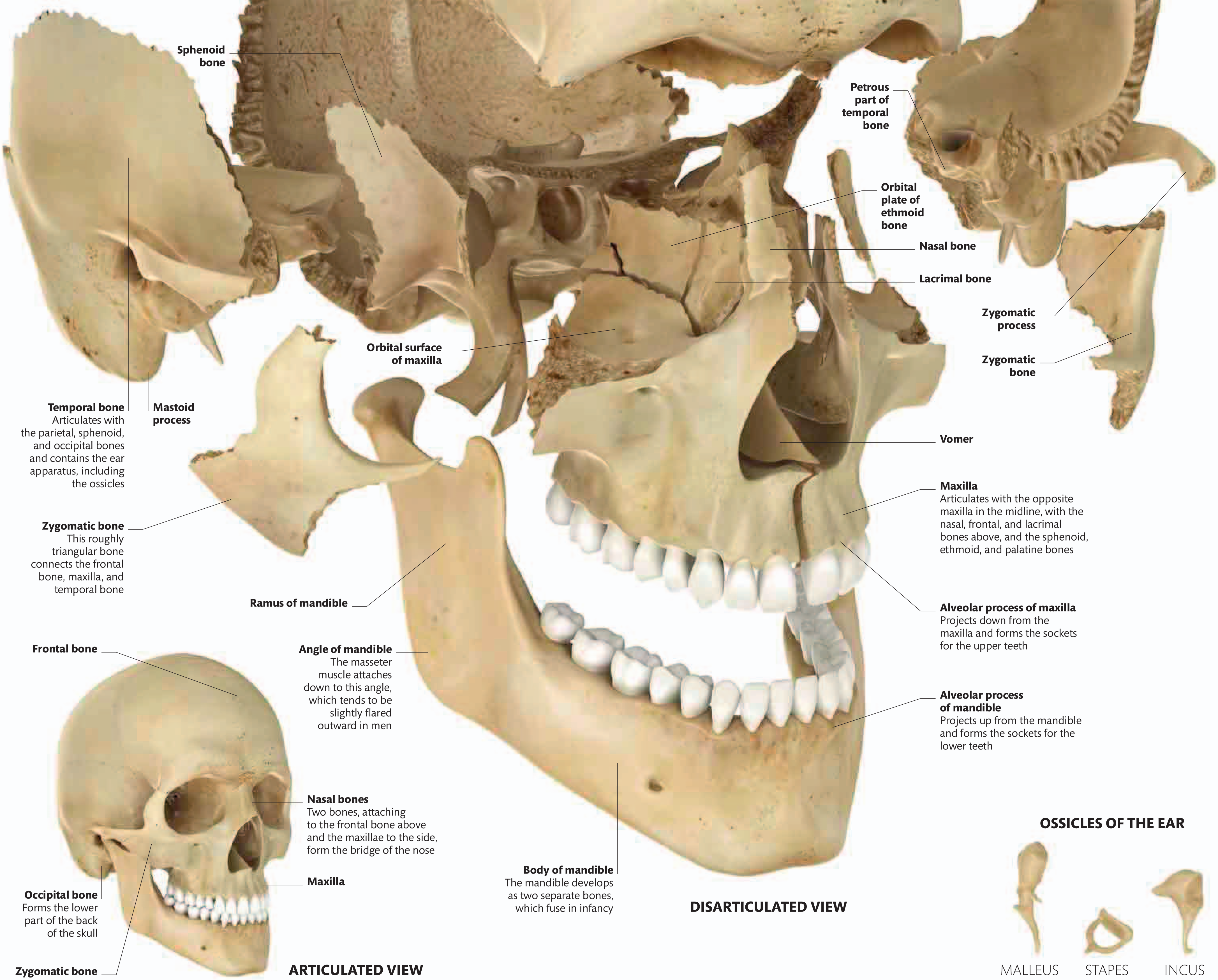
Sphenoid bone
Petrous part of temporal bone
Orbital plate of ethmoid bone
Nasal bone
Lacrimal bone
Zygomatic process
Zygomatic bone
Orbital surface of maxilla
Mastoid process
Vomer
Ramus of mandible
Frontal bone
Maxilla
Zygomatic bone
MALLEUS
STAPES
INCUS

Temporal bone Articulates with the parietal, sphenoid, and occipital bones and contains the ear apparatus, including the ossicles

Zygomatic bone This roughly triangular bone connects the frontal bone, maxilla, and temporal bone

Maxilla Articulates with the opposite maxilla in the midline, with the nasal, frontal, and lacrimal bones above, and the sphenoid, ethmoid, and palatine bones

Alveolar process of maxilla Projects down from the maxilla and forms the sockets for the upper teeth

Alveolar process of mandible Projects up from the mandible and forms the sockets for the lower teeth

Angle of mandible The masseter muscle attaches down to this angle, which tends to be slightly flared outward in men

Body of mandible The mandible develops as two separate bones, which fuse in infancy

DISARTICULATED VIEW

Nasal bones Two bones, attaching to the frontal bone above and the maxillae to the side, form the bridge of the nose

Occipital bone Forms the lower part of the back of the skull

ARTICULATED VIEW

OSSICLES OF THE EAR

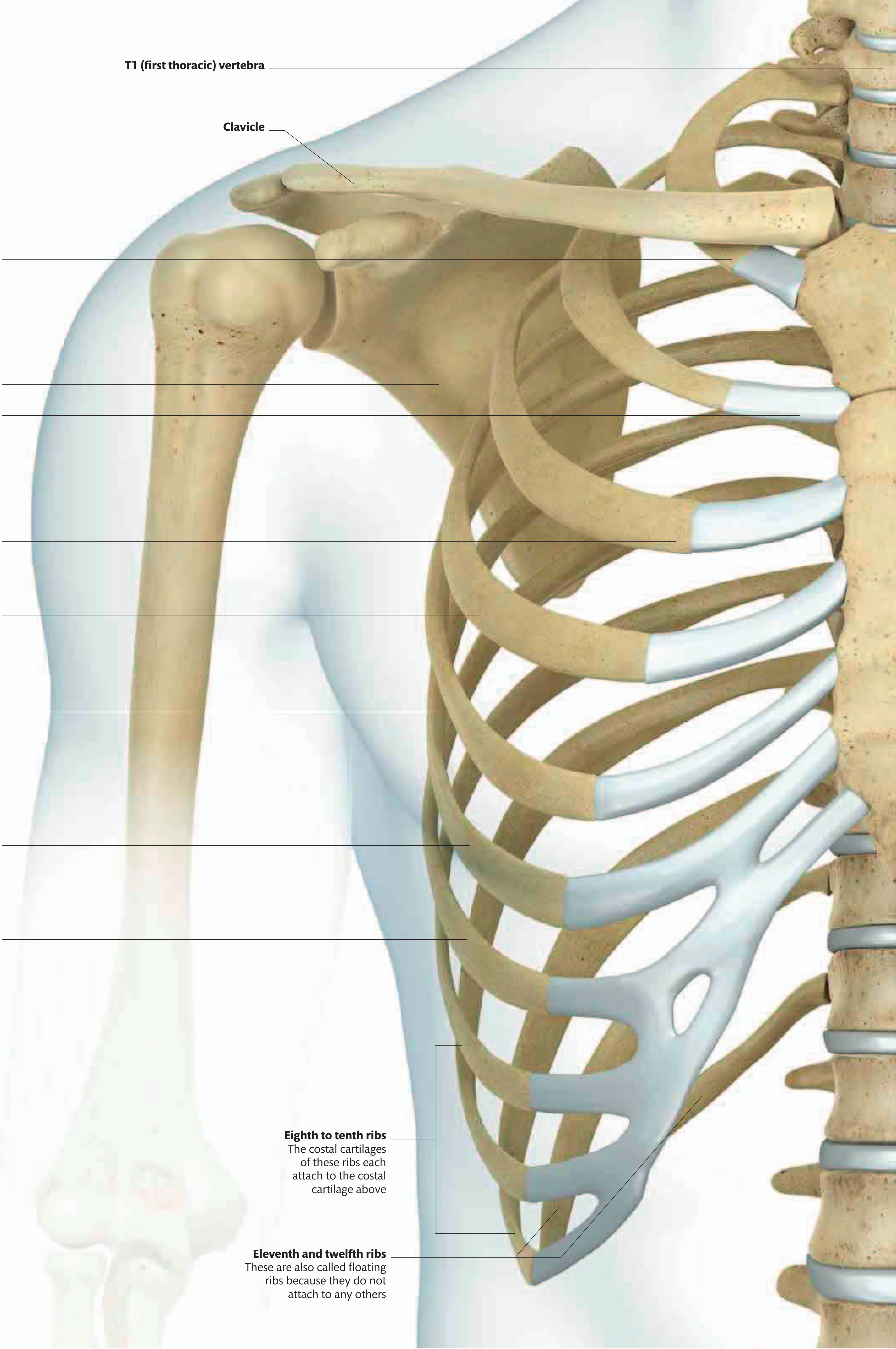
T1 (first thoracic) vertebra
Clavicle
First rib
Scapula
Second costal cartilage
Third rib
Fourth rib
Fifth rib
Sixth rib
Seventh rib
Eighth to tenth ribs
Eleventh and twelfth ribs

First rib
Smaller and more curved than all the other ribs; the thoracic inlet is formed by the first rib on each side, together with the manubrium sterni and the body of the T1 vertebra

Second costal cartilage
The upper seven ribs are true ribs, and all attach directly to the sternum via costal cartilages

Eighth to tenth ribs
The costal cartilages of these ribs each attach to the costal cartilage above

Eleventh and twelfth ribs
These are also called floating ribs because they do not attach to any others

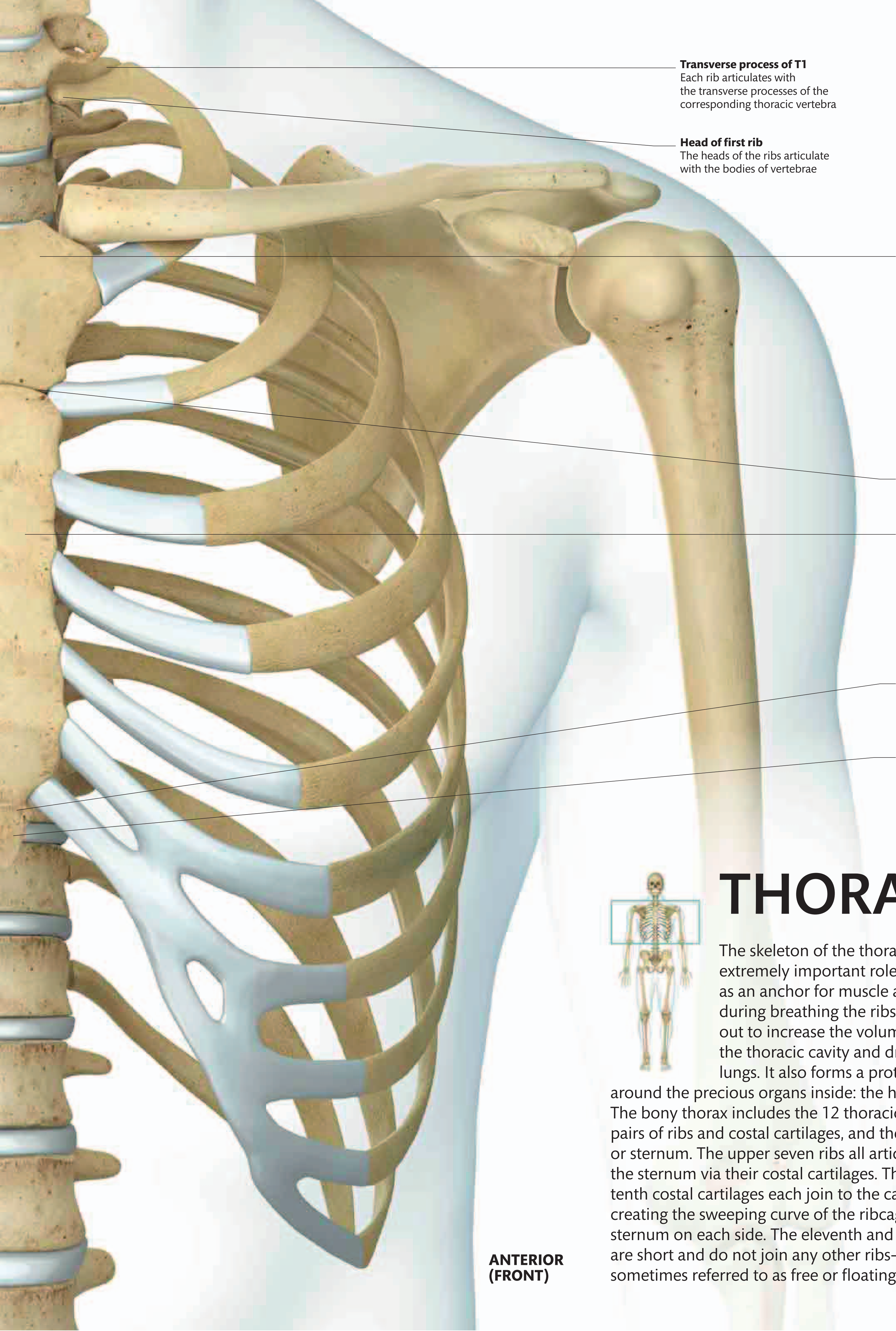

Transverse process of T1
Each rib articulates with the transverse processes of the corresponding thoracic vertebra

Head of first rib
The heads of the ribs articulate with the bodies of vertebrae

Manubrium sterni
The sternum is shaped like a dagger or short sword; manubrium means handle or hilt in Latin

Manubriosternal joint

Body of sternum
Sternum comes from the Greek for breastbone

Xiphisternal joint

Xiphoid process
The tip of the sternum takes its name from the Greek word for sword

ANTERIOR (FRONT)

THORAX

The skeleton of the thorax plays several extremely important roles. It not only acts as an anchor for muscle attachment, but during breathing the ribs move up and out to increase the volume inside the thoracic cavity and draw air into the lungs. It also forms a protective cage around the precious organs inside: the heart and lungs. The bony thorax includes the 12 thoracic vertebrae, 12 pairs of ribs and costal cartilages, and the breastbone, or sternum. The upper seven ribs all articulate with the sternum via their costal cartilages. The eighth to the tenth costal cartilages each join to the cartilage above, creating the sweeping curve of the ribcage below the sternum on each side. The eleventh and twelfth ribs are short and do not join any other ribs—they are sometimes referred to as free or floating ribs.

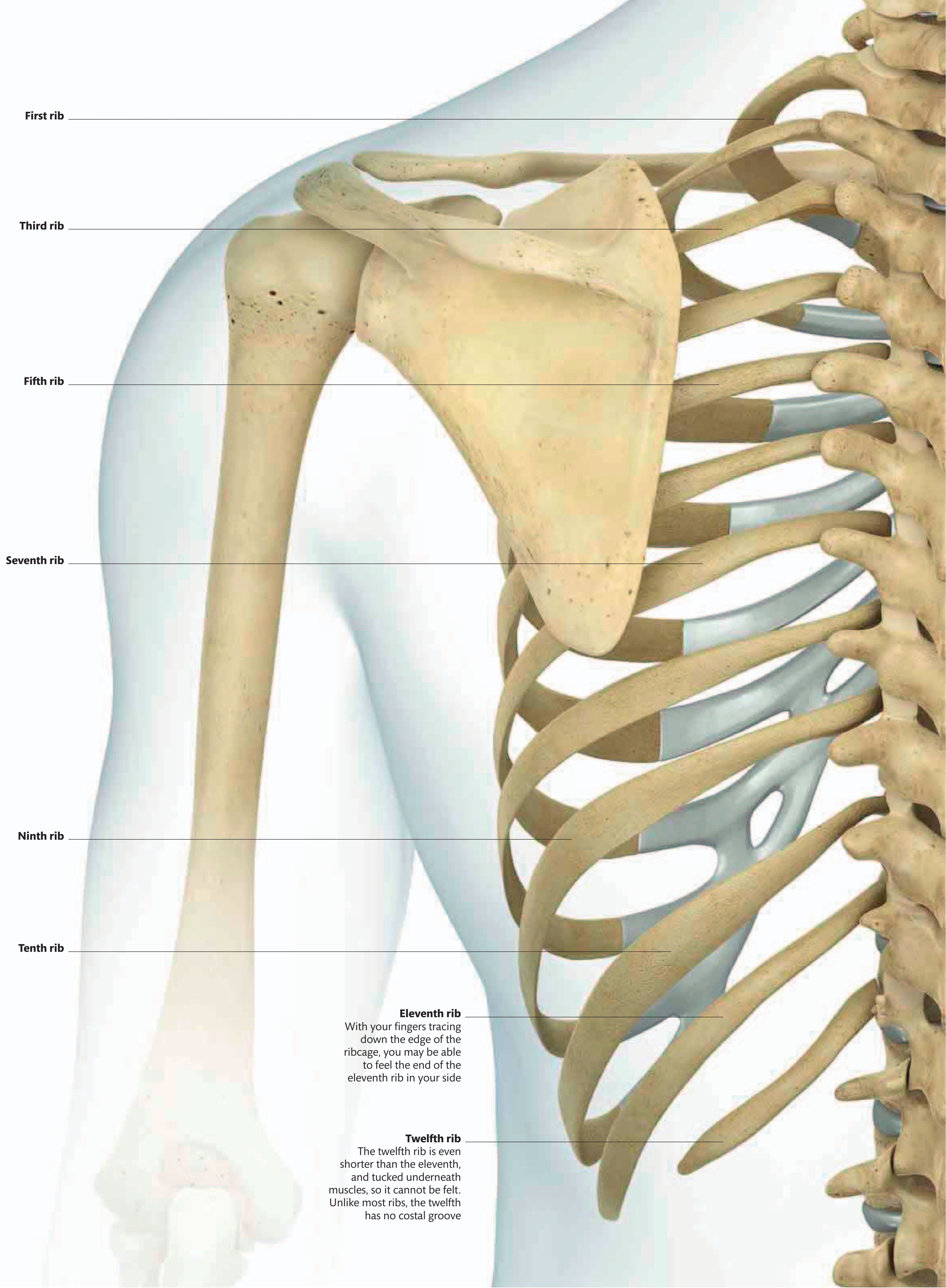
First rib
Third rib
Fifth rib
Seventh rib
Ninth rib
Tenth rib

Eleventh rib
With your fingers tracing down the edge of the ribcage, you may be able to feel the end of the eleventh rib in your side

Twelfth rib
The twelfth rib is even shorter than the eleventh, and tucked underneath muscles, so it cannot be felt. Unlike most ribs, the twelfth has no costal groove

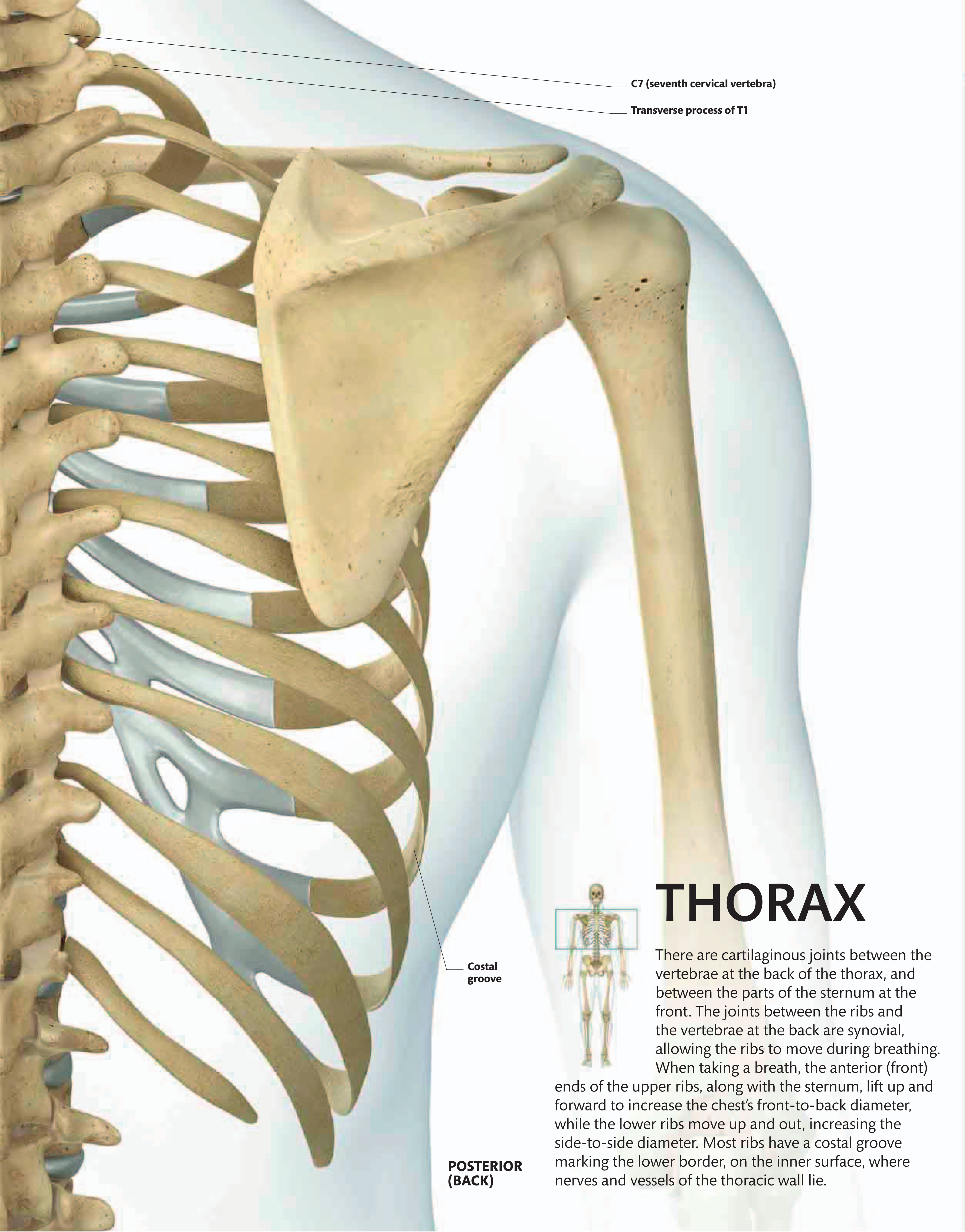
C7 (seventh cervical vertebra)
Transverse process of T1
Costal groove

POSTERIOR (BACK)

THORAX

There are cartilaginous joints between the vertebrae at the back of the thorax, and between the parts of the sternum at the front. The joints between the ribs and the vertebrae at the back are synovial, allowing the ribs to move during breathing. When taking a breath, the anterior (front) ends of the upper ribs, along with the sternum, lift up and forward to increase the chest's front-to-back diameter, while the lower ribs move up and out, increasing the side-to-side diameter. Most ribs have a costal groove marking the lower border, on the inner surface, where nerves and vessels of the thoracic wall lie.

SPINE

The spine, or vertebral column, occupies a central position in the skeleton, and plays several extremely important roles: it supports the trunk, encloses and protects the spinal cord, provides sites for muscle attachment, and contains blood-forming bone marrow. The entire vertebral column is about 28in (70cm) long in men, and 24in (60cm) long in women. About a quarter of this length is made up by the cartilaginous intervertebral discs between the vertebrae. The number of vertebrae varies from 32 to 35, mostly due to variation in the number of small vertebrae that make up the coccyx. Although there is a general pattern for a vertebra—most possess a body, a neural arch, and spinous and transverse processes—there are recognizable features that mark out the vertebrae of each section of the spine.

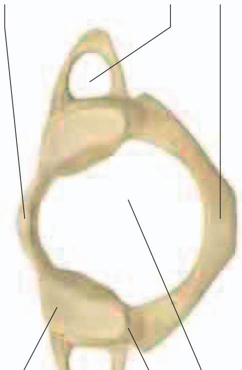

ATLAS (C1)

Anterior arch The atlas has no body, but it has an anterior arch that forms a joint with the dens of the axis

Transverse foramen

Posterior arch

Superior articular facet Articulates with the condyle of the occipital bone, on the base of the skull

Lateral mass

Vertebral foramen

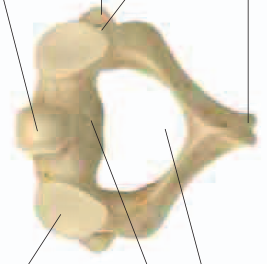

AXIS (C2)

Dens (odontoid peg) This projection sticks up to articulate with the atlas

Transverse process

Transverse foramen

Spinous process

Superior articular facet

Body

Vertebral foramen

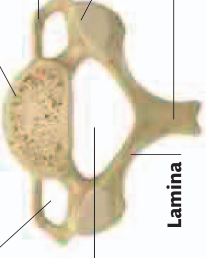

CERVICAL

Body Made of cancellous bone containing blood-making bone marrow

Transverse process For neck muscle attachment

Superior articular facet

Spinous process Tends to be small and forked; for the attachment of back muscles

Lamina

Transverse foramen The vertebral artery passes through here

Vertebral foramen Large compared with the size of the body; contains the spinal cord

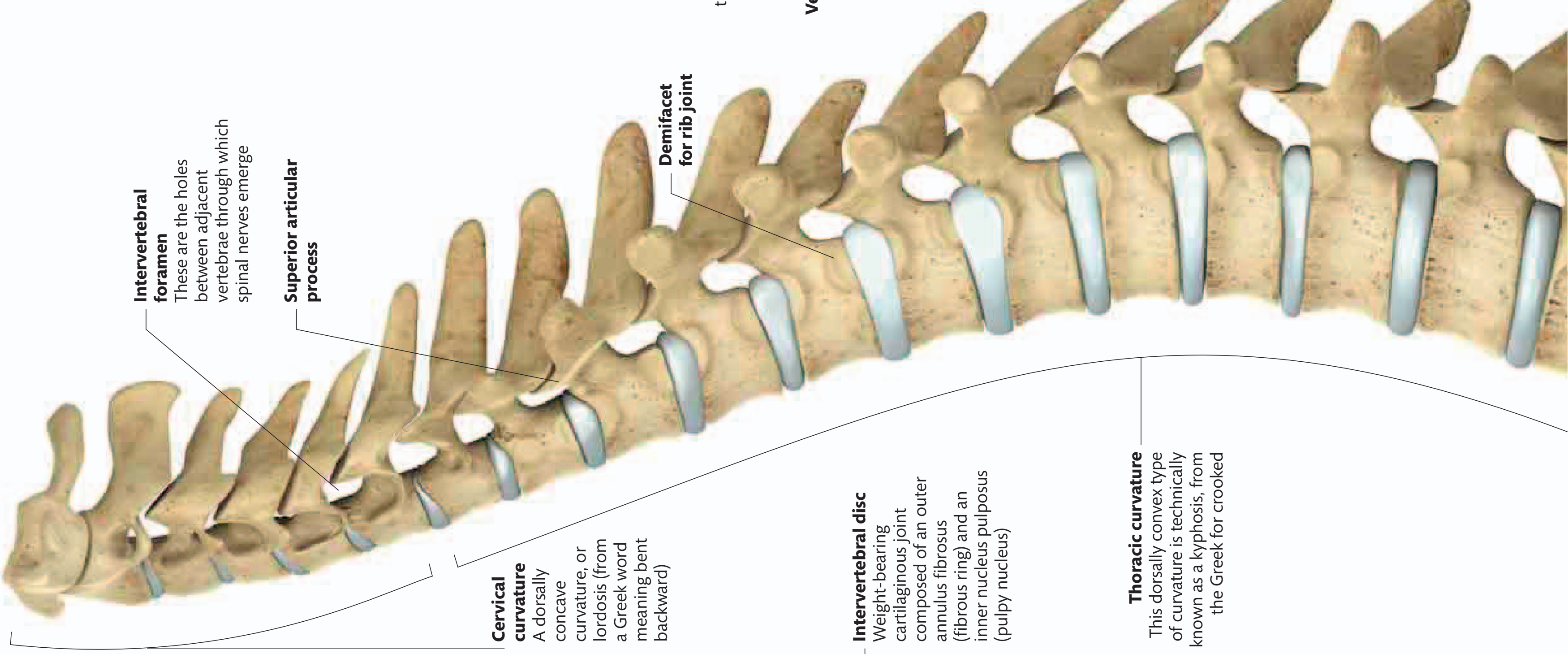

Intervertebral foramen These are the holes between adjacent vertebrae through which spinal nerves emerge

Superior articular process

Demifacet for rib joint

Cervical curvature A dorsally concave curvature, or lordosis (from a Greek word meaning bent backward)

Intervertebral disc Weight-bearing cartilaginous joint composed of an outer annulus fibrosus (fibrous ring) and an inner nucleus pulposus (pulpy nucleus)

Thoracic curvature This dorsally convex type of curvature is technically known as a kyphosis, from the Greek for crooked

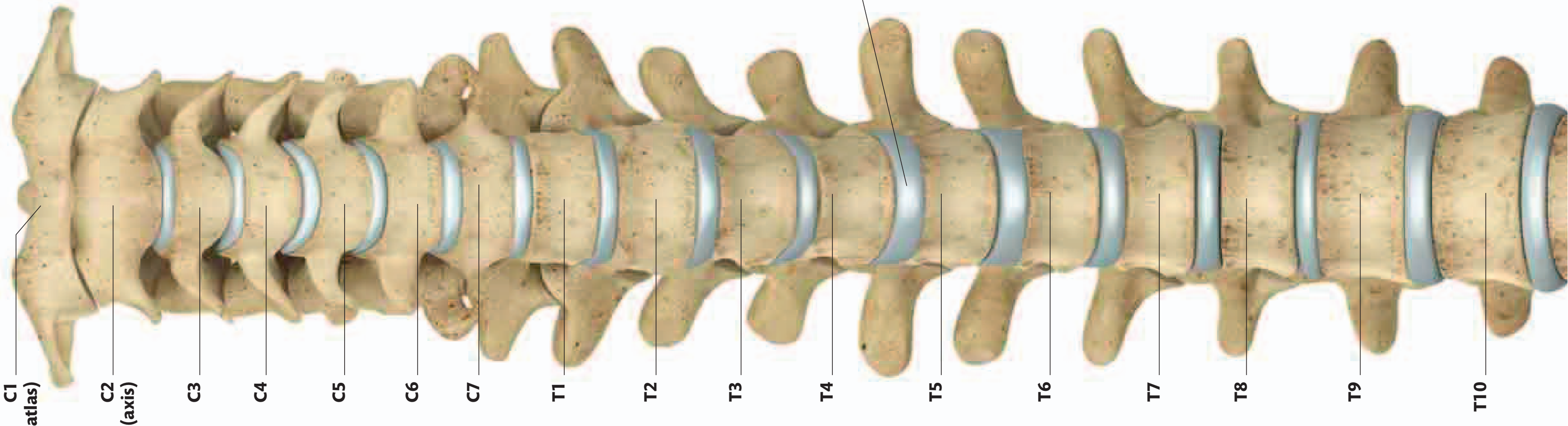

Cervical spine
(Seven vertebrae make up the spine in the neck)

Thoracic spine
(Twelve vertebrae, providing attachment for twelve pairs of ribs)

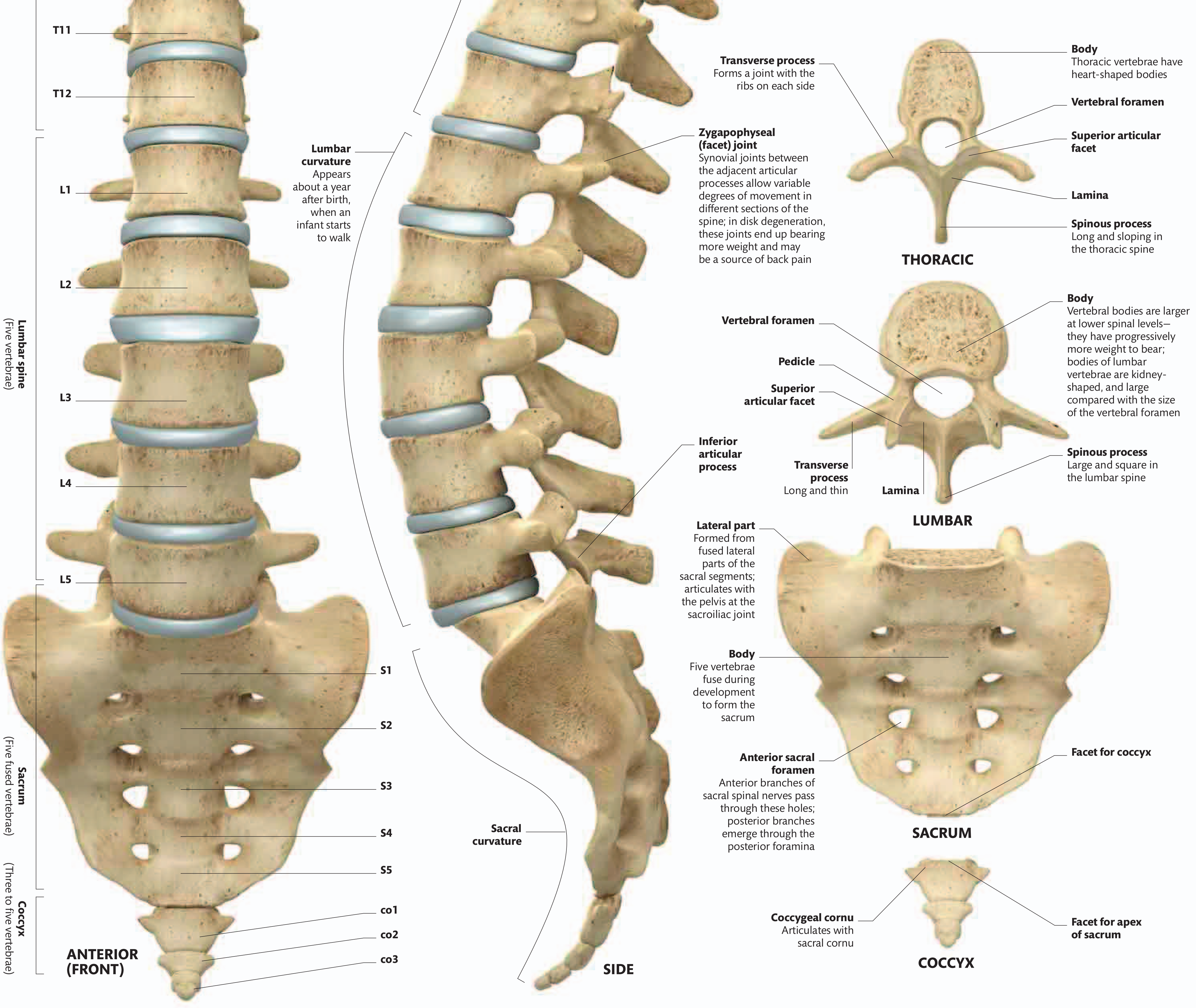
T11
T12
L1
L2
L3
L4
L5
S1
S2
S3
S4
S5
co1
co2
co3
Lumbar spine (Five vertebrae)
Sacrum (Five fused vertebrae)
Coccyx (Three to five vertebrae)
ANTERIOR (FRONT)
Lumbar curvature Appears about a year after birth, when an infant starts to walk
Sacral curvature
SIDE
Zygapophyseal (facet) joint Synovial joints between the adjacent articular processes allow variable degrees of movement in different sections of the spine; in disk degeneration, these joints end up bearing more weight and may be a source of back pain
Inferior articular process
Transverse process Forms a joint with the ribs on each side
Body Thoracic vertebrae have heart-shaped bodies
Vertebral foramen
Superior articular facet
Lamina
Spinous process Long and sloping in the thoracic spine
THORACIC
Vertebral foramen
Pedicle
Superior articular facet
Transverse process Long and thin
Lamina
Body Vertebral bodies are larger at lower spinal levels—they have progressively more weight to bear; bodies of lumbar vertebrae are kidney-shaped, and large compared with the size of the vertebral foramen
Spinous process Large and square in the lumbar spine
LUMBAR
Lateral part Formed from fused lateral parts of the sacral segments; articulates with the pelvis at the sacroiliac joint
Body Five vertebrae fuse during development to form the sacrum
Anterior sacral foramen Anterior branches of sacral spinal nerves pass through these holes; posterior branches emerge through the posterior foramina
Facet for coccyx
SACRUM
Coccygeal cornu Articulates with sacral cornu
Facet for apex of sacrum
COCCYX

ABDOMEN AND PELVIS

The bony boundaries of the abdomen include the five lumbar vertebrae at the back, the lower margin of the ribs above, and the pubic bones and iliac crest of the pelvic bones below. The abdominal cavity itself extends up under the ribcage, as high as the gap between the fifth and sixth ribs, due to the domed shape of the diaphragm. This means that some abdominal organs, such as the liver, stomach, and spleen are, in fact, largely tucked up under the ribs. The pelvis is a basin shape, and is enclosed by the two pelvic (or innominate) bones, at the front and sides, and by the sacrum at the back. Each pelvic bone is made of three fused bones: the ilium at the rear, the ischium at the lower front, and the pubis above it.

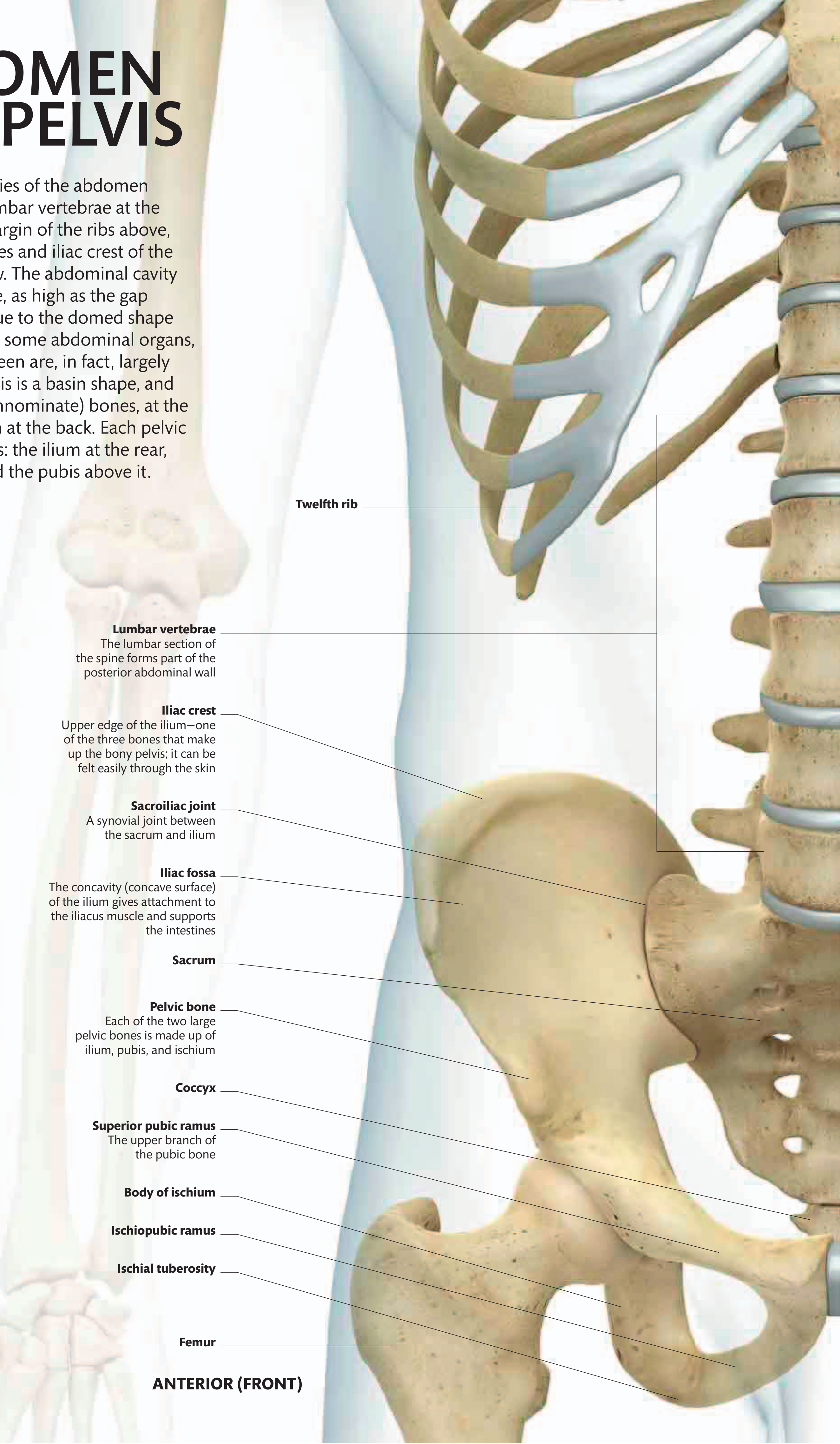

Lumbar vertebrae
The lumbar section of the spine forms part of the posterior abdominal wall

Iliac crest
Upper edge of the ilium—one of the three bones that make up the bony pelvis; it can be felt easily through the skin

Sacroiliac joint
A synovial joint between the sacrum and ilium

Iliac fossa
The concavity (concave surface) of the ilium gives attachment to the iliacus muscle and supports the intestines

Pelvic bone
Each of the two large pelvic bones is made up of ilium, pubis, and ischium

Superior pubic ramus
The upper branch of the pubic bone

ANTERIOR (FRONT)

CARTILAGINOUS JOINTS

Semi-movable cartilaginous joints are formed by bones separated by a disc of resilient and compressible fibrocartilage, which allows limited movement. Cartilaginous joints include the junctions between ribs and costal cartilages, joints between the components of the sternum, and the pubic symphysis. The intervertebral discs are also specialized cartilaginous joints.

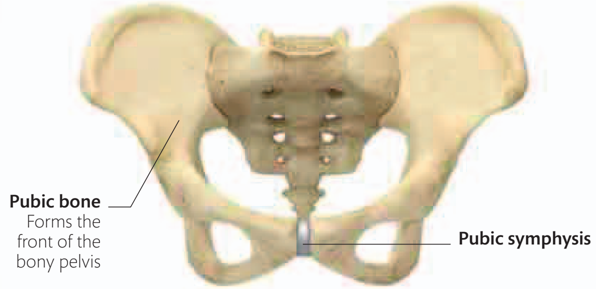
Pubic bone Forms the front of the bony pelvis
Pubic symphysis

PELVIS

Pubic symphysis
At the front of the bony pelvis, the two pubic bones meet each other. The articular surface of each is covered with hyaline cartilage, with a pad of fibrocartilage joining them in the middle.

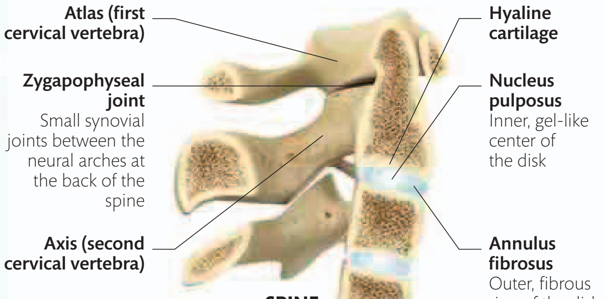
Atlas (first cervical vertebra)
Zygapophyseal joint Small synovial joints between the neural arches at the back of the spine
Axis (second cervical vertebra)
Hyaline cartilage
Nucleus pulposus Inner, gel-like center of the disk
Annulus fibrosus Outer, fibrous ring of the disk

SPINE

Intervertebral disc
Each fibrocartilage pad or disk between vertebrae is organized into an outer annulus fibrosus and an inner nucleus pulposus.

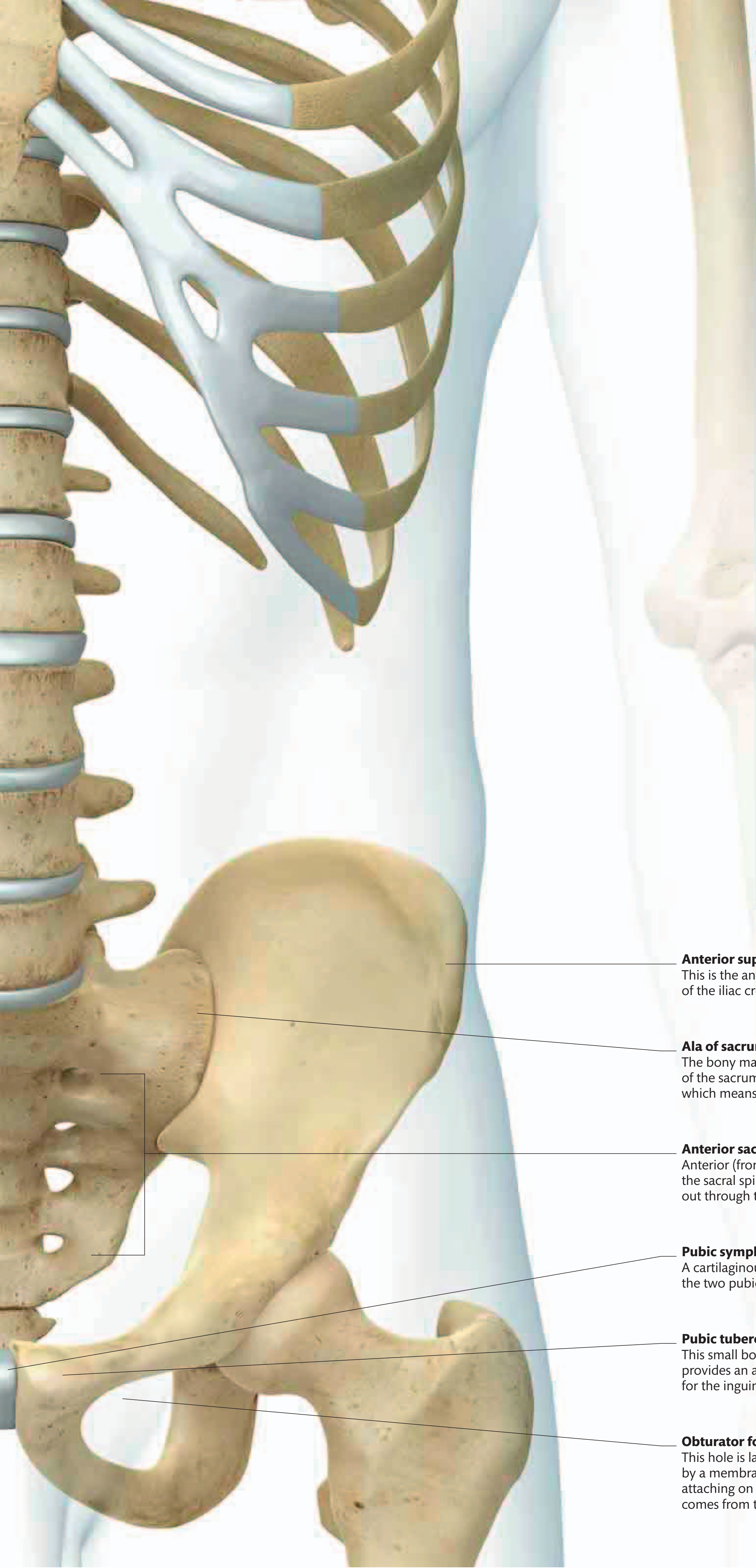

Anterior superior iliac spine
This is the anterior (front) end of the iliac crest

Ala of sacrum
The bony masses to the sides of the sacrum are called the alae, which means wings in Latin

Anterior sacral foramina
Anterior (frontal) branches of the sacral spinal nerves pass out through these holes

Pubic symphysis
A cartilaginous joint between the two pubic bones

Pubic tubercle
This small bony projection provides an attachment point for the inguinal ligament

Obturator foramen
This hole is largely closed over by a membrane, with muscles attaching on either side; its name comes from the Latin for stopped up

ABDOMEN AND PELVIS

The orientation of the facet joints (the joints between the vertebrae) of the lumbar spine means that rotation of the vertebrae is limited, but flexion and extension can occur freely. There is, however, rotation at the lumbosacral joint, which allows the pelvis to swing during walking. The sacroiliac joints are unusual in that they are synovial joints (which are usually very movable), yet they are particularly limited in their movement. This is because strong sacroiliac ligaments around the joints bind the ilium (part of the pelvic bone) tightly to the sacrum on each side. Lower down, the sacrospinous and sacrotuberous ligaments, stretching from the sacrum and coccyx to the ilium, provide additional support and stability.

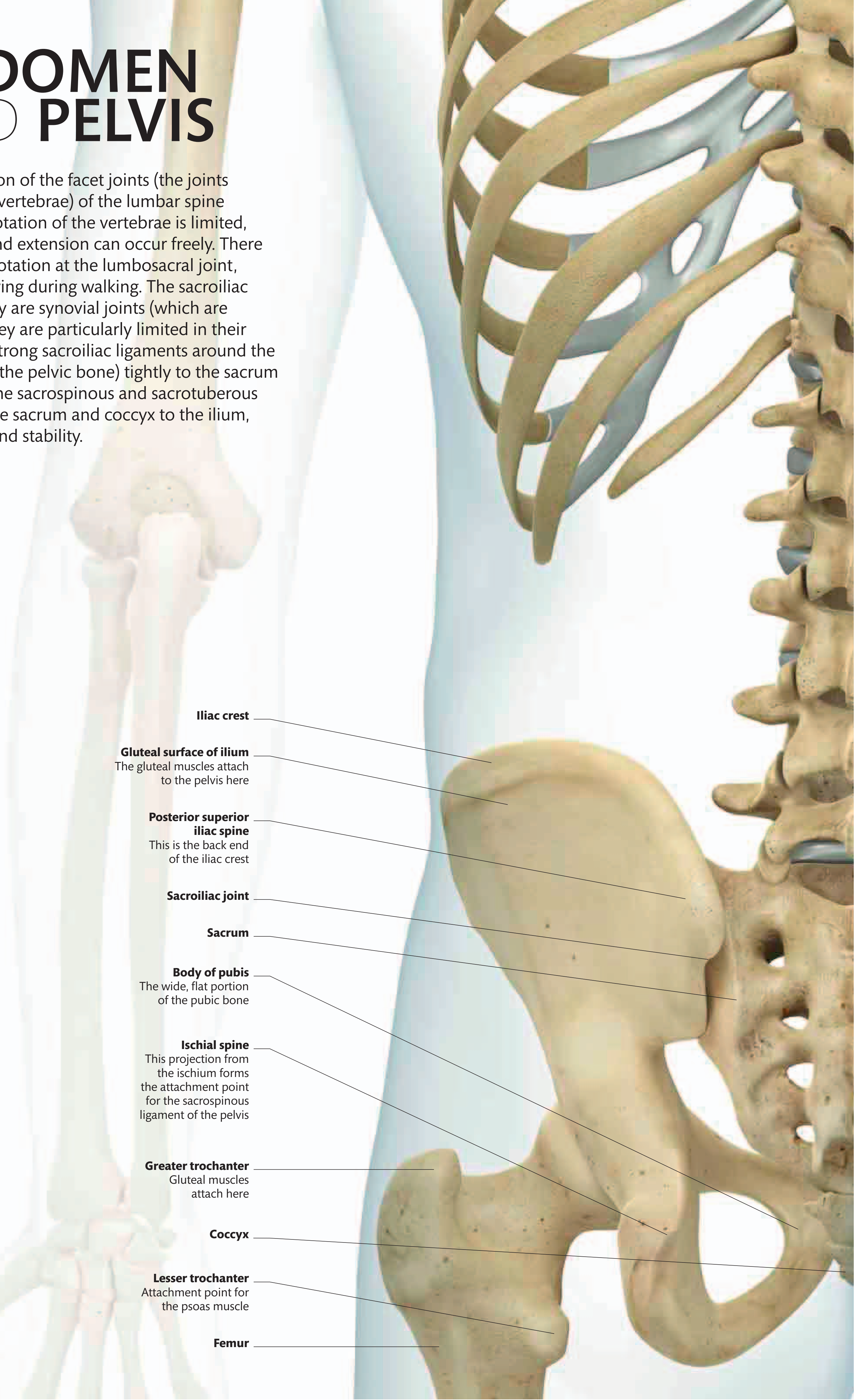
Iliac crest
Gluteal surface of ilium
The gluteal muscles attach to the pelvis here
Posterior superior iliac spine
This is the back end of the iliac crest
Sacroiliac joint
Sacrum
Body of pubis
The wide, flat portion of the pubic bone
Ischial spine
This projection from the ischium forms the attachment point for the sacrospinous ligament of the pelvis
Greater trochanter
Gluteal muscles attach here
Coccyx
Lesser trochanter
Attachment point for the psoas muscle
Femur

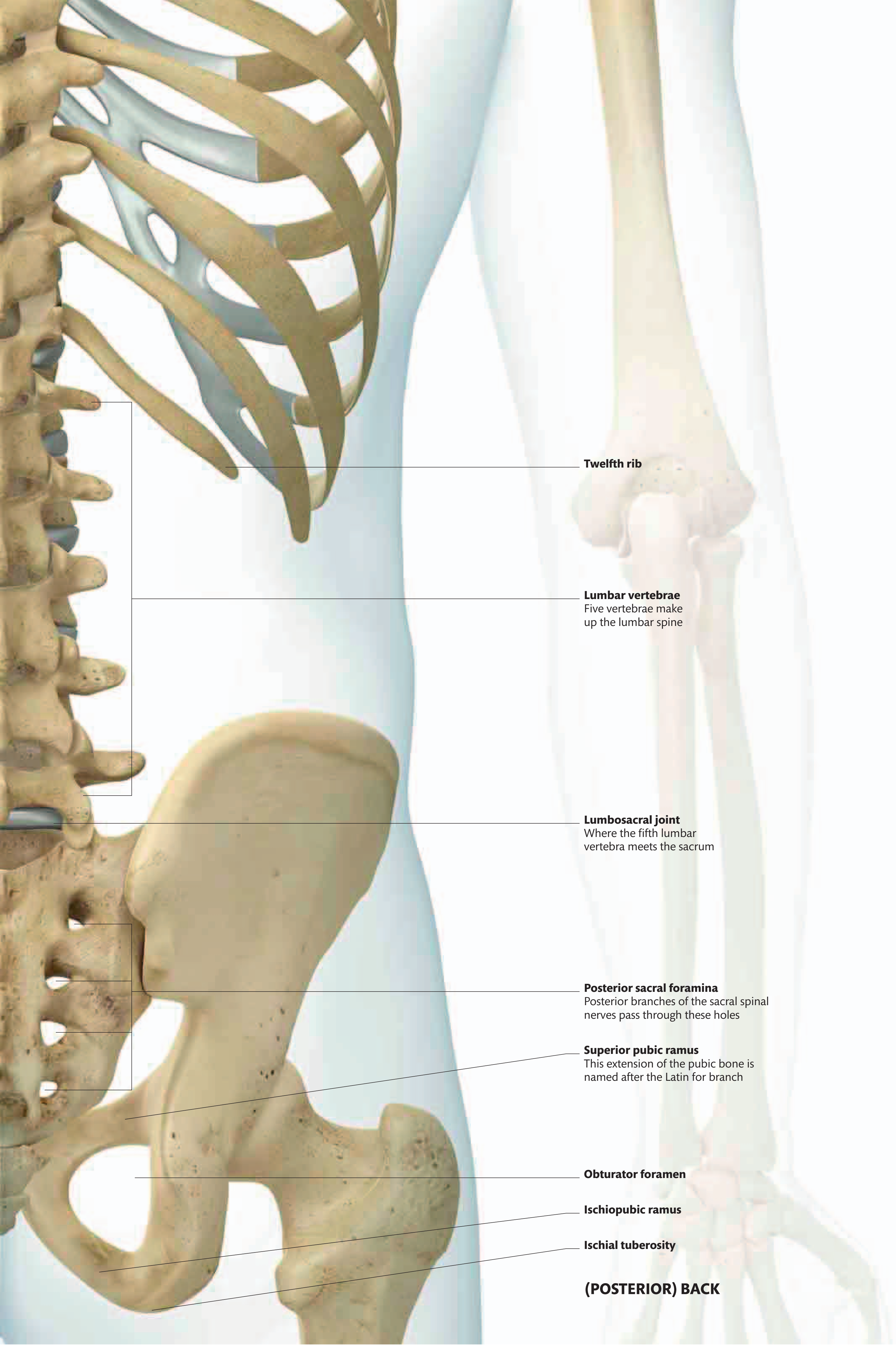

Twelfth rib

Lumbar vertebrae
Five vertebrae make up the lumbar spine

Lumbosacral joint
Where the fifth lumbar vertebra meets the sacrum

Posterior sacral foramina
Posterior branches of the sacral spinal nerves pass through these holes

Superior pubic ramus
This extension of the pubic bone is named after the Latin for branch

Obturator foramen

Ischiopubic ramus

Ischial tuberosity

(POSTERIOR) BACK

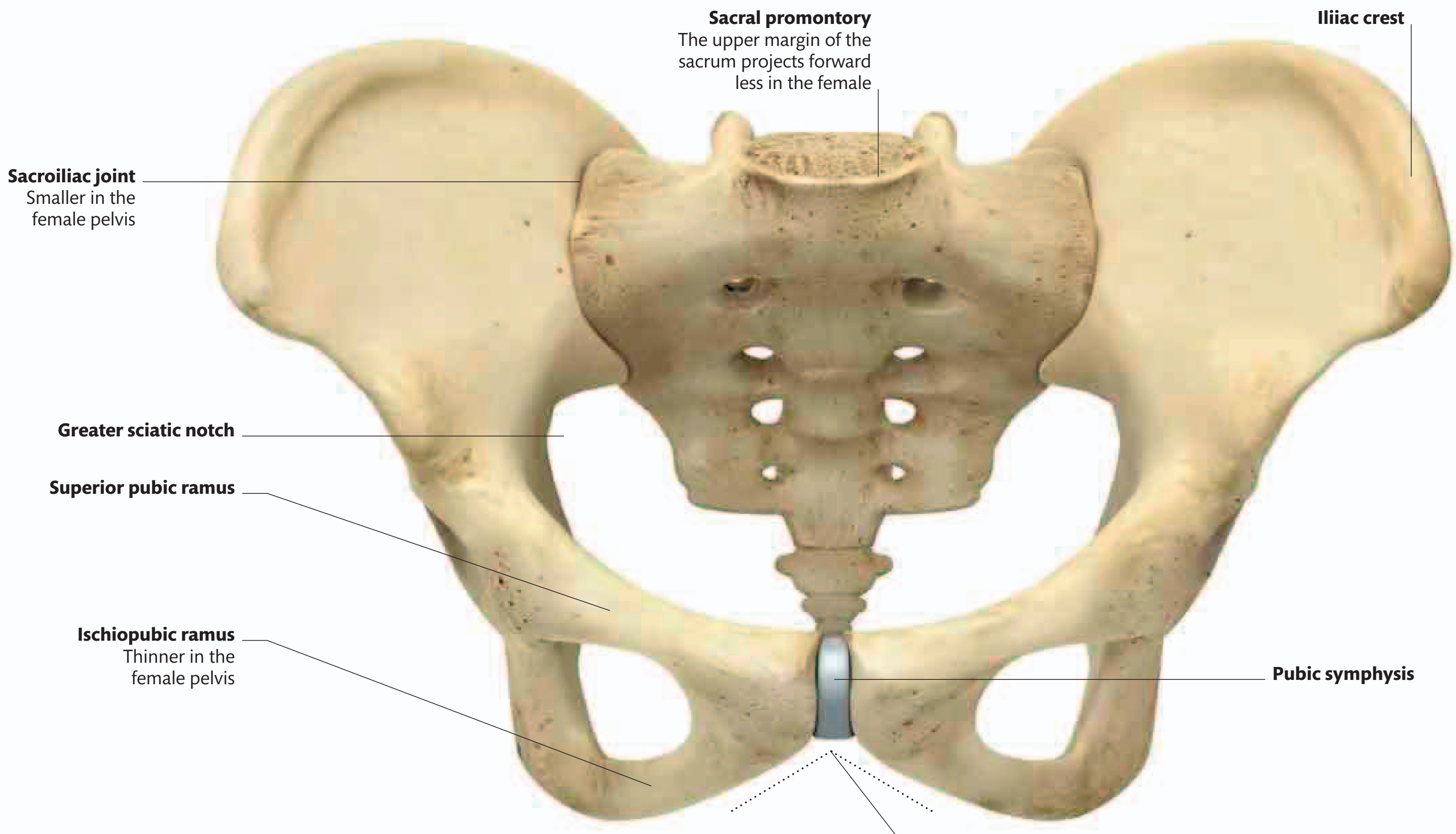
Sacral promontory
The upper margin of the sacrum projects forward less in the female
Iliiac crest
Sacroiliac joint
Smaller in the female pelvis
Greater sciatic notch
Superior pubic ramus
Ischiopubic ramus
Thinner in the female pelvis
Pubic symphysis
Subpubic angle
Much wider in the female pelvis

FEMALE PELVIS ANTERIOR (FRONT)

PELVIS

The bony pelvis is the part of the skeleton that is most different between the sexes, because the pelvis in the female has to accommodate the birth canal, unlike the male pelvis. Comparing the pelvic bones of a man and a woman, there are obvious differences between the two. The shape of the ring formed by the sacrum and the two pelvic bones—the pelvic brim—tends to be a wide oval in the woman and much narrower and heart-shaped in a man. The subpubic angle, underneath the joint between the two pubic bones, is much narrower in a man than it is in with a woman. As with the rest of the skeleton, the pelvic bone also tends to be more chunky or robust in a man, with more obvious ridges where muscles attach.

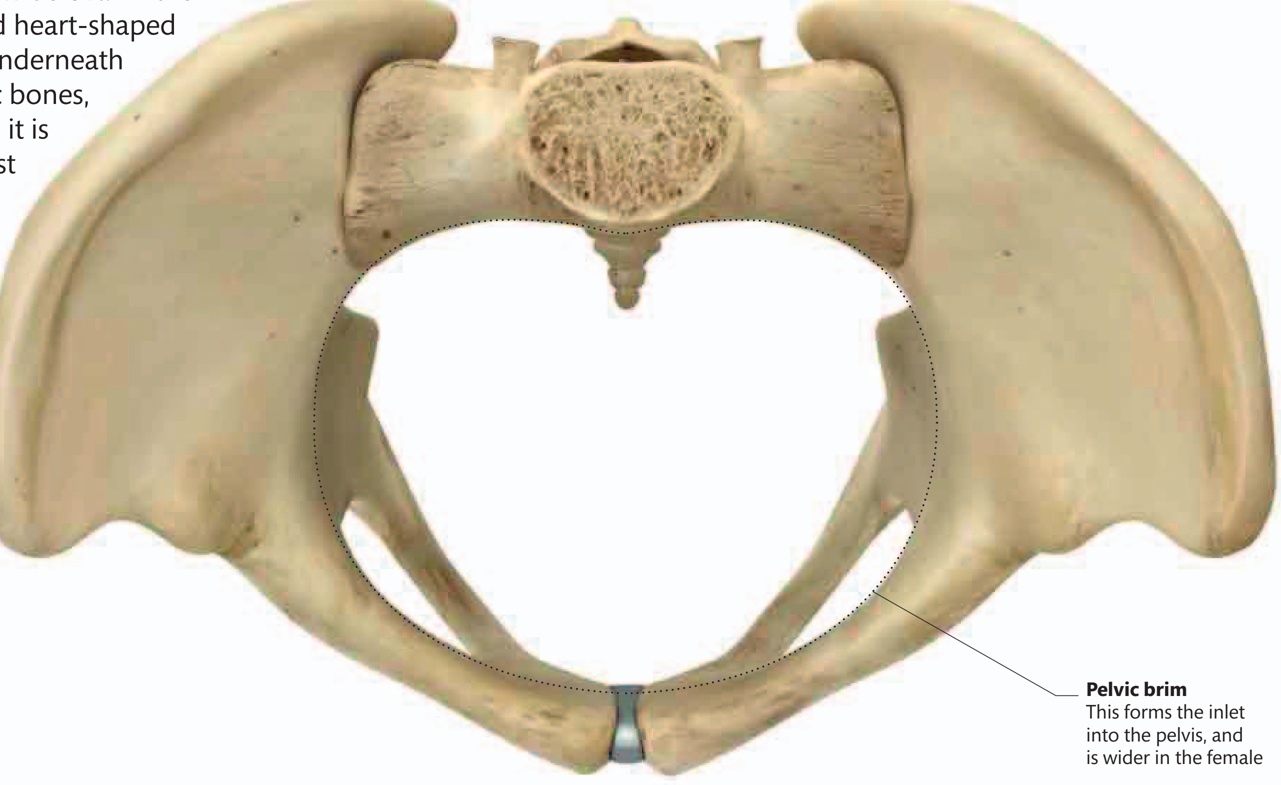
Pelvic brim
This forms the inlet into the pelvis, and is wider in the female

FEMALE PELVIS VIEWED FROM ABOVE

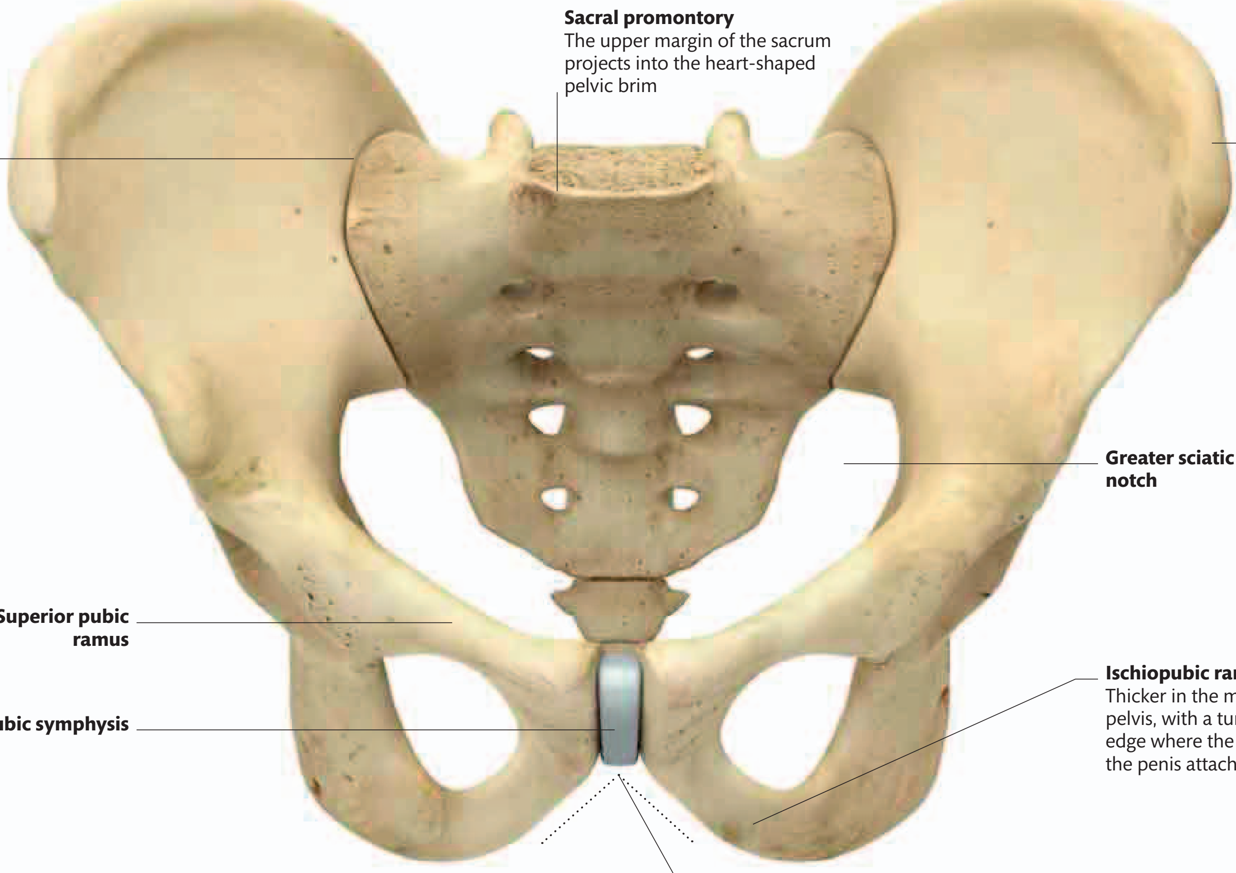

Sacral promontory
The upper margin of the sacrum projects into the heart-shaped pelvic brim

Sacroiliac joint
Male joints tend to be larger than those of the female, and this one is no exception

Iliac crest
Gives attachment to the muscles of the abdominal wall and is more robust or chunky in the male

Greater sciatic notch

Superior pubic ramus

Ischiopubic ramus
Thicker in the male pelvis, with a turned-out edge where the crus of the penis attaches

Pubic symphysis

Subpubic angle

MALE PELVIS ANTERIOR (FRONT)

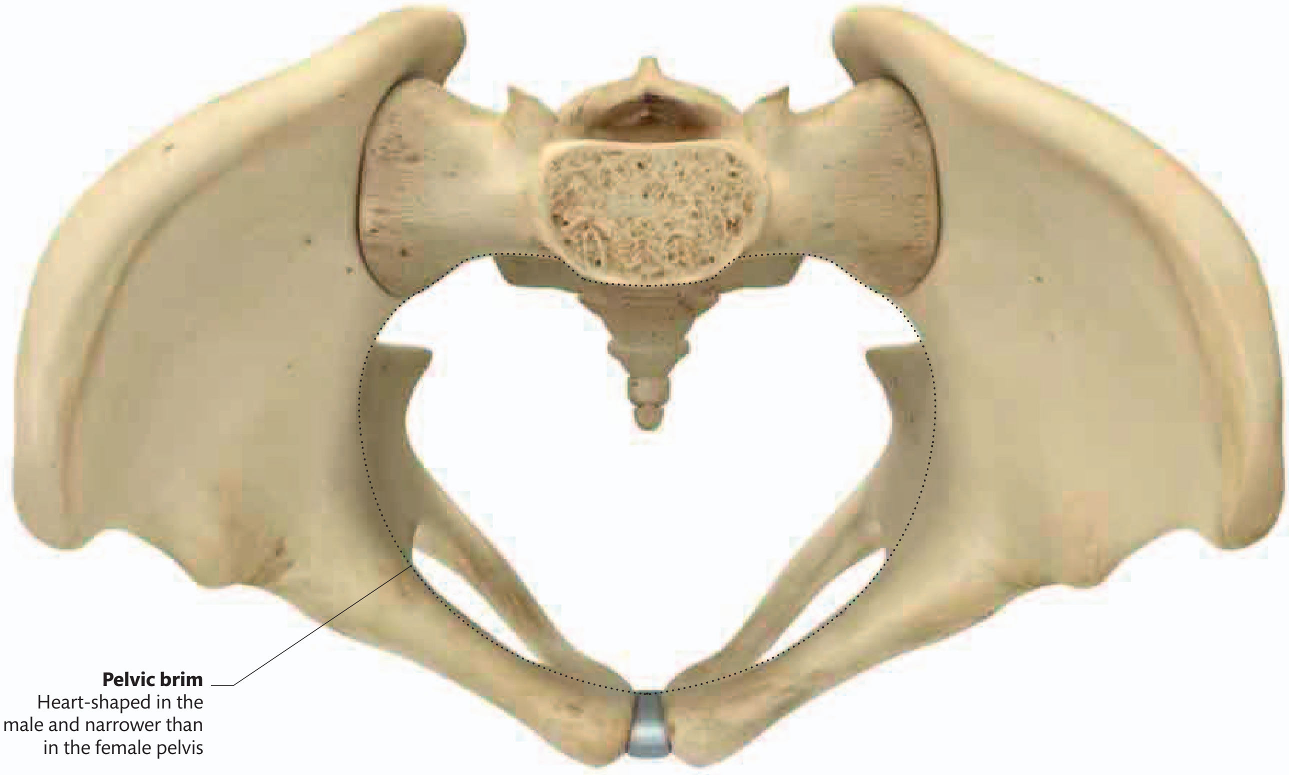

Pelvic brim
Heart-shaped in the male and narrower than in the female pelvis

MALE PELVIS VIEWED FROM ABOVE

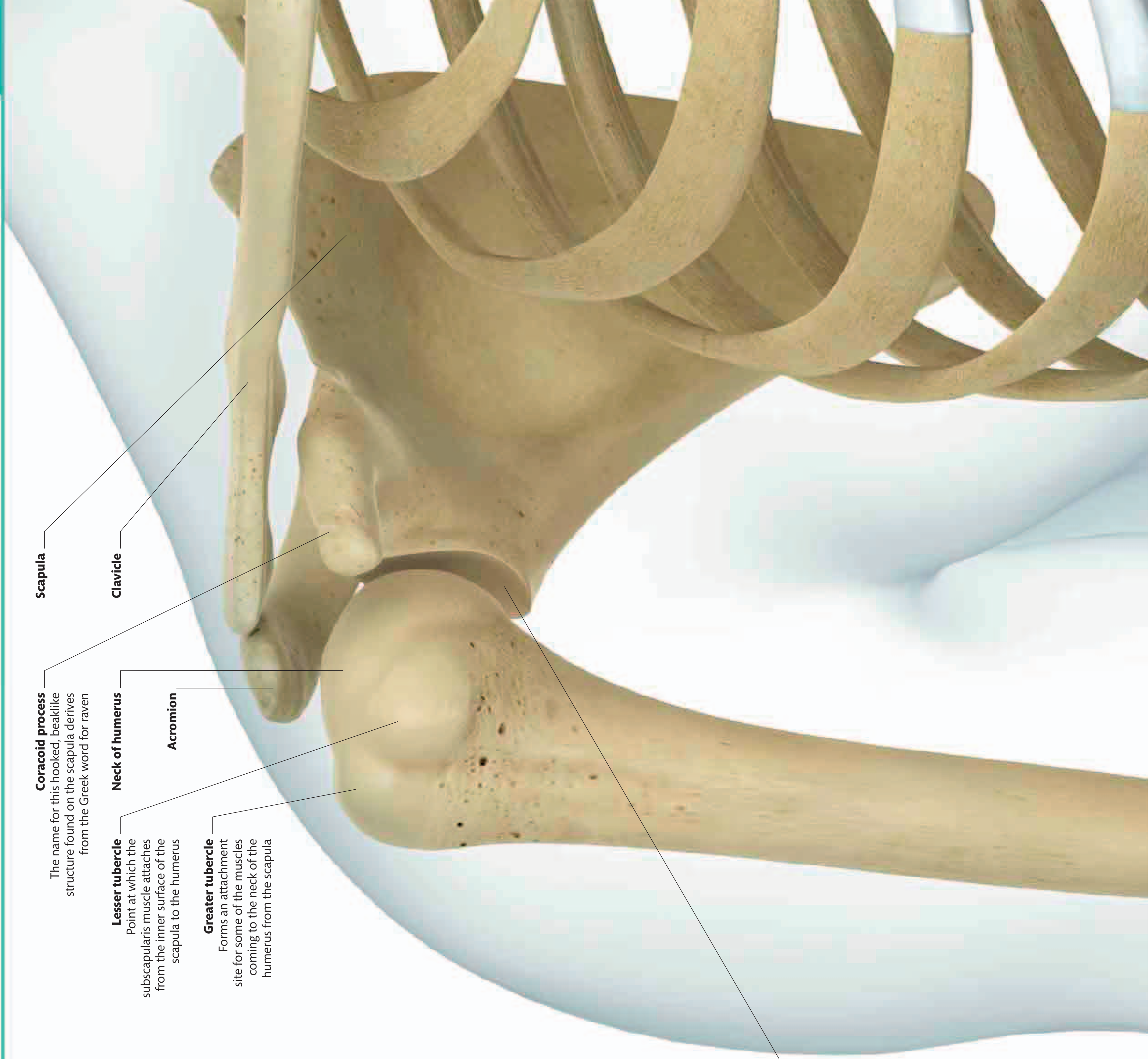
Scapula
Clavicle
Coracoid process The name for this hooked, beaklike structure found on the scapula derives from the Greek word for raven
Neck of humerus
Acromion
Lesser tubercle Point at which the subscapularis muscle attaches from the inner surface of the scapula to the humerus
Greater tubercle Forms an attachment site for some of the muscles coming to the neck of the humerus from the scapula
Glenoid fossa Shallow area that articulates with the head of the humerus, forming part of the shoulder socket

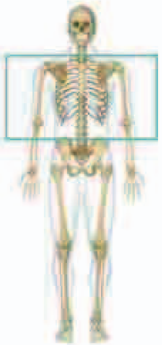

SHOULDER AND **UPPER ARM**

The scapula and clavicle make up the shoulder girdle, which anchors the arm to the thorax. This is a very mobile attachment—the scapula “floats” on the ribcage, attached to it by muscles only (rather than by a true joint) that pull the scapula around on the underlying ribs, altering the position of the shoulder joint. The clavicle has joints—it articulates with the acromion of the scapula laterally (at the side) and the sternum at the other end—and helps hold the shoulder out to the side while allowing the scapula to move around. The shoulder joint, the most mobile joint in the body, is a ball-and-socket joint, but the socket is small and shallow, allowing the ball-shaped head of the humerus to move freely.

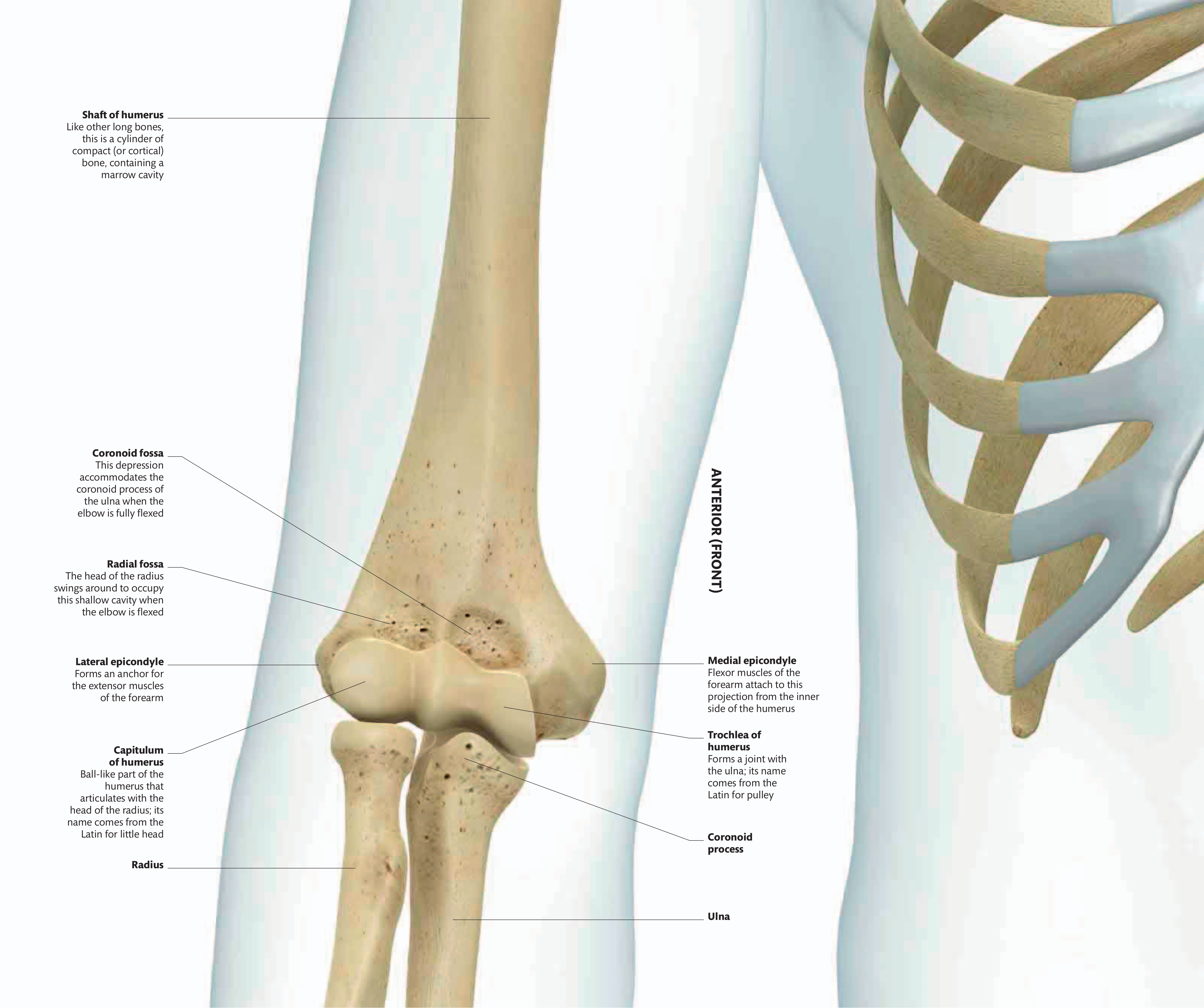

ANTERIOR (FRONT)

Shaft of humerus
Like other long bones, this is a cylinder of compact (or cortical) bone, containing a marrow cavity

Coronoid fossa
This depression accommodates the coronoid process of the ulna when the elbow is fully flexed

Radial fossa
The head of the radius swings around to occupy this shallow cavity when the elbow is flexed

Lateral epicondyle
Forms an anchor for the extensor muscles of the forearm

Capitulum of humerus
Ball-like part of the humerus that articulates with the head of the radius; its name comes from the Latin for little head

Radius

Medial epicondyle
Flexor muscles of the forearm attach to this projection from the inner side of the humerus

Trochlea of humerus
Forms a joint with the ulna; its name comes from the Latin for pulley

Coronoid process

Ulna

SHOULDER AND UPPER ARM

The back of the scapula is divided into two sections by its spine. The muscles that attach above this spine are called supraspinatus; those that attach below are called infraspinatus. They are part of the rotator cuff muscle group, which enables shoulder movements and stabilizes the shoulder joint. The spine of the scapula runs to the side and projects out above the shoulder joint to form the acromion, which can be easily felt on the top of the shoulder. The scapula rests in the position shown here when the arm is hanging at the side of the body. If the arm is abducted (raised to the side), the entire scapula rotates so that the glenoid cavity points upward and the inferior angle moves outward.

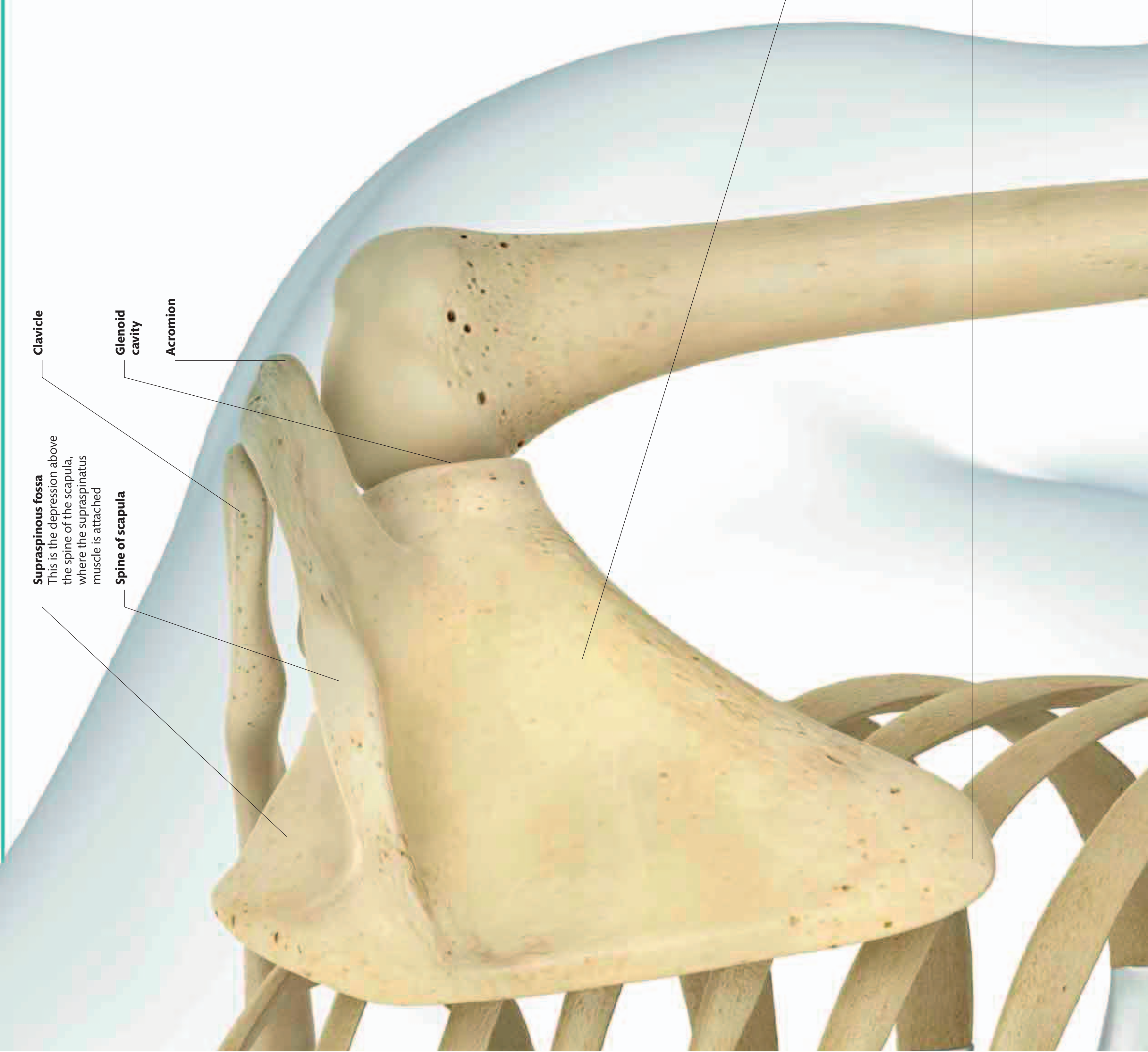

Supraspinous fossa
This is the depression above the spine of the scapula, where the supraspinatus muscle is attached

Spine of scapula

Clavicle

Glenoid cavity

Acromion

Infraspinous fossa
The infraspinatus muscle attaches to this part of the scapula—below its spine

Inferior angle

Spiral groove
This faint line marks where the radial nerve spirals around the posterior aspect of the humerus

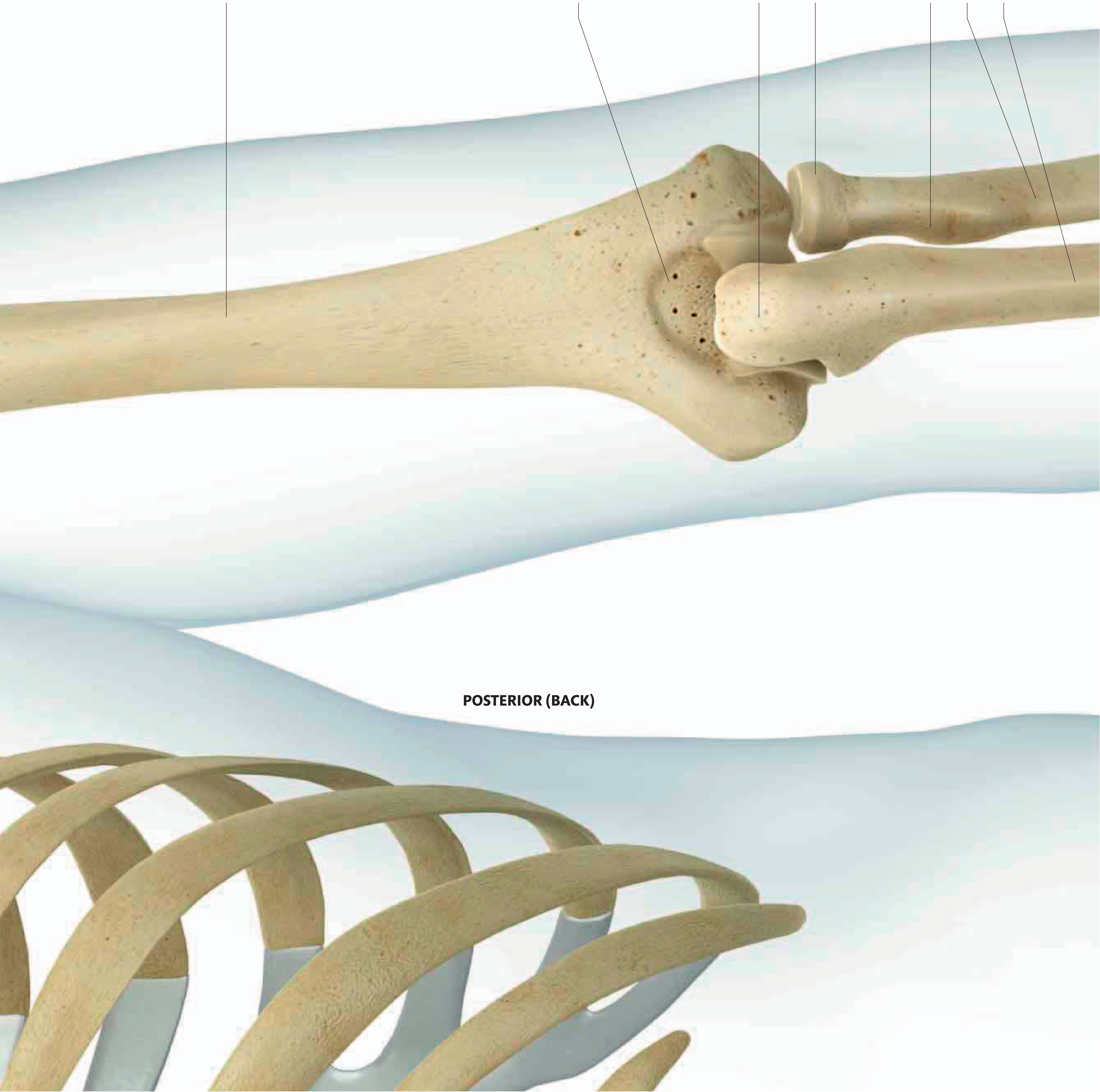

Shaft of humerus

Olecranon fossa A deep cavity on the posterior surface of the humerus; it accommodates the olecranon of ulna when the elbow is fully extended—as shown here

Olecranon of ulna

Head of radius

Radial tuberosity

Shaft of radius

Shaft of ulna

POSTERIOR (BACK)

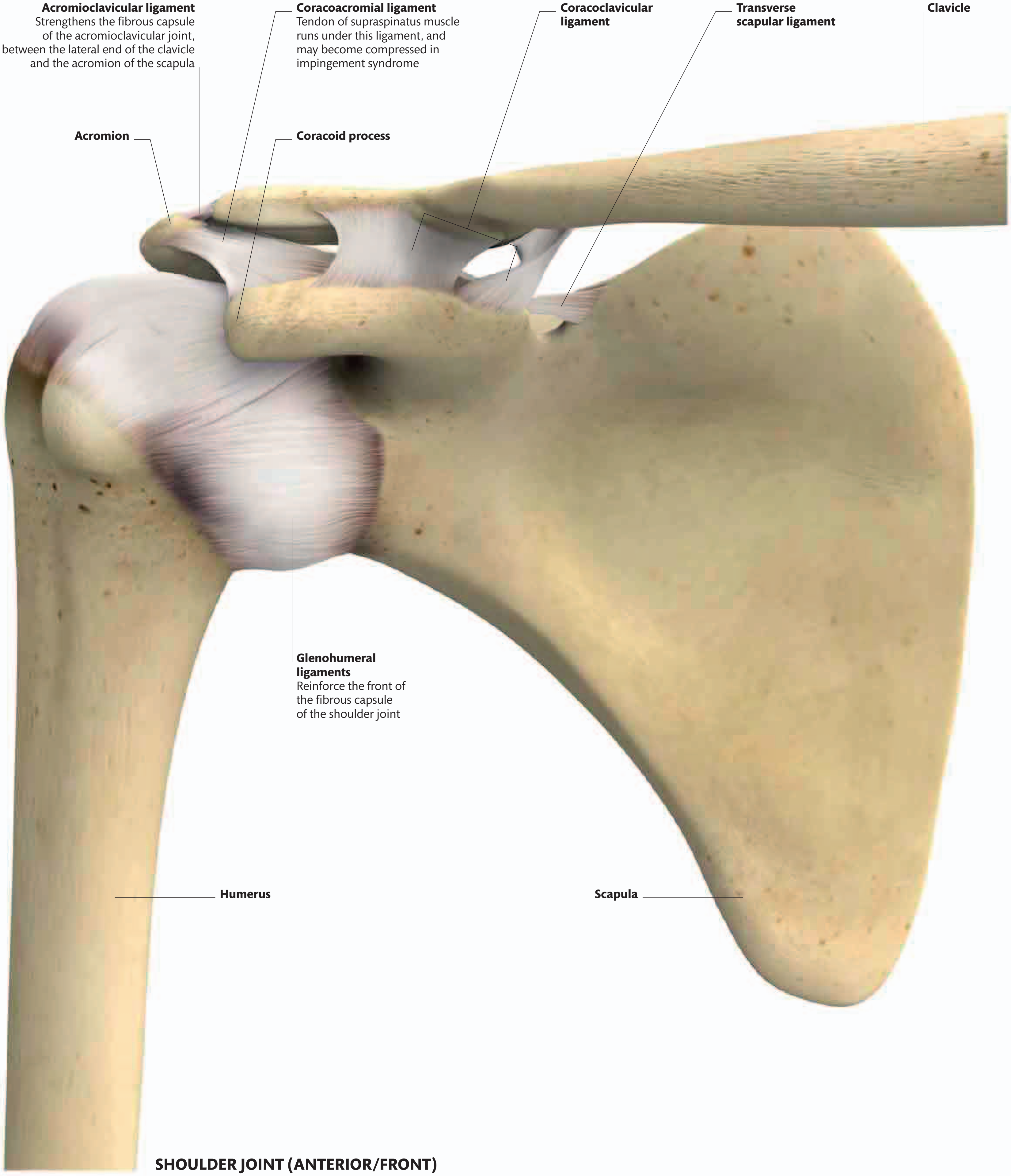

Acromioclavicular ligament
Strengthens the fibrous capsule of the acromioclavicular joint, between the lateral end of the clavicle and the acromion of the scapula
Coracoacromial ligament
Tendon of supraspinatus muscle runs under this ligament, and may become compressed in impingement syndrome
Coracoclavicular ligament
Transverse scapular ligament
Clavicle
Acromion
Coracoid process
Glenohumeral ligaments
Reinforce the front of the fibrous capsule of the shoulder joint
Humerus
Scapula

SHOULDER JOINT (ANTERIOR/FRONT)

SHOULDER AND UPPER ARM

In any joint, there is always a play off between mobility and stability. The extremely mobile shoulder joint is therefore naturally unstable, and so it is not surprising that this is the most commonly dislocated joint in the body. The coracoacromial arch, formed by the acromion and coracoid process of the scapula with the strong coracoacromial ligament stretching between them, prevents upward dislocation; when the head of the humerus dislocates, it usually does so in a downward direction. The elbow joint is formed by the articulation of the humerus with the forearm bones: the trochlea articulates with the ulna, and the capitulum with the head of the radius. The elbow is a hinge joint, stabilized by collateral ligaments on each side.

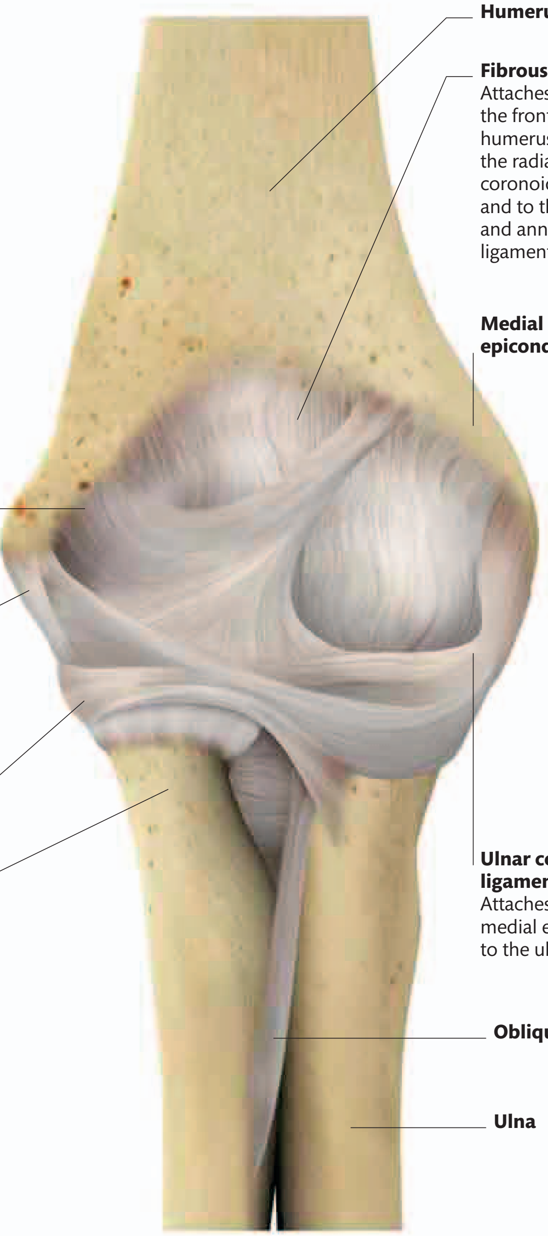

Fibrous capsule Attaches to the front of the humerus above the radial and coronoid fossae, and to the ulna and annular ligament below

Radial collateral ligament Attaches from the lateral epicondyle to the annular ligament

Ulnar collateral ligament Attaches from the medial epicondyle to the ulna

ELBOW (ANTERIOR/FRONT)

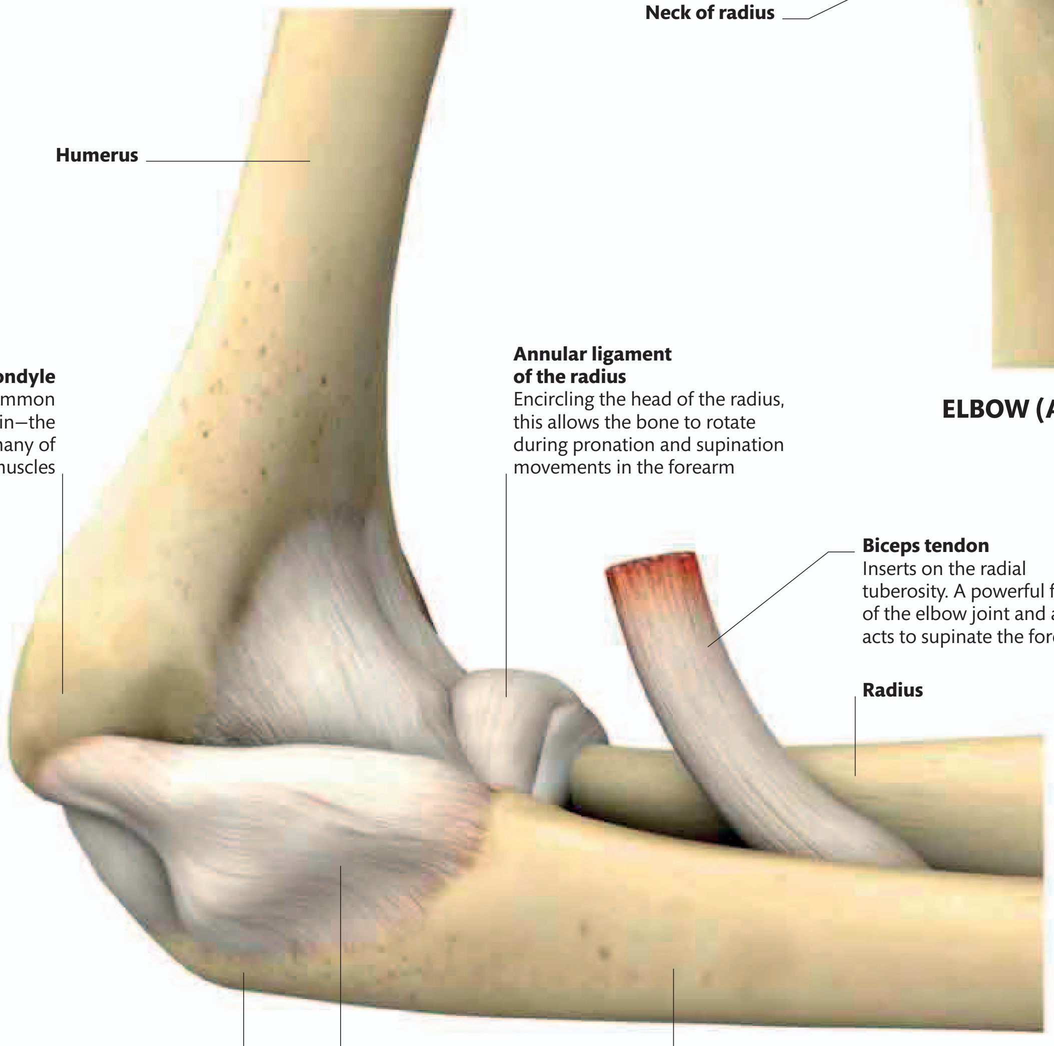

Medial epicondyle Also forms the common flexor origin—the attachment of many of the forearm flexor muscles

Annular ligament of the radius Encircling the head of the radius, this allows the bone to rotate during pronation and supination movements in the forearm

Biceps tendon Inserts on the radial tuberosity. A powerful flexor of the elbow joint and also acts to supinate the forearm

ELBOW (LATERAL/ OUTER SIDE)

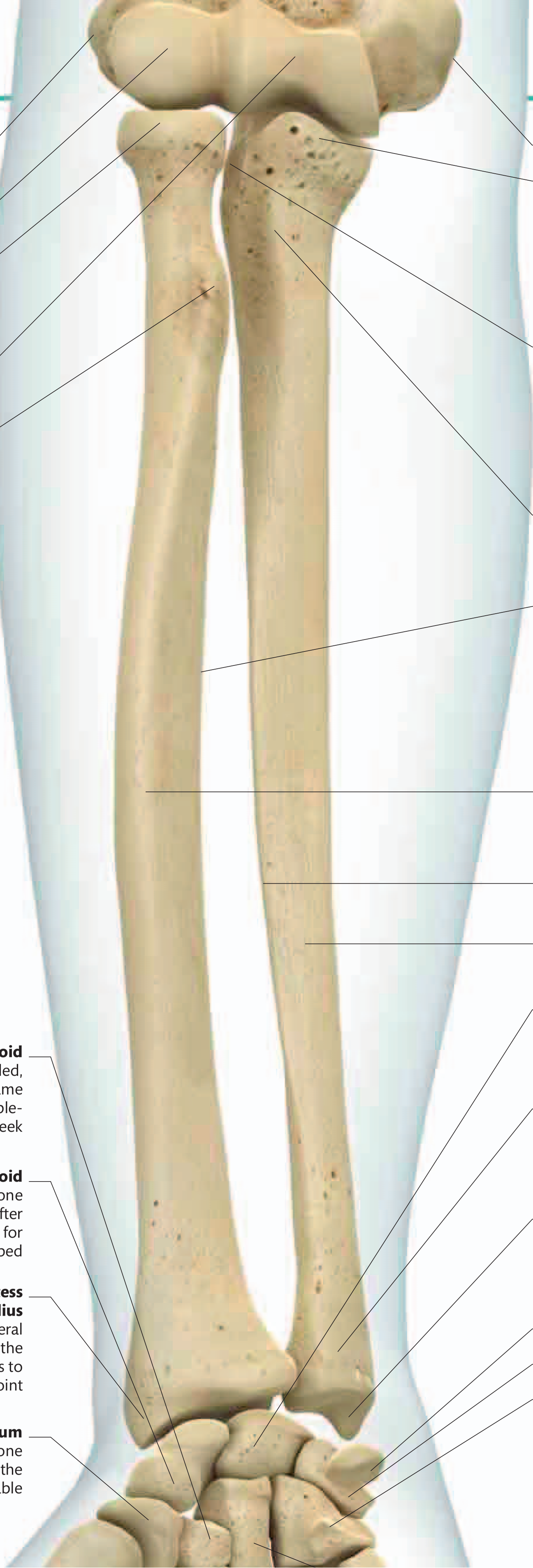
Medial epicondyle
Coronoid process
Forms anterior margin of the trochlear notch of the ulna, which accommodates the trochlea of the humerus
Radial notch of ulna
This concave surface articulates with the head of the radius, forming the proximal radioulnar joint
Tuberosity of ulna
Brachialis muscle attaches here
Interosseous border of radius
Sharp ridges on facing edges of the radius and ulna provide attachment for the forearm's interosseous membrane
Shaft of radius
Like the ulna, this is triangular in cross section
Interosseous border of ulna
Shaft of ulna
Lunate
Crescent-shaped bone named after the Latin for moon
Head of ulna
Articulates with lower end of the radius, at the distal radioulnar joint
Styloid process of ulna
Where the ulnar collateral ligament attaches
Pisiform
Triquetral
Hamate
One of the carpal bones, along with the other bones between the radius and ulna
Lateral epicondyle
Capitulum of humerus
Head of radius
Bowl-shaped surface articulates with the capitulum of humerus
Trochlea of humerus
Radial tuberosity
Biceps tendon attaches here
Trapezoid
Also four-sided, this bone's name means table-shaped in Greek
Scaphoid
Convex bone named after the Greek for boat-shaped
Styloid process of radius
The radial collateral ligament of the wrist attaches to this sharp point
Trapezium
Four-sided bone named after the Greek for table

POSTERIOR (BACK)

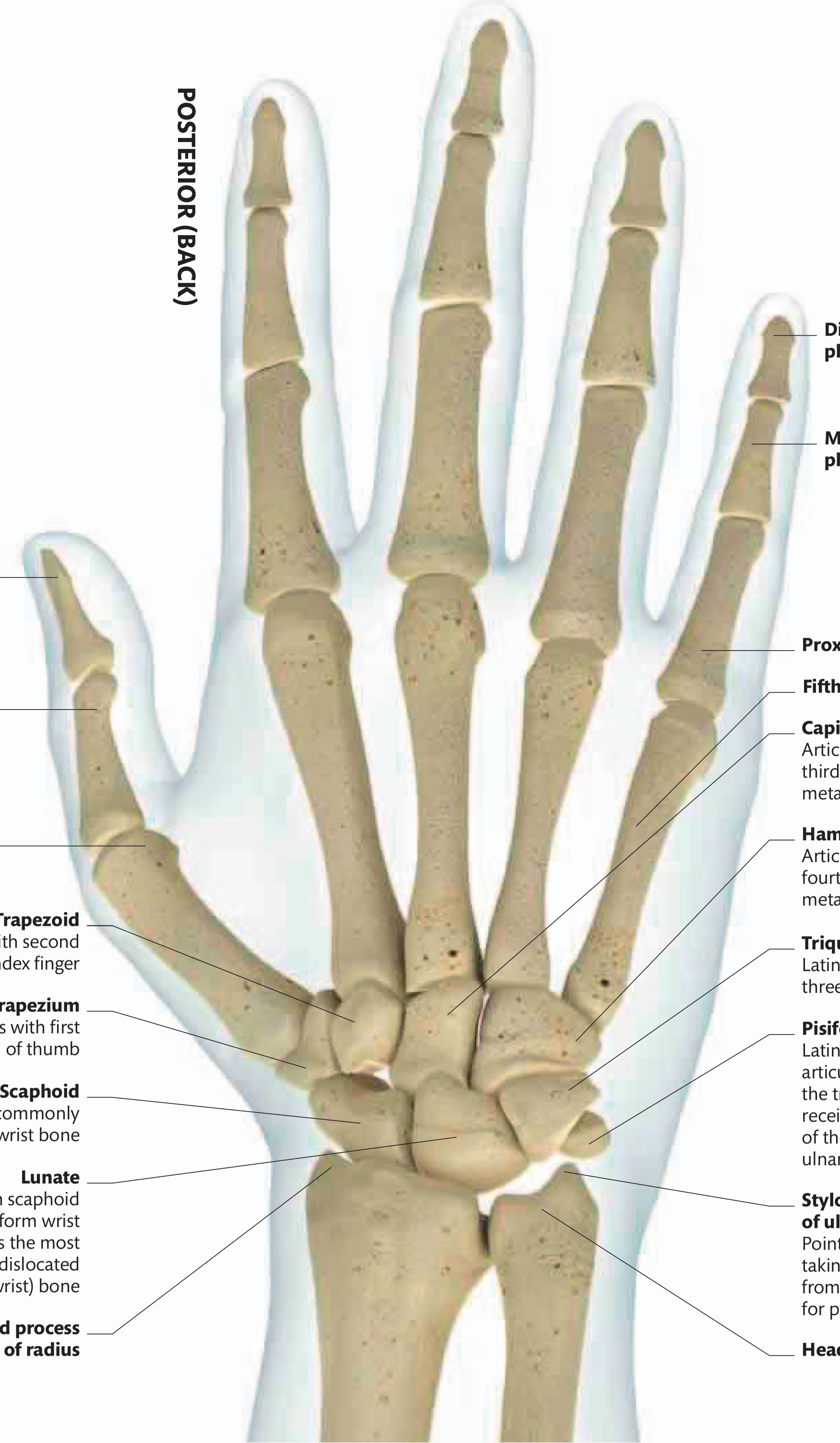
Distal phalanx
Middle phalanx
Proximal phalanx
Fifth metacarpal
Capitate
Articulates with third and fourth metacarpals
Hamate
Articulates with fourth and fifth metacarpals
Triquetral
Latin for three-cornered
Pisiform
Latin for pea-shaped; articulates with the triquetral, and receives the tendon of the flexor carpi ulnaris muscle
Styloid process of ulna
Pointed projection taking its name from the Greek for pillar-shaped
Head of ulna
Distal phalanx
Proximal phalanx
First metacarpal
Trapezoid
Articulates with second metacarpal of index finger
Trapezium
Articulates with first metacarpal of thumb
Scaphoid
The most commonly fractured wrist bone
Lunate
Articulates with scaphoid and radius to form wrist joint; this is the most commonly dislocated carpal (wrist) bone
Styloid process of radius

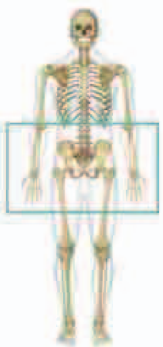

LOWER ARM AND HAND

The two forearm bones, the radius and ulna, are bound together by a flat sheet of ligament called the interosseous membrane, and by synovial joints between the ends of the two bones. Known as radioulnar joints, these joints allow the radius to move around the ulna. Hold your hand out in front of you, palm upward. Now turn your hand so that the palm faces the ground. This movement is called pronation, and is achieved by bringing the radius to cross over the ulna. The movement that returns the palm to an upward-facing position is called supination. Since the forearm bones are bound together by ligaments, joints, and muscles, it is common for both bones to be involved in a serious forearm injury. Often, one bone is fractured and the other dislocated. The skeleton of the hand comprises the eight carpal bones (bones between the radius and ulna), five metacarpals, and fourteen phalanges.

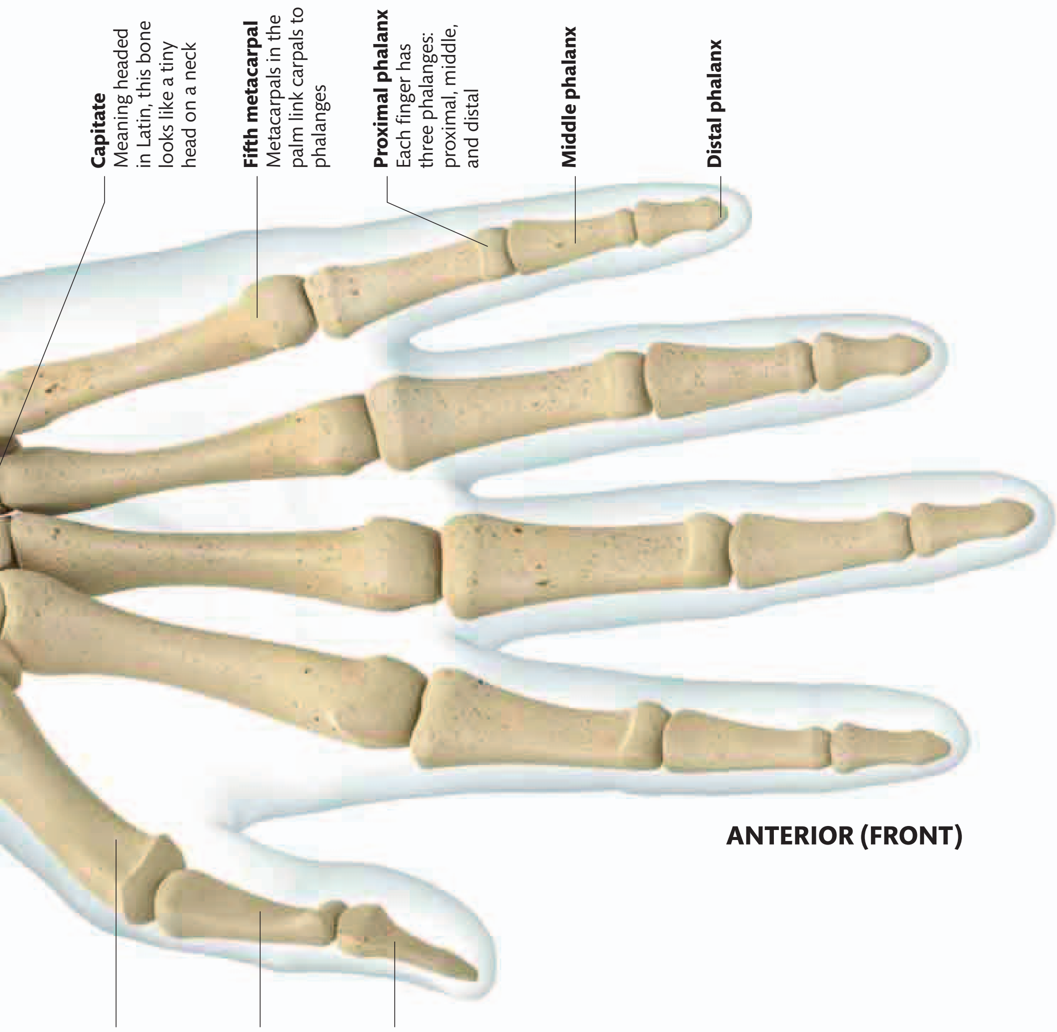
Capitate Meaning headed in Latin, this bone looks like a tiny head on a neck
Fifth metacarpal Metacarpals in the palm link carpals to phalanges
Proximal phalanx Each finger has three phalanges: proximal, middle, and distal
Middle phalanx
Distal phalanx
First metacarpal
Proximal phalanx
Distal phalanx of thumb The thumb has just two phalanges: proximal and distal

ANTERIOR (FRONT)

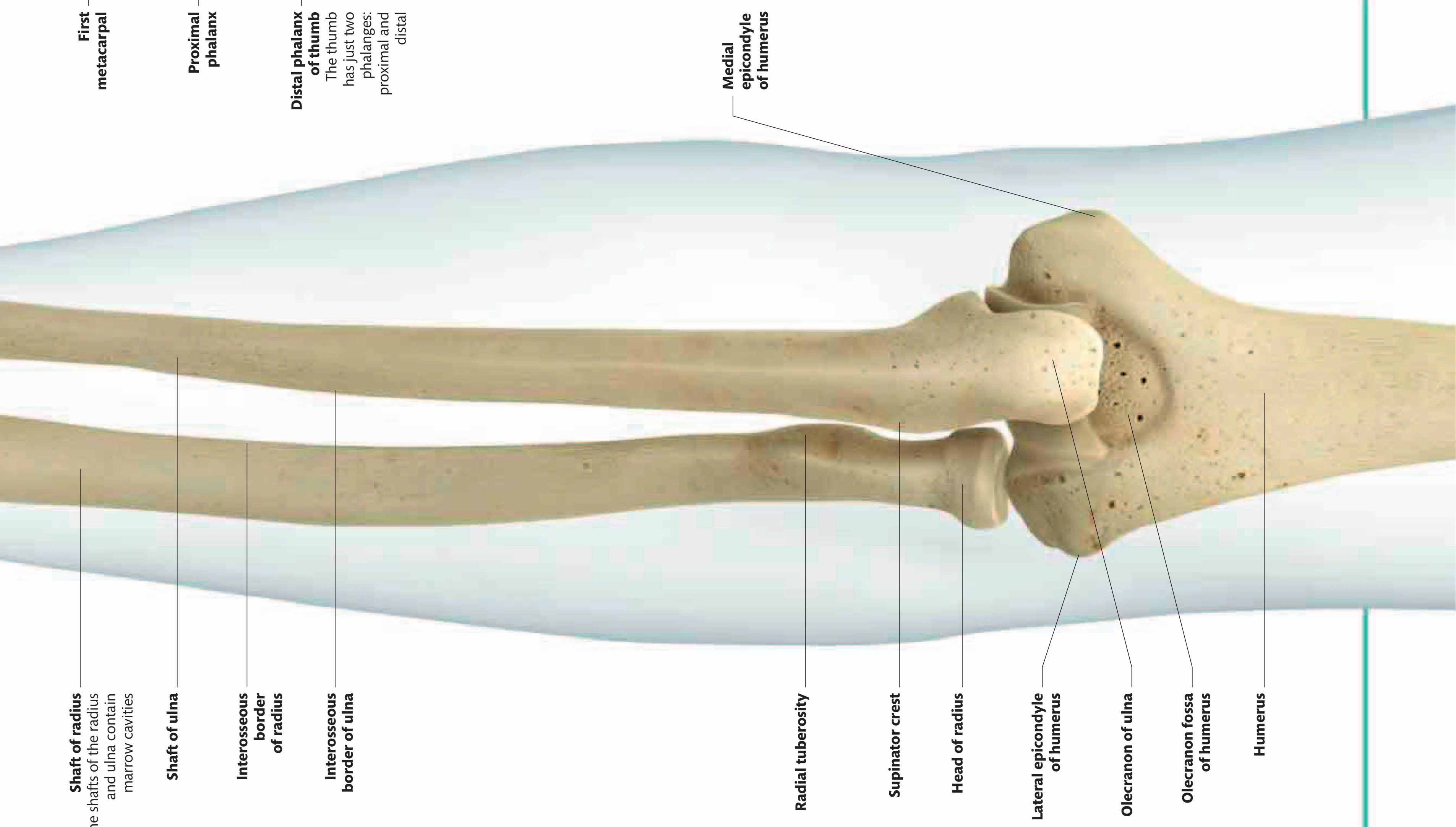
Medial epicondyle of humerus
Shaft of radius The shafts of the radius and ulna contain marrow cavities
Shaft of ulna
Interosseous border of radius
Interosseous border of ulna
Radial tuberosity
Supinator crest
Head of radius
Lateral epicondyle of humerus
Olecranon of ulna
Olecranon fossa of humerus
Humerus

HAND AND WRIST JOINTS

The radius widens out at its distal (lower) end to form the wrist joint with the closest two carpal bones, the lunate and scaphoid. This joint allows flexion, extension, adduction, and abduction (see pp.16–17). There are also synovial joints (see p.60) between the carpal bones in the wrist, which increase the range of motion during wrist flexion and extension. Synovial joints between metacarpals and phalanges allow us to spread or close our fingers, as well as flexing or extending the whole finger. Joints between the individual finger bones, or phalanges, enable fingers to bend and straighten. In common with many other primates, humans have opposable thumbs. The joints at the base of the thumb are shaped differently from those of the fingers. The joint between the metacarpal of the thumb and the wrist bones is especially mobile and allows the thumb to be brought across the palm of the hand so that the tip of the thumb can touch the other fingertips.

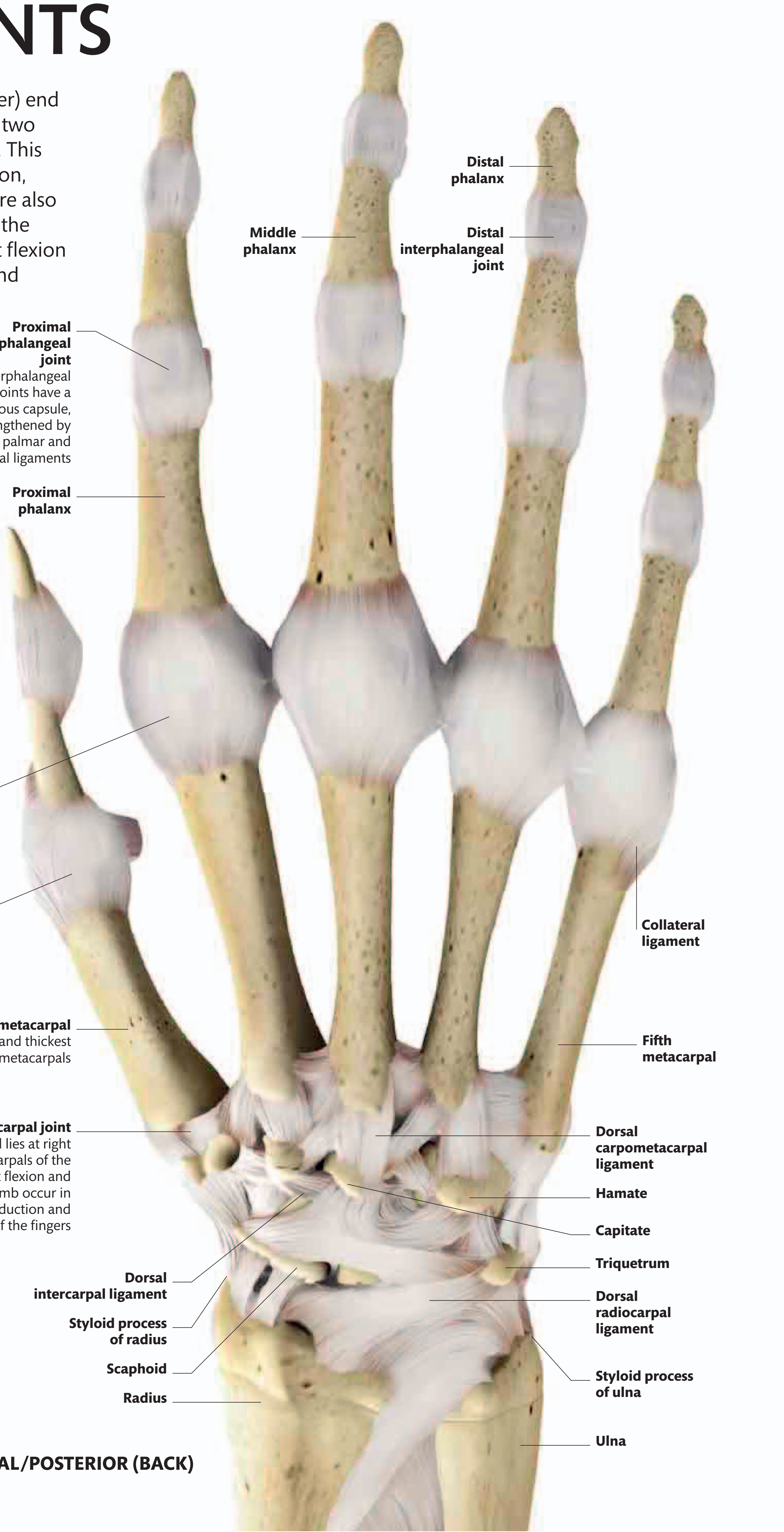
Distal phalanx
Middle phalanx
Distal interphalangeal joint
Proximal interphalangeal joint
Proximal phalanx
Second metacarpophalangeal joint
First metacarpophalangeal joint
Collateral ligament
First metacarpal
Fifth metacarpal
First carpometacarpal joint
Dorsal carpometacarpal ligament
Hamate
Capitate
Triquetrum
Dorsal intercarpal ligament
Dorsal radiocarpal ligament
Styloid process of radius
Scaphoid
Styloid process of ulna
Radius
Ulna

Proximal interphalangeal joint The interphalangeal joints have a fibrous capsule, strengthened by palmar and collateral ligaments

Second metacarpophalangeal joint These joints allow about 90 degrees of flexion, a very small amount of extension, and about 30 degrees of abduction and adduction of the metacarpals

First metacarpophalangeal joint Allows about 60 degrees of flexion, a little extension, as well as abduction and adduction

First metacarpal The shortest and thickest of the metacarpals

First carpometacarpal joint The first metacarpal lies at right angles to the metacarpals of the fingers, so that flexion and extension of the thumb occur in the same plane as abduction and adduction of the fingers

DORSAL/POSTERIOR (BACK)

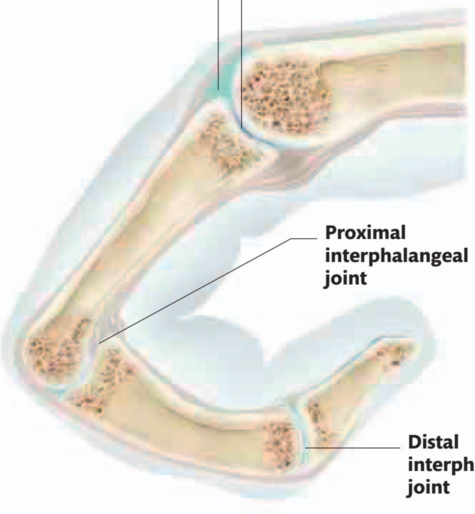
Joint capsule
Metacarpophalangeal joint
Proximal interphalangeal joint
Distal interphalangeal joint

FINGER (SAGITTAL SECTION)

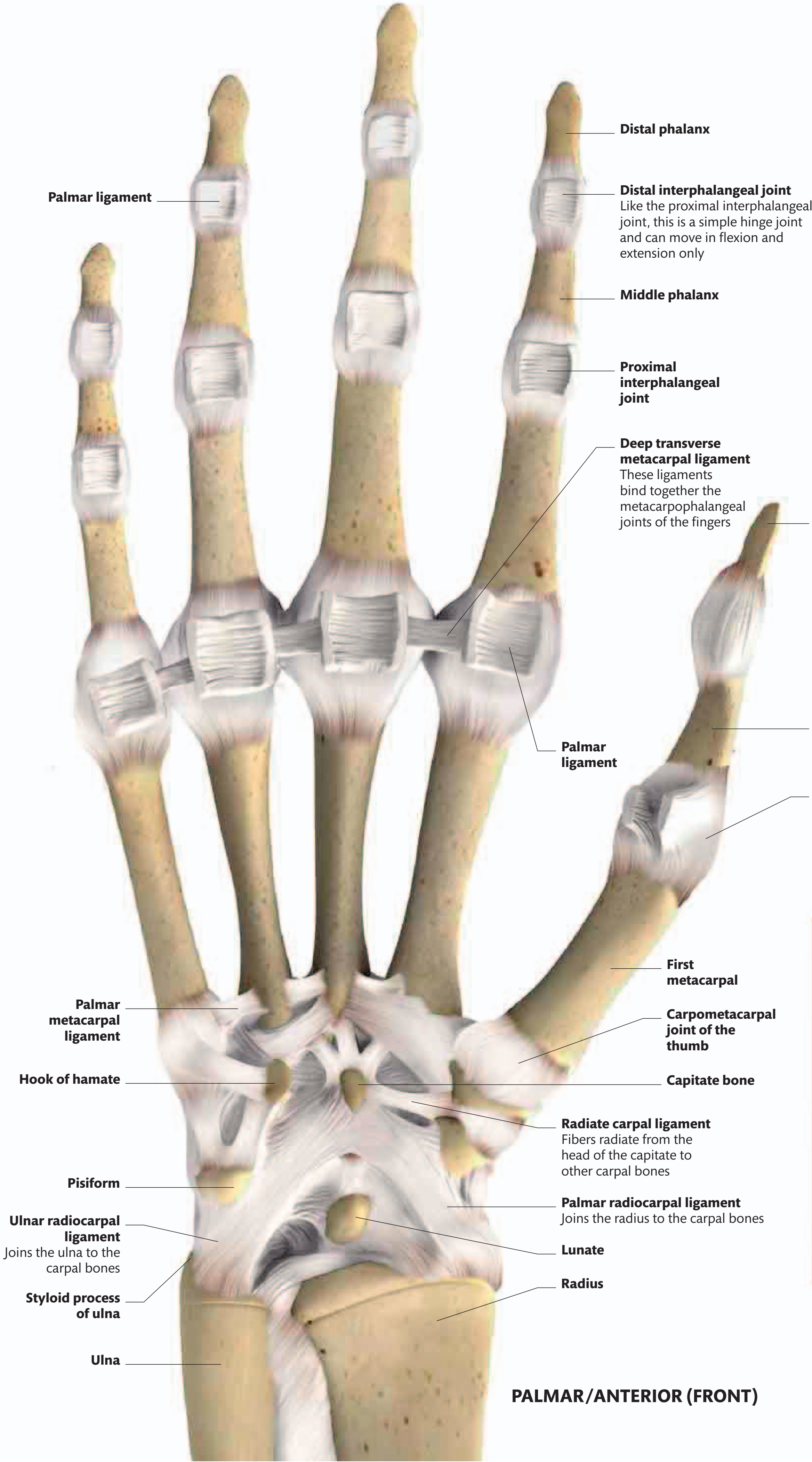
Palmar ligament
Distal phalanx
Distal interphalangeal joint
Like the proximal interphalangeal joint, this is a simple hinge joint and can move in flexion and extension only
Middle phalanx
Proximal interphalangeal joint
Deep transverse metacarpal ligament
These ligaments bind together the metacarpophalangeal joints of the fingers
Distal phalanx
Proximal phalanx
Palmar ligament
Metacarpophalangeal joint
First metacarpal
Palmar metacarpal ligament
Carpometacarpal joint of the thumb
Hook of hamate
Capitate bone
Radiate carpal ligament
Fibers radiate from the head of the capitate to other carpal bones
Pisiform
Palmar radiocarpal ligament
Joins the radius to the carpal bones
Ulnar radiocarpal ligament
Joins the ulna to the carpal bones
Lunate
Radius
Styloid process of ulna
Ulna

PALMAR/ANTERIOR (FRONT)

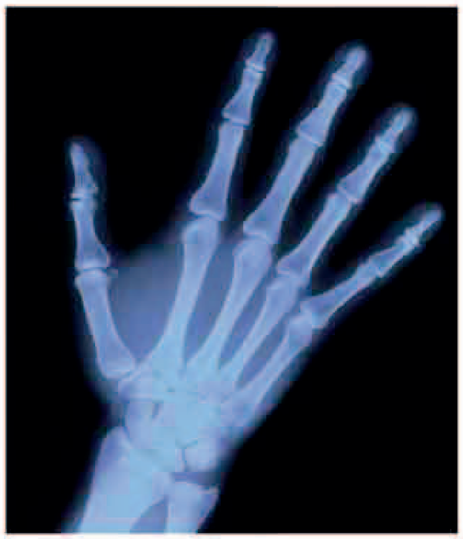

Hand X-ray
This X-ray of the hand clearly shows the carpal bones in the wrist and the joints between them. Near the metacarpophalangeal joint of the thumb, the thumb's tiny sesamoid bones, embedded in tendons, are also visible.

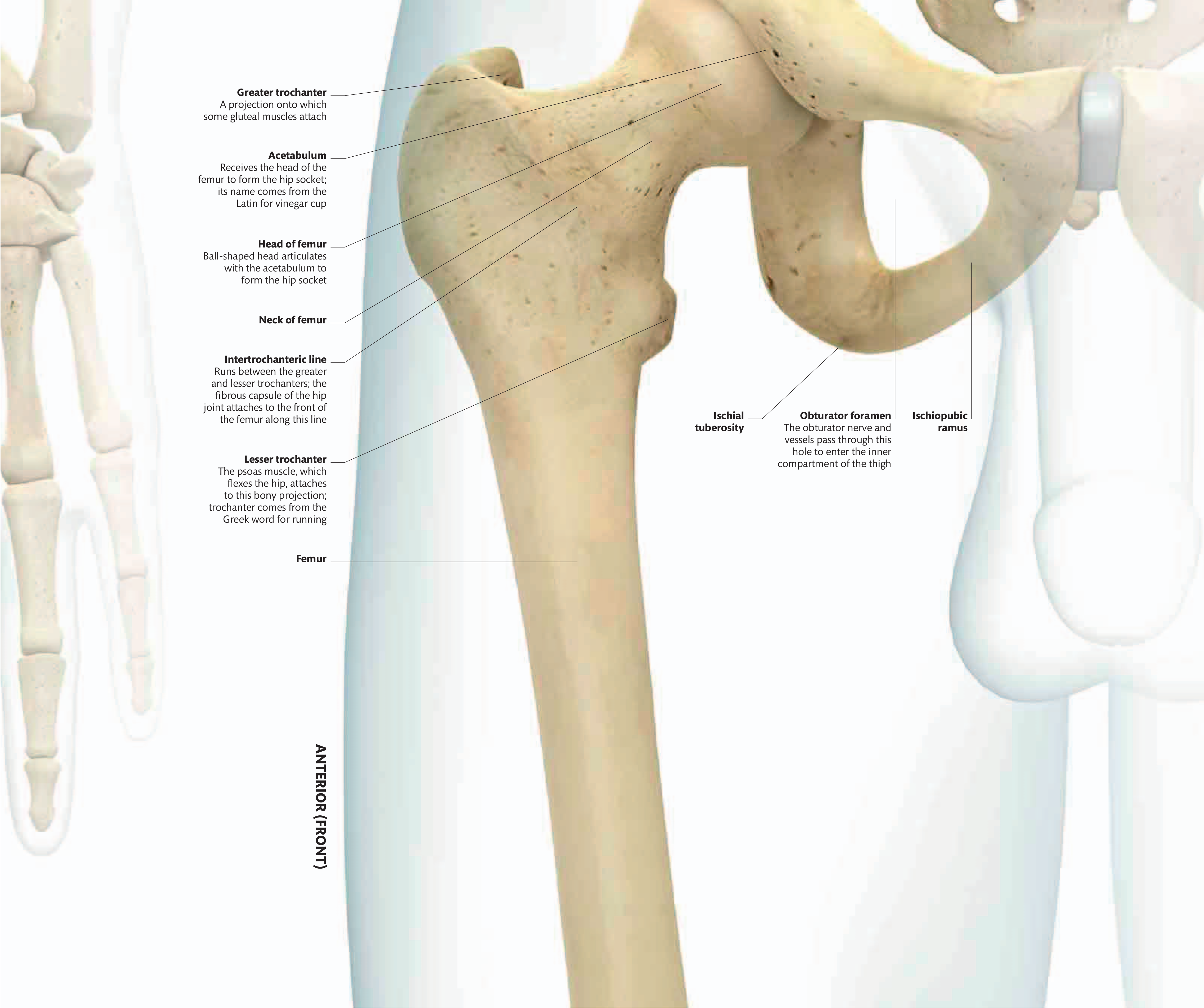

Greater trochanter
A projection onto which some gluteal muscles attach

Acetabulum
Receives the head of the femur to form the hip socket; its name comes from the Latin for vinegar cup

Head of femur
Ball-shaped head articulates with the acetabulum to form the hip socket

Neck of femur

Intertrochanteric line
Runs between the greater and lesser trochanters; the fibrous capsule of the hip joint attaches to the front of the femur along this line

Lesser trochanter
The psoas muscle, which flexes the hip, attaches to this bony projection; trochanter comes from the Greek word for running

Femur

Ischial tuberosity

Obturator foramen
The obturator nerve and vessels pass through this hole to enter the inner compartment of the thigh

Ischiopubic ramus

ANTERIOR (FRONT)

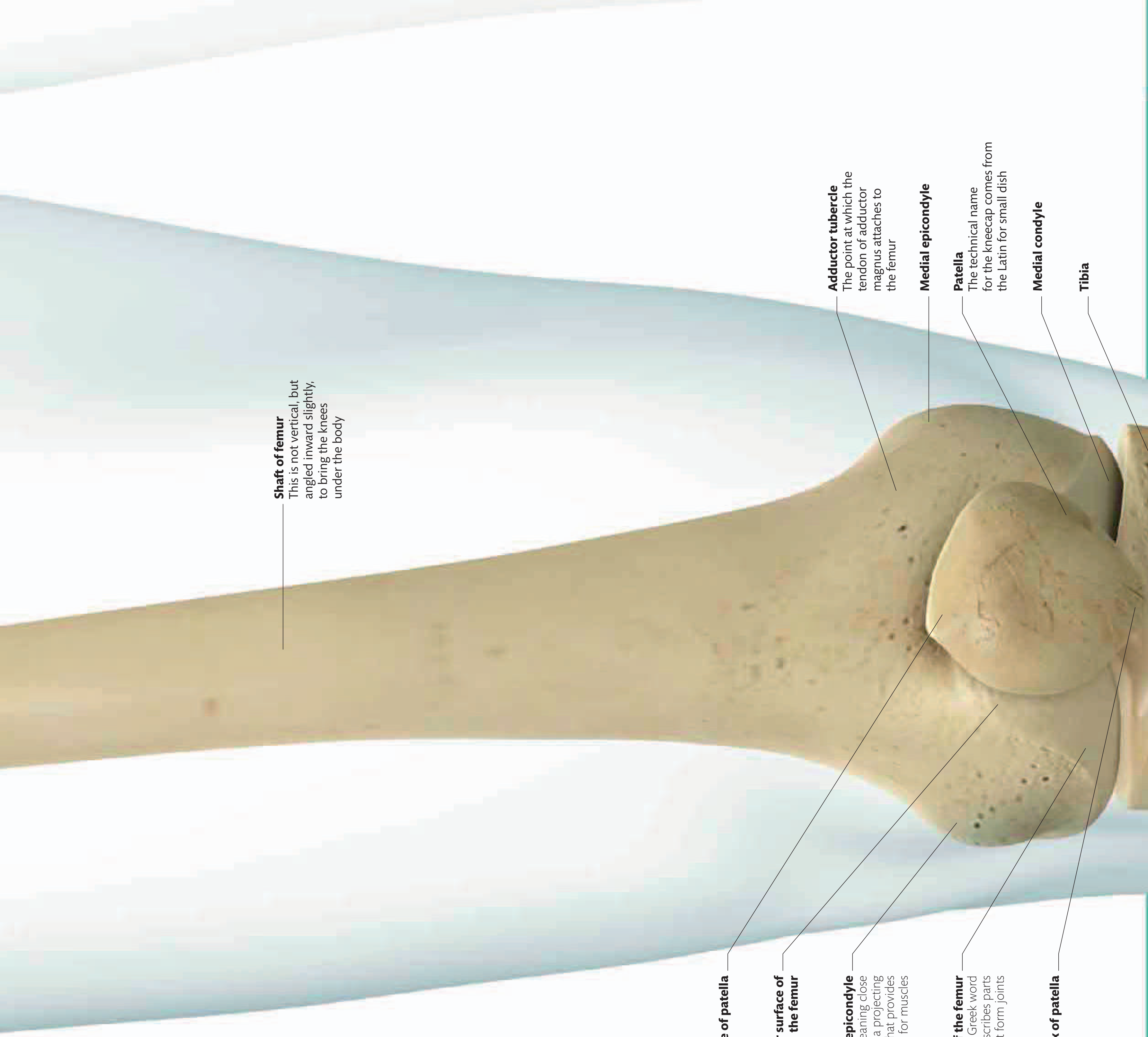

Shaft of femur This is not vertical, but angled inward slightly, to bring the knees under the body

Adductor tubercle The point at which the tendon of adductor magnus attaches to the femur

Medial epicondyle

Patella The technical name for the kneecap comes from the Latin for small dish

Medial condyle

Tibia

Base of patella

Patellar surface of the femur

Lateral epicondyle The term epicondyle (meaning close to the condyle) describes a projecting part of bone near a joint that provides a point of attachment for muscles

Lateral condyle of the femur Condyle comes from the Greek word for knuckle; the term describes parts of the ends of bones that form joints

Apex of patella

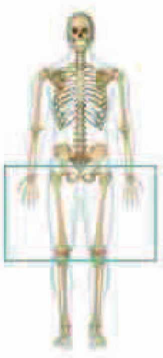

HIP AND THIGH

The leg or, to be anatomically precise, the lower limb, is attached to the spine by the pelvic bones. This is a much more stable arrangement than that of the shoulder girdle, which anchors the arm, because the legs and pelvis must bear our body weight as we stand or move around. The sacroiliac joint provides a strong attachment between the ilium of the pelvis and the sacrum, and the hip joint is a much deeper and more stable ball-and-socket joint than that in the shoulder. The neck of the femur joins the head at an obtuse angle. A slightly raised diagonal line on the front of the neck (the intertrochanteric line) shows where the fibrous capsule of the hip joint attaches to the bone.

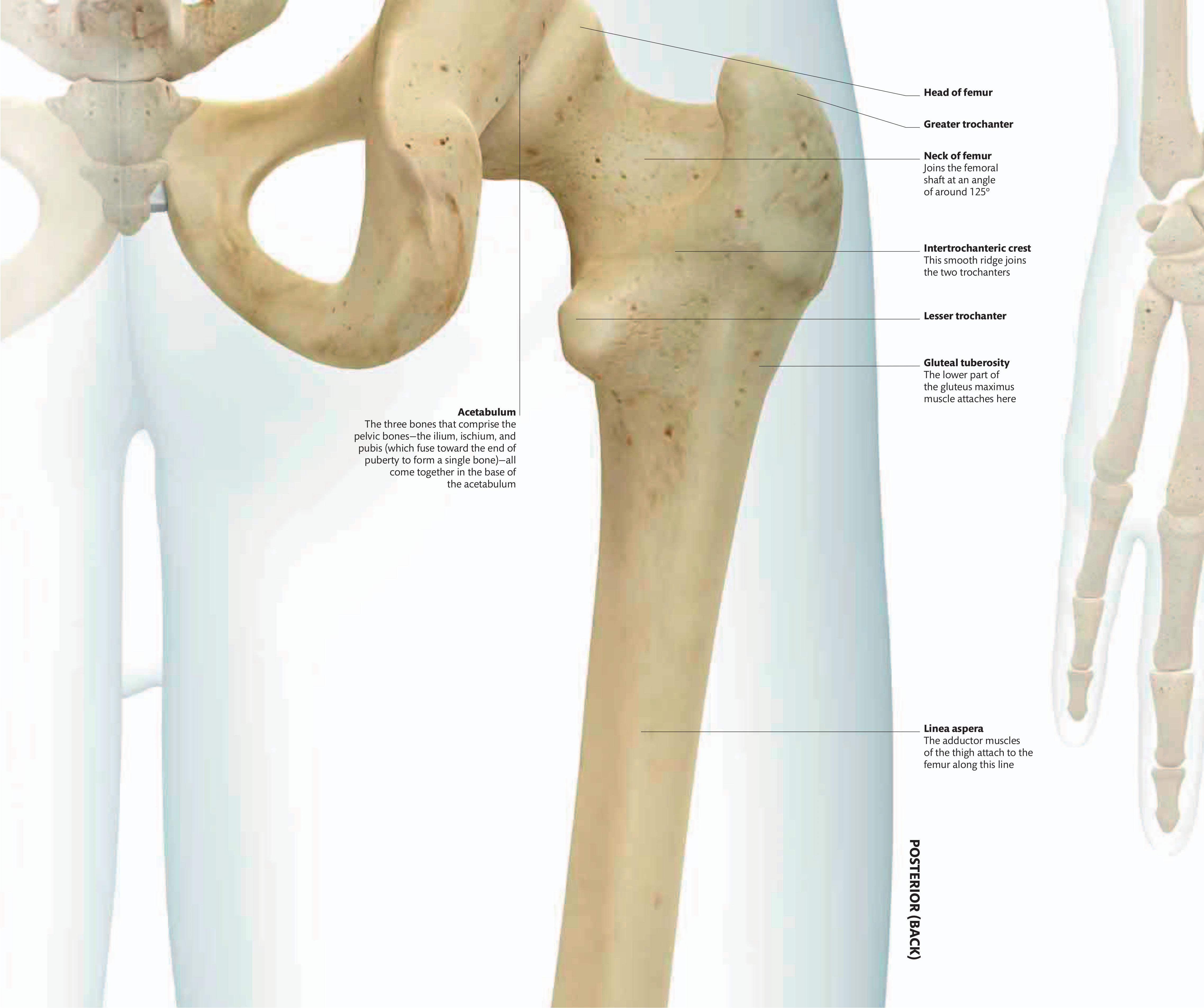

Head of femur

Greater trochanter

Neck of femur
Joins the femoral shaft at an angle of around 125°

Intertrochanteric crest
This smooth ridge joins the two trochanters

Lesser trochanter

Gluteal tuberosity
The lower part of the gluteus maximus muscle attaches here

Acetabulum
The three bones that comprise the pelvic bones—the ilium, ischium, and pubis (which fuse toward the end of puberty to form a single bone)—all come together in the base of the acetabulum

Linea aspera
The adductor muscles of the thigh attach to the femur along this line

POSTERIOR (BACK)

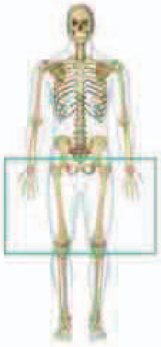

HIP AND THIGH

The shaft of the femur (thighbone) is cylindrical, with a marrow cavity. The linea aspera runs down along the back of the femoral shaft. This line is where the inner thigh's adductor muscles attach to the femur. Parts of the quadriceps muscle also wrap right around the back of the femur to attach to the linea aspera. At the bottom—or distal—end, toward the knee, the femur widens to form the knee joint with the tibia and the patella. From the back, the distal end of the femur has a distinct double-knuckle shape, with two condyles (rounded projections) that articulate with the tibia.

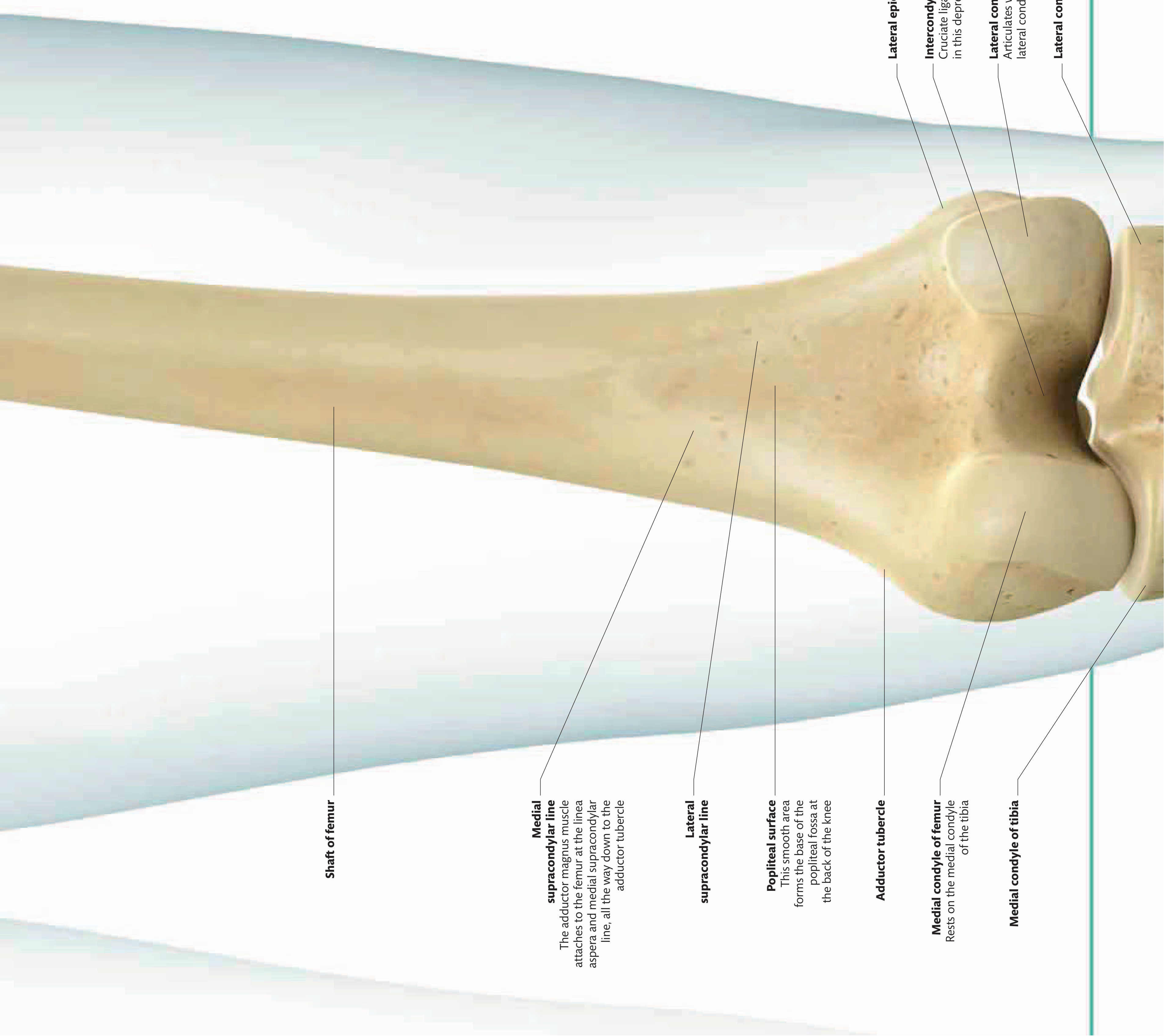
Lateral epicondyle
Intercondylar fossa Cruciate ligaments attach to the femur in this depression between the condyles
Lateral condyle of femur Articulates with the slightly concave lateral condyle of the tibia
Lateral condyle of tibia
Shaft of femur
Medial supracondylar line The adductor magnus muscle attaches to the femur at the linea aspera and medial supracondylar line, all the way down to the adductor tubercle
Lateral supracondylar line
Popliteal surface This smooth area forms the base of the popliteal fossa at the back of the knee
Adductor tubercle
Medial condyle of femur Rests on the medial condyle of the tibia
Medial condyle of tibia

HIP AND KNEE

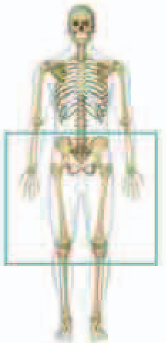

The hip joint is very stable. Its fibrous capsule is strengthened by ligaments that attach from the neck of the femur to the pelvic bone. These are the iliofemoral and pubofemoral ligaments at the front and the ischiofemoral ligament at the back. Inside the joint capsule, a small ligament attaches from the edge of the acetabulum (hip socket) to the head of the femur. The knee joint is formed by the articulation of the femur with the tibia and patella. Although primarily a hinge joint, the knee also permits some rotation to occur. These complex movements are reflected by the complexity of the joint: there are crescent-shaped articular disks (menisci) inside the joint, powerful collateral ligaments on either side of the joint, as well as crossed-over cruciate ligaments binding the femur to the tibia, and numerous extra pockets of synovial fluid, called bursae, that lubricate tendons around the joint.

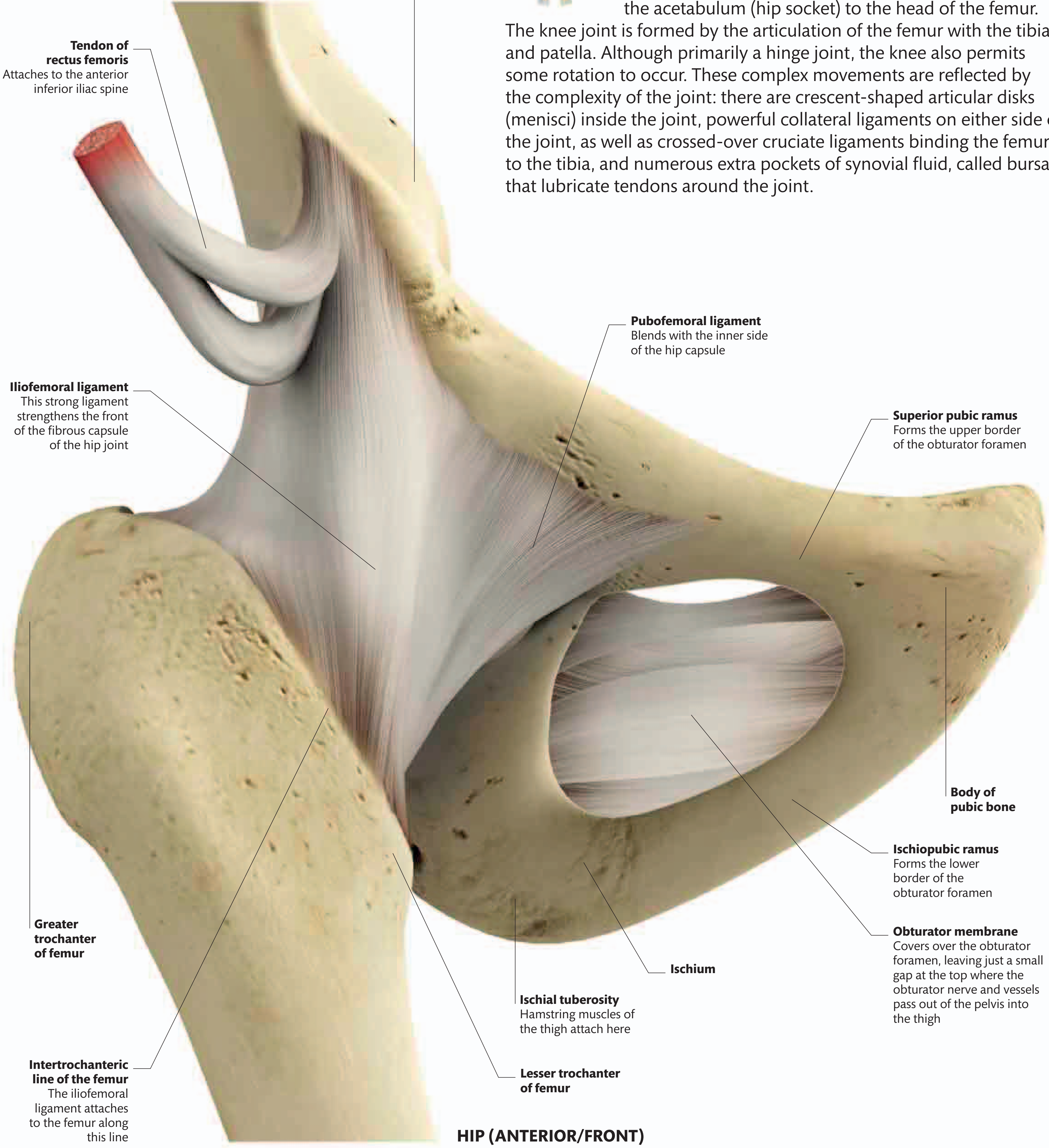
Ilium
Ilium, pubis, and ischium of the pelvis all meet in the acetabulum or hip socket
Tendon of rectus femoris
Attaches to the anterior inferior iliac spine
Iliofemoral ligament
This strong ligament strengthens the front of the fibrous capsule of the hip joint
Pubofemoral ligament
Blends with the inner side of the hip capsule
Superior pubic ramus
Forms the upper border of the obturator foramen
Body of pubic bone
Ischiopubic ramus
Forms the lower border of the obturator foramen
Obturator membrane
Covers over the obturator foramen, leaving just a small gap at the top where the obturator nerve and vessels pass out of the pelvis into the thigh
Ischium
Ischial tuberosity
Hamstring muscles of the thigh attach here
Lesser trochanter of femur
Greater trochanter of femur
Intertrochanteric line of the femur
The iliofemoral ligament attaches to the femur along this line

HIP (ANTERIOR/FRONT)

SYNOVIAL JOINTS

The majority of the body's 320 or so joints, including those in the finger, knee, and shoulder, are free-moving synovial joints. The joint surfaces are lined with smooth hyaline cartilage to reduce friction, and contain lubricating synovial fluid.

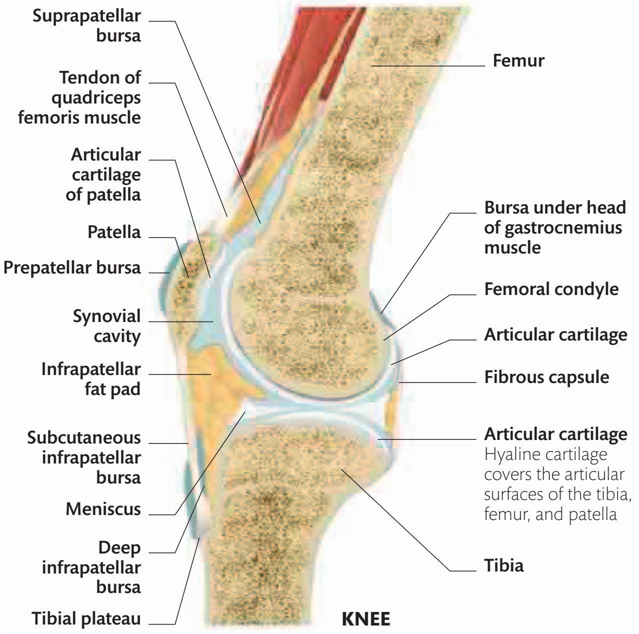
Suprapatellar bursa
Tendon of quadriceps femoris muscle
Articular cartilage of patella
Patella
Prepatellar bursa
Synovial cavity
Infrapatellar fat pad
Subcutaneous infrapatellar bursa
Meniscus
Deep infrapatellar bursa
Tibial plateau
Femur
Bursa under head of gastrocnemius muscle
Femoral condyle
Articular cartilage
Fibrous capsule
Articular cartilage
Hyaline cartilage covers the articular surfaces of the tibia, femur, and patella
Tibia
KNEE

Complex joint

A complex synovial joint, such as the knee, has articular discs or menisci inside the synovial cavity. The knee is also a compound hinge joint, as it involves more than two bones. Its complex anatomy allows it to move in flexion and extension, but some sliding and axial rotation of the femur on the tibia also occurs.

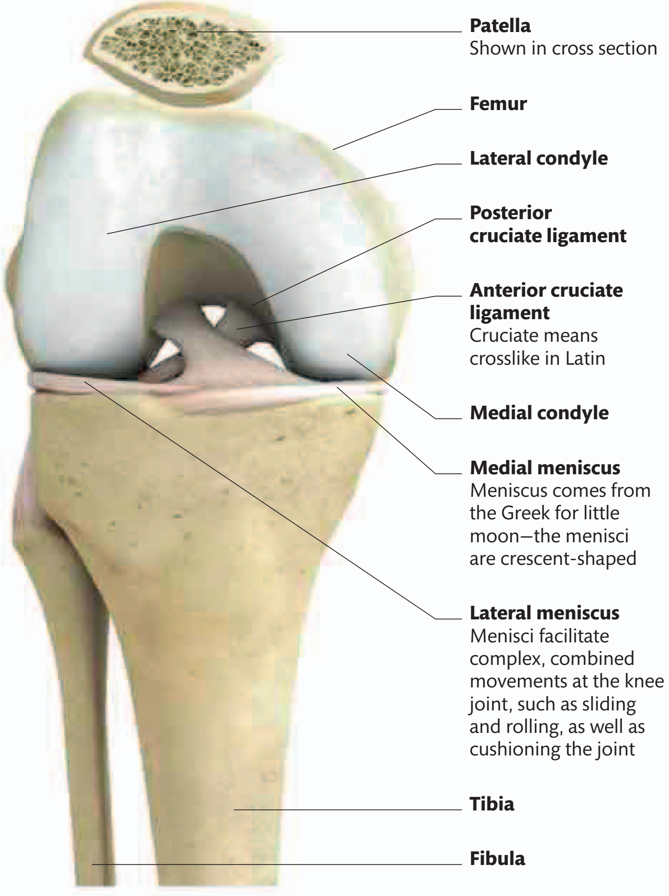
Patella
Shown in cross section
Femur
Lateral condyle
Posterior cruciate ligament
Anterior cruciate ligament
Cruciate means crosslike in Latin
Medial condyle
Medial meniscus
Meniscus comes from the Greek for little moon—the menisci are crescent-shaped
Lateral meniscus
Menisci facilitate complex, combined movements at the knee joint, such as sliding and rolling, as well as cushioning the joint
Tibia
Fibula

KNEE (FLEXED)

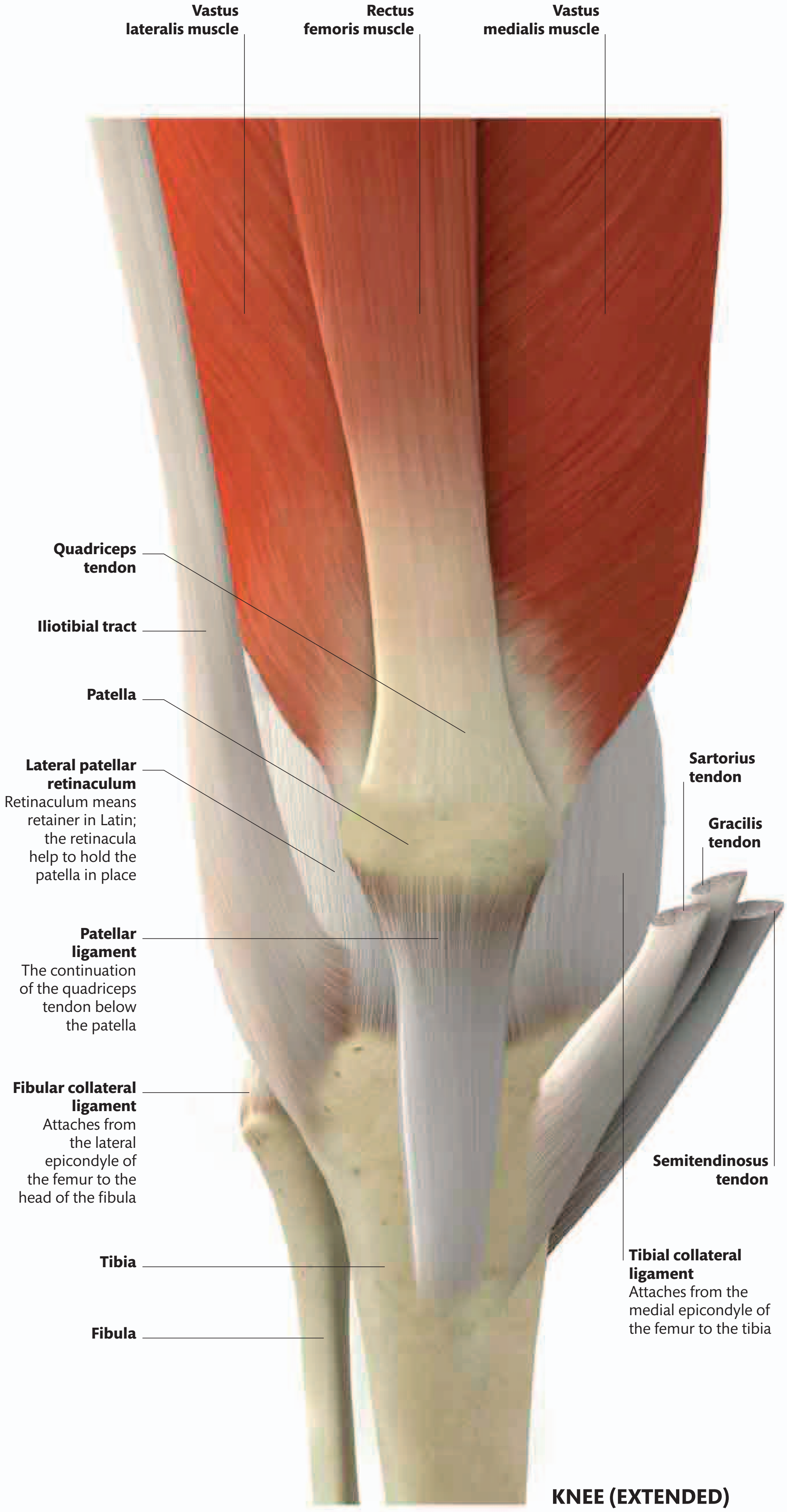
Vastus lateralis muscle
Rectus femoris muscle
Vastus medialis muscle
Quadriceps tendon
Iliotibial tract
Patella
Lateral patellar retinaculum
Retinaculum means retainer in Latin; the retinacula help to hold the patella in place
Patellar ligament
The continuation of the quadriceps tendon below the patella
Fibular collateral ligament
Attaches from the lateral epicondyle of the femur to the head of the fibula
Tibia
Fibula
Sartorius tendon
Gracilis tendon
Semitendinosus tendon
Tibial collateral ligament
Attaches from the medial epicondyle of the femur to the tibia

KNEE (EXTENDED)

LOWER LEG AND FOOT

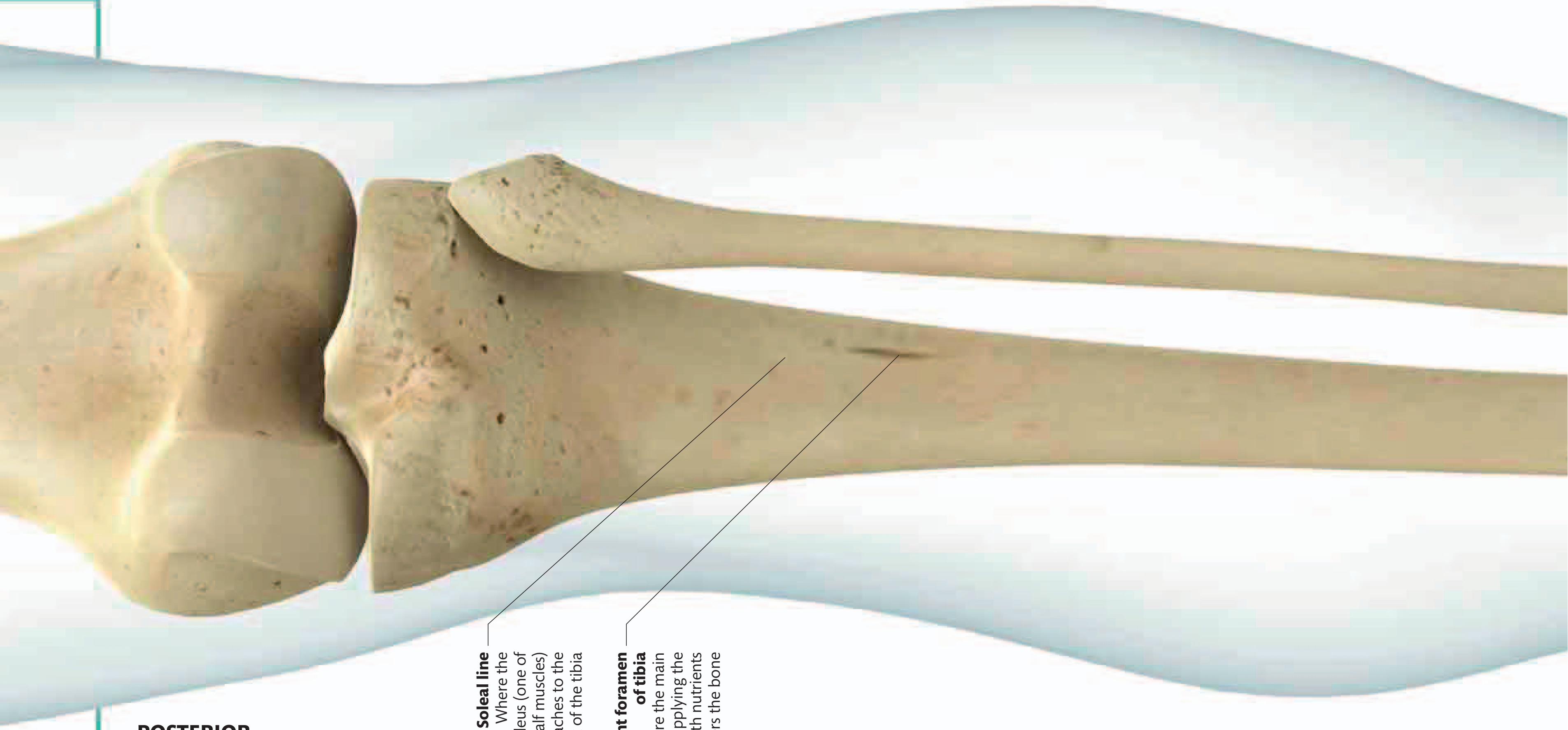

POSTERIOR (BACK)

Soleal line
Where the soleus (one of the calf muscles) attaches to the back of the tibia

Nutrient foramen of tibia
Where the main artery supplying the tibia with nutrients enters the bone

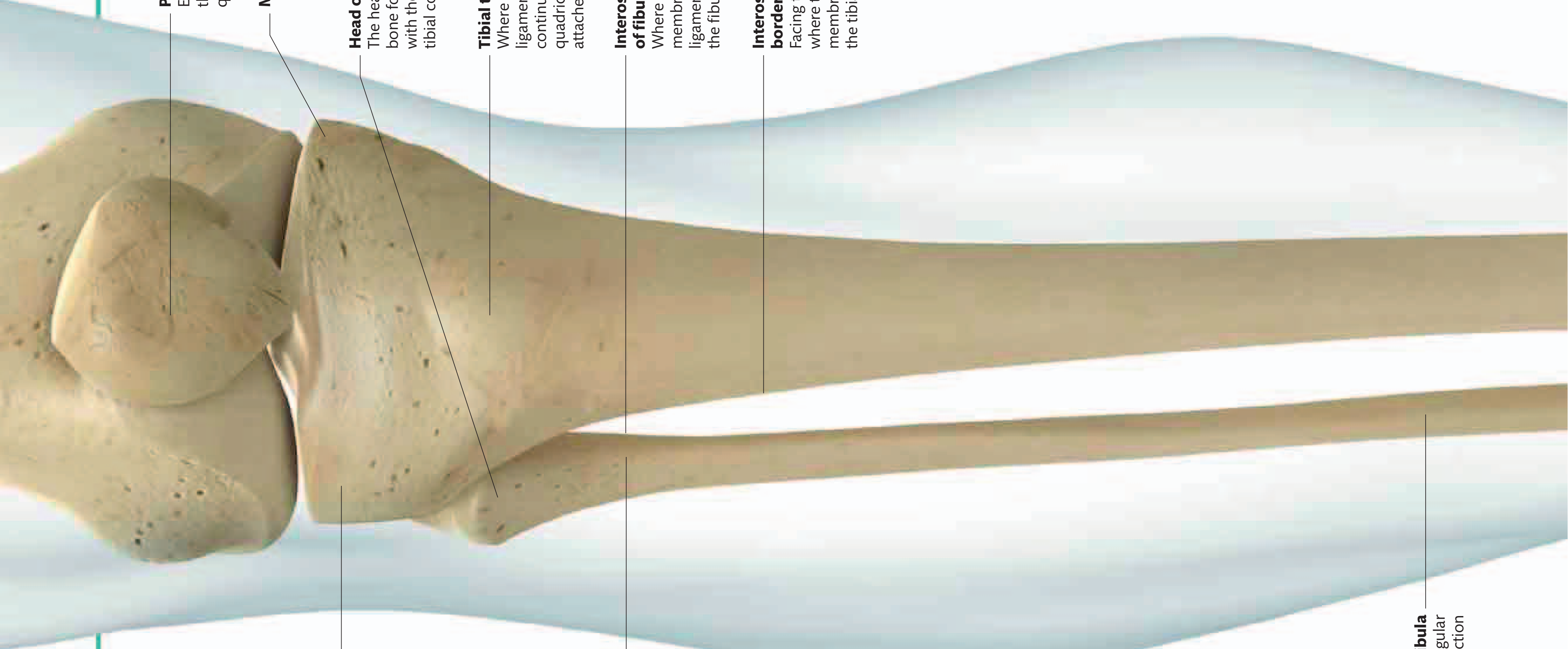

ANTERIOR (FRONT)

Patella
Embedded in the tendon of the quadriceps femoris

Medial condyle of tibia

Head of fibula
The head of this slender bone forms a synovial joint with the side of the lateral tibial condyle

Tibial tuberosity
Where the patellar ligament (the continuation of quadriceps tendon) attaches to the tibia

Interosseous border of fibula
Where the interosseous membrane—a sheetlike ligament—attaches to the fibula

Interosseous border of tibia
Facing the fibula, this is where the interosseous membrane attaches to the tibia

Lateral condyle of tibia

Neck of fibula

Shaft of fibula
Roughly triangular in cross section

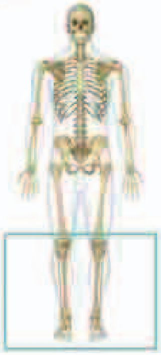

The tibia is the main weight-bearing bone of the lower leg. The fibula, which attaches to the tibia below the knee joint, provides extra areas for the attachment of muscles in the shin and calf and also forms part of the ankle joint. The foot comprises the tarsal bones, metatarsals, and phalanges. The arrangement of these bones is very similar to that of the carpals, metacarpals, and phalanges in the hand. In fact, each limb can be seen to be constructed to a common plan, with a limb girdle providing attachment to the thorax or spine, a single long bone in the first segment, two long bones in the second, a collection of small bones (at the wrist or ankle), and a fan of long, slender bones forming fingers or toes.

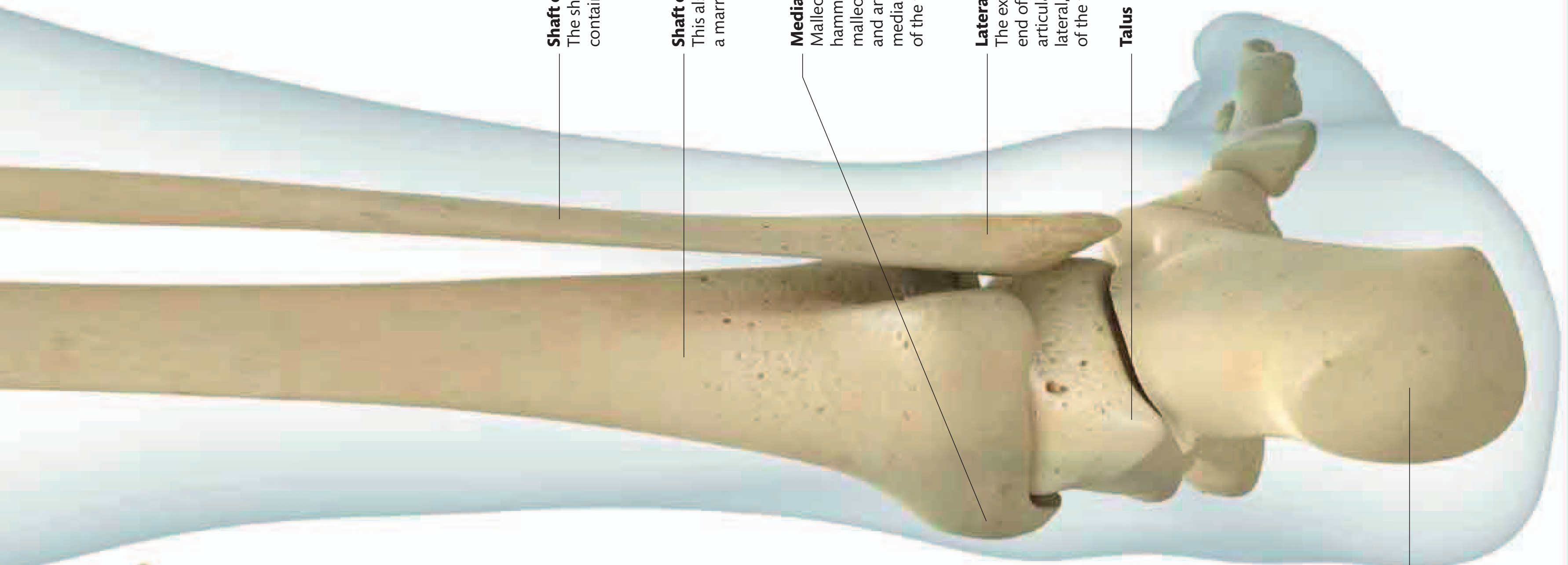

Shaft of fibula The shaft of the fibula contains a marrow cavity

Shaft of tibia This also contains a marrow cavity

Medial malleolus Malleolus means small hammer in Latin; the medial malleolus is part of the tibia, and articulates with the medial, or inner, surface of the talus

Lateral malleolus The expanded lower end of the fibula, articulating with the lateral, or outer, side of the talus

Talus

Calcaneus Meaning heel bone in Latin, this is the largest tarsal bone, projecting posteriorly to form a lever to which the Achilles tendon attaches

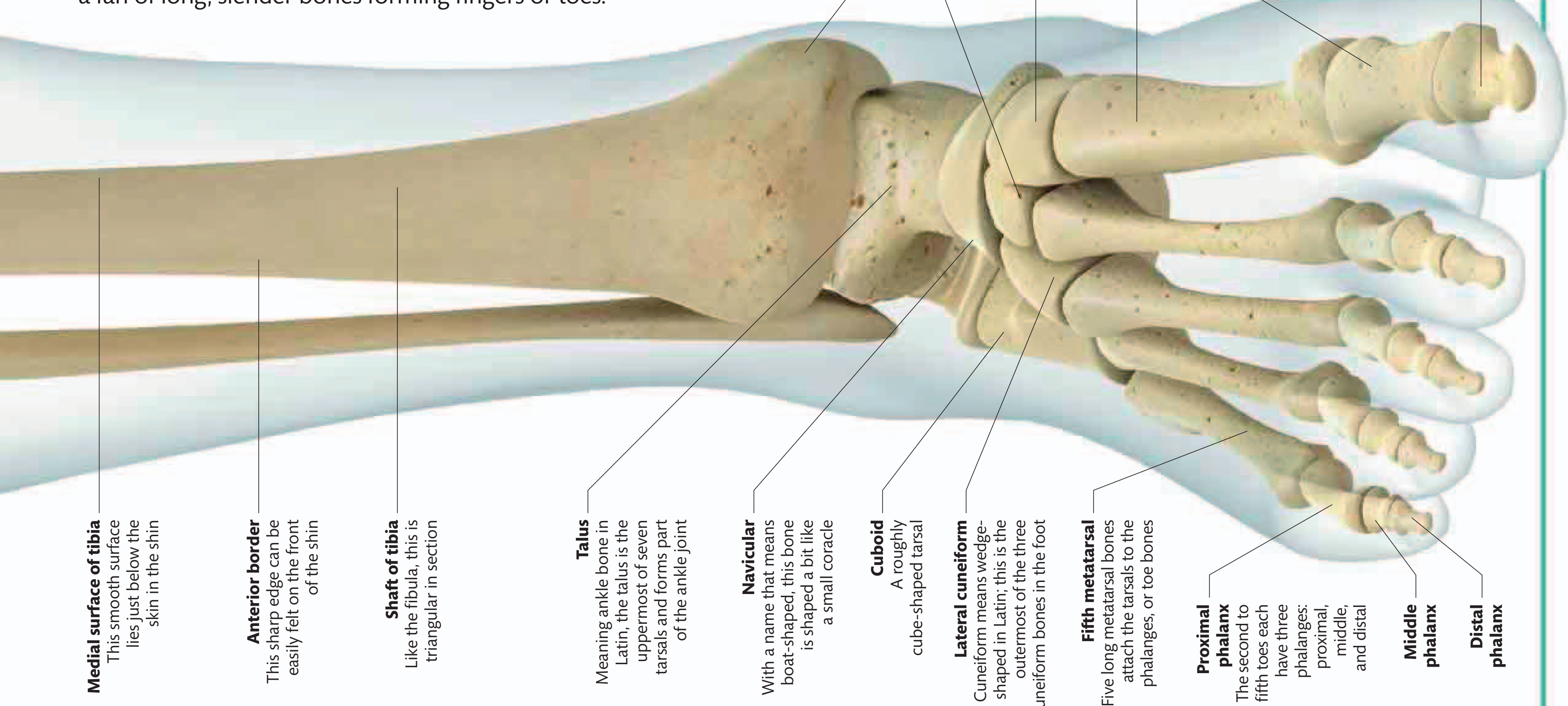

Medial malleolus

Intermediate cuneiform

Medial cuneiform

First metatarsal

Proximal phalanx Phalanx comes from a Greek word for a line of infantry, and it refers to both the finger and toe bones; the big toe has just two phalanges: proximal and distal

Distal phalanx

Medial surface of tibia This smooth surface lies just below the skin in the shin

Anterior border This sharp edge can be easily felt on the front of the shin

Shaft of tibia Like the fibula, this is triangular in section

Talus Meaning ankle bone in Latin, the talus is the uppermost of seven tarsals and forms part of the ankle joint

Navicular With a name that means boat-shaped, this bone is shaped a bit like a small coracle

Cuboid A roughly cube-shaped tarsal

Lateral cuneiform Cuneiform means wedge-shaped in Latin; this is the outermost of the three cuneiform bones in the foot

Fifth metatarsal Five long metatarsal bones attach the tarsals to the phalanges, or toe bones

Proximal phalanx The second to fifth toes each have three phalanges: proximal, middle, and distal

Middle phalanx

Distal phalanx

LATERAL (OUTSIDE)

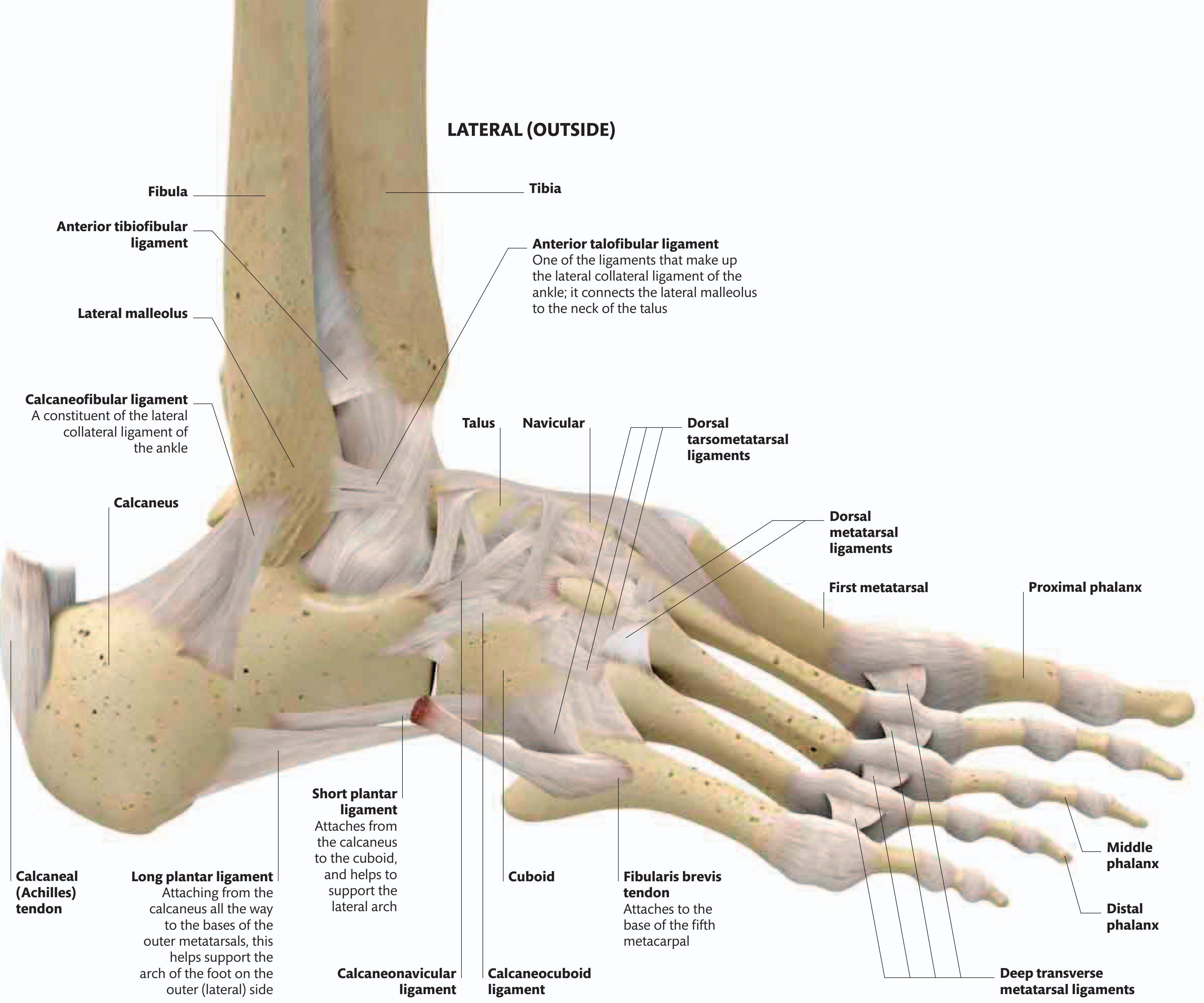
Fibula
Tibia
Anterior tibiofibular ligament
Lateral malleolus
Calcaneus
Talus
Navicular
Dorsal tarsometatarsal ligaments
Dorsal metatarsal ligaments
First metatarsal
Proximal phalanx
Middle phalanx
Distal phalanx
Calcaneal (Achilles) tendon
Cuboid
Calcaneonavicular ligament
Calcaneocuboid ligament
Deep transverse metatarsal ligaments

Anterior talofibular ligament One of the ligaments that make up the lateral collateral ligament of the ankle; it connects the lateral malleolus to the neck of the talus

Calcaneofibular ligament A constituent of the lateral collateral ligament of the ankle

Long plantar ligament Attaching from the calcaneus all the way to the bases of the outer metatarsals, this helps support the arch of the foot on the outer (lateral) side

Short plantar ligament Attaches from the calcaneus to the cuboid, and helps to support the lateral arch

Fibularis brevis tendon Attaches to the base of the fifth metacarpal

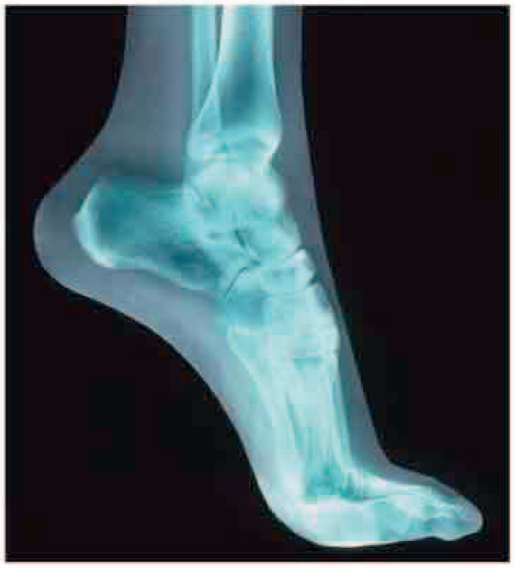

X-ray on tiptoe
This X-ray shows the foot in action. The calf muscles are pulling up on the lever of the calcaneus to flex the ankle down (plantarflex), while the metatarsophalangeal joints are extended.

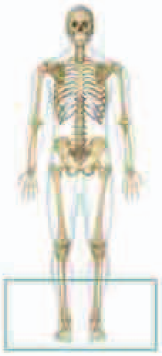

FOOT AND ANKLE

The ankle joint is a simple hinge joint. The lower ends of the tibia and fibula are firmly bound together by ligaments, forming a strong fibrous joint, and making a wrench shape that neatly sits around the nut of the talus. The joint is stabilized by strong collateral ligaments on either side. The talus forms synovial joints (see p.61) with the calcaneus beneath it, and the navicular bone in front of it. Level with the joint between the talus and the navicular is a joint between the calcaneus and the cuboid. These joints together allow the foot to be angled inward or outward—these movements are called inversion and eversion respectively. The skeleton of the foot is a sprung structure, with the bones forming arches, held together by ligaments and also supported by tendons.

POSTERIOR (BACK)

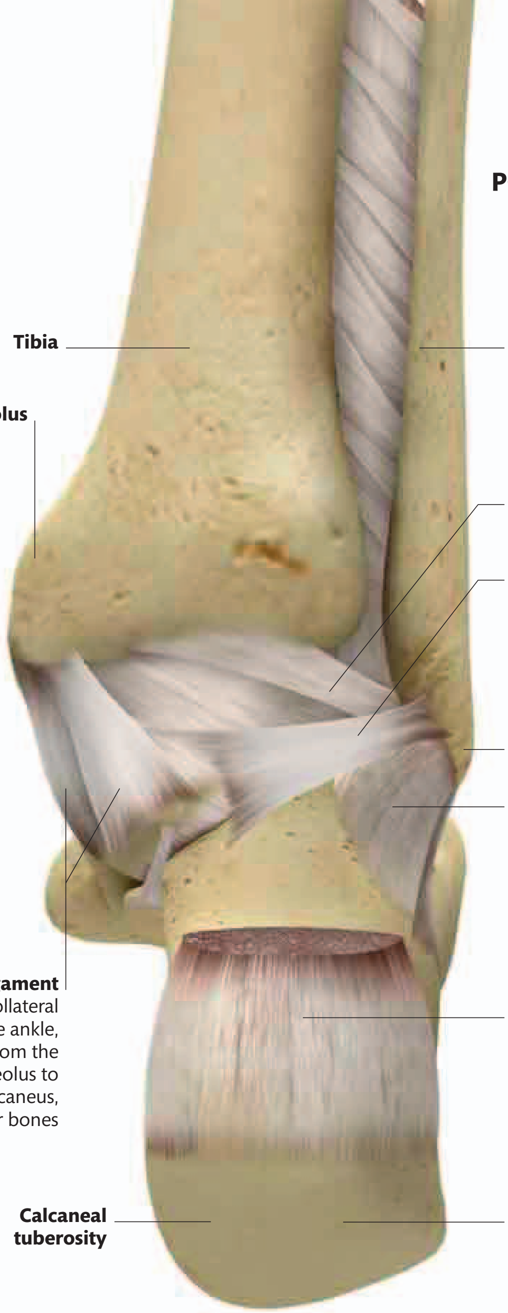
Tibia
Fibula
Medial malleolus
Posterior tibiofibular ligament
Posterior talofibular ligament
Part of the lateral collateral ligament of the ankle
Lateral malleolus
Calcaneofibular ligament
Deltoid ligament
The medial collateral ligament of the ankle, attaching from the medial malleolus to the talus, calcaneus, and navicular bones
Calcaneal (Achilles) tendon
Calcaneal tuberosity
Calcaneus

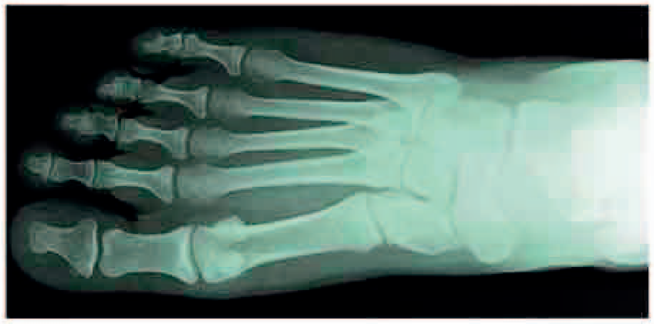

Top view of the foot bones

This is a dorsal-plantar X-ray of the foot, showing the bones as if you were looking down at your right foot. The two small bones near the head of the first metatarsal are sesamoid bones, embedded in the tendons of the short muscles operating the big toe.

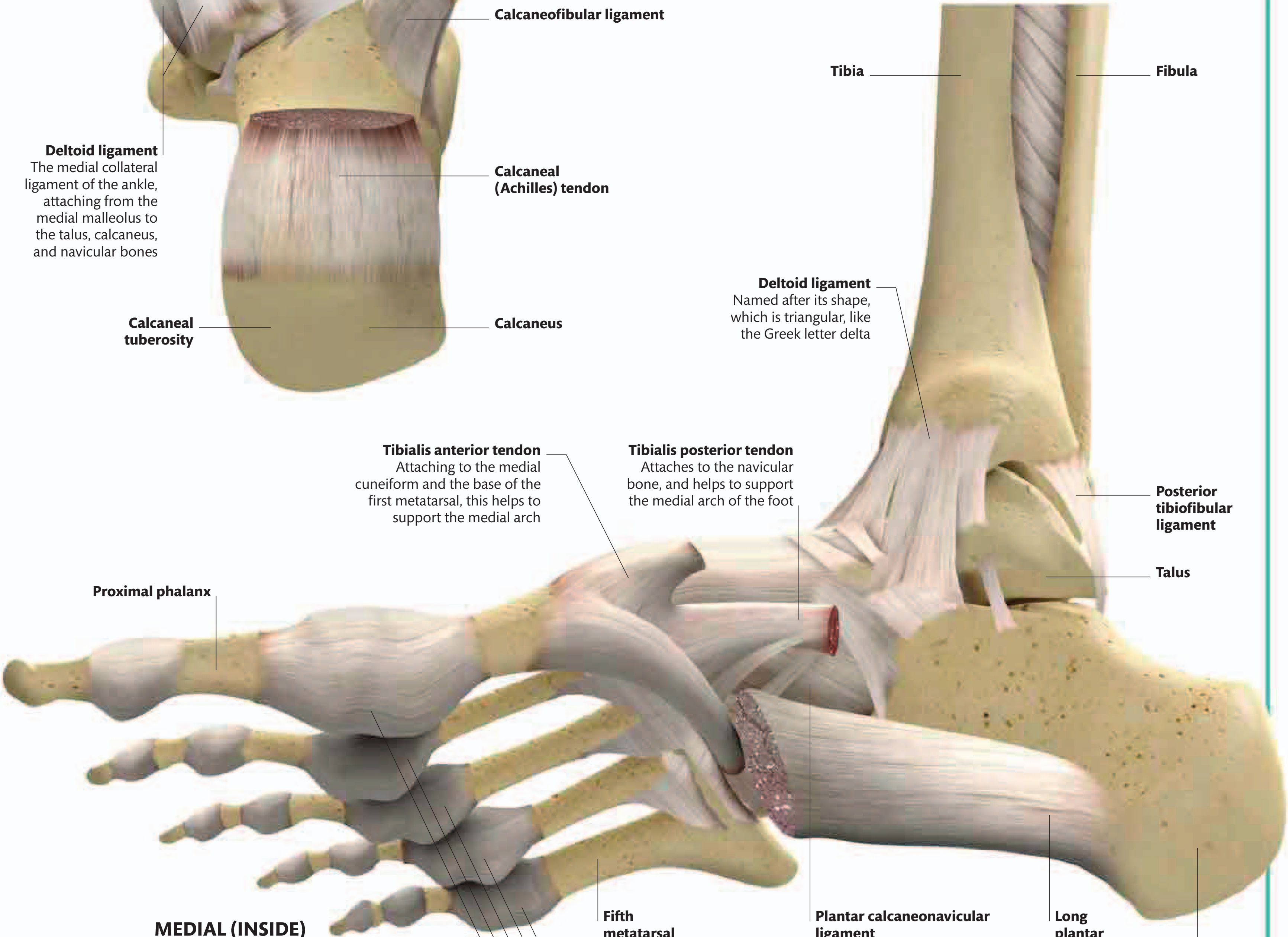
Tibia
Fibula
Deltoid ligament
Named after its shape, which is triangular, like the Greek letter delta
Tibialis anterior tendon
Attaching to the medial cuneiform and the base of the first metatarsal, this helps to support the medial arch
Tibialis posterior tendon
Attaches to the navicular bone, and helps to support the medial arch of the foot
Posterior tibiofibular ligament
Talus
Proximal phalanx
Fifth metatarsal
Plantar tarsometatarsal ligaments
Plantar calcaneonavicular ligament
Also known as the "spring ligament", this is very important in supporting the head of the talus and maintaining the medial arch of the foot
Long plantar ligament
Calcaneus

MEDIAL (INSIDE)

MUSCULAR SYSTEM OVERVIEW

There are three types of muscle in the body: skeletal, smooth, and cardiac. The main role of skeletal muscles is to generate movement. A muscle's movement, or "action" is produced when it contracts. The force it generates depends on the shape of the muscle. For instance, long, thin muscles contract a lot but exert low forces. Muscles attach to the skeleton by means of tendons, aponeuroses, and connective tissue called fascia. While muscles are well supplied with blood vessels and appear reddish, tendons have a sparse vascular supply and look white. The muscles in our body are located at varied depths. The deep layer sits closest to the bone, while the superficial one lies beneath the skin.

SKELETAL MUSCLE STRUCTURE

Skeletal muscle includes familiar muscles such as biceps or quadriceps. It is built up of parallel bundles of muscle fibers, which are conglomerations of many cells. These muscles are supplied by somatic motor nerves, which are part of the peripheral nervous system and are generally under conscious control.

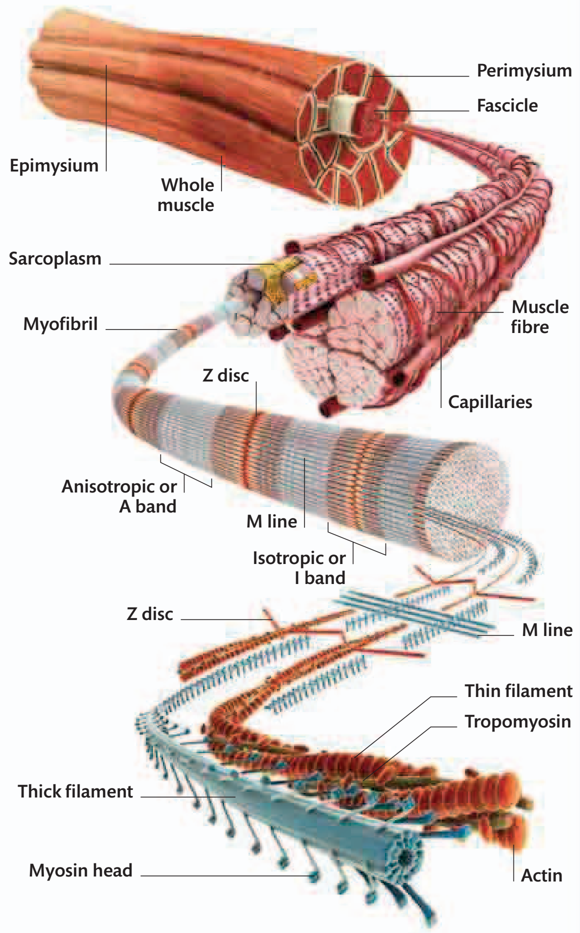

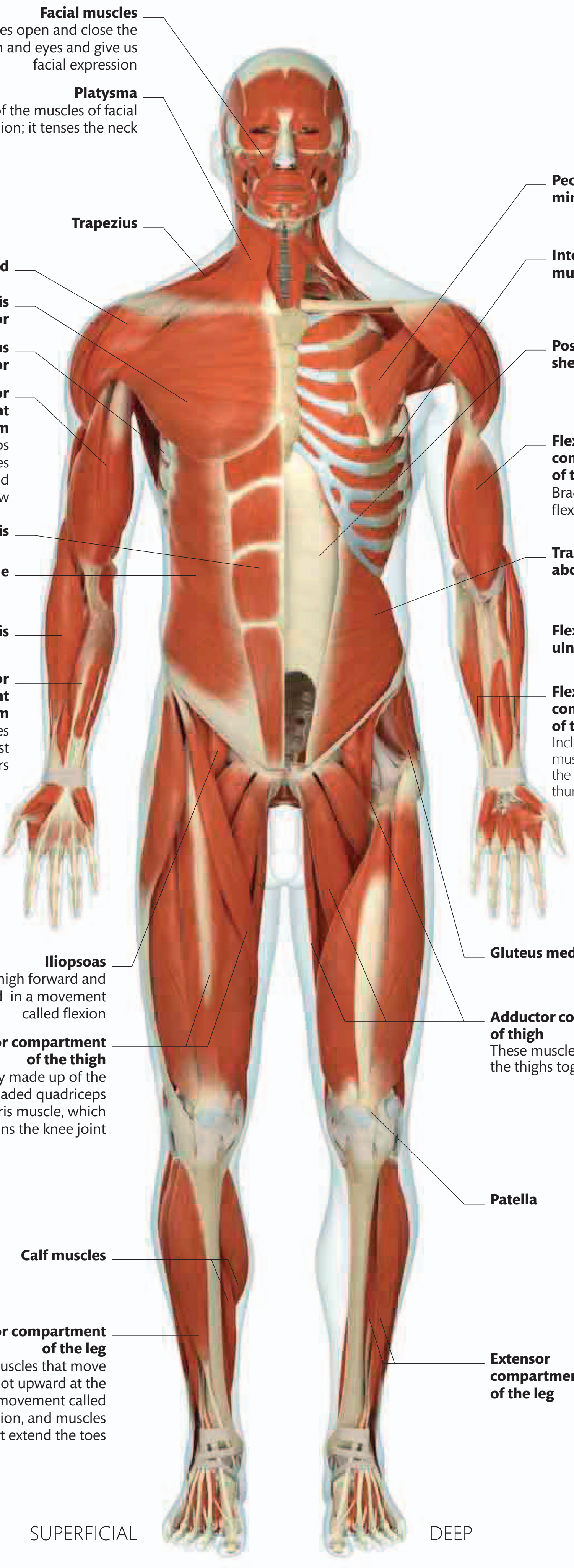

Facial muscles
Muscles open and close the mouth and eyes and give us facial expression

Platysma
Part of the muscles of facial expression; it tenses the neck

Trapezius

Deltoid

Pectoralis major

Serratus anterior

Flexor compartment of the arm
Contains biceps brachii, which flexes the shoulder and the elbow

Rectus abdominis

External oblique

Brachioradialis

Flexor compartment of the forearm
Includes muscles that flex the wrist and fingers

Iliopsoas
Bends the thigh forward and upward in a movement called flexion

Extensor compartment of the thigh
Largely made up of the four-headed quadriceps femoris muscle, which straightens the knee joint

Calf muscles

Extensor compartment of the leg
Includes muscles that move the foot upward at the ankle, a movement called dorsiflexion, and muscles that extend the toes

Pectoralis minor

Intercostal muscle

Posterior rectus sheath

Flexor compartment of the arm
Brachialis muscle flexes the elbow

Transversus abdominis

Flexor carpi ulnaris

Flexor compartment of the forearm
Includes deep muscles that flex the fingers and thumb

Gluteus medius

Adductor compartment of thigh
These muscles bring the thighs together

Patella

Extensor compartment of the leg

ANTERIOR (FRONT)

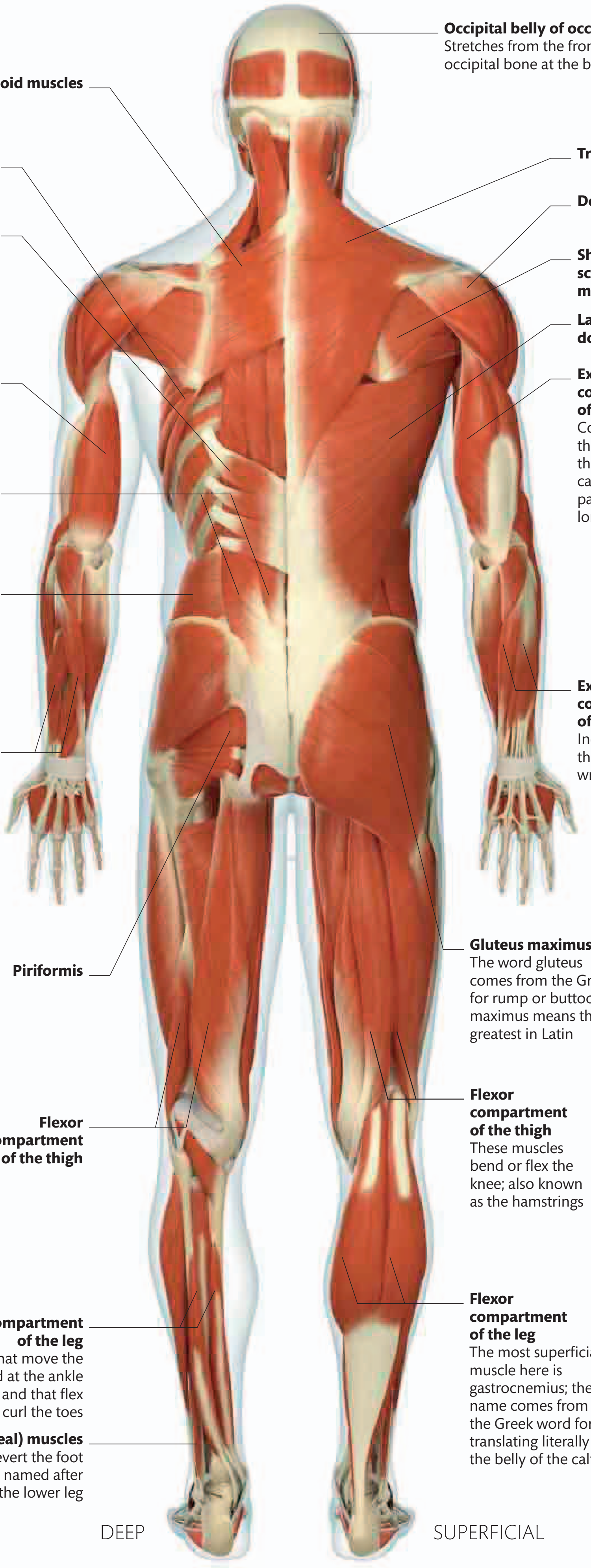

Rhomboid muscles

Serratus anterior

Serratus posterior inferior

Extensor compartment of the arm

Erector spinae
This muscle group, as its name suggests, helps to keep the spine erect

Transversus abdominis
The innermost of three sheet-like anterolateral (front-side) abdominal muscles

Extensor compartment of the forearm
Includes deep muscles that extend the thumb or move it out to the side

Piriformis

Flexor compartment of the thigh

Flexor compartment of the leg
Deep muscles that move the foot downward at the ankle (plantarflexion), and that flex or curl the toes

Fibular (peroneal) muscles
Two muscles that evert the foot (move it outward), named after the fibula bone in the lower leg

Occipital belly of occipitofrontalis
Stretches from the frontal bone to the occipital bone at the back of the skull

Trapezius

Deltoid

Short scapular muscles

Latissimus dorsi

Extensor compartment of the arm
Contains one muscle, the triceps, which means three-headed; here we can see the two superficial parts of the muscle, the long and lateral heads

Extensor compartment of the forearm
Includes muscles that extend the wrist and fingers

Gluteus maximus
The word gluteus comes from the Greek for rump or buttock; maximus means the greatest in Latin

Flexor compartment of the thigh
These muscles bend or flex the knee; also known as the hamstrings

Flexor compartment of the leg
The most superficial muscle here is gastrocnemius; the name comes from the Greek word for calf, translating literally as the belly of the calf

POSTERIOR (BACK)

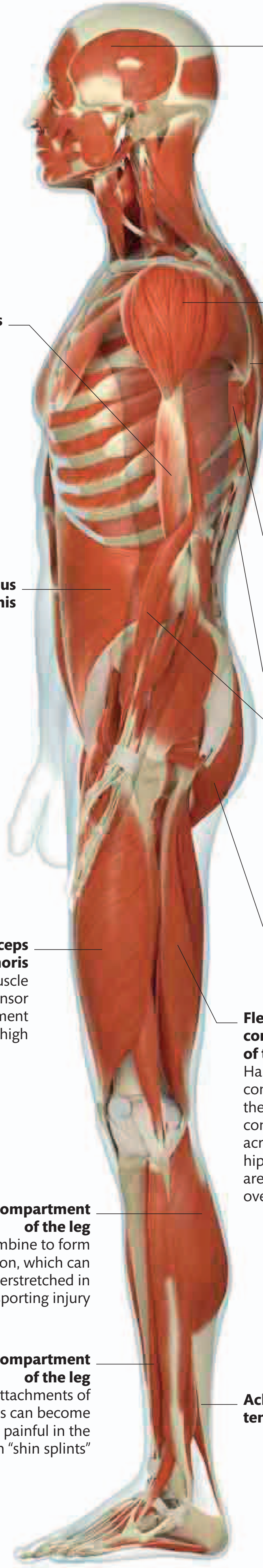

Brachialis

Transversus abdominis

Quadriceps femoris
Largest muscle in the extensor compartment of the thigh

Flexor compartment of the leg
These muscles combine to form the Achilles tendon, which can be ruptured if overstretched in a sporting injury

Extensor compartment of the leg
The bony attachments of these muscles can become inflamed and painful in the condition "shin splints"

Temporalis

Deltoid

Short scapular muscles
Problems with these muscles can lead to osteoarthritis at the shoulder joint

Serratus anterior
Anchors the scapula against the chest wall and also helps to move it

Erector spinae

Flexor compartment of the forearm
Some of these muscles originate from the medial epicondyle of the humerus

Gluteus maximus

Flexor compartment of the thigh
Hamstring injuries are common in athletes: the long muscles in this compartment stretch across two joints—the hip and the knee—and are at risk of tearing if overstretched

Achilles tendon

SIDE

HEAD AND NECK

The muscles of the face have very important functions. They open and close the apertures in our faces—our eyes, noses, and mouths. But they also play an extremely important role in communication, and this is why these muscles are often known, collectively, as “the muscles of facial expression”. These muscles are attached to bone at one end and skin at the other. It is these muscles that allow us to raise our eyebrows in surprise, frown, or knit our brows in concentration, to scrunch up our noses in distaste, to smile gently or to grin widely, and to pout. As we age, and our skin forms creases and wrinkles, these reflect the expressions we have used throughout our lives. The wrinkles and creases lie perpendicular to the direction of the underlying muscle fibers.

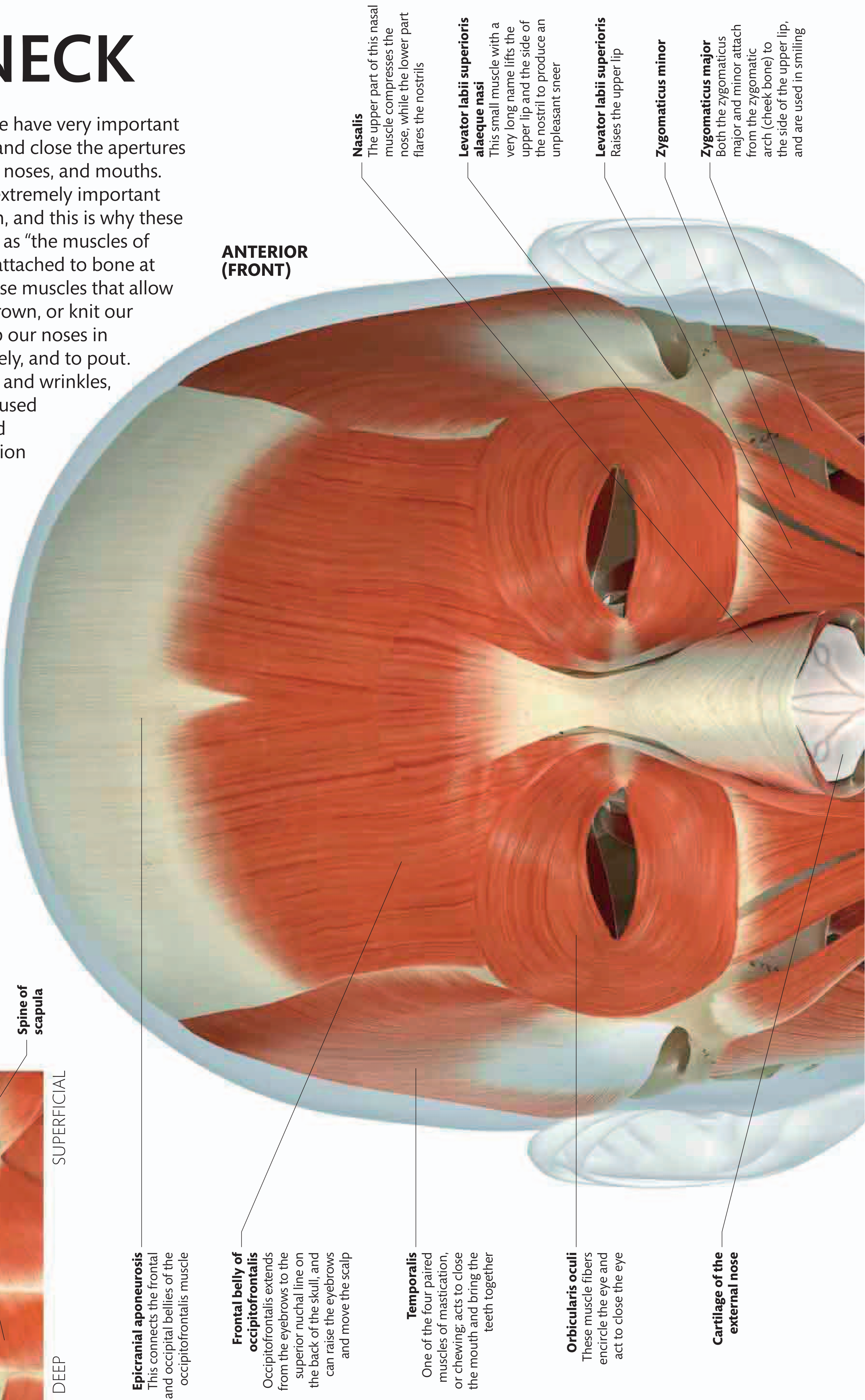

ANTERIOR (FRONT)

Nasalis The upper part of this nasal muscle compresses the nose, while the lower part flares the nostrils

Levator labii superioris alaeque nasi This small muscle with a very long name lifts the upper lip and the side of the nostril to produce an unpleasant sneer

Levator labii superioris Raises the upper lip

Zygomaticus minor

Zygomaticus major Both the zygomaticus major and minor attach from the zygomatic arch (cheek bone) to the side of the upper lip, and are used in smiling

Epicranial aponeurosis This connects the frontal and occipital bellies of the occipitofrontalis muscle

Frontal belly of occipitofrontalis Occipitofrontalis extends from the eyebrows to the superior nuchal line on the back of the skull, and can raise the eyebrows and move the scalp

Temporalis One of the four paired muscles of mastication, or chewing; acts to close the mouth and bring the teeth together

Orbicularis oculi These muscle fibers encircle the eye and act to close the eye

Cartilage of the external nose

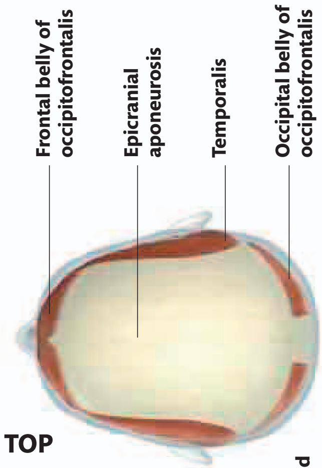

Frontal belly of occipitofrontalis
Epicranial aponeurosis
Temporalis
Occipital belly of occipitofrontalis
TOP

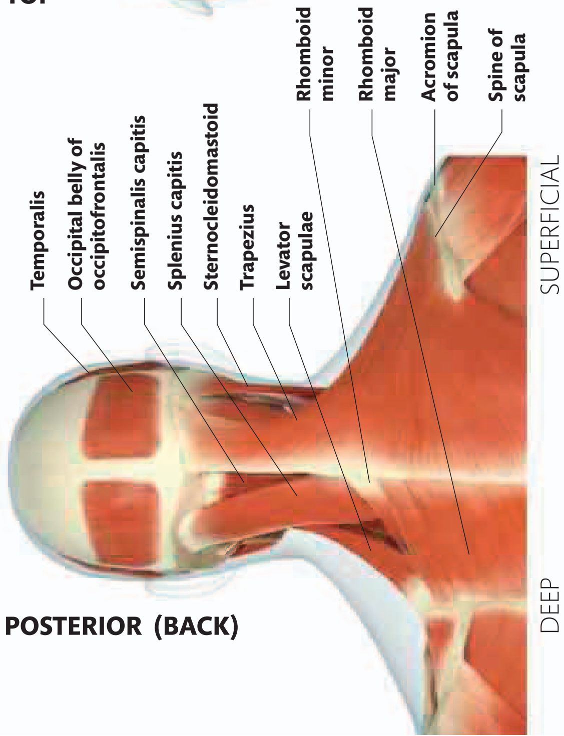

Temporalis
Occipital belly of occipitofrontalis
Semispinalis capitis
Splenius capitis
Sternocleidomastoid
Trapezius
Levator scapulae
Rhomboid minor
Rhomboid major
Acromion of scapula
Spine of scapula
SUPERFICIAL
DEEP
POSTERIOR (BACK)

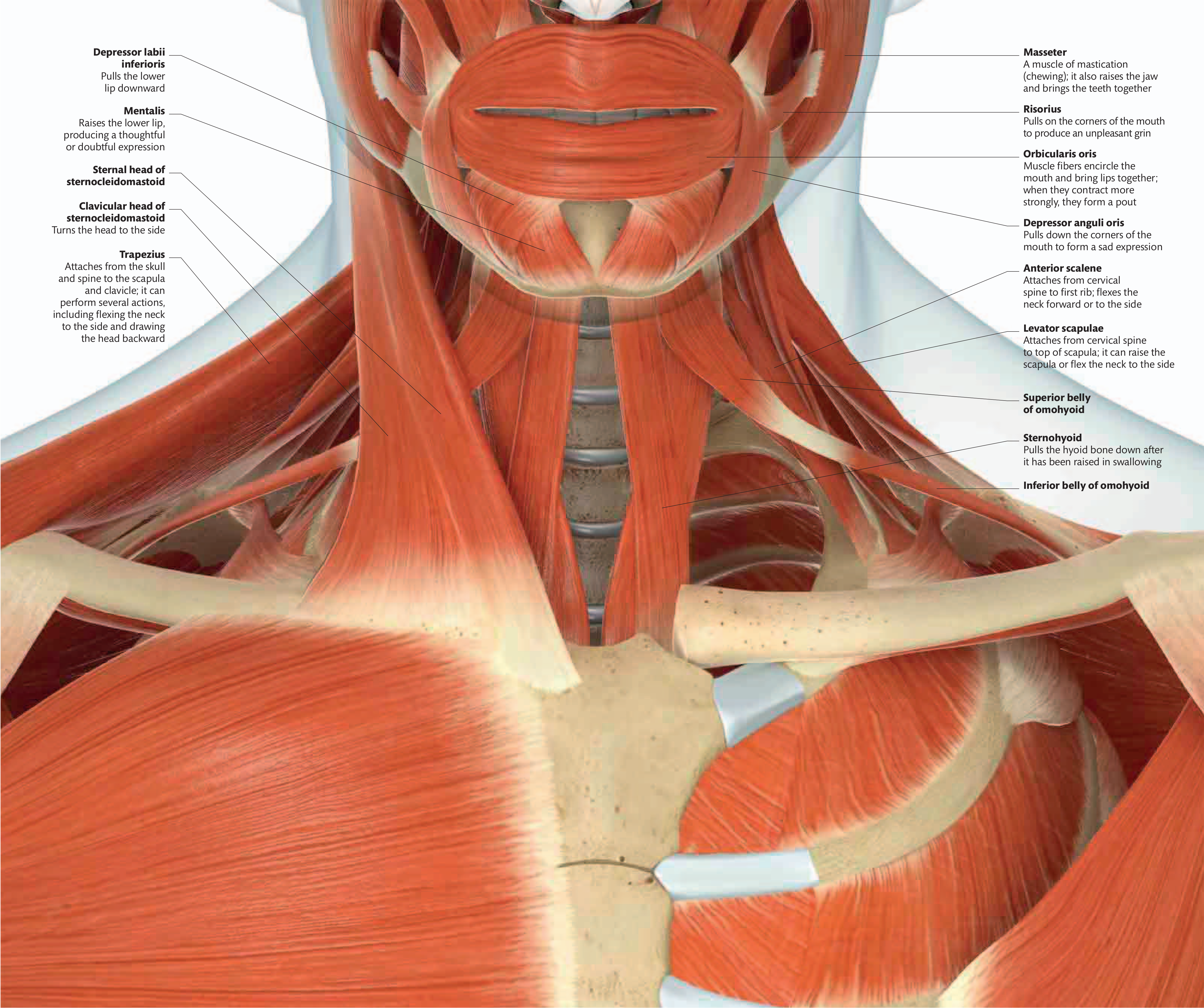

Depressor labii inferioris
Pulls the lower lip downward

Mentalis
Raises the lower lip, producing a thoughtful or doubtful expression

Sternal head of sternocleidomastoid

Clavicular head of sternocleidomastoid
Turns the head to the side

Trapezius
Attaches from the skull and spine to the scapula and clavicle; it can perform several actions, including flexing the neck to the side and drawing the head backward

Masseter
A muscle of mastication (chewing); it also raises the jaw and brings the teeth together

Risorius
Pulls on the corners of the mouth to produce an unpleasant grin

Orbicularis oris
Muscle fibers encircle the mouth and bring lips together; when they contract more strongly, they form a pout

Depressor anguli oris
Pulls down the corners of the mouth to form a sad expression

Anterior scalene
Attaches from cervical spine to first rib; flexes the neck forward or to the side

Levator scapulae
Attaches from cervical spine to top of scapula; it can raise the scapula or flex the neck to the side

Superior belly of omohyoid

Sternohyoid
Pulls the hyoid bone down after it has been raised in swallowing

Inferior belly of omohyoid

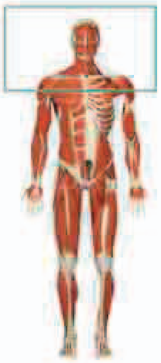

HEAD AND NECK

The muscles of mastication (chewing) attach from the skull to the mandible (jawbone), operating to open and shut the mouth, and to grind the teeth together to crush the food we eat. In this side view, we can see the two largest muscles of mastication, the temporalis and masseter muscles. Two smaller muscles attach to the inner surface of the mandible. Human jaws don't just open and close, they also move from side to side, and these four muscles act in concert to produce complex chewing movements. In this view, we can also see how the frontal bellies (fleshy central parts) of the occipitofrontalis muscle are connected to occipital bellies at the back of the head by a thin, flat tendon, or aponeurosis. This makes the entire scalp movable on the skull.

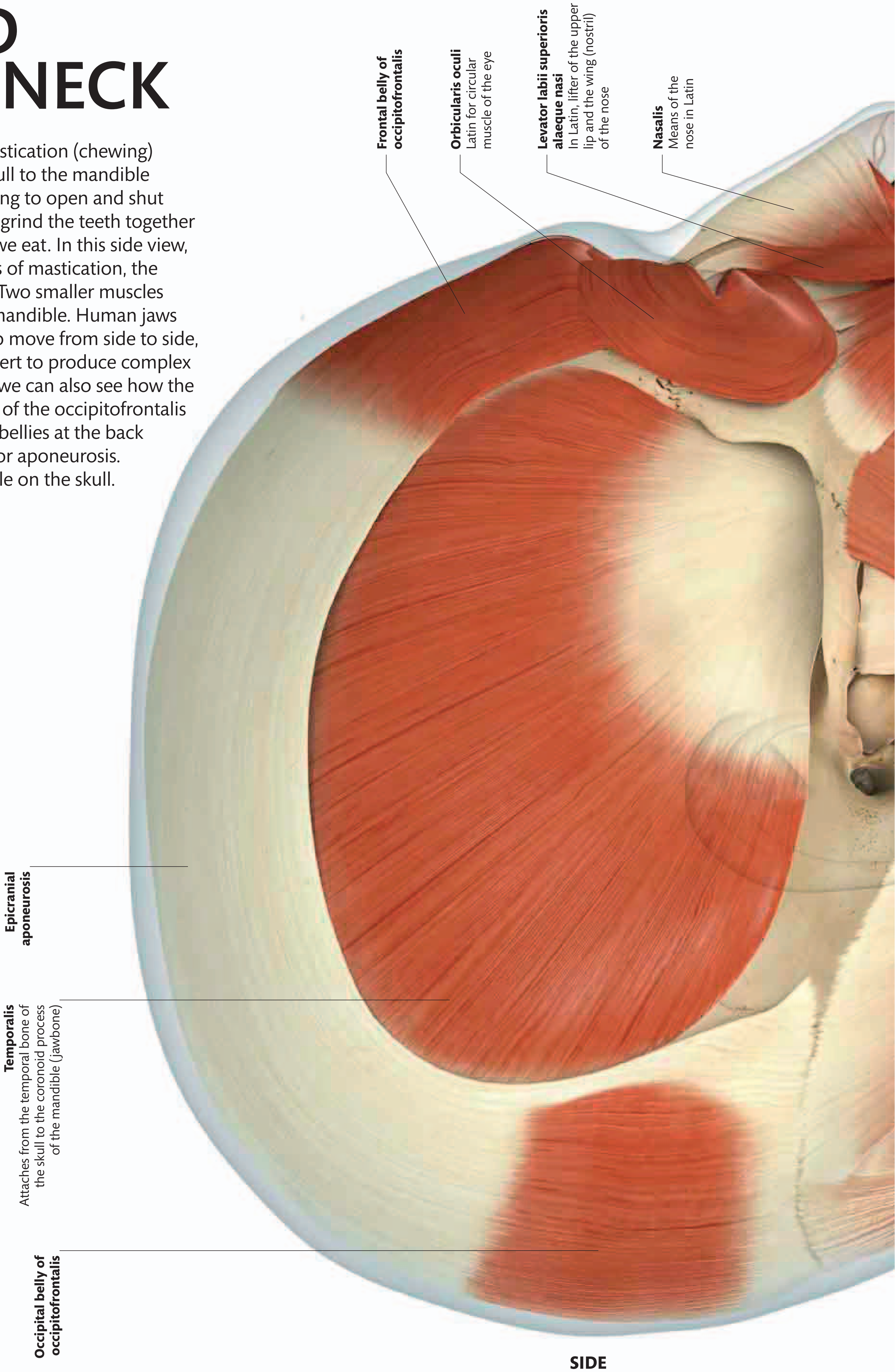
Frontal belly of occipitofrontalis
Orbicularis oculi Latin for circular muscle of the eye
Levator labii superioris alaeque nasi In Latin, lifter of the upper lip and the wing (nostril) of the nose
Nasalis Means of the nose in Latin
Epicranial aponeurosis
Temporalis Attaches from the temporal bone of the skull to the coronoid process of the mandible (jawbone)
Occipital belly of occipitofrontalis

SIDE

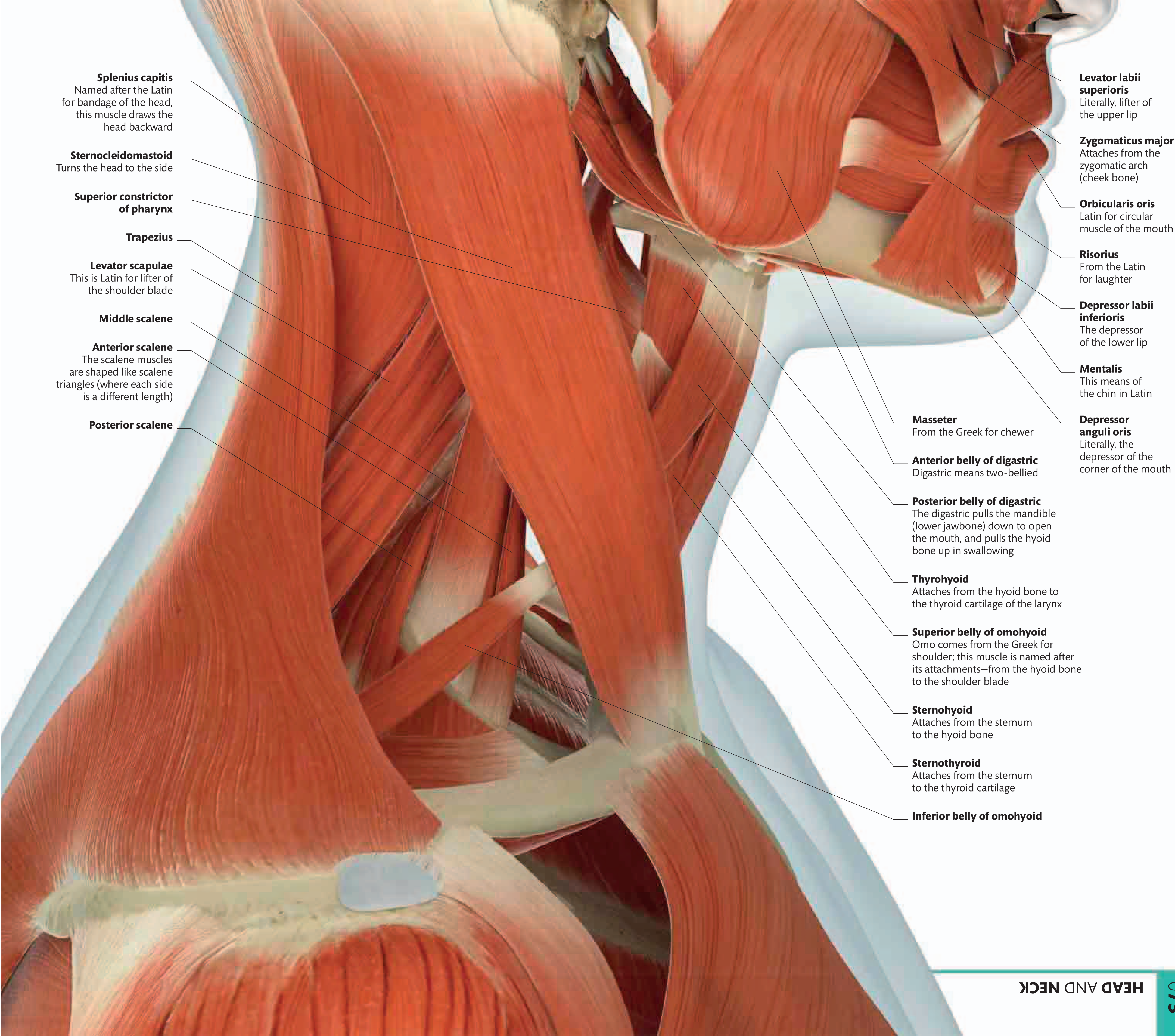
Splenius capitis
Named after the Latin for bandage of the head, this muscle draws the head backward
Sternocleidomastoid
Turns the head to the side
Superior constrictor of pharynx
Trapezius
Levator scapulae
This is Latin for lifter of the shoulder blade
Middle scalene
Anterior scalene
The scalene muscles are shaped like scalene triangles (where each side is a different length)
Posterior scalene
Levator labii superioris
Literally, lifter of the upper lip
Zygomaticus major
Attaches from the zygomatic arch (cheek bone)
Orbicularis oris
Latin for circular muscle of the mouth
Risorius
From the Latin for laughter
Depressor labii inferioris
The depressor of the lower lip
Mentalis
This means of the chin in Latin
Depressor anguli oris
Literally, the depressor of the corner of the mouth
Masseter
From the Greek for chewer
Anterior belly of digastric
Digastric means two-bellied
Posterior belly of digastric
The digastric pulls the mandible (lower jawbone) down to open the mouth, and pulls the hyoid bone up in swallowing
Thyrohyoid
Attaches from the hyoid bone to the thyroid cartilage of the larynx
Superior belly of omohyoid
Omo comes from the Greek for shoulder; this muscle is named after its attachments—from the hyoid bone to the shoulder blade
Sternohyoid
Attaches from the sternum to the hyoid bone
Sternothyroid
Attaches from the sternum to the thyroid cartilage
Inferior belly of omohyoid

SAGITTAL SECTION

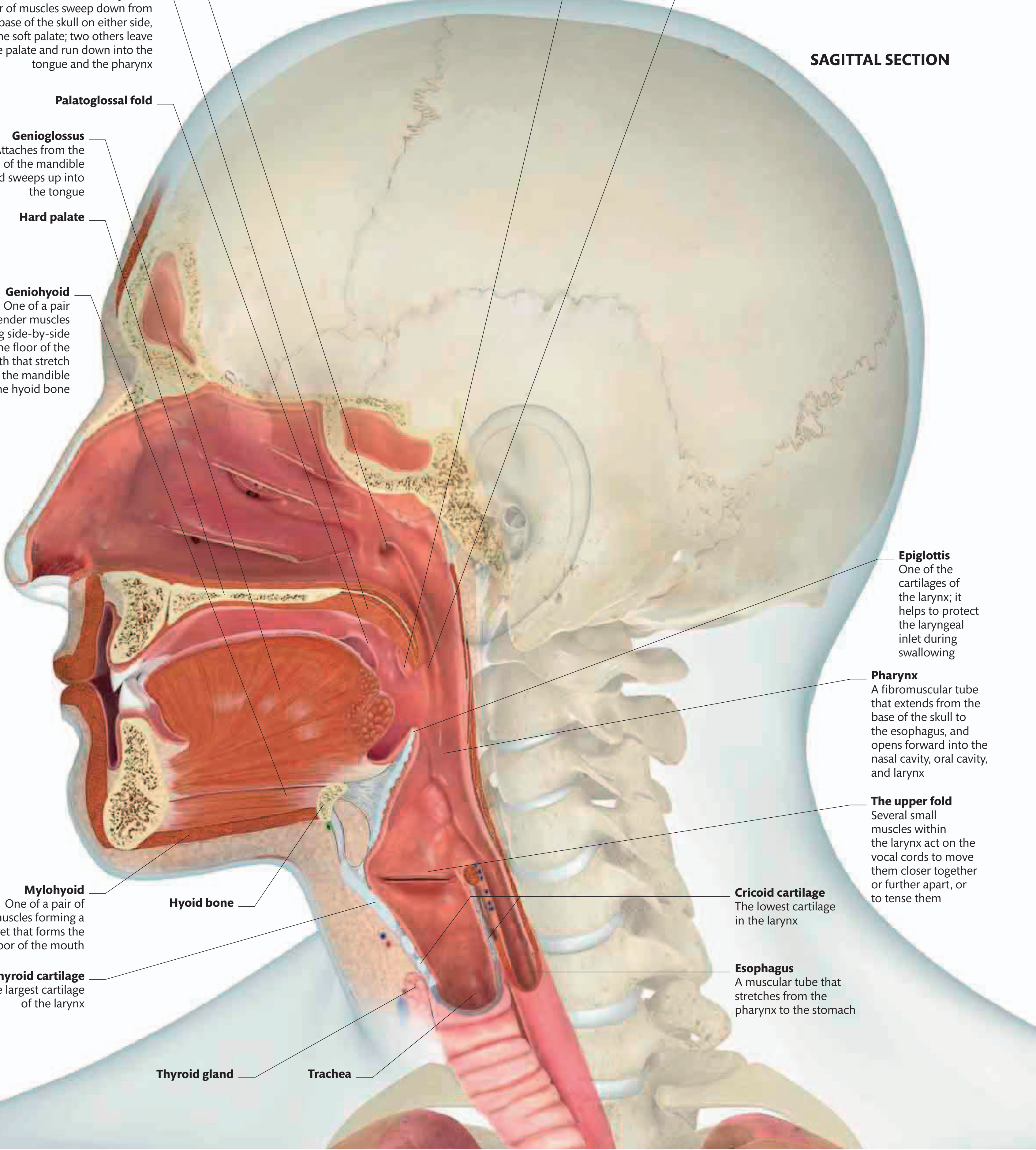

Pharyngotympanic tube
Also known as the auditory or Eustachian tube; a slender muscle called salpingopharyngeus descends from its cartilage to contribute to the side wall of the pharynx

Soft palate
A pair of muscles sweep down from the base of the skull on either side, into the soft palate; two others leave the palate and run down into the tongue and the pharynx

Palatoglossal fold

Genioglossus
Attaches from the inside of the mandible and sweeps up into the tongue

Hard palate

Geniohyoid
One of a pair of slender muscles lying side-by-side in the floor of the mouth that stretch from the mandible to the hyoid bone

Mylohyoid
One of a pair of muscles forming a sheet that forms the floor of the mouth

Hyoid bone

Thyroid cartilage
The largest cartilage of the larynx

Thyroid gland

Trachea

Palatine tonsil

Palatopharyngeal fold

Epiglottis
One of the cartilages of the larynx; it helps to protect the laryngeal inlet during swallowing

Pharynx
A fibromuscular tube that extends from the base of the skull to the esophagus, and opens forward into the nasal cavity, oral cavity, and larynx

The upper fold
Several small muscles within the larynx act on the vocal cords to move them closer together or further apart, or to tense them

Cricoid cartilage
The lowest cartilage in the larynx

Esophagus
A muscular tube that stretches from the pharynx to the stomach

HEAD AND NECK

In the section through the head (opposite), we see the soft palate, tongue, pharynx, and larynx, all of which contain muscles. The soft palate comprises five pairs of muscles. When relaxed, it hangs down at the back of the mouth but, during swallowing, it thickens and is drawn upward to block off the airway. The tongue is a great mass of muscle, covered in mucosa. Some of its muscles arise from the hyoid bone and the mandible, and anchor it to these bones and move it around. Other muscle fibers are entirely within the tongue and change its shape. The pharyngeal muscles are important in swallowing, and the laryngeal muscles control the vocal cords. The muscles that move the eye can be seen on p.122.

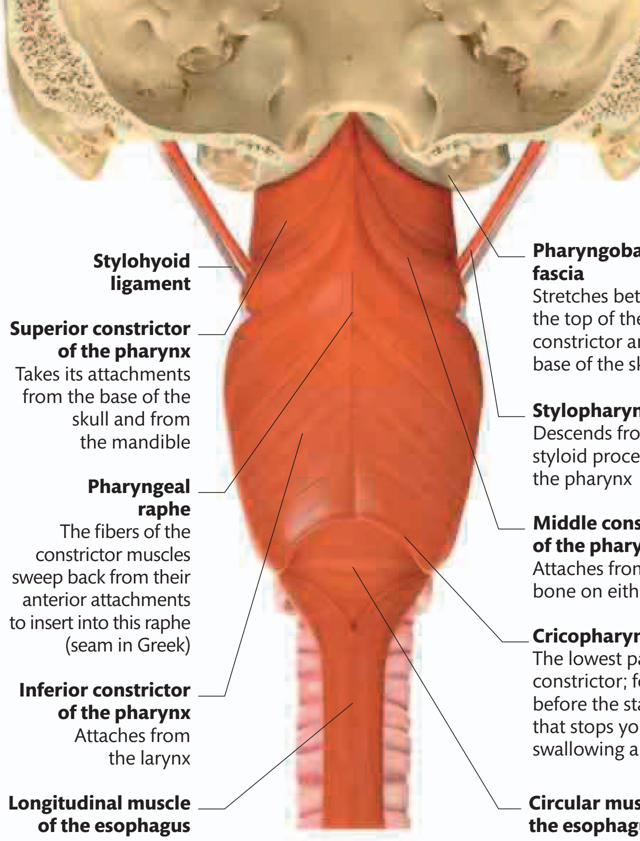

Stylohyoid ligament

Superior constrictor of the pharynx
Takes its attachments from the base of the skull and from the mandible

Pharyngeal raphe
The fibers of the constrictor muscles sweep back from their anterior attachments to insert into this raphe (seam in Greek)

Inferior constrictor of the pharynx
Attaches from the larynx

Longitudinal muscle of the esophagus

Pharyngobasilar fascia
Stretches between the top of the superior constrictor and the base of the skull

Stylopharyngeus
Descends from the styloid process into the pharynx

Middle constrictor of the pharynx
Attaches from the hyoid bone on either side

Cricopharyngeus
The lowest part of the inferior constrictor; forms a sphincter just before the start of the esophagus that stops you from continually swallowing air as you breathe

Circular muscle of the esophagus

PHARYNX POSTERIOR (BACK)

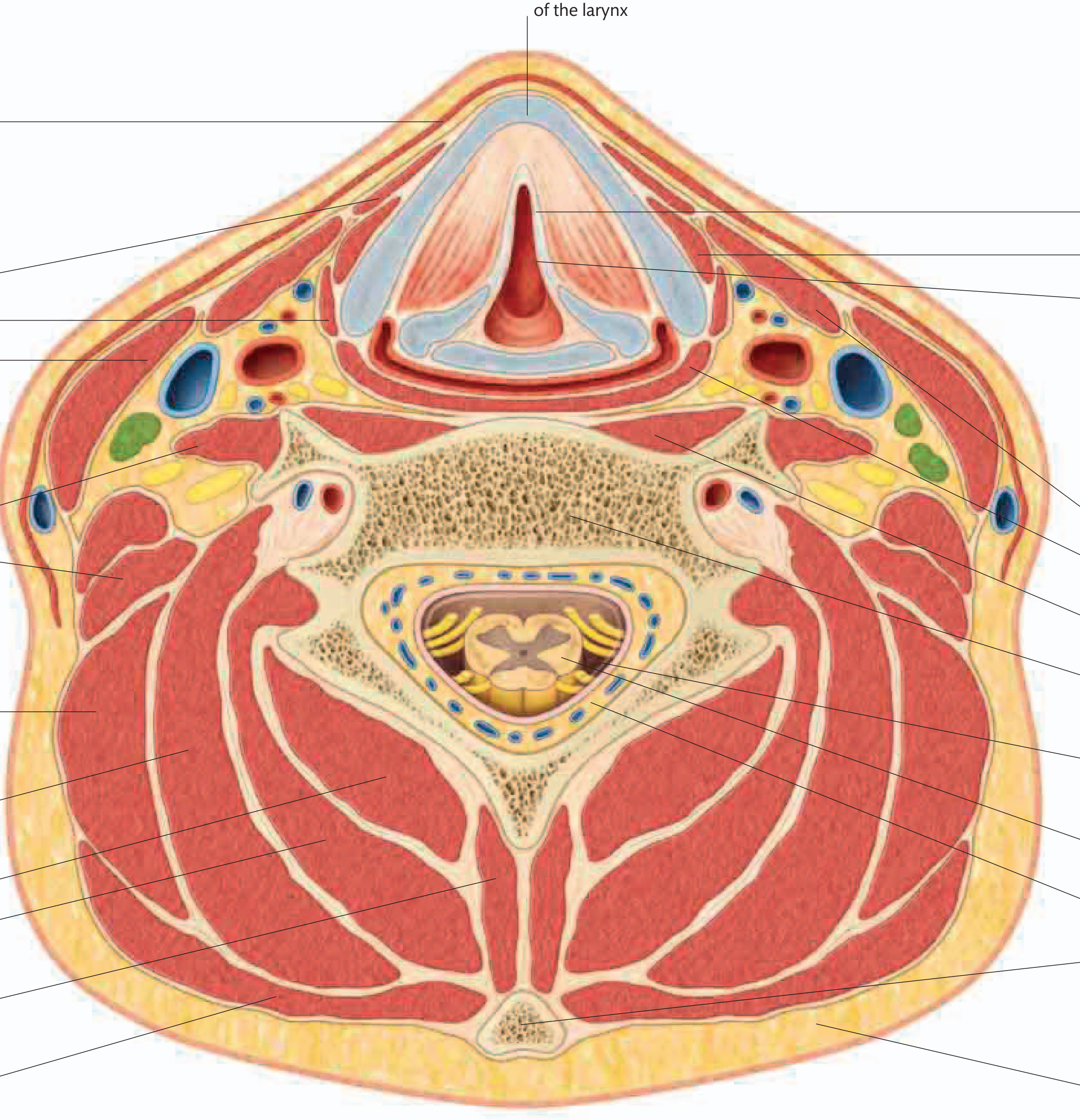

Thyroid cartilage
The largest cartilage of the larynx

Platysma
This very thin sheet of muscle lies in the superficial fascia over the front of the neck, and produces a grimace

Sternohyoid

Sternothyroid

Sternocleidomastoid
Connecting the mastoid process of the skull above to the clavicle and sternum below, this muscle turns the head to the side

Anterior scalene

Middle scalene
The scalene muscles attach from the cervical spine down to the upper ribs; they flex the neck forward or to the side

Levator scapulae
Attaches from the cervical spine to the scapula

Splenius capitis
Acts to draw the head backward

Multifidus

Semispinalis capitis
Extends or tips back the head on the neck

Semispinalis cervicis
Extends or arches the neck backward

Trapezius

Vocal cord

Thyrohyoid

Glottis
The gap between the vocal cords; muscles of the larynx act to change the position and tension of the vocal cords

Omohyoid

Inferior constrictor

Longus colli

Body of cervical vertebra

Subarachnoid space

Spinal cord

Epidural space

Spinous process of cervical vertebra

Subcutaneous fat

TRANSVERSE SECTION OF THE NECK AT THE VOCAL CORDS

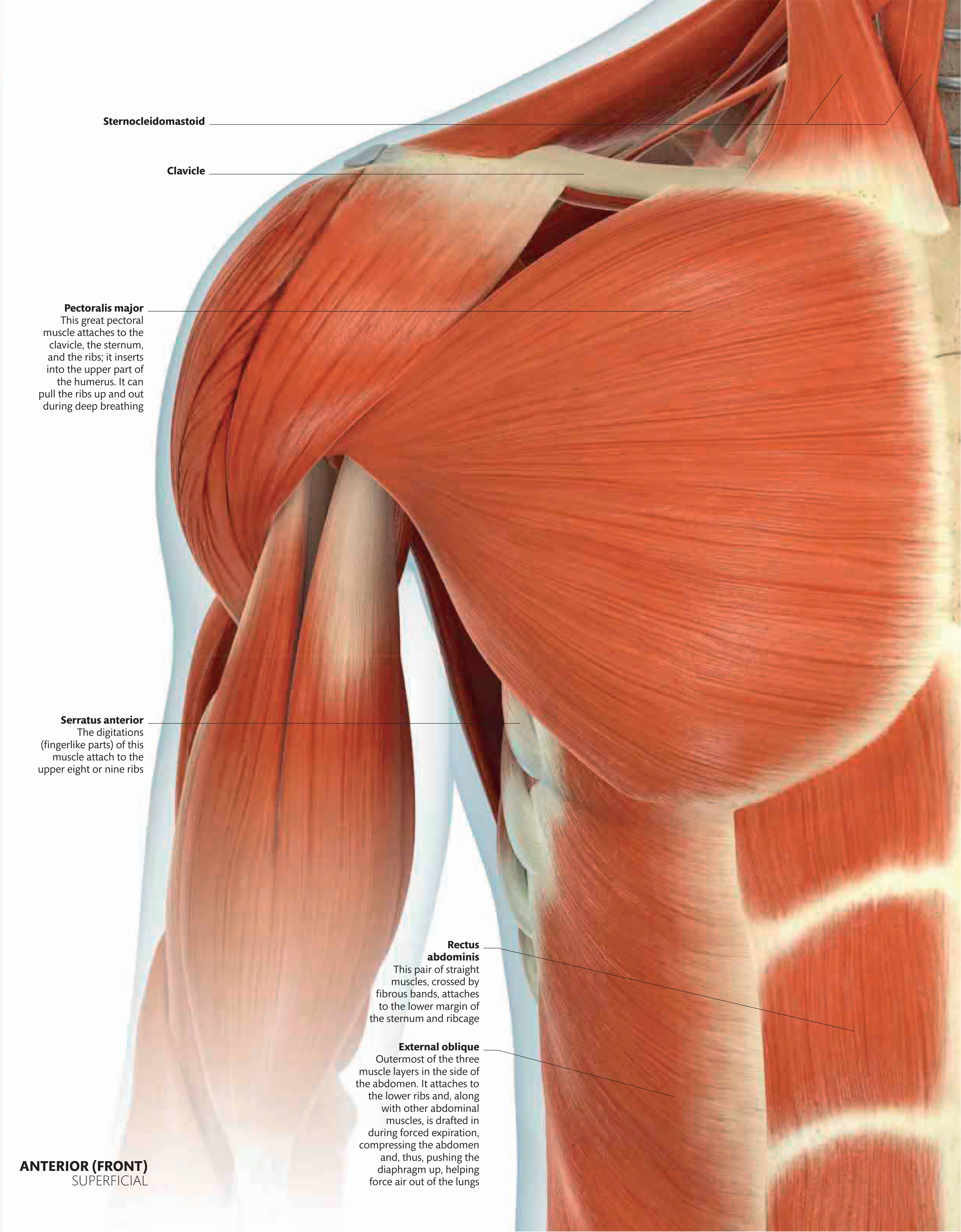

Sternocleidomastoid

Clavicle

Pectoralis major
This great pectoral muscle attaches to the clavicle, the sternum, and the ribs; it inserts into the upper part of the humerus. It can pull the ribs up and out during deep breathing

Serratus anterior
The digitations (fingerlike parts) of this muscle attach to the upper eight or nine ribs

Rectus abdominis
This pair of straight muscles, crossed by fibrous bands, attaches to the lower margin of the sternum and ribcage

External oblique
Outermost of the three muscle layers in the side of the abdomen. It attaches to the lower ribs and, along with other abdominal muscles, is drafted in during forced expiration, compressing the abdomen and, thus, pushing the diaphragm up, helping force air out of the lungs

ANTERIOR (FRONT)
SUPERFICIAL

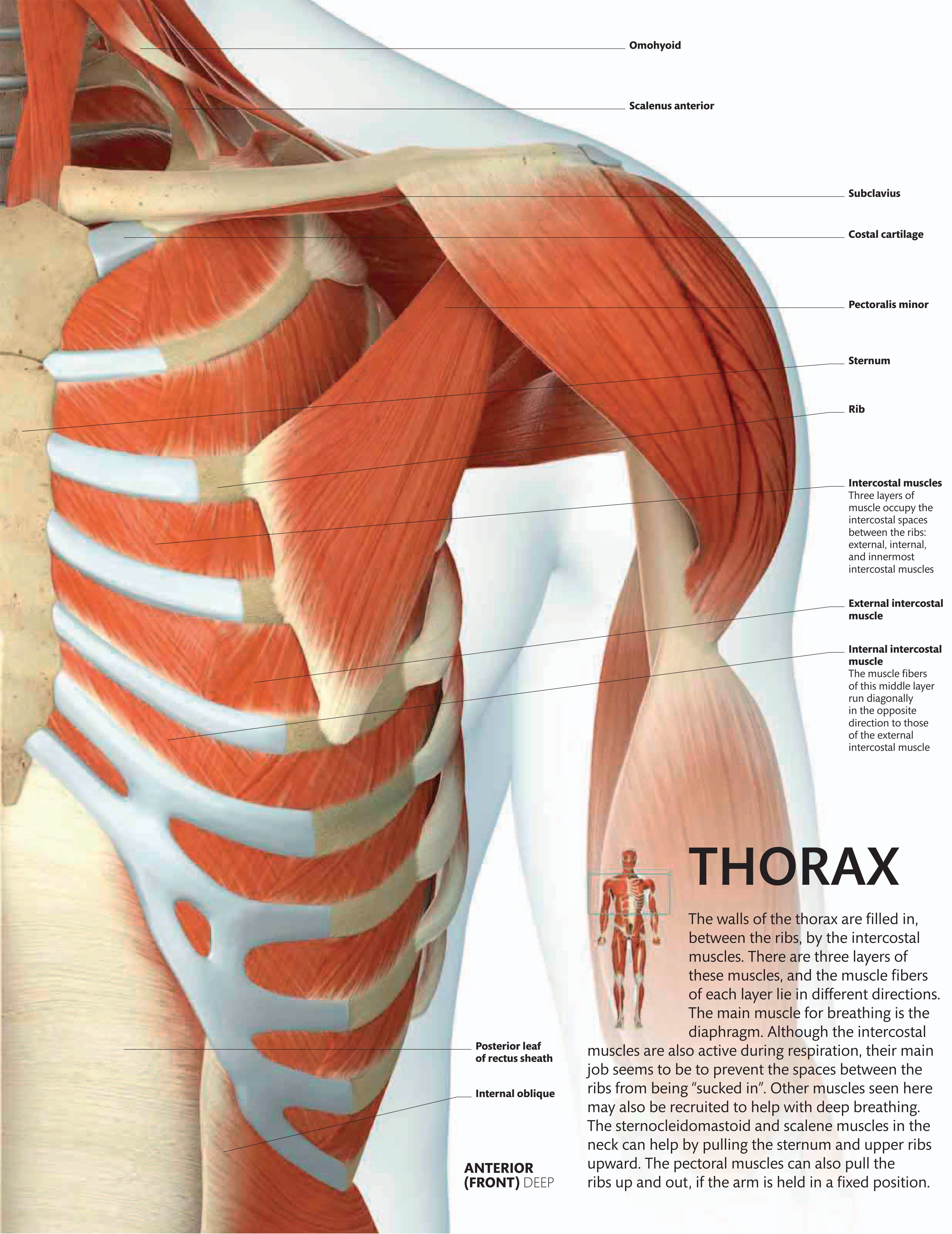
Omohyoid
Scalenus anterior
Subclavius
Costal cartilage
Pectoralis minor
Sternum
Rib
Intercostal muscles
External intercostal muscle
Internal intercostal muscle
Posterior leaf of rectus sheath
Internal oblique

Intercostal muscles Three layers of muscle occupy the intercostal spaces between the ribs: external, internal, and innermost intercostal muscles

Internal intercostal muscle The muscle fibers of this middle layer run diagonally in the opposite direction to those of the external intercostal muscle

ANTERIOR (FRONT) DEEP

THORAX

The walls of the thorax are filled in, between the ribs, by the intercostal muscles. There are three layers of these muscles, and the muscle fibers of each layer lie in different directions. The main muscle for breathing is the diaphragm. Although the intercostal muscles are also active during respiration, their main job seems to be to prevent the spaces between the ribs from being "sucked in". Other muscles seen here may also be recruited to help with deep breathing. The sternocleidomastoid and scalene muscles in the neck can help by pulling the sternum and upper ribs upward. The pectoral muscles can also pull the ribs up and out, if the arm is held in a fixed position.

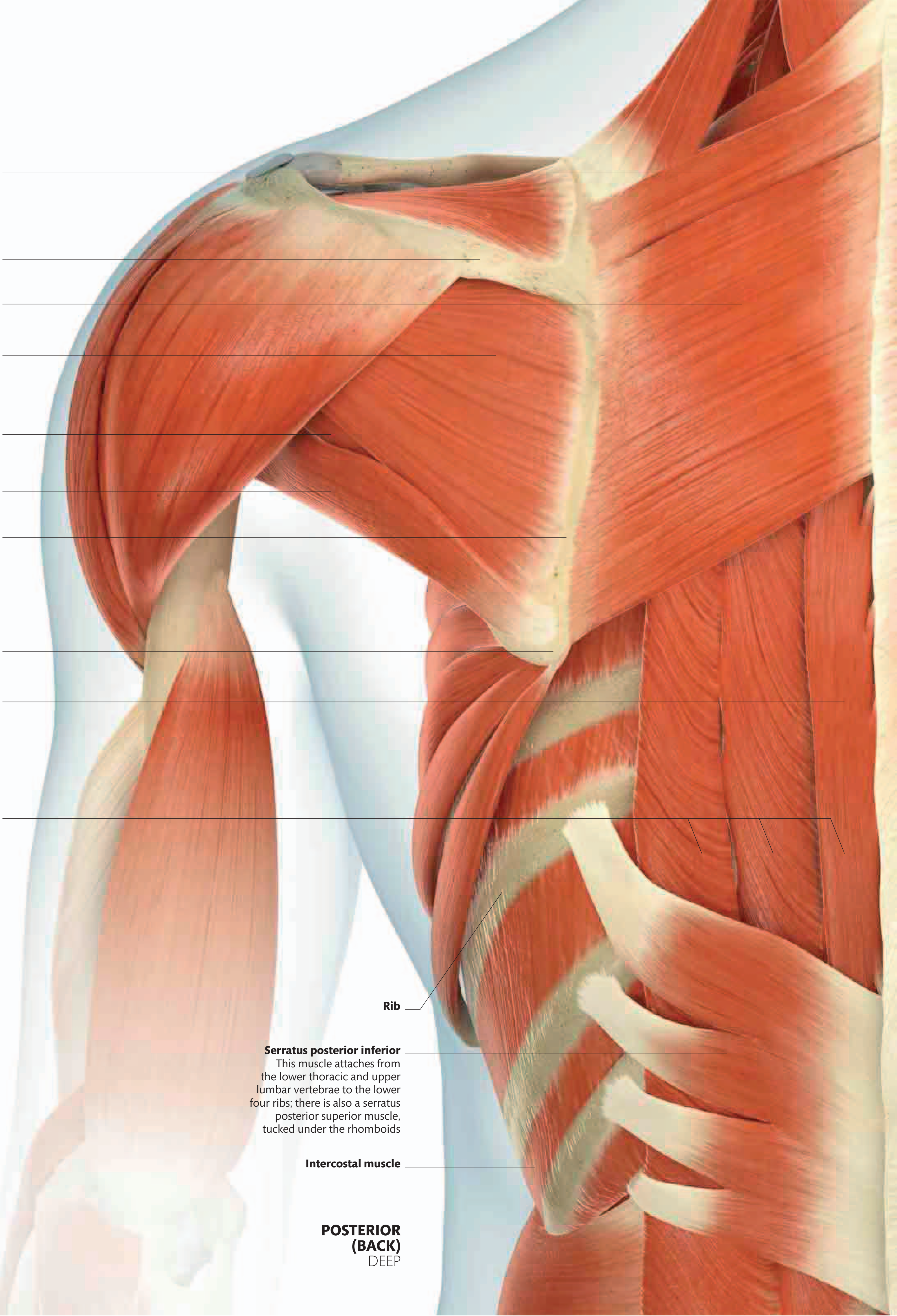

Rhomboid minor
The four-sided rhomboid muscles act to pull the scapulae toward the midline

Spine of scapula

Rhomboid major

Infraspinatus
One of the rotator cuff, or short scapular muscles

Teres minor

Teres major

Vertebral (medial) border of scapula

Inferior angle of scapula

Spinalis
The innermost (most medial) part of the erector spinae; it attaches to the spinous processes of the vertebrae

Erector spinae muscle group

Rib

Serratus posterior inferior
This muscle attaches from the lower thoracic and upper lumbar vertebrae to the lower four ribs; there is also a serratus posterior superior muscle, tucked under the rhomboids

Intercostal muscle

POSTERIOR (BACK)
DEEP

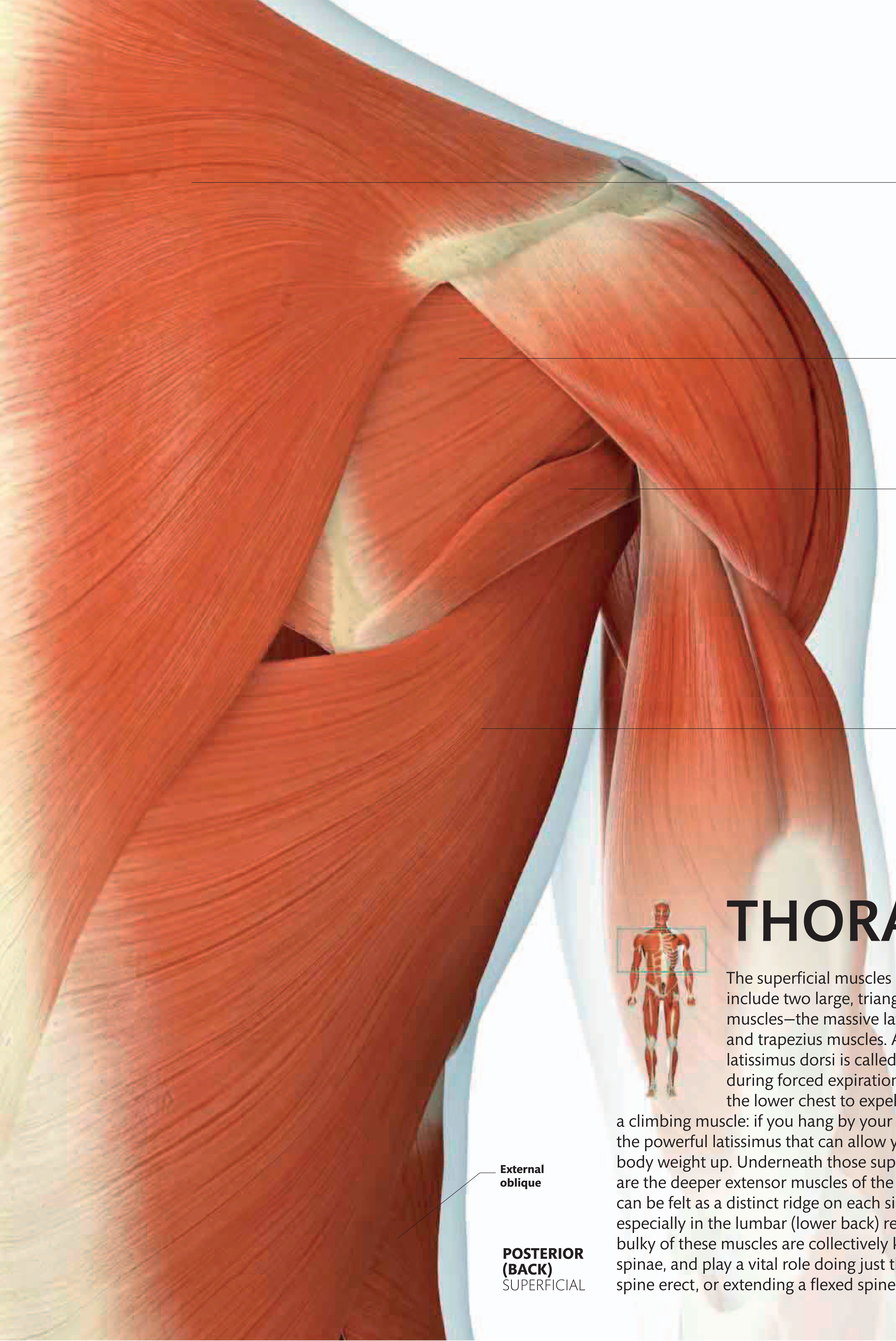

Teres major
This tapering muscle takes its name from the Latin for rounded off

Latissimus dorsi
This huge muscle sweeps up from the lower part of the back to attach to the humerus

POSTERIOR (BACK)
SUPERFICIAL

THORAX

The superficial muscles of the back include two large, triangular-shaped muscles—the massive latissimus dorsi and trapezius muscles. Although latissimus dorsi is called into action during forced expiration, squeezing the lower chest to expel air, it is really a climbing muscle: if you hang by your arms, it is largely the powerful latissimus that can allow you to pull your body weight up. Underneath those superficial muscles are the deeper extensor muscles of the spine, which can be felt as a distinct ridge on each side of the spine, especially in the lumbar (lower back) region. The most bulky of these muscles are collectively known as erector spinae, and play a vital role doing just that—keeping the spine erect, or extending a flexed spine.

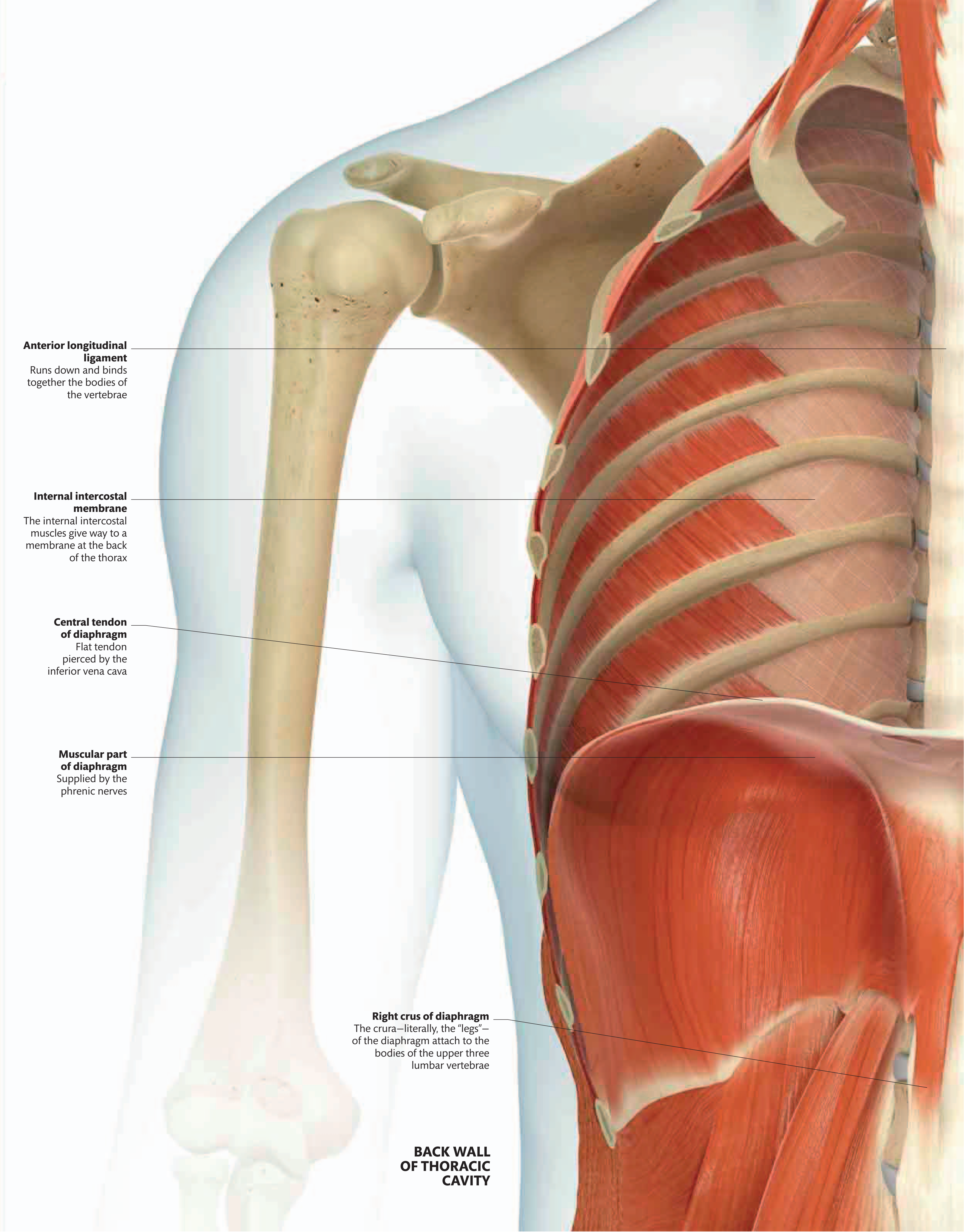

Anterior longitudinal ligament
Runs down and binds together the bodies of the vertebrae

Internal intercostal membrane
The internal intercostal muscles give way to a membrane at the back of the thorax

Central tendon of diaphragm
Flat tendon pierced by the inferior vena cava

Muscular part of diaphragm
Supplied by the phrenic nerves

Right crus of diaphragm
The crura—literally, the "legs"—of the diaphragm attach to the bodies of the upper three lumbar vertebrae

BACK WALL OF THORACIC CAVITY

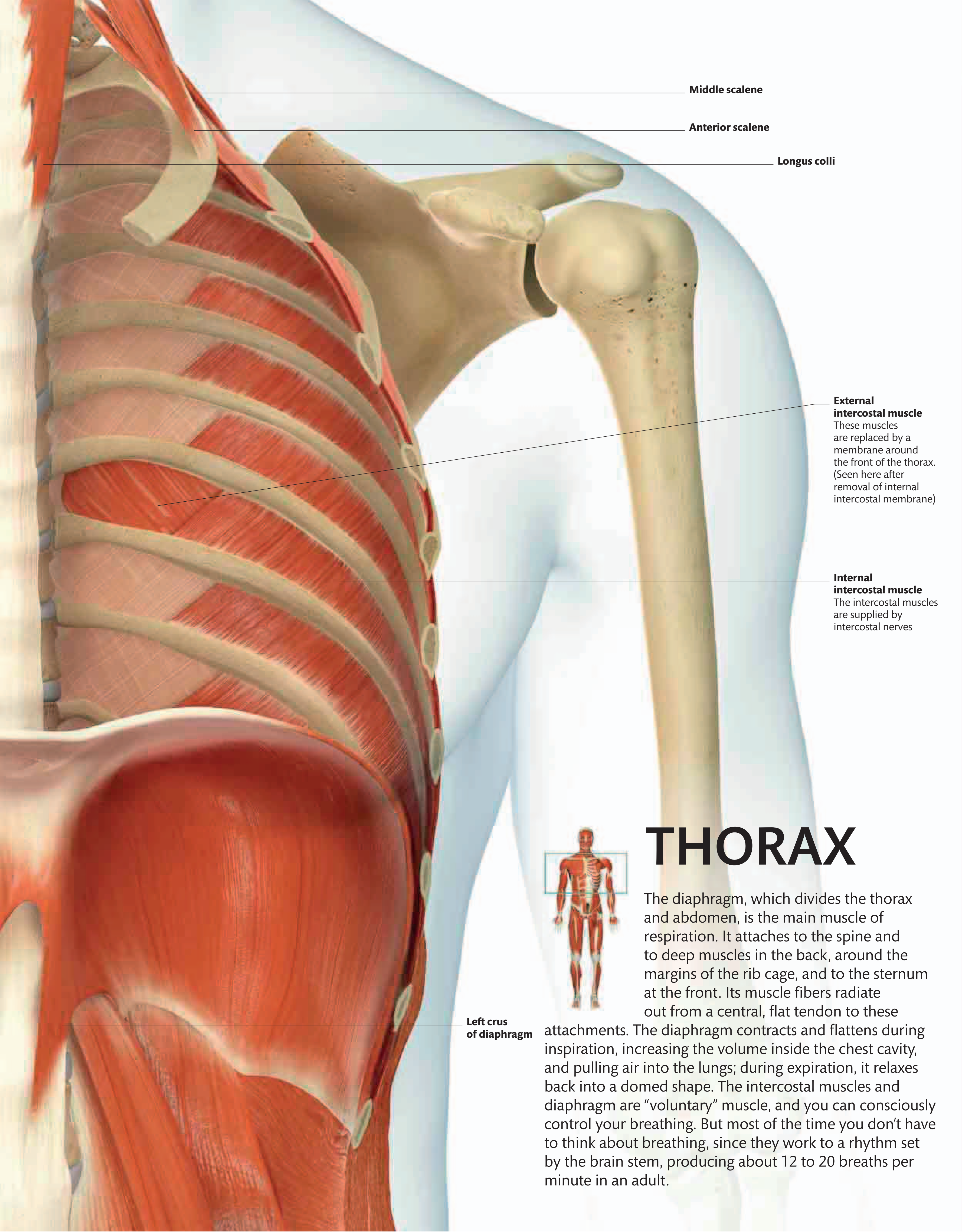

External intercostal muscle These muscles are replaced by a membrane around the front of the thorax. (Seen here after removal of internal intercostal membrane)

Internal intercostal muscle The intercostal muscles are supplied by intercostal nerves

THORAX

The diaphragm, which divides the thorax and abdomen, is the main muscle of respiration. It attaches to the spine and to deep muscles in the back, around the margins of the rib cage, and to the sternum at the front. Its muscle fibers radiate out from a central, flat tendon to these attachments. The diaphragm contracts and flattens during inspiration, increasing the volume inside the chest cavity, and pulling air into the lungs; during expiration, it relaxes back into a domed shape. The intercostal muscles and diaphragm are "voluntary" muscle, and you can consciously control your breathing. But most of the time you don't have to think about breathing, since they work to a rhythm set by the brain stem, producing about 12 to 20 breaths per minute in an adult.

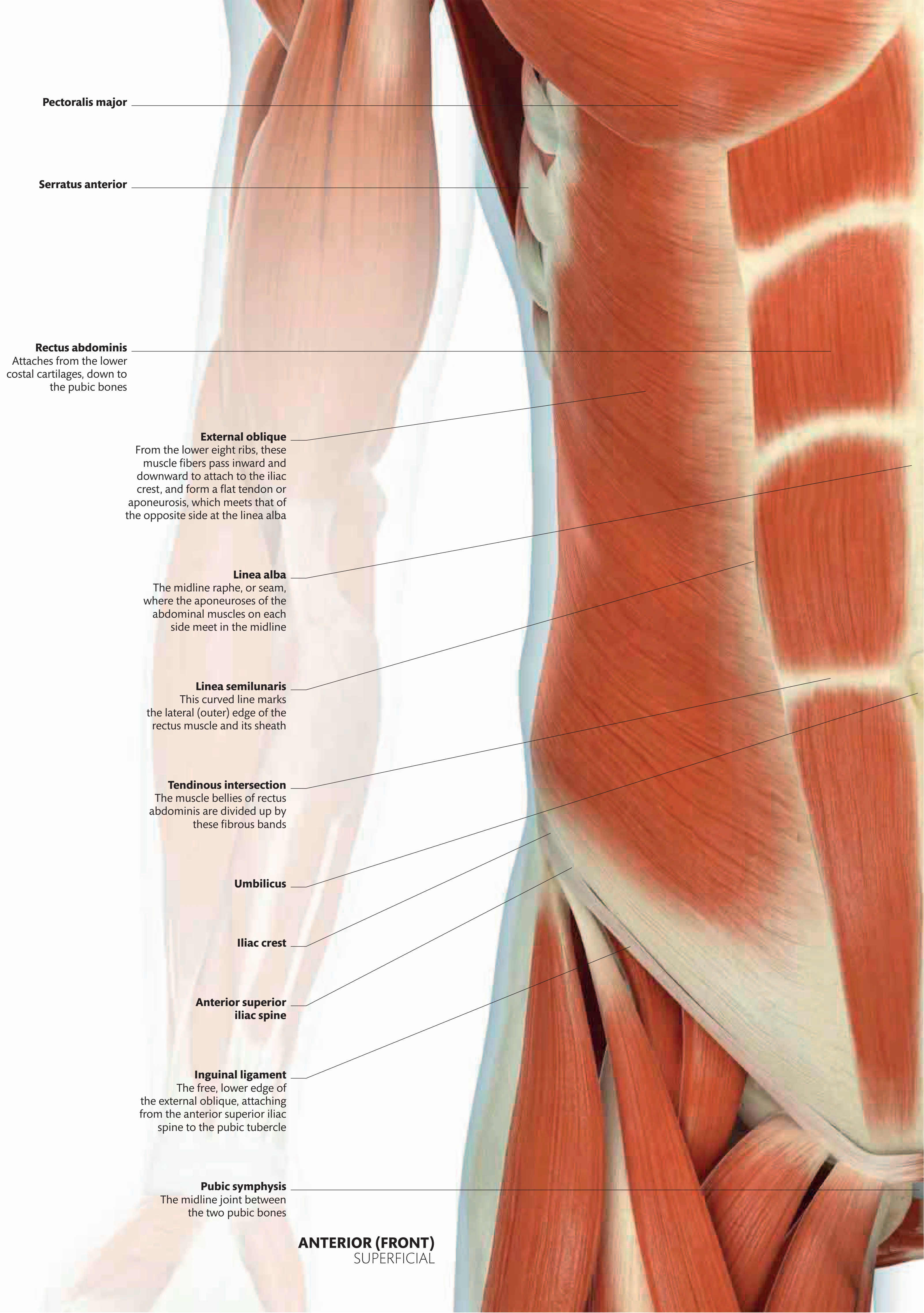

Pectoralis major

Serratus anterior

Rectus abdominis
Attaches from the lower costal cartilages, down to the pubic bones

External oblique
From the lower eight ribs, these muscle fibers pass inward and downward to attach to the iliac crest, and form a flat tendon or aponeurosis, which meets that of the opposite side at the linea alba

Linea alba
The midline raphe, or seam, where the aponeuroses of the abdominal muscles on each side meet in the midline

Linea semilunaris
This curved line marks the lateral (outer) edge of the rectus muscle and its sheath

Tendinous intersection
The muscle bellies of rectus abdominis are divided up by these fibrous bands

Umbilicus

Iliac crest

Anterior superior iliac spine

Inguinal ligament
The free, lower edge of the external oblique, attaching from the anterior superior iliac spine to the pubic tubercle

Pubic symphysis
The midline joint between the two pubic bones

ANTERIOR (FRONT)
SUPERFICIAL

ABDOMEN AND PELVIS

The abdominal muscles can move the trunk—flexing the spine to the front or to the side, or twisting the abdomen from side to side. They are very important muscles in posture, helping support the upright spine when we are standing or sitting, and are also called into action when we lift heavy objects. Because they compress the abdomen and raise the pressure internally, they are involved during defecation, micturition (emptying the bladder), and in forced expiration of air from the lungs. Right at the front, lying either side of the midline, there are two straight, straplike rectus abdominis muscles. These muscles are each broken up by horizontal tendons: in a well-toned, slim person, this creates the much-sought-after "six-pack" appearance. Flanking the recti muscles on each side are three layers of broad, flat muscles.

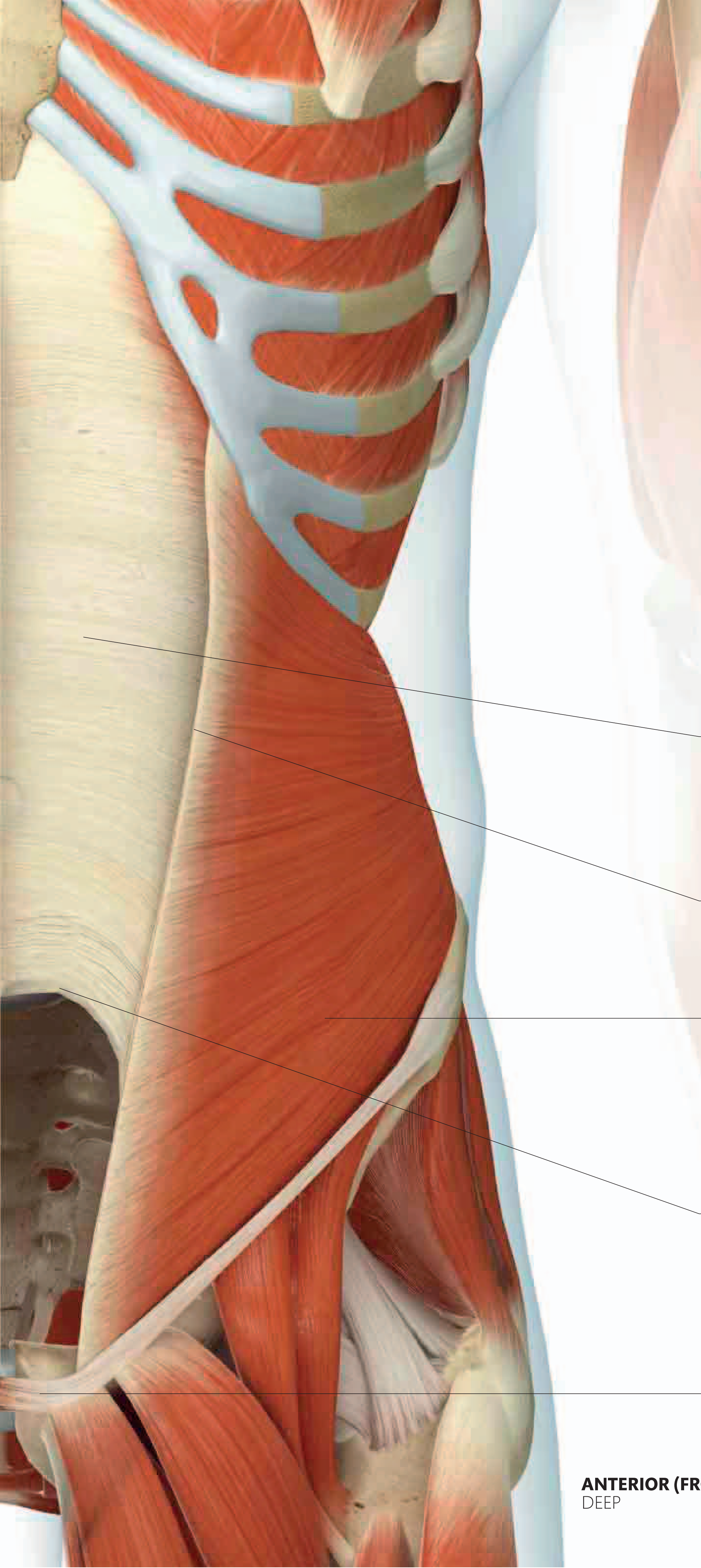

Posterior layer of rectus sheath
The rectus sheath is formed by the aponeuroses of the muscles to the sides: the external oblique, the internal oblique, and the transversus abdominis

Aponeurosis of internal oblique (cut edge)

Internal oblique
Lying underneath the external oblique, these muscle fibers spring from the inguinal ligament and iliac crest and fan inward and upward, attaching to the lower ribs and to each other in the midline

Arcuate line
At this point, all the aponeuroses of the lateral muscles swap to lie in front of the rectus abdominis muscles, leaving only a layer of fascia behind that muscle

Pubic tubercle

ANTERIOR (FRONT)
DEEP

ABDOMEN AND PELVIS

The most superficial muscle of the lower back is the incredibly broad latissimus dorsi. Underneath this, lying along the spine on each side, there is a large bulk of muscle that forms two ridges in the lumbar region in a well-toned person. This muscle mass is collectively known as the erector spinae, and its name suggests its importance in keeping the spine upright. When the spine is flexed forward, the erector spinae can pull it back into an upright position, and even take it further, into extension. The muscle can be divided up into three main strips on each side: iliocostalis, longissimus, and spinalis. Most of the muscle bulk of the buttock comes down to just one muscle: the fleshy gluteus maximus, which extends the hip joint. Hidden beneath the gluteus maximus are a range of smaller muscles that also move the hip.

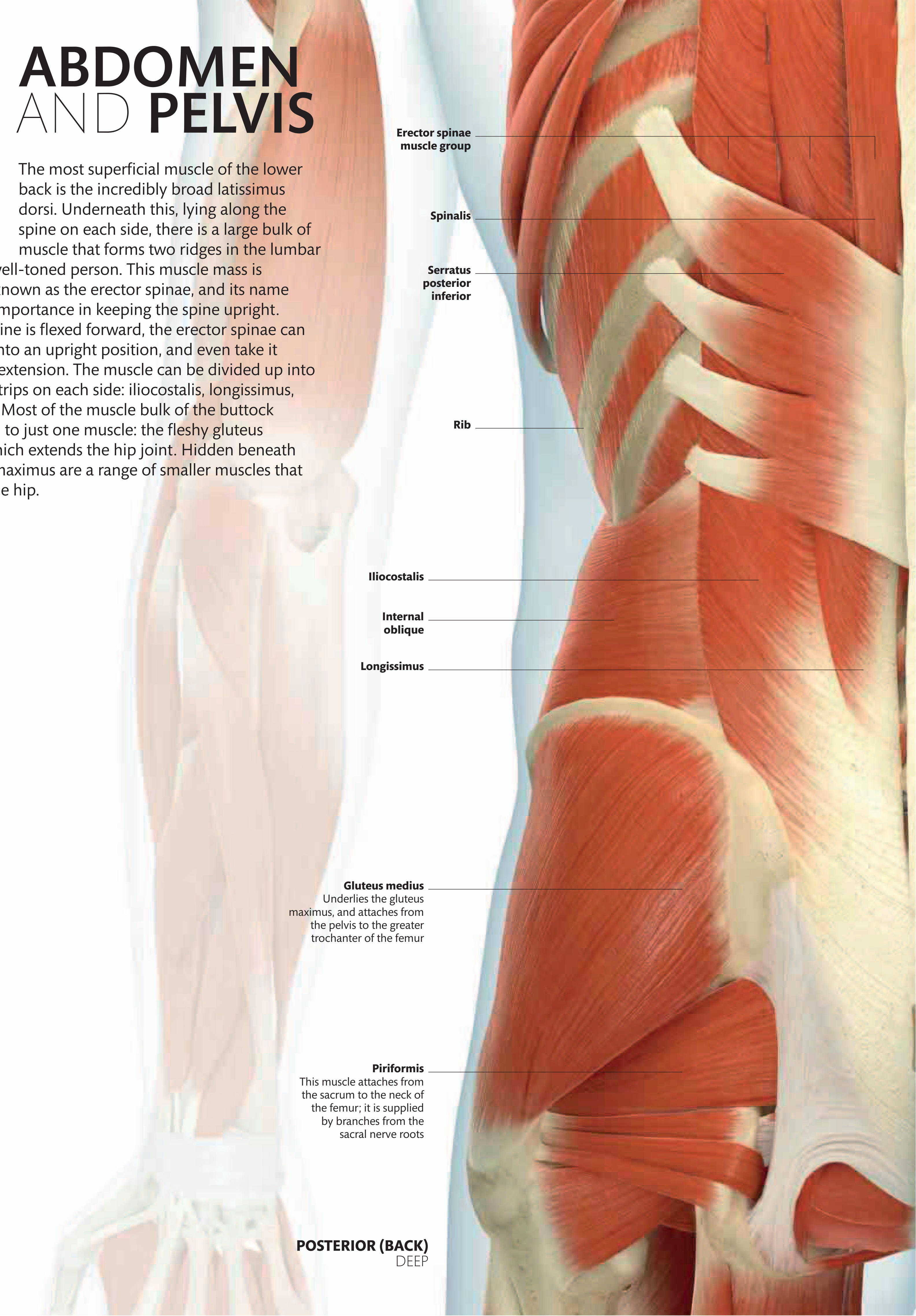
Erector spinae muscle group
Spinalis
Serratus posterior inferior
Rib
Iliocostalis
Internal oblique
Longissimus
Gluteus medius
Piriformis

Gluteus medius Underlies the gluteus maximus, and attaches from the pelvis to the greater trochanter of the femur

Piriformis This muscle attaches from the sacrum to the neck of the femur; it is supplied by branches from the sacral nerve roots

POSTERIOR (BACK)
DEEP

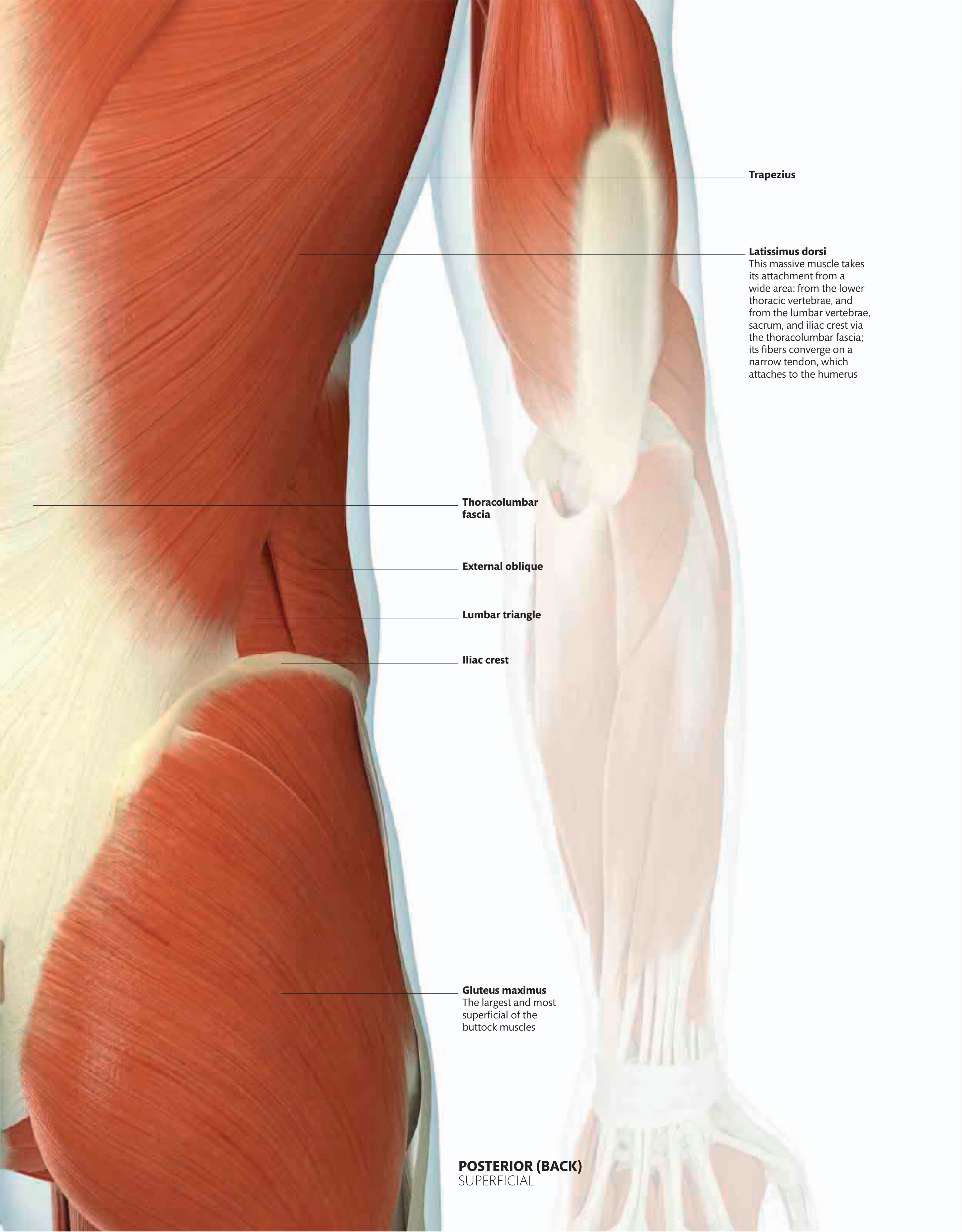

Trapezius

Latissimus dorsi
This massive muscle takes its attachment from a wide area: from the lower thoracic vertebrae, and from the lumbar vertebrae, sacrum, and iliac crest via the thoracolumbar fascia; its fibers converge on a narrow tendon, which attaches to the humerus

Thoracolumbar fascia

External oblique

Lumbar triangle

Iliac crest

Gluteus maximus
The largest and most superficial of the buttock muscles

POSTERIOR (BACK)
SUPERFICIAL

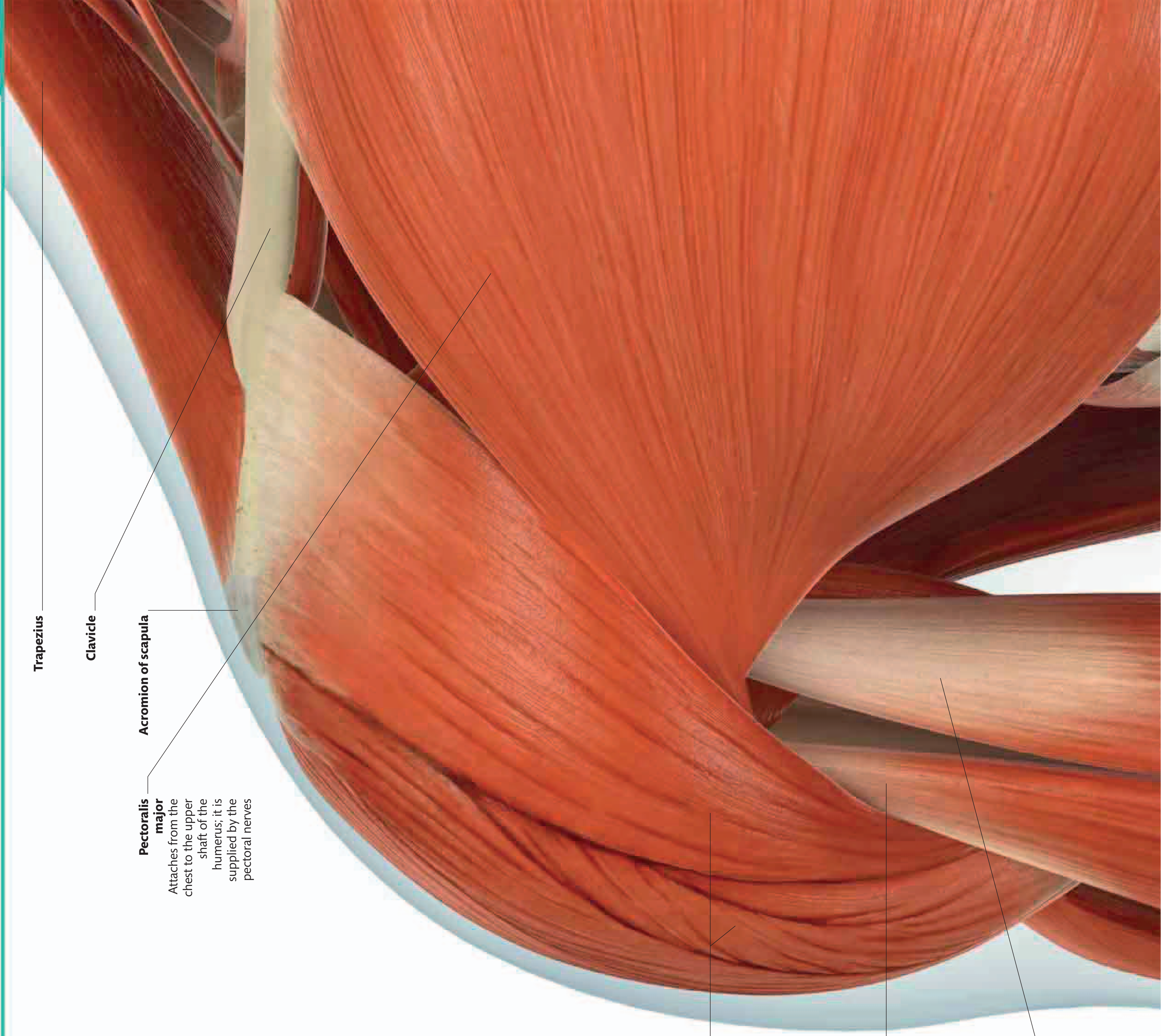
Trapezius
Clavicle
Acromion of scapula
Pectoralis major
Attaches from the chest to the upper shaft of the humerus; it is supplied by the pectoral nerves
Deltoid
This powerful muscle attaches from the clavicle, acromion, and spine of the scapula to the deltoid tuberosity on the side of the humerus
Long head of biceps
This tendon disappears under the deltoid sooner than the short head, so it appears to be the shorter of the two, but it runs right over the head of the humerus to attach to the scapula above the glenoid fossa
Short head of biceps
Attaches to the coracoid process of the scapula

SHOULDER AND UPPER ARM

The triangular deltoid muscle lies over the shoulder. Acting as a whole, this muscle raises the arm to the side (abduction), but the fibers of the deltoid attaching to the front of the clavicle can also move the arm forward. The pectoralis major muscle can also act on the shoulder joint, flexing the arm forward or pulling it in to the side of the chest (adduction). The biceps brachii muscle forms much of the muscle bulk on the front of the arm. The biceps tendon inserts on the radius, and also has an aponeurosis (flat tendon) that fans out over the forearm muscles. The biceps is a powerful flexor of the elbow, and can also rotate the radius to position the lower arm so the palm faces upward (supination).

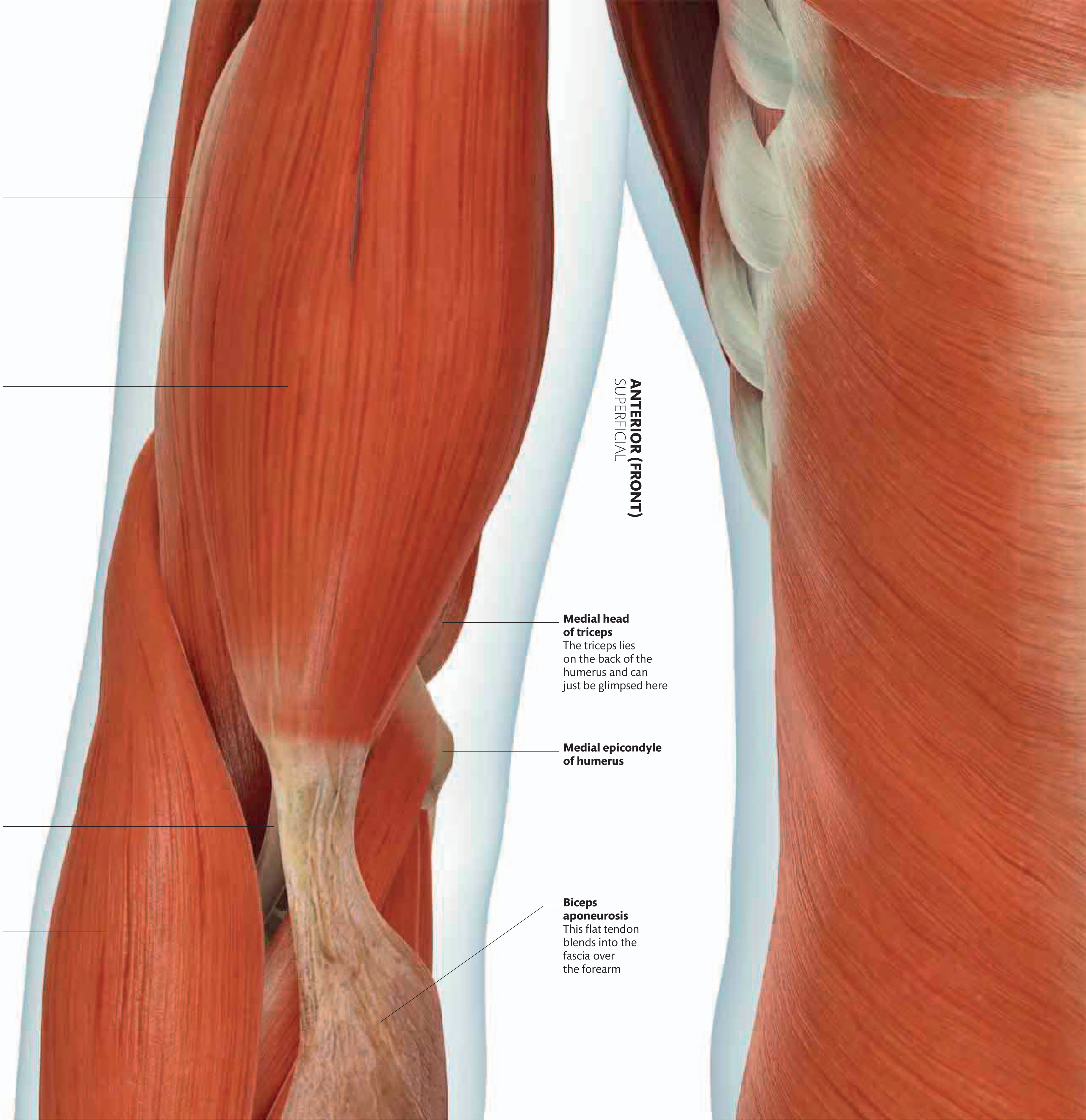

ANTERIOR (FRONT)
SUPERFICIAL

Brachialis
Lying deeper than the biceps, only an edge of brachialis can be seen here

Biceps brachii
Supplied by the musculocutaneous nerve

Medial head of triceps
The triceps lies on the back of the humerus and can just be glimpsed here

Medial epicondyle of humerus

Biceps tendon
Attaches to the radial tuberosity

Biceps aponeurosis
This flat tendon blends into the fascia over the forearm

Brachioradialis

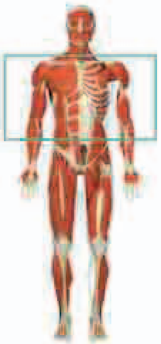

SHOULDER AND UPPER ARM

The posterior fibers of the deltoid attach from the spine of the scapula (shoulder blade) down to the humerus, and this part of the muscle can draw back the arm, or extend it. Latissimus dorsi (a broad muscle attaching from the back of the trunk and ending in a narrow tendon that secures onto the humerus) can also extend the arm. The triceps brachii muscle is the sole extensor of the elbow. In a superficial dissection (represented in this view) only two of the three heads of the triceps can be seen—the long and lateral heads. The triceps tendon attaches to the leverlike olecranon of the ulna, which forms the bony knobble at the back of the elbow.

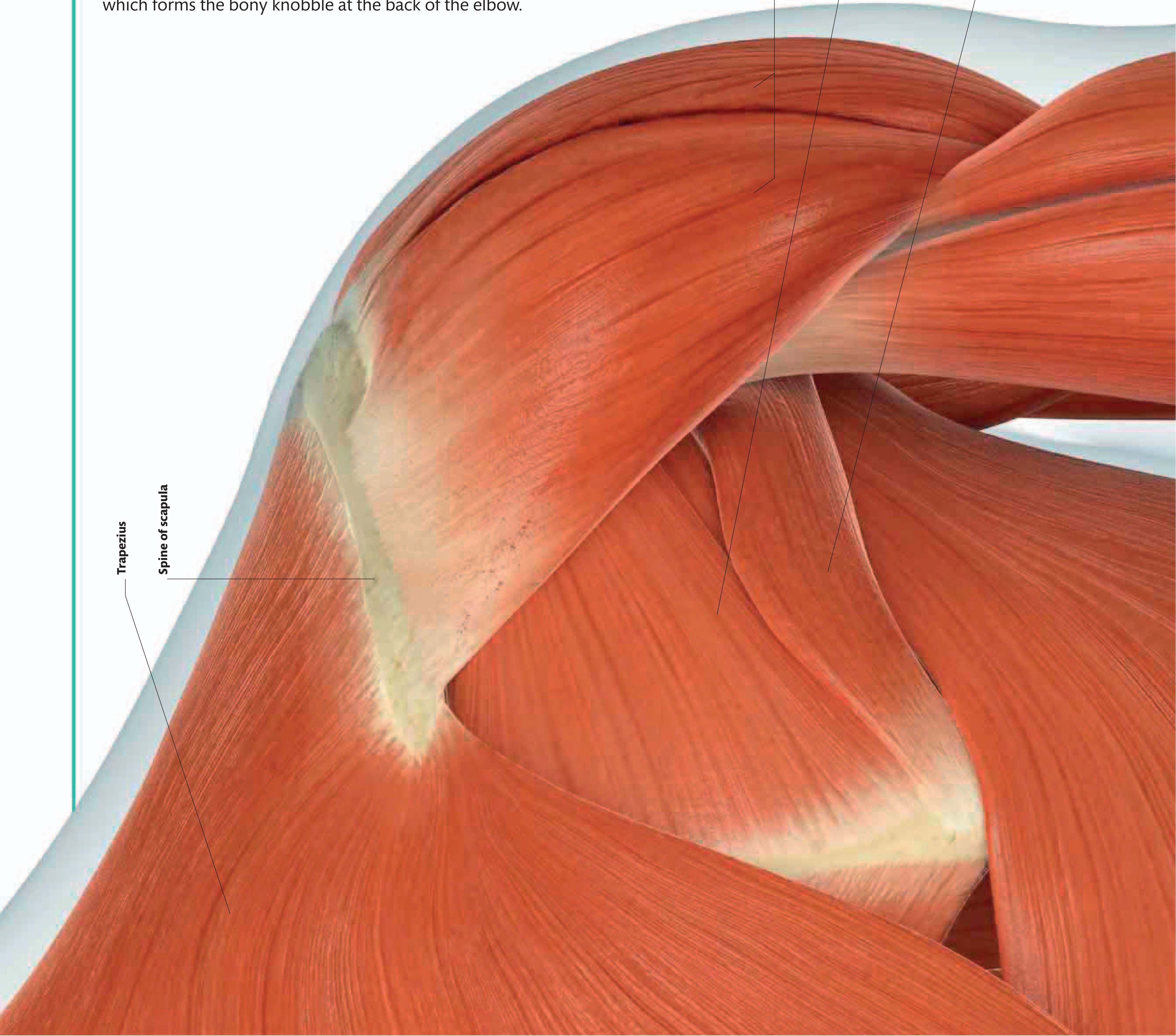

Infraspinatus
Attaches from the infraspinous fossa, below the spine of the scapula, to the back of the neck of the humerus; it can rotate the humerus along its axis, out to the side

Teres major
Attaches from the scapula to the front of the neck of the humerus, and rotates the humerus inward

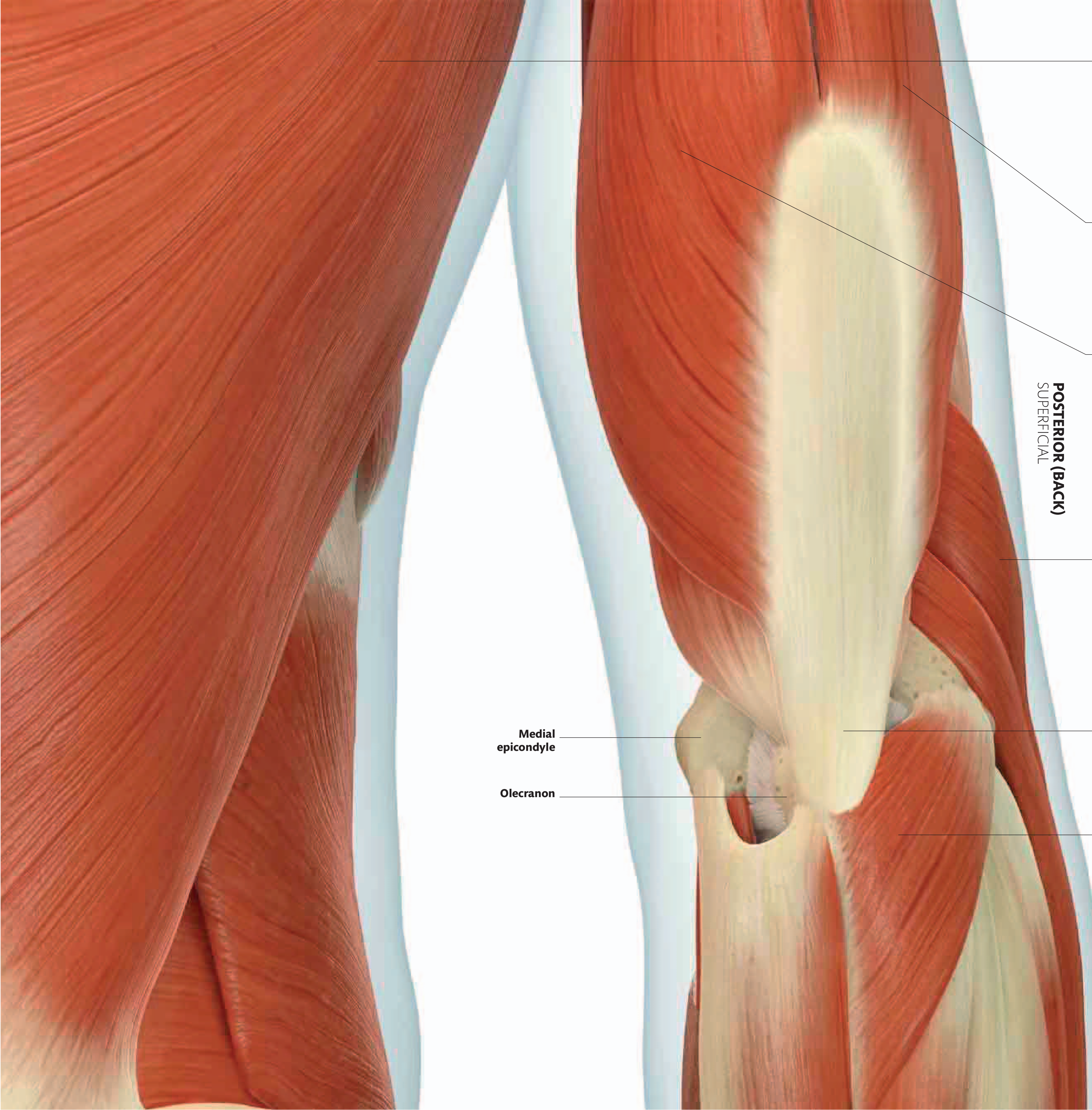

POSTERIOR (BACK)
SUPERFICIAL

Latissimus dorsi
When the arm is extended upward, this large muscle can pull it down to the side of the body or, in the opposite direction, it can pull the weight of the body up, toward the arms (which is important for climbing)

Lateral head of triceps
This, and the long head of the triceps are superficial; the third, medial, head is hidden beneath them; all three are supplied by the radial nerve

Long head of triceps
Attaches to the scapula, just below the glenoid fossa

Brachioradialis
Triceps tendon
Anconeus
Medial epicondyle
Olecranon

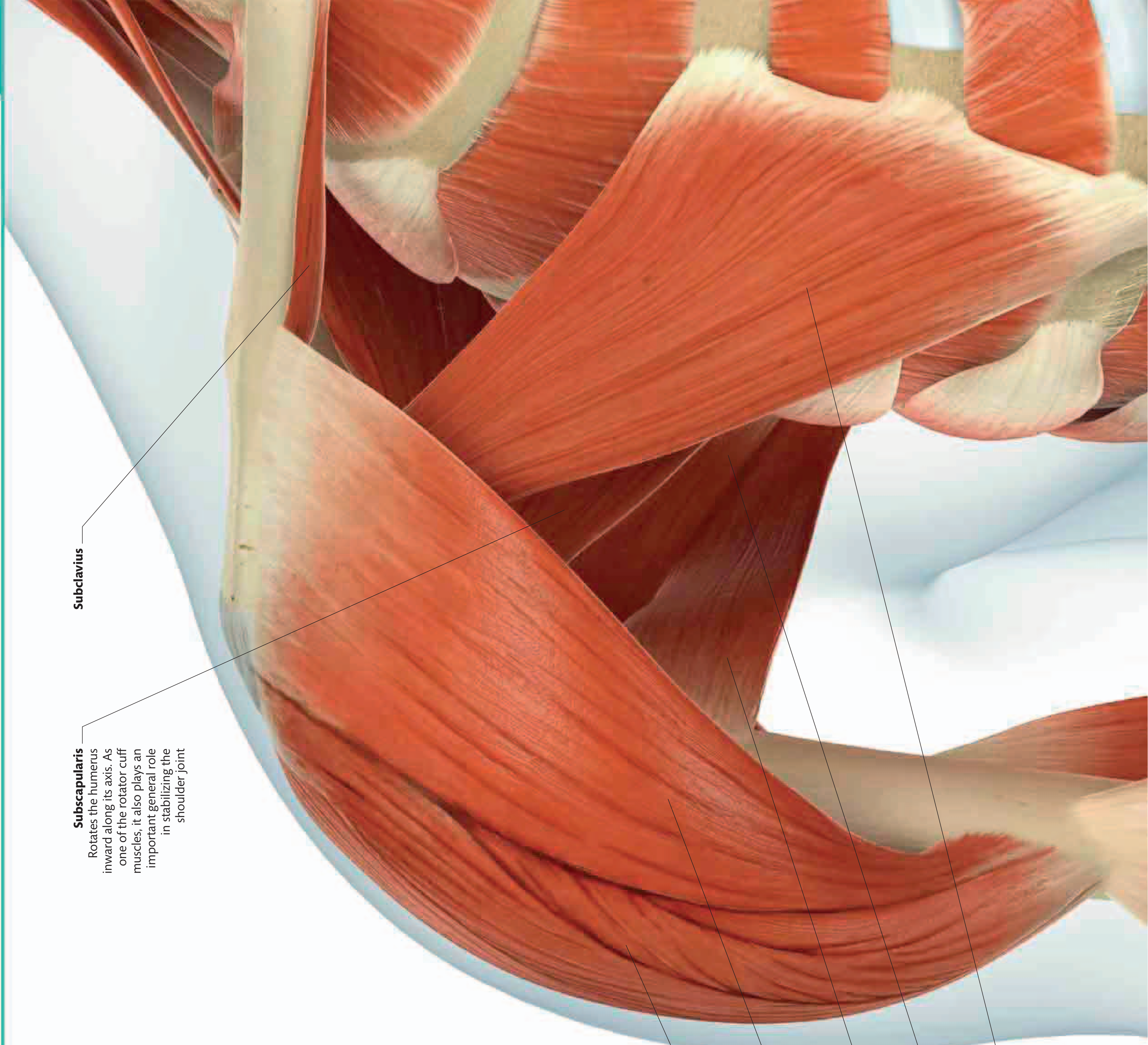
Subclavius
Subscapularis
Middle fibers of deltoid
Anterior fibers of deltoid
Latissimus dorsi
Teres major
Pectoralis minor

Subscapularis
Rotates the humerus inward along its axis. As one of the rotator cuff muscles, it also plays an important general role in stabilizing the shoulder joint

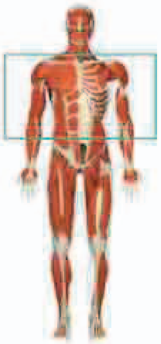

SHOULDER AND UPPER ARM

The deep muscles around the shoulder include the so-called rotator cuff group, two of which can be seen here: the subscapularis (which attaches from the deep surface of the scapula) and the supraspinatus (which runs from the scapula, over the shoulder joint, to attach to the humerus). The supraspinatus tendon passes through a narrow gap between the head of the humerus and the acromion of the scapula, and may become compressed and damaged here, in impingement syndrome. On the front of the humerus, the biceps (see p.85) has been removed to reveal brachialis, which runs from the lower humerus down to the ulna. Like the biceps, brachialis is a flexor of the elbow.

ANTERIOR (FRONT)
DEEP

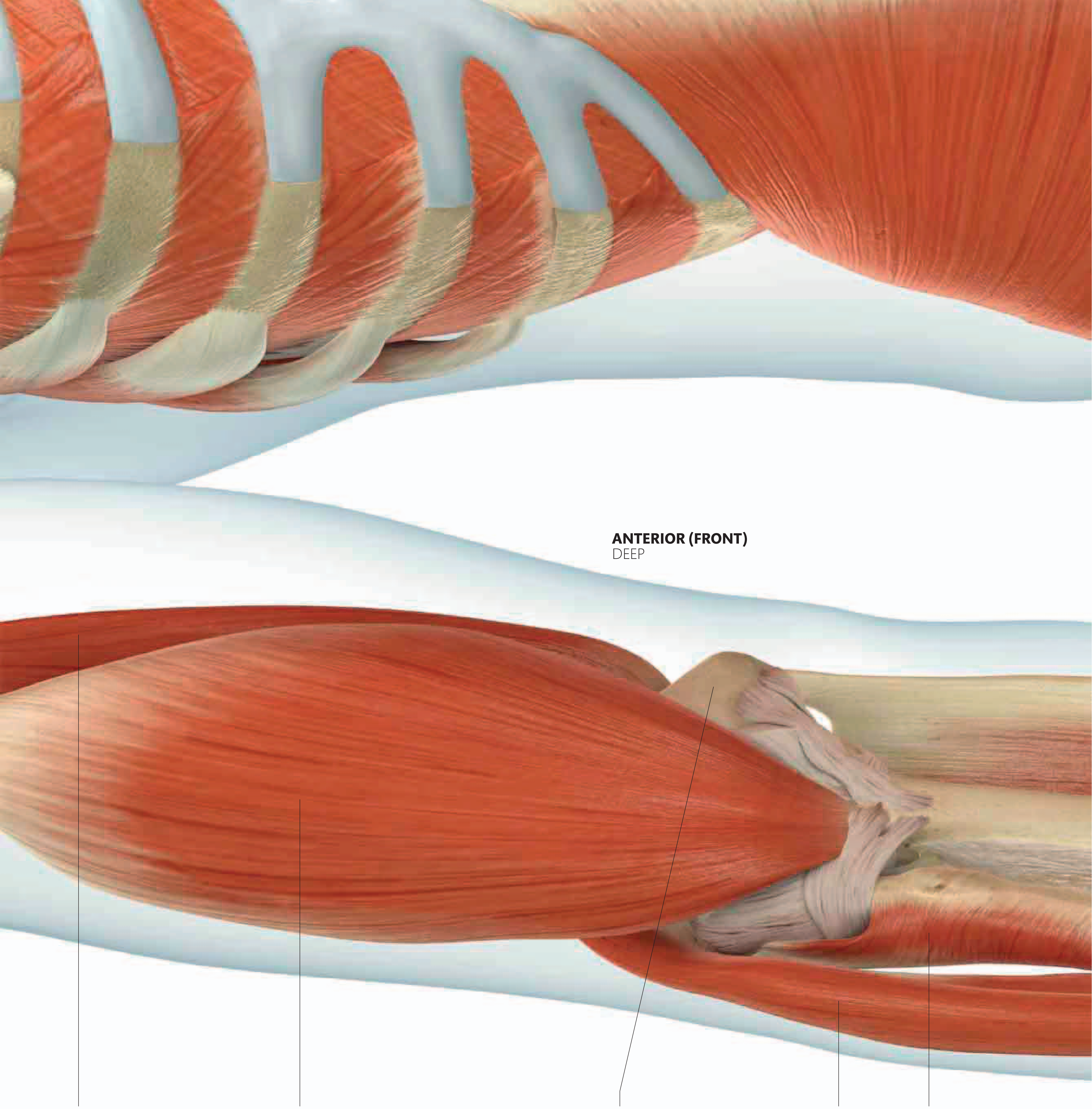

Medial head of triceps
Brachialis
Medial epicondyle of humerus
Brachioradialis
Supinator

Brachialis
The name of this muscle simply means of the arm in Latin. It attaches from the front of the humerus to the tuberosity on the front of the ulna, and it flexes the elbow. It lies underneath the biceps

SHOULDER AND UPPER ARM

More of the rotator cuff muscles—the supraspinatus, infraspinatus, and teres minor—can be seen from the back. In addition to moving the shoulder joint in various directions, including rotation, these muscles are important in helping to stabilize the shoulder joint: they hug the head of the humerus into its socket during movements at the shoulder. On the back of the arm, a deeper view reveals the third, medial head of the triceps, which attaches from the back of the humerus. It joins with the lateral and long heads to form the triceps tendon, attaching to the olecranon. Most of the forearm muscles take their attachment from the epicondyles of the humerus, just above the elbow, but the brachioradialis and extensor carpi radialis longus have higher origins from the side of the humerus, as shown here.

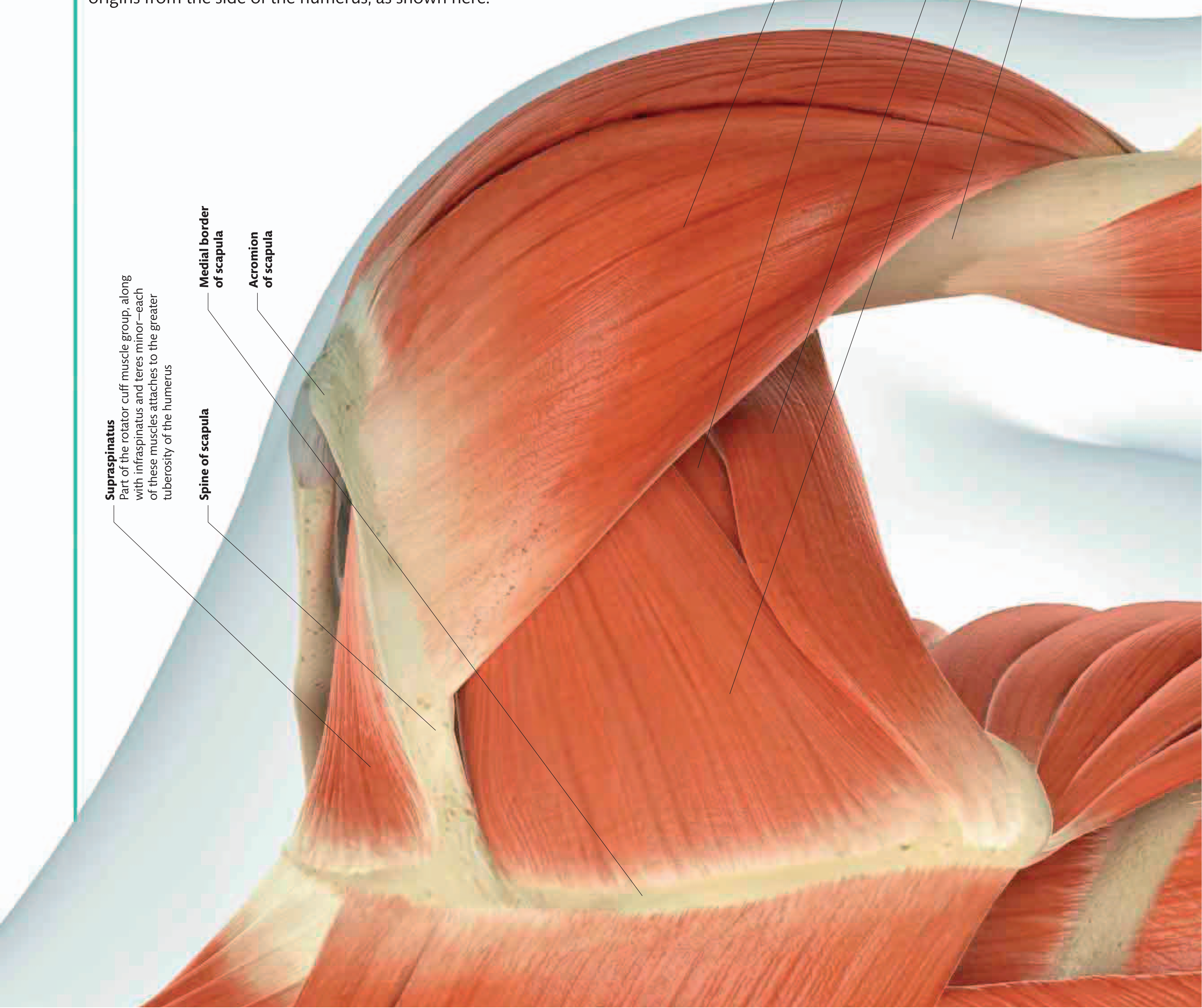
Posterior fibers of deltoid
Teres minor Like infraspinatus, this muscle can laterally rotate the humerus along its axis
Teres major
Infraspinatus
Shaft of humerus
Supraspinatus Part of the rotator cuff muscle group, along with infraspinatus and teres minor—each of these muscles attaches to the greater tuberosity of the humerus
Spine of scapula
Medial border of scapula
Acromion of scapula

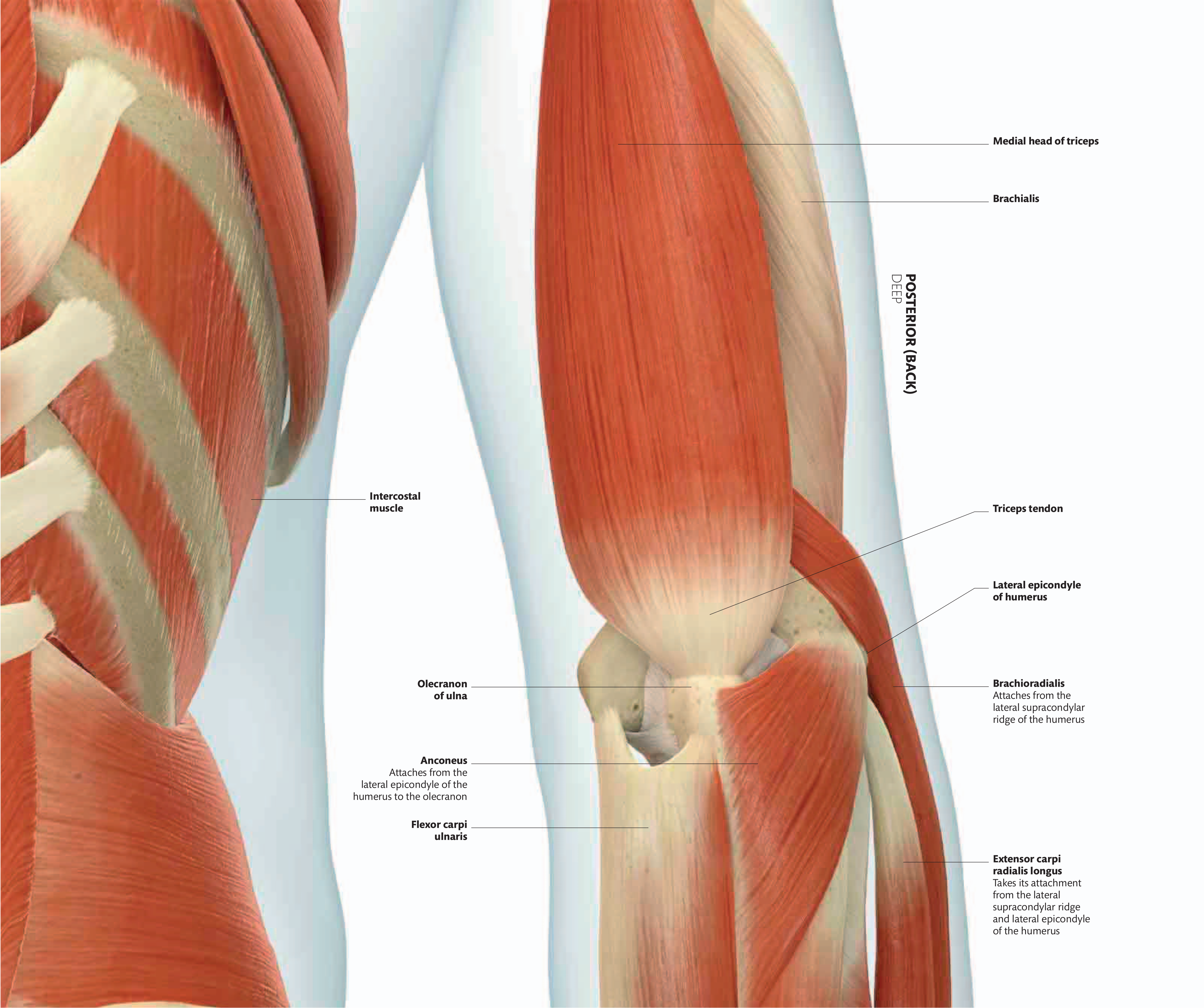

POSTERIOR (BACK)
DEEP

Medial head of triceps
Brachialis
Intercostal muscle
Triceps tendon
Lateral epicondyle of humerus
Olecranon of ulna
Brachioradialis Attaches from the lateral supracondylar ridge of the humerus
Anconeus Attaches from the lateral epicondyle of the humerus to the olecranon
Flexor carpi ulnaris
Extensor carpi radialis longus Takes its attachment from the lateral supracondylar ridge and lateral epicondyle of the humerus

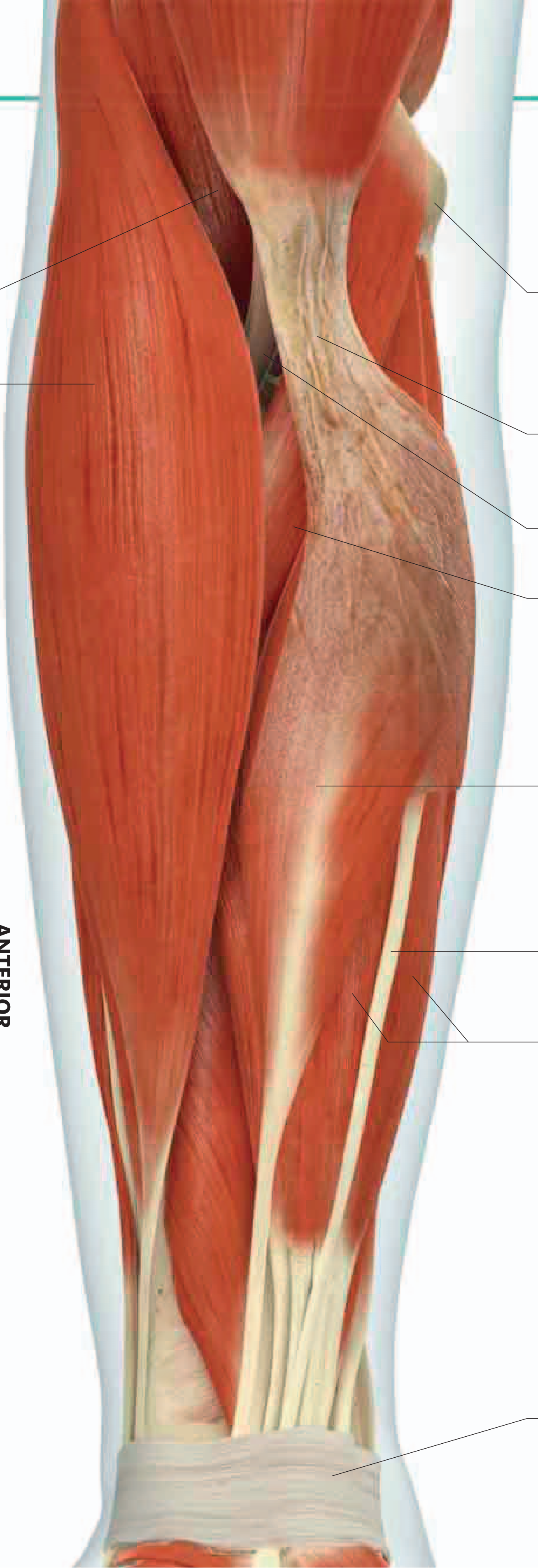

Medial epicondyle of humerus
Also called the common flexor origin; many of the superficial flexor muscles attach from this point

Biceps aponeurosis

Biceps tendon

Pronator teres
Attaches from the humerus and ulna down to the outer edge of the radius; it pronates the forearm, rotating the lower end of the radius around the ulna

Flexor carpi radialis
Radial extensor of the wrist; it arises from the medial epicondyle of the femur and secures on the base of the second metacarpal; it flexes the wrist and abducts the hand

Palmaris longus tendon

Flexor digitorum superficialis
Takes its attachment from the humerus, ulna, and radius, and splits into four tendons that run across the wrist into the hand, to flex the fingers

Flexor retinaculum
This fibrous band keeps the flexor tendons close to the wrist and stops them bow-stringing outward

Brachialis

Brachioradialis
Runs along the outer edge of the forearm and attaches to the end of the radius; it flexes and stabilizes the elbow

ANTERIOR (FRONT)
SUPERFICIAL

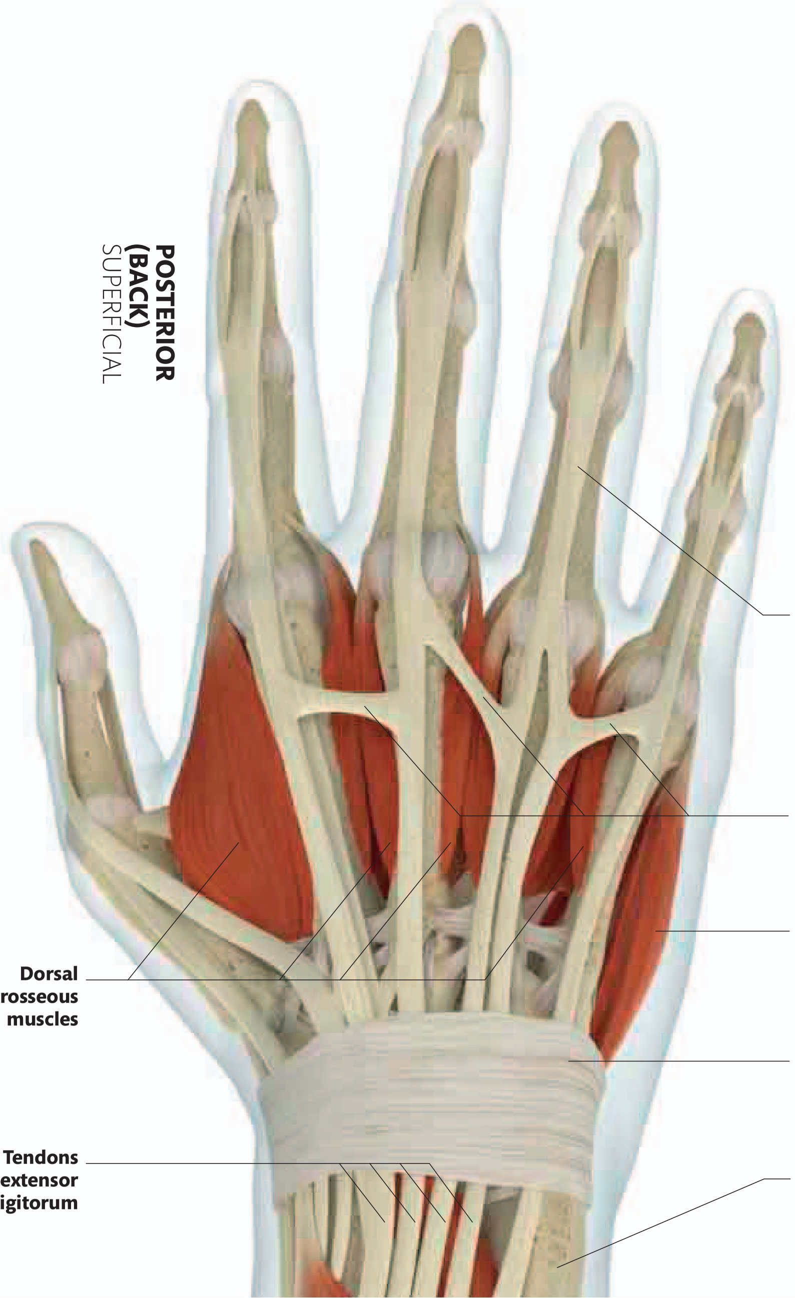

Extensor expansion

Intertendinous connections

Abductor digiti minimi

Extensor retinaculum
This fibrous band keeps the extensor tendons close to the wrist

Ulna

Dorsal interosseous muscles

Tendons of extensor digitorum

POSTERIOR (BACK)
SUPERFICIAL

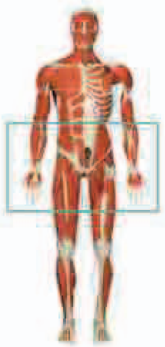

LOWER ARM AND HAND

There are five superficial muscles on the front of the forearm, all taking their attachment from the medial epicondyle of the humerus. Pronator teres attaches across to the radius, and can pull this bone into pronation (held with the palm turned downward). The other muscles run farther down the forearm, becoming slender tendons that attach around the wrist, or continue into the hand. Flexor digitorum superficialis splits into four tendons, one for each finger. On the back of the forearm, seven superficial extensor muscles attach to the lateral epicondyle of the humerus. Most of these tendons run down to the wrist or into the hand.

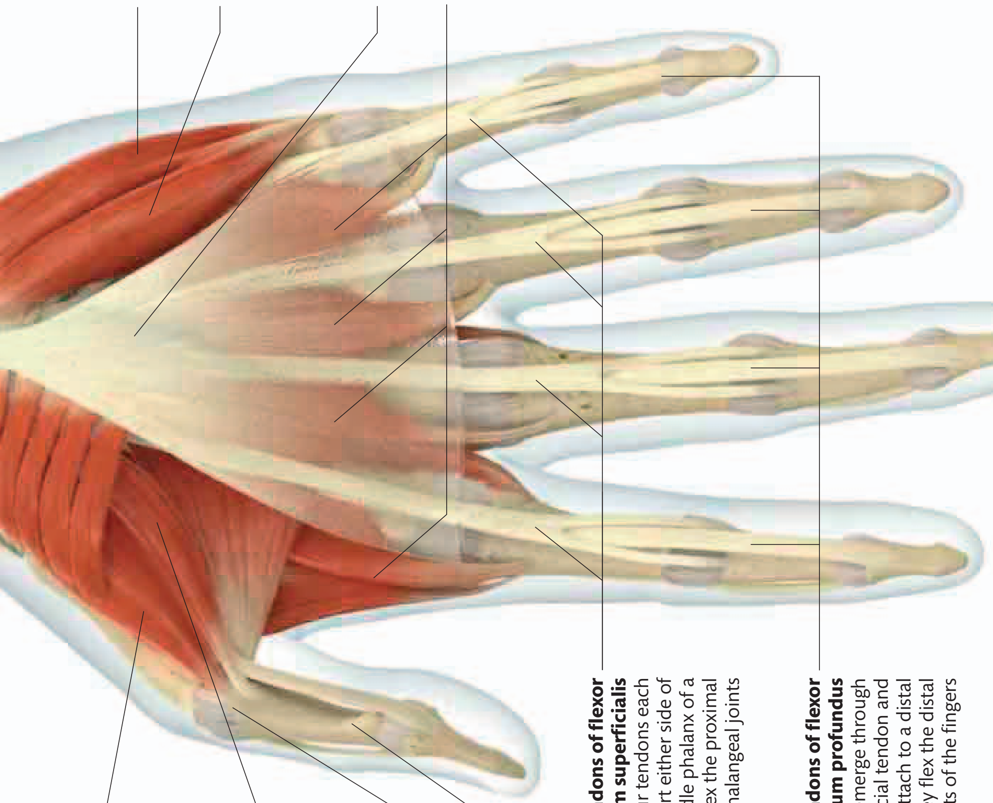
Abductor digiti minimi
Flexor digiti minimi brevis Short flexor of the little finger; it flexes the little finger's metacarpophalangeal joint
Palmar aponeurosis
Lumbricals These small muscles are named after the Latin for worm
Abductor pollicis brevis Attaches to the outer side of the base of the proximal phalanx of the thumb; with the palm facing up, it pulls the thumb upward, away from the palm and fingers
Flexor pollicis brevis Attaches to the base of the proximal phalanx of the thumb; it flexes the thumb's metacarpophalangeal joint
Metacarpophalangeal joint
First proximal phalanx
Tendons of flexor digitorum superficialis These four tendons each split to insert either side of the middle phalanx of a finger; they flex the proximal interphalangeal joints
Tendons of flexor digitorum profundus These tendons emerge through the superficial tendon and continue on, to attach to a distal phalanx; they flex the distal interphalangeal joints of the fingers

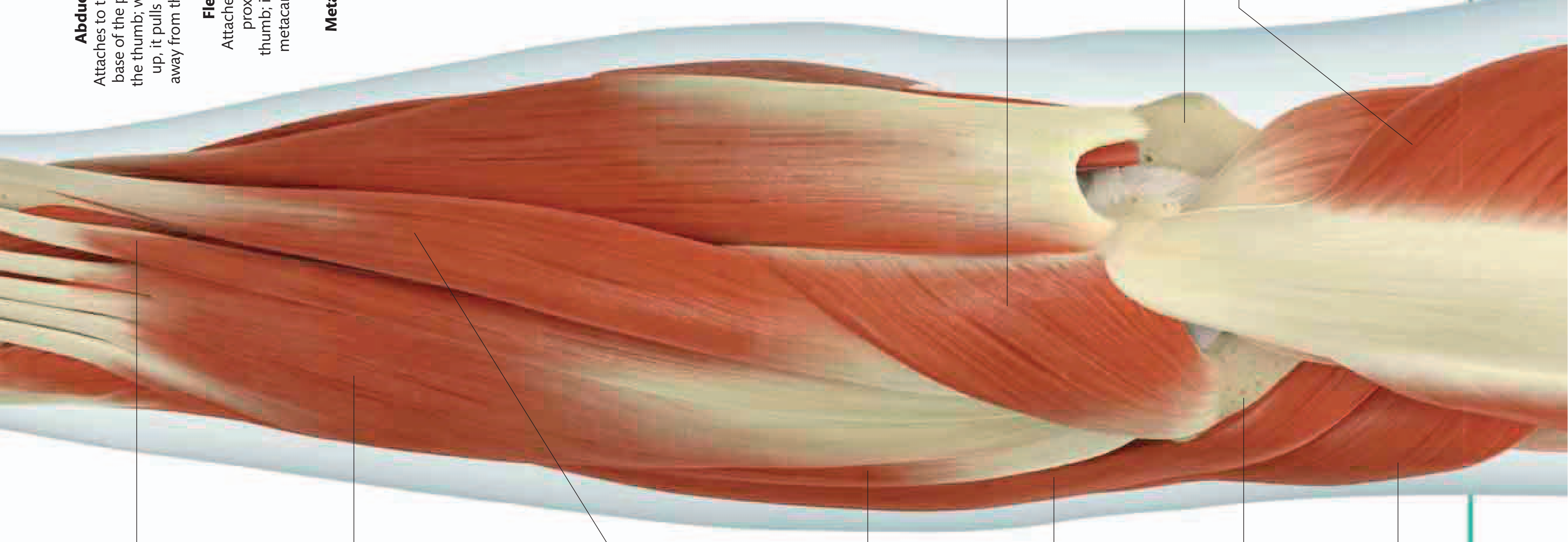
Anconeus Acts with the triceps to extend the elbow joint
Olecranon
Triceps
Extensor digiti minimi The tendon of this extensor of the little finger joins the tendon of the extensor digitorum on the back of the little finger
Extensor digitorum Extensor of the fingers; it takes its attachment from the lateral epicondyle and becomes four tendons that fan out over the back of the fingers, forming the "extensor expansion"
Extensor carpi ulnaris Ulnar extensor of the wrist; it arises from the lateral epicondyle and attaches to the base of the fifth metacarpal; it extends the wrist and adducts the hand
Extensor carpi radialis brevis Short extensor of the wrist; attaches from the lateral epicondyle to the third metacarpal in the hand
Extensor carpi radialis longus Long extensor of the wrist; it attaches from the lateral supracondylar ridge all the way down to the base of the second metacarpal
Lateral epicondyle of humerus Referred to as the common extensor origin–many forearm extensor muscles attach here
Brachioradialis

ANTERIOR (FRONT)
DEEP

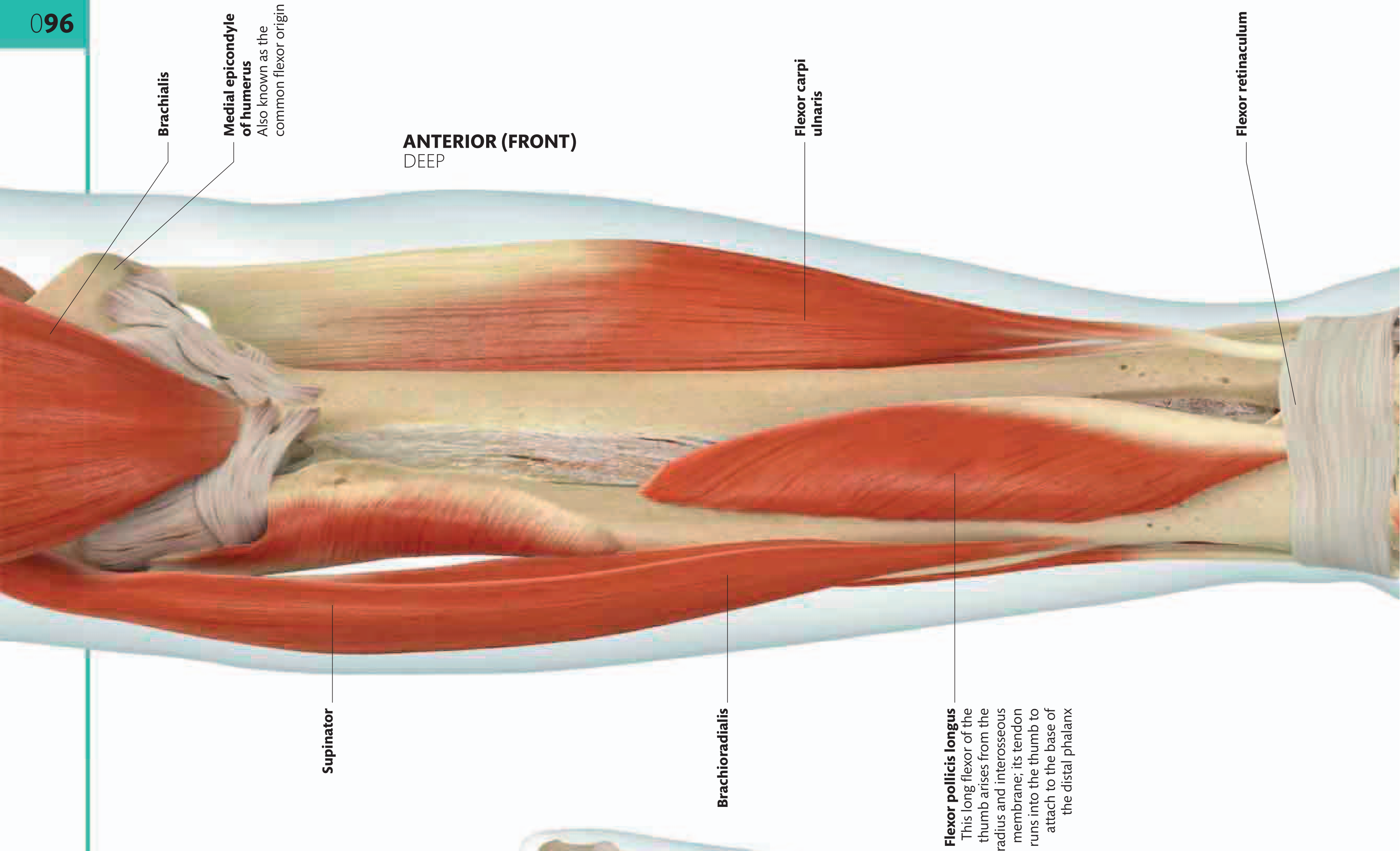
Brachialis
Medial epicondyle of humerus Also known as the common flexor origin
Flexor carpi ulnaris
Flexor retinaculum
Supinator
Brachioradialis
Flexor pollicis longus This long flexor of the thumb arises from the radius and interosseous membrane; its tendon runs into the thumb to attach to the base of the distal phalanx

POSTERIOR (BACK)
DEEP

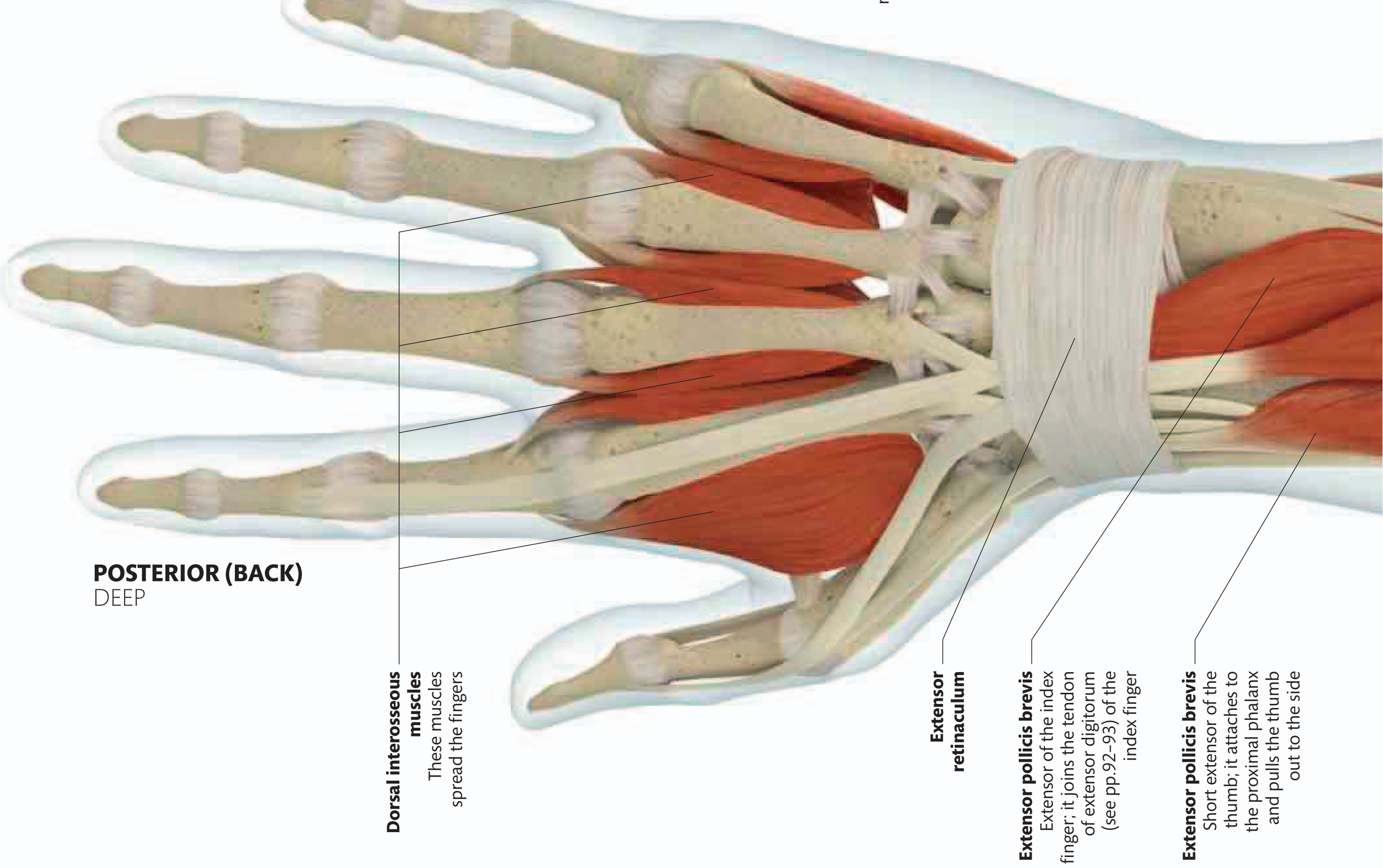
Dorsal interosseous muscles These muscles spread the fingers
Extensor retinaculum
Extensor pollicis brevis Extensor of the index finger; it joins the tendon of extensor digitorum (see pp.92–93) of the index finger
Extensor pollicis brevis Short extensor of the thumb; it attaches to the proximal phalanx and pulls the thumb out to the side

LOWER ARM AND HAND

Stripping away the superficial muscles on the front of the forearm reveals a deeper layer attaching to the radius and ulna, and to the interosseous membrane between the bones. The long, quill-like flexor of the thumb (flexor pollicis longus) can be seen clearly. Deep muscles on the back of the forearm include the long extensors of the thumb and index finger and the supinator, which pulls on the radius to rotate the pronated arm (held with palm facing downward) into supination (with palm facing up). In the hand, a deep dissection reveals the interosseous muscles that act on the metacarpophalangeal joints in order to either spread or close the fingers.

Opponens digiti minimi Opposer of the little finger; it draws the metacarpal of the little finger in towards the palm

Metacarpal of little finger

Palmar interosseous muscles Interosseous means between bones; these muscles lie in the gaps between the metacarpals. They adduct (bring together) the fingers

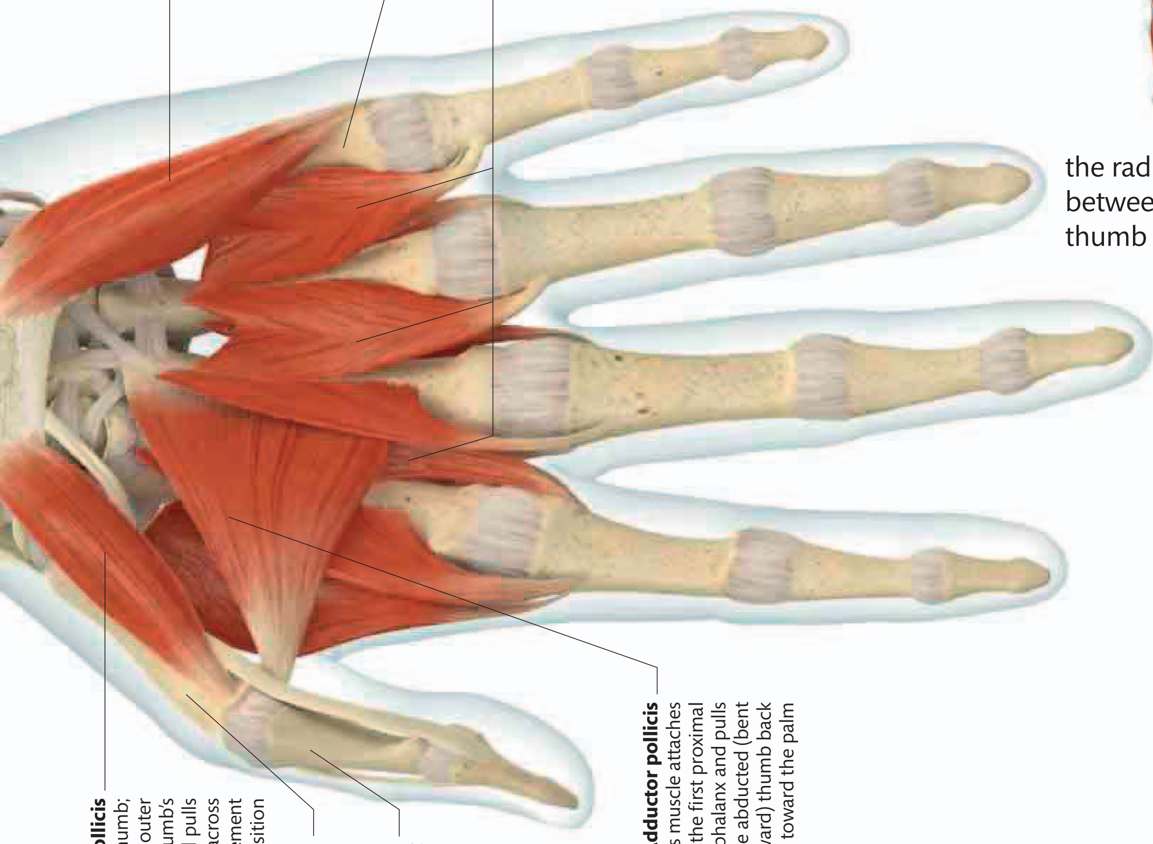

Opponens pollicis Opposer of the thumb; it attaches to the outer border of the thumb's metacarpal and pulls that metacarpal across the palm—in a movement known as opposition

Metacarpal of thumb

First proximal phalanx

Adductor pollicis This muscle attaches to the first proximal phalanx and pulls the abducted (bent outward) thumb back in toward the palm

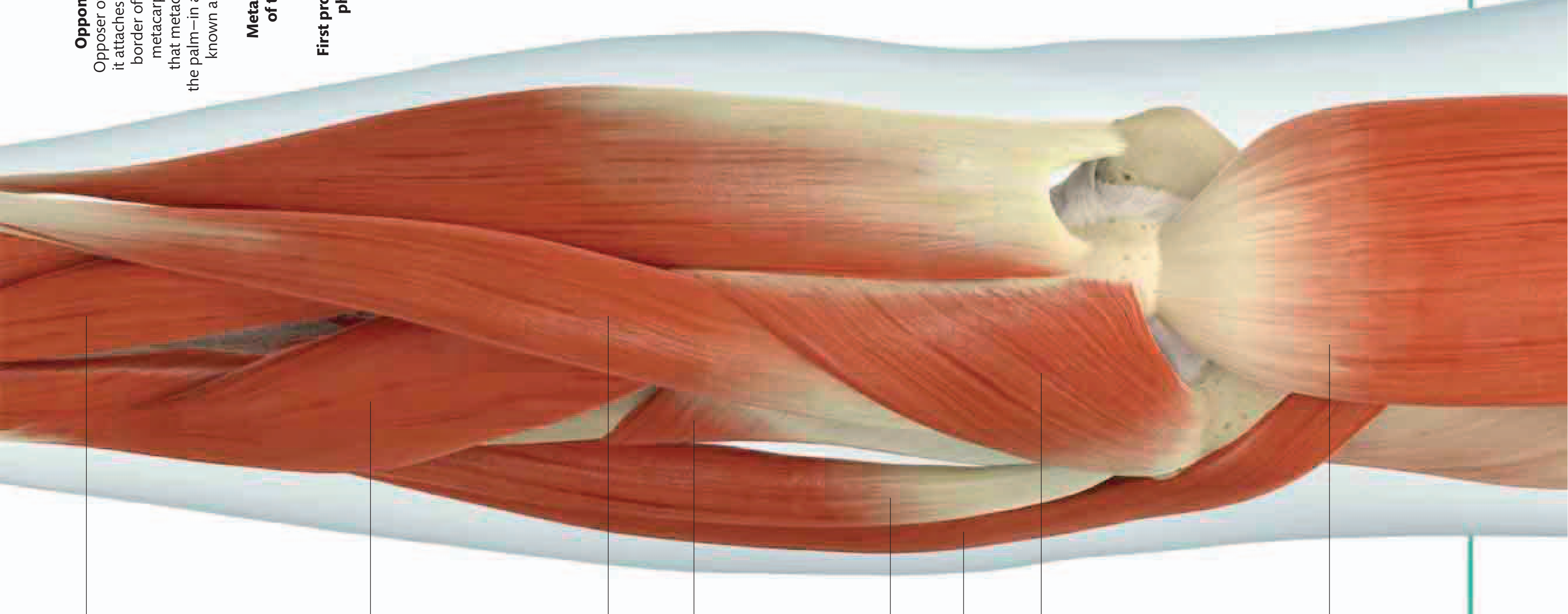

Extensor pollicis longus Long extensor of the thumb; it attaches onto the thumb's distal phalanx

Abductor pollicis longus Long abductor of the thumb; it attaches to the base of the thumb's metacarpal

Extensor carpi ulnaris

Supinator Arises from the lateral epicondyle of the humerus and wraps around the radius; it pulls the pronated forearm back into supination

Extensor carpi radialis brevis

Extensor carpi radialis longus

Anconeus

Triceps

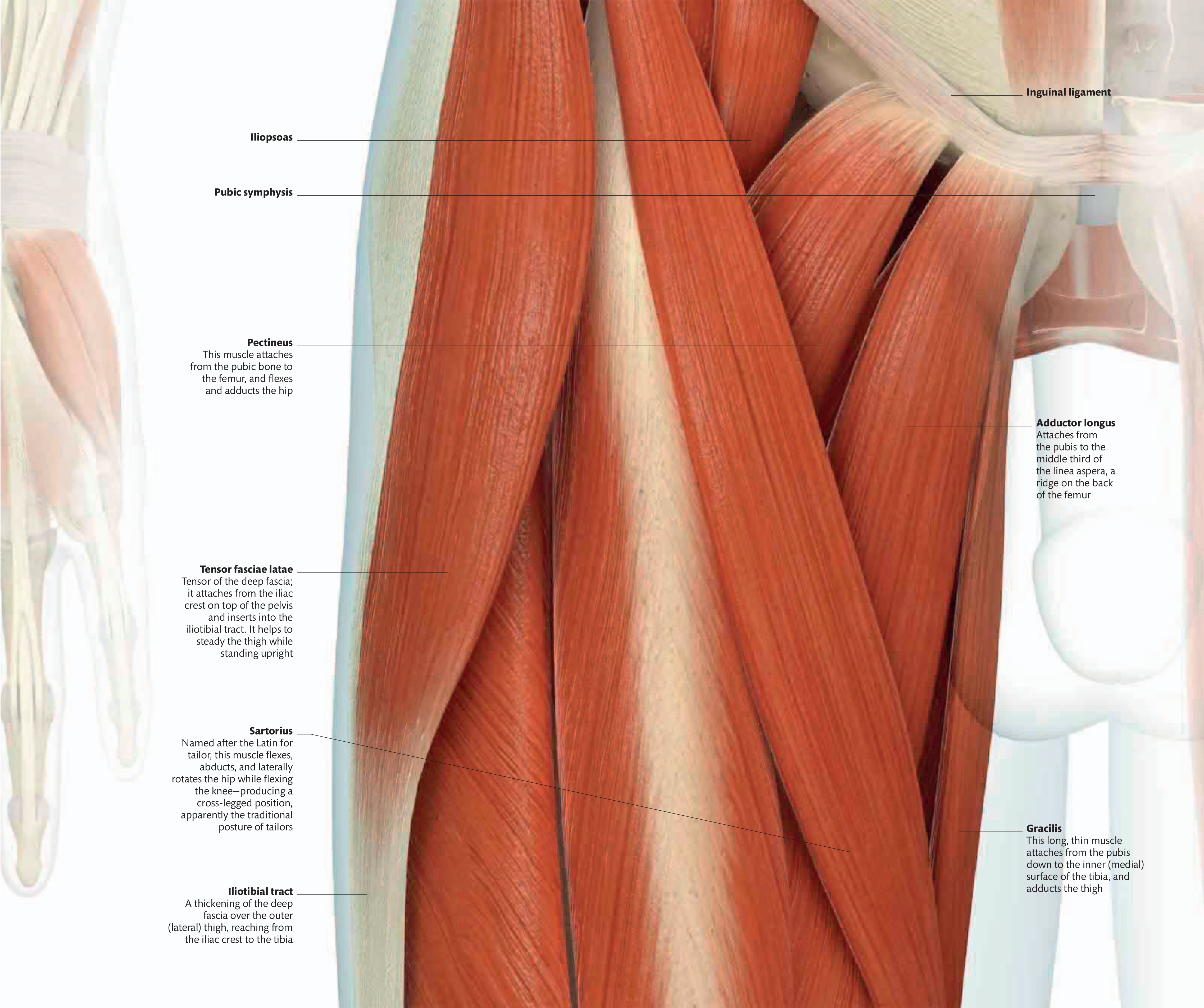

Inguinal ligament

Iliopsoas

Pubic symphysis

Pectineus
This muscle attaches from the pubic bone to the femur, and flexes and adducts the hip

Adductor longus
Attaches from the pubis to the middle third of the linea aspera, a ridge on the back of the femur

Tensor fasciae latae
Tensor of the deep fascia; it attaches from the iliac crest on top of the pelvis and inserts into the iliotibial tract. It helps to steady the thigh while standing upright

Sartorius
Named after the Latin for tailor, this muscle flexes, abducts, and laterally rotates the hip while flexing the knee—producing a cross-legged position, apparently the traditional posture of tailors

Gracilis
This long, thin muscle attaches from the pubis down to the inner (medial) surface of the tibia, and adducts the thigh

Iliotibial tract
A thickening of the deep fascia over the outer (lateral) thigh, reaching from the iliac crest to the tibia

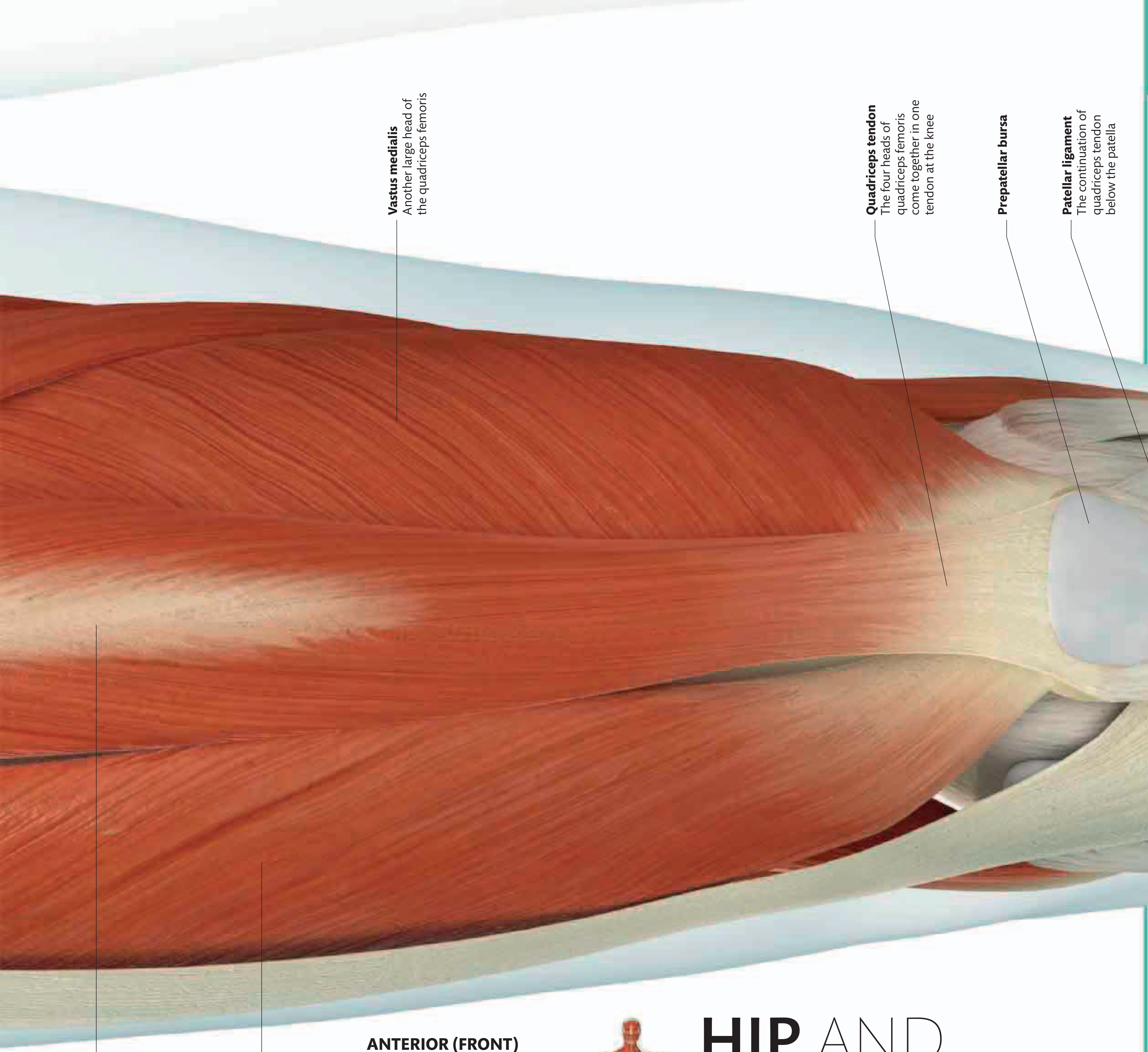
Vastus medialis Another large head of the quadriceps femoris
Quadriceps tendon The four heads of quadriceps femoris come together in one tendon at the knee
Prepatellar bursa
Patellar ligament The continuation of quadriceps tendon below the patella
Rectus femoris The part of the quadriceps that can flex the hip as well as extend the knee
Vastus lateralis The name of this part of the quadriceps reflects its impressive size

ANTERIOR (FRONT)
SUPERFICIAL

HIP AND THIGH

Most of the muscle bulk on the front of the leg is the four-headed quadriceps femoris. Three of its heads can be seen in a superficial dissection of the thigh: the rectus femoris, vastus lateralis, and vastus medialis. The quadriceps extends the knee, but it can also flex the hip, since the rectus femoris part has an attachment from the pelvis, above the hip joint. The patella is embedded in the quadriceps tendon; this may protect the tendon from wear and tear, but it also helps to give the quadriceps good leverage in extending the knee. The part of the tendon below the patella is usually called the patellar ligament. Tapping this with a tendon hammer produces a reflex contraction in the quadriceps—the “knee jerk”.

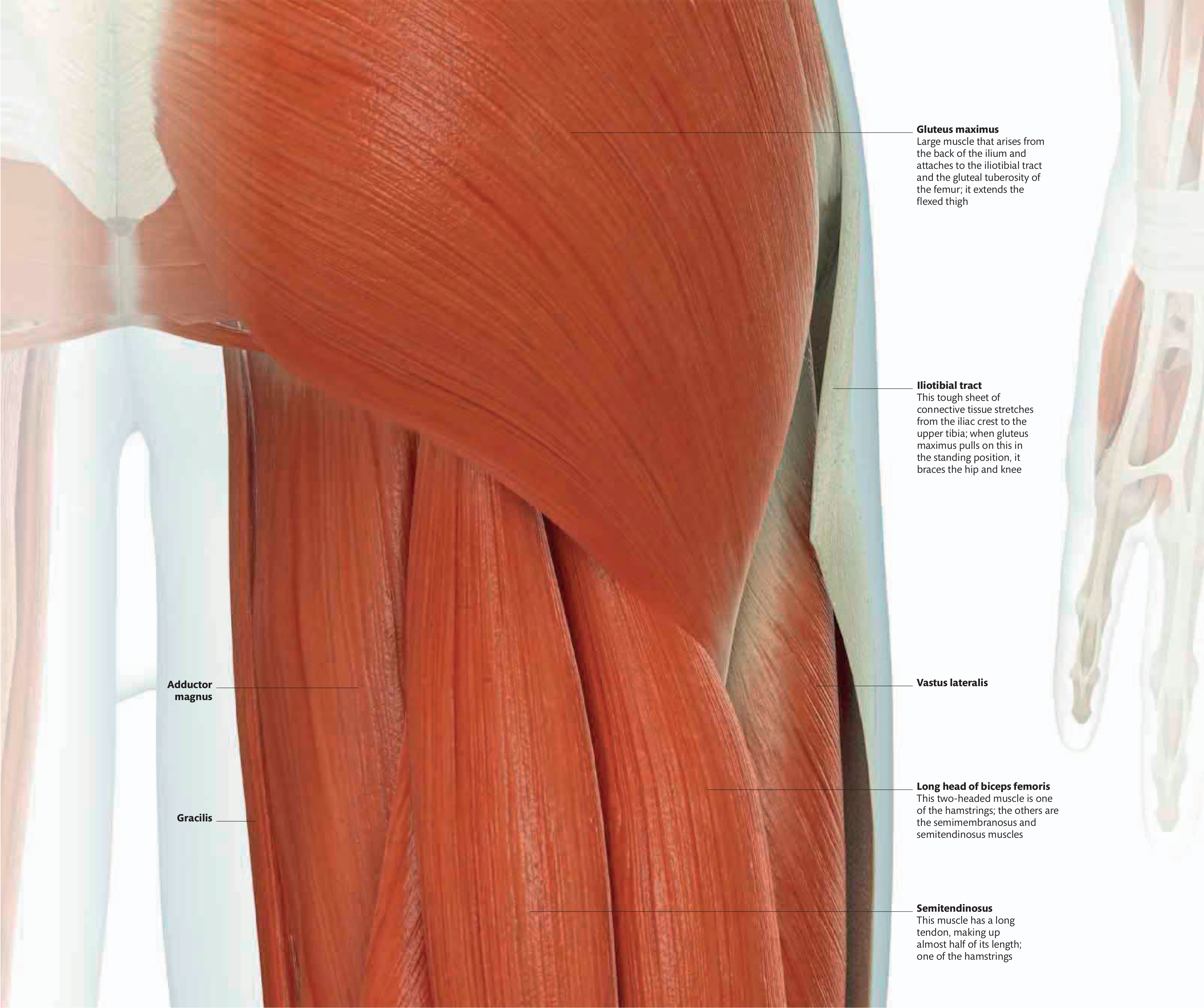

Gluteus maximus
Large muscle that arises from the back of the ilium and attaches to the iliotibial tract and the gluteal tuberosity of the femur; it extends the flexed thigh

Iliotibial tract
This tough sheet of connective tissue stretches from the iliac crest to the upper tibia; when gluteus maximus pulls on this in the standing position, it braces the hip and knee

Long head of biceps femoris
This two-headed muscle is one of the hamstrings; the others are the semimembranosus and semitendinosus muscles

Semitendinosus
This muscle has a long tendon, making up almost half of its length; one of the hamstrings

HIP AND THIGH

On the back of the hip and thigh, a superficial dissection reveals the large gluteus maximus, an extensor of the hip joint, and the three hamstrings. The gluteus maximus acts to extend the hip joint, swinging the leg backward. While it doesn't really contribute to gentle walking, it is very important in running, and also when the hip is being extended from a flexed position, such as when getting up from sitting on the floor or when climbing the stairs. The hamstrings—the semimembranosus, semitendinosus, and biceps femoris muscles—attach from the ischial tuberosity of the pelvis and sweep down the back of the thigh to the tibia and fibula. They are the main flexors of the knee.

POSTERIOR (BACK)
SUPERFICIAL

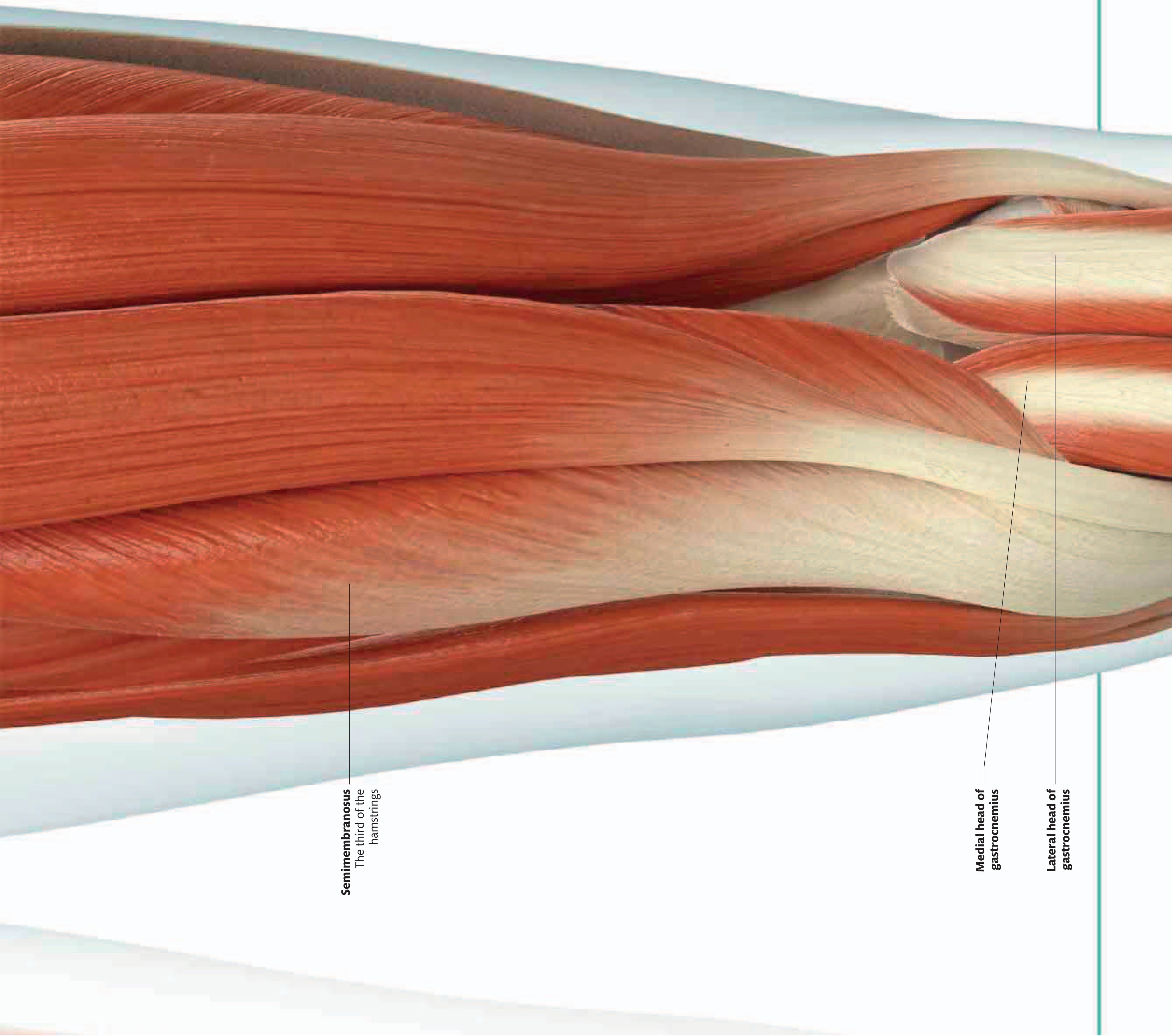
Semimembranosus
The third of the hamstrings
Medial head of gastrocnemius
Lateral head of gastrocnemius

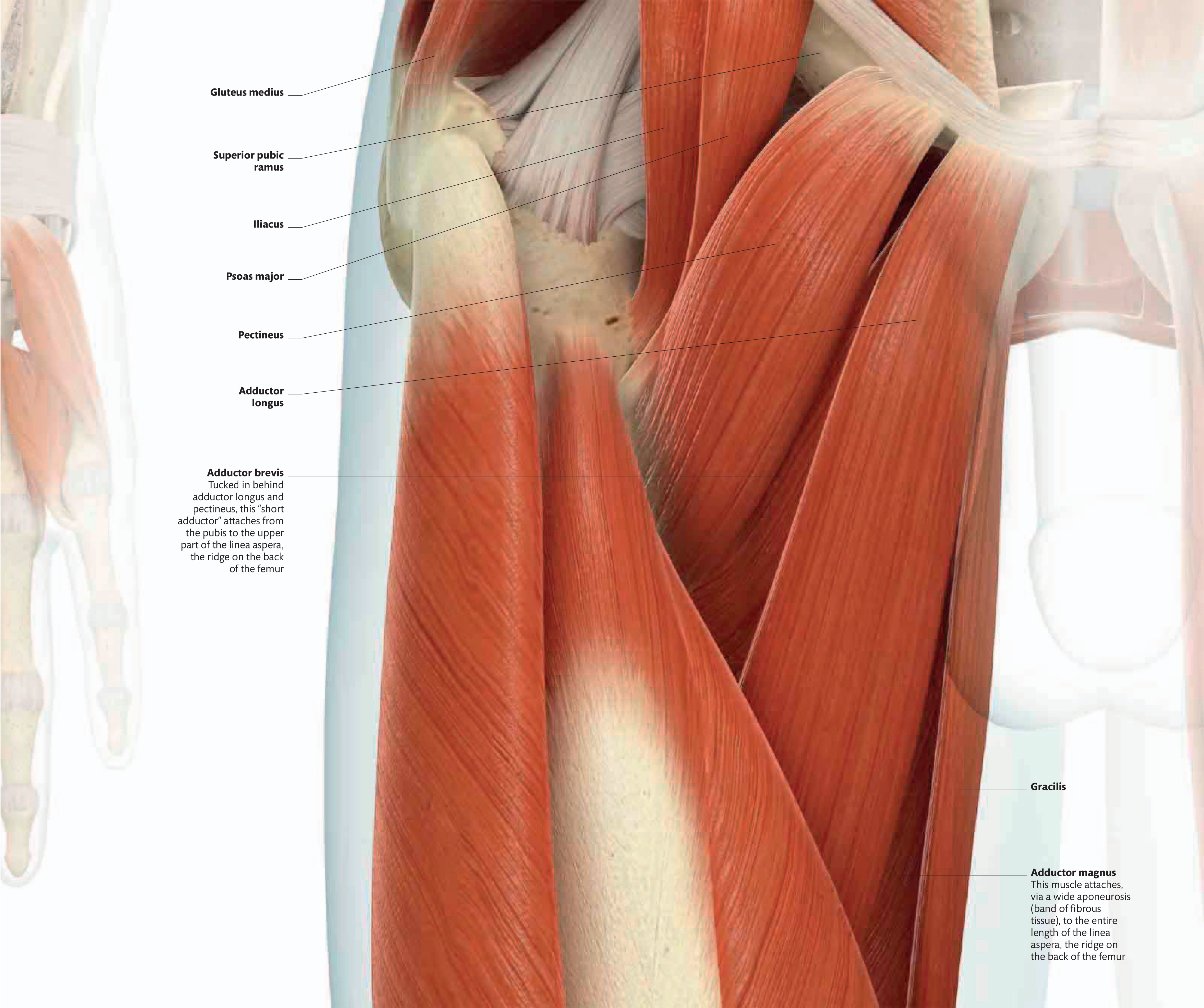
Gluteus medius
Superior pubic ramus
Iliacus
Psoas major
Pectineus
Adductor longus
Gracilis

Adductor brevis
Tucked in behind adductor longus and pectineus, this "short adductor" attaches from the pubis to the upper part of the linea aspera, the ridge on the back of the femur

Adductor magnus
This muscle attaches, via a wide aponeurosis (band of fibrous tissue), to the entire length of the linea aspera, the ridge on the back of the femur

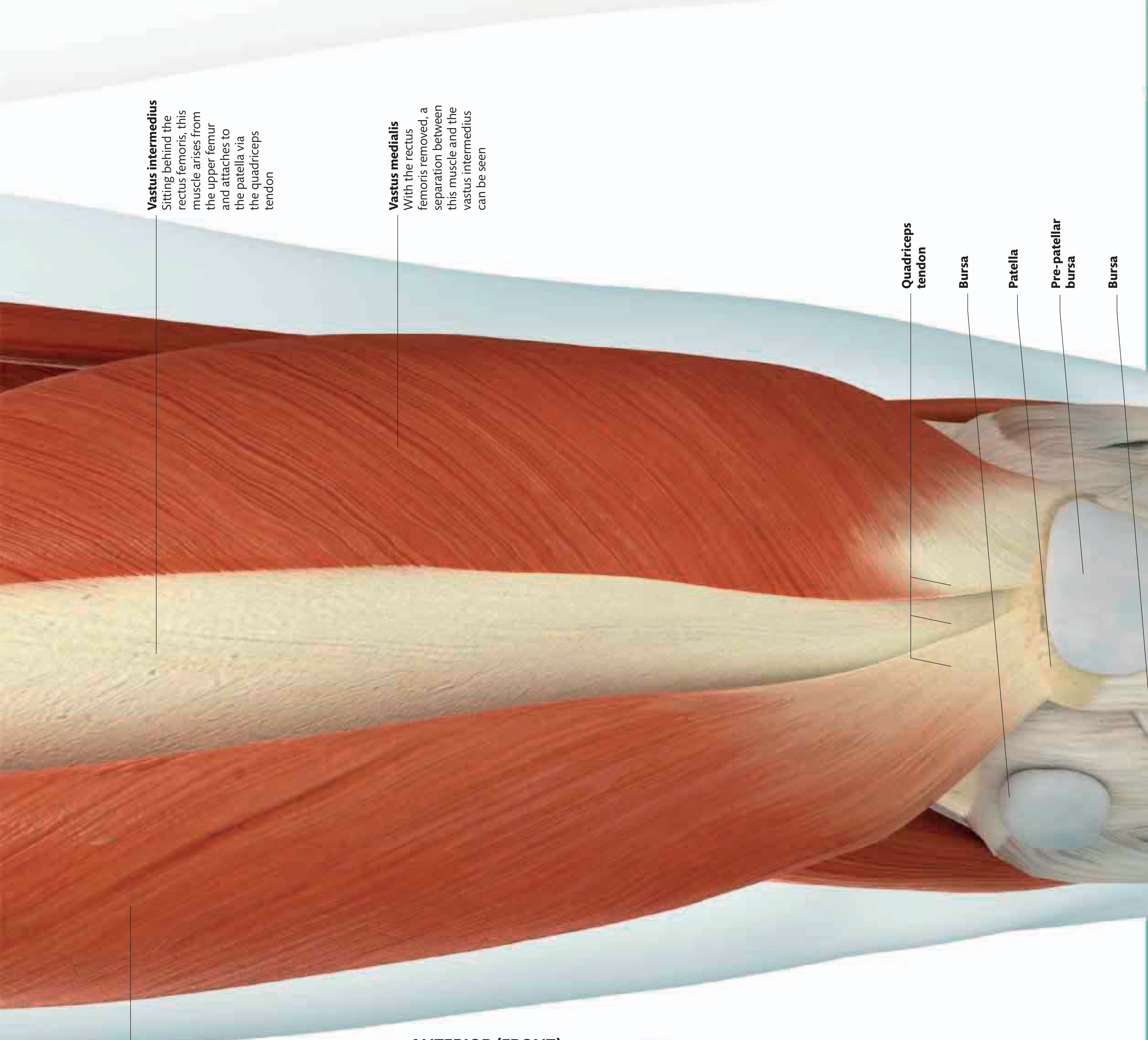
Vastus intermedius Sitting behind the rectus femoris, this muscle arises from the upper femur and attaches to the patella via the quadriceps tendon
Vastus medialis With the rectus femoris removed, a separation between this muscle and the vastus intermedius can be seen
Quadriceps tendon
Bursa
Patella
Pre-patellar bursa
Bursa
Vastus lateralis This muscle is the largest part of the quadriceps

ANTERIOR (FRONT)
DEEP

HIP AND THIGH

With the rectus femoris and sartorius muscles stripped away, we can see the deep, fourth head of the quadriceps, known as vastus intermedius. The adductor muscles that bring the thighs together can also be seen clearly, including the gracilis, which is long and slender, as its name suggests. The largest adductor muscle—the adductor magnus—has a hole in its tendon, through which the main artery of the leg (the femoral artery) passes. The adductor tendons attach from the pubis and ischium of the pelvis, and the sporting injuries referred to as "groin pulls" are often tears in these particular tendons.

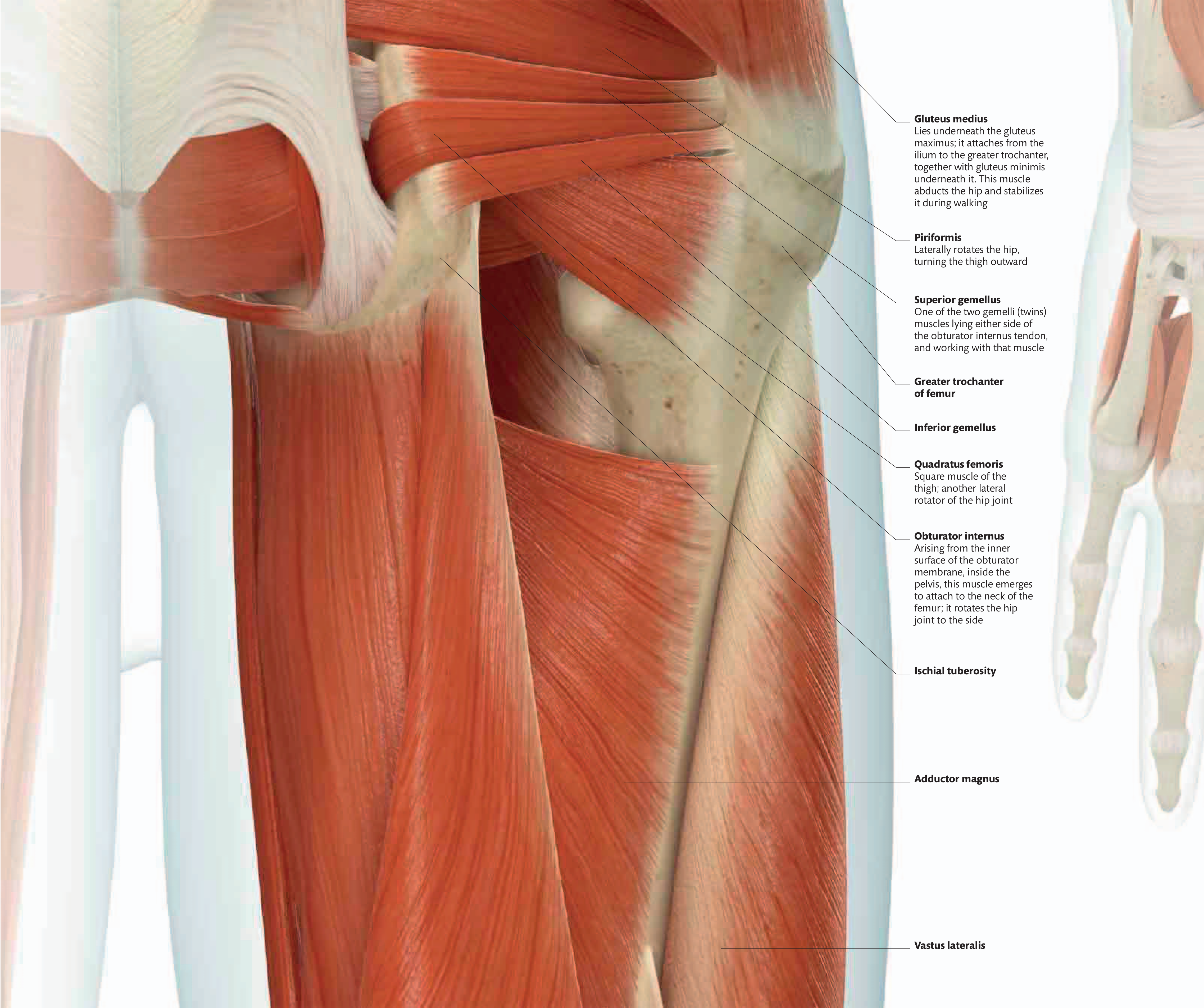

Gluteus medius
Lies underneath the gluteus maximus; it attaches from the ilium to the greater trochanter, together with gluteus minimis underneath it. This muscle abducts the hip and stabilizes it during walking

Piriformis
Laterally rotates the hip, turning the thigh outward

Superior gemellus
One of the two gemelli (twins) muscles lying either side of the obturator internus tendon, and working with that muscle

Greater trochanter of femur

Inferior gemellus

Quadratus femoris
Square muscle of the thigh; another lateral rotator of the hip joint

Obturator internus
Arising from the inner surface of the obturator membrane, inside the pelvis, this muscle emerges to attach to the neck of the femur; it rotates the hip joint to the side

Ischial tuberosity

Adductor magnus

Vastus lateralis

HIP AND THIGH

On the back of the hip, with the gluteus maximus removed, the short muscles that rotate the hip out to the side are clearly revealed. These include the piriformis, obturator internus, and quadratus femoris muscles. With the long head of the biceps femoris removed, we can now see the deeper, short head attaching to the linea aspera on the back of the femur. The semitendinosus muscle has also been cut away to reveal the semimembranosus underneath it, with its flat, membranelike tendon at the top. Popliteus muscle is also visible at the back of the knee joint, as is one of the many fluid filled bursae around the knee.

POSTERIOR (BACK)
DEEP

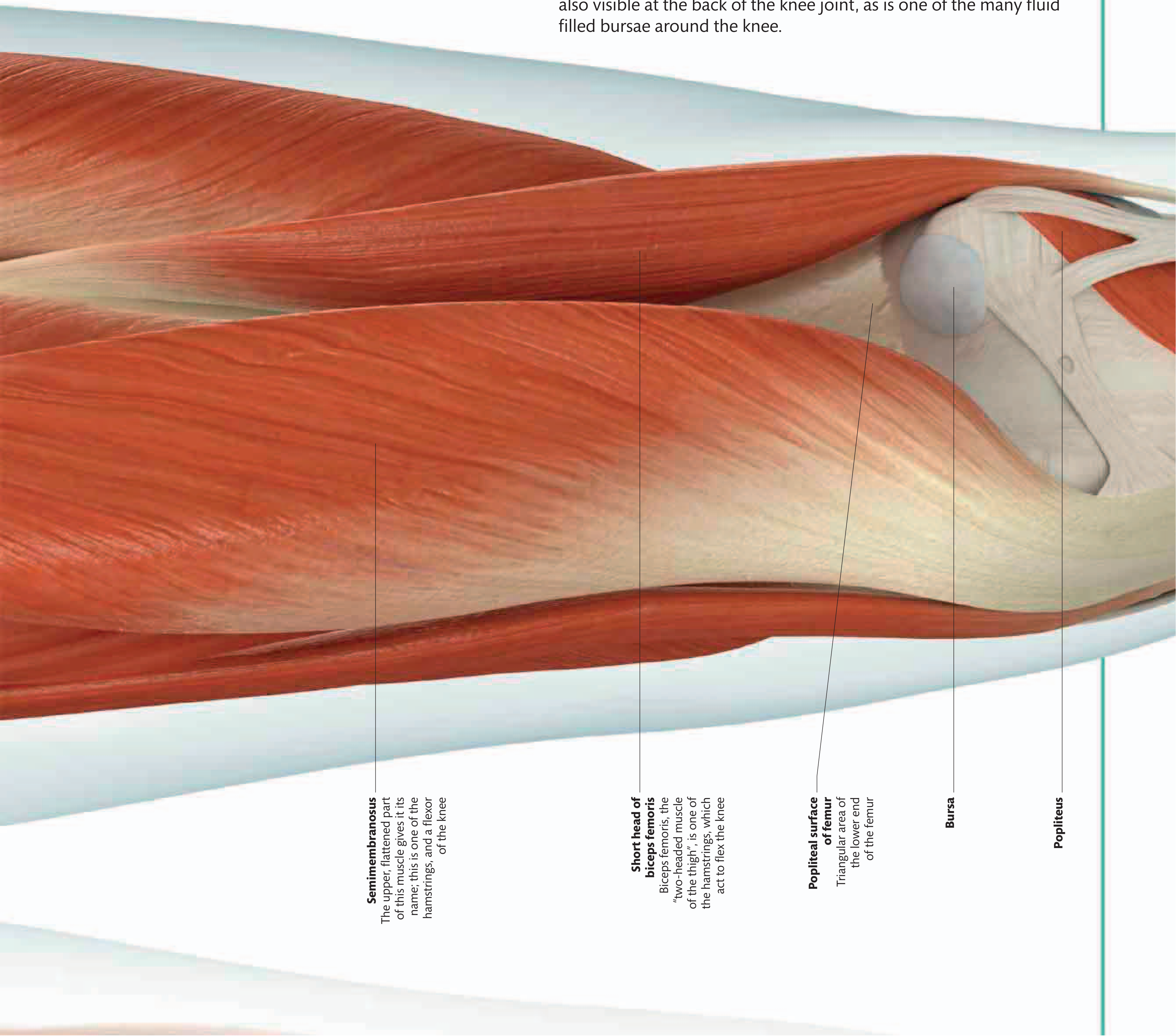
Semimembranosus The upper, flattened part of this muscle gives it its name; this is one of the hamstrings, and a flexor of the knee
Short head of biceps femoris Biceps femoris, the "two-headed muscle of the thigh", is one of the hamstrings, which act to flex the knee
Popliteal surface of femur Triangular area of the lower end of the femur
Bursa
Popliteus

POSTERIOR (BACK)
SUPERFICIAL

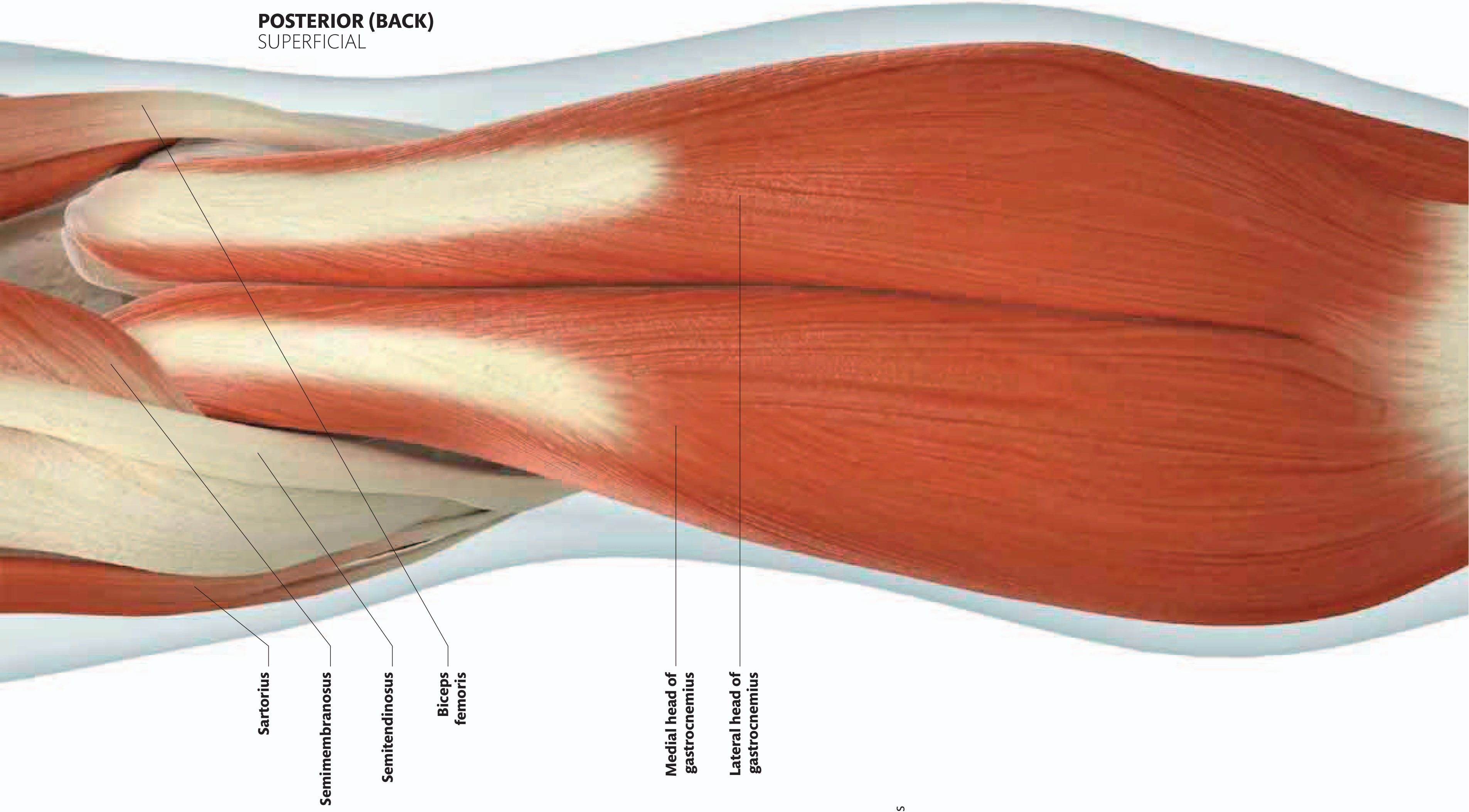
Sartorius
Semimembranosus
Semitendinosus
Biceps femoris
Medial head of gastrocnemius
Lateral head of gastrocnemius

ANTERIOR (FRONT)
SUPERFICIAL

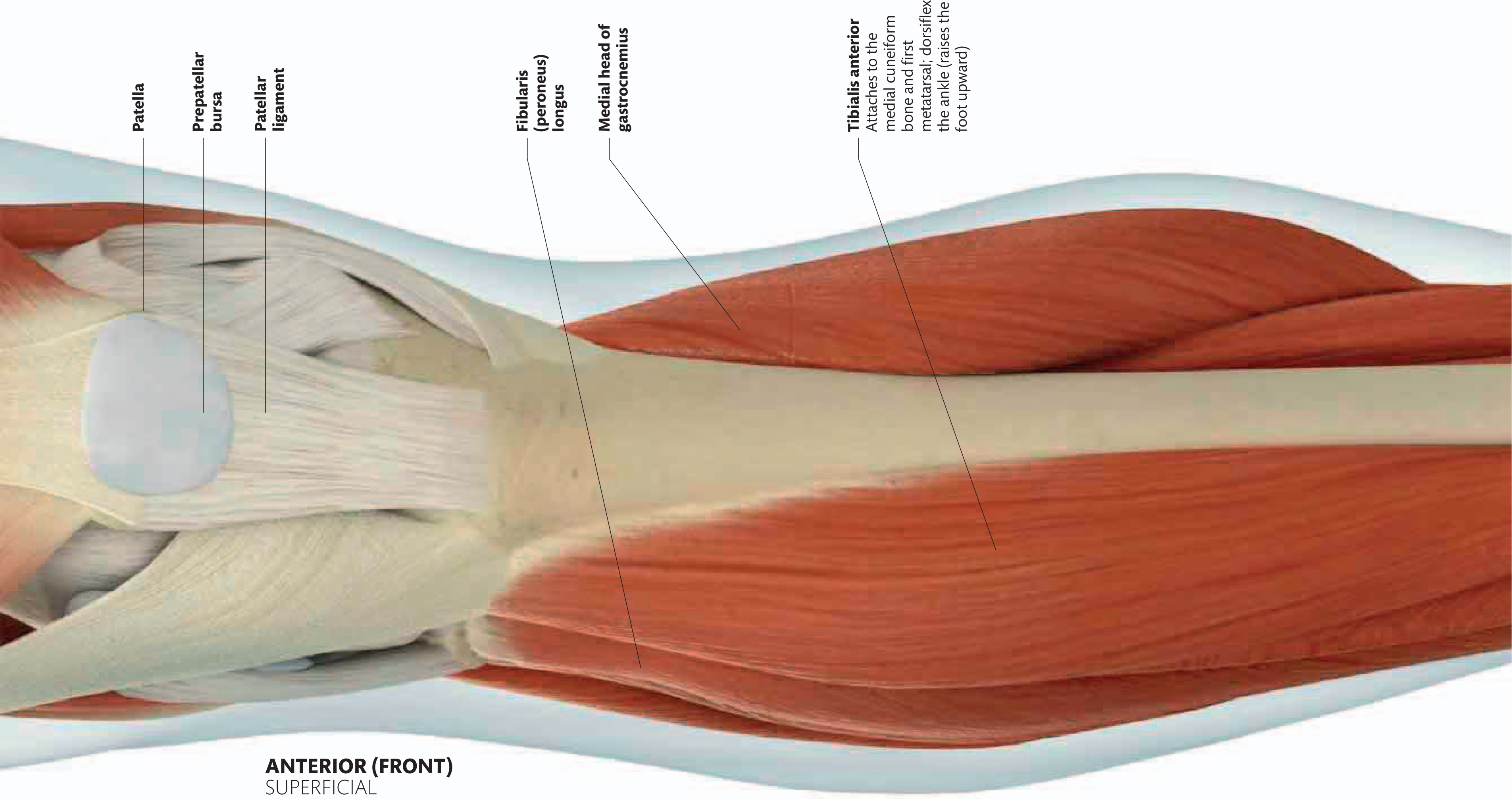
Patella
Prepatellar bursa
Patellar ligament
Fibularis (peroneus) longus
Medial head of gastrocnemius
Tibialis anterior

Tibialis anterior Attaches to the medial cuneiform bone and first metatarsal; dorsiflexes the ankle (raises the foot upward)

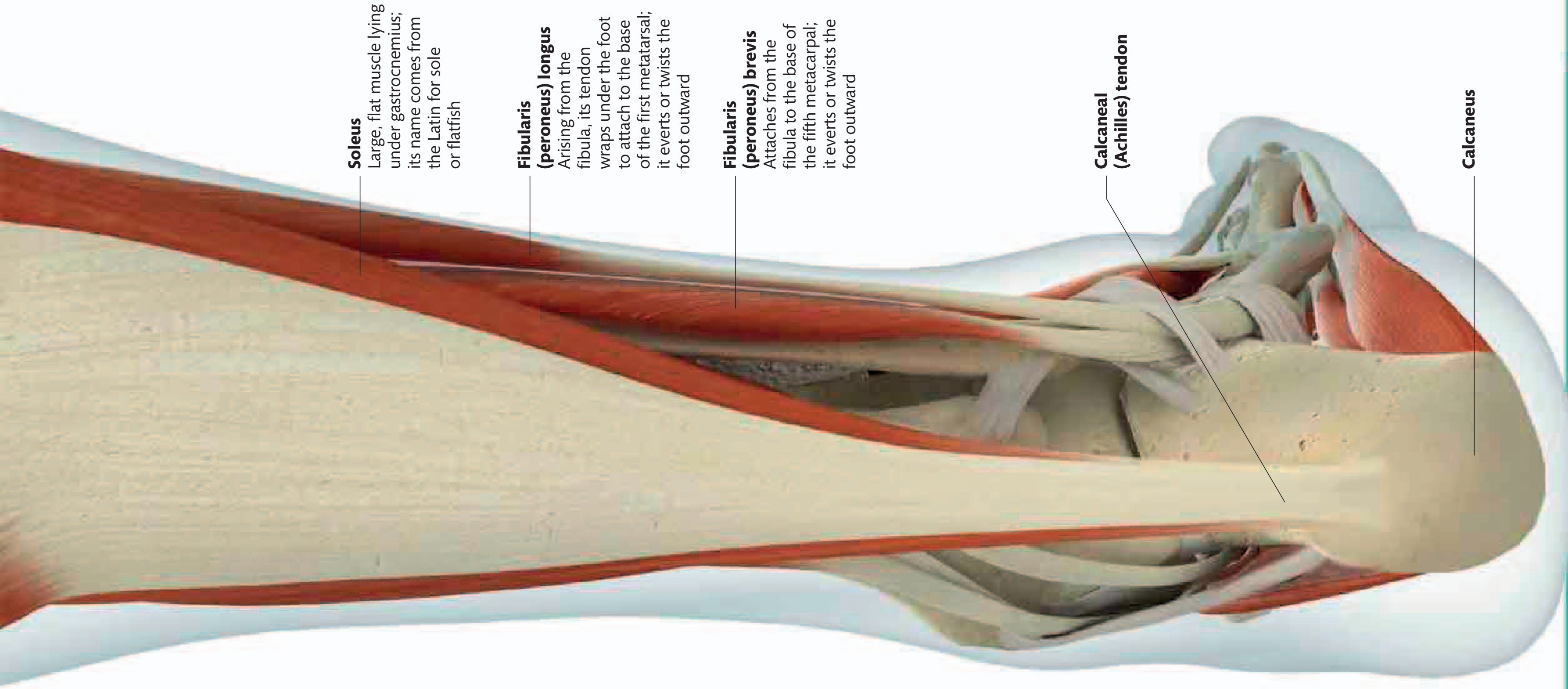
Soleus Large, flat muscle lying under gastrocnemius; its name comes from the Latin for sole or flatfish
Fibularis (peroneus) longus Arising from the fibula, its tendon wraps under the foot to attach to the base of the first metatarsal; it everts or twists the foot outward
Fibularis (peroneus) brevis Attaches from the fibula to the base of the fifth metacarpal; it everts or twists the foot outward
Calcaneal (Achilles) tendon
Calcaneus

LOWER LEG AND FOOT

You can feel the medial surface of the tibia easily, just under the skin on the front of your lower leg, on the inner side. Move your fingers outward, and you feel the sharp border of the bone, and then a soft wedge of muscles alongside it. These muscles have tendons that run down to the foot. They can pull the foot upward at the ankle, in a movement called dorsiflexion. Some extensor tendons continue all the way to the toes. There are much bulkier muscles on the back of the leg, and these form the calf. The gastrocnemius, and soleus underneath it, are large muscles that join together to form the Achilles tendon. They pull up on the lever of the calcaneus, pushing the ball of the foot down. They are involved as the foot pushes off from the ground during walking and running.

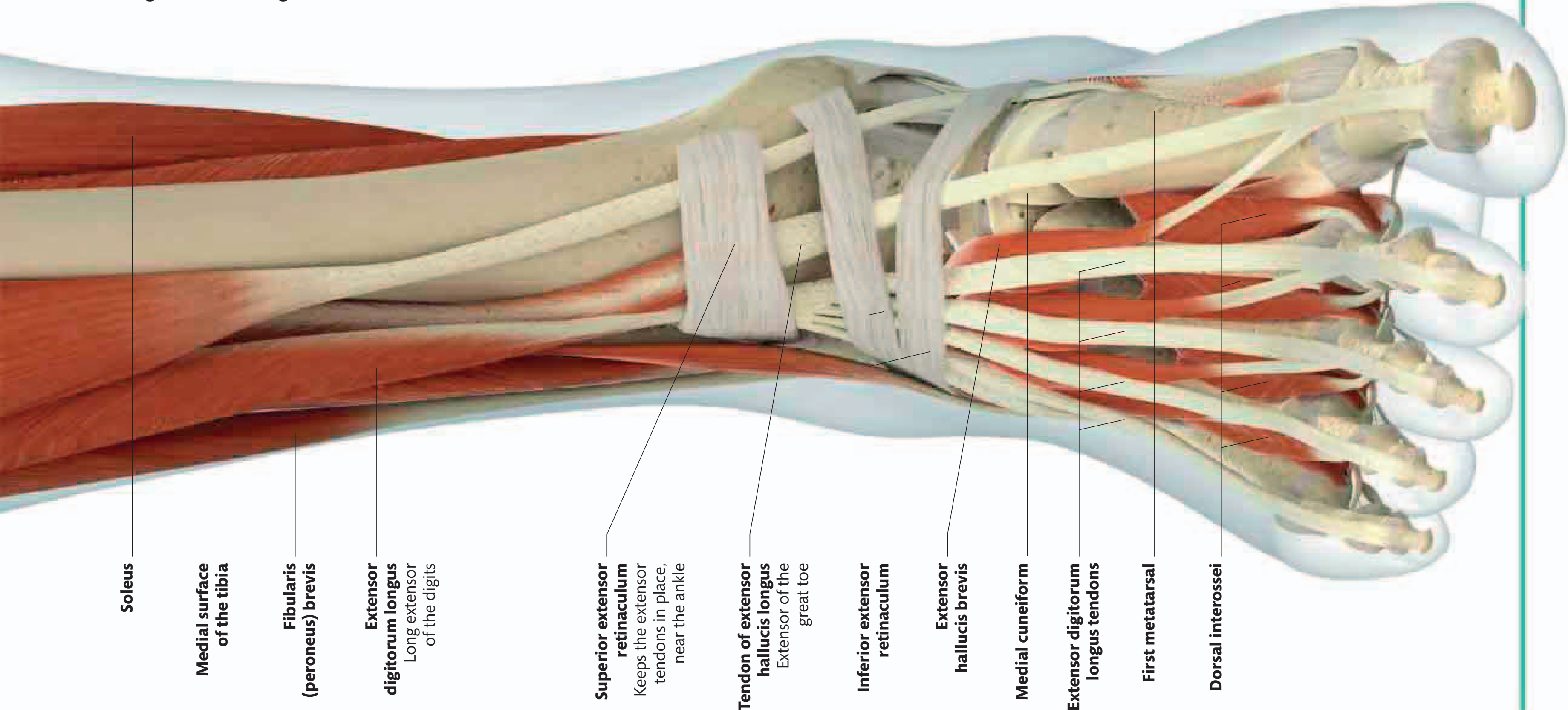
Soleus
Medial surface of the tibia
Fibularis (peroneus) brevis
Extensor digitorum longus Long extensor of the digits
Superior extensor retinaculum Keeps the extensor tendons in place, near the ankle
Tendon of extensor hallucis longus Extensor of the great toe
Inferior extensor retinaculum
Extensor hallucis brevis
Medial cuneiform
Extensor digitorum longus tendons
First metatarsal
Dorsal interossei

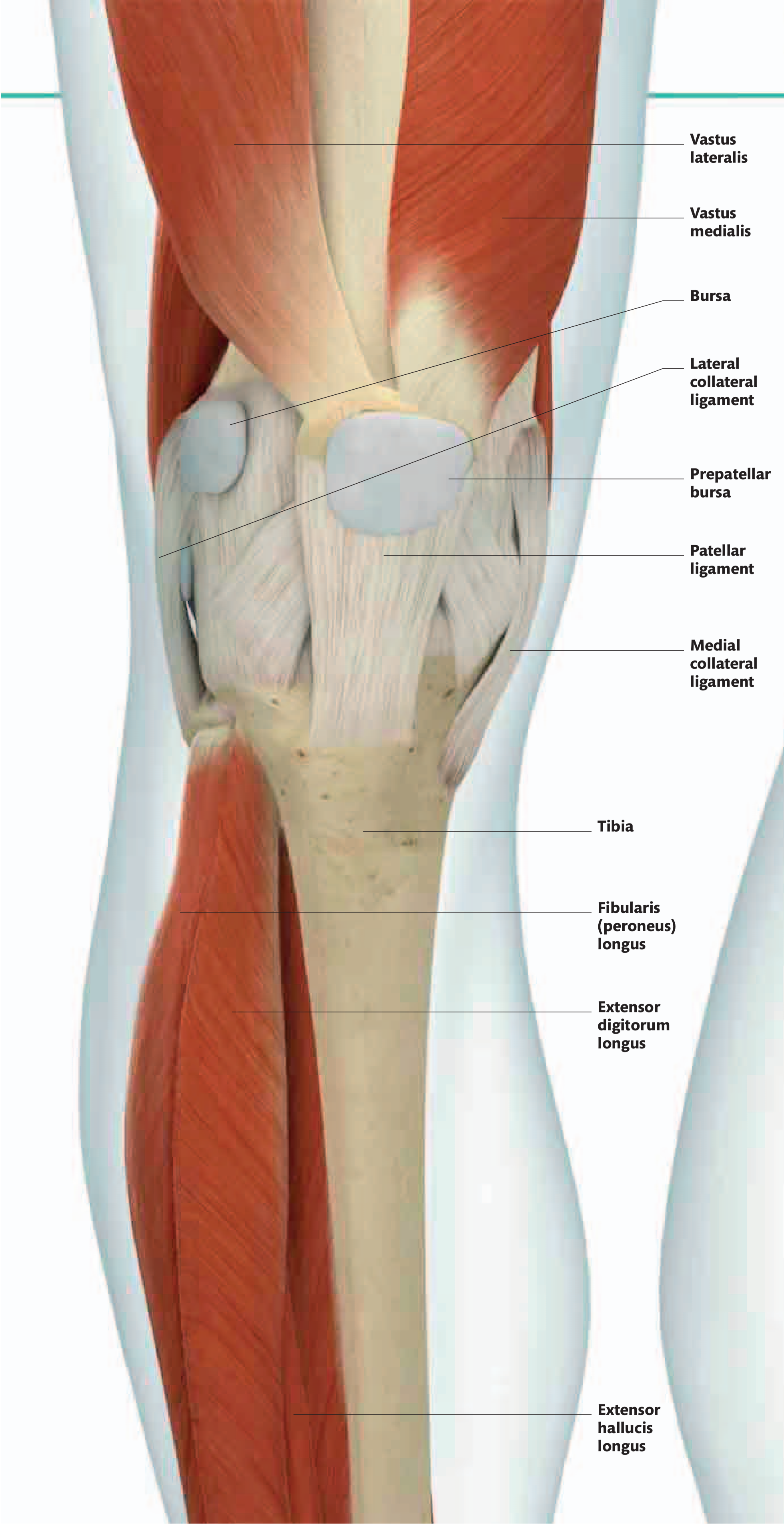
Vastus lateralis
Vastus medialis
Bursa
Lateral collateral ligament
Prepatellar bursa
Patellar ligament
Medial collateral ligament
Tibia
Fibularis (peroneus) longus
Extensor digitorum longus
Extensor hallucis longus

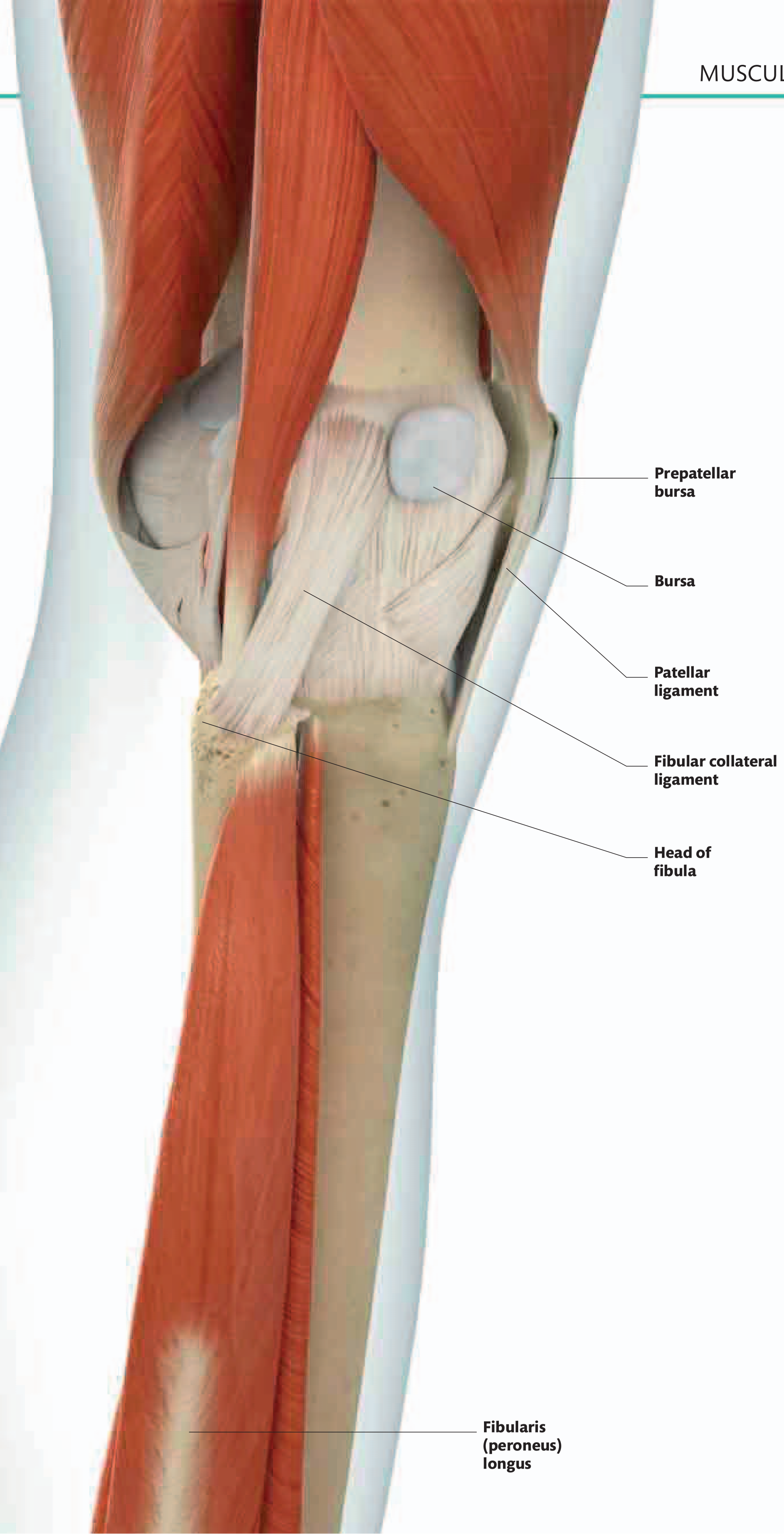
Prepatellar bursa
Bursa
Patellar ligament
Fibular collateral ligament
Head of fibula
Fibularis (peroneus) longus

LOWER LEG AND FOOT

Two muscles run along the outer, or lateral, side of the leg, down into the foot: the fibularis longus and fibularis brevis (see pp.104–05). These muscles pull the outer side of the foot upward, in a movement called eversion. The tendon of fibularis longus runs right underneath the foot, to attach on the inner side, and helps to maintain the transverse arch of the foot. The flexor hallucis longus arises from the fibula and interosseous membrane, and sends its tendon down, behind the medial malleolus and into the sole of the foot, to attach to the distal phalanx of the big toe.

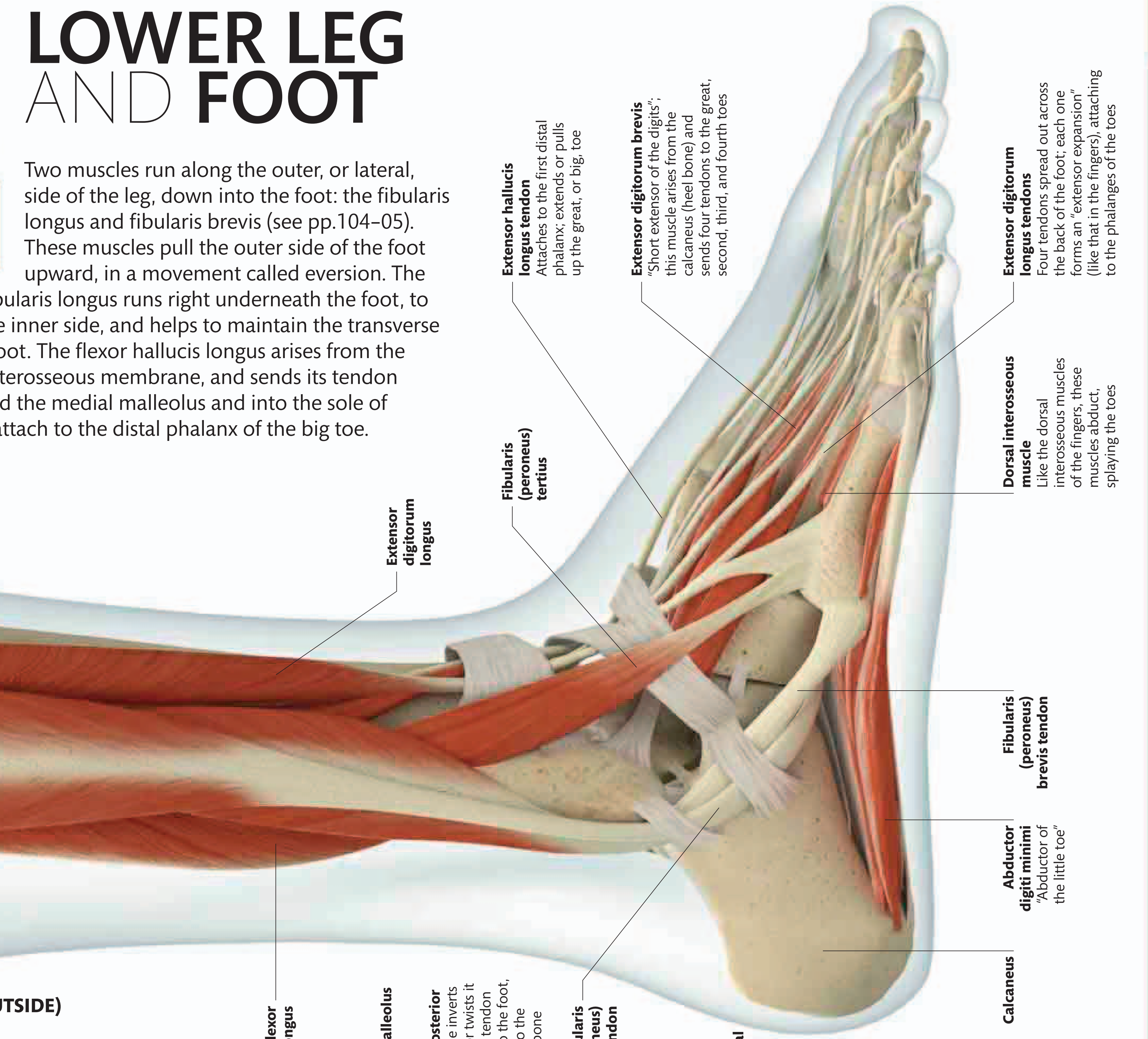
Extensor hallucis longus tendon
Attaches to the first distal phalanx; extends or pulls up the great, or big, toe
Extensor digitorum brevis
"Short extensor of the digits"; this muscle arises from the calcaneus (heel bone) and sends four tendons to the great, second, third, and fourth toes
Extensor digitorum longus tendons
Four tendons spread out across the back of the foot; each one forms an "extensor expansion" (like that in the fingers), attaching to the phalanges of the toes
Dorsal interosseous muscle
Like the dorsal interosseous muscles of the fingers, these muscles abduct, splaying the toes
Fibularis (peroneus) tertius
Extensor digitorum longus
Fibularis (peroneus) brevis tendon
Abductor digiti minimi
"Abductor of the little toe"
Calcaneus
Flexor hallucis longus
Fibularis (peroneus) longus tendon

LATERAL (OUTSIDE)
DEEP

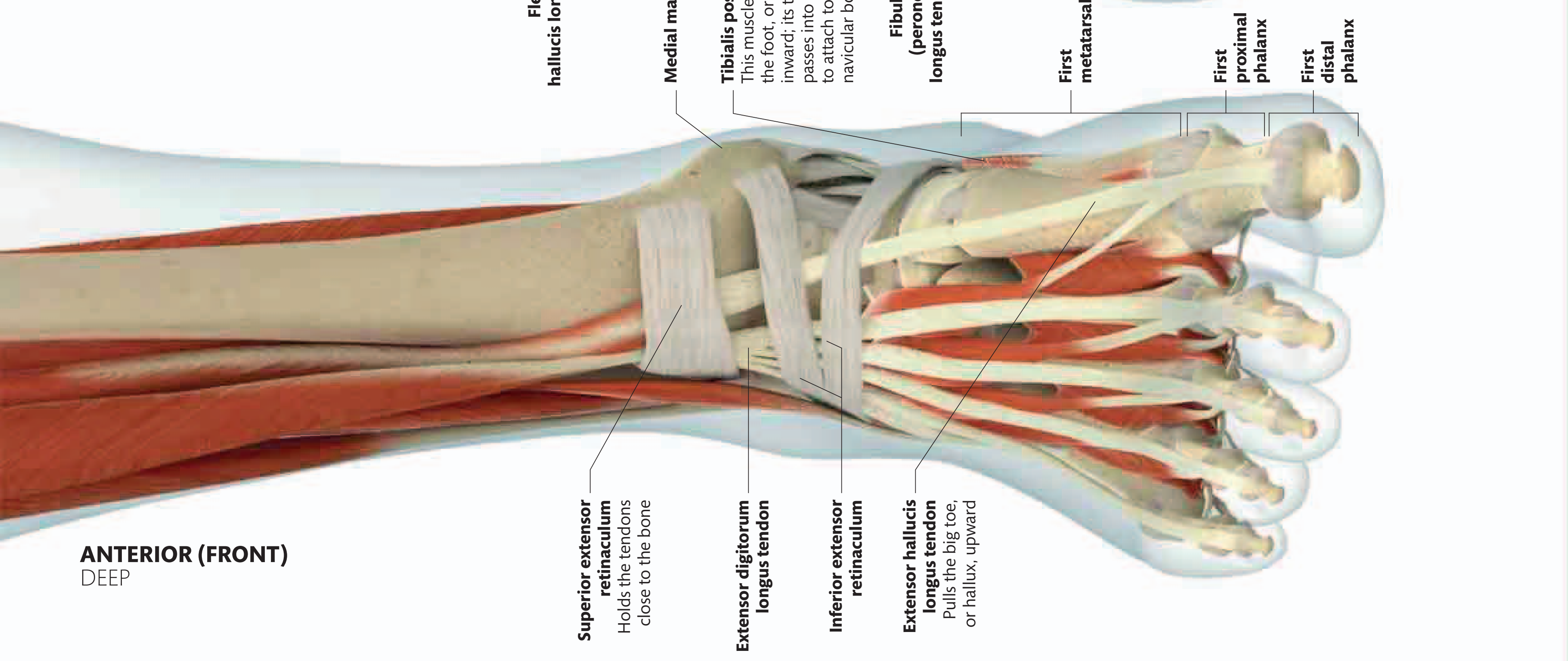
Medial malleolus
Tibialis posterior
This muscle inverts the foot, or twists it inward; its tendon passes into the foot, to attach to the navicular bone
First metatarsal
First proximal phalanx
First distal phalanx
Superior extensor retinaculum
Holds the tendons close to the bone
Extensor digitorum longus tendon
Inferior extensor retinaculum
Extensor hallucis longus tendon
Pulls the big toe, or hallux, upward

ANTERIOR (FRONT)
DEEP

NERVOUS SYSTEM OVERVIEW

The nervous system contains billions of intercommunicating nerve cells, or neurons. It can be broadly divided into the central nervous system (brain and spinal cord) and the peripheral nervous system (cranial and spinal nerves and their branches). The brain and spinal cord are protected by the skull and vertebral column respectively. Twelve cranial nerves emerge from the brain and exit through holes in the skull to supply the head and neck; thirty-one pairs of spinal nerves leave via gaps between vertebrae to supply the rest of the body. You can also divide the nervous system by function. The part that deals more with the way we sense and interact with our surroundings is called the somatic nervous system. The part involved with sensing and controlling our internal environments—affecting glands or heart rate, for example—is the autonomic nervous system.

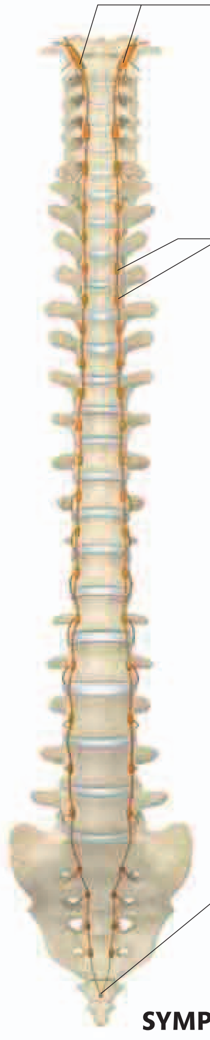

Sympathetic trunks
Part of the autonomic nervous system, the sympathetic trunks extend from the base of the skull to the end of the vertebral column, one on either side

Sympathetic ganglia
Collections of nerve cell bodies form ganglia along each trunk

Ganglion impar
The two sympathetic trunks converge and end in this single, unpaired ganglion, lying on the inner surface of the coccyx

SYMPATHETIC TRUNK

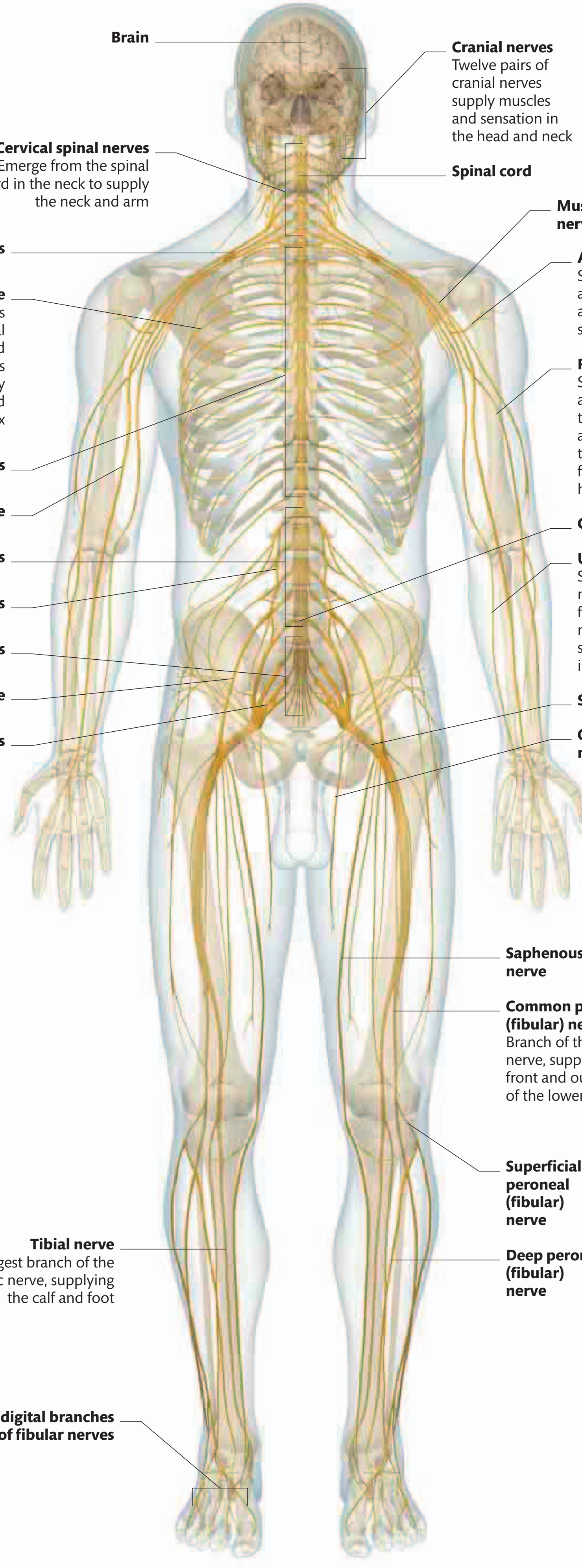

Brain

Cervical spinal nerves
Emerge from the spinal cord in the neck to supply the neck and arm

Brachial plexus

Intercostal nerve
Anterior branches of the thoracic spinal nerves travel forward between the ribs as intercostal nerves; they supply the muscles and skin of the thorax

Thoracic spinal nerves

Median nerve

Lumbar spinal nerves

Lumbar plexus

Sacral spinal nerves

Femoral nerve

Sacral plexus

Tibial nerve
Largest branch of the sciatic nerve, supplying the calf and foot

Dorsal digital branches of fibular nerves

Cranial nerves
Twelve pairs of cranial nerves supply muscles and sensation in the head and neck

Spinal cord

Musculocutaneous nerve

Axillary nerve
Supplies muscles and sensation around the shoulder

Radial nerve
Supplies muscles and sensation on the back of the arm (including the triceps, forearm, and hand)

Cauda equina

Ulnar nerve
Supplies two muscles in the forearm, and many of the small muscles in the hand

Sciatic nerve

Obturator nerve

Saphenous nerve

Common peroneal (fibular) nerve
Branch of the sciatic nerve, supplying the front and outer side of the lower leg

Superficial peroneal (fibular) nerve

Deep peroneal (fibular) nerve

ANTERIOR (FRONT)

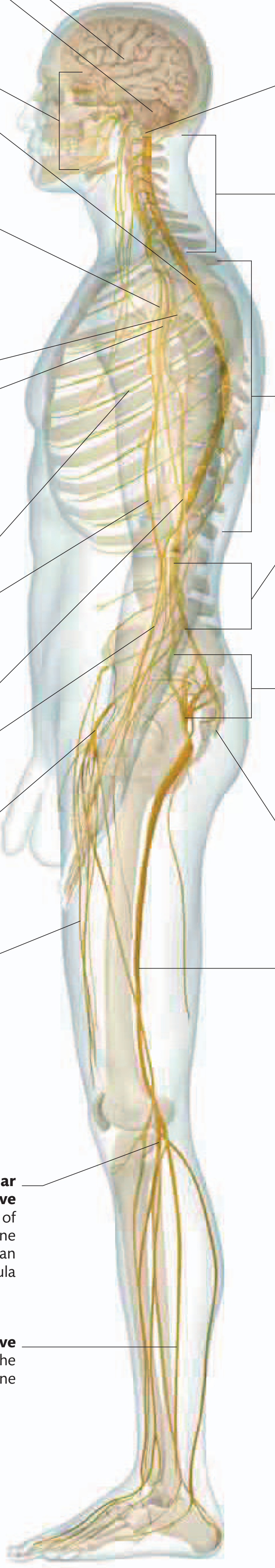

Cerebrum

Cerebellum
Literally little brain in Latin, this part of the brain is involved with balance and coordination of movement

Cranial nerves

Spinal cord
The continuation of the brain stem, lying protected within the vertebral canal of the spine

Musculocutaneous nerve
Supplies the muscles in the front of the upper arm, (including the biceps), as well as sensation to the skin of the outer side of the forearm

Axillary nerve

Brachial plexus
Anterior branches of the lower cervical spinal nerves, together with the first thoracic spinal nerve, form a network, or plexus, from which branches emerge to supply the arm, forearm, and hand

Intercostal nerve

Median nerve
Supplies most of the muscles in the front of the forearm, and also some in the hand

Radial nerve

Ulnar nerve
This nerve lies on the ulnar, or inner, side of the arm and forearm

Femoral nerve
Supplies sensation over the thigh and inner leg, and muscles in the front of the thigh, including the quadriceps

Obturator nerve
Supplies the muscles and skin of the inner thigh

Common fibular (peroneal) nerve
Lies on the outer side of the leg and is named after the bone around which it wraps; perona is an alternative Latin name for fibula

Tibial nerve
Named after the other bone of the lower leg—the tibia, or shinbone

Brainstem
Emerges from the foramen magnum in the base of the skull

Cervical spinal nerves
Cervical means of the neck; cervix is Latin for neck

Thoracic spinal nerves
Thorax is Latin for chest so the term thoracic means of the chest

Lumbar plexus
Anterior branches of the lumbar spinal nerves form a network here, from which nerves emerge to supply the leg

Sacral plexus
Anterior branches of sacral spinal nerves come together here as a network; the network provides nerves to the buttock and leg

Cauda equina
Below the end of the spinal cord, the lumbar and sacral nerve roots continue for some way inside the vertebral canal, before emerging from the spine

Sciatic nerve
Largest nerve in the body, which supplies the hamstrings in the back of the thigh; its branches supply muscles and sensation in the lower leg and foot

SIDE

NEURON STRUCTURE

A single neuron such as the cell shown below from the central nervous system can make contact with hundreds of other neurons, creating an incredibly complex network of connections. Each neuron's cell body has projections or, dendrites. One is usually longer and thinner than the rest, and this is the axon. Some axons within the brain are less than 1⁄32in (1mm) in length; others, stretching from the spinal cord to limb muscles, can measure over 39in (1m) long.

NEURON

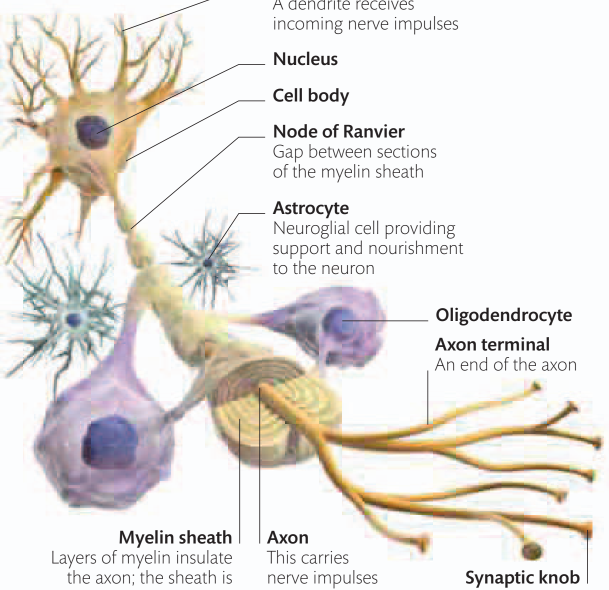

Dendrite
A dendrite receives incoming nerve impulses

Nucleus

Cell body

Node of Ranvier
Gap between sections of the myelin sheath

Astrocyte
Neuroglial cell providing support and nourishment to the neuron

Oligodendrocyte

Axon terminal
An end of the axon

Myelin sheath
Layers of myelin insulate the axon; the sheath is made of fat-filled cells wrapped around the axon

Axon
This carries nerve impulses away from the cell body

Synaptic knob
Transmits the impulse via a synapse

BRAIN

Compared to other animals, humans have massive brains for the size of our bodies. The human brain has grown larger and larger over the course of evolution, and it is now so overblown that the frontal lobes of the brain lie right over the top of the orbits that contain the eyes. Think about any other mammal, perhaps a dog or a cat for easy reference, and you will quickly realize what an odd shape the human head is—and most of that is a result of our huge brains. Looking at a side view of the brain, you can see all the lobes that make up each cerebral hemisphere: the frontal, parietal, temporal, and occipital lobes (individually colored, below). Tucked under the cerebral hemispheres at the back of the brain is the cerebellum (Latin for little brain). The brain stem leads down, through the foramen magnum of the skull, to the spinal cord.

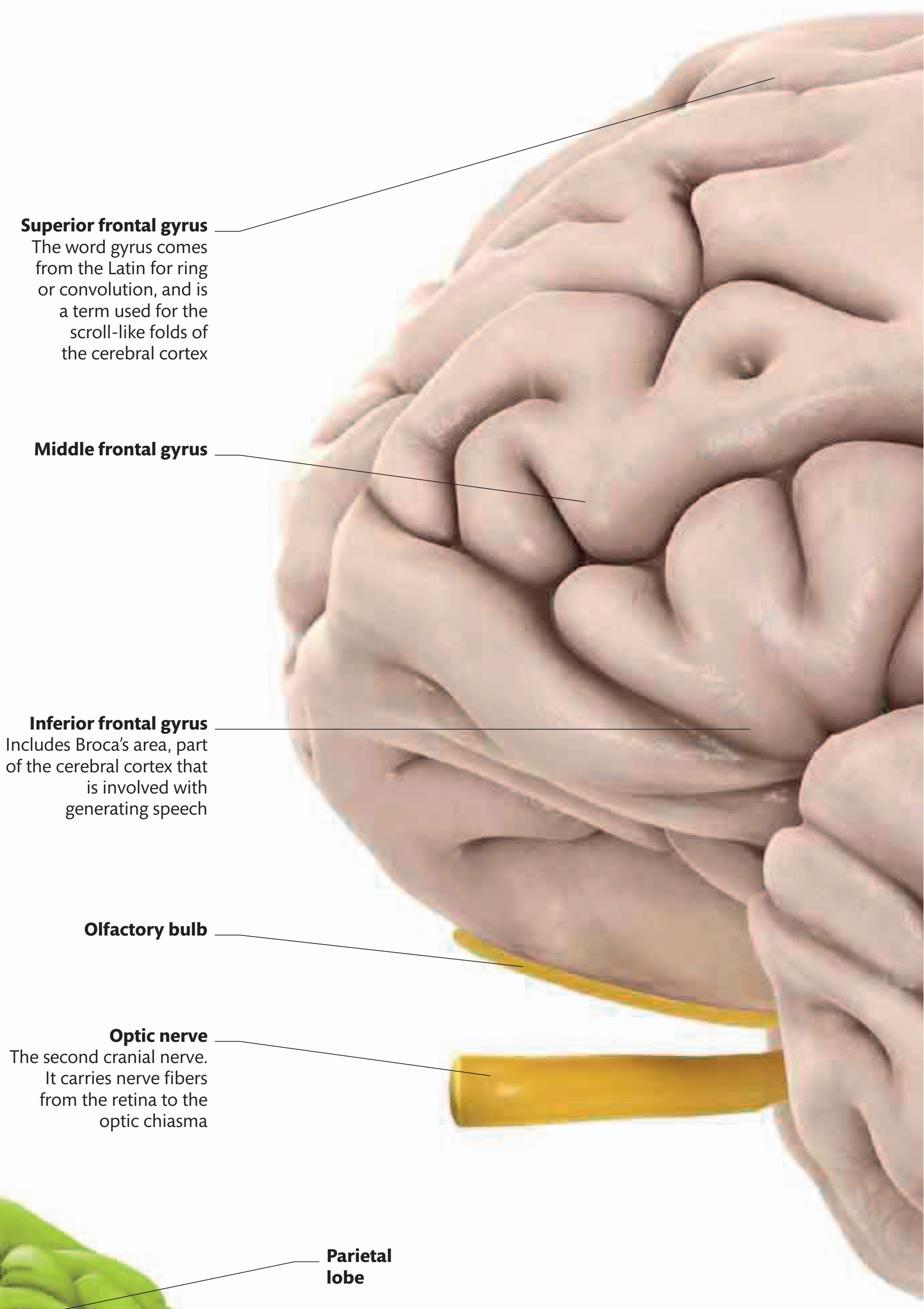

Superior frontal gyrus
The word gyrus comes from the Latin for ring or convolution, and is a term used for the scroll-like folds of the cerebral cortex

Middle frontal gyrus

Inferior frontal gyrus
Includes Broca's area, part of the cerebral cortex that is involved with generating speech

Olfactory bulb

Optic nerve
The second cranial nerve. It carries nerve fibers from the retina to the optic chiasma

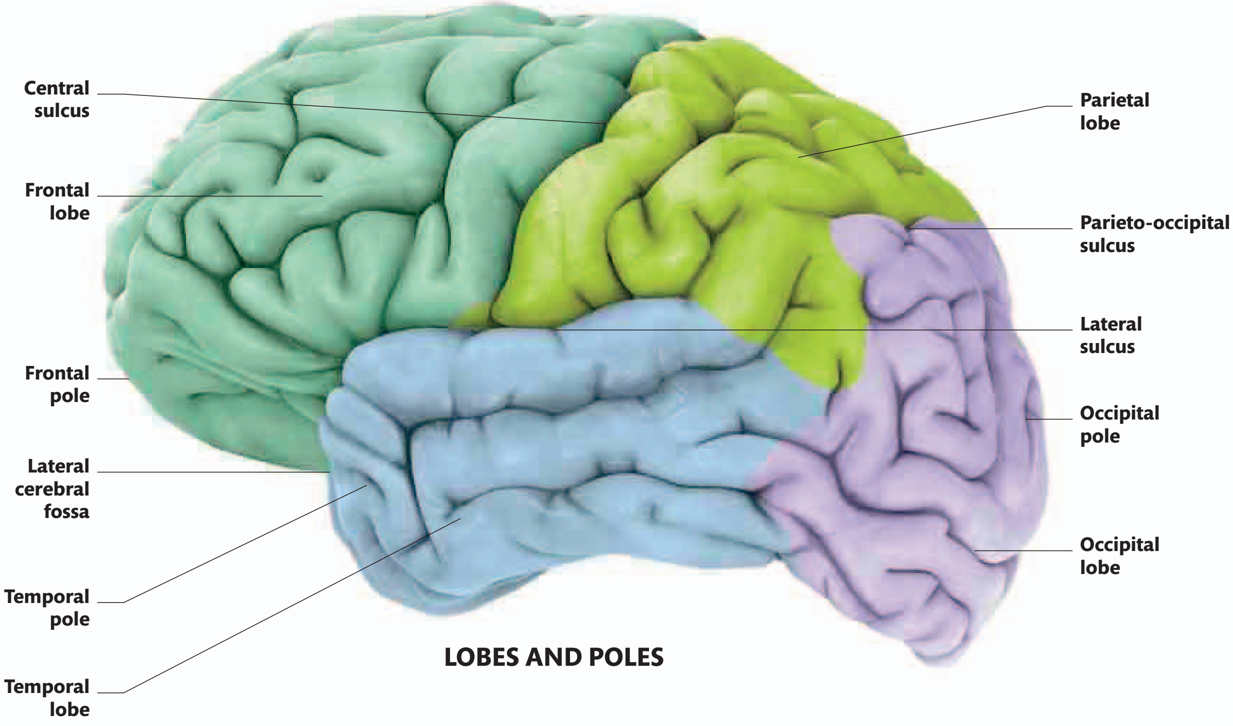
Central sulcus
Frontal lobe
Frontal pole
Lateral cerebral fossa
Temporal pole
Temporal lobe
Parietal lobe
Parieto-occipital sulcus
Lateral sulcus
Occipital pole
Occipital lobe

LOBES AND POLES

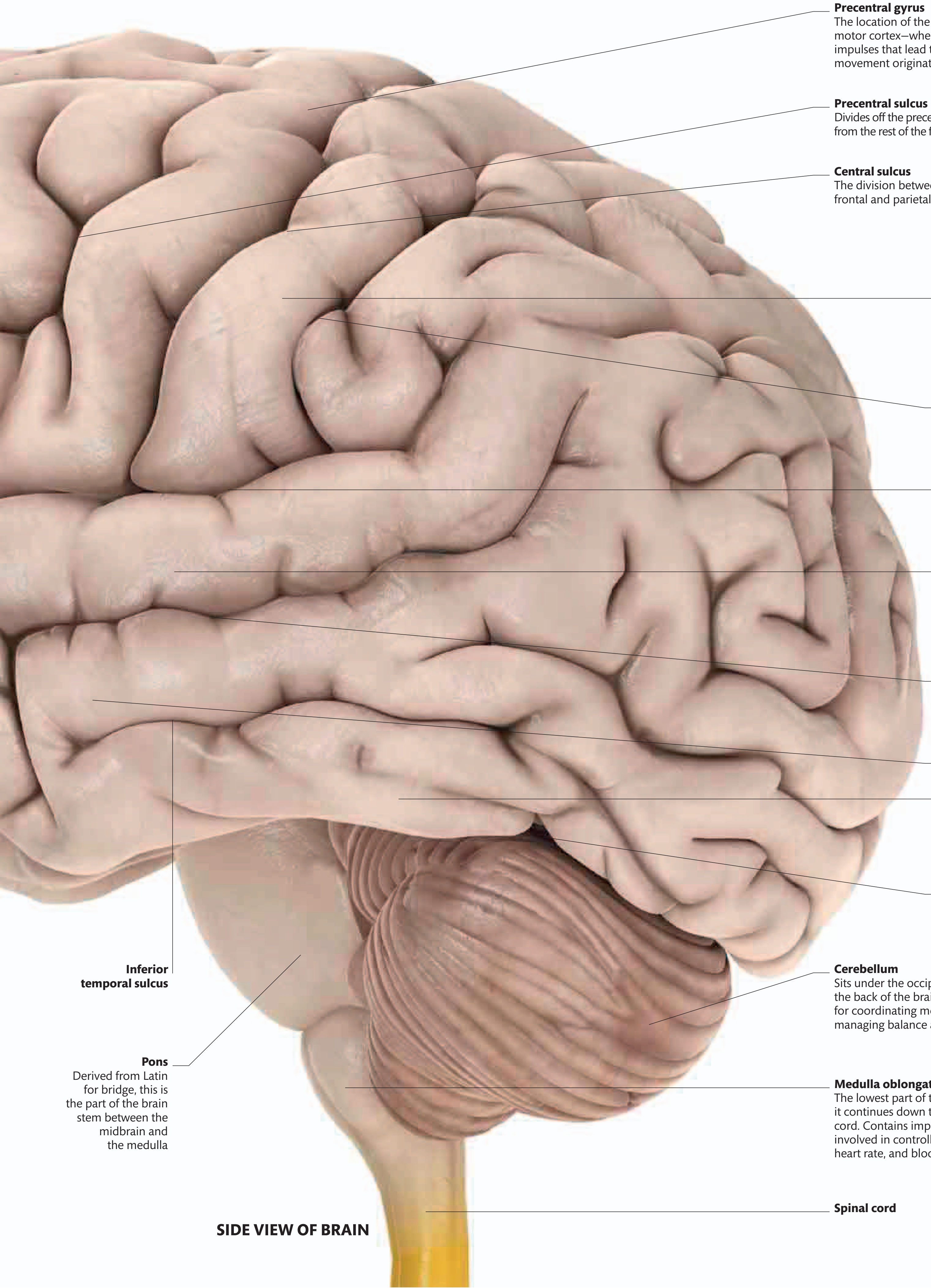

Precentral gyrus
The location of the primary motor cortex—where nerve impulses that lead to muscle movement originate

Precentral sulcus
Divides off the precentral gyrus from the rest of the frontal lobe

Central sulcus
The division between the frontal and parietal lobes

Postcentral gyrus
Lies just behind the central sulcus. The primary somatosensory cortex, which receives sensory information from all over the body

Postcentral sulcus
Separates the postcentral gyrus from the rest of the parietal lobe

Lateral sulcus
A deep cleft dividing the frontal and parietal lobes from the temporal lobe below

Superior temporal gyrus
Includes the primary auditory cortex, where sensory information related to hearing is received

Superior temporal sulcus
Sulcus is a Latin word meaning groove or furrow

Middle temporal gyrus

Inferior temporal gyrus

Preoccipital notch

Inferior temporal sulcus

Cerebellum
Sits under the occipital lobes at the back of the brain; responsible for coordinating movement and managing balance and posture

Pons
Derived from Latin for bridge, this is the part of the brain stem between the midbrain and the medulla

Medulla oblongata
The lowest part of the brain stem; it continues down to form the spinal cord. Contains important centers involved in controlling breathing, heart rate, and blood pressure

Spinal cord

SIDE VIEW OF BRAIN

BRAIN

From an anatomist's point of view, the brain is quite an ugly and unprepossessing organ. It looks rather like a large, pinkish gray, wrinkled walnut—especially when viewed from above. The outer layer of gray matter, called the cortex, is highly folded. Underneath the brain we see some more detail, including some of the cranial nerves that emerge from the brain itself. To the naked eye, there is little to suggest that the brain is the most complicated organ in the human body. Its true complexity is only visible through a microscope, revealing billions of neurons that connect with each other to form the pathways that carry our senses, govern our actions, and create our minds.

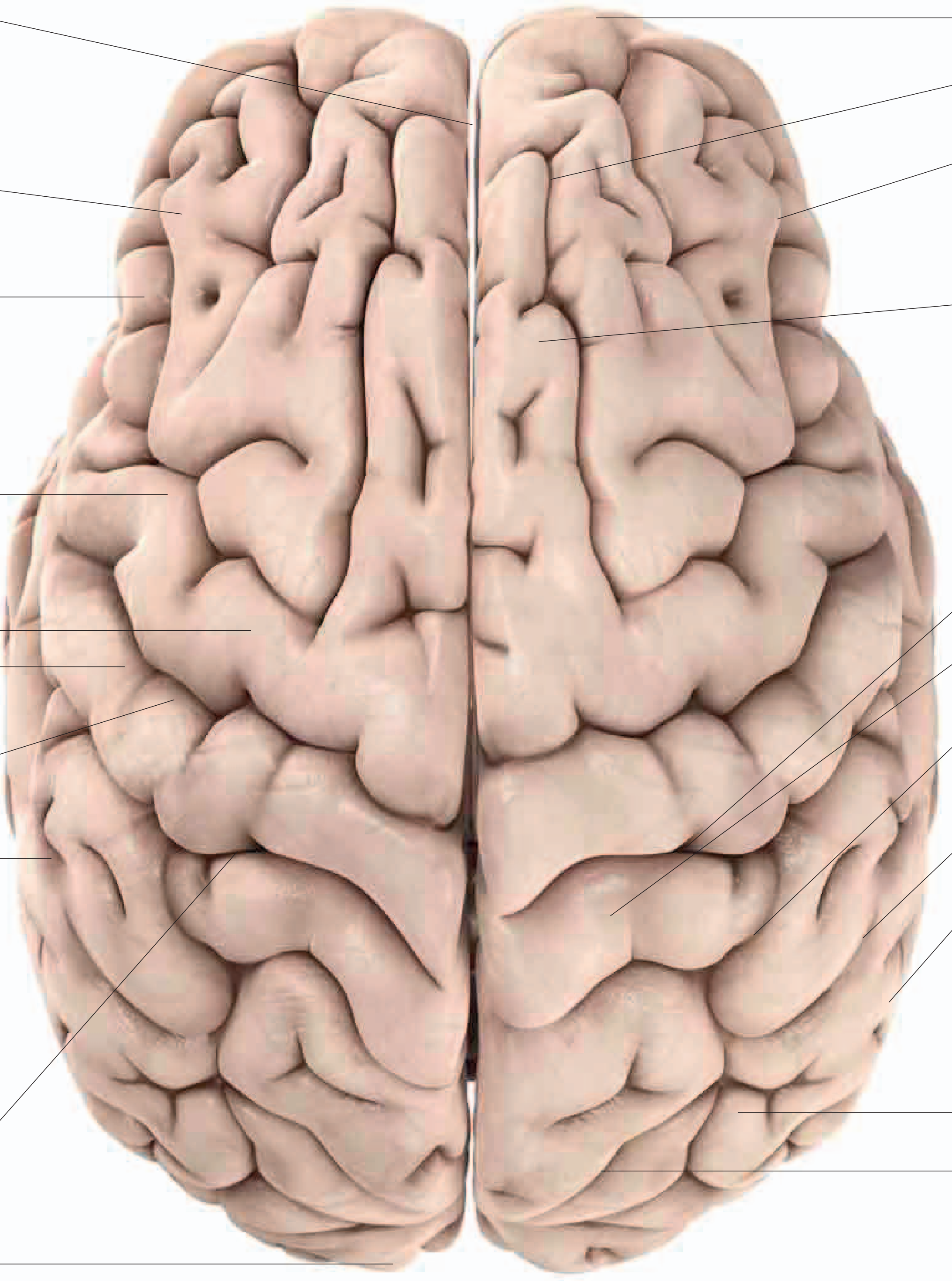
Middle frontal gyrus
Inferior frontal gyrus
Precentral sulcus
Precentral gyrus
Postcentral gyrus
Central sulcus
Postcentral sulcus
Occipital pole
Frontal pole
Superior frontal sulcus
Inferior frontal sulcus
Superior frontal gyrus
Cingulate sulcus
Superior parietal lobule
Superior temporal sulcus
Inferior parietal lobule

Longitudinal (cerebral) fissure A deep cleft dividing the two cerebral hemispheres

Supramarginal gyrus Many parts of the cortex are "association areas", involved with processing sensory information and perception. This gyrus, on the left, has been shown to be important in understanding spoken language, learning new vocabulary, and reading

Intraparietal sulcus Divides the superior from the inferior parietal lobule

Angular gyrus Turns a corner around the end of the superior temporal sulcus. Studies of brain function suggest this area may be important in mathematical problem-solving and understanding metaphors

Parieto-occipital sulcus Divides the parietal and occipital lobes

TOP VIEW OF BRAIN

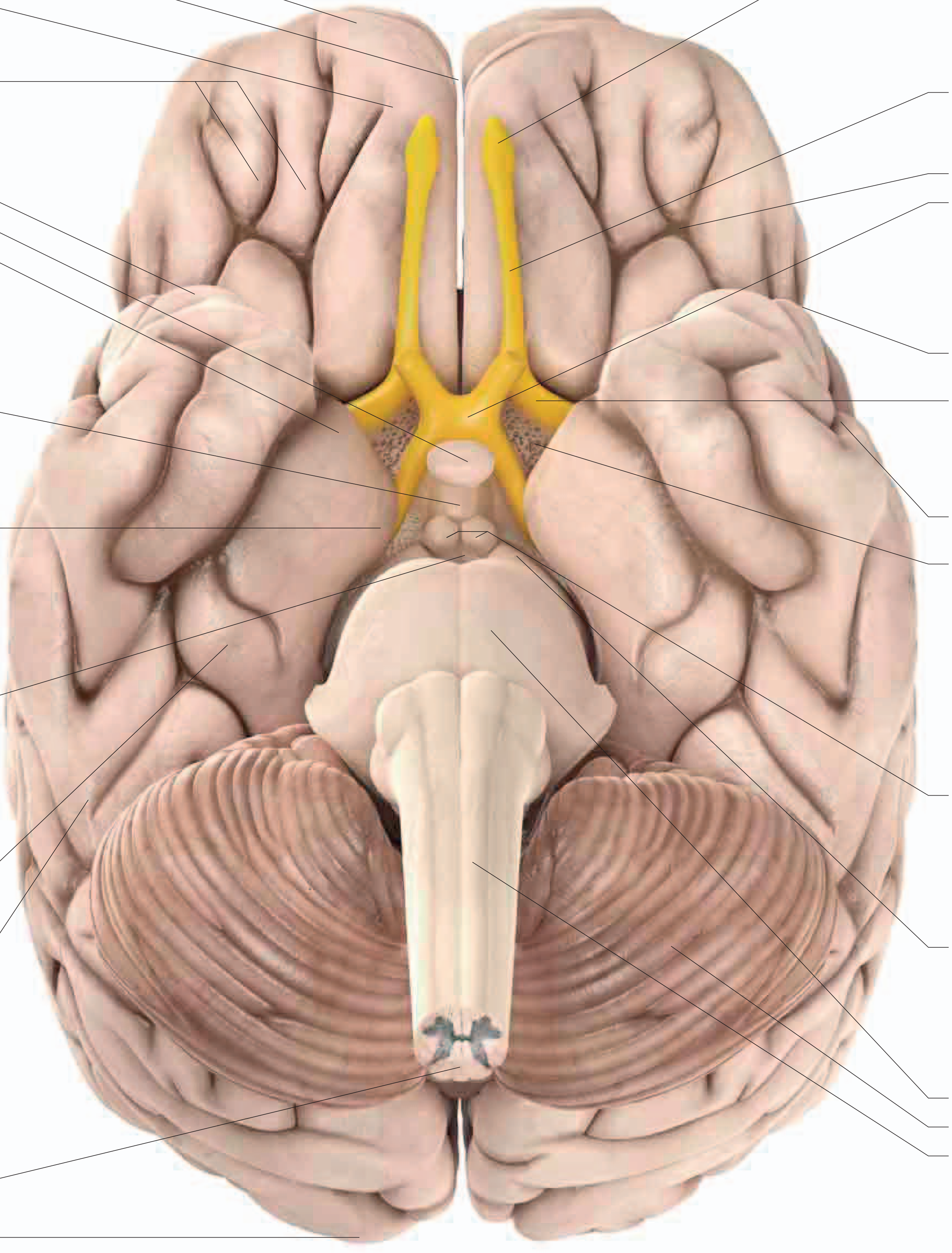

Frontal pole

Longitudinal (cerebral) fissure

Straight gyrus

Orbital gyri
Lie around the edges of the H-shaped orbital sulcus, and seem to play some role in empathy

Temporal pole

Pituitary gland

Parahippocampal gyrus
This part of the cortex, close to the hippocampus, plays an important role in memory and recognition

Tuber cinereum
Small bump of gray matter under the brain; part of the hypothalamus (see p.116)

Uncus
Hooked-under end of the parahippocampal gyrus; contains the primary olfactory cortex, receiving olfactory (smell) information

Interpeduncular fossa
Area enclosed by the cerebral peduncles on each side, by the optic chiasma in front, and the pons of the brain stem behind

Occipitotemporal fusiform gyrus

Inferior temporal gyrus

Spinal cord

Occipital pole

Olfactory bulb
Receives olfactory nerves, which have emerged from the top of the nasal cavity through the cribiform plate of the ethmoid bone, to enter the inside of the skull

Olfactory tract
Carries olfactory (smell) information back to the uncus

Orbital sulcus

Optic chiasma
Where the two optic nerves meet and swap fibers with each other, to form the optic tracts; chiasma means a cross

Lateral cerebral fossa

Olfactory trigone
The olfactory tract splays out into this triangular shape, just in front of the anterior perforated substance

Inferior temporal sulcus

Anterior perforated substance
Area of gray matter between the olfactory trigone, the optic chiasma, and the uncus; pierced by small arteries from the anterior and middle cerebral arteries

Mammillary bodies
Two breast-like bumps that are part of the limbic system, which is involved in memory, emotions, and behavior

Cerebral peduncle
"Stalk" of the brain, containing motor nerve fibers that descend from the cerebral cortex to the brain stem and spinal cord

Pons

Cerebellum

Pyramid
A prominence on the front of medulla containing motor nerve fibers that run from the cortex of the brain to the spinal cord

UNDERSIDE OF BRAIN

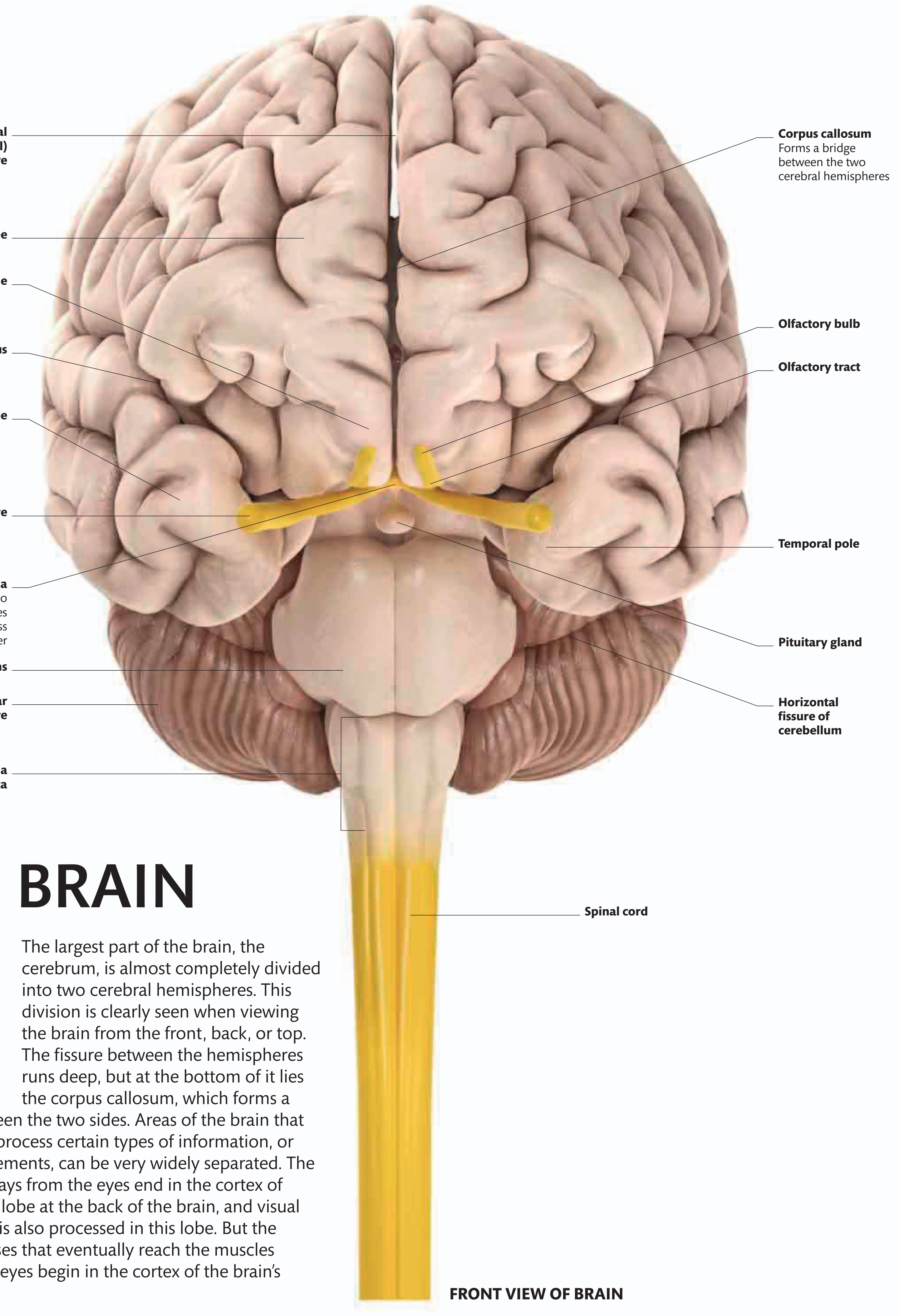
Longitudinal (cerebral) fissure
Frontal lobe
Frontal pole
Lateral sulcus
Temporal lobe
Optic nerve
Optic chiasma Where the two optic nerves partially cross each other
Pons
Cerebellar hemisphere
Medulla oblongata
Corpus callosum Forms a bridge between the two cerebral hemispheres
Olfactory bulb
Olfactory tract
Temporal pole
Pituitary gland
Horizontal fissure of cerebellum
Spinal cord

FRONT VIEW OF BRAIN

BRAIN

The largest part of the brain, the cerebrum, is almost completely divided into two cerebral hemispheres. This division is clearly seen when viewing the brain from the front, back, or top. The fissure between the hemispheres runs deep, but at the bottom of it lies the corpus callosum, which forms a bridge between the two sides. Areas of the brain that receive and process certain types of information, or govern movements, can be very widely separated. The visual pathways from the eyes end in the cortex of the occipital lobe at the back of the brain, and visual information is also processed in this lobe. But the nerve impulses that eventually reach the muscles to move the eyes begin in the cortex of the brain's frontal lobe.

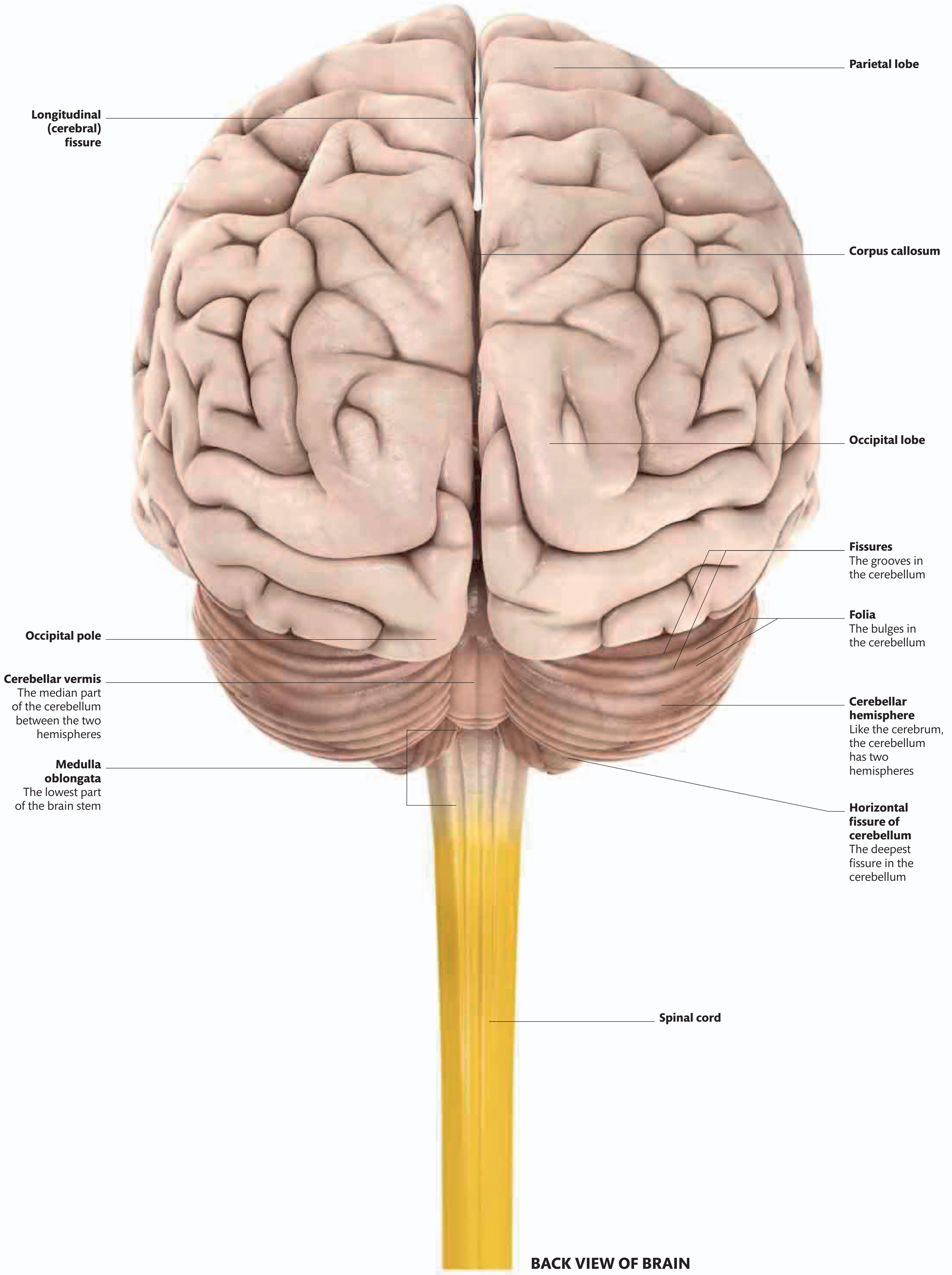
Parietal lobe
Longitudinal (cerebral) fissure
Corpus callosum
Occipital lobe
Fissures
The grooves in the cerebellum
Folia
The bulges in the cerebellum
Occipital pole
Cerebellar vermis
The median part of the cerebellum between the two hemispheres
Cerebellar hemisphere
Like the cerebrum, the cerebellum has two hemispheres
Medulla oblongata
The lowest part of the brain stem
Horizontal fissure of cerebellum
The deepest fissure in the cerebellum
Spinal cord

BACK VIEW OF BRAIN

Body of corpus callosum
The largest commissure (or bundle of connecting nerve fibers) between the two hemispheres, this forms the roofs of the lateral ventricles

Superior frontal gyrus

Cingulate gyrus
Cingulum is the Latin for girdle and this gyrus wraps closely around the corpus callosum; it is part of the limbic system, which is involved with emotional responses and behaviors

Septum pellucidum
This translucent partition is a thin dividing wall between the two lateral ventricles

Genu of corpus callosum
The anterior (front) end of the corpus callosum is bent over—genu means knee in Latin

Anterior commissure
A bundle of nerve fibers connecting parts of the two cerebral hemispheres

Optic chiasma
The crossover point where the two optic nerves meet and swap fibers, then part company as the optic tracts, which continue on each side of the brain toward the thalamus

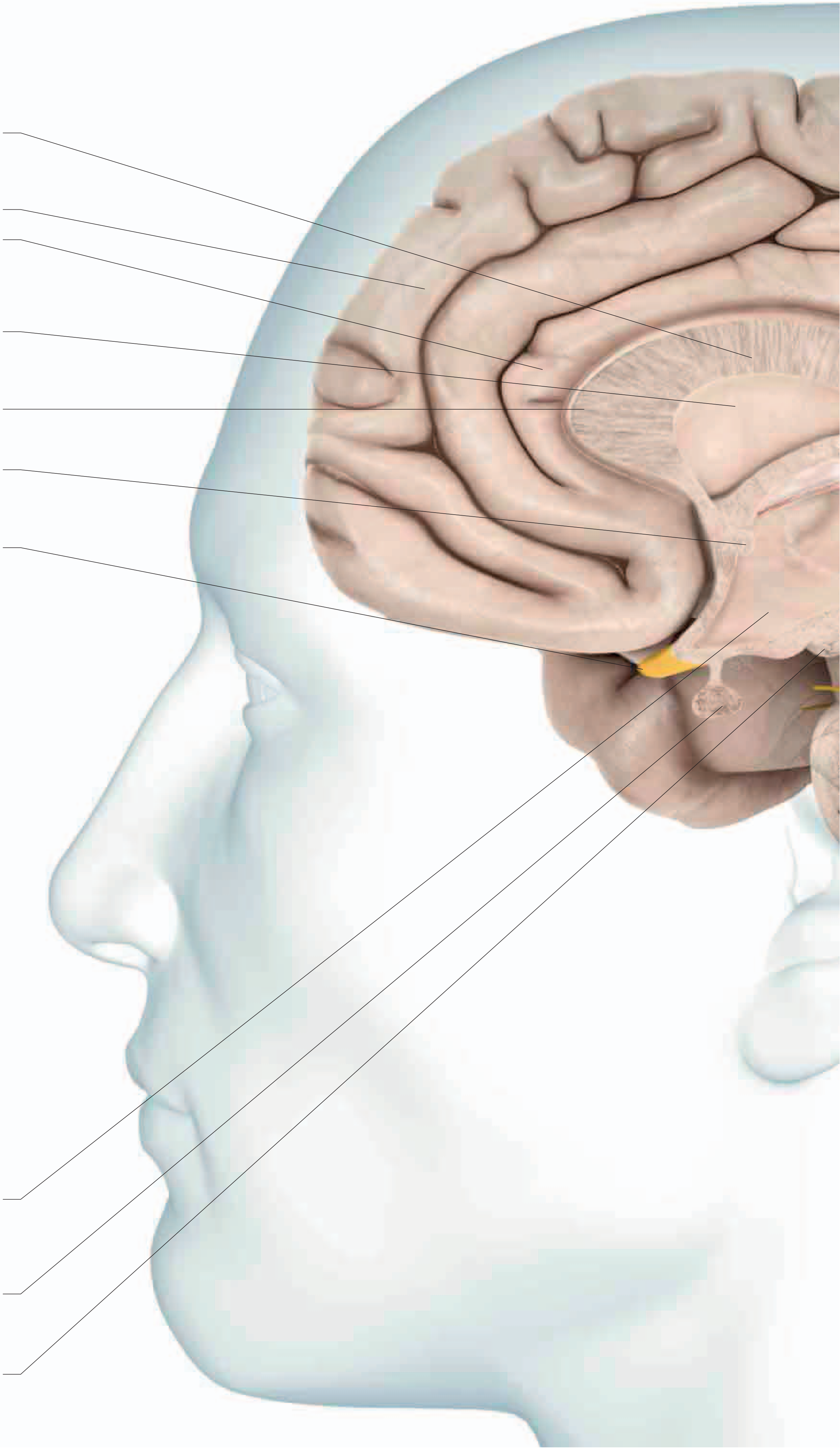

SAGITTAL SECTION THROUGH BRAIN

Hypothalamus
Plays an important role in regulating the internal environment of the body, by keeping a check on body temperature, blood pressure, and blood sugar level, for instance

Pituitary gland
Produces many hormones and forms a link between the brain and endocrine system

Mammillary body
Part of the limbic system of the brain

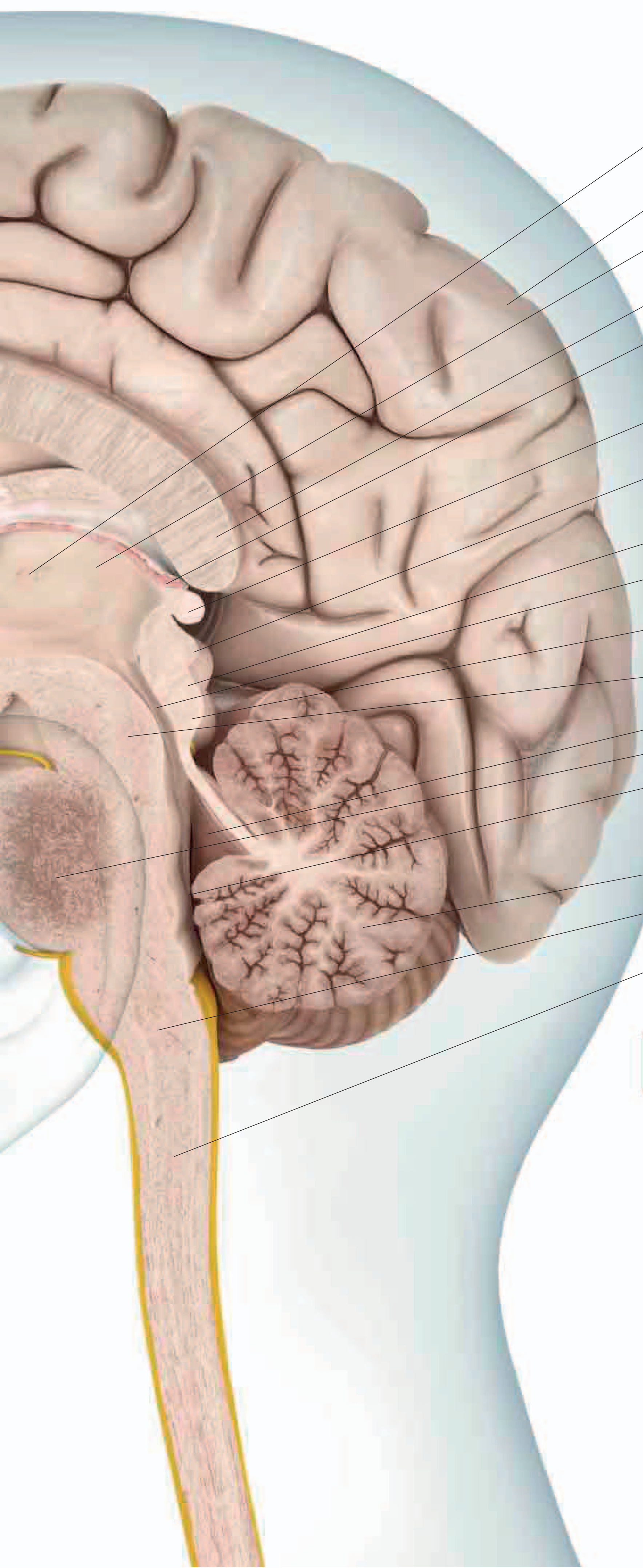

Interthalamic adhesion
Connection between the thalami on each side of the brain

Cerebrum
The largest part of the brain, consisting of the two cerebral hemispheres

Thalamus
Processes and relays sensory and motor information to higher brain centers

Splenium of corpus callosum
The posterior end of the corpus callosum

Choroid plexus of the third ventricle
A choroid plexus is formed where the inner and outer membranes of the brain come together; it is full of capillaries and produces cerebrospinal fluid, which flows into the ventricle

Pineal gland
Produces the hormone melatonin and is involved in the regulation of sleep-wake cycles

Superior colliculus
Involved in visual reflex pathways, including the pupillary light reflex, which makes the pupils constrict when bright light hits the retina

Tectum of the midbrain
The roof of the midbrain

Cerebral aqueduct
A narrow channel connecting the third and fourth ventricles

Inferior colliculus
Involved with auditory pathways, including reflex responses to loud noises

Tegmentum of midbrain

Fourth ventricle

Pons

Median aperture of the fourth ventricle
Cerebrospinal fluid escapes from the fourth ventricle via this opening in the midline, as well as through an opening on each side, into the subarachnoid space around the brain and spinal cord

Cerebellum

Medulla oblongata

Spinal cord

HEAD AND NECK

This median sagittal section—a vertical slice right through the middle of the brain—shows clearly the corpus callosum, which links the two hemispheres. We also see that the brain is not solid: there are cavities within it. Two spaces (or ventricles) lie inside each hemisphere, while the third and fourth ventricles are located on the midline. These spaces are full of cerebrospinal fluid. Beneath and behind the cerebrum sits the cerebellum. The gray cortex of the cerebellum is more finely folded than that of the cerebrum, with fissures separating its leaves (or folia). Sliced through this way, the inside of the cerebellum reveals a beautiful, treelike pattern. In this section, we can also see clearly all the parts of the brain stem—the midbrain, pons, and medulla.

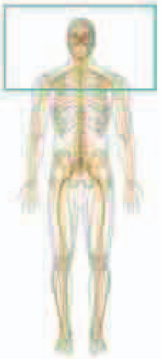

BRAIN

The brain is protected by three membranes called the meninges (which become inflamed in meningitis). The tough dura mater layer is the outermost covering, which surrounds the brain and the spinal cord. Under the dura mater is the cobweblike arachnoid mater layer. The delicate pia mater is a thin membrane on the surface of the brain. Between the pia mater and the arachnoid mater there is a slim gap—the subarachnoid space—which contains cerebrospinal fluid (CSF). Mainly produced by the choroid plexus in the brain's lateral ventricles, CSF flows through the third ventricle into the fourth, where it can escape via small apertures into the subarachnoid space.

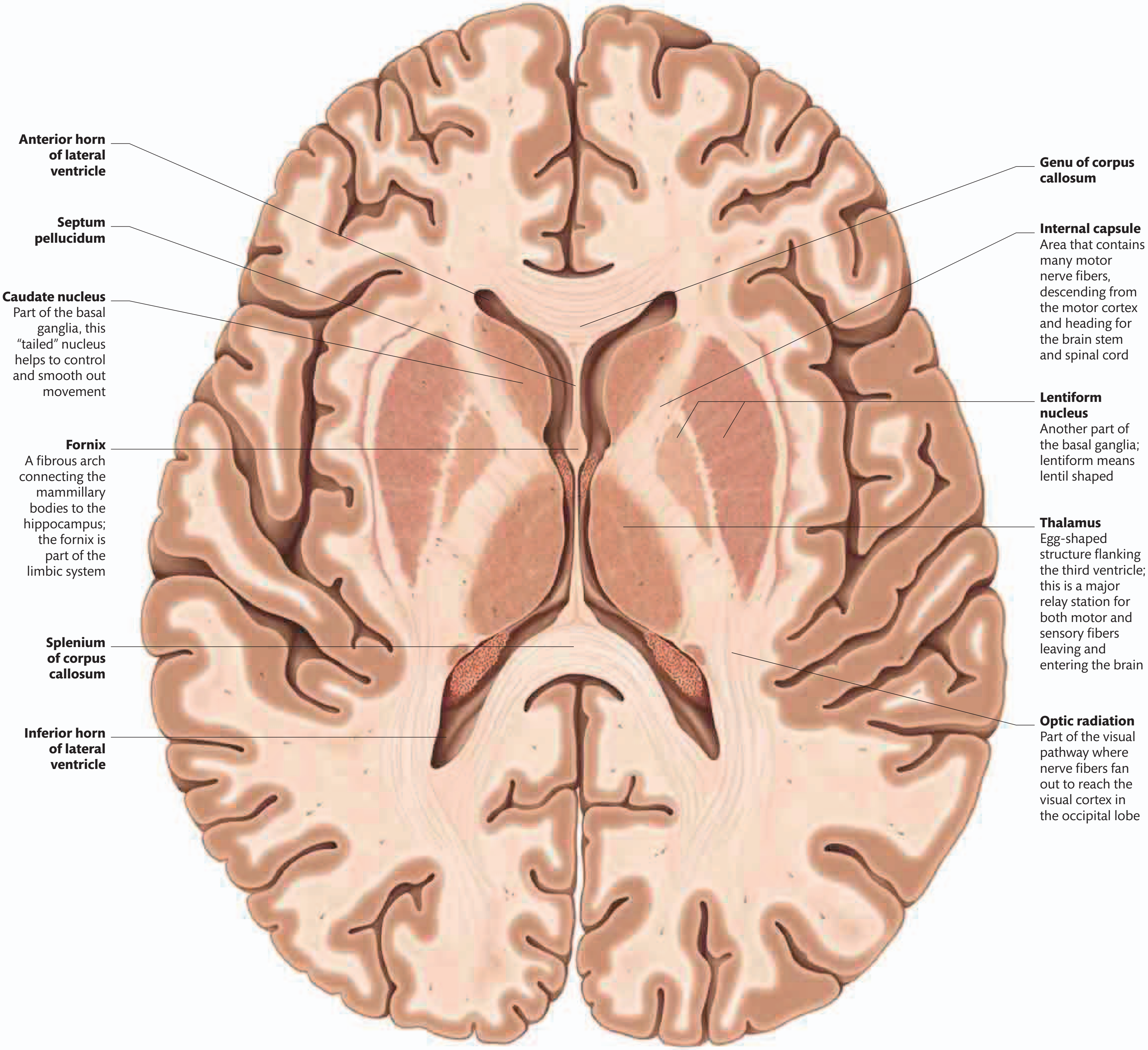

Anterior horn of lateral ventricle

Septum pellucidum

Caudate nucleus Part of the basal ganglia, this "tailed" nucleus helps to control and smooth out movement

Fornix A fibrous arch connecting the mammillary bodies to the hippocampus; the fornix is part of the limbic system

Splenium of corpus callosum

Inferior horn of lateral ventricle

Genu of corpus callosum

Internal capsule Area that contains many motor nerve fibers, descending from the motor cortex and heading for the brain stem and spinal cord

Lentiform nucleus Another part of the basal ganglia; lentiform means lentil shaped

Thalamus Egg-shaped structure flanking the third ventricle; this is a major relay station for both motor and sensory fibers leaving and entering the brain

Optic radiation Part of the visual pathway where nerve fibers fan out to reach the visual cortex in the occipital lobe

TRANSVERSE SECTION OF BRAIN

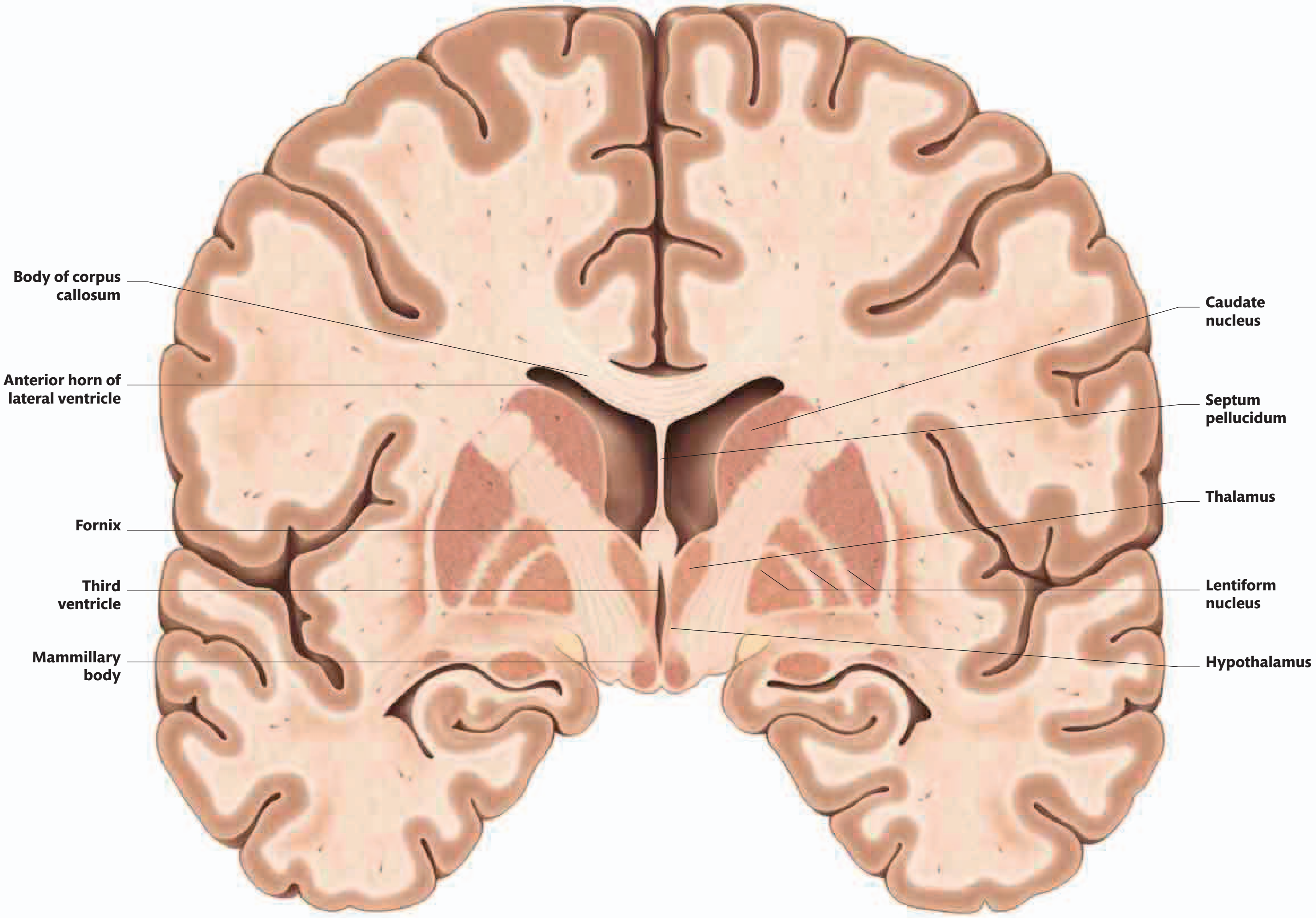
Body of corpus callosum
Anterior horn of lateral ventricle
Fornix
Third ventricle
Mammillary body
Caudate nucleus
Septum pellucidum
Thalamus
Lentiform nucleus
Hypothalamus

CORONAL SECTION OF BRAIN

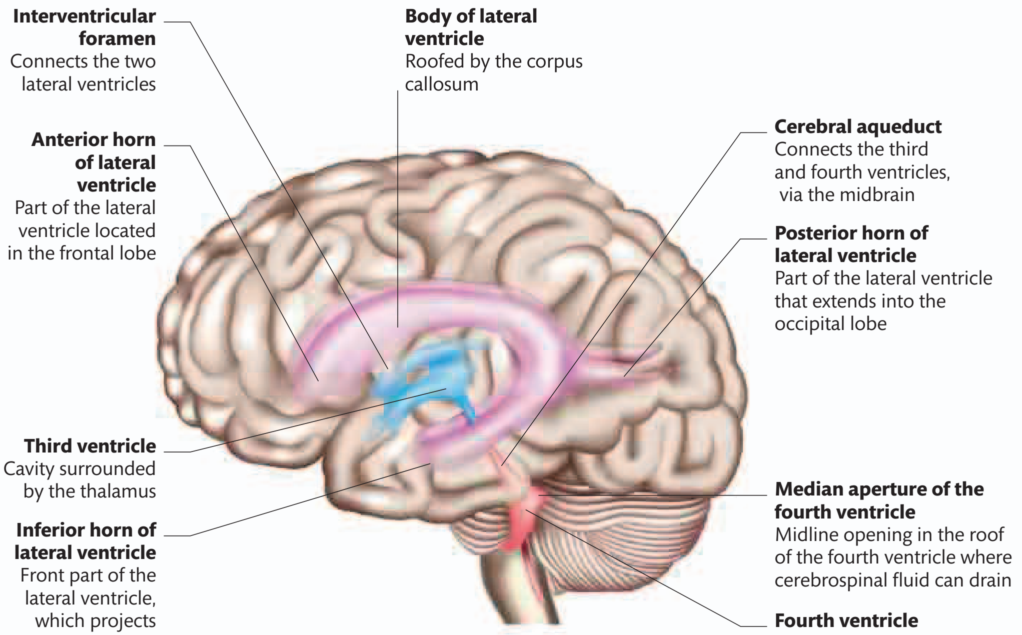
Interventricular foramen
Connects the two lateral ventricles
Body of lateral ventricle
Roofed by the corpus callosum
Anterior horn of lateral ventricle
Part of the lateral ventricle located in the frontal lobe
Cerebral aqueduct
Connects the third and fourth ventricles, via the midbrain
Posterior horn of lateral ventricle
Part of the lateral ventricle that extends into the occipital lobe
Third ventricle
Cavity surrounded by the thalamus
Median aperture of the fourth ventricle
Midline opening in the roof of the fourth ventricle where cerebrospinal fluid can drain
Inferior horn of lateral ventricle
Front part of the lateral ventricle, which projects down into the temporal lobe
Fourth ventricle
Cavity that lies between the pons and the cerebellum

VENTRICLES OF BRAIN

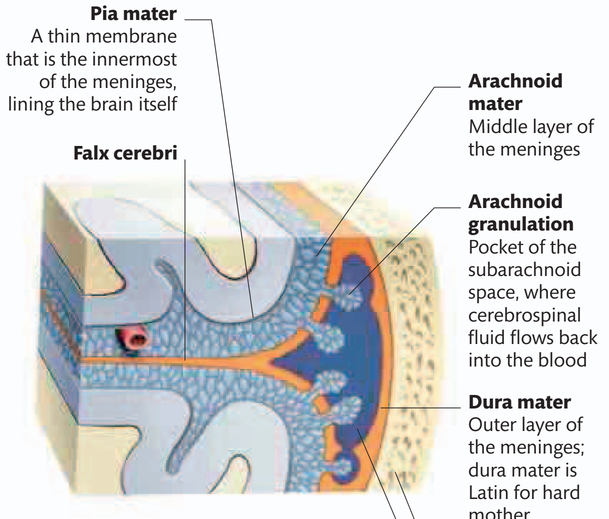
Pia mater
A thin membrane that is the innermost of the meninges, lining the brain itself
Falx cerebri
Arachnoid mater
Middle layer of the meninges
Arachnoid granulation
Pocket of the subarachnoid space, where cerebrospinal fluid flows back into the blood
Dura mater
Outer layer of the meninges; dura mater is Latin for hard mother
Skull
Superior sagittal sinus

MENINGES SECTION

HEAD AND NECK

The 12 pairs of cranial nerves (the standard abbreviation for which is CN) emerge from the brain and brain stem, leaving through holes, or "foramina", in the base of the skull. Some nerves are purely sensory, some just have motor functions, but most contain a mixture of motor and sensory fibers. A few also contain autonomic nerve fibers. The olfactory nerve and the optic nerve attach to the brain itself. The other 10 pairs of cranial nerves emerge from the brain stem. All the cranial nerves supply parts of the head and neck, except the vagus nerve. This has branches in the neck, but then continues on to supply organs in the thorax and down to the abdomen. Careful testing of cranial nerves, including tests of sight, eye and head movement, taste, and so on, can help doctors to pinpoint neurological problems in the head and neck.

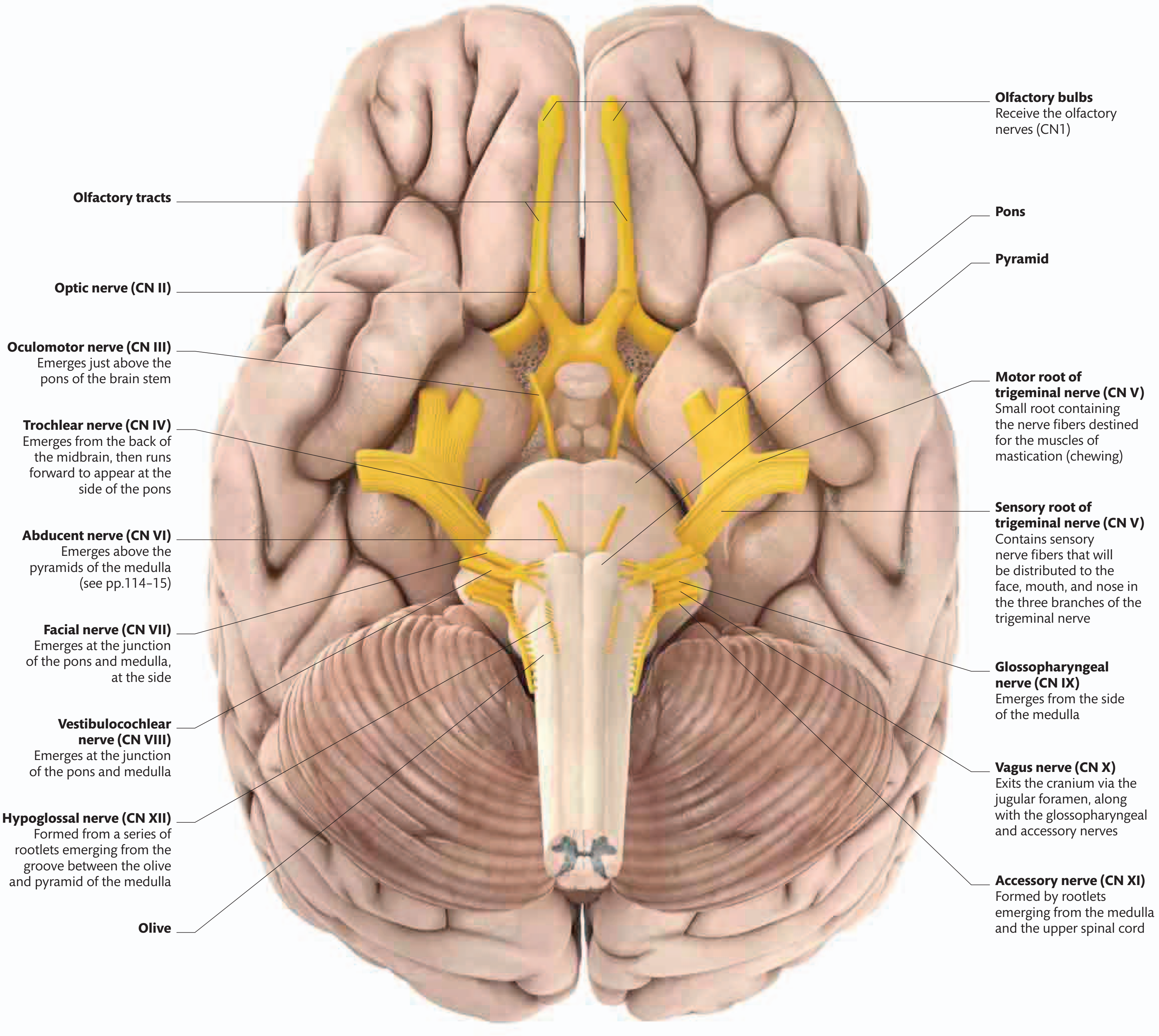
Olfactory bulbs
Receive the olfactory nerves (CN1)
Olfactory tracts
Pons
Pyramid
Optic nerve (CN II)
Oculomotor nerve (CN III)
Emerges just above the pons of the brain stem
Motor root of trigeminal nerve (CN V)
Small root containing the nerve fibers destined for the muscles of mastication (chewing)
Trochlear nerve (CN IV)
Emerges from the back of the midbrain, then runs forward to appear at the side of the pons
Sensory root of trigeminal nerve (CN V)
Contains sensory nerve fibers that will be distributed to the face, mouth, and nose in the three branches of the trigeminal nerve
Abducent nerve (CN VI)
Emerges above the pyramids of the medulla (see pp.114–15)
Facial nerve (CN VII)
Emerges at the junction of the pons and medulla, at the side
Glossopharyngeal nerve (CN IX)
Emerges from the side of the medulla
Vestibulocochlear nerve (CN VIII)
Emerges at the junction of the pons and medulla
Vagus nerve (CN X)
Exits the cranium via the jugular foramen, along with the glossopharyngeal and accessory nerves
Hypoglossal nerve (CN XII)
Formed from a series of rootlets emerging from the groove between the olive and pyramid of the medulla
Accessory nerve (CN XI)
Formed by rootlets emerging from the medulla and the upper spinal cord
Olive

ORIGIN OF CRANIAL NERVES (UNDERSIDE OF BRAIN)

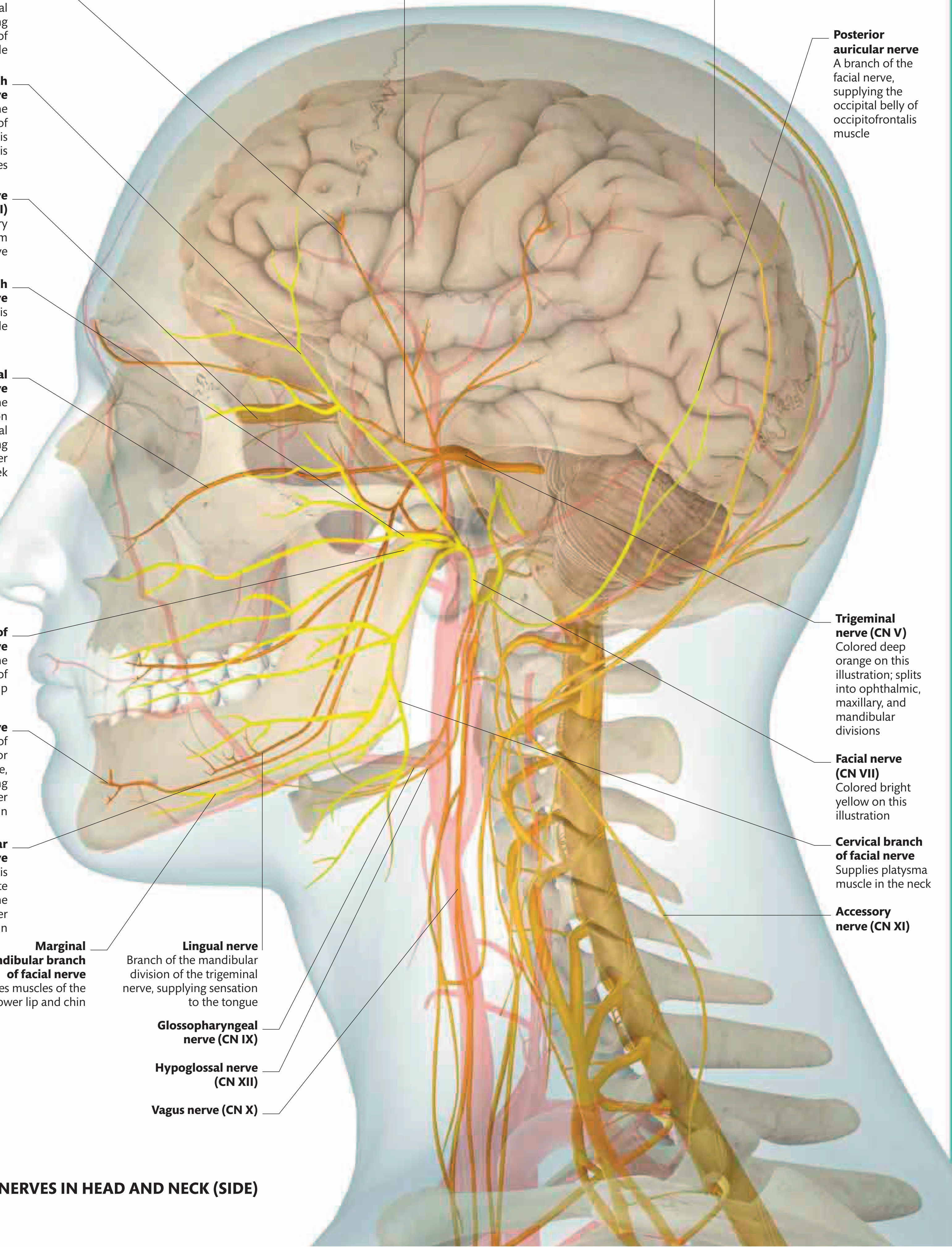

CRANIAL NERVES IN HEAD AND NECK (SIDE)

Auriculotemporal nerve Branch of the mandibular division of the trigeminal nerve, supplying sensation to part of the ear and temple

Temporal branch of facial nerve Supplies the frontal belly of occipitofrontalis and orbicularis oculi muscles

Optic nerve (CN II) Carries sensory information from the retina of the eye

Zygomatic branch of facial nerve Supplies orbicularis oculi muscle

Infraorbital nerve Branch of the maxillary division of the trigeminal nerve, supplying sensation over the cheek

Buccal branch of facial nerve Supplies the muscles of the upper lip

Mental nerve Continuation of the inferior alveolar nerve, supplying sensation over the chin

Inferior alveolar nerve Branches of this nerve innervate the lower teeth, the gums, the lower lip, and the chin

Marginal mandibular branch of facial nerve Supplies muscles of the lower lip and chin

Lingual nerve Branch of the mandibular division of the trigeminal nerve, supplying sensation to the tongue

Glossopharyngeal nerve (CN IX)

Hypoglossal nerve (CN XII)

Vagus nerve (CN X)

Ophthalmic nerve A branch of the trigeminal nerve, supplying sensation to the upper part of the face

Greater auricular nerve A branch of the second cervical nerve, supplying sensation to the skin of the back of the head

Posterior auricular nerve A branch of the facial nerve, supplying the occipital belly of occipitofrontalis muscle

Trigeminal nerve (CN V) Colored deep orange on this illustration; splits into ophthalmic, maxillary, and mandibular divisions

Facial nerve (CN VII) Colored bright yellow on this illustration

Cervical branch of facial nerve Supplies platysma muscle in the neck

Accessory nerve (CN XI)

EYE

The eyes are precious organs. They are well protected inside the eye sockets, or bony orbits, of the skull. They are also protected by the eyelids, and bathed in tears produced by the lacrimal glands. Each eyeball is only 1in (2.5cm) in diameter. The orbit provides an anchor for the muscles that move the eye, and the rest of the space inside the orbit is largely filled up with fat. Holes and fissures at the back of this bony cavern transmit nerves and blood vessels, including the optic nerve, which carries sensory information from the retina to the brain. Other nerves supply the eye muscles and the lacrimal glands, and even continue on to the face to supply sensation to the skin of the eyelids and forehead.

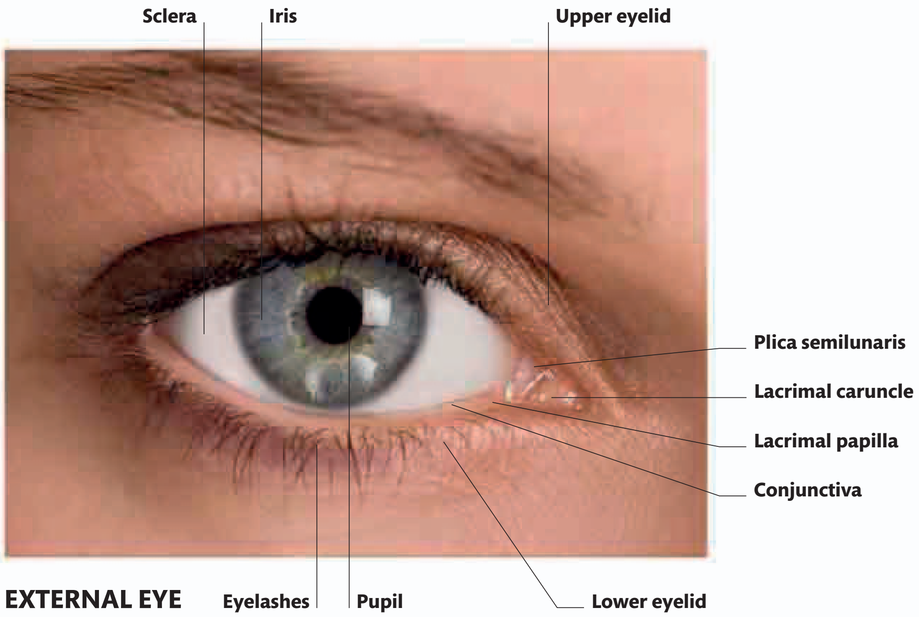

EXTERNAL EYE

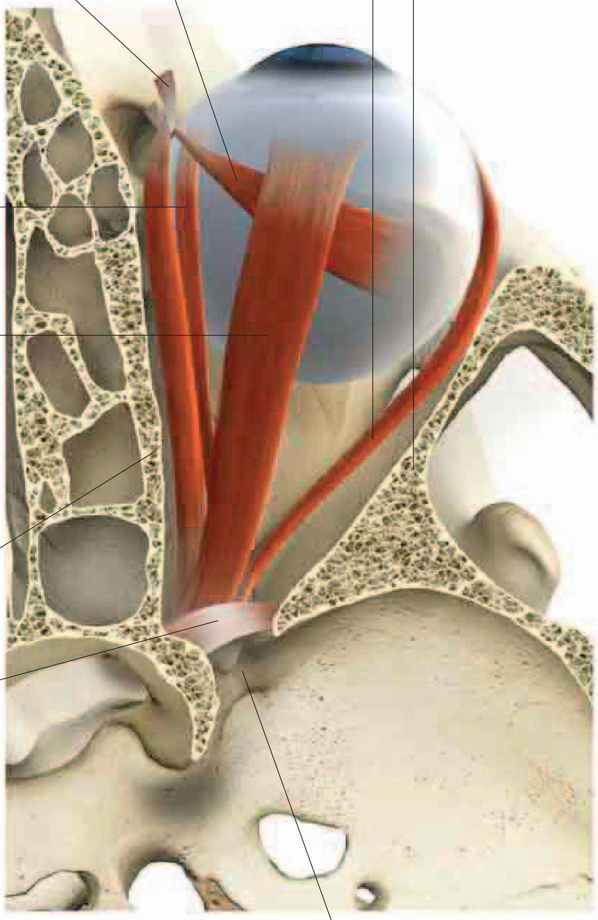

Superior oblique muscle Rotates the eyeball downward and outward, as well as medially; the inferior oblique muscle under the eyeball rotates it upward and inward

Trochlea of superior oblique muscle Trochlea is Greek for pulley; the superior oblique muscle runs through this fibrous loop attached to the frontal bone, which changes the muscle's trajectory

Medial rectus muscle Rotates the eyeball inward (adduction)

Superior rectus muscle Rotates the eyeball upward (elevation); the inferior rectus under the eyeball rotates it downward (depression)

Medial wall of orbit Formed here by the ethmoid bone

Common annular tendon A ringlike tendon anchored to the edges of the optic canal and superior orbital fissure, to which the four rectus (straight) muscles of the eye attach

Lateral rectus muscle Rotates the eyeball outward (abduction)

Lateral wall of orbit Formed here by the zygomatic bone

Superior orbital fissure Hole in the sphenoid bone at the back of the orbit

MUSCLES OF THE EYE (FROM ABOVE)

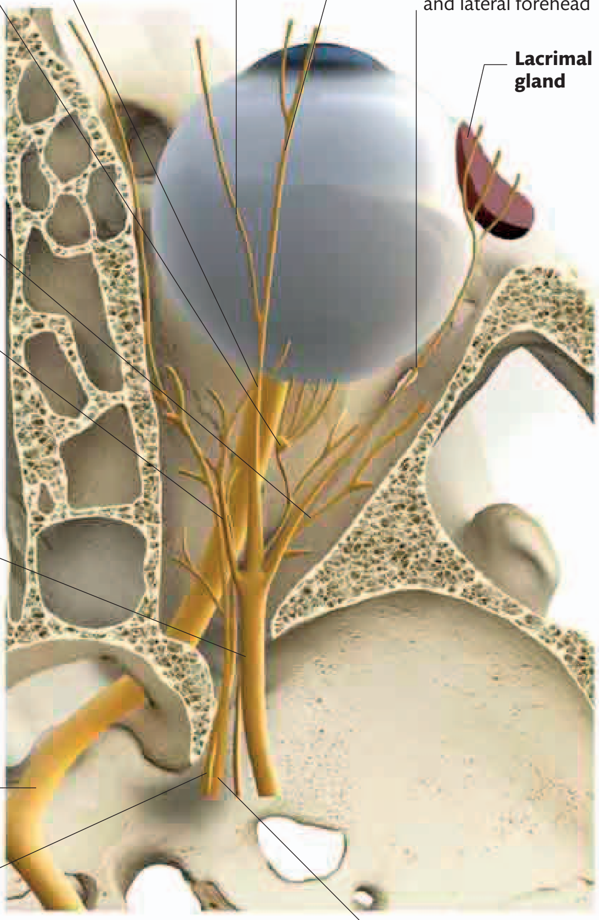

Frontal nerve Large branch of the ophthalmic nerve; splits into supraorbital and supratrochlear branches

Ciliary ganglion Receives parasympathetic nerve fibers from the oculomotor nerve and sends them into the eyeball via the short ciliary nerves, to supply the muscles of the iris and lens

Abducent nerve Supplies the lateral rectus muscle

Nasociliary nerve Part of the ophthalmic nerve; its branches supply sensation to the ethmoidal sinuses, the nasal cavity, and the eyeball

Ophthalmic nerve Branch of the trigeminal nerve; supplies sensation to the eyeball, the conjunctiva, and part of the lining of the nose, as well as the eyelids and forehead

Optic nerve Carries sensory nerve fibers from the retina

Oculomotor nerve Supplies all muscles that move the eye, apart from the superior oblique and lateral rectus muscles

Supratrochlear nerve Runs over the eyeball and up, out of the orbit, to supply sensation to the middle of the forehead

Supraorbital nerve Runs forward, out of the orbit, and turns upward on the frontal bone to supply the upper eyelid

Lacrimal nerve Supplies skin over the upper eyelid and lateral forehead

Lacrimal gland

Trochlear nerve Supplies the superior oblique muscle

NERVES OF THE ORBIT (FROM ABOVE)

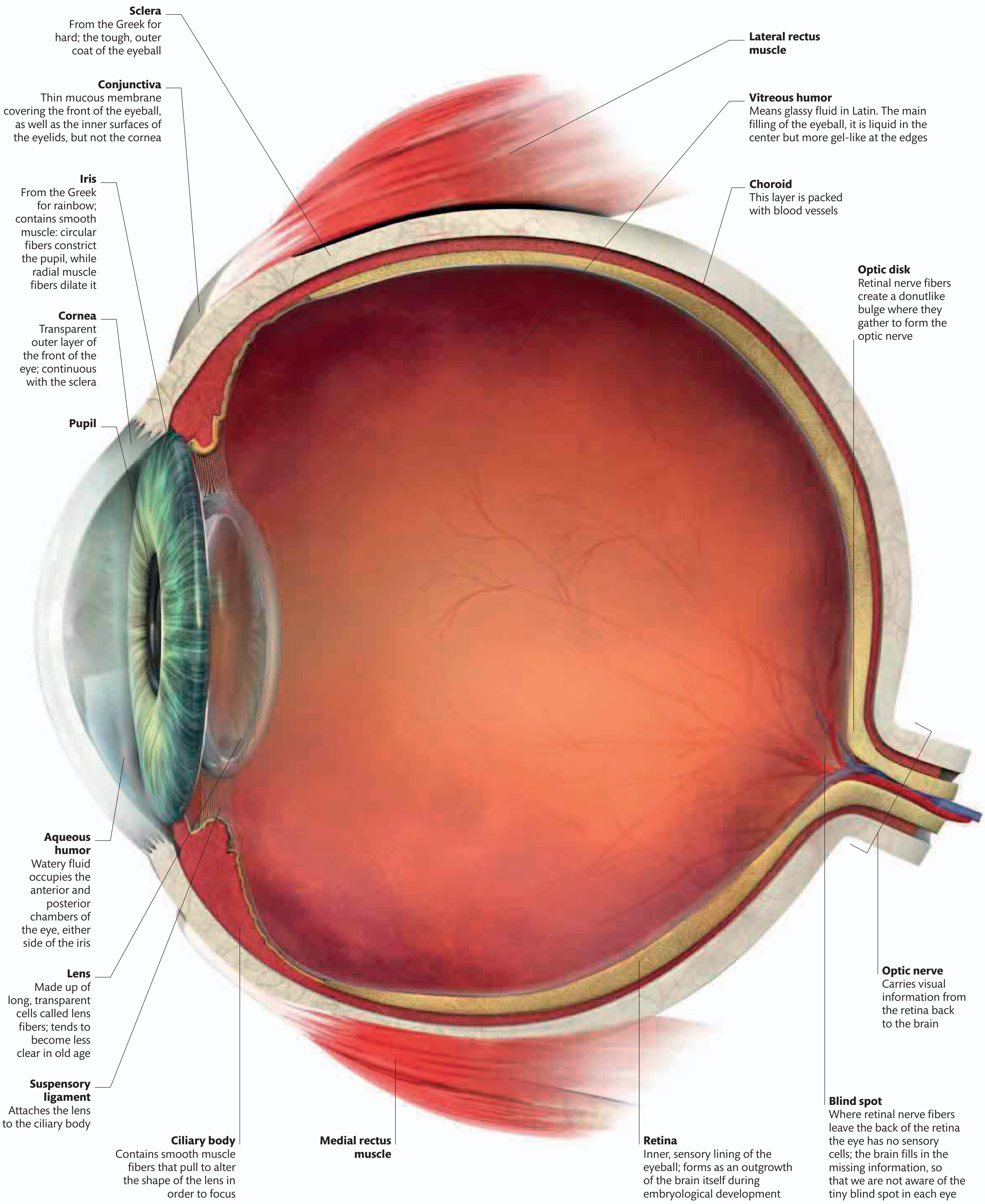
Sclera
From the Greek for hard; the tough, outer coat of the eyeball
Conjunctiva
Thin mucous membrane covering the front of the eyeball, as well as the inner surfaces of the eyelids, but not the cornea
Iris
From the Greek for rainbow; contains smooth muscle: circular fibers constrict the pupil, while radial muscle fibers dilate it
Cornea
Transparent outer layer of the front of the eye; continuous with the sclera
Pupil
Aqueous humor
Watery fluid occupies the anterior and posterior chambers of the eye, either side of the iris
Lens
Made up of long, transparent cells called lens fibers; tends to become less clear in old age
Suspensory ligament
Attaches the lens to the ciliary body
Ciliary body
Contains smooth muscle fibers that pull to alter the shape of the lens in order to focus
Medial rectus muscle
Retina
Inner, sensory lining of the eyeball; forms as an outgrowth of the brain itself during embryological development
Blind spot
Where retinal nerve fibers leave the back of the retina the eye has no sensory cells; the brain fills in the missing information, so that we are not aware of the tiny blind spot in each eye
Optic nerve
Carries visual information from the retina back to the brain
Optic disk
Retinal nerve fibers create a donutlike bulge where they gather to form the optic nerve
Choroid
This layer is packed with blood vessels
Vitreous humor
Means glassy fluid in Latin. The main filling of the eyeball, it is liquid in the center but more gel-like at the edges
Lateral rectus muscle

HORIZONTAL SECTION THROUGH THE EYEBALL

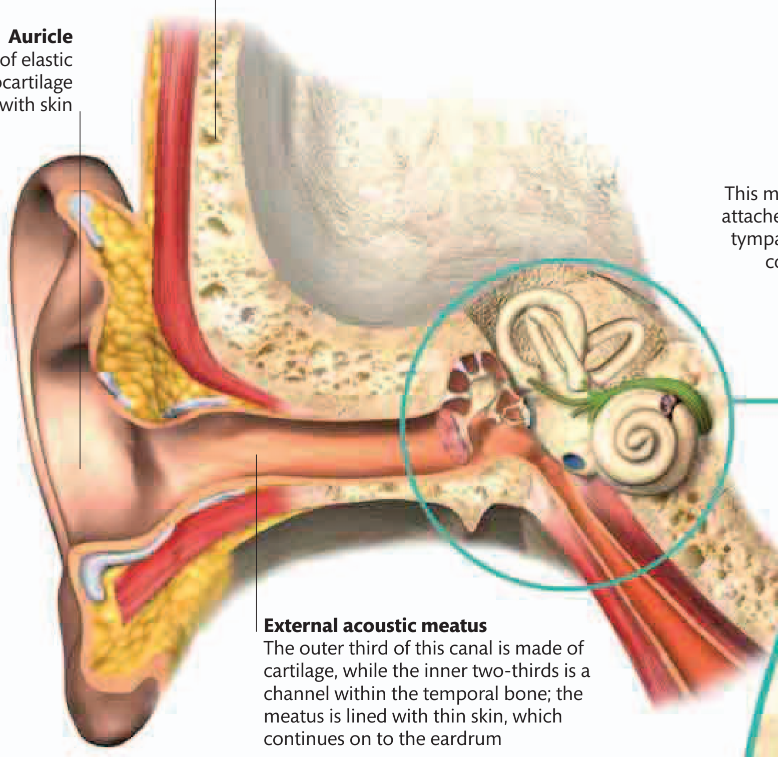

Temporal bone
Forms part of the side wall and base of the skull; houses the workings of the ear

Auricle
Made of elastic fibrocartilage covered with skin

External acoustic meatus
The outer third of this canal is made of cartilage, while the inner two-thirds is a channel within the temporal bone; the meatus is lined with thin skin, which continues on to the eardrum

EXTERNAL EAR

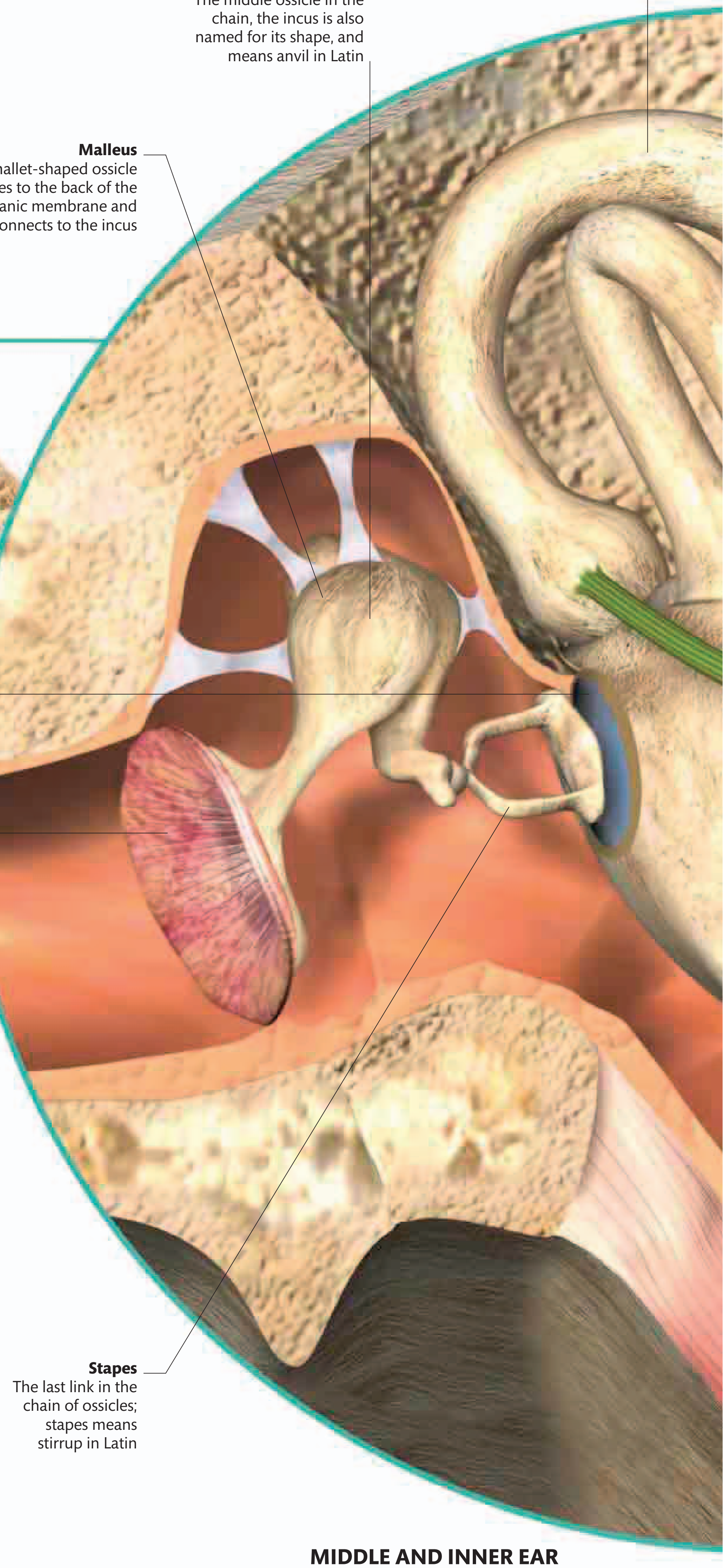

Posterior semicircular canal
The semicircular canals are each less than ¾in (2cm) in length, with a diameter of less than 1⁄32in (1mm); this canal is positioned vertically

Incus
The middle ossicle in the chain, the incus is also named for its shape, and means anvil in Latin

Malleus
This mallet-shaped ossicle attaches to the back of the tympanic membrane and connects to the incus

Oval window
Where the stapes attaches to the base of the cochlea, transmitting its vibrations to the fluid inside the cochlea

Tympanic membrane
The eardrum vibrates as sound waves buffet it; the ossicles (the malleus, incus, and stapes) carry those vibrations through the middle ear to the inner ear

Stapes
The last link in the chain of ossicles; stapes means stirrup in Latin

MIDDLE AND INNER EAR

EAR

The ear can be divided up into external, middle, and internal parts. The external ear includes the auricle on the outside of the head, and the external acoustic meatus—the canal that leads to the eardrum, or tympanic membrane. The middle ear is an air space inside the temporal bone. It contains the ossicles (ear bones) and is linked to the pharynx by the pharyngotympanic, or Eustachian, tube. Minute hair cells inside the inner ear convert vibrations in the fluid within the cochlea into an electrical nerve impulse. Similar hair cells in the vestibular apparatus (the semicircular canals, utricle, and saccule) convert mechanical stimuli, produced by motions of the head, into nerve impulses. The sensory nerves leaving the inner ear join to form the vestibulocochlear nerve.

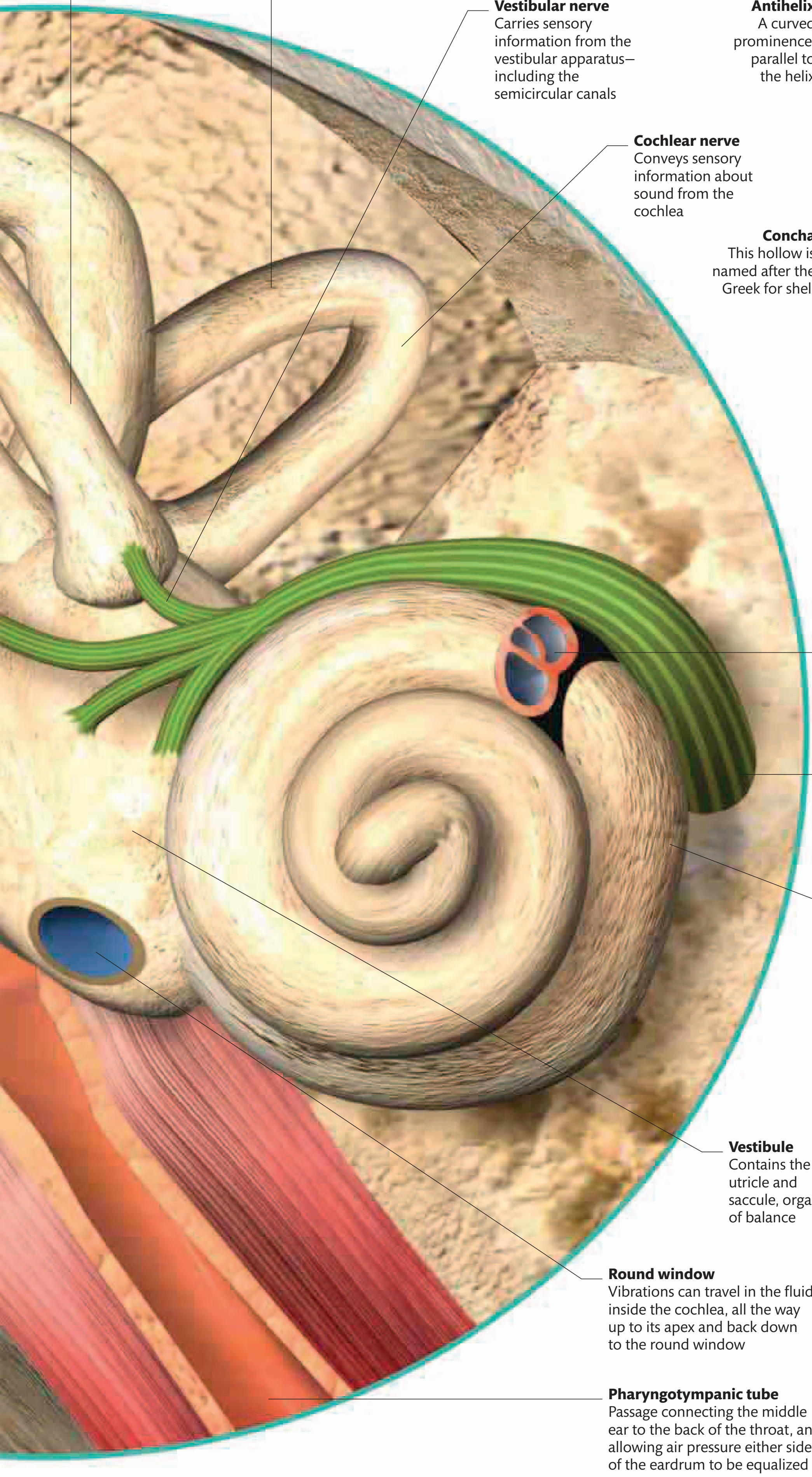

Lateral semicircular canal This is positioned horizontally

Anterior semicircular canal Positioned vertically, but at right angles to the plane of the posterior semicircular canal

Vestibular nerve Carries sensory information from the vestibular apparatus—including the semicircular canals

Cochlear nerve Conveys sensory information about sound from the cochlea

Section cut from cochlea From top to bottom shows vestibular canal, cochlear duct, and tympanic canal

Vestibulocochlear nerve The cochlear nerve conveys sensory information about sound from the cochlea. It joins the vestibular nerve to the vestibulocochlear nerve

Cochlea Not surprisingly, cochlea means snail in Latin

Vestibule Contains the utricle and saccule, organs of balance

Round window Vibrations can travel in the fluid inside the cochlea, all the way up to its apex and back down to the round window

Pharyngotympanic tube Passage connecting the middle ear to the back of the throat, and allowing air pressure either side of the eardrum to be equalized

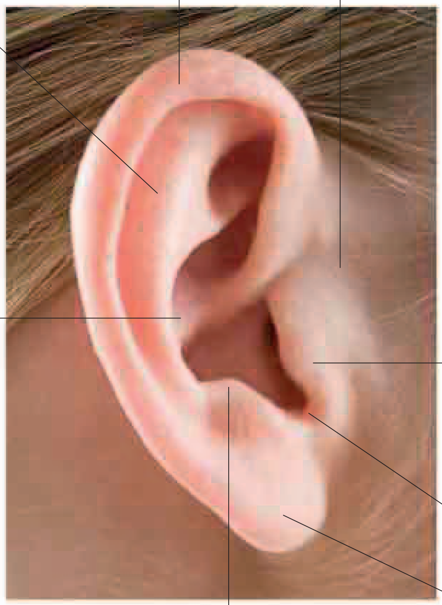

Helix The outer rim of the auricle

External acoustic meatus

Antihelix A curved prominence, parallel to the helix

Concha This hollow is named after the Greek for shell

Tragus This little flap overlaps the external acoustic meatus

Intertragic notch

Lobule

Antitragus A small tubercle opposite the tragus

AURICLE

Tympanic membrane As seen with an otoscope, a healthy eardrum has a pearly, almost translucent appearance

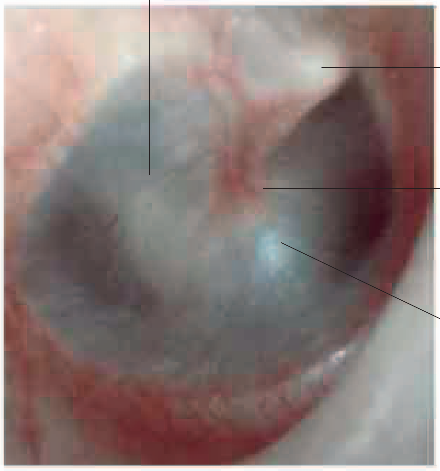

Lateral process of malleus

Handle of malleus

Cone of light Light is reflected in the front, lower quadrant of the eardrum

EARDRUM

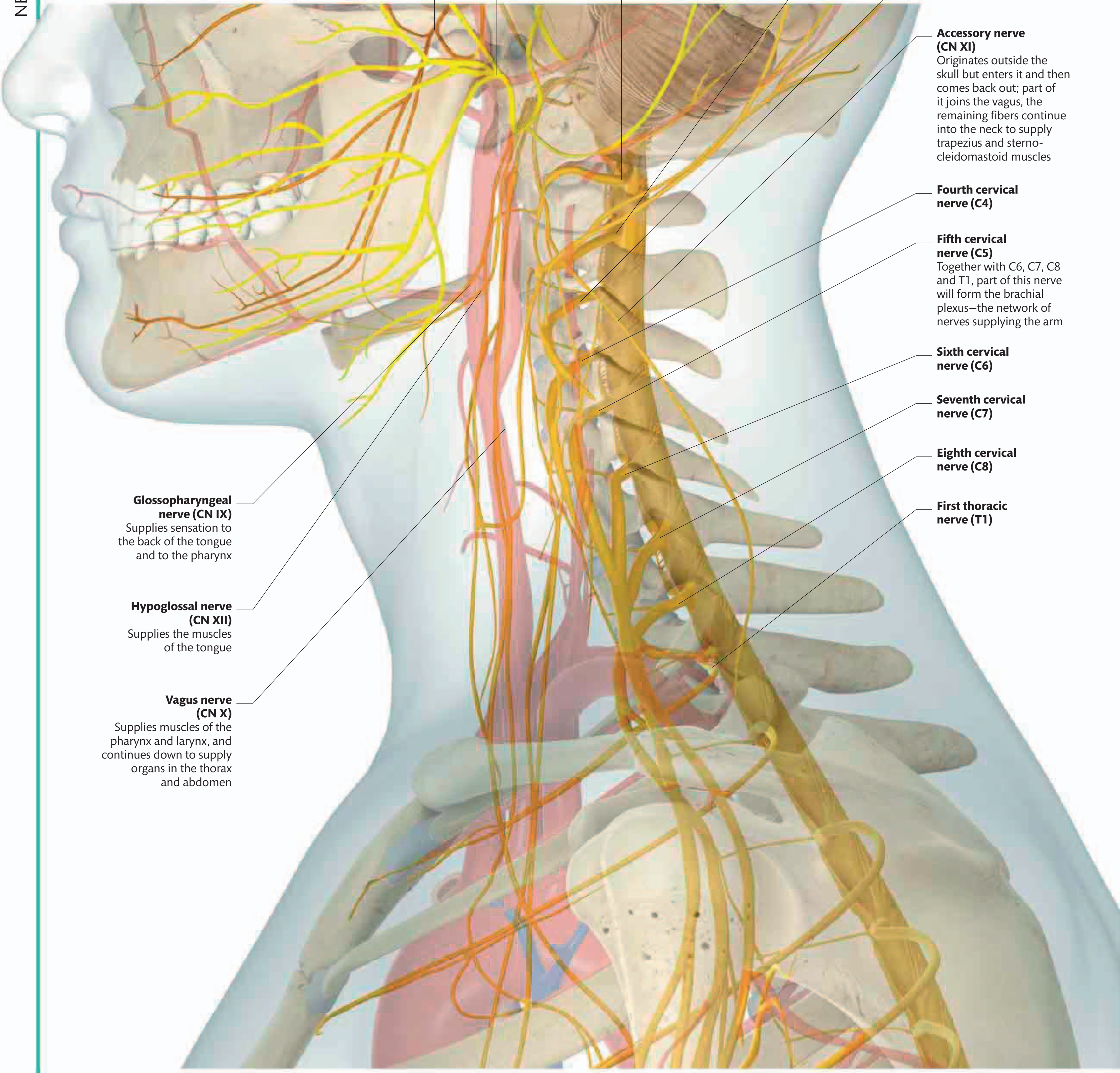

Mandibular division of trigeminal nerve (CN V)

Facial nerve (CN VII)

First cervical nerve (C1)
The very first spinal nerve; its branches supply some muscles in the upper neck

Second cervical nerve (C2)
Along with C3 and C4, this nerve supplies sensation to the skin of the neck as well as supplying a range of muscles in the neck

Third cervical nerve (C3)

Accessory nerve (CN XI)
Originates outside the skull but enters it and then comes back out; part of it joins the vagus, the remaining fibers continue into the neck to supply trapezius and sternocleidomastoid muscles

Fourth cervical nerve (C4)

Fifth cervical nerve (C5)
Together with C6, C7, C8 and T1, part of this nerve will form the brachial plexus—the network of nerves supplying the arm

Sixth cervical nerve (C6)

Seventh cervical nerve (C7)

Eighth cervical nerve (C8)

First thoracic nerve (T1)

Glossopharyngeal nerve (CN IX)
Supplies sensation to the back of the tongue and to the pharynx

Hypoglossal nerve (CN XII)
Supplies the muscles of the tongue

Vagus nerve (CN X)
Supplies muscles of the pharynx and larynx, and continues down to supply organs in the thorax and abdomen

NERVES OF THE NECK (SIDE)

NECK

The last four cranial nerves all appear in the neck. The glossopharyngeal nerve supplies the parotid gland and the back of the tongue, then runs down to the pharynx. The vagus nerve is sandwiched between the common carotid artery and the internal jugular vein, and it gives branches to the pharynx and larynx before continuing down into the thorax. The accessory nerve supplies the sternocleidomastoid and trapezius muscles in the neck, while the last cranial nerve, the hypoglossal, dips down below the mandible, then curves back up to supply the muscles of the tongue. We can also see spinal nerves in the neck. The upper four cervical nerves supply neck muscles and skin, while the lower four contribute to the brachial plexus and are destined for the arm.

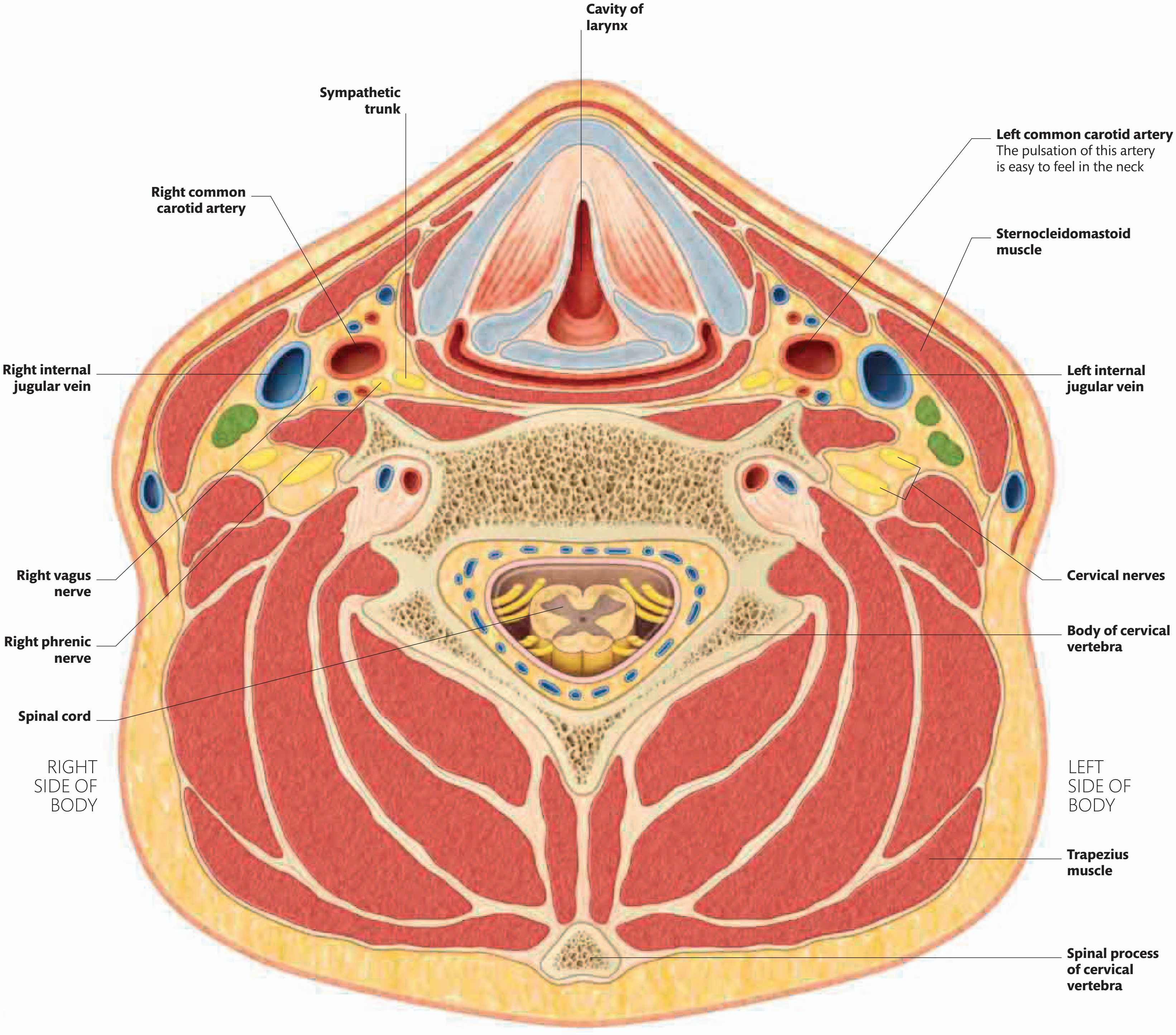
Cavity of larynx
Sympathetic trunk
Right common carotid artery
Left common carotid artery
The pulsation of this artery is easy to feel in the neck
Sternocleidomastoid muscle
Right internal jugular vein
Left internal jugular vein
Right vagus nerve
Cervical nerves
Right phrenic nerve
Body of cervical vertebra
Spinal cord
RIGHT SIDE OF BODY
LEFT SIDE OF BODY
Trapezius muscle
Spinal process of cervical vertebra

TRANSVERSE SECTION OF THE NECK

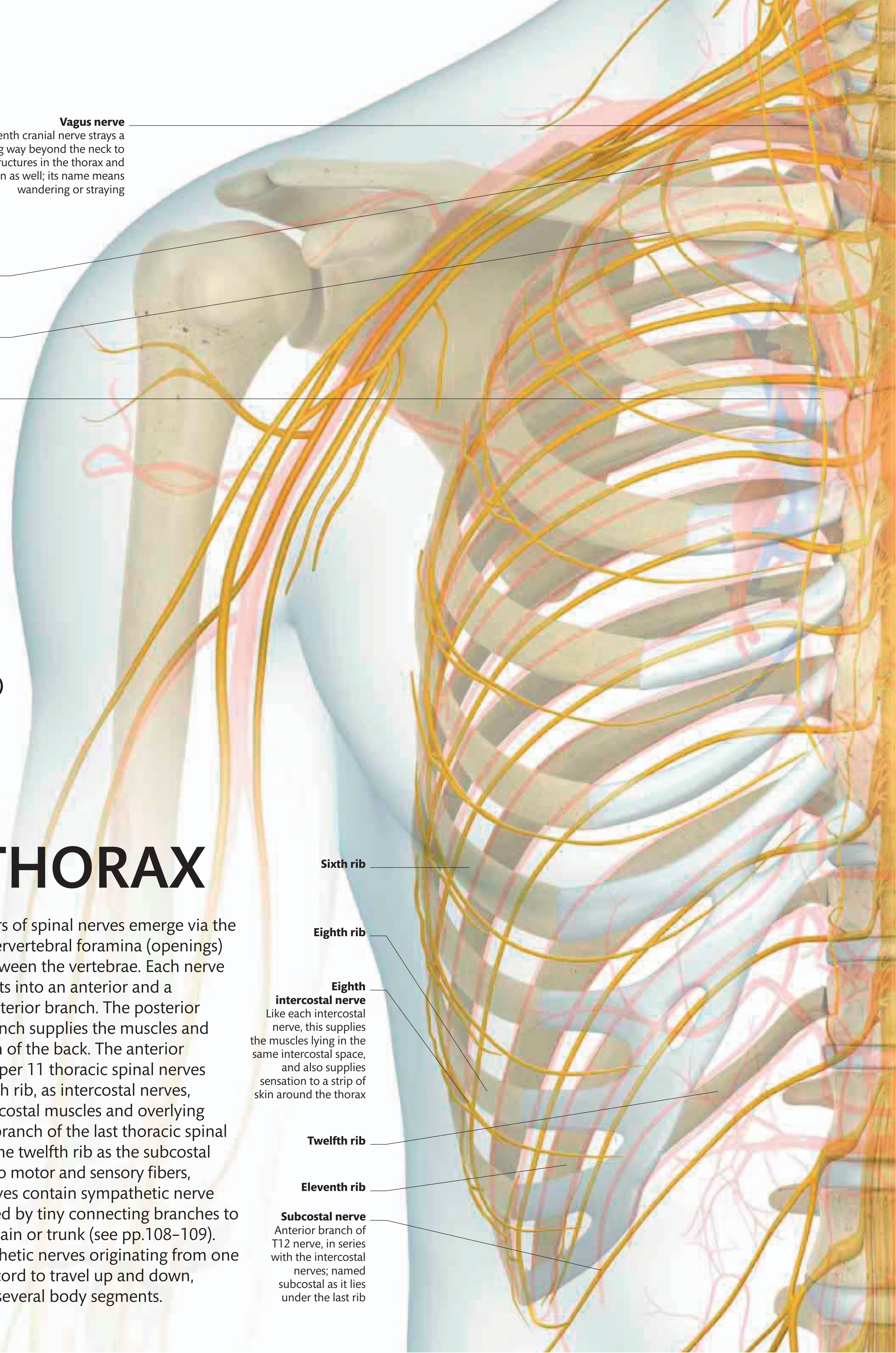

Vagus nerve
The tenth cranial nerve strays a long way beyond the neck to supply structures in the thorax and abdomen as well; its name means wandering or straying

First rib

First intercostal nerve
Anterior branch of T1 (first thoracic) spinal nerve

Phrenic nerve
Comes from the third, fourth, and fifth cervical nerves; supplies the muscle of the diaphragm and the membranes lining either side of it—the pleura on the thoracic side and peritoneum on the abdominal side

ANTERIOR (FRONT)

Sixth rib

Eighth rib

Eighth intercostal nerve
Like each intercostal nerve, this supplies the muscles lying in the same intercostal space, and also supplies sensation to a strip of skin around the thorax

Twelfth rib

Eleventh rib

Subcostal nerve
Anterior branch of T12 nerve, in series with the intercostal nerves; named subcostal as it lies under the last rib

THORAX

Pairs of spinal nerves emerge via the intervertebral foramina (openings) between the vertebrae. Each nerve splits into an anterior and a posterior branch. The posterior branch supplies the muscles and skin of the back. The anterior branches of the upper 11 thoracic spinal nerves run, one under each rib, as intercostal nerves, supplying the intercostal muscles and overlying skin. The anterior branch of the last thoracic spinal nerve runs under the twelfth rib as the subcostal nerve. In addition to motor and sensory fibers, thoracic spinal nerves contain sympathetic nerve fibers that are linked by tiny connecting branches to the sympathetic chain or trunk (see pp.108–109). This allows sympathetic nerves originating from one level of the spinal cord to travel up and down, and spread out to several body segments.

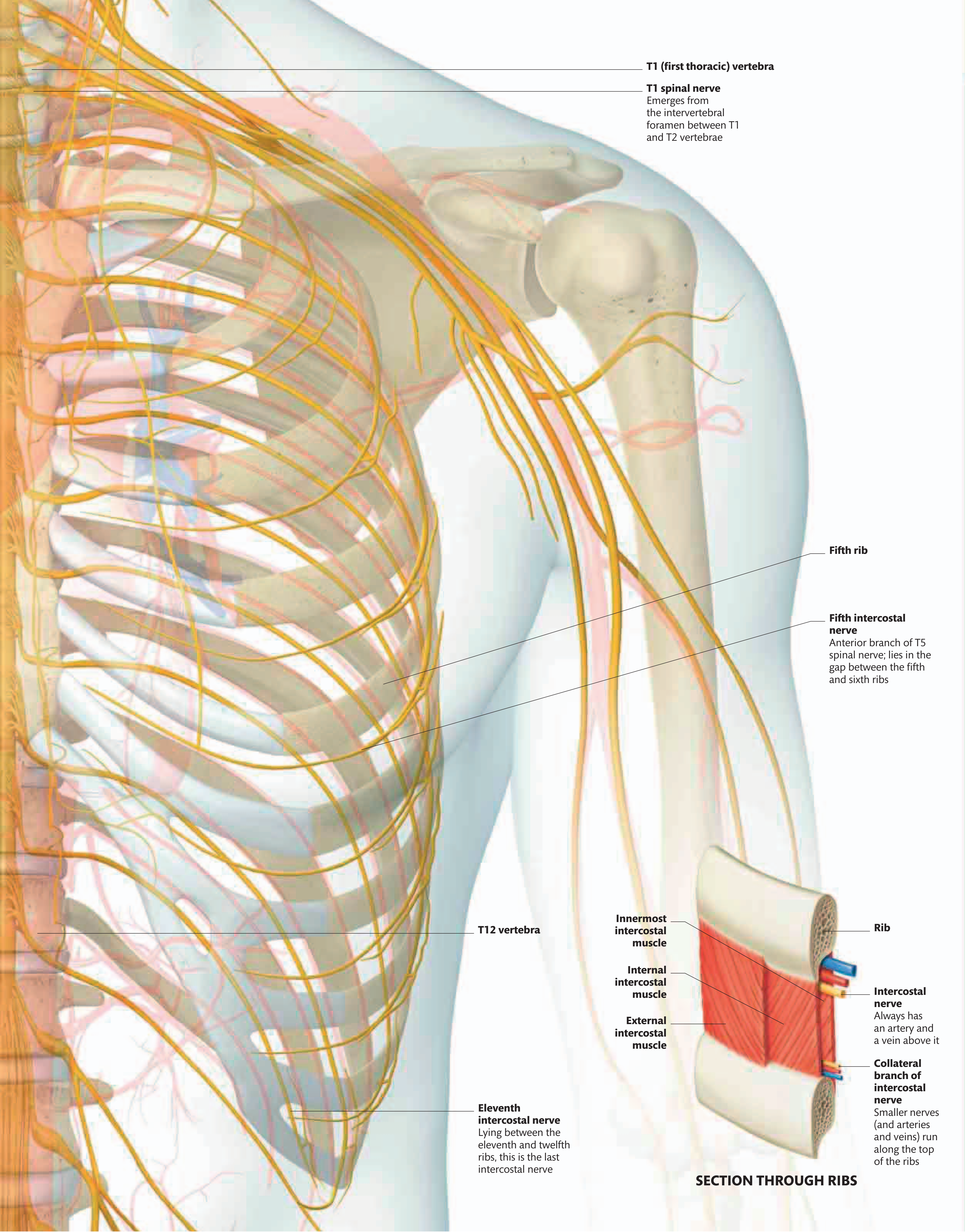
T1 (first thoracic) vertebra
T1 spinal nerve
Emerges from the intervertebral foramen between T1 and T2 vertebrae
Fifth rib
Fifth intercostal nerve
Anterior branch of T5 spinal nerve; lies in the gap between the fifth and sixth ribs
T12 vertebra
Eleventh intercostal nerve
Lying between the eleventh and twelfth ribs, this is the last intercostal nerve
Innermost intercostal muscle
Internal intercostal muscle
External intercostal muscle
Rib
Intercostal nerve
Always has an artery and a vein above it
Collateral branch of intercostal nerve
Smaller nerves (and arteries and veins) run along the top of the ribs

SECTION THROUGH RIBS

ABDOMEN AND PELVIS

The lower intercostal nerves continue past the lower edges of the ribcage at the front to supply the muscles and skin of the abdominal wall. The lower parts of the abdomen are supplied by the subcostal and iliohypogastric nerves. The abdominal portion of the sympathetic trunk receives nerves from the thoracic and first two lumbar spinal nerves, and sends nerves back to all the spinal nerves. The lumbar spinal nerves emerge from the spine and run into the psoas major muscle at the back of the abdomen. Inside the muscle, the nerves join up and swap fibers to form a network, or plexus. Branches of this lumbar plexus emerge around and through the psoas muscle and make their way into the thigh. Lower down, branches of the sacral plexus supply pelvic organs and enter the buttock. One of these branches, the sciatic nerve, is the largest nerve in the entire body. It supplies the back of the thigh, as well as the rest of the leg and foot.

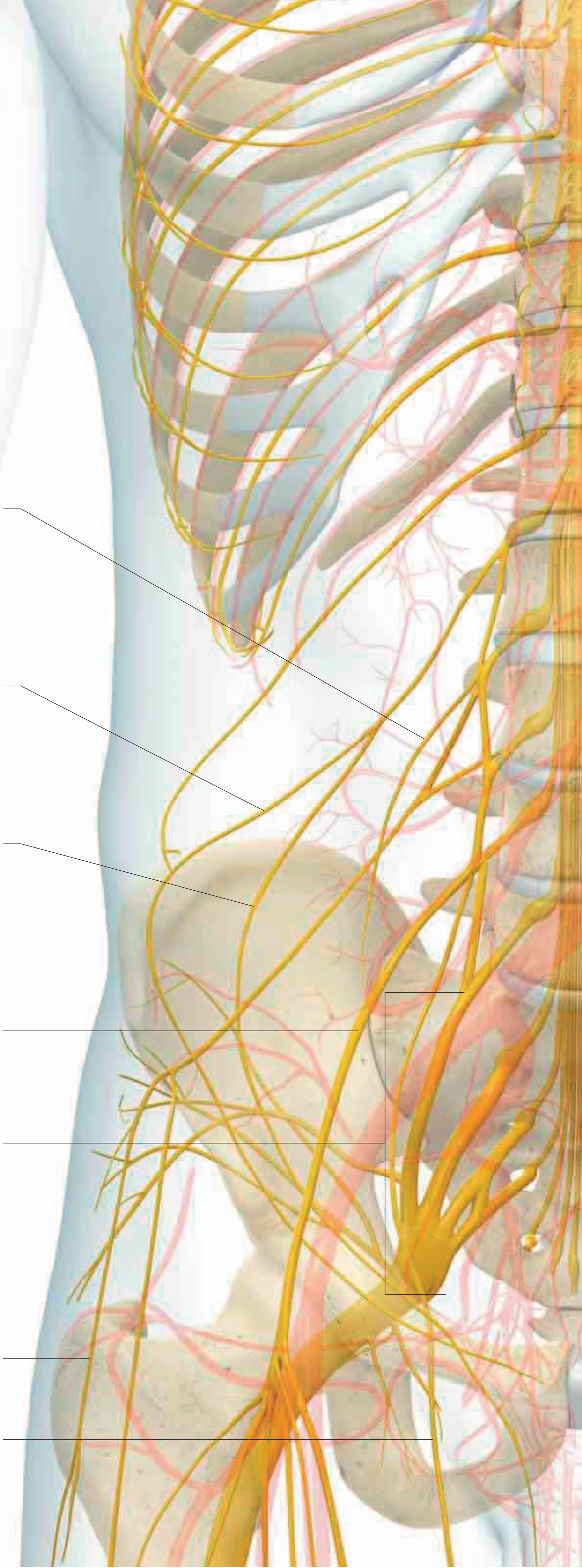

Genitofemoral nerve
Splits into two branches: the genital branch supplies some of the scrotum or labium majus, while the femoral branch supplies a small patch of skin at the top of the thigh

Iliohypogastric nerve
Runs around the side of the lower abdomen to supply the lowest parts of the muscles and skin of the abdominal wall

Ilioinguinal nerve
Travels through the layers of the abdominal wall, then down to supply sensation in the front of the scrotum in the male, or the labia majora in the female

Femoral nerve
Supplies the front of the thigh

Sacral plexus
Nerve roots from the fourth and fifth lumbar nerves join the upper four sacral nerves to form this network. Pelvic splanchnic nerves come from the second to fourth sacral nerve roots, and convey parasympathetic nerve fibers to the pelvic organs, via the pelvic plexus on each side

Lateral cutaneous nerve of the thigh
Supplies the skin of the side of the thigh

Obturator nerve
Travels along the inside of the pelvis, then emerges through the obturator foramen to supply the inner thigh

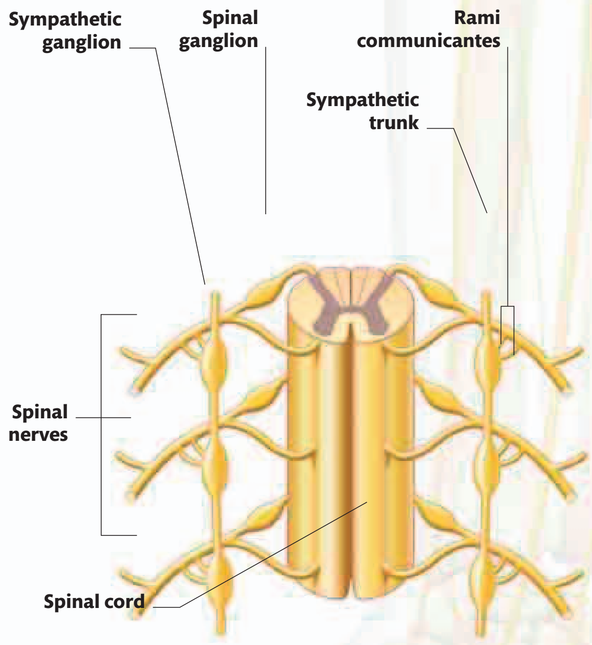

Section of sympathetic trunk and spinal cord
Branches from the sympathetic trunk innervate the organs of the abdomen and pelvis

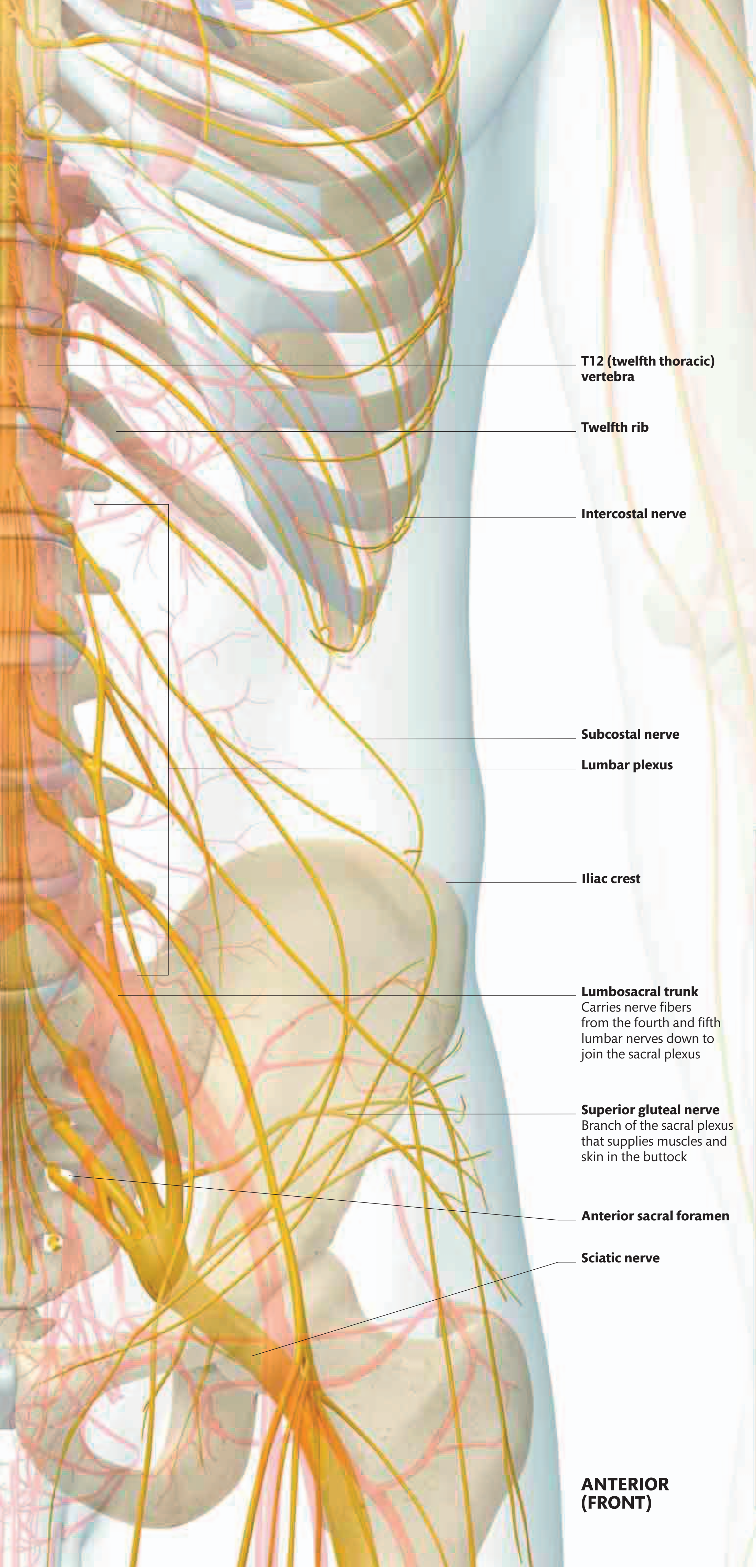
T12 (twelfth thoracic) vertebra
Twelfth rib
Intercostal nerve
Subcostal nerve
Lumbar plexus
Iliac crest
Lumbosacral trunk Carries nerve fibers from the fourth and fifth lumbar nerves down to join the sacral plexus
Superior gluteal nerve Branch of the sacral plexus that supplies muscles and skin in the buttock
Anterior sacral foramen
Sciatic nerve

ANTERIOR (FRONT)

STRUCTURE OF THE SPINAL CORD

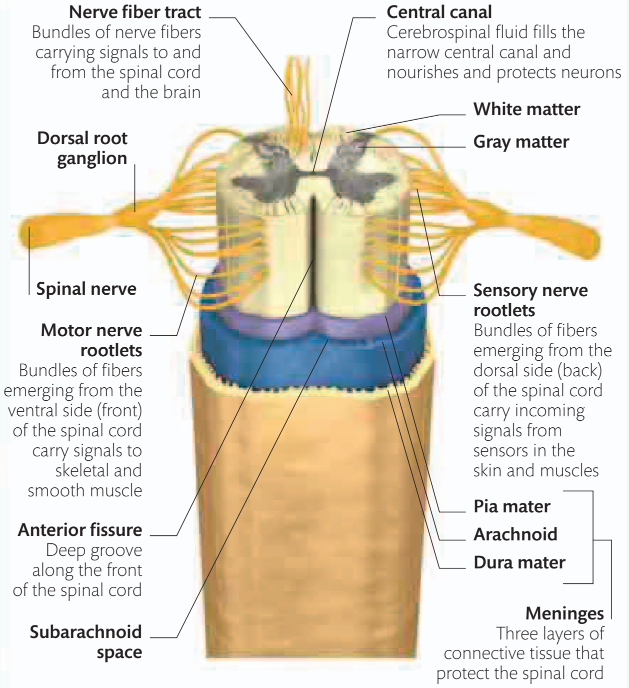
Nerve fiber tract Bundles of nerve fibers carrying signals to and from the spinal cord and the brain
Central canal Cerebrospinal fluid fills the narrow central canal and nourishes and protects neurons
White matter
Gray matter
Dorsal root ganglion
Spinal nerve
Motor nerve rootlets Bundles of fibers emerging from the ventral side (front) of the spinal cord carry signals to skeletal and smooth muscle
Sensory nerve rootlets Bundles of fibers emerging from the dorsal side (back) of the spinal cord carry incoming signals from sensors in the skin and muscles
Anterior fissure Deep groove along the front of the spinal cord
Pia mater
Arachnoid
Dura mater
Meninges Three layers of connective tissue that protect the spinal cord
Subarachnoid space

SPINAL CORD

The walls of the abdomen and pelvis are supplied by nerves emerging from the spinal cord. Like the brain, the spinal cord contains gray matter (mostly neuron cell bodies) and white matter (axons), and is covered in the same three layers of meninges: dura mater, arachnoid, and pia mater (see p.119).

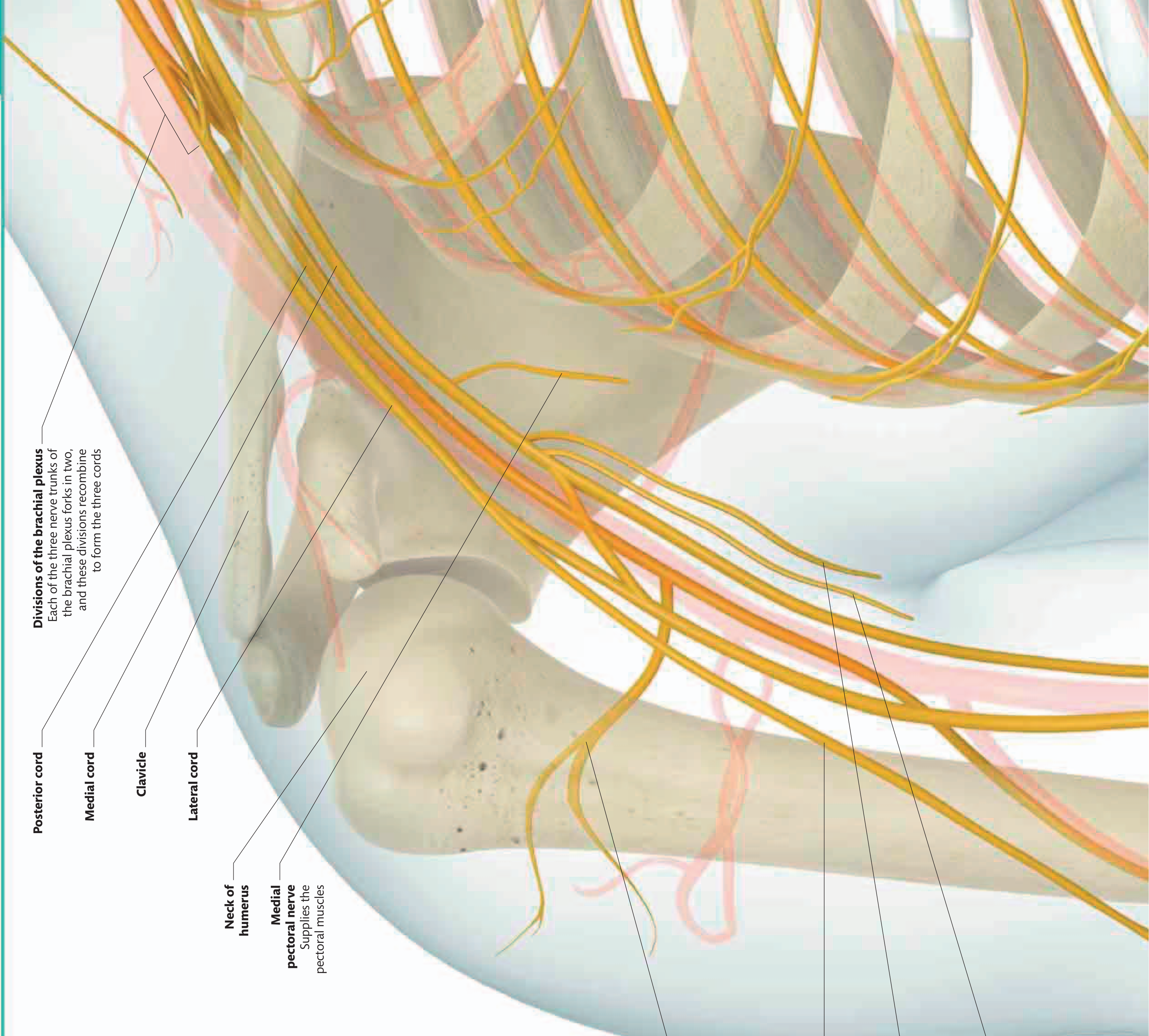
Divisions of the brachial plexus Each of the three nerve trunks of the brachial plexus forks in two, and these divisions recombine to form the three cords
Posterior cord
Medial cord
Clavicle
Lateral cord
Neck of humerus
Medial pectoral nerve Supplies the pectoral muscles
Axillary nerve Supplies deltoid and teres minor muscles, as well as sensation to skin over the outer shoulder, and to the shoulder joint itself
Musculocutaneous nerve Supplies coracobrachialis, biceps, and brachialis muscles in the upper arm
Medial cutaneous nerve of the arm Supplies the skin on the lower, inner part of the upper arm (shown cut here)
Medial cutaneous nerve of the forearm Supplies the skin of the anterior and medial surfaces of the forearm (shown cut here)

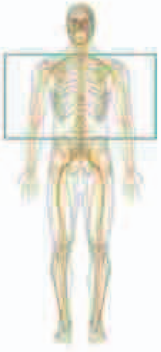

SHOULDER AND UPPER ARM

The upper limb (shoulder to hand) is supplied by five thick nerve roots that branch from the last four cervical and first thoracic spinal nerves. Emerging between the scalene muscles of the neck, they link up to form a complex skein (network) called the brachial plexus that dives under the clavicle to enter the axilla—the space between the upper arm and the chest. At this point, the plexus comprises three cords lying around the axillary artery. The network's five major nerves—musculocutaneous, median, ulnar, axillary, and radial—provide sensation to the upper limb and supply its muscles. The musculocutaneous nerve supplies the muscles in the front of the arm: the biceps, brachialis, and coracobrachialis.

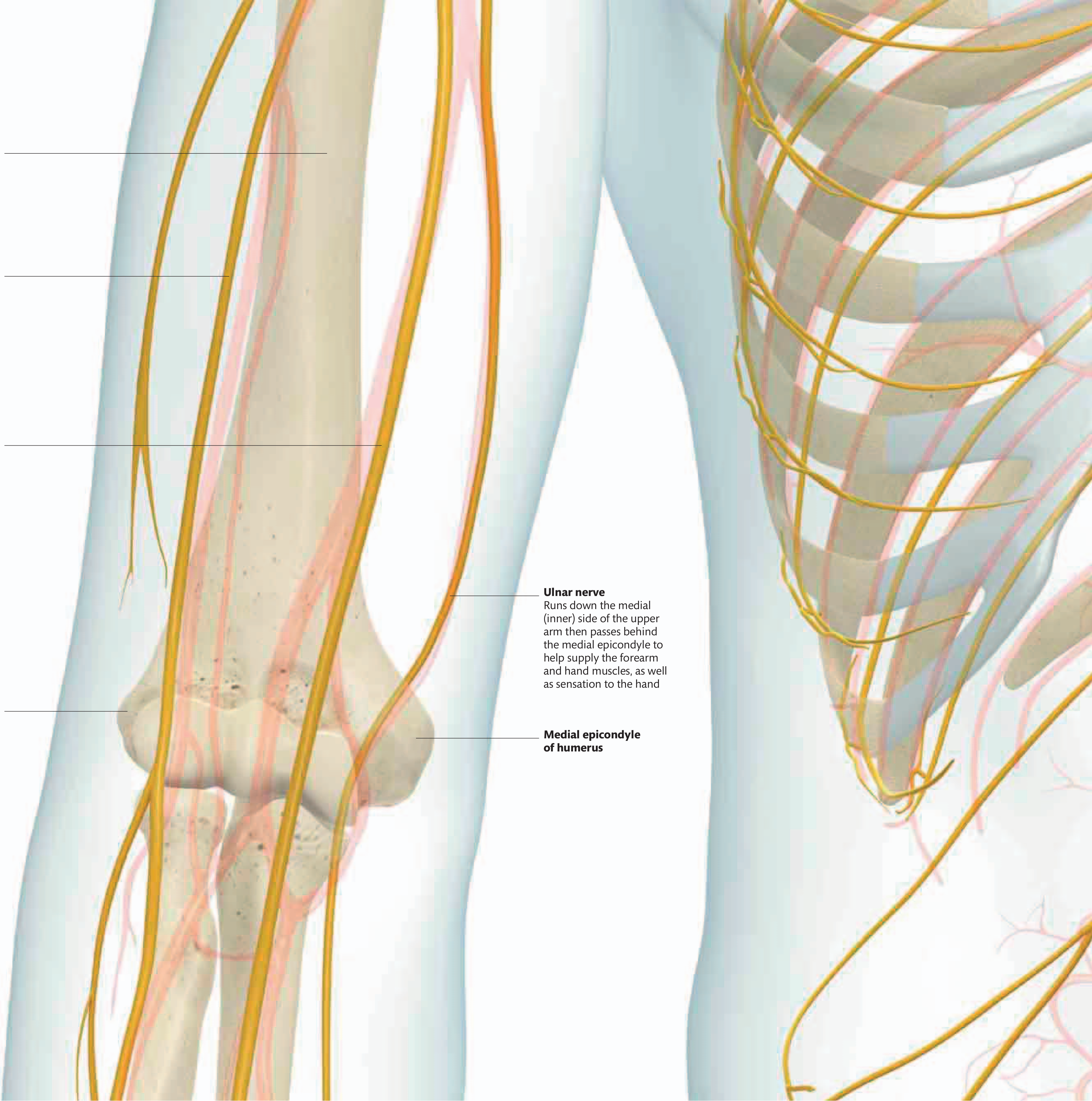
Shaft of humerus
Radial nerve
Largest branch of the brachial plexus; wraps all the way around the back of the humerus, then comes to lie in front of the lateral epicondyle; supplies muscles and sensation on the back of the arm
Median nerve
Pulls in nerve fibers from both the medial and lateral cords of the brachial plexus; lies close to the brachial artery as it runs down the upper arm, on its way to supply the forearm and hand
Ulnar nerve
Runs down the medial (inner) side of the upper arm then passes behind the medial epicondyle to help supply the forearm and hand muscles, as well as sensation to the hand
Lateral epicondyle of humerus
Medial epicondyle of humerus
ANTERIOR (FRONT)

SHOULDER AND UPPER ARM

The axillary and radial nerves emerge from the back of the brachial plexus and run behind the humerus. The axillary nerve wraps around the neck of the humerus, just underneath the shoulder joint, and supplies the deltoid muscle. The radial nerve—the largest branch of the brachial plexus—supplies all the extensor muscles in the upper arm and in the forearm. It spirals around the back of the humerus, lying right against the bone, and sends branches to supply the heads of the triceps. The radial nerve then continues in its spiral, running forward to lie just in front of the medial epicondyle of the humerus at the elbow.

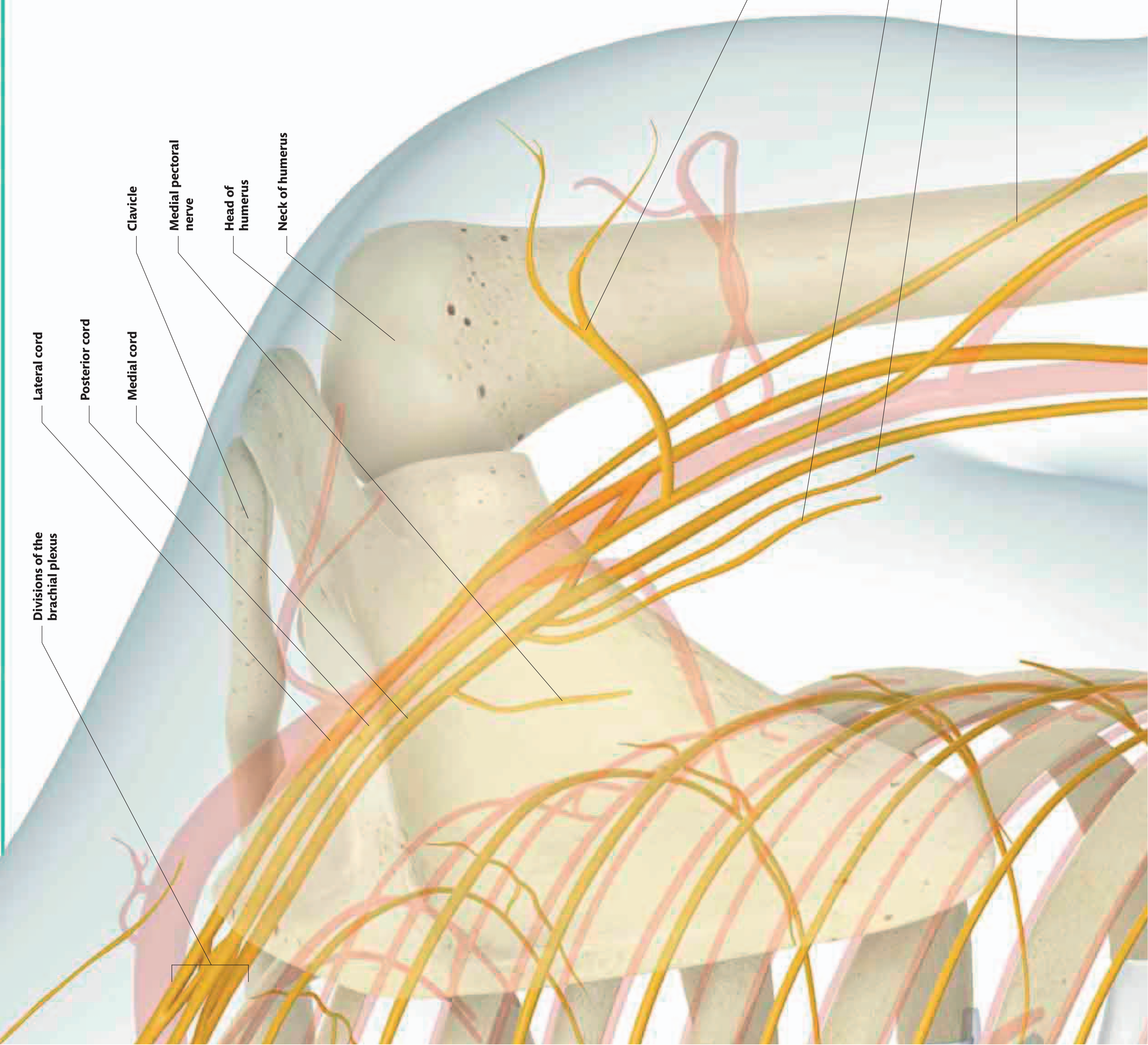
Divisions of the brachial plexus
Lateral cord
Posterior cord
Medial cord
Clavicle
Medial pectoral nerve
Head of humerus
Neck of humerus

Axillary nerve
Wraps around the neck of the humerus, just below the shoulder joint, and is at risk of damage if the shoulder dislocates

Medial cutaneous nerve of the arm

Medial cutaneous nerve of the forearm

Musculocutaneous nerve
After supplying muscles in the front of the arm, this continues as a cutaneous nerve, which will supply sensation to the skin of the lateral (outer) forearm

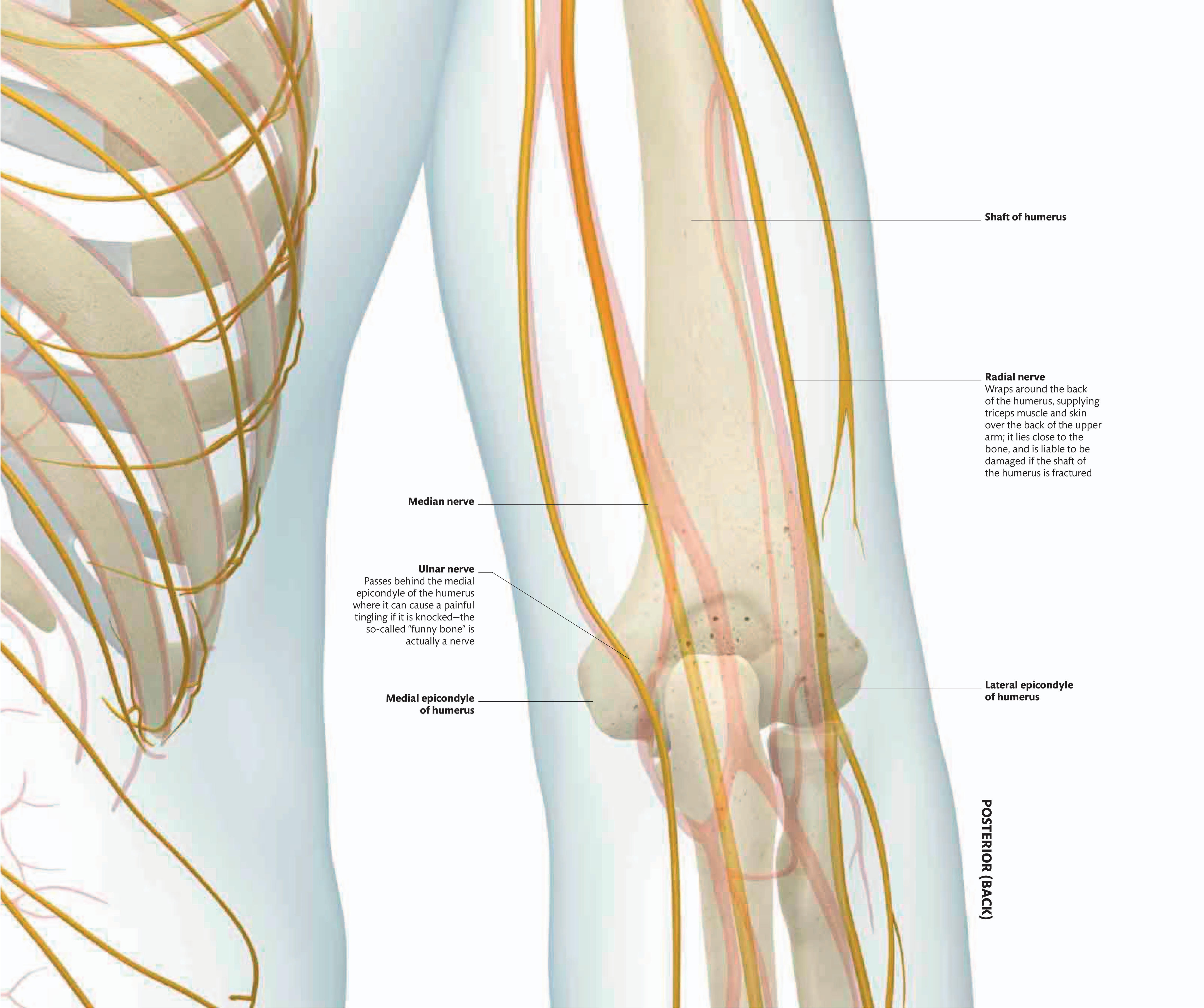

POSTERIOR (BACK)

Shaft of humerus

Radial nerve
Wraps around the back of the humerus, supplying triceps muscle and skin over the back of the upper arm; it lies close to the bone, and is liable to be damaged if the shaft of the humerus is fractured

Median nerve

Ulnar nerve
Passes behind the medial epicondyle of the humerus where it can cause a painful tingling if it is knocked—the so-called "funny bone" is actually a nerve

Medial epicondyle of humerus

Lateral epicondyle of humerus

POSTERIOR (BACK)

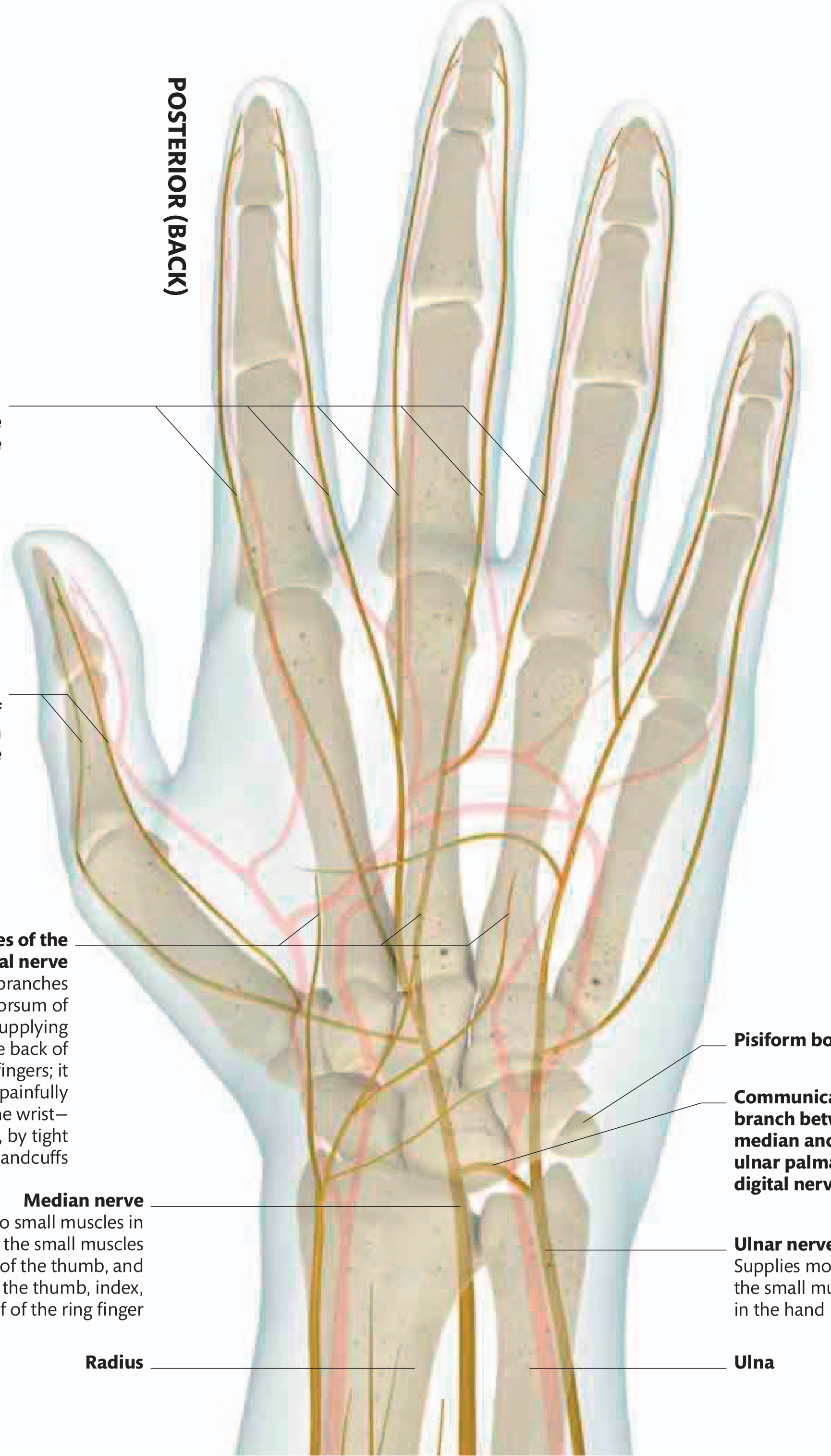

Digital branches of the median nerve

Digital branches of the median nerve

Branches of the superficial radial nerve
This nerve branches across the dorsum of the hand, supplying sensation to the back of the hand and fingers; it can become painfully compressed at the wrist—for example, by tight bracelets or handcuffs

Median nerve
Supplies two small muscles in the hand, the small muscles in the ball of the thumb, and sensation to the thumb, index, middle, and half of the ring finger

Radius

Pisiform bone

Communicating branch between median and ulnar palmar digital nerves

Ulnar nerve
Supplies most of the small muscles in the hand

Ulna

ANTERIOR (FRONT)

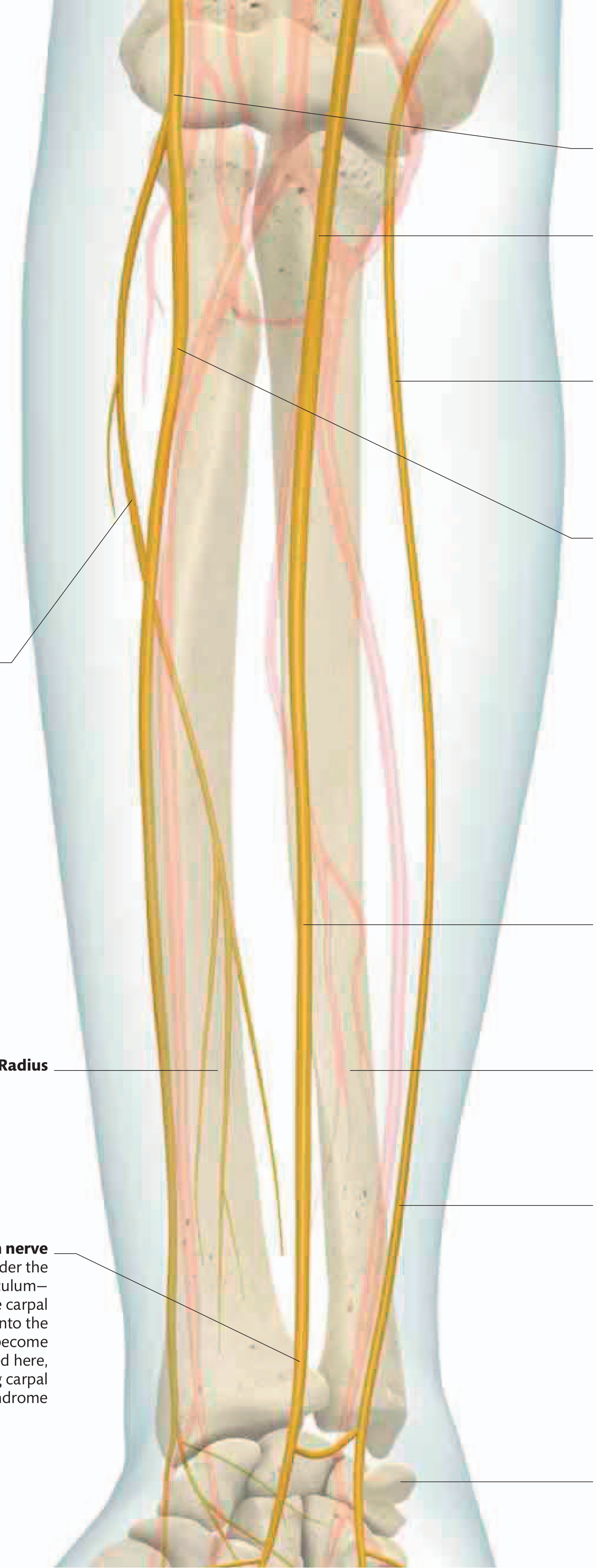

Posterior interosseous nerve
Branch of the radial nerve that passes backward around the radius

Radius

Median nerve
Passes under the flexor retinaculum—through the carpal tunnel—into the wrist; it may become compressed here, producing carpal tunnel syndrome

Radial nerve
Branches at the elbow

Median nerve
Passes through a triangular area on the inside of the elbow known as the cubital fossa

Ulnar nerve
Supplies flexor carpi ulnaris and some of flexor digitorum profundus muscles in the forearm

Superficial radial nerve
Branch of the radial nerve that runs down to the wrist, just tucked underneath brachioradialis muscle on the outer (radial) side of the forearm

Median nerve
Supplies most of the flexor muscles in the front of the forearm

Ulna

Ulnar nerve
Passes into the hand, close to the pisiform bone, where it can get compressed—for example, when using vibrating machinery or when holding a motorbike handlebar

Pisiform bone

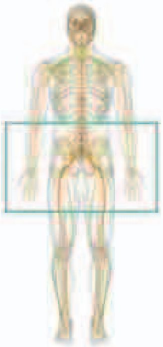

LOWER ARM AND HAND

The front of the forearm is supplied by the musculocutaneous, median, and ulnar nerves. The musculocutaneous nerve supplies sensation to the lateral side of the forearm. The median nerve runs down the middle of the forearm, supplying most of the flexor muscles. It then travels over the wrist and into the hand to supply some of the thumb muscles, as well as sensation to the palm, thumb, and some fingers. The ulnar nerve courses down the inner side of the forearm, where it supplies just two muscles. It continues on to supply most of the small muscles in the hand and provide sensation to the inner side of the ring finger, and also the little finger. On the back of the forearm, the radial nerve and its branches supply all the extensor muscles. Branches of the radial nerve fan out over the back of the hand, where they provide sensation.

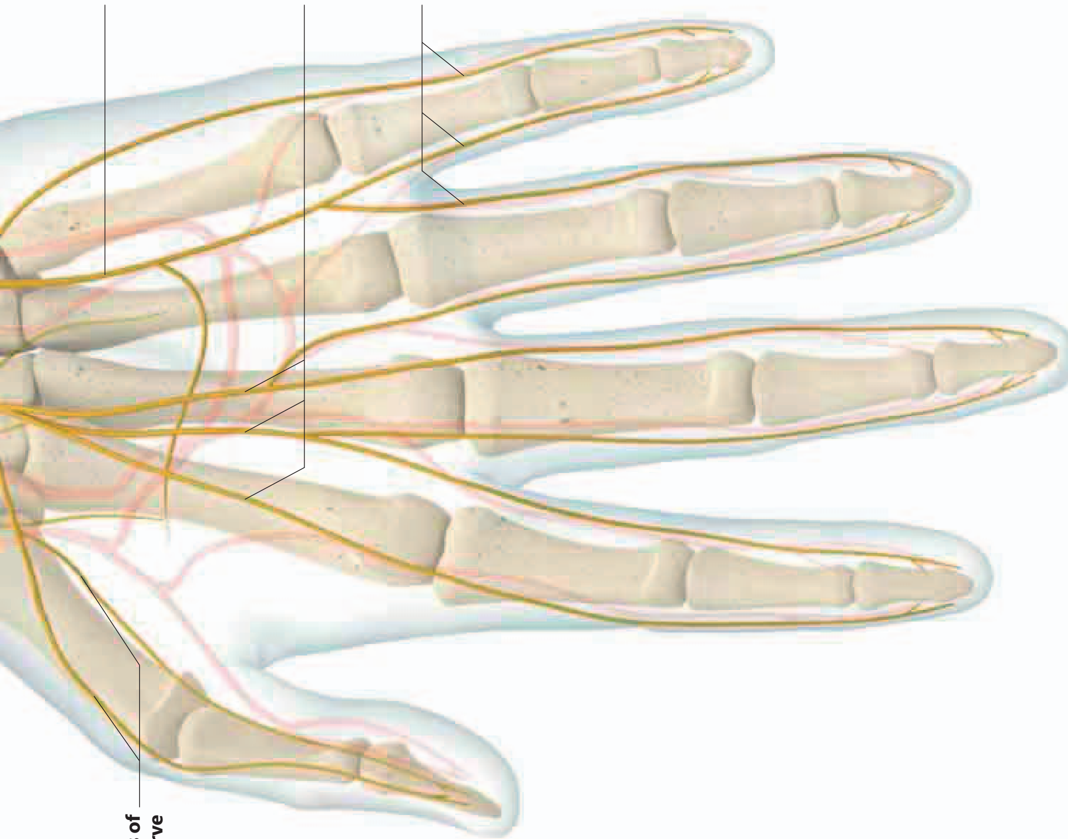
Palmar branch of the ulnar nerve Supplies skin over the palm, and palmaris brevis muscle
Digital branches of median nerve
Digital branches of ulnar nerve
Digital branches of median nerve

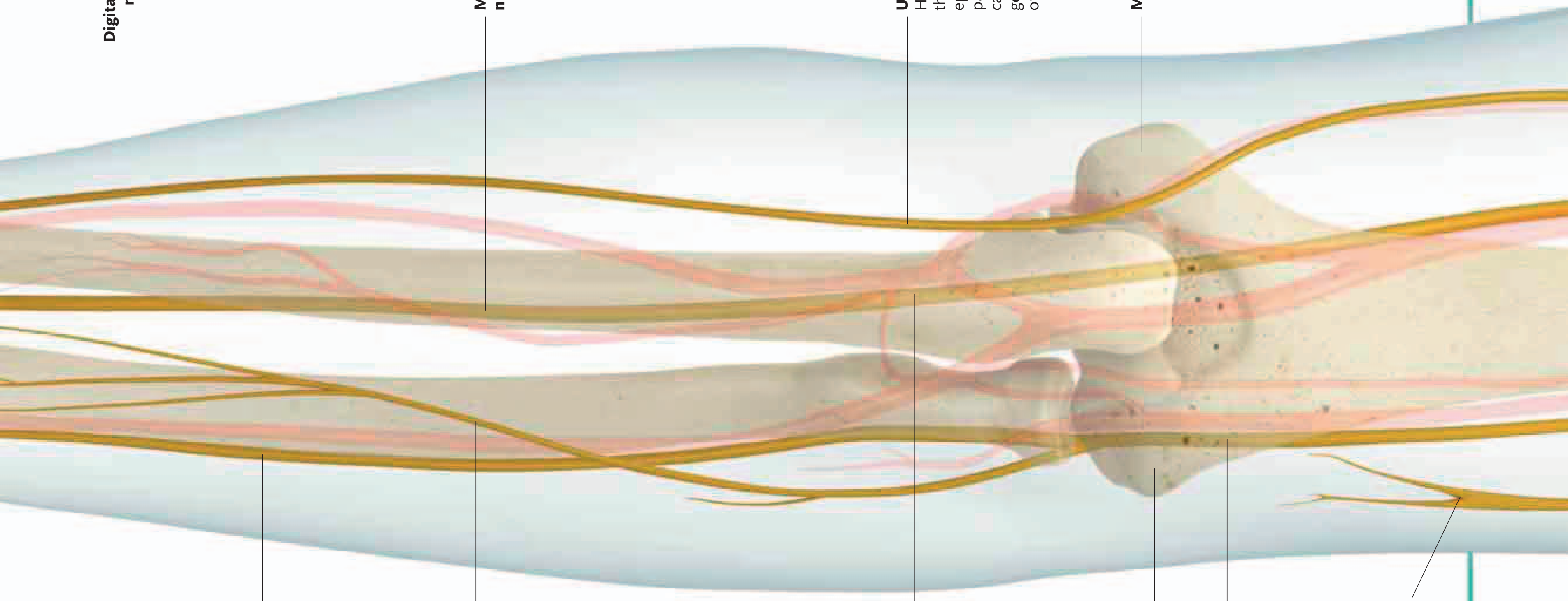
Median nerve
Ulnar nerve Having passed around the back of the medial epicondyle, this nerve passes through flexor carpi ulnaris muscle to get back into the front of the forearm
Medial epicondyle
Superficial branch of radial nerve
Posterior interosseous nerve Lies on the back of the interosseous membrane; supplies the extensor muscles and skin of the back of the forearm
Median nerve Lies on the medial (inner) side of the brachial artery at the elbow
Lateral epicondyle
Radial nerve Passes along the lateral (outer) side of the elbow
Musculocutaneous nerve

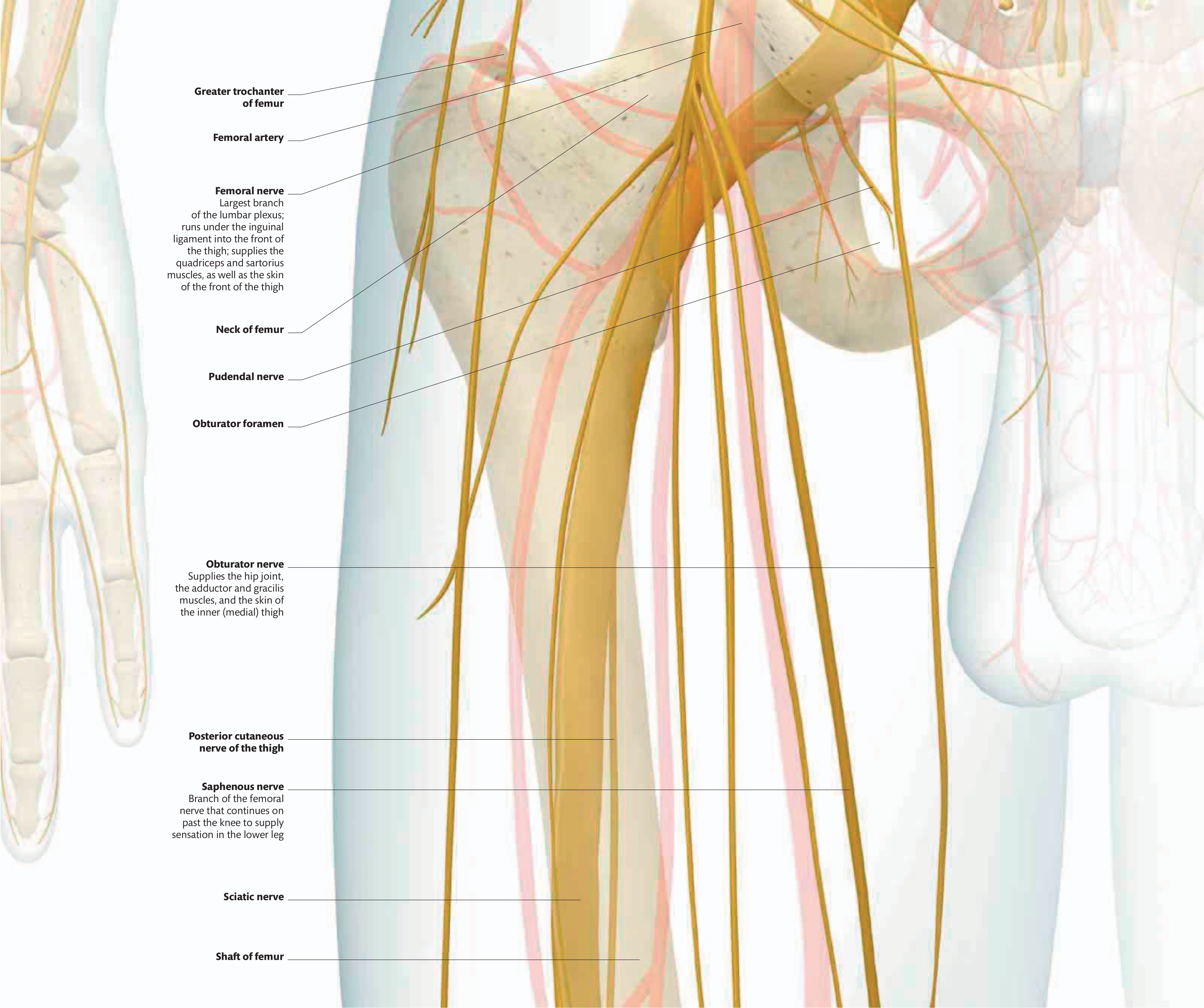

Greater trochanter of femur

Femoral artery

Femoral nerve
Largest branch of the lumbar plexus; runs under the inguinal ligament into the front of the thigh; supplies the quadriceps and sartorius muscles, as well as the skin of the front of the thigh

Neck of femur

Pudendal nerve

Obturator foramen

Obturator nerve
Supplies the hip joint, the adductor and gracilis muscles, and the skin of the inner (medial) thigh

Posterior cutaneous nerve of the thigh

Saphenous nerve
Branch of the femoral nerve that continues on past the knee to supply sensation in the lower leg

Sciatic nerve

Shaft of femur

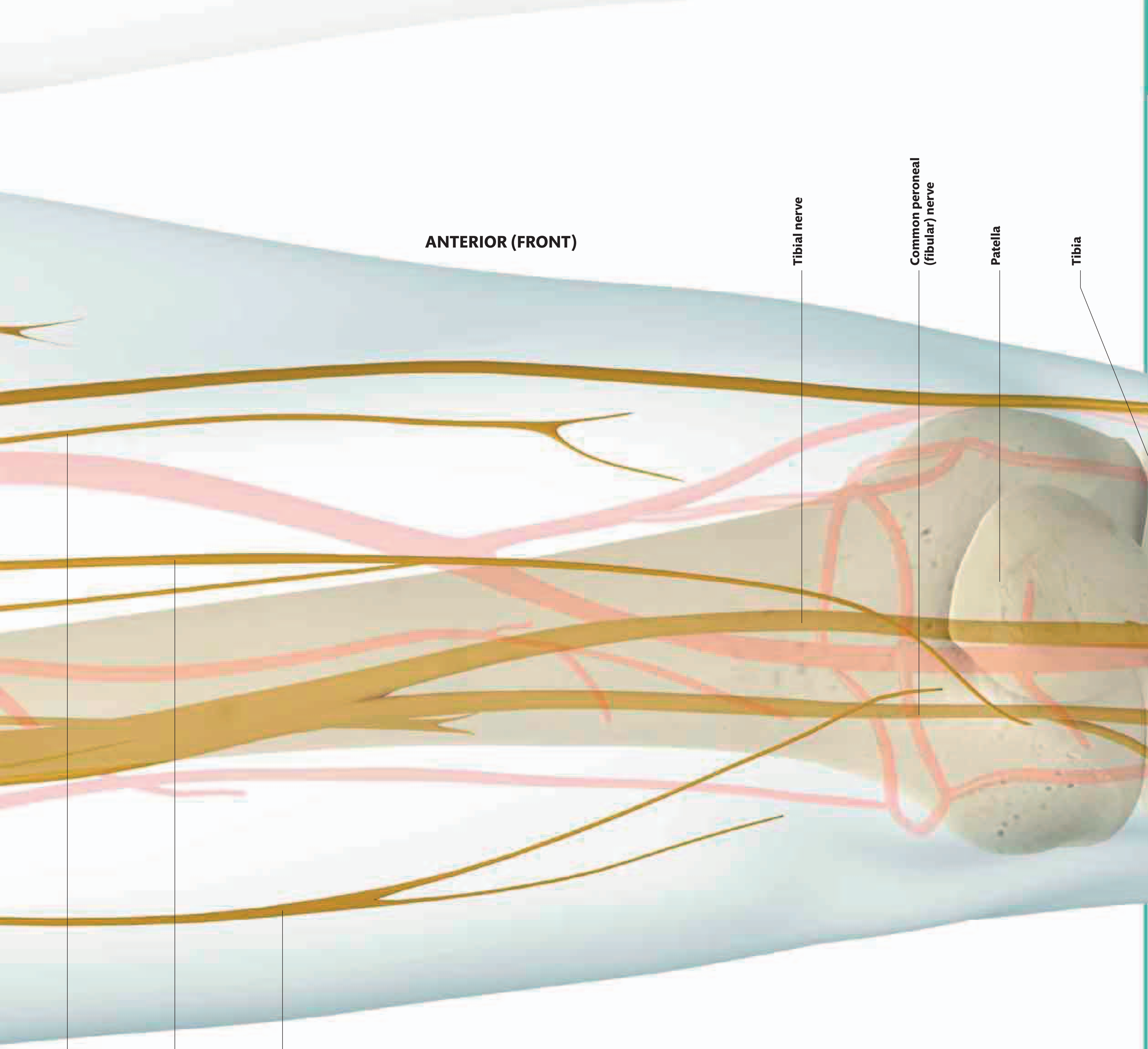
ANTERIOR (FRONT)
Tibial nerve
Common peroneal (fibular) nerve
Patella
Tibia
Medial femoral cutaneous nerve
A branch of the femoral nerve
Intermediate femoral cutaneous nerve
Also a branch of the femoral nerve
Lateral femoral cutaneous nerve
Emerges under or through the inguinal ligament, to supply the skin of the upper, outer thigh

HIP AND THIGH

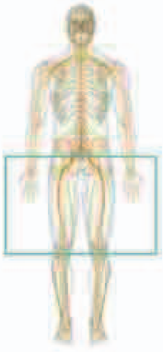

The lower limb (hip, thigh, leg, and foot) receives nerves from the lumbar and sacral plexuses. Three main nerves supply the thigh muscles: the femoral, obturator, and sciatic nerves (the last in the back). The femoral nerve runs over the pubic bone to supply the quadriceps and sartorius muscles in the front. The saphenous nerve, a slender branch of the femoral, continues past the knee and supplies skin on the inside of the lower leg and the inner side of the foot. The obturator nerve passes through the obturator foramen in the pelvic bone to supply the adductor muscles of the inner thigh and provide sensation to the skin there. Some smaller nerves just supply skin, such as the femoral cutaneous nerves.

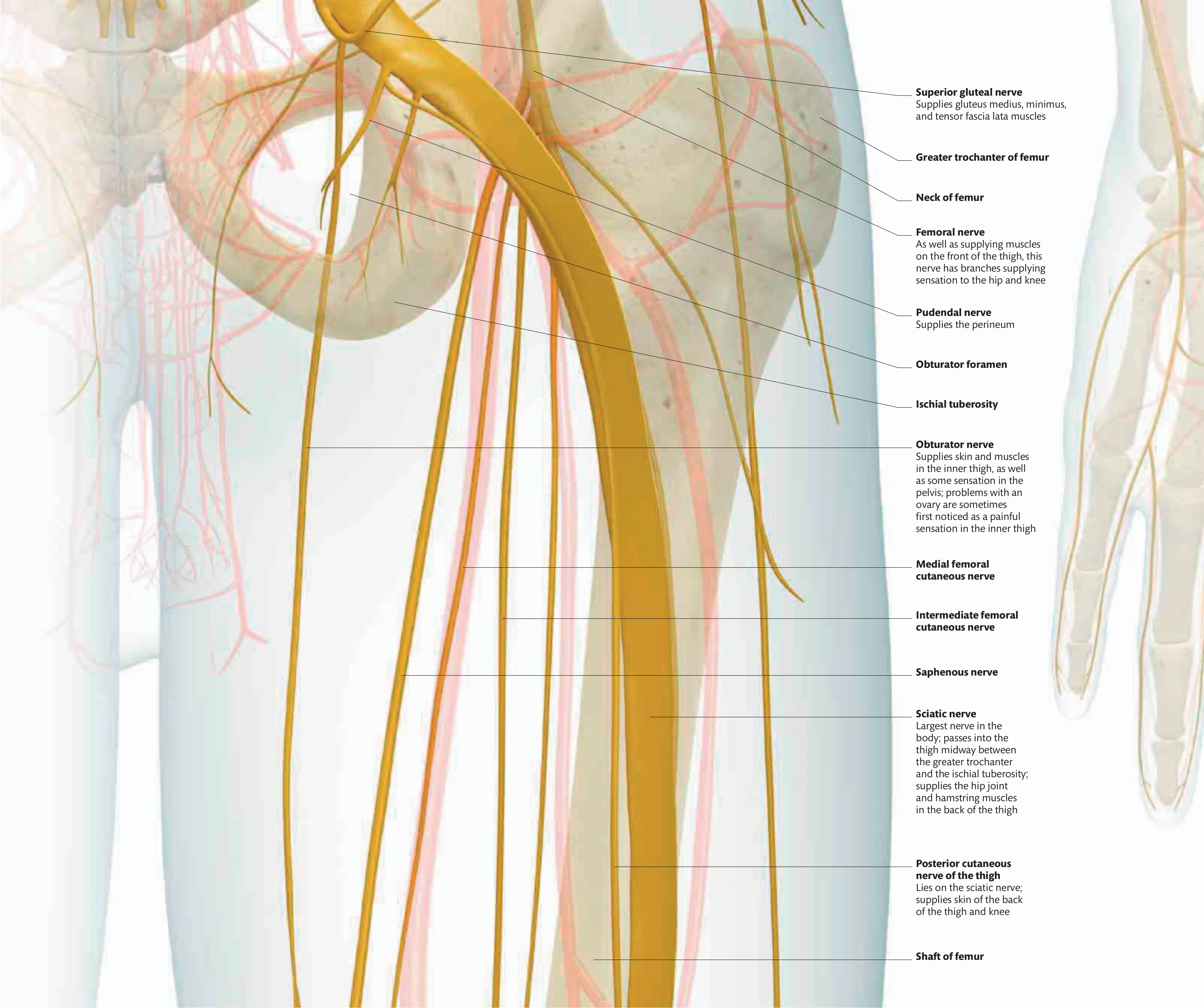

Superior gluteal nerve
Supplies gluteus medius, minimus, and tensor fascia lata muscles

Greater trochanter of femur

Neck of femur

Femoral nerve
As well as supplying muscles on the front of the thigh, this nerve has branches supplying sensation to the hip and knee

Pudendal nerve
Supplies the perineum

Obturator foramen

Ischial tuberosity

Obturator nerve
Supplies skin and muscles in the inner thigh, as well as some sensation in the pelvis; problems with an ovary are sometimes first noticed as a painful sensation in the inner thigh

Medial femoral cutaneous nerve

Intermediate femoral cutaneous nerve

Saphenous nerve

Sciatic nerve
Largest nerve in the body; passes into the thigh midway between the greater trochanter and the ischial tuberosity; supplies the hip joint and hamstring muscles in the back of the thigh

Posterior cutaneous nerve of the thigh
Lies on the sciatic nerve; supplies skin of the back of the thigh and knee

Shaft of femur

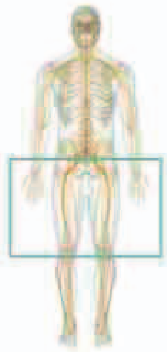

HIP AND THIGH

Gluteal nerves from the sacral plexus emerge via the greater sciatic foramen, at the back of the pelvis, to supply the muscles and skin of the buttock. The sciatic nerve also emerges through the greater sciatic foramen into the buttock. The gluteus maximus is a good site for injections into a muscle, but these should always be given in the upper, outer part of the buttock to make sure the needle is well away from the sciatic nerve. The sciatic nerve runs down the back of the thigh, supplying the hamstrings. In most people, the sciatic nerve runs halfway down the thigh then splits into two branches, the tibial and common peroneal nerves. These continue into the popliteal fossa (back of the knee) and on into the lower leg.

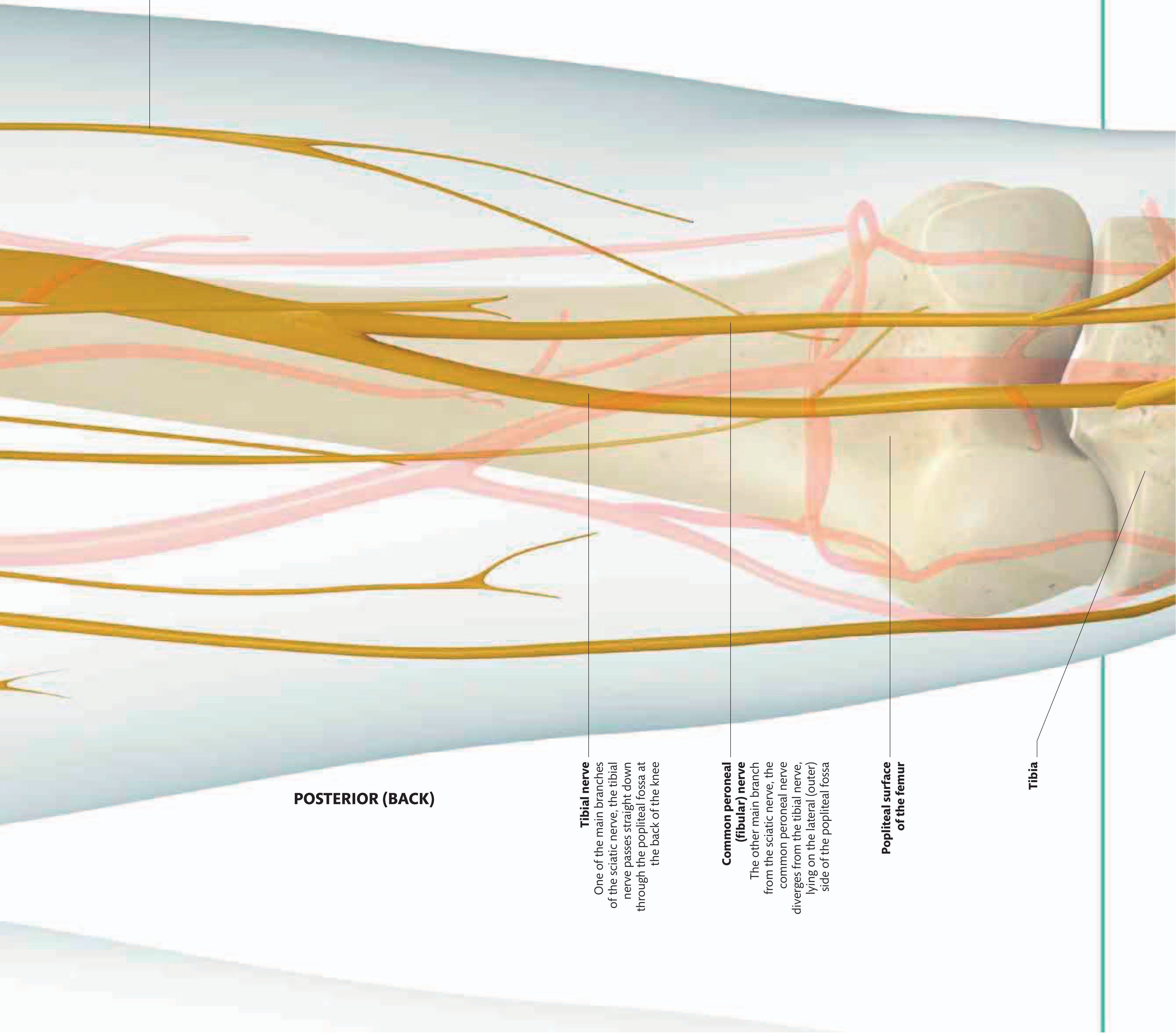

Lateral femoral cutaneous nerve
May become compressed at the inguinal ligament, causing a painful tingling in the thigh, called meralgia paresthetica

Tibial nerve
One of the main branches of the sciatic nerve, the tibial nerve passes straight down through the popliteal fossa at the back of the knee

Common peroneal (fibular) nerve
The other main branch from the sciatic nerve, the common peroneal nerve diverges from the tibial nerve, lying on the lateral (outer) side of the popliteal fossa

Popliteal surface of the femur

Tibia

POSTERIOR (BACK)

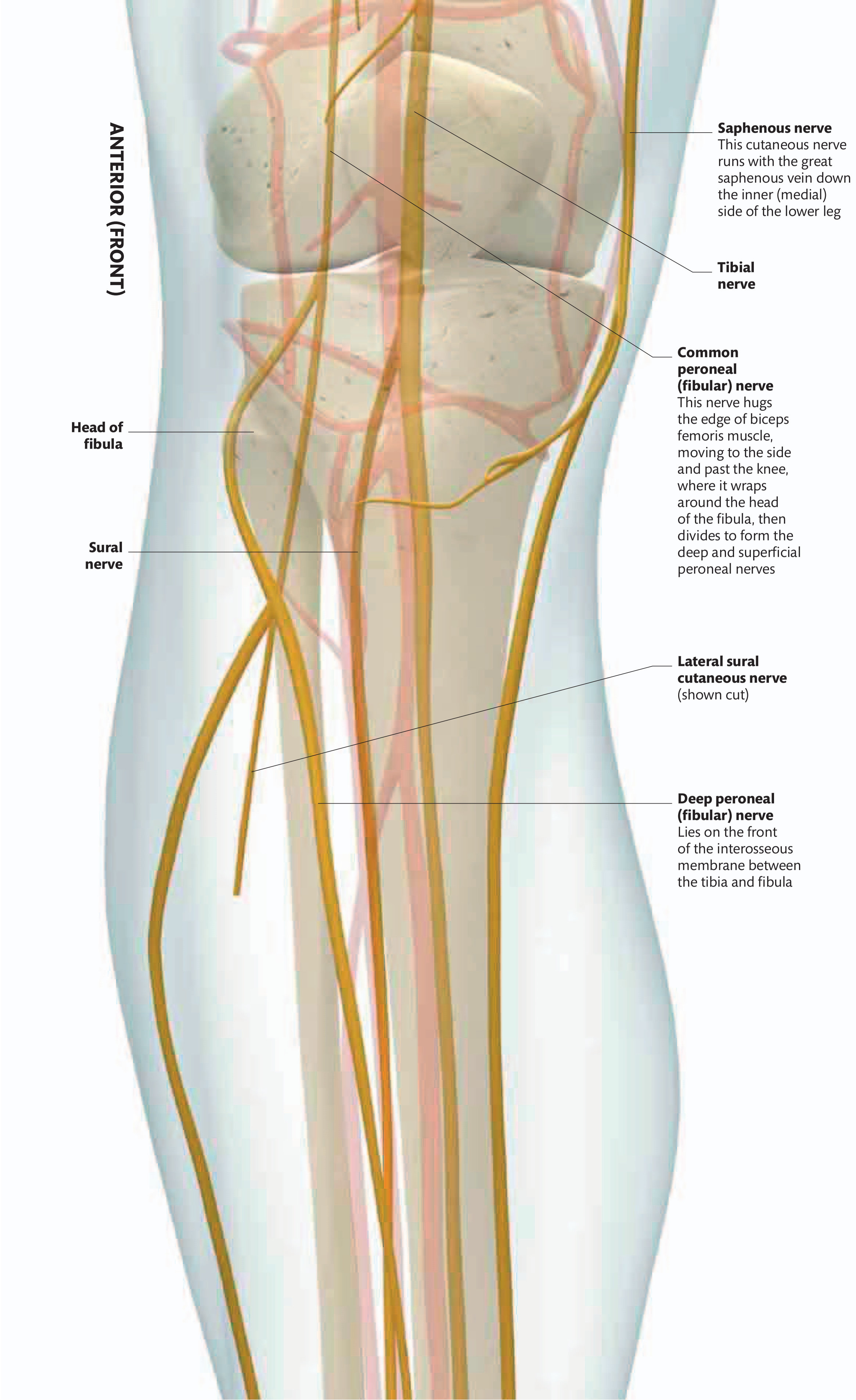
ANTERIOR (FRONT)
Saphenous nerve
This cutaneous nerve runs with the great saphenous vein down the inner (medial) side of the lower leg
Tibial nerve
Common peroneal (fibular) nerve
This nerve hugs the edge of biceps femoris muscle, moving to the side and past the knee, where it wraps around the head of the fibula, then divides to form the deep and superficial peroneal nerves
Head of fibula
Sural nerve
Lateral sural cutaneous nerve
(shown cut)
Deep peroneal (fibular) nerve
Lies on the front of the interosseous membrane between the tibia and fibula

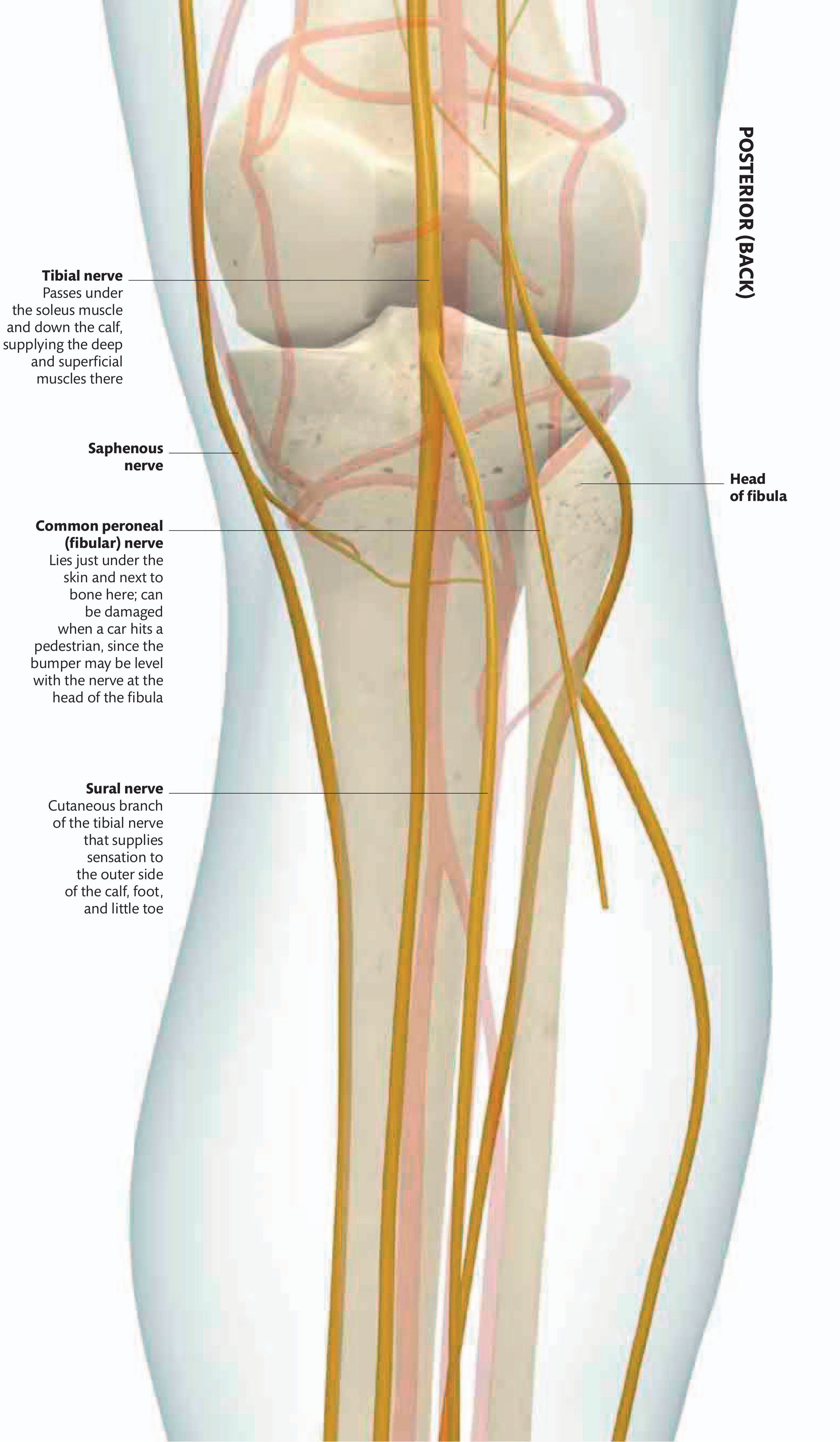
POSTERIOR (BACK)
Tibial nerve
Passes under the soleus muscle and down the calf, supplying the deep and superficial muscles there
Saphenous nerve
Head of fibula
Common peroneal (fibular) nerve
Lies just under the skin and next to bone here; can be damaged when a car hits a pedestrian, since the bumper may be level with the nerve at the head of the fibula
Sural nerve
Cutaneous branch of the tibial nerve that supplies sensation to the outer side of the calf, foot, and little toe

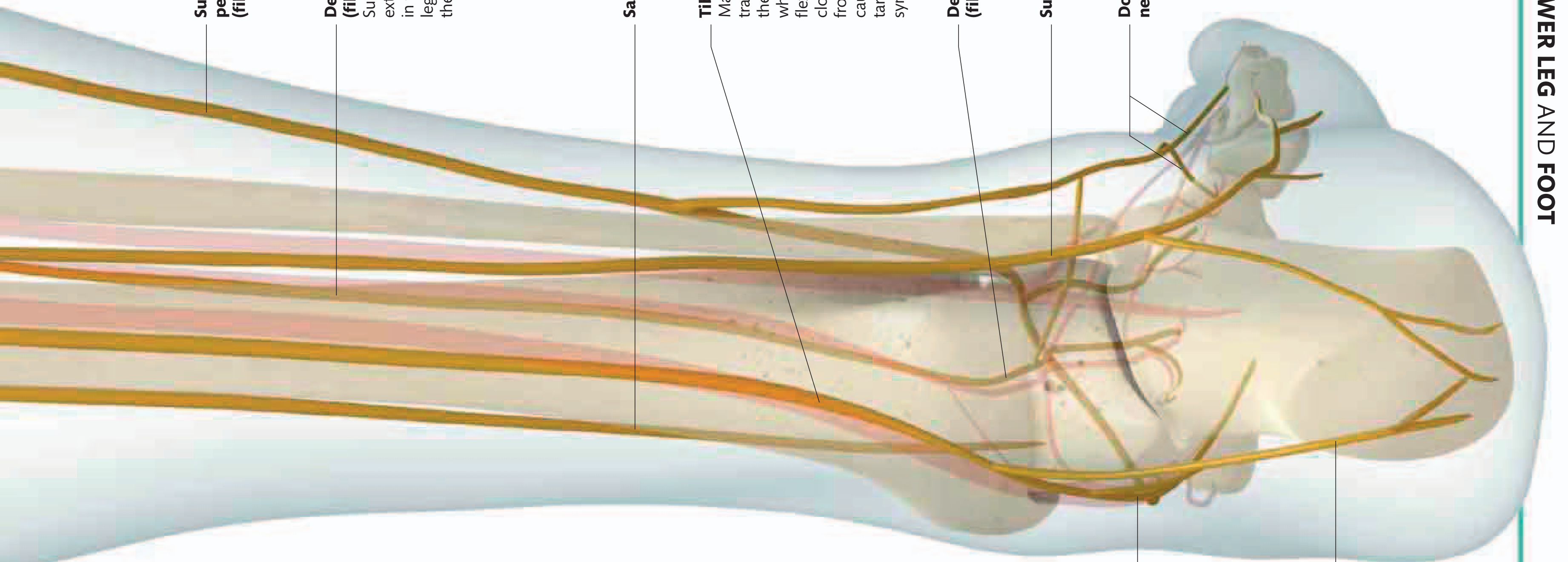

Superficial peroneal (fibular) nerve

Deep peroneal (fibular) nerve Supplies the extensor muscles in the front of the leg, as well as the ankle joint

Saphenous nerve

Tibial nerve May become trapped under the retinaculum, which holds the flexor tendons close to the front of the ankle, causing the rare tarsal tunnel syndrome

Deep peroneal (fibular) nerve

Sural nerve

Dorsal digital nerves

Medial plantar nerve One of the terminal branches of the tibial nerve, supplying the sole and toes

Calcaneal branch of tibial nerve Supplies the heel and inner (medial) sole

LOWER LEG AND FOOT

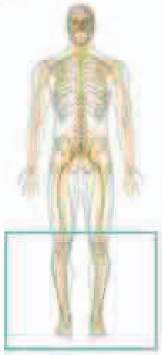

The common peroneal nerve runs past the knee and wraps around the neck of the fibula. Then it splits into the deep and superficial peroneal nerves. The deep peroneal nerve supplies the extensor muscles of the shin, then fans out to provide sensation to the skin at the back of the foot. The superficial peroneal nerve stays on the side of the leg and supplies the peroneal muscles. The tibial nerve runs through the popliteal fossa (back of the knee), under the soleus muscle, and between the deep and superficial calf muscles, which it supplies. It continues behind the medial malleolus and under the foot, then splits into two plantar nerves that supply the small muscles of the foot and the skin of the sole.

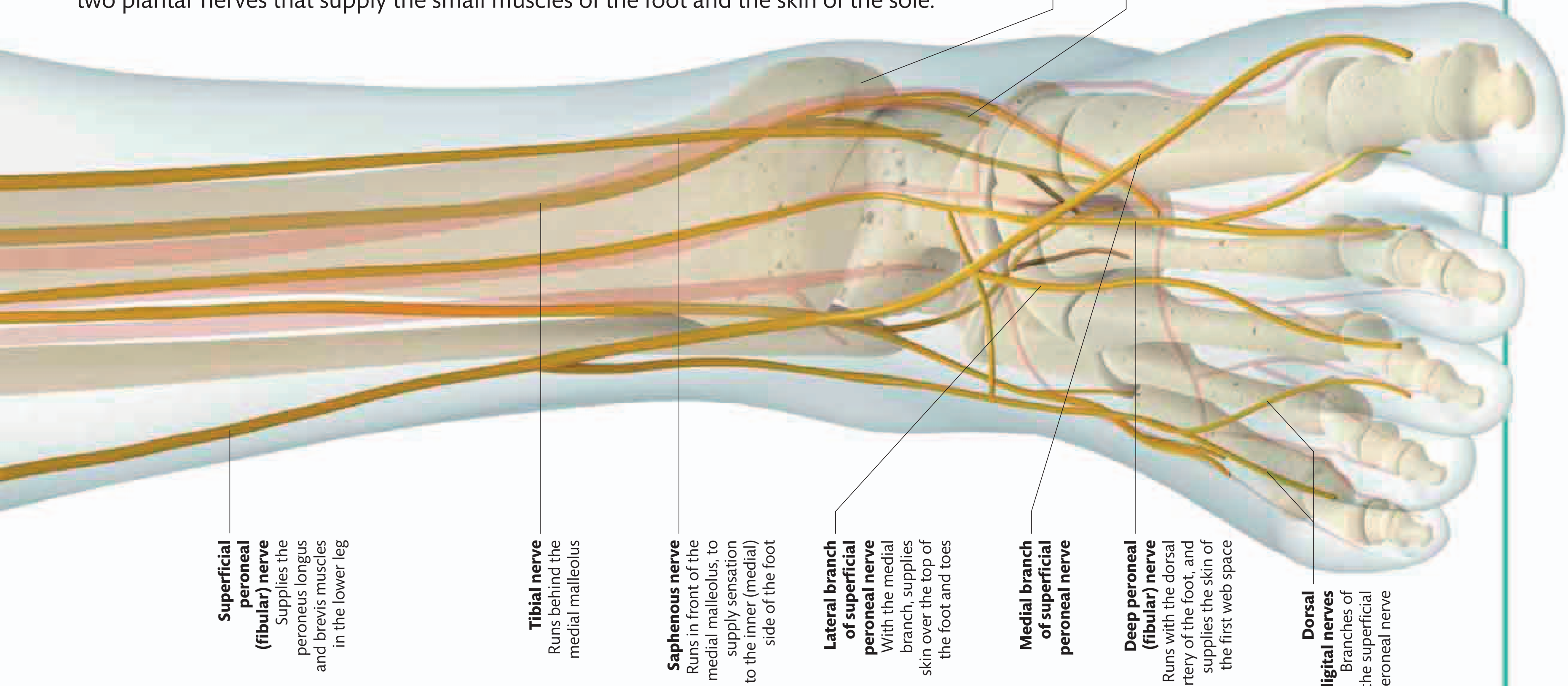

Medial malleolus

Lateral plantar nerve With the medial plantar nerve, supplies the muscles and skin of the sole and toes

Superficial peroneal (fibular) nerve Supplies the peroneus longus and brevis muscles in the lower leg

Tibial nerve Runs behind the medial malleolus

Saphenous nerve Runs in front of the medial malleolus, to supply sensation to the inner (medial) side of the foot

Lateral branch of superficial peroneal nerve With the medial branch, supplies skin over the top of the foot and toes

Medial branch of superficial peroneal nerve

Deep peroneal (fibular) nerve Runs with the dorsal artery of the foot, and supplies the skin of the first web space

Dorsal digital nerves Branches of the superficial peroneal nerve

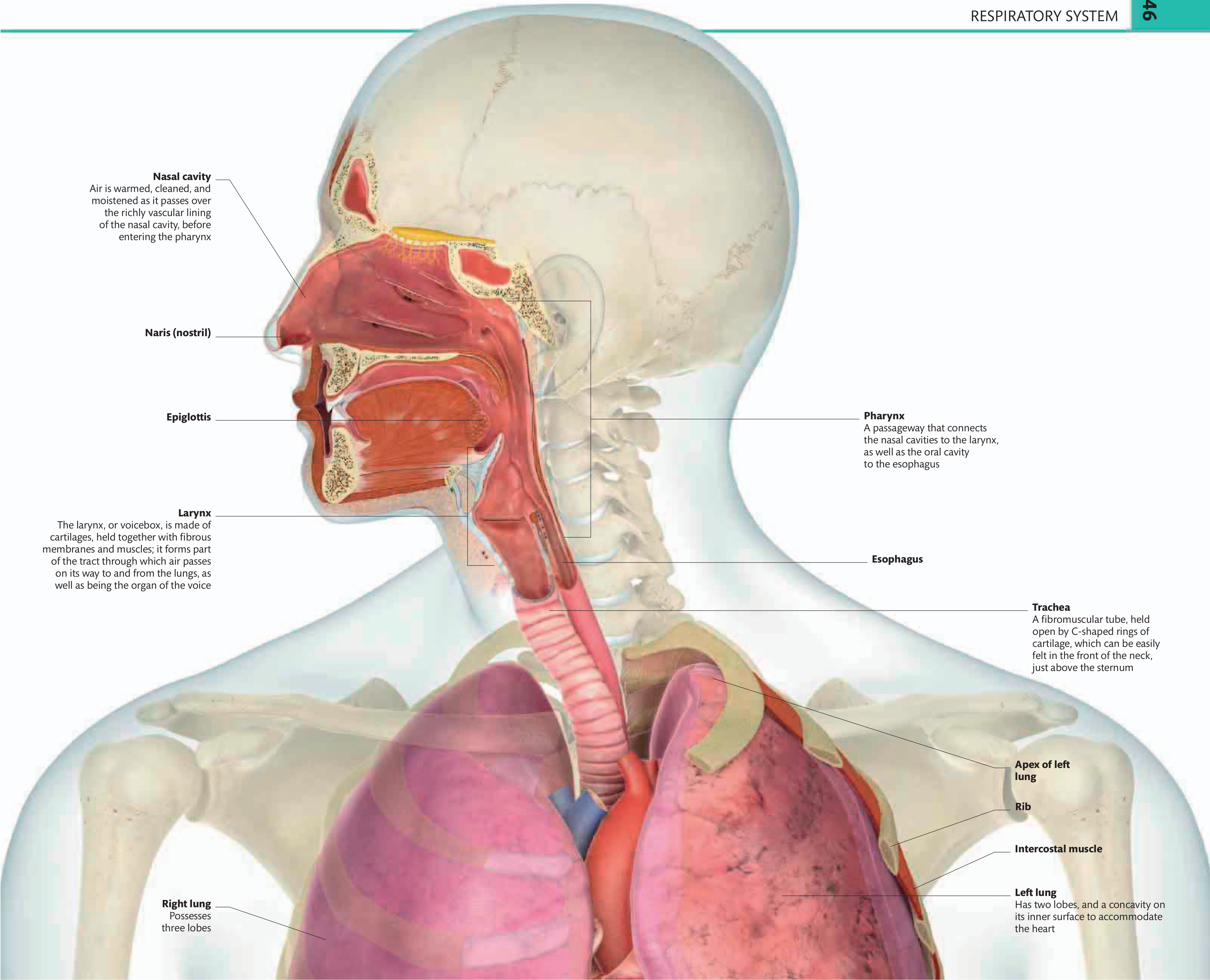

Nasal cavity
Air is warmed, cleaned, and moistened as it passes over the richly vascular lining of the nasal cavity, before entering the pharynx

Naris (nostril)

Epiglottis

Larynx
The larynx, or voicebox, is made of cartilages, held together with fibrous membranes and muscles; it forms part of the tract through which air passes on its way to and from the lungs, as well as being the organ of the voice

Right lung
Possesses three lobes

Pharynx
A passageway that connects the nasal cavities to the larynx, as well as the oral cavity to the esophagus

Esophagus

Trachea
A fibromuscular tube, held open by C-shaped rings of cartilage, which can be easily felt in the front of the neck, just above the sternum

Apex of left lung

Rib

Intercostal muscle

Left lung
Has two lobes, and a concavity on its inner surface to accommodate the heart

RESPIRATORY SYSTEM OVERVIEW

Every cell in the human body needs to get oxygen, and to get rid of carbon dioxide. These gases are transported around the body in the blood, but the actual transfer of gases between the air and the blood occurs in the lungs. The lungs have extremely thin membranes that allow the gases to pass across easily. But air also needs to be regularly drawn in and out of the lungs, to expel the building carbon dioxide and to bring in fresh oxygen, and this is brought about by respiration—commonly called breathing. The respiratory system includes the airways on the way to the lungs: the nasal cavities, parts of the pharynx, the larynx, the trachea, and the bronchi (see p.149).

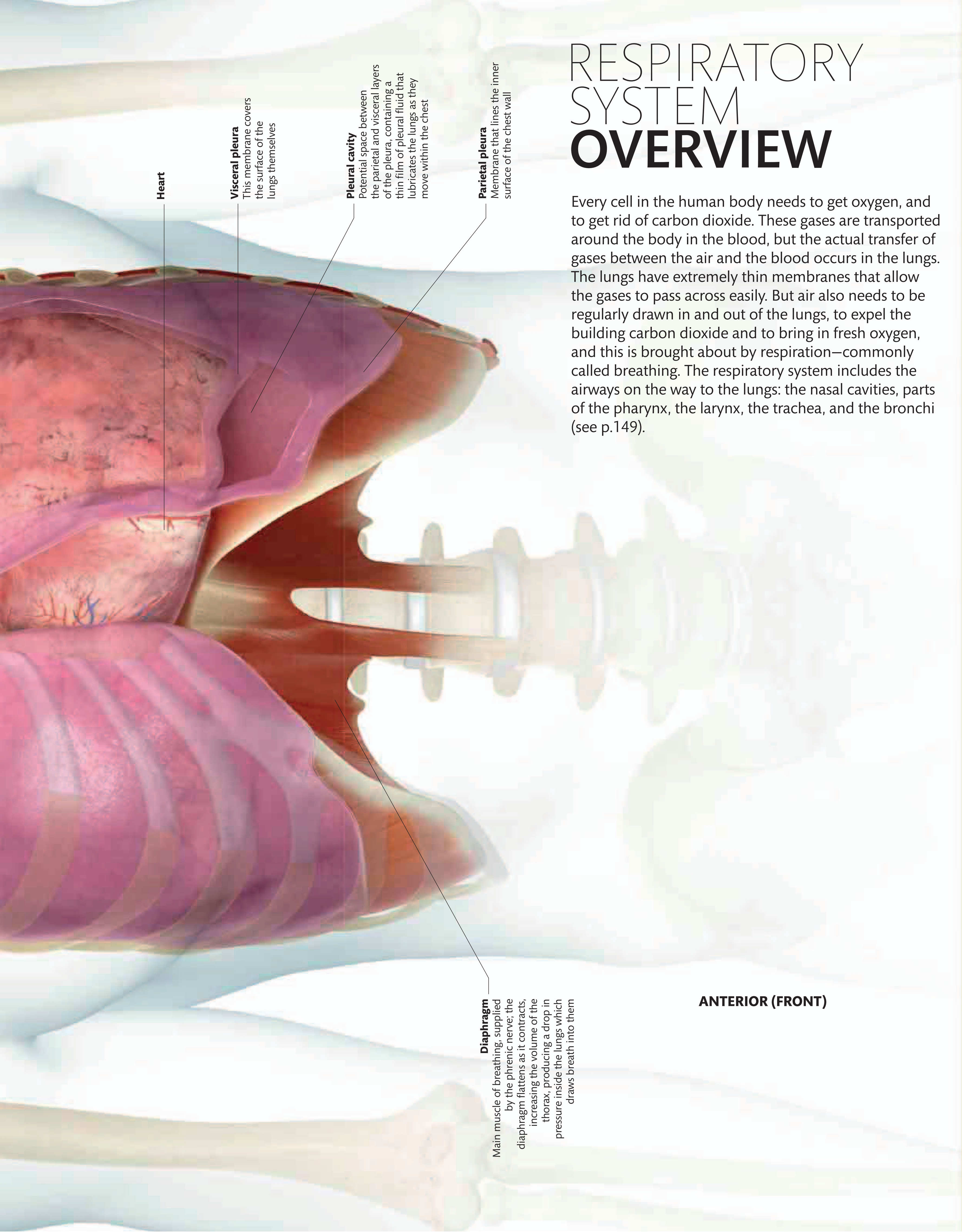
Heart
Visceral pleura This membrane covers the surface of the lungs themselves
Pleural cavity Potential space between the parietal and visceral layers of the pleura, containing a thin film of pleural fluid that lubricates the lungs as they move within the chest
Parietal pleura Membrane that lines the inner surface of the chest wall
Diaphragm Main muscle of breathing, supplied by the phrenic nerve; the diaphragm flattens as it contracts, increasing the volume of the thorax, producing a drop in pressure inside the lungs which draws breath into them

ANTERIOR (FRONT)

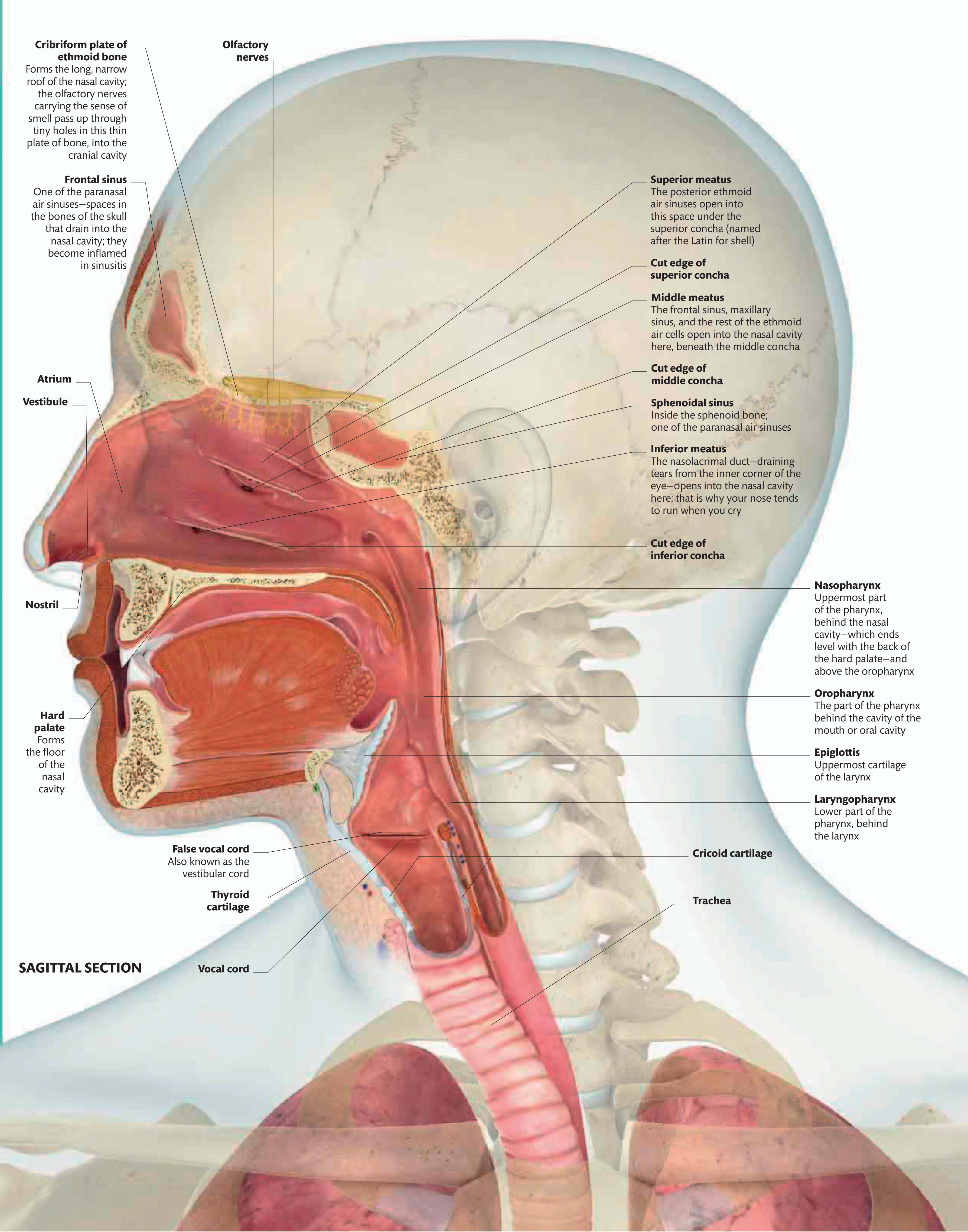

Cribriform plate of ethmoid bone
Forms the long, narrow roof of the nasal cavity; the olfactory nerves carrying the sense of smell pass up through tiny holes in this thin plate of bone, into the cranial cavity

Olfactory nerves

Frontal sinus
One of the paranasal air sinuses—spaces in the bones of the skull that drain into the nasal cavity; they become inflamed in sinusitis

Atrium

Vestibule

Nostril

Hard palate
Forms the floor of the nasal cavity

False vocal cord
Also known as the vestibular cord

Thyroid cartilage

Vocal cord

SAGITTAL SECTION

Superior meatus
The posterior ethmoid air sinuses open into this space under the superior concha (named after the Latin for shell)

Cut edge of superior concha

Middle meatus
The frontal sinus, maxillary sinus, and the rest of the ethmoid air cells open into the nasal cavity here, beneath the middle concha

Cut edge of middle concha

Sphenoidal sinus
Inside the sphenoid bone; one of the paranasal air sinuses

Inferior meatus
The nasolacrimal duct—draining tears from the inner corner of the eye—opens into the nasal cavity here; that is why your nose tends to run when you cry

Cut edge of inferior concha

Nasopharynx
Uppermost part of the pharynx, behind the nasal cavity—which ends level with the back of the hard palate—and above the oropharynx

Oropharynx
The part of the pharynx behind the cavity of the mouth or oral cavity

Epiglottis
Uppermost cartilage of the larynx

Laryngopharynx
Lower part of the pharynx, behind the larynx

Cricoid cartilage

Trachea

HEAD AND NECK

When we take a breath, air is pulled in through our nostrils, into the nasal cavities. Here the air is cleaned, warmed, and moistened before its onward journey. The nasal cavities are divided by the thin partition of the nasal septum, which is composed of plates of cartilage and bone. The lateral walls of the nasal cavity are more elaborate, with bony curls (conchae) that increase the surface area over which the air flows. The nasal cavity is lined with mucosa, which produces mucus. This often undervalued substance does an important job of trapping particles and moistening the air. The nasal sinuses, also lined with mucosa, open via tiny orifices into the nasal cavity. Below and in front of the pharynx is the larynx—the organ of speech. The way that air passes through this can be modulated to produce sound.

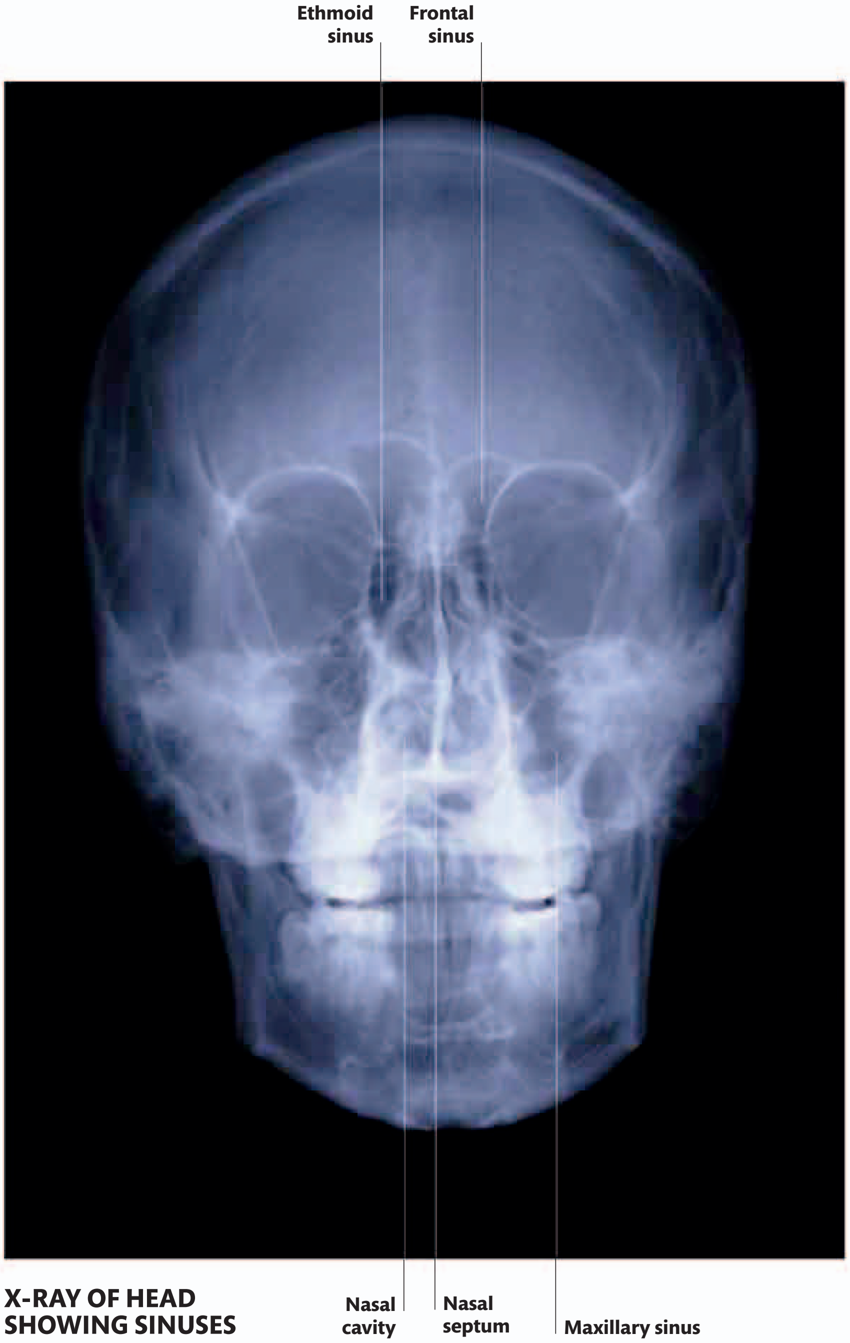
Ethmoid sinus
Frontal sinus
Nasal cavity
Nasal septum
Maxillary sinus

X-RAY OF HEAD SHOWING SINUSES

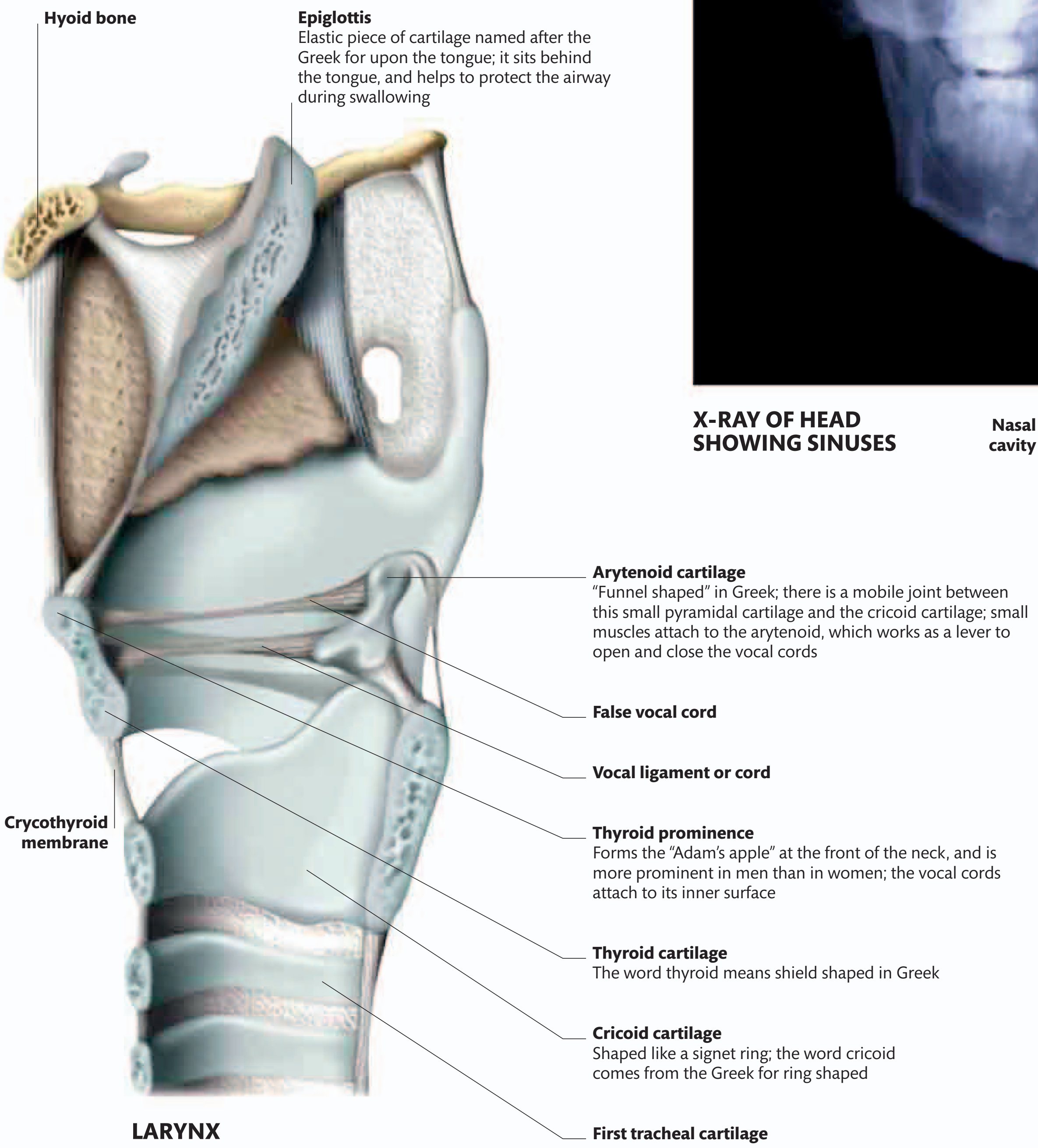
Hyoid bone
Crycothyroid membrane
False vocal cord
Vocal ligament or cord
First tracheal cartilage

Epiglottis
Elastic piece of cartilage named after the Greek for upon the tongue; it sits behind the tongue, and helps to protect the airway during swallowing

Arytenoid cartilage
"Funnel shaped" in Greek; there is a mobile joint between this small pyramidal cartilage and the cricoid cartilage; small muscles attach to the arytenoid, which works as a lever to open and close the vocal cords

Thyroid prominence
Forms the "Adam's apple" at the front of the neck, and is more prominent in men than in women; the vocal cords attach to its inner surface

Thyroid cartilage
The word thyroid means shield shaped in Greek

Cricoid cartilage
Shaped like a signet ring; the word cricoid comes from the Greek for ring shaped

LARYNX

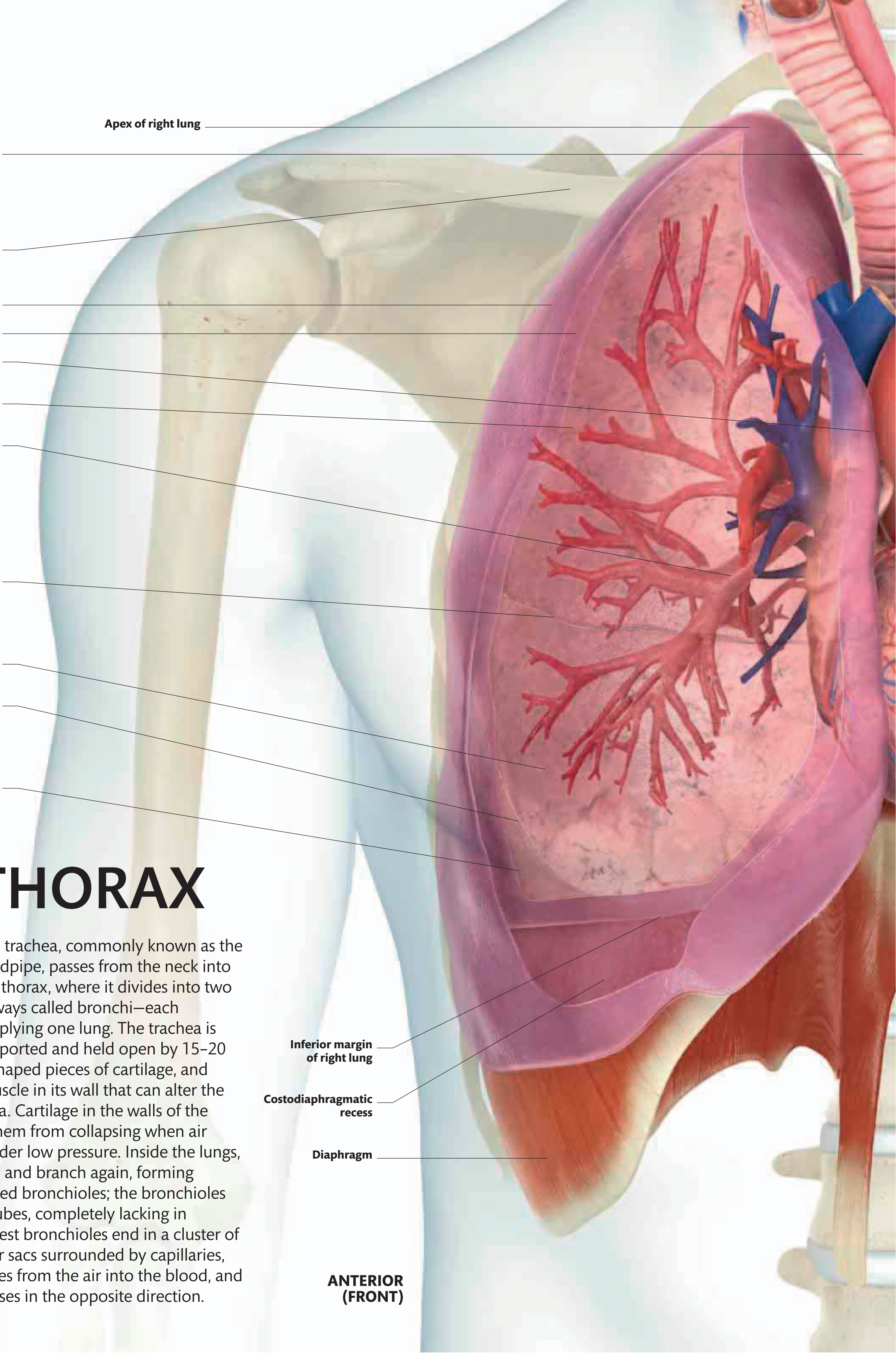

Apex of right lung

Trachea
Named after the Greek for rough vessel, the trachea is about $4^{3}/_{4}$in (12cm) long and $^{1}/_{2}$–$^{3}/_{4}$in (1.5–2cm) wide in an adult

Right clavicle (cut away to show lung behind)

Parietal pleura

Visceral pleura

Anterior margin of right lung

Superior lobe of right lung

Right main bronchus
Several smaller bronchi branch off the two main bronchi that enter the lung by bifurcating from the trachea; confusingly, the word bronchus comes from the Greek for windpipe

Horizontal fissure
Deep cleft that separates the superior (upper) and middle lobes of the right lung

Middle lobe of right lung

Oblique fissure of right lung
Separates the middle and inferior (lower) lobes of the right lung

Inferior lobe of right lung

Inferior margin of right lung

Costodiaphragmatic recess

Diaphragm

ANTERIOR (FRONT)

THORAX

The trachea, commonly known as the windpipe, passes from the neck into the thorax, where it divides into two airways called bronchi—each supplying one lung. The trachea is supported and held open by 15–20 C-shaped pieces of cartilage, and there is smooth muscle in its wall that can alter the width of the trachea. Cartilage in the walls of the bronchi prevents them from collapsing when air enters the lungs under low pressure. Inside the lungs, the bronchi branch and branch again, forming smaller airways called bronchioles; the bronchioles are just muscular tubes, completely lacking in cartilage. The smallest bronchioles end in a cluster of alveoli, these are air sacs surrounded by capillaries, where oxygen passes from the air into the blood, and carbon dioxide passes in the opposite direction.

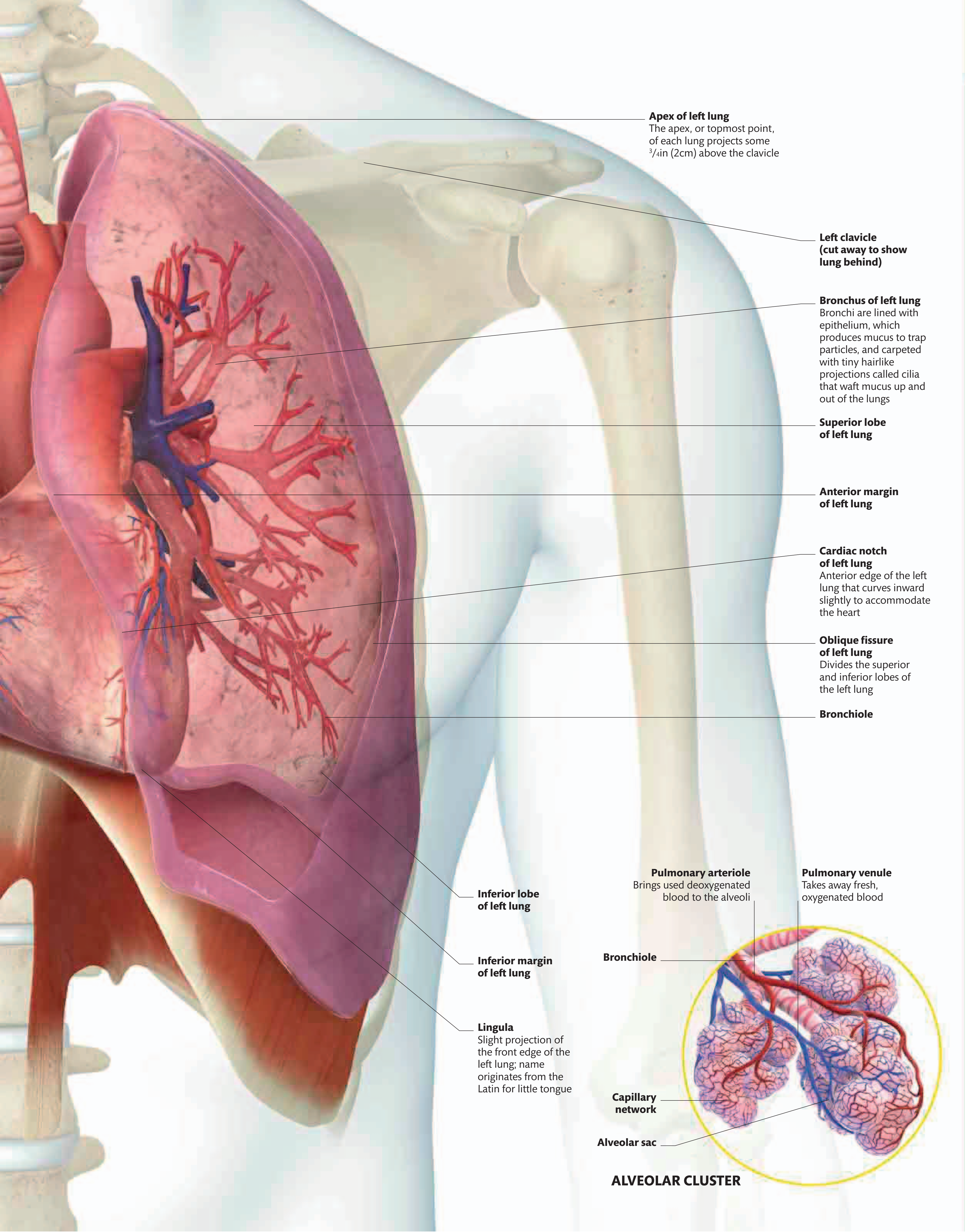

Apex of left lung
The apex, or topmost point, of each lung projects some 3/4in (2cm) above the clavicle

Left clavicle (cut away to show lung behind)

Bronchus of left lung
Bronchi are lined with epithelium, which produces mucus to trap particles, and carpeted with tiny hairlike projections called cilia that waft mucus up and out of the lungs

Superior lobe of left lung

Anterior margin of left lung

Cardiac notch of left lung
Anterior edge of the left lung that curves inward slightly to accommodate the heart

Oblique fissure of left lung
Divides the superior and inferior lobes of the left lung

Bronchiole

Inferior lobe of left lung

Inferior margin of left lung

Lingula
Slight projection of the front edge of the left lung; name originates from the Latin for little tongue

Pulmonary arteriole
Brings used deoxygenated blood to the alveoli

Pulmonary venule
Takes away fresh, oxygenated blood

ALVEOLAR CLUSTER

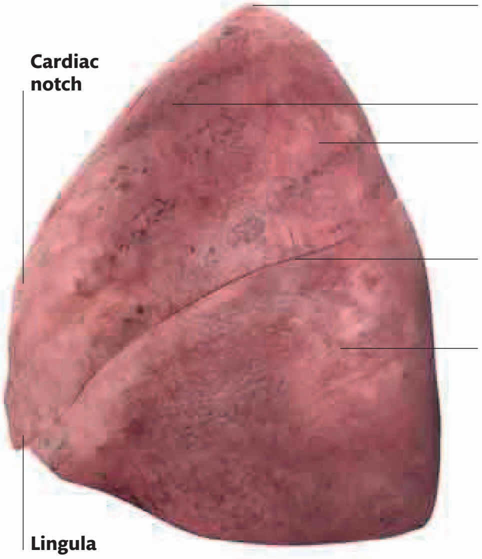
Apex
Cardiac notch
Costal surface
Superior lobe
Oblique fissure
Inferior lobe
Lingula

LEFT LUNG (LATERAL VIEW)

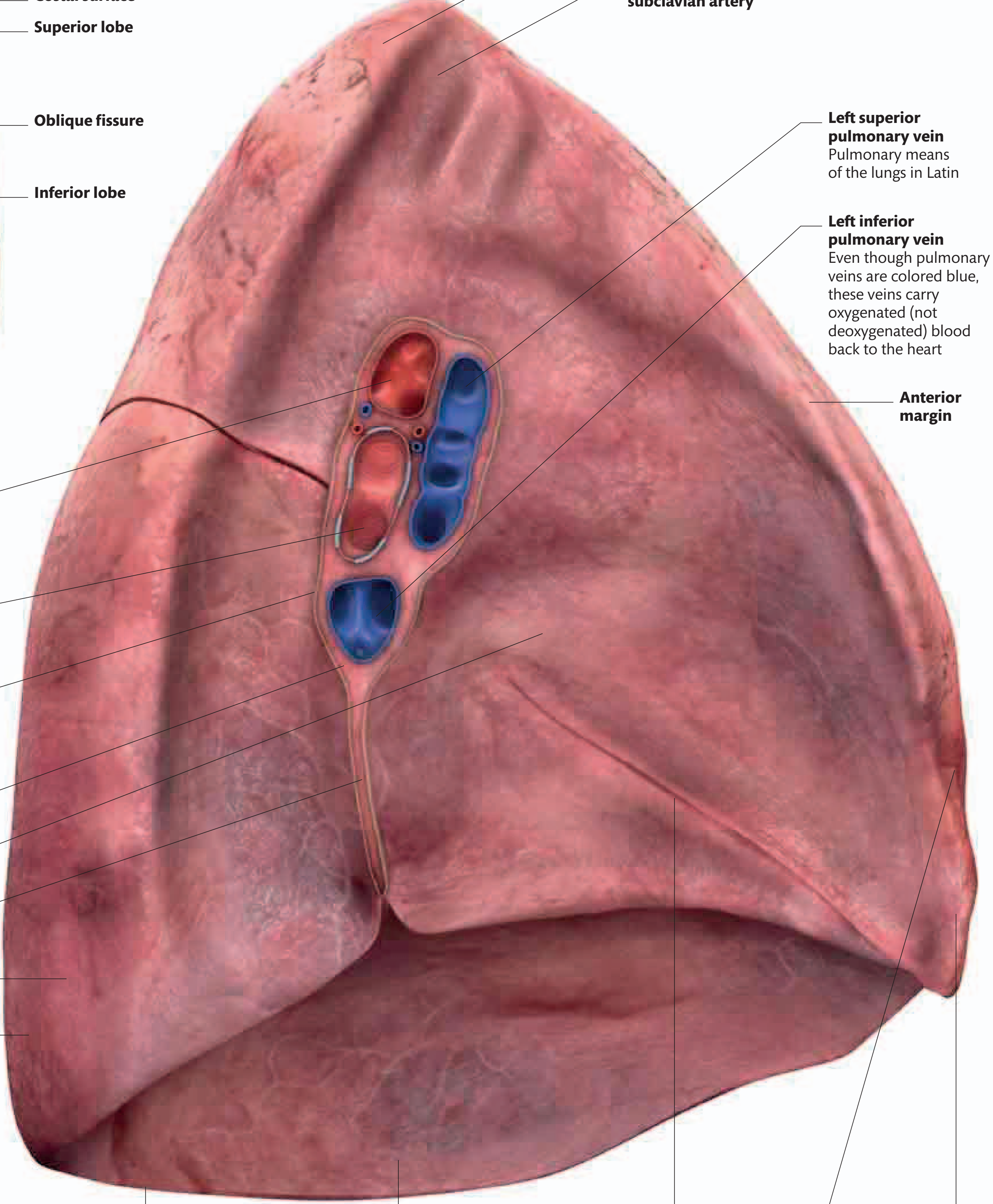
Superior lobe
Groove for left subclavian artery
Anterior margin
Hilum
Cardiac impression
Pulmonary ligament
Costal surface of lung
Inferior lobe
Diaphragmatic surface of lung
Oblique fissure
Cardiac notch
Lingula

Left superior pulmonary vein
Pulmonary means of the lungs in Latin

Left inferior pulmonary vein
Even though pulmonary veins are colored blue, these veins carry oxygenated (not deoxygenated) blood back to the heart

Left pulmonary artery
Brings deoxygenated blood to the lungs to be oxygenated; the pulmonary arteries are the only arteries to carry deoxygenated blood

Left main bronchus
Just before it divides into the superior and the inferior lobar bronchi

Pleura
The membrane lining the lungs; pleura comes from the Greek for rib or side of the body

Inferior margin
This sharp lower edge of the lung fits down into the cleft between the edge of the dome of the diaphragm and the chest wall; the bottom of the pleural cavity extends a couple more inches below the edge of the lung

LEFT LUNG (MEDIAL VIEW)

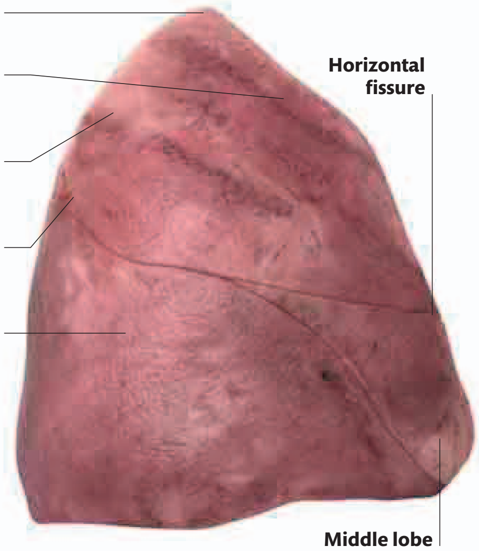
Apex
Costal surface
Superior lobe
Oblique fissure
Inferior lobe
Horizontal fissure
Middle lobe

RIGHT LUNG (LATERAL VIEW)

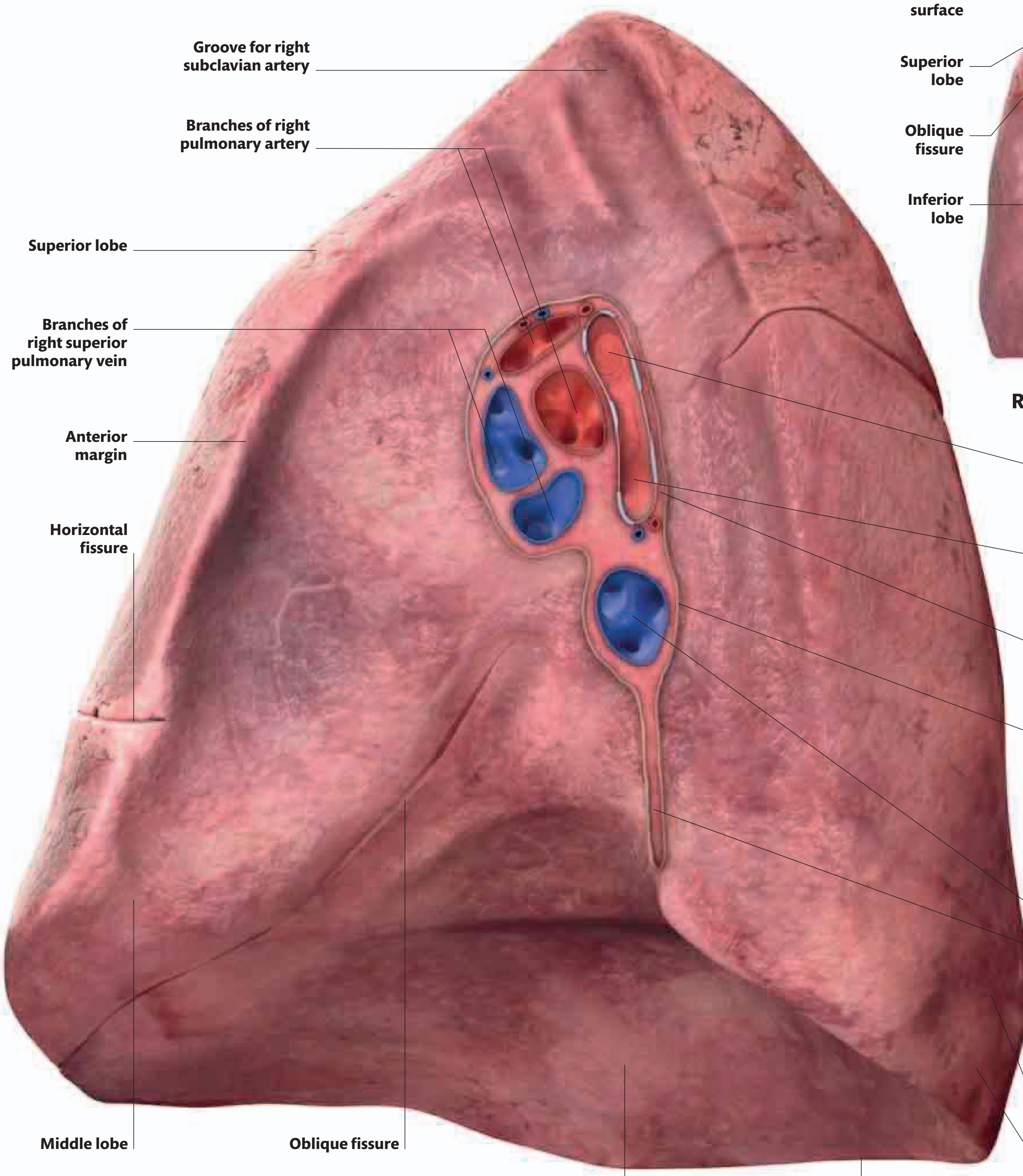
Groove for right subclavian artery
Branches of right pulmonary artery
Superior lobe
Branches of right superior pulmonary vein
Anterior margin
Horizontal fissure
Middle lobe
Oblique fissure
Diaphragmatic surface
Inferior margin
Right inferior pulmonary vein
Costal surface
Inferior lobe

Superior lobar bronchus
The structure of the wall of the bronchus, containing rings of cartilage, can be seen in cross section here

Right main bronchus
Having already given off the superior lobar bronchus, this will branch to form the middle and inferior lobar bronchi

Hilum
Depression where the bronchi and major pulmonary vessels enter and leave the lung

Pleura
Around the hilum, the visceral pleura doubles back on itself to form the parietal pleura, which continues on to the inner surface of the ribcage and the upper surface of the diaphragm; the visceral and parietal pleura thus form a continuous, closed envelope around the lung

Pulmonary ligament
The line where the visceral pleura doubles back on itself to become the parietal pleura hangs down below the hilum, forming this "ligament", which allows the pulmonary veins to expand freely

RIGHT LUNG (MEDIAL VIEW)

LUNGS

Each lung fits snugly inside its half of the thoracic cavity. The surface of each lung is covered with a thin pleural membrane (visceral pleura), and the inside of the chest wall is also lined with pleura (parietal pleura). Between the two pleural layers lies a thin film of lubricating fluid that allows the lungs to slide against the chest wall during breathing movements, but it also creates a fluid seal, effectively sticking the lungs to the ribs and the diaphragm. Because of this seal, when you inhale, the lungs are pulled outward in all directions, and air rushes into them. The bronchi and blood vessels enter each lung at the hilum on its inner or medial surface. Although the two lungs may appear to be similar at first glance, there is some asymmetry. The left lung is concave to fit around the heart and has only two lobes, whereas the right lung has three lobes, marked out by two deep fissures.

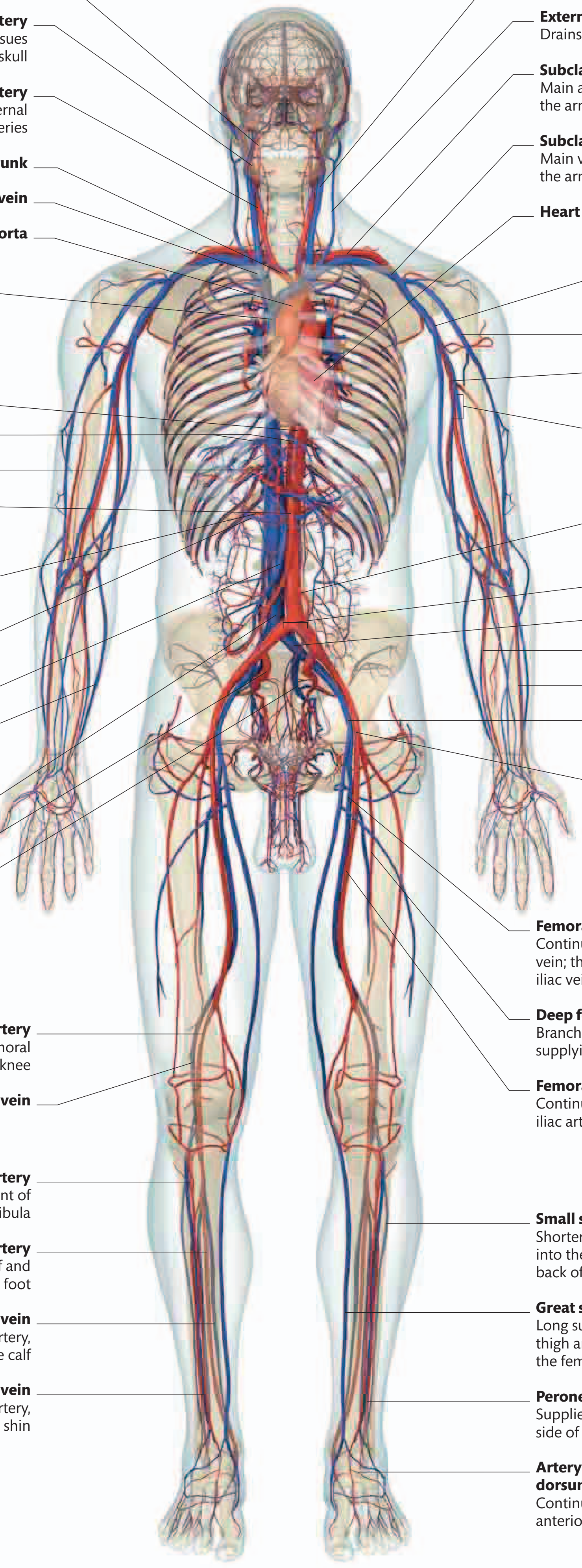

Internal carotid artery
Supplies blood to the brain

External carotid artery
Supplies the neck and tissues of the head outside the skull

Common carotid artery
Divides to form the external and internal carotid arteries

Brachiocephalic trunk

Brachiocephalic vein

Arch of aorta

Superior vena cava
Large vein formed by the joining of the two brachiocephalic veins, returning blood from the head, arms, and chest wall to the heart

Descending aorta

Hepatic veins

Portal vein

Superior mesenteric artery
Branch of the abdominal aorta supplying the small intestine and part of the large intestine

Renal artery
Carries blood to the kidneys

Renal vein
Drains the kidney

Superior mesenteric vein

Basilic vein
Superficial vein draining the medial side of the arm, forearm, and hand

Inferior vena cava

Internal iliac artery

Internal iliac vein

Popliteal artery
Continuation of the femoral artery, at the back of the knee

Popliteal vein

Anterior tibial artery
Supplies muscles in front of the tibia and fibula

Posterior tibial artery
Supplies the calf and sole of the foot

Posterior tibial vein
Runs with the posterior tibial artery, draining deep tissues in the calf

Anterior tibial vein
Runs with the anterior tibial artery, draining deep tissues in the shin

Internal jugular vein
Drains blood from the brain, and from the face and neck

External jugular vein
Drains blood from the face and scalp

Subclavian artery
Main artery supplying blood to the arm, forearm, and hand

Subclavian vein
Main vein draining blood from the arm, forearm, and hand

Heart

Axillary artery
Continuation of the subclavian artery in the axilla, or armpit

Cephalic vein

Brachial artery
Continuation of the axillary artery, in the upper arm

Brachial veins
A pair of veins that run with the brachial artery

Inferior mesenteric artery
Supplies the lower half of the large intestine and the rectum

Common iliac arteries

Common iliac vein

Ulnar artery

Radial artery

External iliac vein
Main vein carrying blood back from the thigh, leg, and foot

External iliac artery
Main artery supplying the thigh, leg, and foot

Femoral vein
Continuation of the popliteal vein; this becomes the external iliac vein at the groin

Deep femoral artery
Branch of the femoral artery supplying the muscles of the thigh

Femoral artery
Continuation of the external iliac artery, in the thigh

Small saphenous vein
Shorter superficial vein drains into the popliteal vein at the back of the knee

Great saphenous vein
Long superficial vein of the thigh and leg, ending in the femoral vein

Peroneal artery
Supplies muscles in the side of the lower leg

Artery of the dorsum of the foot
Continuation of the anterior tibial artery

ANTERIOR (FRONT)

CARDIOVASCULAR SYSTEM **OVERVIEW**

The cardiovascular system consists of the heart, blood, and blood vessels. The heart—a muscular pump—contracts to push blood through the body's network of vessels in order to deliver oxygen, nutrients, white blood cells, and hormones to the tissues of the body. The blood also removes waste products and takes them to other organs—mainly the liver and kidneys—for excretion. The circulatory system can be divided in two: pulmonary circulation carries blood pumped by the right side of the heart to the lungs, and systemic circulation carries blood pumped by the more powerful left side of the heart to the rest of the body. Pressure in the pulmonary circulation is relatively low, to prevent fluid being forced out of capillaries into the alveoli of the lungs. Pressure in the systemic circulation is much higher, to push blood up to the brain, into all other organs, and out into the fingers and toes.

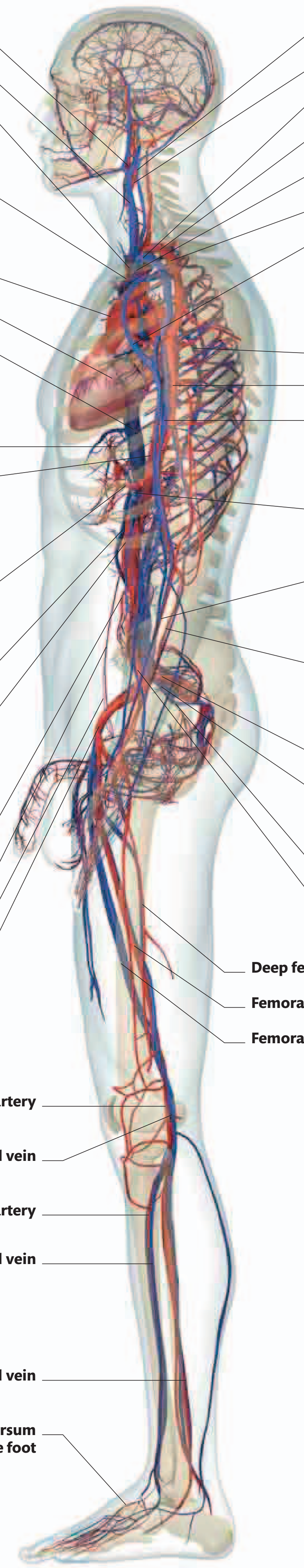

External carotid artery

External jugular vein

Brachiocephalic trunk
Divides to form the right common carotid and subclavian arteries

Brachiocephalic vein
Formed by the union of the internal jugular and subclavian veins

Arch of aorta

Heart

Inferior vena cava
Large vein draining blood from the lower body and returning it to the heart

Hepatic vein

Descending aorta
The arch of the aorta becomes the descending aorta, which runs down through the thorax and into the abdomen

Celiac trunk
The name of this artery comes from the Greek for belly or bowels

Superior mesenteric vein

Superior mesenteric artery
Branches of this run in the mesentery—the membrane surrounding the intestines

Inferior mesenteric artery

Gonadal vein

Gonadal artery

External iliac artery

Popliteal artery

Popliteal vein

Anterior tibial artery

Anterior tibial vein

Posterior tibial vein

Artery of the dorsum of the foot

Internal jugular vein

Internal carotid artery

Subclavian vein

Subclavian artery

Superior vena cava

Axillary artery

Cephalic vein
Superficial vein, lying in the subcutaneous tissue, draining the lateral side of the arm, forearm, and hand

Azygos vein

Brachial artery

Brachial vein
One of two veins that run with the brachial artery in the upper limb

Portal vein
Carries blood to the porta hepatis, or "gateway to the liver"

Radial artery
Takes its name from the outer or lateral forearm bone—the radius

Ulnar artery
Branches off the brachial artery to supply the inner forearm and the hand

Internal iliac artery

Common iliac vein
A pair of veins that unite to form the inferior vena cava

Internal iliac vein

Common iliac artery
A pair of arteries formed by the division of the aorta

Deep femoral artery

Femoral artery

Femoral vein

SIDE

BLOOD VESSELS

The heart contracts to keep blood moving through a vast network of blood vessels—arteries, arterioles, capillaries, venules, and veins. A thick elastic wall in arteries helps them to carry high-pressure blood from the heart to organs and tissues; veins contain valves that prevent backflow of blood when carrying it back to the heart. Arteries and smaller vessels branch into capillaries—the smallest blood vessels. The endothelial wall of a capillary is one cell thick.

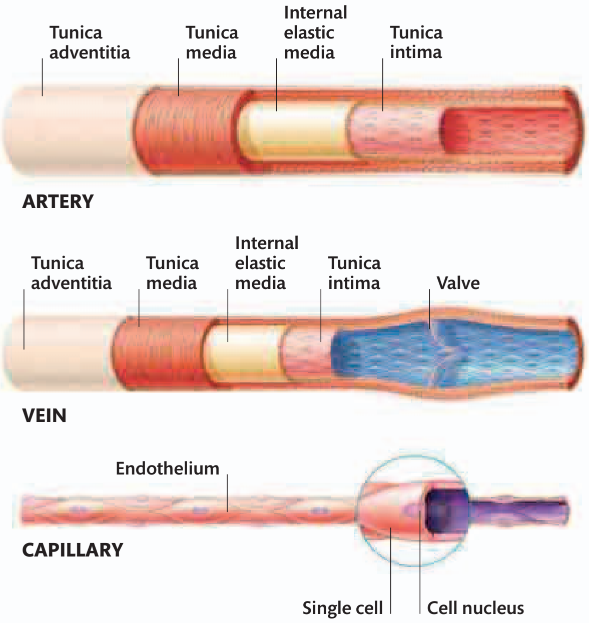

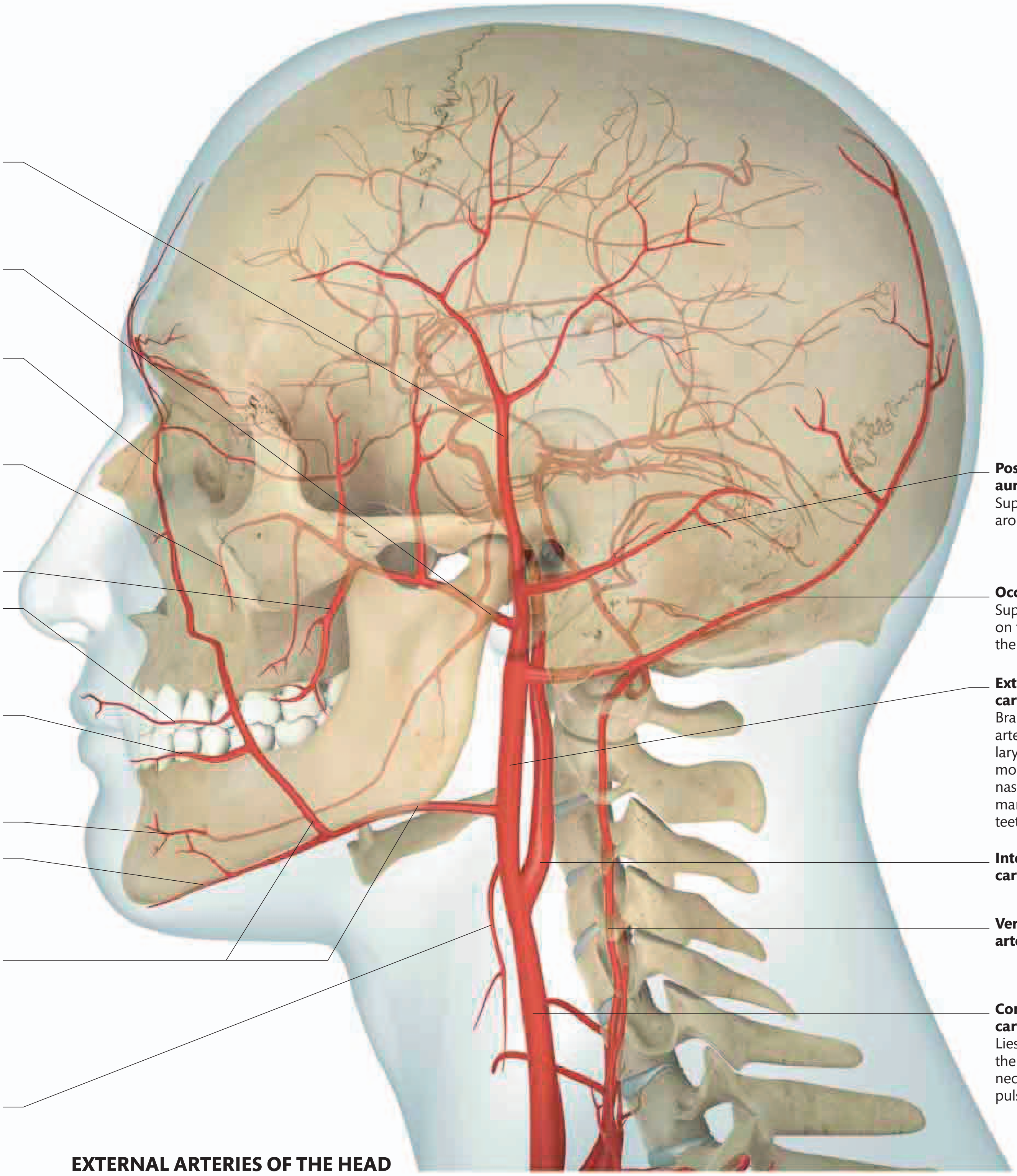

Superficial temporal artery Supplies the scalp over the side of the head

Maxillary artery Supplies the maxilla, mandible, palate, nose, and teeth

Angular artery The continuation of the facial artery, lying near the inner angle of the eye

Infraorbital artery Emerges through the infraorbital foramen just under the orbit

Buccal artery

Superior labial artery Branches from the facial artery to supply the upper lip

Inferior labial artery Branches from the facial artery to supply the lower lip

Mental artery

Submental artery A branch of the facial artery that runs along the underside of the chin

Facial artery Wraps under the lower border of the mandible—where its pulse can be felt—and runs up to supply the face

Superior thyroid artery Supplies the thyroid gland and muscles in the front of the neck

Posterior auricular artery Supplies an area around the ear

Occipital artery Supplies the scalp on the back of the head

External carotid artery Branches of this artery supply the larynx, thyroid gland, mouth, tongue, nasal cavity, face, mandible, maxillae, teeth, and scalp

Internal carotid artery

Vertebral artery

Common carotid artery Lies to the side of the trachea in the neck, where its pulse may be felt

EXTERNAL ARTERIES OF THE HEAD

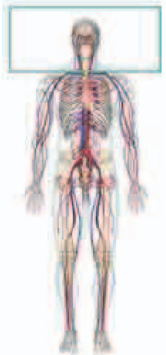

HEAD AND NECK

The main vessels supplying oxygenated blood to the head and neck are the common carotid and vertebral arteries. The vertebral artery runs up through holes in the cervical vertebrae and eventually enters the skull through the foramen magnum. The common carotid artery runs up the neck and divides in two—the internal carotid artery supplies the brain, and the external carotid artery gives rise to a profusion of branches, some of which supply the thyroid gland, the mouth, tongue, and nasal cavity. Veins of the head and neck come together like river tributaries, draining into the large internal jugular vein, behind the sternocleidomastoid muscle, and into the subclavian vein, low in the neck.

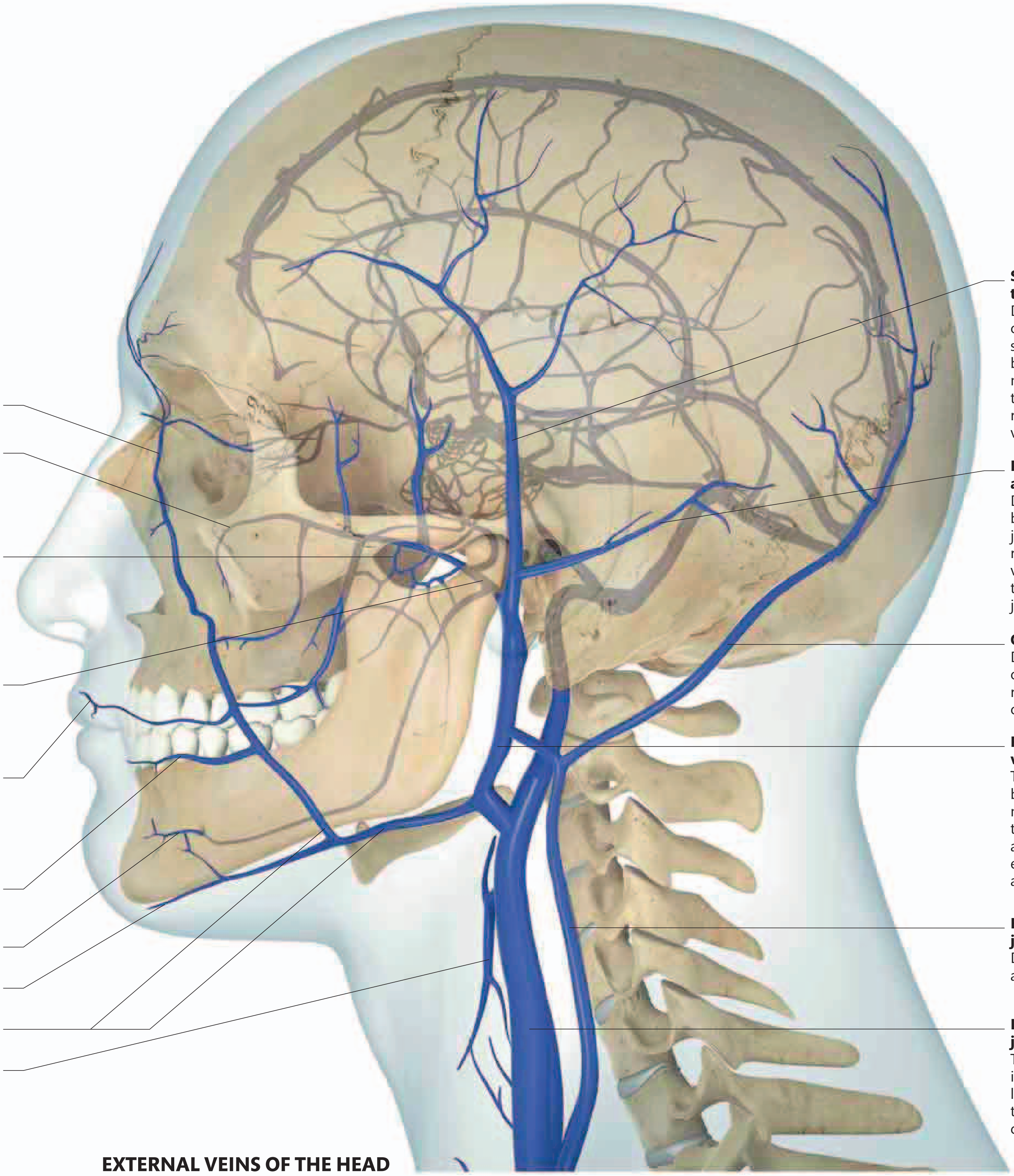

EXTERNAL VEINS OF THE HEAD

Angular vein

Infraorbital vein

Pterygoid venous network
A network of veins lying under the ramus of the mandible

Maxillary vein
Drains the pterygoid venous network

Superior labial vein
Drains from the upper lip into the facial vein

Inferior labial vein
Drains the lower lip

Mental vein

Submental vein

Facial vein

Superior thyroid vein

Superficial temporal vein
Drains a network of veins in the scalp and ends by joining the maxillary vein to form the retromandibular vein

Posterior auricular vein
Drains the scalp behind the ear; joins the retromandibular vein to form the external jugular vein

Occipital vein
Drains the back of the scalp and runs deep to join other veins

Retromandibular vein
Travels down behind the mandible, through the parotid gland alongside the external carotid artery

External jugular vein
Drains the face and scalp

Internal jugular vein
The largest vein in the neck; lies close to the common carotid artery

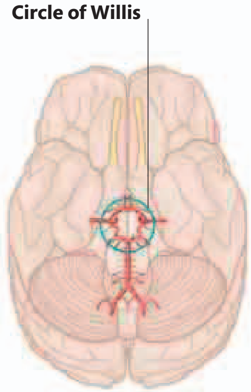

LOCATION OF CIRCLE OF WILLIS

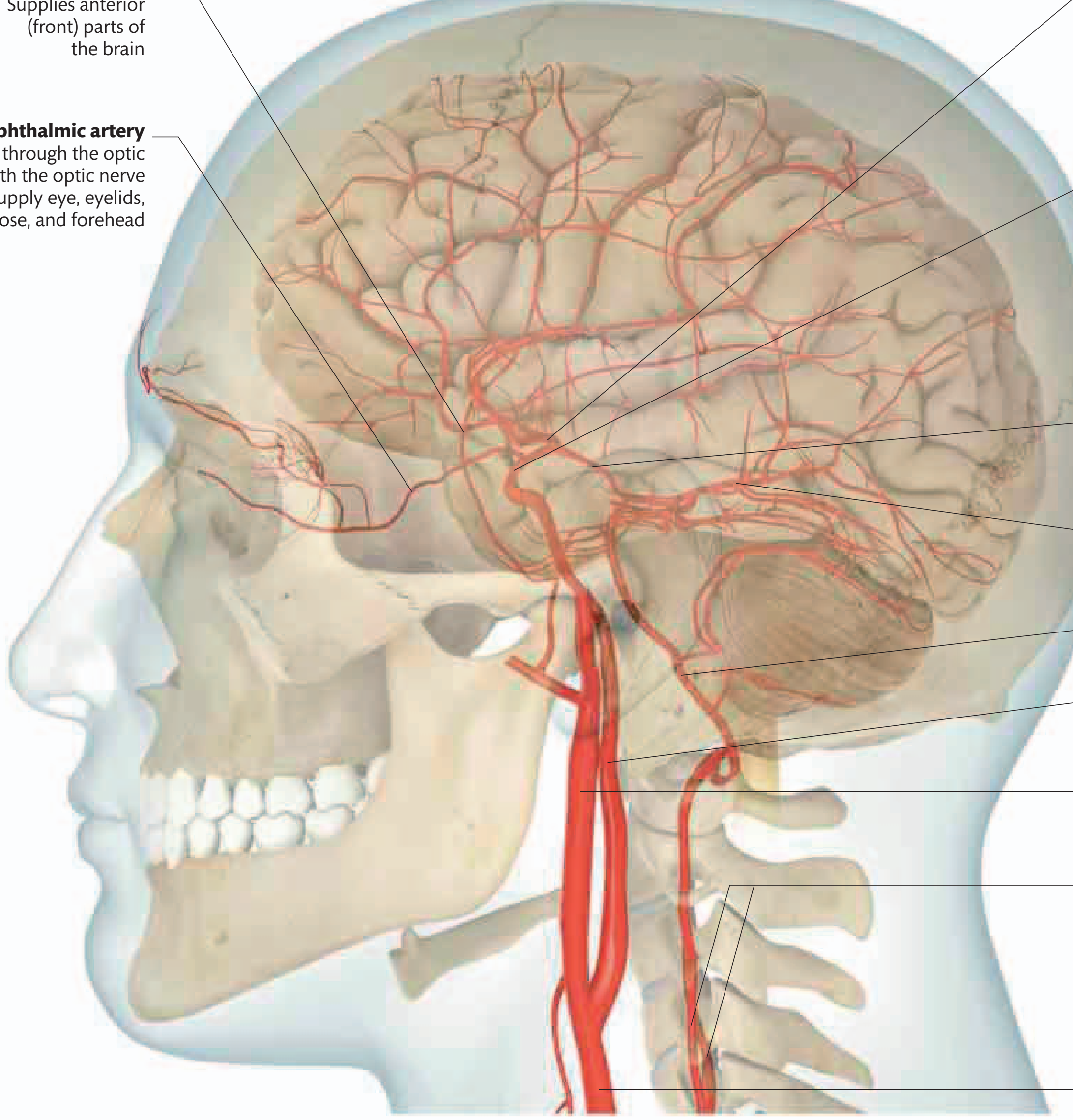

Anterior cerebral artery Supplies anterior (front) parts of the brain

Ophthalmic artery Runs through the optic canal with the optic nerve to supply eye, eyelids, nose, and forehead

Middle cerebral artery Provides branches to the cortex of the frontal, parietal, and temporal lobes of the brain

Cavernous part of the internal carotid artery Passes into the carotid canal and emerges inside the skull, travelling through the cavernous sinus

Posterior communicating artery

Posterior cerebral artery

Basilar artery

Internal carotid artery

External carotid artery

Vertebral arteries Travel up through the foramina in the cervical vertebrae, and enter the skull through the foramen magnum

Common carotid artery

ARTERIES AROUND THE BRAIN

HEAD AND NECK

The brain has a rich blood supply, which arrives via the internal carotid and vertebral arteries. The vertebral arteries join together to form the basilar artery. The internal carotid arteries and basilar artery join up on the undersurface of the brain to form the Circle of Willis. From there, three pairs of cerebral arteries make their way into the brain. The veins of the brain and the skull drain into venous sinuses, which are enclosed within the dura mater (the outermost layer of the meninges) and form grooves on the inner surface of the skull. The sinuses join up and eventually drain out of the base of the skull, into the internal jugular vein.

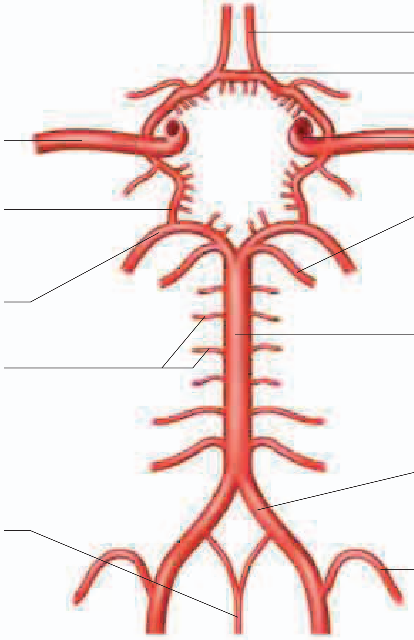

Middle cerebral artery

Posterior communicating artery

Posterior cerebral artery

Pontine arteries Branch from the basilar artery to supply the pons

Anterior spinal artery Supplies the medulla and spinal cord

Anterior cerebral artery

Anterior communicating artery

Internal carotid artery

Superior cerebellar artery Highest of three pairs of arteries supplying the cerebellum of the brain

Basilar artery Carries blood from vertebral arteries to the Circle of Willis and supplies the midbrain

Vertebral artery Joins the other vertebral artery to form the basilar artery

Posterior inferior cerebellar artery Supplies the cerebellum and the choroid plexus of the fourth ventricle in the brain

CIRCLE OF WILLIS

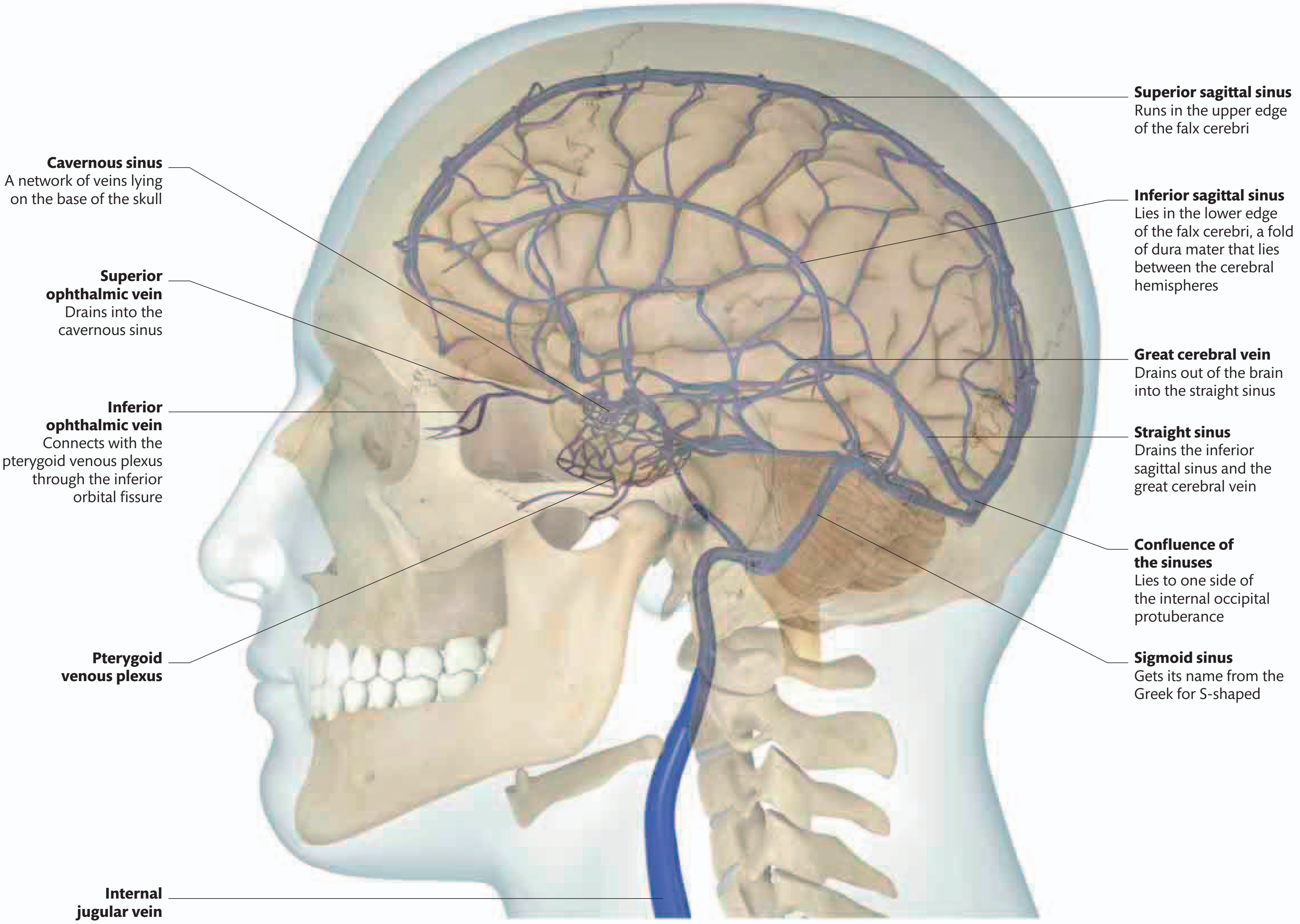
Cavernous sinus
A network of veins lying on the base of the skull
Superior ophthalmic vein
Drains into the cavernous sinus
Inferior ophthalmic vein
Connects with the pterygoid venous plexus through the inferior orbital fissure
Pterygoid venous plexus
Internal jugular vein
Superior sagittal sinus
Runs in the upper edge of the falx cerebri
Inferior sagittal sinus
Lies in the lower edge of the falx cerebri, a fold of dura mater that lies between the cerebral hemispheres
Great cerebral vein
Drains out of the brain into the straight sinus
Straight sinus
Drains the inferior sagittal sinus and the great cerebral vein
Confluence of the sinuses
Lies to one side of the internal occipital protuberance
Sigmoid sinus
Gets its name from the Greek for S-shaped

VEINS AROUND THE BRAIN

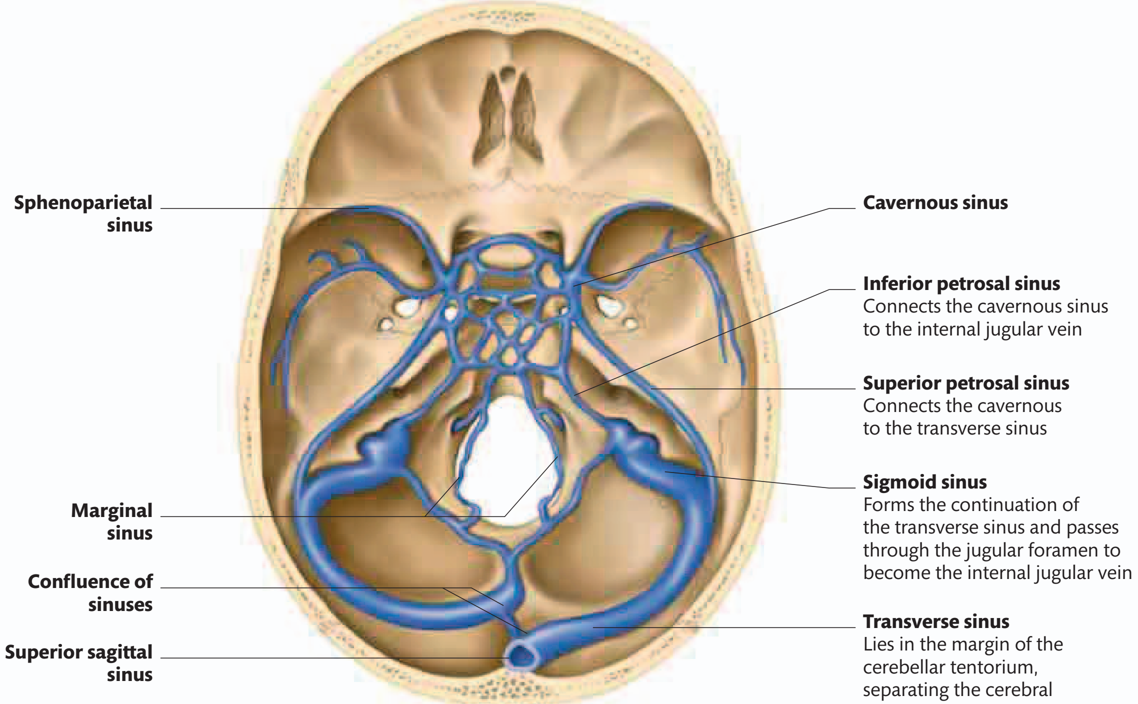
Sphenoparietal sinus
Marginal sinus
Confluence of sinuses
Superior sagittal sinus
Cavernous sinus
Inferior petrosal sinus
Connects the cavernous sinus to the internal jugular vein
Superior petrosal sinus
Connects the cavernous to the transverse sinus
Sigmoid sinus
Forms the continuation of the transverse sinus and passes through the jugular foramen to become the internal jugular vein
Transverse sinus
Lies in the margin of the cerebellar tentorium, separating the cerebral hemispheres from the cerebellum

DURAL VENOUS SINUSES

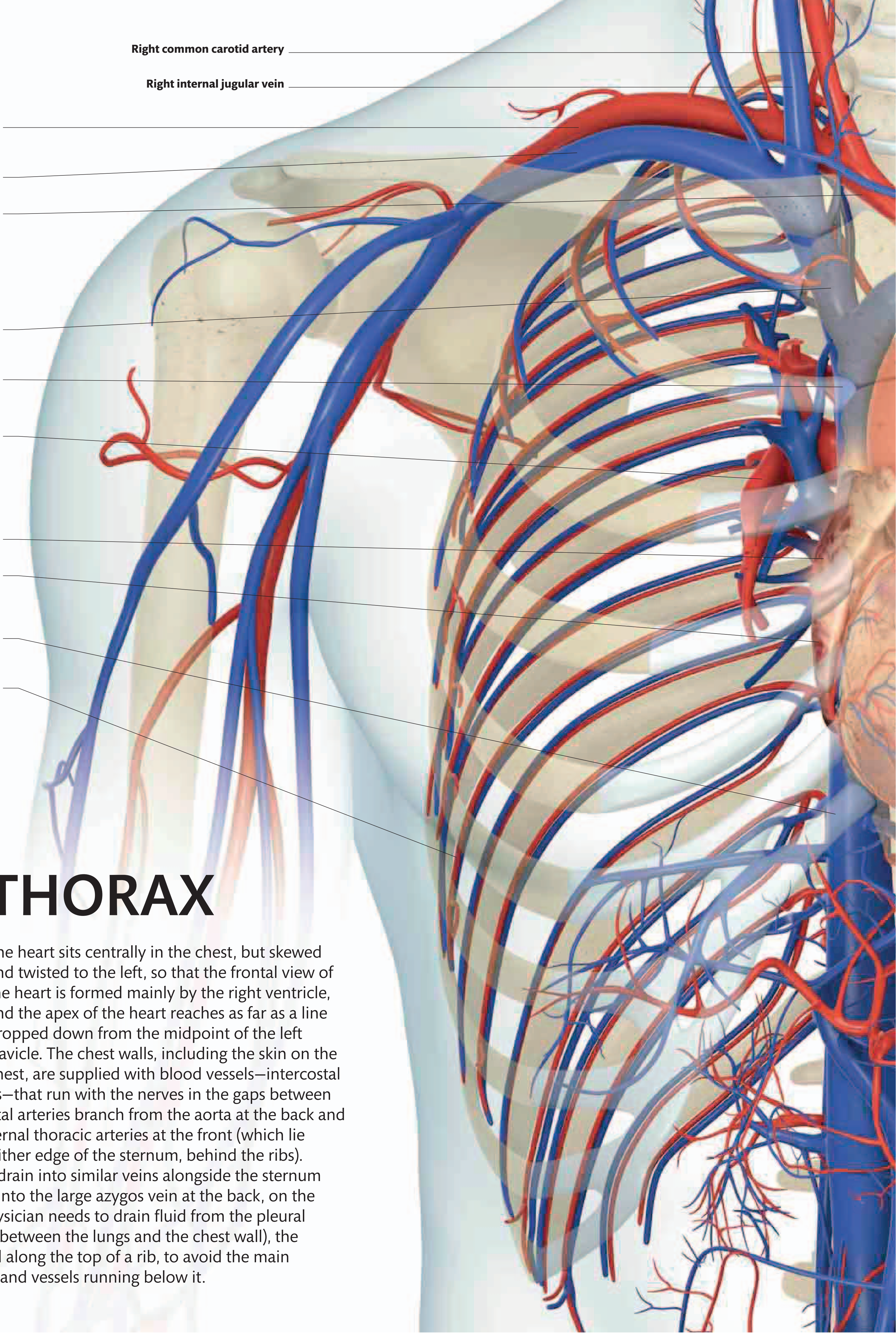

Right common carotid artery

Right internal jugular vein

Right subclavian artery

Right subclavian vein

Brachiocephalic trunk
On the right side of the body, the brachiocephalic trunk divides to form the common carotid and subclavian arteries

Right brachiocephalic vein

Superior vena cava

Right pulmonary artery
The pulmonary arteries carry deoxygenated blood from the heart to the lungs

Right auricle

Right atrium
Forms the border of the heart on the right side

Inferior vena cava

Intercostal blood vessels
An artery and vein run below the length of each rib; for clarity this illustration shows them only around the back of the ribcage

THORAX

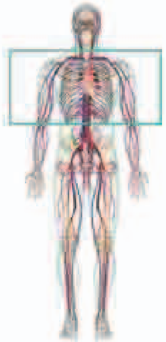

The heart sits centrally in the chest, but skewed and twisted to the left, so that the frontal view of the heart is formed mainly by the right ventricle, and the apex of the heart reaches as far as a line dropped down from the midpoint of the left clavicle. The chest walls, including the skin on the chest, are supplied with blood vessels—intercostal arteries and veins—that run with the nerves in the gaps between the ribs. Intercostal arteries branch from the aorta at the back and from the two internal thoracic arteries at the front (which lie vertically along either edge of the sternum, behind the ribs). Intercostal veins drain into similar veins alongside the sternum at the front, and into the large azygos vein at the back, on the right side. If a physician needs to drain fluid from the pleural cavity (the space between the lungs and the chest wall), the needle is inserted along the top of a rib, to avoid the main intercostal nerve and vessels running below it.

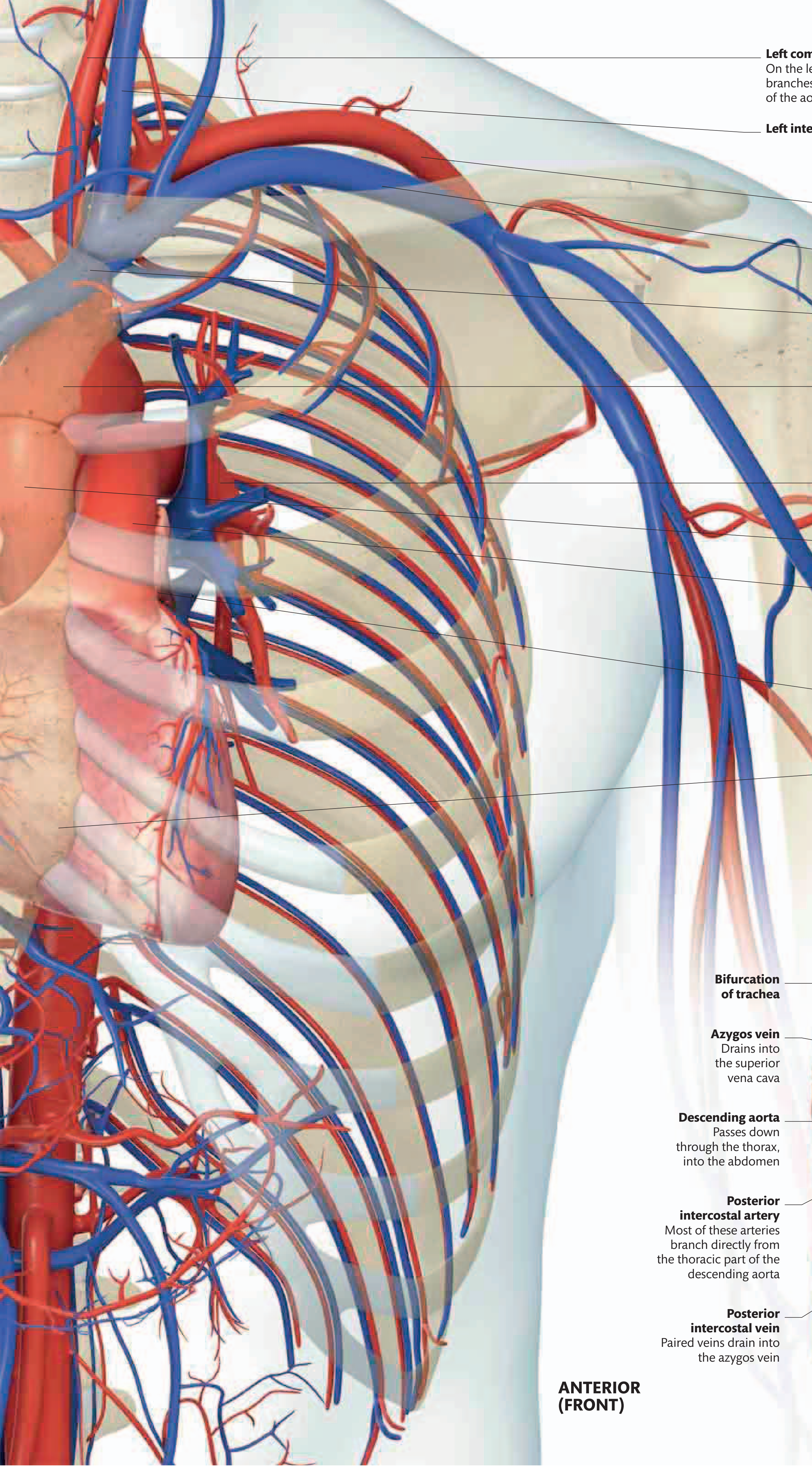
Left common carotid artery
On the left side of the body this branches directly from the arch of the aorta (compare with right)
Left internal jugular vein
Left subclavian artery
Left subclavian vein
Left brachiocephalic vein
Arch of aorta
The main artery of the body exits the heart and arches over it
Left pulmonary artery
Ascending aorta
Pulmonary trunk
Branches into the right and left pulmonary arteries, under the arch of the aorta
Left auricle
Right ventricle

ANTERIOR (FRONT)

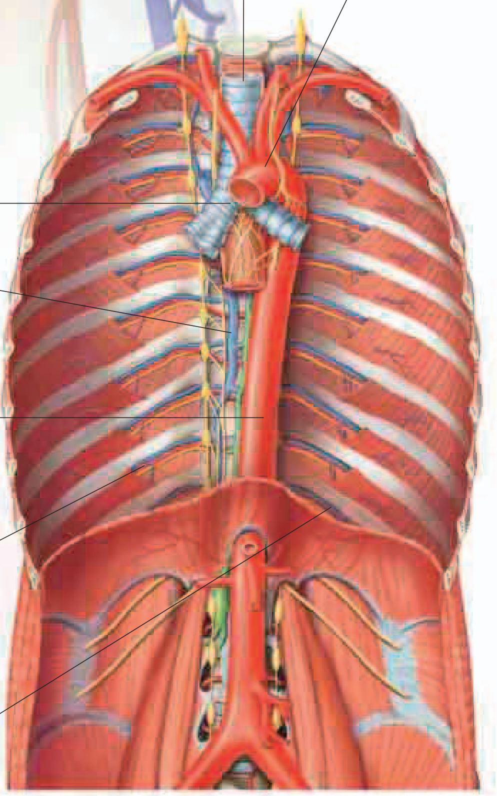
Trachea
Arch of aorta
Bifurcation of trachea
Azygos vein
Drains into the superior vena cava
Descending aorta
Passes down through the thorax, into the abdomen
Posterior intercostal artery
Most of these arteries branch directly from the thoracic part of the descending aorta
Posterior intercostal vein
Paired veins drain into the azygos vein

BACK OF THORACIC CAVITY (HEART REMOVED)

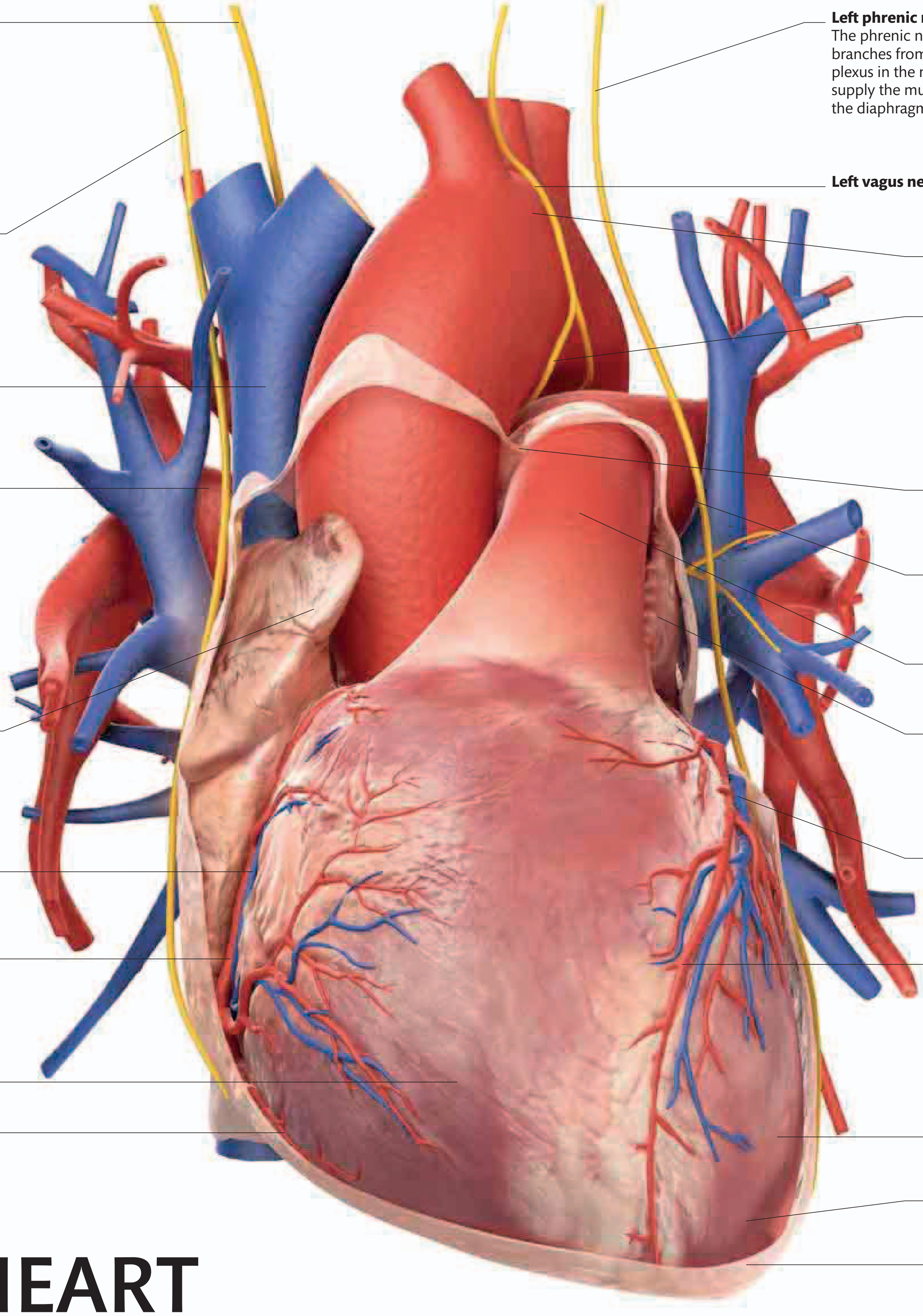

Right vagus nerve The vagus nerves travel through the thorax, giving branches to the heart and lungs, then travel close to the esophagus, behind the heart, as they descend into the abdomen

Right phrenic nerve

Superior vena cava

Right pulmonary artery

Right auricle A pocketlike projection from the right atrium; its name comes from the Latin for little ear

Small cardiac vein Drains into the coronary sinus

Right coronary artery The word coronary comes from the Latin for crown; the right and left coronary arteries encircle the heart

Right ventricle

Marginal artery A branch of the right coronary artery

Left phrenic nerve The phrenic nerves are branches from the cervical plexus in the neck; they supply the muscle of the diaphragm

Left vagus nerve

Arch of aorta

Left recurrent laryngeal nerve This branch of the left vagus nerve loops under the arch of the aorta before travelling back up to the neck to supply the larynx

Cut edge of pericardium

Left pulmonary artery

Pulmonary trunk

Left auricle Similar to the right auricle, this is a projection from the left atrium

Great cardiac vein Drains into the coronary sinus

Anterior interventricular artery A branch of the left coronary artery that runs down between the two ventricles on the front of the heart

Left ventricle

Apex

Pericardium

ANTERIOR (FRONT)

HEART

The heart is encased in the pericardium. This has a tough outer layer that is fused to the diaphragm below and to the connective tissue around the large blood vessels above the heart. Lining the inside of this cylinder (and the outer surface of the heart) is a thin membrane called the serous pericardium. Between these two layers is a thin film of fluid that lubricates the movement of the heart as it beats. Inflammation of this membrane, known as pericarditis, can be extremely painful. Branches of the right and left coronary arteries, which spring from the ascending aorta, supply the heart muscle itself. The heart is drained by cardiac veins, most of which drain into the coronary sinus.

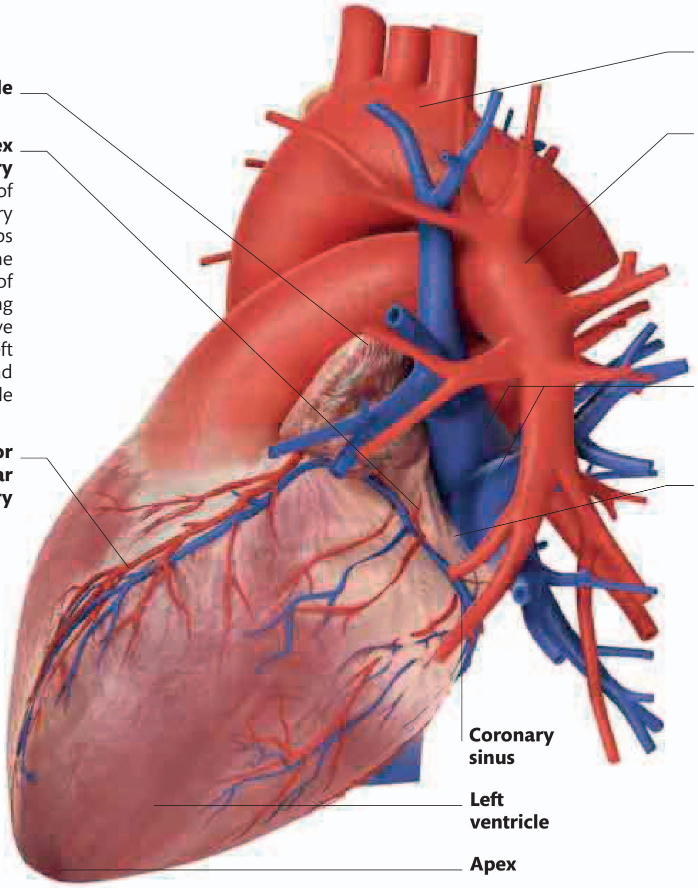
Left auricle
Circumflex artery
Anterior interventricular artery
Arch of aorta
Left pulmonary artery
Left pulmonary veins
Left atrium
Coronary sinus
Left ventricle
Apex

Circumflex artery This branch of the left coronary artery wraps around the left side of the heart, lying in the groove between the left atrium and left ventricle

LEFT

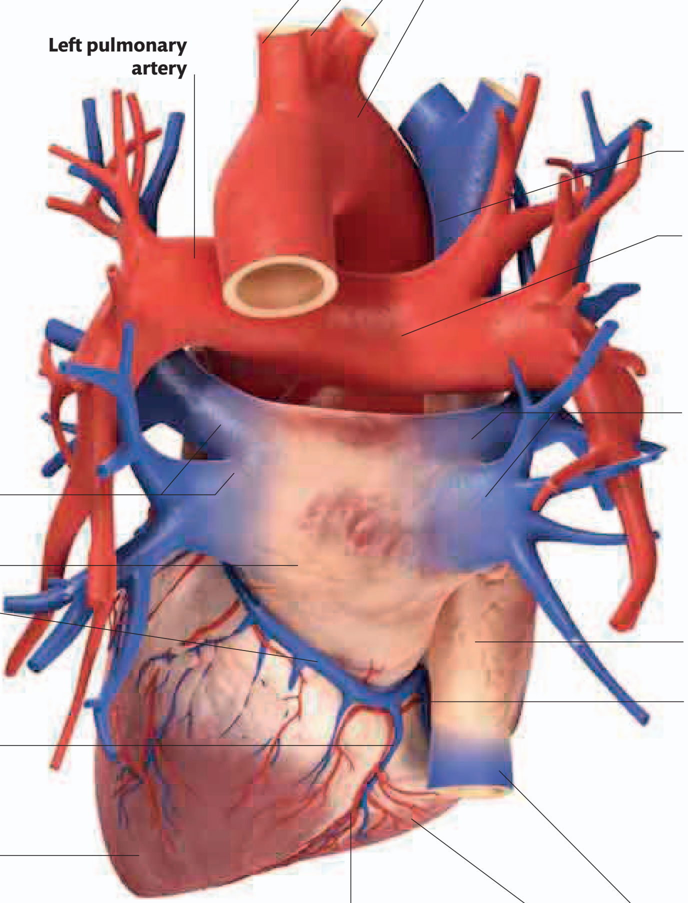
Left subclavian artery
Left common carotid artery
Brachiocephalic trunk
Arch of aorta
Left pulmonary artery
Superior vena cava
Right pulmonary artery
Right pulmonary veins
Left pulmonary veins
Left atrium
Right atrium
Coronary sinus
Right coronary artery
Middle cardiac vein
Left ventricle
Inferior vena cava
Right ventricle
Posterior interventricular artery

Coronary sinus This large vein receives many of the cardiac veins and empties into the right atrium

Middle cardiac vein Drains into the coronary sinus

Right coronary artery Wraps around to the back of the heart, lying in the groove between the right atrium and right ventricle

Posterior interventricular artery This large branch of the right coronary artery runs down between the two ventricles on the underside of the heart

POSTERIOR (BACK)

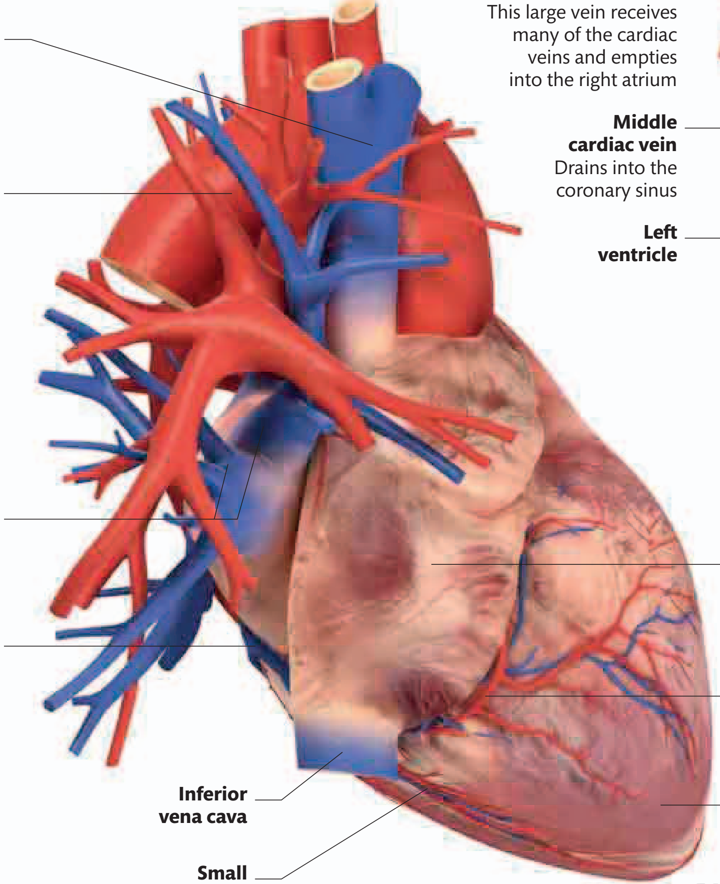
Superior vena cava
Arch of aorta
Right pulmonary veins
Coronary sinus
Inferior vena cava
Small cardiac vein
Right atrium
Right coronary artery
Right ventricle

Coronary sinus Drains into the right atrium

RIGHT

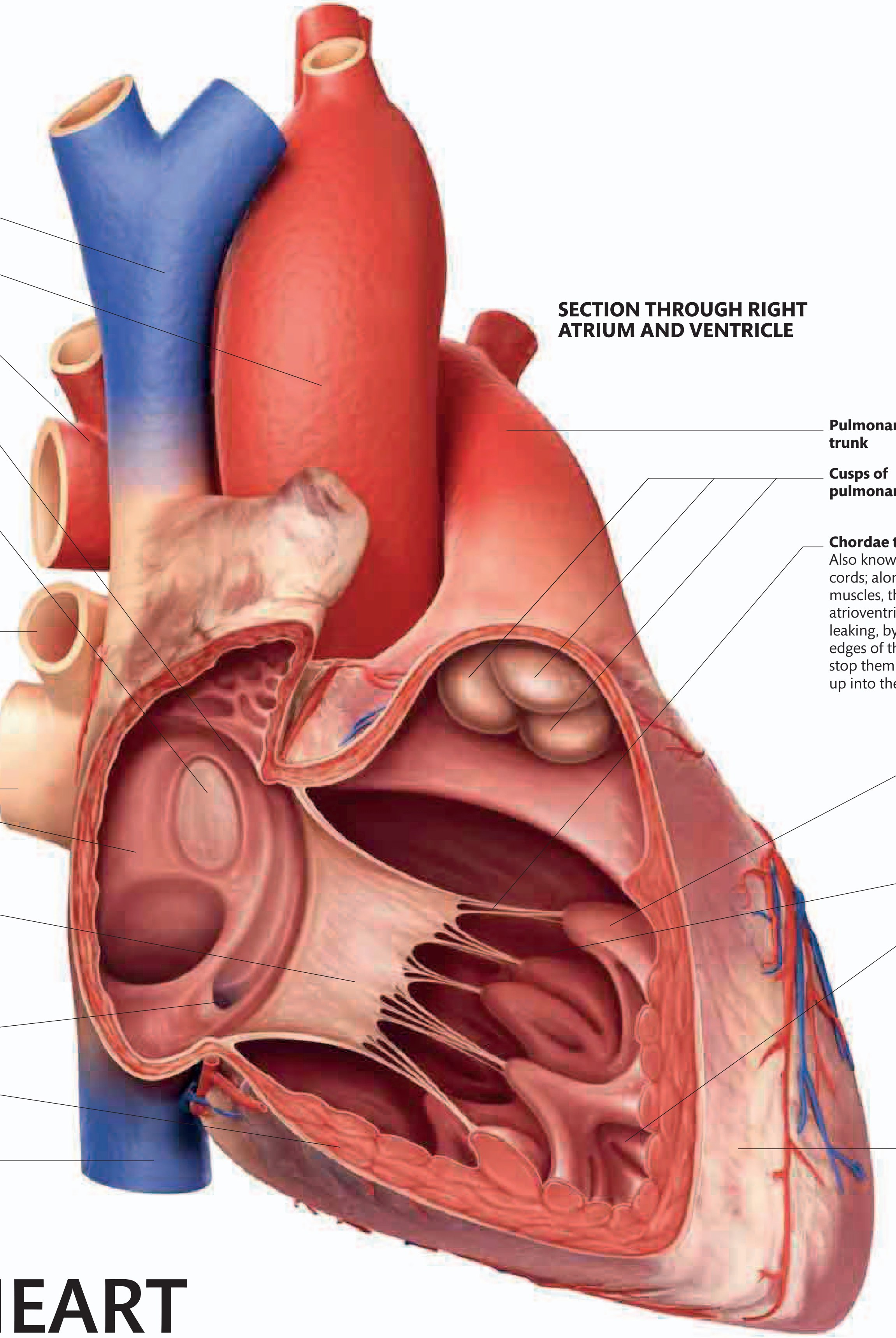

SECTION THROUGH RIGHT ATRIUM AND VENTRICLE

Superior vena cava

Ascending aorta

Branch of right pulmonary artery

Interatrial septum The dividing wall between the two atria

Oval fossa In the fetal heart, a valvelike opening allowing blood to flow between the right and left atria; it closes at birth—failure to close results in a "hole in the heart" defect

Right pulmonary veins

Right atrium Atrium in Latin means hall or court

Tricuspid valve Prevents blood flowing back into the right atrium when the ventricle contracts

Opening of coronary sinus

Myocardium Muscular wall of heart

Inferior vena cava

Pulmonary trunk

Cusps of pulmonary valve

Chordae tendineae Also known as tendinous cords; along with the papillary muscles, these prevent the atrioventricular valve from leaking, by pulling on the edges of the valve cusps to stop them from flapping back up into the atrium

Papillary muscle Named after the shape—papilla means nipple in Latin

Right ventricle

Trabeculae carneae This means fleshy beams in Latin; these muscular columns and ridges are characteristic of the inner surfaces of the ventricles

Serous pericardium Forms the outer layer of the heart; the word pericardium comes from the Greek for around the heart, while serous comes from the Latin word serum, meaning whey or watery fluid

HEART

The heart receives blood from veins and pumps it out through arteries. It has four chambers: two atria and two ventricles. The heart's left and right sides are separate. The right side receives deoxygenated blood from the body via the superior and inferior venae cavae, and pumps it to the lungs through the pulmonary trunk. The left gets oxygenated blood from the lungs via the pulmonary veins, and pumps it into the aorta for distribution. Each atrium opens into its corresponding ventricle via a valve (on the right, the tricuspid valve, and the bicuspid valve on the left), which shuts when the ventricle contracts, to stop blood flowing back into the atrium. The aorta and pulmonary trunk also have valves.

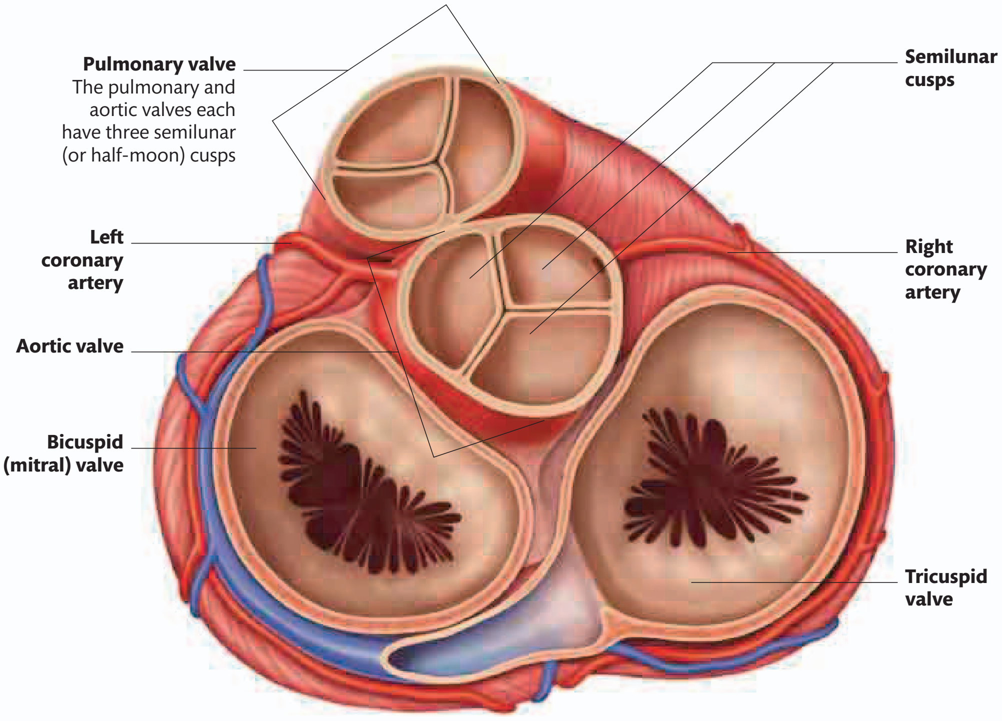
Pulmonary valve
The pulmonary and aortic valves each have three semilunar (or half-moon) cusps
Left coronary artery
Aortic valve
Bicuspid (mitral) valve
Semilunar cusps
Right coronary artery
Tricuspid valve

TRANSVERSE SECTION SHOWING VALVES

CARDIAC MUSCLE STRUCTURE

Cardiac muscle (myocardium) is essential to the heart's function as a pump. When the myocardium contracts, blood is squeezed out of the heart. The myocardium is a network of interconnected fibers, which contract rhythmically and spontaneously. Autonomic nerves can adjust the rate of contraction, matching the heart's output to the body's need.

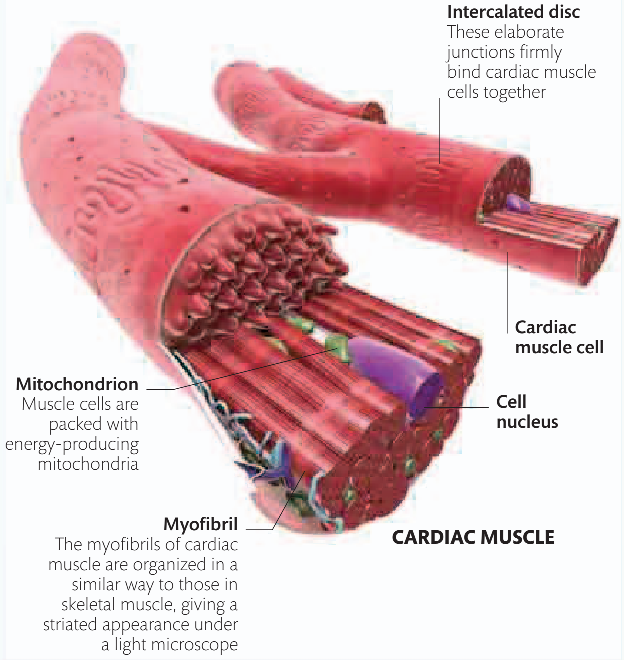
Intercalated disc
These elaborate junctions firmly bind cardiac muscle cells together
Cardiac muscle cell
Cell nucleus
Mitochondrion
Muscle cells are packed with energy-producing mitochondria
Myofibril
The myofibrils of cardiac muscle are organized in a similar way to those in skeletal muscle, giving a striated appearance under a light microscope

CARDIAC MUSCLE

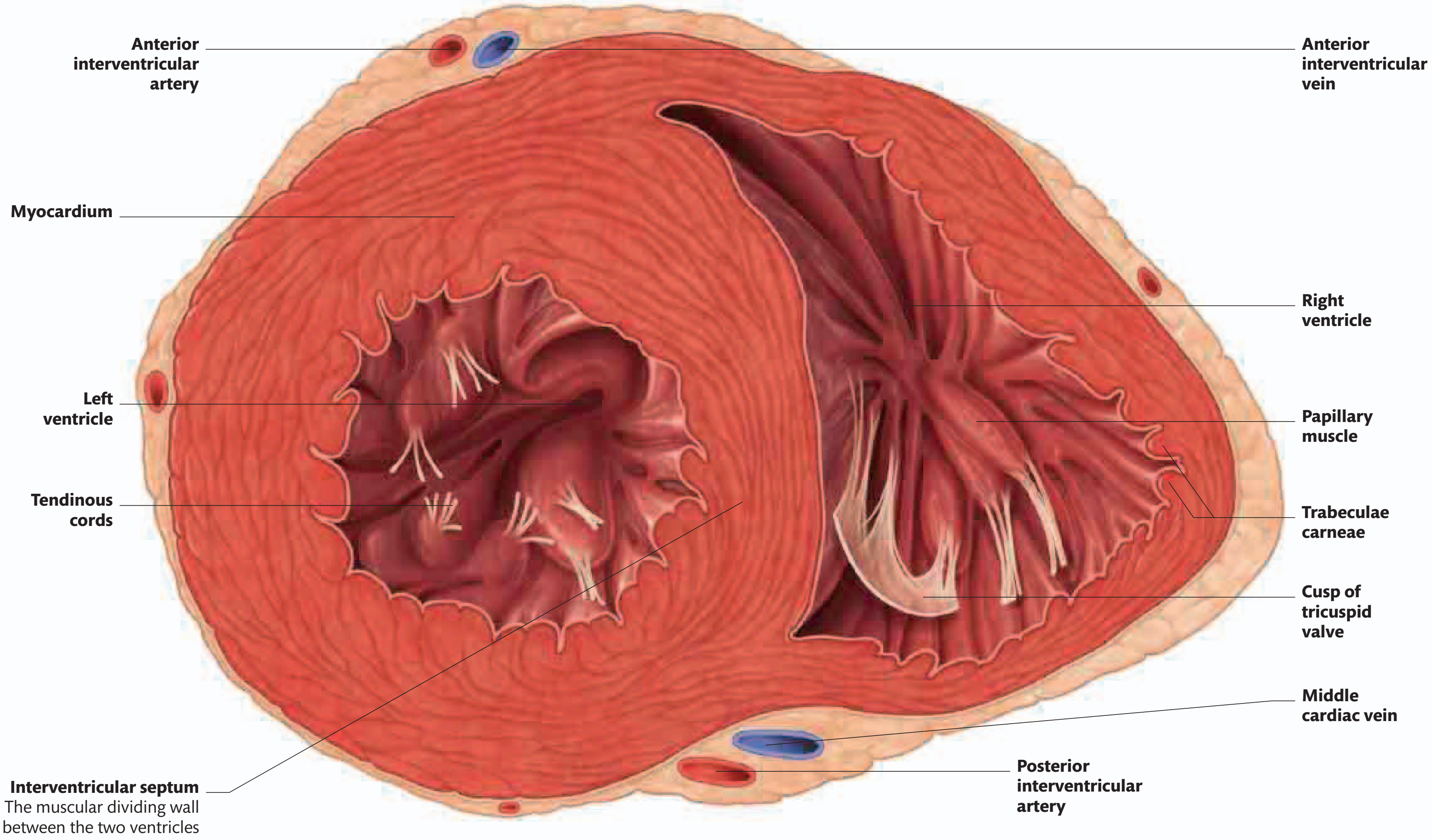
Anterior interventricular artery
Myocardium
Left ventricle
Tendinous cords
Interventricular septum
The muscular dividing wall between the two ventricles
Anterior interventricular vein
Right ventricle
Papillary muscle
Trabeculae carneae
Cusp of tricuspid valve
Middle cardiac vein
Posterior interventricular artery

TRANSVERSE SECTION THROUGH VENTRICLES

ABDOMEN AND PELVIS

The aorta passes behind the diaphragm, level with the twelfth thoracic vertebra, and enters the abdomen. Pairs of arteries branch from the sides of the aorta to supply the walls of the abdomen, the kidneys, suprarenal glands, and the testes or ovaries with oxygenated blood. A series of branches emerge from the front of the abdominal aorta to supply the abdominal organs: the celiac trunk gives branches to the liver, stomach, pancreas, and spleen, and the mesenteric arteries provide blood to the gut. The abdominal aorta ends by splitting into two, forming the common iliac arteries. Each of these then divides, in turn, forming an internal iliac artery (which supplies the pelvic organs) and an external iliac artery (which continues into the thigh, becoming the femoral artery). Lying to the right of the aorta is the major vein of the abdomen: the inferior vena cava.

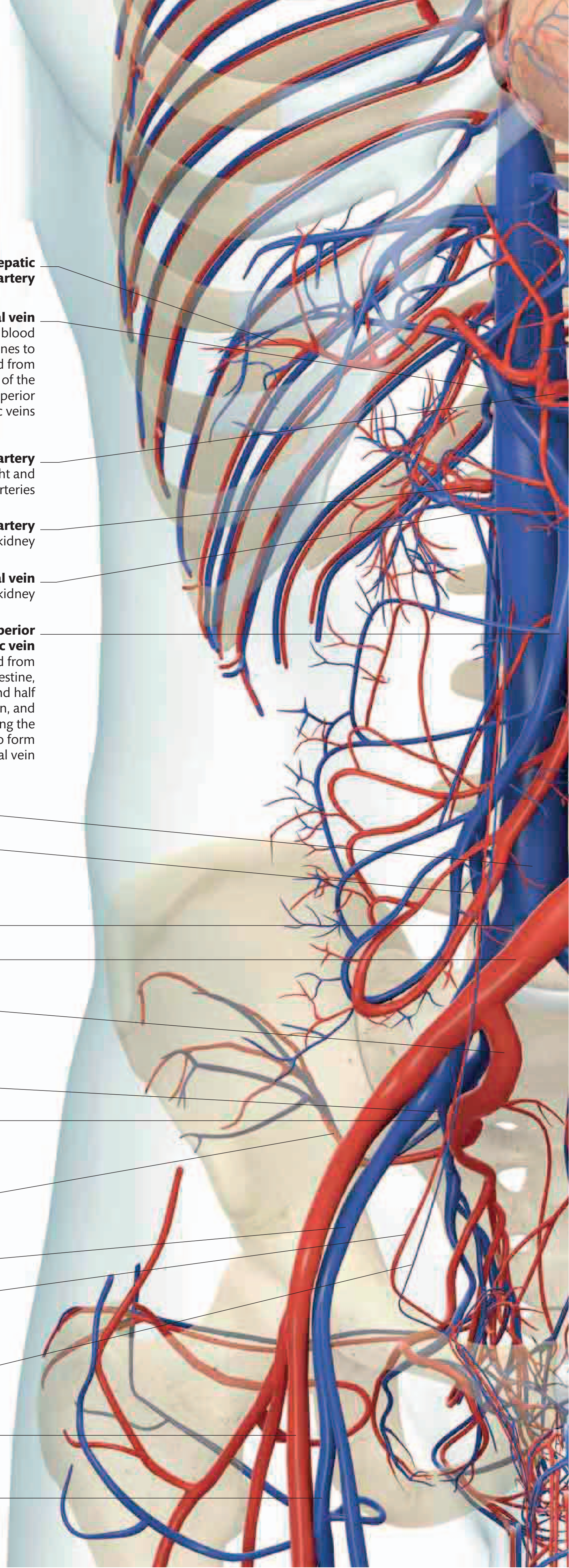

Right hepatic artery

Portal vein
Carries blood from the intestines to the liver; formed from the joining of the splenic and superior mesenteric veins

Common hepatic artery
Branches into right and left hepatic arteries

Right renal artery
Supplies the right kidney

Right renal vein
Drains the right kidney

Superior mesenteric vein
Drains blood from the small intestine, cecum, and half of the colon, and ends by joining the splenic vein to form the portal vein

Inferior vena cava

Ileocolic artery
Branch of the superior mesenteric artery supplying the end of the ileum, the cecum, the start of the ascending colon, and the appendix

Right common iliac vein

Right common iliac artery
Divides into the right external and internal iliac arteries

Right internal iliac artery
Provides branches to the bladder, rectum, perineum, and external genitals, muscles of the inner thigh, bone of the ilium and sacrum, and the buttock, as well as the uterus and vagina in a woman

Right internal iliac vein

Right external iliac artery
Gives a branch to the lower part of the anterior abdominal wall before passing over the pubic bone and under the inguinal ligament to become the femoral artery

Right superior gluteal artery
The largest branch of the internal iliac artery; passes out through the back of the pelvis to supply the upper buttock

Right external iliac vein

Right gonadal artery
In a woman, supplies the ovary on each side; in a man, extends to the scrotum to supply the testis

Right gonadal vein
Drains the ovary or testis and ends by joining the inferior vena cava

Right femoral artery
The main artery of the leg; the continuation of the external iliac artery in the thigh

ANTERIOR (FRONT)

Right femoral vein

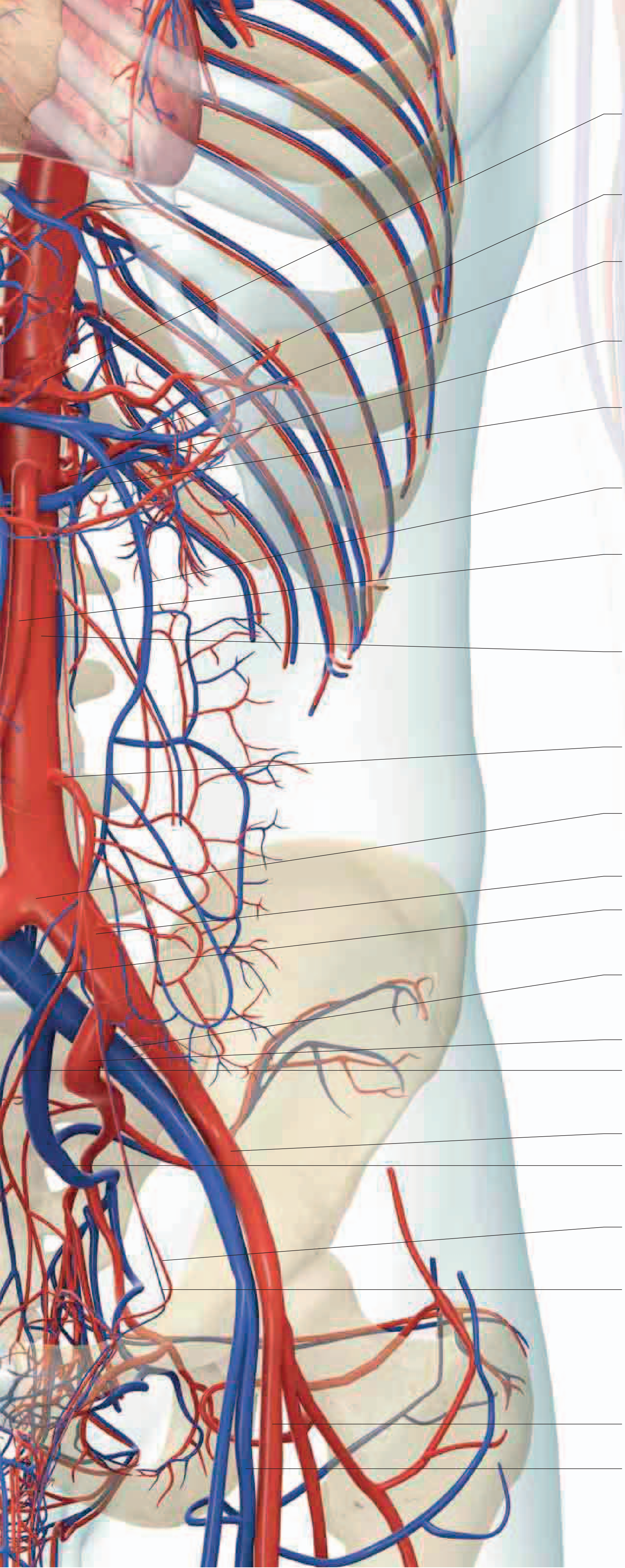

Celiac trunk
Only just over ⅜in (1cm) long, it quickly branches into the left gastric, splenic, and common hepatic arteries

Splenic artery
Supplies the spleen, as well as most of the pancreas, and the upper part of the stomach

Splenic vein
Drains the spleen and receives other veins from the stomach and pancreas, as well as the inferior mesenteric vein

Left renal artery
Shorter than the right renal artery, this supplies the left kidney

Left renal vein
Longer than its counterpart on the right, this drains the left kidney and receives the left gonadal vein

Inferior mesenteric vein
Drains blood from the colon and rectum and ends by emptying into the splenic vein

Superior mesenteric artery
Branches within the mesentery to supply a great length of intestine, including all of the jejunum and ileum and half of the colon

Abdominal aorta
The thoracic aorta becomes the abdominal aorta as it passes behind the diaphragm, level with the twelfth thoracic vertebra

Inferior mesenteric artery
Supplies the last third of the transverse colon, the descending and sigmoid colon, and the rectum

Bifurcation of aorta
The abdominal aorta divides in front of the fourth lumbar vertebra

Left common iliac artery

Left common iliac vein
Formed from the union of the external and internal iliac veins

Left external iliac vein
The continuation of the femoral vein, after it has passed into the pelvis

Left internal iliac artery

Superior rectal artery
The last branch of the inferior mesenteric artery passes down into the pelvis to supply the rectum

Left external iliac artery

Left internal iliac vein
Drains the pelvic organs, perineum, and buttock

Left gonadal artery
Gonadal arteries branch from the aorta just below the renal arteries

Left gonadal vein
Drains the ovary or testis, and empties into the left renal vein

Left femoral artery

Left femoral vein
The main vein from the leg; becomes the external iliac vein

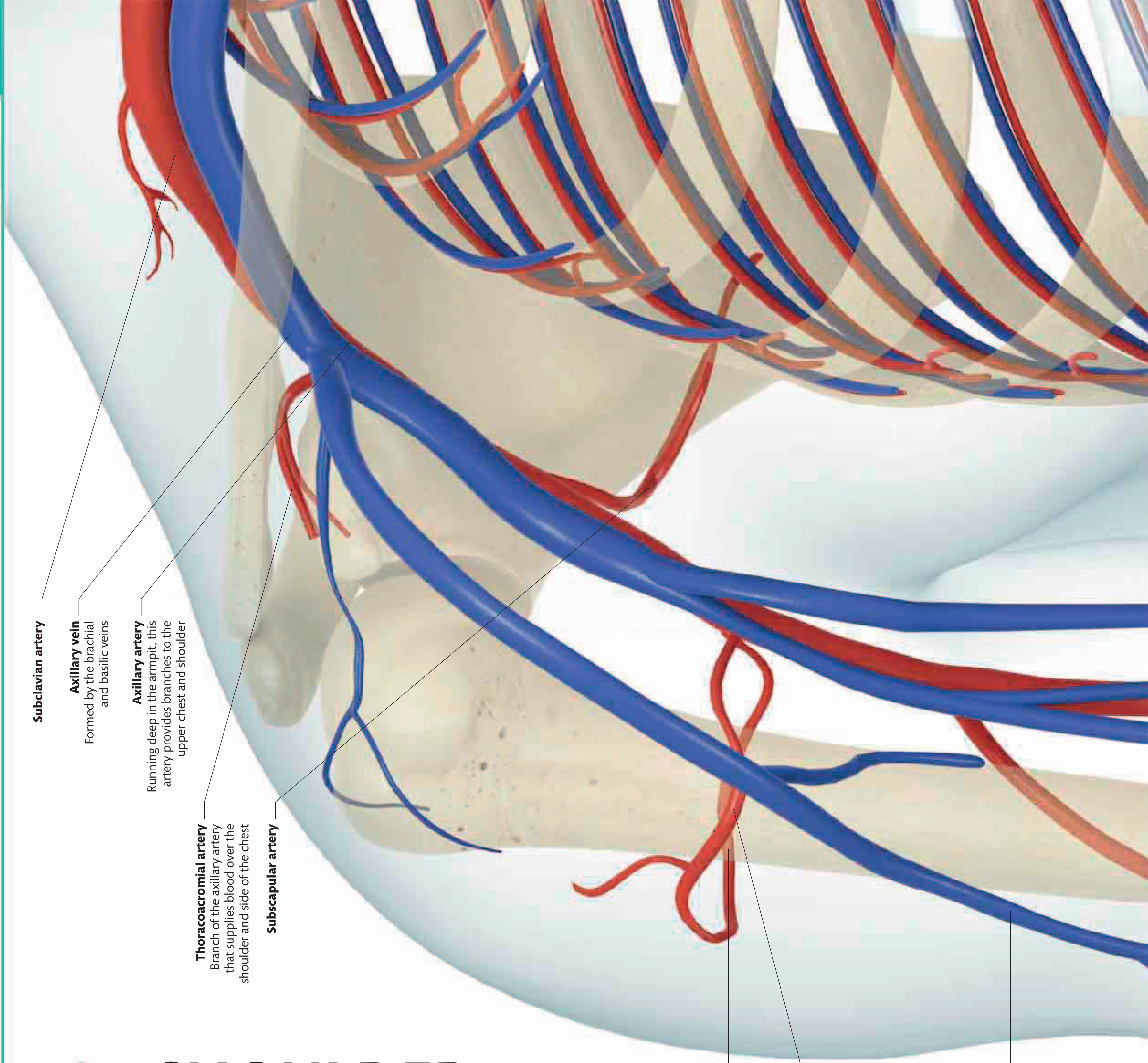
Subclavian artery
Axillary vein
Formed by the brachial and basilic veins
Axillary artery
Running deep in the armpit, this artery provides branches to the upper chest and shoulder
Thoracoacromial artery
Branch of the axillary artery that supplies blood over the shoulder and side of the chest
Subscapular artery
Posterior circumflex humeral artery
Circumflex means bent around in Latin
Anterior circumflex humeral artery
Looping in front of the neck of the humerus to join up with the posterior circumflex humeral artery, this artery supplies the shoulder joint and muscles
Cephalic vein
Courses up the outer side of the upper arm, just under the skin; runs deep under the clavicle to join the axillary vein

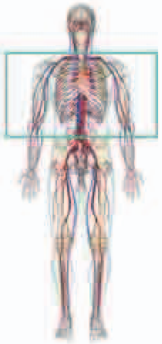

SHOULDER AND UPPER ARM

The subclavian artery is the main arterial supply to the upper limb. When this artery passes under the clavicle and into the axilla (armpit), it becomes the axillary artery. Several branches spring off in this region, running backward toward the scapula, up to the shoulder, and around the humerus. Beyond the armpit, the name of the axillary artery changes to the brachial artery, which runs down the front of the arm, usually accompanied by a pair of companion veins. Two superficial veins that drain blood from the back of the hand end in the arm by draining into deep veins: the basilic vein drains into brachial veins; the cephalic vein runs up to the shoulder, then plunges deeper to join the axillary vein.

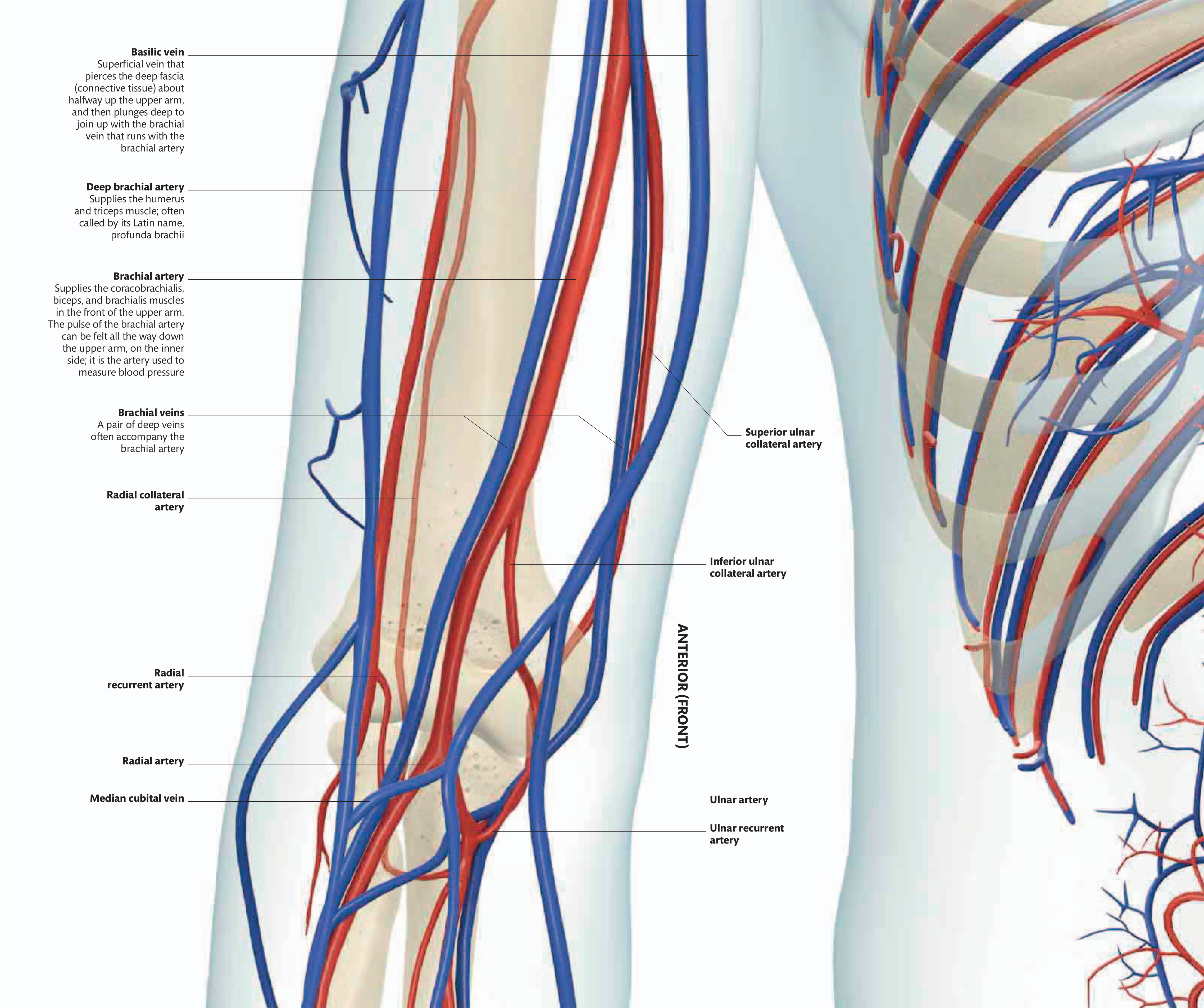

ANTERIOR (FRONT)

Basilic vein
Superficial vein that pierces the deep fascia (connective tissue) about halfway up the upper arm, and then plunges deep to join up with the brachial vein that runs with the brachial artery

Deep brachial artery
Supplies the humerus and triceps muscle; often called by its Latin name, profunda brachii

Brachial artery
Supplies the coracobrachialis, biceps, and brachialis muscles in the front of the upper arm. The pulse of the brachial artery can be felt all the way down the upper arm, on the inner side; it is the artery used to measure blood pressure

Brachial veins
A pair of deep veins often accompany the brachial artery

Radial collateral artery

Radial recurrent artery

Radial artery

Median cubital vein

Superior ulnar collateral artery

Inferior ulnar collateral artery

Ulnar artery

Ulnar recurrent artery

SHOULDER AND UPPER ARM

Various branches from the axillary and brachial arteries supply the back of the shoulder and upper arm. The posterior circumflex humeral artery, which runs with the axillary nerve, curls around the upper end of the humerus. The deep brachial artery runs with the radial nerve, spiraling around the back of the bone. From this artery, and from the brachial artery itself, collateral branches run down the arm and join up, or anastomose, with recurrent branches running back up from the ulnar and radial arteries of the forearm. There are also anastomoses (links) between branches of the subclavian and axillary arteries around the shoulder. Anastomoses like this, where branches from different regions join up, can provide alternative routes through which blood can flow if the main vessel becomes squashed or blocked.

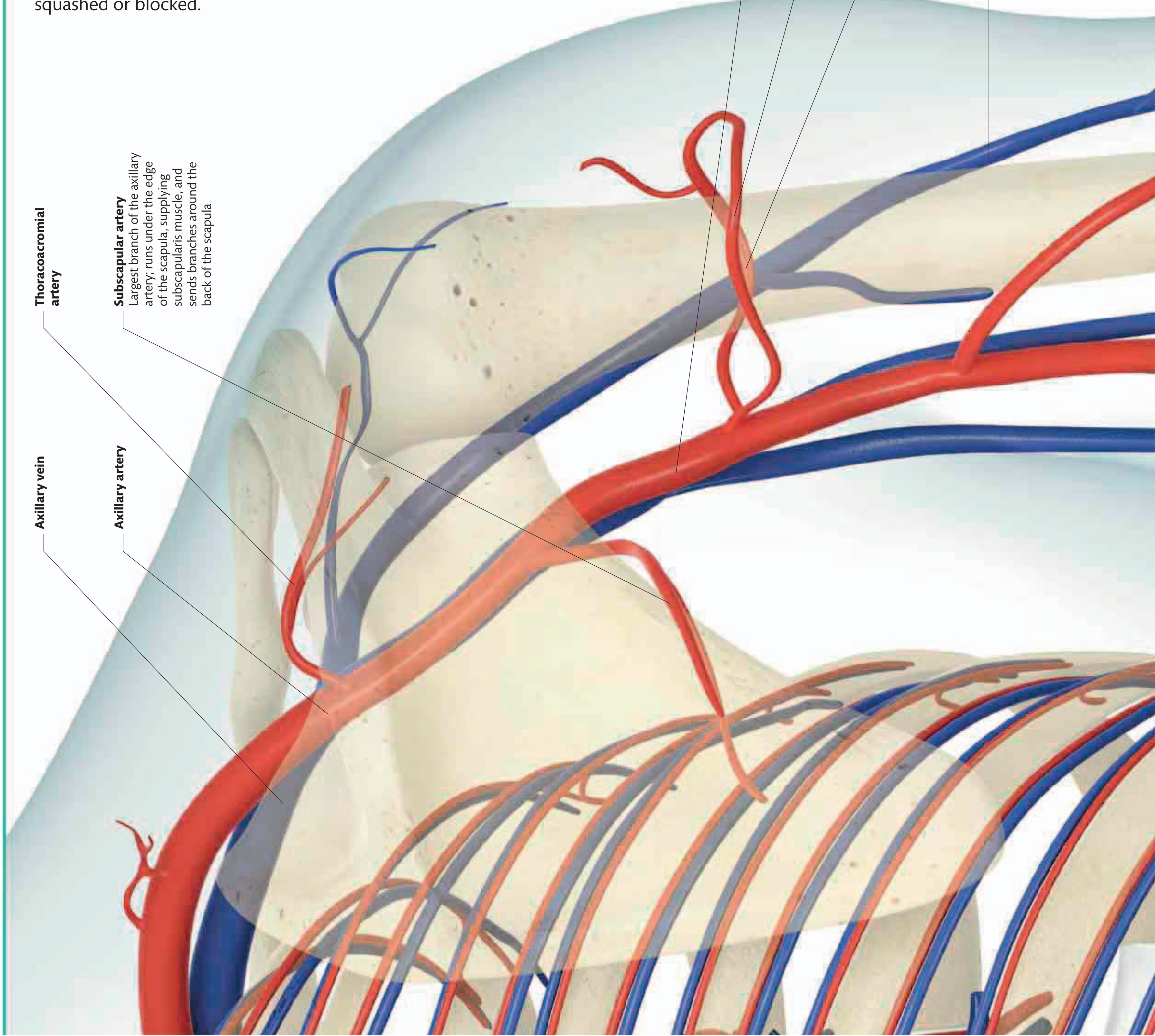
Brachial artery
Posterior circumflex humeral artery
Anterior circumflex humeral artery
Cephalic vein
Thoracoacromial artery
Subscapular artery Largest branch of the axillary artery; runs under the edge of the scapula, supplying subscapularis muscle, and sends branches around the back of the scapula
Axillary vein
Axillary artery

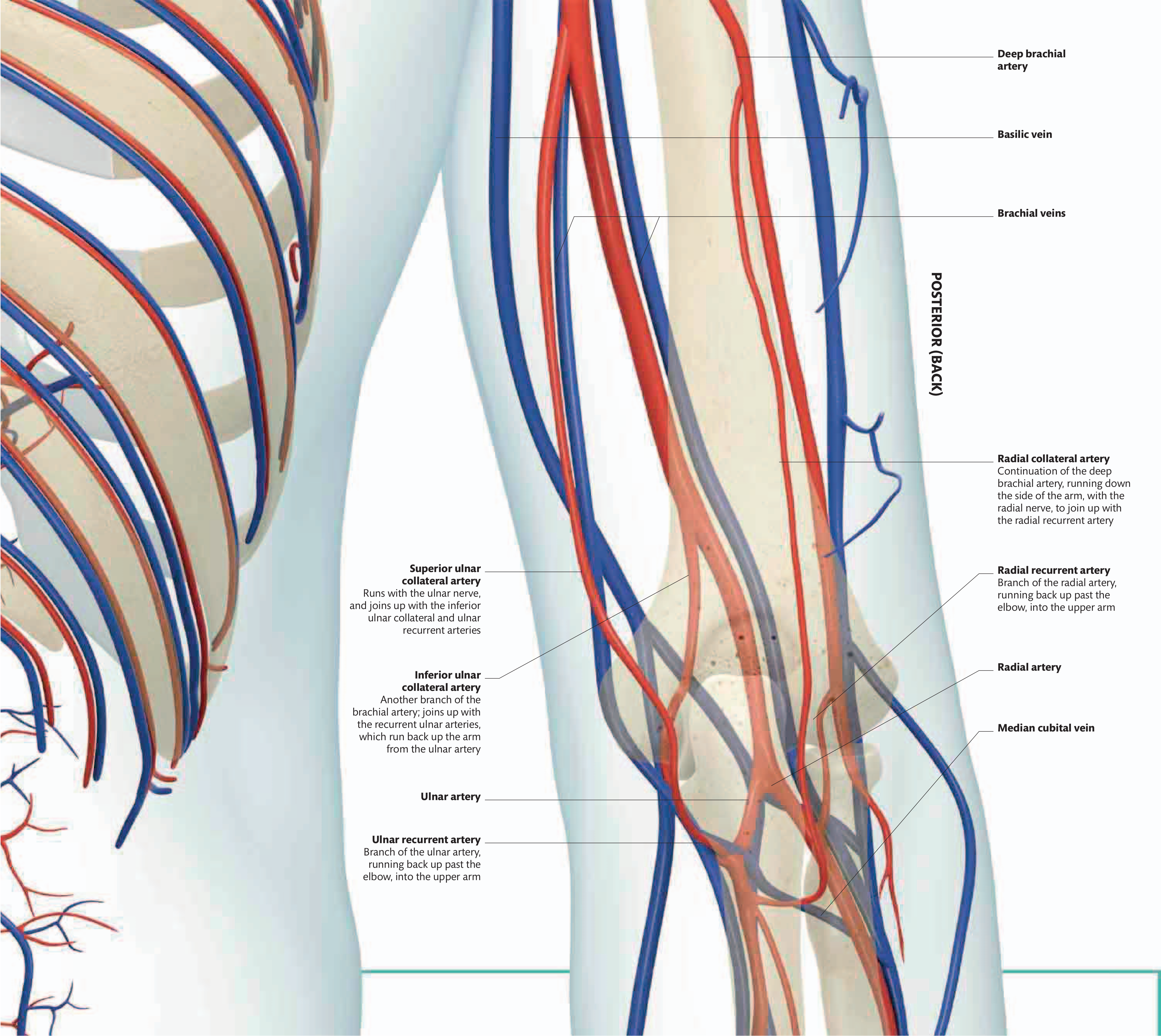

POSTERIOR (BACK)

Deep brachial artery

Basilic vein

Brachial veins

Radial collateral artery Continuation of the deep brachial artery, running down the side of the arm, with the radial nerve, to join up with the radial recurrent artery

Radial recurrent artery Branch of the radial artery, running back up past the elbow, into the upper arm

Radial artery

Median cubital vein

Superior ulnar collateral artery Runs with the ulnar nerve, and joins up with the inferior ulnar collateral and ulnar recurrent arteries

Inferior ulnar collateral artery Another branch of the brachial artery; joins up with the recurrent ulnar arteries, which run back up the arm from the ulnar artery

Ulnar artery

Ulnar recurrent artery Branch of the ulnar artery, running back up past the elbow, into the upper arm

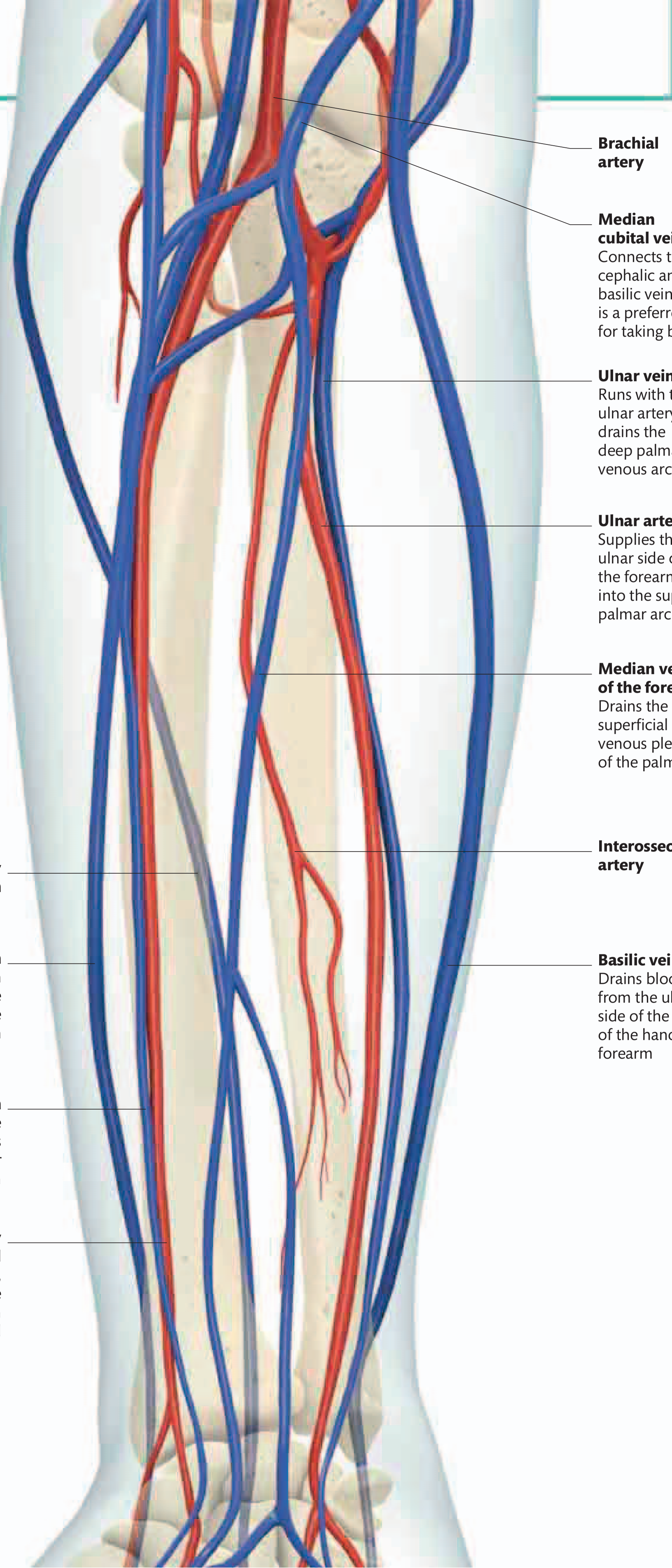
ANTERIOR (FRONT)
Brachial artery
Median cubital vein Connects the cephalic and basilic veins; is a preferred site for taking blood
Ulnar vein Runs with the ulnar artery; drains the deep palmar venous arch
Ulnar artery Supplies the ulnar side of the forearm; feeds into the superficial palmar arch
Median vein of the forearm Drains the superficial venous plexus of the palm
Interosseous artery
Basilic vein Drains blood from the ulnar side of the back of the hand and forearm
Accessory cephalic vein
Cephalic vein Drains blood from the radial side of the back of the hand and forearm
Radial vein Runs with the radial artery; drains the superficial palmar venous arch
Radial artery Supplies the radial side of the forearm, and feeds into the deep palmar arch of the hand

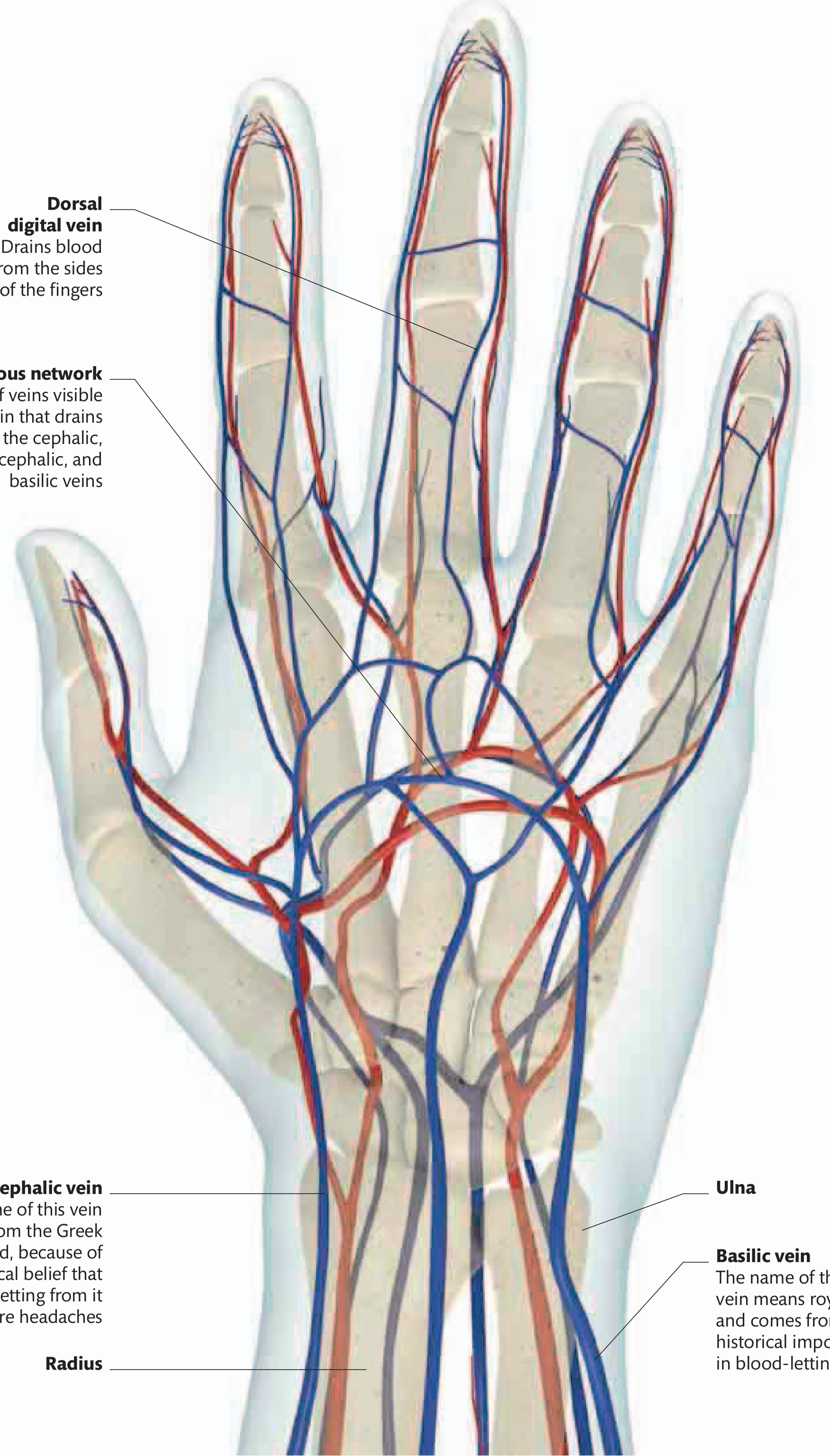
Dorsal digital vein Drains blood from the sides of the fingers
Dorsal venous network A plexus of veins visible under the skin that drains blood into the cephalic, accessory cephalic, and basilic veins
Cephalic vein The name of this vein comes from the Greek for head, because of the historical belief that blood-letting from it could cure headaches
Radius
Ulna
Basilic vein The name of this vein means royal and comes from its historical importance in blood-letting

LOWER ARM AND HAND

The brachial artery divides into two arteries, which take their names from the bones of the forearm: the radial and ulnar arteries. The radial artery can be felt at the wrist, and this is the most common place for taking a pulse as the strong pulsations are easy to feel when the artery is pressed against the bone beneath it. Furthermore, taking a pulse here doesn't even require any undressing. The radial and ulnar arteries end by joining up to form arterial arches in the wrist and palm. Digital arteries, destined for the fingers, spring off from the palmar arch. Superficial veins are concentrated on the back of the hand, rather than on the palm—otherwise those thin-walled vessels would be compressed every time a person gripped something. The dorsal venous network of the hand drains into two main vessels: the basilic and cephalic veins.

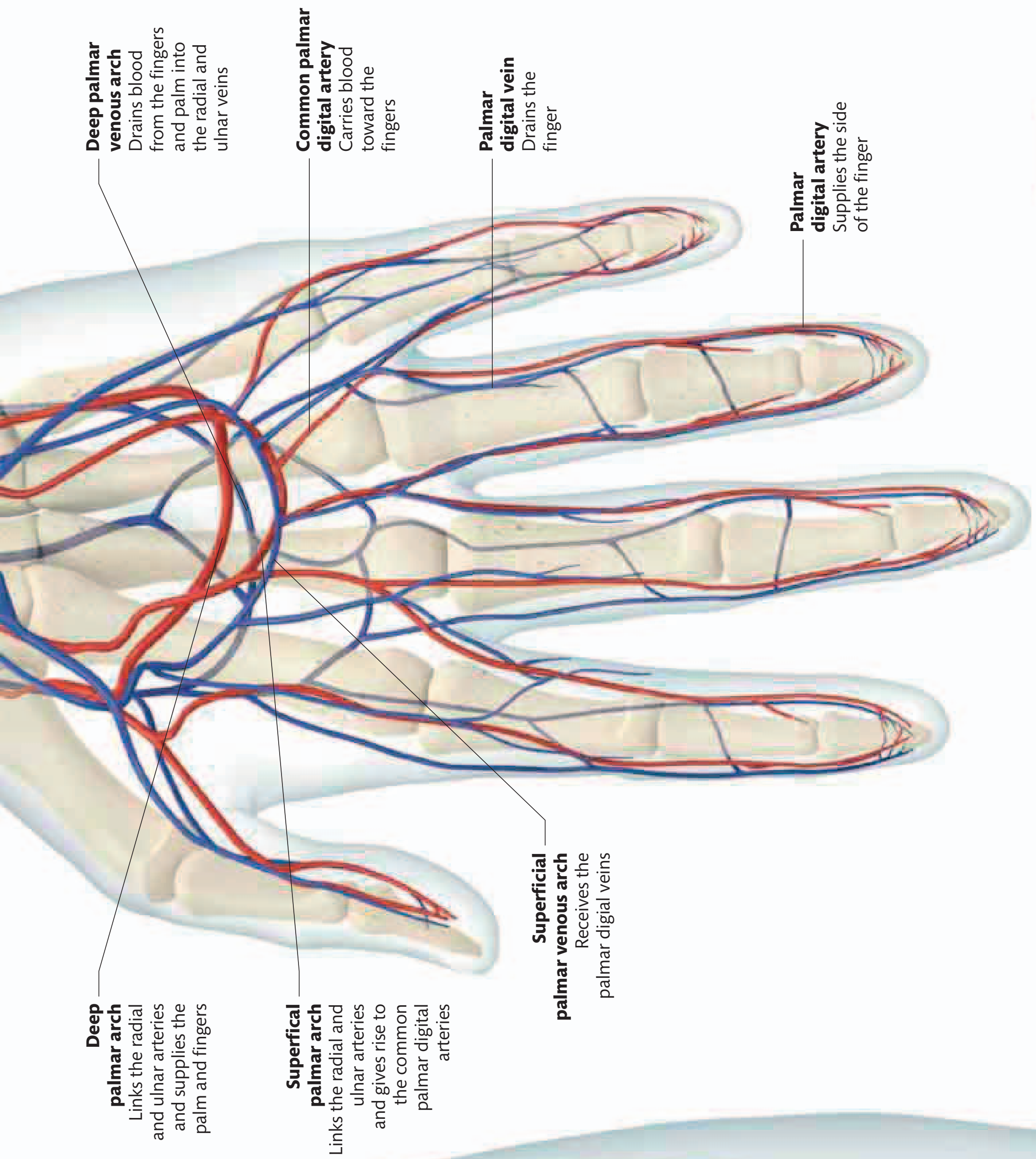
Deep palmar venous arch Drains blood from the fingers and palm into the radial and ulnar veins
Common palmar digital artery Carries blood toward the fingers
Palmar digital vein Drains the finger
Palmar digital artery Supplies the side of the finger
Deep palmar arch Links the radial and ulnar arteries and supplies the palm and fingers
Superfical palmar arch Links the radial and ulnar arteries and gives rise to the common palmar digital arteries
Superficial palmar venous arch Receives the palmar digial veins

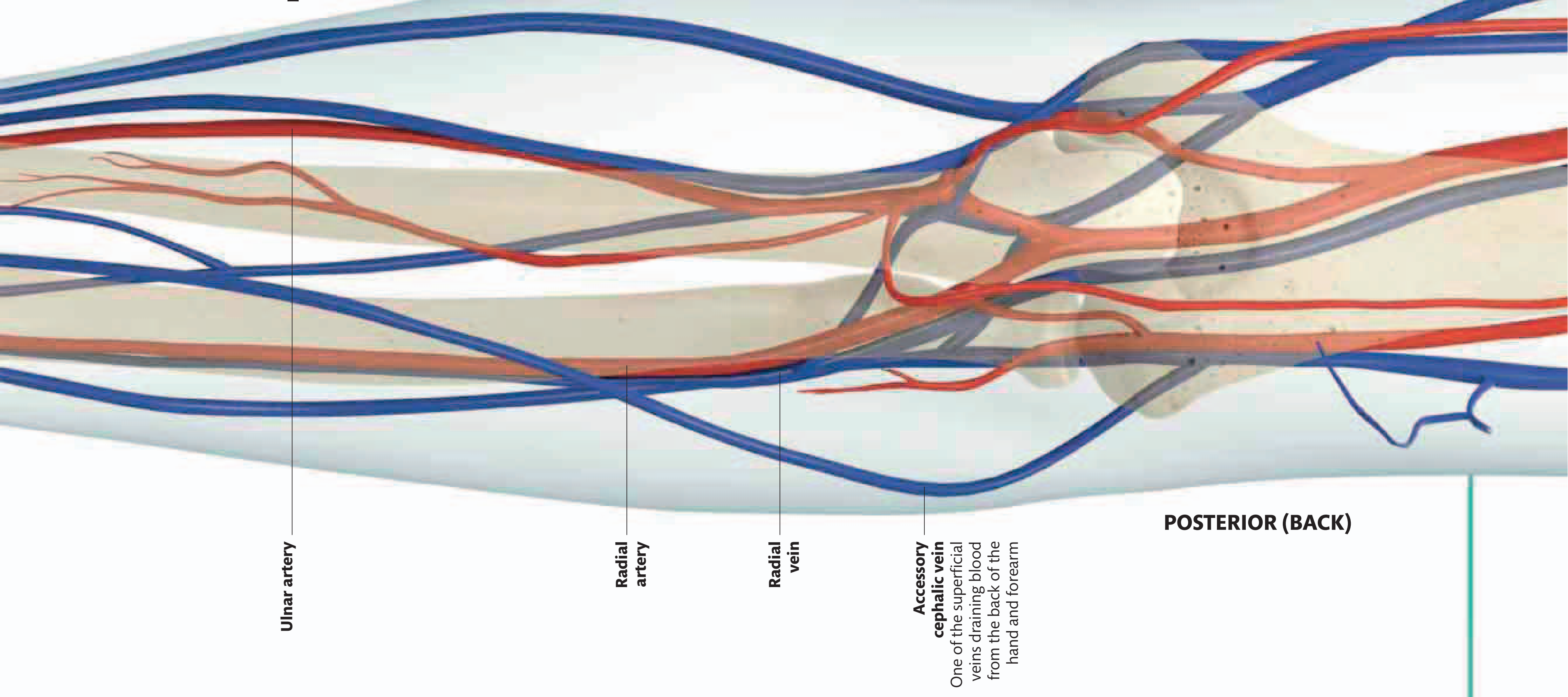
Ulnar artery
Radial artery
Radial vein
Accessory cephalic vein One of the superficial veins draining blood from the back of the hand and forearm
POSTERIOR (BACK)

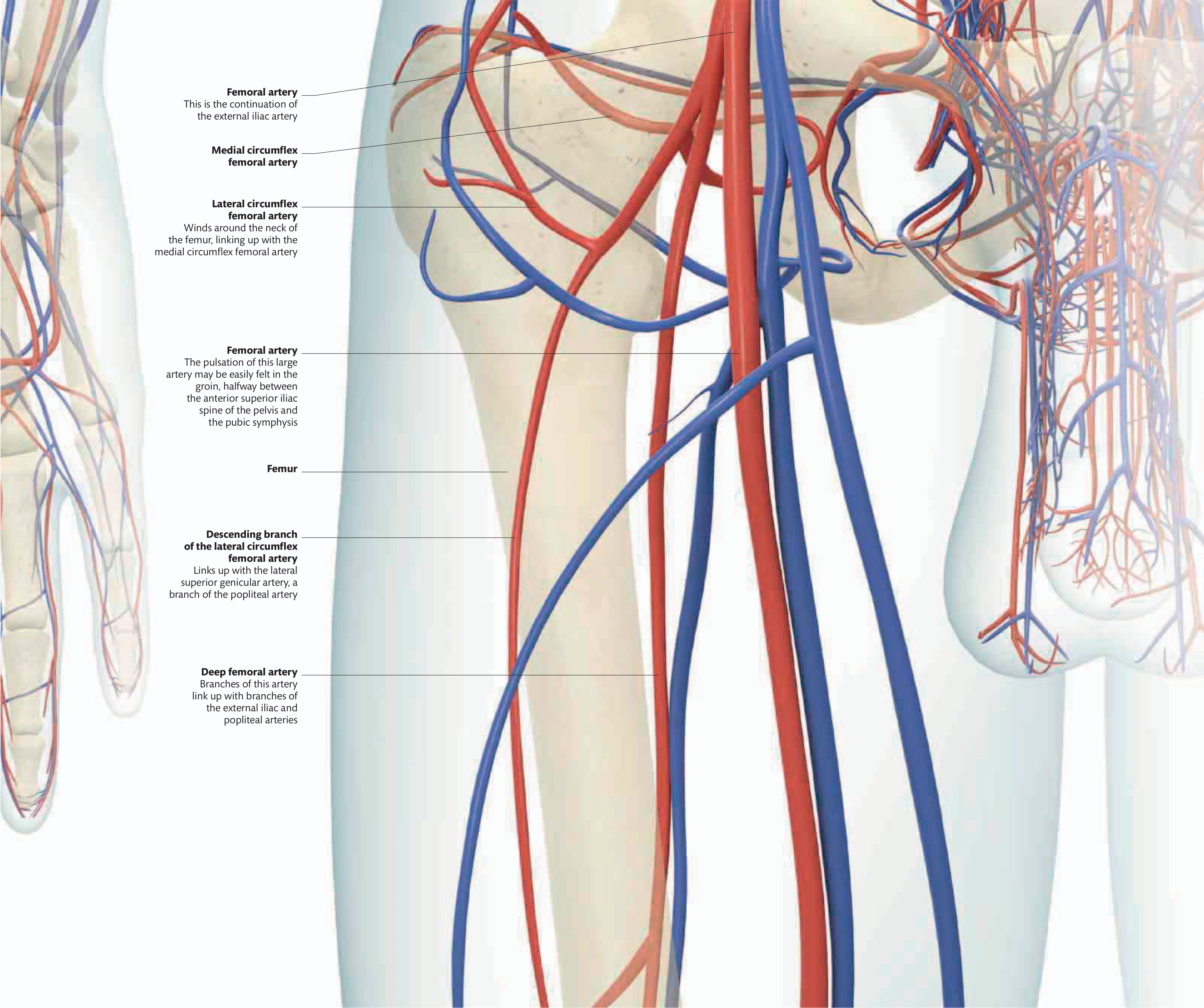

Femoral artery
This is the continuation of the external iliac artery

Medial circumflex femoral artery

Lateral circumflex femoral artery
Winds around the neck of the femur, linking up with the medial circumflex femoral artery

Femoral artery
The pulsation of this large artery may be easily felt in the groin, halfway between the anterior superior iliac spine of the pelvis and the pubic symphysis

Femur

Descending branch of the lateral circumflex femoral artery
Links up with the lateral superior genicular artery, a branch of the popliteal artery

Deep femoral artery
Branches of this artery link up with branches of the external iliac and popliteal arteries

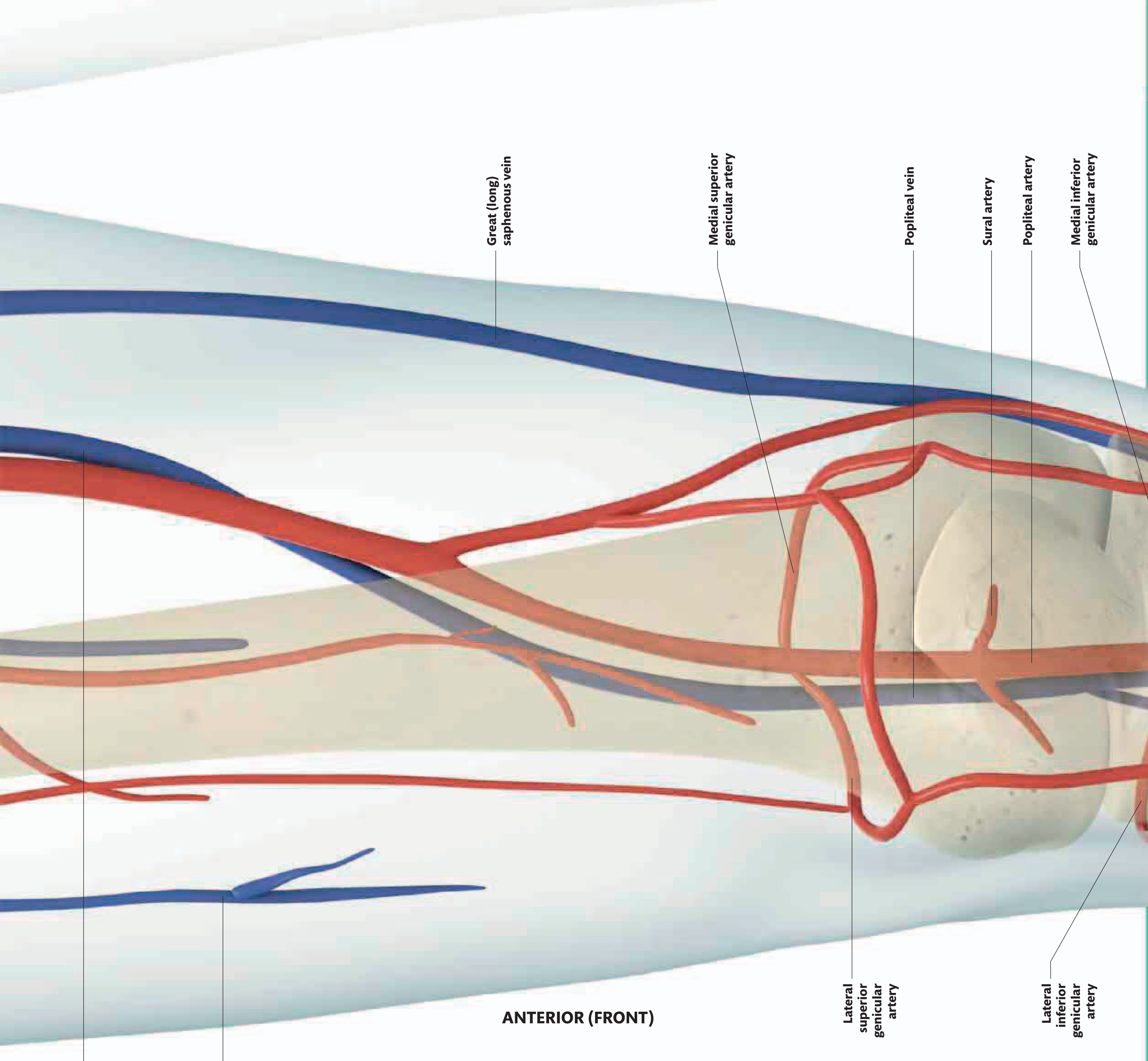
Great (long) saphenous vein
Medial superior genicular artery
Popliteal vein
Sural artery
Popliteal artery
Medial inferior genicular artery
Femoral vein
Accessory saphenous vein
ANTERIOR (FRONT)
Lateral superior genicular artery
Lateral inferior genicular artery

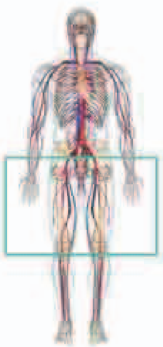

HIP AND THIGH

As the external iliac artery runs over the pubic bone and underneath the inguinal ligament, it changes its name to the femoral artery—the main vessel carrying blood to the lower limb. The femoral artery lies exactly halfway along a line between the anterior superior iliac spine of the pelvis and the pubic symphysis. It has a large branch, the deep femoral artery, that supplies the muscles of the thigh. The femoral artery then runs toward the inner thigh, passing through the hole in the adductor magnus tendon, where its name changes to the popliteal artery. Deep veins run with the arteries, but—just as in the arm—there are also superficial veins. The great (or long) saphenous vein drains up the inner side of the leg and thigh, and ends by joining the femoral vein near the hip.

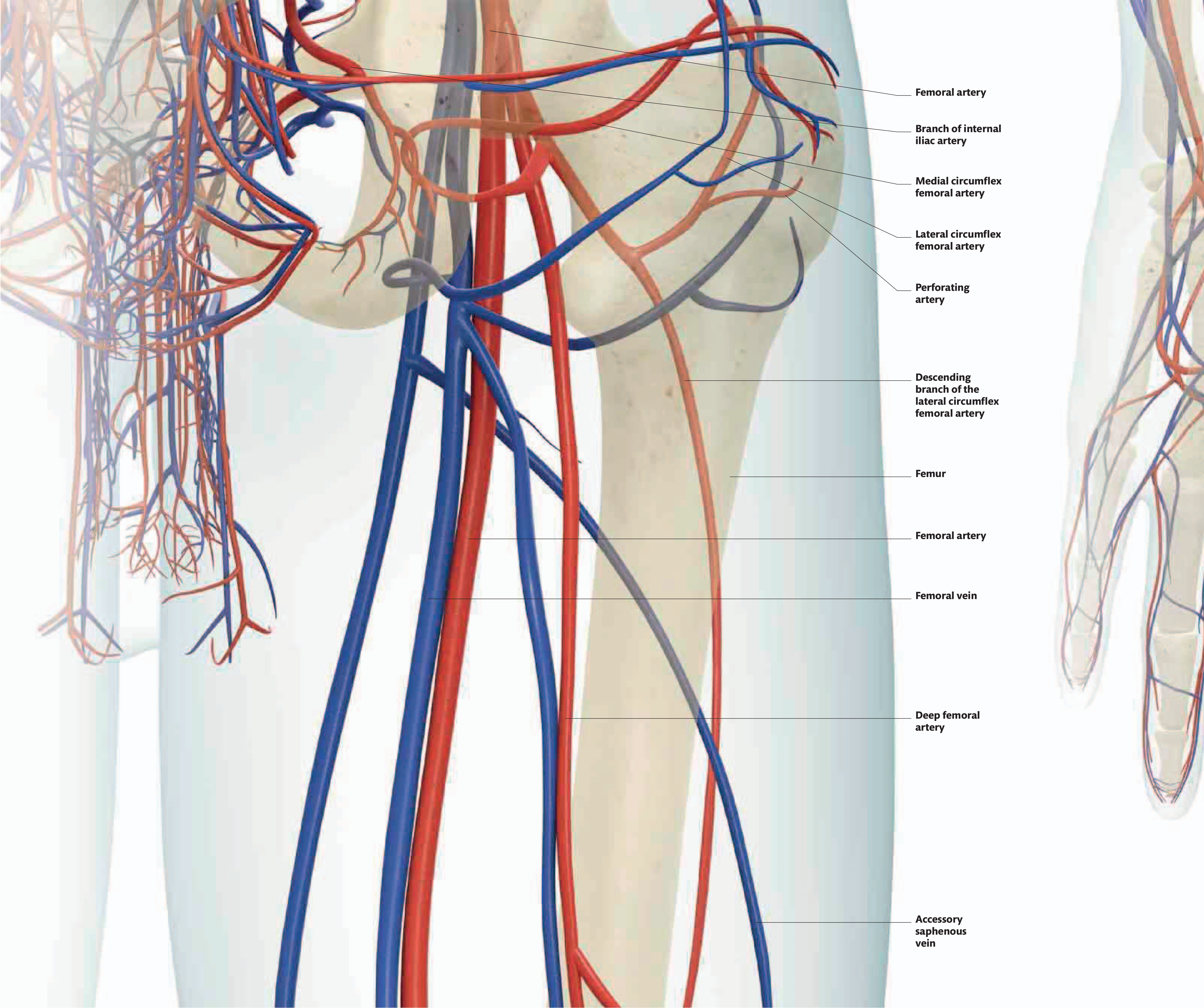
Femoral artery
Branch of internal iliac artery
Medial circumflex femoral artery
Lateral circumflex femoral artery
Perforating artery
Descending branch of the lateral circumflex femoral artery
Femur
Femoral artery
Femoral vein
Deep femoral artery
Accessory saphenous vein

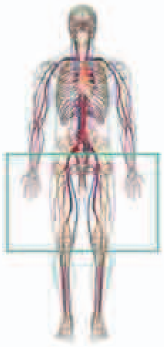

HIP AND THIGH

In this back view, gluteal branches of the internal iliac artery can be clearly seen, emerging through the greater sciatic foramen to supply the buttock. The muscles and skin of the inner part and back of the thigh are supplied by branches of the deep femoral artery. These are known as the perforating arteries because they pierce through the adductor magnus muscle. Higher up, the circumflex femoral arteries encircle the femur. The popliteal artery, formed after the femoral artery passes through the hiatus (gap) in adductor magnus, lies on the back of the femur, deep to the popliteal vein.

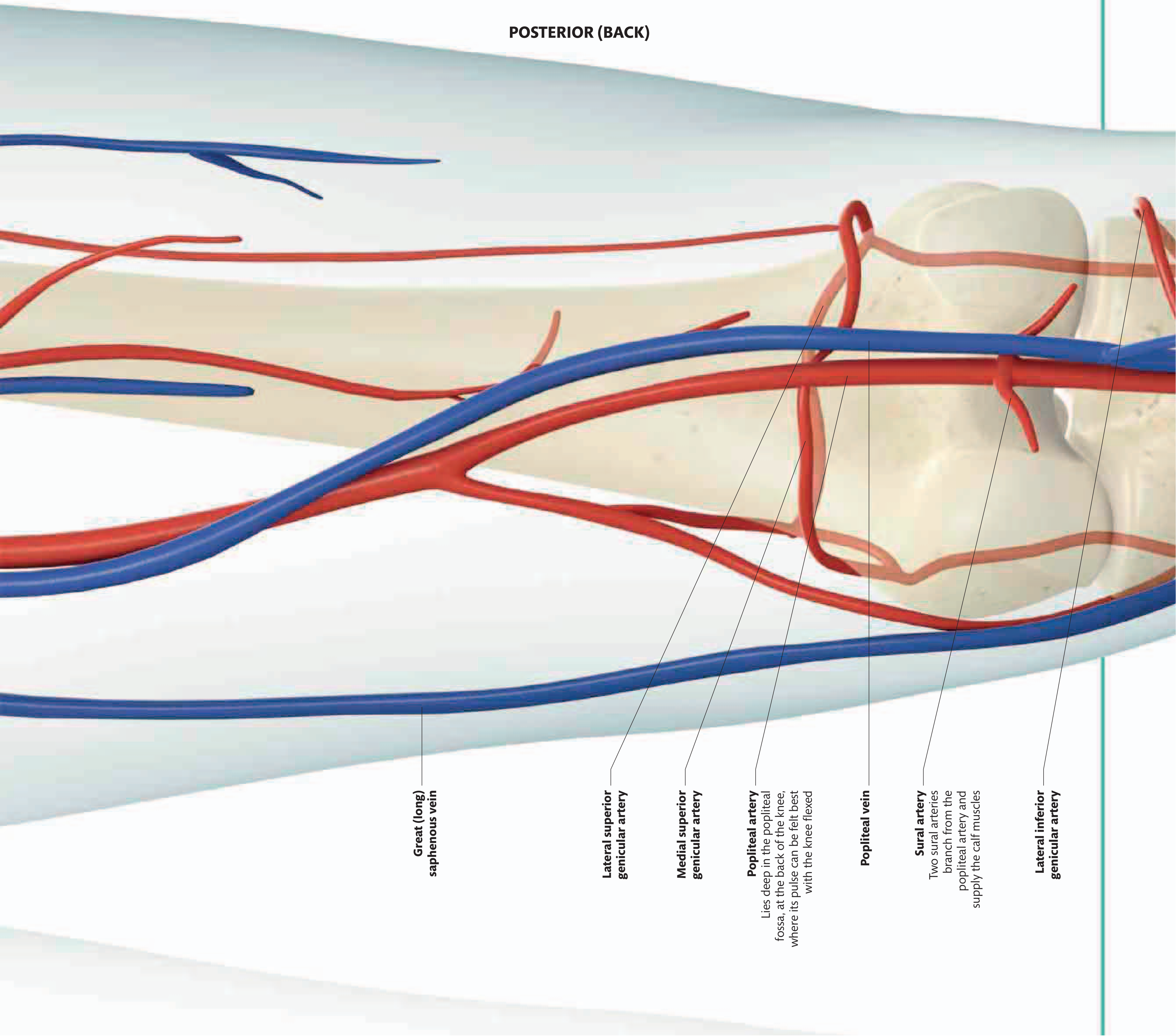
POSTERIOR (BACK)
Great (long) saphenous vein
Lateral superior genicular artery
Medial superior genicular artery
Popliteal artery
Lies deep in the popliteal fossa, at the back of the knee, where its pulse can be felt best with the knee flexed
Popliteal vein
Sural artery
Two sural arteries branch from the popliteal artery and supply the calf muscles
Lateral inferior genicular artery

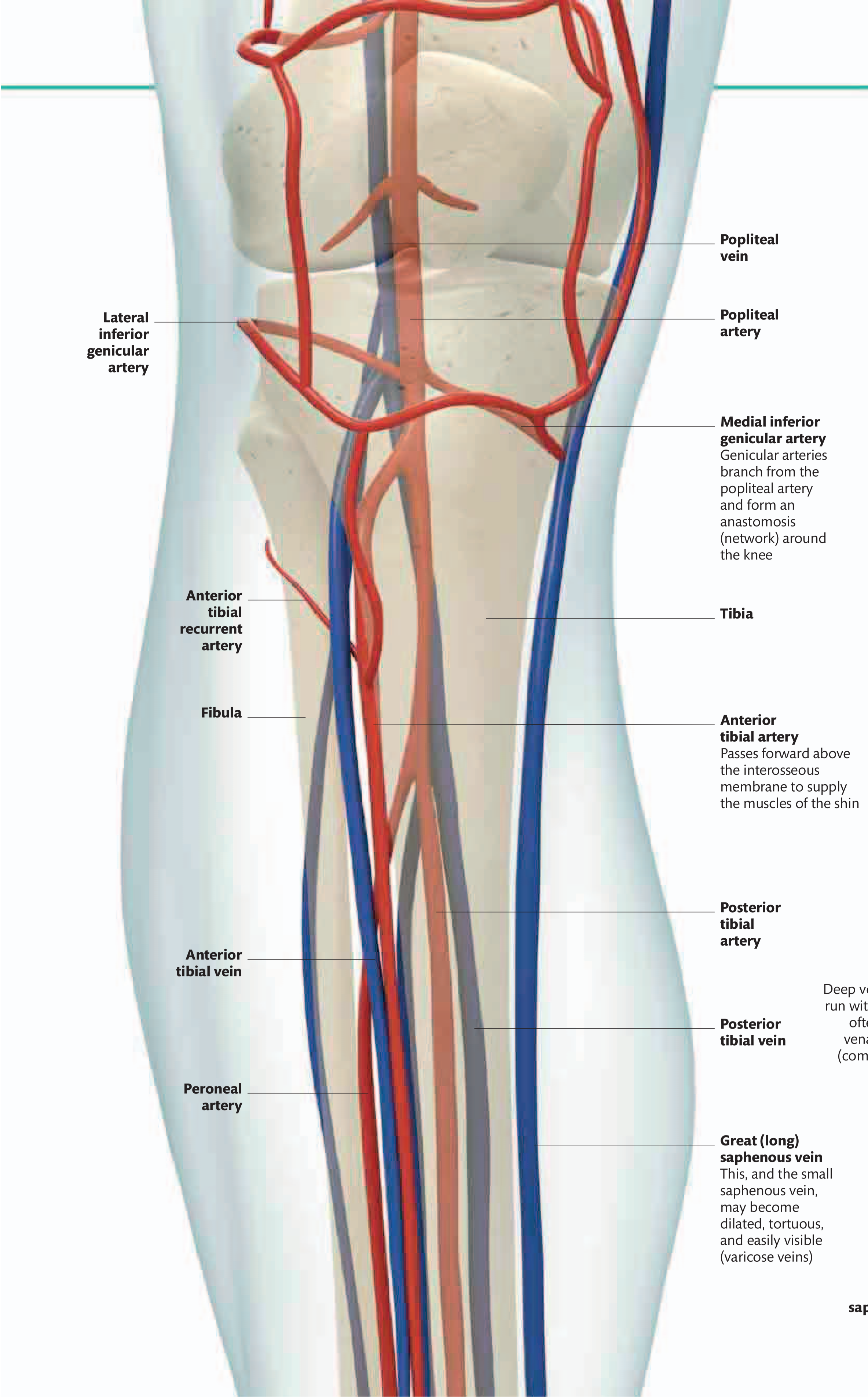
Lateral inferior genicular artery
Popliteal vein
Popliteal artery
Anterior tibial recurrent artery
Tibia
Fibula
Anterior tibial vein
Posterior tibial artery
Posterior tibial vein
Peroneal artery

Medial inferior genicular artery Genicular arteries branch from the popliteal artery and form an anastomosis (network) around the knee

Anterior tibial artery Passes forward above the interosseous membrane to supply the muscles of the shin

Great (long) saphenous vein This, and the small saphenous vein, may become dilated, tortuous, and easily visible (varicose veins)

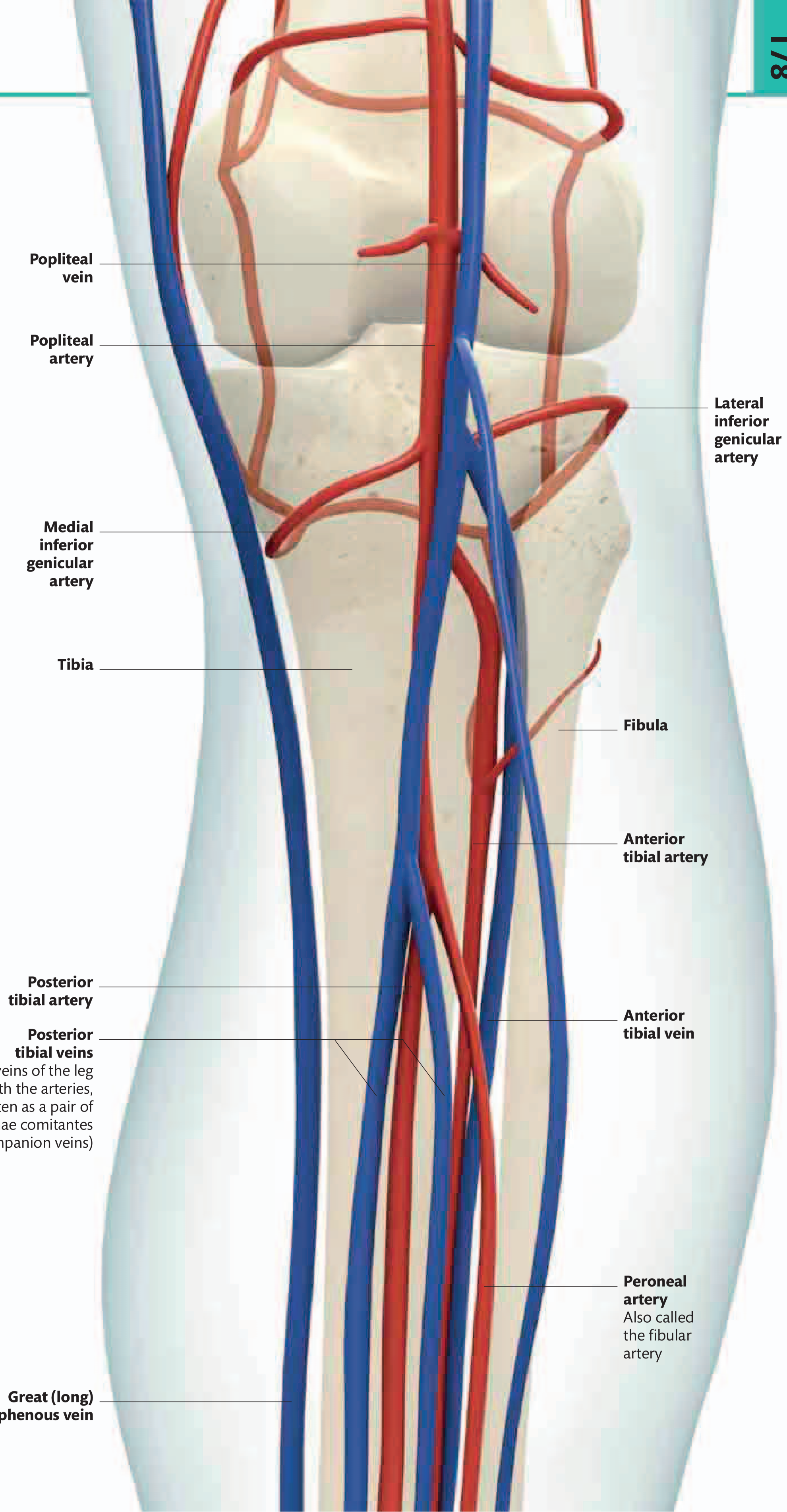
Popliteal vein
Popliteal artery
Lateral inferior genicular artery
Medial inferior genicular artery
Tibia
Fibula
Anterior tibial artery
Posterior tibial artery
Anterior tibial vein
Great (long) saphenous vein

Posterior tibial veins Deep veins of the leg run with the arteries, often as a pair of venae comitantes (companion veins)

Peroneal artery Also called the fibular artery

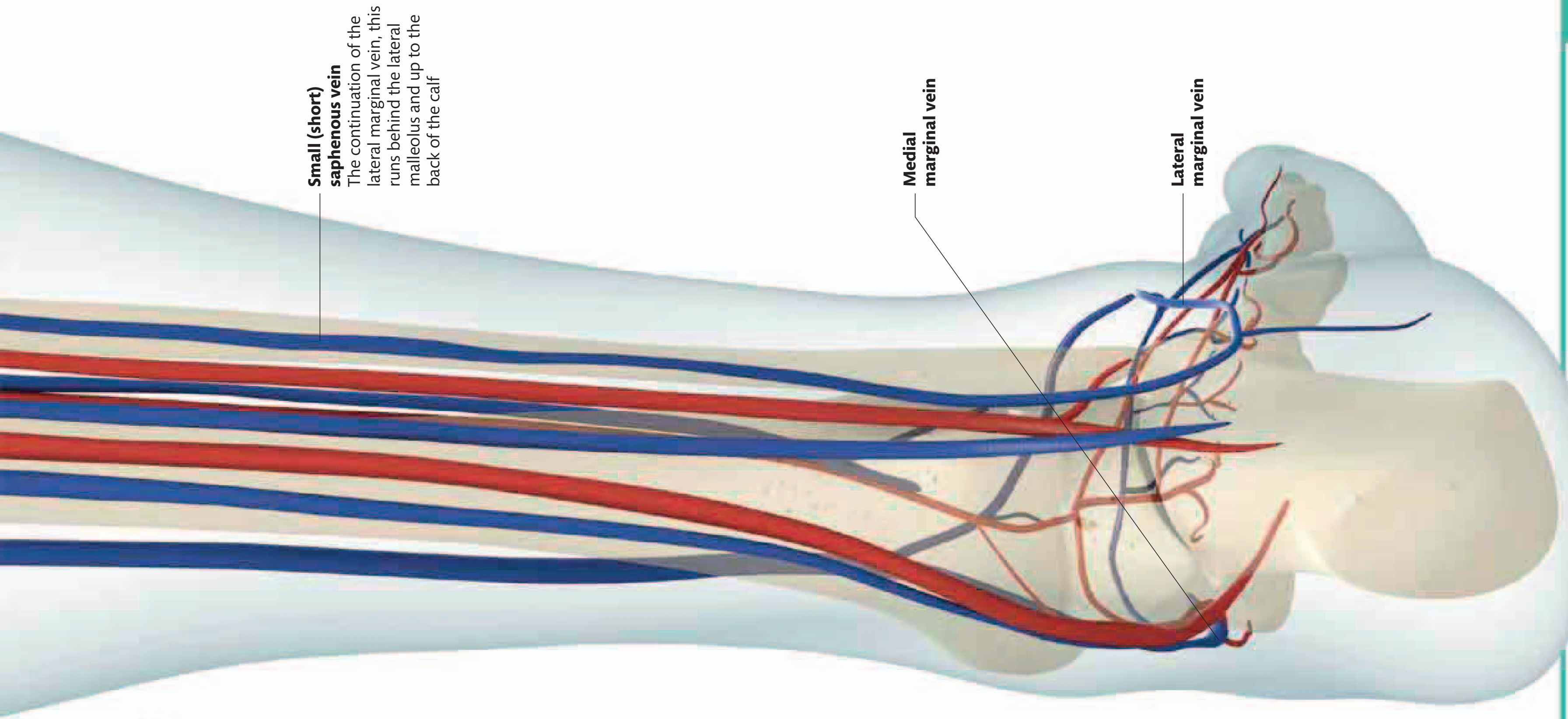
Small (short) saphenous vein The continuation of the lateral marginal vein, this runs behind the lateral malleolus and up to the back of the calf
Medial marginal vein
Lateral marginal vein

POSTERIOR (BACK)

LOWER LEG AND FOOT

The popliteal artery runs deep across the back of the knee, dividing into two branches: the anterior and posterior tibial arteries. The former runs forward, piercing the interosseous membrane between the tibia and fibula, to supply the extensor muscles of the shin. It runs down past the ankle, onto the top of the foot, as the dorsalis pedis artery. The latter gives off a peroneal branch, supplying the muscles and skin on the leg's outer side. The posterior tibial artery itself continues in the calf, running with the tibial nerve and, like the nerve, divides into plantar branches to supply the sole of the foot. A network of superficial veins on the back of the foot is drained by the saphenous veins.

ANTERIOR (FRONT)

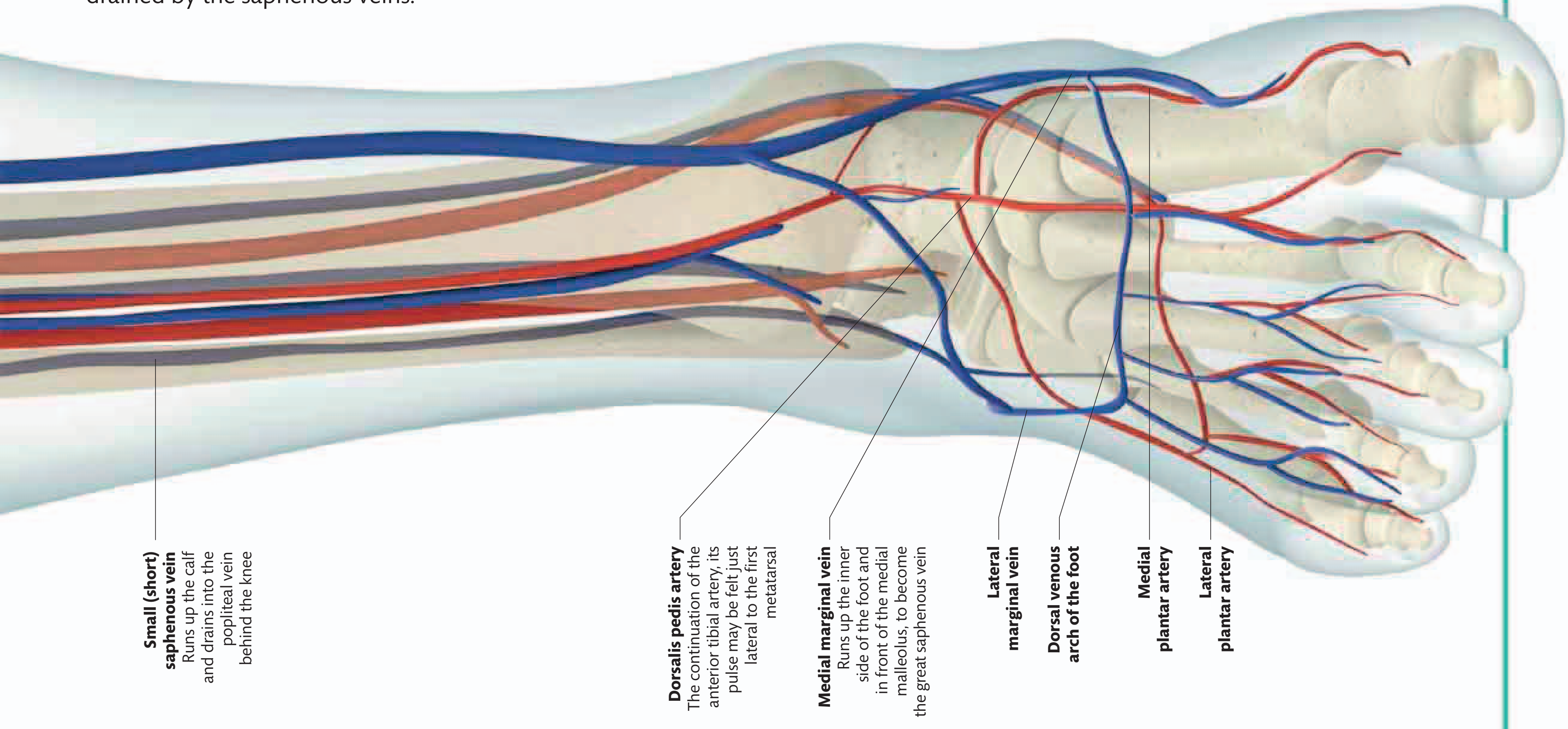
Small (short) saphenous vein Runs up the calf and drains into the popliteal vein behind the knee
Dorsalis pedis artery The continuation of the anterior tibial artery, its pulse may be felt just lateral to the first metatarsal
Medial marginal vein Runs up the inner side of the foot and in front of the medial malleolus, to become the great saphenous vein
Lateral marginal vein
Dorsal venous arch of the foot
Medial plantar artery
Lateral plantar artery

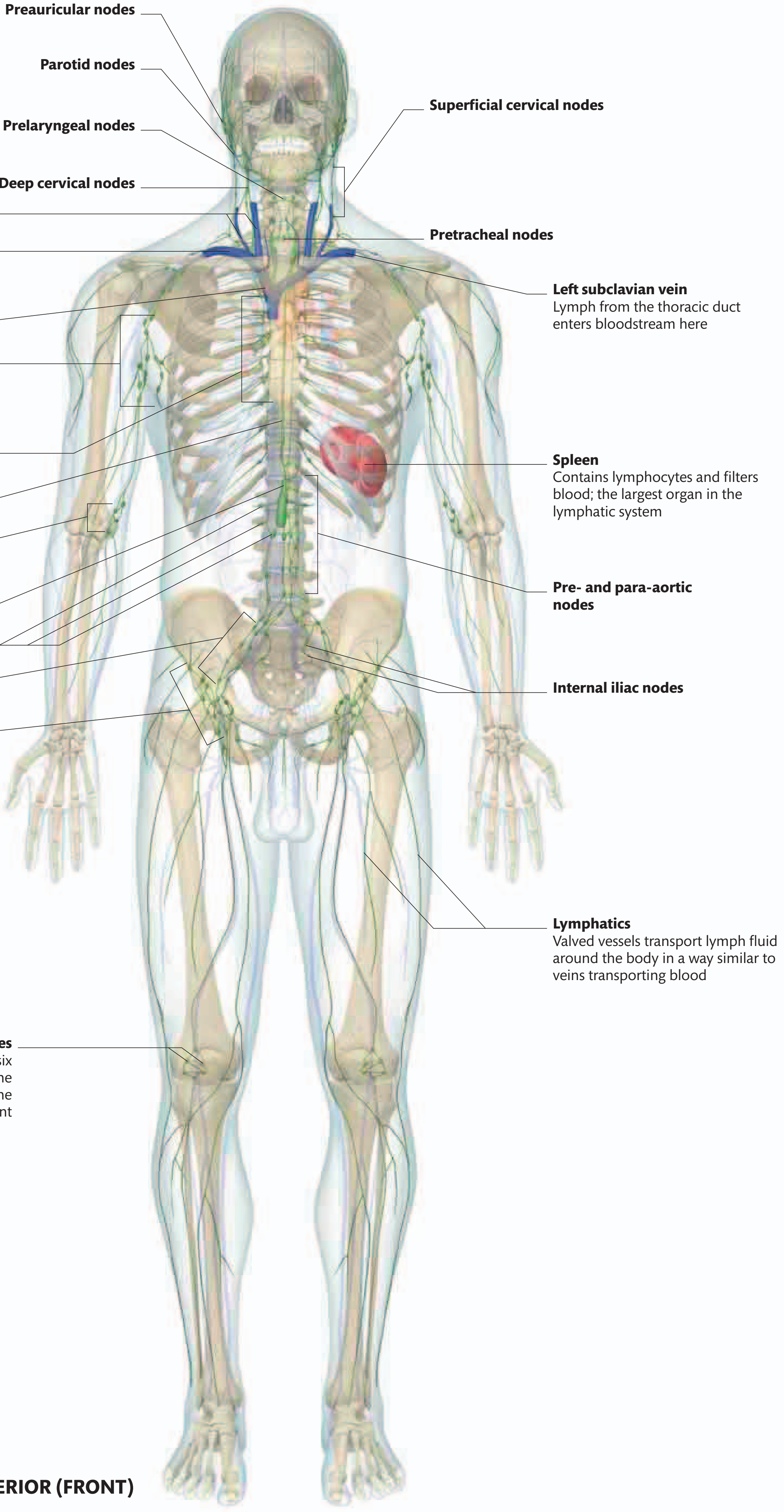
Preauricular nodes
Parotid nodes
Prelaryngeal nodes
Deep cervical nodes
Jugular veins
Right subclavian vein
Lymph from right arm, and right side of head and chest enters bloodstream here
Superior vena cava
Axillary nodes
Drain upper trunk and arm
Parasternal nodes
Thoracic duct
Supratrochlear nodes
Lymph from the hand and forearm drains to nodes at the elbow
Cisterna chyli
Lateral aortic nodes
External iliac nodes
Inguinal nodes
Popliteal nodes
A group of around six nodes sit within the popliteal fossa, at the back of the knee joint
Superficial cervical nodes
Pretracheal nodes
Left subclavian vein
Lymph from the thoracic duct enters bloodstream here
Spleen
Contains lymphocytes and filters blood; the largest organ in the lymphatic system
Pre- and para-aortic nodes
Internal iliac nodes
Lymphatics
Valved vessels transport lymph fluid around the body in a way similar to veins transporting blood

ANTERIOR (FRONT)

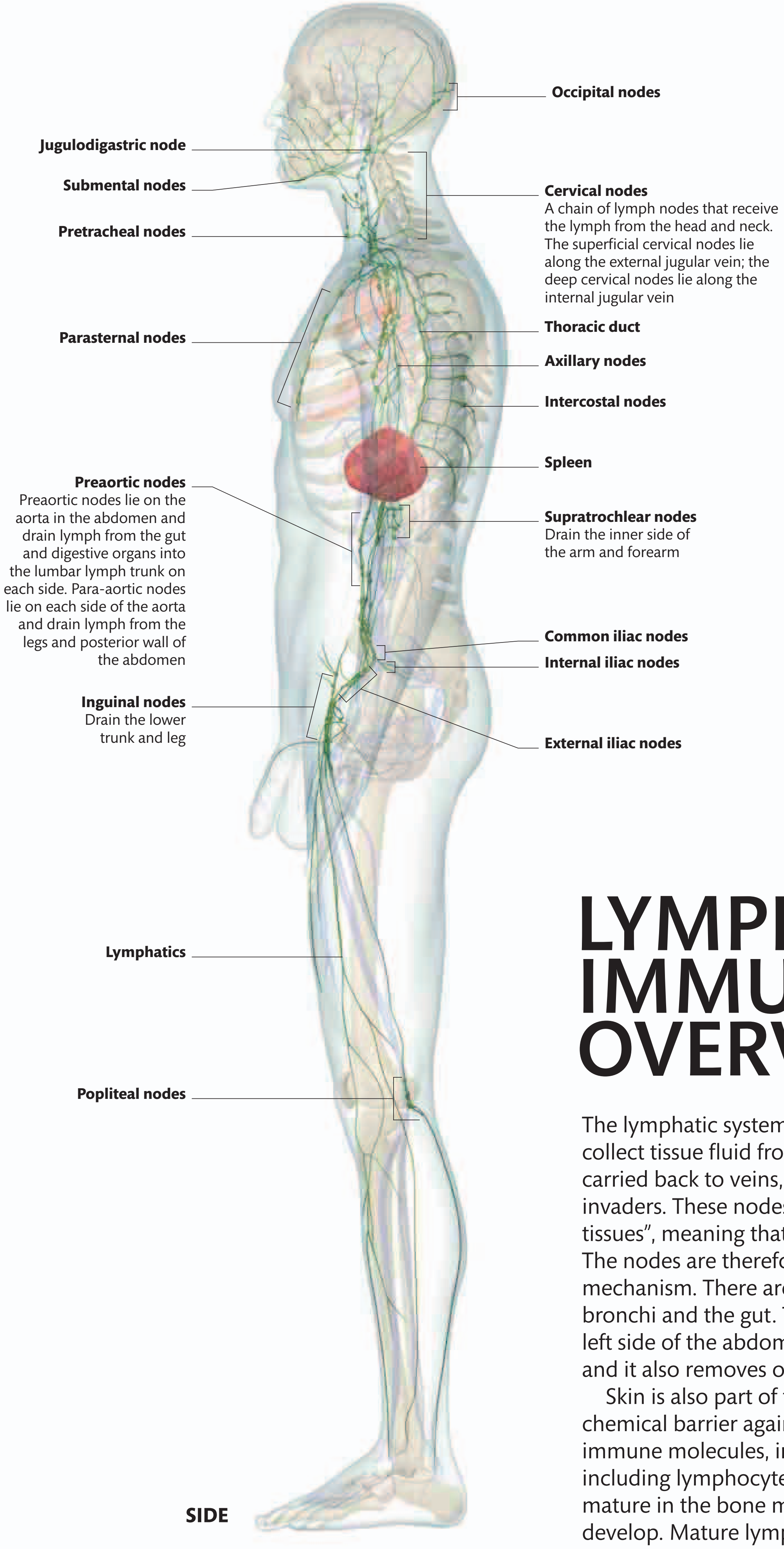
Occipital nodes
Jugulodigastric node
Submental nodes
Pretracheal nodes
Thoracic duct
Parasternal nodes
Axillary nodes
Intercostal nodes
Spleen
Common iliac nodes
Internal iliac nodes
External iliac nodes
Lymphatics
Popliteal nodes
SIDE

Cervical nodes
A chain of lymph nodes that receive the lymph from the head and neck. The superficial cervical nodes lie along the external jugular vein; the deep cervical nodes lie along the internal jugular vein

Preaortic nodes
Preaortic nodes lie on the aorta in the abdomen and drain lymph from the gut and digestive organs into the lumbar lymph trunk on each side. Para-aortic nodes lie on each side of the aorta and drain lymph from the legs and posterior wall of the abdomen

Supratrochlear nodes
Drain the inner side of the arm and forearm

Inguinal nodes
Drain the lower trunk and leg

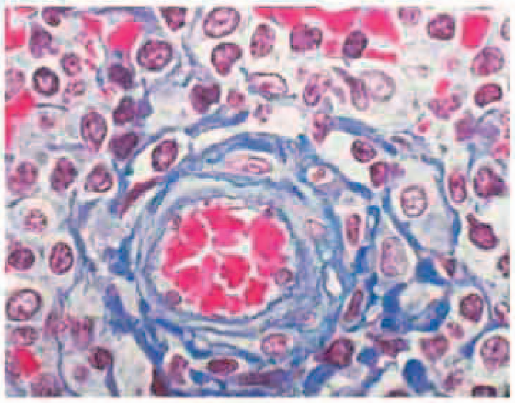

Lymphoid tissue
At a high magnification, individual lymphocytes (purple) can be seen in a section of lymphoid tissue. The blue circle in the image is an arteriole, packed full of blood cells (stained pink).

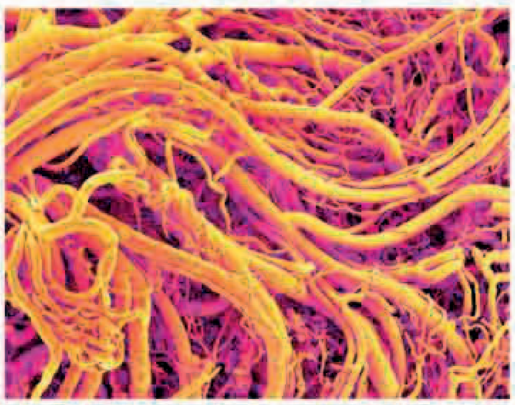

Blood vessels of lymph node
This image, produced using a scanning electron microscope, shows a resin cast of the dense network of tiny blood vessels inside a lymph node.

LYMPHATIC AND IMMUNE SYSTEM OVERVIEW

The lymphatic system consists of a network of lymphatic vessels that collect tissue fluid from the spaces between cells. Before this fluid is carried back to veins, it is delivered to lymph nodes to check for potential invaders. These nodes, like the tonsils, spleen, and thymus, are "lymphoid tissues", meaning that they contain immune cells known as lymphocytes. The nodes are therefore part of the immune system, the body's defense mechanism. There are also patches of lymphoid tissue in the walls of the bronchi and the gut. The spleen, which lies tucked under the ribs on the left side of the abdomen, has two important roles: it is a lymphoid organ, and it also removes old red blood cells from circulation.

Skin is also part of the immune system as it forms a physical and chemical barrier against infections. The formation of some important immune molecules, including antibodies, and a range of immune cells, including lymphocytes, happens in the bone marrow. Some lymphocytes mature in the bone marrow, whereas others move to the thymus to develop. Mature lymphocytes stay in the lymph nodes, where they perform their function.

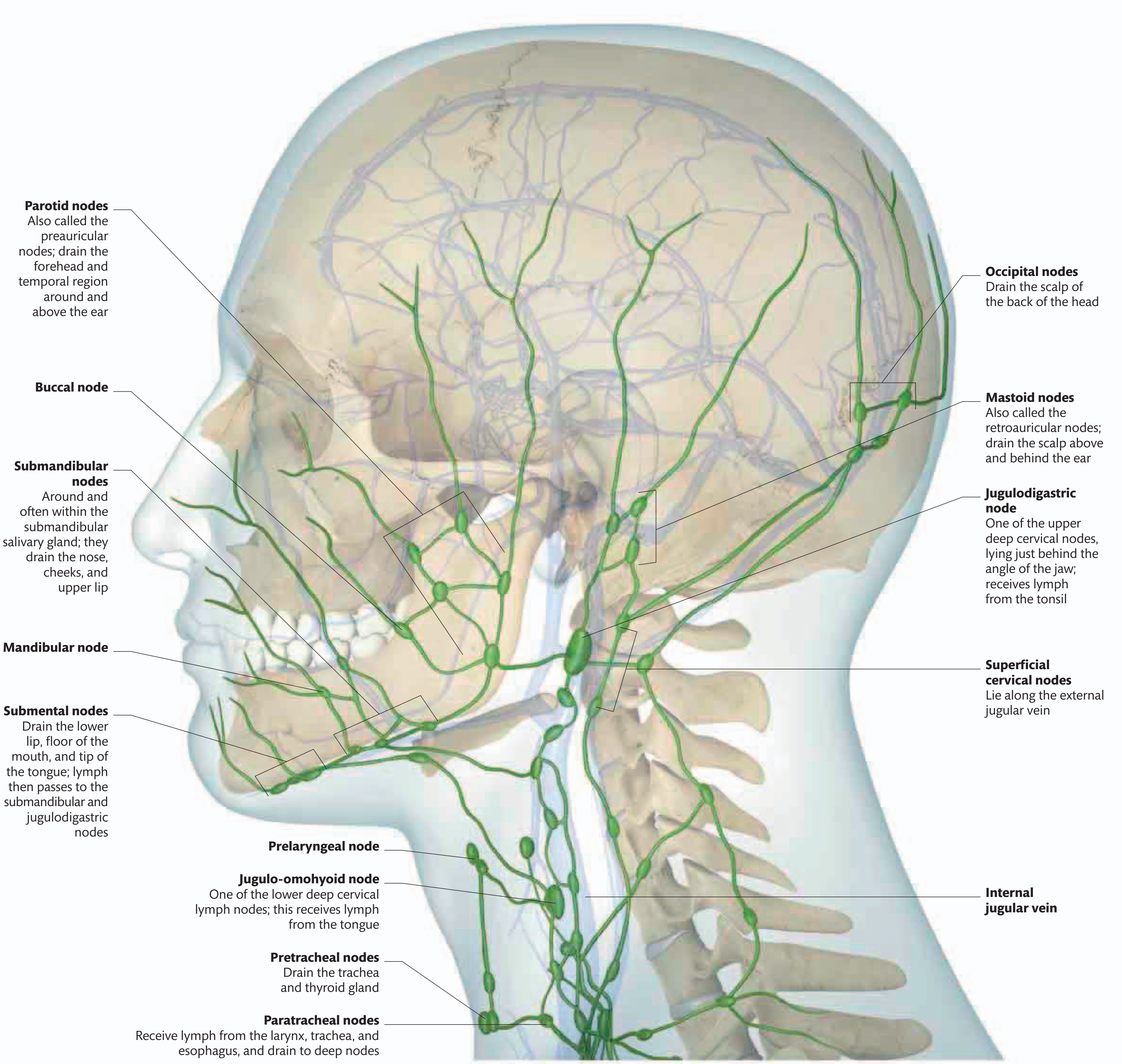
Parotid nodes
Also called the preauricular nodes; drain the forehead and temporal region around and above the ear
Buccal node
Submandibular nodes
Around and often within the submandibular salivary gland; they drain the nose, cheeks, and upper lip
Mandibular node
Submental nodes
Drain the lower lip, floor of the mouth, and tip of the tongue; lymph then passes to the submandibular and jugulodigastric nodes
Prelaryngeal node
Jugulo-omohyoid node
One of the lower deep cervical lymph nodes; this receives lymph from the tongue
Pretracheal nodes
Drain the trachea and thyroid gland
Paratracheal nodes
Receive lymph from the larynx, trachea, and esophagus, and drain to deep nodes
Occipital nodes
Drain the scalp of the back of the head
Mastoid nodes
Also called the retroauricular nodes; drain the scalp above and behind the ear
Jugulodigastric node
One of the upper deep cervical nodes, lying just behind the angle of the jaw; receives lymph from the tonsil
Superficial cervical nodes
Lie along the external jugular vein
Internal jugular vein

LYMPH NODES OF HEAD

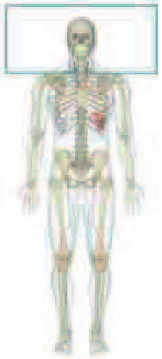

HEAD AND NECK

A ring of lymph nodes lies close to the skin where the head meets the neck, from the occipital nodes (against the skull at the back) to the submandibular and submental nodes (which are tucked under the jaw). Superficial nodes lie along the sides and front of the neck, and deep nodes are clustered around the internal jugular vein, under cover of sternocleidomastoid muscle. Lymph from all other nodes passes to these deep ones, then into the jugular lymphatic trunk before draining back into veins in the base of the neck. Lymphoid tissue, in the form of the palatine, pharyngeal, and lingual tonsils, forms a protective ring around the upper parts of the respiratory and digestive tracts.

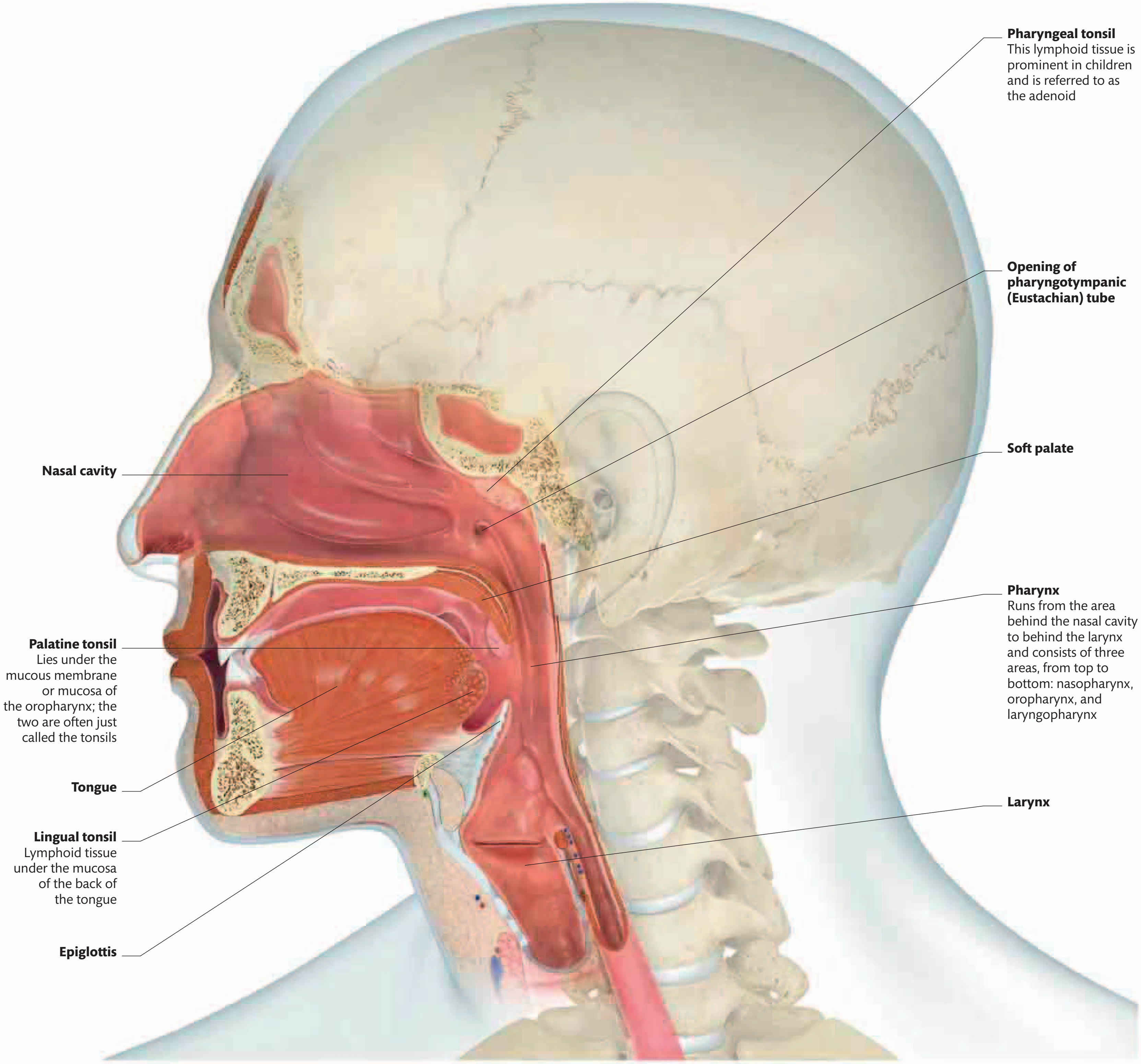
Nasal cavity
Palatine tonsil
Lies under the mucous membrane or mucosa of the oropharynx; the two are often just called the tonsils
Tongue
Lingual tonsil
Lymphoid tissue under the mucosa of the back of the tongue
Epiglottis
Pharyngeal tonsil
This lymphoid tissue is prominent in children and is referred to as the adenoid
Opening of pharyngotympanic (Eustachian) tube
Soft palate
Pharynx
Runs from the area behind the nasal cavity to behind the larynx and consists of three areas, from top to bottom: nasopharynx, oropharynx, and laryngopharynx
Larynx

LOCATION OF TONSILS

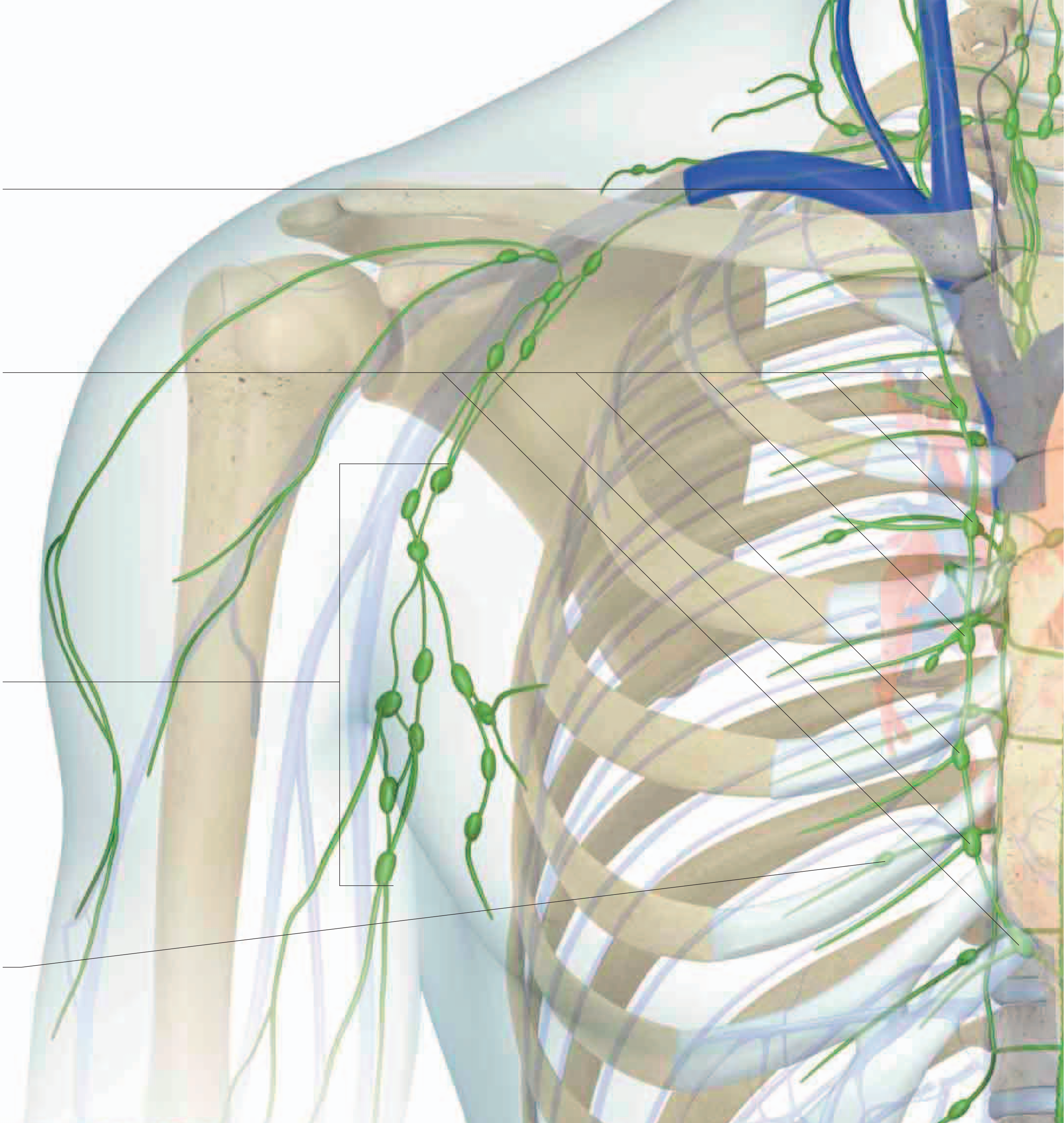

Right lymphatic duct
Lymph from the right arm and the right side of the neck and thorax drains into the junction of the right internal jugular and subclavian veins

Parasternal nodes
Also called internal thoracic nodes; these lie in the gaps between the ribs, either side of the sternum on the inside of the ribcage; they drain some of the lymph from the front of the thorax – including from the breast in a woman

Axillary nodes
Receive lymph from superficial tissues of the thorax, upper limb, and breast

Intercostal nodes
Sitting in the intercostal spaces between the ribs at the back of the ribcage, these drain lymph from the deeper tissues at the sides and back of the thorax

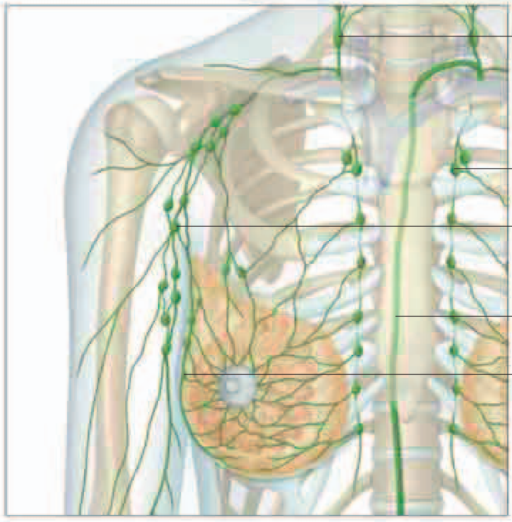
Supraclavicular nodes
Parasternal nodes
Axillary nodes
Thoracic duct
Paramammary node

ANTERIOR (FRONT) / FEMALE

ANTERIOR (FRONT) / MALE

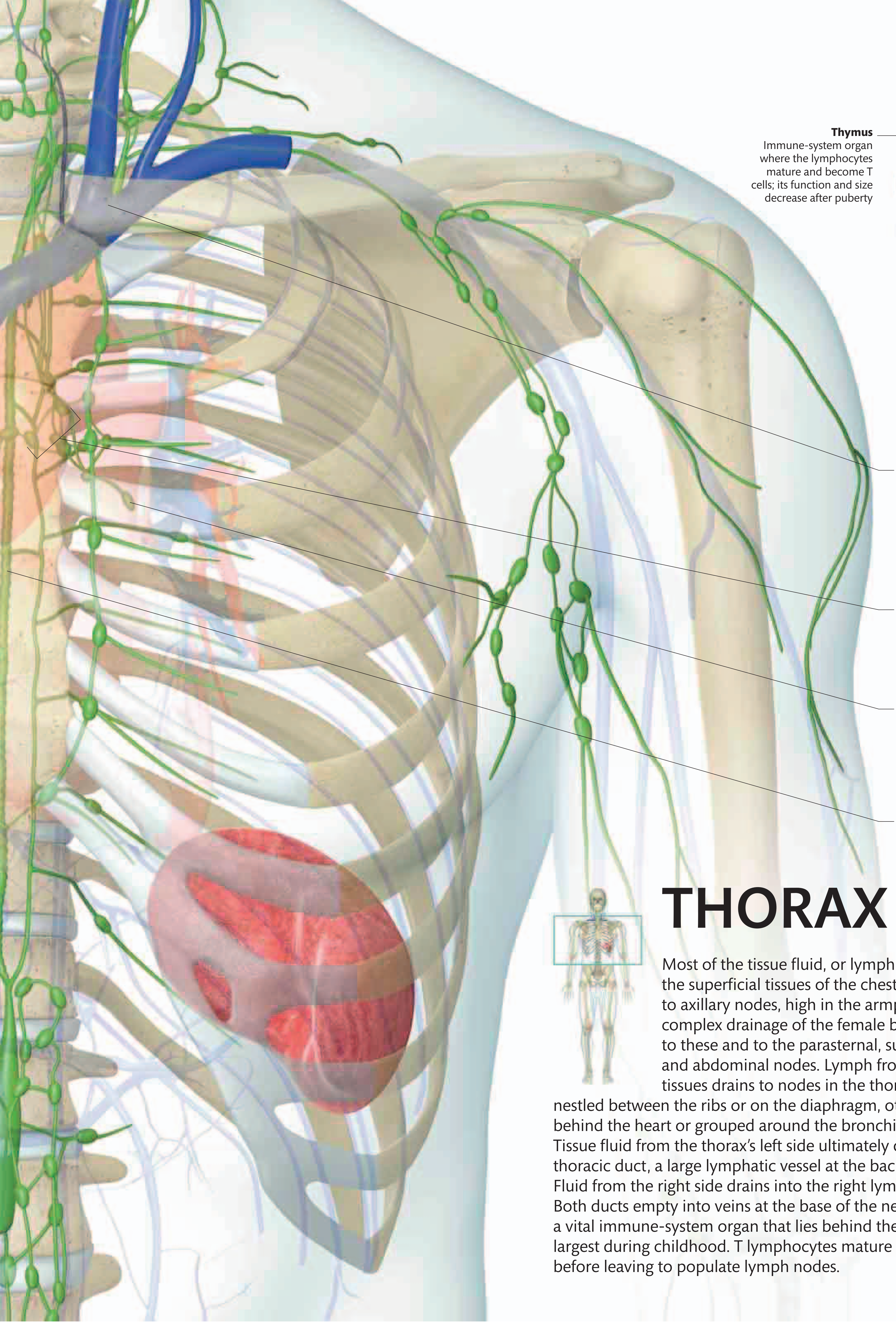

Termination of thoracic duct
The thoracic duct ends by draining into the point where the left internal jugular and subclavian veins join at the bottom of the neck

Tracheobronchial nodes
Clustered around the lower trachea and bronchi, these nodes drain the lungs

Posterior mediastinal node
Tucked behind the heart, these receive lymph from the heart, esophagus, and diaphragm

Thoracic duct
Lies right at the back of the thorax, against the spine

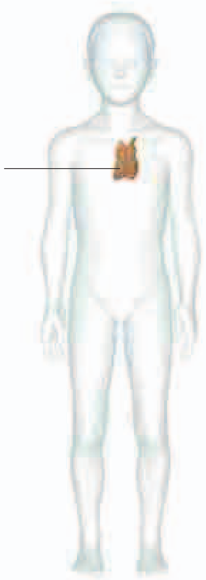

Thymus
Immune-system organ where the lymphocytes mature and become T cells; its function and size decrease after puberty

THYMUS IN CHILD

THORAX

Most of the tissue fluid, or lymph, from the superficial tissues of the chest drains to axillary nodes, high in the armpits. The complex drainage of the female breast passes to these and to the parasternal, supraclavicular, and abdominal nodes. Lymph from deeper tissues drains to nodes in the thorax, some nestled between the ribs or on the diaphragm, others tucked behind the heart or grouped around the bronchi and trachea. Tissue fluid from the thorax's left side ultimately drains into the thoracic duct, a large lymphatic vessel at the back of the thorax. Fluid from the right side drains into the right lymphatic duct. Both ducts empty into veins at the base of the neck. The thymus, a vital immune-system organ that lies behind the sternum, is largest during childhood. T lymphocytes mature in the thymus before leaving to populate lymph nodes.

ABDOMEN AND PELVIS

The deep lymph nodes of the abdomen are clustered around arteries. Nodes lying along each side of the aorta receive lymph from paired structures, such as the muscles of the abdominal wall, the kidneys and adrenal glands, and the testes or ovaries. Iliac nodes collect lymph returning from the legs and pelvis. Nodes clustered around the branches on the front of the aorta collect lymph from the gut and abdominal organs. Eventually, all this lymph from the legs, pelvis, and abdomen passes into a swollen lymphatic vessel called the cisterna chyli; this narrows down to become the thoracic duct, which runs up into the chest. Most lymph nodes are small, bean-sized structures, but the abdomen also contains a large and important organ of the immune system—the spleen.

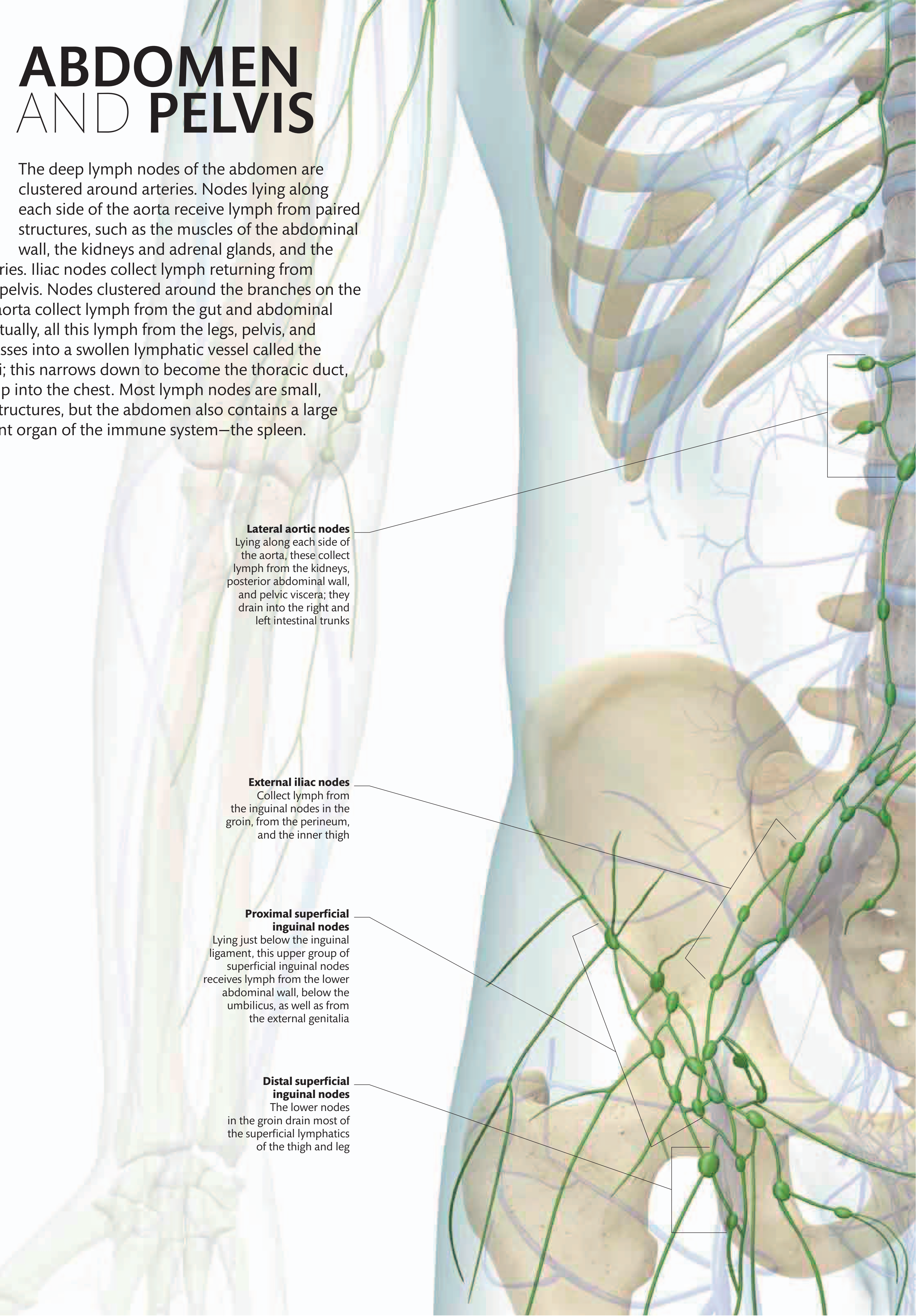

Lateral aortic nodes
Lying along each side of the aorta, these collect lymph from the kidneys, posterior abdominal wall, and pelvic viscera; they drain into the right and left intestinal trunks

External iliac nodes
Collect lymph from the inguinal nodes in the groin, from the perineum, and the inner thigh

Proximal superficial inguinal nodes
Lying just below the inguinal ligament, this upper group of superficial inguinal nodes receives lymph from the lower abdominal wall, below the umbilicus, as well as from the external genitalia

Distal superficial inguinal nodes
The lower nodes in the groin drain most of the superficial lymphatics of the thigh and leg

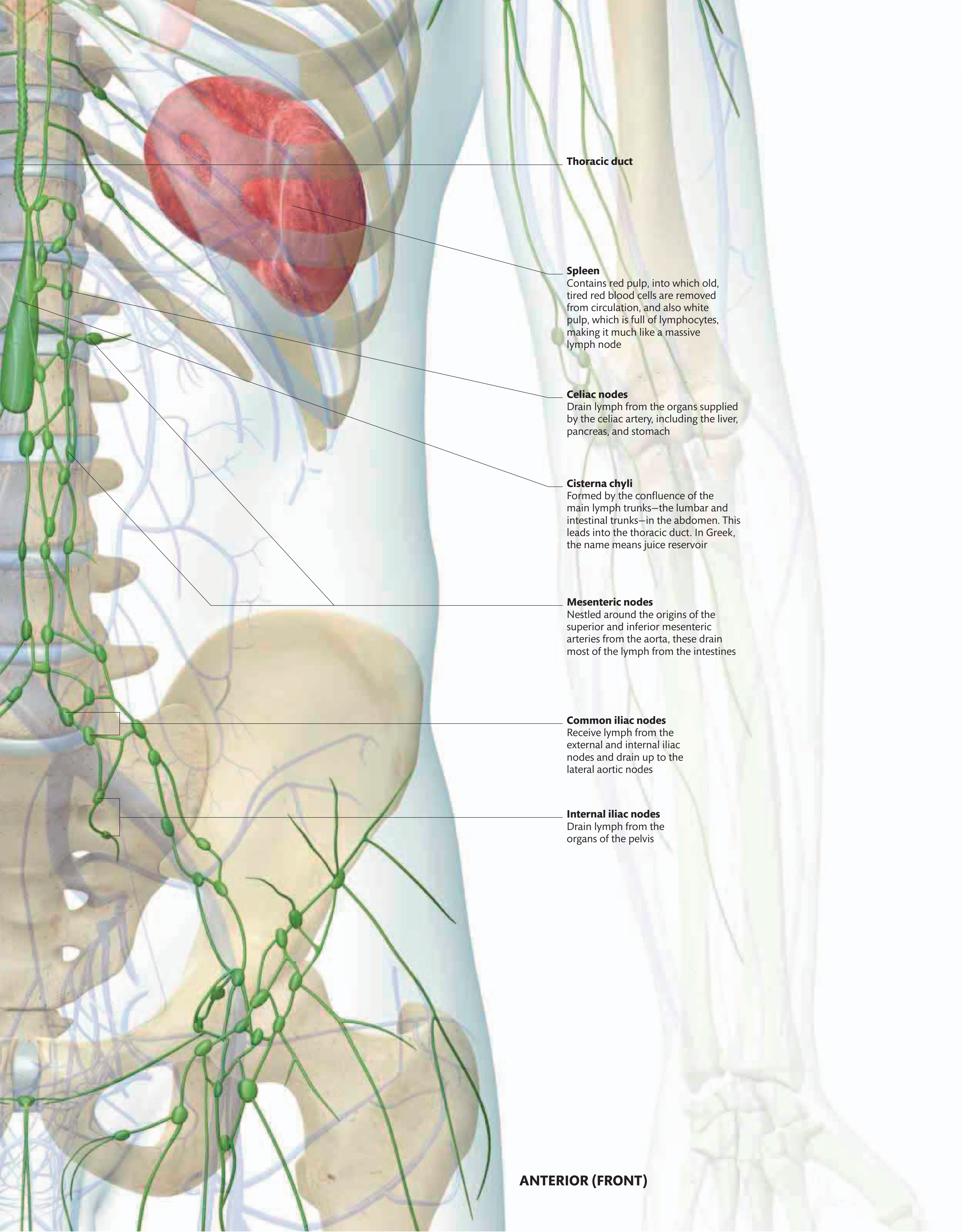

Thoracic duct

Spleen
Contains red pulp, into which old, tired red blood cells are removed from circulation, and also white pulp, which is full of lymphocytes, making it much like a massive lymph node

Celiac nodes
Drain lymph from the organs supplied by the celiac artery, including the liver, pancreas, and stomach

Cisterna chyli
Formed by the confluence of the main lymph trunks—the lumbar and intestinal trunks—in the abdomen. This leads into the thoracic duct. In Greek, the name means juice reservoir

Mesenteric nodes
Nestled around the origins of the superior and inferior mesenteric arteries from the aorta, these drain most of the lymph from the intestines

Common iliac nodes
Receive lymph from the external and internal iliac nodes and drain up to the lateral aortic nodes

Internal iliac nodes
Drain lymph from the organs of the pelvis

ANTERIOR (FRONT)

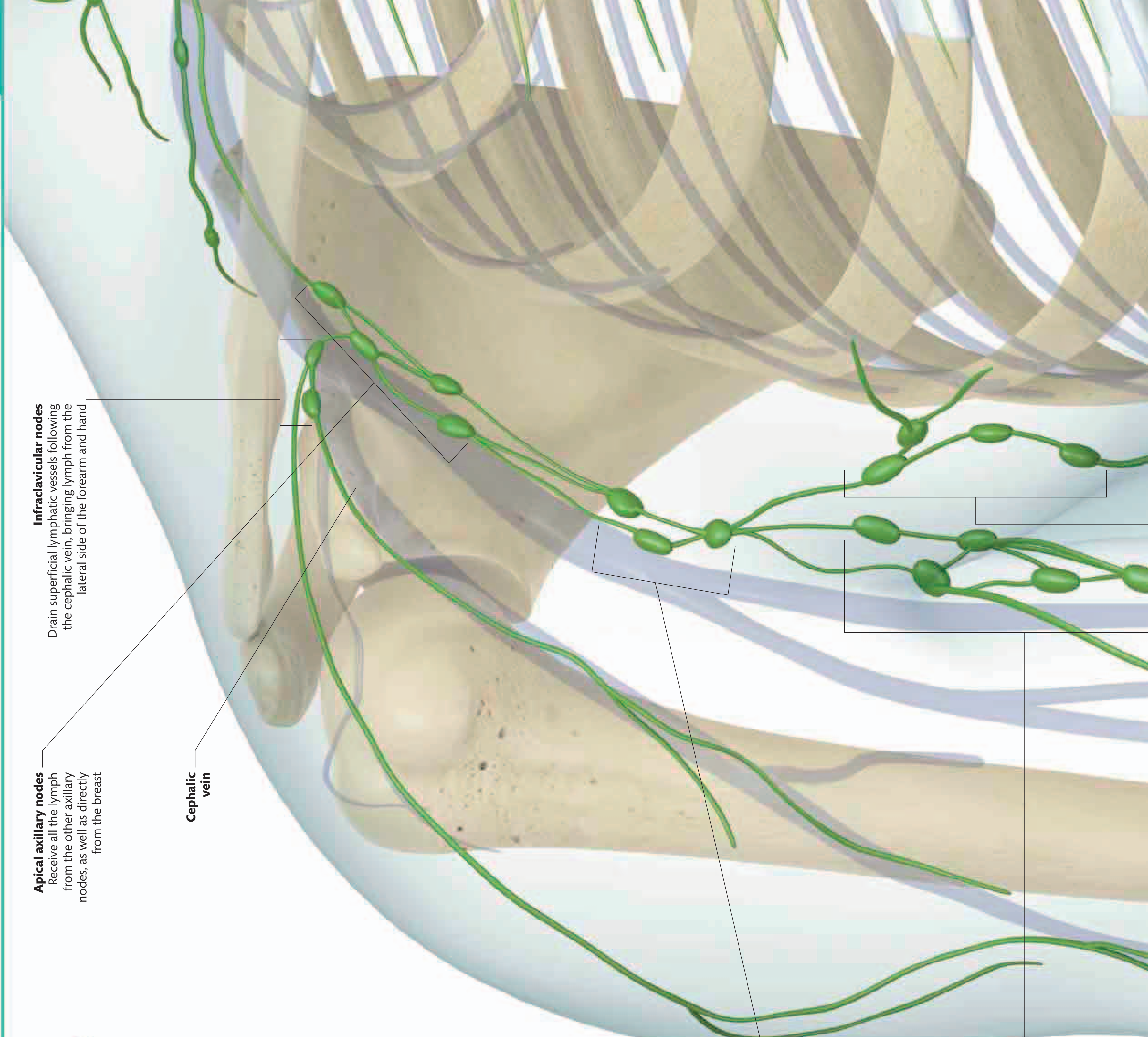
Infraclavicular nodes
Drain superficial lymphatic vessels following the cephalic vein, bringing lymph from the lateral side of the forearm and hand
Apical axillary nodes
Receive all the lymph from the other axillary nodes, as well as directly from the breast
Cephalic vein
Central axillary nodes
Receive lymph from the anterior and lateral axillary nodes; also from the posterior axillary nodes, which drain the back of the neck and trunk
Lateral axillary nodes
Receive deep and superficial lymphatics from most of the upper limb, apart from those following the cephalic vein

SHOULDER AND UPPER ARM

Ultimately, all the lymph from the hand, forearm, and arm drains to the axillary nodes in the armpit. But there are a few nodes, lower in the arm, that lymph may pass through on its way to the axilla. The supratrochlear nodes lie in the subcutaneous fat on the inner arm, above the elbow. They collect lymph that has drained from the medial side of the hand and forearm. The infraclavicular nodes, lying along the cephalic vein, below the clavicle, receive lymphatics draining from the thumb and the lateral side of the forearm and arm. Axillary nodes drain lymph from the arm and receive it from the chest wall. They may become infiltrated with cancerous cells spreading from a tumor in the breast.

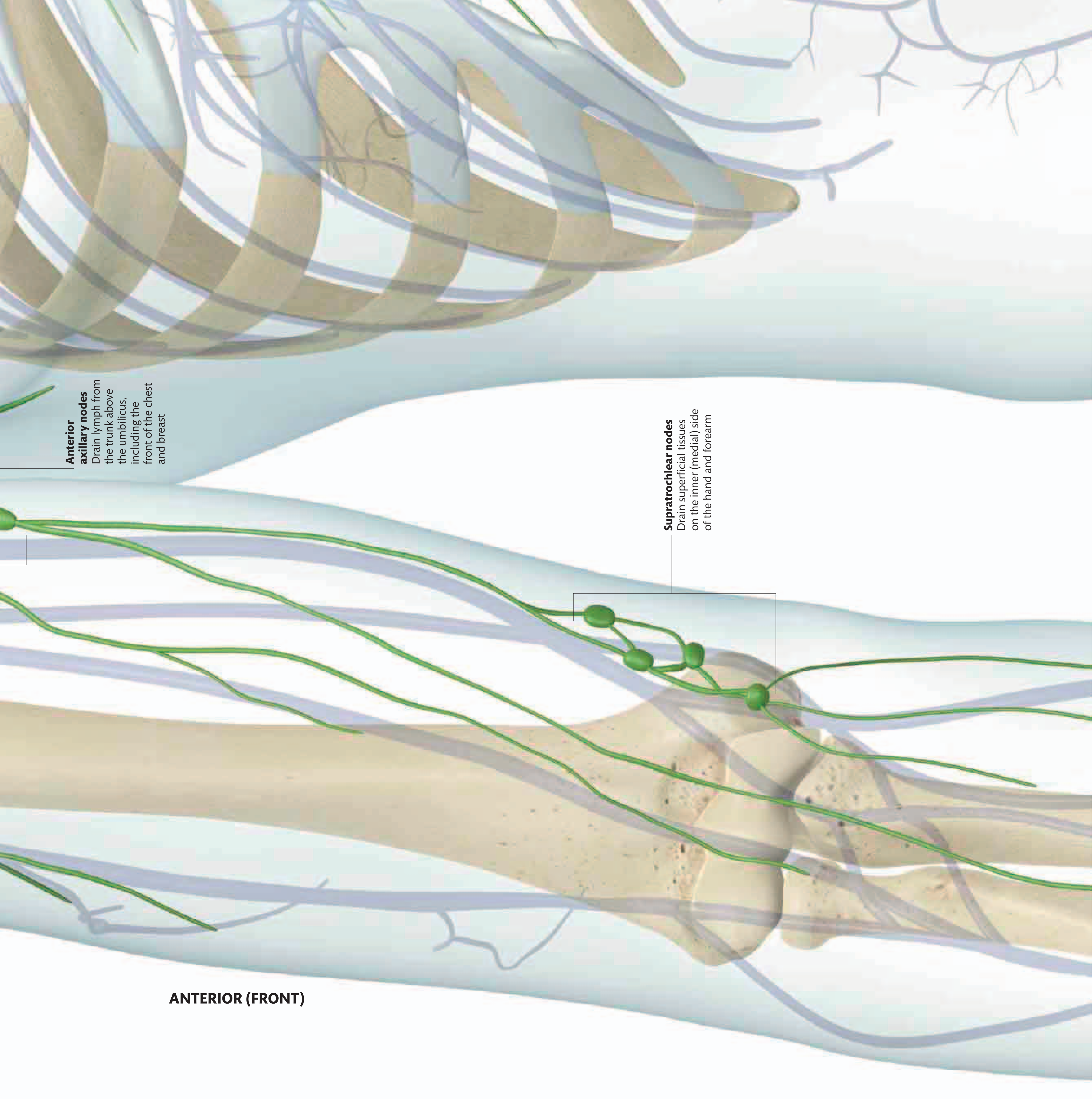

Anterior axillary nodes Drain lymph from the trunk above the umbilicus, including the front of the chest and breast

Supratrochlear nodes Drain superficial tissues on the inner (medial) side of the hand and forearm

ANTERIOR (FRONT)

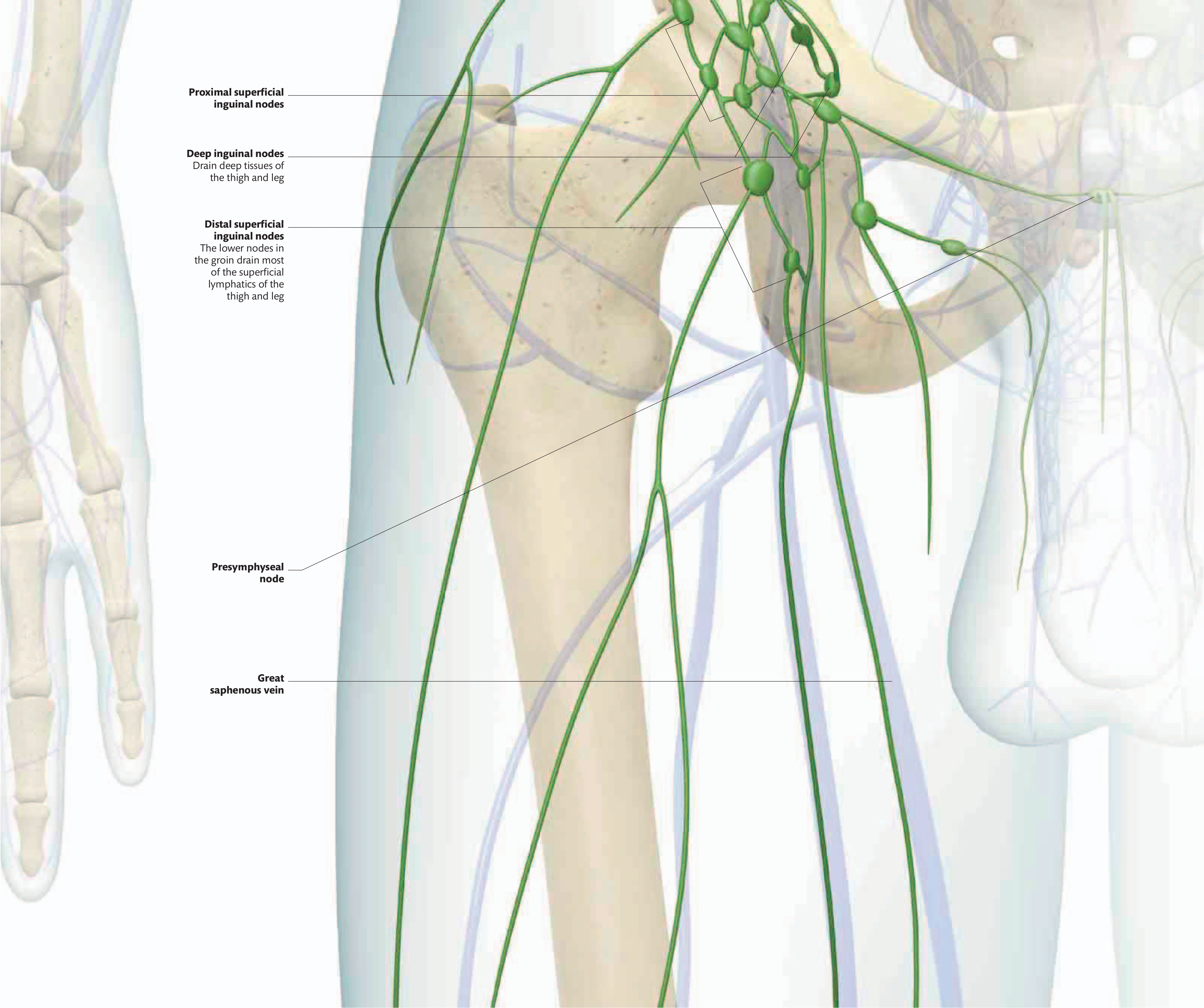

Proximal superficial inguinal nodes

Deep inguinal nodes Drain deep tissues of the thigh and leg

Distal superficial inguinal nodes The lower nodes in the groin drain most of the superficial lymphatics of the thigh and leg

Presymphyseal node

Great saphenous vein

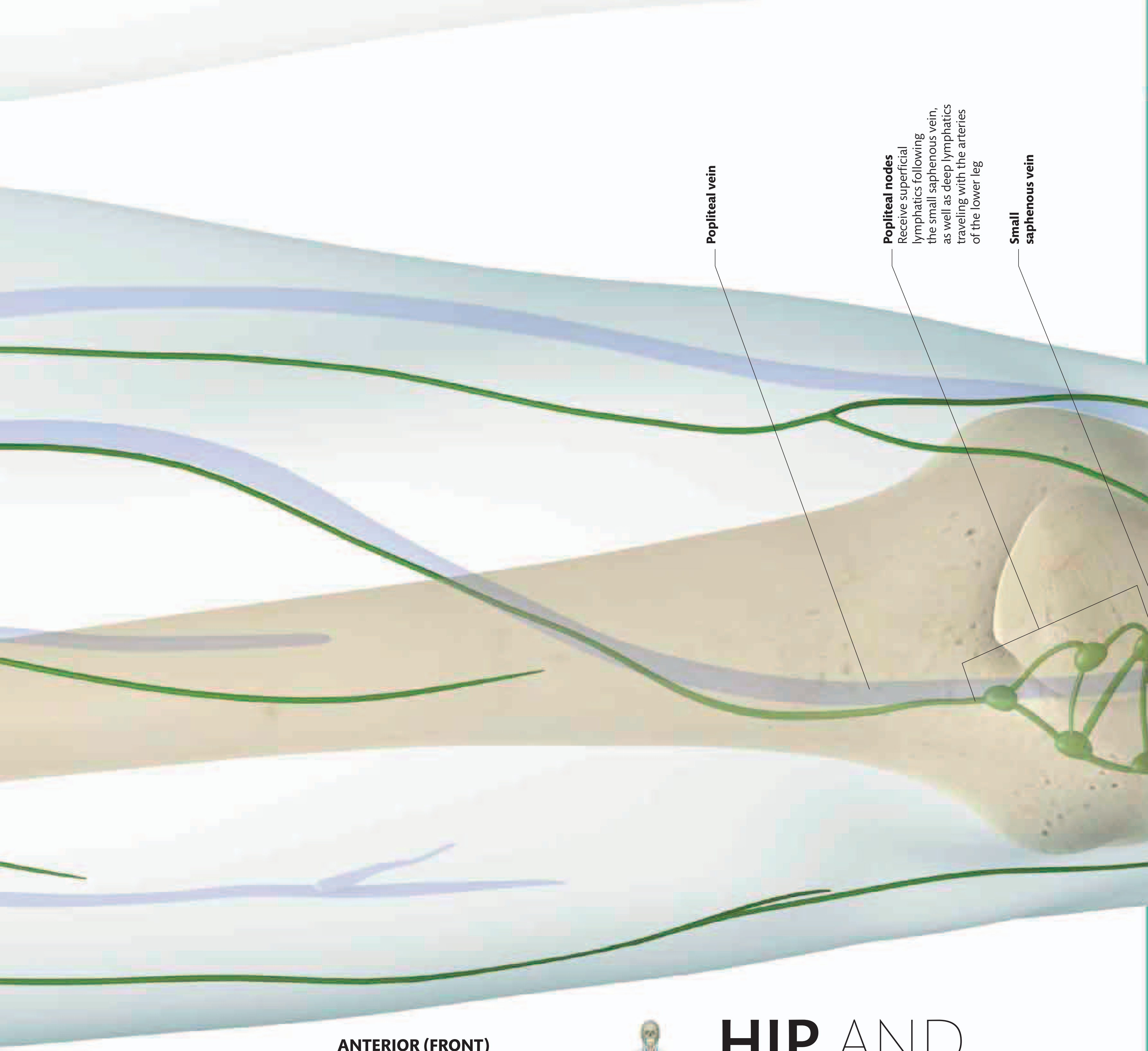
Popliteal vein
Popliteal nodes
Receive superficial lymphatics following the small saphenous vein, as well as deep lymphatics traveling with the arteries of the lower leg
Small saphenous vein

ANTERIOR (FRONT)

HIP AND THIGH

Most lymph from the thigh, leg, and foot passes through the inguinal group of lymph nodes, which are in the groin. But lymph from the deep tissues of the buttock passes straight to nodes inside the pelvis (see pp.184–85), along the internal and common iliac arteries. Eventually, all the lymph from the leg reaches the lateral aortic nodes, on the back wall of the abdomen. As in the arm, there are groups of nodes clustered around points at which superficial veins drain into deep veins. Popliteal nodes are close to the drainage of the small saphenous vein into the popliteal vein, while the superficial inguinal nodes lie close to the great saphenous vein, just before it empties into the femoral vein.

DIGESTIVE SYSTEM OVERVIEW

The digestive system comprises the organs that enable us to take in food, break it down physically and chemically, extract its useful nutrients, and excrete what we don't need. This process begins in the mouth, where the teeth, tongue, and saliva work together to form food into a moist ball that can be swallowed. The mouth, pharynx, stomach, intestines, rectum, and anal canal form a long tube that is referred to as the digestive tract. It usually takes between one and two days for ingested food to travel all the way from the mouth to the anus. Other organs—including the salivary glands, liver, gallbladder, and pancreas—complete the digestive system.

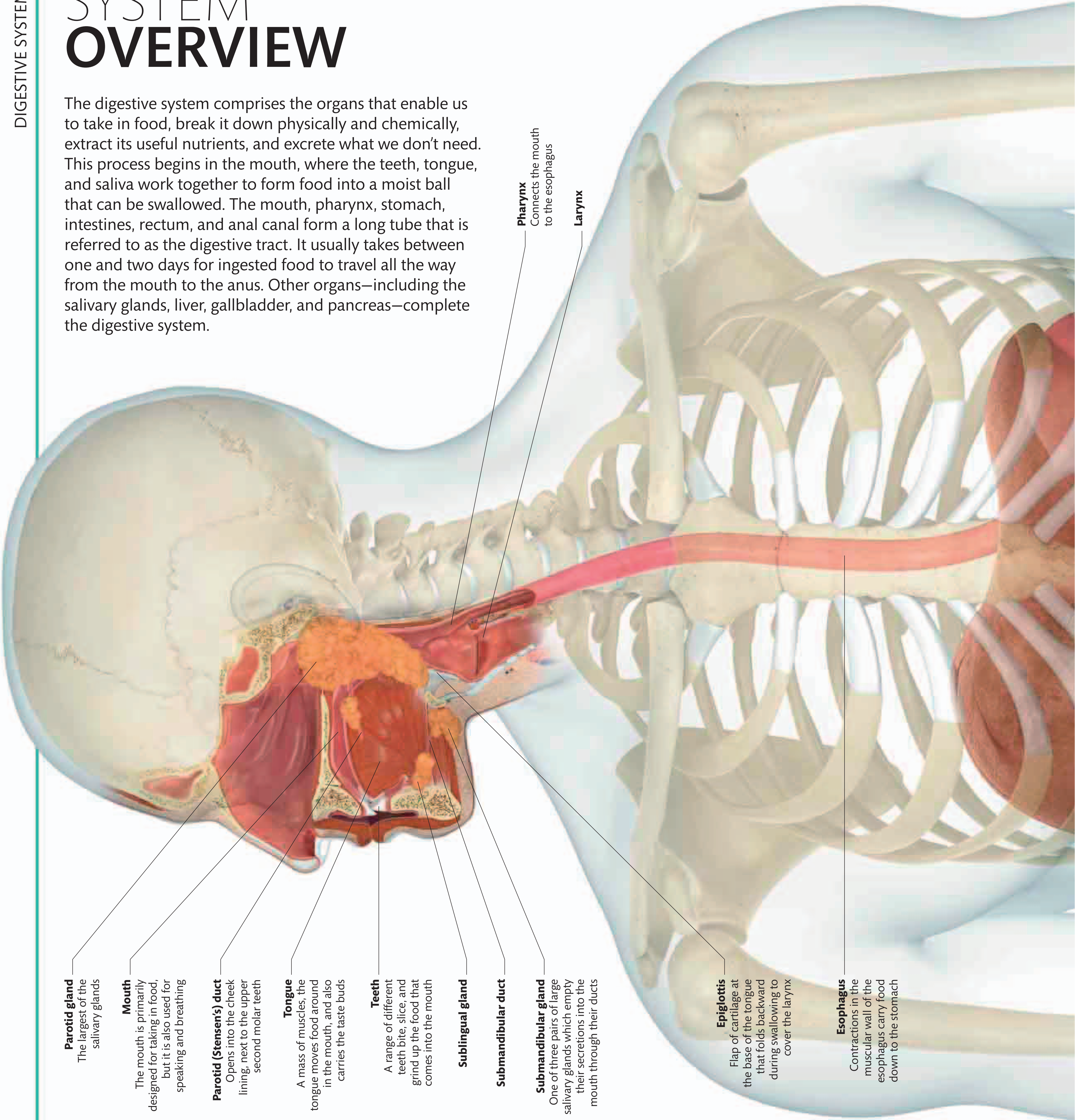
Pharynx Connects the mouth to the esophagus
Larynx
Parotid gland The largest of the salivary glands
Mouth The mouth is primarily designed for taking in food, but it is also used for speaking and breathing
Parotid (Stensen's) duct Opens into the cheek lining, next to the upper second molar teeth
Tongue A mass of muscles, the tongue moves food around in the mouth, and also carries the taste buds
Teeth A range of different teeth bite, slice, and grind up the food that comes into the mouth
Sublingual gland
Submandibular duct
Submandibular gland One of three pairs of large salivary glands which empty their secretions into the mouth through their ducts
Epiglottis Flap of cartilage at the base of the tongue that folds backward during swallowing to cover the larynx
Esophagus Contractions in the muscular wall of the esophagus carry food down to the stomach

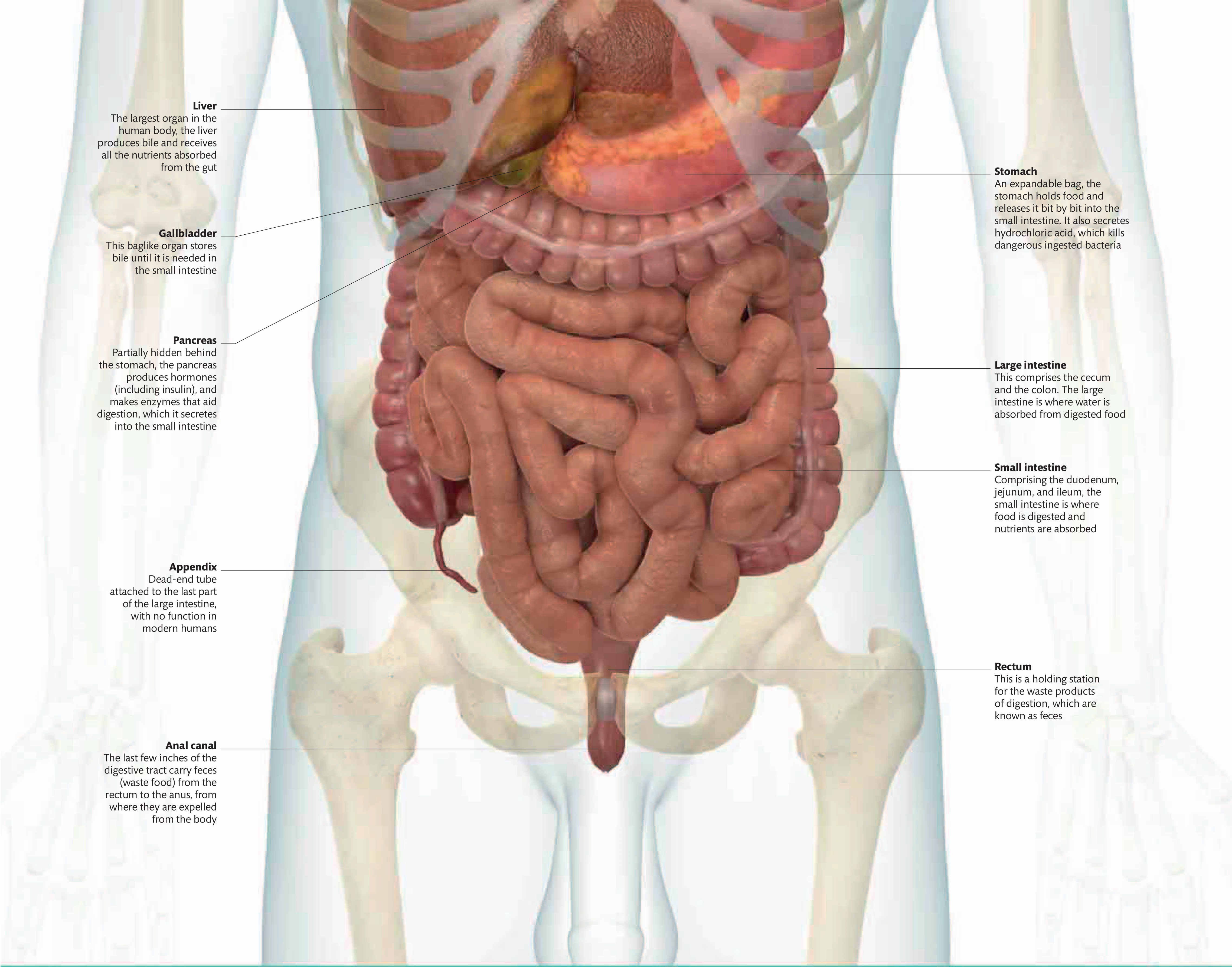

Liver
The largest organ in the human body, the liver produces bile and receives all the nutrients absorbed from the gut

Gallbladder
This baglike organ stores bile until it is needed in the small intestine

Pancreas
Partially hidden behind the stomach, the pancreas produces hormones (including insulin), and makes enzymes that aid digestion, which it secretes into the small intestine

Appendix
Dead-end tube attached to the last part of the large intestine, with no function in modern humans

Anal canal
The last few inches of the digestive tract carry feces (waste food) from the rectum to the anus, from where they are expelled from the body

Stomach
An expandable bag, the stomach holds food and releases it bit by bit into the small intestine. It also secretes hydrochloric acid, which kills dangerous ingested bacteria

Large intestine
This comprises the cecum and the colon. The large intestine is where water is absorbed from digested food

Small intestine
Comprising the duodenum, jejunum, and ileum, the small intestine is where food is digested and nutrients are absorbed

Rectum
This is a holding station for the waste products of digestion, which are known as feces

SAGITTAL SECTION

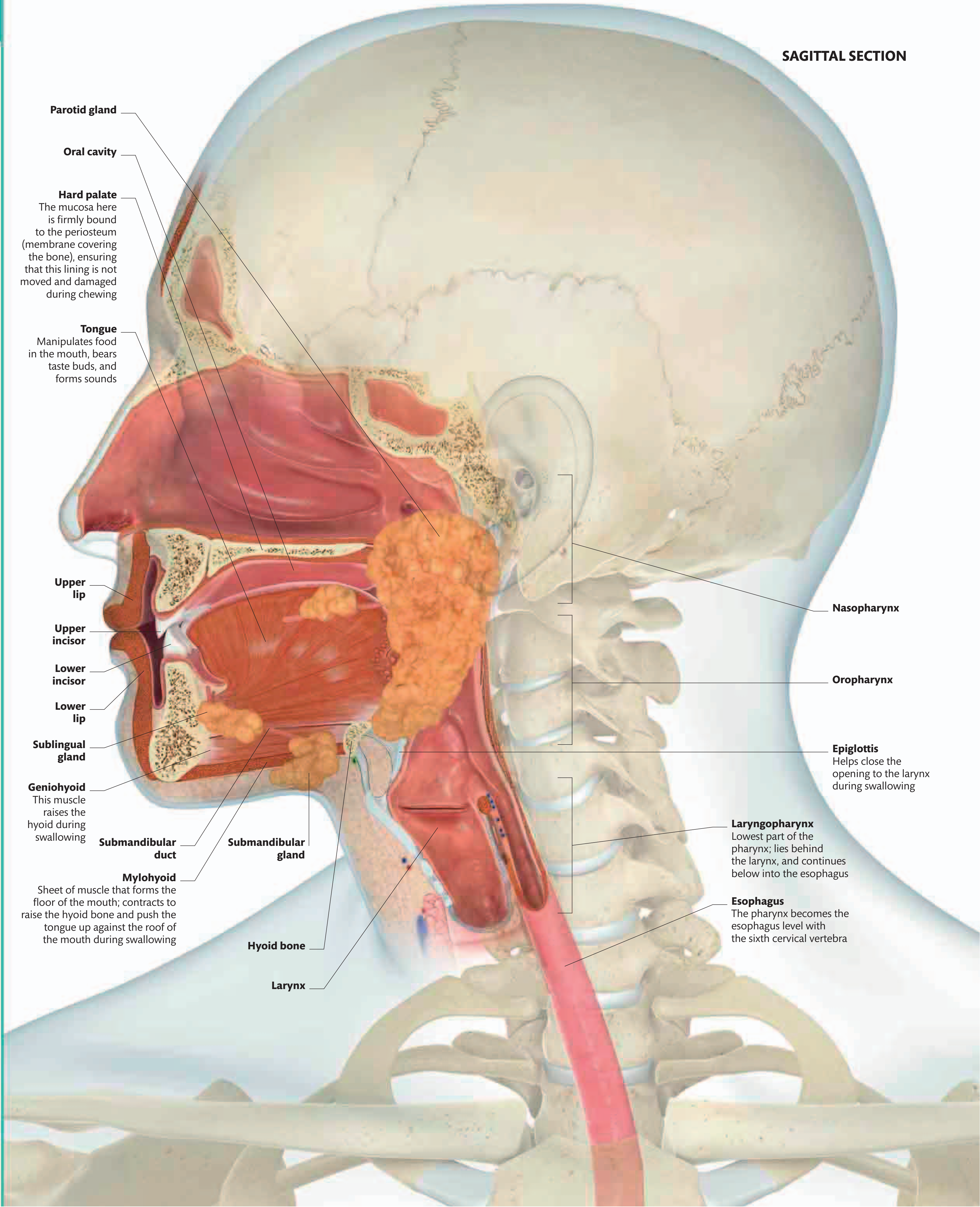

Parotid gland

Oral cavity

Hard palate
The mucosa here is firmly bound to the periosteum (membrane covering the bone), ensuring that this lining is not moved and damaged during chewing

Tongue
Manipulates food in the mouth, bears taste buds, and forms sounds

Upper lip

Upper incisor

Lower incisor

Lower lip

Sublingual gland

Geniohyoid
This muscle raises the hyoid during swallowing

Submandibular duct

Submandibular gland

Mylohyoid
Sheet of muscle that forms the floor of the mouth; contracts to raise the hyoid bone and push the tongue up against the roof of the mouth during swallowing

Hyoid bone

Larynx

Nasopharynx

Oropharynx

Epiglottis
Helps close the opening to the larynx during swallowing

Laryngopharynx
Lowest part of the pharynx; lies behind the larynx, and continues below into the esophagus

Esophagus
The pharynx becomes the esophagus level with the sixth cervical vertebra

HEAD AND NECK

The mouth is the first part of the digestive tract, and it is here that the processes of mechanical and chemical digestion get underway. Your teeth grind each mouthful, and you have three pairs of major salivary glands—parotid, submandibular, and sublingual—that secrete saliva through ducts into the mouth. Saliva contains digestive enzymes that begin to chemically break down the food in your mouth. The tongue manipulates the food, and also has taste buds that allow you to quickly make the important distinction between delicious food and potentially harmful toxins. As you swallow, the tongue pushes up against the hard palate, the soft palate seals off the airway, and the muscular tube of the pharynx contracts in a wave to push the ball of food down into the esophagus, ready for the next stage of its journey.

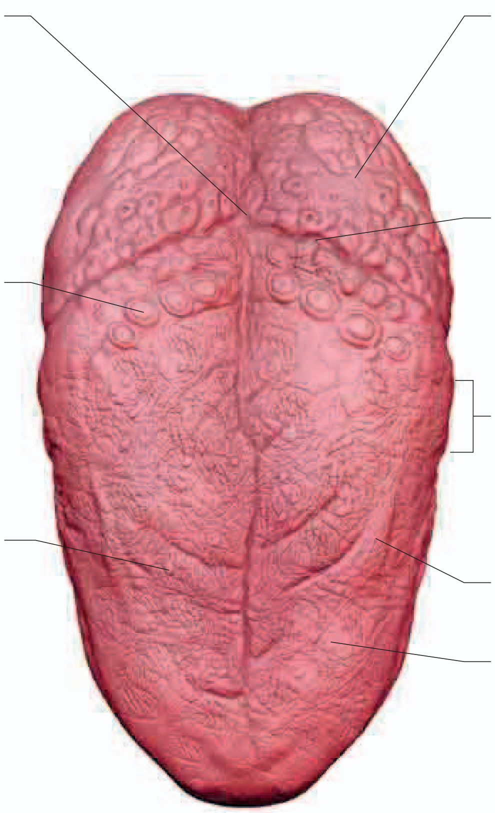

Foramen cecum This small, blind hole at the back of the tongue is a remnant of where the thyroid gland started to develop in the embryo, before it dropped down into the neck

Vallate papillae There are around a dozen of these large papillae at the back of the tongue; each one is surrounded by a circular furrow that contains taste buds

Fungiform papilla Literally means mushroom-shaped; these are scattered over the tongue like mushrooms across the lawn of filiform papillae; fungiform papillae also bear taste buds

Pharyngeal part of tongue Lymphoid tissue underlies the mucosa here, forming the lingual tonsil

Sulcus terminalis Border between the pharyngeal and oral parts of the tongue, lying in the oropharynx and oral cavity respectively

Foliate papillae Leaf-shaped papillae that form a series of ridges on each side of the back of the tongue

Oral part of the tongue

Filiform papilla Tiny, hair-shaped papillae that give the tongue a velvety texture

TONGUE

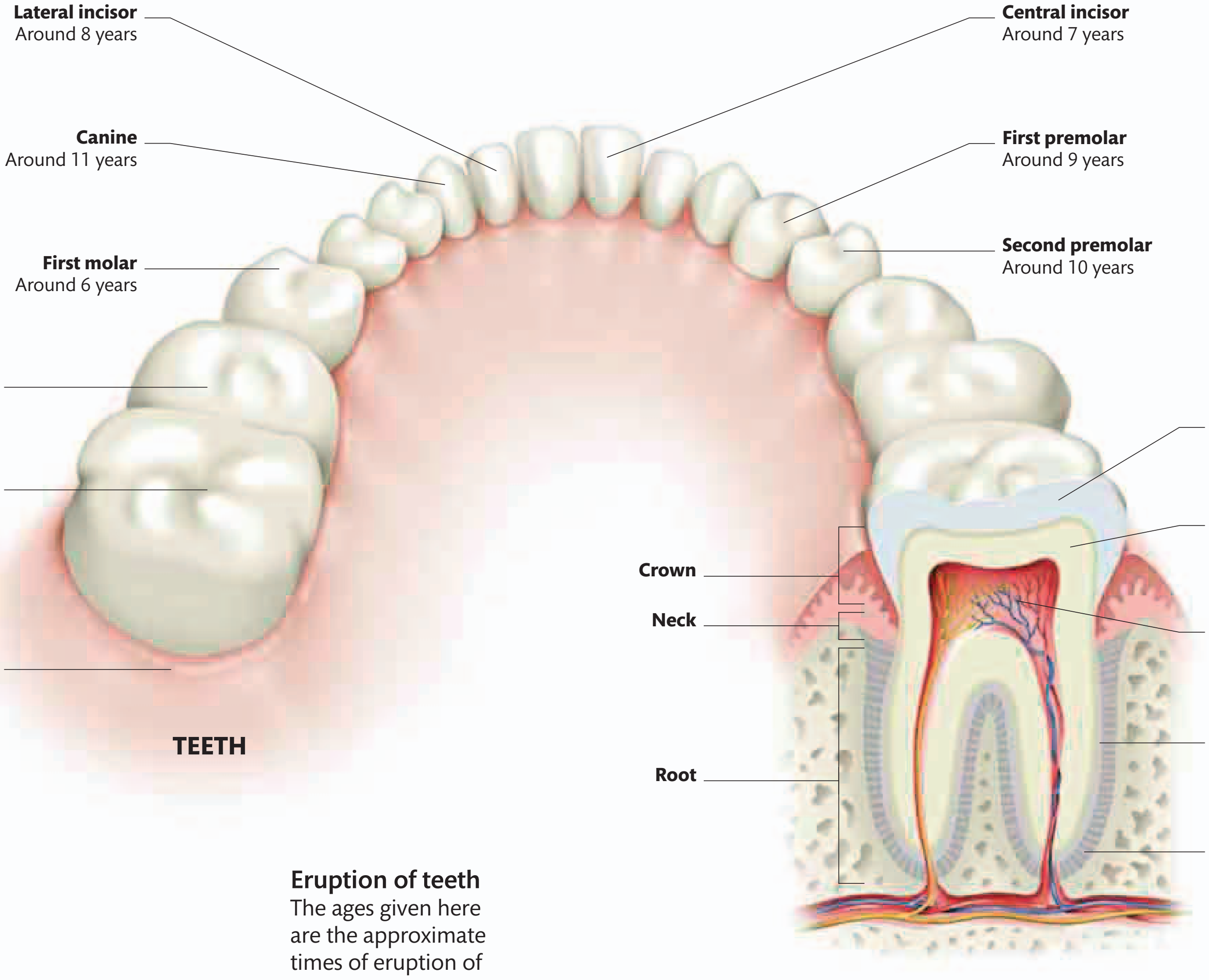

Lateral incisor Around 8 years

Canine Around 11 years

First molar Around 6 years

Second molar Around 12 years

Third molar From 17 to 21 years (although these "wisdom teeth" may stay, unerupted, in the jaw, or may even be absent)

Gingivae The gums—connective tissue containing blood vessels, covered with mucosa

Central incisor Around 7 years

First premolar Around 9 years

Second premolar Around 10 years

Crown

Neck

Root

Enamel The hardest tissue in the human body

Dentine Hard tissue forming the bulk of the tooth

Pulp cavity Connective tissue containing nerves and blood vessels

Cementum This bonelike tissue covers the roots of the tooth

Periodontal ligament Collagen fibers bind the tooth roots to the bone of the socket

TEETH

Eruption of teeth The ages given here are the approximate times of eruption of the permanent teeth.

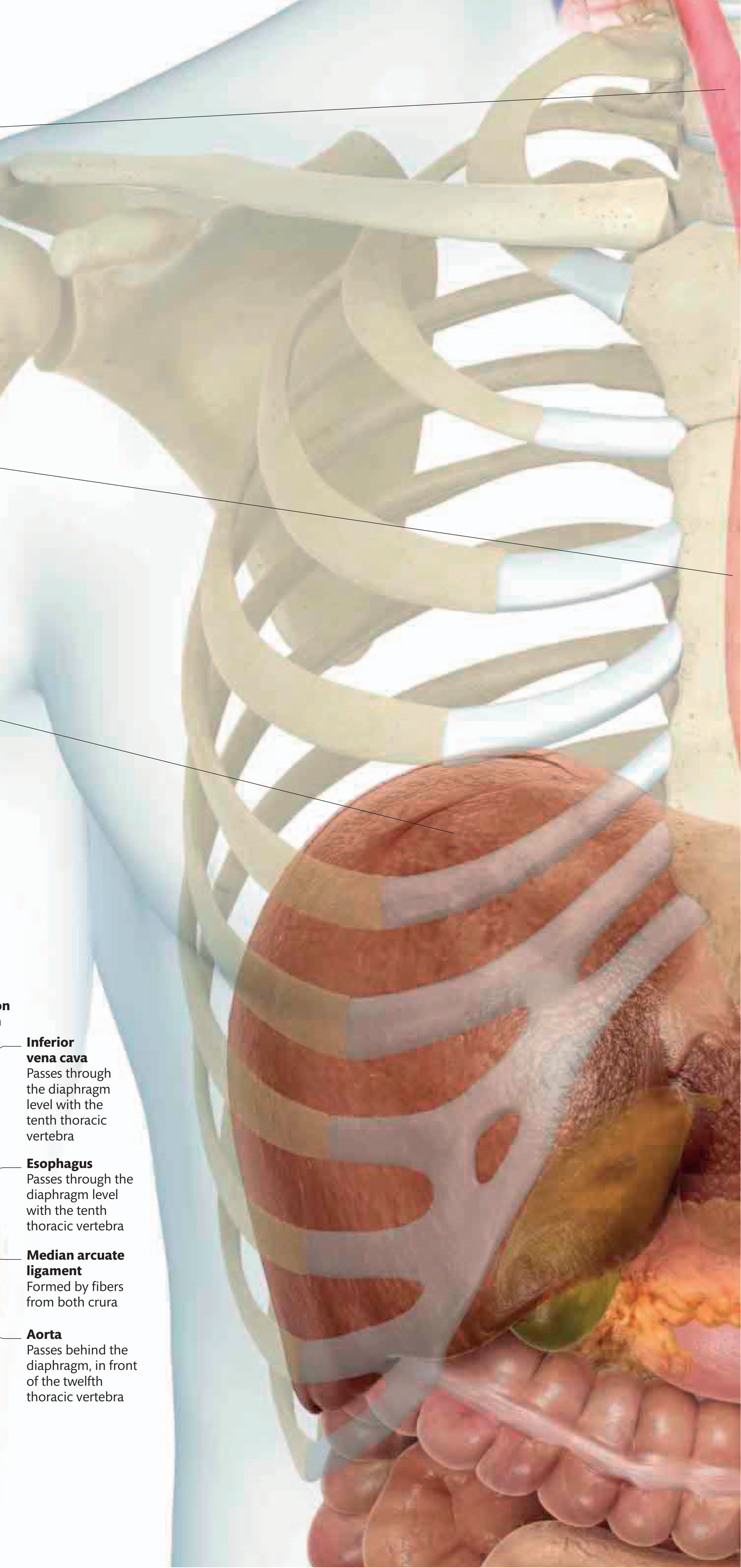

Esophagus
In the neck, the esophagus lies behind the trachea

Thoracic part of the esophagus
The esophagus is slightly constricted here by the left main bronchus, which crosses in front of it

Liver
Lies under the right dome of the diaphragm, and largely under cover of the ribs

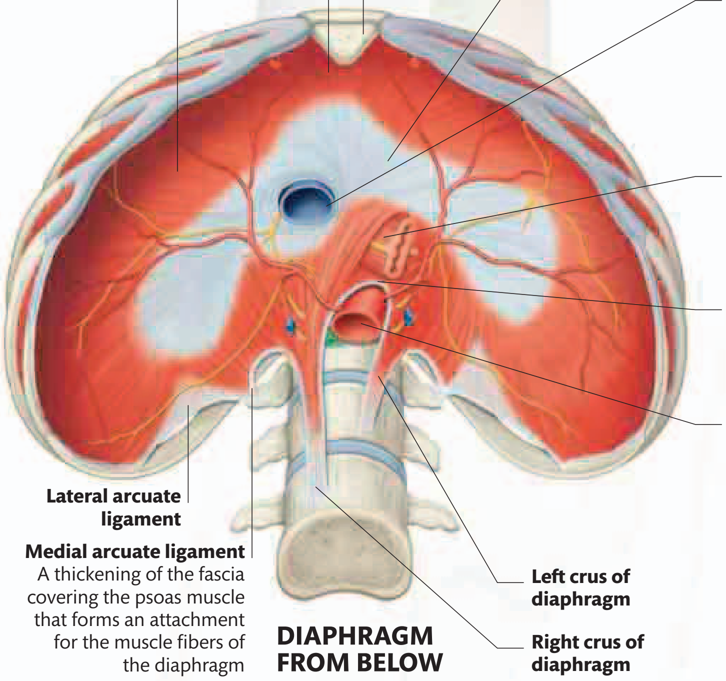

DIAPHRAGM FROM BELOW

Muscular part of diaphragm

Sternal part of diaphragm

Xiphoid process

Central tendon of diaphragm

Inferior vena cava
Passes through the diaphragm level with the tenth thoracic vertebra

Esophagus
Passes through the diaphragm level with the tenth thoracic vertebra

Median arcuate ligament
Formed by fibers from both crura

Aorta
Passes behind the diaphragm, in front of the twelfth thoracic vertebra

Lateral arcuate ligament

Medial arcuate ligament
A thickening of the fascia covering the psoas muscle that forms an attachment for the muscle fibers of the diaphragm

Left crus of diaphragm

Right crus of diaphragm

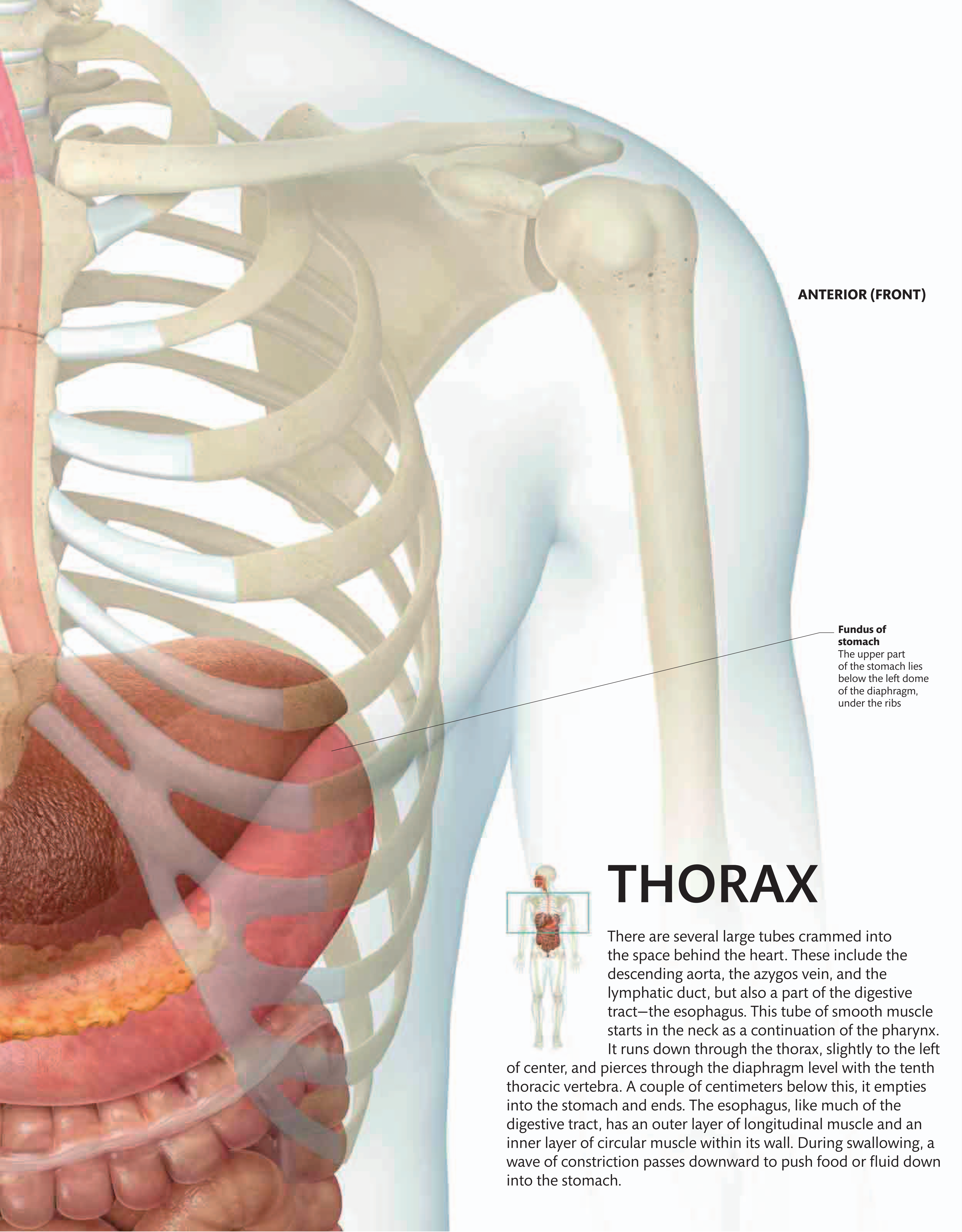

ANTERIOR (FRONT)

Fundus of stomach
The upper part of the stomach lies below the left dome of the diaphragm, under the ribs

THORAX

There are several large tubes crammed into the space behind the heart. These include the descending aorta, the azygos vein, and the lymphatic duct, but also a part of the digestive tract—the esophagus. This tube of smooth muscle starts in the neck as a continuation of the pharynx. It runs down through the thorax, slightly to the left of center, and pierces through the diaphragm level with the tenth thoracic vertebra. A couple of centimeters below this, it empties into the stomach and ends. The esophagus, like much of the digestive tract, has an outer layer of longitudinal muscle and an inner layer of circular muscle within its wall. During swallowing, a wave of constriction passes downward to push food or fluid down into the stomach.

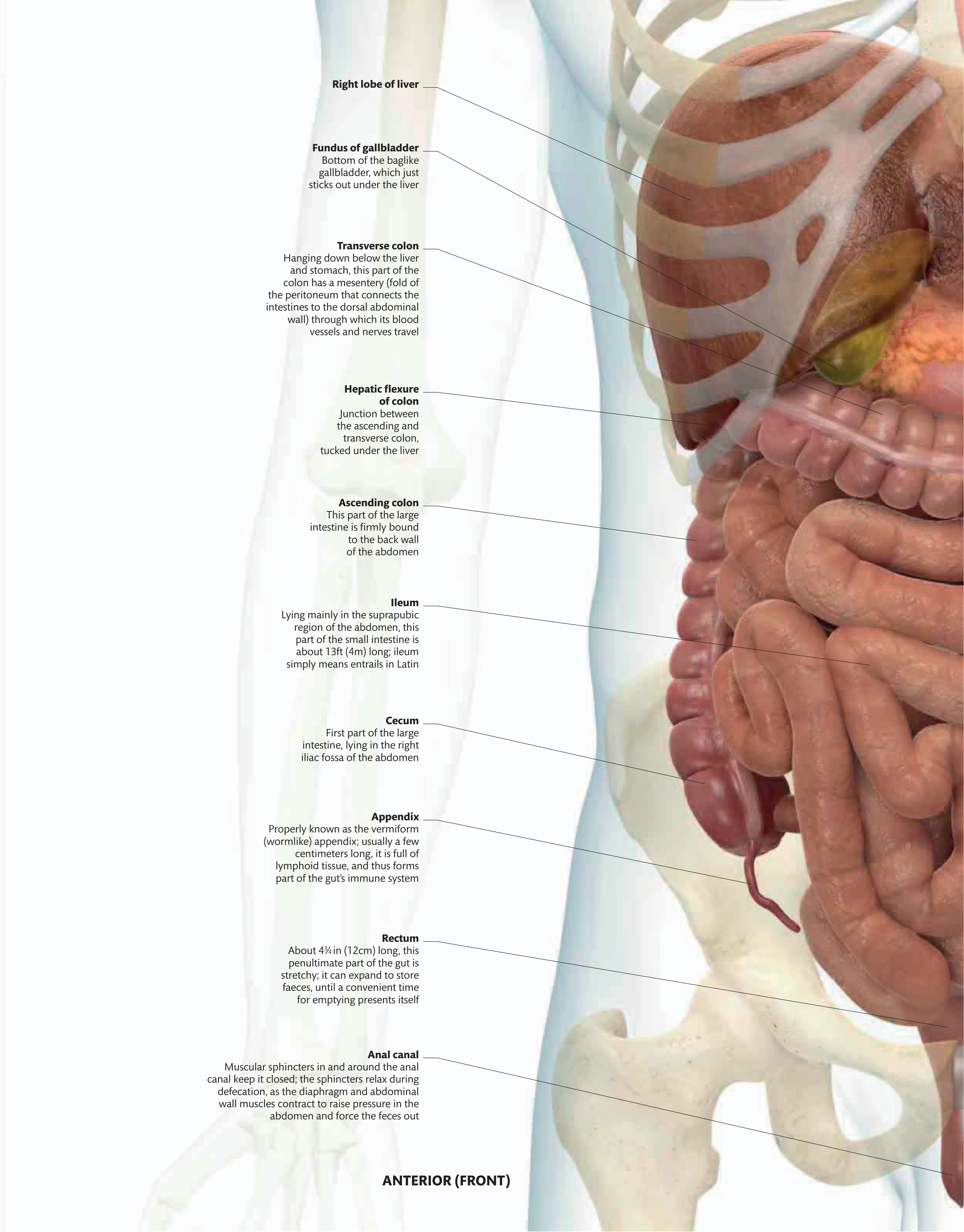

Right lobe of liver

Fundus of gallbladder
Bottom of the baglike gallbladder, which just sticks out under the liver

Transverse colon
Hanging down below the liver and stomach, this part of the colon has a mesentery (fold of the peritoneum that connects the intestines to the dorsal abdominal wall) through which its blood vessels and nerves travel

Hepatic flexure of colon
Junction between the ascending and transverse colon, tucked under the liver

Ascending colon
This part of the large intestine is firmly bound to the back wall of the abdomen

Ileum
Lying mainly in the suprapubic region of the abdomen, this part of the small intestine is about 13ft (4m) long; ileum simply means entrails in Latin

Cecum
First part of the large intestine, lying in the right iliac fossa of the abdomen

Appendix
Properly known as the vermiform (wormlike) appendix; usually a few centimeters long, it is full of lymphoid tissue, and thus forms part of the gut's immune system

Rectum
About 4¾ in (12cm) long, this penultimate part of the gut is stretchy; it can expand to store faeces, until a convenient time for emptying presents itself

Anal canal
Muscular sphincters in and around the anal canal keep it closed; the sphincters relax during defecation, as the diaphragm and abdominal wall muscles contract to raise pressure in the abdomen and force the feces out

ANTERIOR (FRONT)

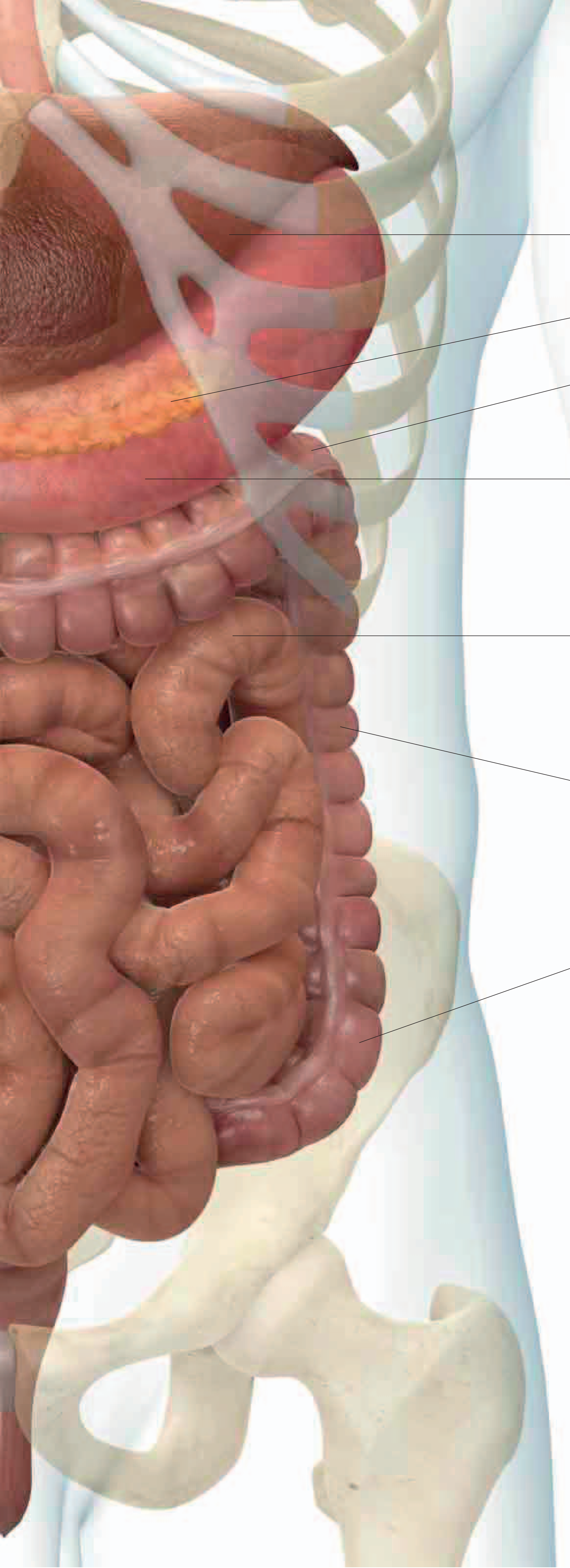

Left lobe of liver

Pancreas

Splenic flexure of colon
Junction between the transverse and descending colon, close to the spleen (spleen not shown here)

Stomach
The name comes originally from the Greek for gullet, but has come to mean this baglike part of the digestive system, just below the diaphragm

Jejunum
About 6½ft (2m) long, this part of the small intestine is more vascular (so slightly redder) than the ileum, and lies mainly in the umbilical region of the abdomen; its name comes from the Latin for empty—perhaps because food passes through here quickly

Descending colon
Like the ascending colon, this part of the large intestine has no mesentery, and is firmly bound to the back wall of the abdomen

Sigmoid colon
This S-shaped part of colon has a mesentery

ABDOMEN AND PELVIS

With the organs in situ, it is clear how much the abdominal cavity extends up under the ribs. The upper abdominal organs—the liver, stomach, and spleen—are largely under cover of the ribcage. This gives them some protection, but it also means that they are vulnerable to injury if a lower rib is fractured. The large intestine forms an M shape in the abdomen, starting with the cecum low down on the right, and the ascending colon running up the right flank and tucking under the liver. The transverse colon hangs down below the liver and stomach, and the descending colon runs down the left side of the abdomen. This becomes the S-shaped sigmoid colon, which runs down into the pelvis to become the rectum. The coils of the small intestine occupy the middle of the abdomen.

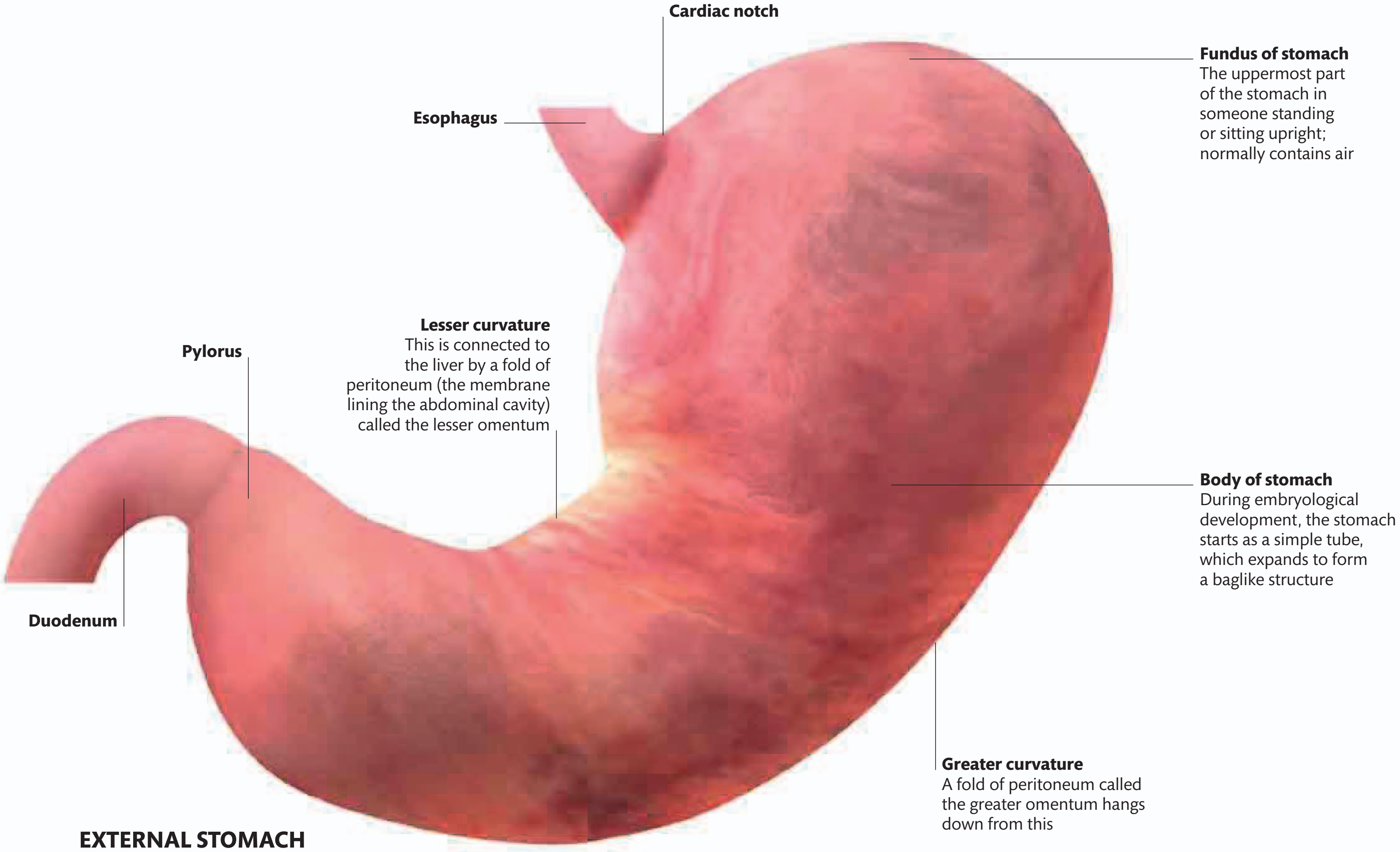
Cardiac notch
Fundus of stomach The uppermost part of the stomach in someone standing or sitting upright; normally contains air
Esophagus
Lesser curvature This is connected to the liver by a fold of peritoneum (the membrane lining the abdominal cavity) called the lesser omentum
Pylorus
Body of stomach During embryological development, the stomach starts as a simple tube, which expands to form a baglike structure
Duodenum
Greater curvature A fold of peritoneum called the greater omentum hangs down from this

EXTERNAL STOMACH

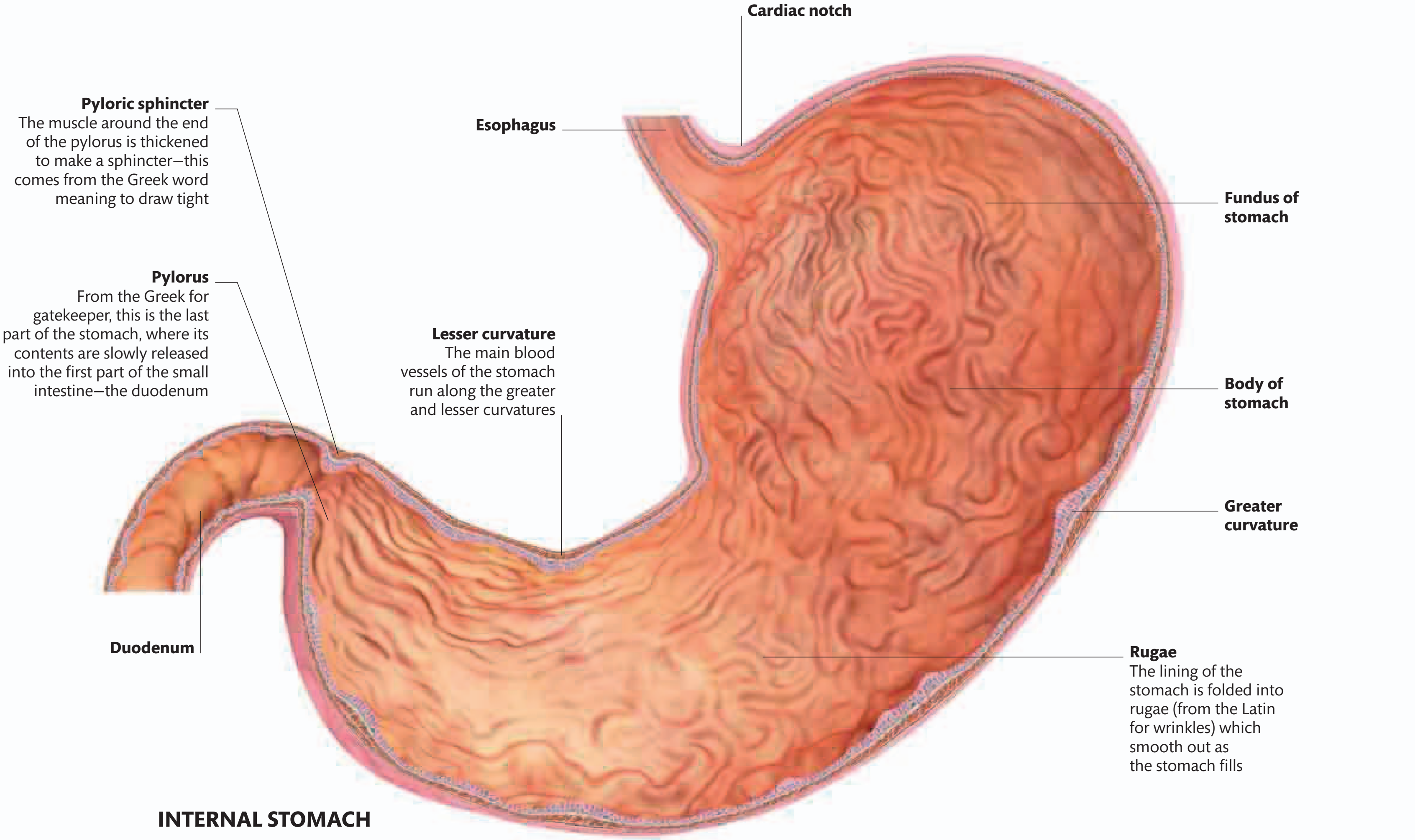
Cardiac notch
Pyloric sphincter The muscle around the end of the pylorus is thickened to make a sphincter—this comes from the Greek word meaning to draw tight
Esophagus
Fundus of stomach
Pylorus From the Greek for gatekeeper, this is the last part of the stomach, where its contents are slowly released into the first part of the small intestine—the duodenum
Lesser curvature The main blood vessels of the stomach run along the greater and lesser curvatures
Body of stomach
Greater curvature
Duodenum
Rugae The lining of the stomach is folded into rugae (from the Latin for wrinkles) which smooth out as the stomach fills

INTERNAL STOMACH

STOMACH AND INTESTINES

The stomach is a muscular bag, where food is held before moving on to the intestines. Inside the stomach, food is exposed to a cocktail of hydrochloric acid, which kills off bacteria, and protein-digesting enzymes. The layered muscle of the stomach wall contracts to churn up its contents. Semidigested food is released from the stomach into the first part of the small intestine, the duodenum, where bile and pancreatic juices are added. Contractions in the intestine wall then push the liquid food into the jejunum and ileum, where digestion continues. What is left passes into the cecum, the beginning of the large intestine. In the colon, the next part of the large intestine, water is absorbed so that the gut contents become more solid. The resulting feces pass into the rectum, where they are stored until excretion.

SMOOTH MUSCLE STRUCTURE

Functions of the gut, blood vessels, and respiratory tract are carried out involuntarily, at a subconscious level, with the help of a special type of muscle called smooth muscle. This is supplied by autonomic motor nerves.

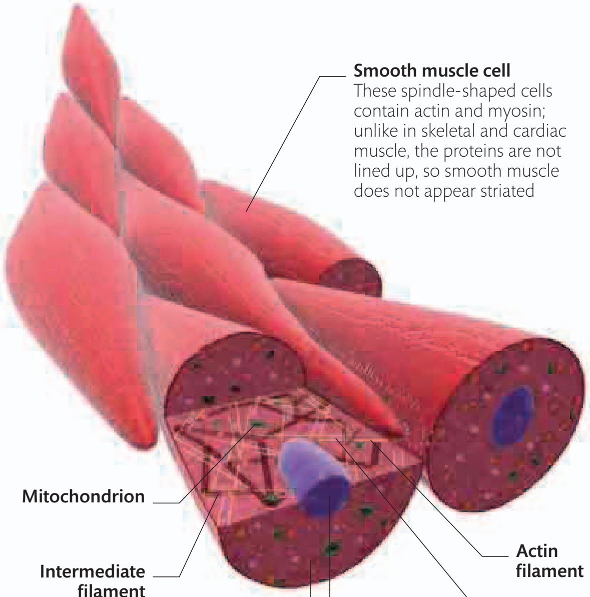

Smooth muscle cell
These spindle-shaped cells contain actin and myosin; unlike in skeletal and cardiac muscle, the proteins are not lined up, so smooth muscle does not appear striated

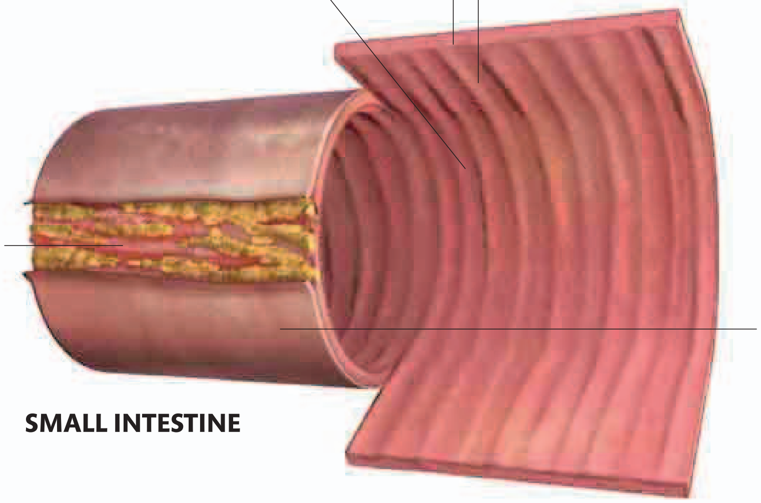

Mucosa
The epithelium lining is packed with mucus-producing glands

Circular folds
These ridges help to increase the surface area available for absorption of nutrients

Serous lining of the small intestine
This is formed by the mesentery (membranous folds) enveloping the gut tube

SMALL INTESTINE

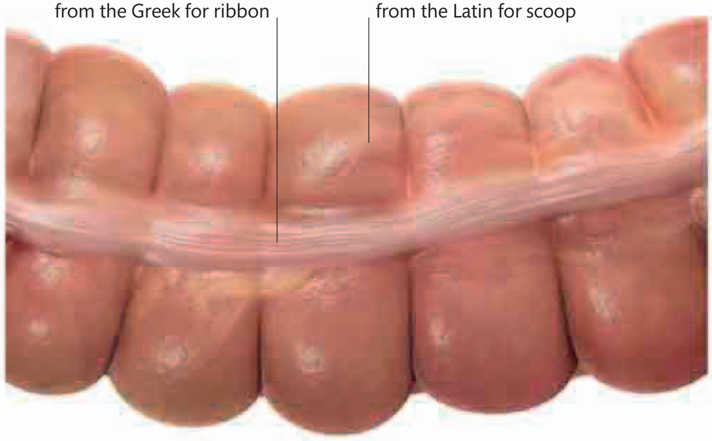

Taenia coli
The longitudinal muscle coat is condensed into three bands, or taeniae; the name comes from the Greek for ribbon

Haustra
This is the name given to the sacculations (pouches) of the large intestine; it comes from the Latin for scoop

LARGE INTESTINE

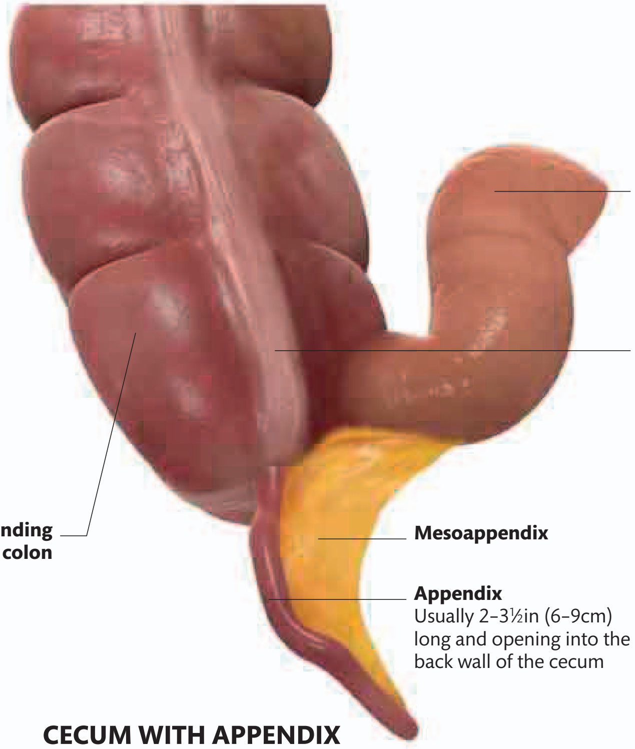

Taenia coli
These ribbonlike bands of longitudinal muscle converge on the base of the appendix

Appendix
Usually 2–3½in (6–9cm) long and opening into the back wall of the cecum

CECUM WITH APPENDIX

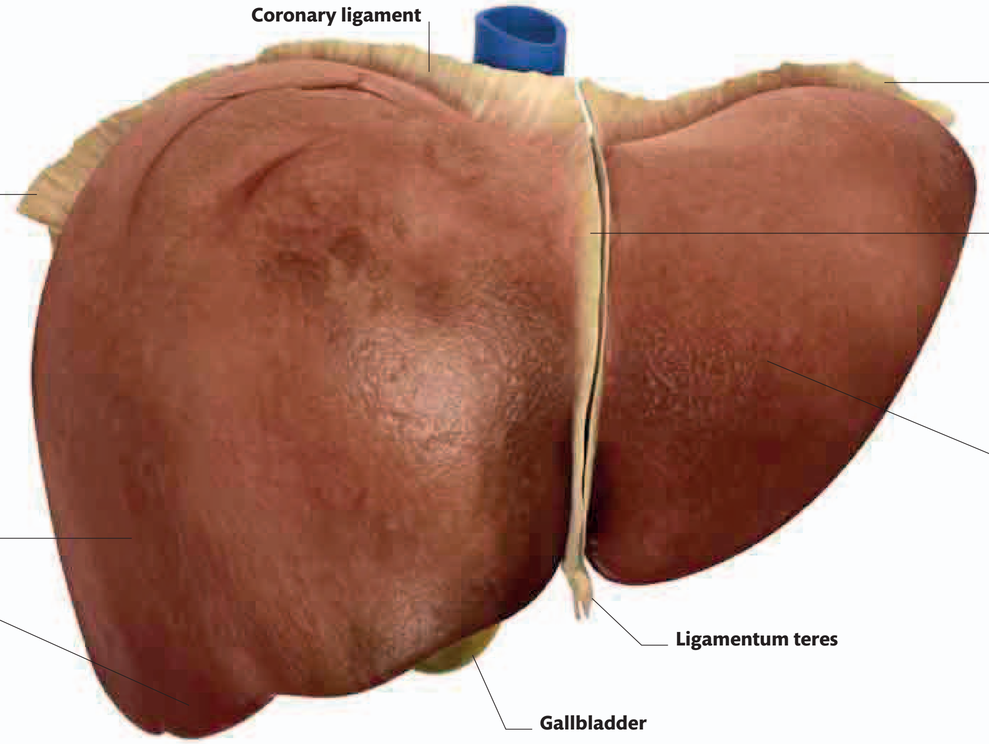
Coronary ligament
Left triangular ligament
Right triangular ligament
Falciform ligament
Left lobe of liver
Right lobe of liver
Inferior margin
Ligamentum teres
Gallbladder

Right triangular ligament The peritoneum, which lines the walls of the abdominal cavity and the organs within it, is one continuous sheet of serous membrane; the parts of it that form connections between the abdominal wall and organs are known variously as ligaments, mesenteries, or omenta

Falciform ligament A fold of peritoneum attaching the liver to the anterior abdominal wall and diaphragm

Inferior margin This sharp edge separates the dome-shaped upper surface of the liver, which lies under the diaphragm, from the lower surface, which lies against the other abdominal organs

LIVER ANTERIOR (FRONT)

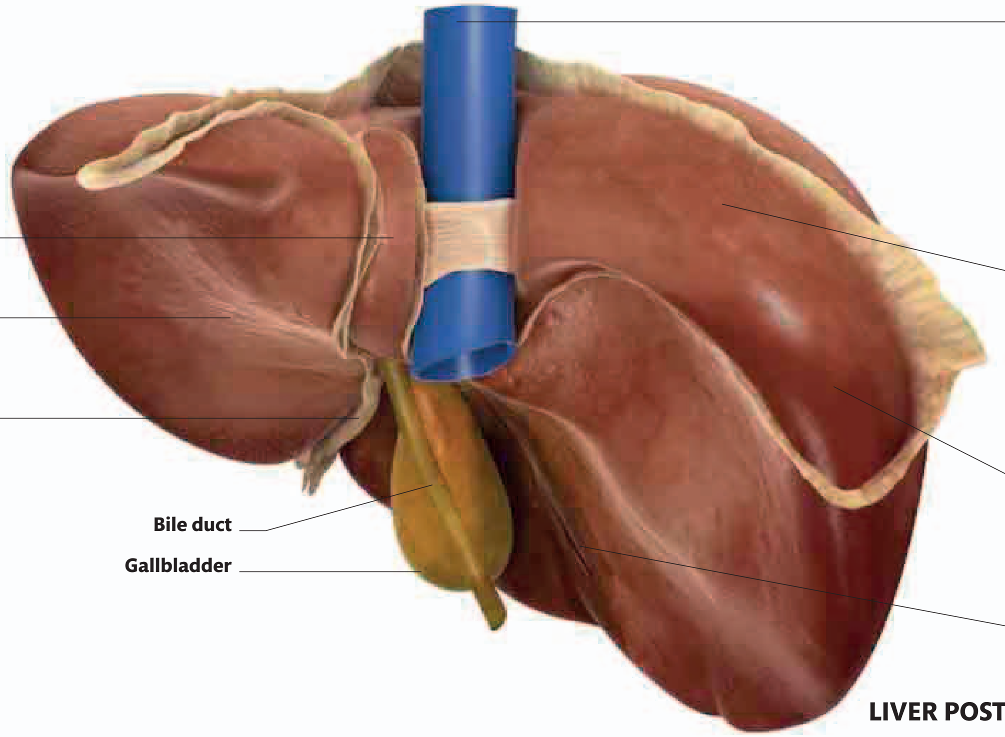
Inferior vena cava
Caudate lobe
Left lobe of liver
Bare area
Ligamentum venosum
Right lobe of liver
Bile duct
Gallbladder
Quadrate lobe

Inferior vena cava This large vein is partly embedded in the back of the liver; the three hepatic veins drain directly into it

Bare area This area of the liver is not covered with the peritoneum

Ligamentum venosum A remnant of what was the umbilical vein in the fetus

LIVER POSTERIOR (BACK)

LIVER, PANCREAS, AND GALLBLADDER

The liver, the largest internal organ, can weigh up to 6lb (3kg). It does hundreds of jobs simultaneously, many of them related to digestion. It produces bile, which is stored in the gallbladder and helps to digest fats. It also receives nutrients from the gut via the portal vein and processes them. It breaks down or builds up proteins, carbohydrates, and fats according to need; detoxifies or deactivates substances such as alcohol and drugs; and plays a role in the immune system. The pancreas, a long, thin, leaf-shaped gland lying under the liver and behind the stomach, produces hormones that are secreted into the blood, and makes pancreatic juice, full of digestive enzymes, which it empties into the duodenum.

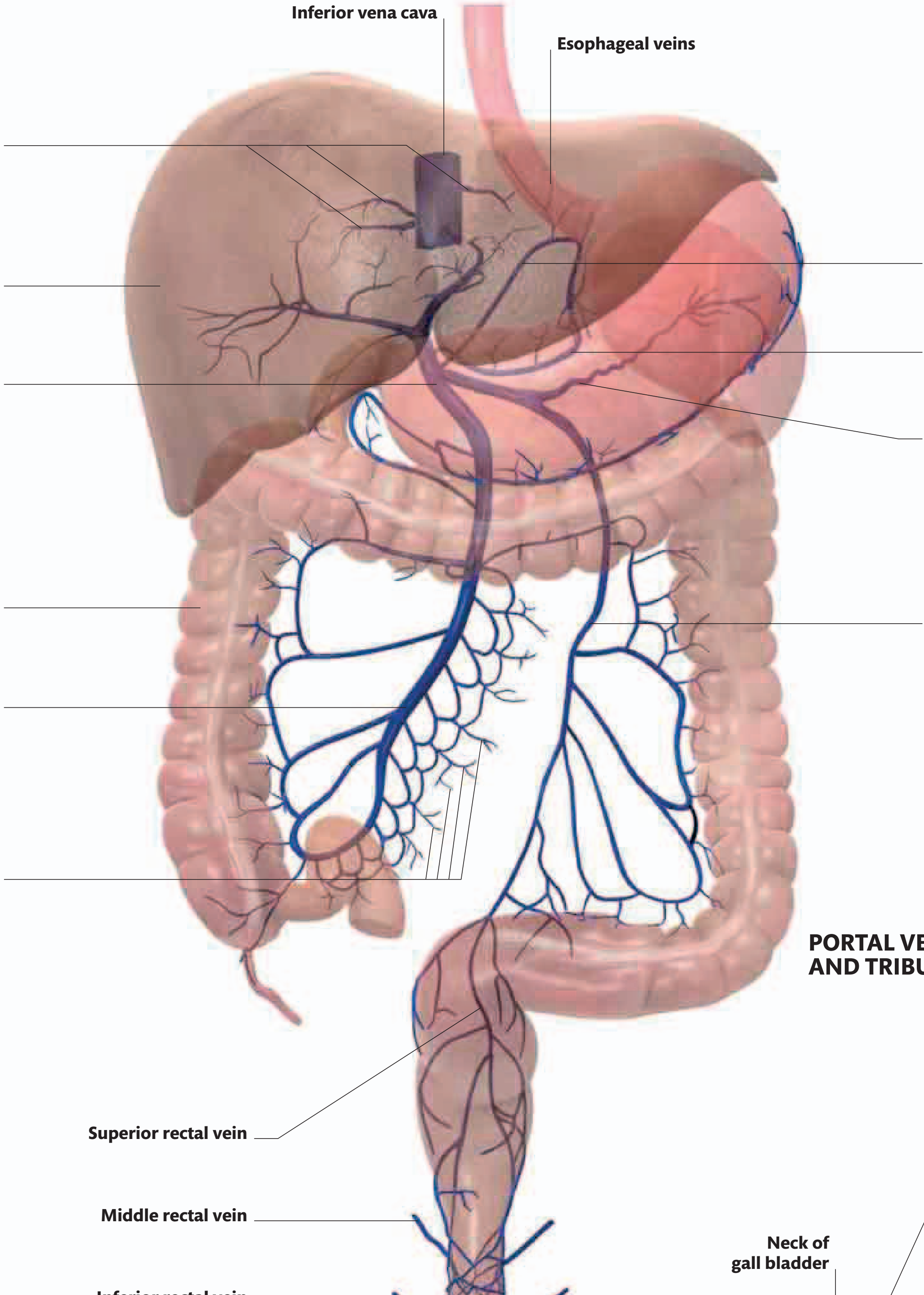

Inferior vena cava

Esophageal veins

Hepatic veins
Drain into inferior vena cava

Liver

Left gastric vein

Right gastric vein

Hepatic portal vein
This is about 3in (8cm) long and is where veins from every part of the digestive tract converge and enter the liver

Splenic vein
Drains blood from the spleen, pancreas, and stomach, and receives the inferior mesenteric vein

Large intestine

Inferior mesenteric vein
Drains the rectum, sigmoid colon, and descending colon

Superior mesenteric vein
Drains the small intestine, cecum, ascending and transverse colon

Jejunal and ileal veins

Superior rectal vein

Middle rectal vein

Inferior rectal vein

PORTAL VEINS AND TRIBUTARIES

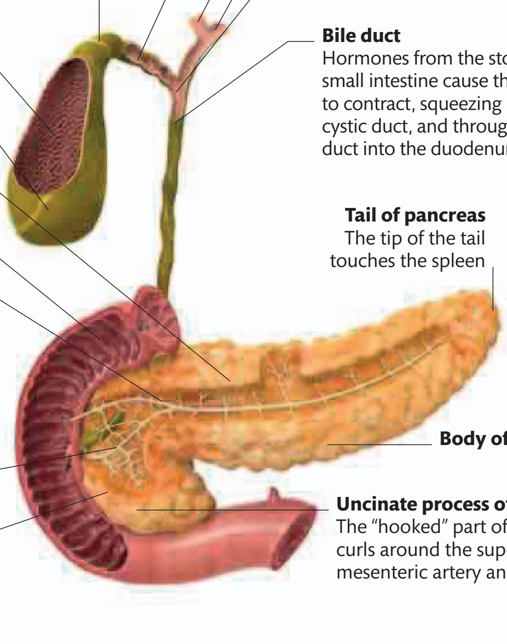

Cystic duct
Between meals, bile is diverted up the cystic duct into the gallbladder for storage

Right hepatic duct

Left hepatic duct

Common hepatic duct
Drains the bile produced by the liver

Neck of gall bladder

Body of gallbladder

Bile duct
Hormones from the stomach and small intestine cause the gallbladder to contract, squeezing bile down the cystic duct, and through the bile duct into the duodenum

Fundus of gallbladder

Neck of pancreas

Duodenum

Tail of pancreas
The tip of the tail touches the spleen

Main pancreatic duct
If the duct becomes blocked, the digestive enzymes build up inside the pancreas and may damage it

Body of pancreas

Accessory pancreatic duct

Uncinate process of pancreas
The "hooked" part of the pancreas curls around the superior mesenteric artery and vein

Head of pancreas
The head lies tucked into the C-shaped curve of the duodenum

GALLBLADDER AND PANCREAS

URINARY SYSTEM
OVERVIEW

The urinary system comprises the kidneys, ureters, bladder, and urethra. The kidneys lie high up in the abdomen, on its back wall. The upper part of both kidneys is tucked under the twelfth rib. The kidneys filter the blood and ensure that it stays at exactly the right volume and concentration to keep all the cells in the body working properly. They also get rid of unwanted substances from the blood, playing an important role in excreting nitrogen-containing urea, for example. The urine made by the kidneys is carried by the ureters down to the bladder, which lies in the pelvis. The urethra runs from the bottom of the bladder and opens to the outside world. In a woman, the urethra is short—only a few inches long—and opens at the perineum, between the legs. The urethra of a man is longer, running through the length of the penis to open on the tip.

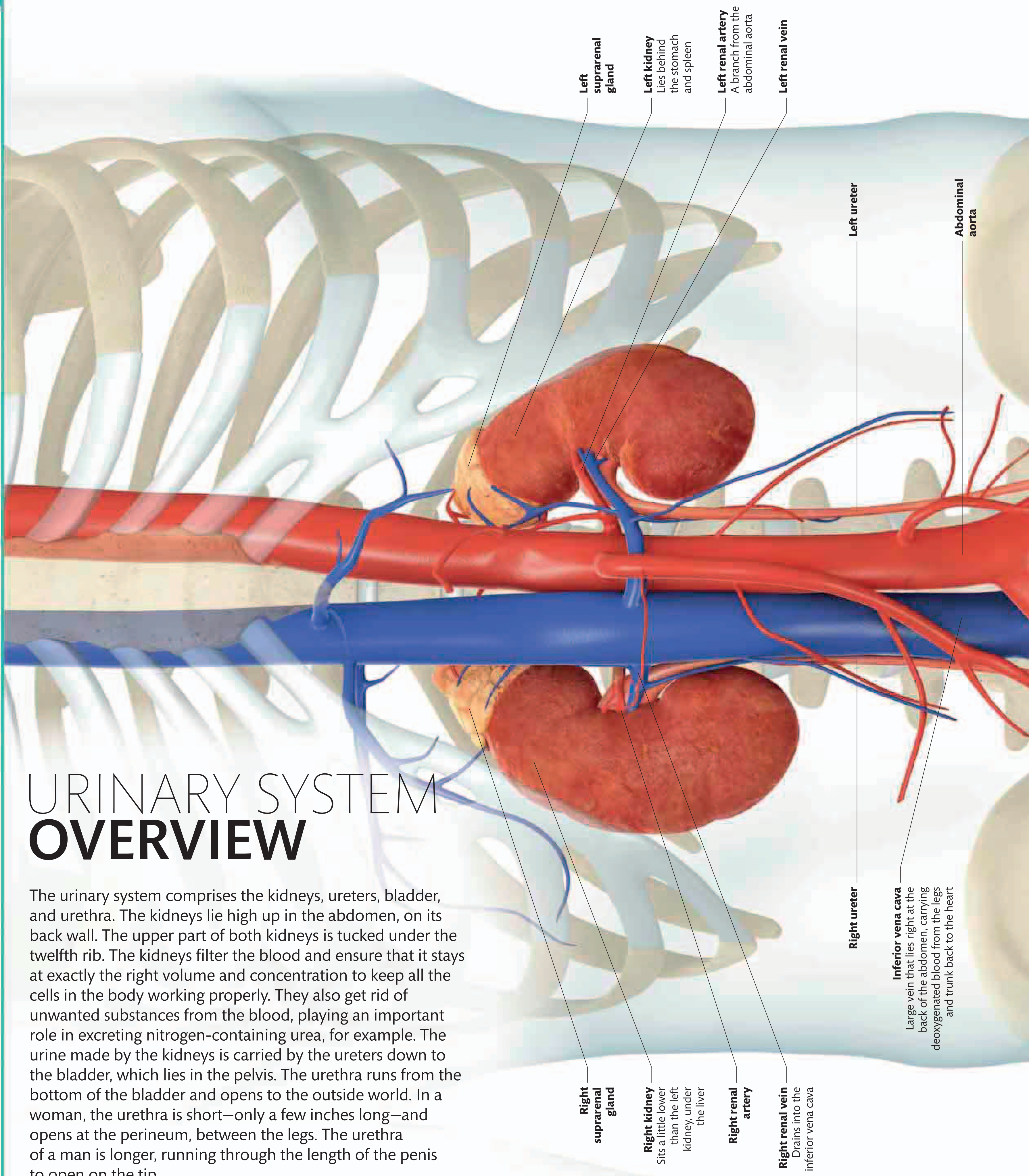
Left suprarenal gland
Left kidney Lies behind the stomach and spleen
Left renal artery A branch from the abdominal aorta
Left renal vein
Left ureter
Abdominal aorta
Right suprarenal gland
Right kidney Sits a little lower than the left kidney, under the liver
Right renal artery
Right renal vein Drains into the inferior vena cava
Right ureter
Inferior vena cava Large vein that lies right at the back of the abdomen, carrying deoxygenated blood from the legs and trunk back to the heart

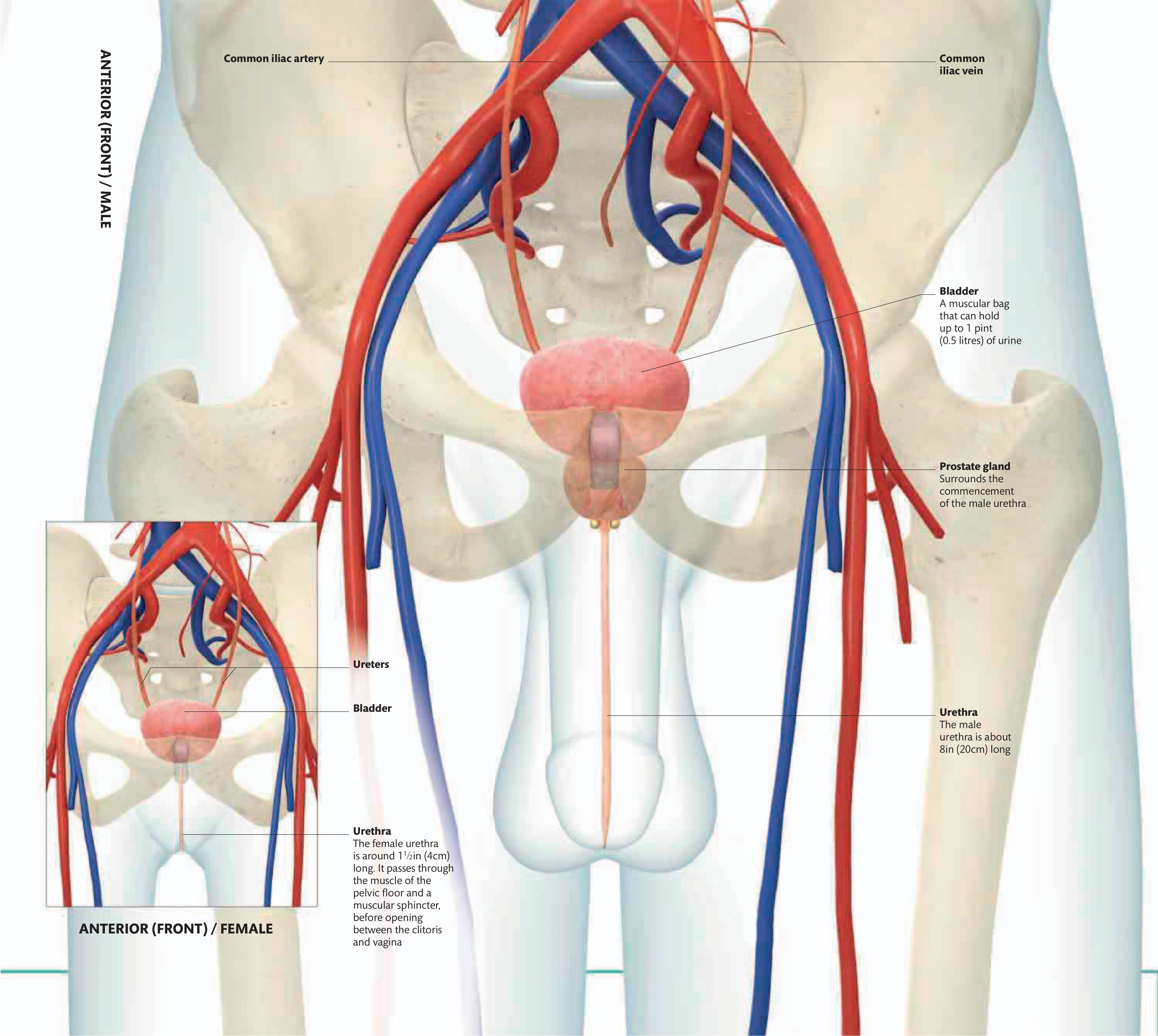

ANTERIOR (FRONT) / MALE

ANTERIOR (FRONT) / FEMALE

Common iliac artery
Common iliac vein
Ureters
Bladder

Bladder A muscular bag that can hold up to 1 pint (0.5 litres) of urine

Prostate gland Surrounds the commencement of the male urethra

Urethra The male urethra is about 8in (20cm) long

Urethra The female urethra is around 1½in (4cm) long. It passes through the muscle of the pelvic floor and a muscular sphincter, before opening between the clitoris and vagina

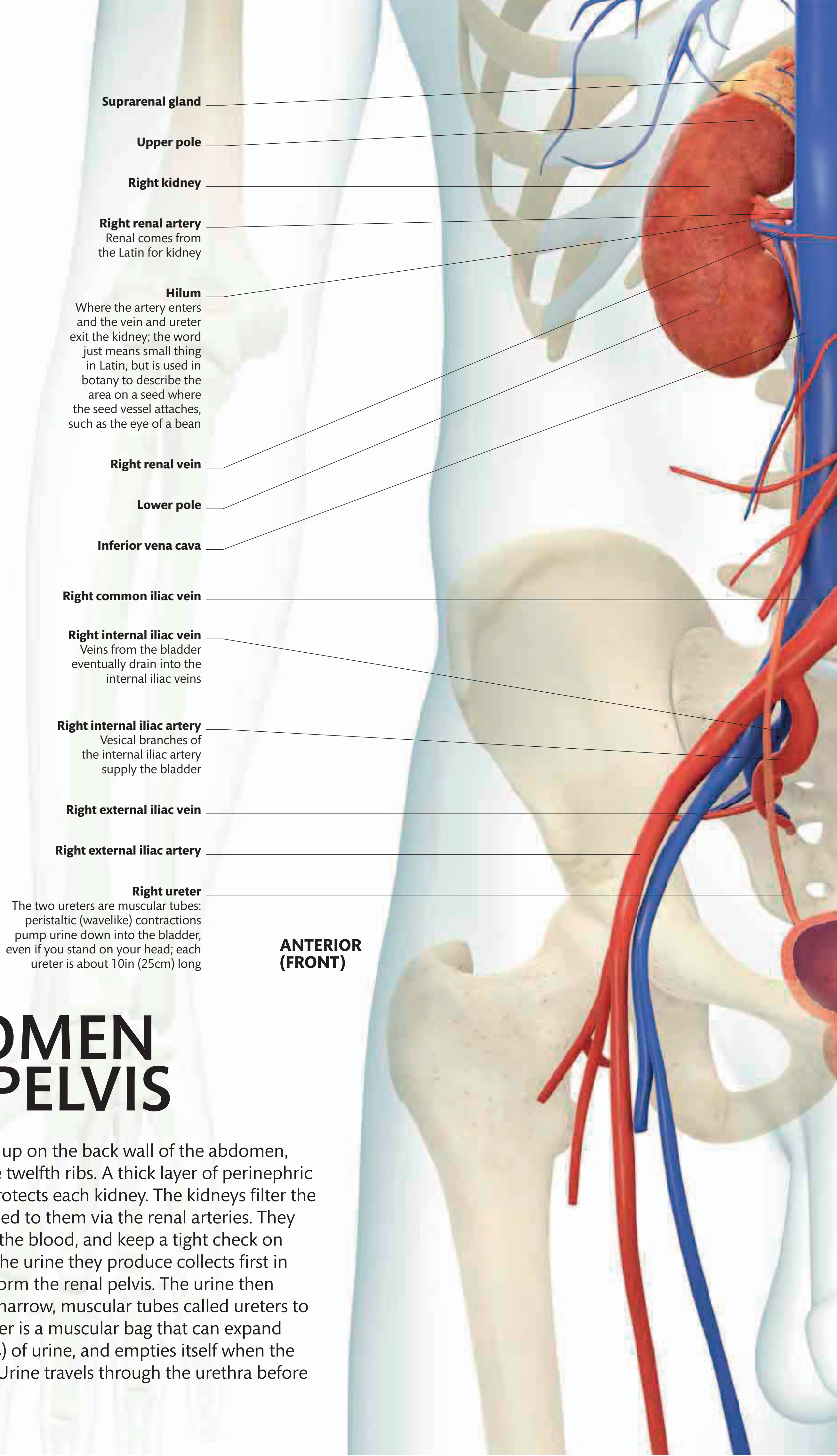

Suprarenal gland

Upper pole

Right kidney

Right renal artery
Renal comes from the Latin for kidney

Hilum
Where the artery enters and the vein and ureter exit the kidney; the word just means small thing in Latin, but is used in botany to describe the area on a seed where the seed vessel attaches, such as the eye of a bean

Right renal vein

Lower pole

Inferior vena cava

Right common iliac vein

Right internal iliac vein
Veins from the bladder eventually drain into the internal iliac veins

Right internal iliac artery
Vesical branches of the internal iliac artery supply the bladder

Right external iliac vein

Right external iliac artery

Right ureter
The two ureters are muscular tubes: peristaltic (wavelike) contractions pump urine down into the bladder, even if you stand on your head; each ureter is about 10in (25cm) long

ANTERIOR (FRONT)

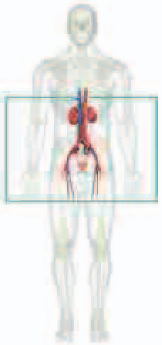

ABDOMEN AND PELVIS

The kidneys lie high up on the back wall of the abdomen, tucked up under the twelfth ribs. A thick layer of perinephric fat surrounds and protects each kidney. The kidneys filter the blood, which is carried to them via the renal arteries. They remove waste from the blood, and keep a tight check on blood volume and concentration. The urine they produce collects first in cup-shaped calyces, which join to form the renal pelvis. The urine then flows out of the kidneys and down narrow, muscular tubes called ureters to the bladder in the pelvis. The bladder is a muscular bag that can expand to hold up to about 1 pint (0.5 litres) of urine, and empties itself when the individual decides it is convenient. Urine travels through the urethra before leaving the body.

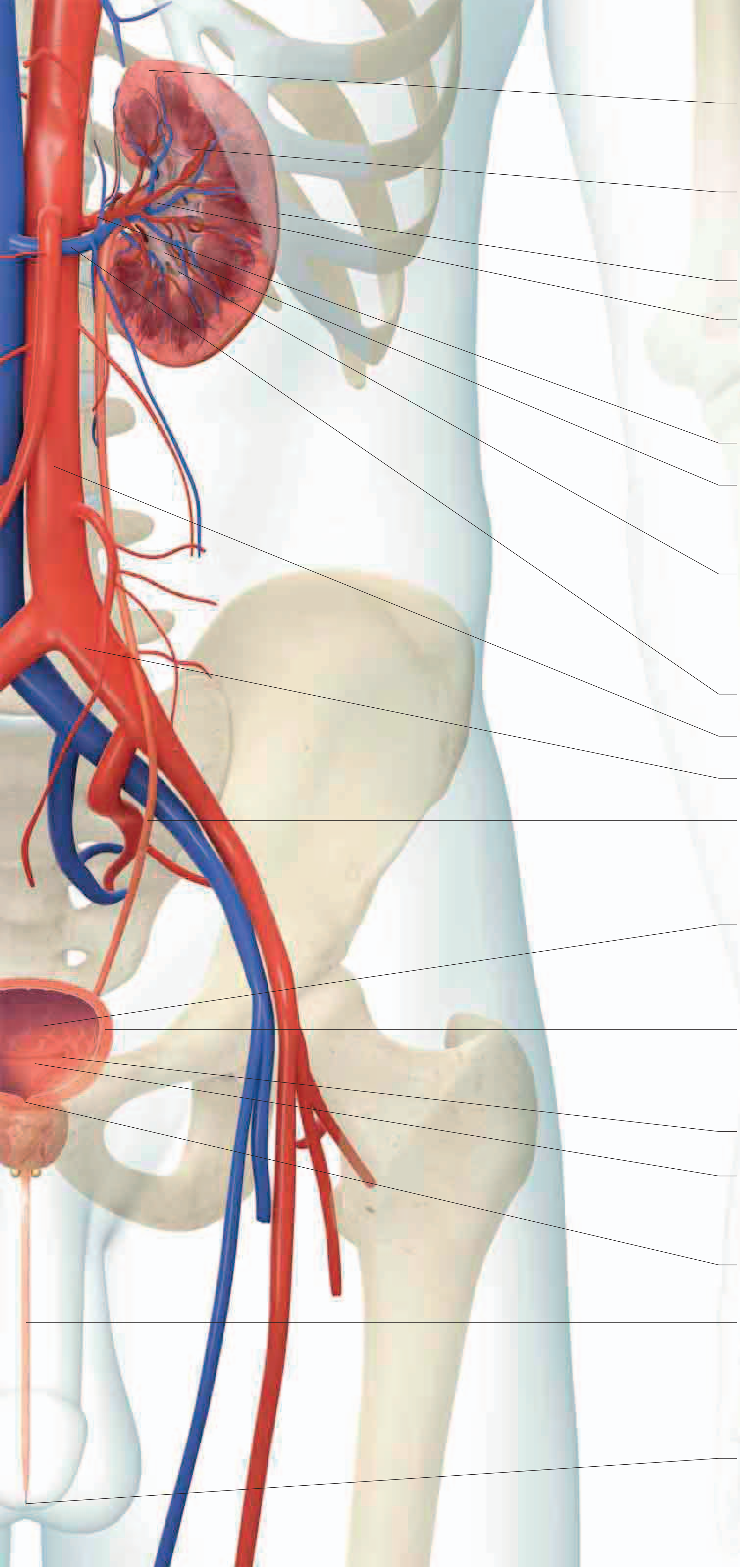

Renal cortex
Cortex means rind or bark; this is the outer tissue of the kidney

Renal medullary pyramid
Medulla means marrow or pith; this core tissue of the kidney is arranged as pyramids, which look triangular in cross section

Left kidney

Renal pelvis
Collects all urine from the kidney, and empties into the ureter; pelvis means basin in Latin, and the renal pelvis should not be confused with the bony pelvis—also shaped like a large basin

Left renal artery

Major calyx
The major calyces collect urine from the minor calyces, then themselves join together to form the renal pelvis

Minor calyx
Calyx originally meant flower covering in Greek, but because it is similar to the Latin word for cup it is used to describe cup-shaped structures in biology; urine from the microscopic collecting tubules of the kidney flows out into the minor calyces

Left renal vein

Abdominal aorta

Left common iliac artery

Left ureter
This name comes from the Greek for to make water; the two ureters carry urine from the kidneys to the bladder

Bladder
The empty bladder lies low down, in the true pelvis, behind the pubic symphysis; as the bladder fills, it expands up into the abdomen

Detrusor muscle
The crisscrossing smooth muscle bundles of the bladder wall give the inner surface of the bladder a netlike appearance

Ureteric orifice

Trigone
The three-cornered region of the back wall of the bladder, between the ureteric orifices and the internal urethral orifice

Internal urethral orifice

Urethra
From the Greek for urinate; this tube carries urine from the bladder to the outside world, a distance of around 1½in (4cm) in women, and about 8in (20cm) in men (as it travels the length of the penis)

External urethral orifice
The male urethra opens at the tip of the glans penis

REPRODUCTIVE SYSTEM OVERVIEW

Most organs in the body are similar in men and women. However, when it comes to the reproductive organs, there is a world of difference. In a woman, the ovaries, which produce eggs and female sex hormones, are tucked away, deep inside the pelvis. Also located within the pelvis are the vagina, uterus, and paired oviducts, or fallopian tubes, in which eggs are conveyed from the ovaries to the uterus. The woman's reproductive system also includes the mammary glands, which are important in providing milk for the newborn.

In a man, the testes, which produce sperm and sex hormones, hang well outside the pelvis, in the scrotum. The rest of the male reproductive system consists of a pair of tubes called the vasa deferentia (singular, vas deferens), the accessory sex glands (the seminal vesicles and the prostate), and the urethra.

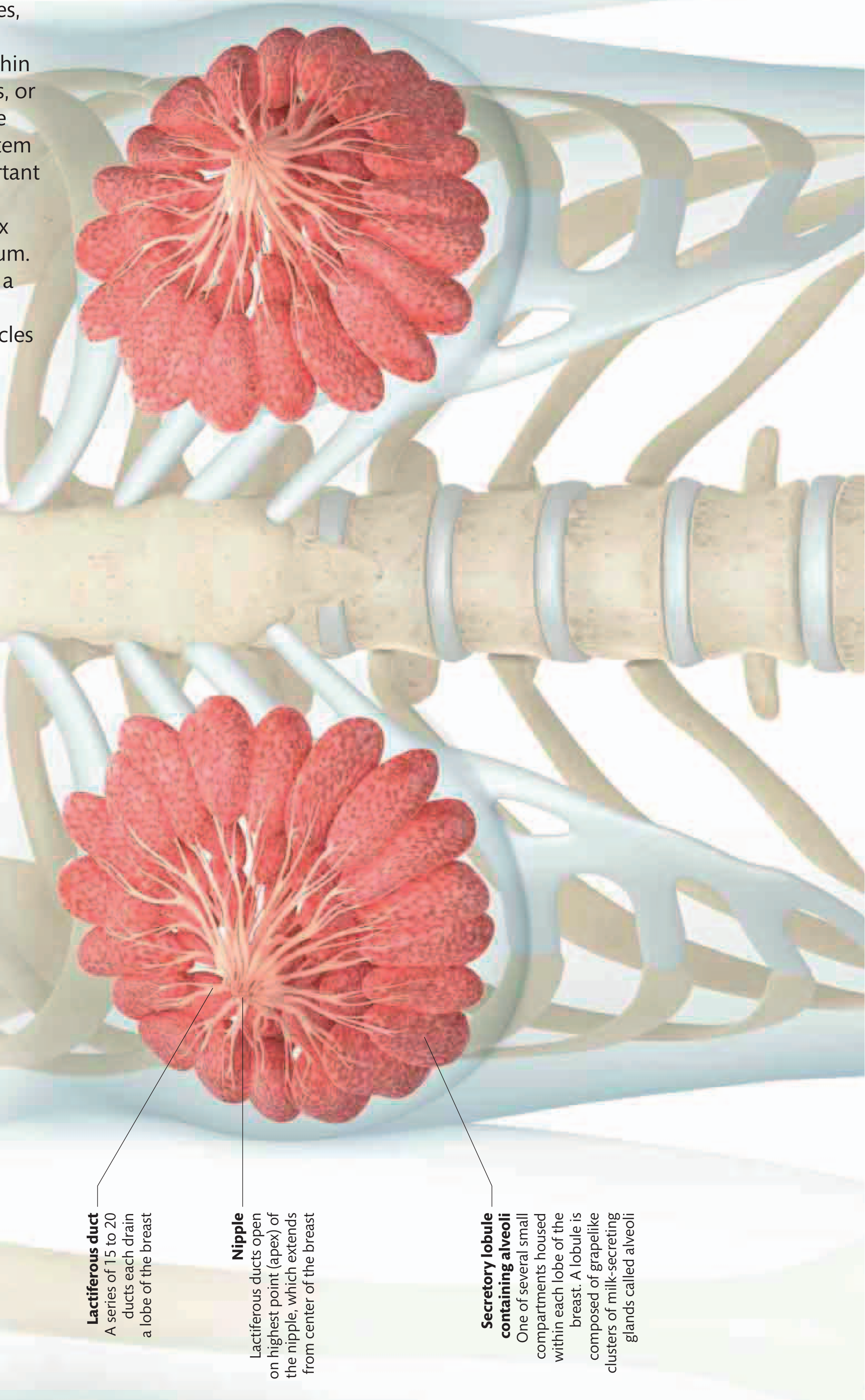
Lactiferous duct
A series of 15 to 20 ducts each drain a lobe of the breast
Nipple
Lactiferous ducts open on highest point (apex) of the nipple, which extends from center of the breast
Secretory lobule containing alveoli
One of several small compartments housed within each lobe of the breast. A lobule is composed of grapelike clusters of milk-secreting glands called alveoli

ANTERIOR (FRONT) / FEMALE

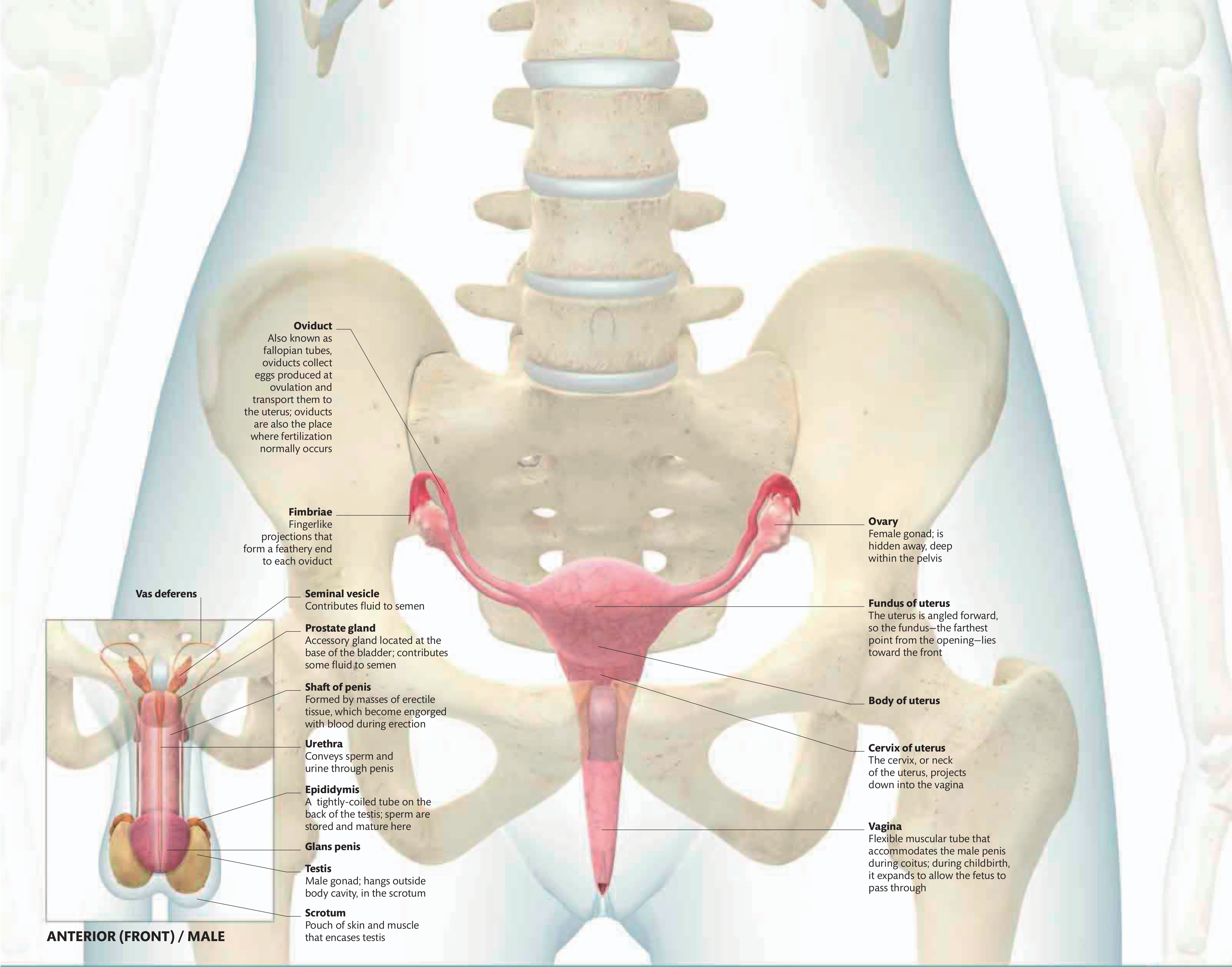

Oviduct
Also known as fallopian tubes, oviducts collect eggs produced at ovulation and transport them to the uterus; oviducts are also the place where fertilization normally occurs

Fimbriae
Fingerlike projections that form a feathery end to each oviduct

Ovary
Female gonad; is hidden away, deep within the pelvis

Fundus of uterus
The uterus is angled forward, so the fundus—the farthest point from the opening—lies toward the front

Body of uterus

Cervix of uterus
The cervix, or neck of the uterus, projects down into the vagina

Vagina
Flexible muscular tube that accommodates the male penis during coitus; during childbirth, it expands to allow the fetus to pass through

Vas deferens

Seminal vesicle
Contributes fluid to semen

Prostate gland
Accessory gland located at the base of the bladder; contributes some fluid to semen

Shaft of penis
Formed by masses of erectile tissue, which become engorged with blood during erection

Urethra
Conveys sperm and urine through penis

Epididymis
A tightly-coiled tube on the back of the testis; sperm are stored and mature here

Glans penis

Testis
Male gonad; hangs outside body cavity, in the scrotum

Scrotum
Pouch of skin and muscle that encases testis

ANTERIOR (FRONT) / MALE

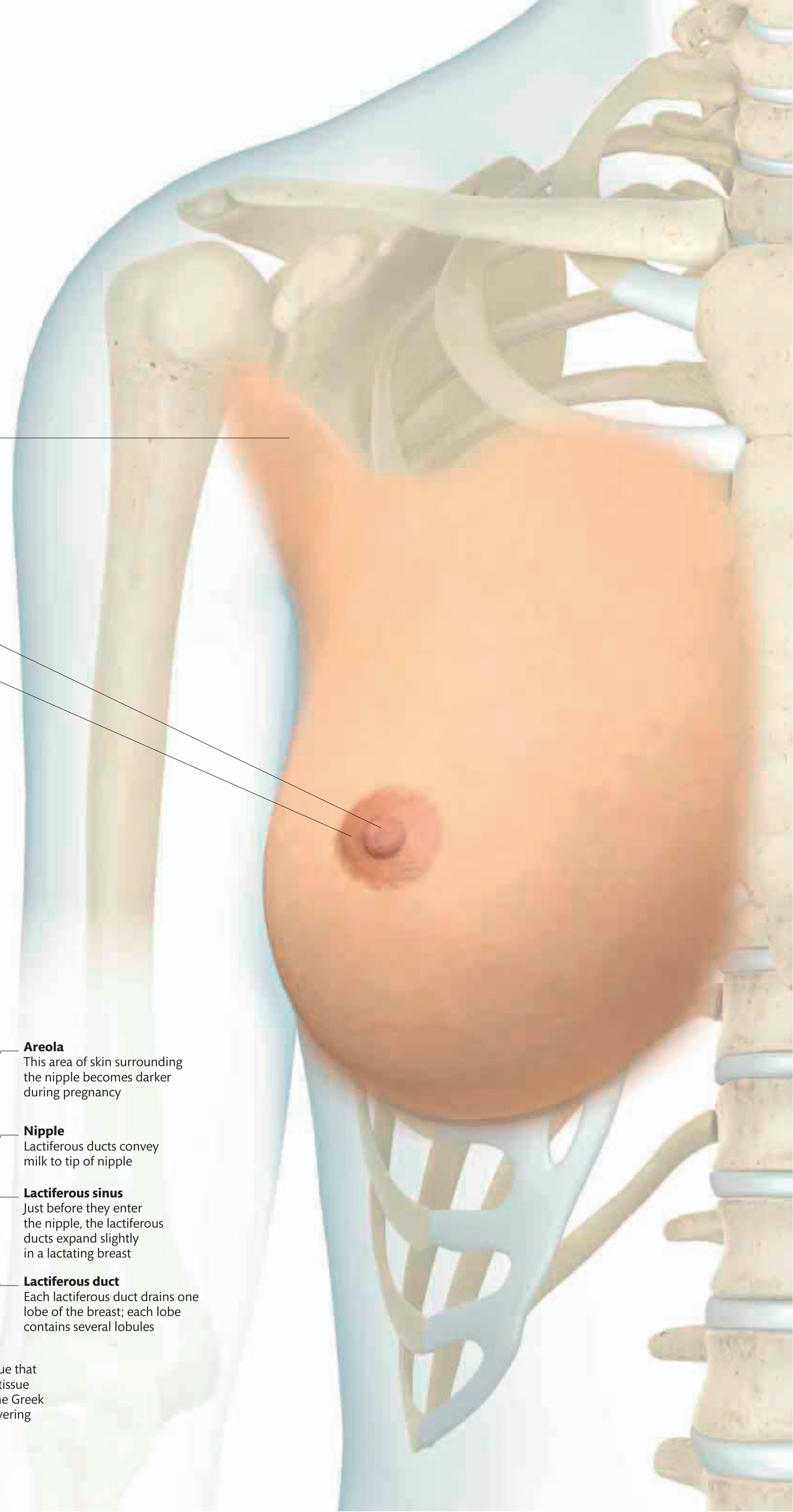

Axillary tail
This part of the illustration shows the extent of the breast; breast tissue continues upward and outward on the chest wall—right up into the axilla (armpit)

Nipple

Areola
Means small area in Latin

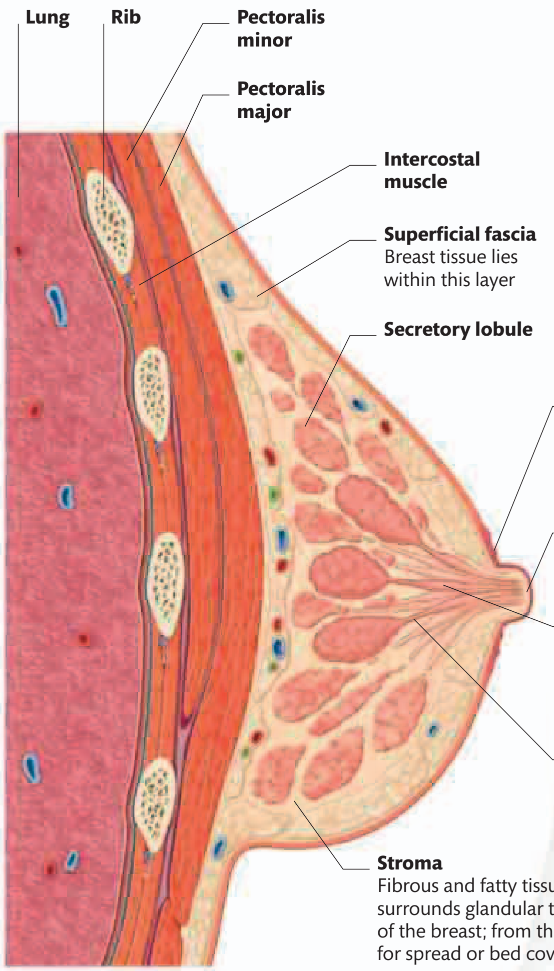

Intercostal muscle

Superficial fascia
Breast tissue lies within this layer

Secretory lobule

Areola
This area of skin surrounding the nipple becomes darker during pregnancy

Nipple
Lactiferous ducts convey milk to tip of nipple

Lactiferous sinus
Just before they enter the nipple, the lactiferous ducts expand slightly in a lactating breast

Lactiferous duct
Each lactiferous duct drains one lobe of the breast; each lobe contains several lobules

Stroma
Fibrous and fatty tissue that surrounds glandular tissue of the breast; from the Greek for spread or bed covering

CROSS SECTION OF BREAST

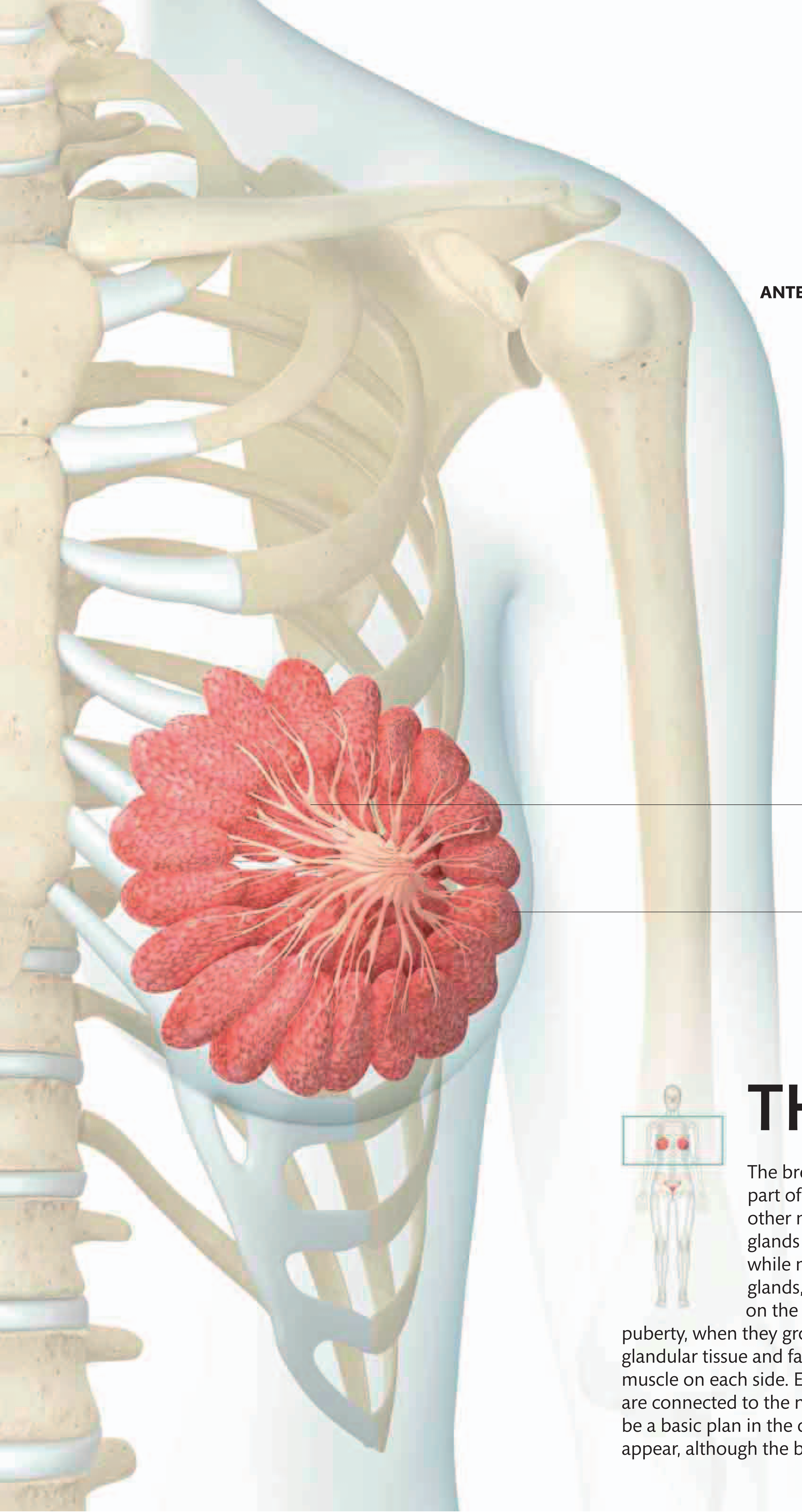

ANTERIOR (FRONT) / FEMALE

Lactiferous duct
From the Latin for milk-carrying duct

Secretory lobule
Lactiferous ducts branch during puberty to form lobules, where milk is produced and secreted

THORAX

The breasts, or mammary glands, are an important part of the reproductive system in women. Like all other mammals, human females have mammary glands to provide the newborn with milk. But while many mammals have multiple mammary glands, humans (and other apes) have just two, on the front of the chest. The breasts develop at puberty, when they grow due to the increased production of glandular tissue and fat. The breasts lie on the pectoralis major muscle on each side. Each breast contains 15 to 20 lobes, which are connected to the nipple by lactiferous ducts. There seems to be a basic plan in the developing embryo, so that male nipples appear, although the breast does not form.

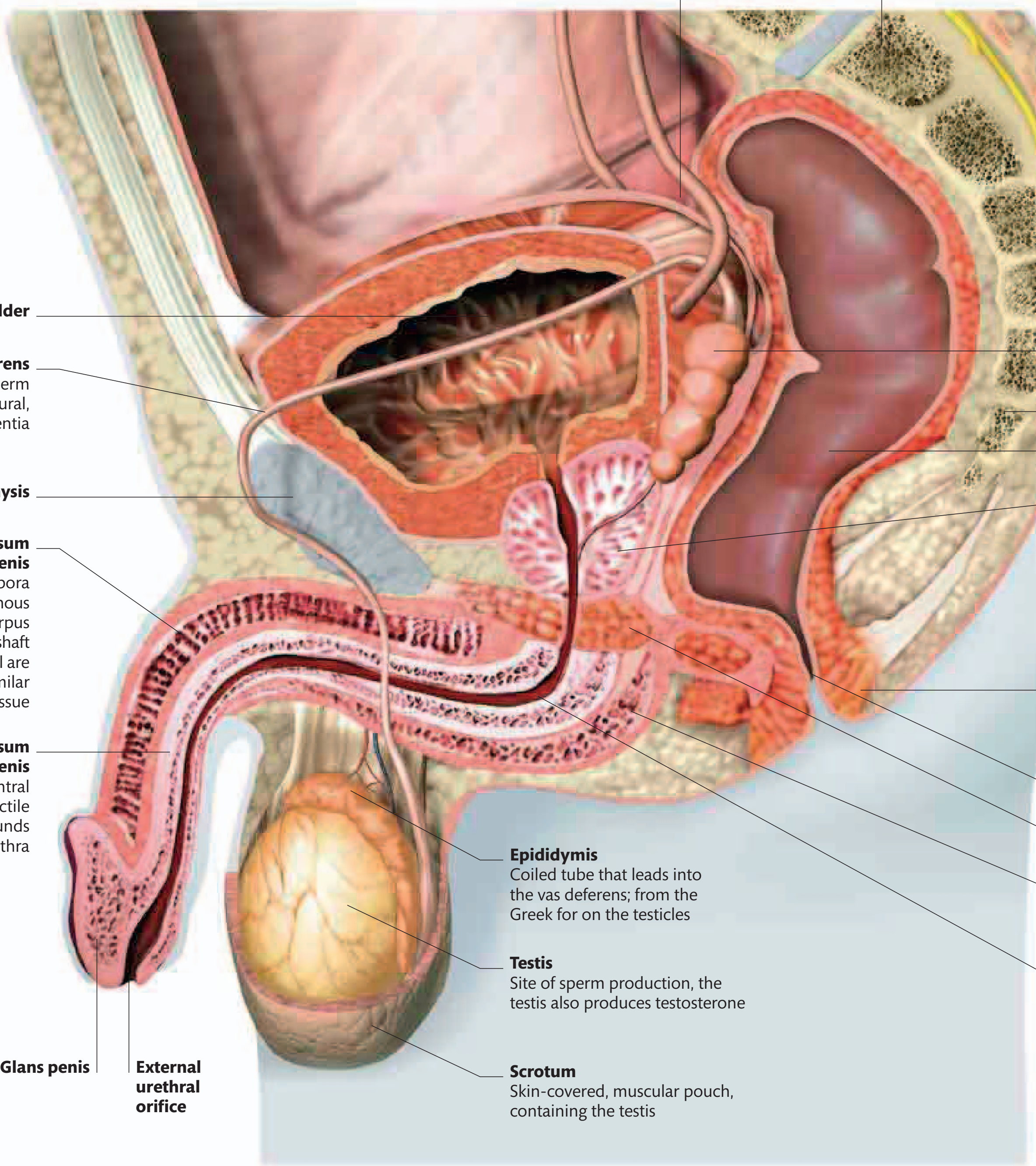

Rectovesical pouch A pocket of peritoneal cavity between the rectum and the bladder, in a man

Vas deferens Conveys sperm from the testis; plural, vasa deferentia

Corpus cavernosum of penis A pair of corpora cavernosa, or cavernous bodies, join the corpus spongiosum in the shaft of the penis; all are made of similar spongy tissue

Corpus spongiosum of penis Spongy, central column of erectile tissue; surrounds the urethra

Prostate gland Lies just in front of the rectum; can enlarge in later life and block the urethra

Epididymis Coiled tube that leads into the vas deferens; from the Greek for on the testicles

Testis Site of sperm production, the testis also produces testosterone

Scrotum Skin-covered, muscular pouch, containing the testis

Bulb of penis Becomes the corpus spongiosum; contains the urethra

Urethra Travels from the bladder, through the prostate, the external urethral sphincter, and the perineal membrane (a sheet of fascia within the perineum); continues through the penis to the external urethral orifice, on the tip of the glans penis

SAGITTAL SECTION / MALE

ABDOMEN AND PELVIS

The male and female reproductive systems are both comprised of a series of internal and external organs, although structurally these are very different. It is true that both sexes possess gonads (ovaries in women and testes in men) and a tract, or set of tubes, but the similarity ends there. When we look in detail at the anatomy of the pelvis in each sex, the differences are obvious. The pelvis of a man contains only part of the reproductive tract, as well as the lower parts of the digestive and urinary tracts, including the rectum and bladder. Beneath the bladder is the prostate gland; this is where the vasa deferentia, which bring sperm from the testis, empty into the urethra. A woman's pelvic cavity contains more of the reproductive tract than a man's. The vagina and uterus are situated between the bladder and rectum in the pelvis.

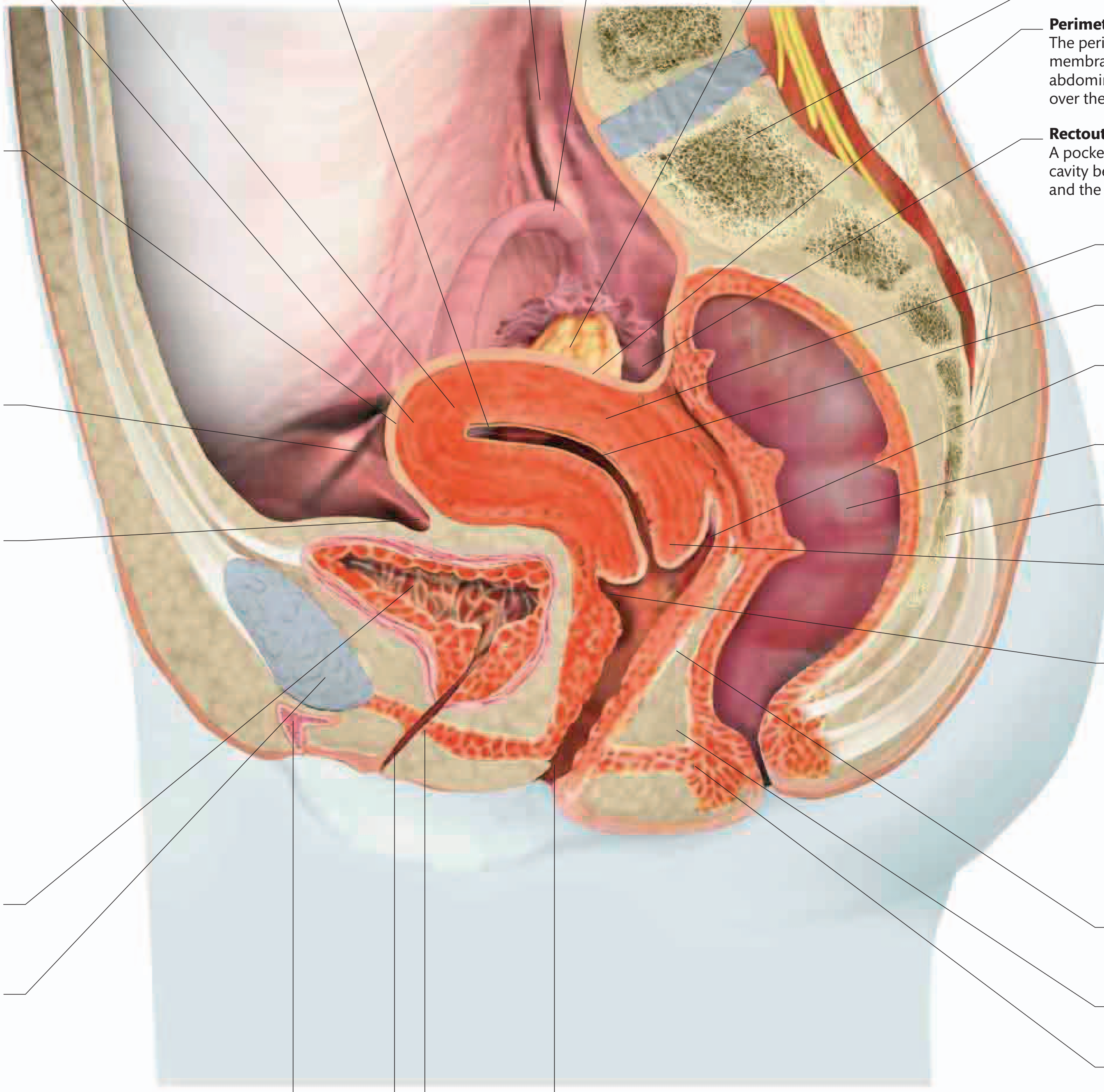

Myometrium
Thick, smooth muscle layer of uterus

Endometrium
Lining of the uterus; innermost layer of endometrium is shed during menstruation. From the Greek for within the womb

Suspensory ligament of ovary
Carries ovarian arteries and veins to and from the ovary

Oviduct
Literally, egg-duct; each one is about 4in (10cm) long

Ovary
Means egg place in Latin; each of the two ovaries lies on the pelvic side wall, in the angle between the internal and external iliac arteries

Fundus of uterus
The top portion of the uterus, farthest from the cervix

Uterus
The uterus is shaped like a flattened pear, and normally lies in the position shown here—bent forward over an empty bladder

Round ligament of the uterus

Vesicouterine pouch
A pocket of peritoneal cavity between the bladder and the uterus. The peritoneal cavity is a potential space between the peritoneum lining the abdominal walls, and the abdominal and pelvic organs

Bladder

Pubic symphysis
Cartilage joint at front of pelvis; softens during pregnancy then widens slightly during childbirth

Sacrum

Perimetrium
The peritoneum (serous membrane lining the abdominal cavity) lies over the uterus

Rectouterine pouch
A pocket of peritoneal cavity between the rectum and the uterus

Body of uterus

Cavity of uterus

Posterior fornix of vagina

Rectum

Coccyx

Cervix of uterus
Literally, the neck of the uterus

Anterior fornix of vagina
Fornices are gutter-like areas that form as the cervix projects down into the vagina, known as anterior, lateral, and posterior fornices; fornix means arch or vault in Latin

Rectovaginal septum

Anal canal

External anal sphincter

Clitoris
Contains spongy, erectile tissue similar to that in the penis

Urethra

External urethral sphincter

Vagina
Tube, around 3½in (9cm) long, with walls of fibrous and muscular tissue

SAGITTAL SECTION / FEMALE

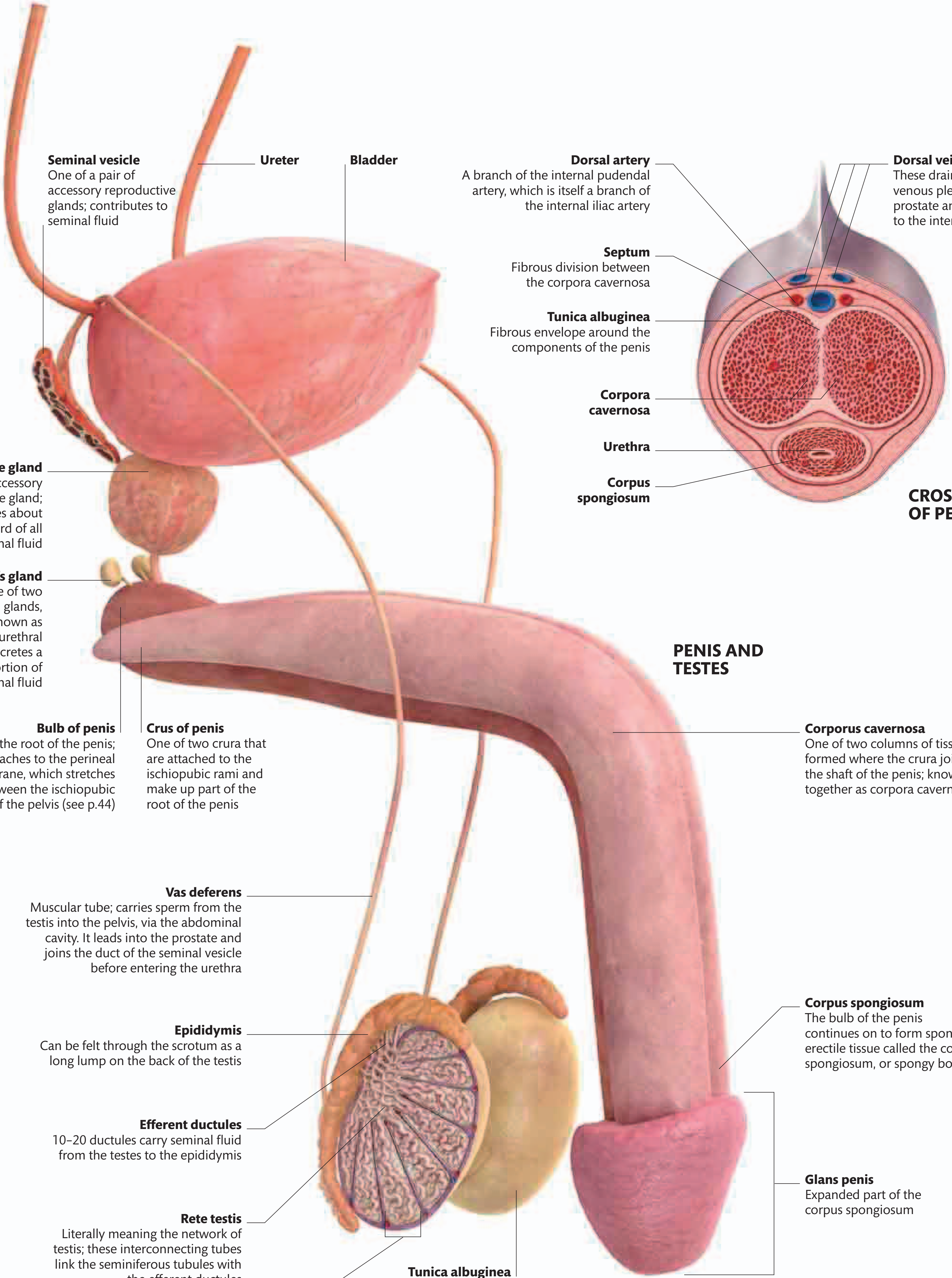

PENIS AND TESTES

Seminal vesicle
One of a pair of accessory reproductive glands; contributes to seminal fluid

Ureter

Bladder

Prostate gland
Accessory reproductive gland; contributes about a third of all seminal fluid

Cowper's gland
One of two pea-sized glands, also known as bulbourethral glands; secretes a tiny portion of seminal fluid

Bulb of penis
Part of the root of the penis; attaches to the perineal membrane, which stretches between the ischiopubic rami of the pelvis (see p.44)

Crus of penis
One of two crura that are attached to the ischiopubic rami and make up part of the root of the penis

Corporus cavernosa
One of two columns of tissue, formed where the crura join at the shaft of the penis; known together as corpora cavernosa

Vas deferens
Muscular tube; carries sperm from the testis into the pelvis, via the abdominal cavity. It leads into the prostate and joins the duct of the seminal vesicle before entering the urethra

Epididymis
Can be felt through the scrotum as a long lump on the back of the testis

Efferent ductules
10–20 ductules carry seminal fluid from the testes to the epididymis

Rete testis
Literally meaning the network of testis; these interconnecting tubes link the seminiferous tubules with the efferent ductules

Lobule of testis
200–300 lobules lie in each testis; each contains 1–3 tightly-packed seminiferous tubules, in which sperm are made

Tunica albuginea
Literally meaning egg-white-like coat; the outer covering of the testis

Corpus spongiosum
The bulb of the penis continues on to form spongy erectile tissue called the corpus spongiosum, or spongy body

Glans penis
Expanded part of the corpus spongiosum

CROSS SECTION OF PENIS

Dorsal artery
A branch of the internal pudendal artery, which is itself a branch of the internal iliac artery

Dorsal veins
These drain up to the venous plexus around the prostate and, ultimately, to the internal iliac veins

Septum
Fibrous division between the corpora cavernosa

Tunica albuginea
Fibrous envelope around the components of the penis

Corpora cavernosa

Urethra

Corpus spongiosum

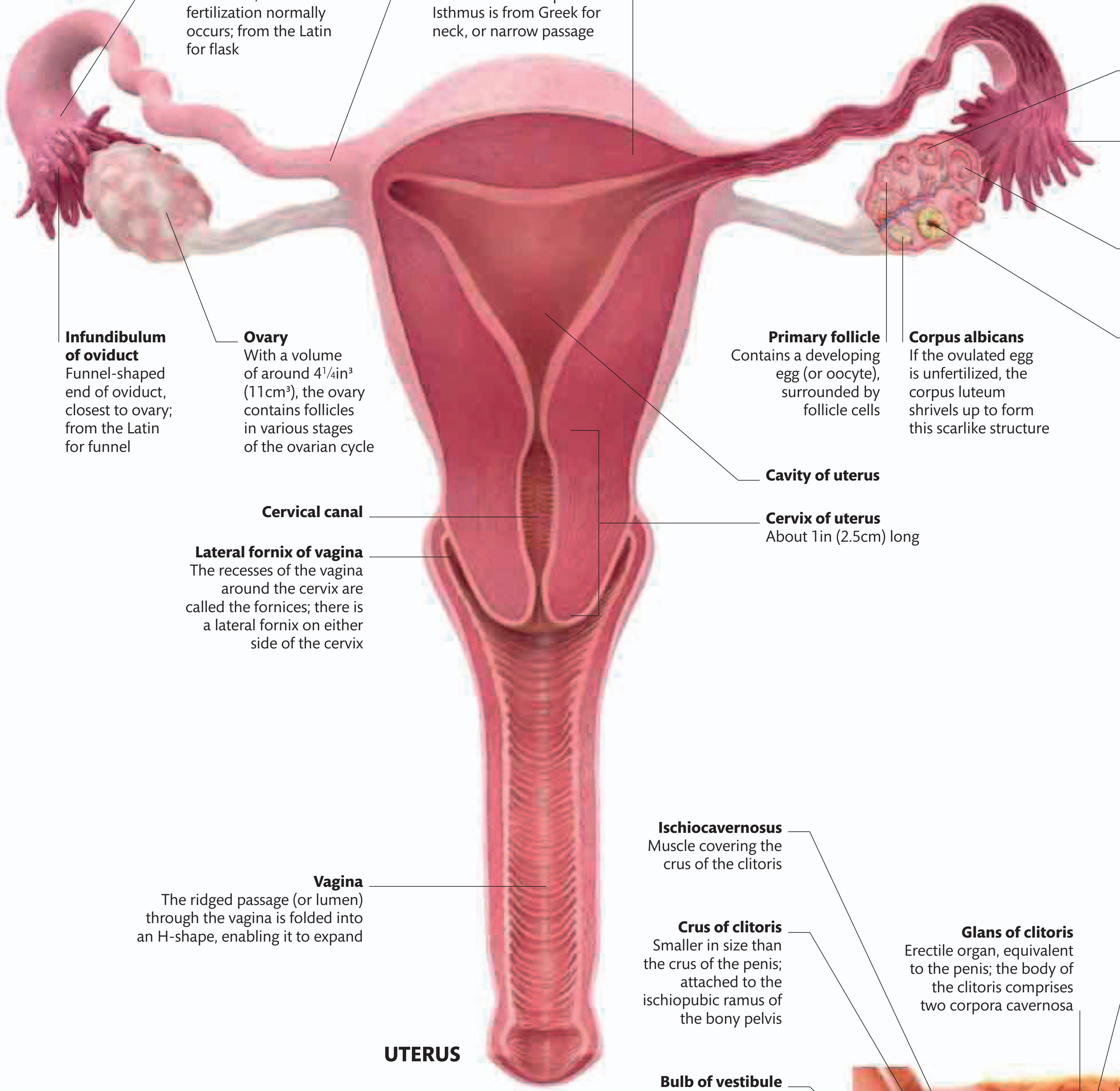
Ampulla of oviduct Slightly widened part of oviduct, where fertilization normally occurs; from the Latin for flask
Isthmus of oviduct Last third of oviduct; narrower than ampulla. Isthmus is from Greek for neck, or narrow passage
Body of uterus
Secondary follicle Follicle that has started to accumulate fluid
Fimbriae Fingerlike projections, help pick up the ovulated egg. From the Latin for fringe
Mature follicle Filled with fluid; this will burst to release the egg at ovulation
Corpus luteum Remains of follicle after ovulation, literally meaning yellow body
Infundibulum of oviduct Funnel-shaped end of oviduct, closest to ovary; from the Latin for funnel
Ovary With a volume of around $4\frac{1}{4}$in^3 (11cm^3), the ovary contains follicles in various stages of the ovarian cycle
Primary follicle Contains a developing egg (or oocyte), surrounded by follicle cells
Corpus albicans If the ovulated egg is unfertilized, the corpus luteum shrivels up to form this scarlike structure
Cavity of uterus
Cervical canal
Cervix of uterus About 1in (2.5cm) long
Lateral fornix of vagina The recesses of the vagina around the cervix are called the fornices; there is a lateral fornix on either side of the cervix
Vagina The ridged passage (or lumen) through the vagina is folded into an H-shape, enabling it to expand

UTERUS

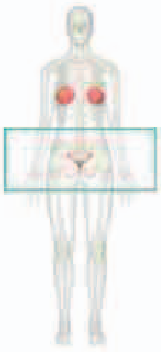

ABDOMEN AND PELVIS

At a very fundamental level, the reproductive systems of man and woman must work together to allow eggs and sperm to meet. These views of the isolated organs and reproductive tracts show clearly how the anatomy is arranged to achieve this. The ovaries, where eggs (or ova) are produced, are deep inside the female pelvis. The eggs are collected from the ovaries by a pair of tubes, the oviducts, and it is usually here that fertilization takes place. The fertilized egg then moves along the oviduct, dividing into a ball of cells. The embryo eventually reaches the uterus, which is designed to accommodate and support the growing fetus. The vagina provides both a way for sperm to get in, and the route for the baby to get out at birth.

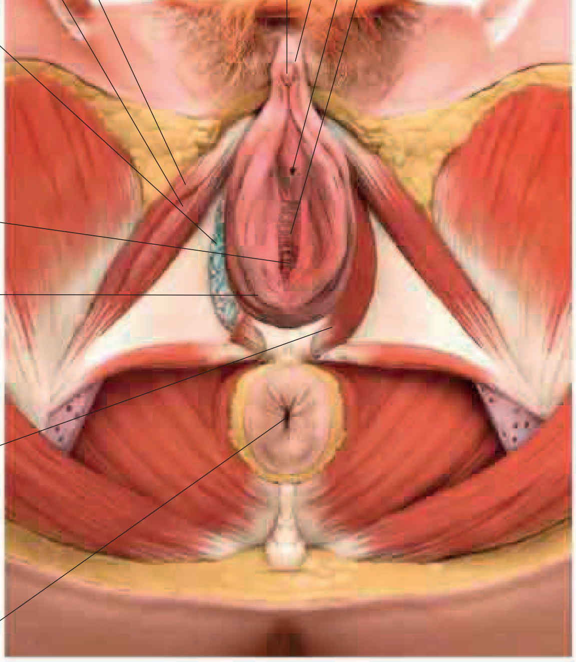
Ischiocavernosus Muscle covering the crus of the clitoris
Crus of clitoris Smaller in size than the crus of the penis; attached to the ischiopubic ramus of the bony pelvis
Glans of clitoris Erectile organ, equivalent to the penis; the body of the clitoris comprises two corpora cavernosa
Prepuce, overlying body of clitoris
External urethral orifice
Vestibule Area between the labia minora; Latin for entrance court
Bulb of vestibule One of a pair of structures equivalent to the single bulb of the penis; made of spongy erectile tissue
Vaginal orifice
Labia minora Folds of skin either side of the vestibule; singular is labium minus
Bulbospongiosus Muscle covering the bulb of vestibule; helps increase pressure in the underlying spongy tissue
Anus

EXTERNAL FEMALE GENITALIA

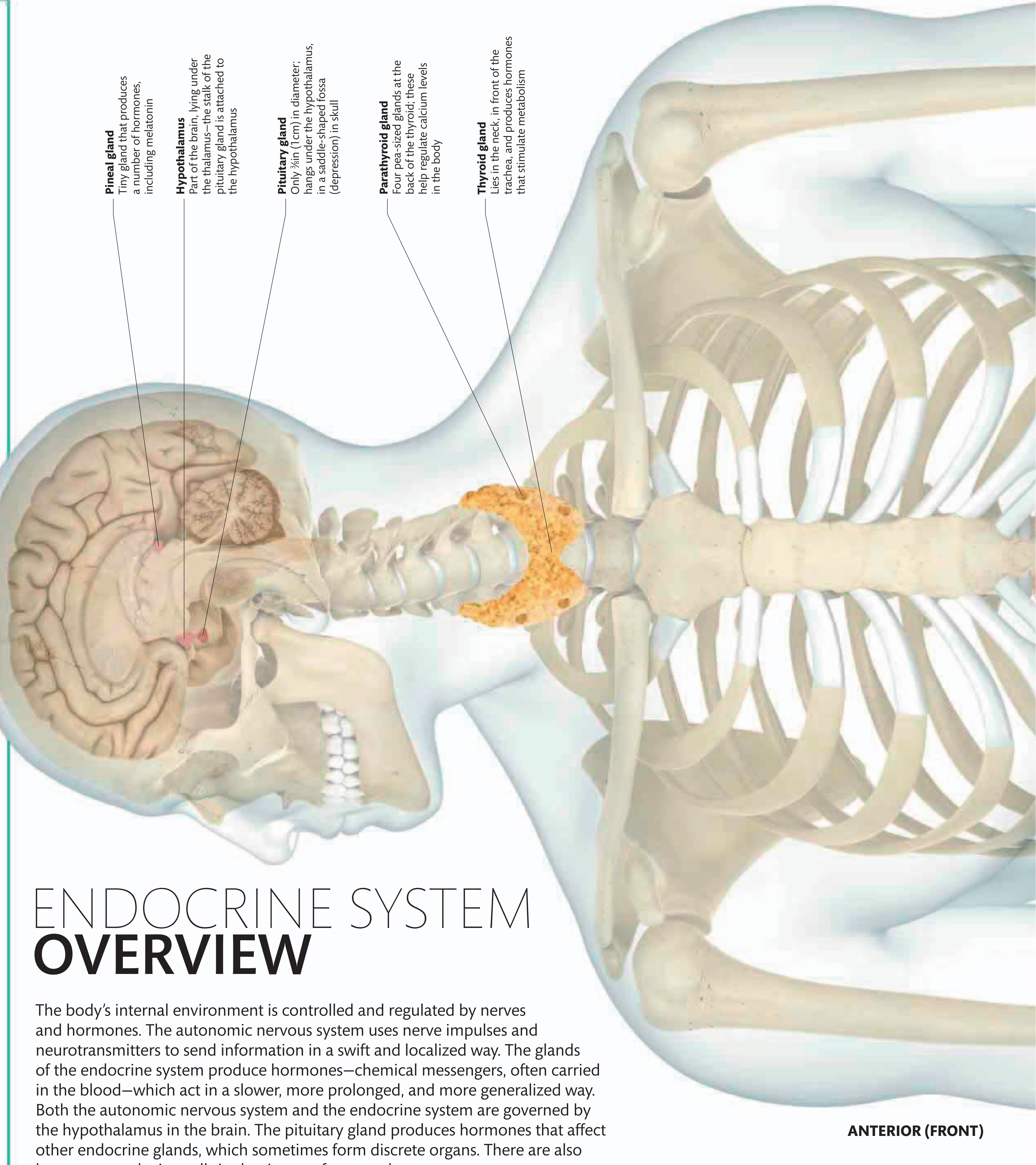
Pineal gland Tiny gland that produces a number of hormones, including melatonin
Hypothalamus Part of the brain, lying under the thalamus—the stalk of the pituitary gland is attached to the hypothalamus
Pituitary gland Only ⅜in (1cm) in diameter; hangs under the hypothalamus, in a saddle-shaped fossa (depression) in skull
Parathyroid gland Four pea-sized glands at the back of the thyroid; these help regulate calcium levels in the body
Thyroid gland Lies in the neck, in front of the trachea, and produces hormones that stimulate metabolism
ANTERIOR (FRONT)

ENDOCRINE SYSTEM **OVERVIEW**

The body's internal environment is controlled and regulated by nerves and hormones. The autonomic nervous system uses nerve impulses and neurotransmitters to send information in a swift and localized way. The glands of the endocrine system produce hormones—chemical messengers, often carried in the blood—which act in a slower, more prolonged, and more generalized way. Both the autonomic nervous system and the endocrine system are governed by the hypothalamus in the brain. The pituitary gland produces hormones that affect other endocrine glands, which sometimes form discrete organs. There are also hormone-producing cells in the tissues of many other organs.

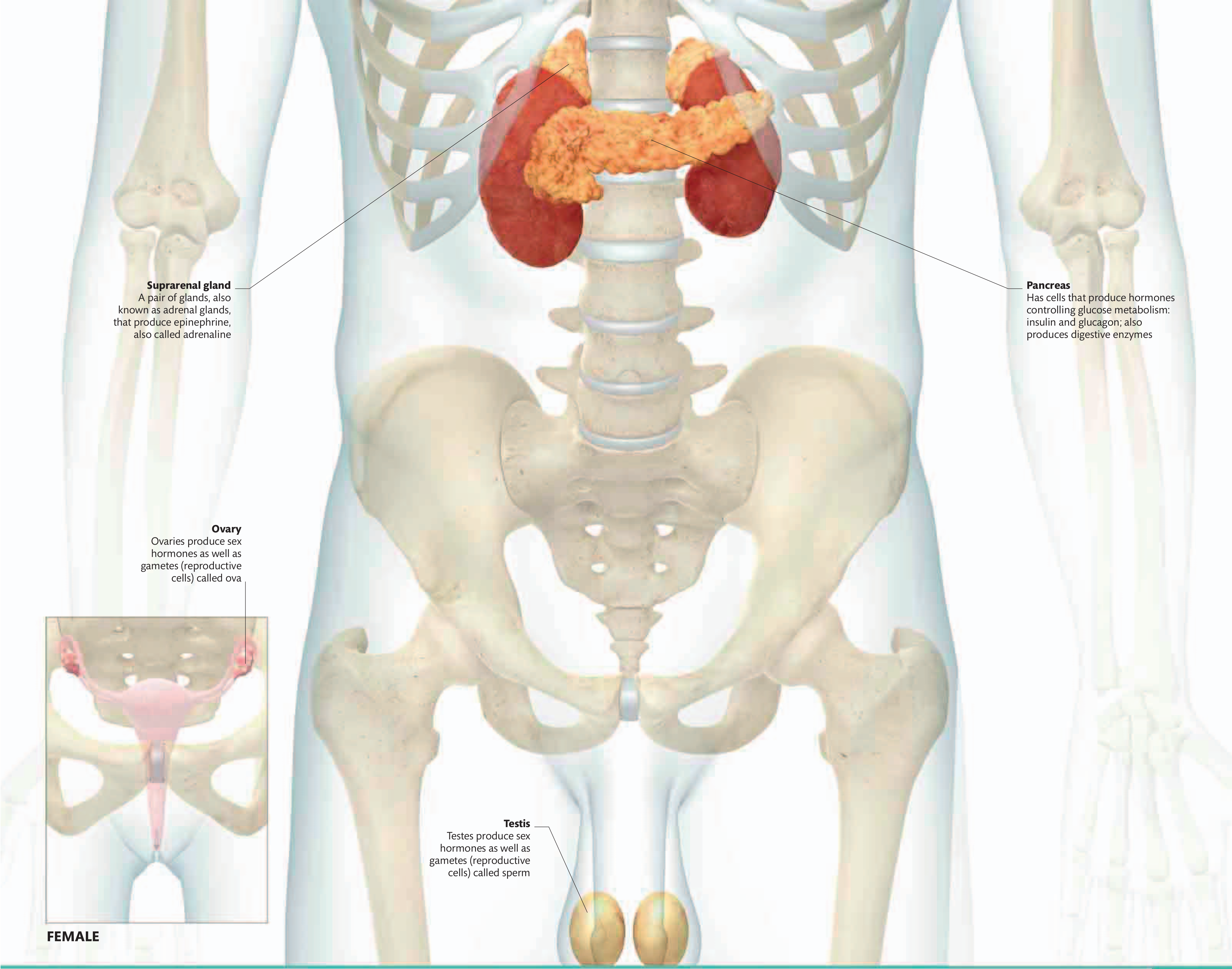
Suprarenal gland
A pair of glands, also known as adrenal glands, that produce epinephrine, also called adrenaline
Pancreas
Has cells that produce hormones controlling glucose metabolism: insulin and glucagon; also produces digestive enzymes
Ovary
Ovaries produce sex hormones as well as gametes (reproductive cells) called ova
Testis
Testes produce sex hormones as well as gametes (reproductive cells) called sperm
FEMALE

HEAD AND NECK

The insides of our bodies are regulated by the autonomic nervous and endocrine systems. There is overlap between these two systems, and their functions are integrated and controlled within the hypothalamus of the brain. The pituitary gland has two lobes; its posterior lobe develops as a direct extension of the hypothalamus. Both lobes of the pituitary gland secrete hormones into the bloodstream, in response to nerve signals or blood-borne releasing factors from the hypothalamus. Many of the pituitary hormones act on other endocrine glands, including the thyroid gland in the neck, the suprarenal glands on top of the kidneys, and the ovaries or testes.

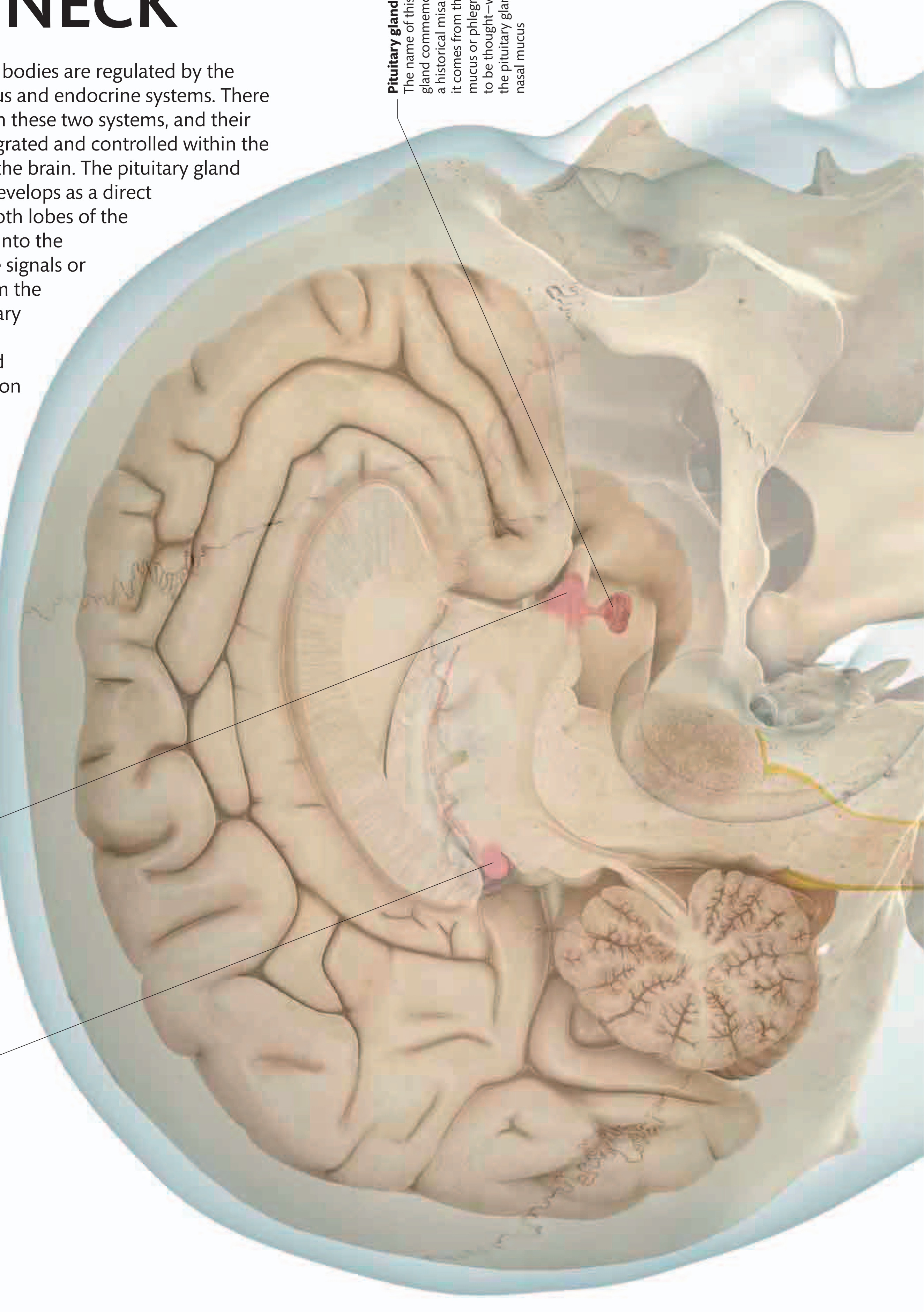

Pituitary gland The name of this gland commemorates a historical misapprehension: it comes from the Latin for mucus or phlegm, as it used to be thought—wrongly—that the pituitary gland secreted nasal mucus

Hypothalamus

Pineal gland This tiny gland is about $\frac{5}{16}$in (8mm) long, and shaped a bit like a pine nut; it has links to the visual pathway, and is involved in regulating circadian rhythms—the daily sleep-wake cycle

SIDE

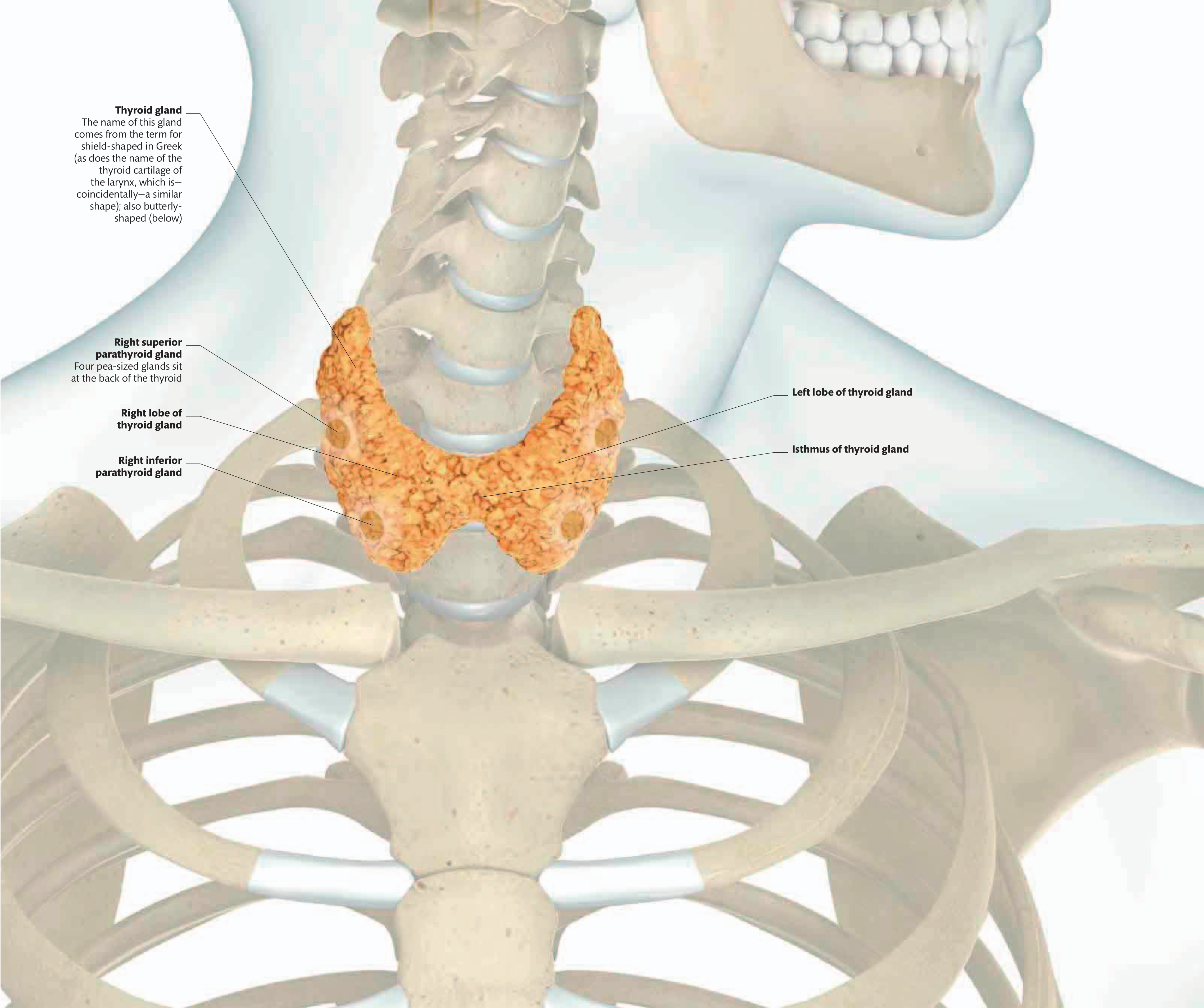

Thyroid gland
The name of this gland comes from the term for shield-shaped in Greek (as does the name of the thyroid cartilage of the larynx, which is—coincidentally—a similar shape); also butterfly-shaped (below)

Right superior parathyroid gland
Four pea-sized glands sit at the back of the thyroid

Right lobe of thyroid gland
Right inferior parathyroid gland
Left lobe of thyroid gland
Isthmus of thyroid gland

03

Imaging the Body

The human body is a “living machine” with many complex working parts. To understand how the body functions, and to cure the various ailments that afflict it, it is crucial for medical professionals to examine it in detail. Advances in technology have made it possible to view human anatomy without dissecting the body. Techniques such as magnetic resonance imaging (MRI) reveal the inside of the body with great accuracy and allow us to build up a complete picture of our anatomy from every possible angle.

222 Imaging techniques
224 Head and neck
226 Thorax
228 Abdomen and pelvis
230 Lower arm and hand
232 Lower limb and foot

IMAGING TECHNIQUES

Imaging is vital to diagnose illness, unravel disease processes, and evaluate treatments. Modern techniques provide detailed information with minimum discomfort to the patient and have largely replaced surgery in establishing the presence and extent of disease. Imaging has also helped advance biological research.

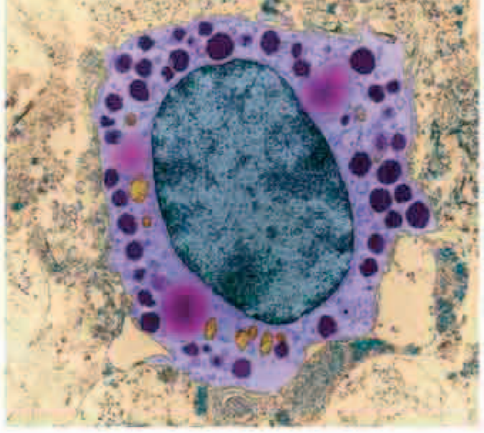

TEM of mast cell
Magnifications of several million times are possible using transmission electron microscopy (TEM). This image shows a mast cell with granules (dark purple) that it releases when it is damaged or is fighting microbes.

The invention of the X-ray in 1895 made the development of noninvasive medicine possible. Without the ability to see inside the body, many internal disorders could only be found after major surgery. Computerized imaging now helps doctors make early diagnoses, which at times greatly increase the likelihood of recovery. Computers process and enhance raw data to aid our visual ability. However, sometimes direct observation is essential. Viewing techniques have also become less invasive with the development of instruments such as the endoscope (see opposite).

MICROSCOPY

Light microscopy (LM) uses magnifying lenses to focus light rays. The light passes through a thin section of material and enlarges it up to 2,000 times. Higher magnifications are achieved with beams of electrons (subatomic particles). In scanning electron microscopy (SEM), the beam runs across a specimen coated with gold film and bounces off the surface to create a three-dimensional image.

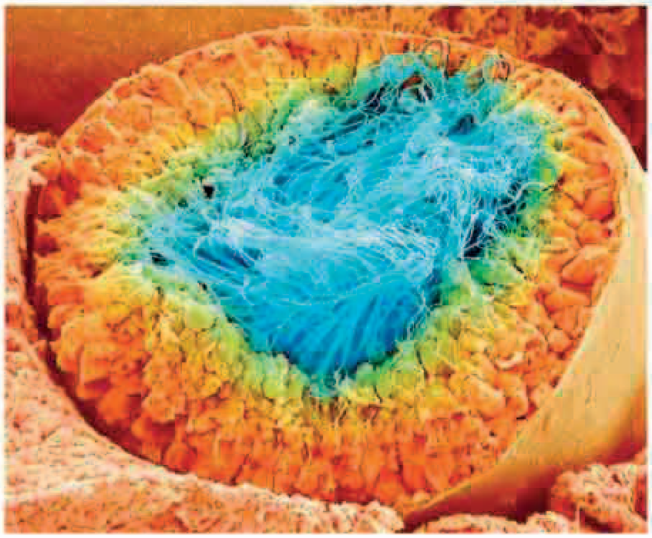

SEM of seminiferous tubule
This freeze-fracture image—in which the specimen is frozen and then cracked open before being scanned—shows sperm heads buried in Sertoli cells (orange) with the tails (blue) projecting into the tubule's lumen.

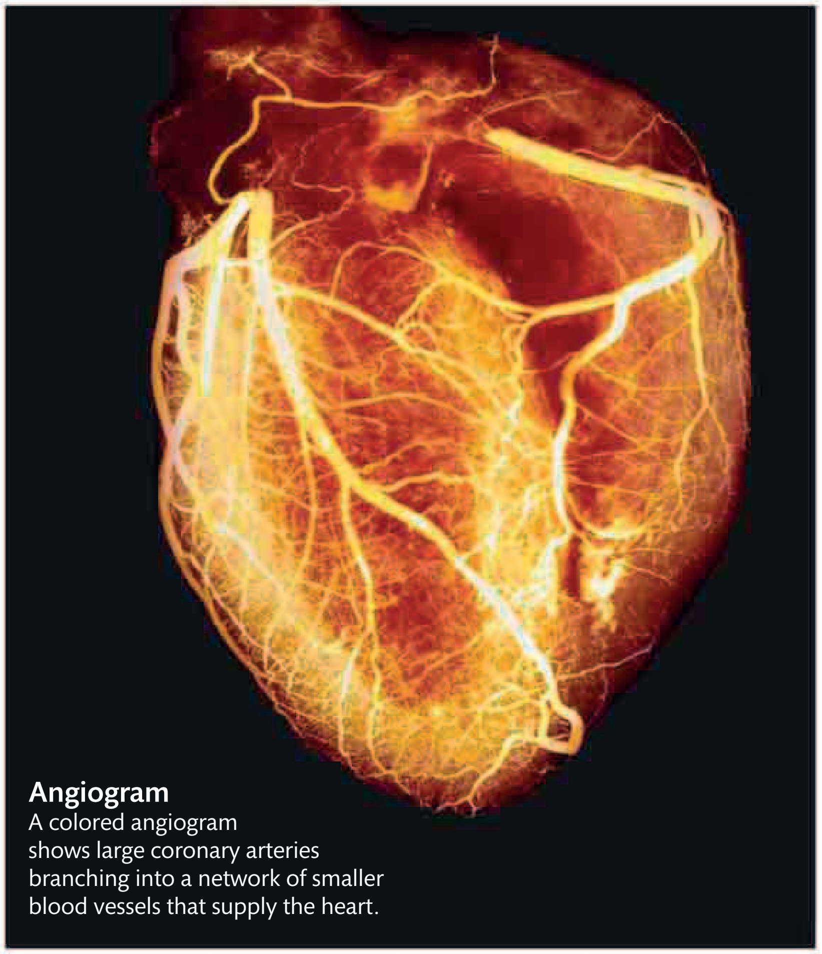

Angiogram
A colored angiogram shows large coronary arteries branching into a network of smaller blood vessels that supply the heart.

X-RAY

Like light rays, X-rays are electromagnetic energy, but of short wavelength. When passed through the body to strike photographic film, they create shadow images (radiographs). Dense structures, such as bone, absorb more X-rays and show up as white, while soft tissues, such as muscle, appear gray. Air-filled spaces, such as the lungs, appear black. The spaces inside the digestive tract, or within blood vessels, may be visualized by filling them with a contrast medium—such as iodine or barium—that absorbs X-rays. A contrast X-ray image of blood vessels is known as an angiogram.

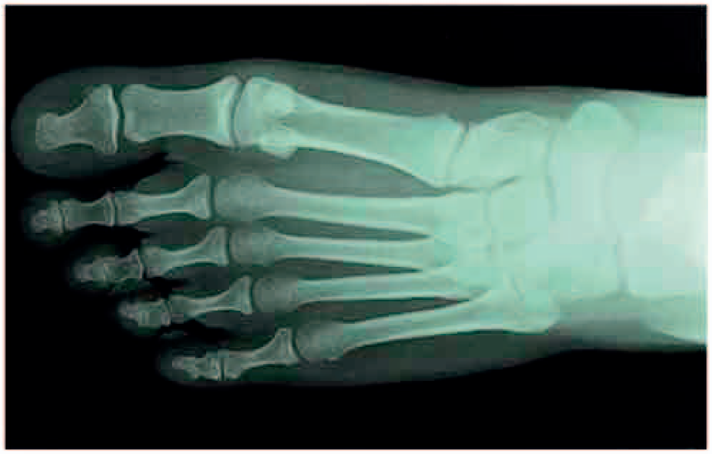

X-ray of foot
This X-ray image shows the foot bones of an adult from above. X-rays are especially useful for viewing dense tissue, such as bone.

RADIONUCLIDE AND PET SCANNING

In radionuclide imaging, a radioactive substance is injected into the body and is absorbed by the area to be imaged. As the substance decays, it emits gamma rays, which a computer forms into an image. Positron emission tomography (PET) is a type of radionuclide scanning where the injected chemical emits radioactive particles called positrons. PET gives data about how the brain functions rather than anatomy.

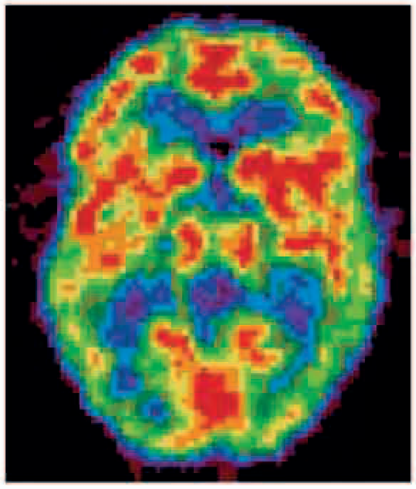

PET brain scan
This scan of the brain shows that the organ is active even in sleep mode. Areas in red, orange, and yellow represent high levels of activity.

ULTRASOUND

In ultrasound, a device called a transducer emits very high-frequency sound waves as it is passed over the body part being examined. The sound waves echo back to the device based on the density of the tissues they encounter. A computer analyses the reflected waves and creates an image.

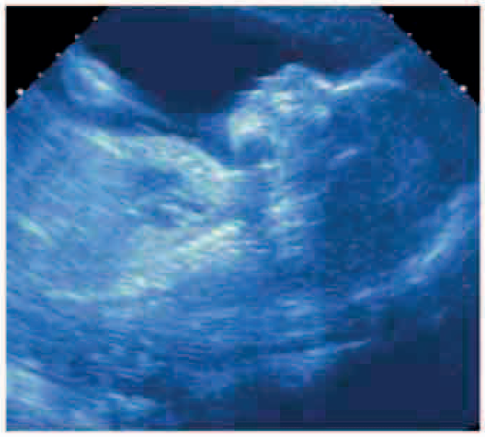

Fetal ultrasound
Low-intensity ultrasound is a safe way to monitor fetal development. In this scan the fetus's head can be seen clearly in profile on the right.

ENDOSCOPY

Telescopelike endoscopes are inserted through natural orifices or incisions to image the body's interior. They can be bent and controlled, and may carry instruments for other purposes as well, such as surgery and biopsy. Endoscopes have been designed to fit various body parts—bronchoscope for airways, gastroscope for the stomach, laparoscope for abdomen, and proctoscope for lower bowel.

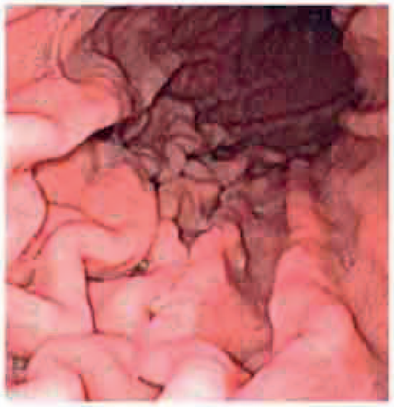

Endoscopic view of stomach
The gastric mucosa (inner lining) of a healthy stomach as seen through an endoscope. This procedure may be performed to investigate upper digestive tract disorders.

MRI AND CT SCANNING

Computerized tomography (CT) and magnetic resonance imaging (MRI) detail various tissue types. In CT, a scanner using X-rays rotates around the patient as a computer records the levels of electromagnetic energy passing through tissues of different densities. A cross section is built from layers of data. In MRI, a person lies in a magnetic chamber that causes hydrogen atoms in the body to align. A pulse of radiowaves is released, throwing the atoms out of alignment. As they realign, the atoms emit radio signals that are used to create an image.

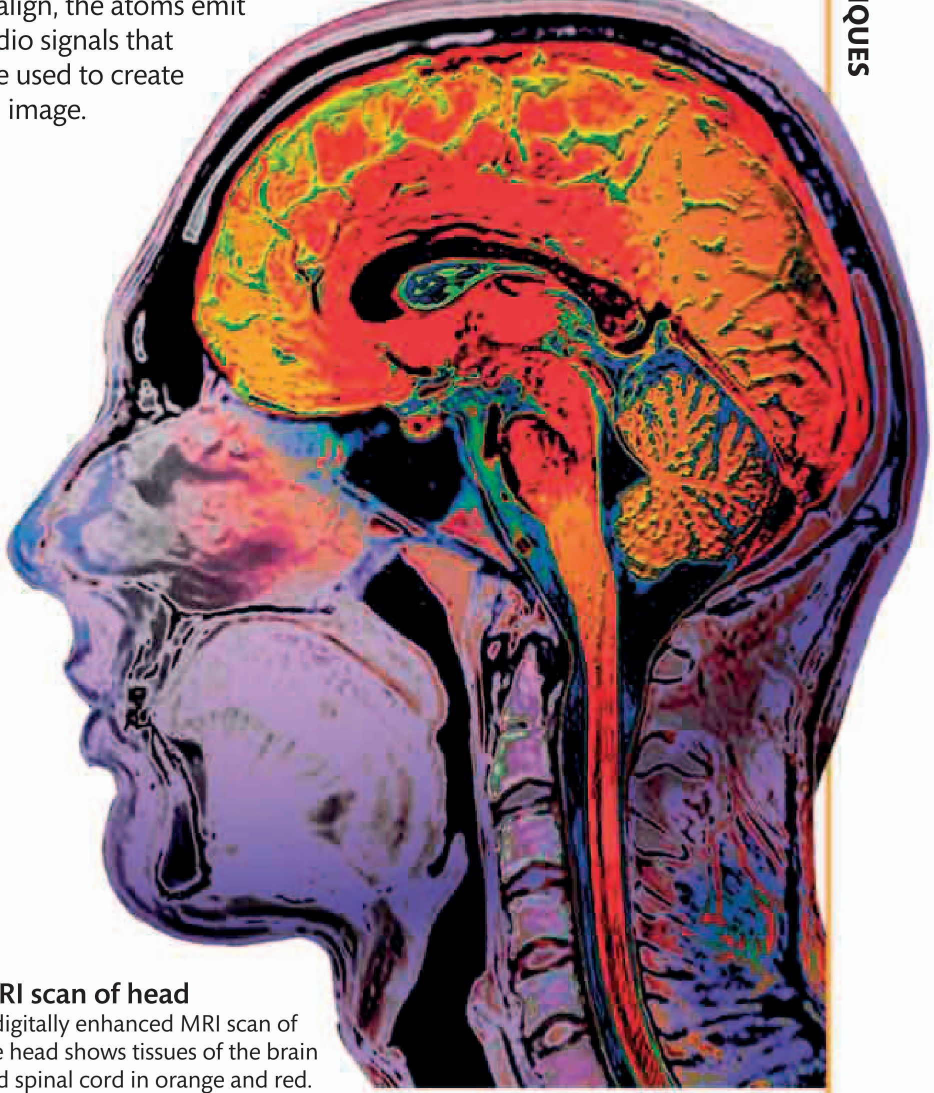

MRI scan of head
A digitally enhanced MRI scan of the head shows tissues of the brain and spinal cord in orange and red.

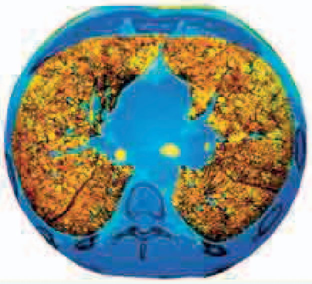

CT scan of lungs
In a horizontal slice through the chest, the spongy tissues and airways of the healthy lungs (oranges and yellows) show up clearly. The heart and major blood vessels between the lungs are mid-blue; the vertebrae, ribs, and sternum are dark blue.

ELECTRICAL ACTIVITY

Monitoring electrical activity in the body can reveal whether it is functioning normally. Signals coming from muscles and nerves are detected by applying sensor pads to the skin. The signals are sent to a computer that coordinates, amplifies, and displays them as a real-time trace—usually a spiked or wavy line. Examples of this technique include electrocardiography (ECG) of the heart, electromyography (EMG) of skeletal muscles, and electroencephalography (EEG) of the brain's nerve activity.

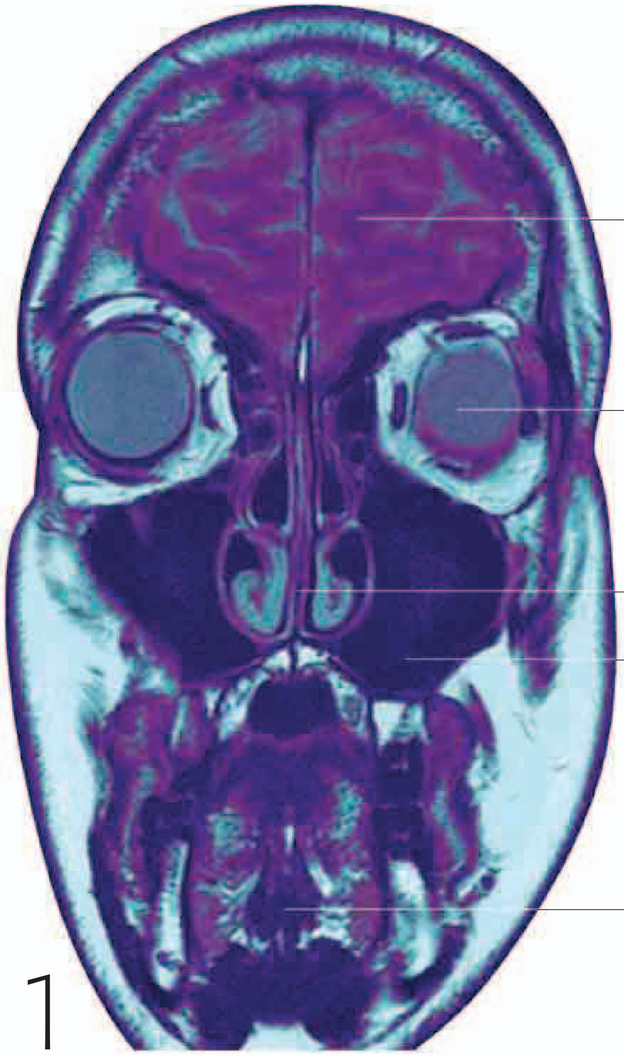
Cerebral cortex
Eye
Nasal septum
Maxillary sinus
Tongue
1

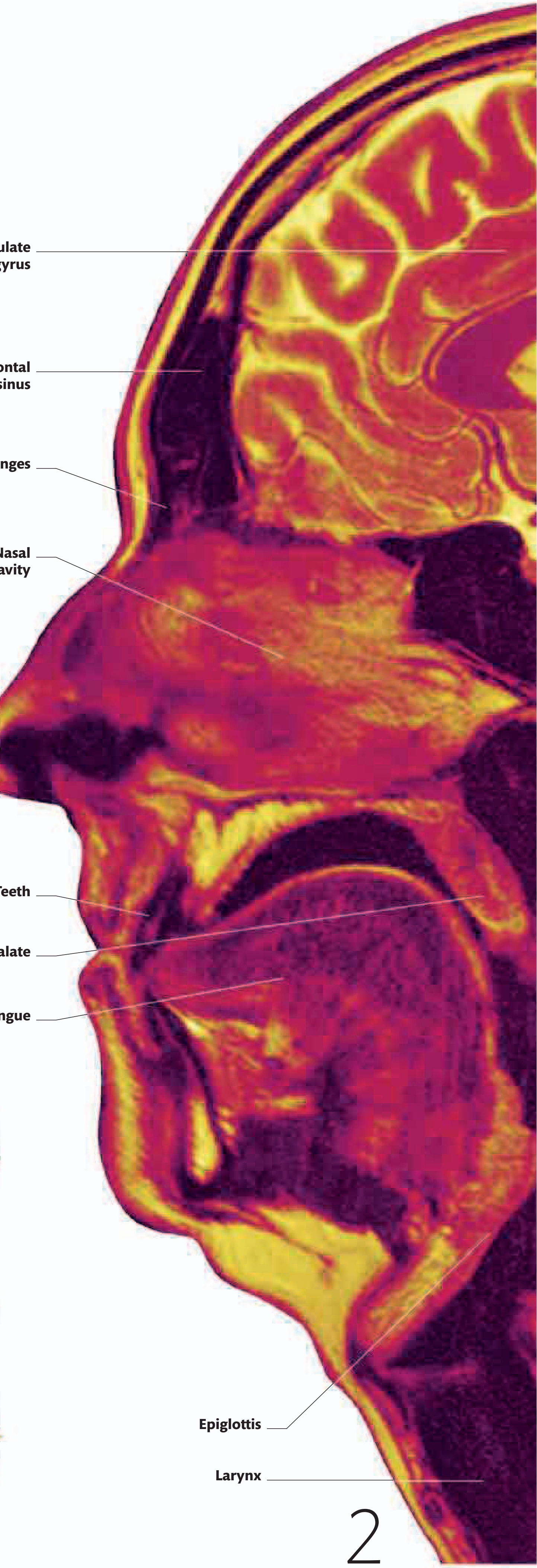
Cingulate gyrus
Frontal sinus
Meninges
Nasal cavity
Teeth
Soft palate
Tongue
Epiglottis
Larynx
2

HEAD AND NECK

The discovery of X-rays at the end of the 19th century suddenly created the possibility of looking inside the human body—without having to physically cut it open. Medical imaging is now an important diagnostic tool, as well as being used for the study of normal anatomy and physiology. In computed tomography (CT), X-rays are used to produce virtual sections or slices through the body. Another form of sectional imaging, using magnetic fields rather than X-rays to create images, is magnetic resonance imaging (MRI), as shown here. MRI is very useful for looking in detail at soft tissue, for instance, muscle, tendons, and the brain. Also seen clearly in these sections are the eyes (1 and 3), the tongue (1 and 2), the larynx, vertebrae, and spinal cord (2 and 5).

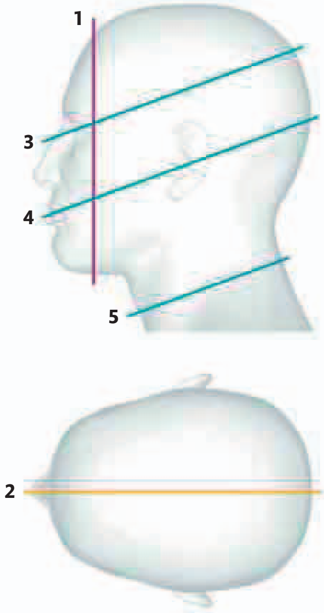

LEVELS OF SCANS

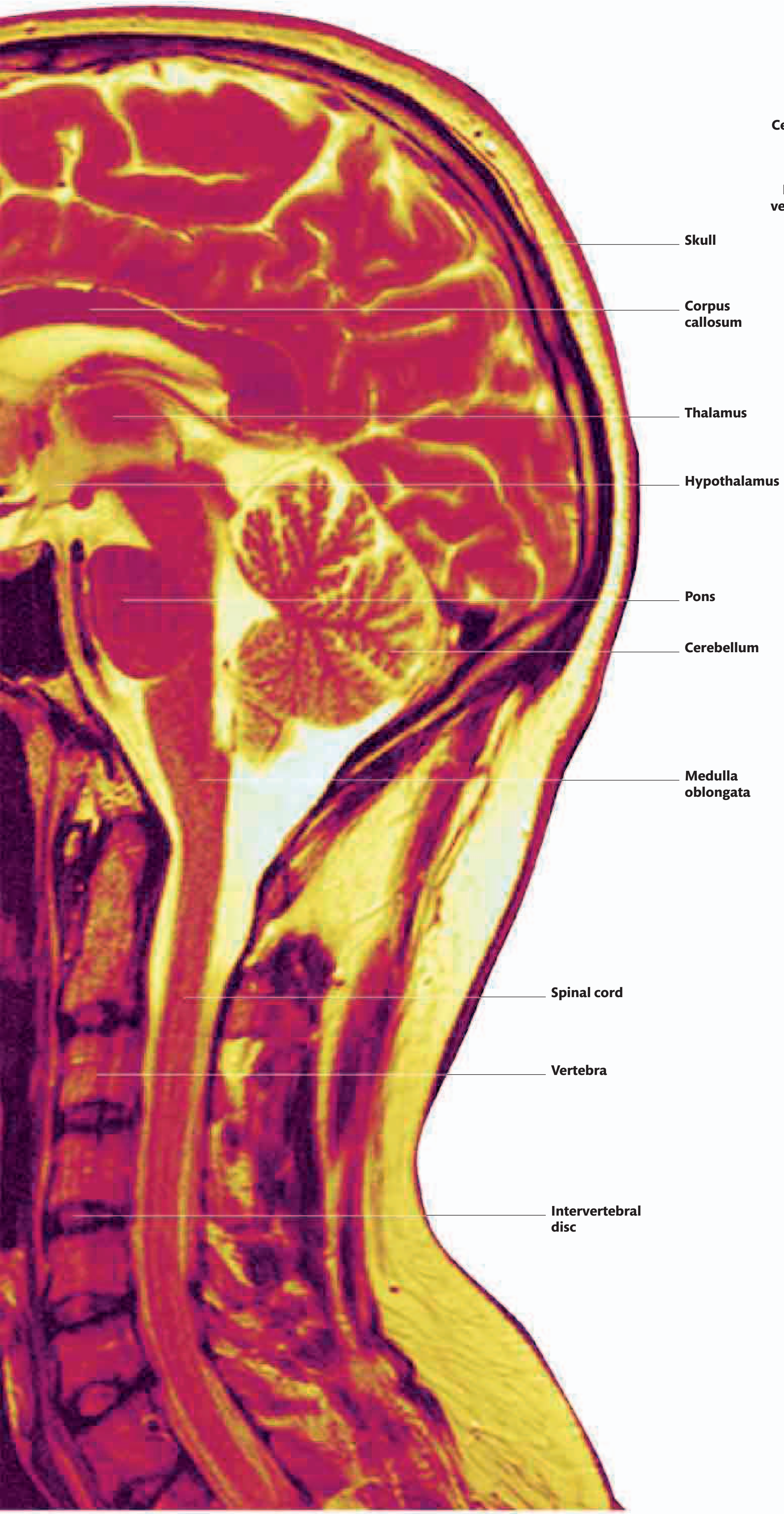
Skull
Corpus callosum
Thalamus
Hypothalamus
Pons
Cerebellum
Medulla oblongata
Spinal cord
Vertebra
Intervertebral disc

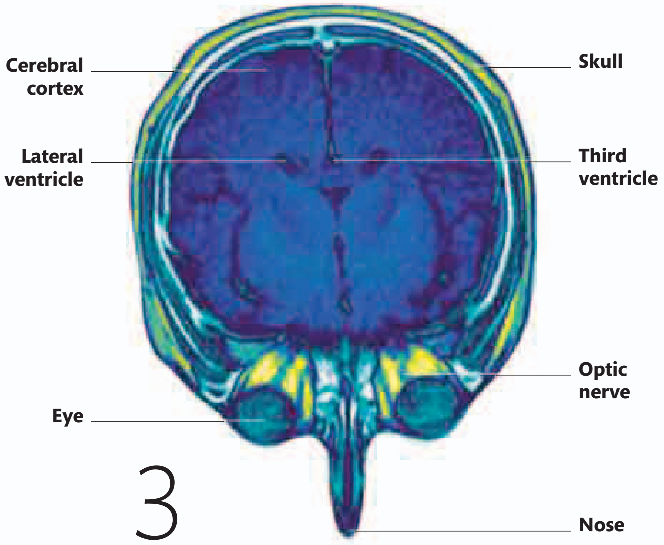
Cerebral cortex
Lateral ventricle
Eye
Skull
Third ventricle
Optic nerve
Nose

3

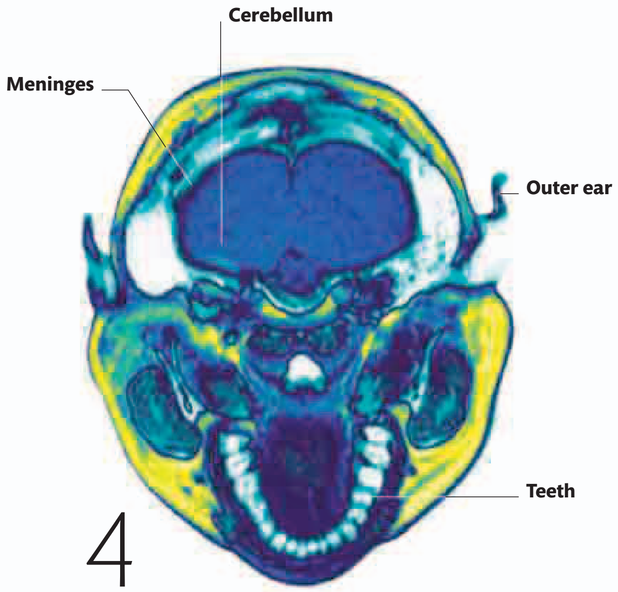
Cerebellum
Meninges
Outer ear
Teeth

4

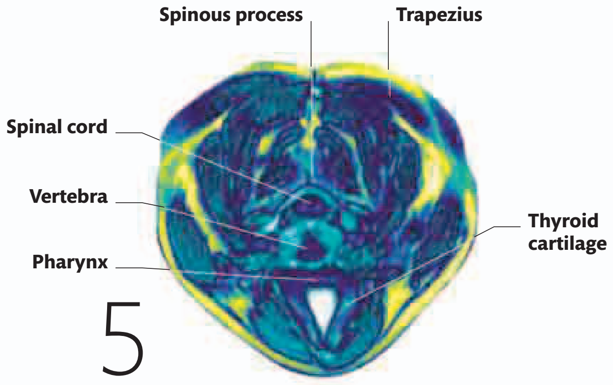
Spinous process
Trapezius
Spinal cord
Vertebra
Pharynx
Thyroid cartilage

5

1

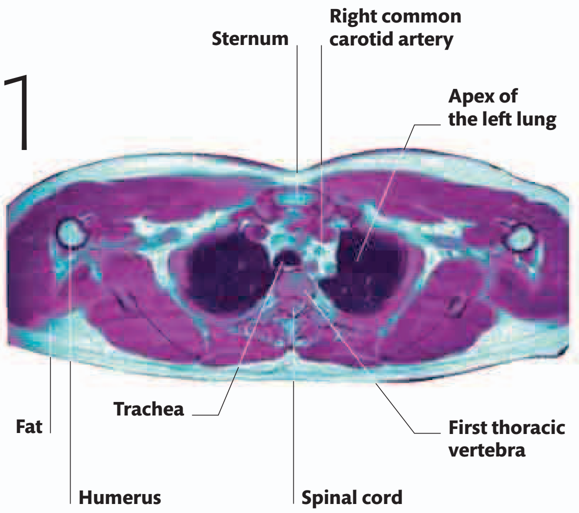
Sternum
Right common carotid artery
Apex of the left lung
Fat
Trachea
First thoracic vertebra
Humerus
Spinal cord

2

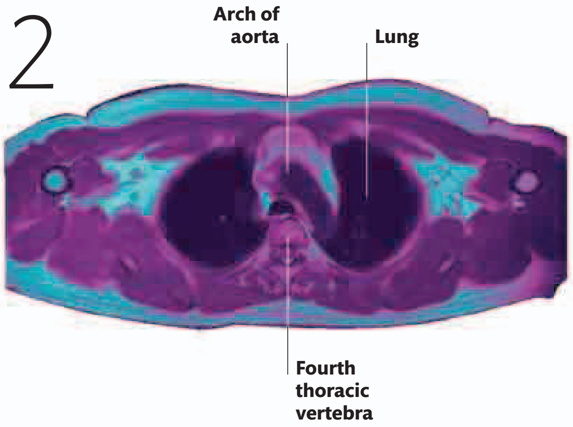
Arch of aorta
Lung
Fourth thoracic vertebra

3

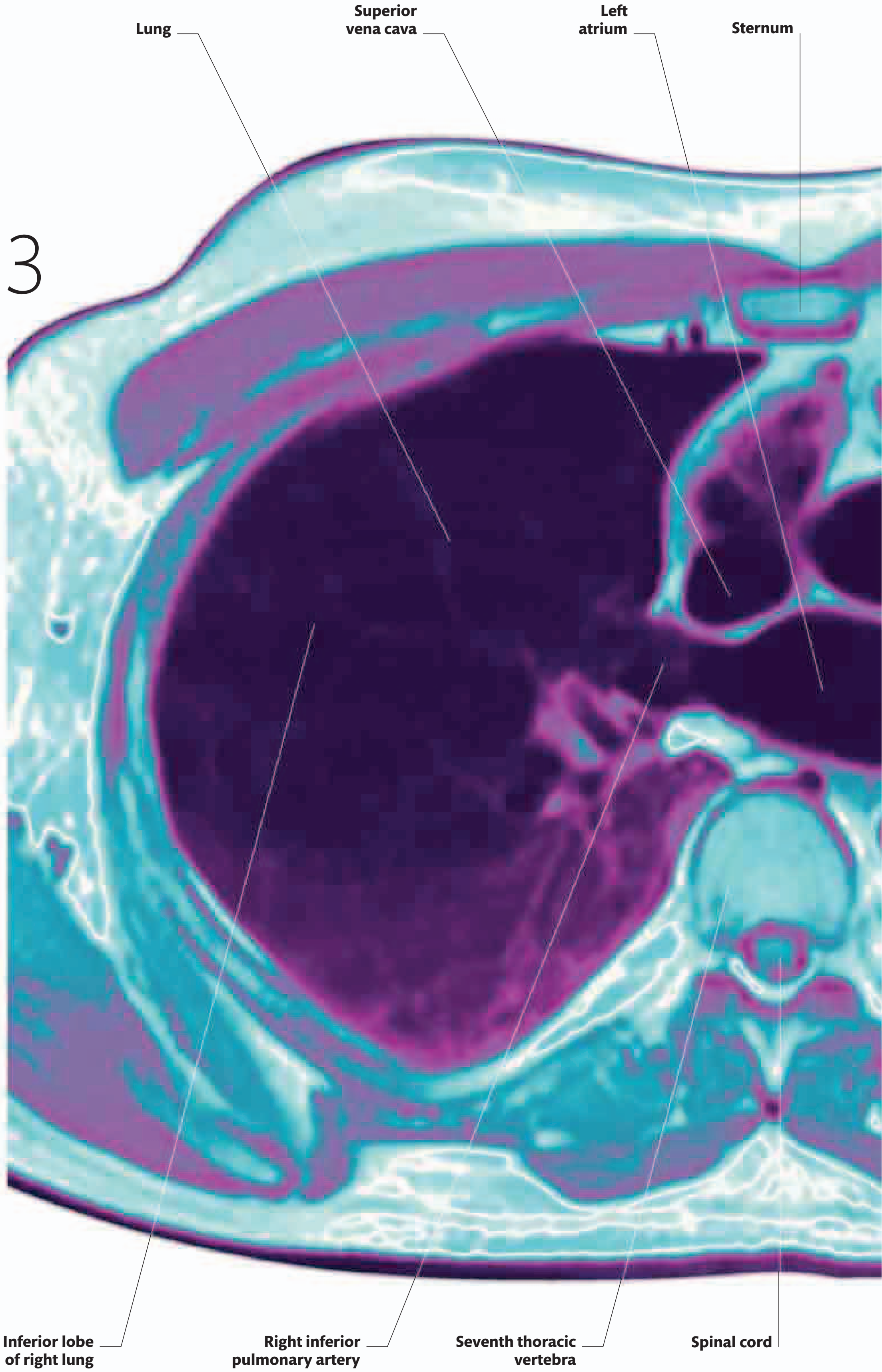
Lung
Superior vena cava
Left atrium
Sternum
Inferior lobe of right lung
Right inferior pulmonary artery
Seventh thoracic vertebra
Spinal cord

4

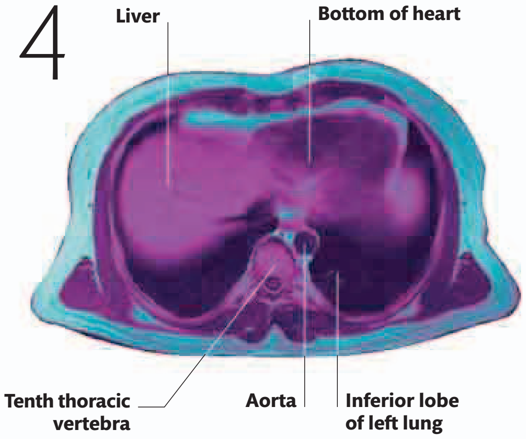
Liver
Bottom of heart
Tenth thoracic vertebra
Aorta
Inferior lobe of left lung

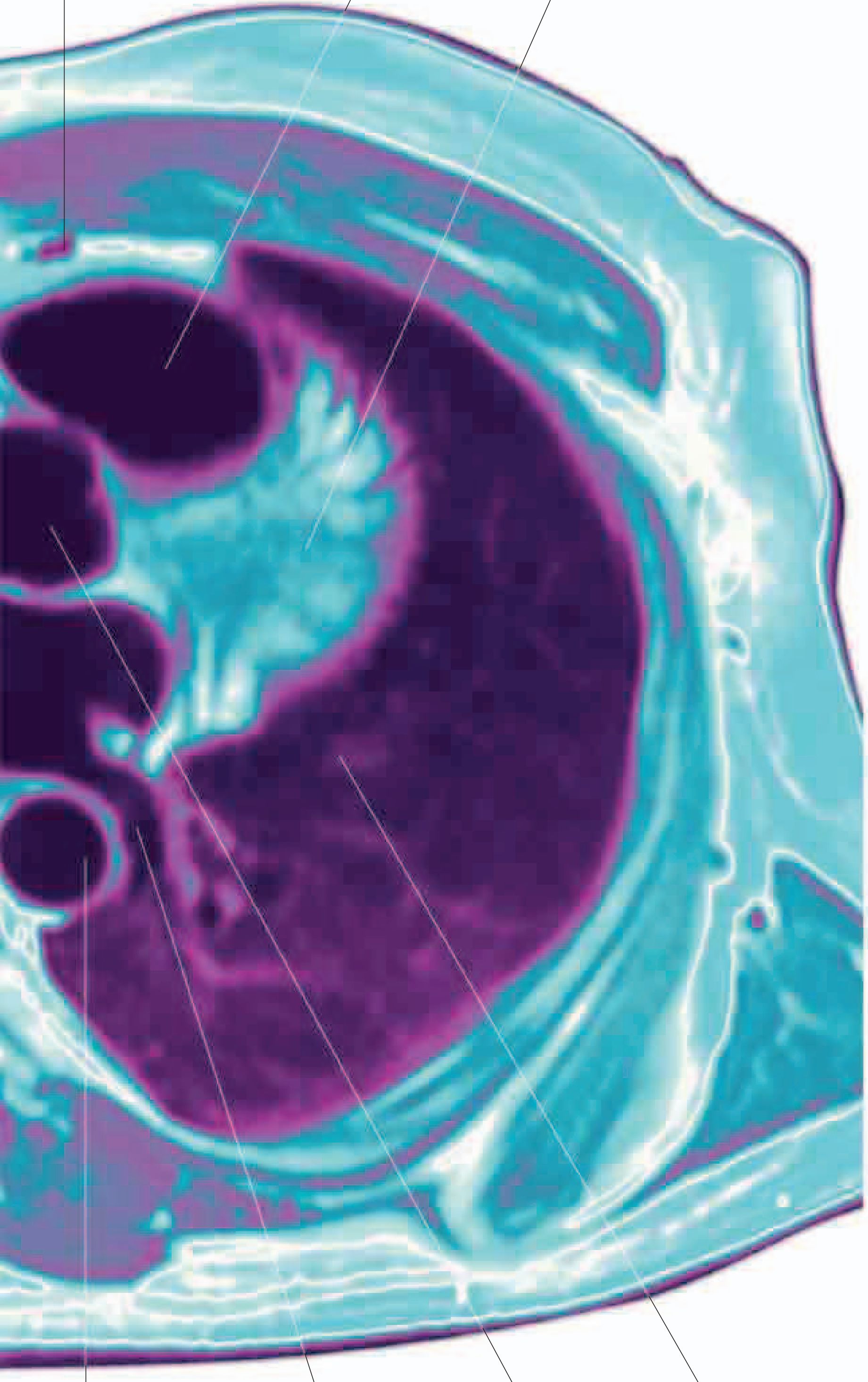
Internal thoracic vessels
Right ventricle
Muscle of left ventricle
Descending aorta
Left inferior pulmonary artery
Ascending aorta
Inferior lobe of left lung

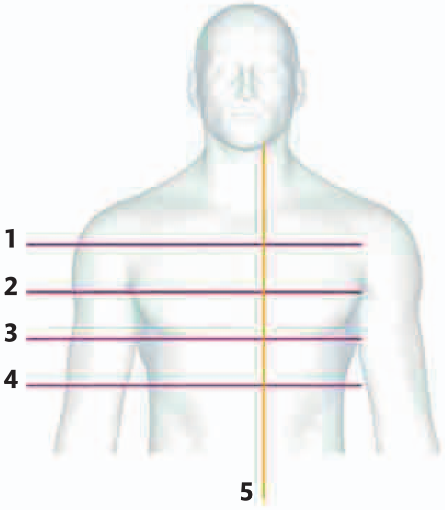

LEVELS OF SCANS

THORAX

The axial, or transverse, sections through the chest (sections 1–4) show the heart and large blood vessels lying centrally within the thorax, flanked by the lungs, and all set within the protective, bony casing of the ribcage. Section 1 shows the clavicles, or collarbones, joining the sternum at the front, the apex (top) of the lungs, and the great vessels passing between the neck and the thorax. Section 2 is lower down in the chest, just above the heart, while section 3 shows the heart with detail of its different chambers. The aorta appears to be to the right of the spine in this image, rather than to the left, but this is the usual way in which scans are viewed. You need to imagine yourself standing at the foot of the bed, looking down at the patient. This means that the left side of the body appears on the right side of the image as you view it. Section 4 shows the very bottom of the heart, and the inferior lobes of the lungs.

5

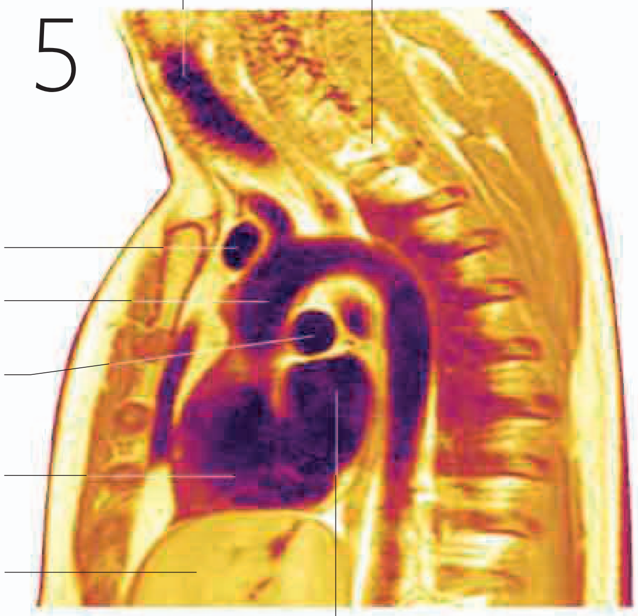
Infraglottic cavity
Spine
Left brachiocephalic vein
Ascending aorta
Left pulmonary artery
Right ventricle
Liver
Left atrium

1

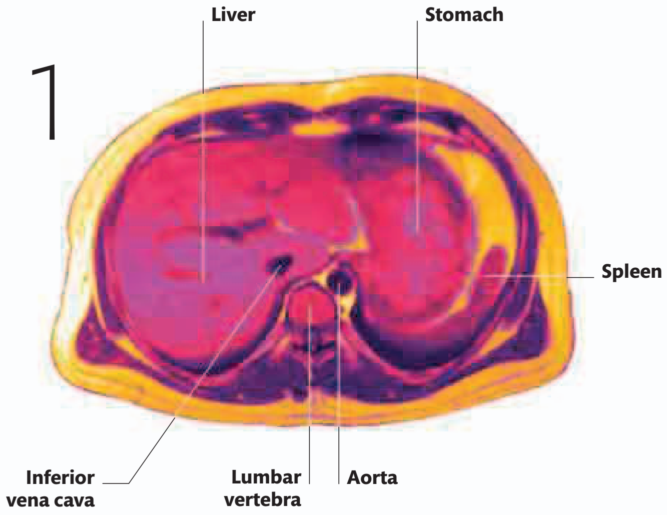
Liver
Stomach
Spleen
Inferior vena cava
Lumbar vertebra
Aorta

3

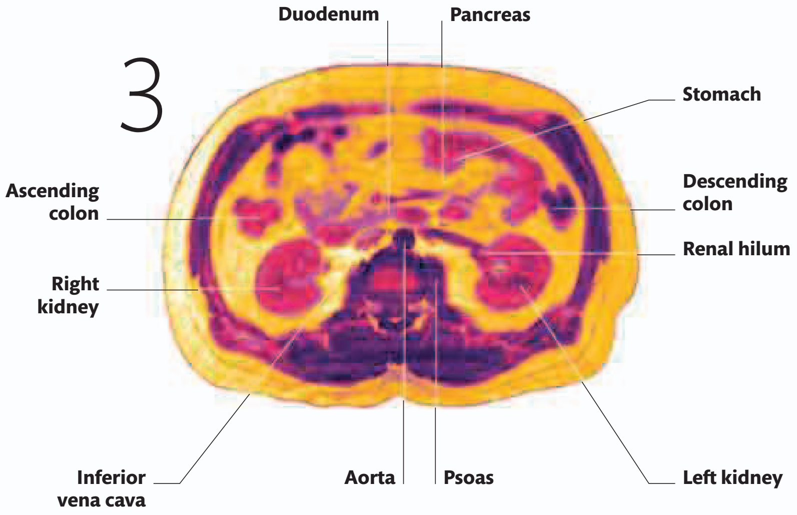
Duodenum
Pancreas
Stomach
Ascending colon
Descending colon
Renal hilum
Right kidney
Inferior vena cava
Aorta
Psoas
Left kidney

2

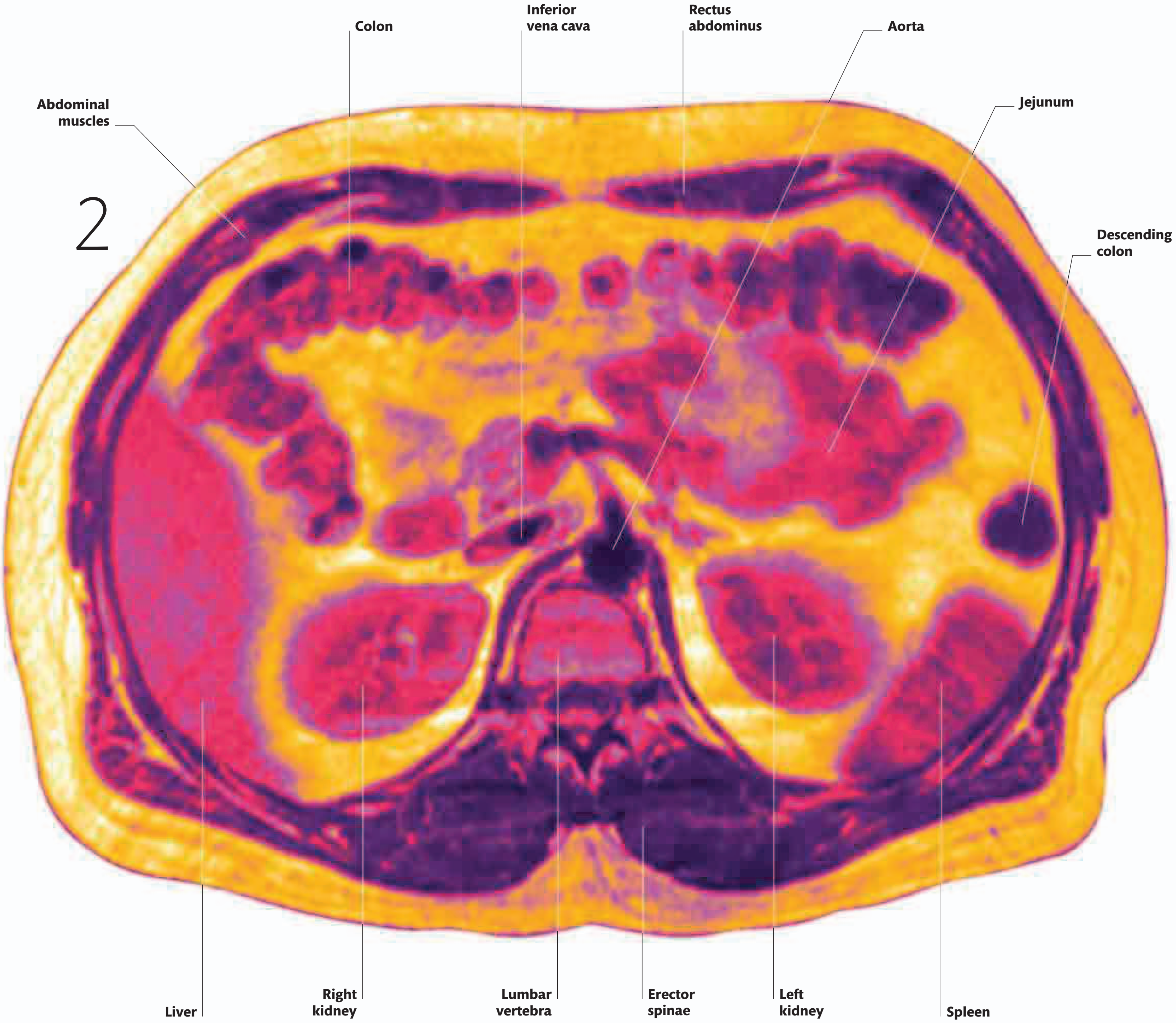
Colon
Inferior vena cava
Rectus abdominus
Aorta
Abdominal muscles
Jejunum
Descending colon
Liver
Right kidney
Lumbar vertebra
Erector spinae
Left kidney
Spleen

4

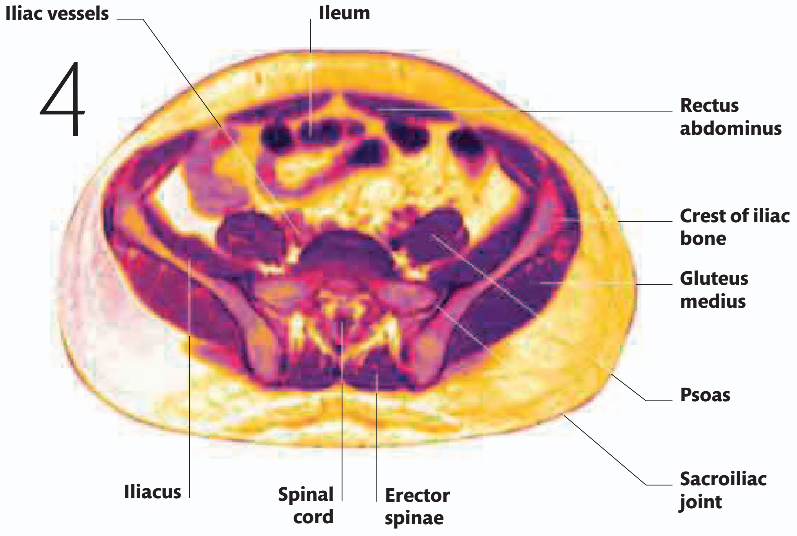
Iliac vessels
Ileum
Rectus abdominus
Crest of iliac bone
Gluteus medius
Psoas
Sacroiliac joint
Iliacus
Spinal cord
Erector spinae

5

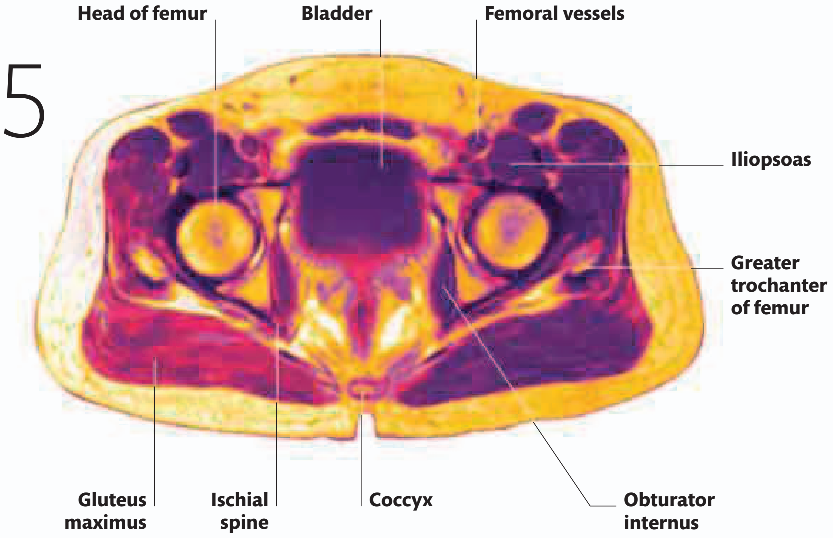
Head of femur
Bladder
Femoral vessels
Iliopsoas
Greater trochanter of femur
Gluteus maximus
Ischial spine
Coccyx
Obturator internus

ABDOMEN AND PELVIS

MRI is a useful way of looking at soft tissues—and for visualizing the organs of the abdomen and pelvis, which only appear as subtle shadows on a standard X-ray. In the series of axial or transverse sections through the abdomen and pelvis, we can clearly see the dense liver, and blood vessels branching within it (section 1); the right kidney lying close to the liver, and the left kidney close to the spleen (section 2); the kidneys at the level where the renal arteries enter them (section 3), with the stomach and pancreas lying in front; coils of small intestine, the ileum, resting in the lower part of the abdomen, cradled by the iliac bones (section 4); and the organs of the pelvis at the level of the hip joints (section 5). The sagittal view (section 6) shows how surprisingly shallow the abdominal cavity is, in front of the lumbar spine. In a slim person, it is possible to press down on the lower abdomen and feel the pulsations of the descending aorta—right at the back of the abdomen.

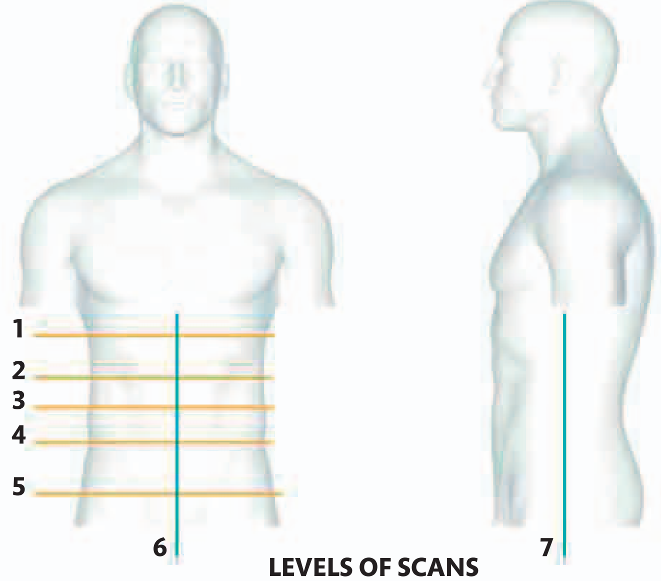

LEVELS OF SCANS

6

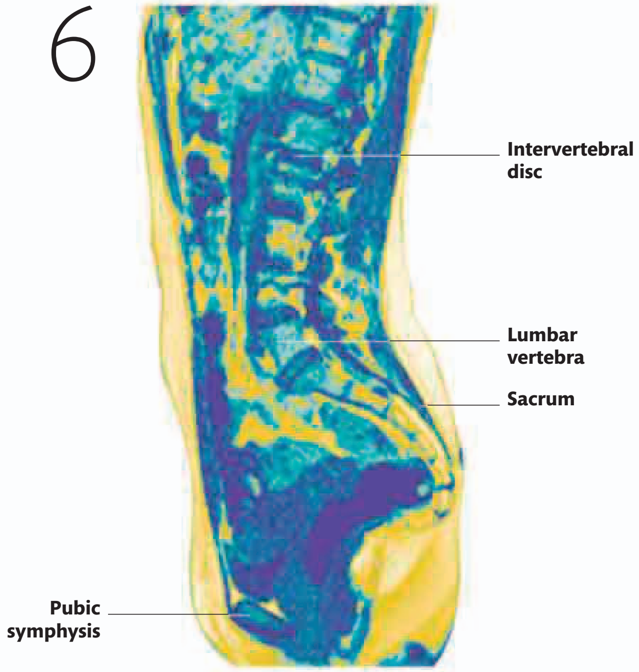
Intervertebral disc
Lumbar vertebra
Sacrum
Pubic symphysis

7

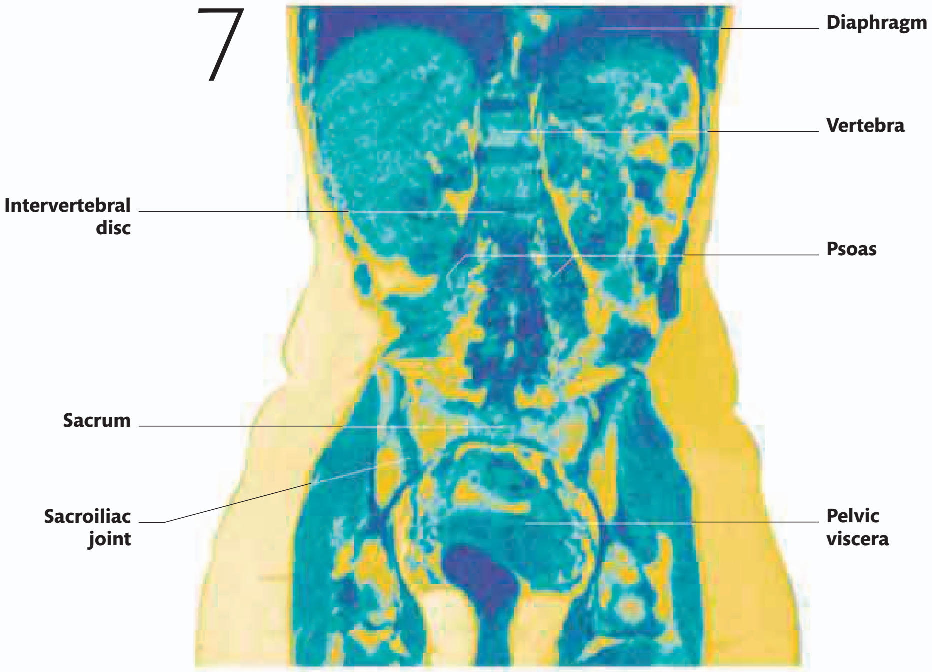
Diaphragm
Vertebra
Intervertebral disc
Psoas
Sacrum
Sacroiliac joint
Pelvic viscera

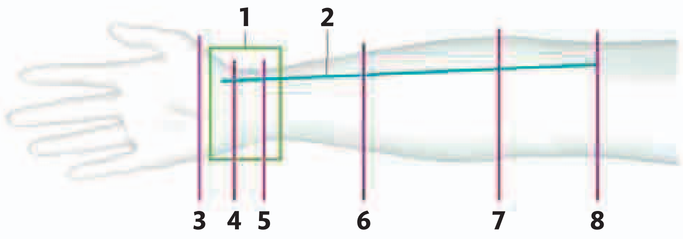

LEVELS OF SCANS

LOWER ARM AND HAND

These scans of the arm, forearm, and hand show how tightly packed the structures are. Section 1 reveals the bones of the wrist—the carpals—interlocking like a jigsaw. The wrist joint itself is the articulation between the radius and the scaphoid and lunate bones. In section 2, part of the elbow joint is visible, with the bowl-shaped head of the radius cupping the rounded end of the humerus. Muscles in the forearm are grouped into two sets, flexors on the front and extensors behind the forearm bones and interosseous membrane. Compare sections 3–8 with sections through the leg (see pp.234–35)—both limbs have a single bone (humerus or femur) in the upper part, two bones in the lower part (radius and ulna in the forearm; tibia and fibula in the lower leg), a set of bones in the wrist and ankle (carpals and tarsals), fanning out to five digits at the end of the limb. Evolutionarily, these elements developed from the rays of a fish fin.

1

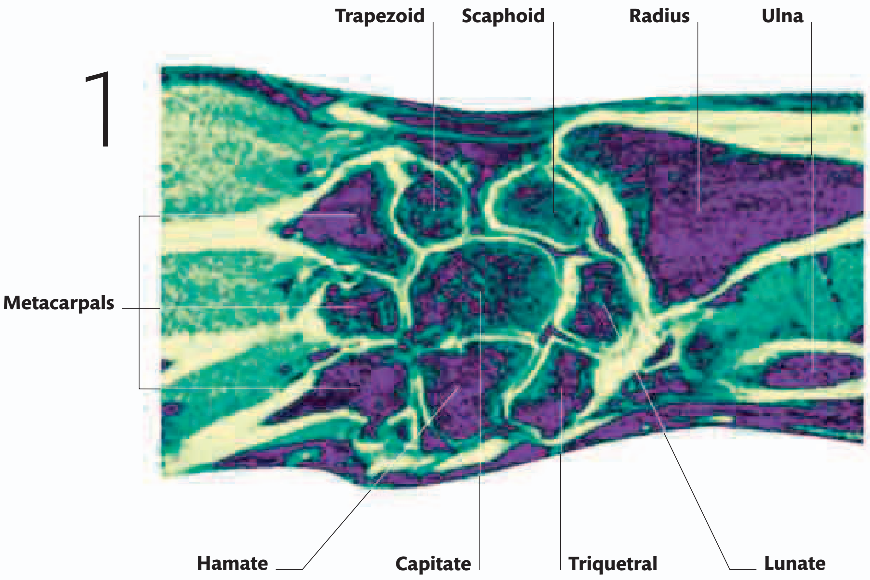
Trapezoid
Scaphoid
Radius
Ulna
Metacarpals
Hamate
Capitate
Triquetral
Lunate

2

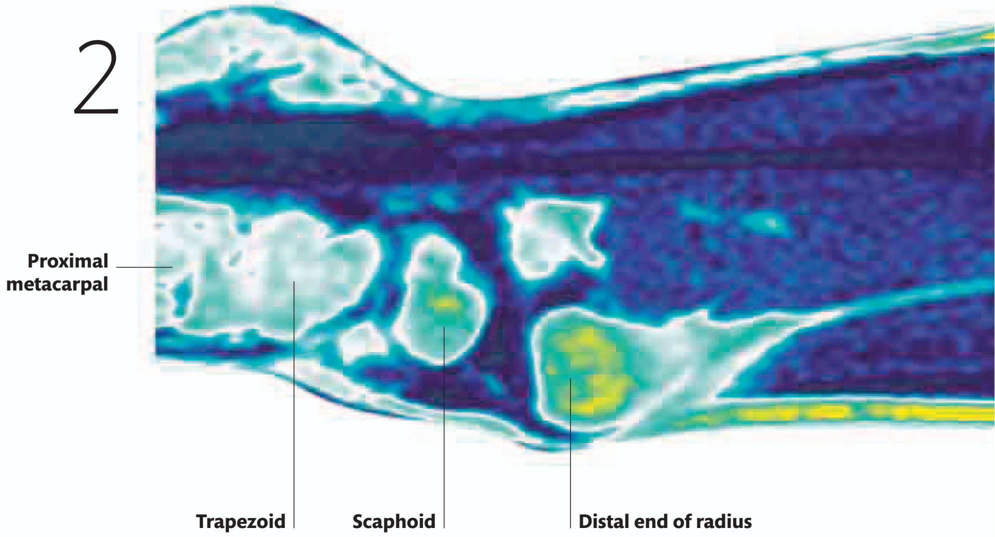
Proximal metacarpal
Trapezoid
Scaphoid
Distal end of radius

3

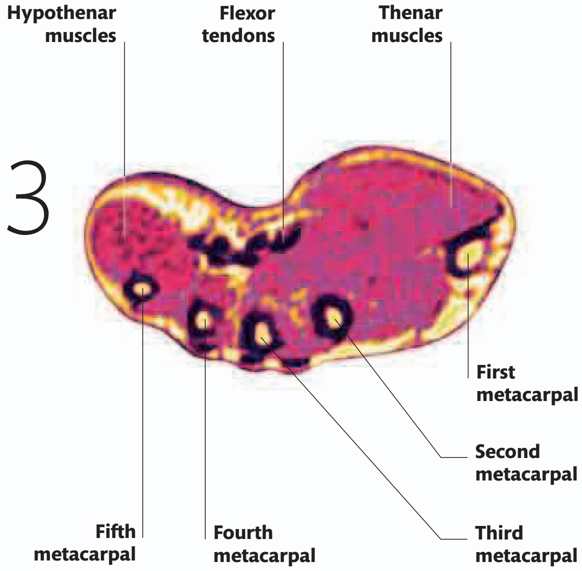
Hypothenar muscles
Flexor tendons
Thenar muscles
First metacarpal
Second metacarpal
Fifth metacarpal
Fourth metacarpal
Third metacarpal

4

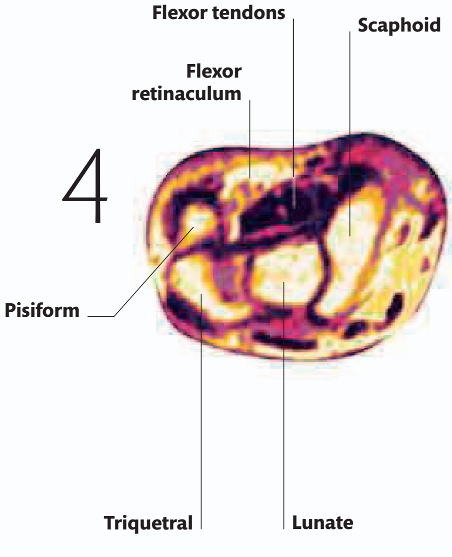
Flexor tendons
Scaphoid
Flexor retinaculum
Pisiform
Triquetral
Lunate

5

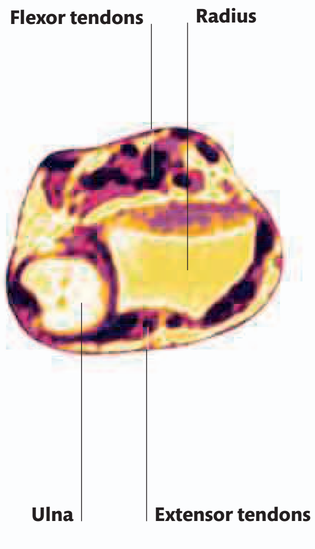
Flexor tendons
Radius
Ulna
Extensor tendons

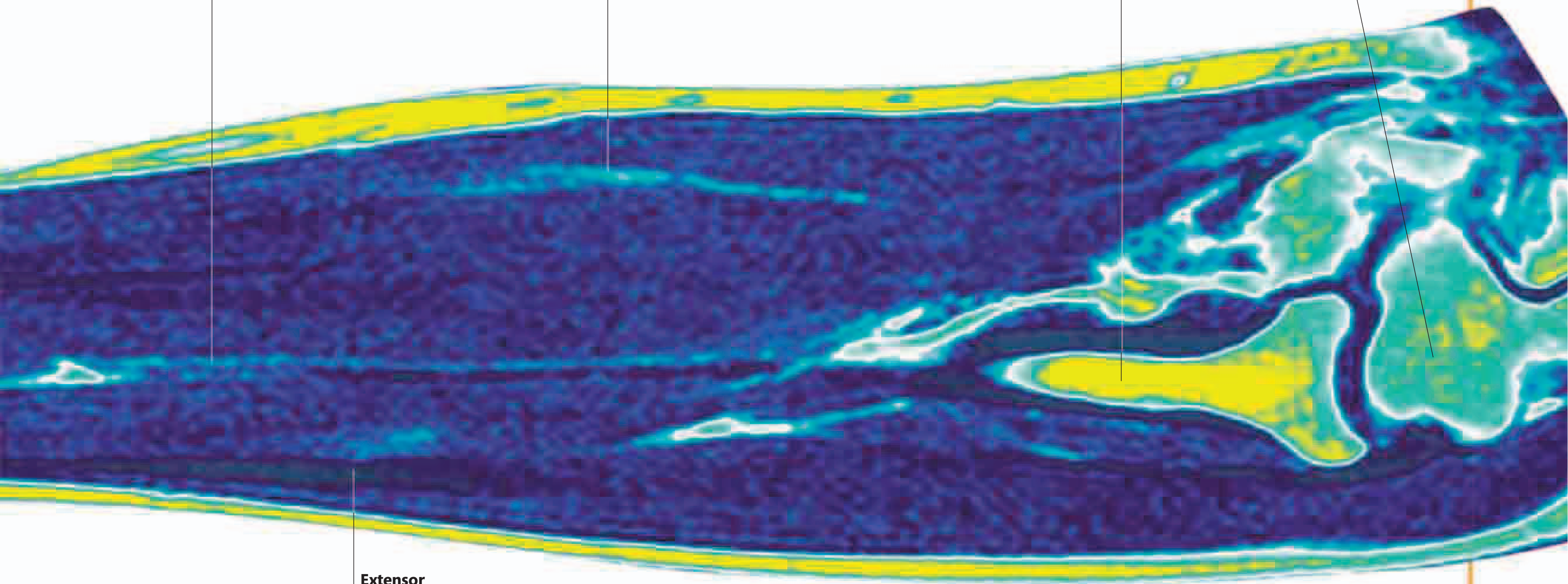
Interosseous membrane
Flexor compartment of forearm
Radius
Humerus
Extensor compartment of forearm

6

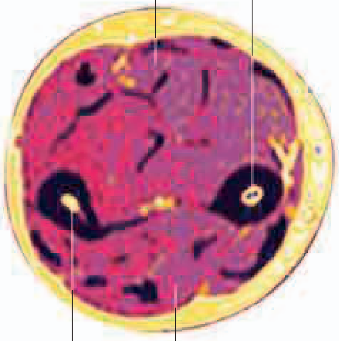
Flexor compartment of forearm
Radius
Ulna
Extensor compartment of forearm

7

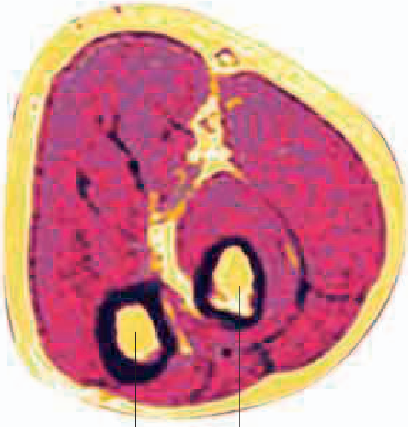
Ulna
Radius

8

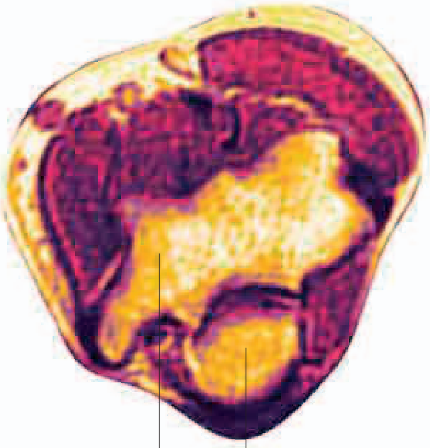
Ulna
Humerus

1

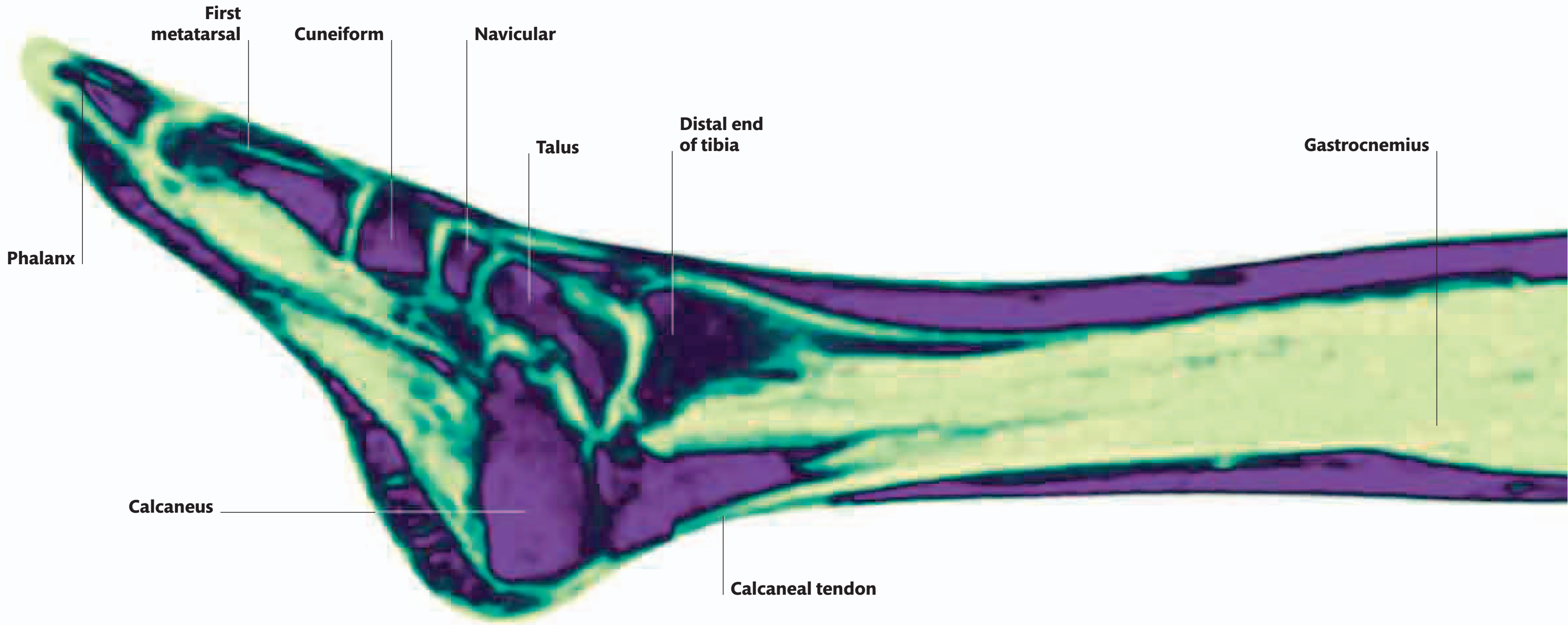
First metatarsal
Cuneiform
Navicular
Talus
Distal end of tibia
Gastrocnemius
Phalanx
Calcaneus
Calcaneal tendon

2

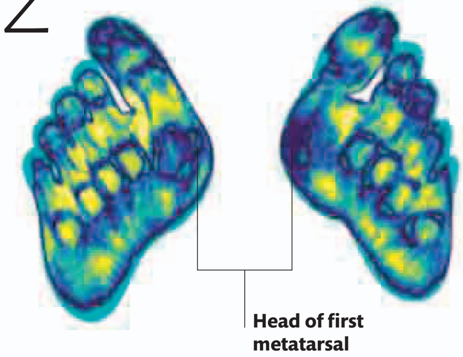
Head of first metatarsal

3

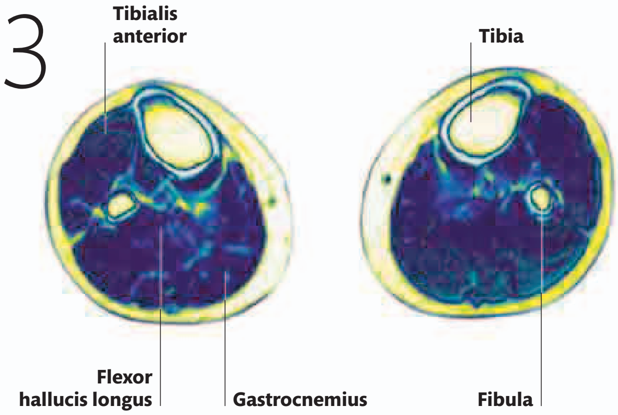
Tibialis anterior
Tibia
Flexor hallucis longus
Gastrocnemius
Fibula

4

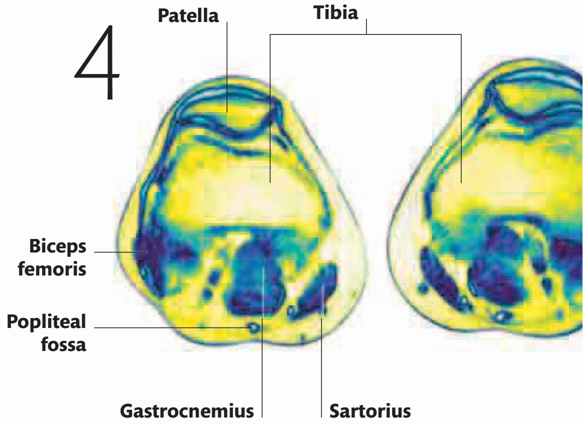
Patella
Tibia
Biceps femoris
Popliteal fossa
Gastrocnemius
Sartorius

LOWER LIMB AND FOOT

The sequence of axial and transverse sections through the thigh and lower leg show how the muscles are arranged around the bones. Groups of muscles are bound together with fascia—fibrous packing tissue—forming three compartments in the thigh (the flexor, extensor, and adductor muscles), and three in the lower leg (flexor, extensor, and peroneal or fibular muscles). Nerves and deep blood vessels are also packaged together in sheaths of fascia, forming “neurovascular bundles”. Section 2 shows the bones of the forefoot, while the tightly packed muscles surrounding the tibia and fibula in the lower leg are visible in section 3. At the knee joint, shown in section 4, the patella can be seen to fit neatly against the reciprocal shape of the femoral condyles. The neurovascular bundle is clearly visible here, at the back of the knee, in a space known as the popliteal fossa—with the hamstring muscles on either side. Sections 5 and 6, through the middle and upper thigh, show the powerful quadriceps and hamstring muscles surrounding the thigh bone, or femur.

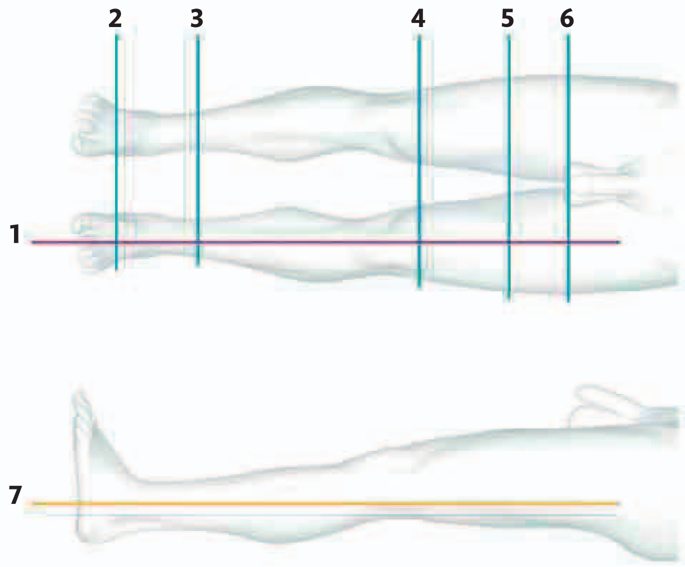

LEVELS OF SCANS

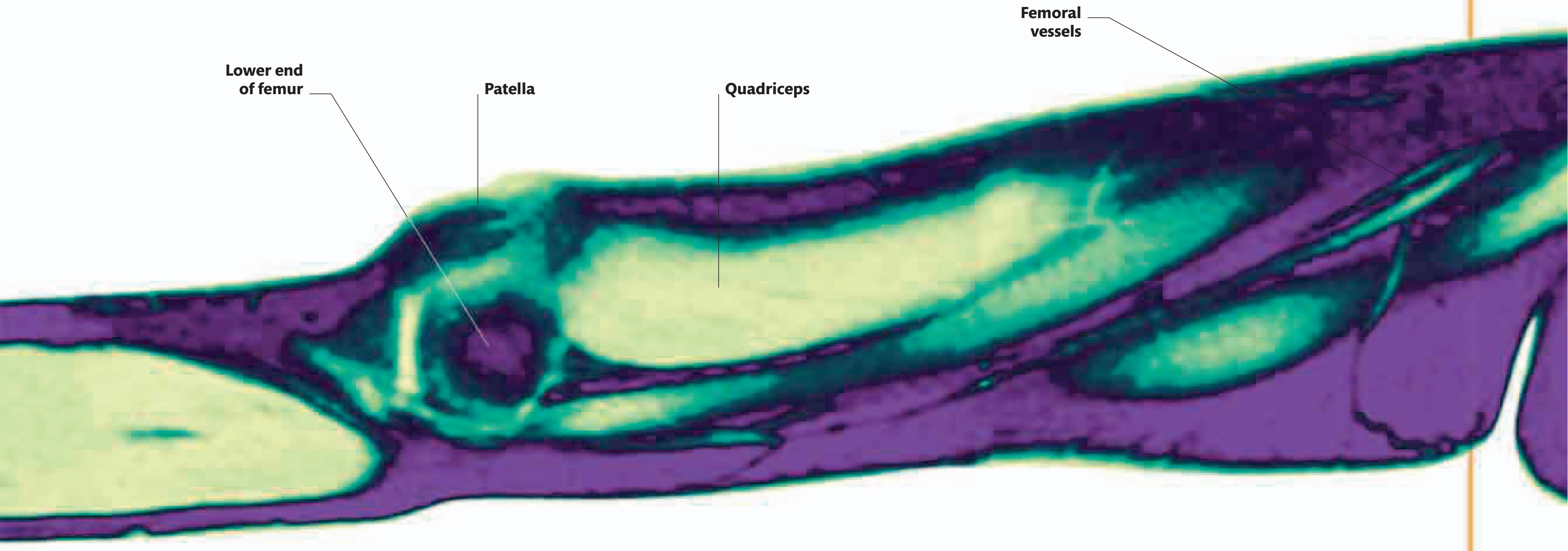
Lower end of femur
Patella
Quadriceps
Femoral vessels

5

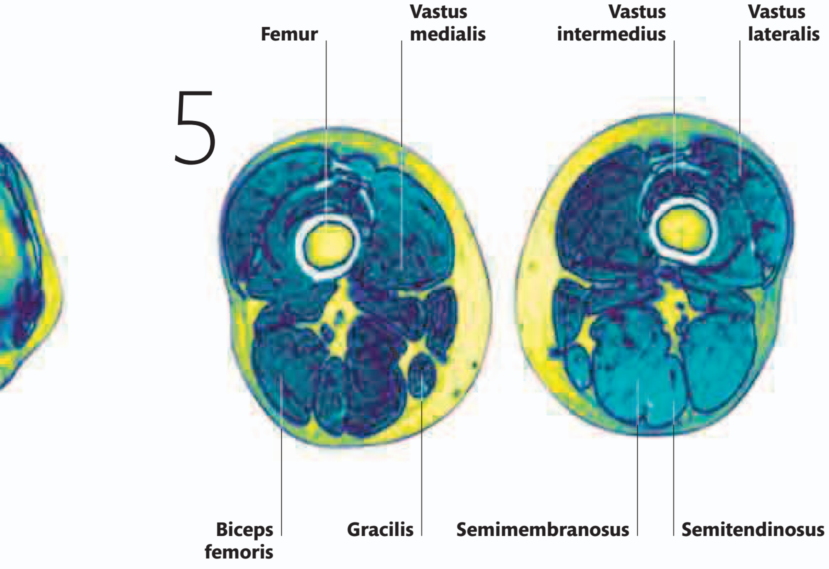
Femur
Vastus medialis
Vastus intermedius
Vastus lateralis
Biceps femoris
Gracilis
Semimembranosus
Semitendinosus

6

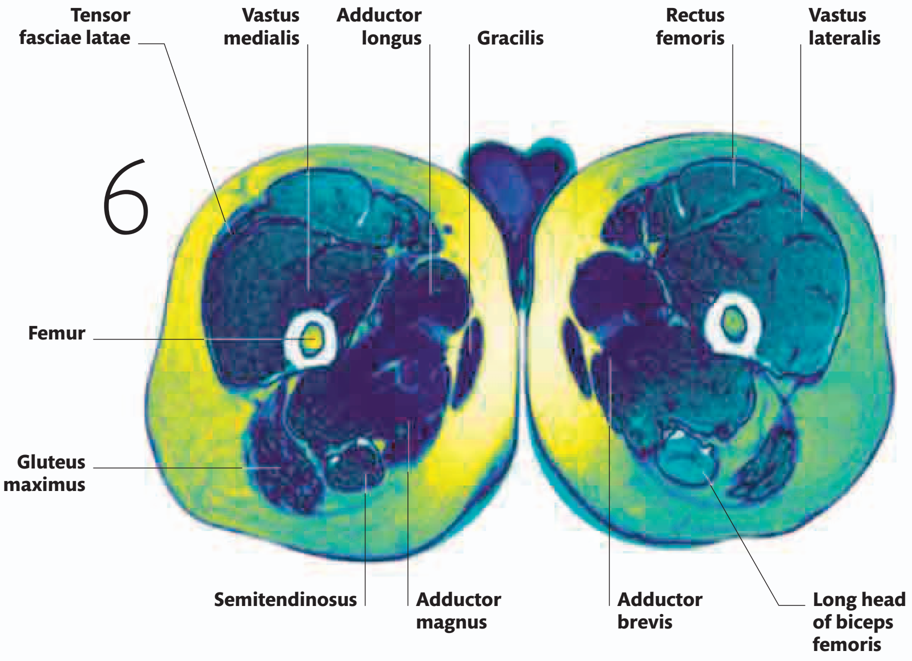
Tensor fasciae latae
Vastus medialis
Adductor longus
Gracilis
Rectus femoris
Vastus lateralis
Femur
Gluteus maximus
Semitendinosus
Adductor magnus
Adductor brevis
Long head of biceps femoris

7

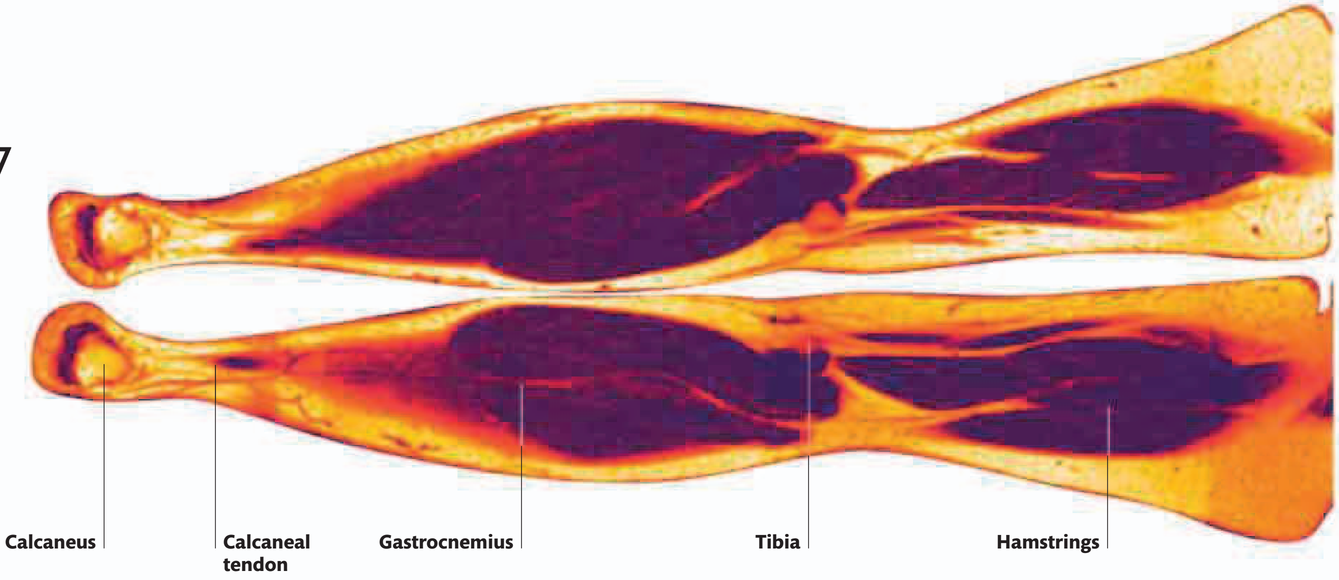
Calcaneus
Calcaneal tendon
Gastrocnemius
Tibia
Hamstrings

Get Full Access and More at **ExpertConsult.com**

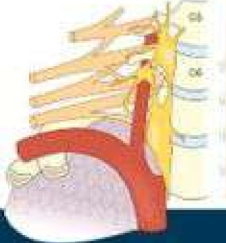

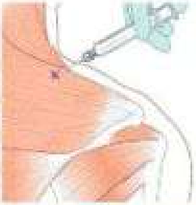

ATLAS OF PAIN INJECTION TECHNIQUES

SECOND EDITION

Therese O'Connor & Stephen Abram

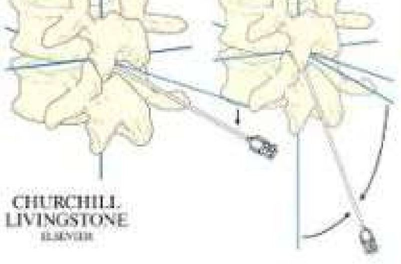

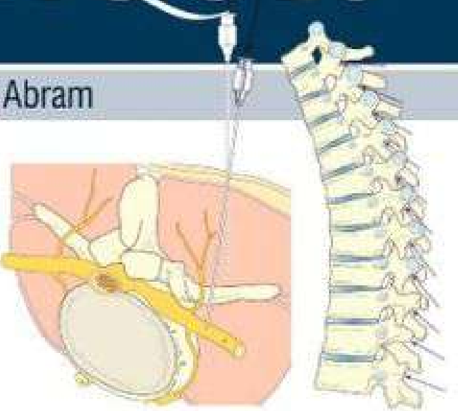

CHURCHILL LIVINGSTONE
ELSEVIER

JOINT INJECTIONS

2

2.1 LUMBAR FACET JOINT INJECTION

Anatomy

The zygopophyseal or facet joints (**Fig. 2.1.1**) are paired articular surfaces between the posterior aspects of adjacent vertebrae. In the cervical region, rotation and flexion are possible as the joint surfaces lie midway between the coronal and the axial planes. Rotation is prevented in the lumbar region but flexion is possible as the anterior portion of the lumbar facet joints lie in the coronal plane and the posterior portions in the sagittal plane. In the thoracic region, the joints' inferior and superior articular surfaces overlap each other in an almost vertical incline.

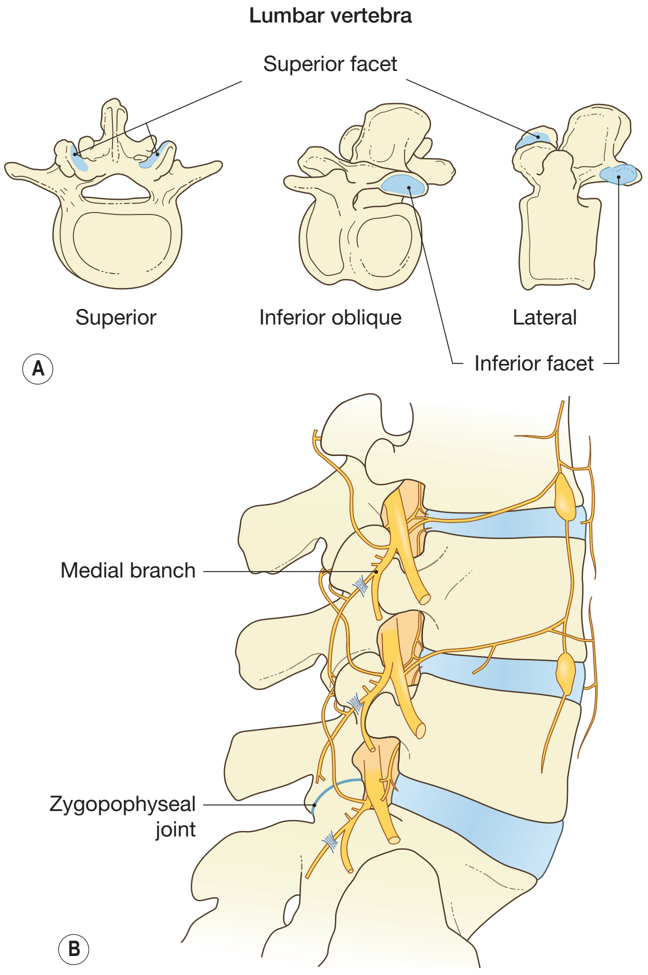

Fig. 2.1.1

The facet joints bear most of the shear forces when the spine is flexed. In addition, when the intervertebral discs are degenerated, the facet joints carry increased load and weight, especially when the spine is extended. Innervation of the facet joints is via the medial branches of the dorsal rami of the spinal nerves. These nerves also innervate the muscles and ligaments surrounding the joints. Each medial branch divides into proximal and distal branches (**Fig. 2.1.2**). The proximal branch innervates the

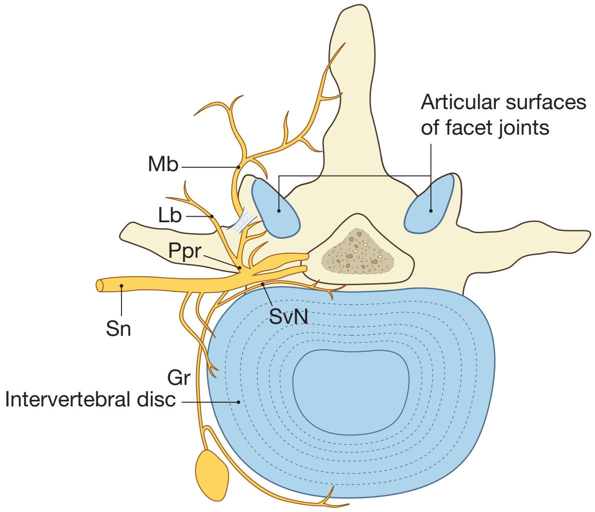

Fig. 2.1.2 Lumbar spine innervation. Innervation of the lumbar spinal structures in the transverse view. Note the posterior primary ramus (Ppr) leaving the spinal nerve (Sn) and splitting into a lateral branch (Lb) and a medial branch (Mb). The medial branch passes under the mamillo-accessory ligament to innervate the facet joint and capsule, the spinous process and the multifidus muscles. Sensory fibers traveling with the gray rami (Gr) form the sinu-vertebral nerve (SvN) and provide sensory function to the disc annulus. (Reproduced with permission from *Cousins and Bridenbaugh's Neural Blockade in Clinical Anesthesia and Management of Pain*, 4th edition, Wolters Kluwer/Lippincott Williams & Wilkins, 2009.)

adjacent facet joint, and the distal branch innervates the next facet joint below. The medial branch also innervates the interspinous ligaments and the multifidus muscles and the lateral branch innervates other adjacent muscles. Thus, pain from irritation of a joint may cause generalized sensitization of the dorsal rami with secondary hyperactivity and spasm of the innervated muscles and may be difficult to localize.

The facet joints contain vascular, highly innervated intra-articular synovial inclusions, which may become trapped and inflamed when the joint is injured, causing pain.

Equipment

- 2 ml and 10 ml syringes
- 25 G needle
- 22 G spinal needle, end-opening
- Non-ionic radio-opaque contrast medium
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- C-arm fluoroscopy or ultrasound

Drugs

- Lidocaine (lignocaine) 1% 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 25 mg (or its equivalent)
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Pillow under anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 2.1.3*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The lumbar midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution.
- The spinous processes of the vertebrae are marked.
- **The insertion point of the needle lies 2–3 cm lateral to the cephalic end of the spinous process of the vertebra** (*Fig. 2.1.4 a,b*).
- C-arm fluoroscopy is positioned at an angle of about 30°, tilted towards the side of the joint to be injected. The angle is adjusted until the joint is well visualized. A radio-opaque object, e.g. the tip of a hemostat, is positioned over the joint and the skin is marked.
- Thereby, with the aid of fluoroscopy, the insertion point is identified.
- A skin wheal is raised and the area is infiltrated with lidocaine (lignocaine) 1%.
- A spinal needle is introduced in a **vertical direction to the skin,** until the needle is observed to enter the joint space, preferably near the lowest aspect of the joint (inferior recess). Confirmation of intra-articular placement is made by **observation of the needle tip remaining on the joint line as the fluoroscope is rotated laterally** (*Fig. 2.1.5*).
- After negative aspiration, 0.5 ml of non-ionic radio-opaque contrast medium (that is compatible with nerve tissue) is injected.
- The correct placement is indicated by outlining the joint with non-ionic radio-opaque contrast medium, visible on anteroposterior and oblique views (*Fig. 2.1.6 a,b*).

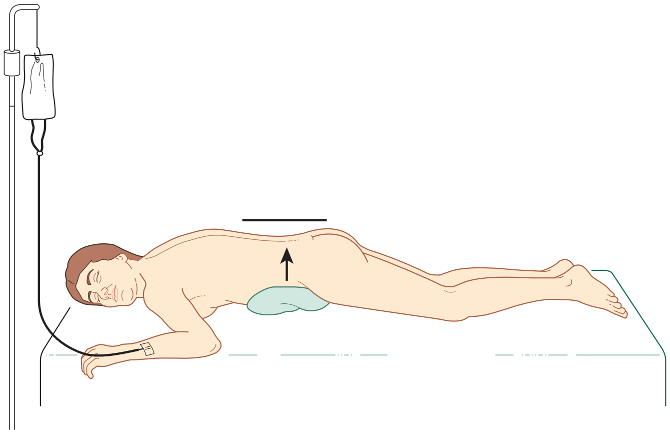

Fig. 2.1.3

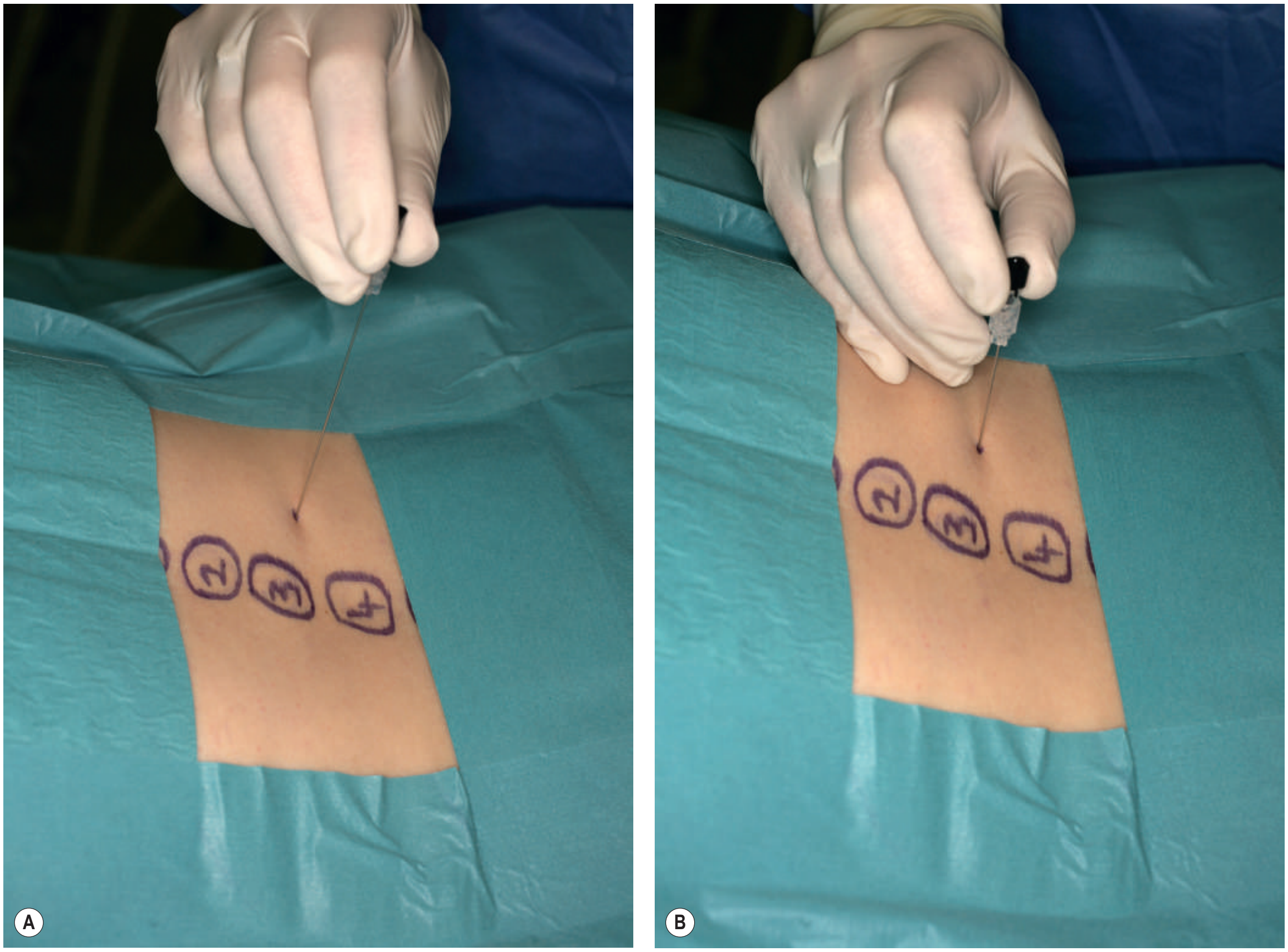

Fig. 2.1.4

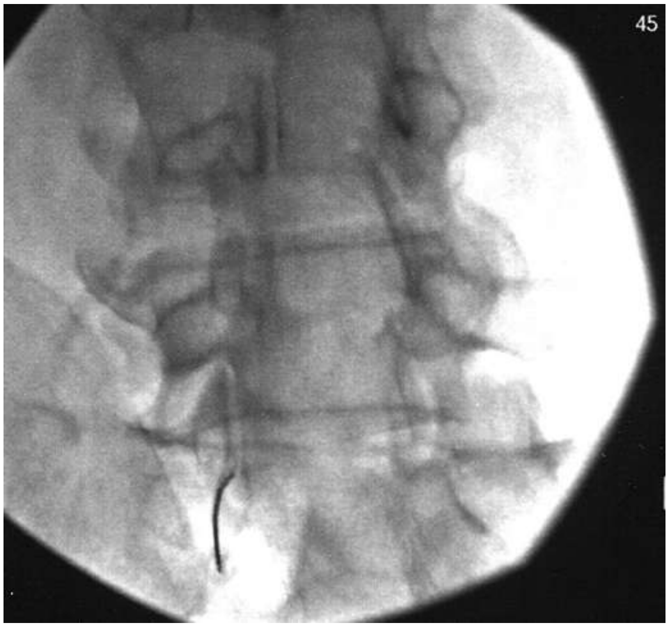

Fig. 2.1.5

Ultrasound may also be helpful in identifying the facet joint (*Fig. 2.1.7*).

- When the correct placement of the needle is confirmed, lidocaine (lignocaine) 1% 0.5 ml plus corticosteroid, e.g. triamcinolone diacetate 25 mg, may be injected and the needle removed while clearing with lidocaine (lignocaine) 1% 1 ml.

Confirmation of a successful injection

- Relief of pain.

Tips

- Care must be taken to inject only a small amount of volume as described above. A total volume of more than 1 ml may damage the joint. If the joint is disrupted anteriorly, drug may spread to the epidural space.

Potential problems

- Solution may spread to the epidural space via the anteromedial capsule.
- Nerve root injection.

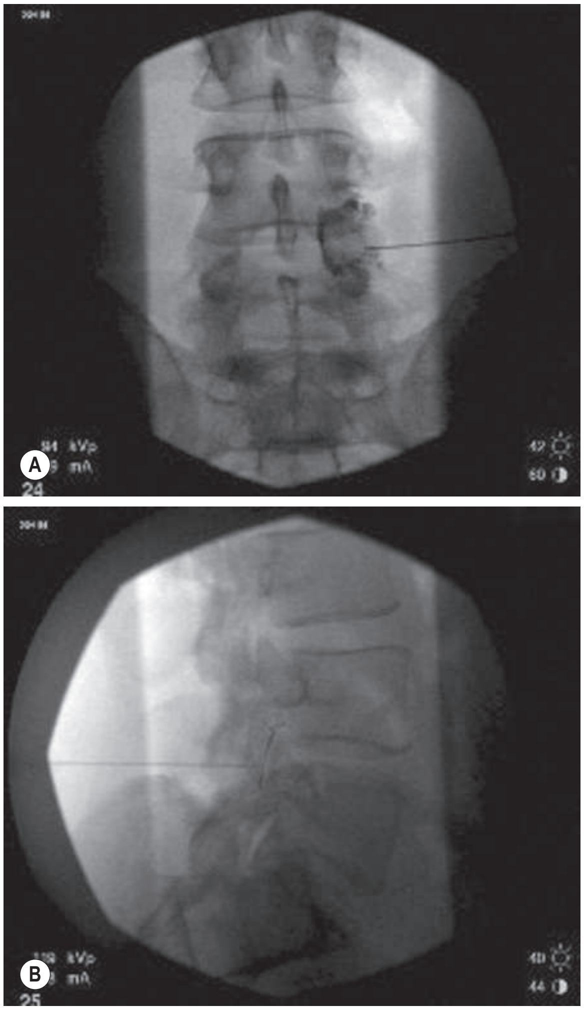

Fig. 2.1.6

Lumbar Facet Nerve Injection

- Facet nerve injection may be carried out by placing a spinal needle at the point where the superior articular and transverse processes join as the median branch passes over the cephalad edge of the transverse process (*Fig. 2.1.8*).
- The direct posterior approach should be avoided as the needle placement may be obstructed by the superior portion of the facet joint.
- Approach to the target site from a **lateral oblique angle 30° to skin** is recommended.
- The needle is advanced towards the target site (the posterior-superior edge of the transverse process) until bone is encountered.
- It is recommended that the transverse process be approached first, to determine depth.
- The needle is then repositioned medially until the lateral edge of the facet joint is reached.
- The needle is then moved superiorly until it just "falls off" the superior edge of the transverse process (*Fig. 2.1.9*).
- The optimum position is obtained by repositioning the needle to the postero-superior edge of the transverse process.
- The patient may now report reproduction of back pain.
- Injection of lidocaine (lignocaine) 1% 0.5 ml plus triamcinolone diacetate 25 mg may be carried out for therapeutic effect. Diagnostic blockade may be unreliable as anesthesia of a facet joint means that both nerves supplying the joint should be blocked. However, this means that the joint above and the joint below will also be partially blocked and therefore diagnosis of pain in a particular joint using nerve block is not feasible.

Potential problems

- The same potential problems may occur as described for lumbar facet joint injection (see above).

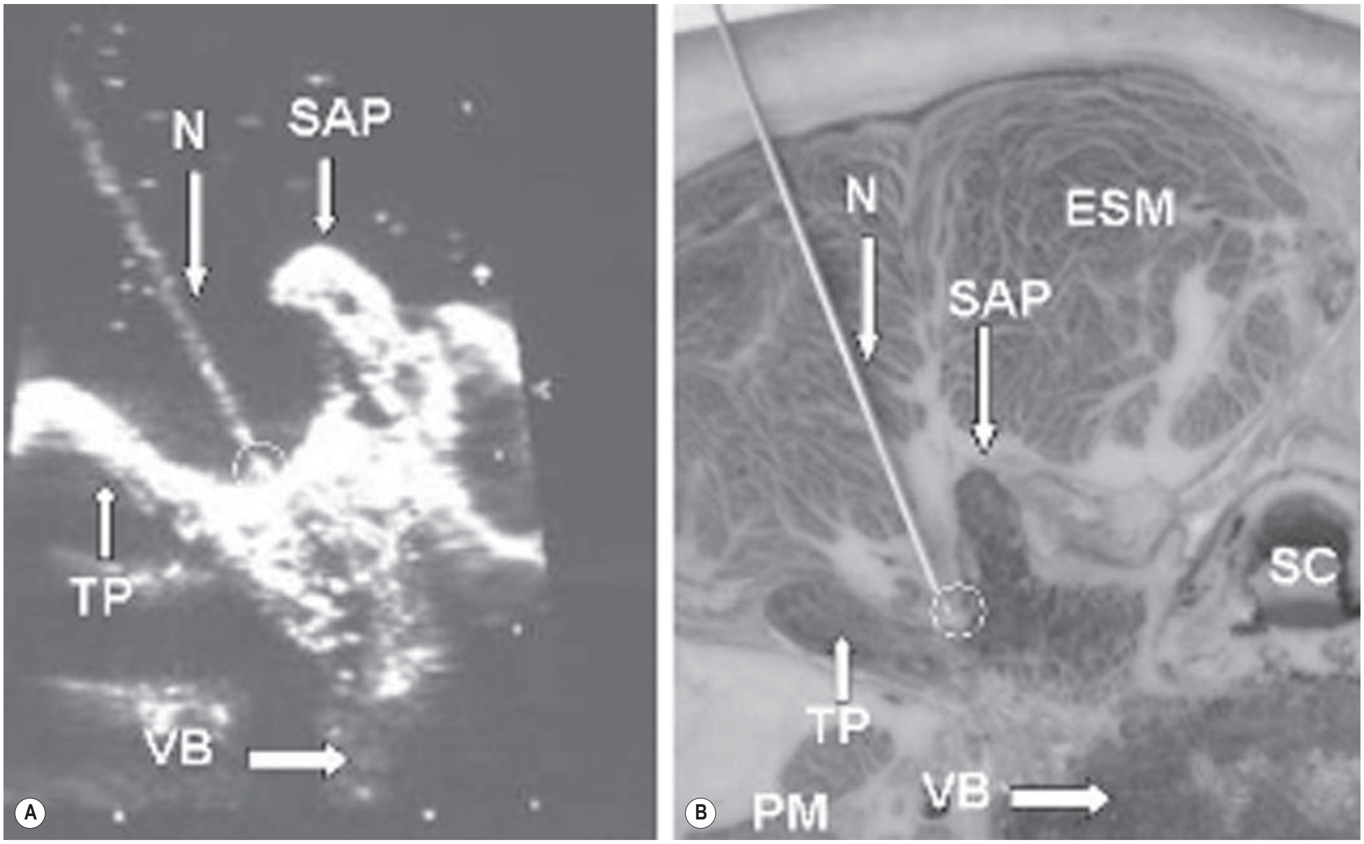

Fig. 2.1.7 **A** High-resolution sonogram (15-MHz linear transducer) of vertebral bone L3 immersed in water in the cross-axis view. **B** Corresponding anatomic cross-sectional cadaver preparation. Circles indicate targets. ESM erector spinae muscle; N needle; PM psoas muscle; SAP superior articular process; SC spinal channel; TP transverse process; VB vertebral body. (From Greher M, Scharbert G, Kamolz LP, et al, Ultrasound-guided lumbar facet nerve block: a sonoanatomic study of a new methodologic approach. Anesthesiology 2004; 100:1242–8 © 2004 American Society of Anesthesiologists, Inc. Lippincott Williams & Wilkins, Inc.)

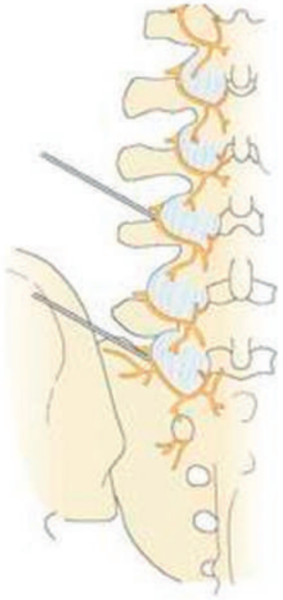

Fig. 2.1.8

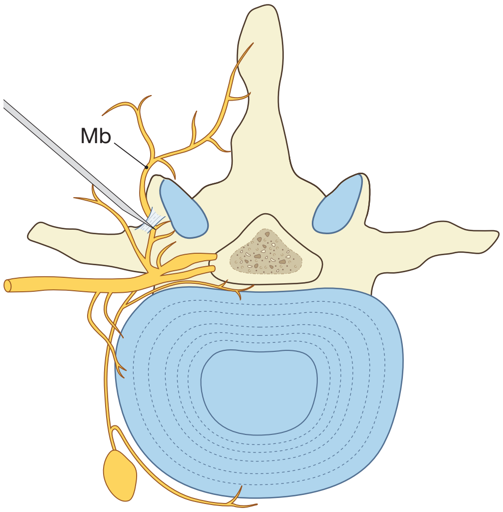

Fig. 2.1.9

2.2 CERVICAL FACET JOINT INJECTION

Anatomy

The anatomy relevant to injection of the cervical facet joints is similar to that relevant to the lumbar facet joints. The cervical facet joints below the C2–3 level are innervated by the medial branches of the cervical posterior primary rami. These divide into lateral and medial branches after leaving the posterior spinal canal and the splenius capitis muscles cover the medial branch posteriorly. The medial branches lie in close proximity to the vertebral artery and the epidural space is in close proximity to the anterior joint capsule (**Fig. 2.2.1**). The C2–3 facet joint is innervated by the medial branch of the third occipital nerve, which travels beneath the tendonous origin of the splenius capitis muscle where it may be accessed for local anesthetic blockade (**Fig. 2.2.2**).

Equipment

- 2 ml and 10 ml syringes
- 25 G needle
- 22 G spinal needle, end-opening
- Radio-opaque contrast medium
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- C-arm fluoroscopy or ultrasound

Drugs

- Lidocaine (lignocaine) 1%, 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 25 mg (or its equivalent)
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Neck slightly flexed (*Fig. 2.2.3*).

Needle puncture and technique

Caution: injection of 0.5–1 ml of lidocaine (lignocaine) 1% into the vertebral artery may result in immediate convulsion and/or loss of consciousness with possible cardiovascular system (CVS) collapse.

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The cervical midline and an area of 7 cm × 5 cm laterally is cleaned with antiseptic solution.
- The spinous processes are marked.

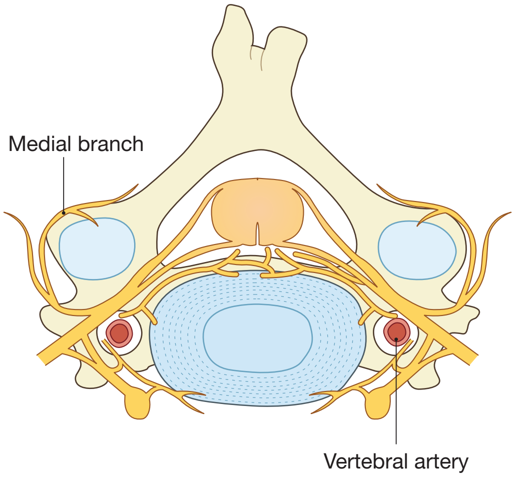

Fig. 2.2.1

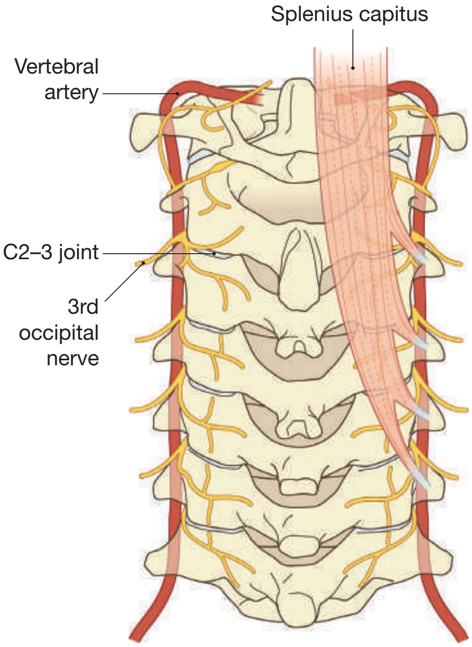

Fig. 2.2.2

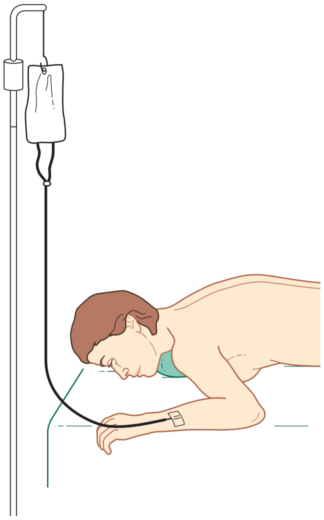

Fig. 2.2.3

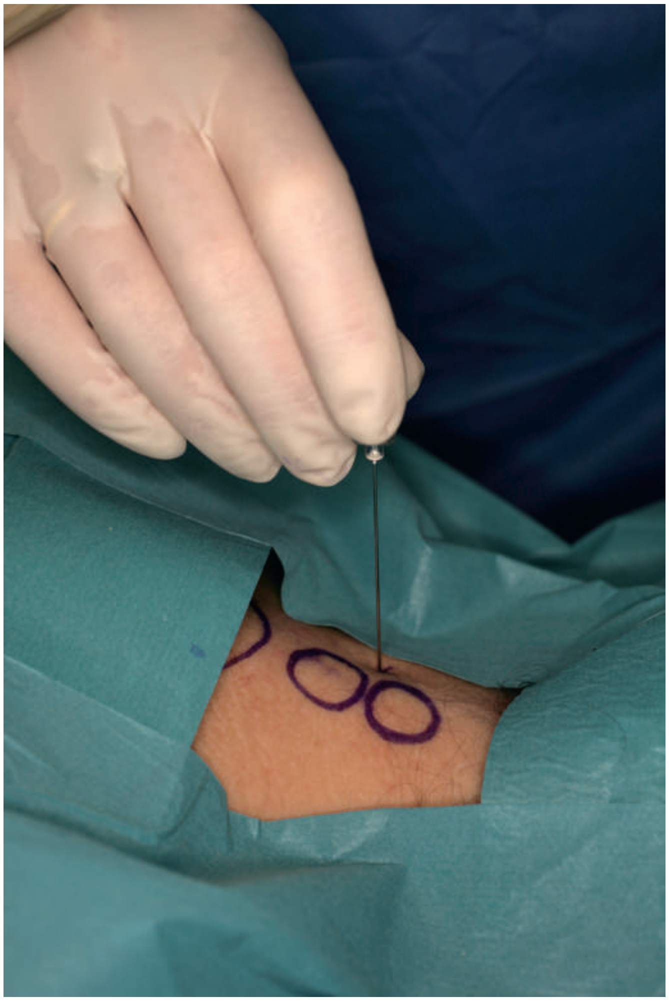

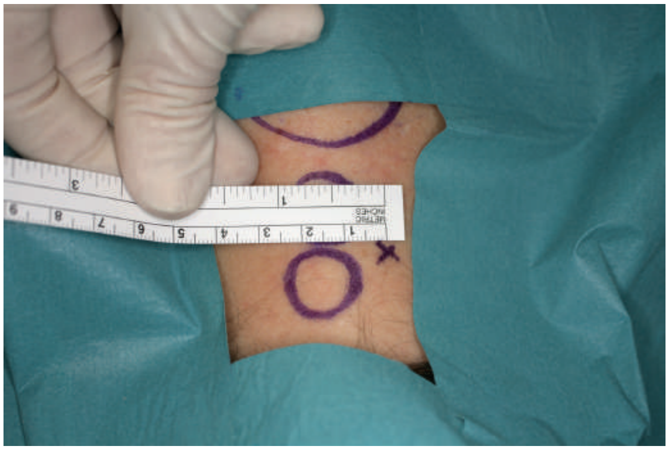

Fig. 2.2.4

Fig. 2.2.5

- **The insertion point of the needle lies 2–3 cm lateral to the cephalic end of the spinous process of the vertebra** (*Fig. 2.2.4*).
- C-arm fluoroscopy is positioned at an angle of about 30°, tilted towards the side of the joint to be injected. The angle is adjusted until the joint is well visualized. A radio-opaque object, e.g. the tip of a hemostat, is positioned over the joint and the skin is marked.
- Thereby, with the aid of fluoroscopy, the insertion point is identified.
- A skin wheal is raised and the area is infiltrated with lidocaine (lignocaine) 1%.
- A spinal needle is introduced in a vertical direction to the skin, until the needle is observed to enter the joint space (*Fig. 2.2.5*). Confirmation of intra-articular placement is made by observation of the needle tip remaining on the joint line as the fluoroscope is rotated (*Fig. 2.2.6*) or on ultrasound.
- After negative aspiration, 0.5 ml of non-ionic radio-opaque contrast medium (that is compatible with nerve tissue) is injected.
- The correct placement is indicated by outlining the joint with non-ionic radio-opaque contrast medium, visible on anteroposterior and oblique views.
- When the correct placement of the needle is confirmed, lidocaine (lignocaine) 1% 0.5 ml plus corticosteroid, e.g. triamcinolone diacetate 25 mg, may be injected and the needle removed while clearing with lidocaine (lignocaine) 1% 1 ml.

Confirmation of a successful injection

- Relief of pain.

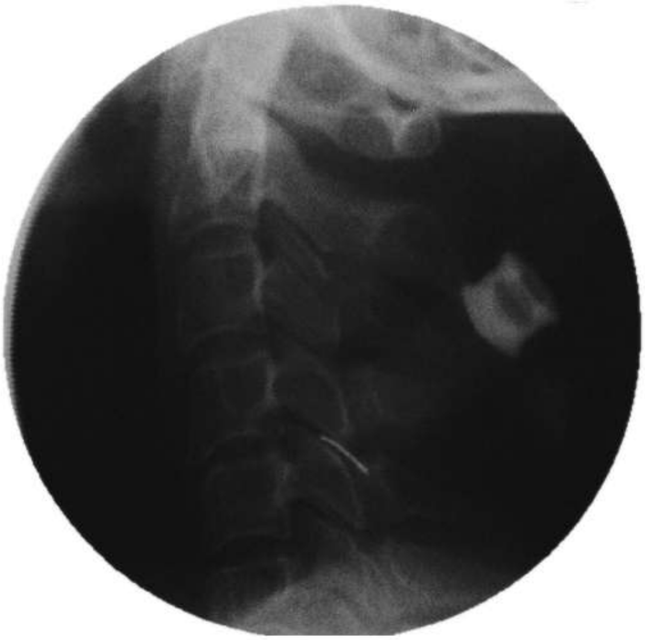

Fig. 2.2.6

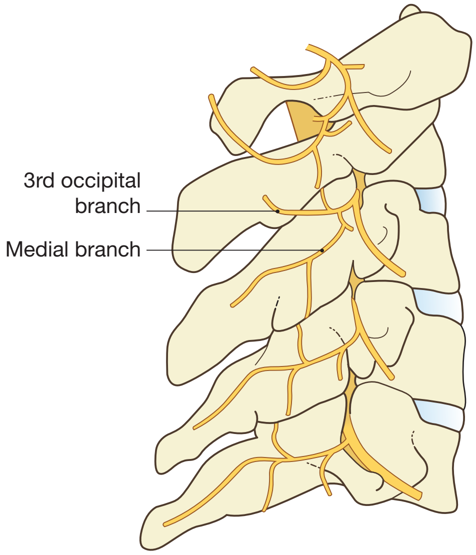

Fig. 2.2.7

Tips

- Care must be taken to inject only a small amount of volume, as described above. A total volume of more than 1 ml may damage the joint. If the joint is disrupted anteriorly, a drug may spread to the epidural space.

Potential problems

- Solution may spread to the epidural space via the anteromedial capsule.
- Nerve root injection.
- Intrathecal injection resulting in spinal anesthesia may occur if local anesthetic is inadvertently injected into the nerve root sleeve. Prompt recognition of this complication is vital during cervical procedures, because the patient's breathing may be arrested and there may be immediate convulsion and/or loss of consciousness with CVS collapse requiring immediate resuscitation. In addition, the patient's head should be immediately elevated after the injection to ensure that the lidocaine (lignocaine) flows inferiorly. Some practitioners elevate the head of the table during all cervical injections to help prevent this complication. Intravenous injection may be harmless, but it results in a suboptimal or false-negative result.
- Intra-arterial injection may result in immediate convulsion and/or loss of consciousness with possible CVS collapse. Intrarterial injection can be dangerous if the agent is injected into the vertebral artery or radicular branches that enter the neural foramina at various levels and, rarely, persisting paraplegia or paraparesis have been reported after cervical facet joint injection or nerve root block. These complications may be due to embolism from intra-arterial injection of particulate corticosteroid. However, even when contrast injection prior to steroid infiltration confirms extravascular needle placement, nerve damage may occur, which suggests an alternative cause for the complication, such as vasospasm or direct arterial injury from the needle-tip. Regardless of the cause, contrast injection is recommended to at least potentially reduce the risk of intravascular injection. Ultrasound will have limitations in this regard.
- Hematoma may occur (avoid performing block on patients who have coagulopathy).

CERVICAL FACET NERVE INJECTION

- Facet nerve injection may be carried out by placing a spinal needle at the point where the superior articular and transverse processes join as the median branch passes over the cephalad edge of the transverse process (*Fig. 2.2.7*).
- The direct posterior approach should be avoided as the needle placement may be obstructed by the superior portion of the facet joint.
- Approach to the target site from a **lateral oblique angle 30° to the skin** is recommended.
- The needle is advanced towards the target site (the posterior–superior edge of the transverse process) until bone is encountered.
- It is recommended that the transverse process be approached first, to determine depth.

- The needle is then repositioned medially until the lateral edge of the facet joint is reached.
- The needle is then moved superiorly until it just “falls off” the superior edge of the transverse process.
- The optimum position is obtained by repositioning the needle to the posterosuperior edge of the transverse process.
- The patient may now report reproduction of back pain.
- Injection of lidocaine (lignocaine) 1% 0.5 ml plus triamcinolone diacetate 25 mg may be carried out for therapeutic effect. Diagnostic blockade may be unreliable as anesthesia of a facet joint means that both nerves supplying the joint should be blocked. However, this means that the joint above and the joint below will also be partially blocked and therefore diagnosis of pain in a particular joint using nerve block is not feasible.

2.3 SACRO-ILIAC JOINT INJECTION

Anatomy

The surfaces of the sacrum and ilium form a synovial joint, the sacro-iliac joint. Ligaments and connective tissue surround the joint, conferring stability and preventing excessive movement of the joint (**Figs 2.3.1**, **2.3.2**). The joint is innervated by L4, L5, S1 (the superior gluteal nerve), S2, and L3. Localization of the pain is therefore difficult due to this wide nerve supply to the joint.

Equipment

- 2 ml and 10 ml syringes
- 25 G needle
- 22 G spinal needle, end-opening
- Radio-opaque contrast medium
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- C-arm fluoroscopy or ultrasound

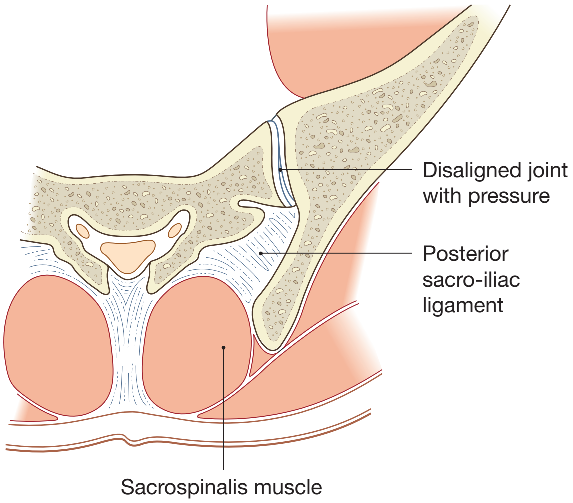
Disaligned joint with pressure
Posterior sacro-iliac ligament
Sacrospinalis muscle

Fig. 2.3.1

Drugs

- Lidocaine (lignocaine) 1%, 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 25 mg (or its equivalent)
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Pillow under anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 2.3.3*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The sacral area is prepared antiseptically.
- An AP image is obtained, centered over the joint to be injected.
- Two joint lines are observed. The posterior joint line is located more medial in a direct AP view (*Fig. 2.3.4a*). The image intensifier (positioned above the patient) is rotated toward the opposite side until the two joint lines

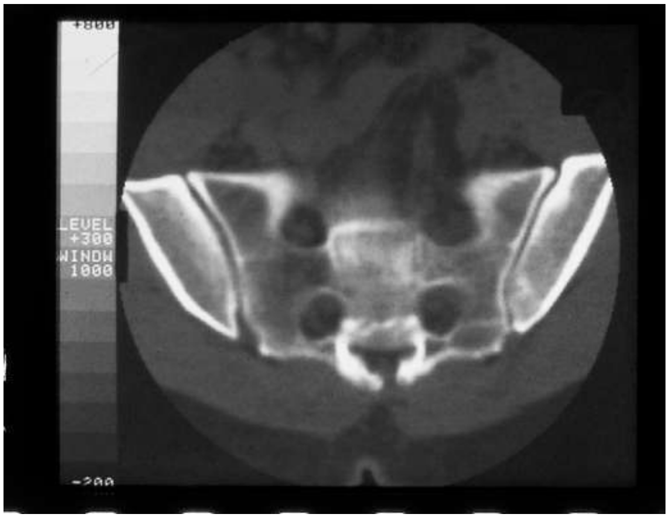

Fig. 2.3.2

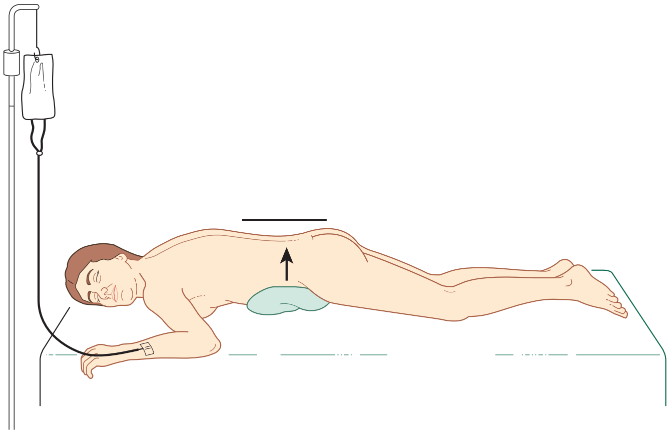

Fig. 2.3.3

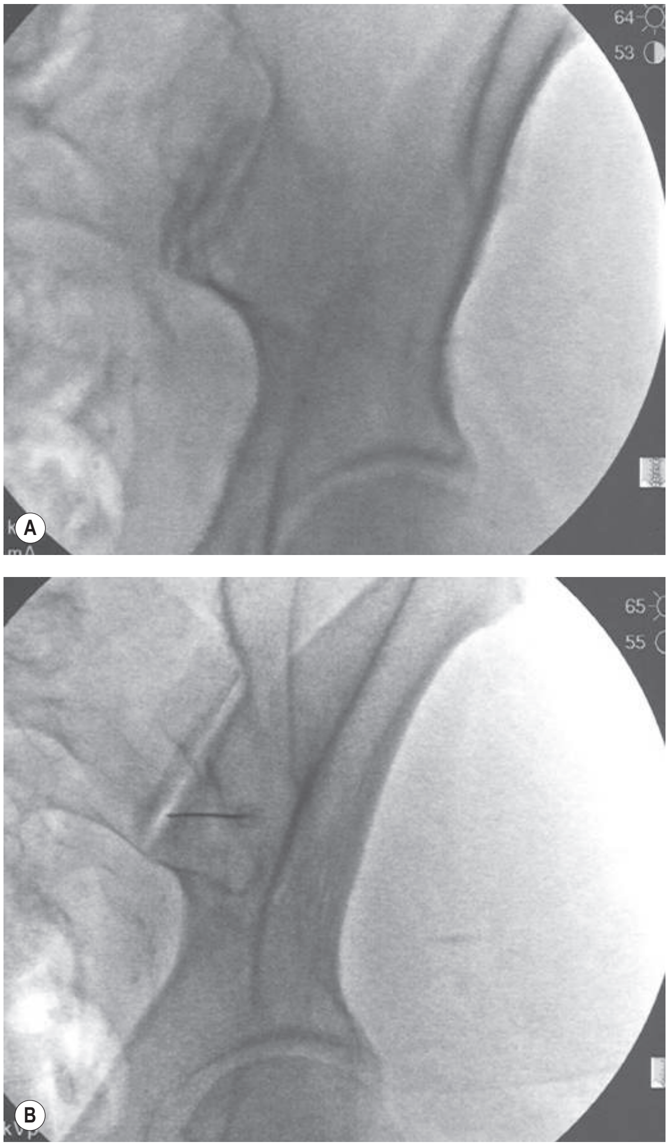

Fig. 2.3.4

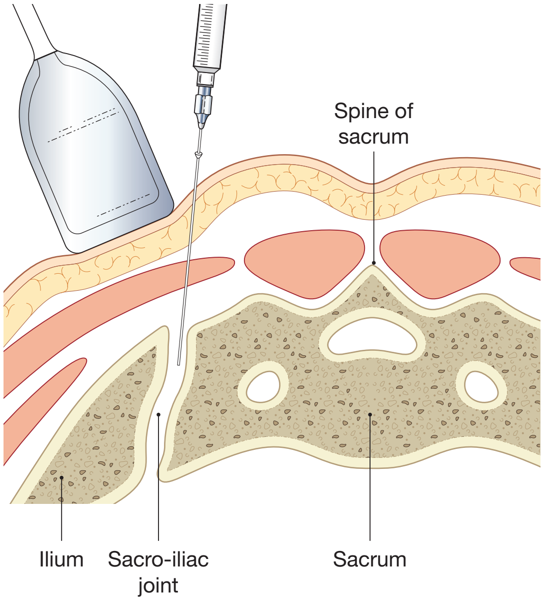

Fig. 2.3.5

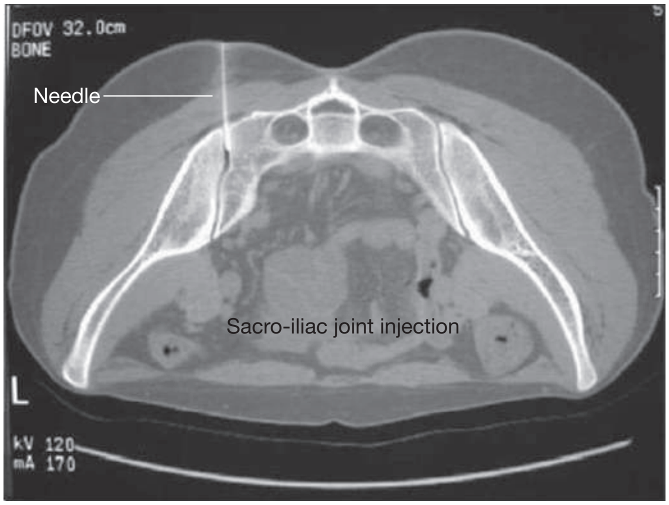

Fig. 2.3.6

are superimposed (usually about 10–20°) (*Fig. 2.3.4b*). The skin is marked and a skin wheal is raised. The area is infiltrated with lidocaine (lignocaine) 1% **over the joint line 1 cm above the most caudal point of the joint.**

- A 22 or 25 G 3½ in spinal needle is advanced no more than 1 cm into the joint. Some resistance is usually felt as the needle contacts the joint.
- A lateral view is then obtained. The needle should traverse no more than half the distance across the sacrum, and should never be advanced beyond the anterior cortex.
- Contrast dye, 0.5 to 1 ml, may be injected to ensure intra-articular spread. Intravascular injection is best detected during "live" fluoroscopy injection. In the AP view, dye should be seen within the joint space. Some extravasation outside the joint is common. If extensive, the needle should be repositioned.
- 1–2 ml lidocaine 1% is injected alone for diagnostic purposes. Reproduction of the patient's pain during needle positioning and injection as well as pain relief following the block will help confirm the sacro-iliac joint as the pain generator (avoid sedation with opioids for diagnostic procedures).
- Corticosteroid, e.g. triamcinolone diacetate 25 mg, plus 1–2 ml 1% lidocaine may be injected for therapeutic effect.
- Ultrasound may also be used to identify the joint (*Fig. 2.3.5*).
- CT scan may also be used to identify the joint but is not usually necessary (*Fig. 2.3.6*).

Confirmation of a successful injection

- Reproduction of pain during injection and relief of pain following injection confirms correct placement.
- Radiologic assessment of the X-ray image after injection of contrast medium may demonstrate tears in the joint capsule.

Tips

- While the joint may be easily entered, injection can be difficult where the joint is heavily invested with connective tissue and ligaments. This is especially true in elderly patients, where the joint is rigid and the joint space cannot expand to accommodate a volume of liquid. In such cases it may be possible to inject only as the needle is being removed from the joint.
- Sometimes, injection into the deep sacro-iliac ligaments around the joint may be helpful for pain relief. Introduction of a spinal needle just above the midline of the upper sacrum and advanced at 45° to the skin, under the rim of the ilium and in the direction of the joint, will access these ligaments. Lidocaine (lignocaine) 1% 4 ml plus triamcinolone diacetate 25 mg may then be injected.

Potential problems

- Discomfort on injection.
- Epidural injection.
- Sacral nerve root blockade.
- Subperiosteal injection (painful in the awake patient).

EPIDURAL INJECTION

3

Intervertebral disc disease may produce inflammation of spinal nerve roots, which may be the cause of radicular pain. The L5 and S1 nerve roots are most commonly affected, probably because they exit the bony canal through a narrow lateral bony recess, therefore increasing the likelihood of nerve compression and irritation. Lumbo-sacral radiculopathy consists of low-back pain that radiates a varying distance into the lower extremity, and which may be associated with motor and sensory loss consistent with damage to the affected nerve root. If bowel and bladder symptoms of dysfunction are present, large midline disc protrusion is suspected and prompt surgical intervention is indicated. Otherwise, if severe pain exists after treatment with immobilization and mild analgesics, epidural steroid injection may be carried out. Similarly, pain of thoracic or cervical disc origin may respond to epidural steroid injection.

Triamcinolone diacetate is the most commonly administered preparation and injection should be carried out as close to the affected nerve root as possible. Injection of a small amount of local anesthetic with the steroid will help to confirm drug placement and provide analgesia. In patients with S1 pathology the drug may not spread to the affected nerve root using the lumbar approach and the caudal approach to the epidural space may be required. Cervical epidural injection accesses the cervical spinal nerve roots, while in the thoracic region a paramedian approach is usually more successful.

3.1 LUMBAR EPIDURAL BLOCK

Anatomy

Structures encountered when inserting an epidural needle include skin, subcutaneous tissue, supraspinous ligament, interspinous ligament, ligamentum flavum (5–6 mm thick in the midline of the lumbar region, 3–5 mm thick in the midline of the thoracic region), prior to reaching the epidural space itself (**Fig. 3.1.1**). Beyond this space lies the dura mater, the arachnoid mater and intrathecal space containing the cerebrospinal fluid. The spinal cord usually ends at the L2 level (**Fig. 3.1.2**).

One should expect a distance of 3.5–6 cm from skin to the epidural space using a midline approach. In the lumbar region the spinous processes are generally perpendicular to the vertebral bodies (**Fig. 3.1.3**). In the thoracic region the spinous processes lie at an angle of 30–45° to the thoracic vertebral body, thus making midline epidural injection a little more difficult, and sometimes necessitating a paravertebral approach. Other relevant anatomy of the vertebral bodies is illustrated in **Fig. 3.1.4**.

Equipment

- 2 ml and 10 ml syringes
- 18 G, 20 G, and 25 G needles
- ECG, BP, and SpO_2 monitors
- 18 G epidural set (*Fig. 3.1.5*)
- Filter aspiration needle
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy (optional)
- Ultrasound (optional)

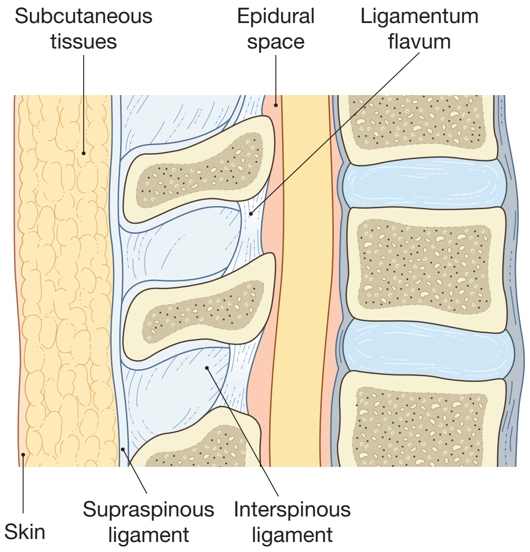
Subcutaneous tissues
Epidural space
Ligamentum flavum
Skin
Supraspinous ligament
Interspinous ligament

Fig. 3.1.1

Drugs

- Lidocaine (lignocaine) 1%, 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 50 mg (or its equivalent)
- Saline (NaCl) 10 ml
- Resuscitation drugs (*see Appendix 3*)

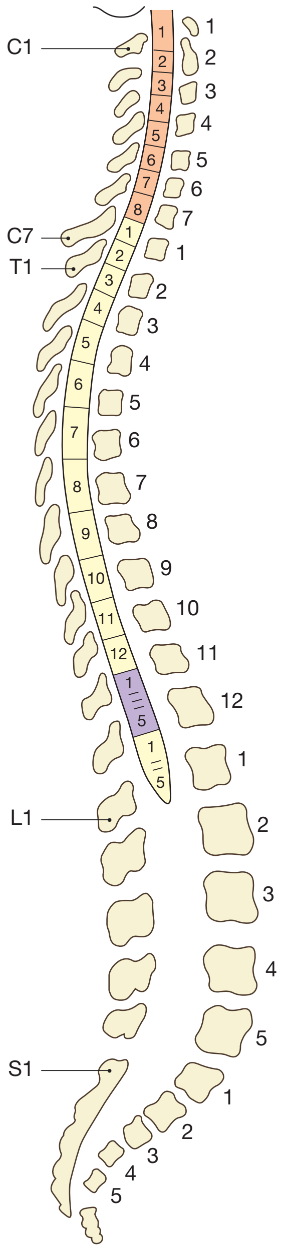
C1
C7
T1
L1
S1
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
9
10
11
12
1
5
1
5
1
2
3
4
5
6
7
1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4
5
1
2
3
4
5

Fig. 3.1.2

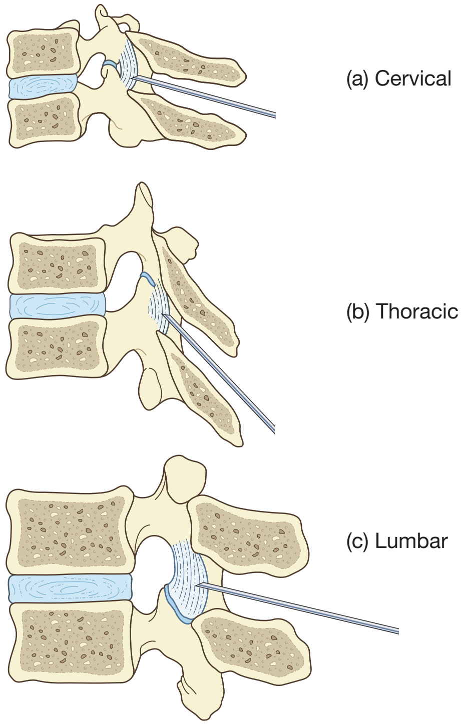
(a) Cervical
(b) Thoracic
(c) Lumbar

Fig. 3.1.3

Position of patient

- Lateral, usually lying on the side of the radiculopathy.
- Shoulders and buttocks parallel to the edge of the bed, perpendicular to the floor, with spine flexed.

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- Lidocaine (lignocaine) 1% 2 ml is drawn up into three 2 ml syringes.
- Lidocaine (lignocaine) 1% 2 ml, plus corticosteroid, e.g. triamcinolone diacetate 50 mg is drawn up into the 10 ml syringe.
- NaCl 10 ml is drawn up into the 10 ml loss-of-resistance syringe.
- The iliac crest is palpated and the intercrestal line (this corresponds with the inferior aspect of the spinous process of L4 or may lie in the L4–5 interspace) is identified (*Fig. 3.1.6*).
- The spinous processes are palpated, and the level requiring injection is identified.
- This may be confirmed by fluoroscopy or ultrasound.

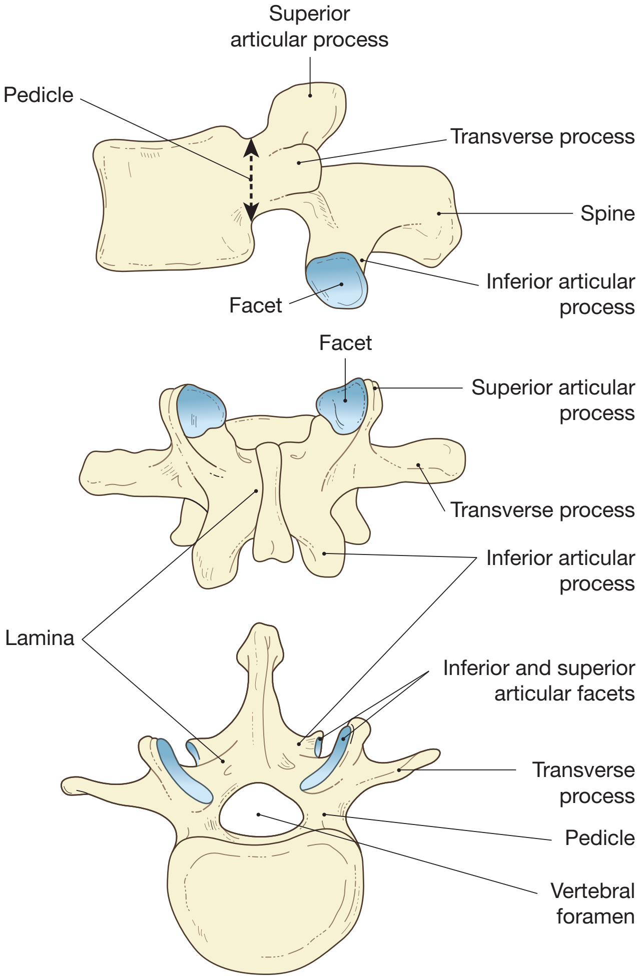
Superior articular process
Pedicle
Transverse process
Spine
Inferior articular process
Facet
Facet
Superior articular process
Transverse process
Inferior articular process
Lamina
Inferior and superior articular facets
Transverse process
Pedicle
Vertebral foramen

Fig. 3.1.4

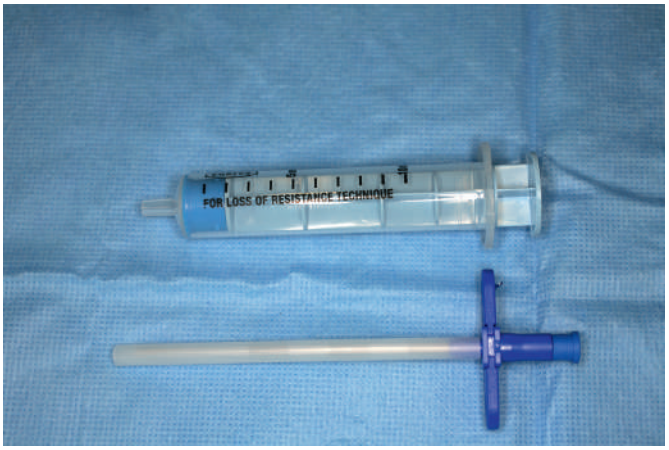
FOR LOSS OF RESISTANCE TECHNIQUE

Fig. 3.1.5

MIDLINE APPROACH FOR THE RIGHT-HANDED OPERATOR

With the left hand

- The fore- and middle fingers are placed each side of the interspace.

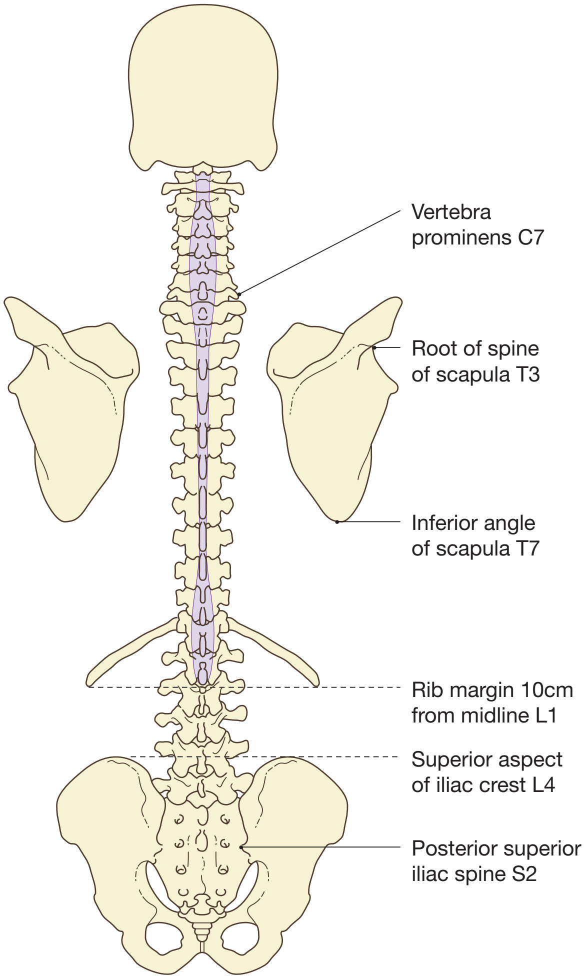
Vertebra prominens C7
Root of spine of scapula T3
Inferior angle of scapula T7
Rib margin 10cm from midline L1
Superior aspect of iliac crest L4
Posterior superior iliac spine S2

Fig. 3.1.6

- These fingers are kept in place until the epidural needle is gripped by the interspinous ligament.

With the right hand

- The interspinous ligament is infiltrated with lidocaine (lignocaine) 1% 2 ml.
- **The epidural needle is inserted, bevel facing the side of the radiculopathy, between the fore- and middle fingers of the left hand in a direction 60° cephalad, perpendicular to the spine, parallel to the floor, until it is gripped by the interspinous ligament** (*Fig. 3.1.7 a,b*). In the case of thoracic epidural injection (midline approach), the point of entry of the needle should be as close as possible to the caudal end of the interspinous space and the needle directed 30–45° cephalad to enter between the spinous processes.
- The hub of the needle is gripped with the fore- and middle fingers of the left hand and this hand is steadied by leaning the wrist against the patient's back.
- The stylet is removed and the loss-of-resistance syringe is applied.

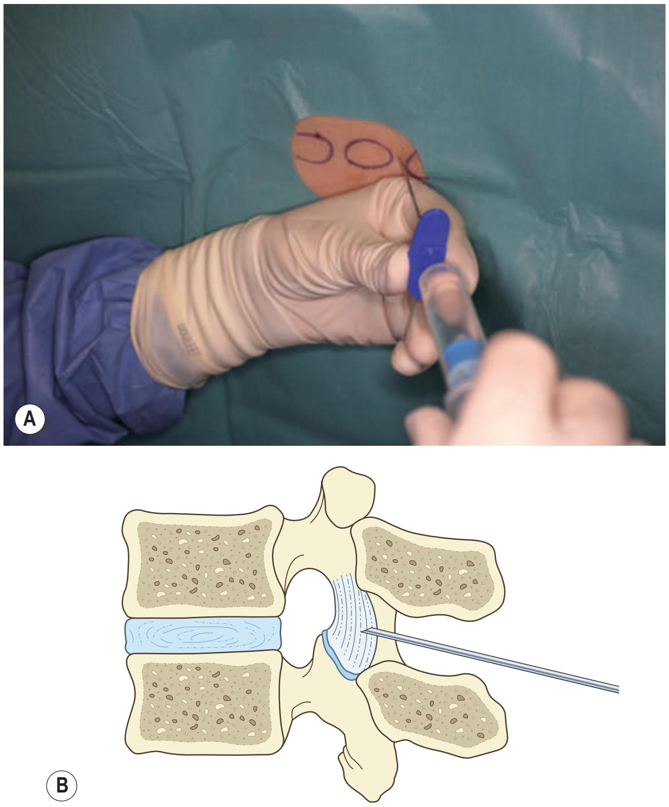

Fig. 3.1.7

- The needle is slowly and carefully advanced, while constant pressure is applied to the plunger, the left hand aiding the advance, while at the same time applying a brake if required (*Fig. 3.1.8 a,b*).
- At the point at which the needle enters the ligamentum flavum, absolute resistance to injection is experienced.
- At this point the needle is advanced very slowly until a sudden loss of resistance to the pressure on the plunger is experienced, the point at which the epidural space is entered.
- After negative aspiration for blood or cerebrospinal fluid (CSF), lidocaine (lignocaine) 1% 3 ml is injected.
- After 5 minutes the patient is questioned about any changes in sensation or power, and any changes in heart rate or blood pressure are noted.
- If the injection is for diagnostic purposes only, the needle may be removed at this point.
- If therapeutic effect is required, lidocaine (lignocaine) 1% 2 ml, plus corticosteroid if indicated, e.g. triamcinolone diacetate 50 mg (or its equivalent), may be injected. Alternatively, a catheter may be inserted through the needle if indicated.
- The needle is flushed with NaCl 1 ml and removed.
- The patient is allowed to lie in the lateral position, on the side of the pain.

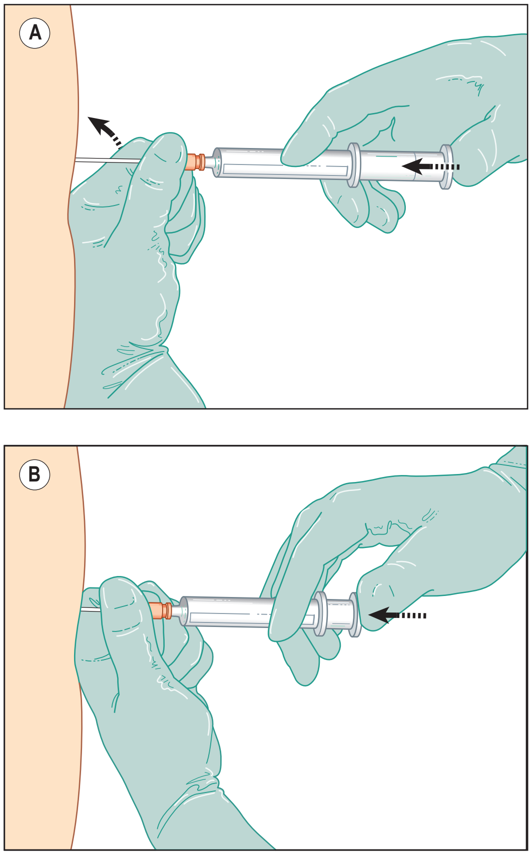

Fig. 3.1.8

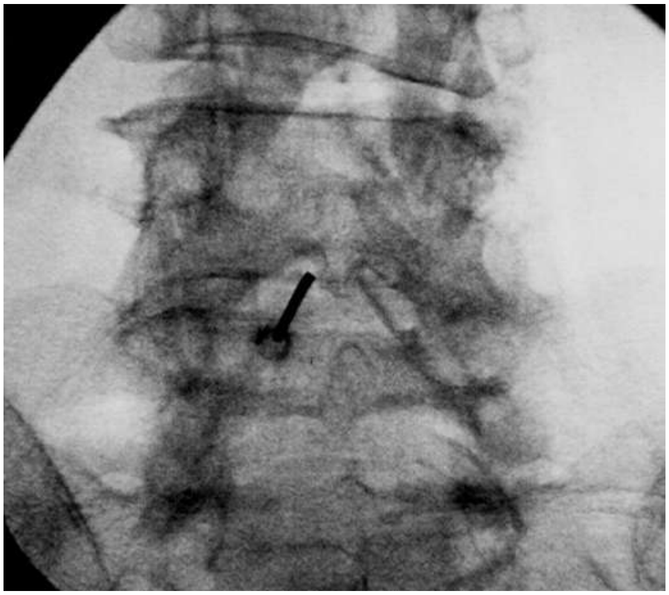

Fig. 3.1.9

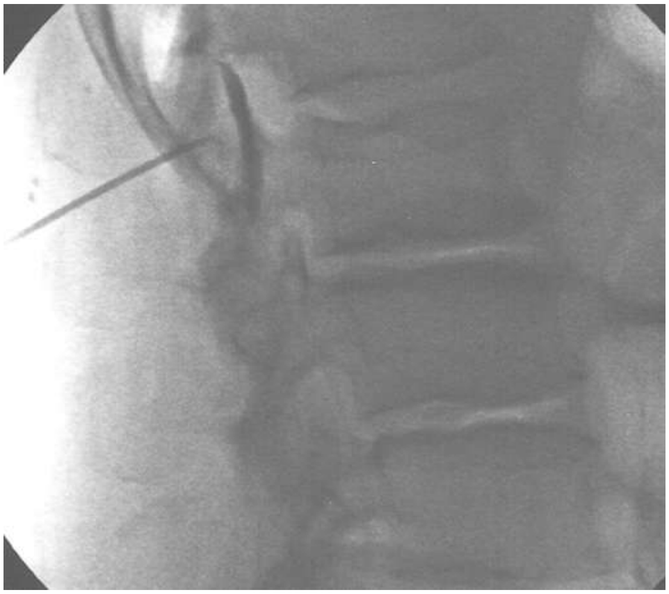

Fig. 3.1.10

- Monitors should be left attached and i.v. access left in situ for at least 30 minutes.
- The patient is advised to contact the hospital should the anesthesia remain after several hours.

FLUOROSCOPIC GUIDED LUMBAR EPIDURAL INJECTION

- Position the patient prone with a pillow under the lower abdomen to increase lumbar flexion.
- The interlaminar space for the desired segmental level is identified fluoroscopically using a straight AP view. Angling the fluoroscope slightly cephalad may open the space if it appears very narrow.
- Prepare skin with antiseptic and sterile drape.
- Raise a local anesthetic skin wheal just below the interlaminar space, about 0.5 cm from the midline toward the symptomatic side.
- Advance the Tuohy needle, angling slightly toward the midline until resistance of the ligamentum flavum is encountered, keeping the trajectory just lateral to the midline to avoid contacting the spinous process. Repeat imaging periodically to ensure that the needle is approaching the space, not the lamina or spinous process (see *Fig. 3.1.9*). Advance the needle through the ligamentum flavum using a loss of resistance technique with air or saline. Once loss of resistance is achieved, obtain a lateral view to ensure the needle is barely into the bony spinal canal.
- Attach a low volume extension set to the needle, aspirate to ensure there is no blood return, and inject 0.5–1 ml of non-ionic contrast medium during live fluoroscopy. Dye should be seen spreading within the bony canal (see *Fig. 3.1.10*). Prior to injecting local anesthetic or steroid, obtain an AP view to reconfirm epidural spread.

- Aspirate again, then inject local anesthetic (1–2 ml) and 50 mg triamcinolone diacetate or equivalent.
- Loss of resistance is occasionally encountered with the needle superficial to the epidural space. Dye will be seen dorsal to the epidural space on the lateral view, and spread lateral to the spinal canal will be seen on the AP view. The needle can then be advanced through the ligamentum flavum, again using loss of resistance, followed by dye confirmation.

Confirmation of a successful block

- Relief of pain.
- Anesthesia in the distribution of the blocked nerves.
- For lumbar epidural steroid injection, improvement in straight-leg raising may be evident.

Tips

- Air may be used instead of NaCl to determine loss of resistance. If this technique is used it is advisable to avoid constant pressure on the plunger, as the air is compressible; instead it should be bounced intermittently with the thumb to test for resistance and loss of resistance.
- Advocates claim identification of CSF is easier with this technique.
- Advocates of the use of NaCl point out that absolute resistance to pressure identifies the ligamentum flavum, and that by applying constant pressure to the plunger one can identify loss of resistance earlier, thereby more easily avoiding the possibility of dural tap.
- An epidural catheter may be inserted through the needle and the needle removed, taking care not to withdraw the catheter when removing the needle. However, a test dose of lidocaine (lignocaine) 1% 4 ml with/without epinephrine (adrenaline) 1:200 000 is given after insertion, before any injection through the catheter is carried out.
- Ultrasound may guide the insertion of the needle as spinous proccesses are easily visible on ultrasound (optional) (*Fig. 3.1.11*).

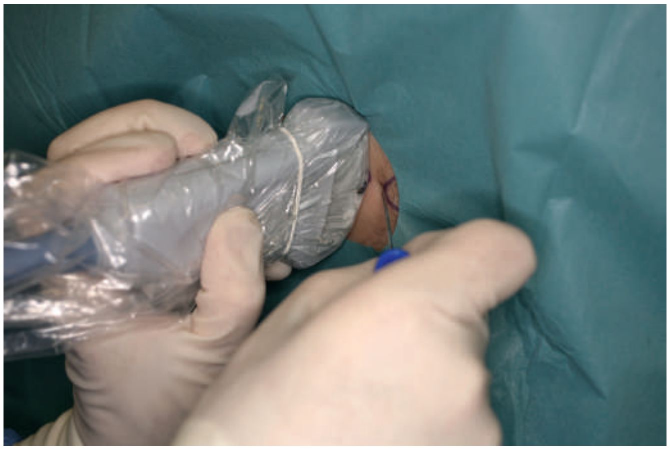

Fig. 3.1.11

- Injection of radio-opaque dye under direct fluoroscopy can confirm epidural placement.
- Insertion of a radio-opaque epidural catheter may be carried out also under fluoroscopy.

Potential problems

IMMEDIATE

- Failure to locate epidural space (sitting position may be successful).
- Intravascular injection (test dose important); addition of epinephrine (adrenaline) to test dose may help identification of intravascular injection.
- Intrathecal injection (test dose important).
- Hypotension due to sympathetic blockade (give i.v. fluid; consider ephedrine).
- Headache (possible dural puncture).
- Allergic reaction.

LATER

- Infection (epidural abscess; bacterial meningitis).
- Aseptic meningitis; usually a result of intrathecal injection (test dose important).
- Cushingoid symptoms; usually as a result of repeated steroid injections.

3.2 THORACIC EPIDURAL BLOCK

Anatomy

In the thoracic region, the spinous processes lie at an angle of 30–45° to the thoracic vertebral body (**Fig. 3.2.1 a,b**) thus making midline epidural injection a little more difficult, and sometimes necessitating a paramedian approach (**Fig. 3.2.2 a–d**).

Equipment

- 2 ml and 10 ml syringes
- 18 G, 20 G, and 25 G needles
- ECG, BP, and SpO_2 monitors
- 18 G epidural set
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy or ultrasound (optional)

Drugs

- Lidocaine (lignocaine) 1%, 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 50 mg (or its equivalent)
- Saline (NaCl) 10 ml
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Lateral, usually lying on the side of the radiculopathy.
- Shoulders and buttocks parallel to the edge of the bed, perpendicular to the floor, with spine flexed.

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- Lidocaine (lignocaine) 1% 2 ml is drawn up into three 2 ml syringes.
- Lidocaine (lignocaine) 1% 2 ml, plus corticosteroid, e.g. triamcinolone diacetate 50 mg, is drawn up into the 10 ml syringe.
- NaCl 10 ml is drawn up into the 10 ml loss-of-resistance syringe.
- The spinous processes are palpated, and the level requiring injection is identified.
- This may be confirmed by fluoroscopy or ultrasound.

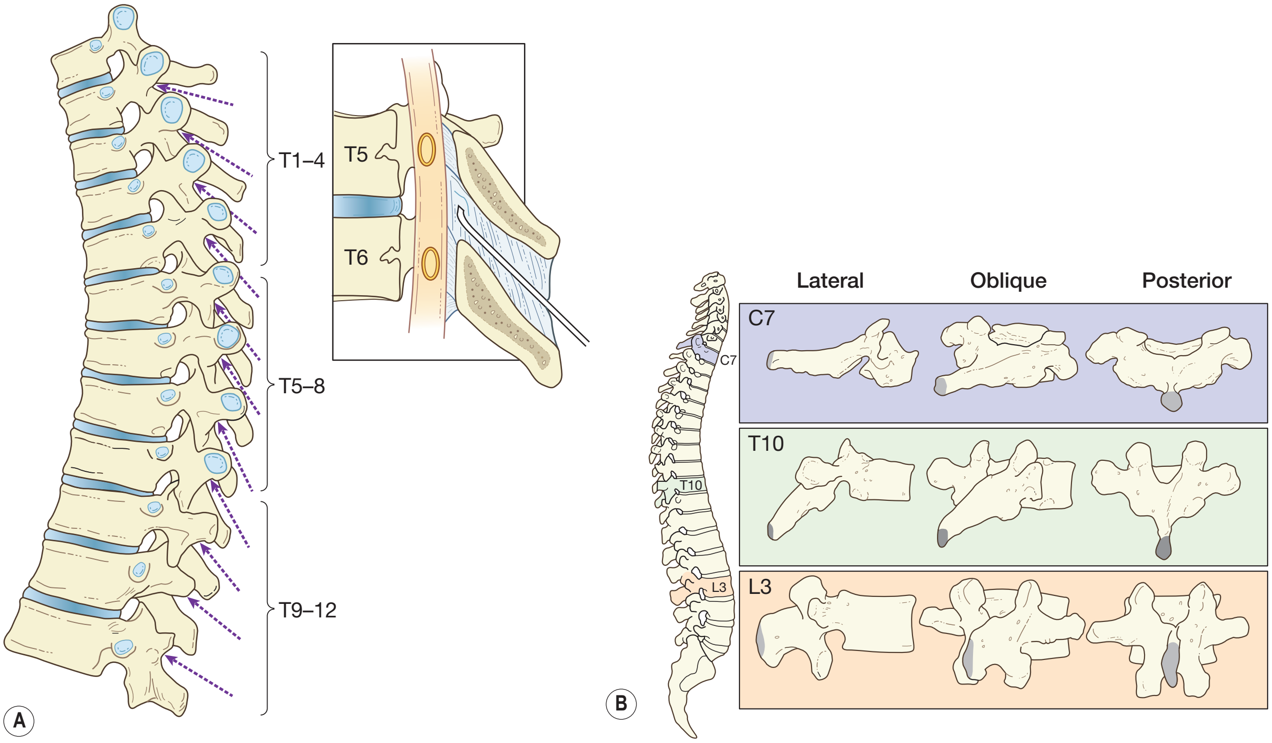

Fig. 3.2.1

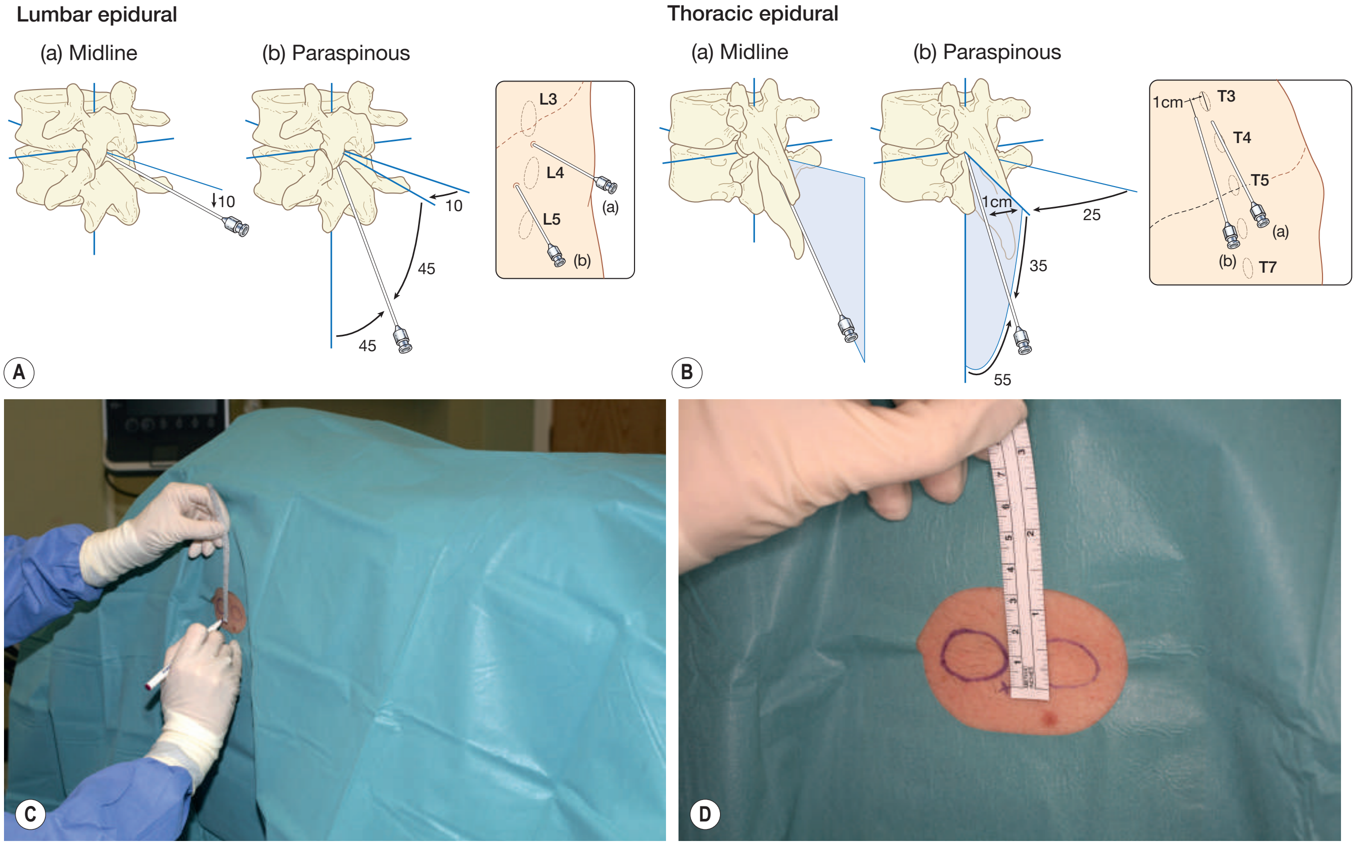

Fig. 3.2.2

Paramedian approach

FOR THE RIGHT-HANDED OPERATOR

With the left hand

- The fore- and middle fingers are placed each side of the interspace.
- These fingers are kept in place until the epidural needle is gripped by the interspinous ligament.

With the right hand

- The interspinous ligament is infiltrated with lidocaine (lignocaine) 1% 2 ml.
- **The insertion point of the epidural needle in the paravertebral approach lies 1 cm lateral to the midline, at the lower border of the spinous process** (*Fig. 3.2.3 a,b*). The epidural needle is inserted, bevel facing the side of radiculopathy, between the fore- and middle fingers of the left hand, perpendicular to the spine, parallel to the floor, until it is gripped by the interspinous ligament.
- **The direction of the needle is 130° cephalad and 15° medial to the midline.** Care must be taken as the ligamentum flavum is not as thick laterally, and may not be identified as easily. Therefore, it is usually easiest to first identify the lamina and **walk the needle off the lamina in a cephalad direction** until the needle enters the ligamentum flavum. At that point the loss-of-resistance technique may be performed.
- The hub of the needle is gripped with the fore- and middle fingers of the left hand and this hand is steadied by leaning the wrist against the patient's back.
- The stylet is removed and the loss-of-resistance syringe is applied.
- The needle is slowly and carefully advanced until the osseous endpoint of the lamina is encountered.
- It is then walked off the lamina in the cephalad direction until it enters the ligamentum flavum.
- At the point at which the needle enters the ligamentum flavum, absolute resistance to injection is experienced.
- It is then carefully advanced further while constant pressure is applied to the plunger, the left hand aiding the advance, while at the same time applying a brake if required (*Fig. 3.2.4*, viewed from above).
- The needle is advanced very slowly until a sudden loss of resistance to the pressure on the plunger is experienced, the point at which the epidural space is entered.
- After negative aspiration for blood or cerebrospinal fluid, lidocaine (lignocaine) 1% 2 ml is injected.

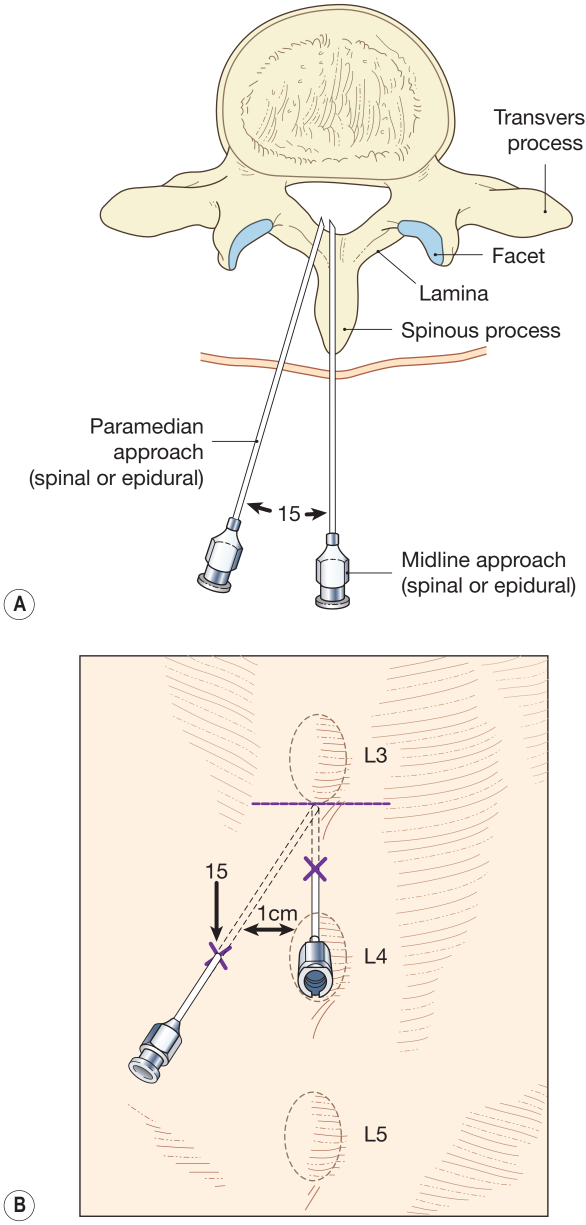
Transverse process
Facet
Lamina
Spinous process
Paramedian approach (spinal or epidural)
15
Midline approach (spinal or epidural)
A
L3
15
1cm
L4
L5
B

Fig. 3.2.3

- After 5 minutes the patient is questioned about changes in sensation or power, and any changes in heart rate or blood pressure are noted.
- If the injection is for diagnostic purposes only, the needle may be removed at this point.
- If therapeutic effect is required, lidocaine (lignocaine) 1% 2 ml plus corticosteroid, e.g. triamcinolone diacetate 50 mg, may be injected.
- The patient is allowed to lie in the lateral position, on the side of the pain.
- Monitors should be left attached and i.v. access should remain in situ for at least 30 minutes.

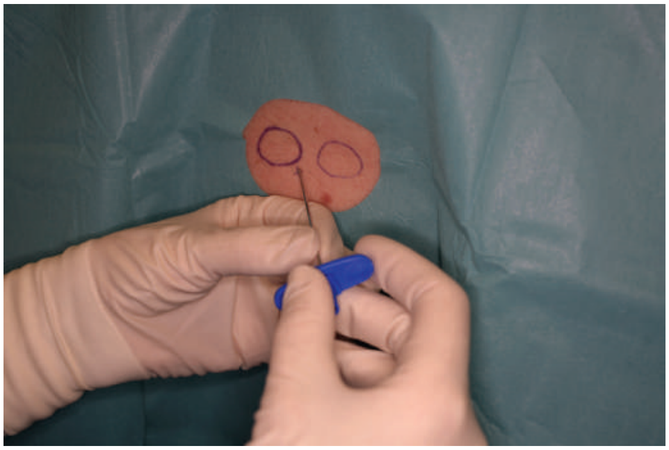

Fig. 3.2.4

- The patient is advised to contact the hospital should anesthesia remain after several hours.

Confirmation of a successful block

- Relief of pain.
- Anesthesia in the distribution of blocked nerves.
- For lumbar epidural steroid injection, improvement in straight-leg raising may be evident.

Tips

- Air may be used instead of NaCl to determine loss of resistance. If this technique is used it is advisable to avoid constant pressure on the plunger, as the air is compressible; instead the plunger should be bounced intermittently with the thumb to test for resistance and loss of resistance.
- Advocates claim identification of CSF is easier with this technique.
- Advocates of the use of NaCl point out that absolute resistance to pressure identifies the ligamentum flavum, and that by applying constant pressure to the plunger one can identify loss of resistance more immediately, thereby avoiding the possibility of dural tap more easily.
- An epidural catheter may be inserted through the needle and the needle removed, taking care not to withdraw the catheter when removing the needle. However, a test dose of lidocaine (lignocaine) 1% 4 ml with epinephrine (adrenaline) 1:200 000 is given after insertion, before any injection through the catheter is carried out.
- Identification of the insertion point may be aided by ultrasound (*Fig. 3.2.5*).
- Injection of radio-opaque dye under direct fluoroscopy can confirm epidural placement.
- Insertion of a radio-opaque epidural catheter may be carried out also under fluoroscopy.

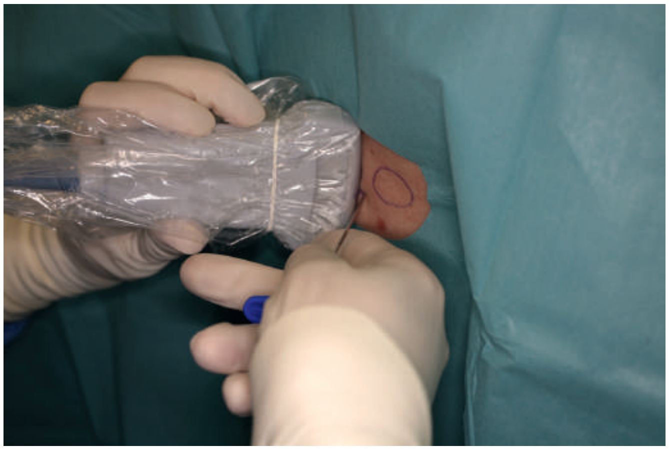

Fig. 3.2.5

Potential problems

IMMEDIATE

- Failure to locate epidural space (sitting position may be successful).
- Intravascular injection (test dose important); addition of epinephrine (adrenaline) to test dose may help identification of intravascular injection.
- Intrathecal injection (test dose important).
- Hypotension due to sympathetic blockade (give i.v. fluid, consider ephedrine).
- Headache (possible dural puncture).
- Allergic reaction.
- Spinal cord injury may occur if the epidural space is not recognized. Deep sedation should be avoided during needle insertion and drug injection.

LATER

- Infection (epidural abscess, bacterial meningitis).
- Aseptic meningitis, usually the result of intrathecal injection (test dose important).
- Cushingoid symptoms (usually as a result of repeated injections).
- Epidural hematoma. This complication should be suspected when sensory or motor function loss occurs minutes to hours after the procedure. Immediate diagnostic imaging (CT or MRI) is essential. Prompt surgical decompression may be required.

3.3 CERVICAL EPIDURAL BLOCK

Anatomy

In the cervical region, the spinous processes are almost perpendicular to the vertebral bodies, especially in the lower part. They also widen and become bifid. As a result, insertion of the needle is often easy. However, it must be remembered that the epidural space is relatively narrow in this area (2–4 mm), and that the spinal cord lies very close to it (**Fig. 3.3.1 a,b**). Most workers prefer to use the "hanging drop" technique when accessing the cervical epidural space, as there exists a significant negative pressure in the cervical region in the sitting position.

Equipment

- 2 ml, 5 ml, and 10 ml syringes
- 18 G, 20 G, and 25 G needles
- ECG, BP, and SpO_2 monitors
- 18 G epidural set
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy or ultrasound (optional)

Drugs

- Lidocaine (lignocaine) 1%, 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 50 mg (or its equivalent)
- Saline (NaCl) 10 ml
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Sitting.
- Head flexed forward.

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- Lidocaine (lignocaine) 1% 2 ml is drawn up into two 2 ml syringes.
- Lidocaine (lignocaine) 1% 1 ml plus corticosteroid, e.g. triamcinolone diacetate 50 mg, is drawn up into a 5 ml syringe.
- NaCl 10 ml is drawn up into a 10 ml syringe.
- The patient is allowed to sit up straight for a moment, and the spinous process of T3, which lies opposite the root of the spine of the scapula, is identified and marked. The prominent spinous process of C7 (vertebra prominens) is identified (*Fig. 3.3.2*) and marked (*Fig. 3.3.3 a,b*). Ultrasound can be used to guide needle placement (*Fig. 3.3.3 c,d*). The interspace to be used for epidural injection is also marked.

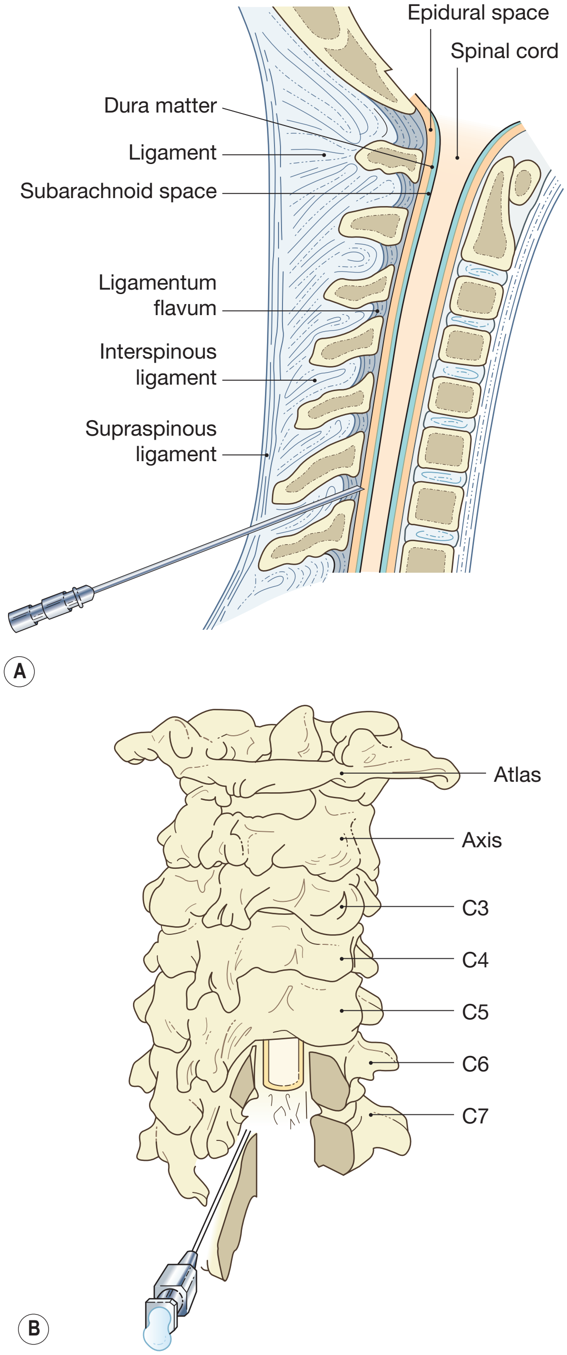

Fig. 3.3.1

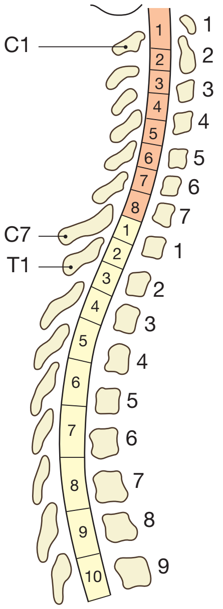

Fig. 3.3.2

FOR THE RIGHT-HANDED OPERATOR

With the left hand

- The fore- and middle fingers are placed each side of the interspace.
- These fingers are kept in place until the epidural needle is gripped by the interspinous ligament.

With the right hand

- The interspinous ligament is infiltrated with lidocaine (lignocaine) 1% 2 ml.
- **The epidural needle is inserted, bevel facing caudad, between the fore- and middle fingers of the left hand in a direction 60° cephalad, until it is gripped firmly by the interspinous ligament.**
- The hub of the needle is gripped with the fore- and middle fingers of the left hand and this hand is steadied by leaning the wrist against the patient's spine.
- The stylet is removed.
- The hub of the epidural needle is filled with saline (*Fig. 3.3.4*) such that a "hanging drop" appears (*Fig. 3.3.5*).

With both hands

- The wings of the epidural needle are gripped with each hand, steadying the hands by resting the wrists against the posterior thoracic wall (*Fig. 3.3.6*).
- The needle is slowly and carefully advanced with both hands.
- It is prudent periodically to confirm high resistance of the needle in the ligament by testing with an air-filled syringe, then replace the stylet to make sure there is no tissue blocking the needle before resuming the "hanging drop" technique.

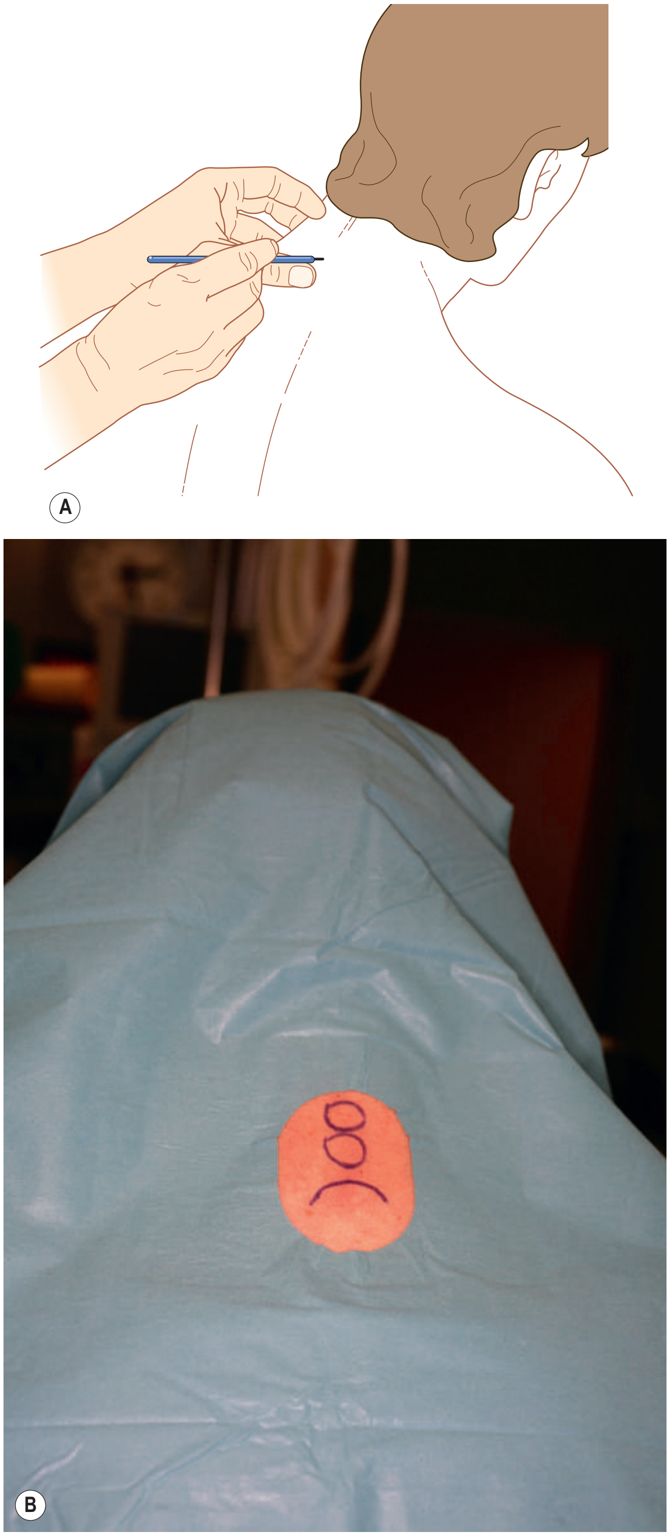

Fig. 3.3.3

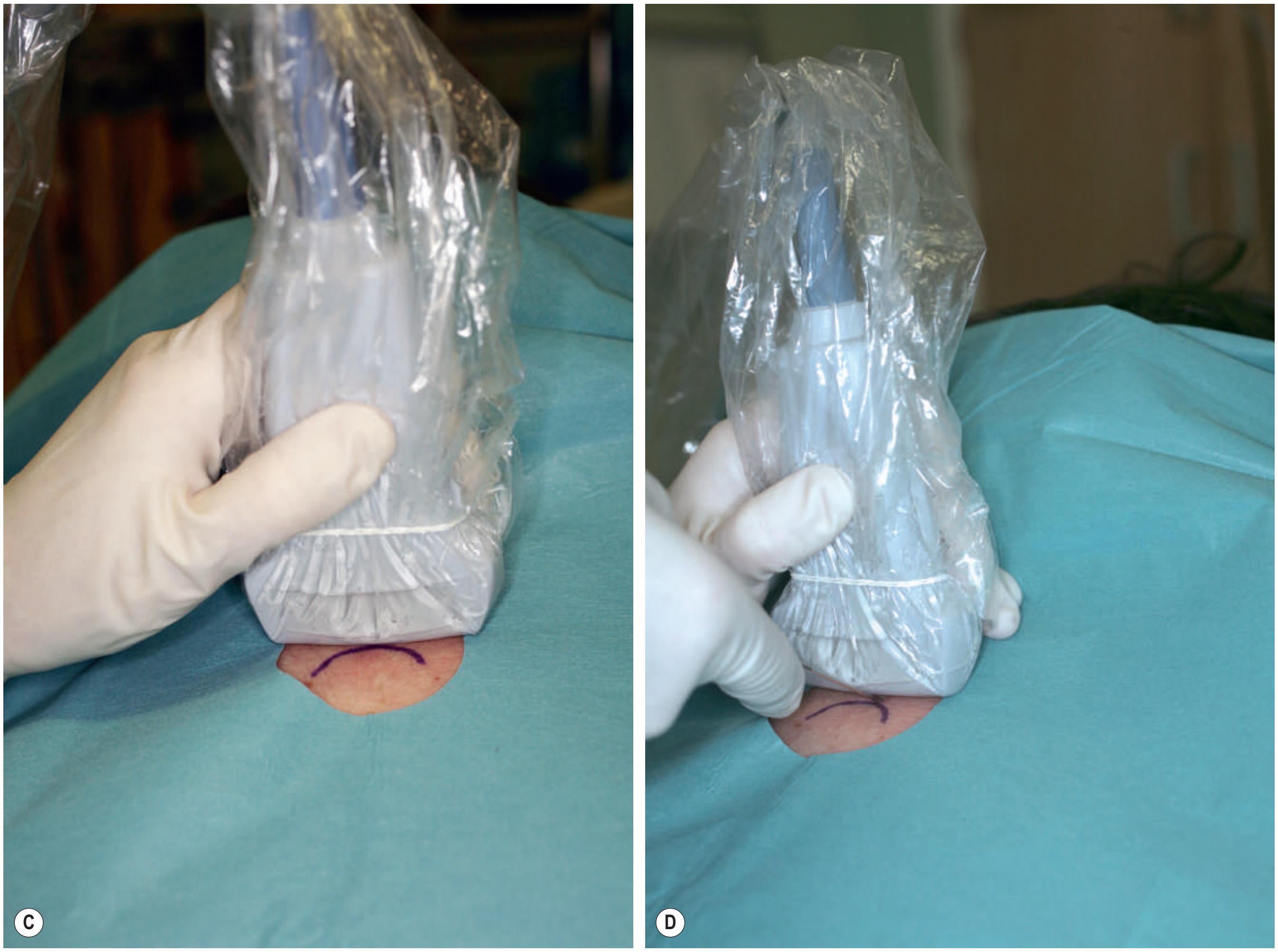

Fig. 3.3.3, cont'd

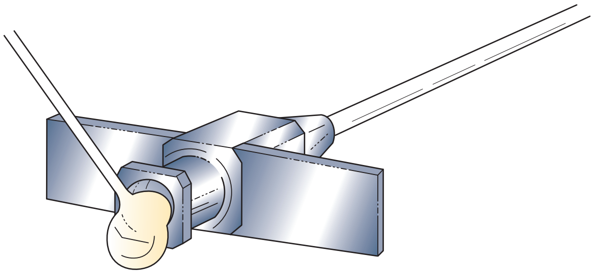

Fig. 3.3.4

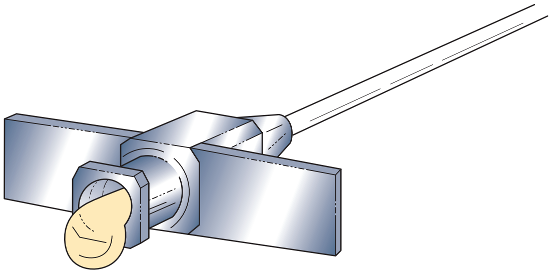

Fig. 3.3.5

- The "hanging drop" at the hub of the needle is watched closely, and the patient is asked periodically about the presence of paresthesia.
- At the point at which the epidural needle enters the epidural space, the drop should appear to be sucked into the needle (*Figs 3.3.7*, *3.3.8*).
- After negative aspiration for blood or CSF, lidocaine (lignocaine) 1% 2 ml is injected.
- After 1–2 minutes the patient is questioned about changes in sensation or power, and any changes in heart rate or blood pressure are noted.
- If the injection is for diagnostic purposes only, the needle may be removed at this point.
- If therapeutic effect is required, lidocaine (lignocaine) 1% 1 ml plus corticosteroid, e.g. triamcinolone diacetate 50 mg, may be injected.
- The patient is allowed to lie in the lateral position, on the side of the pain.
- Monitors should be left attached and i.v. access kept in situ for at least 30 minutes.

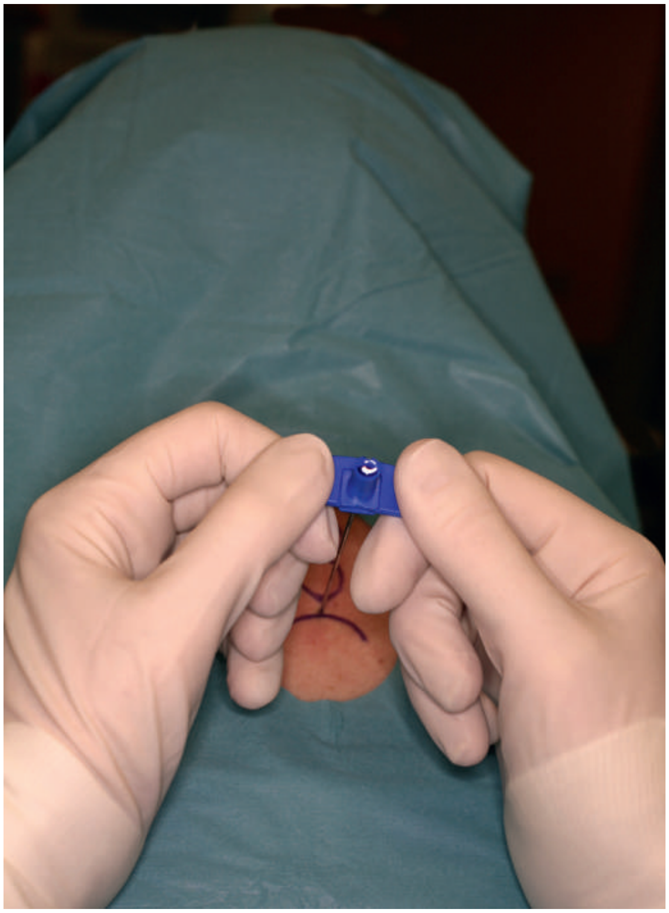

Fig. 3.3.6

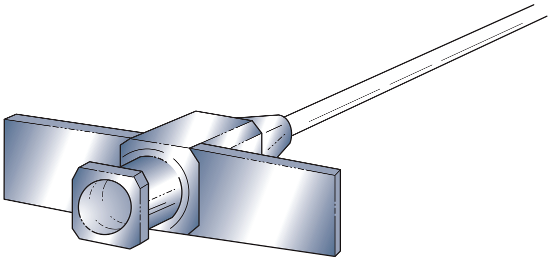

Fig. 3.3.7

- Ultrasound may aid in identifying the interspinous space as spinous processes are easily visible.

FLUOROSCOPIC GUIDED CERVICAL EPIDURAL INJECTION

- Check MRI to ensure that the spinal cord is not displaced posteriorly. If the posterior epidural space is compromised or the spinal cord is shifted posteriorly, it is safer to enter the upper thoracic epidural space and advance a radio-opaque catheter to the low cervical level.
- Position patient prone, with pillow under shoulders, neck flexed, arms at sides, shoulders as far downward as possible.

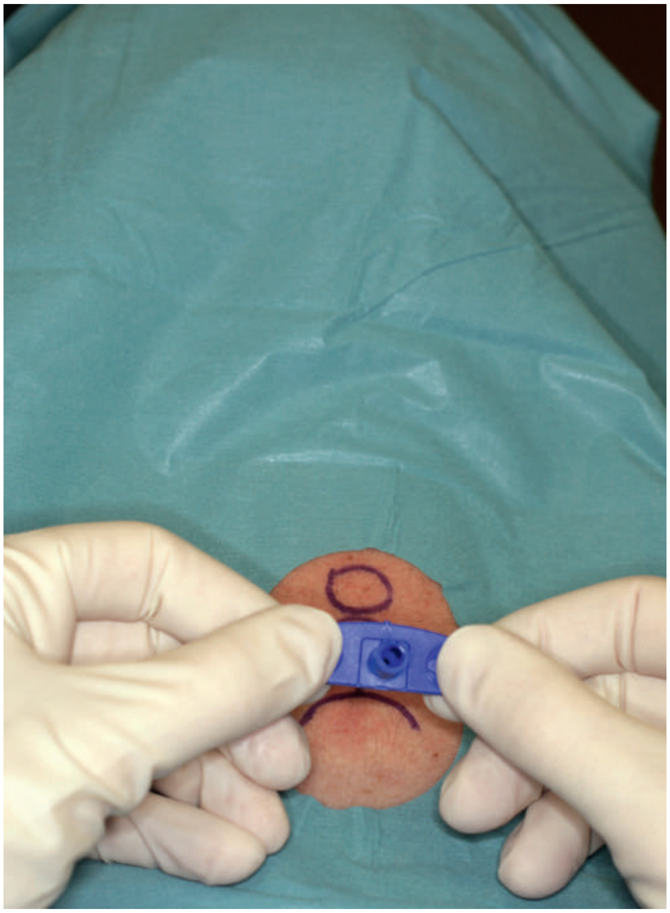

Fig. 3.3.8

- Identify targeted interlaminar space (T1–2, C7–T1, or C6–7) in direct AP fluoroscopic view. Adjust angle upward or downward slightly to maximize view of interlaminar space.
- Mark skin over lower border of T1–2, C7–T1 (preferred) or C6–7 interlaminar space just lateral to the midline. Do not perform epidural injection above C6–7 because of absence of midline epidural fat at higher levels.
- Prepare skin with antiseptic and sterile drape.
- Provide only minimal sedation or no sedation. Instruct patient to report any pain or paresthesia during the procedure.
- Infiltrate skin and subcutaneous tissue with lidocaine (lignocaine) 1%.
- Advance the Tuohy needle through the insertion point into the interspinous ligament and check the fluoroscopy image to ensure the needle tip is directed toward the midline (*Fig. 3.3.9*).
- Begin to advance through the ligament using loss of resistance with air or saline. Check lateral view if possible (may be obscured by the shoulders). Proceed with needle advancement. When loss of resistance

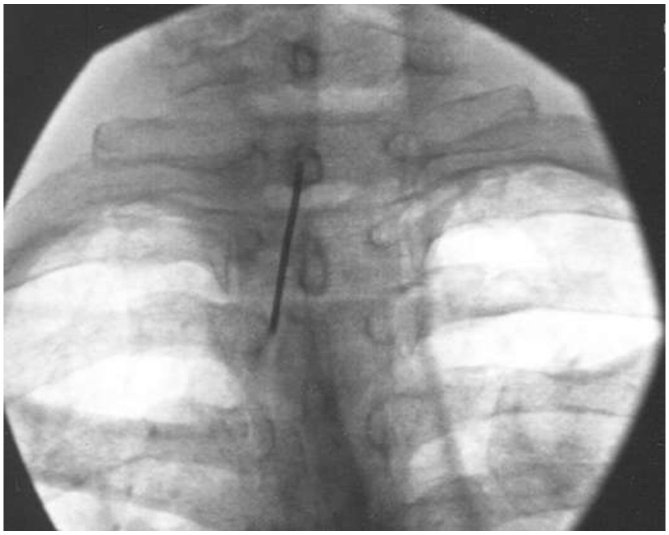

Fig. 3.3.9

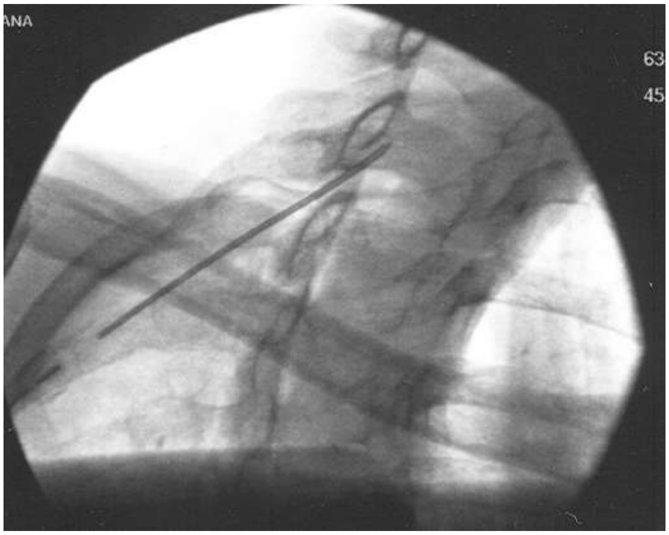

Fig. 3.3.10

occurs, recheck the lateral view. If unable to visualize the spinal canal, obtain slightly oblique view (*Fig. 3.3.10*).

- Attach a low volume extension set to the needle and inject a small volume (1 ml or less) of non-ionic contrast medium under live fluoroscopy, preferably in lateral view (AP or oblique view is used if this is not possible) (*Fig. 3.3.11*), then observe the AP view to confirm epidural dye spread.
- Aspirate to ensure there is no blood return, then inject 1–2 ml lidocaine (lignocaine) 1% followed by 25–50 mg triamcinolone diacetate or equivalent.

Confirmation of a successful block

- Relief of pain.

Tips

- Loss-of-resistance techniques may also be used to access the cervical epidural space.
- The steroid may be given soon after the test dose, as hypotension may be a problem if the patient remains in the sitting position.

Potential problems

IMMEDIATE

- Failure to locate epidural space (lateral position with loss of resistance technique may be successful).
- Pain on injection (caution: close proximity to spinal cord).
- Intrathecal injection (test dose important).
- Hypotension ± bradycardia due to sympathetic blockade (maximum 3 ml local anesthetic administered in this technique).
- Vasovagal syncope is common in young adult patients.

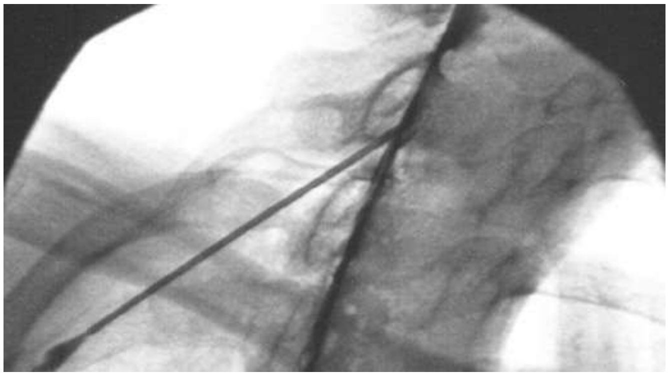

Fig. 3.3.11

- Intravascular injection; addition of epinephrine (adrenaline) 1:200 000 to the test dose may aid identification of intravascular injection.
- Headache (possible dural puncture).
- Allergic reaction.
- Spinal cord injury may occur if the epidural space is not recognized. Deep sedation should be avoided during needle insertion and drug injection.

LATER

- Infection (epidural abscess, bacterial meningitis).
- Aseptic meningitis, usually the result of intrathecal injection (test dose important).
- Cushingoid symptoms (usually as a result of repeated injections).
- Epidural hematoma. This complication should be suspected when sensory or motor function loss occurs minutes to hours after the procedure. Immediate diagnostic imaging (CT or MRI) is essential. Prompt surgical decompression may be required.

3.4 CAUDAL EPIDURAL BLOCK

Anatomy

Injection of anesthetic through the sacral hiatus allows access to the sacral epidural space or caudal anesthesia. The sacrum is roughly triangular in shape and is made up of five fused sacral vertebrae (**Fig. 3.4.1**). Its dorsal aspect is convex and there is a midline sacral canal that allows passage of the sacral nerves through four pairs of foramen, anteriorly and posteriorly. At the caudal end lies the coccyx, and at the cephalad end lies the fifth lumbar vertebra. The posterior wall of S5, and sometimes S4, is unfused. The thick fibrous sacro-coccygeal membrane or sacral hiatus covers the defect. This may be variable in size as the posterior wall of other sacral vertebrae may also be unfused. Penetration of this membrane allows access to the sacral epidural space.

Equipment

- 2 ml, 5 ml, and 20 ml syringes
- 18 G, 20 G, and 25 G needles
- 22 G, 3–5 cm short-bevel needle, with stylet
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy or ultrasound (optional)

Drugs

- Lidocaine (lignocaine) 1%, 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 50 mg (or its equivalent)
- Saline (NaCl) 10 ml
- Resuscitation drugs (*see Appendix 3*)

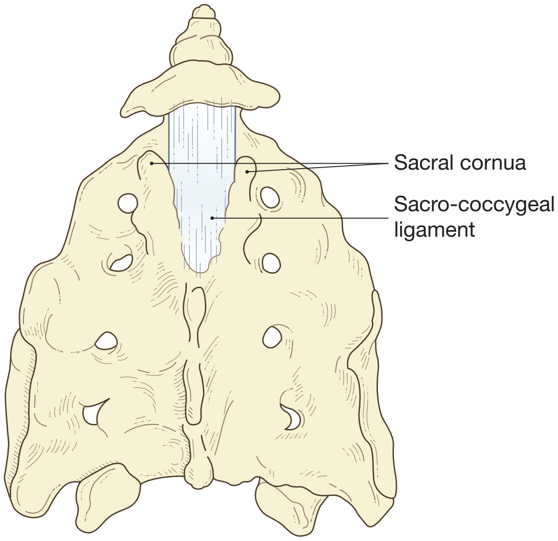

Fig. 3.4.1

Position of patient

- Prone.
- Pillow under abdomen and/or operating table broken to allow flexion of the lumbo-sacral spine.
- Lower limbs abducted 15°, toes rotated to point towards the opposite foot (*Fig. 3.4.2*).

Needle puncture and technique (adult)

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.

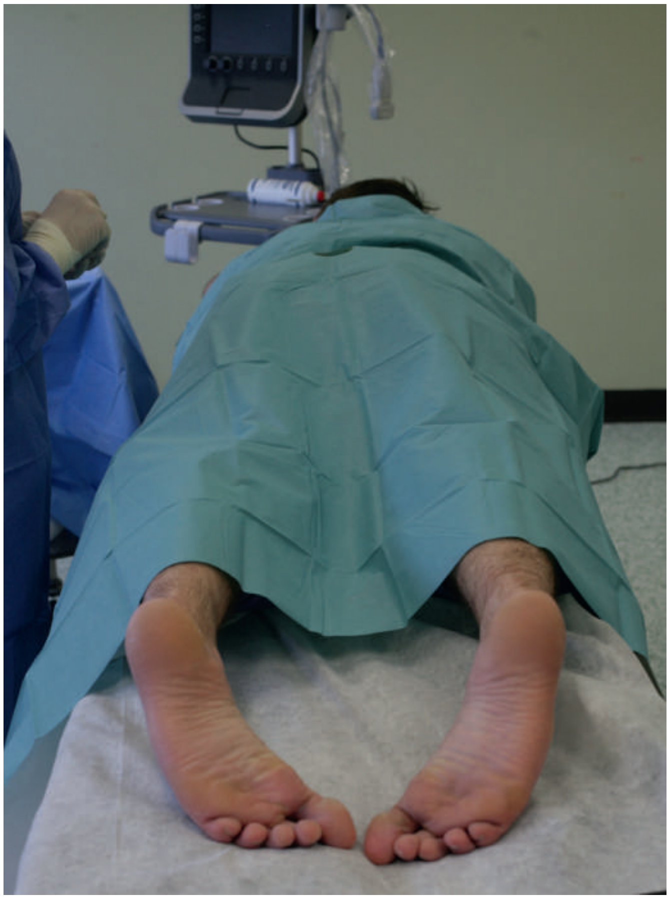

Fig. 3.4.2

- Lidocaine (lignocaine) 1% 2 ml is drawn up into three 2 ml syringes.
- Lidocaine (lignocaine) 1% 15 ml plus corticosteroid, e.g. triamcinolone diacetate 50 mg, is drawn up into the 20 ml syringe.
- NaCl 10 ml is drawn up into the 10 ml syringe.

FOR THE RIGHT-HANDED OPERATOR

With the left hand

- The posterior superior iliac spines are identified.
- The sacral cornua (the unfused spinous processes of S5) are also identified and marked.
- Between the cornua lies the base of the sacral hiatus, a roughly triangular fibroelastic structure.
- The index and middle fingers of the left hand are placed on each of the sacral cornua (*Fig. 3.4.3*).
- **The insertion point lies between these two fingers.**

With the right hand

- The insertion point is infiltrated with lidocaine (lignocaine) 1% 2 ml.
- The 22 G short-bevel needle with stylet is inserted at an angle of 45° to the skin (*Fig. 3.4.4. a,b*). Ultrasound can be used to guide needle placement (*Fig. 3.4.4c*).
- As the needle passes through the fibroelastic sacral hiatus, a “pop” may be experienced, although this is not always evident in adults, and bone may be contacted.
- After passing through the sacral hiatus, the needle is withdrawn a little, and redirected to an angle to the skin of 15–20° (*Fig. 3.4.5*). This should allow further advancement of 1–2 cm, as the needle enters the long axis of the caudal epidural space.
- After negative aspiration, lidocaine (lignocaine) 1% 2 ml is injected.
- After 5 minutes the patient is questioned about changes in sensation or power of the lower limbs, and any changes in heart rate or blood pressure are noted.
- Then lidocaine (lignocaine) 1% 5–15 ml plus corticosteroid, e.g. triamcinolone diacetate 50 mg, may be injected in order to promote spread to upper sacral and lower lumbar segments (a volume of at least 10 ml should be used if the symptoms are at the level of S1 nerve root or higher) (*Fig. 3.4.6*). The needle is then

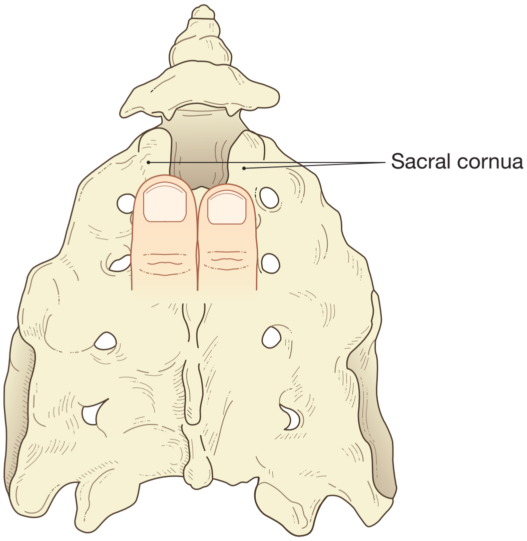

Fig. 3.4.3

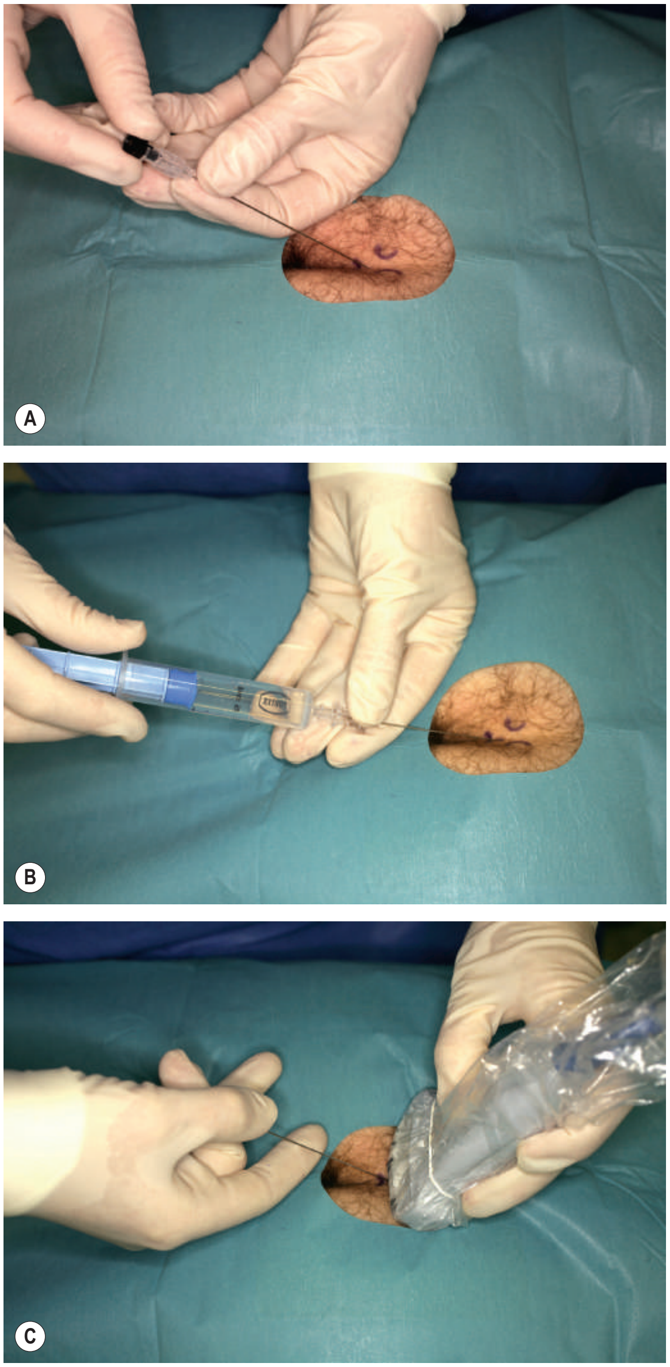

Fig. 3.4.4

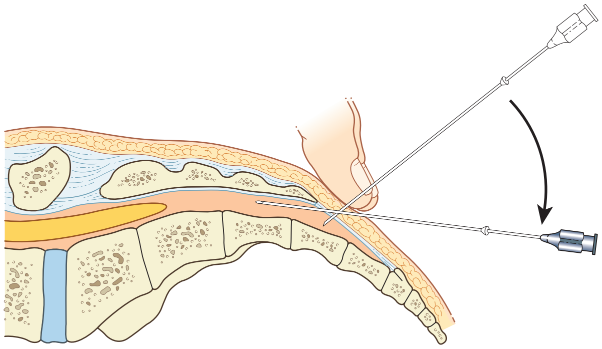

Fig. 3.4.5

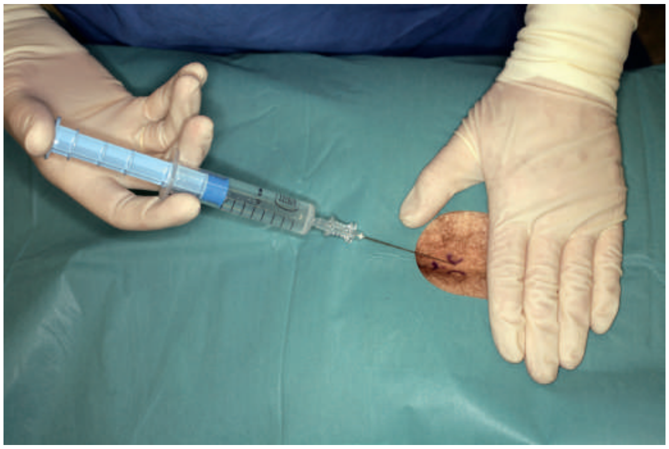

Fig. 3.4.6

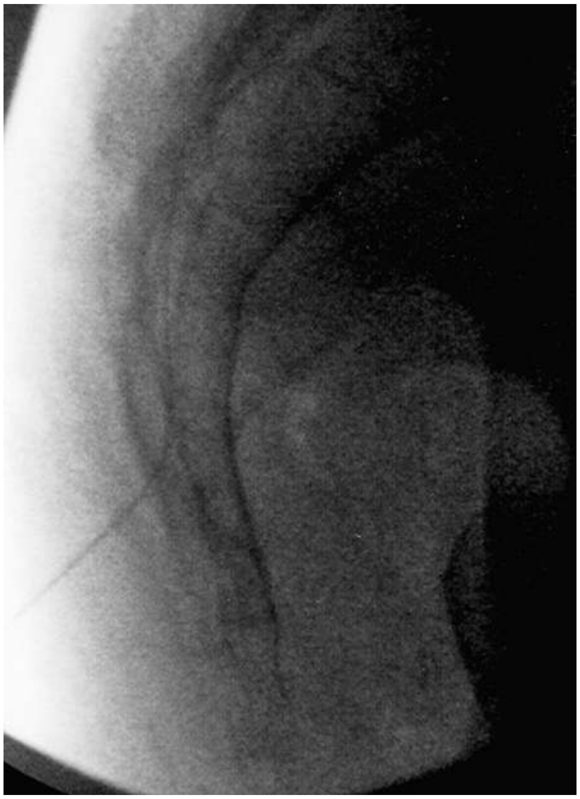

Fig. 3.4.7

removed while clearing it with lidocaine (lignocaine) 1% 2 ml.

- Monitors should be left attached and i.v. access kept in situ for at least 30 minutes.
- It is prudent to warn the patient about possible loss of sensation and or power of one or both lower limbs.

FLUOROSCOPIC GUIDED CAUDAL EPIDURAL INJECTION

- This approach may be used for treating L5 or S1 radiculopathy. It is a reasonable alternative to the interlaminar approach when surgery has disrupted the posterior spinal anatomy.
- Position the patient prone.
- Locate the sacral hiatus using the sacral cornua as landmarks.
- Prepare skin with antiseptic and sterile drape.
- Place the tip of a sterile blunt instrument over the sacral hiatus and obtain a lateral fluoroscopic view of the sacrum.
- Raise a skin wheal with 1% lidocaine (lignocaine) just below the sacral hiatus and infiltrate with lidocaine down to the sacral hiatus with a small gauge needle.
- Advance a Tuohy needle at a 45° angle to the skin through the sacral hiatus, checking a lateral fluoroscopic view to make sure the needle has entered the spinal canal. Lower the needle angle and advance the needle slightly. Recheck lateral fluoroscopy to ensure the needle is in the spinal canal.
- Aspirate to rule out intravascular placement and inject 0.5–1 ml contrast medium under live fluoroscopy. Check a lateral and AP image to ensure epidural spread.
- If no epidural catheter is used, inject a mixture of local anesthetic and steroid. Inject 10 ml 0.5% lidocaine plus 50 mg triamcinolone diacetate. This volume should be sufficient to reach the L5 or S1 nerve roots.
- Alternatively, a radio-opaque catheter can be inserted through the needle and advanced to the desired level. Check the catheter position in both AP and lateral views (*Fig. 3.4.7*). Attach the injection hub to the catheter and inject 0.5–1 ml contrast medium under live fluoroscopy, rechecking dye spread in both AP and lateral views (*Fig. 3.4.8*). Inject 1–2 ml 1% lidocaine plus 50 mg triamcinolone diacetate or equivalent.

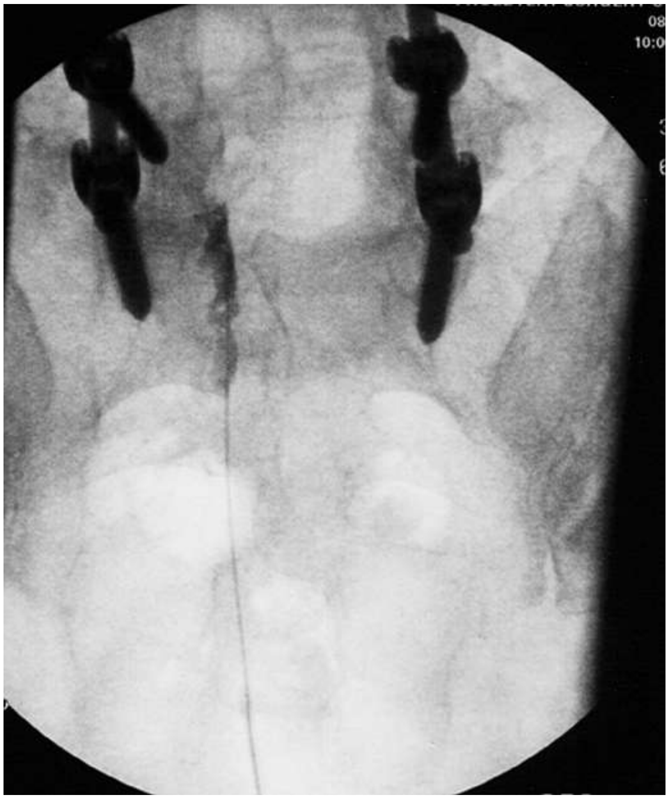

Fig. 3.4.8

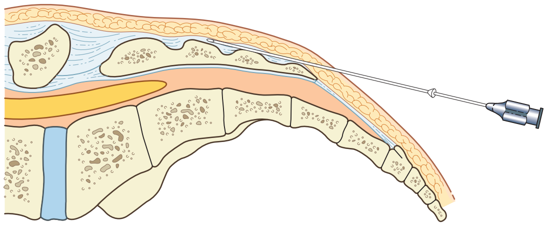

Fig. 3.4.9

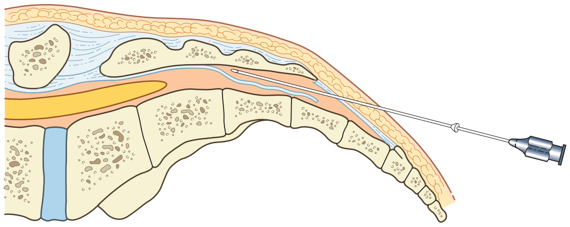

Fig. 3.4.10

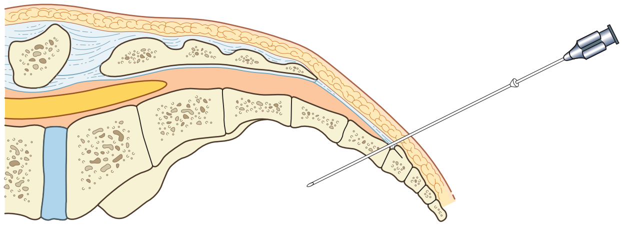

Fig. 3.4.11

Confirmation of a successful block

- Relief of pain.
- Anesthesia or diminished sensation in distribution of blocked nerves.
- Improvement in straight-leg raising (although for sacral nerve root-related pain, this may not have been positive prior to caudal blockade).

Tips

- After location of the caudal epidural space, the left hand may be placed over the sacral hiatus while 10 ml saline is rapidly injected (*Fig. 3.4.6*). Misplacement of the needle in the subcutaneous tissue (*Fig. 3.4.9*) should be evident if the injection is appreciated by the palpating left hand.
- Subperiosteal injection in the awake patient will cause pain (*Fig. 3.4.10*). The needle angle is important as the tip may come to lie anterior to the sacrum (*Fig. 3.4.11*).
- Anatomic variations exist in many patients, making access to the caudal epidural space difficult or impossible. Fluoroscopy using lateral views is useful to confirm the epidural needle position. Ultrasound may also be helpful in identifying the sacro-coccygeal membrane.

Potential problems

IMMEDIATE

- Failure to locate epidural space.
- Intravascular injection (test dose important); addition of epinephrine (adrenaline) to test dose may help identification of intravascular injection.
- Intrathecal injection, rare but possible (test dose important).
- Hypotension due to sympathetic blockade (give i.v. fluid, consider ephedrine).
- Transient exacerbation of radiculopathic pain (caution patient).
- Headache (possible dural puncture, rare).
- Allergic reaction.

LATER

- Infection (epidural abscess, bacterial meningitis).
- Aseptic meningitis, usually the result of intrathecal injection (test dose important).
- Cushingoid symptoms (usually as a result of repeated injections).
- Infection (epidural abscess).
- Epidural hematoma.

3.5 LONG-TERM EPIDURAL CATHETER INSERTION

Anatomy

As described for lumbar epidural block in Section 3.1.

Equipment

- 2 ml and 10 ml syringes
- 18 G, 20 G, and 25 G needles
- ECG, BP, and SpO_2 monitors
- 18 G epidural set
- Epidural catheter passer
- A surgical pack, including small scalpel and suture set
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy or ultrasound (optional)

Drugs

- Lidocaine (lignocaine) 1% (preservative free) 20 ml (or its equivalent)
- Lidocaine (lignocaine) 1% (preservative free) 4 ml plus epinephrine (adrenaline) 1:200 000
- Saline (NaCl) 10 ml

Position of patient

- Lateral, lying on side of radiculopathy (*Fig. 3.5.1*).
- Shoulders and buttocks parallel to the edge of the bed, perpendicular to the floor, with spine flexed.

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- Lidocaine (lignocaine) 1%, 2 ml is drawn up into three 2 ml syringes.
- Lidocaine (lignocaine) 1% 10 ml is drawn up into one 10 ml syringe.
- NaCl 10 ml is drawn up into the 10 ml loss-of-resistance syringe.
- An epidural catheter is inserted through the needle as previously described, 5–6 cm into epidural space (see Section 3.1).
- The needle is withdrawn 1–1.5 cm, but *is not removed.*
- After negative aspiration, a test dose of lidocaine (lignocaine) 1% 4 ml, with epinephrine (adrenaline) 1:200 000, is given. After 5 minutes the patient is questioned about changes in sensation or power, and any changes in heart rate or blood pressure are noted.
- After a negative reaction lidocaine (lignocaine) 1% 10 ml is injected slowly over 10 minutes. Assessment of the level of blockade is carried out after a further 15 minutes.
- Subcutaneous infiltration around the epidural needle with lidocaine (lignocaine) 1% is carried out. Note: a small incision is made to include the epidural needle (*Fig. 3.5.2*).
- A purse string suture is placed around the epidural needle, *but is not tied* (*Fig. 3.5.3*).
- Another small incision in the lateral abdominal wall is made after subcutaneous infiltration with lidocaine (lignocaine) 1% (*Fig. 3.5.4*). A catheter passer is tunneled through the subcutaneous tissue between the two incision sites.
- A catheter is manually bent to a curve and tunneled through the subcutaneous tissue between the two incision sites in the direction from the abdominal site to the epidural needle site (*Fig. 3.5.5*).
- The epidural needle is carefully removed (*Figs 3.5.6, 3.5.7*).
- The catheter is secured to subcutaneous tissue in the midline by tightening the purse string suture (*Fig. 3.5.8*).
- The catheter is further secured using an angle piece and then threaded in a lateral direction through the catheter

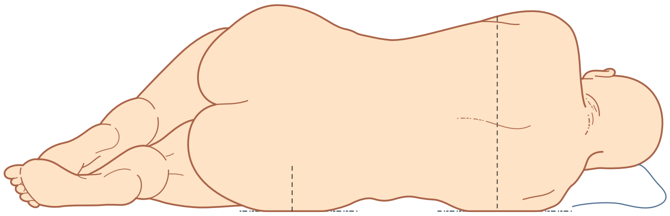

Fig. 3.5.1

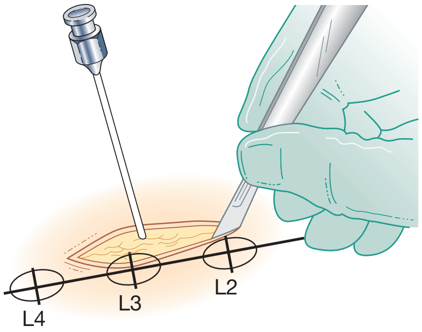

Fig. 3.5.2

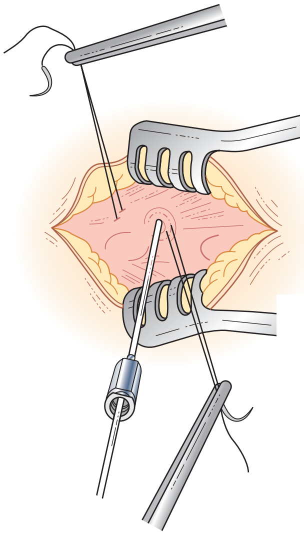

Fig. 3.5.3

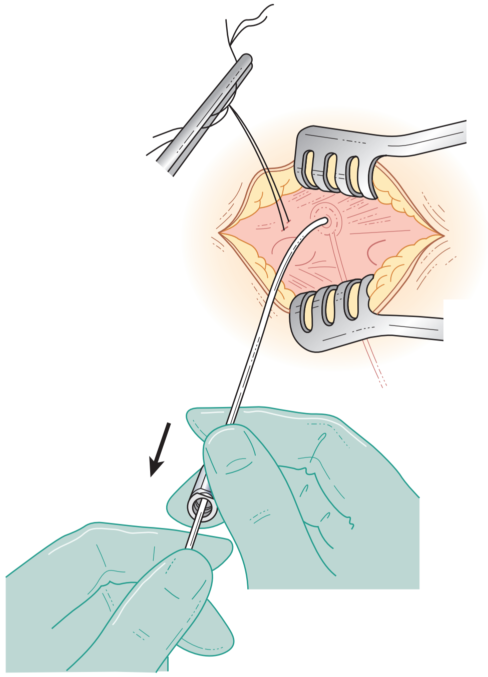

Fig. 3.5.6

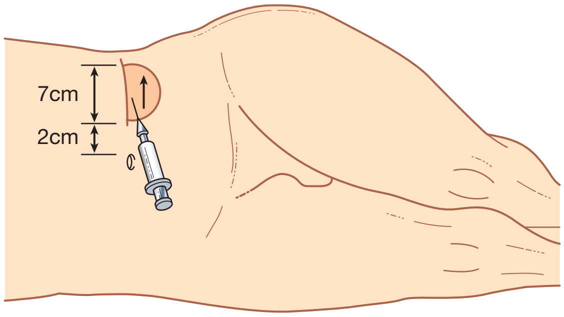

Fig. 3.5.4

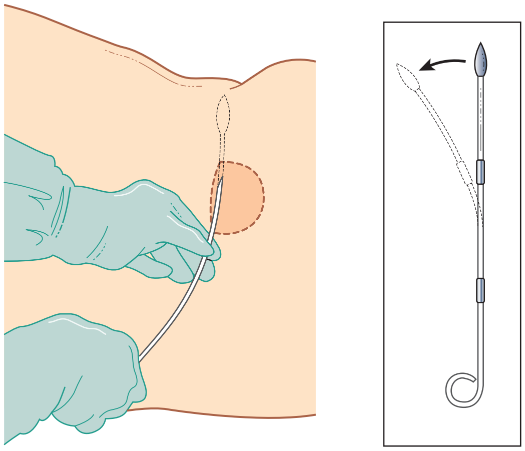

Fig. 3.5.5

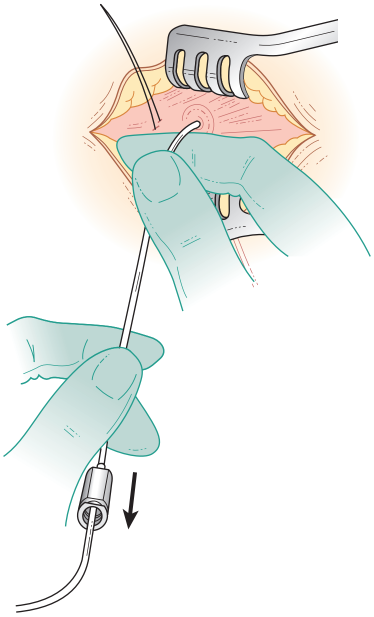

Fig. 3.5.7

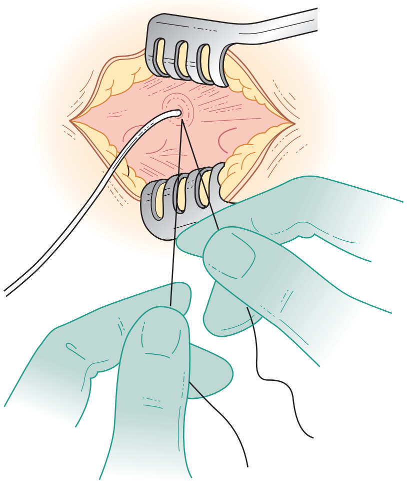

Fig. 3.5.8

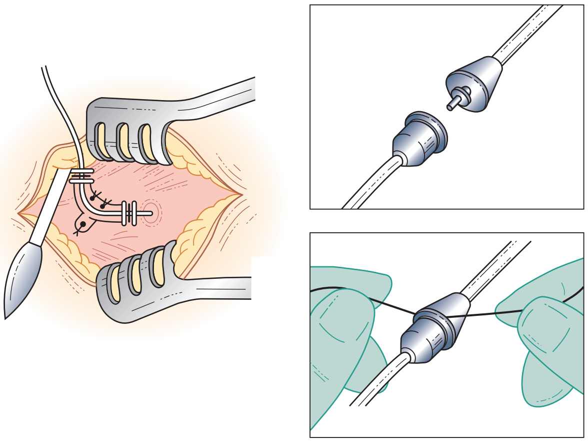

Fig. 3.5.9

passer from the needle site to the abdominal wall site and connections are secured (*Fig. 3.5.9*).

- A pump may be placed in the abdominal wall site and the skin incisions closed.

Confirmation of a successful block

- Relief of pain.
- Anesthesia or diminished sensation in the distribution of affected nerves.

Tips

- Injection of radio-opaque dye under direct fluoroscopy can confirm epidural placement.
- Insertion of a radio-opaque epidural catheter may be carried out also under fluoroscopy.
- Ultrasound may aid insertion

Potential problems

- As described for lumbar epidural block in Section 3.1.
- However, in view of the long-term nature of epidural catheter implantation, any symptoms of infection should be immediately investigated and treated.
- Epidural catheters should not be inserted or removed during anticoagulation. Coagulation and platelet function should be normalized before catheter removal.

SOMATIC NERVE BLOCKADE

4

Mechanical nerve root compression was originally thought to be the cause of pain in discogenic radiculopathy. However, it has been found that many asymptomatic patients demonstrate substantial disc protrusion on magnetic resonance (MR) imaging, myelography and subsequent autopsy examination. In addition, surgical decompression does not result in uniform success in the relief of such pain. Following a period of mechanical nerve-root compression it is likely that an acute inflammatory process ensues, resulting in intraneural accumulation of serum proteins and fluid, raised intraneural pressure, ischemia and axonal degeneration. Degenerating glycoprotein material from the nucleus pulposis may also contribute to the inflammatory process.

There are many situations in which injection of spinal nerve roots with local anesthetic may be helpful in the diagnosis of radicular pain. These include those where investigations including electromyography, computer tomography (CT) or MR imaging are not consistent with the clinical findings, where there are multiple levels of pathology, and after spinal surgery with subsequent scarring in the area of the surgery. In addition, the contribution of the somatic nerve root may be elucidated in pain of uncertain origin, e.g. chest pain or abdominal pain, by specific nerve root local anesthetic injection.

It may be used therefore to determine the level of surgery, if indicated, and the addition of steroid may produce longer-lasting pain relief.

4.1 INTERCOSTAL NERVE BLOCK

Anatomy

The intercostal nerve is made up of several types of nerves: sympathetic white and grey rami communicantes, cutaneous and motor fibers supplied by dorsal rami, sensory fibers to the chest wall, anterior and posterior, via the lateral cutaneous branch, and further sensory fibers to the anterior chest wall via the anterior cutaneous branch. The lateral cutaneous branch exits just distal to the angle of the rib. Just below the inferior edge of the rib, in the intercostal groove, lie the intercostal nerve, artery and vein, the latter lying superior to the nerve. The optimal site to block the intercostal nerve is the most posterior point at which the rib is palpable, usually the angle of the rib (**Fig. 4.1.1 a,b**).

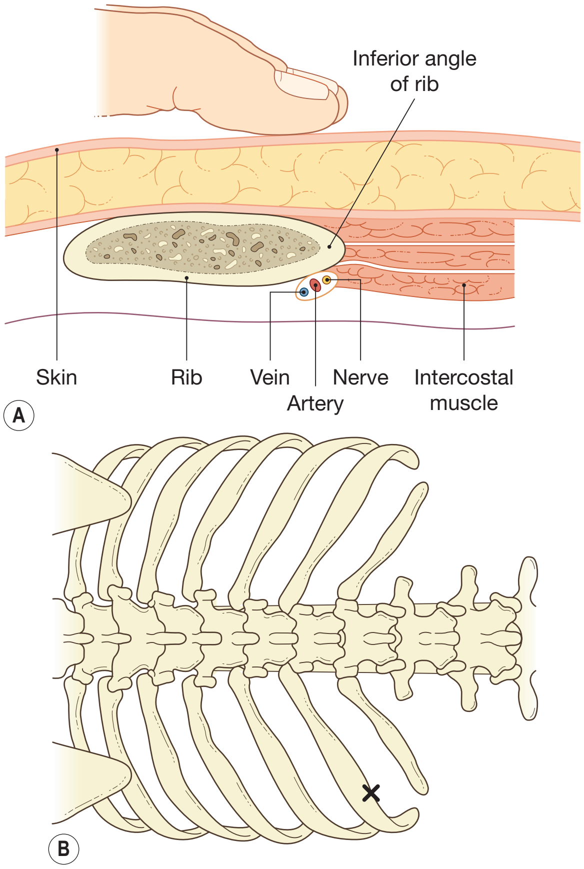

Fig. 4.1.1

Equipment

- 2 ml and 5 ml syringes
- 30 G needle
- 22 G 3–4 cm short-bevel needle
- Extension set (optional)
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- Ultrasound (optional)

Drugs

- Lidocaine (lignocaine) 1% 2 ml for skin infiltration
- Lidocaine (lignocaine) 1% 5 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 50 mg (or its equivalent)
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone (this allows best access, although a lateral or supine position may also be used).
- Pillow under mid-abdomen to widen the intercostal spaces.
- Arms hanging over sides of table to rotate the scapulae laterally.

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- Sedation may be administered if multiple blocks are being performed.
- The midline and an area 10 cm × 10 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- The midline is palpated and marked.
- **The inferior edge of the rib is palpated and marked at the most posterior point at which the rib is palpable; this is the insertion point** (*Fig. 4.1.2*). If multiple blocks are planned these marks will form a line that becomes more medial towards the cephalad end as the scapulae are avoided laterally (*Fig. 4.1.2*).

FOR THE RIGHT-HANDED OPERATOR

With the left hand

- The inferior edge of the rib is palpated with the fore- and middle fingers (*Fig. 4.1.3*).

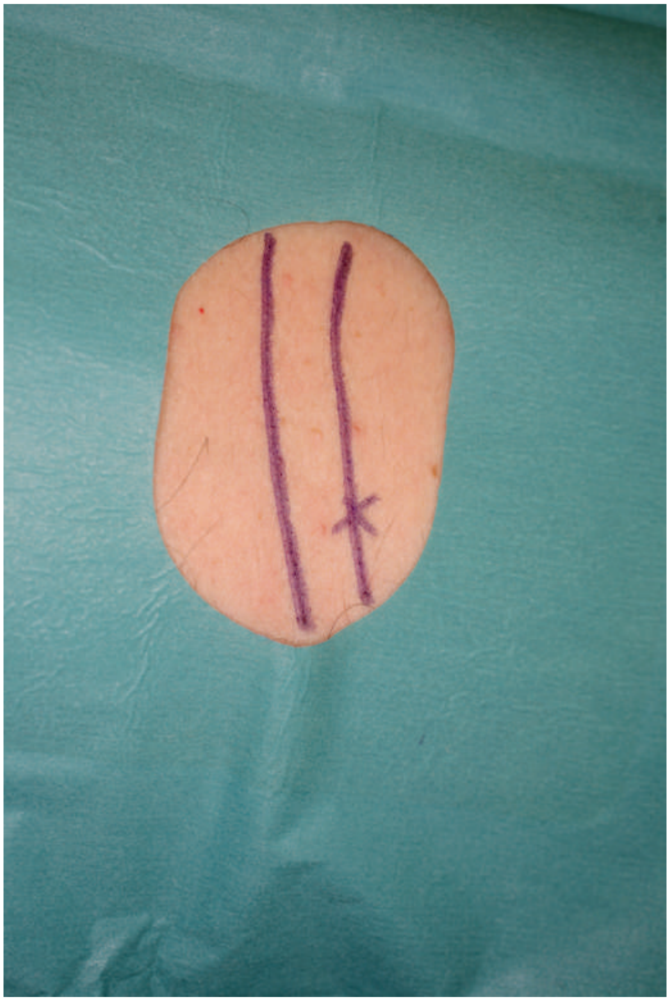

Fig. 4.1.2

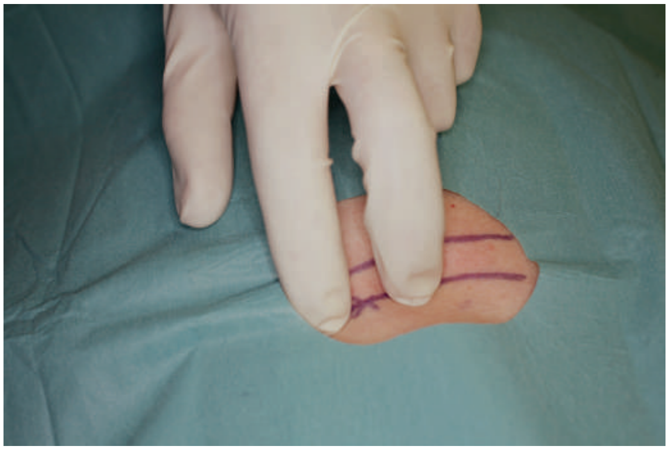

Fig. 4.1.3

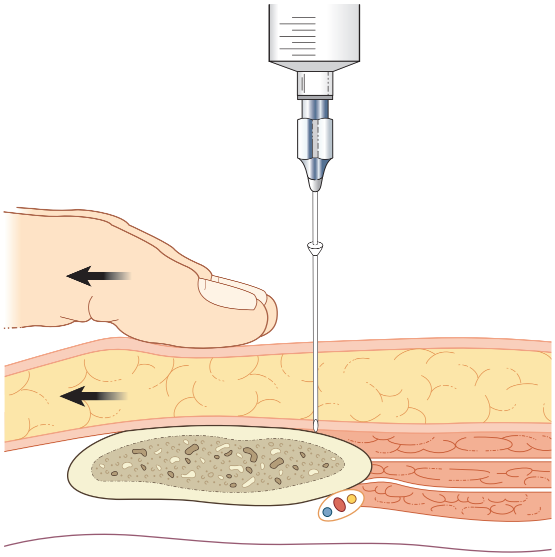

Fig. 4.1.4

- The skin is drawn up over the rib itself.
- The fingers of the left hand will grip the needle-hub for controlled advancement of the needle once contact with the rib is made during injection.

With the right hand

- The insertion point is infiltrated with lidocaine (lignocaine) 1% using a 2 ml syringe and a 30 G needle.
- The 22 G short-bevel needle with syringe attached is inserted between fore- and middle finger of the left hand in a direction 15–20° cephalad, until it makes contact with the rib (*Figs 4.1.4, 4.1.5*).
- The needle-hub is gripped with the fingers of the left hand and this hand is steadied by leaning the wrist against the patient's posterior chest wall.
- With the right and left hands acting as one unit, the needle is walked off the edge of the rib until it enters the intercostal space immediately below the rib (*Fig 4.1.6*). Alternatively, a catheter may be inserted between the needle and syringe and injection may then be carried out by a second operator, while the first maintains the needle steady in the correct position (*Fig. 4.1.7*).
- It is then advanced 2 mm.
- After negative aspiration, 3–4 ml of local anesthetic, plus corticosteroid if indicated, is injected and the needle is withdrawn.
- Monitors are left attached and i.v. access left in situ for at least 30 minutes.
- Chest X-ray is performed if pneumothorax is suspected.

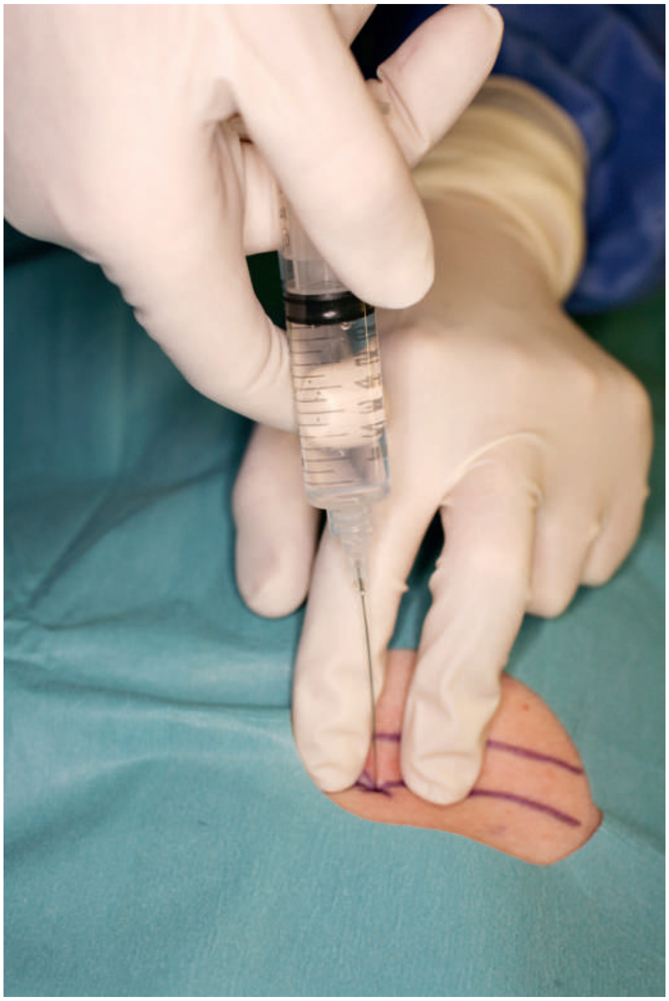

Fig. 4.1.5

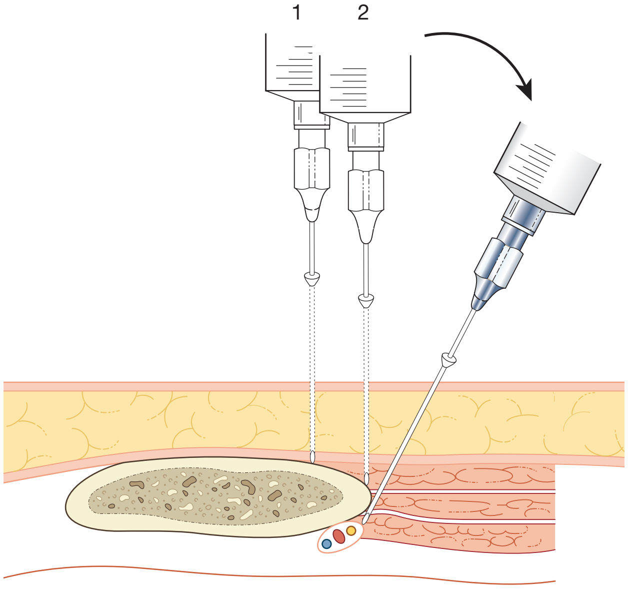

Fig. 4.1.6

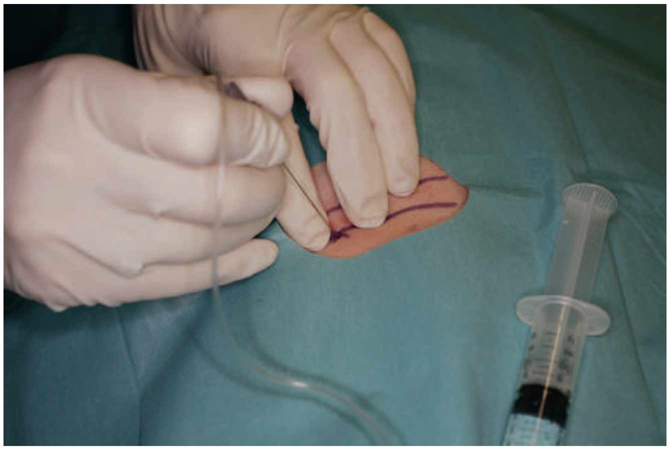

Fig. 4.1.7

Confirmation of a successful block

- Relief of pain.
- Anesthesia in the distribution of the blocked nerve.

Tips

- The approximate depth to the rib may be determined with the left fore- and middle fingers before insertion of the needle.
- Insertion of the needle > 2 mm deeper than the rib when intercostal space is reached is avoided. This will minimize the risk of pneumothorax, as the average distance from the rib to the pleura is 8 mm. If patient coughs on injection, pneumothorax is suspected.
- Neurolytic intercostal nerve block, e.g. with alcohol 50% 3 ml (made up by combining equal parts of alcohol 100% and lidocaine (lignocaine) 1% or its equivalent), or phenol 6%, may be carried out after local anesthetic block confirms accurate placement of the needle as described above. However, it is important to note that injection may result in subarachnoid spread of a neurolytic agent with resultant possible permanent spinal cord damage.
- Ultrasound may aid accurate placement of the needle (*Fig. 4.1.8*). Injection of non-ionic radio-constrast medium may also aid accurate placement of the needle in neurolytic block (*Fig. 4.1.9*).
- Placement of a catheter into the intercostal space can be achieved by threading 3 cm of catheter through an 18 G epidural needle after the intercostal neurovascular bundle has been identified as above.

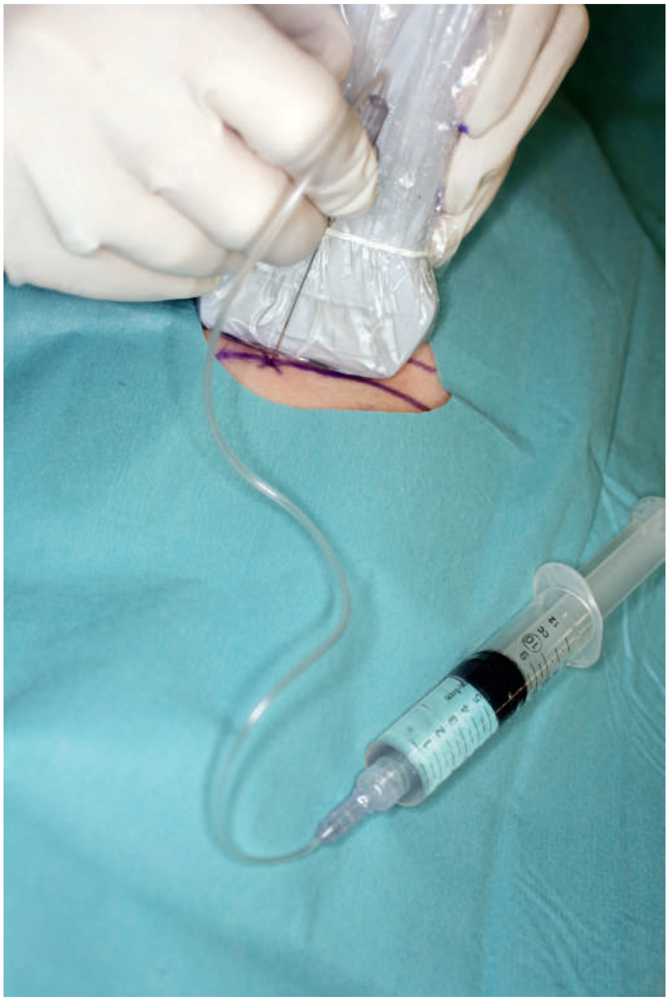
Fig. 4.1.8

Radiofrequency lesioning

- Radiofrequency lesioning of the intercostal nerve is simple and has a low level of side effects. The lesioning is carried out using the same method of placement of the needle as described for intercostal nerve block with local anesthetic.
- However, after placement of the needle and confirmation of accuracy by fluoroscopy (as described above), a trial of stimulation is carried out using 2 V at 50 Hz. If the needle has been placed accurately the patient should experience paresthesiae in the distribution of that intercostal nerve. A pulsed radiofrequency lesion may then be produced at 40–45 °C for 5 minutes or 49–60 °C for 90 seconds.

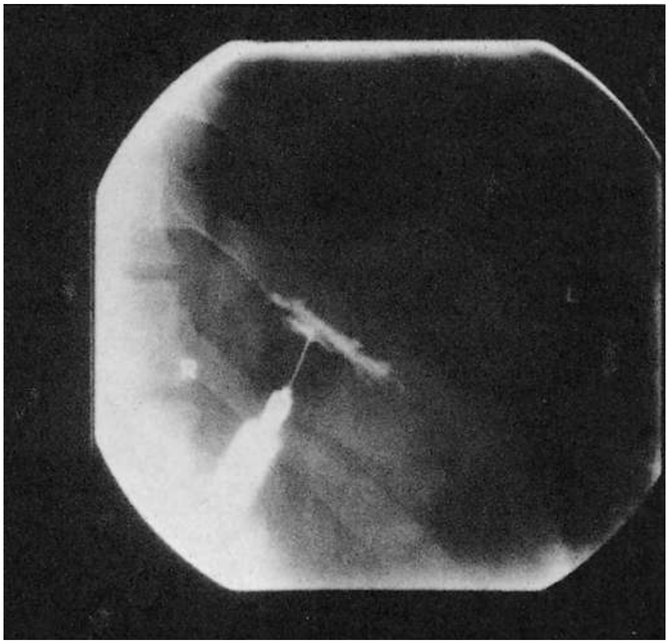
Fig. 4.1.9

Potential problems

- Injection within the nerve sheath can result in the spread of anesthetic to the subarachnoid space.
- Intercostal block in patients with severe respiratory problems should be avoided as there is risk of a pneumothorax. Careful observation of a small pneumothorax is usually all that is required but failure to re-expand the lung may require chest tube insertion.
- Because of the vascularity of the intercostal space, there may be rapid absorption of local anesthetic and systemic effects can occur quickly, especially with multiple blocks. However, peak plasma concentration of local anesthetic may occur 15–20 minutes after the block is performed, when systemic toxicity effects may develop. Addition of epinephrine (adrenaline) to the anesthetic solution may decrease the peak plasma concentration of local anesthetic.
- If aspiration of blood occurs, the needle should be removed, keeping the left fore- and middle fingers in place. The needle is cleared, reinserted to contact the rib again, and the block is continued as above.

- Care of the airway must be remembered if sedation is administered to a patient in the prone position.
- Complications of neurolytic intercostal nerve block include pneumothorax, infection (especially in the immunocompromised patient) as well as post-lesioning intercostal nerve neuritis. The frequency of the latter in radiofrequency lesioning increases as higher temperatures are used. Intercostal nerve neuritis usually responds to local injection of lidocaine (lignocaine) 1% 3 ml plus triamcinolone diacetate 20 mg to the lesion site.

4.2 INTERPLEURAL BLOCK

Anatomy

The parietal pleura lines the thoracic wall, the thoracic surface of the diaphragm and the lateral mediastinum. The visceral pleura completely covers the surface of the lung. Both the pleural layers become contiguous at the root of the lung. Between the two pleural layers lies the interpleural space (**Fig. 4.2.1**). Injection of local anesthetic into this space produces an interpleural block by topical contact with free nerve endings within the pleura, and by local diffusion to nerves in the vicinity of the injection site. These include the intercostal nerves, the sympathetic chain, and the inferior part of the brachial plexus. Local anesthetic solution may also track to the epidural and subarachnoid spaces producing blockade.

Equipment

- 2 ml and 10 ml syringes
- 18 G, 20 G, and 25 G needles
- ECG, BP, and SpO_2 monitors
- 18 G epidural set
- Well-lubricated 5 ml glass syringe
- Resuscitation equipment (*see Appendix 3*)

Drugs

- Lidocaine (lignocaine) 1% 10 ml
- Levobupivacaine 0.25%, 20 ml (or its equivalent)
- Saline (NaCl) 10 ml
- Resuscitation drugs (*see Appendix 3*)

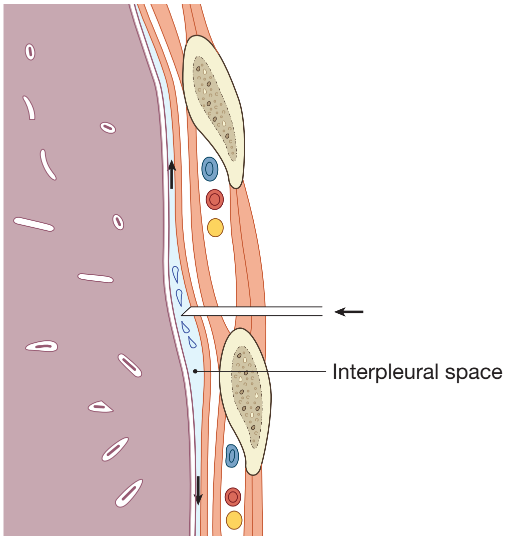

Fig. 4.2.1

Position of patient

- The technique described here relies on negative interpleural pressure to identify the interpleural space. Therefore the patient should be breathing spontaneously for this technique.
- Semi-prone.
- Side to be blocked uppermost, supported by a pillow (*Fig. 4.2.2*).
- The arm should be allowed to fall forwards in front of the body to rotate the scapula anterolaterally.

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The midline and an area 15 cm × 12 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- Lidocaine (lignocaine) 2%, 2 ml is drawn up.
- Levobupivacaine 0.25% 20 ml is drawn up.
- Air is drawn up into a well-lubricated 5 ml syringe.
- The seventh and eighth ribs are palpated and marked.

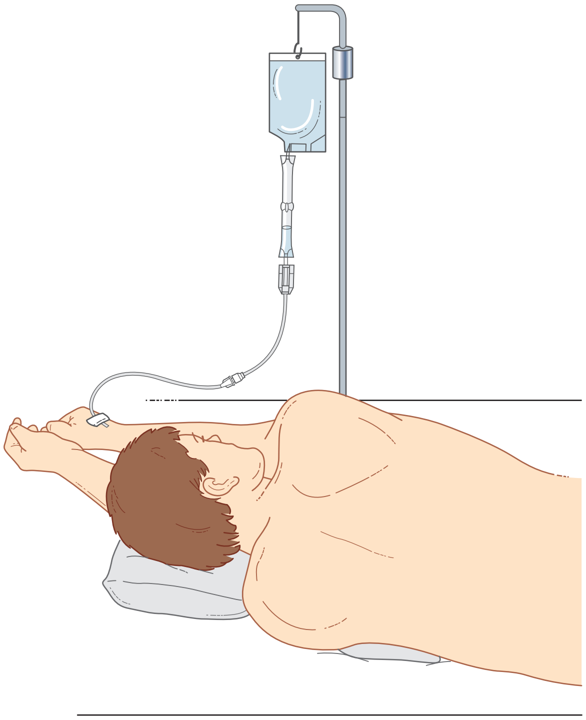

Fig. 4.2.2

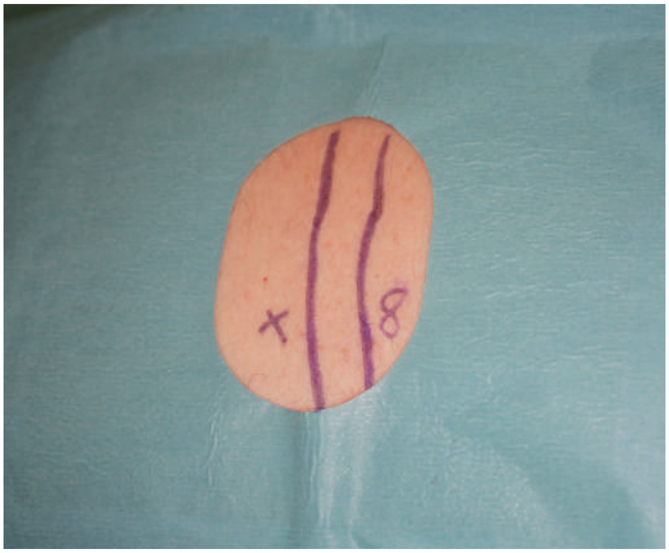

Fig. 4.2.3

- **A point approximately 10 cm from the midline, immediately superior to the eighth rib, is marked; this is the insertion point** (*Fig. 4.2.3*).

FOR THE RIGHT-HANDED OPERATOR

With the left hand

- The fore- and middle fingers are placed each side of the insertion point, palpating the superior aspect of the eighth rib.
- These fingers are kept in place until the epidural needle passes through the subcutaneous tissue.

With the right hand

- A skin wheal is raised at the insertion point.
- The epidural needle is inserted between the fore- and middle fingers of the left hand, taking care that the point of entry of the needle is as close as possible to the superior aspect of the eighth rib. This helps to avoid damage to the neurovascular bundle, which lies immediately inferior to the seventh rib.
- After passage through the subcutaneous tissue, the hub of the needle is gripped with the fore- and middle fingers of the left hand and this hand is steadied by leaning the wrist against the patient's posterior chest wall.
- The stylet is removed and the well-lubricated glass syringe containing 3 ml air is applied (*Fig. 4.2.4*).
- The needle is slowly and carefully advanced, with no pressure applied to the plunger, the left hand aiding the advance, while at the same time applying a brake if required.
- Resistance from the tissues prevents the plunger from advancing.
- At the point at which the needle enters the interpleural space a definite click is experienced, negative pressure draws the air in and the barrel drops (*Fig. 4.2.5*).

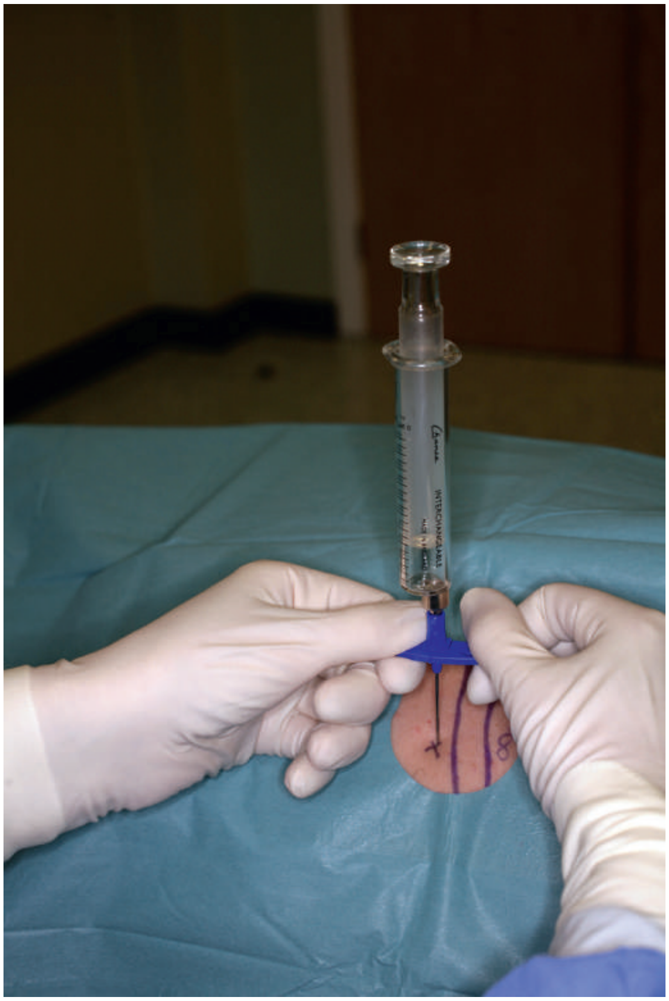

Fig. 4.2.4

- Taking care not to allow air entry into the interpleural cavity, a catheter is inserted approximately 10 cm into the interpleural space (*Fig. 4.2.6*). **Once the catheter is in position it is best to place the patient supine, tilted slightly, with the side to be blocked upwards** (*Fig. 4.2.7 a,b*). **(If blockade of the upper thoracic segments is required the patient is tilted head-down).**
- After negative aspiration and a test dose, 10–15 ml levobupivacaine 0.25% (or its equivalent) in divided doses of 5 ml is injected.
- Infusion of local anesthetic may be set up for continuous analgesia.
- Monitors should be left attached and i.v. access left in situ while the catheter is in place.
- Chest radiograph may be performed to rule out pneumothorax.

Confirmation of a successful block

- Relief of pain.
- Anesthesia in the distribution of the blocked nerves.

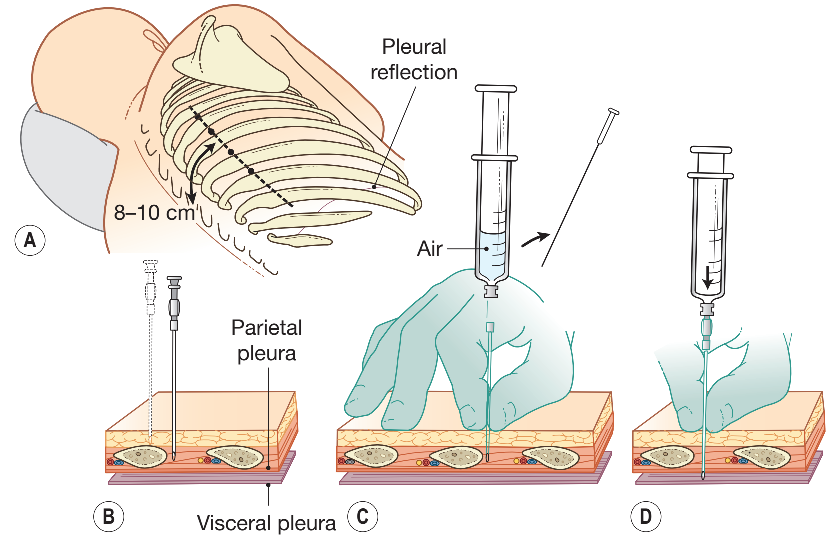
Pleural
reflection
8–10 cm
A
Air
Parietal
pleura
B
Visceral pleura
C
D

Fig. 4.2.5

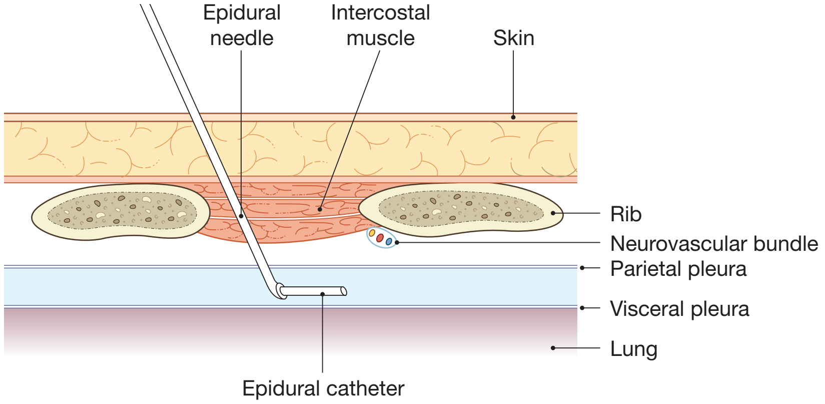
Epidural
needle
Intercostal
muscle
Skin
Rib
Neurovascular bundle
Parietal pleura
Visceral pleura
Lung
Epidural catheter

Fig. 4.2.6

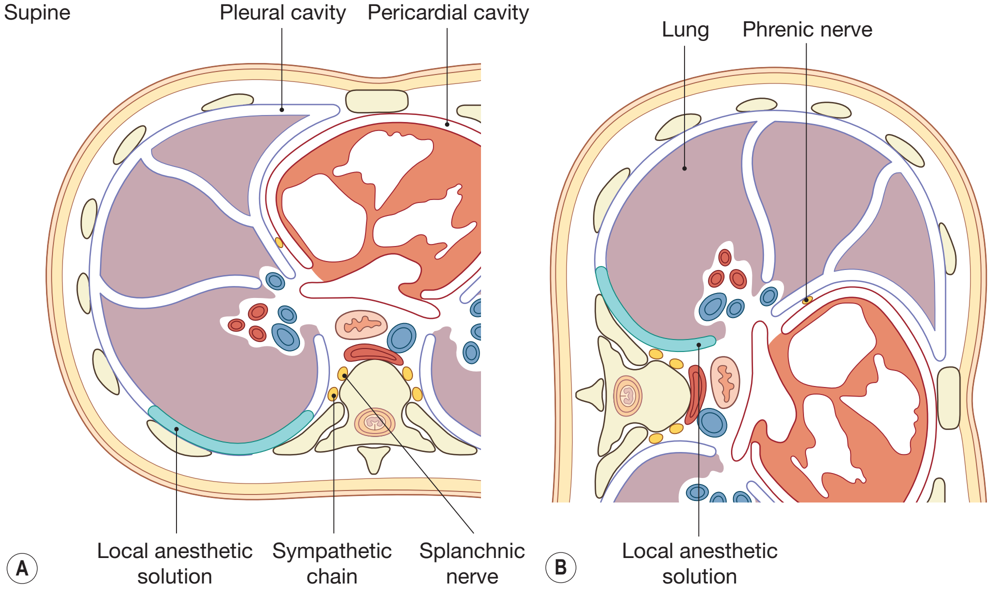
Supine
Pleural cavity
Pericardial cavity
Lung
Phrenic nerve
A
Local anesthetic
solution
Sympathetic
chain
Splanchnic
nerve
B
Local anesthetic
solution

Fig. 4.2.7

Tips

- As an alternative technique, the barrel of the syringe may be removed and the open syringe filled with NaCl or local anesthetic and advanced until the fluid level begins to fall as the solution is sucked into the interpleural space (*Fig. 4.2.8*).
- Also, a bag containing NaCl 500 ml may be attached to the epidural needle via a giving set, and drops observed on entry to the interpleural space (*Fig. 4.2.9*).
- Ultrasound may aid placement of the needle (*Fig. 4.2.10*).

Potential problems

- Pneumothorax.
- Unpredictable analgesia. The mechanism of action of interpleural block is still unproven and spread of the local anesthetic solution may be unpredictable. The duration of the block may be decreased when a thoracotomy drainage tube is present.

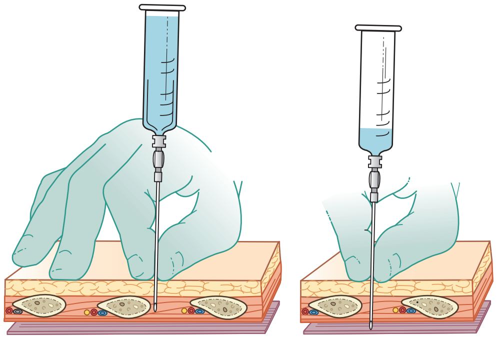

Fig. 4.2.8

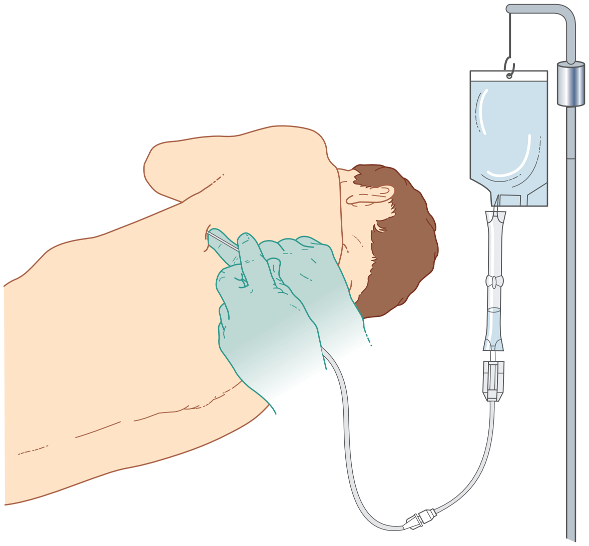

Fig. 4.2.9

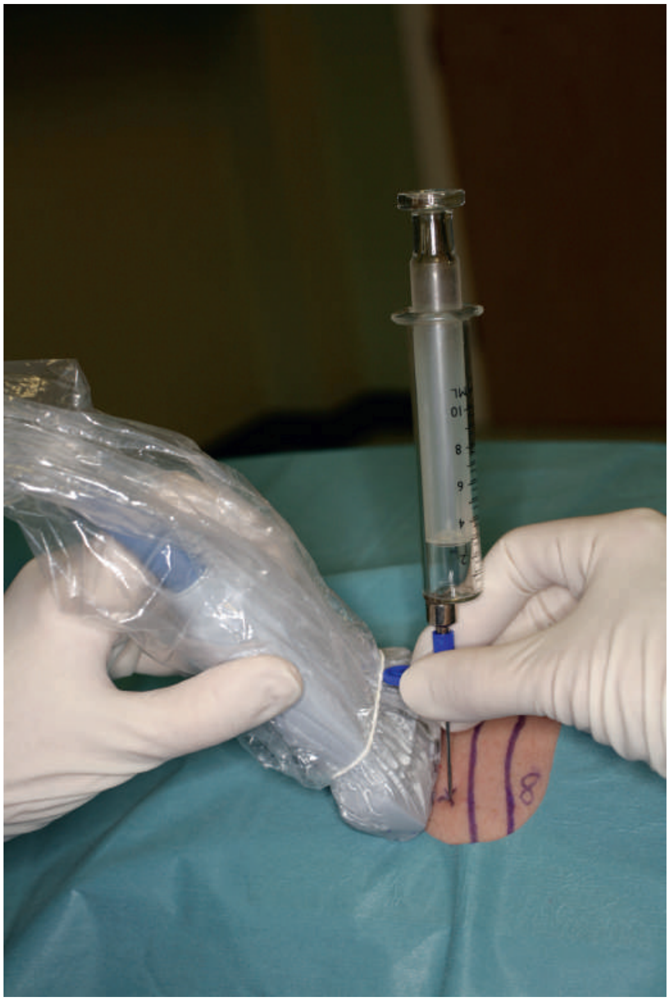

Fig. 4.2.10

4.3 LUMBAR NERVE ROOT BLOCK

Anatomy

The lumbar nerves are made up of sensory and motor fibers to the trunk and lower limbs, and sympathetic white and grey rami communicantes. Each lumbar nerve exits via the intervertebral foramen which lies just inferior to the caudad edge of the transverse process of the respective vertebral body, and passes anteriorly over the lateral aspect of the transverse process of the vertebral body below (**Fig. 4.3.1**). It then branches into posterior and anterior branches. The posterior branch supplies the paravertebral muscles and cutaneous fibers to the back. The anterior branch passes through the substance of the psoas muscle and branches further, communicating with the other anterior branches to form the lumbar plexus. As a result there is significant overlap of nerve supply. The fascial layers of the psoas muscle prevent spread of local anesthetic to the sympathetic lumbar chain. It may be helpful to consider blockade of a lumbar nerve root to be similar to intercostal nerve block except that the transverse process is present instead of a rib and the insertion site is therefore more medial.

Equipment

- 2 ml and 10 ml syringes
- 25 G needle
- 22 G spinal needle, end-opening
- Radio-opaque contrast medium
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy or ultrasound

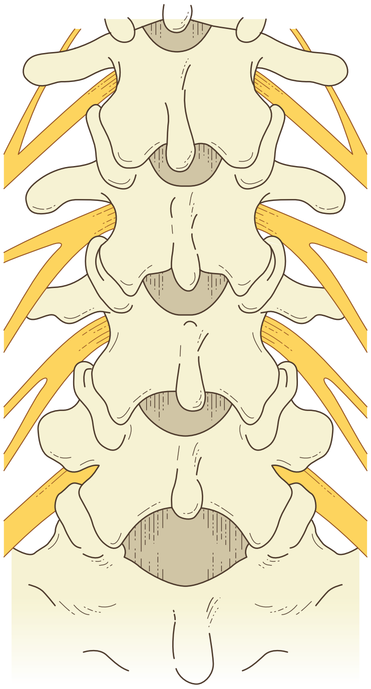

Fig. 4.3.1

Drugs

- Lidocaine (lignocaine) 1% 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 25 mg (or its equivalent)
- Resuscitation drugs

Position of patient

- Prone.
- Pillow under the anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 4.3.2*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The lumbar midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- The iliac crests are palpated and the intercrestal line is identified. This corresponds with the inferior aspect of the spinous process of L4 or may lie in the L4–5 interspace (*Fig. 4.3.3 a*).
- The spinous processes are counted until the level to be blocked is identified and confirmed with fluoroscopy.
- The spinous processes of the vertebrae are marked (*Fig. 4.3.3 b*).
- **The insertion point of the needle lies 2–3 cm lateral to the cephalic end of the spinous process of the vertebra. The nerve corresponding to each vertebra emerges just below the transverse process of that vertebra at this site** (*see Appendix 6*).
- Therefore, with the aid of fluoroscopy, the insertion point is identified.
- A skin wheal is raised and the area is infiltrated with lidocaine (lignocaine) 1%.
- A spinal needle is introduced in a vertical direction to the skin, until the needle contacts bone at an approximate depth of 3–5 cm, the transverse process of that vertebra (*Fig. 4.3.4 a,b*).

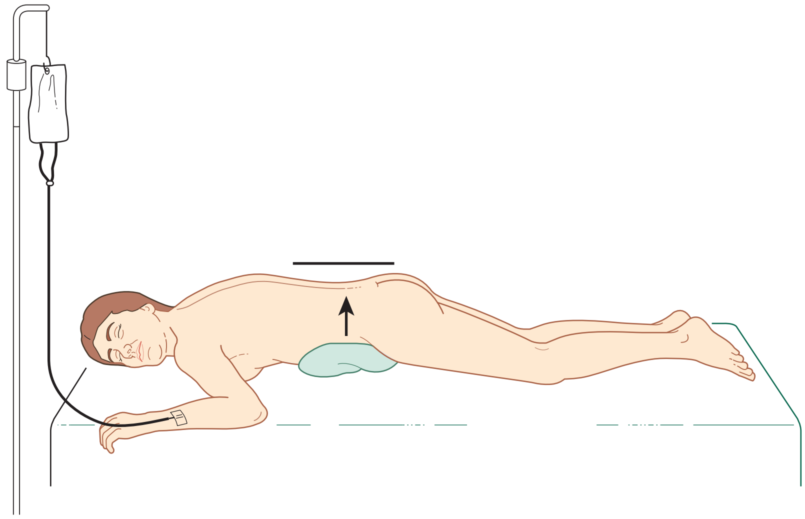

Fig. 4.3.2

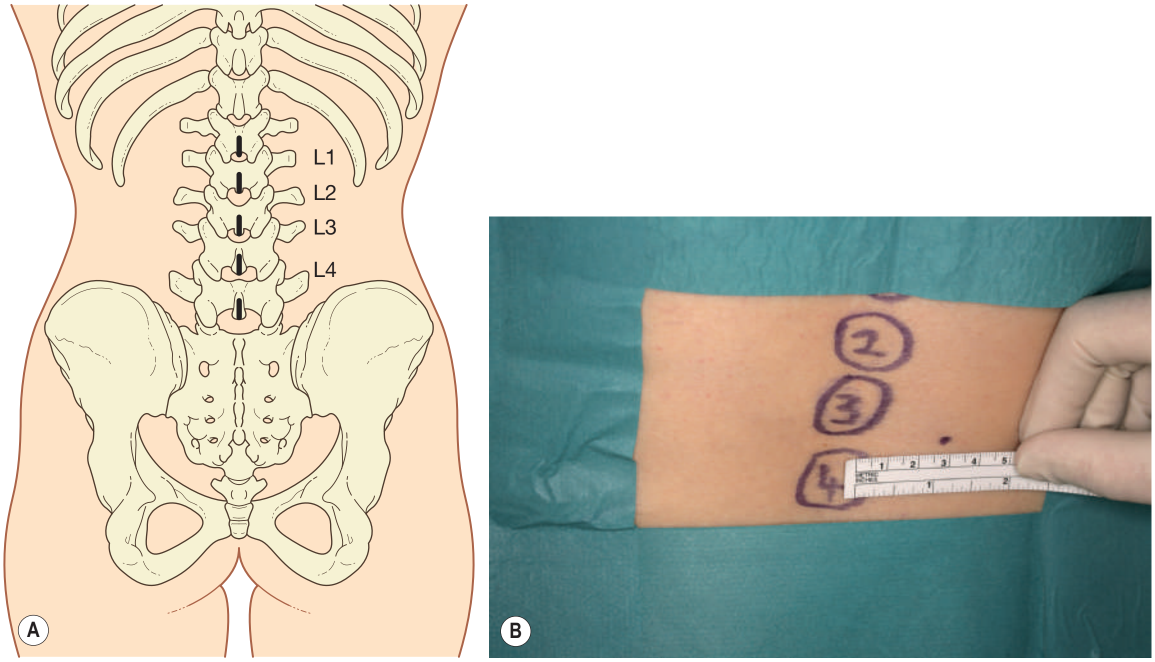

Fig. 4.3.3

- The needle is then walked off the transverse process in the caudad direction and advanced 1.5–2 cm, the site of the emerging nerve root (*Fig. 4.3.5*).
- It is useful to confirm the needle tip over the intervertebral foramen with fluoroscopy.
- Paresthesia in the distribution of the nerve may be experienced.
- After aspiration, non-ionic radio-opaque contrast medium 1 ml is injected.
- The correct placement is indicated by outlining the nerve root with non-ionic radio-opaque contrast medium, visible on anteroposterior and lateral fluoroscopic views (*Fig. 4.3.6*).
- After further aspiration, lidocaine (lignocaine) 1% 0.5–1 ml is injected.
- After 5 minutes the patient is questioned about changes in pain, sensation and power of the lower limb.
- For diagnostic nerve root blockade, the needle may be removed when the level causing pain is identified.
- Ultrasound may also aid placement of the needle (*Fig. 4.3.7*).

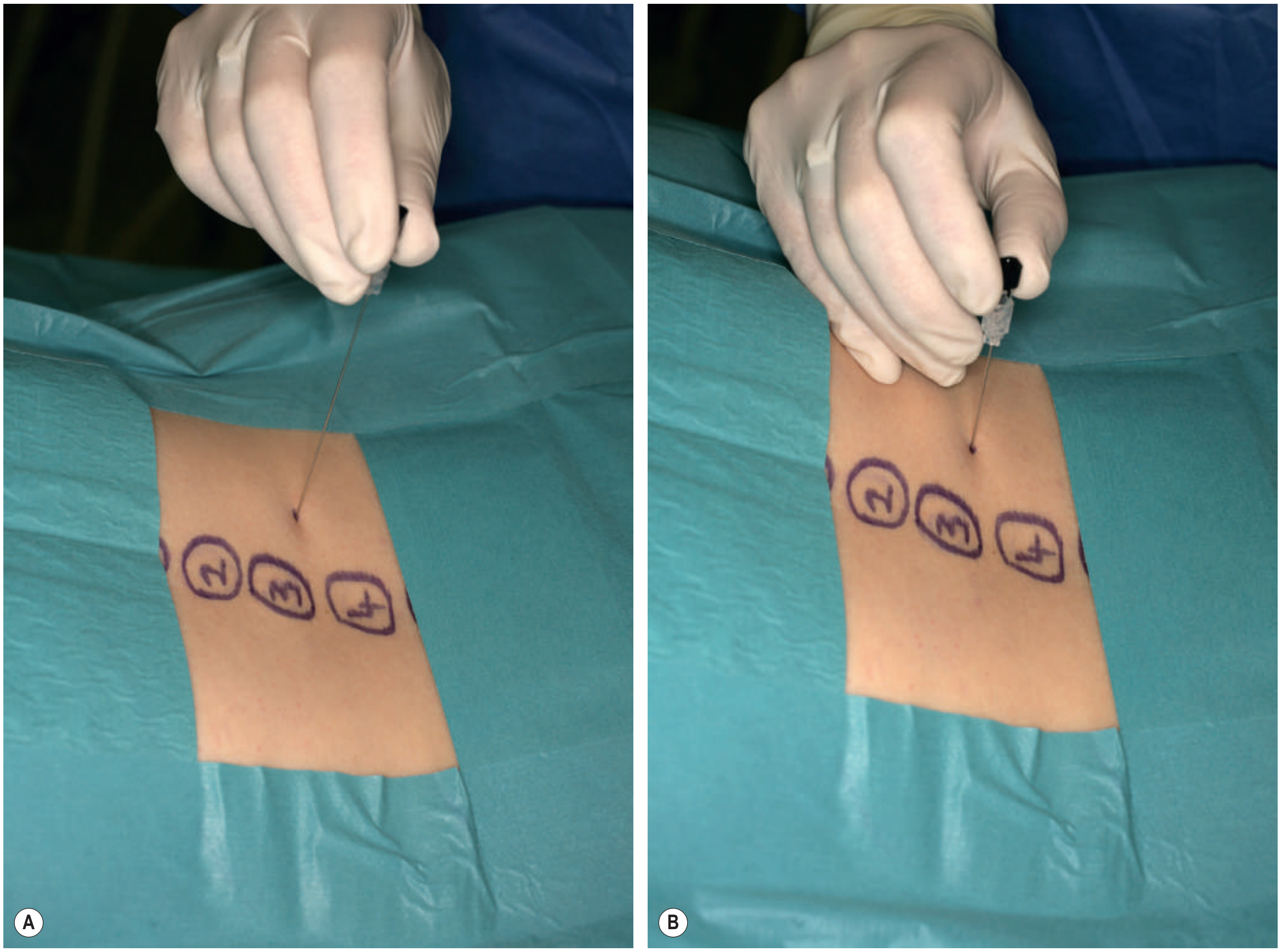

Fig. 4.3.4

Confirmation of a successful block

- Relief of pain and anesthesia in distribution of the blocked nerve.

Tips

- As in the case of thoracic nerve root block, it has also been recommended that the needle is angled 20° medially after entering the paravertebral space. However care must be taken, with the aid of fluoroscopy, not to inject local anesthetic solution into the nerve sheath allowing tracking of the solution centrally to produce intrathecal blockade.

Potential problems

- Intrathecal injection.
- Epidural blockade usually occurs with this block, but this is not a problem once low volumes are used. However, even small volumes of epidural spread may confound the diagnostic value of the block.
- Sympathetic blockade is unlikely, but it may occur and cause hemodynamic changes.
- Intravascular injection. Injection of particulate steroids into a radicular artery can cause spinal cord infarction. Particulate steroids should never be injected near the foramen unless intravascular placement has been ruled out using live fluoroscopy contrast dye injection, preferably with digital subtraction technique.

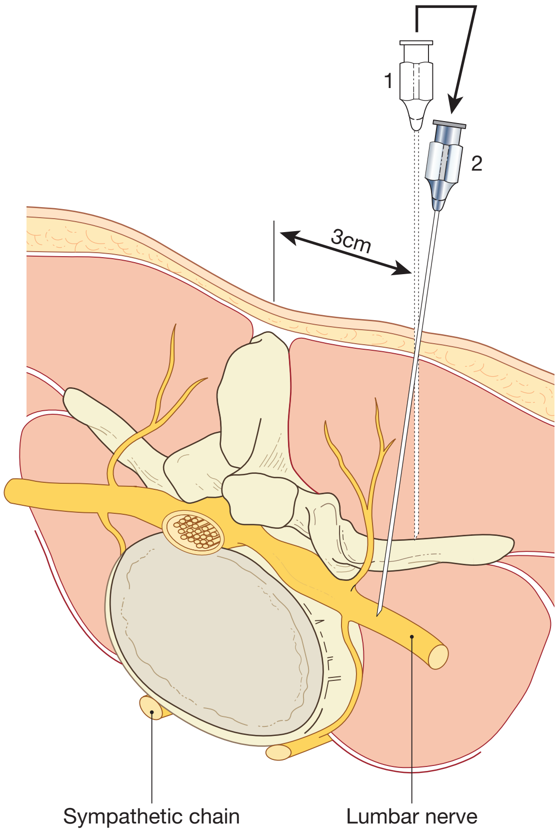
1
2
3cm
Sympathetic chain
Lumbar nerve

Fig. 4.3.5

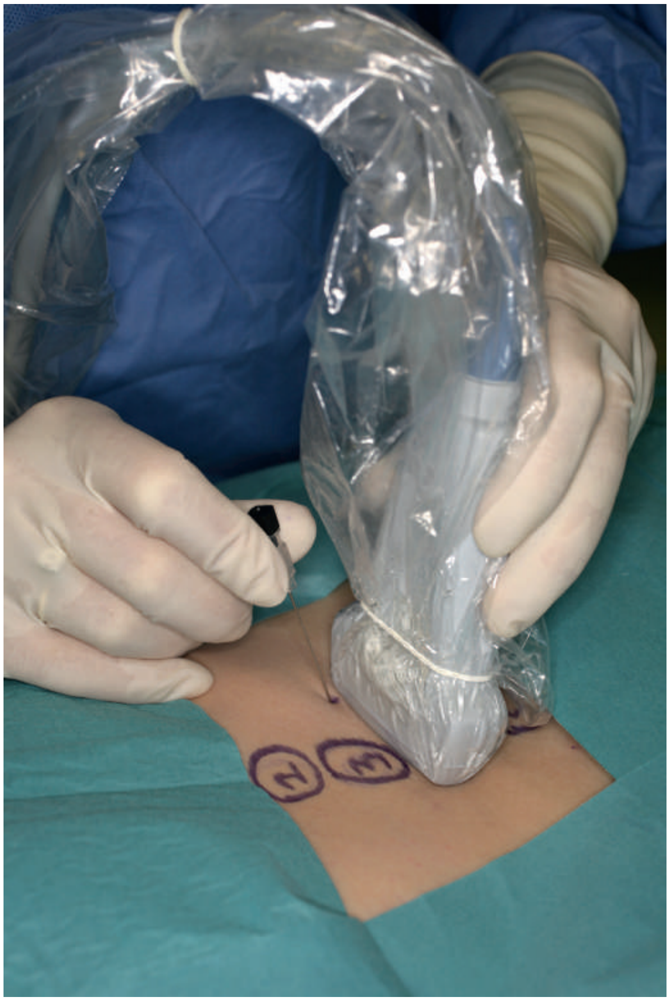

Fig. 4.3.7

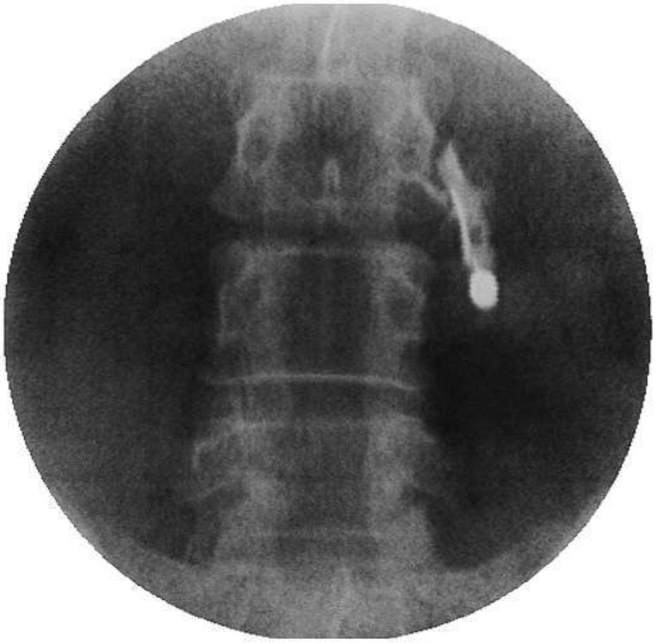

Fig. 4.3.6

4.4 THORACIC NERVE ROOT BLOCK

Anatomy

The anatomy relevant to thoracic paravertebral nerve root blockade is very similar to the anatomy relevant to lumbar paravertebral nerve root block, except that:

- ribs are present instead of the rudimentary ribs of the lumbar spine, the transverse processes;
- the lung and pleura are in close proximity to the injection site, therefore the risk of pneumothorax is significant;
- unlike the lumbar region, where the needle passes through the substance of the psoas muscle, the needle passes through connective tissue only in the thoracic region (*Fig. 4.4.1 a,b*);
- the fascial layers of the psoas muscle prevent spread to the sympathetic lumbar chain. However, such layers do not exist in the thoracic region, and paravertebral injection usually results in sympathetic blockade.

Equipment

- 2 ml and 10 ml syringes
- 25 G needle
- 22 G spinal needle, end-opening
- Radio-opaque contrast medium
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment and drugs (*see Appendix 3*)
- Fluoroscopy or ultrasound

Drugs

- Lidocaine (lignocaine) 1% 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 25 mg (or its equivalent)
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone (the block requires localization of the transverse process and can be performed in a lateral, sitting or prone position) (*Fig. 4.4.2*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The thoracic midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution.

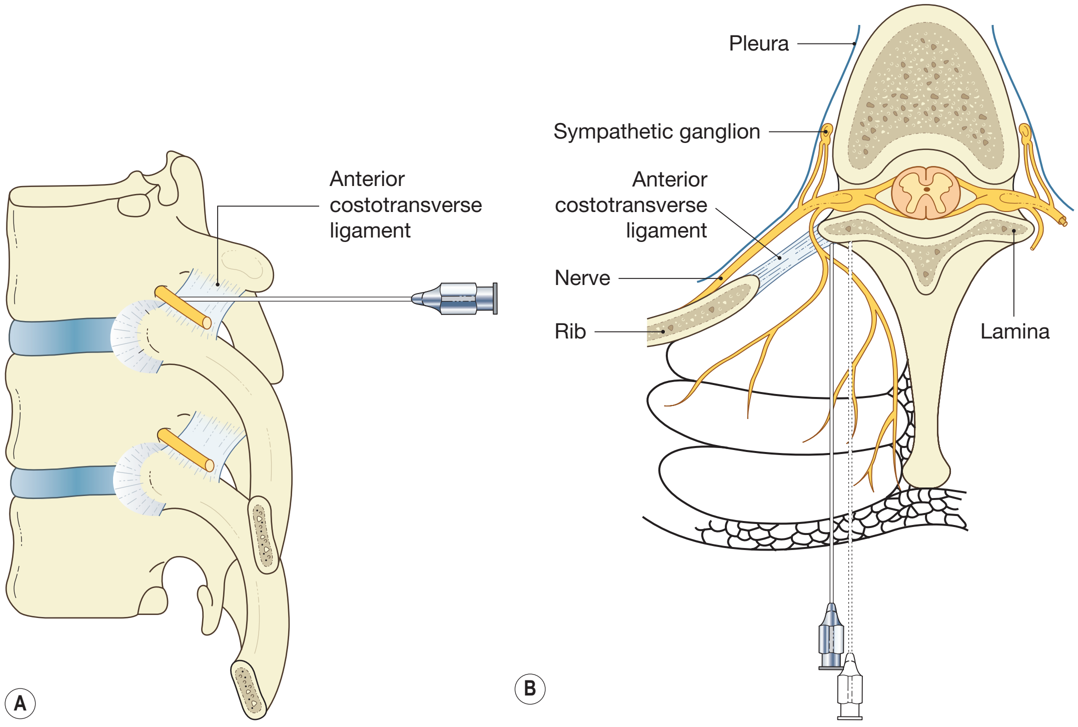
Anterior costotransverse ligament
A
Pleura
Sympathetic ganglion
Anterior costotransverse ligament
Nerve
Rib
Lamina
B

Fig. 4.4.1

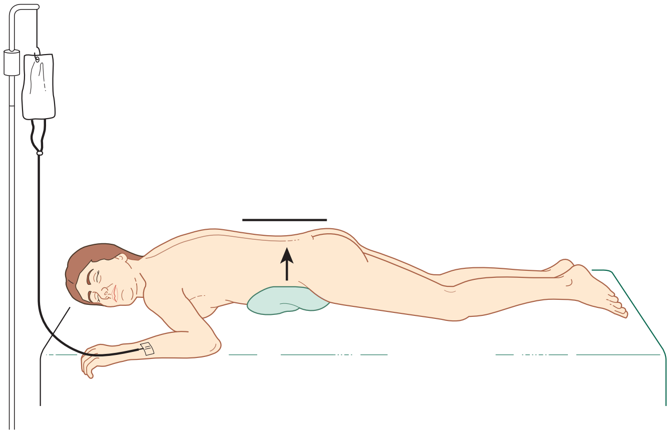

Fig. 4.4.2

- The inferior angle of the scapula is identified; this lies at the level of the spinous process of T7.
- The root of the spine of the scapula is identified; this lies at the level of the spinous process of T3.
- The spinous processes are counted until the level to be blocked is identified, and confirmed with fluoroscopy.
- The spinous processes of vertebrae are then marked.
- **The insertion point of the needle lies 1.5–3 cm lateral to the cephalic end of the spinous process of the vertebra, usually midway between the ribs** (*Fig. 4.4.3 a,b*).
- The nerve corresponding to each vertebra emerges just below the transverse process of that vertebra at this site.
- Therefore, with the aid of fluoroscopy, the insertion point is identified.
- A skin wheal is raised and the area is infiltrated with lidocaine (lignocaine) 1%.
- The transverse process is identified under fluoroscopy and the spinal needle is introduced in a direction perpendicular to the skin until the needle contacts bone.
- A slightly mesiad inclination avoids the pleura and increases the chances of placing the needle tip near the nerve root.
- Care must be taken to avoid the pleura by not advancing the needle any further than is necessary to locate the transverse process, approximately 3 cm. If the needle does not contact the transverse process by 3 cm, it should be withdrawn and re-advanced in a more caudad, and subsequently more cephalic, direction, again taking care not to advance beyond 3 cm.
- The needle is then walked off the caudad edge of the transverse process until it slips off the edge (*Fig. 4.4.4 a,b*). It is then advanced 1 cm, the site of the emerging nerve root at that level.
- It is useful to confirm the needle tip over the intervertebral foramen with fluoroscopy.
- Paresthesia in the distribution of the nerve may be experienced.

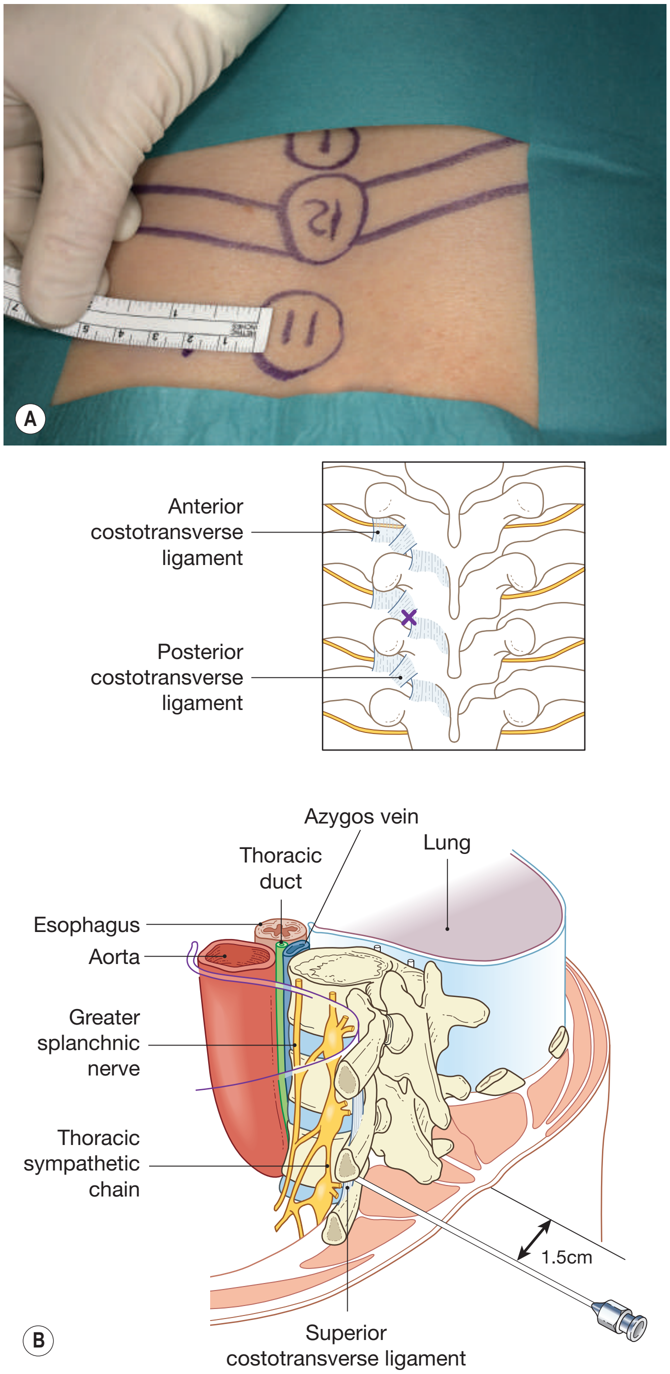
A
12
11
Anterior costotransverse ligament
Posterior costotransverse ligament
Azygos vein
Lung
Thoracic duct
Esophagus
Aorta
Greater splanchnic nerve
Thoracic sympathetic chain
1.5cm
B
Superior costotransverse ligament

Fig. 4.4.3

- After aspiration, radio-opaque contrast medium 1 ml, is injected.
- The correct placement is indicated by outlining the nerve root with non-ionic radio-opaque contrast medium, visible on anteroposterior and lateral views (as in the description of lumbar somatic nerve injection in Section 4.3).
- Ultrasound may aid placement of needle.

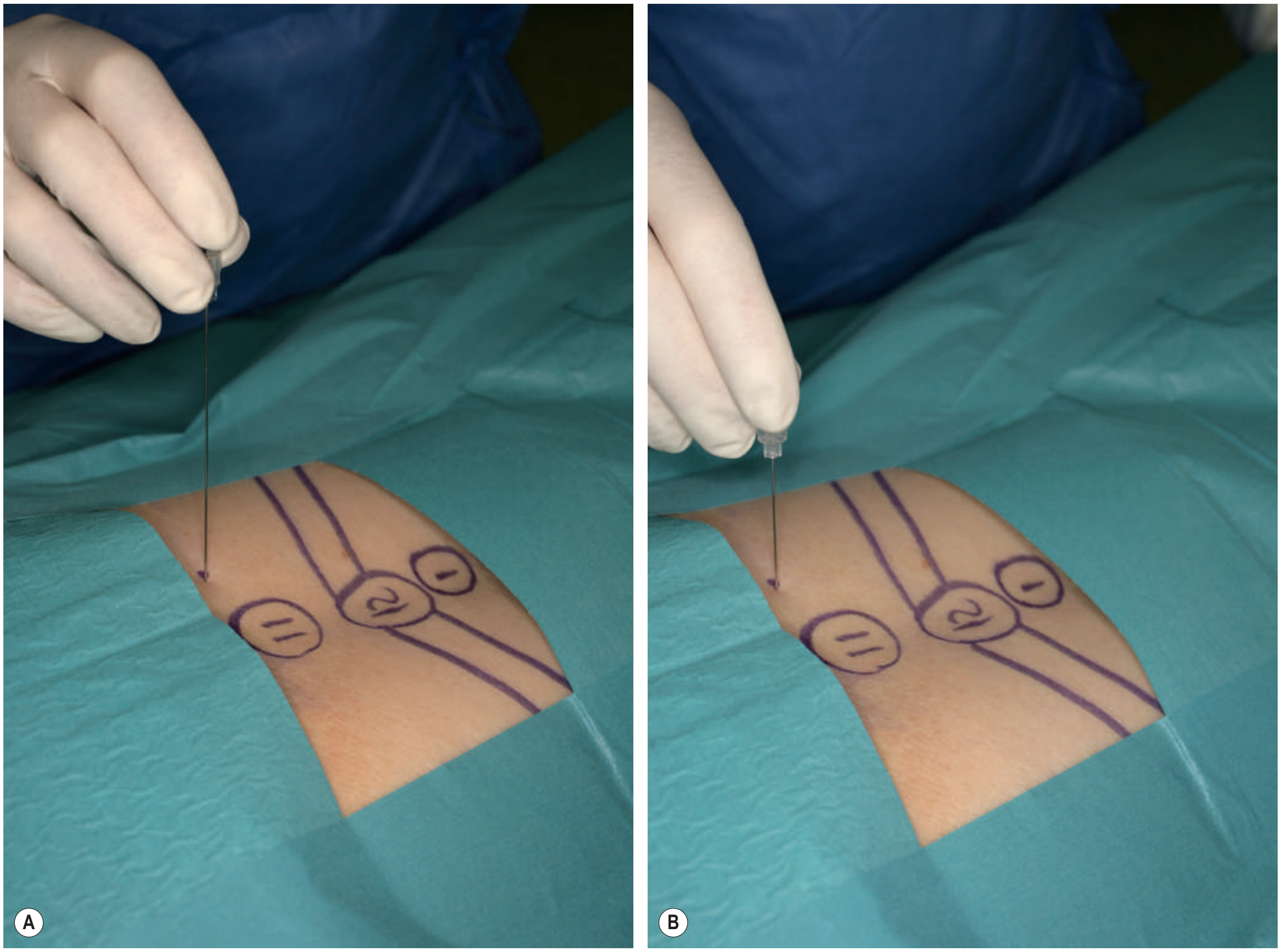

Fig. 4.4.4

- After further aspiration, lidocaine (lignocaine) 1% 0.1–1 ml is injected.
- After 5 minutes the patient is questioned about changes in pain and sensation in the distribution of the nerve root.
- For diagnostic nerve root blockade the needle may be removed when the level causing pain is identified.

Confirmation of a successful block

- Relief of pain and anesthesia in distribution of the blocked nerve.

Tips

- An alternative approach is to advance the needle in a mesiad direction until the lamina is contacted. The needle is inserted more medially, 1.5 cm lateral to the cephalad edge of the spinous process and then walked laterally off the edge of the lamina until it slips into the costovertebral ligament and is advanced 1 cm.
- Some workers advocate applying an air-filled loss-of-resistance syringe to the needle after it has been walked off the transverse process or lamina, and advancing the needle while applying pressure to the plunger. Loss of resistance has been described as the needle pops through the costotransverse ligament to enter the thoracic paravertebral space.
- A catheter may be passed into the paravertebral space via an epidural needle, with the bevel medial, by using this technique.

Potential problems

- Pneumothorax.
- While it has been recommended that the needle be angled 20° medially after entering the paravertebral space, care must be taken with the aid of fluoroscopy that the local anesthetic solution is not injected into the nerve sheath causing the solution to track centrally to produce intrathecal blockade.

- Epidural blockade usually occurs with this block, but this is not a problem once low volumes are used. However, even small volumes of epidural spread may confound the diagnostic value of the block.
- Sympathetic blockade may cause hemodynamic changes.
- Neuritis may occur with catheter placement.
- Intravascular injection. Injection of particulate steroids into a radicular artery can cause spinal cord infarction. Particulate steroids should never be injected near the foramen unless intravascular placement has been ruled out using live fluoroscopy contrast dye injection, preferably with digital subtraction technique.

4.5 SACRAL NERVE ROOT BLOCK

Anatomy

Each of the five sacral nerves is accessible by passing a needle into the sacral foramen via the posterior opening at the level of the nerve. The sacral canal is the caudal extension of the epidural space and nerves of the cauda equina leave via the sacral foramina (**Fig. 4.5.1**). The distal dural sac ends at S2, the level of the posterior superior iliac spines. The epidural space ends at the sacral hiatus (**Fig. 4.5.2**). While variability in the bony anatomy of the sacrum is common, this occurs usually in the midline.

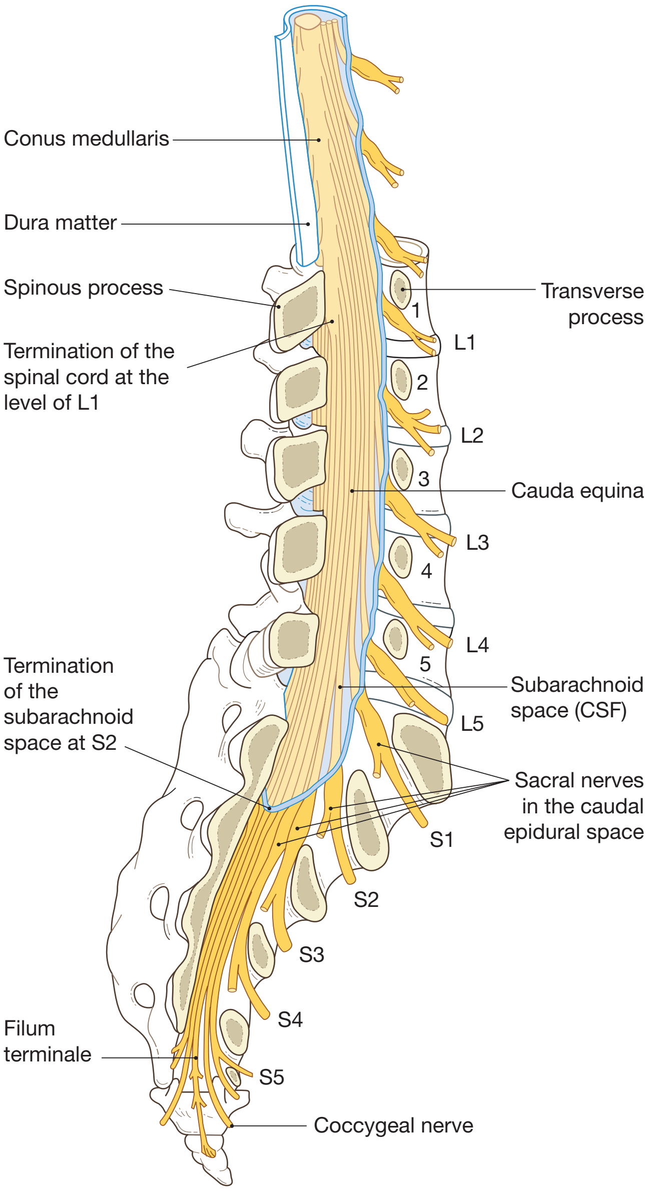
Conus medullaris
Dura matter
Spinous process
Termination of the spinal cord at the level of L1
Transverse process
1
L1
2
L2
3
Cauda equina
L3
4
L4
5
Subarachnoid space (CSF)
L5
Termination of the subarachnoid space at S2
Sacral nerves in the caudal epidural space
S1
S2
S3
S4
S5
Filum terminale
Coccygeal nerve

Fig. 4.5.1

Equipment

- 2 ml and 10 ml syringes
- 25 G needle
- 22 G spinal needle, end-opening
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*) and drugs
- Fluoroscopy or ultrasound

Drugs

- Lidocaine (lignocaine) 1% 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 50 mg (or its equivalent)
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Pillow under anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 4.5.3*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The sacral midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.

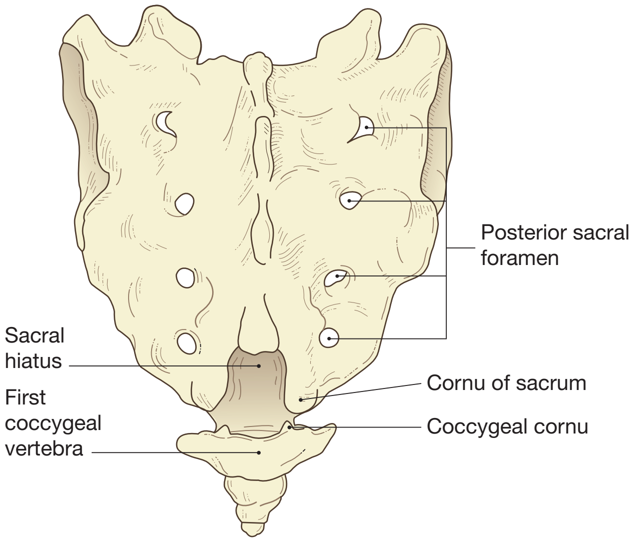
Posterior sacral foramen
Sacral hiatus
Cornu of sacrum
First coccygeal vertebra
Coccygeal cornu

Fig. 4.5.2

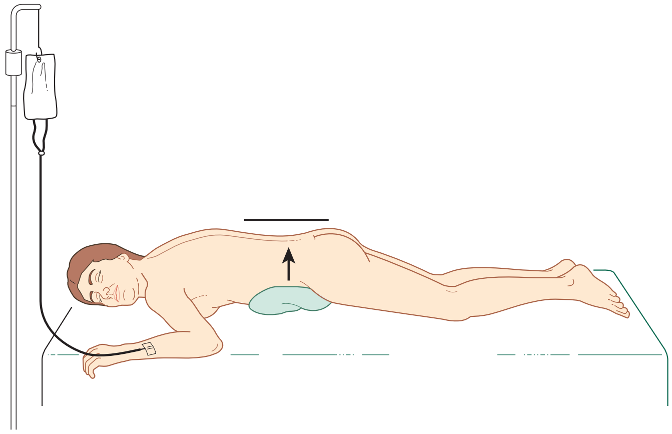

Fig. 4.5.3

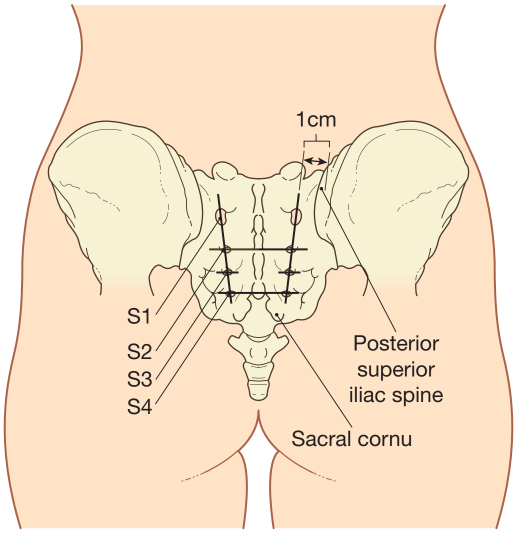

Fig. 4.5.4

- Iliac crests are palpated and an intercrestal line is identified.
- This corresponds with the inferior aspect of the spinous process of L4 or may lie in the L4–5 interspace.
- For this block, it is almost essential that the posterior opening of the sacral foramina is identified with fluoroscopy. The X-ray beam is used to guide the needle tip into the foramen. It is important to note that when the anteroposterior X-ray view is used, it is usually the anterior opening of the foramen that is the most prominent.
- **The insertion point of the needle lies 2–3 cm lateral to the midline (variable) and approximately 1 cm medial to the posterior iliac spine** (*Figs 4.5.4, 4.5.5 a,b*).
- Therefore, with the aid of fluoroscopy, the insertion point is identified.
- A skin wheal is raised and the area is infiltrated with lidocaine (lignocaine) 1%.

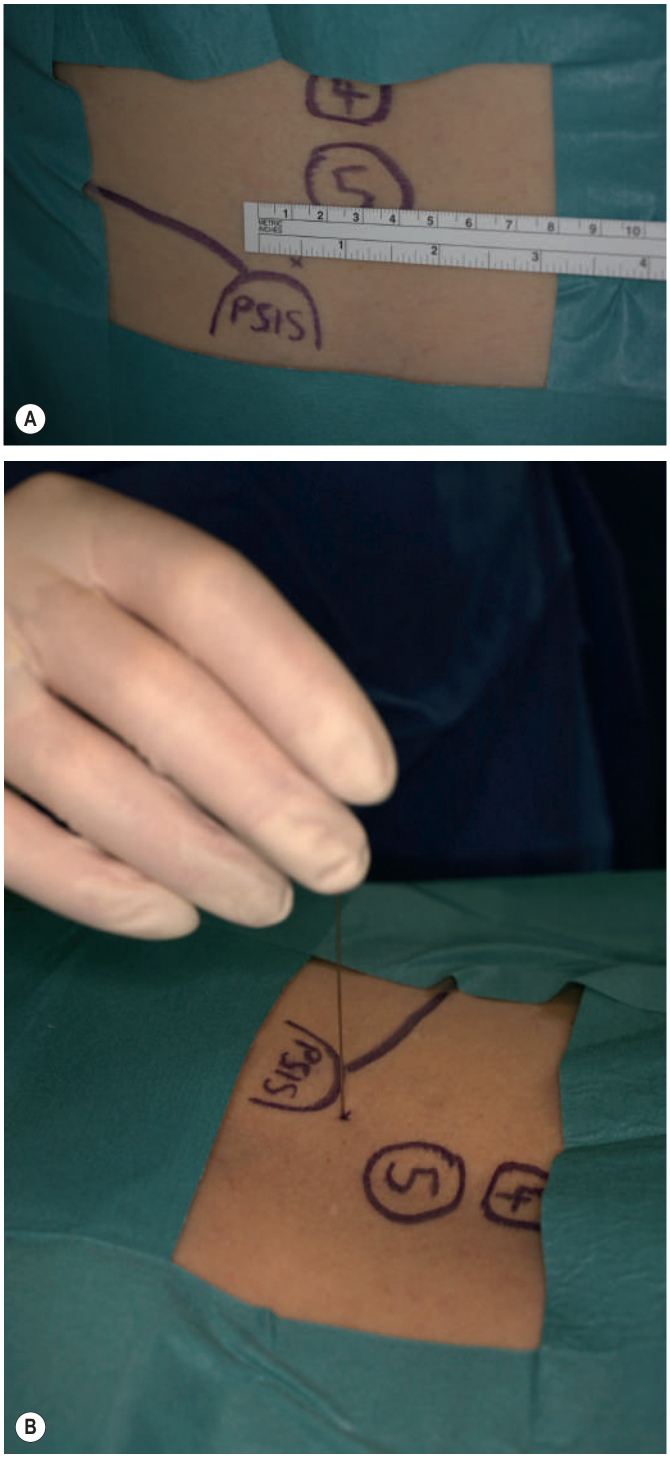

Fig. 4.5.5

- A spinal needle is introduced in a vertical direction to the skin, and aimed slightly cephalad until bone is contacted (*Fig. 4.5.6*).
- The needle is then walked off the sacrum in a caudad direction until it slips into the foramen. After confirmation with the aid of fluoroscopy it is then advanced 1 cm.

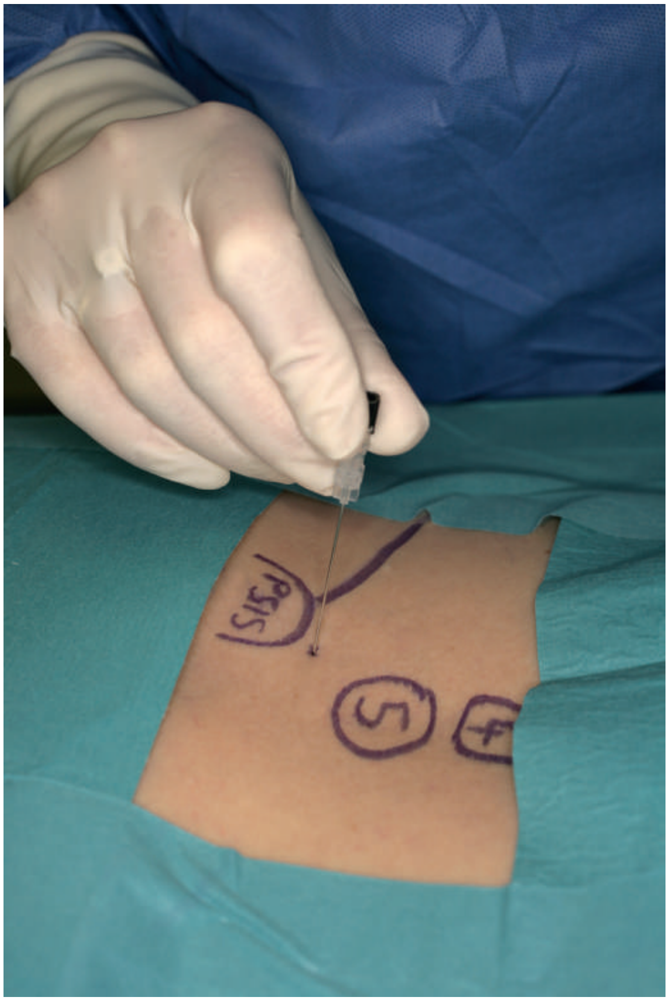

Fig. 4.5.6

- Paresthesia may be produced.
- After aspiration, non-ionic radio-opaque contrast medium 0.5 ml is injected.
- The correct placement is indicated by a needle tip flush with the anterior surface of the spinal canal in the lateral fluoroscopic view.
- Injection of 0.5 ml non-ionic contrast should spread diagonally along the S1 spinal nerve (*Fig. 4.5.7*).
- After further aspiration, lidocaine (lignocaine) 1% 0.5 ml is injected.
- After 5 minutes, the patient is questioned about changes in pain, sensation and power of the lower limb.
- For diagnostic nerve root blockade the needle may be removed when the level causing pain is identified.
- Ultrasound may also aid placement.

Confirmation of a successful block

- Relief of pain.
- Anesthesia in the distribution of the blocked nerve.

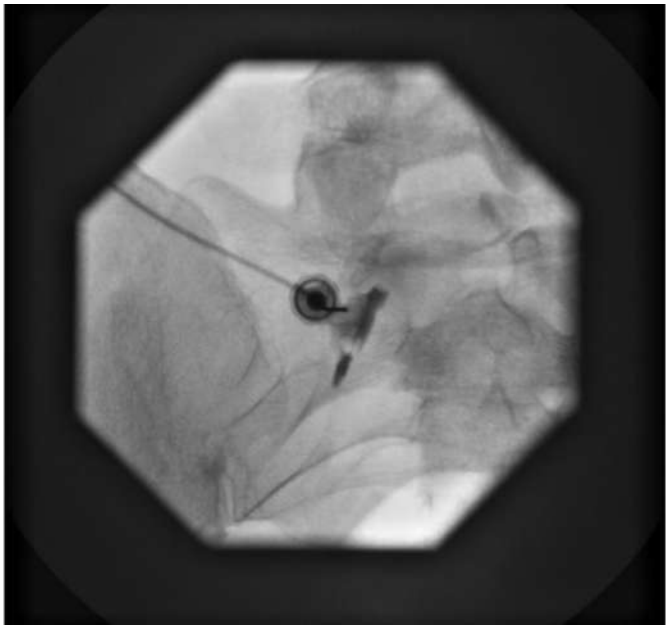

Fig. 4.5.7

Tips

- Some workers advocate drawing a line from a point 2–3 cm medial to the posterior superior iliac spine to a point 1–2 cm lateral to the sacral cornua. The sacral foramina usually lie along this line.
- It is best to angle the X-ray beam caudally, thereby perpendicular to the sacrum, superimposing the anterior and posterior sacral foramina. Consequently, when the needle is introduced in a direction perpendicular (*Fig. 4.5.8*) to the skin, fluoroscopic guidance is easier. Optimally, the needle makes gentle contact with the spinal nerve in the middle of the canal (*Figs 4.5.9*, *4.5.10*) and produces paresthesia in the distribution of the nerve.
- While all sacral nerve roots are accessible, using this technique for blockade of S5 is achieved by walking the needle caudally off the inferior edge of the bony plate of the sacrum and advancing the needle 1 cm.

Potential problems

- Caudal epidural blockade may occur with this block, but this is not a problem once low volumes are used. However, even small volumes of epidural spread may confound the diagnostic value of the block.
- Intravascular injection. Injection of particulate steroids into a radicular artery can cause spinal cord infarction. Particulate steroids should never be injected near the foramen unless intravascular placement has been ruled out using live fluoroscopy contrast dye injection, preferably with digital subtraction technique.

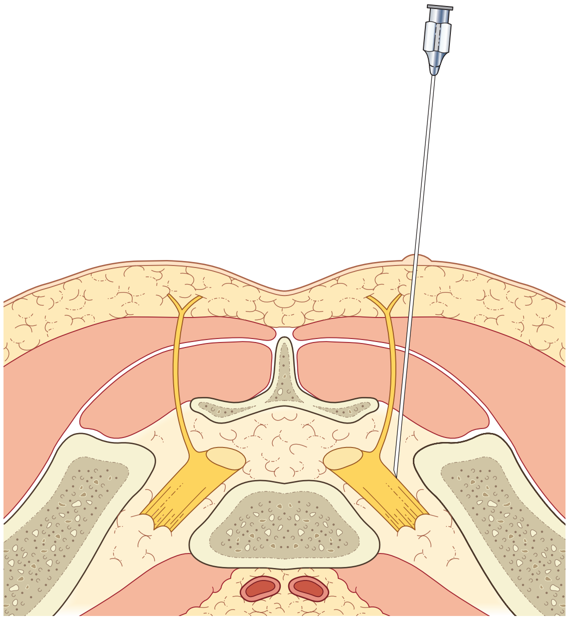

Fig. 4.5.8

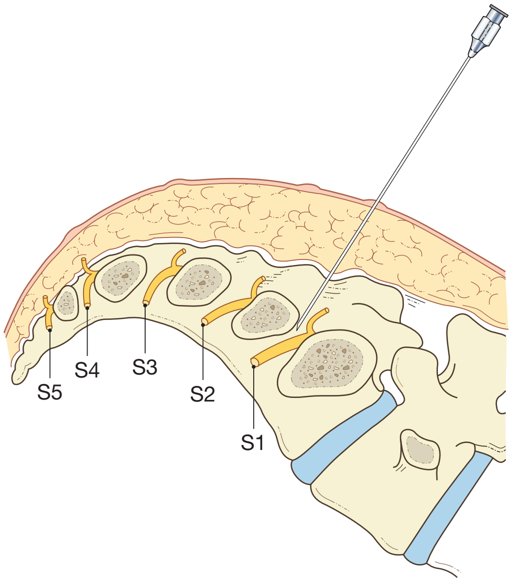
S5
S4
S3
S2
S1

Fig. 4.5.9

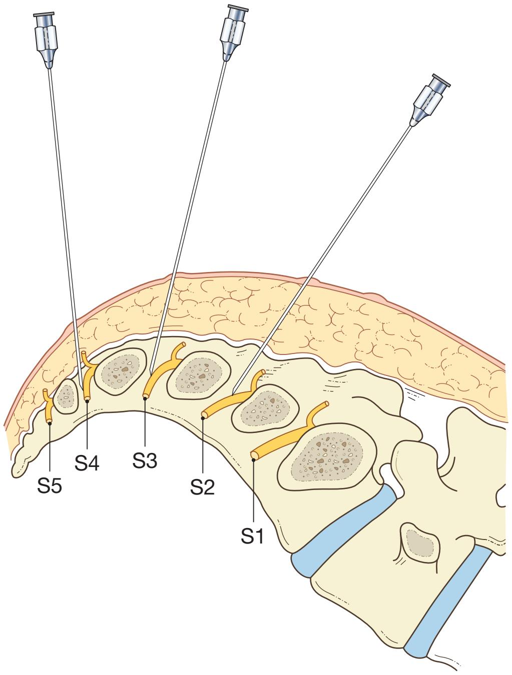
S5
S4
S3
S2
S1

Fig. 4.5.10

4.6 OCCIPITAL NERVE BLOCK

Anatomy

The greater occipital nerve arises from the dorsal rami of the second cervical nerve. From here it passes through the muscles of the neck and becomes subcutaneous at the superior nuchal line, where it emerges with the occipital artery (**Fig. 4.6.1**). The superior nuchal line extends from the mastoid process to the greater occipital protuberance bilaterally (**Fig. 4.6.2**).

Equipment

- 10 ml syringe
- 25 G needle
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)

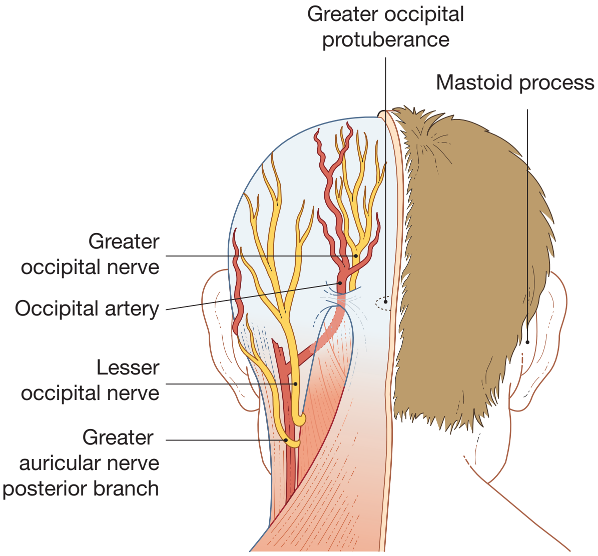
Greater occipital protuberance
Mastoid process
Greater occipital nerve
Occipital artery
Lesser occipital nerve
Greater auricular nerve posterior branch

Fig. 4.6.1

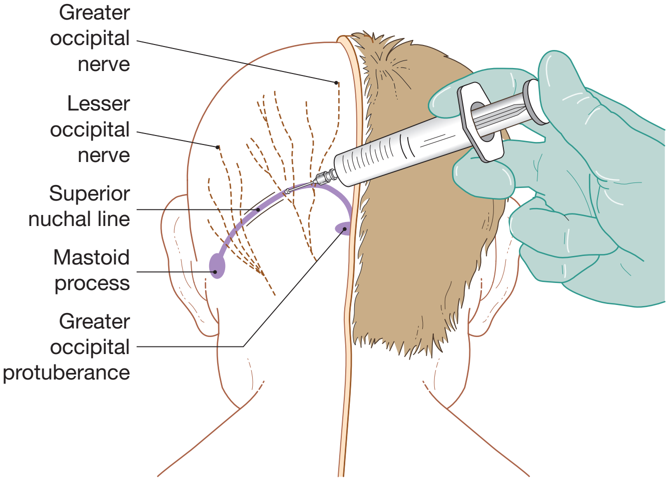
Greater occipital nerve
Lesser occipital nerve
Superior nuchal line
Mastoid process
Greater occipital protuberance

Fig. 4.6.2

Drugs

- Lidocaine (lignocaine) 1% 10 ml (or its equivalent)
- Corticosteroid if indicated, e.g. triamcinolone diacetate 50 mg (or its equivalent)
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Sitting.
- Neck flexed.

Needle puncture and technique

- The superior nuchal line is cleaned with antiseptic solution (no drape is required) (*Fig. 4.6.3*).

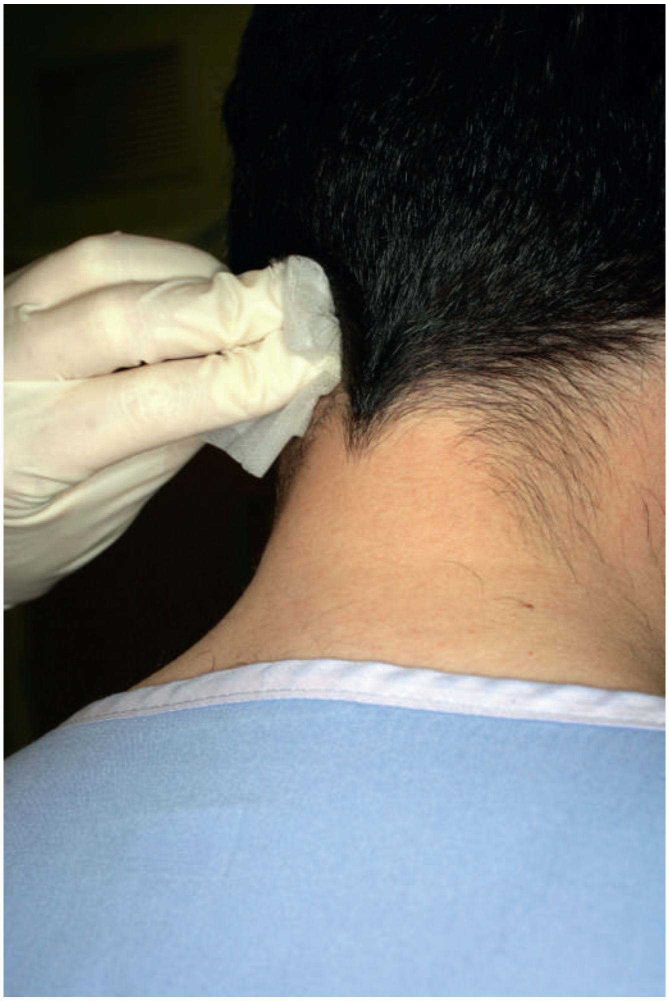

Fig. 4.6.3

- The occipital artery is palpated 2 cm lateral to the greater occipital protuberance on the superior nuchal line (*Fig. 4.6.4*).
- A 25 G needle is inserted subcutaneously at this point.
- After negative aspiration, 3–5 ml of lidocaine (lignocaine) 1% or its equivalent, plus corticosteroid if indicated, is injected to surround the occipital artery.
- It is often difficult to feel an occipital artery pulse. If this is the case, it is best to pick a point midway between the occipital protuberance and the mastoid bone and fan out the injection in both directions, medially and laterally from that site (*Fig. 4.6.5*).

Confirmation of a successful block

- Relief of pain and anesthesia in distribution of nerve.

Tips

- Bone should be contacted at a depth no greater than 1–2 cm.
- The lesser occipital nerve is blocked by redirecting the needle towards the mastoid process along the greater nuchal line and injecting a further 3 ml of solution.

Potential problems

- Injection into the cerebrospinal fluid (CSF) of the cisterna magna is possible and will produce a total spinal block.

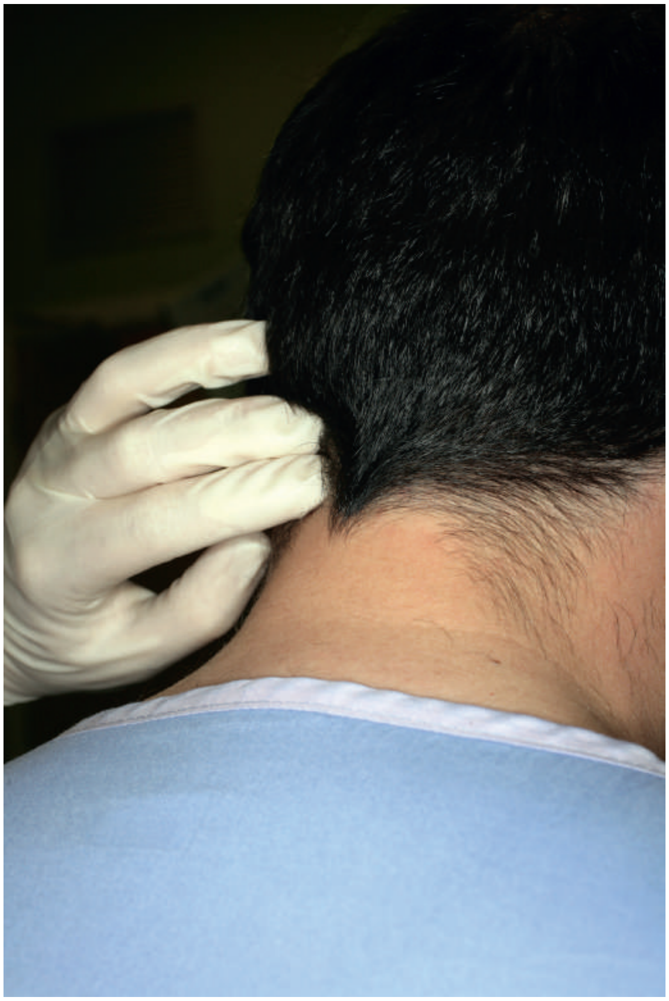

Fig. 4.6.4

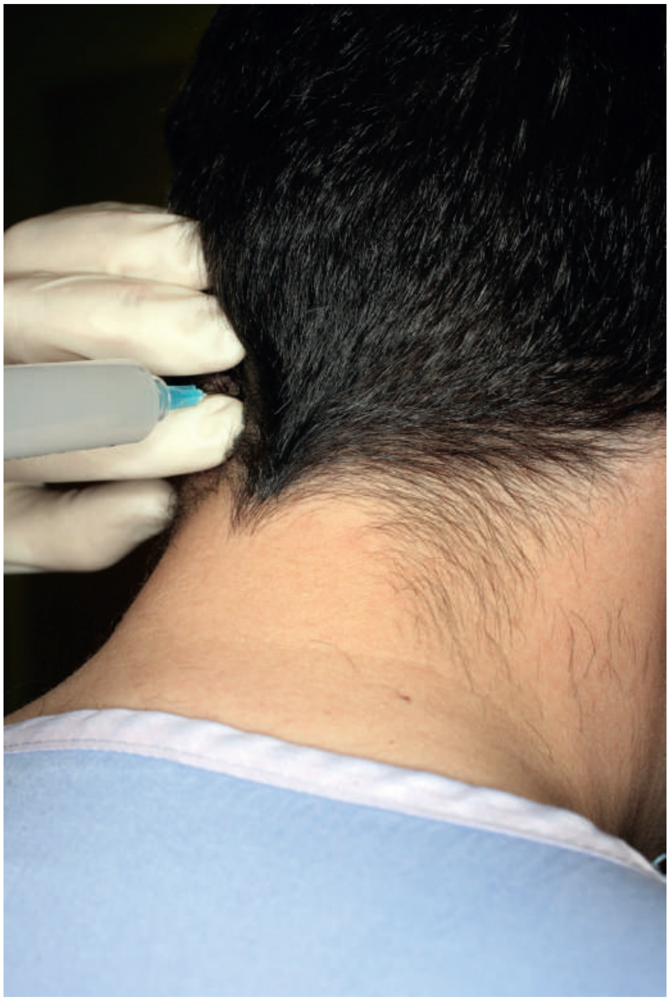

Fig. 4.6.5

4.7 TRIGEMINAL GANGLION (GASSERIAN) BLOCK

Anatomy

The trigeminal ganglion gives rise to the fifth cranial nerve and divides into three branches, the ophthalmic, maxillary, and mandibular nerves (Fig. 4.7.1). These provide the sensory nerve supply to the ipsilateral face and the anterior two-thirds of the head (Fig. 4.7.2). The mandibular nerve also provides motor supply to the muscles of mastication.

The trigeminal ganglion is located at the apex of the petrous temporal bone in a fold of dura mater, "Meckel's cave". This dural invagination covers the posterior two-thirds of the ganglion and contains cerebrospinal fluid (CSF). Posterior to Meckel's cave lies the brainstem, superior to it lies the temporal lobe, and medially lies the cavernous sinus which contains the internal carotid artery and cranial nerves III, IV, and VI. Accordingly, extreme care must be taken when carrying out this block, especially if neurolytic agents are used. Blockade of the ganglion is carried out by passage of a needle through the foramen ovale, which lies immediately below it (Fig. 4.7.3).

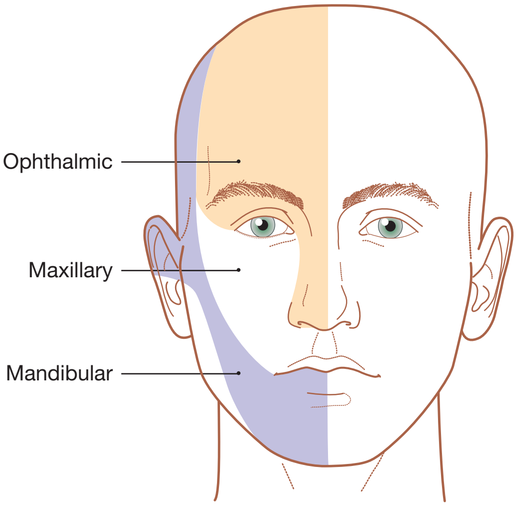
Ophthalmic
Maxillary
Mandibular

Fig. 4.7.1

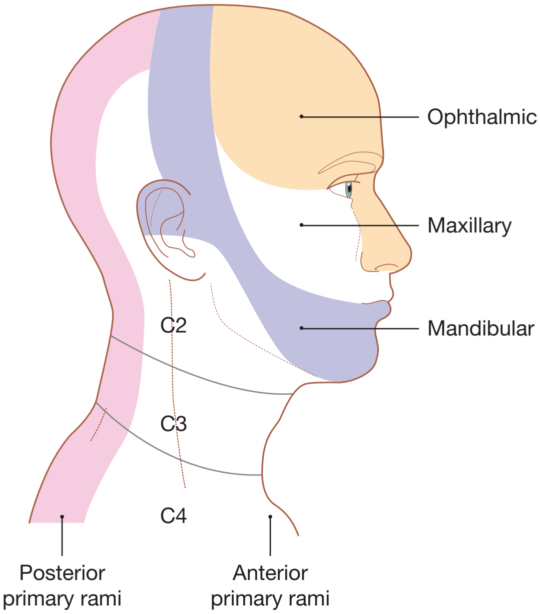
Ophthalmic
Maxillary
Mandibular
C2
C3
C4
Posterior primary rami
Anterior primary rami

Fig. 4.7.2

Equipment

- 2 ml and 10 ml syringes
- 25 G needle
- 22 G 8–10 cm needle
- Non-ionic radio-opaque contrast medium
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy

Drugs

- Lidocaine (lignocaine) 2% 10 ml
- Lidocaine (lignocaine) 1% 10 ml (or its equivalent)
- Neurolytic agent, e.g. phenol 6% plus glycerol (or its equivalent)
- Sedative, e.g. midazolam, propofol
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Supine.
- Eyes directed straight ahead.

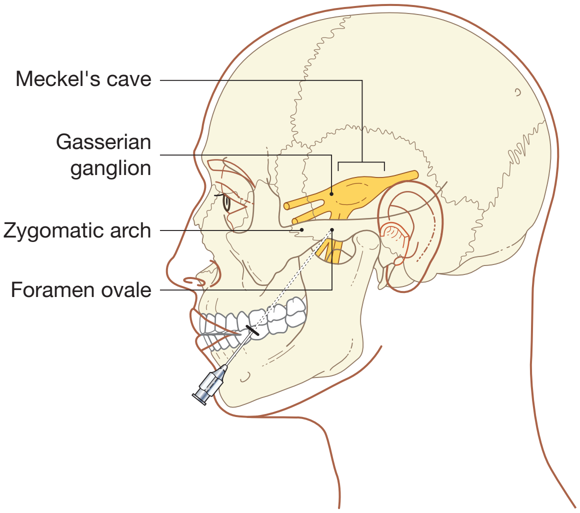
Meckel's cave
Gasserian ganglion
Zygomatic arch
Foramen ovale

Fig. 4.7.3

Needle puncture and technique

Caution: Injection of 0.25 ml of lidocaine (lignocaine) 1% into the CSF may result in immediate convulsion and/or loss of consciousness with possible cardiovascular system (CVS) collapse.

- The cheek on the side of the block is cleaned with antiseptic solution or saline.
- Mild sedation is induced.
- It is best to stand on the side of the block, just below the shoulder.
- **The insertion point lies 1–3 cm posterior to the lateral margin of the mouth, at the medial edge of the masseter muscle (located by asking the patient to clench the jaw) and is marked.**
- **In edentulous patients the insertion point should be more caudad** as sufficient angle towards the infratemporal surface of the sphenoid bone may not be achieved.
- One finger is placed inside the upper lip to avoid injection into the buccal cavity and possible bacterial contamination, and a skin wheal is raised at this site.
- Viewed from above, a 22 G 8–10 cm needle is advanced towards the ipsilateral pupil (*Figs 4.7.4–4.7.6*), or viewed from the side the needle advances towards the mid-point of the zygomatic arch (see Anatomy above) until bone is contacted; the roof of the infratemporal fossa. This lies just anterior to the foramen ovale and lateral to the base of the pterygoid process. The location of the needle tip is confirmed with fluoroscopy.
- A depth mark is set and the needle is withdrawn to the subcutaneous tissue. With the aid of fluoroscopy the needle is reinserted to walk off the bone and enter the foramen ovale (*Fig. 4.7.7*).
- Paresthesia in the distribution of the mandibular nerve (sometimes the other branches of the trigeminal nerve) or contraction of the muscles of mastication may be experienced at this point.
- The needle is advanced a further 1 cm to bring the tip to lie in the trigeminal ganglion. The correct placement is indicated by a visible outline of Meckel's cave on injection of 0.25 ml non-ionic radio-opaque contrast medium under fluoroscopy (*Figs 4.7.8*, *4.7.9*).
- The patient is allowed to awaken from sedation and is questioned about the presence and distribution of paresthesia and pain.
- A stimulating device may aid placement in patients who are not able to locate the paresthesia with accuracy.
- If necessary, analgesia may be administered, although this may affect accurate assessment of the blockade.
- Adjustment of the needle may be required to place the needle near the appropriate nerve division.

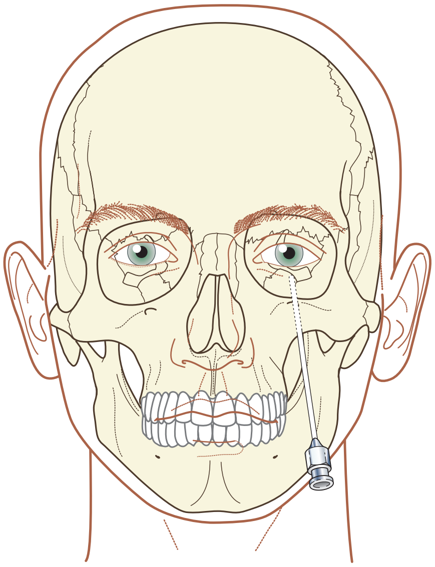

Fig. 4.7.4

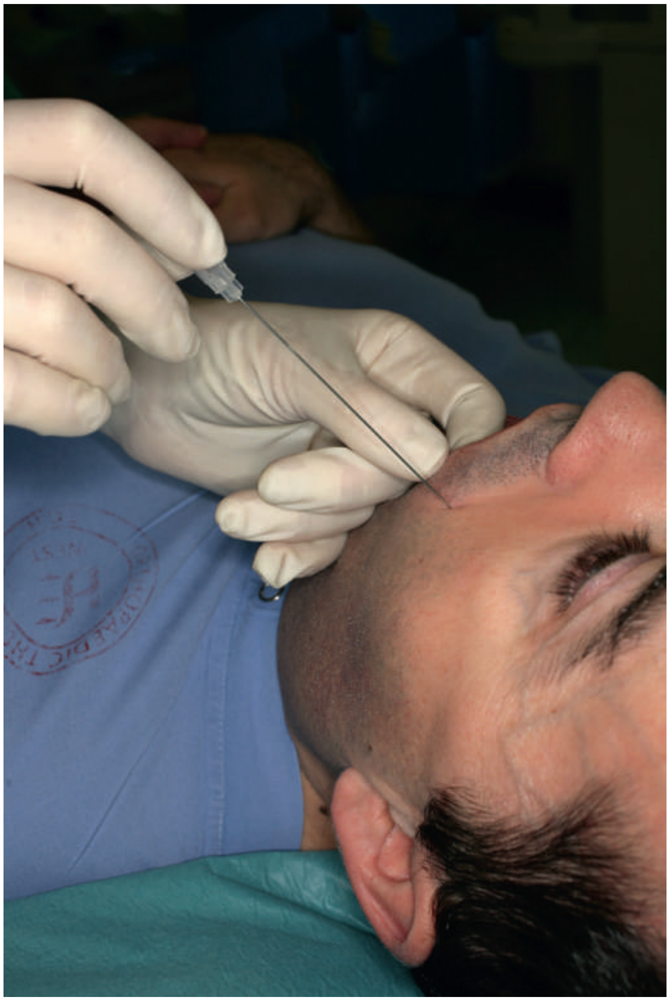

Fig. 4.7.5

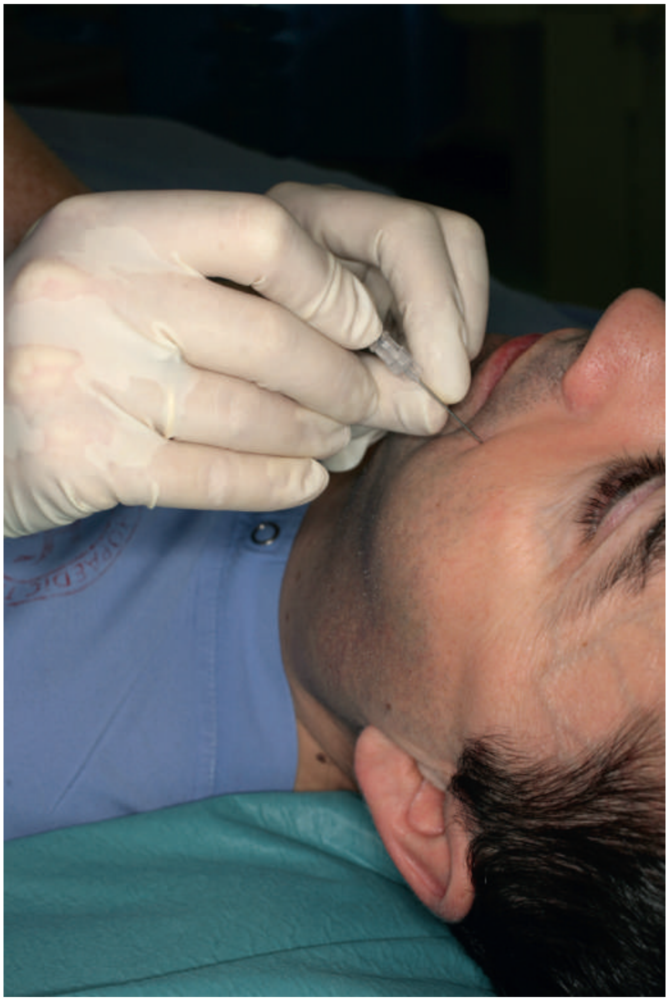

Fig. 4.7.6

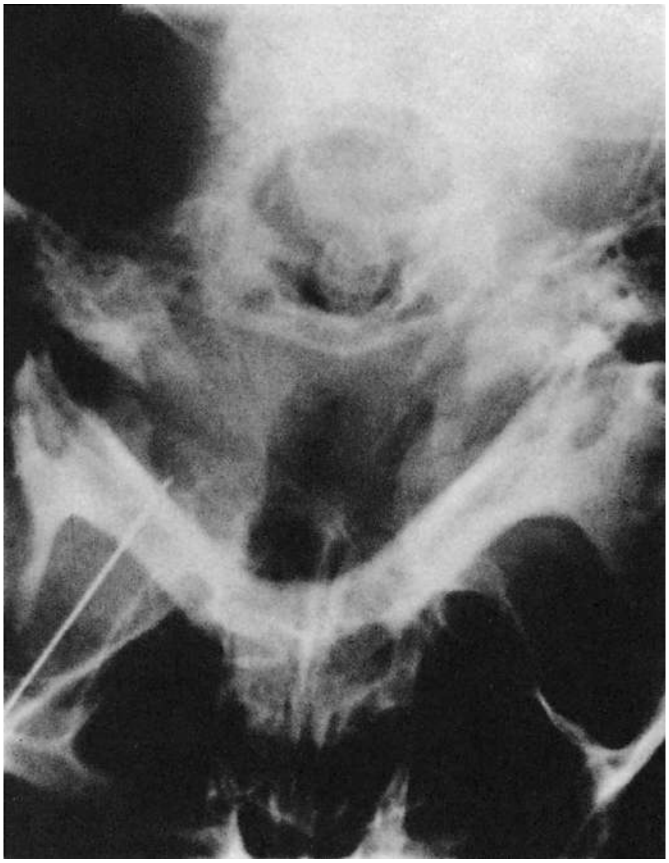

Fig. 4.7.8

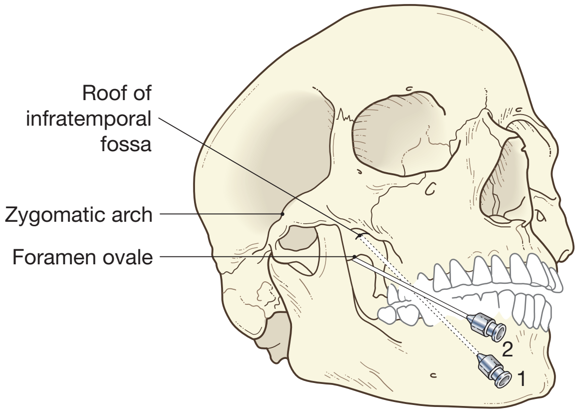
Roof of infratemporal fossa
Zygomatic arch
Foramen ovale
2
1

Fig. 4.7.7

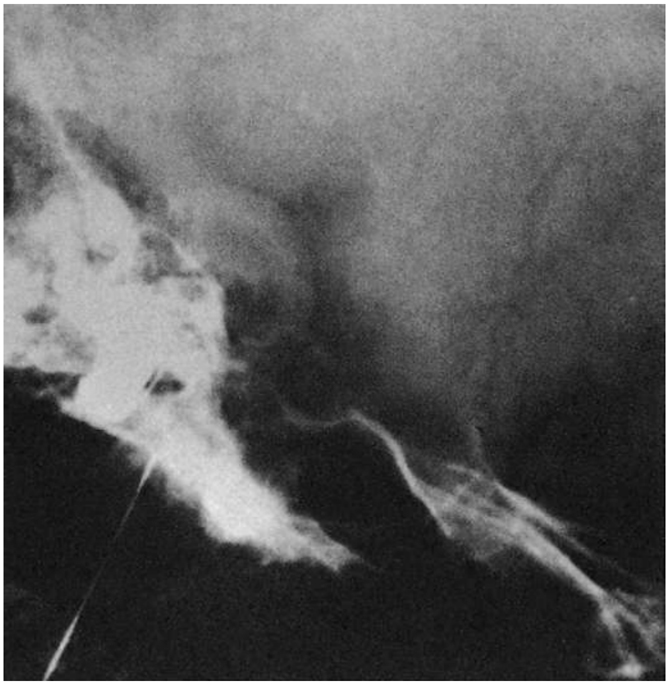

Fig. 4.7.9

- After careful negative aspiration for CSF or blood, lidocaine (lignocaine) 1% 0.25 ml is injected (***Caution:*** ***injection into CSF may cause loss of consciousness***.) This is followed by further boluses of lidocaine (lignocaine) 1% 0.25 ml until a total of 1 ml is given.
- After 5 minutes the patient is questioned about pain relief and changes in sensation.
- When the desired analgesia has been achieved for diagnostic blockade the needle may be removed.

Confirmation of a successful block

- Relief of pain and anesthesia in the distribution of the trigeminal nerve or its desired branches.

Tips

- Intravenous anesthesia using a short-acting agent, e.g. propofol, may be induced to allow placement of the needle. The patient is then allowed to awaken and a stimulating device may aid accurate placement of the needle.
- The needle is advanced beyond the infratemporal bone by 0.5 cm for location of the mandibular division, 1.0 cm for the maxillary division, and 1.5 cm for the ophthalmic division.
- Gangliolysis using thermocoagulation may be employed for trigeminal-nerve division destruction after location of the ganglion using this technique. Further intravenous anesthesia using a short-acting agent, e.g. propofol, may be induced after placement of the insulated needle to facilitate this painful procedure.
- Injection of glycerol alone may produce pain relief with this injection technique. This involves placement of the needle in the cul-de-sac of CSF, positioning the patient face-down or supine, with the head extended to prevent spill-over into the posterior cranial fossa. After entry into the CSF, and positive aspiration of CSF, 0.1–0.3 ml of glycerol may be injected.

Potential problems

- Injection of 0.25 ml of lidocaine (lignocaine) 1% into the CSF may result in immediate convulsion and/or loss of consciousness with possible CVS collapse.
- Spread of hyperbaric neurolytic solution may immediately affect cranial nerves VI, VIII, IX, X, XI, and XII.
- Spread of hypobaric neurolytic solution may immediately affect the oculomotor and trochlear nerves.
- Neurolytic blocks of the trigeminal ganglion commonly produce corneal and hemifacial anesthesia.

AUTONOMIC BLOCKADE

5

Autonomic blockade is useful in the diagnosis and treatment of pain of autonomic origin. In cases of thoracic, abdominal or pelvic pain, it is often difficult to distinguish between that of visceral origin and that of somatic. Pain of visceral origin, e.g. pancreatic cancer or pelvic cancer, may cause pain that responds to celiac or hypogastric plexus blockade, respectively. A prognostic block may be carried out prior to neurolytic blockade for relief of cancer pain. Pain of the upper abdominal viscera may also be relieved by celiac plexus block, proceeding to neurolytic blockade as appropriate for cancer-related pain. The retrocrural approach to the celiac plexus also targets the splanchnic nerves to produce a splanchnic nerve block if required. Chest pain may be of somatic origin, e.g. intercostal neuralgia and costochondritis, or visceral origin, e.g. pulmonary or cardiac-related pain. Stellate ganglion blockade may be helpful in the diagnosis and management of the latter.

In addition, increased sympathetic activity is thought to contribute to a large number of pain states. These are generally grouped under the term Complex Regional Pain Syndrome Type I and II. In these cases, trophic changes and alterations in blood flow are often obvious but the pathophysiologic origin is not. Blockade of sympathetic innervation may therefore help in diagnosis and management of such pain. This may also indicate other therapies that could be beneficial, e.g. sympathetically active drugs or destructive therapies. Similarly, these therapies would not be indicated if sympathetic blockade failed to relieve the pain. If blockade did succeed in relieving this type of pain, further blocks may effect lasting relief.

There are a number of other conditions in which the diagnosis is clear but there is a possible contribution of sympathetic activity in the pathogenesis of the pain. These conditions include central pain, post-herpetic neuralgia, trigeminal neuralgia, peripheral vascular disease and others. Blockade of sympathetic activity may help to clarify the sympathetic contribution to the pain and therefore help to indicate management options.

5.1 STELLATE GANGLION BLOCK—C6 (CLASSIC) APPROACH

Anatomy

The cervical sympathetic trunk—the superior, middle, and stellate ganglia—supplies the sympathetic innervation of the head, neck, and upper limbs. The stellate ganglion is made up of a combination of the lower cervical and first thoracic ganglia. It lies on the prevertebral fascia of the seventh cervical and first thoracic vertebrae (**Fig. 5.1.1**). However, as the sixth cervical anterior tubercle (Chassaignac's tubercle) is easy to palpate, injection of a large volume of local anesthetic is made at this level and allowed to track caudally along the prevertebral fascia to block the stellate ganglion. The vertebral and carotid arteries, the pleura, and the brachial plexus are in close proximity to the stellate ganglion.

Equipment

- 10 ml syringe
- 22 G short-bevel needle
- Extension set (optional)
- ECG, BP, SpO_2 monitors
- Skin temperature monitor
- Resuscitation equipment (*see Appendix 3*)
- Ultrasound (optional)

Drugs

- Lidocaine (lignocaine) 1%, 10 ml
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Supine.
- Thin pillow under head.
- Roll under neck.
- Eyes directed at ceiling.
- Mouth slightly open.

Needle puncture and technique

Caution: Injection of 0.5–1 ml of lidocaine (lignocaine) 1% into the vertebral artery may result in immediate convulsion and/or loss of consciousness with possible cardiovascular system (CVS) collapse.

- Intravenous access is inserted.
- Monitors are attached, each temperature probe is attached to the palmar aspect of the middle finger of each hand (*Fig. 5.1.2*).
- Resuscitation equipment and drugs are checked and made ready for use.
- The side of the neck is cleaned with antiseptic solution.
- It is best to stand at the same side of the neck as the ganglion to be blocked.

FOR THE RIGHT-HANDED OPERATOR

With the left hand

- The thyroid cartilage is located and marked.
- The cricoid cartilage is identified and marked (*Figs 5.1.3, 5.1.4*).

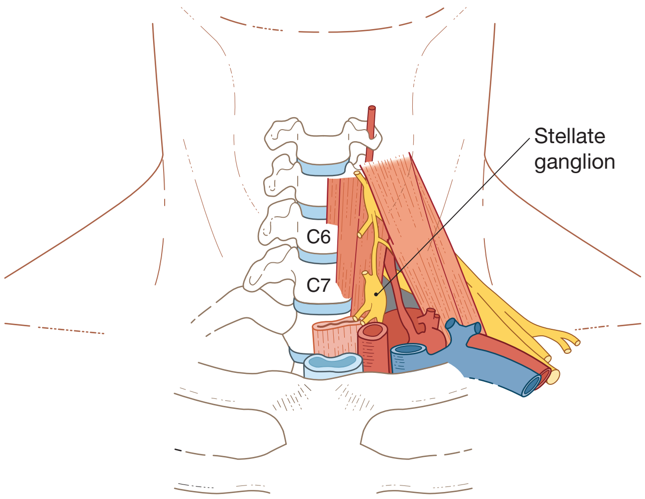

Fig. 5.1.1

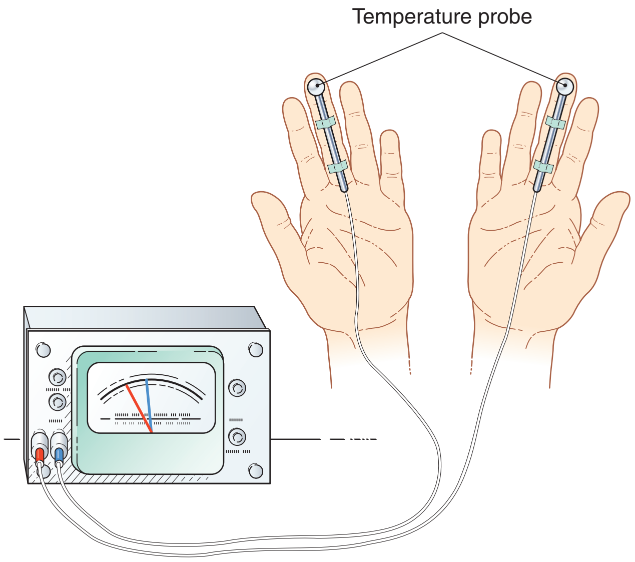

Fig. 5.1.2

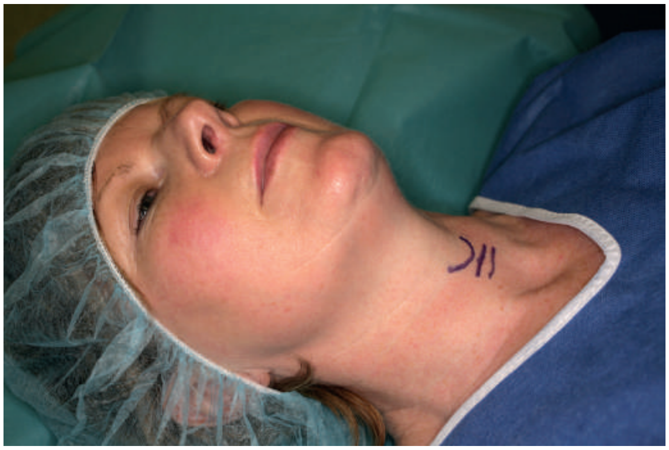

Fig. 5.1.3

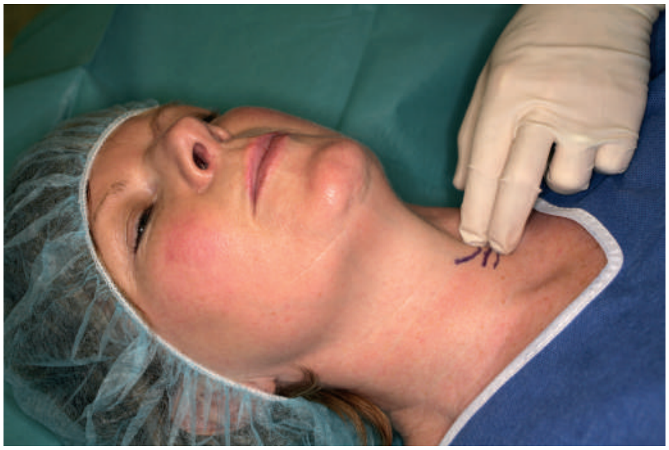

Fig. 5.1.6

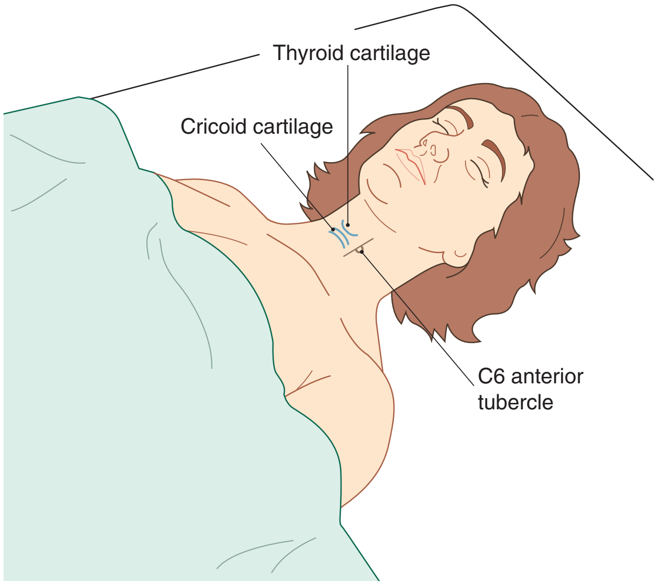

Fig. 5.1.4

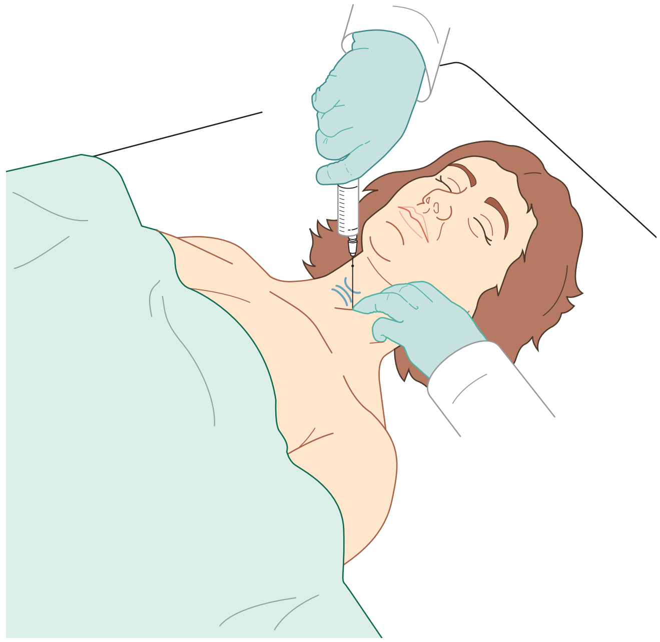

Fig. 5.1.7

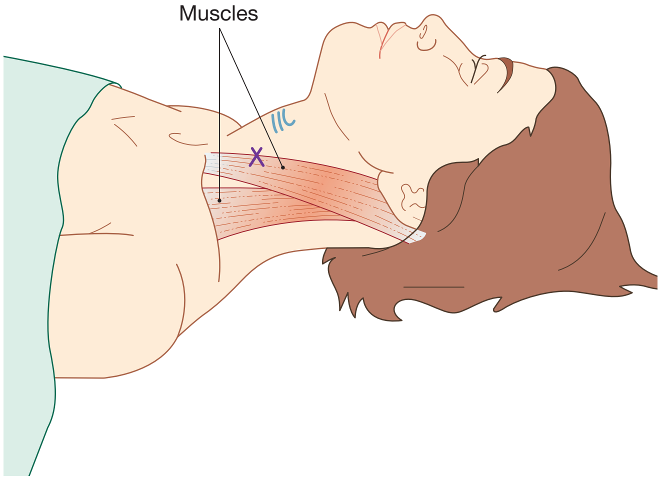

Fig. 5.1.5

- Chassaignac's tubercle is palpated with the middle finger, just lateral to the cricoid cartilage (*Fig. 5.1.5*).
- The sternocleidomastoid (SCM) muscle is gently pulled laterally and the carotid pulse is palpated (*Fig. 5.1.6*).
- Chassaignac's tubercle is palpated again and positioned between the fore- and middle fingers.

With the right hand

- The needle is inserted between the fore- and middle fingers of the other hand, directly perpendicular to the floor, aiming for Chassaignac's tubercle (*Figs 5.1.7, 5.1.8*).
- When contact with the tubercle is reached, the injecting hand is steadied and the needle is withdrawn 2 mm.
- The hub of the needle is held in place with the other hand.

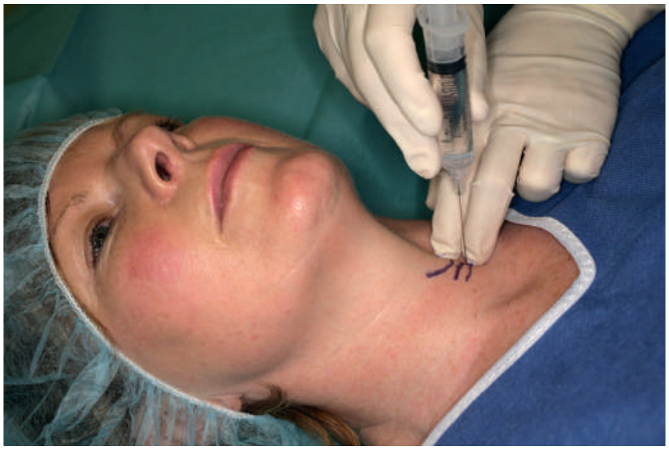

Fig. 5.1.8

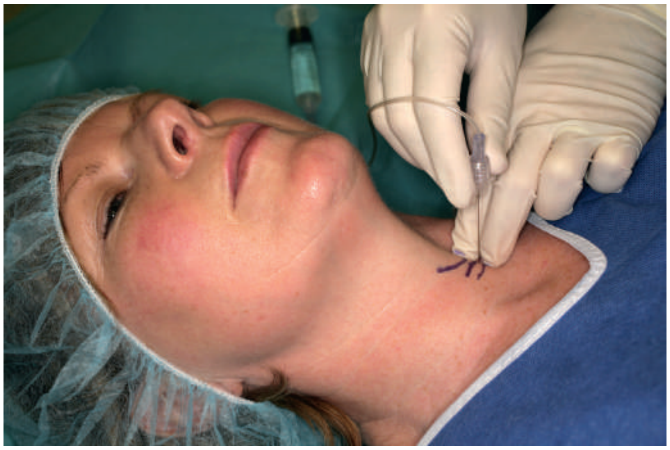

Fig. 5.1.9

- After negative aspiration, lidocaine (lignocaine) 1%, 0.5 ml, is injected.
- The patient is questioned about sensation and any change in level of consciousness is noted.
- If negative, the same procedure is repeated as 0.5 ml boluses are given until 10 ml is injected.
- The needle is withdrawn and the patient is immediately put in the sitting position.
- Monitors should be left attached and i.v. access left in situ for at least 30 minutes. The patient is requested not to eat or drink for 2 hours, as the recurrent laryngeal nerve may be blocked.
- Note: if aspiration of blood occurs during the block the needle is removed and cleared, keeping left fore- and middle fingers in place. It is then reinserted and the block is continued as above.
- If hematoma occurs before the solution is injected it may be worth performing the block at the C7 level.
- If there is pain on injection and/or paresthesia, it is likely that the brachial plexus may have been contacted, the needle is withdrawn and the landmarks are rechecked.

Confirmation of a successful block

- Skin temperature, measured over the palmar aspect of the hand or fingers on the blocked side, should begin to rise within 2–3 minutes. Extensive sympathetic blockade is confirmed by a rise in skin temperature to >33 °C.
- Ptosis of eyelid.
- Miosis of pupil.
- Unilateral blockage of nose on side of block.
- Unilateral flushing of conjunctiva of eye on side of block.
- Blockade of the upper sympathetic chain can occur in the absence of sympathetic denervation of the upper extremity, resulting in Horner's syndrome without a rise in skin temperature in the hand.

Tips

- The external jugular vein usually crosses the SCM muscle at the level of C6.
- Skin infiltration prior to block should be avoided if possible, as this will make landmarks more difficult to locate.
- If palpation is painful or difficult it may be helpful to try to bounce the middle finger off the tubercle during identification.
- An extension set may be inserted between the needle and syringe for better stability of needle (*Fig. 5.1.9*), but an assistant is then required to continue the procedure as described above.
- Consideration should be given to performance of the block under fluoroscopy or CT control if the landmarks are difficult to locate.
- Lidocaine (lignocaine) 1% 15 ml may be given if a previous block failed to relieve sympathetically maintained pain in the presence of correctly placed solution. This may improve tracking of the solution caudally to produce more effective blockade of the stellate ganglion.
- Ultrasound may aid placement of needle (*Fig. 5.1.10 a,b*).

Potential problems

- Intra-arterial injection or intrathecal injection may result in immediate convulsion and/or loss of consciousness with possible CVS collapse.

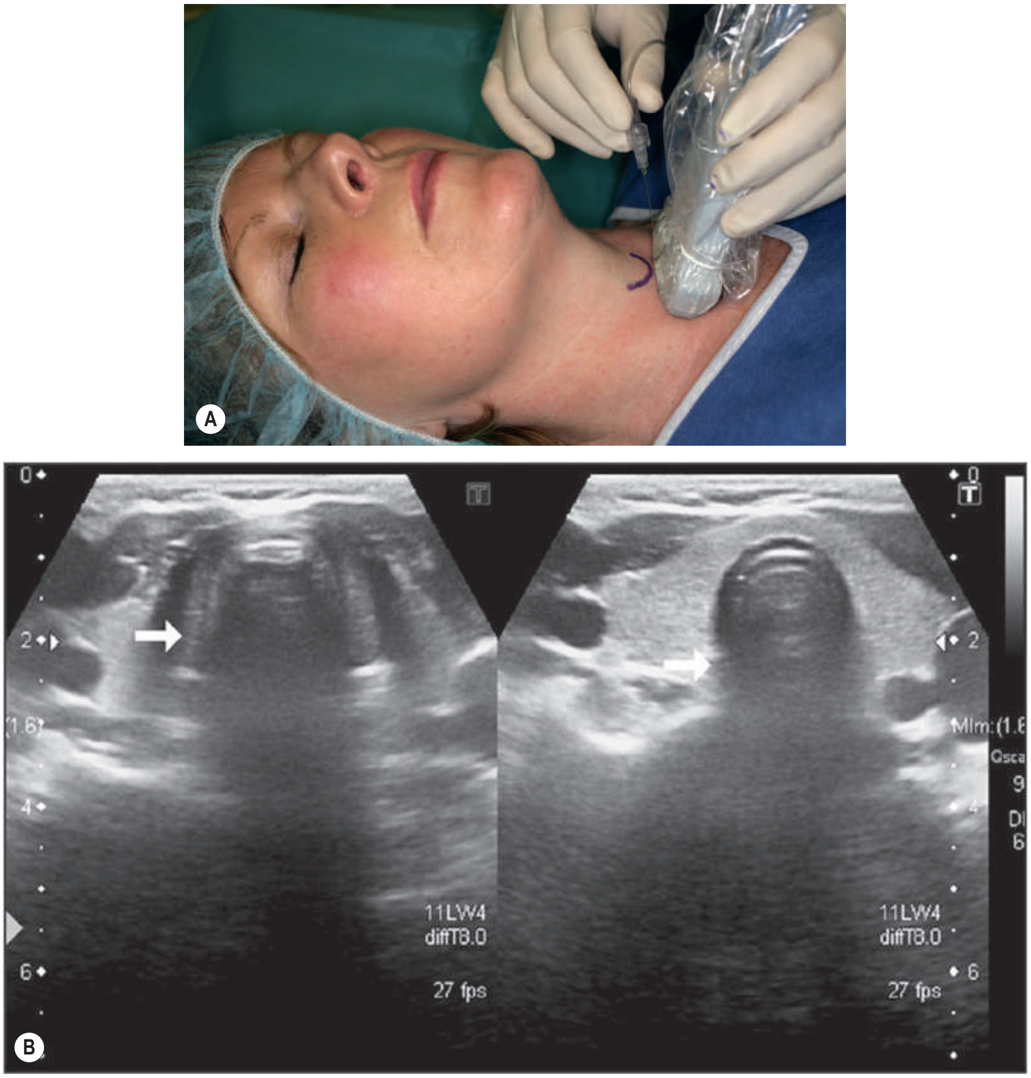

Fig. 5.1.10 From Gupta Prashant K, Gupta Kumkum, Dwivedi Amit Nandan D, Jain Manish. Potential role of ultrasound in anesthesia and intensive care, Anesthesia Essays and Research, 2011 Volume 5, Issue Number 1, Page: 11-19.

- Hematoma may occur (avoid performing block on patients who have coagulopathy).
- Pneumothorax may occur.
- Recurrent laryngeal nerve block may occur and it is prudent to advise the patient about possible hoarseness post-blockade. Bilateral stellate ganglion blockade should be avoided for the same reason.
- Phrenic nerve block may occur and it is prudent to caution the patient about possible shortness of breath post blockade of the stellate ganglion.

5.2 STELLATE GANGLION BLOCK—C7 APPROACH

Anatomy

As for stellate ganglion block (C6 approach) in Section 5.1 (**Figs 5.2.1 a,b**).

Equipment

- 10 ml syringe
- 22 G short-bevel needle
- Extension set (optional)
- ECG, BP, SpO_2 monitors
- Skin temperature monitor
- Fluoroscopy or CT (optional)

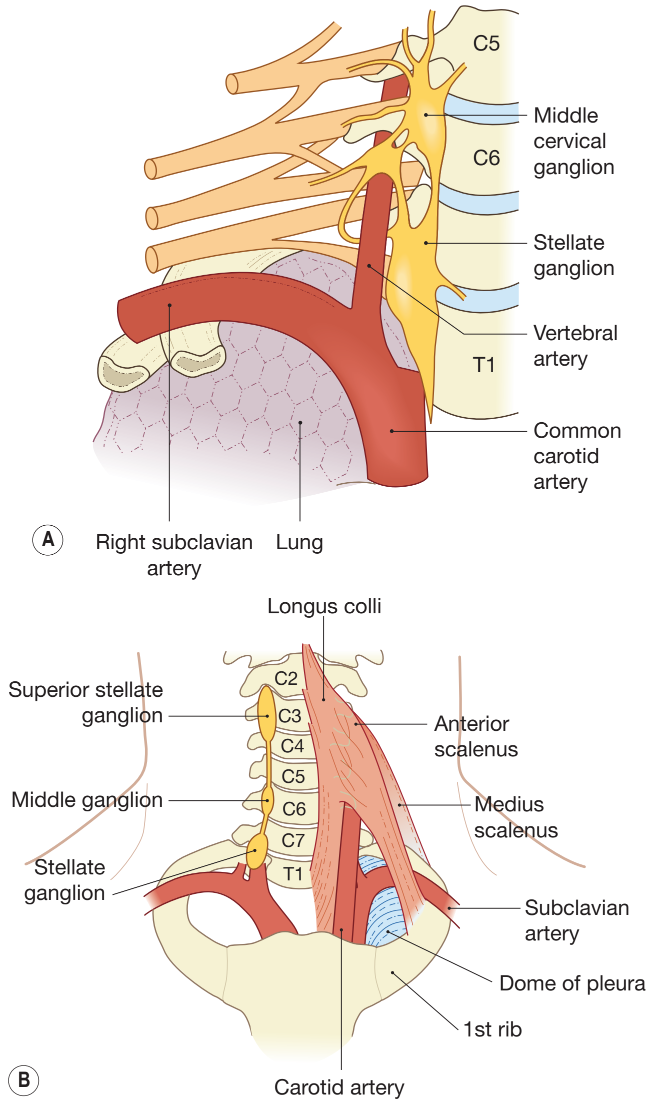
C5
Middle cervical ganglion
C6
Stellate ganglion
Vertebral artery
T1
Common carotid artery
A
Right subclavian artery
Lung
Longus colli
Superior stellate ganglion
C2
C3
C4
C5
C6
C7
T1
Anterior scalenus
Middle ganglion
Medius scalenus
Stellate ganglion
Subclavian artery
Dome of pleura
1st rib
B
Carotid artery

Fig. 5.2.1

- Resuscitation equipment (*see Appendix 3*)
- Ultrasound (optional)

Drugs

- Lidocaine (lignocaine) 1%, 10 ml
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Supine.
- Thin pillow under head.
- Roll under neck.
- Eyes directed at ceiling.
- Mouth slightly open.

Needle puncture and technique

Caution: injection of 0.5–1 ml of lidocaine (lignocaine) 1% into the vertebral artery may result in immediate convulsion and/or loss of consciousness with possible cardiovascular (CVS) collapse. The risk of pneumothorax is greater with this approach.

- Intravenous access is inserted.
- Monitors are attached.
- Each temperature probe is attached to the palmar aspect of the middle finger of each hand. (*Fig. 5.2.2*).
- Resuscitation equipment and drugs are checked and made ready for use.
- The side of the neck is cleaned with antiseptic solution.

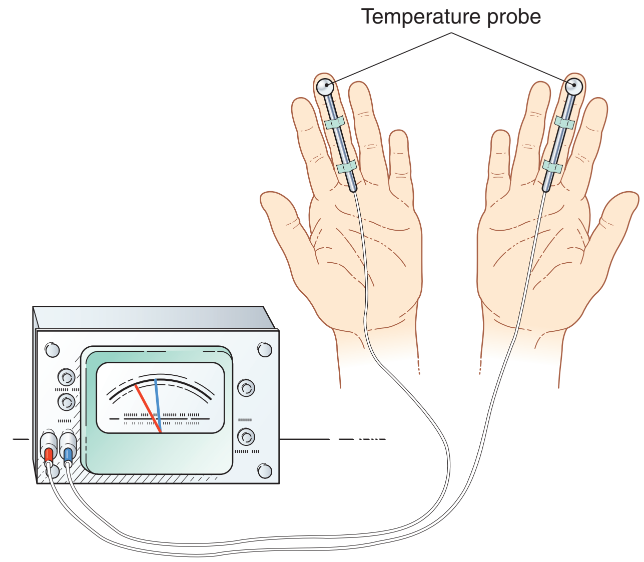
Temperature probe

Fig. 5.2.2

- It is best to stand at the same side of the neck as the ganglion to be blocked.

FOR THE RIGHT-HANDED OPERATOR

With the left hand

- The thyroid cartilage is located and marked.
- The cricoid cartilage is identified and marked (*Figs 5.2.3, 5.2.4*).
- The sternoclavicular junction is palpated and marked (*Fig. 5.2.5*).
- The SCM muscle is gently pulled laterally and the carotid pulse is palpated (*Fig. 5.2.6*).
- **The site of insertion of the needle lies 3 cm above the sternoclavicular junction or one to two finger-breadths below the level of the cricoid cartilage.**

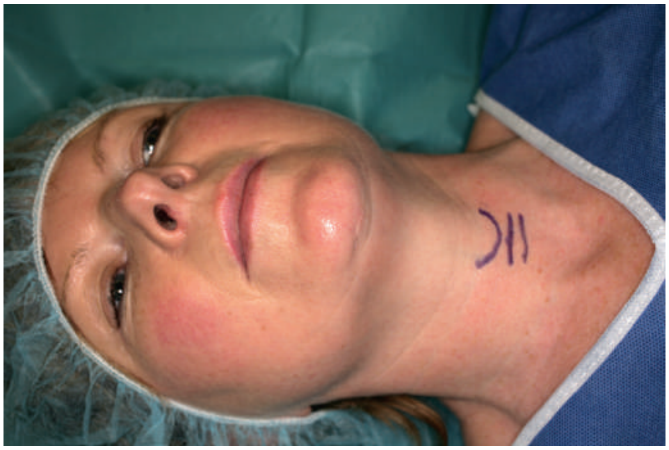

Fig. 5.2.3

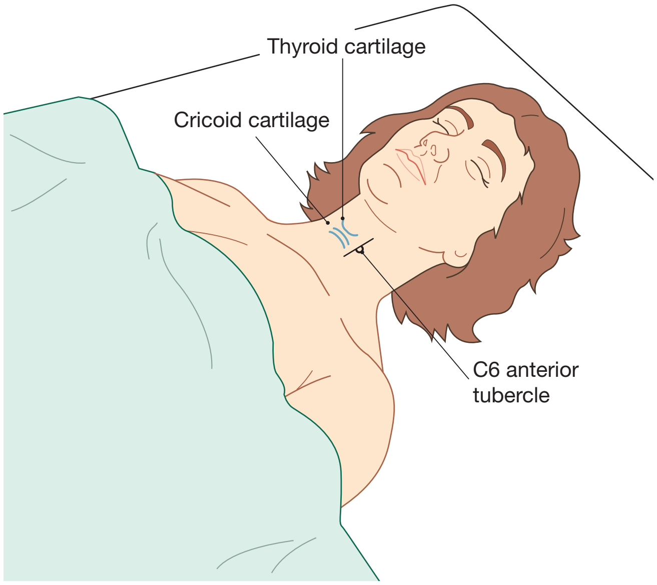

Fig. 5.2.4

With the right hand

- The patient is requested to exhale deeply before needle insertion to minimize the risk of pneumothorax.
- The needle is inserted between the fore- and middle fingers of the other hand, directly perpendicular to the floor.
- When contact with the transverse process of C7 is reached, the injecting hand is steadied and the needle is withdrawn 2 mm.
- The hub of the needle is held in place with the other hand.
- After negative aspiration, lidocaine (lignocaine) 1%, 0.5 ml, is injected.
- The patient is questioned about sensation, and any change in level of consciousness is noted.
- If negative, the same procedure is repeated and 0.5 ml boluses are given until 5–8 ml is injected.
- The needle is withdrawn and the patient is immediately put in the sitting position.

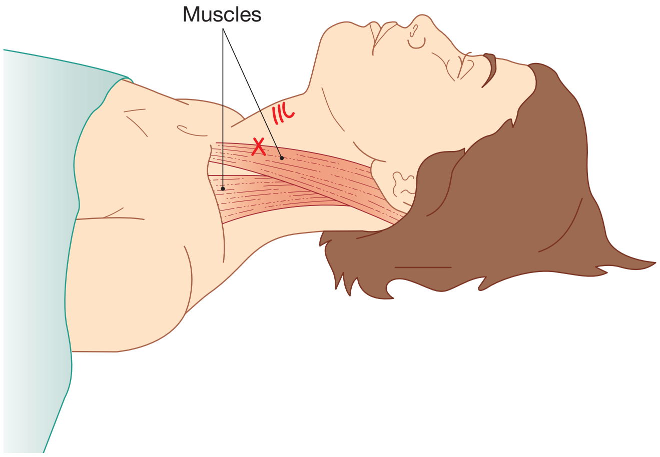

Fig. 5.2.5

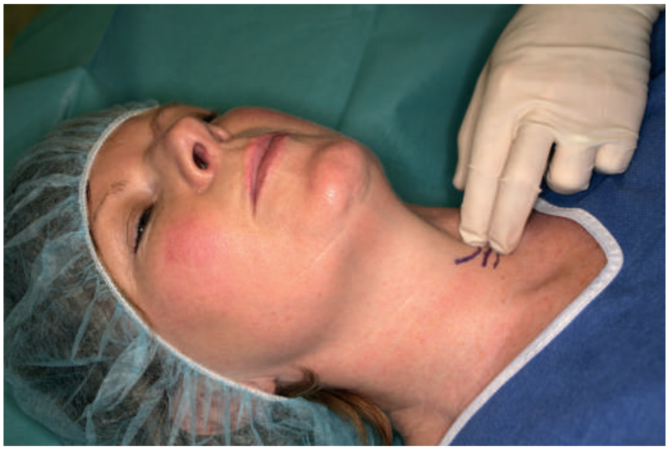

Fig. 5.2.6

- Monitors should be left attached and i.v. access left in situ for at least 30 minutes. Blockade of the recurrent laryngeal nerve is less likely with the C7 approach but it is wise to advise the patient not to eat or drink for 2 hours.
- Note: if aspiration of blood occurs during the block, the needle is removed and cleared, keeping the left fore- and middle fingers in place. It is then reinserted and the block is continued as above.
- If there is pain on injection and/or paresthesia, it is likely that the brachial plexus may have been contacted, the needle is withdrawn and the landmarks are rechecked.
- Ultrasound may aid placement of the needle.

Confirmation of a successful block

- Temperature increase >1° on the side of block. The temperature should begin to rise in the finger of the blocked side within 3 minutes of injection.
- Ptosis of eyelid.
- Miosis of pupil.
- Unilateral blockage of nose on side of block.
- Unilateral flushing of conjunctiva of eye on side of block.
- Relief of sympathetically maintained pain.

Tips

- Some workers advocate targeting the **ventrolateral aspect of the C7 vertebral body instead of its transverse process.** The needle is directed 15–20° medially. With the aid of fluoroscopy, ultrasound or CT, the vertebral body is contacted just medial to the insertion of the longus colli muscle. The needle is then withdrawn 2 mm, stabilized, and 1 ml of non-ionic contrast medium is injected.
- The external jugular vein usually crosses the SCM muscle at the level of C6.
- Skin infiltration prior to block should be avoided if possible, as this will make landmarks more difficult to locate.
- An extension set may be inserted between the needle and syringe for better stability of needle, but an assistant is then required to continue the procedure as described above.
- Lidocaine (lignocaine) 1% 10 ml may be given if a previous block failed to relieve sympathetically maintained pain in the presence of correctly placed solution. This may improve tracking of the solution caudally to produce more effective blockade of the stellate ganglion.

Potential problems

- Intra-arterial injection or intrathecal injection may result in immediate convulsion and/or loss of consciousness with possible CVS collapse.
- Hematoma may occur (avoid performing block on patients who have coagulopathy).
- Pneumothorax may occur (more likely with C7 approach).
- Recurrent laryngeal nerve block may occur and it is prudent to advise the patient about possible hoarseness post blockade. Bilateral stellate ganglion blockade should be avoided for the same reason.
- Phrenic nerve block may occur and it is prudent to caution the patient about possible shortness of breath following blockade of the stellate ganglion.

5.3 LUMBAR SYMPATHETIC BLOCK

Anatomy

The lumbar sympathetic chain is located in the prevertebral fascia, which lies on the anterolateral aspects of the vertebral bodies. The psoas muscle separates the lumbar sympathetic chain from the lumbar somatic nerves. A single injection of local anesthetic at the level of L2 will usually provide a complete block of postganglionic sympathetic efferents to the lower extremity because the lowest preganglionic sympathetic outflow to the chain is at the level of L2.

Equipment

- 2 ml, 5 ml, and 10 ml syringes
- 30 G needle
- Two 15 cm 22 G needles
- Extension set (optional)
- ECG, BP, SpO_2 monitors
- Skin temperature monitor (two probes)
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy

Drugs

- Lidocaine (lignocaine) 1%, 5 ml for skin infiltration
- Lidocaine (lignocaine) 1%, 15–20 ml (or its equivalent) for block
- Phenol 6%
- Radio-opaque contrast medium
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Pillow under anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 5.3.1*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Each temperature probe is attached to the plantar aspect of the big toe (*Fig. 5.3.2*).
- Resuscitation equipment and drugs are checked and made ready for use.
- The thoracolumbar midline and an area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- The twelfth rib is identified and a line is drawn along its inferior border (*Fig. 5.3.3 a,b*).
- The iliac crests are palpated and the intercrestal line is identified (this corresponds with the inferior aspect of the spinous process of L4 or may lie in the L4–5 interspace).
- The spinous processes are counted until L2 is identified and confirmed with fluoroscopy.
- **The insertion point lies 8 cm lateral to the L2 spinous process and is also marked (*Fig. 5.3.4*).**
- A skin wheal is raised at one of the marked sites and the area is infiltrated with lidocaine (lignocaine) 1%.
- At a 30° angle to the frontal plane, a 22 G 15 cm needle is advanced slightly cephalad towards the lower

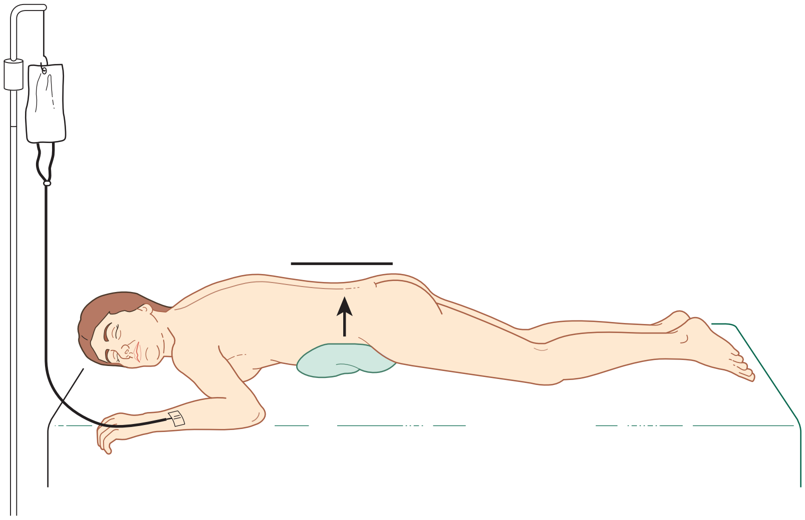

Fig. 5.3.1

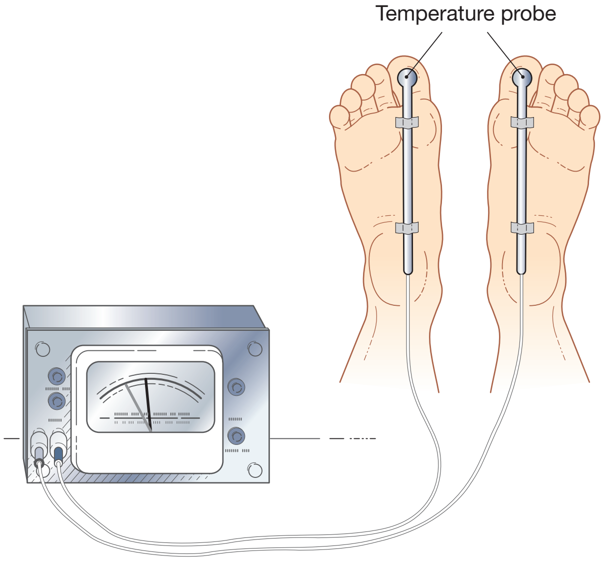

Fig. 5.3.2

portion of the L2 vertebral body (*Fig. 5.3.5*), until its vertebral body is contacted at a depth of about 7–9 cm and confirmed with fluoroscopy. If the needle contacts bone at a more superficial level, it is probable that it has come into contact with the transverse process and it will need to be repositioned.

- The needle depth is noted.
- The needle is then withdrawn to the subcutaneous tissue and, with the aid of fluoroscopy, it is re-advanced, this time at an angle 45° to the frontal plane, until the previous depth (as noted) is reached. It should slip past the vertebral body at a depth about 1–2 cm deeper than the first depth mark (*Fig. 5.3.6*).
- After negative aspiration, the fluoroscopic image is observed as a small amount of non-ionic radio-contrast medium is injected. The correct placement of the needle is indicated by the presence of a layer of contrast medium in a thin line along the anterior border of the vertebral column (*Figs 5.3.7–5.3.9*).
- After further negative aspirations, 5 ml of lidocaine (lignocaine) 1% is injected. The patient is questioned about pain relief and observations of skin temperature are made. There should be little resistance to injection, similar to resistance felt when injecting through an epidural needle. If resistance is encountered, or if the injection is painful, the needle should be repositioned. A unilateral rise in skin temperature indicates a successful block.
- After 10 minutes the patient is questioned about pain relief and any symptoms of somatic nerve blockade. Sensory and motor functioning of the lower extremities is checked. The procedure should be abandoned if there is evidence of somatic blockade.

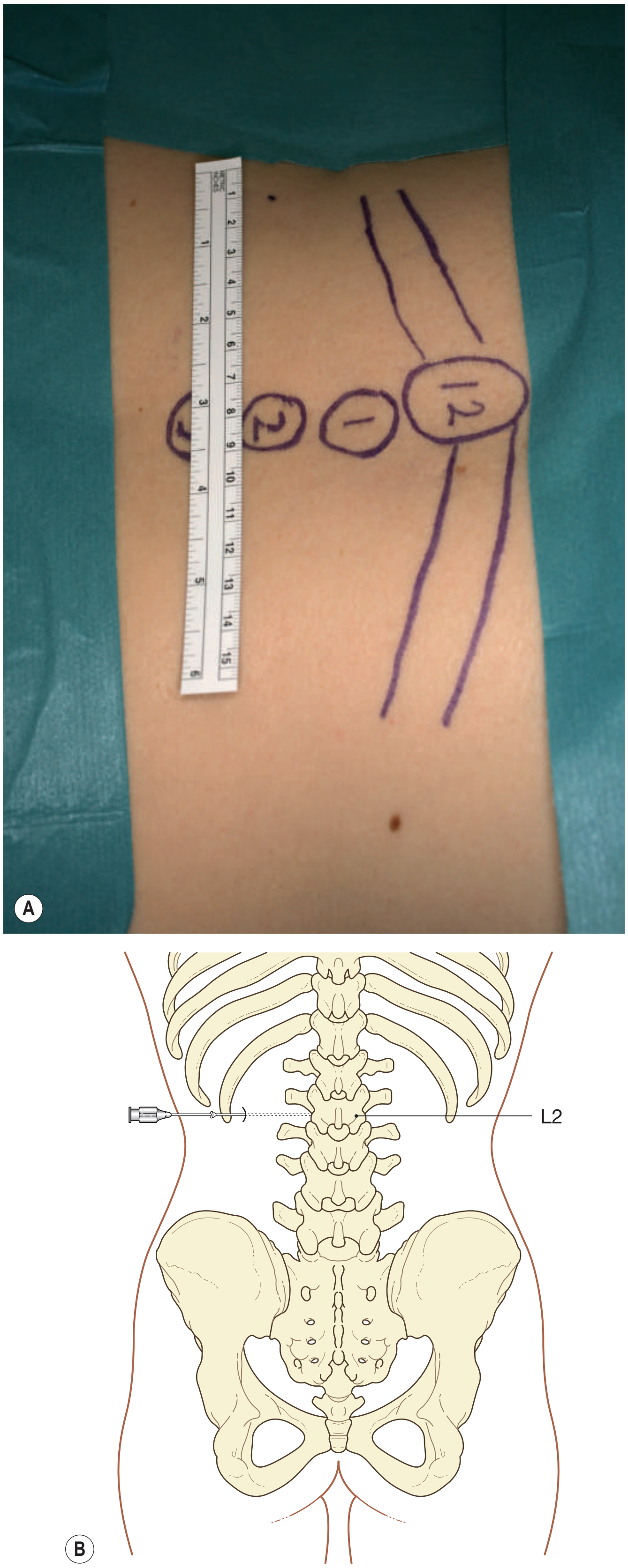

Fig. 5.3.3

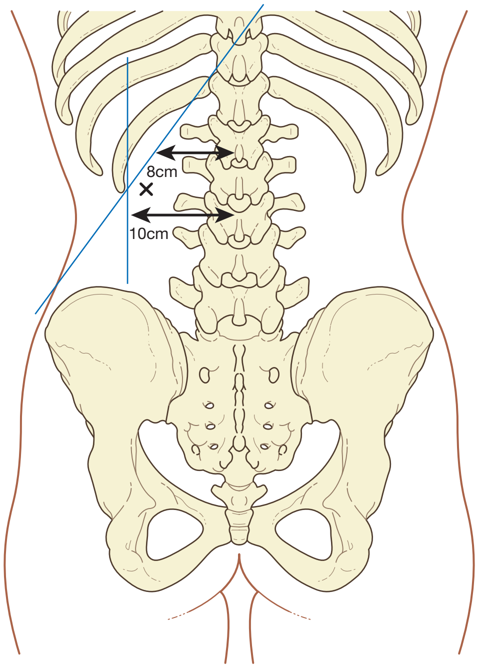

Fig. 5.3.4

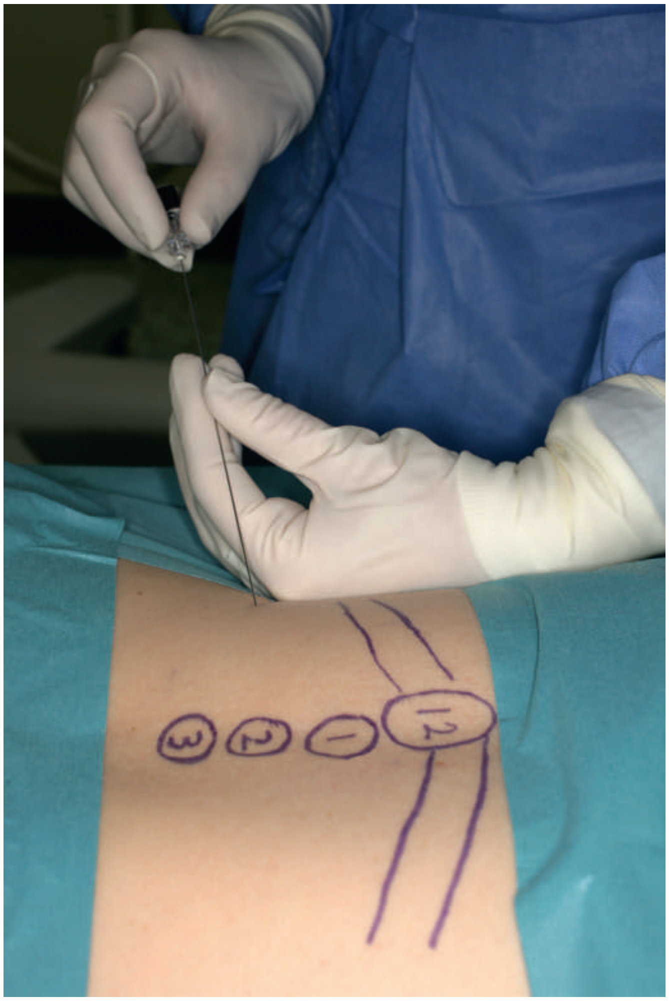

Fig. 5.3.5

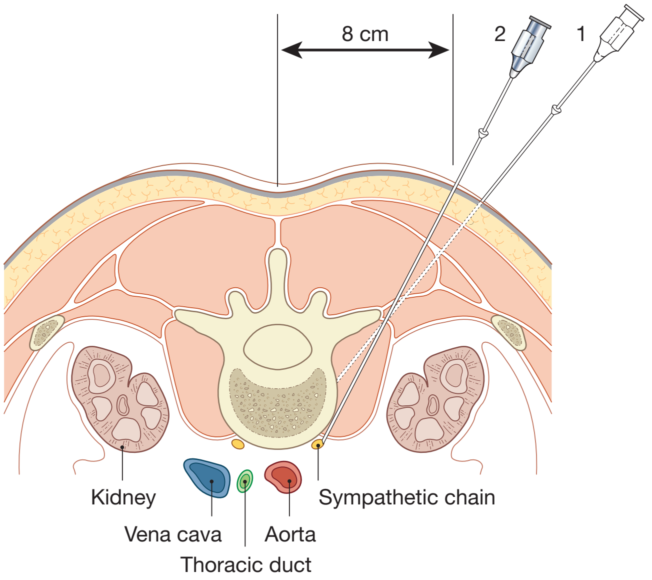

Fig. 5.3.6

- Neurolysis may be achieved by leaving the needle in place after block has been confirmed and injecting 5 ml phenol 6%. To avoid leaving alcohol in the needle tract, the needle is then cleared with air or local anesthetic 1 ml, and removed. However, because of the very high incidence of genitofemoral neuralgia that can occur post-neurolytic lumbar sympathetic block, the benefit versus risk should be considered carefully.
- Monitors should be left attached and i.v. access left in situ for at least 30 minutes.

Confirmation of a successful block

- Increase in skin temperature on the plantar surface of the foot to about 35 °C; temperature should begin to rise in the foot on the side of the block within 3 minutes of injection of local anesthetic.
- Relief of sympathetically maintained pain in the lower limb.

Tips

- If fluoroscopy is not available ultrasound may aid placement of the needle. A line 10 cm from the midline is drawn parallel to the midline; the lowest rib is identified and a line is drawn along its inferior border.
- The point of intersection of these lines should be lateral to the L2 vertebral body.
- Consideration should be given to performance of the block under CT control if the block is unsuccessful.
- Repeated blocks may bring about gradual improvement in sympathetically maintained pain.
- Immediate physiotherapy after blockade may improve the outcome.

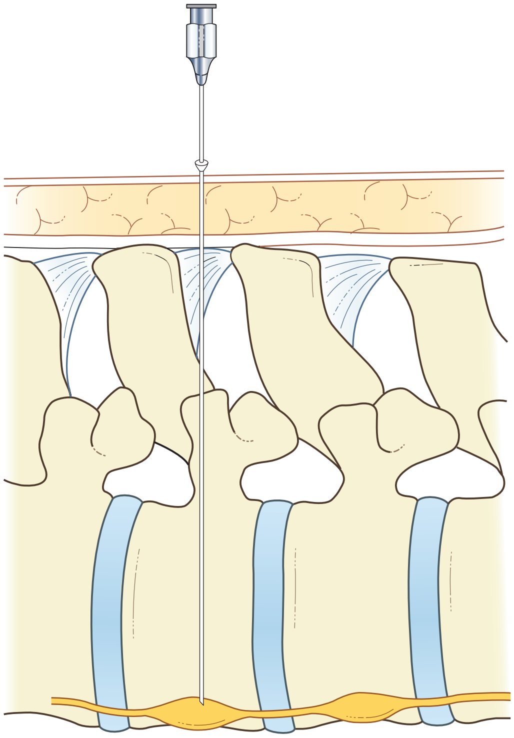

Fig. 5.3.7

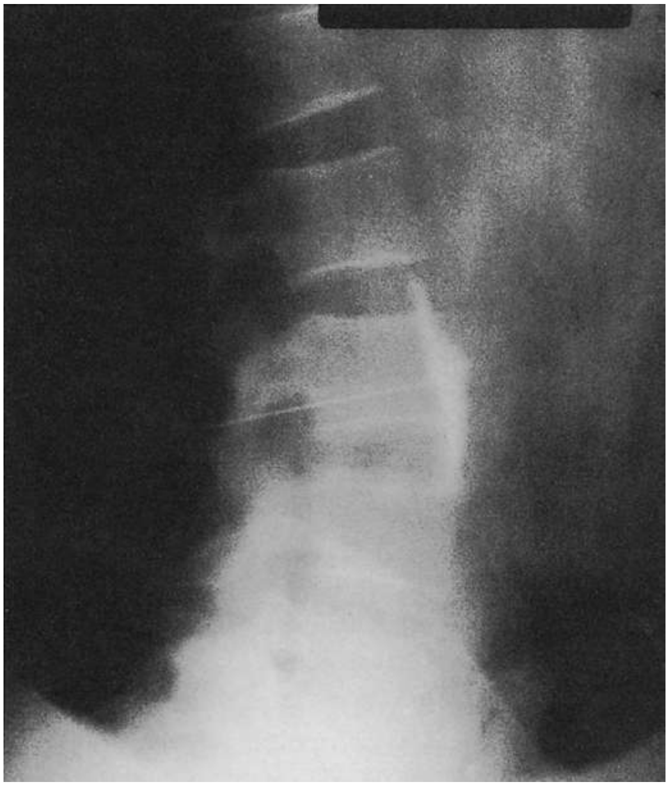

Fig. 5.3.9

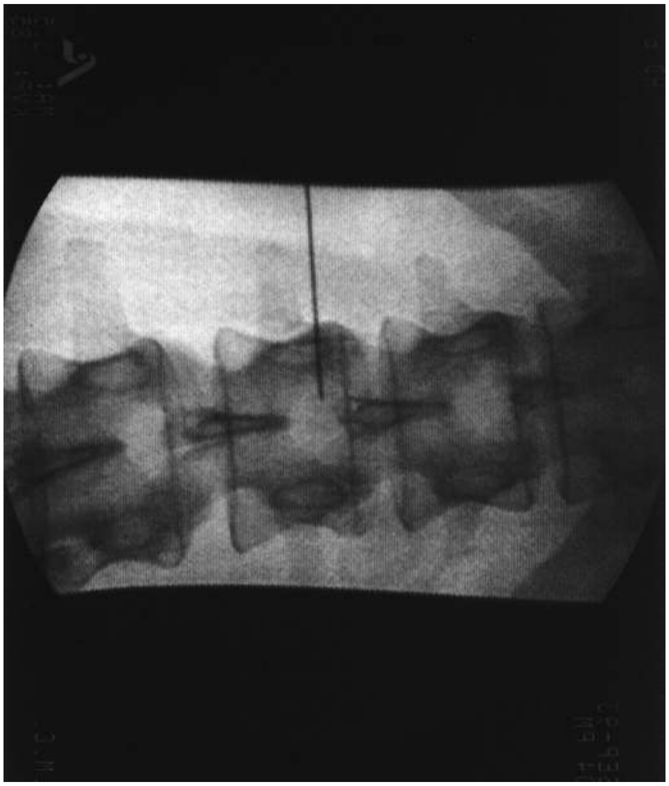

Fig. 5.3.8

Potential problems

- If the needle tip is placed too superficially, the tip may come to lie in the intervertebral foramen and injection may result in a subarachnoid block, an epidural block, or a somatic nerve block. Confirmation of needle position using lateral fluoroscopy is therefore recommended.
- Genitofemoral neuralgia may occur in 5–10% of patients post-neurolytic block causing pain in the groin.
- Perforation of the aorta or the inferior vena cava is possible and retroperitoneal hematoma may occur. Consequently the block should be avoided in patients with coagulopathy.
- Intravascular injection may occur.
- Perforation of the kidney or ureter is usually of no clinical significance unless neurolytic agents are used.
- Perforation of the intervertebral disc may occur. This also is usually of no clinical significance but may produce a septic discitis if bacterial contamination occurs.
- Postural hypotension, secondary to sympathetic blockade, may occur.
- Injection of neurolytic solution into the psoas muscle may cause rhabdomyolysis.
- Patients in the prone position should be monitored carefully when intravenous sedation is administered.

5.4 CELIAC PLEXUS BLOCK—RETROCRURAL APPROACH

Anatomy

The celiac plexus is flat and lies against the crus of the diaphragm, surrounding the root of the celiac and mesenteric arteries and anterior to three vertebral bodies centered at L1. Posteriorly on the left side is the aorta, and on the right is the inferior vena cava. The kidneys lie lateral and the pancreas anterior to the celiac plexus (**Fig. 5.4.1 a,b**).

The celiac plexus is made up of pre- and postganglionic sympathetic and parasympathetic nerve fibers. Postganglionic sympathetic fibers are supplied from the paired celiac ganglia. Preganglionic sympathetic efferents from the thoracic sympathetic chain are supplied via the greater and lesser splanchnic nerves. The intra-abdominal viscera are supplied by postganglionic sympathetic fibers that have synapsed in the celiac ganglia (**Fig. 5.4.2**). The vagus nerve also supplies parasympathetic nerve fibers. Via the celiac plexus dorsal root, ganglion cells innervate the whole of the abdominal viscera, including the liver, spleen, kidneys, suprarenal glands, and intestines, with the exception of the pelvic organs, the rectum, and the left half of the colon.

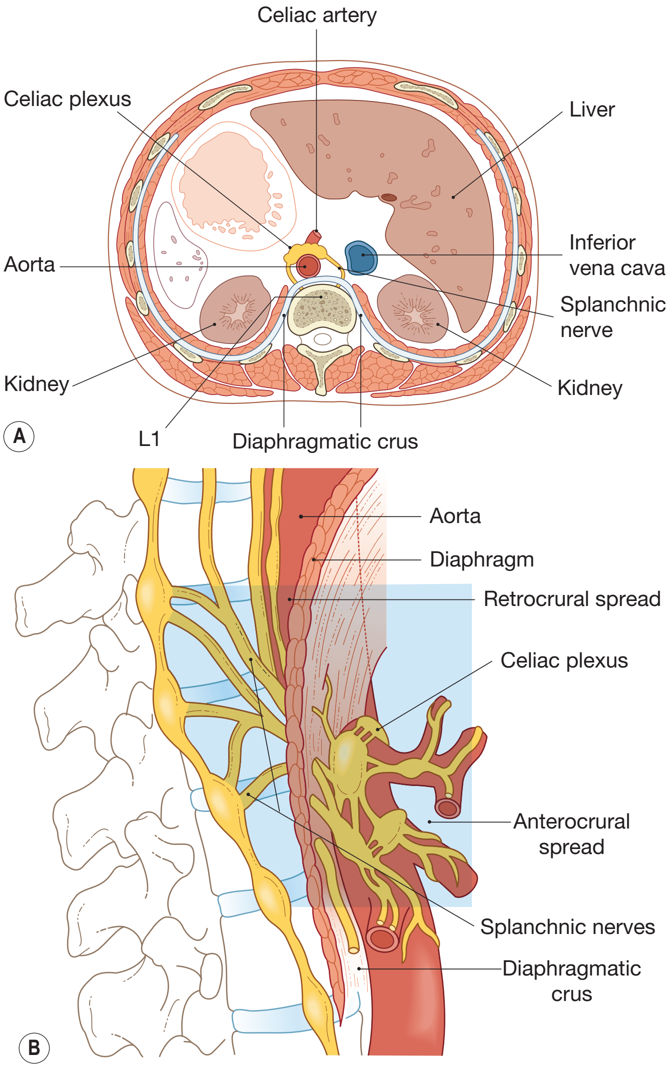
Celiac artery
Celiac plexus
Liver
Aorta
Inferior vena cava
Splanchnic nerve
Kidney
Kidney
A
L1
Diaphragmatic crus
Aorta
Diaphragm
Retrocrural spread
Celiac plexus
Anterocrural spread
Splanchnic nerves
Diaphragmatic crus
B

Fig. 5.4.1

Pain originating from the viscera is often vague and poorly localized as a result of convergence of neurons in the dorsal horn and crossing over the midline of some of the visceral afferents.

There are two main approaches to celiac plexus blockade. One approach places the two needles posterior to the crura of the diaphragm, the retrocrural approach. The retrocrural approach to the celiac plexus also targets the splanchnic nerves to produce a splanchnic nerve block if required. The other approach places a needle anterior to the crus of the diaphragm on the right, the anterocrural approach (Fig. 5.4.3), as discussed in Section 5.5.

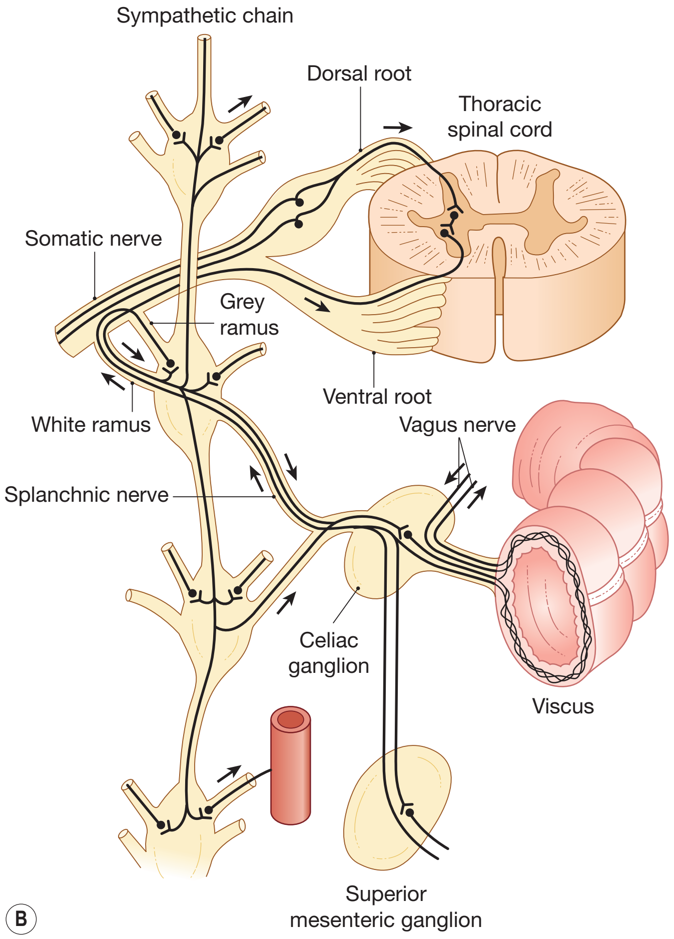
Sympathetic chain
Dorsal root
Thoracic spinal cord
Somatic nerve
Grey ramus
Ventral root
White ramus
Vagus nerve
Splanchnic nerve
Celiac ganglion
Viscus
Superior mesenteric ganglion
B

Fig. 5.4.2

Equipment

- 2 ml, 5 ml, and 10 ml syringes
- 30 G needle
- Two 15 cm 22 G needles
- Extension set (optional)
- ECG, BP, SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy

Drugs

- Mild sedative
- Lidocaine (lignocaine) 1%, 5 ml for skin infiltration
- Lidocaine (lignocaine) 1%, 15–20 ml (or its equivalent) for block
- 6% aqueous phenol or 50–75% alcohol (ethanol). Our suggestion: mix 2 parts absolute alcohol with one part 1% lidocaine. This will help reduce the incidence and severity of pain following injection. In addition, precede all alcohol injections with 3–4 ml 1% lidocaine
- Non-ionic radio-opaque contrast medium
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Pillow under anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 5.4.4*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- Mild sedation may be induced.

RETROCRURAL APPROACH

- A 15 cm 22 G spinal needle is selected. A slight curve at the needle tip, away from the bevel direction, may be created, which allows the needle to be redirected during placement.
- An AP view of the upper lumbar/low thoracic spine is obtained and the C-arm is adjusted to superimpose the T12–L1 endplates.
- **A skin wheal is raised at the lower border of the twelfth rib on the right just above the level of the L1 transverse process** (*Fig. 5.4.5 a*). The needle is inserted at this site and advanced at an angle 30° from perpendicular inward until the L1 body is contacted just below the upper endplate (*Fig. 5.4.5*). The curve of the needle is turned laterally and the needle is advanced along the upper portion of the body. Once the needle has slipped a few millimeters past the lateral aspect of the L1 body, a lateral view is obtained. The curve of the needle is directed inward toward the body and advanced until the tip lies at the anterior border of the body, near the upper endplate, in a direct lateral view.
- Using "live" fluoroscopy, 1 ml non-ionic contrast medium is injected. The dye should remain against the anterior aspect of the bodies in the lateral view (*Fig. 5.4.6 a*).
- If dye is seen spreading dorsally toward the neural foramina (*see Fig. 5.4.6 b*), the needle should be withdrawn and repositioned at a higher level.
- An AP view is obtained, which should demonstrate dye spread against the lateral aspect of the bodies (*Fig. 5.4.6 c*).
- Spread more laterally indicates injection within the psoas muscle, in which case the needle should be repositioned more medially and anteriorly. 3 ml 1% lidocaine is then injected. The dye shadow will

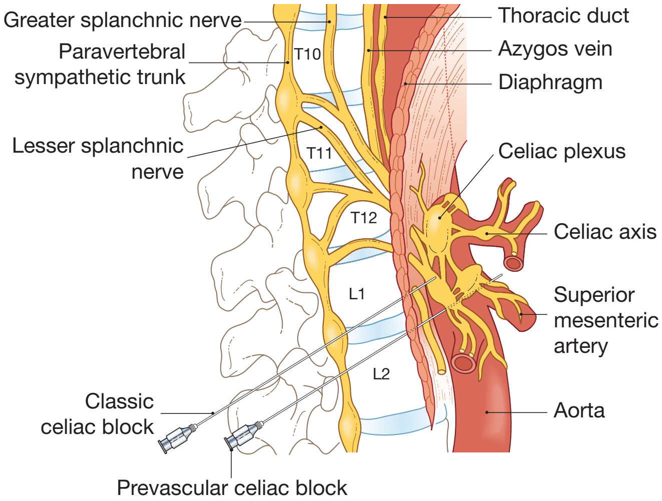

Fig. 5.4.3

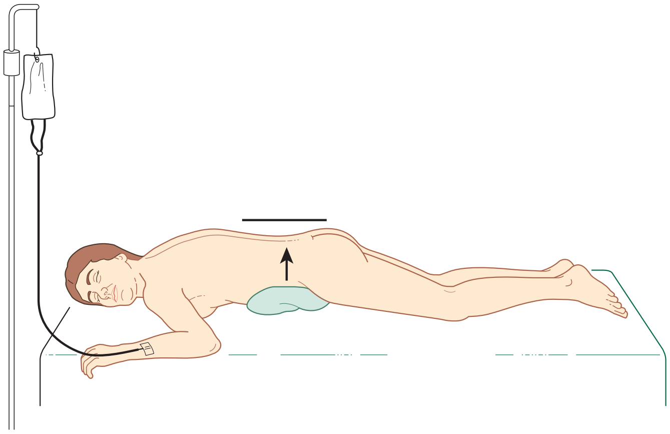

Fig. 5.4.4

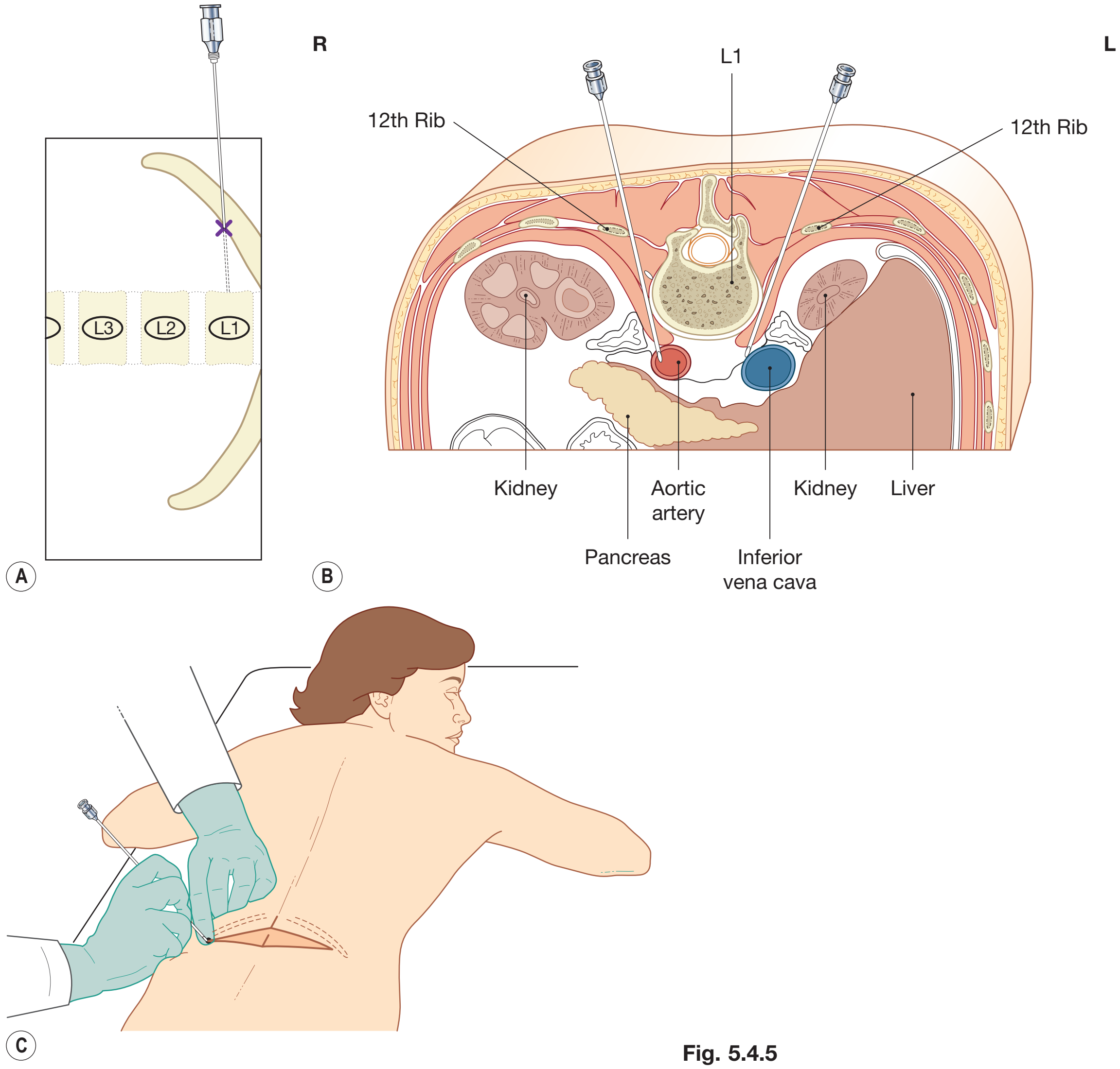
R
L
L1
12th Rib
12th Rib
L3
L2
L1
Kidney
Aortic
artery
Kidney
Liver
Pancreas
Inferior
vena cava
A
B
C

Fig. 5.4.5

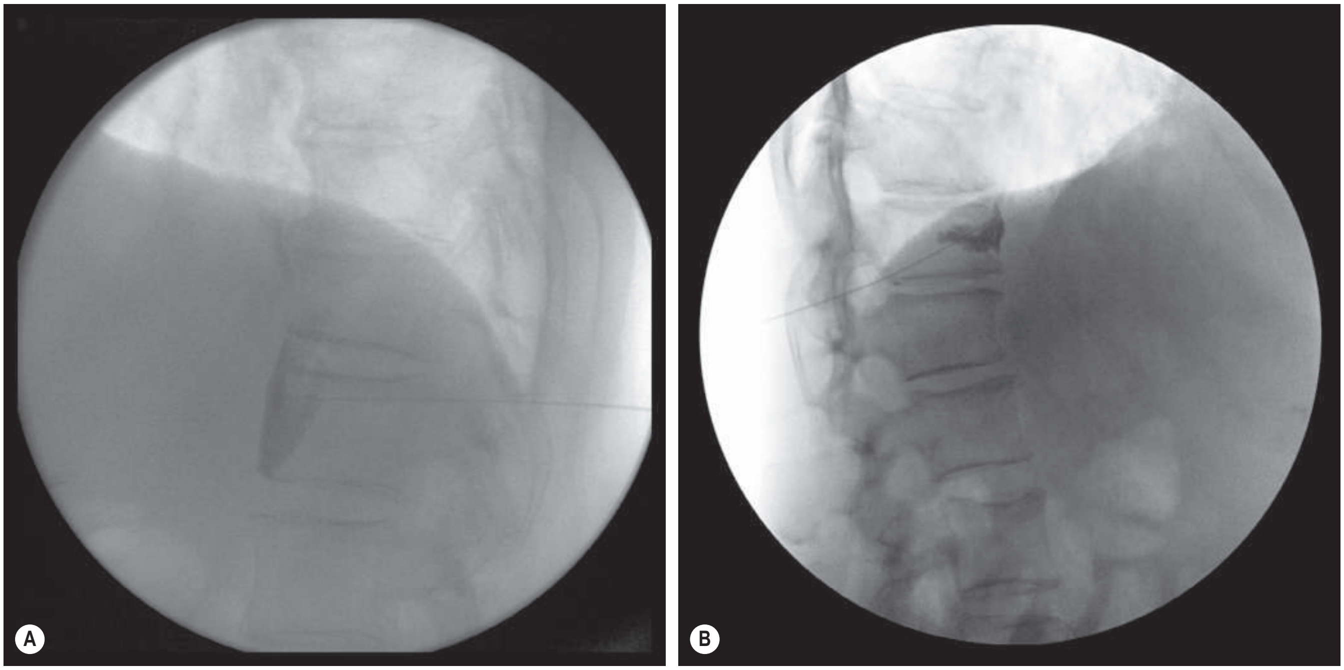

Fig. 5.4.6

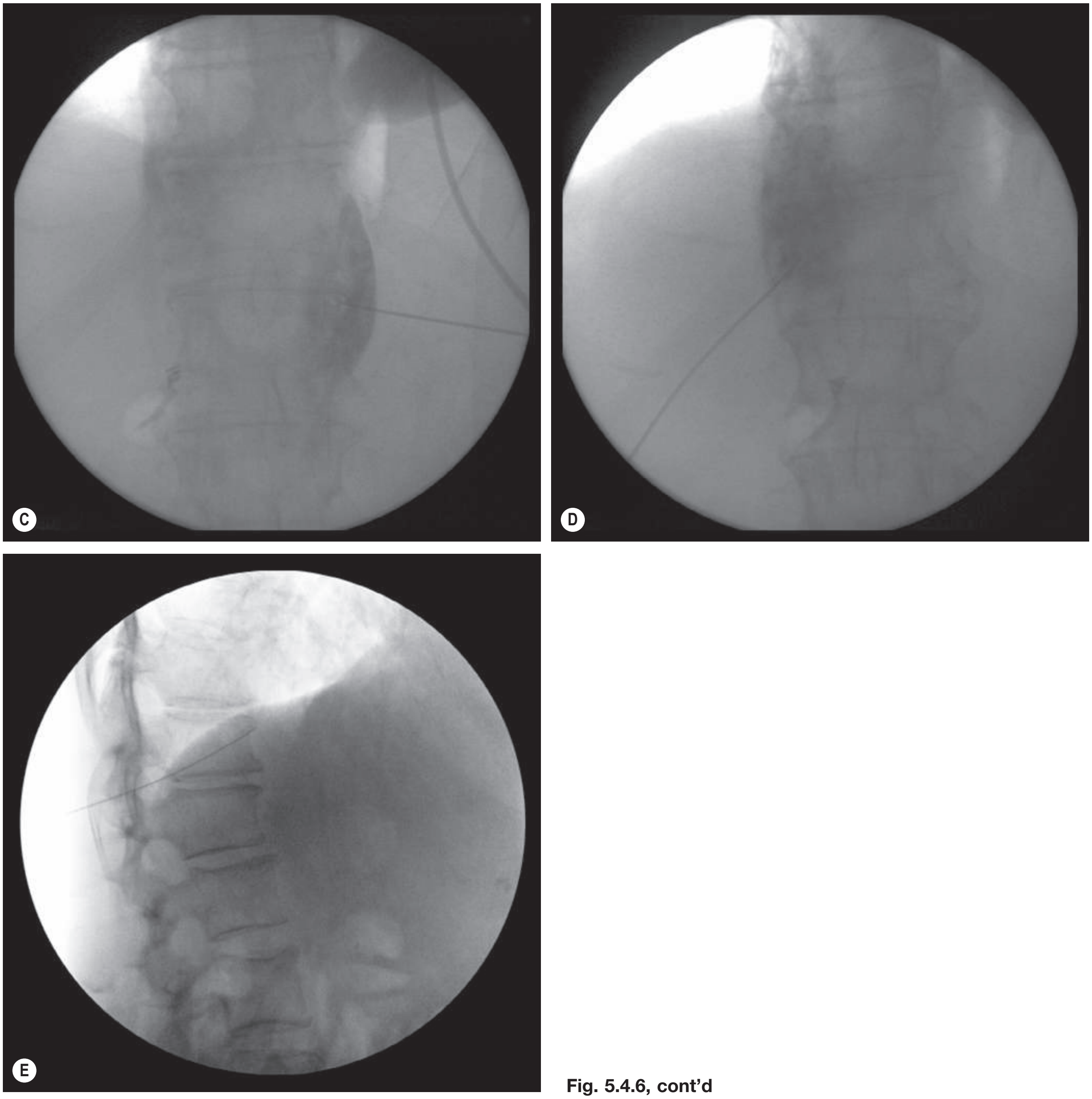

Fig. 5.4.6, cont'd

be seen to expand superiorly, spreading to the thoracic levels to contact the splanchnic nerves (*see Fig. 5.4.6 d*).

- After confirming negative aspiration for blood, 15–20 ml alcohol or phenol is injected. The needle is cleared with 1 ml lidocaine prior to removal.
- Alternatively, the needle can be advanced more cephalad to a position at the anterior border of T12 preferably near either the lower or upper endplate (*Fig. 5.4.6 e*).
- Injection near the mid-point of the body is more likely to result in dorsal spread of the drug toward the neural foramen. More cephalad placement is a bit more difficult technically, but places the needle closer to the splanchnic nerves.
- The procedure is repeated in an identical manner on the left side.

Confirmation of a successful block

- Relief of upper abdominal pain.

Tips

- After injection of non-ionic radio-contrast medium, a blush will indicate injection into muscle. If visible contrast medium disappears immediately it is likely that intravascular injection has occurred.
- Consideration should be given to performance of the block under CT control if the block is unsuccessful.
- Placement of the needle anterior to the diaphragmatic crus can also be achieved via insertion through the abdominal wall.

Potential problems

- The position of each needle tip should always be confirmed with fluoroscopy before injection of neurolytic agent as it may lie in the peritoneal cavity, within a viscus or intravascularly. If a needle tip is placed too superficially, the tip may come to lie in the intervertebral foramen and injection may result in an epidural block or a somatic nerve block. Injection of neurolytic solution into the psoas muscle may cause rhabdomyolysis.
- Perforation of the aorta or the inferior vena cava is possible and consequently the block should be avoided in patients with coagulopathy. Dissection of the aorta may occur as a result of direct damage during the block. Retroperitoneal hematoma may occur and for this reason also the block should be avoided in patients with coagulopathy.
- Orthostatic hypotension may occur as a result of sympathetic blockade for up to 3 days after a neurolytic block. Diarrhea may occur also and hydration of the patient should be monitored.
- Pneumothorax may occur.
- Transient motor paralysis and paraplegia may occur after the block, probably as a result of spasm of the segmental arteries.
- Perforation of the intervertebral disc may occur, but this also is usually of no clinical significance.
- Perforation of the kidney or ureter is usually of no clinical significance unless neurolytic agent is injected.
- The thoracic duct may be damaged (possibly causing chylothorax, or lymphedema).
- Abdominal and chest discomfort may be experienced for 30 minutes after injection of alcohol.
- There may be a detectable odor from the breath after alcohol injection.
- Patients in the prone position should be monitored carefully when intravenous sedation is administered.

5.5 CELIAC PLEXUS BLOCK—ANTEROCRURAL APPROACH

Anatomy

The anterocrural approach places a needle anterior to each crus of the diaphragm. The needles are inserted more medially and directed at a larger angle towards the midline until they come to lie in the retroperitoneal compartment between the aorta, and the inferior vena cava dorsally, and the pancreas ventrally (**Fig. 5.5.1**). Fluoroscopic imaging is necessary for accurate placement of the anterocrural needles using this approach.

Equipment

- 2 ml, 5 ml, and 10 ml syringes
- 30 G needle
- 15 cm 22 G needle (penetration of the diaphragmatic crus is easier with a large gauge needle)
- Extension set (optional)
- ECG, BP, SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy

Drugs

- Mild sedative
- Lidocaine (lignocaine) 1%, 5 ml for skin infiltration
- Lidocaine (lignocaine) 1%, 15–20 ml (or its equivalent) for block
- 6% aqueous phenol or 50–75% alcohol (ethanol). Our suggestion: mix 2 parts absolute alcohol with one part 1% lidocaine. This will help reduce the incidence and severity of pain following injection. In addition, precede all alcohol injections with 3–4 ml 1% lidocaine
- Non-ionic radio-opaque contrast medium
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Pillow under anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 5.5.2*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- Mild sedation may be induced.

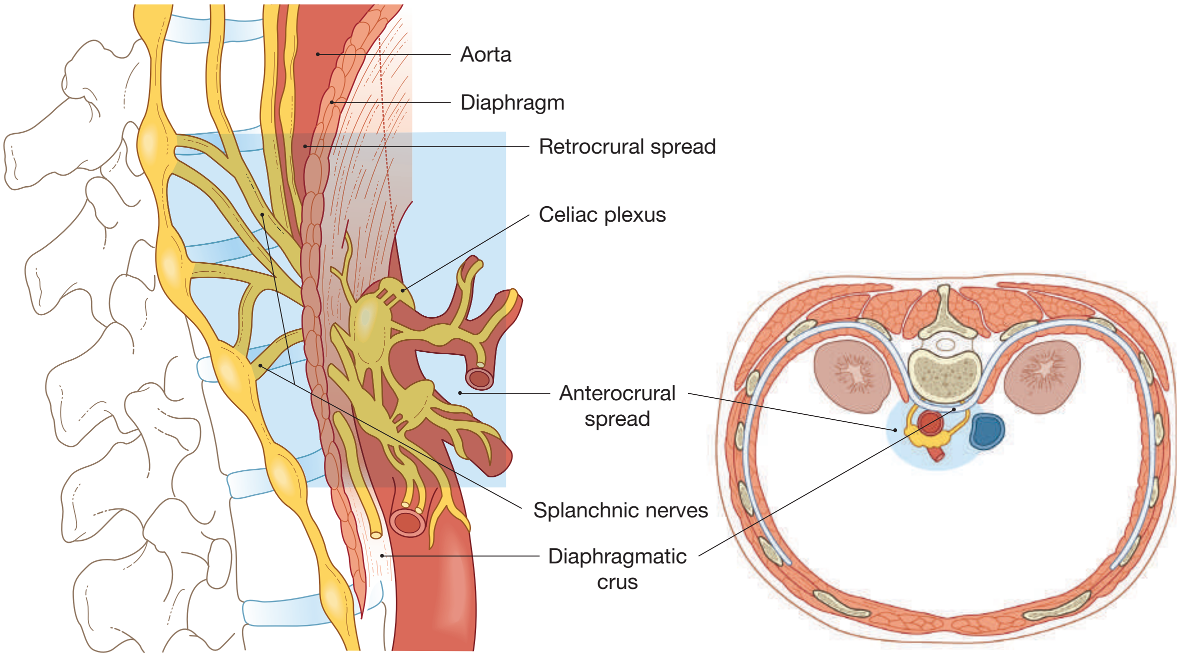

Fig. 5.5.1

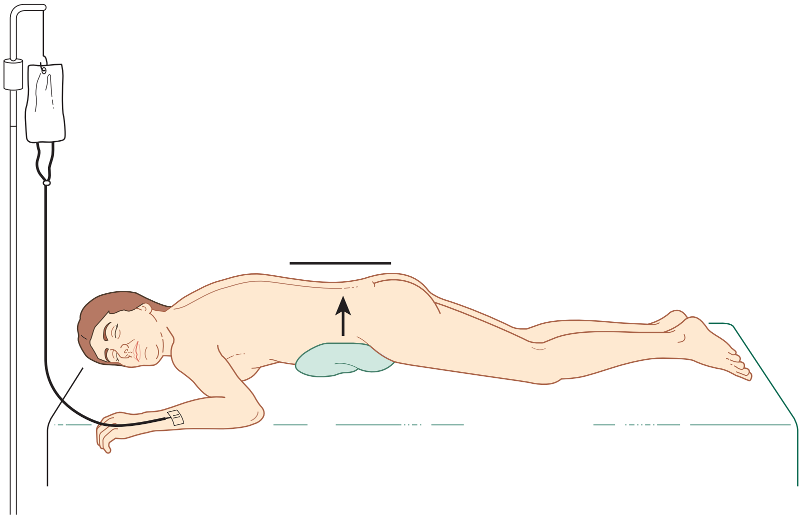

Fig. 5.5.2

- The thoracolumbar midline and area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- The twelfth rib and L1 are identified and confirmed with fluoroscopy.

ANTEROCRURAL APPROACH

Right side

- A 15 cm 22 G spinal needle is selected. A slight curve at the needle tip, away from the bevel direction, may be created which allows the needle to be redirected during placement.
- An AP view of the upper lumbar/low thoracic spine is obtained and the C-arm is adjusted to superimpose the T12–L1 endplates.
- **A skin wheal is raised at the lower border of the twelfth rib on the right just above the level of the L1 transverse process. The needle is inserted at this site and advanced at an angle 30° from perpendicular inward until the L1 body is contacted just below the upper endplate.**
- The curve of the needle is turned laterally and the needle is advanced along the upper portion of the body.
- Once the needle has slipped a few millimeters past the lateral aspect of the L1 body, a lateral view is obtained.
- The needle is advanced until the tip is 1.5–2 cm anterior to the anterior border of the L1 body. The needle is aspirated and if negative, 1 ml non-ionic contrast is injected. Dye spread should be in an amorphous pattern (*Fig. 5.5.3*).
- If aspiration is negative, 3 ml 1% lidocaine (lignocaine) is injected. If no nerve block is noted after 10 minutes, this is followed by 15–20 ml alcohol. If phenol is used, the lidocaine is not needed.
- 1 ml 1% lidocaine is injected before removing the needle to clear it.

Left side

- **The same procedure is repeated on the left.**
- The needle is positioned 1.5–2 cm anterior to the anterior border of the L1 body. It is then usually within the aorta, and aspiration is positive for arterial blood.
- The needle is advanced forward until aspiration is negative for blood (*Fig. 5.5.3 b*).
- 1 ml contrast is injected. The pattern is generally amorphous anteriorly, but a straight border of dye along the anterior surface of the aorta may be seen (*Fig. 5.4.4*).
- Aspiration is repeated and, if negative, 3 ml 1% lidocaine (lignocaine) is injected. If no nerve block is noted after 10 minutes, this is followed by 15–20 ml alcohol. If phenol is used, the lidocaine is not needed.
- 1 ml 1% lidocaine is injected before removing the needle to clear it.
- Monitors should be left attached and i.v. access left in situ for at least 30 minutes.

Confirmation of a successful block

- Relief of upper abdominal pain.

Tips

- After injection of non-ionic radio-contrast medium, a blush will indicate injection into muscle. If visible contrast medium disappears immediately it is likely that intravascular injection has occurred.
- Consideration should be given to performance of the block under CT control if the block is unsuccessful.

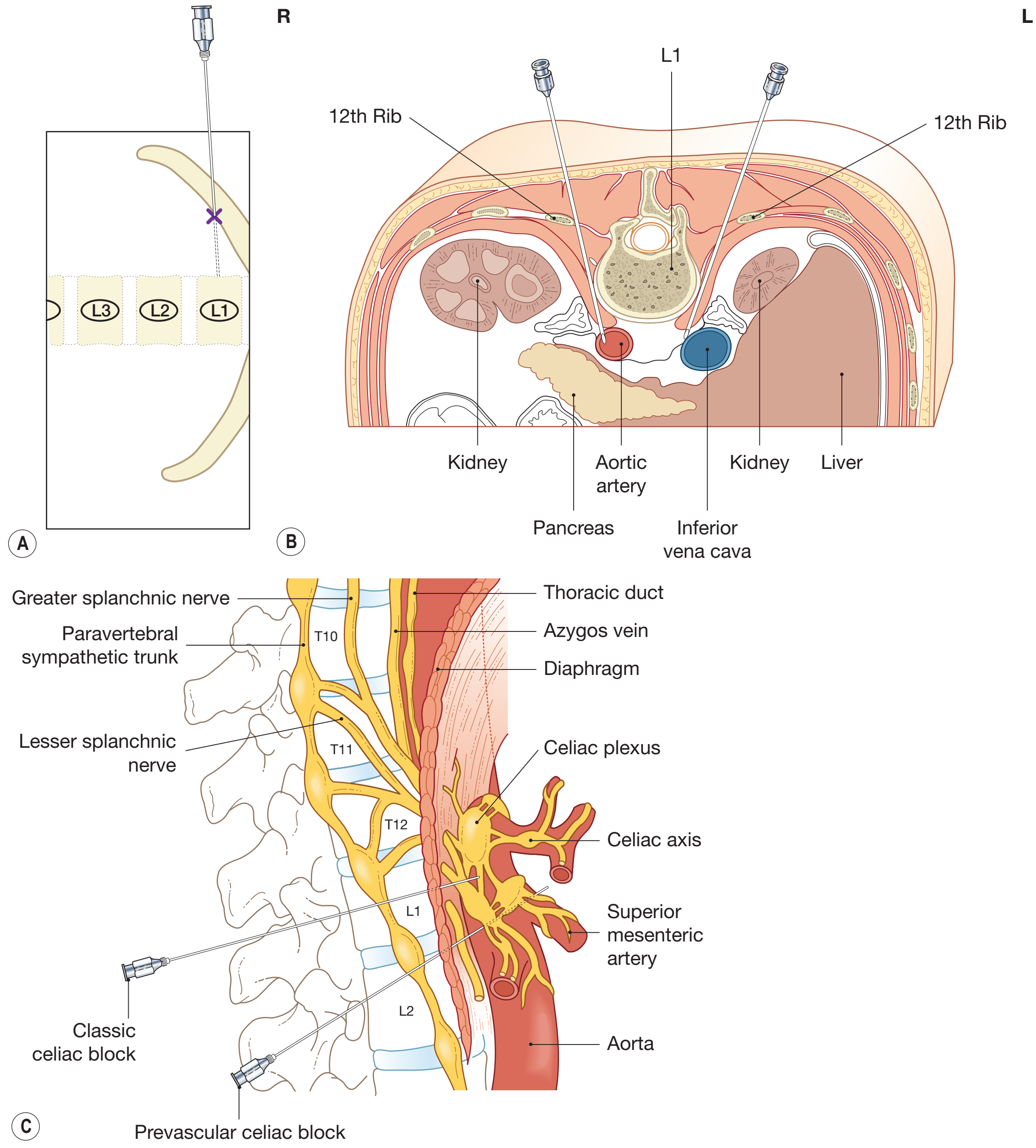
R
L
L3
L2
L1
A
L1
12th Rib
12th Rib
Kidney
Aortic artery
Kidney
Liver
Pancreas
Inferior vena cava
B
Greater splanchnic nerve
Paravertebral sympathetic trunk
T10
Lesser splanchnic nerve
T11
T12
L1
L2
Thoracic duct
Azygos vein
Diaphragm
Celiac plexus
Celiac axis
Superior mesenteric artery
Aorta
Classic celiac block
Prevascular celiac block
C

Fig. 5.5.3

Potential problems

- The position of each needle tip should always be confirmed with fluoroscopy prior to injection of neurolytic agent as it may lie in the peritoneal cavity, within a viscus or intravascularly. If a needle tip is placed too superficially, the tip may come to lie in the intervertebral foramen and injection may result in epidural block or a somatic nerve block. Injection of neurolytic solution into the psoas muscle may cause rhabdomyolysis.
- Perforation of the aorta or the inferior vena cava is possible and consequently the block should be avoided in patients with coagulopathy. Dissection of the aorta may occur as a result of direct damage during the block. Retroperitoneal hematoma may occur and for this reason also the block should be avoided in patients with coagulopathy.
- Orthostatic hypotension may occur as a result of sympathetic blockade for up to three days after a neurolytic block. Diarrhea may occur also and hydration of the patient should be monitored.

- Pneumothorax may occur.
- Transient motor paralysis and paraplegia may occur after the block, probably as a result of spasm of segmental arteries.
- Perforation of the intervertebral disc may occur, but this also is usually of no clinical significance.
- Perforation of the kidney or ureter is usually of no clinical significance unless neurolytic agents are injected.
- The thoracic duct may be damaged (possibly causing chylothorax, or lymphedema).
- Abdominal and chest discomfort may be experienced for 30 minutes after injection of alcohol.
- There may be a detectable odor from the breath after alcohol injection.
- Patients in the prone position should be monitored carefully when intravenous sedation is administered.

5.6 HYPOGASTRIC PLEXUS BLOCK

Anatomy

The superior hypogastric plexus is formed from pelvic sympathetic fibers of the aortic plexus and L2 and L3 splanchnic nerves. These afferent and efferent fibers innervate the pelvic viscera, including the uterus, bladder, vagina, and prostate. The plexus is located between the upper third of the first sacral vertebral body and the lower third of the fifth lumbar vertebral body, at the sacral promontory, in the retroperitoneal space (**Fig. 5.6.1 a,b**). Parasympathetic nerve fibers from S2–S4 pass through the inferior hypogastric plexus.

Equipment

- 2 ml, 5 ml, and 10 ml syringes
- 30 G needle
- Two 15 cm 22 G needles
- Extension set (optional)
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)
- Fluoroscopy

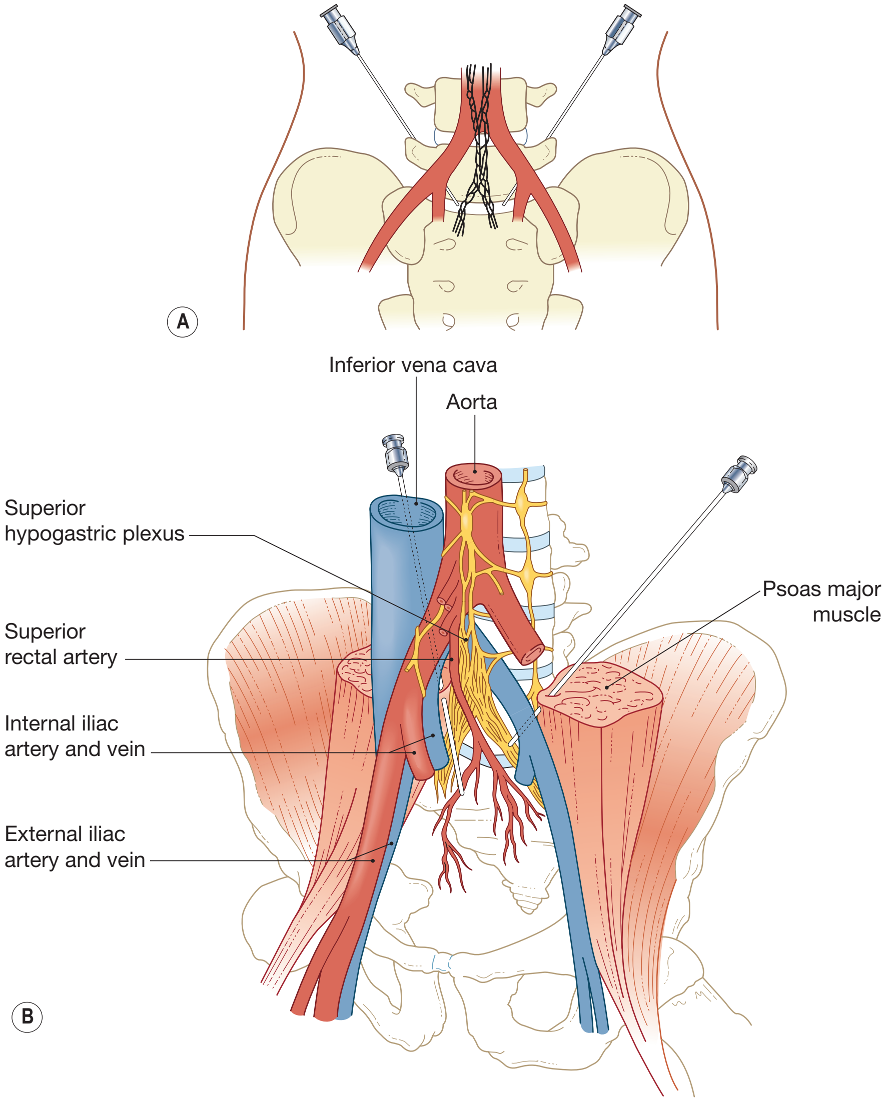
A
Inferior vena cava
Aorta
Superior hypogastric plexus
Superior rectal artery
Internal iliac artery and vein
External iliac artery and vein
Psoas major muscle
B

Fig. 5.6.1

Drugs

- Lidocaine (lignocaine) 1%, 5 ml for skin infiltration
- Lidocaine (lignocaine) 1%, 15–20 ml (or its equivalent) for block
- Phenol 6%
- Non-ionic radio-opaque contrast medium
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Pillow under anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 5.6.2*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The lumbosacral midline and area 10 cm × 5 cm laterally is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- The iliac crests are palpated and the intercrestal line is identified (this corresponds with the inferior aspect of the spinous process of L4 or may lie in the L4–5 interspace).
- The spinous processes are counted until the L5–S1 interspace is identified and confirmed with fluoroscopy.
- **The insertion points lie 2 cm lateral and 2 cm cephalad to the space between the L5 transverse process and the sacrum** (*Fig. 5.6.3 a,b*).
- A skin wheal is raised at one of the marked sites and the area is infiltrated with lidocaine (lignocaine) 1%.
- At a 30° angle to the frontal plane, a 22 G 15 cm needle is advanced, aimed slightly caudad towards the L5–S1 interspace (*Fig. 5.6.3 c*), until the lower part of the L5 vertebral body is contacted at a depth of about 7–9 cm and confirmed with fluoroscopy. If the needle contacts bone at a more superficial level, it is probable that it has come into contact with the L5 transverse process or the sacrum and needs to be repositioned.
- The needle depth is noted.
- The needle is then withdrawn to the subcutaneous tissue and, with the aid of fluoroscopy, it is re-advanced, this time at an angle 45° to the frontal plane (or with slight concavity of the needle) until the previous depth (as noted) is reached. It should slip past the vertebral body at a depth about 1–2 cm deeper than the first depth mark, to lie just anterior to the upper portion of the sacrum (*Fig. 5.6.4*).
- After negative aspiration, the fluoroscopic image is observed as a small amount of non-ionic radio-contrast medium is injected (*Fig. 5.6.5*). The correct placement of the needle is indicated by the presence of a collection of contrast medium just anterior to the upper portion of the sacrum or the L5–S1 interspace (*Fig. 5.6.6*). The contrast medium usually spreads in all directions, not usually along the sacrum.
- The procedure is repeated on the other side in a mirrored fashion (*Figs 5.6.7, 5.6.8*).
- After further negative aspirations, 5 ml of lidocaine (lignocaine) 1% is injected bilaterally and the patient is questioned about pain relief. There should be little resistance to injection, similar to that felt when injecting through an epidural needle. If resistance is encountered, or if the injection is painful, the needle should be repositioned.
- After 10 minutes the patient is questioned about pain relief and any symptoms of somatic nerve blockade. Sensory and motor functioning of the lower extremities is checked. The procedure should be abandoned if there is evidence of somatic blockade.
- After confirmation of pain relief and lack of somatic block, 6 ml phenol 6% is injected through each needle using glass syringes. The needles are then cleared with air or local anesthetic 1 ml, and removed.
- Monitors should be left attached and i.v. access left in situ for at least 30 minutes.
- Ultrasound may help placement of the needle (*Figs 5.6.9–5.6.10*).

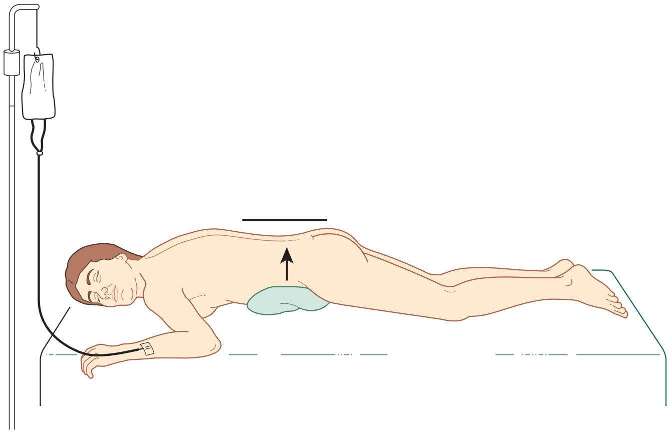

Fig. 5.6.2

Confirmation of a successful block

- Relief of lower abdominal pain.

Tips

- After injection of non-ionic radio-contrast medium, a blush will indicate injection into muscle. If this disappears immediately it is likely that intravascular injection has occurred.

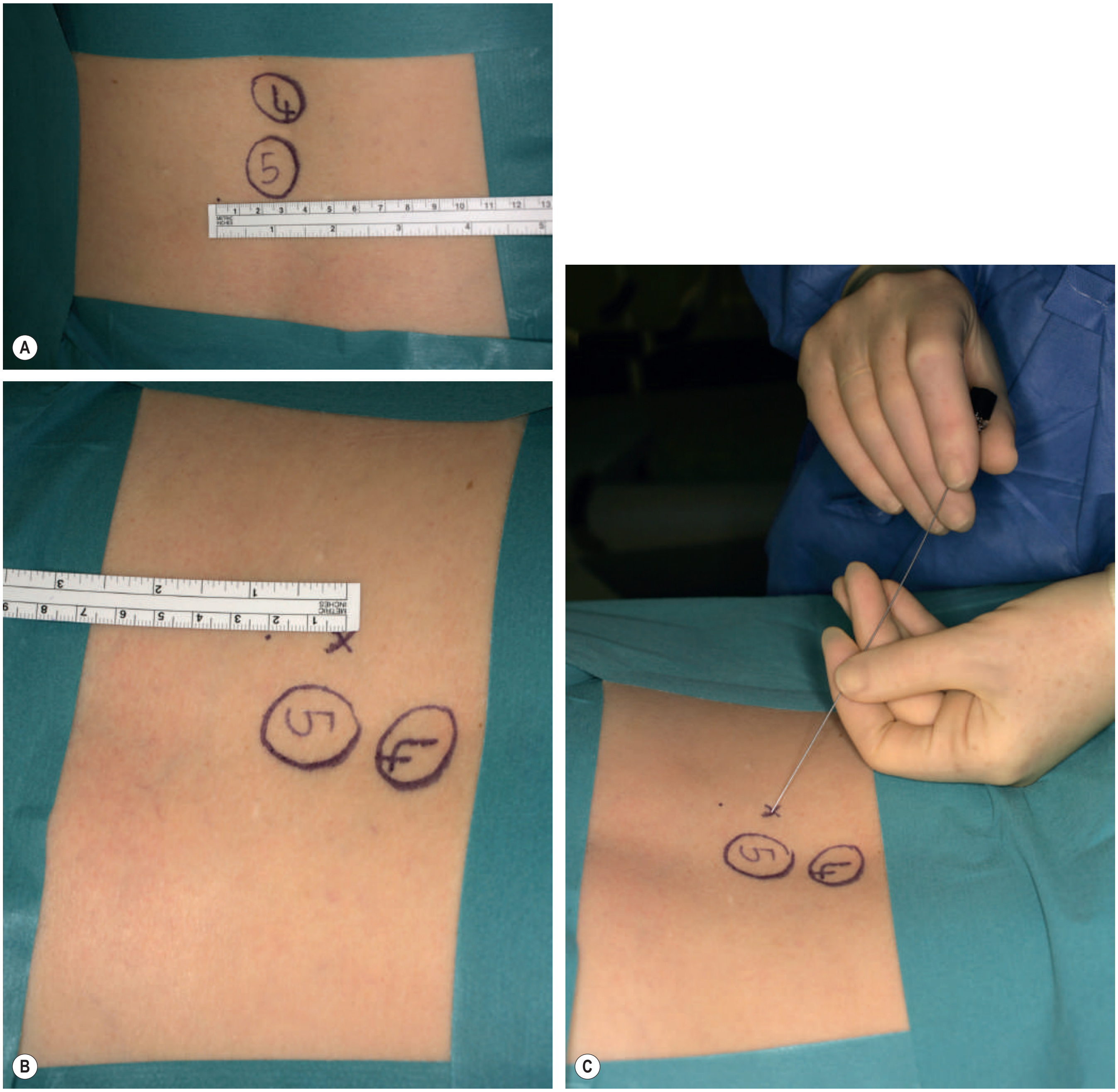

Fig. 5.6.3

- Consideration should be given to performance of the block under CT control if the block is unsuccessful (*Fig. 5.6.10*).

Potential problems

- The position of each needle tip should always be confirmed with fluoroscopy prior to injection of neurolytic agents as it may lie in the peritoneal cavity, within a viscus or intravascularly. If a needle tip is placed too superficially, the tip may come to lie in the intervertebral foramen and injection may result in an epidural block or a somatic nerve block. Injection of neurolytic solution into the psoas muscle may cause rhabdomyolysis.
- Perforation of the aorta or the inferior vena cava is possible and consequently the block should be avoided in patients with coagulopathy. Dissection of the aorta may occur as a result of direct damage during the block. Retroperitoneal hematoma may occur and for this reason also the block should be avoided in patients with coagulopathy.

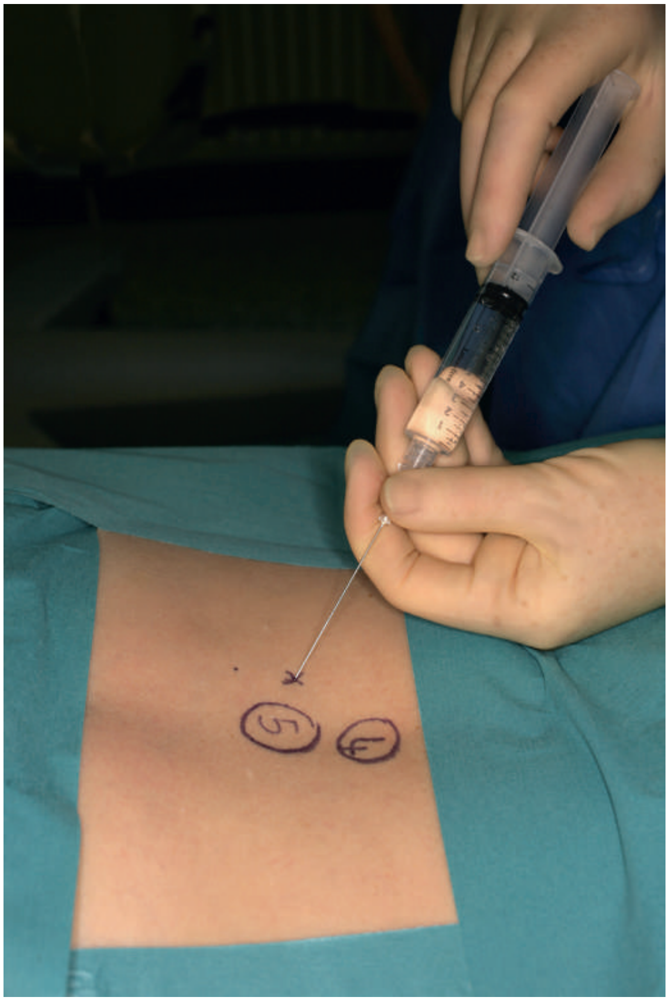

Fig. 5.6.4

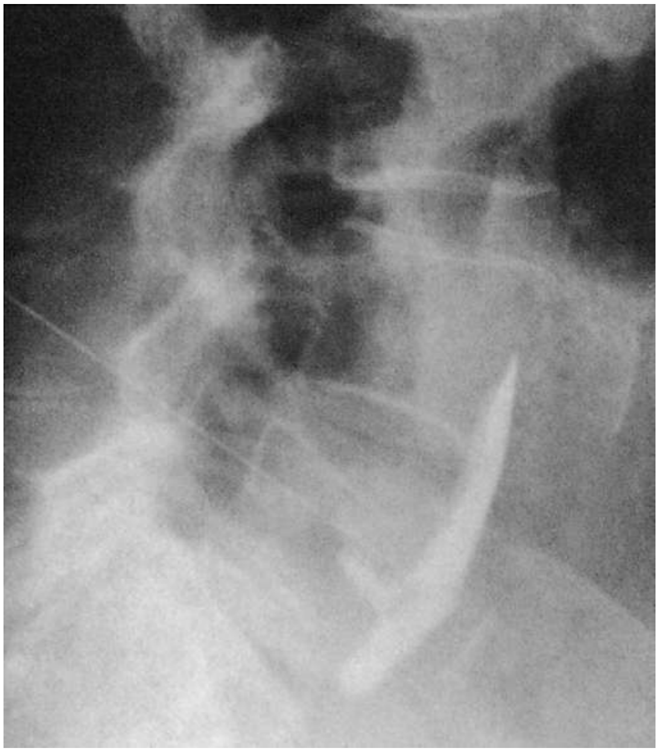

Fig. 5.6.6

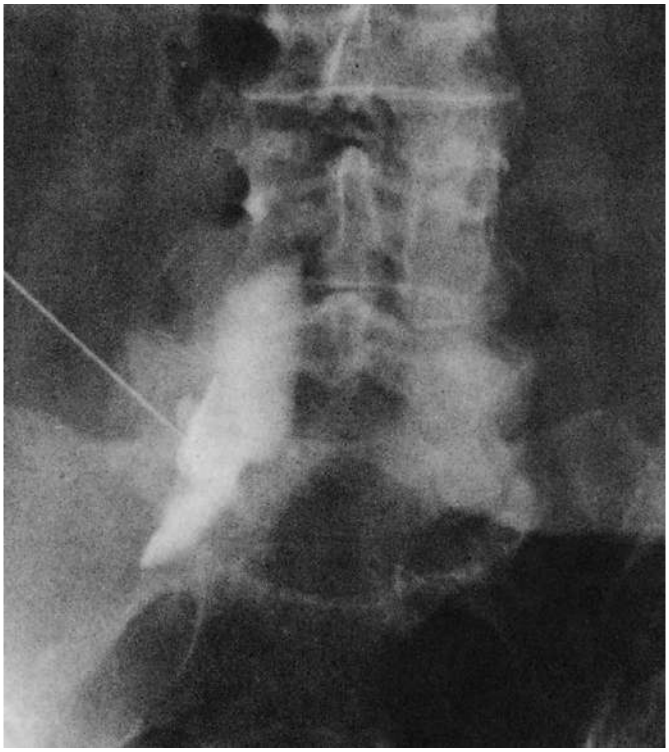

Fig. 5.6.5

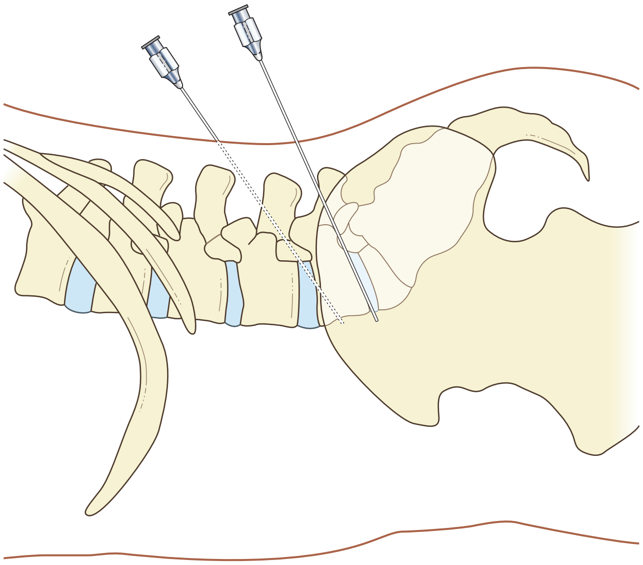

Fig. 5.6.7

- Orthostatic hypotension may occur as a result of sympathetic blockade for up to three days after a neurolytic block. Diarrhea may occur also and hydration of the patient should be monitored.
- Transient motor paralysis and paraplegia may occur after the block, probably as a result of spasm of segmental arteries.
- Perforation of the intervertebral disc may occur, but this is usually of no clinical significance.

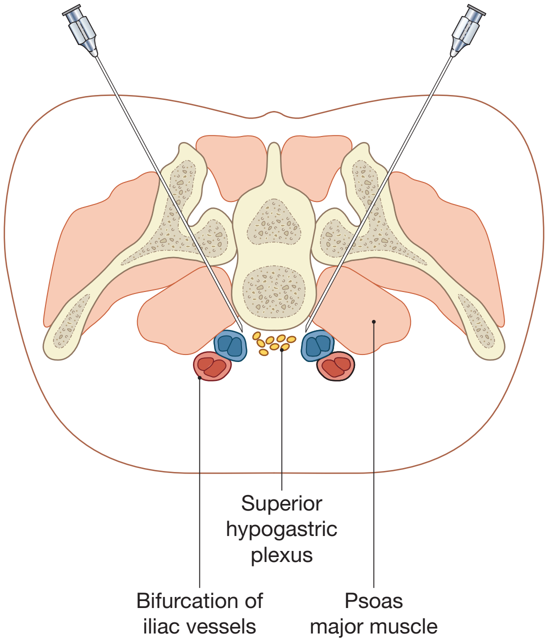
Superior hypogastric plexus
Bifurcation of iliac vessels
Psoas major muscle

Fig. 5.6.8

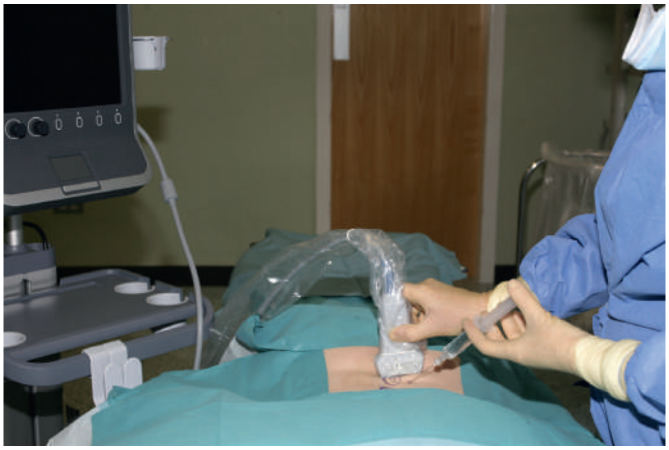

Fig. 5.6.9

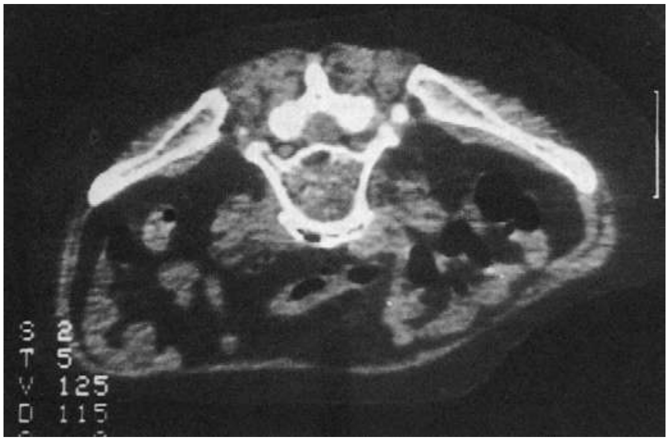
S 2
T 5
V 125
D 115

Fig. 5.6.10

5.7 GANGLION IMPAR BLOCK

Anatomy

The ganglion impar is a retroperitoneal sympathetic ganglion located at the level of the sacrococcygeal junction (**Fig. 5.7.1**). Above the level of this ganglion the sympathetic chains are paired. Sympathetic afferents from the perineum, distal rectum and anus, distal urethra, vulva and the distal third of the vagina converge in the ganglion impar.

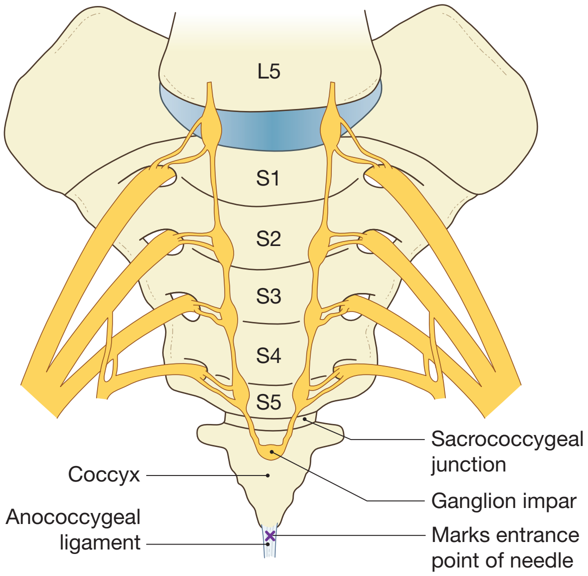

Fig. 5.7.1

Equipment

- 2ml, 5 ml, and 10 ml syringes
- 30 G needle
- 22 G spinal needle
- Extension set (optional)
- ECG, BP, and SpO_2 monitors
- Resuscitation equipment (*see Appendix 3*)

Drugs

- Lidocaine (lignocaine) 1%, 5 ml for skin infiltration
- Lidocaine (lignocaine) 1%, 15–20 ml (or its equivalent) for block
- Phenol 6%
- Non-ionic radio-opaque contrast medium
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Prone.
- Pillow under anterior superior iliac spine to flatten the normal lumbar lordosis (*Fig. 5.7.2*).

Needle puncture and technique

- Intravenous access is inserted.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The midline along the intergluteal groove and an area 10 cm × 5 cm laterally is cleaned with antiseptic

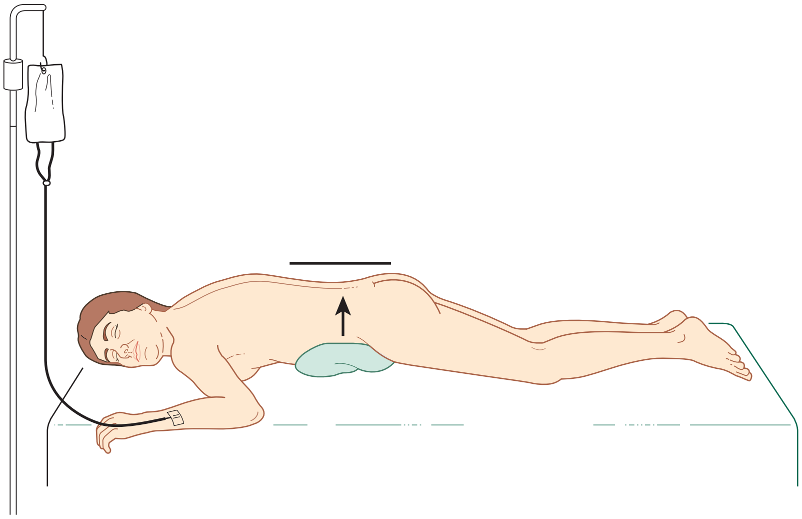

Fig. 5.7.2

solution and a fenestrated drape is placed over the sterile area.

- A skin wheal is raised at the superior aspect of the intergluteal groove, just above the anus, over the anococcygeal ligament (*Fig. 5.7.3*).
- The stylet from the 22 G spinal needle is removed, and the needle is bent with the fingers to form a 30° angle, approximately 2 cm from the hub.
- The needle is inserted through the skin wheal, with the concave curvature facing posteriorly.
- With the aid of fluoroscopy, the needle is advanced deep into the coccyx, closely approximating its anterior surface, until the tip reaches the level of the sacrococcygeal junction (*Fig. 5.7.4*).
- After negative aspiration, the fluoroscopic image is observed as a small amount of non-ionic radio-contrast medium is injected. The correct placement of the needle is indicated by the presence a small round blob of contrast medium at the anterior border of the vertebral column (*Fig. 5.7.5 a,b*).
- Lidocaine (lignocaine) 1% 5 ml is injected for ganglion blockade.

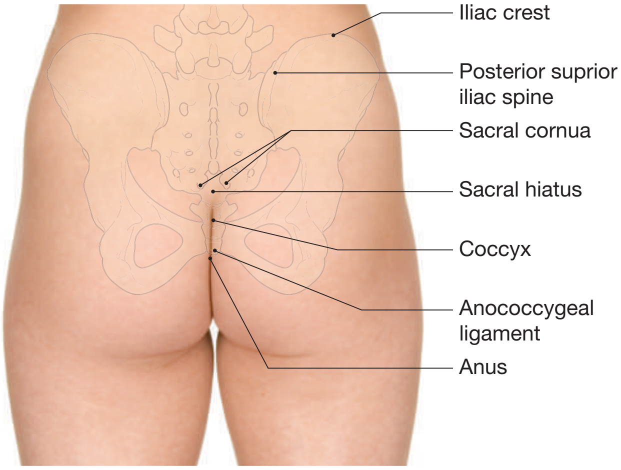
Iliac crest
Posterior suprior iliac spine
Sacral cornua
Sacral hiatus
Coccyx
Anococcygeal ligament
Anus

Fig. 5.7.3

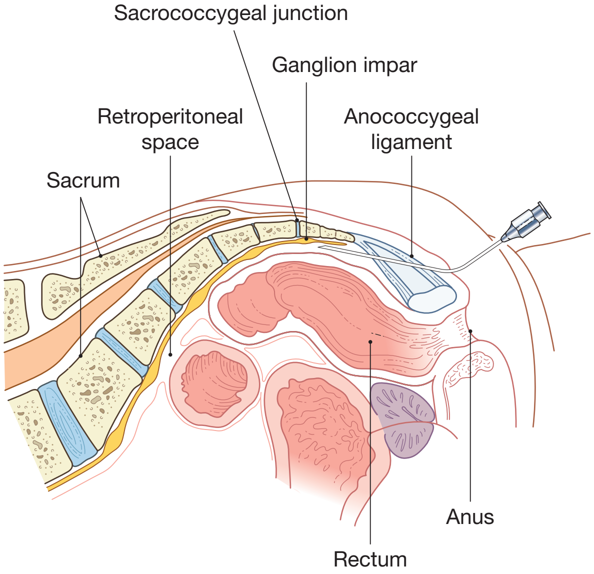
Sacrococcygeal junction
Ganglion impar
Retroperitoneal space
Anococcygeal ligament
Sacrum
Anus
Rectum

Fig. 5.7.4

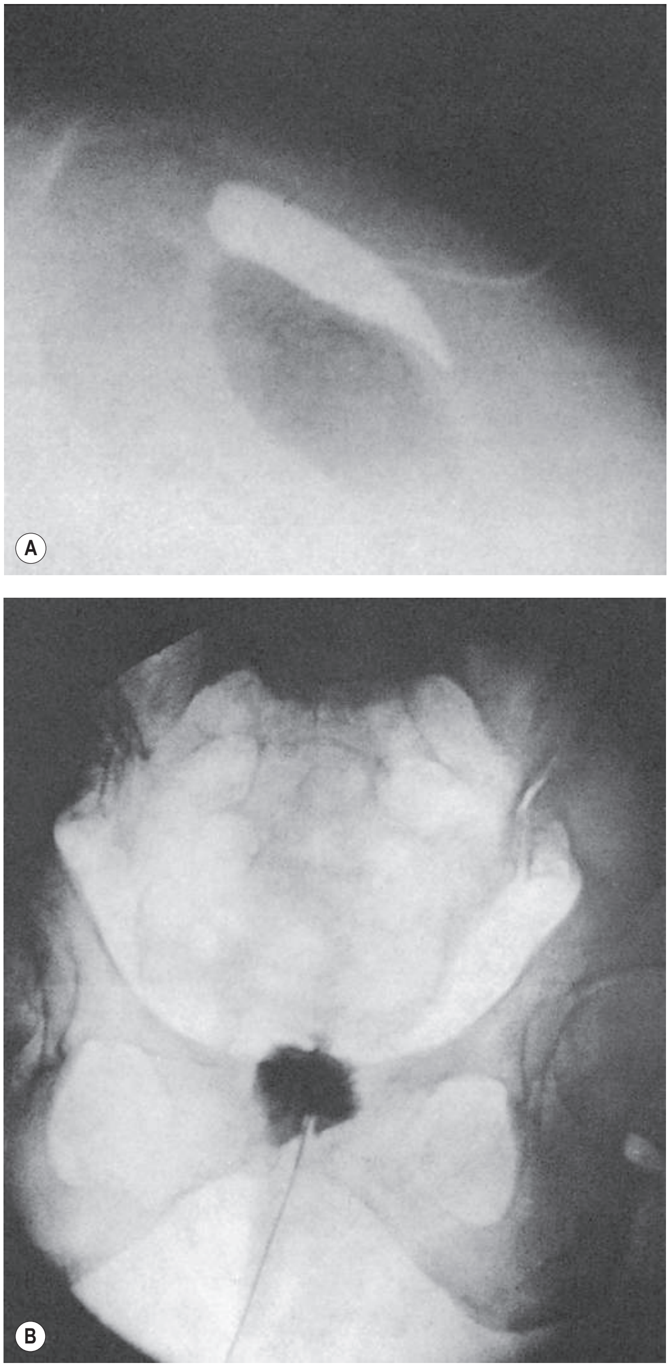

Fig. 5.7.5

- After 10 minutes the patient is questioned about pain relief and any somatic blockade. Sensory and motor functioning of the lower extremities is checked. The procedure should be abandoned if there is evidence of somatic blockade.
- After confirmation of pain relief and lack of somatic block, 5 ml of phenol 6% is injected using a glass syringe. To avoid leaving alcohol in the needle tract the needle is then cleared with air or local anesthetic 1 ml, and removed.
- Monitors should be left attached and i.v. access left in situ for at least 30 minutes.

Confirmation of a successful block

- Relief of perineal pain.

Tips

- To aid access to the anococcygeal ligament an assistant may be asked to retract the skin of the buttock; after penetration of the skin, this is no longer required.
- Exaggerated anterior curvature of the sacrococcygeal vertebral column may inhibit access and it may be necessary to bend the needle to a more acute angle.

Potential problems

- The position of each needle tip should always be confirmed with fluoroscopy prior to injection of neurolytic agent as it may lie in the peritoneal cavity, within a viscus or intravascularly. Caudal epidural placement of the needle is possible, therefore it is essential that spread of contrast material is observed to be restricted to the retroperitoneum, and that a test dose produces no somatic nerve blockade. Perforation of the rectum or periosteal injection is also possible.
- Local tumor invasion may inhibit spread of solution.
- Retroperitoneal hematoma may occur and the block should be avoided in patients with coagulopathy. Diarrhea may occur also and hydration of the patient should be monitored.
- There may be a detectable odor from the breath after alcohol injection.
- Patients in the prone position should be monitored carefully when intravenous sedation is administered.

5.8 INTRAVENOUS REGIONAL SYMPATHETIC BLOCK—UPPER LIMB

Anatomy

Peripheral sympathetic blockade is achieved by limiting the effect of the sympatholytic agent to the tissues of the affected limb using a tourniquet. Intravenous injection of an agent that releases endogenous norepinephrine (noradrenaline) from sympathetic nerve endings causes depletion of this neurotransmitter, and thereby chemical sympathetic blockade.

Equipment

- 20 ml syringe
- Two i.v. cannulae
- Pneumatic tourniquet
- ECG, BP, and SpO_2 monitors
- Skin temperature monitor
- Resuscitation equipment (*see Appendix 3*)

Drugs

- Lidocaine (lignocaine) 0.5% *without epinephrine/adrenaline*, or its equivalent
- Bretylium 1.5 mg/kg (or its equivalent, e.g. guanethedine 0.25 mg/kg)
- Saline (NaCl) 20 ml
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Supine.

Technique

- Intravenous access is inserted in the contralateral limb.
- Peripheral i.v. access is inserted in the limb to be blocked (*Fig. 5.8.1 a*).
- Monitors are attached.

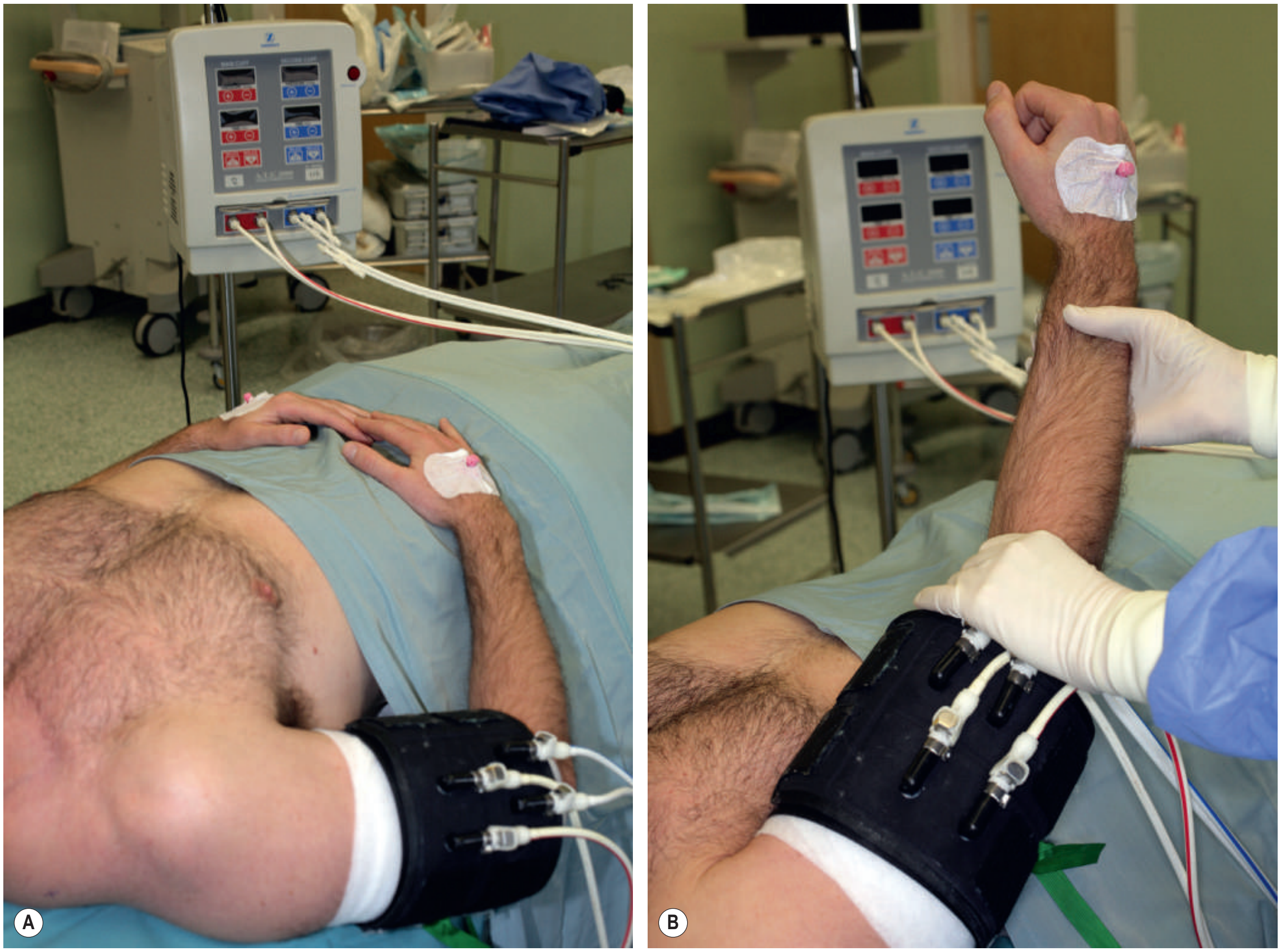

Fig. 5.8.1

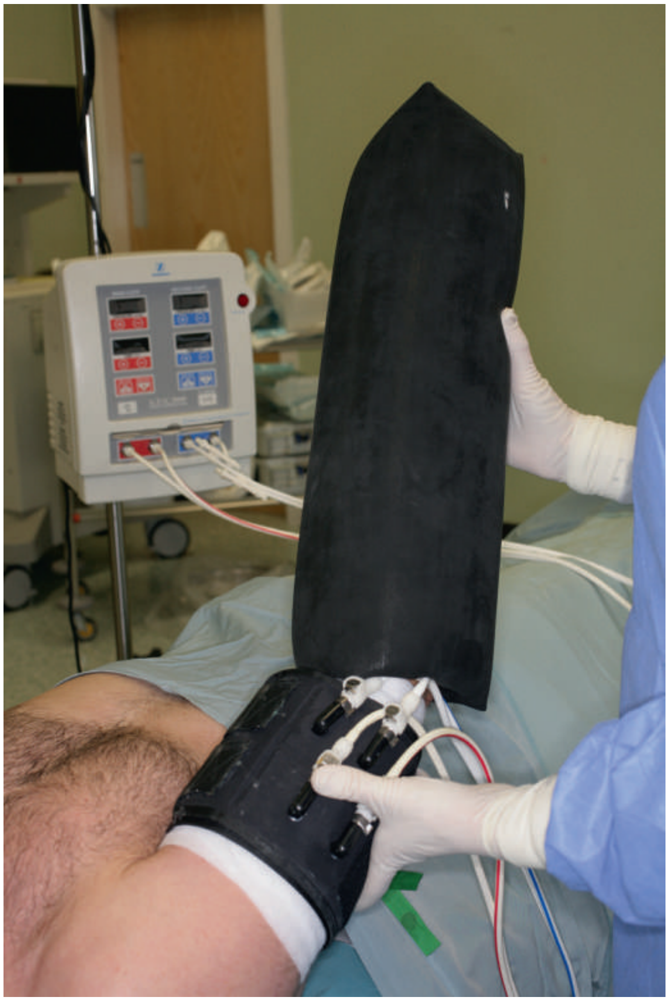

Fig. 5.8.2

- Resuscitation equipment and drugs are checked and made ready for use.
- The limb is raised above the level of the heart for 2 minutes (*Fig. 5.8.1 b*).
- With the limb raised, it is exsanguinated by applying a tight wrap, e.g. Esmarch bandage.
- A thin layer of padding is applied, e.g. Velband, under the tourniquet site.
- The tourniquet is applied and the cuff is inflated to a pressure 100 mmHg higher than the systolic blood pressure (*Fig. 5.8.2*).
- The limb is then lowered. A mixture of lidocaine (lignocaine) 0.5% 15 ml (*without epinephrine/adrenaline*), bretylium 1.5 mg/kg (or guanethidine 0.25 mg/kg), and NaCl to make a total volume of 40 ml (a final lidocaine (lignocaine) solution of 0.25%), is injected through the i.v. cannula in the affected limb.
- The tourniquet is allowed to remain inflated for at least 30 minutes.

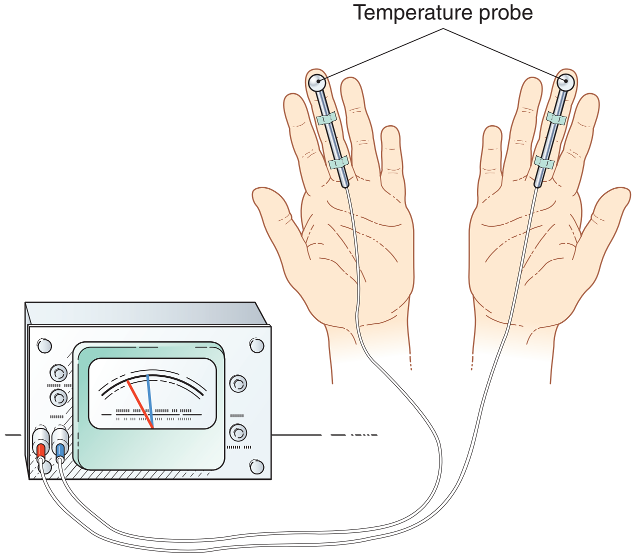

Fig. 5.8.3

- It is then deflated in one step, but left in place. Re-inflation may be required if there is a precipitous change in blood pressure.
- Monitors should be left attached and i.v. access left in situ for at least 30 minutes.

Confirmation of a successful block

- Relief of sympathetically maintained pain.
- Measurements of skin temperature of the affected limb before and after the block should demonstrate temperature increase. However, the sympatholytic effect of the drug may not be immediate (*Fig. 5.8.3*).

Tips

- If i.v. access to the affected limb is difficult due to vasoconstriction, a smear of glycerol trinitrate cream on the dorsum of the hand will usually aid i.v. insertion.
- A single or double cuff may be employed for this block but a double tourniquet may make the block more comfortable. The proximal cuff is inflated first. A few minutes after injection the distal cuff is inflated and **when inflation is complete the proximal cuff is released.**
- Retrograde cannulation, i.e. towards the periphery (*Figs 5.8.4, 5.8.5*) rather than proximally (*Figs 5.8.6, 5.8.7*), may help direct the spread of bretylium to the periphery.
- Active or passive movements of the limb may hasten the distribution of bretylium to the periphery.
- If the tourniquet inflation is painful, inhalation of nitrous oxide–oxygen mixture may improve comfort.
- Repeated blocks may bring about gradual improvement in sympathetically maintained pain.

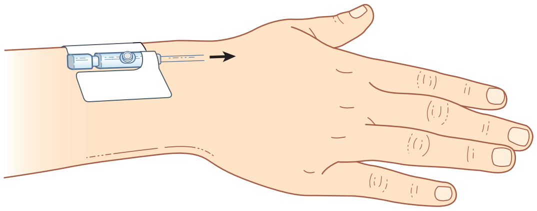

Fig. 5.8.4

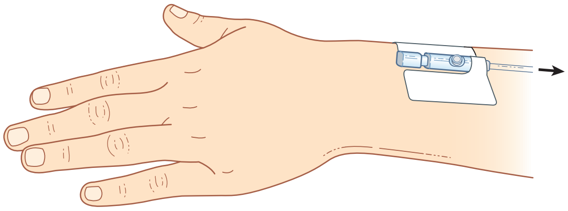

Fig. 5.8.6

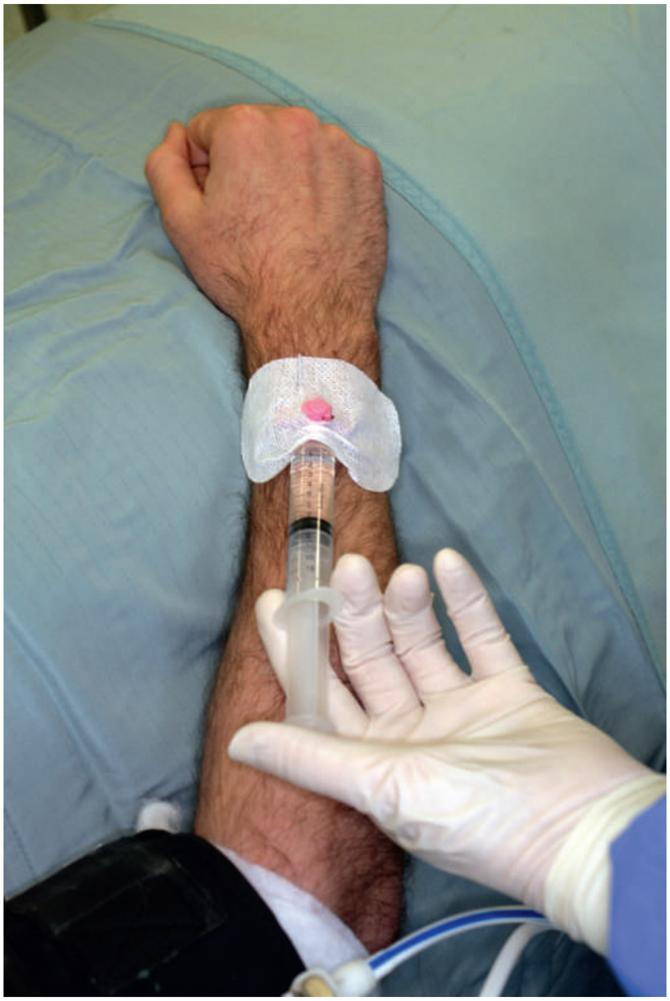

Fig. 5.8.5

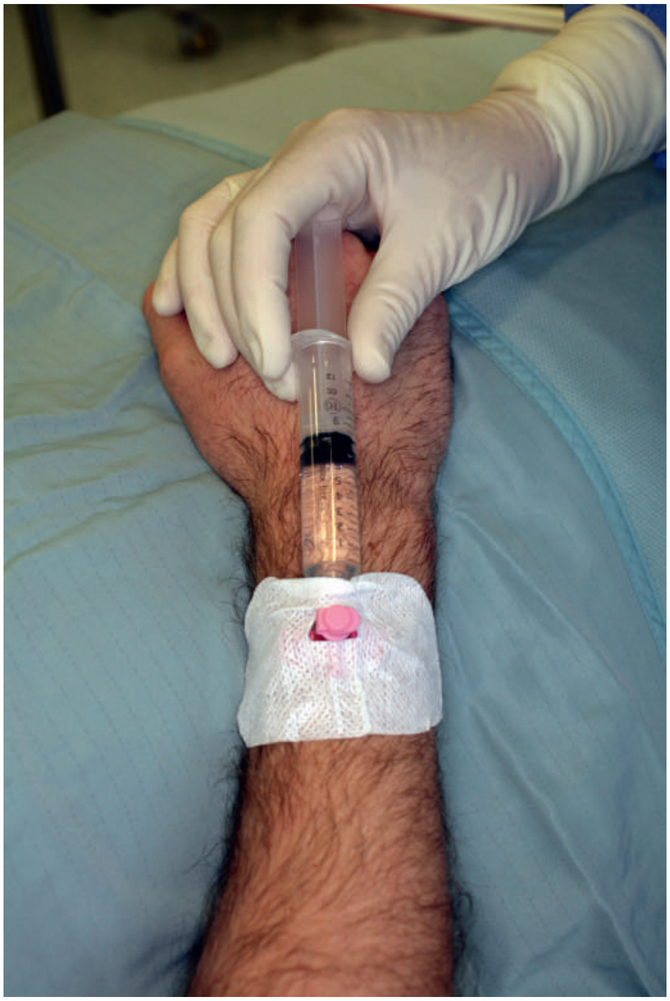

Fig. 5.8.7

- Immediate physiotherapy after block may improve outcome.

Potential problems

- Accidental deflation of the tourniquet early in the procedure may cause a precipitous rise in blood pressure due to the general release of endogenous norepinephrine/noradrenaline when unfixed bretylium enters the circulation. Systemic toxicity of lidocaine (lignocaine) may also occur, possibly causing seizures. Blood pressure may decrease after deflation of the cuff later in the procedure.
- The tourniquet inflation may be painful.
- A sensation of burning may occur after injection due to release of endogenous norepinephrine.
- Neuropraxia may occur (rarely) with a very tight tourniquet.
- Avoid in sickle cell anemia.

5.9 INTRAVENOUS REGIONAL SYMPATHETIC BLOCK—LOWER LIMB

Anatomy

As in the case of the upper limb, peripheral sympathetic blockade is achieved by limiting the effect of the sympatholytic agent to the tissues of the affected limb using a tourniquet. Intravenous injection of an agent releases endogenous norepinephrine (noradrenaline) from sympathetic nerve endings, which causes depletion of this neurotransmitter, and may bring about a chemical sympathetic block.

Equipment

- 50 ml syringe
- Two i.v. cannulae
- Pneumatic tourniquet
- ECG, BP, and SpO_2 monitors
- Skin temperature monitor
- Resuscitation equipment (*see Appendix 3*)

Drugs

- Lidocaine (lignocaine) 0.5% *without epinephrine (adrenaline)*, or its equivalent
- Bretylium 1.5 mg/kg (or its equivalent, e.g. guanethedine 0.5 mg/kg)
- Saline (NaCl) 30 ml
- Resuscitation drugs (*see Appendix 3*)

Position of patient

- Supine.

Technique

- Intravenous access is inserted in the contralateral limb.
- Peripheral i.v. access is inserted in the limb to be blocked.
- Monitors are attached.
- Resuscitation equipment and drugs are checked and made ready for use.
- The limb is raised above the level of the heart for 2 minutes (*Fig. 5.9.1*).
- With the limb raised, it is exsanguinated by applying a tight wrap (*Fig. 5.9.2*).
- A thin layer of padding is applied, e.g. Velband, under the tourniquet site.
- The tourniquet is applied and inflated to a pressure 100 mmHg higher than the systolic blood pressure. A second tourniquet may be applied to the calf of patients with no known predispositions to deep venous thrombosis. Inflation of this second cuff may aid limitation of spread of sympatholytic agent to the periphery (*Fig. 5.9.3*).
- The limb is then lowered and a mixture of lidocaine (lignocaine) 0.5% 25 ml (*without epinephrine*), bretylium 1.5 mg/kg (or its equivalent) and NaCl to make a total volume of 40 ml (a final lidocaine/lignocaine solution of 0.25%) is injected through the i.v. cannula in the affected limb.
- The tourniquet is allowed to remain inflated for at least 30 minutes.
- It is then deflated in one step, but left in place. Re-inflation may be required if there is a precipitous change in blood pressure.

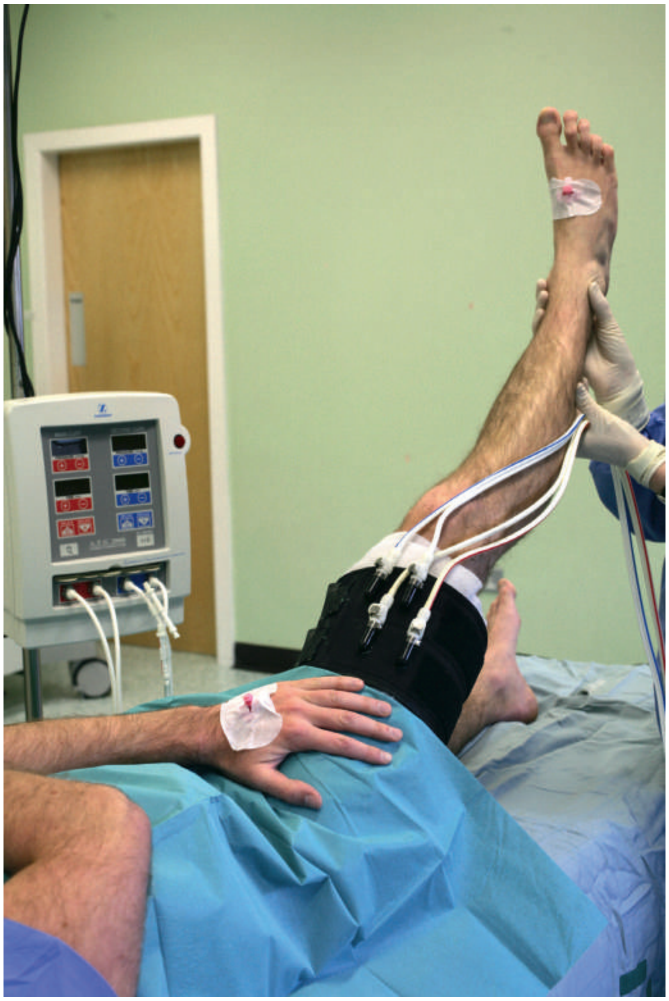

Fig. 5.9.1

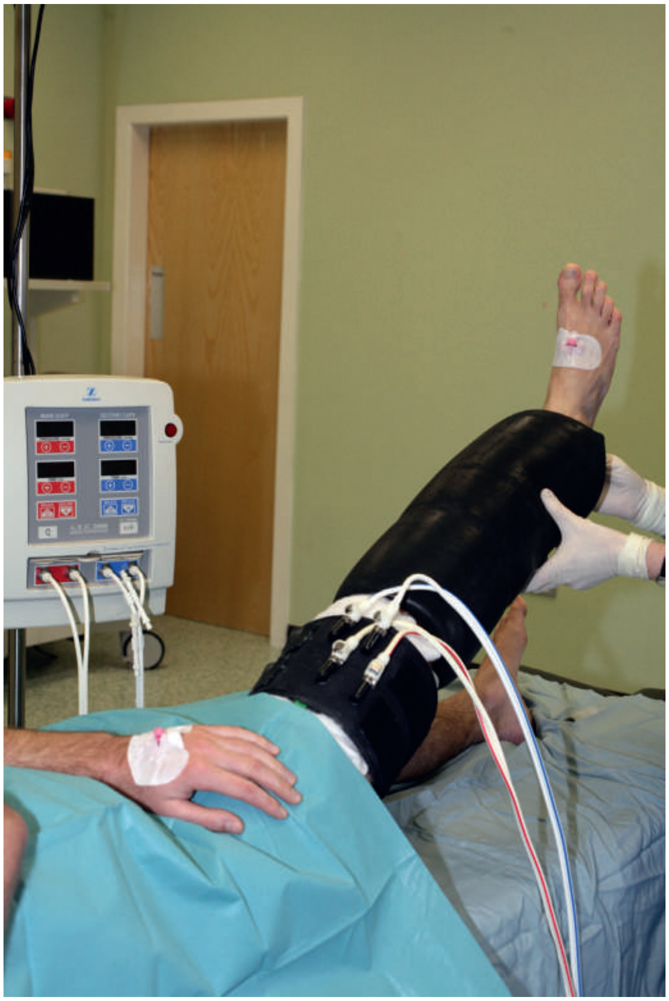

Fig. 5.9.2

- Monitors should be left attached and i.v. access left in situ for at least 30 minutes.

Confirmation of a successful block

- Relief of sympathetically maintained pain.
- Measurements of skin temperature of the affected limb before and after the block should demonstrate temperature increase (*Fig. 5.9.4*).

Tips

- If i.v. access to the affected limb is difficult due to vasoconstriction, a smear of glycerol trinitrate cream on the dorsum of the foot will usually aid i.v. insertion.
- A single or double cuff may be employed for this block but a double tourniquet may make the block more comfortable. The proximal cuff is inflated first. A few minutes after injection the distal cuff is inflated and when inflation is complete the proximal cuff is released.
- Retrograde cannulation, i.e. towards the periphery (*Fig. 5.9.5*) rather than proximally (*see Fig. 5.9.3*), may help direct the spread of bretylium to the periphery.
- Active or passive movements of the limb may hasten the distribution of bretylium to the periphery.
- If the tourniquet inflation is painful, inhalation of nitrous oxide–oxygen mixture may improve comfort.
- Repeated blocks may bring about gradual improvement in sympathetically maintained pain.
- Immediate physiotherapy after block may improve outcome.

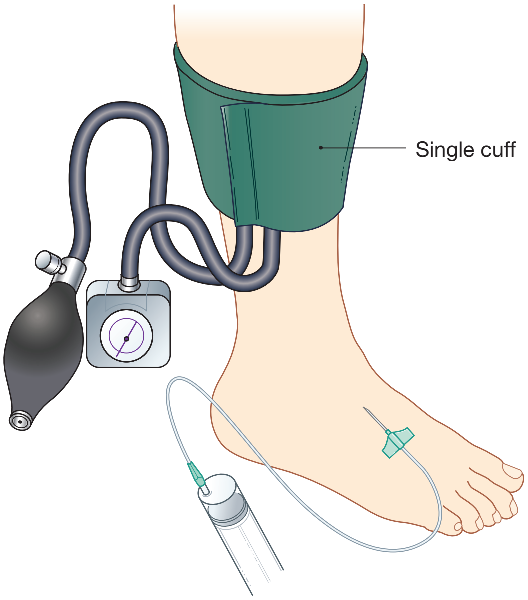

Fig. 5.9.3

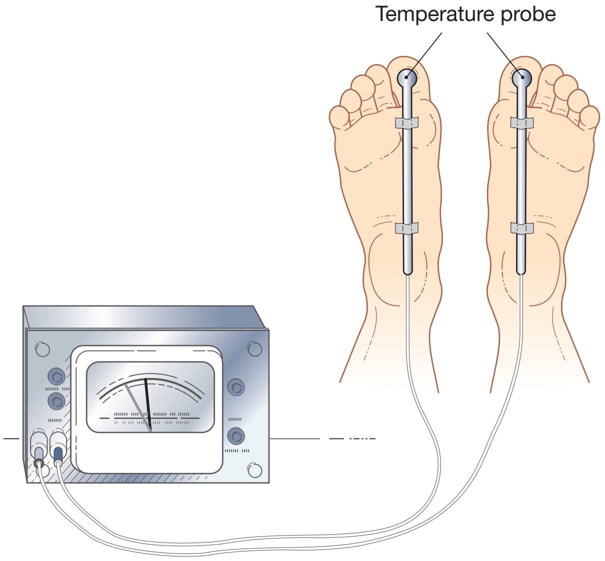

Fig. 5.9.4

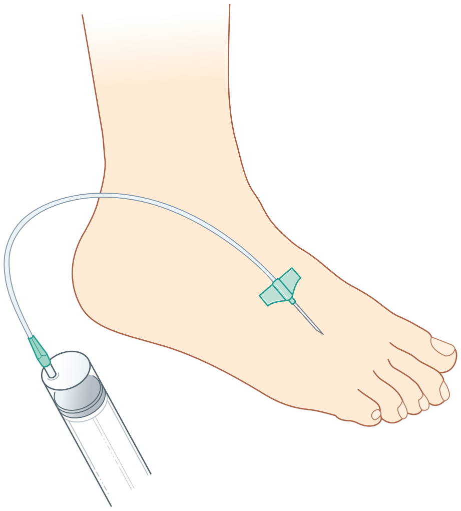

Fig. 5.9.5

Potential problems

- Accidental deflation of the tourniquet early in the procedure may cause a precipitous rise in blood pressure due to the general release of endogenous norepinephrine or epinephrine when unfixed bretylium enters the circulation. Systemic toxicity of lidocaine (lignocaine) may also occur in high doses possibly causing seizures. Blood pressure may decrease after deflation of the cuff later in the procedure.
- The tourniquet inflation may be painful.
- A sensation of burning may occur after injection due to release of endogenous norepinephrine.
- Neuropraxia may occur (rarely) with a very tight tourniquet.
- Avoid in sickle cell anemia.

MUSCLE INJECTIONS

6

Myofascial pain occurs commonly in the muscles of the upper and lower back. It is characterized by pain associated with movement of the affected muscles that develop areas of extreme tenderness, termed trigger points. Palpation of these points is usually perceived as a tight band or firm nodule in the muscle and reproduces pain that may be referred some distance from the site of palpation. Involuntary muscular contraction can occur on palpation, and snapping palpation can result in a local twitch response. Electromyography (EMG) is not reliable in diagnosing myofascial pain syndrome and it is worth remembering that this syndrome may occur in association with underlying painful disorders of the spine.

Injection of trigger points with local anesthetic, especially if repeated several times and combined with stretching exercises, may have a beneficial therapeutic effect on the pain of myofascial pain syndrome. Pain reproduction during injection, followed by relief of pain after injection, that lasts at least as long as the expected local anesthetic effect, indicates that these painful points contribute to myofascial pain syndrome.

Fibromyalgia is a pain syndrome characterized by widespread, diffuse and usually symmetrical tender areas of muscles. Bony structures, such as costochondral junctions and lateral epicondyles, produce local pain, but not referred, on palpation of tender points. Injection of these tender areas typically does not improve the pain of fibromyalgia.

Usually a dilute solution of local anesthetic suffices for beneficial effect. Bupivacaine produces more muscle degeneration than any other local anesthetic when injected into a muscle, and consequently it is usually avoided, lidocaine (lignocaine) being the usual local anesthetic of choice.

The optimum number of trigger-point injections required to produce pain relief is variable. The injection sites may themselves be painful after the local anesthetic wears off. This may exacerbate muscle spasm if too many trigger-point injections are performed. Consideration should be given to the severity of the muscle spasm, the number of trigger points, and to the sensitivity of the patient to pain when deciding on the number of injections.

6.1 TRIGGER-POINT INJECTIONS—NECK AND THORAX

Anatomy

The muscles most often involved in myofascial pain syndrome of the neck include the trapezius, rhomboid minor and major, latissimus dorsi, levator scapulae and splenius capitis (**Fig. 6.1.1**; see also Fig. 2.1.2).

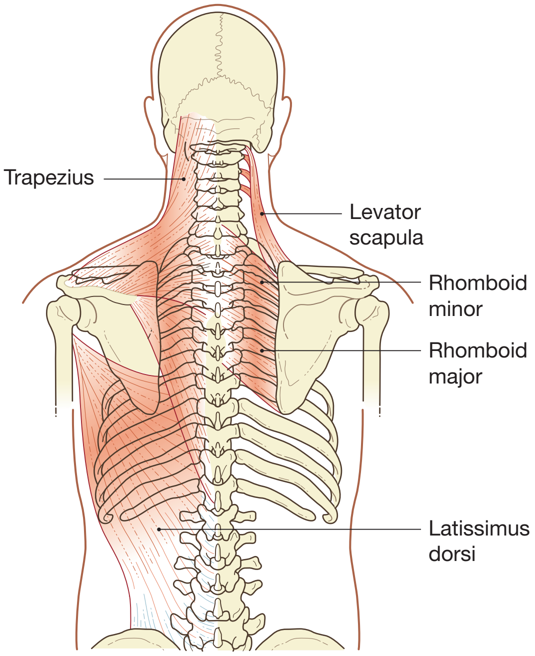

Fig. 6.1.1

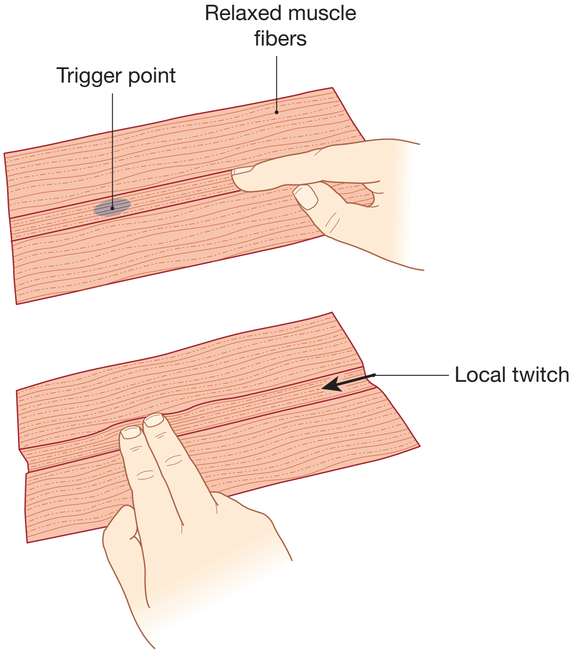

Fig. 6.1.2

Equipment

- 10 ml syringe
- 25 G needle

Drugs

- Lidocaine (lignocaine) 1% 10 ml

Position of patient

- Prone.
- Pillow under chest to allow the neck to flex.
- The sitting position is also used, but vasovagal response may follow trigger-point injections especially in young adults, and it is probably more prudent to use the prone position.

Needle puncture and technique

- The neck, shoulders, and upper posterior thorax are cleaned with antiseptic solution.
- **Trigger points in the muscles are palpated** (**Fig. 6.1.2**) **and marked** (*Fig. 6.1.3*).
- A 25 G needle with syringe attached is inserted into a trigger point (*Fig. 6.1.4*).
- After negative aspiration, 2–3 ml of lidocaine (lignocaine) 1% is injected into the trigger point while moving the needle back and forth through the muscle.
- After injection, the next trigger point is injected in the same manner (*Fig. 6.1.5 a,b*).

Confirmation of a successful injection

- Pain reproduction when the needle enters the muscle confirms correct placement.

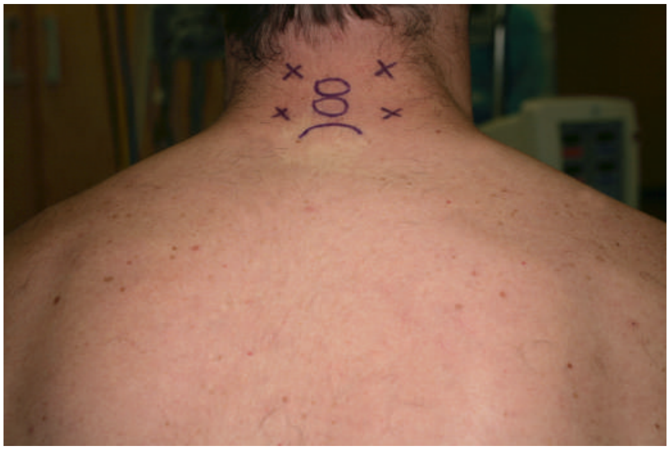

Fig. 6.1.3

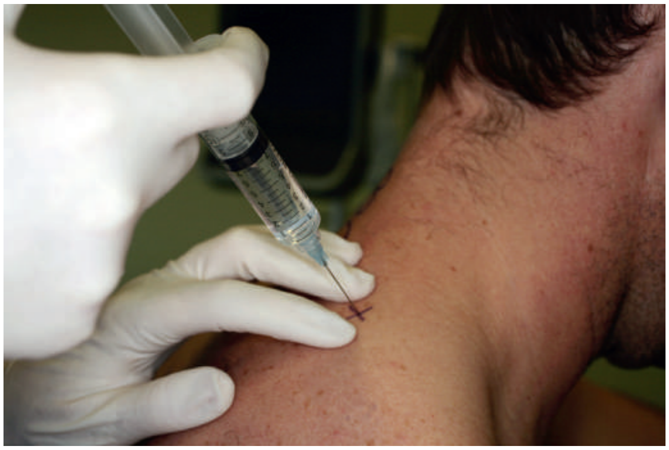

Fig. 6.1.4

Tips

- For best results, injection is carried out in a fan-like manner by repeatedly withdrawing the needle slightly and redirecting it.
- Stretching of the involved muscles by physiotherapy within the duration of the local anesthesia improves results.
- Some workers advocate massage of the area immediately after injection.

Potential problems

- Pain on injection.
- Vasovagal response (especially in young adults in the sitting position).
- Pneumothorax (especially in thin patients).

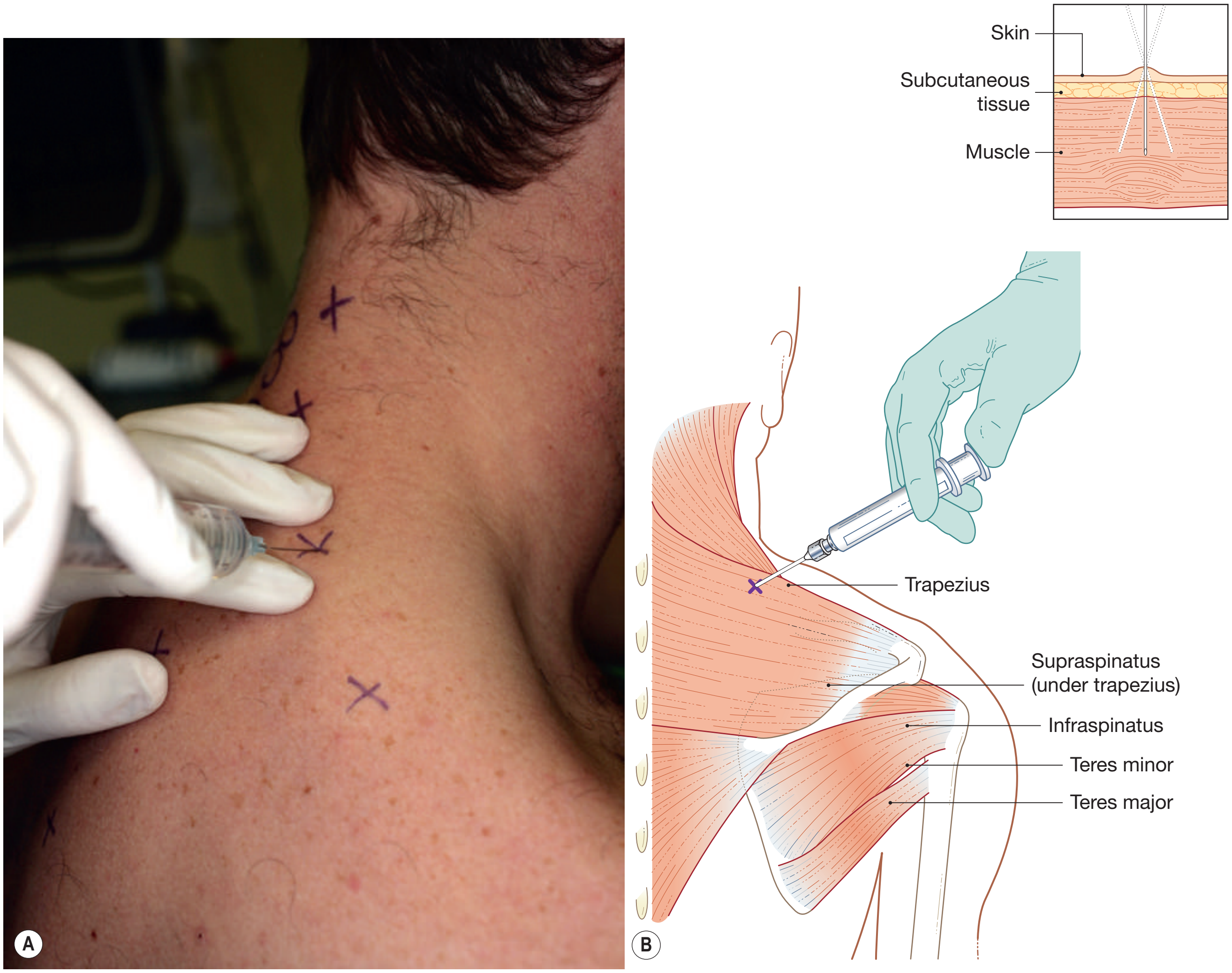
A
Skin
Subcutaneous tissue
Muscle
Trapezius
Supraspinatus (under trapezius)
Infraspinatus
Teres minor
Teres major
B

Fig. 6.1.5

6.2 TRIGGER-POINT INJECTIONS—BACK

Anatomy

The muscles most often involved in myofascial pain syndrome of the back include the erector spinae (the longissimus, iliocostalis, and spinalis columns) and the deep transversospinal (semispinalis, multifidus, and rotatores) muscles (**Fig. 6.2.1 a,b**). In the buttocks, spasm of the gluteus medius muscle may also cause significant pain.

Equipment

- 10 ml syringe
- 25 G needle

Drugs

- Lidocaine (lignocaine) 1% 10 ml

Position of patient

- Prone.
- Pillow under abdomen to straighten the normal lumbar lordosis (*Fig. 6.2.2 a*).
- The sitting position is also used, but vasovagal response may follow trigger-point injections, and it is probably more prudent to use the prone position.
- Alternatively, the semiprone position will also allow access to affected muscles (*Fig. 6.2.2 b*).

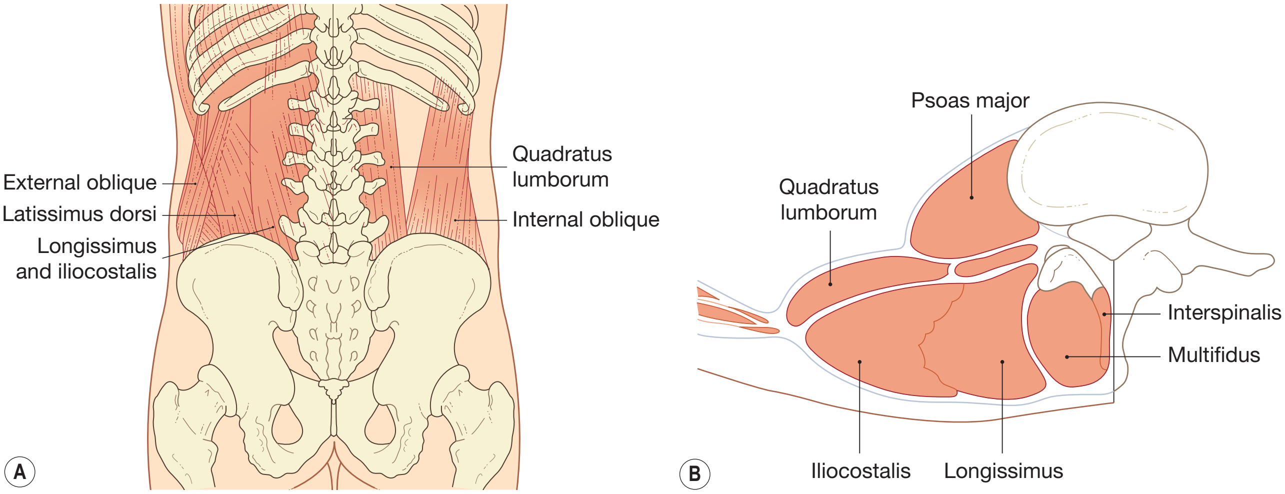
External oblique
Latissimus dorsi
Longissimus
and iliocostalis
Quadratus
lumborum
Internal oblique
A
Psoas major
Quadratus
lumborum
Interspinalis
Multifidus
Iliocostalis
Longissimus
B

Fig. 6.2.1

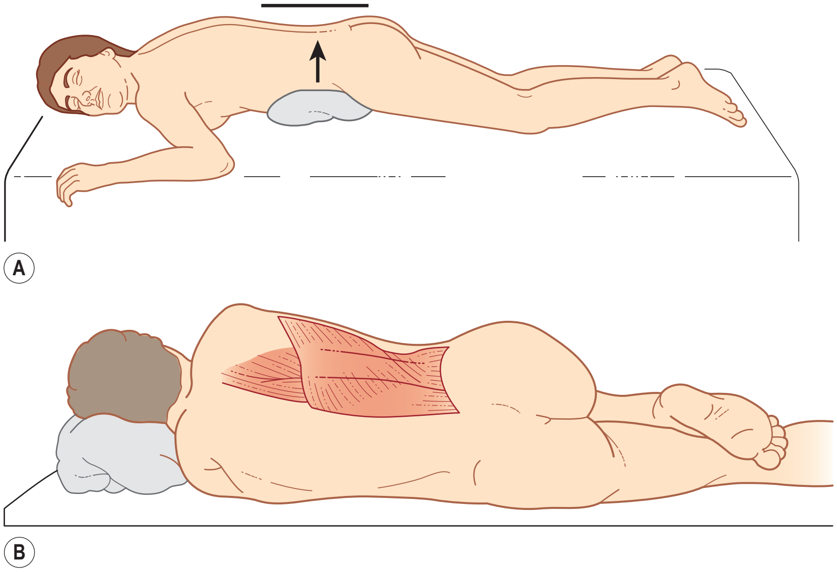

Fig. 6.2.2

Needle puncture and technique

- The midline and the surrounding area are cleaned with antiseptic solution.
- **Trigger points in the muscles are identified by palpation and marked** (*Fig. 6.2.3*).
- A 25 G needle with syringe attached is inserted into a trigger point (*Fig. 6.2.4*).
- After negative aspiration, 2–3 ml of lidocaine (lignocaine) 1% is injected into the trigger point while moving the needle back and forth through the muscle (*Fig. 6.2.5*).

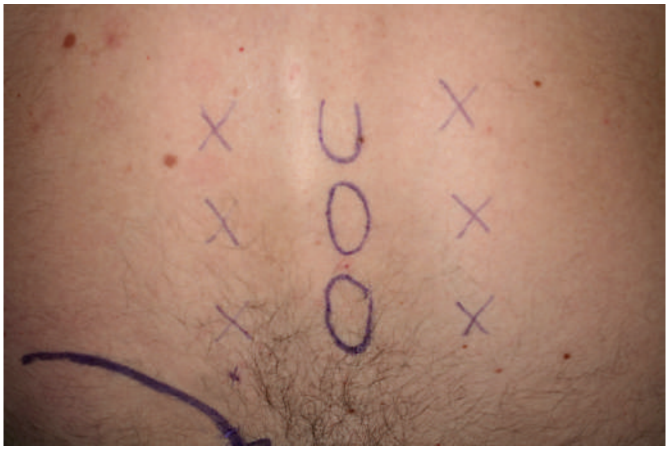

Fig. 6.2.3

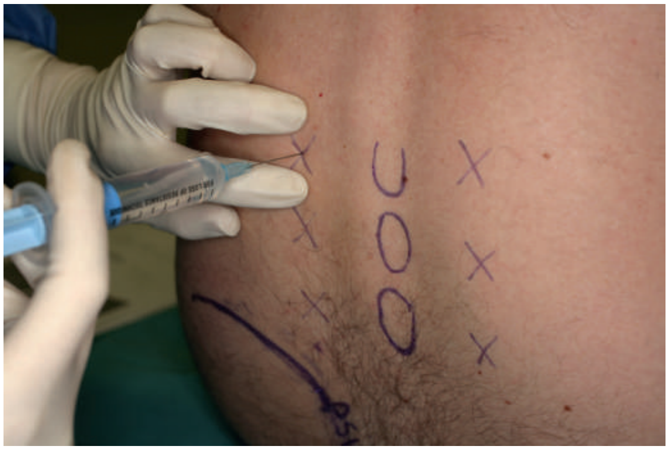

Fig. 6.2.4

- After injection, the next trigger point is injected in the same manner.

Confimation of a successful injection

- Pain reproduction when the needle enters the muscle confirms correct placement.

Tips

- For best results, injection is carried out in a fan-like manner by repeatedly withdrawing the needle slightly and redirecting it.
- Stretching of the involved muscles by physiotherapy within the duration of the local anesthetic improves results.
- Some workers advocate massage of the area immediately after injection.

Potential problems

- Pain on injection.
- Vasovagal response (especially young adults in the sitting position).
- Pneumothorax (especially in thin patients) is also a possibility when injecting the upper back.

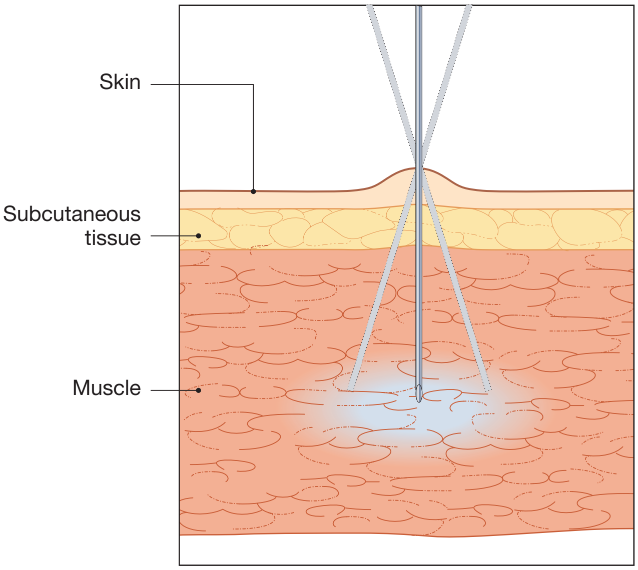

Fig. 6.2.5

6.3 GLUTEUS MEDIUS INJECTION

Anatomy

When the buttock muscles are relaxed the quadratus femoris, gemelli and gluteus medius muscles can be palpated. Spasm of the gluteus medius muscle (**Fig. 6.3.1**) may be the source of buttock pain and may respond to trigger-point injection. Unlike pirformis muscle spasm, this does not produce symptoms of sciatic nerve irritation but causes localized pain, often referred to the posterior thigh and calf.

Equipment

- 10 ml syringe
- 22 G needle

Drugs

- Lidocaine (lignocaine) 1% 10 ml

Position of patient

- Prone.
- Pillow under abdomen to flatten the normal lumbar lordosis (*Fig. 6.3.2*).

Needle puncture and technique

- The surface of the buttock and hip is cleaned with antiseptic solution.
- The posterior superior iliac spine is palpated and marked.
- The greater trochanter is palpated.
- **The insertion point of the needle lies approximately 2 cm medial and superior to the greater trochanter** (*Figs 6.3.3, 6.3.4*).
- A 22 G needle is introduced in a direction vertical to the skin and advanced until it is felt to be gripped by the tense muscle (*Fig. 6.3.5*).

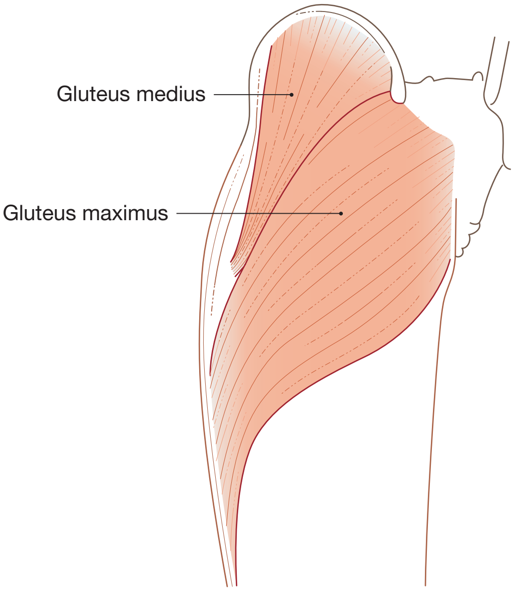

Fig. 6.3.1

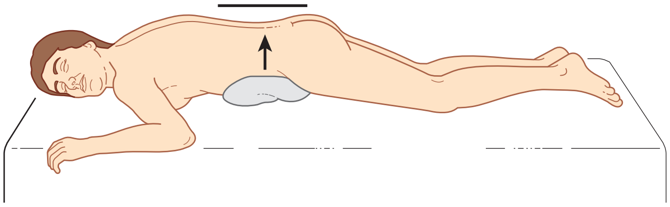

Fig. 6.3.2

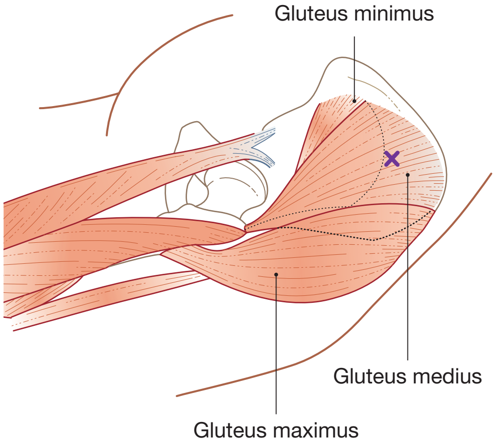

Fig. 6.3.3

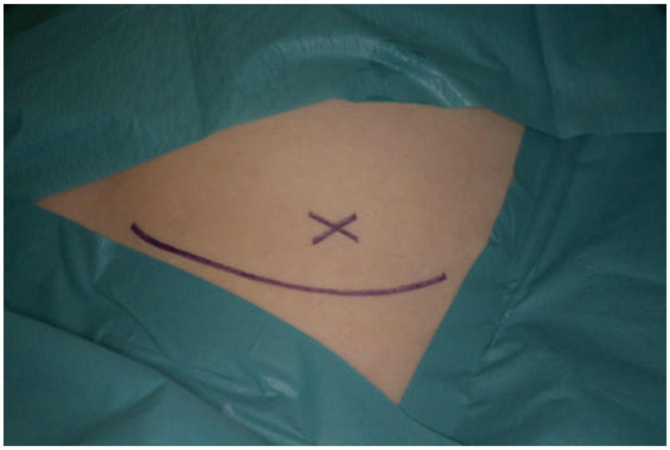

Fig. 6.3.4

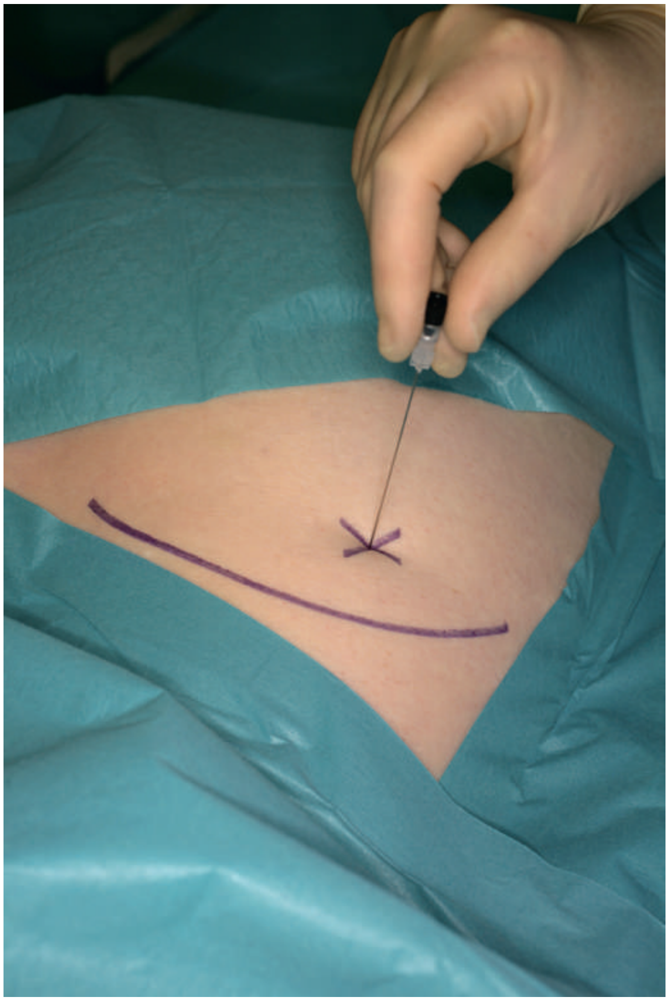

Fig. 6.3.5

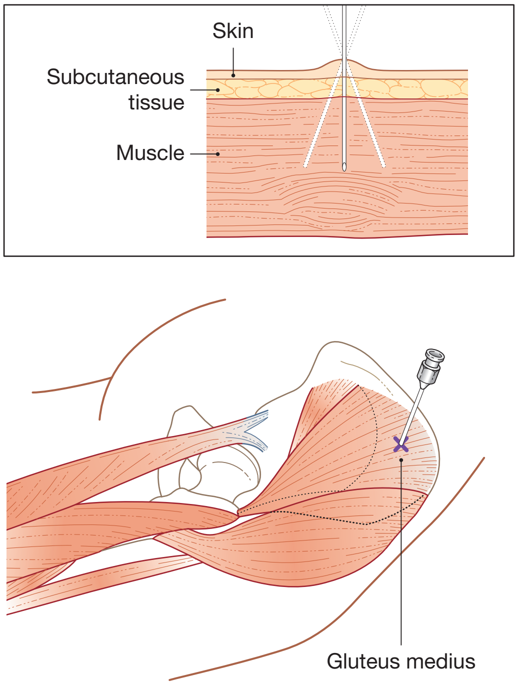

Fig. 6.3.6

- After negative aspiration, lidocaine (lignocaine) 1% 3 ml is injected in the substance of the muscle while moving the needle back and forth in the muscle (*Fig. 6.3.6*).
- The procedure is repeated if other trigger points are present in the muscle.

Confirmation of a successful injection

- Pain reproduction when the needle enters the muscle confirms correct placement of the needle.
- Relief of pain on abduction of the hip.

Tips

- For best results injection is carried out in a fan-like manner by repeatedly withdrawing the needle slightly and redirecting it (*see inset in Fig. 6.3.6*).

Potential problems

- Sciatic nerve block: although this is unusual because the injection site is not very close to the sciatic notch, it is prudent to warn the patient of the possibility.
- Infection or abscess formation.

6.4 PIRIFORMIS INJECTION

Anatomy

The piriformis muscle inserts into the pelvic surface of the sacrum from the second to the fourth segments, lateral to the anterior sacral foramina, and passes out of the pelvis through the greater sciatic foramen to insert into the superior aspect of the greater trochanter (**Fig. 6.4.1**). It overlies the sciatic nerve in the greater sciatic foramen. Contraction contributes to abduction of the lower limb. Spasm of the muscle in myofascial pain syndrome often causes pain referred to the posterior thigh and calf.

Equipment

- 2 ml syringe and two 5 ml syringes
- 25 G needle
- 22 G spinal needle, end-opening

Drugs

- Lidocaine (lignocaine) 1% 10 ml (or its equivalent)

Position of patient

- Prone.
- Pillow under abdomen to flatten the normal lumbar lordosis (*Fig. 6.4.2*).

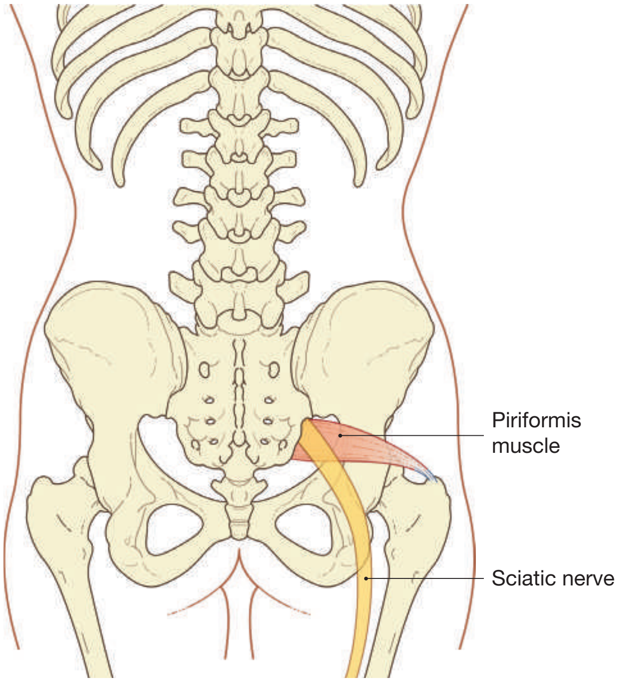

Fig. 6.4.1

Needle puncture and technique

- The surface of the buttock and hip is cleaned with antiseptic solution and a fenestrated drape is placed over the sterile area.
- The posterior superior iliac spine is palpated and marked.
- The greater trochanter is palpated and marked.
- **The insertion points of the needle lie at the points one-third and two-thirds along, and 1–3 cm below the line connecting these two marks** (*Figs 6.4.3 a,b,c*).
- The first insertion point, the medial one, is infiltrated with lidocaine (lignocaine) 1% 2 ml.
- A 22 G spinal needle is introduced in a direction vertical to the skin and advanced until it is felt to be gripped by the tense piriformis muscle, or until bone is contacted (*Figs 6.4.4, 6.4.5*).
- The end-point is a fascial click at a depth of about 4–5 cm, depending on the thickness of adipose tissue.
- The patient is questioned about the presence of pain, paresthesia, and changes in sensation in the distribution of the sciatic nerve, while the needle is being advanced.
- If these symptoms arise, the needle may be in contact with the sciatic nerve and should be repositioned. It is also possible that spasm of the muscle on needle insertion may produce these symptoms and often reproduction of pain occurs with entry into the muscle.
- After negative aspiration, lidocaine (lignocaine) 1% 5 ml is injected.
- Ultrasound may aid placement of the needle (*Fig. 6.4.6 a,b*)

Confirmation of a successful injection

- Relief of pain on abduction of the lower limb against pressure on the lateral knee in the sitting position.

Tips

- If injection is not successful in relieving the pain, it may be repeated at the lateral insertion point. This lies at a

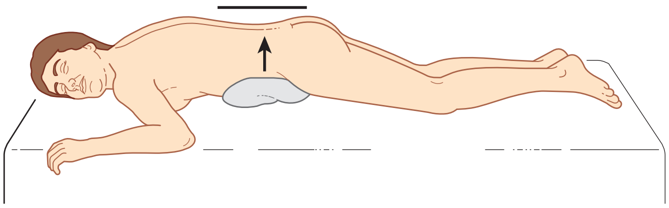

Fig. 6.4.2

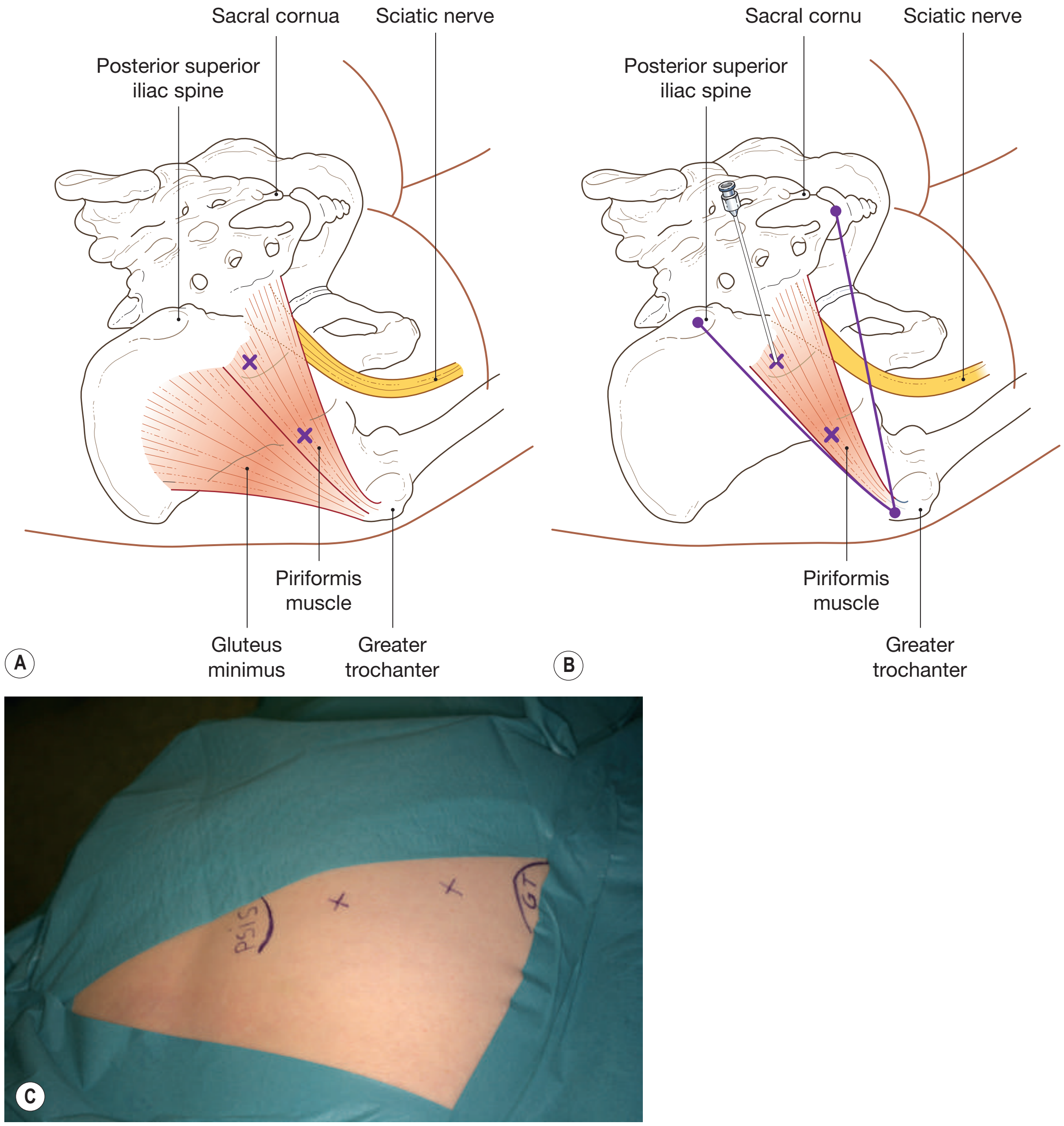
Sacral cornua
Sciatic nerve
Posterior superior iliac spine
Piriformis muscle
Gluteus minimus
Greater trochanter
A
Sacral cornu
Sciatic nerve
Posterior superior iliac spine
Piriformis muscle
Greater trochanter
B
PSIS
GT
C

Fig. 6.4.3

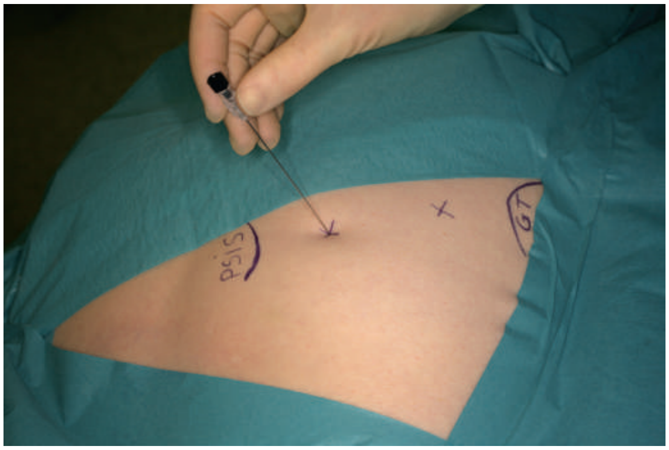

Fig. 6.4.4

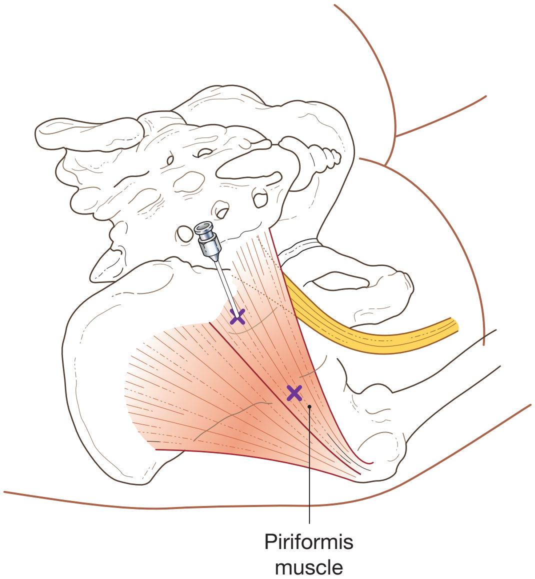
Piriformis muscle

Fig. 6.4.5

point two-thirds along and 1–3 cm below the line joining the posterior superior iliac spine and the greater trochanter.

Potential problems

- Sciatic nerve block.
- Infection or abscess may occur (rarely).

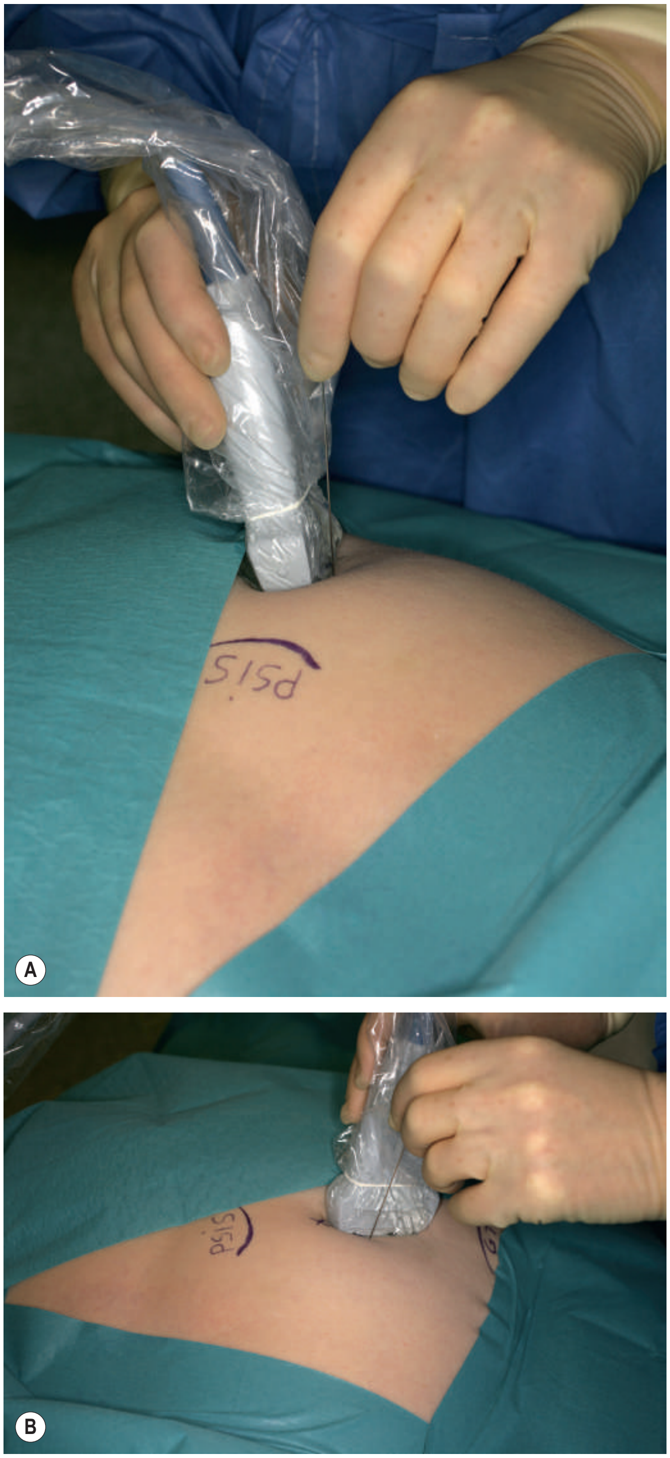
PSIS
A
PSIS
B

Fig. 6.4.6

TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS)

7

Transcutaneous electrical nerve stimulation is thought to modify pain appreciation by stimulation of large fibers thereby blocking (or "closing the gate" to) smaller C-fibers carrying nociceptive impulses. There is also evidence that high-frequency stimulation of the skin increases latency and decreases maximum firing rates in small afferent fibers. This can produce conduction blockade in C-fibers as the current is increased, probably via potassium efflux from the axon. It is thought that a combination of these actions is responsible for the analgesia derived from the use of TENS. This is probably not related to opiate-mediated mechanisms when conventional parameters are used.

Not all pain responds to TENS. If the usual parameters do not produce pain relief, low frequency, high intensity stimulation may be tried. This means that the current amplitude is increased to a level that produces mild discomfort and muscle stimulation. Analgesia from this type of stimulation may be due to opiate-mediated mechanisms. Burst stimulation means short bursts of high frequency stimulation delivered at 1–2 Hz and may also relieve pain that is not responsive to conventional TENS.

A TENS trial may be carried out prior to giving the unit to the patient to use at home. This allows the patient to become familiar with the use of TENS, and to ensure that the pain is not aggravated by its use. A minimum of one hour is recommended as the trial period. This will indicate whether the patient is likely to respond to TENS. However, failure to respond within this time period does not necessarily mean that there will be no response if used for longer periods, or with different settings. It is important to allow the patient to use the TENS at home for a period of at least 14 days.

The TENS stimulator is a battery-operated pulse generator which has several controls. These include an on/off switch plus amplitude control, frequency control, mode selector, and width control. In addition, multichannel units have amplitude controls for each channel. The pulse generator connects to leads that then connect to electrodes, which are applied to the skin. Electrodes are applied in pairs, and are positioned so that they lie along the direction of the nerves in the area being treated, e.g. longitudinally in the limbs, but dermatomally in the trunk.

Control settings (*Figs 7.1.1, 7.1.2*)

CONTINUOUS STIMULATION

- Amplitude set to zero.
- Pulse width set to midrange.
- Switch to continuous mode.
- Increase pulse amplitude level gradually to the maximum level for comfort (sensation should be strong but not painful).
- Adjust pulse frequency to maximum level for comfort (amplitude may be reduced as pulse width is increased).
- Adjust pulse width to maximum level for comfort.
- Maintain for 45–60 minutes.

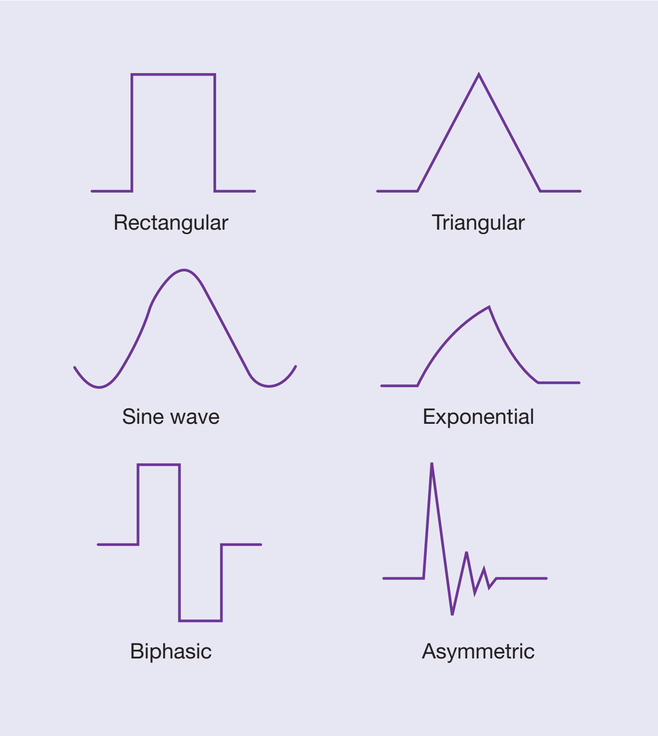

Fig. 7.1.1

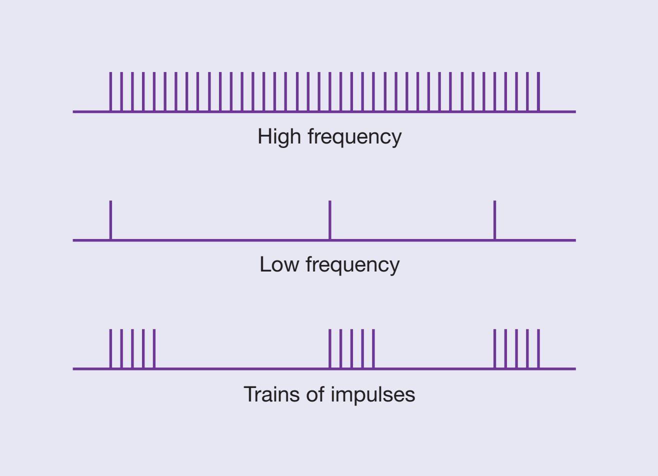

Fig. 7.1.2

Modulated settings

BURST STIMULATION

- All controls set to zero.
- Switch to pulsed mode.
- Frequency set to 1–2 Hz.
- Increase amplitude and pulse width as with continuous mode described above.
- Maintain for 45–60 minutes.

LOW-FREQUENCY HIGH-INTENSITY STIMULATION

- All controls set to zero.
- Increase amplitude to level where the muscle underlying the electrodes twitches visibly but not painfully.
- Increase frequency to 2–4 Hz.
- Maintain for short period (5–15 minutes).

All the types of stimulation should be tried for each pain, and the effects on the pain should be compared. The optimum parameters must be found by trial and error. The patient is usually advised to begin by using TENS for at least one hour three times a day. Once the effect of TENS on the pain is known it is recommended that stimulation should be discontinued after 30 minutes if the patient experiences one or more hours of analgesia from a single application. If pain relief is achieved only during stimulation, the unit can be kept on constantly. However, electrode sites should be changed every 24 hours. Occasionally, skin rash under the electrodes may occur and this problem may be minimized by frequent rotation of the electrode sites and with topical steroids. However, very few side effects are associated with the use of TENS. Electrical skin burns may occur if TENS is applied to skin with poor innervation and it is necessary to ensure that there is normal sensation prior to applying the electrodes. Allergic reaction to the electrodes or the adhesive tape has also been described, but is not common.

Use of TENS is contraindicated on areas over the anterior neck (stimulation of carotid sinus, larynx), over the pregnant uterus or in the presence of a cardiac pacemaker.

Note: Description of the insertion technique of a spinal cord stimulator or peripheral nerve stimulator is outside the scope of this text as the techniques are specific to the different types of stimulator.

APPENDIX 1

SUGGESTED CORTICOSTEROIDS

Drug name	**Duration of action**	**Equivalent dosage (mg)**	**Anti-inflammatory potency (relative)**	**Mineralocorticoid potency (relative)**
Triamcinolone[a]	12–36 h	4	5	0
Methylprednisolone	12–36 h	4	5	0.5
Dexamethasone	48 h	0.75	25	0
Hydrocortisone	12 h	20	1	2

[a] *Triamcinolone diacetate recommended for central neuroaxial injections.*

Corticosteroid injection side effects

LOCAL SIDE EFFECTS

- Atrophy of subcutaneous tissue
- Rupture of injected tendon
- Depigmentation of skin
- Infection

SYSTEMIC SIDE EFFECTS (HIGHER INCIDENCE WITH LARGER DOSES)

- Skin flushing
- Irregularity of the menstrual cycle
- Impaired glucose tolerance
- Osteoporosis
- Muscle wasting and myopathy
- Arthropathy
- Suppression of adrenal function
- Psychologic upset

APPENDIX 2

SUGGESTED NEUROLYTIC AGENTS

- Aqueous phenol 6%.
- Alcohol 100% may be diluted to 50% (pain on injection may be experienced and it is recommended that the nerve is blocked with local anesthetic prior to injection).

APPENDIX 3

RECOMMENDED RESUSCITATION DRUGS AND EQUIPMENT

SUGGESTED RESUSCITATION DRUGS

Drug	Suggested dosage (70 kg adult)	Indication
Atropine	0.2–0.4 mg i.v. increments	Bradycardia from vagal dominance
Ephedrine	5–10 mg i.v. increments	Hypotension from sympathetic block
Lidocaine (lignocaine)	50–100 mg i.v. bolus	Ventricular arrhythmias
Midazolam	1–3 mg i.v. increments	Local anesthetic; seizure activity
Diazepam	2.5–5 mg i.v. increments	Local anesthetic; seizure activity
Thiopental (thiopentone)	50–100 mg i.v. increments	Local anesthetic; seizure activity
Succinylcholine	50–100 mg i.v. bolus	Muscle relaxation; airway control

It is also recommended that the full range of drugs required for advanced cardiac life support (ACLS), including pre-filled syringes, be available in the operating room.

Suggested resuscitation equipment

- Oxygen source
- Bag and masks (full range)
- Breathing system for positive pressure ventilation
- Oro- and nasopharangeal airways (full range)
- Laryngoscopes and blades (full range)
- Endotracheal tube stylets and forceps, e.g. Magill's forceps

APPENDIX 4

DERMATOMES

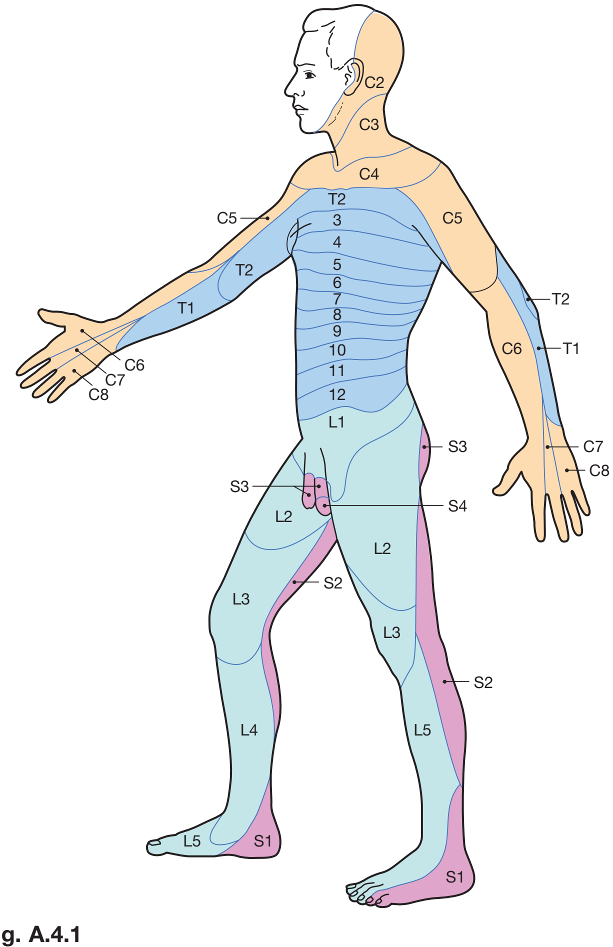
C2
C3
C4
C5
T2
3
4
5
6
7
8
9
10
11
12
L1
C5
T2
T1
C6
C7
C8
T2
C6
T1
C7
C8
S3
S3
S4
L2
L2
S2
L3
L3
S2
L4
L5
L5
S1
S1

Fig. A.4.1

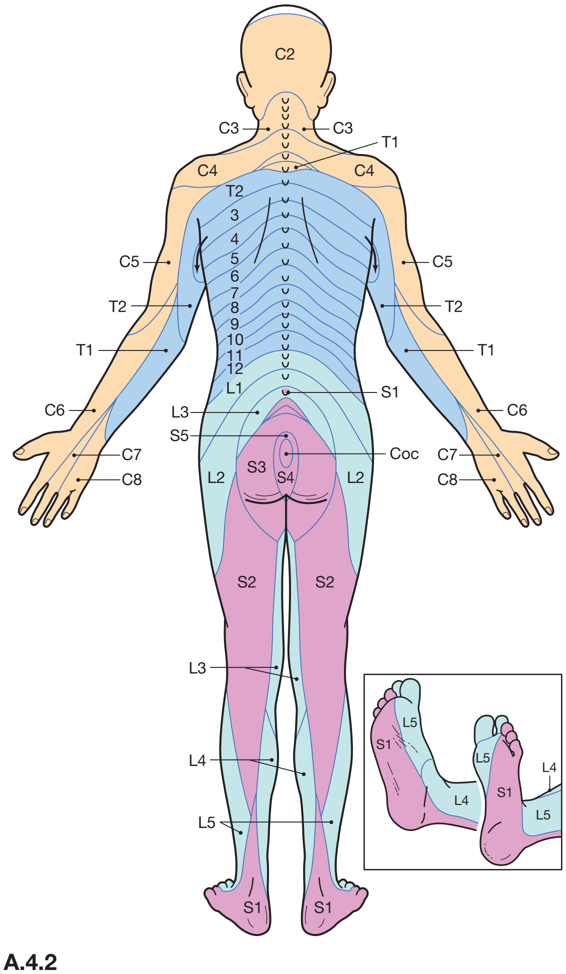
C2
C3
C3
T1
C4
C4
T2
3
4
5
6
7
8
9
10
11
12
L1
C5
C5
T2
T2
T1
T1
S1
C6
C6
L3
S5
Coc
C7
C8
C7
C8
L2
S3
S4
L2
S2
S2
L3
L4
L5
S1
S1
L5
S1
L5
L4
L4
S1
L5

Fig. A.4.2

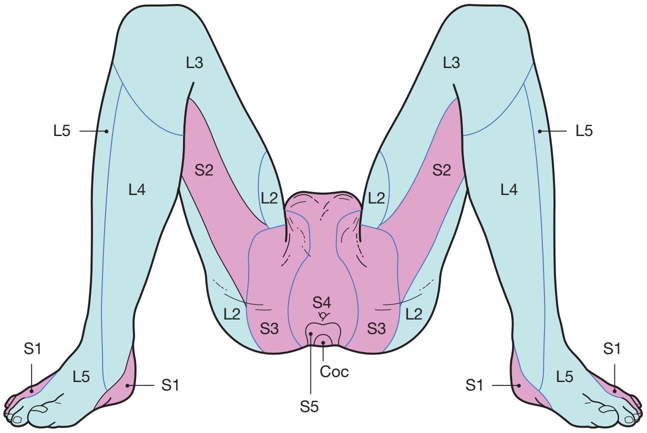
L3
L3
L5
L5
S2
S2
L4
L4
L2
L2
S4
L2
L2
S3
S3
S1
S1
L5
S1
S1
L5
Coc
S5

Fig. A.4.3

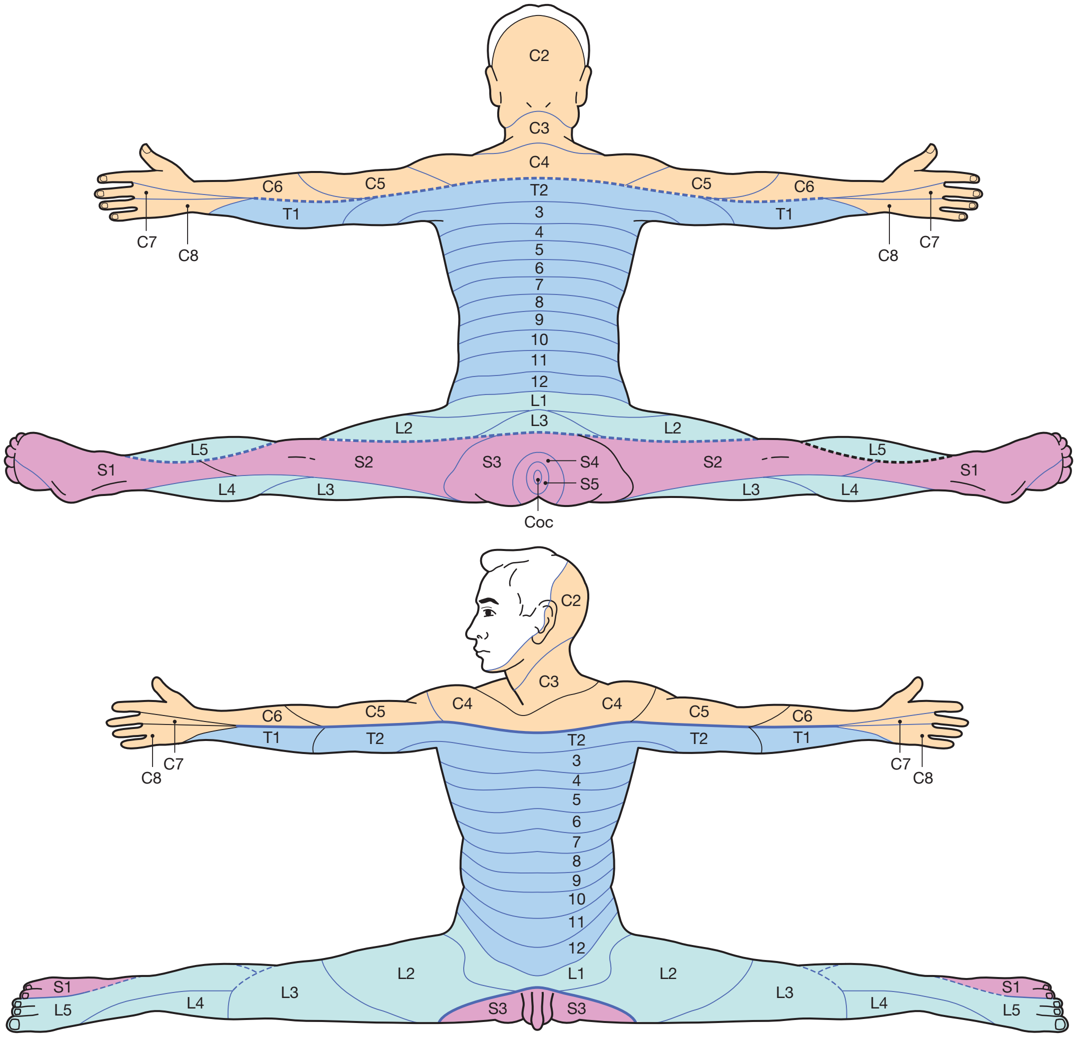
C2
C3
C4
C6
C5
T2
C5
C6
T1
3
T1
C7
C8
4
C8
C7
5
6
7
8
9
10
11
12
L1
L2
L3
L2
L5
L5
S1
S2
S3
S4
S2
S1
S5
L4
L3
L3
L4
Coc
C2
C3
C4
C4
C6
C5
C5
C6
T1
T2
T2
T2
T1
C8
C7
C7
C8
3
4
5
6
7
8
9
10
11
12
L2
L1
L2
S1
L3
L3
S1
L5
L4
S3
S3
L4
L5

Fig. A.4.4

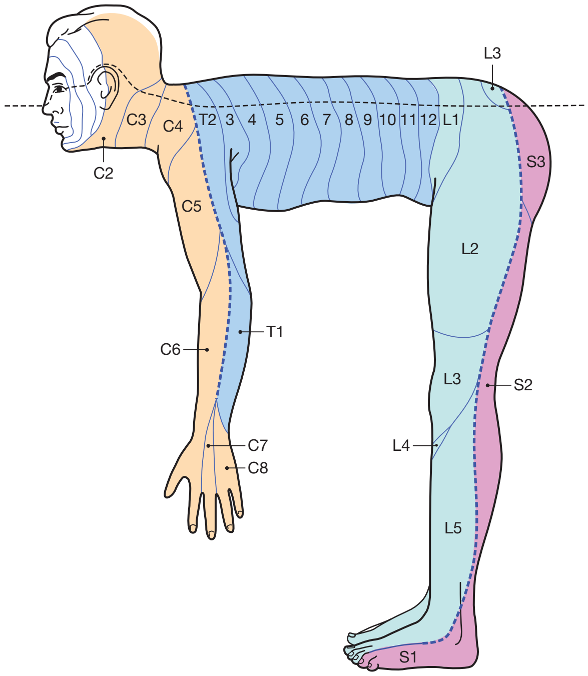
L3
C3
C4
T2
3
4
5
6
7
8
9
10
11
12
L1
C2
S3
C5
L2
T1
C6
L3
S2
C7
C8
L4
L5
S1

Fig. A.4.5

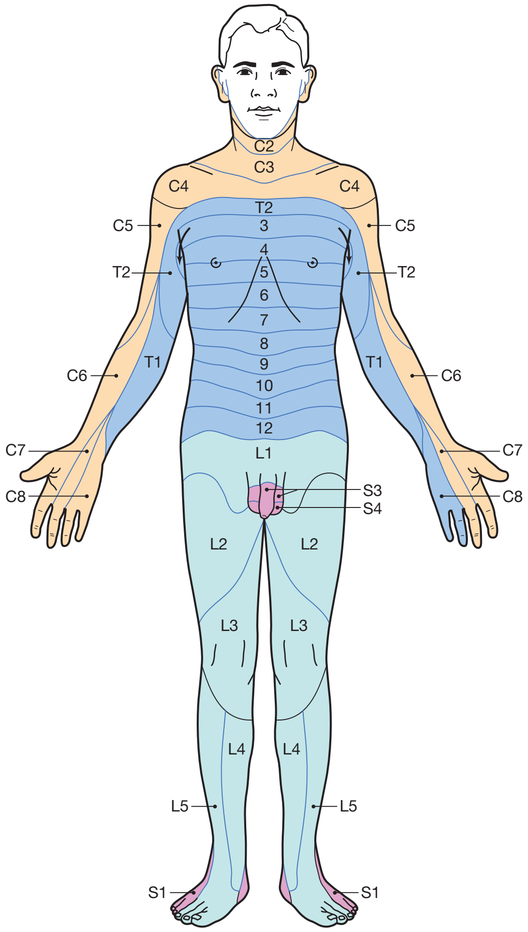
C2
C3
C4
C4
T2
3
C5
C5
4
T2
5
T2
6
7
8
C6
T1
9
T1
C6
10
11
12
C7
C7
L1
S3
C8
S4
C8
L2
L2
L3
L3
L4
L4
L5
L5
S1
S1

Fig. A.4.6

APPENDIX 5

SPINAL CORD SEGMENTAL MYOTOMES

Each muscle in the body is supplied by a particular level or segment of the spinal cord and by its corresponding spinal nerve.

- C5 also supplies the shoulder muscles and the muscle that we use to bend our elbow.
- C6 is for bending the wrist back.
- C7 is for straightening the elbow.
- C8 bends the fingers.
- T1 spreads the fingers.
- T1–T12 supplies the chest wall and abdominal muscles.
- L2 bends the hip.
- L3 straightens the knee.
- L4 pulls the foot up.
- L5 wiggles the toes.
- S1 pulls the foot down.
- S3, S4 and S5 supply the bladder, bowel and sex organs, and the anal and other pelvic muscles.

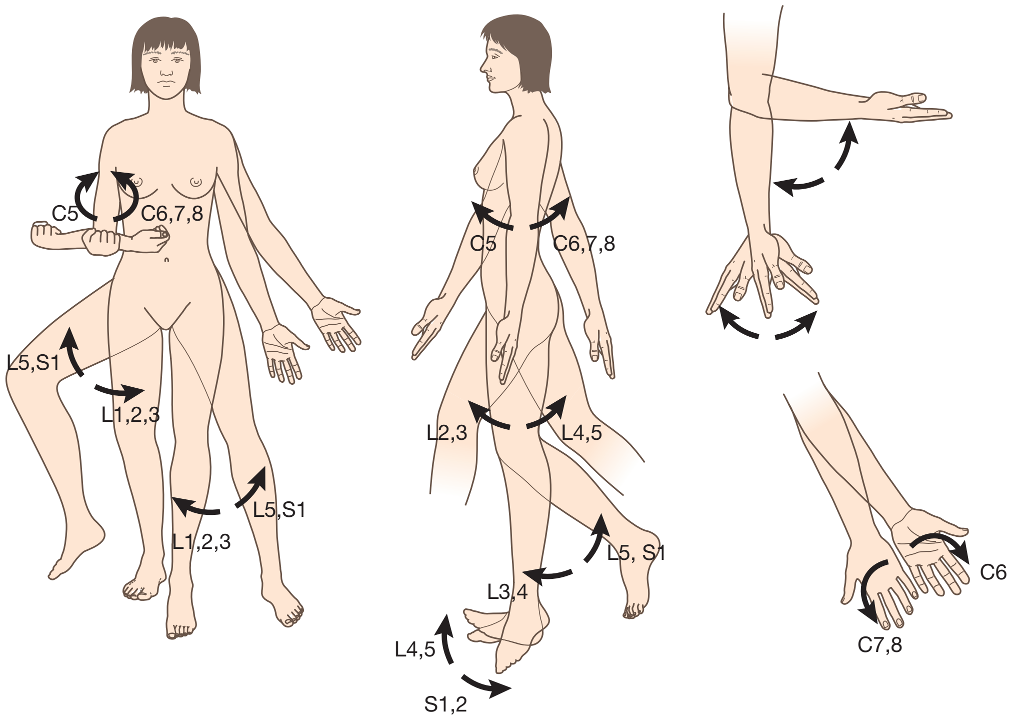

Fig. A.5.1

APPENDIX 6

LUMBO-SACRAL SPINE ANATOMY

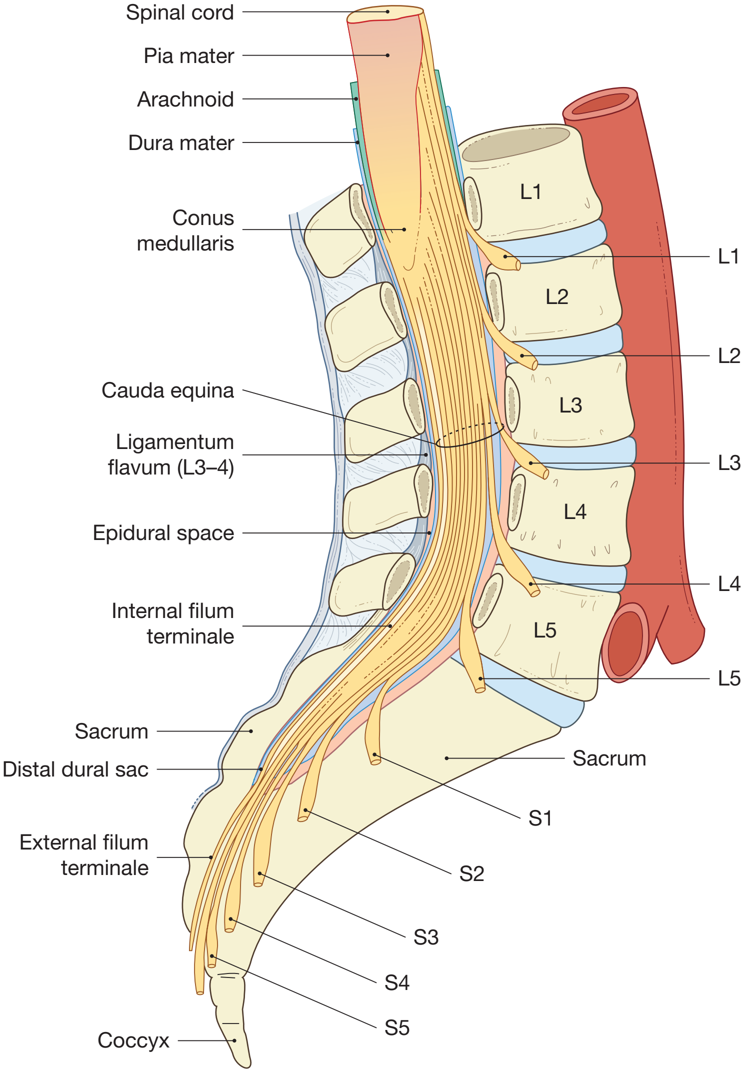
Spinal cord
Pia mater
Arachnoid
Dura mater
Conus medullaris
Cauda equina
Ligamentum flavum (L3–4)
Epidural space
Internal filum terminale
Sacrum
Distal dural sac
External filum terminale
Coccyx
L1
L2
L3
L4
L5
L1
L2
L3
L4
L5
Sacrum
S1
S2
S3
S4
S5

Fig. A.6.1

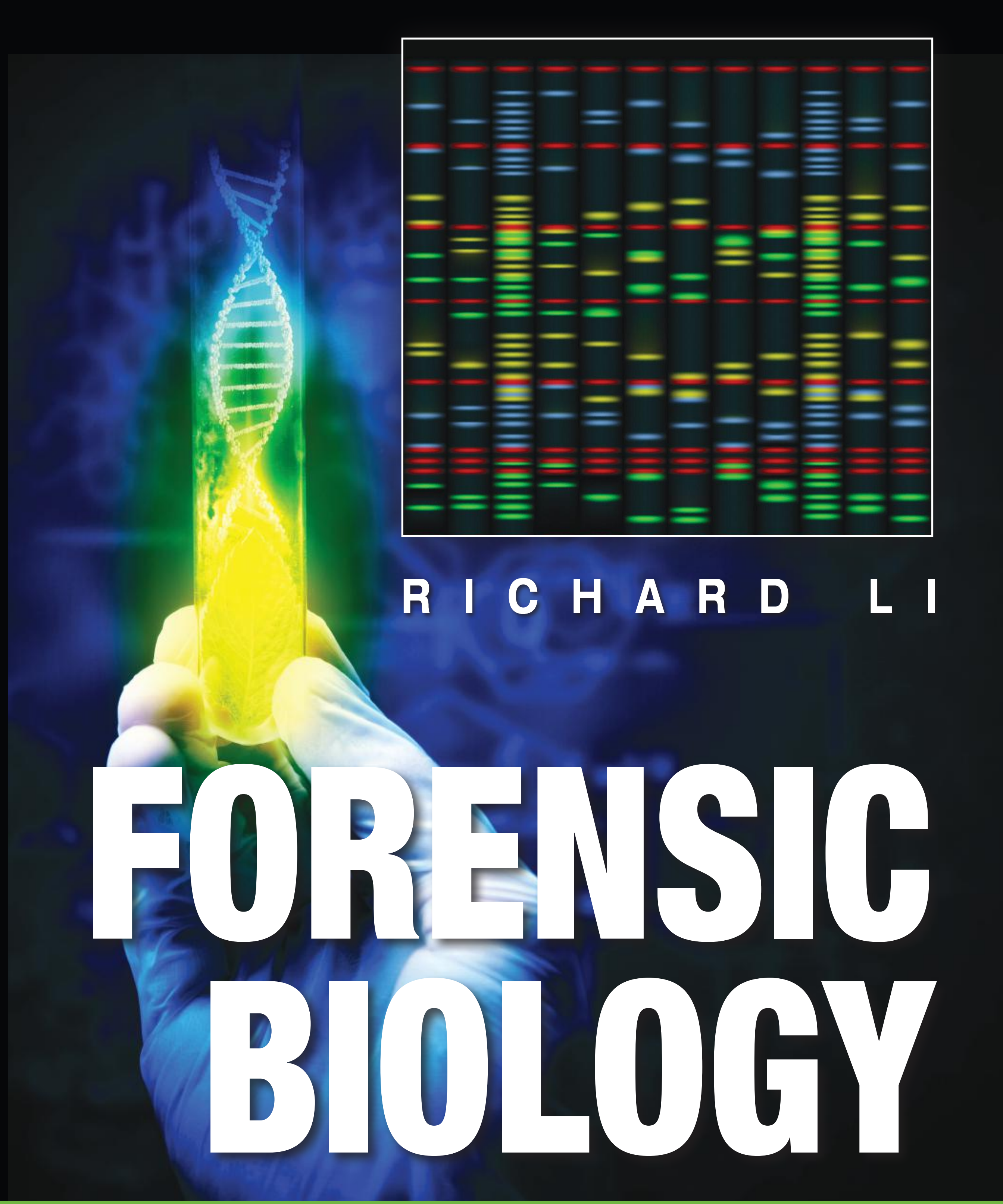

RICHARD LI

FORENSIC BIOLOGY

SECOND EDITION

CRC Press
Taylor & Francis Group

Contents

Preface ... xv
Acknowledgments ... xvii
Author ... xix
Introduction ... xxi

SECTION I Biological Evidence

Chapter 1 Crime Scene Investigation of Biological Evidence ... 3
- 1.1 Protection of Crime Scene ... 3
- 1.2 Recognition of Biological Evidence ... 3
- 1.3 Searches ... 7
- 1.4 Documentation ... 16
- 1.5 Chain of Custody ... 17
- 1.6 Collection of Biological Evidence ... 18
- 1.7 Marking Evidence ... 20
- 1.8 Packaging and Transportation ... 20
- 1.9 Final Survey and the Release of the Crime Scene ... 25
- 1.10 Crime Scene Reconstruction ... 29
- Bibliography ... 30

Chapter 2 Crime Scene Bloodstain Pattern Analysis ... 35
- 2.1 Basic Biological Properties of Human Blood ... 35
- 2.2 Formation of Bloodstains ... 35
- 2.3 Chemical Enhancement and Documentation of Bloodstain Evidence ... 37
- 2.4 Analyzing Spatter Stains ... 39
  - 2.4.1 Velocity of Blood Droplets ... 39
  - 2.4.2 Determining the Directionality of the Stains ... 39
  - 2.4.3 Determining Angles of Impact ... 40
  - 2.4.4 Determining Area of Origin ... 41
- 2.5 Types of Bloodstain Patterns ... 44
  - 2.5.1 Passive Bloodstains ... 44
  - 2.5.2 Transfer Bloodstains ... 46
  - 2.5.3 Projected Bloodstains ... 46
- Bibliography ... 51

Chapter 3 Forensic Biology: A Subdiscipline of Forensic Science 53
3.1 Common Disciplines of Forensic Laboratory Services 53
3.2 Laboratory Analysis of Biological Evidence 53
3.2.1 Identification of Biological Evidence 53
3.2.2 Comparison of Individual Characteristics of Biological Evidence 54
3.2.3 Reporting Results and Expert Testimony 54
3.3 Forensic Science Services Related to Forensic Biology 57
3.3.1 Forensic Pathology 58
3.3.2 Forensic Anthropology 58
3.3.3 Forensic Entomology 60
3.3.4 Forensic Odontology 64
3.4 Brief History of the Development of Forensic Biology 65
3.4.1 Antigen Polymorphism 65
3.4.2 Protein Polymorphism 66
3.4.3 DNA Polymorphism 66
3.4.3.1 Genes and Related Sequences 66
3.4.3.2 Intergenic Noncoding Sequences 66
3.4.3.3 Human DNA Polymorphic Markers 68
3.4.3.4 Forensic DNA Polymorphism Profiling 69
Bibliography 70

Chapter 4 Sources of Biological Evidence 77
4.1 Bodily Fluids 77
4.1.1 Extracellular Nucleic Acids 77
4.2 Cells 81
4.2.1 Cell Surface Markers 81
4.2.2 Nucleated Cells 81
4.2.3 Mitochondria and Other Organelles 83
4.2.4 Cytosol 84
4.2.4.1 Messenger RNAs 84
4.2.4.2 MicroRNAs 85
4.3 Tissues 87
4.3.1 Skin 87
4.3.1.1 Biology of Skin 87
4.3.1.2 Skin as Source of DNA Evidence 87
4.3.2 Hair 88
4.3.2.1 Biology of Hair 89
4.3.2.2 Hair as Source of DNA Evidence 91
4.3.3 Bone 92
4.3.3.1 Biology of Bone 92
4.3.3.2 Bone as Source of DNA Evidence 94
4.3.4 Teeth 98
4.3.4.1 Biology of Teeth 98
4.3.4.2 Teeth as Source of DNA Evidence 100
Bibliography 102

SECTION II Basic Techniques in Forensic Biology

Chapter 5 Nucleic Acid Extraction....111
5.1 Basic Principles of DNA Extraction....111
5.1.1 Cell and Tissue Disruption....111
5.1.2 Lysis of Cellular and Organelle Membranes....114
5.1.3 Removal of Proteins and Cytoplasmic Constituents....114
5.1.4 Storage of DNA Solutions....115
5.1.5 Contamination....115
5.2 Methods of DNA Extraction....115
5.2.1 Extraction with Phenol–Chloroform....115
5.2.1.1 Cell Lysis and Protein Digestion....115
5.2.1.2 Extraction with Organic Solvents....115
5.2.1.3 Concentrating DNA....115
5.2.2 Extraction by Boiling Lysis and Chelation....117
5.2.2.1 Washing....117
5.2.2.2 Boiling....117
5.2.2.3 Centrifugation....117
5.2.3 Silica-Based Extraction....118
5.2.3.1 Cell Lysis and Protein Digestion....118
5.2.3.2 DNA Adsorption onto Silica....118
5.2.3.3 Washing....119
5.2.3.4 Elution of DNA....119
5.2.4 Differential Extraction....121
5.3 Essential Features of RNA....122
5.4 Methods of RNA Extraction....124
5.4.1 RNA–DNA Coextraction....124
5.4.2 miRNA Extraction....124
Bibliography....127

Chapter 6 DNA Quantitation....133
6.1 Slot Blot Assay....133
6.2 Fluorescent Intercalating Dye Assay....134
6.3 Quantitative PCR Assay....137
6.3.1 Real-Time Quantitative PCR....137
6.3.1.1 TaqMan Method....138
Bibliography....140

Chapter 7 Amplification by Polymerase Chain Reaction....143
7.1 Denaturation and Renaturation of DNA....143
7.2 Basic Principles of Polymerase Chain Reaction....143
7.3 Essential PCR Components....145
7.3.1 Thermostable DNA Polymerases....145
7.3.2 PCR Primers....146
7.3.3 Other Components....147

7.4 Cycle Parameters....147
7.5 Factors Affecting PCR....149
  7.5.1 Template Quality....149
  7.5.2 Inhibitors....149
  7.5.3 Contamination....149
7.6 Reverse Transcriptase PCR for RNA-Based Assays....150
  7.6.1 Reverse Transcription....150
  7.6.2 Oligodeoxynucleotide Priming....151
  7.6.3 Reverse Transcriptase PCR....151
Bibliography....153

Chapter 8 DNA Electrophoresis....159

8.1 Basic Principles....159
8.2 Supporting Matrices....159
  8.2.1 Agarose....160
  8.2.2 Polyacrylamide....161
    8.2.2.1 Polymerization Reaction....161
    8.2.2.2 Cross-Linking Reaction....161
    8.2.2.3 Denaturing Polyacrylamide Electrophoresis....163
8.3 Apparatus and Forensic Applications....163
  8.3.1 Slab Gel Electrophoresis....163
    8.3.1.1 Agarose Gel Electrophoresis....163
    8.3.1.2 Polyacrylamide Gel Electrophoresis....164
  8.3.2 Capillary Electrophoresis....164
  8.3.3 Microfluidic Devices....165
    8.3.3.1 Modular Microfluidic Devices....167
    8.3.3.2 Integrated Microfluidic Devices....168
8.4 Estimation of DNA Size....168
  8.4.1 Relative Mobility....168
  8.4.2 Local Southern Method....170
Bibliography....171

Chapter 9 Detection Methods....175

9.1 Direct Detection of DNA in Gels....175
  9.1.1 Fluorescent Intercalating Dye Staining....175
  9.1.2 Silver Staining....175
9.2 Detection of DNA Probes in Hybridization-Based Assays....177
  9.2.1 Radioisotope Labeled Probes....177
  9.2.2 Enzyme-Conjugated Probe with Chemiluminescence Reporting System....180
  9.2.3 Biotinylation of DNA with Colorimetric Reporting Systems....180
    9.2.3.1 Biotin....180
    9.2.3.2 Enzyme-Conjugated Avidin....180
    9.2.3.3 Reporter Enzyme Assay....181
9.3 Detection Methods for PCR-Based Assays....182
  9.3.1 Fluorescence Labeling....182

9.3.1.1 Fluorescent Dyes 182
9.3.1.2 Labeling Methods....182
9.3.1.3 Fluorophore Detection 185
Bibliography 186

Chapter 10 Serology Concepts 189

10.1 Serological Reagents 189
10.1.1 Immunogens and Antigens....189
10.1.2 Antibodies....189
10.1.2.1 Polyclonal Antibodies....191
10.1.2.2 Monoclonal Antibodies....191
10.1.2.3 Antiglobulins 192
10.2 Strength of Antigen–Antibody Binding 193
10.3 Antigen–Antibody Binding Reactions....193
10.3.1 Primary Reactions 194
10.3.2 Secondary Reactions 194
10.3.2.1 Precipitation....195
10.3.2.2 Agglutination.... 197
Bibliography 198

Chapter 11 Serology Techniques: Past, Current, and Future.... 199

11.1 Introduction to Forensic Serology.... 199
11.1.1 The Scope of Forensic Serology.... 199
11.1.2 Class Characteristics and Individual Characteristics of Biological Evidence 199
11.1.3 Presumptive and Confirmatory Assays....200
11.1.4 Primary and Secondary Binding Assays....200
11.2 Primary Binding Assays 201
11.2.1 Enzyme-Linked Immunosorbent Assay (ELISA) 201
11.2.2 Immunochromatographic Assays 202
11.3 Secondary Binding Assays....204
11.3.1 Precipitation-Based Assays....204
11.3.1.1 Immunodiffusion....204
11.3.1.2 Immunoelectrophoretic Methods....206
11.3.2 Agglutination-Based Assays....208
11.4 DNA Methylation Assays for Bodily Fluid Identification....210
11.5 Forensic Applications of RNA-Based Assays and RNA Profiling.... 212
11.5.1 Messenger RNA-Based Assays....212
11.5.2 MicroRNA-Based Assays....214
11.6 Proteomic Approaches Using Mass Spectrometry for Bodily Fluid Identification 215
11.6.1 Mass Spectrometric Instrumentation for Protein Analysis 215
11.6.2 Analysis Strategies for Protein Identification....216
11.7 Microbial DNA Analysis for Bodily Fluid Identification....217
11.8 Nondestructive Assays for the Identification of Bodily Fluids.... 220
Bibliography 221

SECTION III Identification of Biological Evidence

Chapter 12 Identification of Blood 231

- 12.1 Biological Properties 231
  - 12.1.1 Red Blood Cells....231
  - 12.1.2 White Blood Cells....231
  - 12.1.3 Platelets....231
- 12.2 Presumptive Assays for Identification 232
  - 12.2.1 Mechanisms of Presumptive Assays 232
    - 12.2.1.1 Oxidation–Reduction Reactions.... 234
  - 12.2.2 Colorimetric Assays 234
    - 12.2.2.1 Phenolphthalin Assay.... 234
    - 12.2.2.2 Leucomalachite Green (LMG) Assay 234
    - 12.2.2.3 Benzidine and Derivatives 234
  - 12.2.3 Chemiluminescence and Fluorescence Assays.... 236
    - 12.2.3.1 Luminol (3-Aminophthalhydrazide).... 237
    - 12.2.3.2 Fluorescin 238
  - 12.2.4 Factors Affecting Presumptive Assay Results.... 238
    - 12.2.4.1 Oxidants 238
    - 12.2.4.2 Plant Peroxidases.... 239
    - 12.2.4.3 Reductants.... 239
- 12.3 Confirmatory Assays for Identification 239
  - 12.3.1 Microcrystal Assays.... 239
    - 12.3.1.1 Hemochromagen Crystal Assay.... 239
    - 12.3.1.2 Hematin Crystal Assay.... 240
  - 12.3.2 Other Assays.... 240
- Bibliography 241

Chapter 13 Species Identification 245

- 13.1 General Considerations 245
  - 13.1.1 Types of Antibodies.... 245
  - 13.1.2 Titration of Antibodies 246
  - 13.1.3 Antibody Specificity.... 246
  - 13.1.4 Optimal Conditions for Antigen–Antibody Binding.... 247
- 13.2 Assays 247
  - 13.2.1 Immunochromatographic Assays 247
    - 13.2.1.1 Identification of Human Hemoglobin Protein 247
    - 13.2.1.2 Identification of Human Glycophorin A Protein 248
  - 13.2.2 Double Immunodiffusion Assays....251
    - 13.2.2.1 Ring Assay....251
    - 13.2.2.2 Ouchterlony Assay 251
  - 13.2.3 Crossed-Over Electrophoresis 251
- Bibliography 255

Chapter 14 Identification of Semen 257

14.1 Biological Characteristics 257
14.1.1 Spermatozoa 257
14.1.2 Acid Phosphatase 259
14.1.3 Prostate-Specific Antigen 259
14.1.4 Seminal Vesicle–Specific Antigen 259
14.2 Analytical Techniques for Identifying Semen 260
14.2.1 Presumptive Assays 260
14.2.1.1 Lighting Techniques for Visual Examination of Semen Stains 260
14.2.1.2 Acid Phosphatase Techniques 261
14.2.2 Confirmatory Assays 264
14.2.2.1 Microscopic Examination of Spermatozoa 264
14.2.2.2 Identification of Prostate-Specific Antigen 265
14.2.2.3 Identification of Seminal Vesicle–Specific Antigen 269
14.2.2.4 RNA-Based Assays 270
Bibliography 271

Chapter 15 Identification of Saliva 277

15.1 Biological Characteristics of Saliva 277
15.1.1 Amylases 277
15.2 Analytical Techniques for Identification of Saliva 279
15.2.1 Presumptive Assays 279
15.2.1.1 Visual Examination 279
15.2.1.2 Determination of Amylase Activity 279
15.2.2 Confirmatory Assays 281
15.2.2.1 Identification of Human Salivary α-Amylase 281
15.2.2.2 RNA-Based Assays 285
Bibliography 286

Chapter 16 Identification of Vaginal Secretions and Menstrual Blood 289

16.1 Identification of Vaginal Stratified Squamous Epithelial Cells 289
16.1.1 Lugol's Iodine Staining and Periodic Acid–Schiff Method 290
16.1.2 Dane's Staining Method 293
16.2 Identification of Vaginal Acid Phosphatase 293
16.3 Identification of Vaginal Bacteria 294
16.4 Outlook for Confirmatory Assays of Vaginal Secretions 295
16.5 Menstruation 295
16.5.1 Uterine Cycle 296
16.5.2 Uterine Endometrial Hemostasis 298
16.6 D-dimer Assay 299
16.7 Lactate Dehydrogenase Assay 301
16.8 RNA-Based Assays 302
Bibliography 303

Chapter 17 Identification of Urine, Sweat, Fecal Matter, and Vomitus........................ 307
17.1 Identification of Urine.. 307
17.1.1 Urine Formation.. 307
17.1.2 Presumptive Assays.. 308
17.1.2.1 The Identification of Urea 308
17.1.2.2 Identification of Creatinine 314
17.1.3 Confirmative Assays..315
17.1.3.1 Identification of Tamm–Horsfall Protein.............315
17.1.3.2 Identification of 17-Ketosteroids............................315
17.2 Identification of Sweat..316
17.2.1 Biology of Perspiration...316
17.2.2 Sweat Identification Assays 317
17.3 Identification of Fecal Matter...318
17.3.1 Fecal Formation 319
17.3.2 Fecal Matter Identification Assays 319
17.3.2.1 Macroscopic and Microscopic Examination........ 320
17.3.2.2 Urobilinoids Tests 322
17.3.2.3 Fecal Bacterial Identification................................ 324
17.4 Identification of Vomitus.. 324
17.4.1 Biology of Gastric Fluid.. 324
17.4.2 Vomitus Identification Assays 325
Bibliography 326

SECTION IV Individualization of Biological Evidence

Chapter 18 Blood Group Typing and Protein Profiling 331
18.1 Blood Group Typing..331
18.1.1 Blood Groups...331
18.1.2 ABO Blood Group System..331
18.1.2.1 Biosynthesis of Antigens 333
18.1.2.2 Molecular Basis of the ABO System.......................333
18.1.2.3 Secretors 335
18.1.2.4 Inheritance of A and B Antigens 336
18.1.3 Forensic Applications of Blood Group Typing.................... 336
18.1.4 Blood Group Typing Techniques... 336
18.1.4.1 Lattes Crust Assay.. 336
18.1.4.2 Absorption–Elution Assay.................................... 338
18.2 Forensic Protein Profiling.. 338
18.2.1 Methods 339
18.2.1.1 Matrices Supporting Protein Electrophoresis...... 340
18.2.1.2 Separation by Molecular Weight........................... 340
18.2.1.3 Separation by Isoelectric Point.............................. 341
18.2.2 Erythrocyte Protein Polymorphisms.................................. 342
18.2.2.1 Erythrocyte Isoenzymes 342
18.2.2.2 Hemoglobin 342

18.2.3 Serum Protein Polymorphisms 344
Bibliography 345

Chapter 19 Variable Number Tandem Repeat Profiling 353

19.1 Restriction Fragment Length Polymorphism 353
19.1.1 Restriction Endonuclease Digestion 354
19.1.2 Southern Transfer 356
19.1.3 Hybridization with Probes 357
19.1.3.1 Multilocus Probe Technique 357
19.1.3.2 Single-Locus Probe Technique 358
19.1.4 Detection 359
19.1.5 Factors Affecting RFLP Results 360
19.1.5.1 DNA Degradation 360
19.1.5.2 Restriction Digestion–Related Artifacts 361
19.1.5.3 Electrophoresis and Blotting Artifacts 362
19.2 Amplified Fragment Length Polymorphism 363
Bibliography 364

Chapter 20 Autosomal Short Tandem Repeat Profiling 369

20.1 Characteristics of STR Loci 369
20.1.1 Core Repeat and Flanking Regions 369
20.1.2 Repeat Unit Length 370
20.1.3 Repeat Unit Sequences 370
20.2 STR Loci Commonly Used for Forensic DNA Profiling 371
20.3 Forensic STR Analysis 375
20.3.1 Determining the Genotypes of STR Fragments 376
20.3.2 Interpretation of STR Profiling Results 376
20.3.2.1 Inclusion (Match) 376
20.3.2.2 Exclusion 376
20.3.2.3 Inconclusive Result 376
20.4 Factors Affecting Genotyping Results 376
20.4.1 Mutations 377
20.4.1.1 Mutations at STR Core Repeat Regions 377
20.4.1.2 Chromosomal and Gene Duplications 380
20.4.1.3 Point Mutations 381
20.4.2 Amplification Artifacts 382
20.4.2.1 Stuttering 382
20.4.2.2 Nontemplate Adenylation 383
20.4.2.3 Heterozygote Imbalance 383
20.4.2.4 Allelic Dropout 383
20.4.3 Electrophoretic Artifacts 384
20.4.3.1 Pull-Up Peaks 384
20.4.3.2 Spikes 384
20.5 Genotyping of Challenging Forensic Samples 384
20.5.1 Degraded DNA 384

20.5.2 Low Copy Number DNA Testing 386
20.5.3 Mixtures 386
Bibliography 389

Chapter 21 Sex Chromosome Haplotyping and Gender Identification 407

21.1 Y Chromosome Haplotyping 407
21.1.1 Human Y Chromosome Genome 407
21.1.1.1 Pseudoautosomal Region 407
21.1.1.2 Male-Specific Y Region 407
21.1.1.3 Polymorphic Sequences 408
21.1.2 Y-STR 409
21.1.2.1 Core Y-STR Loci 410
21.1.2.2 Multilocal Y-STR Loci 411
21.1.2.3 Rapidly Mutating Y-STR 411
21.2 X Chromosome Haplotyping 412
21.3 Sex Typing for Gender Identification 416
21.3.1 Amelogenin Locus 416
21.3.2 Other Loci 418
Bibliography 419

Chapter 22 Single Nucleotide Polymorphism Profiling 437

22.1 Basic Characteristics of SNPs 437
22.2 Forensic Applications of SNP Profiling 438
22.2.1 *HLA-DQA1* Locus 438
22.2.1.1 DQα AmpliType and Polymarker Assays 438
22.2.1.2 Allele-Specific Oligonucleotide Hybridization 440
22.2.2 Current and Potential Applications of SNP Analysis 443
22.2.2.1 Application of SNP Analysis for Forensic Identification 443
22.2.2.2 Potential Applications of SNPs for Phenotyping 444
22.3 SNP Techniques 444
22.3.1 Next-Generation Sequencing Technologies 444
22.3.2 DNA Samples, Sequencing Library, and Template Preparation 446
22.3.3 NGS Chemistry 447
22.3.4 NGS Coverage 447
Bibliography 451

Chapter 23 Mitochondrial DNA Profiling 461

23.1 Human Mitochondrial Genome 461
23.1.1 Genetic Contents of Mitochondrial Organelle Genomes 461
23.1.2 Maternal Inheritance of mtDNA 463
23.2 mtDNA Polymorphic Regions 463
23.2.1 Hypervariable Regions 463
23.2.2 Heteroplasmy 463

23.3 Forensic mtDNA Testing 465
    23.3.1 General Considerations 465
    23.3.2 mtDNA Screen Assay 466
    23.3.3 mtDNA Sequencing 467
        23.3.3.1 PCR Amplification 467
        23.3.3.2 DNA Sequencing Reactions 468
        23.3.3.3 Electrophoresis, Sequence Analysis, and Mitotype Designations 471
    23.3.4 Interpretation of mtDNA Profiling Results 472
        23.3.4.1 Exclusion 474
        23.3.4.2 Cannot Exclude 474
        23.3.4.3 Inconclusive Result 474
Bibliography 474

SECTION V Forensic Issues

Chapter 24 Forensic DNA Databases: Tools for Crime Investigations 485

24.1 Brief History of Forensic DNA Databases 485
24.2 Infrastructure of CODIS 485
24.3 Indexes of CODIS 487
24.4 Database Entries 489
24.5 Database Expansion 490
24.6 DNA Profiles 491
24.7 Routine Database Searches for Forensic Investigations 492
    24.7.1 Case-to-Offender Searches 492
    24.7.2 Case-to-Case Searches 493
    24.7.3 Search Stringency and Partial Matches 493
24.8 Familial Searches 495
    24.8.1 Legal and Ethical Issues of Familial Search 495
    24.8.2 Familial Search Strategies 496
        24.8.2.1 Identity-by-State and Kinship Index Method 496
        24.8.2.2 Focusing on Rare Alleles 496
        24.8.2.3 Excluding Candidates through Y-STR Screening 498
Bibliography 498

Chapter 25 Evaluation of the Strength of Forensic DNA Profiling Results 503

25.1 A Review of Basic Principles of Genetics 503
    25.1.1 Mendelian Genetics 503
    25.1.2 Population Genetics 504
        25.1.2.1 Allele Frequency 504
        25.1.2.2 Genotype Frequency 504
        25.1.2.3 Heterozygosity 504
        25.1.2.4 Hardy–Weinberg Principle 505
        25.1.2.5 Testing for HW Proportions of Population Databases 505
        25.1.2.6 Probability of Match 511

25.2 Statistical Analysis of DNA Profiling Results....512
  25.2.1 Genotypes....513
    25.2.1.1 Profile Probability....513
    25.2.1.2 Likelihood Ratio....516
  25.2.2 Haplotypes....516
    25.2.2.1 Mitotypes Observed in Database....518
    25.2.2.2 Mitotype Not Observed in Database....518
Bibliography....518

Chapter 26 Quality Assurance and Quality Control....523

26.1 US Quality Standards....523
26.2 International Quality Standards....524
26.3 Laboratory Accreditation....525
26.4 Laboratory Validation....525
26.5 Proficiency Testing....526
26.6 Certification....526
26.7 Forensic DNA Analyst Qualifications....527
26.8 Code of Ethics of Forensic Scientists....528
Bibliography....529

Index....533

1

Crime Scene Investigation of Biological Evidence

A forensic investigation involving biological evidence usually begins at the crime scene. The crime scene investigation process includes maintaining scene security, preparing documentation, and collecting and preserving physical evidence. A crime scene investigation requires teamwork and effort. Each team member should be assigned specific tasks (Figure 1.1).

1.1 Protection of Crime Scene

A crime scene investigation begins with the initial response to a scene (Figure 1.2). Securing and protecting the scene are important steps in a crime scene investigation (Figures 1.3 through 1.6), and this task is usually carried out by the first responding officer arriving on the scene. The entry of authorized personnel admitted to the scene should be documented using a log sheet (Figure 1.7). Suspects, witnesses, and living victims should be evacuated from the scene. If a victim is wounded, medical attention should be sought.

Appropriate supplies and devices should be used to prevent the contamination of evidence by investigators. Protective wear and devices including a face mask or shield, safety eyeglasses, a disposable coverall bodysuit, gloves, shoe covers, and a hairnet should be used (Figures 1.8 and 1.9). Exposure to bodily fluids may occur during a crime scene investigation. An investigator can be exposed to bodily fluids through the mucous membranes, skin exposure, and needlestick injuries (especially when investigating a clandestine drug laboratory scene). Therefore, biosafety procedures must be followed for the protection of personnel from infectious blood-borne pathogens such as the human immunodeficiency virus (HIV), hepatitis B virus (HBV), and hepatitis C virus (HCV); infectious aerosol tuberculosis pathogens; and other biohazardous materials.

1.2 Recognition of Biological Evidence

A preliminary survey should be carried out to evaluate potential evidence. In particular, the recognition of evidence plays a critical role in solving or prosecuting crimes. The priority of the potential evidence at crime scenes should be assessed based on each item's relevance to the solution of the case. Higher priority should be assigned to evidence with probative value to the case. For example, the evidence related to a *corpus delicti* is considered to be of the highest priority. *Corpus delicti* is a Latin term meaning "body of crime." In Western law, it primarily refers to the principle that in order for an individual to be convicted, it is necessary to prove the occurrence

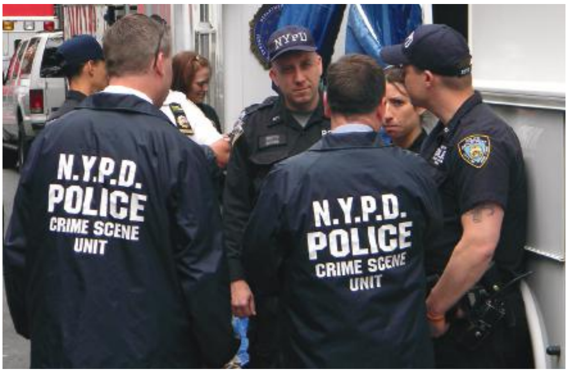
N.Y.P.D.
POLICE
CRIME SCENE
UNIT

Figure 1.1 A crime scene investigation team. (© Richard C. Li.)

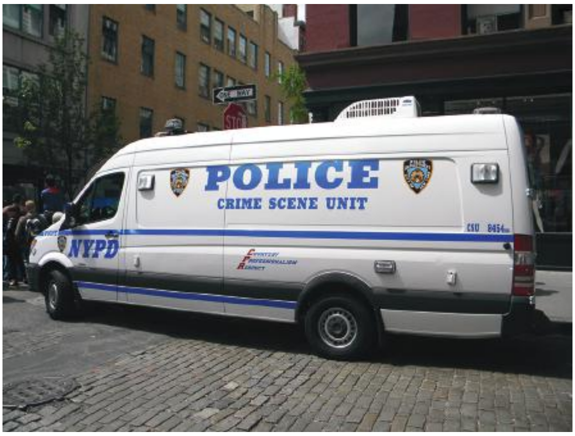
POLICE
CRIME SCENE UNIT
NYPD

Figure 1.2 A crime scene unit vehicle that is used to respond to crime scenes. This type of vehicle is usually outfitted with devices and supplies that investigators need when processing a crime scene, as well as evidence packaging materials, fingerprint collection kits, and DNA collection kits. Additionally, it can be equipped with a workstation for computer access, a refrigerator for storing chemicals, and a compact fuming hood for processing latent fingerprints, as well as equipment cabinets and drawers. (© Richard C. Li.)

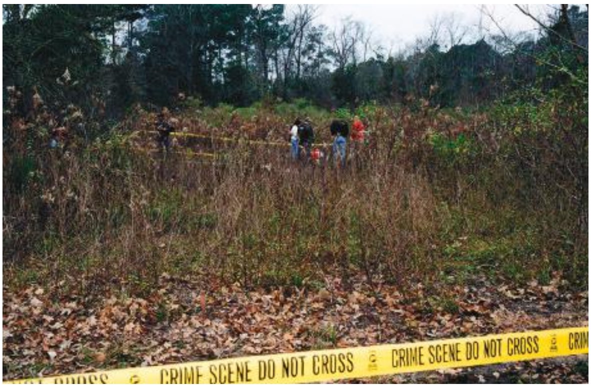
CRIME SCENE DO NOT CROSS

Figure 1.3 Crime scene barrier tape is used to ensure that only investigators are admitted to the scene. (Courtesy of H. Brewster.)

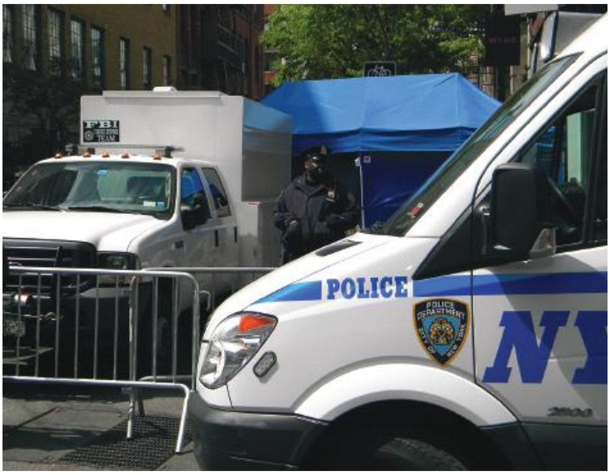
FBI
POLICE
POLICE DEPARTMENT CITY OF NEW YORK
NY

Figure 1.4 A police officer guards the crime scene. (© Richard C. Li.)

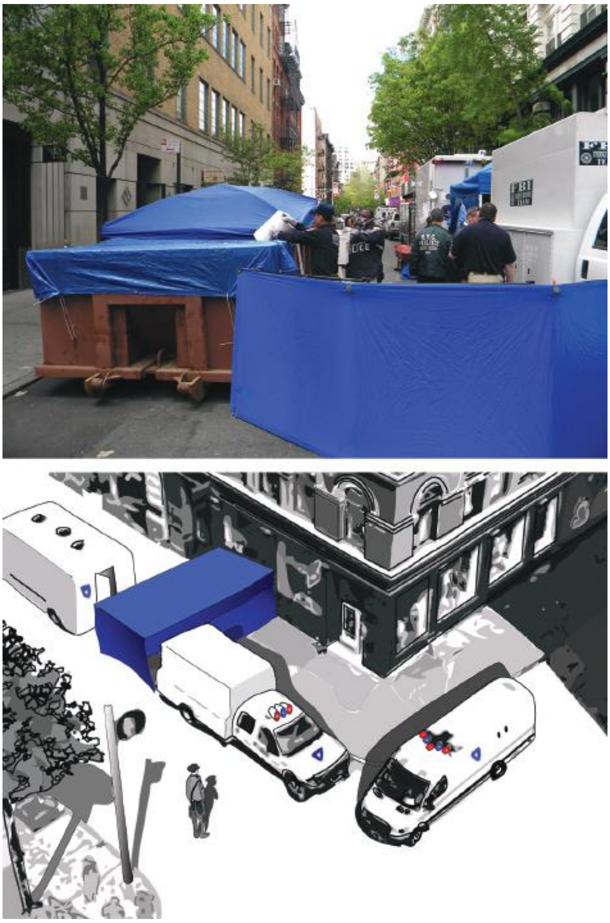

Figure 1.5 Crime scene privacy screen and tent. The screen (top) and the tent (bottom) are useful devices for shielding the evidence or body from viewing by unauthorized personnel. (© Richard C. Li.)

Figure 1.6 Barricades are set up to keep crowds at a distance from the scene. (© Richard C. Li.)

of the crime. In a forensic investigation, it also refers to the physical evidence proving that a crime was committed. For example, when an individual is missing, a missing persons investigation is usually initiated. If *corpus delicti*, such as a dead body or a victim's blood at a crime scene (Figure 1.10), is discovered during the investigation, a homicide case can be established and a suspect can be charged with homicide. Higher priority should also be attached to evidence that can establish connections such as *victim-to-perpetrator linkage*. For example, items found in a perpetrator's possession may be linked to a victim. This also applies to transfer evidence based on the principles of transfer theory, also known as the *Locard exchange principle*, which theorizes that the cross-transfer of evidence occurs when a perpetrator has any physical contact with an object or another person (Figures 1.11 through 1.14). Thus, trace evidence, such as hairs and fibers, may be transferred from a perpetrator to a victim or vice versa. This explains why it is important to ensure that perpetrators and their belongings are thoroughly searched for trace evidence. Likewise, victims and their belongings should be examined for the same reason.

Victim-to-scene and perpetrator-to-scene linkages can also be established. Blood belonging to a perpetrator or a victim found at a crime scene can establish such a linkage (Figure 1.15). Additionally, reciprocal transfers of trace evidence from crime scenes can be used to link a suspect or a victim to a crime scene. A perpetrator may present a unique *modus operandi* (MO). *Modus operandi*, a Latin term commonly used in criminal investigations, refers to a particular pattern of characteristics and the manner in which a crime is committed. For example, Richard Cottingham, a serial killer known as "the torso killer," dismembered his victims and took their limbs and heads with him but left their torsos at the scene. He then set the rooms on fire before

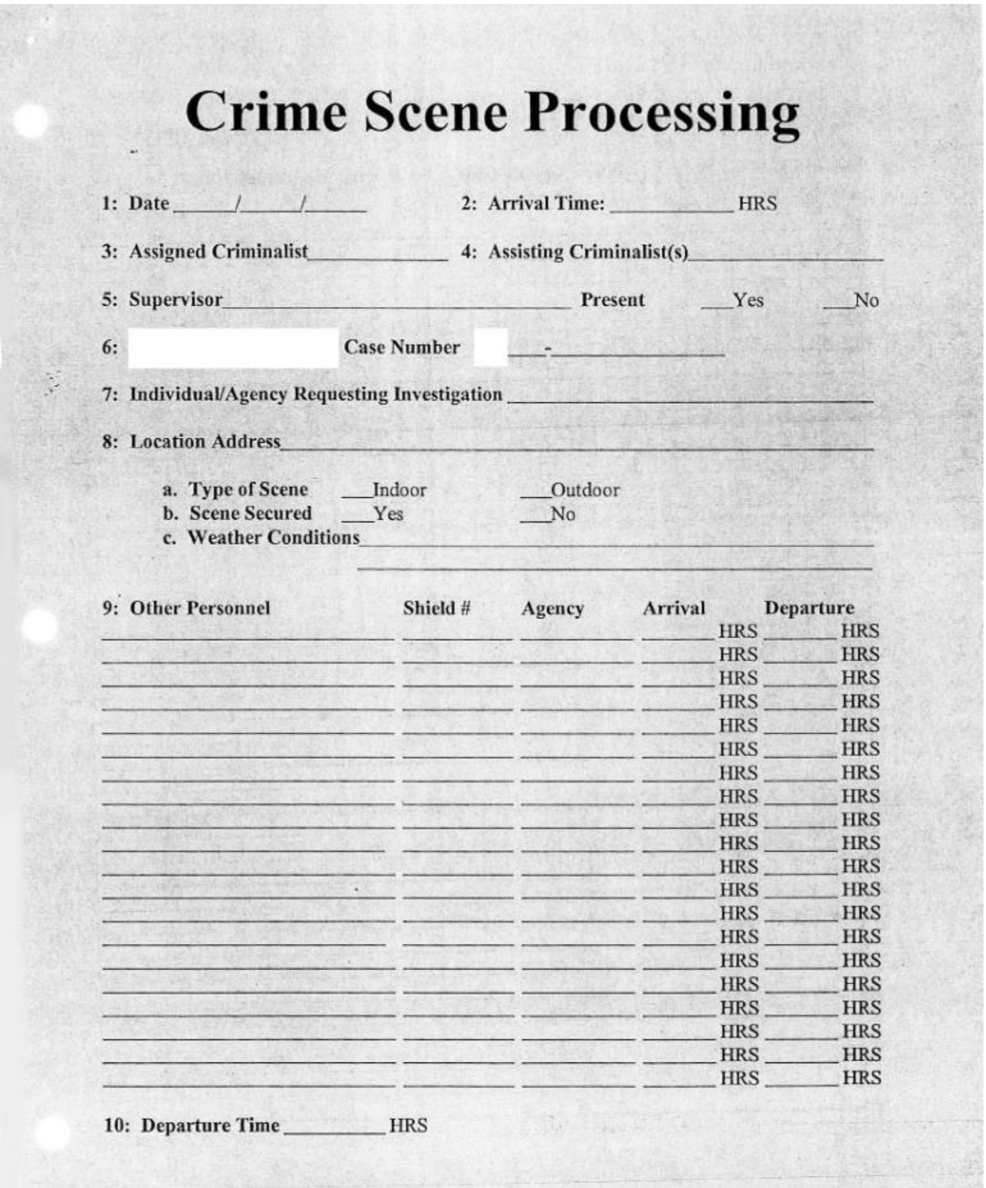

Crime Scene Processing

1: Date ______/______/______ **2: Arrival Time:** ____________ HRS

3: Assigned Criminalist ______________ **4: Assisting Criminalist(s)** ________________

5: Supervisor __________________________________ **Present** ___Yes ___No

6: [redacted] **Case Number** [redacted] ____-______________

7: Individual/Agency Requesting Investigation ______________________________

8: Location Address ____

- **a. Type of Scene** ___Indoor ___Outdoor
- **b. Scene Secured** ___Yes ___No
- **c. Weather Conditions** ____

9: Other Personnel	**Shield #**	**Agency**	**Arrival**	**Departure**
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS
			______ HRS	______ HRS

10: Departure Time __________ HRS

Figure 1.7 An example of a log sheet for documenting authorized personnel at a crime scene. (© Richard C. Li.)

fleeing the scenes. Evidence that provides information on the MO is also vital to an investigation. A distinct MO can establish a *case-to-case linkage* for serial offender cases.

1.3 Searches

Some investigations require a search for specific items of evidence such as biological stains, human remains, and all relevant evidence. A search usually has a specific purpose. Thus, the use of search patterns can be helpful, especially in cases involving large outdoor crime scenes.

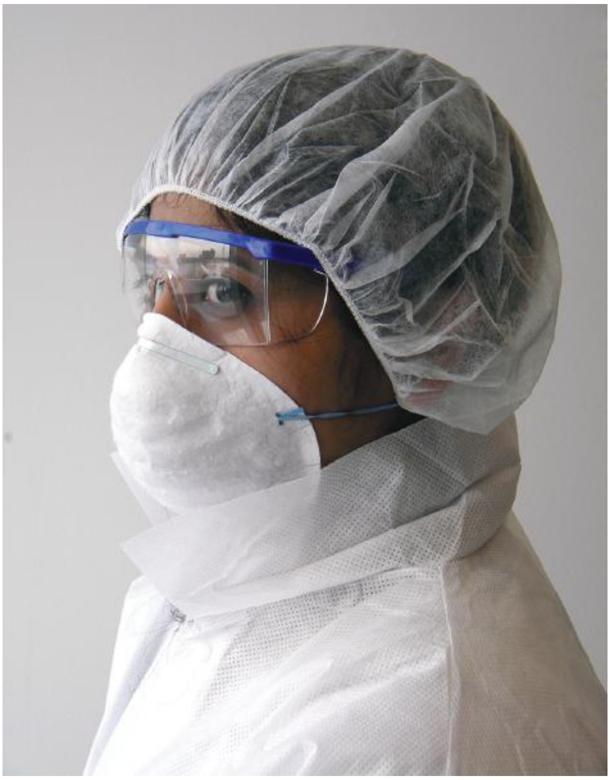

Figure 1.8 Personal protection wear and devices that are used at crime scenes. (© Richard C. Li.)

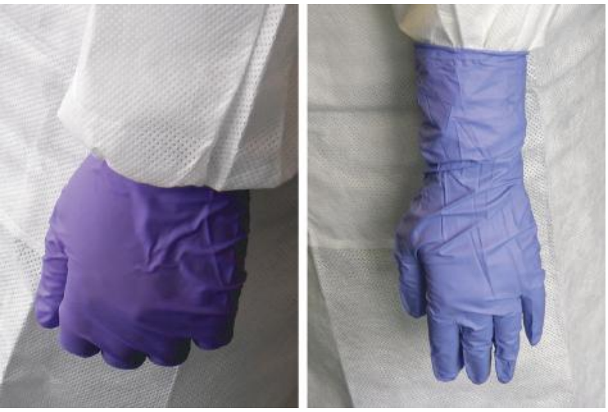

Figure 1.9 Disposable glove (left) and glove with extended cuff (right). (© Richard C. Li.)

Search patterns may include a grid, line, or zone (Figures 1.16 and 1.17). The method that is ultimately used depends on the type and size of the scene (Figures 1.18 through 1.20). Additionally, the points of entry and exit and the paths followed by a perpetrator should also be searched.

Searching for biological stains usually utilizes devices such as an alternate light source (ALS); see Figures 1.21 and 1.22. An ALS either produces a single specific wavelength of light or a desired wavelength by using specific filters. Biological materials such as blood, semen, and saliva emit fluorescent light under an ALS, which can facilitate the locating of biological materials. Additionally,

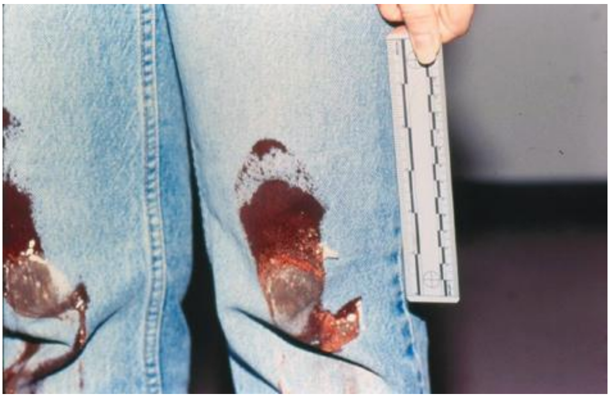

Figure 1.10 Photographic documentation of bloodstains on clothing. (© Richard C. Li.)

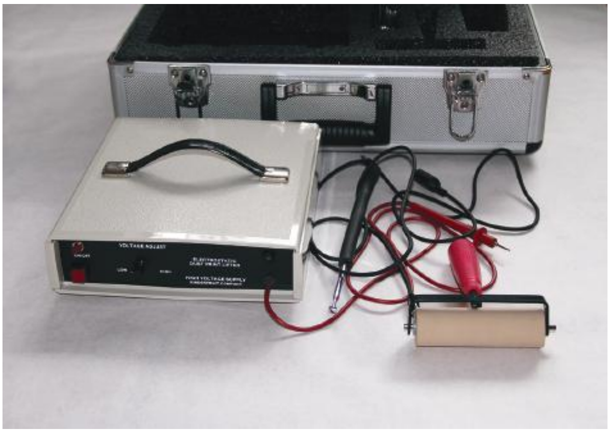

Figure 1.11 An electrostatic dust print lifting device can be utilized for processing impression evidence such as footprints and tire tracks. (© Richard C. Li.)

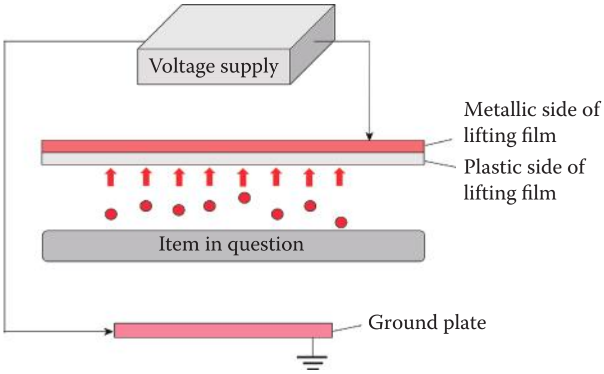
Voltage supply
Metallic side of lifting film
Plastic side of lifting film
Item in question
Ground plate

Figure 1.12 Basic components of an electrostatic dust print lifting device. The lifting film is placed on top of the item in question with the plastic side against the surface and the metallic side facing up. A ground plate is placed directly on the ground. The lifting film and ground plate are connected to the voltage supply apparatus. Once the charging voltage is turned on, the static charge transfers the dust particles from the surface to the plastic side of the lifting film. (© Richard C. Li.)

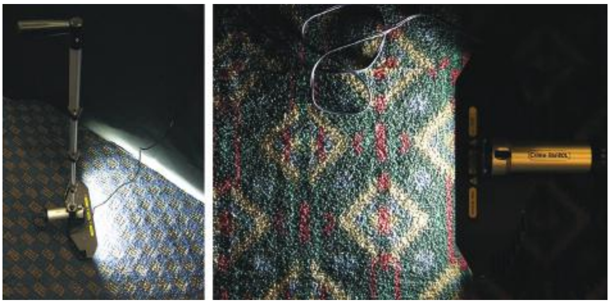

Figure 1.13 A high-intensity light-emitting diode (LED) device for locating evidence at a crime scene is particularly effective in highlighting trace evidence such as hairs, fibers, and shoe prints. (© Richard C. Li.)

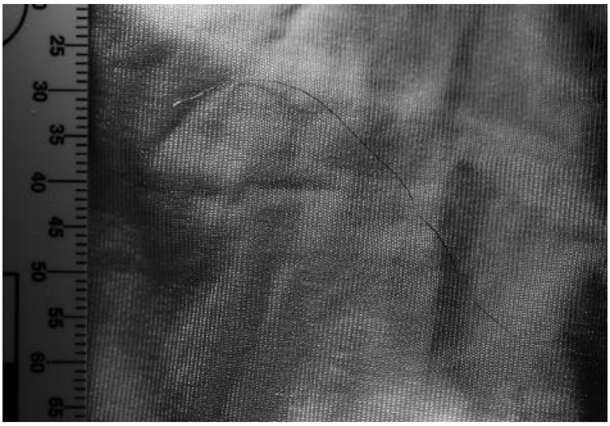

Figure 1.14 A hair found on a victim's clothing can be transferred evidence from a suspect. (© Richard C. Li.)

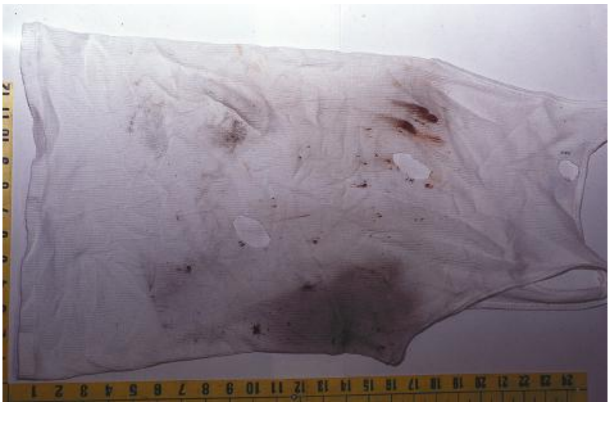

Figure 1.15 Finding a victim's blood on a suspect's clothing can establish a link between them. (© Richard C. Li.)

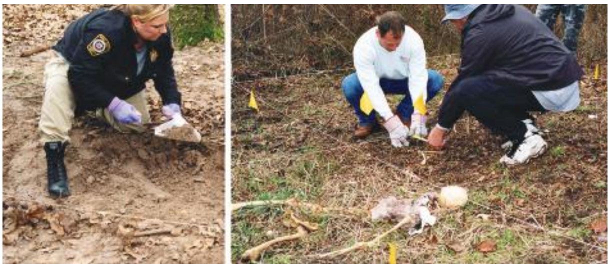

Figure 1.16 Grid search pattern for an outdoor scene. The investigators and anthropologists present are searching for human bone evidence within the grid. (Courtesy of H. Brewster.)

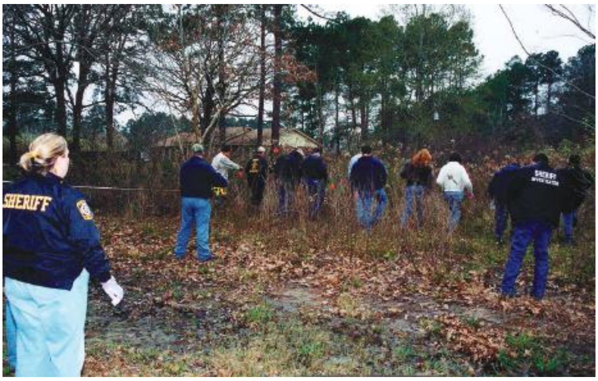

Figure 1.17 Line search pattern for an outdoor scene. (Courtesy of H. Brewster.)

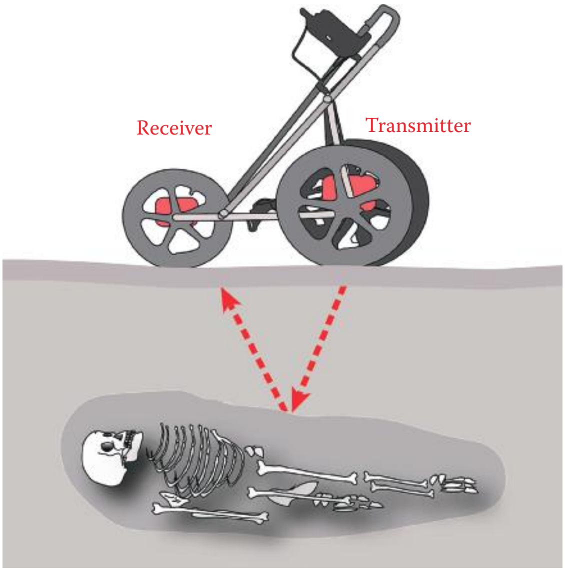

Figure 1.18 Using ground-penetrating radar (GPR) to locate clandestine graves of homicide victims. GPR uses electromagnetic waves emitted from a transmitter, which are detected by a receiver to locate clandestine burials and buried objects such as weapons embedded in soils. Images of the potential evidence are typically obtained by moving the antenna of the GPR device over the surface of the ground. (© Richard C. Li.)

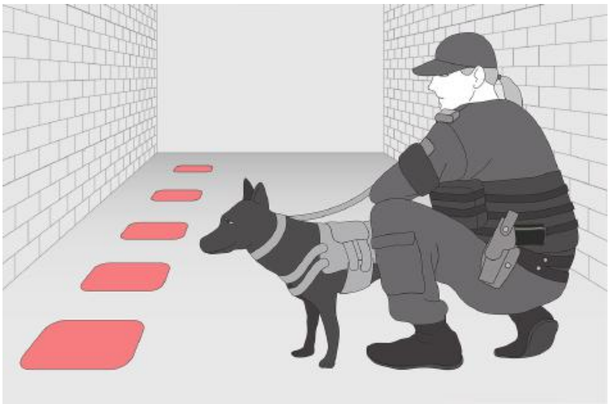

Figure 1.19 Using cadaver-sniffing dogs to alert investigators to the presence of buried bodies. The odor produced by the decomposition of the human body may be sensed by cadaver-sniffing dogs. Odor-absorbing pads absorb and retain the scent of decomposed remains and can be placed at the scene for several days. Upon sniffing the pads, a cadaver dog may indicate the presence of buried human remains or may indicate that human remains were once buried in that location. However, it is not clear that this technique is reliable. (© Richard C. Li.)

MADISON
CONSTABLE SAR K9

Figure 1.20 Tracking dog for searching suspects. In a situation of processing a recent crime scene when the suspect may be in close vicinity to the scene, a tracking dog can be used. A tracking dog, such as a bloodhound, can potentially follow the scent from items left at the scene to locate a suspect nearby. (© Richard C. Li.)

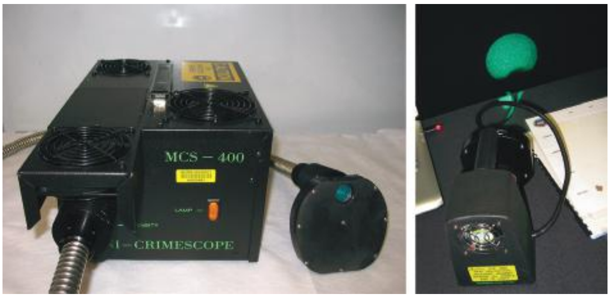

Figure 1.21 Compact alternate light source devices, which are intended specifically for use at crime scenes, reduce search time and improve the recovery of evidence such as biological evidence and chemically enhanced latent fingerprints. (© Richard C. Li.)

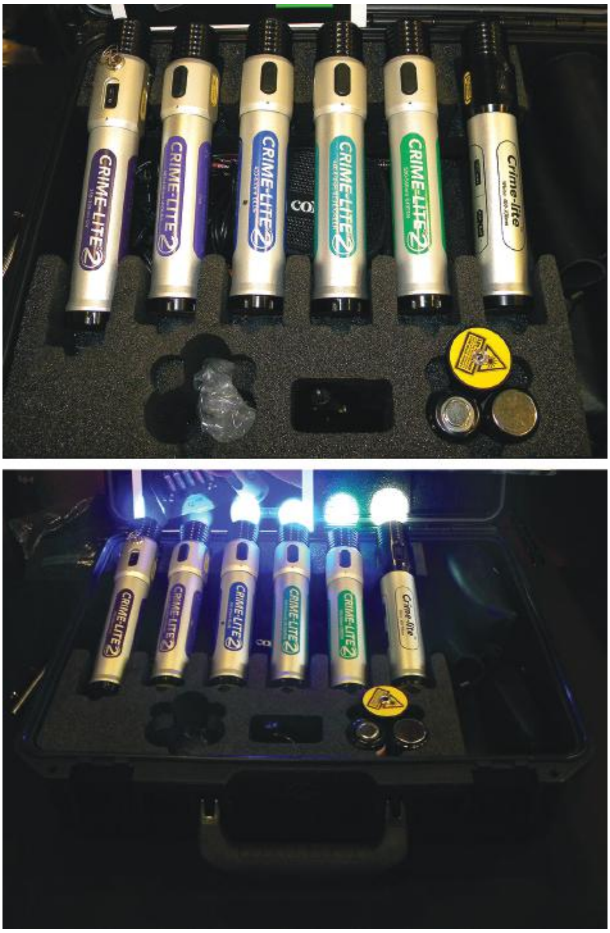

Figure 1.22 LED light sources provide illumination for locating evidence such as bodily fluid stains, hair, or fibers. (© Richard C. Li.)

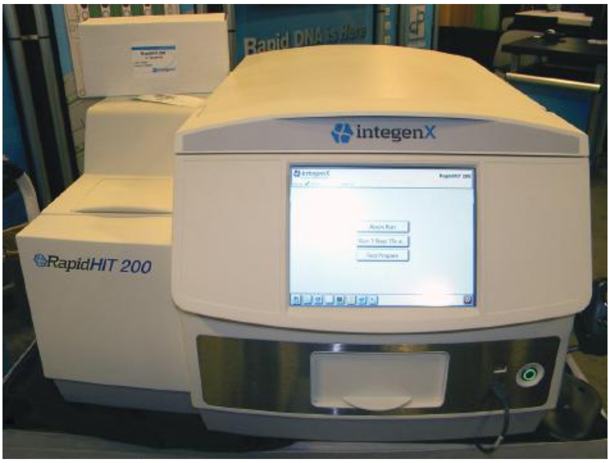

Figure 1.23 An example of a compact Rapid DNA device for processing DNA evidence in the field. (© Richard C. Li.)

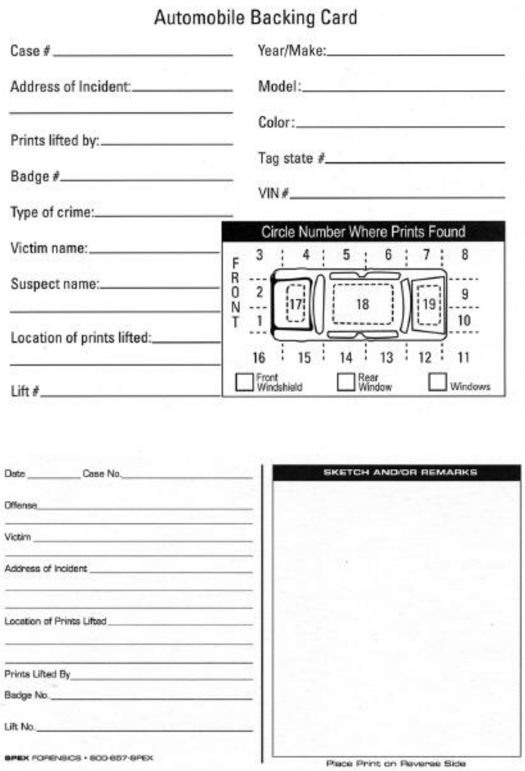

Automobile Backing Card

Case # ____________ Year/Make: ____________

Address of Incident: ____________ Model: ____________

____

Color: ____________

Prints lifted by: ____________

Tag state # ____________

Badge # ____________

VIN # ____________

Type of crime: ____________

Victim name: ____________

Suspect name: ____________

____

Location of prints lifted: ____________

____

Lift # ____________

Circle Number Where Prints Found

FRONT
3
4
5
6
7
8
2
17
18
19
9
1
10
16
15
14
13
12
11

☐ Front Windshield ☐ Rear Window ☐ Windows

Date ________ Case No. ____________

Offense ____________

Victim ____________

Address of Incident ____________

Location of Prints Lifted ____________

Prints Lifted By ____________

Badge No. ____________

Lift No. ____________

SPEX FORENSICS • 800-657-SPEX

SKETCH AND/OR REMARKS

Place Print on Reverse Side

Figure 1.24 An example of sketch documentation. (© Richard C. Li.)

field tests and enhancement reagents can be used to facilitate crime scene searching (Chapter 12). These reagents can detect and identify biological evidence. The tests are very simple, rapid, and sensitive, and thus can be used at crime scenes. For example, phenolphthalin and leucomalachite green tests can be used for detecting blood evidence. Sometimes, minute amounts of blood may be

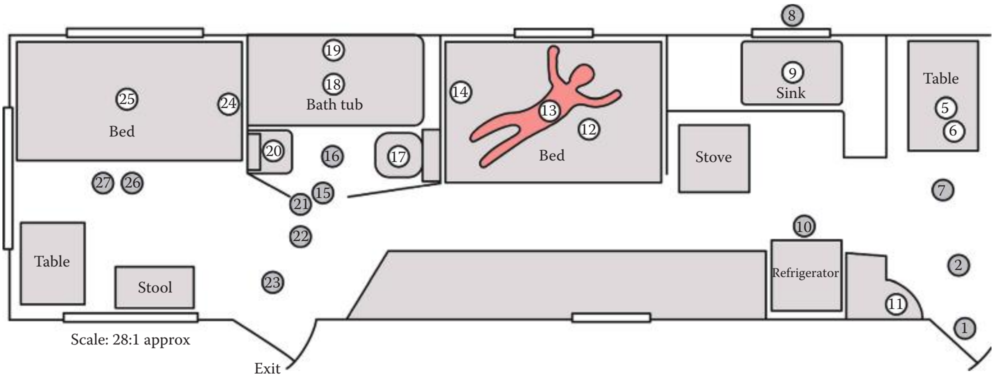

Figure 1.25 Sketches for documenting the sites of fingerprints. (© Richard C. Li.)

present at the scene as a result of attempts to clean up blood prior to the investigation. These stains may not be visible with the naked eye. Enhancement reagents such as luminol and fluorescein, which emit chemiluminant and fluorescent light upon reacting with certain biological materials, respectively, can be used. Additionally, the enhancement reagents can detect faint blood-containing pattern evidence such as faint bloody fingerprints, footprints, and other pattern evidence of physical contact such as drag marks in blood. However, precaution should be taken since these reagents are not usually very specific to blood. Certain substances such as bleach, various metals, and plants may also lead to chemical reactions with the field tests and the enhancement reagents. In these cases, the evidence collected is further tested with laboratory examination and analysis.

Recently, portable and field-deployable instruments have been developed that are capable of processing buccal swabs and potentially other evidence to produce a DNA profile on-site (Figure 1.23). It is a fully automated process, using the Rapid DNA technology (Chapter 8), that

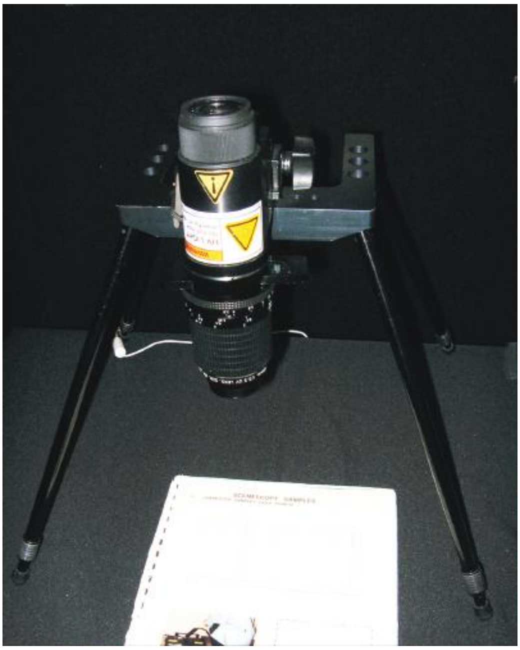

Figure 1.26 A reflected ultraviolet imaging system (RUVIS) imager for documenting close-up views of evidence such as latent fingerprints. (© Richard C. Li.)

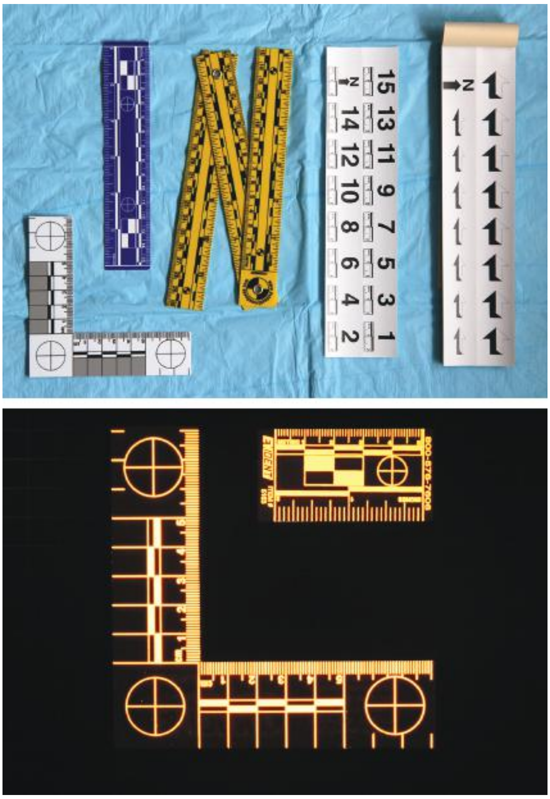

Figure 1.27 Scales for photographic documentation. Regular scales (top) can be used with visible light sources. Fluorescence scales (bottom) can be used for the documentation of fluorescing evidence after certain treatment, such as chemical enhancement, using alternate light sources. (© Richard C. Li.)

can be completed within 2 h by a trained crime scene investigator or police officer. These instruments may provide a new tool for expediting the identification of suspects and developing investigative leads at the scene. Additionally, this technology can enable law enforcement agents to rapidly determine whether the crimes were isolated incidents or part of serial crimes committed by the same offender, such as in serial burglary and arson cases. It can also be used in the identification of human remains in mass disasters.

1.4 Documentation

The conditions at a crime scene, including both the individual items of evidence and the overall scene, must be documented to provide vital information for investigators and for the courts. The most common documentation methods are drawing sketches and taking photographs and videographs. The sketch is to reflect the positions and the spatial relationships of items and persons with measurements using a scale. An investigator usually prepares a rough sketch first and later converts the rough sketch into a finished sketch (Figures 1.24 and 1.25). If bloodstains are present at the scene, the location of bloodstain patterns should be emphasized. Prior to handling and moving evidence, photographs should be taken with different views: an overall view of the entire scene, a medium-range view showing the positions and the relationships of items, and a close-up

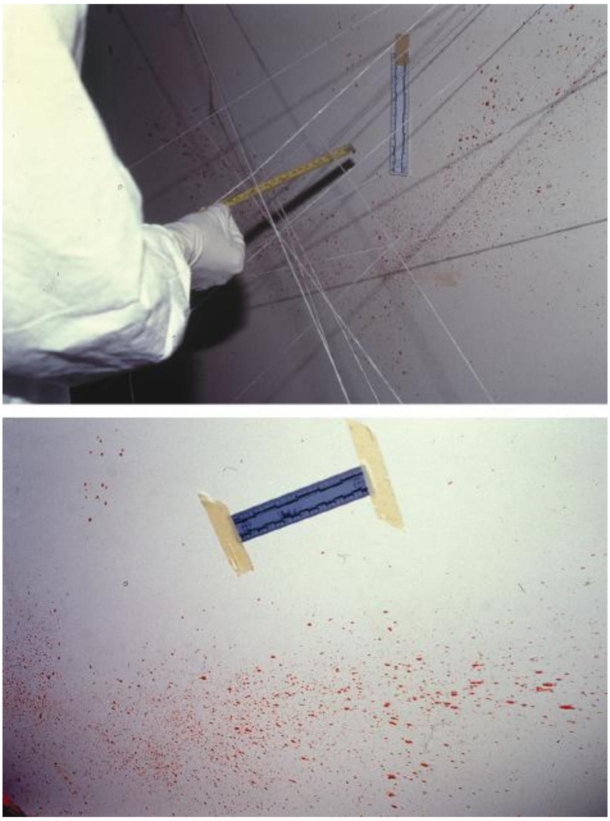

Figure 1.28 Photographic documentation of bloodstain patterns. (© Richard C. Li.)

view showing details of the evidence (Figure 1.26). Photographs should also include a measuring device such as a scale (Figure 1.27) to accurately depict the sizes of items such as bloodstains or bite marks. This can be achieved simply by placing a ruler adjacent to the evidence when it is photographed (see Figure 1.28). A photograph log sheet can be used to record the chronological order of crime scene photographs and to note filming conditions and any additional relevant information (Figure 1.29). Similar documentation should be prepared for videographs when appropriate. Additionally, written or audio-recorded notes can be used. Notes should include complete and accurate information of a crime scene investigation, such as the case identifier number, the identities of the investigators, and a description of the scene or items (e.g., location, size, and shape). Additionally, any disturbance of evidence occurring during crime scene processing should be noted.

1.5 Chain of Custody

Custody information should be recorded at each event when evidence is handled or transferred by authorized personnel. Usually, a custody form listing a specific evidence item is used to document the chain (Figure 1.30). Each individual who acquires custody of the evidence must sign a chain of custody document. An incomplete chain of custody may lead to an inference of possible tampering or contamination of evidence. As a result, the evidence may not be admissible in court.

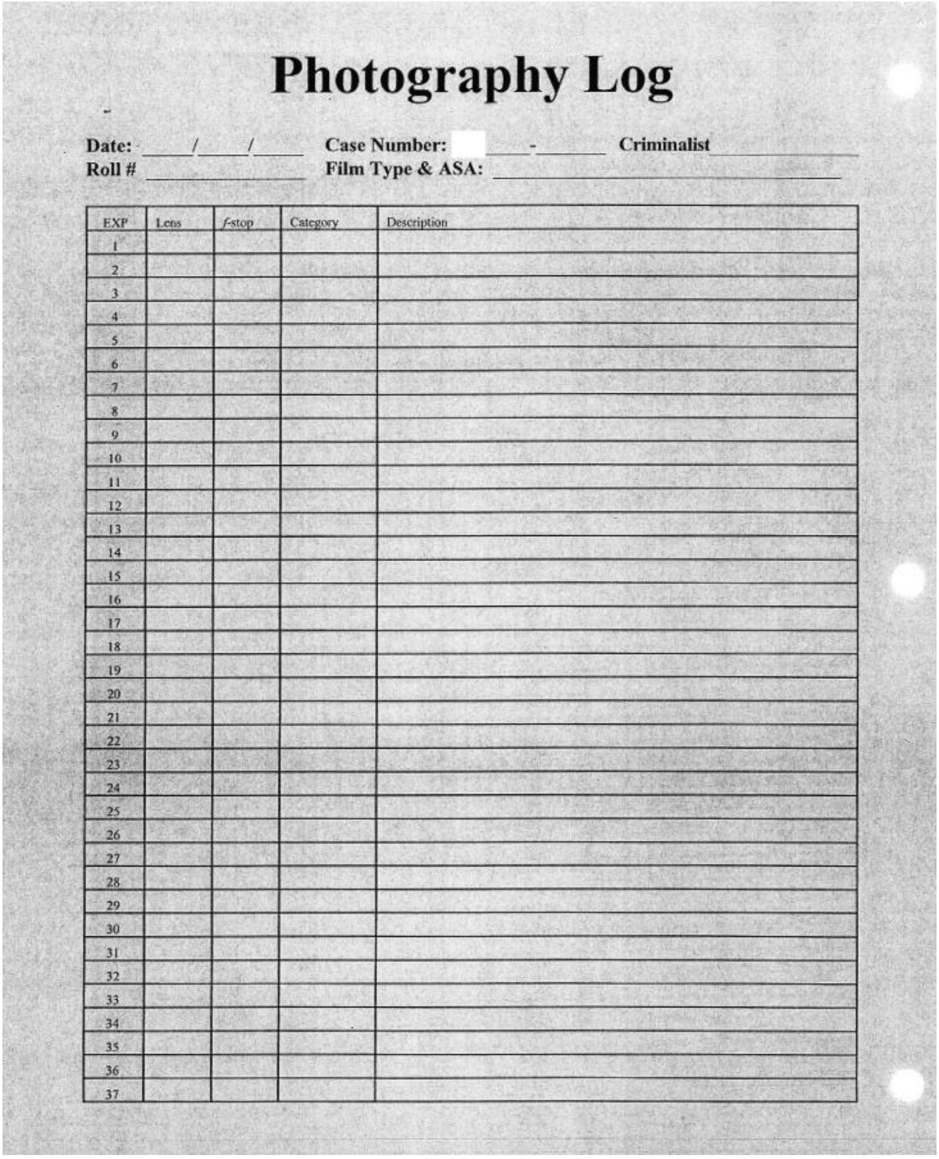

Photography Log

Date: ____/____/____ Case Number: [redacted] ____-____ Criminalist ________
Roll # ________ Film Type & ASA: ________

EXP	Lens	f-stop	Category	Description
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				

Figure 1.29 Crime scene photography log. (© Richard C. Li.)

1.6 Collection of Biological Evidence

After the crime scene documentation is completed, the collection of evidence can be initiated. Small or portable items, such as bloodstained knives, can be collected and submitted to a crime laboratory (Figure 1.31). Large or unmovable items of evidence (Figure 1.32) can be collected and submitted in sections, such as a section of wall where bloodstains are located. Table 1.1 and Figures 1.33 through 1.38 summarize and illustrate representative collection techniques. Specific care is required for the collection of biological evidence in the following situations:

- *Bloodstain pattern evidence*: It is especially important to thoroughly document the bloodstain pattern evidence at a crime scene prior to collection. Bloodstain patterns can be especially useful in crime scene reconstruction.

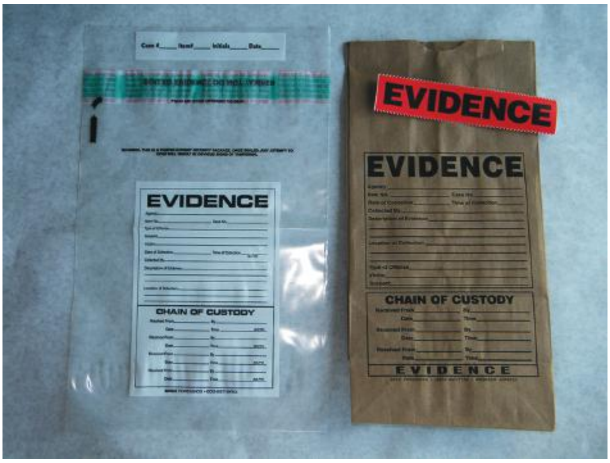
EVIDENCE
EVIDENCE
CHAIN OF CUSTODY
EVIDENCE
CHAIN OF CUSTODY
EVIDENCE

Figure 1.30 Labels with the chain of custody that are used for marking the evidence contained in the packaging. (© Richard C. Li.)

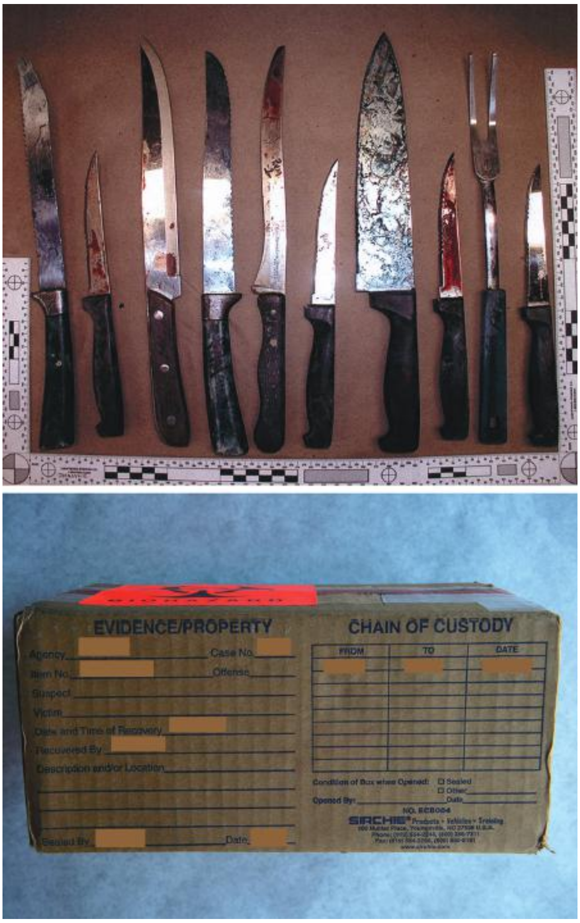
EVIDENCE/PROPERTY
CHAIN OF CUSTODY
FROM
TO
DATE
SIRCHIE

Figure 1.31 Handling sharp objects. Bloodstained knives collected and submitted to laboratories (top) and a box for packing sharp objects (bottom). (© Richard C. Li.)

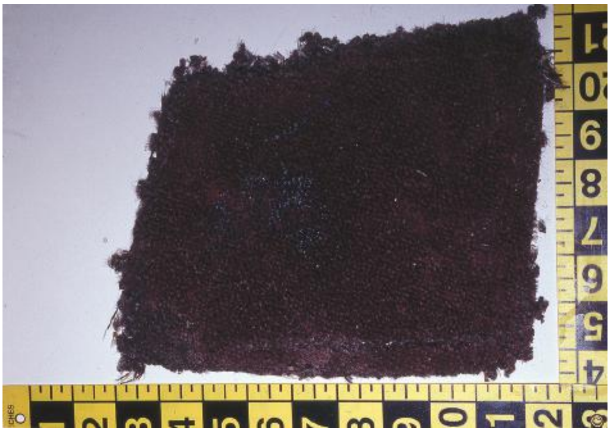

Figure 1.32 A section of bloodstained carpet is collected. (© Richard C. Li.)

- *Multiple analysis of evidence*: If multiple analyses are needed for a single item of evidence, nondestructive analyses should be carried out first. For example, a bloody fingerprint should be collected for ridge detail analysis prior to collecting blood for DNA analysis.
- *Trace evidence*: Trace evidence such as hairs and fibers can be present in bloodstained evidence and should be identified and properly collected.
- *Control samples*: Control (known or blank) samples should be collected from a control area (e.g., unstained area near a collected stain).
- *Size of stain*: Polymerase chain reaction (PCR)-based forensic DNA techniques are highly sensitive and allow for the successful analysis of very small bloodstains. All bloodstains, even if they are barely visible, should be collected at a crime scene.
- *Wet evidence*: Wet evidence should be air-dried (without heat) prior to packaging to prevent the degradation of proteins and nucleic acids, which are used for forensic serological and DNA analysis.

1.7 Marking Evidence

The marking of evidence is necessary for identification purposes so that it can be quickly recognized even years later (Figure 1.39). An investigator's initials, the item number, and the case number are usually included in marking. Information can be marked on a tag, a label attached to the item, or directly on garment evidence. The marking of evidence should not be proximal to bullet holes or biological stains to prevent the mark from interfering with analyses.

1.8 Packaging and Transportation

Packaging is intended to protect and preserve evidence. All evidence should be secured and protected from possible contamination. Fragile items should be protected to prevent any damage during transportation. Exposure to heat and humidity should be avoided to protect biological evidence from degradation during transport. Various packaging methods are available

Table 1.1 Methods for Collecting Biological Evidence

Type of Evidence	Condition	Method of Collection	Procedure
Blood	Dry	Swab	Best on nonabsorbent surfaces; lightly moisten sterile swab with distilled or sterile water, rub over stain while rotating; allow to air-dry; a combination of first a moistened swabbing followed by second a dry swabbing (both swabs submitted) is recommended
		Cutting	Cut stain from item
		Scraping	Scrape bloodstain into a clean piece of paper using a clean blade; wrap sample using druggist's fold
		Lifting	Works for nonabsorbent surfaces; use fingerprint lifting tape that does not interfere with DNA testing to lift stain; lifted stain should be covered with a piece of lifter's cover
		Collect entire item	Collect if item contains bloodstain pattern; difficult to swab; requires multiple exams
	Wet	Swab	Absorb blood sample onto sterile cotton swabs; stain should be concentrated on tip and allowed to air-dry
		FTA paper	Use a sterile disposable pipet to collect liquid blood; spot on FTA paper; allow to air-dry
	Reference liquid blood sample	Venous blood collection	Collect blood in a purple-topped vacutainer tube containing ethylenediaminetetraacetic acid (EDTA) anticoagulant; refrigerate but never freeze
Semen	Dry or wet	See "Blood" above	See "Blood" above
	Condom	Collect entire item	Secure condom with a tie and place in a container in a refrigerator; submit to laboratory as soon as possible
	Various conditions	Victim sexual assault kit	Standardized kit to collect biological evidence from the body of a victim includes swabs, microscope slides, and envelopes
	Various conditions	Suspect standard kit	Standardized kit to collect biological evidence from the body of a suspect includes swabs, microscope slides, and envelopes

(*continued*)

Table 1.1 (Continued) Methods for Collecting Biological Evidence

Type of Evidence	Condition	Method of Collection	Procedure
Victim vaginal fluid	Dry or wet	See “Blood” above	Often collected from a suspect’s pubic area or fingers
Saliva	Dry or wet	See “Blood” above	Often collected from a bite mark or a victim’s pubic area
	Reference saliva (buccal) samples	Swab	Swab the inside of the cheek using two swabs, rotating them during collection; allow swabs to air-dry
		Filter paper	Place donor saliva sample on marked area of filter paper; allow to air-dry
Hairs	Head and pubic hairs	Lifting	Refer to dry bloodstain lifting method
		Transfer	Use forceps to transfer hair onto a piece of paper that can be folded
		Vacuum	Vacuum can be used if necessary; generally not recommended
Fingernails and scrapings	Various conditions	Clipping	Use a clean clipper to clip nails onto clean paper; wrap samples using druggist’s fold
		Scraping	Scrape undersides of nails onto clean paper; wrap samples using druggist’s fold
Bones	Various conditions	Freeze in container	Collect if blood is not available; collect bone with marrow if available; rib bone and vertebrae are preferred; place specimen in a container and freeze if wet
Teeth	Various conditions	Container	Collect teeth with dental pulp if possible into a container
Tissues and organs	Wet	Freeze in a container	Collect if blood is not available; place specimen in a container and freeze
DNA database samples	Various conditions	Various	Follow jurisdiction protocol for collecting samples from arrested individuals or convicted offenders or both

Source: Fisher, B., *Techniques of Crime Scene Investigation*, 7th edn., 2004, CRC Press, Boca Raton; Lee, H.C. et al., *Henry Lee’s Crime Scene Handbook*, 2001, Academic Press, San Diego; National Institute of Justice. Using DNA to solve cold cases, Special Report, 2002, US Department of Justice, Office of Justice Programs, Washington, DC.

Note: Samples requiring refrigeration or freezing are noted; other samples may be stored at room temperature. Dry evidence should be packed in a porous material such as paper (envelope, bag, box) as described in the text.

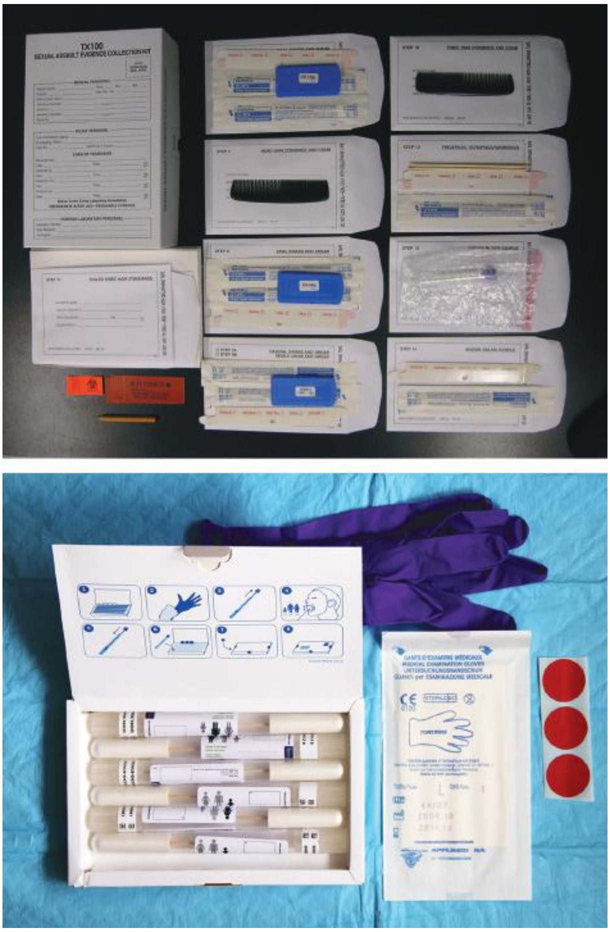

Figure 1.33 Evidence collection kits. Sexual assault evidence collection kit (top). Paternity evidence collection kit (bottom). (© Richard C. Li.)

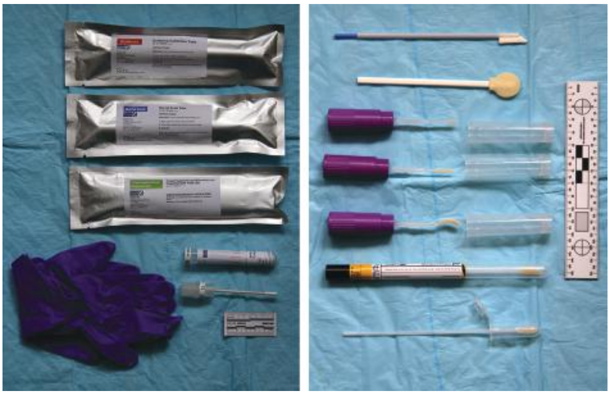

Figure 1.34 Various types of swabs that are used for collecting biological evidence. (© Richard C. Li.)

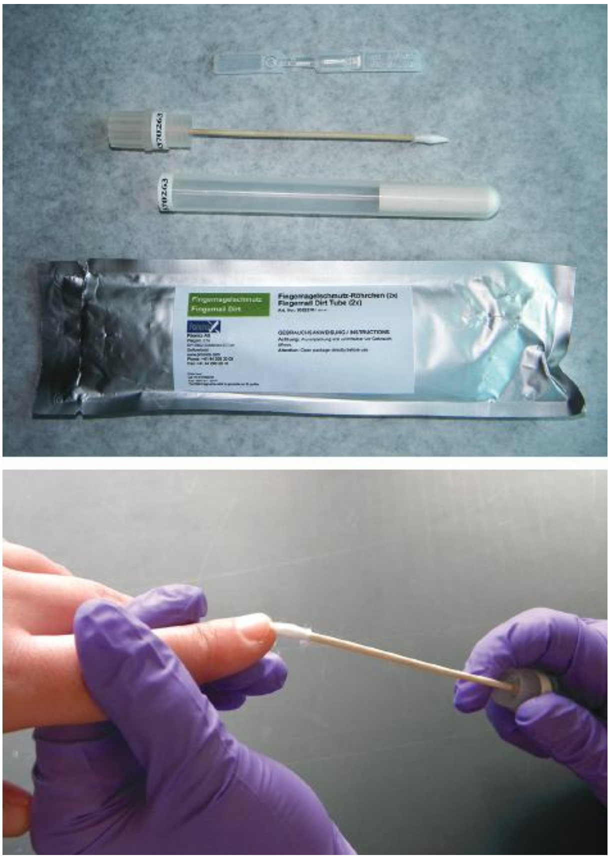

Figure 1.35 Fingernail swabbing for recovering evidence. Fingernail swabs are often collected from individuals who are involved in a struggle in violent crimes and in digital penetration in sexual assault cases. Fingernail swabs are sampled separately from both hands. (© Richard C. Li.)

depending on the type of evidence handled (Figures 1.40 through 1.42). The following are general considerations related to the packaging of evidence:

- *Evidence from different sources*: To prevent the transfer of evidence from different sources, items of evidence should not be grouped in a single package. However, evidence may be packed in a single container if the items were found together.
- *Folding of evidence*: Folding of clothing, especially items with wet bloodstains, can transfer evidence from one part of a garment to another. If a large, dry garment must be folded, a piece of clean paper should be placed between different parts of the garment to avoid direct contact between the different parts of the garment, thereby preventing the transfer of evidence.
- *Packing materials*: Envelopes, bags, and boxes that are made of porous materials such as paper are appropriate for packaging dry biological evidence. Dry, bloodstained evidence should not be sealed in plastic bags or containers that trap moisture.
- *Liquid evidence*: Tubes containing liquid such as blood should not be frozen because the volume of a liquid expands in freezing temperatures and this expansion may lead to cracking. Tubes should be placed in plastic bags to prevent leaks in case of

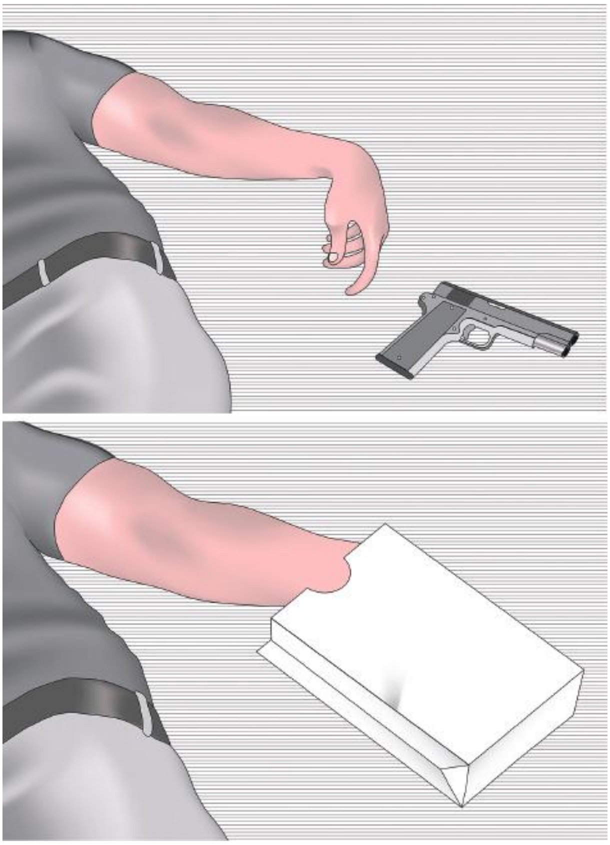

Figure 1.36 Hand bags for protecting the hands of a decedent in an alleged suicide. Gunshot residue can be found on hands after firing a weapon (top). The bagging of the hands, using paper bags and wide rubber bands, prevents the loss of gunshot residue during the transportation of the body (bottom). (© Richard C. Li.)

accidental breakage. Liquid evidence should be transported and submitted to a laboratory as soon as possible after the collection of evidence.

- *Trace evidence*: All such evidence should be wrapped in paper with a druggist's fold (Figure 1.43). The wrapped trace evidence can be packed in an envelope.

Packaged evidence should be properly labeled with a description of the evidence and sealed with evidence tape. It is important for the person packaging the evidence to initial and date across the seal to show authenticity (Figure 1.44). A seal should not be cut when a sealed evidence bag is opened. Instead, an opening can be created by cutting at an area distal from the existing seal. After analysis is complete, the evidence packaging should be resealed. Table 1.1 summarizes additional steps for packaging evidence.

1.9 Final Survey and the Release of the Crime Scene

During a final survey, a discussion with all personnel in the crime scene investigation team should be carried out to thoroughly review all aspects of the search. It is important to ensure

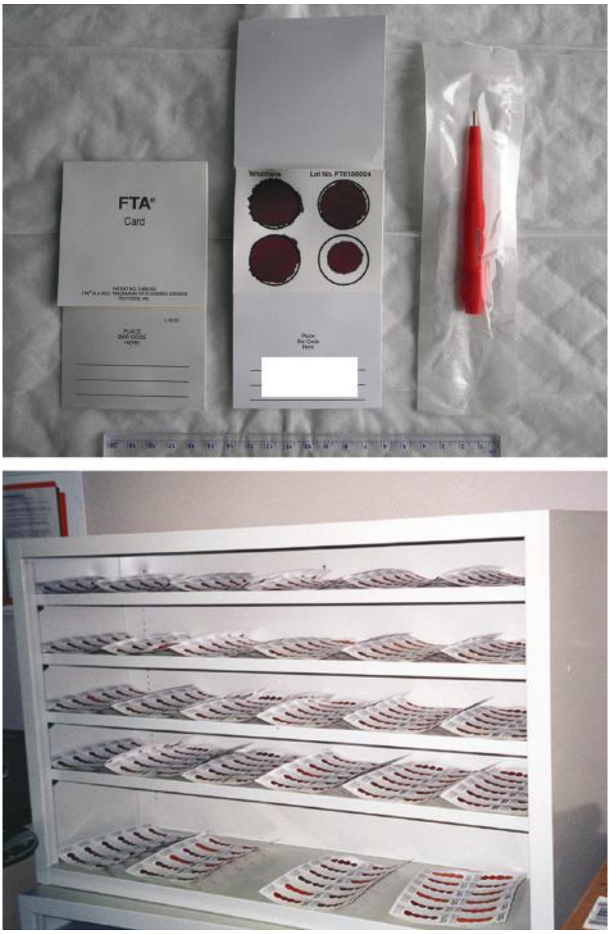

Figure 1.37 Blood cards are typically used for collecting blood evidence from a known source such as a suspect or a victim (top). A manual hole punch can be used to create a blood card punch for DNA extraction. Blood samples are air-dried on blood cards for storage (bottom). (© Richard C. Li.)

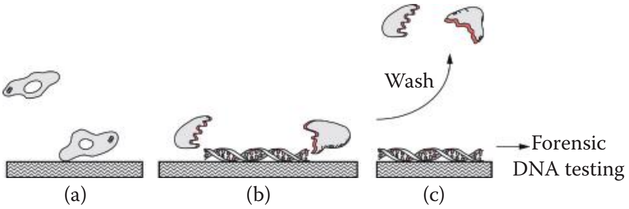

Figure 1.38 Application of Flinders Technology Associates (FTA) filter paper for the collection of biological evidence. (a) Biological fluid with cells is applied to FTA paper. (b) Cells are lysed and DNA is immobilized on FTA paper. (c) Cellular materials are washed away and DNA remaining on the FTA paper can be used for forensic testing. (© Richard C. Li.)

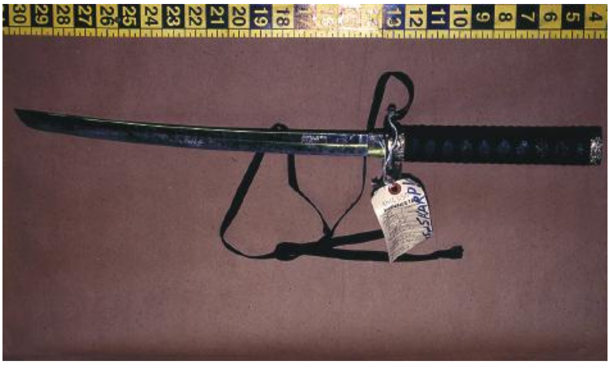

Figure 1.39 Photographic documentation of a knife. Note the evidence tag. (© Richard C. Li.)

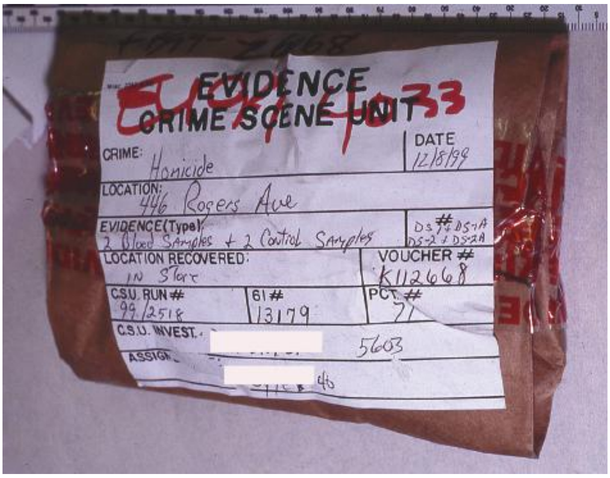

EVIDENCE
CRIME SCENE UNIT

CRIME: Homicide
DATE 12/8/99
LOCATION: 446 Rogers Ave
EVIDENCE(Type): 2 Blood Samples + 2 Control Samples
DS # DS-1 + DS-1A DS-2 + DS-2A
LOCATION RECOVERED: IN Store
VOUCHER # K112668
C.S.U. RUN # 99/2518
61 # 13179
PCT. # 77
C.S.U. INVEST. [redacted] 5603
ASSIGN [redacted] 40

Figure 1.40 Evidence containing dried bodily fluid stains is packed in a paper bag. (© Richard C. Li.)

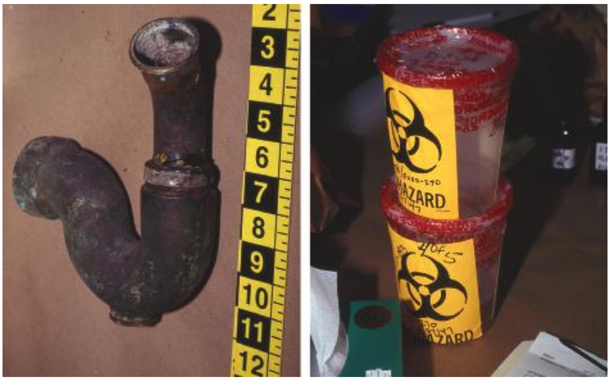

Figure 1.41 Alleged diluted blood collected from a pipe (left) and placed in plastic containers (right). (© Richard C. Li.)

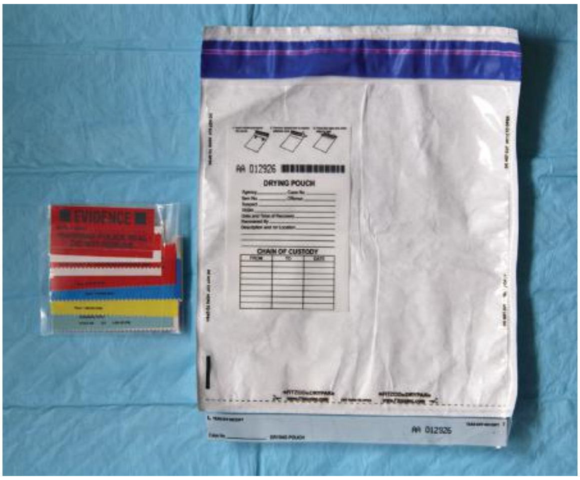
EVIDENCE
AA 012926
DRYING POUCH
CHAIN OF CUSTODY
AA 012926

Figure 1.42 Evidence pouch. The front of the pouch is transparent for viewing the content. The back of the pouch is made of breathable materials allowing wet evidence, such as swabs, to dry inside the pouch. (© Richard C. Li.)

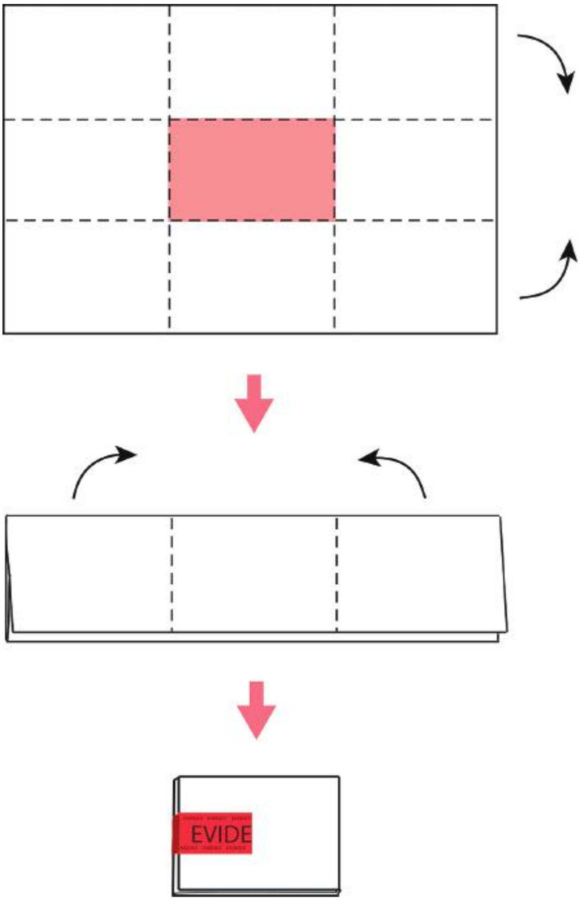

Figure 1.43 An example of a druggist's fold. Trace evidence should be deposited in the center (colored area) of the paper. (© Richard C. Li.)

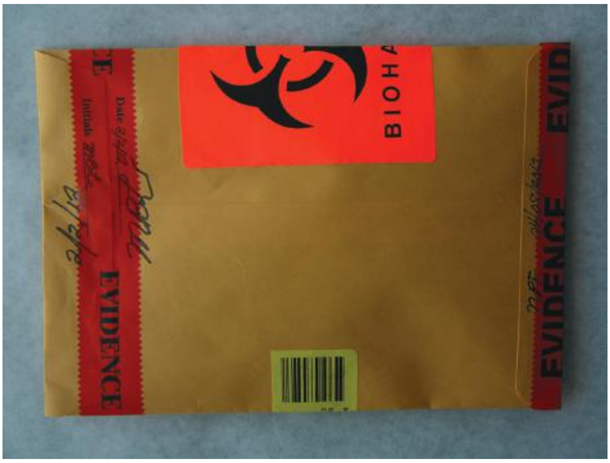

Figure 1.44 Proper marking of sealed evidence. Note that the evidence packaging was cut a second time and resealed at a different location than that of the preexisting seal. (© Richard C. Li.)

that the scene has been searched correctly and completely, and that no area has been missed or overlooked. All documentation including the chain of custody document must be complete and all evidence should be collected, packed, documented, and marked. Photographs of the final condition of the scene should be taken. Once the final survey is completed, the crime scene can be released. Reentry into the crime scene may require a search warrant after the scene is released. Crime scene release documentation usually includes the time and date of release, to whom it is released, and by whom it is released.

1.10 Crime Scene Reconstruction

Crime scene reconstruction is the scientific process of determining the sequence of events and actions that occurred prior to, during, and after a crime. Reconstruction is carried out based on the information from the crime scene observations and the laboratory examination of physical evidence. The overall scientific process in reconstruction usually involves several steps. The process usually begins with the formulation of questions related to the problems that need to be solved. The questions can refer to the explanation of the specifics of the crime, for instance, "Where was the shooter's position when the shooting occurred?," "Where was the victim's position when shot?," and "What is the muzzle-to-target distance during the shooting?" In order to conduct a thorough crime scene reconstruction, all useful information is collected for review, such as photographs, videotapes, notes, sketches, autopsy reports, and analysis reports of the physical evidence. A hypothesis is then constructed based on the information obtained, which may explain the events and actions involved in a crime. The next step is making predictions that determine the logical consequences of the hypothesis. One or more predictions are selected for testing. The hypothesis is tested by conducting reconstruction experiments. One example of a reconstruction test is bloodstain pattern reconstruction in violent crimes (Chapter 2). Other examples of reconstruction may include trajectory and shooting, glass fracture, and accident reconstruction. The final step is to analyze experimental data and draw a conclusion. The experimental data are analyzed to see if the hypothesis is true or false. Additionally, the interpretation of physical evidence analysis, witness and confession statements, and investigative information should also be considered.

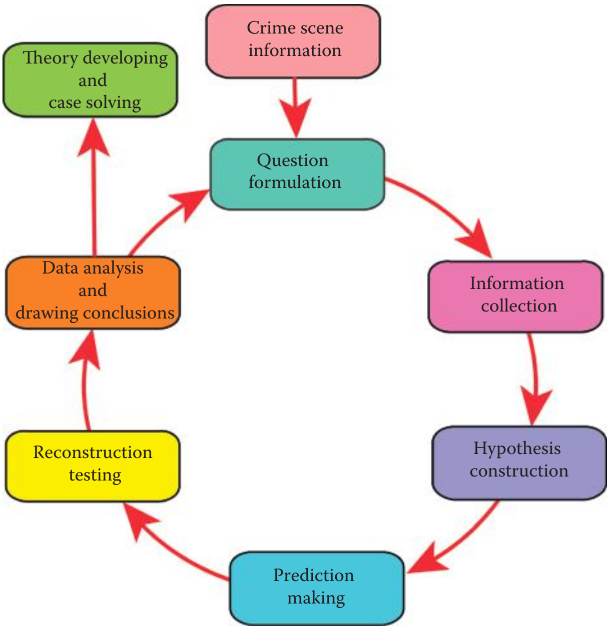
Crime scene information
Theory developing and case solving
Question formulation
Data analysis and drawing conclusions
Information collection
Reconstruction testing
Hypothesis construction
Prediction making

Figure 1.45 The scientific process of crime scene reconstruction. (© Richard C. Li.)

If the results of the experiment are consistent with the hypothesis, a theory can be developed that is intended to provide valuable information to the investigation and future prosecution of a case. Sometimes, forensic scientists may find that the hypothesis is inconsistent with the test results. In that case, an alternative hypothesis needs to be constructed to initiate another reconstruction process. Figure 1.45 illustrates the scientific process of crime scene reconstruction.

Bibliography

Ackermann, K., K.N. Ballantyne, and M. Kayser, Estimating trace deposition time with circadian biomarkers: A prospective and versatile tool for crime scene reconstruction. *Int J Legal Med*, 2010, **124**(5): 387–395.

Adderley, R. and J.W. Bond, The effects of deprivation on the time spent examining crime scenes and the recovery of DNA and fingerprints. *J Forensic Sci*, 2008, **53**(1): 178–182.

Barash, M., A. Reshef, and P. Brauner, The use of adhesive tape for recovery of DNA from crime scene items. *J Forensic Sci*, 2010, **55**(4): 1058–1064.

Benschop, C.C., et al., Post-coital vaginal sampling with nylon flocked swabs improves DNA typing. *Forensic Sci Int Genet*, 2010, **4**(2): 115–121.

Bond, J.W. and C. Hammond, The value of DNA material recovered from crime scenes. *J Forensic Sci*, 2008, **53**(4): 797–801.

Brownlow, R.J., K.E. Dagnall, and C.E. Ames, A comparison of DNA collection and retrieval from two swab types (cotton and nylon flocked swab) when processed using three QIAGEN extraction methods. *J Forensic Sci*, 2012, **57**(3): 713–717.

Buck, U., et al., Accident or homicide—Virtual crime scene reconstruction using 3D methods. *Forensic Sci Int*, 2013, **225**(1–3): 75–84.

Byard, R.W., et al., "Murder-suicide" or "murder-accident"? Difficulties with the analysis of cases. *J Forensic Sci*, 2010, **55**(5): 1375–1377.

Castello, A., F. Frances, and F. Verdu, Solving underwater crimes: Development of latent prints made on submerged objects. *Sci Justice*, 2013, **53**(3): 328–331.

2

Crime Scene Bloodstain Pattern Analysis

Bloodstain pattern analysis is the application of scientific knowledge to the examination and the interpretation of the morphology, the sequence, and the distribution of bloodstains associated with a crime. These analyses may determine the sequence of events; the approximate blood source locations; the positioning of the victim; and the position, the intensity, and the number of impacts applied to the blood source. They also can assist in the determination of the manner of death and can distinguish between accidents, homicides, and suicides. Bloodstain pattern analysis provides critical information for crime scene reconstructions in violent crime investigations (Chapter 1).

2.1 Basic Biological Properties of Human Blood

Blood is a bodily fluid circulating within the body. An average adult has a blood volume of approximately 8% of his or her body weight. Blood consists of a cellular portion as well as a liquid portion known as plasma (Chapter 12). The cellular portion consists of blood cells and platelets. The plasma is mostly composed of water and other substances such as proteins, inorganic salts, and other organic substances. The mass density of blood is only slightly greater than that of water. Blood can form clots (or *thrombi*) that are the result of blood coagulation (Chapter 16). Coagulation begins after an injury occurs, stopping blood loss from a damaged vessel. The normal coagulation time for 1 mL of venous blood in a glass tube is 5–15 min. The coagulation time can be affected by many factors such as blood volume and mechanical disturbance.

2.2 Formation of Bloodstains

The formation of a blood droplet is a complex event that is influenced by viscosity, surface tension, cohesion force, and gravity. Blood is viscous, and blood viscosity is a measure of the blood's resistance to flow. The viscosity of blood is approximately five times greater than that of water. During the formation of a drop of blood, blood leaks out from a blood source. The surface tension of the blood causes it to hang from the opening of a blood source and to form a pendant drop of blood (Figure 2.1). The molecules of a blood drop are held together by the cohesion force to maintain the shape of a blood drop (Figure 2.2). Surface tension causes liquids to minimize their surface. As a result, the formed blood drop is spherical. As the volume of the drop gradually increases and exceeds a certain size, it detaches itself and falls. The falling drop is also held

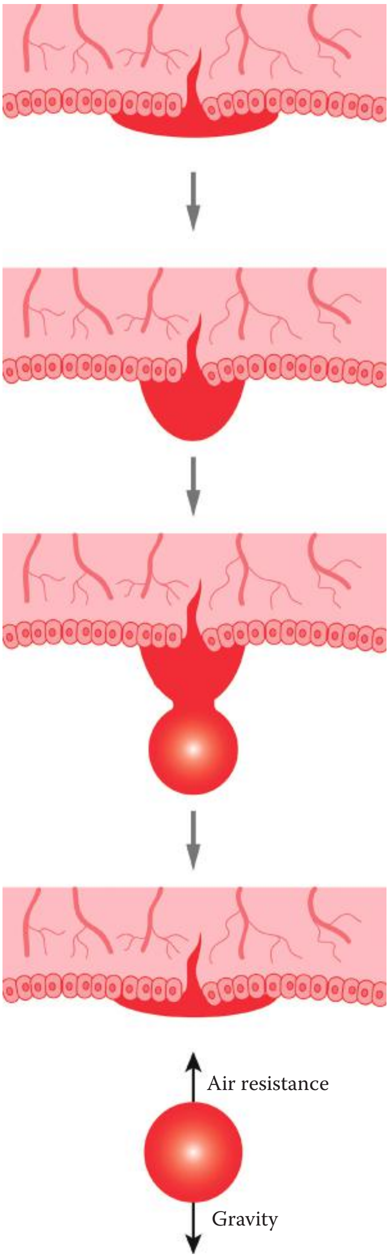

Figure 2.1 Forming a blood drop from a blood source. The blood that leaks out of the blood source forms a pendant drop of blood. As the volume of the pendant drop increases, the drop stretches in a downward direction. Eventually, the drop detaches and falls. The falling drop is largely influenced by the force of gravity and air resistance. (© Richard C. Li.)

together by surface tension. A falling blood drop is influenced by the downward force of gravity acting on the drop and the air resistance that acts in the opposite direction as the drop is in motion (Figure 2.1).

When a bloodstain lands on a surface, the shape and the size of the bloodstain is affected by the texture of the target surface. Bloodstains that land on porous or rough surfaces usually

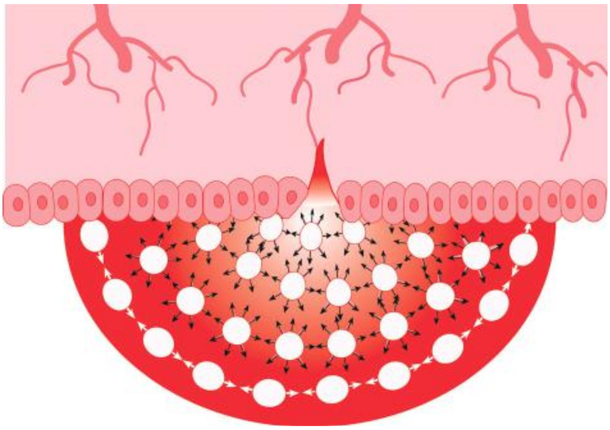

Figure 2.2 The particles of blood are attracted to each other by cohesive forces that are responsible for surface tension. As a result, a formed blood drop is spherical in order to minimize its surface area. Black arrow, cohesive force; white arrow, surface tension. (© Richard C. Li.)

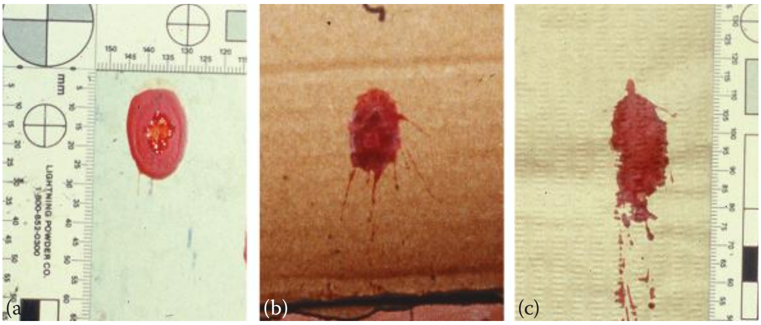

Figure 2.3 The respective morphologies of falling blood drops that land on surfaces with different textures at a 30° angle. (a) Tile, (b) cardboard, and (c) paper towel. (© Richard C. Li.)

have more distortion around the edges of the stains than those that land on smooth surfaces. A comparison of blood dropped onto different textures of target surfaces is shown in Figure 2.3.

2.3 Chemical Enhancement and Documentation of Bloodstain Evidence

Many chemical reagents react with blood to exhibit a color, a chemiluminescent light, or a fluorescent light (Chapter 12). These tests are extremely sensitive and thus are used as chemical enhancement reagents for detecting bloodstains. For bloodstain pattern analysis, the enhancement reagent is primarily used for detecting latent bloodstains such as diluted bloodstains that are visible on enhancement. A commonly used chemical reagent is luminol, which can be used for locating bloodstains at the scene. Other reagents such as phenolphthalein, leucomalachite green, and tetramethylbenzidine are not often used as enhancement reagents but rather as presumptive tests for blood. The positive reactions of all these reagents indicate the presence of blood.

Documenting bloodstain patterns at the scene is a major task of the investigation. Documenting bloodstain evidence can be done using a combination of photography, note-taking, and sketching. The general principle of crime scene documentation is described in Chapter 1. In bloodstain pattern analysis, special attention must obviously be given to bloodstains. The photographic documentation of bloodstains may be performed by multiple means, including film and digital photography, as well as videotaping. Photographs should be taken with an overall view followed by a medium-range and a close-up view of the bloodstain patterns. A scale of measurement must be included in the photograph, which is critical for bloodstain analysis. To avoid any distortion, the photographs should be taken with the camera lens parallel to the target surface where the bloodstains are located. An overall photograph provides an overall view of the scene including the bloodstain evidence (Figure 2.4a). A midrange photograph provides more details of the bloodstain pattern compared with that of the overall photograph (Figure 2.4b). Single bloodstains should be visible in midrange images. A close-up photograph, usually taken with a macro lens, provides a detailed image of single bloodstains, which is useful for spatter pattern analysis (Figure 2.4c).

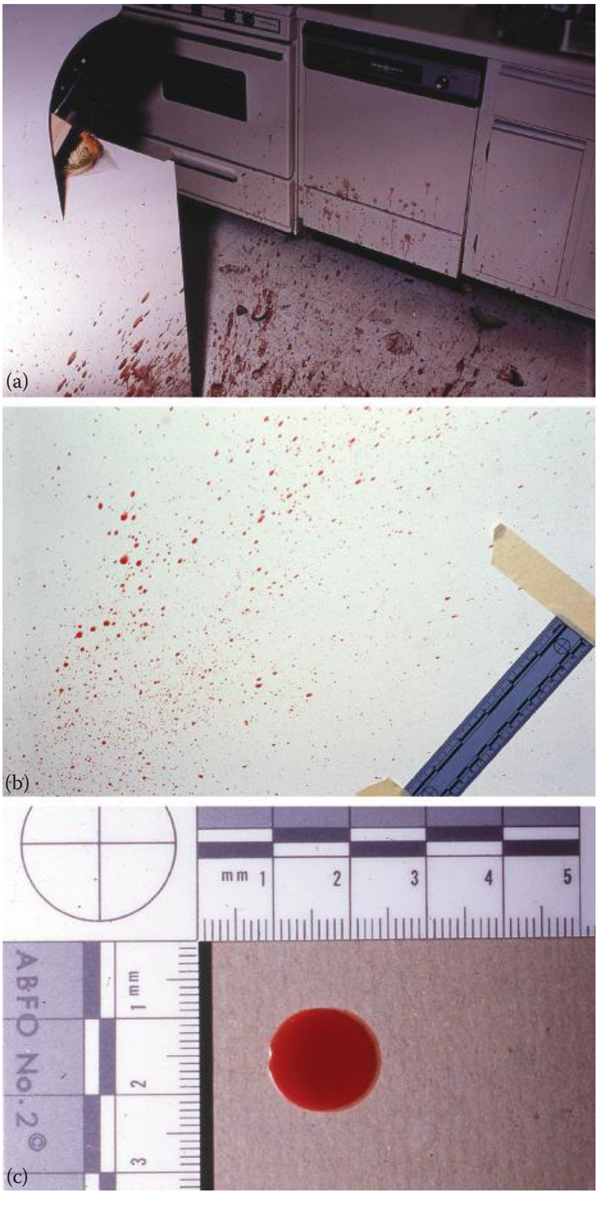

Figure 2.4 Crime scene photographic documentation. (a) Overall, (b) midrange, and (c) close-up photographs. (© Richard C. Li.)

2.4 Analyzing Spatter Stains

A *spatter stain*, based on the recommended terminology of the Scientific Working Group on Bloodstain Pattern Analysis (SWGSTAIN), is “a bloodstain resulting from a blood drop dispersed through the air due to an external force applied to a source of liquid blood.” The patterns of spatter stains, including the shape and the size of the stains, are affected by the direction and the angle of impact (discussed in detail in Section 2.5.3) of the spatter stains that are projected. This information can be obtained from an analysis of the patterns of the spatter stains. Thus, it is possible to determine the area of origin (discussed in detail in Section 2.4.4) where an external force was directly applied to the blood source.

2.4.1 Velocity of Blood Droplets

The sizes of bloodstains are affected by the external force that is directly applied on a blood source. Increasing the energy of the external force will reduce the surface tension, thus decreasing the size of the droplets. Since these travelling blood droplets are driven by the energy derived from the external force, the higher the energy, the higher the velocities of the droplets. Bloodstains can be divided into three categories based on different travelling speeds. Low-velocity impact spatter is formed when a blood droplet is travelling at <1.5 m/s. The resulting stains are usually >4 mm in diameter (Figure 2.5a). As the travelling speed of blood droplets increases, the size of the spatter stain decreases. Medium-velocity impact spatter is formed when a blood source is subjected to a force associated with beatings or stabbings. The resulting stains range from 1 to 4 mm in diameter (Figure 2.5b). High-velocity impact spatter is formed when a blood source is subjected to a force associated with shooting using firearms. The resulting stains are usually <1 mm in diameter.

2.4.2 Determining the Directionality of the Stains

In this analysis, the effects of the directionality of the spatter stains projected are examined. SWGSTAIN defines the directionality to be “the characteristic of a bloodstain that indicates the direction blood was moving at the time of deposition.” This analysis is applicable when the blood source is projected onto a surface at an angle of between 0° and 90°. Under this condition, the resulting spatter stain is an elongated ellipse (Figure 2.6), which is known as the *parent stain*. Additionally, *satellite stains* in the vicinity of the parent stain can be observed. As defined by SWGSTAIN, a satellite stain is “a smaller bloodstain that originated during the formation of the parent stain as a result of blood impacting a surface.” More importantly, a spine is observed, which is the pointed edge away from the parent stain. When such a pattern is observed, the pointed end of the spine always points toward the direction of travel of the bloodstains.

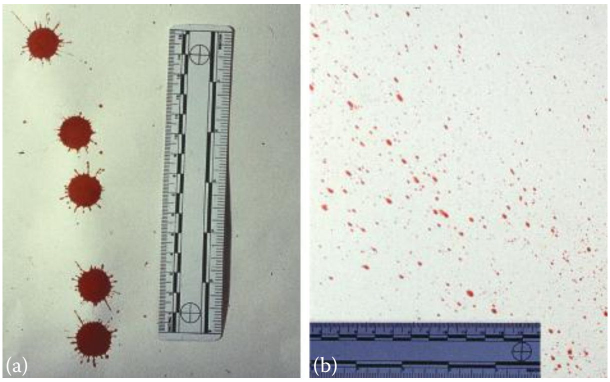

Figure 2.5 Bloodstains can be categorized based on their travelling velocities. (a) An example of a low-velocity impact spatter stain and (b) an example of a medium-velocity impact spatter stain. (© Richard C. Li.)

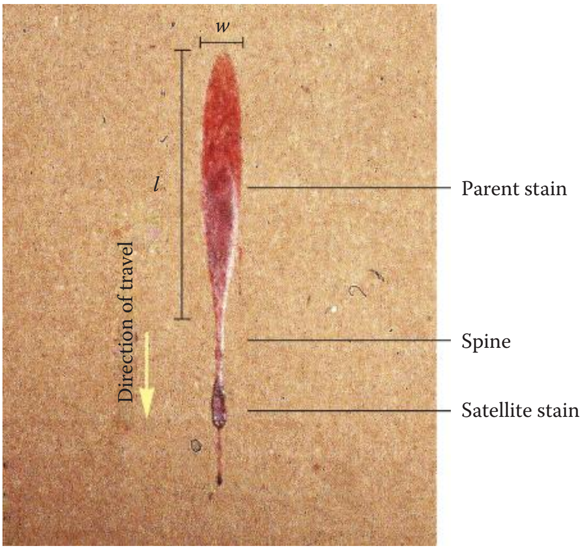

Figure 2.6 The morphology and directionality of a blood spatter stain. The arrow indicates the direction of travel. (© Richard C. Li.)

2.4.3 Determining Angles of Impact

SWGSTAIN defines the *angles of impact* to be “the acute angle (alpha), relative to the plane of a target at which a blood drop strikes the target.” The shapes of the spatter stains are affected by the angle of impact. When a blood drop lands on a surface at a perpendicular angle (90°), a circular parent stain is formed (Figure 2.7), where the length and the width of the stain are equal. When a blood drop is projected onto a surface at an angle of between 0° and 90°, the stain is elongated. As the impact angle decreases, the shape of the spatter stain is more elongated (Figure 2.7), in which the length of the stain is greater than the width. It is observed that the ratio of the width and the length of the parent stain is proportional to the sine of the impact angle, which is summarized in the following trigonometric equation:

$$\sin \alpha = \frac{w}{l}$$

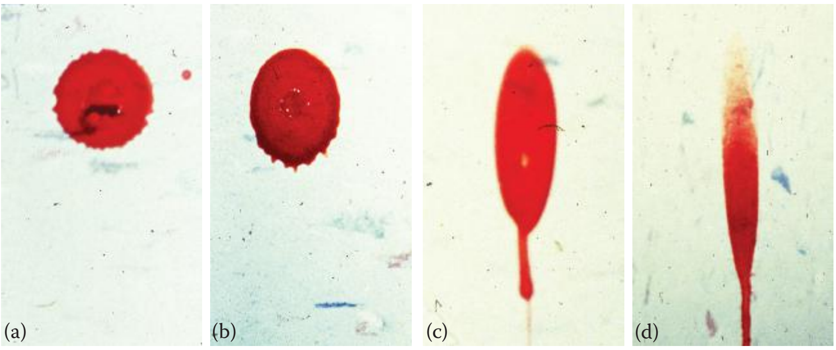

Figure 2.7 The effects of the impact angle on the shapes of blood spatter stains. Spatter stains are projected onto the surface of a ceramic tile at: (a) 90°, (b) 50°, (c) 20°, and (d) 10°. (© Richard C. Li.)

In this equation, α is the angle of impact, *l* is the length of the parent stain (major axis), and *w* is the width of the parent stain (minor axis). Thus, the angle of impact can be determined based on the relationship between the length and the width of the stain (Figure 2.8). Obviously, the measurement of the stain's axes is critical to the accuracy of the calculation of the angle of impact. To produce accurate and reproducible measurements, bloodstain pattern analysis software can be used, which superimposes a scaled close-up image of an individual bloodstain and calculates the angle of impact.

2.4.4 Determining Area of Origin

SWGSTAIN defines the *area of origin* to be "the three-dimensional location from which spatter originated." Using simple trigonometry, the area of origin can be determined based on the measurements from multiple elongated spatter stains (Figure 2.9). This can be accomplished by using the string method or the tangent method.

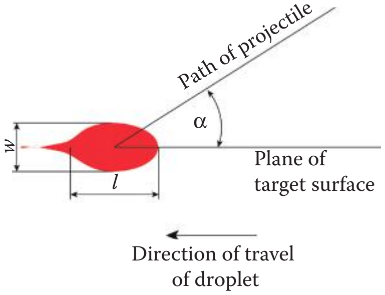

Figure 2.8 Impact angle. The angle between the path of a projectile and the plane of the target surface is shown. α, the impact angle; *l*, the length of the parent stain; and *w*, the width of the parent stain. (© Richard C. Li.)

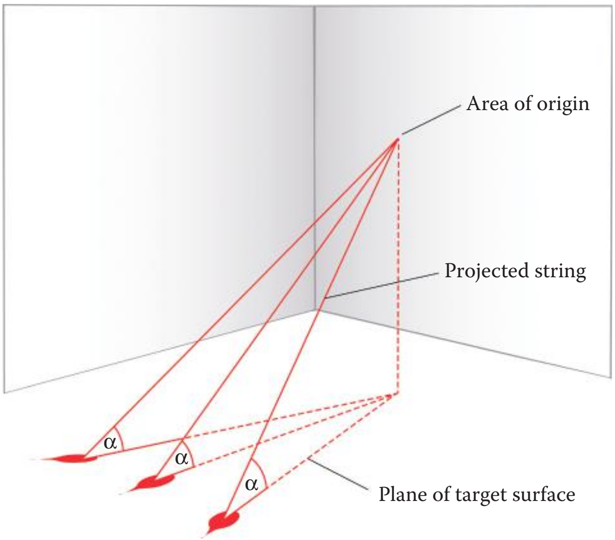

Figure 2.9 Area of origin. The area of origin is determined using the string method. Only three representative bloodstains are shown. α, impact angle. (© Richard C. Li.)

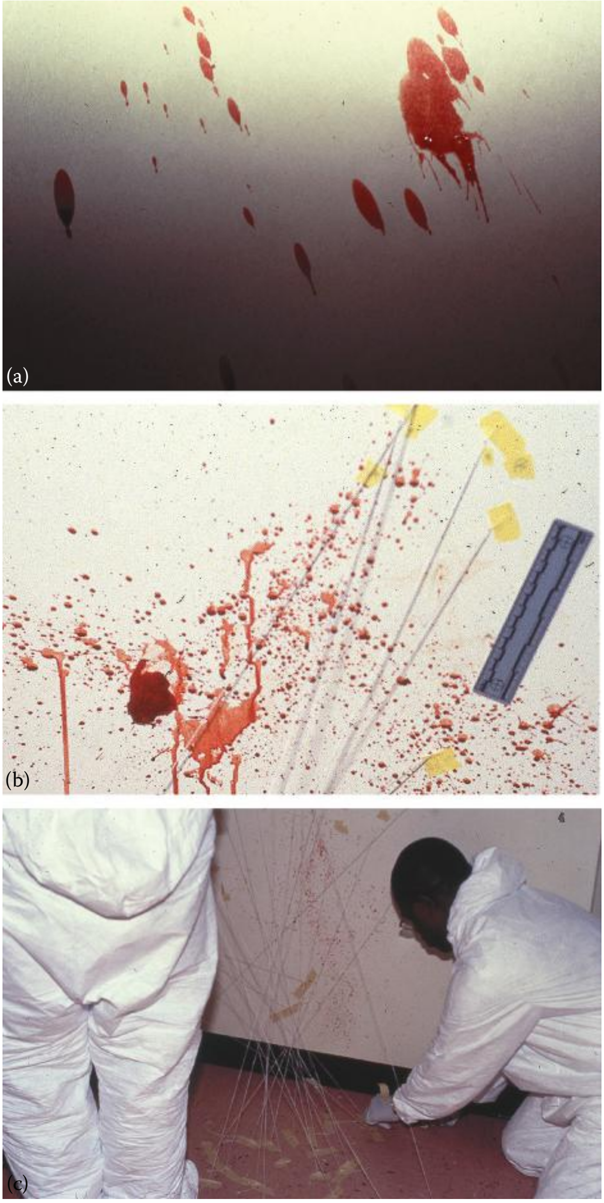

Figure 2.10 Determining the area of origin using the string method. (a) Selecting elongated spatter stains, (b) connecting strings, and (c) setting the path of the strings. (© Richard C. Li.)

In the string method, multiple (approximately two dozen) well-formed, elongated spatter stains are selected for analysis (Figure 2.10a). For each stain, the angle of impact is calculated. A piece of string is then connected between the stain and a surface with one end of the string precisely attached to the spatter stain (Figure 2.10b). The path of the string, indicating the trajectory of the stain, is set using a protractor based on the calculated angle of impact (Figures 2.10c and 2.11). This process is repeated until all the stains that have been selected are processed. For

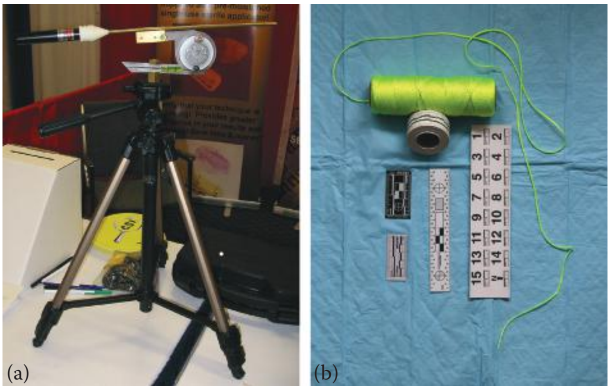

Figure 2.11 Tools for finding the area of origin. (a) Laser trajectory pointer and (b) string and scales. (© Richard C. Li.)

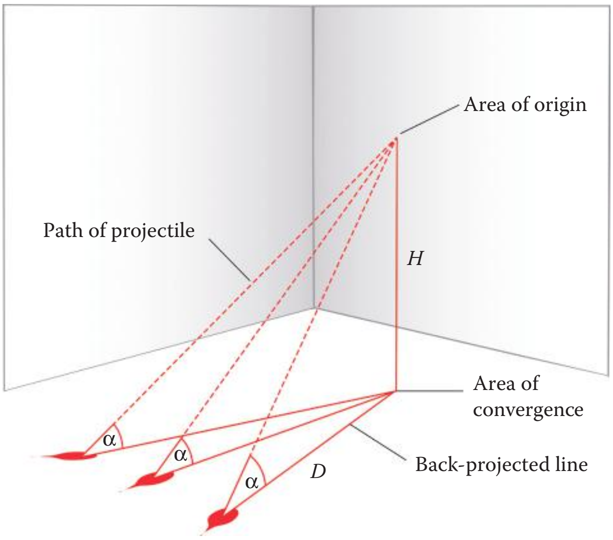

Figure 2.12 Determining the area of origin using the tangent method. Only three representative bloodstains are shown. α, the angle of impact; *H*, the height of the area of origin; and *D*, the distance from the spatter stain to the area of convergence. (© Richard C. Li.)

a spatter pattern generated from a single impact event, the strings converge. The area where the strings meet is the area of origin.

In the tangent method, the directionality of a single stain is determined first. A line is then back projected through the major axis of the bloodstain. For a single impact event, approximately two dozen stains are processed to determine the *area of convergence*. Based on SWGSTAIN's definition, the area of convergence is "the area containing the intersections generated by lines drawn through the long axes of individual stains that indicates in two dimensions the location of the blood source." Next, the angle of impact of each stain is calculated. The distance from the bloodstain to the area of convergence is measured (Figure 2.12). The height of the area of origin is calculated using the tangent function as shown:

$$H = D \cdot \tan \alpha$$

In the equation, α is the angle of impact, *H* is the height of the area of origin, and *D* is the distance from the spatter stain to the area of convergence.

2.5 Types of Bloodstain Patterns

Bloodstain patterns can be classified into three basic categories: passive, transfer, and projected bloodstains.

2.5.1 Passive Bloodstains

A *passive bloodstain* is formed due to bleeding from wounds, and the blood is deposited on a surface by the influence of the force of gravity alone. For example, a *drip stain* is formed when a falling drop of blood from an exposed wound or a blood-bearing object lands on a surface. If a blood source is moving, a *drip trail* is formed. A *drip pattern*, which is distinct from a drip stain, is formed when a liquid drips into another liquid, where one or both of the liquids are blood (Figure 2.13). As a result, secondary spatter stains are generated. As the dropping distance of the blood increases, the number of secondary spatter stains usually increases, and the size of these stains decreases. An approximate estimation of the dropping distance is possible. A *splash pattern* is formed when a volume of blood spills onto a surface (Figure 2.14). Splash patterns usually

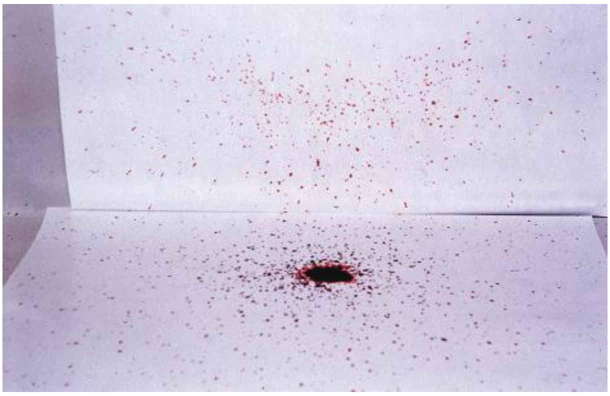

Figure 2.13 A drip pattern. The secondary spatter stains are shown. (© Richard C. Li.)

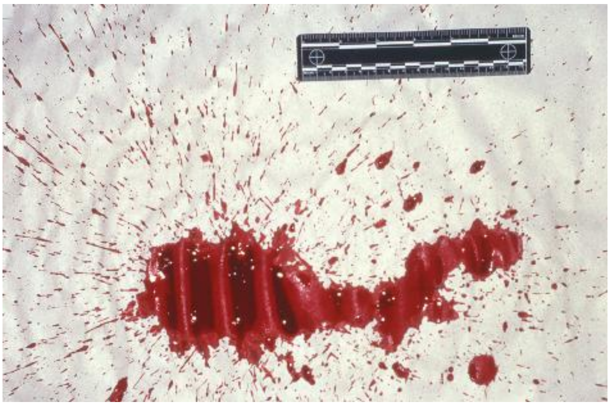

Figure 2.14 A splash pattern. Peripheral, elongated bloodstains are shown. (© Richard C. Li.)

have large stains surrounded by numerous, peripheral, elongated bloodstains. A *flow pattern* is caused by the movement of a large volume of blood on a surface either due to gravity or to the movement of the target such as a victim or postmortem disturbance. A *pool* is a bloodstain resulting from the accumulation of liquid blood on a surface (Figure 2.15). Sometimes, air bubbles in the blood may cause a *bubble ring* pattern (Figure 2.15). If blood is coagulated, gelatinous *blood clots* can be observed. Additionally, a *serum stain*, which consists of the liquid portion of the blood after a clot is formed, may also be present.

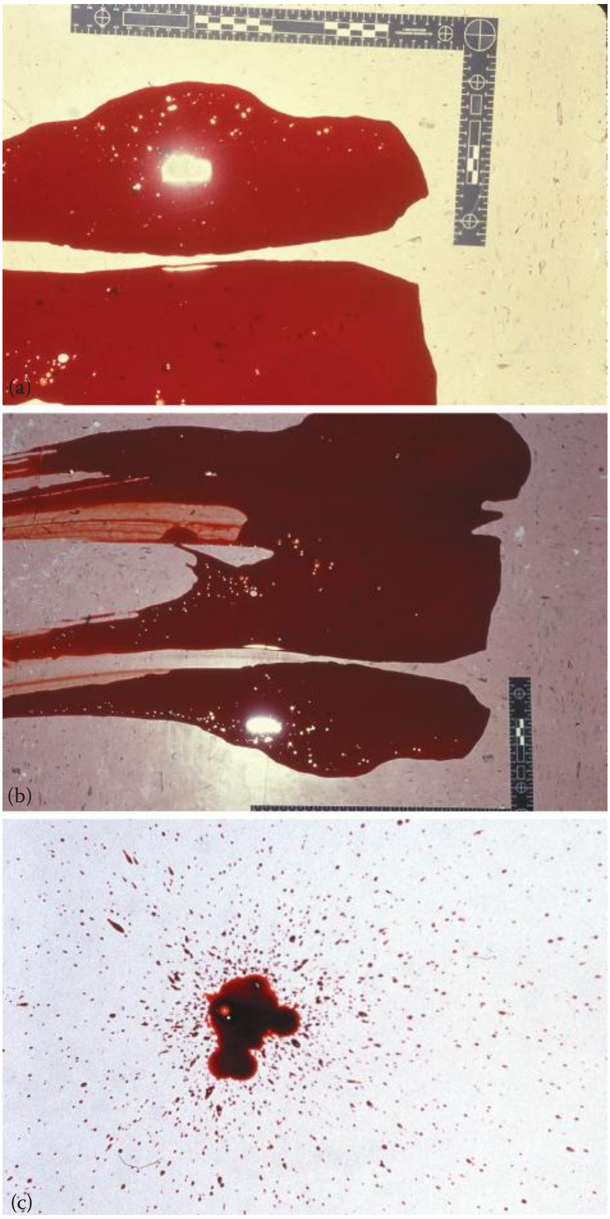

Figure 2.15 Pool and bubble ring patterns. (a) A pool pattern, (b) a disturbed pool pattern, and (c) a splash pattern with a bubble ring. Bubble rings are also present in (a) and (b). (© Richard C. Li.)

2.5.2 Transfer Bloodstains

A *transfer bloodstain*, based on SWGSTAIN, is "a bloodstain resulting from contact between a blood-bearing surface and another surface." For example, a *swipe pattern* is "a bloodstain pattern resulting from the transfer of blood from a blood-bearing surface onto another surface, with characteristics that indicate relative motion between the two surfaces." For example, bloody impressions can provide information about the shape, the size, and the pattern of the objects such as finger ridges, hands, and shoe soles. Examples of hand and shoe swipe patterns are shown in Figures 2.16 and 2.17. A *wipe pattern* is "an altered bloodstain pattern resulting from an object moving through a preexisting wet bloodstain." Examples of wipe patterns are shown in Figures 2.18 and 2.19. A *perimeter* stain, a type of wipe pattern, is a bloodstain that is disturbed before it is dried but it maintains the peripheral characteristics of the original stain (Figure 2.20). Perimeter stain patterns can be useful for the estimation of sequential events of acts. The pattern can also be used to estimate a time frame between the time of bleeding and the subsequent act. However, the drying time of a blood drop varies based on the surrounding conditions. Therefore, it is necessary to carry out a crime scene reconstruction under similar conditions to those of the scene to make such estimations.

2.5.3 Projected Bloodstains

A *projected bloodstain* is formed when a volume of blood is deposited on a surface under a pressure or a force that is greater than the force of gravity. For example, an *impact pattern* is

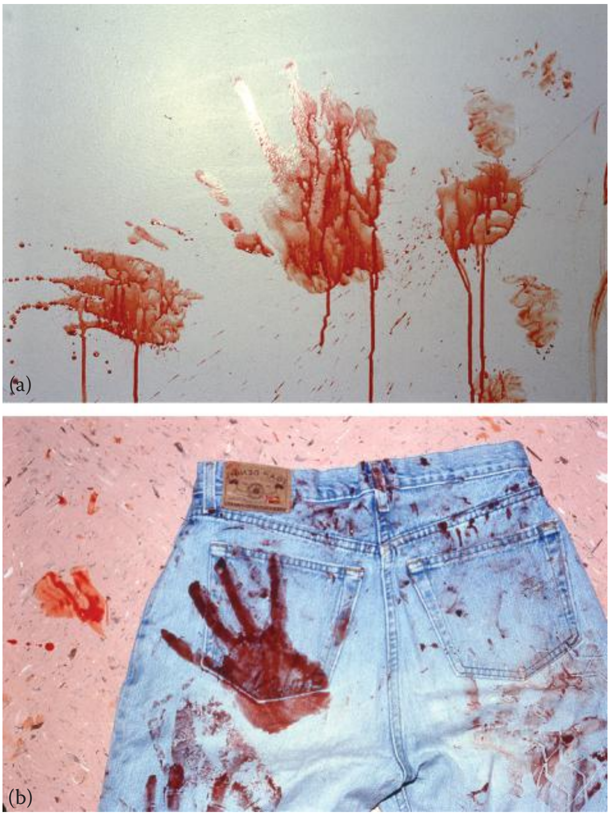

Figure 2.16 Bloody impressions. Bloody handprints are present on (a) a wall and (b) fabric. (© Richard C. Li.)

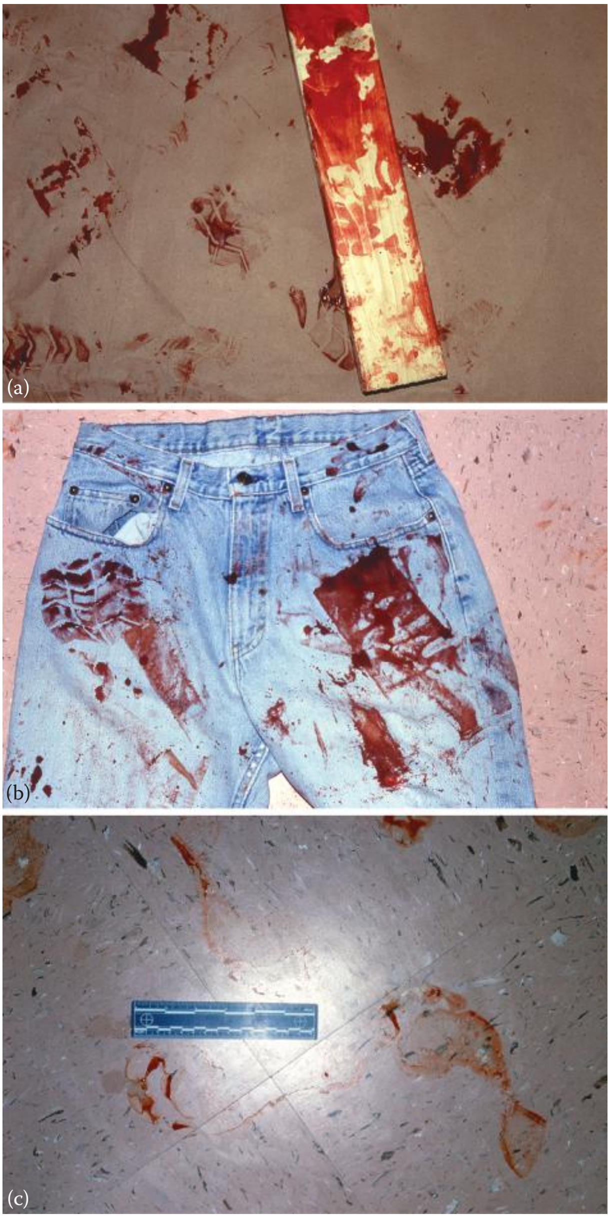

Figure 2.17 Bloody impressions. Bloody shoe prints on (a) paper and (b) fabric; and (c) bloody footprints on tile. (© Richard C. Li.)

formed when an object strikes liquid blood (Figure 2.21). A *cast-off pattern* is formed when blood drops are released from a moving blood-bearing object (Figure 2.22). Some spatter patterns are often associated with a wound penetrated by a projectile (Figure 2.23). A *forward spatter* is formed when blood drops travel from an exit wound in the same direction as a projectile, while a *back spatter* is formed when blood drops travel from an entry wound in the opposite direction of a projectile. Sometimes, internal bleeding caused by an injury may block the airway. An *expiration pattern* is formed when blood is forced by airflow through the

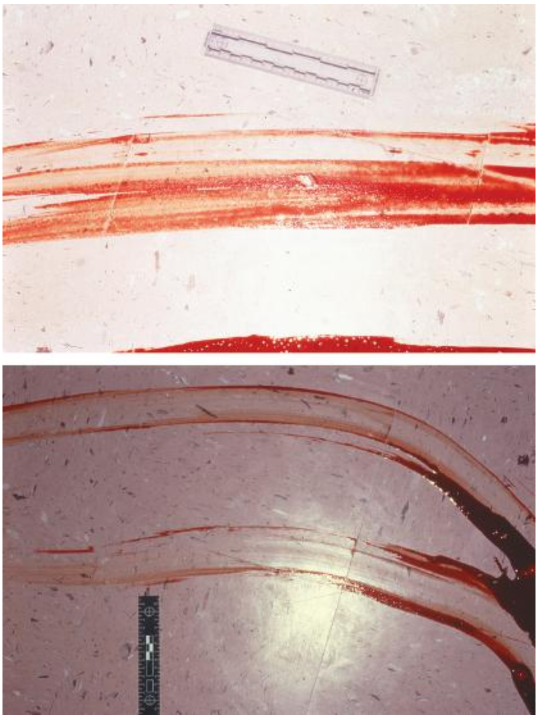

Figure 2.18 Wipe patterns. (© Richard C. Li.)

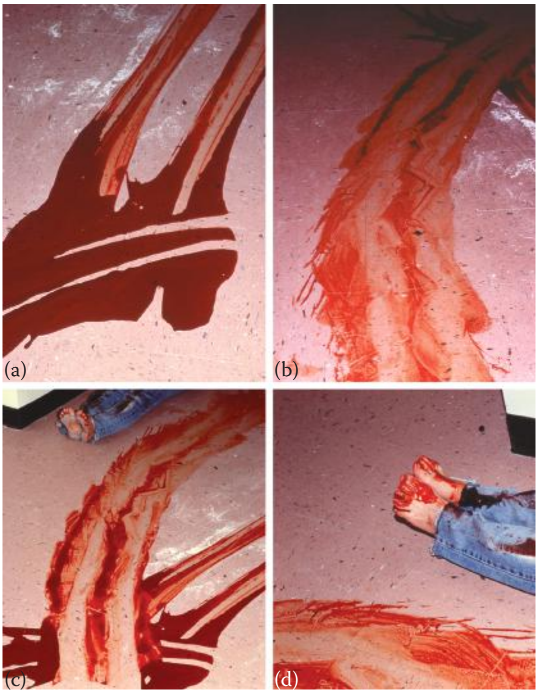

Figure 2.19 A wipe pattern caused by dragging a body through a pool of blood. (a) A pool of blood. Sections of the wipe pattern caused by dragging are shown in (b), (c), and (d). (© Richard C. Li.)

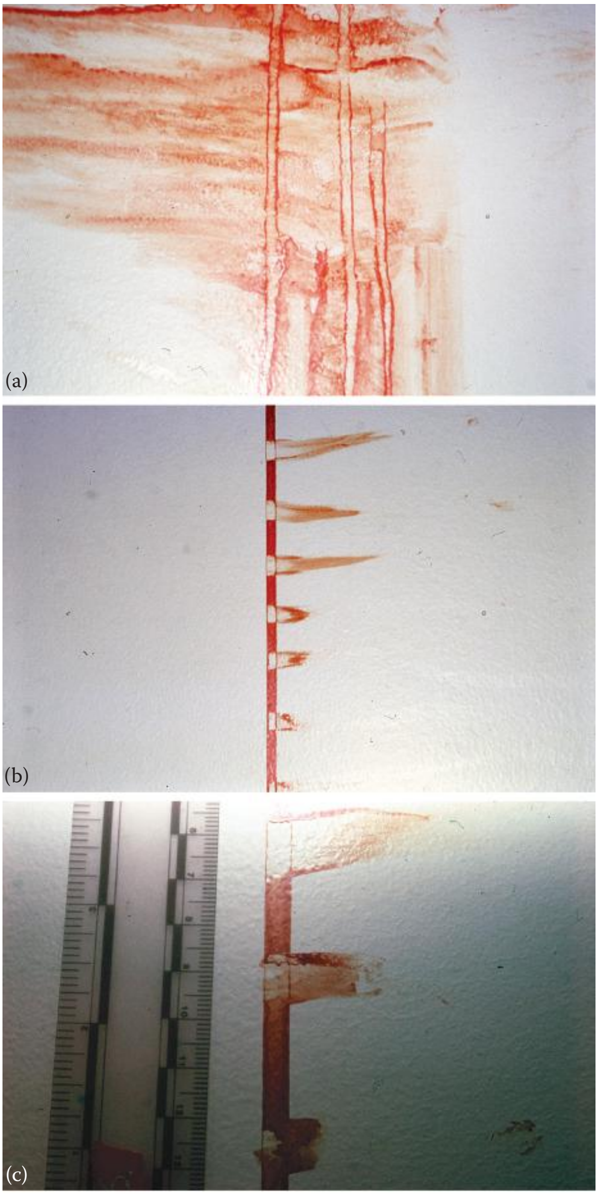

Figure 2.20 Perimeter stains. (a) Peripheral characteristics of the original stains are shown. Perimeter stains were created at different periods of time after the original stain was formed: (b) midrange view and (c) close-up view. (© Richard C. Li.)

trachea and out of the nose or mouth (Figure 2.24). An *arterial spurt pattern* is associated with wounds damaging arterial blood vessels where bloodstains are driven by arterial pressure. Although the shape of arterial patterns varies, these patterns usually have a series of large spurts with fluctuations corresponding to the systolic and the diastolic blood pressures. At a crime scene, if the projectile of bloodstains is blocked by an object, a *void pattern* is formed, which exhibits an area where there is an absence of blood surrounded by continuously distributed bloodstains.

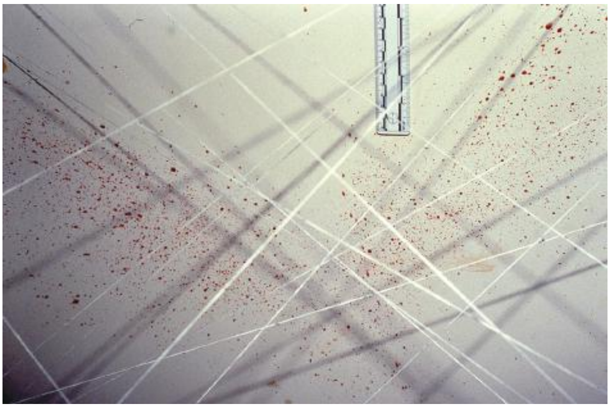

Figure 2.21 An impact bloodstain pattern as a result of using blunt force. (© Richard C. Li.)

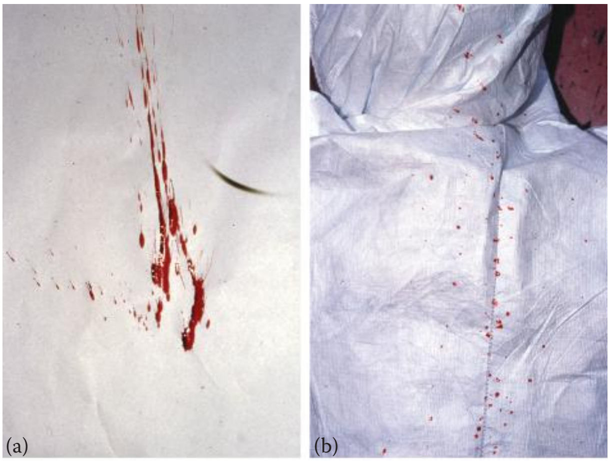

Figure 2.22 Cast-off patterns. Spatter stains are projected onto (a) a covered wall and (b) a lab coat. (© Richard C. Li.)

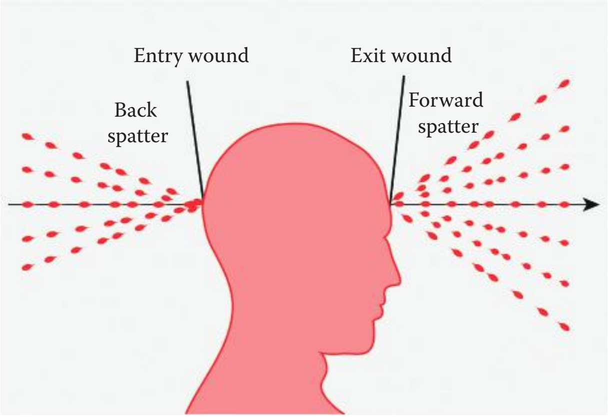
Entry wound
Exit wound
Back spatter
Forward spatter

Figure 2.23 Forward and back spatter patterns. Arrow, the direction of a projectile. (© Richard C. Li.)

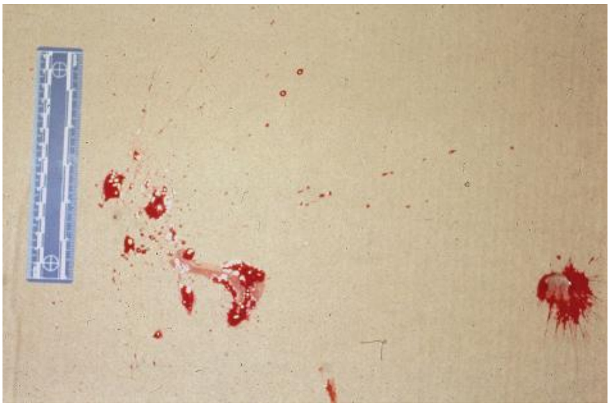

Figure 2.24 An expiration pattern. (© Richard C. Li.)

Bibliography

Adam, C.D., Fundamental studies of bloodstain formation and characteristics. *Forensic Sci Int*, 2012, **219**(1–3): 76–87.

Adelson, L., The coroner of Elsinore. *New Engl J Med*, 1960, **262**(5): 229–234.

Attinger, D., et al., Fluid dynamics topics in bloodstain pattern analysis: Comparative review and research opportunities. *Forensic Sci Int*, 2013, **231**(1–3): 375–396.

Behrooz, N., L. Hulse-Smith, and S. Chandra, An evaluation of the underlying mechanisms of bloodstain pattern analysis error. *J Forensic Sci*, 2011, **56**(5): 1136–1142.

Benecke, M. and L. Barksdale, Distinction of bloodstain patterns from fly artifacts. *Forensic Sci Int*, 2003, **137**(2–3): 152–159.

Buck, U., et al., 3D bloodstain pattern analysis: Ballistic reconstruction of the trajectories of blood drops and determination of the centres of origin of the bloodstains. *Forensic Sci Int*, 2011, **206**(1–3): 22–28.

Byard, R.W., et al., Blood stain pattern interpretation in cases of fatal haemorrhage from ruptured varicose veins. *J Forensic Leg Med*, 2007, **14**(3): 155–158.

Camana, F., Determining the area of convergence in bloodstain pattern analysis: A probabilistic approach. *Forensic Sci Int*, 2013, **231**(1–3): 131–136.

Connolly, C., M. Illes, and J. Fraser, Affect of impact angle variations on area of origin determination in bloodstain pattern analysis. *Forensic Sci Int*, 2012, **223**(1–3): 233–240.

Davidson, P.L., et al., Physical components of soft-tissue ballistic wounding and their involvement in the generation of blood backspatter. *J Forensic Sci*, 2012, **57**(5): 1339–1342.

de Bruin, K.G., R.D. Stoel, and J.C. Limborgh, Improving the point of origin determination in bloodstain pattern analysis. *J Forensic Sci*, 2011, **56**(6): 1476–1482.

de Castro, T., et al., Interpreting the formation of bloodstains on selected apparel fabrics. *Int J Legal Med*, 2013, **127**(1): 251–258.

Denison, D., et al., Forensic implications of respiratory derived blood spatter distributions. *Forensic Sci Int*, 2011, **204**(1–3): 144–155.

Donaldson, A.E., et al., Using oral microbial DNA analysis to identify expirated bloodspatter. *Int J Legal Med*, 2010, **124**(6): 569–576.

Donaldson, A.E., et al., Characterising the dynamics of expirated bloodstain pattern formation using high-speed digital video imaging. *Int J Legal Med*, 2011, **125**(6): 757–762.

Emes, A., Expirated blood—A review. *J Can Soc Forensic Sci*, 2001, **34**(4): 197–203.

Finnis, J., J. Lewis, and A. Davidson, Comparison of methods for visualizing blood on dark surfaces. *Sci Justice*, 2013, **53**(2): 178–186.

Fujikawa, A., et al., Changes in the morphology and presumptive chemistry of impact and pooled bloodstain patterns by *Lucilia sericata* (Meigen) (Diptera: Calliphoridae). *J Forensic Sci*, 2011, **56**(5): 1315–1318.

Illes, M. and M. Boue, Robust estimation for area of origin in bloodstain pattern analysis via directional analysis. *Forensic Sci Int*, 2013, **226**(1–3): 223–229.

Karger, B., et al., Bloodstain pattern analysis—Casework experience. *Forensic Sci Int*, 2008, **181**(1–3): 15–20.

Kettner, M., F. Ramsthaler, and A. Schnabel, "Bubbles"—A spot diagnosis. *J Forensic Sci*, 2010, **55**(3): 842–844.

Kunz, S.N., H. Brandtner, and H. Meyer, Unusual blood spatter patterns on the firearm and hand: A backspatter analysis to reconstruct the position and orientation of a firearm. *Forensic Sci Int*, 2013, **228**(1–3): e54–e57.

Larkin, B.A. and C.E. Banks, Preliminary study on the effect of heated surfaces upon bloodstain pattern analysis. *J Forensic Sci*, 2013, **58**(5): 1289–1296.

Larkin, B.A.J., et al., Crime scene investigation III: Exploring the effects of drugs of abuse and neurotransmitters on bloodstain pattern analysis. *Anal Methods*, 2012, **4**(3): 721–729.

Lee, W.C., et al., Statistical evaluation of alternative light sources for bloodstain photography. *J Forensic Sci*, 2013, **58**(3): 658–663.

Liesegang, J., Bloodstain pattern analysis—Blood source location. *J Can Soc Forensic Sci*, 2004, **37**(4): 215–222.

Maclean, B., K. Powley, and D. Dahlstrom, A case study illustrating another logical explanation for high velocity impact spatter. *J Can Soc Forensic Sci*, 2001, **34**(4): 191–195.

Makovicky, P., et al., The use of trigonometry in bloodstain analysis. *Soud Lek*, 2013, **58**(2): 20–25.

Maloney, A., et al., One-sided impact spatter and area-of-origin calculations. *J Forensic Ident*, 2011, **61**(2): 123–135.

Nikolic, S. and V. Zivkovic, Bloodstain pattern in the form of gushing in a case of fatal exsanguination due to ruptured varicose vein. *Med Sci Law*, 2011, **51**(1): 61–62.

Peschel, O., et al., Blood stain pattern analysis. *Forensic Sci Med Pathol*, 2011, **7**(3): 257–270.

Pizzola, P.A., et al., Commentary on "3D bloodstain pattern analysis: Ballistic reconstruction of the trajectories of blood drops and determination of the centres of origin of the bloodstains" by Buck et al. [Forensic Sci. Int. 206 (2011) 22–28]. *Forensic Sci Int*, 2012, **220**(1–3): e39–e40; author reply e41.

Power, D.A., et al., PCR-based detection of salivary bacteria as a marker of expirated blood. *Sci Justice*, 2010, **50**(2): 59–63.

Ramsthaler, F., et al., Drying properties of bloodstains on common indoor surfaces. *Int J Legal Med*, 2012, **126**(5): 739–746.

Randall, B., Blood and tissue spatter associated with chainsaw dismemberment. *J Forensic Sci*, 2009, **54**(6): 1310–1314.

Raymond, M.A., E.R. Smith, and J. Liesegang, Oscillating blood droplets—Implications for crime scene reconstruction. *Sci Justice*, 1996, **36**(3): 161–171.

Raymond, M.A., E.R. Smith, and J. Liesegang, The physical properties of blood—Forensic considerations. *Sci Justice*, 1996, **36**(3): 153–160.

Sauvageau, A., et al., Bloodstain pattern analysis in a case of fatal varicose vein rupture. *Am J Forensic Med Pathol*, 2007, **28**(1): 35–37.

Stephens, B.G. and T.B. Allen, Back spatter of blood from gunshot wounds. Observations and experimental simulation. *J Forensic Sci*, 1983, **28**(2): 437–439.

Striman, B., et al., Alteration of expirated bloodstain patterns by *Calliphora vicina* and *Lucilia sericata* (Diptera: Calliphoridae) through ingestion and deposition of artifacts. *J Forensic Sci*, 2011, **56**(Suppl 1): S123–S127.

Taylor, M.C., et al., The effect of firearm muzzle gases on the backspatter of blood. *Int J Legal Med*, 2011, **125**(5): 617–628.

Thanakiatkrai, P., A. Yaodam, and T. Kitpipit, Age estimation of bloodstains using smartphones and digital image analysis. *Forensic Sci Int*, 2013, **233**(1–3): 288–297.

Trombka, J.I., et al., Crime scene investigations using portable, non-destructive space exploration technology. *Forensic Sci Int*, 2002, **129**(1): 1–9.

White, B., Bloodstain patterns on fabrics: The effect of drop volume, dropping height and impact angle. *J Can Soc Forensic Sci*, 1986, **19**(1): 3–36.

Yen, K., et al., Blood-spatter patterns: Hands hold clues for the forensic reconstruction of the sequence of events. *Am J Forensic Med Pathol*, 2003, **24**(2): 132–140.

Yi, R. and Y. Wang, The experimental observation of secondary spatter bloodstain in morphology. *Chin J Forensic Med*, 2011, **26**(1): 26–29.

3

Forensic Biology

A Subdiscipline of Forensic Science

Forensic laboratories provide scientific analysis, evidence evaluation, and consultations to various criminal justice agencies for the investigation of criminal cases. Additionally, forensic laboratories provide expert testimony related to the resolution of criminal cases to the courts.

3.1 Common Disciplines of Forensic Laboratory Services

Many of the disciplines of the forensic laboratory services are commonly practiced in various municipal, county, state, and federal forensic laboratories in the United States. Forensic biology is a subdiscipline of forensic science. A full range of forensic laboratory services, known as "full service," usually includes: crime scene investigation (Figure 3.1); latent print examination (Figure 3.2); forensic biology (Figure 3.3); controlled substance analysis (Figure 3.4); postmortem toxicology (Figure 3.5); questioned document examination (Figure 3.6); firearm, toolmark, and other impression evidence examination (Figures 3.7 through 3.10); explosive and fire debris examination (Figure 3.11); and transfer (trace) evidence examination (Figures 3.12 and 3.13). Table 3.1 describes the services that are normally provided by a forensic laboratory with their respective analyses.

3.2 Laboratory Analysis of Biological Evidence

Forensic biology uses scientifically accepted protocols to analyze biological evidence. Laboratory analysis (Figures 3.14 and 3.15) utilizes scientific techniques for the examination of evidence, the reconstruction of a crime scene, the identification of biological fluids, and the comparison of individual characteristics of biological evidence.

3.2.1 Identification of Biological Evidence

The identification of biological evidence is the first step that is performed before further analyses are carried out. This includes the identification of biological fluids such as blood, saliva, and semen; this process is discussed in more detail in subsequent chapters (Chapters 12 through 14). The *identification* is based on a comparison of *class characteristics*—a set of characteristics that allows a sample to be placed in a category with similar materials. By comparing the

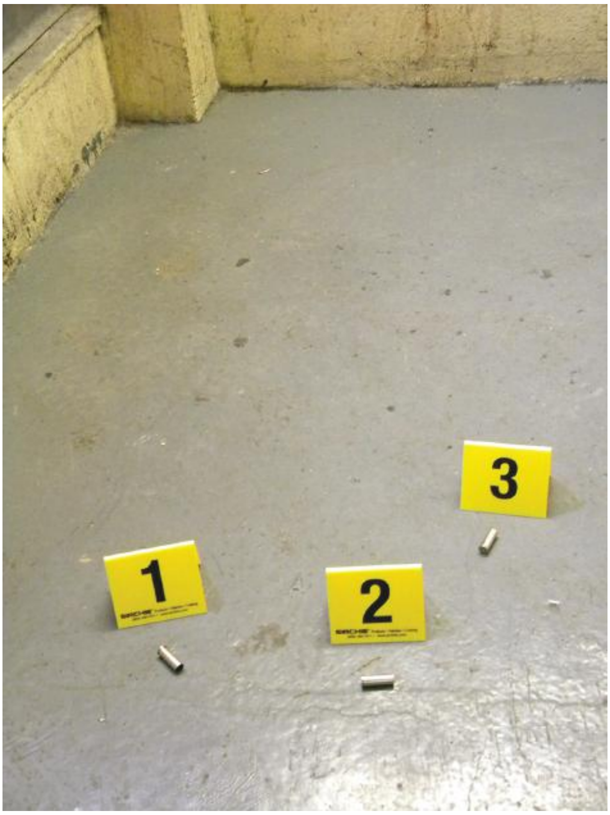

Figure 3.1 Crime scene investigation. Recovery of fired casings at the scene can aid in determining the position of the shooter. (© Richard C. Li.)

class characteristics of a sample with known standards of its class, biological samples can be identified.

3.2.2 Comparison of Individual Characteristics of Biological Evidence

Individual characteristics refer to the unique characteristics of both the evidence and a reference sample such as fingerprints, which share a common origin to a high degree of certainty. An example of biological evidence possessing individual characteristics is DNA polymorphisms. In the case of biological evidence, current forensic DNA profiling can compare individual characteristics of DNA evidence with a known reference sample. It is possible to determine that a biological stain originated from a particular individual, which is useful for human identification. The examination of individual characteristics of evidence can also exclude the possibility of a common origin. The specific methods utilized for the individualization of evidence are also discussed in subsequent chapters (Chapters 19 through 23).

3.2.3 Reporting Results and Expert Testimony

After the analysis of evidence is completed, a report is prepared based on the results of the analysis, which may include sections discussing the specific evidence analyzed, the method of analysis used, the results obtained, and the conclusions drawn. In the case of DNA evidence, the strength of the conclusion is usually evaluated via statistical computations (Chapter 25). A forensic scientist often serves as an expert witness whose testimony

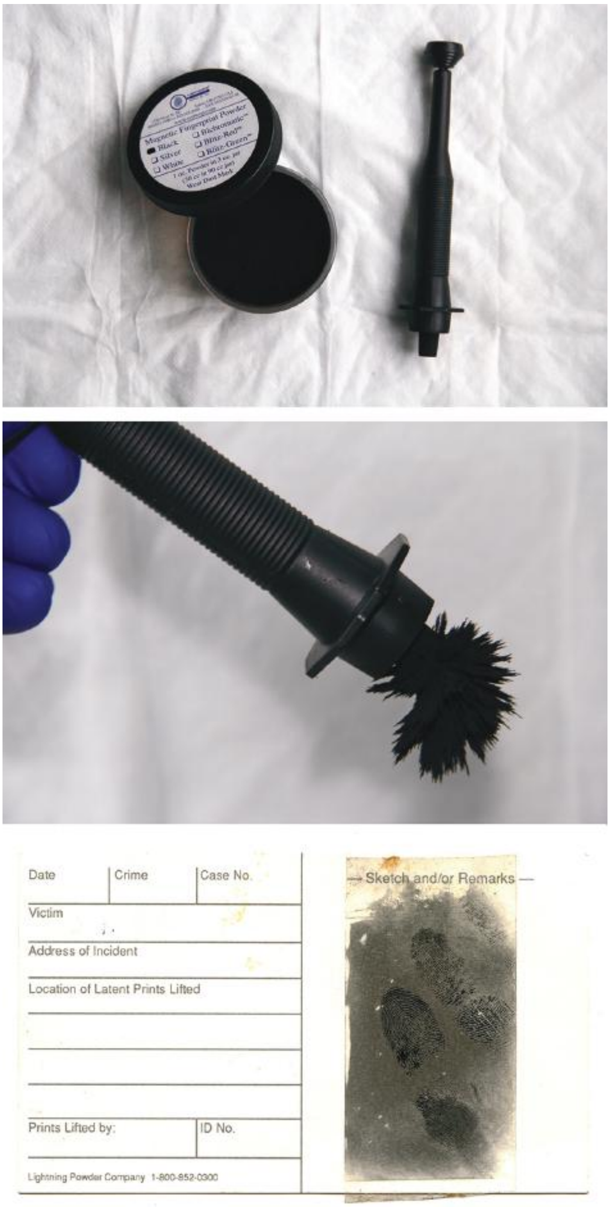

Date | Crime | Case No.
Victim
Address of Incident
Location of Latent Prints Lifted
Prints Lifted by: | ID No.
Lightning Powder Company 1-800-852-0300

— Sketch and/or Remarks —

Figure 3.2 Developing latent fingerprints by dusting methods. Using magnetic fingerprint powder (top) that is held by a magnetic applicator (middle). Fingerprints (bottom) dusted by magnetic fingerprint powder, tape lifted and preserved on a fingerprint card. (© Richard C. Li.)

provides professional opinions about the evidence analyzed. Based on the federal rules of evidence, an expert witness is qualified based on his or her knowledge, skill, experience, training, or education, and may give an opinion to the court that is relevant to the analyses conducted. However, forensic scientists must also communicate their findings to attorneys, judges, and members of a jury. This requires the translation of technical information into layman's terms.

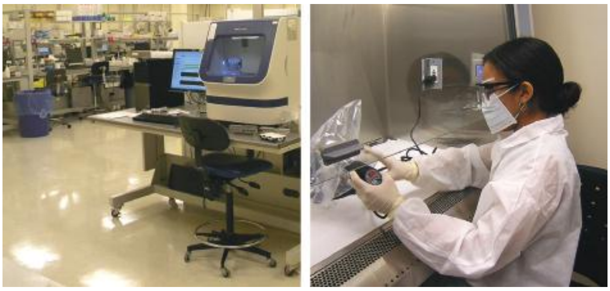

Figure 3.3 A section of a forensic biology laboratory showing an automated electrophoresis instrument used for forensic DNA profiling (left). Processing biological evidence in a biosafety cabinet (right). (© Richard C. Li.)

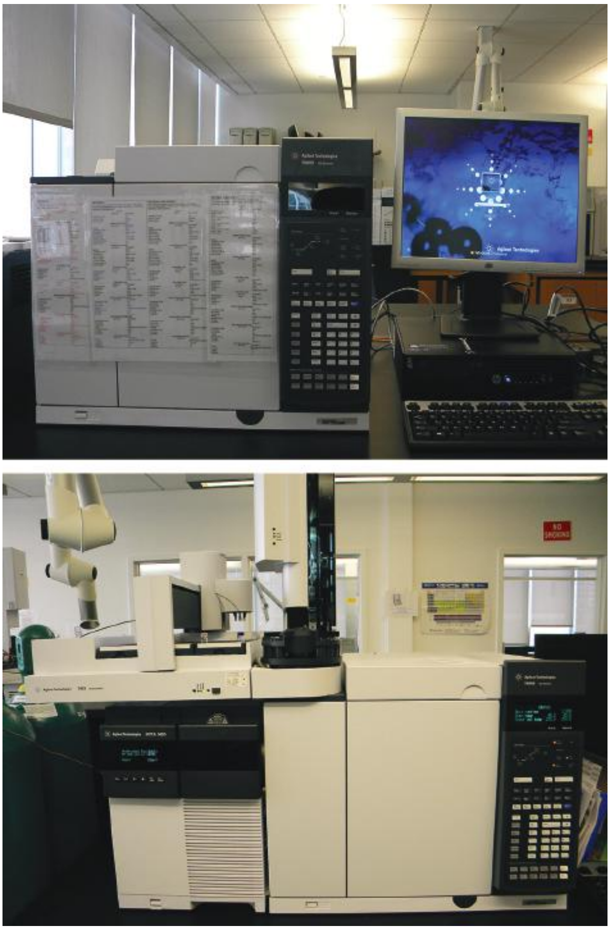

Figure 3.4 Gas chromatograph (top) and gas chromatograph-mass spectrometer (bottom) used for controlled substance analysis. (© Richard C. Li.)

Figure 3.5 Tissue samples (left) and gas chromatograph-mass spectrometer (right) for forensic postmortem toxicological analysis. (© Richard C. Li.)

Figure 3.6 A digital imaging system, using multiple illumination sources ranging from ultraviolet to infrared wavelengths, for examining altered and counterfeit documents (left). An electrostatic imaging system for detecting indented writing on questioned documents (right). This device generates an electrostatic image of indented writing, which is then visualized using charge-sensitive toners. (© Richard C. Li.)

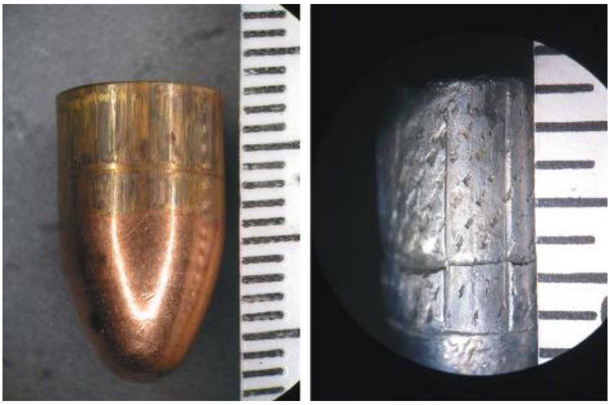

Figure 3.7 Striation marks on fired bullets can be analyzed to match a bullet to a gun. (Courtesy of P. Diaczuk.)

3.3 Forensic Science Services Related to Forensic Biology

A number of specialized forensic science services beyond those provided by forensic laboratories are routinely available to law enforcement agencies. For example, forensic services related to biological evidence and those involving more specialized analysis are available. These services are important aids to a criminal investigation and require the expertise of individuals who have highly specialized skills.

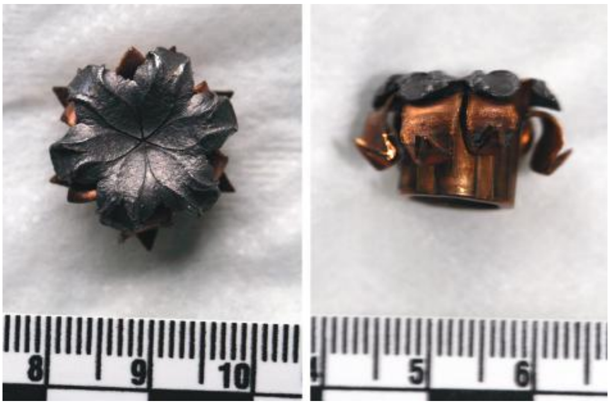

Figure 3.8 Fired hollowpoint bullets (.45 caliber ACP Winchester). Striation marks are visible on the side view of a bullet (right). (© Richard C. Li.)

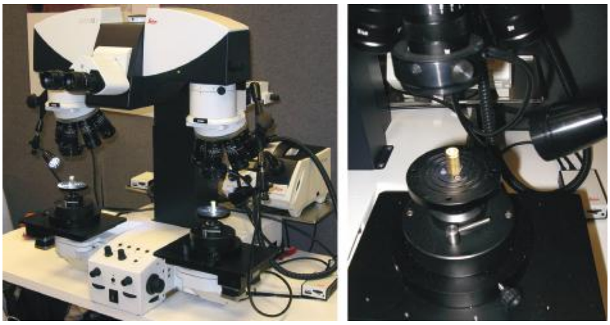

Figure 3.9 A comparison microscope is used for the simultaneous comparison of two items of firearm evidence side by side. (© Richard C. Li.)

3.3.1 Forensic Pathology

When a death is deemed suspicious or unexplained, medical examiners frequently perform autopsies to determine the exact cause (Figures 3.16 and 3.17). The manner of death is classified into one of five categories based on the circumstances: natural, homicide, suicide, accident, or undetermined. Additionally, a medical examiner participating in a criminal investigation is often responsible for estimating the time of death.

3.3.2 Forensic Anthropology

Forensic anthropology is the identification and the examination of human skeletal remains (Figures 3.18 and 3.19). Skeletal remains can reveal a number of individual characteristics that can be useful in attempting to identify an individual. An examination of bones may reveal an individual's origin, sex, approximate age, race, and the presence of a skeletal injury. A forensic anthropologist may also assist in creating facial reconstructions to aid in the identification of skeletal remains or may be called on to help collect and organize bone fragments in the course of identifying victims of mass disasters such as plane crashes as well as victims in mass graves discovered after wars or genocides.

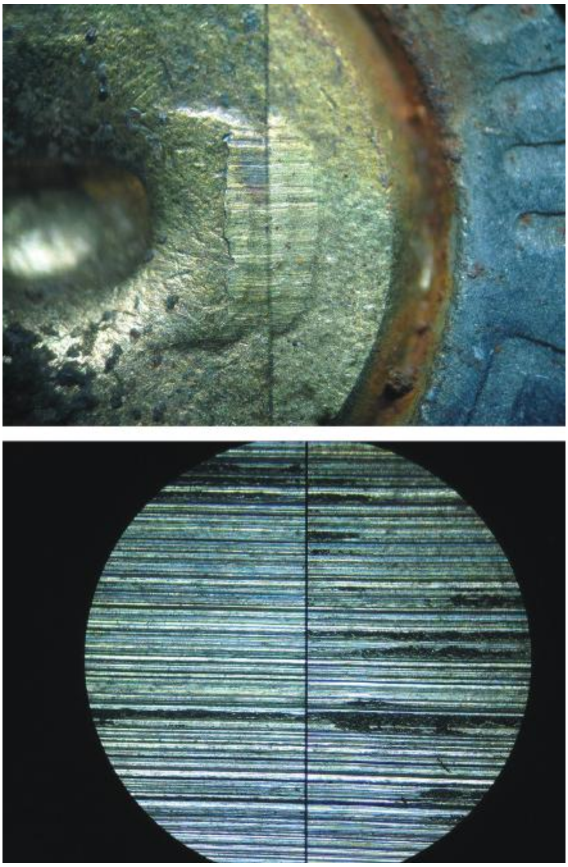

Figure 3.10 Comparing the striations between evidence and reference samples using a comparison microscope: casings (top) and bullets (bottom). (Courtesy of P. Diaczuk.)

Figure 3.11 Scanning electron microscope used in the analysis of gunshot residue and explosives. (© Richard C. Li.)

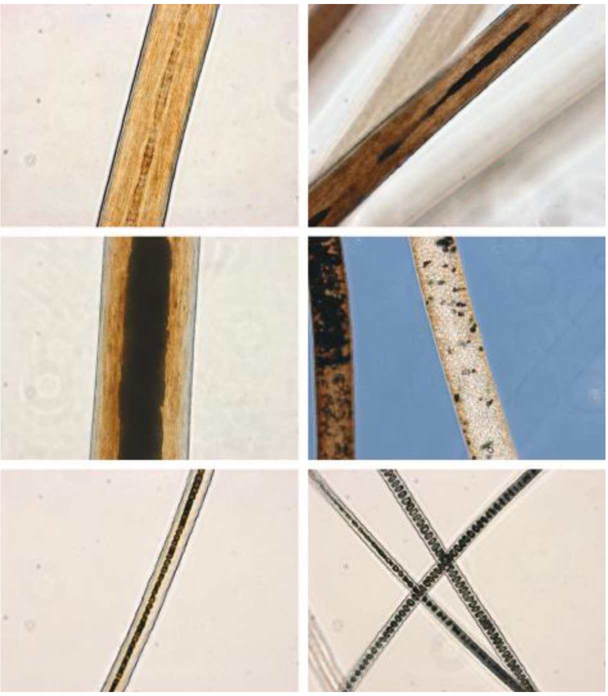

Figure 3.12 Trace evidence such as hairs can be transferred during the acts of a violent crime. The analysis and comparisons of these types of trace evidence can potentially establish a link between a suspect or a victim and a crime scene. Hairs from human (top), horse (middle left), deer (middle right), dog (bottom left), and cat (bottom right). (© Richard C. Li.)

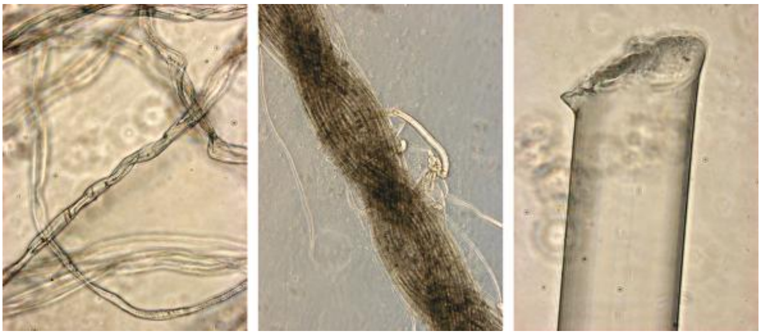

Figure 3.13 Trace evidence: fibers. Cotton (left), nylon (middle), and polyester (right). (© Richard C. Li.)

3.3.3 Forensic Entomology

The study of insects in relation to a criminal investigation is known as forensic entomology. This forensic discipline is valuable for estimating the time of death when the circumstances surrounding the crime are otherwise unknown. The stages of development of certain insect species present in or on a body can be identified and allow a forensic entomologist to approximate how long the body was left exposed (Figure 3.20).

Table 3.1 Common Services Provided by US Forensic Laboratories

Service	Function	Method
Crime scene investigation	Evidence recognition, documentation, collection, and preservation	Crime scene responses and related endeavors are diverse and vary with case and type of evidence
Latent print examination	Analysis of friction ridge detail in fingerprints Activities include visualization, recording, comparison, storage, and recovery of latent prints	Alternate light sources, physical (powder) and chemical enhancements Direct lifts, photography, and digital imaging Use of an Automated Fingerprint Identification System (AFIS) database
Forensic biology	Identification of biological fluids (blood, semen, and saliva) DNA profiling for individualization	Serological and biochemical methods Polymerase chain reaction (PCR)-based methods Automated electrophoresis platforms Use of Combined DNA Index System (CODIS)
Controlled substance analysis	Identification and quantification of drugs present in submitted evidence	Microscopic, chemical, chromatographic, and spectroscopic methodologies Gas chromatography–mass spectrometry or infrared spectrophotometry
Postmortem toxicology	Determination of concentrations of substances and their metabolites in biological fluids or tissues	Immunoassays and chemical methods Confirmatory techniques such as gas and liquid chromatography–mass spectrometry
Questioned document examination	Investigation of forgeries, tracings, disguised handwritings, computer manipulation of images, and recovery of altered documents Analysis of papers, inks, toners, word processors, typewriters, copiers, and printers	Macroscopic and microscopic comparisons Chromatographic and spectroscopic methods
Firearm and toolmark examination	Identification of firearms, tools, and other implements (expertise achieved predominantly through experience)	Microscopic comparisons of questioned and authenticated impressions Comparison of striae on recovered bullets Use of National Integrated Ballistics Information Network (NIBIN)
Explosive and fire debris examination	Identification, recovery, and detection of bulk explosives, residues, debris, and accelerants	Microscopic, spectroscopic, and chromatographic methods Gas chromatography–mass spectrometry may be needed to adequately characterize sample
Trace evidence examination	Analysis of transferred evidence such as hairs, fibers, soil, paints, and glass	Microscopic analysis of evidence with gas chromatograph-mass spectrometers, FTIR microscopes, scanning electron microscopes, basic and advanced microscopy, and capillary electrophoresis

Source: Adapted from US Department of Justice, Office of Justice Programs, National Institute of Justice (NIOJ), *Forensic Science: Review of Status and Needs*, 1999, US Department of Justice, Washington, DC.

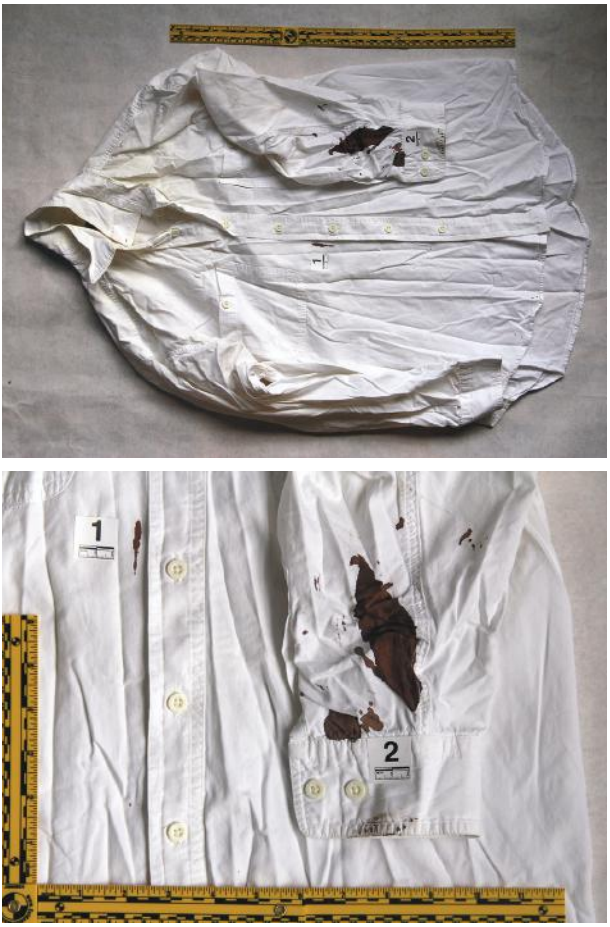

Figure 3.14 Photographic documentation of a bloodstained shirt after visual examination. (© Richard C. Li.)

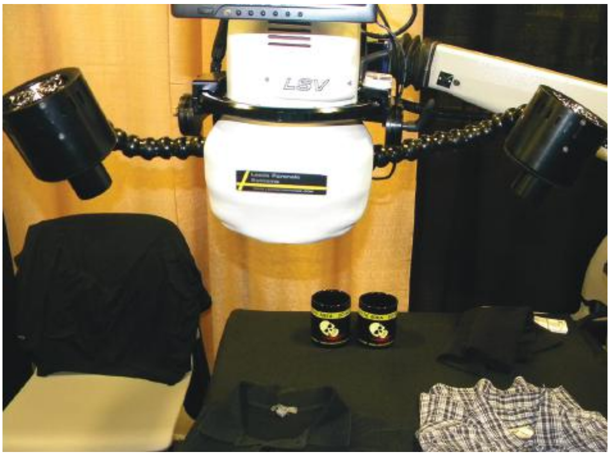

Figure 3.15 A multiwavelength viewing and imaging device used for the examination of various types of evidence, including bodily fluid stains. (© Richard C. Li.)

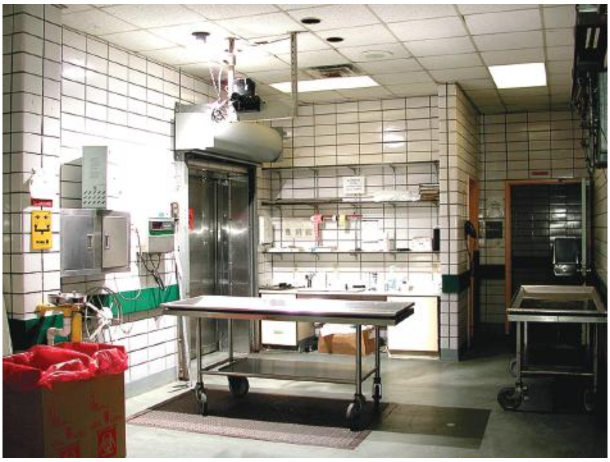

Figure 3.16 View of a forensic pathology facility. (Courtesy of G. Ledwell.)

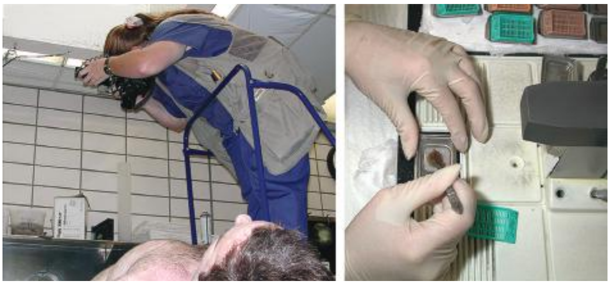

Figure 3.17 Photographic documentation prior to autopsy (left). Preparing specimens for histological sections for forensic pathological examination (right). (Courtesy of G. Ledwell.)

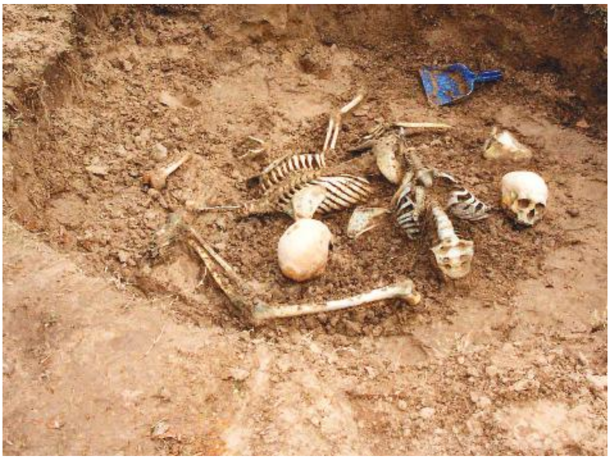

Figure 3.18 Buried human skeletons recovered by forensic anthropologists. (Courtesy of H. Brewster.)

Figure 3.19 Human skull recovered at the scene. (Courtesy of H. Brewster.)

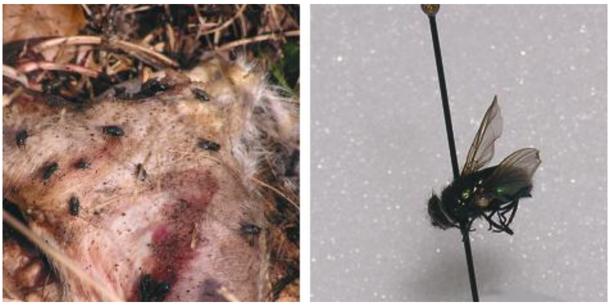

Figure 3.20 Insects found on a dead animal (left). Blow fly specimen. This insect is commonly encountered at crime scenes (right). (Courtesy of K. Wendler.)

3.3.4 Forensic Odontology

Practitioners of forensic odontology participate in the identification of victims whose bodies are left in an unrecognizable state. The characteristics of teeth, their alignment, and the overall structure of the mouth provide evidence that can identify a specific person. Dental records such as x-rays and dental casts allow a forensic odontologist to compare a set of dental remains with an alleged victim. Another application of forensic odontology in a criminal investigation is bite mark analysis (Figure 3.21). A forensic odontologist can analyze the marks left on a victim and compare them with the tooth structures of a suspect to make a comparison.

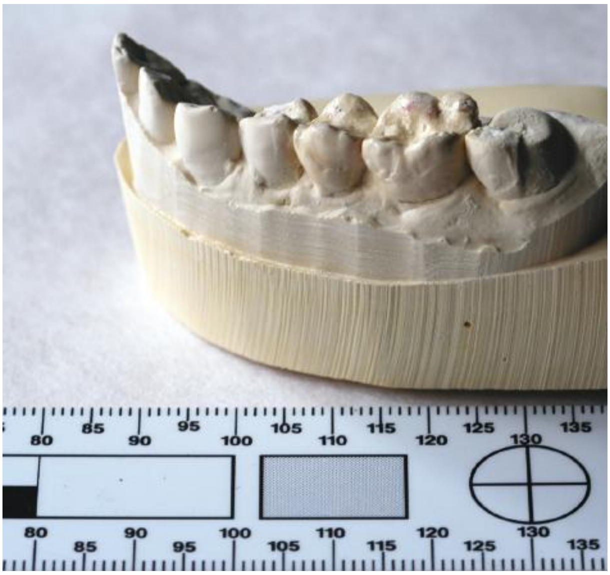
80
85
90
95
100
105
110
115
120
125
130
135

Figure 3.21 A dental cast can be utilized for identifying bite marks. (© Richard C. Li.)

3.4 Brief History of the Development of Forensic Biology

The developmental history of modern forensic biology spans three stages: (1) antigen polymorphism, (2) protein polymorphism, and (3) DNA polymorphism. Figure 3.22 illustrates this history.

3.4.1 Antigen Polymorphism

The human ABO blood groups were discovered in 1900 by Karl Landsteiner in a study of the causes of blood transfusion reactions. Landsteiner's discovery made blood transfusions feasible, and he received the Nobel Prize in 1930 when he revealed the four groups of human blood cells designated A, B, AB, and O. By the 1960s, a dozen more blood group systems had been characterized, and at least 29 systems are currently known (Chapter 18).

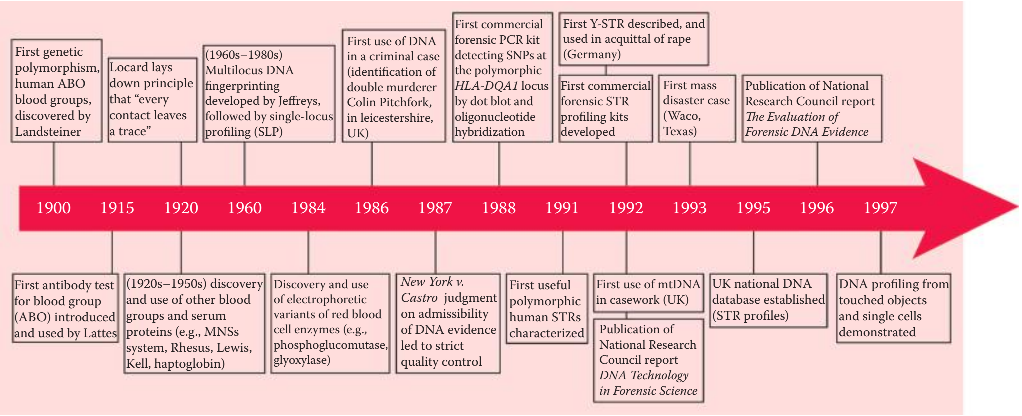
First genetic polymorphism, human ABO blood groups, discovered by Landsteiner
Locard lays down principle that "every contact leaves a trace"
(1960s–1980s) Multilocus DNA fingerprinting developed by Jeffreys, followed by single-locus profiling (SLP)
First use of DNA in a criminal case (identification of double murderer Colin Pitchfork, in leicestershire, UK)
First commercial forensic PCR kit detecting SNPs at the polymorphic *HLA-DQA1* locus by dot blot and oligonucleotide hybridization
First Y-STR described, and used in acquittal of rape (Germany)
First commercial forensic STR profiling kits developed
First mass disaster case (Waco, Texas)
Publication of National Research Council report *The Evaluation of Forensic DNA Evidence*
1900
1915
1920
1960
1984
1986
1987
1988
1991
1992
1993
1995
1996
1997
First antibody test for blood group (ABO) introduced and used by Lattes
(1920s–1950s) discovery and use of other blood groups and serum proteins (e.g., MNSs system, Rhesus, Lewis, Kell, haptoglobin)
Discovery and use of electrophoretic variants of red blood cell enzymes (e.g., phosphoglucomutase, glyoxylase)
New York v. Castro judgment on admissibility of DNA evidence led to strict quality control
First useful polymorphic human STRs characterized
First use of mtDNA in casework (UK)
Publication of National Research Council report *DNA Technology in Forensic Science*
UK national DNA database established (STR profiles)
DNA profiling from touched objects and single cells demonstrated

Figure 3.22 A brief history of the development of forensic biology. (From Jobling, M.A. and Gill, P., *Nat Rev Genet*, 5, 739–751, 2004. With permission.)

Subsequent studies found that the blood types in the ABO system were inherited, and the frequencies with which the four types appeared in specific human populations were found to differ. This led to the discovery of the first antigen polymorphic marker for use in human identification in forensic cases. In the past, forensic laboratories utilized blood group systems in a discipline known as forensic serology. While it is possible to exclude a suspect through the use of blood group typing, the evidence for the inclusion of a suspect is weak due to the high probability of a coincidental match between two unrelated persons.

3.4.2 Protein Polymorphism

Because of the limitations of antigen polymorphism, protein polymorphism was introduced for forensic identification (Chapter 18). Initially, a few polymorphisms in serum proteins and erythrocyte enzymes were reported. By the 1980s, however, approximately a hundred protein polymorphisms had been discovered. A few systems were commonly used in forensic laboratories, including the polymorphisms of erythrocyte enzymes, serum proteins, and hemoglobin. Blood groups and protein polymorphism analysis were combined in forensic investigations to lower the probability of a match between two unrelated individuals. However, more powerful methods were still sought.

3.4.3 DNA Polymorphism

The human genome contains all the necessary biological information for cellular and organ structure and function. It consists of the *nuclear genome* and the *mitochondrial genome* (Chapter 23 discusses the mitochondrial genome). The human nuclear genome, a set of 23 chromosomes, contains approximately 3 billion base pairs (bp). The Human Genome Project was initiated in 1990 to sequence the entire human nuclear genome. In 2003, 99% of the human genome, including the most important parts of the genome, was sequenced. Further analyses on the human genome sequences continue. The genome contains genes and intergenic noncoding sequences.

3.4.3.1 Genes and Related Sequences

Approximately 20,000–25,000 genes have been identified in the human genome, which encode the information for the synthesis of proteins. The functions of nearly half of these genes have been identified. Most encode the proteins that are responsible for the maintenance of the genome, the functioning of the cells, the immune response, and the structural proteins of cells.

Most human genes are discontinuous. The coding regions of genes are called *exons* and are separated by *introns*. During gene expression, the precursor messenger RNA transcript (pre-mRNA), consisting of both the exons and introns, is produced. The mRNA is a template for protein synthesis in which the sequence is based on a complementary strand of DNA. Through the process of splicing, the introns are removed and the exons are joined, producing the spliced mRNA form, which can be used for protein synthesis via the translation process. Other gene-related sequences include those responsible for gene transcription such as *promoter sequences*; those responsible for gene regulation such as *cis-regulatory sequences* (or enhancers); and *untranslated sequences*, which are transcribed but do not encode proteins. Figure 3.23 depicts the features of a representative human gene.

3.4.3.2 Intergenic Noncoding Sequences

More than 90% of the human genome sequence consists of intergenic noncoding sequences located between genes. The functions of these sequences are yet to be discovered. The intergenic noncoding sequences contain large quantities of various types of *repetitive DNA*, which falls into two categories: tandem repeats and genome-wide or interspersed repeats.

Tandem repeats are repeat units placed next to each other in an array. One type is called *satellite DNA* because of the observation of satellite bands containing DNA with tandem repeats

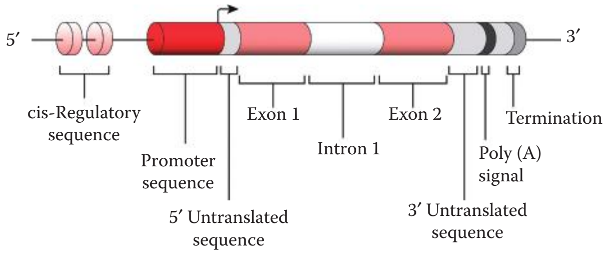
5′
3′
cis-Regulatory sequence
Promoter sequence
5′ Untranslated sequence
Exon 1
Intron 1
Exon 2
3′ Untranslated sequence
Poly (A) signal
Termination

Figure 3.23 Gene structure. Transcription, which can be regulated by the cis-regulatory sequence, is initiated at the transcription start site (arrow) near the promoter. The exons, noncoding introns, and the untranslated sequences are also shown. (© Richard C. Li.)

during density gradient centrifugation. Satellite DNA can be found at centromeres and telomeres consisting of regions composed of long stretches of tandem repeats. Minisatellites and microsatellites are two other types of shorter tandem repeats. *Minisatellites*, also known as *variable number tandem repeats* (VNTRs), form arrays of tandem repeats with a *repeat unit length* from several to hundreds of base pairs. In a *microsatellite*, also known as a *short tandem repeat* (STR) or a *simple sequence repeat* (SSR), the repeat unit length can be 2–6 bp long.

Mobile elements (*interspersed repeats*) are randomly located throughout the human genome (Figure 3.24). Two human types have been characterized: *DNA transposons* and *retrotransposons*. The mobile elements change their locations, a process called *transposition*, by which these sequences are inserted into a new site in the genome. The transposition of DNA transposons is through a "cut-and-paste" mechanism. During transposition, DNA transposons are excised from one site and inserted at a new site in the genome. In contrast, retrotransposons duplicate themselves during transposition and propagate throughout the genome, which is a copy-and-paste mechanism: a copy of the original retrotransposons is generated at the new site and the original copy is retained. Additionally, the transposition of retrotransposons requires an RNA intermediate, a process called *retrotransposition*. Retrotransposons have two subtypes: *long terminal repeat* (LTR) and *non-LTR retroposons*. The non-LTR retroposons can be further divided into two subtypes: *long interspersed elements* (LINEs) and *short interspersed elements* (SINEs). Alu elements are the most abundant type of human SINE. There are more than one million copies of Alu elements in the human genome (Figure 3.25). Some members of the Alu elements are

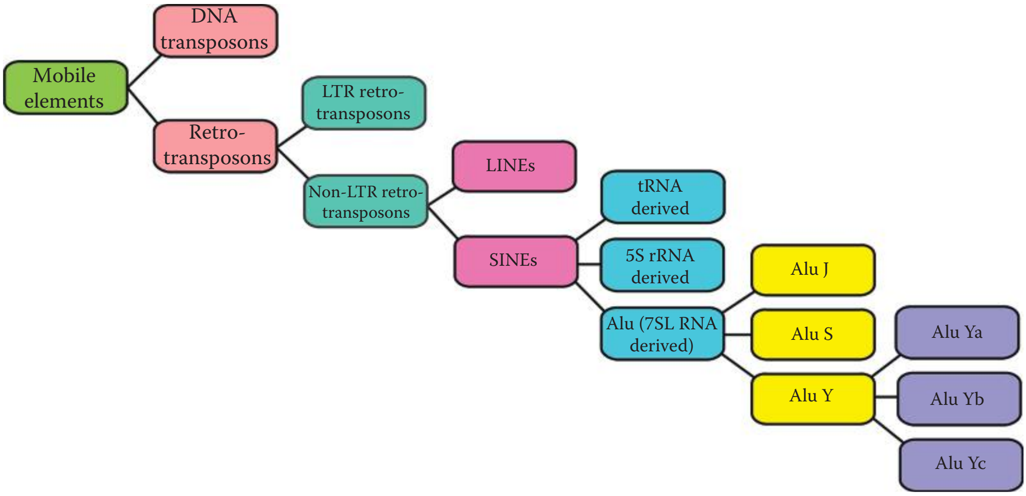
Mobile elements
DNA transposons
Retro-transposons
LTR retro-transposons
Non-LTR retro-transposons
LINEs
SINEs
tRNA derived
5S rRNA derived
Alu (7SL RNA derived)
Alu J
Alu S
Alu Y
Alu Ya
Alu Yb
Alu Yc

Figure 3.24 Mobile elements classification. (© Richard C. Li.)

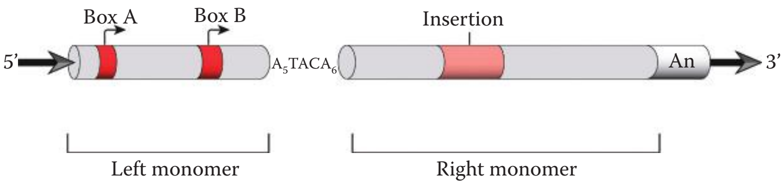

Figure 3.25 Structure of Alu elements. The name, Alu elements, was given since these elements usually contain DNA sequences cleaved by the restriction enzyme *Alu*I. The Alu element is a dimeric structure including two regions separated by an A-rich sequence. The 5′ region has boxes A and B containing RNA polymerase III promoter sequences. The 3′ region contains a 31-nucleotide insertion. There is a short polyadenylation tail located at the 3′ end of the Alu element (terminal A-stretch). Full-length Alu elements are approximately 300 bp long. (© Richard C. Li.)

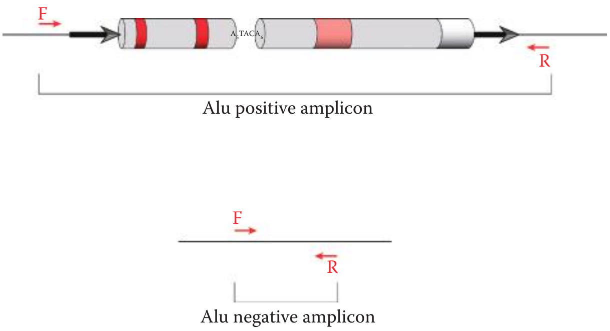

Figure 3.26 Detecting Alu element insertion polymorphism for human identification. The Alu element insertion site is amplified using the polymerase chain reaction assay. The presence and absence of the Alu element at the site can be detected based on the length of the DNA fragment analyzed. F, forward primer; R, reverse primer. (© Richard C. Li.)

polymorphic for the presence or absence of insertion that can potentially be used for forensic human identification (Figure 3.26).

3.4.3.3 Human DNA Polymorphic Markers

Most individual human genome sequences are very similar. However, variations in sequences do occur. The differences between individual genomes that occur at the DNA level are called *DNA polymorphisms*. In particular, a DNA polymorphism with alternative forms of a chromosomal locus that differ in nucleotide sequence is known as a *sequence polymorphism*. A DNA polymorphism that differs in the numbers of tandem repeat units is known as a *length polymorphism*. A DNA polymorphism can occur anywhere in a genome including genes and other chromosomal locations. Many DNA polymorphisms are useful for genetic mapping studies and hence are called *DNA markers*. DNA polymorphisms form the basis of forensic DNA profiling. The focus of this text is on the human genome, but polymorphisms also occur in the genomes of other organisms.

Most DNA polymorphisms are single nucleotide polymorphisms (SNPs) involving a single-base-pair change or a point mutation. Over one million SNPs have been identified. SNPs arise by spontaneous mutation. Most SNPs occur in noncoding regions of the genome, although some appear in coding regions as well. Other forms of DNA polymorphisms are tandem repeats such as STRs and VNTRs. Although their biological functions are unknown, STRs and VNTRs are very useful for forensic DNA analysis. Many are highly polymorphic, and the number of repeat units varies greatly among different individuals of a population. It is unlikely that two unrelated individuals will have exactly the same combination of STR or VNTR polymorphisms if sufficient markers are examined. Thus, a resulting genetic profile can be used for human identification.

Alternative forms of DNA polymorphisms are called *alleles*. The same allele present in both homologous chromosomes is referred to as *homozygous*. Two different alleles present in homologous chromosomes are referred to as *heterozygous*. A combination of alleles at a given locus is a *genotype*. In forensic analysis, the genotype for a panel of analyzed loci is called the *DNA profile*.

3.4.3.4 Forensic DNA Polymorphism Profiling

In 1984, Sir Alec Jeffreys (Figure 3.27) developed a DNA profiling technique using a VNTR technique involving multilocus profiling and later followed by single-locus profiling (Chapter 19). This technique led to the solving of a double murder that had been committed in Leicestershire (United Kingdom) in the 1980s. The case was the first to apply DNA evidence to a criminal investigation. During the investigation, DNA profiling not only identified the true perpetrator but it also excluded an innocent suspect. This case demonstrated DNA profiling's great potential in forensic investigations.

DNA profiling offered a number of advantages compared with earlier systems. The most important is the ability of the technique to reveal far greater individual variability in DNA than can be revealed by antigen and protein polymorphic markers. The probability of two unrelated individuals having the same DNA profile is very low. The great variability of DNA polymorphisms has made it possible to offer strong support for concluding that DNA from a suspect and from a crime scene originated from the same source. This technique was subsequently implemented in forensic laboratories worldwide.

In the mid-1980s, Kary Mullis and his coworkers developed the polymerase chain reaction (PCR) technique, which amplifies a small quantity of DNA. Mullis's invention had a powerful impact on molecular biology and earned him a Nobel Prize in 1993. Since the introduction of PCR, new techniques have been developed for forensic DNA testing purposes.

Figure 3.27 Sir Alec Jeffreys. (© Richard C. Li.)

The application of PCR-based assays makes forensic DNA analysis possible when only minute quantities of DNA can be recovered from a crime scene, for example, from hairs and cigarette butts. These assays have greatly increased the sensitivity of forensic DNA testing. The first forensic application of a PCR-based assay utilizing SNPs at the *HLA-DQA1* locus (formerly called DQα) was developed in 1986 (Chapter 22). One major disadvantage of the assay was the high probability (approximately 1 in 4000) of a match between two unrelated persons. Amplified fragment length polymorphism (AFLP) at the D1S80 locus has also been implemented in forensic laboratories. The D1S80 locus is a small-size VNTR marker that can be amplified by PCR. The *HLA-DQA1* and AFLP assays were used for some years until the introduction of STR assays.

In the late 1990s, forensic laboratories started utilizing STR loci. STRs have a number of advantages compared with VNTRs. For example, STRs can be amplified by PCR because of their smaller size, which greatly increases the sensitivity of the assay. Furthermore, STR markers are as highly variable as VNTRs. With the application of multiple STR loci, the probability of a match between two unrelated persons becomes extremely low. As a result of DNA testing, perpetrators have been identified and wrongly convicted innocent people have been exonerated (Chapter 20).

In 1995, the United Kingdom established the first national DNA database for criminal investigations. By the end of 1998, several other countries, including the United States, had created their own national DNA databases. The United States has selected 13 STR loci for the Combined DNA Index System (CODIS). These national DNA databases play important roles in solving criminal cases.

Another technique known as mitochondrial DNA (mtDNA) profiling has also been used for forensic testing. mtDNA is maternally inherited genetic material and is therefore particularly useful for human identification. Each cell contains multiple copies of mtDNA. Thus, mtDNA typing can be useful for analysis when nuclear DNA is severely degraded or present in very limited amounts, such as in cases involving decomposed human remains. Alternatively, polymorphic markers at the Y chromosome have also been utilized for forensic DNA testing. Y chromosomal markers are paternally inherited so they can be used for paternity testing. These markers are also very useful in analyzing DNA from multiple contributors in sexual assault cases.

Bibliography

Aleksander, A., Forensic expertise and judicial practice: Evidence or proof? *J Eval Clin Pract*, 2012, **18**(6): 1147–1150.

Alvarez-Cubero, M.J., et al., Genetic identification of missing persons: DNA analysis of human remains and compromised samples. *Pathobiology*, 2012, **79**(5): 228–238.

Andelinovic, S., et al., Twelve-year experience in identification of skeletal remains from mass graves. *Croat Med J*, 2005, **46**(4): 530–539.

Apostolov, A., S. Hristov, and E. Angelova, DNA identification of biological traces and interpretation in a sexual assault case. *Am J Forensic Med Pathol*, 2009, **30**(1): 57–60.

Aquila, I., et al., The role of forensic botany in crime scene investigation: Case report and review of literature. *J Forensic Sci*, 2014, **59**(3): 820–824.

Asari, M., et al., Multiplex PCR-based Alu insertion polymorphisms genotyping for identifying individuals of Japanese ethnicity. *Genomics*, 2012, **99**(4): 227–232.

Auvdel, M.J., Comparison of laser and ultraviolet techniques used in the detection of body secretions. *J Forensic Sci*, 1987, **32**(2): 326–345.

Baraybar, J.P., When DNA is not available, can we still identify people? Recommendations for best practice. *J Forensic Sci*, 2008, **53**(3): 533–540.

Bauer, C.M., et al., Molecular genetic investigations on Austria's patron saint Leopold III. *Forensic Sci Int Genet*, 2013, **7**(2): 313–315.

Bogdanowicz, W., et al., Genetic identification of putative remains of the famous astronomer Nicolaus Copernicus. *Proc Natl Acad Sci U S A*, 2009, **106**(30): 12279–12282.

4

Sources of Biological Evidence

Biological evidence analysis is one of the standard forensic examinations in the investigation of a wide variety of crimes. In particular, DNA evidence facilitates investigators' efforts to link offenders to crime scenes by matching DNA profiles. DNA evidence can also be used to eliminate suspects. The DNA evidence that is routinely encountered at crime scenes can often be categorized into several groups or types. Table 4.1 lists sources of DNA that are frequently found on personal items. Figures 4.1 and 4.2 illustrate representative types of evidence that are processed and their success rates.

4.1 Bodily Fluids

Bodily fluids and their stains are useful biological evidence for forensic, serological, and DNA analysis and may be useful in solving crimes. The most common bodily fluids in forensic analysis are blood, seminal fluid, and saliva. Blood evidence, such as peripheral blood that circulates through the heart, arteries, veins, and capillaries, is one of the most common types of biological evidence that is found at crime scenes (Chapter 12). The fluid portion of blood is called *plasma*, a subcompartment of extracellular fluid, which is the bodily fluid outside cells. Blood contains various suspended blood cells. The cellular portion of the blood consists of *erythrocytes* (also known as red blood cells), *leucocytes* (also known as white blood cells), and *platelets*. Because mature human erythrocytes and platelets do not have nuclei, they are not useful sources of nuclear DNA. For forensic DNA profiling, the nuclear DNA in blood samples (Figure 4.3) is primarily isolated from leucocytes, which are nucleated. Besides peripheral blood, menstrual blood can be analyzed to investigate the possibility of the occurrence of a sexual assault crime (Chapter 16).

Other bodily fluids are *transcellular fluids*. These fluids are considered to be a subcompartment of the extracellular fluid that is contained within epithelial-lined extracellular spaces. For example, seminal fluid (Chapter 14) and saliva (Chapter 15) stains as well as vaginal secretions are analyzed for the investigation of sexual assault crimes. Sometimes, urine stains and fecal materials (Chapter 16) are related to assault crimes as well. Sweat, which is secreted from the eccrine and apocrine sweat glands in the skin (Chapter 17), and cerumen, also known as earwax (a waxy substance secreted in the ear canal), can be potentially used for human identification. Fluids present in vomitus (Chapter 17) can be potentially important for forensic investigations of violent crimes.

4.1.1 Extracellular Nucleic Acids

Blood plasma and other various bodily fluids usually contain small amounts of nucleic acids (DNA or RNA) known as extracellular nucleic acids. The nucleic acids circulating in plasma

Table 4.1 Common Items of Evidence

Evidence	Possible DNA Location	Source of DNA
Baseball bat	Handle	Skin cells, sweat, blood, tissue
Hat, bandana, mask	Inside surfaces	Sweat, hair, skin cells, dandruff, saliva
Eyeglasses	Nose, ear piece, lens	Sweat, skin cells
Facial tissue, cotton swab	Surface	Mucus, blood, sweat, semen, earwax
Dirty laundry	Surface	Blood, sweat, semen, saliva
Toothpick	Surface	Saliva
Used cigarette	Butt (filter area)	Saliva
Used stamp, envelope seal	Moistened area	Saliva
Tape or ligature	Inside or outside surface	Skin cells, sweat, saliva
Bottle, can, glass	Mouthpiece, rim, outer surface	Saliva, sweat, skin cells
Used condom	Inside surface, outside surface	Semen, vaginal cells, rectal cells
Bed linen	Surface	Sweat, hair, semen, saliva, blood
Through-and-through bullet	Outside surface	Blood, tissue
Bite mark	Skin surface	Saliva
Fingernail, partial fingernail	Scrapings	Blood, sweat, tissue, skin cells

Source: National Institute of Justice. Using DNA to solve cold cases, Special Report, 2002, US Department of Justice, Office of Justice Programs, Washington, DC.

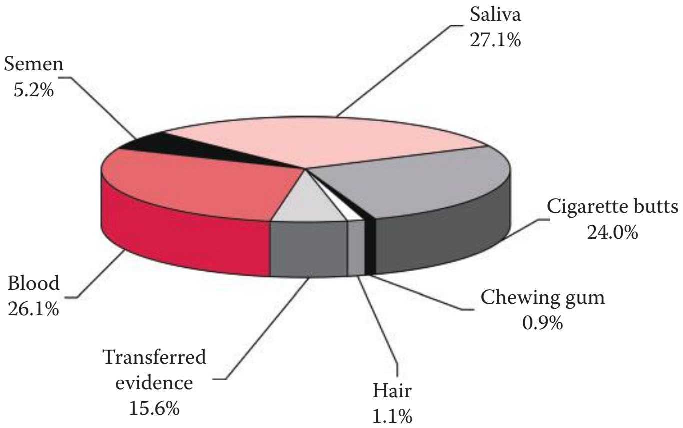

Figure 4.1 Representative types of evidence samples. Data compiled from the third quarter 2005 (July–September) results for all police forces in England and Wales. (Adapted from Bond, J.W., *J Forensic Sci*, 52, 128–136, 2007.)

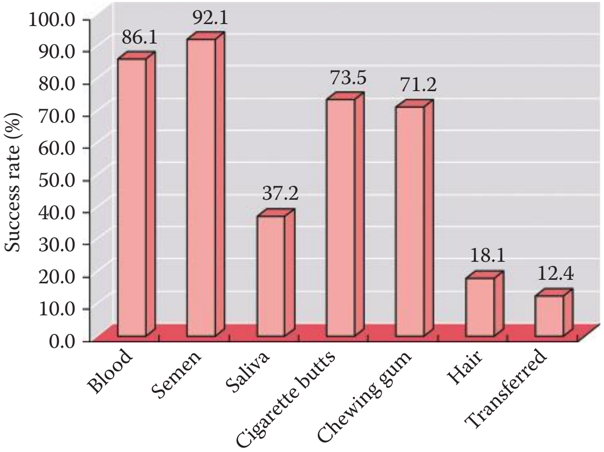

Figure 4.2 Success rate of obtaining suitable profiles from processed samples for submitting to DNA databases. Data are from the third-quarter 2005 (July–September) results for all police forces in England and Wales. (Adapted from Bond, J.W., *J Forensic Sci*, 52, 128–136, 2007.)

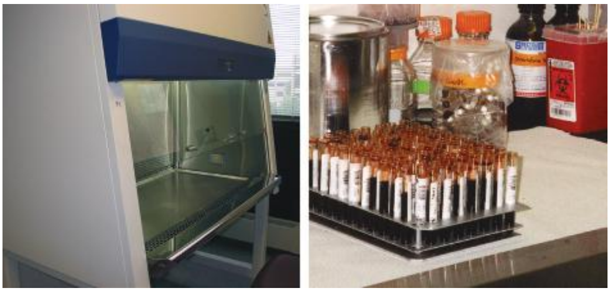

Figure 4.3 Biosafety cabinet for the extraction of DNA from biological evidence (left). Blood samples to be processed in a biosafety cabinet for DNA isolation (right). (© Richard C. Li.)

are referred to as *circulating* nucleic acids; the nucleic acids that are found in other body fluids, such as saliva and urine, are referred to as *cell-free* nucleic acids. Extracellular nucleic acids can remain soluble or form complexes with proteins and lipids.

The potential sources of extracellular nucleic acids are *extracellular vesicles* (EVs). EVs are endogenous vesicular structures, containing proteins and nucleic acids that are secreted by most eukaryotic cells (Figure 4.4). There are many types of EVs, including exosomes and microvesicles, which can be detected in bodily fluids. *Exosomes* are one potential source of extracellular nucleic acids. Exosomes are derived from multivesicular bodies (MVBs), which are intracellular organelles of the endocytic pathway (Figure 4.4a). MVBs fuse with the plasma membrane. As a result, the vesicles are released into the extracellular compartment as exosomes. *Microvesicles*, also called shed vesicles or ectosomes, are another possibility. Microvesicles shed from the plasma membrane and thus carry along membrane and cytosolic materials including nucleic acids (Figure 4.4b). *Apoptotic bodies* are a special type of microvesicle that is formed in apoptotic cells. During apoptosis, a process of programmed cell death, cells shrink and eventually form apoptotic bodies (Figure 4.5). Apoptotic bodies contain fragmented DNA by nucleolytic degradation that resembles similar characteristics to those observed in extracellular nucleic acids. However, messenger RNA (mRNA) within apoptotic bodies is protected from RNase degradation. These

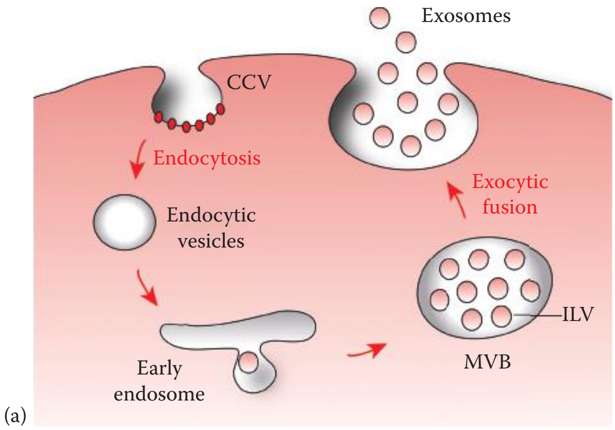

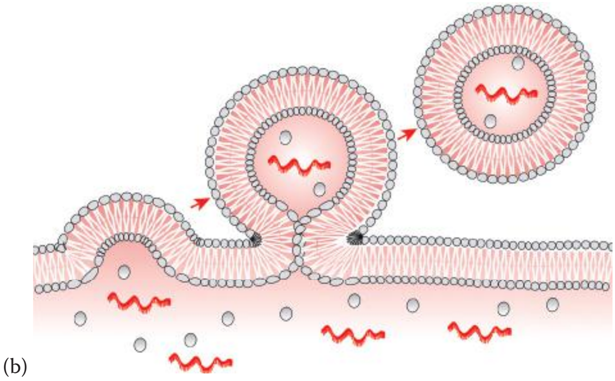

Figure 4.4 Exosomes and microvesicles released from healthy cells. (a) Exosomes are derived from the endocytic pathway. In the endocytic pathway, clathrin-coated vesicles (CCV) are formed at the plasma membrane. The endocytic vesicles are then formed through endocytosis and are transported to early endosomes. Multivesicular bodies (MVBs) containing intraluminal vesicles (ILVs) are then developed from early endosomes. Through the exocytic fusion of the MVB membrane with the plasma membrane of the cell, exosomes, containing cellular components including nucleic acids, are released to the extracellular space. (b) Microvesicles are derived from the budding of vesicles from the plasma membrane. Microvesicles are then shed from the plasma membrane carrying along cellular components including nucleic acids that are present in the cytoplasm. (© Richard C. Li.)

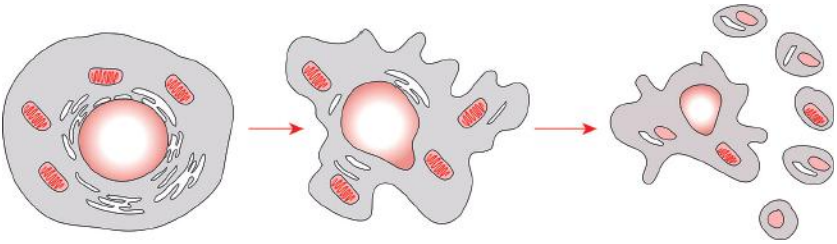

Figure 4.5 Apoptotic bodies released during apoptosis. During apoptosis, the preapoptotic cell (left) undergoes morphological changes. The early apoptotic cell (middle) shows deformation: the cell shrinks, the membrane blebs, and chromatin condense. The late apoptotic cells (right) are fragmented, releasing apoptotic bodies, which contain cytosol and organelles, and nuclear fragments. In addition, proteolytic enzymes are activated that cleave the genomic DNA into fragments. (© Richard C. Li.)

extracellular nucleic acids are potentially useful sources for the forensic identification of bodily fluids and for forensic DNA profiling.

4.2 Cells

4.2.1 Cell Surface Markers

Many of the forensic markers that are analyzed in forensic laboratories are from cells, the building blocks of the human body. All cells have membranes, also known as plasma membranes, which constitute their outer boundaries (Figure 4.6). The functions of cell membranes include exchanges with the environment, signal transduction, and structural support. The cell membrane is a phospholipid bilayer containing lipids, proteins, and carbohydrates. Membrane proteins can act as enzymes, receptors, or ion channels. Many cells also have carbohydrate-rich molecules, including proteoglycans, glycoproteins, and glycolipids, on their membrane surfaces (Figure 4.6). Many of these molecules act as cell surface antigens (Chapter 18).

4.2.2 Nucleated Cells

Probative biological evidence usually contains nucleated cells. The *nucleus* (Figure 4.7) is surrounded by a nuclear envelope and contains chromosomes and a nucleolus. A nucleolus is a dense, non-membrane-bound structure due to its high RNA content. The function of a nucleolus is to transcribe ribosomal RNA and to form ribosomes.

There are two types of cells in the body: sex cells (sperm and oocytes) and somatic cells (all other types). Spermatozoa and ova, which are formed by germ cells, are called *gametes*. In humans, each gamete is haploid, containing 22 *autosomes* (chromosomes other than sex chromosomes) plus one *sex chromosome*. In ova, the sex chromosome is always an X, while in spermatozoa it may be an X or a Y. After fertilization, the zygote, which is the fertilized egg cell formed when two gamete cells are joined, becomes diploid as a result of the fusion of the haploid spermatozoon and ovum. Most of the other cells of the body, known as *somatic cells*, are diploid. This means that they have two copies of each autosome plus two sex chromosomes, XX for females or XY for males. This results in a total of 46 chromosomes per diploid cell. The two chromosomes of a pair in a diploid cell are *homologous chromosomes*. One of the homologous chromosomes is inherited from the spermatozoon and the other from the ovum.

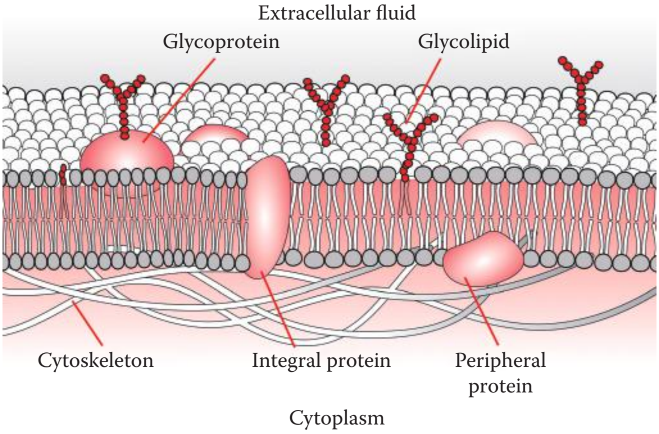

Figure 4.6 Cell membrane and carbohydrate-containing glycoproteins and glycolipids. (© Richard C. Li.)

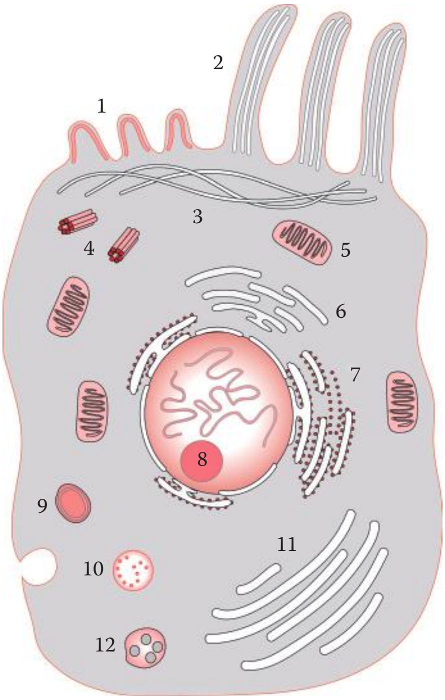

Figure 4.7 A cross-sectional view of a cell. 1, Microvilli; 2, cilia; 3, cytoskeleton; 4, centrioles; 5, mitochondrium; 6, smooth endoplasmic reticulum; 7, rough endoplasmic reticulum; 8, nucleolus; 9, peroxisome; 10, vesicle; 11, Golgi apparatus; and 12, lysosome. (© Richard C. Li.)

Although most somatic cells are diploid, exceptions exist. Some differentiated cells such as red blood cells and platelets have no nuclei and are designated *nulliploid*. A few other cells have more than two sets of chromosomes as a result of DNA replication without cell division and are referred to as *polyploid*. For example, the regenerating cells of the liver and other tissues are naturally tetraploid, while the giant megakaryocytes of the bone marrow may contain 8, 16, or even 32 copies of chromosomes.

The nuclear *chromosomes* of humans consist of complexes of DNA, histone proteins, and nonhistone chromosomal proteins. Each chromosome consists of one linear, double-stranded DNA molecule. The large amounts of DNA present in the human chromosome are compacted by their association with histones into nucleosomes and even further compacted by higher levels of folding of the nucleosomes into *chromatin* fibers. Each chromosome contains a large number of looped domains of chromatin fibers attached to a protein scaffold.

The degree of DNA packing varies throughout the cell cycle. During the metaphase of mitosis and meiosis of the cell cycle, chromatin is the most condensed. Two forms of chromatin have been defined on the basis of their chromosome-staining properties. *Euchromatin* regions are areas of chromosomes that undergo normal chromosome condensation and decondensation during the cell cycle. The intensity of staining of euchromatin is darkest in the metaphase and lightest in the synthesis (S) phase. Euchromatic regions account for most of the genome and lack repetitive DNA. Usually, the genes within the euchromatin can be expressed. *Heterochromatin* comprises the chromosomal regions that usually remain condensed throughout the cell cycle. It contains repetitive DNA and can be found at centromeres, much of the Y chromosome long arm, and the short arms of the acrocentric chromosomes (chromosomes with centromeres near one end). Genes within heterochromatic DNA are usually inactive.

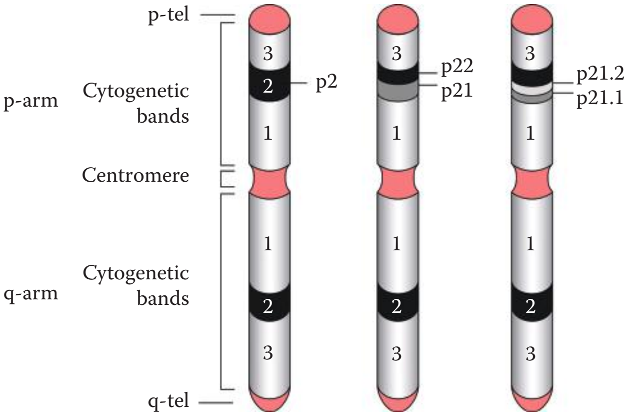

Figure 4.8 The ideogram of a human chromosome and cytogenetic banding nomenclature. The cytogenetic banding pattern is shown after chemical staining such as G-banding, which treats chromosomes with Giemsa dye. The short arm is designated p; the long arm is q. The centromere, p telomere (p-tel), and q telomere (q-tel) are also shown. G-bands at increasing resolution (from left to right) are shown. (© Richard C. Li.)

Each human chromosome has a short arm, designated p (for petit), and a long arm, designated q (for queue), separated by a centromere (Figure 4.8). *Centromeres* are the DNA sequences that are found near the points of attachment of mitotic or meiotic spindle fibers. The centromere region of each chromosome is responsible for accurately segregating the replicated chromosomes to daughter cells during cell divisions. The ends of the chromosome are called *telomeres* and they help stabilize the chromosome and play a role in the replication of DNA in the chromosome.

Chemical staining of metaphase chromosomes results in an alternating dark and light banding pattern (cytogenetic banding) that can be seen under a microscope. Each chromosome arm is divided into regions based on the cytogenic bands. This process is known as *cytogenetic mapping*. The cytogenetic bands are labeled p1, p2, p3, q1, q2, q3, and so on, counting from the centromere out toward the telomeres. At higher resolutions, subbands can be observed. For example, the cytogenetic map location of a gene termed *AMELY* (amelogenin, Y-linked; Chapter 21) is Yp11.2, which indicates its location on chromosome Y, p arm, band 11, subband 2. The visually distinct banding pattern gives each chromosome a unique appearance. Recently, the cytogenic map has been integrated with the human genome sequence to allow the determination of the positions of cytogenetic bands within the DNA sequence.

Chromosomes can be identified on the basis of the size and the positions of the centromeres and cytogenetic banding patterns. The chromosome constitution is described as a *karyotype* and can be displayed as a *karyogram*, which includes the total number of chromosomes and the sex chromosome composition (Figure 4.9). Chromosomes are numbered in order of their size, with chromosome 1 as the largest (except chromosome 21 is smaller than 22). In the cases of chromosomal abnormality, the karyotype can also reflect the type of abnormality and allow visualization of the affected chromosome bands.

4.2.3 Mitochondria and Other Organelles

The cytoplasm contains the cytosol fluid in which organelles are suspended (Figure 4.7). Multiple copies of *mitochondria* are located within the cytoplasm. Mitochondria are surrounded by phospholipid membranes that separate them from the cytosol. The mitochondria are responsible for energy production through aerobic metabolism by producing molecules containing high-energy bonds, such as adenosine triphosphate (ATP). Mitochondria have their own genome, which can be analyzed for human identification (Chapter 23).

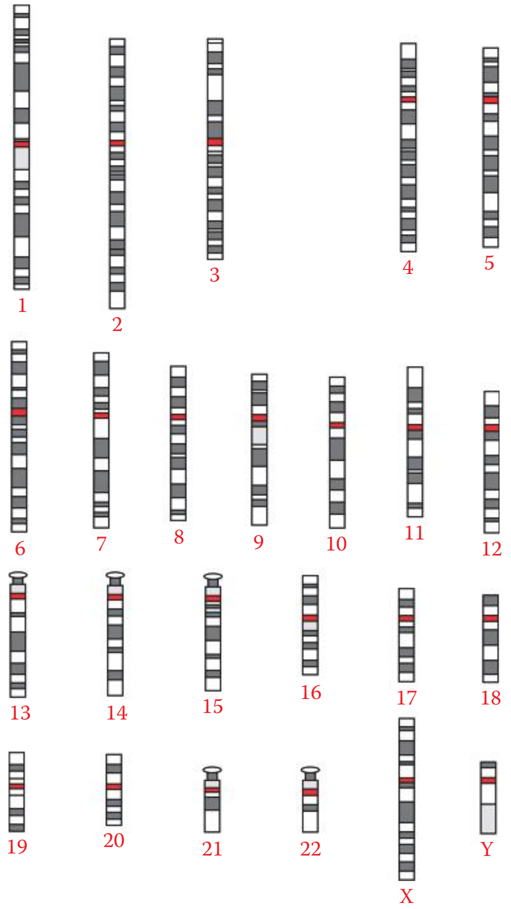

Figure 4.9 Human karyogram. The chromosomes are numbered. Cytogenic patterns show alternating dark and light bands (From www.genome.gov). Centromeres (in red) are shown.

Two types of *endoplasmic reticulum* (ER) can exist within the cytoplasm: the smooth ER (SER) is involved in lipid synthesis; the rough ER (RER) contains ribosomes on its outer surface and forms transport vesicles. The *Golgi apparatus* is responsible for the production of secretory vesicles and new membrane components, and also for the packaging of *lysosomes* (vesicles containing digestive enzymes for the degradation of injured cells). *Peroxisomes* carry enzymes that neutralize potentially harmful free radicals. Other organelles found within the cytoplasm of eukaryotic cells include the cytoskeleton, microvilli, centrioles, and cilia.

4.2.4 Cytosol

4.2.4.1 Messenger RNAs

The chromosomal DNA contains genes that encode for specific proteins. The genetic code is read as an array of triplet codes, a sequence of three bases that specifies the identity of a single amino acid. As gene expression is activated, transcription occurs in which precursor mRNA (pre-mRNA) is produced from a DNA template. After transcription, the pre-mRNA is capped, polyadenylated, and spliced to form matured mRNA. Only the matured mRNA is transported from the nucleus to the cytoplasm. Tissue-specific mRNA can be potentially used for the identification of biological evidence (Chapter 11). The proteins are synthesized in a process known as translation, in which amino acids are assembled based on the codons derived from the triplet code of the DNA contained in the sequence of the mRNA strand. Various components including the ribosomal complex are involved in translation. The cytosol also contains many proteins that can be used for the identification of bodily fluids (Chapters 11 and 12).

4.2.4.2 MicroRNAs

MicroRNAs (miRNAs) are short RNA molecules that are 21–23 nucleotides in length. In eukaryotic organisms, miRNAs function as negative regulators of gene expression. They play roles in development and cell differentiation. Additionally, the altered expression of miRNAs can be detected in many human diseases. The biological function and the potential application of miRNAs in forensic biology are discussed in Chapter 11.

It is estimated that the human genome may encode approximately 1000 miRNAs. Based on their genomic location and structure, miRNAs can be characterized into three types: intergenic, intronic, and exonic (Figure 4.10). *Intergenic* miRNA genes are distinct transcriptional units that are found in genomic regions. In humans, most of the intergenic miRNA genes have a transcription start site and a polyadenylation site. *Intronic* miRNAs reside within the introns of protein-coding and noncoding genes. The orientation of the intronic miRNAs can be the same as that of the sense (coding) strand of a host gene or that of the antisense strand (complementary to the sense strand) of a host gene. Sense intronic miRNAs are transcribed from the same promoter as their host genes. The antisense intronic miRNAs are transcribed from their own promoters. *Exonic* miRNAs are rare in eukaryotic genomes and reside in genomic regions that overlap with an exon and an intron of a pseudogene, which is a noncoding gene or is no longer transcribed. These miRNAs are also transcribed from their host gene promoter. miRNAs can also be monocistronic or polycistronic. A *monocistronic* miRNA has a single transcriptional unit with its own promoter. In *polycistronic* miRNAs, several miRNAs reside as a cluster of transcriptional units with a shared promoter.

The biogenesis of miRNAs begins in the nucleus. miRNAs are transcribed by RNA polymerase II. Nascent transcripts, referred to as primary transcripts (pri-miRNAs), can be several hundreds to thousands of nucleotides in length. The pri-miRNAs have a hairpin secondary

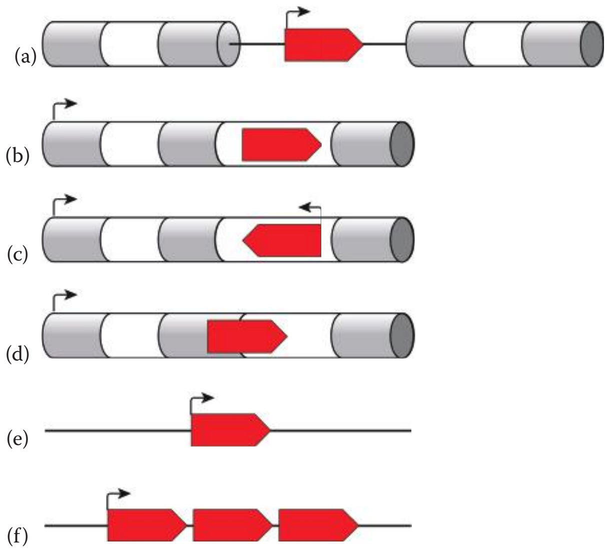

Figure 4.10 The genomic location and the structure of miRNAs. (a) An intergenic miRNA gene, (b) a sense intronic miRNA gene, (c) an antisense intronic miRNA gene, (d) an exonic miRNA gene, (e) a monocistronic miRNA gene, and (f) polycistronic miRNA genes. Exons (gray), introns (white), and miRNA genes (red) are shown. (© Richard C. Li.)

structure that is approximately 70 nucleotides long with imperfect base pairing in the stem. miRNA processing is initiated in the nucleus (Figure 4.11). The hairpin region of a pri-miRNA is then cleaved from the pri-miRNA by the nuclear RNase III endonuclease, Drosha. As a result, a pre-miRNA is formed. The pre-miRNA is transported from the nucleus to the cytoplasm by a nuclear transporter, Exportin 5. In the cytoplasm, a cytoplasmic RNAse III–like endonuclease, Dicer, cleaves the pre-miRNA to generate a double-stranded miRNA that is approximately 21–23 nucleotides in length. The mature miRNA strand is bound to the Argonaut protein to assemble the RNA-induced silencing complex (RISC). The complementary RNA strand is degraded.

A specific miRNA is designated with a prefix. The prefix for a mature miRNA is a capitalized "miR," while the prefix for a pre-miRNA is an uncapitalized "mir." The prefix "miR" is followed by a dash and a number (e.g., miR-135). Experimentally confirmed miRNAs are sequentially numbered and are deposited in the miRBase, which is a database that archives miRNA sequences and annotations.

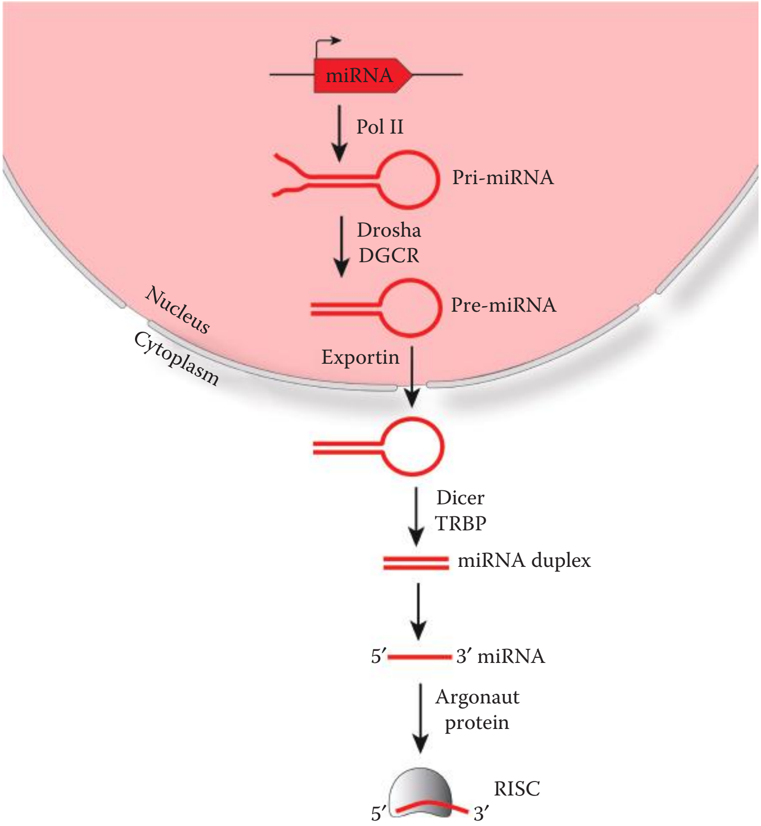

Figure 4.11 The biogenesis of miRNAs. In the nucleus, the pri-miRNA (primary transcript) is processed by Drosha (nuclear RNase for double-stranded RNA) and DGCR (nuclear double-stranded RNA-binding protein), forming the pre-miRNA. The pre-miRNA is transported from the nucleus to the cytoplasm by the exportin. In the cytoplasm, the pre-miRNA is further processed by Dicer (cytoplasmic RNase for double-stranded RNA) and TRBP (cytoplasmic double-stranded RNA-binding protein). The mature miRNA strand is bound to the Argonaut protein to assemble the RISC (RNA-induced silencing complex). (© Richard C. Li.)

4.3 Tissues

4.3.1 Skin

4.3.1.1 Biology of Skin

The biological evidence related to skin is important in forensic investigations. For example, fingerprints are ridge skin impressions that are usually collected at crime scenes, and shed skin tissue is a source of DNA for human identification. The skin covers the entire body surface. Additionally, the skin contains specialized structures that include sebaceous and sweat glands, hair follicles, and nails. The thickness of skin varies throughout the body. The skin of the dorsal area of the body is usually thicker than that of the ventral area of the body. The skin consists of different layers (Figure 4.12). The *epidermis* is the outer layer of the skin. The epidermis also contains melanocytes that produce the skin pigment melanin. The *dermis* is the middle layer of the skin. It is filled with fibrous collagen proteins secreted by fibroblasts and contains hair follicles and sweat glands. Additionally, it contains blood, lymph vessels, and nerves. The *subcutaneous layer* is the deepest layer of the skin. The subcutaneous layer consists of collagen networks and adipose tissue to prevent loss of heat.

The epidermis is a multilayered tissue that includes a number of morphologically distinct zones: the basal, the spinous, the granular, and the cornified layers (Figure 4.12). The epidermis renews continually through the proliferation and differentiation of keratinocytes. The *basal layer* contains newly formed keratinocytes that are proliferative. Epidermal differentiation begins with the migration of the keratinocytes from the basal layer toward the outer layer of skin. Once the migrating keratinocytes reach the *spinous* and *granular* layers, the keratinocytes become nonproliferating and partially differentiated. As the cells reach the *cornified* layer, these cells are filled with keratin filaments and are differentiated into corneocytes, which are dead, and terminally differentiated keratinocytes. In the course of differentiation, the cells are flattened, and all organelles including the nucleus are lost. The corneocytes are then shed from the skin surface.

4.3.1.2 Skin as Source of DNA Evidence

Evidence from skin contact, also referred to as *touched evidence*, can be collected and used for forensic DNA analysis. One example of this evidence is shed skin cells that are found on worn clothing, which is frequently encountered in crime scene investigations. For instance, a perpetrator's shed skin cells that are deposited on worn clothing are potential evidence to be collected. After collection, a DNA profile can be obtained from shed skin cells, providing the forensic evidence or "lead" that is required for the criminal investigation. Touched evidence becomes more

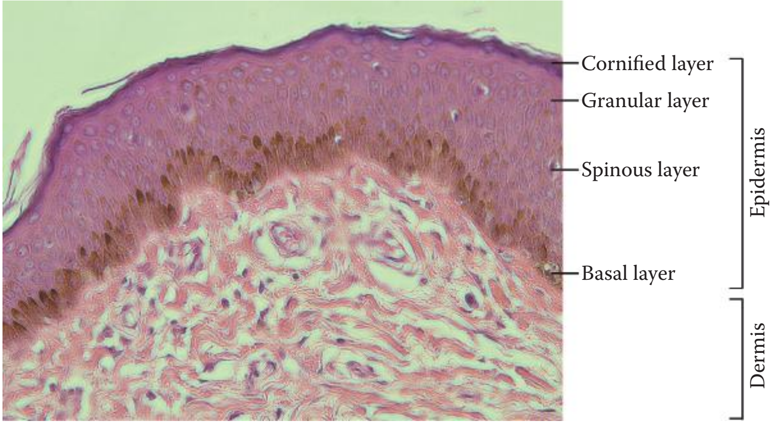

Figure 4.12 A sectional view of the skin. Multilayered epidermis tissues are shown. (© Richard C. Li.)

useful and important when no other type of evidence (e.g., a fingerprint, bloodstain, or seminal stain) can be collected at the crime scene, especially for the investigation of certain criminal cases such as cold cases and property crimes such as theft and burglary (Figures 4.13 through 4.15). Touched evidence usually contains minute quantities of nuclear DNA. DNA recovered from touched evidence is referred to as *transfer DNA*. However, the source of DNA transferred through physical contact is not well understood. Although the shed skin cells from touched evidence lack nuclei, these cells may contain DNA remnants from partial degradation during the differentiation process. These can be a possible DNA source. Additionally, the sweat glands in the skin produce sweat. When the skin makes contact with an item, a residue of sweat is left on the surface of the item. It is known that sweat contains cell-free DNA. Thus, cell-free DNA from sweat is another possible source of DNA. Furthermore, a small number of nucleated cells can be observed in touched evidence. These cells possibly originated from sweat glands and ducts.

4.3.2 Hair

Hairs, including scalp and pubic hairs, frequently constitute biological evidence that is found at crime scenes, and their identification can be of great forensic importance. Formerly, the

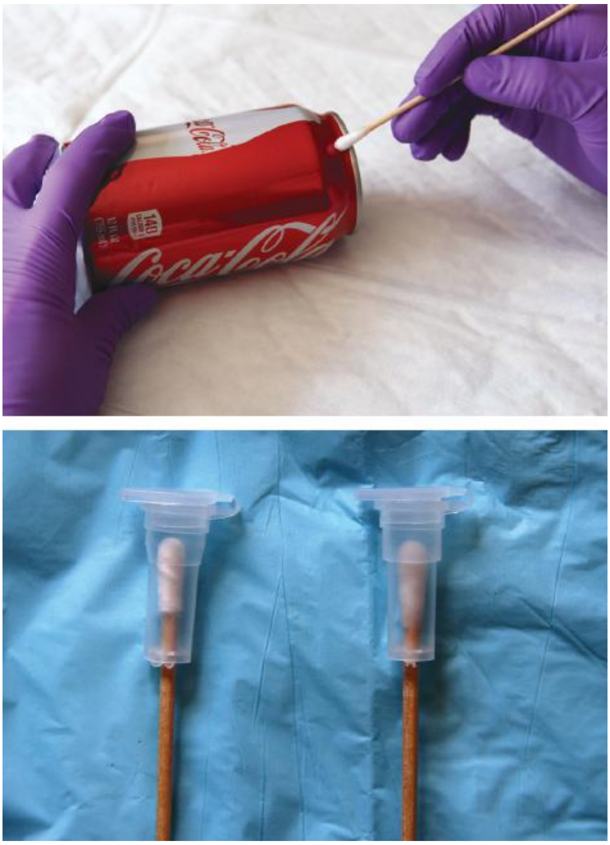

Figure 4.13 Evidence collection using swabbing. The evidence from skin contact for forensic DNA analysis is usually collected by swabbing. More DNA can be recovered when evidence is collected with a double swab method than with a single swab. The double swab method involves applying a moistened cotton swab followed by a second dry cotton swab onto the same target surface of evidence. The target surface is swabbed using a moistened swab first. The moisture left by the first swab is absorbed by the second dry swab. Both swabs can be pooled for DNA extraction. (© Richard C. Li.)

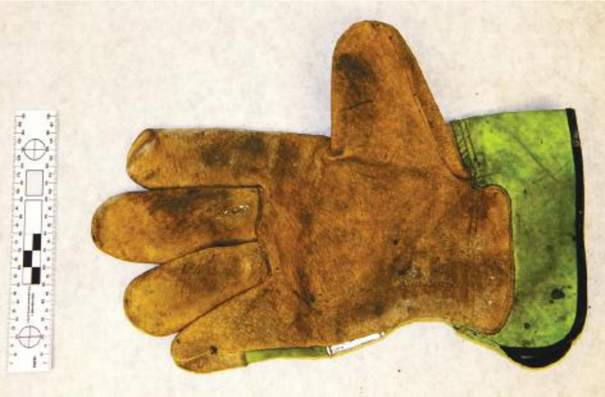

Figure 4.14 A worn glove as a source of DNA evidence. Shed skin cells on a worn glove can potentially generate DNA profiles. (© Richard C. Li.)

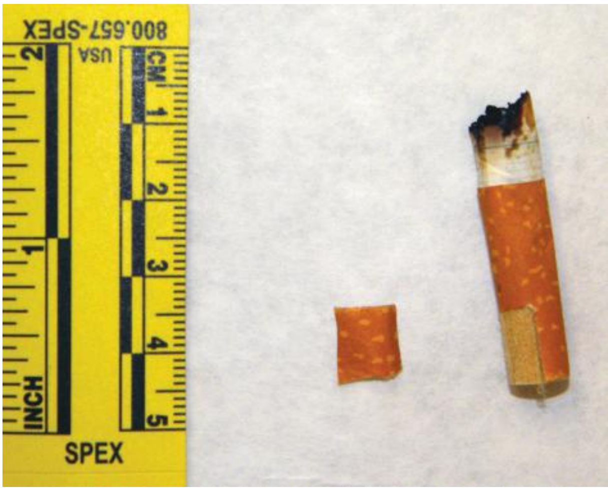

Figure 4.15 A cigarette butt as a source of DNA evidence. Shed cells left behind on a cigarette butt can be a source of DNA. A portion (~1 cm^2) of the filter paper of a smoked cigarette butt can be cut for isolating DNA. (© Richard C. Li.)

principal methods that were utilized in forensic hair analysis were limited to morphological analysis and comparisons. Since then, protein polymorphisms provide some potential for identifying individuals from single hairs. However, human hairs contain DNA and as a DNA source they may be used for forensic analysis. The development of the polymerase chain reaction (PCR) amplification technique made it possible to analyze very small quantities of DNA in hair, and the use of hair as evidence of identification has become more significant.

4.3.2.1 Biology of Hair

The human hair shaft is a keratinized cylindrical structure (Figure 4.16). The center or core of the hair is called the *medulla*, which is present in the majority of hairs. The medulla is surrounded by a *cortex*, which is the outer layer of the hair shaft. The *cuticle* consists of overlapping layers of flattened keratinized cells that protect the hair. Hairs are produced in *hair follicles* (Figures 4.17 and 4.18). Each hair follicle is located deep in the dermis (a skin layer beneath the epidermis) and opens onto the surface of the epidermis (the outer layer of the skin). The hair follicle is composed

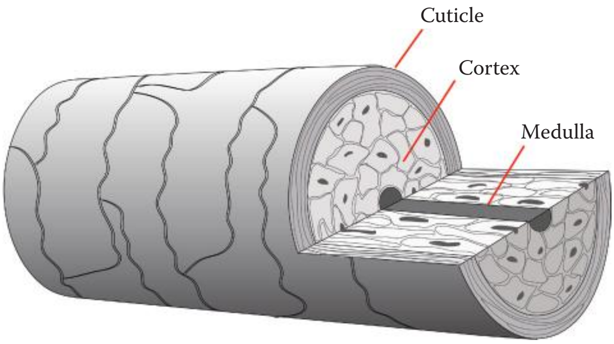

Figure 4.16 A sectional view of a hair shaft. (© Richard C. Li.)

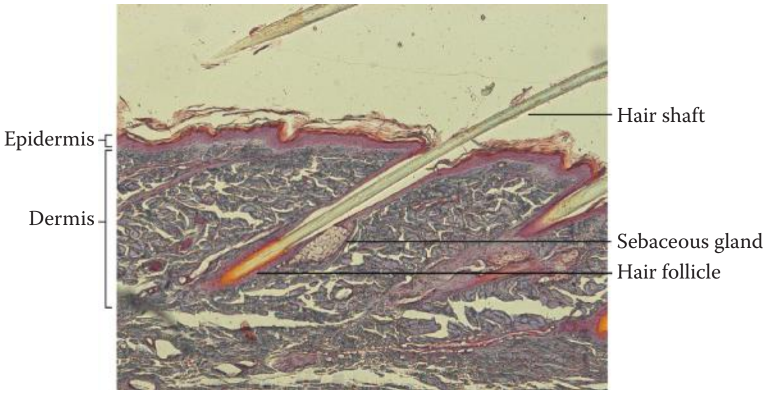

Figure 4.17 Longitudinal section view of a scalp hair follicle with accessory structures. (© Richard C. Li.)

of the *bulge* and the *bulb* regions. The bulge exports the stem cells that migrate down and give rise to bulb cells. The bulge also produces the stem cells that migrate up to form skin cells. The matrix cells, which generate the hair shaft cells, are located at the lower portion of the bulb. The *dermal papilla* is situated at the base of the bulb and contains cells that regulate hair growth, which are nourished by blood vessels and nerves. The growing hair shaft is surrounded by two concentric layers of cells, which are referred to as the inner root sheath and the outer root sheath, respectively. The entire hair follicle is surrounded by a connective tissue sheath.

Human scalp hairs grow for a few years and shed according to the hair follicle cycle (Figures 4.19 through 4.21). A human scalp hair can grow at its highest rate of approximately 1 mm per day. The growing phase of hair is called the *anagen* phase. The matrix cells undergo rapid proliferation and eventually become differentiated cells such as hair shaft cells. As a hair grows, it is pushed toward the surface of the skin and becomes longer. During the migration upward, keratinization occurs as cells are filled with fibril keratin proteins. In these keratinizing cells, nuclei are absent. However, mitochondrial remnants can be observed. By the time a hair approaches the skin surface, cell death occurs at the medulla, cortex, and cuticle. At the end of the anagen phase, the matrix cells enter the *catagen* phase and undergo cell death, thus leading to the regression of the bulb. Hair follicles then enter the *telogen* phase: the stage of rest. When another cycle begins, the follicle produces a new bulb and the telogenic hair, also known as the club hair, is pushed to the surface and shed. On average, an adult loses approximately dozens of hairs daily.

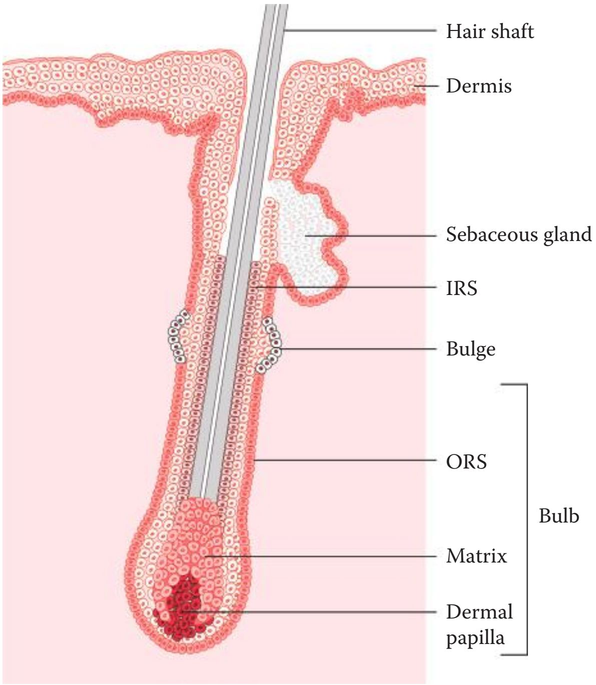

Figure 4.18 Structure of an anagen-phase hair follicle. The diagram shows the hair follicle structure including the dermal papilla, sebaceous gland, bulge, bulb, and hair shaft. Concentric layers of the outer root sheath (ORS), the inner root sheath (IRS), and the hair shaft are also shown. (© Richard C. Li)

4.3.2.2 Hair as Source of DNA Evidence

The isolation of DNA from intact hair roots is routinely used in nuclear DNA analysis. In comparison, the quantities of DNA in telogen hair roots are considerably less than the DNA that is found in the roots of anagen hairs. Nuclear DNA analysis is usually accomplished by using freshly plucked hair roots (Figures 4.22 and 4.23) because cells at the root region may contain nuclear DNA. Unfortunately, most human hairs recovered from crime scenes are shed naturally (in the telogen phase) and contain little nuclear DNA (Figure 4.24). Thus, multiple telogen hairs with roots are necessary to isolate enough nuclear DNA. However, shed hairs that are found at crime scenes may be derived from different individuals. Therefore, the ability to perform a forensic DNA analysis of a single shed hair would be highly desirable. Nuclear DNA isolation from hair shafts is still far less reliable because hair shafts contain very low amounts of nuclear DNA. In addition, variations in the amounts of DNA isolated from hair shafts are observed in a comparison between different hairs from the same head and hairs from different individuals.

A hair follicle cell contains multiple copies of mitochondria. As a result, mitochondrial DNA (mtDNA) can be successfully isolated from hair roots. Additionally, mtDNA is embedded in the keratin matrix of hair shaft cells, which protects the mtDNA molecules from degradation. Thus, mtDNA can also be isolated from hair shafts. A sequence polymorphism analysis of mtDNA from hair can be carried out. mtDNA is maternally inherited, which is useful to identify maternal relatives but cannot be used to perform paternity testing. Additionally, the mtDNA profiling results are not as discriminating as nuclear DNA profiling. Furthermore, mtDNA analysis is time-consuming. Therefore, the typing of nuclear DNA from hair would be preferable for forensic DNA analysis.

Sometimes, a mixture of more than one mtDNA sequence in the same individual is observed. This heterogeneous pool of mtDNA molecules is referred to as *heteroplasmy*. In hair, it is

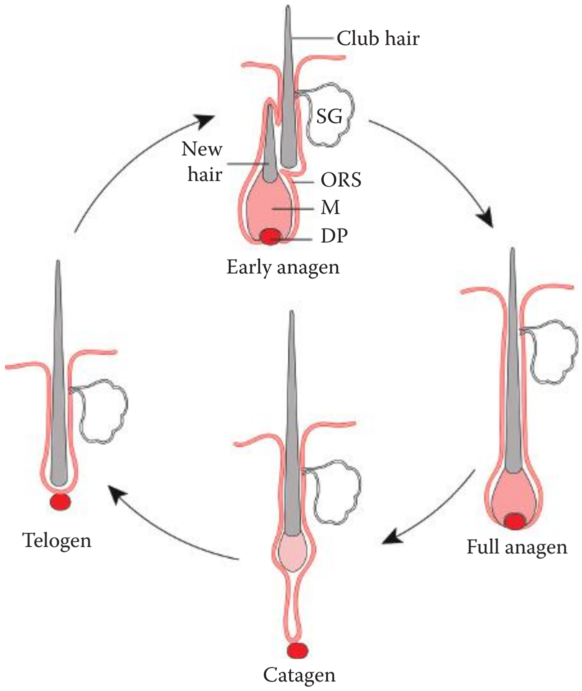

Figure 4.19 Diagrammatic representation of the scalp hair cycle. The morphological characteristics of a hair follicle are shown in three distinct phases encompassing the entire hair cycle: anagen (including early and full anagen), catagen, and telogen. DP, dermal papilla; M, matrix; ORS, outer root sheath; SG, sebaceous gland. (© Richard C. Li.)

believed that it is due to a mixture of mtDNA molecules from keratinocyte and melanocyte-derived mitochondria (Figure 4.25). Hair follicle melanocytes are formed at the beginning of each hair cycle and die at the end of the cycle. Melanocytes are located in the bulb hair follicle. Hair melanocytes play roles in hair pigmentation, which determine hair color. Melanin, which is produced by the melanocytes, is contained in an organelle called the *melanosome*. The melanosomes are transferred to neighboring keratinocytes though *dendritic processes*. In addition to melanosomes, the melanocyte mitochondria can also be transferred to keratinocytes. Thus, the keratinized cells in the hair shaft may carry more than one type of mitochondria, one from the keratinocytes and the other from the melanocytes. As a result, heteroplasmy of a mixture of different mtDNA molecules, with different DNA sequences, can occur in hairs.

4.3.3 Bone

4.3.3.1 Biology of Bone

The bodies of human remains begin to decompose shortly after death. The rate of decomposition of human remains varies greatly with environmental conditions such as climate, bacterial growth, and the presence of insects and other animal scavengers. However, soft tissues may be lost first while more stable bone tissues may remain. Identifying human skeletal remains can be applied in a variety of cases including mass fatality incidents, missing persons, fires, explosions, and violent crime cases involving skeletal remains.

An adult human skeleton consists of 206 bones (Figures 4.26 and 4.27). The shaft of a long bone, such as an arm or a leg bone, consists largely of an outer layer of *cortical* (or compact) bone, which is solid and strong. The shaft of a long bone forms a *marrow cavity*, which is filled with a specialized type of connective tissue called bone marrow. The portion at each end of a

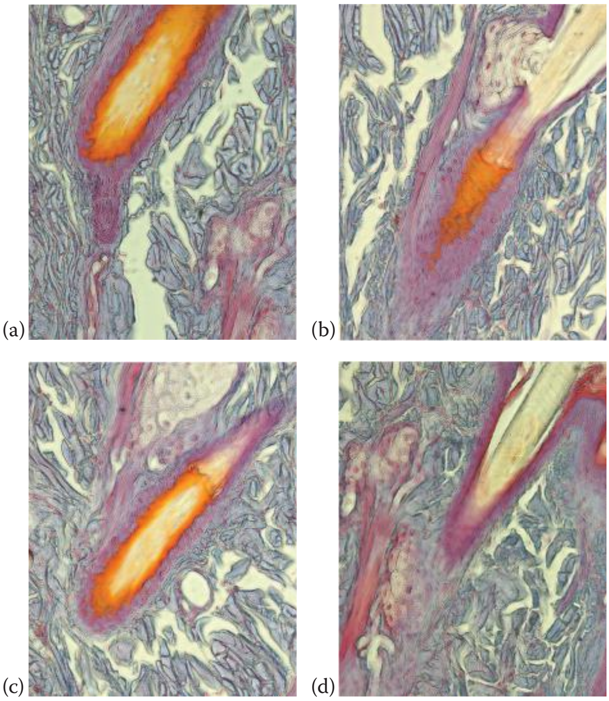

Figure 4.20 Longitudinal section view of a scalp hair follicle during the hair cycle. (a) Early anagen, (b) full anagen, (c) catagen, and (d) telogen. (© Richard C. Li.)

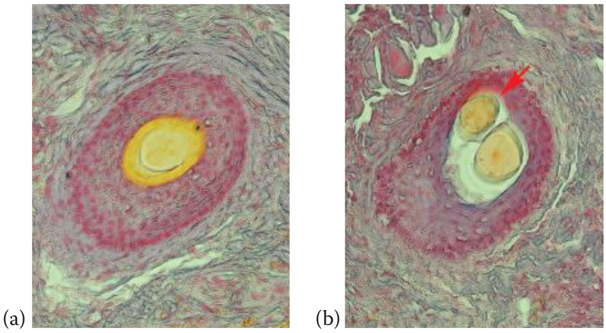

Figure 4.21 Cross-sectional view of hair follicles. (a) Early anagen and (b) full anagen. Note that the new hair (arrow) and the club hair are shown. (© Richard C. Li.)

long bone is called the *epiphysis*, which is composed largely of *cancellous* (or spongy) bone, and can bear the force of compression. A flat bone can have primarily either cortical or cancellous bone. For instance, a rib consists of primarily cancellous bone surrounded by a thin layer of cortical bone. A skull bone usually consists of largely cortical bone.

Bone, which is a connective tissue, contains a matrix and cells. The bone matrix consists of an inorganic and an organic matrix. Calcium and phosphorus are the major components of the inorganic matrix, which consists mainly of hydroxyapatite crystals, $Ca_{10}(PO_4)_6(OH)_2$. The organic matrix consists of collagens, primarily type I collagen, which are insoluble fibrous proteins. With the deposition of calcium hydroxyapatite crystals around the collagen fibrils, bone becomes a weight-bearing hard tissue.

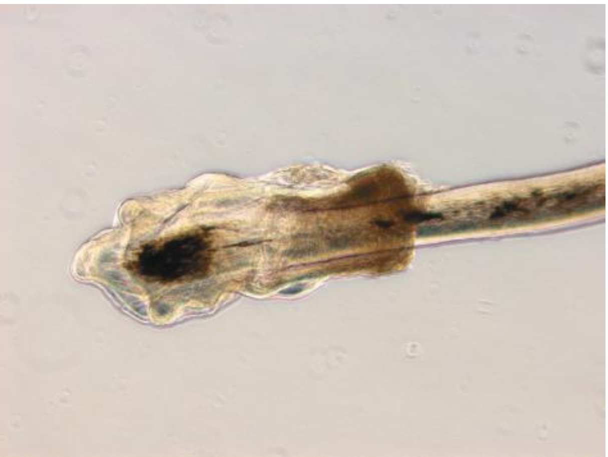

Figure 4.22 Hair root of a pulled hair with visible soft tissue attached. (© Richard C. Li.)

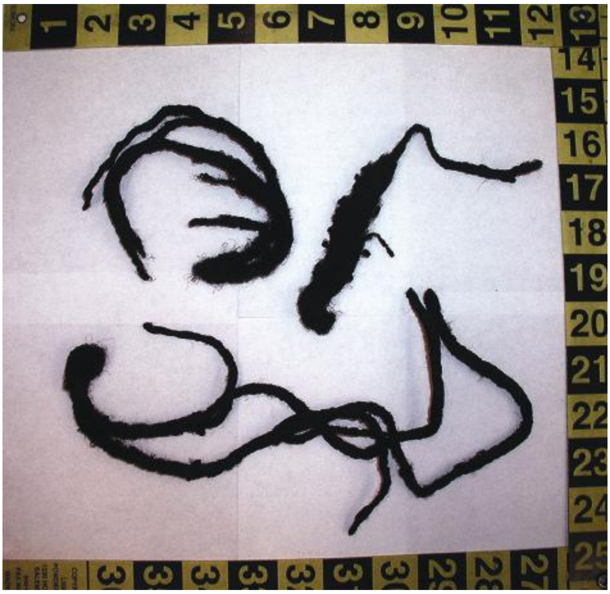

Figure 4.23 Pulled dreadlocks recovered from a crime scene. (© Richard C. Li.)

Developing bones contain small numbers of *osteoprogenitor* cells. These cells can divide to produce cells that differentiate into *osteoblasts*. Osteoblast cells regulate the calcification of the bone matrix. Osteoblasts that are embedded in the bone matrix are termed *osteocytes* and are the most abundant cells in bone. Osteocytes play a role in maintaining the surrounding matrix and repairing damaged bone. Another type of cell that can be found in bone tissues is *osteoclasts*. These cells are giant cells containing 50 or more nuclei and are responsible for dissolving and recycling the bone matrix.

4.3.3.2 Bone as Source of DNA Evidence

A number of methods are used to identify human remains, for example: the identification of facial characteristics; the recognition of individualizing scars, marks, and other special body

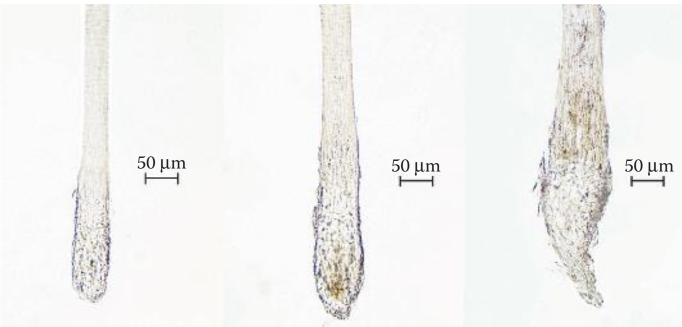

Figure 4.24 Telogen hair roots. From left to right: telogen club root with no visible soft tissue; telogen hair root with some visible soft tissue; telogen hair root with visible soft tissue. (From Bourguignon, L. et al., *Forensic Sci Int Genet*, 3, 27–31, 2008. With permission.)

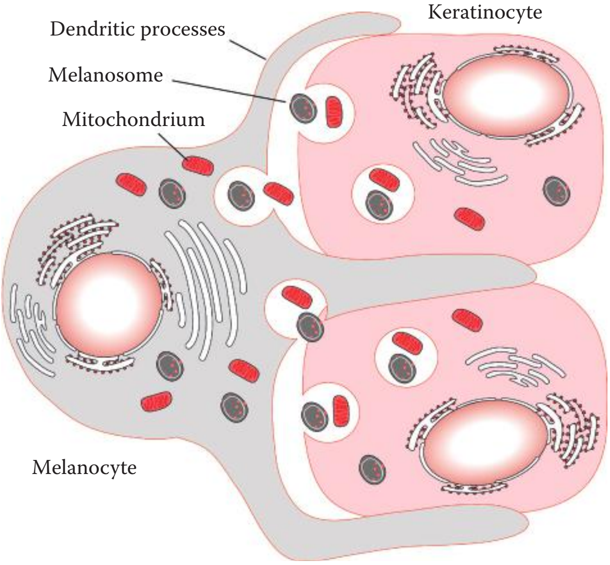

Figure 4.25 Scheme diagram of the melanosome transportation. Melanosomes are released into the extracellular space from the melanocyte dendrites through exocytosis and subsequent endocytosis by keratinocytes. During the process, mitochondria originating from the melanocytes can be potentially transported to the keratinocytes. (© Richard C. Li.)

features; the matching of dentition with premortem dental x-rays; and the comparison of fingerprints. In some situations, these methods cannot be used because of the extensive decomposition of the remains. The mass fatality terrorist attack on the World Trade Center in 2001 (Figure 4.28) serves as an example of a situation where common identification techniques may not be useful. Large quantities of compromised human skeletal fragments were recovered at the fatality site. In these cases, DNA typing is a powerful tool for identifying human remains.

Most DNA in cortical bone is located in the osteocytes. It has been estimated that there are approximately 20,000 osteocytes per cubic millimeter of calcified bone matrix. As a result, microgram quantities of DNA can potentially be obtained from a gram of bone. Thus, compact bone tissue should contain sufficient amounts of nuclear DNA for analysis. However, the skeletal fragments recovered from burial sites are often subjected to decomposition (Figures 4.29 and 4.30). During the decomposition process, both nuclear and mtDNA can be degraded.

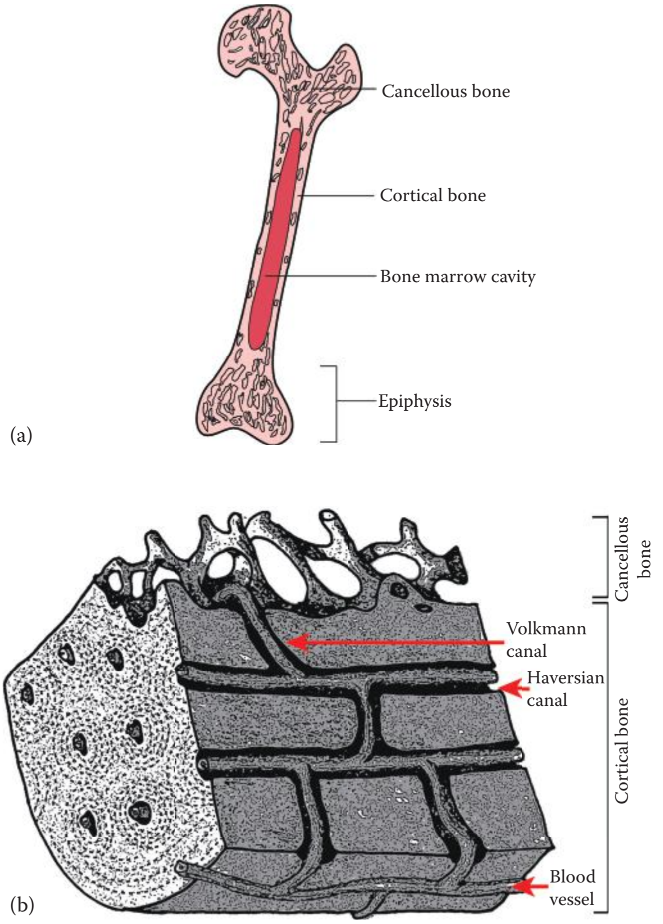

Figure 4.26 Bone structure. An adult human consists of 206 bones. A long bone, such as an arm or a leg bone, consists of an outer cylinder of cortical bone surrounding a marrow cavity. Each end of a long bone is called the epiphysis, which is composed largely of cancellous bone. Flat bones have variable structures; for example, the skull consists mainly of cortical bone whereas the spine consists mainly of cancellous bone. (a) Diagram of a long bone. (b) A bone fragment. Cortical and cancellous bones are shown. Blood vessels can be found in the Volkmann and Haversian canals. (© Richard C. Li.)

Additionally, burial conditions with high humidity and temperature promote the degradation of DNA. Thus, the identification of partial DNA profiles or the complete failure to obtain DNA profiles can occur after samples from decomposed remains are analyzed.

Processing bone samples for DNA extraction is a time-consuming task (Figure 4.31). Due to the potential for commingled remains and contaminants that interfere with forensic analysis, a bone sample initially must be cleaned prior to isolating DNA. The outer surfaces of bone fragments are usually removed by using a mechanical method such as sanding. However, to avoid cross-contamination of samples, the bone dust that is generated by sanding must be removed. Additionally, special protective equipment and safety procedures are necessary to protect analysts from exposure to blood-borne pathogens.

To obtain adequate quality and quantity of DNA from a bone sample, a high-yield DNA extraction method should be selected. The bone samples can be ground to powder to aid in DNA extraction (Chapter 5). The osteocytes containing DNA are embedded in a calcified bone matrix, which is a barrier for extracting DNA from the osteocytes. The bone matrix must be

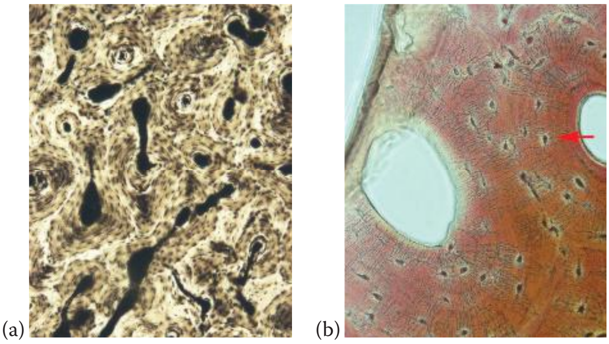

Figure 4.27 Cross-sectional view of cortical bone. (a) The functional unit of cortical bones is a cylindrical structure known as osteon. Haversian canals are shown in the center of the osteon. (b) Detailed view of an osteon. Osteocytes (arrows) are shown within osteons. (© Richard C. Li.)

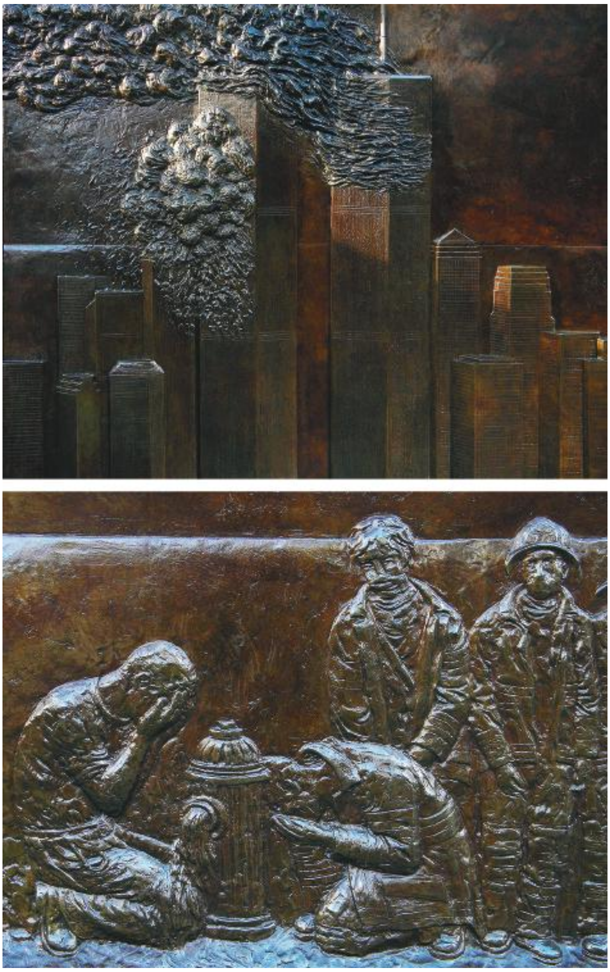

Figure 4.28 Sections of the Fire Department of New York City (FDNY) Memorial Wall. Memorial to the Fallen Firefighters of 9/11, at FDNY Engine Co. 10 on Liberty Street, New York, by Rambusch Studios. (© Richard C. Li.)

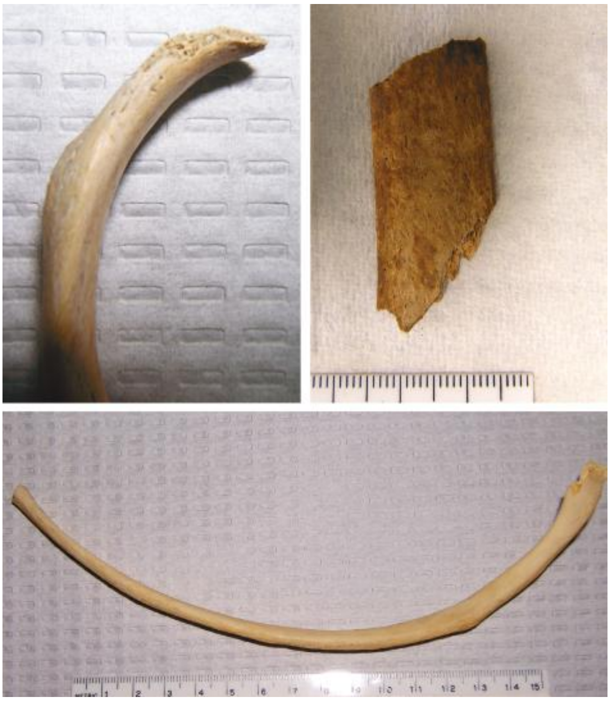

Figure 4.29 Human rib bone fragments recovered from burial site. (© Richard C. Li.)

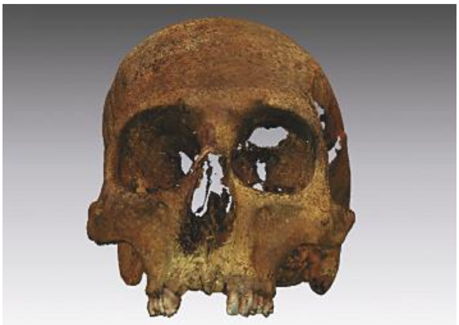

Figure 4.30 Skeletal remains, exhumed from Frombork Cathedral in Poland, are thought to be those of astronomer Nicolaus Copernicus (1473–1543). (From Bogdanowicz, W., et al., *Proc Natl Acad Sci U S A*, 106, 12279–12282, 2009. With permission.)

removed to improve the yield of DNA. A decalcification method can be utilized to dissolve calcium ions to soften the bone tissue. Additionally, the application of proteinase can be used to digest the matrix proteins, thus increasing the yield of DNA that is harvested from osteocytes.

4.3.4 Teeth

4.3.4.1 Biology of Teeth

During embryonic development, two sets of teeth begin to form. The first to appear are the *deciduous teeth* or *primary teeth*. Most children have 20 deciduous teeth, which are later replaced by 32 teeth known as the *secondary dentition* or *permanent dentition* (Figure 4.32).

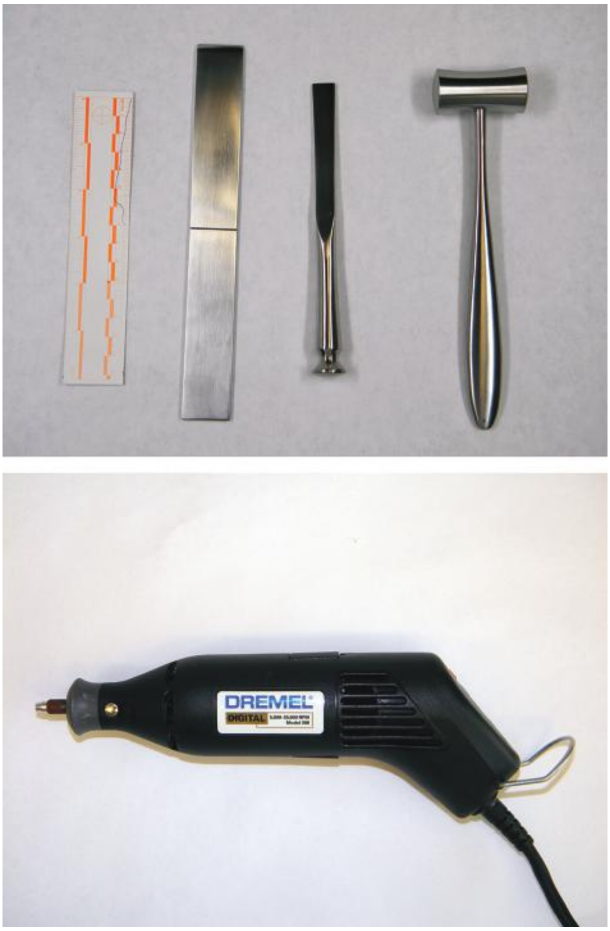

Figure 4.31 Tools for cutting bone samples. Osteotomes and a mallet (top), and a rotary device (bottom). (© Richard C. Li.)

The bulk of each tooth consists of a calcified connective tissue called *dentin* (Figure 4.33). The dentin of the crown is covered by a layer of *enamel*. The rest of the dentin is referred to as radicular dentin and is covered by a layer of *cementum*, which separates the tooth from the surrounding jawbone.

Similar to bone, tooth tissue contains a matrix. The inorganic matrix of tooth contains hydroxyapatite, a calcium phosphate in a crystalline form. The organic matrices of dentin and cementum are primarily collagens. In enamel, amelogenin is the major protein of the organic matrix. Other proteins include ameloblastin and enamelin, which are also the components of the enamel organic matrix.

The interior chamber within the tooth surrounded by dentin is known as the pulp cavity. The *dental pulp*, found within the pulp cavity, is the connective tissue made up of nerve fibers, blood vessels, and various cells. The blood vessels and nerves in the pulp cavity are innervated through the root canal, a narrow tunnel located at the root of the tooth. Incisor and cuspid teeth have single roots. Bicuspids have one or two roots. Molars typically have three or more roots.

The columnar cell bodies of *odontoblasts* are located along the peripheral dental pulp. A single *odontoblast process*, arising from each cell body of the odontoblasts, projects into the dentinal tubule (Figure 4.34). Odontoblasts play important roles in the formation of dentin. Odontoblasts secrete collagens and ground substances that are the components of the

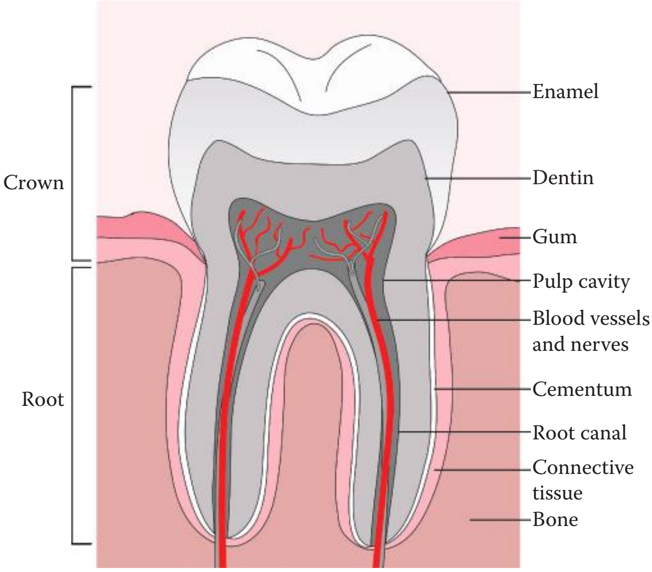
Crown
Root
Enamel
Dentin
Gum
Pulp cavity
Blood vessels and nerves
Cementum
Root canal
Connective tissue
Bone

Figure 4.32 A sectional view of an adult tooth. (© Richard C. Li.)

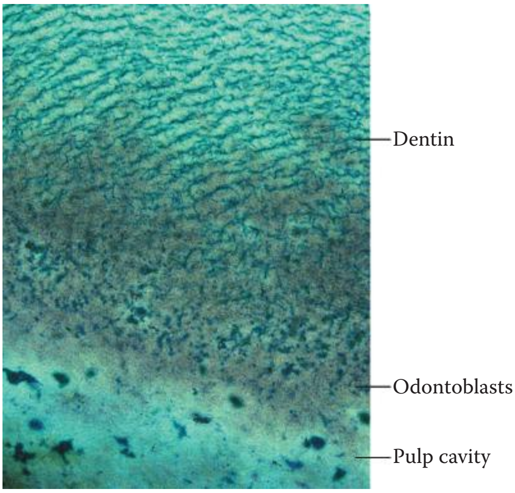
Dentin
Odontoblasts
Pulp cavity

Figure 4.33 Cross-sectional view of a tooth. (© Richard C. Li.)

dentinal matrix. Additionally, odontoblasts regulate the calcification of the matrix of dentin. *Cementoblasts* are cells that play roles in forming the cementum. Cementoblasts secrete collagens and ground substances to form the extracellular matrix of the cementum. Through the process of forming cementum, cementoblasts become trapped in the extracellular matrix. The cementoblasts embedded in the cementum are referred to as *cementocytes*. *Ameloblasts* are cells that play a role in producing enamel and are subsequently lost during tooth eruption.

4.3.4.2 Teeth as Source of DNA Evidence

The characteristics of teeth, their alignment, and the overall structure of the mouth provide information for identifying a person. The use of dental records such as x-rays and dental casts can allow dental remains to be connected to a victim. Particularly in circumstances such as

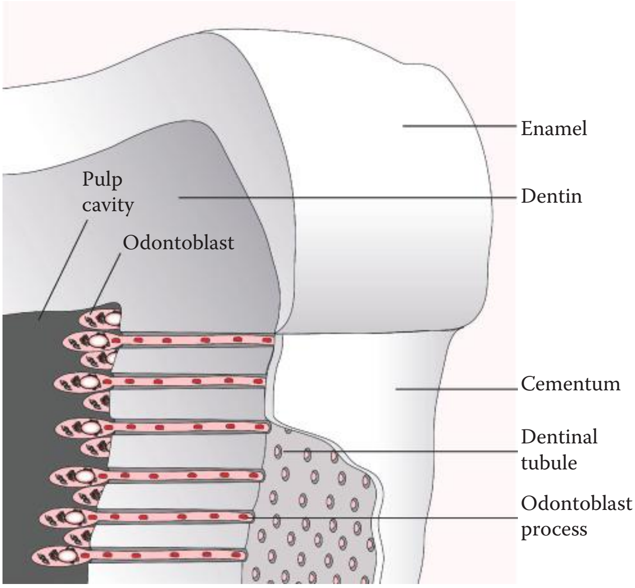

Figure 4.34 Odontoblast processes. See Section 4.3.4.1. (© Richard C. Li.)

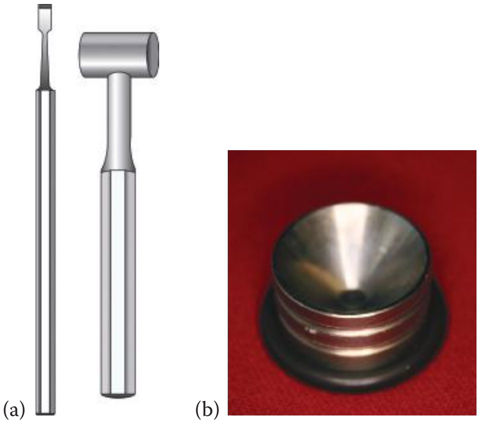

Figure 4.35 Tools for dissecting teeth. (© Richard C. Li.) (a) A dental chisel and a mallet. (© Richard C. Li.) (b) An amalgam well. (Courtesy of Dr. Ken Hermsen.)

decomposition, an odontological comparison is possible since the dental evidence often remains intact. When no antemortem dental record is available for comparison, forensic DNA testing can be carried out for postmortem human identification. The mineralized dental structure protects DNA from degradation in cases where it may be degraded in other tissues of the body. Thus, teeth are an excellent source of DNA for forensic DNA analysis under such conditions.

Dental pulp tissue contains various cells and is a suitable source of DNA. However, when tooth evidence has been exposed to high temperature and humid environments, decomposition of the pulp tissue can occur. Thus, cementoblasts within the cementum, containing both nuclei and mitochondria, can then be utilized as a source of DNA. Additionally, odontoblast processes within dentin, containing mitochondria, can also be used. Several different methods are used to obtain dental tissues for DNA isolation (Figure 4.35). A vertical section is cut along the longitudinal axis of the tooth, which allows the dissection of the pulp, dentin, and cementum tissues

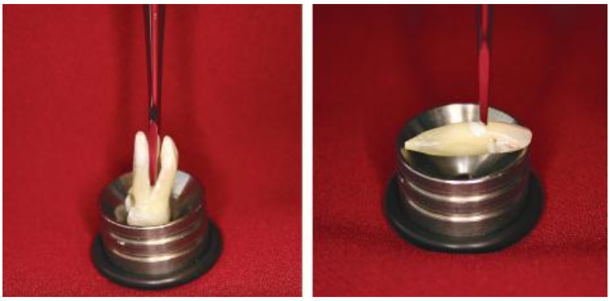

Figure 4.36 Dissecting teeth for DNA isolation. Vertical cutting and dissecting of the pulp, dentin, and cementum tissues (left) and horizontal cutting and dissecting of the root portion (right). (Courtesy of Dr. Ken Hermsen.)

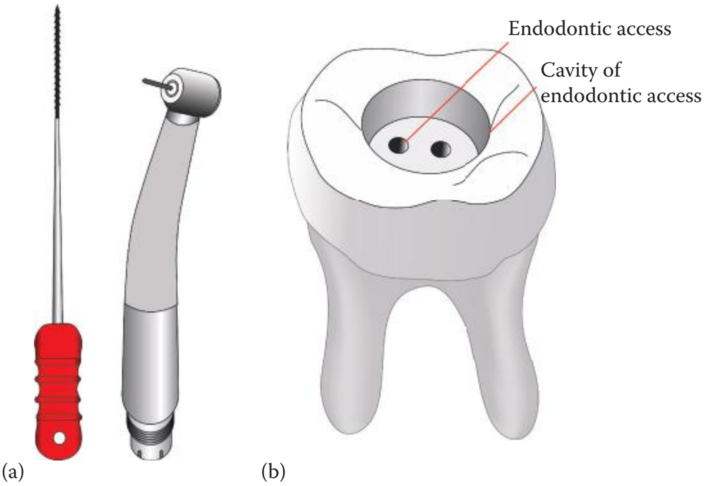

Figure 4.37 Extraction of pulp tissue using endodontic access procedures: trepanning the occlusal surface of a tooth using a dental bur mounted on a turbine (a), creating a cavity for endodontic access (b), and extracting pulp tissue using a nerve broach (a). (© Richard C. Li.)

(Figure 4.36). A horizontal section is cut through the cementum–enamel junction of the tooth (Figure 4.36). The root portion can be pulverized to a fine powder for DNA isolation. The crown can be preserved for forensic odontological comparisons if needed. Additionally, the extraction of pulp tissue can be carried out by standard endodontic access (Figure 4.37). For calcified tissues such as dentin and cementum, a decalcification step is needed to soften the dental matrix, which facilitates DNA isolation.

Bibliography

Abaz, J., et al., Comparison of the variables affecting the recovery of DNA from common drinking containers. *Forensic Sci Int*, 2002, **126**(3): 233–240.

Aditya, S., C.N. Bhattacharyya, and K. Chaudhuri, Generating STR profile from "Touch DNA". *J Forensic Leg Med*, 2011, **18**(7): 295–298.

Alakoc, Y.D. and P.S. Aka, "Orthograde entrance technique" to recover DNA from ancient teeth preserving the physical structure. *Forensic Sci Int*, 2009, **188**(1–3): 96–98.

5

Nucleic Acid Extraction

5.1 Basic Principles of DNA Extraction

Deoxyribonucleic acid (DNA) is a linear polynucleotide consisting of four types of monomeric nucleotides. Each nucleotide contains three components: a deoxyribose, a nitrogenous base, and a phosphate group (Figure 5.1). The four bases for DNA are adenine (A), cytosine (C), guanine (G), and thymine (T) (Figure 5.2). The deoxyribose is attached to the nitrogen of a base. The phosphate group is attached to the deoxyribose. In a *polynucleotide*, individual nucleotides are linked by phosphodiester bonds (Figure 5.3).

Interestingly, the first attempt to isolate DNA from humans, which involved the use of human leucocytes, was accomplished by the Swiss physician Friedrich Miescher in 1869. This study led to the discovery of the DNA molecule, which he referred to as "nuclein." Over the years, various methods for isolating DNA from blood samples were developed. However, many protocols still use the basic principles Miescher developed more than 100 years ago.

This chapter introduces basic techniques for DNA extraction used in forensic laboratories (Table 5.1). The commonly used methods isolate total cellular DNA, which is suitable for most forensic DNA analyses. However, DNA extraction procedures may vary according to the type of biological evidence, which can include cell types, substrates, the quantity of biological evidence collected, and the type of test that is being performed. Specific DNA extraction procedures for a sample of interest can be found in the literature. The extraction method of choice yields an optimal quantity, quality, and purity of DNA to satisfy forensic DNA testing needs. A sufficient quantity of DNA ensures the generation of a complete DNA profile. Poor quality of DNA, such as fragmentation due to DNA degradation, may result in a partial DNA profile or a failure to obtain a profile. Poor purity of DNA may cause interference during subsequent DNA testing. For example, DNA polymerase inhibitors interfere with DNA amplification (Chapter 7). Additional criteria for selecting proper DNA extraction methods include adaptability to automation, throughput potential, simplicity, the reduction of contamination risks, and cost-effectiveness. The most common DNA extraction protocols include basic components discussed below.

5.1.1 Cell and Tissue Disruption

In most DNA extraction protocols, enzymatic digestions, such as those with proteinase K, are used for cell and tissue disruption. The disruption process can also be carried out by boiling and by using alkali treatment and mechanical methods (Figures 5.4 and 5.5). Materials such as bone and teeth can be frozen in liquid nitrogen and then ground to a fine powder with a mortar or cryogenic grinder such as the SPEX CertiPrep® freezer mill (Figure 5.6). In such specimens, cells containing DNA

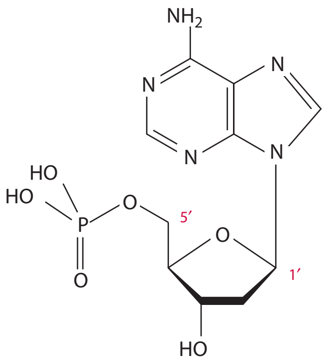
NH2
N
N
N
N
HO
HO
P
O
O
5′
O
1′
HO

Figure 5.1 Structure of a nucleotide. Each nucleotide in a DNA polymer is made up of three components: a deoxyribose, a nitrogenous base (adenine is shown as an example), and a phosphate group. The 1′ carbon of the deoxyribose is attached to the nitrogen of a nitrogenous base. A phosphate group is attached to the 5′ carbon of the deoxyribose.

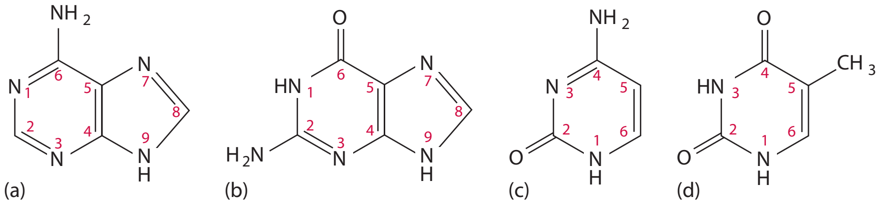
NH 2
N
N
N
NH
O
HN
H 2N
N
NH
NH 2
N
O
NH
O
HN
CH 3
O
NH
(a)
(b)
(c)
(d)

Figure 5.2 Chemical structures of nitrogenous bases: (a) adenine, (b) guanine, (c) cytosine, and (d) thymine.

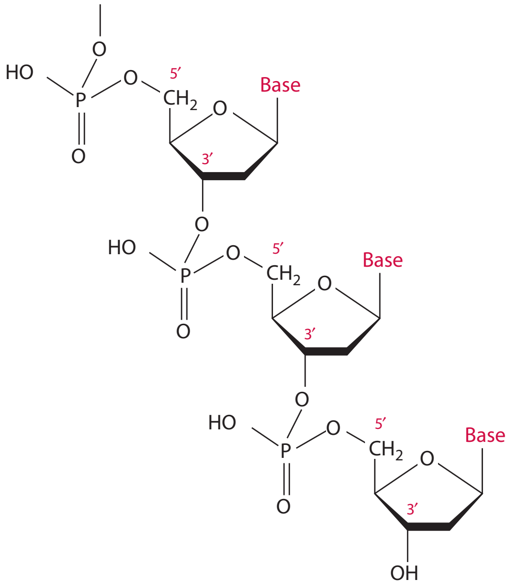
HO
P
O
O
O
5′
CH2
O
Base
3′
HO
5′
Base
3′
HO
5′
Base
3′
OH

Figure 5.3 DNA polynucleotide chain. Individual nucleotides are linked by phosphodiester bonds between their 5′ and 3′ carbons.

Table 5.1 Three Basic DNA Extraction Methods

Method	Forensic Application	Purity	DNA Strand	DNA Size	Throughput	Note
Solvent-based	RFLP- and PCR-based assay	High	Double stranded	Large	Time-consuming; difficult to adapt for automation	Uses toxic solvent; multiple transfer steps between tubes
Boiling	PCR-based assay	Low	Single stranded	Small	~30 min per sample	No transfer required
Silica-based	RFLP- and PCR-based assay	High	Double stranded	Large	~1 h per sample; amenable to automation	Minimum transfer required

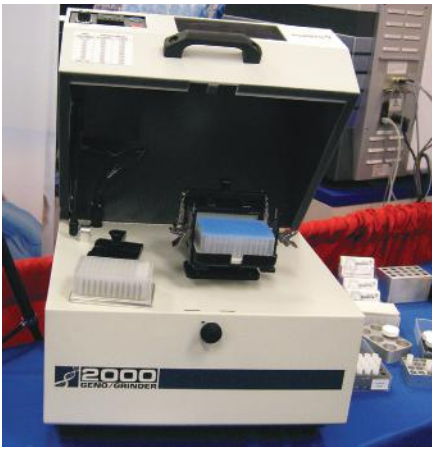

Figure 5.4 An automated mechanical disruption device for efficient disruption of tissues. It can be used for high-throughput applications involving sample preparation for the isolation of nucleic acids. (© Richard C. Li.)

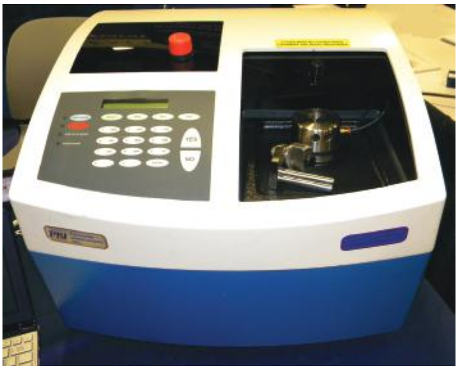

Figure 5.5 Tissue disruption using a pressure-generating instrument also known as a barocycler. During the process, a barocycler is utilized to apply alternating cycles of high and low pressures onto specimens placed in single-use processing containers. The technique is known as pressure-cycling technology (PCT). (© Richard C. Li.)

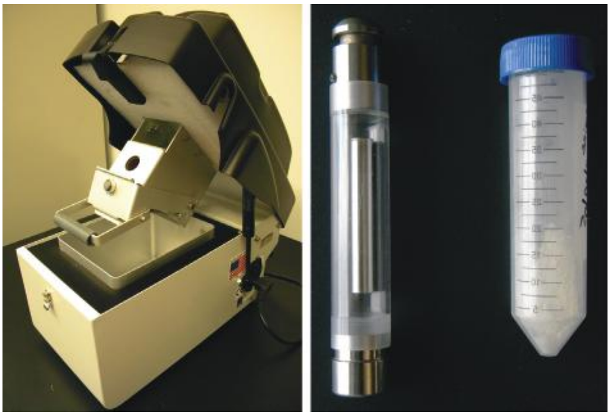

Figure 5.6 Tissue disruption using cryogenic grinding. A cryogenic mill (left), also known as freezer mill, can be used for the pulverization of hard tissue such as bone samples into fine powder prior to DNA extraction. During a grinding cycle, an impactor (right) moves back and forth, under a magnetic field, inside a grinding vial, grinding the sample to a fine powder. The grinding process is carried out at low temperatures (approximately −196°C), using liquid nitrogen to protect temperature-sensitive analytes such as DNA. (© Richard C. Li.)

are embedded in a calcified matrix that is responsible for the rigid structure of bone and teeth. In order to achieve a high yield of DNA extraction, it is necessary to remove the calcified matrix. A decalcification process removes calcium ions from the matrix, thus making the specimen suitable for DNA extraction. The most commonly used decalcifying agents for DNA extraction purposes are chelating agents such as ethylenediaminetetraacetic acid (EDTA), which sequesters the calcium ions. Decalcification is a lengthy procedure. Bone or teeth powder is usually treated with the decalcifying agent overnight or for a longer period of time, depending on the size and source of the sample.

5.1.2 Lysis of Cellular and Organelle Membranes

During or after tissue disruption, membranes—including those of cells, nuclei, and mitochondria—are lysed in order to release DNAs, including nuclear and mitochondrial DNA. The lysis can be carried out using salts and chaotropic agents (Section 5.2.3) such as guanidinium salts and detergents such as sarkosyl and sodium dodecyl sulfate (SDS). These substances can destroy membranes, denature proteins, and dissociate proteins such as histones from DNA. The lysis procedure is usually carried out in a buffer, such as Tris, in order to maintain a pH where endogenous *deoxyribonucleases* (DNases) remain inactive. DNases are a type of nuclease that catalyzes the cleavage of phosphodiester bonds of DNA. Endogenous DNases are located in cytoplasmic lysosomes and play a role in degrading the DNA of invading viruses. When cells are lysed, the DNases that are also released can degrade the extracted DNA. Chelating agents such as EDTA or Chelex® (Section 5.2.2) can therefore be used to chelate the divalent cations that are the cofactors of DNases, in order to inhibit DNase activities. Furthermore, reducing agents such as mercaptoethanol or dithiothreitol (DTT) can be used to inhibit the oxidization processes that can damage DNA.

5.1.3 Removal of Proteins and Cytoplasmic Constituents

After lysis, cytoplasmic constituents, such as proteins and liquids that interfere with DNA extraction, are removed. Proteins and lipids are usually removed by one or more rounds of extraction with organic solvents such as phenol–chloroform mixtures (Section 5.2.1). Another strategy to remove cytoplasmic constituents is to utilize the reversible binding of DNA to a solid material such as silica, which selectively binds DNA in chaotropic salt solutions. The proteins and cytoplasmic constituents can then be removed through washing steps.

5.1.4 Storage of DNA Solutions

Purified high-molecular-weight DNA is usually stored in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). EDTA is usually included in storage solutions to chelate divalent cations and thereby inhibit DNases. Such a DNA solution may be stored at 4°C or −20°C. For long-term storage, −80°C is recommended. Frequent freezing or thawing cycles should be avoided because temperature fluctuations may cause breaks of single- and double-stranded DNA. Additionally, DNA containing impurities such as extracts generated via the Chelex method are much less stable. The presence of heavy metals, such as cadmium and cobalt, can cause breakage of phosphodiester bonds in the molecules. Additionally, unsuccessful removal of exogenous or intracellular free radicals, such as a hydroxyl radical (HO·), which have unpaired valence electrons, can cause DNA damage such as strand breaks. Furthermore, contamination by DNases may lead to subsequent degradation of DNA.

5.1.5 Contamination

Contamination is usually caused by the introduction of exogenous DNA to an evidence sample. It can occur between samples, between an individual and a sample, between other organisms and a sample, or between amplified DNA and a sample. Certain procedures can be utilized to prevent the occurrence of contamination. For example, evidence and reference samples should be processed separately in different rooms to avoid sample-to-sample contamination. In situations where space is limited, the evidence sample should be processed before the reference sample. Additionally, evidence samples should be processed and extracted for isolating DNA in separate areas (or at different times) from DNA amplification areas.

Solutions and test tubes used for extraction should be DNA-free, and aerosol-resistant pipet tips should be used during the extraction process. Additionally, the levels of contamination should be monitored by using *extraction reagent blanks* (having reagents but no samples), which monitor contamination from the extraction.

5.2 Methods of DNA Extraction

5.2.1 Extraction with Phenol–Chloroform

This method is also called organic extraction. Major steps include the following:

5.2.1.1 Cell Lysis and Protein Digestion

These steps can be achieved by digestion with proteolytic enzymes such as proteinase K before extraction with organic solvents.

5.2.1.2 Extraction with Organic Solvents

The removal of proteins is carried out by extracting aqueous solutions containing DNA with a mixture of phenol:chloroform:isoamyl alcohol (25:24:1). Phenol is used to extract the proteins from the aqueous solution. Although phenol has a slightly higher density than water, it is sometimes difficult to separate it from the aqueous phase. Therefore, chloroform is utilized as it has a higher density than phenol. As a result, the phenol–chloroform mixture forms the organic phase at the bottom of the tube and is easily separated from the aqueous phase. Isoamyl alcohol is often added to the phenol–chloroform mixture to reduce foaming. During partition, DNA is solubilized in the aqueous phase, while lipids are solubilized in the organic phase. Proteins are located at the interface between the two phases (Figure 5.7).

5.2.1.3 Concentrating DNA

Two common methods for concentrating DNA are ethanol precipitation and ultrafiltration. In the first method, the DNA is precipitated from the aqueous solution with ethanol and salts. Ethanol depletes the hydration shell of DNA, thus exposing its negatively charged phosphate groups. The

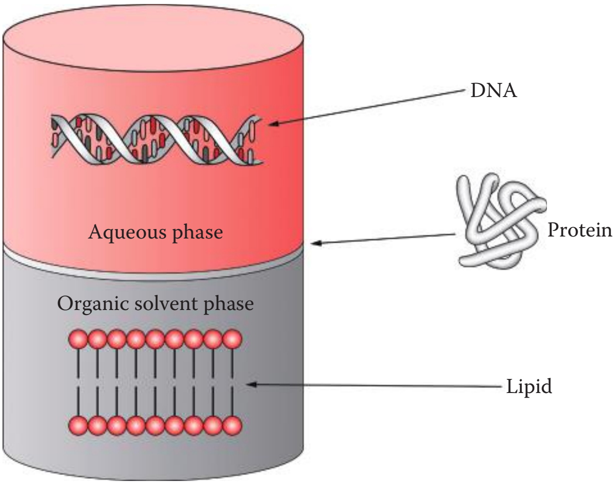
DNA
Protein
Aqueous phase
Organic solvent phase
Lipid

Figure 5.7 DNA extraction using organic solvent. The DNA is contained in the aqueous phase, while cellular materials such as lipids are contained in the organic-solvent phase. Proteins remain in the barrier between the two phases. (© Richard C. Li.)

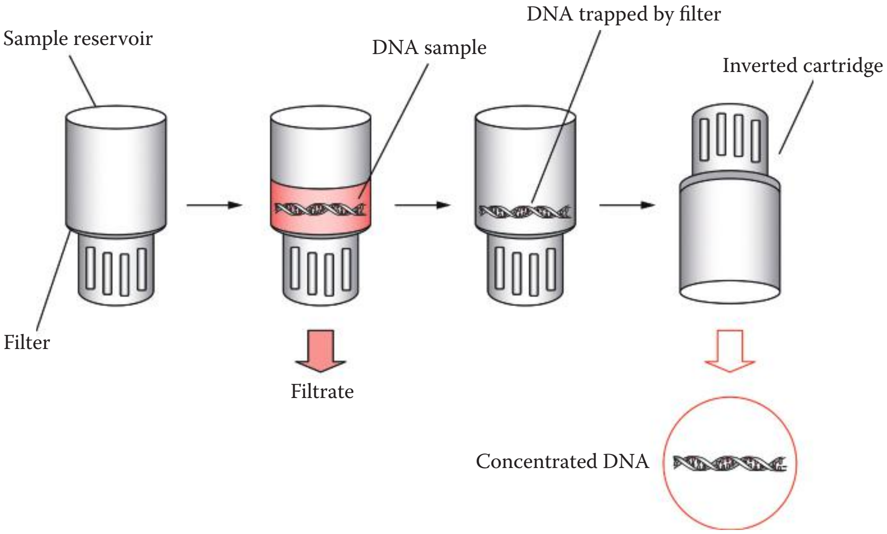
DNA trapped by filter
Sample reservoir
DNA sample
Inverted cartridge
Filter
Filtrate
Concentrated DNA

Figure 5.8 Concentrating DNA solutions using filtration devices. DNA samples are loaded into the reservoir. The liquid is filtered by centrifugation and the DNA becomes trapped in the membrane. The cartridge is then inverted to recover the trapped DNA by centrifugation. (© Richard C. Li.)

precipitation can only occur if sufficient quantities of cations are present in the solution. The most commonly used cations, such as ammonium, lithium, and sodium, neutralize the charges on the phosphate residues of DNA, forming a precipitate. Ultrafiltration is an alternative to ethanol precipitation for concentrating DNA solutions. The Microcon® and Amicon® are centrifugal ultrafiltration devices that can concentrate DNA samples (Figure 5.8). A proper Microcon® unit can be selected with a nucleotide cutoff equal to or smaller than the molecular weight of the DNA fragment of interest. Usually, the cutoff size is 100 kDa for forensic DNA samples.

Phenol–chloroform extraction yields large, double-stranded DNA and can be used for either restriction fragment length polymorphism (RFLP)-based or polymerase chain reaction

(PCR)-based analysis. However, the organic extraction method is time-consuming, involves the use of hazardous reagents, and requires the transfer of samples among tubes.

5.2.2 Extraction by Boiling Lysis and Chelation

This technique, also called the Chelex® extraction method, was introduced in the early 1990s to forensic DNA laboratories. It usually includes the following steps:

5.2.2.1 Washing

This step removes contaminants and inhibitors that may interfere with DNA amplification. For example, heme compounds found in blood samples should be removed from blood samples because it inhibits DNA amplification.

5.2.2.2 Boiling

Cells are suspended in a solution and incubated at 56°C, where DNases are not active, for 20 min. This preboiling step softens cell membranes and separates clumps of cells from each other. The cells are then lysed by heating to boiling temperature in order to break open the membranes and to release the DNA. Additionally, the lysis of cells releases all of their cellular constituents, including DNases. The DNase degradation of the extracted DNA can be blocked by applying a chelating resin (Chelex® 100) during the extraction process. Chelex® 100 is an ion-exchange resin composed of styrene divinylbenzene copolymers. The paired iminodiacetate ion groups in Chelex® 100 act as chelators by binding to divalent metal ions such as magnesium. Magnesium is a cofactor of endogenous DNases. Thus, sequestering magnesium in the solution using Chelex® 100 protects DNA from degradation by DNases (Figure 5.9).

5.2.2.3 Centrifugation

Brief centrifugation is performed to pull the Chelex® 100 resin and cellular debris to the bottom of the tube. The supernatant is used for DNA analysis. Carrying the Chelex® 100 resin over into the DNA amplification solutions should be avoided because the resin chelates magnesium, which is a necessary cofactor for DNA polymerases used for amplification.

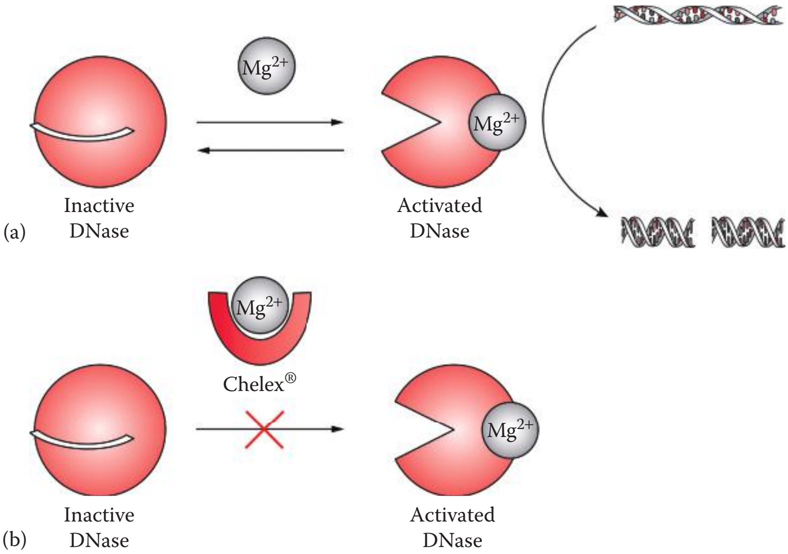

Figure 5.9 Utilizing Chelex® for preparing DNA samples. (a) Cations such as Mg^{2+} are required for the activity of endogenous DNase, which degrades DNA. (b) Chelex® prevents DNA degradation from endogenous DNase by sequestering the Mg^{2+}. (© Richard C. Li.)

This method is simple and rapid and uses only a single tube for extraction, thus reducing the risks of contamination and sample mix-ups. However, the heating step of this method disrupts and denatures proteins and also affects the chromosomal DNA. The resulting DNA extracted from the solution is fragmented single-stranded DNA. Thus, the DNA extracted is not suitable for RFLP analysis because RFLP requires double-stranded DNA samples. The DNA obtained by lysis and chelation can only be used for PCR-based DNA analysis.

5.2.3 Silica-Based Extraction

The method of adsorbing DNA molecules to solid silica surfaces is used for extraction (Figures 5.10 and 5.11). The method is based on the phenomenon that DNA is reversibly adsorbed to silica—silicon dioxide (SiO_2)—in the presence of high concentrations of chaotropic salts. Chaotropic salts can disrupt hydrogen bonding, affecting the three-dimensional structures of macromolecules. These salts are used to denature proteins. Additionally, chaotropic salts can facilitate the adsorption of DNA to silica.

In aqueous solutions, the hydration shells of nucleic acids shield the negative charges of phosphates at the phosphodiester backbone of nucleic acids. As a result, nucleic acids are usually hydrophilic in aqueous solutions. In the presence of chaotropic salts, nucleic acids become hydrophobic. The dehydrating effect caused by chaotropic salts allows the phosphate residues to become available for adsorption to the silica surface. Common chaotropic salts utilized for DNA extraction include guanidinium salts such as guanidinium thiocyanate (GuSCN) and guanidinium hydrochloride (GuHCl). GuSCN is a more potent chaotropic salt and also facilitates cell lysis and DNA adsorption. This technique usually includes the following steps:

5.2.3.1 Cell Lysis and Protein Digestion

This is carried out by proteinase K digestion. The cell membranes are broken open, and DNA is released.

5.2.3.2 DNA Adsorption onto Silica

This step utilizes silica as the stationary phase in a membrane configuration to which the DNA in the cell lysate binds. Adsorption of the DNA to the silica occurs in the presence of high concentrations of chaotropic agents (some protocols adjust pH conditions to enhance adsorption). Under these conditions, cellular materials and other contaminants that can inhibit DNA amplification reactions are not retained on the silica membrane. The adsorbed DNA is largely double stranded.

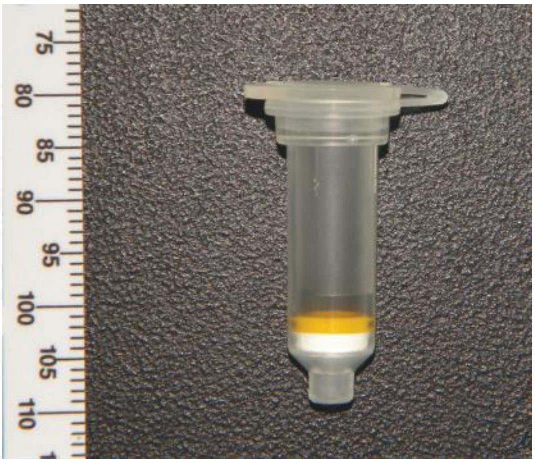

Figure 5.10 A silica-membrane spin column. (© Richard C. Li.)

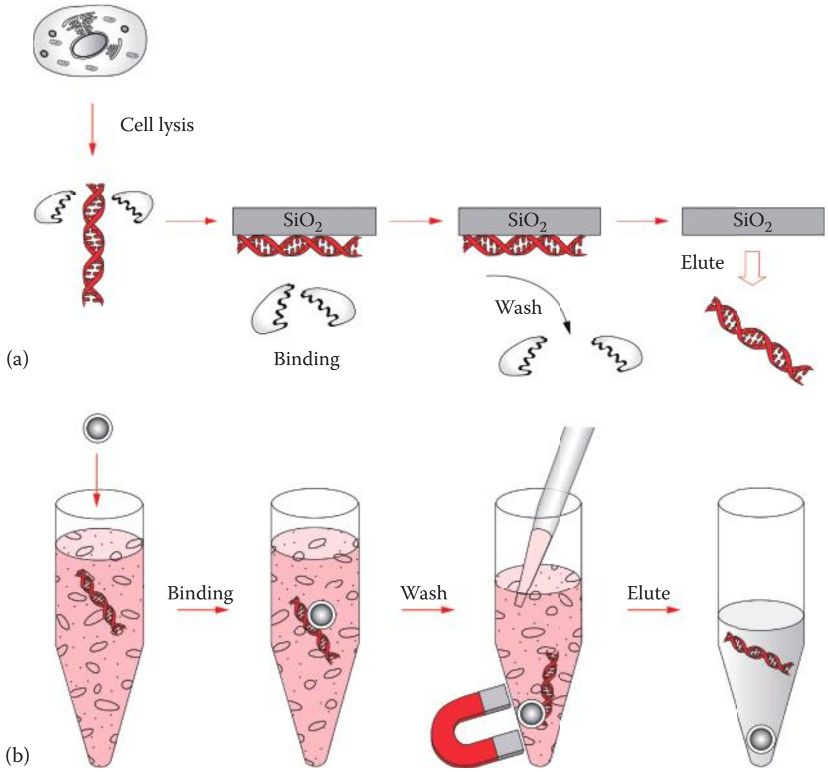

Figure 5.11 Silica-based DNA extractions. (a) Cells are lysed in the presence of proteinase. The DNA then binds to the silica matrix. A washing step removes unbound cellular materials and salts from the matrix. The purified DNA is then eluted for use in downstream applications. (b) Using silica-coated paramagnetic particles for DNA extraction. The particle is added into a lysate, and then DNA binds to the silica surface of the particle. The particle is then captured by a magnetic field. Afterwards, the supernatant is removed, and cellular materials are washed away. DNA elution follows. See Section 5.2.3. (© Richard C. Li.)

5.2.3.3 Washing

This step removes chaotropic agents and other contaminants. An ethanol-based wash solution is used. This wash solution does not remove DNA from the silica. The chaotropic agents and contaminants that are present in the solution can be removed using the ethanol-based wash solution.

5.2.3.4 Elution of DNA

The adsorbed DNA can be eluted by rehydration with aqueous low-salt solutions. The eluted DNA is double stranded and can be used for a wide variety of applications.

Silica-based extraction methods yield high-quality DNA. Silica membrane devices can also be adapted for automation; for example, this can be done by using 96-well silica membrane plates and a variety of robotic platforms. Another type of device utilizes silica-coated paramagnetic particles that adsorb DNA in the solution (Figure 5.11b). A magnet is used for particle capture instead of centrifugation or vacuum filtration. The magnetic particles can be resuspended during the wash steps, and the solution containing contaminants and cellular materials is then discarded. DNA is eluted after washing. This device can also be adapted to automated, high-throughput methods.

Over the years, high-throughput silica-based procedures have been developed to process large numbers of samples in parallel. Some of these methods are adapted for automated DNA extraction platforms (Figures 5.12 and 5.13).

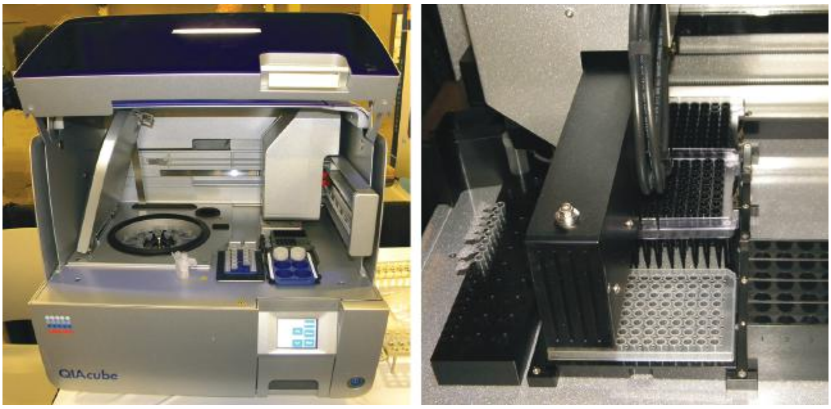

Figure 5.12 Automated bench-top DNA purification systems that enable the isolation of genomic DNA from a wide variety of forensic samples. A low-throughput sample-preparation system that can process up to 12 samples (left) and a low-to-medium throughput sample-preparation system (right). (© Richard C. Li.)

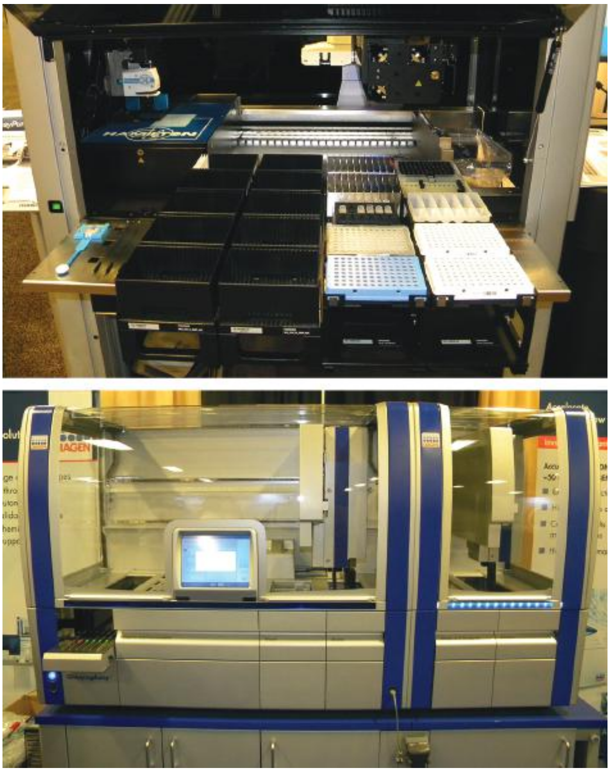

Figure 5.13 Integrated platforms enabling automated DNA sample preparation and liquid handling for subsequent assay setup. These systems utilize silica-coated bead chemistries and can achieve moderate to high throughput of processing (1–96 samples) with bar coding for sample tracking. A Hamilton (top) and a Qiagen system (bottom). (© Richard C. Li.)

5.2.4 Differential Extraction

This method is very useful for the extraction of DNA from biological evidence derived from sexual assault cases, such as vaginal swabs and bodily fluid stains. These types of evidence often contain mixtures of spermatozoa from a male contributor and nonsperm cells such as epithelial cells from a female victim. Mixtures of individual DNA profiles can complicate data interpretation.

This method selectively lyses the nonsperm and spermatozoa in separate steps based on the differences in cell-membrane properties of spermatozoa and other types of cells. Thus, the DNA from spermatozoa and nonsperm cell fractions can be sequentially isolated.

First, the differential extraction procedure involves preferentially lysing the nonsperm cells with proteolytic degradation using proteinase. Sperm plasma membrane contains proteins cross-linked by disulfide bonds. The membrane exhibits a much higher mechanical stability than nonsperm cells and is thus resistant to proteolytic degradation. The nonsperm DNA is released into the supernatant and the liquid containing it (the nonsperm fraction) is extracted, yielding a fraction that predominantly contains DNA from nonsperm cells.

To lyse the sperm cells, it is necessary to cleave the disulfide bonds in addition to proteolytic digestion. The application of DTT, a reducing agent, is an approach that can be used for cleavage. In the presence of DTT and proteinase K, the sperm plasma membrane is then lysed (Figure 5.14). Subsequently, DNA from the sperm cells can be extracted (Figure 5.15).

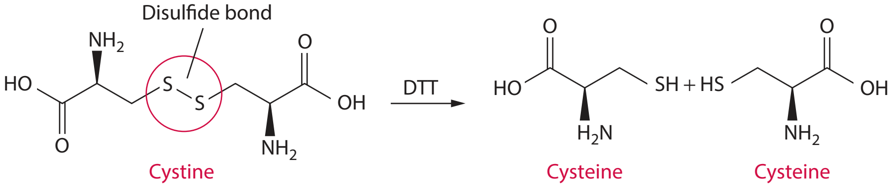

Figure 5.14 DTT reaction. The breaking of disulfide bonds in cystine residues is carried out by adding a reducing agent such as DTT.

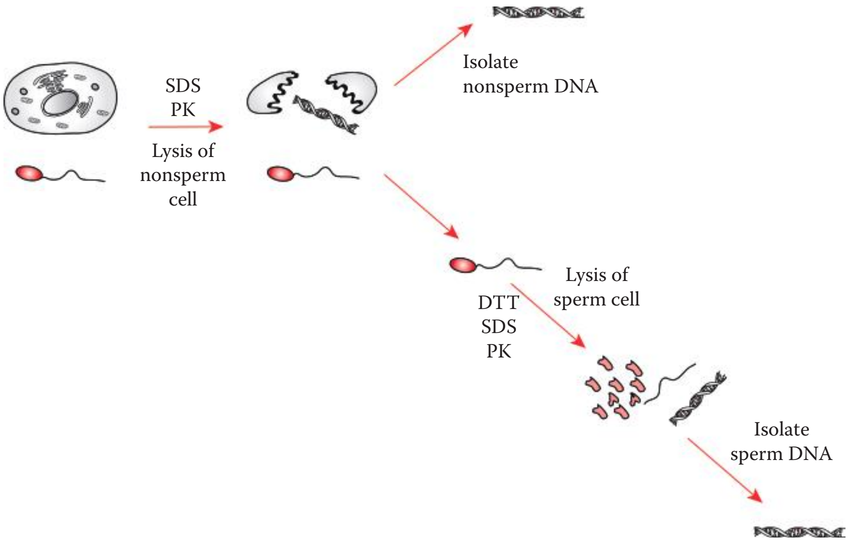

Figure 5.15 Differential extraction process used to separate sperm cells from nonsperm cells. Nonsperm cells are lysed in the presence of SDS and proteinase K (PK); sperm cells are resistant to such conditions. The nonsperm cell DNA is extracted. The sperm cells are then lysed separately in the presence of SDS and proteinase K plus DTT to extract the sperm DNA. (© Richard C. Li.)

This process isolates sperm and nonsperm cell DNA separately for obtaining DNA profiles from male and female contributors, respectively. However, the non-sperm-cell DNA and sperm-cell DNA may not be completely separated from one another; this can happen, for example, if the sperm cells have already lysed due to poor sample conditions. Some sperm DNA may be present in the non-sperm-cell fraction. Additionally, if a mixture has an abundance of non-sperm cells and fewer sperm cells, non-sperm-cell DNA may be detected in the sperm fraction. Thus, new methods that can overcome these problems are highly desired.

5.3 Essential Features of RNA

Like DNA, *ribonucleic acid* (RNA) is a linear molecule containing four types of nucleotides linked by phosphodiester bonds (Figures 5.16 and 5.17). However, certain properties of RNA differ from those of DNA. Unlike DNA containing deoxyribose (see Figure 5.18), the sugar residue

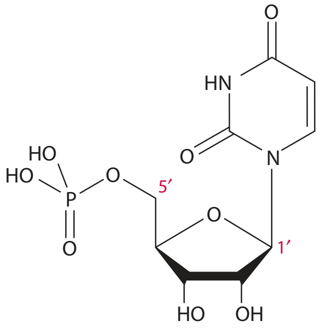

Figure 5.16 The structure of ribonucleotide. Each ribonucleotide in a RNA polymer consists of three components: a ribose, a nitrogenous base (uracil is shown as an example), and a phosphate group. The 1′ carbon of the ribose is attached to the nitrogen of the nitrogenous base. A phosphate group is attached to the 5′ carbon of the ribose.

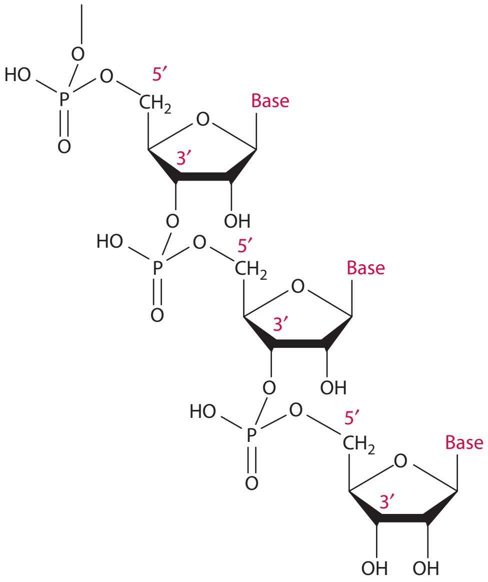

Figure 5.17 Chemical structure of RNA (polyribonucleotide) chain.

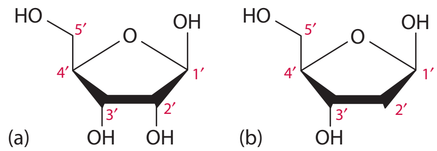

Figure 5.18 Comparison of (a) ribose and (b) deoxyribose. The hydroxyl group attached to the 2′ carbon of ribose is replaced by a hydrogen group in deoxyribose.

of RNA is a ribose, which has a hydroxyl (OH) group at the 2′ carbon position. Therefore, RNA has a lower pKa, which means that it is more acidic than DNA. RNA contains uracil (U) in the place of thymine in DNA (Figure 5.19). RNA is typically found in cells as a single-stranded molecule, while DNA is double stranded. Polyribonucleotides can form complementary helices with DNA strands by base pairing (with the exception of uracil pairs with adenine) (see Figure 5.20).

Total RNA contains all the RNA of the cells, including RNAs involved in protein synthesis and posttranscriptional modification, as well as regulatory RNAs. The RNA that carries codes from a DNA template is called messenger RNA (mRNA) and usually contains a cap and polyadenine tail at the 5′ and 3′ ends of the molecule, respectively (Figure 5.21).

The stability of RNA is an issue for mRNA-based forensic analysis. Single-stranded RNA is chemically less stable than DNA. The 2′ OH group of RNA can react with its phosphodiester backbone, potentially causing the nonenzymatic hydrolysis of an RNA molecule. Several factors, such as moisture, UV light, high temperature, and extreme pH, can facilitate the nonenzymatic

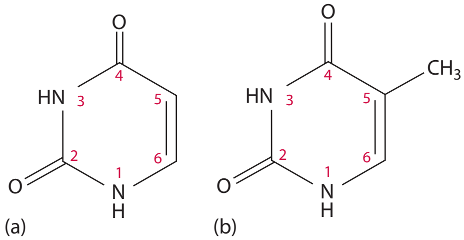

Figure 5.19 Comparison of (a) uracil and (b) thymine. Thymine has a methyl group attached at the 5-carbon position.

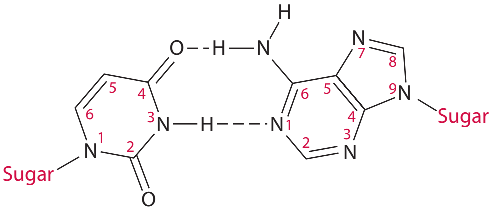

Figure 5.20 Base pairing of uracil (left) and adenine (right).

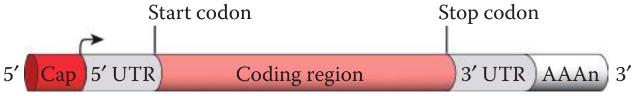

Figure 5.21 Structure of mRNA. Eukaryotic mRNA is modified at the 5′ end (cap) and 3′ end (polyadenine tail) needed for protein synthesis. Start and stop codons are required for the initiation and termination of protein synthesis. (© Richard C. Li.)

hydrolysis of RNA phosphodiesters, which leads to the degradation of RNA. Moreover, endogenous *ribonucleases* (RNases), present within cells, represent the major factor causing RNA degradation. Environmental microorganisms are also a common source of RNase contamination. RNases are very stable and usually do not require cofactors to carry out enzymatic reactions. Small amounts of RNases are sufficient to degrade RNA. RNA can be protected by adding RNase inhibitors such as diethylpyrocarbonate. Additionally, using RNase-free laboratory supplies and reagents and wearing disposable gloves while handling samples and reagents can reduce the risks of RNase contamination.

5.4 Methods of RNA Extraction

5.4.1 RNA–DNA Coextraction

For most forensic applications, total RNA is usually isolated. The procedure for isolating total RNA is simpler than that for mRNA. The total RNA isolated usually contains a sufficient amount of mRNA for subsequent reverse transcriptase PCR (RT-PCR; see Chapter 7). However, if the quantity of a target mRNA is very low, the procedures specifically for isolating mRNA should be used. There are a variety of protocols for isolating total RNA. However, RNA–DNA coextraction methods allow for the simultaneous extraction of high-quality DNA and RNA for forensic DNA analysis and bodily-fluid identification (Chapter 11), respectively, without the consumption of additional samples when extracted separately.

Biological samples are first lysed in lysis buffer. The lysis buffer usually contains chaotropic salts that can facilitate the lysis of cells and denature proteins, thus inactivating endogenous RNases to protect RNA. The lysate is then passed through a silica membrane column. A high-salt environment in the lysate allows selective binding of silica to genomic DNA over RNA. The column is washed, and purified DNA is then eluted.

The extraction of RNA is achieved by utilizing a high concentration of chaotropic salts that are already present in the lysate, along with ethanol, in order to decrease the hydrophilic property of RNA and increase its affinity for silica. After ethanol is added to the flow-through that passed the first column, the sample is then applied to a second silica column, where total RNA molecules, longer than 200 nucleotides, bind to the silica membrane. RNA including mRNA is then eluted (Figure 5.22). This method does not isolate small RNAs such as miRNA, rRNA, and tRNA, which comprise 15%–20% of total RNA.

5.4.2 miRNA Extraction

miRNAs (Chapter 4) are low-molecular-weight RNAs ranging between 15 and 30 nucleotides in length. Conventional methods routinely used for extracting total RNA do not effectively recover small RNAs; thus, they are not suitable for isolating miRNA. One approach to extracting small RNA molecules including miRNA involves two steps: organic-solvent extraction to isolate total RNA and solid-phase extraction to enrich small RNA.

To begin the process, tissues are disrupted and cells are lysed. The lysis reagents also inactivate RNases to protect RNA. The lysate is then extracted with an organic solvent, such as phenol and chloroform, with a high concentration of chaotropic salts such as GuSCN. The partitioning of DNA and RNA between the organic phase and the aqueous phase is determined by the pH during the organic extraction. At an acidic pH, DNA partitions to the organic phase. The pKa of RNA is usually lower than DNA; thus, RNA remains hydrophilic and is retained in the aqueous phase under this condition. This step removes most of the DNA, proteins, and other cellular components from the lysate into the organic phase.

The second step, silica-based extraction, further purifies and enriches small RNAs. It is achieved by utilizing the high concentration of chaotropic salt that is already present in the lysate, along with ethanol. The size-fractioning of RNAs is achieved using different ethanol

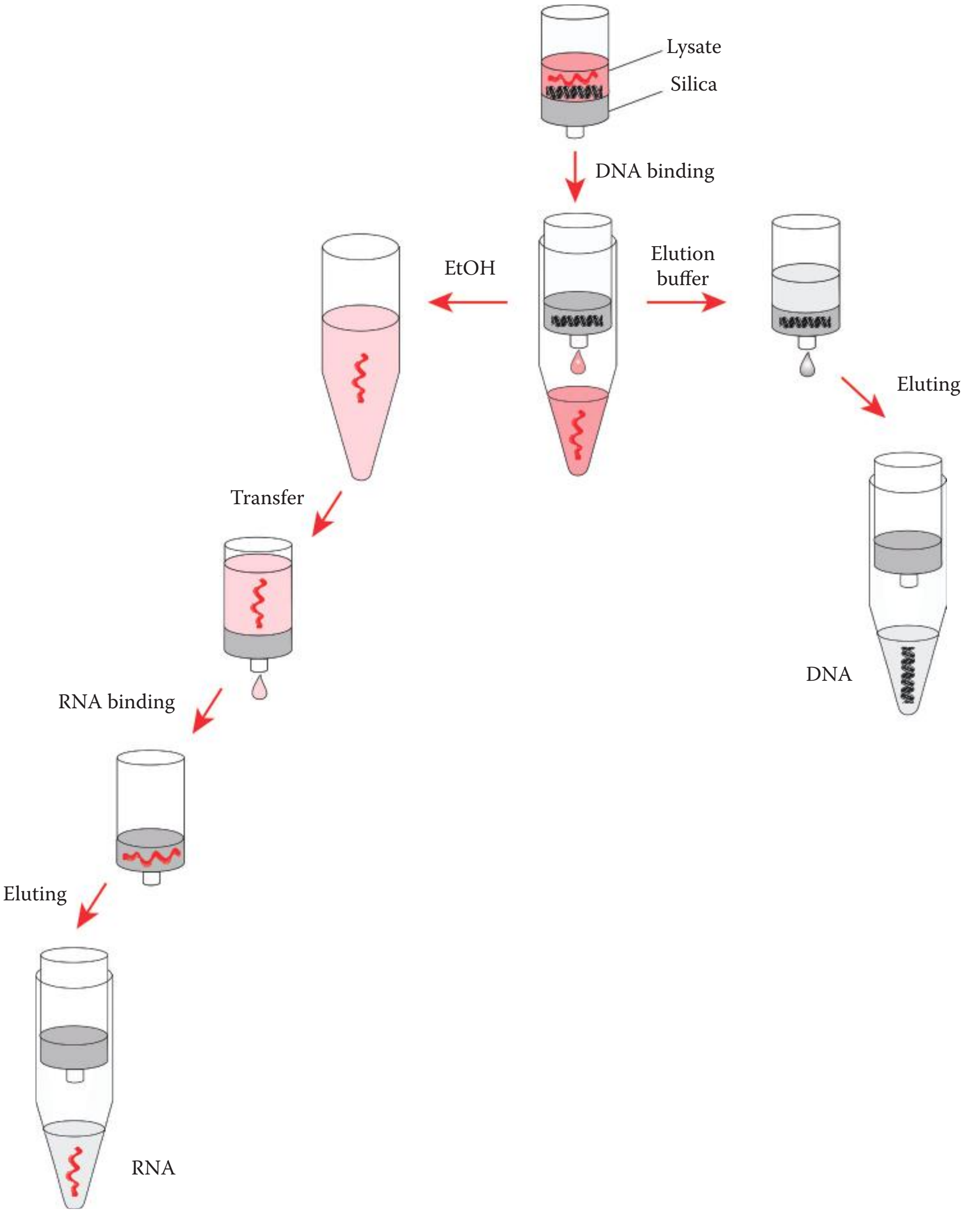
Lysate
Silica
DNA binding
EtOH
Elution
buffer
Eluting
Transfer
DNA
RNA binding
Eluting
RNA

Figure 5.22 RNA and DNA coextraction using a silica-based method. See Section 5.4.1. (© Richard C. Li.)

concentrations. At a low ethanol concentration, the solvent-extracted lysate is passed through a silica membrane filter. Large RNAs are retained on the silica membrane, while small RNAs are not and are collected in the filtrate. The filtrate is passed through a second silica membrane filter. At a high ethanol concentration, the small RNAs are retained on the silica membrane. The small RNAs are then eluted in a low ionic strength solution (Figure 5.23). RNA less than 200 nucleotides, including miRNA, can be obtained.

After RNA is extracted, the small amounts of RNA can be quantified using a quantitative RT-PCR (Chapter 7) or fluorescent intercalating dye assay (Chapter 6). Purified RNA can be

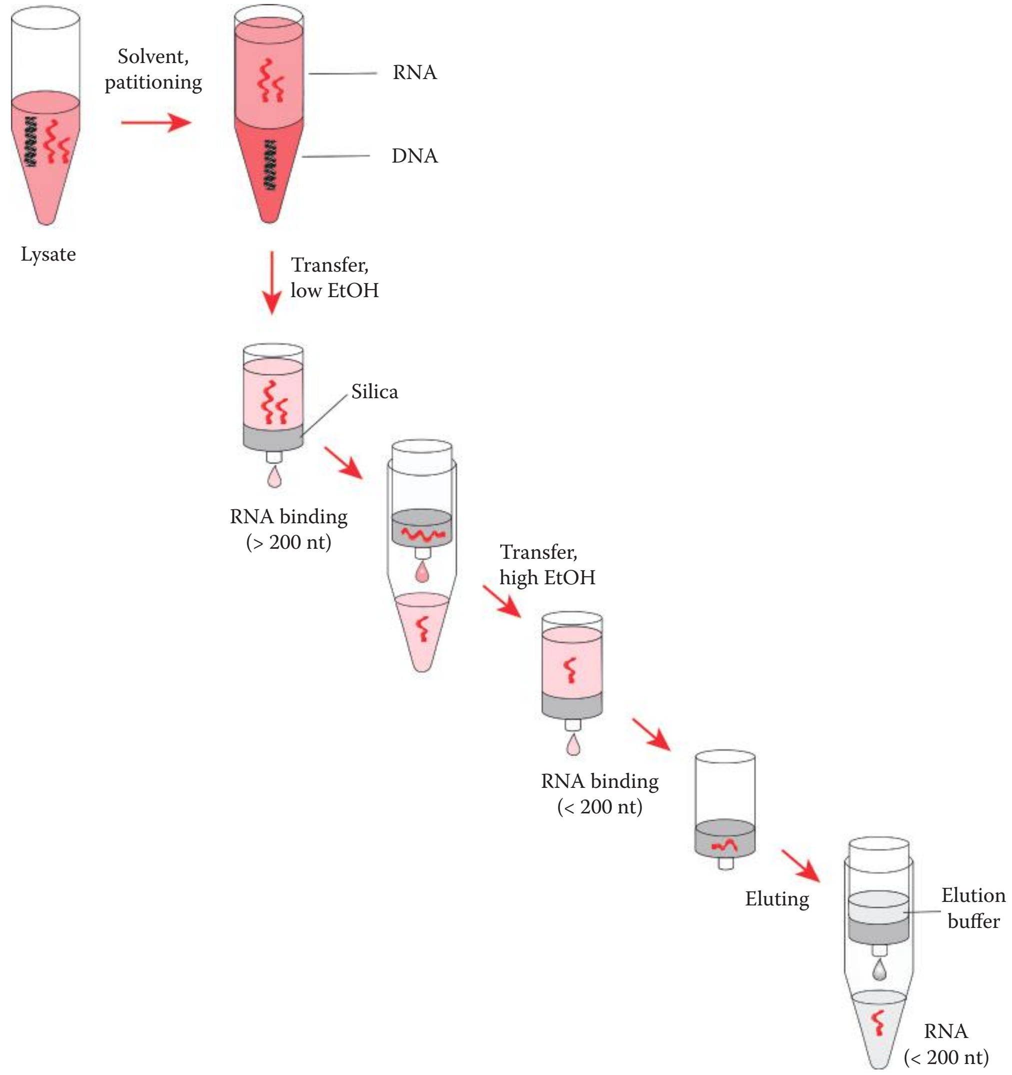
Solvent,
patitioning
RNA
DNA
Lysate
Transfer,
low EtOH
Silica
RNA binding
(> 200 nt)
Transfer,
high EtOH
RNA binding
(< 200 nt)
Eluting
Elution
buffer
RNA
(< 200 nt)

Figure 5.23 Schematic illustration of an miRNA extraction method. See Section 5.4.2. (© Richard C. Li.)

stored at −80°C in RNase-free water or Tris-EDTA buffer (pH 7). Single-use aliquots are preferred when possible to avoid multiple freeze–thaw cycles.

The integrity of extracted RNA is a great concern for RNA-based forensic analysis. RNA quality is traditionally assessed using the 28S:18S rRNA ratio. There are some conditions where this method has been shown to be inconsistent; that is, in which the 28S:18S rRNA ratio does not reflect the true level of RNA degradation. Recently, the *RNA integrity number* (RIN) has provided an effective method for determining RNA quality. RIN is a software algorithm that takes the electrophoretic RNA measurements into account in order to assign integrity values to RNA samples. RIN ranges from 1 to 10, with 10 being the most intact. Thus, the RIN method facilitates the assessment of the integrity of RNA samples. As reflected by RIN, RNA degradation has a negative influence on the reproducibility of the results of RNA-based forensic analysis.

6

DNA Quantitation

Determining the amount of DNA in a sample is essential for polymerase chain reaction (PCR)-based DNA testing (Chapter 7). Because PCR-based DNA testing is very sensitive, a narrow concentration range is required for amplification to be successful. If too much DNA template is used in PCR, the resulting artifacts may interfere with data analysis and interpretation. On the other hand, very low amounts of DNA template may result in a partial DNA profile or failure to attain a profile. Samples with poor purity may contain PCR inhibitors that lead to a failure of DNA amplification. Thus, a test that can measure the quality and quantity of the DNA template of a sample is desirable.

Additionally, forensic samples containing DNA are often mixtures that may include nonhuman DNA. For instance, microbial DNA may be present. Thus, it is necessary to use human-specific DNA quantitation methods to selectively determine the amount of human DNA present. In the United States, quality-assurance guidelines require the use of a quantitation method that estimates the amount of human nuclear DNA in crime-scene evidence samples.

In this chapter, the slot blot, intercalating dye, and quantitative PCR methods are introduced. The quantitative PCR method is the most sensitive of the three methods. It is the only method that can detect PCR inhibitors. Both the slot blot and the quantitative PCR methods can detect human and nonhuman primate DNA. Current technology cannot distinguish between human and nonhuman primate DNA.

6.1 Slot Blot Assay

Historically, the slot blot assay was used to detect human genomic DNA in a sample. Methodically, the slot blot assay works based on the following principles. Prior to the quantitation, an alkaline solution is added to the genomic DNA sample, which denatures DNA. Generating single-stranded DNA is necessary for DNA to be cross-linked onto a nitrocellulose membrane. The DNA sample is then spotted, using a slot blot device, onto a nitrocellulose membrane (Figure 6.1). The single-stranded DNA is then immobilized onto a nylon membrane. The targeted sequence is revealed by hybridization with a labeled 40-nucleotide probe complementary to a primate-specific α-satellite DNA sequence at the D17Z1 locus (Figure 6.2). In humans, the α-satellite DNA sequences are highly repetitive sequences located near the centromeres of chromosomes. These sequences are usually distinct for each chromosome.

Three detecting schemes of the slot blot assay have been developed. Initially, the D17Z1 probe was labeled with radioisotopes that could be visualized by exposing the slot blot membrane to x-ray film. The hazardous radioisotope detection method was then replaced by alkaline phosphatase-labeled and biotinylated probes. The alkaline phosphatase-labeled probe can be coupled with chemiluminescent detection (Lumi-Phos Plus kit, Lumigen, Inc.). The biotinylated probe

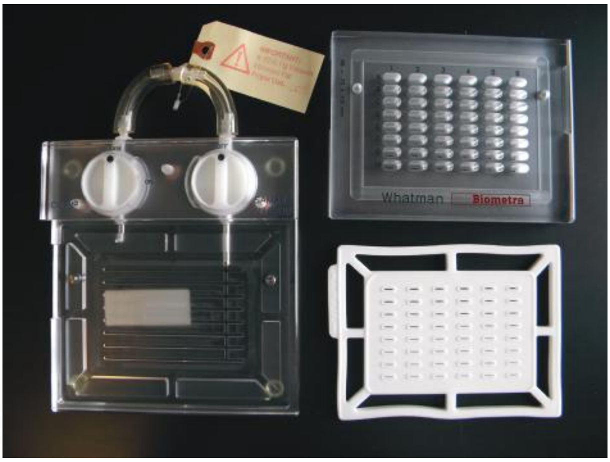

Figure 6.1 A slot blot device. (© Richard C. Li.)

(QuantiBlot Human DNA Quantitation Kit, Applied Biosystems) can be coupled with either colorimetric or chemiluminescent detection.

In colorimetric detection, the biotin moiety of the probe is bound to streptavidin. The streptavidin is conjugated with horseradish peroxidase, which catalyzes the oxidation reaction of tetramethylbenzidine (TMB), a substrate, forming a blue precipitate. With the chemiluminescent detection method, the horseradish peroxidase catalyzes the oxidation reaction of a substrate such as luminol, emitting photons that can be detected by exposure to x-ray film. The sensitivity of chemiluminescent detection is slightly higher than that of the colorimetric detection. The detection mechanisms will be discussed in Chapter 9.

The detected signal intensity is proportional to the concentration of the DNA sample in question. Quantitative measurements can be made by comparing an unknown sample to a set of standards with known DNA concentrations (Figure 6.3).

The assay typically quantifies DNA over the range of 150 pg–10 ng. However, the quantitation results are manually read, and conclusions are based on subjective judgments. Additionally, the D17Z1 sequences of humans and other primates share homology. The probe cannot distinguish between human and other primate genomic DNA. This is not a great concern because cases involving nonhuman primates are rare. Nevertheless, the probe does not cross-react with all other species. This assay has been replaced by recently developed quantitative PCR assays (see Section 6.3).

6.2 Fluorescent Intercalating Dye Assay

Small quantities of DNA can also be quantified by using a fluorescent intercalating dye method. Intercalating dyes, usually planar molecules, can slide themselves in between base pairs of DNA without breaking the DNA double helix. The Quant-iT™ PicoGreen® dsDNA reagent (Invitrogen) is a fluorescent intercalating dye that stains double-stranded DNA (dsDNA) for quantitation in a sample (Figure 6.4a). The detection limit of this method is approximately 250 pg. Intercalating dyes, not specific to human DNA, bind to all DNA molecules. Therefore, fluorescent intercalating dye assay can be utilized for the quantitation of known reference samples. For instance, DNA database samples from known sources can be quantified using this method. The assay has also been adapted for automation and is, thus, a high-throughput method.

DNA samples are simply added to a solution containing the fluorescent intercalating dye. The fluorescence, proportional to the quantities of DNA (Figure 6.4b), is measured using a standard spectrofluorometer with excitation and emission wavelengths of the light source. A standard

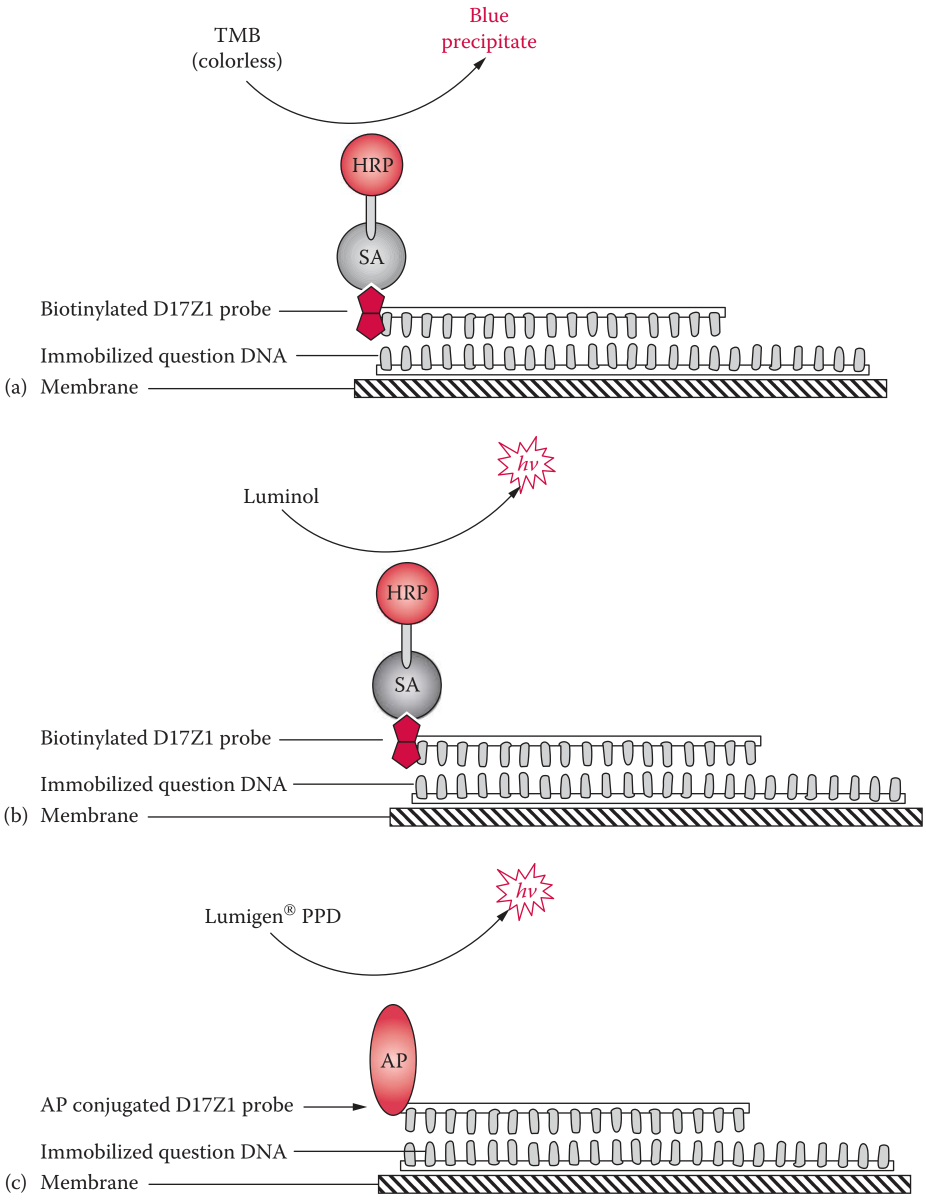
TMB
(colorless)
Blue
precipitate
HRP
SA
Biotinylated D17Z1 probe
Immobilized question DNA
(a) Membrane
Luminol
hv
HRP
SA
Biotinylated D17Z1 probe
Immobilized question DNA
(b) Membrane
Lumigen® PPD
hv
AP
AP conjugated D17Z1 probe
Immobilized question DNA
(c) Membrane

Figure 6.2 Slot blot assay. Questioned DNA is immobilized onto a solid-phase membrane then hybridized with a biotinylated D17Z1 probe. The detection of the hybridization is carried out by (a) streptavidin (SA) and horseradish peroxidase (HRP) conjugate, and a colorimetric reaction is catalyzed by HRP using tetramethylbenzidine (TMB) as a substrate; (b) SA and HRP conjugate, and a chemiluminescent reaction is catalyzed by HRP using Luminol as a substrate; (c) immobilized DNA is hybridized with an alkaline phosphatase (AP)-labeled D17Z1 probe. The detection of the hybridization is carried out by a chemiluminescent reaction catalyzed by AP using Lumigen® PPD as a substrate. (© Richard C. Li.)

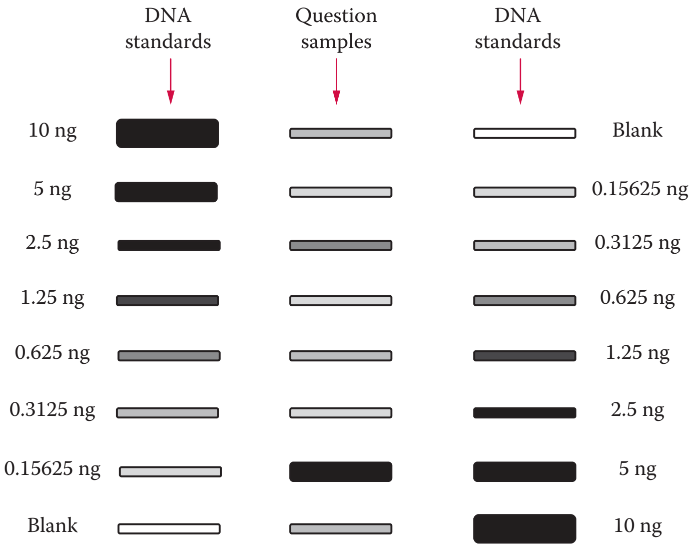
DNA standards
Question samples
DNA standards
10 ng
5 ng
2.5 ng
1.25 ng
0.625 ng
0.3125 ng
0.15625 ng
Blank
Blank
0.15625 ng
0.3125 ng
0.625 ng
1.25 ng
2.5 ng
5 ng
10 ng

Figure 6.3 Human DNA quantitation using the slot blot assay. Standards with known amounts of human DNA are applied, and unknown samples and a set of standards are compared. The quantities in the unknown samples are estimated by visual comparison to the standards. (© Richard C. Li.)

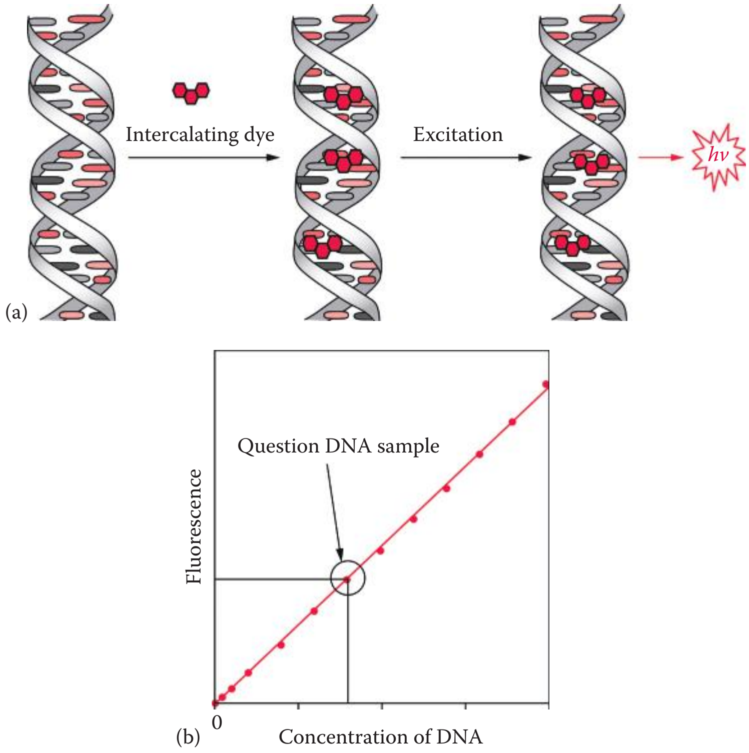
Intercalating dye
Excitation
hv
(a)
Question DNA sample
Fluorescence
0
Concentration of DNA
(b)

Figure 6.4 DNA quantitation using intercalating dye. (a) Fluorescent dye intercalates into DNA. The fluorescence can be measured upon applying an excitation light source. (b) A standard curve can be constructed using known amounts of DNA standards. The amount of questioned DNA can be determined by comparing the standard curve. (© Richard C. Li.)

curve is first created using samples containing known amounts of DNA. The assay is then performed for the unknown samples and the quantities of DNA in the samples are determined by comparing the results to the standard curve.

6.3 Quantitative PCR Assay

Based on the principle of PCR amplification (Chapter 7), the amount of PCR product amplified correlates with the initial concentration of DNA templates. Thus, the DNA concentration of a sample can be determined. There are two types of quantitative PCR methods. *End-point PCR* methods measure the quantity of amplified product at the end of PCR. Usually, the fluorescence is emitted by the dyes that intercalate into the double-stranded DNA. The quantity of amplified DNA is measured from the amount of fluorescence emitted from dyes such as SYBR (Figures 6.5 and 9.1b).

Real-time PCR methods can quantify the amplified DNA during the exponential phase of PCR (Chapter 7). The quantitation result is not affected to a significant extent by slight variations in PCR conditions. Thus, the precision of the quantitation of target sequences is improved with this method.

6.3.1 Real-Time Quantitative PCR

Real-time quantitative PCR (qPCR) was developed in the early 1990s, and it analyzes the amplification of a target sequence at each cycle of PCR. A fluorescent reporter is used to monitor the accumulation of amplified products during PCR. The fluorescence signals of the reporter molecule increase as amplified products accumulate with each cycle of PCR. qPCR is commonly used because of the following advantages:

- Better objectivity than the QuantiBlot method
- Increased sensitivity with a large dynamic range (30 pg–100 ng)
- More accurate measurements of small quantities of DNA in samples
- Fewer laboratory manipulations; amenable to automation
- Ability to detect PCR inhibitors (Section 6.3.1.1)

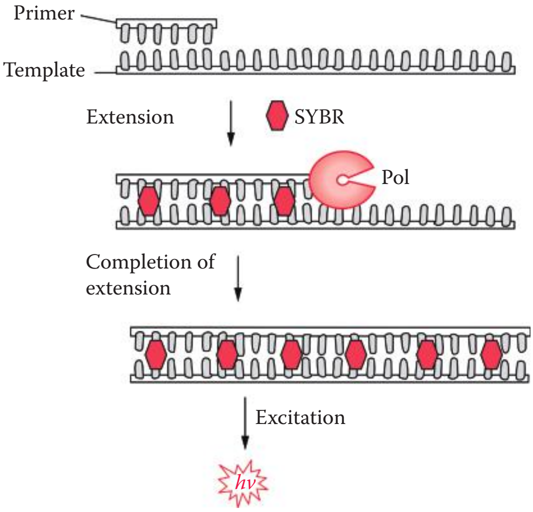

Figure 6.5 End-point PCR using SYBR Green detection. During the extension phase of PCR, in which DNA synthesis occurs, the dye binds to the double-stranded amplicons. Upon excitation, the emission intensity of the dye can be measured. Pol represents Taq polymerase. (© Richard C. Li.)

The technique is amenable to multiplexing to detect more than one type of DNA target sequence in a single reaction. Commercial qPCR kits for human DNA and Y-chromosome DNA quantitation are available. Additionally, the qPCR method for mtDNA quantitation is possible. qPCR uses commercially available fluorescence-detecting thermocyclers to amplify specific DNA sequences and measure their concentrations simultaneously. The fluorescent reporter can be a nonspecific intercalating double-stranded DNA-binding dye or a sequence-specific fluorescently labeled oligonucleotide probe (Chapter 9). The target sequences are amplified and detected by the same instrument, and the reporter fluorescence is monitored externally. Thus the reaction tubes do not need to be opened. This minimizes aerosol contamination and reduces the risk of false-positive results. A widely used qPCR probe technique is the TaqMan method (Applied Biosystems).

6.3.1.1 TaqMan Method

This method utilizes the 5′ exonuclease activity of *Taq* polymerase to cleave the probe during PCR (also known as the 5′ exonuclease assay). The probe is designed to anneal to the target sequence between the upstream and downstream primers and is added to the PCR mixture together with primers (Figure 6.6). The probe T_m (melting temperature; see Chapter 7) should be higher than the amplification primer T_m. A minor groove binder (MGB), such as dihydrocyclopyrroloindole tripeptide, is often linked at the 3′ end of the probe (Figure 6.6). A conjugated MGB binds to the minor groove of a B-form DNA helix (Figure 6.7), which is stabilized by van

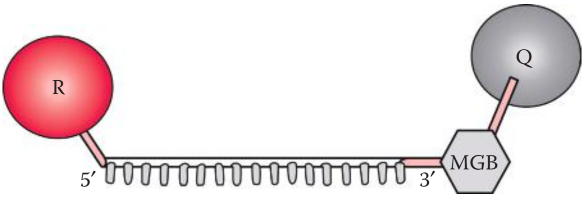

Figure 6.6 TaqMan probe. Each probe is labeled with a reporter dye (R) on the 5′ end and a fluorescence quencher (Q) on the 3′ end. MGB represents a minor groove binder. (© Richard C. Li.)

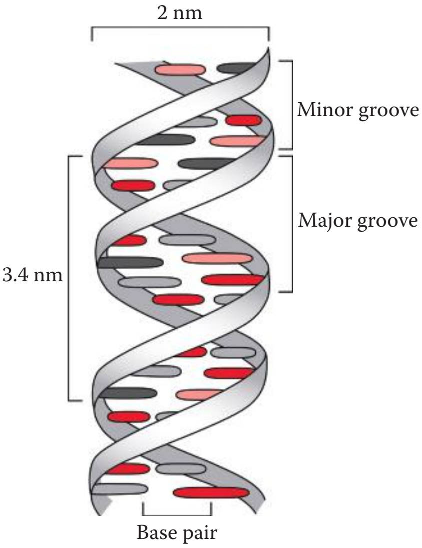

Figure 6.7 DNA double helix. The double-helical structure of B-DNA, the most common form of DNA, is shown. With a helical diameter of 2 nm, each turn of the helix takes 3.4 nm, which corresponds to 10 base pairs per turn. The major and minor grooves are shown. (© Richard C. Li.)

der Waals forces. As a result, conjugating an MGB into a probe increases the T_m values, allowing for the use of shorter probes. As a result, having the probe T_m higher than that of the primers ensures that the probe is fully hybridized during primer extension. The oligonucleotide probe is labeled with both a *reporter* fluorescent dye, usually 6-carboxyfluorescein (6-FAM) or tetrachlorofluorescein (TET) at the 5′ end; and a nonfluorescent *quencher* moiety, such as tetramethylrhodamine (TAMRA), usually at the 3′ end or any thymine position. While the probe is intact, the quencher greatly reduces the fluorescence emitted by the reporter via fluorescent resonance energy transfers (FRET). FRET is a distance-dependent interaction between two molecules in which the excitation energy is transferred from a photon donor molecule (reporter) to an acceptor molecule (quencher) without emission of a photon. During the extension phase of the PCR cycle, the 5′ exonuclease activity of Taq polymerase cleaves the reporter dye from the probe. Because the reporter dye is no longer in close proximity to the quencher, the FRET is disrupted and the probe begins to fluoresce (Figure 6.8). The intensity of fluorescence can be measured (Chapter 9) and is proportional to the amount of target DNA synthesized during the PCR.

In this assay, the rate of accumulation of amplified DNA over the entire course of a PCR is generated. The greater the initial concentration of target templates in a sample, the fewer cycles required to reach a particular quantity of amplified product. The initial concentration of target templates can be expressed using the *cycle threshold* (C_T). C_T is defined as the number of PCR cycles required for the fluorescent signal to cross a threshold of amplification where the signal exceeds background level or baseline noise. A plot of C_T against the $\log_{10}$ of the initial concentration of a set of DNA standards yields a straight line as a standard curve (Figure 6.9). The target sequences in an unknown sample can be quantified by comparing to the standard curve. Additionally, qPCR has the ability to detect PCR inhibitors (Chapter 7) that interact with DNA or with DNA polymerases. The presence of PCR inhibitors in the DNA extracts can be measured by monitoring the amplification of the internal positive control (IPC). Most human DNA quantitation kits contain a known amount of exogenous DNA as IPC that can be fortified to the sample and amplified. Monitoring the amplification of IPC enables the detection of PCR failure due to inhibition when the IPC's C_T value is higher than that of an uninhibited PCR reaction.

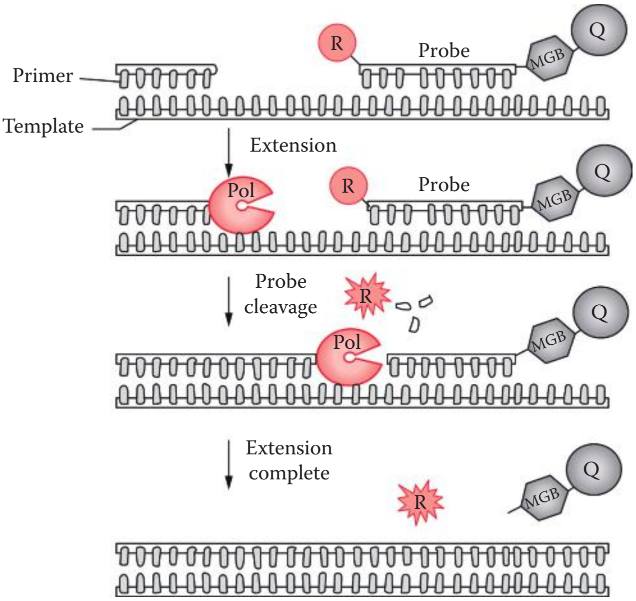

Figure 6.8 Real-time PCR with TaqMan detection. The TaqMan probe is shown. During extension, the 5′ nuclease activity of Taq polymerase (Pol) cleaves the probe. Reporter dye is released during each cycle of PCR. (© Richard C. Li.)

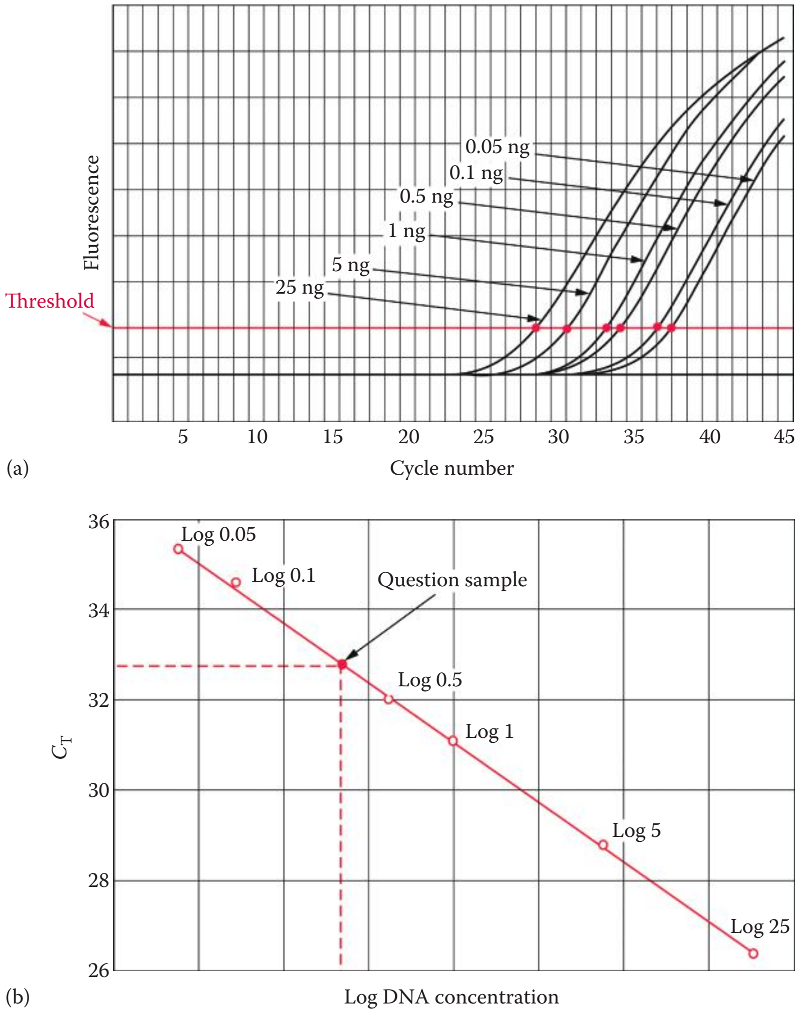

Figure 6.9 Real-time quantitative PCR. (a) Amplification curves for a dilution series of standards with known quantities of DNA. C_T is the cycle threshold at which the amplification curve crosses the threshold, as indicated by the red line. (b) A standard curve based on data obtained from the amplification curves. The quantity of DNA in a questioned sample can be determined from the standard curve. (© Richard C. Li.)

Bibliography

Ahn, S.J., J. Costa, and J.R. Emanuel, PicoGreen quantitation of DNA: Effective evaluation of samples pre- or post-PCR. *Nucleic Acids Res*, 1996, **24**(13): 2623–2625.

Alonso, A. and P. Martin, A real-time PCR protocol to determine the number of amelogenin (X-Y) gene copies from forensic DNA samples. *Methods Mol Biol*, 2005, **297**: 31–44.

Alonso, A., et al., Specific quantification of human genomes from low copy number DNA samples in forensic and ancient DNA studies. *Croat Med J*, 2003, **44**(3): 273–280.

Alonso, A., et al., Real-time PCR designs to estimate nuclear and mitochondrial DNA copy number in forensic and ancient DNA studies. *Forensic Sci Int*, 2004, **139**(2–3): 141–149.

Andreasson, H. and M. Allen, Rapid quantification and sex determination of forensic evidence materials. *J Forensic Sci*, 2003, **48**(6): 1280–1287.

7

Amplification by Polymerase Chain Reaction

7.1 Denaturation and Renaturation of DNA

The DNA double helix is stabilized by chemical interactions. Base pairing of the two strands involves the formation of hydrogen bonds that provide weak electrostatic attractions between electronegative atoms. An adenine always pairs with a thymine (two hydrogen bonds), and a cytosine pairs with a guanine (three hydrogen bonds) (see Figure 7.1). Base stacking involves hydrophobic interactions between adjacent base pairs and provides stability to the double helix.

Double-stranded DNA is maintained by hydrogen bonding between the bases of complementary pairs. *Denaturation* occurs when the hydrogen bonds of DNA are disrupted, and the strands are separated. A *melting curve* can be obtained from measuring DNA denaturation by slowly heating a solution of DNA. As shown in Figure 7.2, an increasing temperature increases the percentage of the DNA that is denatured. The temperature at which 50% of DNA strands are denatured is defined as the *melting temperature* (T_m). The value of T_m is affected by the salt concentration of the solution, but can also be affected by nucleotide content, high pH, and length of the molecule.

Nucleotide content affects the value of T_m because GC pairs are joined by three hydrogen bonds while the AT pairs are joined by only two. Increasing the GC content of a DNA molecule increases the T_m. Excessively high pH causes the hydrogen bonds to break and the paired strands to separate. Finally, the length of the molecule also affects the T_m simply because a longer molecule of DNA requires more energy to break more bonds than a shorter molecule.

The single strands in a solution of denatured DNA can, under certain conditions, reanneal into double-stranded DNA. The process is called *renaturation* and two requirements must be met for it to occur. First, sufficient amounts of charged molecules, such as salts, must be present in the solution to neutralize the negative charges of the phosphate groups in DNA. This prevents the complementary strands from repelling each other. Additionally, the temperature must be high enough to disrupt hydrogen bonds that formed randomly between the bases of DNA strands. However, excessively elevated temperatures can disrupt the base pairs between the complementary DNA strands.

7.2 Basic Principles of Polymerase Chain Reaction

The *polymerase chain reaction* (*PCR*) allows the exponential amplification of specific sequences of DNA to yield sufficient amplified products, also known as *amplicons*, for various downstream

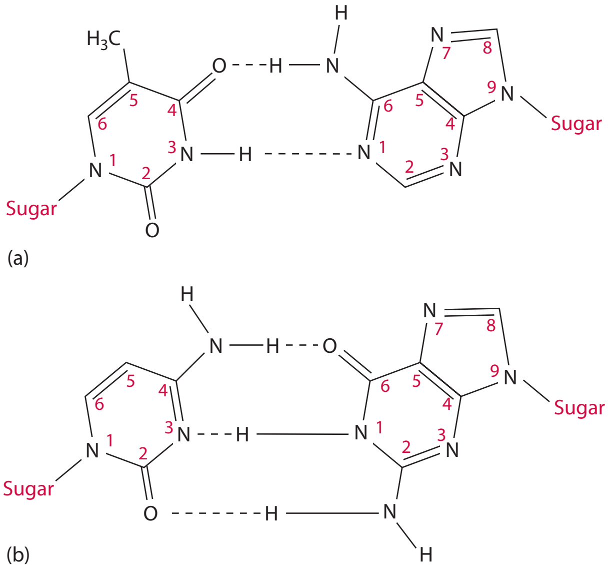

Figure 7.1 Base pairing between two DNA strands. (a) An adenine (right) pairs with a thymine (left). (b) A cytosine (left) pairs with a guanine (right).

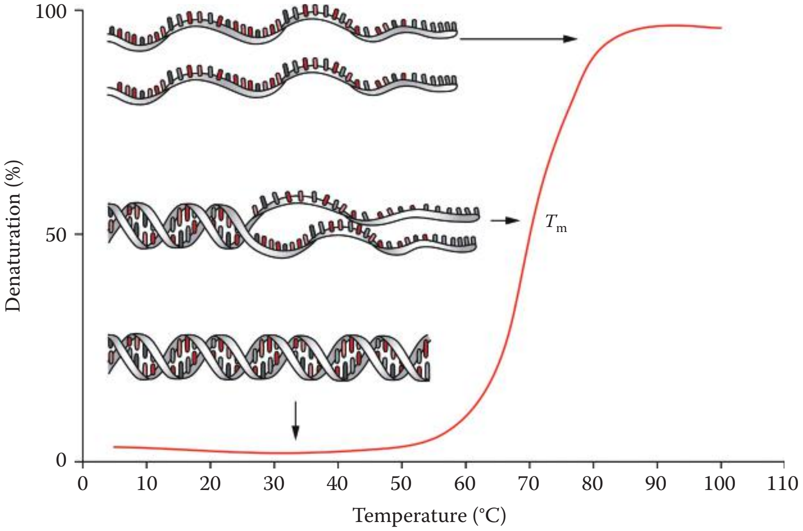

Figure 7.2 Melting curve of DNA. The degree of DNA denaturation is increased by increasing the temperature. T_m and possible shapes of a DNA molecule are shown. (© Richard C. Li.)

applications. The technique is highly sensitive and can amplify very small quantities of DNA. Therefore, it can be utilized for the analysis of samples of limited quantity. PCR-based assays are rapid and robust. Thus, PCR forms the basis of many forensic DNA assays such as DNA quantitation (Chapter 6), short tandem repeat (STR) profiling (Chapter 20), and mitochondrial DNA (mtDNA) sequencing (Chapter 23).

The concept of synthesizing DNA by a cycling process was first proposed in the early 1970s. In the mid-1980s, PCR technology was finally developed by Kary Mullis and his coworkers (Cetus Corporation) to amplify the β-globin gene for the diagnosis of sickle-cell anemia. In the late 1980s, a

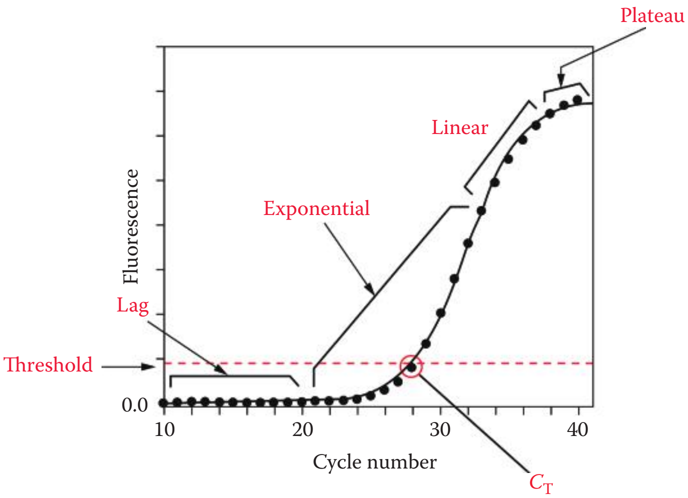

Figure 7.3 PCR amplification curve. A fluorescence signal results from the accumulation of amplicon. The location of the signal threshold is indicated as a dashed line. The C_T value for the curve is the cycle at which the amplification curve crosses the threshold. (© Richard C. Li.)

thermostable polymerase from *Thermus aquaticus* was utilized for PCR. This step greatly increased efficiency and allowed the process to be automated. The result was a powerful impact on molecular biology. In 1993, Mullis was awarded the Nobel Prize for the invention of PCR technology.

In the early 1990s, the technique for the simultaneous amplification and detection of the accumulation of amplicons at each PCR cycle was developed, and the concept of *real-time PCR* was born. This process allows the monitoring of amplicon production at each cycle of the PCR process. The fundamental processes were studied by characterizing the amplification kinetics of PCR using a graph that plotted the amount of amplicon yield at each cycle versus the cycle number (Figure 7.3).

An S-shaped amplification curve is obtained and divided into an exponential phase, a linear phase, and a plateau. During the exponential phase, the amplicon accumulates exponentially. It was revealed that the amplicon accumulation during PCR was correlated to the starting copy number of DNA template. Thus, the amount of amplicon produced during the exponential amplification phase can be used to determine the amount of starting material. This relationship can be further examined using a plot of cycle numbers versus a log scale of the serial dilution of the starting concentration of DNA template, which results in a linear relationship (see Chapter 6). It demonstrates that fewer cycles are needed if larger quantities of starting DNA template are present. The slope of this linear curve (Figure 6.9b) is known as the *amplification efficiency*. The exponential phase continues until one or more of the components (Section 7.3) in the reaction become limited. At this point, the amplification efficiency decreases, the amplicon no longer accumulates exponentially, and PCR enters the linear phase of the curve. At the plateau phase, no more amplicon is accumulated due to the exhaustion of reagents and polymerase.

7.3 Essential PCR Components

A PCR reaction requires thermostable DNA polymerases, primers, and other components, as described below (Figure 7.4).

7.3.1 Thermostable DNA Polymerases

A wide variety of DNA polymerases are available. They vary in fidelity, efficiency, and ability to synthesize longer DNA fragments. Nonetheless, *Taq* polymerase is the most commonly used

HAMILTON
microLAB
NIMBUS
NIMBUS PCR
Automated Assay Setup

Figure 7.4 A liquid handling workstation for automated PCR assay setup. (© Richard C. Li.)

enzyme for routine PCR applications (0.5–5 units per reaction). Currently, AmpliTaq Gold™ DNA polymerase (Applied Biosystems) is the most common DNA polymerase for forensic applications.

PCR reactions are usually set up at room temperature. Nonspecific annealing between primers and template DNA can occur, resulting in the formation of nonspecific amplicons. Additionally, annealing between the primers can occur to form primer dimers. Nonspecific annealing interferes with PCR amplification by reducing the amplification efficiency of the specific sequences of interest.

Such interference can be minimized by a *hot-start PCR* approach. The AmpliTaq Gold™ DNA polymerase, a modified enzyme, remains in an inactive form until activated with a pH below 7 prior to the PCR cycling in which the inhibitory motif is inactivated. The pH of the buffer system used in the PCR reaction is temperature sensitive; increasing the temperature decreases the pH of the solution. Thus, the activation of the enzyme can actually be carried out by a heating step at 95°C prior to the start of the cycling. During the heating process, the DNA strands also denature, which can prevent the formation of nonspecific PCR products.

7.3.2 PCR Primers

PCR primers are the oligonucleotides that are complementary to the sequences that flank the target region of the template. A pair (forward and reverse) of primers (typically 0.1–1 μM) is required. Properly designed primers are critical to the success of a PCR reaction. Computer software such as Primer3 is available to assist and optimize the designing of primers.

A primer must be specific to the target sequence; otherwise, nonspecific products that might interfere with the proper interpretation of a DNA profile might be produced. The primers within a pair should have similar T_m values. The estimated T_m values of a primer pair should not differ by more than 5°C. The T_m of an oligonucleotide primer can be predicted and calculated using the following equation:

$$T_m = 81.5°\text{C} + 16.6\left(\log_{10}[K^+]\right) + 0.41\left(\%\ [\text{G}+\text{C}]\right) - \left(\frac{675}{n}\right)$$

where:

$[K^+]$ = concentration of the potassium ion
$[\text{G}+\text{C}]$ = GC content (%) of the oligonucleotide
n = number of bases in the oligonucleotide

This equation shows that the T_m can be affected by primer base composition (GC content) and primer length. The GC content of an oligonucleotide primer should be 40%–60%, and the length of an oligonucleotide primer should be 15–25 base pairs, although longer primers can be used.

A primer should not contain self-complementary sequences that may form hairpin structures interfering with the annealing of primers and the template. Additionally, the primers in a pair should not share similar sequences, to avoid the primers from annealing to each other. These annealed primers may then be amplified during PCR, creating products known as *primer dimers*, which compete with the target DNA template for PCR components.

Multiplex PCR is the simultaneous amplification of more than one region of a DNA template in a single reaction to achieve high-throughput analysis. Multiplex PCR consists of multiple primer sets in a single reaction to produce amplicons of multiple target DNA regions. The primers should be designed to yield proper sizes of amplicons to be resolvable in downstream separation and detection procedures such as electrophoresis. To prevent preferential amplification of one target sequence over another, the annealing temperatures should be similar among multiplex PCR primer pairs. Additionally, the primers should lead to similar amplification efficiencies among the loci to be tested. Forensic applications of multiplex PCR, such as autosomal and Y-chromosomal short tandem repeat analysis, are discussed in Chapters 20 and 21.

7.3.3 Other Components

Essential components include template DNA with target sequences in either linear form (nuclear genomic DNA) or circular form (mitochondrial DNA). Both single- and double-stranded DNA can be used as a template for PCR. Typically, 1–2.5 ng of template DNA is utilized for forensic applications using PCR.

Deoxynucleoside triphosphates (dNTPs) are the substrates for DNA synthesis. A PCR assay usually contains equal molar amounts (typically, 200 μM) of dATP, dCTP, dGTP, and dTTP.

Divalent cations, such as Mg^{2+}, are required for the enzymatic activity of DNA polymerases. A PCR assay usually contains 1.5–2.5 mM Mg^{2+}. Monovalent cations, such as K^+ (50 mM), are usually recommended, and a buffer is often utilized to maintain pH between 8.3 and 8.8 at room temperature.

Controls should be used to monitor the effectiveness of PCR amplification. A *positive control* shows that PCR components such as reagents and PCR cycle parameters are working properly during a PCR. A standard DNA template should be used as a positive control and amplified with the same PCR components used on the rest of the samples. The amplification negative control and extraction reagent blank are discussed in Section 7.5.4.

7.4 Cycle Parameters

PCR cycling protocols may vary according to the type of analysis. Figure 7.5 shows representative PCR cycling conditions commonly used by forensic DNA laboratories. PCR utilizes a number of cycles for the replication of a specific region of a DNA template. During each cycle, a copy of the target DNA sequence is synthesized. A PCR cycle consists of three elements: denaturation, annealing, and extension. Precise and accurate temperatures at denaturation, annealing, and extension are critical to achieve a successful amplification. At the beginning of each cycle, the two complementary DNA template strands are separated at high temperatures (94°C–95°C) in a process called *denaturation*. The temperature is then decreased to allow *annealing* between the oligonucleotide primers and the template. The temperature for annealing is usually 3°C–5°C lower than the T_m of the oligonucleotide primer.

The annealing temperature is critical. If it is too high, a very low quantity of amplicon is yielded because of the failure of annealing between the primer and the template. If the annealing temperature is too low, nonspecific amplification can occur. Next, optimal temperature for DNA polymerase is reached, thus allowing for DNA replication (*extension*). By the end of each cycle, the copy number of the amplicon is nearly doubled (Figure 7.6).

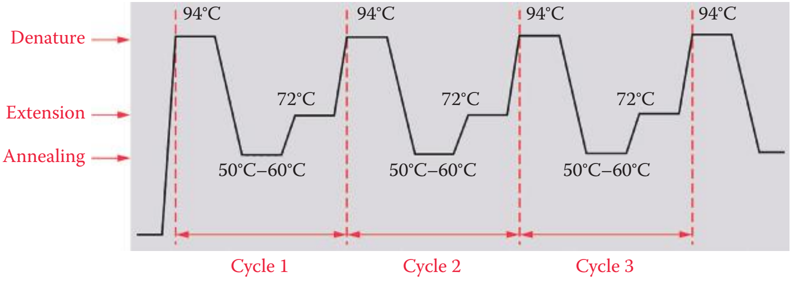

Figure 7.5 Temperature parameters during thermal cycling of the PCR process. The first three cycles are shown. (© Richard C. Li.)

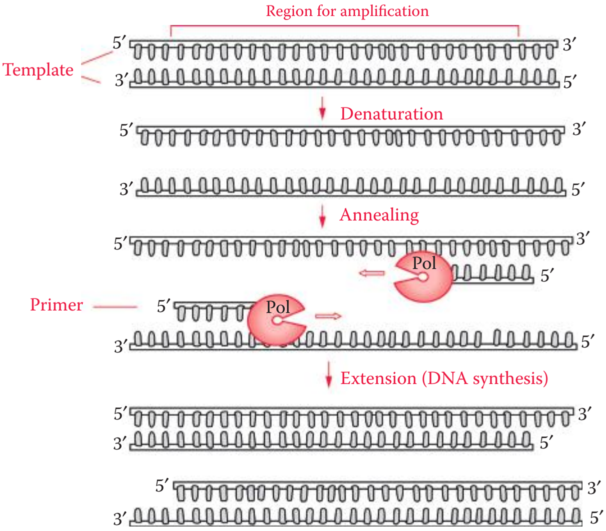

Figure 7.6 First cycle of PCR. The amplified region is determined by the positions of the primers. The direction (5′ to 3′) of DNA synthesis is indicated by arrows. Pol represents *Taq* polymerase. (© Richard C. Li.)

The number of cycles needed for PCR depends primarily on the number of copies of starting DNA template. The relationship can be expressed as the following equation:

$$N_x = N_0\left(1 + E\right)^x$$

where:

X = number of PCR cycles
N_x = copy number of the amplicon after x cycles of PCR
N_0 = initial copy number of the template
E = amplification efficiency of the *Taq* polymerase

For example, in a 28-cycle PCR amplification, the DNA template can theoretically be amplified by a factor of approximately 10^8. If the cycle number is increased to 34, a factor of 10^{10} can theoretically be reached.

PCR amplification can be carried out using an instrument known as a *thermal cycler* (Figure 7.7). Various types of thermal cyclers differ in the number of samples they can process, the sizes of the sample tubes, and temperature control features.

Figure 7.7 Thermal cyclers. A thermal cycler can provide rapid temperature changes as desired to carry out PCR. A PCR thermal cycler (left) and a real-time PCR instrument (right). (© Richard C. Li.)

7.5 Factors Affecting PCR

7.5.1 Template Quality

It is important to prevent degradation of the DNA during the collection and processing of evidence. Degradation causes DNA to break into smaller fragments. If the damage occurs at a region to be amplified, the result can be failure in PCR amplification. In a degraded sample, the longer the amplicon length, the higher the risk of failure in PCR. Low copy number (LCN) of DNA template is often encountered in forensic samples. When amplifying very low levels (approximately 100 pg of DNA) of a template, the following phenomenon is often observed: one of the two alleles fails to be detected from a heterozygote and can falsely be identified as a homozygote. This phenomenon is also known as the *stochastic effect* in which the two alleles in a heterozygous individual are unequally detected at a low level of starting DNA template. Approaches such as increasing the cycle number, from 28 to 34 (Section 7.4), have been introduced to address the LCN problem.

7.5.2 Inhibitors

Inhibitors, if present, can interact with the DNA template or polymerase, causing PCR amplification failure. The presence of PCR inhibitors can be detected using an internal positive control (Section 6.3.1.1). A number of PCR inhibitors commonly encountered in evidence samples include heme molecules from blood, indigo dyes from fabrics, and melanin from hair samples. Thus, it is important to remove PCR inhibitors during DNA extraction. If PCR inhibitors are not eliminated during the extraction process, additional procedures such as the use of centrifugal filtration devices can be used. Centrifugal filtration devices can separate molecules by size. After the centrifugation step, small molecular weight inhibitors are filtered by passing through the membrane and are discarded. Alternatively, increasing the amount of DNA polymerase or adding bovine serum albumin (BSA) in the reaction can overcome the inhibition effects.

7.5.3 Contamination

PCR is a highly sensitive method; therefore, procedures that minimize the risk of contamination are necessary. To prevent contamination, pre- and post-PCR samples should be processed in separate areas or at different times. Additionally, reagents, supplies, and equipment used for pre- and post-PCR steps should be separated as well. Protective gear should include laboratory coats and disposable gloves. Facial masks and hair caps may be used if necessary. Aerosol-resistant pipet tips and DNA-free solutions and test tubes should also be used.

The levels of contamination must be monitored using controls. Extraction *reagent blanks*, which contain all extraction reagents but no sample, monitor contamination from extraction to PCR. Contamination detected in an extraction reagent blank but not in an amplification-negative control indicates that the reagents used for extraction are contaminated. Amplification-*negative* *controls*, which contain all PCR reagents and no DNA template, monitor contamination of the amplification. Contamination observed in an amplification-negative control but not in an extraction reagent blank indicates that the contamination occurred during the amplification step. Contamination observed in both an amplification reagent blank and an extraction-negative control indicates that amplification reagents are contaminated. A collection of DNA profiles of each member of a laboratory should be readily available for comparisons. Sources of laboratory contamination can be identified by comparing results with an analyst's DNA profile.

7.6 Reverse Transcriptase PCR for RNA-Based Assays

The pathway for the flow of genetic information is called the *central dogma*—a term coined by Francis Crick in 1956. According to the central dogma, parental DNA serves as the template for *DNA replication*. With RNA synthesis or *transcription*, the process is carried out using the DNA as a template. Conversely, RNA chains can be used as templates for the synthesis of a DNA strand of complementary sequence, in which the end product is referred to as *complementary DNA (cDNA)*. Protein synthesis, also known as *translation*, is directed by an RNA template (Figure 7.8).

The flow of genetic information from RNA to DNA is referred to as *reverse transcription*. It was discovered independently by David Baltimore and Howard Temin in 1970, and they shared the Nobel Prize for their work. Reverse transcription is carried out by a *reverse transcriptase* that forms the basis of reverse transcriptase PCR as described below.

Reverse transcriptase PCR (RT-PCR) is highly sensitive and can be used to detect very small quantities of mRNA. It can be utilized to measure levels of gene expression even when the RNA of interest is expressed at very low levels. Detecting mRNAs of tissue-specific genes can be utilized for bodily fluid identification in forensic investigations (Chapter 11). During an RT-PCR process, a single-stranded cDNA is synthesized from a template mRNA using reverse transcription. The cDNA is then amplified by PCR for detection and analysis.

7.6.1 Reverse Transcription

The synthesis of single-stranded cDNA from an mRNA template is catalyzed by reverse transcriptase. Reverse transcriptases share many features in structure and function with DNA polymerases. The catalytic function of cDNA synthesis requires a primer that anneals to a complementary mRNA template. The primer can be either RNA or DNA; however, DNA primers are more efficient than RNA primers. During the elongation of the primer, reverse transcriptase incorporates the corresponding deoxyribonucleotide triphosphate according to the rules of base pairing with the RNA template. The RNA template is then degraded by an intrinsic RNase H activity of reverse transcriptase during the reverse transcription reaction. The retroviral RNase H is a domain of the viral reverse transcriptase enzyme. It is a nonspecific endonuclease that cleaves RNA.

Several reverse transcriptases derived from retroviruses can be used to generate cDNA from an RNA template. The most common reverse transcriptases used for cDNA synthesis are

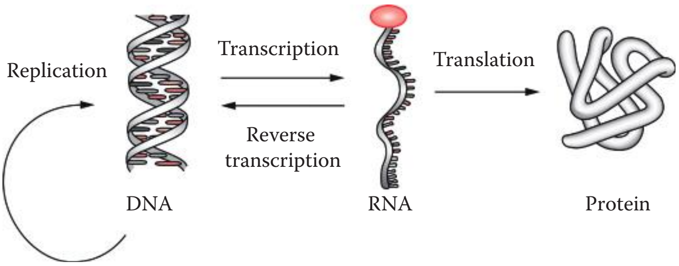

Figure 7.8 Pathway for the flow of genetic information. (© Richard C. Li.)

encoded by the *pol* gene from the avian myeloblastosis virus (AMV) and the Moloney strain of the murine leukemia virus (MMLV). These enzymes are genetically engineered reverse transcriptases but lack RNase H activity. Genetically engineered reverse transcriptases that are stable at higher temperatures (up to 60°C) are also produced. Increased thermal stability of reverse transcriptase allows the reverse transcription to be carried out at a higher than ambient temperature, which eliminates RNA secondary structures and improves the specificity of the reaction and the yields of the synthesis of full-length cDNA.

7.6.2 Oligodeoxynucleotide Priming

Oligodeoxynucleotide primers are an essential component for a reverse transcriptase reaction. Different priming strategies can be utilized to synthesize cDNA from a particular target mRNA or from all mRNA in a sample. *Gene-specific primers* are designed to hybridize to a particular mRNA sequence for the conversion of a specific gene sequence into cDNA. *Universal primers* can hybridize to any mRNA sequence in a sample to convert all mRNAs to cDNA. Two types of universal primers can be used: oligo (dT) and random hexamer primers. An *oligo (dT) primer* can hybridize to the 3′ termini poly (A) tails of eukaryotic mRNAs. Reverse transcriptases with oligo (dT) only synthesize cDNA from transcribed genes. *Random hexamer primers* are another type of universal primer. Random hexamers are nonspecific primers that can hybridize, at multiple sites, to any RNA sequence including non-mRNA templates such as ribosomal RNA.

7.6.3 Reverse Transcriptase PCR

During an RT-PCR process, a single-stranded cDNA is synthesized from a template mRNA using reverse transcription. The cDNA is then amplified by PCR with a pair of oligonucleotide primers corresponding to a specific sequence in the cDNA (Figure 7.9).

Two strategies of RT-PCR exist: a one-step and a two-step RT-PCR (Figure 7.10). One-step RT-PCR combines the reverse transcription reaction and PCR in a single tube. Only gene-specific

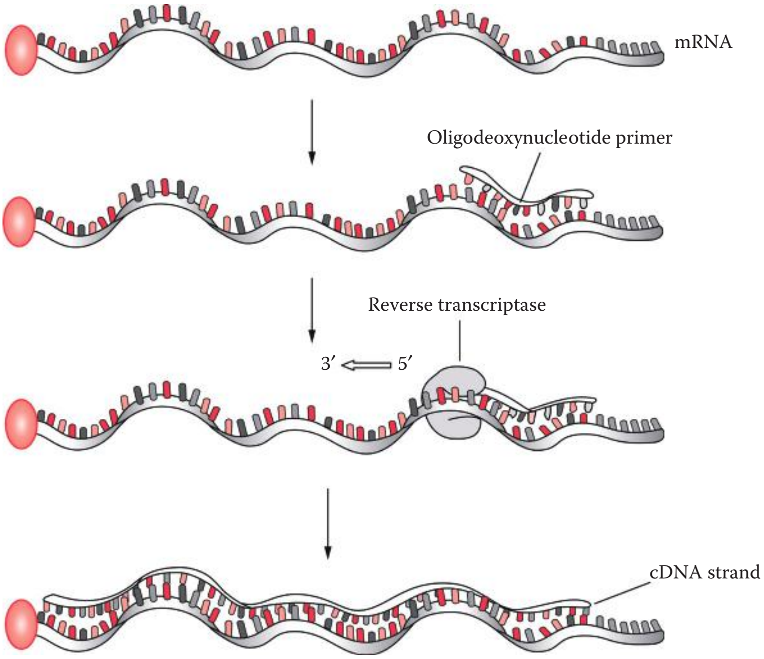

Figure 7.9 cDNA synthesis. The synthesis of a DNA strand transcribed from mRNA can be carried out using a primer and reverse transcriptase. A gene-specific primer that amplifies a specific target sequence is shown. (© Richard C. Li.)

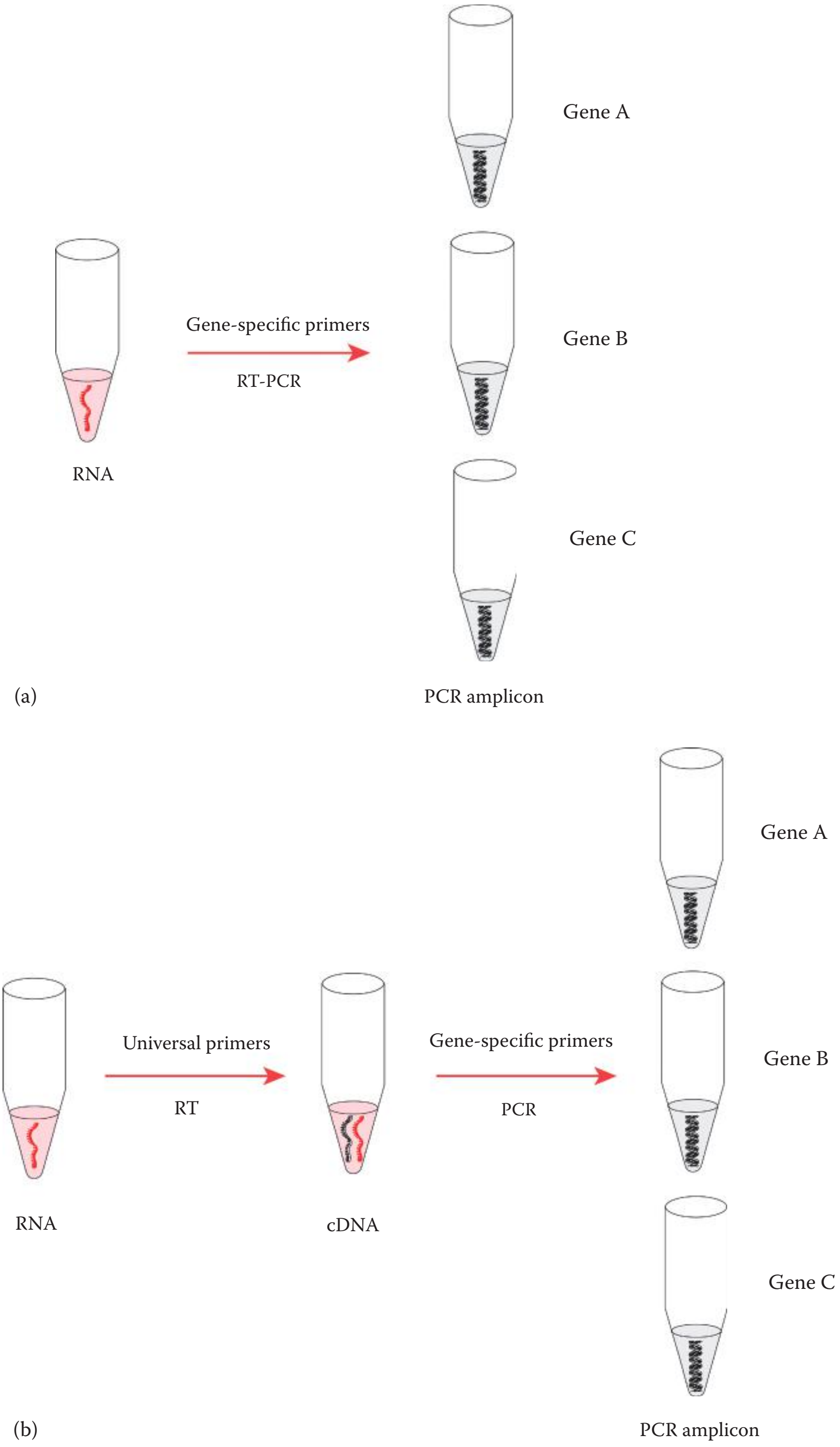
Gene A
Gene-specific primers
RT-PCR
Gene B
RNA
Gene C
(a)
PCR amplicon
Gene A
Universal primers
RT
Gene-specific primers
PCR
Gene B
RNA
cDNA
Gene C
(b)
PCR amplicon

Figure 7.10 RT-PCR strategies. (a) One-step RT-PCR and (b) two-step RT-PCR. (© Richard C. Li.)

primers can be used for the reverse transcription reaction and PCR. RNAs from either total RNA or mRNA can be used. This approach simplifies the reaction setup as it is useful for processing large numbers of samples. The tubes are not opened between reverse transcription and PCR, thus minimizing the risk of pipetting errors and carryover contamination.

During a two-step PCR, the reverse transcription reaction and PCR are carried out in separate tubes. Following the reverse transcription reaction, the cDNA is transferred to a separate tube for the PCR amplification. Either oligo (dT) or random hexamer primers can be used for the reverse transcription reaction. Total RNA or polyadenylated RNA can be used for the reverse transcriptase reaction. This allows for the ability to convert all the messages in an RNA sample into cDNA. The two-step RT-PCR is useful for analyzing multiple mRNAs from a single sample.

Two types of PCR methods can be utilized for analyzing amplified products. The end-point PCR method measures the amount of amplified product synthesized during PCR at the end of the PCR amplification. The detection of the amplified product indicates the presence of the mRNA of interest. With the real-time PCR method, the amplified product is quantified during the exponential phase of PCR. The hot-start PCR strategy is typically used to increase sensitivity, specificity, and yield. Usually, the hot-start strategy is carried out at a high temperature, for example 94°C, prior to PCR cycling. Under this condition, *Taq* DNA polymerase is activated. Additionally, the RNA-cDNA hybrid is denatured, and reverse transcriptase is inactivated at 94°C. Multiplex RT-PCR assays have been developed that can detect multiple bodily fluid mRNAs from single or mixed stains (Chapter 11). The commonly used RT-PCR for forensic identification of bodily fluids is the two-step RT-PCR. The end-point PCR is usually utilized with the forward primers labeled with florescent dyes at the 5′ end. The amplified products are separated and detected in a standard capillary electrophoresis instrument.

Bibliography

Aboud, M., H.H. Oh, and B. McCord, Rapid direct PCR for forensic genotyping in under 25 min. *Electrophoresis*, 2013, **34**(11): 1539–1547.

Afonina, I.A., et al., Minor groove binder-conjugated DNA probes for quantitative DNA detection by hybridization-triggered fluorescence. *Biotechniques*, 2002, **32**(4): 940–944, 946–949.

Alaeddini, R., Forensic implications of PCR inhibition—A review. *Forensic Sci Int Genet*, 2012, **6**(3): 297–305.

Ballantyne, K.N., R.A. van Oorschot, and R.J. Mitchell, Comparison of two whole genome amplification methods for STR genotyping of LCN and degraded DNA samples. *Forensic Sci Int*, 2007, **166**(1): 35–41.

Ballantyne, K.N., R.A. van Oorschot, and R.J. Mitchell, Increased amplification success from forensic samples with locked nucleic acids. *Forensic Sci Int Genet*, 2011, **5**(4): 276–280.

Ballantyne, K.N., R.A. van Oorschot, and R.J. Mitchell, Locked nucleic acids in PCR primers increase sensitivity and performance. *Genomics*, 2008, **91**(3): 301–305.

Ballantyne, K.N., et al., Decreasing amplification bias associated with multiple displacement amplification and short tandem repeat genotyping. *Anal Biochem*, 2007, **368**(2): 222–229.

Bassam, B.J., et al., Nucleic acid sequence detection systems: Revolutionary automation for monitoring and reporting PCR products. *Aust Biotechnol*, 1996, **6**: 285–294.

Bekaert, B., et al., Automating a combined composite-consensus method to generate DNA profiles from low and high template mixture samples. *Forensic Sci Int Genet*, 2012, **6**(5): 588–593.

Birch, D.E., et al., Simplified hot start PCR. *Nature*, 1996, **318**: 2.

Bloch, W., A biochemical perspective of the polymerase chain reaction. *Biochemistry*, 1991, **30**(11): 2735–2747.

Bright, J.A., et al., The effect of cleaning agents on the ability to obtain DNA profiles using the Identifiler and PowerPlex® Y multiplex kits. *J Forensic Sci*, 2011, **56**(1): 181–185.

Brotherton, P., et al., Preferential access to genetic information from endogenous hominin ancient DNA and accurate quantitative SNP-typing via SPEX. *Nucleic Acids Res*, 2010, **38**(2): e7.

Budowle, B., A.J. Eisenberg, and A. van Daal, Validity of low copy number typing and applications to forensic science. *Croat Med J*, 2009, **50**(3): 207–217.

8

DNA Electrophoresis

It is necessary to separate various sizes of DNA fragments so that the DNA fragment in question can be identified and analyzed. This can be achieved by electrophoresis, a process in which fragments are separated based on the migration of charged macromolecules in an electric field.

8.1 Basic Principles

DNA is a negatively charged molecule in an aqueous environment, with the phosphate groups of DNA nucleotides carrying negative charges. DNA molecules migrate from the negative electrode (cathode) toward the positive electrode (anode) in an electric field during electrophoresis. The *electrical potential*, a measure of the work required to move a charged molecule in an electric field, is the force responsible for moving the charged macromolecules during electrophoresis.

The *electrophoretic mobility* of macromolecules is primarily determined by their *charge-to-mass ratios* and their shapes. However, because the phosphate group of every DNA nucleotide carries a negative charge, the charge-to-mass ratio of DNA molecules is almost the same even if the length of the DNA fragment varies. Additionally, forensic DNA testing usually analyzes linear DNA molecules such as double-stranded linear DNA for variable number tandem repeat (VNTR) analysis (Chapter 19) and single-stranded DNA for short tandem repeat (STR) analysis (Chapter 20). The shapes of linear DNA molecules are similar. Therefore, the electrophoretic separation of different-sized DNA fragments, through a series of pores of the supporting matrix in which the fragments travel, is based more on their sizes than their shapes.

It is obviously easier for smaller molecules to migrate through the pores of a supporting matrix than larger molecules; this is why smaller molecules migrate faster through the matrix. Hence, the electrophoretic mobility increases as the size of the DNA molecule decreases. Conversely, larger DNA molecules migrate much more slowly because they experience more friction and collisions as they travel through the net of pores in a matrix. Therefore, the separation of DNA molecules with different sizes can be accomplished. Figure 8.1 depicts the models for DNA electrophoresis.

8.2 Supporting Matrices

Although the actual separation occurs in an aqueous phase, most variants of electrophoresis use a physical support material also called a *matrix*. As discussed in Section 8.1, the matrix can be used as a molecular sieve for the separation of DNA molecules, and it also reduces diffusion and convection during electrophoresis. Agarose and polyacrylamide are commonly used electrophoresis matrices because of their good reproducibility, reliability, and versatility.

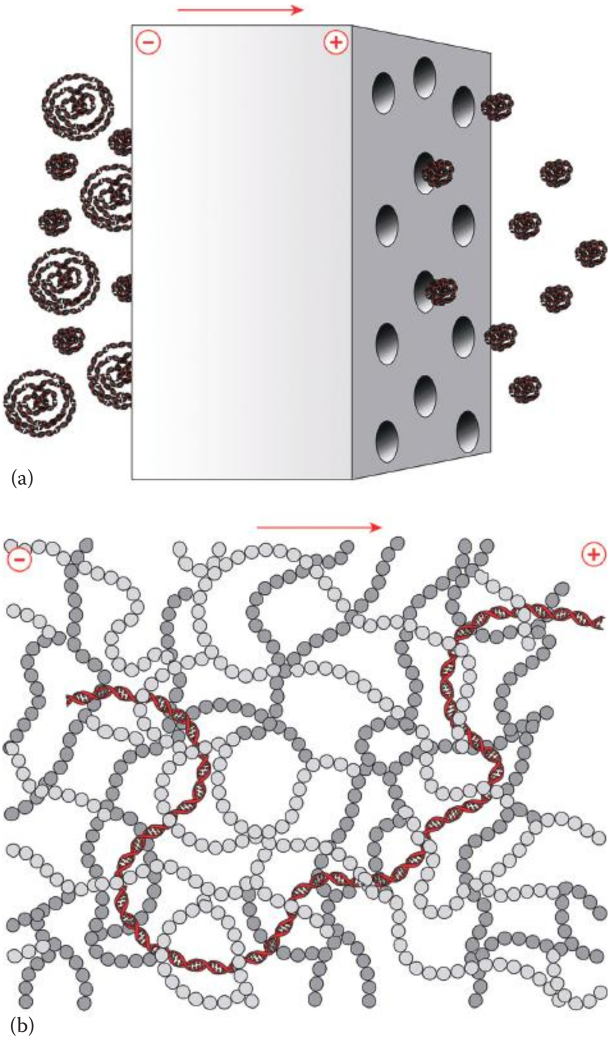

Figure 8.1 Schematic illustrations showing some of the nucleic acid separation mechanisms. (a) Ogston sieving model for the separation of DNA fragments through electrophoresis. Based on this model, the matrix used for gel electrophoresis consists of randomly distributed pores. DNA molecules are considered to behave as globular objects. During the migration, smaller DNA molecules migrate faster than their larger counterparts through the matrix. However, this model does not apply to DNA molecules with a radius much larger than the pore size of the matrix. (b) A reptation model presents an alternative mechanism of gel electrophoretic separation of DNA molecules. Based on this model, DNA molecules are flexible and can migrate through the pores of the gel matrix through a reptation process that is driven by an applied electric field. This model can also apply to capillary electrophoresis separation of DNA molecules in entangled polymer solutions. +, anode; –, cathode; arrow, direction of migration. (© Richard C. Li.)

8.2.1 Agarose

Agarose is a linear polymer composed of alternating residues of D- and L-galactose (Figure 8.2). A gelatinized agarose forms a three-dimensional sieve with pores from 50 to 200 nm in diameter (Figure 8.3). DNA fragments ranging from 50 to 20,000 base pairs in size are best resolved in agarose gels.

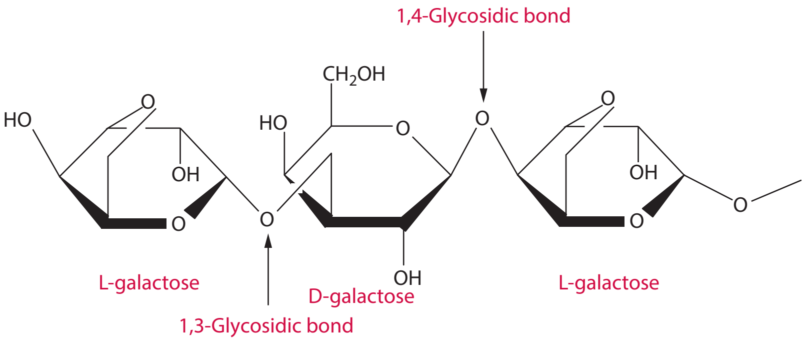

Figure 8.2 Chemical components of agarose, a linear polysaccharide composed of alternating units of D- and L-galactose. Often, the L-galactose residue has an anhydro bridge between the 3 and 6 positions and is called 3,6-anhydro-L-galactose.

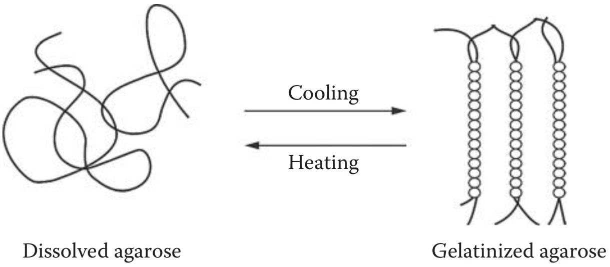

Figure 8.3 Agarose chains. Gelatinized agarose consists of a three-dimensional network in which chains of agarose form helical fibers that aggregate into supercoiled structures. (© Richard C. Li.)

The electrophoretic mobility of double-stranded DNA through agarose gel matrices is inversely proportional to the $\log_{10}$ of the size of DNA fragments ranging from 50 to 20,000 base pairs. The electrophoretic mobility of a linear DNA fragment is also inversely proportional to the concentration of agarose in the gel. A low concentration of agarose in a gel forms larger pores, allowing the separation of larger DNA fragments.

8.2.2 Polyacrylamide

A polyacrylamide gel matrix is very effective for the separation of smaller fragments of DNA (5–500 base pairs). Additionally, a single nucleotide difference in the length of a DNA fragment can be resolved with this type of gel. Polyacrylamide produces much smaller pore sizes than agarose gels and thus has a much higher resolving power than agarose gels for low-molecular-weight DNA molecules. A polyacrylamide gel matrix is formed by polymerization and cross-linking reactions.

8.2.2.1 Polymerization Reaction

Long linear chains of polyacrylamide are polymerized from acrylamide monomers (Figure 8.4). This polymerization reaction is initiated in the presence of free radicals that are generated from the reduction of ammonium persulfate (APS) by *N*,*N*,*N*′,*N*′-tetramethylethylene diamine (TEMED). See Figure 8.5.

8.2.2.2 Cross-Linking Reaction

As shown in Figure 8.5, three-dimensionally cross-linked polyacrylamide chains can be formed with the use of cross-linking agents, such as *N*,*N*′-methylenebisacrylamide (BIS; Figure 8.4).

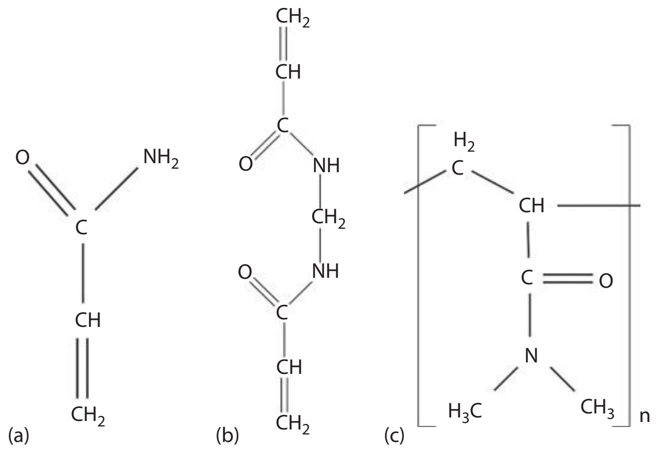
(a)
(b)
(c)
O
NH2
C
CH
CH2
CH2
CH
C
O
NH
CH2
NH
O
C
CH
CH2
H2
C
CH
C
O
N
H3C
CH3
n

Figure 8.4 Chemical structures of (a) acrylamide, (b) *N*,*N*′-methylenebisacrylamide (BIS), and (c) polydimethylacrylamide (a linear polymer for capillary electrophoresis).

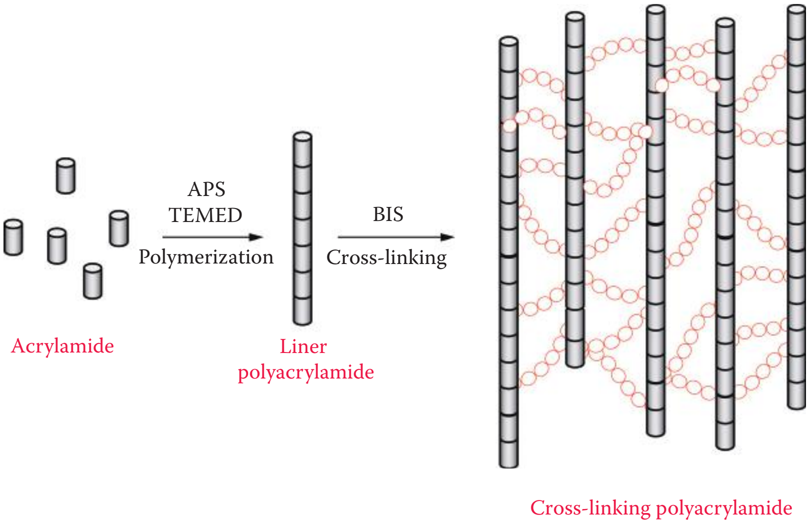
APS
TEMED
Polymerization
BIS
Cross-linking
Acrylamide
Liner polyacrylamide
Cross-linking polyacrylamide

Figure 8.5 Formation of polyacrylamide gel. Polymerization is followed by cross-linking in which the pore size of the gel is determined by its degree of polymerization and cross-linking. (© Richard C. Li.)

The porosity of the resulting gel is determined by the lengths of these chains and the degree of cross-linking between them. Therefore, the sizes of the pores that are formed can be adjusted by altering the concentrations of the acrylamide monomer and the cross-linking reagent.

Polyacrylamide can also be used in a capillary electrophoresis matrix. It is very difficult to insert a cross-linked polyacrylamide matrix into a capillary. Therefore, a solution of a linear polymer (non-cross-linked) of polydimethylacrylamide (Figure 8.4) is used as a matrix for capillary electrophoresis. For instance, POP-4 (4% un-cross-linked polydimethylacrylamide polymer) is used for fragment analysis, such as forensic STR analysis (Chapter 18), in which fluorescently labeled DNA amplicons are separated using electrophoresis and are identified based on their sizes by comparing them with a size standard. POP-6 (6% un-cross-linked polydimethylacrylamide polymer) is used for sequencing, such as mitochondrial DNA (mtDNA) sequencing (Chapter 21). These un-cross-linked polymers are commercially available (Applied Biosystems).

8.2.2.3 Denaturing Polyacrylamide Electrophoresis

Polyacrylamide-based electrophoresis can be carried out with both double- and single-stranded DNA. Electrophoresis that is performed under the conditions of single-stranded DNA is called *denaturing electrophoresis*. Denatured DNA migrates through the gel as linear molecules at a rate independent of the base composition and sequence. With short single-stranded fragments, molecules differing in size by a single nucleotide can be separated. This extraordinary sensitivity to size is extremely useful for the separation of small DNA fragments that are used for DNA sequencing and STR analysis in forensic applications. The denaturing condition can be best achieved by adding chemicals such as urea or formamide to the matrix or by increasing the temperature or pH during electrophoresis.

8.3 Apparatus and Forensic Applications

Agarose electrophoresis is carried out in a slab gel. Agarose gel can be prepared in a variety of shapes and sizes and can be run in different configurations. The most common gel for forensic use is the horizontal slab. Polyacrylamide can be used either in a slab gel or in a capillary electrophoretic apparatus.

8.3.1 Slab Gel Electrophoresis

8.3.1.1 Agarose Gel Electrophoresis

One forensic application of agarose gel electrophoresis is restriction fragment length polymorphism (RFLP) analysis of VNTR loci (see Chapter 19). An agarose gel is used to separate the DNA fragments by size ranging from 500 to 20,000 base pairs of commonly used VNTR loci for forensic testing. This type of electrophoresis is done under nondenatured conditions (double-stranded DNA).

Figure 8.6 shows an example of the apparatus used for gel electrophoresis of DNA. A slab (3–4 mm) of agarose gel is prepared by allowing liquid agarose to gelatinize in a cast. The gel is submerged in a buffer tank filled with an electrophoretic buffer with proper ionic strength, which is necessary to achieve efficient electrical conductance. The gel contains small wells for loading samples. The DNA samples are mixed with a gel loading buffer prior to loading them into the gel. The gel loading buffer contains dyes that add color to the sample to facilitate the process of loading. Dyes such as bromophenol blue and xylene cyanol FF migrate toward the anode during electrophoresis. In most forensic applications, these dyes migrate faster than DNA fragments. The dyes are visible and can thus be used to track the progression of electrophoresis. The electrophoresis can be stopped as the dye front reaches the bottom of the gel (anode side).

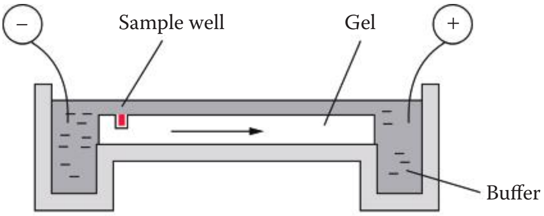

Figure 8.6 Horizontal agarose slab gel apparatus. The gel can be prepared by heating an agarose suspension to dissolve it. The agarose solution can then be poured into a gel cast and allowed to cool until gelatinized. An electrophoresis buffer is poured over the gel to submerge it. The samples, usually mixed with loading buffer containing dye to allow visualization, are loaded into the wells of the submerged gel using a pipet. The electrodes are indicated. (© Richard C. Li.)

The samples are loaded into the sample wells situated proximal to the cathode. An electric field is applied and the negatively charged DNA molecules migrate through the agarose toward the anode. Slab gel electrophoresis is capable of running multiple samples simultaneously. The sample is made visible by using a fluorescent intercalating reagent that can make DNA fluoresce under ultraviolet (UV) light (Chapter 9), either using the fluorescent intercalating reagent incorporated into the gel or staining the gel after electrophoresis. The separated DNA samples appear as bands and the sizes of the DNA fragments can be estimated by comparing them with the sizes of standards run concurrently.

8.3.1.2 Polyacrylamide Gel Electrophoresis

The forensic application of this apparatus is the separation of STR fragments and DNA sequencing reaction products of mtDNA. The sizes of DNA fragments that can be separated range from 100 to 500 base pairs—much smaller than what can be separated efficiently with agarose gels. Single-nucleotide resolution to distinguish similarly sized fragments can be achieved with this technique under denatured conditions with only single-stranded DNA. Polyacrylamide thin slab gels (0.75–1.5 mm) are usually used. Samples are loaded into wells flanked by two pieces of glass. Therefore, polyacrylamide gels are usually run in a vertical configuration (Figure 8.7). The detection of DNA bands in polyacrylamide gels will be described in Chapter 9. The sizes of DNA fragments can be calculated by including an internal size standard. As with agarose gel, polyacrylamide gel electrophoresis is capable of running multiple samples simultaneously and high throughput (a measurement of the rate that a sample is processed by a given analysis) can be achieved. However, caution should be taken to ensure that cross-contamination does not occur from the spilling of samples from adjacent wells. Additionally, polyacrylamide gels are more difficult to prepare than agarose gels.

8.3.2 Capillary Electrophoresis

Capillary electrophoresis (Figures 8.8 through 8.11) is a newer method than the slab gel method, and can be utilized to separate charged macromolecules, such as DNA, RNA, polysaccharides, and proteins.

One essential component of the capillary electrophoresis instrument is the capillary, a thin hollow tube made of fused silica (which typically has a diameter between 50 and 100 μm and a length between 10 and 50 cm). The capillary contains a translucent detection window for the instrument to detect signals from the labeled DNA fragments during electrophoresis.

Linear polydimethylacrylamide is used as the matrix. Capillary electrophoresis is conducted under denatured conditions for forensic applications such as STR analysis and mtDNA sequencing analysis. The denatured condition is achieved by including urea in the electrophoresis matrix and formamide during sample preparation. The injection of samples into the capillary is performed by an autosampler using an *electrokinetic* mechanism (an injection based on the charge of molecules). Only small quantities of sample are required for each injection, and any remaining sample can be saved in case an analysis needs to be repeated.

During electrophoresis, the capillary is connected to buffer reservoirs that are connected to electrodes. The efficient heat dissipation property of thin capillaries allows the separation to be performed at higher voltages, as the electric field (typically 200–300 V/cm) is much higher than that of a slab gel platform. Thus, the separation in capillary electrophoresis is rapid. During electrophoresis, DNA fragments migrate through the capillaries toward the anode. Similar to gel electrophoresis, DNA fragments are separated according to their sizes, with the shorter fragments moving faster than the longer fragments. During capillary electrophoresis, linear polymers in the solution can act as obstacles to the migrating DNA fragments. Thus, electrophoretic separation models (Figure 8.1) developed for gels are applicable for capillary electrophoresis. Currently, capillary electrophoresis can separate DNA fragments of up to 1000 nucleotides with single-nucleotide resolution.

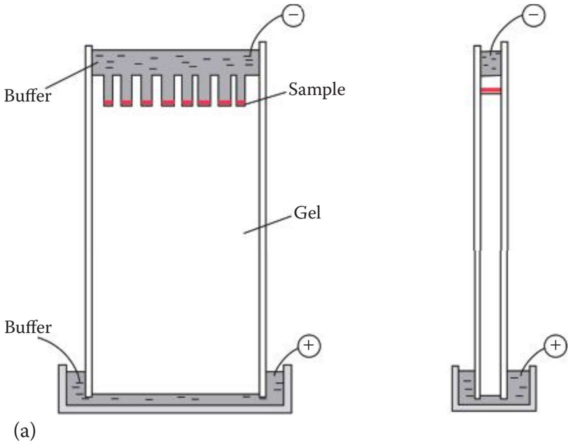

Figure 8.7 Vertical slab polyacrylamide gel. (a) Front and side view of vertical slab polyacrylamide gel apparatus. (b) Automated gel electrophoresis instrument, ABI PRISM 377® Genetic Analyzer, which was used in forensic laboratories but is now discontinued. (© Richard C. Li.)

Capillary electrophoresis instruments are equipped with detection systems utilizing laser excitation sources that excite the fluorescently labeled DNA fragments. They also include fluorescence detectors that record the signals emitted from the labeled DNA fragments (Chapter 9). In capillary electrophoresis, samples can only be analyzed sequentially so the throughput is much more limited than with slab gels. This disadvantage can be overcome by utilizing capillary array systems that can run up to 16 capillaries at one time (Figure 8.10).

8.3.3 Microfluidic Devices

Forensic DNA analysis includes the extraction of DNA from a sample, DNA quantification, polymerase chain reaction (PCR) amplification, capillary electrophoresis, data collection, and genotyping. Various microfluidic devices have been developed for forensic DNA analysis. Microfluidic devices control the movement of samples and reagents in a small, geometrically constrained environment and carry out biochemical reactions and analysis. These devices are made using microfabrication technology. This technology fabricates miniature structures historically used for integrated circuit fabrication in semiconductor manufacturing. The microfluidic

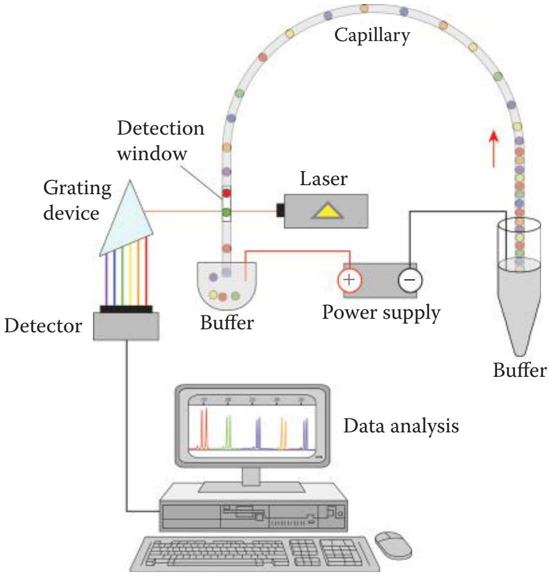
Capillary
Detection window
Grating device
Laser
Detector
Buffer
Power supply
Buffer
Data analysis

Figure 8.8 Essential components of a capillary electrophoresis system. The system includes a capillary, buffer reservoirs, two electrodes, a laser excitation source, a fluorescence detector, and an autosampler that holds the sample. Sample injection, electrophoresis, and data collection are automated and controlled by a computer. (© Richard C. Li.)

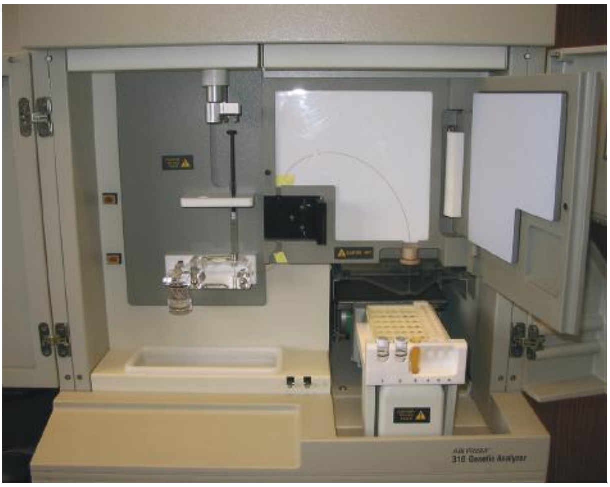
ABI PRISM
310 Genetic Analyzer

Figure 8.9 A single-capillary electrophoresis instrument, ABI PRISM 310® Genetic Analyzer. This was used in forensic laboratories but is now discontinued. (© Richard C. Li.)

devices can be classified as modular and integrated devices. Overall, the microfluidic device has many advantages compared with conventional techniques. The reaction volume required in microfluidic devices is usually in the nanoliter range, which decreases reagent and sample consumption. Additionally, due to the high surface area-to-volume ratio of the system, the efficiency of the system is greatly improved. Lastly, it can be automated, which makes the device a potential new platform for forensic DNA analysis.

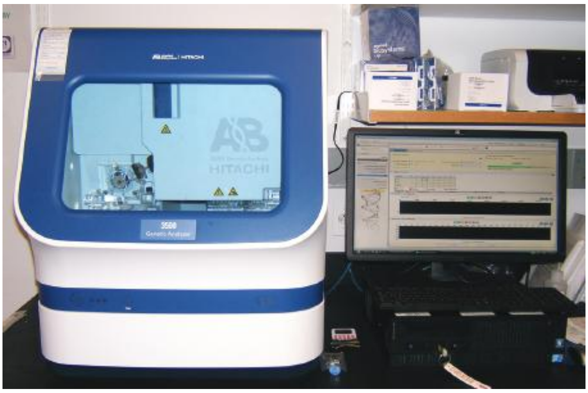

Figure 8.10 Photo of a multicapillary electrophoresis instrument. ABI PRISM 3500® Genetic Analyzer. (© Richard C. Li.)

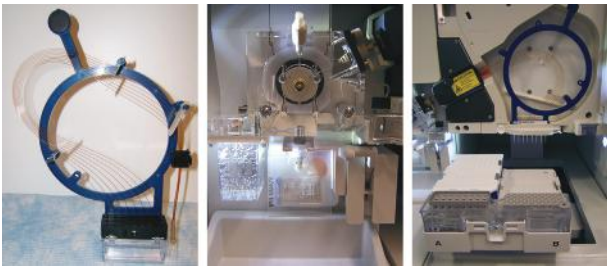

Figure 8.11 Components of the ABI PRISM 3500® Genetic Analyzer. A capillary array (left), electrophoretic buffer chambers (middle), and an autosampler (right). (© Richard C. Li.)

8.3.3.1 Modular Microfluidic Devices

The modular devices are separate devices for DNA extraction, quantitation, amplification, electrophoresis, and so on. Each processing step is carried out on different devices. The modular design is flexible, allowing laboratories to choose the best device suited to the procedure for each step. For example, an electrophoresis microfluidic device has been developed (Figure 8.12). The electrophoretic assay is carried out on a modular microfluidic chip format using the same basic principle of capillary electrophoresis. The single-use chip contains wells for samples, a sieving polymer matrix for electrophoresis, and size standards. The wells are connected through microchannels. When the interconnected wells and channels are filled, the chip becomes an integrated electrical circuit. Each well is then connected to an electrode with an independent power supply. DNA molecules, carrying negative charges, are

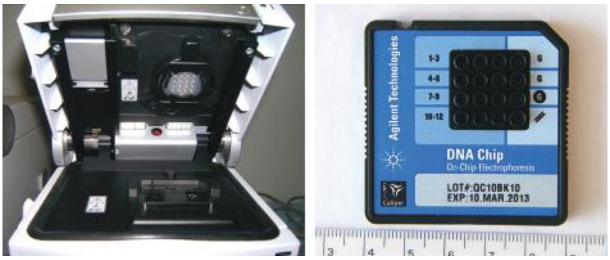

Figure 8.12 A modular instrument for gel electrophoresis. Left: Agilent 2100 Bioanalyzers contain 16-pin electrodes that fit into the wells of a chip. Right: The chip can accommodate sample wells, gel wells, and a well for a size standard. (© Richard C. Li.)

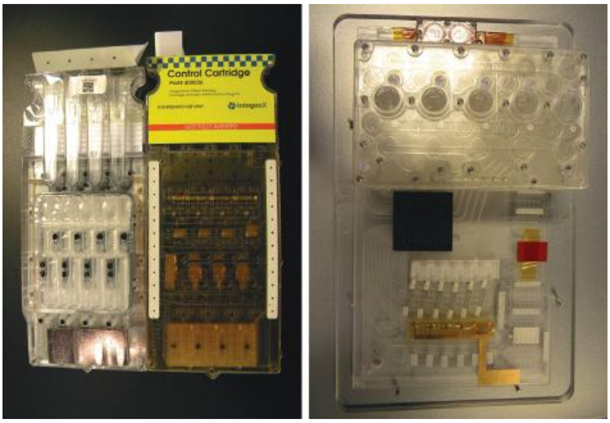

Figure 8.13 Integrated instruments for rapid DNA analysis. Cartridges developed by IntegeneX (left) and Zygem (right). (© Richard C. Li.)

electrophoretically separated based on their size. Smaller molecular weight fragments migrate faster than larger fragments. The polymer matrix contains fluorescent dye molecules that intercalate with DNA in a sample, which can be detected by laser-induced fluorescence. The duration of electrophoresis is approximately 30 min. The sensitivity is approximately 0.1 ng/μL; thus, sample consumption is small. The device provides the sizing and quantitation information of the nucleic acid fragment. In addition to DNA, the device can also be used for the analysis of RNA and protein. In the forensic laboratory, it has been used for the quantitation of amplified mtDNA product to monitor successful amplification and to normalize the quantities of DNA template for cycle sequencing.

8.3.3.2 Integrated Microfluidic Devices

It is possible to integrate several different modular devices into a single device (Figure 8.13). These devices are also known as “micro-total analysis” systems. An integrated system is a fully automated process. It also reduces the risk of contamination from laboratory sources. Full integration of all the constituent steps required for forensic STR DNA analysis has been achieved by a *rapid DNA* instrument. This instrument integrates various microfluidic devices including DNA extraction, PCR amplification, electrophoresis, detection, and genotyping steps into a single process on a cartridge. The cartridge, made from a variety of materials such as polycarbonate, incorporates pumps and valves controlling fluidic movement for the delivery of samples and prefilled reagents to the chambers for processing. DNA extraction usually utilizes silica-coated magnetic particles that can be collected by a magnet for DNA extraction. Microquantitative PCR utilizes the Taqman real-time quantitative PCR technique. During amplification, multiplex PCR is accomplished using commercial STR kits. Additionally, the cartridge also integrates a microcapillary electrophoresis chip for the separation of amplified products. The fluorescently labeled DNA fragments are detected based on laser-induced fluorescence. Data analysis and genotyping are then carried out to produce a DNA profile, which is compatible with the CODIS DNA database (Chapter 24). Rapid DNA instruments are fully automated platforms that are designed to process a single-source sample such as a buccal swab to generate a DNA profile and database search in less than 2 h. The goal is for these instruments to be deployed in field tests at the scene or at the booking stations operated by trained law enforcement agents.

8.4 Estimation of DNA Size

8.4.1 Relative Mobility

The *relative mobility* (R_f) of a DNA molecule during electrophoresis can be calculated as the distance of band migration divided by the distance of tracking dye migration. The DNA

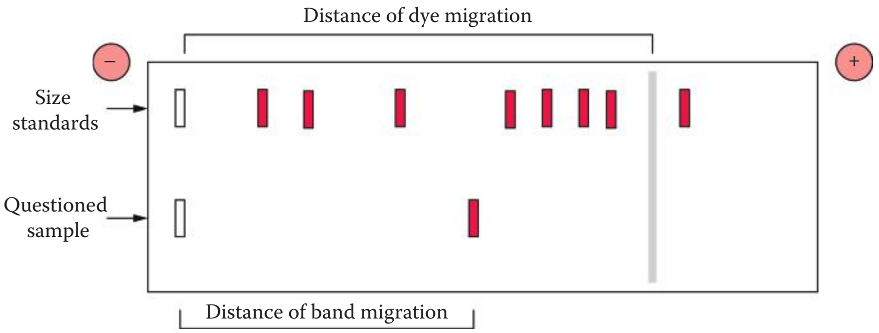

Figure 8.14 Measurement of the electromobility of DNA. Questioned DNA samples are analyzed concurrently with a set of standard DNA fragments of known size. The distances of DNA migration can thus be measured. The samples are usually mixed with a gel loading buffer prior to loading samples to the gel. The buffer contains dyes that add color to the sample to facilitate the process of loading. Dyes such as bromophenol blue and xylene cyanol FF migrate toward the positive electrode during electrophoresis. They can be used for tracking purposes as well. (© Richard C. Li.)

band migration is the distance from the sample origin to the center of the band. The tracking dye migration distance extends from the sample origin to the center of the dye band (Figure 8.14). To estimate the size of the DNA, standards containing DNA of known size and questioned samples are run on the same gel at the same time. The standards can be used to estimate the size of an unknown DNA molecule. A plot of $\log_{10}$ base pair of the standards versus R_f for a given gel can be constructed. A linear relationship, over a size range, between $\log_{10}$ base pair of the DNA molecule and the R_f can be observed. The R_f of the test sample is interpolated on the plot from which the size of an unknown DNA molecule can be determined (Figure 8.15).

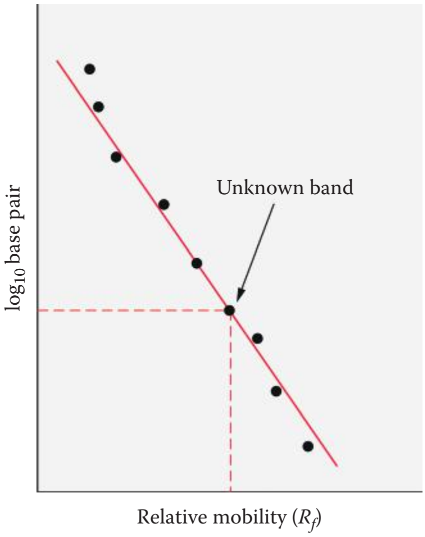

Figure 8.15 Estimation of the size of a DNA fragment. A plot of size standards versus relative migration during electrophoresis is shown. The size of the questioned DNA fragment can be estimated using this plot. (© Richard C. Li.)

8.4.2 Local Southern Method

The size of a DNA fragment is determined by an internal size standard that is a set of synthetic fragments with known molecular weight. The standard is labeled with a different colored dye so that it can be spectrally distinguished from DNA fragments of an unknown size (Figure 8.16). The sample including the internal size standard is then mixed in with DNA samples and is analyzed by electrophoresis. To determine the sizes of DNA fragments, a standard curve using the internal size standards must be established based on the reciprocal relationship between the electrophoretic mobility and the sizes of DNA fragments. However, this relationship is not exactly linear; instead, it appears to be sigmoidal (Figure 8.17). Therefore, the *Local Southern method*, described by Sir Edwin Southern, is used to generate standard curves for determining the sizes of DNA fragments. The equation is as follows:

$$L = \left[\frac{C}{M - M_0}\right] + L_0$$

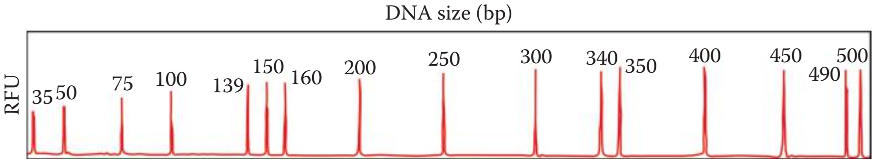

Figure 8.16 Electropherogram of a synthetic molecular weight size standard, GeneScan™ 500 size standard (Applied Biosystems). RFU, relative fluorescence units. (© Richard C. Li.)

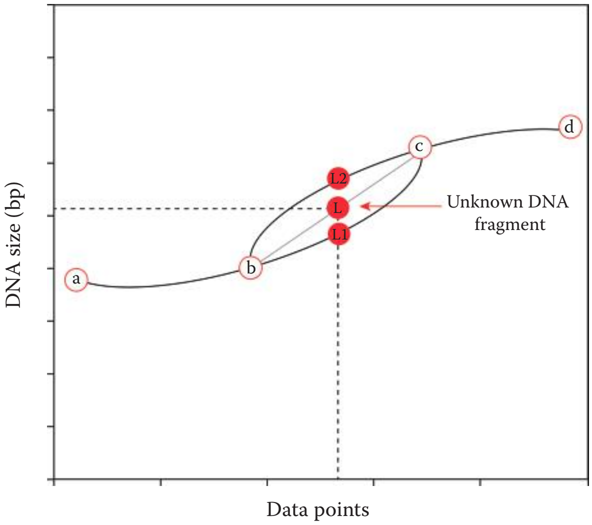

Figure 8.17 Estimation of unknown DNA fragment sizes using the Local Southern method. The migration of ssDNA fragments (20–1200 base pairs) is not linear over the entire fragment size range and is somewhat sigmoidal. To determine the size of a DNA fragment in question, two curves are generated. The first curve is generated using three standard points: two points (a and b) that are smaller and one point (c) that is larger than the questioned fragment. As a result, a fragment size (L_1) is determined. The second curve is then generated using an additional three standard points: one point (b) that is smaller and two points (c and d) that are larger than the questioned fragment. Thus, a fragment size, L_2, is determined. The two values (L_1 and L_2) are averaged to determine the size (L) of the questioned fragment. (Adapted from Southern, E.M., *Anal Biochem*, 100, 319–323, 1979.)

In the equation, L is the size of the unknown fragment; M is the mobility of the fragment; and C, M_0, and L_0 are constants that are obtained from the fragments of known size. To determine the size of an unknown fragment, two curves are generated based on the equation. Each curve utilizes three data points. The first curve utilizes two size standards that are smaller and one size standard that is greater than that of the unknown fragment. As a result, size L_1 is obtained. The second curve utilizes a size standard that is smaller and two size standards that are greater than that of the unknown fragment. Thus, size L_2 is obtained. Finally, the average of L_1 and L_2 is defined as the size of the unknown fragment. In this method, only the size standards that are in the proximity of the unknown fragment are analyzed. However, the accuracy of the standard curve relies on the proper separation of the standards during electrophoresis.

Bibliography

Aboud, M.J., M. Gassmann, and B.R. McCord, The development of mini pentameric STR loci for rapid analysis of forensic DNA samples on a microfluidic system. *Electrophoresis*, 2010, **31**(15): 2672–2679.

Biega, L.A. and B.W. Duceman, Substitution of H_2O for formamide in the sample preparation protocol for STR analysis using the capillary electrophoresis system: The effects on precision, resolution, and capillary life. *J Forensic Sci*, 1999, **44**: 3.

Bienvenue, J.M., et al., An integrated microfluidic device for DNA purification and PCR amplification of STR fragments. *Forensic Sci Int Genet*, 2010, **4**(3): 178–186.

Buel, E., M.B. Schwartz, and M.J. LaFountain, Capillary electrophoresis STR analysis: Comparison to gel-based systems. *J Forensic Sci*, 1998, **43**(1): 164–170.

Butler, J.M., C.M. Ruitberg, and P.M. Vallone, Capillary electrophoresis as a tool for optimization of multiplex PCR reactions. *Fresenius J Anal Chem*, 2001, **369**(3–4): 200–205.

Butler, J.M., et al., Rapid analysis of the short tandem repeat HUMTH01 by capillary electrophoresis. *Biotechniques*, 1994, **17**(6): 1062–1064, 1066, 1068, passim.

Butler, J.M., et al., Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis*, 2004, **25**(10–11): 1397–1412.

Chen, Y., et al., Sample stacking capillary electrophoretic microdevice for highly sensitive mini Y short tandem repeat genotyping. *Electrophoresis*, 2010, **31**(17): 2974–2980.

Choi, J.Y. and T.S. Seo, An integrated microdevice for high-performance short tandem repeat genotyping. *Biotechnol J*, 2009, **4**(11): 1530–1541.

Fernandez, C. and A. Alonso, Microchip capillary electrophoresis protocol to evaluate quality and quantity of mtDNA amplified fragments for DNA sequencing in forensic genetics. *Methods Mol Biol*, 2012, **830**: 367–379.

Frazier, R.R., et al., Validation of the applied biosystems Prism 377 automated sequencer for the forensic short tandem repeat analysis. *Electrophoresis*, 1996, **17**(10): 1550–1552.

French, D.J., et al., Interrogation of short tandem repeats using fluorescent probes and melting curve analysis: A step towards rapid DNA identity screening. *Forensic Sci Int Genet*, 2008, **2**(4): 333–339.

Gale, N., et al., Rapid typing of STRs in the human genome by HyBeacon melting. *Org Biomol Chem*, 2008, **6**(24): 4553–4559.

Gill, P., P. Koumi, and H. Allen, Sizing short tandem repeat alleles in capillary array gel electrophoresis instruments. *Electrophoresis*, 2001, **22**(13): 2670–2678.

Greenspoon, S.A., et al., A forensic laboratory tests the Berkeley microfabricated capillary array electrophoresis device. *J Forensic Sci*, 2008, **53**(4): 828–837.

Hagan, K.A., et al., A valveless microfluidic device for integrated solid phase extraction and polymerase chain reaction for short tandem repeat (STR) analysis. *Analyst*, 2011, **136**(9): 1928–1937.

Horsman, K.M., et al., Forensic DNA analysis on microfluidic devices: A review. *J Forensic Sci*, 2007, **52**(4): 784–799.

Hurth, C., et al., An automated instrument for human STR identification: Design, characterization, and experimental validation. *Electrophoresis*, 2010, **31**(21): 3510–3517.

Isenberg, A.R., et al., Analysis of two multiplexed short tandem repeat systems using capillary electrophoresis with multiwavelength fluorescence detection. *Electrophoresis*, 1998, **19**(1): 94–100.

9

Detection Methods

A variety of techniques are available for the detection of DNA fragments in forensic DNA analysis. For example, the direct detection of DNA fragments in a gel can be achieved via staining. The detection of DNA probes in a hybridization-based assay can be performed with radioisotopes, colorimetric assays, and chemiluminescence labeling. For polymerase chain reaction (PCR)-based assays, DNA primers can be labeled directly with fluorescent dyes.

9.1 Direct Detection of DNA in Gels

This section describes two simple and rapid detection methods used for detecting DNA: staining agarose gels with fluorescent intercalating dyes and staining denatured polyacrylamide gels with silver.

9.1.1 Fluorescent Intercalating Dye Staining

The location of DNA in an agarose gel can be determined directly by staining with low concentrations of fluorescent intercalating dyes (Chapter 6) such as ethidium bromide (Figures 9.1 through 9.3). These techniques allow the detection of DNA bands as small as 10 ng in agarose gels. Staining of DNA in agarose gels can be achieved by including an intercalating dye in the gel or staining the gel after electrophoresis in a dye-containing solution, followed by a washing step known as de-staining to reduce nonspecific staining. Historically, ethidium bromide-stained gels were photographed with a standard ultraviolet transilluminator at approximately 300 nm using a camera with an orange filter, whereas gels can currently be documented using digital techniques. When examining a DNA-containing gel under a UV lamp, the eyes and the skin should be protected from UV exposure. Ethidium bromide is a mutagen and a potential carcinogen. Ethidium bromide should be handled according to the Material Safety Data Sheet and safety protection wear should be used while handling the chemical. Ethidium bromide waste is usually disposed of as hazardous waste and is handled in accordance with laboratory guidelines. Alternatives to ethidium bromide are available, for example, fluorescent dyes such as SYBR® stains (Molecular Probes; Figures 9.2 and 9.3). Some of these alternatives are less mutagenic and have better performance than ethidium bromide.

9.1.2 Silver Staining

Electrophoretically separated DNA fragments can also be detected with silver nitrate staining. Silver staining of polyacrylamide gels has been used for the amplified fragment length polymorphism (AFLP; see Chapter 19) method of variable number tandem repeat (VNTR; see

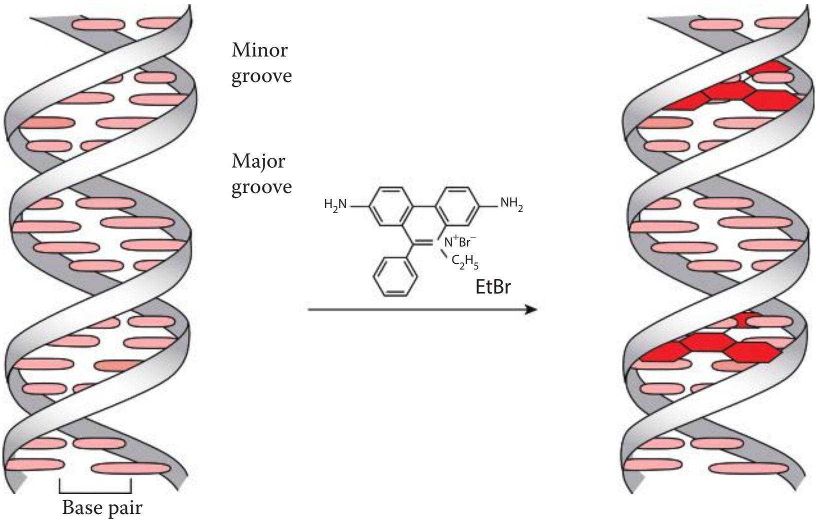

Figure 9.1 Schematic illustration showing binding of an intercalating agent, ethidium bromide (EtBr) to DNA. It intercalates into the minor groove of DNA. (© Richard C. Li.)

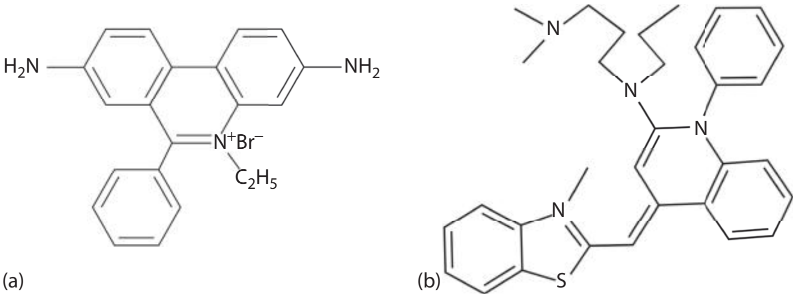

Figure 9.2 Intercalating agents used as nucleic acid stains. (a) Chemical structure of ethidium bromide. Ethidium bromide has UV absorbance. The emission maximum of the DNA–dye complex in aqueous solution is approximately 590 nm. (b) Chemical structure of SYBR Green 1. The stain is a cyanine dye that binds to DNA and is used as a nucleic acid stain. The excitation and emission maxima of the DNA–dye complex are 494 and 521 nm, respectively. The stain preferentially binds to double-stranded DNA rather than single-stranded DNA and RNA.

Chapter 19) profiling. The sensitivity of silver staining is approximately 100 times higher than that obtained with ethidium bromide, and silver staining is less hazardous than ethidium bromide detection. Also, the developing chemicals are readily available at low cost.

Silver staining involves processing a gel followed by exposure to a series of chemicals. First, the gel is submerged in a silver nitrate solution. Silver ions are positively charged, and DNA is negatively charged. Therefore, silver ions bind to the DNA and are subsequently reduced using formaldehyde to form a deposit of metallic silver on the DNA in the gel (Figure 9.4). A photograph of the gel with images of the silver-stained DNA strands is kept as a permanent record. Alternatively, the gels may be sealed and preserved for record purposes. One disadvantage of this method is that silver stains RNA and proteins along with DNA. The presence of restriction enzymes and polymerase should therefore be minimized. Additionally, bands from both complementary DNA strands may be detected in a denatured polyacrylamide gel, which leads to a two-band pattern.

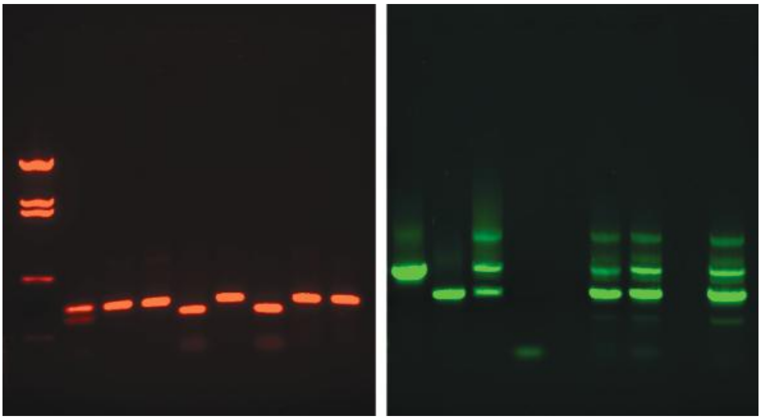

Figure 9.3 A DNA–dye complex emitting fluorescence on UV-light exposure. Ethidium bromide–containing agarose gel (left). SYBR Green 1–containing agarose gel (right). (© Richard C. Li.)

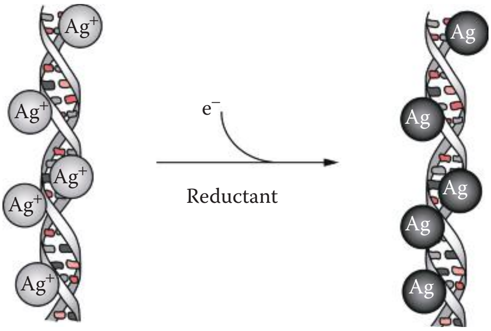

Figure 9.4 Silver staining of DNA. Silver (Ag^+) ions bind to DNA and are reduced to metallic silver (Ag) to form dark particles. (© Richard C. Li.)

9.2 Detection of DNA Probes in Hybridization-Based Assays

9.2.1 Radioisotope Labeled Probes

Radioisotope probe labeling was used for early versions of VNTR testing and DNA quantitation. Labeling can be accomplished in several ways (Figure 9.5). For example, nick translation incorporates labeled deoxyribonucleotides (dNTPs) into double-stranded DNA. DNase I is used to introduce single-strand nicks within the DNA fragment to be labeled. Next, DNA Polymerase I recognizes the nicks and replaces the preexisting nucleotides with new strands containing labeled dNTPs, resulting in the generation of ^{32}P-labeled double-stranded DNA molecules. ^{32}P, with a half-life of approximately 14 days, is the most common radioisotope used in this technique. Nick translation can utilize any dNTP labeled with ^{32}P.

Prior to hybridization, the probe is denatured into single-stranded fragments by boiling for a few minutes followed by rapid cooling on ice. After the hybridization process, these probes can be visualized by exposing the DNA-containing membrane to an x-ray film. The radioactive object is commonly placed in an x-ray cassette. The energy released from the decay of the radioisotopes is absorbed by the silver halide grains in the film emulsion and forms a latent image. A chemical development process amplifies the latent image and renders it visible on film (Figure 9.6). Because most ^{32}P emissions pass through the thin film emulsion without

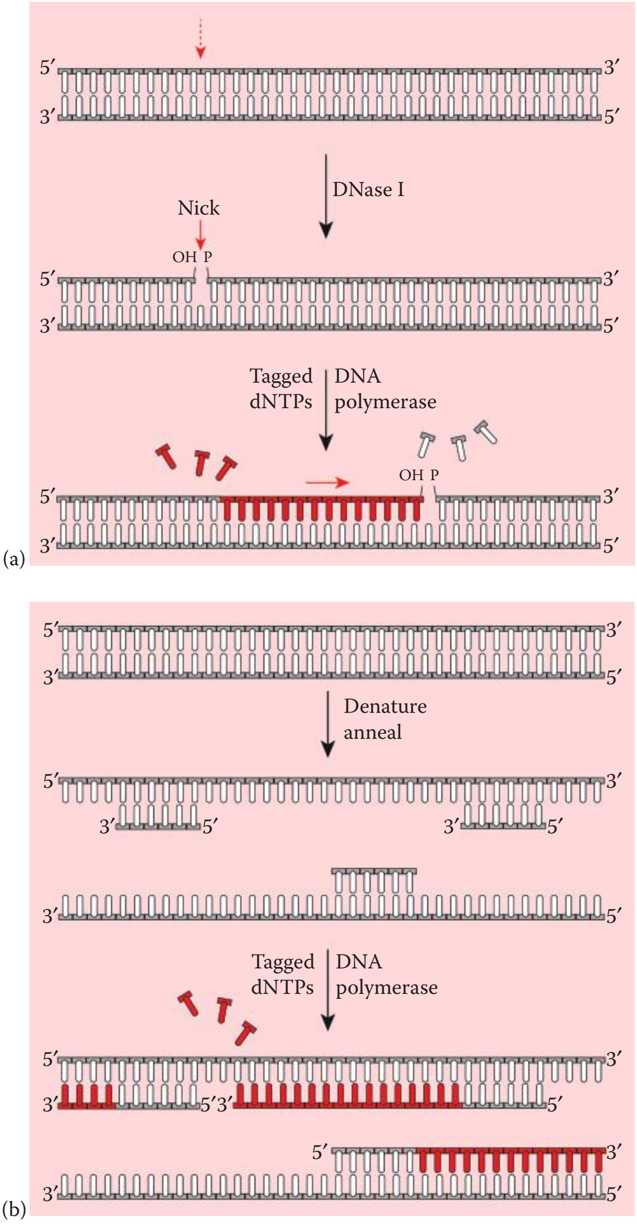
5′
3′
3′
5′
DNase I
Nick
OH P
5′
3′
3′
5′
Tagged dNTPs
DNA polymerase
OH P
5′
3′
3′
5′
(a)
5′
3′
3′
5′
Denature anneal
5′
3′
3′
5′
3′
5′
3′
5′
Tagged dNTPs
DNA polymerase
5′
3′
3′
5′3′
5′
5′
3′
3′
5′
(b)

Figure 9.5 DNA probe labeling using dNTPs. (a) Nick translation method. A DNase is utilized to create nicks along a DNA strand. Tagged nucleotides (one or more than one of the dNTPs) are then incorporated into the DNA strand using a DNA polymerase. DNA synthesis extends from the 5′ to 3′ end and the original strand is degraded. (b) Random primer labeling method. A template DNA is denatured and separated into single strands. A random mixture of hexameric deoxynucleotide fragments, as primers, is then annealed to the template strand. The tagged nucleotides are then incorporated by the Klenow fragment of the DNA polymerase. (© Richard C. Li.)

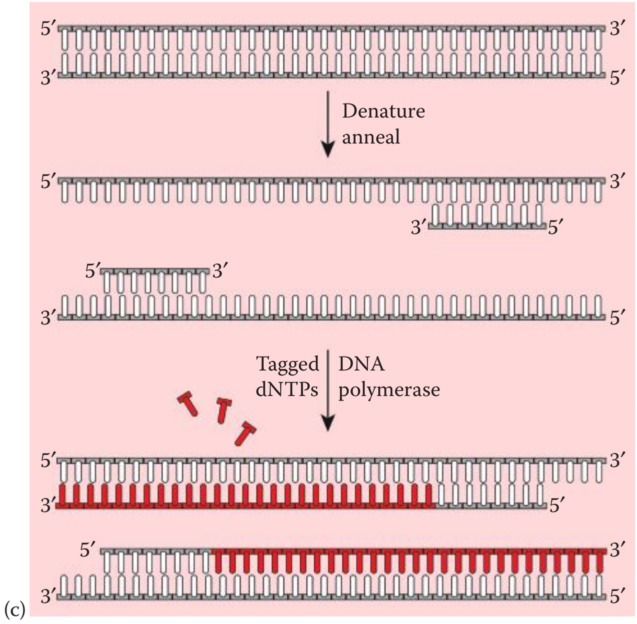

Figure 9.5 (Continued) (c) PCR method. DNA templates are denatured and then annealed to primers flanking the region of interest. During the DNA synthesis phase of a PCR cycle (see Chapter 7), tagged nucleotides are incorporated into the DNA strand using a Taq DNA polymerase, producing new double-stranded DNA. The process is repeated for many cycles. Only the first cycle is shown. (© Richard C. Li.)

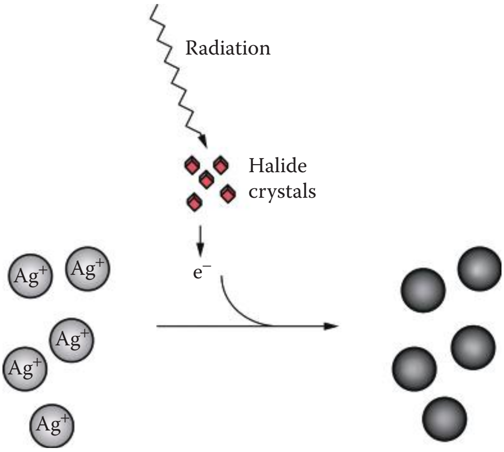

Figure 9.6 Autoradiography. Exposure to radiation causes halide crystals to release electrons, thus reducing silver (Ag^+) ions to metallic silver (Ag). (© Richard C. Li.)

contributing to the final image, the detection process may require long exposure times. Signal intensity can be enhanced, however, by using intensifying screens at low temperatures. The screens emit photons upon receiving radioactive β-particles, thus further enhancing signals. Additionally, highly sensitive x-ray film can be used. One disadvantage of using ^{32}P is that it is a safety hazard. Additionally, autoradiography is a lengthy process. Therefore, nonradioisotopic detection methods have become popular alternatives.

9.2.2 Enzyme-Conjugated Probe with Chemiluminescence Reporting System

The use of alkaline phosphatase (AP)-conjugated probes with chemiluminescent substrates comprises a highly sensitive nonradioisotopic detection system.

Alkaline phosphatase can cleave the phosphate groups from a variety of substrate molecules. Its enzymatic activity can be measured using dioxetane-based chemiluminescent substrates such as Lumigen® PPD (Figure 9.7). The Lumi-Phos Plus kit of Lumigen Inc. contains this substrate and can serve as a detection system for slot blot assays for DNA quantitation (Chapter 6) and RFLP assays for VNTR profiling (Chapter 19). AP catalyzes the cleavage of the phosphate ester of Lumigen® PPD, resulting in the release of a photon (Figure 9.8). The Lumigen® PPD substrate yields a long-lasting light emission that can be detected by exposure to x-ray film. This system provides a highly sensitive chemiluminescent detection method for AP-conjugated DNA probes in solution or on a solid matrix such as a membrane (Table 9.1).

9.2.3 Biotinylation of DNA with Colorimetric Reporting Systems

9.2.3.1 Biotin

Biotin, also known as vitamin H, is a water-soluble molecule found in egg yolk (Figure 9.9). It can be incorporated onto oligonucleotide probes without interfering with the ability of probes to hybridize because of its small size (molecular weight: 244.31 u). Signals from a biotinylated probe can be detected with an enzyme-conjugated avidin system. Two steps are required to detect biotin-labeled probes. First, an avidin conjugate consisting of a reporter enzyme is added. Then, the reporter enzyme is assayed with substrates.

9.2.3.2 Enzyme-Conjugated Avidin

Avidin is a glycoprotein found in egg white; it binds to biotin with extremely high affinity. Thus, a biotin–avidin complex is very stable. However, avidin detection has a high background due to nonspecific binding. The nonspecific binding can be reduced by replacing avidin with its

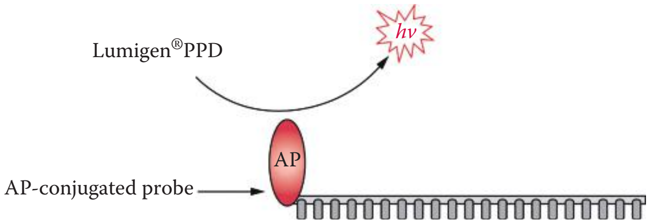

Figure 9.7 Detection system using AP-conjugated probe. Chemiluminescence is generated by using the Lumigen® PPD as an AP substrate. (© Richard C. Li.)

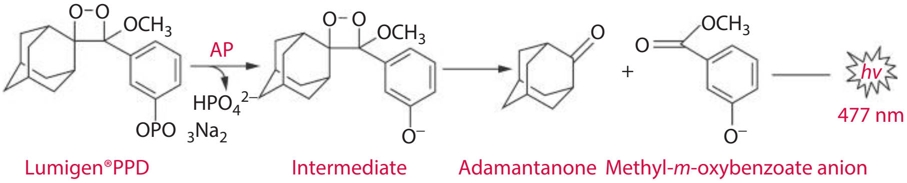

Figure 9.8 Lumi-Phos Plus contains Lumigen® PPD (4-methoxy-4-[3-phosphatephenyl]spiro[1,2-dioxetane-3,2′-adamantane], disodium salt). AP catalyzes the removal of the phosphate group of Lumigen® PPD and generates a chemiluminescent intermediate that is subsequently broken down to the excited state of methyl-*m*-oxybenzoate. The decay of the excited state end-product releases a photon.

Table 9.1 Forensic Applications of Enzyme Reporting Systems for Detecting DNA

Reporter Enzyme	Labeling	Detection Mechanism	Substrate	Forensic Application
AP	AP-conjugated probe	Chemiluminescent	Lumigen®PPD	RFLP; blot assay for DNA quantitation
Horseradish peroxidase (HRP)	Biotinylated probe, recognized by streptavidin conjugated with HRP	Colorimetric/ chemiluminescent	TMB/luminol	Blot assay for DNA quantitation
	Biotinylated primer, recognized by streptavidin conjugated with HRP	Colorimetric	TMB	Reverse blot assay for DQA1 typing and for mtDNA typing

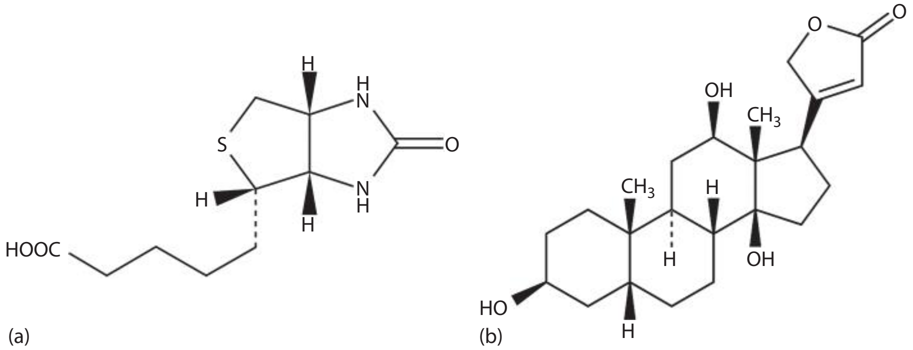

Figure 9.9 Nonradioactive tags for labeling and detecting nucleic acids. (a) Chemical structure of biotin, also known as vitamin H. Molecular formula: $C_{10}H_{16}N_2O_3S$. Molecular weight: 244.31. (b) Chemical structure of digoxigenin. Molecular formula: $C_{23}H_{34}O_5$. Molecular weight: 390.51.

streptavidin counterpart from *Streptomyces avidinii*. To detect binding, an enzyme-conjugated streptavidin such as horseradish peroxidase (HRP)-conjugated streptavidin can be used. HRP isolated from horseradish roots contains heme residues that catalyze the oxidation reactions of substrates (Chapter 12).

9.2.3.3 Reporter Enzyme Assay

HRP can be assayed with colorimetric, chemiluminescent, or fluorogenic substrates. One common colorimetric substrate for forensic DNA testing is 3,3′,5,5′-tetramethylbenzidine (TMB), which is oxidized by the peroxidase to form an insoluble precipitate of intense blue at an acidic pH (Figure 9.10). Because the colored precipitate is difficult to remove from membranes, TMB is not suitable if reprobing for RFLP analysis is required. This technique has been used for forensic DNA testing such as slot blot assays for DNA quantitation (Table 9.1). A chemiluminescent substrate such as a luminol-based reagent can also be utilized with

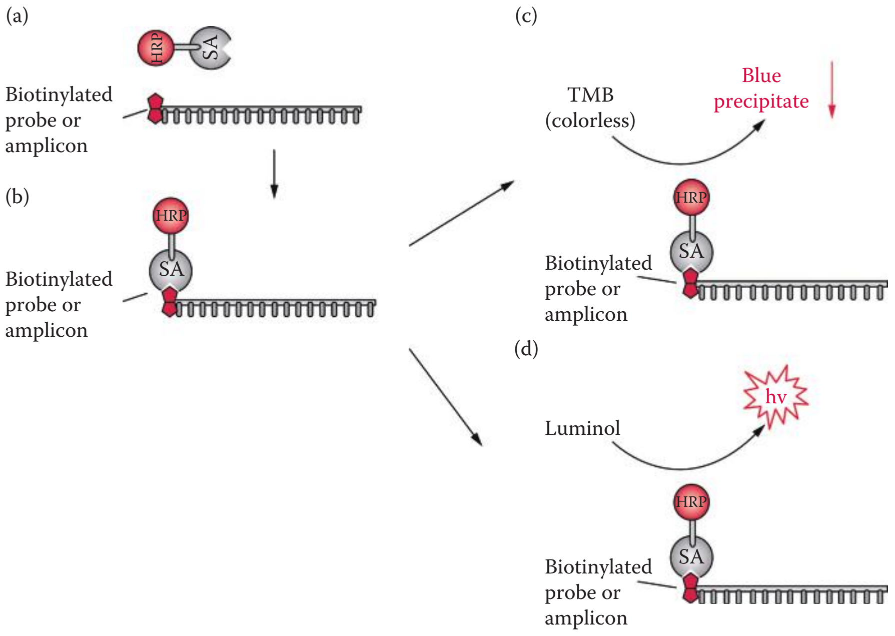
(a)
HRP
SA
Biotinylated probe or amplicon
(b)
HRP
SA
Biotinylated probe or amplicon
(c)
TMB (colorless)
Blue precipitate
HRP
SA
Biotinylated probe or amplicon
(d)
Luminol
hv
HRP
SA
Biotinylated probe or amplicon

Figure 9.10 Detection system using biotinylated DNA probes with colorimetric and chemiluminescent reactions. (a) Biotinylated probe is incubated with a streptavidin (SA) and horseradish peroxidase (HRP) conjugate complex. (b) Biotin is recognized by the complex. Reporter enzyme assays can be carried out using either a colorimetric reaction with a TMB substrate (c) or a chemiluminescent reaction using luminol analogs (d). (© Richard C. Li.)

HRP. The peroxidase catalyzes the oxidation of luminol to form a chemiluminescent product (Figure 9.10).

9.3 Detection Methods for PCR-Based Assays

9.3.1 Fluorescence Labeling

9.3.1.1 Fluorescent Dyes

The advantages of fluorescence detection methods include a higher sensitivity and broader dynamic range than comparable colorimetric detection methods. Furthermore, they have the capacity for simultaneous analysis of complex samples such as multiplex PCR products with different fluorescent labels (Figure 9.11), allowing the distinction of various amplicons. Commonly used fluorescent dyes in DNA labeling emit fluorescence in the range of 400–600 nm (Figures 9.12 and 9.13).

9.3.1.2 Labeling Methods

Fluorescent dye labeling can be incorporated into a DNA fragment using a 5′-end fluorescently labeled oligonucleotide primer (Figures 9.14 and 9.15). The dye-labeled primer method is usually used for STR profiling (see Chapter 20) in which only one primer from each primer pair is labeled; therefore, only one strand can be detected. The two-band pattern observed with silver staining does not appear with this method. Additionally, dye-labeled primers allow multiplex PCR amplifications in the same tube.

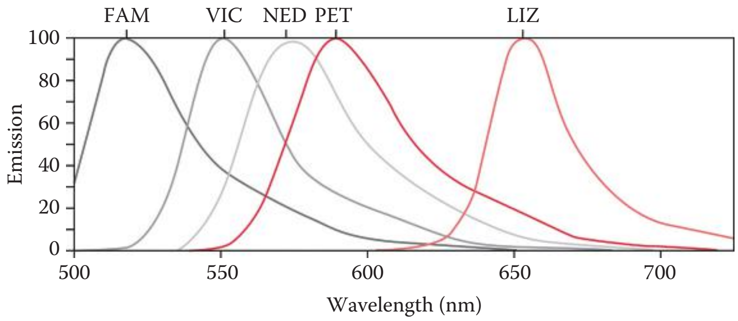

Figure 9.11 Emission spectra of common fluorescent dyes used for forensic DNA analysis. (© Richard C. Li.)

Color	Fluorescent dye	Globalfiler	AmpFlSTR Yfiler	PowerPlex fusion system	PowerPlex Y23 system
Purple dye	SID	☑			
Blue dyes	FAM	☑	☑		
	Fluorescein			☑	☑
Green dyes	JOE			☑	☑
	VIC	☑	☑		
Yellow dyes	NED	☑	☑		
	TMR (TAMRA)			☑	☑
Orange dyes	CC5			☑	☑
	LIZ	☑	☑		
Red dyes	PET		☑		
	ROX (CXR)			☑	☑
	TAZ	☑			

Figure 9.12 Examples of fluorescent dyes and their applications in STR kits. *Note*: Globalfiler and AmpFlSTR Yfiler are manufactured by Applied Biosystems. PowerPlex Fusion system and PowerPlex® Y23 System are manufactured by Promega.

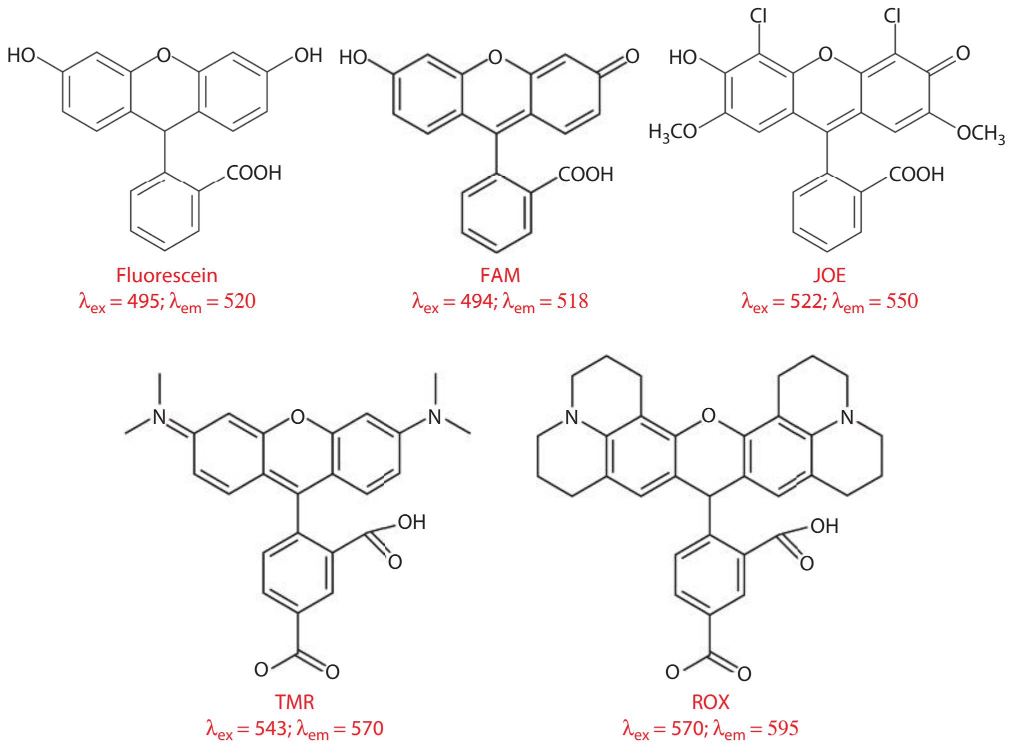
HO
OH
O
COOH
HO
O
O
COOH
Cl
Cl
HO
O
O
H_3CO
OCH_3
COOH
N
O
N
OH
O
O
O
N
O
N
OH
O
O
O

Fluorescein
$\lambda_{ex} = 495$; $\lambda_{em} = 520$

FAM
$\lambda_{ex} = 494$; $\lambda_{em} = 518$

JOE
$\lambda_{ex} = 522$; $\lambda_{em} = 550$

TMR
$\lambda_{ex} = 543$; $\lambda_{em} = 570$

ROX
$\lambda_{ex} = 570$; $\lambda_{em} = 595$

Figure 9.13 Chemical structures of representative fluorescent dyes. Fluorescein, 3′,6′-dihydroxyspiro[2-benzofuran-3,9′-xanthene]-1-one; FAM, 5(6)-carboxyfluorescein; JOE, 6-carboxy-4′,5′-dichloro-2′,7′-dimethoxy-fluorescein; TMR, 5-carboxytetramethylrhodamine; ROX, 5(6)-carboxy-X-rhodamine; λ_{ex}, peak excitation wavelength (nm); λ_{em}, peak emission wavelength (nm).

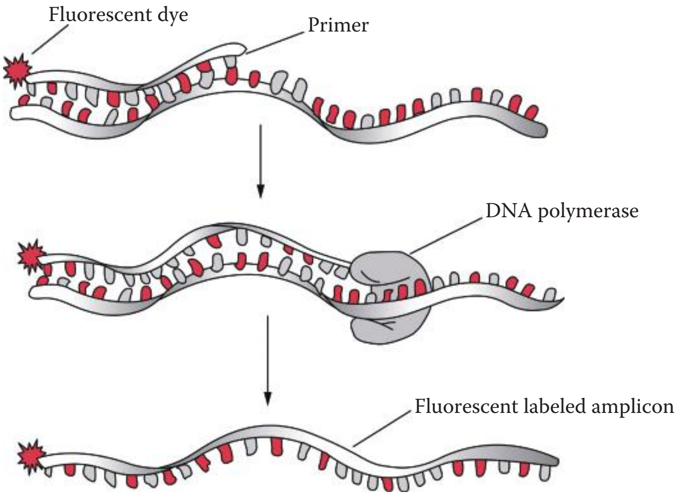
Fluorescent dye
Primer
DNA polymerase
Fluorescent labeled amplicon

Figure 9.14 A fluorescent dye–labeled primer can be used for the amplification of DNA. The dye is conjugated at the 5′ end of the primer. The amplified product is fluorescently labeled. (© Richard C. Li.)

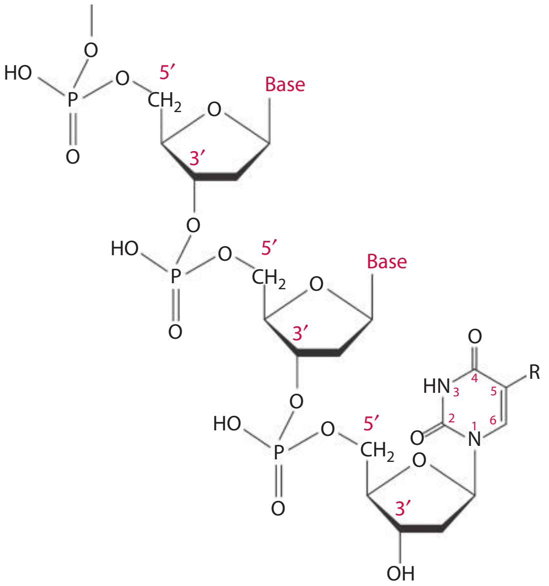

Figure 9.15 Direct labeling of nucleic acids with fluorescent dye. Labeled cytosine is shown as an example. R, fluorescent dye.

Alternatively, fluorescent dye labeling of DNA fragments can be carried out by incorporating fluorescently labeled dideoxynucleotides (ddNTPs) in the PCR product. This labeling method is usually used in DNA sequencing such as mtDNA sequence profiling (Chapter 23).

9.3.1.3 Fluorophore Detection

The *fluorophore* is a component of a florescent dye molecule that causes the molecule to be fluorescent. First, a laser strikes a fluorophore covalently linked to the end of a DNA fragment. An electron of the fluorophore is then excited, rising to an excited state from a ground state. The excited electron then descends to the ground state and releases a photon (Figure 9.16). The emitted photon has a longer wavelength (emission spectrum) than that of the excitation photon (excitation spectrum). The wavelengths of excitation and emission spectra (Figure 9.17) are largely dependent on the chemical structure of the fluorophore.

Lasers are commonly used as *excitation sources* because laser light emissions have high intensity and are monochromatic (single wavelength). The argon ion gas laser is frequently used in applications such as fluorescence-labeled STR and mtDNA sequence analysis because the excitation wavelength of commonly used fluorescent dyes matches the wavelength of the argon laser.

Optical filters are used to filter out undesired light and to allow only one particular wavelength to pass through. An essential optical filter consists of three components: an excitation filter, a dichroic beam splitter, and an emission filter. The excitation filter selectively transmits light from an excitation source. The light is then directed by the dichroic beam splitter to DNA molecules labeled with fluorescent dyes. The light emitted from a fluorophore is also transmitted by the dichroic beam splitter toward the detector. The emission filter selectively blocks undesired light, thus transmitting a specific wavelength of the emitted fluorescence. The light intensity emitted from a fluorophore is detected using a photosensitive device such as a charge-coupled device (CCD). The signal from the fluorophore is collected and converted to an electronic signal expressed in an arbitrary unit such as a *relative fluorescence unit* (RFU). Signals from multiple fluorophores in the same sample can be recorded separately using optical filters and a mathematical *matrix* (fluorophore separation algorithm). The function of a matrix is to subtract the

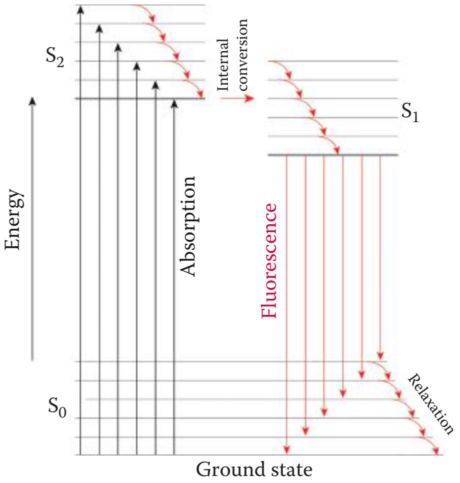

Figure 9.16 A Jablonski diagram illustrating the absorption of a single photon and subsequent emission of fluorescence. During an excitation stage, a single photon with the appropriate energy excites a molecule to the first excited energy level (S_2). In fluorescence reactions, the energy is supplied by an external light source such as a laser, while in chemiluminescence reactions, the energy is derived from a chemical reaction. The excited molecule is not stable and relaxes to its lowest vibrational level in the excited state (S_1). The molecule then returns to the ground state (S_0), emitting a photon. (© Richard C. Li.)

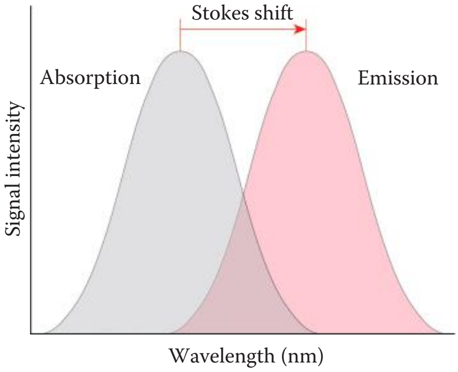

Figure 9.17 A fluorescence excitation spectrum. In fluorescence reactions, the energy of the photon emitted is lower than that of the excitation photon. As a result, the wavelength of an emission photon is longer than that of an excitation photon. The difference in wavelength is known as the Stokes shift. (© Richard C. Li.)

color overlaps of the various fluorescent dyes. The matrix can be established by the calibration of fluorescently labeled standards.

Bibliography

Bassam, B.J., G. Caetano-Anolles, and P.M. Gresshoff, Fast and sensitive silver staining of DNA in polyacrylamide gels. *Anal Biochem*, 1991, **196**(1): 80–83.

Bregu, J., et al., Analytical thresholds and sensitivity: Establishing RFU thresholds for forensic DNA analysis. *J Forensic Sci*, 2013, **58**(1): 120–129.

Bronstein, I. and P. McGrath, Chemiluminescence lights up. *Nature*, 1989, **338**(6216): 599–600.

10

Serology Concepts

10.1 Serological Reagents

10.1.1 Immunogens and Antigens

A foreign substance that is capable of eliciting antibody formation when introduced into a host is called an *immunogen*. Natural immunogens are usually macromolecules such as proteins and polysaccharides. Other molecular structures can also act as immunogens, for example, glycolipids (such as A, B, and O blood group antigens) and glycoproteins (such as Rh and Lewis antigens). However, they must be foreign to their hosts. The molecular structure of an immunogen, usually a small portion recognized by an antibody, is called an *epitope* or determinant site. An immunogen usually consists of multiple epitopes and is thus considered *multivalent* (Figure 10.1). Each epitope can elicit the production of its own corresponding antibody.

An *antigen* is a foreign substance that is capable of reacting with an antibody. All immunogens can be considered antigens, but not all antigens can elicit antibody formation. *Hapten* is one example of a substance that is antigenic but not immunogenic. Haptens are chemical compounds that are too small to elicit antibody production when they are introduced to a host animal. However, a hapten can be coupled to a *carrier*, usually a macromolecule, to produce antibodies. A hapten-conjugated carrier can become immunogenic to elicit the formation of an antibody specific to the hapten. The resulting antibody can bind to free haptens. Certain controlled substances such as cocaine and amphetamines are haptens and can be detected through corresponding antibodies for forensic toxicological analysis.

10.1.2 Antibodies

Antibodies, also known as *immunoglobulins*, are capable of binding specifically to antigens and are designated with an Ig prefix. The five major classes of immunoglobulins are designated IgG, IgA, IgM, IgD, and IgE. Additionally, IgG immunoglobulins can be further divided into four subclasses (IgG1–IgG4) and IgA immunoglobulins can be further divided into two subclasses (IgA1 and IgA2). Thus, there are a total of nine immunoglobulin isotypes in humans. *Isotypes* are the immunoglobulins that differ based on the molecular variations in the constant domains of the heavy and light chains (Figure 10.2). IgD, IgE, and IgG are usually monomers. IgM can be a membrane-bound monomer or a cross-linked pentamer (secreted form). IgA can be a monomer, dimer, or trimer. In an immune response, an initial exposure to an immunogen elicits a primary response, producing IgM immunoglobulins. Further exposure to the immunogen can elicit a secondary response, producing IgG, IgA, IgE, and IgD immunoglobulins. IgG is the most abundant immunoglobulin in serum. A majority of serology tests are based on the IgG immunoglobulins.

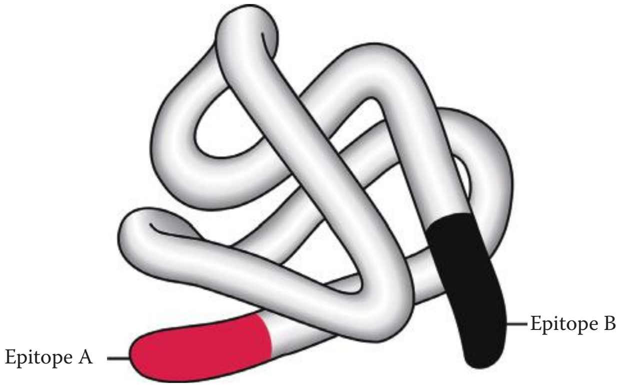

Figure 10.1 Multivalent immunogen. A protein with two different epitopes is shown. (© Richard C. Li)

Immunoglobulins have many similarities in their molecular structures. Figure 10.2 illustrates a diagram of the IgG molecule. The structures of immunoglobulins were first revealed by Gerald Edelman and Rodney Porter who shared a Nobel Prize in 1972. Immunoglobulins are composed of four polypeptide chains: two heavy (H) chains and two light (L) chains. The polypeptide chains are linked by disulfide bonds into a Y-shaped complex. The H chain can be divided into fragment antigen-binding (*Fab*) and fragment crystallizable (*Fc*) fragments. The L chain consists of a Fab fragment only. A typical antibody has two identical antigen-binding sites and is thus considered *bivalent*. The antigen-binding activity is located within the Fab fragments. In particular, the N-terminal ends of the L and H chains together form antigen-binding sites. At the amino acid sequence level, both H and L chains have *variable and constant domains* (Figure 10.2). The variable domains are located at the N-terminal ends of the immunoglobulins. Additionally, three small *hypervariable regions* are located within the variable domain of each chain.

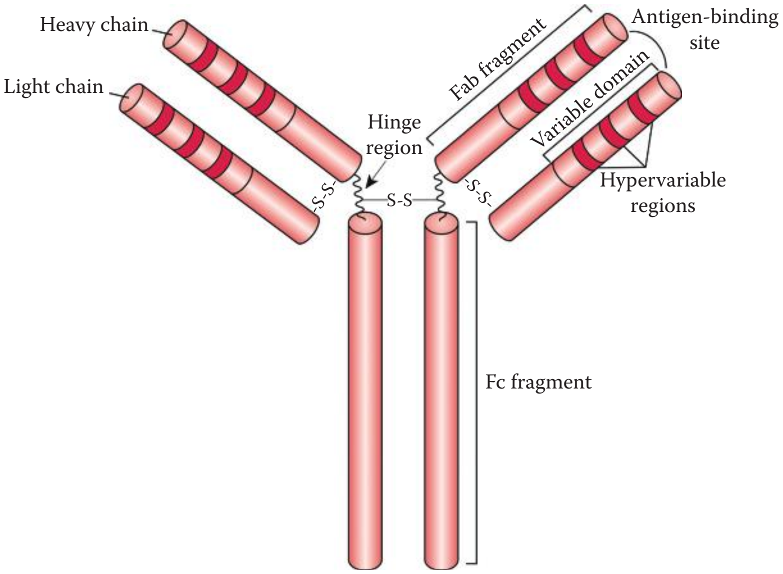

Figure 10.2 Immunoglobulin IgG structure. Immunoglobulin IgG is composed of light and heavy chains that contain variable domains. The remaining portions of chains form the constant domains. The variable domains of both light and heavy chains contain three hypervariable regions that form the antigen-binding site of the immunoglobulin. (© Richard C. Li.)

The diversity in the amino acid sequences of these hypervariable regions determines the specificity of the antigen-binding sites. The hinge regions provide flexibility to the antibody molecule and are important for the efficiency of the binding and cross-linking reactions. The basal portion of the H chain consists of *Fc* fragment.

The binding affinity and specificity of antibodies make them useful reagents for serological testing. Two types of antibodies are commonly used: polyclonal and monoclonal antibodies.

10.1.2.1 Polyclonal Antibodies

To produce an antibody, an immunogen is usually introduced into a host animal. A multivalent immunogen is capable of eliciting a mixture of antibodies with diverse specificities for the immunogen. As a result, a polyclonal antibody is produced by different B lymphocyte clones in response to the different epitopes of the immunogen. Antibodies can be circulating (in blood or other bodily fluids) or tissue bound (in cell surface antibodies). Circulating immunoglobulins are referred to as humoral antibodies.

The blood from an immune host is drawn and allowed to clot, resulting in the formation of a solid consisting largely of blood cells and a liquid portion known as *serum* containing antibodies (Figure 10.3). Such a preparation of humoral antibodies is also called a *polyclonal antiserum.* Depending on the type of animal used, the antibodies produced are classified as avian (B), rabbit (R), or horse (H) type. The characteristics of polyclonal antibodies may vary if they are produced from different individual host animals of the same species. Variations in reactions among different sources of antibodies should be monitored by quality-control procedures.

10.1.2.2 Monoclonal Antibodies

To produce a monoclonal antibody, spleen cells are harvested from a host animal, such as a mouse, inoculated with an immunogen (Figure 10.4). Next, the plasma cells of the spleen, which produce antibodies, are fused with myeloma cells. Since only a small population of cells fuse, a selection step is needed to allow only fused cells to grow (Figure 10.4). The fused cells, called *hybridoma* cells, are immortal (proliferate indefinitely) in cell cultures. Pools of hybridoma cells are diluted into single clones and are allowed to proliferate. The clones are then screened for the specific antibody of interest.

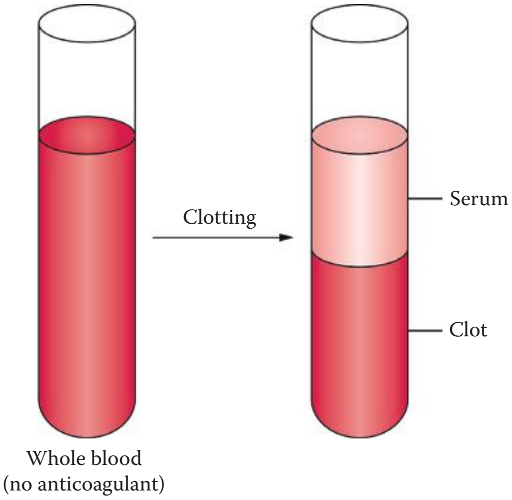

Figure 10.3 Serum component of blood. The blood of an immunized animal is collected in the absence of an anticoagulant and is allowed to clot. The resulting liquid portion of the blood is serum. (© Richard C. Li.)

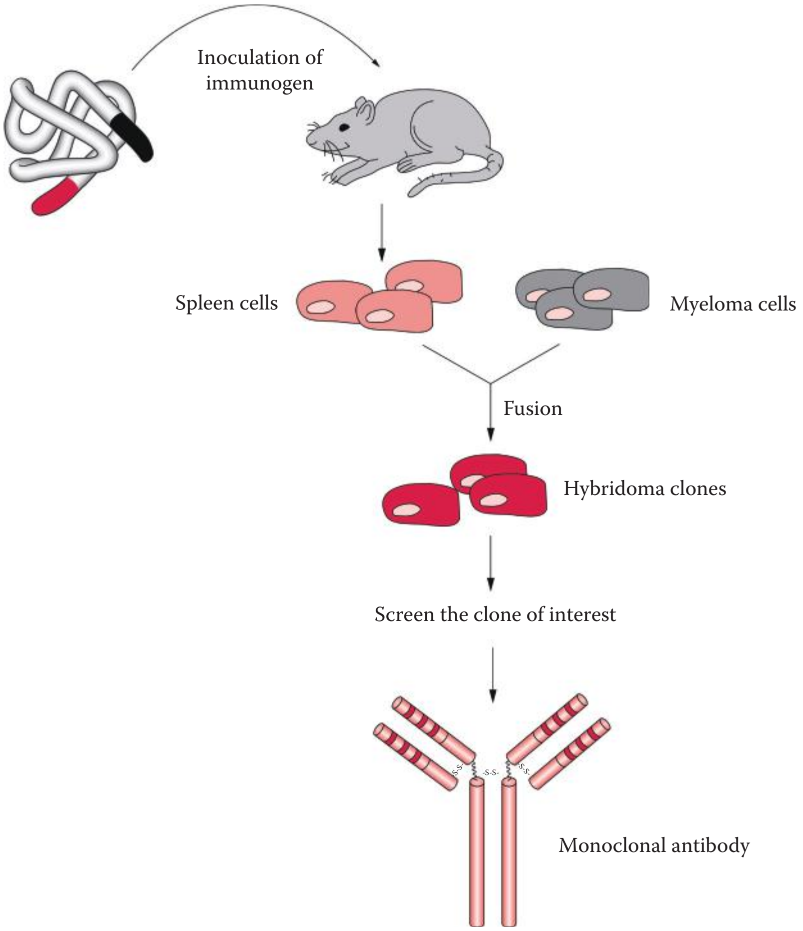

Figure 10.4 Preparation of a monoclonal antibody. A mouse is immunized with an immunogen, and its spleen cells are fused with myeloma cells to generate hybridoma cells. The clone that synthesizes and secretes the monoclonal antibody of interest can then be identified. (© Richard C. Li.)

The desired hybridoma clone can be maintained indefinitely, and it produces a monoclonal antibody that reacts with a single epitope. The hybridoma-derived monoclonal antibodies are specific and homogenous, and can be obtained in unlimited quantities. Monoclonal antibodies have been utilized in many serological assays, as discussed in Chapter 11. However, they have certain limitations in serology assays. For instance, monoclonal antibodies react with only a single epitope of a multivalent antigen and, therefore, cannot form cross-linked networks in precipitation assays (see Section 10.3.2.1).

10.1.2.3 Antiglobulins

Immunoglobulins are proteins that can also function as immunogens. If a purified foreign immunoglobulin or a fragment of a foreign immunoglobulin is introduced into a host, the antibodies produced are known as *antiglobulins*. Antiglobulins that are specific to a particular isotype can be produced in laboratories. In addition to specific antiglobulins, it is possible to produce nonspecific antiglobulins, which recognize an epitope that is common to all isotypes of an immunoglobulin class, such as the Fc portion of the heavy chain of all subclasses of human IgG. Antiglobulins are important reagents in many serological tests.

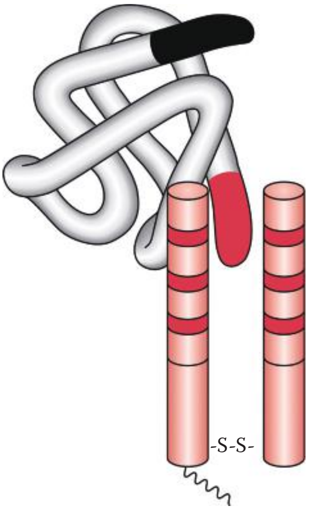

Figure 10.5 Affinity is a measure of the interaction between a single epitope on an antigen and a single binding site on an antibody. (© Richard C. Li.)

10.2 Strength of Antigen–Antibody Binding

The binding of an antigen to its specific antibody is mediated by the interaction between the epitope of an antigen and the binding site of its antibody. Noncovalent bonds can be formed during antigen–antibody binding. Various forces act cooperatively during antigen–antibody binding. These include hydrogen bonding, hydrophobic interactions that exclude water molecules from the area of contact, and van der Waals forces arising from the asymmetric distribution of the charges of electrons. The binding process occurs rapidly and the formation of the antigen–antibody complex is reversible. Such binding occurs at short distances when the antigen and antibody are in close proximity. Additionally, the strongest binding occurs only if the shape of the epitope fits the binding site of the antibody. The strength of the interaction between the antigen and the antibody depends on two characteristics, designated affinity and avidity.

Affinity is the energy of the interaction of a single epitope on an antigen and a single binding site on a corresponding antibody (Figure 10.5). The strength of the interaction depends on the specificity of the antibody for the antigen. Nevertheless, antibodies can bind with lower strength to antigens that are structurally similar to the immunogen. Such binding is known as *cross-reaction*.

Avidity is the overall strength of the binding of an antibody and an antigen (Figure 10.6). Since an antigen is usually considered multivalent and an antibody is bivalent, the avidity reflects the combined synergistic strength of the binding of all the binding sites of antigens and antibodies rather than the sum of individual affinities. It also reflects the overall stability of an antigen–antibody complex that is essential for many serological assays.

10.3 Antigen–Antibody Binding Reactions

The binding of an antigen to an antibody is an equilibrium reaction consisting of three types of reactions. The primary and secondary reactions form the basis for many forensic serological assays and will be discussed in the following subsections. The third type is called the tertiary reaction. It is used to measure *in vivo* immune responses such as inflammation and phagocytosis. Because most forensic serology tests are *in vitro* assays, the tertiary reaction is not commonly utilized in forensic serology testing and will not be discussed here.

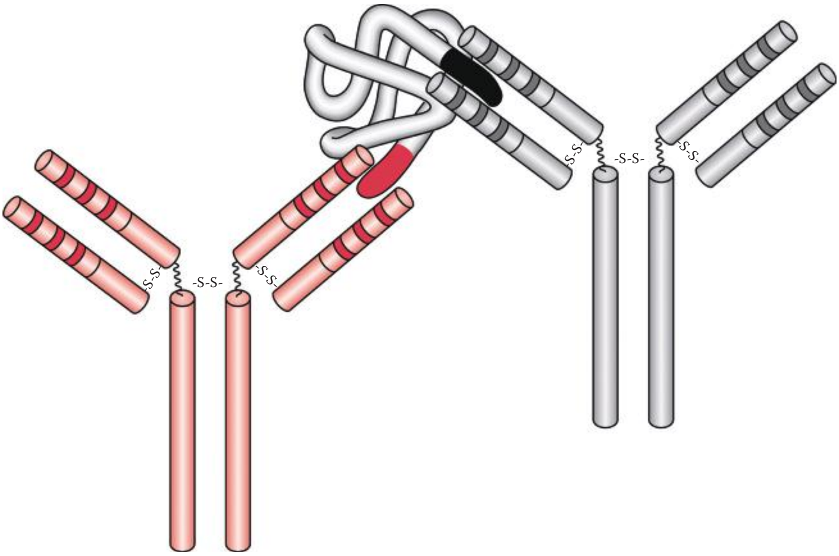

Figure 10.6 Avidity is a measure of the overall strength of the binding between antigens and antibodies. (© Richard C. Li.)

10.3.1 Primary Reactions

A primary reaction is the initial binding of a single epitope of an antigen (Ag) and a single binding site of an antibody (Ab) to form an antigen–antibody complex (Figure 10.7). This rapid and reversible binding reaction can be expressed as

$$\mathrm{Ag} + \mathrm{Ab} \rightleftarrows \mathrm{AgAb}$$

At equilibrium, the strength of the interaction can be expressed as the affinity constant (K_a) that reflects the affinity of binding, where:

$$K_a = \frac{[\mathrm{AgAb}]}{[\mathrm{Ag}][\mathrm{Ab}]}$$

The square brackets indicate the concentration of each component at equilibrium. K_a is the reciprocal of the concentration of free epitopes when half the antibody-binding sites are occupied. Thus, a higher K_a corresponds to a stronger binding interaction.

Techniques such as enzyme immunoassays, immunofluorescence assays, radioimmunoassays, and dye-labeled immunochromatography can measure the concentrations of antigen–antibody complexes formed by primary reactions (Chapter 11). These techniques are the most sensitive for detecting the presence of an antigen and an antibody in a sample. Additionally, many forensic serology assays are based on the detection of primary reactions and will be discussed in Chapters 12 through 17.

10.3.2 Secondary Reactions

The primary reaction between an antigen and an antibody is often followed by a secondary reaction. The three types of secondary reactions are precipitation, agglutination, and complement fixation. The techniques that detect secondary reactions are usually less sensitive but easier to perform than primary reaction assays. The precipitation and agglutination reactions form the basis for many serologic assays performed in forensic laboratories. These reactions will be discussed in the following subsections in detail. The third type of reaction is called *complement fixation*. If an antigen is located on a cell surface, the binding of the antigen and the antibody

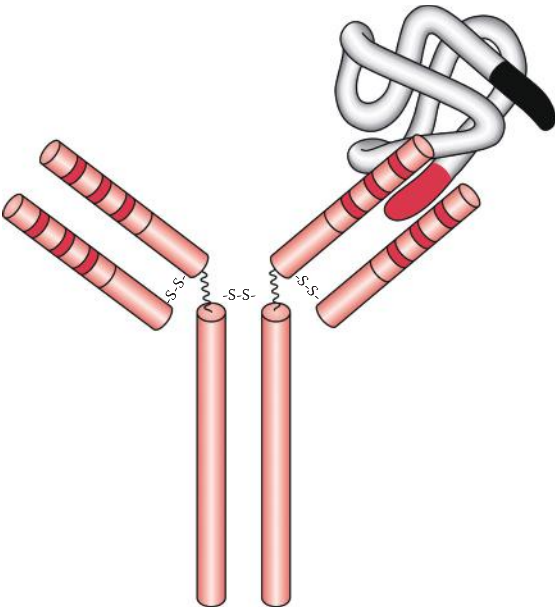

Figure 10.7 Primary reaction. Initial binding forms an antigen–antibody complex. (© Richard C. Li.)

may activate the classical complement pathway and lead to cell lysis, also known as a complement fixation reaction. The detection of this type of reaction is not commonly used in forensic serology.

10.3.2.1 Precipitation

If a soluble antigen is mixed and incubated with its antibody, the antigen–antibody complexes can form cross-linked complexes at the optimal ratio of antigen-to-antibody concentration. The cross-linked complex is insoluble and eventually forms a precipitate that settles to the bottom of a test tube. Antibodies that produce such precipitation are also called *precipitins*.

This precipitation reaction can be characterized by examining the effect of varying the relative ratio of antigen and antibody. If an increasing amount of soluble antigen is mixed with a constant amount of antibody, the amount of precipitate formed can be plotted. A precipitin curve (Figure 10.8)

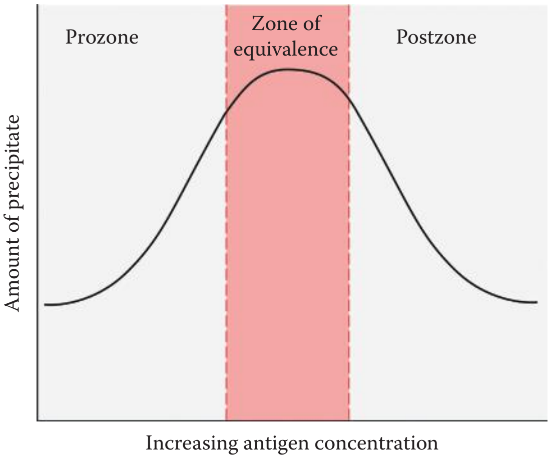

Figure 10.8 Precipitin curve. (© Richard C. Li.)

illustrates the results observed when antigens and antibodies are mixed in various concentration ratios. The curve can be divided into three zones known as the prozone, the zone of equivalence, and the postzone.

10.3.2.1.1 Prozone

At this zone, the ratio of antigen–antibody concentration is low. In other words, the antibody is in excess. Each antigen molecule is rapidly saturated with antibody, thus preventing cross-linking (Figure 10.9a). No precipitate is formed at the prozone stage.

10.3.2.1.2 Zone of Equivalence

As the concentrations of antigen increase, the amount of precipitate increases until it reaches a maximum. The amount of precipitation depends on the relative proportions of antigens and antibodies present. The maximum precipitation occurs in what is called the zone of equivalence. In the zone of equivalence, the ratio of antibody to antigen concentration is optimal and

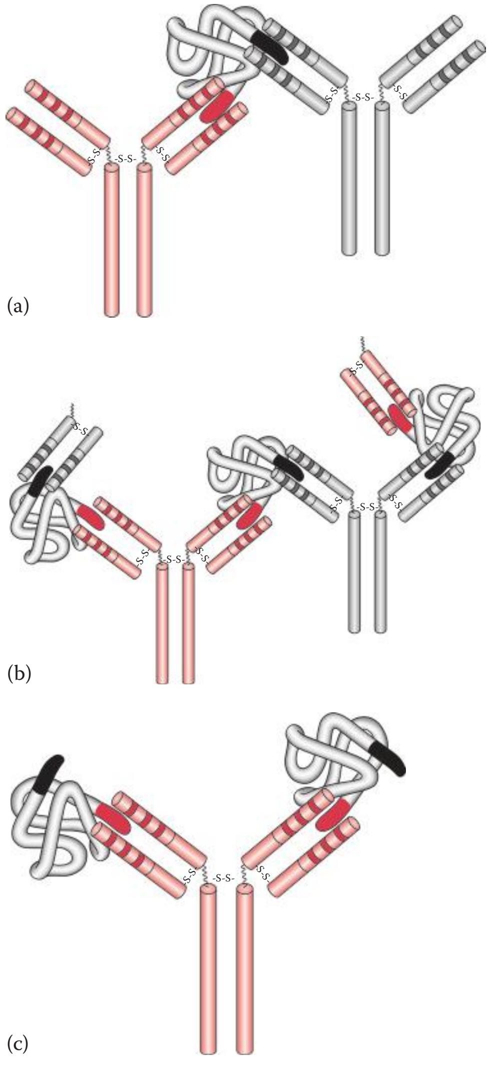

Figure 10.9 Antigen–antibody binding in (a) prozone, (b) zone of equivalence, and (c) postzone. (© Richard C. Li.)

precipitation occurs as a result of forming cross-linked networks (Figure 10.9b). Precipitation assays are usually carried out under the condition of the zone of equivalence, forming a sufficient quantity of precipitation to be detected.

10.3.2.1.3 Postzone

With the addition of more antigens, the ratio of antigen–antibody concentrations is high. In other words, the antigen is in excess. The amount of precipitate decreases and eventually diminishes. Each antibody molecule is saturated with antigen molecules (Figure 10.9c). Cross-linkage cannot form, and precipitation does not occur.

10.3.2.2 Agglutination

As discussed in the previous subsection, precipitation reactions involve soluble antigens. If the antigens are located on the surfaces of cells or carriers (carrier cells such as sheep erythrocytes,

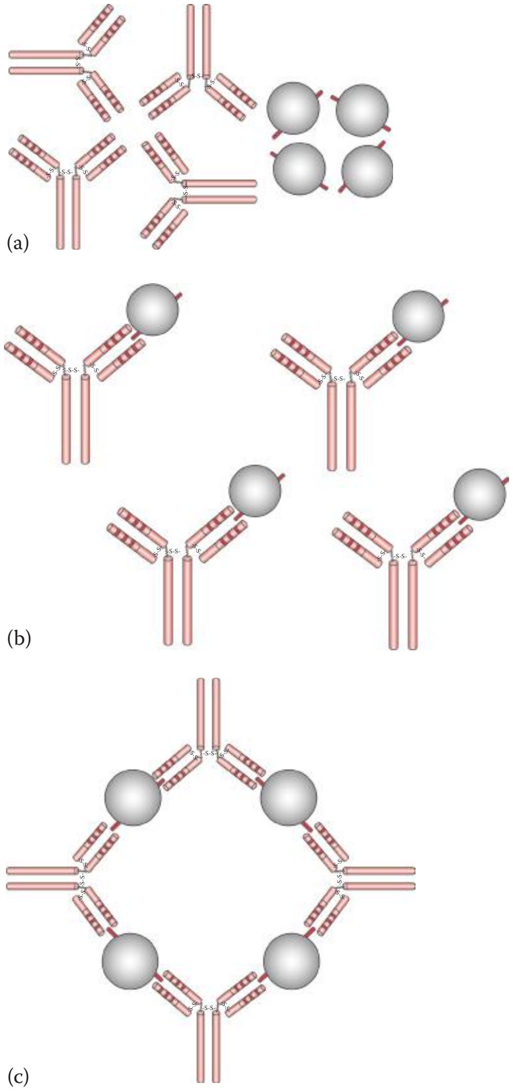

Figure 10.10 Agglutination reaction: (a) antigens are mixed with antibodies, (b) an antigen–antibody complex is formed during initial binding, and (c) the lattice is formed. (© Richard C. Li.)

bacteria, or latex particles), antibodies can bind to the surface antigens and can form cross-links among cells or carriers, causing them to aggregate. This aggregation is referred to as *agglutination*. In agglutination, a visible clumping can be observed as an indicator of the reaction of the antigen and the antibody. If the antigen is located on an erythrocyte, the agglutination reaction is designated *hemagglutination*. Agglutination is a two-step process that includes initial binding and lattice formation (Figure 10.10a).

10.3.2.2.1 Initial Binding

The first step of the reaction involves antigen–antibody binding at a single epitope on the cell surface (Figure 10.10b). This initial binding is rapid and reversible.

10.3.2.2.2 Lattice Formation

The second step involves the formation of a cross-linked network resulting in visible aggregates that constitute a lattice (Figure 10.10c). This involves an antibody binding to multiple epitopes because each antibody has two binding sites and antigens are multivalent. Lattice formation is a much slower process than the initial binding step. The cross-linking of cells requires physical contact. Additionally, an antibody must bind to epitopes on two different cells. The ability to cross-link cells depends on the nature of the antibody. Antibodies that produce such reactions are often called *agglutinins*.

Additionally, a *complete antibody* is capable of carrying out both primary and secondary interactions that result in agglutination. An antibody that can carry out initial binding but fails to form agglutination is called an *incomplete antibody*. This type of antibody is believed to have only one active antigen-binding site and is thus not capable of agglutination. It is potentially caused by the presence of steric obstruction due to the conformation of the incomplete antibody molecule, preventing the binding of antigens at the second binding site. However, other incomplete antibodies have two active sites but cannot bridge the distance between cells, thus failing to form lattices. Certain antibodies such as IgG are small and lack flexibility at the hinge region, and this may prevent agglutination. In contrast, the large IgM antibodies produce agglutination much more easily than IgG. Agglutination reactions have a wide variety of applications in the detection of antigens and antibodies. Such assays have high degrees of sensitivity and have been used for many years in forensic serology.

Bibliography

Baxter, S.J. and B. Rees, The use of anti-human haemoglobin in forensic serology. *Med Sci Law*, 1974, **14**(3): 159–162.

Dodd, B.E., Some recent advances in forensic serology. *Med Sci Law*, 1972, **12**(3): 195–199.

Fletcher, S. and M.J. Davie, Monoclonal antibody: A major new development in immunology. *J Forensic Sci Soc*, 1980, **20**(3): 163–167.

Giusti, G.V., Leone Lattes: Italy's pioneer in forensic serology. *Am J Forensic Med Pathol*, 1982, **3**(1): 79–81.

Goldsby, R., Kindt, T, and Osborne, B., *Immunology*, 6th edn. 2007, New York: W.H. Freeman and Company.

Lee, H.C., et al., Bits and pieces: Serology, criminalistics, anthropology and odontology. *J Forensic Sci Soc*, 1991, **31**(2): 293–296.

Patzelt, D., History of forensic serology and molecular genetics in the sphere of activity of the German Society for Forensic Medicine. *Forensic Sci Int*, 2004, **144**(2–3): 185–191.

Tietz, N. (ed.), *Fundamentals of Clinical Chemistry*, 1987. Philadelphia, PA: W.B. Saunders Company.

11

Serology Techniques

Past, Current, and Future

11.1 Introduction to Forensic Serology

11.1.1 The Scope of Forensic Serology

Forensic serology is the component of forensic biology that deals with the examination and identification of biological evidence. In particular, it focuses on determining the presence and identification of various bodily fluids such as blood, semen, and saliva in a questioned sample.

Bodily fluid stains are commonly associated with violent criminal cases. For instance, the identification of blood evidence (Chapter 12) is often necessary for investigating cases involving homicide, aggravated assault, sexual assault, and burglary. Proving the presence of blood has probative value or may corroborate allegations of violent acts. This evidence can also be used for investigative purposes. For example, the forensic DNA analysis of evidence such as a victim's blood on a suspect's weapon can establish a link between a victim and a suspect. The identification of semen (Chapter 14) and saliva (Chapter 15) is important for the investigation of a sexual assault case. For instance, a stain from a victim's clothing would be processed with forensic DNA analysis in a sexual assault case so that the DNA profile of the stain could be compared with that of an alleged suspect. Matching DNA profiles prove that the suspect's DNA was found on the victim's clothing, establishing a link between the suspect and the victim. Additionally, the identification of a biological stain as a semen stain through forensic serology testing proves that semen was found on the clothing taken from the victim and that the suspect is the source of the DNA from the semen stain. This evidence can then be used in court to support allegations that a sexual act occurred. Likewise, the presence of a suspect's saliva stains on a victim's genital area may corroborate an alleged oral copulation. The identification of bodily fluids other than blood, semen, and saliva will be discussed in Chapters 16 and 17.

11.1.2 Class Characteristics and Individual Characteristics of Biological Evidence

The identification of an unknown fluid sample is based on a comparison of the class characteristics (Section 3.2.1) of a sample with known standards of its class. Forensic identification typically involves bodily fluids, such as blood, semen, and saliva. If the presence of the bodily fluid is confirmed, the individual characteristics (Section 3.2.2) of the biological evidence are then determined to find out whether or not a bodily fluid sample has come from a particular individual. Today, this analysis can be achieved through forensic DNA analysis. The most commonly utilized forensic DNA analysis is the short tandem repeat (STR) analysis used for human identification. Additionally, Y-chromosomal STR analysis is often utilized for the investigation of sexual assault crimes. Furthermore, mitochondrial DNA analysis is used for the identification of human remains. Therefore, establishing the probative value of a sample requires both the identification of its class characteristics and the individualization of its contributor. Therefore, the identification of bodily fluids cannot be replaced by forensic DNA analysis.

11.1.3 Presumptive and Confirmatory Assays

The identification of bodily fluids can be carried out using presumptive and confirmatory assays to identify the type of bodily fluid in question. The advantages of *presumptive assays* are that these assays are sensitive, rapid, and simple. A positive reaction of a presumptive assay indicates the possibility of the presence of the bodily fluid in question. However, presumptive assays are not very specific. Therefore, they should not be considered conclusive for the presence of a type of bodily fluid. In contrast, a negative assay indicates that the questioned bodily fluid is absent. Thus, presumptive assays can be used as a screening method and for narrowing down bodily fluid stains prior to other types of analyses, such as forensic DNA testing. Moreover, these assays can be used as a search method to locate bodily fluid stains at the crime scene. Additional assays, such as confirmatory assays, should be conducted afterward if necessary.

Confirmatory assays are more specific for the bodily fluid in question. These assays are utilized to identify bodily fluids with higher certainty than presumptive assays. For example, bloodstains are commonly associated with criminal investigations. A reddish-brown stain that has been identified through visual examination is usually tested by using presumptive assays. If the result of the presumptive assay on the alleged bloodstain is positive, the stain is then further analyzed by forensic DNA analysis. This approach *indicates* the presence of blood. Confirmatory assays are performed when a sample has to be *identified* as blood (Figure 11.1). Additionally, the human or animal origin of blood evidence can be determined if necessary and the techniques to do so will be discussed in Chapter 13.

11.1.4 Primary and Secondary Binding Assays

Traditionally, the detection and measurement of the antigen–antibody binding reactions serve as the bases of forensic serology. These assays fall into two categories: primary and secondary binding assays. Recall that the primary binding assays involve the initial binding between a single epitope of an antigen and a single binding site of an antibody (see Section 10.3.1). Therefore, they are very sensitive. Although the secondary binding assays are less sensitive than the primary binding assays, they are easier to perform than primary assays. The secondary assays consist of precipitation-based and agglutination-based assays. Precipitation-based assays have been used for species identification. Agglutination-based assays are more sensitive

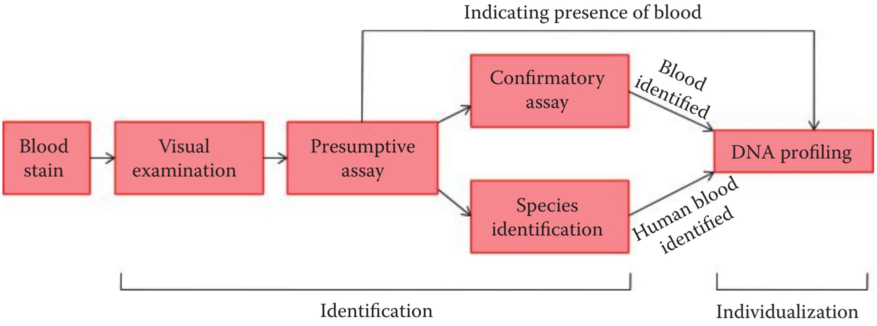

Figure 11.1 An example of work flow for processing blood samples. (© Richard C. Li.)

than precipitation-based assays. Agglutination-based assays, which detect antigens located on the surface of cells or carriers, are normally applied to blood group typing. Recently, emerging techniques such as RNA, proteomic, and DNA methylation assays, and fluorescence as well as Raman spectroscopy can potentially be used for the identification of bodily fluids.

11.2 Primary Binding Assays

11.2.1 Enzyme-Linked Immunosorbent Assay (ELISA)

An enzyme-linked immunosorbent assay (ELISA) is an immunoenzyme assay that can be used to detect and measure the antibody or antigen in question. The most common ELISA that is used in forensic serology is the antibody-sandwich ELISA (Figure 11.2). It is utilized to detect the prostate-specific antigen (PSA) to identify seminal stains and amylase for the identification of saliva (Chapters 13 and 14).

An antibody (usually monoclonal) coating is formed by nonspecific adsorption onto a solid phase such as the wells of a polystyrene plate. A sample containing the antigen to be tested is then added and binds to the solid-phase antibody. Subsequently, a second antibody is added to form an antibody–antigen–antibody sandwich complex. The second antibody binds to different epitopes of the antigen. Next, an enzyme-labeled antiglobulin (Chapter 10) is added to bind the sandwich. Subsequently, the excess enzyme-labeled antiglobulin is removed by washing and the bound antiglobulin can be detected.

A number of enzymes such as alkaline phosphatase and horseradish peroxidase have been used as reporting enzymes (Chapter 9) to label the antiglobulin for ELISA. The enzyme catalyzes the substrate and produces colorimetric or fluorometric signals. The intensity of the signals can

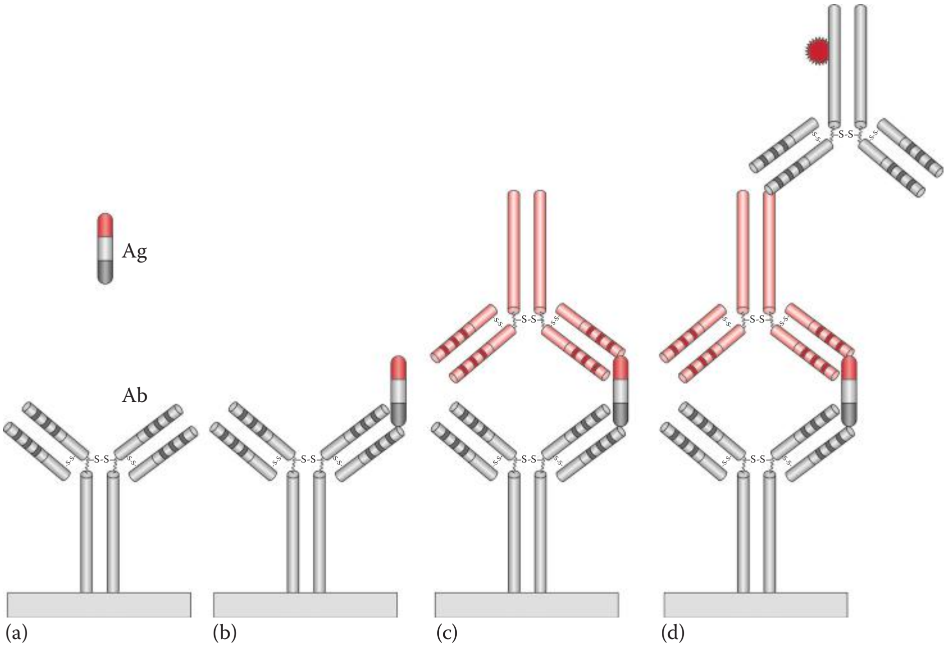

Figure 11.2 ELISA assay. (a) Sample containing Ag (antigen) is applied to a sample well where Ab (antibody) is immobilized. (b) Ag binds to an immobilized Ab to form an Ag–Ab complex. (c) A second Ab to a different epitope is added to form an Ab–Ag–Ab sandwich. (d) Labeled antiglobulin binds to the sandwich. The bound antiglobulin can be detected by various reporting schemes. (© Richard C. Li.)

be detected spectrophotometrically and is proportional to the amount of bound antigen. The amount of antigen can be quantified by comparing the standard with known concentrations. Recall that antiglobulin can recognize an epitope that is common to all isotypes within an antibody class (Section 10.1.2.3); the enzyme-labeled antiglobulin is a universal reagent regardless of the antibody used, as long as they are of the same class.

Alternatively, antibodies in a sample can also be detected and may be quantified by an ELISA system in which the antigen is bound to a solid phase instead of the antibody. After an antibody in a sample binds to the solid-phase antigen, an enzyme-labeled antiglobulin is added to bind to the bound antibody. The bound antiglobulin can be detected and measured by the addition of an enzyme substrate. The enzymatic catalytic reaction is similar to the antibody sandwich ELISA procedure previously described.

11.2.2 Immunochromatographic Assays

Figure 11.3 depicts a test using an immunochromatographic membrane device. A dye-labeled monoclonal antibody is contained in a sample well and a polyclonal antibody for the antigen (or a second monoclonal antibody to a different epitope of the antigen) is immobilized onto a test zone of a nitrocellulose membrane. An antiglobulin that recognizes the antibody is immobilized onto a control zone.

The assay is carried out by loading a sample into the sample well. The antigen in the sample binds to the dye-labeled antibody already in the sample well to form an antigen–antibody complex. The complex then diffuses across the nitrocellulose membrane until it reaches the test zone. The antibody immobilized at the test zone traps the antigen–antibody complex to form an antibody–antigen–antibody sandwich. The presence of the antigen in the sample results in a colored vertical line at the test zone. (Figure 11.4)

The immunochromatographic device also utilizes a control zone to ensure that the device works properly and that the sample has diffused completely along the test strip. Unbound

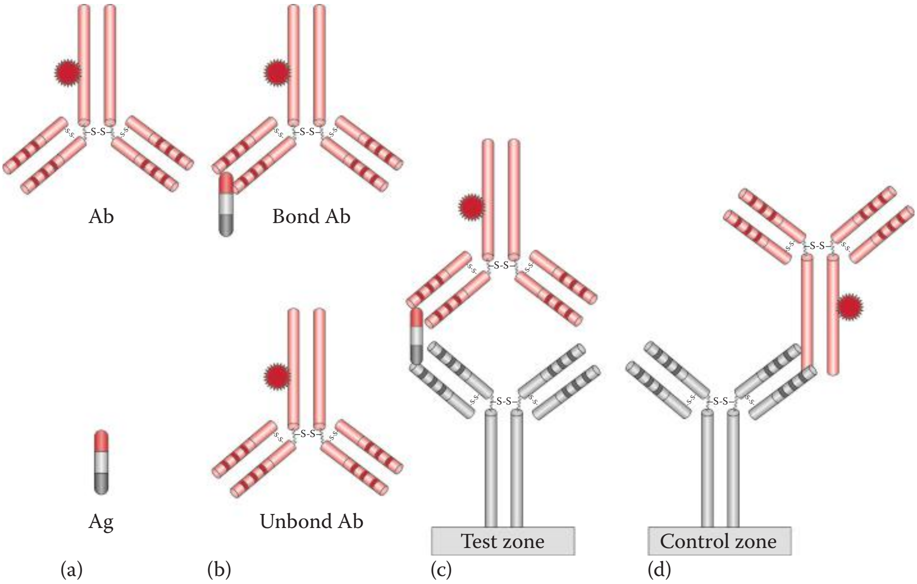

Figure 11.3 Immunochromatographic assay. (a) Sample containing Ag (antigen) is loaded in a sample well. (b) Ag binds to a labeled Ab (antibody) to form a labeled Ab–Ag complex. (c) At the test zone, the labeled Ab–Ag complex binds to an immobilized Ab to form a labeled Ab–Ag–Ab sandwich. (d) At the control zone, a labeled Ab binds to an immobilized antiglobulin and is captured at the control zone. (© Richard C. Li.)

monoclonal antibodies diffuse across the membrane until they reach the control zone where they are trapped by the immobilized antiglobulin. This antibody–antiglobulin complex at the control zone also results in a colored vertical line (Figure 11.4). The test is considered valid only if the line in the control zone is observed. The presence of an antigen results in a line at both the test zone and the control zone, while the absence of an antigen results in a line in the control zone only (Figure 11.4). This method is rapid and simple and thus can be used as a screening test in laboratories and as a field test at crime scenes to identify semen, saliva, and species (Table 11.1).

However, a false-negative result may be obtained if a sample contains a very high concentration of an antigen (Figure 11.5). Under this condition, the dye-labeled antibody will become saturated with the antigen to form antigen–antibody complexes. The unbound antigen will diffuse along with the

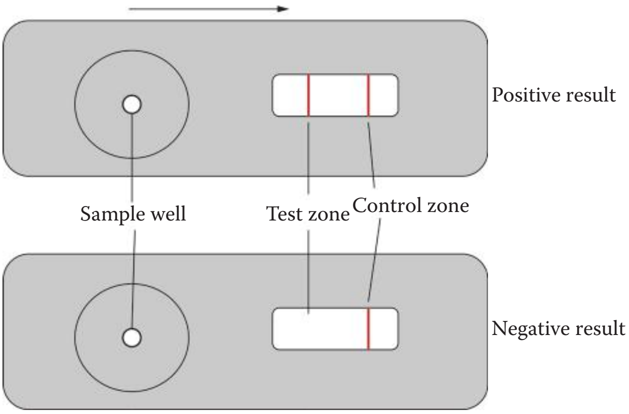

Figure 11.4 Immunochromatographic device. Positive and negative results are shown. (© Richard C. Li.)

Table 11.1 Common Immunochromatographic Assays for Forensic Applications

Assay	Antigen	Labeled Antibody	Immobilized Antibody	Forensic Application
ABAcard® HemaTrace® (Abacus Diagnostics)	Hemoglobin (Hb)	Monoclonal antihuman Hb antibody	Polyclonal antihuman Hb antibody	Blood and species identification
RSID™-Blood (Independent Forensics)	Glycophorin A (GPA)	Monoclonal antihuman GPA antibody	Monoclonal antihuman GPA antibody	Blood and species identification
RSID™-Saliva (Independent Forensics)	Human salivary α-amylase (HAS)	Monoclonal antihuman HAS antibody	Monoclonal antihuman HAS antibody[a]	Saliva identification
One-Step ABAcard PSA® (Abacus Diagnostics)	Prostate-specific antigen (PSA)	Monoclonal antihuman PSA antibody	Polyclonal antihuman PSA antibody	Semen identification
RSID™-Semen (Independent Forensics)	Semenogelin (Sg)	Monoclonal antihuman Sg antibody	Monoclonal antihuman Sg antibody[a]	Semen identification

[a] The epitope recognized by the immobilized antibody is different from that of the labeled antibody.

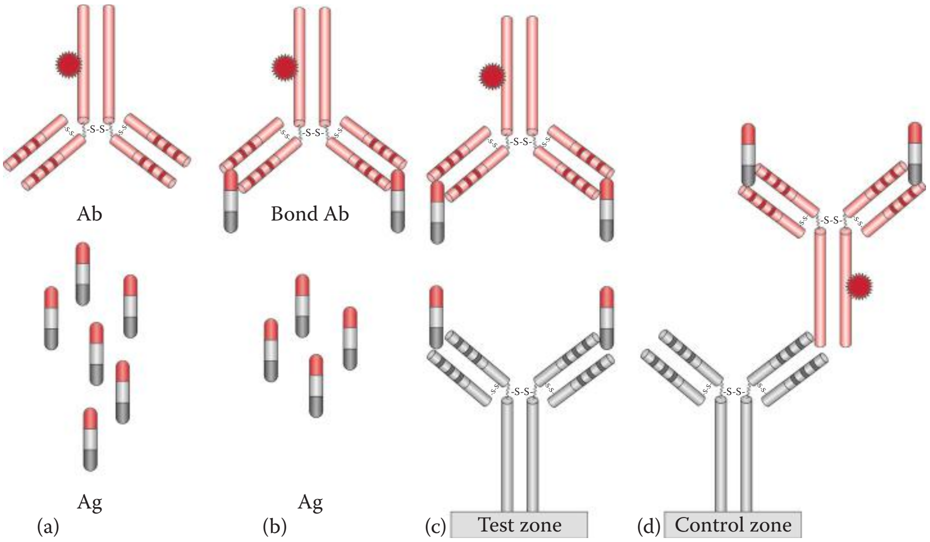

Figure 11.5 High-dose hook effect of immunochromatographic assay. (a) Ag (antigen) in sample is loaded in the sample well. (b) Ag binds to a labeled Ab (antibody) to form a labeled Ag–Ab complex. (c) At the test zone, free Ag binds to an immobilized Ab to form an unlabeled Ab–Ag complex and prevents the formation of a labeled Ab–Ag–Ab sandwich. (d) At the control zone, a labeled Ab binds to immobilized antiglobulin and is captured at the control zone. (© Richard C. Li.)

antigen–antibody complex toward the test zone. At the test zone, the unbound antigen will compete with the antigen–antibody complexes for the antibody immobilized at the test zone. Since the antigen is in excess, the unbound antigen binds to the immobilized antibody, preventing the formation of the antibody–antigen–antibody sandwich. The resulting reading appears as a negative result. This artifact is known as the *high-dose hook* effect. To prevent the high-dose hook effect, a smaller volume of sample can be applied, or the sample can be diluted to reduce the amount of antigen applied.

11.3 Secondary Binding Assays

11.3.1 Precipitation-Based Assays

These techniques are based on the precipitation reaction and are used primarily for species identification in forensic laboratories.

11.3.1.1 Immunodiffusion

Immunodiffusion is a passive method in which an antigen or an antibody or both are allowed to diffuse and therefore a gradient, from low to high concentration, is established for an antigen or an antibody or both. As a result, precipitation occurs due to the interaction between an antigen and an antibody. The assay can be carried out in a liquid or a semisolid medium such as agarose gel. The semisolid medium can stabilize the diffusion process and reduce interference such as convection, which is the movement of molecules within liquids. The two types are single immunodiffusion and double immunodiffusion.

11.3.1.1.1 Single Immunodiffusion

A concentration gradient is established for either an antigen or an antibody. One serology technique based on this principle is radial immunodiffusion—a single diffusion method in which a concentration gradient is established for an antigen. The antibody is uniformly distributed in the gel matrix (Figure 11.6). The antigen is loaded into a sample well and allowed to diffuse from the well into the gel until a precipitation reaction occurs. The precipitate ring around the well is observed; the area within the ring of precipitate is proportional to the amount of antigen loaded in the well. Standards using known concentrations of antigen can be included in the same assay along with the samples, and a standard curve can be plotted. The amounts of antigen in samples can then be quantified by comparing the results with the standard curve.

11.3.1.1.2 Double Immunodiffusion

The second type of assay is double diffusion in which a concentration gradient is established for both an antigen and an antibody. The most common examples are the ring assay and the Ouchterlony assay. The *ring assay* can be performed in a test tube or capillary tube (Figure 11.7). An antiserum, a denser phase, is placed in a small tube. An antigen solution is carefully layered on top of an antibody solution without mixing. Both the antigen and the antibody will diffuse toward each other. In a positive reaction, a ring of precipitate can be observed at the interface of the two solutions. A negative reaction is indicated by a lack of precipitation. The assay requires positive and negative controls along with questioned samples.

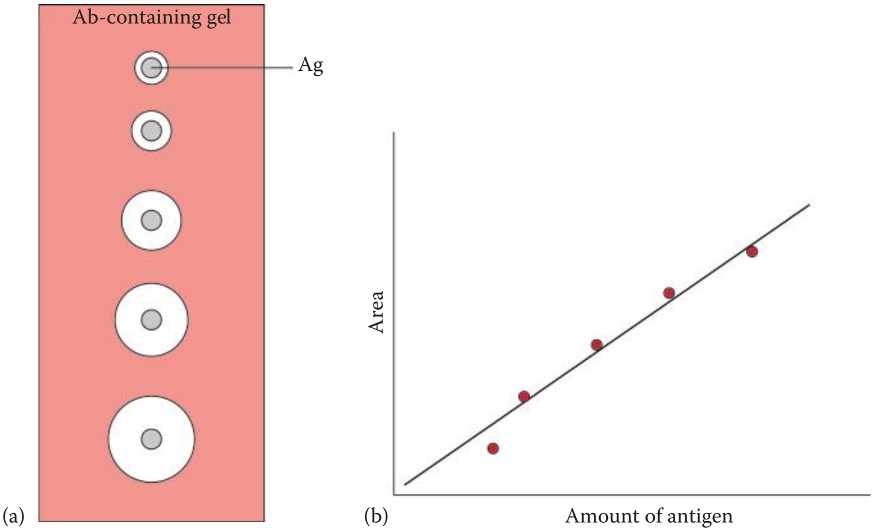

Figure 11.6 Radial immunodiffusion assay. (a) Immunodiffusion of antigens from a well into an antibody-containing gel. (b) Standard curve based on results of standards with known amounts of antigens. (© Richard C. Li.)

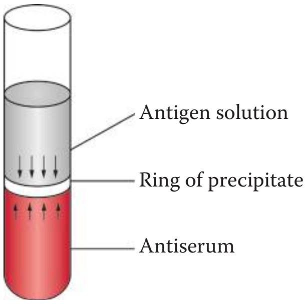

Figure 11.7 Ring assay. (© Richard C. Li.)

The *Ouchterlony assay* is named after the Swedish immunologist, Örjan Ouchterlony, who developed it. The assay can be performed in an agarose gel supported by a glass slide or polyester film (Figure 11.8). Wells are created by punching holes in the gel layer at desired locations. Often, a pattern with six wells surrounding a center well is used. The antibody is loaded in the central well while the questioned samples and the controls are loaded in the surrounding wells. The double diffusion of the antigen and the antibody from the wells is allowed to occur during incubation. If the reaction is positive, a precipitate line between wells can be observed at the end of incubation. The precipitate can be stained, enhancing visibility to aid observation. A single assay can compare more than one antigen to determine whether the antigens in question react the same way or differently with the antibody (see Chapter 7). This method is sometimes used to determine whether samples have come from the same or different origins.

11.3.1.2 Immunoelectrophoretic Methods

Diffusion techniques can be combined with electrophoresis to enhance test results. Electrophoresis separates molecules according to the differences in their electrophoretic mobility.

11.3.1.2.1 Immunoelectrophoresis

Immunoelectrophoresis (IEP) is a two-step procedure that can analyze a wide range of antigens. This technique uses electrophoresis to separate the antigen mixture prior to immunodiffusion (Figure 11.9). In the first step, the antigens in a sample are separated using agarose gel electrophoresis.

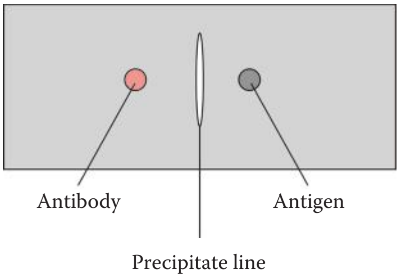

Figure 11.8 Ouchterlony assay. (© Richard C. Li.)

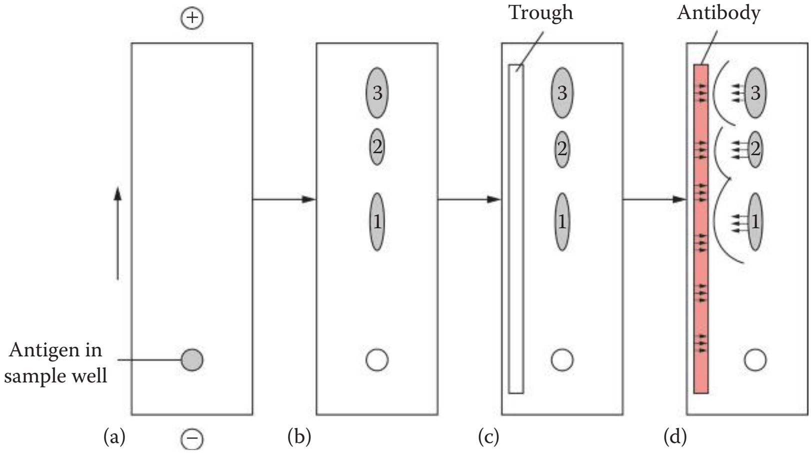

Figure 11.9 IEP assay. (a) Electrophoresis of antigens is carried out. (b) Various antigens are separated after electrophoresis. (c) Trough is cut. (d) Antibody is applied to the trough allowing diffusion to occur and forming precipitate lines. +, anode; −, cathode. (© Richard C. Li.)

Then, a trough is cut in the gel parallel to the array of antigens separated by electrophoresis. In the second step, an antibody is loaded in the trough and the gel is incubated for double diffusion. An arc-shaped precipitate line can be observed for a positive reaction. Multiple precipitate lines can occur if more than one antigen reacts with the antibody. The shapes, intensities, and locations of the precipitate lines of a known control and a questioned sample can be compared.

11.3.1.2.2 Crossed Immunoelectrophoresis

Crossed immunoelectrophoresis (CRIE), also known as two-dimensional IEP, is a modification of IEP. The first step utilizes electrophoresis to separate antigens contained in a sample (Figure 11.10). A strip of gel containing separated antigens is cut for the second round of electrophoresis. The gel including the gel strip is turned at a 90° angle and is further separated by a second-dimension electrophoresis. This drives the antigens from the gel strip into an agarose gel that contains uniformly distributed antibodies. Following the second-dimension electrophoresis, an arc-shaped precipitate line is formed. The area of the arc can be measured, and a sample can be identified by comparison with a known standard. This technique is more sensitive than IEP.

11.3.1.2.3 Rocket Immunoelectrophoresis

An antibody-containing agarose gel is used. The antigen is loaded into the well (Figure 11.11). Electrophoresis then drives the antigen from the well into the agarose gel. In a positive reaction,

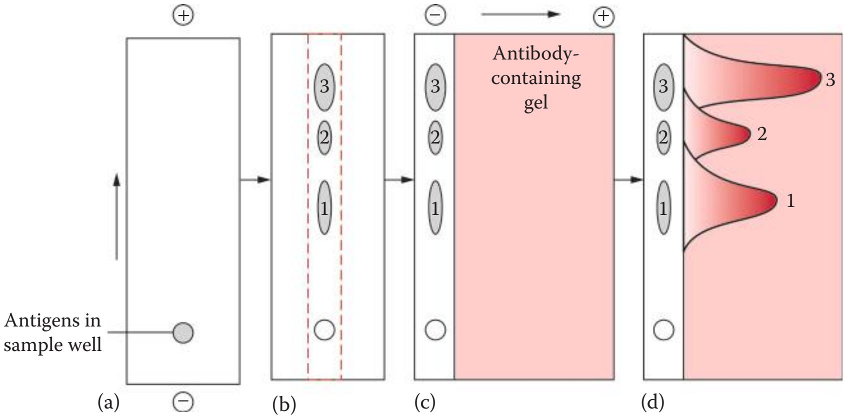

Figure 11.10 CRIE assay. (a) Electrophoresis of antigens is carried out. (b) Various antigens are separated after electrophoresis. A strip of gel containing separated antigens is excised and then used for second-dimension electrophoresis. (c) As shown, second-dimension electrophoresis is carried out to drive the antigens into an antibody-containing gel; (d) arc-shaped precipitate lines are formed as a result. (© Richard C. Li.)

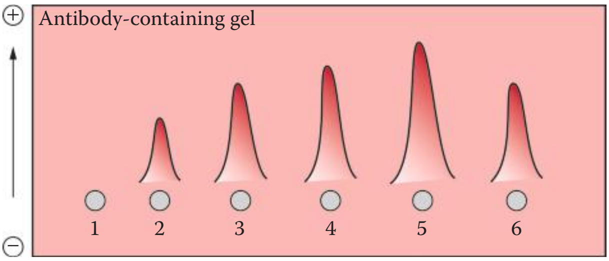

Figure 11.11 Rocket immunoelectrophoresis assay. Antigens are driven from the wells into an antibody-containing gel forming precipitate lines. (© Richard C. Li.)

a rocket-shaped precipitate line can be observed. The height of the rocket is in proportion to the amount of antigen in the sample. Quantitation can be achieved by comparing standards and the sample in the same gel.

11.3.1.2.4 Crossed-Over Immunoelectrophoresis

This technique is also known as counterimmunoelectrophoresis (CIE). Two arrays of opposing wells are created by punching holes in the agarose gel (Figure 11.12). The antibody and samples are loaded in opposing wells arranged by pairs. Electrophoresis is used to drive the antigen and the antibody toward each other. The wells containing the antibody should be proximal to the anode and the wells containing the samples should be proximal to the cathode. During gel electrophoresis, the antigen, which is usually negatively charged, migrates toward the anode. The antibody migrates in the opposite direction as a result of electroendosmosis—a phenomenon in which the movement of molecules is caused by fluid flow. A precipitate line is formed between the opposing wells if the antigen reacts with its specific antibody.

11.3.2 Agglutination-Based Assays

Agglutination reactions can be used as forensic serological assays such as for blood group typing (Chapter 18) and menstrual blood identification (Chapter 16). Agglutination assays are qualitative, indicating the absence or presence of antigens or antibodies. Semiquantitative results can be obtained by titration (diluting the antigen or antibody). Many types of agglutination reactions are available; only those used in forensic serology are discussed next.

Direct agglutination assays involve reactions in which an antibody interacts with antigens originally located on cell surfaces. In a *hemagglutination* reaction (Figure 11.13a), an antibody binds to the antigens located on erythrocytes. This method is used for the identification of blood types, for example, the testing of erythrocytes for ABO blood group typing. The assay can be carried out on a glass slide and the agglutination, as indicated by forming cell clumps, can be observed under a microscope (Chapter 18). The assay can also be carried out in a test tube. In a positive reaction, agglutinated cells are formed on the bottom of the test tube. Test tubes can also be centrifuged and then swirled to determine whether the cell clumps can be resuspended. Agglutinated cells cannot be resuspended. In a negative reaction, unagglutinated cells can be resuspended.

In *agglutination inhibition assays*, the presence of an antigen in question is indirectly detected (Figure 11.13b). If a known antigen is added to a mixture consisting of antigen-containing cells and antibodies, the added antigen will compete with the antigen located on the cell surfaces for antibody binding and inhibit the agglutination reaction. Another indirect agglutination assay is the absorption–elution assay (Chapter 18), which can be used for ABO blood group typing.

Passive agglutination assays are different from direct agglutination assays, in which the antigen is coated on the surface of carrier cells such as tannic acid–treated sheep erythrocytes. Tannic acid is considered to be a fixative. Treating with tannic acid can stabilize the carrier cells

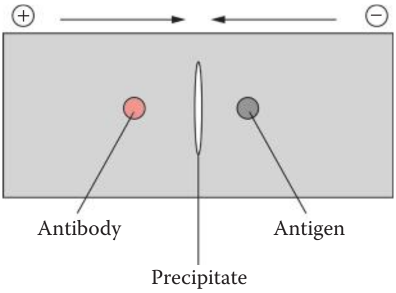

Figure 11.12 Crossed-over immunoelectrophoresis assay. (© Richard C. Li.)

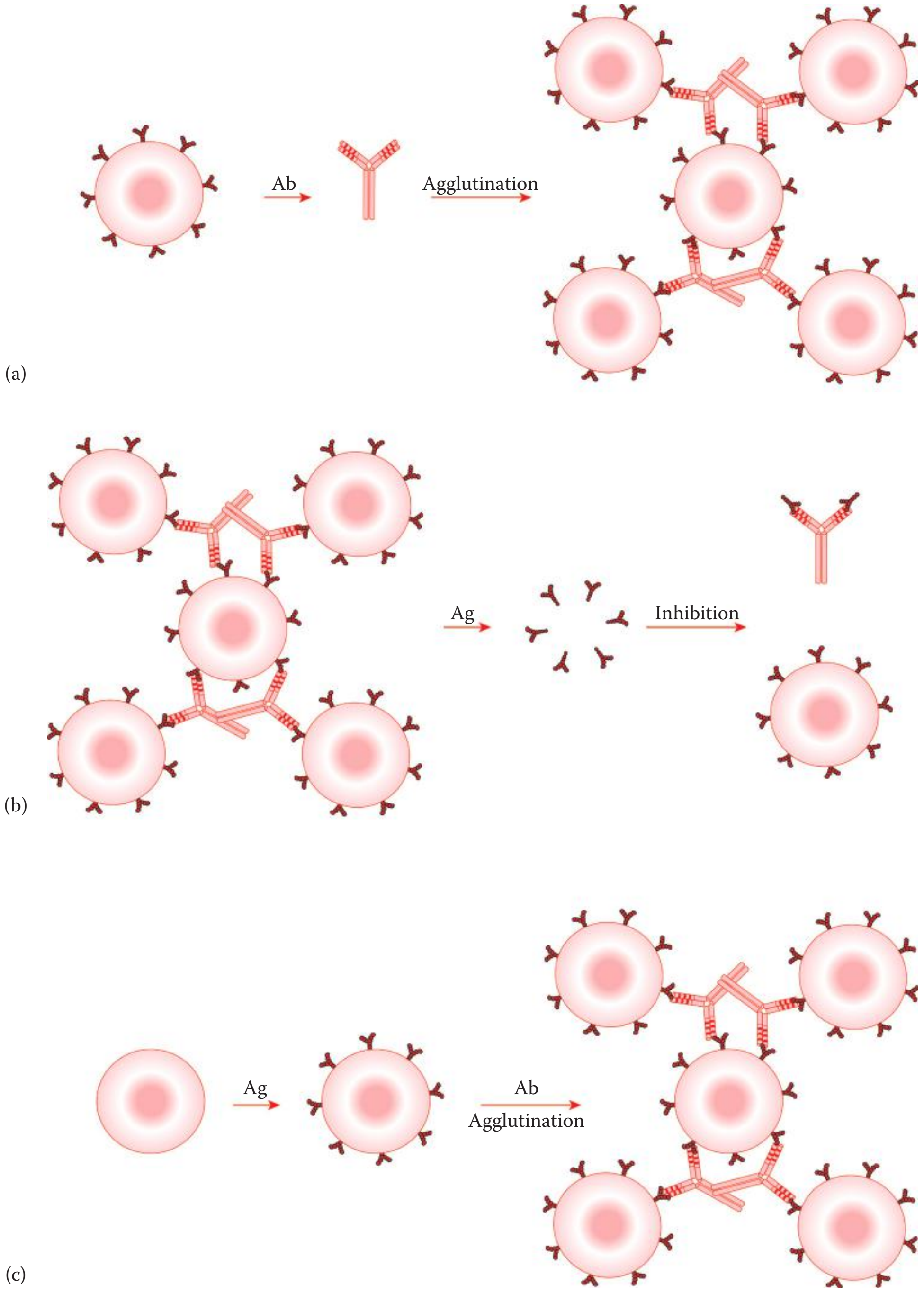
Ab
Agglutination
(a)
Ag
Inhibition
(b)
Ag
Ab
Agglutination
(c)

Figure 11.13 Agglutination-based assays. (a) Direct agglutination assay and (b) agglutination inhibition assay. Hemagglutination reactions are shown; antigens are located on erythrocytes. (c) Passive agglutination assay. (© Richard C. Li.)

for a subsequent coating of antigens. The carrier cells are then incubated with samples containing antigens. A coating of antigens can be formed on the surfaces of carriers. Such antigen-coated carrier cells can then be agglutinated using an antibody that is specific to the absorbed antigens (Figure 11.13c).

11.4 DNA Methylation Assays for Bodily Fluid Identification

In the eukaryotic genome, methylation occurs at the cytosine residues commonly in the CpG dinucleotide sequences of both DNA strands. The "p" in CpG refers to the phosphodiester bond between cytosine and guanine. Cytosine methylation is carried out *in vivo* by methyltransferase (DNMT). DNMT catalyzes the transfer of a methyl group (CH_3) from S-adenosylmethionine (SAM) to the C^5 position of cytosine (Figure 11.14). As a result, cytosine is converted to 5-methylcytosine. Cytosine methylation is observed throughout the eukaryotic genomes in both the coding and intergenic regions located between genes. However, cytosine methylation is rare at CpG islands, which are stretches of DNA containing a high frequency of CpG dinucleotides in the regulatory regions of genes (Figure 11.15). It is implicated that cytosine methylation alters the chromatin structure and causes chromatin condensation. As a result, gene expression is usually inhibited by cytosine methylation.

Genomic loci, known as *tissue-specific differentially methylated regions* (TDMRs), have been identified. TDMRs show differential DNA methylation patterns between different tissues. For example, semen-specific TDMRs are consistently hypomethylated in spermatozoa, which can be used to identify semen samples. Other TDMRs display varying degrees of methylation in different types of bodily fluids. This suggests that differential DNA methylation assays may be potentially useful for bodily fluid identification.

The detection of DNA methylation can be carried out by a number of methods. *Methylation-sensitive restriction enzyme digestion polymerase chain reaction* (MSRE-PCR) can be used for the

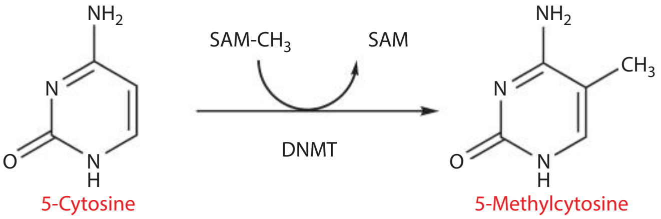

Figure 11.14 Eukaryotic cytosine methylation catalyzed by methyltransferase.

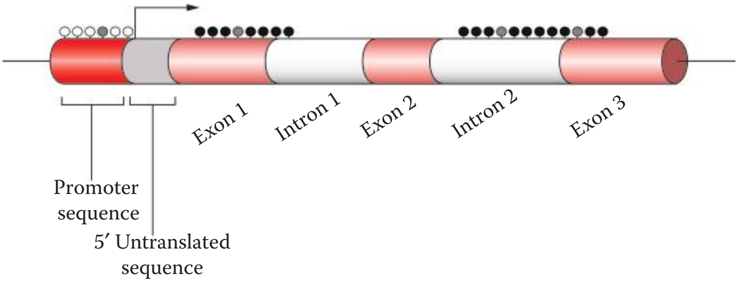

Figure 11.15 Methylation at the cytosine residues in CpG dinucleotide sequences. Note that the CpG dinucleotide sequences are usually unmethylated in the promoter regions. Black circle: methylated. Clear circle: unmethylated. Gray circle: partial methylated. (© Richard C. Li.)

rapid detection of DNA methylation. This method utilizes the *methylation-sensitive restriction enzyme* (MSRE). MSREs cleave DNA at unmethylated restriction sites, but they are not able to cleave once a cytosine residue is methylated. Thus, the methylated DNA remains intact. After PCR amplification, the methylated DNA is amplified and the amplicon is detected. In contrast, the unmethylated DNA is cleaved by the MSRE and cannot be amplified (Figure 11.16). *Bisulfite sequencing* is another widely used method. The genomic DNA is treated with sodium bisulfite, which catalyzes the hydrolytic deamination of unmethylated cytosines (Figure 11.17). As a result, unmethylated cytosines are converted to uracils, while methylated cytosines are resistant to conversion and remain unchanged. Next, uracils are replaced by thymines during PCR. After sequencing, the methylation sites can be deduced at a single-nucleotide resolution by comparing them with the reference sequence (Figure 11.18). *Methyl-DNA immunoprecipitation* (MeDIP) is a useful method for isolating methylated DNA. In this method, DNA is first sheared into fragments. The fragmented DNA is then treated with an antimethylcytosine antibody that binds to methylated cytosines. Methylated DNA fragments can then be isolated by immunoprecipitation for further analysis (Figure 11.19).

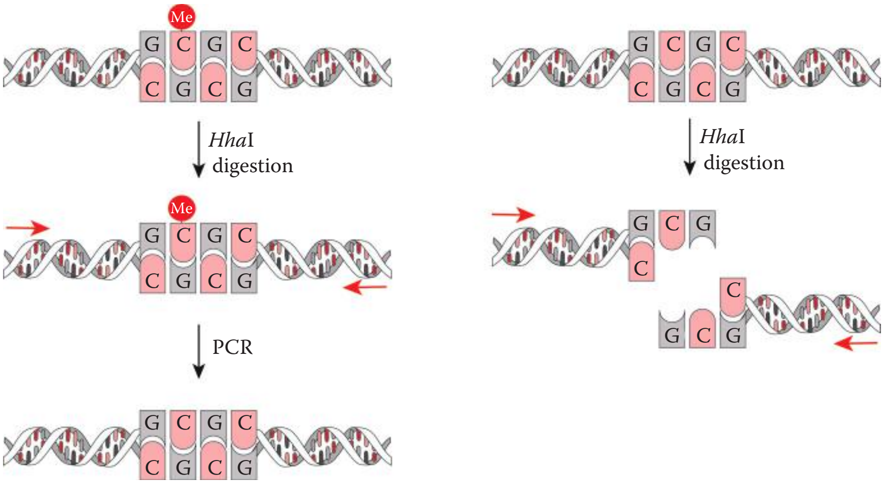

Figure 11.16 Methylation-sensitive restriction enzyme digestion PCR (MSRE-PCR) for the detection of DNA methylation. This method utilizes the methylation-sensitive restriction enzyme (MSRE) such as *Hha*I. *Hha*I cleaves DNA at the unmethylated sequence GCGC but is not able to cleave once a cytosine residue is methylated (GCmGC). The *Hha*I-treated DNA is then PCR amplified. The *Hha*I uncleaved DNA results in an amplicon, indicating the presence of a methylated cytosine in the sequence. The *Hha*I cleaved DNA cannot be amplified, indicating the absence of methylated cytosine in the sequence. Me, 5-methylcytosine site; red arrow, PCR primers. (Adapted from Frumkin, D. et al., *Forensic Sci Int Genet*, 5, 517–524, 2011.)

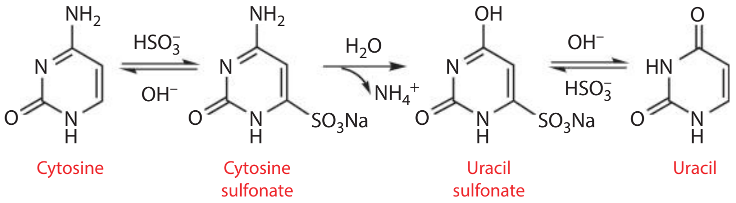

Figure 11.17 Bisulfite conversion of 5-methylcytosine.

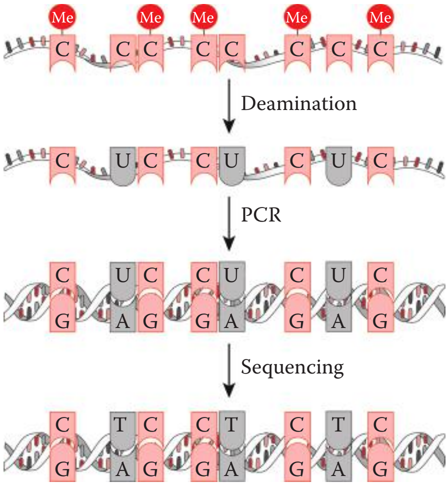

Figure 11.18 Schematic illustration of bisulfite sequencing of cytosine methylation. In a deamination reaction, unmethylated cytosines are converted to uracil. PCR and DNA sequencing are then carried out. The retention of C indicates that the site was methylated; the conversion of C to T indicates that the site was unmethylated in the DNA sample. Me, 5-methylcytosines site. (© Richard C. Li)

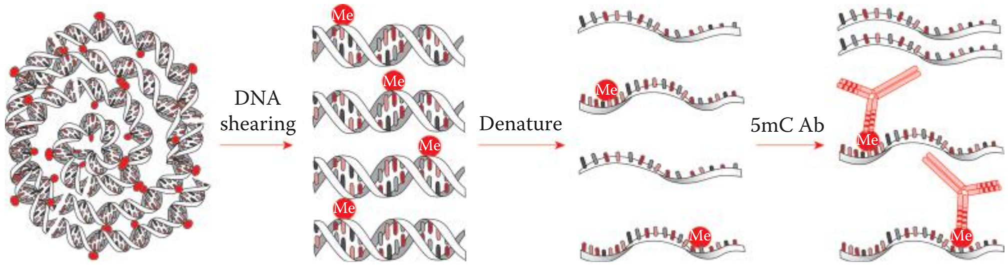

Figure 11.19 Methylated DNA immunoprecipitation (MeDIP) for the detection of methylated DNA. Genomic DNA is randomly and mechanically sheared into fragments. Fragments are then denatured to single-stranded DNA. The fragments containing the methylated region can be immunoprecipitated using an antibody that specifically recognizes 5-methylcytosine. The methylated DNA sequence in the immunoprecipitated fragments can then be determined by DNA sequencing and methylation analysis. Me, 5-methylcytosine site; 5mC Ab, 5-methylcytosine antibody. (© Richard C. Li.)

11.5 Forensic Applications of RNA-Based Assays and RNA Profiling

11.5.1 Messenger RNA-Based Assays

mRNA-based assays have been developed to identify bodily fluids for forensic investigation. In forensic pathological investigation, they can potentially be used for wound age estimation and the age of biological stains. The assays are based on the expression of certain genes in certain cell or tissue types. Candidate tissue-specific genes that may be useful for the identification of bodily fluids have been identified. Thus, the techniques that are used in the identification of bodily fluids are based on the detection of specific types of mRNA that are expressed exclusively in certain cells. The *tissue-specific genes* that are utilized for bodily fluid identification are summarized in Table 11.2. Additionally, *reference genes* that are constitutively expressed housekeeping genes are utilized as internal controls. The amount of mRNA can be assessed by normalizing the target gene to the expression level of the reference genes. Compared with conventional assays that are used for bodily fluid identification, the mRNA-based assay has higher specificity and is amenable to automation. However, one limitation is that the mRNA is unstable due to degradation by

Table 11.2 Representative Markers of mRNA-Based Assays for Bodily Fluid Identification

Bodily Fluid	Gene Symbol	Description
Blood	*ALAS2*	Aminolevulinatesynthase 2
	AMICA1	Adhesion molecule, interacts with CXADR antigen 1
	ANK1	Ankyrin1
	CD3G	CD3 gammamolecule
	CD93	CD93 molecule
	HBA1	Alpha 1 hemoglobin
	HBB	Beta hemoglobin
	PBGD	Porphobilinogen deaminase
	SPTB	Beta spectrin
Saliva	*HTN3*	Histatin 3
	MUC7	Mucin 7
	STATH	Statherin
Semen	*KLK3*	Kallikrein 3 (prostate-specific antigen)
	PRM1	Protamine 1
	PRM2	Protamine 2
	SEMG1	Semenogelin 1
	TGM4	Transglutaminase 4
Vaginal secretions	*CYP2B7P1*	Cytochrome P450, family 2, subfamily B, polypeptide 7 pseudogene 1
	DKK4	Dickkopf homolog 4
	FUT6	Fucosyltransferase 6
	HBD1	Beta defensin 1
	IL19	Interleukin 19
	MUC4	Mucin 4
	MYOZ1	Myozenin 1
	SFTA2	Surfactant associated 2
Menstrual blood	*MMP7*	Matrix metalloproteinase 7
	MMP11	Matrix metalloproteinase 11
Skin	*CDSN*	Corneodesmosin
	LOR	Loricrin

(*continued*)

Table 11.2 (Continued) Representative Markers of mRNA-Based Assays for Bodily Fluid Identification

Bodily Fluid	Gene Symbol	Description
Reference gene	*18S rRNA*	18S ribosomal RNA
	ACTB	Beta actin
	G6PDH	Hexose-6-phosphate dehydrogenase
	RPS15	Ribosomal protein S15

Sources: Adapted from Bauer, M. and Patzelt, D., *J Forensic Sci*, 47, 1278–1282, 2002; Hanson, E.K. and Ballantyne, J., *Sci Justice*, 53, 14–22, 2013; Lindenbergh, A., Maaskant, P., and Sijen, T., *Forensic Sci Int Genet*, 7, 159–166, 2013; Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., *Forensic Sci Int*, 157, 181–186, 2006.

endogenous ribonucleases. Additionally, bodily fluid stains collected from crime scenes are often exposed to ultraviolet (UV) light, moisture, and high temperature, which can promote mRNA degradation. Nevertheless, the successful detection of mRNA from aged samples is possible.

11.5.2 MicroRNA-Based Assays

The biological function of microRNAs (miRNAs) is to regulate gene expression. The mature miRNA strand associated with the RNA-induced silencing complex (RISC) binds to its target mRNA. In animal cells, the mature miRNA forms a base pairing with its complementary sequence of the *miRNA responsive element* (MRE) within the 3′ untranslated regions (3′UTRs) of the target mRNA (Figure 11.20). A single miRNA can bind different mRNA transcripts encoded by multiple genes. Animal miRNAs usually form a base pairing with their target mRNAs through partial complementary sequences, which lead to translation repression to inhibit protein synthesis. If the base pairing is exactly complementary to its target mRNA sequence, then the cleavage and the degradation of the target mRNA occurs. As a result, the synthesis of the protein encoded by the mRNA is reduced. Thus, miRNAs play important roles in cellular function.

miRNAs have potential applications in forensic identification. Recent studies have revealed that miRNAs can be detected from various bodily fluids such as blood, saliva, semen, vaginal secretions, and menstrual blood. The type of miRNA expressed is not exclusively tissue- or cell-type specific. However, the expression patterns of specific miRNAs are unique for various bodily fluids. Thus, a bodily fluid in question can potentially be identified using multiple differentially expressed miRNAs as markers (Table 11.3). miRNA markers for bodily fluid identification have their advantages compared with those of mRNA. Mature miRNAs are much

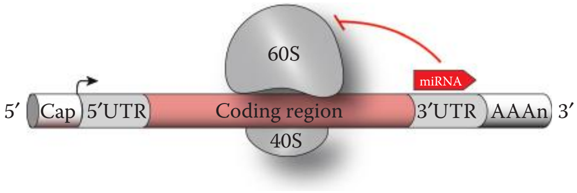

Figure 11.20 Schematic illustration of the roles of miRNA in translational inhibition. In animal cells, miRNAs bind to 3′UTR of their target mRNA with partially complementary Watson–Crick base pairing. As a result, the binding of miRNA to the target mRNA inhibits translation of the target mRNA. Eukaryotic 40S and 60S ribosomal subunits are shown. (© Richard C. Li.)

Table 11.3 Differentially Expressed miRNA Markers for Bodily Fluid Identification

Biological Fluid	miRNA Set
Blood	miR16; miR451
	miR16; miR486
	miR20a; miR106a; miR144; miR185
	miR126; miR150; miR451
Semen	miR10b; miR135b
	miR10a; miR135a; miR943; miR507; miR891a
Saliva	miR205; miR658
	miR200c; miR203; miR205
Vaginal secretions	miR124a; miR372
Menstrual blood	miR214
	miR412; miR451
Reference gene	*RNU6-2* (U6 small nuclear 2 RNA)

Sources: Adapted from Courts, C. and Madea, B., *J Forensic Sci*, 56, 1464–1470, 2011; Hanson, E.K., Lubenow, H., and Ballantyne, J., *Anal Biochem*, 387, 303–314, 2009; Wang, Z. et al., *Forensic Sci Int Genet*, 7, 116–123, 2013; Zubakov, D. et al., *Int J Legal Med*, 124, 217–226, 2010.

smaller, approximately 21–23 nucleotides in length, than mRNA. As a result, miRNAs are less susceptible to degradation and more stable than mRNAs. Additionally, the sensitivity of the miRNA-based method is much higher than that of the mRNA-based methods.

11.6 Proteomic Approaches Using Mass Spectrometry for Bodily Fluid Identification

The forensic application of proteomic approaches is to identify biomarker proteins derived from bodily fluids to determine the type of bodily fluid in question for forensic investigations. Mass spectrometry (MS) is a highly sensitive and rapid technique for protein identification from complex biological samples.

11.6.1 Mass Spectrometric Instrumentation for Protein Analysis

A mass spectrometer is an analytical technique to identify a molecule by measuring the ratio of the mass (m) to the charge (z) of a charged molecule (Figure 11.21). A typical mass spectrometer instrument analyzes ionized molecules while in its gas phase. A mass spectrometer consists of an *ion source* that converts analytes into gaseous phase ions, through a process known as *ionization*, by gaining a positive or a negative charge from a neutral species. The ions are then introduced into the *mass analyzer* of a mass spectrometer, where ions are accelerated under electric fields and separated based on their m/z ratio. The separated ions are detected by an *ion detector* that records the number of ions at each m/z value. The result of molecular ionization,

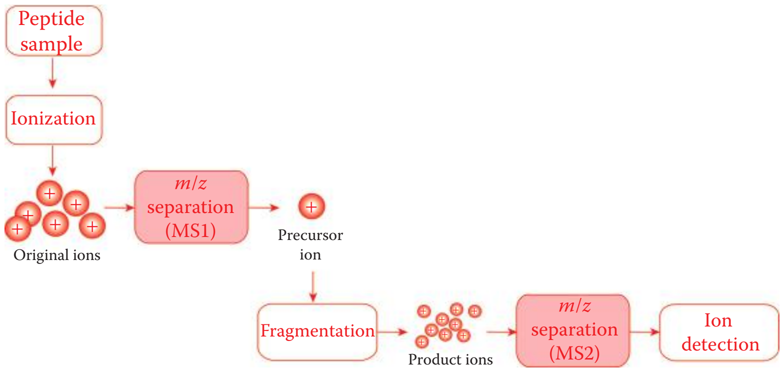

Figure 11.21 Diagram of tandem mass spectrometry. For the initial mass analysis of thermolabile analytes including peptides and proteins, soft ionization techniques such as matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) can be used to produce original ions. MALDI utilizes a laser beam to trigger ionization and facilitate the vaporization of the analyte. In the ESI method, a fine mist of droplets is formed in the presence of a high electric field, producing charged ions. In the first round of mass analysis, ions are separated based on their *m/z* ratio. A particular ion, called a precursor ion, is then selected for fragmentation, generating product ions. The fragmentation can be carried out using collision techniques such as collision-induced dissociation (CID) by multiple collisions with rare gas atoms. Additionally, a new fragmentation technique, electron-capture dissociation (ECD), can be used in which the fragmentation is induced through the capture of a thermal electron by a protonated peptide cation. The product ions are analyzed by the second round of mass analysis and detected by an ion detector. MS, mass analyzer. (© Richard C. Li.)

ion separation, and ion detection is a spectrum that can be used to determine the mass (or molecular weight) and potentially the structure of the molecule.

Furthermore, biomarker proteins can be identified by *peptide sequencing* using the tandem mass spectrometry (MS/MS) mode of operation in which a mass spectrometer uses two or more mass analyzers. After an initial mass analysis, individual peptide ions in the mass spectrum can be selected. The selected ions are known as *precursor ions*. A precursor ion is then subjected to a second round of fragmentation through collision and is broken into smaller ions known as *product ions*. The resulting product ions are further analyzed by MS. Thus, the corresponding protein can be identified through searching a protein database, based on the mass of a specific peptide obtained.

11.6.2 Analysis Strategies for Protein Identification

To date, various MS-based protein identification strategies have been developed, which can be divided into two categories: "top-down" and "bottom-up" strategies (Figure 11.22). In the *top-down* strategy, intact proteins in a complex mixture are fractionated and separated into less complex protein mixtures or single proteins. Intact proteins are then analyzed using MS. A liquid chromatography–mass spectrometry (LC–MS) system can be used for such an application. The advantage of this approach is that an entire protein is analyzed for identification purposes. However, the throughput and efficiency of this approach is still a major challenge for large-scale sample analysis in forensic applications.

The *bottom-up* strategy is commonly used for high-throughput analysis of small peptides derived from highly complex samples. Current MS usually cannot resolve the mass measurement of large proteins; rather, MS is ideal for the analysis of small peptides. One approach is the *sort-then-break* approach in which proteins in a complex mixture are first isolated and then enzymatically cleaved. The target proteins are isolated from their complex biological source using 1-D or 2-D gel electrophoresis (Chapter 18). The isolated proteins are then enzymatically

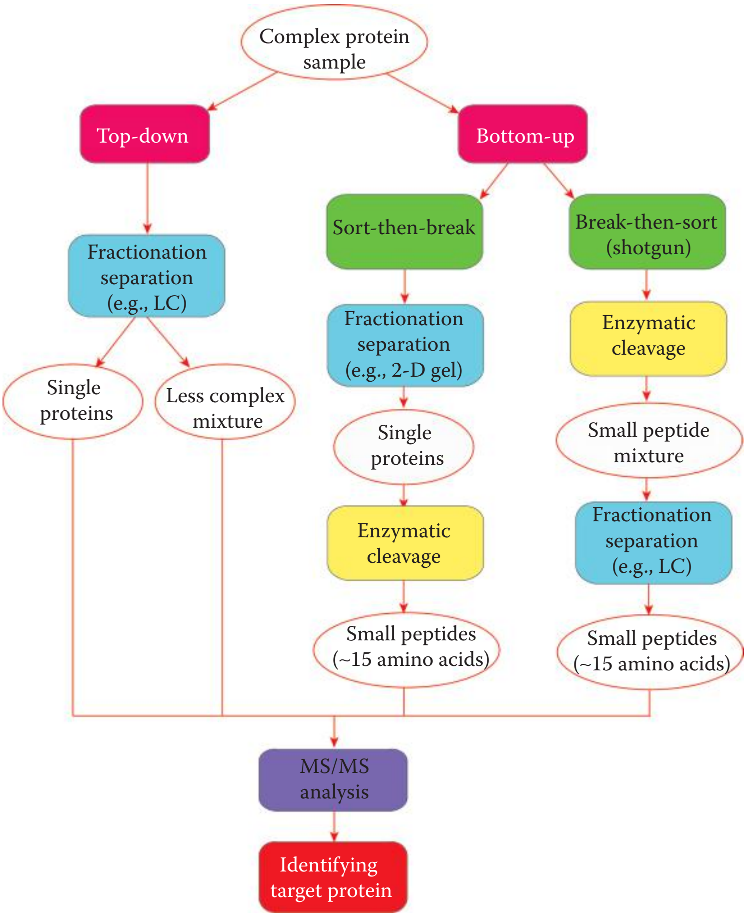
Complex protein sample
Top-down
Bottom-up
Fractionation separation (e.g., LC)
Single proteins
Less complex mixture
Sort-then-break
Fractionation separation (e.g., 2-D gel)
Single proteins
Enzymatic cleavage
Small peptides (~15 amino acids)
Break-then-sort (shotgun)
Enzymatic cleavage
Small peptide mixture
Fractionation separation (e.g., LC)
Small peptides (~15 amino acids)
MS/MS analysis
Identifying target protein

Figure 11.22 Top-down and bottom-up strategies for MS-based protein identification. LC, liquid chromatography; MS/MS, tandem mass spectrometry. (© Richard C. Li.)

cleaved into small peptides, approximately 15 amino acids in length, using proteases such as trypsin. These peptides are then analyzed by peptide sequencing using tandem MS to identify the target protein.

Break-then-sort, also known as the *shotgun* approach, is another approach of the bottom-up strategy. In this approach, proteins are cleaved first and the resulting peptide mixture is then separated by LC and analyzed by tandem MS (usually coupled to the LC). In a complex mixture, a peptide mass measurement alone is not sufficient for the identification of a target protein. Database searches are carried out utilizing software that compares the observed MS peptide spectra against all candidate spectra of the proteins that are present in a protein sequence (or translated from genomic DNA sequences) database for possible matches. Thus, a biomarker protein can be identified.

11.7 Microbial DNA Analysis for Bodily Fluid Identification

Human bodies harbor a large number of microbes. It is estimated that there are over 10 times more microbial cells than human cells in and on the human body, in the nose, mouth, throat, intestine, vagina, and skin. This microbial community is referred to as the human *microbiota*.

It is known that the microbiota has an important influence on human health. A decade ago, the concept of the *microbiome* was proposed by Nobel laureate Joshua Lederberg to identify and characterize the microbiota, their genomes, and their environmental interactions in a defined community. Traditionally, the studies of microbial genomes rely on cultivated individual species. The disadvantage of cultivation-based methods, however, is that large numbers of microbial species cannot be successfully isolated because many microbial species require specific growth conditions. Based on the advances in DNA sequencing technologies, a new approach, known as *metagenomics*, has been developed. Instead of studying the genome of an individual cultivated microbial species, metagenomics allows the study of the microbial genome of complete microbial communities harvested directly from natural environments. Metagenomics has been a powerful tool to analyze the human microbiome. Recently, human microbiota have been surveyed, revealing that microbial diversity is greatest in the teeth and the intestines, intermediate on the skin and the inside surface of the cheek, and lowest in the vagina (Table 11.4). More interestingly, each habitat is characterized by a small number of highly predominant bacterial taxa. The predominant taxa of each habitat can potentially be utilized for the forensic analysis of biological evidence. Bodily fluids identification is an important task in the forensic analysis of biological evidence. The detection of predominant taxa of bacteria can aid in the forensic identification of a particular type of bodily fluid (Table 11.4).

Table 11.4 Representative Microbiota that are Potentially Can Be Used for Forensic Identification

Body Habitats	Estimated Number of Microbial Species	Predominent Taxa	Potential Forensic Applications	DNA Marker	Further Reading
Cheek	800	*Streptococcus* spp.	Saliva identification	*gtf*	Nakanishi et al. (2011)
			Expirated blood	*gtf*	Donaldson et al. (2010)
				16S rRNA	Power et al. (2010)
Vagina	300	*Lactobacillus* spp.	Vaginal secretion identification	16S rRNA	Akutsu et al. (2012)
				16S–23S ISR	Fleming and Harbison (2010)
					Giampaoli et al. (2012)
Gastrointestinal tract	4000	*Bacteroides* spp.	Fecal matter identification	*rpoB*	Nakanishi et al. (2013)
Skin	1000	*Staphylococcus* spp.	Touched evidence identification	16S rRNA	Fierer et al. (2010)

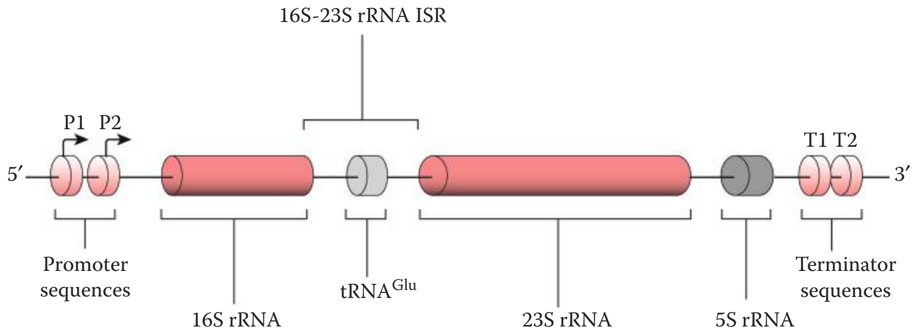

Figure 11.23 Schematic diagram of an *Escherichia coli* rRNA operon. *E. coli* has seven rRNA operons. A typical rRNA operon is shown. The rRNA operon encodes for three ribosomal RNA (rRNA) genes: the 5S and 23S genes code for the large ribosomal subunit, while the 16S gene encodes for the small ribosomal subunit. *E. coli* encodes one tRNA gene in the rRNA operon. The intergenic spacer regions (ISRs) and the promoter and terminator sequences are also shown. The rRNA operon is transcribed into a single precursor RNA transcript that is cleaved into separate rRNA and tRNA transcripts. The diagram is not to scale to the physical map of rRNA operons. (© Richard C. Li.)

The detection of bacteria can be achieved by utilizing DNA markers. The most commonly used DNA markers are the essential genes present in all bacterial species. These markers have highly variable regions that can be used to distinguish various bacterial species. The DNA sequences within these variable regions can also suggest taxonomic relationships between different taxa. Additionally, DNA markers contain conserved regions that are necessary for use as primer-binding sites for PCR-based detection.

For example, the rRNA genes (Figure 11.23), encoding for the ribosomal RNA, are conserved in all living organisms including bacteria. The bacterial rRNA operon contains three rRNA genes: the 5S and 23S rRNA genes encode the RNA components of the large subunit of the ribosome, while the 16S rRNA gene encodes the RNA component of the small subunit of the ribosome. The 16S rRNA gene is the most commonly used marker for the taxonomic identification of bacteria. The bacterial 16S rRNA genes are similar in length, approximately 1.5 kb. They contain highly conserved regions that can be used as primer-binding sites for the amplification of the adjacent DNA regions. Additionally, they contain variable regions that allow for taxonomic identification. In forensic applications, they are used as a marker for the identification of vaginal bacteria as an indicator of vaginal secretions (Chapter 16). However, the number of variable regions in the 16S rRNA genes of some bacteria species is limited. Furthermore, closely related species have high levels of sequence similarity to the 16S rRNA genes. Thus, the 16S rRNA gene marker is not adequate for distinguishing closely related species.

DNA markers of bacteria identification can also be located in noncoding regions of DNA. The *intergenic spacer region* (ISR) between the 16S rRNA and 23S rRNA genes in the rRNA operon is another commonly used marker. The 16S–23S rRNA ISR contains both length and sequence variations between species. These variations are partially caused by the number and type of tRNA genes that this region of operon contains. For example, most gram-negative bacteria contain both a copy of the tRNA^{ala} and the tRNA^{ile} genes, while some contain only a copy of the tRNA^{glu} gene. In contrast, most gram-positive bacteria have no tRNA gene; some contain either a copy of the tRNA^{ala} gene or the tRNA^{ile} gene or both. Thus, heterogeneity in the length and in the sequence of the 16S–23S rRNA ISR allows bacteria to be identified at the species level.

In addition to rRNA genes, many other genes are utilized as DNA markers for bacterial species identification. For example, the *rpoB* gene encodes the β subunit of bacterial RNA polymerase and is used as a marker for the forensic identification of fecal matter (Chapter 17). The sequence variation of the *rpoB* gene allows the distinguishing of species when they are not distinguishable, using 16S rRNA gene sequences. Several oral streptococcal species that produce extracellular polysaccharides contain the *gtf* gene, which encodes glucosyltransferase. Thus, the *gtf* gene is a useful marker for oral streptococcal species identification as a potential indicator of saliva in forensic investigations. However, this marker cannot be used to identify bacteria other than *streptococcus*. Nevertheless, combining several DNA markers can be used for the identification of bacteria species.

11.8 Nondestructive Assays for the Identification of Bodily Fluids

Most of the bodily fluid identification techniques mentioned previously consume a portion of the evidence. When the amount of sample evidence is very limited, such destructive assays may consume the evidence that is also needed for subsequent forensic DNA analysis. Moreover, most of these tests are used for a single type of bodily fluid. Tests for multiple bodily fluids in question will consume additional amounts of the evidence. Therefore, nondestructive identification techniques are highly desired. Additionally, it is important to develop an assay that can be utilized to test multiple types of bodily fluids. Furthermore, a portable device that can identify bodily fluids at a crime scene is very useful. It provides investigators with test results immediately to aid the crime scene investigations. Two techniques have been developed that can potentially be used for nondestructive bodily fluid identification: fluorescence spectroscopy and Raman spectroscopy.

Fluorescence spectroscopy is one example of a nondestructive technique. Fluorescence is the emission of light by a fluorophore. A fluorophore is a moiety in a molecule that fluoresces on absorbing the energy from an excitation light source or radiation. The emitted light usually has a longer wavelength and lower energy than the absorbed radiation. Constituents present in bodily fluids, such as nucleic acids, proteins, lipids, and metabolites, can exhibit fluorescence. The unique composition of a bodily fluid emits characteristic emission spectra, thus making it identifiable. Using multiple wavelengths of an excitation light source, fluorescent emissions at a wide range of wavelengths can be detected, which allows for the identification of various bodily fluids. This technique is sensitive and rapid. It does not utilize chemical reagents and the detection does not require physical contact with a sample. After identification, the same biological sample can be used for forensic DNA analysis. Additionally, portable fluorescence instruments can potentially be used at a crime scene for bodily fluid identification. However, it is not clear if exposure to an excitation light source can damage DNA evidence.

Raman spectroscopy is another example of a nondestructive technique. Raman spectroscopy utilizes a near-infrared excitation light source and measures the scattering of laser light caused by the vibrating molecules of a sample. A typical Raman spectrum provides information about the molecular structure and the "signature" based on the properties of the constituents of a sample. Thus, it is possible to obtain unique Raman spectra for a particular type of bodily fluid. Raman spectroscopy is highly sensitive. The measurements can be carried out on very small amounts of sample, ranging from a few picoliters to femtoliters. This technique does not require any chemical reagent. Additionally, it does not consume the sample, which can be used for subsequent DNA analysis. Raman spectroscopy has been utilized for various forensic purposes including the identification of drugs, trace evidence such as fibers, and questioned document evidence such as inks and paints. It can potentially be used for the nondestructive identification of various bodily fluids. A portable Raman spectrometer will allow bodily fluids to be identified at crime scenes.

12

Identification of Blood

12.1 Biological Properties

Blood constitutes about 8% of the human body weight of a healthy individual. *Plasma* is the fluid portion of the blood. The cellular portion of the blood consists of *red blood cells*, *white blood cells*, and *platelets*, all of which are suspended in the plasma (Figures 12.1 and 12.2).

12.1.1 Red Blood Cells

These cells are also called *erythrocytes*. Their life span in humans is approximately 3–4 months. Additionally, mature human erythrocytes do not have nuclei, and therefore lack nuclear DNA. Erythrocytes consist of hemoglobin—proteins that are responsible for the transportation of oxygen. Most adult human hemoglobin consists of four polypeptide chains, two α chains and two β chains. Thus, adult hemoglobin is designated as $\alpha_2\beta_2$. Other forms of hemoglobin will be discussed in Chapter 18. Under normal physiological conditions, each hemoglobin subunit contains a heme moiety that binds to oxygen (Figure 12.3). A heme molecule consists of an organic component known as protoporphyrin IX and a *ferrous* (Fe^{2+}) iron ion (Figure 12.4). A heme molecule is also known as *ferroprotoporphyrin*. The ferrous ion of heme forms four bonds with the nitrogens of protoporphyrin IX, along with a fifth bond with a hemoglobin chain and a sixth bond with a molecule of O_2. Other chemicals such as carbon monoxide and cyanide also bind to the ferrous iron of the heme molecule and can cause chemical asphyxia. Heme groups are also present in the blood of various animals and in other proteins such as myoglobin in muscles and neuroglobin in the brain.

12.1.2 White Blood Cells

Also called leucocytes, white blood cells are subdivided into three types: granulocytes, lymphocytes, and monocytes. White blood cells are involved in defending the body against infection. They have nuclei and thus represent the main sources of nuclear DNA from the blood.

12.1.3 Platelets

These cells are also known as thrombocytes, and they play a role in blood clotting (Chapter 16). Platelets aggregate at sites of vascular and blood vessel injury. Like erythrocytes, they lack nuclei.

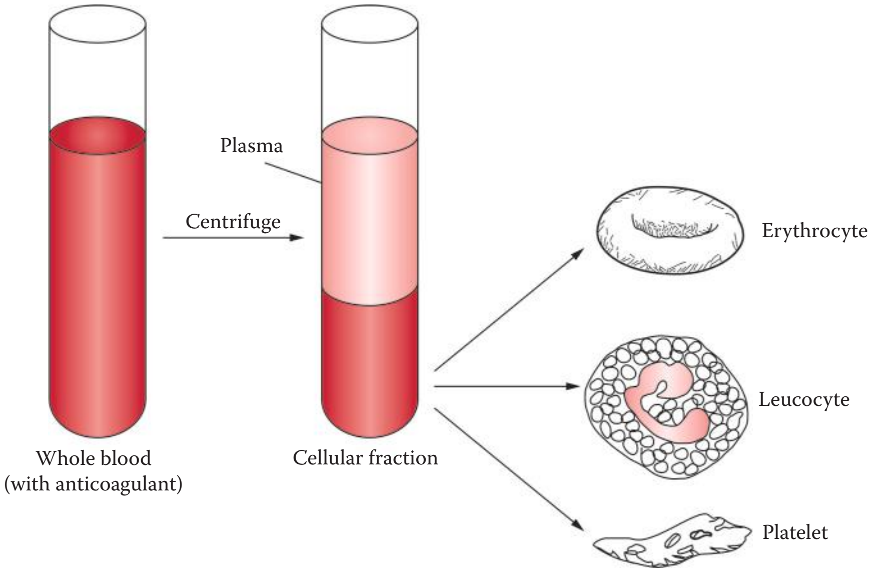

Figure 12.1 Basic composition of blood. Blood can be separated into two phases in the presence of an anticoagulant. The liquid portion called plasma accounts for approximately 55% of blood volume. The cellular elements include erythrocytes, leucocytes, and platelets. (© Richard C. Li.)

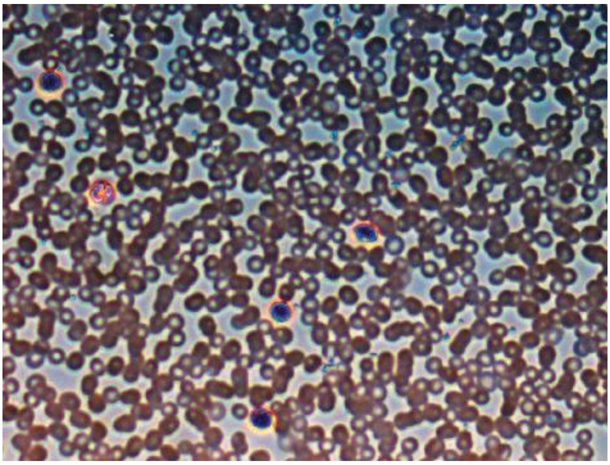

Figure 12.2 Blood cells. (© Richard C. Li.)

12.2 Presumptive Assays for Identification

12.2.1 Mechanisms of Presumptive Assays

Presumptive blood assays are designed to detect traces of blood. These assays are based on the basic principle of the oxidation–reduction reaction catalyzed by the heme moiety of the hemoglobin. As a result, colorless substrates catalyzed by heme undergo an oxidation reaction, causing either chemiluminescence, fluorescence, or a change of color. These assays are very sensitive and can detect blood in samples with 10^{-5}–10^{-6}-fold dilutions. A positive reaction indicates the possible presence of blood. Additionally, most of these assays do not interfere with forensic DNA analysis.

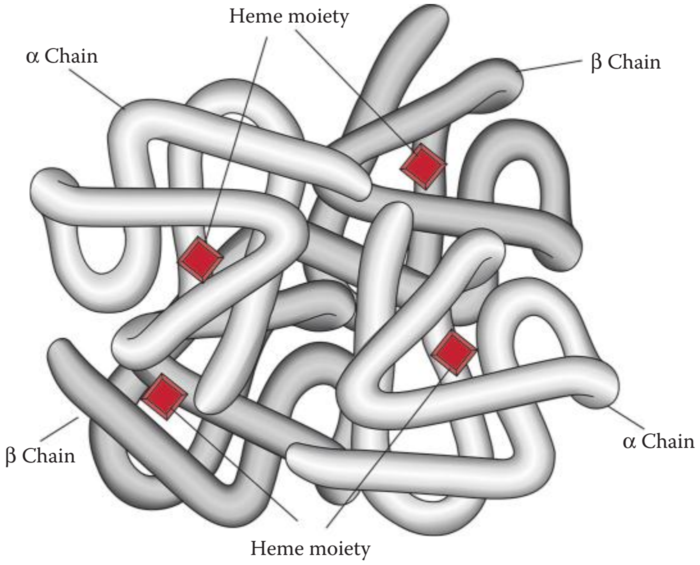
Heme moiety
α Chain
β Chain
β Chain
α Chain
Heme moiety

Figure 12.3 Human adult hemoglobin. Four subunits, two α and two β chains, of human adult hemoglobin are shown. Each hemoglobin subunit contains a heme moiety. (© Richard C. Li.)

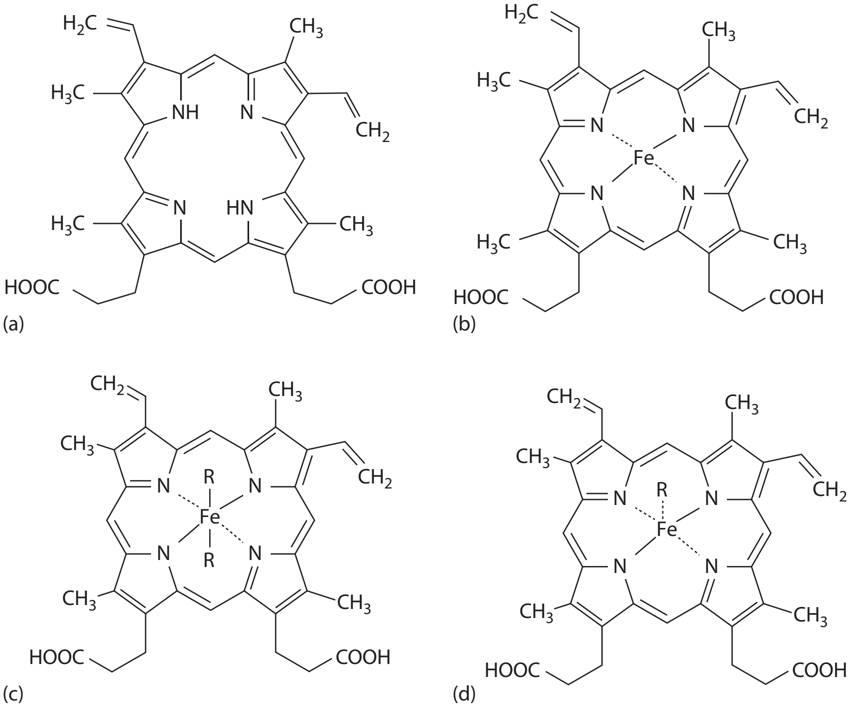

Figure 12.4 Chemical structures of heme and its precursor and derivatives. (a) Protoporphyrin IX. (b) Heme (ferroprotoporphyrin). (c) Hemochromagen, R = pyridine (pyridineferroprotoporphyrin). (d) Hematin hydroxide, R = OH (ferriprotoporphyrin hydroxide); hematin chloride, R = Cl (ferriprotoporphyrin chloride).

12.2.1.1 Oxidation–Reduction Reactions

An oxidation–reduction reaction involves a change in the oxidation state of a molecule. Specifically, the *oxidation* of a molecule means that the molecule has lost electrons, and the *reduction* of a molecule means that the molecule has gained electrons. Chemicals that can be reduced and therefore gain electrons from other molecules are called *oxidants*. In contrast, *reductants* are chemicals that can be oxidized and therefore lose electrons to other molecules. In biochemical reactions, oxidation often coincides with a loss of hydrogen. Figure 12.5 depicts an example of an oxidation–reduction reaction for blood identification. In presumptive assays, heme is utilized as the catalyst, and hydrogen peroxide is utilized as the oxidant for the reaction. In the presence of heme, a colorless substrate is oxidized, yielding a product with color, chemiluminescence, or fluorescence.

12.2.2 Colorimetric Assays

Many procedures are available for detecting heme in blood through color reactions. The most common agents are phenolphthalin, leucomalachite green, and benzidine derivatives. The color reactions produced by these assays can be observed immediately with the naked eye.

12.2.2.1 Phenolphthalin Assay

Phenolphthalein, a member of a class of indicators and dyes, is used in titrations of mineral and organic acids as well as most alkalis. The phenolphthalin assay for blood identification is also known as the Kastle–Meyer test. Kastle published a study in 1901, presenting the results of a reaction in which phenolphthalin, a colorless compound, is catalyzed by heme with hydrogen peroxide as the oxidant (Figures 12.6 and 12.7). The oxidized derivative is phenolphthalein, which appears pink under alkaline conditions.

12.2.2.2 Leucomalachite Green (LMG) Assay

Malachite green is a triphenylmethane dye. The leuco base form of malachite green is colorless and can be oxidized by the catalysis of heme to produce a green color. The reaction is carried out under acid conditions with hydrogen peroxide as the oxidant (Figures 12.8 and 12.9).

12.2.2.3 Benzidine and Derivatives

Historically, *benzidine* was used as an intermediate in dye manufacturing (Figure 12.10). Subsequently, it was used as a presumptive assay for the presence of blood after the discovery that the oxidation of benzidine can be catalyzed by heme to produce a blue to dark blue color (carried out in an acid solution). Since the blue color may eventually turn brown, the reaction must be read

$$\underset{\text{(Colorless)}}{AH_2} + H_2O_2 \xrightarrow{\text{Heme}} \underset{\text{(Color)}}{A} + 2H_2O$$

Figure 12.5 Oxidation–reduction reaction as the basis for presumptive assays for blood identification. AH_2, substrate; A, oxidized substrate.

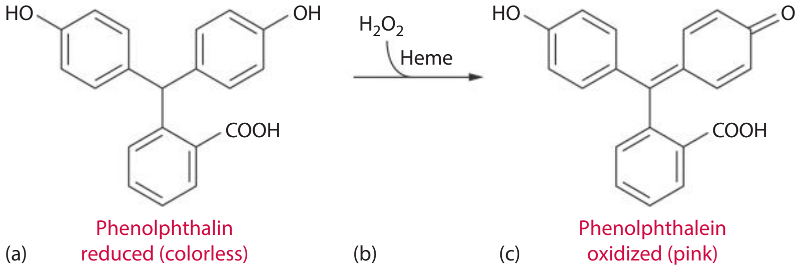

(a) Phenolphthalin reduced (colorless) (b) (c) Phenolphthalein oxidized (pink)

Figure 12.6 Chemical reaction of phenolphthalin assay.

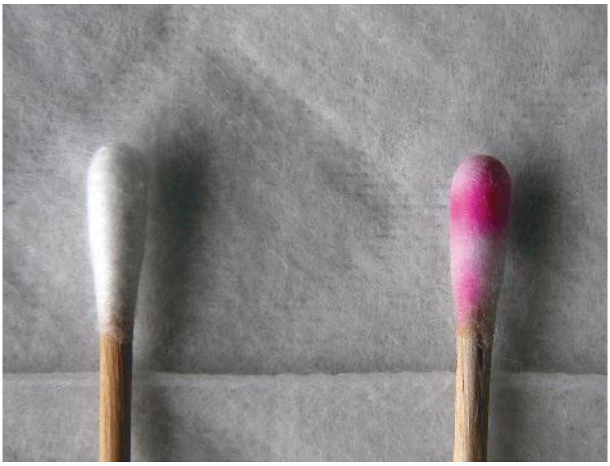

Figure 12.7 Photograph of phenolphthalin assay results. Negative (left) and positive (right) reaction. (© Richard C. Li.)

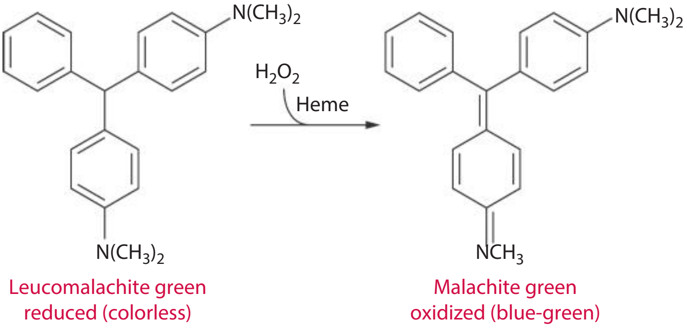
N(CH3)2
N(CH3)2
H2O2
Heme
N(CH3)2
NCH3
Leucomalachite green
reduced (colorless)
Malachite green
oxidized (blue-green)

Figure 12.8 Chemical reaction of leucomalachite green assay.

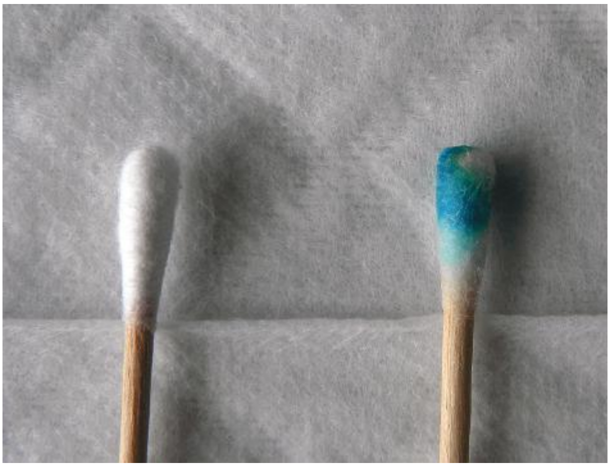

Figure 12.9 Photograph of leucomalachite green assay results. Negative (left) and positive (right) reaction. (© Richard C. Li.)

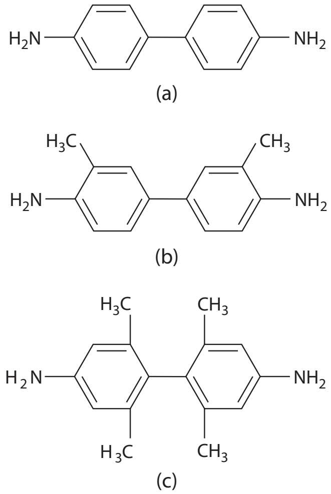

Figure 12.10 Chemical structures of benzidine and derivatives: (a) benzidine; (b) orthotolidine; (c) tetramethylbenzidine.

immediately. Benzidine was found to be a carcinogen and is therefore no longer used for forensic testing. *Orthotolidine* is a dimethyl derivative of benzidine. Its oxidation reaction can be catalyzed by heme to produce a blue color reaction under acidic conditions (Figure 12.11). Orthotolidine is also considered a potential carcinogen based on animal studies, and for this reason it has been replaced by tetramethylbenzidine. *Tetramethylbenzidine* (TMB) is a tetramethyl derivative of benzidine. The oxidation of TMB can be catalyzed by heme to produce a green to blue-green color under acidic conditions. TMB continues to be used. The Hemastix® assay kit (Miles Laboratories) is a TMB-based assay that utilizes a TMB-containing strip device. A test is carried out by applying a moistened sample to a Hemastix® strip. The appearance of a green or a blue-green color indicates the presence of blood.

12.2.3 Chemiluminescence and Fluorescence Assays

Other organic compounds whose oxidation products have chemiluminescent or fluorescent properties are utilized for testing. In the *chemiluminescence assay*, light is emitted as a product of a chemical reaction. In this category, luminol produces chemiluminescence when blood is present. In contrast, a *fluorescence assay* requires the exposure of an oxidized product, such as fluorescin, to a particular wavelength of an excitation light source. The fluorescence is then emitted at longer wavelengths than that of the excitation light source.

One advantage of chemiluminescent and fluorescent reagents is that they can be sprayed over large areas where latent bloodstains are potentially located. A positive reaction identifies blood and also reveals the patterns of bloody impressions such as footprints and fingerprints. These methods are very sensitive and can pinpoint the locations of even small traces of blood. Additionally, they are useful for detecting blood at crime scenes that have been cleaned and show no visible staining. One disadvantage of chemiluminescent and fluorescent reagents is that precautions must be taken if stains are very small or have been washed. The spraying of

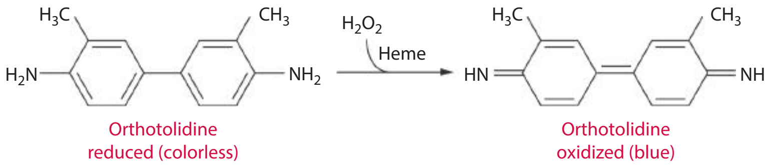

Figure 12.11 Chemical reaction of orthotolidine assay.

presumptive assay reagents may further dilute a sample and thus lead to difficulty in isolating sufficient amounts of DNA for forensic DNA analysis.

12.2.3.1 Luminol (3-Aminophthalhydrazide)

Luminol is usually utilized as a chemiluminescent reagent. The oxidation reaction of luminol catalyzed by heme produces light in the presence of an oxidant (Figures 12.12 through 12.14). The light emitted from a positive reaction can only be observed in the dark, which limits the

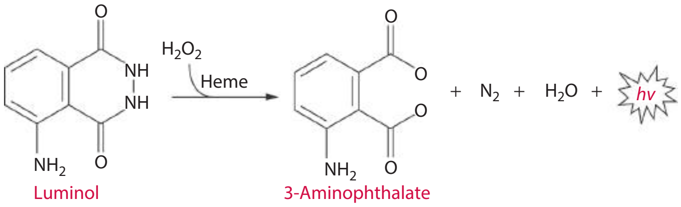

Figure 12.12 Chemical reaction of luminol assay.

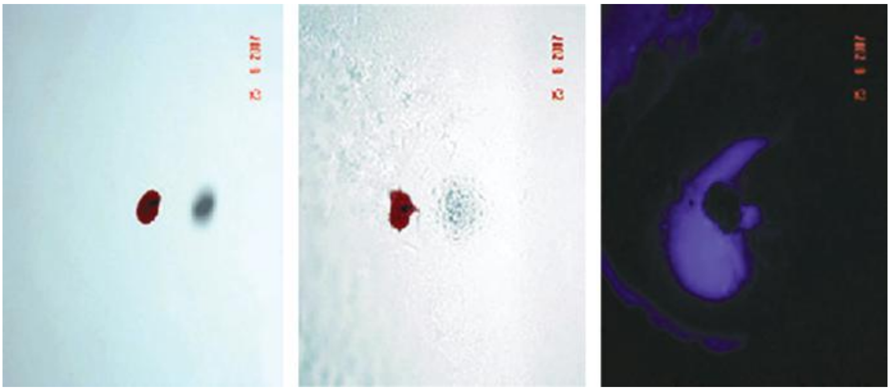

Figure 12.13 Results obtained from testing a drop of blood on a glass plate with luminol. From left to right: drop of blood; drop of blood after spraying with luminol; drop of blood after spraying with luminol and viewing in the dark. A blue chemiluminescence indicates the presence of blood. (From Bergervoet, P.W., et al., *J Hosp Infect*, 68, 329–333, 2008. With permission.)

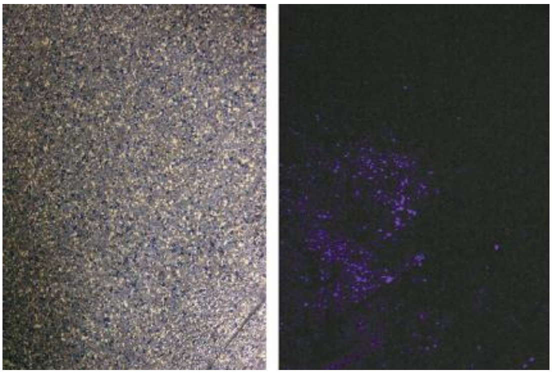

Figure 12.14 Detecting latent bloodstains on a floor with luminol. Left: with the light on; right: immediately after the light is switched off. A blue chemiluminescence indicates the presence of blood traces. (From Bergervoet, P.W., et al., *J Hosp Infect*, 68, 329–333, 2008. With permission.)

applications of luminol. The photodocumentation of a luminol-enhanced pattern should be done immediately before it fades away.

12.2.3.2 Fluorescin

Fluorescin is another reagent that is used to test for the presence of bloodstains at a crime scene (Figures 12.15 and 12.16). When oxidized and catalyzed by heme, fluorescin demonstrates fluorescent properties. Usually, fluorescin-sprayed stains are exposed to light in the range of 425–485 nm using an alternate light source device. In a positive reaction, the oxidized fluorescin emits an intense yellowish-green fluorescent light, which indicates the presence of a bloodstain. The light emitted from fluorescin-sprayed stains lasts longer than that of luminol.

12.2.4 Factors Affecting Presumptive Assay Results

The catalytic assays discussed in the previous sections are not specific to blood only, which can possibly lead to the observation of false-positive or false-negative results (Figure 12.17).

12.2.4.1 Oxidants

Chemicals that are strong oxidants may cause a false-positive reaction. Such chemicals can catalyze the oxidation reaction even in the absence of heme and result in a false-positive reaction. Certain metal salts, such as copper and nickel salts, household bleaches and cleaners that contain hypochlorite ions, and hair-coloring products that contain hydrogen peroxide, work as oxidants. To address this problem, a two-step catalytic assay should be performed. The substrate

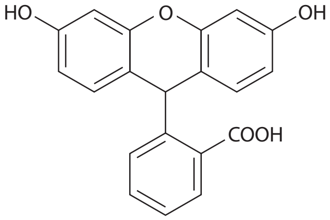

Figure 12.15 Chemical structure of fluorescein.

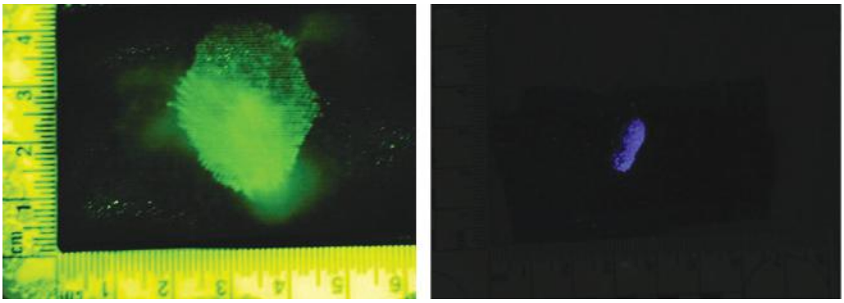

Figure 12.16 Detecting diluted bloodstains on black cotton fabric with fluorescin (left) and Bluestar (right). (From Finnis, J., Lewis, J., and Davidson, A., *Sci Justice*, 53, 178–186, 2013. With permission.)

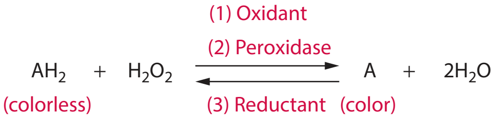

$$\underset{\text{(colorless)}}{AH_2} + H_2O_2 \underset{\text{(3) Reductant}}{\overset{\text{(1) Oxidant}}{\overset{\text{(2) Peroxidase}}{\rightleftarrows}}} \underset{\text{(color)}}{A} + 2H_2O$$

Figure 12.17 Factors affecting presumptive assay results. Strong oxidant and peroxidase may cause false-positive results; reductant may cause false-negative results.

is applied first to the sample in question. A color change occurring before the addition of hydrogen peroxide indicates a false-positive result due to a possible oxidant in the sample. If a color change is observed after the addition of hydrogen peroxide, the result is a true positive.

12.2.4.2 Plant Peroxidases

Many types of plants such as horseradish contain peroxidases. Plant peroxidases may also catalyze oxidation reactions and lead to false-positive results. However, plant peroxidases are usually heat sensitive and may be inactivated by high temperatures. Because the heme molecule is relatively stable at high temperatures, samples can be retested after heating. This will inactivate any plant peroxidases in a sample.

12.2.4.3 Reductants

Although not common, a false-negative result can occur when a strong reductant is present in a sample. Strong reductants such as certain metal ions including lithium and zinc may inhibit the oxidation reaction.

12.3 Confirmatory Assays for Identification

12.3.1 Microcrystal Assays

Microcrystal assays apply chemicals to treat bloodstains, forming crystals of heme molecules. The morphologies of the resulting crystals are distinctive for heme and can be compared with a known standard using a microscope. A positive microcrystal assay strongly indicates the presence of blood. However, confirmatory assays are usually not as sensitive as presumptive assays. Additionally, these assays cannot distinguish between human and animal blood.

12.3.1.1 Hemochromagen Crystal Assay

Hemochromagens are heme derivatives in which the ferrous iron of the heme forms two bonds with nitrogenous bases (Figure 12.4). The method for forming hemochromagen crystals was documented in 1864. Since then, various modifications have been reported. The *Takayama crystal assay*, published in 1912, has been the method preferred by many forensic laboratories. A bloodstain is treated with pyridine and glucose (a reducing sugar that is capable of reducing ferric ion) under alkaline conditions to form crystals of pyridine ferroprotoporphyrin (Figure 12.18).

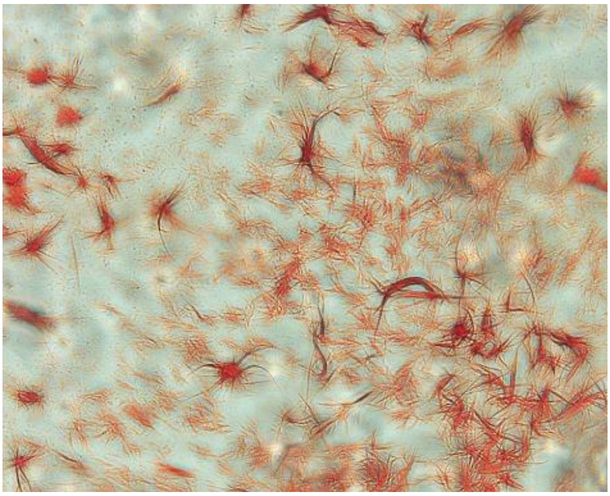

Figure 12.18 Microcrystal assays using the Takayama method. (© Richard C. Li.)

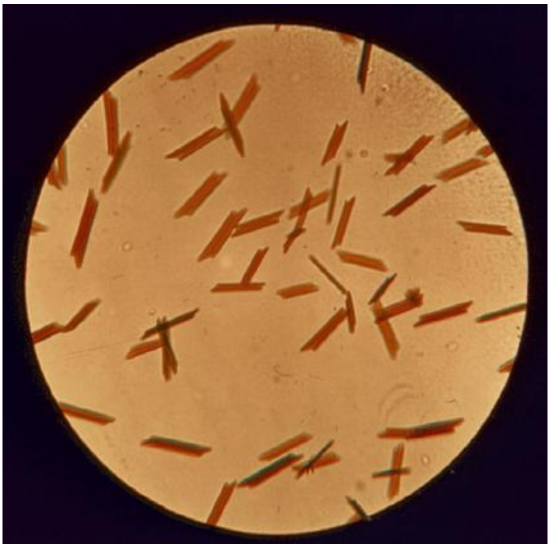

Figure 12.19 Microcrystal assays using the Teichmann method. (From James, S.H., Nordby, J.J., and Bell, S., *Forensic Science: An Introduction to Scientific and Investigative Techniques*, 4th edn., CRC Press, Boca Raton, 2014. With permission.)

12.3.1.2 Hematin Crystal Assay

This assay is also known as the *Teichmann crystal assay*. In 1853, Teichmann documented a method of forming crystals of blood specimens. When blood specimens are treated with glacial acetic acid and salts, and subsequently heated, hematin chloride (ferriprotoporphyrin chloride), a prismatic brown-colored crystal, is formed (Figure 12.19). Hematin (Figure 12.4) is a heme derivative; its iron is in the *ferric* (Fe^{+3}) state. This hematin assay has a similar sensitivity and specificity as hemochromagen assays. The hematin assay has the advantage of being more reliable than hemochromagen assays for aged blood samples.

12.3.2 Other Assays

Additional techniques may be used to confirm the presence of hemoglobin. For example, *chromatographic* and *electrophoretic methods* can identify hemoglobin by its mobility characteristics. *Spectrophotometric methods* for identifying hemoglobin are based on measurements of the

Table 12.1 Application of RT-PCR Assay for Blood Identification

Gene Symbol	Gene Product	Description	Further Reading
HBA1	Hemoglobin α1	Hemoglobin α1 chain (abundant in erythrocytes)	Waye and Chui (2001)
PBGD[a]	Porphobilinogen deaminase	Erythrocyte-specific isoenzyme of heme biosynthesis pathway	Gubin and Miller (2001)
SPTB	β-Spectrin	Subunit of major protein of erythrocyte membrane skeleton	Amin et al. (1993)

Source: Adapted from Juusola, J. and Ballantyne, J., *Forensic Sci Int*, 152, 1–12, 2005; Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., *Forensic Sci Int*, 157, 181–186, 2006.

[a] Also known as hydroxymethylbilane synthase (HMBS).

characteristic light spectra, with peak absorbance at 400–425 nm, absorbed by hemoglobin and its derivatives. Finally, *immunological methods* utilize antihuman hemoglobin antibodies. This antibody can be used to detect human hemoglobin and thus indicate the presence of human blood (see Chapter 13).

Recently, *RNA-based assays* have been developed to identify blood. These assays are based on the fact that certain genes are specifically expressed in certain cell types (Chapter 11). Thus, the techniques used in the identification of blood are based on the detection of specific types of messenger RNA (mRNA) that are expressed exclusively in erythrocytes. These assays utilize reverse transcriptase polymerase chain reaction (RT-PCR; see Chapter 7) methods to detect the gene expression levels of mRNAs for blood identification. Table 12.1 summarizes the tissue-specific genes utilized for blood identification. Compared with conventional assays that are used for blood identification, the RNA-based assays have higher specificity and are amenable to automation. However, one limitation is that RNA is unstable due to degradation by endogenous and environmentally born ribonucleases.

Bibliography

Amin, K.M., et al., The exon-intron organization of the human erythroid beta-spectrin gene. *Genomics*, 1993, **18**(1): 118–125.

Andersen, J. and S. Bramble, The effects of fingermark enhancement light sources on subsequent PCR-STR DNA analysis of fresh bloodstains. *J Forensic Sci*, 1997, **42**(2): 303–306.

Anderson, S.E., G.R. Hobbs, and C.P. Bishop, Multivariate analysis for estimating the age of a bloodstain. *J Forensic Sci*, 2011, **56**(1): 186–193.

Anderson, S., et al., A method for determining the age of a bloodstain. *Forensic Sci Int*, 2005, **148**(1): 37–45.

Andrasko, J., The estimation of age of bloodstains by HPLC analysis. *J Forensic Sci*, 1997, **42**(4): 601–607.

Anslinger, K., et al., Ninhydrin treatment as a screening method for the suitability of swabs taken from contact stains for DNA analysis. *Int J Legal Med*, 2004, **118**(2): 122–124.

Arany, S. and S. Ohtani, Age estimation of bloodstains: A preliminary report based on aspartic acid racemization rate. *Forensic Sci Int*, 2011, **212**(1–3): e36–e39.

Barnett, P.D., et al., Discussion of "Effects of presumptive test reagents on the ability to obtain restriction fragment length polymorphism (RFLP) patterns from human blood and semen stains". *J Forensic Sci*, 1992, **37**(2): 369–370.

Bauer, M., S. Polzin, and D. Patzelt, Quantification of RNA degradation by semi-quantitative duplex and competitive RT-PCR: A possible indicator of the age of bloodstains? *Forensic Sci Int*, 2003, **138**(1–3): 94–103.

Bergervoet, P.W., et al., Application of the forensic Luminol for blood in infection control. *J Hosp Infect*, 2008, **68**: 329–333.

Botoniic-Sehic, E., et al., Forensic application of near-infrared spectroscopy: Aging of bloodstains. *Spectroscopy* (Santa Monica), 2009, **24**(2): 42–48.

Boyd, S., et al., Highly sensitive detection of blood by surface enhanced Raman scattering. *J Forensic Sci*, 2013, **58**(3): 753–756.

Bremmer, R.H., et al., Remote spectroscopic identification of bloodstains. *J Forensic Sci*, 2011, **56**(6): 1471–1475.

Bremmer, R.H., et al., Forensic quest for age determination of bloodstains. *Forensic Sci Int*, 2012, **216**(1–3): 1–11.

Budowle, B., et al., The presumptive reagent fluorescein for detection of dilute bloodstains and subsequent STR typing of recovered DNA. *J Forensic Sci*, 2000, **45**(5): 1090–1092.

Caldwell, J.P., W. Henderson, and N.D. Kim, ABTS: A safe alternative to DAB for the enhancement of blood fingerprints. *J Forensic Sci*, 2000, **45**(4): 785–794.

Cheeseman, R., Direct sensitivity comparison to the fluorescein and luminol bloodstain enhancement techniques. *J Forensic Ident*, 1999, **49**(3): 8.

Cheeseman, R. and L.A. DiMeo, Fluorescein as a field-worthy latent bloodstain detection system. *J Forensic Ident*, 1995, **45**(6): 1.

13

Species Identification

Chapter 12 discussed the principles of the identification of blood. If a stain is identified as blood, the evidence can be tested to determine whether the blood is of human origin. If the bloodstain is nonhuman, further analysis is usually not necessary.

Before forensic DNA techniques were implemented, species identification was largely determined by serological methods. Currently, most forensic laboratories perform DNA quantitation prior to DNA profile analysis. The quantitation method specifically detects higher-primate DNA. The presence of DNA measured by the quantitation assays concurrently identifies a sample as being of human origin (since crimes involving primate blood are extremely rare). Thus, species identification is usually not performed in forensic laboratories.

Nevertheless, species identification assays can be useful for screening to exclude or eliminate nonhuman samples unrelated to an investigation. Thus, it is practical for small laboratories to eliminate unnecessary analyses due to time and budget constraints. Additionally, species identification kits such as immunochromatographic devices allow field testing by crime scene investigators.

In cases involving the killing, trading, and possession of products derived from species that are protected from illegal hunting, it may be necessary to identify the animal species prior to further analysis. Species identification using DNA analysis can be performed using commonly used loci at the mitochondrial cytochrome b gene (*Cytb*), the cytochrome c oxidase I gene (*COI*), and the D-loop region. This type of identification is usually within the scope of wildlife forensic science and is thus not discussed here.

13.1 General Considerations

Most assays for species identification are based on serological techniques, including primary and secondary binding assays. The most common primary binding assays are immunochromatographic assays. The most commonly used secondary binding assays are precipitation-based assays that rely on the binding of an antigen to an antibody, causing the formation of visible precipitation. These precipitation-based assays include ring assays, Ouchterlony assays, and crossed-over immunoelectrophoresis. These assays utilize antihuman and antianimal antibodies to identify human and animal species, respectively.

13.1.1 Types of Antibodies

An antihuman antibody that is used in the identification of human samples can be made by introducing human serum into a host animal, which then produces specific antibodies against the human serum proteins. Antibodies produced from different species of host animals may produce variations in the characteristics of reactions. Since albumin is the most abundant protein in human serum, the antihuman antibody that is produced reacts strongly with human albumin. Albumin is a protein that plays important roles in the maintenance of the vascular circulating fluid and the transportation of various substances such as nutrients, hormones, and metabolic products. Blood is drawn from the host animal and the serum portion is collected. The collected serum is a polyclonal antihuman antiserum containing a mixture of antibodies against various human serum proteins. Likewise, an antibody against animal serum proteins can also be made to identify animal species of interest.

Other antibodies such as antihuman hemoglobin (Hb) antibodies can also be used to identify the human origin of a sample. Hb is an oxygen-transport protein that is found in erythrocytes (Chapter 12). Purified Hb can be used to generate monoclonal and polyclonal antihuman Hb antibodies. Likewise, antibodies recognizing glycophorin A (GPA), a human erythrocyte membrane antigen (see Section 13.2.1.2), can also be produced in a similar manner.

13.1.2 Titration of Antibodies

Recall that the ratio of antigen to antibody is critical for the success of a secondary reaction (Chapter 10). An extreme excess of antigen or antibody concentrations can inhibit secondary reactions. The prozone and postzone phenomena must be considered, and the concentrations of antigen and antibody must be carefully determined for forensic serology assays. For instance, in the prozone situation, a false-negative reaction may occur due to the presence of a high concentration of antibody.

Quality-control procedures can be used to estimate the amount of a specific antibody that is present, often via titration (Figure 13.1). To titrate an antiserum, a series of dilutions are made and each dilution is then tested for activity using precipitation or agglutination methods. The reciprocal of the highest dilution giving a positive reaction is known as the *titer*. This reflects the amount of antibody in the antiserum. Additionally, the polyclonal antiserum is a mixture of antibodies; thus, the reaction of the antiserum may vary from animal to animal (of the same species). Each lot or batch of antiserum must be validated by titration.

13.1.3 Antibody Specificity

In addition to the titer, the specificity of the antihuman antibody must be tested. Most antihuman antibodies usually have cross-reactivity with higher primates. This is not a great

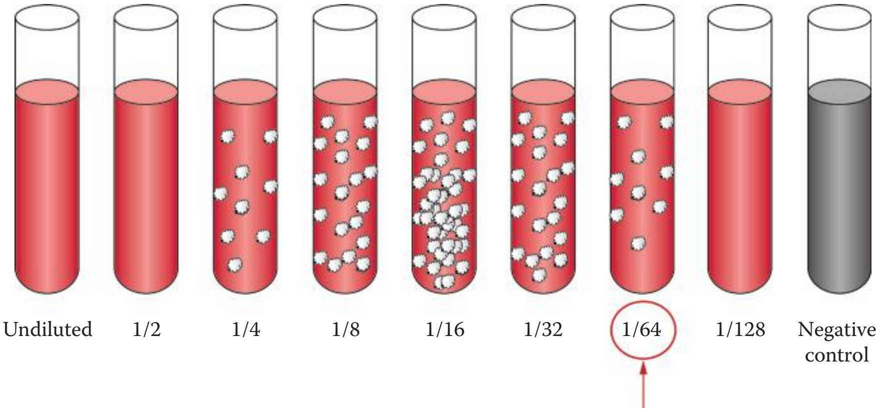

Figure 13.1 Titration of antibodies. Serum is serially diluted and a constant amount of antigen is applied to each tube. The mixture is incubated, allowing agglutination to occur. The reciprocal of the highest dilution giving a positive agglutination reaction is 64 (the titer). (© Richard C. Li.)

concern because crimes involving nonhuman primates are very rare. Nevertheless, the antihuman antibody must not cross-react with other commonly encountered animals. Antisera and positive control samples must be validated for cross-reactivity. Tissue specificity must also be validated. The antiserum against human serum is usually reactive with other human biological fluids such as semen and saliva.

13.1.4 Optimal Conditions for Antigen–Antibody Binding

A number of factors can affect antigen–antibody binding. For example, increasing ionic strength can inhibit the binding of an antigen and an antibody. Stronger inhibition is usually observed for ions with large ionic radii and small radii of hydration. It is believed that the lower degree of hydration permits interactions of ions and the antibody-binding site, leading to inhibition. A proper buffer system must be selected in serological assays to ensure reliable results. The introduction of polymers can facilitate precipitation in secondary binding reactions because the presence of a polymer in a solution decreases the solubility of proteins. Linear hydrophilic polymers with high molecular weights (e.g., polyethylene glycol) are preferred. Additional factors such as temperature and pH can also affect antigen–antibody binding.

13.2 Assays

Samples can be prepared by cutting out a portion of a stain or scraping stains from a surface. A sample is usually extracted with a small volume of saline or buffer. The extracted sample can be tested using the assays described in the following subsections. Controls should be included, for example, by using a known human serum as a positive control and an extraction blank as a negative control.

13.2.1 Immunochromatographic Assays

Immunochromatographic assays are rapid, specific, and sensitive and can be used in both laboratory and field tests for species identification. Two types of assays are discussed, including those based on the detection of human erythrocyte proteins. Chapter 11 discusses the principle of immunochromatographic assays in more detail.

13.2.1.1 Identification of Human Hemoglobin Protein

Commercially produced immunochromatographic kits such as the Hexagon OBTI (Human Gesellschaft für Biochemica und Diagnostica mbH, Wiesbaden) and the ABAcard HemaTrace® (Abacus Diagnostics, California) are available. They utilize the antibody–antigen–antibody sandwich method by using antibodies that recognize human Hb. The ABAcard HemaTrace assay utilizes a labeled monoclonal antihuman Hb antibody contained in a sample well, and a polyclonal antihuman Hb antibody immobilized at a test zone of a nitrocellulose membrane. Additionally, an antiglobulin that recognizes the antibody is immobilized onto a control zone (Figure 13.2).

A sample can be prepared by cutting a small portion (2 mm diameter) of a stain or a swab. Each sample is extracted for 5 min in 2 mL of extraction buffer. A longer extraction time may be used for older stains. The samples are loaded into the sample well, and the antigen in the sample binds to the labeled antibody in the well to form an antigen–antibody complex, which then diffuses across the nitrocellulose membrane. At the test zone, the solid-phase antihuman Hb antibody binds to the antigen–antibody complex to form a labeled antibody–antigen–antibody sandwich.

The ABAcard HemaTrace® uses a pink dye that is visualized in a positive result as a pink horizontal line at the test zone (Figure 13.3). In the control zone, unbound labeled antihuman

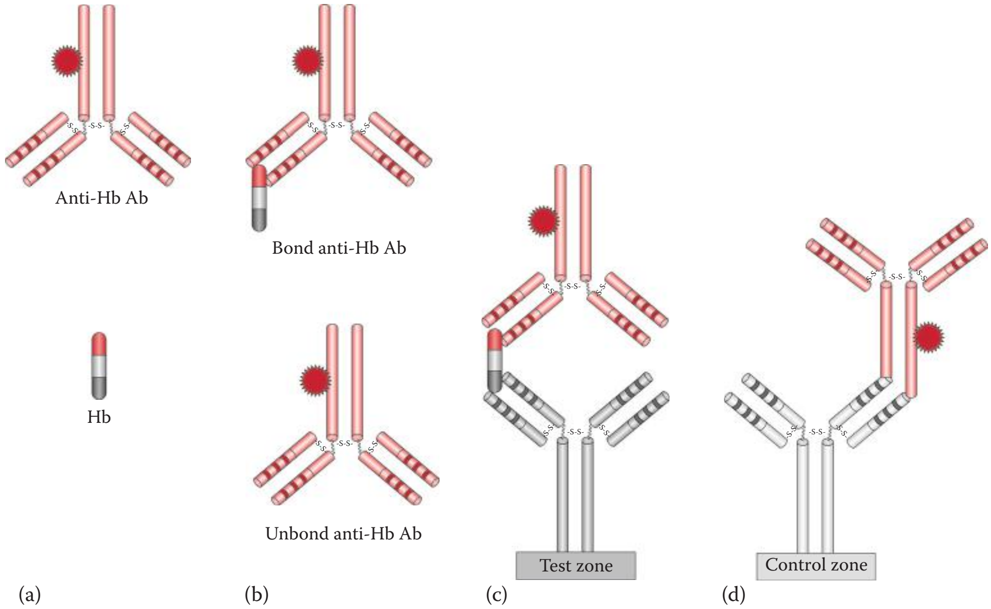

Figure 13.2 Immunochromatographic assays for the identification of Hb in human blood. (a) In a sample well, Hb in a blood sample is mixed with a labeled anti-Hb Ab. (b) The Hb binds to the labeled anti-Hb Ab to form a labeled Ab–Hb complex. The complex diffuses toward the test zone. (c) At the test zone, the labeled Ab–Hb complex binds to an immobilized anti-Hb Ab to form a labeled Ab–Hb–Ab sandwich. (d) At the control zone, the labeled anti-Hb Ab binds to an immobilized antiglobulin and is captured. Ab and Hb represent the antibody and hemoglobin, respectively. (© Richard C. Li.)

Hb antibody binds to the solid-phase antiglobulin. This antibody–antiglobulin complex at the control zone also produces a pink horizontal line. The test is considered valid only if the line in the control zone is observed. The presence of human Hb results in a pink line at both the test and control zones. The absence of human Hb results in a pink line in the control zone only. A positive result can appear in less than a minute.

Validation studies have revealed that the sensitivity of the ABAcard HemaTrace® can be as low as 0.07 µg/mL of Hb. The normal blood Hb concentration is 14–18 and 12–16 g/dL among males and females, respectively. This assay is more sensitive than the Kastle–Meyer assay (Chapter 12). Additionally, the assay is responsive to aged stains and degraded materials. Specificity studies have shown that it is specific for blood of higher primates, including humans. However, it is also responsive to seminal stains, and oral, vaginal, anal, and rectal swabs. It is believed that these biological fluids contain very low amounts of Hb, which can still be detected by highly sensitive assays. However, if the concentration of blood is too high, a false negative can result due to the high-dose hook effect described in Chapter 11.

13.2.1.2 Identification of Human Glycophorin A Protein

Commercially produced immunochromatographic kits such as RSID™-Blood (Independent Forensics, Hillside, IL) use antibodies that recognize human GPA (Figure 13.4). A labeled monoclonal antihuman GPA antibody is contained in a sample well, and a second monoclonal antihuman GPA antibody, to a different epitope of GPA, is immobilized onto a test zone of the membrane. An antiglobulin that recognizes the antibody is immobilized onto a control zone (Figure 13.5).

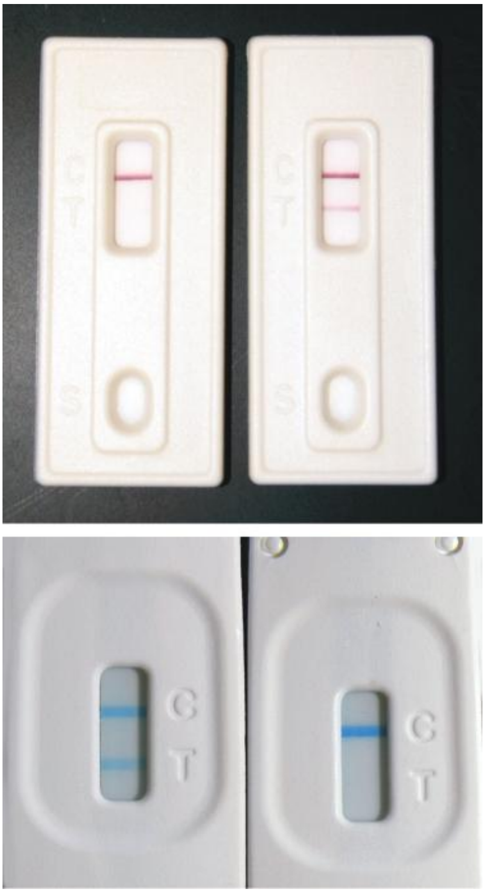

Figure 13.3 Human blood identification using immunochromatic devices. Top: In an assay using an ABAcard HemaTrace device, the negative (left) and positive (right) results are shown. The "C" band indicates that the test is valid. The "T" band indicates the presence of human blood. The sample well is labeled "S." (© Richard C. Li.) Bottom: The positive (left) and negative (right) results are shown using a Hexagon-OBTI device. The "C" band indicates that the test is valid. The "T" band indicates the presence of human blood. (From Ramsthaler, F., et al., Postmortem interval of skeletal remains through the detection of intraosseal hemin traces. A comparison of UV-fluorescence, luminol, Hexagon-OBTI(R), and Combur(R) tests. *Forensic Sci Int*, 209, 59–63, 2011. With permission.)

The sample can be collected by cutting out a small portion of a stain or a swab. The sample is then extracted overnight in an extraction buffer. The extract is removed and mixed with a running buffer. The assay is carried out by loading the extracted sample into the sample well. Again, the presence of GPA results in a pink line at both the test zone and the control zone, while the absence of GPA results in a pink line in the control zone only. The test is considered valid only if the line in the control zone is observed. A result can be read after 10 min.

Validation studies revealed that the sensitivity of the RSID kit can be as low as 100 nL of human blood. Species specificity studies showed no cross-reactivity with various animal species, including nonhuman primates. Biological fluid specificity studies revealed that the kit is not responsive to other human biological fluids such as semen, saliva, urine, milk, and amniotic and vaginal fluid. No high-dose hook effects were observed in samples containing up to 5 μL of blood.

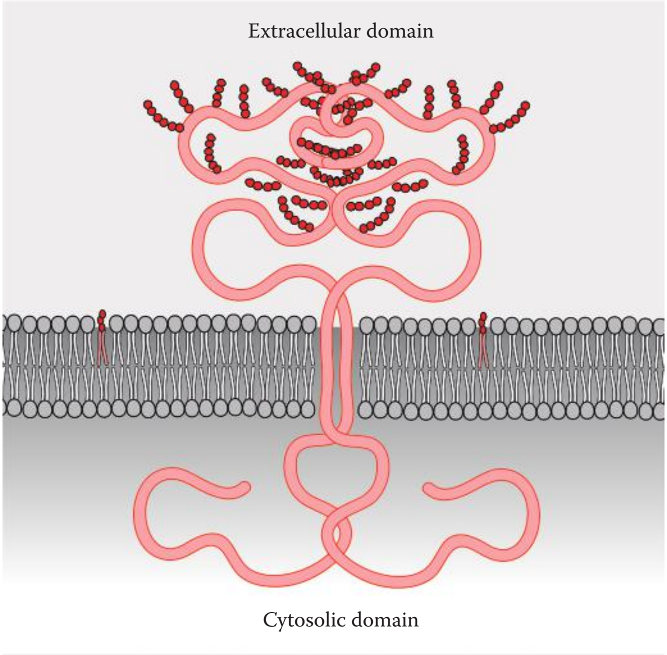
Extracellular domain
Cytosolic domain

Figure 13.4 Diagram of the structure of glycophorin (GPA) protein. GPA is a transmembrane protein on the human erythrocyte membrane. A GPA dimer is shown. The extracellular domain of the GPA is glycosylated with carbohydrate side chains (red). Various GPA epitopes are antigenic determinants of several blood group systems. (© Richard C. Li.)

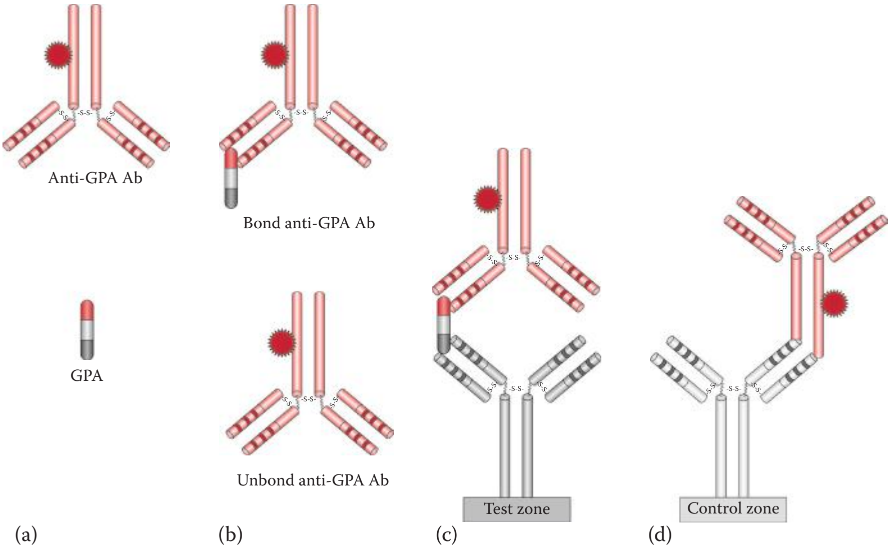
Anti-GPA Ab
GPA
Bond anti-GPA Ab
Unbond anti-GPA Ab
Test zone
Control zone
-S-S-
(a)
(b)
(c)
(d)

Figure 13.5 Immunochromatographic assays for the identification of GPA in human blood. (a) In a sample well, GPA in a blood sample is mixed with a labeled anti-GPA Ab. (b) The GPA binds to the labeled anti-GPA Ab to form a labeled Ab–GPA complex. (c) At the test zone, the labeled Ab–GPA complex binds to an immobilized anti-GPA Ab to form a labeled Ab–GPA–Ab sandwich. (d) At the control zone, the labeled anti-GPA Ab binds to an immobilized antiglobulin and is captured. Ab and GPA represent the antibody and GPA protein, respectively. (© Richard C. Li.)

13.2.2 Double Immunodiffusion Assays

13.2.2.1 Ring Assay

Chapter 11 discussed the basic principle of ring assay in detail. In this double immunodiffusion assay, an antihuman antibody reagent is placed at the bottom of a test tube and a bloodstain extract is placed on top of the bottom layer, as illustrated in Figure 13.6. The procedure is described in Box 13.1. In a positive reaction, a white precipitate between the two layers observed after several minutes indicates that a sample is of human origin. If the bloodstain extract is not human, no precipitation should appear.

13.2.2.2 Ouchterlony Assay

The basic principle of this double immunodiffusion assay has also been discussed in Chapter 11. The procedure of the Ouchterlony assay is described in Box 13.2. In a positive reaction, a line of precipitate will form between each antigen well and antibody well. This assay can also determine the similarity of the antigens (Figure 13.7). During the diffusion process, different antigen–antibody complexes migrate at different rates. Consequently, a separate line of precipitate will appear in the gel for each antigen–antibody complex. In an assay in which two antigens are loaded in adjacent wells and an antibody in the third well, the following results can be observed:

- If the two antigens are identical, the two lines will become fused. This phenomenon is referred to as *identity*.
- If the two antigens are totally unrelated, the lines will cross each other but not fuse; this is known as *nonidentity*.
- If the two antigens are related (share a common epitope) but are not identical, the lines will merge with spur formation. The spurs are continuations of the line formed by the antigen due to its unique epitope. This phenomenon is known as *partial identity*.

Thus, in a species test, a positive result is noted when the precipitate lines for the positive controls and the samples fuse. No spur formation should be observed.

13.2.3 Crossed-Over Electrophoresis

This method is a combination of immunodiffusion and electrophoresis (also see Chapter 11). The procedure for the assay is described in Box 13.3. With this technique, a sharp precipitate band is visualized in a positive reaction (Figure 13.8). However, false-negative results can occur due to the postzone phenomenon, in which excess antigen may inhibit precipitation. In this

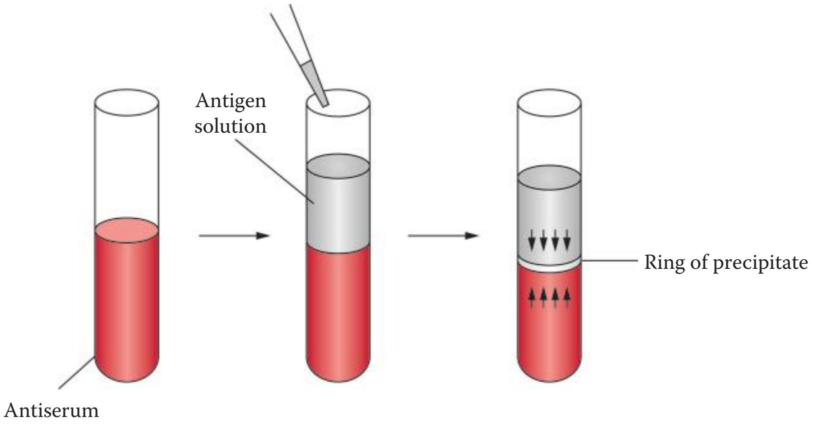

Figure 13.6 Ring assay. The antigen solution is carefully applied over the antiserum solution. Incubation may be necessary in order to form a ring of precipitate. (© Richard C. Li.)

BOX 13.1 RING ASSAY PROCEDURE

Sample preparation and extraction

- Extract a portion of a stain with saline at 4°C overnight.

Controls

- Include a positive control (known human serum sample) and a negative control (extraction blank).

Loading of antibody and samples

- Spin the antihuman antibody in a microfuge and transfer the supernatant into test tubes or capillary tubes (depending on the volume of the stain and the antiserum extracted).
- Place the sample carefully over the top of the antiserum solution, which is usually denser than the sample.

Immunodiffusion reaction

- Carry out the reaction at room temperature.
- In a positive reaction, white precipitate between the two layers can be observed after several minutes. This indicates that the sample is of human origin. No precipitate is formed if a bloodstain extract is from a nonhuman origin.

BOX 13.2 OUCHTERLONY ASSAY PROCEDURE

Sample preparation and extraction

- Cut out a small portion (approximately 5×5 mm) of a stain or a portion of a swab.
- Extract at room temperature in 100 μL of water for 30 min. The extract can be diluted if necessary. Alternatively, a very small piece of the stain or swab can be inserted directly into the well.

Controls

- Positive (known serum)
- Negative (extraction blank)
- Substrate controls (extraction of substrate from unstained area) if applicable

Agarose gel preparation

Heat a suspension of agarose (4%) until liquefied. Cool the solution in a water bath at 55°C. Pour the agarose onto a piece of glass slide and let the gel solidify to a thickness of about 2–3 mm. Alternatively, a polyester support film such as GelBond (Cambrex, New Zealand) can be used as a gel support. The agarose should be poured onto the hydrophilic side of a piece of GelBond film (6×9 cm). Punch wells consisting of a central well surrounded by four wells using a template.

Loading antibodies and samples

- Apply antihuman antibody to the central well. Apply the positive control to one of the surrounding wells.
- Apply the sample(s) in question next to a positive control.
- Apply negative and substrate controls to the remaining wells; only one negative control is needed per gel.

Immunodiffusion reaction

Incubate the plate overnight in a moisture chamber at 37°C.

Staining

Soak the gel overnight in saline solution and then soak it in deionized water for 10 min. Repeat once. Dry the gel between paper towels with a weight on top for 30 min. Dry in an oven for 30 min. Stain the gel with Coomassie blue. Stained precipitate bands appear blue.

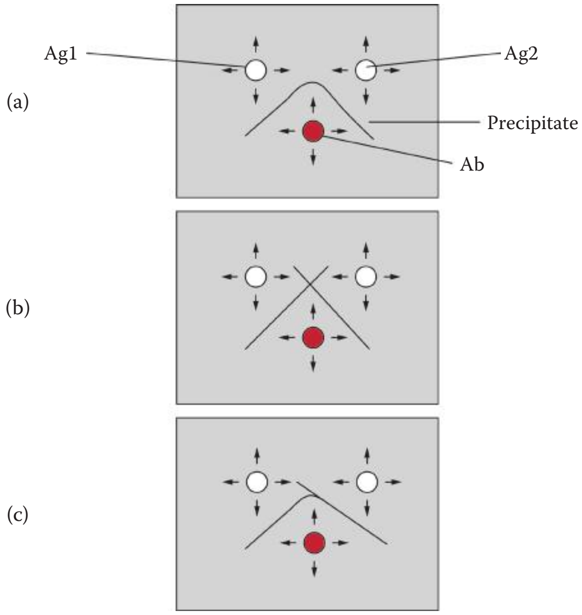

Figure 13.7 Results of Ouchterlony assay. Ab indicates an antibody. Ag1 and Ag2 are the antigen samples in question. (a) Identity. The two antigens are identical (fused line). (b) Nonidentity. The two antigens are unrelated (spur). (c) Partial identity. The two antigens are related but not identical (fused line with spur). (© Richard C. Li.)

situation, the sample can be diluted and the assay can be repeated. False-negative results can also occur due to simple mistakes made during electrophoresis:

- Electrophoresis is carried out in the opposite direction, which results in samples running off the gel.
- Electrophoresis is carried out using an incorrect buffer system, affecting antigen–antibody binding. The amount of current applied during the electrophoresis is too strong and generates heat and denatures proteins.

BOX 13.3 CROSSED-OVER ELECTROPHORESIS PROCEDURE

Sample preparation and extraction

- Cut out a small portion (~5×5 mm) of a stain or a portion of a swab.
- Extract at room temperature in 100 μL of water for 30 min. The extract can be diluted if necessary. Alternatively, a very small piece of the stain or swab can be inserted directly into the well.

Controls

- Positive (known serum)
- Negative (extraction blank)
- Substrate (extraction of substrate from unstained area) if applicable

Agarose gel preparation

Heat a suspension of agarose (4%) until liquefied. Cool the solution in a water bath at 55°C. Pour the agarose onto a piece of glass slide and let it solidify. Alternatively, a polyester support film such as GelBond can be used as a gel support. The agarose should be poured onto the hydrophilic side of a piece of GelBond (6×9 cm). Punch small wells (about 1–2 mm) in rows using a template.

Loading antibodies and samples

- Apply antihuman antibody in one row of wells.
- Apply samples in the other row of wells. Apply the positive, negative, and substrate controls.

Electrophoresis

Submerge the agarose gel in an electrophoresis tank in proper orientation. The wells containing antihuman antibody should be closest to the anode (positive electrode) and the wells containing samples should be closest to the cathode (negative electrode). During electrophoresis, the antibody in the antiserum should migrate toward the cathode while the antigen migrates toward the anode. Electrophoresis is carried out at 10 V/cm for 20 min.

Staining

Soak the gel overnight in a saline solution and then soak it in deionized water for 10 min. Repeat once. Dry the gel between paper towels with a weight on top for 30 min. Dry in an oven for 30 min. Stain the gel with Coomassie blue. Stained precipitate bands appear blue.

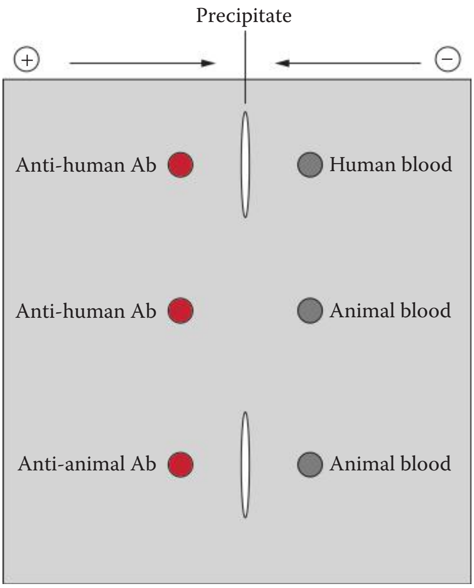

Figure 13.8 Results of crossed-over electrophoresis. A precipitate line is formed between a human blood sample and an antihuman antibody. No precipitate line is formed when the antihuman antibody is tested for an animal blood sample. A precipitate line is formed between the animal blood sample and the antianimal antibody. (© Richard C. Li.)

Bibliography

Allison, A.C. and J.A. Morton, Species specificity in the inhibition of antiglobulin sera: A technique for the identification of human and animal bloods. *J Clin Pathol*, 1953, **6**(4): 314–319.

Berger, B., et al., Validation of two canine STR multiplex-assays following the ISFG recommendations for non-human DNA analysis. *Forensic Sci Int Genet*, 2014, **8**(1): 90–100.

Bhatia, R.Y., Specificity of some plant lectins in the differentiation of animal blood. *Forensic Sci*, 1974, **4**(1): 47–52.

Coomber, N., et al., Validation of a short tandem repeat multiplex typing system for genetic individualization of domestic cat samples. *Croat Med J*, 2007, **48**(4): 547–555.

Dawnay, N., et al., Validation of the barcoding gene *COI* for use in forensic genetic species identification. *Forensic Sci Int*, 2007, **173**(1): 1–6.

DeForest, P., R. Gaensslen, and H.C. Lee, *Forensic Science: An Introduction to Criminalistics*, 1983. New York: McGraw-Hill.

Dorrill, M. and P.H. Whitehead, The species identification of very old human blood-stains. *Forensic Sci Int*, 1979, **13**(2): 111–116.

El-Sayed, Y.S., et al., Using species-specific repeat and PCR-RFLP in typing of DNA derived from blood of human and animal species. *Forensic Sci Med Pathol*, 2010, **6**(3): 158–164.

Grobbelaar, B.G., D. Skinner, and H.N. van de Gertenbach, The anti-human globulin inhibition test. In the identification of human blood stains. *J Forensic Med*, 1970, **17**(3): 103–111.

Hochmeister, M.N., et al., Validation studies of an immunochromatographic 1-step test for the forensic identification of human blood. *J Forensic Sci*, 1999, **44**(3): 597–602.

Imaizumi, K., et al., Development of species identification tests targeting the 16S ribosomal RNA coding region in mitochondrial DNA. *Int J Legal Med*, 2007, **121**(3): 184–191.

Johnston, S., J. Newman, and R. Frappier, Validation study of the Abacus Diagnostics ABAcard® HemaTrace® membrane test for the forensic identification of human blood. *Can Soc Forensic Sci*, 2003, **36**(3): 173–183.

Juusola, J. and J. Ballantyne, Messenger RNA profiling: A prototype method to supplant conventional methods for body fluid identification. *Forensic Sci Int*, 2003, **135**(2): 85–96.

Kanthaswamy, S., et al., Quantitative real-time PCR (qPCR) assay for human-dog-cat species identification and nuclear DNA quantification. *Forensic Sci Int Genet*, 2012, **6**(2): 290–295.

Karlsson, A.O. and G. Holmlund, Identification of mammal species using species-specific DNA pyrosequencing. *Forensic Sci Int*, 2007, **173**(1): 16–20.

Kun, T., et al., Developmental validation of Mini-DogFiler for degraded canine DNA. *Forensic Sci Int Genet*, 2013, **7**(1): 151–158.

Lawton, M.E. and J.G. Sutton, Species identification of deer blood by isoelectric focusing. *J Forensic Sci Soc*, 1982, **22**(4): 361–366.

Lee, H.C., Identification and grouping of bloodstains, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 267–337, 1982. Englewood Cliffs, NJ: Prentice Hall.

Lee, H.C. and P.R. De Forest, A precipitin-inhibition test on denatured bloodstains for the determination of human origin. *J Forensic Sci*, 1976, **21**(4): 804–810.

Lowenstein, J.M., et al., Identification of animal species by protein radioimmunoassay of bone fragments and bloodstained stone tools. *Forensic Sci Int*, 2006, **159**(2–3): 182–188.

Nakaki, S., et al., Study of animal species (human, dog and cat) identification using a multiplex single-base primer extension reaction in the cytochrome b gene. *Forensic Sci Int*, 2007, **173**(2–3): 97–102.

Ogden, R., Forensic science, genetics and wildlife biology: Getting the right mix for a wildlife DNA forensics lab. *Forensic Sci Med Pathol*, 2010, **6**(3): 172–179.

Parson, W., et al., Species identification by means of the cytochrome b gene. *Int J Legal Med*, 2000, **114**(1–2): 23–28.

Ramsthaler, F., et al., Postmortem interval of skeletal remains through the detection of intraosseal hemin traces. A comparison of UV-fluorescence, luminol, Hexagon-OBTI(R), and Combur(R) tests. *Forensic Sci Int*, 2011, **209**: 59–63.

Sato, I., et al., Forensic hair analysis to identify animal species on a case of pet animal abuse. *Int J Legal Med*, 2010, **124**(3): 249–256.

Shaler, R.C., Modern forensic biology, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 525–614, 2002. Upper Saddle River, NJ: Pearson Education.

Sivaram, S., et al., Differentiation between stains of human blood and blood of monkey. *Forensic Sci*, 1975, **6**(3): 145–152.

Spear, T.F. and S.A. Binkley, The HemeSelect test: A simple and sensitive forensic species test. *J Forensic Sci Soc*, 1994, **34**(1): 41–46.

Tabata, N. and M. Morita, Immunohistochemical demonstration of bleeding in decomposed bodies by using anti-glycophorin A monoclonal antibody. *Forensic Sci Int*, 1997, **87**(1): 1–8.

Tanaka, Y., et al., Enzyme-linked immunospot assay for detecting cells secreting antibodies against human blood group A epitopes. *Transplant Proc*, 2003, **35**(1): 555–556.

Tobe, S. and A. Linacre, Species identification of human and deer from mixed biological material. *Forensic Sci Int*, 2007, **169**(2–3): 278–279.

Tobe, S.S., et al., Recovery of human DNA profiles from poached deer remains part 2: Improved recovery protocol without the need for LCN analysis. *Sci Justice*, 2013, **53**(1): 23–27.

Tsutsumi, H. and Y. Katsumata, Forensic study on stains of blood and saliva in a chimpanzee bite case. *Forensic Sci Int*, 1993, **61**(2–3): 101–110.

van Asch, B. and A. Amorim, Capillary electrophoresis analysis of a 9-plex STR assay for canine genotyping. *Methods Mol Biol*, 2012, **830**: 231–240.

van Asch, B., et al., A new autosomal STR nineplex for canine identification and parentage testing. *Electrophoresis*, 2009, **30**(2): 417–423.

van Asch, B., et al., Forensic analysis of dog (*Canis lupus familiaris*) mitochondrial DNA sequences: An inter-laboratory study of the GEP-ISFG working group. *Forensic Sci Int Genet*, 2009, **4**(1): 49–54.

Virkler, K. and I.K. Lednev, Blood species identification for forensic purposes using Raman spectroscopy combined with advanced statistical analysis. *Anal Chem*, 2009, **81**(18): 7773–7777.

Whitehead, P.H. and A. Brech, A micro-technique involving species identification and ABO grouping on the same fragment of blood. *J Forensic Sci Soc*, 1974, **14**(2): 109–110.

Wictum, E., et al., Developmental validation of DogFiler, a novel multiplex for canine DNA profiling in forensic casework. *Forensic Sci Int Genet*, 2013, **7**(1): 82–91.

Wilson-Wilde, L., et al., Current issues in species identification for forensic science and the validity of using the cytochrome oxidase I (COI) gene. *Forensic Sci Med Pathol*, 2010, **6**(3): 233–241.

14

Identification of Semen

14.1 Biological Characteristics

A typical ejaculation releases 2–5 mL of semen, which contains seminal fluid and sperm cells (*spermatozoa*)**.** A normal sperm count ranges from 10^7 to 10^8 spermatozoa per milliliter of semen. The spermatozoa are formed from spermatogonia in the seminiferous tubules of the testes. This process of generating spermatozoa is referred to as spermatogenesis (Figure 14.1). The spermatozoa are then transported and stored in the tubular network of the epididymis where they undergo functional maturation (spermatogenesis and maturation take approximately 3 months). The epididymis joins the ductus deferens, which transports matured sperm from the epididymis to the ejaculatory duct. From there, spermatozoa follow the ejaculatory ducts into the prostatic urethra where they are joined with secretions from the prostate. Figure 14.2 illustrates the anatomy of the male reproductive system.

Seminal fluid is a complex mixture of glandular secretions. A typical sample of seminal fluid contains the combined secretions of several accessory glands. Seminal vesicle fluid accounts for approximately 60% of the ejaculate. Various proteins secreted by the seminal vesicles play a role in the coagulation of the ejaculate. Additionally, seminal vesicle fluid contains flavin, which causes semen to fluoresce under ultraviolet light, often utilized when searching for semen-stain evidence.

Prostatic fluid secretions account for approximately 30% of the ejaculate. The components of this fluid are complex as well. This portion of semen contains high concentrations of *acid phosphatase* (AP) and *prostate-specific antigen* (PSA). Both are useful markers for the identification of semen in forensic laboratories. The epididymis and the bulbourethral secretions each account for approximately 5% of the ejaculate.

A *vasectomy* is the surgical removal of a bilateral segment of the ductus deferens. The surgery prevents spermatozoa from reaching the distal portions of the male reproductive tract. However, a vasectomized male can still produce ejaculate that contains only seminal vesicle fluid, prostatic fluid, and bulbourethral fluid. The condition by which males have abnormally low sperm counts is known as *oligospermia***.** *Azoospermia* is a condition that causes males to produce no spermatozoa. However, the secretion of seminal fluid is not affected in males who have these conditions. DNA derived from epithelial cells can be isolated from the seminal fluids of these individuals.

14.1.1 Spermatozoa

A human spermatozoon has three morphologically distinct structures: the head, the middle piece, and the tail (Figure 14.3). The head contains a nucleus with densely packed chromosomes. At the tip of the head is the *acrosomal cap*, which is a membranous compartment

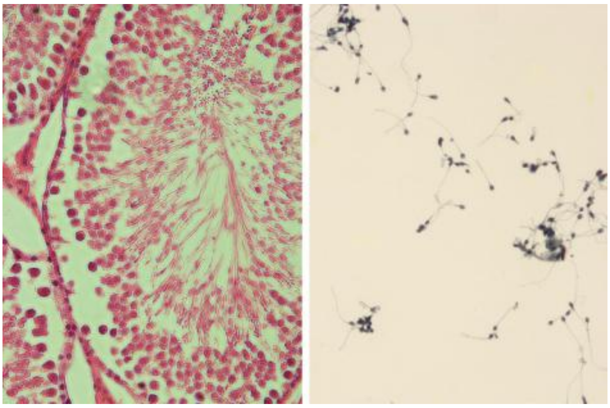

Figure 14.1 Spermatogenesis. In the seminiferous tubules (left) of the testes, spermatogonia (located at peripheral area of the seminiferous tubules) are differentiated to spermatids (located at the center of the seminiferous tubules). The spermatids are eventually differentiated to matured spermatozoa (right). (© Richard C. Li.)

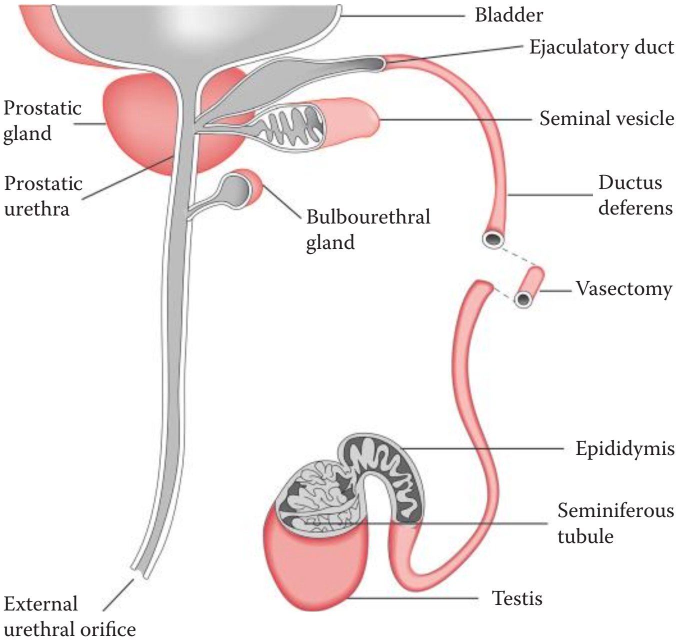
Bladder
Ejaculatory duct
Prostatic gland
Seminal vesicle
Prostatic urethra
Ductus deferens
Bulbourethral gland
Vasectomy
Epididymis
Seminiferous tubule
Testis
External urethral orifice

Figure 14.2 Male reproductive system and accessory glands (unilateral view). (© Richard C. Li.)

containing enzymes essential for fertilization. The head is attached to the middle piece through a short neck where the mitochondria that provide the energy for moving the tail are located. The tail or flagellum is responsible for spermatozoon motility. In contrast to other cell types, a mature spermatozoon lacks various intracellular organelles such as an endoplasmic reticulum, Golgi apparatus, lysosomes, and peroxisomes. In a normal male, at least 60% of spermatozoa have normal morphology, so morphological abnormalities can often be observed.

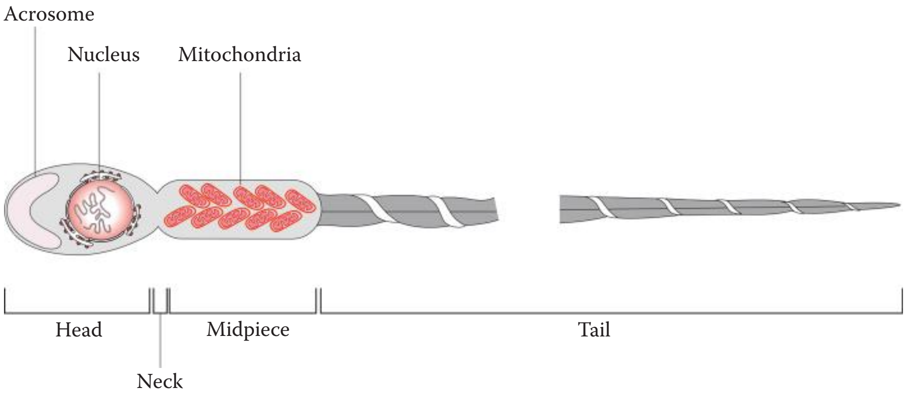

Figure 14.3 Structure of spermatozoon. (© Richard C. Li.)

14.1.2 Acid Phosphatase

Acid phosphatase (AP) consists of a group of phosphatases with optimal activity in an acidic pH environment. The greatest forensic importance of AP is that the prostate-derived AP contributes most of the AP activity present in semen. AP levels in semen are not affected by vasectomies. AP isoenzymes are also found in other tissues (Section 16.2).

The half-life of AP activity at 37°C is 6 months. However, the half-life is decreased if a sample is stored in a wet environment. AP activity can be detected from dry seminal stains stored at –20°C up to 1 year. Low levels of prostatic AP are present in the sera of healthy males. Elevated levels of prostatic AP found in serum are useful in diagnosing and monitoring prostate carcinoma. Many AP tests utilized in clinical testing may be used to identify semen for forensic applications.

14.1.3 Prostate-Specific Antigen

Prostate-specific antigen (PSA) is a major protein present in seminal fluid at concentrations of 0.5–2.0 mg/mL. PSA is produced in the prostate epithelium and secreted into the semen. PSA can also be found in the paraurethral glands, perianal glands, apocrine sweat glands, and mammary glands. Thus small quantities can be detected in urine, fecal material, sweat, and milk. PSA can also be found at much lower levels in the bloodstream. An elevated plasma PSA is present in prostate cancer patients, and it is widely used as a screening test for this disease. PSA is also elevated in cases of benign prostatic hyperplasia and prostatitis. The synthesis of PSA is stimulated by androgen, a steroid hormone.

PSA is a protein that has a molecular weight of 30 kDa and is thus also known as *P30*. It is responsible for hydrolyzing *semenogelin* (Sg), which mediates gel formation in semen (Section 14.1.4). PSA is a member of the tissue *kallikrein* (serine protease) family and is encoded by the *KLK3* locus located on chromosome 19. In addition to PSA, other tissue kallikreins encoded by *KLK2* and *KLK4* loci are expressed in the prostate. The half-life for PSA in a dried semen stain is about 3 years at room temperature. The half-life is greatly reduced when a sample is stored in wet conditions.

14.1.4 Seminal Vesicle–Specific Antigen

Human *seminal vesicle–specific antigen* (SVSA) includes two major types, semenogelin I (SgI) and semenogelin II (SgII), and constitutes the major seminal vesicle–secreted protein in semen. On ejaculation, SVSA forms a coagulum that is liquefied after a few minutes due to the degradation of SVSA by PSA. In humans, both SgI and SgII are present in a number of tissues of the male reproductive system, including the seminal vesicles, ductus deferens, prostate, and epididymis. They are also present in several other tissues such as skeletal muscle, kidney, colon, and trachea. They have also been found in the sera of lung cancer patients.

The use of Sg as a marker for semen identification instead of PSA presents certain advantages. The concentration of Sg in seminal fluid is much higher than that of PSA, and this is beneficial for the sensitivity of detection. Sg is present in seminal fluid and absent in urine, milk, and sweat, where PSA can be found. Although Sg compounds are present in skeletal muscle, kidney, and colon, this is not a great concern because these tissue samples are not routinely collected for semen detection in sexual assault cases.

14.2 Analytical Techniques for Identifying Semen

The location of semen stains is usually carried out through visual examination. Particularly, the application of alternate light sources (ALSs) can facilitate searches for semen stains. The presumptive identification of semen is largely based on the detection of the presence of prostatic AP activity in a sample. However, most presumptive assays cannot completely distinguish prostatic AP from nonprostatic AP. Confirmatory assays for the identification of semen are available, including the microscopic examination of spermatozoa, the identification of PSA and SVSA, and the RNA-based assay.

14.2.1 Presumptive Assays

14.2.1.1 Lighting Techniques for Visual Examination of Semen Stains

Lighting techniques can be used to aid in searching for semen stains. A dried semen stain fluoresces under certain light sources such as ALSs or argon lasers. ALSs are most commonly utilized for the visual examination of semen stains (Chapter 1; Figures 14.4 and 14.5). Excitation wavelengths between 450 and 495 nm can be used, allowing for the visualization of fluorescence with orange goggles. However, this approach is not specific for semen. Other bodily fluid stains, such as saliva and urine stains, can also fluoresce with less intensity. Additionally, the intensity of the fluorescence can be affected by different colors of substrates, and the material, such as clothing, where semen stains have been deposited.

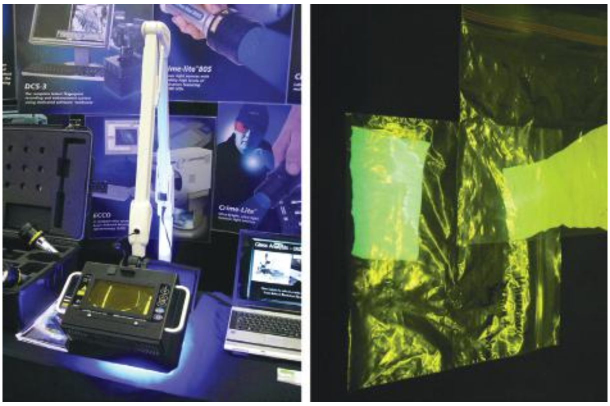

Figure 14.4 A tabletop ALS device (left) for the detection of semen stains (right). (© Richard C. Li.)

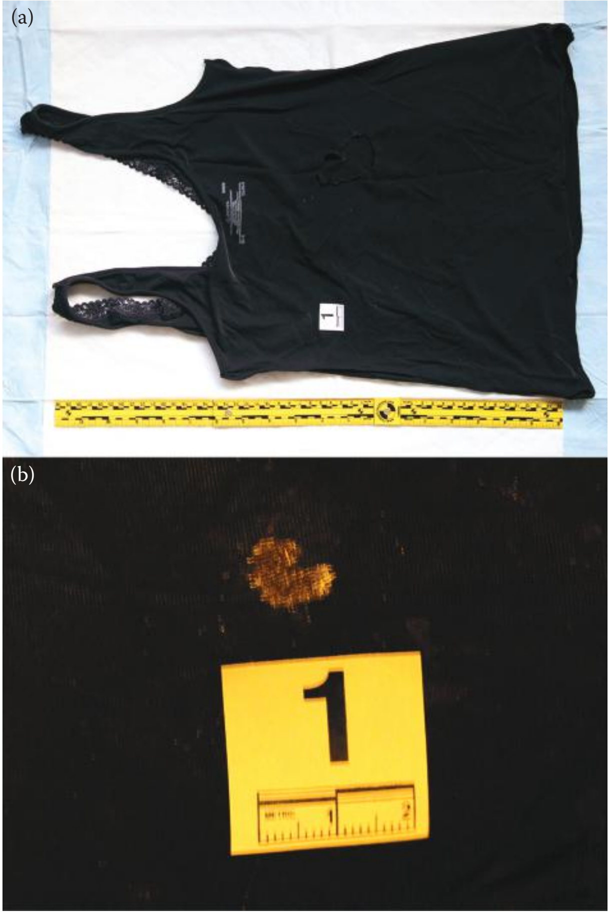

Figure 14.5 Examining a garment for semen stains. (a) A potential semen stain (labeled) is found on a garment, and (b) the stain fluoresces when irradiated with an ALS device. (© Richard C. Li.)

14.2.1.2 Acid Phosphatase Techniques

14.2.1.2.1 Colorimetric Assays

Colorimetric assays can be used for the presumptive identification of semen. The AP contained in semen can hydrolyze a variety of phosphate esters. It catalyzes the removal of the phosphate group from a substrate (Figure 14.6). Subsequently, an insoluble colored precipitate at sites of acid phosphatase activity is formed with a stabilized diazonium salt (usually in the form of zinc double salts).

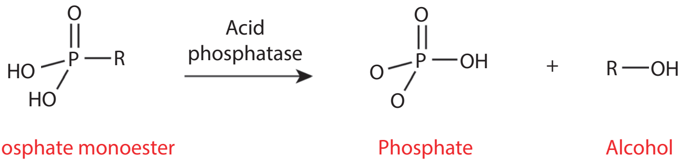

Figure 14.6 Reaction catalyzed by acid phosphatases (EC 3.1.3.2). The optimal pH of the reaction is usually under pH 7.

However, interference during a test by nonprostatic AP isoenzymes (multiple forms of AP), such as contamination by AP commonly present in vaginal secretions (Chapter 16), can create problems in specimens collected from victims. Thus, it is desirable to be able to increase the specificity of the assay for prostatic AP. One solution is the application of substrates that are hydrolyzed rapidly by the prostatic enzyme and at a slower rate by the other forms of AP isoenzymes. For example, α-naphthylphosphate and thymolphthalein monophosphate are more specific to prostatic AP than phenyl phosphate and 4-nitrophenyl phosphate (Figure 14.7). The most common method for forensic applications is the use of α-naphthyl phosphate as a substrate. In the presence of AP, α-naphthylphosphate is hydrolyzed to phosphate and α-naphthol. Subsequently, the Fast Blue B, a stabilized diazonium salt, is added to carry out an azo coupling reaction, producing a purple azo dye (Figures 14.8 through 14.10).

Prostatic AP is water soluble. Thus, a moistened cotton swab or piece of filter paper can be used to transfer a small amount of sample from a stain by briefly pressing onto the questioned stain area. The α-naphthylphosphate reagent is added to the swab or filter paper followed by the addition of Fast Blue B reagent. If a purple coloration develops within 1 min, the test is considered a positive indication for semen. Color that develops after more than 1 min may arise from the activity of nonprostatic AP.

Additionally, the prostatic enzyme is strongly inhibited by dextrorotatory tartrate ions. Thus, these inhibitors, particularly tartrate, allow a distinction to be made between prostatic AP and other AP isoenzymes. Prostate and vaginal acid phosphatase can also be distinguished by using gel electrophoresis (Chapter 16).

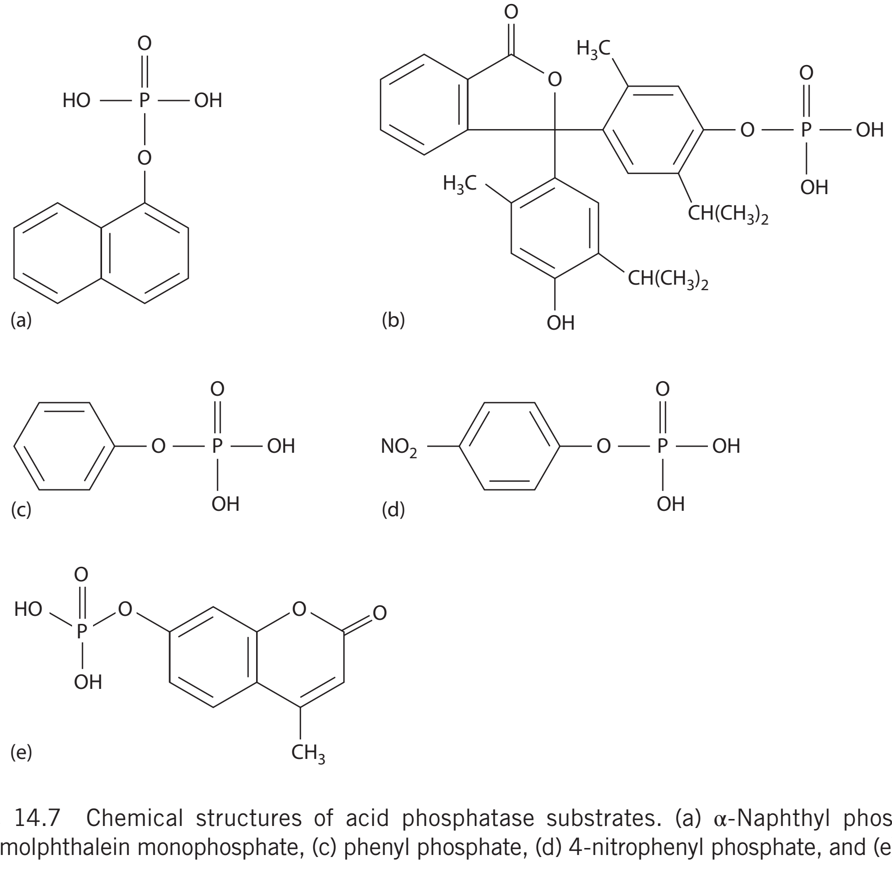

Figure 14.7 Chemical structures of acid phosphatase substrates. (a) α-Naphthyl phosphate, (b) thymolphthalein monophosphate, (c) phenyl phosphate, (d) 4-nitrophenyl phosphate, and (e) MUP.

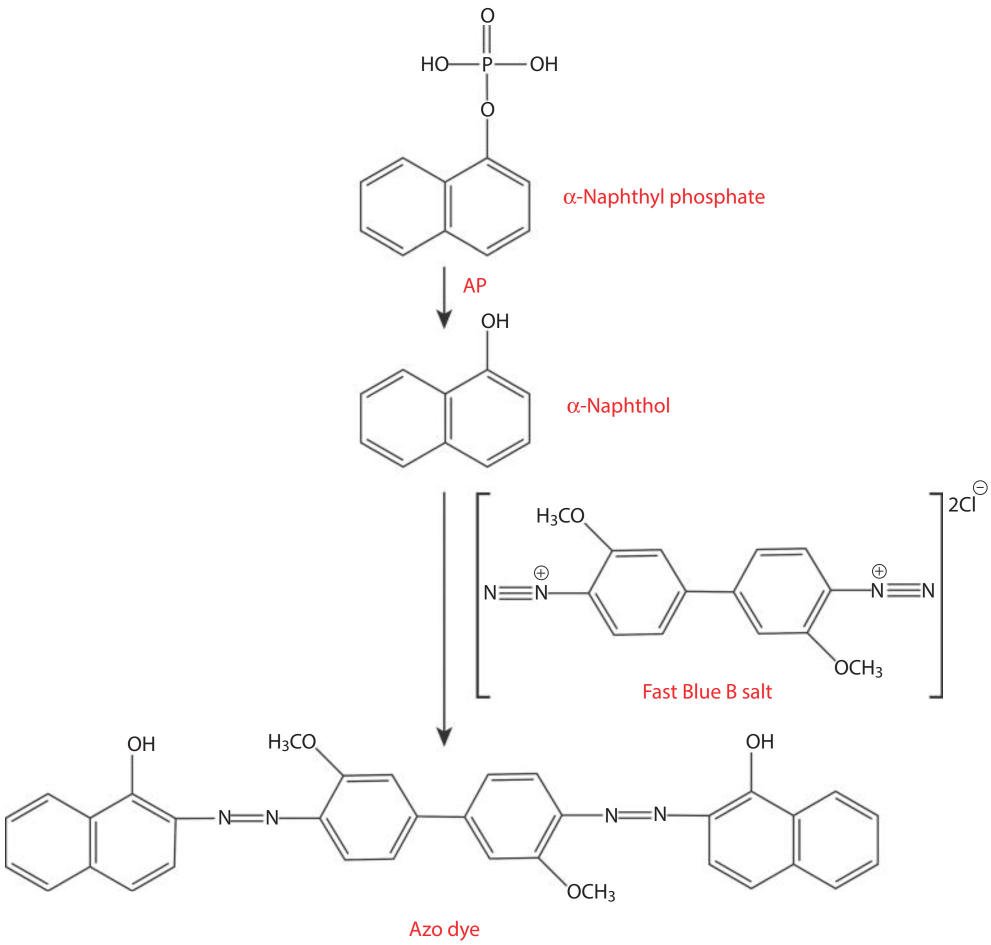
HO
P
OH
O
O
α-Naphthyl phosphate
AP
OH
α-Naphthol
H_3CO
N
N
2Cl
OCH_3
Fast Blue B salt
OH
H_3CO
N
N
N
N
OH
OCH_3
Azo dye

Figure 14.8 A colorimetric acid-phosphatase assay. In this assay, α-naphthylphosphate is hydrolyzed by acid phosphatase to phosphate and α-naphthol. The α-naphthol is subsequently converted into a purple azo dye with a diazonium salt such as Fast Blue B salt. AP, acid phosphatase.

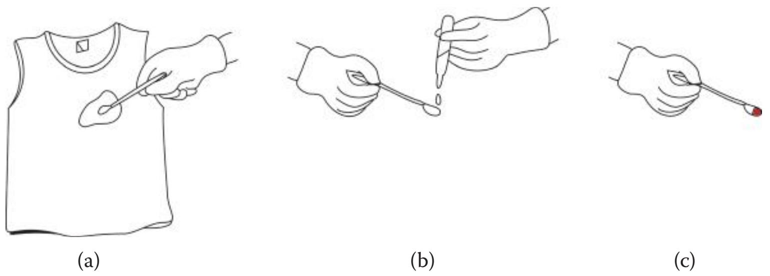

Figure 14.9 AP colorimetric assay. (a) A small amount of sample from the stain is transferred using a moistened cotton swab. (b) The substrate reagent is applied, followed by adding Fast Blue B reagent. (c) Purple coloration indicates a positive reaction. (© Richard C. Li.)

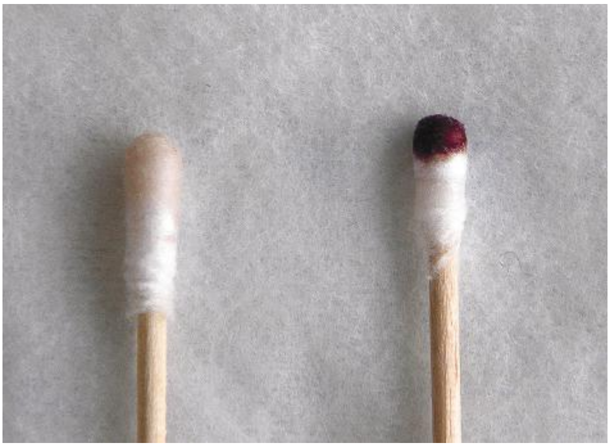

Figure 14.10 Photo of a colorimetric acid-phosphatase assay using α-naphthylphosphate as a substrate. (© Richard C. Li.)

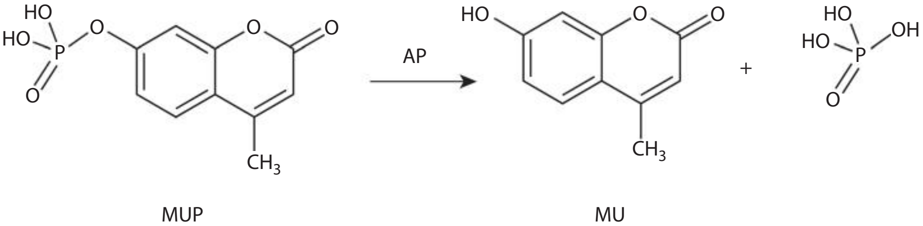

Figure 14.11 The principle of MUP assay for detecting acid phosphatase activity. In the presence of acid phosphatase, 4-methylumbelliferone phosphate (MUP) is hydrolyzed, forming phosphate and 4-methylumbelliferone (MU), which fluoresces. AP, acid phosphatase.

14.2.1.2.2 Fluorometric Assays

Fluorometric methods are more sensitive than the colorimetric detection of AP and are used for semen stain mapping. AP catalyzes the removal of the phosphate residue on a 4-methylumbelliferone phosphate (MUP) substrate (Figure 14.11), a reaction that generates fluorescence under ultraviolet light. A piece of moistened filter paper, marked for proper orientation and identification, is used for transferring the prostatic AP. The evidence to be tested, a garment for example, is covered by the filter paper. Gloved hands are used to press the filter paper onto the stained area, ensuring that the evidence is in close contact with the paper. The filter paper is lifted from the evidence and examined in a dark room using long-wave ultraviolet light to detect any background fluorescence, which is then marked on the paper. The paper can then be sprayed with MUP reagent in a fume hood. The AP reaction on the paper can be visualized immediately. Areas where semen is present can be visualized as fluorescent areas on the filter paper (Figure 14.12).

14.2.2 Confirmatory Assays

14.2.2.1 Microscopic Examination of Spermatozoa

The cells from a questioned stain on an absorbent material can be transferred to a microscope slide by extracting a small portion of a stain with water, followed by gentle vortexing. The suspension is then transferred to a slide and evaporated at room temperature or fixed with low heat. Alternatively, it can be transferred by dampening the stain with water and rubbing or rolling it onto a microscope slide.

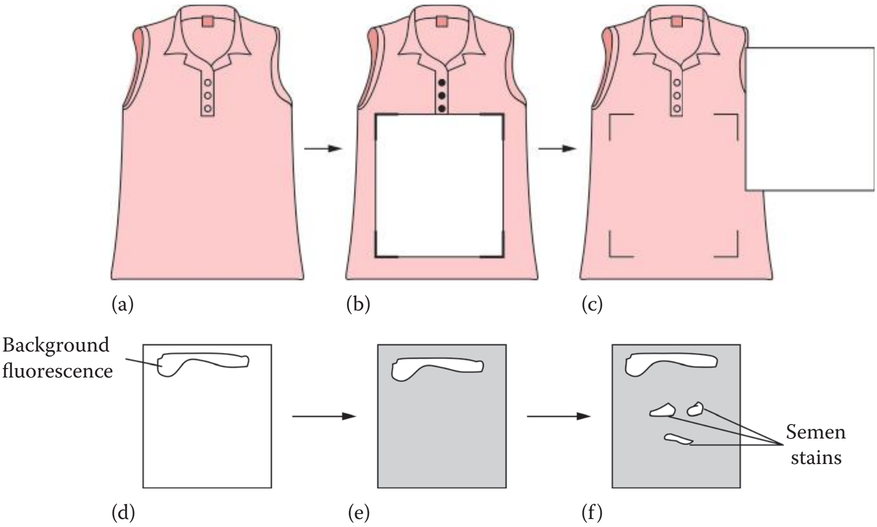

Figure 14.12 Fluorometric assay of acid phosphatase for locating semen stains. Evidence item (a) is closely covered by a piece of moistened filter paper (b) to allow transfer of a small amount of stain. The orientation of the paper is marked (c). The paper is lifted. The background fluorescence is marked under UV light (d). The filter paper is treated with MUP (e). The presence of fluorescence under ultraviolet light indicates semen stains (f). (© Richard C. Li.)

Microscopic identification of spermatozoa provides the proof of a seminal stain. Histological staining can facilitate microscopic examination. The most common staining technique is the *Christmas tree stain* (Figure 14.13). The red component known as Nuclear Fast Red (NFR) is a dye used for staining the nuclei of spermatozoa in the presence of aluminum ions. The green component, picroindigocarmine (PIC), stains the neck and tail portions of the sperm. The acrosomal cap and the nucleus stain pink-red, and the sperm tails and the midpiece stain blue-green. Epithelial cells, if present in the sample, appear blue-green and have red nuclei. Additionally, fluorescent detection utilizing SPERM HY-LITER Fluorescent Staining Kit can facilitate the identification of spermatozoa.

Laser capture microdissection (LCM) has been shown to be an effective technique for separating spermatozoa from nonsperm cells (i.e., epithelial cells from the victim) on a glass slide (Figure 14.14). This technique involves using a thin layer of a thermosensitive polymer that is placed on the surface of an LCM cap. Once spermatozoa are identified on the slide under a microscope, a polymer-containing LCM cap is placed over the spermatozoa on the slide. An infrared laser melts the polymer and causes it to adhere only to the targeted spermatozoa. The spermatozoa are then lifted off the slide. This allows spermatozoa to be separated and placed into snap-cap tubes for forensic DNA analysis.

14.2.2.2 Identification of Prostate-Specific Antigen

Over the years, a number of methods have been utilized to detect PSA: immunodiffusion, immunoelectrophoresis, enzyme-linked immunosorbent assay (ELISA), and immunochromatographic assays. ELISA and immunochromatographic assays have been found to be the most sensitive methods (Chapter 11).

14.2.2.2.1 Immunochromatographic Assays

Commercially produced immunochromatographic kits such as the PSA-check-1 (VED-LAB, Alencon), Seratec® PSA Semiquant (Seratec Diagnostica, Göttingen), and One Step ABAcard

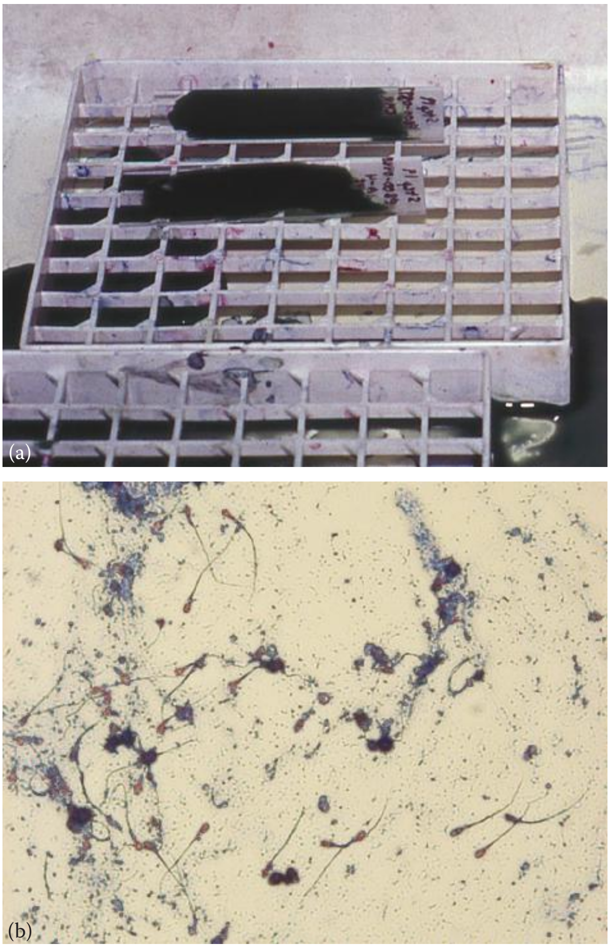

Figure 14.13 Human spermatozoa stained with Christmas tree stain. (a) Staining spermatozoa on a microscope slide and (b) stained spermatozoa. (© Richard C. Li.)

Figure 14.14 A microscopic device for fluorescent detection utilizing SPERM HY-LITER Fluorescent Staining Kit (left) and a laser-capture microdissection device for separating sperm cells from other types of cells (right). (© Richard C. Li.)

PSA® (Abacus Diagnostics, California) are available. These devices utilize antihuman PSA antibodies. In the ABAcard PSA® assay, a labeled monoclonal antihuman PSA antibody is contained in a sample well, a polyclonal antihuman PSA antibody is immobilized on a test zone of a nitrocellulose membrane, and an antiglobulin that recognizes the antibody is immobilized on a control zone (Figures 14.15 and 14.16).

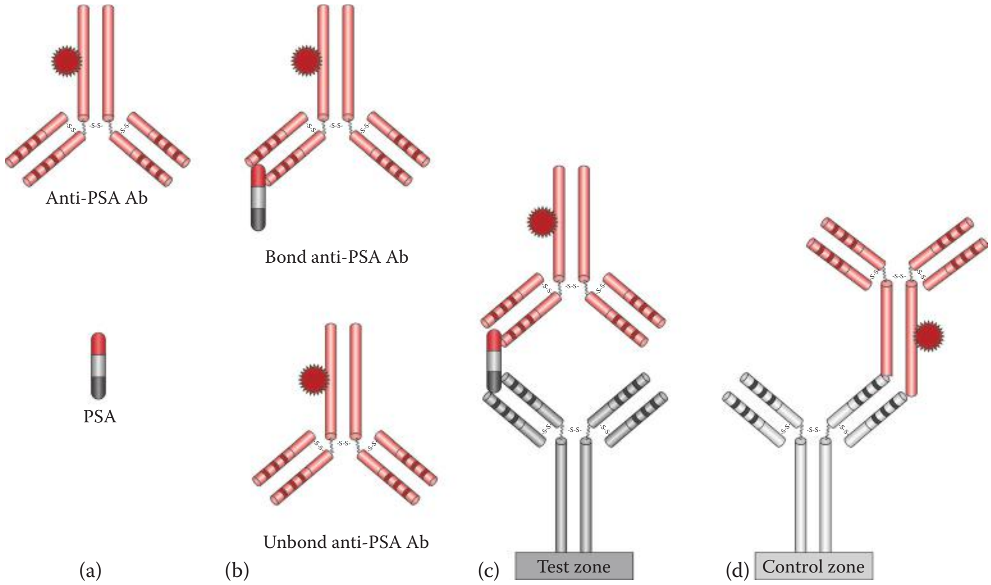

Figure 14.15 Immunochromatographic assays for identification of PSA in semen. (a) In a sample well, PSA in a semen sample is mixed with labeled anti-PSA Ab. (b) The PSA binds to the labeled anti-PSA Ab to form a labeled Ab–PSA complex. (c) At the test zone, the labeled complex binds to an immobilized anti-PSA Ab to form a labeled Ab–PSA–Ab sandwich. (d) At the control zone, the labeled anti-PSA Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab and PSA represent antibody and prostate-specific antigen, respectively. (© Richard C. Li.)

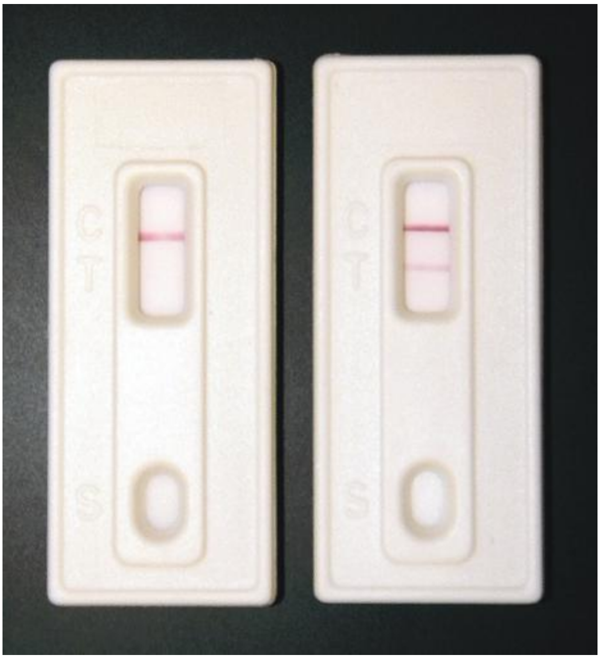

Figure 14.16 Semen identification using an immunochromatic device (ABAcard PSA). The negative (left) and positive (right) results are shown. The "C" band indicates that the test is valid. The "T" band indicates the presence of human blood. The sample well is labeled as "S". (© Richard C. Li.)

The assay is carried out by loading an extracted sample into the sample well. The antigen in the sample binds to the labeled antibody in the sample well to form an antigen–antibody complex. The complex then diffuses across the nitrocellulose membrane. At the test zone, the immobilized antihuman PSA antibody binds with the antigen–antibody complex to form an antibody–antigen–antibody sandwich. The ABAcard PSA® uses a pink dye that allows for the visualization of a positive test with a pink line at the test zone. In the control zone, unbound labeled antihuman PSA antibody binds to the immobilized antiglobulin. This antibody–antiglobulin complex at the control zone also results in a pink line. The test is considered valid only if the line in the control zone is observed.

The presence of human PSA results in a pink line at both the test and control zones. The absence of human PSA produces a pink line in the control zone only. A positive result can appear within 1 min; a negative result is read after 10 min. However, the high-dose hook effect, an artifact that may cause false-negative results (Chapter 10), occurs when high quantities of seminal fluid are tested.

14.2.2.2.2 ELISA

The ELISA method can be used to detect PSA with anti-PSA antibodies. The most common method used in forensic serology is antibody sandwich ELISA, in which an antibody–antigen–antibody sandwich complex is formed (Figure 14.17). The intensity of the signal can be detected and is proportional to the amount of bound antigen. The amount of PSA can also be quantified by comparing a standard with known concentrations. Although this method is specific and highly sensitive, it is time-consuming. Chapter 11 discusses the principle of ELISA in further detail.

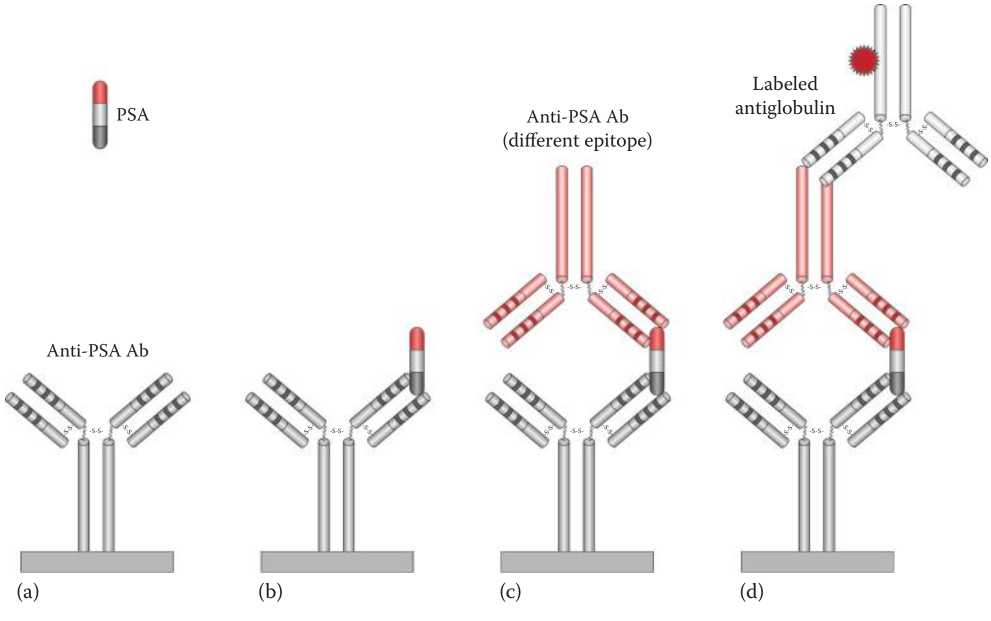

Figure 14.17 Use of ELISA for identification of PSA in semen. (a) Sample containing PSA is applied to polystyrene tubes in which anti-PSA Ab is immobilized. (b) The PSA binds to immobilized Ab to form a PSA–Ab complex. (c) A second anti-PSA Ab, specific for a different epitope of PSA, is added to form an Ab–PSA–Ab sandwich. (d) A labeled antiglobulin then binds to the Ab–PSA–Ab sandwich. The bound antiglobulin can be detected by various reporting schemes. Ab and PSA represent antibody- and prostate-specific antigen, respectively. (© Richard C. Li.)

14.2.2.3 Identification of Seminal Vesicle–Specific Antigen

14.2.2.3.1 Immunochromatographic Assays

Commercially produced immunochromatographic kits include the RSID®-Semen test (Independent Forensics, Hillside, IL) and the Nanotrap Sg. In the RSID®-Semen assay, a labeled monoclonal anti-Sg antibody is contained in a sample well, and a second monoclonal anti-Sg antibody, to a different epitope of Sg, is immobilized on the test zone of the membrane. An antiglobulin that recognizes the antibody is immobilized on a control zone (Figure 14.18).

The sample can be prepared by cutting a small portion of a stain or a swab and is extracted for 1–2 h in an extraction buffer (200–300 μL). Approximately 10% of the extract is removed and mixed with the running buffer. The assay is carried out by loading an extracted sample into the sample well. The antigen in the sample binds to the labeled anti-Sg antibody in the sample well to form a labeled antibody–antigen complex that then diffuses across the membrane. At the test zone, the solid-phase anti-Sg antibody binds with the labeled complex to form a labeled antibody–antigen–antibody sandwich. The antigen in the sample produces a pink line at the test zone. In the control zone, unbound labeled anti-Sg antibody binds to the solid-phase antiglobulin. This labeled antibody–antiglobulin complex at the control zone also results in a pink line. The presence of Sg generates a pink line at both the test and control zones. The absence of Sg results in a pink line in the control zone only. Results may be read after 10 min.

Validation studies have revealed that the sensitivity of the RSID-Semen kit for detecting seminal fluid can be as low as a 5×10^4-fold dilution. Species specificity studies have shown no cross-reactivity with various animal species including ruminants and small mammals. Bodily fluid specificity studies have also shown that the assay is not responsive to human blood, saliva,

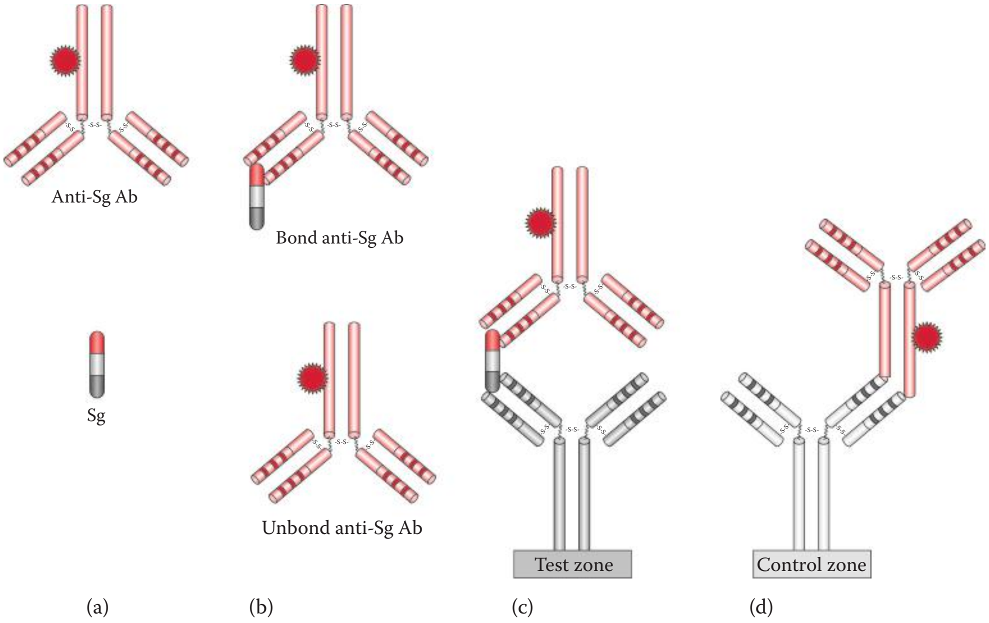

Figure 14.18 Immunochromatographic assays for identification of semenogelin (Sg) in semen. (a) In a sample well, Sg in a semen sample is mixed with labeled anti-Sg Ab. (b) Sg binds to the labeled anti-Sg Ab to form a labeled Ab–Sg complex. (c) At the test zone, the labeled complex binds to an immobilized anti-Sg Ab to form a labeled Ab–Sg–Ab sandwich. (d) At the control zone, the labeled anti-Sg Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab represents antibody. (© Richard C. Li.)

urine, sweat, fecal matter, milk, or vaginal secretions. The assay results are not affected by condom lubricants or spermicides such as nonoxynol-9 and menfegol. However, the high-dose hook effect occurs when more than 3 μL of seminal fluid is tested.

14.2.2.3.2 ELISA

Identification of Sg for semen detection has also been carried out with ELISA. Anti-Sg antibodies are utilized. An antibody–antigen–antibody sandwich complex is formed (Figure 14.19). The intensity of the colorimetric or fluorometric signals can be detected spectrophotometrically and is proportional to the amount of bound antigen. The amount of Sg can be quantified by comparing a standard with known concentrations.

14.2.2.4 RNA-Based Assays

RNA-based assays (Chapter 11) have been developed to identify semen. The assays are based on the expression of certain genes in certain cell or tissue types. Thus, the techniques used in the identification of semen are based on the detection of specific types of mRNA expressed exclusively in spermatozoa and in certain cells of male accessory glands. These assays utilize reverse transcriptase polymerase chain reaction (RT-PCR; see Chapter 7) methods to detect gene expression levels of mRNAs for semen identification. Table 14.1 lists the tissue-specific genes utilized for semen identification. Compared to conventional assays used for semen identification, the RNA-based assay has higher specificity and is amenable to automation. However, one limitation is that the RNA is unstable due to degradation by endogenous ribonucleases.

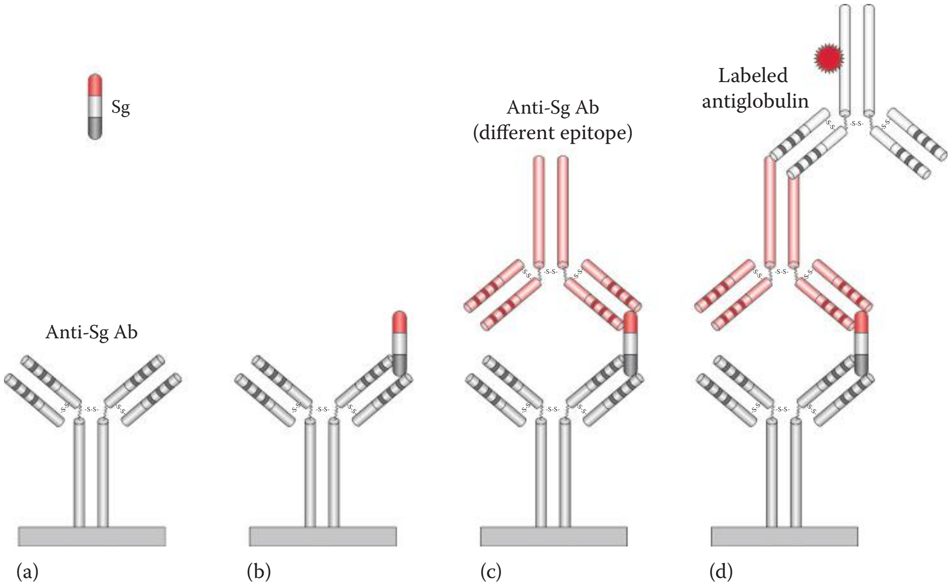

Figure 14.19 Use of ELISA for identification of semenogelin (Sg) in semen. (a) Sample containing Sg is applied to polystyrene tubes in which anti-Sg Ab is immobilized. (b) Sg binds to immobilized Ab to form a Sg–Ab complex. (c) A second anti-Sg Ab, specific for a different epitope of Sg, is added to form an Ab–Sg–Ab sandwich. (d) A labeled antiglobulin then binds to the Ab–Sg–Ab sandwich. The bound antiglobulin can be detected by various reporting schemes. Ab represents antibody. (© Richard C. Li.)

Table 14.1 Application of RT-PCR for Semen Identification

Gene Symbol	Gene Product	Description	Further Reading
KLK3	Kallikrein 3	Also called prostate-specific antigen (PSA)	Gelmini et al. (2001)
PRM1	Protamine 1	DNA-binding proteins involved in condensation of sperm chromatin	Steger et al. (2000)
PRM2	Protamine 2	DNA-binding proteins involved in condensation of sperm chromatin	Steger et al. (2000)

Source: Adapted from Juusola, J. and Ballantyne, J., *Forensic Sci Int*, 152, 1–12, 2005; Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., *Forensic Sci Int*, 157, 181–186, 2006.

Bibliography

Ablett, P.J., The identification of the precise conditions for seminal acid phosphatase (SAP) and vaginal acid phosphatase (VAP) separation by isoelectric focusing patterns. *J Forensic Sci Soc*, 1983, **23**(3): 255–256.

Allard, J.E., The collection of data from findings in cases of sexual assault and the significance of spermatozoa on vaginal, anal and oral swabs. *Sci Justice*, 1997, **37**(2): 99–108.

Allard, J.E., et al., A comparison of methods used in the UK and Ireland for the extraction and detection of semen on swabs and cloth samples. *Sci Justice*, 2007, **47**(4): 160–167.

Allen, S.M., An enzyme linked immunosorbent assay (ELISA) for detection of seminal fluid using a monoclonal antibody to prostatic acid phosphatase. *J Immunoassay*, 1995, **16**(3): 297–308.

Allery, J.P., et al., Cytological detection of spermatozoa: Comparison of three staining methods. *J Forensic Sci*, 2001, **46**(2): 349–351.

Anoruo, B., et al., Isolating cells from non-sperm cellular mixtures using the PALM microlaser micro dissection system. *Forensic Sci Int*, 2007, **173**(2–3): 93–96.

Anslinger, K., et al., Sex-specific fluorescent labelling of cells for laser microdissection and DNA profiling. *Int J Legal Med*, 2007, **121**(1): 54–56.

Astrup, B.S., et al., Detection of spermatozoa following consensual sexual intercourse. *Forensic Sci Int*, 2012, **221**(1–3): 137–141.

Auvdel, M.J., Comparison of laser and high-intensity quartz arc tubes in the detection of body secretions. *J Forensic Sci*, 1988, **33**(4): 929–945.

Baechtel, S., The identification and individualization of semen stains, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 369–374, 1988. Englewood Cliffs, NJ: Prentice Hall.

Bauer, M. and D. Patzelt, Protamine mRNA as molecular marker for spermatozoa in semen stains. *Int J Legal Med*, 2003, **117**(3): 175–179.

Baxter, S.J., Immunological identification of human semen. *Med Sci Law*, 1973, **13**(3): 155–165.

Berti, A., et al., Expression of seminal vesicle–specific antigen in serum of lung tumor patients. *J Forensic Sci*, 2005, **50**(5): 1114–1145.

Bitner, S.E., False positives observed on the Seratec(R) PSA SemiQuant cassette test with condom lubricants. *J Forensic Sci*, 2012, **57**(6): 1545–1548.

Blake, E.T. and G.F. Sensabaugh, Genetic markers in human semen: A review. *J Forensic Sci*, 1976, **21**(4): 785–796.

Chapman, R.L., N.M. Brown, and S.M. Keating, The isolation of spermatozoa from sexual assault swabs using proteinase K. *J Forensic Sci Soc*, 1989, **29**(3): 207–212.

15

Identification of Saliva

15.1 Biological Characteristics of Saliva

The human salivary glands produce 1.0–1.5 L of saliva daily. About 70% of saliva is produced from the submandibular salivary glands, 25% from the parotids, and 5% from the sublingual salivary glands (Figure 15.1). Although a continuous basal level of saliva secretion is maintained, a large amount of saliva is produced during eating. Saliva is largely water containing small quantities of electrolytes, proteins, antibodies, and enzymes. The digestion of carbohydrates in the diet begins in the oral cavity, where amylase in the saliva breaks down carbohydrates such as starch. Thus, detecting amylase indicates the presence of saliva.

15.1.1 Amylases

Amylases are enzymes that cleave polysaccharides such as starches, which are composed of D-glucose units connected by α1→4 linkages. Starches contain two types of glucose polymers: *amylose* and *amylopectin* (Figure 15.2). Amylose consists of long, linear chains of glucose residues connected by α1→4 linkages. Amylopectin is highly branched and consists of linear chains of glucose residues connected by α1→4 linkages with the branch points connected by α1→6 linkages. Both linear amylose and amylopectin can be hydrolyzed by amylase by cleaving the chains at alternate α1→4 linkages. Amylase cleaves off one maltose (two glucose units) at a time. However, the α1→6 linkages at the branch points are not cleaved by the amylase.

Two types of amylases are characterized. β-*Amylases* found in plant and bacterial sources cleave only at the terminal reducing end of a polysaccharide chain. The end of a chain with a free anomeric carbon (not involved in a glycosidic bond) is called the reducing end. Human α-*amylases* cleave at α1→4 linkages randomly along the polysaccharide chain.

Human α-amylases have two major isoenzymes (multiple forms that differ in their amino acid sequences). Human *salivary* α-*amylase* (HSA) is encoded by the *Amy1* locus, synthesized at the salivary glands and secreted into the oral cavity. Human *pancreatic* α-amylase (HPA), encoded by the *Amy2* locus, is synthesized by the pancreas and secreted into the duodenum through the pancreatic duct. The amino acid sequences of the HSA and HPA are highly homologous. Therefore, monoclonal antibodies against HSA also cross-react with HPA. However, HSA is inactivated by acids in the stomach, while most HPA is inactivated in the lower portions of the intestine, and some amylase activity remains in the feces.

Amylase activity is found in various bodily fluids including semen, tears, milk, perspiration, and vaginal secretions. Most amylase present in normal serum consists of HPA and HSA. The amylases are small molecules and can pass through the glomeruli of the kidney (Chapter 17).

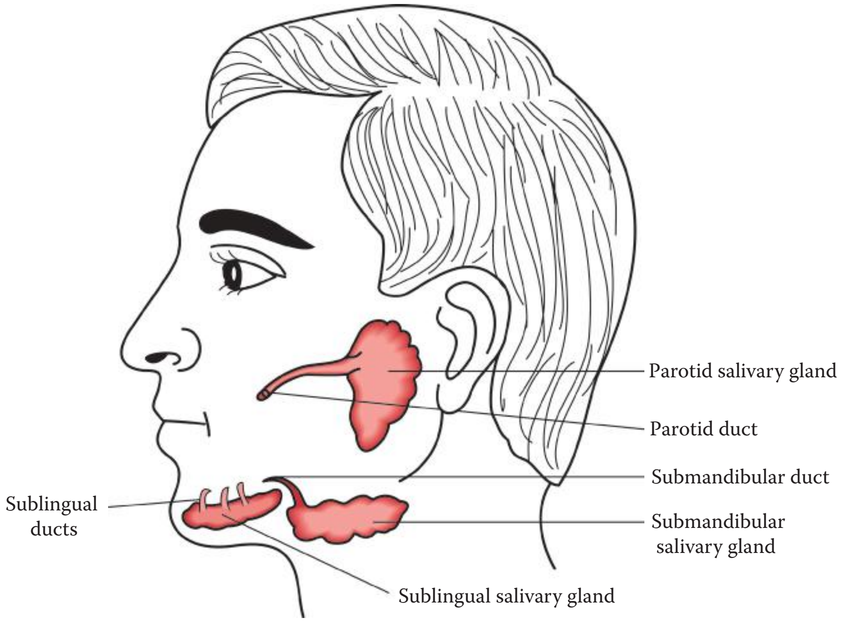
Parotid salivary gland
Parotid duct
Submandibular duct
Submandibular salivary gland
Sublingual ducts
Sublingual salivary gland

Figure 15.1 Human salivary glands. (© Richard C. Li.)

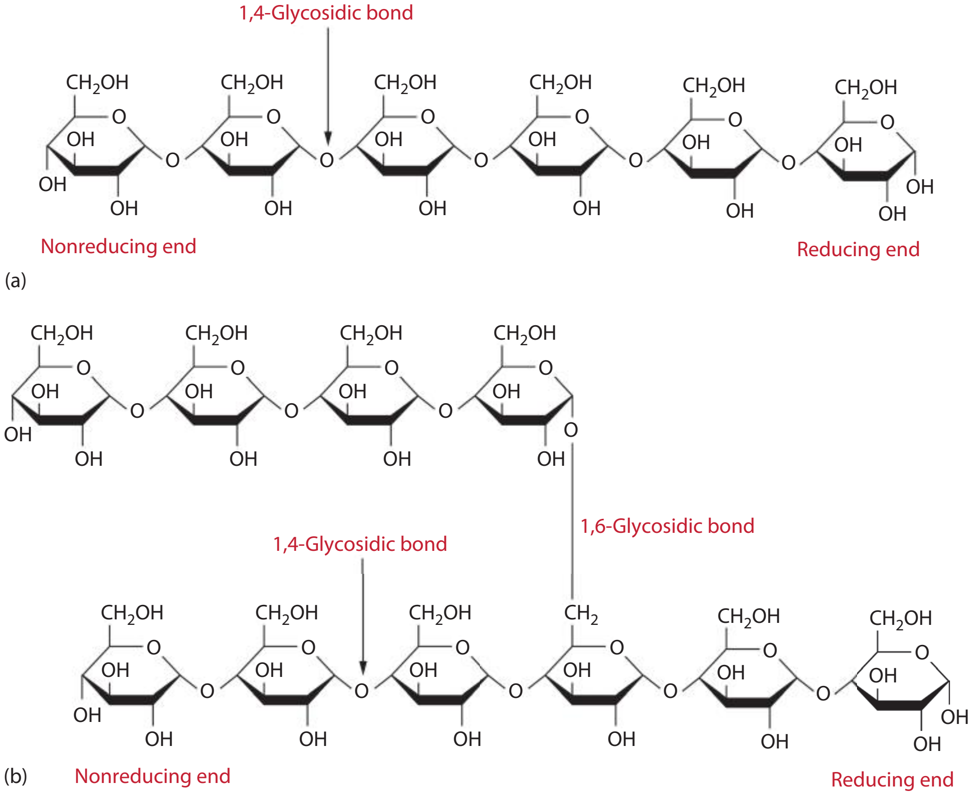
1,4-Glycosidic bond
CH_2OH
O
OH
Nonreducing end
Reducing end
(a)
1,6-Glycosidic bond
1,4-Glycosidic bond
CH_2
(b)

Figure 15.2 Chemical structures of polysaccharides found in starch: (a) amylose and (b) amylopectin.

Thus, the amylase present in urine is derived from plasma. Amylase can be inactivated under boiling temperatures and strong acidic and alkaline conditions. Based on various studies, its stability varies from a few weeks to several months.

15.2 Analytical Techniques for Identification of Saliva

The use of alternate light sources facilitates the search for and the visual examination of saliva stains. The identification of saliva is largely based on detecting the presence of amylase in a sample. Two types of amylase assays can be utilized. The first type measures the enzymatic activity of total amylase. This type of assay cannot distinguish HSA from other amylases including HPA and nonhuman amylases, such as those from plants, animals, and microorganisms. The second type of assay, which includes direct detection of HSA proteins and RNA-based assays, is more confirmatory than enzymatic assays.

15.2.1 Presumptive Assays

15.2.1.1 Visual Examination

The lighting techniques used to search for semen stains can be utilized in searching for saliva stains. For example, a 470 nm excitation wavelength can be used with orange goggles to allow visualization of fluorescence. However, the fluorescence of a saliva stain is usually less intense than that of a seminal stain (Figure 15.3). Microscopic examination with proper histological staining can also be performed to identify the buccal epithelial cells, indicating the presence of a saliva stain (Figure 15.4).

15.2.1.2 Determination of Amylase Activity

15.2.1.2.1 Starch–Iodine Assay

Iodine (I_2) is used to test for the presence of starch. The amylose in starch reacts strongly with iodine to form a dark blue complex, while amylopectin develops a reddish-purple color. In the presence of amylases, starch is broken down to mono- or disaccharides. Consequently, such colors do not develop when iodine is added.

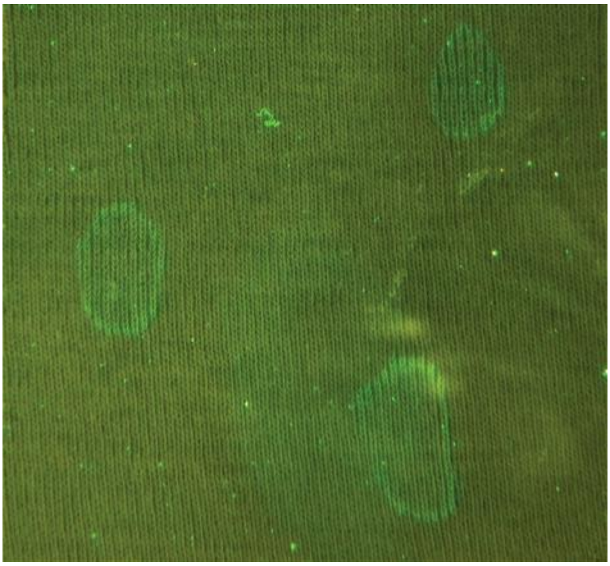

Figure 15.3 Examining saliva stains using an ALS technique. (© Richard C. Li.)

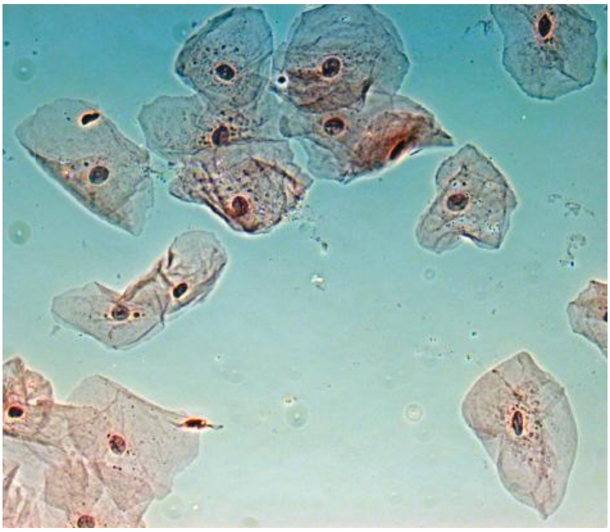

Figure 15.4 Buccal epithelial cells. (© Richard C. Li.)

A common configuration of the method is the radial diffusion assay (Figure 15.5). An agar gel containing starch is prepared. A sample well is created by punching a hole in the gel and an extract of the questioned sample is placed into the well. If amylase is present in the sample, it diffuses from the sample well and hydrolyzes the starch in the gel. The gel is then stained using an iodine solution. A clear area in the gel indicates amylase activity, and the size of the clear area is proportional to the amount of amylase in the sample. A linear standard curve

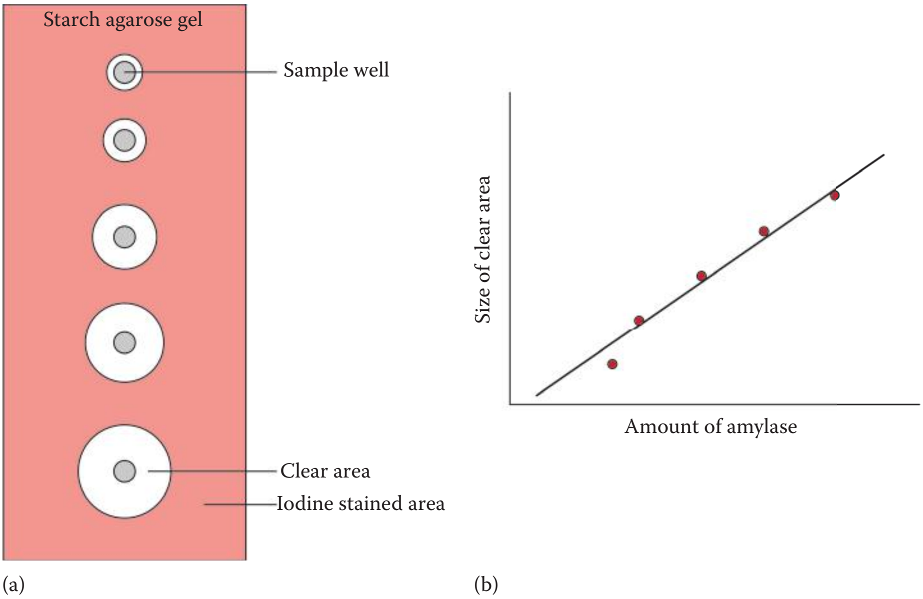

Figure 15.5 Radial diffusion assay for the identification of amylase. (a) Known amounts of amylase standards are applied to the well and allowed to diffuse. (b) The size of the clear area arising from amylase activity is plotted. The standard curve can be used to determine the amount of amylase in a questioned sample. (© Richard C. Li.)

(in log scale) can be prepared using a series of standard amylases with known concentrations. The amount of amylase can be quantified by comparing the results with the standard curve. However, this assay is not specific to HSA and can produce false-positive results. Moreover, the quantities of amylase can be used for determining the amount of sample needed for Y-STR analysis (Chapter 19), which now can be determined by using Y-chromosome-specific quantitative PCR assays (Chapter 6).

15.2.1.2.2 Colorimetric Assays

Dye-labeled amylase substrates such as dye-conjugated amylose or amylopectin are utilized. These substrates are not soluble in water. In the presence of amylase, the dye-containing moieties are cleaved and are soluble in water to produce a color. The degree of coloration, which can be measured colorimetrically by spectrophotometric methods, is proportional to the amount of amylase in the sample. Most of these assays are not HSA specific, although their specificity can be tested by using inhibitors that preferentially inhibit HSA, such as α-amylase inhibitors derived from the seeds of the wheat plant, *Triticum aestivum*. These amylase activity assays are considered presumptive, which means they are not conclusive for the presence of saliva in a sample.

While many substrates are available, Phadebas reagent (Pharmacia) is usually used in forensic laboratories. Produced in a tablet form, it is used to detect α-amylase in specimens for clinical diagnostic purposes. A small portion (approximately 3 mm^2) of a sample is cut and placed in a tube and incubated for 5 min at 37°C. One Phadebas tablet is added to each tube and mixed. Samples are then incubated for 15 min at 37°C, and the reaction is stopped at an alkaline pH by the addition of sodium hydroxide. The amylase substrate is an insoluble blue dye conjugated to starch. Amylase hydrolyzes the substrate to generate a blue color that can be measured at 620 nm using a spectrophotometer. The optical density of the supernatant is read and can be converted to amylase units by comparing to a standard curve.

The amylase assay can also be used for amylase mapping as a method of searching for possible saliva stains (Figures 15.6 and 15.7). These assays are based on the principle that amylase is water soluble and can be transferred from evidence to filter paper and then analyzed via colorimetric assay. This procedure is also referred to as a *press test*. The sensitivity of the method is similar to that of the test tube method.

The substrate can be prepared by evenly spraying the Phadebas reagent on a sheet of filter paper and allowing it to air-dry. The dried substrate-containing paper can be used immediately or stored until needed. To perform amylase mapping, a piece of paper is placed over the entire area to be tested (the item to be tested must be fairly flat to ensure good contact with the paper). The paper is dampened slightly by spraying with distilled water. An outline may be drawn on the paper to aid in locating stains. A piece of plastic wrap is placed on top to prevent the paper from drying during the assay, and a weight is applied to ensure close contact of substrate and evidence. The test is observed every minute for the first 10 min, and every 5 min thereafter up to 40 min, when a positive reaction should appear as a light blue area.

The SALIgAE® kit (Abacus Diagnostics), another commercially available colorimetric assay, has been validated for saliva identification (Figure 15.8). Its manufacturer also produces the SALIgAE spray kit, which can be used for amylase mapping.

15.2.2 Confirmatory Assays

15.2.2.1 Identification of Human Salivary α-Amylase

15.2.2.1.1 Immunochromatographic Assays

Commercially produced immunochromatographic kits include the RSID®-Saliva kit (Independent Forensics). A labeled monoclonal anti-HSA antibody is contained in a sample well. A second monoclonal anti-HSA antibody is immobilized onto a test zone of a membrane, and an antiglobulin that recognizes the antibody is immobilized onto a control zone (Figures 15.9 and 15.10).

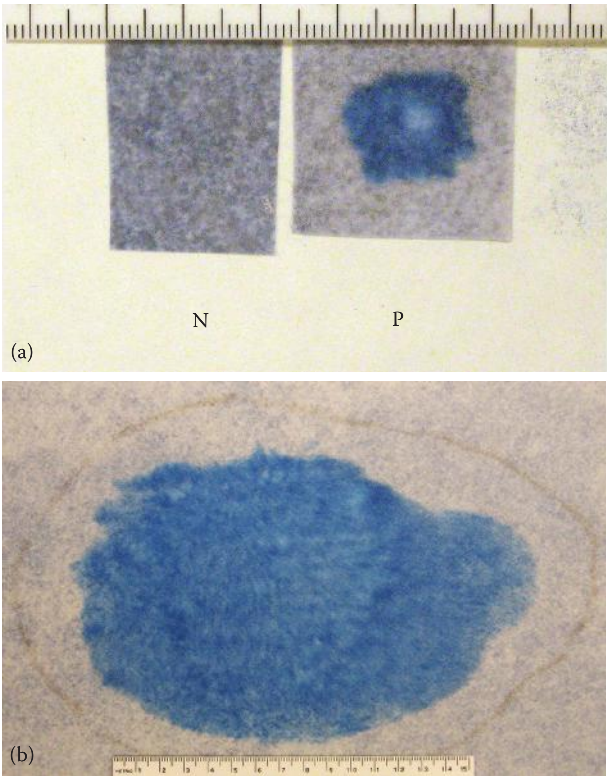

Figure 15.6 Amylase colorimetric assay using Phadebas reagent. (a) A spot test for saliva and (b) amylase mapping result showing a saliva-stained area. N, a negative result; P, a positive result. (© Richard C. Li.)

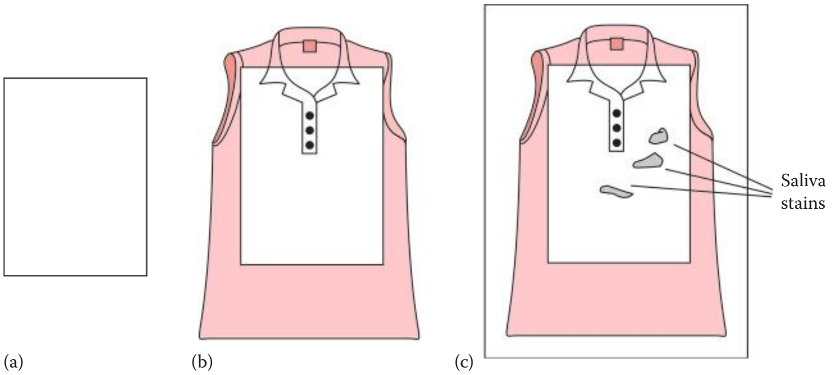

Figure 15.7 Amylase mapping for saliva stains. (a) Amylase substrate is sprayed on a sheet of filter paper. (b) Substrate-containing paper is placed over the area to be tested. The orientation of the filter paper is marked to aid in locating the stain. (c) The filter paper is dampened by spraying it with water, and plastic wrap is placed on top to prevent the paper from drying. Blue color indicates a positive reaction. (© Richard C. Li.)

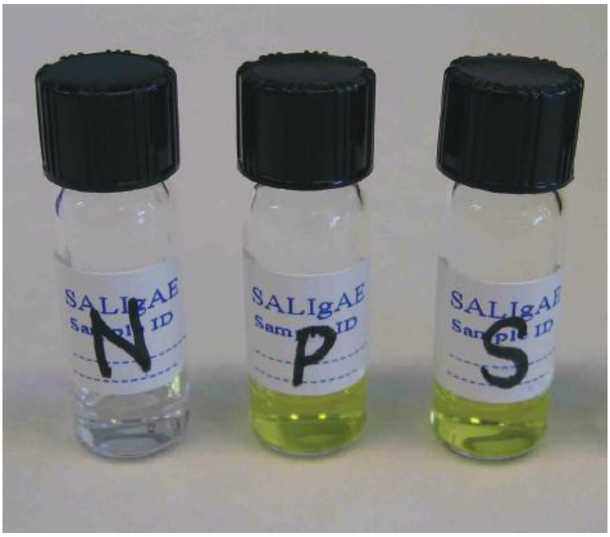

Figure 15.8 Saliva identification using SALIgAE reagent. P, a positive result; N, a negative result; S, a positive control. (© Richard C. Li.)

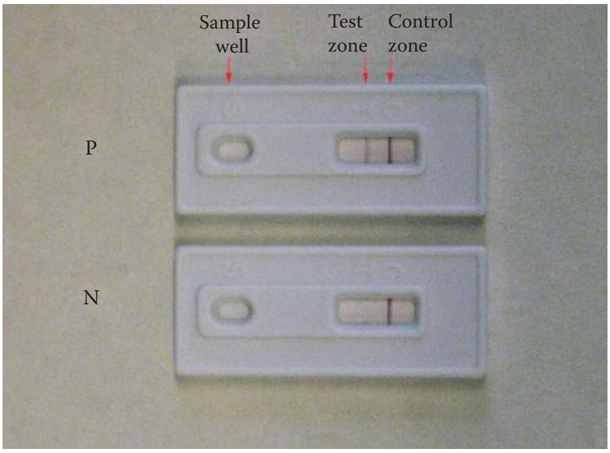

Figure 15.9 Identification of saliva using immunochromatographic assays. P, a positive result; N, a negative result. (© Richard C. Li.)

A sample can be prepared by cutting out a small portion of a stain or a swab. Each sample is extracted for 1–2 h in 200–300 μL of an extraction buffer. Approximately 10% of the extract is removed and mixed with a running buffer. The assay is carried out by loading an extracted sample into the sample well where the antigen in the sample binds to the labeled anti-HSA antibody in the well to form a labeled antibody–antigen complex. The complex then diffuses across the membrane to the test zone, where the solid-phase anti-HSA antibody binds with the labeled complex to form a labeled antibody–antigen–antibody sandwich.

The presence of antigen in the sample results in a pink line at the test zone. In the control zone, unbound labeled anti-HSA antibody binds to the solid-phase antiglobulin. The labeled antibody–antiglobulin complex at the control zone also results in a pink line. The test is

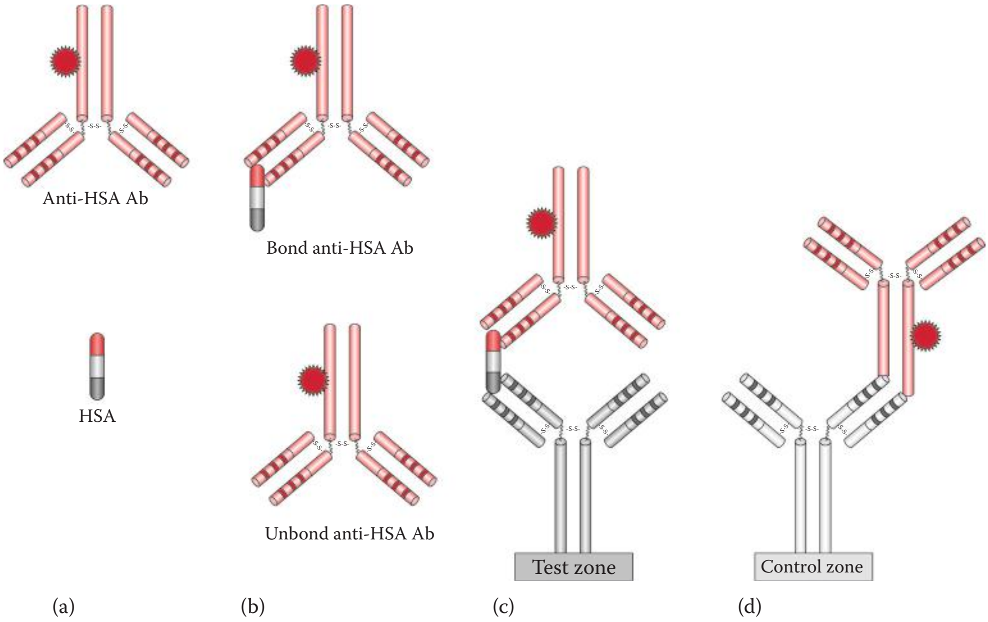

Figure 15.10 Immunochromatographic assay for the identification of HSA in saliva. (a) Sample containing amylases is loaded in a sample well. (b) Antigen binds to a labeled anti-HSA Ab to form a labeled Ab–HSA complex. (c) At the test zone, the labeled complex binds to an immobilized antihuman HSA Ab to form a labeled Ab–HSA–Ab sandwich. (d) At the control zone, the labeled anti-HSA Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab, an antibody; HSA, human salivary α-amylase. (© Richard C. Li.)

considered valid only if the line in the control zone is observed. The presence of HSA results in a pink line at both the test zone and the control zone, while the absence of HSA results in a pink line in the control zone only. A result can be read after 10 min.

The sensitivity of the RSID®-Saliva kit can be as low as 1 μL of saliva. Additionally, the assay is responsive to samples extracted from saliva stains on both smooth and porous surfaces. In terms of species specificity, the kit has no cross-reactivity with various animal species, including monkeys (tamarin and callimico). As for bodily fluid specificity, it has also been shown that it is not responsive to human blood, semen, or urine. The high-dose hook effect, which creates an artifact that may cause false-negative results as described previously (Chapter 11), is not observed when up to 50 μL of saliva is tested. This method is rapid, specific, and sensitive and can be used in both laboratory and field analysis.

15.2.2.1.2 Enzyme-Linked Immunosorbent Assay (ELISA)

This method can be used to detect and to quantify a sample with the use of an anti-HSA antibody. The most common configuration in forensic serology is the antibody–antigen–antibody sandwich (Figure 15.11). ELISA utilizes reporting enzymes to produce colorimetric or fluorometric signals. The intensity of the signal can be detected spectrophotometrically and is proportional to the amount of bound antigen. The amount of HSA can be quantified by comparison with a standard of known concentration. This method is specific and highly sensitive in detecting HSA, but it is time-consuming. Chapter 11 discusses the ELISA principle in further detail.

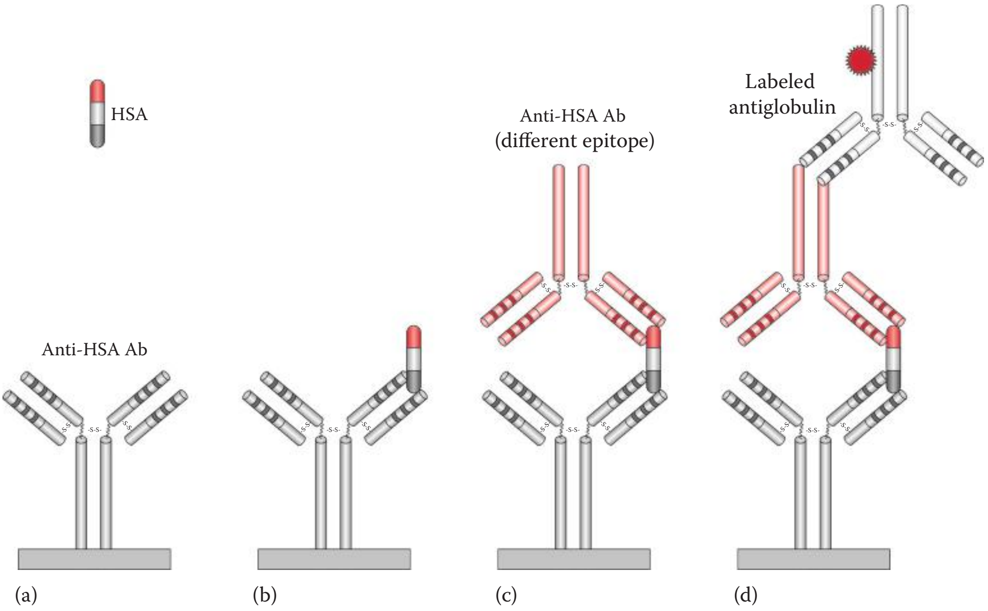

Figure 15.11 ELISA for identification of HSA in saliva. (a) Sample containing antigen is applied to polystyrene tubes where anti-HSA Ab is immobilized. (b) Antigen binds to immobilized Ab to form HSA–Ab complex. (c) Second anti-HSA Ab, specific for a different epitope of HSA, is added to form Ab–HSA–Ab sandwich. (d) Labeled antiglobulin then binds to the sandwich. The bound antiglobulin can be detected by various reporting schemes. Ab, an antibody; HSA, human salivary α-amylase. (© Richard C. Li.)

15.2.2.2 RNA-Based Assays

RNA-based assays (Chapter 11) have been developed recently for the identification of saliva. They are based on the expression of certain genes in certain cell or tissue types. Thus, the techniques used in the identification of saliva are based on the detection of specific types of mRNA expressed exclusively in certain cells in the oral cavity. These assays utilize reverse transcriptase polymerase chain reaction (RT-PCR; see Chapter 7) methods to detect gene expression levels of mRNAs for saliva identification. Table 15.1 summarizes the tissue-specific genes utilized for saliva identification. Compared to conventional assays used for saliva identification, RNA-based assays present higher specificity and are amenable to automation. However, one limitation is that the RNA is unstable because of degradation by endogenous ribonucleases.

Table 15.1 Application of RT-PCR Assay for Saliva Identification

Gene Symbol	Gene Product	Description	Further Reading
HTN3	Histatin 3	Histidine-rich protein involved in nonimmune host defense in oral cavity	Sabatini et al. (1993)
STATH	Statherin	Inhibitor of precipitation of calcium phosphate salts in oral cavity	Sabatini et al. (1990)

Source: Adapted from Juusola, J. and Ballantyne, J., *Forensic Sci Int*, 152, 1–12, 2005.

16

Identification of Vaginal Secretions and Menstrual Blood

The identification of vaginal secretions and menstrual blood is important for the investigation of sexual assault cases. Such identification can help to corroborate allegations of sexual assault. For example, in a sexual assault investigation, a stain was observed after examining the suspect's clothing. Subsequently, forensic DNA analysis revealed that the DNA of the stain originated from the victim, establishing a link between the victim and the suspect. However, the defendant may assert that the victim's DNA originated from a sweat stain as a result of casual contact and deny any criminal act. If vaginal secretions were found in the stain, the evidence has probative value to corroborate an allegation of a sexual act. Additionally, postcoital drainage stains on clothing or bedding are often recovered at crime scenes. These stains usually consist of a mixture of semen and vaginal secretions. In such cases, the presence of vaginal secretions in these stains indicates the occurrence of sexual intercourse. Sometimes, vaginal secretions can be transferred onto a perpetrator during a sexual assault. For example, the presence of vaginal secretions on a suspect's genital area can indicate the occurrence of sexual intercourse. Furthermore, the presence of vaginal secretions on an object can corroborate an allegation of vaginal rape with a foreign object.

If a sexual assault victim is in menses when an assault occurred, blood evidence, such as the victim's bloodstains located on the suspect's clothing, may be recovered at the scene. The defense may argue that the bloodstains resulted from an injury and deny that any sexual act occurred. In this case, the identification of menstrual blood would corroborate an alleged rape. Therefore, it is necessary to distinguish between peripheral and menstrual blood in investigations of sexual assault where blood evidence is found at the scene.

16.1 Identification of Vaginal Stratified Squamous Epithelial Cells

A normal human vagina is covered by the *squamous mucosa*, which is composed of stratified squamous epithelial tissue (Figure 16.1a). Lying under the squamous mucosa is the *submucosa*, which contains an abundance of connective tissue and capillaries. Below the submucosa is the *muscularis*, which is made up of smooth muscle.

The squamous mucosa consists of multiple layers of cells (Figure 16.1b). At the *basal layer* of the squamous mucosa, basal cells are anchored to the basement membrane that separates the squamous mucosa from the submucosa. The basal cells are small in size with relatively large nuclei and are highly proliferative. As the cells migrate up from the basal layer to the *parabasal layer*, the cells undergo differentiation. At the *intermediate layer*, the cells are flattened and their nuclei are compressed. As the epithelial cells reach the apical layer, the *superficial layer*, the cells are fully differentiated with small and dense nuclei.

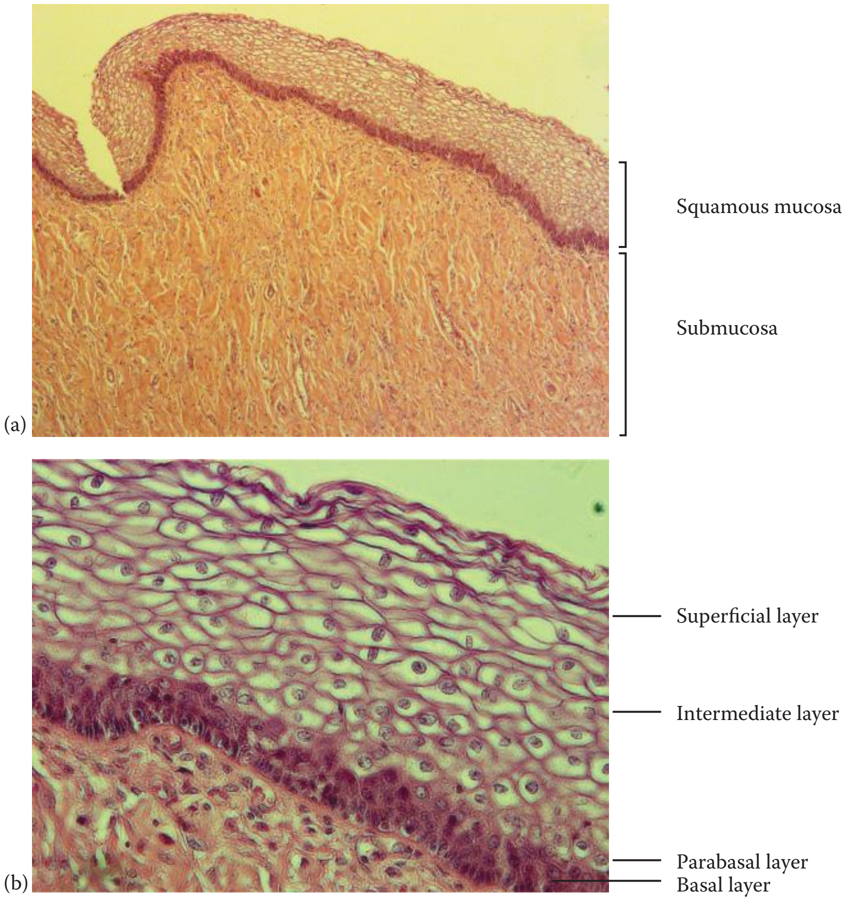

Figure 16.1 Histological structure of human vaginal tissues. (a) The squamous mucosa and the submucosa. (b) Multiple layers of cells in the squamous mucosa. (© Richard C. Li.)

Vaginal epithelial cells do not accumulate keratin, which is different from skin cells (Chapter 4). The intermediate and superficial layer cells contain abundant glycogen in their cytoplasm. The presence of glycogen in these cells is an indication of normal development and the differentiation of the vaginal epithelial cells. The apical surface of the vaginal squamous mucosa is usually covered by mucus that is secreted from glands that are located deep in the epithelium. The cells of the apical layers are eventually sloughed and are continuously replaced by the cells of deeper layers. When the lining of the vagina is swabbed during the collection of evidence, glycogenated cells with small numbers of parabasal cells are usually recovered (Figure 16.2). In addition to the vagina, glycogenated squamous epithelial cells are found in the linings of the oral cavity, pharynx, esophagus, anus, and the apex of the urethra. The differentiation of the vaginal epithelial cells requires estrogen. However, in premenarche and postmenopausal women, estrogen levels are very low; thus, the vaginal epithelial cells only differentiate to the parabasal cells. Primarily, parabasal cells are found in specimens from these individuals.

16.1.1 Lugol's Iodine Staining and Periodic Acid–Schiff Method

Lugol's iodine solution, named after the French physician Jean Lugol, is originally used as an antiseptic that is applied to skin or tissue to prevent infection. In forensic applications, it is utilized for the identification of glycogenated vaginal epithelial cells. The technique is based on the

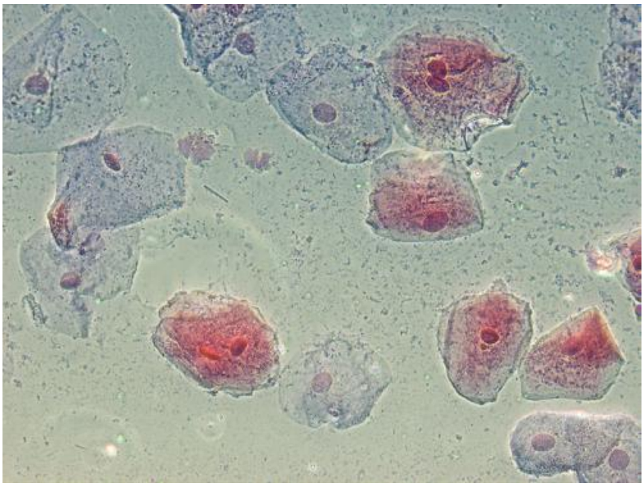

Figure 16.2 Vaginal stratified squamous epithelial cells. (© Richard C. Li.)

principle that iodine reacts with intracellular glycogen to exhibit a color. Glycogen is the principal storage form of glucose in animal and human cells and is found in the granules in the cytoplasm of the cells of many tissues. In addition to squamous epithelial cells, glycogen is also found in hepatocytes, which have the highest glycogen content, as well as muscle cells. Glycogen is a polysaccharide composed of D-glucose units. Similar to the amylopectin in plant cells (Chapter 15), glycogen is a branched polysaccharide consisting of linear and branched chains (Figure 16.3). The D-glucose residues of the linear chain are connected by α1→4 glycosidic bonds, while the branching points are connected by α1→6 glycosidic bonds. However, glycogen is more branched than

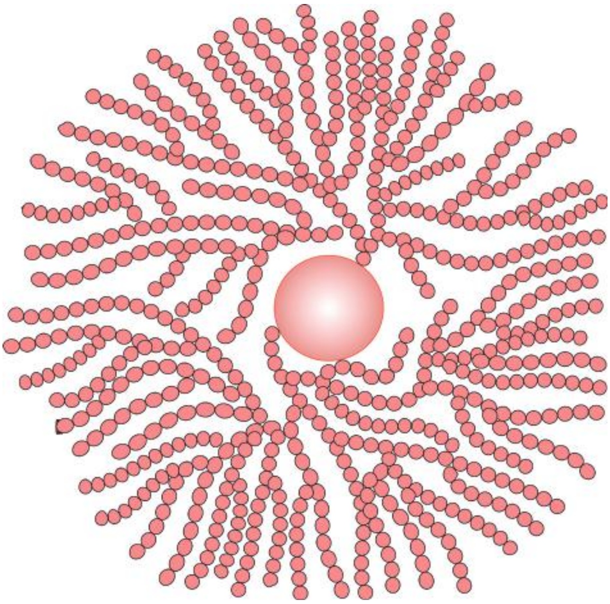

Figure 16.3 Structure of glycogen. Glycogen is a highly branched polysaccharide composed of glucose. Glycogen has a similar structure to amylopectin in starch (see Chapter 15). However, glycogen is more branched than starch. Additionally, glycogen contains a protein known as glycogenin at the center of its structure. (© Richard C. Li.)

amylopectin. Branch points occur approximately every 10 glucose residues in glycogen while they occur every 25 glucose residues in amylopectin. Lugol's stock solution is an aqueous solution of 5% iodine (I_2) and 10% potassium iodide (KI). Potassium iodide allows the iodine to be soluble in water through the formation of the triiodide ion. For staining the vaginal epithelial cells, 5% of the stock solution is usually used as the working solution. Iodine atoms fit into the helices of glycogen to form a dark brown glycogen–iodine complex (Figures 16.4 and 16.5). Vaginal epithelial cells can also be stained using the periodic acid–Schiff method. The cytoplasm of the vaginal epithelial cells is stained magenta, and the nucleus is stained purple.

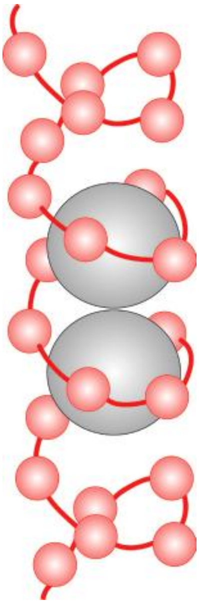

Figure 16.4 Diagram of a glycogen–iodine complex. In the complex, the glycogen chain forms a helix structure with six monosaccharide residues (red) per turn. Iodine molecules (gray) fit in the helix to form the glycogen–iodine complex, exhibiting a color. (© Richard C. Li.)

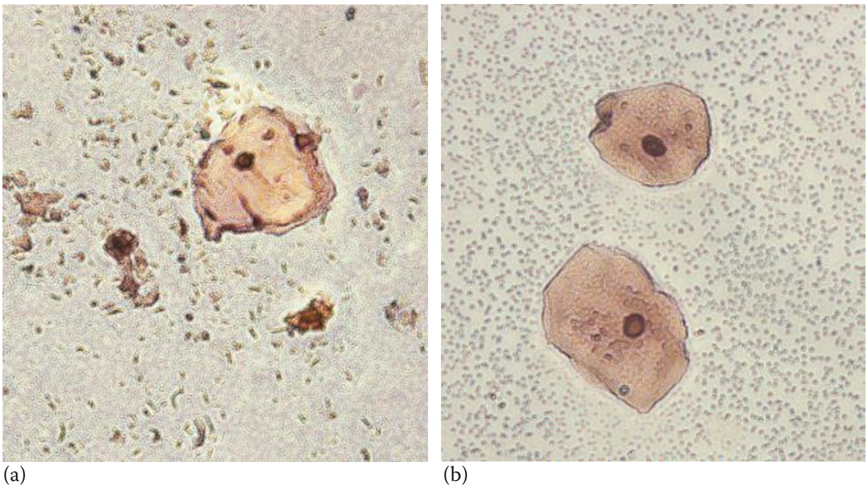

(a) (b)

Figure 16.5 Epithelial cells stained with Lugol's iodine solution. (a) Vaginal and (b) buccal epithelial cells. (© Richard C. Li.)

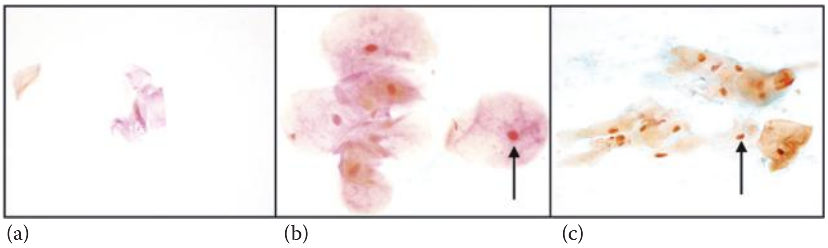

Figure 16.6 Staining of epithelial cells with Dane's method. The results obtained using methanol fixation followed by Dane's staining on skin (a), buccal (b), and vaginal epithelial (c) cells are shown. The arrows indicate red nuclei in buccal and orange nuclei in vaginal cells. (From French, C.E., et al., *Forensic Sci Int*, 178, 1–6, 2008. With permission.)

16.1.2 Dane's Staining Method

Evidence containing skin, buccal, and vaginal epithelial cells is often recovered in forensic investigations, particularly in sexual assault cases. Thus, it is necessary to distinguish between these cells. Skin, buccal, and vaginal epithelial cells belong to stratified squamous epithelium. Differentiated skin epithelial cells are keratinized and are classified as keratinizing squamous epithelial cells, while buccal and vaginal epithelial cells are nonkeratinizing squamous epithelial cells. Additionally, the skin epithelial cells lose nuclei and other cellular organelles during differentiation. In contrast, buccal and vaginal cells contain nuclei. Based on their morphology, it is possible to distinguish skin epithelial cells from buccal and vaginal epithelial cells. However, buccal and vaginal cells are morphologically indistinguishable from each other. Although Lugol's iodine solution and periodic acid–Schiff can stain vaginal epithelial cells, these stains cannot distinguish vaginal from other glycogenated epithelial cells in the oral mucosa and the urinary tract. Recently, Dane's staining method has been developed to distinguish all three types of cells. Skin cells are stained red and orange; buccal cells are stained predominantly orange-pink with red nuclei; and vaginal cells are stained bright orange with orange nuclei (Figure 16.6).

16.2 Identification of Vaginal Acid Phosphatase

Acid phosphatases are a group of enzymes that are capable of hydrolyzing a variety of small organic phosphomonoesters under acidic conditions. To date, at least five different acid phosphatase isoenzymes have been identified in human tissues: erythroid acid phosphatase (encoded by the *ACP1* gene), lysosomal acid phosphatase (encoded by the *ACP2* gene), prostate acid phosphatase (encoded by the *ACPP* gene, also known as *ACP3*), macrophage acid phosphatase (encoded by the *ACP5* gene), and testicular acid phosphatase (encoded by the *ACPT* gene). Human prostatic acid phosphatase is found in large quantities in seminal fluid and is used as a biomarker for semen identification (Chapter 14). The prostatic acid phosphatase is a homodimer containing two identical subunits with a molecular weight of 50 kDa. Small amounts of acid phosphatase can be detected in vaginal fluid, which is produced in normal cervical epithelial cells. However, the molecular characteristics of vaginal acid phosphatase are still not known. Historically, vaginal acid phosphatase has been used as a biomarker for the identification of vaginal secretions using acid phosphatase catalytic assays (Chapter 14). In sexual assault investigations, it is important to distinguish vaginal acid phosphatase from prostate acid phosphatase originating from semen exposure. These two enzymes have identical molecular weights, enzymatic specificities, and responses to the same inhibitors. Nevertheless, vaginal and prostate acid phosphatases can be distinguished using agarose electrophoresis. Based on their electrophoretic mobility, bands of vaginal and prostate acid phosphatases can be separated. The prostate acid phosphatase

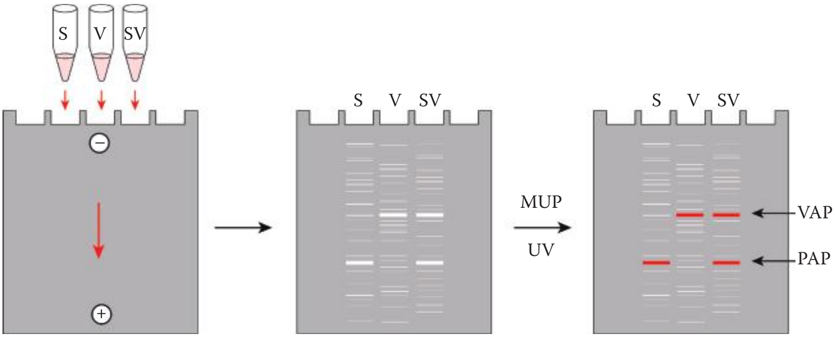

Figure 16.7 Distinguishing prostate and vaginal acid phosphatases using polyacrylamide gel electrophoresis. Bands of acid phosphatases are detected using 4-methylumbelliferyl phosphate (MUP). MUP is a substrate of acid phosphatases, resulting in a product that fluoresces in UV illumination (see Chapter 14). The prostate acid phosphatase has higher electrophoretic mobility toward an anode (a positively charged electrode) than vaginal acid phosphatase. S, semen; V, vaginal secretions; SV, mixture of semen and vaginal secretions; PAP, prostate acid phosphatase; VAP, vaginal acid phosphatase; MUP, 4-methylumbelliferyl phosphate; UV, ultraviolet illumination. (© Richard C. Li.)

has higher electrophoretic mobility toward an anode (a positively charged electrode) than the vaginal acid phosphatase (Figure 16.7). Thus, the presence of vaginal acid phosphatase can be determined.

16.3 Identification of Vaginal Bacteria

Lactobacillus can be found in the respiratory, the gastrointestinal, and the urogenital tract of healthy humans and animals. *Lactobacillus* taxa are the predominant bacteria in the vagina of women of reproductive age, and they play an important role in protecting the host against invasive pathogenic organisms. *Lactobacillus* consists of rod-shaped, nonmotile, and non-spore-forming gram-positive bacteria (Figure 16.8). Since these bacteria survive on carbohydrates, *Lactobacillus* bacteria produce lactic acid. As a result, a low pH environment is established in

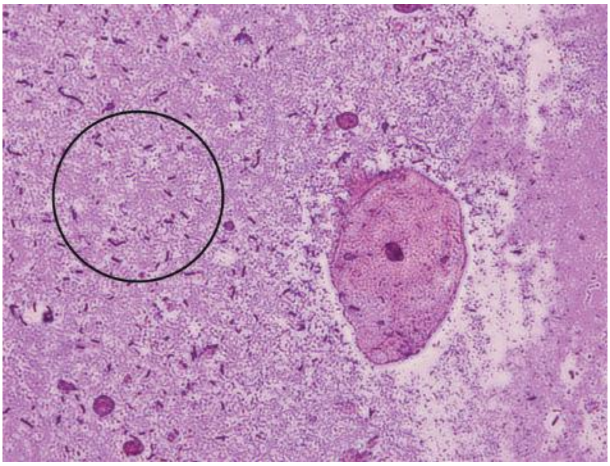

Figure 16.8 Vaginal *Lactobacillus* bacteria (circled). (© Richard C. Li.)

the lumen of the vagina, which restricts the growth of pathogenic organisms. Such vaginal lactic acid–producing bacteria were thought to be *Lactobacillus acidophilus*. In the 1980s, it was determined that *L. acidophilus* was not a single species, but actually a group of related species known as the *L. acidophilus* complex. The species of the complex can now be distinguished based on their DNA sequences. The most frequently occurring *Lactobacillus* species found in the vagina are *L. iners*, *L. crispatus*, *L. gasseri*, and *L. jensenii*. Among them, *L. iners* is the most common species of *Lactobacillus* in women. *Lactobacillus* can be identified based on the sequences of DNA markers such as the 16S rRNA gene and the intergenic spacer region between the 16S rRNA and 23S rRNA genes (Chapter 11). In forensic applications, *L. iners*, *L. crispatus*, *L. gasseri*, and *L. jensenii* can be detected in vaginal secretions. Thus, the identification of *Lactobacillus* taxa can potentially be utilized for the forensic identification of vaginal fluid. However, the presence of *Lactobacillus* taxa is not specific enough to vaginal fluid. Some studies show that *Lactobacillus* taxa are present in semen while others show that *Lactobacillus* can be found in female urine. The openings of the female urogenital system are in close proximity; thus, it is possible to have partially overlapping microbiota between the urine and vaginal secretions. Thus, the identification of multiple bacterial species, at least the four *Lactobacillus* species previously mentioned, should be carried out to distinguish these samples.

16.4 Outlook for Confirmatory Assays of Vaginal Secretions

The identification of the vaginal stratified squamous epithelial cells provides important probative evidence in forensic investigations. However, the existing methods described earlier can sometimes give false-negative or false-positive results. In some situations, these assays also cross-react with other types of bodily fluids. Thus, none of these assays is confirmatory. A useful identification method for vaginal secretions should be able to distinguish vaginal secretions from other bodily fluids and should be easy to perform. For example, nondestructive confirmatory identification methods such as fluorescence spectroscopy and Raman spectroscopy (Chapter 11) can potentially be useful for the identification of vaginal secretions. Recently, the analysis of tissue-specific gene expression has been utilized for the identification of vaginal secretions. Using the reverse transcription polymerase chain reaction (RT-PCR) technique (Chapter 7), the mRNAs of the tissue-specific genes of vaginal epithelial cells can be detected. For example, two commonly used markers for vaginal secretion identification are *MUC4* and *HBD1*. *MUC4* encodes a mucin protein that is a major component of vaginal mucus, and *HBD1*, the human β defensin 1, encodes a vaginal antimicrobial peptide. Both *MUC4* and *HBD1* are expressed in vaginal epithelial cells and are considered reliable markers of vaginal fluid. Additional mRNA and miRNA markers are described in Tables 16.1 and 16.2.

16.5 Menstruation

Menstruation is the periodic discharge of blood and the elimination of the degenerated lining of the endometrium from the uterus of nonpregnant women. From menarche to menopause, women may menstruate up to 400 times during their reproductive age. The uterus plays an important role in preparing the uterine endometrium for the possible implantation of a developing embryo. The linings of the uterus are composed of the myometrium and the endometrium. The *myometrium* consists of the muscle fibers of the uterus. The *endometrium* consists of the simple columnar epithelium and the stroma (Figure 16.9). The simple columnar epithelium is formed by single-layered elongated cells located at the apical surface of the endometrium. The stroma consists of connective tissues as well as spiral arteries. Spiral arteries are small arteries that ascend through the endometrium and form a coil-like structure, which supplies blood to the endometrium.

Table 16.1 Representative Markers of mRNA-Based Assays for Vaginal Secretions and Menstrual Blood Identification

Biological Fluid	Gene Symbol	Description
Vaginal secretions	*CYP2B7P1*	Cytochrome P450, family 2, subfamily B, polypeptide 7, pseudogene 1
	DKK4	Dickkopf homolog 4
	FUT6	Fucosyltransferase 6
	HBD1	β Defensin 1
	IL19	Interleukin 19
	MUC4	Mucin 4
	MYOZ1	Myozenin 1
	SFTA2	Surfactant associated 2
Menstrual blood	*MMP7*	Matrix metalloproteinase 7
	MMP11	Matrix metalloproteinase 11

Source: Adapted from Bauer, M. and Patzelt, D., *J Forensic Sci*, 47, 1278–1282, 2002; Hanson, E.K. and Ballantyne, J., *Sci Justice*, 53, 14–22, 2013; Juusola, J. and Ballantyne, J., *Forensic Sci Int*, 152, 1–12, 2005; Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., *Forensic Sci Int*, 157, 181–186, 2006.

Table 16.2 miRNA Markers for Vaginal Secretions and Menstrual Blood Identification

Bodily Fluid	miRNA Marker	Sequence
Vaginal secretions	miR124*a*	UAAGGCACGCGGUGAAUGCC
	miR372	AAAGUGCUGCGACAUUUGAGCGU
	miR617	AGACUUCCCAUUUGAAGGUGGC
	miR891a	UGCAACGAACCUGAGCCACUGA
Menstrual blood	miR214	UGCCUGUCUACACUUGCUGUGC
	miR412	ACUUCACCUGGUCCACUAGCCGU
	miR451	AAACCGUUACCAUUACUGAGUU

Source: Adapted from Hanson, E.K., Lubenow, H., and Ballantyne, J., *Anal Biochem*, 387, 303–314, 2009; Wang, Z., et al., *Forensic Sci Int Genet*, 7, 116–123, 2013.

16.5.1 Uterine Cycle

The endometrium can be divided into two zones: the functionalis and the basalis (Figure 16.10). The *functionalis* is the luminal part of the endometrium. It is the zone of cyclic changes in the endometrium and is shed during menstruation. The *basalis* is the basal part of the endometrium and is not shed during menstruation. This zone produces cells to regenerate the functionalis during the next cycle.

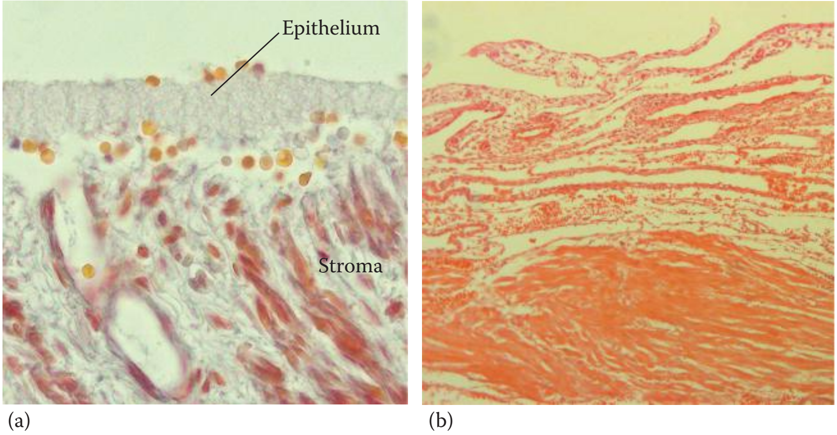
Epithelium
Stroma
(a)
(b)

Figure 16.9 Uterus endometrium. (a) The proliferative phase and (b) the menstrual phase. (© Richard C. Li.)

During the uterine cycle, repetitive physiological changes occur in the functionalis (Figure 16.10). The cycle is divided into three phases: the menstrual, proliferative, and secretory phases. The first day of menstrual bleeding is considered the onset of the menstrual phase. During the menstrual phase, the functionalis degenerates and is sloughed off from the uterine wall and bleeding occurs, known as menses. During the proliferative phase, the functionalis begins regeneration in which the spiral arteries are proliferated. During the secretory phase, the spiral arteries are further developed and coiled. In the absence of pregnancy, a decrease in the progesterone level leads to the constriction of the spiral arteries. As a result, the functionalis becomes ischemic (insufficient blood flow), leading to hypoxia (low levels of oxygen in tissue). In addition, the activation of an enzymatic degradation process causes the destruction and

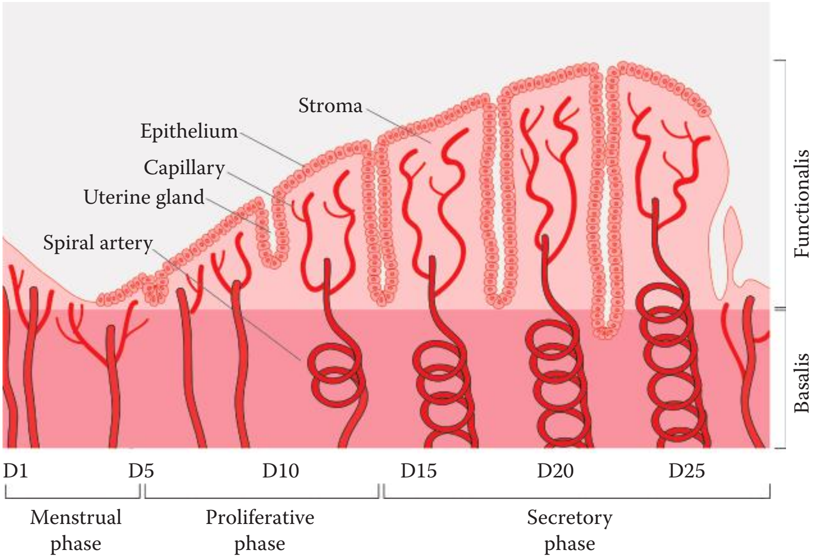
Stroma
Epithelium
Capillary
Uterine gland
Spiral artery
Functionalis
Basalis
D1
D5
D10
D15
D20
D25
Menstrual phase
Proliferative phase
Secretory phase

Figure 16.10 Changes in the functionalis of the uterine mucosa during a uterine cycle. D, day. (© Richard C. Li.)

shedding of the functionalis during menses. Menstrual fluid contains blood, the functionalis layer tissue, and mucus.

16.5.2 Uterine Endometrial Hemostasis

The cessation of menstrual bleeding is achieved by endometrial *hemostasis* that is initiated when injury occurs due to the shedding of the endometrium (Figure 16.11). Hemostasis begins with platelet activation and aggregation to form platelet plugs at the site of injury. Additionally, the blood coagulation cascade is activated to produce *thrombin*. Thrombin, a serine protease, converts soluble fibrinogen into fibrin. Fibrin, a protein involved in blood clotting, aggregates with the platelet plugs and leads to the cessation of bleeding by forming blood clots, known as *thrombi*. Under normal physiological conditions, uterine endometrial hemostasis is a balanced process between blood coagulation and clot dissolution to control blood loss and to prevent clot accumulation within the uterus. As a result, the balance of these two processes allows the removal of tissue fragments from the uterus cavity in order to reduce the risks of infection. Blood clots are prevented from accumulating during menstruation by forming low amounts of platelet plugs and synthesizing coagulation factor inhibitors that inhibit blood coagulation. Additionally, *fibrinolysis* is activated, during which thrombus is broken down by a protease known as *plasmin*. Plasmin cleaves fibrin, generating soluble degradation products. As a result, fibrinolysis can inhibit blood clot formation.

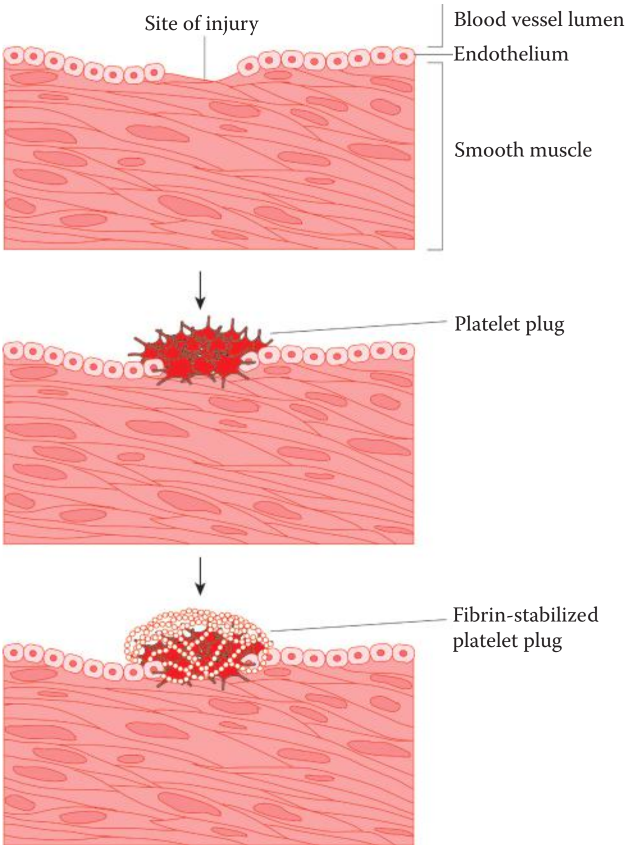

Figure 16.11 Diagram of hemostasis. Hemostasis occurs at the site of injury of a blood vessel. A platelet plug is formed as a result of the aggregation of platelets. The plug is further stabilized by the formation of a fibrin clot over the platelet plug. (© Richard C. Li.)

16.6 D-dimer Assay

During fibrinolysis, cross-linked fibrin is cleaved by plasmin, producing a degradation product known as *D-dimer* (Figure 16.12). Assays for D-dimer fragments have been utilized in the forensic identification of menstrual blood. A number of different formats of D-dimer assays can be used. In an enzyme-linked immunosorbent assay (ELISA; Chapter 11), antibodies bind to the D-dimer antigens on the solid phase. The D-dimer–antibody complex is subsequently analyzed using an antibody-based detection system. This method is highly sensitive, but is time-consuming. Latex agglutination assays are based on the interaction of antibodies and D-dimers that are located on carriers to form aggregates during the agglutination process (Chapter 11). However, the magnitude of the agglutination response is manually read and conclusions are based on subjective judgments. Immunochromatographic assays (Chapter 11) utilize monoclonal antibodies specific to D-dimers, which have been developed recently (Figures 16.13 and 16.14). This immunochromatographic assay is very specific, sensitive, and rapid, and can be completed within minutes. The immunochromatographic devices are portable, and thus can potentially be used at crime scenes. The D-dimer assays can positively identify menstrual blood samples. Although peripheral blood contains low levels of D-dimer, these assays do not show positive reactions with peripheral blood. Thus, menstrual blood can be distinguished from peripheral blood using D-dimer assays. However, postmortem blood also contains these D-dimers, which are detected by these assays. Although the detection of postmortem blood would complicate the interpretation of results, postmortem blood is not often encountered in sexual assault cases.

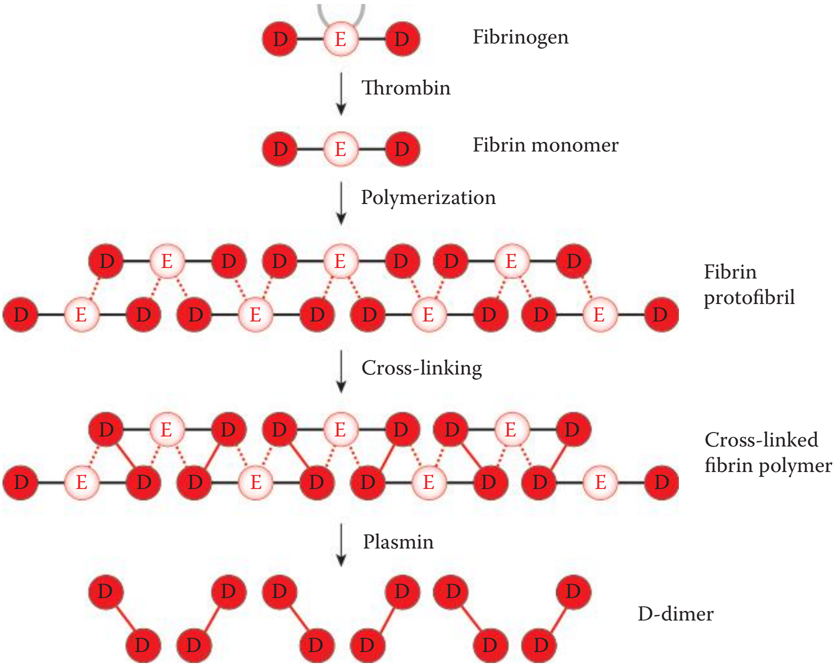

Figure 16.12 The formation and the degradation of a fibrin polymer. The activation of coagulation ultimately generates thrombin, which catalyzes the conversion of fibrinogen to fibrin by cleaving the fibrinopeptides (gray). A fibrin monomer contains an E domain and two D domains. The fibrin monomers are held together by noncovalent bonds (dotted red) between the D domains and E domain to form a fibrin polymer. Fibrin polymers are then covalently linked (solid red) to form a cross-linked fibrin polymer, which plays a role in forming clots. During fibrinolysis, plasmin cleaves the cross-linked fibrin at multiple sites giving rise to fibrin degradation products including D-dimer. (© Richard C. Li.)

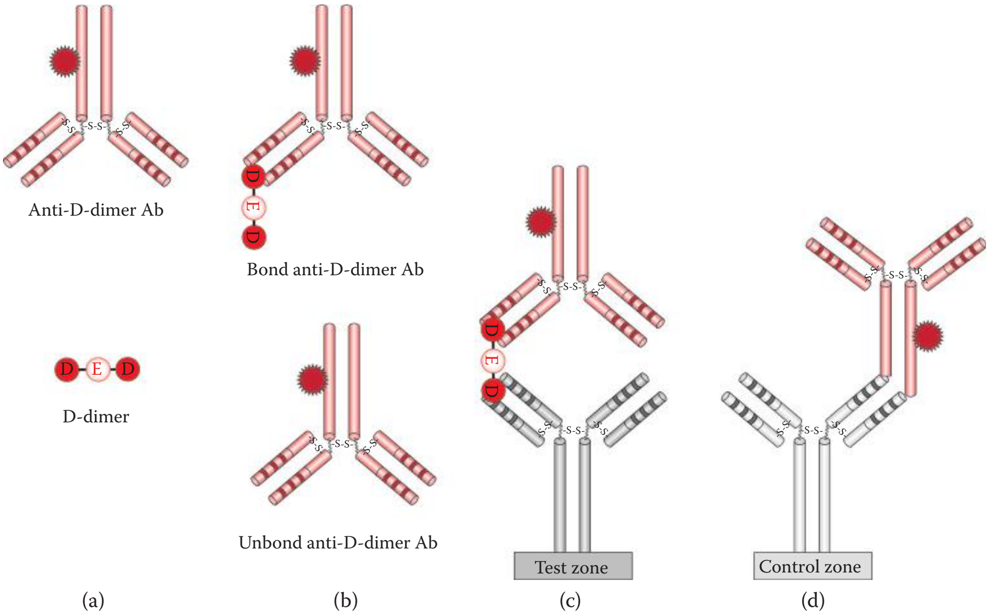
Anti-D-dimer Ab
D
E
D
D-dimer
Bond anti-D-dimer Ab
Unbond anti-D-dimer Ab
S-S
Test zone
Control zone
(a)
(b)
(c)
(d)

Figure 16.13 Immunochromatographic assays for the identification of D-dimer in menstrual blood. (a) In a sample well, D-dimer antigen in a menstrual blood sample is mixed with labeled anti-D-dimer Ab. (b) The D-dimer binds to the labeled anti-D-dimer Ab to form a labeled Ab–antigen complex. (c) At the test zone, the labeled complex binds to an immobilized anti-D-dimer Ab to form a labeled Ab–antigen–Ab sandwich. (d) At the control zone, the labeled anti-D-dimer Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab, antibody. (© Richard C. Li.)

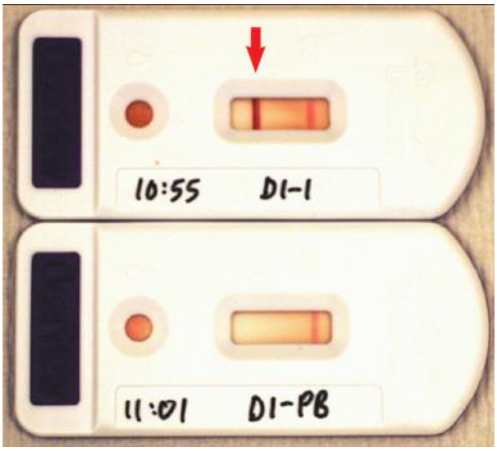
10:55
DI-I
11:01
DI-PB

Figure 16.14 Detecting D-dimer using immunochromatographic assays. The results obtained using the Clearview Simplify D-dimer devices (Alere, Cheshire, UK) on menstrual (top) and peripheral blood (bottom) are shown. A positive result (arrow) indicates the presence of D-dimer. (From Baker, D.J., Grimes, E.A., and Hopwood, A.J., *Forensic Sci Int*, 212, 210–214, 2011. With permission.)

16.7 Lactate Dehydrogenase Assay

Lactate dehydrogenase (LDH) is an enzyme that plays an important role in glycolysis. In humans, LDH catalyzes the reversible reduction of pyruvate into lactate when the amount of oxygen is limited. LDH is a tetrameric enzyme consisting of three different types of subunits. The A subunit, also known as the M subunit, is encoded by the *LDHA* gene and is primarily expressed in skeletal muscle. The B subunit, also known as the H subunit, is encoded by the *LDHB* gene and is primarily expressed in cardiac muscle. The C subunit, encoded by the *LDHC* gene, is expressed restrictively in the testes. LDHs are found in various human tissues. Five isoenzymes can be found in blood (Figure 16.15). LDH1 consists of four identical B subunits; LDH2 consists of one A and three B subunits; LDH3 consists of two A and two B subunits; LDH4 consists of three A and one B subunits; and LDH5 consists of four identical A subunits. The five isoenzymes can be separated using electrophoresis (Figure 16.16) and detected using a colorimetric assay (Figures 16.17 and 16.18). According to their electrophoretic mobility, five bands can be identified. LDH1 has the highest electrophoretic mobility (toward an anode that is a positively charged electrode) and LDH5 has the lowest electrophoretic mobility. In peripheral blood, LDH1, 2, and 3 are the predominant forms of the isoenzymes and LDH4 and 5 are the minor forms of the isoenzymes. In contrast, LDH4 and 5 are consistently the predominant isoenzymes in menstrual blood, while the amounts of LDH1, LDH2, and LDH3 vary. Thus, menstrual blood can be distinguished from peripheral blood. LDH was historically used as a marker for the forensic identification of menstrual blood.

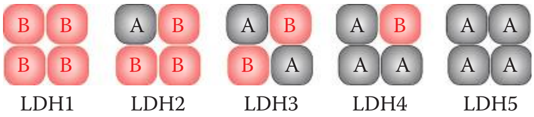

Figure 16.15 Isozymes of lactate dehydrogenase. Human lactate dehydrogenases are composed of four subunits. Five types of isozymes with their subunits are shown. A, LDH A subunit; B, LDH B subunit.

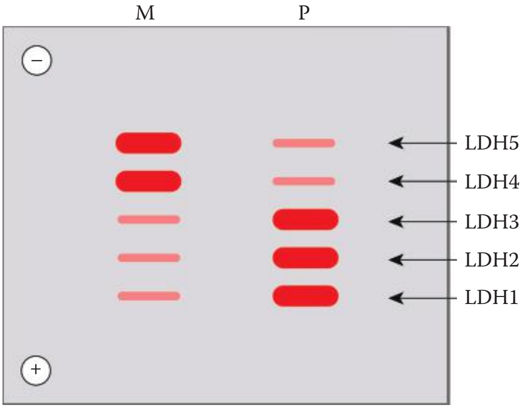

Figure 16.16 Electrophoretic separation of lactate dehydrogenase (LDH) isozymes. Extracted blood samples are loaded onto a cellulose acetate membrane. Electrophoresis is then carried out. Separated bands of LDH isozymes are detected using a colorimetric assay (see Figure 16.15). M, menstrual blood; P, peripheral blood. (© Richard C. Li.)

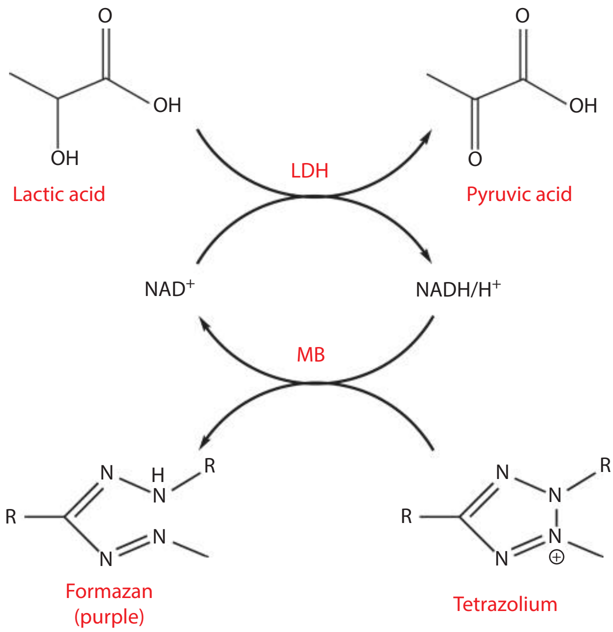

Figure 16.17 Detecting LDH activity using a colorimetric assay forming a colored dye. At the start of a cascade reaction, the LDH catalyzes the conversion of lactic acid to pyruvic acid. Nicotinamide adenine dinucleotide (NAD$^+$) is then reduced to NADH/H$^+$. Subsequently, Meldola's Blue (MB; 8-dimethylamino-2-benzophenoxazine), an electron carrier, transfers H/H$^+$ from NADH/H$^+$ to a tetrazolium salt. The pale yellow tetrazolium salt, nitroblue tetrazolium (NBT), is reduced to form a purple formazan dye.

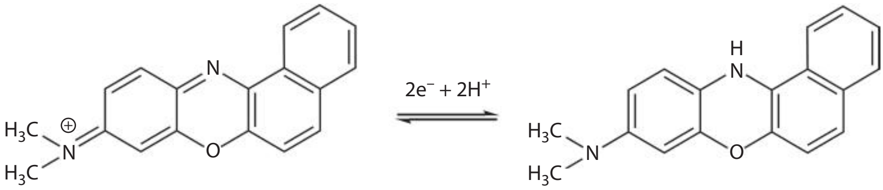

Figure 16.18 The reduction and oxidation of Meldola's Blue. Meldola's Blue is an electron carrier, which transfers hydrogen atoms (coupled to electron transfer).

16.8 RNA-Based Assays

Matrix metalloproteinase (MMP) genes are considered tissue-specific markers for human endometrium tissues. MMPs are zinc-dependent endopeptidases that degrade extracellular matrix components. Additionally, they cleave other proteins such as cytokines, chemokines, and growth factors. The *extracellular matrix* (ECM) is the extracellular space of tissue that is filled by macromolecules such as collagens, laminins, fibronectins, and proteoglycans. The ECM can be divided into two categories: the interstitial matrix and the basement membrane. The interstitial matrices are located in the intercellular spaces. The basement membranes are thin layers of macromolecule fibers that usually lie under the epithelium and the endothelium. Both the interstitial matrices and the basement membranes provide structural support to the cells. It is proposed that MMPs play an important role in the degradation of ECM, leading to the destruction of endometrium tissues during the uterine cycle. To date, a total of 23 MMPs, divided into five subgroups, have been found in humans. The most commonly used markers for the forensic identification of menstrual blood are *MMP7* and *MMP11* (Table 16.1).

Many MMPs are expressed in human endometrium throughout the uterine cycle. The patterns of the MMP gene expressions are correlated with their functions in endometrium tissue breakdown during menstruation. *MMP7* is predominantly expressed in epithelial cells, while *MMP11* is expressed in the stromal cells of the endometrium. Both *MMP7* and *MMP11* mRNA expressions are elevated at the menstrual phase and remain at high levels during the proliferative phase. It is also known that MMPs' mRNA may be elevated in postpartum, wound healing, and metastatic cancer conditions, which may potentially lead to a false-positive identification of menstrual blood.

In menstrual blood samples, among all of the MMP genes tested, *MMP11* is the most sensitive and specific marker for distinguishing menstrual blood from peripheral blood. Using the RT-PCR technique (Chapter 7), *MMP11* mRNA can be detected in menstrual blood from the first to the eighth day of menstruation but it is absent in peripheral blood and vaginal secretions. Therefore, *MMP11* can be used as a marker for the identification of menstrual blood. Likewise, *MMP7* is also a useful marker for menstrual blood identification. However, the *MMP7* mRNA level in menstrual blood is less than that of *MMP11*. Additional markers such as miRNA markers for menstrual blood identification are included in Table 16.2.

Bibliography

Ablett, P.J., The identification of the precise conditions for seminal acid phosphatase (SAP) and vaginal acid phosphatase (VAP) separation by isoelectric focusing patterns. *J Forensic Sci Soc*, 1983, **23**(3): 255–256.

Adams, E.G. and Wraxall, B.G., Phosphatases in body fluids: The differentiation of semen and vaginal secretion. *Forensic Sci*, 1974, **3**: 57–62.

Akutsu, T., et al., Evaluation of latex agglutination tests for fibrin-fibrinogen degradation products in the forensic identification of menstrual blood. *Leg Med (Tokyo)*, 2012, **14**(1): 51–54.

Akutsu, T., et al., Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid. *Leg Med (Tokyo)*, 2012, **14**(3): 160–162.

Albrecht, K., et al., Immunohistochemical distribution of cyclic nucleotide phosphodiesterase (PDE) isoenzymes in the human vagina: A potential forensic value? *J Forensic Leg Med*, 2007, **14**(5): 270–274.

Alsawaf, K. and A.T. Tu, Isotachophoretic analysis of bloodstains: Differentiation of human, menstrual, bovine, and ovine bloods. *J Forensic Sci*, 1985, **30**(3): 922–930.

Baker, D.J., E.A. Grimes, and A.J. Hopwood, D-dimer assays for the identification of menstrual blood. *Forensic Sci Int*, 2011, **212**(1–3): 210–214.

Bauer, M. and D. Patzelt, Evaluation of mRNA markers for the identification of menstrual blood. *J Forensic Sci*, 2002, **47**(6): 1278–1282.

Bauer, M. and D. Patzelt, Identification of menstrual blood by real time RT-PCR: Technical improvements and the practical value of negative test results. *Forensic Sci Int*, 2008, **174**(1): 55–59.

Benschop, C.C., et al., Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? *Int J Legal Med*, 2012, **126**(2): 303–310.

Divall, G.B. and M. Ismail, Lactate dehydrogenase isozymes in vaginal swab extracts: A problem for the identification of menstrual blood. *Forensic Sci Int*, 1983, **21**(2): 139–147.

Ferri, G., et al., Successful identification of two years old menstrual bloodstain by using MMP-11 shorter amplicons. *J Forensic Sci*, 2004, **49**(6): 1387.

Fleming, R.I. and S. Harbison, The use of bacteria for the identification of vaginal secretions. *Forensic Sci Int Genet*, 2010, **4**(5): 311–315.

French, C.E., et al., A novel histological technique for distinguishing between epithelial cells in forensic casework. *Forensic Sci Int*, 2008, **178**(1): 1–6.

Fujita, Y., I. Tokunaga, and S.I. Kubo, Forensic identification of a vaginal fluid and saliva mixture through DNA analysis. *Acta Criminologiae et Medicinae Legalis Japonica*, 2003, **69**(2): 48–52.

17

Identification of Urine, Sweat, Fecal Matter, and Vomitus

17.1 Identification of Urine

The identification of the presence of urine is useful for the forensic investigation of an alleged sexual assault and harassment involving urination. More importantly, the identification of urine stains can aid in the investigation of homicides involving either ligature or manual strangulation. In these incidents, strangulation victims often involuntarily excrete urine prior to death. The locations of urine stains at the crime scene provide useful information to determine the site where the violence has occurred.

17.1.1 Urine Formation

In humans, the urinary excretory system eliminates soluble toxic wastes that are cellular metabolic by-products. The urinary system consists of the kidney, the ureter, the urinary bladder, and the urethra (Figure 17.1). The formation of urine takes place in the kidneys and, in particular, in the nephrons. The *nephron* is the basic structural and functional unit of the kidney (Figures 17.2 and 17.3). Each nephron is composed of a glomerulus, a Bowman's capsule, and a renal tubule. The glomerulus is formed by a network of capillaries and is surrounded by a Bowman's capsule. *Filtration* is the first step in urine formation. As blood flows through the glomeruli, much of its fluid, except cells and large molecules, is filtered through the capillaries into the Bowman's capsule. The glomerular filtrate, the preliminary form of urine, consists of water, salts (largely sodium and potassium ions), glucose, and waste products such as urea. The filtrate is then passed through the renal tubule where *reabsorption* occurs. During the reabsorption process, water, glucose, nutrients, and ions such as sodium are reabsorbed back into the blood. The last process of urine formation is *secretion*. Secretion occurs at the distal and the collecting tubules of the nephron where ions (such as hydrogen and potassium ions), ammonia, and certain metabolites are secreted from the blood into the lumen of the renal tubule to be eliminated in the urine. The urine is drained from the kidneys through the ureters and is stored in the bladder before it is finally excreted through the urethra.

Urine is an aqueous solution consisting largely of water. Urea is the most abundant waste product in urine, resulting from the elimination of ammonia that is produced from the metabolic process of amino acids. The average urea concentration in human urine is approximately 9 g/L. Other major components are creatinine, uric acid, and ions such as phosphate, sulfate,

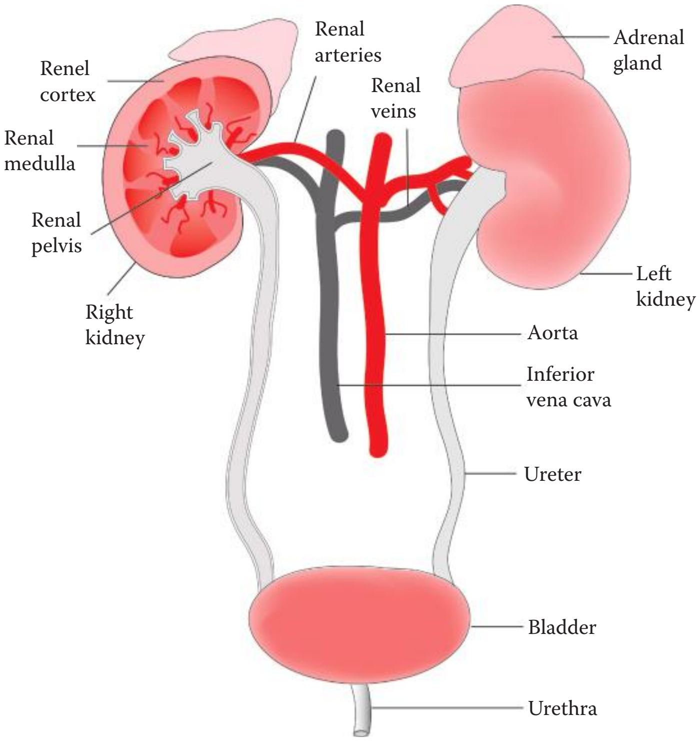

Figure 17.1 Human urinary system. A longitudinal view of the right kidney is shown. (© Richard C. Li.)

chloride, sodium, and potassium. Healthy individuals also excrete a small amount of protein. Urine has a characteristic yellow color. Subsequent to excretion from the body, urine acquires an odor that results from the ammonia that is released from the breakdown of urea by bacteria.

17.1.2 Presumptive Assays

In forensic identification, urine stains can be located by visual examination based on the characteristic yellow color of urine and the detection of the distinctive odor of urine stains. Under alternative light sources, urine stains emit a fluorescent light that facilitates the locating of urine stains in clothing and bedding. Chemical analysis can be carried out to detect the major inorganic anions in urine such as phosphate and sulfate as well as the major organic compounds in urine such as urea, creatinine, and uric acid. These assays are summarized in Table 17.1. However, these assays are not specific to urine. Other bodily fluids, such as sweat, also contain these chemical components.

17.1.2.1 The Identification of Urea

The *para-dimethylaminocinnamaldehyde* (DMAC) assay is simple and rapid and is the most commonly used presumptive assay for the forensic identification of urine stains (Figure 17.4). The DMAC assay can be performed using two different methods: the colorimetric and the fluorometric methods (Figure 17.5).

In the colorimetric method, a portion of a stain (~1 cm^2) is cut and is extracted with 1 mL of distilled water. The extraction is transferred onto a piece of filter paper and is allowed to dry. One drop of 0.1% DMAC solution is then added to the filter paper. DMAC reacts specifically with urea, if present, producing a pink-colored (or magenta-colored) product. DMAC does not react with creatinine, ammonia, or uric acid. The appearance of a pink color within 30 min after applying the DMAC reagent is considered a positive reaction. No color change within 30 min is

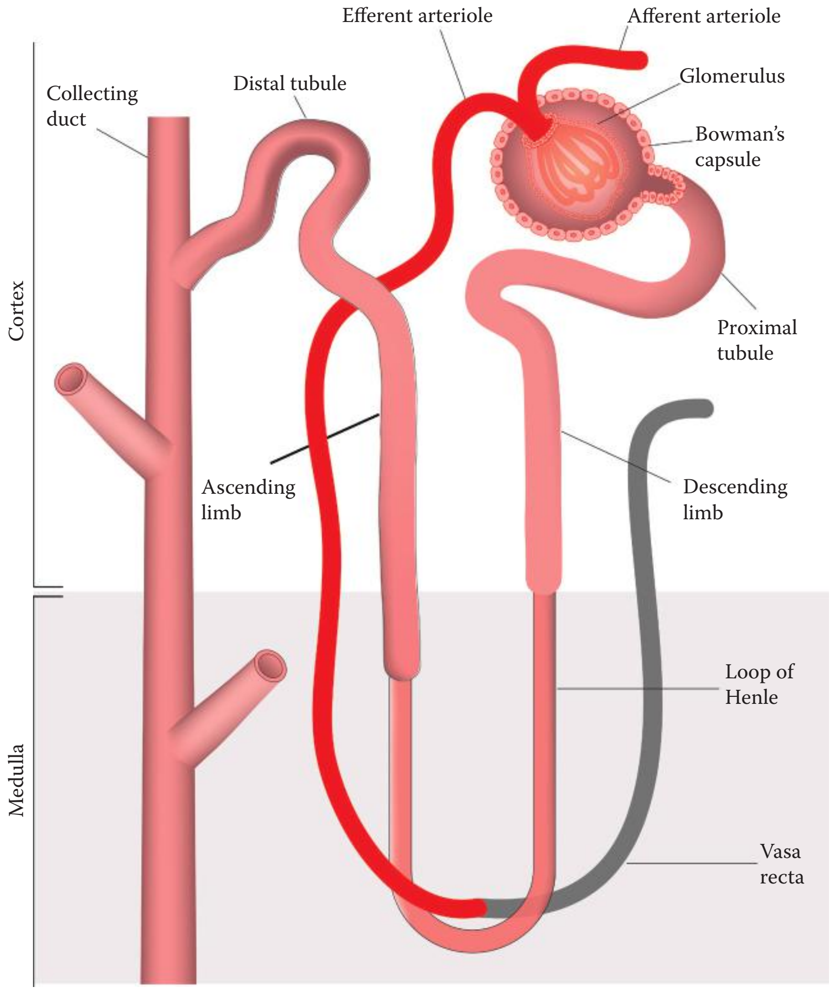
Efferent arteriole
Afferent arteriole
Collecting duct
Distal tubule
Glomerulus
Bowman's capsule
Cortex
Proximal tubule
Ascending limb
Descending limb
Medulla
Loop of Henle
Vasa recta

Figure 17.2 Diagram of the human nephron. (© Richard C. Li.)

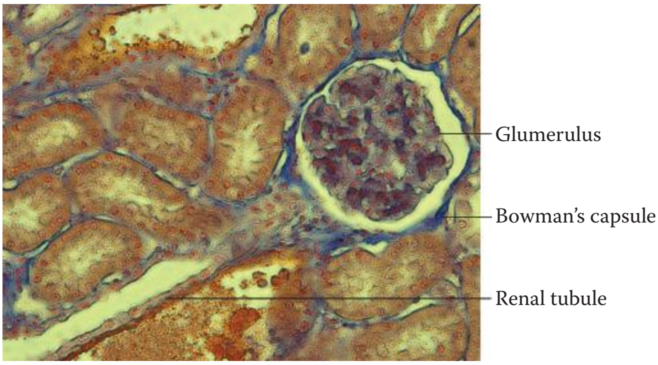
Glumerulus
Bowman's capsule
Renal tubule

Figure 17.3 Section view of the human nephron. (© Richard C. Li.)

Table 17.1 Assays for Identification of Urine and Sweat

Biological Fluid	Assays	Identification	Assay	Method	Further Reading
Urine	Presumptive assays	Locating stains	Visual examination	Alternate light source	Wawryk and Odell (2005)
		Urea	DMAC assay	Colorimetric method	Ong et al. (2012)
				Fluorometric method	Farrugia et al. (2012)
			Microscopic crystal assay	Microscopic examination of urea crystals and urea nitrate crystals converted from urea	Nickolls (1956)
				Microscopic examination of crystals using xanthydrol	Biles and Ziobro (1998)
			Urease assay	The enzyme catalyzes the breakdown of urea and releases ammonia and carbon dioxide. The ammonia is detected using an acid-base indicator, bromthymol blue, which exhibits a blue color	Bedrosian et al. (1984)
				The ammonia is detected using Mn and Ag nitrates, which exhibit a black color	Gaensslen (1983)
			Chromatography	Gas chromatograph–mass spectrometer, thin-layer chromatography	Lum (1991)
		Creatinine	Colorimetric assay	Jaffe's test, using picric acid to convert creatinine to a red creatinine picrate compound	Greenwald (1930)
				Using potassium ferricyanide to produce a blue nitroprusside compound	O'Leary et al. (1992)
			Chromatography	Gas chromatograph–mass spectrometer	Lum (1991)

		Uric acid	Uricase assay	Measuring the disappearance of the absorption of uric acid at 293 nm after treatment with uricase that catalyzes the oxidation of uric acid to 5-hydroxyisourate	Kageyama (1971)
	Confirmatory assays	Tamm–Horsfall glycoprotein	Immunological assays	RIA, crossover electrophoresis, Ouchterlony, ELISA, and immunochromatographic methods	Tsutsumi et al. (1988)
		17-ketosteroids	Mass spectrometer	Liquid chromatograph–mass spectrometer	Nakazono et al. (2002)
Sweat	Presumptive assays	Lactate, lactic acid, urea, and single amino acids	Various	SEM coupled with EDX and Raman microspectroscopy	Sikirzhytski et al. (2011)
	Confirmatory assays	Dermcidin	Immunological assay	ELISA	Sagawa et al. (2003)
			RNA-based assay	RT-PCR	Sakurada et al. (2010)

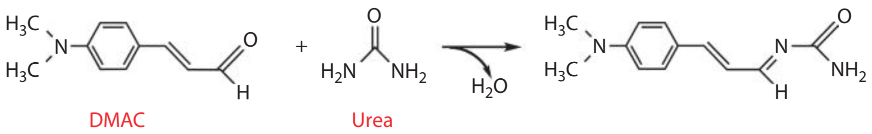

Figure 17.4 Chemical reaction of DMAC assay.

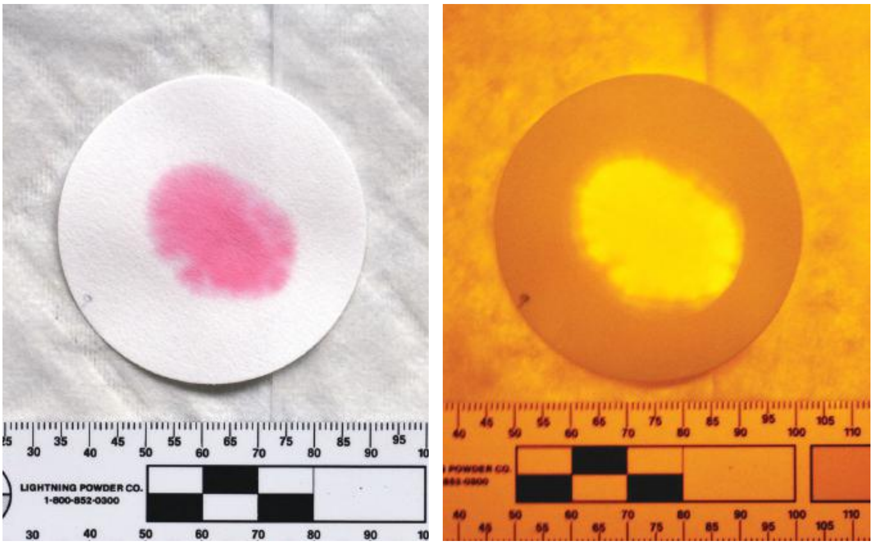

(a)

(b)

Figure 17.5 Positive reaction of DMAC assay of a urine stain. (a) Colorimetric and (b) fluorometric method. (© Richard C. Li.)

considered a negative reaction. However, this method is not specific to urine, as other bodily fluids such as saliva, semen, sweat, and vaginal secretions can also give positive reactions. A diluted DMAC solution, from 0.1% to 0.05%, can maintain appropriate sensitivity to urine stains and minimize false-positive reactions caused by the detection of low levels of urea that are present in other bodily fluids.

The fluorometric method is useful for locating urine stains on large pieces of evidence such as clothing and bedding (Figure 17.6). Additionally, this method can detect patterns of urine stains (Figure 17.7), which can be useful in crime scene reconstructions. In the fluorometric method, the evidence to be examined, such as a garment, is covered by a sheet of filter paper that has been preabsorbed with the DMAC solution. The evidence and the filter are then wrapped together in a sheet of aluminum foil and are left overnight in a press, ensuring that the evidence is in close contact with the filter paper. Alternatively, the layers can be heated for 30 s using an iron. Using a light source at 473–548 nm, the DMAC-treated urine stain fluoresces. The fluoresced stain is best observed with a 549 nm filter. However, colored fabrics interfere with the assay since dyes and pigments can inhibit the fluorescence.

In addition to the DMAC test, the identification of urea that is present in urine stains can be carried out by urease assays. Ureases catalyze the breakdown of urea, thus releasing ammonia and carbon dioxide (Figure 17.8). The ammonia is detected using an acid-base indicator, bromthymol blue, which exhibits a blue color. Alternatively, the ammonia can be detected by manganese and silver nitrates, which exhibit a black color. Additional assays such as microscopic crystal and chromatographic assays are summarized in Table 17.1.

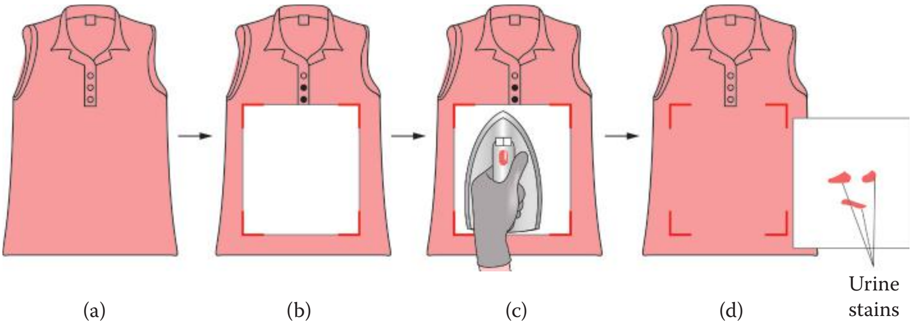

Figure 17.6 Fluorometric DMAC assay of urine stains. Evidence item (a) is closely covered by a piece of moistened filter paper containing DMAC. The orientation of the paper is marked (b). The layers are pressed to allow the transfer of a small amount of the stain onto the paper and are heated using an iron (c). The paper is then lifted. The DMAC-treated urine stain fluoresces under a light source at 473–548 nm (d). (© Richard C. Li.)

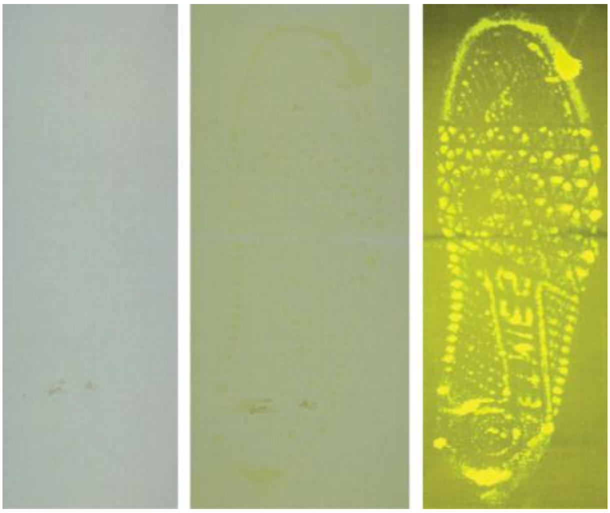

Figure 17.7 Fluorometric DMAC enhancement of a footwear impression in urine on white cotton fabric. From left to right: urine impression, DMAC-treated urine stains under white light, and DMAC-treated urine stains using an excitation light source. (From Farrugia, K.J., et al., *Forensic Sci Int*, 214, 67–81, 2012. With permission.)

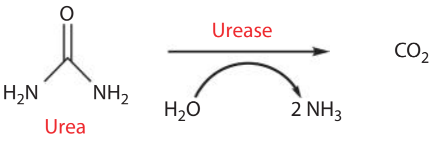

Figure 17.8 Biochemical reaction of urease.

17.1.2.2 Identification of Creatinine

Creatinine is produced during normal muscle cell metabolism (Figure 17.9). During this metabolism, phosphocreatine, an energy-storing molecule in muscle cells, breaks down to form creatine. Creatine is then metabolized to creatinine, which is released from the muscle cells into the blood. Serum creatinine is largely filtrated by the renal glomeruli. A small amount of creatinine is secreted by the renal distal tubules. The amount of creatinine excreted in urine is proportional to the muscle mass of an individual. However, creatinine is not only present in urine. It is also present in other bodily fluids such as blood and semen.

The creatinine present in urine can be detected using the Jaffe color test (Figure 17.10). In this test, picric acid is used to convert creatinine, under alkaline conditions, to form creatinine picrate, which is a bright red product. Additional tests for urine creatinine are summarized in Table 17.1. Recently, a Uritrace device (Abacus Diagnostics) has been made commercially available for the detection of creatinine.

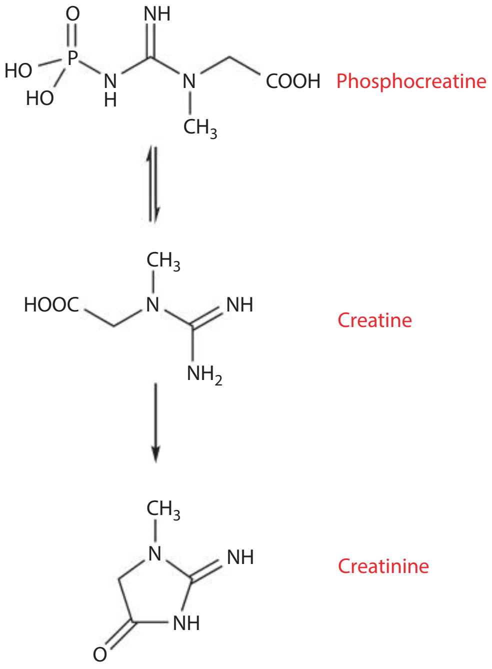

Figure 17.9 Schematic view of the formation of creatinine.

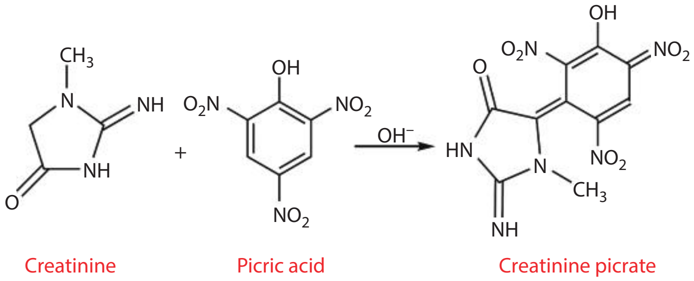

Figure 17.10 Chemical reaction of Jaffe's assay.

17.1.3 Confirmative Assays

17.1.3.1 Identification of Tamm–Horsfall Protein

Tamm–Horsfall protein (THP), also known as *uromodulin*, is the most abundant protein in urine, and accounts for 40% of the urine proteins. THP is exclusively synthesized in the epithelial cells of Henle's loop. THP is secreted from the apical plasma membrane of the epithelial cells into the lumen. Under physiological conditions, an adult excretes 20–100 mg of THP into urine daily. The biological function of THP is not fully understood. It is speculated that it prevents the body from contracting urinary tract infections and from forming renal stones.

THP is a urine-specific biomarker for forensic urine identification. It can be detected using an enzyme-linked immunosorbent assay (ELISA). More recently, it can be detected using an immunochromatographic assay, RSID-Urine, which utilizes a polyclonal rabbit antibody that is specific to THP (Figures 17.11 and 17.12). This test is rapid and simple and thus can be used as a screening test in laboratories and as a field test at crime scenes to identify urine. The detection limit of RSID-Urine for THP is 0.5 μL of urine. Although the sensitivity of RSID-Urine is lower than that of ELISA detection of THP, it is sufficient for forensic applications. These methods are specific to THP that is present only in urine while no detection is found in other bodily fluids such as plasma, saliva, semen, vaginal fluid, or sweat.

17.1.3.2 Identification of 17-Ketosteroids

In humans, urine contains derivatives of 17-ketosteroids such as androsterone, dehydroepiandrosterone (DHEA), and etiocholanolone. *Androsterone* is a steroid hormone with a weak potency of testosterone. *DHEA* is a metabolic intermediate in the biosynthesis of the gonadal steroids. DHEA also has a potential function as a steroid hormone. *Etiocholanolone* is a metabolite of testosterone. These compounds are present in urine as conjugates, in which 17-ketosteroids

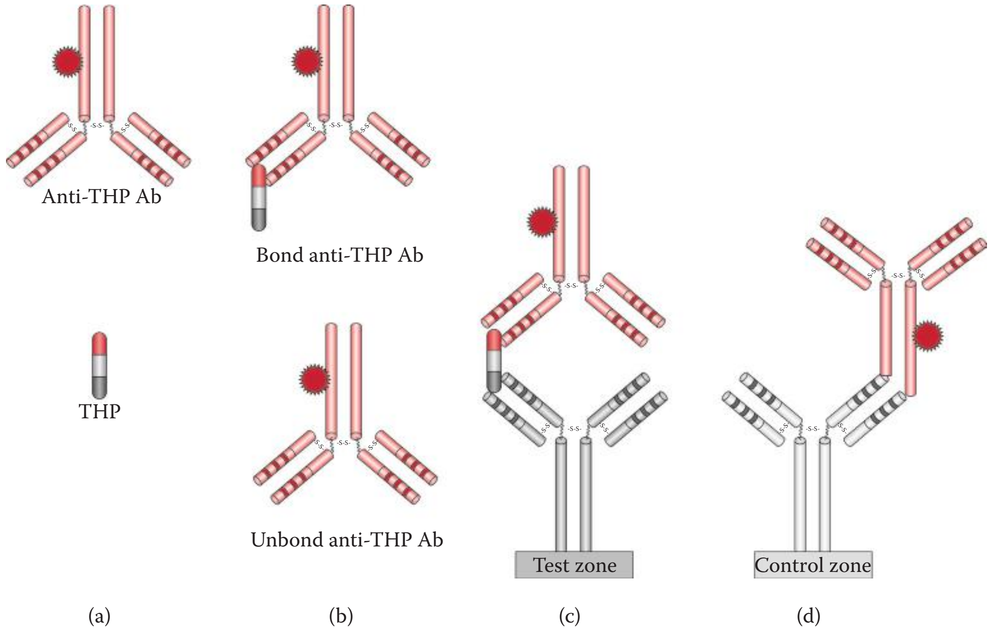

Figure 17.11 Immunochromatographic assays for the identification of THP in urine. (a) In a sample well, THP in a urine sample is mixed with labeled anti-THP Ab. (b) The THP binds to the labeled anti-THP Ab to form a labeled Ab–THP complex. (c) At the test zone, the labeled complex binds to an immobilized anti-THP Ab to form a labeled Ab–THP–Ab sandwich. (d) At the control zone, the labeled anti-THP Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab, antibody; THP, Tamm–Horsfall protein antigen. (© Richard C. Li.)

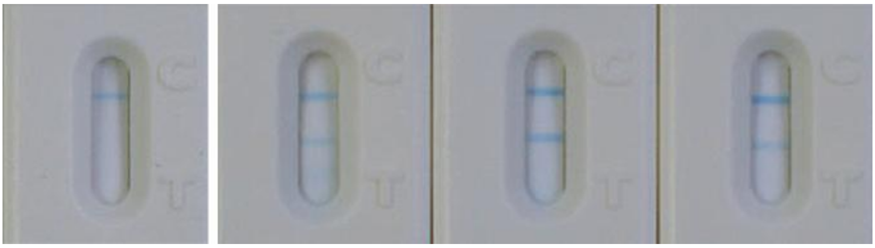

Figure 17.12 Results obtained using immunochromatographic assays (RSID™-Urine; Independent Forensics, Lombard, IL) for the identification of THP in urine. From left to right: negative control, undiluted, 1:10 diluted, and 1:100 diluted urine samples. A visible blue line at the test (T) zone and control (C) zone indicates a positive result. The blue line at the control zone only indicates a negative result. (From Akutsu, T., Watanabe, K., and Sakurada, K., *J Forensic Sci*, 57, 1570–1573, 2012. With permission.)

are modified in the liver through glucuronidation and sulfation. In glucuronidation, glucuronic acids are added to 17-ketosteroids while in sulfation, sulfates are transferred to 17-ketosteroids. As a result, these conjugates, containing charged moieties, are more water soluble than the nonconjugated 17-ketosteroids. Thus, they are excreted into urine to be eliminated from the body. The five major components of the 17-ketosteroid conjugates present in human urine are androsterone glucuronide, androsterone sulfate, DHEA sulfate, etiocholanolone glucuronide, and etiocholanolone sulfate (Figure 17.13). Thus, the analysis of 17-ketosteroid conjugates in urine stains is useful for the identification of human urine stains. These five 17-ketosteroid conjugates can be identified using liquid chromatography–mass spectrometry (LC-MS). However, some of these 17-ketosteroid conjugates are also detected in serum. Therefore, the presence of all five conjugated 17-ketosteroids in a sample is required to identify a urine stain. Additionally, the profiles of the 17-ketosteroid conjugates are human specific and are distinguishable between humans and animals.

17.2 Identification of Sweat

Sweat is the least common bodily fluid analyzed in forensic laboratories compared with others that have been mentioned in previous chapters. However, sweat identification is still useful for forensic investigations. For example, forensic DNA analysis allows the generation of DNA profiles from trace biological evidence such as fingerprints. Identifying sweat can be important for the analysis of these types of evidence.

17.2.1 Biology of Perspiration

Humans have two types of secretory sweat glands: the eccrine and the apocrine sweat glands (Figure 17.14). *Eccrine* sweat glands are distributed almost all over the body and are controlled by the sympathetic nervous system. Eccrine sweat glands play a role in regulating body temperature. When the body temperature rises, eccrine sweat glands secrete a watery sweat to the skin's surface where heat is carried away through the evaporation of the sweat to maintain normal body temperature. In humans, *apocrine* sweat glands, which are associated with hair follicles, are usually restricted to the underarm and genital areas and are controlled by emotional stress. Apocrine sweat glands are inactive until puberty. This type of sweat gland secretes an oily sweat that is odorous after being processed by skin bacteria. The majority of sweat evidence that is analyzed in forensic laboratories is sweat stains secreted from eccrine glands. Sweat contains water, minerals, lactate, and urea. Its biochemical composition varies

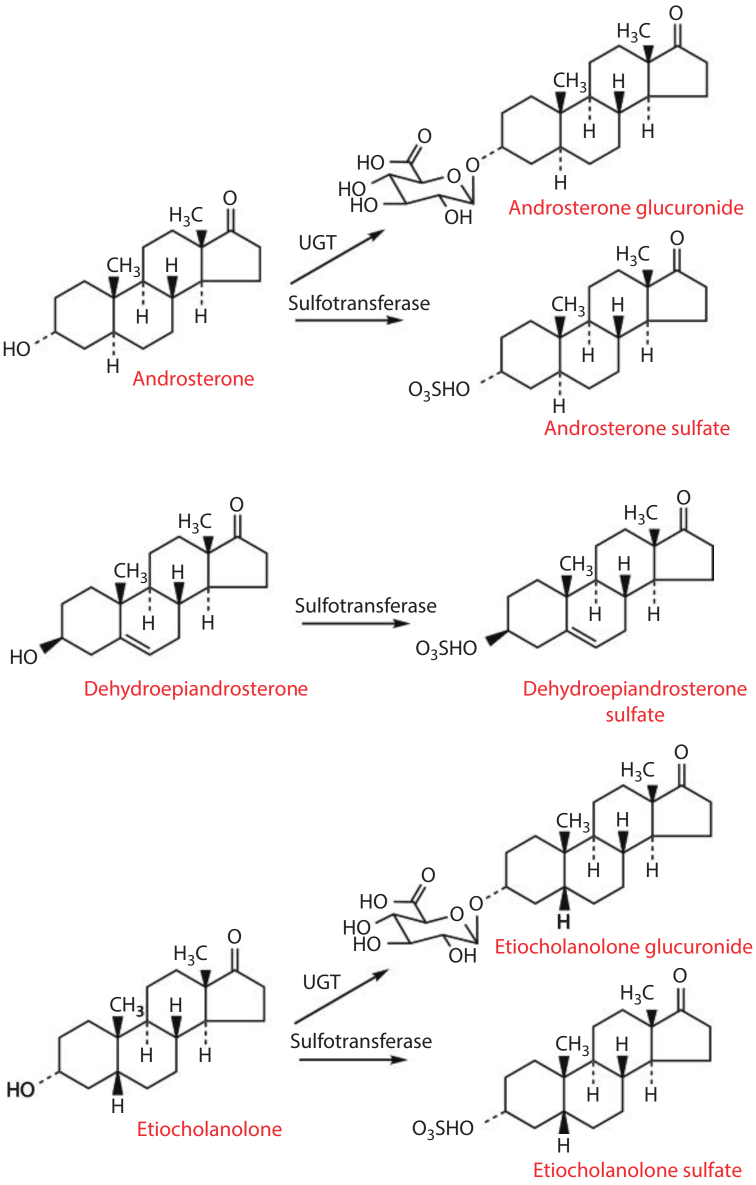
UGT
Sulfotransferase
Androsterone
Androsterone glucuronide
Androsterone sulfate
Dehydroepiandrosterone
Dehydroepiandrosterone sulfate
Etiocholanolone
Etiocholanolone glucuronide
Etiocholanolone sulfate
HO
O
OH
H
H_3C
CH_3
O_3SHO

Figure 17.13 Structural formula of 17-ketosteroids and 17-ketosteroid conjugates present in human urine. UGT, UDP glucuronosyltransferase. (© Richard C. Li.)

among individuals and their physical activities. Sweat contains low levels of constituents that are also present in other bodily fluids such as urine. Thus, it has been considered a difficult bodily fluid to identify.

17.2.2 Sweat Identification Assays

Sweat evidence has been analyzed using presumptive assays (Table 17.1) such as elemental analysis using scanning electron microscopes coupled with energy dispersive x-ray spectroscopy in the detection of lactic acid. Since sweat contains many inorganic and organic compounds that are also present in other bodily fluids, these assays are not specific to sweat. Raman

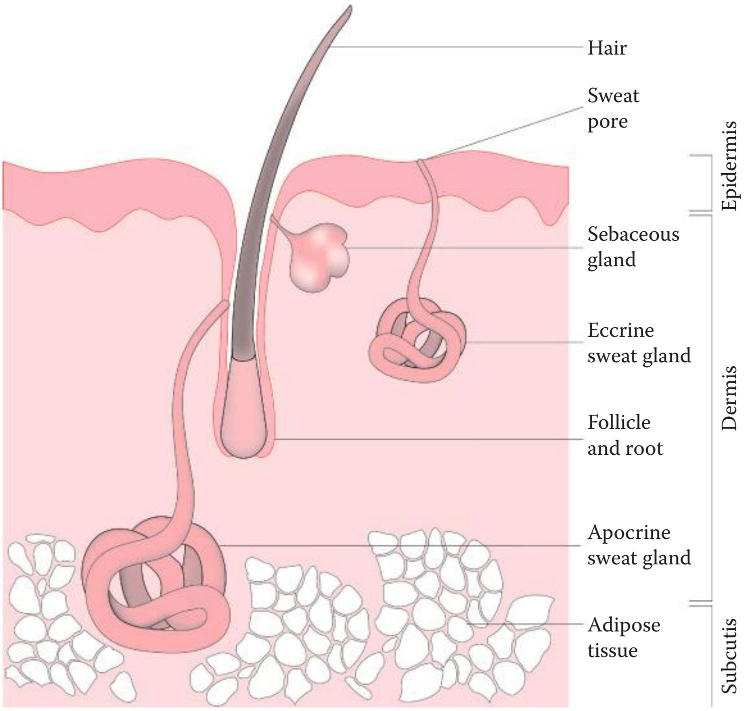

Figure 17.14 Diagram of the human eccrine and apocrine sweat glands. (© Richard C. Li.)

microspectroscopy is potentially useful for the identification of sweat for forensic purposes, which is based largely on the profiles of lactate, lactic acid, urea, and single amino acids in urine.

Recently, *dermcidin* has been identified as a potential biomarker of human sweat. Dermcidin belongs to a class of human antimicrobial peptides of the innate immune defense system and plays an important role in protecting epithelial barriers from infections. Dermcidin, specifically expressed in eccrine sweat glands (Figure 17.15), is secreted into the sweat and is transferred to the epidermal surface. In forensic applications, dermcidin can potentially be utilized as a biomarker for the confirmatory assay of sweat identification. The detection of dermcidin in sweat stains can be performed using ELISA assays utilizing antibodies specific to human dermcidin. This method is highly sensitive and is able to detect dermcidin in sweat samples that are diluted 10,000-fold. Dermcidin is encoded by the *DCD* gene. Its mRNA can be detected using reverse transcription polymerase chain reaction (RT-PCR) assays (Chapter 7) that can detect *DCD* mRNA in 10 μL of sweat sample. Dermcidin assays are also specific to sweat as dermcidin is not detected in other bodily fluids such as semen, saliva, and urine.

17.3 Identification of Fecal Matter

The examination of feces has been used in criminal investigations for over a century. Specifically in 1948, a burglary case was reported using fecal analysis to link the shoes of a suspect to a crime scene. One of the aspects of fecal analysis is to determine a common origin of the reference sample and the fecal evidence, thus potentially linking a suspect to a crime scene. Today, the individual characteristics of a fecal sample can be effectively determined using forensic DNA analysis of sloughed intestinal epithelial cells that are present in fecal matter. The identification of fecal matter is valuable in providing important information for a criminal investigation. For example, the presence of fecal matter may corroborate a sexual

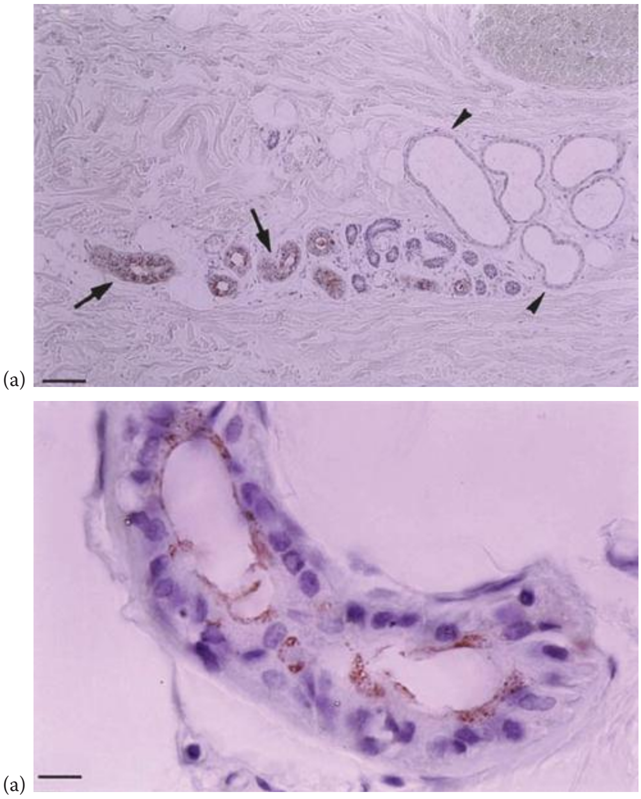

Figure 17.15 Localization of the dermcidin peptide in the sweat glands. Skin tissue sections are treated with a dermcidin antibody and stained. (a) Both the eccrine and apocrine sweat glands are shown. Only cells of the eccrine gland in the skin express dermcidin (brown stained). Arrows: eccrine sweat glands. Arrowheads: apocrine sweat glands. Scale bar: 100 µm. (b) Close-up view of the eccrine gland. The presence of the dermcidin peptide is observed (brown). Scale bar: 10 µm. (From Sagawa, K., et al., *Int J Legal Med*, 117, 90–95, 2003. With permission.)

assault involving sodomy, assault with fecal matter, vandalism, and burglary during which the perpetrator defecated at the scene.

17.3.1 Fecal Formation

Feces are a type of waste matter that is the direct result of food that has been processed by the digestive system (Figure 17.16). Human feces contain undigested foodstuffs, sloughed intestinal epithelial cells, intestinal bacteria, bile pigments, electrolytes, and water. Feces are formed in the intestines during the last phase of digestion. Feces first enter the colon in liquid form. Most of the nutrients are absorbed on the surface area of the small intestine. In the large intestine, water, sodium, and chloride are absorbed on the surface of the lumen. The remaining luminal contents are converted into feces. Food stays for approximately 2–6 h in the stomach. It takes an additional 3–5 h to travel through the small intestine and 12–24 h to travel through the large intestine.

17.3.2 Fecal Matter Identification Assays

A fecal analysis includes macroscopic and microscopic examination, chemical tests, and fecal bacterial identification.

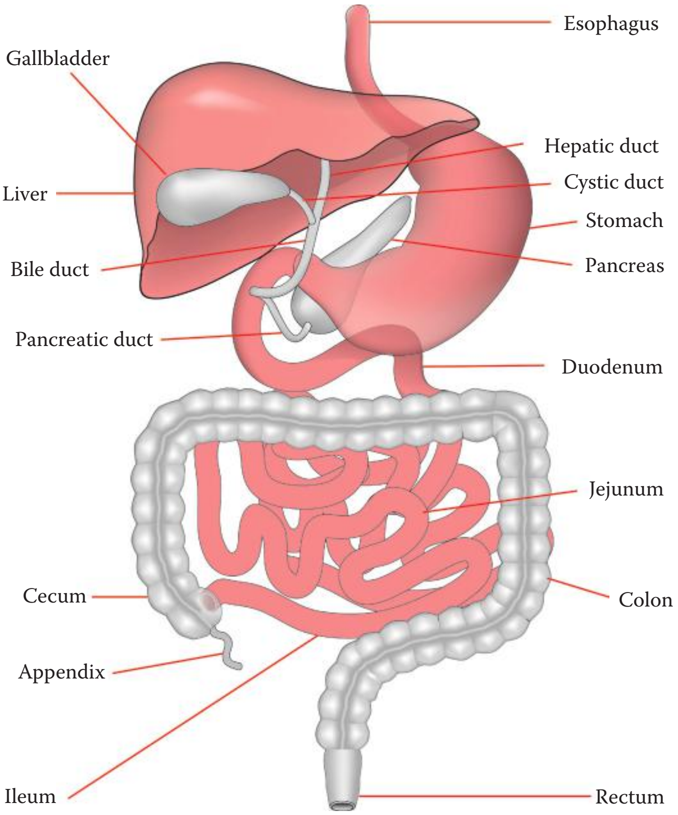

Figure 17.16 Diagram of the human digestive system. The esophagus, stomach, liver, gallbladder, pancreas, small intestine (duodenum, jejunum, and ileum), and large intestine (cecum, colon, appendix, and rectum) are shown. (© Richard C. Li.)

17.3.2.1 Macroscopic and Microscopic Examination

The color and odor of human feces are useful characteristics for fecal identification. The normal brown color of feces primarily results from the presence of urobilinoids, which are heme catabolic by-products. The characteristic odor of feces is caused by the metabolic by-products of the intestinal bacterial flora. Indole, skatole, and hydrogen sulfide are the compounds that are responsible for the odor of feces.

When fecal stains are analyzed, the microscopic examination of the feces can be performed on an aliquot of fecal suspension. The presence of characteristic undigested foodstuffs can indicate human feces. Fecal matter can be transferred from samples of clothing by scraping with a sterile stainless steel spatula. The fecal matter is then hydrated in 6% formalin solution for 24–48 h prior to microscopic examination. Undigested foodstuffs such as vegetable fragments and meat fibers are often present in human feces. Vegetable fragments (Figure 17.17) are often undigested vegetable dermal tissues that cover and protect the plant and fragments of vascular tissues that play roles in transporting water and nutrients throughout the plant. The types of vegetables can be identified by comparing the observed fragments with known plants. Meat fibers (Figure 17.18) are undigested animal skeletal muscle fibers. These fibers have characteristic striations, usually rectangular in shape, that are used for comparison with known animal tissues.

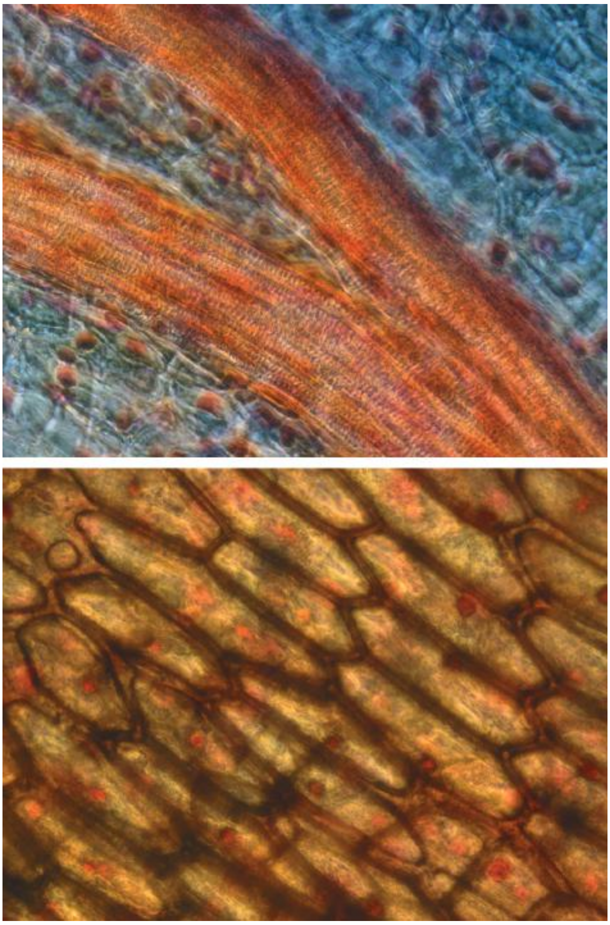

Figure 17.17 Scallion vascular tissue (top) and dermal tissue (bottom). (© Richard C. Li.)

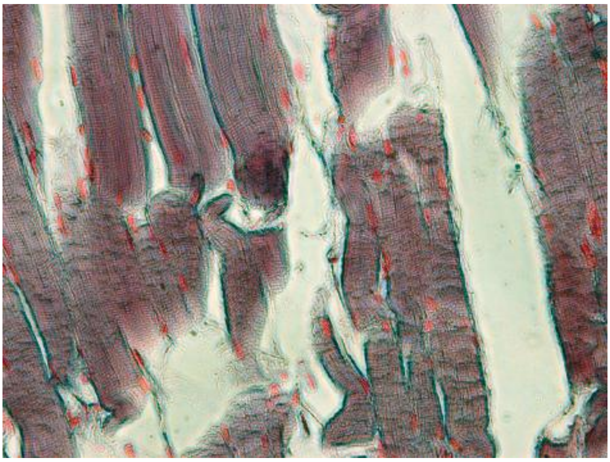

Figure 17.18 Cattle meat fibers. (© Richard C. Li.)

17.3.2.2 Urobilinoids Tests

The forensic analysis of fecal matter often involves the identification of fecal stains on swabs and clothing from which small amounts of sample may be available. Chemical tests can be useful for the analysis of fecal stains. By far the most common chemical test performed on feces is the detection of urobilinoids. *Urobilinoids*, including urobilin and stercobilin, are generated from the degradation of heme and are excreted into feces.

The average lifetime of erythrocytes is approximately 3–4 months. Erythrocytes are continuously undergoing hemolysis in which erythrocytes are naturally broken down and are usually processed in the reticuloendothelial system of the spleen. Hemoglobin is released daily during the hemolysis process and is degraded into heme, globin, and iron. Other sources of heme are derived from the degradation of erythrocyte precursors in bone marrow and other heme-containing proteins such as myoglobin and cytochromes. In the peripheral tissues, heme undergoes catabolism to form *bilirubin* (Figures 17.19 and 17.20). Bilirubin is further converted to *urobilinogen* in the intestine. A portion of urobilinogen is reduced to *stercobilinogen*. In the large intestine, the spontaneous oxidation of urobilinogen and stercobilinogen results in the formation of

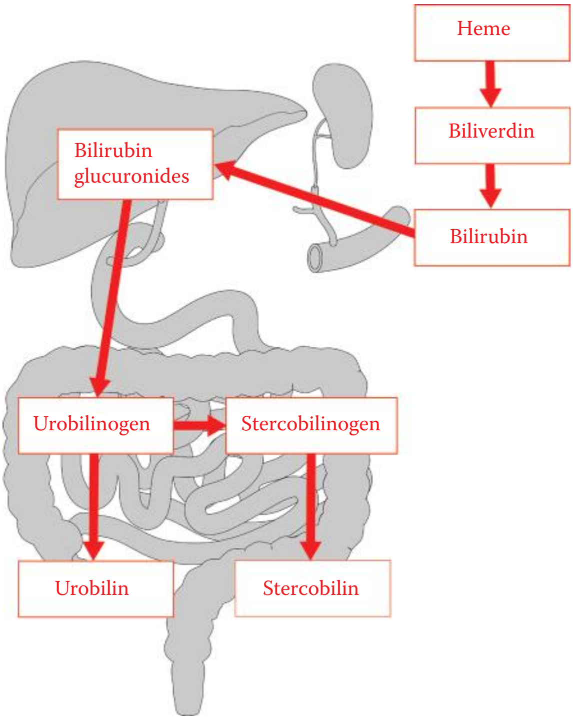

Figure 17.19 The formation of urobilinoids. Aged erythrocytes are disposed in the spleen, releasing hemoglobin that is broken down to heme. The heme is converted to biliverdin and is subsequently reduced to bilirubin. The bilirubin is then released into the bloodstream where it is bound to albumin, which cannot be filtrated at the glomeruli. The bilirubin is transported through the bloodstream to the liver where it is conjugated with glucuronic acid, forming water-soluble bilirubin monoglucuronide and diglucuronide. The bilirubin glucuronides are excreted into the bile and are subsequently excreted into the small intestine. In the intestines, the glucuronic acid of the conjugated bilirubin is removed. The unconjugated bilirubin is metabolized by intestinal bacteria, forming urobilinogen. A portion of the urobilinogen is further metabolized to stercobilinogen. The urobilinogen and the stercobilinogen are oxidized by intestinal bacteria, forming urobilin and stercobilin, respectively, which are excreted into the feces. (© Richard C. Li.)

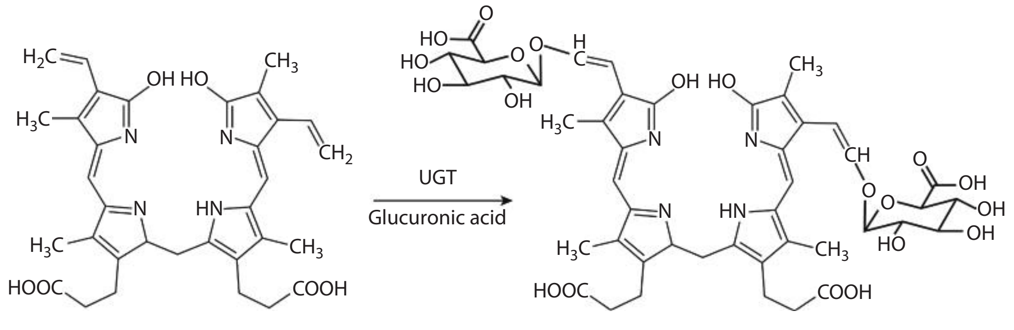

Bilirubin

Bilirubin diglucuronide

Figure 17.20 The formation of bilirubin glucuronides in the liver. UGT, UDP glucuronosyltransferases.

urobilin and stercobilin (Figure 17.21), respectively. These compounds are brown colored and are responsible for the characteristic color of feces.

The urobilinoids can be detected using the Schlesinger and Edelman tests. In the *Schlesinger* test, a sample is mixed with saturated zinc acetate in ethanol solution to form aurobilinoid–zinc chelation complex that emits a characteristic green fluorescence under ultraviolet light. The *Edelman* test is a variation of the Schlesinger test. A sample is treated with a mercuric salt solution to yield a pink-colored compound. Further treatment with a zinc salt produces fluorescence. However, less fluorescence is observed in the Edelman test than in the Schlesinger test. Inconclusive and inconsistent results are often obtained using these tests where fecal material sometimes gives no visible fluorescence. Additionally, the intensity of the fluorescence observed varies between samples. The reliability and selectivity of the tests can be increased using a spectrometric measurement of the fluorescence detection of fecal urobilinoids based on the principle of the Schlesinger test. A dry sample is treated with 1 mL of zinc acetate solution (1% zinc acetate methoxyethanol solution and 0.2% Tris). The suspension is then sonicated for 5 min, heated at 100°C for 10 min, cooled, and centrifuged. The presence of urobilinoids can be detected using excitation and emission maxima at 507 and 514 nm, respectively.

The disadvantages of the Schlesinger and the Edelman tests are their low species specificity as both tests cannot distinguish between human and other mammalian fecal materials. Moreover, under normal circumstances, up to 5% of urobilinogen is transported to the kidney and oxidized to urobilin in urine. Under some pathological conditions such as hepatic function disorders, the level of urobilin in the urine can be very high. Both the Schlesinger and Edelman tests respond to the urobilin of urine stains as well.

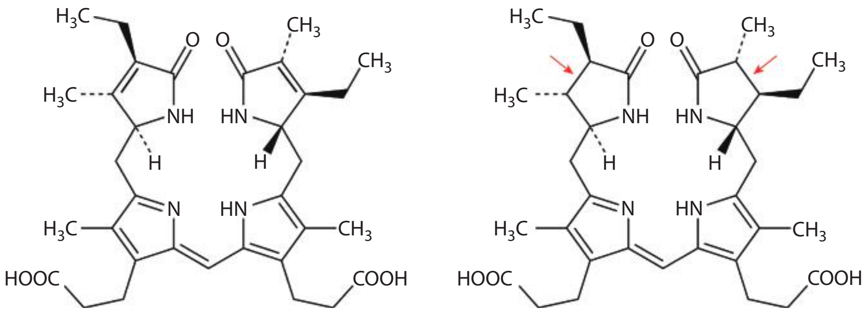

Urobilin

Stercobilin

Figure 17.21 Urobilinoids: urobilin and stercobilin. Arrow, site of reduction.

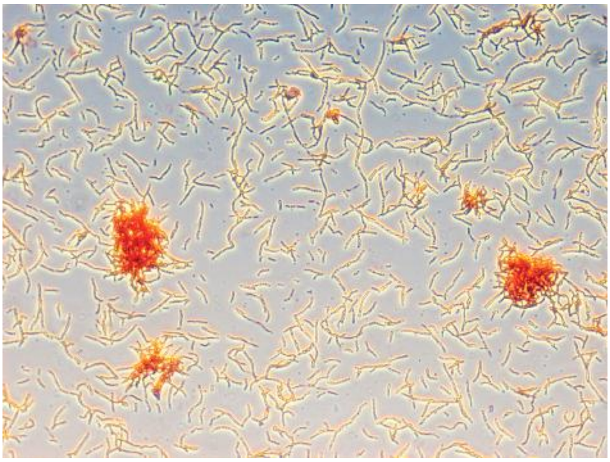

Figure 17.22 Intestinal bacteria. (© Richard C. Li.)

17.3.2.3 Fecal Bacterial Identification

The human intestinal microbiota contains more than 4000 bacterial species (Figure 17.22). Traditional methods for detecting fecal materials utilize the cultivation of fecal indicator bacteria such as *Escherichia coli* or *Enterococci* spp. However, these fecal indicator bacteria constitute only a small portion of the fecal microbiota, and thus may not be adequate for forensic identification purposes. *Bacteroides*, however, accounts for approximately 30% of fecal microbiota that are the predominant bacteria in human feces. *Bacteroides* can potentially be used for the forensic identification of feces. *Bacteroides* is a genus of rod-shaped, anaerobic gram-negative bacteria. These bacteria play a role in digesting complex carbohydrates and other substances that cannot be digested by human enzymes. The identification of *bacteroides* can be carried out by detecting specific DNA sequences of the *rpoB* gene, which encodes the β subunit of bacterial RNA polymerase (Chapter 11). The presence of the species-specific DNA sequence can be detected by RT-PCR utilizing primers that are specific to the target species but not to other species. Thus, only a targeted species can be amplified if it is present. Two fecal predominant bacteroides species, *B. uniformis* and *B. vulgatus*, can be detected in feces. *B. uniformis* is not detectable in blood, saliva, semen, urine, vaginal fluids, or on skin surfaces. Therefore, *B. uniformis* is considered as a specific indicator bacterium for forensic fecal identification. Sometimes, *B. vulgatus* can also be detected in vaginal fluid samples. Therefore, precaution should be taken in interpreting the results obtained using a *B. vulgatus* assay. Additionally, this method alone cannot discriminate between human and animal feces. Furthermore, fecal microbial populations can be affected by the host's diet. Individuals who consume saturated fats and proteins, which are abundant in Western diets, have predominantly *Bacteroides* species in their feces. However, individuals who consume a low-fat and carbohydrate-rich diet have predominantly *Prevotella* species, also a genus of gram-negative bacteria, in their feces.

17.4 Identification of Vomitus

17.4.1 Biology of Gastric Fluid

Gastric fluid can be found in stains derived from stomach wounds. Most often, gastric fluid is from vomitus found at a crime scene or as dried stains on clothing. Vomiting is the forceful expulsion of the contents of the stomach through the mouth (Figure 17.23). It is usually preceded by salivation, sweating, and the sensation of nausea. Vomiting usually begins with a deep

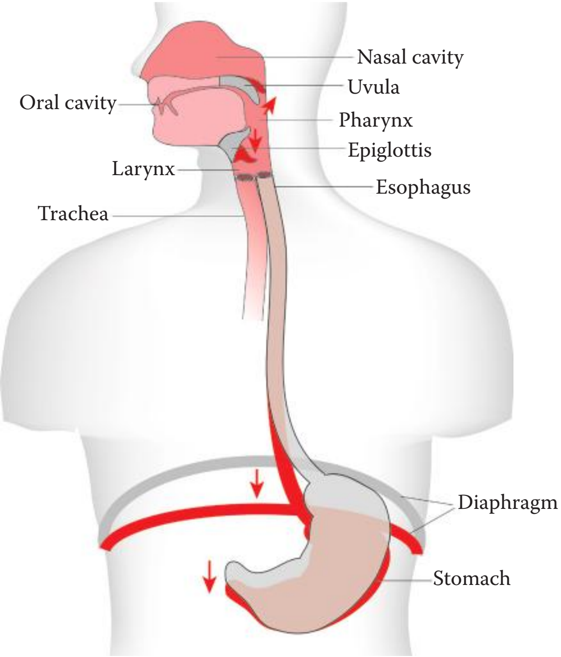

Figure 17.23 Act of vomiting. The contraction of the diaphragm and the stomach during the phase of vomiting is shown (in red). Additionally, the positions of the epiglottis and the uvula during vomiting are shown (in red). As a result, the stomach contents are expelled through the mouth. (© Richard C. Li.)

inhalation and the closure of the glottis. Subsequent contractions of the diaphragm and the abdominal muscles compress the stomach. The gastric contents are then forced upward through the relaxed sphincters and the esophagus, and are expelled through the mouth. Since the glottis is closed, vomitus usually does not enter the respiratory tract. Although the uvula is usually raised to close the nasal cavity, vomitus sometimes enters the nose. Vomiting can be caused by a wide variety of pathological conditions. Vomiting may be a specific response to acute intoxication in homicidal poisoning cases or it may be caused by trauma in a violent assault. Thus, the forensic identification of the gastric fluid can corroborate a criminal act.

The stomach stores ingested food until it can be emptied into the small intestines. When food enters the stomach, hydrochloric acid is secreted in large quantities, which facilitates the initial degradation of proteins. The stomach also secretes mucus that lubricates the gastric surface to protect the epithelium from acidic environments. Hormones such as gastrin, which are found in the gastric fluid, regulate acid secretion and gastric movement. A number of enzymes are secreted into the gastric fluid, including lipase, which plays a role in lipid hydrolysis, and gelatinase, which can hydrolyze gelatin. The stomach also secretes pepsinogens, which are enzyme precursors, into the gastric fluid. In the stomach, pepsinogens are activated by hydrochloric acid into pepsin, which is largely responsible for the digestion of proteins.

17.4.2 Vomitus Identification Assays

Vomitus is highly acidic and tends to be malodorous. The color of vomitus may be of forensic interest. Fresh blood in the vomit is usually bright red and suggests bleeding due to injuries, while dark red blood clots suggest bleeding in the stomach due to pathological conditions such as an ulcer. A microscopic examination (Section 17.3.2.1) can be performed to identify recently ingested food particles that are present in a sample. The presence of gastric fluid in vomitus can also be identified by the detection of pepsins secreted from the stomach. This identification test

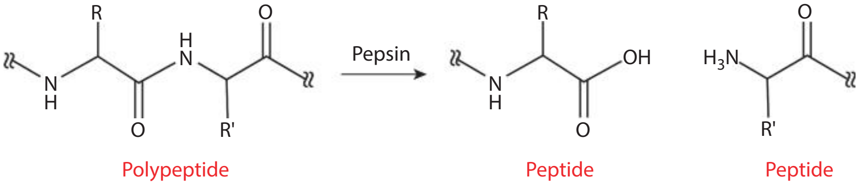

Figure 17.24 The cleavage of polypeptide by pepsin. Pepsin cleaves polypeptide at sites of aromatic amino acids. R and R′: tryptophan, phenylalanine, or tyrosine.

is based on the proteolytic activity of pepsins. Pepsins are endopeptidases that cleave primarily on peptide bonds in the middle of the protein. Aromatic amino acids such as tryptophan, phenylalanine, and tyrosine are the preferred targeted amino acids for the cleavage reaction by pepsins (Figure 17.24). In the pepsin-proteolytic assay, fibrin blue is used as a substrate for pepsin. Fibrin blue is an insoluble protein–dye complex that is colorless. In the presence of vomitus, pepsin cleaves fibrin blue and releases a chromophore that is soluble in water and exhibits a blue color. In the assay, a fibrin blue–containing agarose gel is utilized. The sample from vomitus is loaded onto the gel plate. After incubation, a blue ring around the sample can be observed as a result of the enzymatic reactivity of pepsin. The amounts of pepsin in a sample can be quantified. The results on the gel plate can be photographed and the dried gel plate can also be preserved as evidence. This method can determine the pepsin content of fresh and aged forensic samples. Bodily fluids other than vomitus do not show positive reactions with the use of this method. However, this method cannot distinguish the vomitus of humans from that of other vertebrates.

Bibliography

Akutsu, T., K. Watanabe, and K. Sakurada, Specificity, sensitivity, and operability of RSID-urine for forensic identification of urine: Comparison with ELISA for Tamm–Horsfall protein. *J Forensic Sci*, 2012, **57**(6): 1570–1573.

Akutsu, T., et al., Evaluation of Tamm–Horsfall protein and uroplakin III for forensic identification of urine. *J Forensic Sci*, 2010, **55**(3): 742–746.

Bedrosian, J.L., M.D. Stolorow, and M.A. Tahir, Development of a radial gel diffusion technique for the identification of urea in urine stains. *J Forensic Sci*, 1984, **29**(2): 601–606.

Biles, P.V. and G.C. Ziobro, Identification of the source of reagent variability in the xanthydrol/urea method. *J Assoc Official Anal Chem Int*, 1998, **81**(6): 1155–1161.

Bock, J.H., M.A. Lane, and D.O. Norris, *Identifying Plant Food Cells in Gastric Contents for Use in Forensic Investigations: A Laboratory Manual*, 1988. Washington, DC: US Department of Justice, National Institute of Justice.

Brunzel, N.A., *Fundamentals of Urine and Body Fluid Analysis*, 2nd edn., 2004. Philadelphia, PA: W.B. Saunder.

De Araújo, W.R., M.O. Salles, and T.R.L.C. Paixão, Development of an enzymeless electroanalytical method for the indirect detection of creatinine in urine samples. *Sensor Actuat B-Chem*, 2012, **173**: 847–851.

Farrugia, K.J., et al., Chemical enhancement of footwear impressions in urine on fabric. *Forensic Sci Int*, 2012, **214**(1–3): 67–81.

Farrugia, K.J., et al., A comparison of enhancement techniques for footwear impressions on dark and patterned fabrics. *J Forensic Sci*, 2013, **58**(6): 1472–1485.

Fujishiro, M., et al., Identification of human urinary stains by the uric acid/creatinine quotient and HPLC chromatogram. *J Showa Med Assoc*, 2008, **68**(3): 175–181.

Gaensslen, R.E., Identification of urine, in *Sourcebook in Forensic Serology, Immunology, and Biochemistry*, pp. 191–195, 1983. Washington, DC: US Government Printing Office.

Greenwald, I., The chemistry of Jaffe's reaction for creatinine: VI. A compound of picric acid with two molecules of creatinine. Its combinations with acid and alkali. *J Biol Chem*, 1930, **86**: 333–343.

18

Blood Group Typing and Protein Profiling

18.1 Blood Group Typing

18.1.1 Blood Groups

For the purposes of this text, blood groups are defined as antigen polymorphisms present on erythrocyte surfaces. Human erythrocyte surface membranes contain a variety of blood group antigens. Transfusion reactions occur when an incompatible type of blood is transfused into an individual, which can lead to severe symptoms or even death. Karl Landsteiner discovered the first blood group, known as the ABO system, in the early 1900s, while studying transfusion and transplantation. The discovery made blood transfusions feasible, and Landsteiner was awarded the Nobel Prize in 1930.

The International Society of Blood Transfusion currently recognizes 29 blood group systems, which include hundreds of antigen polymorphisms (Table 18.1). From the 1950s to the 1970s, the structures and biosynthesis pathways of many blood group antigens were determined. The genes for most of these blood group systems have been identified as well. The isolation of blood group genes has made it possible to understand the molecular mechanisms of the antigenic characteristics of the blood group systems.

The ABO system of antigens in human erythrocytes is the most commonly used blood group system for forensic applications. Forensic laboratories also use others, including the Rh, MNS, Kell, Duffy, and Kidd systems.

18.1.2 ABO Blood Group System

In the ABO blood group system, two types of antigens, designated A and B, give rise to four blood types:

- Type A individuals have the A antigen.
- Type B individuals have the B antigen.
- Type AB individuals have both A and B antigens.
- Type O individuals have neither A nor B antigens.

The antigens may be found in other bodily fluids as well as blood, such as amniotic fluid, saliva, and semen as well as many organs including the kidney, pancreas, liver, and lungs.

Table 18.1 Blood Group Systems

Number	Name	Symbol	Number of Antigens	Gene Names	Chromosomal Location
001	ABO	ABO	4	*ABO*	9q34.2
002	MNS	MNS	43	*GYPA, GYPB, GYPE*	4q31.21
003	P	P1	1		22q11.2-qter
004	Rh	RH	49	*RHD, RHCE*	1p36.11
005	Lutheran	LU	20	*LU*	19q13.32
006	Kell	KEL	25	*KEL*	7q34
007	Lewis	LE	6	*FUT3*	19p13.3
008	Duffy	FY	6	*FY*	1q23.2
009	Kidd	JK	3	*SLC41A1*	18q12.3
010	Diego	DI	21	*SLC4A1*	17q21.31
011	Yt	YT	2	*ACHE*	7q22.1
012	Xg	XG	2	*XG, MIC2*	Xp22.33, Yp11.3
013	Scianna	SC	5	*ERMAP*	1p34.2
014	Dombrock	DO	5	*DO*	12p12.3
015	Colton	CO	3	*AQP1*	7p14.3
016	Landsteiner-Wiener	LW	3	*ICAM4*	19p13.2
017	Chido/Rodgers	CH/RG	9	*C4A, C4B*	6p21.3
018	H	H	1	*FUT1*	19q13.33
019	Kx	XK	1	*XK*	Xp21.1
020	Gerbich	GE	8	*GYPC*	2q14.3
021	Cromer	CROM	12	*DAF*	1q32.2
022	Knops	KN	8	*CR1*	1q32.2
023	Indian	IN	2	*CD44*	11p13
024	Ok	OK	1	*BSG*	19p13.3
025	Raph	RAPH	1	*CD151*	11p15.5
026	John Milton Hagen	JMH	1	*SEMA7A*	15q24.1
027	I	I	1	*GCNT2*	6p24.2
028	Globoside	GLOB	1	*B3GALT3*	3q26.1
029	Gill	GIL	1	*AQP3*	9p13.3

Source: Adapted from Daniels, G.L., et al., *Vox Sang*, 87, 304–316, 2004.

18.1.2.1 Biosynthesis of Antigens

All individuals generate the *O antigen*, also known as the *H antigen*. The O antigen is synthesized by *fucosyltransferase*, a fucose transferase encoded by the *FUT* genes, which adds a fucose on the end of a glycolipid (in erythrocytes) or glycoprotein (in tissues). An additional monosaccharide (Figure 18.1) is then transferred to the O antigen by a transferase encoded by the *ABO* locus. The specificity of this enzyme determines the ABO blood type (Figure 18.2):

- The *A* allele produces the A-transferase, which transfers N-acetylgalactosamine to the O antigen and thus synthesizes the A antigen.
- The *B* allele produces the B-transferase, which transfers galactose to the O antigen and thus synthesizes the B antigen.
- The *O* allele has a mutation (small deletion), which eliminates transferase activity, and no modification of the O antigen occurs.

As a result, the A and B antigens differ in their terminal sugar molecules. Subgroups of blood types A and B have been described. The most important are the A_1 and A_2 antigens. Both A_1 and A_2 (and A_1B and A_2B) cells react with anti-A antibodies. However, A_1 cells react more strongly than A_2 cells. The apparent difference between A_1 and A_2 is that each A_1 cell contains more copies of the A antigen than A_2 cells.

18.1.2.2 Molecular Basis of the ABO System

A- and B-transferases are encoded by a single gene, *ABO*, on chromosome 9. The *ABO* gene (approximately 20 kb) is organized into seven exons. Most of its coding regions are located in exons 6 and 7 of the *ABO* locus, including the domain responsible for catalytic activity (Figure 18.3). The gene products of the *A* and *B* alleles differ by four amino acid substitutions (Table 18.2). In particular, amino acid residues at positions 266 and 268 are more important in determining the enzymatic property of a transferase.

The A^1 allele and A^2 allele differ in a single nucleotide deletion upstream from the translation stop codon. The resulting reading-frame shift in the A^2 allele abolishes the stop codon, yielding a product with an extra 21–amino acid residue at the C-terminus.

Subgroups of blood types O have also been reported. The sequence of the O^1 allele has a deletion of a single nucleotide at exon 6. This nucleotide deletion leads to a reading-frame shift generating a truncated protein, which lacks the catalytic domain. While the O^{1y} allele also has a single nucleotide deletion, it differs from O^1 by nine nucleotides within the coding sequence. O^1 and O^{1y} have identical phenotypes. There is also an O^2 allele, which is inactivated by a substitution

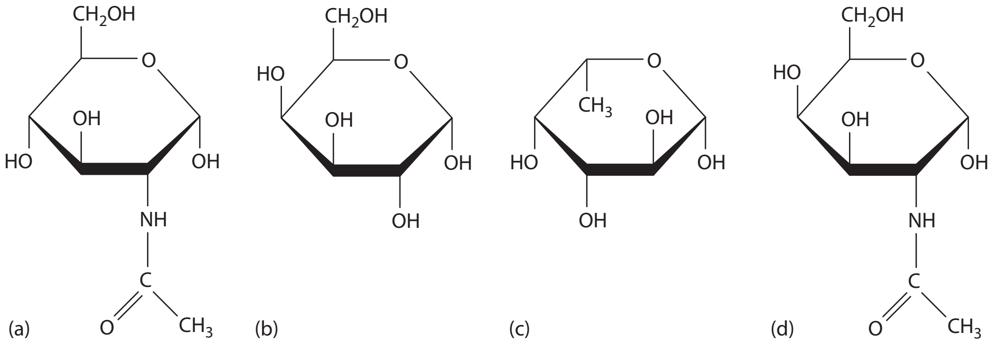

Figure 18.1 Chemical structures of (a) *N*-acetylglucosamine, (b) galactose, (c) fucose, and (d) *N*-acetylgalactosamine.

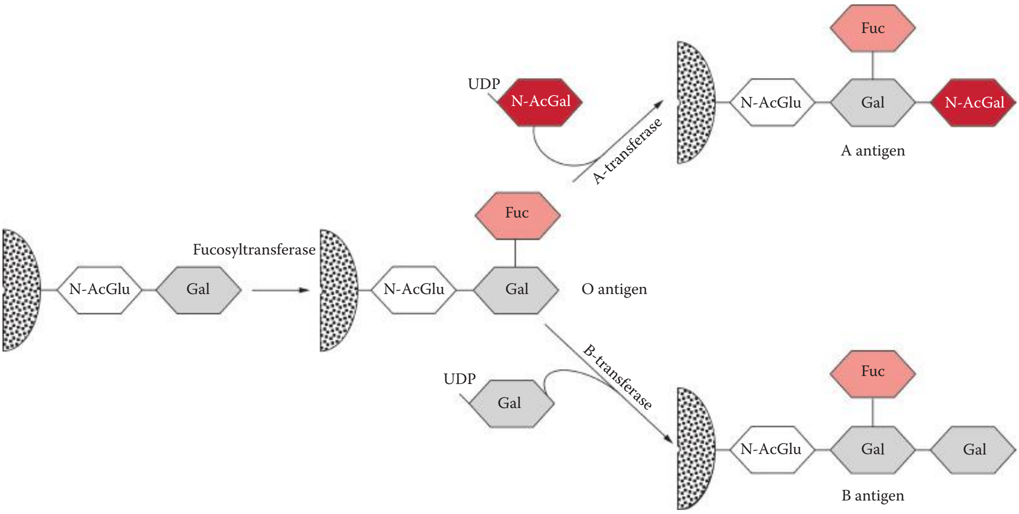
N-AcGlu
Gal
Fucosyltransferase
N-AcGlu
Gal
Fuc
O antigen
UDP
N-AcGal
A-transferase
N-AcGlu
Gal
Fuc
N-AcGal
A antigen
UDP
Gal
B-transferase
N-AcGlu
Gal
Fuc
Gal
B antigen

Figure 18.2 Biosynthesis of ABO antigens. O-antigen biosynthesis is catalyzed by fucosyltransferase. A-antigen biosynthesis is catalyzed by the A-transferase that transfers the N-acetylgalactosamine from the donor and uridine diphosphate (UDP)-*N*-acetylgalactosamine to the O antigen. B-antigen biosynthesis is catalyzed by the B-transferase that transfers the galactose from UDP-galactose to the O antigen. N-AcGlu, *N*-acetytglucosamine; Gal, galactose; Fuc, fucose; N-AcGal, *N*-acetylgalactosamine. (© Richard C. Li.)

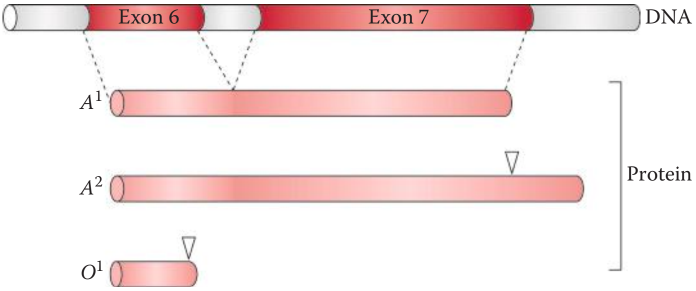
Exon 6
Exon 7
DNA
A1
A2
O1
Protein

Figure 18.3 Structure of ABO gene and variants. Exons 6 and 7 are shown. The deletion mutation in *A*² and *O*¹ variants is indicated by an inverted triangle and leads to their A_2 and O phenotypes, respectively. (© Richard C. Li.)

Table 18.2 Amino Acid Substitutions at Four Positions in Human ABO Variants *A*¹, *B*, and *O*²

ABO Variant	Amino Acid Position			
	176	235	266	268
*A*¹	Arginine	Glycine	Leucine	Glycine
B	Glycine	Serine	Methionine	Alanine
*O*²	Glycine	Glycine	Leucine	Arginine

mutation at glycine (position 268) by arginine. Additionally, a few dozen other rare *O* alleles, which yield inactive proteins, have also been documented.

18.1.2.3 Secretors

In addition to erythrocytes, individuals whose A, B, and O antigens can be found in other types of bodily fluids are referred to as *secretors*. Eighty percent of Caucasians are secretors. As described earlier, the O antigen is the substrate for the A- and B-transferase because the A- and B-transferase can only utilize a fucosylated substrate. The O antigen is synthesized by fucosylation of the terminal galactosyl residue catalyzed by the fucosyltransferase, which is encoded by *FUT* genes.

Chromosome 19 contains two homologous genes: *FUT1* and *FUT2*. *FUT1* is expressed in tissues of mesodermal origin (embryonic tissues that serve as precursors of hemopoietic tissues, muscle, the skeleton, and internal organs) and is responsible for the synthesis of the O antigen in erythrocytes. *FUT2* is expressed in tissues of endodermal origin (embryonic tissues that are precursors of the gut and other internal organs); it is responsible for the synthesis of the O antigen in secretions.

About 20% of Caucasian individuals (called *nonsecretors*) are homozygous for a nonsense mutation in *FUT2* at amino acid position 143, resulting in a truncated protein. Bodily fluids such as the semen of type A or B nonsecretors (who carry homozygous *FUT2* mutations) contain no A or B antigens despite containing active A- or B-transferases (Figure 18.4). This can be a problem in investigating sexual assault cases when the blood type of the seminal evidence needs to be determined. However, nonsecretors have O antigens on erythrocytes synthesized by *FUT1* and thus have A or B antigens in blood. Individuals carrying very rare homozygous *FUT1* mutations produce erythrocyte O-deficient phenotypes in which the erythrocytes express no O antigens and thus express neither A nor B antigens, regardless of *ABO* genotype. Individuals

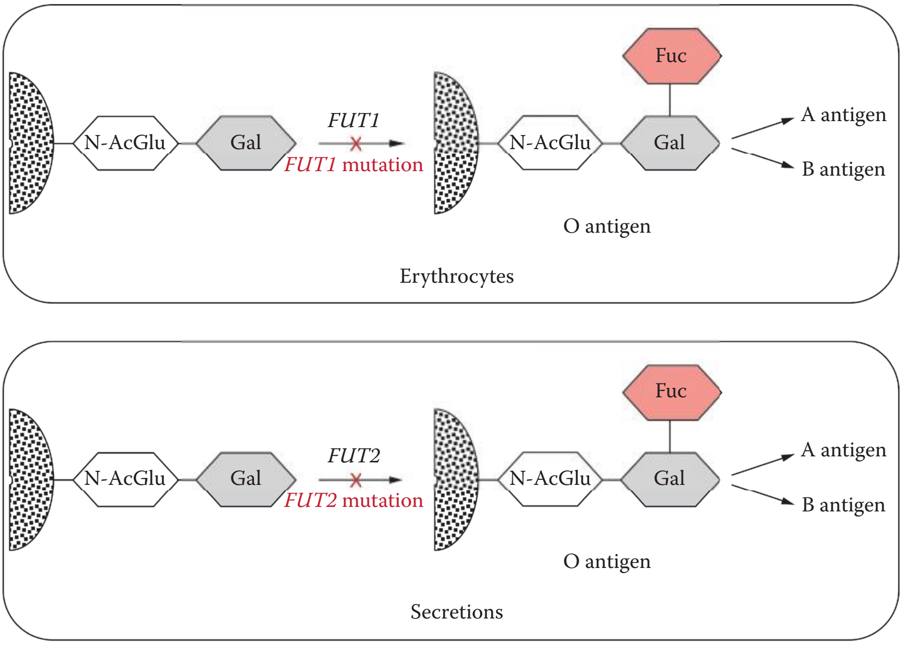

Figure 18.4 Tissue-specific O-antigen biosynthesis by *FUT1* and *FUT2* gene products. Tissue-specific O-antigen biosynthesis in erythrocytes is catalyzed by the *FUT1* gene product; in secretions, it is catalyzed by the *FUT2* gene product. The mutations abolishing the biosynthesis of O antigens are indicated. The *FUT2* mutation produces a nonsecretor phenotype. (© Richard C. Li.)

who carry both *FUT1* and *FUT2* mutations have no O antigens (nor A nor B antigens) in their erythrocytes and other bodily fluids and are known as Bombay (O_h) phenotypes.

18.1.2.4 Inheritance of A and B Antigens

A and *B* alleles are dominant. For *AO* and *BO* heterozygotes, the corresponding transferase synthesizes the A or B antigen. *A* and *B* alleles are codominant in *AB* heterozygotes because both transferase activities are expressed. The *OO* homozygote produces neither transferase activity and therefore lacks both antigens. The inheritance of *A* and *B* alleles obeys Mendelian principles (Chapter 25). For example, an individual with type B blood may have inherited a *B* allele from each parent or a *B* allele from one parent and an *O* allele from the other; thus, an individual whose phenotype is B may have the *BB* (homozygous) or *BO* (heterozygous) genotype. Conversely, if the blood types of the parents are known, the possible genotypes of their children can be determined. When both parents are type B (heterozygous), they may produce children with the genotype *BB* (B antigens from both parents), *BO* (B antigen from one parent, O from the other heterozygous parent), or *OO* (O antigens from parents who are both heterozygous). Thus, blood group typing can be used for paternity testing.

18.1.3 Forensic Applications of Blood Group Typing

The application and usefulness of blood typing in forensic identification are based on the ability to group individuals into four different types using the ABO blood system, allowing individuals to be identified. For example, if one crime scene blood sample is type B and a suspect has type A, the crime scene sample must have a different origin. However, if both the sample and the suspect are type A, the sample may have come from the same origin or from a different origin that happened to be type A.

Unfortunately, the probability that any two randomly chosen individuals have an identical blood type is very high. Approximately 42% of Caucasians have type A blood. The frequency of other blood types within the ABO system is shown in Figure 18.5. Multiple blood group systems were utilized to decrease the probability of a coincident match.

The A and B antigens are very stable and can be identified in dried blood even after many years. They can also be found in semen and other bodily fluids of secretors. Thus, in sexual assault cases, for example, the ABO type of a semen sample can be examined to identify a perpetrator.

18.1.4 Blood Group Typing Techniques

The most common assays used in forensic serology involve agglutination and include the Lattes crust and absorption–elution assays.

18.1.4.1 Lattes Crust Assay

In the early 1900s, Karl Landsteiner used his blood and blood obtained from his laboratory coworkers to test the effects of serum on erythrocytes. He discovered that naturally occurring

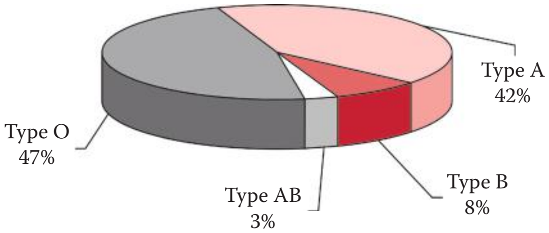

Figure 18.5 Frequency distributions of ABO types observed in American Caucasians. Different human populations may exhibit different frequencies of the four blood types. (© Richard C. Li.)

antibodies in serum caused agglutination of certain erythrocytes, and the agglutination patterns observed were designated A, B, and O. Each pattern indicated the presence or absence of a particular antigen on erythrocytes.

Shortly after birth, newborn infants develop antibodies against antigens that are not present in their own bodies. For example, type A individuals develop anti-B antibodies, type B individuals develop anti-A antibodies, type O individuals develop both types of antibodies, and type AB individuals do not develop anti-A or anti-B antibodies. When the plasma of a type A individual is mixed with type B cells, the anti-B antibodies from the type A individual cause the type B cells to agglutinate. This result forms the basis for blood group typing.

The Lattes crust assay relies on the principles of Landsteiner's experiments. It is an agglutination-based assay that utilizes the A, B, and O indicator cells to test the agglutination reaction with its corresponding naturally occurring serum antibodies in a questioned sample. The procedure for the Lattes crust assay is described in Box 18.1 and illustrated in Figure 18.6. Typical results are summarized in Table 18.3 and illustrated in Figure 18.7. Type A blood contains naturally occurring anti-B antibodies that agglutinate only with B cells. Likewise, type B blood agglutinates only with A cells, type O blood agglutinates with both A and B cells, and type AB blood does not agglutinate with any cells.

The Lattes crust assay is simple and rapid. However, one limitation is that the assay is not very sensitive and requires a large quantity of blood. Recall that successful agglutination reactions usually require intact cells. The agglutination assay of forensic samples is, therefore, difficult to carry out because blood cells lyse when they are dry. Therefore, this method is not reliable for testing old stains.

BOX 18.1 LATTES CRUST ASSAY PROCEDURE

1. Place small quantities of blood crust from a specimen on a microscopic slide and place a cover slide over the crusts. Prepare slides for A, B, and O cells separately.
2. Prepare cell suspensions with saline (0.85% NaCl in phosphate buffer, pH 7.4) for the A, B, and O cells separately.
3. Apply a few drops of the A-cell suspension and allow the cells to diffuse under the cover slip. Repeat this step for B cells and O cells.
4. Incubate the slides in a moisture chamber at room temperature for 2 h.
5. Examine results under a microscope.

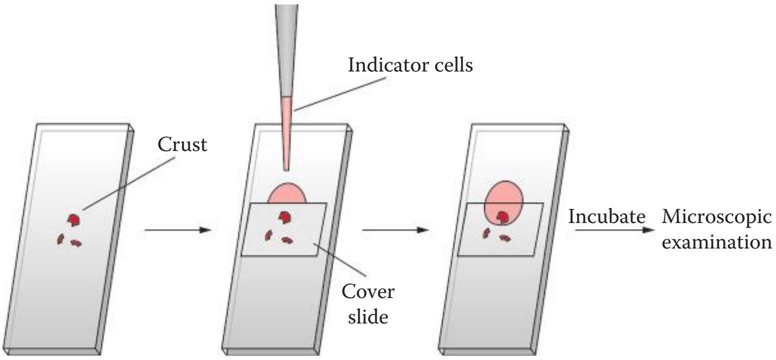

Figure 18.6 Lattes crust assay. (© Richard C. Li.)

Table 18.3 Representative Results of Lattes Crust Assay

Blood Type	Serum Antibody	Agglutination Reaction Observed
A	Anti-B	B cells
B	Anti-A	A cells
O	Anti-A, anti-B	A cells, B cells
AB	None	None

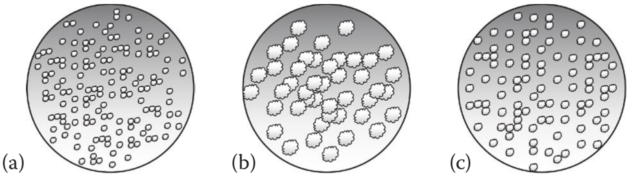

Figure 18.7 Diagram of Lattes crust assay results. (a) Indicator cells added before incubation. (b) Strong agglutination: large clumps are observed after incubation. (c) Negative agglutination: a cloudy background may be observed after incubation.

18.1.4.2 Absorption–Elution Assay

The absorption–elution assay is highly sensitive and can be used for testing dried bloodstains. This method indirectly detects the presence of antigens. The antigens are immobilized in a solid phase (Figure 18.8). At low temperatures, the antigens bind to their corresponding antibodies: anti-A antibodies, anti-B antibodies, or anti-O lectins. (The anti-O lectin is isolated from plants and reacts strongly with the O antigen present in type O blood, but has some cross-reaction with the A antigen). The excess unbound antibodies are removed by washing, and the bound antibodies are then eluted at higher temperatures (recall that antigen–antibody binding can be affected by temperature; Chapter 13). The eluted antibodies can then be identified by an agglutination assay using A, B, and O indicator cells.

Typical results of an absorption–elution assay are summarized in Table 18.4. The bloodstains containing the A antigen can bind to anti-A antibodies. The eluted anti-A antibody can form agglutination with A cells. Likewise, for type B blood, the eluted anti-B antibody can form agglutination with B cells; for type AB blood, the eluted antibodies can form agglutination with both A and B cells; and with type O blood, the eluted anti-O lectins can form agglutination with O cells.

18.2 Forensic Protein Profiling

Because of the limitations of blood group systems, inherited protein polymorphic markers have been utilized to decrease the chances of matches between two unrelated individuals. The amino acid sequences of many proteins vary in the human population. An estimated 20%–30% of the proteins in humans are polymorphic. Some of the variations in amino acid sequences affect the function of proteins, but many of them exert little or no effect on protein function. Thus, individuals can be divided into groups based on the types of protein polymorphisms. A combination of the blood group systems and protein polymorphic markers can be used for

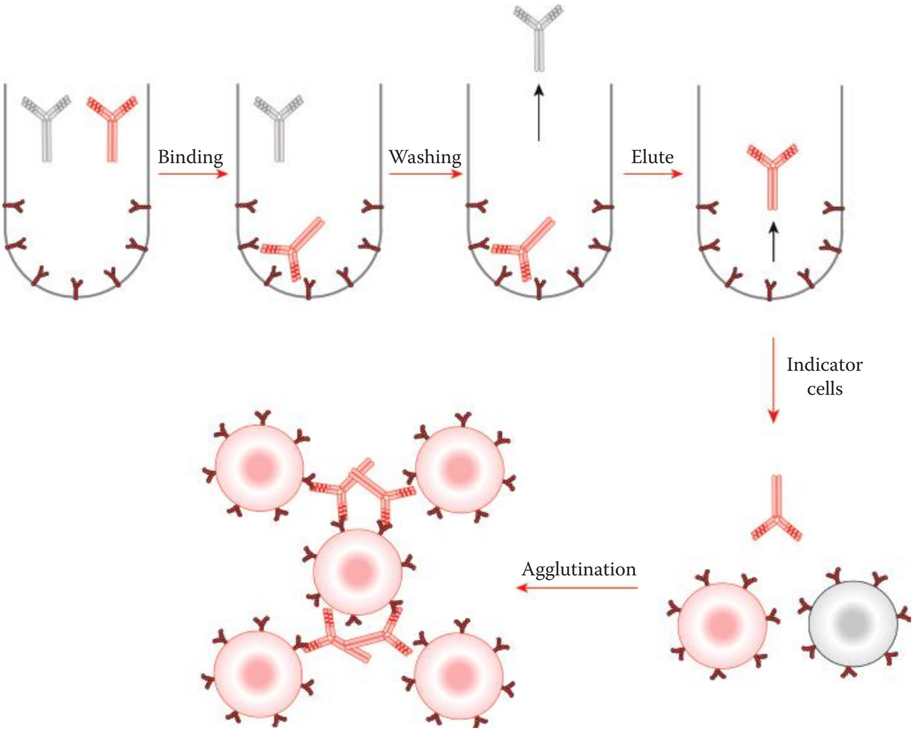

Figure 18.8 Absorption–elution assay. An antigen sample is immobilized on a solid-phase matrix. Antibodies (two different types are shown) are added. The antibody that is specific to the antigen binds. Unbound antibody is washed away. The bound antibody is then eluted. The eluted antibody is tested with indicator cells (two different types are shown). The indicator cell that is specific to the antibody eluted shows a positive agglutination reaction. (© Richard C. Li.)

Table 18.4 Representative Results of Absorption–Elution Assay

Blood Type (Stain)	Antibody Bound and Eluted	Agglutination Reaction Observed
A	Anti-A	A cells
B	Anti-B	B cells
O	Anti-O	O cells
AB	Anti-A, anti-B	A cells, B cells

criminal investigations and paternity testing. The probability that results for two unrelated persons would match is decreased to one in several hundred through use of the blood-typing and protein-profiling techniques.

18.2.1 Methods

Identification of protein polymorphisms is performed through electrophoretic separation based on the molecular weights (*Mr*) and charges of the protein variants.

18.2.1.1 Matrices Supporting Protein Electrophoresis

Electrophoresis of proteins is generally carried out in a support material, also called the matrix, to separate various macromolecules. The matrix also reduces the effects of diffusion and convection on the macromolecules. Historically, protein profiling for forensic application utilizes two types of matrices: papers such as cellulose acetate; and gels composed of starch, agar, agarose, or polyacrylamide. The first polymorphic protein marker, phosphoglucomutase, was characterized by starch-gel electrophoresis. However, agarose and polyacrylamide became more commonly used in electrophoresis due to good reproducibility and reliability (Table 18.5).

18.2.1.2 Separation by Molecular Weight

An electrophoretic method is frequently utilized to resolve various proteins based on their molecular weights. Native electrophoresis, also known as nondenaturing electrophoresis, can be used to isolate proteins for studying the functions of proteins. Biological activity of the protein can be retained for further analysis. However, some proteins are not well separated in electrophoresis in their native form. Thus, it may be necessary to denature the proteins in order for them to be resolved better during separation. This process is called *denaturing protein electrophoresis*. The following additives can be used:

18.2.1.2.1 Reducing Agents

It is common to include reducing agents such as mercaptoethanol (ME), dithiothreitol (DTT), or sodium mercaptoethane sulfonate (MESNA) to denature proteins. Reducing agents cleave the disulfide bonds of proteins. As a result, protein shape becomes unfolded and linear. These agents can be used during sample preparation and can also be added to the electrophoresis buffer.

18.2.1.2.2 Detergents

Detergents disrupt noncovalent interactions within the structures of native proteins. The procedure is generally performed with sodium dodecylsulfate (SDS), a strong anion detergent that binds to most proteins in amounts proportional to the molecular weight of the protein (approximately one molecule of SDS for two amino acids). The bound SDS contributes a large net negative charge on the protein, which masks any surface charges of the native protein. As a result, the charge-to-mass ratio of the protein becomes a constant. As with reducing agents, the various native conformations of proteins change to a more uniformly linear shape when SDS is bound.

Table 18.5 Properties of Matrices Supporting Protein Electrophoresis

Supporting Matrix	Pore Size	EEO	Reproducibility	Strength	Preparation	Toxicity
Cellulose acetate	Large	High	Poor	Good	Simple	Nontoxic
Starch	Large	High	Poor	Fragile	Simple	Nontoxic
Agar	Large	High	Poor	Fragile	Simple	Nontoxic
Agarose	Large	Low	Good	Fragile	Simple	Nontoxic
Polyacrylamide	Small	Very Low	Good	High; tolerates high electronic field	Complex	Toxic

Note: Electroendosmosis (EEO) occurs when fixed charges of the supporting matrix cause liquid flow toward the electrodes. A matrix with high EEO may affect the mobilities and separation performances of proteins during electrophoresis.

Electrophoretic mobility in the presence of SDS, therefore, becomes based on *M*r rather than both *M*r and the charge. Smaller proteins move through the pores of the gel matrix more rapidly than larger proteins. As a result, the larger the size of the protein, the smaller its electrophoretic mobility.

SDS gel electrophoresis can also be used to determine the *M*r of an unidentified protein based on its electrophoretic mobility on the gel. Standard marker proteins of known molecular weight are run on the same gel and allow the estimation of the *M*r of an unknown protein. A linear plot of log *M*r values of marker proteins versus relative migration during electrophoresis allows the molecular weight of the unknown protein to be determined from the graph.

18.2.1.3 Separation by Isoelectric Point

The *isoelectric focusing (IEF)* technique can be used to separate proteins according to their *isoelectric points (pI)*. The pI is the pH value at which the net electric charge of an amino acid is zero. All proteins are composed of amino acids, and each has its own characteristic pI at which its net electric charge is zero and does not migrate in an electric field.

In IEF electrophoresis, a pH gradient is created in a gel between the electrodes, and a protein sample is placed in a well on the gel. With an applied electric field, proteins enter the gel and migrate until they reach a pH equivalent to their pI values, at which they lose mobility (Figure 18.9). IEF, based on molecular charge, is capable of producing sharper bands than denaturing protein electrophoresis and thus has a higher resolving power. The technique can detect very low quantities of proteins in samples. A pH gradient in the gel is established by utilizing materials such as carrier ampholytes or immobilines that are dispersed in the gel. Carrier ampholytes are synthetic amphoteric compounds that contain multiple weak ionizable moieties acting as either acids or bases. To establish a pH gradient, a mixture of ampholytes with slightly different pIs is directly added to an IEF gel. The pH gradient is generated by applying an electric field on the ampholyte-containing gel. Under the electric field, the negatively charged ampholytes migrate toward the anode, and the positively charged ampholytes migrate to the cathode. As a result, a gradual pH gradient is created between the anodal end of the gel (acidic) and the cathodal side of the gel (basic). Immobilines are a series of modified acrylamide monomers that can be acidic or basic. The pH gradient of an IEF gel can also be established using a gradient-forming device that changes the proportion of the immobilines added to the gel matrix mixture as it is loaded into the gel-casting apparatus.

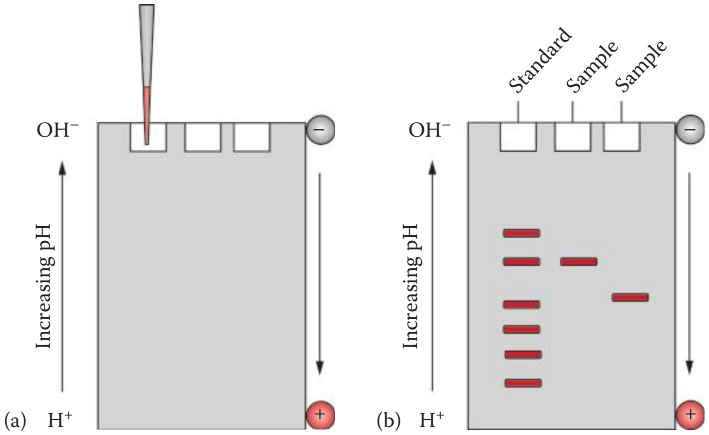

Figure 18.9 Isoelectric focusing. (a) A pH gradient is established by allowing low-molecular-weight organic acids and bases to distribute themselves in an electric field across a gel. A sample containing a protein mixture is loaded into a sample well. (b) During the electrophoresis, each protein migrates until it matches its pI. Proteins with different pIs are separated. (© Richard C. Li.)

18.2.2 Erythrocyte Protein Polymorphisms

18.2.2.1 Erythrocyte Isoenzymes

The human erythrocyte contains a number of *isoenzymes*, which are multiple forms of an enzyme that catalyze the same reaction but differ in their amino acid sequences. Individuals can be divided into groups on the basis of the different isoenzymes present in their erythrocytes. The isoenzyme type is also inherited according to Mendelian principles.

The polymorphism of erythrocyte phosphoglucomutase (PGM) was first described in the 1960s and was later successfully applied to the testing of bloodstains. PGM, an important metabolic enzyme, catalyzes the reversible conversion of glucose-1-phosphate and glucose-6-phosphate. The PGM found in erythrocytes is encoded at the *PGM1* locus at chromosome 1. The PGM encoded by *PGM1* can also be found in semen and thus can be utilized for the testing of semen samples in sexual assault cases. The protein polymorphisms of the PGM have two alleles, which result in three different phenotypes, depending on the combination of the two alleles. The success in the forensic application of PGM led to the similar use of many other erythrocyte isoenzyme polymorphisms. The most commonly used erythrocyte isoenzyme systems are listed in Table 18.6.

18.2.2.2 Hemoglobin

Recall that the use of hemoglobin (Hb) in screening and confirmatory blood tests was discussed in Chapter 12. Adult human Hb consists of two α chains and two β chains. Each polypeptide chain contains a heme group involved in oxygen binding. A very small portion of blood possesses a form of the human adult Hb consisting of two α chains and two δ chains.

More than 200 Hb variants have been identified and can be useful as markers for forensic applications. In particular, two types of human Hb variants are important in forensic testing: fetal Hb and sickle-cell Hb (Hb S). Hb S is the factor responsible for sickle-cell disease (Figure 18.10). Hb variants can be resolved using electrophoresis (Figure 18.11).

18.2.2.2.1 Fetal Hemoglobin

Humans have three forms of Hb during their development: embryonic, fetal, and adult Hb. In adults, the Hb tetramer consists of two identical α and two identical β chains. Embryonic erythrocytes contain Hb tetramers that are different from the adult form. Each embryonic Hb consists of two identical α-like chains and two identical β-like chains. The embryonic Hb is gradually replaced during pregnancy (approximately 3 months after conception) by fetal Hb, which comprises approximately 70% of the Hb in fetal blood. The fetal Hb has two identical α chains and two identical γ chains.

TABLE 18.6 Common Isoenzymes Used for Forensic Protein Profiling

Erythrocyte Isoenzyme	Protein Symbol	Number of Alleles
Phosphoglucomutase	PGM	2[a]
Erythrocyte acid phosphatase	ACP/EAP	3
Esterase D	ESD	2
Adenylate kinase	AK	2
Glyoxalase I	GLO	2
Adenosine deaminase	ADA	2

[a] Ten alleles can be observed using IEF electrophoresis.

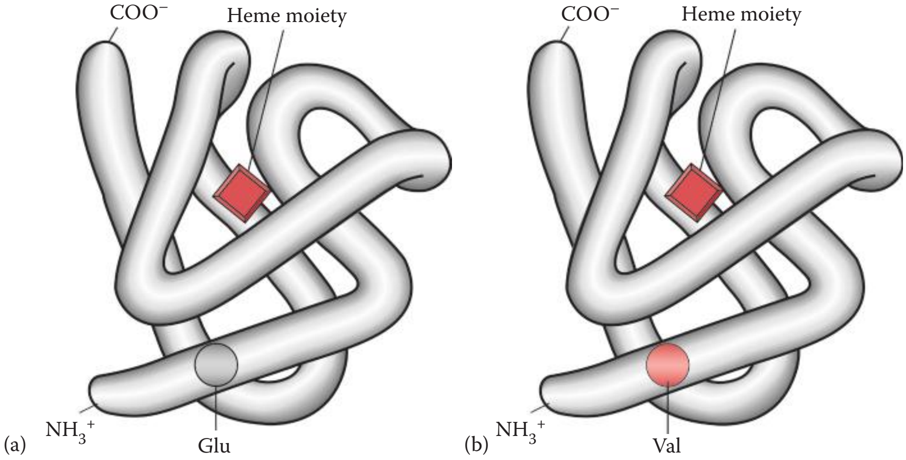

Figure 18.10 Normal and sickle-cell hemoglobin β chains. (a) Normal hemoglobin β chain contains a glutamic acid residue (Glu) at position 6 of the N-terminal of the protein. (b) At position 6 of the sickle-cell hemoglobin β chain, the glutamic acid residue is replaced by a valine (Val). (© Richard C. Li.)

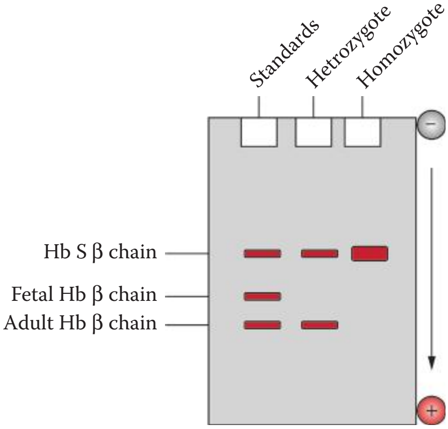

Figure 18.11 Hemoglobins resolved by isoelectric focusing electrophoresis. (© Richard C. Li.)

The embryonic and fetal Hbs have higher affinities to oxygen, required to provide the embryo and fetus with sufficient amounts of oxygen taken from maternal blood. Fetal Hb is replaced by adult Hb approximately 6 months after birth. These Hbs are encoded by their corresponding genes located at the globin gene clusters. The detection of fetal Hb in a bloodstain via electrophoresis can provide important evidence in cases of infanticide and concealed birth.

18.2.2.2.2 Hemoglobin S

Hb S polymorphism has forensic importance in identifying individuals. The Hb S polymorphism is observed in high frequencies among those of African heritage and some Hispanic populations. Such a protein polymorphic marker can provide investigational leads for the indication of the ethnic origin of a perpetrator. Hb S transports oxygen much less efficiently than normal Hb. Individuals who are homozygous for Hb S usually die early after suffering from sickle-cell anemia and related complications. However, a heterozygous individual (an individual with a copy of the wild-type Hb allele from one parent and a copy of the Hb S allele from the other) can survive. This condition is known as the sickle-cell trait.

In the 1950s, Vernon Ingram of Cambridge University discovered the molecular mechanism of the Hb S defect. His work revealed that the Hb S bears a mutation, which changes the glutamic acid in wild type to a valine at the sixth amino acid from the N-terminal end of the β chain. This substitution of amino acids causes a major change in the structure of the β chain, which in turn results in sickle-cell anemia.

18.2.3 Serum Protein Polymorphisms

The serum portion of blood consists of a large number of proteins. The work on serum proteins for forensic purposes started in the 1950s, when variations in serum proteins were found useful for distinguishing individuals. Over the years, a number of serum proteins were characterized and applied for forensic testing. Haptoglobin (Hp) was the most widely used of the polymorphic serum proteins in forensic biology (Figure 18.12). Haptoglobin is a protein that binds and transports Hb from the bloodstream to the liver for the recycling of the iron contained in the Hb.

Immunoglobulin (Chapter 10) accounts for approximately 15% of serum protein and has been found to be highly variable. Two immunoglobulin proteins are utilized for forensic application. The γ chain protein (G_m) is the heavy chain of immunoglobulin G and the κ chain protein (K_m) is one of two types of the light chain of all immunoglobulins. Table 18.7 lists common serum group systems. All exhibit genetic variations and can be detected in bloodstains. The variants of these proteins can be determined by electrophoresis or serological methods.

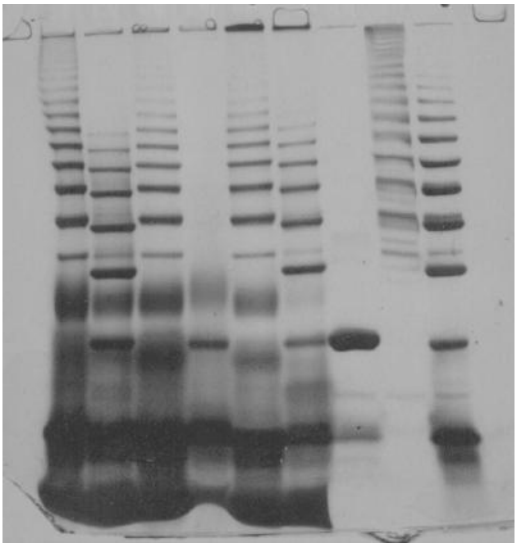

Figure 18.12 Polyacrylamide gel electrophoresis of haptoglobin proteins. From left to right: Hp2, Hp2-1, Hp2, Hp1, Hp2, Hp2-1, Hp1, Hp2, Hp2-1. An anode is at the bottom. (From James, S. and Nordby, J.J., *Forensic Science: An Introduction to Scientific and Investigative Techniques*, CRC Press, Boca Raton, FL, 2005. With permission.)

TABLE 18.7 Serum Proteins Used for Forensic Protein Profiling

Serum Protein	Protein Symbol	Gene Symbol	Chromosomal Location	Number of Amino Acids	Number of Alleles
Haptoglobin	Hp	*HP*	16q22.1	328/387	2
Group-specific component	Gc	*GC*	4q12–13	458	2
Transferrin	Tf	*TF*	3q22.1	679	3
Protease inhibitor (α1-antitrypsin)	Pi	*PI*	14q32.1	394	Many

Source: Adapted from Yuasa, I. and Umetsu, K., *Leg Med*, 7, 251–254, 2005.

Bibliography

Blood Groups

Achermann, F.J., et al., Soluble type A substance in fresh-frozen plasma as a function of ABO and secretor genotypes and Lewis phenotype. *Transfus Apher Sci*, 2005, **32**(3): 255–262.

Aki, K., A. Izumi, and E. Hosoi, The evaluation of histo-blood group ABO typing by flow cytometric and PCR-amplification of specific alleles analyses and their application in clinical laboratories. *J Med Invest*, 2012, **59**(1–2): 143–151.

Allen, F.H. Jr. and R.E. Rosenfield, Review of Rh serology. Eight new antigens in nine years. *Haematologia* (Budap), 1972, **6**(1): 113–120.

Anstee, D.J. and M.J. Tanner, Separation of ABH, I, Ss antigenic activity from the MN-active sialoglycoprotein of the human erythrocyte membrane. *Vox Sang*, 1975, **29**(5): 378–389.

Asari, M., et al., A new method for human ABO genotyping using a universal reporter primer system. *J Forensic Sci*, 2010, **55**(6): 1576–1581.

Bargagna, M. and M. Pereira, A study of absorption–elution as a method of identification of Rhesus antigens in dried bloodstains. *J Forensic Sci Soc*, 1967, **7**(3): 123–130.

Bargagna, M., M. Sabelli, and C. Giacomelli, The detection of Rh antigens (D,C,c,E,e) on bloodstains by a micro-elution technique using low ionic strength solution (LISS) and papain-treated red cells. *Forensic Sci Int*, 1982, **19**(2): 197–203.

Bassler, G., Determination of the Lewis blood group substances in stains of forensically relevant body fluids. *Forensic Sci Int*, 1986, **30**(1): 29–35.

Boorman, K.E., B.E. Dodd, and P.J. Lincoln, *Blood Group Serology*, 5th edn., vol. 4, p. 495, 1977. London: Churchill Livingstone.

Brauner, P., DNA typing and blood transfusion. *J Forensic Sci*, 1996, **41**(5): 895–897.

Brewer, C.A., P.L. Cropp, and L.E. Sharman, A low ionic strength, hemagglutinating, autoanalyzer for rhesus typing of dried bloodstains. *J Forensic Sci*, 1976, **21**(4): 811–815.

Busuttil, A., et al., Assessment of Lewis blood group antigens and secretor status in autopsy specimens. *Forensic Sci Int*, 1993, **61**(2–3): 133–140.

Camp, F.R. Jr., Forensic serology in the United States. I. Blood grouping and blood transfusion—Historical aspects. *Am J Forensic Med Pathol*, 1980, **1**(1): 47–55.

Cartron, J.P., et al., "Weak A" phenotypes. Relationship between red cell agglutinability and antigen site density. *Immunology*, 1974, **27**(4): 723–727.

Cayzer, I. and P.H. Whitehead, The use of sensitized latex in the identification of human bloodstains. *J Forensic Sci Soc*, 1973, **13**(3): 179–181.

Chatterji, P.K., A simplified mixed agglutination technique for ABO grouping of dried bloodstains using cellulose acetate sheets. *J Forensic Sci Soc*, 1978, **17**(2–3): 143–144.

Chester, M.A. and M.L. Olsson, The ABO blood group gene: A locus of considerable genetic diversity. *Transfus Med Rev*, 2001, **15**(3): 177–200.

19

Variable Number Tandem Repeat Profiling

Tandem repeats are abundant in the human genome. *Minisatellites* were first defined as a class of tandem repeats in the 1980s. Some of these repeats share a GC-rich core sequence. Subsequently, tandem repeats with higher AT contents of core sequence have also been characterized. The minisatellites are also called *variable number tandem repeats* (*VNTRs*), as shown in Figure 19.1. The repeat unit length of a VNTR can range from several to hundreds of base pairs (bp). The tandem repeat arrays can be kilobases (kb, corresponding to 10^3 bp) long, and the numbers of tandem repeat units in some VNTR loci are highly variable, leading to variable lengths of DNA fragments. A genotype is defined by a particular number of tandem repeat units at a given locus.

Table 19.1 lists the common VNTR loci used for forensic testing. To achieve high discriminating power, the VNTR loci should not be linked, which means that they should be inherited independently of each other. For example, loci located on different chromosomes or far apart on the same chromosome can be used (Section 21.2). Many VNTR loci used for forensic applications are highly polymorphic, and as many as hundreds of different genotypes per locus can be observed among the population. The discriminating power of VNTR loci used for forensic testing can be measured by *population match probability* (P_m; Chapter 25). The lower the P_m, the less likely a match will occur between two randomly chosen individuals. A P_m of up to 10^{-12} can be achieved by testing several VNTR loci.

19.1 Restriction Fragment Length Polymorphism

VNTR profiling utilizes RFLP—the first historical method used in forensic DNA testing (Figure 19.2). It utilizes *restriction endonucleases* that recognize and cleave specific sites along the DNA sequence. Cleavage of a DNA sample with a particular restriction endonuclease results in a reproducible set of restriction fragments of various lengths. Appropriate restriction endonucleases should be selected so that the genomic DNA is cleaved at sites that flank the VNTR core repeat region. The resulting fragments are then separated according to their sizes by gel electrophoresis through a standard agarose gel (Chapter 9).

The DNA is then processed using the *Southern transfer and hybridization* technique. The DNA is denatured and transferred from the gel to a supporting matrix such as a nylon or nitrocellulose membrane. The DNA immobilized on the membrane is then hybridized with a labeled *probe*. Only bands of DNA that have complementary sequences to the probe are recognized by detection systems such as autoradiography (Section 19.1.4). Using the RFLP technique, the

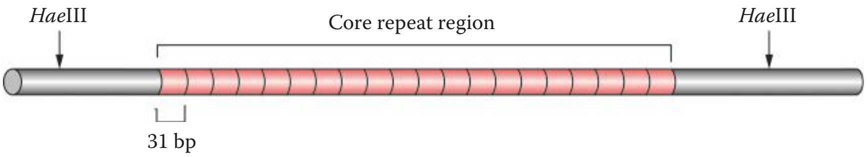

Figure 19.1 VNTR locus D2S44 (2q21.3–2q22). Each repeat unit consists of 31 bp. *Hae*III represents the *Hae*III restriction site.

Table 19.1 Common VNTR Loci

Locus	Chromosome Location	Repeat Unit Length (bp)	*Hae*III Fragment Size (kb)	Probe
D1S7	1	9	0.5–12	MS1
D2S44	2	31	0.7–8.5	yNH24
D4S139	4	31	2–12	pH30
D10S28	10	33	0.4–10	pTBQ7
D14S13	14	15	0.7–12	pCMM101
D16S85	16	17	0.2–5	3′HVR
D17S26	17	18	0.7–11	pEFD52
D17S79	17	38	0.5–3	V1

Source: Adapted from Budowle, B., et al., *DNA Typing Protocols: Molecular Biology and Forensic Analysis*, Eaton Publishing, Natick, MA, 2000; Office of Justice Programs, Future of forensic DNA testing: Predictions of the Research and Development Working Group. National Institute of Justice, US Department of Justice, 2000.

length variations among restriction sites can be detected. Most forensic applications focus on the length variations of VNTR regions located between two restriction sites.

In summary, the RFLP method includes several steps: (1) genomic DNA preparation, (2) restriction endonuclease digestion of the genomic DNA into fragments, (3) agarose gel electrophoretic separation of the DNA fragments according to size, (4) transfer of DNA fragments using Southern transfer, (5) hybridization with locus-specific probes, and (6) detection of locus-specific bands by autoradiography or chemiluminescence.

19.1.1 Restriction Endonuclease Digestion

Restriction endonucleases are enzymes that cleave the phosphodiester bond of DNA at or near specific recognition nucleotide sequences known as restriction sites. A restriction site usually is a short motif that is 4–8 bp in length. It often has a specific palindromic recognition sequence, that is, a segment of double-stranded DNA in which the nucleotide's sequence is identical with an inverted sequence in the complementary strand. Thus, double-stranded DNA is required to be cleaved by most restriction endonucleases. As a result, both sticky ends and blunt ends of restriction fragments can be generated after the cleavage (Figure 19.3).

To date, hundreds of restriction endonucleases have been described. They are traditionally classified into three types on the basis of subunit composition and enzymatic properties. Type II restriction endonucleases are most commonly used in molecular biology applications. Type II restriction endonucleases, requiring magnesium as a cofactor, usually cleave DNA at defined positions within their recognition sequences. The Enzyme Commission (EC) number of type II

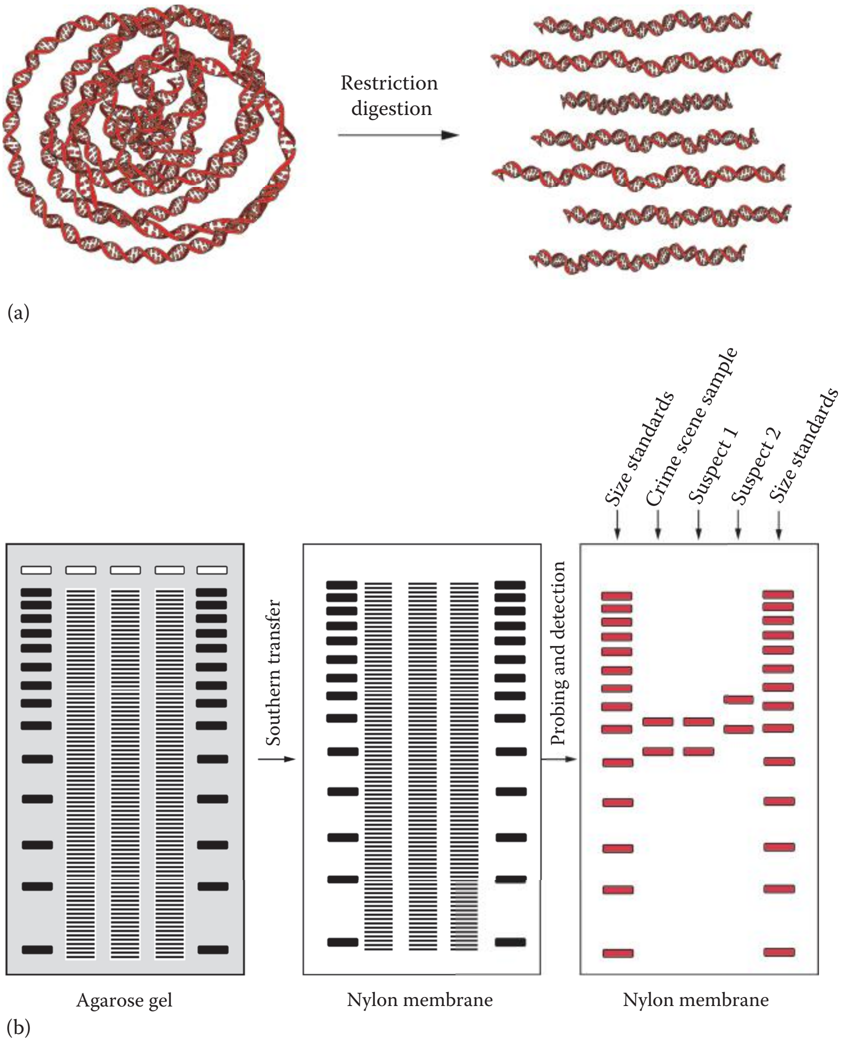
Restriction
digestion
(a)
Size standards
Crime scene sample
Suspect 1
Suspect 2
Size standards
Southern transfer
Probing and detection
Agarose gel
Nylon membrane
Nylon membrane
(b)

Figure 19.2 RFLP. (a) Restriction digestion generates restriction fragments with various lengths of genomic DNA. (b) Restriction fragments are separated by gel electrophoresis. DNA is transferred to a solid phase and probed. The signal is detected and the DNA fragment of interest can be observed. Band patterns of heterozygous loci of individuals are shown. (© Richard C. Li.)

restriction endonucleases is EC 3.1.21.4. The EC number is a numerical classification system of nomenclature based on the chemical reaction that is catalyzed by the enzyme. In contrast, type I (EC 3.1.21.3) and type III (EC 3.1.21.5) restriction endonucleases cleave at sites remote from their recognition site, which are not utilized for RFLP applications.

Restriction endonucleases are isolated from various bacteria. Each enzyme is named using a nomenclature system after the bacterium from which it was isolated. For example, in the restriction enzyme *Hae*III, H is from the genus name *Haemophilus*, ae is from the species name *aegypticus*, and III stands for the third endonuclease isolated from the *Haemophilus aegyptius* bacteria.

The restriction endonucleases play a role in protecting the bacteria from phage (bacterial virus) infections by using their endonucleases to destroy foreign DNA molecules. Bacterial DNA is usually methylated. In prokaryotes, the *Dam* methylase transfers a methyl group to the N6 position of the adenine in the sequence GmATC and the *Dcm* methylase transfers a methyl

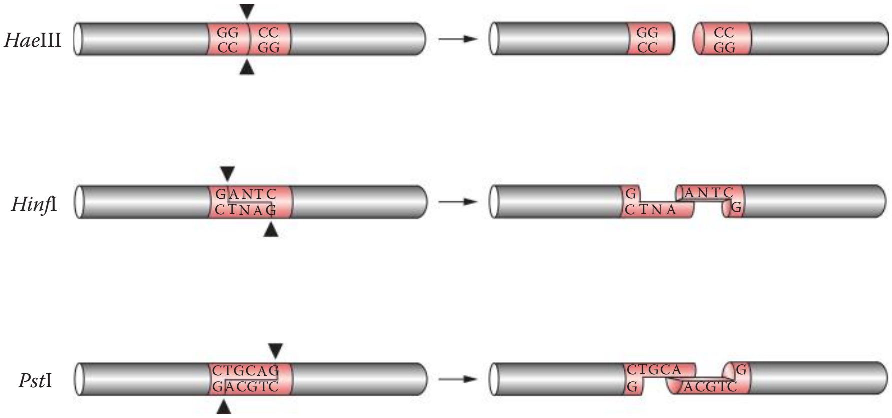

Figure 19.3 Restriction sites for *Hae*III, *Hinf*I, and *Pst*I. *Hae*III digestion produces a blunt end DNA fragment. *Hinf*I and *Pst*I digestions produce sticky ends. *N* represents any nucleotide. (© Richard C. Li.)

group to the C5 position of cytosine in the sequences CmCAGG and CmCTGG. The activities of restriction endonucleases can be influenced by DNA methylation. Many restriction endonucleases cannot cleave methylated DNA. Therefore, bacterial DNA is distinguishable from foreign DNA by the bacteria's restriction endonucleases. This phenomenon protects bacterial DNA from digestion by their own endonucleases. Note that because DNA methylation also occurs in the human genome (Chapter 11), it is important to choose the restriction endonucleases that are not affected by the methylation of human genomic DNA for RFLP analysis.

The type II restriction endonucleases were used in RFLP analysis for forensic DNA testing (Figure 19.3). In order to perform this analysis, the preferred restriction endonucleases for RFLP were those that cleave at the flanking regions of VNTR repeat units but not within the core repeat sequences of the VNTR. Most forensic laboratories used a single restriction endonuclease for a panel of VNTR loci because the DNA in evidence samples was often insufficient for performing multiple tests with different restriction endonucleases. For instance, *Hinf*I-based RFLP was commonly used in European forensic laboratories, and *Hae*III-based RFLP was used in North American and some European forensic laboratories. Other restriction endonuclease-based RFLPs such as *Pst*I were also used.

The use of these common restriction endonucleases allows the comparison of data of various laboratories. Several VNTR loci are suitable for RFLP analysis with these restriction endonucleases. For example, *Hae*III presents several advantages for forensic RFLP analysis. It recognizes a four-base sequence, 5′-GGCC-3′, and cleaves the DNA between the internal G and C residues of the recognition site (GG/CC). *Hinf*I recognizes a five-base restriction site, and *Pst*I recognizes a six-base restriction site. Hypothetically, four-base restriction sites are likely to occur more often than five- and six-base restriction sites in the human genome. Thus, *Hae*III restriction sites occur more frequently than *Hinf*I and *Pst*I sites. As a result, *Hae*III-cleaved DNA fragments are smaller than those of *Hinf*I and *Pst*I. The *Hae*III-generated VNTR allele sizes are easier to separate using conventional agarose gels, also called analytic gel electrophoresis. After electrophoresis, a smear of various sizes of DNA fragments can be observed. The analytic gel is then processed for Southern transfer. Moreover, the enzymatic activity of *Hae*III is not affected by the methylation of human genomic DNA. Its enzymatic activities also appear unaffected when a reaction proceeds under nonoptimal conditions. Additionally, low star activity is observed (Section 19.1.5.2.2).

19.1.2 Southern Transfer

Also known as *Southern blotting*, this technique was named after Sir Edwin Southern, who developed it in the United Kingdom in the mid-1970s. The method can be used to transfer

DNA from an agarose gel to a solid matrix so that it can be detected with a hybridization probe (Section 19.1.3). This method is still used today in many research laboratories. Prior to the transfer of DNA, the DNA in the gel must be denatured, under alkaline conditions such as treatment with sodium hydroxide, into single-stranded DNA. The single-stranded DNA in the gel is then transferred by capillary action to a solid matrix such as a piece of nylon membrane. The single-stranded DNA fragments transferred can be immobilized on a nylon membrane by an ultraviolet cross-linking process (Figure 19.4).

19.1.3 Hybridization with Probes

A hybridization probe of RFLP is a small segment of labeled DNA that is usually several hundred to a thousand bases in length containing the VNTR sequence. It is utilized to detect the presence of much longer target DNA sequences, in this case the VNTR sequences that are complementary to the nucleotide sequences of the probe. The probe is first denatured by heating or by exposure to alkaline conditions into single-stranded DNA. The hybridization process allows complementary pairing between the probe and the target sequence. Two types of probe techniques were developed for VNTR analysis: the multilocus probe and single-locus probe techniques.

19.1.3.1 Multilocus Probe Technique

The *multilocus probe* (*MLP*) technique can detect multiple VNTR loci simultaneously (Figure 19.5). Some VNTRs in the human genome share a short GC-rich core sequence of 10–15 bp. The MLP consists of this core sequence and hybridizes to multiple VNTRs that share these core sequences. As a result, the utilization of MLP produces a complex bar-code-like band pattern from alleles of multiple VNTR loci (Figure 19.6).

The MLP technique was pioneered by Sir Alec Jeffreys in 1984 at the University of Leicester in the United Kingdom and was called *DNA fingerprinting*. Because of its excellent discriminating power, the method was used for parentage testing in immigration disputes with great success.

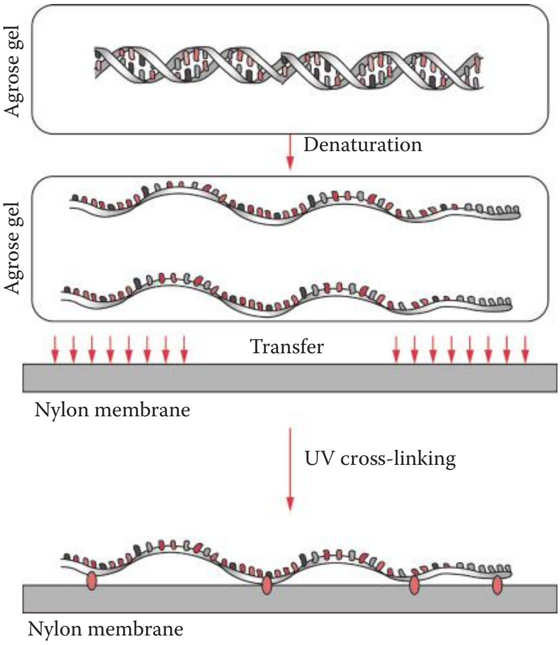

Figure 19.4 Southern blotting. DNA in agarose gel is denatured into single-stranded DNA and transferred to a solid-phase membrane where the single-stranded DNA is immobilized by ultraviolet cross-linking. (© Richard C. Li.)

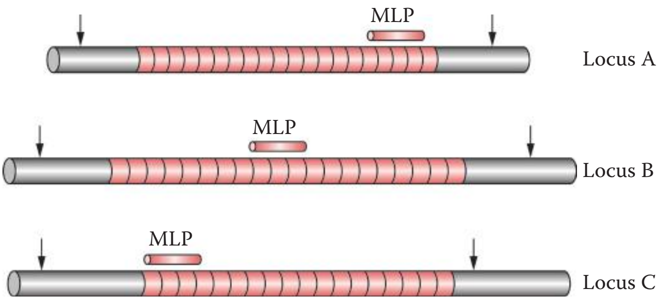

Figure 19.5 VNTR analysis using the MLP method. The technique can detect multiple VNTR loci simultaneously. Restriction sites are indicated by arrows. (© Richard C. Li.)

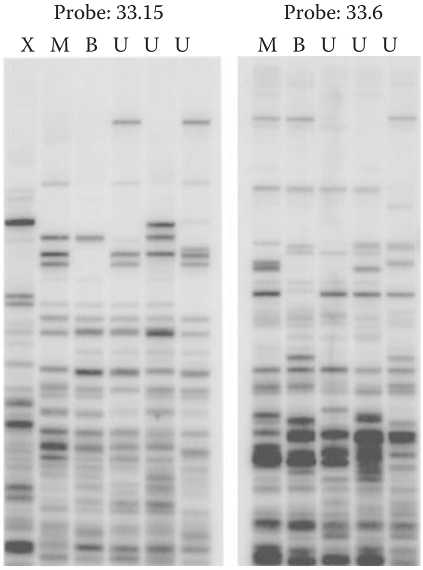

Figure 19.6 First application of DNA fingerprinting. The MLP method was used to analyze samples for an immigration case. M, mother; U, three undisputed children; B, male child in dispute; X, an unrelated individual. All bands in B can be traced back to M or U. (From Jeffreys, A.J., *Nat Med*, 11, 1035, 2005. With permission.)

However, one of the disadvantages of the MLP approach is that the interpretation of a mixed DNA sample from more than one individual is nearly impossible due to its complex DNA fingerprinting patterns. Therefore, MLP analysis was not widely utilized in forensic DNA laboratories.

19.1.3.2 Single-Locus Probe Technique

To resolve the disadvantages encountered in the MLP technique, probes that recognize the genomic DNA at the flanking regions of specific VNTR loci can be used. The probe only hybridizes to a single VNTR locus and the technique is called the *single-locus probe* (*SLP*), as depicted in Figure 19.7. SLP generates a simple pattern called a *DNA profile*, consisting of one band for a homozygote and two bands for a heterozygote per locus. In order to improve the discriminating power of the test, SLP analyses of different VNTR loci can be performed by using different probes sequentially with a single locus at one time. SLP can analyze mixed DNA samples from two or more contributors. The sizes of fragments can be estimated and converted into a numerical form suitable for databasing. Therefore, DNA profiles can be compared among different laboratories.

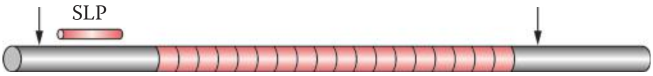

Figure 19.7 VNTR analysis using the SLP method. The technique can detect a single VNTR locus. Restriction sites are indicated by arrows. (© Richard C. Li.)

This technique led to the solving of a double murder case in Leicestershire in the 1980s. The case was the first to apply DNA evidence to a criminal investigation. DNA profiling identified the true perpetrator and also excluded an innocent suspect (Figure 19.8). In 1983 and 1986, two girls were raped and murdered. Crime scene evidence suggested that the two cases were committed by the same perpetrator. A young local man, Richard Buckland, was the suspect. However, the DNA evidence revealed that the semen samples from both crimes did not originate from Buckland. To solve the crimes, an investigation was carried out in which 5000 local men were asked to volunteer DNA samples for testing. Several months after the investigation, a witness tipped off police that a man named Pitchfork had paid someone for giving a blood sample as Pitchfork's. In 1987, Pitchfork was arrested. It was discovered that Pitchfork's DNA profile matched that of the crime scene evidence. He was sentenced to life imprisonment. This case demonstrated the great potential of DNA profiling in forensic investigations. Consequently, SLP became a common method in most forensic laboratories in the late 1980s–1990s.

19.1.4 Detection

To detect VNTR loci, a labeled SLP probe is hybridized to the target sequence of DNA, which has been immobilized on a solid matrix such as a piece of nylon membrane (see Southern

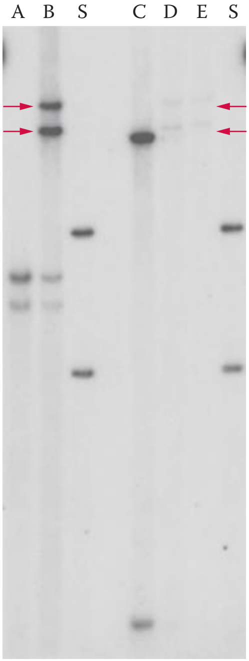

Figure 19.8 First application of DNA profiling in a criminal investigation using the SLP method. A, hair roots from the first victim; B, a mixture of semen and vaginal fluid from the first victim; C, blood from second victim; D, a vaginal swab from the second victim; E, a semen stain on clothing from the second victim; S, blood from the suspect. Alleles (arrows) are matched with the profiles of the two cases but not with the suspect profile. (From Jeffreys, A.J., *Nat Med*, 11, 1035, 2005. With permission.)

transfer, Section 19.1.2). Any unbound probes are washed away so that they do not interfere with the signal. Two types of detection systems are used for VNTR analysis. Radioisotope labeling, such as with a ^{32}P-labeled probe, can be used. The hybridized probe can be detected by exposing the membrane to a sheet of x-ray film to generate an autoradiograph. Alternatively, an enzyme-conjugated probe can also be used. Alkaline phosphatase (Section 9.2.2) is an example of an enzyme used in this type of probe. Its enzymatic activity can be detected with chemiluminescent substrates. A chemiluminescent signal can be detected by exposure to x-ray film as well.

In RFLP analysis, several loci are commonly analyzed sequentially using the same membrane. This approach has the advantage of not consuming additional DNA samples, which are often limited in forensic cases. When probing for multiple loci, multiple probes for each locus are sequentially hybridized and removed one at a time. Once the analysis of the first probe is completed, the probe is removed by a procedure called probe stripping, which is carried out under conditions such as high-temperature washing to denature the DNA strands (Chapter 17). The probe for the next locus to be analyzed is then hybridized to the same membrane, and the process is repeated for each probe to be tested.

Typically, a size standard is utilized on each gel. Band sizes can thus be estimated by comparison to these standards. However, the VNTR alleles that differ by only one or two repeat units are usually not distinguishable. For this reason, genotypes can be determined by bins but not discrete alleles. A *bin* is a range of DNA fragments that differ by only a few repeat units. A sample with a known VNTR genotype is also utilized on each gel as a positive control where historically a genomic DNA sample from cell line K562 (a human erythroleukemic cell line) was used for the positive control. DNA samples to be compared can be loaded side by side on the same gel. As a result, the patterns of VNTR fragments can be compared from sample to sample. The following possible conclusions can be made. If the VNTR fragments are at corresponding positions (profiles match), they are considered to be a match (inclusion). Chapter 25 evaluates and discusses the strengths of the results. If the profiles are different, the two DNA samples are considered to have come from different origins (exclusion).

19.1.5 Factors Affecting RFLP Results

The accuracy of VNTR profiling results can be affected by certain factors such as sample conditions, genetic mutations, and experimental artifacts appearing during the procedure. Consequently, these factors can impact data interpretation and are explained in the following sections.

19.1.5.1 DNA Degradation

RFLP analysis requires the genomic DNA to be intact. DNA degradation results in damage such as creating nicks and breaks in the strand. The more severe the degradation, the smaller the average size of the DNA fragments. When the average size of DNA fragments becomes too small, the allele may not be detected. Many VNTR tandem arrays can span several kilobases in length. In theory, large alleles are more likely to be affected by degradation than smaller alleles at a different locus.

A two-banded heterozygous profile can be observed as a one-banded homozygous profile if the larger band is not detected due to degradation. This artifact could lead to a false determination of exclusion. However, DNA degradation can be detected prior to conducting RFLP by the use of agarose gel electrophoresis, also known as a *yield gel*, used for evaluating the yield and integrity of the isolated genomic DNA. High-molecular-weight genomic DNA bands are usually observed for a typical genomic DNA sample. In contrast, a smear of low-molecular-weight DNA bands can be observed if DNA is degraded. The sizes of the DNA can be estimated by comparison to a size standard run on the same gel. Additionally, the yield of DNA can be estimated by comparing the intensity of the size standards with a known quantity of DNA.

19.1.5.2 Restriction Digestion–Related Artifacts

19.1.5.2.1 Partial Restriction Digestion

Complete restriction digestion should be achieved for RFLP analysis. If partial digestion occurs, the partially cleaved DNA strands are longer than the cleaved fragments (Figure 19.9). Thus, partial digestion results in a mixture of fragments with correct sizes and slightly larger fragments. Under these conditions, a larger uncleaved band, usually lower in intensity than the true bands, can be observed. The multibanded pattern due to the partial digestion can be observed at multiple loci analyzed in the same nylon membrane.

Detection of more than two bands in an RFLP profile may lead to a false interpretation and be incorrectly concluded to be a mixture. However, partial digestion can be detected after restriction digestion. DNA cleavage by restriction endonuclease digestion can be examined using agarose gel electrophoresis. A small portion of a sample can be analyzed. After separation using electrophoresis, a smear of various sizes of cleaved DNA fragments can usually be observed if restriction digestion is completed. Conversely, high-molecular-weight genomic DNA can still be observed from partially digested DNA samples. Additionally, comparisons can be made between the sample and uncleaved and completely cleaved standard samples of DNA. This quick assay is also called a *test gel*, used to determine if the DNA was cleaved to completion.

If partial digestion occurs, procedures such as additional purification of the DNA sample can be carried out. Additionally, optimal amounts of DNA, restriction enzymes, buffer, and proper incubation conditions should be used in achieving complete digestion.

19.1.5.2.2 Star Activity

Star activity refers to a deviation of the specificity of a cleavage site of a restriction endonuclease under certain conditions, such as a high concentration ratio of enzyme to DNA, the use of non-optimal buffers for restriction digestion, prolonged digestion time, the substitution of Mg^{2+} with other divalent cations, and the presence of organic solvents such as ethanol and concentrated glycerol. For instance, *Hae*III cleaves at the GGCC DNA sequence. When star activity occurs, the enzymatic specificity is reduced and cleaves at a sequence slightly different from GGCC. If the start site is presented at an internal location of a VNTR locus, the enzyme would cleave the GGCC sequences and additionally cleave at the internal star site (Figure 19.10). This would result in an additional band smaller than the true alleles, although the intensity of this band is usually not the same as other bands. Thus, a multiband pattern is observed. However, the star

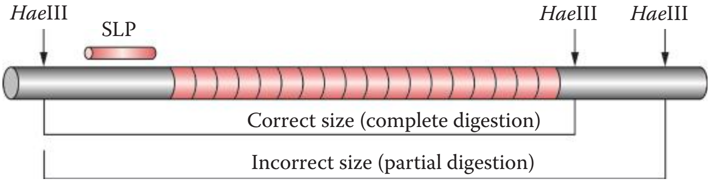
*Hae*III
SLP
*Hae*III
*Hae*III
Correct size (complete digestion)
Incorrect size (partial digestion)

Figure 19.9 Effects of partial restriction digestion on the RFLP profile. Only the restriction fragments detectable by the probe are shown. *Hae*III restriction sites are indicated by arrows. (© Richard C. Li.)

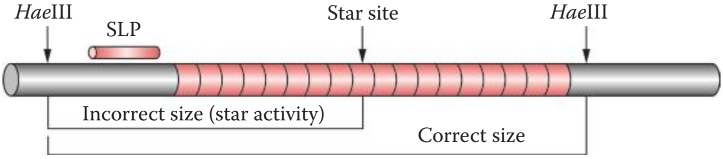

Figure 19.10 Star effects on RFLP profiles. Only the restriction fragments detectable by the probe are shown. *Hae*III restriction sites and star sites are indicated by arrows. (© Richard C. Li.)

site presented externally to a VNTR locus cannot be detected by the probe and does not affect the profiling results. Star activity can usually be avoided by carrying out restriction digestion reactions under conditions recommended by their manufacturers.

19.1.5.2.3 Point Mutations

A *point mutation* is caused by the substitution, deletion, or insertion of a single nucleotide. Point mutations at a restriction site within flanking regions may abolish the site, and the result is a band slightly larger than the true allele (Figure 19.11). The point mutation may also be present internally in a VNTR sequence. If such a point mutation creates a restriction enzyme site, the enzyme will cleave at the regular site and at the mutation site and yield two smaller bands. If the created restriction site is located internal to the probe binding region, both bands will be detected for that allele. These rare mutations obey Mendelian inheritance.

19.1.5.3 Electrophoresis and Blotting Artifacts

19.1.5.3.1 Partial Stripping

If more than one VNTR locus is analyzed sequentially using the same membrane, a probe must be removed by the stripping process before the application of the next probe. Any probe remaining on the membrane due to partial stripping may generate additional bands when the next probe is analyzed. However, bands due to partial stripping are usually faint and have the same electrophoretic mobility as the previous autoradiograph.

19.1.5.3.2 Separation Resolution Limits and Band Shifting

Agarose gel electrophoresis cannot resolve restriction fragments that differ by one or a few repeat units, especially for high-molecular-weight fragments. These bands may not be

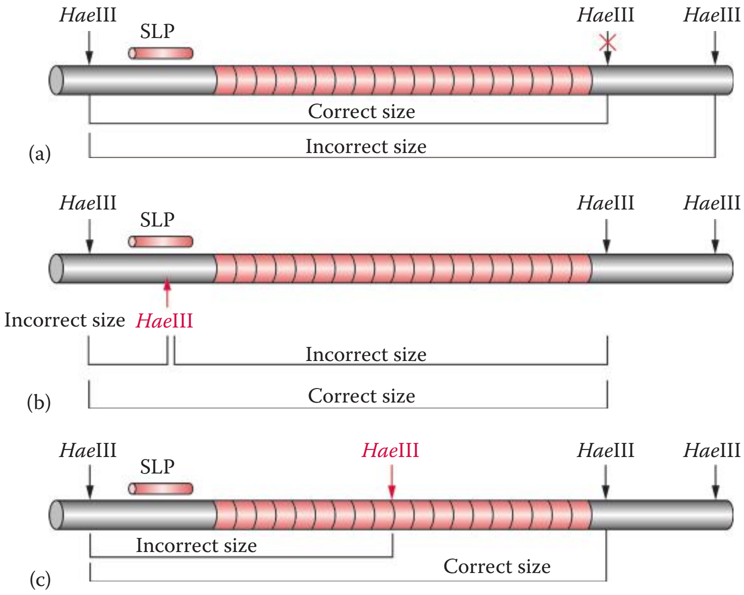

Figure 19.11 Effects of point mutations on the RFLP profile. Only the restriction fragments detectable by the probe are shown. *Hae*III restriction sites are indicated by arrows. (a) Point mutation (in red) abolishes the *Hae*III restriction site. (b) A point mutation (in red) creates an internal *Hae*III restriction site residing within the probe-binding region. (c) A point mutation (in red) creates an internal *Hae*III restriction site residing outside the probe-binding region. (© Richard C. Li.)

separated and will appear as a single band. This may lead to a false interpretation as a homozygous profile. Additionally, minor variations in the electrophoretic mobility of DNA fragments, known as band shifting, can cause two samples from the same individual to appear different.

19.1.5.3.3 Bands Running off Gel

The commonly used VNTR loci generate bands from hundreds of base pairs to 20 kb in length. The small bands have higher electrophoretic mobility and may run off the front edge of a gel during electrophoresis and fail to be detected. This phenomenon may also lead to a false interpretation as a homozygous profile. To prevent these DNA fragments from running off the gel into the buffer tank, a longer gel can be used. Alternatively, the electrophoresis can be stopped before the dye front, the furthest extent that dyes migrate, reaches the front edge of the gel.

19.2 Amplified Fragment Length Polymorphism

The RFLP analysis of VNTR profiling does not perform well for degraded or limited quantities of DNA from crime scene samples. For these reasons, an improved VNTR method was developed. Some VNTR loci have relatively short alleles (<1 kb). These loci are suitable for PCR amplification. This technique is called *amplified fragment length polymorphism* (*AFLP*). One locus, D1S80, was used by forensic DNA laboratories for AFLP analysis. Fragments in the range of 14–42 repeat units (16 bp per repeat) were amplified using the AFLP method (Figure 19.12). The amplified DNA fragments were commonly separated according to size using polyacrylamide gel electrophoresis and detected using a silver stain (Figure 19.13).

D1S80 loci are detected as discrete alleles and thus can be compared directly to an allelic ladder (a collection of common alleles used as a standard) on the same gel. This technique represented an improvement over the RFLP system. RFLP allele sizing cannot be performed with precision and the resolution limits of agarose gel electrophoresis are much lower than those of the polyacrylamide gels.

The AFLP technique requires less DNA than the RFLP method and performs better for degraded samples. The AFLP method at the D1S80 locus can be analyzed in a multiplex fashion with an amelogenin locus (Chapter 21). The amelogenin gene is used for forensic sex-typing applications. Typing the amelogenin gene enables the determination of the sex of the contributor of a biological sample.

Due to the wide variation in allele sizes at the D1S80 locus, preferential amplification may be observed. Under certain conditions, the larger alleles may not be as consistently amplified as the small alleles, which may cause lower signal intensity of the larger allele. Additionally, only one locus is analyzed in this system. Furthermore, the D1S80 locus contains two alleles that are very common in some populations. Thus, the discriminating power is reduced compared to RFLP. D1S80 was gradually replaced by multiplex STR systems in the late 1990s.

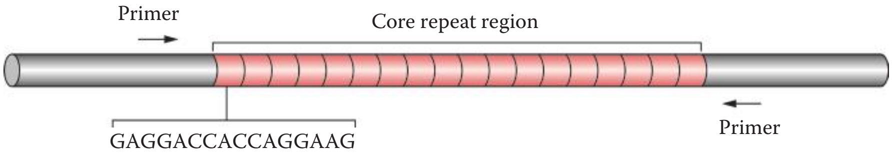

Figure 19.12 VNTR locus D1S80 (chromosome 1p). Each repeat unit is 16 bp long. PCR primers are indicated to amplify the core repeat region.

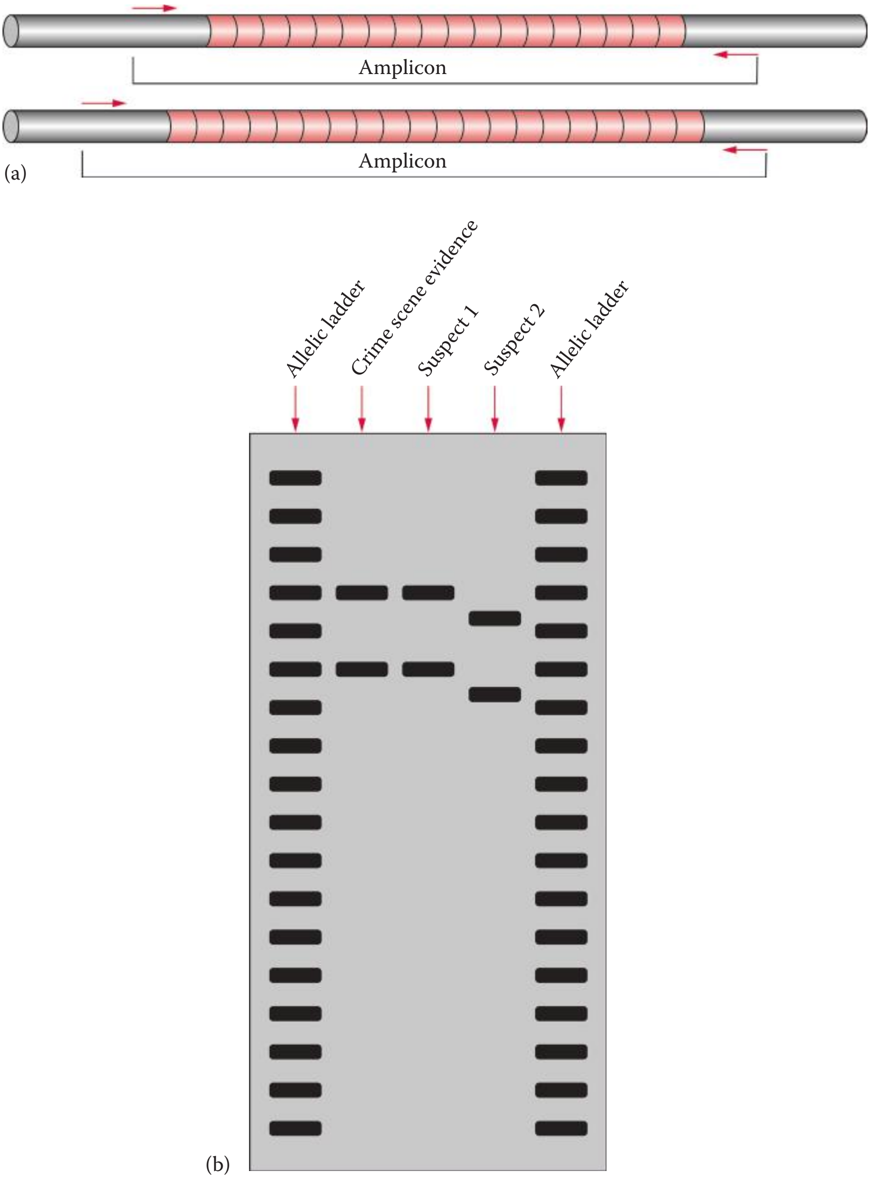
Amplicon
Amplicon
(a)
Allelic ladder
Crime scene evidence
Suspect 1
Suspect 2
Allelic ladder
(b)

Figure 19.13 AFLP analysis. (a) Heterozygous D1S80 loci amplified by PCR. PCR primers are indicated as *arrows*. (b) Silver-stained polyacrylamide gel showing D1S80 amplicons along with allelic ladders. (© Richard C. Li.)

Bibliography

RFLP

Balazs, I., et al., Human population genetic studies of five hypervariable DNA loci. *Am J Hum Genet*, 1989, **44**(2): 182–190.

Benzinger, E.A., et al., Time course and inhibitors of *Hae*III digestion in the forensic laboratory. *Appl Theor Electrophor*, 1995, **4**(4): 179–188.

Benzinger, E.A., et al., Products of partial digestion with *Hae*III. Part 1. Characterization, casework experience and confirmation of the theory of three-, four- and five-banded RFLP pattern origins using partial digestion. *J Forensic Sci*, 1997, **42**(5): 850–863.

20

Autosomal Short Tandem Repeat Profiling

A *short tandem repeat* (STR) is a region of genomic DNA containing an array of short repeating sequences. STRs are also called *microsatellites* or *simple sequence repeats*. A STR repeat unit can be several base pairs (bp) in length. Arrays range from several to approximately a hundred *repeat units*, which are the component of repetition. The number of STR repeat units varies among individuals. The most commonly used STR loci are 100–500 bp in length, which are shorter than the smallest variable number tandem repeats (VNTRs) (approximately 1000 bp). Thus, STR loci have many advantages compared to VNTR loci:

- STR loci can be amplified by PCR.
- STR profiling can be carried out for degraded DNA samples.
- Preferential amplification is reduced at STR loci.
- The resolution of electrophoretic separation of STR fragments is superior.
- STR loci are suitable for multiplex amplification.

Additionally, STR profiling, as with VNTR profiling, is suitable for the interpretation of mixed DNA profiles from multiple individuals. Thus, STR loci are better candidates for forensic DNA testing than VNTR loci. This chapter will discuss autosomal STR profiling. Male-specific Y chromosomal STR will be discussed in Chapter 21.

20.1 Characteristics of STR Loci

More than 10^5 STRs exist in the human genome. Many STRs have been characterized and used in various types of studies such as genetic mapping and linkage analysis. Some STRs have been characterized specifically for forensic DNA profiling.

20.1.1 Core Repeat and Flanking Regions

The *core repeat region* of each STR locus contains tandemly repeated sequences. The designation of genotypes for human identification is based on the number of tandem repeat units at a STR locus, which varies among human individuals (Figure 20.1). The *flanking regions* surrounding the core repeat region are also needed for STR analysis. PCR primers complementary to these flanking regions are used, allowing the core repeat regions to be amplified.

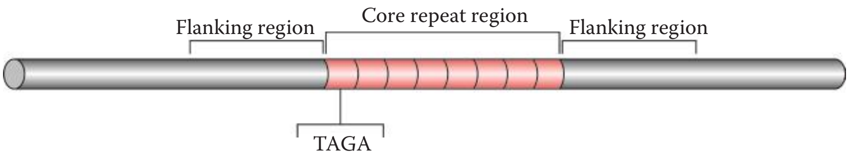

Figure 20.1 Core repeat and flanking regions of CSF1PO STR locus. It consists of eight repeating units of tetrameric nucleotides (TAGA); thus, it is designated as allele 8. (© Richard C. Li.)

20.1.2 Repeat Unit Length

Repeat unit length is the number of nucleotides in a single repeat unit. Dimeric, trimeric, tetrameric, pentameric, and hexameric repeat units appear in the human genome. For example, dimeric and trimeric repeats are very abundant, but they are not usually used for forensic applications. High frequencies of stutter peaks (Section 20.4.2.1) that interfere with genotype interpretation are observed when dimeric and trimeric repeats are amplified. On the contrary, only a few thousand pentameric repeats and a few hundred hexameric repeats exist in the human genome. The pentameric and hexameric repeats are very polymorphic. Only a few pentameric and hexameric repeats are used for forensic applications because they are less abundant in the human genome. The human genome has at least 10^4 tetrameric repeats representing approximately 9% of the total STRs. The STRs with tetrameric repeats are very polymorphic. When they are amplified by PCR, this category of STRs exhibits fewer frequencies of stutter peaks than STRs with dimeric and trimeric repeats. Therefore, the most commonly used STR loci for forensic DNA profiling are the STRs with tetrameric repeats.

20.1.3 Repeat Unit Sequences

STR loci compatible for forensic use can be divided into several classes based on their repeat unit sequences. Figure 20.2 shows representative examples of core repeat sequences. *Simple repeats* consist of tandem repeats with identical repeat unit sequences (Figure 20.2a). Allele designation is based on the number of repeat units in the core repeat region. For example, a D5S818 allele consisting of ten repeating units of the tetrameric nucleotide sequence AGAT is designated as allele 10. *Compound repeats* consist of more than one type of simple repeat (Figure 20.2b). *Complex repeats* contain several clusters of different tandem repeats with intervening sequences (Figure 20.2c).

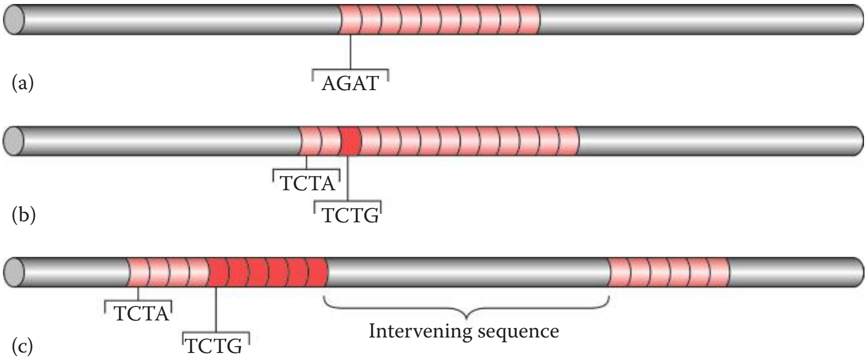

Figure 20.2 Examples of core repeat sequences. (a) A simple repeat in which D5S818 $[AGAT]_{10}$ is designated as allele 10, consisting of 10 repeating units of the tetrameric nucleotides, AGAT. (b) Compound repeats. Allele 14 of D8S1179 consists of two types of repeating units: $[TCTA]_2$, $[TCTG]_1$, and $[TCTA]_{11}$. (c) Complex repeats. Allele 24 of D21S11 contains several clusters of different tandem repeats, $[TCTA]_4$, $[TCTG]_6$, and $[TCTA]_6$, with a 43 bp intervening sequence: $[TCTA]_3$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA. (© Richard C. Li.)

Nonconsensus alleles with partial repeat units also appear in the population. These nonconsensus alleles, also known as *microvariants*, differ from common alleles by one or more nucleotides. They are designated by the number of consensus repeats, followed by a decimal point and the number of nucleotides of the partial repeat, for example, the TH01 allele 9.3 is 1 nucleotide shorter than allele 10.

Another type of nonconsensus allele can result from a limitation of STR analysis. These alleles have the same number of tandem repeats as common alleles but contain different sequences. These microvariants cannot be distinguished by STR profiling because their length is identical to the lengths of common alleles.

20.2 STR Loci Commonly Used for Forensic DNA Profiling

In the early 1990s, STR loci were initially utilized for genetic studies and were later applied to forensic DNA profiling. The first STR multiplex system, known as the quadruplex, was developed by Forensic Science Services in the United Kingdom. It consisted of four STR loci (F13A1, FES, TH01, and VWA) with a *population match probability* (P_m) of 10^{-4} (Figure 20.3). P_m measures the discriminating power of an STR locus used for forensic DNA analysis. The lower the P_m (i.e., the higher the discriminating power), the less likely a match will occur between two randomly chosen profiles from different individuals (Chapter 25). In 1995, the first national DNA database was established in the United Kingdom. It contained six STR loci, also known as the *second-generation multiplex* (SGM), consisting of D8S1179, D18S51, D21S11, FGA, TH01, and VWA, with a P_m value of 10^{-7}. The SGM system also included the amelogenin locus (Section 21.3.1) for determining the sexes of DNA contributors. Subsequently, four additional loci were added to SGM with a P_m as low as 10^{-13} (SGM Plus).

To allow for international data exchange, the European DNA Profiling Group (EDNAP) recommended the use of TH01 and VWA loci for all participating European laboratories in 1996. In 1998, the *European Standard Set* (ESS) of loci was established and included TH01, VWA, FGA, and D21S11 for forensic use in Europe. Thereafter, D3S1358, D8S1179, and D18S51 loci were added to the ESS. Other loci are used as well, such as SE33, which is used in Germany's database. In 1998, the US Federal Bureau of Investigation established the *Combined DNA Index System* (CODIS). It contains 13 core STR loci plus the amelogenin sex-typing locus with a P_m

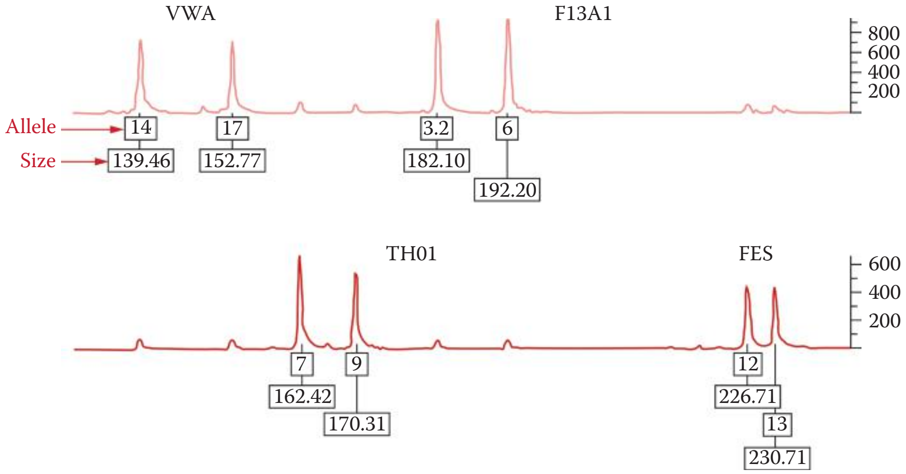

Figure 20.3 DNA profile obtained using the first STR multiplex system: the quadruplex. F13A1, FES, TH01, and VWA loci are shown. (© Richard C. Li.)

Table 20.1 Common STR Loci

Locus Symbol	Repeat Motif	Repeat Category	Cytogenetic Map Location on Chromosome	Distance from pter (Mb)	Structural Gene
CSFIPO	TAGA	Simple	5q33.1	149.4	Intron 6 of c-fms protooncogene
FGA	[CTTT] [CCTT]	Compound	4q31.3	155.5	Intron 3 of fibrinogen α chain gene
TH01	AATG	Simple	11p15.5	2.2	Intron 1 of tyrosine hydroxylase gene
TPOX	TGAA	Simple	2p25.3	1.5	Intron 10 of thyroid peroxidase gene
SE33/ ACTBP2	AAAG	Complex	6q14	89	5′ Flanking sequence of β-actin-related pseudogene 2 gene
VWA	[TCTG] [TCTA]	Compound	12p13.31	6.1	Intron 40 of von Willebrand factor gene
D1S1656	[TAGA] [TAGC]	Compound	1q42.2	230.9	Anonymous
D2S441	[TCTA] [TCAA]	Compound	2p14	68.2	Anonymous
D2S1338	[TGCC] [TTCC]	Compound	2q35	218.9	Anonymous
D3S1358	[TCTG] [TCTA]	Compound	3p21.31	45.6	Anonymous
D5S818	AGAT	Simple	5q23.2	123.1	Anonymous
D7S820	GATA	Simple	7q21.11	83.8	Anonymous
D8S1179	[TCTA] [TCTG]	Compound	8q24.13	125.9	Anonymous
D10S1248	GGAA	Simple	10q26	–[a]	Anonymous
D12S391	[AGAT] [AGAC]	Compound	12p13.2	12.5	Anonymous
D13S317	TATC	Simple	13q31.1	82.7	Anonymous
D16S539	GATA	Simple	16q24.1	86.4	Anonymous
D18S51	AGAA	Simple	18q21.33	60.9	Anonymous
D19S433	[AAGG] [TAGG]	Compound	19q12	30.4	Anonymous
D21S11	[TCTA] [TCTG]	Complex	21q21.1	20.6	Anonymous
D22S1045	ATT	Simple	22q12.3	37.5	Anonymous

Source: Ensembl Homo sapiens version 75.37 (GRCh37). Mb, megabase.

[a] Locus is not mapped to the assembly in the current Ensembl database.

of 10^{-15}. As the database grows rapidly, the chances of finding incidental matches among DNA profiles are increasing. Recently, additional loci have been added to CODIS and ESS core loci to reduce the likelihood of adventitious matches, as well as to facilitate international data sharing among law enforcement agencies and to improve the discrimination power of forensic STR analysis.

Commonly used STR loci characterized for forensic DNA profiling are summarized in Tables 20.1 and 20.2. To achieve low P_m in forensic STR profiling, desired STR loci should possess certain characteristics as described below. First, the alleles of STR loci selected should be highly variable among individuals. Second, if more than one locus is selected, the loci should not be linked to each other or inherited together (Section 21.2). The STR loci utilized are usually located at different chromosomes to ensure that they are not linked. However, loci that are far

TABLE 20.2 Core STR Loci

Locus	SGM	SGM Plus	ESS	ESS-Extended	CODIS	CODIS-Extended
Amel	☑	☑	☑	☑	☑	☑
CSFIPO					☑	☑
D1S1656				☑		☑
D2S441				☑		☑
D2S1338		☑				☑
D3S1358		☑	☑	☑	☑	☑
D5S818					☑	☑
D7S820					☑	☑
D8S1179	☑	☑	☑	☑	☑	☑
D10S1248				☑		☑
D12S391				☑		☑
D13S317					☑	☑
D16S539		☑			☑	☑
D18S51	☑	☑	☑	☑	☑	☑
D19S433		☑				☑
D21S11	☑	☑	☑	☑	☑	☑
D22S1045				☑		2
DYS391						☑
FGA	☑	☑	☑	☑	☑	☑
THO1	☑	☑	☑	☑	☑	☑
TPOX					☑	1
VWA	☑	☑	☑	☑	☑	☑
SE33						3

Source: Hares, D.R., *Forensic Sci Int Genet*, 6, e135, 2012; Gill, P., et al., *Forensic Sci Int*, 156, 242–244, 2006.

☑, minimum required loci; 1–3, recommended loci in ranked order of preference.

enough apart on the same chromosome can still be used (Figures 20.4 and 20.5), since they are not linked. Additionally, STR loci with fewer amplification artifacts such as stutter products are desired. Stutters can complicate the interpretation of profiles derived from a mixed DNA sample from more than one contributor. Moreover, STR loci with short amplicon (amplified product) lengths are preferred for multiplex STR analysis and the testing of degraded DNA samples.

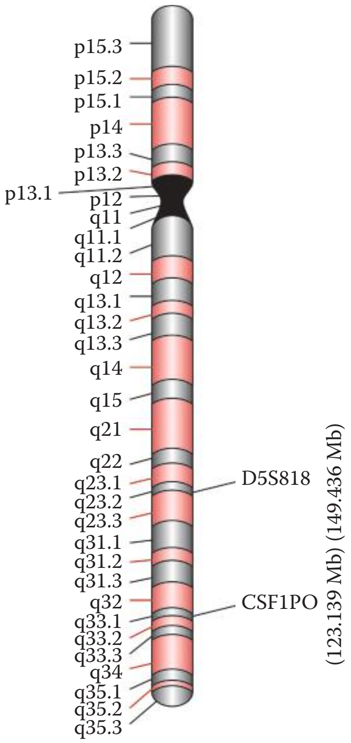
p15.3
p15.2
p15.1
p14
p13.3
p13.2
p13.1
p12
q11
q11.1
q11.2
q12
q13.1
q13.2
q13.3
q14
q15
q21
q22
q23.1
q23.2
q23.3
q31.1
q31.2
q31.3
q32
q33.1
q33.2
q33.3
q34
q35.1
q35.2
q35.3
D5S818
CSF1PO
(123.139 Mb) (149.436 Mb)

Figure 20.4 Cytogenetic map showing the locations of STR markers on chromosome 5. CSF1PO and D5S818 are separated by 26 Mb (megabases). (© Richard C. Li.)

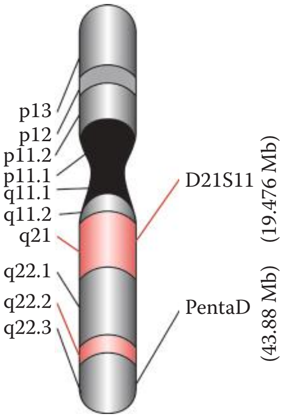
p13
p12
p11.2
p11.1
q11.1
q11.2
q21
q22.1
q22.2
q22.3
D21S11
PentaD
(43.88 Mb) (19.476 Mb)

Figure 20.5 Cytogenetic map showing the locations of STR markers on chromosome 21. D21S11 and PentaD are separated by 24 Mb (megabases). (© Richard C. Li.)

20.3 Forensic STR Analysis

STR loci are amplified using fluorescent dye-labeled primers. A multiplex STR system utilizes multiple fluorescent dyes to label each amplicon. The amplicons are separated via electrophoresis. The different fluorescent dye colors are resolved by the detector, and the signals corresponding to each DNA fragment are identified using specialized computer software. The data collection process generates an *electropherogram* that shows a profile of peaks corresponding to each DNA fragment. The positions of these peaks represent the electrophoretic mobility of the DNA fragments. A small fragment, migrating faster than a large one, peaks earlier in the electropherogram than a longer fragment. The DNA fragments are sized by comparison to an internal size standard (Figure 20.6; Chapter 8). Figure 20.7 summarizes the work flow of a forensic STR analysis.

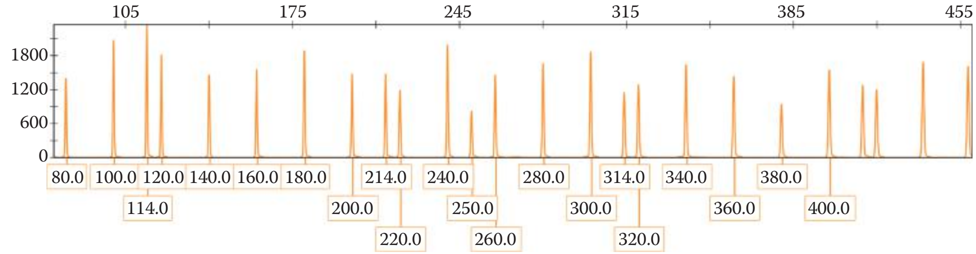

Figure 20.6 Electropherogram of GeneScan™ 500 size standard (Applied Biosystems). RFU represents relative fluorescence unit. (© Richard C. Li.)

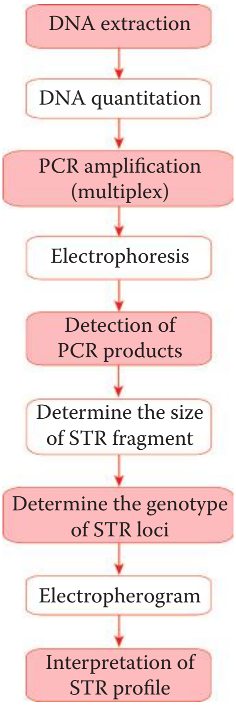
DNA extraction
DNA quantitation
PCR amplification (multiplex)
Electrophoresis
Detection of PCR products
Determine the size of STR fragment
Determine the genotype of STR loci
Electropherogram
Interpretation of STR profile

Figure 20.7 Capillary electrophoresis separation of amplified STR products. Fluorescent dye-labeled amplification products are separated and subsequently detected. Various fluorescent dye colors are resolved by the detector. The peaks corresponding to each DNA fragment are identified. (© Richard C. Li.)

The area or amplitude of the peak, expressed as relative fluorescence units (RFU) (Section 9.3.1.3), reflects the fluorescent signal intensity. The RFU value of peak height is proportional to the amount of DNA amplicons being analyzed. When the RFU value is very low, it is difficult to distinguish a signal from background noise. The manufacturer of commonly used instruments for forensic DNA analysis recommends 150 RFU as the threshold of detection. Peaks below 150 RFU should be interpreted with caution. Some forensic laboratories use lower thresholds, as low as 50 RFU, based on their own validation studies. In contrast, when the RFU value is too high, it saturates the sensitivity of an instrument as well as causing artifacts such as pull-up signals (Section 20.4.3.1). The maximum RFU allowed is usually 6000 RFU.

20.3.1 Determining the Genotypes of STR Fragments

As noted earlier, electropherograms are usually plotted as fluorescent signal intensity (in RFU) versus the sizes of the DNA fragments. The data in an electropherogram can then be converted into a genotype. The genotype for a specific STR locus is defined as the number of repeat units of the allele. STR genotype data generated from different laboratories can be compared easily and are suitable for databasing.

The genotype is determined by using an allelic ladder, which is important to achieve accurate genotype profiling. An allelic ladder is a collection of synthetic fragments corresponding to common alleles observed in the human population for a given set of STR loci (Figures 20.8 and 20.9). The ladders are compared to data obtained from an electropherogram of a questioned sample to determine the genotype. Thus, each allele in a ladder must be resolved properly in order to determine correct STR alleles for a sample. The sizes of DNA fragments of a sample are correlated to sizes of fragments for each allele in an allelic ladder in order to determine the allele designation (genotype) of a questioned sample (Figures 20.10 and 20.11). If a rare allele fails to match alleles within an allelic ladder, it is considered an *off-ladder allele*. If an off-ladder allele is present, the sample should be reanalyzed. The presence of a rare allele can be confirmed by repeating the electrophoresis process based on the characteristic electrophoretic mobility of the rare allele.

20.3.2 Interpretation of STR Profiling Results

General guidelines for the interpretation and the reporting of STR profile results were set by the Scientific Working Group on DNA Analysis Methods (SWGDAM) and the DNA Commission of the International Society of Forensic Genetics (ISFG). Typically, conclusions are categorized as inclusion, exclusion, or inconclusive result.

20.3.2.1 Inclusion (Match)

Peaks of compared STR loci, such as those between the profiles of suspect and crime scene evidence or victim and crime scene evidence, show identical genotypes. The strength of this conclusion can be evaluated via statistical analysis and is usually cited in the case report (Chapter 25).

20.3.2.2 Exclusion

The genotypes of two or more samples differ, and the profile of the sample is determined to be an exclusion, meaning that the profiles originated from different sources.

20.3.2.3 Inconclusive Result

The data do not support a conclusion of inclusion or exclusion. In other words, insufficient information is available to reach a conclusion.

20.4 Factors Affecting Genotyping Results

A number of genetics-, amplification-, and electrophoresis-related factors may affect the accuracy of genotypic profiles.

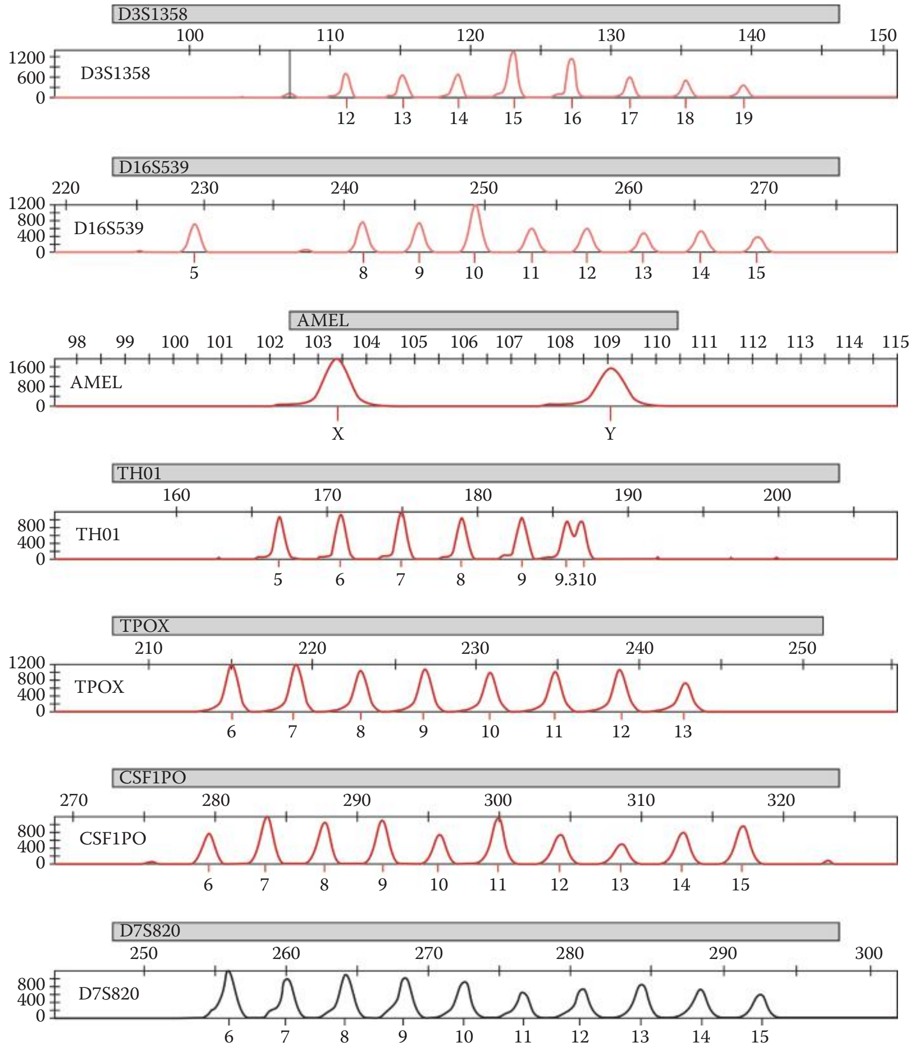

Figure 20.8 Electropherogram of the allelic ladder of the AmpFlSTR® COfiler® PCR Amplification Kit (Applied Biosystems). AMEL, CSF1PO, D3S1358, D7S820, D16S539, TH01, and TPOX loci are shown. (© Richard C. Li.)

20.4.1 Mutations

STR loci with low mutation frequencies are desired, in particular, for human identification in mass disasters and for missing person and paternity cases. However, STR mutations do occur, which can affect the profiling results.

20.4.1.1 Mutations at STR Core Repeat Regions

Mutations, usually resulting in a gain or a loss of a single repeat unit, are observed at STR loci. If a mutation occurs in the germ cells (cells that form gametes), the mutant allele will be transmitted to and be present in all cell types of the progeny. This type of inheritable mutation in germ cell lineage is called a *germ-line mutation*. The frequency of germ-line mutation can be measured by the *mutation rate*, expressed as the number of mutations per generation (germ-line transmission). The average mutation rate of commonly used STR loci is about 10^{-4} mutations per germ-line transmission. However, the mutation rate may vary among different STR loci.

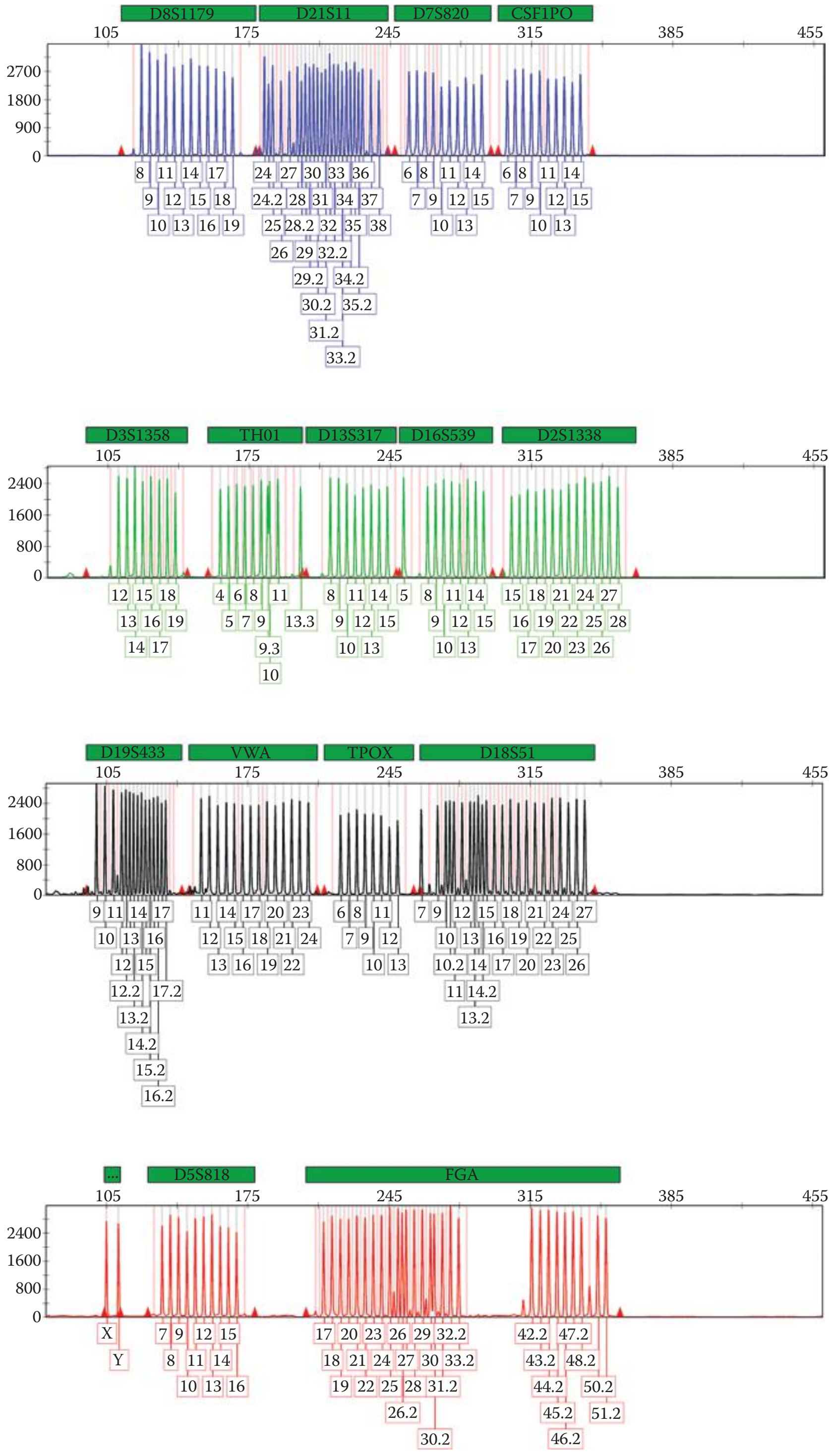

Figure 20.9 Electropherogram of the allelic ladder of the AmpFlSTR® Identifiler® Plus Kit (Applied Biosystems). AMEL, CSF1PO, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, TH01, TPOX, and VWA loci are shown. (© Richard C. Li.)

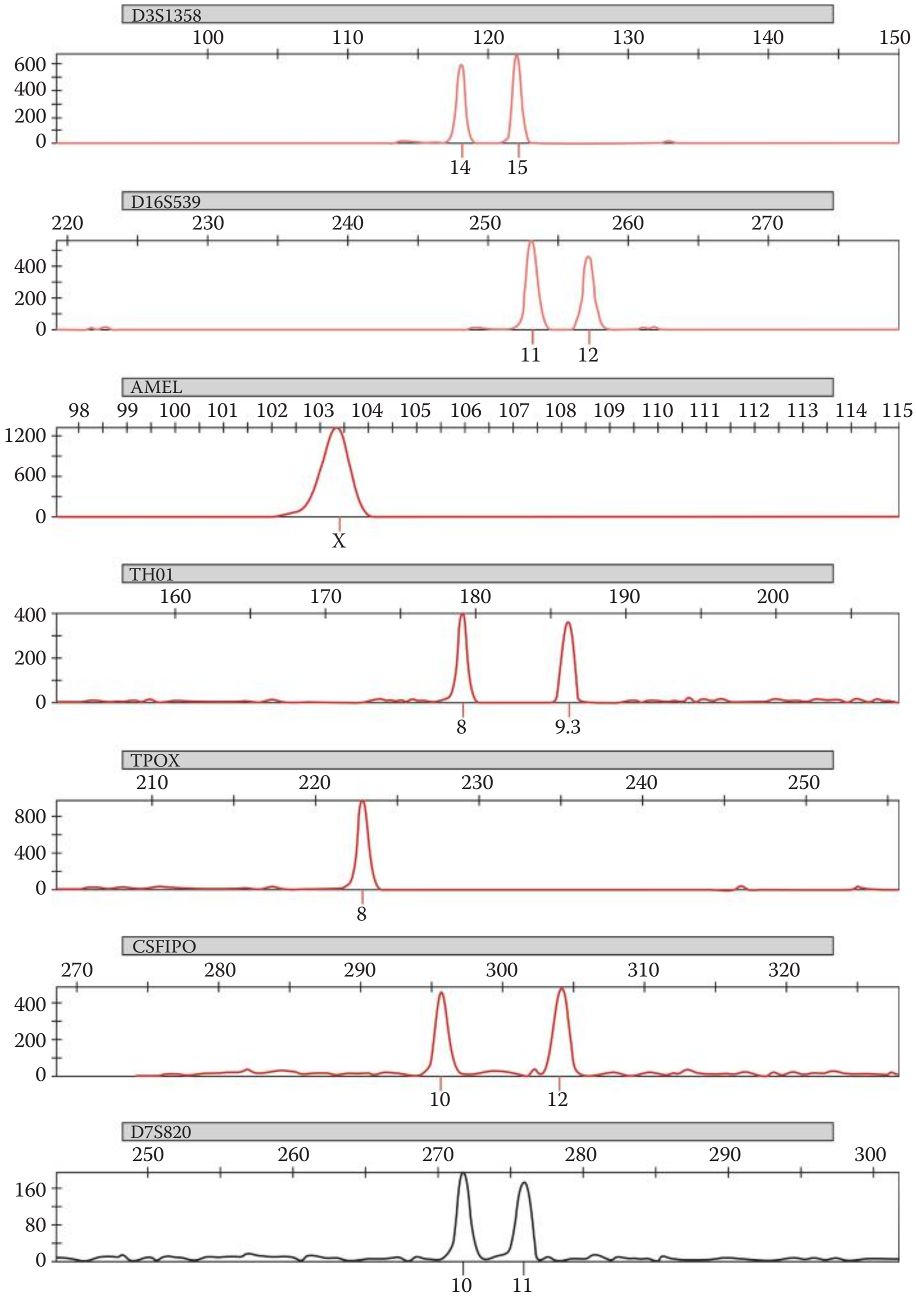

Figure 20.10 Individual DNA profile (Cofiler). The genotype of the DNA profile is shown. AMEL: X, X. CSF1PO: 10, 12. D3S1358: 14, 15. D16S539: 11, 12. D7S820:10, 11. TH01:8, 9.3. TPOX: 8, 8. (© Richard C. Li.)

In contrast, *somatic mutations* involve the mutation of only somatic cells. The germ cells are not affected, and thus a mutant allele is not transmitted to the progeny. A somatic mutation occurring at the core repeat region of an STR locus can be detected and compared to the wild-type allele. The ratio of the signal intensities of the wild-type and mutant alleles varies, depending on the number

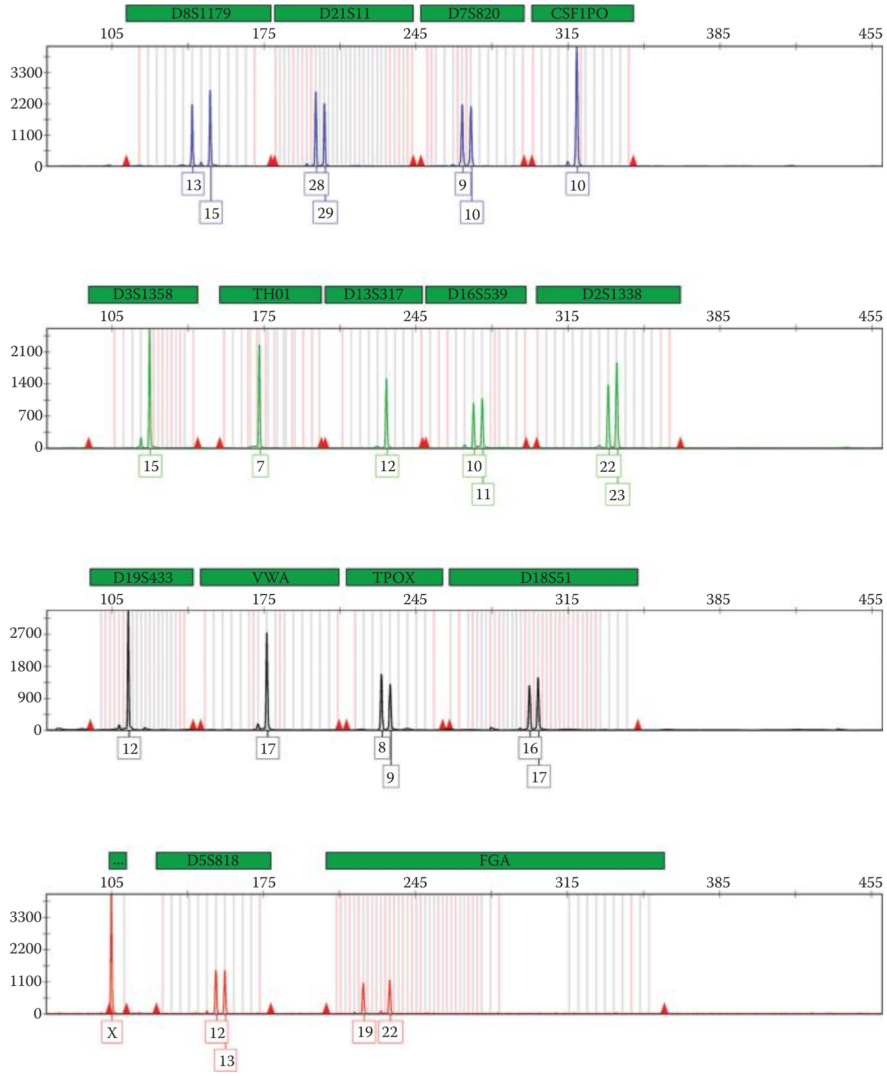

Figure 20.11 Individual DNA profile (Identifiler). The genotype of the DNA profile is shown. AMEL: X, X. CSF1PO: 10, 10. D2S1338: 22, 23. D3S1358: 15, 15. D5S818: 12, 13. D7S820: 9, 10. D8S1179: 13, 15. D13S317: 12, 12. D16S539: 10, 11. D18S51: 16, 17. D19S433: 12, 12. D21S11: 28, 29. FGA: 9, 22. TH01: 7, 7. TPOX: 8, 9. VWA: 17, 17. (© Richard C. Li.)

of mutation-carrying cells in the tissue. Somatic mutations are usually tissue specific. STR profiles from different tissues of the same individual can be compared if a somatic mutation is suspected.

20.4.1.2 Chromosomal and Gene Duplications

Duplicating one of the homologous chromosomes results in a total of three copies of a particular chromosome. This condition, called *trisomy*, is rare and often associated with genetic diseases such as Down's syndrome (chromosome 21 duplication). Duplications have also been observed in chromosomes 13, 18, and X. Other anomalies include duplication of a portion of a chromosome and a single or group of genes instead of an entire chromosome.

A duplication bearing a mutation within the STR core repeat region can affect the number of tandem repeat units. If the duplicated locus is mutated, a *triallelic* or three-peaks pattern can be detected at a single locus, but not at loci located at other chromosomes in a multiplex STR profile (Figure 20.12). The three alleles usually have equal signal intensity (peak amplitude or peak area). Triallelic patterns at STR loci commonly used for forensic DNA analysis have been documented. Many occur at the TPOX, FGA, and CSF1PO loci. If the duplicate locus is not mutated, only two alleles will be observed in a heterozygote. However, the ratio of the peak amplitude of the alleles will be 1:2 (one copy vs. two copies including the duplicate) at that particular locus. However, STR profiles at the loci located at other chromosomes are not affected.

20.4.1.3 Point Mutations

Point mutations involve the changing of a nucleotide sequence through nucleotide substitution, insertion, or deletion. Insertion or deletion mutations affect the lengths of the core repeat regions and the amplified flanking regions of STR loci and thus affect STR profiles. Nucleotide substitution mutations (except those residing within the primer-binding regions) do not affect the length of DNA and thus do not affect STR profiles.

However, mutations occurring at the primer-binding sequences of the flanking regions of STR loci may affect genotype results. If a mutation at a primer-binding sequence prevents the primer from annealing to the template, this leads to a complete failure of the amplification of the allele. This phenomenon is known as a *null allele* or silent allele (Figure 20.13). To overcome the consequences of a null allele, an alternative primer annealing to a flanking region away from the mutated sequence can be used. Additionally, a primer with the sequence that is complementary to the known mutation can also be used. If the mutation does not completely prevent the primer from annealing but reduces the efficiency of the amplification, the resulting signal intensity of the allele is usually decreased. This problem may be solved by modifying the condition of amplification for the mutant allele.

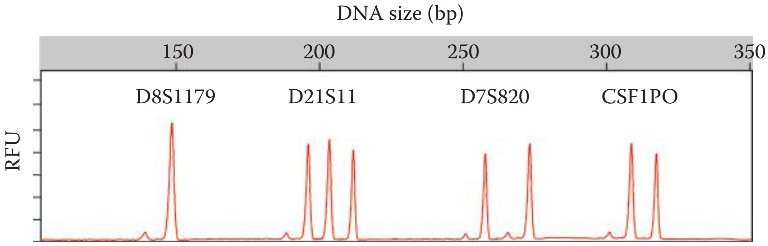

Figure 20.12 Triallele. In this example, the triallele is observed only at D21S11 and not at other STR loci. (© Richard C. Li.)

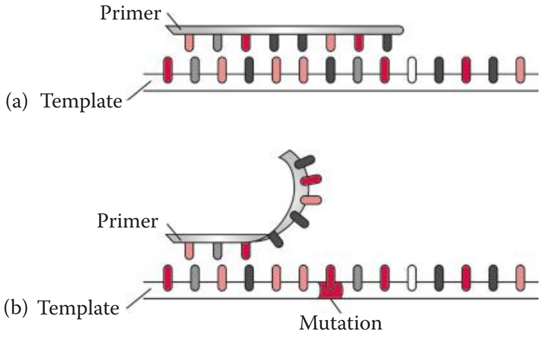

Figure 20.13 Null allele. An allele present in the sample failed to be amplified by one of the primer sets as a result of a rare mutation at the primer-binding sequence of the flanking region: (a) wild type and (b) mutation. (© Richard C. Li.)

20.4.2 Amplification Artifacts

20.4.2.1 Stuttering

A stutter is a minor allele peak, also known as a stutter peak, whose repeat units are shorter or longer than the parental allele peak (Figure 20.14). Less stuttering is observed with pentameric and hexameric repeat unit loci compared to shorter repeat unit loci. The loci that contain complex repeat sequences usually exhibit reduced stuttering. At a given STR locus, large alleles appear to yield more stutter than smaller alleles.

Commonly observed stutters are one repeat unit shorter than the parental allele. It is believed that stuttering is due to the slippage of polymerase, which may have occurred during amplification reactions (Figure 20.15). Stutters with repeat units longer than the parental allele peak can also be observed, but are very rare. The *stutter ratio* is defined as the area of the stutter peak divided by the area of the parental peak. The stutter ratio is usually less than 0.15. A ratio of greater than 0.15 should be interpreted with caution due to the potential presence of DNA from more than one contributor.

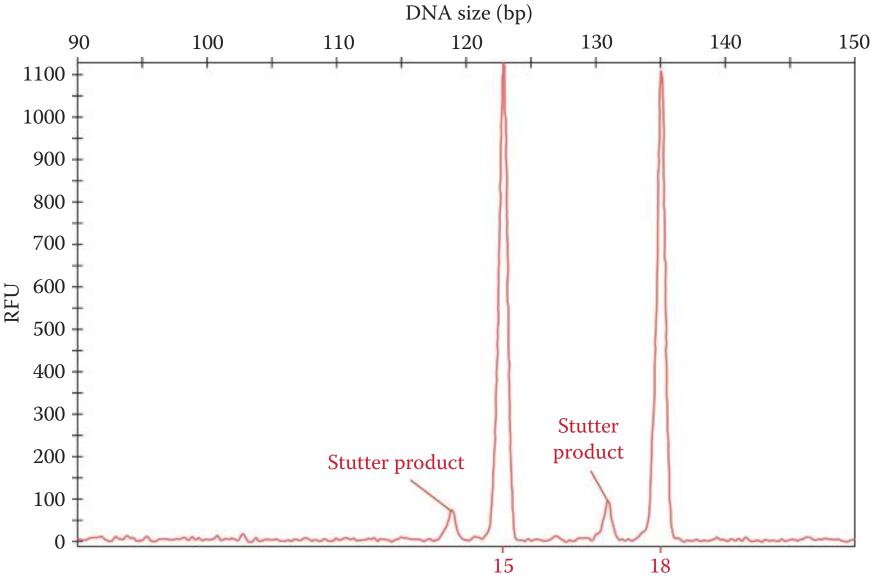

Figure 20.14 Stutter products. (© Richard C. Li.)

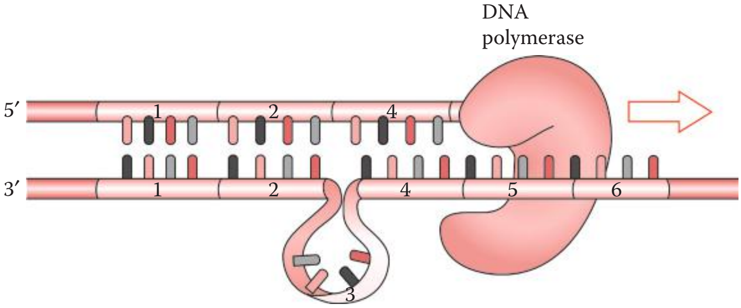

Figure 20.15 Proposed mechanism for stutter products. During the DNA synthesis step of PCR amplification, a DNA polymerase slips, and a region of the primer–template complex becomes unpaired, causing the template strand to form a loop. The consequence of this one-repeat loop is a shortened PCR product smaller than the template by a single repeat unit. (© Richard C. Li.)

20.4.2.2 Nontemplate Adenylation

During PCR amplification, DNA polymerase often adds an extra nucleotide, usually an adenosine, to the 3′-end of an amplicon. Such a phenomenon is referred to as a *nontemplate addition* resulting in an amplicon that is one base pair longer than the parental allele (designated the +A peak), as shown in Figure 20.16. As a result, an amplicon has both the –A amplicon, which corresponds to the size of the parental allele, and the +A amplicon, which represents the amplicon with the nontemplate addition. To simplify the analysis, the most commonly used multiplex STR kits utilize amplification conditions that favor the adenylation of amplicons. Thus, most amplicons in a sample contain an additional adenosine at the 3′ end (+A peak). However, partial nontemplate addition can occur when too much DNA template is utilized in PCR amplification. As a result, a mixture of –A and +A peaks is usually observed.

20.4.2.3 Heterozygote Imbalance

Heterozygote imbalance occurs when one of the alleles has greater peak area or amplitude than the other allele within the same locus in which the two alleles of a heterozygote are compared (Figure 20.17). It is believed that heterozygote imbalance may arise if the DNA sample contains unequal copies of DNA template of the two alleles for the heterozygote, or the two alleles of a heterozygote are unequally amplified, a condition known as *preferential amplification*. Preferential amplification usually refers to an event where a smaller allele is amplified more efficiently than larger ones. As a result, the presence of heterozygote imbalance interferes with the interpretation of samples with a DNA mixture derived from more than one contributor.

20.4.2.4 Allelic Dropout

Allelic dropout occurs when an allele, usually one of the heterozygote alleles, fails to be detected. To date, our understanding of what causes the dropout is very limited. The occurrence of allelic dropout can be the result of an extreme situation of preferential amplification or heterozygote imbalance. Additionally, certain mutations leading to amplification failure (Section 20.4.1.3) can cause allelic dropout.

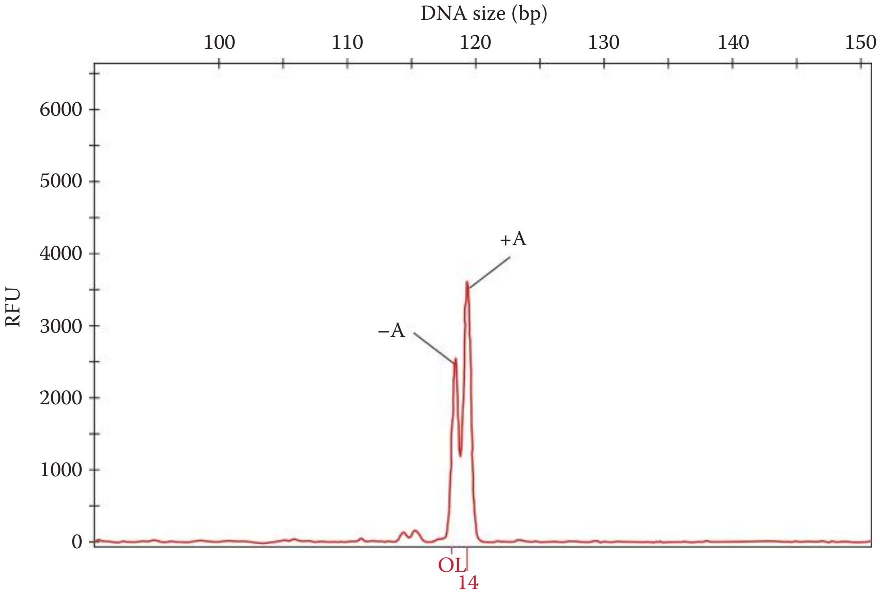

Figure 20.16 Nontemplate adenylation. OL represents the off-ladder allele. (© Richard C. Li.)

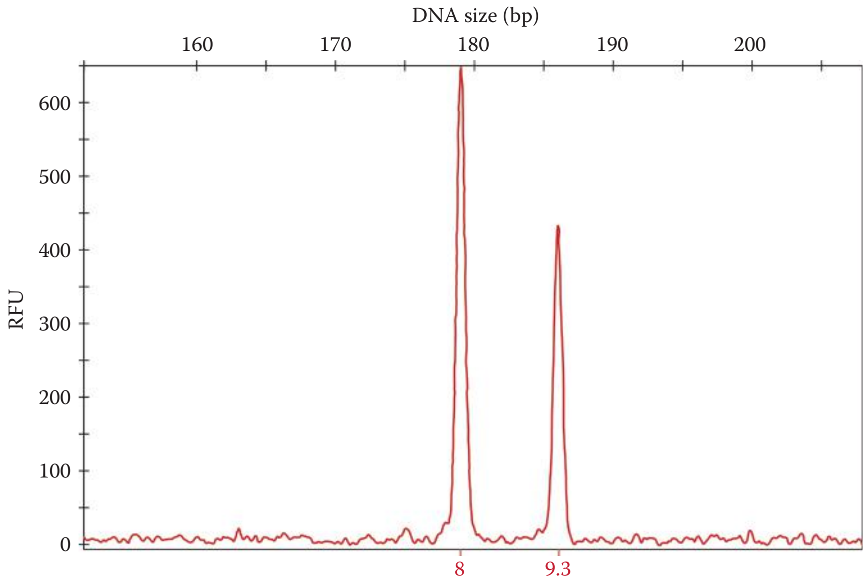

Figure 20.17 Heterozygote imbalance. The signal intensity of one allele is greater than that of the other allele within the same locus. (© Richard C. Li.)

20.4.3 Electrophoretic Artifacts

20.4.3.1 Pull-Up Peaks

A *pull-up peak* occurs when a minor peak of one color on an electropherogram is pulled up from a major allelic peak in another color (Figure 20.18) when the colors have overlapping spectra. For example, a green peak may pull up a yellow peak, or a blue peak may pull up a green peak. A pull-up peak may contribute to the inaccuracy of a profile if the position of a pull-up peak corresponds to the position of an allele. A pull-up peak often occurs when a sample is overloaded or a matrix file (a spectral calibration) is not updated. Thus, loading a proper amount of sample or installing an appropriate matrix file can prevent the occurrence of pull-up peaks.

20.4.3.2 Spikes

Spikes are sharp peaks, with similar signal intensities, that are present in all color panels of an electropherogram (Figure 20.19). Spikes are caused by air bubbles and urea crystals in the capillary of an electrophoretic platform. Voltage spikes can also contribute to spike peaks. The spikes are electrophoretic artifacts and are not reproducible. Thus, electrophoresis can be repeated to verify that the spikes occurred in a previous electrophoresis result.

20.5 Genotyping of Challenging Forensic Samples

20.5.1 Degraded DNA

Environmental exposure, such as high humidity and temperature, of biological evidence can lead to DNA degradation such as the breaking of DNA molecules into small fragments. The more severe the degradation, the more intensive the fragmentation. In forensic DNA analysis, the size range of STR amplicons is usually 100–500 bp in length. When a sample experiences some degradation, large alleles are less likely to be amplified than small alleles (Figures 20.20 and 20.21). As a result, the dropout of large alleles often occurs, leading to a partial DNA profile or even a failure in obtaining a DNA profile. To address this issue, the PCR primers can be

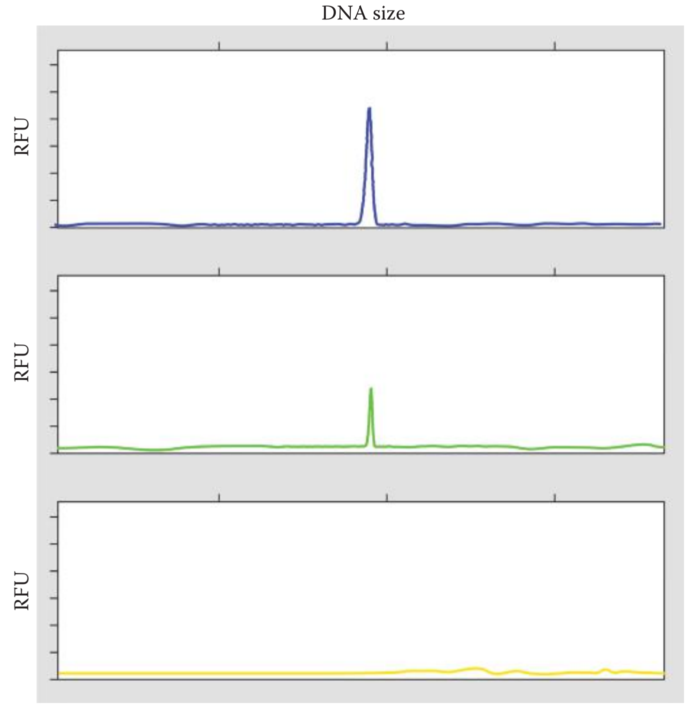

Figure 20.18 Pull-up peaks. The peaks with overlapping spectra observed in the top and middle panels are not observed in the bottom panel. (© Richard C. Li.)

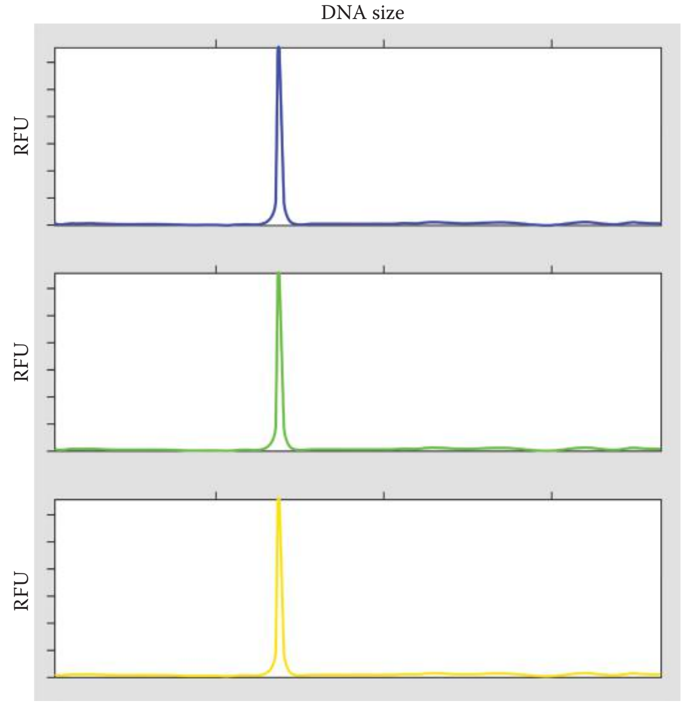

Figure 20.19 Spike peaks can be observed in various intensities. (© Richard C. Li.)

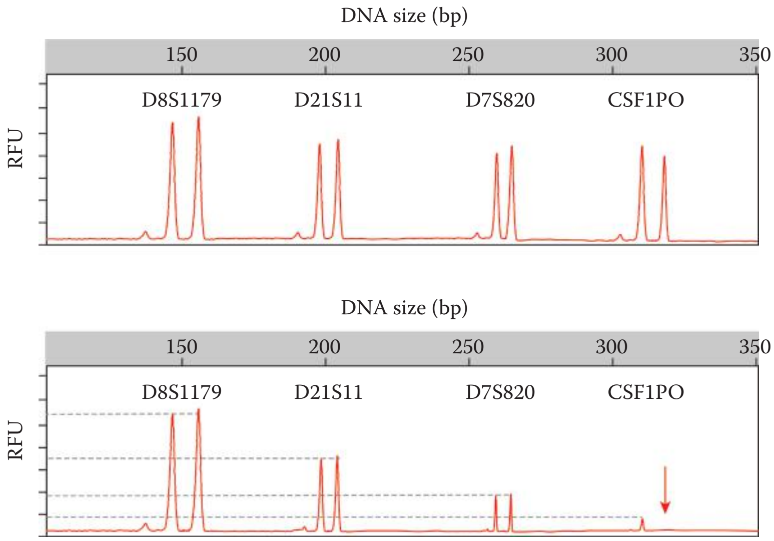

Figure 20.20 Diagram of STR profiles of samples: without DNA degradation (top) and with degradation (bottom). Arrow: allelic dropout. (© Richard C. Li.)

redesigned to anneal more proximally to the STR core repeat region than standard STR primers, yielding small amplicons also known as miniSTRs. Using the miniSTR strategy, more alleles can be detected in degraded DNA samples than using the standard STR primers.

20.5.2 Low Copy Number DNA Testing

Low copy number (LCN) DNA analysis involves the testing of very small amounts of DNA (<100 pg) in a sample. LCN DNA analysis is often needed for samples derived from evidence such as fingerprints and tools and weapons handled by perpetrators. STR analysis of extremely low levels of human DNA can be achieved by increasing the number of PCR cycles (e.g., increasing from 28 to 34 cycles) to improve the yield of amplicons, thus improving the sensitivity of the analysis.

However, this approach also increases the appearance of artifacts that can make interpretation difficult. For instance, the occurrence of allele dropout, heterozygote imbalance, and stuttering is frequently observed in LCN DNA analysis. Additionally, allele drop-in can arise from contamination. The phenomenon of allele drop-in is usually not reproducible. In an LCN DNA analysis, genotypes can be determined if identical alleles can be detected from two independent amplification reactions.

20.5.3 Mixtures

Samples of DNA from two or more contributors are commonly encountered in forensic cases such as sexual assaults in which the evidence recovered from a victim is mixed with a suspect's bodily fluids (Figure 20.22). The interpretation of DNA profiles of mixed stains is known as *mixture interpretation*, which is described below:

1. *To determine the presence of a mixture*: First, determine whether the source of the DNA in the sample came from one or more individuals by examining the number of alleles at multiple loci. The characteristics listed below usually indicate a mixture:
    a. Severe heterozygote imbalance.
    b. Stutter ratio above 0.15.
    c. Presence of three or more alleles per locus at multiple loci.

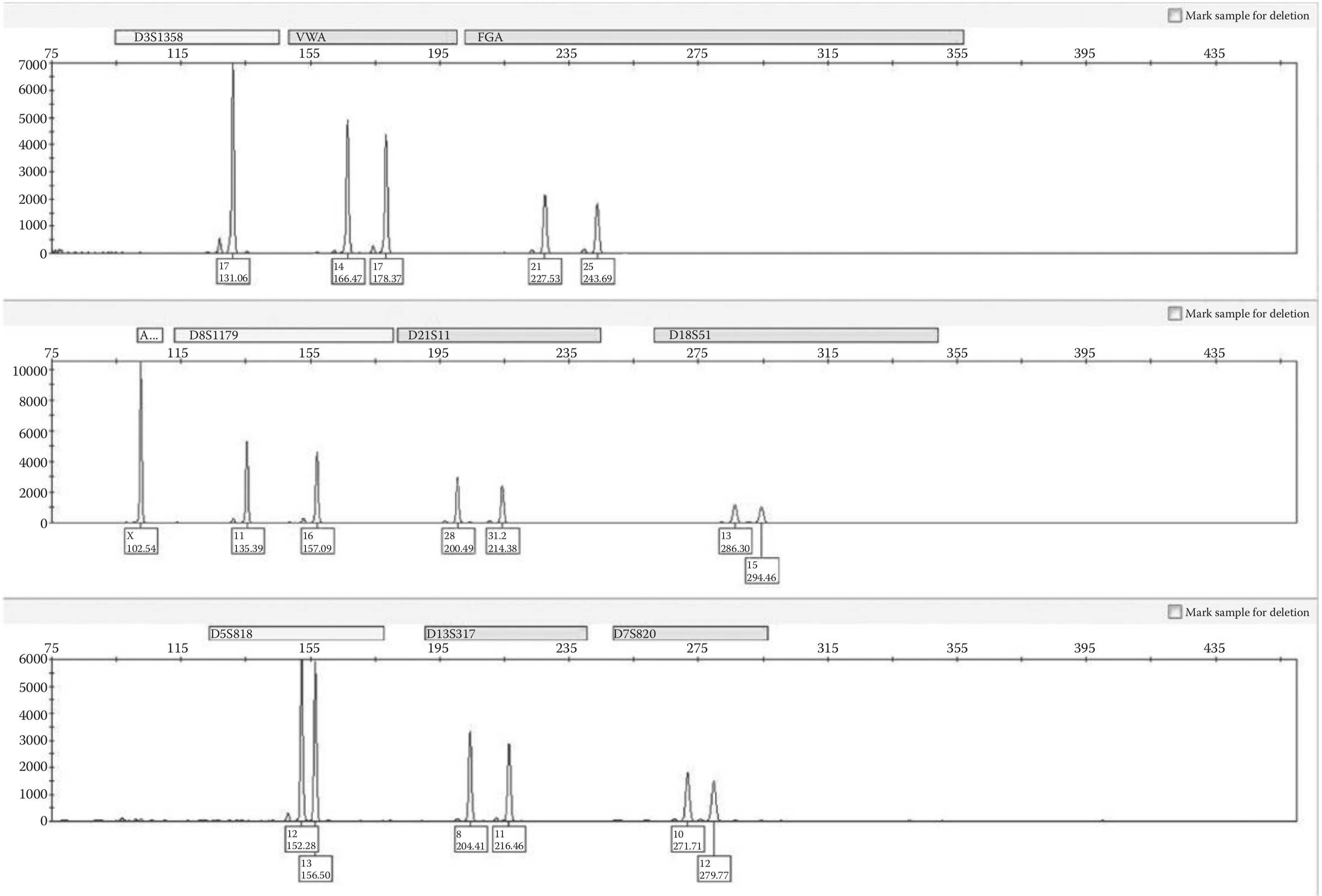

Figure 20.21 DNA profile (AmpFlSTR® Profiler Plus® PCR Amplification Kit; Applied Biosystems) of the degraded sample. (© Richard C. Li.)

Additionally, caution should be taken to distinguish the presence of a mixture and various artifacts such as stutters and nontemplate adenylation.

2. *To determine the genotypes of all alleles and to identify the number of contributors*: Note that the maximum number of alleles at any given locus is two per individual. In the case of homozygous or allele overlap, the number of alleles observed can be less than two per individual.
3. *To estimate the ratios of the contributions*: Determine the relative ratios of the contributions to the mixture made by each individual by comparing the peak areas or

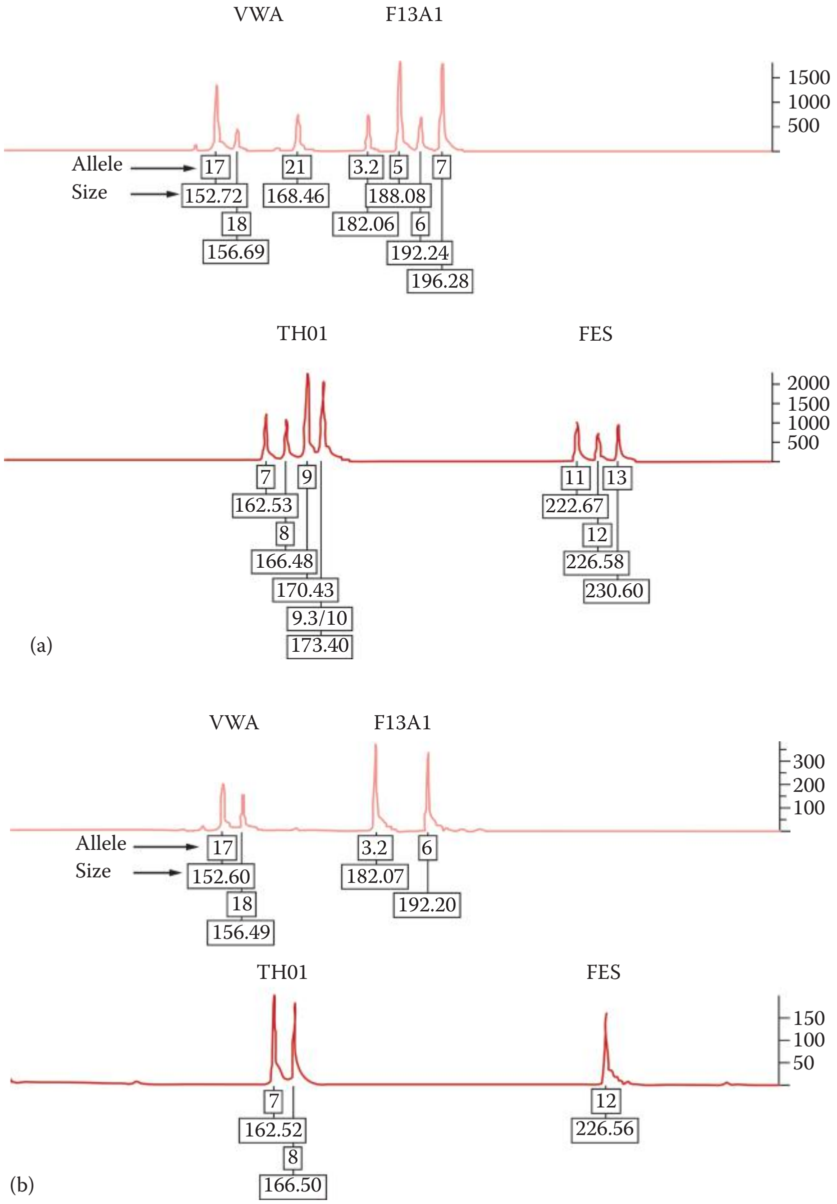

Figure 20.22 DNA profiles of mixed bodily fluids. (a) DNA profile of mixed stains from evidence, (b) DNA profile of the victim. (© Richard C. Li.)

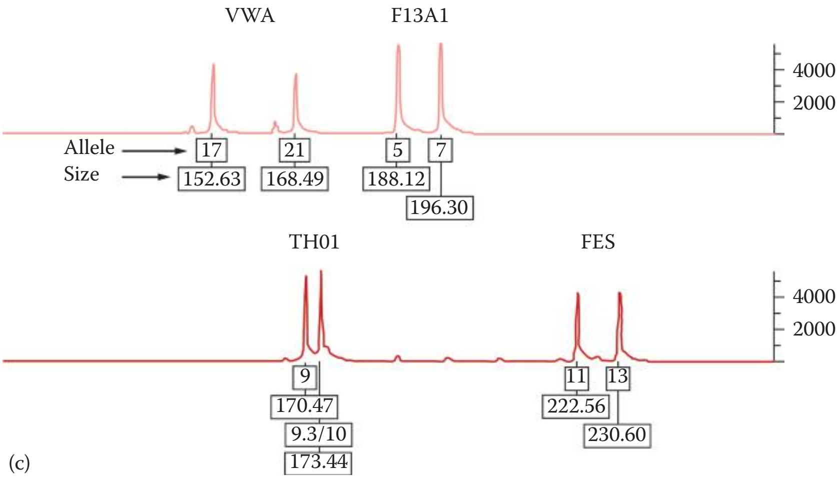

Figure 20.22 (Continued) (c) DNA profile of the suspect.

amplitudes. Amelogenin, a sex-typing marker, is useful in determining the genders of DNA contributors.

4. *To consider all possible genotype combinations*: This may be done by pair-wise comparisons to determine the allele combinations that belong to the minor contributor and those that belong to the major contributor.
5. *To compare reference samples*: The final step is to compare the genotype profiles with the genotypes of reference samples from a suspect and victim. If the DNA profile of the suspect's reference sample matches a major or minor component of the mixture, the suspect cannot be excluded as a contributor.

Detailed guidance for basic steps in mixture interpretation is provided in a number of published guidelines. For example, the Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories was recently established by the Scientific Working Group on DNA Analysis Methods (SWGDAM), which provides direction on DNA profiling and mixture interpretation. Additionally, guidance for the statistical evaluations of the mixture analysis results is also provided.

Bibliography

STR Profiling

Andersen, J., et al., Report on the third EDNAP collaborative STR exercise. *Forensic Sci Int*, 1996, **78**(2): 83–93.

Anker, R., T. Steinbrueck, and H. Donis-Keller, Tetranucleotide repeat polymorphism at the human thyroid peroxidase (hTPO) locus. *Hum Mol Genet*, 1992, **1**(2): 137.

Asamura, H., et al., MiniSTR multiplex systems based on non-CODIS loci for analysis of degraded DNA samples. *Forensic Sci Int*, 2007, **173**(1): 7–15.

Asamura, H., et al., MiniY-STR quadruplex systems with short amplicon lengths for analysis of degraded DNA samples. *Forensic Sci Int Genet*, 2007, **1**(1): 56–61.

Aznar, J.M., et al., I-DNASE21 system: Development and SWGDAM validation of a new STR 21-plex reaction. *Forensic Sci Int Genet*, 2014, **8**(1): 10–19.

Balding, D.J., Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proc Natl Acad Sci U S A*, 2013, **110**(30): 12241–12246.

21

Sex Chromosome Haplotyping and Gender Identification

21.1 Y Chromosome Haplotyping

21.1.1 Human Y Chromosome Genome

The Y chromosome is inherited from the father and is passed on to all male offspring; this is known as *patrilineage* (Figure 21.1). Thus, the Y chromosome is unique to males. The chromosome encodes dozens of genes required for male-specific functions, including sex determination and spermatogenesis. The human Y chromosome genome contains approximately 59 million base pairs (bp) and likely contains 50–60 genes. The Y chromosome can be divided into two regions: the pseudoautosomal region and the male-specific Y region (Figure 21.2).

21.1.1.1 Pseudoautosomal Region

The *pseudoautosomal regions (PARs)* are homologous nucleotide sequences that are present on the X and Y chromosomes. There are two PARs on each X and Y chromosome: PAR1 and PAR2. In the Y chromosome, PAR1 is located on the terminal region of the short arm. The PAR1 comprises 2.6 Mb. Twenty-four genes have been identified within the PAR1. The PAR2 of the Y chromosome is located at the tip of the long arm. The PAR2 comprises 320 kb, with only four genes identified so far.

The PARs play an important role in proper segregation of the X and Y chromosomes during meiosis. During meiosis in males, the PARs allow the Y chromosome to pair with the X chromosome. Crossing over and recombination (Chapter 25) within the PARs of the X and Y chromosomes can occur. As a result, males can inherit an allele originally present on the PARs of the X chromosome, and females can inherit an allele originally present on the PARs of the Y chromosome. In particular, PAR1 plays a major role in X–Y chromosome pairing. The deletion of PAR1 results in failure of pairing and leads to male infertility. PAR2 is much shorter than PAR1, and thus shows a lower frequency of pairing than PAR1. The deletion of PAR2 exhibits less severe phenotypes than that of PAR1.

21.1.1.2 Male-Specific Y Region

The remainder of the Y chromosome is known as the *male-specific Y (MSY)* region. It was previously called the *nonrecombining Y (NRY)* region. The bulk of this region does not participate in homologous recombination. However, certain sections involve intrachromosomal gene conversion, which is the nonreciprocal transfer of genomic DNA between a pair of repeated genes on a single chromosome; in this case, the Y chromosome. About 40 megabases (Mb) within the

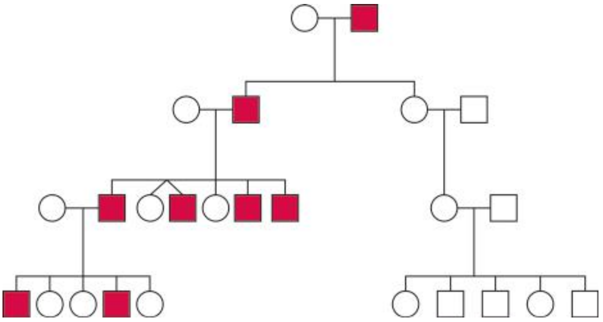

Figure 21.1 Human family pedigree showing inheritance of the Y chromosome. Females and males are denoted by circles and squares, respectively. Red symbols indicate individuals who inherited the same Y chromosome.

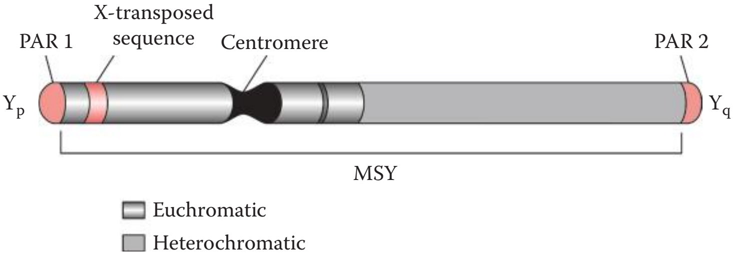

Figure 21.2 Human Y chromosome structure. PAR, the pseudoautosomal region; MSY, male-specific Y region; Y_p, the short arm of the Y chromosome; Y_q, the long arm of the Y chromosome. (© Richard C. Li.)

MSY region are heterochromatic (highly repetitive sequences) including the centromeric region and the bulk of the distal long arm. The euchromatic region, which comprises transcriptional active genes, is approximately 23 Mb. Certain sections of the euchromatic region share some homology with the X chromosome. For instance, X-transposed sequences of the Y chromosome are 99% identical to sequences within Xq21 (a band in the long arm of the X chromosome). The X-transposed sequences are the sequences, a total of approximately 3.4 Mb in length, that were transposed from the X chromosome to the Y chromosome several million years ago. Additionally, dozens of genes located in the euchromatic region share 60%–96% homology with their X chromosome counterparts. The regions sharing homology with the X chromosome should be avoided when selecting Y chromosome–specific markers for forensic DNA profiling.

21.1.1.3 Polymorphic Sequences

The Y chromosome contains an abundance of polymorphic markers. DYF155S1, also known as MSY1, is the first characterized variable number of tandem repeats (VNTR) or minisatellite at the human Y chromosome. It consists of an array of AT-rich repeats at 25 bp per unit repeat. The DYF155S1 locus is highly polymorphic. The unit repeat sequence varies through base substitution. At least five different variant types of unit repeat sequences exist. Additionally, the numbers of these units also vary, ranging from 50 to 115 repeats. Thus, the length of alleles range from approximately 1200 to 2800 bp. DYF155S1 is of considerable interest as a potential marker in forensic testing. However, the analysis method is labor intensive, making it difficult to implement in forensic casework, such as the investigation of sex crimes. Moreover, many single nucleotide polymorphisms (SNPs) and mobile elements exist at the Y chromosome. However, the discrimination power of SNPs and mobile elements is considerably less than that of Y chromosome short tandem repeats (Y-STRs). To date, Y-STRs are usually used for Y chromosome DNA testing due to the high-throughput analysis and good discrimination power (Section 21.1.2).

21.1.2 Y-STR

Y chromosome loci are very important for forensic DNA profiling, and this chapter will discuss such applications. For instance, the Y-STRs used in forensic DNA testing are male specific (for humans and certain higher primates) and are thus useful in the investigation of sexual assault cases involving male suspects. The evidence gathered in such cases usually consists of a mixture of high levels of female DNA and low levels of male DNA. The Y chromosome–specific loci can be examined without interference from large amounts of female DNA; differential extraction of sperm and nonsperm cells may not be needed. Furthermore, the Y-STR system is useful for determining the numbers of unrelated male perpetrators in sexual assault cases. The Y-STR loci used for forensic applications are located in the nonrecombining section of the Y chromosome so that patrilineage can be established. The technique can be used for paternity testing and the identification of missing persons. Finally, data interpretation can be simplified by the use of a single allele per Y-STR locus profile at most loci. Reference databases are available for estimating Y-STR haplotype frequencies among various human populations for statistical analysis of profiling results.

The major disadvantage of Y-STR loci is that their discriminating power is lowerthan that of autosomal loci. Because Y chromosome loci are linked, the product rule for statistical calculations for profile probability does not apply. Chapter 25 discusses the statistical evaluation of the strength of the matches. Also, the current Y-STR profiling cannot distinguish individuals with the same patrilineage.

More than 400 STR loci have been identified in the Y chromosome genome. The precise locations of these loci have been sequentially mapped using human genome sequencing data. Most Y-STR loci, approximately 60% of the 400 identified, are located on the long arm of the chromosome; about 22% are located on the short arm and a few are found in the centromeric region (Figure 21.3). Y-STRs in the telomeric region have yet to be identified. Only about 5% of Y-STRs are located within 5′ untranslated or intron regions of protein-coding genes. The

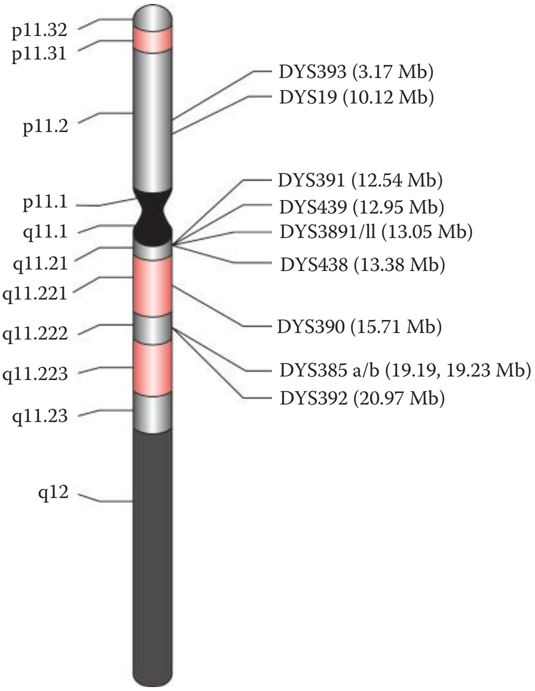

Figure 21.3 Human cytogenetic map of the Y chromosome. The Y-STRs and positions are shown (Mb = megabase). Cytogenetic patterns with alternating dark and light bands are shown. (© Richard C. Li.)

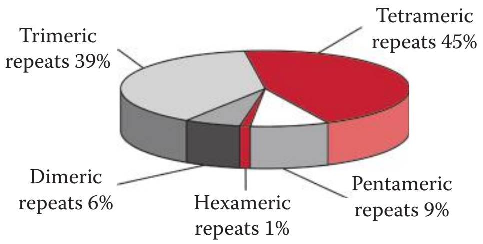

Figure 21.4 Human Y-STRs with different repeat unit length. About 400 Y-STRs have been identified and categorized according to repeat unit length. (Adapted from Hanson, E.K. and Ballantyne, J., *Leg Med (Tokyo)*, 8, 110, 2006.)

repeat unit lengths of identified Y-STRs have been analyzed. Among the 400 Y-STRs, 6% are dimeric repeats, 39% are trimeric, 45% are tetrameric, 9% are pentameric, and 1% are hexameric (Figure 21.4).

Fewer than half of the Y-STRs have been characterized. Some loci are polymorphic and are useful for forensic applications and developing new Y-STR multiplex systems. Since homologous recombination does not occur on the majority of the Y chromosome, alleles of Y-STR loci are linked (inherited together); they are referred to as *haplotypes*. As a result, the discrimination power of Y-STRs is much lower than that of autosomal STRs. The most commonly used Y-STR loci for forensic testing are described below.

21.1.2.1 Core Y-STR Loci

In 1997, the *European minimal haplotype* locus set, also known as the minimal haplotype loci, was recommended by the International Y-STR User Group for forensic applications (Table 21.1; Figure 21.5). This haplotype set includes a core set of nine Y-STR loci: DYS19, DYS385a and b, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393. In 2003, two additional loci were recommended by the Scientific Working Group for DNA Analysis Methods (SWGDAM) for forensic DNA analysis. The SWGDAM loci include the European minimal haplotype loci plus two additional loci: DYS438 and DYS439 (Table 21.1).

The application Y-STR for forensic casework has been expanded with additional Y-STR markers. The use of Y-STR loci has been facilitated by commercially available PCR amplification

Table 21.1 Common Y-STR Core Loci

Locus	EMH	US Haplotype Loci	Repeat Motif
DYS19	☑	☑	TAGA
DYS385a/b	☑	☑	GAAA
DYS389 I	☑	☑	TCTA
DYS389 II	☑	☑	[TCTG] [TCTA]
DYS390	☑	☑	[TCTG] [TCTA]
DYS391	☑	☑	TCTA
DYS392	☑	☑	TAT
DYS393	☑	☑	AGAT
DYS438		☑	TTTTC
DYS439		☑	GATA

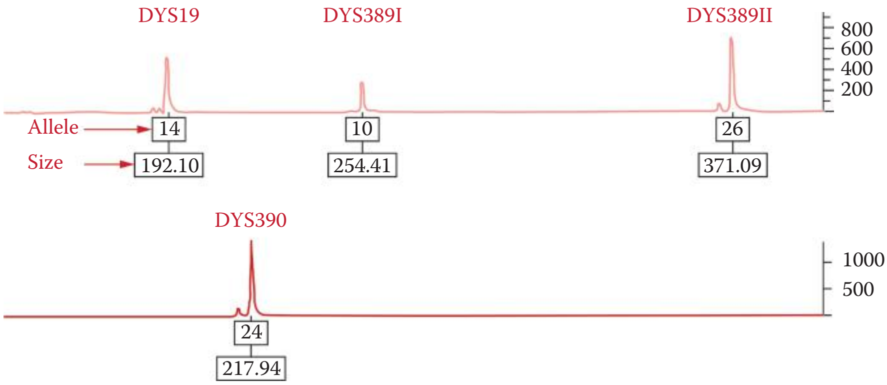

Figure 21.5 Electropherogram of Y-STR profile using an early generation of multiplex of four loci. The genotype of the DNA profile is shown (DYS19: 14. DYS389I: 10. DYS389II: 26. DYS390: 24). (© Richard C. Li.)

kits in multiplex Y-STR systems. To improve discriminating power, multiplex systems including new Y-STR loci are desired. Many new Y-STR loci are being characterized for developing new multiplex systems. Commercially available kits with more Y-STR loci are now available and have been validated for forensic use (Table 21.2; Figures 21.6 and 21.7). Some of the additional Y-STR loci are highly discriminating, allowing for further distinction between unrelated male individuals. Additionally, Y-STR loci, such as DYS570 and DYS576, with high mutation rates (see rapidly mutating Y-STR loci; Section 21.1.2.3) are included, which are useful for the discrimination of related individuals.

21.1.2.2 Multilocal Y-STR Loci

DYS385 and DYS389 are *multilocal Y-STR loci* (*MLL*). The MLL designation refers to the presence of a particular STR at more than one site on the Y chromosome due to duplication. To date, about 50 such MLL Y-STRs have been identified. Further MLL subdivisions are designated bilocal, trilocal, and so on. DYS385 and DYS389 are bilocal.

The DYS385 locus has two inverted duplicated clusters and is separated by a 4×10^4 bp interstitial region (Figure 21.8). It can be amplified by a single set of primers. One allele is observed if the duplicates are the same length. If the duplicated clusters have different lengths, they can generate two different alleles when amplified. The smaller allele is designated “a” and the larger allele is designated “b.” Moreover, the DYS389 locus has two duplicated clusters with the same orientation (Figure 21.9). In a single set of PCR primers, there are two binding sites for the same forward primer at each 5′ flanking sequence of the core repeat region of DYS389. These binding sites between DYS389I and DYS389II are about 120 bp apart. Therefore, two amplicons are produced. DYS389I is designated for the smaller allele, and DYS389II is designated for the larger allele.

21.1.2.3 Rapidly Mutating Y-STR

Current Y-STR loci implemented in forensic casework have adequate resolution of males from different patrilineages. However, most Y-STR sets have limited abilities to differentiate related males who belong to the same patrilineage. Thus, current forensic Y-STR profiling is not able to exclude patrilineal relatives of the suspect. Recently, rapidly mutating Y-STR (RM Y-STR) loci have been identified and characterized. The mutation rates of RM Y-STR are above 10^{-2}, which is considerably higher than the average mutation rates of Y-STRs. The average mutation rate for the majority of the Y-STRs characterized, including the core Y-STR loci currently used in forensic casework, is approximately 10^{-4}–10^{-3} per generation. Due to the high mutation rates, RM Y-STRs can improve patrilineage resolution. For use in forensic casework, the RM Y-STR loci have potential abilities to differentiate both paternally related and unrelated males.

Table 21.2 Y-STR Loci Covered by Representative Commercial Kits

	Loci	Yfiler[a]	PowerPlexY23[b]
Core loci	DYS19	☑	☑
	DYS385a/b	☑	☑
	DYS389 I	☑	☑
	DYS389 II	☑	☑
	DYS390	☑	☑
	DYS391	☑	☑
	DYS392	☑	☑
	DYS393	☑	☑
	DYS438	☑	☑
	DYS439	☑	☑
Additional loci	DYS437	☑	☑
	DYS448	☑	☑
	DYS456	☑	☑
	DYS458	☑	☑
	DYS481		☑
	DYS533		☑
	DYS549		☑
	DYS570		☑
	DYS576		☑
	DYS635	☑	☑
	DYS643		☑
	Y-GATA-H4	☑	☑

[a] AmpFlSTR® Yfiler® PCR Amplification Kit (Applied Biosystems).

[b] PowerPlex® Y23 System (Promega).

21.2 X Chromosome Haplotyping

X-chromosomal STR (X-STR) profiling is a useful tool in kinship testing in forensic investigations. For example, males under usual circumstances have only one X chromosome; thus males are hemizygous for the X-STR loci on the X chromosome. Homologous recombination between the X and Y chromosomes is restricted to the homologous PARs (Figure 21.2). The paternal X chromosome is inherited by daughters as haplotypes. Thus, father–daughter kinship is easier to determine using X-STRs than autosomal STRs. In females, there are two copies of the X chromosomes. Homologous recombination can occur between two X chromosomes in the mother–child transmission. Although it is possible to determine mother–child kinship using X-STR analysis, the level of certainty is less than that of father–daughter kinship analysis. Nevertheless, sharing of rare X-STR haplotypes can strengthen an indication of kinship. The most useful

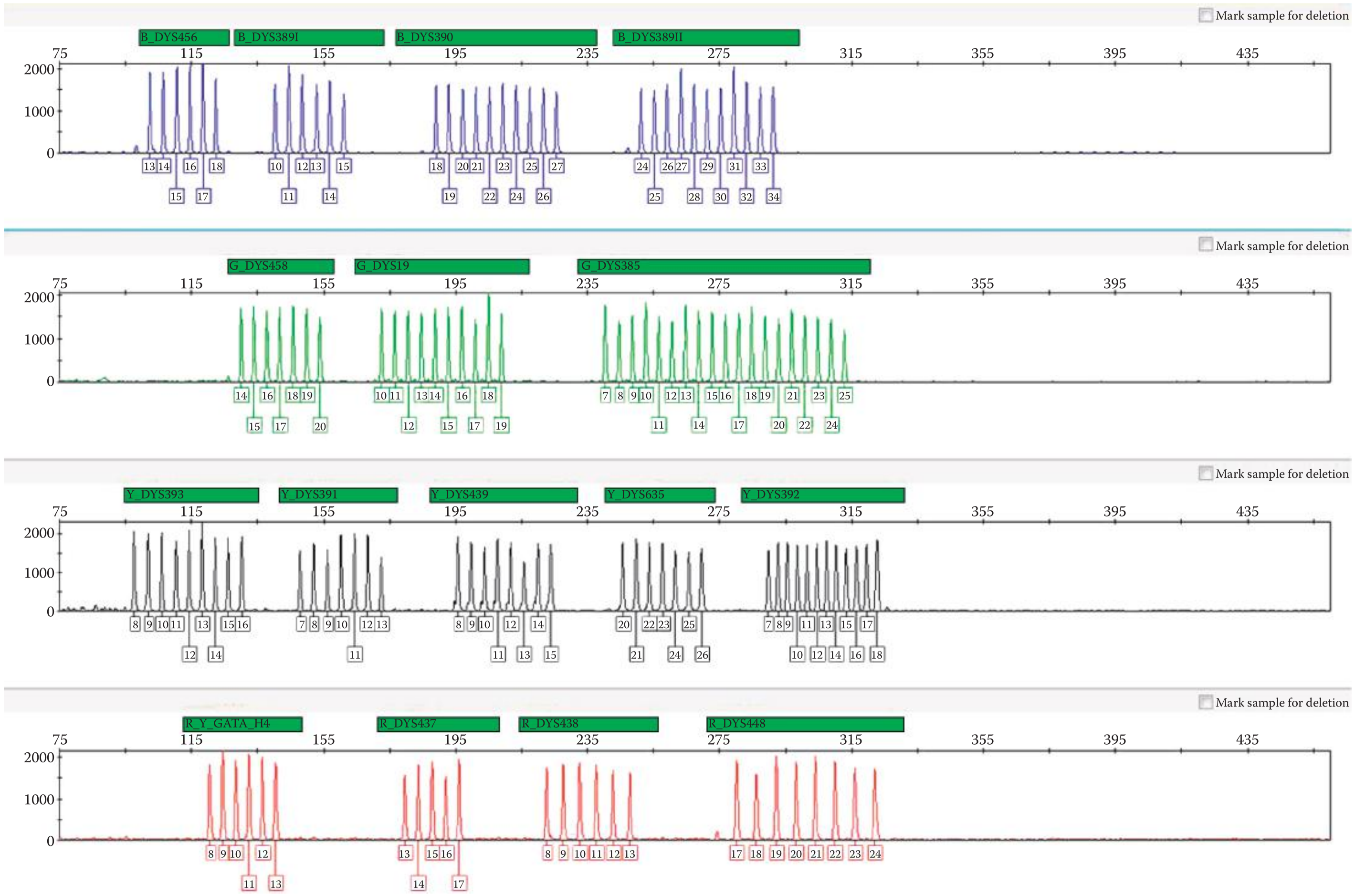

Figure 21.6 Electropherogram of the allelic ladder of AmpFlSTR® Yfiler® PCR Amplification Kit (Applied Biosystems). DYS19, DYS385 a/b, DYS389 I, DYS389 II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and Y-GATA-H4 loci are shown. (© Richard C. Li.)

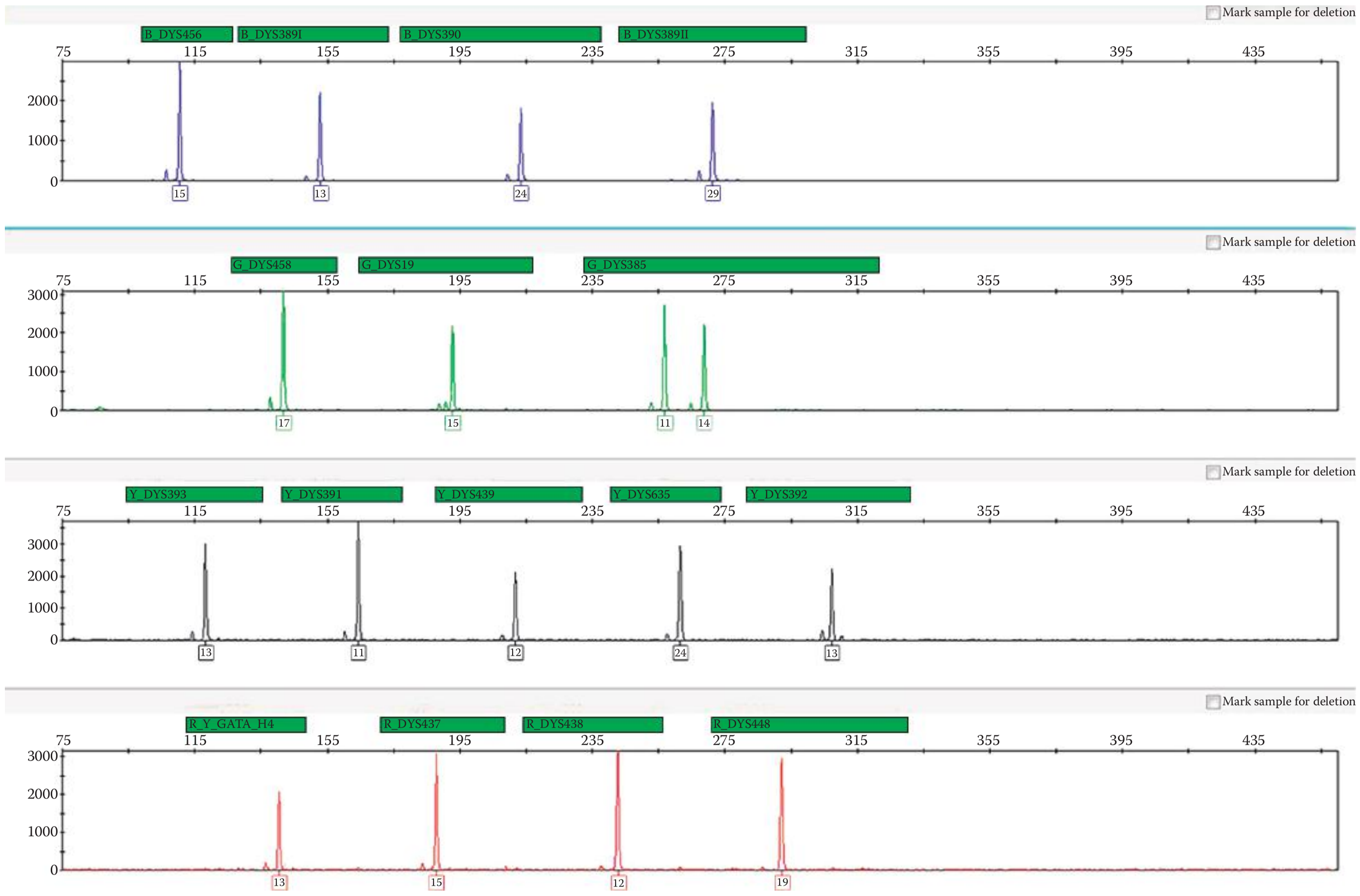

Figure 21.7 Individual Y-STR profile (AmpFlSTR® Yfiler® PCR Amplification Kit). The genotype of the DNA profile is shown. DYS19: 15. DYS385 a/b: 11, 14. DYS389 I: 13. DYS389 II: 29. DYS390: 24. DYS391: 11. DYS392: 13. DYS393: 13. DYS437: 15. DYS438: 12. DYS439: 12. DYS448: 19. DYS456: 15. DYS458: 17. DYS635: 24.Y-GATA-H4: 13. (© Richard C. Li.)

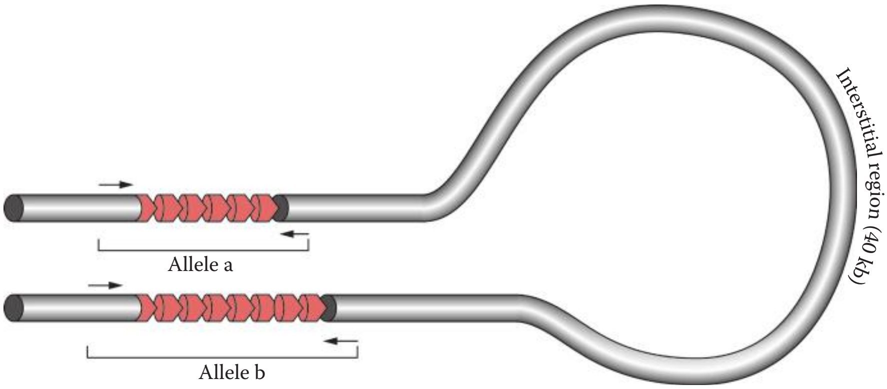

Figure 21.8 MLL Y-STR locus DYS385. At the DYS385 locus, note the two inverted duplicated regions of the Y chromosome with an interstitial region of 40 kilobases (kb). These inverted regions can be amplified with a single pair of primers (indicated by arrows) in one PCR reaction. The allele designations are described in the text. (© Richard C. Li.)

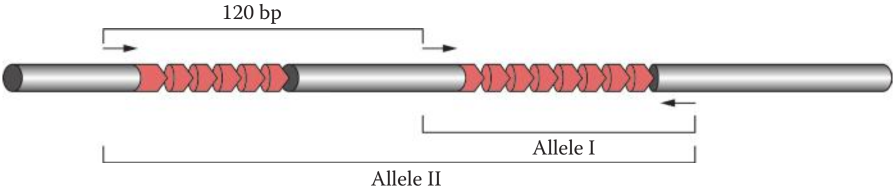

Figure 21.9 MLL Y-STR locus DYS389. At the DYS389 locus, note the two duplicated regions of the Y chromosome with the same orientation. These duplicated regions can be amplified with a single pair of primers (indicated by arrows) in one PCR reaction. The allele designations are described in the text. (© Richard C. Li.)

application of X-STR profiling, however, is in situations where autosomal STR profiling fails to determine kinship with high levels of certainty, and where Y-chromosomal or mitochondrial DNA profiling results are inconclusive.

Many X-STR loci have been identified spanning the entire human X chromosome. Increasing the number of X-STR loci to be analyzed in the inherited region would increase the degree of certainty in determining kinship. However, it is necessary to take into consideration the linkage between X-STR loci on the X chromosome. Based on Gregor Mendel's first law, the different alleles of two loci segregate independently. This is definitely true for loci localized on different chromosomes. The segregation can also be observed for loci located on the same chromosome, including the X chromosome, due to homologous recombination and meiotic crossing-over. Generally speaking, the farther apart the two loci are on the chromosome in physical distance, the more likely those loci are to segregate independently. When two loci are very close to each other, the alleles of the two loci are inherited together as a haplotype. Therefore, including additional closely linked STR loci does not necessarily improve the discrimination power in kinship analysis.

Homologous recombination events at different regions of the genome may occur with different probabilities. Thus, the tightness of linkage between two X-STR loci cannot be accurately measured based on the physical distance. In order to determine how closely two different loci are linked, the *recombination fraction* (or recombination frequency) can be used, which is the percentage of recombinants resulting from chromosomal crossover between two loci during meiosis among all the offspring. The recombination fraction reflects how often the loci are inherited together or how often they are separated. The recombination fraction ranges from 0% to 50%. The minimum recombination fraction is 0, indicating perfect linkage (no recombinants). The maximum is 50%, meaning complete independence of the two loci (they are actually located on different chromosomes or on the same chromosome with great distance between them).

Multiplex X-STR assays have also been developed. Commonly used X-STR loci can be grouped into linkage groups. Each linkage group is presumably transmitted independently from one another. These loosely linked loci can be treated as if they were located on different chromosomes. The distribution of alleles at these X-STR loci should be in accordance with the Hardy–Weinberg principle (Chapter 25). Within each linkage group, the occurrence of recombination is negligible. Thus, loci within groups are generally considered as closely linked and are thus treated as a haplotype.

21.3 Sex Typing for Gender Identification

Sex typing of a biological sample is useful in forensic investigation. For example, the application of sex typing facilitates the identification of the victim and the offender's DNA evidence in sexual assault cases and the remains of victims in mass disasters or missing persons investigations. One commonly used sex-typing marker is the amelogenin (*AMEL*) locus. The use of multiplex PCR systems with an additional amelogenin marker, a non-STR marker, leads to potential gender determination.

21.3.1 Amelogenin Locus

This region encodes extracellular matrix proteins involved in tooth enamel formation (Table 21.3). Mutations in the *AMEL* gene can lead to an enamel defect known as amelogenesis imperfecta. Amelogenesis imperfecta is a disorder that causes abnormal formation of tooth enamel in both primary and permanent teeth. The formed tooth enamel is soft and thin; therefore it is easily damaged. The *AMEL* locus has two homologous genes: *AMELX* (Xp22.2) is located on the human X chromosome (Figure 21.10) and *AMELY* (Yp11.2) is located on the human Y chromosome (Figure 21.11). Although the genes constitute a homologous pair, they differ in size and sequence. The coding sequence of the *AMELX* gene has a size of 2872 bp and the *AMELY* gene has a size of 3272 bp in length.

The most commonly used sex-typing method at the *AMEL* locus is the detection of a 6 bp deletion at intron 3 of *AMELX* (Figures 21.12 and 21.13). This deletion is not present in *AMELY*. Sex typing can be performed using primers specifically amplifying the region of the *AMEL* locus. Primer sets were developed to amplify both alleles in a single PCR. The two most commonly

Table 21.3 Sex-Typing Makers and Homologous Genes on the Human X and Y Chromosomes

Sex-Typing Markers					
X Chromosome			Y Chromosome		
Gene Symbol	Chromosomal Location	Distance from Xpter (Mb)	Gene Symbol	Chromosomal Location	Distance from Ypter (Mb)
AMELX	Xp22.2	11.3	*AMELY*	Yp11.2	6.7
DXYS156	Xq21.31	88.9	DXYS156	Yp11.2	3.2
SOX3	Xq27.1	139.6	*SRY*	Yp11.31	2.7
STS	Xp22.31	7.1–7.3	*STSP1*	Yq11.221	17.7
TSPYL2	Xp11.2	53.1	*TSPY1*	Yp11.2	9.2–9.3

Source: Ensembl Homo sapiens version 75.37 (GRCh37). Mb, megabase.

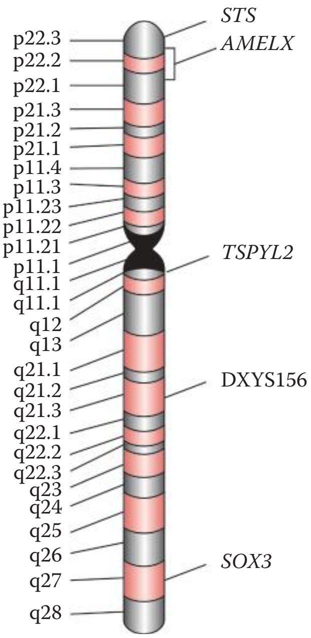
p22.3
p22.2
p22.1
p21.3
p21.2
p21.1
p11.4
p11.3
p11.23
p11.22
p11.21
p11.1
q11.1
q11.1
q12
q13
q21.1
q21.2
q21.3
q22.1
q22.2
q22.3
q23
q24
q25
q26
q27
q28
STS
AMELX
TSPYL2
DXYS156
SOX3

Figure 21.10 Cytogenetic map of the human X chromosome and sex-typing loci shown with physical positions (Mb = megabases). Cytogenetic patterns with alternating dark and light bands are shown. (© Richard C. Li.)

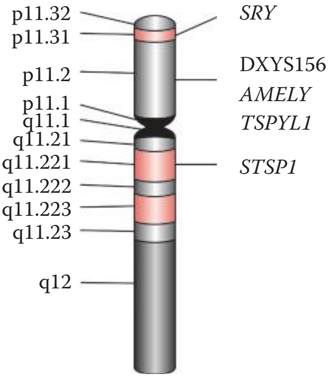
p11.32
p11.31
p11.2
p11.1
q11.1
q11.21
q11.221
q11.222
q11.223
q11.23
q12
SRY
DXYS156
AMELY
TSPYL1
STSP1

Figure 21.11 Cytogenetic map of the human Y chromosome and sex-typing loci shown with physical positions (Mb = megabases). Cytogenetic patterns with alternating dark and light bands are shown. (© Richard C. Li.)

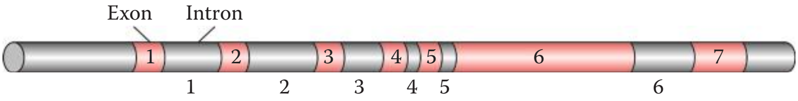
Exon
Intron
1
2
3
4
5
6
7
1
2
3
4
5
6

Figure 21.12 Structure of human *AMELY* gene. Exons 1 through 7 and introns 1 through 6 are numbered. (© Richard C. Li.)

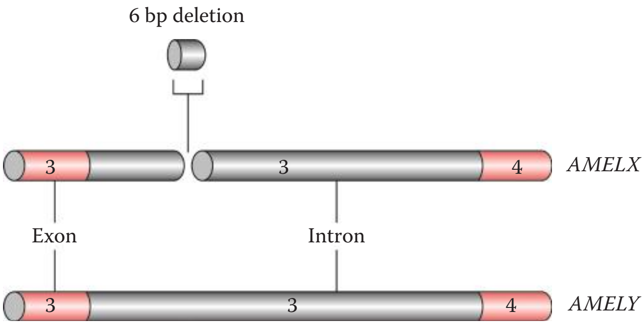

Figure 21.13 Sex typing using *AMEL* markers. A 6-bp deletion in intron 3 is present in *AMELX* but not in *AMELY* and can be resolved using electrophoresis as described in the text. (© Richard C. Li.)

used amelogenin primer sets generate amplicons of 106 and 112 bp or 212 and 218 bp for *AMELX* and *AMELY* loci, respectively. The amplicons generated from *AMELX* and *AMELY* are separated by electrophoresis. The observation of the *AMELX* fragment alone indicates a female, whereas the observation of both *AMELX* and *AMELY* fragments indicates a male (Figure 20.8). Nevertheless, DNA from primates and some rudiments can be amplified as well, but the amplicon sizes vary. The *AMEL* locus has been coamplified with other markers to provide a combined sex and identity test. Such combined tests have been used in historical D1S80 AFLP profiling (Chapter 19) and various current STR multiplex analyses.

Cases of *AMELY* null mutations have been reported where only the *AMELX* fragment is detected in *AMELY* null males. Many of them are phenotypically normal but present the *AMELX* sex types of females. Various interstitial deletions at the Y chromosome short arm and point mutations within the primer-binding sites have been identified as the possible cause of some *AMELY* null sex typing. The frequency of *AMELY* null males is rare, but is higher in Sri Lanka and India. However, the mutations may lead to incorrect identification of sex if the mutations occurred in the DNA evidence of criminal investigations such as sexual assaults or in the identification of human remains in mass disasters. In the case of a female individual, an *AMELX* mutation, which is rare, usually affects one copy of *AMELX*; however, the wild-type copy of the *AMELX* is detected. Thus, the *AMELX* mutation is less problematic in females than the *AMELY* mutation in sex typing in males.

21.3.2 Other Loci

To solve the incorrect identification problem with null *AMEL* mutations, additional candidate genetic markers are used in combination with *AMEL* analysis (Table 21.3). One useful marker is the sex-determining region Y (*SRY*) gene located on the Yp11.31 of the Y chromosome. *SRY* encodes a transcription factor that plays a role in the regulation of sex determination toward male development. The *SRY* protein contains a DNA-binding domain known as the HMG box. *SOX3*, an *SRY*-related HMG box-containing gene, has been identified at Xq27.1 of the X chromosome, which shares sequence homology with *SRY*. *SRY*-specific amplification is achieved using *SRY*-specific primers.

Another marker is the *TSPY* locus located on Yp11.2 of the Y chromosome. The *TSPY* locus encodes the testis-specific protein Y-encoded gene that is only expressed in the testis and may play a role in spermatogenesis. The *TSPY* structural gene is approximately 20 kb in length, consisting of 6 exons and 5 introns. The *TSPY* gene has multiple copies. A single Y chromosome contains several dozen copies of *TSPY* genes that are organized as a tandem array. The *TSPY* gene cluster shows copy number variation (CNV) between individuals in a population ranging from 23 to 64 copies. *DYS14* is a marker utilized for the characterization of the *TSPY* locus. The

DYS14 locus includes the partial sequence of the first exon and intron of the *TSPY* gene and shares approximately 98% sequence homology with other members of the *TSPY* gene cluster.

A *TSPY*-like (*TSPYL*) gene has been identified on the short arm of the X chromosome and is designated as *TSPYL2*. *TSPYL2* is a single-copy gene per X chromosome and is 6.3 kb in length, consisting of 7 exons and 6 introns. *TSPYL2* shares homology with *TSPY*. Additionally, several *TSPYL* genes have been identified on the autosomes. Primers are designed to amplify *TSPY* only and not amplify *TSPYL2* and other *TSPYL* gene sequences. The sensitivity of detecting *TSPY*, due to multiple copies, can be considerably higher than any single copy genetic maker such as *SRY*.

Furthermore, the DXYS156 locus, located at the pseudoautosomal region of both X and Y chromosomes, is another candidate marker used for sex typing. DXYS156 is a polymorphic pentanucleotide STR. The DXYS156 Y alleles have an additional adenine insertion in the repeat units of STR. Thus, the DXYS156 Y alleles can be distinguished from the DXYS156 X alleles because of the insertion. Additionally, the DXYS156 alleles are ethnically distributed. The DXYS156 profiling may potentially indicate the ethnic origin of an individual as an investigative lead.

Moreover, the steroid sulfatase (*STS*) gene encodes an enzyme that catalyzes the conversion of sulfated steroid precursors to biologically active steroids such as estrogens and androgens. The *STS* gene is located on the distal part of the short arm of the X chromosome within Xp22.31. The *STS* gene is 146 kb in length, consisting of 10 exons and 9 introns. The size of exons is between 120 and 160 bp long. However, the sizes of introns vary greatly, ranging from 102 bp to 35 kb. The entire coding sequence of the *STS* gene is 1542 bp in length. Deletions and point mutations of the *STS* gene have been associated with X-linked ichthyosis, a skin disease that affects males. There is an *STS* pseudogene, 100 kb in length, on the long arm of the Y chromosome known as *STSP1*. Although *STSP1* shares some sequence homologies with the *STS* gene, the pseudogene does not encode a functional gene. Primer sets are designed to amplify the target sequences within the first intron (35 kb in length) of the *STS* gene and its homologous *STSP1* pseudogene sequence. Although the homology of these target sequences of the *STS* gene and *STSP1* pseudogene is approximately 80%, amplicons with two different sizes are produced to identify *STS* and *STSP1* alleles for sex typing.

Bibliography

Y Chromosome Haplotyping

Alves, C., et al., Evaluating the informative power of Y-STRs: A comparative study using European and new African haplotype data. *Forensic Sci Int*, 2003, **134**(2–3): 126–133.

Ambers, A., et al., Autosomal and Y-STR analysis of degraded DNA from the 120-year-old skeletal remains of Ezekiel Harper. *Forensic Sci Int Genet*, 2014, **9**: 33–41.

Andersen, M.M., et al., Estimating stutter rates for Y-STR alleles. *Forensic Sci Int Genet Suppl*, 2011, **3**(1): e192–e193.

Andersen, M.M., et al., Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances. *Forensic Sci Int Genet*, 2013, **7**(3): 327–336.

Asamura, H., et al., Evaluation of miniY-STR multiplex PCR systems for extended 16 Y-STR loci. *Int J Legal Med*, 2008, **122**(1): 43–49.

Ayadi, I., et al., Combining autosomal and Y-chromosomal short tandem repeat data in paternity testing with male child: Methods and application. *J Forensic Sci*, 2007, **52**(5): 1068–1072.

Ayub, Q., et al., Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic Acids Res*, 2000, **28**(2): e8.

Balaresque, P., et al., Genomic complexity of the Y-STR DYS19: Inversions, deletions and founder lineages carrying duplications. *Int J Legal Med*, 2009, **123**(1): 15–23.

Ballantyne, K.N., et al., A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Sci Int Genet*, 2012, **6**(2): 208–218.

22

Single Nucleotide Polymorphism Profiling

22.1 Basic Characteristics of SNPs

Human genomes contain sequence polymorphisms, which are the variations in nucleotide sequences among individuals. One type of sequence polymorphism is called *single-nucleotide polymorphism* (*SNP*). An SNP constitutes a single-base-pair change originating from a spontaneous mutation that can be a base substitution, insertion, or deletion at a single site. An estimated 10 million SNPs may exist in the human genome, and among them, over a million SNPs have been identified. Thus, SNPs are the most frequent form of DNA polymorphism observed in humans. Most SNPs appear in noncoding regions and some SNPs are found in coding regions of genes (Figure 22.1).

The vast majority of SNPs are biallelic, although very rare triallelic and tetraallelic SNPs also exist. If an SNP originating from a spontaneous mutation occurs in the germ line, it can be inherited by offspring and can spread in the human population. As a result, both the wild type (the typical form of an allele occurring in nature) and a mutant allele are produced. This type of SNP is referred to as a biallelic SNP. Subsequently, a third mutation, a very rare event, may occur at the same nucleotide site. This SNP has a wild type and two different mutant alleles, which results in a triallelic SNP. It is also possible that a triallelic SNP is created by introducing two of the mutations simultaneously at a single site, which is also a very rare event.

Generally speaking, there are a number of advantages to utilizing SNP loci as markers for forensic applications. First, SNPs are abundant within the human genome; therefore, a sufficient number of SNP loci that are suitable for human identification can be selected. Second, in SNP analysis, amplified fragments (amplicons) are usually 50–100 base pairs in length and are smaller than that in short tandem repeat (STR) analysis (Chapter 20). Therefore, SNP profiling can be more useful than STR for analyzing degraded DNA samples in which genomic DNA is fragmented (Section 20.5.1). Third, SNP loci have lower mutation rates than STRs and are, therefore, useful for specialized human identification such as paternity testing. Lastly, it is possible to achieve high-throughput SNP analysis by utilizing multiplex SNP assays, which are amendable for automation.

The application of SNPs for forensic DNA analysis also has some disadvantages, in that SNP loci are not as polymorphic as STR loci. It is estimated that the analysis of 50–60 SNP loci is needed to achieve a similar level of population match probability (P_m—the lower the P_m, the less likely a match occurs between two randomly chosen individuals; see Chapter 25) using 13

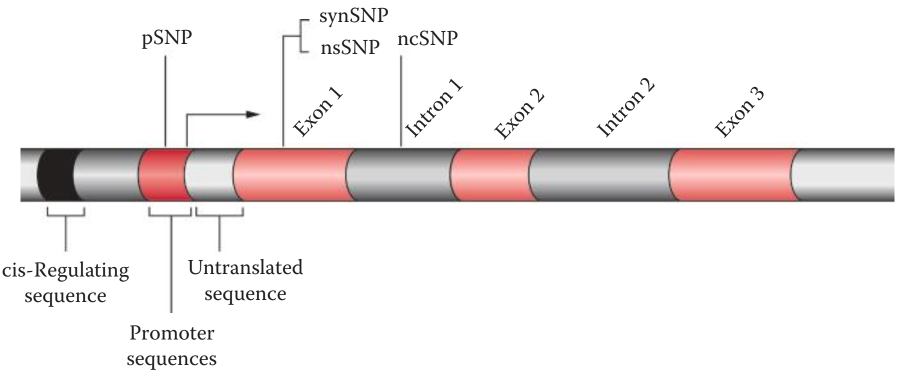

Figure 22.1 SNPs fall into several classes. Most reside in the noncoding regions of DNA and are designated as noncoding SNPs (ncSNPs). A subset of ncSNPs can also be found in introns. SNPs residing in exons are further divided into two types: The synonymous type (synSNP) is an exonic SNP that does not change the amino acid composition of the encoded polypeptide. Conversely, a nonsynonymous type (nsSNP) changes the encoded amino acid. SNPs in the promoter regions of the genome are known as promoter SNPs (pSNPs). Arrow: transcription start site. (© Richard C. Li.)

STR loci in CODIS. Moreover, most SNPs are biallelic; therefore, it is difficult to resolve mixed DNA profiles when a shared common allele exists in a mixture of more than one individual. Furthermore, most DNA databases contain STR profiles instead of SNP profiles, which means that it is not possible to carry out a database search for a matching SNP profile in cases where no suspect has yet been identified.

22.2 Forensic Applications of SNP Profiling

22.2.1 *HLA-DQA1* Locus

The first use of SNP-based profiling for forensic application involved sequence polymorphisms at the *HLA-DQA1* locus (formerly called the DQα locus). The *HLA-DQA1* gene is a member of the human leukocyte antigen (HLA) family, which contains a large number of genes involved in the immune response in humans. The *HLA-DQA1* locus is located within human HLA gene clusters on chromosome 6. The region tested for forensic use is highly polymorphic and is located at the second exon of the gene (Table 22.1).

22.2.1.1 DQα AmpliType and Polymarker Assays

The SNP profiling of the *HLA-DQA1* locus was carried out using the DQα AmpliType® kit. It was the first commercial kit, developed in the late 1980s by Cetus Corporation in Emeryville, CA. The *HLA-DQA1* panel can distinguish the following: alleles 1 (subtyped as 1.1, 1.2, or 1.3), 2, 3, and 4 (subtyped as 4.1 and 4.2/4.3, in which the 4.2 and 4.3 alleles are combined and cannot be distinguished). Therefore, a total of 28 possible genotypes from combinations of these alleles can be distinguished. Although the P_m of this SNP profiling is high (approximately 5×10^{-2}), it is useful as a preliminary test to quickly exclude suspects.

In addition to the *HLA-DQA1* locus, five additional loci—*LDLR, GYPA, HBGG,* D7S8, and *GC*—were utilized for forensic application in 1993 (Table 22.1). These loci were included in the second generation of the kit known as the AmpliType® PM PCR amplification and typing kit (also known as Polymarker), manufactured by Perkin-Elmer (Norwalk, CT). It consisted of one panel for the testing of *HLA-DQAl* and another panel for an additional five loci (Figure 22.2). Among these five additional loci, *LDLR, GYPA*, and D7S8 each have two alleles (designated

Table 22.1 Chromosomal Locations of SNP Loci Used in AmpliType® PM PCR Amplification and Typing Kit

Locus	Gene Product	Chromosome Location	Number of Alleles	Further Reading
HLA-DQA1	HLA-DQA1	6p21.3	7	Gyllensten and Erlich (1988)
LDLR	Low-density lipoprotein receptor	19p 13.1-13.3	2	Yamamoto et al. (1984)
GYPA	Glycophorin A	4q28-31	2	Siebert and Fukuda (1987)
HBGG	Hemoglobin G gammaglobin	11 p 15.5	3	Slightom et al. (1980)
D7S8	Anonymous	7q22-31.1	2	Horn et al. (1990)
GC	Group-specific component	4q11-13	3	Yang et al. (1985)

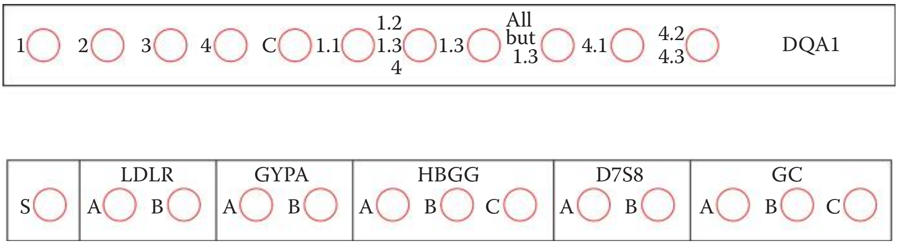

Figure 22.2 Panels of immobilized probes in the Polymarker kit. Top: *HLA-DQA1*. Bottom: additional five loci (*LDLR, GYPA, HBGG*, D7S8, and *GC*). C and S represent threshold control dots. (© Richard C. Li.)

A and B), while *HBGG* and *GC* each have three alleles (A, B, and C). As a result, the P_m of the Polymarker panels decreased to 10^{-4}.

The DNA profiles of the Polymarker system have been accepted in US courts. The Polymarker system has a number of advantages, particularly compared to variable number tandem repeat analysis (VNTR; see Chapter 19). First, the Polymarker system is a PCR-based method and is capable of analyzing a small amount of DNA sample (approximately 2 ng per analysis). Therefore, it is more sensitive than the method used in VNTR analysis. Second, the Polymarker system is an SNP assay with short amplicon sizes and, therefore, can analyze degraded DNA samples, which is not possible in VNTR analysis. Third, the Polymarker system is more rapid and less laborious than VNTR analysis. Lastly, the amplicon sizes of alleles at a given locus have identical lengths and, therefore, this assay does not exhibit preferential amplification as in amplified fragment length polymorphism (Chapter 19). However, the Polymarker system has its limitations compared to VNTR analysis. For example, compared to the VNTR analysis, the P_m of the Polymarker system is high, resulting in poor discrimination power in the comparison of two DNA profiles. Therefore, the Polymarker system is merely useful for excluding a suspect. Moreover, the SNP loci utilized in the Polymarker system are less polymorphic than VNTR loci. The limited number of alleles per SNP locus makes identifying the components of mixtures more difficult than with VNTR analysis in situations when contributors of a mixture sample share common alleles. For these reasons, the Polymarker system was replaced by STR profiling in the late 1990s.

22.2.1.2 Allele-Specific Oligonucleotide Hybridization

Both DQα AmpliType® and Polymarker kits utilize the *allele-specific oligonucleotide (ASO) hybridization* technique. This technique analyzes single-nucleotide variations, such as SNPs, at a given locus. It is based on the principle that ASO probes, usually 14–17 bases in length, hybridize to their complementary DNA sequences to distinguish known polymorphic alleles. ASO probes for multiple alleles at several loci can be arranged on the same panel to establish the presence or absence of specific alleles in PCR-amplified fragments of a DNA sample (Figure 22.3). Thus, the genotypes can be determined.

In the DQα AmpliType® and Polymarker kits, the oligonucleotides representing different alleles are immobilized to a solid matrix consisting of nylon membrane strips. Each immobilized probe at a particular site on the membrane is utilized to detect corresponding SNPs. Since the probe rather than the target DNA (as with regular blot format; see Chapter 9) is immobilized to the solid phase, the configuration used here is known as a *reverse blot* format.

The regions of the DNA in question are amplified by PCR. One of each pair of the primers is conjugated with biotin at the 5′ end (Figure 22.4). Thus, the amplicons are biotinylated for purposes of detection. This kit established multiplexing of a six-locus system, allowing simultaneous amplification of the *HLA-DQA1* locus along with *LDLR, GYPA, HBGG*, D7S8, and *GC* in

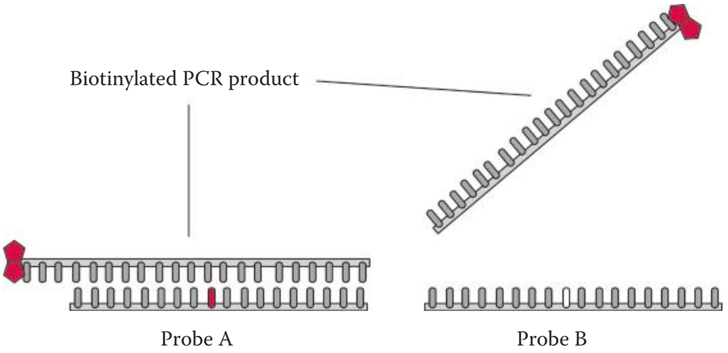

Figure 22.3 Hybridization with ASO. Two probes are incubated with the target DNA containing the SNP. Only the perfectly matched probe–target DNA can hybridize under optimal conditions.

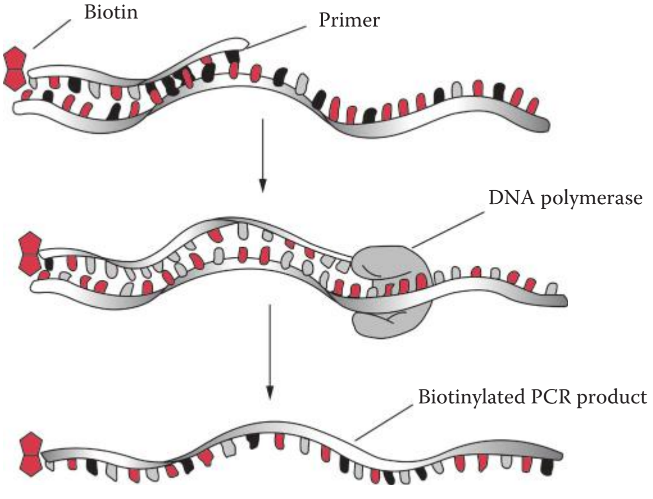

Figure 22.4 Amplification of DNA using biotinylated primers. The biotin is conjugated at the 5′ end of the primer. The amplified products are biotinylated. (© Richard C. Li.)

a single reaction for each sample. Following the denaturation of the PCR product to separate the two complementary DNA strands of the amplicons, the biotinylated strands are hybridized to the immobilized probes. Hybridization and washing conditions are established to ensure proper hybridization of the ASO probe and its target sequence. Unbound amplicons are washed away.

The presence of a PCR product bound to a specific probe can be detected by a colorimetric detection system (Figure 22.5). Since the amplicons are biotinylated, a horseradish peroxidase–conjugated streptavidin complex is utilized to bind to the biotinylated amplicons. The horseradish peroxidase then catalyzes the oxidation reaction of the colorless substrate, tetramethyl benzidine (TMB), into a blue precipitate at a designated location on the membrane strip, allowing the genotype of the sample to be determined (Figures 22.6 and 22.7).

The kit uses a threshold control (C dot in the *HLA-DQA1* panel and S dot in the panel of the other five loci) to distinguish between signal and background noise and to determine whether a

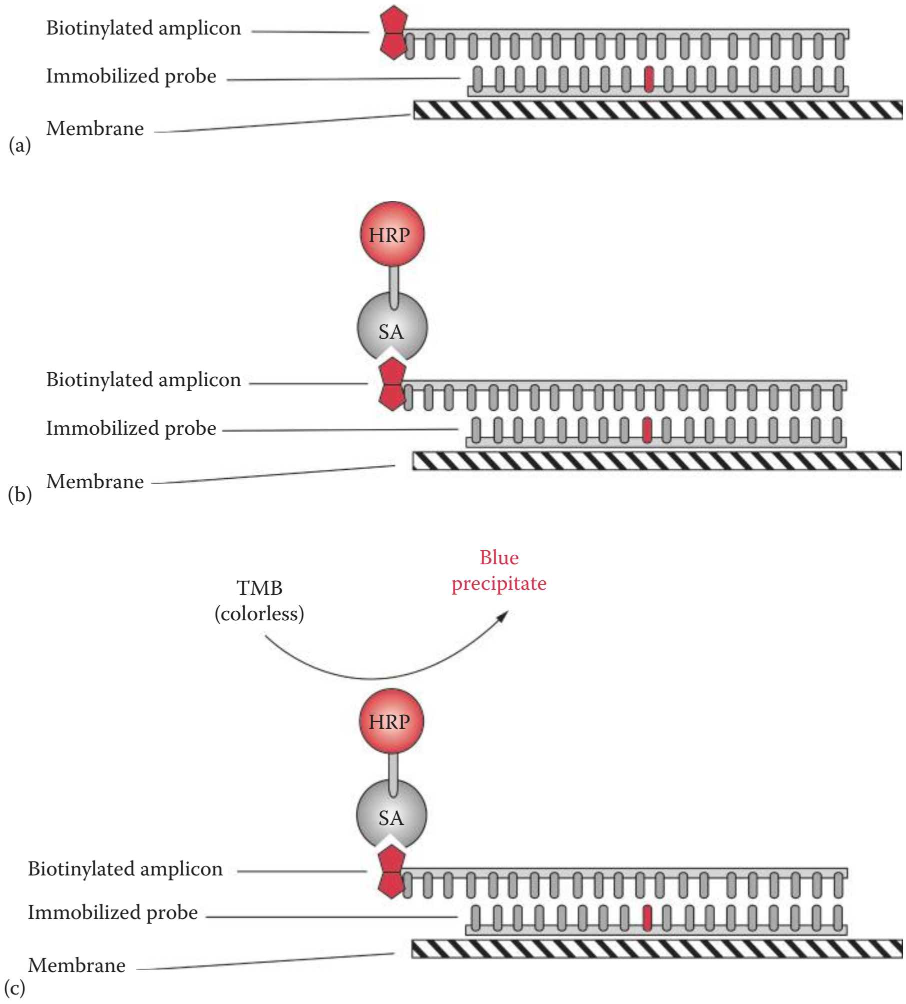
Biotinylated amplicon
Immobilized probe
Membrane
(a)
HRP
SA
Biotinylated amplicon
Immobilized probe
Membrane
(b)
TMB
(colorless)
Blue
precipitate
HRP
SA
Biotinylated amplicon
Immobilized probe
Membrane
(c)

Figure 22.5 Reverse blot assay. (a) The probe is immobilized onto a solid-phase membrane and hybridized with a biotinylated PCR product having the target sequence. (b) The detection of hybridization is carried out by streptavidin (SA) and horseradish peroxidase (HRP) conjugate. (c) Colorimetric reaction is catalyzed by HRP using TMB as a substrate. (© Richard C. Li.)

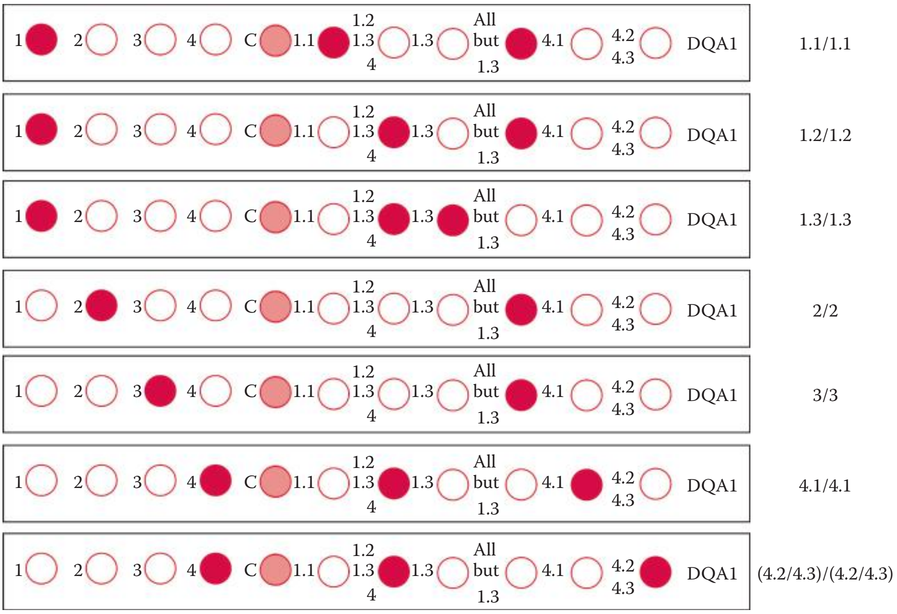
1
2
3
4
C
1.1
1.2
1.3
4
1.3
All
but
1.3
4.1
4.2
4.3
DQA1
1.1/1.1
1.2/1.2
1.3/1.3
2/2
3/3
4.1/4.1
(4.2/4.3)/(4.2/4.3)

Figure 22.6 DNA typing of homozygous alleles at the *HLA-DQA1* locus using the reverse blot assay (Polymarker kit). The genotype of each sample is noted on the right. (© Richard C. Li.)

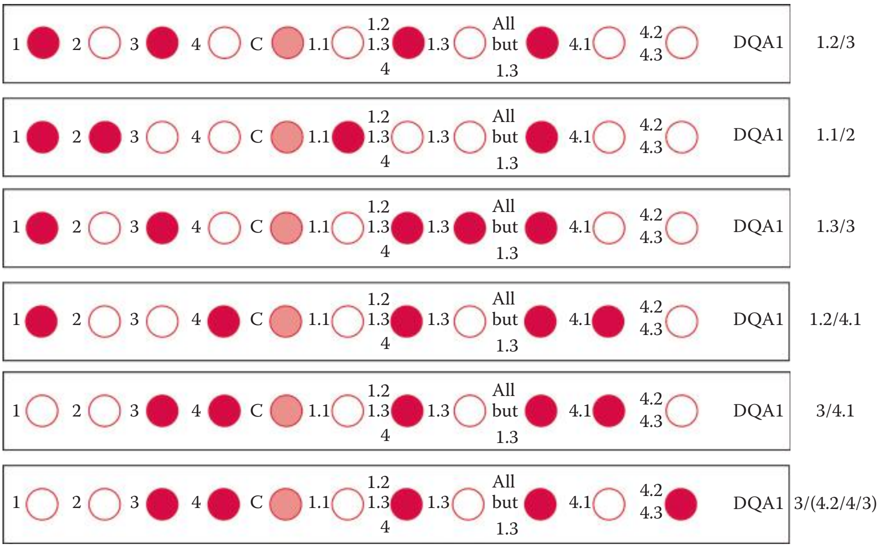
1
2
3
4
C
1.1
1.2
1.3
4
1.3
All
but
1.3
4.1
4.2
4.3
DQA1
1.2/3
1.1/2
1.3/3
1.2/4.1
3/4.1
3/(4.2/4/3)

Figure 22.7 DNA typing of heterozygous alleles at the *HLA-DQA1* locus using the reverse blot assay (Polymarker kit). The representative genotypes of the samples are noted on the right. (© Richard C. Li.)

sufficient amount of DNA has been amplified to detect all the alleles present in a sample. If the signal intensities of allele dots are greater than or equal to the threshold control, the alleles are considered true. If the allele dots are less intense than the threshold control, they are considered inconclusive for the determination of full genotypes because an allele may not have been detected due to a low level of DNA template.

22.2.2 Current and Potential Applications of SNP Analysis

22.2.2.1 Application of SNP Analysis for Forensic Identification

The applications of SNP analysis for forensic human identification are summarized in Table 22.2. Autosomal SNP panels can be used for the most common types of forensic testing, including analysis of degraded DNA samples. The SNP panels for mitochondrial DNA (mtDNA) profiling are useful for identifying human remains by comparing the DNA profiles in question to those of potential relatives. This approach may serve as an alternative method to the DNA sequencing method of mtDNA analysis, which is time-consuming and laborious. SNP loci on the Y chromosome are also potentially useful markers for paternity testing because of low mutation rates. SNP loci such as ancestry informative markers (AIM) can be used to determine the ethnic origins of questioned samples to generate leads for investigations.

Table 22.2 Examples of Forensic Applications of SNP Profiling

Testing	Candidate SNP Loci	Application	Further Reading
Identity	Autosomal SNPs	Human identification	Kidd et al. (2006); Sanchez et al. (2006)
		Human identification via degraded DNA	Budowle (2004)
	mtDNA SNPs	Human identification	Grignani et al. (2006)
	Y chromosome SNPs	Paternity testing	Hammer et al. (2006)
Biogeographical origin	Ancestry informative markers (AIMs)	Ethnic group identification	Frudakis et al. (2003); Shriver and Kittles (2004)
Physical characteristics	*MC1R* (melanocortin 1 receptor gene)	Hair color identification (investigative lead)	Grimes et al. (2001)
	P (gene has role in pigmentation)	Eye color identification (investigative lead)	Rebbeck et al. (2002)
Pathology	*KCNH2* (cardiac potassium channel gene)	Determining cause of sudden death from cardiac arrhythmia long QT syndrome	Lunetta et al. (2002)
	SCN5A (encodes cardiac sodium channel gene)	Determining cause of sudden death from cardiac arrhythmia long QT syndrome	Burke et al. (2005)
Toxicology	*CYP2C19*, *CYP2D6*, *CYP3A4*, *CYP2E1* (drug metabolizing enzyme genes)	Investigation of drug overdose (including death)	Kupiec et al. (2006)

22.2.2.2 Potential Applications of SNPs for Phenotyping

One potential application of SNP analysis is in determining phenotypic information, also known as *phenotyping* (Table 22.2). The relevant SNP loci are usually *nonsynonymous SNPs* (*nsSNPs*); they reside in the exon and change the encoded amino acid, which leads to an altered phenotype. Phenotyping of a questioned sample can reveal physical characteristics of an individual, such as hair and eye color, to provide leads for investigations. A number of SNPs residing within the melanocortin 1 receptor gene (*MC1R*) are associated with red hair, fair skin, and freckles, while SNPs residing within the *P* gene, which play a role in pigmentation, are associated with eye color variations.

Phenotyping can also be employed in the area of forensic pathology. Cardiac arrhythmia long QT syndrome (LQTS) can cause sudden death. A number of LQTS-associated SNPs—for example, SNPs in *KCNH2* and *SCN5A* genes—have been shown to correlate to such deaths. Thus, these SNPs are potentially useful for investigating the causes of death. Finally, phenotyping also has applications in forensic toxicology. A number of SNPs in genes, such as *CYP2D6*, that are responsible for metabolizing drugs can serve as potential markers for postmortem investigations of drug-overdose cases.

22.3 SNP Techniques

Over the years, various techniques of SNP analysis have been developed and can be divided into several groups based on the mechanisms used: allele-specific hybridization, primer extension, oligonucleotide ligation, and invasive cleavage. In *allele-specific hybridization*, allele discrimination is based on an optimal condition allowing only the perfectly matched probe–target hybridization to form. *Primer extension* methods are based on the ability of DNA polymerase to incorporate specific deoxynucleotides (dNTPs) complementary to the sequence of the template DNA. *Allele-specific oligonucleotide ligation* is based on the condition that only the allelic probe perfectly matched to the target is ligated. In the *invasive cleavage* method, allelic discrimination is based on DNA sequence-specific cleavage by endonucleases. A number of detection methods can be utilized in SNP analysis, such as the measurements of fluorescence, luminescence, and molecular mass. Most assays are carried out in solutions or on solid matrices such as glass slides, chips, or beads. Table 22.3 summarizes the representative assays for SNP typing.

For decades, Sanger sequencing, using chain-termination chemistry, has been the standard method for DNA sequencing (Chapter 23). In recent years, *next-generation sequencing* (NGS), a rapidly developing technology, has had a profound impact on biology. Compared to Sanger sequencing, NGS is advantageous in that it can achieve substantially higher throughput, and at lower cost, than the Sanger method. For example, a human genome can now be sequenced in several days using NGS technologies. Although NGS is not yet widely utilized for forensic applications, it has great potential for forensic DNA analysis, particularly in SNP analysis.

22.3.1 Next-Generation Sequencing Technologies

NGS technologies have two categories of application: *de novo* sequencing and resequencing. In *de novo* genome sequencing, uncharacterized genomes or characterized genomes with substantial structural variations are sequenced. Sequence reads are assembled without any reference sequence. In *resequencing* applications, characterized genomes are sequenced. Sequence reads are assembled against an existing reference sequence to identify sequence polymorphisms. Thus, resequencing can potentially be used for forensic applications to detect polymorphisms associated with human identification. *Target resequencing* is a useful method of resequencing that can be utilized for forensic applications. Prior to sequencing, the genomic regions of interest from a DNA sample are selectively isolated through a method known as *enrichment*. Several target-enrichment strategies have been developed. PCR is the most widely used enrichment method.

Table 22.3 Techniques for SNP Assay

Basis of Technique	Representative Assay	Detection	Format	Notes	Further Reading
Allele-specific hybridization	Reverse Blot	Colorimetry	Membrane based	Useful for screen assay	Saiki et al. (1989)
	LightCycler® (Roche)	Fluorescence	Solution based	Applicable for automation; limited multiplexing capability	Lareu et al. (2001)
	TaqMan® (Applied Biosystems)	Fluorescence	Solution based	Applicable for automation; limited multiplexing capability	De la Vega et al. (2005)
	Molecular beacons	Fluorescence	Solution based	Applicable for automation; limited multiplexing capability	Tyagi et al. (1998)
	GeneChip® (Affymetrix)	Fluorescence	Chip based	Designed for genotyping large number of SNPs; exceeds forensic needs	Mei et al. (2000)
Primer extension	SNaPshot™ (Applied Biosystems)	Fluorescence	Solution based	Applicable for capillary electrophoresis and multiplexing	Sanchez et al. (2006)
	PinPoint (Applied Biosystems)	Mass spectrometry	Solution based	Multiplex capability lower than SNaPshot™	Haff and Smirnov (1997)
	Arrayed primer extension (APEX)	Fluorescence	Chip based	Low reproducibility	Pastinen et al. (1997)
	Pyrosequencing	Fluorescence	Solution based	May be applicable for analysis of mixed profiles; not applicable for automation; limited multiplexing capability	Ronaghi et al. (1998)
Allele-specific oligonucleotide ligation	SNPstream® UHT (Orchid-Gene Screen)	Fluorescence	Chip based	Developed for mass disaster cases	Bell et al. (2002)
	SNPlex™ (Applied Biosystems)	Fluorescence	Solution based	Applicable for multiplexing; requires higher amount of target DNA template than other PCR-based methods	De la Vega et al. (2005)
Invasive cleavage	Invader® (Third Wave Technology)	Fluorescence	Solution based	PCR not required; large amount of target DNA required	Lyamichev et al. (1999)

Source: Adapted from Sobrino, B., M. Brion, and A. Carracedo, SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int, 2005. 154(2-3): p. 181–94.; Budowle, B., SNP typing strategies. Forensic Sci Int, 2004. 146 Suppl: p. S139–42.

The PCR-based approach is highly effective in targeting genomic regions that are small in size. Nevertheless, a typical NGS involves several major steps, including sample, library, and template DNA preparation; template amplification; sequencing and detection; and base calling; as well as data analysis (Figure 22.8).

22.3.2 DNA Samples, Sequencing Library, and Template Preparation

For most sequencing applications, micrograms of purified DNA are needed. This requirement is still a challenge for forensic casework applications where nanograms of DNA are often obtained. Human autosomal genome sequencing usually requires converting a DNA sample into a sequencing library. Two types of sequencing libraries are usually used: mate-pair libraries and fragment libraries. A mate-pair library is often used in *de novo* sequencing applications, while the fragment libraries can be used for resequencing and forensic applications. To construct a fragment library, a DNA sample is fragmented using mechanical methods such as sonication, nebulization, or shearing; endonuclease digestion; or a transposon-based method. Subsequently, sequencing adapters containing primer-binding sites for universal PCR primers are ligated to both ends of the DNA fragments. In resequencing applications, multiplex sequencing of pooled samples is often carried out, which can improve the efficiency and reduce the costs of sequencing. For multiplex sequencing, an index tag containing a bar code can be ligated to each DNA fragment to allow it to be identified after sequencing. Indexing can minimize the risk of sample mix-ups and contaminations during the sample preparation.

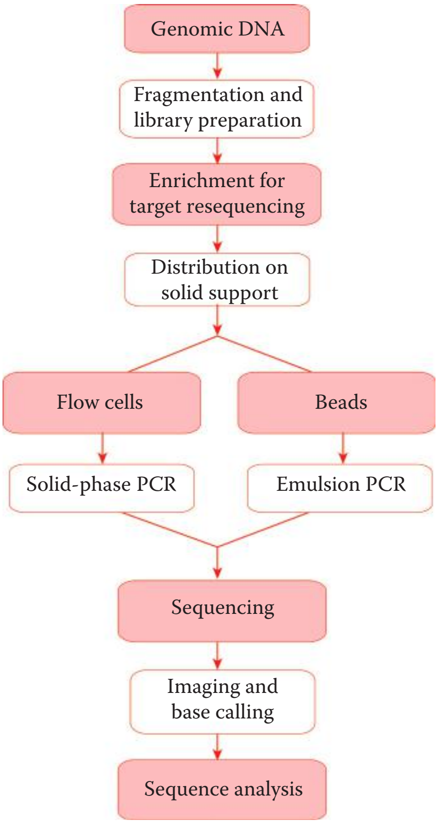
Genomic DNA
Fragmentation and library preparation
Enrichment for target resequencing
Distribution on solid support
Flow cells
Beads
Solid-phase PCR
Emulsion PCR
Sequencing
Imaging and base calling
Sequence analysis

Figure 22.8 NGS work flow for potential forensic applications. (© Richard C. Li.)

Most imaging systems of sequencing platforms are not equipped to detect single-molecule fluorescence. Thus, DNA templates are amplified using either emulsion or solid-phase PCR. In the *emulsion PCR*, single-stranded DNA templates are bound to beads. DNA-bound beads are suspended in a water-in-oil emulsion, where each bead is placed in a single aqueous droplet. DNA fragments are amplified using PCR. As a result, each bead is coated with thousands of copies of the same template sequence. Subsequently, beads carrying amplified DNA are deposited on a glass slide or into individual fiber-optic wells for sequencing. In *solid-phase PCR*, both forward and reverse primers are attached to a slide (Figure 22.9). DNA templates are hybridized to immobilized PCR primers. The extended primer bends and hybridizes to a second immobilized primer, forming a bridge. During the "bridge amplification," the single-stranded template is amplified to form clusters. Solid-phase amplification can produce millions of spatially separated template clusters with free ends, allowing universal sequencing primers to hybridize for sequencing.

22.3.3 NGS Chemistry

The amplified template fragments are then sequenced. At present, there are two major categories of NGS chemistry that are routinely used: *sequencing by synthesis* and *sequencing by ligation*. The *pyrosequencing* technology is one example of NGS using the sequencing-by-synthesis chemistry. During the pyrosequencing, each nucleotide substrate is introduced one at a time. Only the correct nucleotide corresponding to the template is incorporated, and a pyrophosphate is released (Figure 22.10a). Pyrophosphate is then converted to ATP, catalyzed by sulfurylase (Figure 22.10b). ATP is then utilized by luciferase to convert luciferin to oxyluciferin, and the reaction emits light that is detected by a charge-coupled device camera (Figure 22.10c). This technology allows the generation of sequences with long read lengths. *Cyclic reversible termination* is another example of NGS using sequencing-by-synthesis chemistry. In this method, chain terminators (fluorescent-labeled dideoxynucleotides) are used to extend a primer sequence complementary to the template DNA. Each of four nucleotides is labeled with a different fluorophore. After incorporation and fluorescence detection, the terminating and fluorescent moieties are cleaved and removed (Figure 22.11). As a result, the next sequencing cycle is carried out. This process is repeated until the sequence is completed. This is the most commonly used method throughout the field of NGS.

In the sequencing-by-ligation chemistry, probes (eight nucleotides in length) are utilized for ligation reactions. Each probe contains five specific nucleotides that are complementary to the template and three nucleotides that are universal. Probes containing all possible combinations of the first five nucleotides are utilized. Each probe is fluorescently labeled according to the first two bases of the probe so that it can be identified. During the first round of sequencing reaction, only the probe that is complementary to the template sequence can hybridize. The probe is then ligated to the 3′ end of the primer. After detection, three universal nucleotides, including fluorophore, are cleaved. The ligation process is repeated several times using a new set of probes each time. The newly made complementary DNA strand is then stripped off. In the next round of sequencing, new primers that are one nucleotide shorter than the previous primer are utilized, and the ligation reaction is repeated. In this round of sequencing, different nucleotide positions are read. This process is repeated for several rounds with different primers, until all nucleotides in the template have been sequenced twice. As a result, this sequencing method has a low error rate.

22.3.4 NGS Coverage

In NGS, sequencing reads are usually not distributed evenly over the genomic regions of interest. As a result, some nucleotides will be covered by fewer or more reads than the average. Therefore, multiple reads for each nucleotide are necessary to obtain a reliable sequence. The average number of times that each nucleotide in the genomic regions of interest is sequenced is known as the

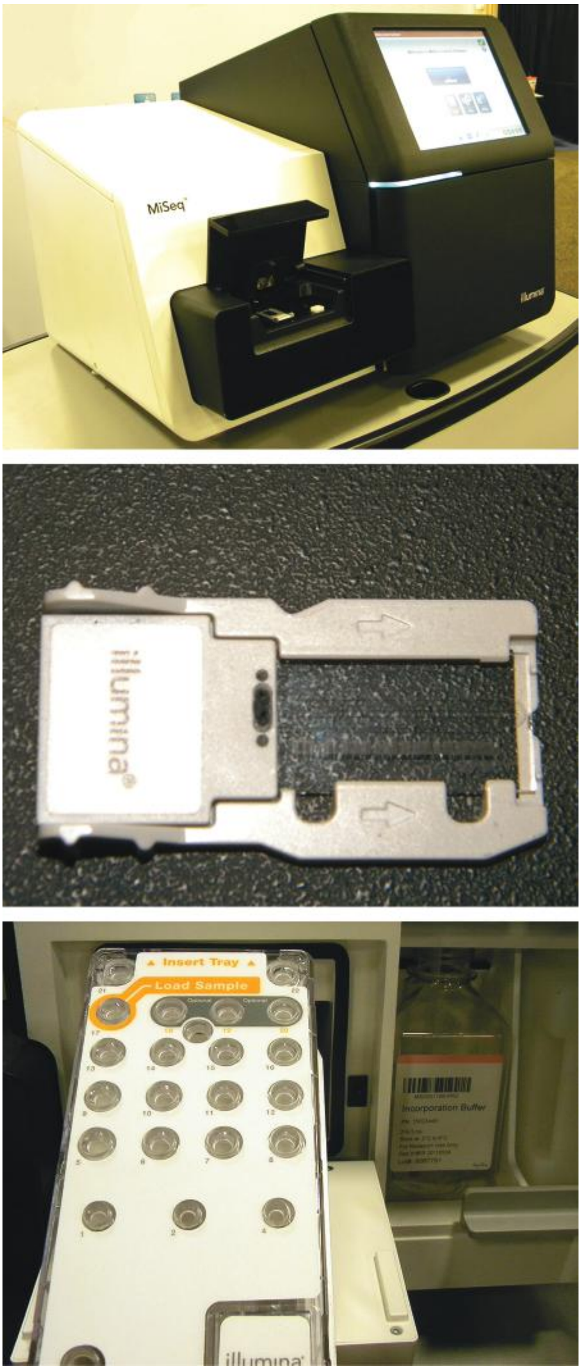
MiSeq
illumina
illumina®
Insert Tray
Load Sample
Incorporation Buffer
illumina

Figure 22.9 A photo of an NGS platform. A MiSeq (Illumna) desktop sequencer (top), a flow-cell device (middle) where DNA templates are distributed for PCR, and a reagent cartridge (bottom) are shown. (© Richard C. Li.)

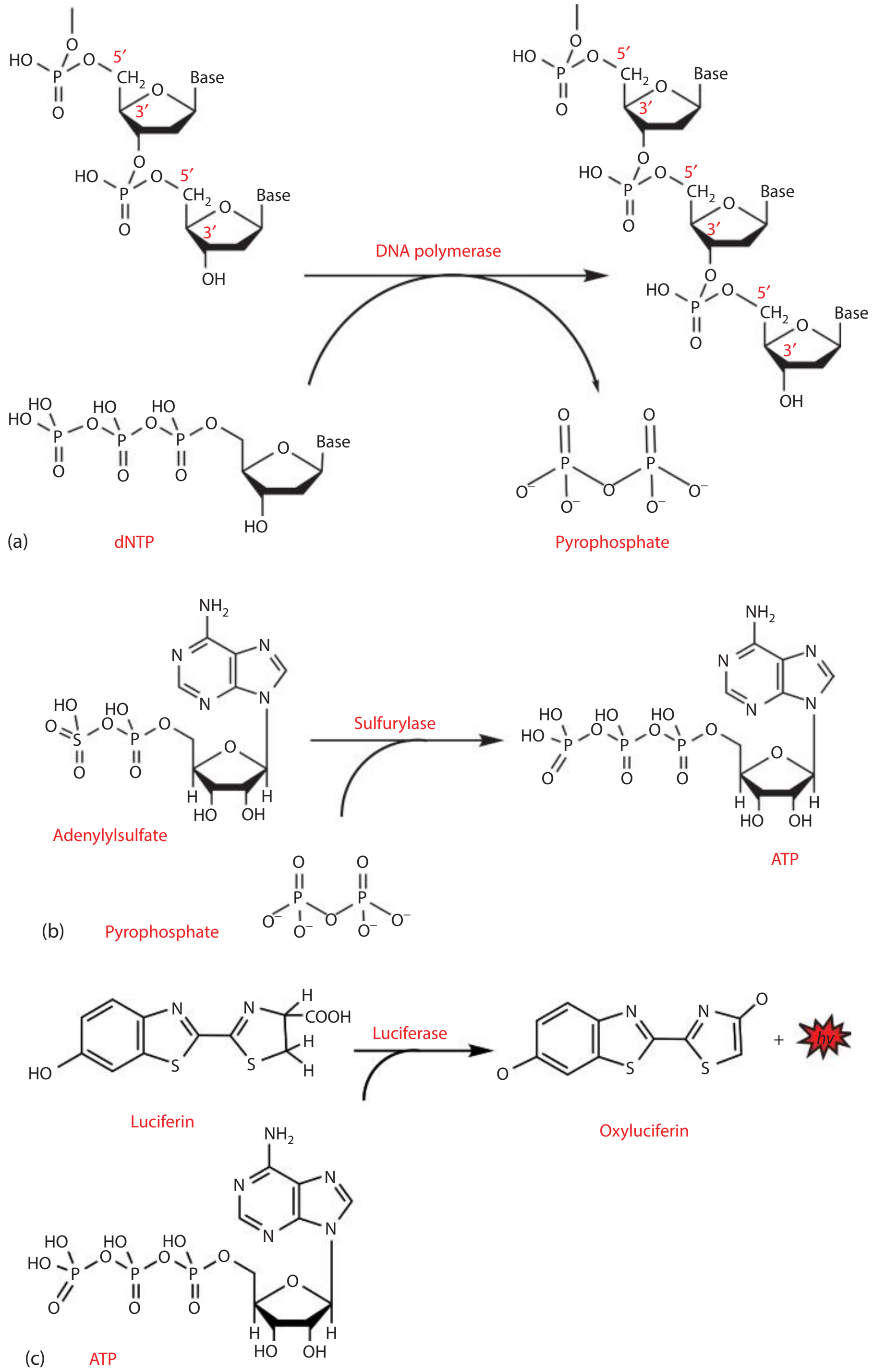
(a)
HO
P
O
5′
3′
CH_2
Base
OH
DNA polymerase
dNTP
Pyrophosphate
O^-
(b)
Adenylylsulfate
S
Sulfurylase
ATP
H
HO
OH
NH_2
N
Pyrophosphate
(c)
Luciferin
COOH
HO
Luciferase
Oxyluciferin
+
hν
ATP

Figure 22.10 Pyrosequencing chemistry. (a) During the pyrosequencing, a nucleotide substrate is incorporated, releasing a pyrophosphate. (b) Pyrophosphate is then converted to ATP. (c) ATP is utilized by luciferase, converting luciferin to oxyluciferin, which emits light.

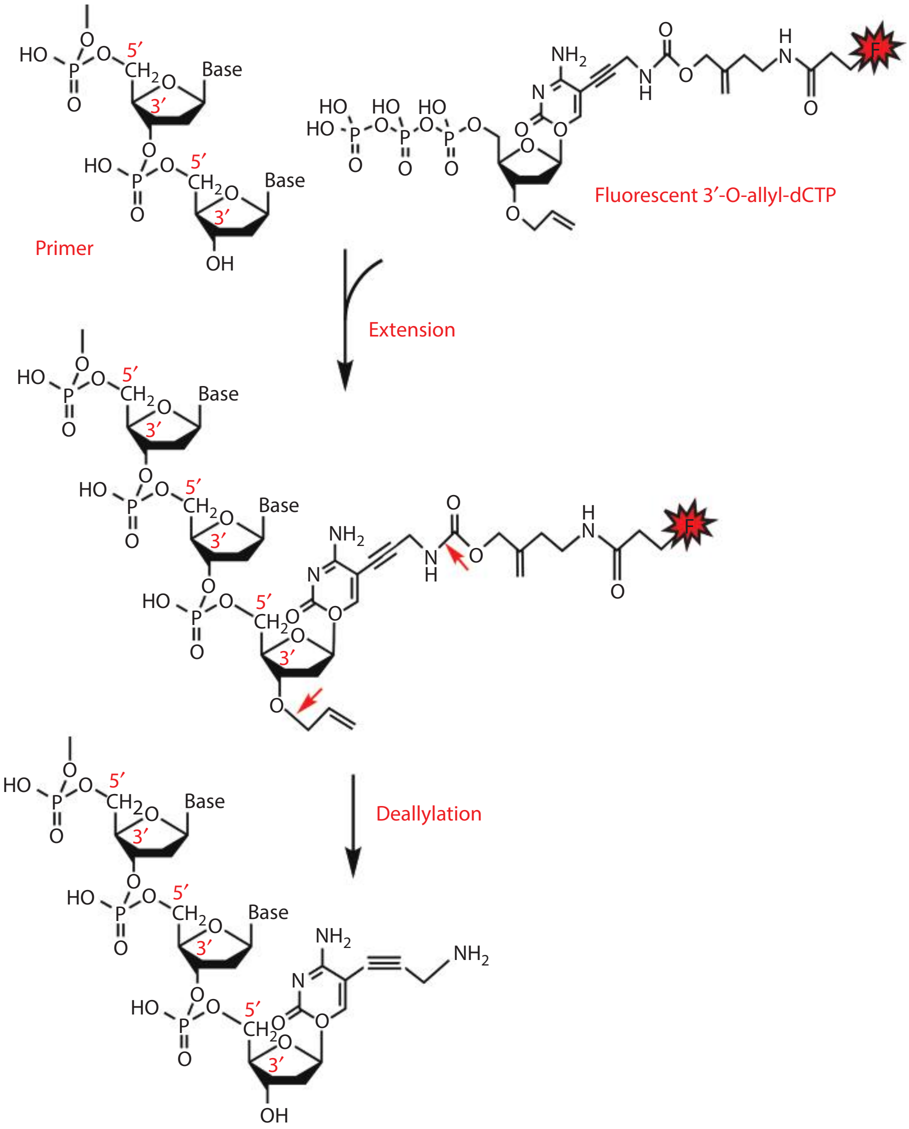
Primer
Fluorescent 3′-O-allyl-dCTP
Extension
Deallylation
Base
5′
3′
CH2
HO
OH
NH2
F

Figure 22.11 Cyclic reversible termination. A single sequencing cycle is shown. The incorporation of a C is illustrated as an example.

coverage or the *sequencing depth*. The level of coverage depends on the types of applications. For example, for human genome SNP analysis, the level of coverage is from 10 to 30. The coverage can be calculated using the following equation:

$$C = \frac{LN}{G}$$

where:

C = coverage
G = haploid genome length
L = read length
N = number of reads

Currently, NGS has not been widely used in forensic applications. One disadvantage of NGS is that the error rate of NGS is higher than that of Sanger sequencing. Although the error rate of NGS is not yet acceptable for forensic casework analysis, it can potentially be overcome by increasing the sequencing coverage.

Bibliography

Ackerman, M.J., D.J. Tester, and D.J. Driscoll, Molecular autopsy of sudden unexplained death in the young. *Am J Forensic Med Pathol*, 2001, **22**(2): 105–111.

Ali, M.E., et al., HLA-A, -B and -DRB1 allele frequencies in the Bangladeshi population. *Tissue Antigens*, 2008, **72**(2): 115–119.

Alkhayat, A., F. Alshamali, and B. Budowle, Population data on the PCR-based loci LDLR GYPA, HBGG, D7S8, Gc, HLA-DQA1, and D1S80 from Arabs from Dubai. *Forensic Sci Int*, 1996, **81**(1): 29–34.

Allah, R., L. Yang, and S.B. Li, SNPs and forensic DNA typing. *Fa Yi Xue Za Zhi*, 2007, **23**(5): 373–379.

Allwood, J.S. and S. Harbison, SNP model development for the prediction of eye colour in New Zealand. *Forensic Sci Int Genet*, 2013, **7**(4): 444–452.

Altshuler, D., et al., An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 2000, **407**(6803): 513–516.

Amigo, J., et al., The SNPforID browser: An online tool for query and display of frequency data from the SNPforID project. *Int J Legal Med*, 2008, **122**(5): 435–440.

Arnestad, M., et al., Prevalence of long-QT syndrome gene variants in sudden infant death syndrome. *Circulation*, 2007, **115**(3): 361–367.

Babol-Pokora, K. and J. Berent, SNP-minisequencing as an excellent tool for analysing degraded DNA recovered from archival tissues. *Acta Biochim Pol*, 2008, **55**(4): 815–819.

Bandelt, H.J. and A. Salas, Current next generation sequencing technology may not meet forensic standards. *Forensic Sci Int Genet*, 2012, **6**(1): 143–145.

Barbaro, A., et al., Genetic variability of the SNPforID 52-plex identification SNP panel in Italian population samples. *Forensic Sci Int Genet*, 2012, **6**(6): e185–e186.

Bell, B., et al., Distribution of types for six PCR-based loci; LDLR, GYPA, HBGG, D7S8, GC and HLA-DQA1 in central Pyrenees and Teruel (Spain). *J Forensic Sci*, 1997, **42**(3): 510–513.

Bell, P.A., et al., SNP stream UHT: Ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. *Bio Techniques*, 2002, **74**(Suppl 70-2): 76–77.

Birngruber, C., F. Ramsthaler, and M.A. Verhoff, The color(s) of human hair—Forensic hair analysis with SpectraCube. *Forensic Sci Int*, 2009, **185**(1–3): e19–e23.

Blake, E., et al., Polymerase chain reaction (PCR) amplification and human leukocyte antigen (HLA)-DQ alpha oligonucleotide typing on biological evidence samples: Casework experience. *J Forensic Sci*, 1992, **37**(3): 700–726.

Blanco-Verea, A., et al., Forensic validation and implementation of Y-chromosome SNP multiplexes. *Forensic Sci Int Genet Suppl*, 2008, **1**(1): 181–183.

Borsting, C. and N. Morling, Mutations and/or close relatives? Six case work examples where 49 autosomal SNPs were used as supplementary markers. *Forensic Sci Int Genet*, 2011, **5**(3): 236–241.

Borsting, C. and N. Morling, Reinvestigations of six unusual paternity cases by typing of autosomal single-nucleotide polymorphisms. *Transfusion*, 2012, **52**(2): 425–430.

Borsting, C., C. Tomas, and N. Morling, SNP typing of the reference materials SRM 2391b 1-10, K562, XY1, XX74, and 007 with the SNPforID multiplex. *Forensic Sci Int Genet*, 2011, **5**(3): e81–e82.

Borsting, C., C. Tomas, and N. Morling, Typing of 49 autosomal SNPs by single base extension and capillary electrophoresis for forensic genetic testing. *Methods Mol Biol*, 2012, **830**: 87–107.

Borsting, C., E. Rockenbauer, and N. Morling, Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard. *Forensic Sci Int Genet*, 2009, **4**(1): 34–42.

Borsting, C., H.S. Mogensen, and N. Morling, Forensic genetic SNP typing of low-template DNA and highly degraded DNA from crime case samples. *Forensic Sci Int Genet*, 2013, **7**(3): 345–352.

Borsting, C., et al., Performance of the SNPforID 52 SNP-plex assay in paternity testing. *Forensic Sci Int Genet*, 2008, **2**(4): 292–300.

23

Mitochondrial DNA Profiling

Forensic mitochondrial DNA (mtDNA) analysis is an important tool for human identification and is especially useful for identifying victims such as missing persons and individuals in mass fatality cases. Because mtDNA is maternally inherited, the mtDNA profiles of these individuals can be compared to those of their maternal relatives, and thus these individuals can be identified. Additionally, cells contain a much higher copy number of the mtDNA genome than that of the nuclear genome. Therefore, mtDNA testing is frequently used to analyze evidence samples, such as hair shafts, that contain low amounts of nuclear DNA. Furthermore, buried bones and decomposed tissues, in which nuclear DNA may be degraded, can be tested with mtDNA analysis.

23.1 Human Mitochondrial Genome

Mitochondria are cellular organelles that serve as the energy-generating components of cells (Figure 23.1). Each cell contains hundreds of mitochondria, which have their own extrachromosomal genomes separated from the nuclear genomes. Although each human mitochondrion contains multiple copies of the mtDNA genome, the exact copy number varies for each cell. However, it is estimated that hundreds of copies of the mtDNA genome exist in most cells.

23.1.1 Genetic Contents of Mitochondrial Organelle Genomes

The human mitochondrial genome was first sequenced by Fred Sanger's laboratory at Cambridge University and was published in 1981; it is known as the *Cambridge reference sequence* (CRS). The sequence was largely derived from a placental sample from an individual of European descent and also partially from HeLa cells (a cell line derived from cervical cancer cells), as well as from bovine cells. It was later discovered, by resequencing the original mtDNA sample, that the CRS contains substitution errors at 10 nucleotide positions. The *revised Cambridge reference sequence* (rCRS) was published in 1999 and presented corrections to these substitution errors. Additionally, CRS contains a cytosine dimer at nucleotide positions 3106 and 3107, which is in fact a single cytosine nucleotide. This error is not corrected in the rCRS in order to retain the original nucleotide numbering system of CRS and thus to avoid inconsistency with the previous literature. The human mitochondrial DNA genome is circular with no beginning and end, which can make sequence comparison a potential problem. Therefore, the rCRS, with its nucleotide numbering system, is used as a reference when aligning with other mitochondrial sequences for comparison purposes.

Organelle genomes are usually much smaller than their nuclear counterparts. The human mtDNA genome consists of 16,569 base pairs (bp) containing 37 genes (Figure 23.2). Thirteen of

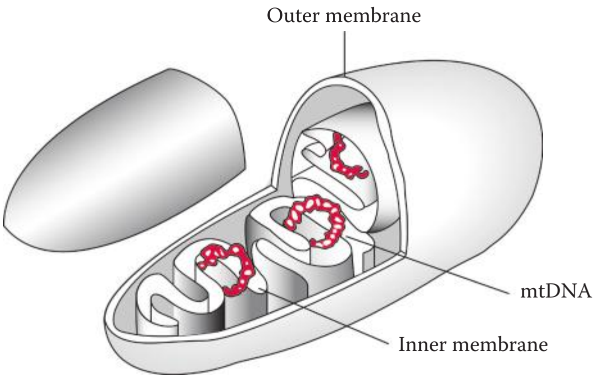
Outer membrane
mtDNA
Inner membrane

Figure 23.1 Mitochondrion. mtDNA, mitochondrial DNA. (© Richard C. Li.)

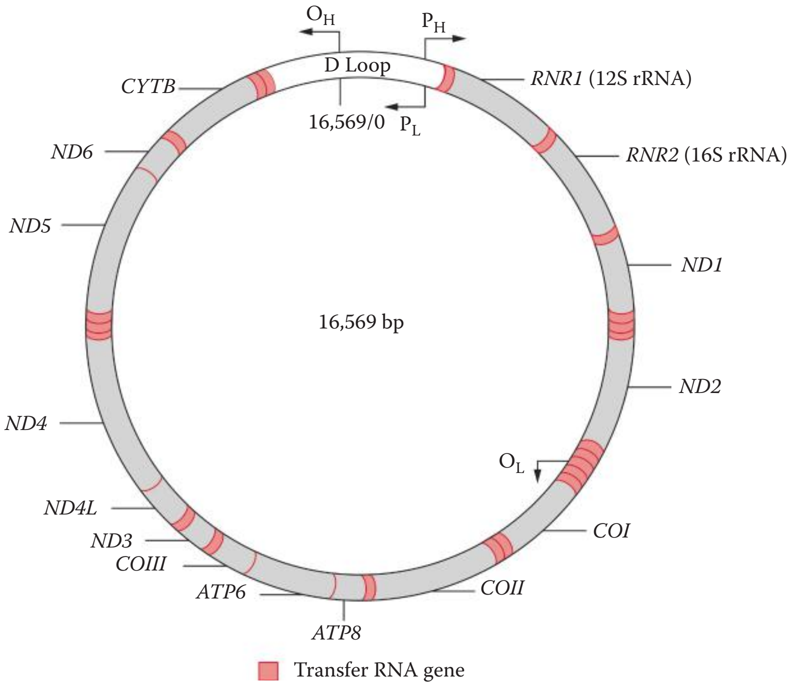
O_H
P_H
D Loop
16,569/0
P_L
CYTB
RNR1 (12S rRNA)
ND6
RNR2 (16S rRNA)
ND5
ND1
16,569 bp
ND2
ND4
O_L
ND4L
COI
ND3
COIII
ATP6
COII
ATP8
🟥 Transfer RNA gene

Figure 23.2 Human circular mitochondrial genome. The transcription direction for the H (heavy) and L (light) strands are indicated by arrows (P_H, P_L, respectively). The origins of replication are labeled O_H for the heavy strand and O_L for the light strand. The mitochondrial DNA genome encodes genes. ND, NADH coenzyme Q oxidoreductase complex; CO, cytochrome c oxidase complex; CYTB, cytochrome b; ATP, ATP synthase; rRNA, ribosomal RNA. Transfer RNA genes are shown as indicated. (© Richard C. Li.)

these genes code for proteins involved in the respiratory complex, a main energy-generating component in mitochondria. The other 24 specify noncoding RNA molecules required for the expression of the mitochondrial genome. The genes in the human mitochondrial genome are much more closely packed than in the nuclear genome and contain no introns. A *control region*, also known as a *displacement loop* (*D loop*), contains the origin of replication for one of the mtDNA strands but does not code for any gene products (Figure 23.2). An asymmetric distribution of

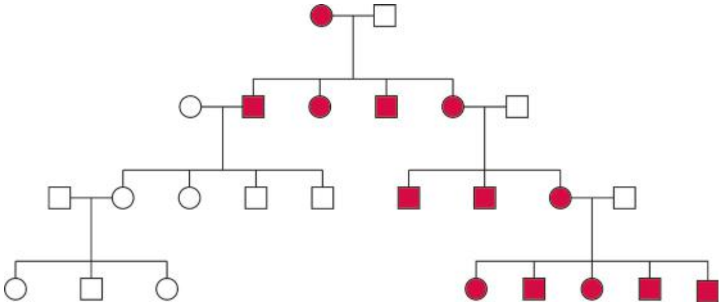

Figure 23.3 Pedigree of a human family showing inheritance of mtDNA. Females and males are denoted by circles and squares, respectively. Red symbols indicate individuals who inherited the same mtDNA.

nucleotides gives rise to light (L) and heavy (H) strands. The H strand contains a greater number of guanine nucleotides and a higher molecular weight in comparison to the L strand.

23.1.2 Maternal Inheritance of mtDNA

Maternal inheritance is typically observed for the mtDNA genome (Figure 23.3), which is inherited differently from nuclear genes. The inheritance of the mtDNA genome does not obey the rules of Mendelian inheritance and is thus called non-Mendelian inheritance.

The mitochondria of a spermatozoon are located in the midpiece (Chapter 14). At conception, only the head portion of a spermatozoon (containing a nucleus but no mitochondria) enters the egg. The fertilized egg contains the maternal mitochondria, which are transmitted to the progeny. Occasionally, the paternal mitochondria can enter the cell. However, paternal mitochondria in the spermatozoon that enter the egg are usually destroyed by the egg cell after fertilization (Figure 23.4). Therefore, the coinheritance of maternal and paternal mtDNA in a single individual is extremely rare in humans. The mtDNA sequence is identical for relatives within the same maternal lineage, a property that is useful when identifying individuals by comparing their mtDNA with that of maternal relatives.

Homologous DNA recombination (Chapter 25) has not been observed in the mtDNA genome. Thus, an mtDNA profile, also referred to as the *mitotype*, is considered a haplotype treated as a single locus. The mitochondrial genome has a higher mutation rate (up to 10 times higher) than its nuclear counterpart. The presence of mutations can be problematic in victim identification when comparing the mtDNA profiles of a victim with the relatives of the victim.

23.2 mtDNA Polymorphic Regions

23.2.1 Hypervariable Regions

The most polymorphic region of mtDNA is located within the D-loop (Figure 23.5). The three *hypervariable regions* in the D-loop are designated hypervariable region I (HV1: 16,024–16,365; 342 bp), hypervariable region II (HV2: 73–340; 268 bp), and hypervariable region III (HV3: 438–574; 137 bp). The most common polymorphic regions of the human mtDNA genome analyzed for forensic purposes are HV1 and HV2.

23.2.2 Heteroplasmy

Heteroplasmy occurs when an individual carries more than one mtDNA haplotype. Heteroplasmy may be observed with one kind of tissue and may be absent in other kinds of tissues; for example,

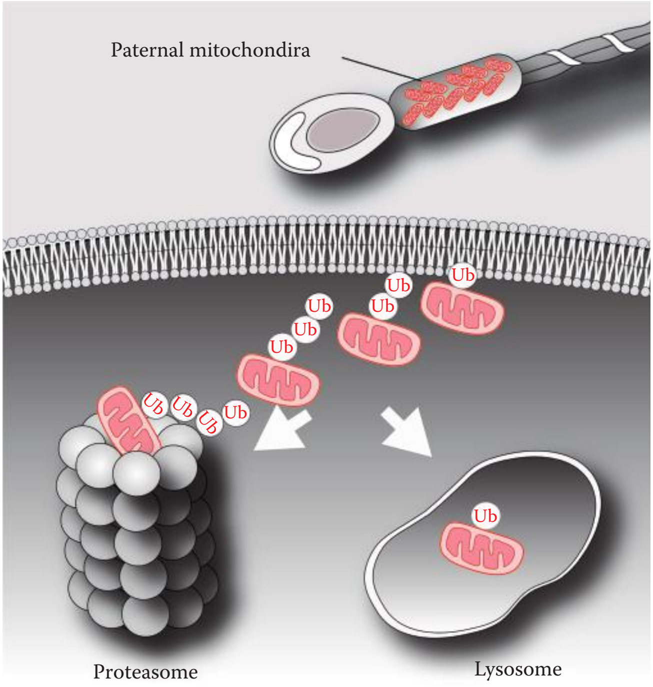

Figure 23.4 A model of uniparental mtDNA inheritance in humans. Paternal mitochondria occasionally enter the egg cell. Paternal mitochondria inside fertilized eggs are tagged by ubiquitin protein (Ub). It is proposed that the ubiquitination of sperm mitochondria leads to the degradation of paternal mitochondria in fertilized eggs. Tagged Ub can be recognized by proteasomes and lysosomes, which are cellular degradation machineries. A polyubiquitin chain with at least four ubiquitin units is needed to be recognized by the proteasomes. Ubiquitin with less than four ubiquitin units can be processed by the lysosomes. (© Richard C. Li.)

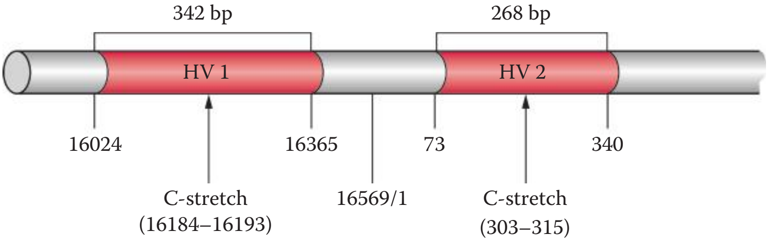

Figure 23.5 Hypervariable regions of the D-loop in mtDNA (with nucleotide positions). (© Richard C. Li.)

mtDNA heteroplasmy is commonly observed in hair samples (Chapter 4). Additionally, an individual may exhibit one mitotype in one tissue and a different mitotype in another. Thus, it is necessary to obtain and process additional samples to confirm the heteroplasmy when it is observed in a questioned sample but not in a known sample or vice versa. The two types of heteroplasmies are sequence and length heteroplasmy.

Sequence heteroplasmy is defined as the presence of two different nucleotides at a single position shown as overlapping peaks in a sequence electropherogram (Figure 23.6). Heteroplasmy usually occurs at one position, but on rare occasions it can be observed at more than one position. Hot spots for heteroplasmy have been documented at both HV1 and HV2 regions. Heteroplasmy may complicate the interpretation of mtDNA results, but its presence can also

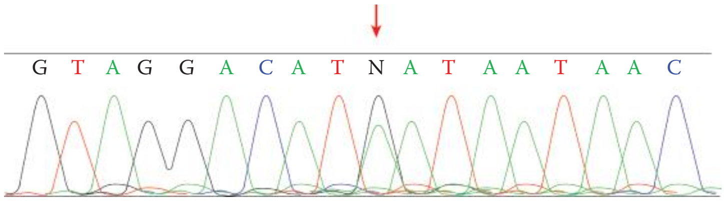

Figure 23.6 Electropherogram showing mtDNA sequence heteroplasmy at position 234R (A/G) as indicated by an arrow. N, unresolved sequence. (© Richard C. Li.)

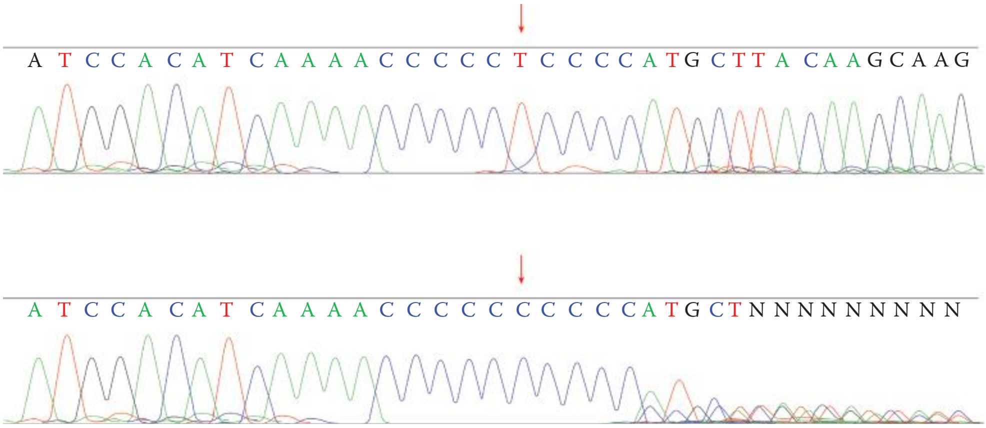

Figure 23.7 Length heteroplasmy. Electropherogram showing mtDNA length heteroplasmy at the C stretch of the HV1 region where position 16,189 is a T (top) and a C (bottom) as indicated by an arrow. N, unresolved sequence. (© Richard C. Li.)

improve the strength of a match. Both HV1 and HV2 of the human mtDNA D-loop region contain homopolymeric cytosine sequences known as *C stretches*. The HVI region contains a C stretch between nucleotide positions 16,184 and 16,193, interrupted by a thymine at position 16,189. If a base transition from T to C occurs at position 16,189 (a variant present in approximately 20% of the population), it results in an uninterrupted C stretch. A similar C stretch resides between positions 303 and 315 of the HV2 region.

Length heteroplasmies are often observed at the uninterrupted C stretches in sequencing, in which sequencing products with various lengths of polymeric cytosine residues are present. As a result, sequences downstream from the C stretch cannot be resolved (Figure 23.7). It is not clear whether the length heteroplasmy is due to replication slippage at the C stretches or results from the presence of a mixture of length variants in the cells. If length heteroplasmy occurs, alternative sequencing primers that anneal at the downstream of C stretches can be used to obtain the downstream sequences of the C stretches.

23.3 Forensic mtDNA Testing

23.3.1 General Considerations

mtDNA analysis is often used on samples derived from skeletal or decomposed remains. The surface of the sample should be cleaned to remove any adhering debris or contaminants. Bones and teeth are pulverized to facilitate extraction of the mtDNA (Chapter 5). Duplicate extractions (e.g., two sections of a single hair) are recommended if sufficient sample material is available.

mtDNA is extracted using a similar method to nuclear DNA (nuclear DNA is coextracted with mtDNA). The amount of mtDNA can therefore be estimated from the quantity of nuclear DNA obtained. Alternatively, mtDNA-specific quantization methods using real-time PCR (Chapter 6) can also be used to directly obtain measurements of mtDNA extracted.

For mtDNA sequencing, the analysis of both strands of the mtDNA in a given region must be performed to ensure accuracy. Due to the high sensitivity of mtDNA analysis, it is essential to minimize risks of contamination during the procedure. Contamination must be strictly monitored using proper controls such as extraction reagent blanks (Section 7.5.3) and amplification negative controls (samples containing all reagents except DNA template).

Finally, a positive control must also be used to monitor the success of the analysis. It should be introduced at the amplification step and remain through the sequencing process. A positive control consists of a DNA template of known sequence, such as DNA purified from the HL60 cell line.

23.3.2 mtDNA Screen Assay

One example of the assay for screening mtDNA variations is the allele-specific oligonucleotide (ASO) assay. It allows initial screening of mtDNA sequence polymorphisms and has the potential to reduce the number of samples required for mtDNA sequencing. This method is also useful for excluding or eliminating suspects from a case. However, HV1 and HV2 sequencing should be performed to obtain complete sequence information for the targeted HV regions to confirm a match.

The commercial Linear Array™ mtDNA HV1/HV2 region sequence typing kit (Roche Applied Sciences, Indianapolis) utilizes reverse ASO configuration with a panel of immobilized ASO probes that detect common polymorphic sites (Figures 23.8 and 23.9). The mtDNA is amplified at both HV1 (444 bp amplicon) and HV2 (415 bp amplicon) regions and the forward primers are biotin labeled at the 5′ ends of the oligonucleotides. Thus, the amplified PCR product (amplicon) is biotinylated. A horseradish peroxidase–conjugated streptavidin complex then binds to the biotinylated amplicon (see Chapter 9). Finally, colorimetric detection is carried out with tetramethylbenzidine (TMB) as the substrate to produce a colored precipitate at the designated location. The typing kit detects sequence variations in 19 positions within the HV1 and HV2 regions.

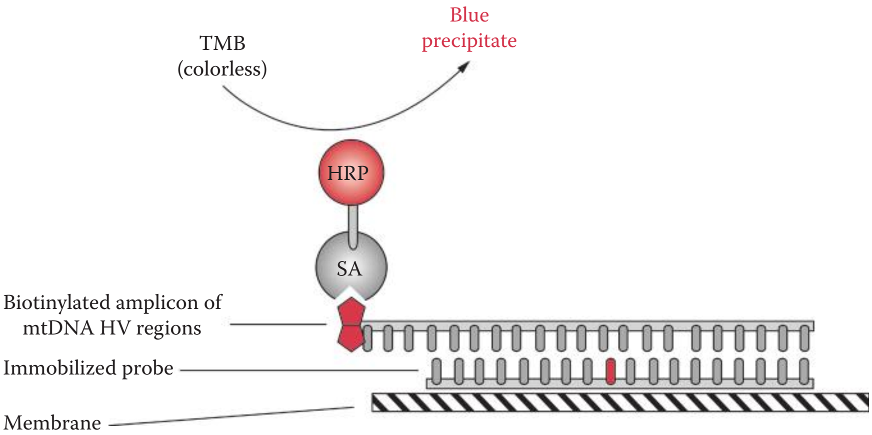

Figure 23.8 Reverse blot assay employed in mtDNA screen. A probe is immobilized onto a solid-phase membrane, and then hybridized with a biotinylated amplicon of the mtDNA HV sequences. Hybridization is detected by a streptavidin (SA) and horseradish peroxidase (HRP) conjugate. A colorimetric reaction is catalyzed by HRP using tetramethylbenzidine (TMB) as a substrate. (© Richard C. Li.)

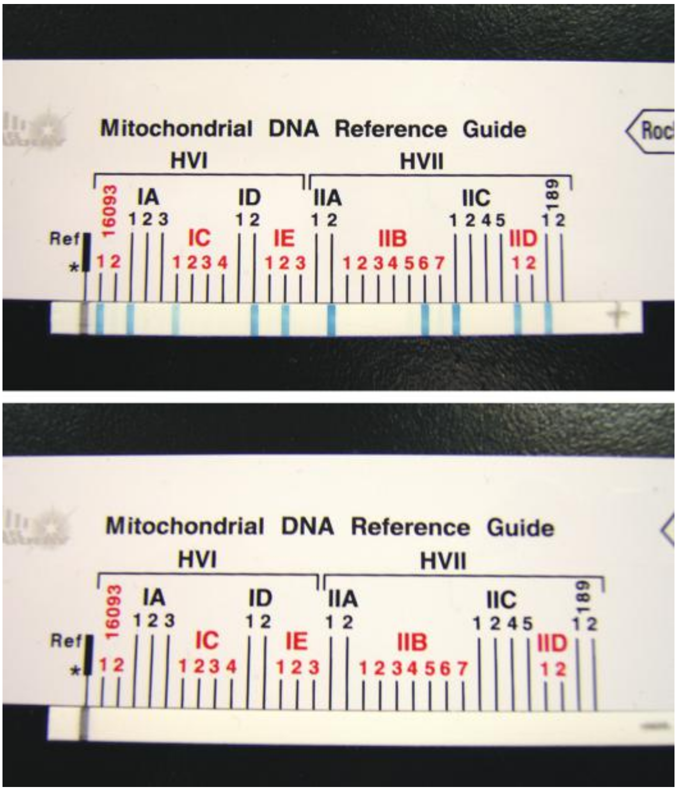

Figure 23.9 Linear array mtDNA assay results (top) and negative control (bottom). (© Richard C. Li.)

23.3.3 mtDNA Sequencing

To sequence a specific region of mtDNA, a combination of PCR amplification and DNA sequencing techniques is utilized to reduce the time and labor needed to obtain DNA sequences from genomic DNA templates. mtDNA sequencing usually consists of (1) PCR amplification, (2) DNA sequencing reactions, (3) separation using electrophoresis, and (4) data collection and sequence analysis (Figure 23.10). Chapter 25 describes the evaluation of the strength of the results via statistical analysis.

23.3.3.1 PCR Amplification

The extracted DNA samples must be amplified to yield sufficient quantities of template for sequencing reactions. PCR amplification of all or a part of the D-loop region can be carried out with various primer sets. If a sample contains intact mtDNA, the HV1 and HV2 regions can be amplified as two amplicons, each of about 350–400 bp in length. If an mtDNA sample is fragmented due to degradation, the hypervariable regions can be amplified as smaller amplicons. PCR amplification of mtDNA is usually done in 34–38 cycles. Protocols for highly degraded DNA specimens sometimes require 42 cycles. The use of higher PCR cycle numbers can improve the yield of the amplicons.

Following mtDNA amplification, a purification step is necessary to remove excess primers and deoxynucleotide triphosphates (dNTPs). This step can be performed using filtration devices such as a Microcon® to remove small molecules from the sample or using nuclease digestion with shrimp alkaline phosphatase or exonuclease I to degrade remaining primers and dNTPs. The concentration of the amplicons is important for an optimal sequencing reaction in the next phase of mtDNA sequencing. The quality and quantity of the mtDNA amplicon must be evaluated to confirm the presence or absence of amplicons and their concentrations. This can be done

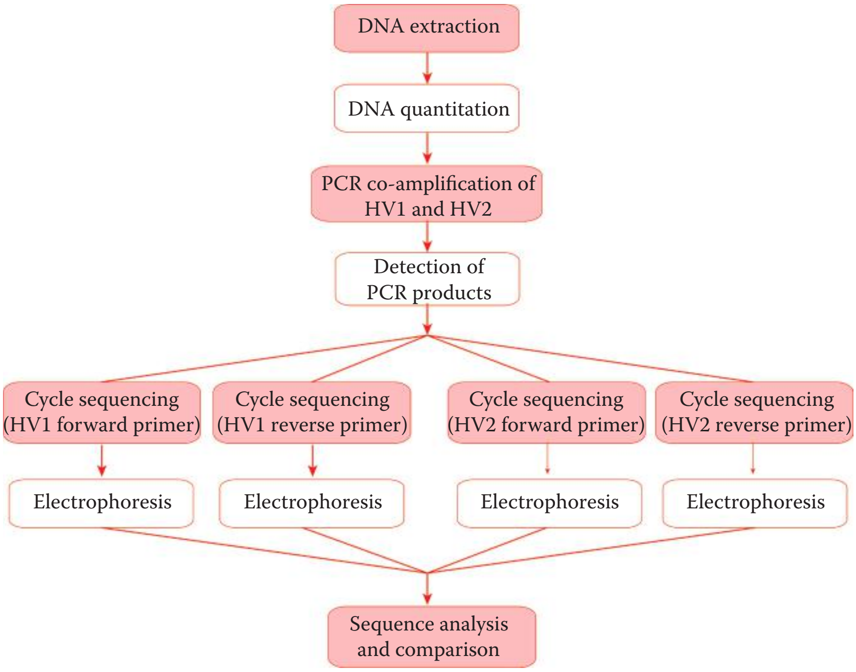
DNA extraction
DNA quantitation
PCR co-amplification of HV1 and HV2
Detection of PCR products
Cycle sequencing (HV1 forward primer)
Cycle sequencing (HV1 reverse primer)
Cycle sequencing (HV2 forward primer)
Cycle sequencing (HV2 reverse primer)
Electrophoresis
Electrophoresis
Electrophoresis
Electrophoresis
Sequence analysis and comparison

Figure 23.10 Sanger sequencing work flow for the mtDNA HV regions. (© Richard C. Li.)

using an agarose yield gel to visualize the amplicons of the sample or via capillary electrophoresis using a modular microfluidic chip device (Figure 8.12), a more informative method, for quantifying amplicons.

23.3.3.2 DNA Sequencing Reactions

The best-known DNA sequencing techniques are the *chain-termination* method and the *chemical degradation* method developed by Sanger and Gilbert (who shared the Nobel Prize for their work), respectively, in 1977. Over the years, the chain-termination method has become more common because it is suitable for automation and does not require the toxic chemicals necessary for the chemical degradation method.

23.3.3.2.1 Chain-Termination or Sanger Method

Using the chain-termination method, an oligonucleotide primer that can anneal to a single-stranded DNA template is utilized. A sequencing reaction contains DNA polymerase and the four dNTPs in order to carry out extension. The reaction also contains small quantities of dideoxynucleotide triphosphates (ddNTPs, Figure 23.11). Thus, a sequencing reaction involves a combination of extension and termination of the chain (Figure 23.12). If a ddNTP molecule is incorporated into a growing DNA chain, the absence of a 3′ OH group in the ddNTP molecule prevents the formation of a phosphodiester bond and thus disrupts the extension of the oligonucleotide chain.

The ratio of ddNTPs to dNTPs has been optimized to result in a collection of DNA fragments varying in length by one nucleotide from the primer length to the full length of the sequencing reaction product. As a result, the products of the sequencing reaction consist of a pool of various lengths of oligonucleotide chains terminated by ddNTPs. By using the four different ddNTPs, populations of DNA fragments are generated that terminate at positions occupied by every A, C, G, or T in the template strand.

The sequencing product of chain termination can be labeled with the *dye-terminator* system, in which the terminator is labeled, or with the *dye-primer* system, in which the primer is labeled.

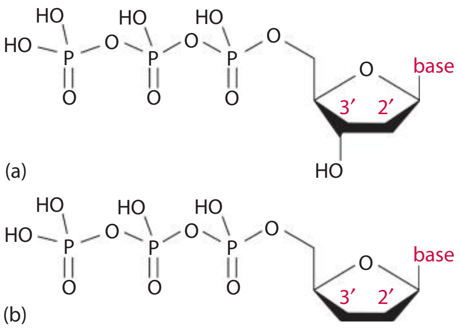
HO
HO
P
O
HO
P
O
HO
P
O
O
O
base
3′
2′
HO
(a)
(b)

Figure 23.11 Chemical structures of dNTP and ddNTP: (a) dNTP; (b) ddNTP. Both hydroxyl groups attached to the 2′ and 3′ carbons of ribose are replaced by hydrogens.

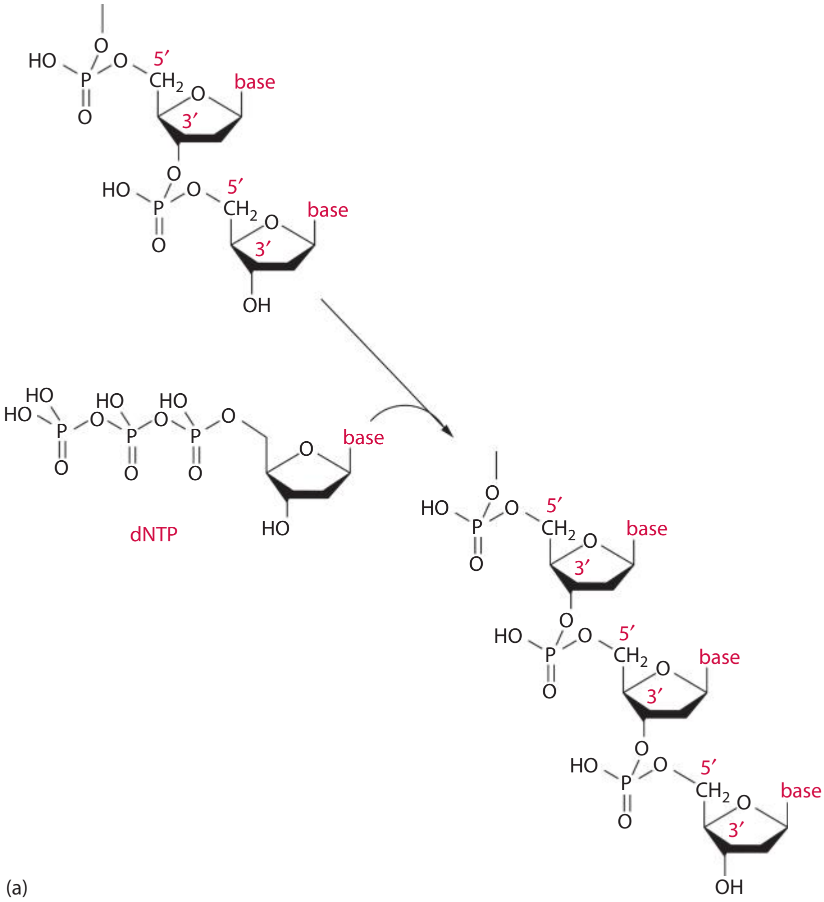
HO
P
O
5′
CH_2
base
3′
OH
dNTP
HO
(a)

Figure 23.12 Competition between extension and termination. (a) Growing chain in extension.

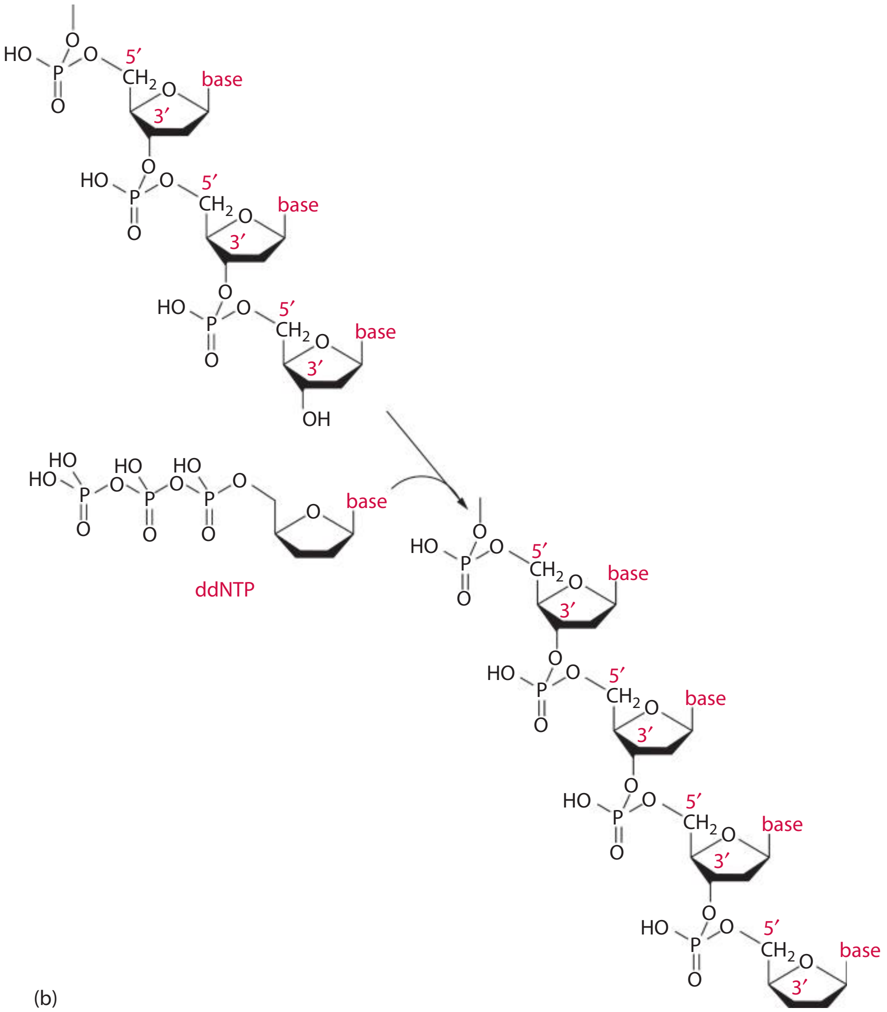

Figure 23.12 (Continued) (b) DNA strand with labeled ddNTP incorporated into a growing DNA chain that interrupts the incorporation of new nucleotides. The ddNTP usually is labeled with a fluorescent dye used for detection.

The dye-terminator system is commonly used for mtDNA sequencing in forensic laboratories. With the dye-terminator system, the ddNTPs are labeled with four different fluorescent dyes, each with a distinct spectrum. Thus, the sequencing with all four ddNTPs can be carried out in a single reaction. The labeled products of sequencing reactions are then resolved during electrophoresis and the sequencing data can be collected (Section 23.3.3.).

23.3.3.2.2 Cycle Sequencing

The chain-termination reaction is carried out using a *cycle sequencing* technique commonly used in forensic laboratories for mtDNA sequencing. Cycle sequencing, developed in the late 1980s, utilizes thermal cycling to generate a single-stranded template for chain-termination sequencing reactions. The application of thermal cycling in a sequencing reaction greatly increases the signal intensity and thus the sensitivity of the sequencing.

The sequencing reactions are carried out with multiple rounds of thermal cycling. Each cycle consists of three steps: denaturation of the double-stranded DNA template, annealing of a sequencing primer to its target sequence, and the extension of the annealed primer by DNA polymerase. Cycle sequencing utilizes only a single primer per reaction. During the extension of cycle sequencing, the extension of the strand is terminated with the incorporation of a ddNTP (Figure 23.13). The resulting partially double-stranded hybrid, consisting of the full-length template strand and its complementary chain-terminated product, is denatured during the first step of the next cycle, thereby liberating the template strand for another round of annealing, extension, and termination.

23.3.3.3 Electrophoresis, Sequence Analysis, and Mitotype Designations

The cycle sequencing products can be separated using electrophoresis in a 4% polyacrylamide denatured gel or a POP-6 polymer (Applied Biosystems) as the matrix for capillary electrophoresis (Figure 23.14). Following data collection, sequence data analysis can be performed with the Sequencher™ software (Gene Codes Corporation, Ann Arbor). Figure 23.15 shows sequence data.

Sequencing of a region of the mtDNA genome should be performed twice. Additionally, both strands of a region of the mtDNA genome should be sequenced to reduce ambiguities in sequence determination. The sequences of evidence samples and reference samples such as that of the victim or suspect can be compared. The nomenclature used in reporting should be compatible with International Union of Pure and Applied Chemistry (IUPAC) codes.

23.3.3.3.1 Reporting Format

The rCRS is used as a reference standard to facilitate the designation of mitotypes. For reporting purposes, sequence differences relative to the rCRS are listed in data format. When a difference between an individual's sequence and that of the rCRS sequence is observed, only the position (designated by a number) and the nucleotide differing from the reference standard are recorded. In this format, nucleotides identical to the rCRS are not listed. For example, at position 228 (HV2), the rCRS has a G. If a mitotype carries an A at position 228, the individual's mtDNA sequence is described as 228A. If an unresolved sequence ambiguity is observed at a position,

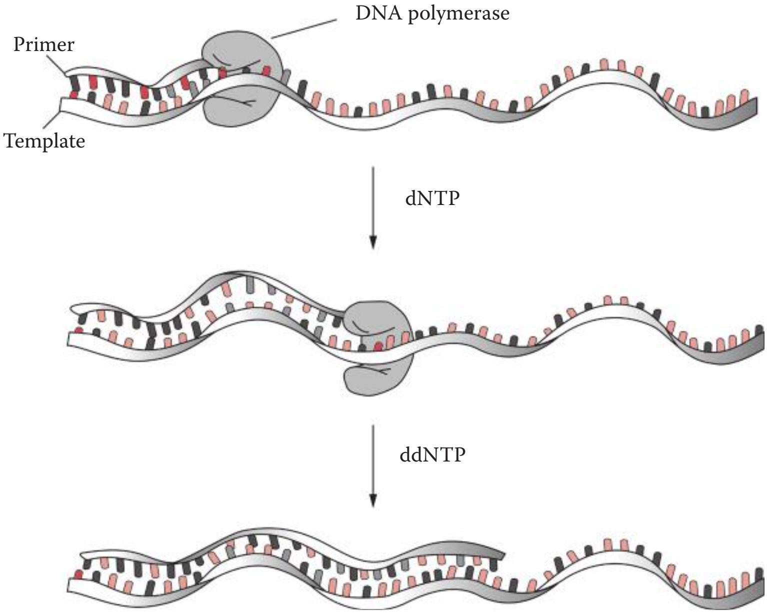

Figure 23.13 Cycle sequencing reaction. The reaction requires DNA polymerase, a template, and a primer. During DNA synthesis, the dNTP is incorporated by a new phosphodiester bond with the primer. The incorporation of ddNTP blocks further DNA synthesis of the growing chain. (© Richard C. Li.)

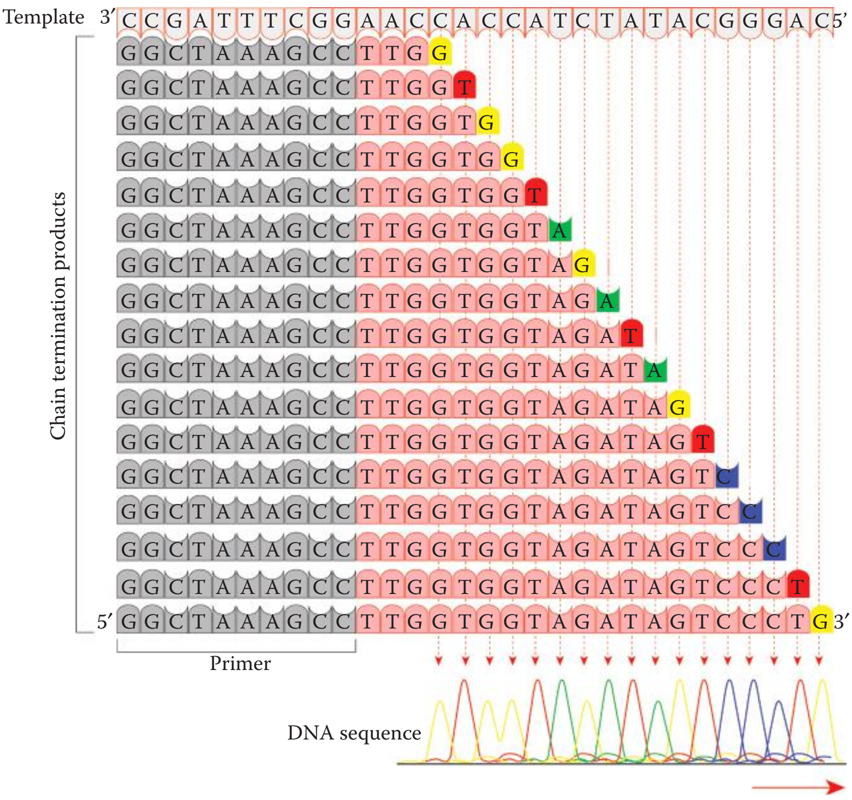
Template 3′ CCGATTTCGGAACCACCATCTATACGGGAC 5′
Chain termination products
GGCTAAAGCCTTG
GGCTAAAGCCTTGGT
GGCTAAAGCCTTGGTG
GGCTAAAGCCTTGGTGG
GGCTAAAGCCTTGGTGGT
GGCTAAAGCCTTGGTGGTA
GGCTAAAGCCTTGGTGGTAG
GGCTAAAGCCTTGGTGGTAGA
GGCTAAAGCCTTGGTGGTAGAT
GGCTAAAGCCTTGGTGGTAGATA
GGCTAAAGCCTTGGTGGTAGATAG
GGCTAAAGCCTTGGTGGTAGATAGT
GGCTAAAGCCTTGGTGGTAGATAGTC
GGCTAAAGCCTTGGTGGTAGATAGTCC
GGCTAAAGCCTTGGTGGTAGATAGTCCC
GGCTAAAGCCTTGGTGGTAGATAGTCCCT
5′ GGCTAAAGCCTTGGTGGTAGATAGTCCCTG 3′
Primer
DNA sequence

Figure 23.14 Diagram of the Sanger sequencing products. The DNA chain lengths are determined by the addition of ddNTP at different positions. The fluorescent dye-labeled cycle-sequencing products are separated using capillary electrophoresis. The fluorescent dyes are resolved by a detector, and the peaks corresponding to each DNA fragment are identified. (© Richard C. Li.)

the base number for the position is listed followed by an N. For example, 228N means an unresolved sequence ambiguity was observed at position 228.

23.3.3.3.2 Insertions

The insertion site is described by noting the position (at 5′ to the insertion) followed by a decimal point and a number. The number indicates the order of the insertions (e.g., 1 indicates the first insertion, 2 indicates the second, etc.). The base calling following the number indicates the inserted nucleotide (e.g., 524.1A, 524.2C).

23.3.3.3.3 Deletions

The deletion site designation is followed by the letter *d*. For example, a deletion at position 16,296 is recorded as 16,296d.

23.3.3.3.4 Heteroplasmic Sites

The IUPAC codes for base calling can be applied to heteroplasmic sites. For example, an A/G heteroplasmy can be designated as R, and a C/T heteroplasmy can be designated as Y.

23.3.4 Interpretation of mtDNA Profiling Results

Interpretation guidelines are used when comparing sequencing results between evidence and reference samples. General guidelines were set forth by the Scientific Working Group on DNA

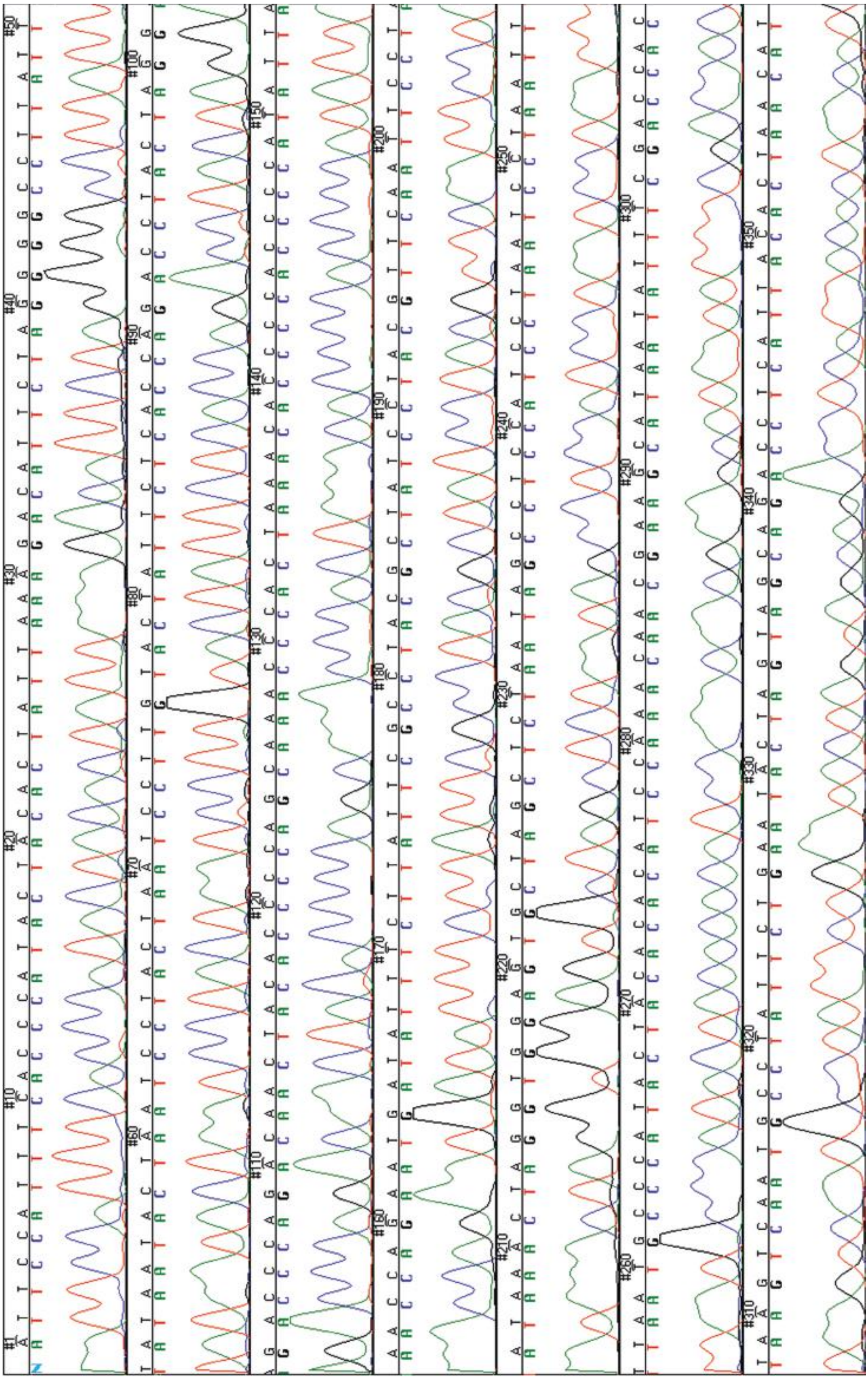

Figure 23.15 Electropherogram representing a DNA sequence presented as peaks with corresponding colors for each of the four bases. The bases associated with each peak are noted. The sequence is read from left to right (5′ to 3′). (© Richard C. Li.)

Analysis Methods (SWGDAM) and the DNA Commission of the International Society of Forensic Genetics (ISFG). In reporting mtDNA profiling results, the most common categories of conclusions are the following: cannot exclude, exclusion, and inconclusive result.

23.3.4.1 Exclusion

If the sequences of questioned and known samples are different, then the samples can be excluded as originating from the same source. It should be taken into account that higher mutation rates are found with the mtDNA genome than are found with the nuclear genome. The SWGDAM's guidelines define that the conclusion of exclusion can be made if there are two or more nucleotide differences between the questioned and known samples. Additionally, mutations seem to be more common in certain tissues. For this reason, the sources of the tissues analyzed should be taken into consideration.

23.3.4.2 Cannot Exclude

If the sequences are the same, the reference sample and evidence cannot be excluded as arising from the same source. When an mtDNA profile cannot be excluded, it is desirable to evaluate the weight of the evidence. In cases where the same heteroplasmy is observed in both questioned and known samples, its presence increases the strength of the evidence. However, if heteroplasmy is observed in a questioned sample but not in a known sample or vice versa, a maternal lineage still cannot be excluded.

23.3.4.3 Inconclusive Result

If the questioned and known samples differ by a single nucleotide, and no heteroplasmy is present, the results are considered to be inconclusive.

Bibliography

Mitochondrial DNA Profiling

Adachi, N., K. Umetsu, and H. Shojo, Forensic strategy to ensure the quality of sequencing data of mitochondrial DNA in highly degraded samples. *Leg Med (Tokyo)*, 2014, **16**(1): 52–55.

Anderson, S., et al., Sequence and organization of the human mitochondrial genome. *Nature*, 1981, **290**(5806): 457–465.

Andreasson, H., et al., Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology. *Forensic Sci Int Genet*, 2007, **1**(1): 35–43.

Andrews, R.M., et al., Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 1999, **23**(2): 147.

Anjum, G.M., et al., Pyrosequencing-based strategy for a successful SNP detection in two hypervariable regions: HV-I/HV-II of the human mitochondrial displacement loop. *Electrophoresis*, 2010, **31**(2): 309–314.

Arnestad, M., et al., A mitochondrial DNA polymorphism associated with cardiac arrhythmia investigated in sudden infant death syndrome. *Acta Paediatr*, 2007, **96**(2): 206–210.

Asari, M., et al., Differences in tissue distribution of HV2 length heteroplasmy in mitochondrial DNA between mothers and children. *Forensic Sci Int*, 2008, **175**(2–3): 155–159.

Bandelt, H.J. and W. Parson, Consistent treatment of length variants in the human mtDNA control region: A reappraisal. *Int J Legal Med*, 2008, **122**(1): 11–21.

Bandelt, H.J., M. van Oven, and A. Salas, Haplogrouping mitochondrial DNA sequences in legal medicine/forensic genetics. *Int J Legal Med*, 2012, **126**(6): 901–916.

Bendall, K.E. and B.C. Sykes, Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet*, 1995, **57**(2): 248–256.

Bendall, K.E., V.A. Macaulay, and B.C. Sykes, Variable levels of a heteroplasmic point mutation in individual hair roots. *Am J Hum Genet*, 1997, **61**(6): 1303–1308.

Bendall, K.E., et al., Heteroplasmic point mutations in the human mtDNA control region. *Am J Hum Genet*, 1996, **59**(6): 1276–1287.

24

Forensic DNA Databases

Tools for Crime Investigations

24.1 Brief History of Forensic DNA Databases

Forensic DNA databases are networks for exchanging information among law enforcement agencies to assist in solving crimes. For example, forensic DNA databases allow forensic laboratories to search DNA profiles against the databases to identify criminals. In 1995, the United Kingdom established the world's first national DNA database, NDNAD, in England and Wales. Scotland and Northern Ireland have their own databases but also submit their profiles to NDNAD. NDNAD demonstrated initial success in solving crimes. Three years later, the United States introduced its national *Combined DNA Index System* or CODIS. By the end of 1998, other countries (such as Austria, Germany, the Netherlands, New Zealand, and Slovenia) had also introduced national DNA databases. Table 24.1 describes some of these national DNA databases. This chapter will focus on CODIS.

24.2 Infrastructure of CODIS

In the United States, a pilot project of DNA databasing was initiated by the Federal Bureau of Investigation (FBI) and 14 participating state and local laboratories. Subsequently in 1994, the Congressional DNA Identification Act authorized the FBI to establish a national DNA database including "DNA identification profiles of persons convicted of crimes, and analyses of DNA samples recovered from crime scenes and from unidentified human remains." By 1997, 13 STR loci were selected and in 1998 were implemented as the core loci for the national database, known today as CODIS. All 50 states, the District of Columbia, the federal laboratories, the US Army Criminal Investigation Laboratory, and Puerto Rico contribute to CODIS.

CODIS has three hierarchical levels: the Local DNA Index System (LDIS), the State DNA Index System (SDIS), and the National DNA Index System (NDIS) (Figure 24.1). The LDIS is maintained at crime laboratories operated by police departments, sheriff's offices, and local agencies. All forensic DNA profiles originating at the local level are stored in the LDIS and are transmitted to the SDIS and NDIS. Each state maintains a SDIS, which is typically operated by a designated state laboratory. An SDIS also stores the DNA profiles generated from state laboratories. The quality assurance standards for a qualified laboratory were set up by the DNA Advisory Board in 1998. The periodic revision of the standards is now carried out by the

Table 24.1 Characteristics of National DNA Databases

Country	Year Established	Suspect Entry Criteria	Convicted Offender Entry Criteria	Removal Criteria
United Kingdom	1995	Any recordable offense that leads to imprisonment	Entered as suspect	Never removed
New Zealand	1996	No suspects entered	Relevant offense (≥7 years in prison)	Never removed unless conviction quashed
Austria	1997	Any recordable offense that leads to imprisonment	Entered as suspect	Only after acquittal
The Netherlands	1997	No suspects entered except when suspect's DNA is tested for case	Offense leading to >4 years in prison	20–30 years after conviction
Germany	1998	Offense leading to >1 year in prison	After court decision	After acquittal or 5–10 years after conviction if prognosis good
Slovenia	1998	Any recordable offense that leads to imprisonment	Entered as suspect	Depends on severity of crime
United States	1998	No suspects entered; under revision	Depends on state law	Depends on state law
Finland	1999	Offense leading to >1 year in prison	Entered as suspect	Only after acquittal
Sweden	2000	No suspects entered	Offense leading to >2 years in prison	10 years after release from prison
Switzerland	2000	Any recordable offense that leads to imprisonment	Entered as suspect	After acquittal or 5–30 years after conviction
France	2001	No suspects entered	Sexual assaults and serious crimes	40 years after conviction

Source: Adapted from Jobling, M.A. and Gill, P., *Nat Rev Genet*, 5, 739–751, 2004.

Scientific Working Group of DNA Analysis and Methods (SWGDAM). The SDIS enables local laboratories within that state to compare DNA profiles. SDIS also serves as the communication path through which the LDIS and NDIS are able to exchange messages. Communication is mediated using a secured network with encryption. Only authorized personnel, approved by the FBI, have access to CODIS network terminals and servers. NDIS is the highest level of the CODIS infrastructure, which contains DNA profiles contributed by participating local and state

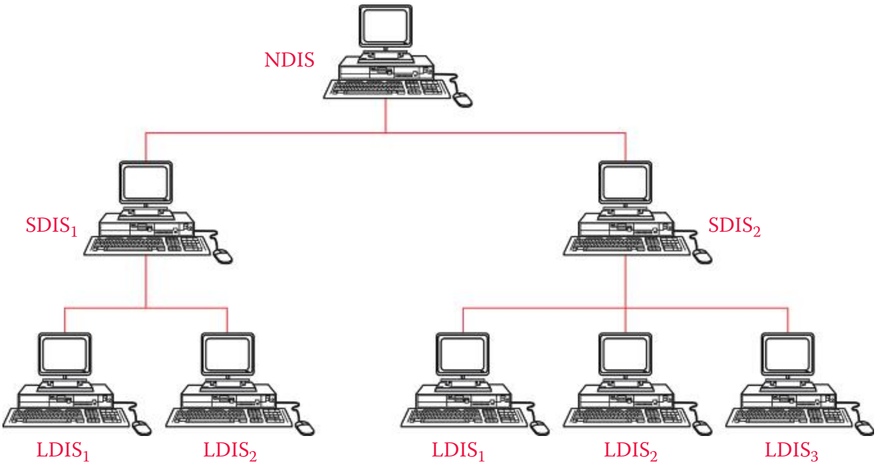

Figure 24.1 CODIS infrastructure. All DNA profiles originate at LDIS, and then enter SDIS and NDIS. SDIS allows laboratories within states to exchange DNA profiles. NDIS is the highest level of the infrastructure. It allows the participating laboratories to exchange DNA profiles on a national level.

laboratories. Additionally, NDIS stores DNA profiles generated by federal laboratories, such as those generated by the FBI and the US Army Criminal Investigation Laboratory. NDIS enables qualified state laboratories to compare DNA profiles. Searches of DNA profiles can be conducted at the national level. All DNA profiles submitted to NDIS are automatically searched weekly against the DNA profiles from other states. NDIS is administered by the FBI, which provides software, training, and support for all participating laboratories. Many law enforcement laboratories from other countries utilize the CODIS software for their own databases. However, laboratories in foreign countries do not have any access to the CODIS system. Nevertheless, a search through the CODIS database may be requested either from the FBI or International Criminal Police Organization (Interpol). Conversely, a search of an international DNA database can be arranged through Interpol.

24.3 Indexes of CODIS

The DNA profiles entered in CODIS are organized into categories known as indexes (Figure 24.2). The Convicted Offender Index contains DNA profiles of individuals convicted of crimes. The Arrestee Index contains DNA profiles of arrested individuals. It varies in SDIS databases based on each state's law permitting the collection of DNA samples from arrestees. The Forensic Index contains DNA profiles, also known as forensic profiles, derived from crime scene evidence, potentially originating from perpetrators but not including suspects.

Additionally, the FBI also established the National Missing Person DNA Database (NMPDD) Program for the identification of missing and unidentified persons at the national level. The DNA profiles entered into the NMPDD are categorized into three indexes that can be searched against each other. The Missing Person Index and the Unidentified Human Remains Index contain DNA profiles from missing persons and unidentified human remains, respectively. The Biological Relatives of the Missing Person Index contains DNA profiles voluntarily contributed from relatives. This index may also store patrilineal or matrilineal DNA profiles from the relative such as a biological father, mother, or child of the missing person to assist investigations. Additionally, a pedigree chart (a diagram showing the relationship between the missing person and relatives) can be created.

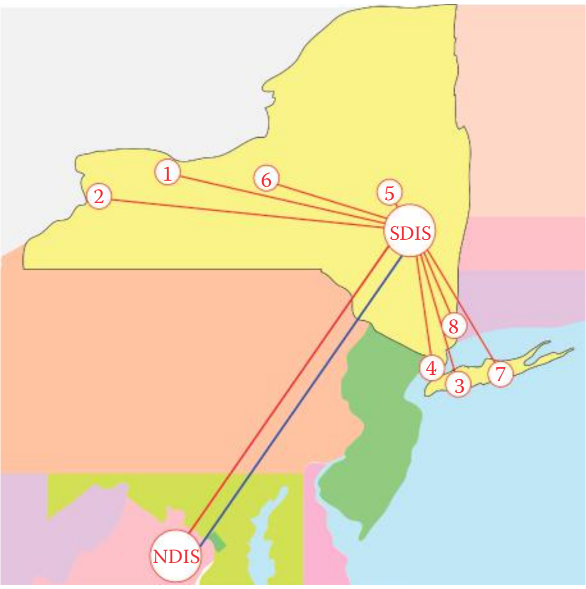

Figure 24.2 Example of an SDIS maintaining the Forensic Index (red) and Convicted Offender Index (blue). In this example, the New York State Police Forensic Investigation Center serves as the SDIS laboratory for New York State. LDIS sites: 1. Erie County Department of Central Police Services Forensic Laboratory, 2. Monroe County Public Safety Laboratory, 3. Nassau County Office of the Medical Examiner Division of Forensic Services, 4. New York City Office of the Chief Medical Examiner Department of Forensic Biology, 5. New York State Police Forensic Investigation Center, 6. Onondaga County Center for Forensic Sciences, 7. Suffolk County Crime Laboratory, and 8. Westchester County Department of Laboratories & Research Division of Forensic Sciences. (© Richard C. Li.)

In 2010, the FBI established a Rapid DNA Program Office for the purpose of developing rapid DNA technology. Rapid DNA technology is a fully automated process of performing STR analysis within 1–2 h to generate a CODIS core STR profile from a reference sample such as a buccal swab (Figure 24.3). The Rapid DNA Index System (RDIS) is the proposed index of NDIS. It shall be an integrated system capable of applying rapid DNA technology and carrying out database searches from police custody or booking units by trained police officers. The entire process, including obtaining any match from a database search, shall take less than 2 h.

The forensic DNA analysis of a reference sample, taken from an individual, takes 2–3 days if processed immediately in a forensic laboratory. The entire process includes the extraction of DNA from a sample, DNA quantification, PCR amplification, electrophoresis, and data collection and analysis. After being transported to a laboratory for processing, samples are usually stored and batched in laboratories prior to analysis. A typical turnaround time for analysis of such a sample is 1–3 months. While the sample is processed, the suspect is often released from custody. Perpetrators released on police bail may commit another crime. Thus, it is desirable to develop new technology that is capable of completing the forensic DNA analysis and database search while the suspect remains in custody. This technology would also facilitate rapid exclusion of a suspect, thus redirecting the investigation.

Rapid DNA instruments are portable, compact instruments designed to be deployed into field testing. These instruments are fully automated for processing reference samples in order to generate a DNA profile in less than 2 h. Several versions of the instruments have recently been developed by manufacturers and some of them are currently commercially available. In addition to rapid DNA instruments, rapid DNA profiling may be achieved through alternative processes. First, it can be achieved through rapid services that offer a quick turnaround time (<2 h). Second,

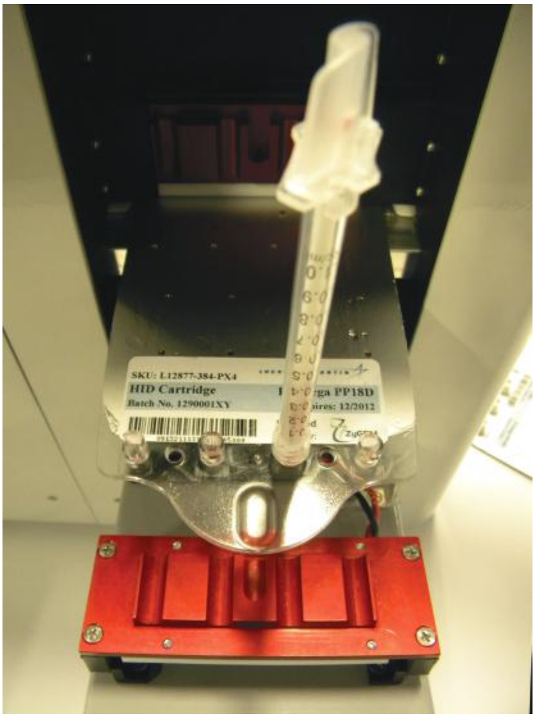

Figure 24.3 A rapid DNA instrument. A compact instrument that is designed to process swab samples and to produce DNA profiles within 2 h in the field. (© Richard C. Li.)

rapid DNA techniques can be applied using standard laboratory equipment but implementing specialized protocols to generate CODIS-compatible profiles in less than 2 h.

The primary goal of the rapid DNA initiative is to produce a CODIS-compatible DNA profile of a sample taken from an arrestee in police custody and to search DNA databases during the booking process in less than 2 h. This technology could be used, for example, to quickly identify other unsolved crimes or eliminate suspects, which can aid in investigative decision making. It can also potentially be useful for other investigations, such as identifying human remains in mass disasters; identifying detainees in counterterrorism applications; and, for immigration and border agents, confirming and verifying individual identifications or family relationships.

24.4 Database Entries

Currently, over 190 public law enforcement laboratories participate in CODIS across the United States. As of April 2013, CODIS contains the DNA profiles of more than 10 million convicted felons, over 1 million arrestees, and half a million crime scene samples. The proportion of the population represented in the database is approximately 3%. The NDIS is one of the largest DNA databases containing DNA profiles in terms of absolute numbers. Each CODIS entry consists of the DNA profile of the sample and the specimen identification number, as well as the information of the laboratory submitting the DNA profile and the laboratory analyst that performed the DNA analysis. The entry does not include case information or the personal information of the offenders or arrestees. Access to CODIS profiles is restricted to criminal justice agencies for law enforcement identification purposes.

The SDIS database retains samples collected from convicted offenders after DNA profiles are obtained. State policies vary in the retention of samples and DNA profiles in situations of dismissed charges, acquittal, or no charges for arrestees. Some states, such as Virginia, destroy the DNA sample and expunge the DNA profile from the database. Other states, such as California, require a petition to destroy the sample and expunge the DNA profile from the database. There are two major reasons for sample retention. First, the sample is needed for confirmatory processes for the purpose of quality assurance and control. In confirmatory processes, the DNA analysis of the original sample is repeated and compared against the prior analysis. Second, the retention of samples allows for possible retesting if new technology becomes available. It also allows retesting of a sample for purposes of updating with expanded loci. Considerable debate has surrounded the retention of samples. It can be argued from the opposite perspective that the database could reveal private genetic information that could then be misused. The objection is based on concern for the protection of privacy rights. If a sample were made available to an unauthorized person, confidential information could be disclosed.

24.5 Database Expansion

Currently, all 50 states have authorized the collection of samples from convicted felons for DNA databasing. Over the years, the demand for the utilization of databases has increased sharply (Figure 24.4). More jurisdictions are incorporating more felonies into the lists of crimes that require DNA profiles, and some jurisdictions plan to include all felonies in such databases. Additionally, the database system is projected to include the profiles of minor criminals, because statistics show that most offenders found guilty of serious crimes were previously convicted for minor crimes. Broadening the size of the database and including samples from more types of crimes could lead to the assumption that the number of crimes solved would also increase. Although state laws vary, each state has its own statute governing the entry criteria of the database samples. More and more states are authorizing the collection of additional types of DNA samples, including individuals convicted of misdemeanor crimes and adult felony arrestees who have not yet been convicted for the offense. It is known that offenders tend to commit multiple

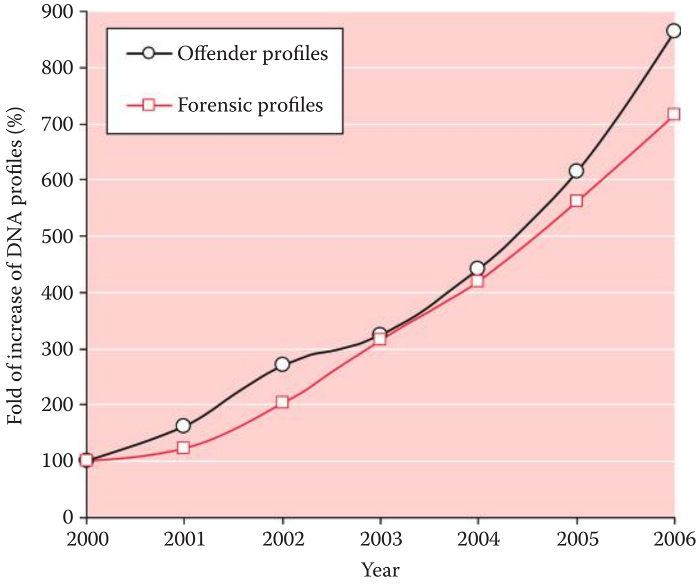

Figure 24.4 Growth of the CODIS database. In 2000, 460,365 offender profiles and 22,484 forensic profiles were entered.

crimes. Thus, including DNA profiles of arrestees in the database is beneficial in catching serial offenders and potentially preventing future crimes. Virginia is the first state to collect DNA from violent-felony arrestees, among other crimes. Over the years, more than half of states in the United States have amended their law to include all felony arrestees or a subset of arrestees for felonies that involve violence or sexual assault.

In 2009, Alonzo King was arrested in Wicomico County, Maryland, on violent assault charges. Maryland's DNA Collection Act authorizes law enforcement agencies to collect DNA samples from an individual who is arrested for violent crimes. During the booking process, under state law, King's buccal swabs were collected for forensic DNA identification. When King's DNA profile was entered into Maryland's SDIS, it matched to a crime scene DNA profile from an unsolved rape case in 2003. Based on the DNA evidence, King was indicted for first-degree rape by a grand jury. The defendant filed a motion to suppress the DNA evidence. King argued that using his DNA to investigate his connection to the 2003 rape was an unreasonable search and seizure under the Fourth Amendment, since the police had no reason to suspect his involvement in the 2003 rape. During the trial at Wicomico County Circuit Court, King's motion was denied. He was convicted of rape and sentenced to life in prison. The Court of Appeals of Maryland, the state's highest court, upheld that the DNA Collection Act was constitutional. The police had probable cause to arrest King on the assault charge, and thus collecting King's DNA sample was a reasonable search. However, the court overturned the low court and reversed King's conviction. It stated that the DNA Collection Act was inappropriately applied to King in this case. Thus, investigating King without probable cause for the unrelated rape was unconstitutional. To determine whether the Fourth Amendment permits states to collect and analyze the DNA of arrestees, a divided Supreme Court ruled 5–4 and reversed the Court of Appeals. The Supreme Court's judgment ruled that Maryland has the right to collect DNA evidence from individuals arrested for serious crimes. The ruling reflects the Supreme Court's view on the balance between the interest of criminal justice systems in solving violent crimes and an individual's interest in the Fourth Amendment, which protects them from unreasonable searches.

Some jurisdictions, such as the United Kingdom, allow DNA samples to be taken from individuals suspected of committing recordable offenses that may lead to prison sentences. In contrast, DNA profiles from suspects are not eligible for entry into CODIS. In some jurisdictions, it has been suggested that databases include many more offenders and suspects, as well as the general public. One advantage of including an entire population in a database is the ability to identify missing, kidnapped, and abducted individuals in addition to victims of major accidents and mass fatalities. Nevertheless, debates concerning the need to balance the benefits and dangers of developing a broader database will inevitably continue.

24.6 DNA Profiles

Currently, the CODIS software supports the storage and search of DNA profiles of short tandem repeat (STR), Y chromosome STR (Y-STR), and mitochondrial DNA (mtDNA). Y-STR and mtDNA profiles may only be searched within NMPDD-related indexes. The CODIS software no longer supports searches of DNA profiles generated by restriction fragment length polymorphism (RFLP) analysis. The 13 core CODIS STR loci are CSF1PO, FGA, THO1, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11. DNA profiles are entered into one of the indexes, such as convicted offender, arrestee, forensic, unidentified human remains, missing person, or a relative of a missing person. There is a minimum loci requirement for each DNA profile entering CODIS: 13 core CODIS loci are required for a DNA profile entering the Convicted Offender Index and the Arrestee Index; 13 core CODIS loci and amelogenin, a sex-typing marker, are required for a DNA profile entering the Relatives of Missing Person Index; at least 10 CODIS loci are required (all 13 core loci must be tested) for the Forensic Index; and at least 8 loci and amelogenin are required (all 13 core loci must be tested) for the Missing Person and Unidentified Human Remains Index. The DNA profiles must be generated by an accredited and audited laboratory in accordance with the FBI's Quality Assurance Standards using approved commercially available kits. As the database is growing rapidly, the chances of finding incidental matches among DNA profiles is increasing. To increase discrimination power, it has recently been proposed that CODIS core loci be expanded to include 20 or more markers (Chapter 20). Additionally, the expanded core loci share additional loci compatible with international standards for forensic DNA analysis, which facilitates the exchange of information with international law enforcement agencies.

24.7 Routine Database Searches for Forensic Investigations

The ultimate goal of the database utilization is to provide investigative leads for law enforcement in solving crimes, particularly in cases where no suspect has yet been identified. Currently, DNA profiles uploaded to NDIS are automatically searched once a week (Figure 24.5). A hit is a match made from the information provided by comparing a target DNA profile against the DNA profiles contained in the database. There are two types of CODIS hits: an *offender hit* provides the identity of a potential suspect of a crime, while a *forensic hit* reveals the linkage between two or more crime scenes. Once a hit is identified, the match is verified by the laboratories that originally processed the evidence. A verified CODIS hit can be utilized as probable cause to allow law enforcement to obtain a court order to collect a DNA sample from a suspect. Collecting DNA with a warrant ensures admissibility in court. *Investigation-aided* cases are those assisted by CODIS hits, including case-to-case matches as well as case-to-offender matches. The number of investigations aided is a useful measure of the successful application of the database (Figure 24.6). As of April 2013, over 207,800 hits had been made with CODIS assistance in more than 199,200 investigations.

24.7.1 Case-to-Offender Searches

Matches of profiles from the Forensic Index and the Offender Index reveal the identities of perpetrators of crimes. For example, in 1998, a Florida man, Leon Dundas, became a suspect in a rape case but refused to provide his DNA reference sample for testing. Dundas was killed a year later in an illegal drug deal. Thus, a postmortem sample of Dundas was obtained and his DNA profile was compared with the Forensic Index of CODIS. It was discovered that Dundas's profile

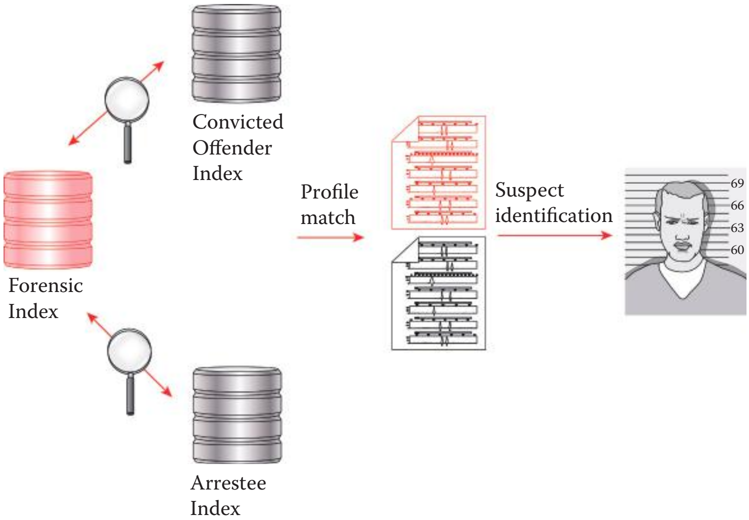

Figure 24.5 Example of a weekly routine CODIS search. (© Richard C. Li.)

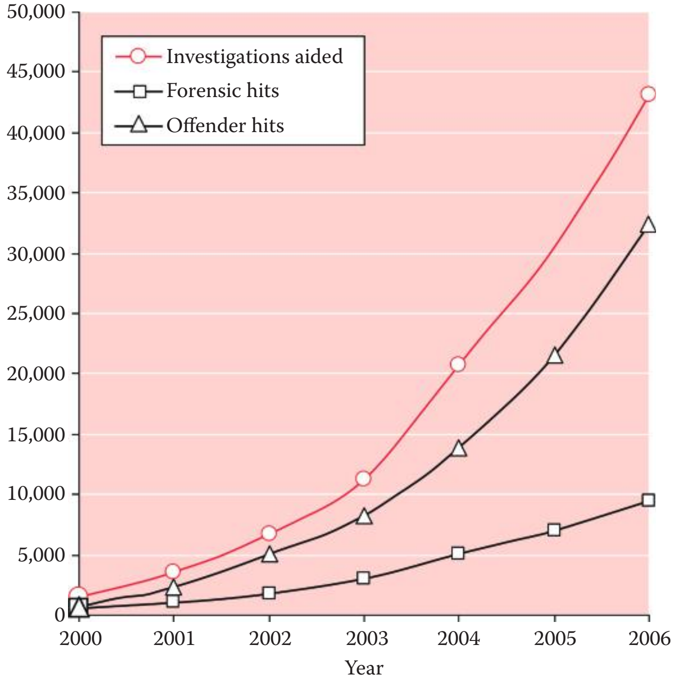

Figure 24.6 Number of hits and investigations aided by the CODIS database.

matched the DNA sample found in the rape case. Additionally, his DNA profile was linked to several other unsolved rapes in Jacksonville and Washington, DC. Such cases can now be solved by utilizing DNA databases.

24.7.2 Case-to-Case Searches

A target DNA profile from a crime scene is also searched against the profiles stored in the Forensic Index of the database. Matches of profiles among the target profile and the profile in the Forensic Index can link separate crime scenes and aid in identifying serial offenders. This helps law enforcement agencies in multiple jurisdictions to coordinate their investigations and share leads. For example, in 1996, two young girls were abducted from bus stops in St. Louis. Both girls were raped and DNA samples were collected. Both DNA profiles pointed to the same perpetrator. In 1999, the St. Louis police decided to reanalyze the samples using new STR technology through the CODIS database. The database found a match to a different rape case, to which the perpetrator, Dominic Moore, had already confessed, thus identifying him as the perpetrator of the 1996 rapes.

24.7.3 Search Stringency and Partial Matches

Database searches are carried out using the CODIS software with three stringency levels, which allow the search of complex forensic profiles against offender profiles (Figure 24.7). A high-stringency match requires an exact match in which all alleles are matched at each locus between the target and candidate profiles. A moderate-stringency match, as defined by the FBI, is a candidate match "between two single source profiles having at each locus all of the alleles of one sample represented in the other sample." In a moderate-stringency match, allelic drop-outs in a target or candidate profile are allowed, possibly resulting in a partial match at some loci. As a result, high-stringency matches are automatically included in a moderate-stringency

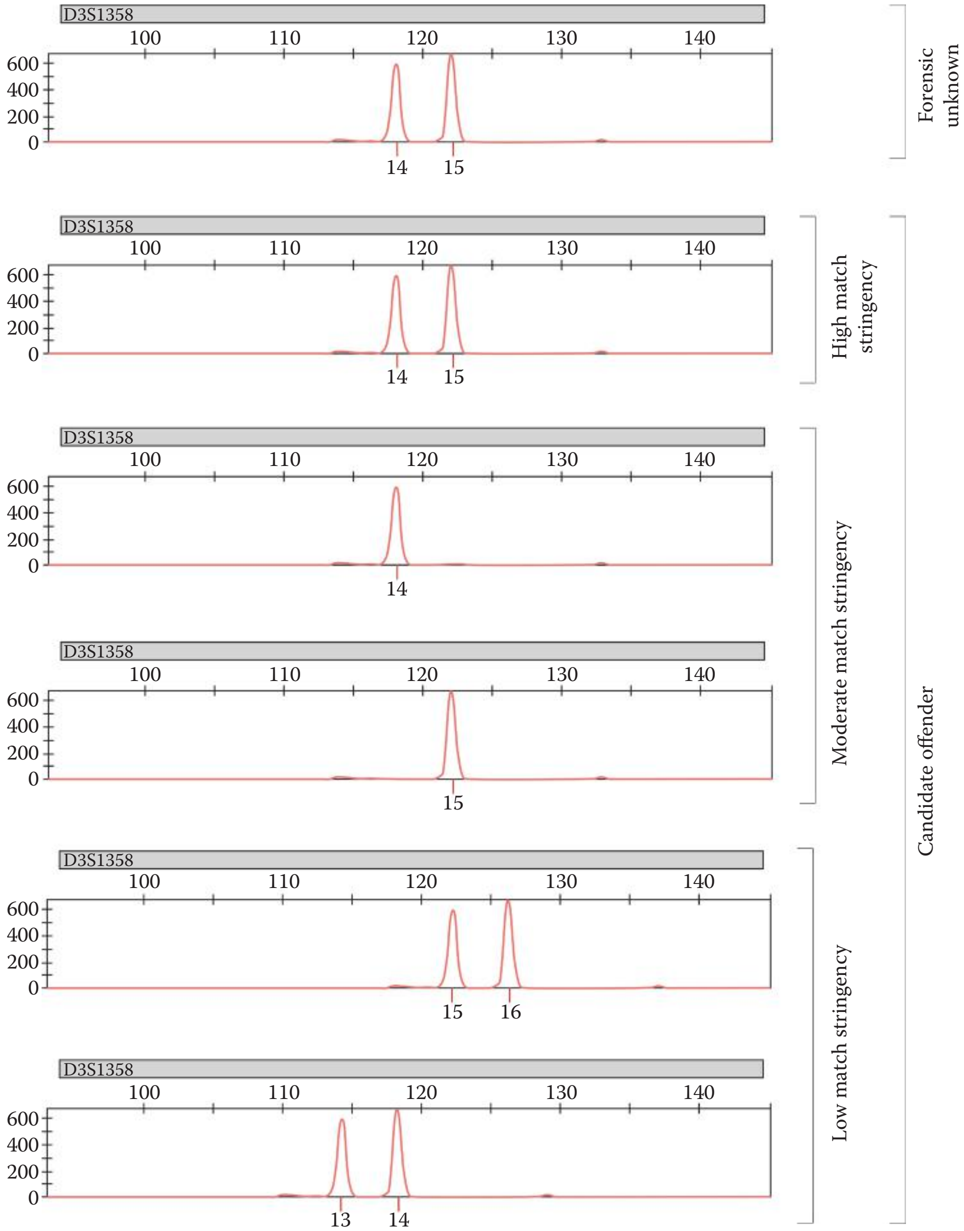

Figure 24.7 Diagram of match stringencies. A heterozygote (allele 14 and 15) at locus D3S1358 is utilized as the genotype for the forensic unknown profile. (© Richard C. Li.)

search. In a low-stringency search, both mismatches and allelic dropouts are allowed. Presently, NDIS searches are only carried out at moderate stringency. During forensic DNA analysis, DNA degradation may prevent full DNA profiles from being processed, producing only partial profiles. Additionally, mixture profiles derived from forensic samples containing DNA contributed by more than one individual may be encountered. Furthermore, due to mutations, null alleles may occur in the profiles produced with some primer sets but not other primers. Therefore, searching at moderate stringency allows the detection of matches under the situations described above.

Such moderate-stringency searches between the Forensic Index and Offender Index may occasionally generate partial match profiles. Since a partial match is not an exact match of the two profiles, further investigation is needed. For instance, additional Y-STR and mtDNA analysis is needed to eliminate unrelated individuals.

24.8 Familial Searches

Familial search, initiated in the United Kingdom, is a new method of applying databases in criminal investigations. It is known that, in the United States, nearly half of prison inmates have close relatives who have also been incarcerated. Familial search is based on the assumption that close relatives share more alleles of DNA profiles than unrelated individuals. Thus, databases may be utilized to identify perpetrators by finding a close relative, if the close relative has been convicted of a crime and is listed in the database. Familial search is an intentional search of a target crime scene profile against an offender database to obtain a list of candidate profiles that are similar to the target profile. This list may include the profile of a close relative of the perpetrator, who is the source of crime scene evidence. These matches most frequently involve siblings, parents, or children. The investigative leads produced by familial searches allow law enforcement to conduct further investigations to identify the perpetrator.

The first familial search leading to a successful prosecution was conducted in Surrey, England. In 2003, a truck driver was killed after a brick was thrown through his windshield from a bridge. The perpetrator's DNA profile was obtained from the brick. A search of the UK's national DNA database revealed no match for the perpetrator. Next, a familial search of the database was conducted. The system identified a close relative that led police to identify the perpetrator, Craig Harman, who was then convicted of manslaughter.

24.8.1 Legal and Ethical Issues of Familial Search

The use of forensic databases involves a balance of individual civil rights and the interests of the criminal justice systems. Many concerns have been raised, including the potential for these searches to violate the privacy of unrelated people whose genetic profiles happen to resemble those of individuals included in the databases. In the United States, the Fourth Amendment protects against unreasonable searches and seizures. The permissibility of familial searching under the Fourth Amendment is yet to be addressed by courts. The collection of biological material for the initial creation of a profile for law enforcement purposes is subject to the Fourth Amendment implications. In terms of familial searches, some legal experts argue that the biological materials are not collected directly from the individuals for a familial search, and that these individuals may thus be protected under the Fourth Amendment. Others argue, however, that the Fourth Amendment protects the initial creation of the profile, including the sample collection, forensic DNA analysis, and the databasing. They argue that the Fourth Amendment may not protect subsequent investigations of DNA profiles during familial searches. In addition, the CODIS database consists of a high percentage of profiles from individuals of racial minority groups, including African Americans. Familial searches can disproportionately focus on a specific racial group.

Familial searches are rare in the United States. Two jurisdictions, Maryland and the District of Columbia, have laws prohibiting the use of familial searches. Familial searches are not conducted at the NDIS. While familial searches are now being performed in several jurisdictions in the United States, policies on familial searches vary among jurisdictions. The major issues relate to criteria for privacy, information release, search approval, and the types of crimes eligible for familial searches. Currently, California, Colorado, Texas, and Virginia have state legislation permitting familial searches. Familial searches are initiated when a specific suspect is not known, and the cause for the search needs to be justified. Familial searches are usually conducted for crimes that pose a substantial threat to public safety, typically those cases involving the most serious offenses. Additionally, familial searches are only conducted after a routine search of a DNA profile has yielded no match in the database.

In 2003, for example, a crime scene DNA sample of a closed rape and murder case, committed decades previously, was reanalyzed using forensic DNA techniques. The crime scene DNA profile was compared to the profile of a man, Darryl Hunt, who was then imprisoned for the crime. Surprisingly, the crime scene DNA profile did not match either that of Hunt or those of convicted felons in the database. However, the database search revealed a close relative of the true perpetrator and thus led law enforcement to identify Willard Brown as the perpetrator. Brown was sentenced to life imprisonment. Additionally, Hunt was exonerated after 18 years in prison.

24.8.2 Familial Search Strategies

The familial search is usually carried out using specially designed software. Although the CODIS software is not designed for familial search, it can be used for familial search through a low-stringency search, which may result in a list of candidate profiles including close relatives of individuals, such as parent–offspring or full-sibling relatives. In a large offender DNA database, similar DNA profiles from unrelated people are often observed due to shared alleles. Therefore, the candidate profile list may also include unrelated individuals whose DNA profiles are similar to the target profile. As a result, a familial search can provide a list of potential candidates consisting of hundreds of profiles, which would be too labor intensive to pursue through further investigations. Several methods can be used to determine a cutoff analytical threshold in order to limit a pool of candidate profiles and exclude unrelated individuals from familial searches (Figure 24.8).

24.8.2.1 Identity-by-State and Kinship Index Method

The *identity-by-state* (IBS) method compares the number of shared alleles and loci between a target forensic profile and the offender profiles in a database but does not take into account allele frequencies. The analytical threshold for a familial search is determined by a preset number of shared alleles or loci in order to prompt further investigation. For example, some states use 15 shared alleles as the analytical threshold to be considered as a candidate, while other states require at least one shared allele for each locus. Additionally, this method can prioritize a pool of candidates based on the highest to the lowest number of shared alleles for investigation.

The *Kinship Index* (KI) method is a likelihood ratio–based method that evaluates the familial match by comparing the probability that two DNA profiles are from related individuals to the probability that they are unrelated. The KI method analyzes the allele frequency data, including all CODIS core loci, to calculate the Combined Kinship Index. The KI may vary based on the allele frequency data across the population. Thus, the accuracy of the KI method relies on the relevance of the population data set analyzed. False inclusions or exclusions may occur if nonrelevant population data are utilized. The KI method also allows the generation of a ranked candidate list according to the probability that the individuals are related. Generally speaking, the accuracy of the KI method is higher than that of the IBS method. However, using both IBS and KI methods is better than using a single method alone for familial search.

24.8.2.2 Focusing on Rare Alleles

During a familial search, target DNA profiles may contain rare alleles with low allelic frequencies. The chance of sharing rare alleles for two close relatives is higher than for two unrelated individuals. If a target profile carries a rare allele, it can be used to narrow the pool of candidates.

For example, Jeffrey Gafoor, then 23, had been living in a suburban neighborhood in Cardiff, Wales. He had a reputation for being a loner. Besides working in his family-owned shop, he spent most of his time at home. On February 13, 1988, Gafoor entered 7 James Street in the Butetown area of Cardiff. The first-floor unit was temporarily occupied by Lynette White, a 20-year-old

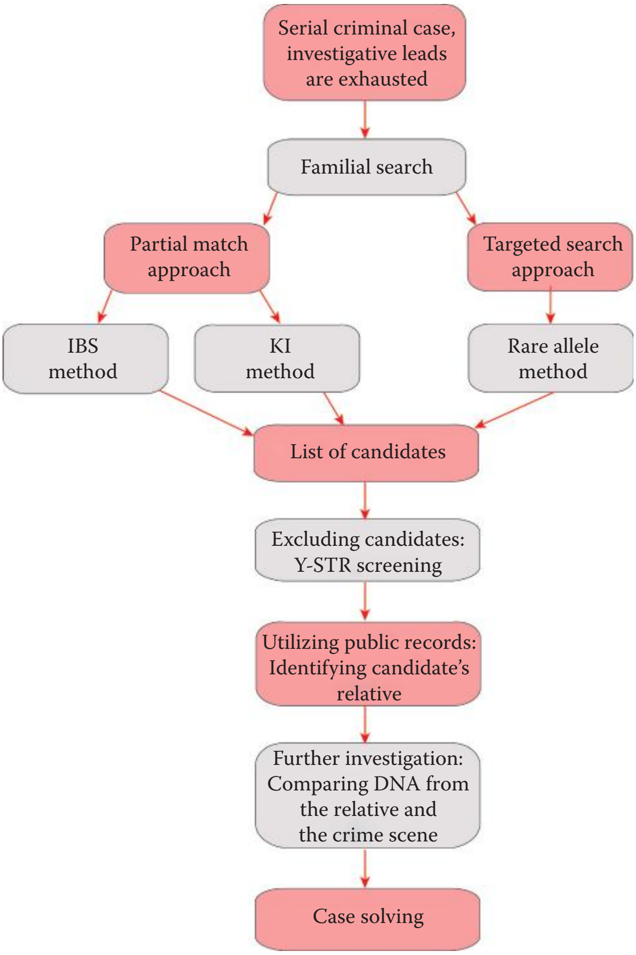
Serial criminal case, investigative leads are exhausted
Familial search
Partial match approach
Targeted search approach
IBS method
KI method
Rare allele method
List of candidates
Excluding candidates: Y-STR screening
Utilizing public records: Identifying candidate's relative
Further investigation: Comparing DNA from the relative and the crime scene
Case solving

Figure 24.8 Example of a familial search workflow. (© Richard C. Li.)

woman who was working as a prostitute in the area. Gafoor went there for the purpose of receiving sexual services. He left the scene in the early hours of Valentine's Day. On Valentine's night, White's body was discovered. She suffered from more than 50 stab, cut, and slash wounds as well as defensive wounds on her hands. The body was dragged to the corner of the room adjacent to the bed, the only piece of furniture in the room.

Gafoor left blood and semen evidence at the scene. He bled during the murder and deposited bloodstains at the scene. Additionally, he left seminal stains at the scene; the semen had no measurable level of sperm and most likely came from an individual who had a medical condition known as azoospermia (Chapter 14). After the murder, Gafoor lived at the same place as he did beforehand and kept the same job for many years.

Two years later, three local men, known as the "Cardiff Three," were convicted of White's murder and sentenced to life in prison. In 1992, however, their convictions were reversed by the court of appeal. Subsequently, the crime scene evidence was tested using new techniques in forensic DNA analysis that were not available in 1988. A DNA profile was obtained from the crime scene. A familial search revealed a similar profile from a 14-year-old boy in the UK National DNA Database. In particular, there was a rare allele match between the crime scene and the candidate DNA profiles. This search led to the further investigation of Gafoor, who was the uncle of the boy. Gafoor was charged with murder.

24.8.2.3 Excluding Candidates through Y-STR Screening

The vast majority of DNA profiles in CODIS come from males. Patrilineal markers, such as Y-STR loci, are successfully utilized to screen the candidates to verify the relationship between two individuals. This analysis helps to identify first-order relatives as well as paternal half-siblings. STR loci beyond the CODIS core loci and maternal lineage makers, such as mitochondrial DNA typing systems, can be used to narrow the pool of candidates and eliminate coincidental partial matches. One or more of these methods can be incorporated to evaluate highly ranked candidates.

For example, Lonnie Franklin lived in a southern Los Angeles neighborhood. Throughout the 1980s, Franklin worked for the city as a maintenance worker and a sanitation truck operator. His neighbors recalled that he often brought prostitutes to a camper parked in the backyard of his house. He also took photos of nude women, which he kept in his garage. Nevertheless, to many of his neighbors, Franklin seemed to be a friendly person who often chatted with them.

Since 1985, the bodies of many women, most of which were prostitutes, had been dumped in an alley running along Western Avenue in a southern Los Angeles neighborhood. The victims had been shot and some had been strangled after sexual contact. DNA and ballistics analysis revealed that at least 10 of the murders had been committed by the same perpetrator. The perpetrator was known to the general public as the “Grim Sleeper” because he had taken a 14-year break during the period in which these 10 murders had taken place. The crime scene DNA profile did not match any profile in the database.

In 2008, the first familial search of the “Grim Sleeper” case was conducted by the California Department of Justice. California state law allows familial searches for high-profile cases if all other leads have been exhausted. However, all candidate DNA profiles were excluded through Y-STR screening. As a result, no candidate was identified as a possible relative of the “Grim Sleeper.” In 2010, a second search targeted partial matches that shared at least 15 alleles. It produced a list of 100 candidate profiles ranked by the likelihood that they were related. Y-STR was used as a screen tool to identify possible patrilineal relatives. This time, the search generated a match to a new DNA profile of a felon, which had recently been entered into California’s SDIS, in 2009. The search suggested that the candidate was a relative of the source of the DNA from the crime scenes. This search led to the identification of Franklin as a suspect in the “Grim Sleeper” murders. Police detectives conducted surveillance and collected a discarded slice of pizza and a cup used by Franklin at a local restaurant for DNA identification. The DNA profile was identical to the crime scene DNA profiles. Franklin was arrested and was charged with ten counts of murder and one of attempted murder.

Bibliography

Alonso, A., et al., Challenges of DNA profiling in mass disaster investigations. *Croat Med J*, 2005, **46**(4): 540–548.

Aoki, Y., et al., Comparison of the likelihood ratio and identity-by-state scoring methods for analyzing sib-pair test cases: A study using computer simulation. *Tohoku J Exp Med*, 2001, **194**(4): 241–250.

Asplen, C. and S.A. Lane, International perspectives on forensic DNA databases. *Forensic Sci Int*, 2004, **146**(Suppl): S119–S121.

Baker, L.E. and E.J. Baker, Reuniting families: An online database to aid in the identification of undocumented immigrant remains. *J Forensic Sci*, 2008, **53**(1): 50–53.

Balding, D.J., The DNA database search controversy. *Biometrics*, 2002, **58**(1): 241–244.

Balding, D.J., et al., Decision-making in familial database searching: KI alone or not alone? *Forensic Sci Int Genet*, 2013, **7**(1): 52–54.

Barash, M., et al., A search for obligatory paternal alleles in a DNA database to find an alleged rapist in a fatherless paternity case. *J Forensic Sci*, 2012, **57**(4): 1098–1101.

Barber, M.D. and B.H. Parkin, Sequence analysis and allelic designation of the two short tandem repeat loci D18S51 and D8S1179. *Int J Legal Med*, 1996, **109**(2): 62–65.

25

Evaluation of the Strength of Forensic DNA Profiling Results

25.1 A Review of Basic Principles of Genetics

25.1.1 Mendelian Genetics

Mendel's first law is the principle of segregation of alleles. Each pair of alleles segregates from others in the formation of gametes (mature reproductive cells such as spermatozoa and oocytes). One-half of the gamete carries one allele, and the other half carries the other allele.

Mendel's second law is the principle of independent assortment of alleles. The segregation of each pair of alleles is independent of the segregation of other pairs during the formation of gametes.

Gametes are formed during a process known as *meiosis*, in which cells with haploid chromosome numbers (23 in humans) are produced by the division of cells with diploid chromosome numbers (46 in humans). A fertilized human egg thus contains a diploid number of chromosomes. A diploid is composed of 22 pairs of autosomes and 2 sex chromosomes (XX in females and XY in males).

Based on Mendelian principles, genes on different chromosomes assort independently of one another in gamete production. Genes residing very closely together on the same chromosome are usually inherited together. Thus, they do not assort independently and are called *linked genes*. Genes distant from each other on the same chromosome are usually inherited separately. This results from an exchange of segments between a pair of *homologous* chromosomes when the chromosomes are paired during the early phases of meiotic division. These types of gene-exchange events on homologous chromosomes are collectively called *crossing over*, which results in the *recombination* of genes in a pair of chromosomes (Figure 25.1).

The Mendelian inheritances of genes can often be measured using *probabilities*. A probability is the ratio of the number of actual occurrences of an event to the number of possible occurrences. Additionally, the probability of two independent events occurring simultaneously is the product of each of their individual probabilities; this is known as the *product rule of probability*.

Mendelian principles apply to the inheritance of loci of the autosomal nuclear DNA genome commonly used for forensic testing. The Y-chromosomal genome is inherited paternally, which does not obey the rules for Mendelian principles. Mitochondria contain their own mitochondrial genomes and are inherited maternally. This maternal inheritance

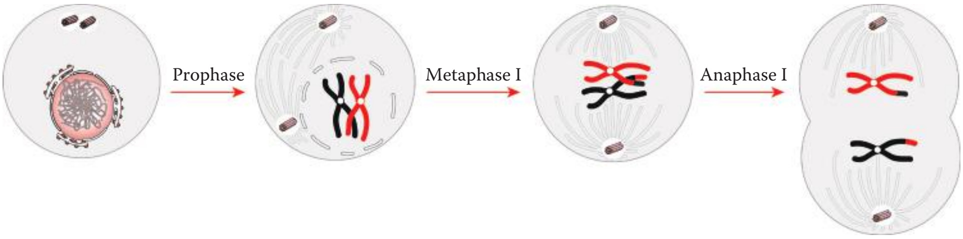

Figure 25.1 Crossing over hypothesis. Chromosomes are replicated prior to the first meiotic division. Each duplicate is called a chromatid, which forms into tetrads in the prophase. A crossover event occurs between the maternal and the paternal chromatids. Each chromatid breaks at the point of the cross and fuses with a portion of its counterpart. Chromosomes are associated with the microtubules during metaphase I and are pulled toward the spindle poles during anaphase I. The maternal and paternal homologs of one chromosome are shown. (© Richard C. Li.)

of mitochondrial genes also does not obey the rules for Mendelian principles. The inheritance of Y-chromosomal and mitochondrial genomes is referred to as non-Mendelian inheritance.

25.1.2 Population Genetics

Population genetics studies the causes of patterns of genetic variations within and among populations.

25.1.2.1 Allele Frequency

Allele frequency (p) can be calculated directly by counting the number of alleles of one type at a given locus and dividing it by the total number of alleles at that locus in a sampled population. This is called the *gene counting* method.

25.1.2.2 Genotype Frequency

Genotype frequency (P) observed at a given locus can be calculated by dividing the number of individuals with one genotype by the total number of individuals in a sampled population. Each genotype at the locus can be calculated separately. The summation of all genotype frequencies at that locus should equal 1.

25.1.2.3 Heterozygosity

Heterozygosity is the proportion of alleles, at a given locus, that are heterozygous and is calculated as

$$h = 1 - \sum_{i} p^2 \tag{25.1}$$

where:

h = heterozygosity
p = allelic frequency of the locus for homozygotes

The amount of heterozygosity at a locus in a sampled population is a measure of genetic variation. The higher the heterozygosity, the more variation there is at a given locus.

25.1.2.4 Hardy–Weinberg Principle

The Hardy–Weinberg (HW) principle, independently discovered by two scholars in the early 1900s, allows predictions of genotype frequencies to be made based on allelic frequencies. However, certain conditions must be met. The population must be large; mate randomly; and lack mutation, migration, and natural selection. If these conditions of the HW principle are met, the population will be in *equilibrium*, and the following results are expected:

1. The frequencies of the alleles will not change from one generation to the next.
2. Genotype frequencies can be predicted by the allelic frequencies (p^2 and q^2 for genotype frequencies of homozygotes and $2pq$ for genotype frequencies of heterozygotes). The sum of the genotype frequencies should equal 1.

$$p^2 + 2pq + q^2 = 1 \tag{25.2}$$

If the observed genotype proportions are different from those expected, one or more of the HW assumptions have been violated.

25.1.2.5 Testing for HW Proportions of Population Databases

To determine whether the genotypes of a population in question obey the HW principle, a population database can be constructed. Samples (usually 100–200 samples for STR loci) are collected and analyzed at the loci of interest. Allelic frequencies are obtained by using the gene counting method. Table 25.1 shows the allelic frequencies of CODIS loci from a population database. The *observed genotype frequencies* at a given locus, as described earlier in Section 25.1.2.2, are calculated by dividing the number of individuals with one genotype by the total number of individuals in the population sampled. The *expected genotype frequencies* are calculated using p^2, $2pq$, and q^2.

The observed and expected genotype frequencies are then compared using a chi-square test. The significance of the differences between observed and expected genotype frequencies can then be determined. The chi-square is calculated using the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i} \tag{25.3}$$

where:

O_i = ith observed genotype frequency
E_i = ith expected genotype frequency
n = total number of genotypes

Chi-square (χ^2) is calculated as the sum of all genotypes of a given locus.

The chi-square value and the degrees of freedom (the number of genotypes minus the numbers of alleles) are then used to obtain a p value (not to be confused with the allelic frequency designated p) from a table of p values, and such tables can be found at the backs of most statistics textbooks. If the p value exceeds 0.05 (5% significance level), the deviation of the expected genotype frequencies from the observed genotype frequencies is not considered statistically significant. Thus, the null hypothesis that the observed genotype frequencies fit the expected genotype frequencies predicted by the HW principle is not rejected if the p value is greater than 0.05.

Table 25.1 Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
D3S1358	(*N*=210)	(*N*=203)	(*N*=209)
<12	0.476	0.000	0.000
12	0.238	0.000	0.000
13	1.190	0.246	0.239
14	12.143	14.039	7.895
15	29.048	24.631	42.584
15.2	0.000	0.000	0.000
16	30.714	23.153	26.555
17	20.000	21.182	12.679
18	5.476	16.256	8.373
19	0.476	0.493	1.435
>19	0.238	0.000	0.239
VWA	(*N*=180)	(*N*=196)	(*N*=203)
11	0.278	0.000	0.246
13	0.556	0.510	0.000
14	6.667	10.204	6.158
15	23.611	11.224	7.635
16	26.944	20.153	35.961
17	18.333	26.276	22.167
18	13.611	22.194	19.458
19	7.222	8.418	7.143
20	2.778	1.020	1.232
21	0.000	0.000	0.000
FGA	(*N* = 180)	(*N* = 196)	(*N* = 203)
<18	0.278	0.000	0.000
18	0.833	3.061	0.246
18.2	0.833	0.000	0.000
19	5.278	5.612	7.882
19.2	0.278	0.000	0.000
20	7.222	14.541	7.143
20.2	0.000	0.255	0.246
21	12.500	17.347	13.054

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
21.2	0.000	0.000	0.246
22	22.500	18.878	17.734
22.2	0.556	1.020	0.493
22.3	0.000	0.000	0.000
23	12.500	15.816	14.039
23.2	0.000	0.000	0.739
24	18.611	13.776	12.562
24.2	0.000	0.000	0.000
24.3	0.000	0.000	0.000
25	10.000	6.888	13.793
26	3.611	1.786	8.374
27	2.222	1.020	3.202
28	1.667	0.000	0.246
29	0.556	0.000	0.000
30	0.278	0.000	0.000
>30	0.278	0.000	0.000
D8S1179	($N = 180$)	($N = 196$)	($N = 203$)
<9	0.278	1.786	0.246
9	0.556	1.020	0.246
10	2.500	10.204	9.360
11	3.611	5.867	6.158
12	10.833	14.541	12.069
13	22.222	33.929	32.512
14	33.333	20.153	24.631
15	21.389	10.969	11.576
16	4.444	1.276	2.463
17	0.833	0.255	0.739
18	0.000	0.000	0.000
D21S11	($N = 179$)	($N = 196$)	($N = 203$)
24.2	0.279	0.510	0.246
24.3	0.000	0.000	0.000

(*continued*)

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
26	0.279	0.000	0.000
27	6.145	4.592	0.985
28	21.508	16.582	6.897
29	18.994	18.112	20.443
29.2	0.279	0.000	0.246
30	17.877	23.214	33.005
30.2	0.838	3.827	3.202
30.3	0.000	0.000	0.000
31	9.218	7.143	6.897
31.2	7.542	9.949	8.621
32	0.838	1.531	1.232
32.1	0.000	0.000	0.000
32.2	6.983	11.224	13.547
33	0.838	0.000	0.000
33.2	3.352	3.061	4.187
34	0.838	0.000	0.000
34.2	0.279	0.000	0.493
35	2.793	0.000	0.000
35.2	0.000	0.255	0.000
36	0.559	0.000	0.000
>36	0.559	0.000	0.000
D18S51	($N = 180$)	($N = 196$)	($N = 203$)
<11	0.556	1.276	0.493
11	0.556	1.276	1.232
12	5.833	12.755	10.591
13	5.556	12.245	16.995
13.2	0.556	0.000	0.000
14	6.389	17.347	16.995
14.2	0.000	0.000	0.000
15	16.667	12.755	13.793
15.2	0.000	0.000	0.000
16	18.889	10.714	11.576

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
17	16.389	15.561	13.793
18	13.056	9.184	5.172
19	7.778	3.571	3.695
20	5.556	2.551	1.724
21	1.111	0.510	1.970
21.2	0.000	0.000	0.000
22	0.556	0.255	0.739
>22	0.556	0.000	1.232
D5S818	(*N* = 180)	(*N* = 195)	(*N* = 203)
7	0.278	0.000	6.158
8	5.000	0.000	0.246
9	1.389	3.077	5.419
10	6.389	4.872	6.650
11	26.111	41.026	42.118
12	35.556	35.385	29.064
13	24.444	14.615	9.606
14	0.556	0.769	0.493
15	0.000	0.256	0.246
>15	0.278	0.000	0.000
D13S317	(*N* = 179)	(*N* = 196)	(*N* = 203)
7	0.000	0.000	0.000
8	3.631	9.949	6.650
9	2.793	7.653	21.921
10	5.028	5.102	10.099
11	23.743	31.888	20.197
12	48.324	30.867	21.675
13	12.570	10.969	13.793
14	3.631	3.571	5.665
15	0.279	0.000	0.000
D7S820	(*N* = 210)	(*N* = 203)	(*N* = 209)
6	0.000	0.246	0.239

(*continued*)

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
7	0.714	1.724	2.153
8	17.381	16.256	9.809
9	15.714	14.778	4.785
10	32.381	29.064	30.622
10.1	0.000	0.000	0.000
11	22.381	20.197	28.947
11.3	0.000	0.000	0.000
12	9.048	14.039	19.139
13	1.905	2.956	3.828
14	0.476	0.739	0.478
CSF1PO	($N = 210$)	($N = 203$)	($N = 209$)
6	0.000	0.000	0.000
7	4.286	0.246	0.239
8	8.571	0.493	0.000
9	3.333	1.970	0.718
10	27.143	25.369	25.359
10.3	0.000	0.246	0.000
11	20.476	30.049	26.555
12	30.000	32.512	39.234
13	5.476	7.143	6.459
14	0.714	1.478	0.957
15	0.000	0.493	0.478
TPOX	($N = 209$)	($N = 203$)	($N = 209$)
6	8.612	0.000	0.478
7	2.153	0.246	0.478
8	36.842	54.433	55.502
9	18.182	12.315	3.349
10	9.330	3.695	3.349
11	22.488	25.369	27.273
12	2.392	3.941	9.330
13	0.000	0.000	0.239
TH01	($N = 210$)	($N = 203$)	($N = 209$)

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
5	0.000	0.000	0.239
6	10.952	22.660	23.206
7	44.048	17.241	33.732
8	18.571	12.562	8.134
8.3	0.000	0.246	0.000
9	14.524	16.502	10.287
9.3	10.476	30.542	24.163
10	1.429	0.246	0.239
D16S539	($N = 209$)	($N = 202$)	($N = 208$)
8	3.589	1.980	1.683
9	19.856	10.396	7.933
10	11.005	6.683	17.308
11	29.426	27.228	31.490
12	18.660	33.911	28.606
13	16.507	16.337	10.337
14	0.957	3.218	2.404
15	0.000	0.248	0.240

Source: Budowle, B., et al., *J Forensic Sci*, 44, 1277–1286, 1999. With permission.

25.1.2.6 Probability of Match

The discriminating power of genetic loci used above can be measured by *population match probability* (P_m). P_m is defined as the probability of having a matching genotype between two randomly chosen individuals. The lower the P_m, the less likely a match between two randomly chosen individuals will occur. This is calculated as follows:

$$P_m = \sum_i \left(p^2\right)^2 + \sum_j \left(2pq\right)^2 \tag{25.4}$$

where:

p and q = the frequencies of two different alleles
P_m can also be used to compare the discriminating powers of different loci

Tables 25.2 through 25.4 show P_m values of loci commonly used for forensic applications, including SNP, VNTR, and STR.

Table 25.2 Heterozygosity and P_m Values of Six SNP Loci

Locus	Allele	Heterozygosity	P_m
DQA1	7	0.828	0.053
LDLR	2	0.493	0.379
GYPA	2	0.498	0.376
HBGG	3	0.508	0.360
D7S8	2	0.476	0.388
GC	3	0.592	0.235
Average		0.566	
Product			2.5×10^{-4} (1 in 4000)

Source: Office of Justice Programs, National Institute of Justice, US Department of Justice, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group*, 2000.

Table 25.3 Heterozygosity and P_m Values of Six VNTR Loci

Locus	Bins	Heterozygosity	P_m
D1S7	28	0.945	0.0058
D2S44	26	0.926	0.0103
D4S139	19	0.899	0.0184
D10S28	24	0.943	0.0063
D14S13	30	0.899	0.0172
D17S79	19	0.799	0.0700
Average		0.902	
Product			8.26×10^{-12} (1 in 1.2×10^{11})

Source: Office of Justice Programs, National Institute of Justice, US Department of Justice, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group,* 2000.

25.2 Statistical Analysis of DNA Profiling Results

It is desirable to evaluate the strength of DNA profiling results, particularly if two DNA profiles match. A DNA profile from crime scene evidence and a profile from a suspect may be the same for two reasons: (1) the crime scene sample may have come from the suspect or (2) the suspect happens to have the same profile as the individual who left the evidence found at the crime scene. The significance of a match between DNA profiles can be evaluated by using statistical calculations that determine the rarity of a specific DNA profile in a relevant population. The statistical evaluation of the significance can be included in a case report

Table 25.4 Heterozygosity and P_m Values for CODIS Loci

Locus	Allele	Caucasian American		African American	
		Heterozygosity	P_m	Heterozygosity	P_m
CSF1PO	11	0.734	0.112	0.781	0.081
TPOX	7	0.621	0.195	0.763	0.090
TH01	7	0.783	0.081	0.727	0.109
VWA	10	0.811	0.062	0.809	0.063
D16S539	8	0.767	0.089	0.798	0.070
D7S820	11	0.806	0.065	0.782	0.080
D13S317	8	0.771	0.085	0.688	0.136
D5S818	10	0.682	0.158	0.739	0.112
FGA	19	0.860	0.036	0.863	0.033
D3S1358	10	0.795	0.075	0.763	0.094
D8S1179	10	0.780	0.067	0.778	0.082
D18S51	15	0.876	0.028	0.873	0.029
D2S11	20	0.853	0.039	0.861	0.034
Average		0.7812		0.7866	
Product			1.738×10^{-15} (1 in 5.753×10^{14})		1.092×10^{-15} (1 in 9.161×10^{14})

Source: Office of Justice Programs, National Institute of Justice, US Department of Justice, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group*, 2000.

(Figure 25.2). Guidelines for DNA profile interpretation such as those issued by the National Research Council, the DNA Advisory Board, and the European DNA Profiling Group can be consulted.

25.2.1 Genotypes

The approaches to performing statistical analysis are (1) calculation of profile probability and (2) use of the likelihood ratio method. Although profile probability is the most commonly used method because of its simplicity, both approaches lead to the same conclusion.

25.2.1.1 Profile Probability

Profile probability can be calculated based on the following steps:

1. The locus genotype frequency can be calculated as follows; p and q are the allelic frequencies observed in the database for a given allele:

   Locus genotype frequency for homozygotes:

$$P_i = p^2 \tag{25.5}$$

Locus genotype frequency for heterozygotes:

$$P_j = 2pq \tag{25.6}$$

2. Profile probability can then be calculated based on the product rule by multiplying all the locus genotype frequencies calculated as above.

The lower the profile probability is, the less likely that an individual chosen at random will have a coincident match with the DNA profile of the evidence sample. Table 25.5 shows calculations of profile probability from a DNA profile.

25.2.1.1.1 Structured Populations

The above calculation of profile probability is based on the assumption that a randomly selected individual is unrelated to a perpetrator. However, it is likely that the individual and the perpetrator are from the same subpopulation (groups within a population) and are thus not completely independent. Mating is more likely to occur within subpopulations than between subpopulations. As a result, the proportion of homozygotes increases and the proportion of heterozygotes decreases in a subpopulation because individuals in a subpopulation appear to be related.

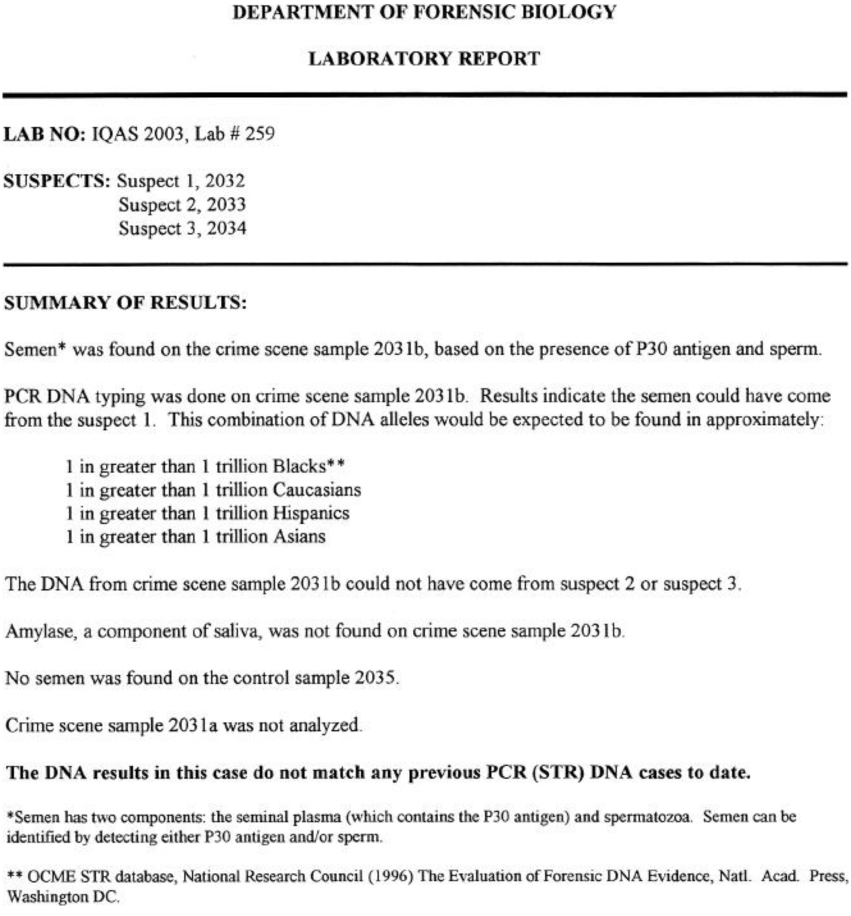

DEPARTMENT OF FORENSIC BIOLOGY

LABORATORY REPORT

LAB NO: IQAS 2003, Lab # 259

SUSPECTS: Suspect 1, 2032
Suspect 2, 2033
Suspect 3, 2034

SUMMARY OF RESULTS:

Semen* was found on the crime scene sample 2031b, based on the presence of P30 antigen and sperm.

PCR DNA typing was done on crime scene sample 2031b. Results indicate the semen could have come from the suspect 1. This combination of DNA alleles would be expected to be found in approximately:

1 in greater than 1 trillion Blacks**
1 in greater than 1 trillion Caucasians
1 in greater than 1 trillion Hispanics
1 in greater than 1 trillion Asians

The DNA from crime scene sample 2031b could not have come from suspect 2 or suspect 3.

Amylase, a component of saliva, was not found on crime scene sample 2031b.

No semen was found on the control sample 2035.

Crime scene sample 2031a was not analyzed.

The DNA results in this case do not match any previous PCR (STR) DNA cases to date.

*Semen has two components: the seminal plasma (which contains the P30 antigen) and spermatozoa. Semen can be identified by detecting either P30 antigen and/or sperm.

** OCME STR database, National Research Council (1996) The Evaluation of Forensic DNA Evidence, Natl. Acad. Press, Washington DC.

Figure 25.2 An example of a laboratory report. (© Richard C. Li.)

Table 25.5 Calculation of Profile Probability of STR Profile with 13 CODIS Loci

Locus	Profile	Allelic Frequency[a]	Formula	Locus Genotype Frequency
CSF1PO	10	0.25369	$2pq$	0.165
	12	0.32512		
D3S1358	14	0.14039	$2pq$	0.0691
	15	0.24631		
D5S818	11	0.41026	p^2	0.168
	11	0.41026		
D7S820	10	0.29064	$2pq$	0.117
	11	0.20197		
D8S1179	13	0.33929	p^2	0.115
	13	0.33929		
D13S317	11	0.31888	p^2	0.102
	11	0.31888		
D16S539	11	0.27228	$2pq$	0.185
	12	0.33911		
D18S51	15	0.12755	$2pq$	0.00911
	19	0.035710		
D21S11	30	0.23214	p^2	0.0539
	30	0.23214		
FGA	23	0.15816	$2pq$	0.0436
	24	0.13776		
TH01	8	0.12562	$2pq$	0.0767
	9.3	0.30542		
TPOX	8	0.54433	p^2	0.296
	8	0.54433		
VWA	17	0.26276	$2pq$	0.117
	18	0.22194		
				Profile probability $= 2.76 \times 10^{-14}$

[a] See Table 25.1.

The effect of population structure should be considered and an appropriate correction should be made to estimate profile probabilities. The correction can be made by using the factor θ. Thus, the locus genotype frequency can be calculated as follows:

Locus genotype frequency for homozygotes:

$$P_i = \frac{\left[2\theta + (1-\theta)p\right]\left[3\theta + (1-\theta)p\right]}{(1+\theta)(1+2\theta)} \tag{25.7}$$

Locus genotype frequency for heterozygotes:

$$P_j = \frac{2\left[\theta + (1-\theta)p\right]\left[\theta + (1-\theta)q\right]}{(1+\theta)(1+2\theta)} \tag{25.8}$$

The θ value is 0.01 for the majority US population and 0.03 for the Native American population. Table 25.6 shows the calculation of profile probability with subpopulation correction using $\theta = 0.01$. The profile probability is approximately three times higher than the value without the correction (Table 25.5). Additional corrections can be calculated for relatives, mixed stains, or database searches using formulas provided by the National Research Council's guidelines.

25.2.1.2 Likelihood Ratio

The *likelihood ratio* (LR) method is an alternative for evaluating the strength of a match. The method allows the calculation of the probability of the DNA profile under two hypothesis:

Hypothesis 1 (H_1) —The evidence and suspect profiles originated from the same source.

Hypothesis 2 (H_2) —The evidence and suspect profiles did not originate from the same source (i.e., the suspect happens to have the same profile as that of the individual who left the evidence).

The LR is the probability of hypothesis H_1 divided by the probability of hypothesis H_2. Where Pr_1 is the probability under hypothesis H_1, and Pr_2 is the probability under hypothesis H_2, this can be expressed as:

$$\mathrm{LR} = \frac{\mathrm{Pr}_1}{\mathrm{Pr}_2} \tag{25.9}$$

The greater the numerator (Pr_1), the greater the likelihood ratio becomes. The result favors hypothesis H_1 (the evidence and suspect profile originated from the same source). Pr_1 is equal to 1 (100%) when a match occurs, and Pr_2 is equal to the profile probability. A LR of 1000 indicates that the evidence is 1000 times as probable if the evidence and suspect profiles originated from the same source.

25.2.2 Haplotypes

The term *haplotype* was first used to describe very closely linked polymorphic loci. During meiosis, alleles at neighboring loci cosegregate (both alleles segregate as a single allele) because of the close linkage of loci. The term also applies to genetic regions within which recombination events are very rare, that is, within mitochondrial and Y-chromosomal DNA. The entire mitochondrial DNA (mtDNA) sequence can be considered to be a single locus or haplotype because of the absence of recombination. Likewise, Y chromosome loci can also be considered haplotypes.

Where recombination is very rare, certain allelic combinations occur in populations much more frequently than would be expected. This phenomenon is called *linkage disequilibrium*. As a result, the HW principle cannot be applied. The two methods for evaluating the strength of

Table 25.6 Calculation of Profile Probability of STR with 13 CODIS Loci and Correction Factor ($\theta=0.01$)

Locus	Profile	Allelic Frequency[a]	Formula	Locus Genotype Frequency
CSF1PO	10	0.25369	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.168
	12	0.32512		
D3S1358	14	0.14039	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.0734
	15	0.24631		
D5S818	11	0.41026	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.180
	11	0.41026		
D7S820	10	0.29064	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.121
	11	0.20197		
D8S1179	13	0.33929	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.126
	13	0.33929		
D13S317	11	0.31888	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.113
	11	0.31888		
D16S539	11	0.27228	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.187
	12	0.33911		
D18S51	15	0.12755	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.012
	19	0.03571		
D21S11	30	0.23214	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.0630
	30	0.23214		
FGA	23	0.15816	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.0473
	24	0.13776		
TH01	8	0.12562	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.0815
	9.3	0.30542		
TPOX	8	0.54433	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.309
	8	0.54433		
VWA	17	0.26276	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.120
	18	0.22194		
				Profile probability $=7.8\times10^{-14}$

[a] See Table 23.1.

a match between haplotypes are (1) mitotype frequency and (2) likelihood ratios. The current most common approach for interpreting mtDNA profiles is mitotype frequency carried out with the gene counting method, that is, the calculation of the number of occurrences of a particular sequence or haplotype. The interpretation of Y chromosome profiles is similar to the interpretation of mtDNA profiles. The estimation of the frequency of a mitotype can be calculated as shown below.

25.2.2.1 Mitotypes Observed in Database

If a mitotype is observed at least once in a database, Equation 25.10 can be used. P_{mt} is the mitotype frequency, x is the number of observations of the haplotype, and n is the size of the database (the number of mitotype entries):

$$P_{\text{mt}} = \frac{x+2}{n+2} \tag{25.10}$$

Any sampling error may be addressed by a confidence interval:

$$P_{\text{mt}} \pm 1.96\sqrt{\frac{P_{\text{mt}}\left(1-P_{\text{mt}}\right)}{n}} \tag{25.11}$$

In this case, P_{mt} is the mitotype frequency, and n is the size of the database. The conservative upper bound of the frequency is usually quoted.

25.2.2.2 Mitotype Not Observed in Database

If a mitotype has not been observed in a database, Equation 25.12 can be used to calculate the mitotype frequency; α is 0.05 for a 95% confidence interval and n is the size of the database.

$$P_{\text{mt}} = 1-\alpha^{1/n} \tag{25.12}$$

Bibliography

Aitken, C.G.G., *Statistics and the Evaluation of Evidence for Forensic Scientists*, p. 260, 1995. New York: Wiley.

Aktulga, H.M., et al., Identifying statistical dependence in genomic sequences via mutual information estimates. *EURASIP J Bioinform Syst Biol*, 2007, **2007**(1): 14741.

Allen, R.W., et al., Considerations for the interpretation of STR results in cases of questioned half-sibship. *Transfusion*, 2007, **47**(3): 515–519.

Amorim, A., A cautionary note on the evaluation of genetic evidence from uniparentally transmitted markers. *Forensic Sci Int Genet*, 2008, **2**(4): 376–378.

Andersen, M.M., P.S. Eriksen, and N. Morling, The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *J Theor Biol*, 2013, **329**: 39–51.

Andersen, M.M., et al., Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Sci Int Genet*, 2013, **7**(2): 264–271.

Balding, D.J., The prosecutor's fallacy and DNA evidence. *Crim Law Rev*, 1994, 711–721.

Balding, D.J., When can a DNA profile be regarded as unique? *Sci Justice*, 1999, **39**(4): 257–260.

Balding, D.J. and P. Donnelly, Inferring identify from DNA profile evidence. *Proc Natl Acad Sci U S A*, 1995, **92**(25): 11741–11745.

Balding, D.J. and P. Donnelly, Evaluating DNA profile evidence when the suspect is identified through a database search. *J Forensic Sci*, 1996, **41**(4): 603–607.

Balding, D.J. and R.A. Nichols, DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci Int*, 1994, **64**(2–3): 125–140.

Beecham, G.W. and B.S. Weir, Confidence interval of the likelihood ratio associated with mixed stain DNA evidence. *J Forensic Sci*, 2011, **56**(Suppl 1): S166–S171.

26

Quality Assurance and Quality Control

Quality assurance (*QA*) for forensic services requires certain processes to ensure that a service will meet laboratory requirements for the integrity of testing. A QA program must include components that address:

- Continuing education, training, and certification of personnel
- Specification and calibration of equipment and reagents
- Documentation and validation of analytic methods
- Use of appropriate standards and controls
- Sample handling procedures
- Proficiency testing
- Data interpretation and reporting
- Audits (internal and external) and laboratory accreditation
- Corrective actions to address deficiencies and assessments for laboratory competence

Over the years, many guidelines for quality assurance in forensic DNA laboratories have been established. These guidelines will be introduced in this chapter.

Quality control (*QC*) for forensic services refers to the operational procedures necessary to meet quality requirements. QC procedures may include maintenance of calibration records for equipment and instruments as well as the testing of chemical reagents and supplies used in analysis to ensure reliable results.

26.1 US Quality Standards

DNA profiling methods were first used in criminal investigations in the 1980s. By the early 1990s, emerging forensic DNA techniques had undergone detailed reviews by the *National Research Council* (*NRC*) of the National Academy of Sciences. In 1992, the first published NRC report included recommendations in the areas of (1) technical considerations, (2) statistical interpretation, (3) laboratory standards, (4) data banks and privacy, (5) legal considerations, and (6) societal and ethical issues related to forensic DNA testing. The NRC report attempted to explain the basic scientific principles of forensic DNA technology and made suggestions for applications and improvements. However, the report received negative criticism from both the forensic and the legal communities.

In 1996, a second NRC committee was formed "to update and clarify discussion of the principles of population genetics and statistics as they apply to DNA evidence." Its report stated:

> The central question that the report addresses is this: What information can a forensic scientist, population geneticist, or statistician provide to assist a judge or jury in drawing inferences from the finding of a match? To answer this question, the committee reviewed the scientific literature and the legal cases and commentaries on DNA profiling, and it investigated the criticisms that have been voiced about population data, statistics, and laboratory error. Much has been learned since the last report. The technology for DNA profiling as well as the methods for estimating frequencies and related statistics have progressed to the point where the reliability and validity of properly collected and analyzed DNA data should not be in doubt. The new recommendations presented here should pave the way to more effective use of DNA evidence. (*The Evaluation of Forensic DNA Evidence* also known as the NRC II report)

The NRC II report consisted of (1) an introduction describing the 1992 report, changes made subsequent to that report, and the validity and application of DNA typing techniques; (2) assurance of high standards of laboratory performance; (3) population genetics issues; (4) statistical issues; and (5) DNA evidence in the legal system.

In 1995, The *DNA Advisory Board* (*DAB*) was formed as a result of the DNA Identification Act (1994) passed by Congress. The DAB served from 1995 to 2000 to develop guidelines for quality assurance in forensic laboratories. During that time, the DAB provided two sets of guidelines: Quality Assurance Standards for Forensic DNA Testing Laboratories (1998) and Quality Assurance Standards for Convicted Offender DNA Databasing Laboratories (1999). These standards describe the requirements to ensure quality and integrity of the data as well as competency of laboratories.

These standards were built on the previous standards set by the *Scientific Working Group on DNA Analysis Methods* (*SWGDAM*). After the DAB's assignment ended in 2000, the SWGDAM became responsible for providing guidelines to the US forensic community. The SWGDAM was established in 1988 by the FBI Laboratory. It was initially called the Technical Working Group on DNA Analysis Methods; the name changed in 1998. The SWGDAM comprises forensic scientists from DNA laboratories in the United States and Canada. Its purpose is to facilitate forensic DNA community discussions regarding necessary laboratory methods and to share protocols for forensic DNA testing. The FBI sponsors and hosts its meetings and plays an important role in its activities.

The SWGDAM established and revised several guidelines including the Guidelines for a Quality Assurance Program for DNA Analysis, published in 1989, 1991, and 1995 (the validation section was revised in 2003); Quality Assurance Standards for Forensic DNA Testing Laboratories (2011); Quality Assurance Standards for DNA Databasing Laboratories (2011); Validation Guidelines for DNA Analysis Methods (2012); and SWGDAM Training Guidelines, published in 2001 and revised in 2013.

The SWGDAM also formed subcommittees to provide guidelines in more specific areas of forensic DNA testing, for example, SWGDAM's Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories (2010), Y-chromosome Short Tandem Repeat (Y-STR) Interpretation Guidelines (2009), and Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories (2013). The SWGDAM also organized a number of interlaboratory and validation studies for new techniques.

26.2 International Quality Standards

In the early stages of forensic DNA testing, the *International Society for Forensic Genetics* (*ISFG*) recognized the potential of DNA testing for criminal investigations and made a number of recommendations related to the forensic application of DNA polymorphisms.

The ISFG provided various recommendations for forensic DNA testing of STR, mtDNA, and Y chromosome markers for the international community. It formed a working group called the *European DNA Profiling Group* (*EDNAP*) in 1991. The EDNAP has investigated systems for DNA profiling, has organized a number of collaborative exercises for the evaluation of new methods, and has published reports of its studies.

The *European Network of Forensic Science Institutes* (*ENFSI*) formed its DNA working group about a decade ago to address issues of quality and standards for forensic DNA testing. The *Interpol European Working Party on DNA Profiling* (*IEWPDP*) also makes recommendations for applying DNA evidence to criminal investigations in Europe. Based on the EDNAP exercises and recommendations by the ENFSI and IEWPDP, the *European Standard Set* (*ESS*) for autosomal STR core loci was established. The Standardization of DNA Profiling Techniques in the European Union (STADNAP) group has been working on the selection of forensic DNA profiling systems, methods for use among European countries, and the maintenance of European population databases.

26.3 Laboratory Accreditation

Accreditation is the process used to assess the qualification of a laboratory to meet established standards. During an accreditation process, the services and performance of a laboratory are evaluated, particularly in the areas of management, operations, personnel, procedures, equipment, physical plants, security, and personnel safety procedures. The accreditation process generally involves several components such as self-evaluation, the preparation of supporting documents, on-site inspection and reports, and accreditation review reports.

Accreditation in the United States is offered by the *Laboratory Accreditation Board* of the *American Society of Crime Laboratory Directors* (*ASCLD/LAB*) for forensic laboratories performing casework. The *American Association of Blood Banks* (*AABB*) provides accreditation for laboratories performing DNA parentage testing according to the AABB's standards.

The accreditation of a forensic laboratory is granted for 5 years. To remain in compliance, a laboratory must undergo audits to evaluate its operation according to established guidelines. The FBI has published the Quality Assurance Standards Audit for Forensic DNA Testing Laboratories (2011) and the Quality Assurance Standards Audit for DNA Databasing Laboratories (2011). The areas of operating protocols, instruments and equipment, and personnel training are evaluated based on guidelines. Problems identified during an audit must be documented and actions to resolve the problems must be addressed. Annual internal audits and external audits during alternate years are required under the guidelines.

26.4 Laboratory Validation

Validation is the process of confirming that a laboratory procedure is sufficiently robust, reliable, and reproducible. A robust method maintains successful performance and can cope with errors. A reliable method produces accurate results. A reproducible method achieves the same or very similar results each time a sample is tested.

Two types of laboratory validations, developmental and internal validations, are used for modifying methods for forensic DNA analysis. An *internal validation* is required when adopting a procedure for forensic applications. Based on the SWGDAM's Validation Guidelines for DNA Analysis Methods, the internal validation is "an accumulation of test data within the laboratory to demonstrate that established methods and procedures perform as expected in the laboratory." The *developmental validation* "is the acquisition of test data and determination of conditions and limitations of a new or novel DNA methodology for use on forensic, database, known or casework reference samples." For example, the characterization of a new genetic marker requires studies in the inheritance, the genomic location, the detection method, and the polymorphism of the genetic marker. However, developmental validation studies in the same subject but from fields other than forensic DNA analysis may be acceptable for forensic applications.

Additionally, the precision and the accuracy of the test are important criteria for developmental validation. Based on the SWGDAM, *precision* "characterizes the degree of mutual agreement among a series of individual measurements, values and/or results. Precision depends only on the distribution of random errors and does not relate to the true value or specified value." The measure of precision can be expressed as a standard deviation of test results. *Accuracy* is referred to as "the ability of a measuring instrument to give responses close to a true value," which can be assessed by a *performance check*, a quality assurance procedure monitoring the performance of instruments and equipment affecting the accuracy of forensic DNA testing.

26.5 Proficiency Testing

Proficiency testing is an important component of quality control and quality assurance. It evaluates a laboratory's performance of DNA analyses according to the laboratory's standard protocols. Proficiency testing also evaluates the quality of performance by individual analysts following laboratory protocols.

Proficiency tests of DNA analysts must be conducted every 6 months based on DNA Advisory Board Standards. The testing usually involves mock forensic case samples including questioned bodily fluid stains and reference samples. The test is assigned to an analyst for processing according to the laboratory procedures. A report must be prepared and is then reviewed. The proficiency test can be administered as either an open or a blind test. In the blind test, the analyst is not aware that he or she is being tested. Blind testing is considered a more effective means of evaluating performance.

The tests may be administered internally or by any of a number of external testing organizations. For example, Orchid Cellmark provides the *International Quality Assessment Scheme* (*IQAS*) DNA proficiency test, and the Collaborative Testing Services (CTS) Forensics Testing Program offers similar proficiency tests. These proficiency tests may include a selection of sample types (neat or mixture) for serological tests and/or DNA analysis (autosomal STR, Y-STR, and mitochondrial DNA). The College of American Pathologists (CAP) has a paternity testing proficiency program offering external proficiency tests. In Europe, the *German DNA Profiling Group* (*GEDNAP*) provides proficiency testing for participating European laboratories.

26.6 Certification

Certification is a voluntary process that recognizes the attainment of professional qualifications needed for practice in forensic services. Certification is not required, but is desired by some laboratories. In 2009, a report entitled Strengthening Forensic Science in the United States: A Path Forward was published by the Committee on Identifying the Needs of the Forensic Science Community at the National Academy of Sciences. The report recommends that certification should be mandatory for forensic science professionals, which includes written examinations, supervised practice, proficiency testing, and compliance to a code of ethics (Section 26.8). In the United States, the *American Board of Criminalistics* (*ABC*) offers three types of certification for forensic scientists. A diplomate must pass a general knowledge examination. ABC also requires a bachelor's degree in a natural science and 2 years of experience in a forensic laboratory. To obtain fellow status (higher than diplomate status), an applicant must have 2 years of experience in his or her specialty and must have met the diplomate requirements in addition to passing a written specialty examination and a proficiency test.

ABC has added a third certification that is completely separate from the other certification programs. The technical specialist certification for molecular biology was created to recognize the qualifications required for the analysis of biological materials through DNA profiling. An applicant must take a specialist examination containing questions from the general knowledge examination and a subset of questions from the forensic biology fellow examination. This certification also requires a bachelor's degree in a natural science and 2 years of experience along with successful completion of a proficiency examination within 12 months of taking the technical specialist certification examination.

26.7 Forensic DNA Analyst Qualifications

DAB's Quality Assurance Standards for Forensic DNA Testing Laboratories require that an examiner or analyst have "at a minimum a BA/BS degree or its equivalent degree in a biology, chemistry or forensic science-related area and must have successfully completed college course work (graduate or undergraduate level) covering the subject areas of biochemistry, genetics and molecular biology," as well as "course work and/or training in statistics and population genetics as it applies to forensic DNA analysis."

Additionally, SWGDAM Training Guidelines discuss the course work requirements in detail. At least nine cumulative semester hours are required for the course work covering the required subject areas. Particularly, the section in Fundamental Scientific Knowledge of the SWGDAM Training Guidelines defines required key elements to be covered in course work. The key elements also aid in evaluating the contents of the course work.

Biochemistry "refers to the nature of biologically important molecules in living systems, DNA replication and protein synthesis, and the quantitative and qualitative aspects of cellular metabolism," and may include but is not limited to:

- Structure and function of cellular components such as proteins, carbohydrates, lipids, nucleic acids, and other biomolecules
- Chemistry of enzyme-catalyzed reactions
- Metabolism
- DNA and RNA
- Protein synthesis
- Cell membrane transport
- Signal transduction

Genetics "refers to the study of inherited traits, genotype/phenotype relationships, and population/species differences in allele and genotype frequencies," and may include but is not limited to:

- Heredity
- Function of genes
- Gene expression
- Recombinant DNA
- Mitosis/meiosis

Molecular biology "covers theories, methods, and techniques used in the study and analysis of gene structure, organization, and function," and may include, but is not limited to:

- Interrelationship of DNA, RNA, and protein synthesis
- Central dogma

- Transcription, translation, replication
- Recombinant DNA techniques
- PCR
- Cloning

Course work that is similar in content and scope to those described above may be qualified as required course work. Compliance with the required course contents can be evaluated through transcripts, syllabi, and letters from the instructors.

DAB's Quality Assurance Standards for Forensic DNA Testing Laboratories also describes requirements for an analyst to meet prior to initiating independent casework analysis using DNA technology. First, an analyst must have at least 6 months of forensic DNA laboratory experience "including the successful analysis of a range of samples typically encountered in forensic casework." Second, an analyst must have "successfully completed a qualifying test before beginning independent casework responsibilities."

26.8 Code of Ethics of Forensic Scientists

Ethical codes for forensic scientists are used as guides for individuals making their decisions in distinguishing the difference between correctness and incorrectness. These codes are usually adopted by forensic organizations to regulate the profession. Failure to comply with a code of forensic practice can raise doubt in an individual's fitness for providing forensic services. Additionally, it may result in the expulsion of an individual from a forensic organization. A typical document of ethical codes for forensic scientists generally contains five sections: (1) ethics relating to the scientific method, (2) ethics relating to opinions and conclusions, (3) ethical aspects of court presentation, (4) ethics relating to the general practice of forensics, and (5) ethical responsibilities to the profession.

For the first section, the ethical application of scientific methods is discussed in detail. True scientific methods should be utilized to adequately analyse all the evidence. Such analyses should not be conducted by secret processes. Conclusions should be drawn from the analyses of evidence that appears representative, typical, or reliable.

Moreover, ethical standards relating to opinions and conclusions are very important to the forensic profession. Conclusions should always be drawn from the application of proven scientific methods. The purpose of experimental design and the interpretation of results is to reveal facts. During an analysis, experimental controls should always be utilized. If necessary, the results of analyses should be verified by repeating the analysis or using additional techniques. Explanations should be provided where inconclusive results are obtained. The opinion provided by a forensic scientist should be unbiased and should not be influenced by matters unrelated to the specific evidence in question. A forensic scientist should not choose the interpretation that is in favor of the side of his or her employer.

Pertaining to the courtroom, ethical standards relating to expert witness testimonies is a crucial component for a forensic scientist. An expert opinion is defined as a formal consideration of a subject within an individual's knowledge, skill, education, training, and experience. A forensic scientist should not extend an opinion beyond his or her competence. Appropriate terms should be used to represent the degrees of certainty of an expert opinion. A forensic scientist should not only present the evidence that supports the view of the side of his or her employer. When testifying as an expert witness, language that can be understood by lay jurors should be used to avoid misinterpretation.

Furthermore, in terms of the general practice of forensic science, a few things should be mentioned here. A forensic scientist in private practice should set a reasonable fee for the services provided, which should not be rendered on a contingency fee basis. When a different opinion is presented by another analyst in the reexamination of evidence, both analysts should resolve their contradiction before the trial. A forensic scientist may serve, in an advisory capacity, either the prosecutor or defense in the cross-examination of another expert.

Lastly, ethical responsibilities to the profession are an important component of ethical codes. Information regarding new developments or techniques of forensic analysis should be made available to other forensic scientists. Likewise, any information regarding methods in use that may appear unreliable should be brought to the attention of others. Individual forensic scientists should not seek publicity for the association of their name with specific cases, developments, publications, or organizations solely for the purpose of gaining personal prestige.

Bibliography

Andersen, J., et al., Report on the third EDNAP collaborative STR exercise. *Forensic Sci Int*, 1996, **78**(2): 83–93.

Balazic, J. and I. Zupanic, Quality control and quality assurance in DNA laboratories: Legal, civil and ethical aspects. *Forensic Sci Int*, 1999, **103**: S1–S5.

Bar, W., et al., DNA recommendations. Further report of the DNA commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics. *Int J Legal Med*, 1997, **110**(4): 175–176.

Buckleton, J., Validation issues around DNA typing of low level DNA. *Forensic Sci Int Genet*, 2009, **3**(4): 255–260.

Budowle, B., et al., A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. *J Forensic Sci*, 2009, **54**(4): 798–809.

Carracedo, A., et al., DNA commission of the International Society for Forensic Genetics: Guidelines for mitochondrial DNA typing. *Forensic Sci Int*, 2000, **110**(2): 79–85.

Carracedo, A., et al., Results of a collaborative study of the EDNAP group regarding the reproducibility and robustness of the Y-chromosome STRs DYS19, DYS389 I and II, DYS390 and DYS393 in a PCR pentaplex format. *Forensic Sci Int*, 2001, **119**(1): 28–41.

Cormier, K., L. Calandro, and D.J. Reeder, Evolution of the quality assurance documents for DNA laboratories. *Forensic Mag*, 2005, 2(1): 16–19.

Council, N.R., *NRC II: The Evaluation of Forensic DNA Evidence*, 1996. Washington, D.C.: National Academy Press.

Coyle, H.M., The importance of scientific evaluation of biological evidence—Data from eight years of case review. *Sci Justice*, 2012, **52**(4): 268–270.

DNA Advisory Board, Federal Bureau of Investigation, US Department of Justice, Quality assurance standards for forensic DNA testing laboratories. *Forensic Sci Commun*, 2000, **2**(3). Retrieved from http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2000/codis2a.htm.

Duewer, D.L., K.L. Richie, and D.J. Reeder, RFLP band size standards: NIST standard reference material 2390. *J Forensic Sci*, 2000, **45**(5): 1093–1105.

Duewer, D.L., et al., NIST mixed stain studies #1 and #2: Interlaboratory comparison of DNA quantification practice and short tandem repeat multiplex performance with multiple-source samples. *J Forensic Sci*, 2001, **46**(5): 1199–1210.

Frank, W.E., et al., Validation of the AmpFlSTR Profiler Plus PCR amplification kit for use in forensic casework. *J Forensic Sci*, 2001, **46**(3): 642–646.

Gill, P., et al., Report of the European DNA profiling group (EDNAP)—Towards standardisation of short tandem repeat (STR) loci. *Forensic Sci Int*, 1994, **65**(1): 51–59.

Gill, P., et al., Report of the European DNA profiling group (EDNAP): An investigation of the complex STR loci D21S11 and HUMFIBRA (FGA). *Forensic Sci Int*, 1997, **86**(1–2): 25–33.

Gill, P., et al., Report of the European DNA profiling group (EDNAP)—An investigation of the hypervariable STR loci ACTBP2, APOAI1 and D11S554 and the compound loci D12S391 and D1S1656. *Forensic Sci Int*, 1998, **98**(3): 193–200.

Gill, P., et al., DNA commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Int J Legal Med*, 2001, **114**(6): 305–309.

FORENSIC SCIENCE

FORENSIC BIOLOGY

SECOND EDITION

Over the last several years, new research and developments in analysis methods and practice have led to rapid advancements in forensic biology. Identifying critical points of knowledge and new methodological approaches in the field, ***Forensic Biology, Second Edition*** focuses on forensic serology and forensic DNA analysis. It provides students and professionals with a scientific grounding in biological evidence—both the techniques utilized to identify it and the methodology to analyze it.

This second edition:

- Introduces the language of forensic biology, enabling students to become comfortable with usage and terminology
- Provides clear explanations of the principles of forensic identification and analysis of biological evidence
- Explains forensic serology and DNA techniques used in the field and the laboratory
- Discusses the benefits and limitations that apply to various forensic biology techniques
- Includes schematic illustrations to clarify concepts
- Presents three new chapters (Chapters 2, 16, and 17) created for this edition
- Adds more than two hundred new full-color figures

Covering the full scope of forensic biology, the book uses an accessible, straightforward style designed to enhance students' education and training so they are prepared, both in the laboratory and in the field.

K14098

ISBN: 978-1-4398-8970-1
90000
9 781439 889701

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

www.crcpress.com

Color Atlas of Forensic Medicine and Pathology

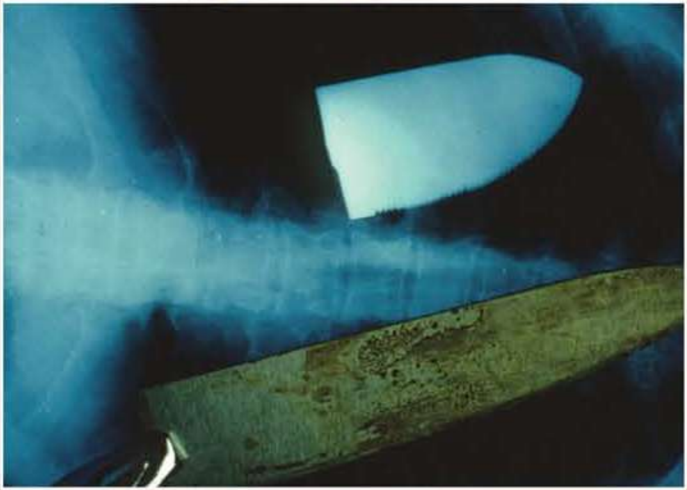

Edited by

Charles A. Catanese

CRC Press
Taylor & Francis Group

Table of Contents

Preface ix

Acknowledgments xi

Editor xiii

Contributors xv

Research Assistants and Photo Imaging Contributors xvii

1 Sudden Natural Death in a Forensic Setting 1
CHARLES A. CATANESE AND AMY V. RAPKIEWICZ

2 Therapeutic Interventions, Complications, and Accidents 57
MICHAEL J. CAPLAN AND CHARLES A. CATANESE

3 Substance Abuse and Poisoning 85
CHARLES A. CATANESE AND LAURA M. LABAY

4 Postmortem Change and Time of Death 111
CHARLES A. CATANESE, BRUCE LEVY, AND GERARD CATANESE

5 Pediatric Forensic Pathology 147
MICHAEL J. CAPLAN AND CHARLES A. CATANESE

6 Blunt-Force Injuries 189
CHARLES A. CATANESE AND JONATHAN LUCAS

7 Sharp-Force Injuries 239
CHARLES A. CATANESE

8 Gunshot Wounds 283
CHARLES A. CATANESE, THOMAS GILSON, AND THOMAS ANDREW

9 Burns 353
CHARLES A. CATANESE AND GERARD CATANESE

10 Asphyxia 373
CHARLES A. CATANESE AND BARBARA K. BOLLINGER

Index 399

Preface

While death investigation, and what we have come to understand as "forensic pathology," has been practiced in one way or another back to antiquity, official designation of forensic pathology as a subspecialty by the American Board of Pathology dates back only to 1956. The subspecialty comprises a small band of trainees in pathology and at any given time there are only 400 to 500 full-time practitioners of the specialty in the United States.

Furthermore, as a recognized subspecialty, forensic pathology is young enough that each of its full-time practitioners can trace his or her roots back to one of the six individuals who sat for that original examination back in 1956. For many of the authors of this volume, that family tree goes from Lester Adelson through Charles Hirsch and ultimately to us. We acknowledge and are proud of the fact that we stand on the shoulders of giants.

Our collective experiences as trainees and staff at the Office of the Chief Medical Examiner in New York City under the tutelage of Dr. Charles S. Hirsch has made us the forensic pathologists we are today. Our mentor's emphasis on precision and accuracy in description of findings, translation of these descriptions into language easily understood by a broad range of end users, and the public health importance of our work has left an indelible imprint that we desire to pass on to others. It is our hope that this atlas reflects these qualities.

With sincerest gratitude,

Thomas Andrew
Fellowship Class of 1992

Michael J. Caplan
Gerard Catanese
Bruce Levy
Fellowship Class of 1993

Thomas Gilson
Charles A. Catanese
Fellowship Class of 1995

Barbara K. Bollinger
Jonathan Lucas
Fellowship Class of 2000

Sudden Natural Death in a Forensic Setting

1

CHARLES CATANESE AND AMY V. RAPKIEWICZ

Introduction

This chapter offers a brief overview of some common and some not so common natural deaths that typically may occur in a medical examiner system. Also demonstrated are examples that may alter appearance of tissue such as formaldehyde fixation and variation due to different types of photographic imagery. There are also examples of normal organs in both a fresh and formaldehyde fixed state that can be used by the reader to compare with diseased organs.

Deaths under this category are often unexpected and sometimes unwitnessed. There is often a suspicion of foul play. Families may say, "But Doctor, he was in fine health. I saw him an hour ago. It cannot be natural. Somebody must have harmed him," etc. Because of the sudden, often unexpected nature of these deaths, it is best to do an autopsy to clarify exactly what happened. This decision to autopsy depends on many factors, including the decedent's age, medical history, family wishes, decedent's wishes (wills, etc.), religious beliefs, circumstance at time of death, resources of a particular system, etc. As one becomes less certain of the cause of death, the level of suspicion will increase. At some point, the decision to autopsy becomes obvious and absolutely necessary. This decision is based on experience, knowledge, and sound judgment. Not infrequently, seemingly natural deaths can have unnatural or traumatic previous circumstances; therefore, when uncertain, an autopsy is best performed. In many medical examiner systems, the majority of deaths end up being certified as natural.

Sudden death is defined in different ways. It may indicate a death that occurs within 24 hours of the onset of symptoms. It may also indicate the death occurred within 1 hour or even within seconds. There are not many diseases that can cause death within minutes of the onset of symptoms. Natural death means the manner of death is exclusively or 100% natural. If there is a 1% component of another manner of death, it is no longer natural. If there are multiple components of different manners of death commingled in a case investigation, the following rule will apply: a homicide overrides all, then an accident, then a therapeutic complication. A suicide requires the establishment of intent to do harm to oneself. For example, someone with end-stage metastatic liver cancer ingests 100 acetaminophen tablets to commit suicide. In the process of waiting to die, he decides to walk to a store. On the way, he trips in a pothole, falls in the street, strikes his head, and has an expanding subdural hemorrhage. While he is lying there waiting for EMS, a stolen car fleeing the scene of a robbery runs him over, lacerates his heart in half, and he dies within seconds. The manner of death in the case would be homicide. The death occurred as a result of being run over by a car during an illegal act. A lacerated heart is universally fatal regardless of the other violent and natural processes. The death certificate should include only the trauma from the car. If the trauma from the car was not lethal by itself, one may add "other finding" to part two of the cause of death, but the manner would remain homicide.

Heart Disease

Heart disease leading to ventricular irritability to create a lethal arrhythmia is the most significant cause of death in this category. The most common arrhythmia leading to sudden cardiac death is ventricular fibrillation. Ventricular tachyarrhythmias are most commonly seen within 12 hours of a myocardial infarction. Critical coronary atherosclerosis and hypertension are by far the leading causes of these processes. Some diseases that contribute to atherosclerosis and arteriosclerosis formation include hyperlipidemia, high blood pressure, diabetes mellitus, obesity, cigarette smoking, stress, and sedentary life style.

Having 75% or greater blockage in any of the epicardial vessels is considered critical stenosis and is consistent with being alive one second and having loss of consciousness leading to death the next. Hypertensive cardiovascular disease is usually essential in origin from an intrinsic abnormality of sodium metabolism. Other significant causes of hypertension include many types of kidney disease including adult polycystic kidney disease and renal artery stenosis. Hypertension may be sporadic and missed on routine doctor appointments. High blood pressure is also associated with small-vessel coronary artery disease, as is diabetes mellitus, which is a reasonable cause of death by itself. Once people reach a pivotal point of myocardial irritability and go into ventricular fibrillation they usually have approximately 15 seconds of consciousness left. Prior to losing consciousness, decedents may reach up to chest or neck and mention a fluttering sensation in the chest. They may have pressure, pain, or no expectation of what is to come. Ventricular irritability associated with coronary artery ischemia is due to lack of oxygen and nutrients reaching the conducting system of the heart. If the heart is not cardioverted back to a normal rhythm within 4–6 minutes, there is usually irreversible brain damage.

Another major cause of ventricular irritability leading to fatal arrhythmia is hypertension. Concentric left ventricular hypertrophy usually defined at autopsy as having a left ventricular wall thickness greater than 1.5 cm for most average-sized adults is a known risk factor for sudden cardiac death. Left ventricular thickness is best measured approximately 2 cm below the mitral valve annulus and excludes trabeculations and papillary muscles. As the disease process causing cardiac hypertrophy advances, heart failure may ensue with chamber dilatation. Although the overall heart size is enlarged, the left ventricle wall thickness may be less than 1.4 cm. Although hypertensive disease is the major risk factor for the development of left ventricular hypertrophy, other risk factors include aortic stenosis, either congenital or acquired. The hearts of patients with hypertensive or arteriosclerotic cardiovascular disease typically show evidence of prior infarction and interstitial fibrosis. Both findings also predispose to myocardial irritability and fatal (tachy)arrythmias.

Complications other than tachyarrhythmia and pump failure of myocardial infarctions can result in sudden cardiac death; the most common include the myocardial rupture syndromes including ventricular wall and papillary wall rupture. Typically, these insults occur approximately 1 week following a myocardial infarction, the point at which there is removal of necrotic myocytes by macrophages. Hemopericardium with ensuing cardiac tamponode can occur following ventricular free wall rupture; this scenario is rapidly fatal in most cases, causing decreased venous return to the heart with jugular venous distention.

In young patients, particularly athletes, hypertrophic cardiomyopathy is not an uncommon cause of sudden death. These patients can be asymptomatic prior to the sudden event or may have past episodes of palpitations or syncope. Typically, macroscopic heart evaluation shows cardiac hypertrophy with significant asymmetry of the subaortic septal region, which poses as an outflow obstruction. Microscopic sections from this region show variable degrees of myocyte disarray, fibrosis, myocyte hypertrophy, and small-vessel disease. The disease is due to an autosomal dominant mutation in the cardiac sarcomere apparatus, most commonly the myosin heavy chain, but many mutations have been described.

Arrhythmogenic right ventricular cardiomyopathy can present with sudden unexpected death. At autopsy, the right ventricle is thinned, with microscopic evaluation showing significant transmural infiltration by fibrofatty tissue.

Myocarditis due to a variety of causes including viral, bacterial, fungal, parasitic, autoimmune, and hypersensitivity can present as sudden death. The degree of activity, myonecrosis, and the location of the inflammation (i.e., conduction system involvement) are important in determining the significance of the infiltrates. Notably, eosinophils are seen quite commonly in hypersensitivity myocarditis and can be a clue to the underlying etiology.

Dilated cardiomyopathy is common, and has many etiologies that include idiopathic arteriosclerotic disease, hypertensive cardiovascular disease, alcoholism, elevated catecholamines, myocarditis, postpartum, doxorubicin, endocrinopathies, and genetic diseases. The heart typically is enlarged with a globoid configuration. The microscopic analysis shows interstitial fibrosis.

Rare infiltrative cardiac disease such as amyloidosis, hemochromatosis, primary or metastatic tumors, and sarcoidosis can result in sudden death. Microscopic evaluation in these cases is necessary, with particular attention to nodal tissues.

S. aureus is the most common organism found in infective endocarditis (IE). *S. aureus* endocarditis is associated with the highest mortality and risk of embolism. Increasing age, periannular abscess, heart failure, and absence of surgical therapy were identified in multivariate analysis as independent poor prognostic factors for increased mortality in patients with *S. aureus* IE. Other risk factors for the development of IE include congenital or acquired anatomic valve abnormalities such as stenosis. Impaired cardiac conductivity and function with heart failure not infrequently develops in patients with multiple septic myocardial emboli and infarcts due to IE, particularly with paravalvular abscess formation. According to a recent study of a cohort of 606 cases of infective endocarditis, 99 cases have embolization, of which 32 cases involve the central nervous system (CNS) with significantly higher mortality (65%) than those without CNS emboli.

Recently, genetic abnormalities have been found to underlie many of the intrinsic abnormalities of conducting systems including Wolff–Parkinson–White syndrome (WPW) and long Q-T syndrome. Sudden death in WPW is thought to occur as a result of an induction of ventricular tachycardia via an atrioventricular reentry pathway. Long-QT syndrome can also present with sudden death. Investigations are ongoing around the association of sudden infant death syndrome with long-QT syndrome. Recent data is suggesting that a genetic basis for the arrhythmogenic disease with the identification of the long-QT genes.

Sudden death related to cardiac valve pathology other than endocarditis is relatively uncommon, as valve replacement surgery has become a standard therapy. Patients with aortic stenosis, especially when acutely symptomatic, can experience sudden cardiac death. Most cases of aortic stenosis are caused by either rheumatic heart disease or valve calcification, which can occur on trileaflet or congentially (uni)bicuspid valves. The mechanism for death in severe aortic stenosis (valve area <1cm2) appears to be through left ventricular hypertrophy and subsequent myocardial instability. In rare instances of severe aortic valve calcification, the deposits can erode the region and involve the conduction system. Mitral valve prolapse has long been associated with sudden cardiac death. The underlying etiology is not well understood, but seems to most frequently involve a severe valve deformity with a redundant, thickened, myxomatous mitral valve and ventricular arrhythmias such as ventricular fibrillation. On histologic sectioning, the mitral valve will show deposition of acid mucopolysacchrides.

Coronary artery anomalies are not uncommon but only certain anomalies result in ischemia such as anomalous origin of a coronary artery from the opposite sinus (ACAOS), anomalous left coronary artery from the pulmonary artery (ALCAPA), ostial atresia/stenosis, and coronary artery fistulas. Left-sided ACAOS can result in acute takeoff angles with an increased risk of sudden death during or shortly after exercise. Besides the acute angle take off, there maybe ridge like defect at the coronary ostea further decreasing blood flow in times of accelerated heart rates with increased oxygen demand. Myocardial tunneling is another anomalous coronary artery distribution that maybe associated with increased arrhythmogenic potential. There is debate about the significance of this anomaly. Some still believe it may be significant when a large portion of the epicardial coronary artery dips deeply into the left ventricle wall for a considerable distance, during times of rapid muscle contraction.

Vascular Disease

Causes of sudden death associated with vascular disease include those that lead to occlusion, narrowing, or rupture of a blood vessel. Atherosclerotic aneurysms can rupture, leading to rapid loss of consciousness and death. These aneurysms can occur just about anywhere, but are by far most common in the abdominal aorta. Most abdominal aortic aneurysms occur below the renal artery. The risk of rupture increases with the size of the aneurism, smoking history, and hypertension. The annual risk of rupture over 7 cm in size is 33%. Retroperitoneal rupture is typically associated with hematoma formation, whereas rupture into the abdominal cavity can be rapidly fatal, with hemoperitoneum and shock. Patients who have a ruptured aortic aneurysm and reach the hospital have a 50% mortality rate, with the overall mortality rate greater than 85%.

Aortic dissection is characterized by an intimal tear followed by a dissection of blood within the wall of the aorta, most commonly the tunica media. Rupture of this dissecting aortic hematoma may lead to hemothoraces, hemopericardium, or fatal arrhythmia. Aortic dissection is a major cause of sudden death, mostly in patients over 50 years of age with the underlying risk factor being essential hypertension. However, pregnant women and patients with connective-tissue diseases such as Marfan's syndrome also make up a significant affected patient population. Aortic dissection can also occur following accidental or iatrogenic trauma to the aortic intima. In younger patients and those with connective tissue disease, microscopy may reveal cystic medial degeneration of the aortic media.

Most spontaneous subarachnoid hemorrhages (SAH) (90%) are caused by ruptured intracranial saccular (berry) aneurysms. SAH occurs at a peak age of 55–60 years. Rupture of an intracranial aneurysm is believed to account for 0.4 to 0.6% of all deaths. SAH is associated with a greater than 50% mortality rate. Some hospital-based studies suggest that approximately 10% of patients with aneurismal SAH die prior to reaching the hospital, 25% die within 24 hours of SAH onset, and about 45% die within 30 days. It is not unusual to perform forensic autopsies where death was almost instantaneous and outside of a hospital. The mechanism of death in such cases is cardiac arrhythmia, which is described in greater depth later. Most intracranial aneurysms (approximately 85%) are located in the anterior circulation, predominately on the circle of Willis. Risk factors for both SAH and intracranial aneurysms are similar and include hypertension, cigarette smoking, and alcohol consumption. Atherosclerosis is an independent risk factor for the development of intracranial aneurysms. The natural history of subarachnoid hemorrhage shows that rupture often occurs when they reach a size over 7 mm. Rupture of an aneurysm releases blood directly into the cerebrospinal fluid (CSF) under arterial pressure. The blood spreads quickly within the CSF, rapidly increasing intracranial pressure. A major symptom associated with SAH includes patients describing the worst headache of one's life. Increased intracranial pressure is associated with the Cushing's triad (hypertension, bradycardia, and abnormal respiration). SAH is associated with cerebral edema and subsequent herniation. Tonsillar and central transtentorial herniation is associated with compression of cardiovascular and respiratory centers in the medulla and as such is rapidly fatal. Other less common causes of subarachnoid hemorrhages include angiomas and arteriovenous malformations. Ruptured berry aneurysms are the most common natural cause of SAH, whereas trauma is the most common overall cause. Ruptured berry aneurysms are a leading cause of sudden death in women during sexual activity, whereas for men it is heart disease.

Cerebrovascular accidents (episodes), which include ischemic or intracerebral hemorrhage, can lead to sudden death. I recommend not using the term "accident" because there is nothing accidental about this process and its use often adds confusion in forensic proceedings. The terms "stroke" or "event" as an alternative is less confusing to nonmedical personnel. Thromboembolic events can underlie ischemic cerebral events and are associated with heart disease, valvular pathology, or carotid artery disease. Hypertension is a major risk factor for intraparenchymal hemorrhage and may lead to increased intracranial pressure, herniation, and death.

The greatest percentage of thrombi resulting in pulmonary embolism is thought to originate in the deep veins of the lower extremities. Deep-venous thrombosis can also occur in the pelvis or other locations. Fragments of blood clot may break off and embolize to the pulmonary arteries. An occlusion greater than 50 to 75% of the large pulmonary vessels results in a rise of the pulmonary artery pressure greater than 40 mmHg. This rise of pulmonary arterial pressure is accompanied by an increase in right ventricular diastolic, right atrial, and systemic venous pressures, with a decrease in cardiac output resulting in sudden death. Patients who have multiple small pulmonary emboli or in situ thrombus formation over time may present with increasing shortness of breath and right-sided heart failure. Because the lungs have dual circulation, infarctions are less common unless there is significant underlying natural disease with decreased cardiac function.

Various types of vasculitis or blood vessel inflammation can cause wall thickening, thrombosis, dissection, and rupture. Mesenteric thrombosis may be associated with polyarteritis nodosum and other autoimmune conditions.

Other Causes of Sudden Death

Rare undiagnosed brain tumors may present with sudden death. Infiltration or edema formation into the key respiratory/cardiac centers of the brain with possible herniation are two mechanisms. Early or late stage malignancies may sometimes metastasize to the heart and interfere with the conducting system, causing a fatal arrhythmia. Other causes of sudden death in patients with malignancies include cardiovascular events such as acute myocardial infarction, therapeutic complications (i.e., anaphylaxis), and metabolic derangements. Rare causes of sudden death in patients with tumors or malignancies include erosion of large vessels or visci with fatal hemorrhage. A colloid cyst of the third ventricle may lead to sudden death and is usually associated with premortem postural headaches. In certain positions, the cyst will act like a ball valve and suddenly block the flow of cerebral spinal fluid, resulting in acute obstructive hydrocephalus. One may be fine standing but develop symptoms when he or she lies down. This buildup of cerebral spinal fluid pressure can cause a fatal arrhythmia. Bacterial pneumonia with the combination of hypoxia and bacterial toxins and end products can cause sudden death.

Status asthmaticus and sudden asphyxic asthma are life-threatening forms of asthma. These cases are not unusual in a forensic setting. Status asthmaticus is defined as an acute attack of respiratory failure due to airway inflammation, edema, and mucous plugging. Sudden asphyxic asthma is due to brochospasm rather than airway inflammation. Viral infections and other causes have been implicated as precipitants of these potentially fatal complications. Grossly in both cases, the lungs may appear so much hyperaerated that at times rib indentations will show. Thick mucus plugs may obstruct the upper airways. Sudden death in asthmatic patients is thought to be secondary to fatal arrhythmia, occurring as a consequence of global hypoxia and right heart failure.

There is a condition known as sudden unexpected death in epilepsy (SUDEP). The mechanism is unclear but this phenomenon occurs in up to 18% of patients with epilepsy, presumably in those with subtherapeutic levels of anticonvulsants. Autonomic dysfunction has been proposed as a mechanism. Other mechanisms for death in patients with epilepsy include accidental/traumatic incidents such as drowning and choking that occur during a seizure. Hypoxia as a result of respiratory compromise can result in ischemic cardiac events. This may be part of the final mechanism of death in epileptic patients experiencing status epilepticus. Another interesting point to remember is that there is often very rapid rigor mortis formation in deaths directly following static epilepticus due to substantial adenosine triphosphate (ATP) depletion associated with prolonged muscle contractions from prolonged convulsions. Usually there are few pathologic findings that explain the sudden death in epileptic patients. Autopsy findings may include bite marks to the tongue with hemorrhage or a voided urinary bladder. There may be no finding at all. These are nonspecific findings and seizure activity may also occur prior to many other nonepilepsy-related deaths.

Fatal anaphylaxis can result from exposure to insect stings, foods, latex, drugs, chemicals, and exercise. This mast cell-mediated systemic reaction results in severe angioedema and bronchoconstriction of the upper respiratory tract along with hypotension resulting in respiratory and circulatory collapse. Death caused by anaphylaxis is primarily due to airway obstruction when laryngeal edema fills the rich lymphatic supply of the epiglottic folds. Increased mast cell tryptase levels in the patient's serum can be detected that peaks approximately 15 to 60 minutes after the onset of anaphylaxis and then declines with a half-life of about 2 hours.

The mortality for gastrointestinal bleeding (GI) in the case of ruptured esophageal varices most commonly encountered in patients with portal hypertension is high. Intra-aortic balloon pumps are lifesaving procedures but only if the patient presents in a timely fashion. Other causes of fatal upper gastrointestinal bleeding include stomach and duodenal ulcers; in this scenario the source is arterial as opposed to venous in esophageal varices. Fatal lower gastrointestinal bleeding can be seen in patients with angiodysplasia, diverticulitis, and carcinoma; however, this scenario is less common than upper GI bleeding.

Mostly complications of morbid obesity are thought to underlie the association with sudden death. Hypertension, left ventricular hypertrophy, and cardiomegaly are all independent risk factors for sudden death. Postural asphyxia may occur as a result of obesity. Morbid obesity is a reasonable cause of death by itself due to stress on the heart. An individual who is three times the expected body weight has roughly three times the vasculature with three times the blood volume to pump. In times of other stress, this can have devastating consequences on the heart, with death by arrhythmia.

Waterhouse–Friderichsen syndrome was first described as occurring in patients with meningococcemia and is characterized by severe bacteremia and bilateral adrenal hemorrhages. This combination results in overwhelming shock and, if untreated, sudden death can occur. Organisms other than *N. meningitis,* such as *E.Coli,* have been reported to produce this syndrome.

Multiorgan failure and death can be seen in sickle cell anemia patients with an acute crisis. Precipitants may include infection, dehydration, hypoxia, physical excretion, vaso-occlusion, or fat embolus following bone infarction. This acute hemolytic sickling crisis results in severe hypoxemia with end organ failure. Patients with sickle cell anemia have auto-infracted spleens and are much more susceptible to encapsulated organisms such as pneumococcal bacteria. Even patients with sickle cell trait may develop crisis in times of great physical exertion with dehydration, such as basic training in the army or boot camp.

Natural disease processes may weaken the body, making fatal traumatic injury more likely. Osteoporosis from aging, Cushing syndrome, steroid use, and other natural disease processes will make bones more fragile and allow fractures to occur more easily.

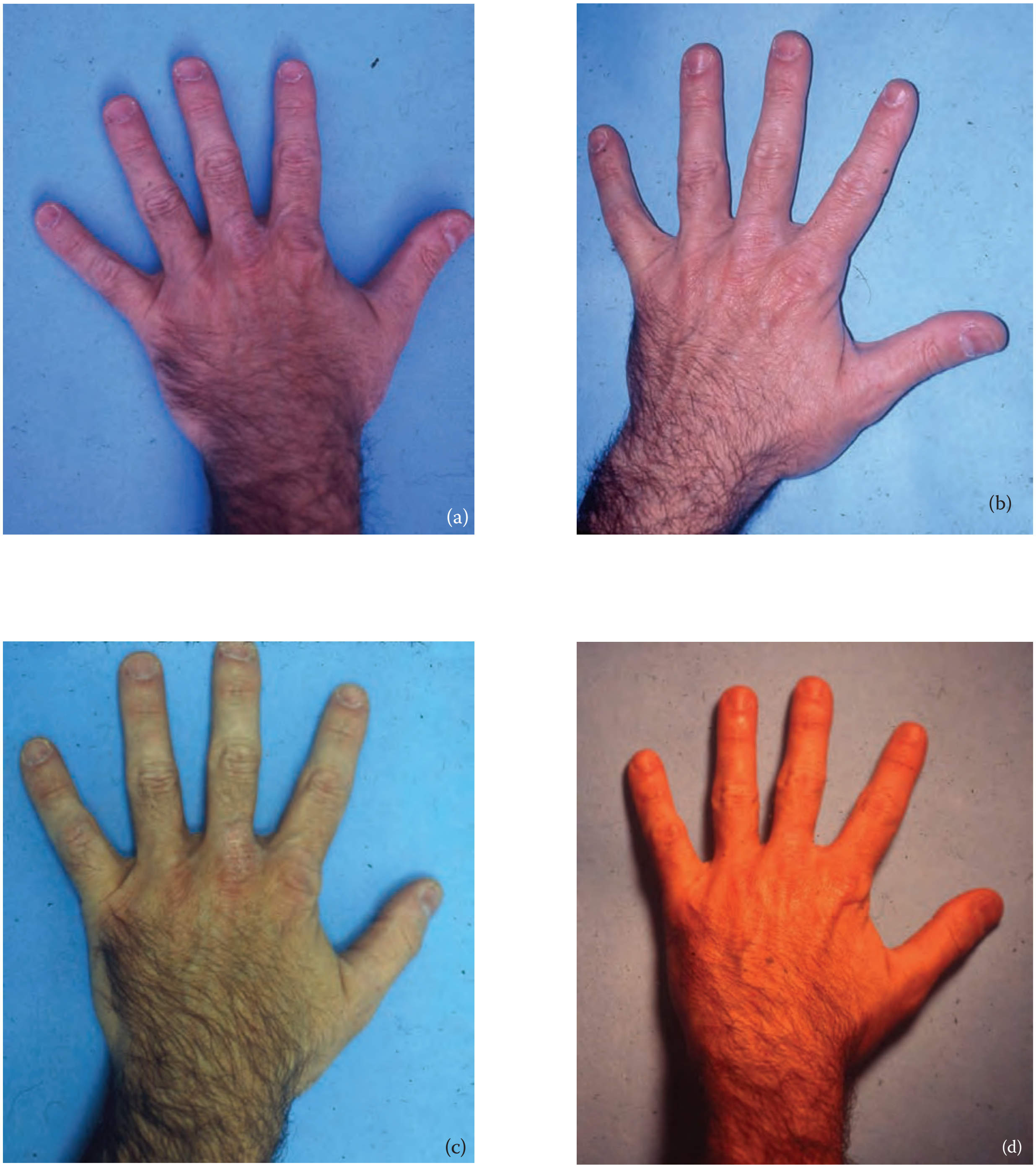

When examining photographic evidence it may be important to have knowledge of the type of camera, film, and lighting used when documenting different disease states. These four photos demonstrate different types of lighting causing variation in picture color. Figure (a) was taken in overcast sunlight, (b) with camera flash, (c) under fluorescent light, and (d) with a Tungsten filament regular light bulb. If your opinion is that a photographic image is not interpretable, it is perfectly acceptable to say that you cannot render an opinion based on this two-dimensional image.

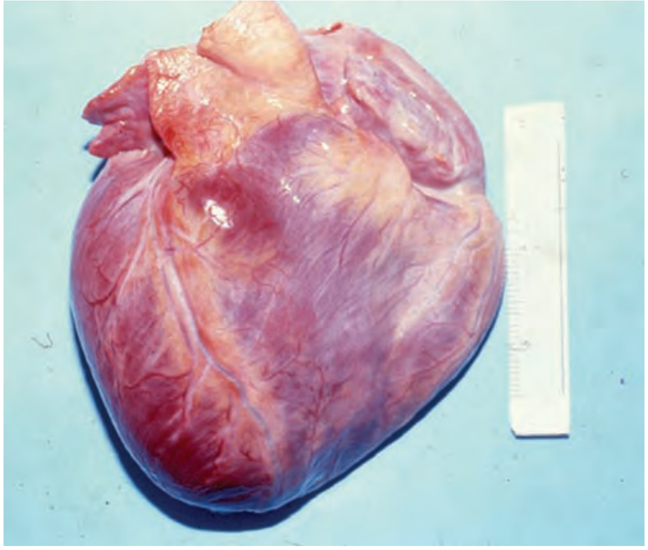

Normal fresh heart.

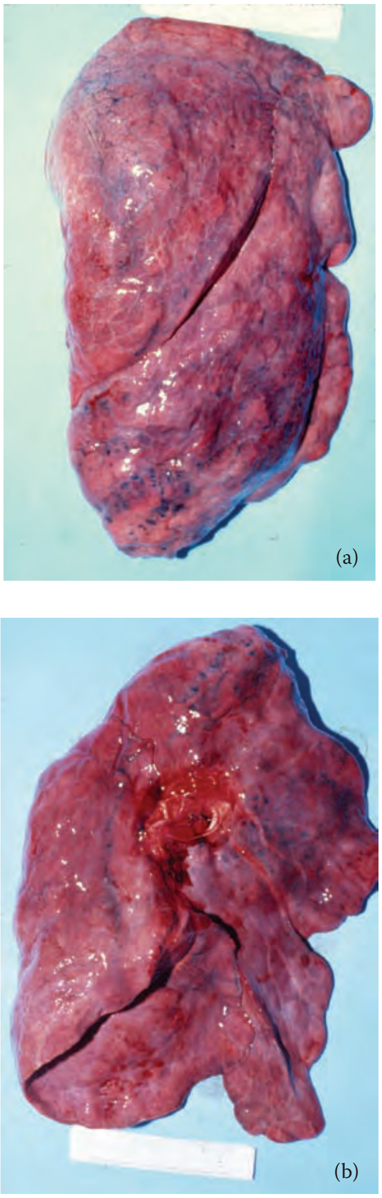

Normal fresh left lung (a) demonstrating two lobes. Right lungs (b) have three lobes.

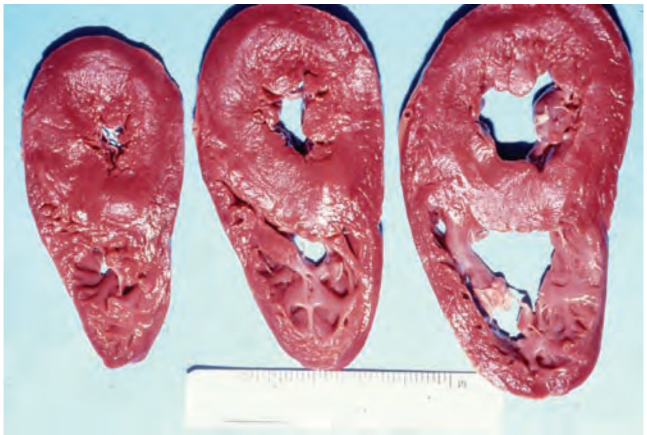

Sections of a normal fresh heart showing right and left ventricle.

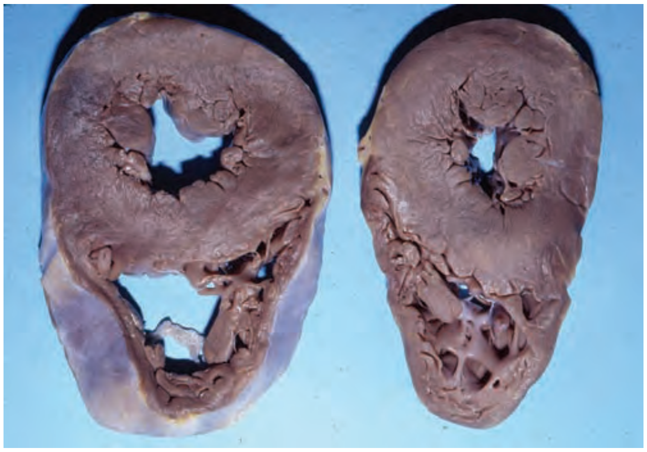

Sections of normal right and left ventricle after formaldehyde fixation.

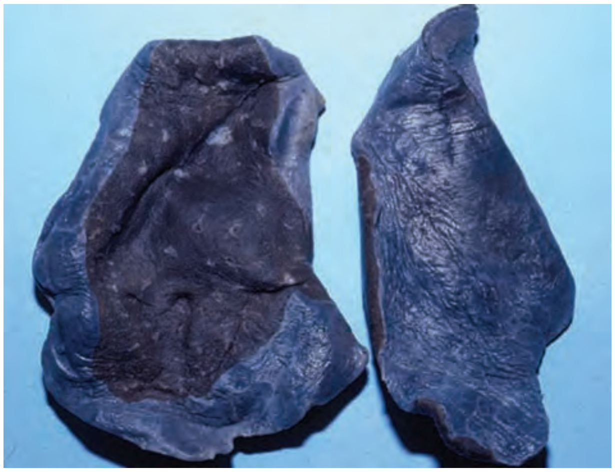

Normal lung fixed in formaldehyde.

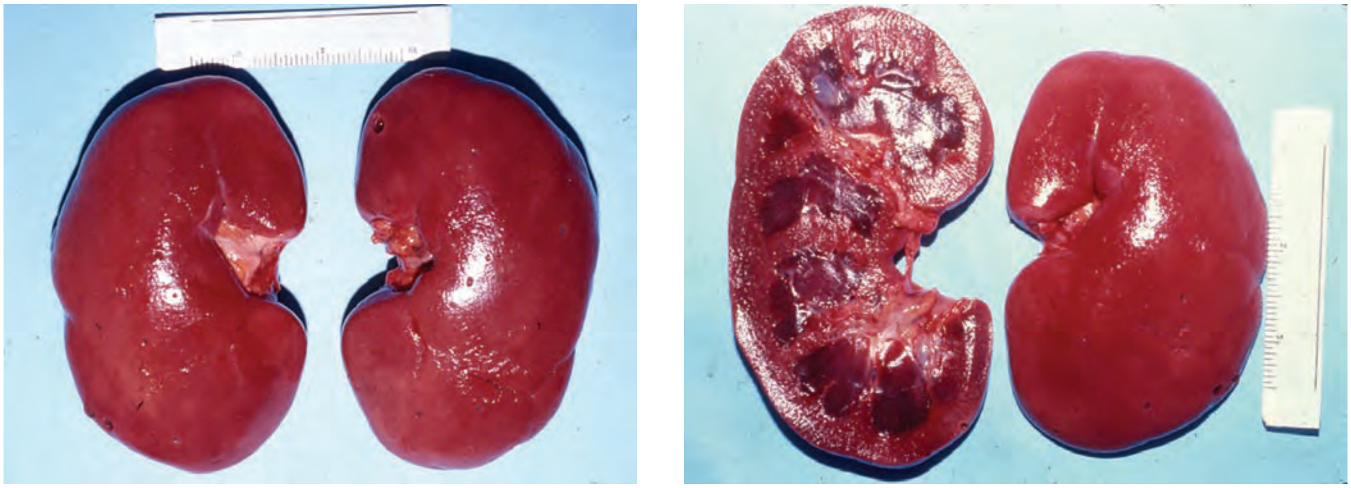

Normal fresh kidneys.

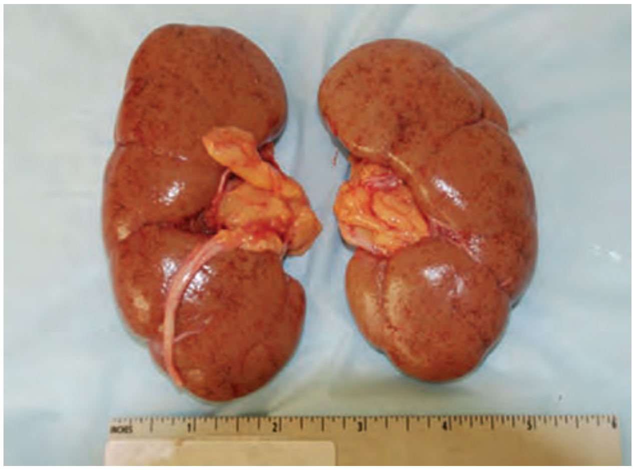

Non-fixed fresh kidneys. Note the pale discoloration resulting from fatal blood loss prior to death due to a gunshot wound.

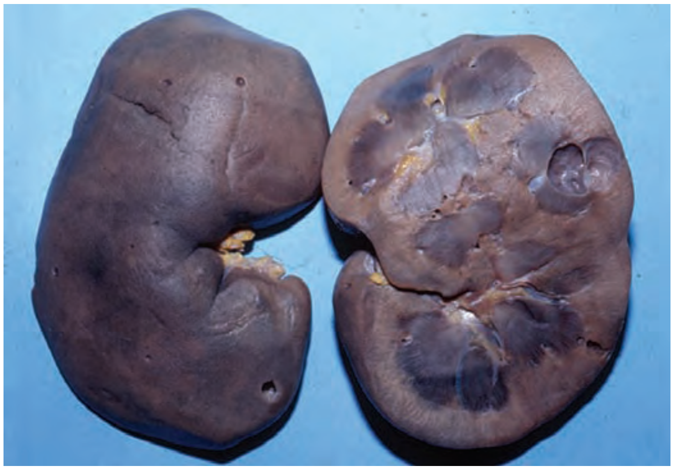

Normal kidney fixed in formaldehyde.

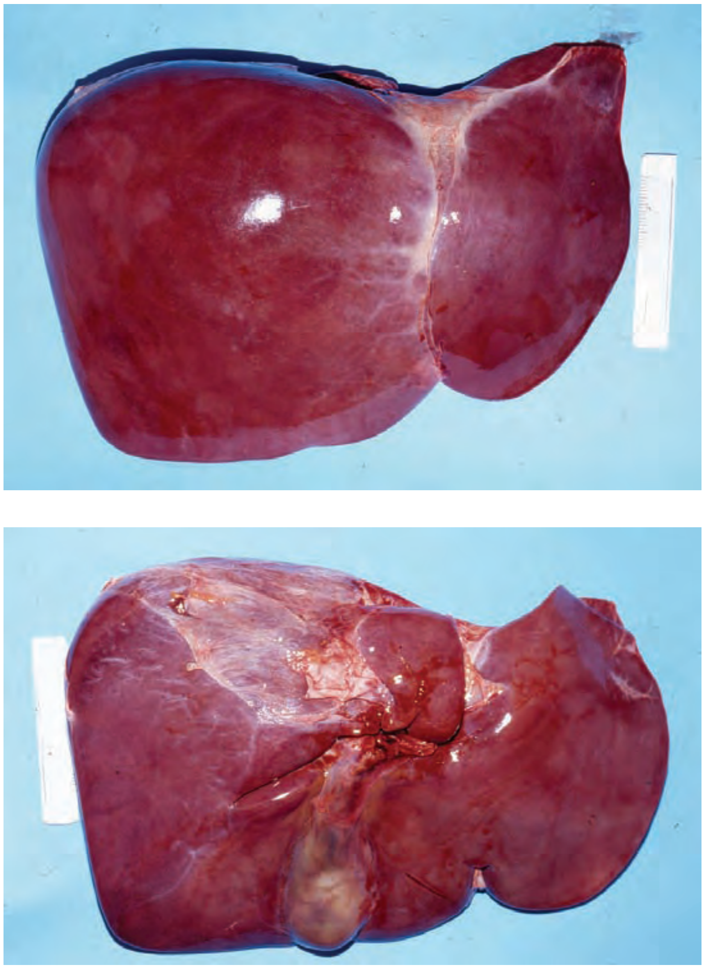

Normal fresh liver.

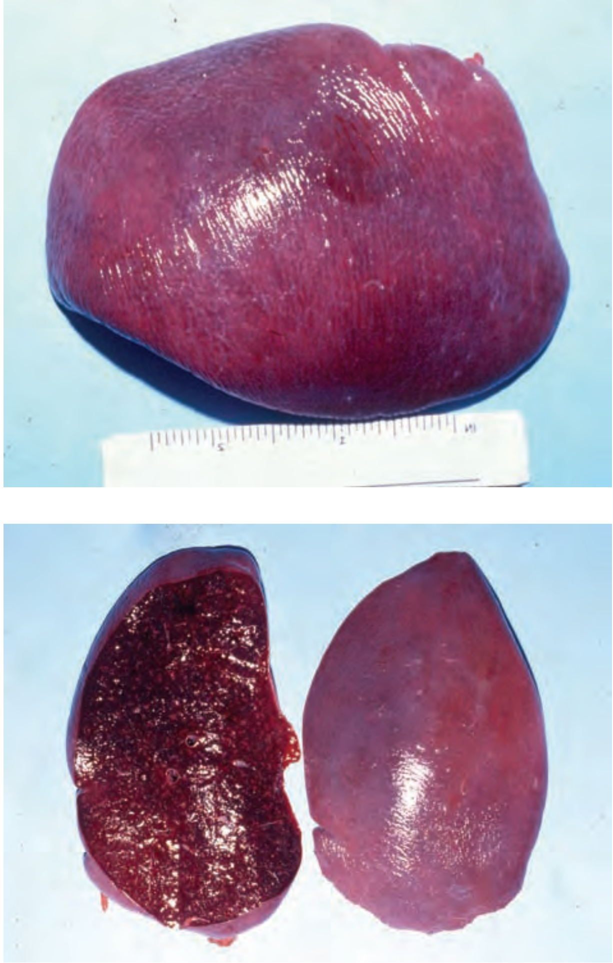

Normal fresh spleen.

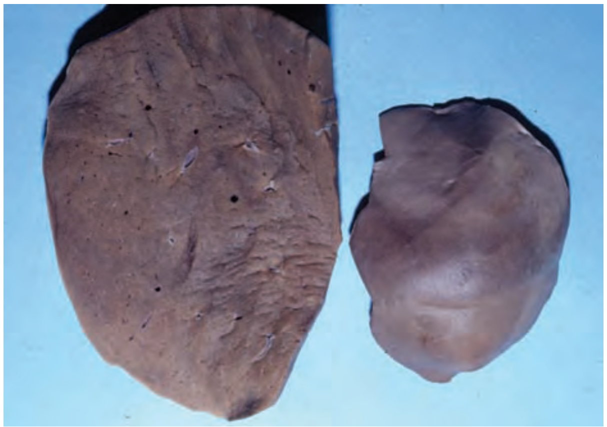

Normal liver fixed in formaldehyde.

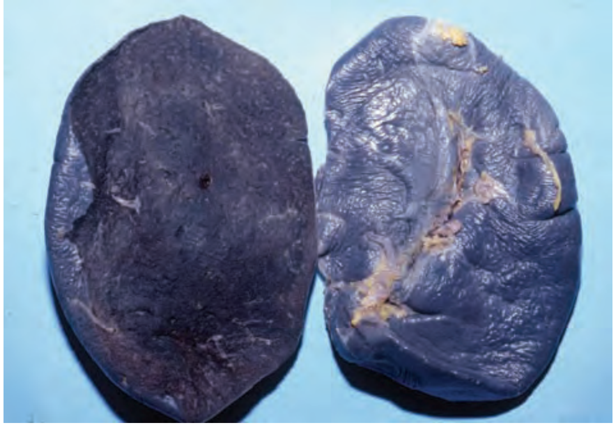

Normal spleen fixed in formaldehyde.

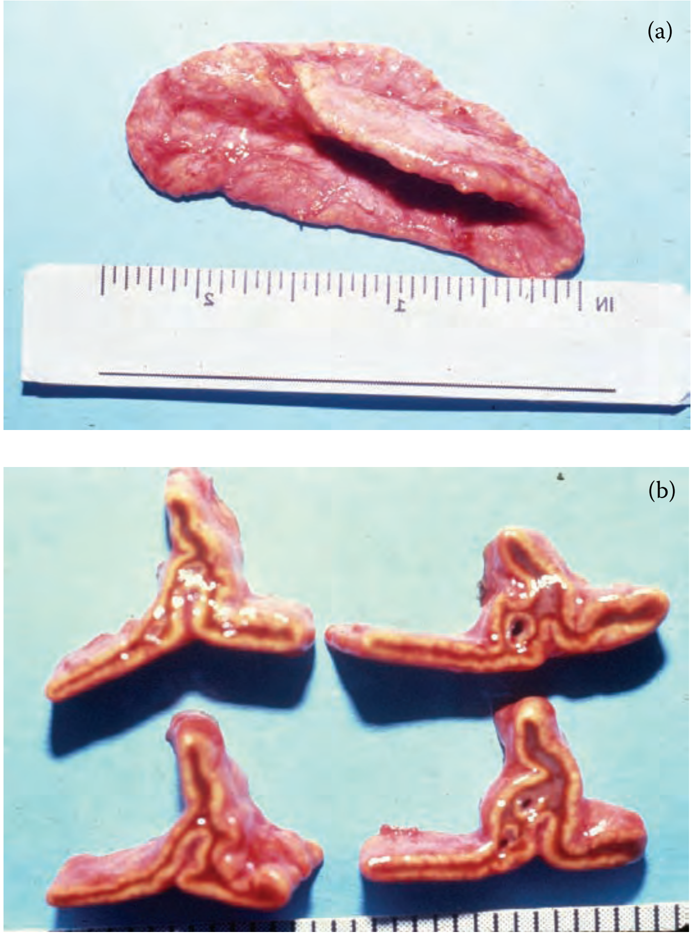

Normal fresh adrenal gland intact (a) and sectioned (b).

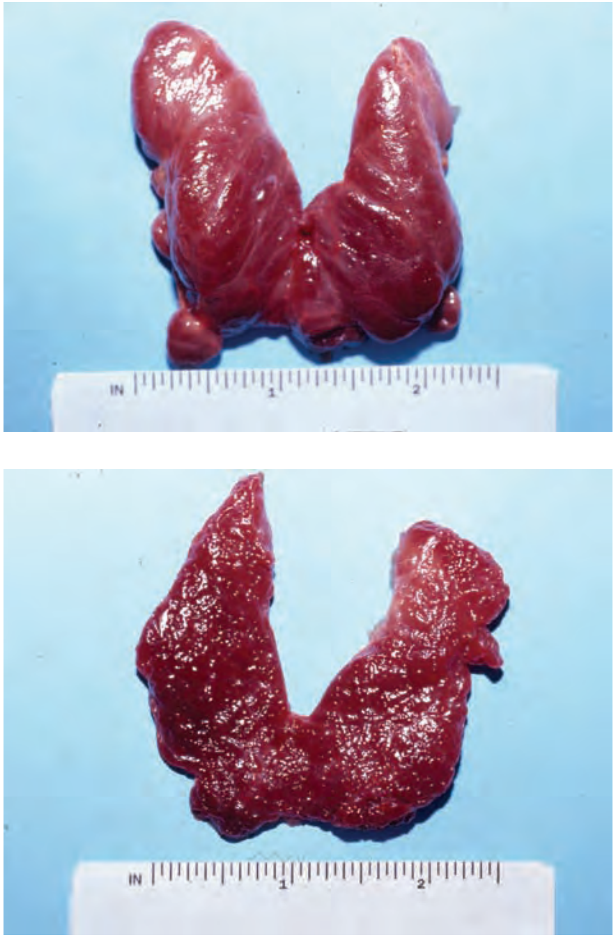

Normal fresh thyroid gland.

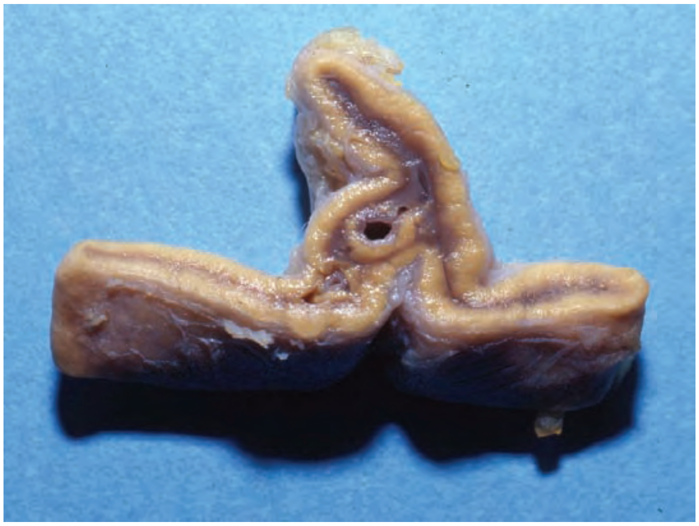

Normal adrenal gland section fixed in formaldehyde.

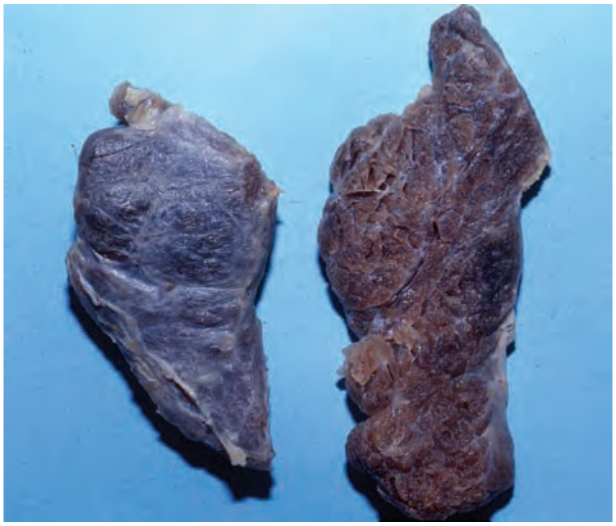

Normal thyroid fixed in formaldehyde.

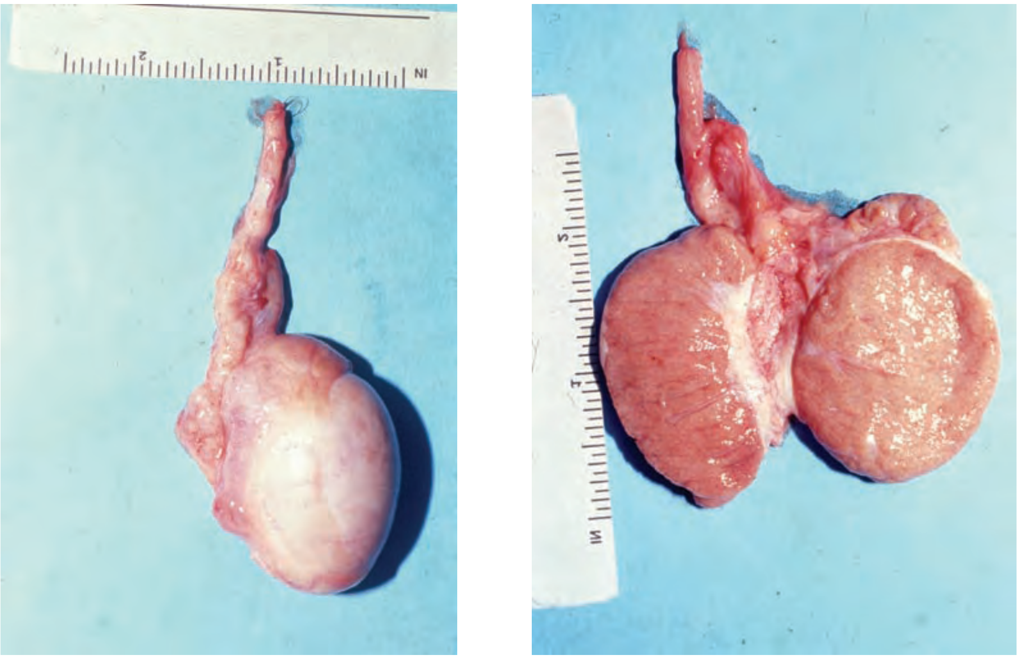

Normal fresh testes.

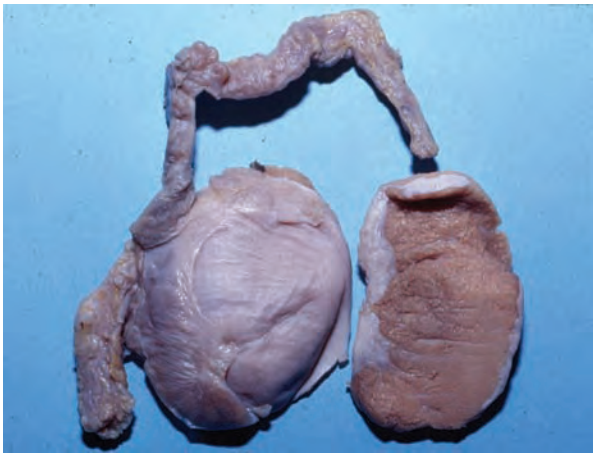

Normal testes fixed in formaldehyde.

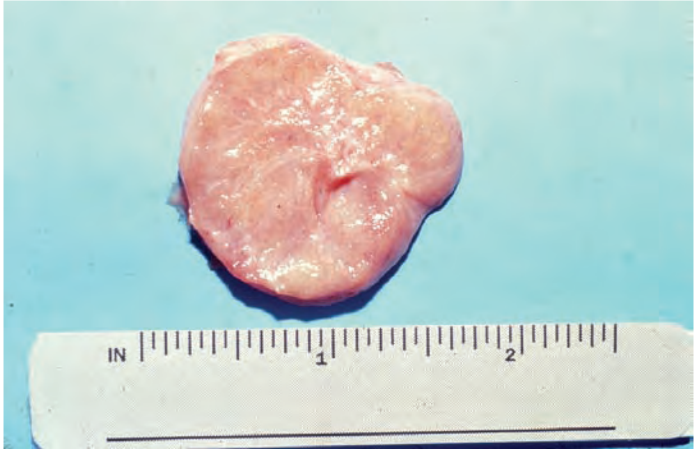

Normal fresh prostate.

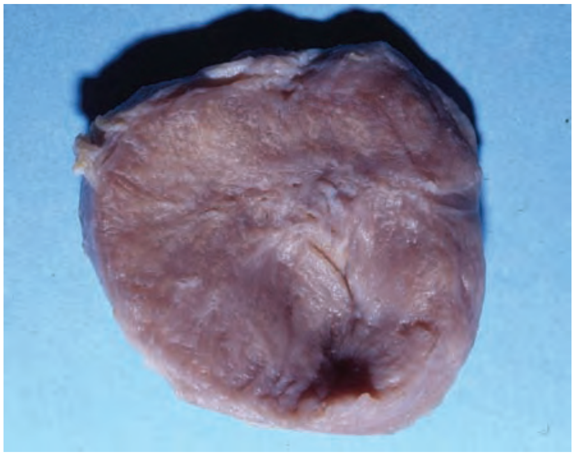

Normal prostate fixed in formaldehyde.

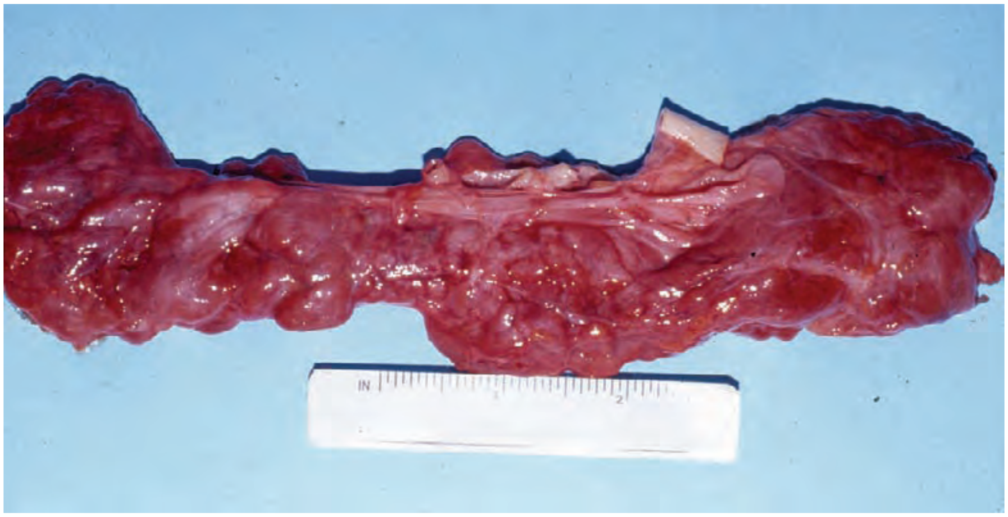

Normal fresh pancreas.

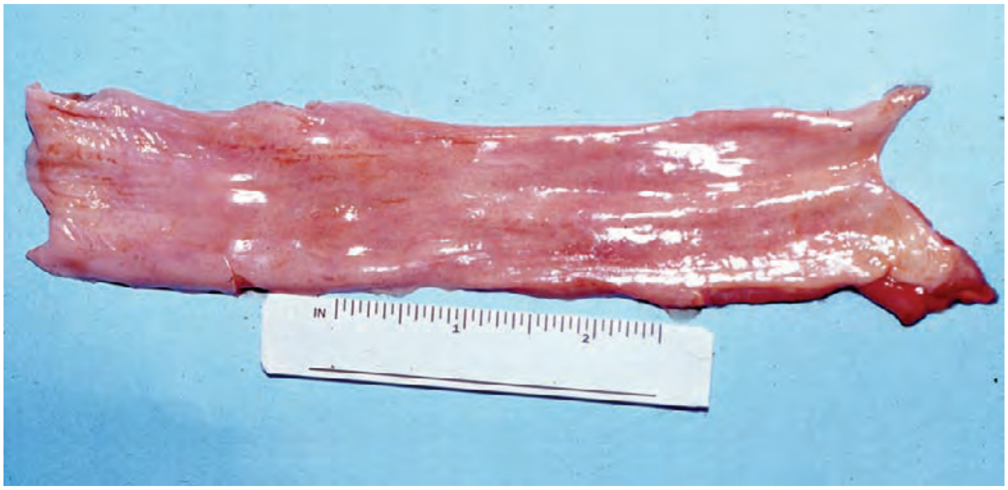

Normal fresh esophagus.

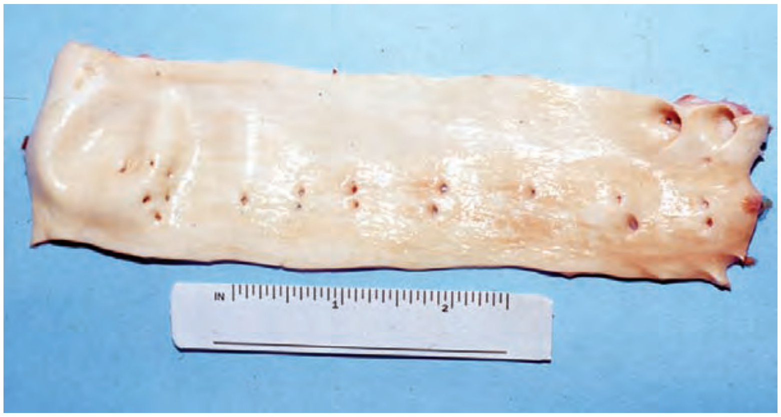

Aorta with slight atherosclerosis. Note the fatty streaks on the intimal surface.

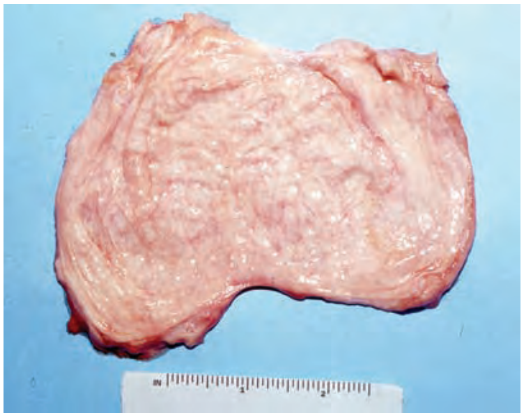

Normal fresh bladder.

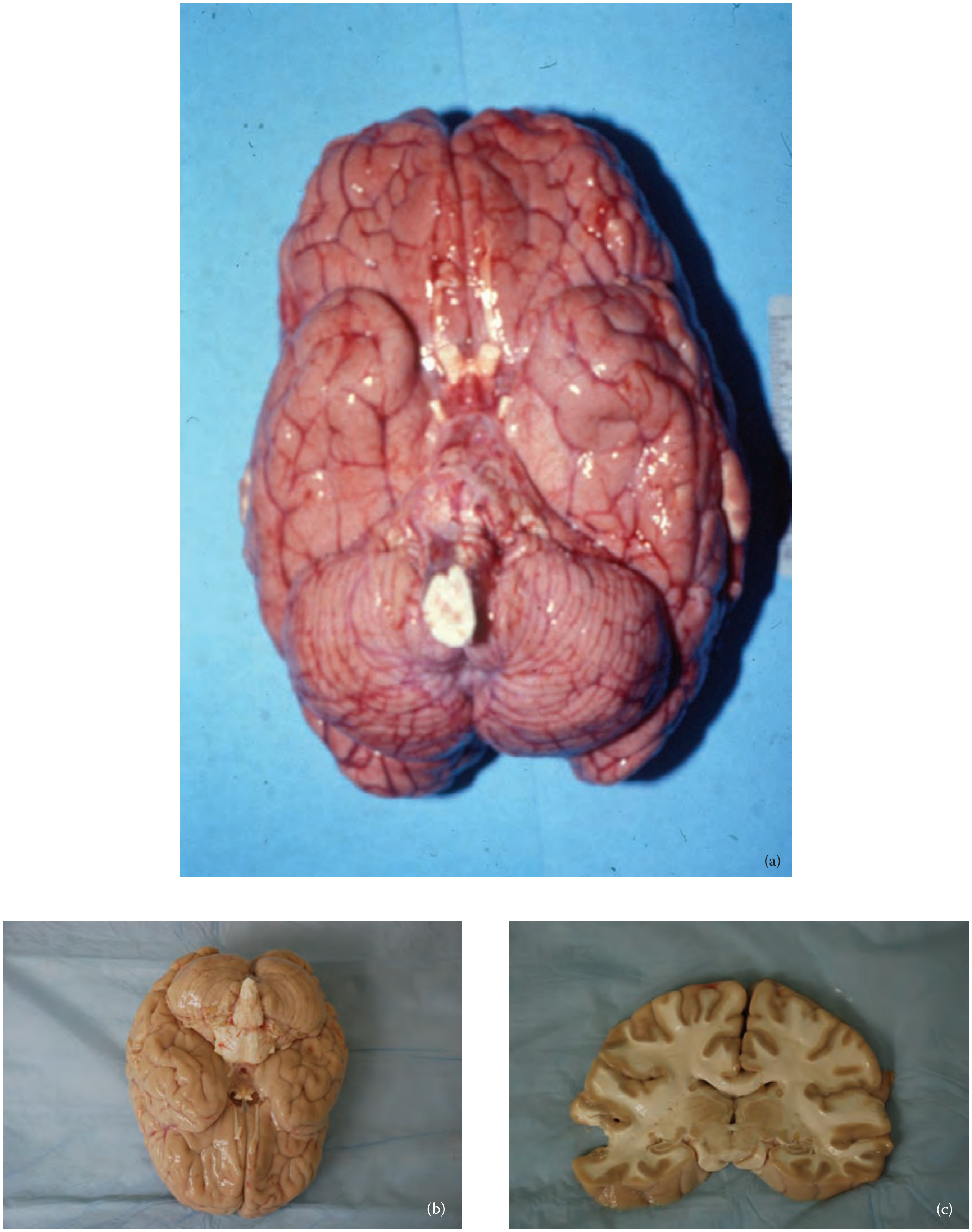

Figure (a) demonstrates congestion of fresh brain. Note the slight pink color. Figures (b) and (c) demonstrate fresh brain in an individual who exsanguinated from a ruptured aortic aneurysm. Note the pale discoloration due to blood loss.

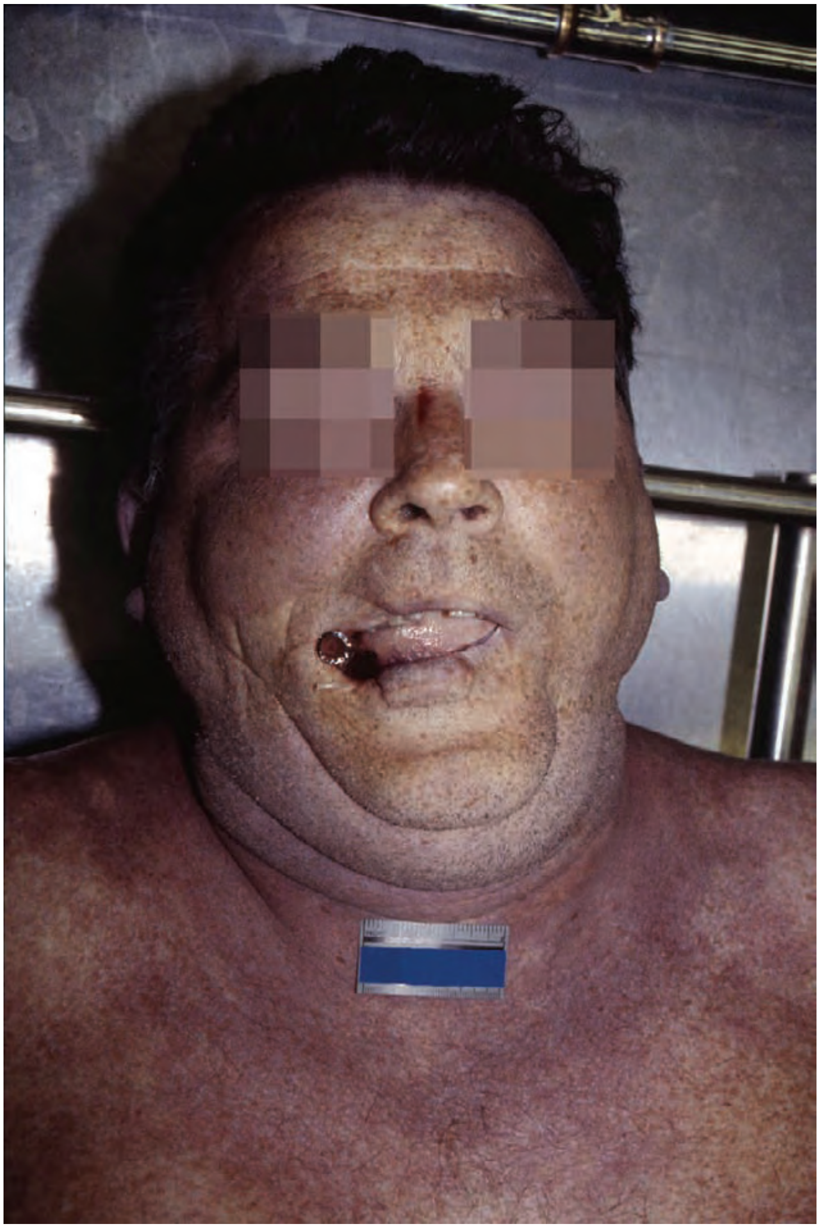

This figure depicts a decedent with an endotracheal tube in the right side of his mouth. It also demonstrates the "purple head sign," a common finding in victims of sudden death, particularly cardiac death. The explanation for this finding is not known in entirety but is attributed to uncontrolled terminal sympathetic nervous system discharges, which open free capillary sphincters and produce a gush of capillary blood.

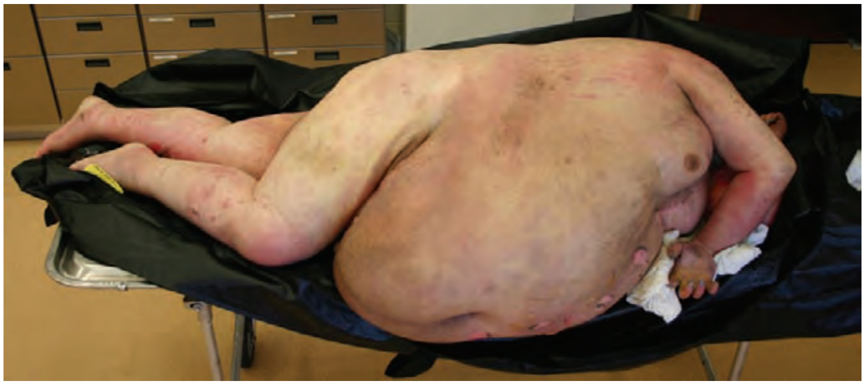

Morbid obesity. This is a legitimate cause of death and can stand alone on a death certificate.

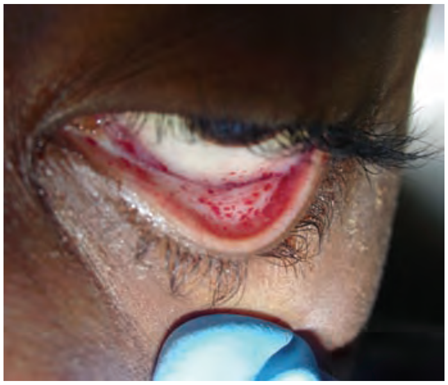

Petechiae associated with heart disease and resuscitation.

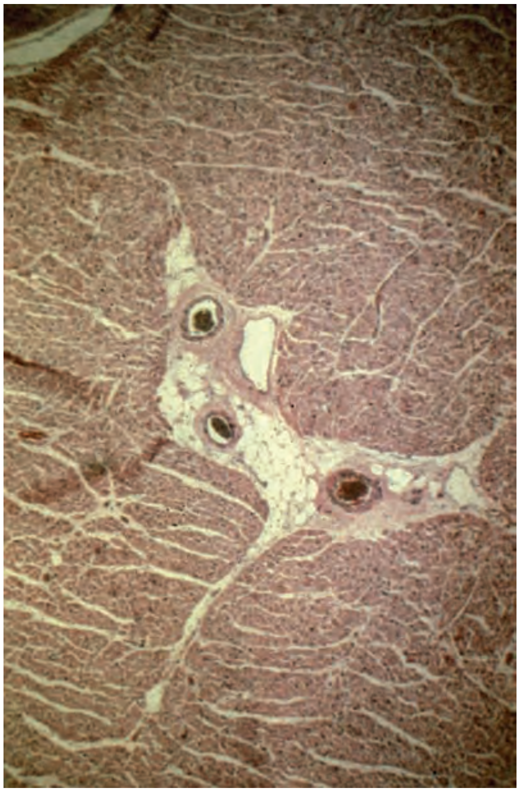

Low-power magnified view of small coronary arteries with thrombosis.

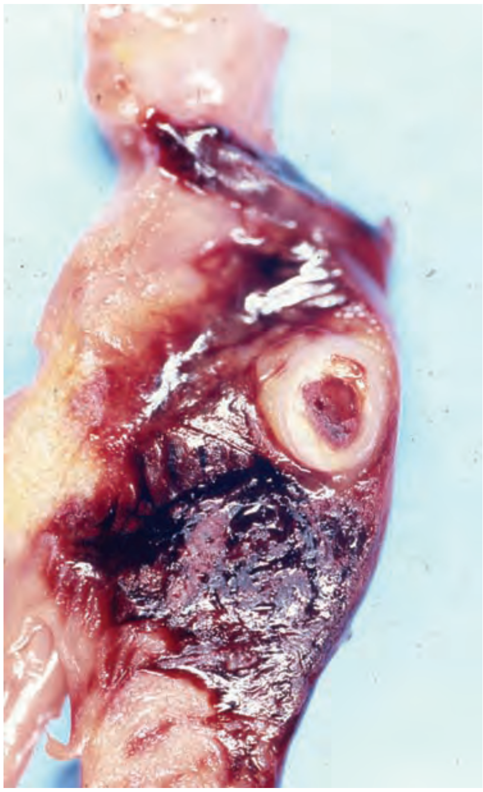

Epicardial vessel with complete occlusion by organizing thrombus. Note the adjacent epicardial hemorrhage.

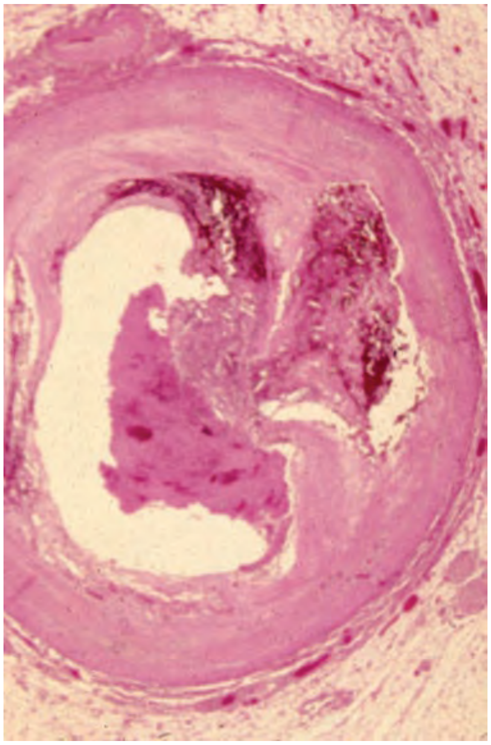

Coronary artery with a ruptured atherosclerotic plaque and thrombus formation.

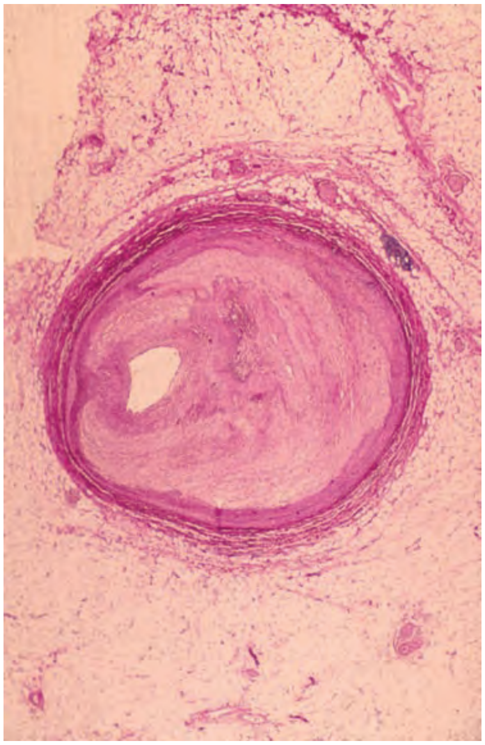

Marked coronary atherosclerosis in an epicardial vessel.

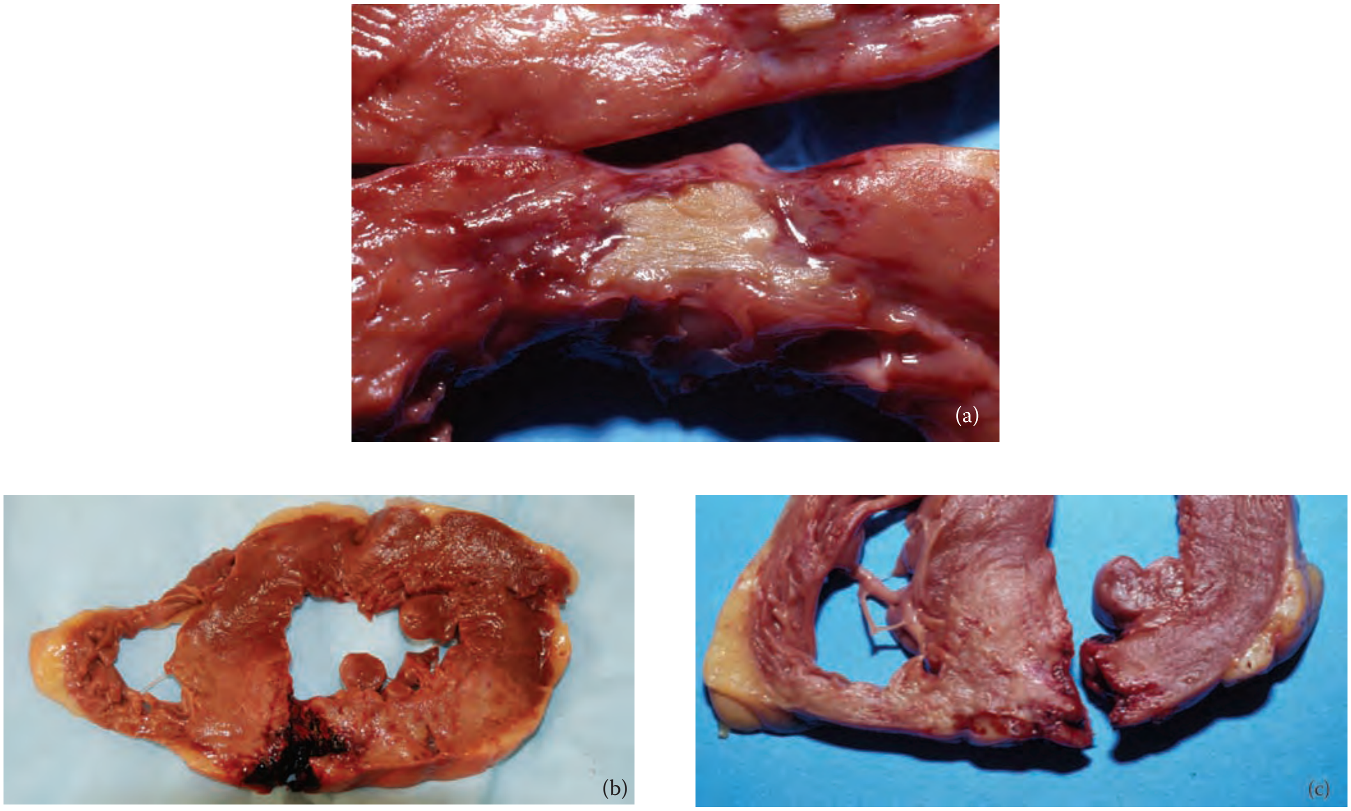

An acute myocardial infarction. Note the yellow discoloration due to necrosis. This infarction is approximately 3 to 6 days old. In Figure (b) and (c), there is a transmural acute myocardial infarction with rupture.

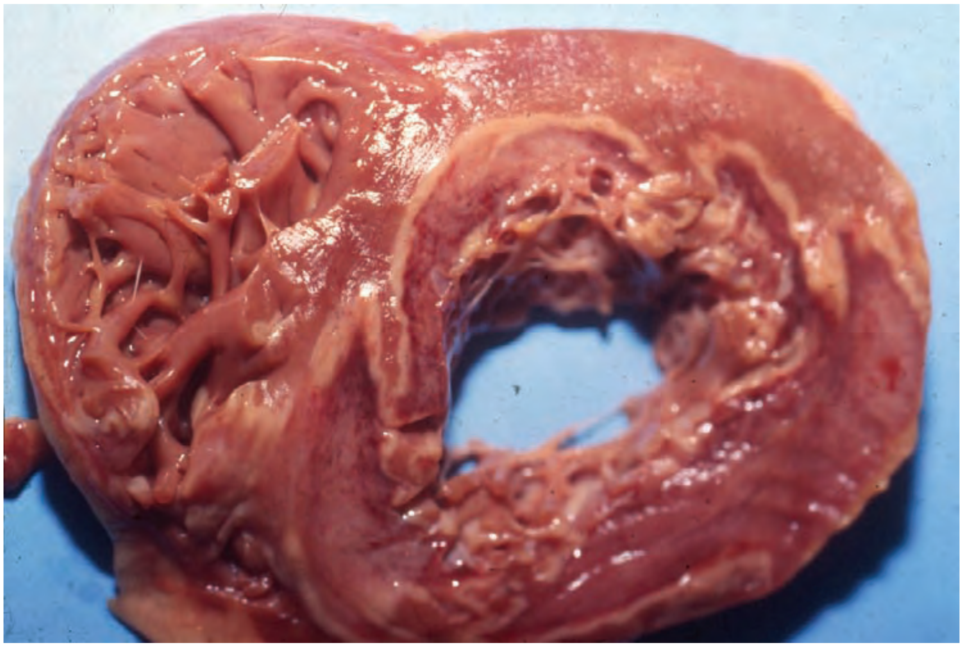

Acute myocardial infarction.

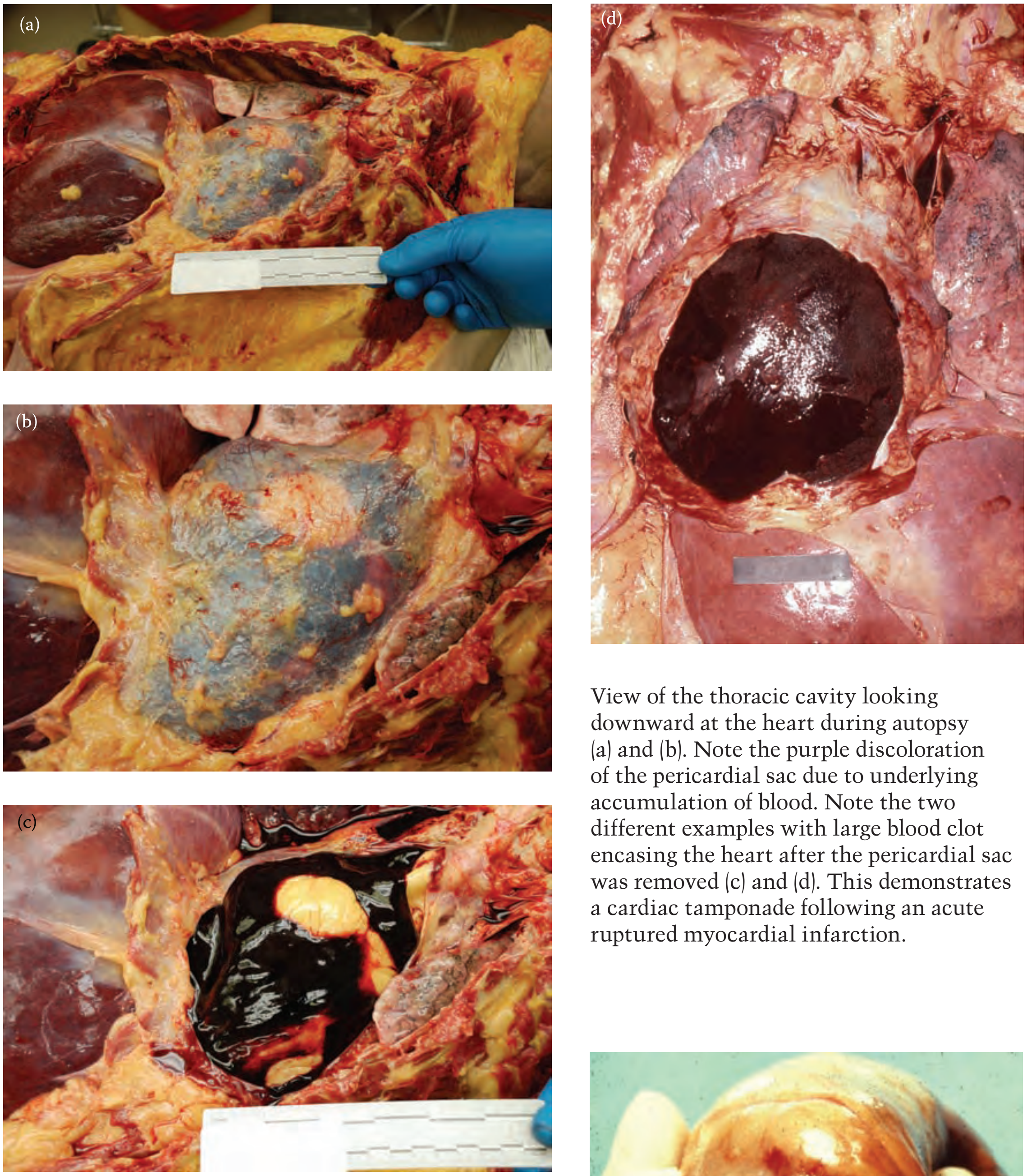

View of the thoracic cavity looking downward at the heart during autopsy (a) and (b). Note the purple discoloration of the pericardial sac due to underlying accumulation of blood. Note the two different examples with large blood clot encasing the heart after the pericardial sac was removed (c) and (d). This demonstrates a cardiac tamponade following an acute ruptured myocardial infarction.

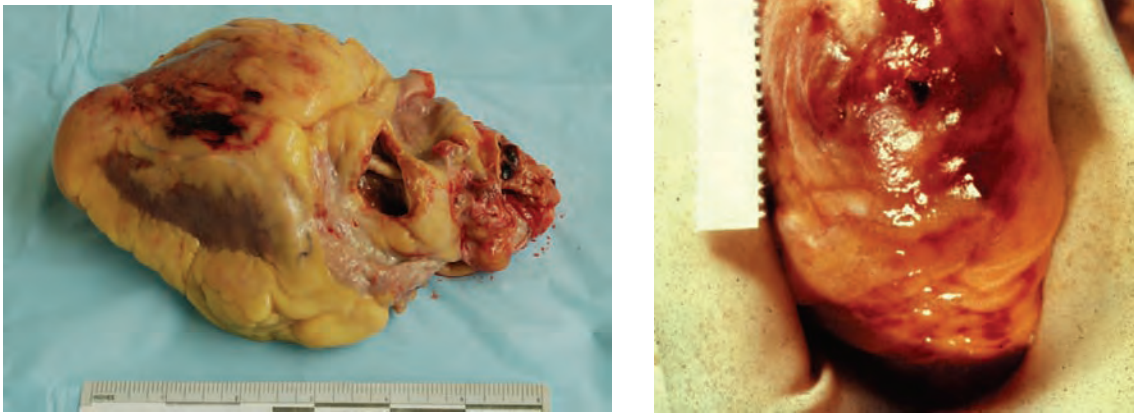

Two examples of hearts demonstrating acute ruptured myocardial infarction. Note the adjacent hemorrhage and perforation site. This resulted in cardiac tamponade and sudden death.

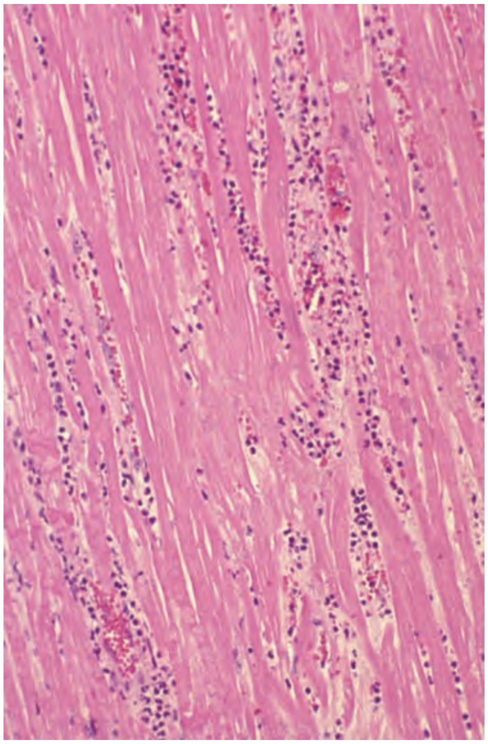

This microscopic section of heart muscle reveals an acute myocardial infarction with hemorrhage, polymorphonuclear cell infiltrates, and myocardial necrosis.

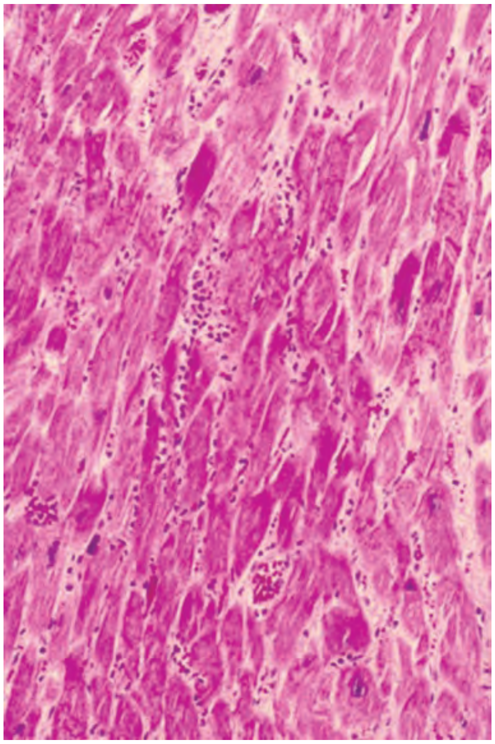

This acute myocardial infarction demonstrates coagulative necrosis with contraction bands. There are also polymorphonuclear cell infiltrates.

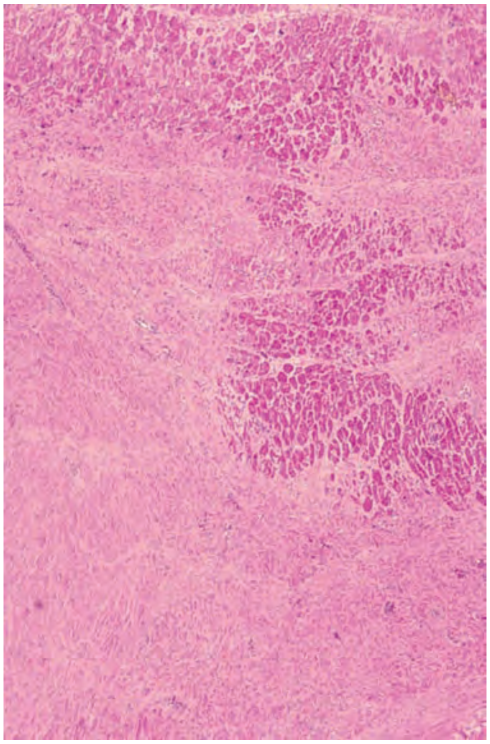

This histopathologic section of myocardium reveals an old infarction with fibrosis.

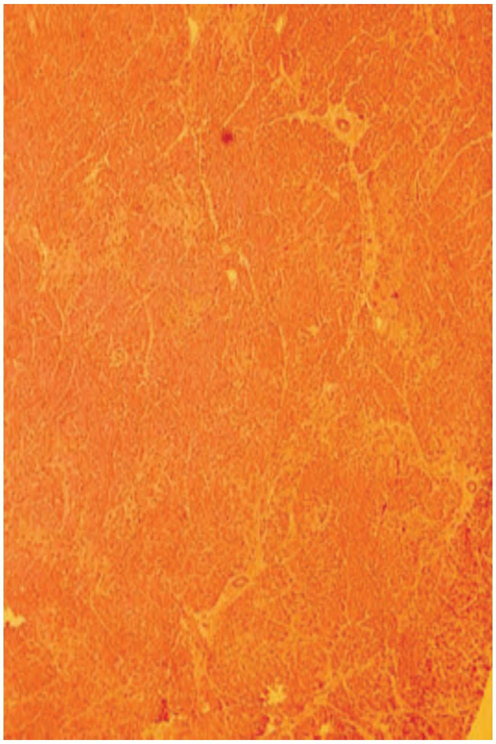

This low-power view of heart muscle reveals patchy areas of fibrosis consisting of healed hypoperfusion infarctions secondary to a remote trauma with severe shock. Also note the perivascular fibrosis.

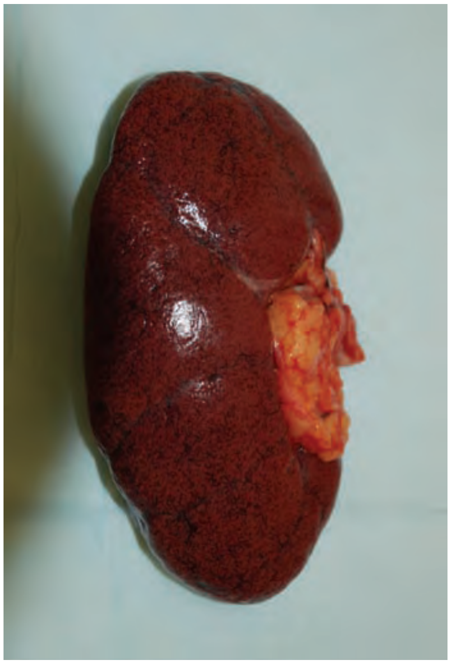

Early to moderate nephro-arteriolosclerosis.

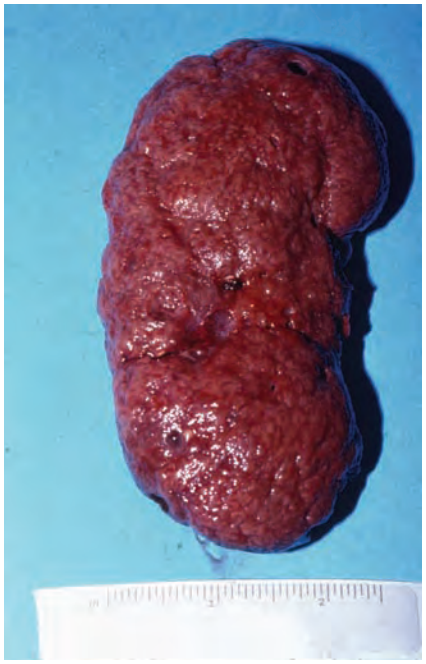

Moderate to marked nephro-arteriolosclerosis.

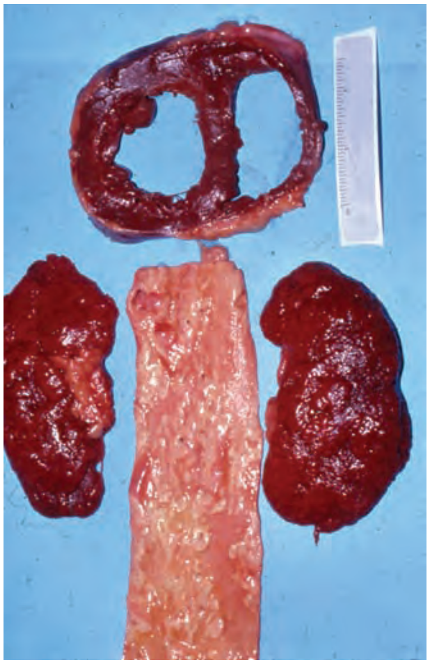

Nephro-arteriolosclerosis associated with hypertensive cardiovascular disease. Note the cardiac hypertrophy and biventricular dilatation in this failing heart. There is also moderate atherosclerosis of the aorta. Note the markedly granular subcapsular kidney surfaces and cortical scarring also associated with this process.

Arteriovenous hemodialysis grafts for treatment of chronic renal failure due to hypertensive cardiovascular disease.

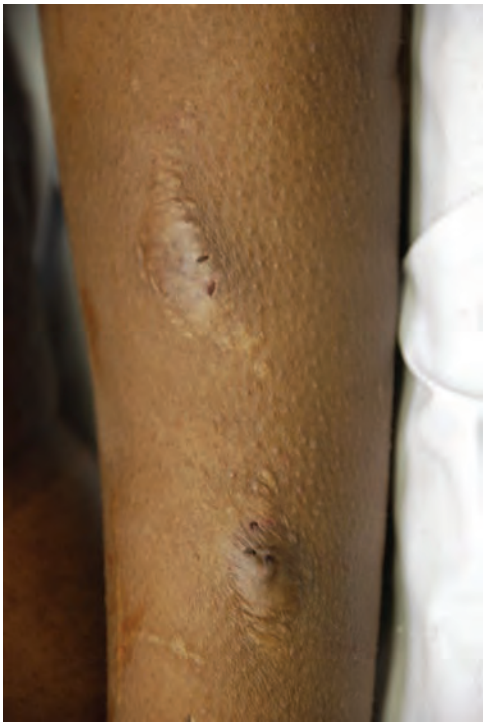

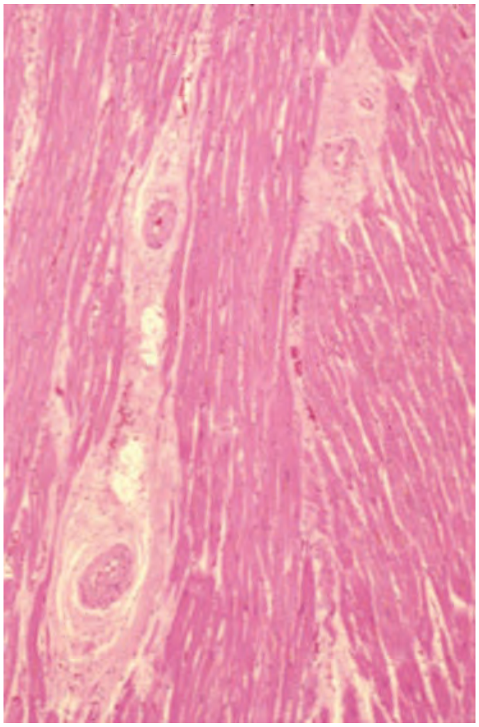

This microscopic image of myocardium reveals significant perivascular fibrosis and small vessel disease in a patient who had a long history of hypertensive cardiovascular disease and diabetes mellitus.

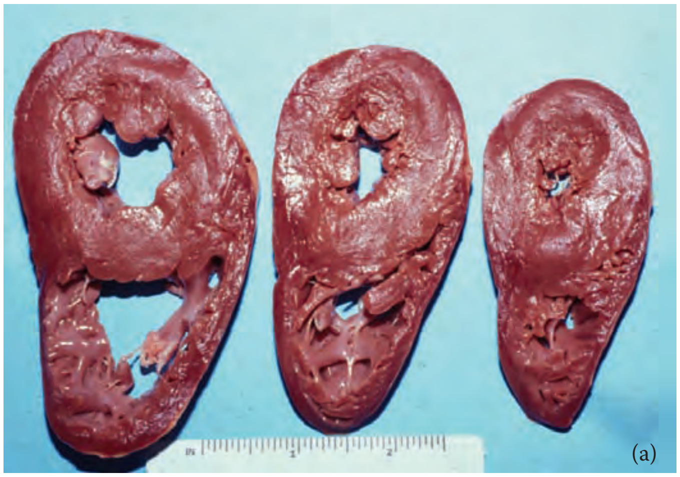

Normal cross sections of heart (a). Compare this image to the ones below. Figure (b) shows a markedly enlarged heart due to hypertensive cardiovascular disease. This heart is diffusely enlarged and shows cardiac hypertrophy. Figure (c) is a markedly hypertrophied heart with extreme concentric left ventricle hypertrophy. This individual had severe hypertension.

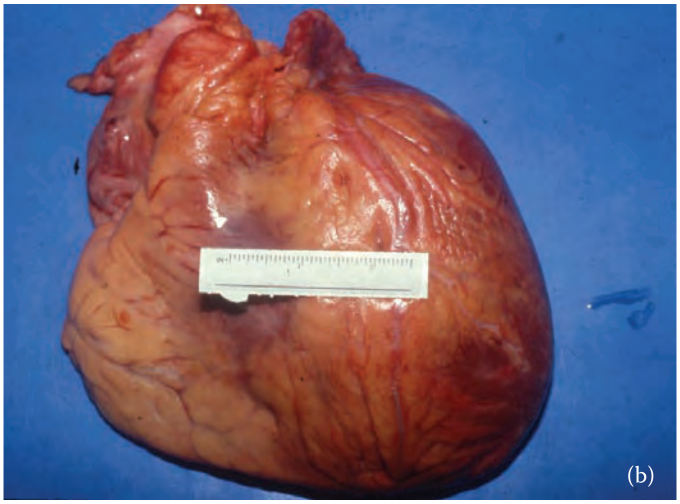

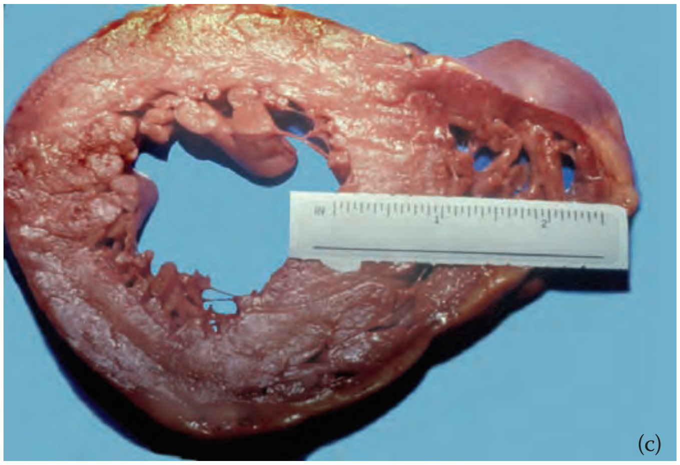

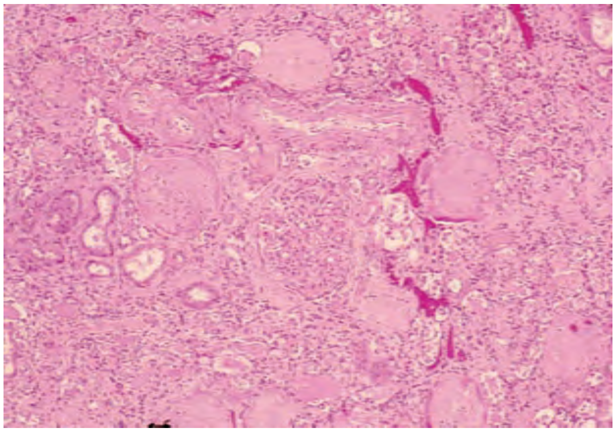

This is a microscopic section of kidney showing fibrosis, arteriolosclerosis, and nephrosclerosis due to hypertensive cardiovascular disease.

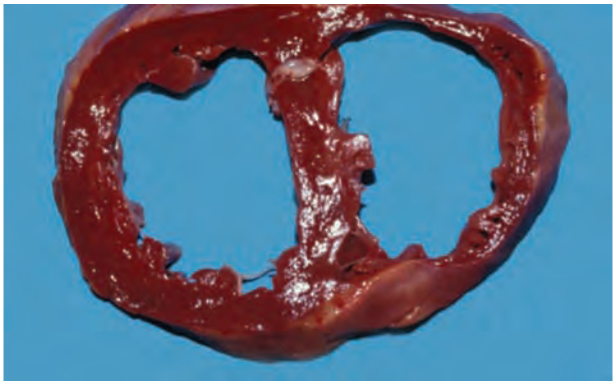

The right and left ventricles show marked dilatation. This individual died of a peripartum cardiomyopathy. A dilated cardiomyopathy or end stage hypertensive cardiovascular disease with cardiac failure will appear the same grossly.

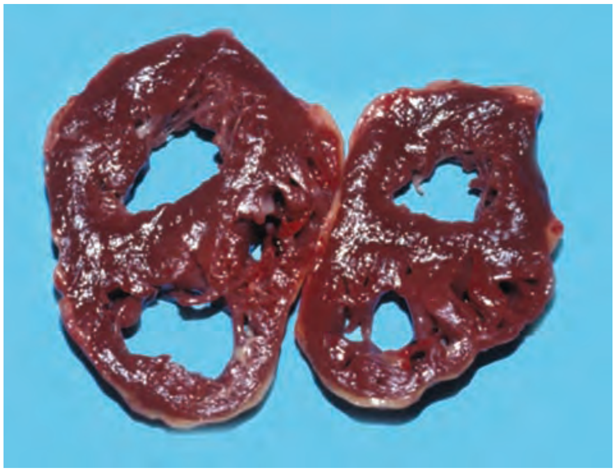

The sections of these ventricles reveal marked right ventricle hypertrophy. This individual had end-stage primary pulmonary fibrosis with cor pulmonale and cardiac failure.

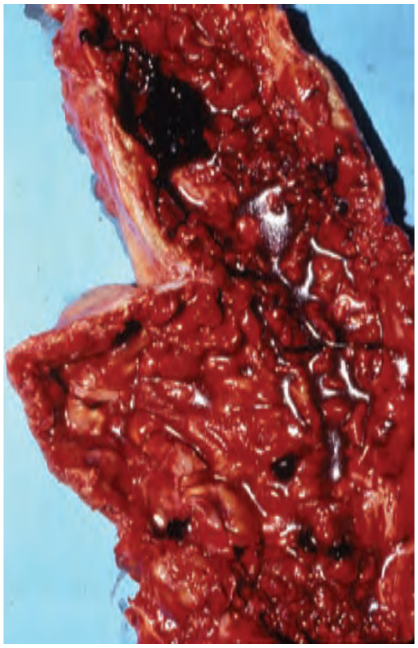

The intimal lining of an aorta with marked atherosclerosis in a decedent with a longstanding history of smoking, diabetes, high cholesterol, and high blood pressure.

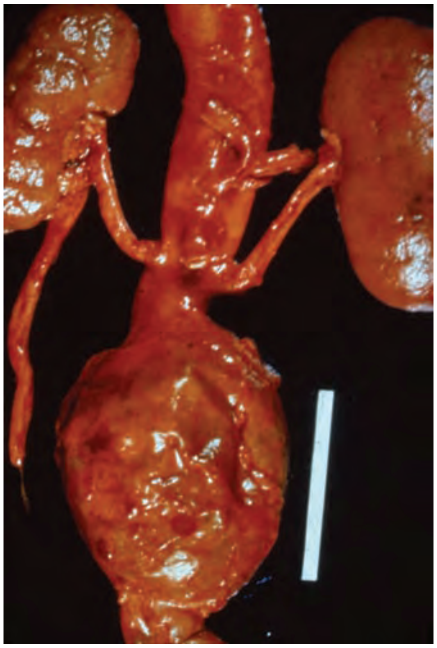

Abdominal aortic atherosclerotic aneurysm shown in its typical location inferior to the renal arteries and above the iliac bifurcation.

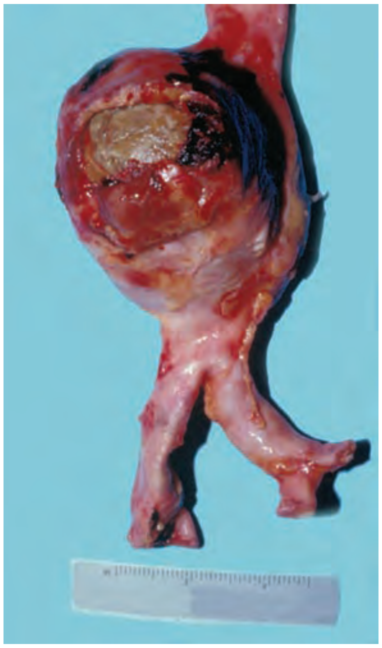

An abdominal atherosclerotic aneurysm with a rupture at its anterior aspect and visible thrombosis.

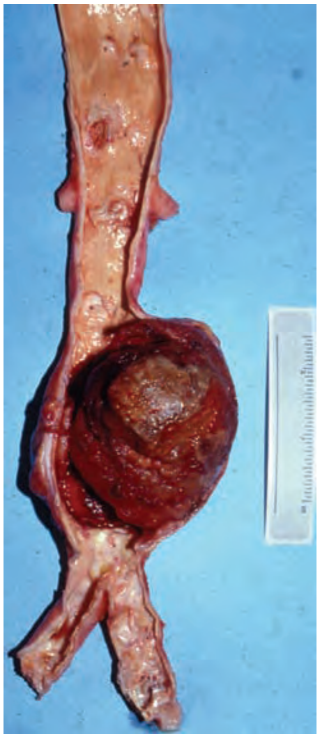

This abdominal aortic aneurysm that has been opened to remove half of the vessel wall and show the underlying intimal surface with moderate atherosclerosis except for the region of the aneurysm, which has marked atherosclerosis and a large overlying thrombus.

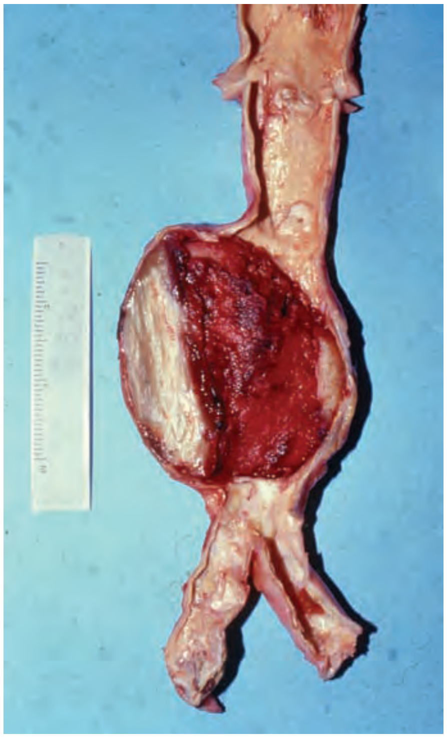

This abdominal aortic aneurysm has been cross sectioned to show the partial obstruction of the aneurysm by organizing thrombosis. The lumen is demonstrated by fresh red blood clot at its surface and the thrombus is demonstrated by the light grey regions adjacent to the right and left wall.

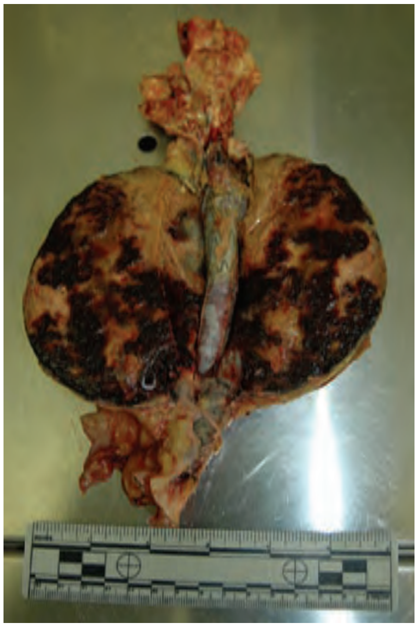

Abdominal aortic aneurysm with vascular graft repair.

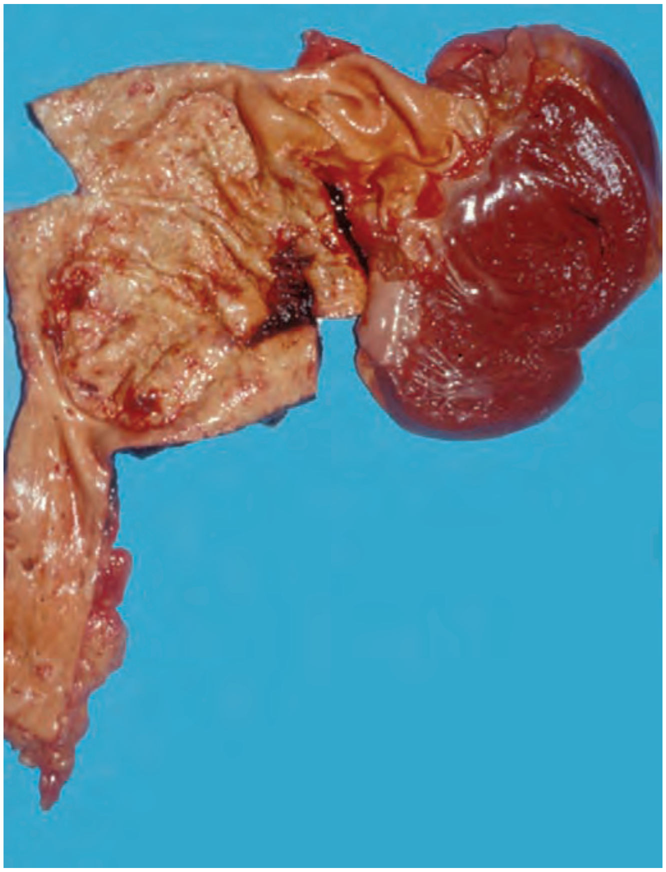

Thoracic aortic atherosclerotic aneurysm. This is not the typical location for such an aneurysm. Also note the aneurysm begins distal to the root of the aorta.

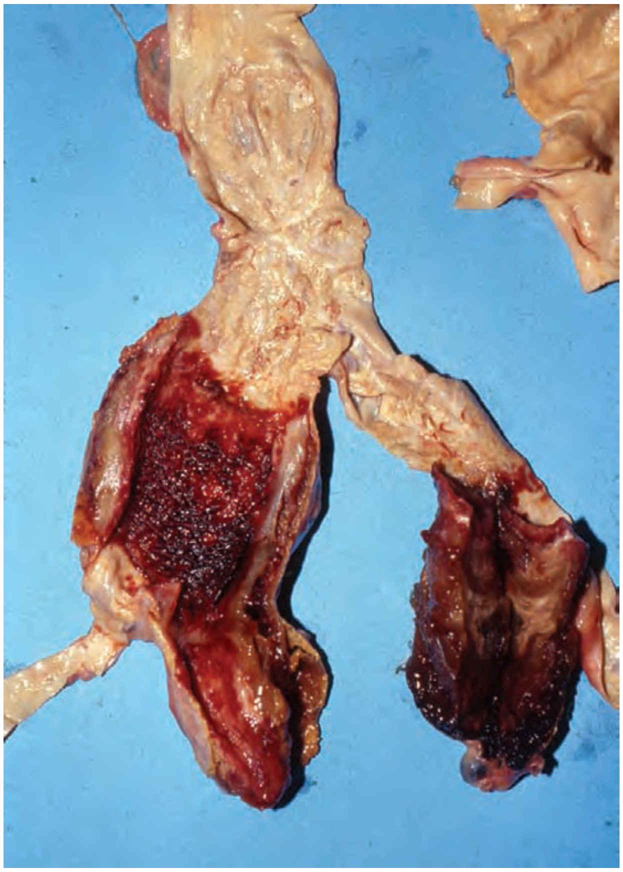

This demonstrates a decedent who had multiple atherosclerotic aneurysms including both iliac arteries.

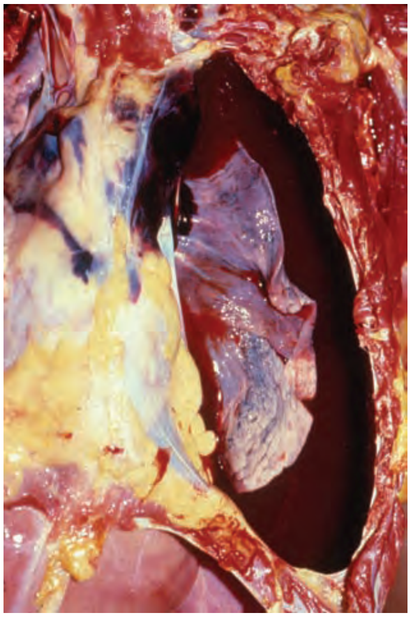

Hemothorax from a ruptured thoracic atherosclerotic aneurysm.

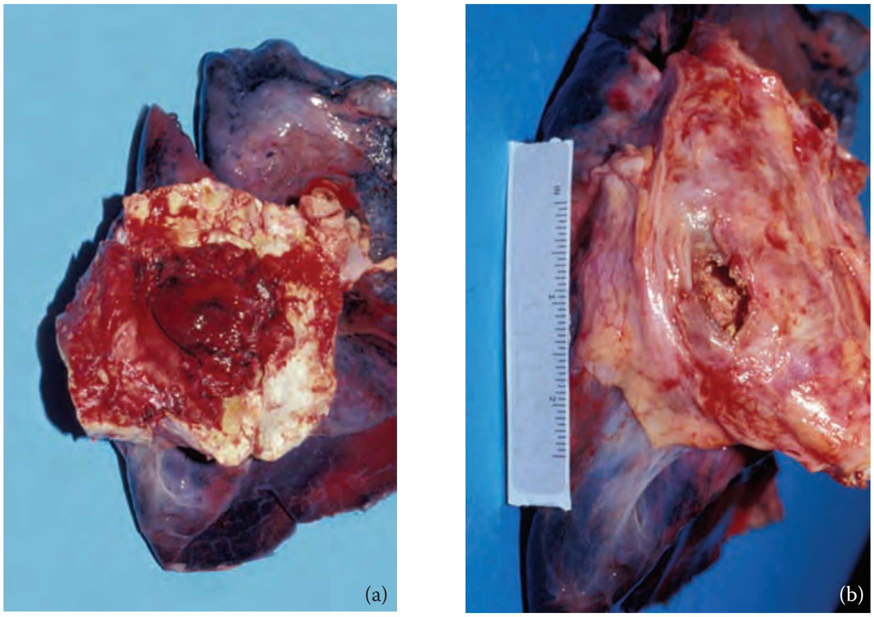

Thoracic atherosclerotic aortic aneurysm adherent to lung with rupture into the lung parenchyma causing massive hemoptysis. Note that Figure (b) demonstrates the rupture site with adherent blood clot removed.

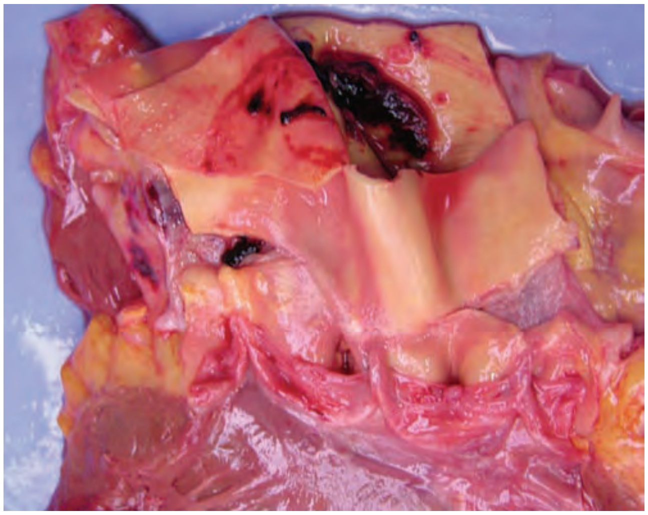

Intimal tear of the ascending aorta with dissection. Note blood tracking through the separated media.

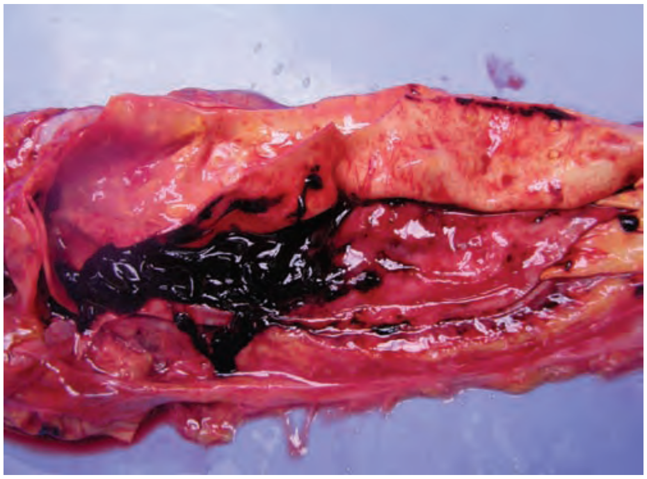

Aortic dissection with exposed separated media with blood clot in a person with Marfan's syndrome.

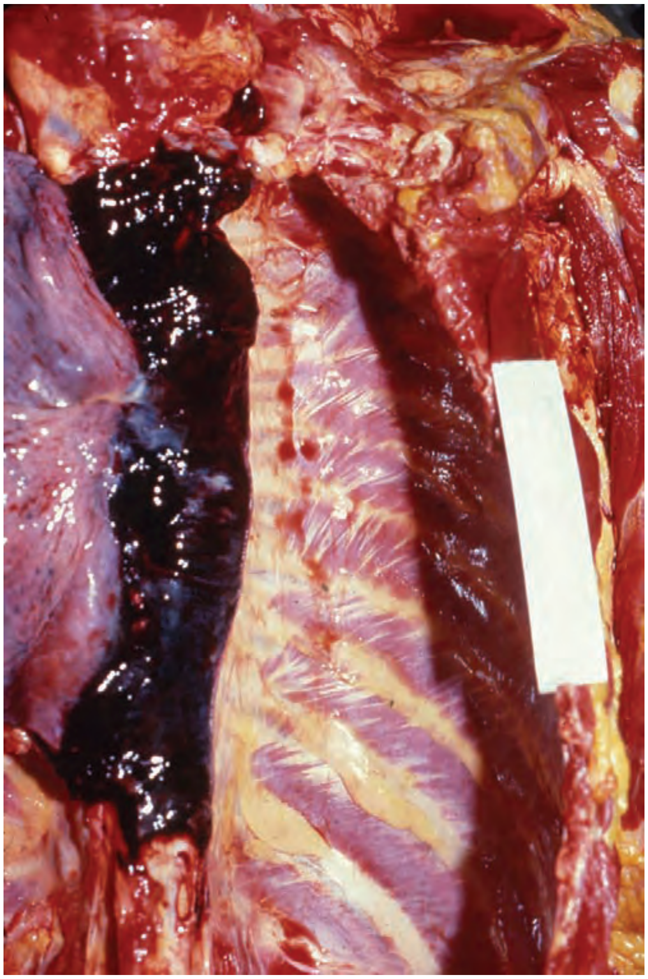

An in situ aortic dissection. Note the hemorrhage extending from the root of the aorta down the paravertebral region shown by dark red hemorrhagic discoloration. This decedent had severe hypertensive cardiovascular disease.

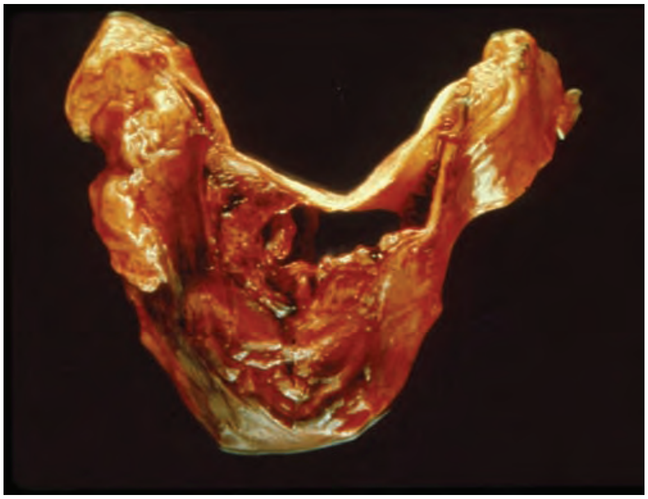

Cross section of aorta revealing a double-barrel lumen. The superior aspect of this figure shows an opened aorta with the exposed lumen. Directly inferior to this is the separated media and adventitia with a second lumen that is partially thrombosed.

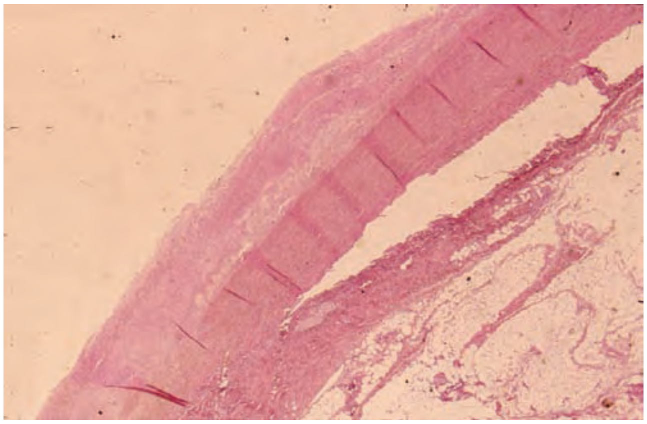

Microscopic view of an aortic dissection.

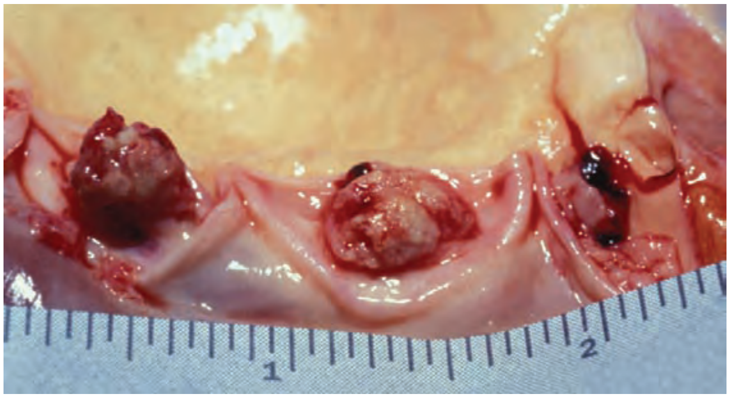

This endocardial surface shows large nonbacterial thrombotic endocarditis associated with a hypercoagulable state from metastatic adenocarcinoma. Special stains were negative for microorganisms.

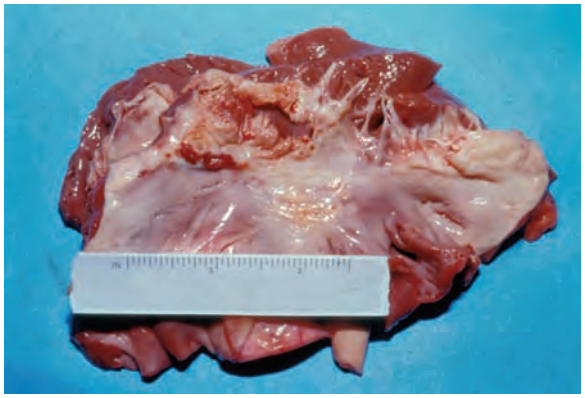

This perforated mitral valve is secondary to acute bacterial endocarditis. This individual first went to the emergency room approximately a day and a half before with the complaint of fever and chest pain. He was sent home with antibiotics and later returned with severe pulmonary edema and died shortly after.

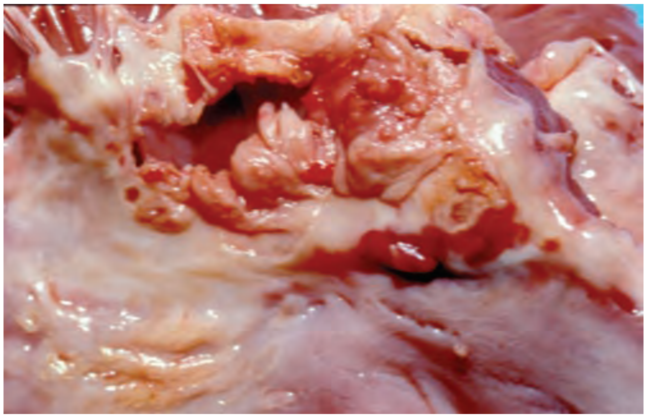

Close-up view of an acute infectious endocarditis with valve perforation. Gram stain revealed numerous gram positive organisms.

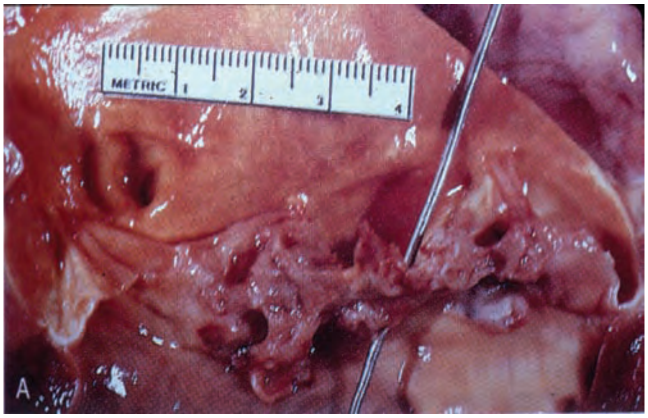

Acute infectious endocarditis with valve leaflet perforation.

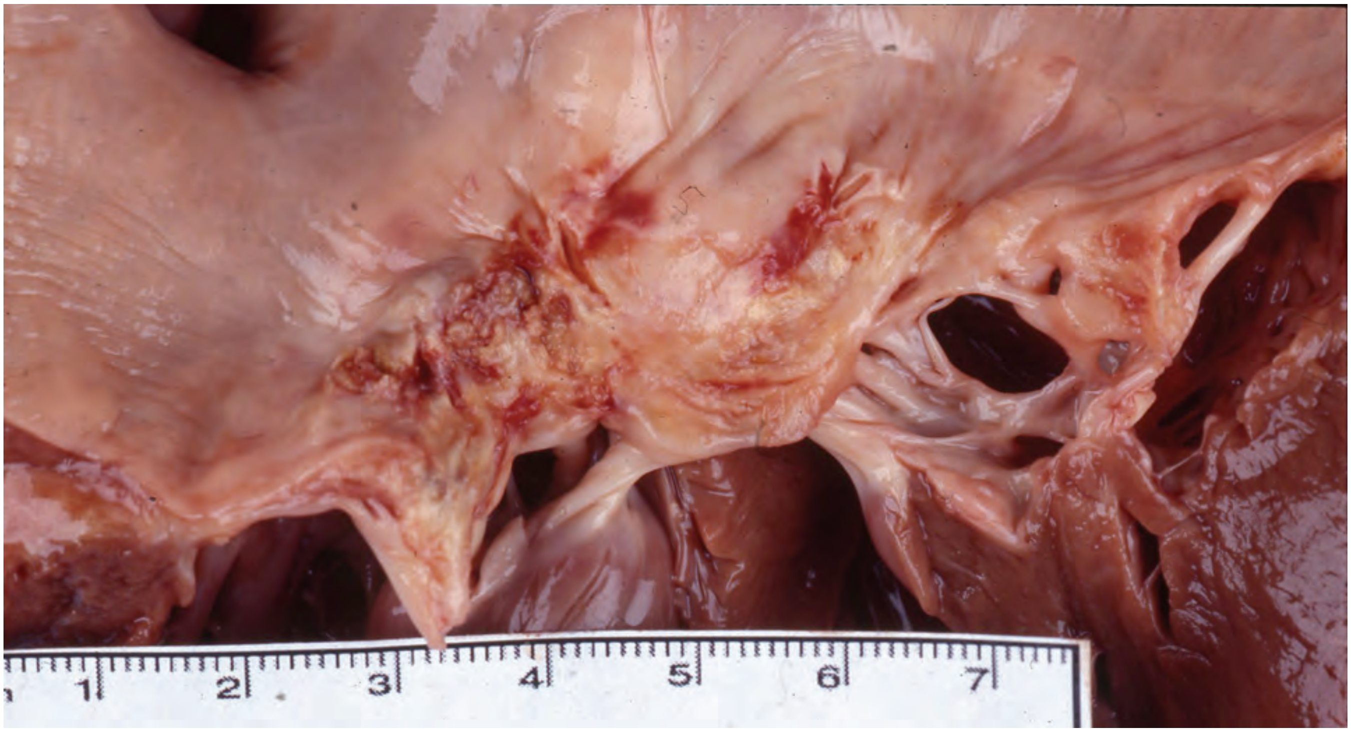

Remote cardiac valve damage from rheumatic fever.

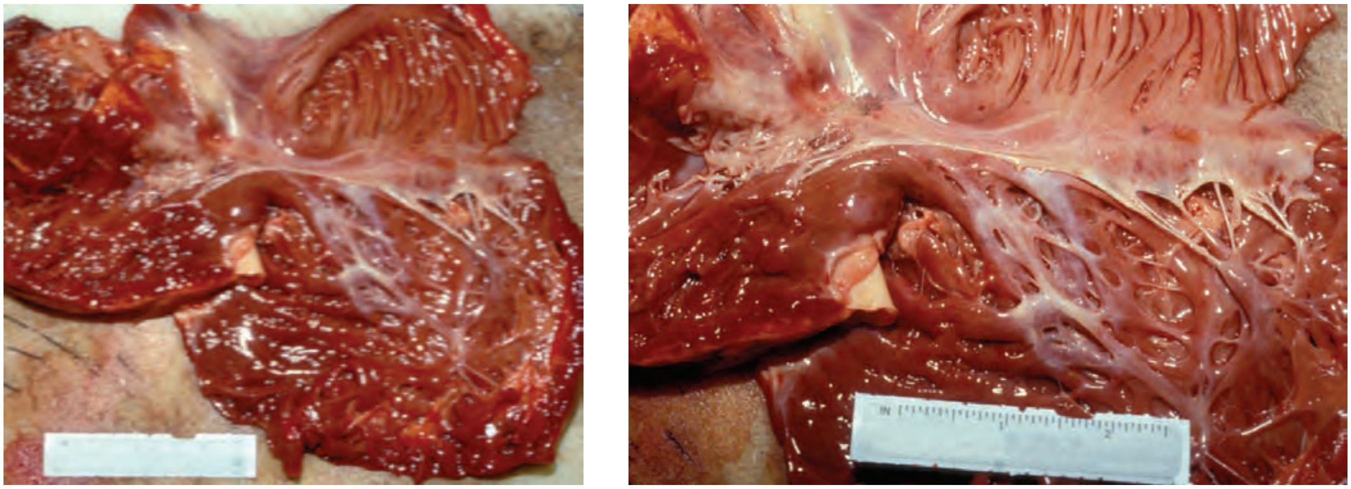

These figures are views of a remotely damaged tricuspid valve secondary to chronic intravenous drug abuse and past endocarditis. Note the fibrosis of the adjacent endocardium secondary to regurgitive turbulent blood flow. The decedent was known to have a longstanding cardiac murmur.

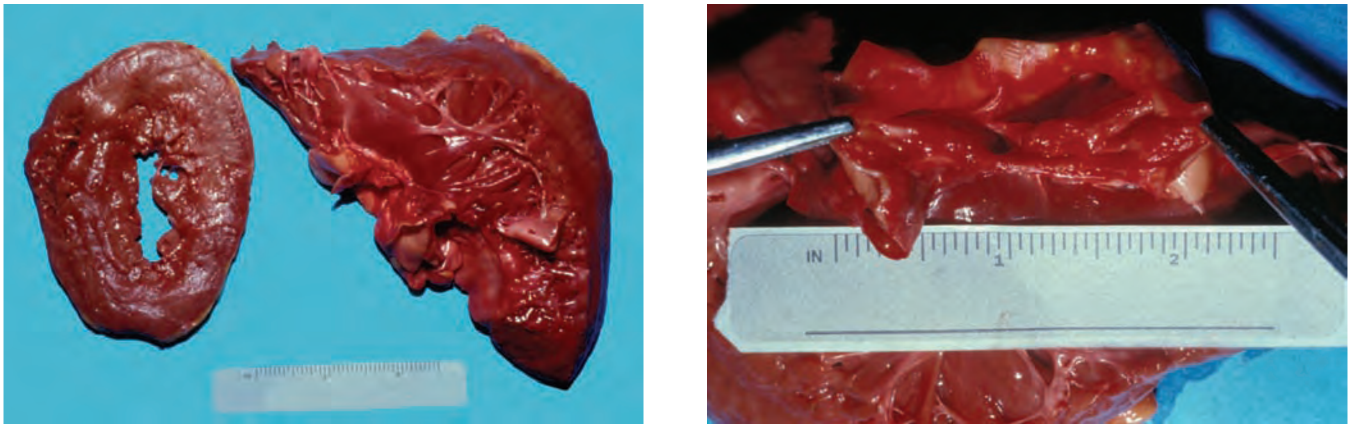

These figures show a markedly hypertrophic heart with concentric left ventricle hypertrophy and a congenital subaortic band causing marked aortic stenosis and sudden cardiac death at age 42 years. The decedent had decided years earlier not to have a valve replacement.

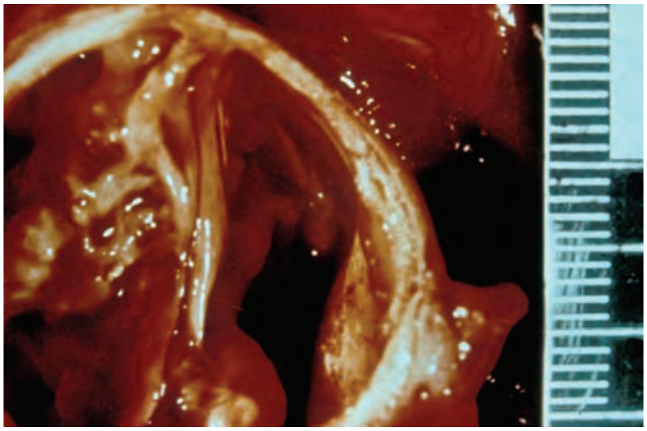

Bicuspid aortic valve. With advancing age and atherosclerosis, these valves may become markedly stenotic and increase one's risk for sudden cardiac death.

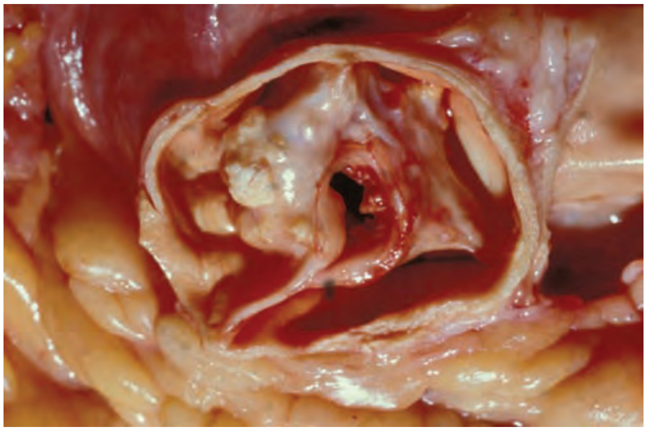

Markedly stenotic bicuspid aortic valve from an older individual with a longstanding history of atherosclerotic cardiovascular disease.

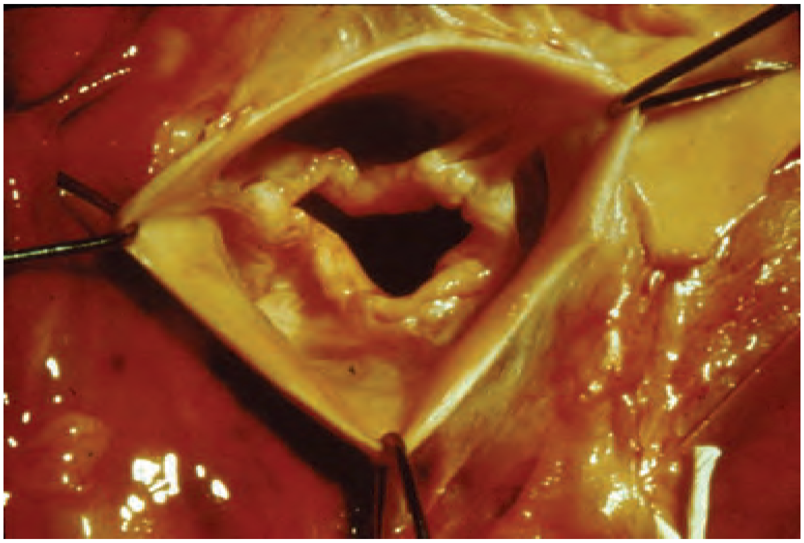

Severely stenotic and insufficient aortic valve associated with childhood rheumatic fever. Correction of aortic stenosis will decrease the risk of sudden cardiac death.

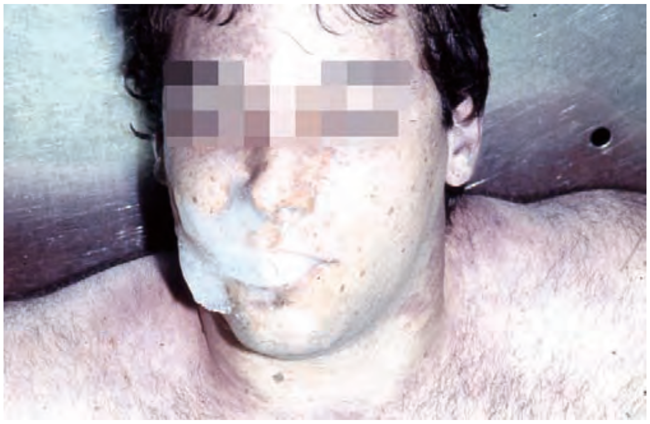

A fulminant pulmonary edema with foam extending from the mouth and nose due to congestive heart failure.

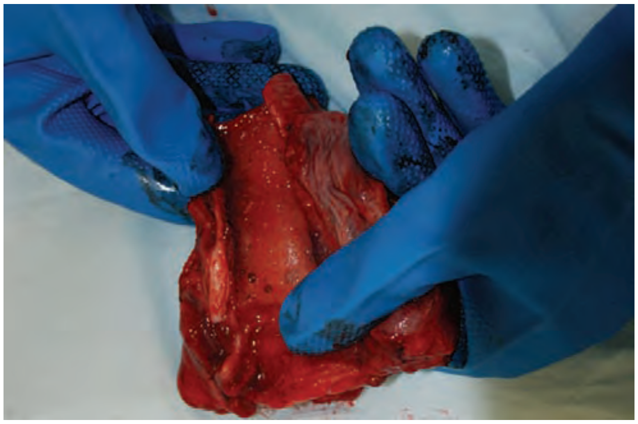

Note the frothy fluid from pulmonary edema extending into the laryngeal airway.

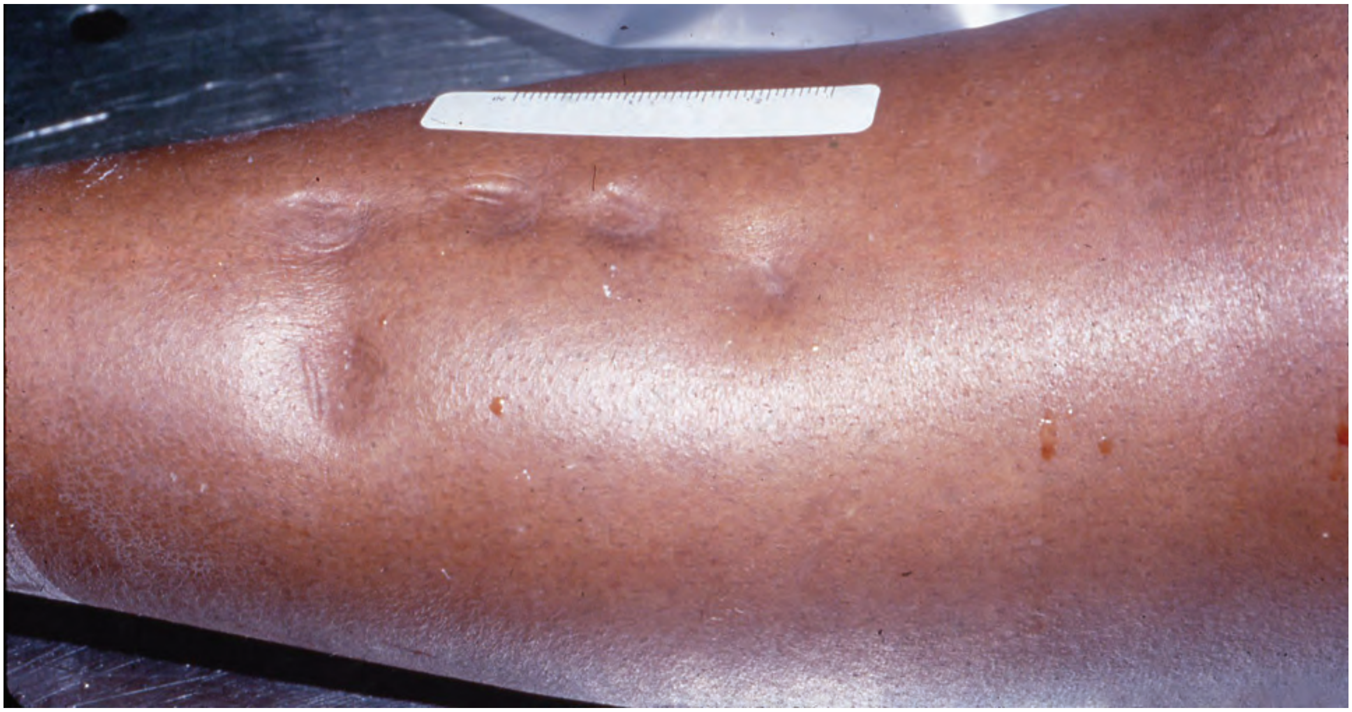

Pitting edema of the leg due to congestive heart failure.

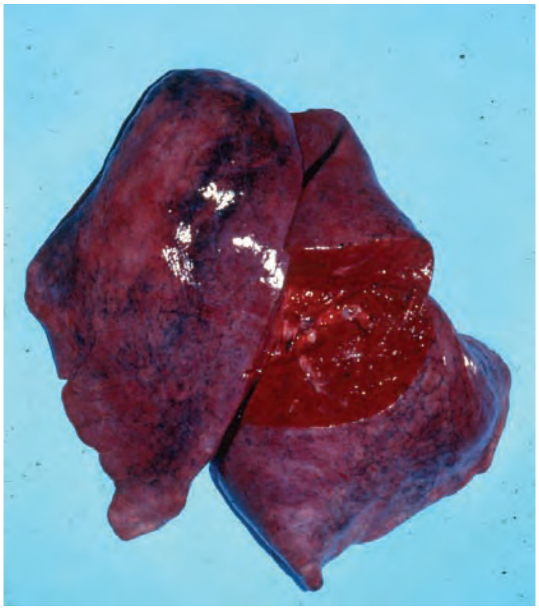

Lung with marked congestion and edema. Note the diffuse purple discoloration.

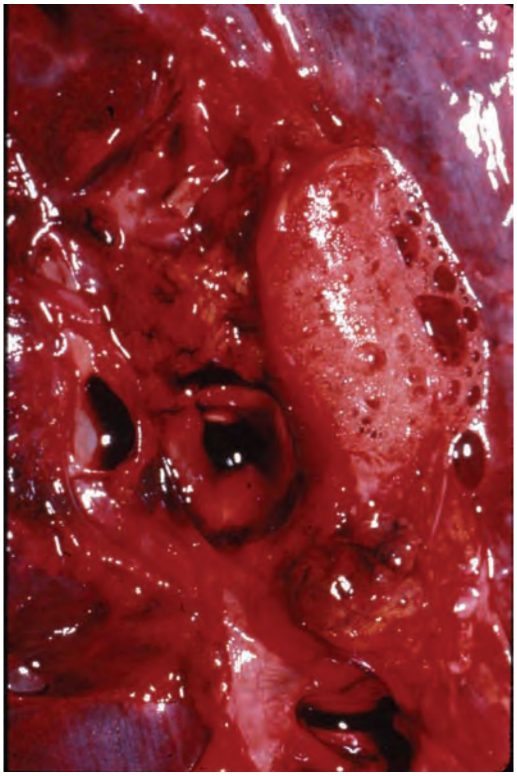

Marked pink to red frothy fluid extending from the cut parenchyma of a lung due to fulminant pulmonary edema.

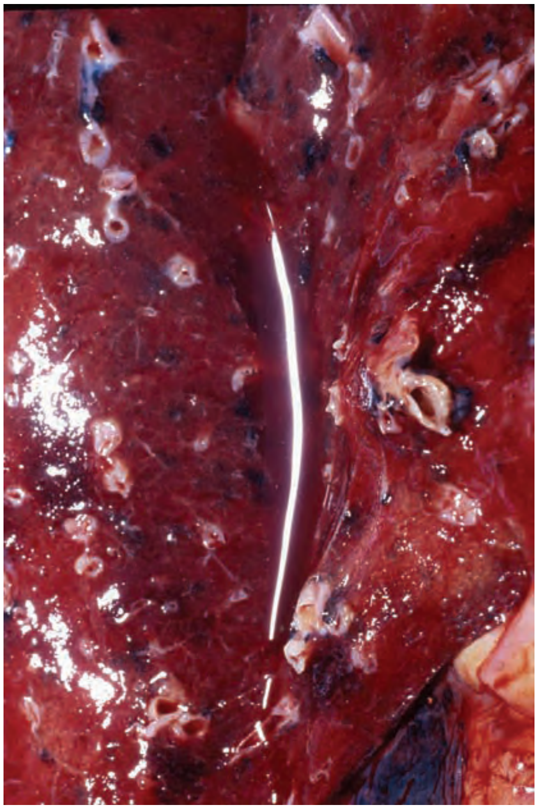

Cut section of a lung with pulmonary edema.

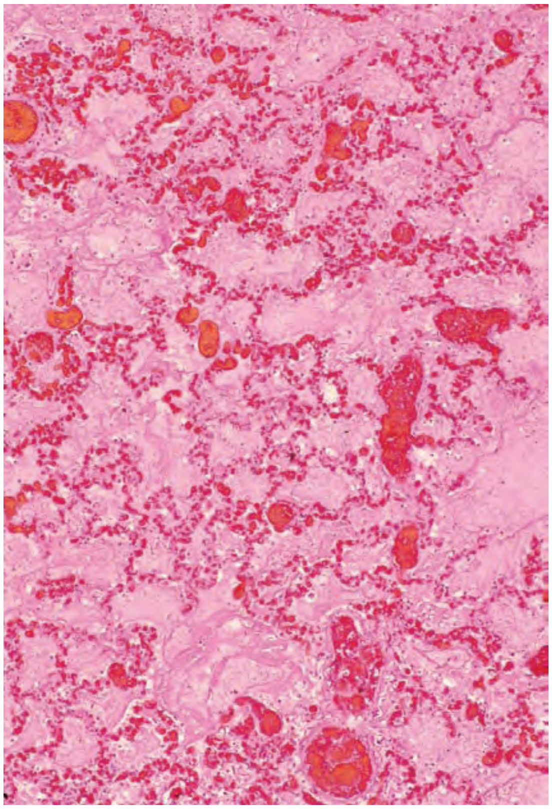

Microscopic view of lung with acute fulminant pulmonary edema.

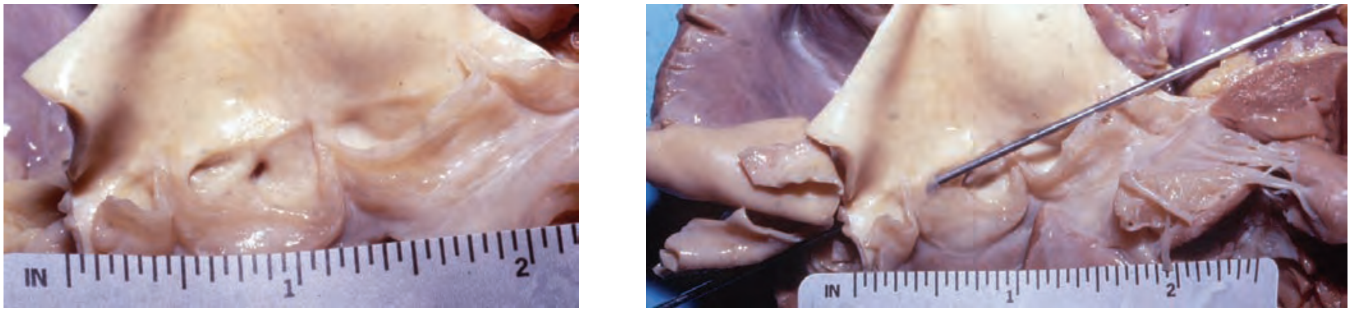

These two figures demonstrate a coronary artery anomaly with acute angle takeoff and luminal narrowing in a 15-year-old who died suddenly during a basketball game. There was no history of blunt impacts to the chest during the game. There was no past history of syncopal episodes or chest discomfort.

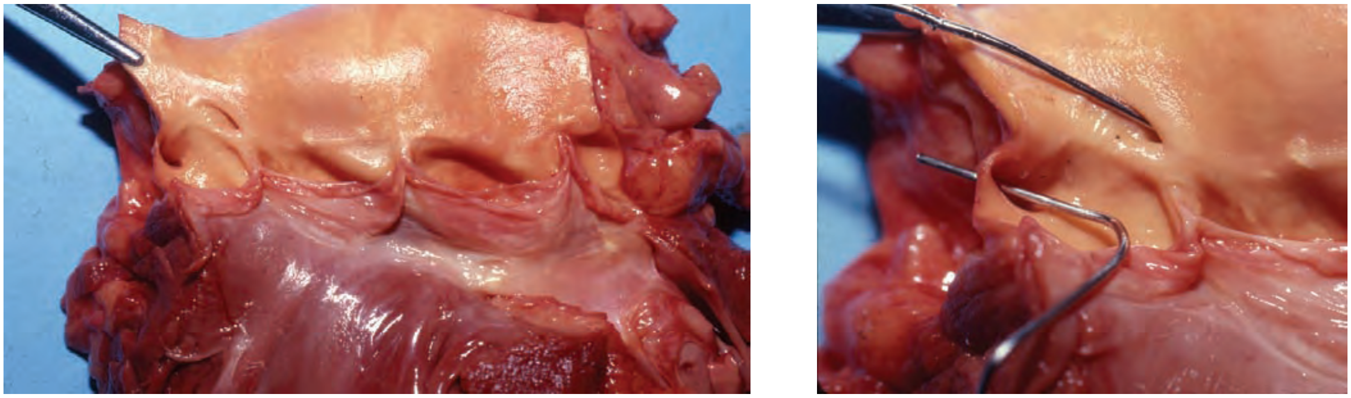

A coronary artery anomaly with acute angle takeoff and luminal narrowing. This individual died a sudden cardiac death shortly after exertion.

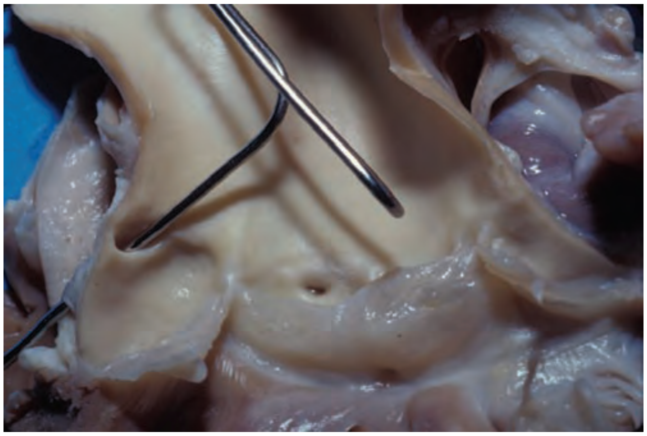

This coronary anomaly reveals a bicuspid aortic valve with superior displacement of one of the coronary ostia. This child died of trauma sustained in a motor vehicle accident and this finding was incidental.

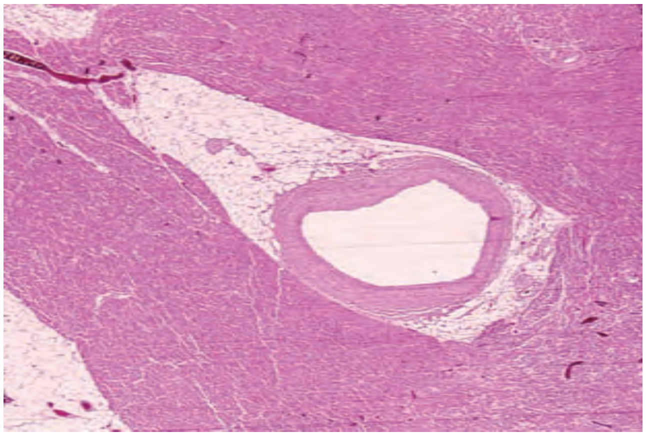

This low-power histopathologic view demonstrates "myocardial tunneling" where the left anterior descending coronary artery dipped deeply into the left ventricle wall.

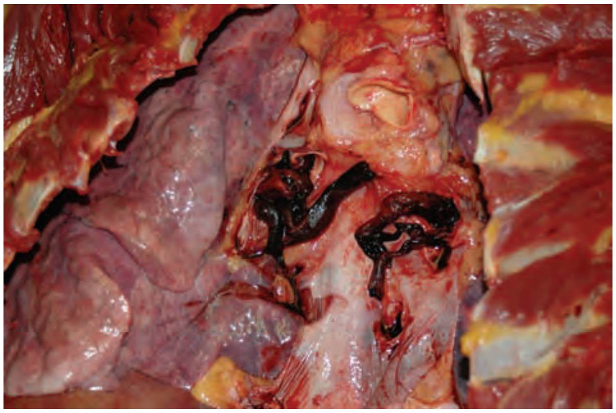

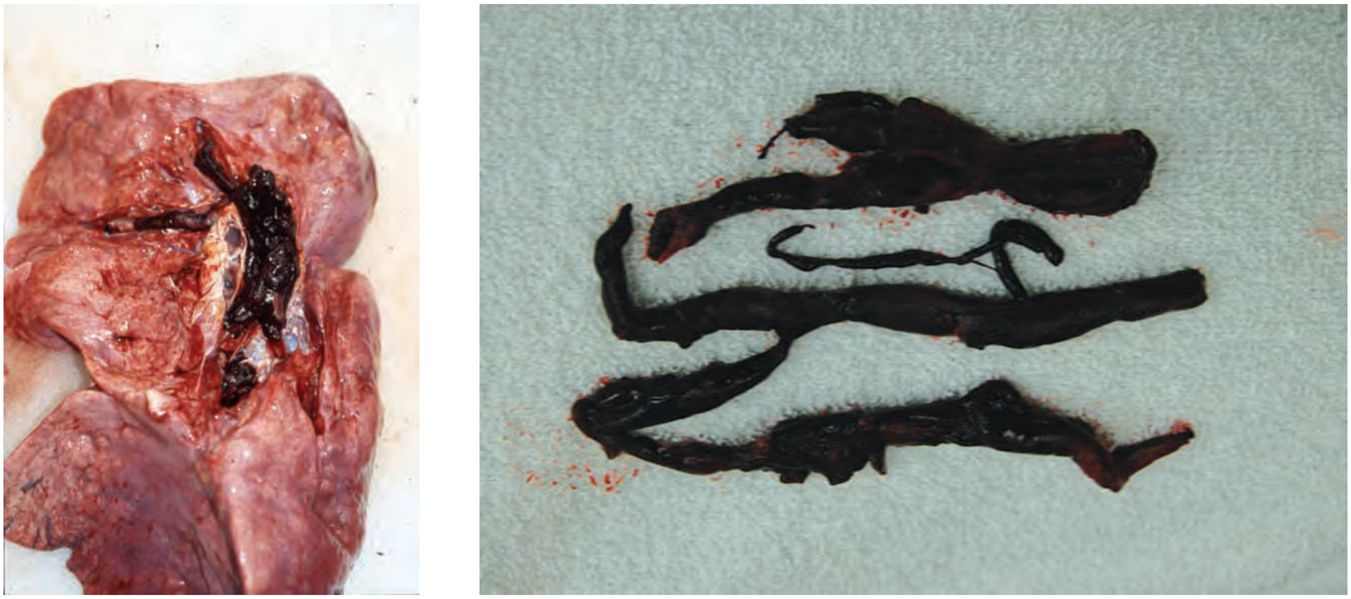

Pulmonary thromboembolus in an obese woman who was on birth control pills and smoked cigarettes.

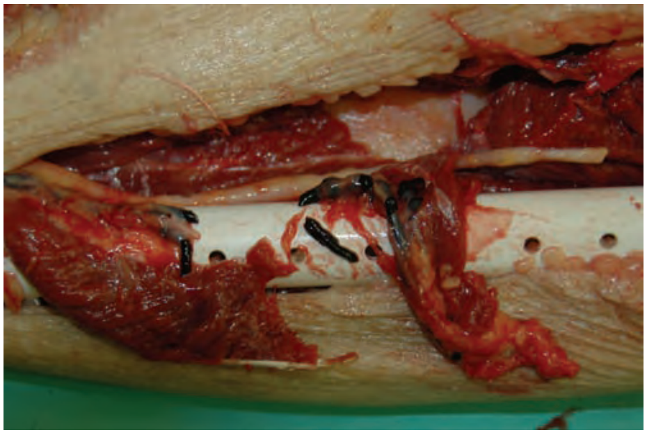

Incision of the lower leg with dissection of the gastrocnemius muscle demonstrating multiple deep venous thrombi in an individual who underwent organ donation with removal of bone and soft tissue from each leg. Note the plastic tubing left by the organ donor network.

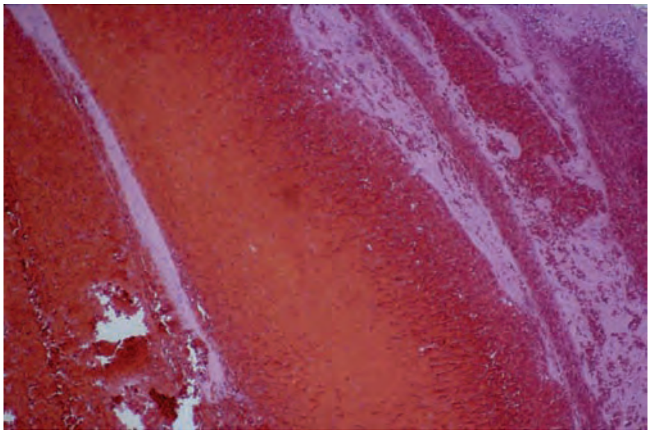

Microscopic view of a pulmonary thromboembolus. Note the laminated alternating lighter fibrin bands referred to as the lines of Zahn.

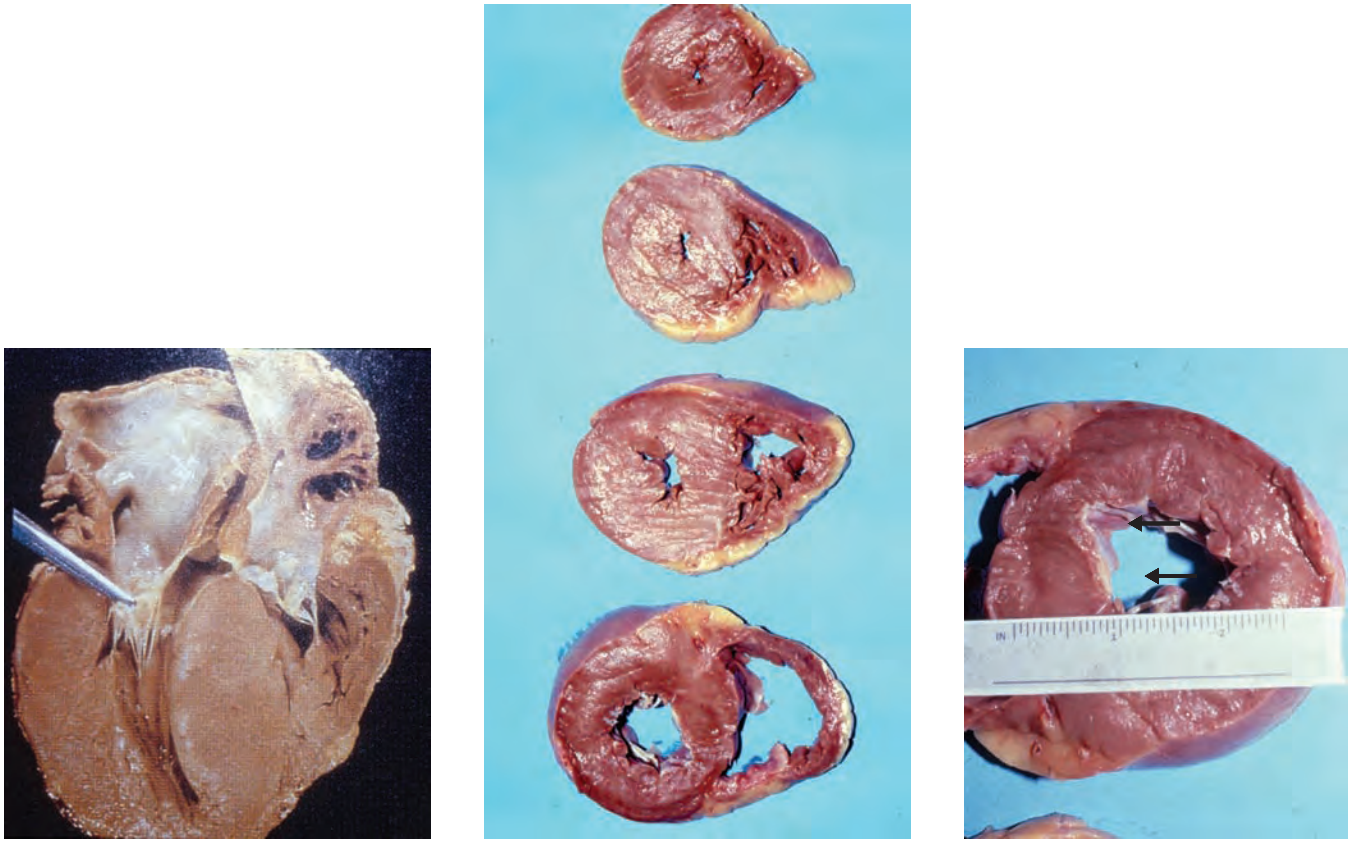

These are different views of two hearts with the classical variant of hypertrophic cardiomyopathy. Note the large degree of asymmetric left ventricle hypertrophy.

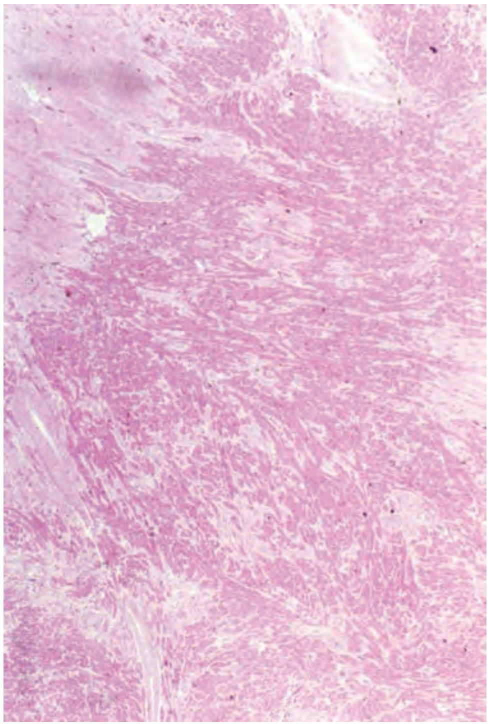

Hypertrophic cardiomyopathy with myocyte fiber disarray and fibrosis.

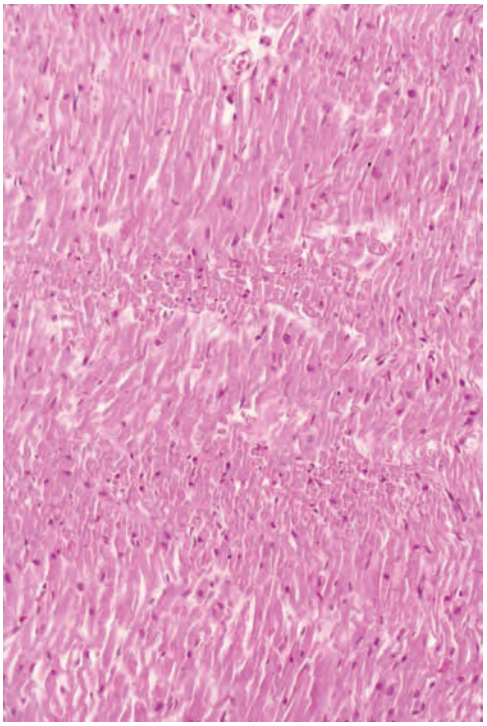

Hypertrophic cardiomyopathy with myocyte fiber bundle disorganization.

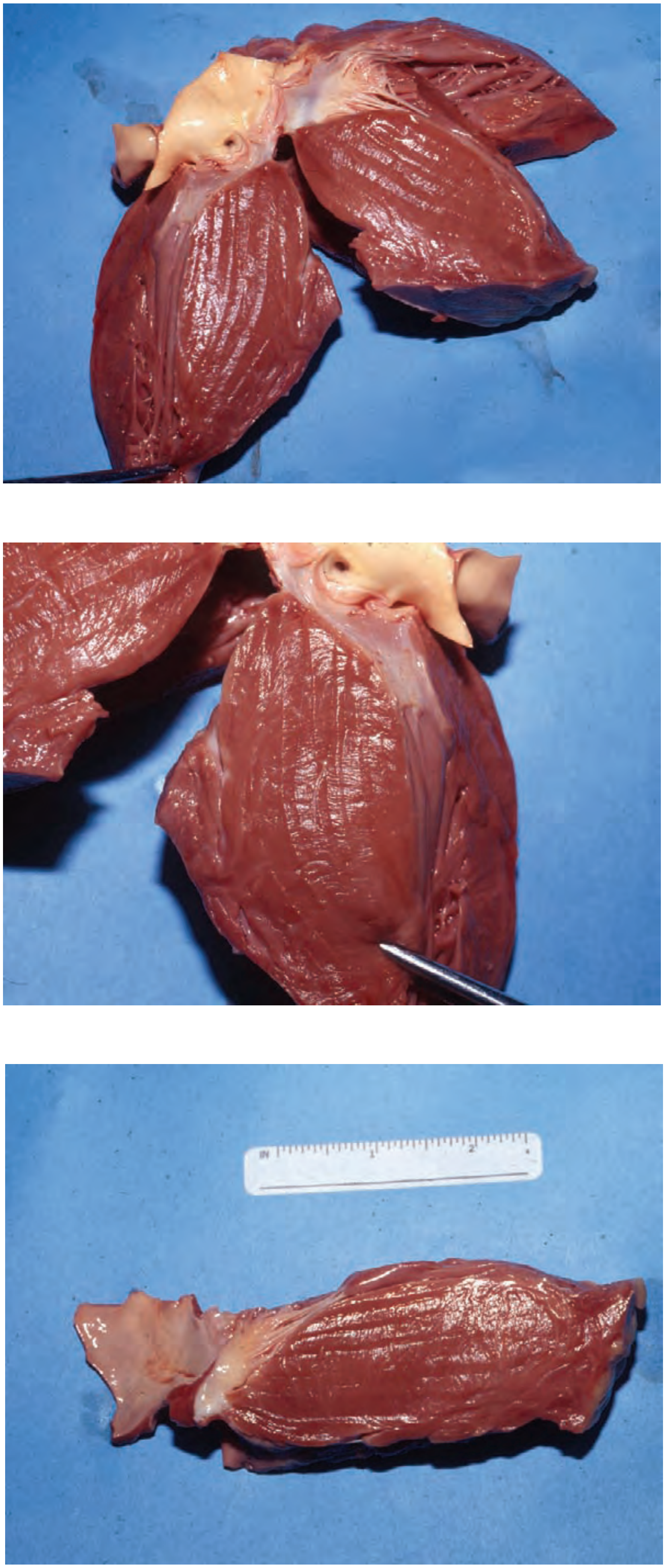

Hypertrophic cardiomyopathy. Note the marked septal hypertrophy with asymmetry.

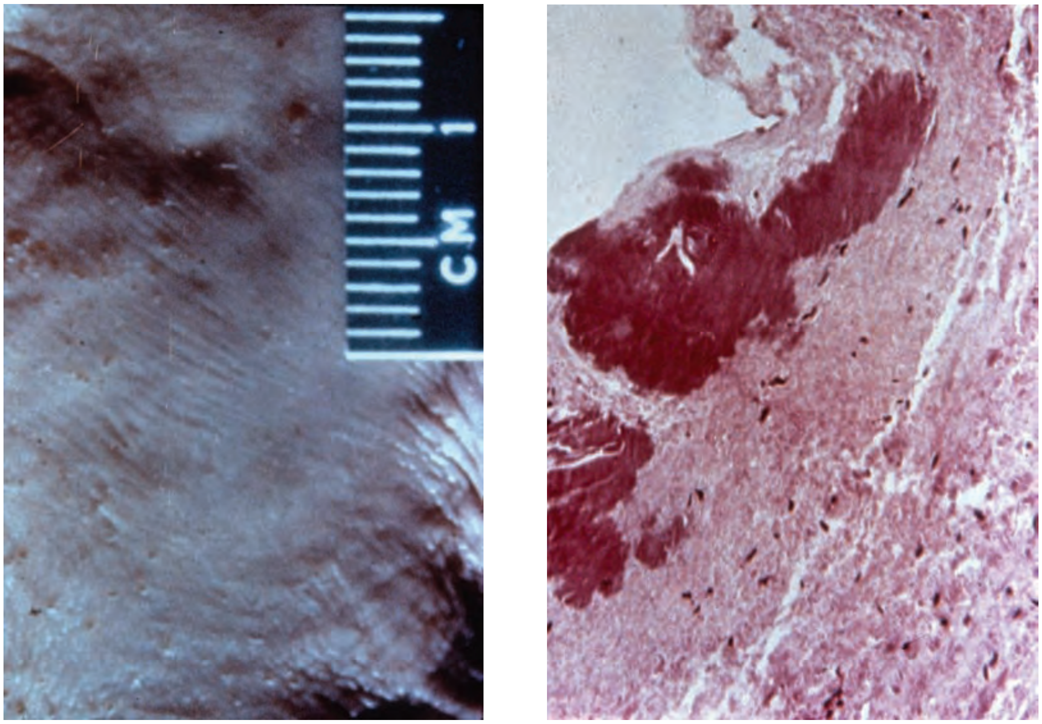

Note the waxy pink myocardial discoloration in this individual with amyloidosis.

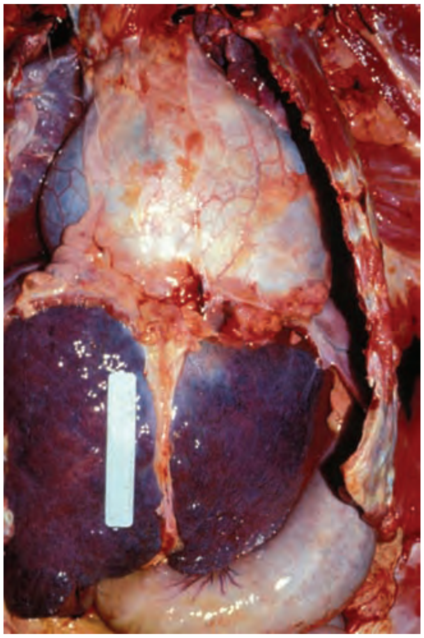

This decedent's pericardial sac was full of several hundred cc's of yellow-pink fluid. The gradual buildup of this fluid over a long period was due to advanced secondary hemochromatosis.

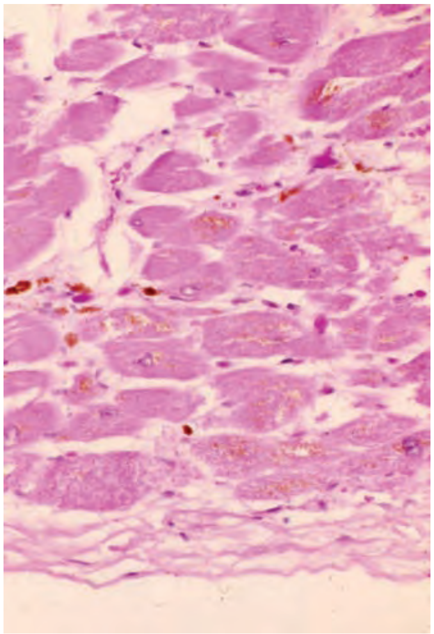

Microscopic section of heart showing iron deposition from secondary hemachromatosis associated with blood transfusions for treatment of beta-thalassemia major.

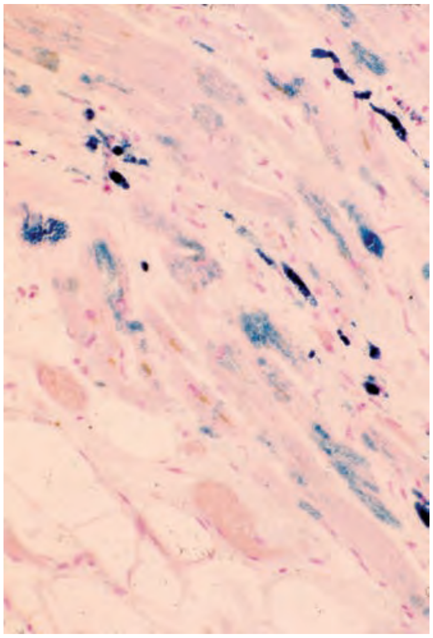

Microscopic section of heart with Prussian blue staining of iron deposits in the heart of the same individual with secondary hemochromatosis.

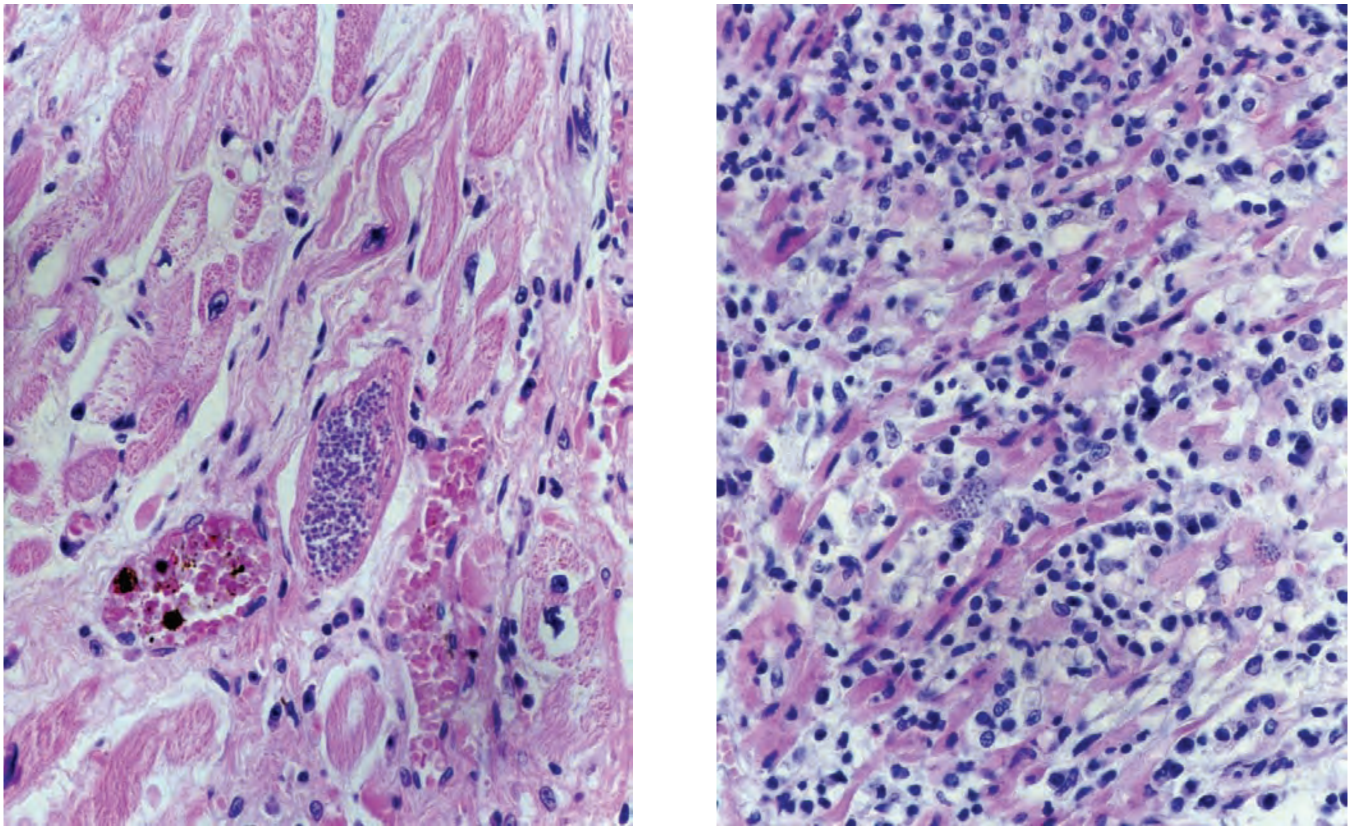

These microscopic myocardial views demonstrate Chagas' disease. Note the extensive chronic inflammation.

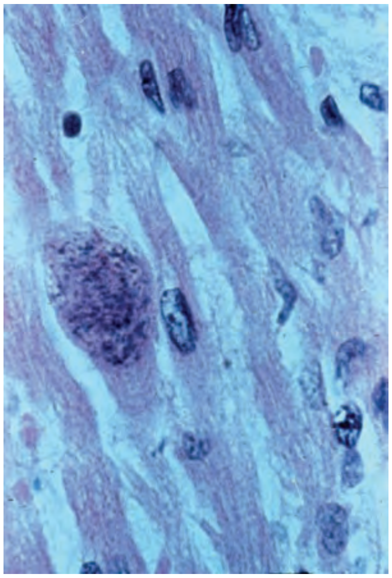

Myocarditis secondary to toxoplasmosis in an individual with acquired immunodeficiency syndrome.

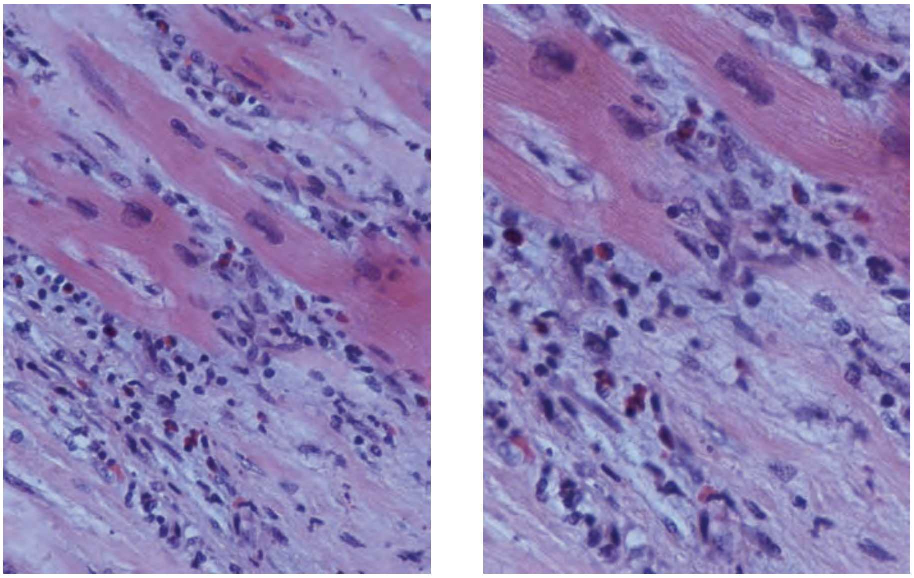

Eosinophilic myocarditis associated with a hypersensitivity drug reaction.

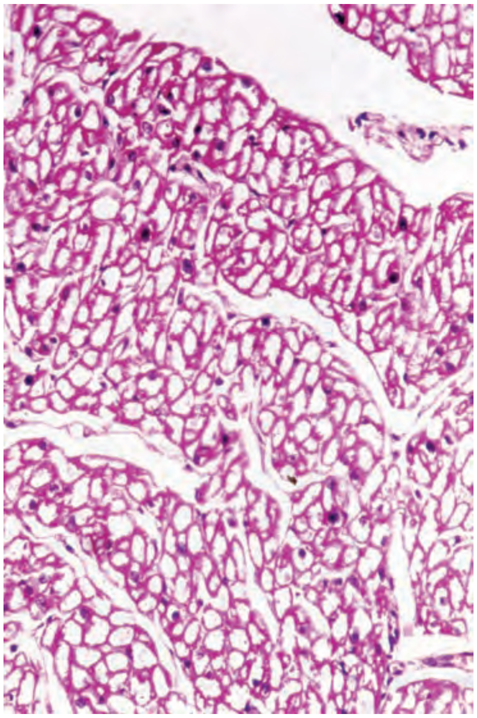

Microscopic view of myocardium in an individual who died of Pompe's disease. Note the many artifactual empty vacuoles following processing leading to glycogen loss.

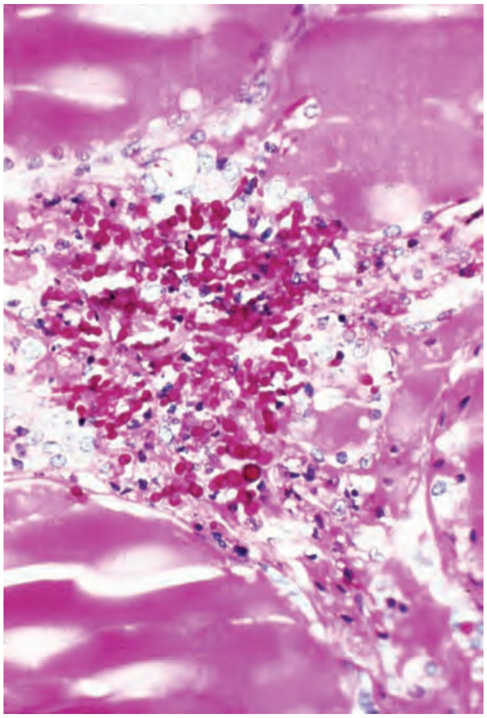

This candidal myocarditis was found in an individual with acquired immunodeficiency syndrome.

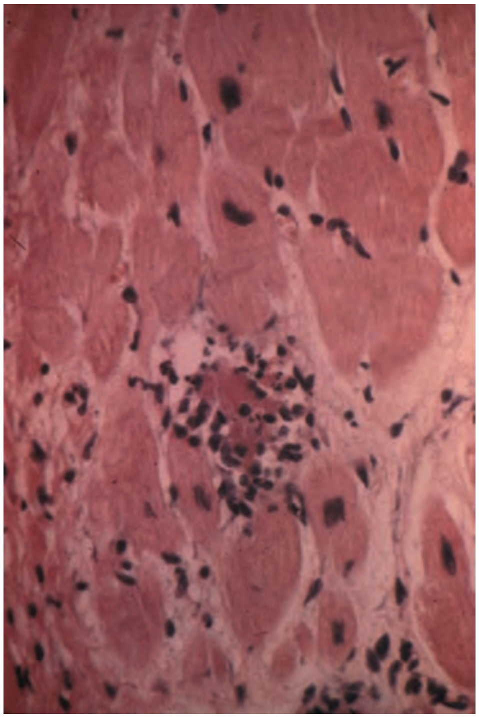

This focal viral myocarditis demonstrates myocardial necrosis with inflammation made up predominately of lymphocytes.

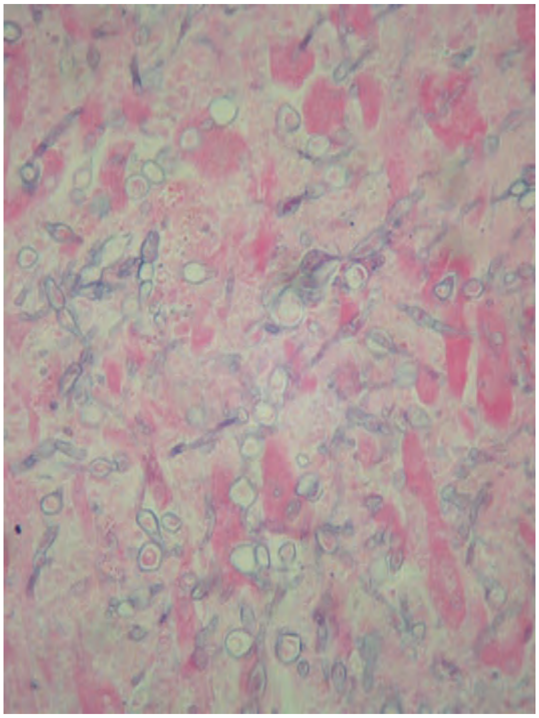

Aspergillosis myocarditis from an immunocompromised individual who underwent a recent bone marrow transplant.

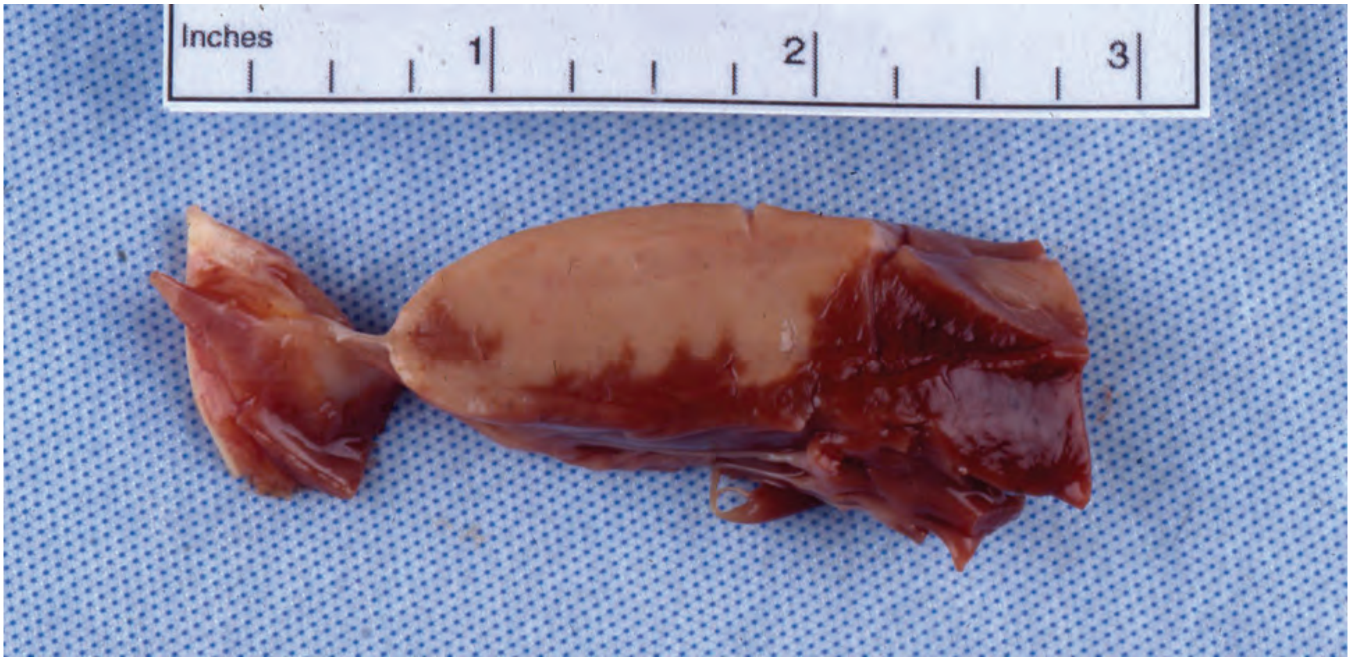

Macroscopic section of myocardium from an individual who died of cardiac sarcoidosis.

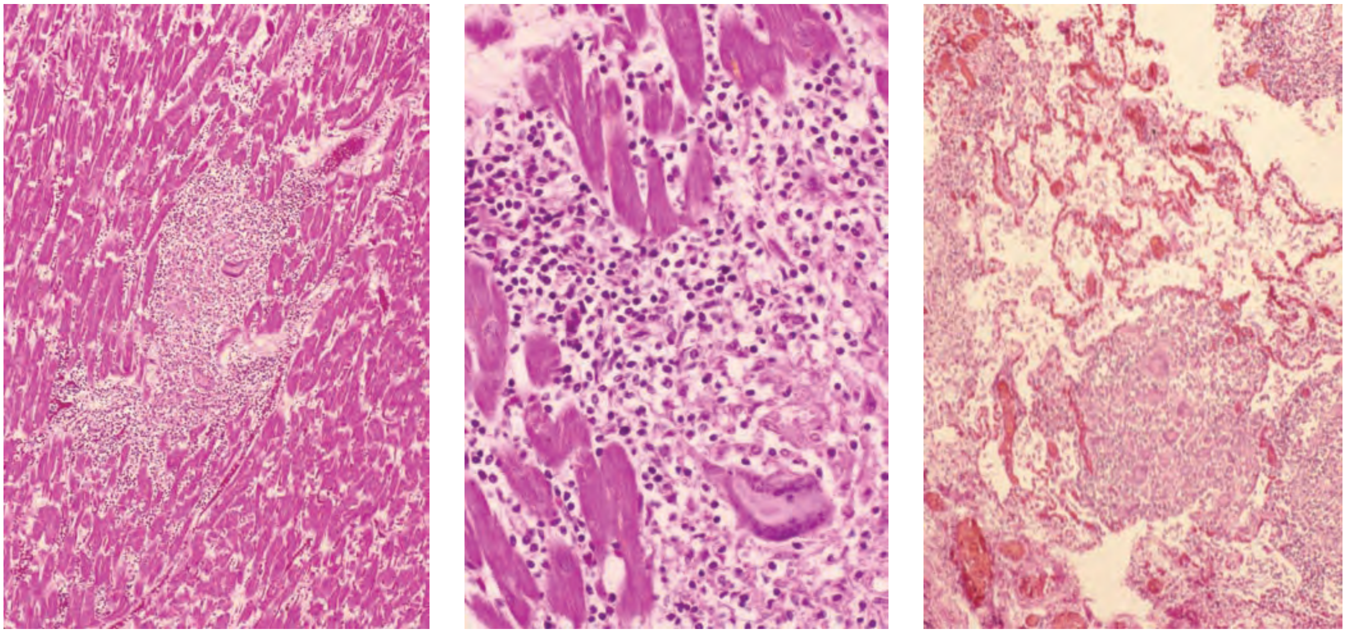

Sarcoidosis. Note the noncaseating granulomas consisting of chronic inflammation, epitheloid cells, lymphocytes, and giant cells found throughout the body. These views demonstrate examples from the heart and lungs.

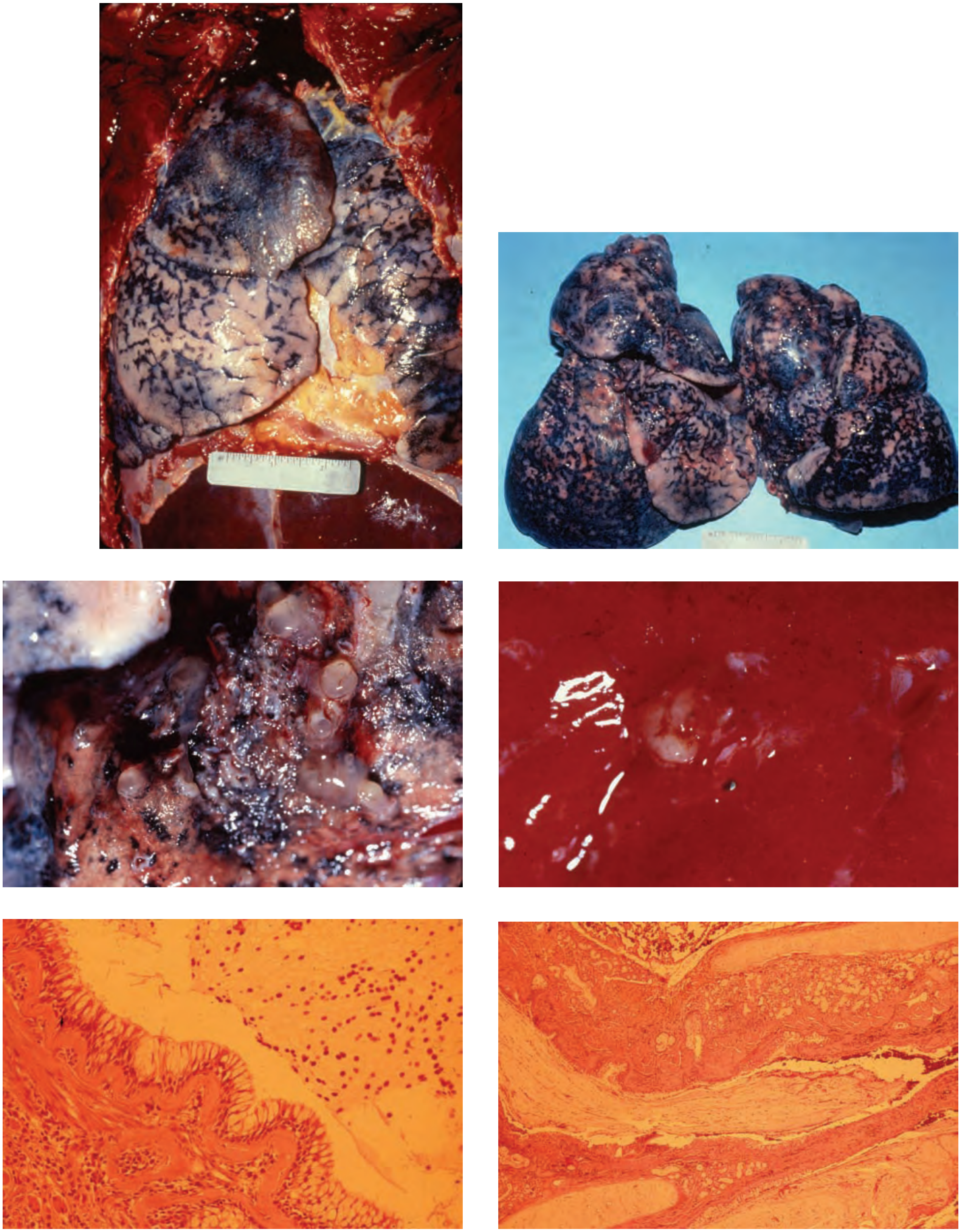

These figures demonstrate acute and chronic bronchial asthma. Note the lungs within the thoracic cavity are hyperaerated and expand to overlie the pericardial sac. Upon removal of the lung, they appear markedly hyperaerated. If these lungs were placed in a water bath they would float almost entirely on the surface. Cut section through the parenchyma reveals thick copious mucoid secretions within the bronchial distribution. Microscopic section reveals increased amounts of goblet cells and smooth muscle. There is a large mucous plug within the airway with numerous eosinophils. The parenchyma also has numerous inflammatory cells, predominantly eosinophils.

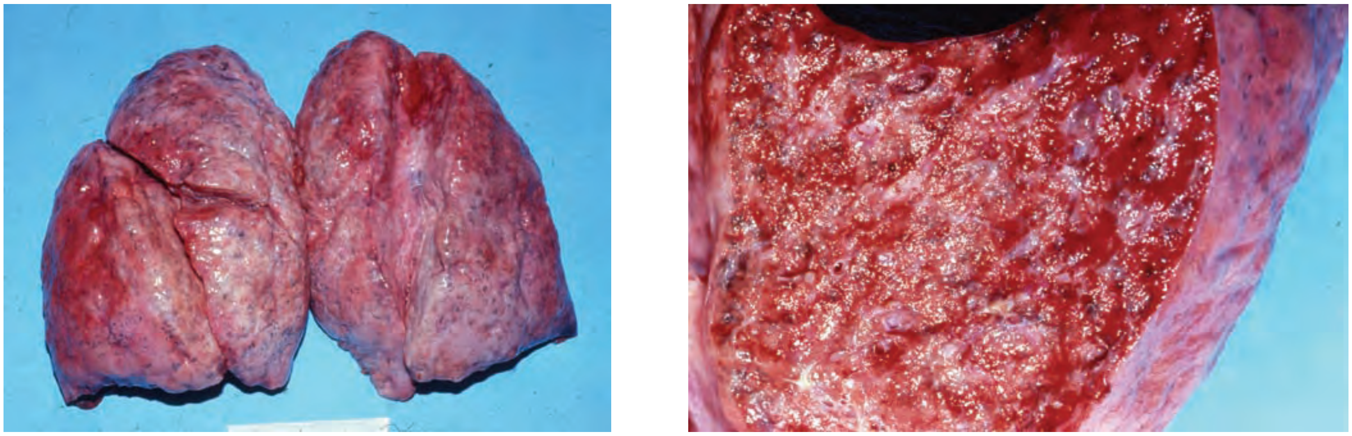

Chronic obstructive pulmonary disease (COPD) with hyperaeration and emphysematous change.

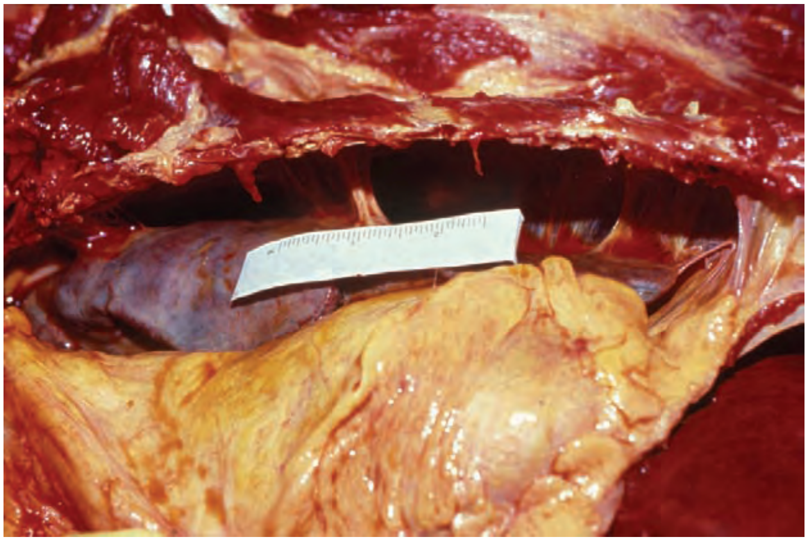

Pleural adhesions most likely due to past bouts of pneumonia.

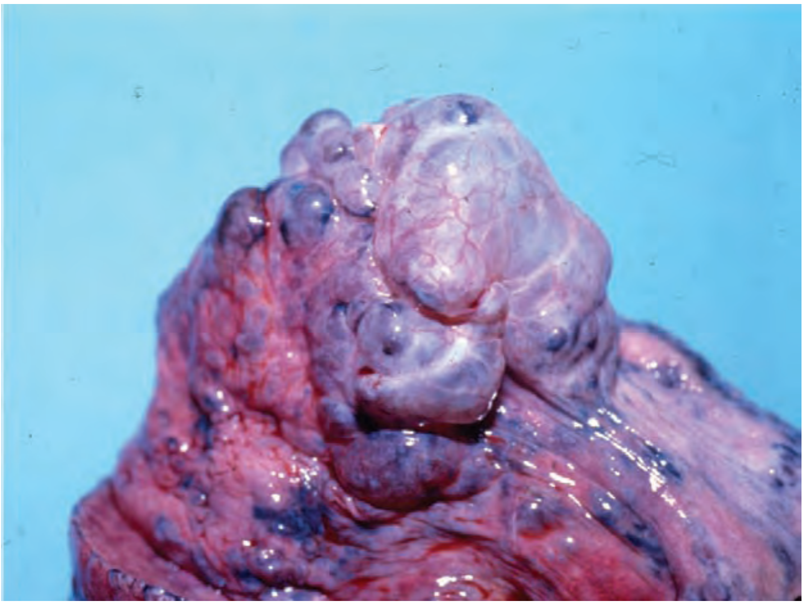

Bolus emphysema in a person with COPD. These bullae may occasionally rupture and cause a pneumothorax.

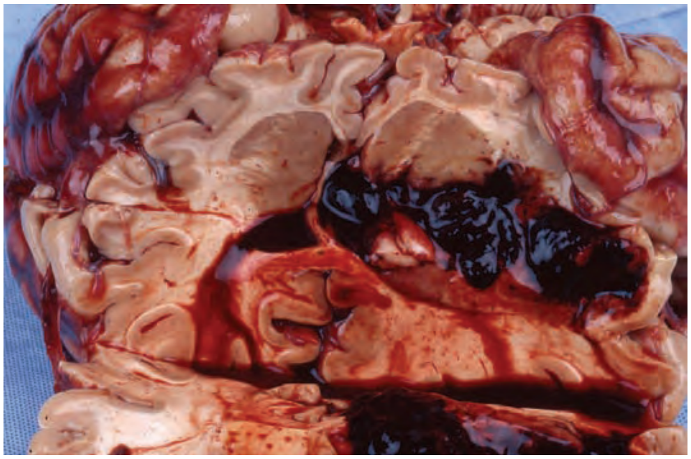

Hypertensive intracerebral hemorrhage. Mostly caused by rupture of a small intraparenchymal vessel.

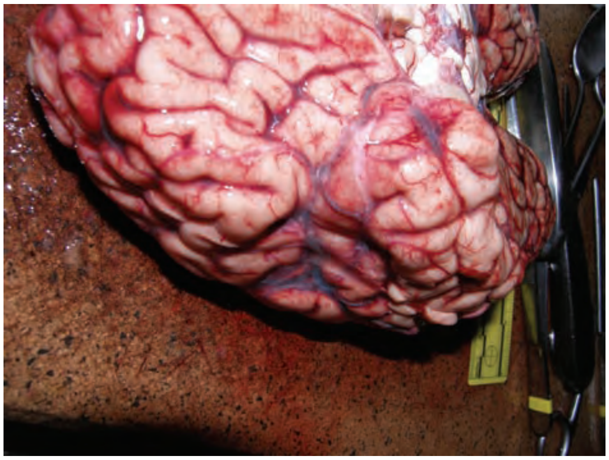

Old healed stroke. Note the indentation of the cerebral hemisphere.

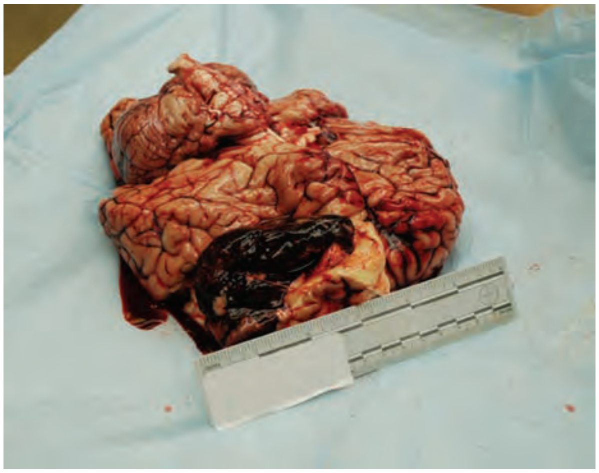

Intracerebral hemorrhage due to ruptured A-V malformation.

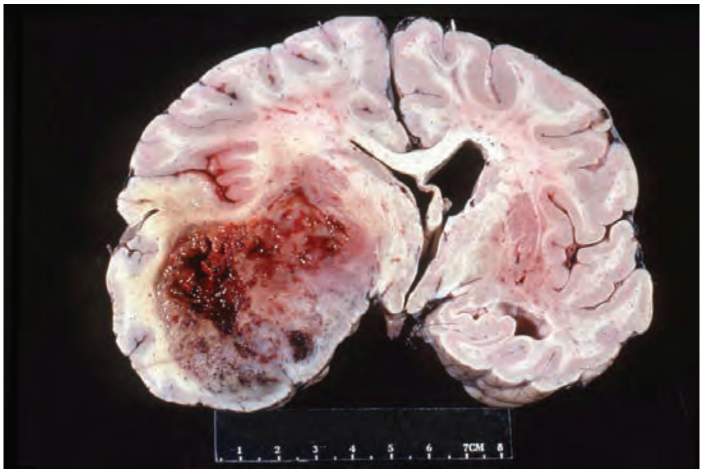

Astrocytoma causing significant compression of the surrounding structures.

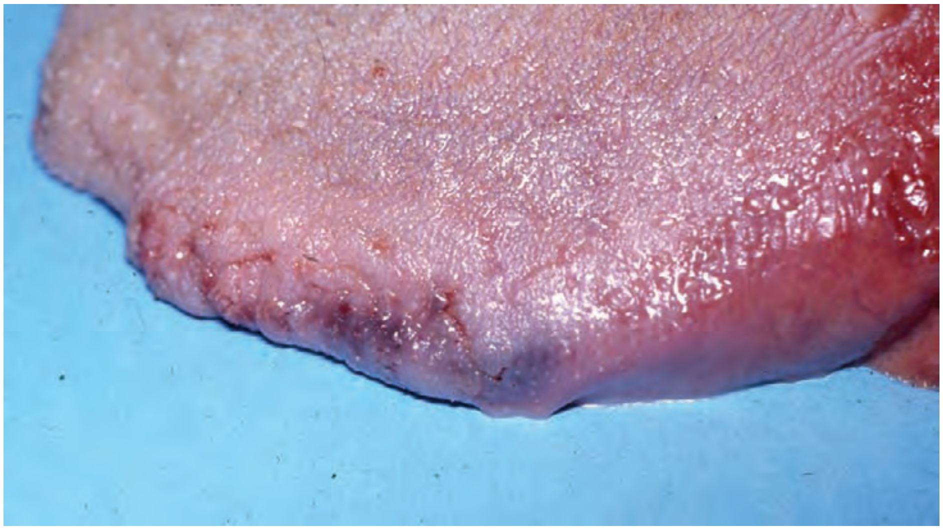

Bite marks to tongue with hemorrhage from an individual who died of epilepsy. This was the only finding at autopsy.

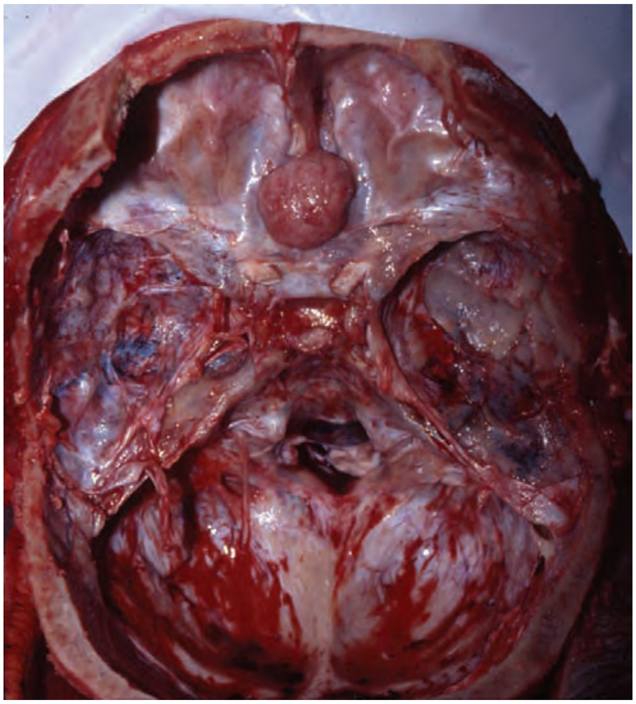

Anterior cranial fossa meningioma.

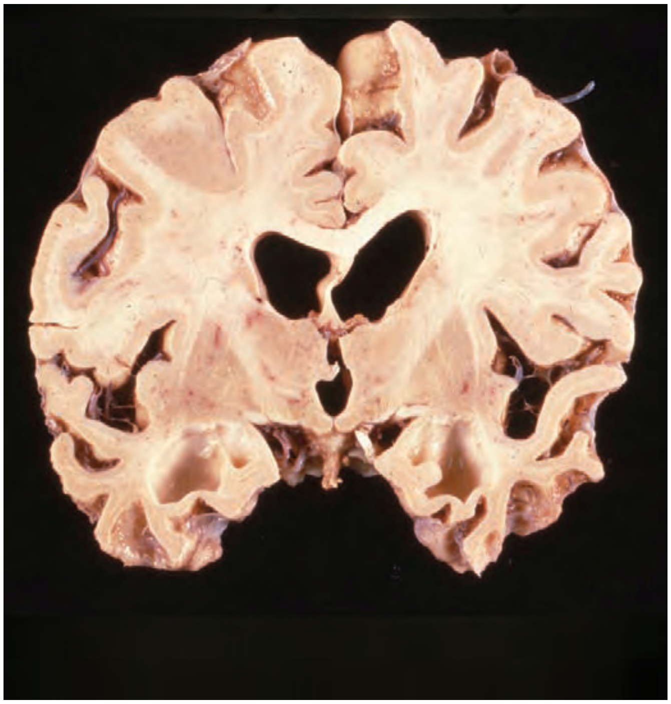

Pick's disease. Note the asymmetric atrophy of the frontal, temporal, and parietal lobes. This form of chronic dementia occurs far less frequently than Alzheimer's disease.

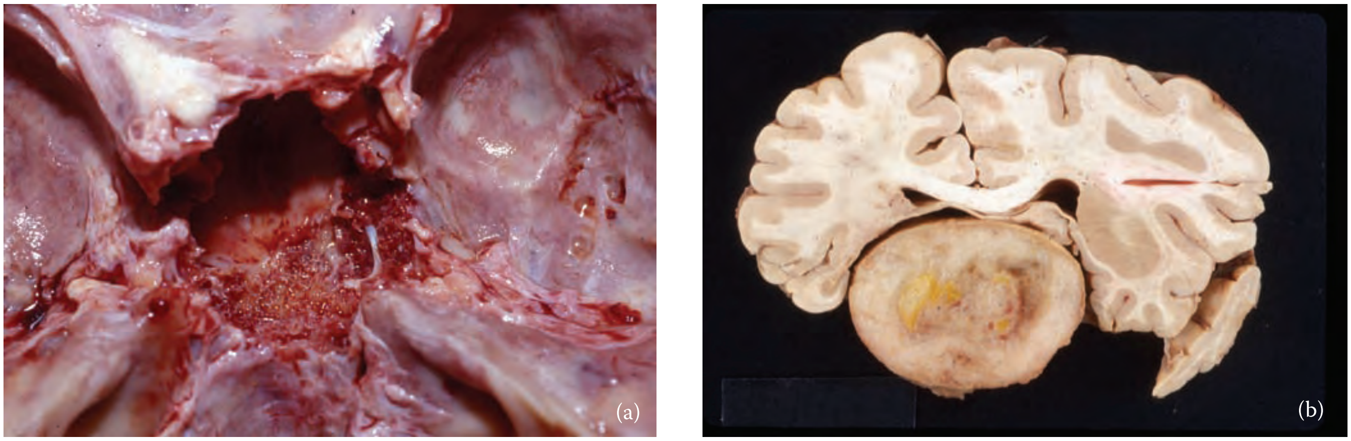

Two separate cases of pituitary adenoma. Figure (a) depicts a large erosion into the sela turcica and (b) demonstrates a large adenoma viewed on a cross section of a cut formaldehyde-fixed brain. These adenomas may vary largely in size.

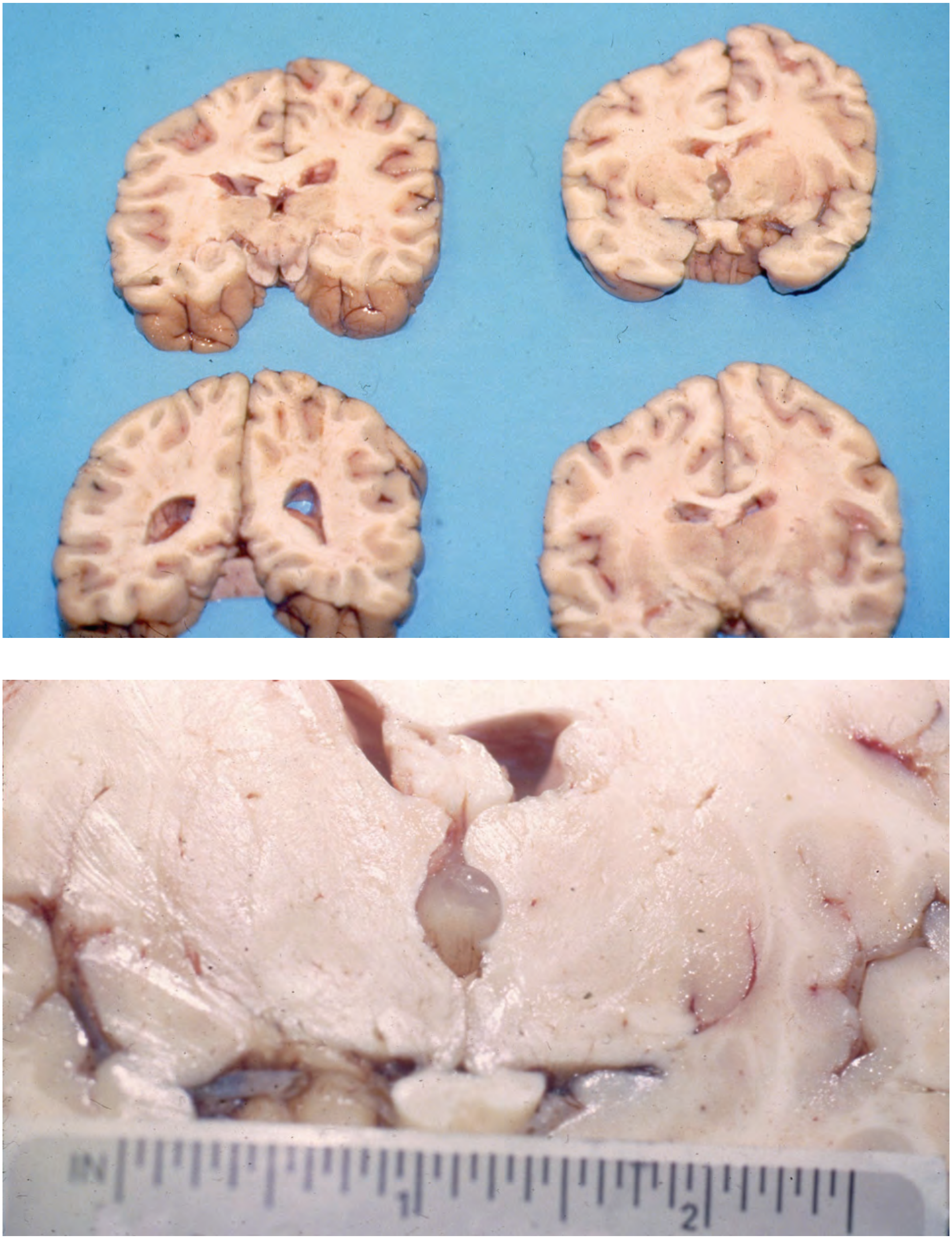

Colloid cyst of the third ventricle. This decedent had a history of severe headache with postural changes. This may be associated with sudden cardiac death following a buildup of cerebrospinal fluid pressures associated with central nervous system cardiac center disruption and fatal arrhythmia.

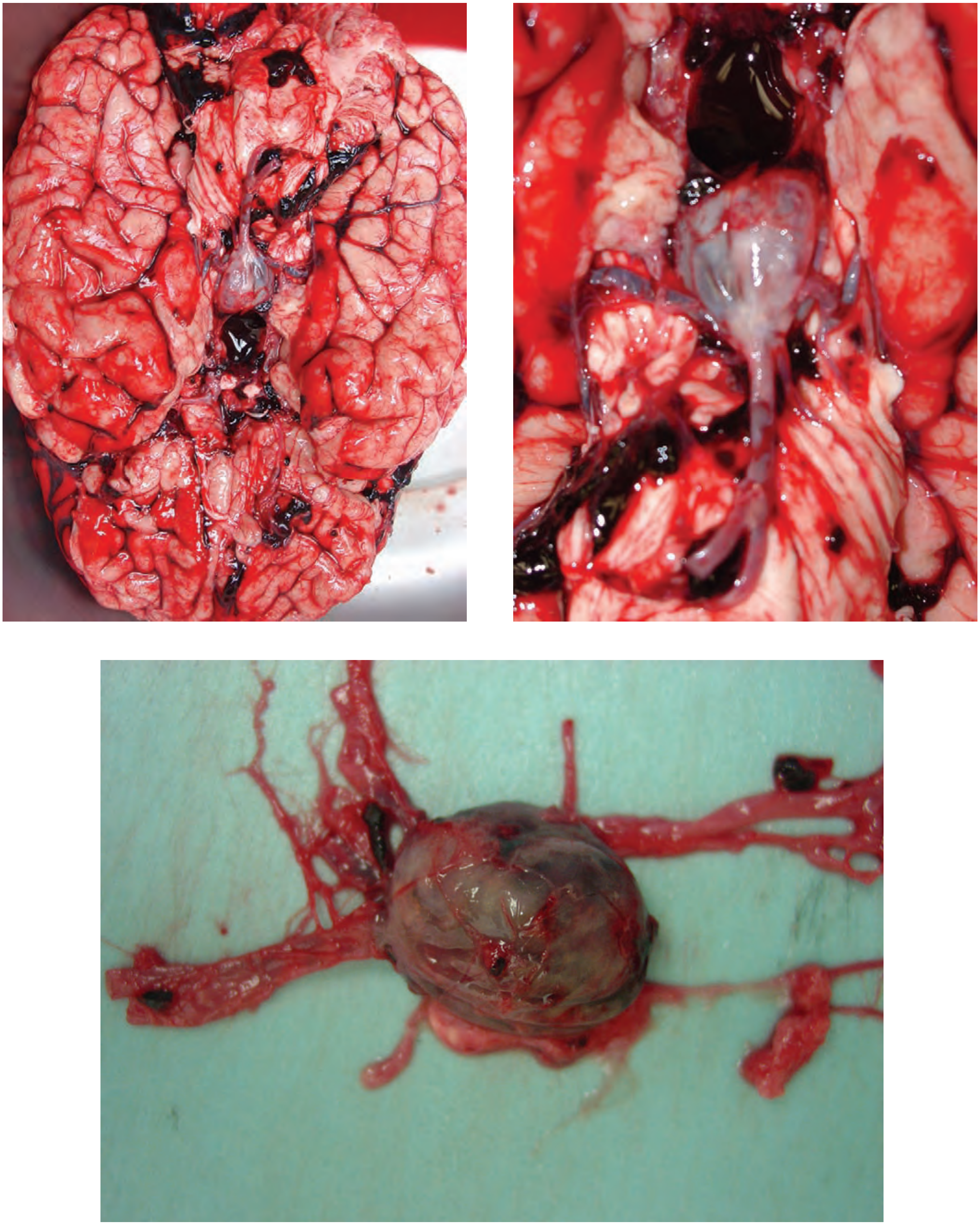

Sudden death associated with ruptured saccular cerebral artery aneurysm with subarachnoid hemorrhage. These are examples of giant berry aneurysms, which are greater than 2.5 centimeters in greatest dimension. Small saccular aneurysms may rupture the same way. The jolt of increased pressure from arterial blood through the subarachnoid space at the base of the brain may disrupt the cardiac centers, causing fatal arrthymia.

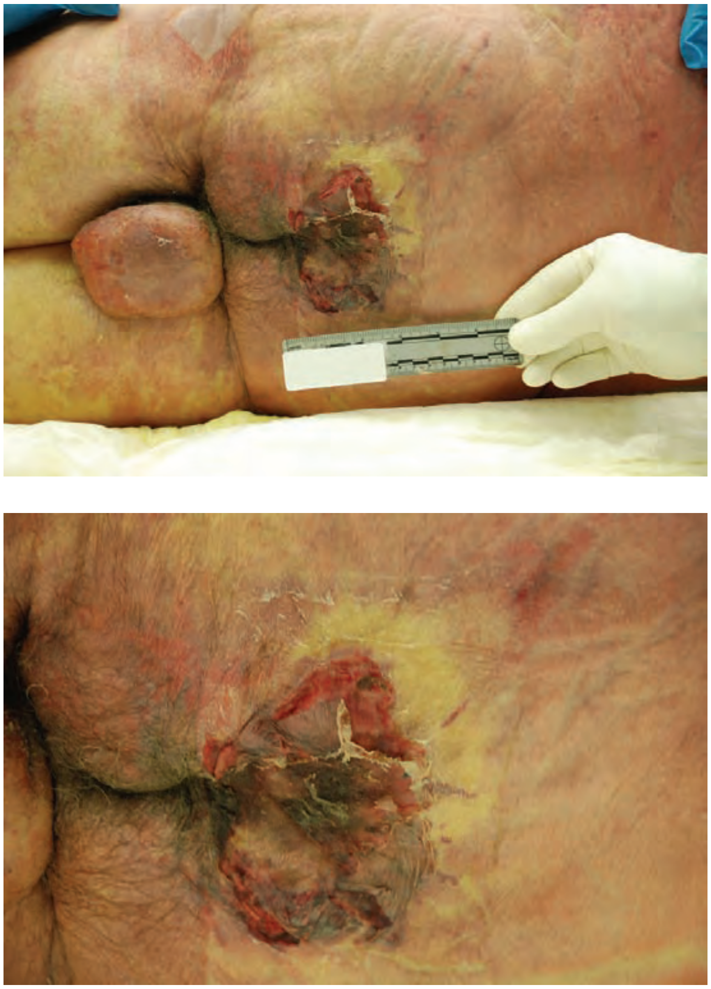

Decubital ulceration or pressure sore. This is often associated with poor patient care.

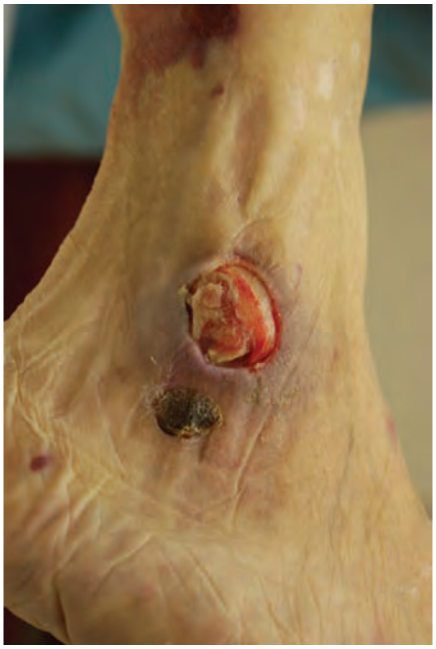

Healing ankle ulceration associated with peripheral vascular disease.

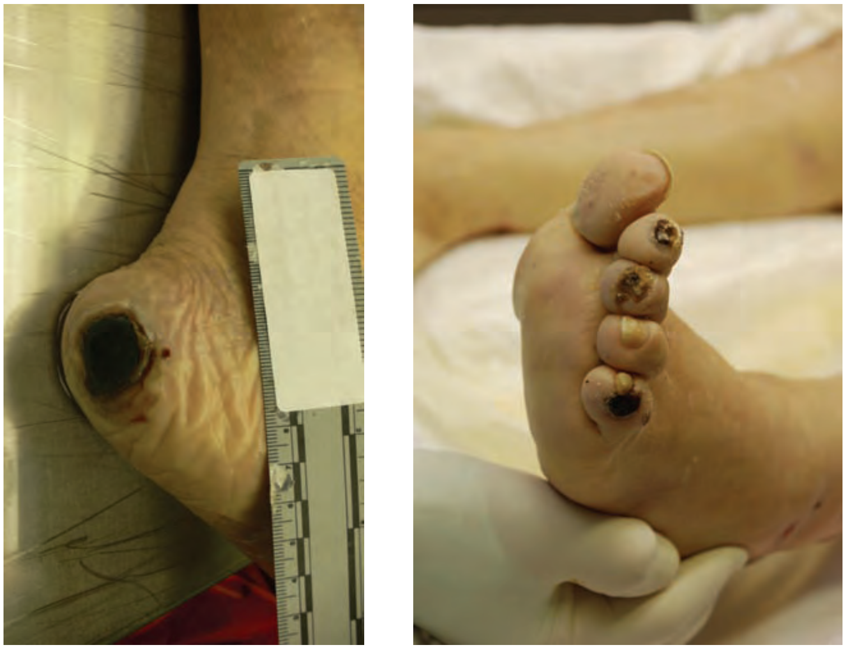

Dry gangrene associated with peripheral vascular disease due to long-term diabetes mellitus.

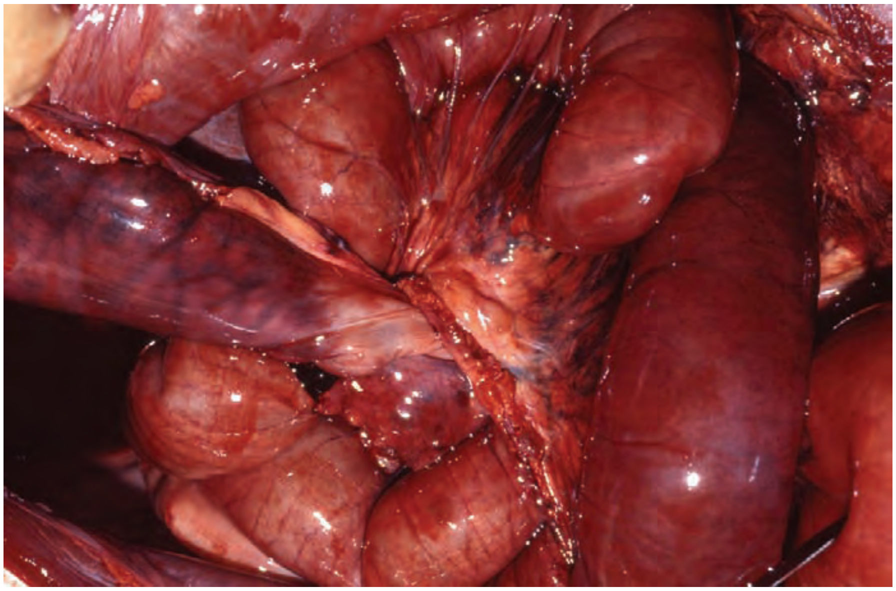

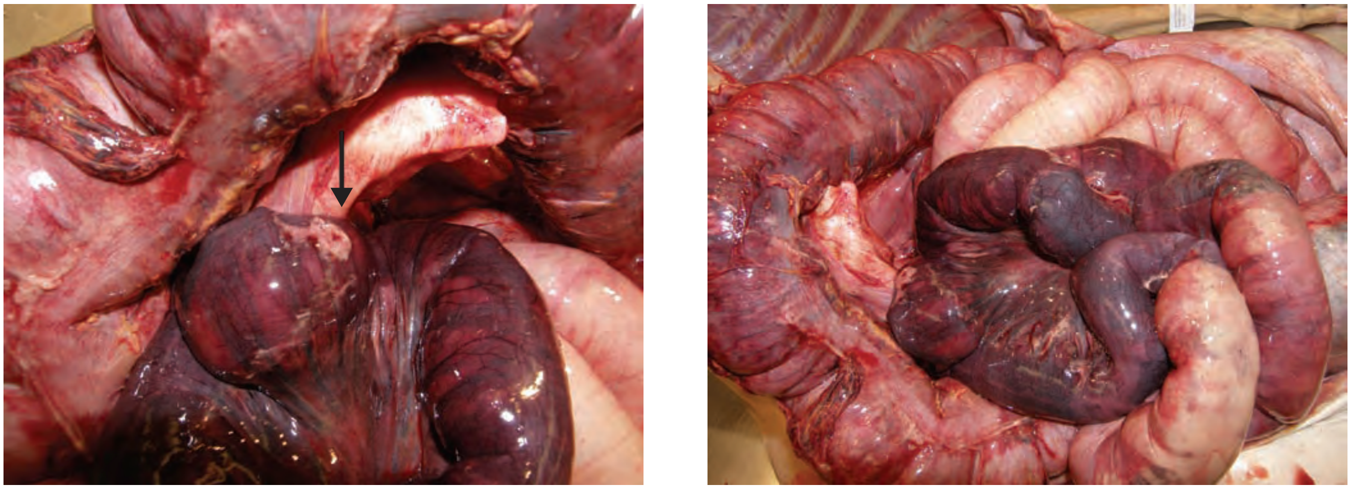

Ischemic bowel due to small intestine volvulus with vascular compromise. The arrow indicates purulent exudate at the serosal surface.

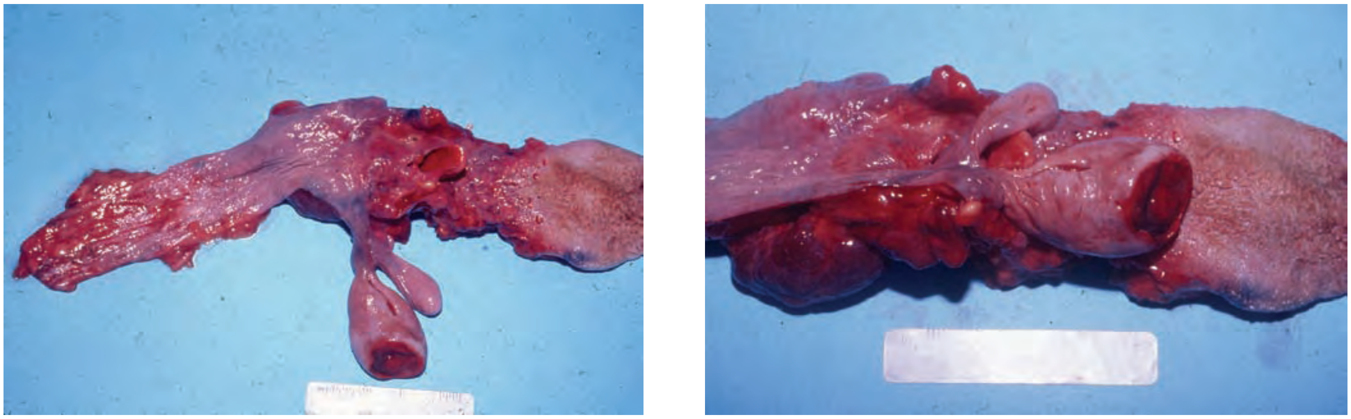

This decedent had a longstanding history of difficulty swallowing. She was seen to gesture as though she could not breathe and then collapsed. Autopsy revealed a large benign esophageal polyp that obstructed the upper airway. Note the ulceration at the tip of the polyp from constant rubbing. During this last episode she was unable to clear the obstruction.

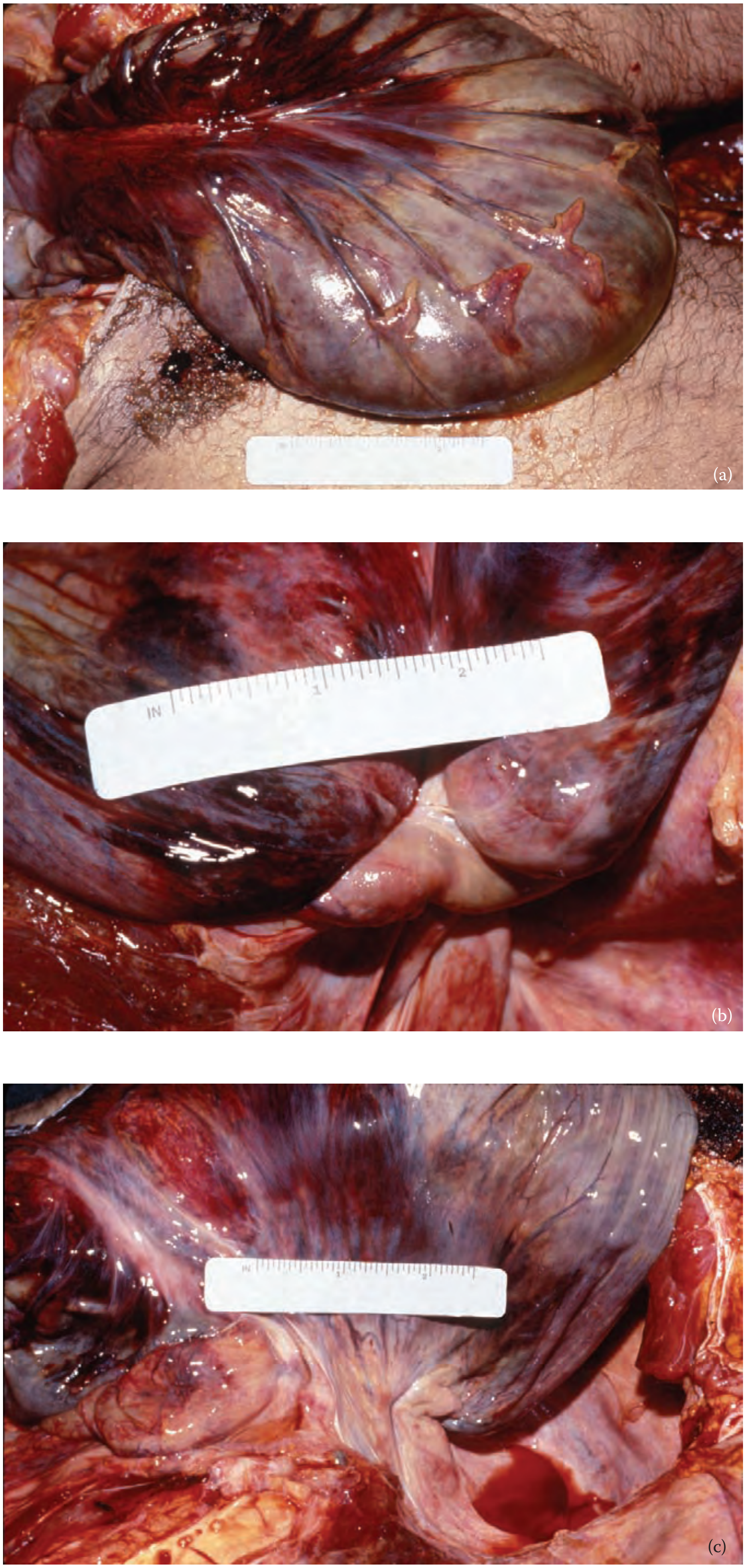

Cecal volvulus. Note the red discoloration of the serosal surface. Figure (a) demonstrates cecal enlargement due to obstruction, (b) demonstrates the cecal volvulus with rotation and obstruction, and (c) demonstrates this region untwisted.

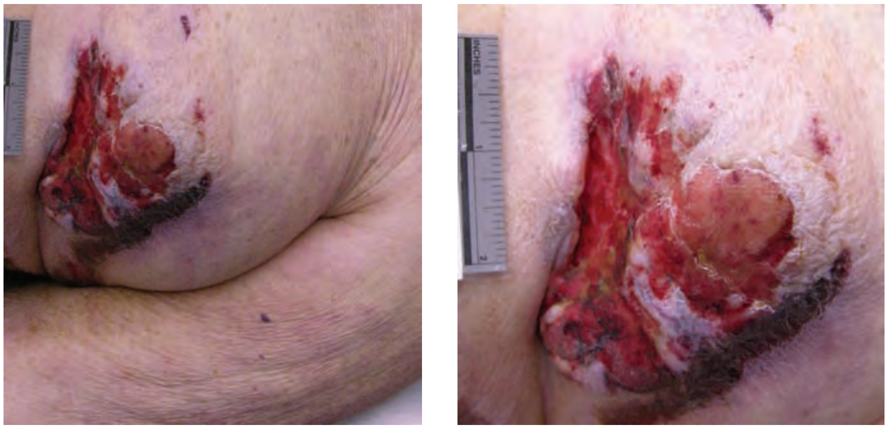

This individual never sought medical attention for her breast carcinoma and died at home.

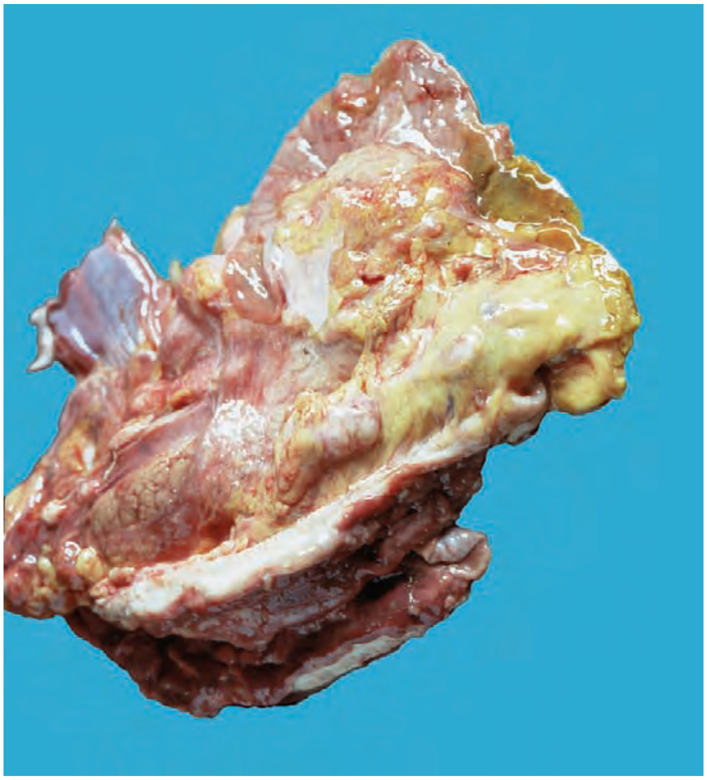

This individual never sought medical attention for his gastric cancer and died at home.

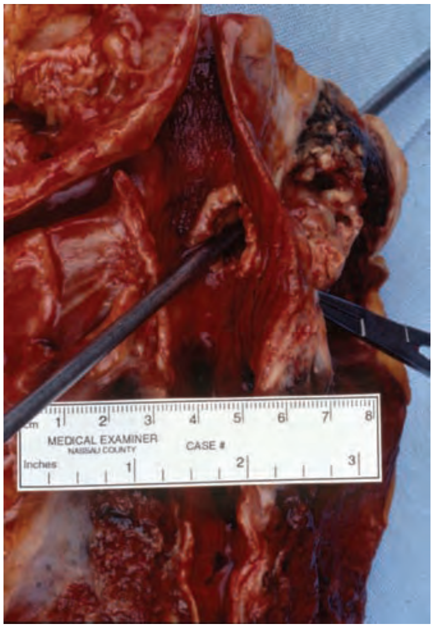
cm
MEDICAL EXAMINER
NASSAU COUNTY
CASE #
Inches

Bronchogenic carcinoma with erosion through the airway.

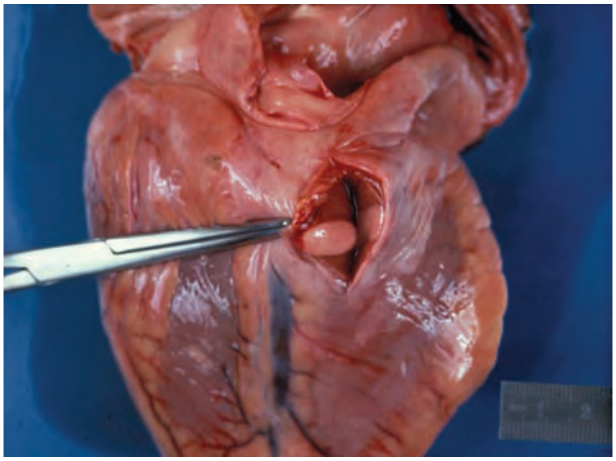

Metastatic bronchogenic carcinoma of the lung. Note the metastatic nodule in the heart muscle causing fatal arrhythmia.

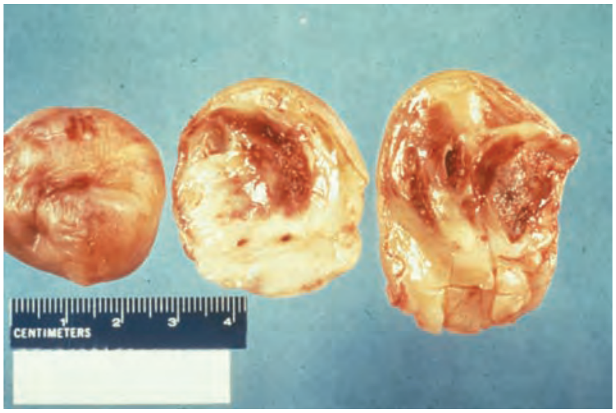
CENTIMETERS

Primary rhabdomyosarcoma of the heart.

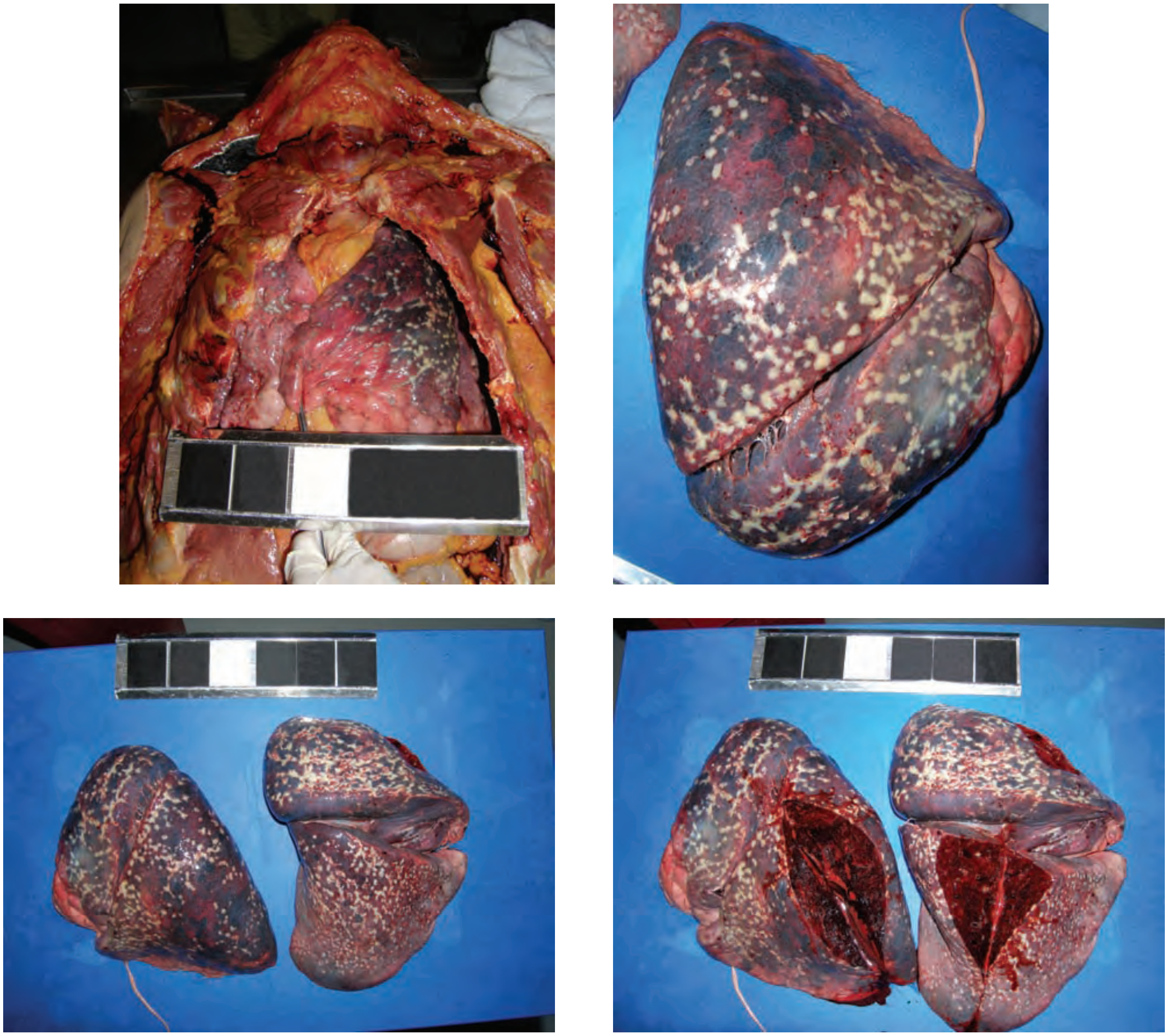

Multiple healed granulomas most likely from a past fungal infection.

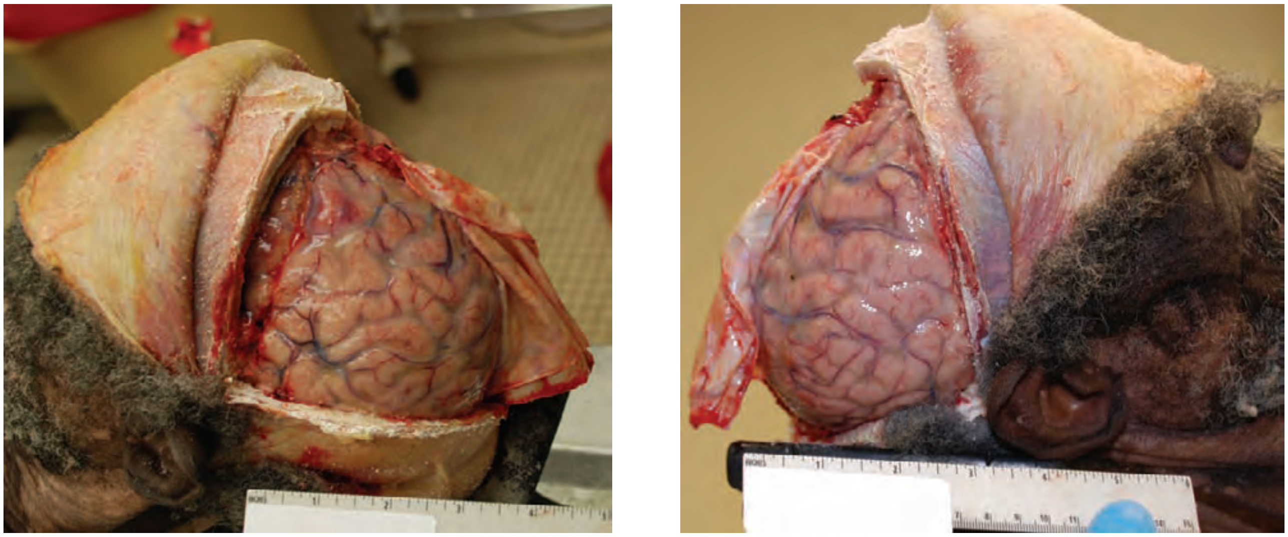

Cryptococcal meningitis in an individual with AIDS caused by intravenous drug use.

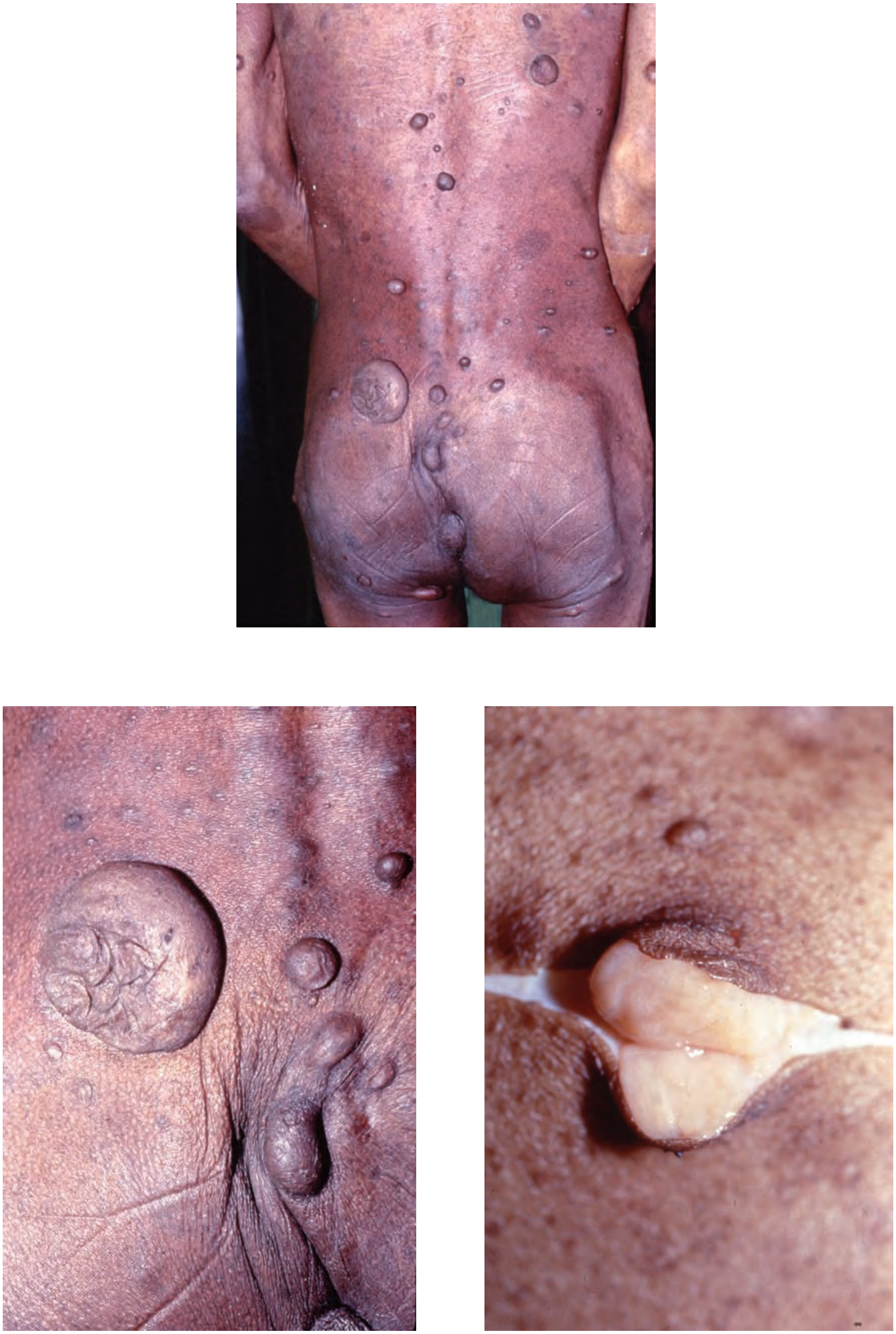

Neurofibromatosis type 1. Note the numerous neurofibromas at the surface of the body. These tumors may be found throughout the body. Death may occur by tumor growth leading to damage of adjacent structures such as the gastrointestinal track with obstruction, central nervous system with compression, renal artery with hypertension, etc.

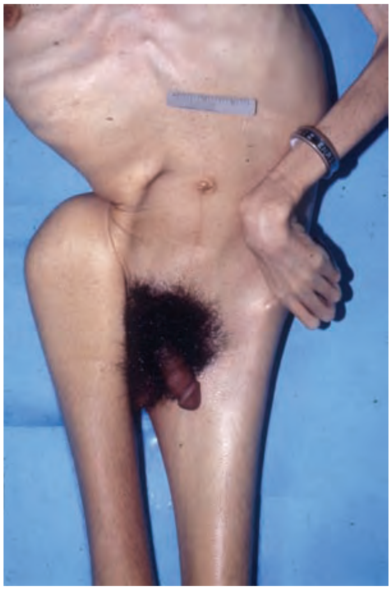

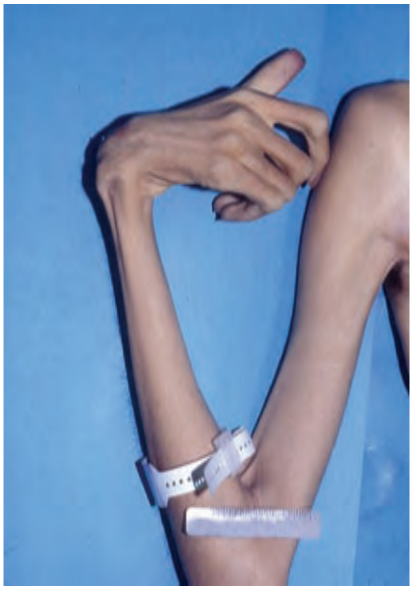

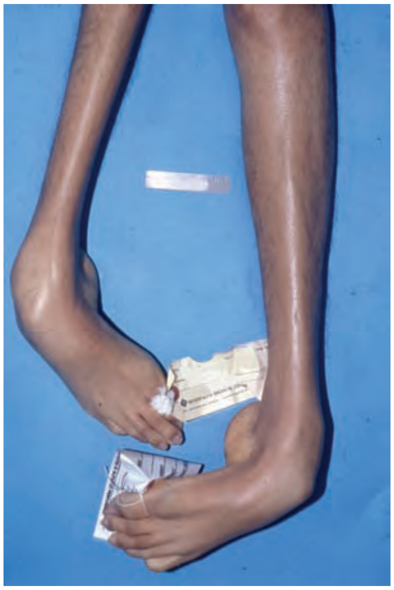

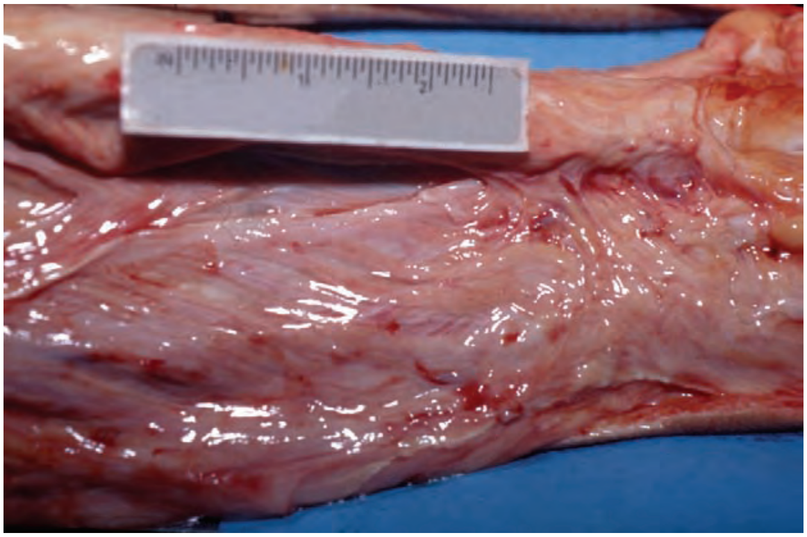

This decedent had Duchenne's muscular dystrophy and was bedridden for many years. He developed severe osteoporosis. While being removed from bed one day by a new caregiver, he sustained bilateral femur fractures and died shortly after.

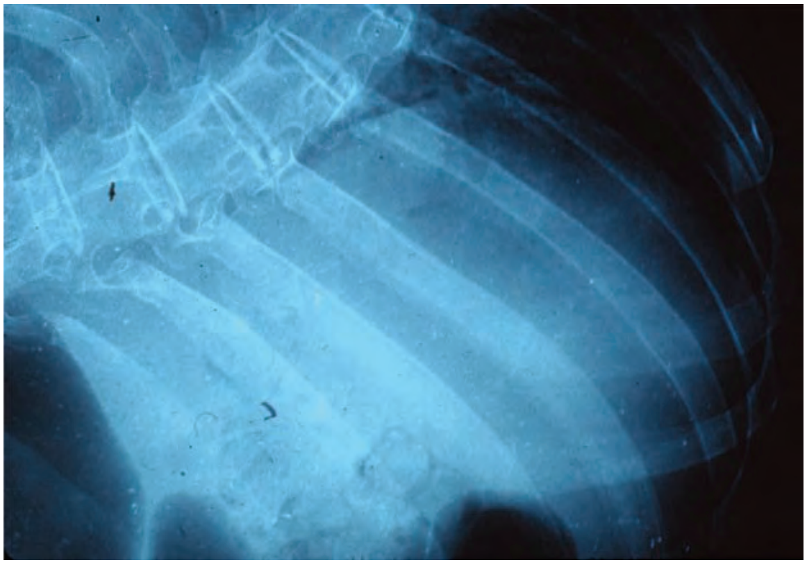

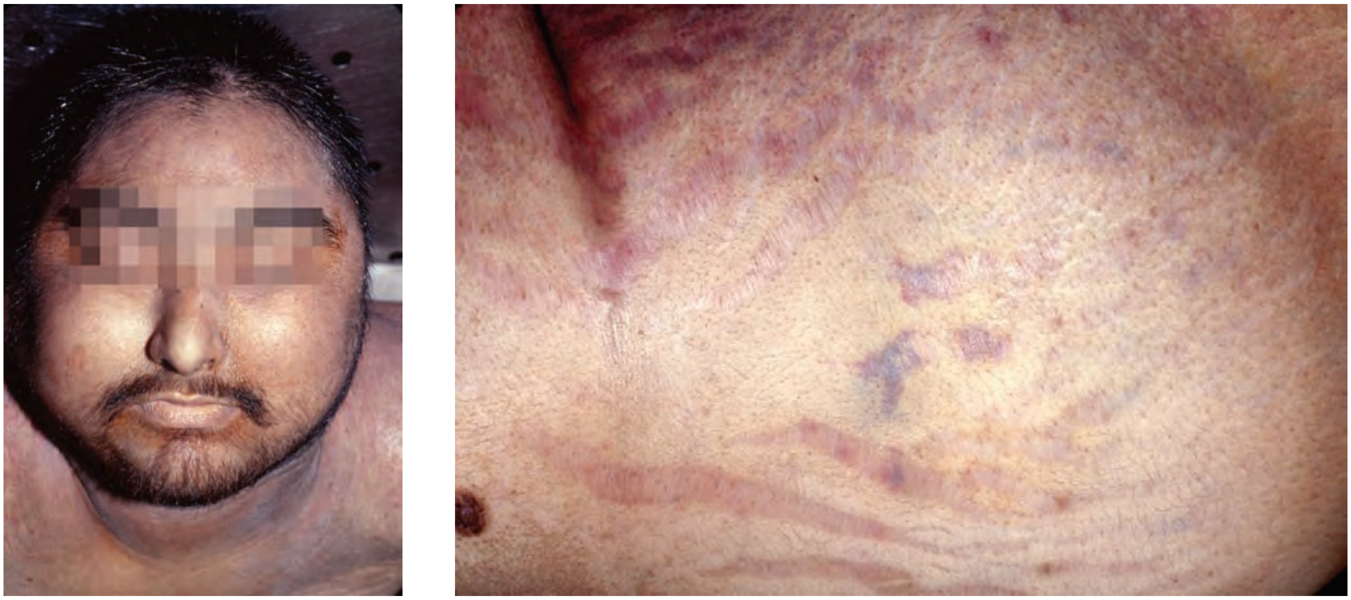

Cushing's syndrome due to hypercortisolism. Note the "moon face," "truncal obesity" and purple striates at the lower abdomen. This syndrome is also associated with glucose intolerance. From a forensic standpoint and interpreting injuries, it is important to know this is associated with osteoporosis, increased bruisability, and poor healing.

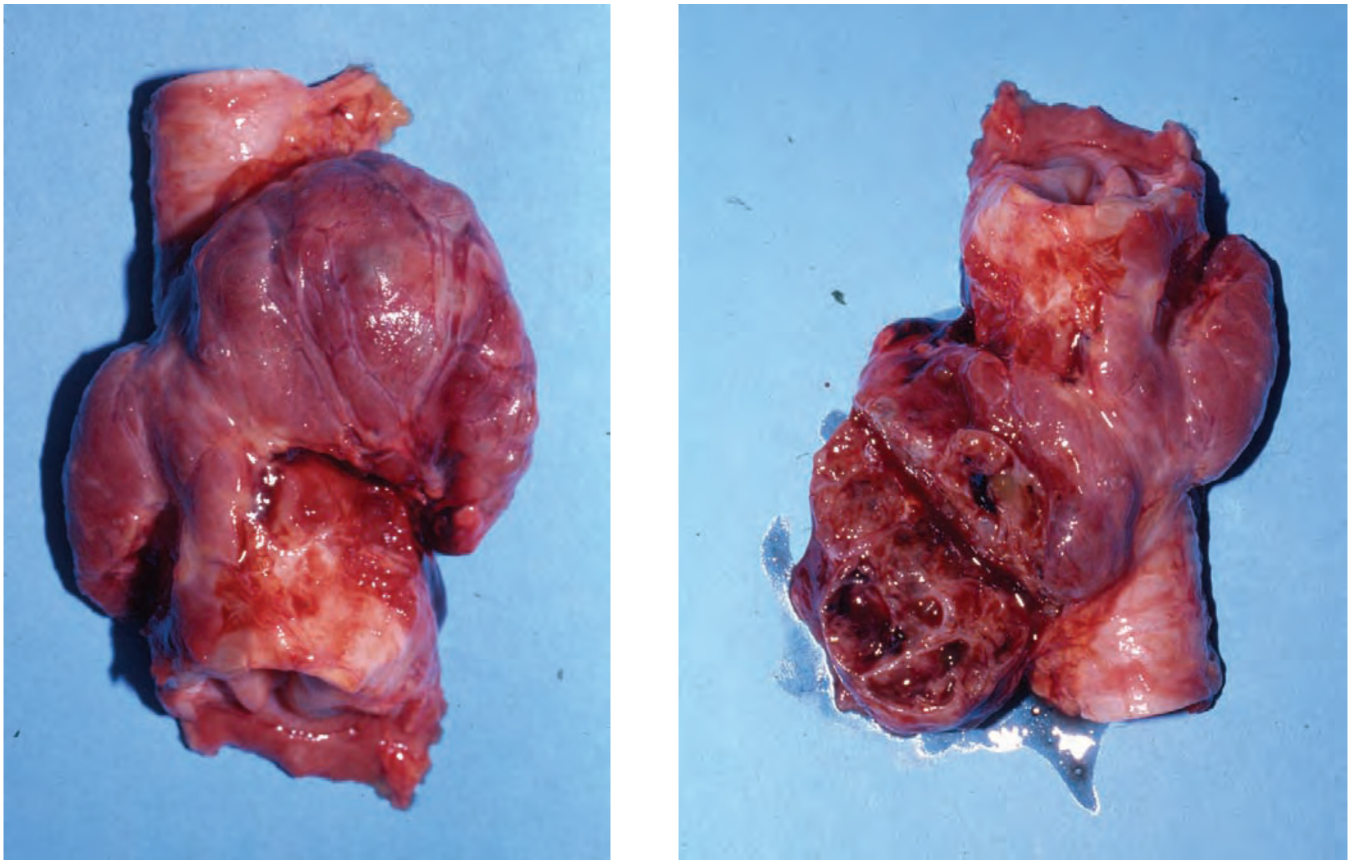

Goiter with hyperthyroidism may lead to thyrotoxicosis and sudden cardiac death due to tachyarrthymia.

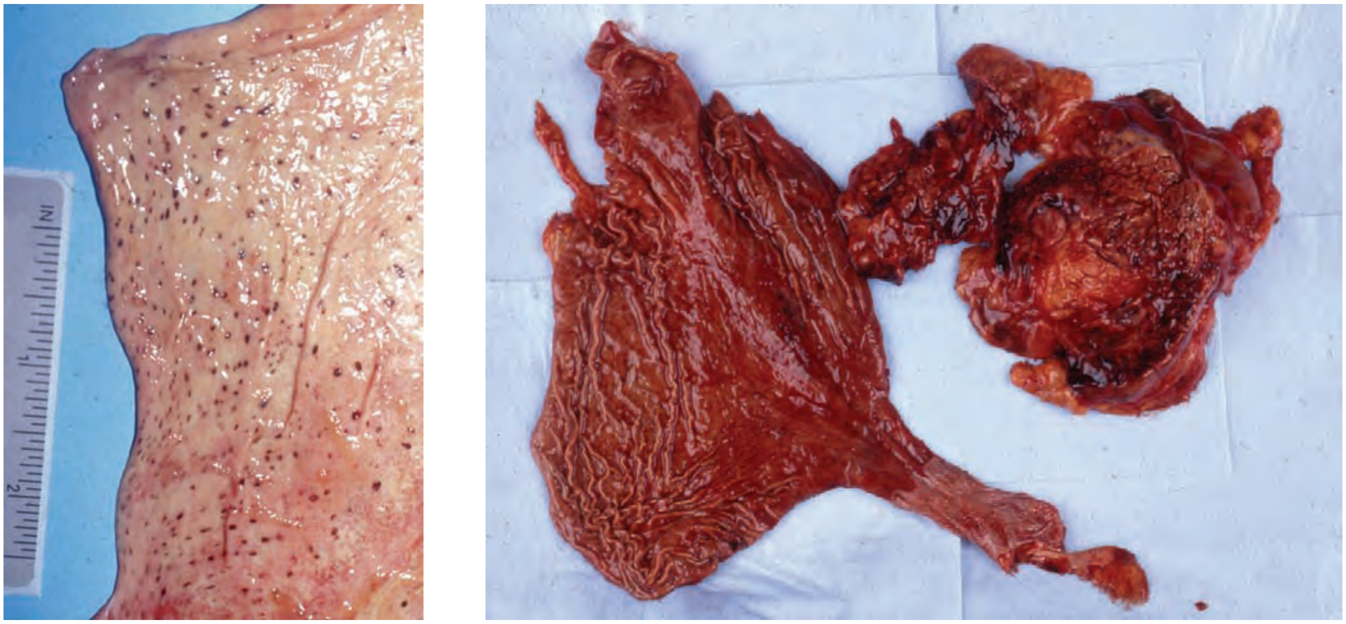

This portion of gastric mucosa reveals multiple small areas of hemorrhage in an individual with hypothermia. "Leopard skinning" of the gastric mucosa and acute pancreatitis are often seen with cases of marked hypothermia followed by short-term survival.

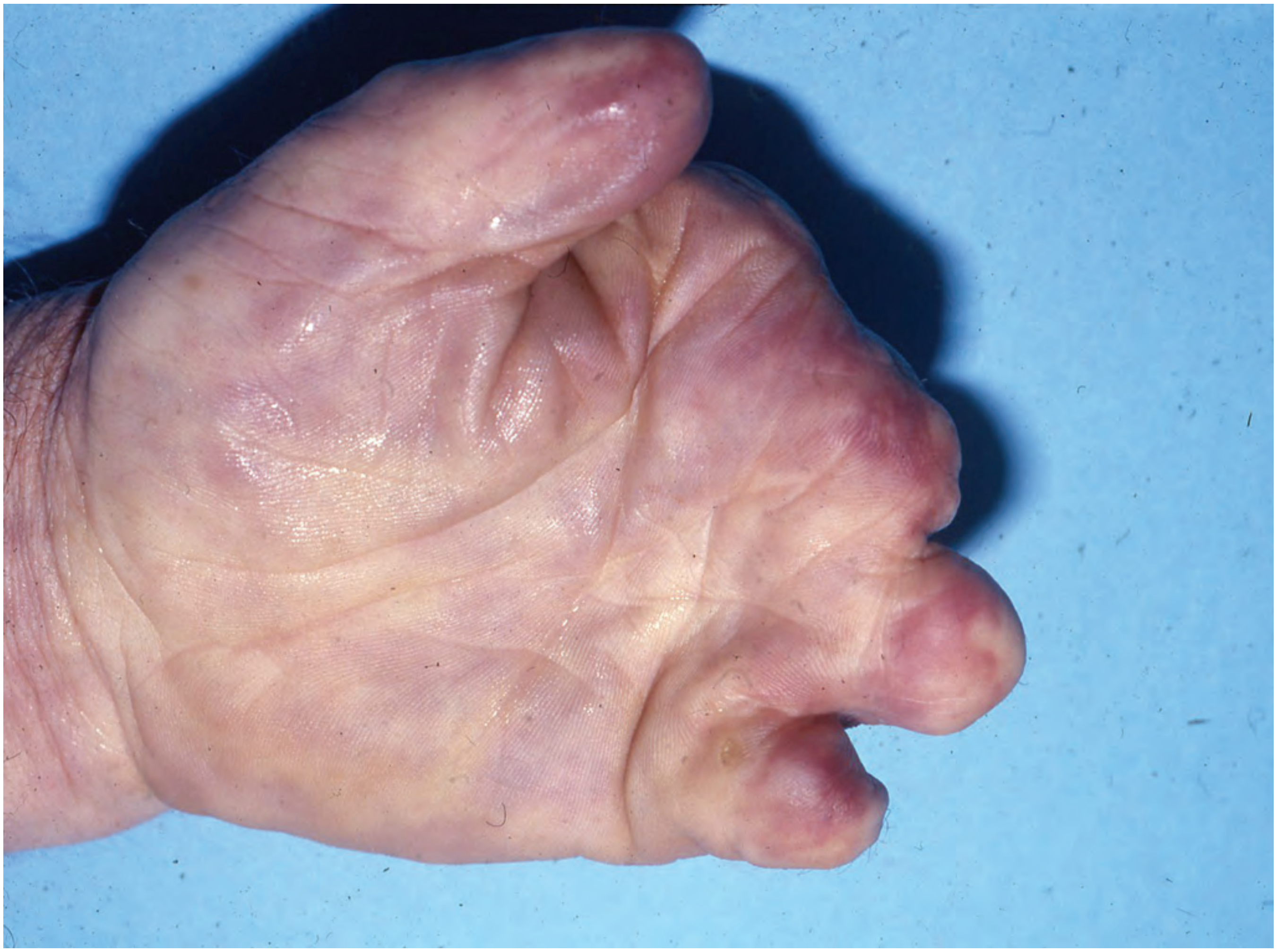

This individual had his fingers and toes amputated years before following extensive hypothermic injury with frostbite.

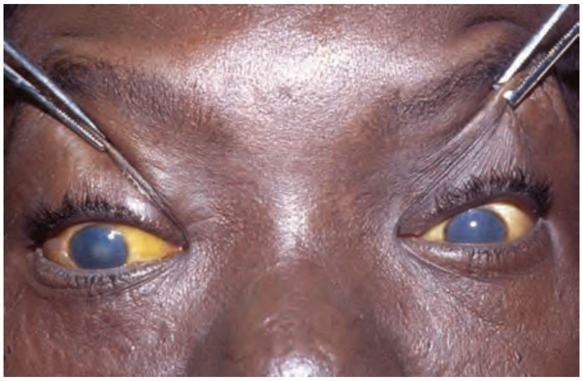

Jaundice due to hemolysis associated with sickle cell crisis.

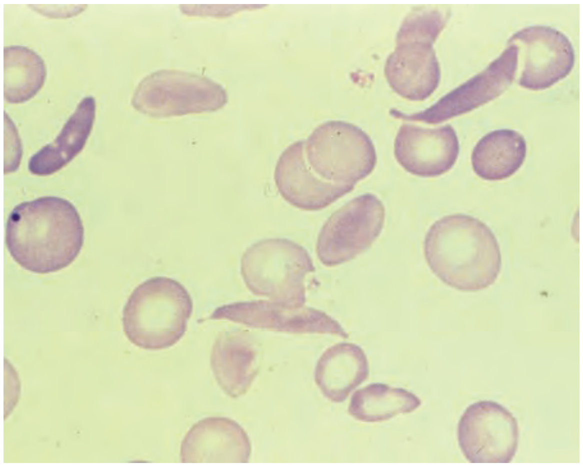

Sickle cell anemia. Peripheral blood smear showing sickled cells.

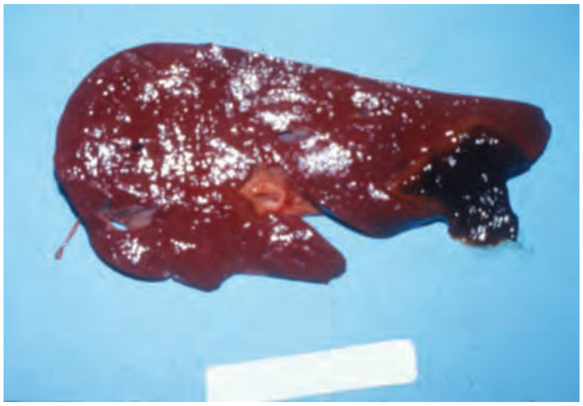

Sickle cell crisis. Note the dark red-black sludge within the gall bladder from hemolysis.

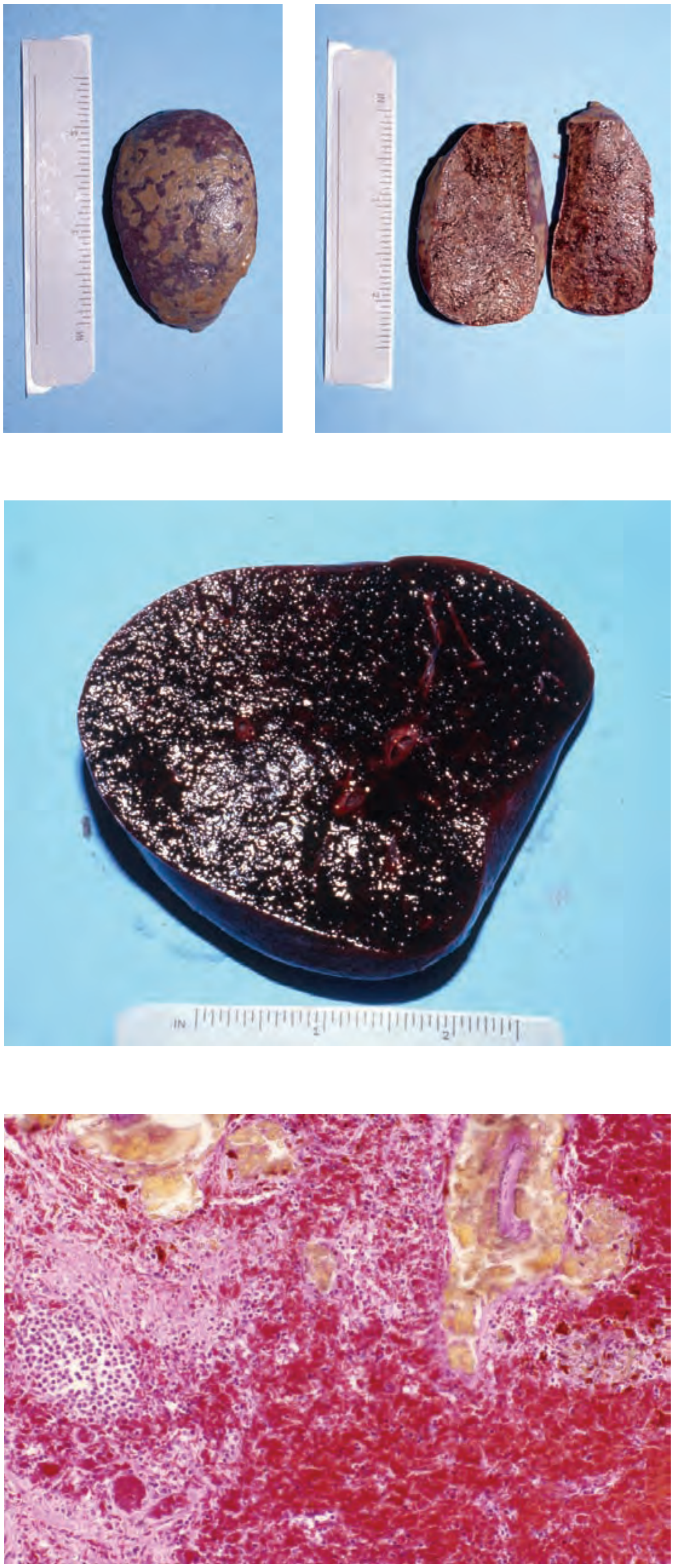

Autoinfarction of spleens from adult individuals with sickle cell anemia. These spleens are typically shrunken and firm. Note the histopathology sections with marked congestion fibrosis and Gamna-Gandy bodies from iron pigments with calcium salts.

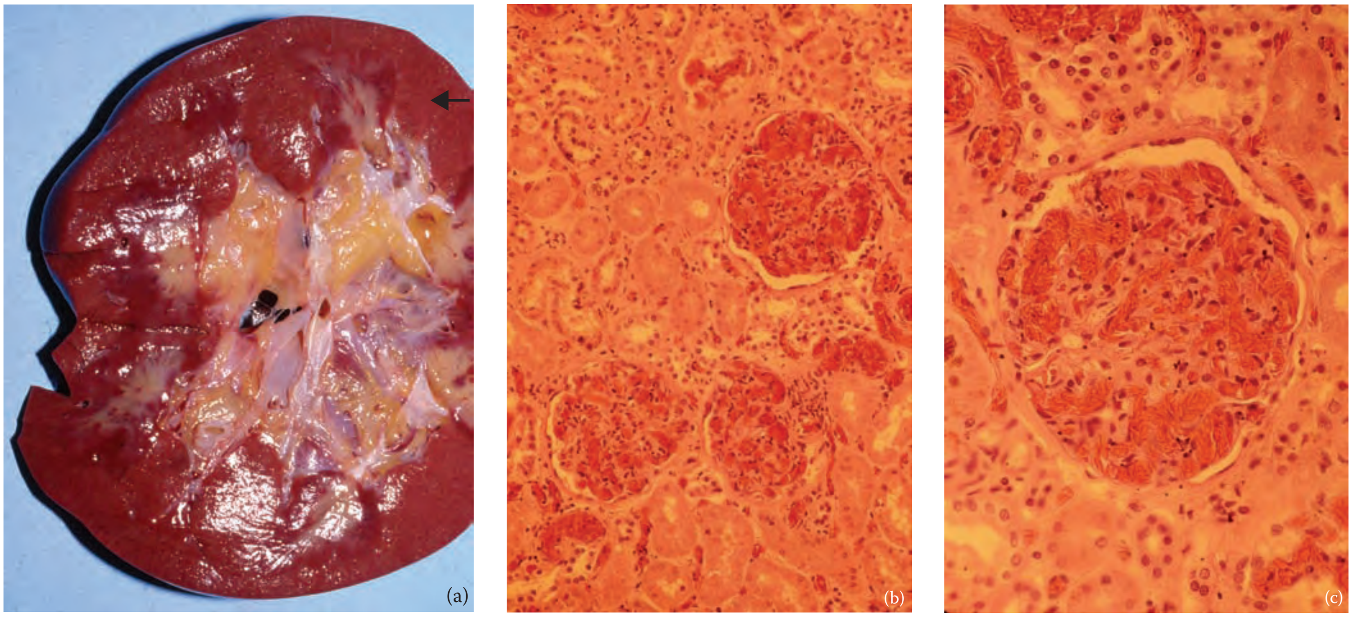

Sickle cell anemia with crisis. Kidney with papillary necrosis (a) and congestion (b) and (c).

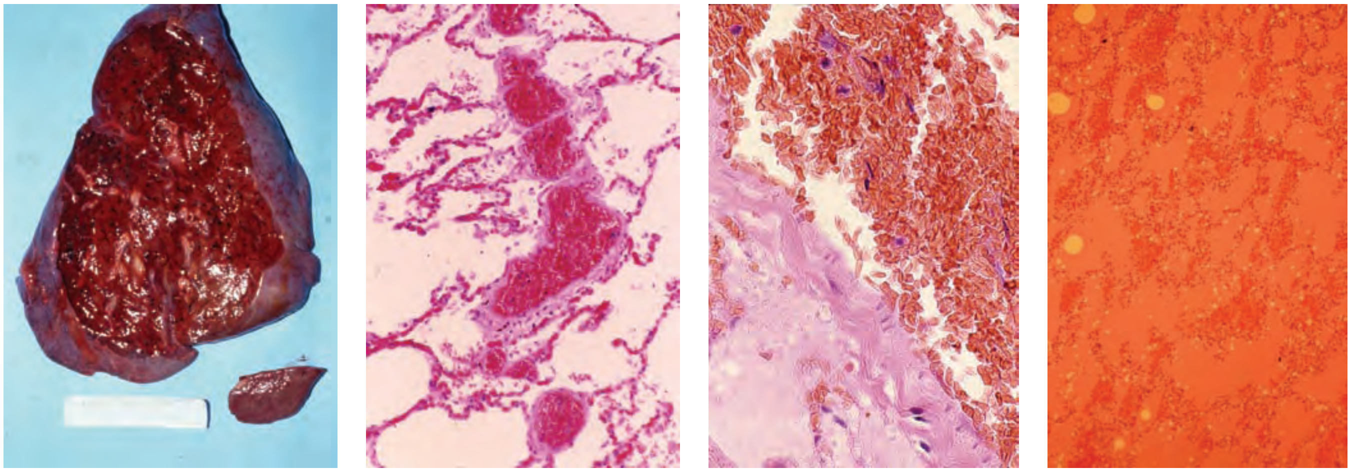

Sickle cell anemia with crisis. Lung with marked congestion and pulmonary edema.

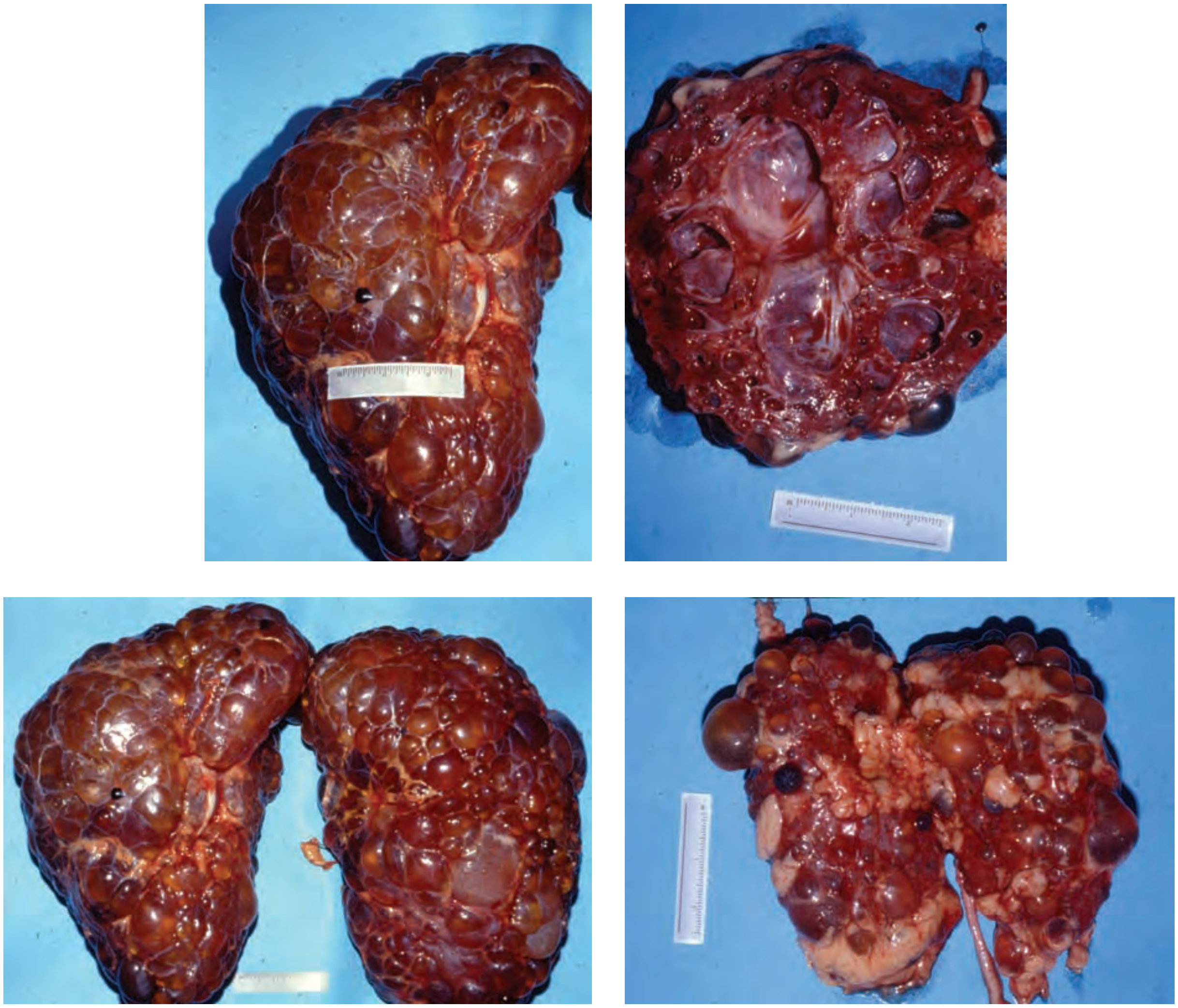

Adult polycystic kidney disease. This autosomal dominant disorder is another significant cause of hypertension. These kidneys weighed over 3000 grams each. There is also an association with cysts in other organs and intracranial berry aneurysms.

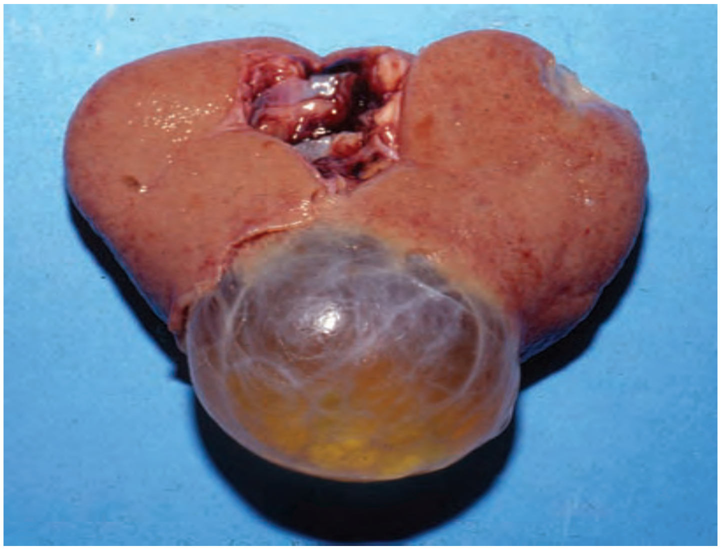

Simple benign cyst.

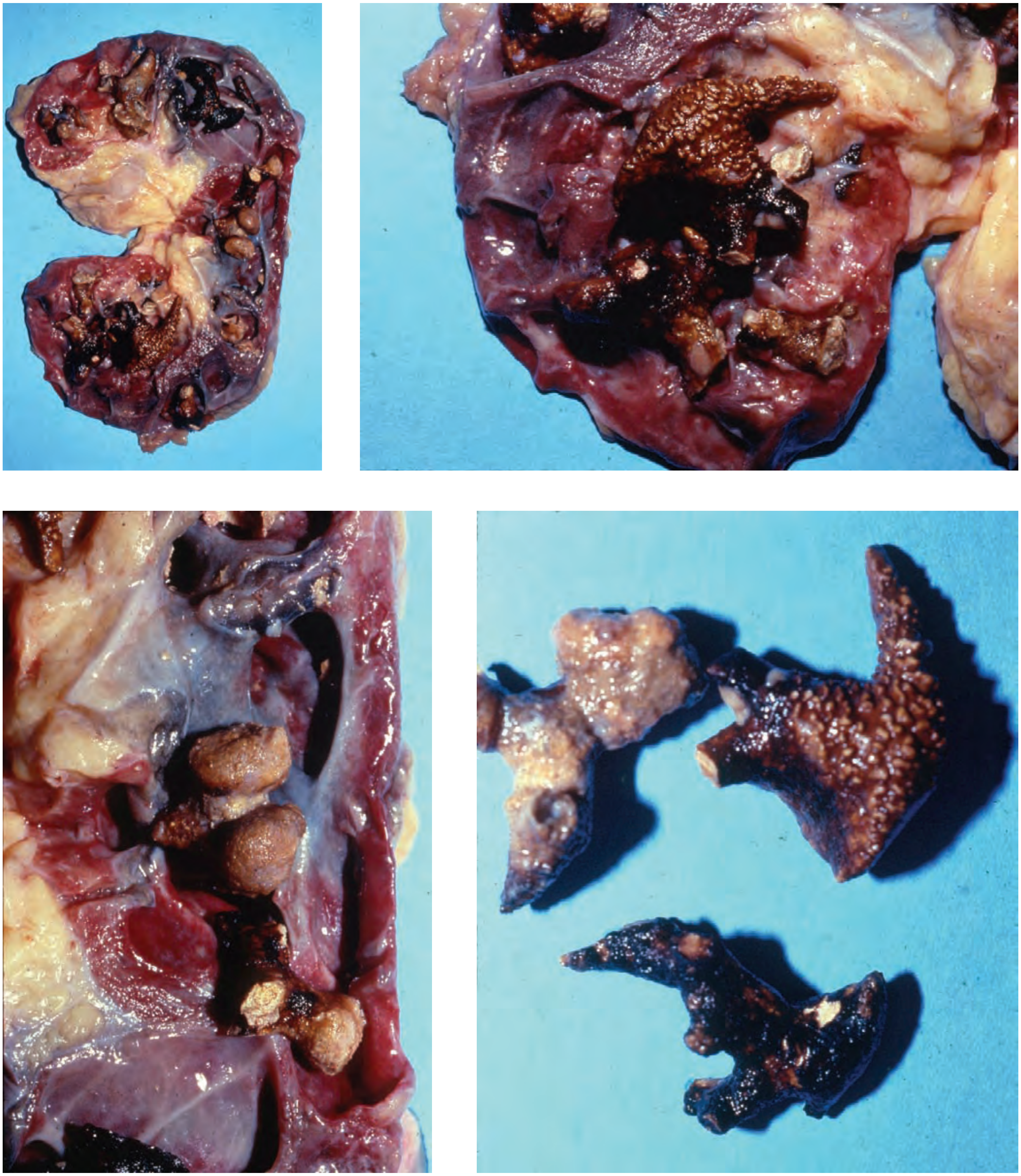

Chronic pyelonephritis with "stag horn calculi" associated with urea splitting bacterial infection such as proteus or staphylococci causing magnesium ammonium phosphate salt precipitation. It is also important to know that acute obstructive urolithiasis with urosepsis can be rapidly fatal.

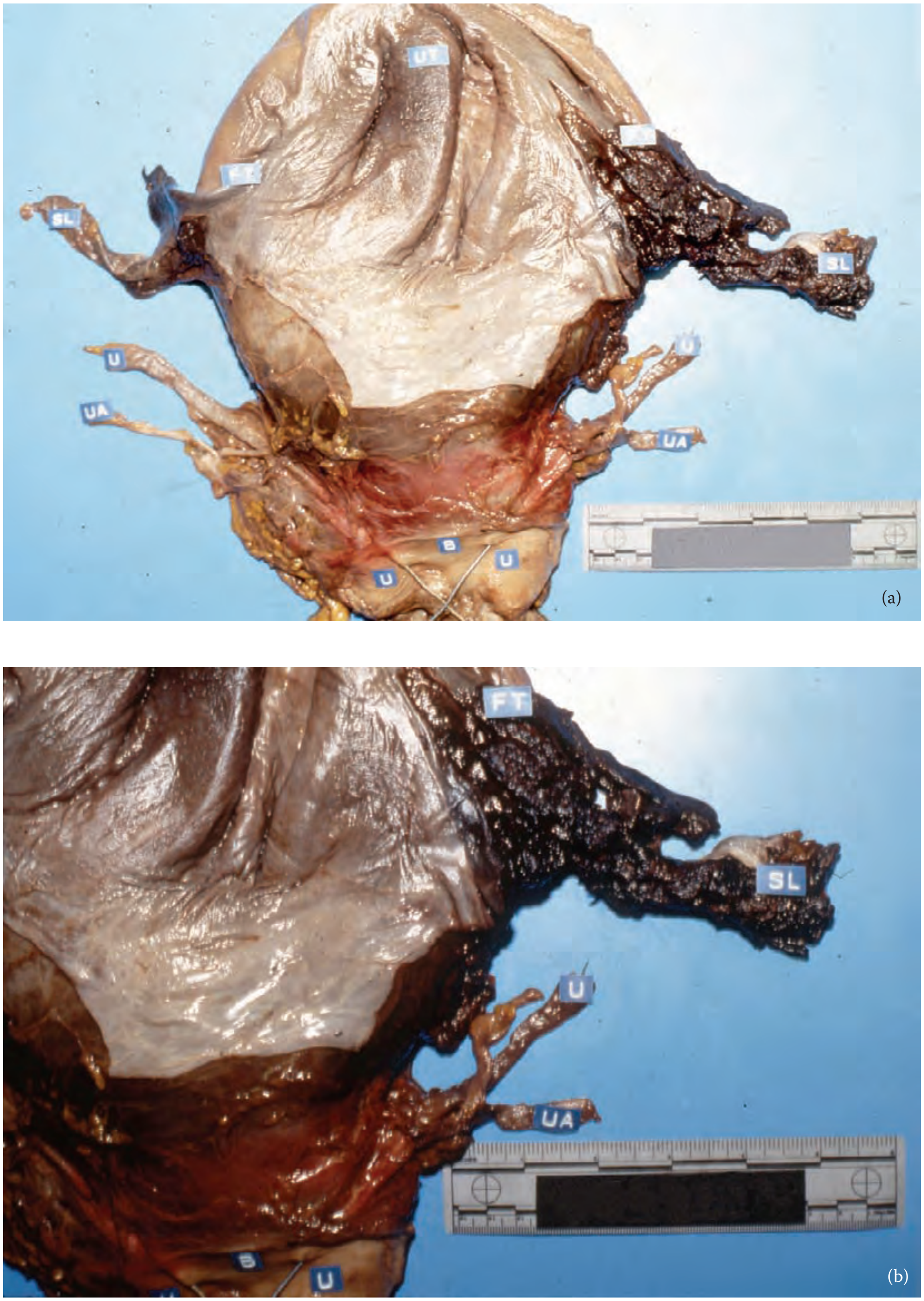

This is a postpartum death involving a 30-year-old nurse. This woman was known to be colonized by group A beta-hemolytic streptococcus (strepococcus pyogenes). Approximately 8 hours following an uneventful delivery, she began to complain of back and pelvic pain. Within 4 hours she was in full-blown shock and was refractory to resuscitative measures. Death was pronounced approximately 14 hours following her delivery. Figure (a) depicts the gravid uterus and confluent brown-black discoloration of the left adnexa, which proved microscopically to show areas of necrosis with numerous clusters of bacterial cocci and scant neutrophilic infiltrates. Postmortem cultures of multiple organs, including lungs, liver, spleen, uterus, and peritoneal fluid, all grew group A beta-hemolytic streptococcus betahemolitic streptococcus. This "toxic shock like" death due to group A streptococcus has given rise to the term "flesh-eating bacteria."

Therapeutic Interventions, Complications, and Accidents

2

MICHAEL J. CAPLAN AND CHARLES A. CATANESE

The evaluation of deaths related to complications of diagnostic and therapeutic procedures has traditionally been a problematic area for forensic pathologists and death investigators. The reasons for this are multiple: (1) these deaths often blur the boundaries dividing medicolegal and hospital death investigation; (2) many forensic pathologists do not feel comfortable approaching the often complex and highly technical issues present today in modern medicine; and (3) at least some forensic pathologists believe that these types of death belong more appropriately within the domain of the hospital autopsy. Despite this ongoing controversy, most forensic pathologists and death investigators will, at some point during their careers, be forced to handle these types of deaths.

If a patient dies either during or sometime following a diagnostic and therapeutic procedure, the first question in the diagnostic algorithm or decision tree becomes: Was the death in any way related to the procedure or was the procedure simply an artifact, with the death attributable to some underlying natural disease process or injury? For example, if a 65-year-old man who has stubbornly refused surgical intervention of a known 8-cm-diameter abdominal aortic aneurysm suddenly collapses and manages to make it to the operating room but dies before repair can be attempted, the "intraoperative death" is simply an artifact of modern medical resuscitation and transport technologies. In such a case, the medical intervention or procedure played no part in the death, and the cause of death would be worded as "ruptured abdominal aortic aneurysm due to atherosclerotic cardiovascular disease, manner: natural." On the other hand, if an otherwise healthy individual underwent an elective procedure, for example, a bunionectomy, and died on the operating table, to ascribe the death to "hallux valgus" would be as ridiculous as it was erroneous. It is for cases like this that the concept of therapeutic complication merits consideration for a place in manner-of-death certification.

Once it is determined that the death was related somehow to the procedure, the next question in the algorithm becomes: Was the death a result of a known or predictable (albeit rare or unusual) complication of a properly performed diagnostic procedure or appropriately administered therapy, or did it ensue from a procedural error? When the former scenario exists, the term *therapeutic complication* is preferred, because that is a neutral designation as opposed to the term "therapeutic misadventure," which is inflammatory and prejudicial. When the latter situation is present, the manner of death is most accurately designated *accident.* Many examples of legitimate therapeutic complications exist. In the case of a 64-year-old man who sustained a non-ST (non-Q wave) segment elevation myocardial infarct, and who was treated with the antiplatelet drug eptifibatide (Integrilin), a platelet glycoprotein IIb-IIIa antagonist, he suffered intractable pulmonary hemorrhage and ultimately succumbed to respiratory failure. Pulmonary hemorrhage has been a recently reported complication of eptifibatide therapy. No error was committed in this case; rather, the decedent experienced an unusual yet known (documented) complication of appropriate therapy for his non-ST elevation MI. Therefore, the cause of death would be worded as: "Complications of pulmonary hemorrhage following eptifibatide (Integrilin) treatment for non-ST (non-Q wave) segment elevation myocardial infarct (due to atherosclerotic heart disease)" and the manner, therapeutic complication. On the other hand, if a middle-aged woman underwent a surveillance colonoscopy and presented days later with peritoneal signs and was found to have *E. coli* bacteremia and sepsis due to peritonitis resulting from inadvertent perforation of the colon during the procedure, the death certificate would read: "*E. coli* sepsis due to peritonitis due to colonic perforation complicating surveillance colonoscopy." However, because this complication is not an accepted sequela of an appropriately performed diagnostic or therapeutic procedure, but rather, an unintentional error, the manner of death in this case is most accurately deemed accident. Admittedly, there are cases that fall within the murky zone between natural and therapeutic complication and between therapeutic complication and accident. In such cases, all that one can do is exercise one's best objective clinical judgment. It is important to recognize and remember that death certificates are not immutable documents that are "written in stone"; rather, they can be amended should more accurate information regarding the circumstances of death become available at a later time.

While the concept of therapeutic complication is an invaluable tool in the assessment of these types of deaths, the practical applicability of the term to the death certificate has enjoyed less success. Only two known jurisdictions—Cuyahoga County, Ohio, and New York City—include “therapeutic complication” as a choice in the manner of death on the death certificate. Therefore, when a death fits the criteria for therapeutic complication as the manner, but the particular jurisdiction does not include it on the death certificate, then the manner defaults to natural. One other option is to list “therapeutic complication” in parentheses after “natural,” but this probably is not common practice. Thus, for the indefinite future, forensic pathologists will be resigned to using the term as a conceptual tool in the evaluation of these most challenging deaths.

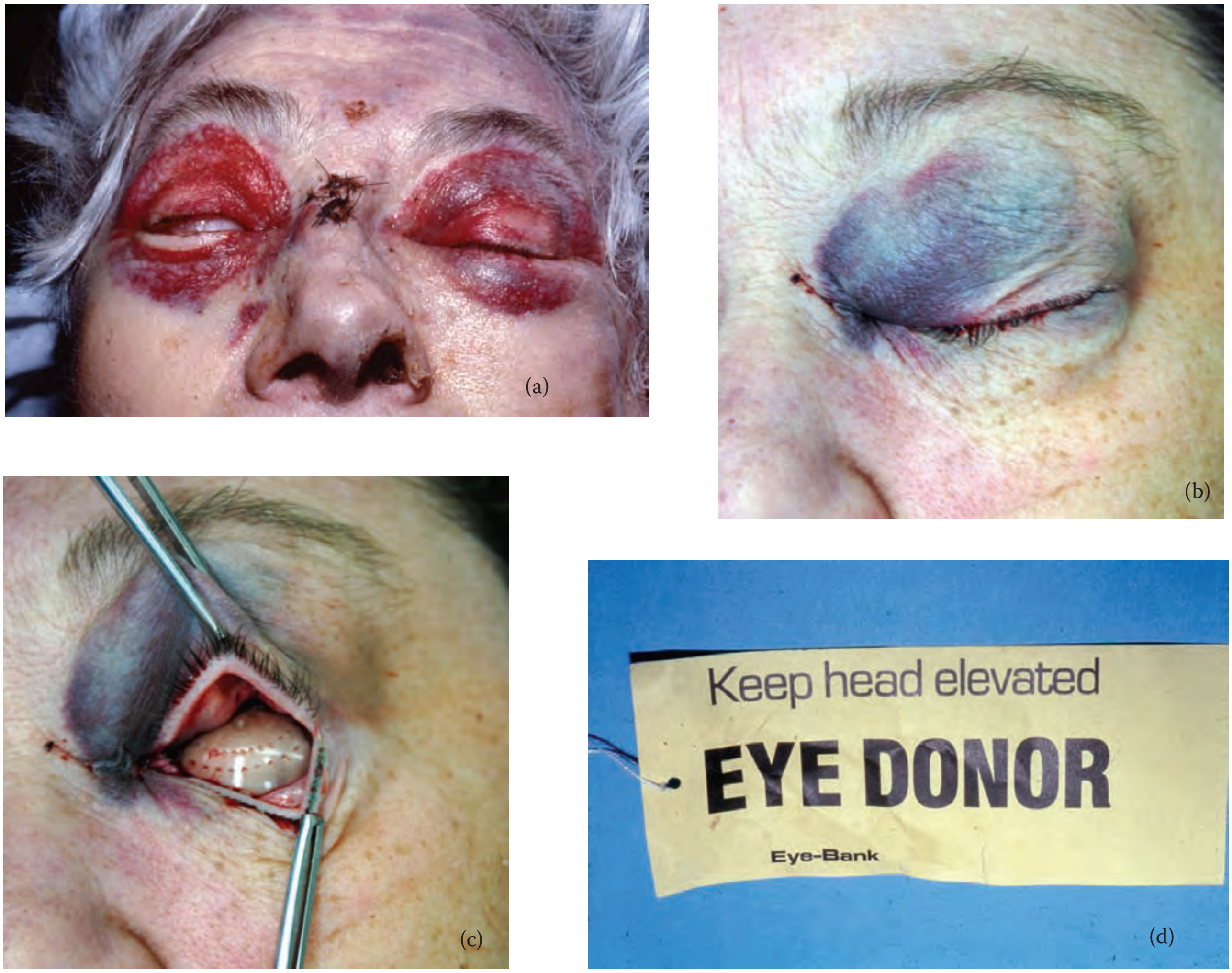

These figures show one of many possible artifacts associated with organ donation that may be mistaken for injuries. Note (a) the hemorrhagic discoloration around each eye (periorbital) and (c) the plastic insert in the opened eye following tissue removal. Medical record review and clinician interview revealed that these periorbital hemorrhages were not present at the time of admission.

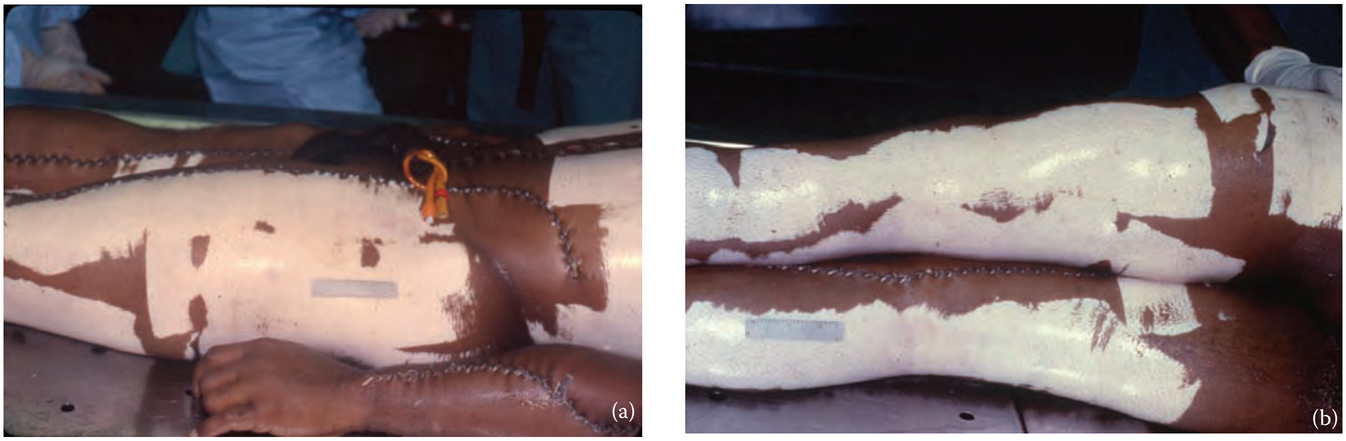

Tissue procurement of superficial skin.

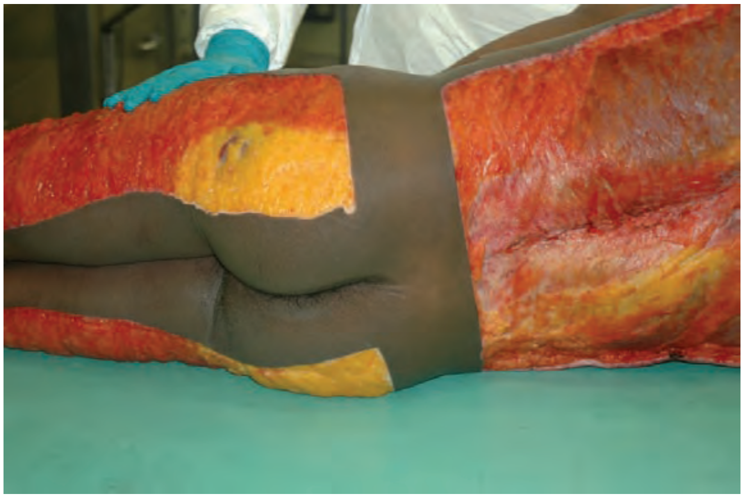

Postmortem tissue procurement and full thickness skin.

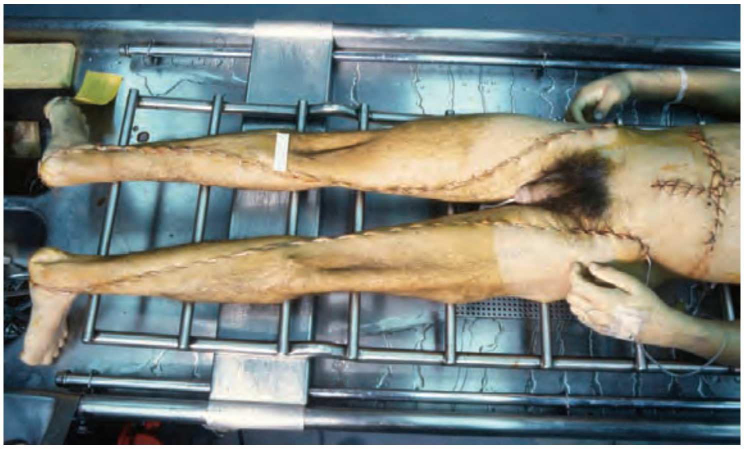

Sutured incisions following long bone, soft tissue, and saphenous veins removal for postmortem tissue procurement.

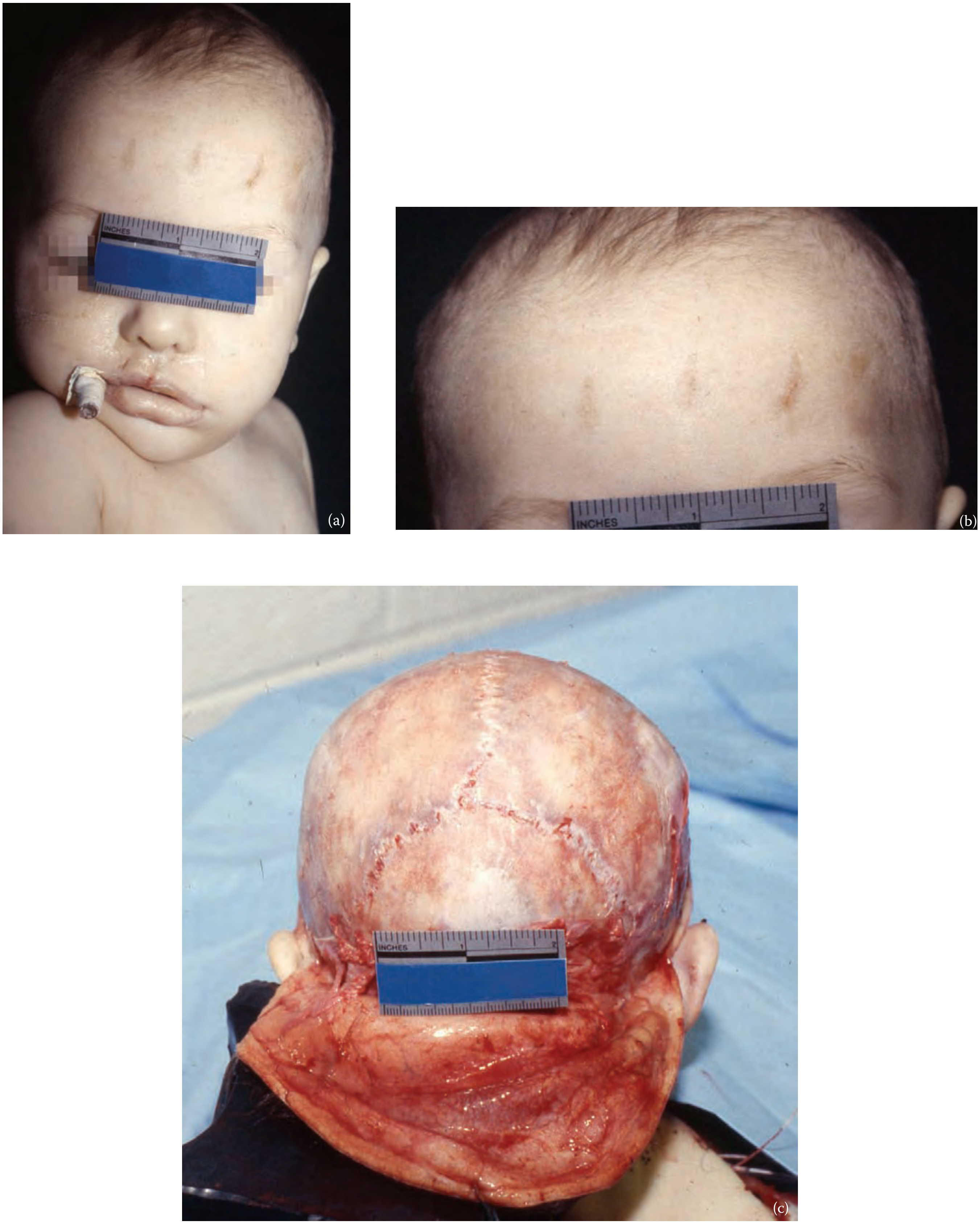

Figures (a) and (b) show ill-defined, parallel, vertically oriented gray-brown marks on the forehead. These are artifacts of transport. The child's head was stabilized with a strap while flown by helicopter from an outside hospital to a tertiary pediatric center. Figure (c) shows reflected scalp and well-defined sagittal and lamboidal sutures without features of subgaleal contusions or skull fractures.

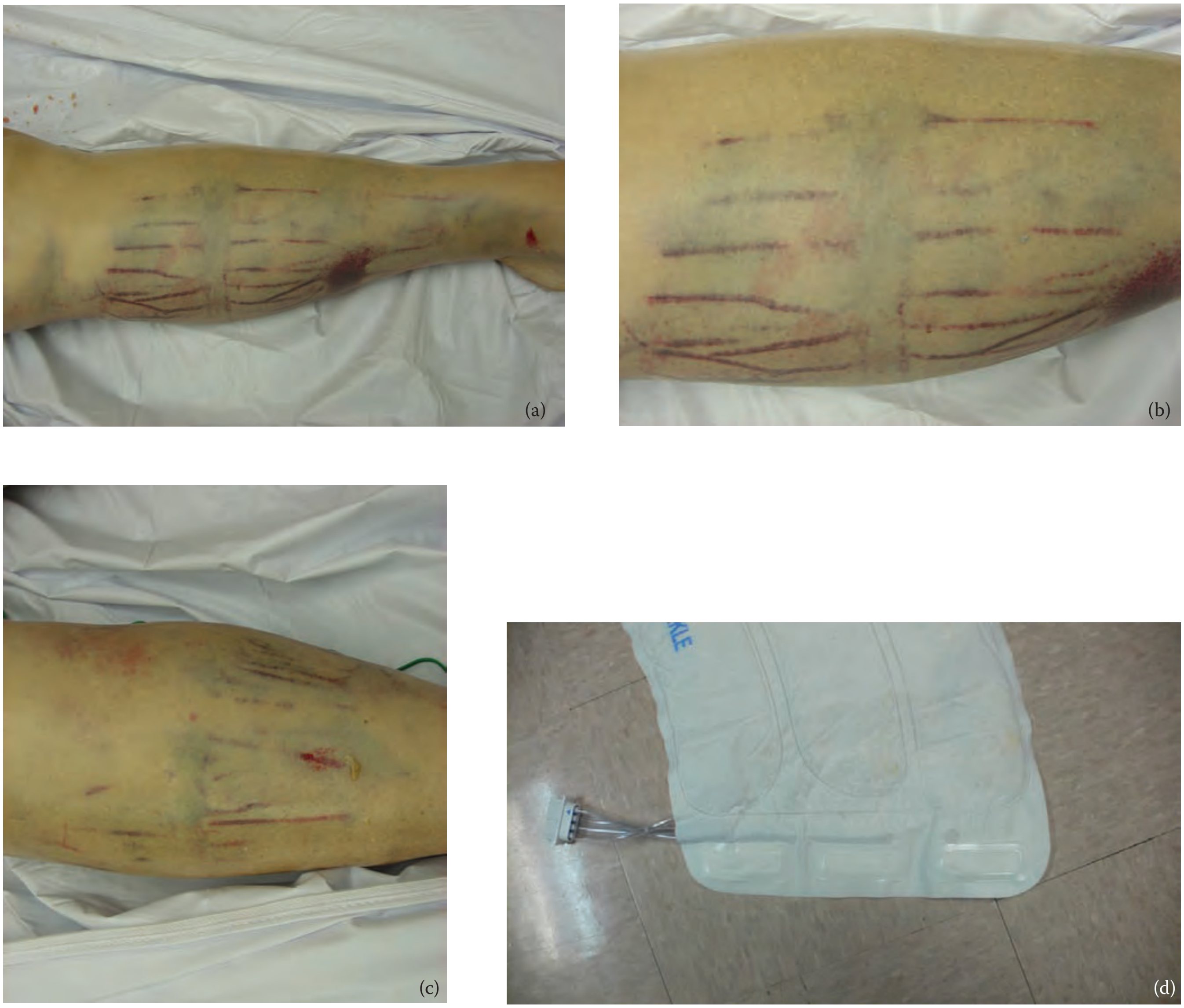

This individual was ejected from a car during a collision. She was taken to the hospital in an unconscious state and remained so until her death. There was initial speculation that she was struck by another motor vehicle due to the "tire-like" patterns on her legs (a–c). Further investigation revealed these injuries were not present at the time of arrival to the emergency department. She survived in the hospital for approximately 2 days. There was a pelvic fracture with blood seeping into the legs, extensive generalized edema, and disseminated intravascular coagulation (b). Pressure boots were placed to decrease the risk of deep venous thrombosis (d). This pattern of ecchymosis was caused by the pressure boots in conjunction with the complications of the injuries.

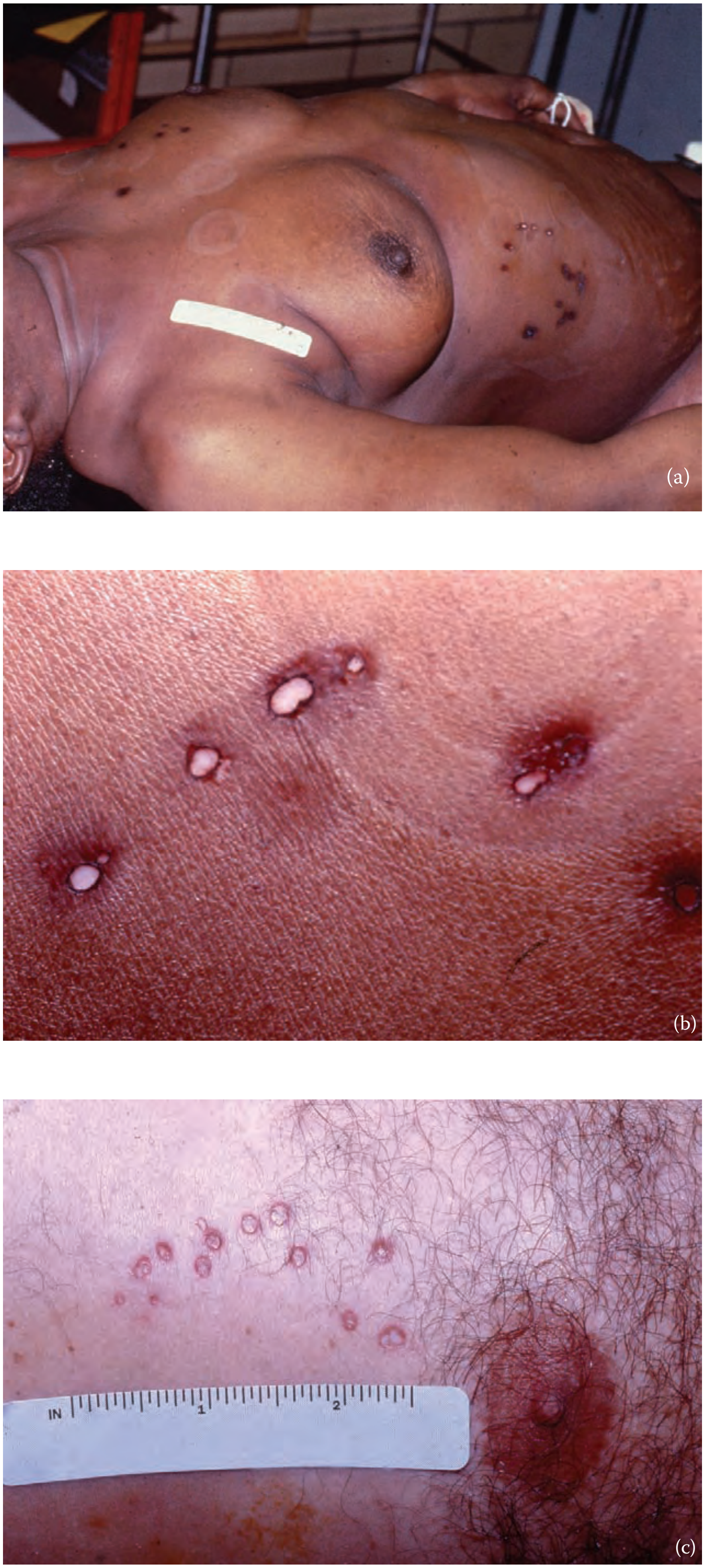

Figures (a), (b), and (c) show electrical burns caused by defibrillation with damaged cardioversion paddles during attempted resuscitation.

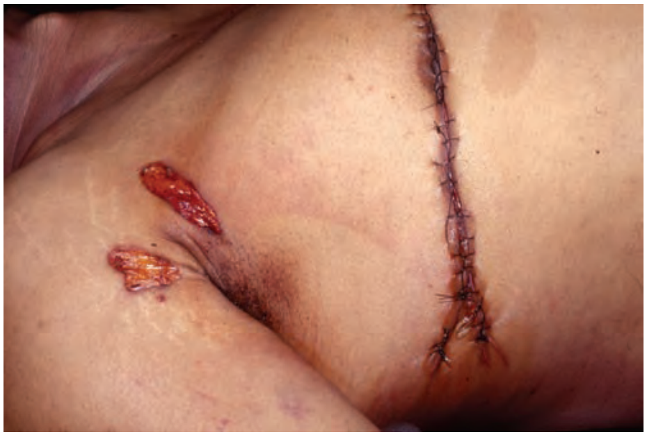

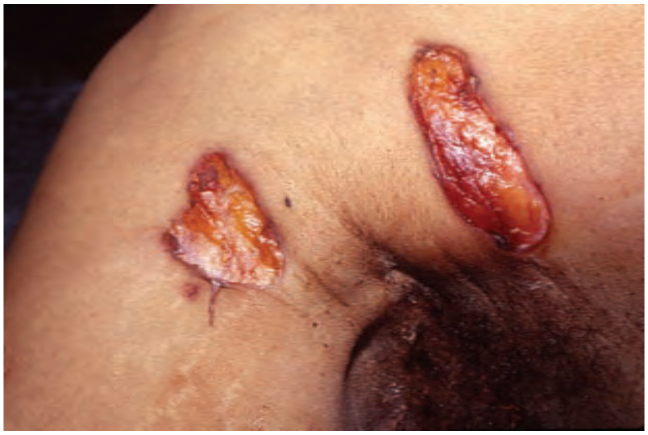

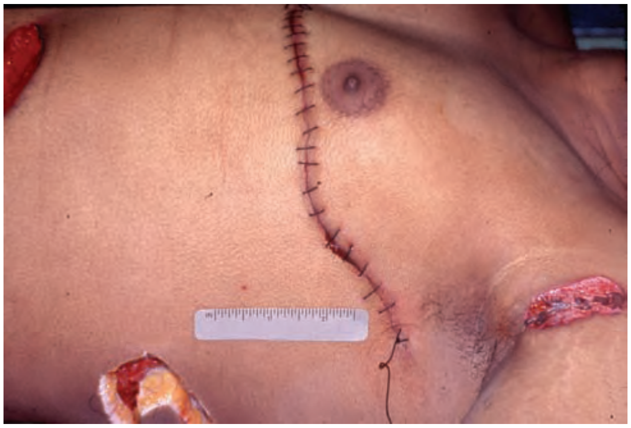

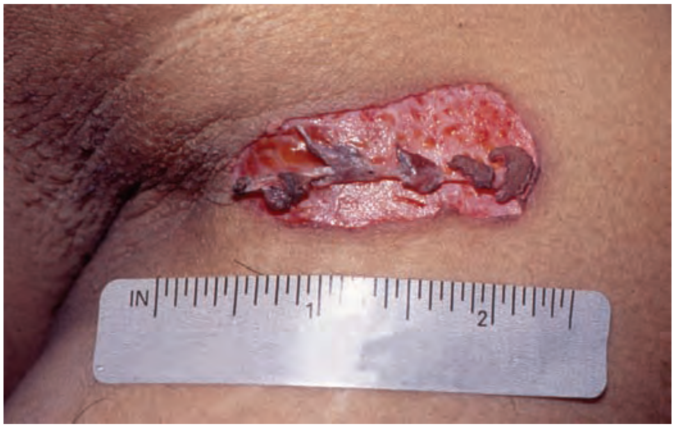

These figures show perimortem injuries to the anterior chest, axillary regions, and the anterior aspects of the upper arms associated with incorrect placement of a rib splitter during a thoracotomy.

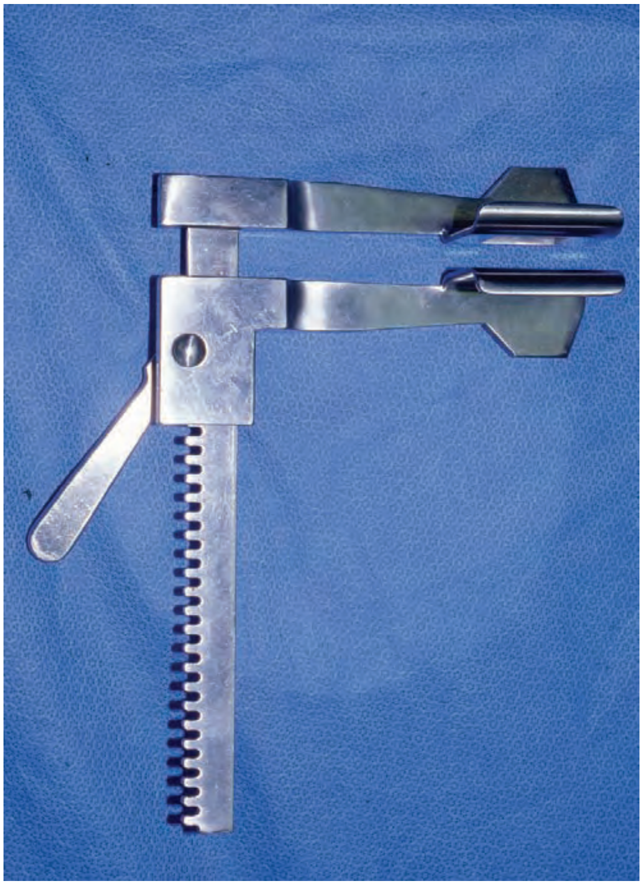

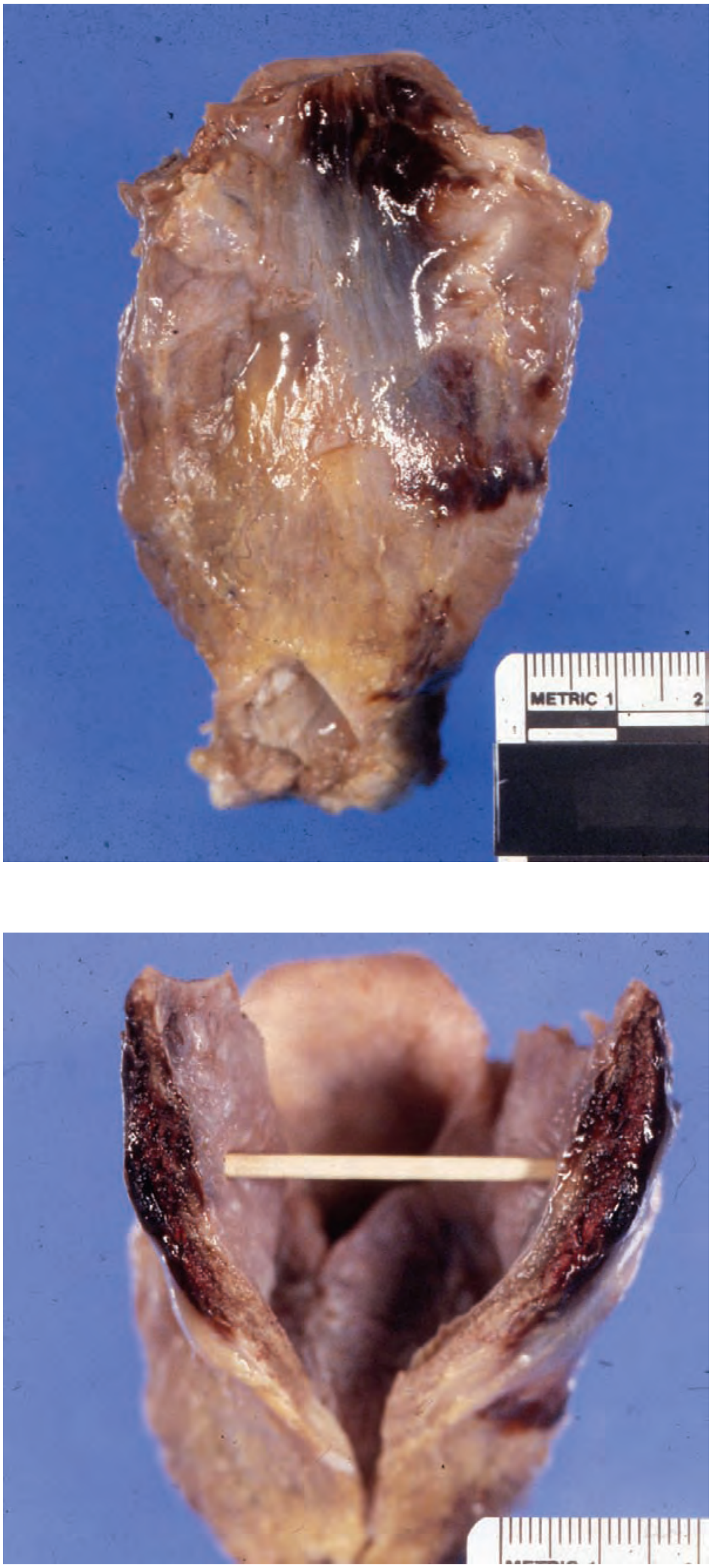

These figures demonstrate the posterior pharyngeal/esophageal cut surface with intramural hemorrhage. This is an artifact encountered frequently during resuscitative measures. The rich esophageal venous plexus tends to be a very hemorrhagic area following intubation.

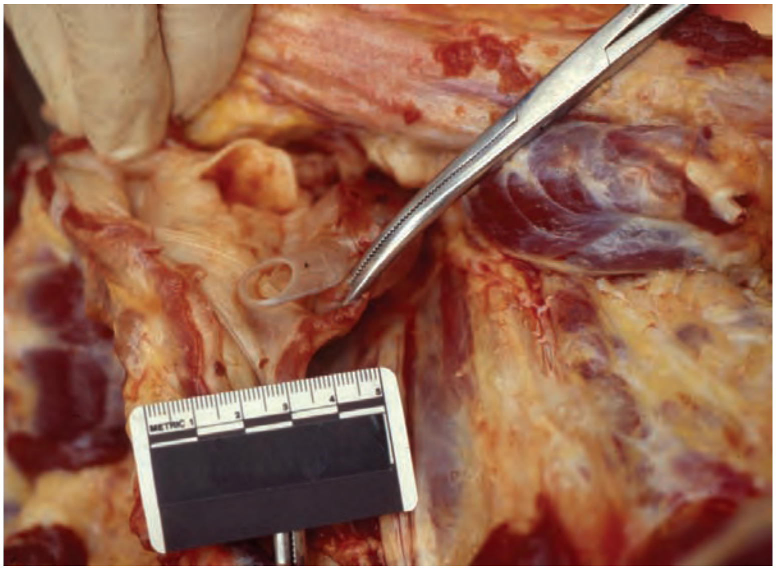

This demonstrates an esophageal intubation with the tip of the endotracheal tube protruding from the esophageal lumen. The epiglottis can be seen slightly behind the endotracheal tube, confirming an esophageal intubation.

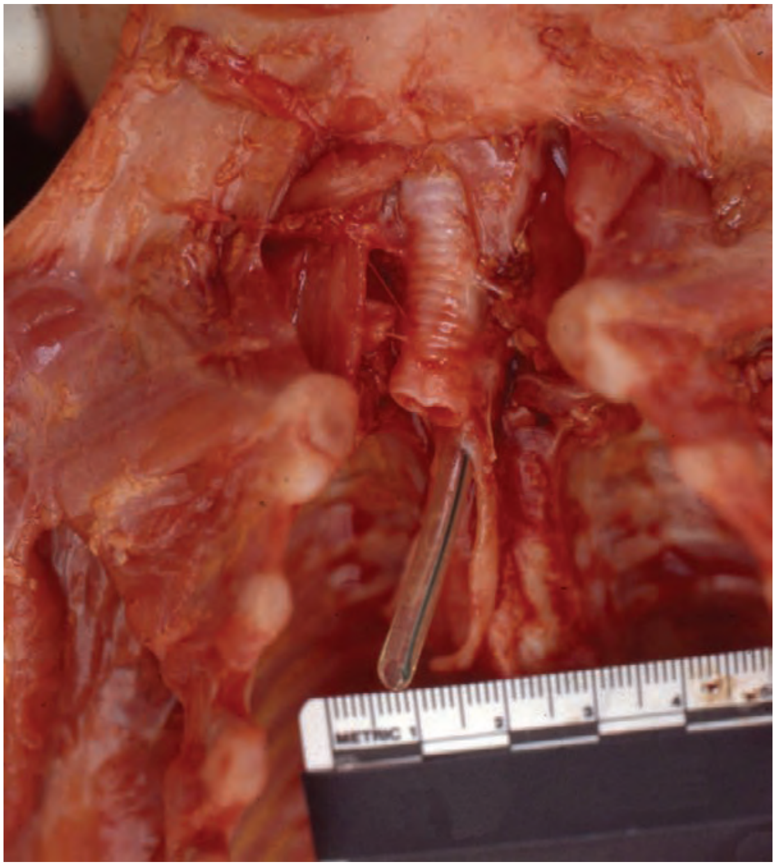

This autopsy involved an infant who succumbed to sudden infant death syndrome (SIDS). The close-up photograph depicts an esophageal intubation, with the endotracheal tube clearly within the esophageal lumen and the concentric tracheal rings visible slightly anterior to and (anatomically) to the right of the tube. Esophageal intubation is not an uncommon finding associated with resuscitation and usually plays no substantial contributory role in the death; however, it should always be documented as part of a complete autopsy and may, in some situations, be a quality control measure for Emergency Medical Services (EMS) personnel and paramedics. Esophageal intubations during an elective procedure, on the other hand, are very important and potentially causal or contributory to the death.

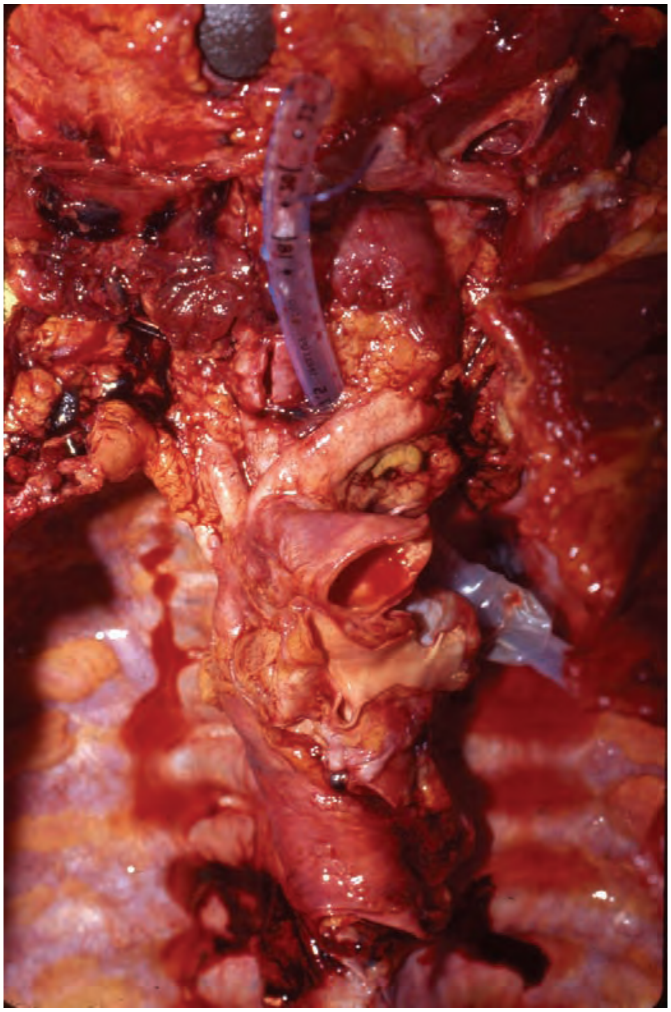

This endotracheal tube was inserted through the trachea and paratracheal soft tissues into the thoracic cavity during resuscitation.

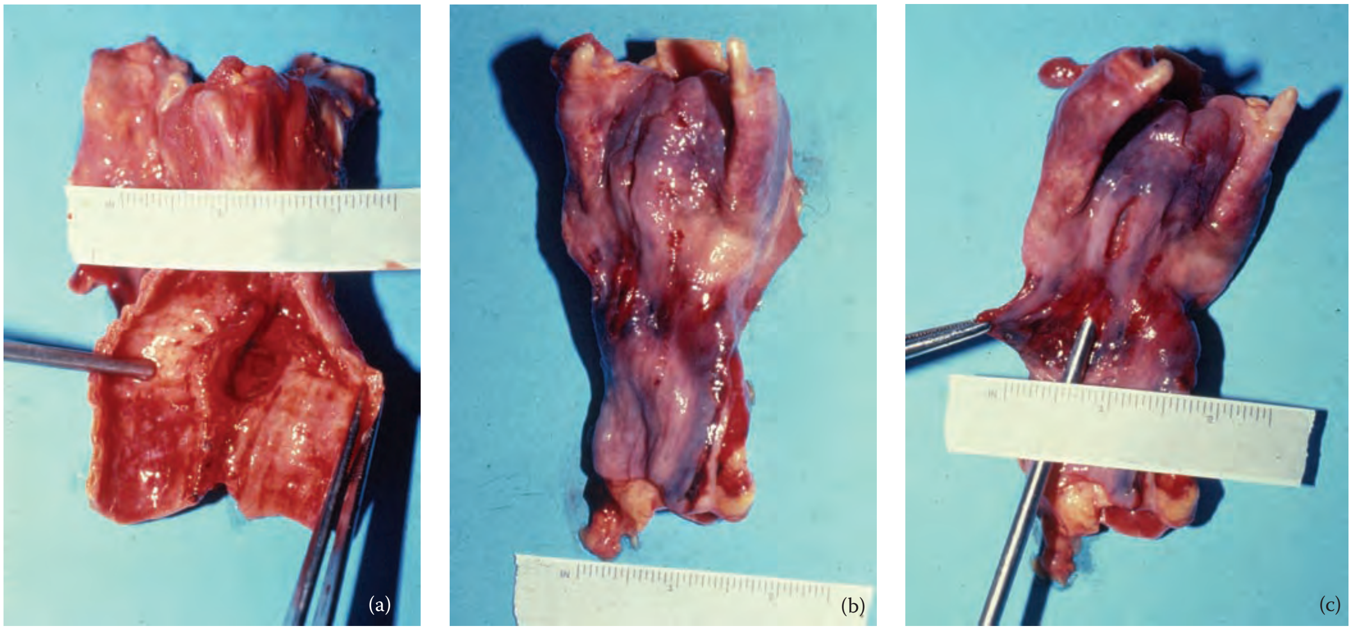

A tracheostomy tube was inserted through the posterior aspect of the trachea (a) into the esophagus (b) (traumatic tracheoesophageal fistula). Note the perforations on the posterior surface of the trachea (a) and the anterior aspect of the esophagus (b).

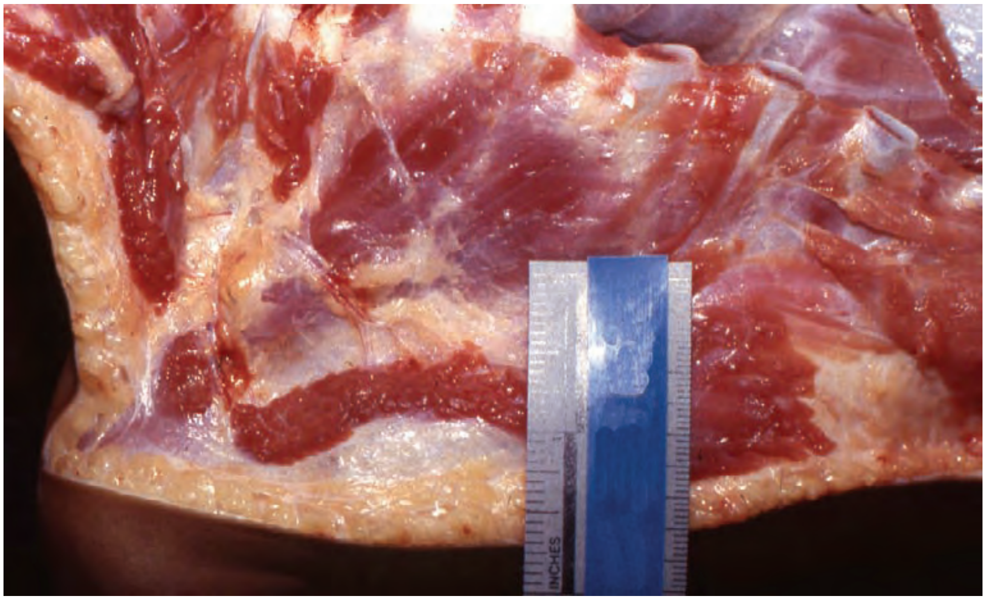

This is a view of the chest wall of an 18-month-old toddler with a history of tracheomalacia and subaortic stenosis who required a tracheostomy. During a bout of crying, the tracheostomy tube became dislodged, and after nurses attempted to reposition it, the baby rapidly developed subcutaneous emphysema, followed by bilateral tension pneumothoraces. Autopsy demonstrated marked subcutaneous emphysema, including periorbital swelling. The tube had been removed before the autopsy, precluding assessment of its placement. This particular view demonstrates air bubbles within the subcutaneous fat. It is important that all tubes remain in the body for objective postmortem (autopsy) evaluation of their placement.

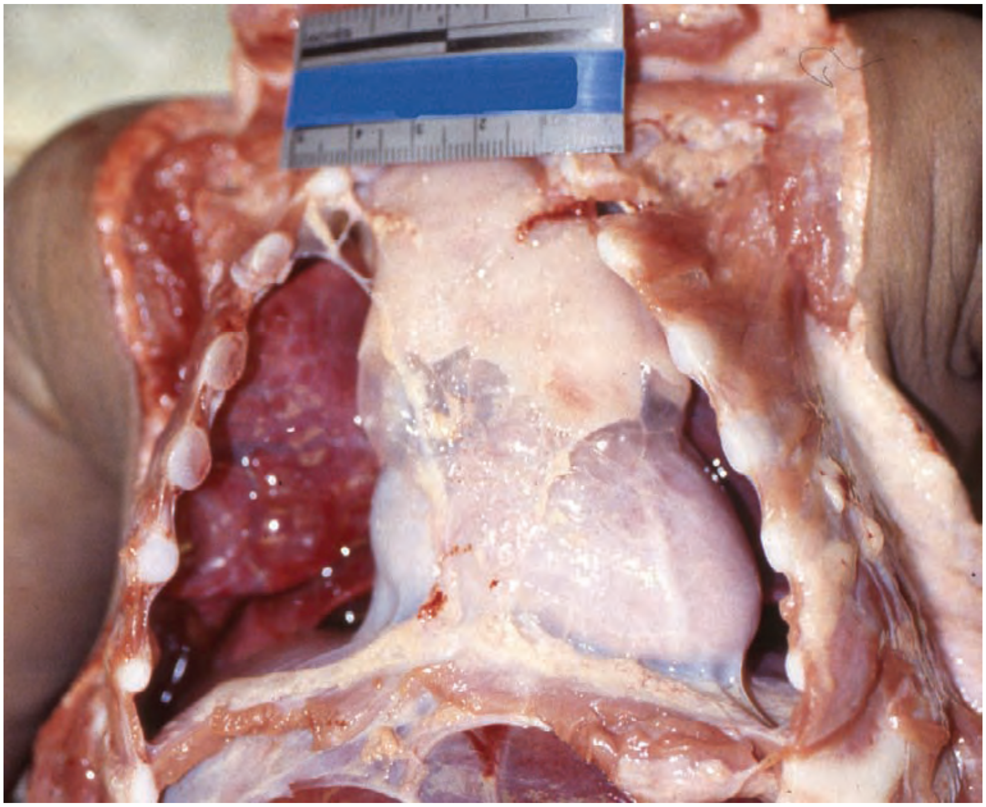

Subcutaneous emphysema. Anterior chest with air bubbles throughout the mediastinal soft tissues. This may be commonly observed with vigorous resuscitation or trauma. Palpation of these regions often reveals crepitus.

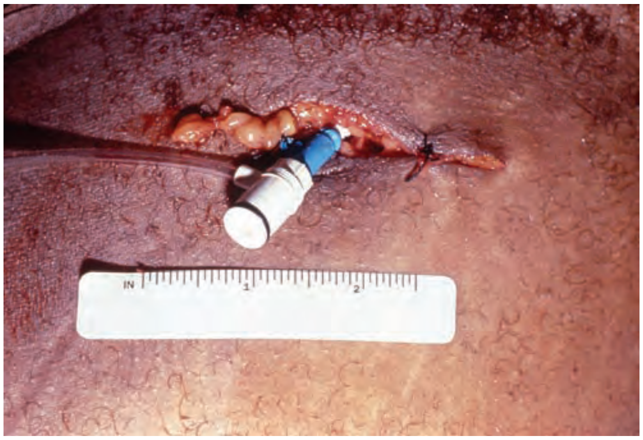

One must be careful while examining injuries altered by therapeutic intervention. This picture demonstrates a catheter placed into an injured blood vessel through a stab wound.

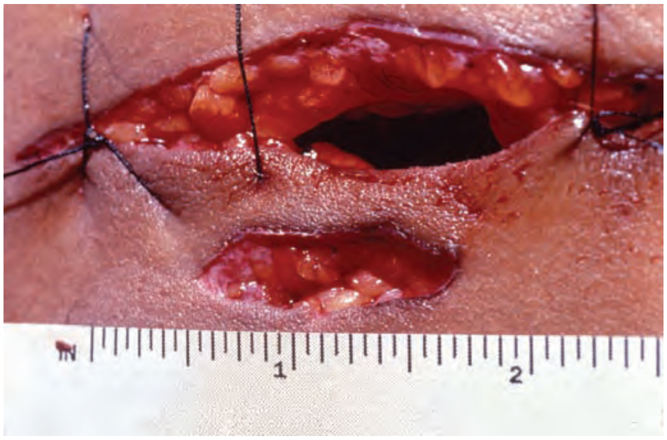

This demonstrates a stab wound adjacent to a sutured chest tube incision. This may make injury interpretation more challenging.

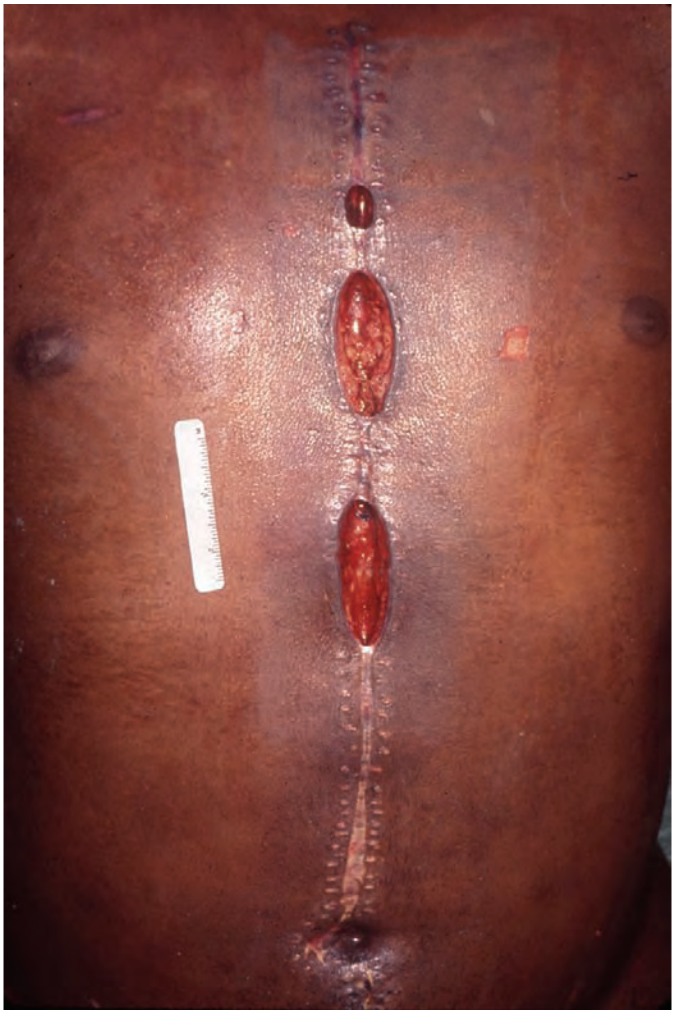

Healing infected incision. This individual died of septic complications of his initial trauma. Had it not been for this trauma, the infection that took his life would not have occurred. The manner of death in this case was ruled accidental.

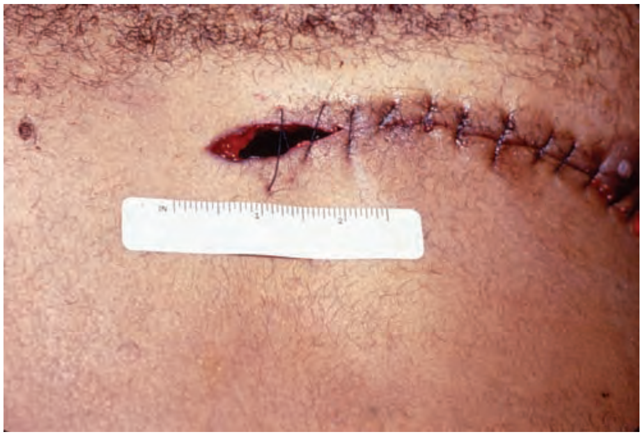

This demonstrates a stab wound that the clinicians converted along one side (edge) into a thoracotomy incision. If at all possible, this should never be done, because it makes injury interpretation much more difficult.

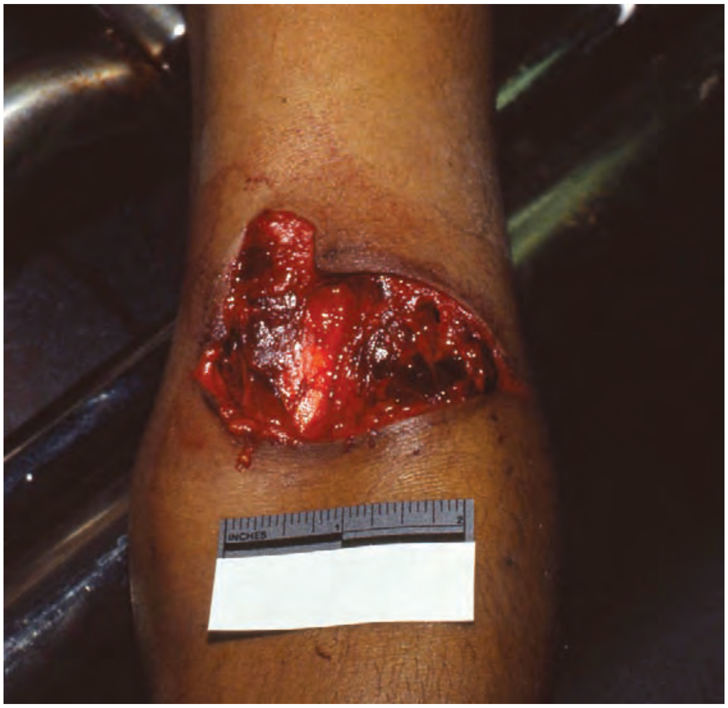

This is a "cutdown site" used for vascular access by a hospital emergency department (ED) physician, created by a transverse incision within the antecubital fossa This decedent had also sustained stab and incised wounds in other parts of the body. Again, this illustrates the importance of having a complete medical record and discussions with the treating ED physicians or trauma surgeons to properly differentiate between therapeutic interventions and injuries sustained before such therapeutic interventions.

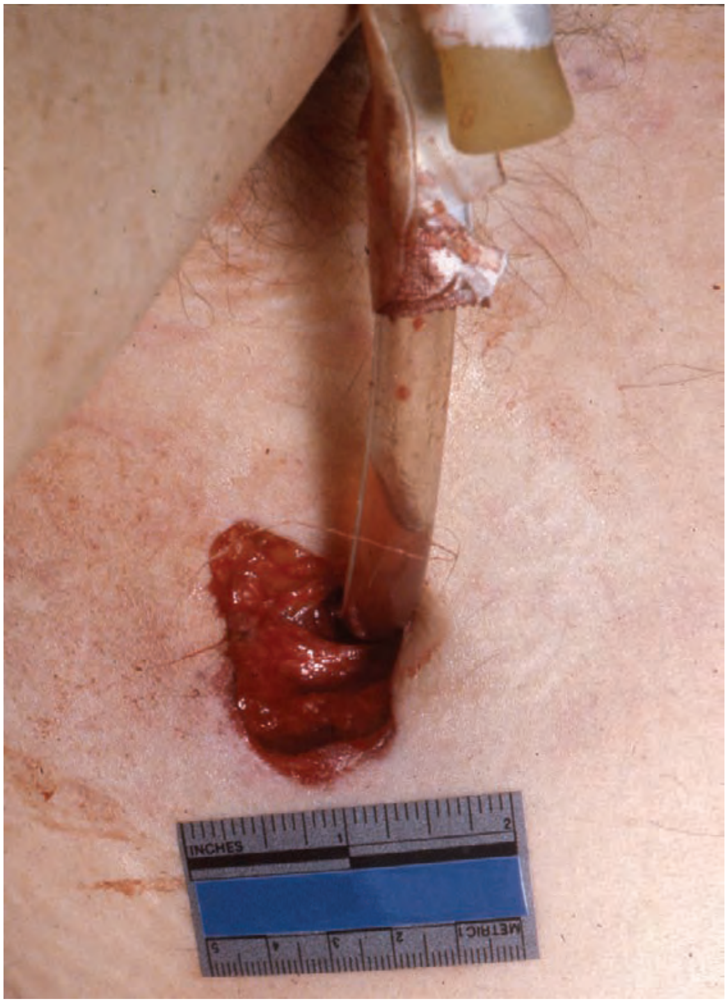

This chest tube was inserted through a previously existing perforation. Notice, however, that along the left margin of this perforation (from the 7–11 o'clock positions), there is a distinct abrasion. Further investigation established that the chest tube was inserted through a previously sustained entrance gunshot wound, thus explaining the abraded margin of this perforation.

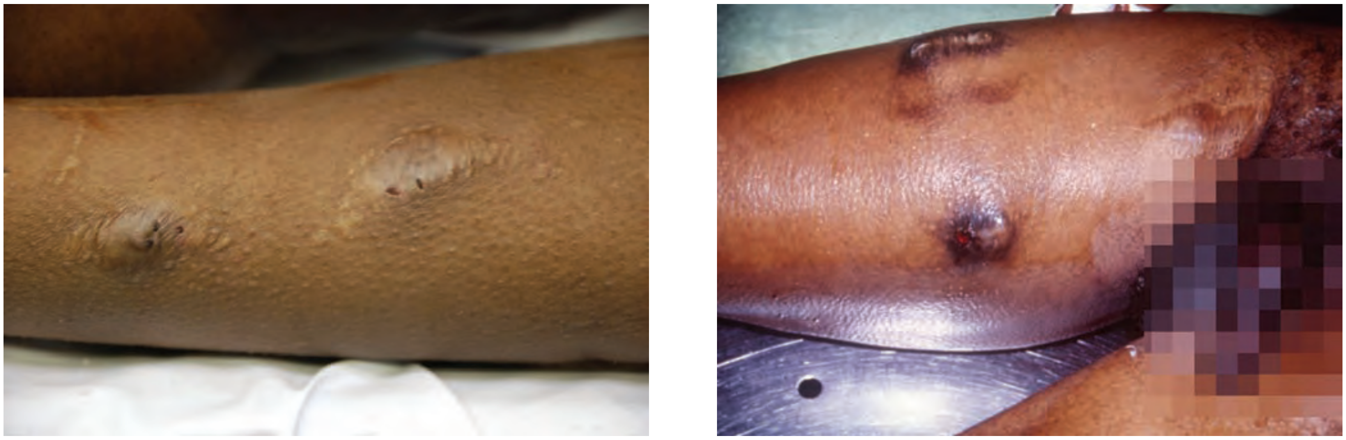

Arteriovenous hemodialysis grafts with healing needle puncture marks.

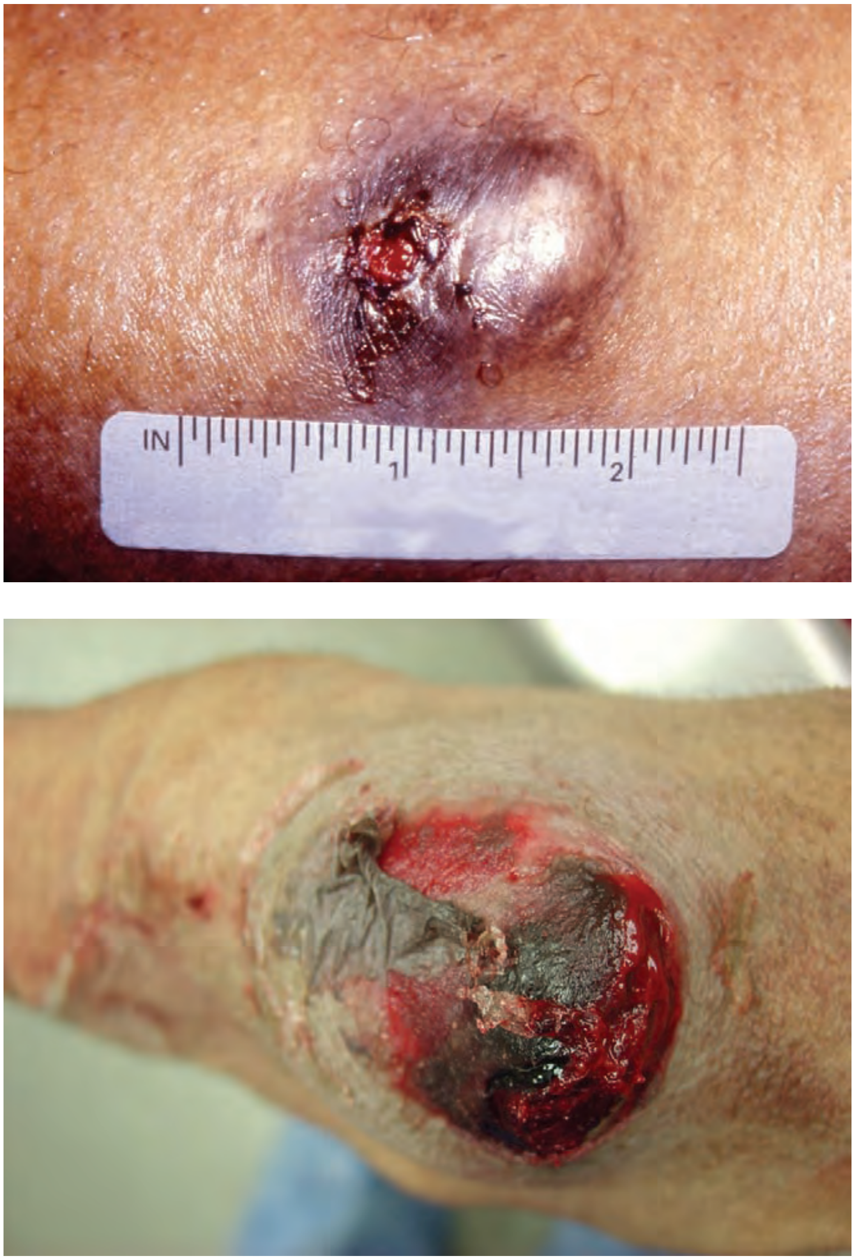

These individuals exsanguinated in their residences following ruptures of their infected arteriovenous dialysis grafts.

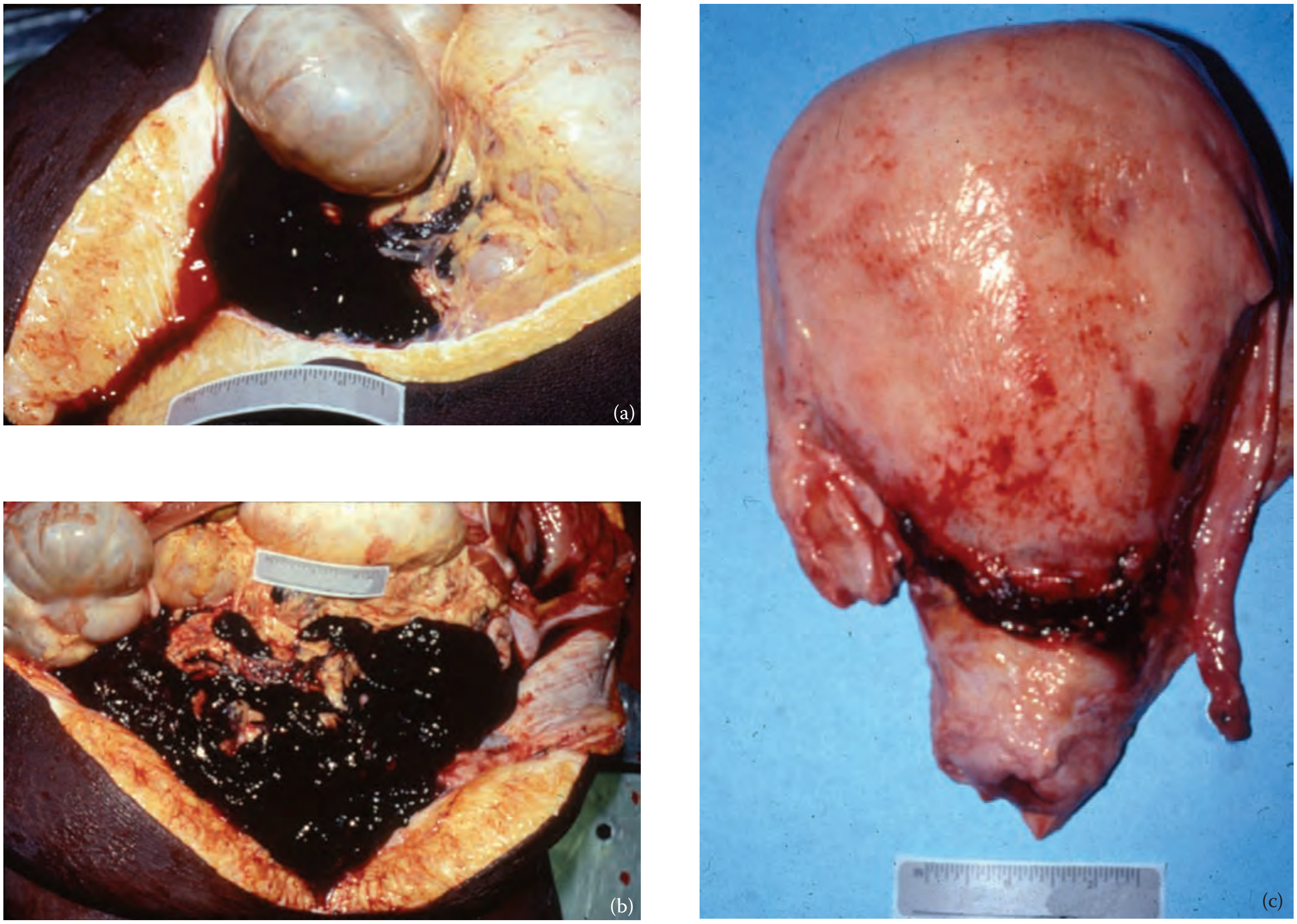

This woman underwent (a) Cesarean section (c) for an otherwise unremarkable term pregnancy hours before going into hemorrhagic shock while in the recovery room. Autopsy disclosed a large hemoperitoneum with clotted blood extending from a bleeding abdominal wall vein (a) and (b).

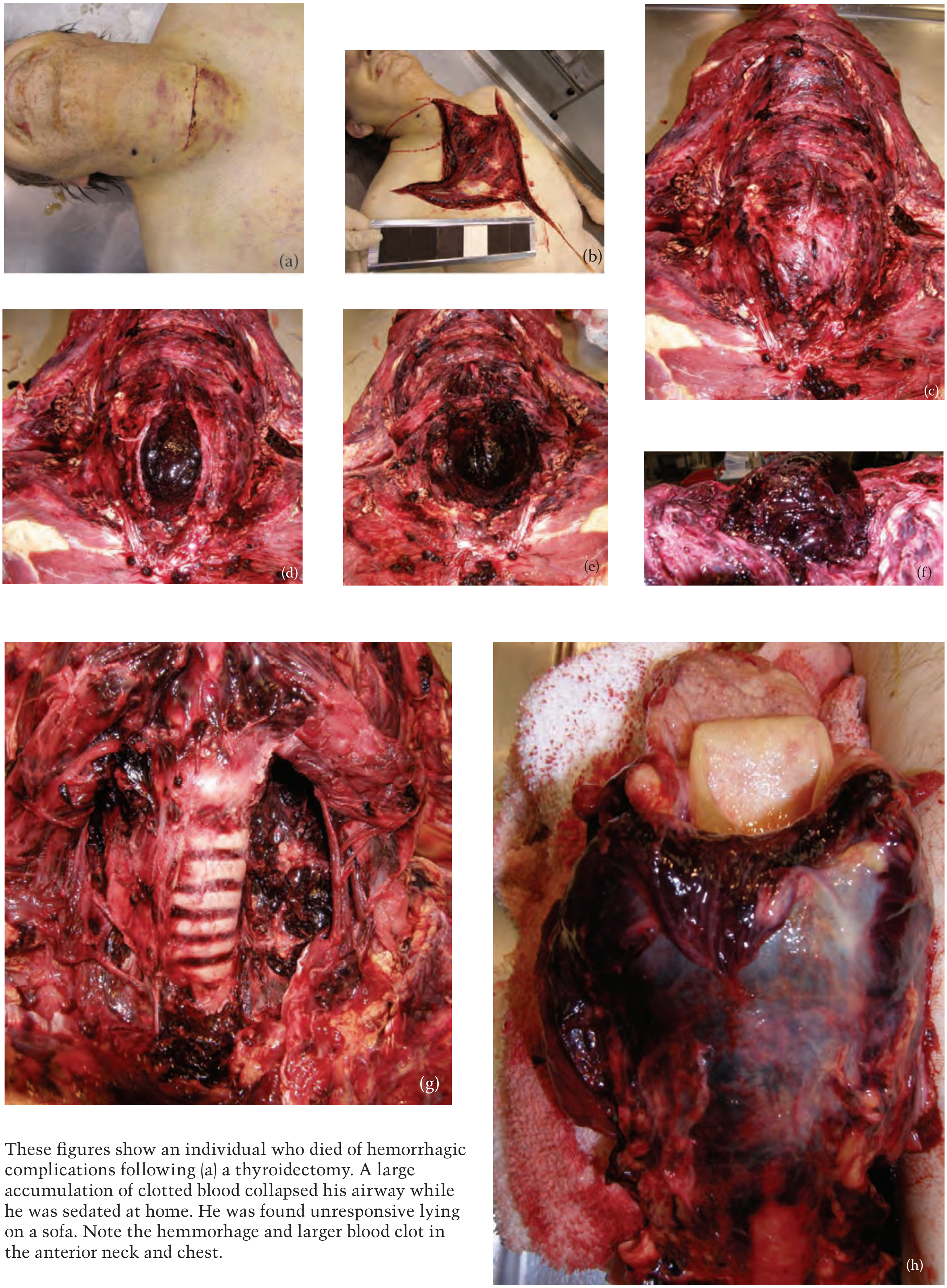

These figures show an individual who died of hemorrhagic complications following (a) a thyroidectomy. A large accumulation of clotted blood collapsed his airway while he was sedated at home. He was found unresponsive lying on a sofa. Note the hemmorhage and larger blood clot in the anterior neck and chest.

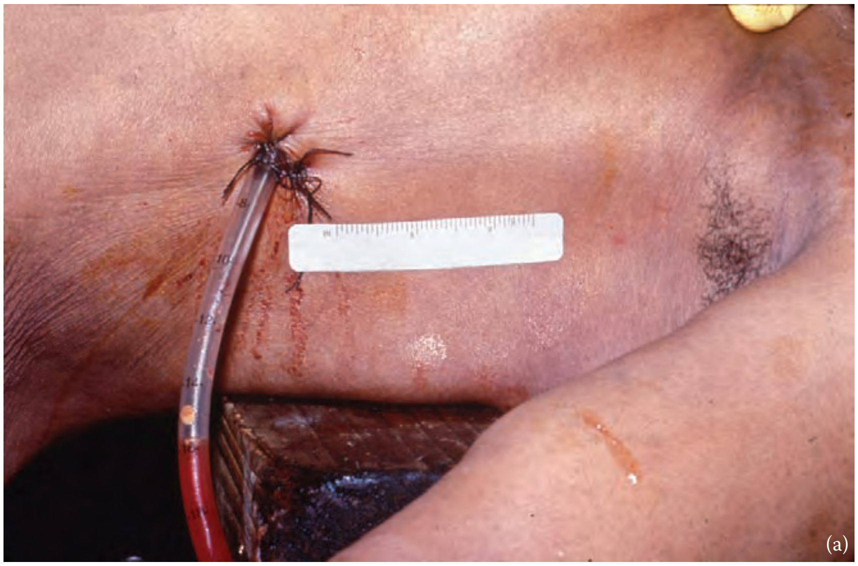

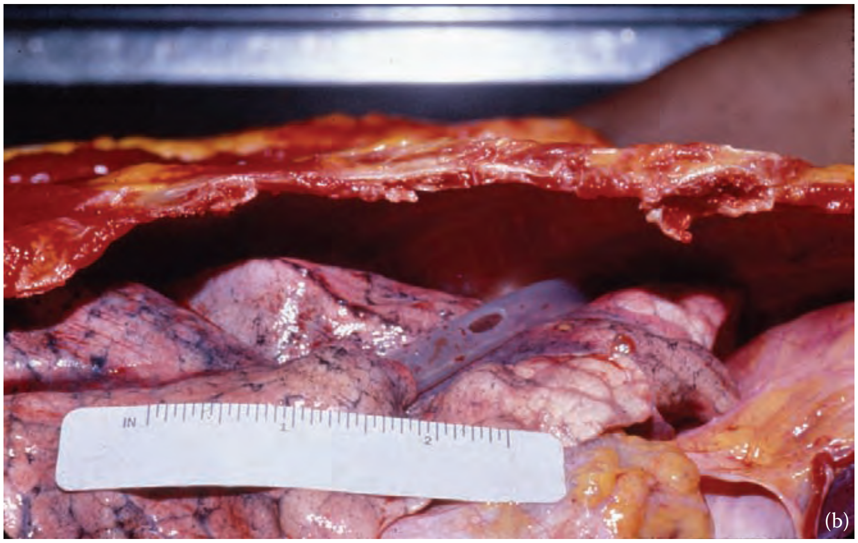

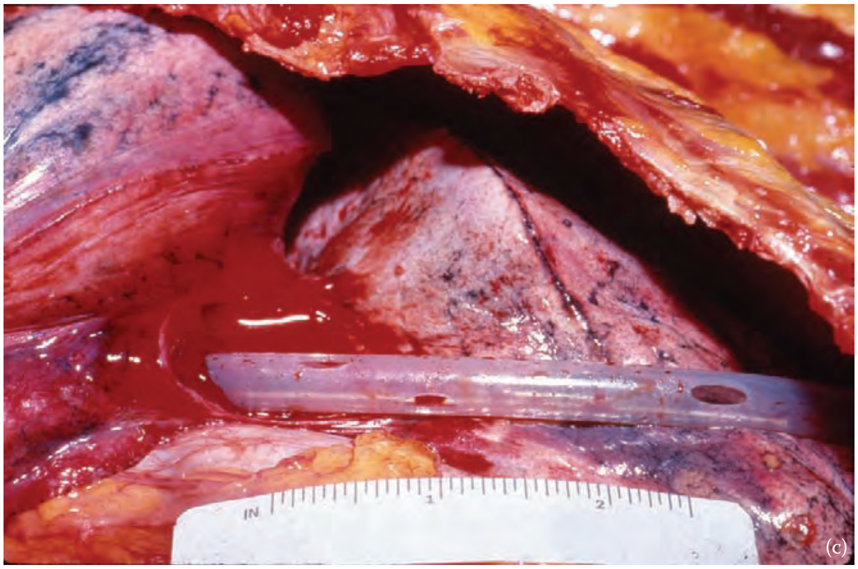

These figures show a patient who was admitted in cardiac arrest following an acute asthma attack. Chest tube placement was inserted through the lung parenchyma during resuscitation. Approximately 150 mL of liquid blood were recovered from the left hemithorax, indicating that this injury was perimortem and iatrogenic.

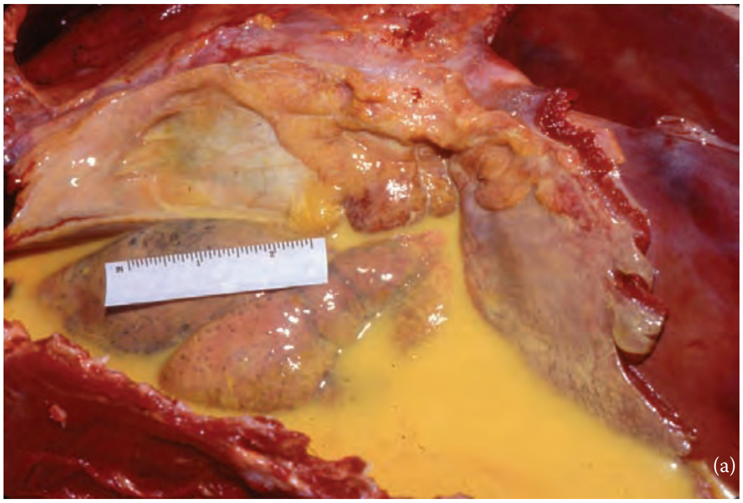

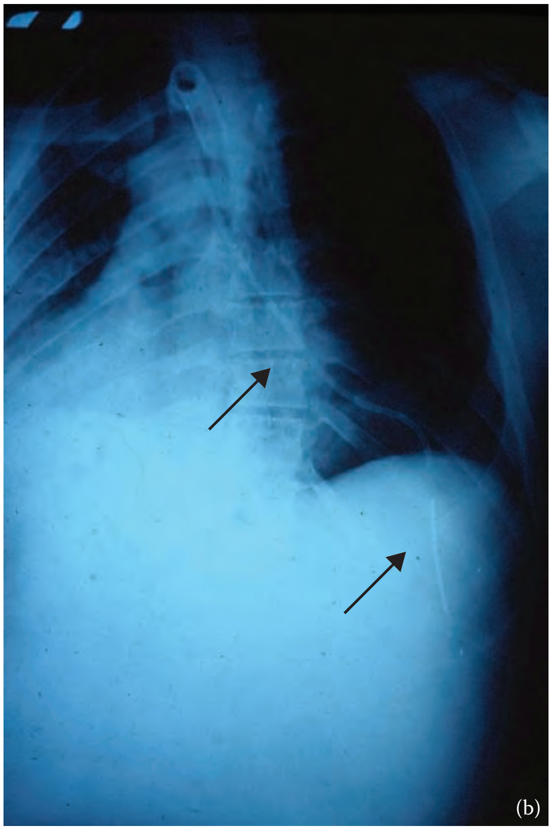

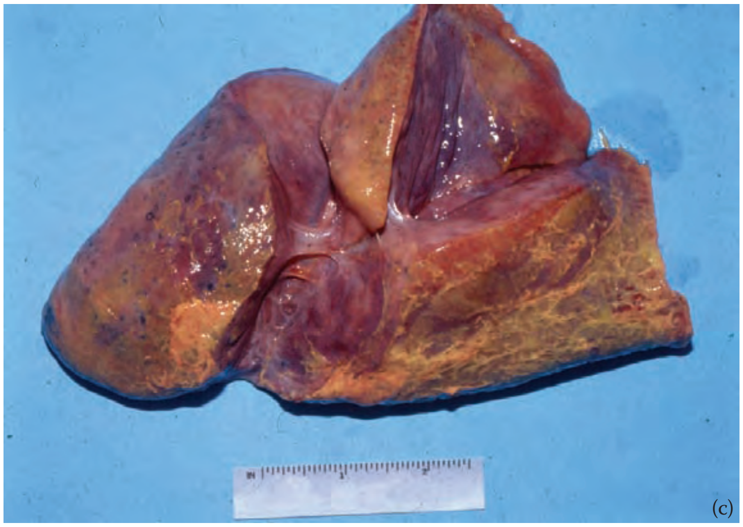

This individual underwent placement of a nasogastric tube that perforated the esophagus and entered the right thoracic cavity (hemithorax). The injury went unnoticed for many hours. The decedent was fed through the nasogastric tube, which is demonstrated by the accumulation of yellow fluid within the thoracic cavity. Autopsy revealed a fibrinous pleuritis and an early pneumonia.

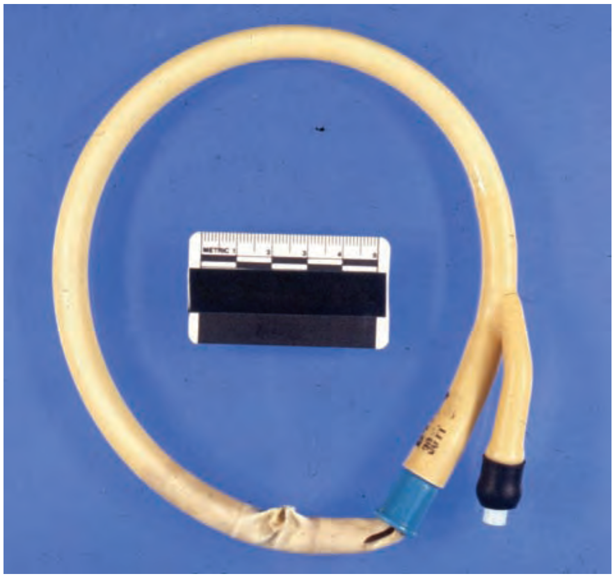

Percutaneous endoscopic gastrostomy feeding tube.

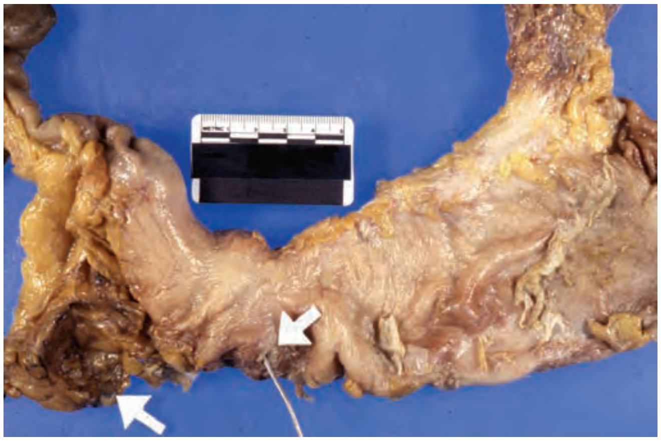

Panoramic view illustrating the anterior gastric wall, with a probe through the previous PEG site, also demarcated by an arrow, and the other arrow demonstrating greater omentum with fibrinous exudate. Similarly appearing exudate can also be seen on the anterior surface of the gastric body and fundus.

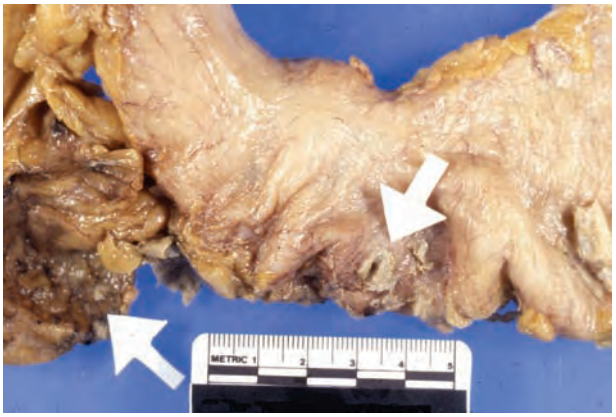

This is a close-up anterior view of the stomach, demonstrating the previous PEG site through the anterior gastric wall, as well fibrinous exudate on the surface of the greater omentum, with both features demarcated by arrows.

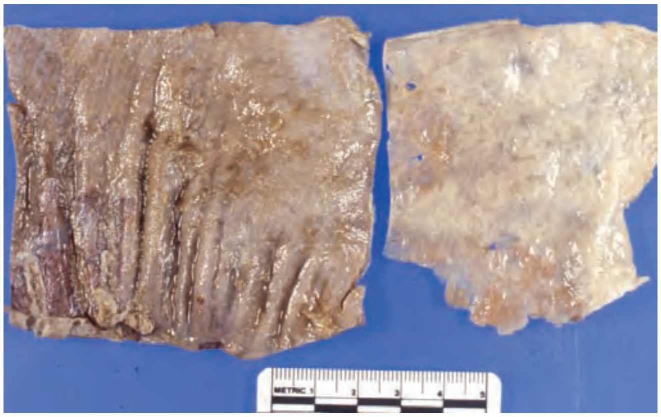

This depicts the parietal peritoneum with a gray, shaggy, fibrinous exudate.

This case involved an elderly woman with a history of dementia who could no longer live independently and was made a ward of the state. She was hospitalized for long-term intermediate care. She had undergone percutaneous endoscopic gastrostomy (PEG) tube insertion due to insufficient oral intake (a). She underwent a tube replacement for leakage. Later that evening, she experienced acute onset of abdominal pain that was complicated by increasing abdominal distension, green discoloration of the abdominal wall site, and purulent drainage from the tube site. The remainder of her course was characterized by respiratory distress, aspiration of gastric contents, hypotension, and tachypnea, culminating in systemic inflammatory response syndrome with lactic acidosis, multisystem organ failure, and hemodynamic instability, refractory to aggressive resuscitative measures. Autopsy revealed intraperitoneal placement of the gastrostomy tube with intraperitoneal administration of enteral feedings and a collection of approximately 700 mL of turbid peritoneal fluid with extensive fibrinous adhesions. Microscopic examination confirmed an exuberant fibrinous exudate with scattered clusters of bacterial cocci and yeast, as well as birefringent material. The cause of death was certified as complications of peritonitis following intraperitoneal placement of gastrostomy tube during tube change, following percutaneous endoscopic gastrostomy insertion for insufficient oral intake related to atherosclerotic cerebral vascular disease and dementia. The manner of death was certified as accident.

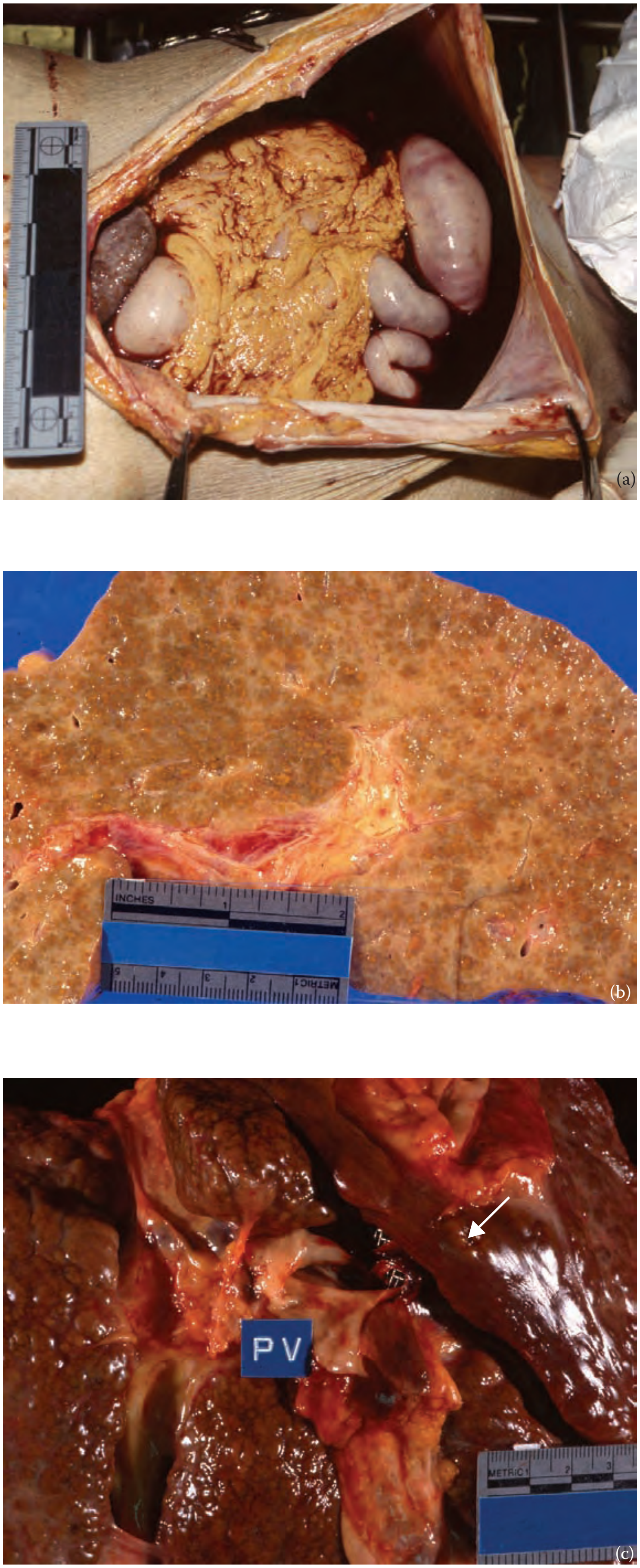

This was a middle-aged woman with primary biliary cirrhosis who underwent a transjugular intrahepatic portosystemic shunt (TIPS) for decompression of high portal venous pressure (portal hypertension). The picture (a) depicts a large hemoperitoneum, with greater omentum, stomach, and intestines floating on top of a pool of blood. There was advanced end-stage cirrhosis (b), with confluent scar enveloping and entrapping regenerative parenchymal nodules. Liver diseases such as this are associated with an increased risk of hemorrhagic complications due to coagulopathy and portal hypertension. This fatal hemorrhage resulted from laceration of a portal vein branch occurring during stent placement (c).

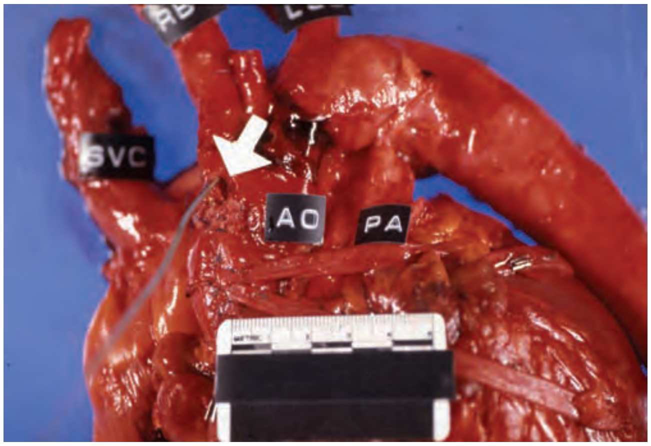

Note the probe inserted through the perforation site, which is also demonstrated with the arrow.

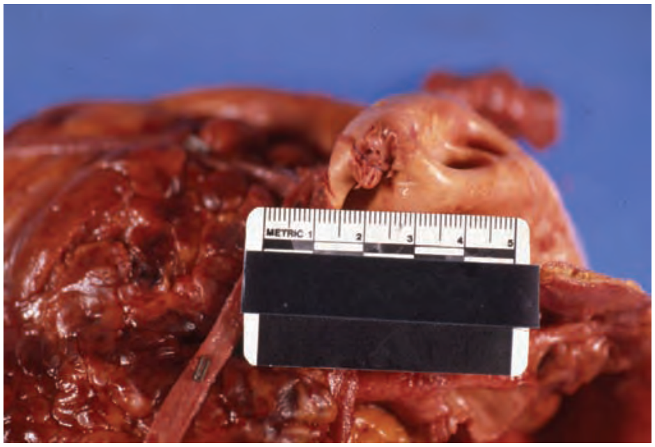

This image demonstrates the sutured cannulation site as well as one of the bypass grafts.

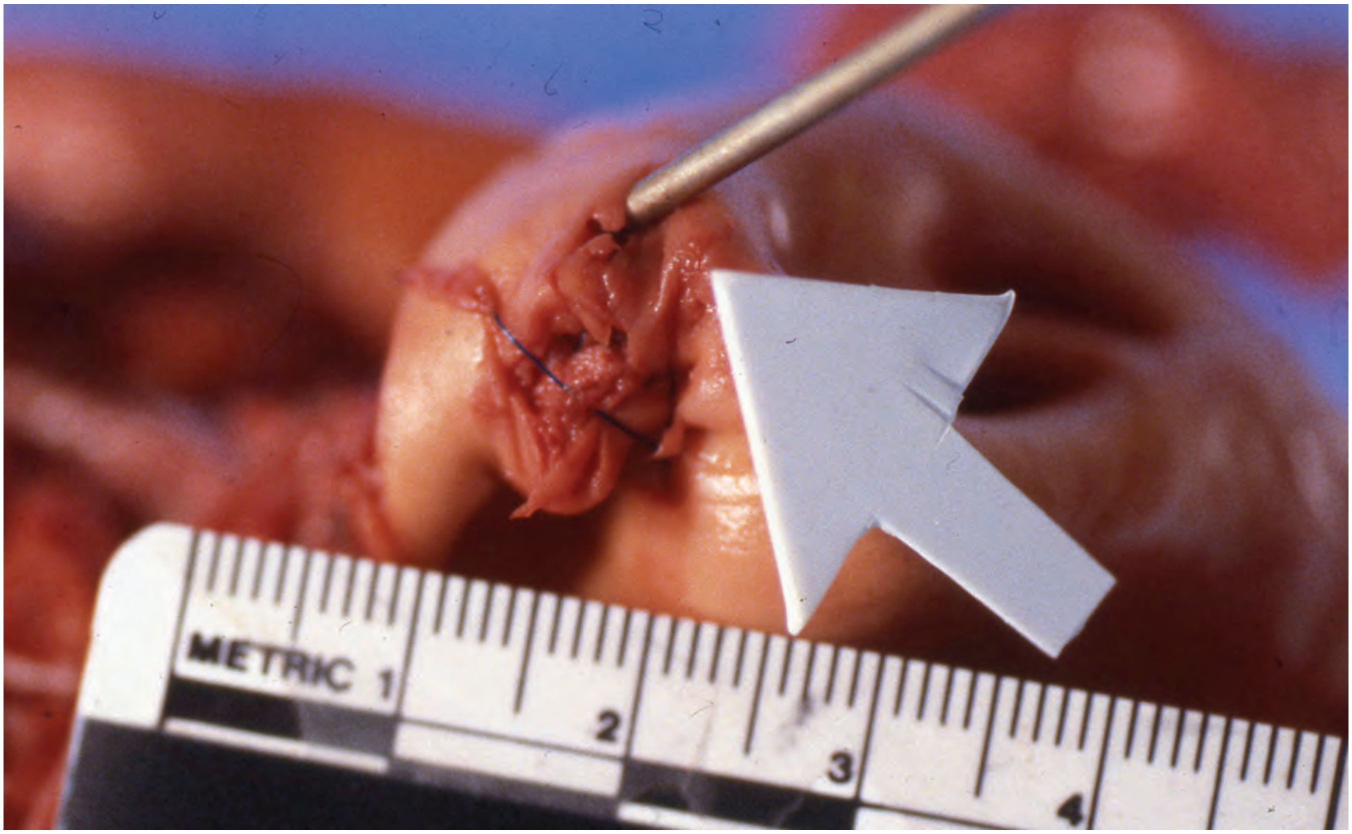

Close-up of perforation site demonstrated by probe and arrow.

This case involved a patient who underwent six-vessel coronary artery bypass grafting that was followed by intractable postoperative hemorrhage. The pericardial and left pleural drainage exceeded 1 liter within the first postoperative hour. The patient became hypotensive, requiring emergent re-exploration that was refractory to resuscitative efforts. Autopsy revealed an approximately 3-mm perforation of the ascending aorta, immediately above and distal to a sutured aortic cardiopulmonary bypass cannulation site, with residual bilateral hemothoraces. (left: 1,100 mL and right: 450 mL). The cause of death was certified as intractable hemorrhage with perforation at the aortic cannulation site complicating coronary artery bypass grafting for atherosclerotic coronary artery disease. Hypertensive heart disease was listed as a contributory cause (heart weight: 690 grams and body weight: 120 kgs). The manner of death was certified as therapeutic complication.

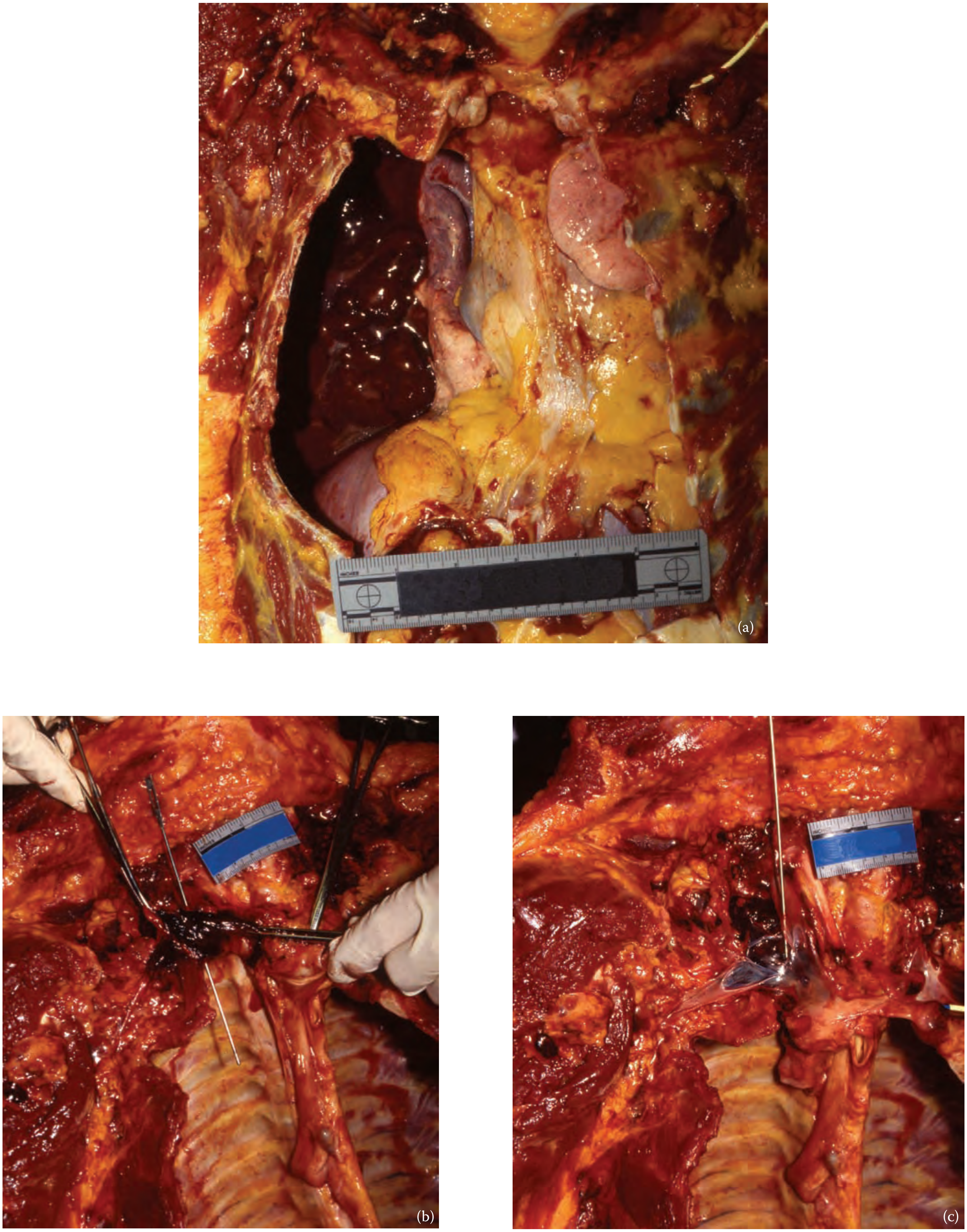

Right hemothorax (a) due to perforation of right internal jugular vein complicating catheter insertion. Note the probe (b) and (c) demonstrating the perforation through the right internal jugular vein and the hemorrhage within the anterior overlying soft tissues.

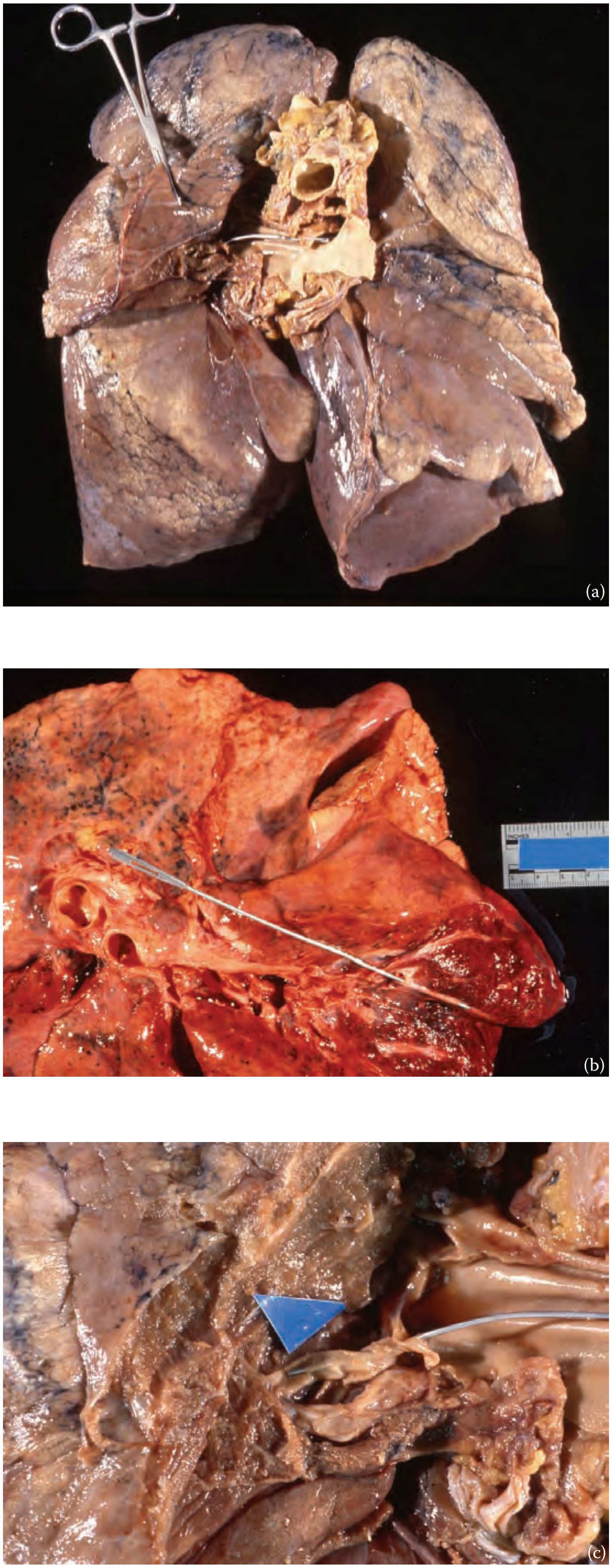

This case involved complications of a transbronchial biopsy in an individual who was HIV-positive and was slightly thrombocytopenic (platelet count: approximately 75,000/uL). The procedure was followed by extensive pulmonary hemorrhage, which culminated in respiratory compromise and death. Figure (a) demonstrates a Swan-Ganz catheter within the pulmonary arterial system. Figures (b) and (c) is a close-up view with a probe through a perforation of one of the large branches of the right main pulmonary artery with the arrow head demarcating the tip of the probe.

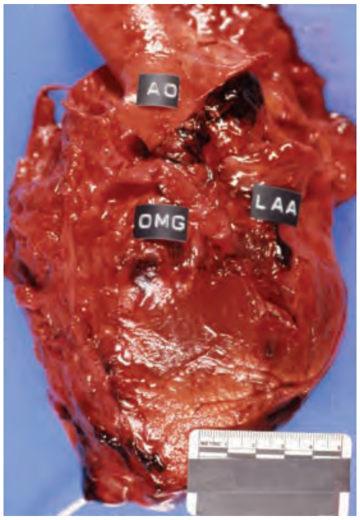

Heart with labels illustrating the aorta (AO), obtuse marginal graft (OMG), and left atrial appendage (LAA).

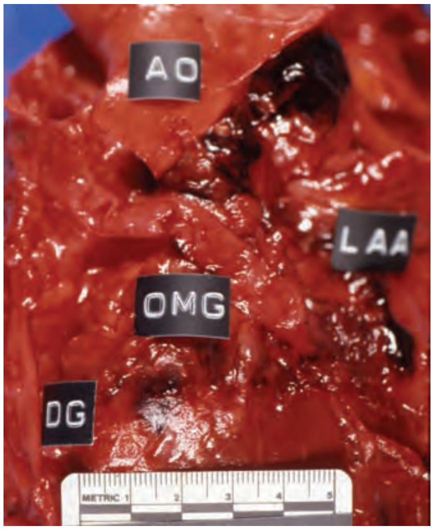

Close-up view of grafts.

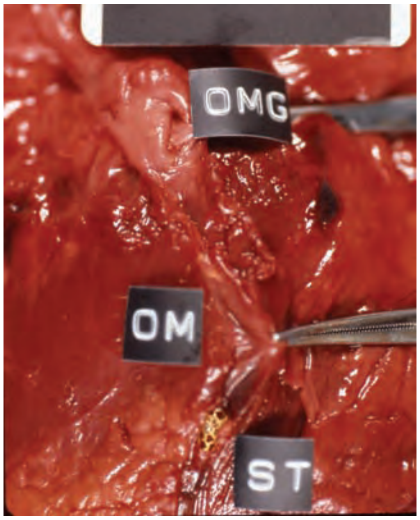

This image demonstrates the relationship of obtuse marginal graft and native obtuse marginal vessel, as well as a stent protruding through the lacerated native obtuse marginal artery.

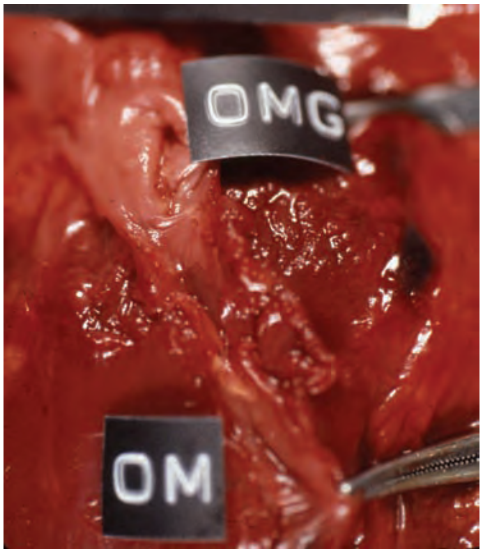

Close-up view of the junction between the obtuse marginal graft and the native obtuse marginal artery demonstrating the irregularity in the arterial wall.

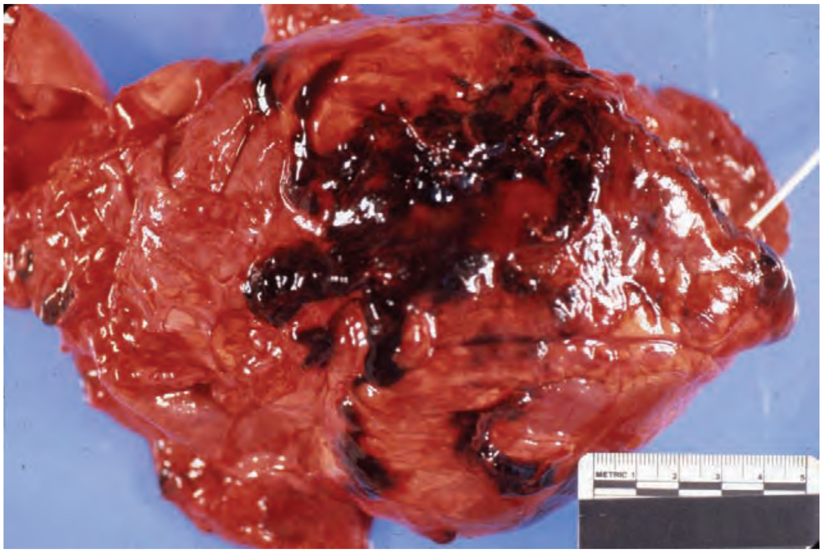

Posterior view of heart demonstrating confluent epicardial hemorrhage associated with this procedure.

These pictures demonstrate a complication of coronary artery bypass grafting. In this case, a stent perforated the native obtuse marginal branch of the left circumflex artery immediately distal to the anastomosis between the graft and native vessel.

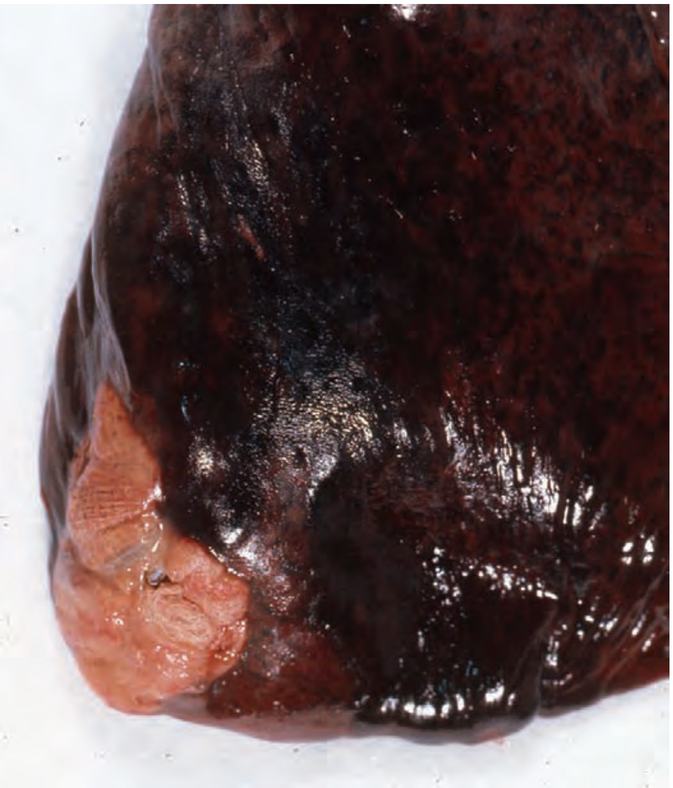

This depicts the lower lobe of the left lung with congestion and pneumonia. Note at the inferolateral aspect of the left lower lobe is a fragment of gauze that was inadvertently left behind during another operation months earlier. The gauze is adherent to the surface with overlying adhesions and adjacent purulent exudate.

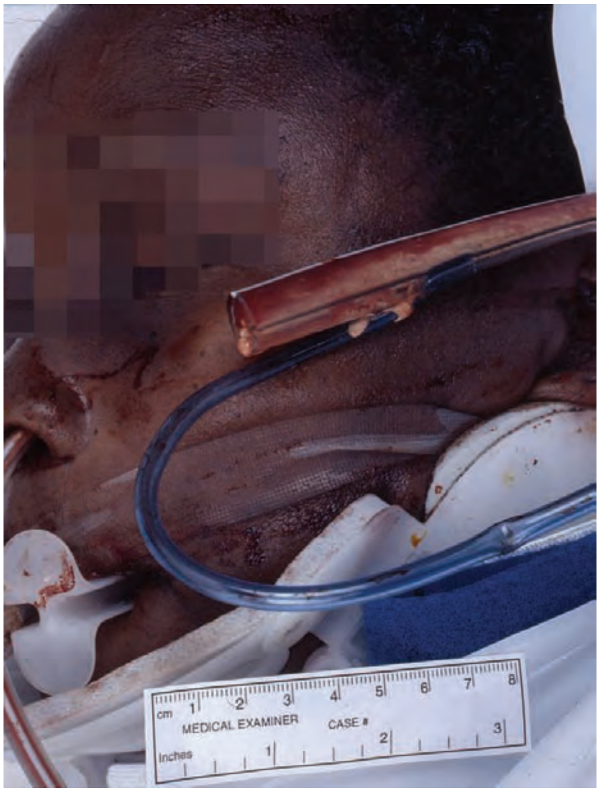

This individual was in a motor vehicle accident and sustained a skull fracture involving the cribriform plate. A nasogastric tube was inserted that inadvertently penetrated the cranium. The tube was placed on intermittent suction. Aspirated brain matter is visible at the end of the tube.

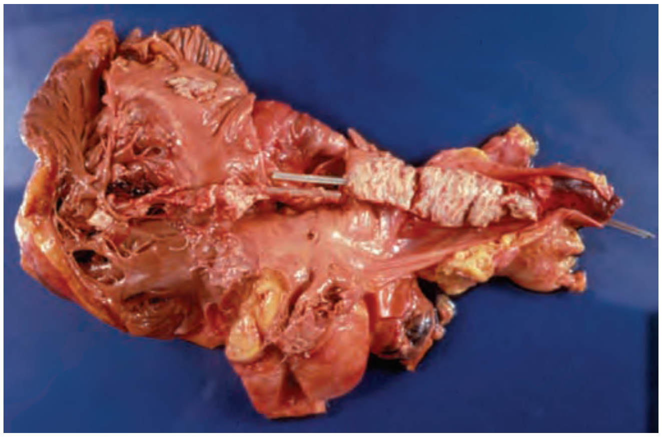

This demonstrates a large gray-tan thrombus encasing a pacemaker lead extending from the heart.

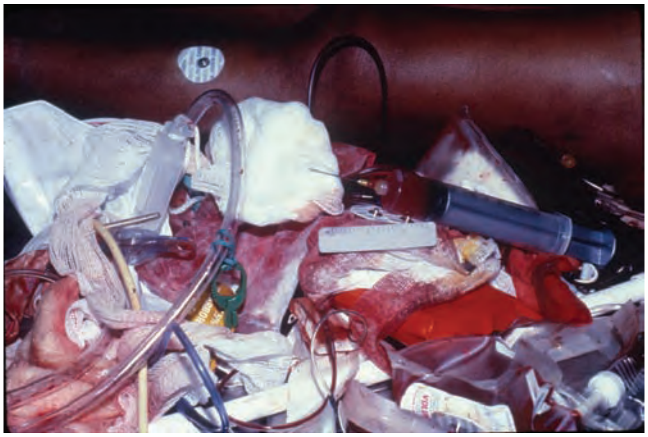

In busy trauma centers, the body is sometimes received with multiple instruments used for resuscitative efforts. This photograph was taken after the body bag was opened at the morgue. It is always important to be careful of sharps that have been inadvertently left behind.

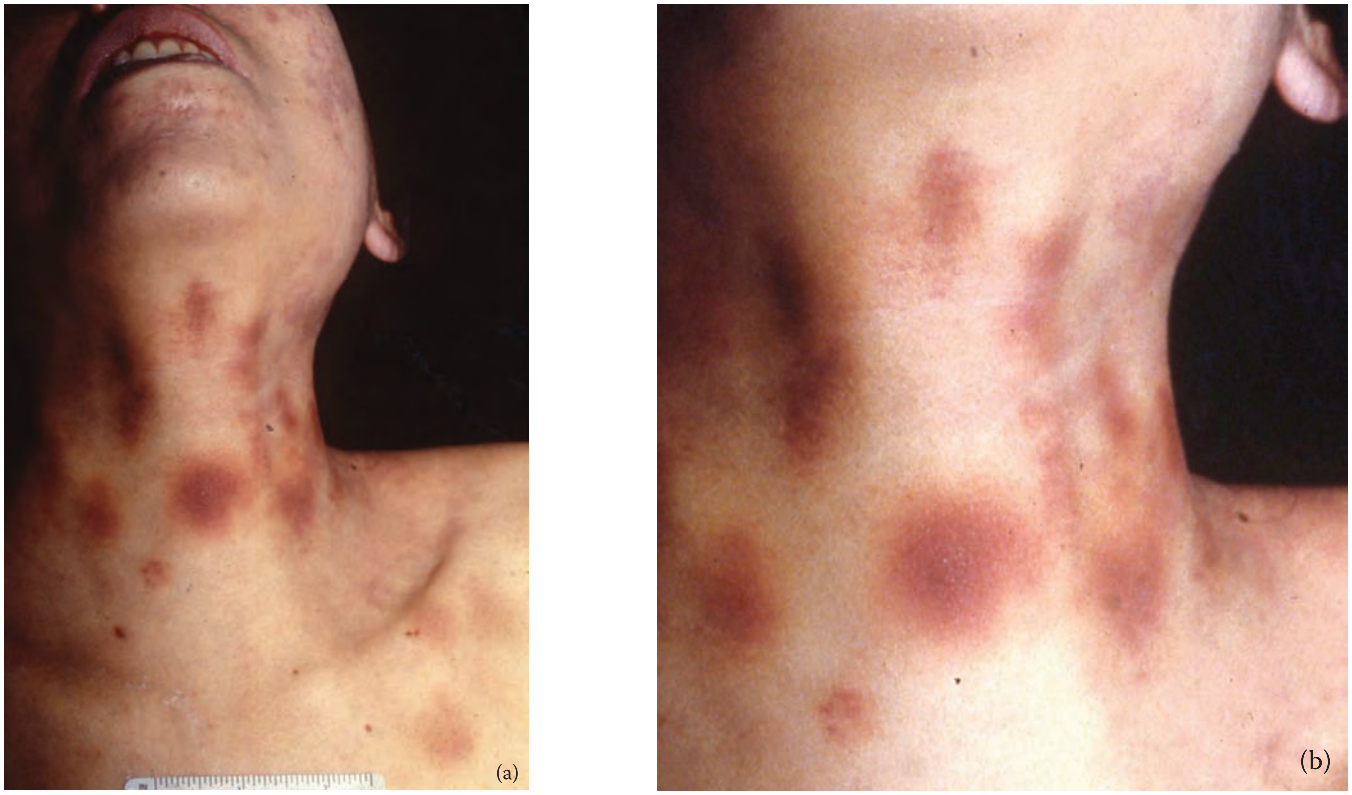

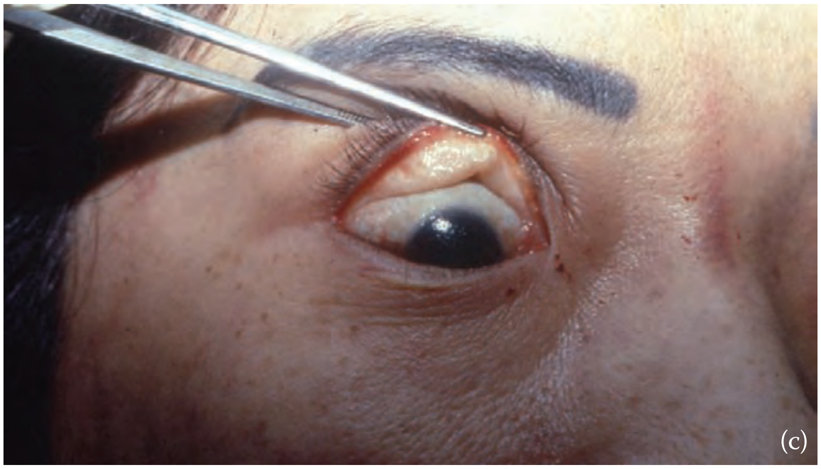

This Vietnamese woman died of Takayasu arteritis that was complicated by a ruptured aortic aneurysm. She was treated at home by her grandmother with Southeast Asian folklore remedies, including coining, which involves rubbing of hot oils and medicine onto skin (d). Note the unusual location of the coin marks at her neck (a) and (b). At first this was mistaken for manual strangulation. There were no internal neck injuries and no scleral or conjunctival hemorrhages (c).

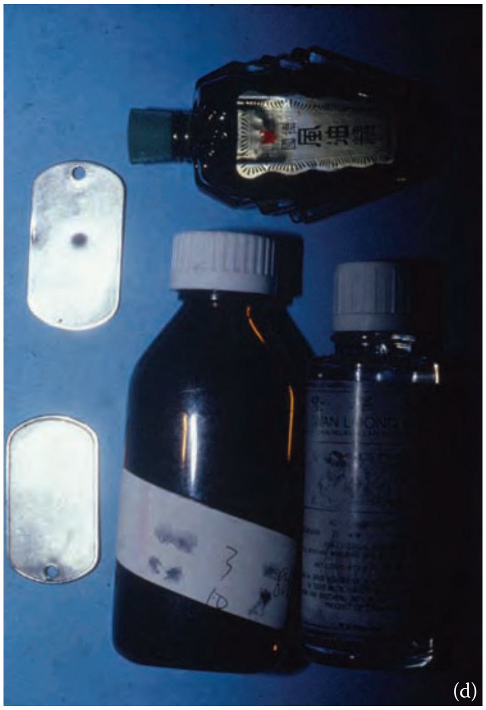

Substance Abuse and Poisoning

3

CHARLES A. CATANESE AND LAURA M. LABAY

Introduction

The term substance abuse is most simply defined as the excessive use of one or numerous drugs. In general, most people typically have some level of access to several types of drugs, including legitimately prescribed medications, and on average, will benefit from their appropriate use. However, the term substance abuse can be applied to a variety of situations or circumstances. For example, one person may intentionally be abusing a multitude of substances including illicit drugs, prescribed medications, and over-the-counter remedies (i.e., supplements, herbals, diet aids, etc.). This is in direct contrast to another individual who may be taking only a single medication, but deliberately using it in a manner that is not consistent with normal therapeutic use. It stands to reason that when individuals begin to abuse one or more substances, and use them in an uncontrolled manner without the appropriate oversight, that they place themselves in a potentially harmful or lethal situation.

In a forensic or medicolegal setting it often becomes a matter of necessity to interpret drug findings and render an opinion regarding the toxicological, physiological, and pathological impact the drugs may have had upon an individual. To this end, it is important to evaluate and consider a host of factors. These factors are not always straightforward or quantifiable, but nevertheless include the drug's inherent physical and chemical properties, the dosage of the drug used, the frequency of the drug intake, the route of drug administration, the concentration of drug and drug metabolite found, the person's tolerance level to the drug and any medical conditions or disease states the person may have experienced.

It is important to recognize that for a particular drug to ultimately produce a toxic or lethal effect, it must be present in an individual at a sufficient concentration for a sufficient length of time. However, while references and other texts are available that help to classify drug concentrations as "therapeutic," "toxic," or "lethal," interpretation is often not so simple. For example, postmortem methadone concentrations are often challenging to interpret because the range of blood concentrations detected in people enrolled in narcotic maintenance programs may overlap the blood concentrations found in overdose or lethal situations.

Also, some drugs, depending on their physical and chemical properties and their concentrations in the body, will exert their most toxic or lethal effects in an acute manner, while others will take a longer period of time to act. In general, deaths associated with substance abuse may be related to acute or chronic complications. *Acute substance abuse* indicates that the death was related to direct toxic effects of the drug shortly after administration. In a *delayed overdose,* the drug produces damage to the body over hours or days following acute intoxication, with complications that may include coma, sepsis, brain swelling, and herniation. This process may evolve over days with the drug(s) metabolized from the body. The underlying process that set off this sequence of events is the acute intoxication.

Depending on the type of drug and its half-life there may be no or minimal amounts of drug left in the blood or other tissues at the time of death. However, the possibility exists that the drug or its metabolite(s) could still be detected and identified in the urine. In these circumstances, it is important to take into account that urine is really a pooled specimen collected in the bladder over a period of time and that a quantitative result represents only the average drug–urine concentration over the period of time that the urine was produced. Therefore, this type of specimen does not accurately reflect the blood–drug concentration at any single point in time. Rather, a positive finding of a drug or drug metabolite in urine indicates prior exposure only to that particular drug and in this regard it is relevant to consider other information (e.g., case history, medical records, anatomical findings, etc.) as well.

If a decedent suffered trauma while intoxicated and developed an *epidural* or *subdural hemorrhage*, you would expect to find the drug present in these samples even after many days. *Chronic sequela* associated with substance abuse include complications of infections such as HIV disease, hepatitis, hepatic cirrhosis, and endocarditis. Injection of ground-up oral medications may lead to pulmonary failure, with multiple foreign body granulomas and fibrosis demonstrated on histopathologic section. There are many other complications, including Wernicke's-Korsakoff's encephalopathy, alcoholic cardiomyopathy, and hardening of the arteries due to cocaine use.

Intravenous drug abuse may present with many different features. If the drugs are injected acutely in the same vein during binges, they appear as multiple fresh needle marks in a row. These are called "track marks" because they resemble a pattern of railroad ties being laid down one after another. It may be possible to detect the parent drug by excising the immediate area surrounding the injection site and submitting this section for toxicological analysis. If several injection sites are observed and a decision is made to submit more than one for testing, make certain to package the specimens in separate containers, identify the sites from which each was taken, and note their age and appearance. Alternatively, the syringe or other drug paraphernalia such as vials, pipes, or spoons, if available, can also be submitted for analysis. This is a good option if, for whatever reason, limited biological specimens exist. Once the nonbiological contents of the items are identified, directed toxicological analysis on the biological specimens can then proceed. If this is done, it is important that the items be packaged in individual containers, away from biological specimens, so that contamination does not occur.

Injecting drugs into blood vessels can lead to the introduction of many types of infections. Sharing of needles may cause transmission of the hepatitis virus or the human immunodeficiency virus. Various types of bacteria can also be introduced during injection, leading to vasculitis, cellulitis, pneumonia, and endocarditis. Acute intravenous fatalities, such as those caused when individuals inject themselves, are usually classified as *accident*. If one can demonstrate the drug was given to the individual in order to purposefully cause his or her death, the manner would be classified as *homicide*. Also, if another individual injects the drug into the decedent's arm the manner of death should be homicide.

As the vein is chronically abused, it will scar and develop a chronic track mark. Chronic track marks appear as linear scars that traverse the path of underlying veins. Histological sections from these regions often show polarizable debris from impurities found in past-injected drugs. If a chronic intravenous drug abuser consistently rotates the injection sites from one vein to another chronic track marks will not develop.

Skin-popping refers to drugs, such as heroin, injected into the subcutaneous tissues. This usually indicates a long history of drug abuse with destruction of the previously accessible peripheral veins. This may appear as fresh needle marks, usually at the surface of the thighs, or in other places without an underlying vein. These sites often get infected and lead to cellulitis, abscess formation, and scarring.

Another route of drug administration includes *inhalation*. Inhalation (i.e., snorting) of drugs may lead to mucosal reddening from direct irritation. Crack pipes may become very hot during use and burn the drug abuser's mouth and lips. Sometimes histopathology sections of the lungs reveal numerous pigment-laden macrophages from constant inhalation of smoke and debris from inhaled burning drugs such as crack. *Chronic snorting* of drugs may lead to a perforated nasal septum.

Keep in mind, it is not only the illicit drugs such as cocaine or heroin that can be inhaled; an individual may choose to *crush* and *"snort" a tablet or pill* as well**.** In some instances, residue may be observed in the nasal or oral cavity, and these areas can be swabbed and the parent drug identified though analytical testing. Individuals may also choose to inhale or "huff" commercial products such as gases, fuels, aerosols, solvents, or propellants as a means of abuse. This may be accomplished by breathing in the fumes from a rag that has been soaked in the material or by placing a bag containing the fumes over the nose and mouth. In some cases, the inhalation may take place directly from the item (i.e., aerosol can, glue bottle), as well. Users have died from hypoxia, pneumonia, cardiac failure, and aspiration of vomit. If the latter situation is relevant to the case, it is important to inform the toxicology lab of this, as most routine toxicology screens do not include these substances in the scope of their analysis. For these cases, lung tissue or tracheal air, in addition to the routine biological samples, can be collected as well.

Transdermal drug delivery systems introduce medications into the general circulation through a slow diffusion process, and may be considered a desirable alternative to taking oral medications or using substances that require repeated injection. In addition, the transdermal systems, perhaps most importantly, help to minimize the extreme blood spikes and trough levels that may be experienced with orally administered medications. Instead, blood–drug concentrations are maintained at more consistent concentrations.

Examples of commonly encountered medications delivered through transdermal systems include nicotine, hormones (i.e., contraceptives) and the potent narcotic analgesic fentanyl. However, similar to any medication, these patch-style systems may be subject to abuse. For instance, a person may make the decision to apply a multiple number of patches to the body, or through manual or chemical means remove the drug from the drug reservoir or from the adhesive matrix. It is also not uncommon for a person to chew or even swallow them. In these cases, remnants of patches may be observed in the oral cavity or the gastric contents. Other signs of this type of abuse would be if a person presented with patches applied in a manner not consistent with normal therapeutic use. For these cases, document the number of patches found on the body, their locations, any writing or markings on the patches and any other descriptive features. Make an attempt to ascertain if all the patches represent a single acute application or if some of the patches may be from remote use.

In these types of death situations where patch use is known or suspected (i.e., old adhesive markings are noted on the body) it is important to make certain that specimens are not collected in proximity to a vein or artery that is in the immediate vicinity of a patch or patch marking. As a specific example, following the death of an individual wearing a fentanyl patch, some drug may still be present in the depot beneath the patch. If a blood specimen is collected from an area immediately beneath or near the location of the patch, the concentration of fentanyl may not accurately reflect the circulating concentration at or around the time of death.

Similar to the patch, suppositories represent another type of drug delivery system. Suppositories contain medicated material for insertion into a bodily passage or cavity such as the rectum. Once the suppository, typically a solid substance, is inserted, it will begin to dissolve over time and subsequently will deliver the medication into the body. Substance abuse using this route of administration is relatively rare.

Substance Abuse

Acute and Chronic Alcoholism

Death occurring from acute alcoholism is usually classified as *natural*, although this may not always be the case. A young adult at a party trying to impress friends may attempt to guzzle a liter of whisky. This is more than enough alcohol at one time to kill most adults. In this case, the manner of death would be accident. Sometimes depressed people with an acute triggering event, tell their family that they are going to “drink themselves to death.” After several days of drinking large quantities of alcohol, they die. If you can establish the intent to do harm, the manner of death would be *suicide*. If someone at a party drinks alcohol from a funnel and a hose, by pouring the alcohol themselves, and then dies, the manner of death would be accident. If someone holds the hose and pours the alcohol for them, the manner of death should be homicide.

The deaths of individuals with a history of chronic alcoholism, including varying degrees of liver disease, intoxications, and withdrawal, are mostly certified as natural. Chronic alcoholics who stop drinking abruptly may get the “shakes” and eventually experience seizures leading to death. Withdrawal of many other drugs without other underlying natural disease will not typically cause death.

It is very difficult to give an opinion about individuals’ state of mind or behavior while they were drinking or using any drug, for that matter. One has no way to know their true thought processes unless their behavior was witnessed or somehow clearly documented. Chronic alcoholics, with a high degree of tolerance, may be able to achieve a high blood alcohol level and not show visible signs or symptoms consistent with intoxication. Some alcoholics become aggressive, loud, and agitated while others become subdued and complacent.

It is important to obtain blood samples with care in cases where there is suspected alcohol use, and trauma played a role in the person’s death. If one obtains a blood sample only from the pericardial sac or chest cavity, where other visceral lacerations may exist, contamination with gastric contents or other bodily fluids is a likely possibility. In turn, the measured blood alcohol level may be markedly increased as compared with the actual circulating blood concentration at and around the time of death.

As a more specific example relating to alcohol, when death is a consequence of multiple traumatic injuries there is the possibility that significant damage to the internal organs occurred. These organs include, but are not necessarily limited to, the stomach, small intestine, and liver. If those organs, now damaged, contained unabsorbed alcohol, a blood specimen collected from the chest area may readily become contaminated. As a result, the concentration of alcohol determined may not accurately reflect the circulating concentration at and around the time of the fatal event. Instead, ideally, for postmortem alcohol analysis and to mitigate some of these issues, an alternative specimen type should be harvested. These specimen types would include a vitreous specimen, a urine specimen, or blood collected from a peripheral site such as the femoral vein. This is because, in those cases where the heart blood may have become contaminated due to events such as trauma, the alternative specimen can be used to help provide an interpretable alcohol concentration.

Alternative specimens, such as vitreous fluid, are vital to collect, especially when alcohol is suspected or involved. It is important to recognize that alcohol may form postmortem, and analytical testing cannot make the distinction between this type of alcohol and alcohol that was present prior to death. Indeed, an alcohol finding may represent a combination of these two circumstances. To prevent or minimize postmortem formation once blood is collected, it should be placed into an appropriate type of specimen collection container, one that contains preservatives that help to inhibit microbial growth. If the postmortem formation of alcohol is a concern, such as in a case where marked decomposition has occurred, a positive alcohol finding may be confirmed in the vitreous fluid, a specimen type that is more resistant to microbial growth.

Post-Mortem Redistribution

Besides contending only with the issue of contamination as caused by trauma, drugs and their metabolites may also be subject to postmortem redistribution. Postmortem redistribution is the movement of drugs among tissues, organs, and bodily fluids after death. The rate and extent of this movement varies according to several factors, including the nature of the drug, and the time interval between death and the postmortem collection time of specimens. Within the torso, the major organs constitute potential drug pools, and the gastrointestinal tract might contain considerable quantities of unabsorbed drug. Therefore, central blood is subject to redistribution from these local organs. In general, redistribution into central vessels is greater than redistribution into peripheral vessels and this is why it is preferred that final quantitative amounts be determined from a peripheral blood source. Therefore, while central blood pools are acceptable for screening purposes, it is always better to get a peripheral blood sample for quantitative confirmation work, if possible.

Cocaine

Cocaine, a Schedule II controlled substance, is found in the leaves of a South American shrub called *Erythroxylon coca* and is one of the most potent of the naturally occurring central nervous system (CNS) stimulants. First isolated in 1855, it has been used medicinally as a local anesthetic. However, because of its high potential for abuse, the use of cocaine for clinical situations has become severely limited. When cocaine is taken for illicit reasons it is either taken as the water-soluble hydrochloride salt by nasal insufflation (“snorting”), intravenous injection, or as the free-base (“crack”) by smoking. Regardless of the chemical form or route of administration, once cocaine is administered it is rapidly absorbed and distributed throughout the body. Once inside the body, the dosage form of cocaine cannot be distinguished and analytical determinations are reported as the free-base form.

Cocaine is rapidly biotransformed in the body to a few major metabolites and products including benzoylecgonine, ecgonine methyl ester, ethylecgonine, and ecgonine. These metabolites are all pharmacologically inactive. Small amounts of an active metabolite, norcocaine, may also be produced. However, this product is rarely detected in blood. Cocaethylene is a pharmacologically active substance formed in the liver when cocaine and ethanol are co-ingested. The most predominant cocaine product detected in the majority of biological specimens is benzoylecgonine. Elimination on half-lives are approximately 4.5 hours for benzoylecgonine and approximately 0.8 hours for cocaine.

Effects displayed by an individual under the influence of cocaine may include dilated pupils, increased blood pressure, increased pulse rate, and increased sense of strength or invincibility. It is believed that cocaine is toxic to the cardiovascular system causing thrombosis, myocardial infarction, tachycardia, or fibrillation that may occur in cases of acute and chronic abuse.

Heroin

Heroin is a Schedule I controlled substance and a synthetic derivative of morphine. It is made by first extracting morphine from opium and then chemically treating the morphine with acetic anhydride, sodium chloride, and hydrochloric acid.

Once heroin is taken into the body, most frequently by injection, smoking, or inhalation, it is rapidly deacetylated to 6-monoacetylmorphine (6-MAM), a product that is then hydrolyzed at somewhat of a slower rate than morphine. Unlike the heroin, which has little affinity for the opiate receptors in brain tissue, both 6-MAM and morphine are pharmacologically active. Because 6-MAM is a specific product of heroin, if it is found to be present in a biological specimen, it can be concluded that the individual either used or was exposed to heroin at some point prior to death. However, due in part to the short half-life of 6-MAM, morphine is most often the predominant species detected in biological specimens. Therefore, in cases where heroin is suspected to be the lethal agent and morphine is found in the blood but 6-MAM is not, it may be of benefit to test an alternate specimen type such as urine for the presence of this heroin specific marker as well.

The primary toxic manifestations of heroin use may last for approximately 4 to 6 hours and include the same effects most commonly associated with other opioids. Some of the more common effects include drowsiness, loss of coordination, decreased blood pressure, decreased pulse and respiration, mental clouding, sedation, and sweating. At sufficiently high levels, the user may slip into a coma and may ultimately stop breathing.

Lysergic acid diethylamide (LSD)

LSD, a Schedule I semisynthetic controlled substance, is manufactured from the main precursor chemicals lysergic acid, lysergic acid amide and ergotamine tartrate. LSD is normally taken by placing a "dot" laced with the material on the tongue. The LSD is then dissolved by the saliva and readily absorbed through the mucous membranes. This method of ingestion allows for its effects to be rapidly felt. Other means of ingestion include mixing the LSD with liquids or adding it to sugar cubes. However, the drug cannot be taken into the body by smoking, as pyrolysis destroys the LSD.

LSD is generally classified as a hallucinogen or psychedelic drug, and may produce both auditory and visual illusions. Approximately 30 to 60 minutes after ingestion, the user will experience the initial effects which, in general, may last for about 8 to 12 hours. Physiological effects are primarily sympathomimetic and may include mydriasis, hyperthermia, seizures, panic, and paranoid reactions. Flashback reactions, a brief recurrence of the LSD experience, are not uncommon in the experienced user, and may occur for weeks, months, or years after the last usage.

Death due to the pharmacological effects of LSD is rare, with most deaths occurring as a result of LSD-induced suicide and accidental trauma.

Marijuana

Marijuana, a Schedule I controlled substance, is a complex mixture of several products obtained from various parts of the *Cannabis sativa* plant, and is the most widely used illicit substance in the United States and the rest of the world. More than 400 chemical substances are found in marijuana. Sixty of these substances are called cannabinoids and are responsible for the psychoactive properties of the plant. The most relevant cannabinoid is tetrahydrocannabinol (THC), as it is the primary psychoactive ingredient in marijuana.

One of the most notable features of this drug is its long half-life, with some metabolic components exceeding 50 hours. This is because the drug is a highly lipid soluble and may undergo significant enterohepatic recirculation. In fact, the redistribution of THC from tissue to blood has been shown to be the rate-limiting step in its metabolism. In the body, THC is metabolized to two major metabolites, 11-hydroxy-THC (11-OH-THC) and tetrahydoxycarboxylic acid (THCC). The former metabolite is pharmacologically active, while the later is devoid of any pharmacological activity.

Marijuana is most frequently smoked, although it can be ingested as well. THC rapidly leaves the blood, even during a smoking period, and falls to below detectable levels within several hours. The most common physical effects are acceleration of heart rate, a moderate increase in blood pressure, a slight decrease in body temperature, reddening of the eyes, and a dryness of the mouth. The psychological effects of marijuana use include a pleasant feeling of well-being and euphoria, distortion of time, reduced ability to concentrate and memorize, and impaired short-term memory. Individuals under the influence of marijuana may have difficulties in tracking movement and demonstrate an inability to appropriately respond to stimuli. In general, this condition may persist for hours after the feelings of intoxication have dissipated, leaving users with a false sense of security concerning their abilities to safely operate a motor vehicle or machinery.

Death strictly due to the pharmacological effects of marijuana is not well documented, with most deaths occurring as a result of accidental trauma.

Methamphetamine

Methamphetamine is a Schedule II controlled substance with a very high potential for addiction and abuse. There are two different chemical forms or isomers of methamphetamine, each producing effects that differ in scope and magnitude. The l-isomer of methamphetamine may be found in over-the-counter nasal inhalers and is used for its vasoconstrictive properties. Compared with d-methamphetamine, it is a weak central nervous system stimulant. In contrast, the d-isomer may represent the licit or illicit forms of methamphetamine. In terms of legitimate or legal use, methamphetamine may be prescribed for a limited number of medical conditions such as weight loss, narcolepsy, and attention deficit disorder. However, because other less addictive and dangerous substances that do not quickly result in patient tolerance to the drug are also available, it is not that frequently prescribed.

Methamphetamine in the body undergoes demethylation to its primary active metabolite amphetamine; in most cases both methamphetamine and amphetamine will be detected. Analytical methods that differentiate the isomers of methamphetamine (and amphetamine) exist and may be employed if warranted.

People who abuse methamphetamine experience certain sequelae of such drug use. In general, the effects of methamphetamine can be broken down into three main stages. The first stage is the "high," where blood concentrations are at their greatest and people are feeling the stimulant effects of methamphetamine. The second stage is the "tweaking" period, where blood levels are on the decline and it is this period where people crave the drug and may behave in an aggressive and violent manner. The third stage is the "crash," where people feel exhausted and drained. High doses of methamphetamine can elicit restlessness, confusion, hallucinations, circulatory collapse, and convulsions.

Methadone

Methadone is a Schedule II controlled substance that is often prescribed during the process of narcotic detoxification, narcotic maintenance, and treatment programs, and to control severe and chronic pain. As compared with morphine, it produces less sedation and euphoria, but cessation of its use may result in withdrawal symptoms—not as severe as those seen with morphine, but longer in duration. Methadone works by decreasing the withdrawal symptoms felt by the narcotic abuser and, when a person attempts to reuse, the desired effects of the illicit drug are minimized.

In the body, methadone is metabolized to EDDP (2-ethylidene-1,5-dimethyl-3,3-diphenylpyrrolidine) and EMDP (2-ethyl-5-methyl-3,3-diphenylpyrroline). These metabolites do not possess any pharmacological activity and do not accumulate to an appreciable extent in plasma during therapy.

Methadone overdose is characterized by stupor, lethargy, pupillary constriction, hypotension, coma, respiratory collapse, and death.

3,4-Methylenedioxymethamphetamine (MDMA)

MDMA is a synthetic sympathomimetic compound with mixed stimulant, psychotropic, and hallucinogenic activities. It was used briefly as an adjunct to psychotherapy, but because of widespread abuse it has now been reclassified as a Schedule I controlled substance. The synthesis of MDMA is both complex and time-consuming. As a direct result, the final product may contain a variety of impurities that may be toxic in their own right. Also, drug manufacturers may add an illicit substance or a stimulant in an attempt to enhance the effects of the MDMA. The drug is available in tablet and powder form and it may be injected, inhaled, or ingested. Tablets of MDMA come in a litany of colors and shapes, and may be imprinted with a variety of images including things such as peace symbols, cartoon characters, butterflies, and angels.

In the body, MDMA is metabolized to many other compounds, but the main metabolite is a demethylated product called methylenedioxyamphetamine (MDA).

The effects of MDMA are related to the dose. A lower-level dose (approximately 50 mg) may result in feelings of enhanced creativity, while mid-level doses (approximately 100 mg) may make the user feel open to improved communication and empathy. The ability of the user to undergo self-exploration and analysis is achieved with the doses typically greater than 125 mg. In general, the psychological effects are much more pronounced than the physical effects. Abusers of the drug have been reported to experience longlasting neurobehavioral disorders following cessation of its use. Symptoms of MDMA toxicity include visual hallucinations, confusion, hypotension, agitation, coma, and death.

Phencyclidine (PCP)

PCP is a Schedule II controlled substance that was developed and marketed in the 1950s for use as an intravenous anesthetic. However, it was discontinued for medicinal use in 1965 as patients frequently would become agitated, delusional, and irrational, and experience the distortion of sights and sounds. In the early 1960s, PCP gained a reputation as a drug that produced certain desired feelings such as detachment and dissociation, but also caused the user to sometimes experience unwanted reactions (i.e., bad trips).

At low doses, PCP may cause changes in body awareness and produce psychological effects of euphoria, an alteration of time and space, confusion, bizarre behavior, and panic. Physical effects produced by PCP include impaired motor skills, shallow breathing, sweating, blank staring, speech disturbance, and an inability to regulate body temperature. At high doses, PCP may result in hallucinations, seizures, coma, and death. Because PCP has sedative-like effects, interactions with other central nervous system depressants, such as alcohol and benzodiazepines, may also lead to a life-threatening situation. In general, chronic PCP users may repeatedly use the drug for days at a time and during this period go without food or sleep. There appears to be no relation between plasma levels of phencyclidine and the degree of intoxication that a person may experience.

Oxycodone

Oxycodone is a Schedule II controlled semisynthetic narcotic analgesic derived from thebaine. It is used to control pain associated with such ailments as bursitis, injuries, simple fractures, and neuralgia, and is often found in combination with other drugs such as acetaminophen and aspirin. The addiction liability of oxycodone is about the same as for morphine.

Oxymorphone is a pharmacologically active metabolite of oxycodone that may be seen in blood in very low concentrations. Of interest is that oxymorphone may be prescribed as a parent drug and has a greater analgesic potency than morphine.

In overdose, oxycodone can produce stupor, coma, muscle flaccidity, severe respiratory depression, hypotension, and cardiac arrest. However, sustained-release preparations appear to produce adverse reactions, up to and including death, at lower-level concentrations, especially in combination with other central nervous system depressants, depending on use pattern and route of administration.

Poisons

Arsenic

Arsenic is a metalloid that is present in all parts of the environment and, for example, may be found in water, soil, and sediment. In broad terms, there are two main forms of arsenic —organic and inorganic. Organic arsenic is present in food, with crustaceans and fish being some of the richest sources. These organic forms of arsenic (arsenobentaine and arsenocholine) are considered to be relatively nontoxic and will be rapidly excreted unchanged in the urine. Inorganic arsenic occurs in two oxidation states: a trivalent form (arsenite) and a pentavalent form (arsenate), with the trivalent form being more toxic than the pentavalent form, which undergoes metabolism to monomethylarsonic acid (MMA) and dimethylarsenic acid (DMA). MMA and DMA are then excreted in the urine.

Arsenic inactivates up to 200 enzymes, most notably those involved in cellular energy pathways, and DNA replication and repair. Unbound arsenic also exerts its toxicity by generating reactive oxygen intermediates that cause lipid peroxidation and DNA damage. Inorganic arsenic binds thiol or sulfhydryl groups in tissue proteins of the liver, lungs, kidney, spleen, gastrointestinal mucosa, and keratin-rich tissues such as the skin, hair, and nails.

The lethal dose of arsenic in acute poisoning ranges from 100 mg to 300 mg. Severe acute arsenic intoxication produces several well-described symptoms. Bloody vomit and diarrhea may occur within 1 to 4 hours of ingestion. Gastrointestinal volume loss is compounded by profound capillary permeability produced by arsenic's interruption of cellular energy metabolism. Cerebral edema, microhemorrhage, encephalopathy, and seizures may also arise from loss of capillary integrity. The rate-corrected QT interval (QT_c) prolongation and tachyarrhythmias may develop.

Subacute arsenic toxicity involves predominately the neurologic and cardiovascular systems. Within days to weeks after ingestion, many untreated or undiagnosed patients describe debilitating peripheral neuropathy characterized by excruciating pain and severe motor weakness. Persistent QT_c prolongation and the accompanying risk of *torsades de pointes*, a specific type of cardiac arrhythmia, occur among patients with clinically significant body burdens of arsenic.

Chronic arsenic toxicity presents itself following months or years of exposure. Some hallmark features of this type of toxicity not described above include hyperpigmentation of the skin, hyperkeratosis of the hands and feet and the appearance of Mee's lines in the fingernails.

It is important to note that death may occur in all of the above situations (e.g., acute, subacute, and chronic) if the person is exposed to a sufficiently high dose of arsenic.

Carbon Monoxide

Carbon Monoxide (CO) is an odorless, colorless gas without taste that forms as the result of the incomplete combustion of carbon-containing material. Motor vehicles, appliances, and heaters that use carbon-based fuels are major sources of exposure. However, it is important to note that natural sources of carbon monoxide also exist. These sources include fire, gases emitted from mines, marine algae, and human metabolism. Carbon monoxide is endogenously produced when hemoglobin, the molecule responsible for oxygen transport, and other heme-containing substances are degraded or broken down. Because of this, endogenous levels of CO, analytically measured as carboxyhemoglobin, are typically less than 1%. It is important to note that CO levels within the body may vary depending on several other factors as well. For example, because cigarette smoke contains CO, a smoker may exhibit carboxyhemoglobin levels as high as 8% saturation.

A person becomes exposed to CO via inhalation with the ultimate biological saturation level dependent upon several factors including CO concentration, duration of exposure and the activity level of the individual. Carbon monoxide poisoning produces hypoxia by two main mechanisms of action. First, CO binds to hemoglobin with an affinity that is greater than 200 times that of oxygen and therefore, by occupying the oxygen binding sites of hemoglobin, CO directly decreases the oxygen-carrying capacity of blood. Second, when CO binds to hemoglobin, the hemoglobin undergoes a change in its configuration so that oxygen release from the hemoglobin is hindered. Early signs of CO poisoning include headache, nausea, and vomiting. As the CO poisoning progresses, the person may experience impaired mental function, an inability to concentrate, and personality changes. Finally, the individual may develop seizures, coma, and death. Classic pathological signs that are most often associated with CO poisoning, although rarely observed, include cherry red skin and retinal hemorrhages.

Cyanide

Cyanide is a potent, rapidly acting lethal poison, and death may occur within minutes following its ingestion.

Common sources of cyanide include industrial manufacturing byproducts, plants, fruit pits, chemicals, and combustion products of certain plastics. Because of the latter, cyanide may play a role in the hypoxic events from fires.

Cyanide exerts its effects by disrupting electron transport at the cytochrome c oxidase step and this in turn adversely impacts the production of adenosine triphosphate (ATP). This break in the oxidative phosphorylation process stops the Kreb's cycle, and ultimately causes a metabolic acidosis as both pyruvic and lactic acid begin to accumulate.

The signs and symptoms of toxicity are dependent upon several factors. These factors include the form of the cyanide (e.g., gas versus solid), the route of exposure (e.g., inhalation versus ingestion), the duration and the extent of the exposure. The minimum lethal dose in an adult has been estimated to be 100 mg for hydrocyanic acid and 200 mg for potassium cyanide.

Cyanide produces a range of symptoms including dizziness, weakness, motor impairment, and mental impairment. These symptoms may progress toward slowed respiration, lactic acidosis, seizures, coma, and death.

It is important to recognize that blood concentrations of cyanide can increase or decrease during storage depending on the length of time, the temperature, and the presence of cyanogenic bacteria.

Ethylene Glycol

Ethylene Glycol is a nonvolatile liquid that is a common ingredient of automotive products such as antifreeze, de-icers, and coolants. It can also be found in some preservatives and as a glycerin substitute.

Following ingestion, ethylene glycol is rapidly absorbed, and manifestations of toxicity may be noted within approximately 30 minutes. Ethylene glycol is metabolized in the liver to several toxic metabolites including glycoaldehyde, glycolic acid, glyoxylic acid, and oxalic acid, and it is these metabolites that then may elicit central nervous system, cardiopulmonary, and renal dysfunction as well as produce a severe metabolic acidosis. The magnitude of the toxicity is dependent upon dose and the onset and success of treatment.

The effects of an ethylene glycol exposure are typically described in three main stages. In the first stage (0.5 to 12 hours post ingestion), neurological symptoms that include signs consistent with ethanol intoxication are manifested. Coma, convulsions, and possibly death may result in this stage. The second stage (12 to 24 hours post-ingestion) is often characterized by cardiopulmonary disturbances including tachycardia, tachypnea, and hypertension. In severe ingestions, congestive heart failure, pulmonary edema, and circulatory collapse may be seen. The end stage of ethylene glycol toxicity is renal failure (24 to 72 hours post-ingestion).

Methanol

Methanol is a type of alcohol that is extremely versatile, and so can be found as a component in industrial solvents, fuels, and antifreeze preparations, and as a denaturant for ethanol. Methanol may be introduced into the body through multiple routes of administration including inhalation and ingestion. Once in the body, methanol is first metabolized to formaldehyde and then to formic acid. Interestingly, although formaldehyde is considered a toxic substance, it has on average a half-life of only several minutes and it is the formic acid that ultimately produces the hallmark features of methanol toxicity, including metabolic acidosis and the loss of visual acuity. Methanol poisoning can be treated by the administration of ethanol, because both methanol and ethanol share a common metabolic pathway. In essence, the ethanol competes for binding to the alcohol dehydrogenase enzyme and indirectly limits the formation of formic acid. Acute methanol exposure may produce severe signs and symptoms of toxicity, including nausea, abdominal pain, and lethargy. In some cases, the poisoning will progress to where the person experiences an anion gap metabolic acidosis leading to coma, seizure, and respiratory collapse. It is important to recognize that methanol may be a component found in embalming fluids and, because of this, tissues or other samples that have come into contact with this type of material may test positive for the presence of methanol.

Poison Hemlock

Poison hemlock (*Conium Maculatum*) is a biennial member of the carrot family that grows wild throughout the United States, especially along roadsides. Socrates is thought to have died from ingestion of poison hemlock. Its toxicity is from several simple piperidine alkaloids including coniine, gamma-coniceine, conhydrine, N-methylconiine, and pseudoconhydrine. Coniine and gamma-coniceine are thought to have the most significant contribution to the toxic effects, which are similar to nicotine poisoning.

Thallium

Thallium is a metal that was once used in rodenticides, insecticides, and depilatories, but was banned from residential use in the 1970s. Today, it is used in the semiconductor industry and may be found in switches and electronic devices. Thallium is readily absorbed from the gastrointestinal tract and may be detected in most of the body's tissues and fluids, including blood, brain, liver, kidney, spleen, bone, hair, and urine. The half-life of thallium in the blood is approximately 2 to 4 days, with one of the hallmark symptoms of this type of poisoning, alopecia or loss of hair, occurring after about 1 to 3 weeks. Symptoms of acute exposure include severe gastrointestinal distress, tingling of the hands and feet, paralysis, and respiratory failure. In people being chronically exposed, additional signs and symptoms include paralysis, hepatic and renal issues, and respiratory failure.

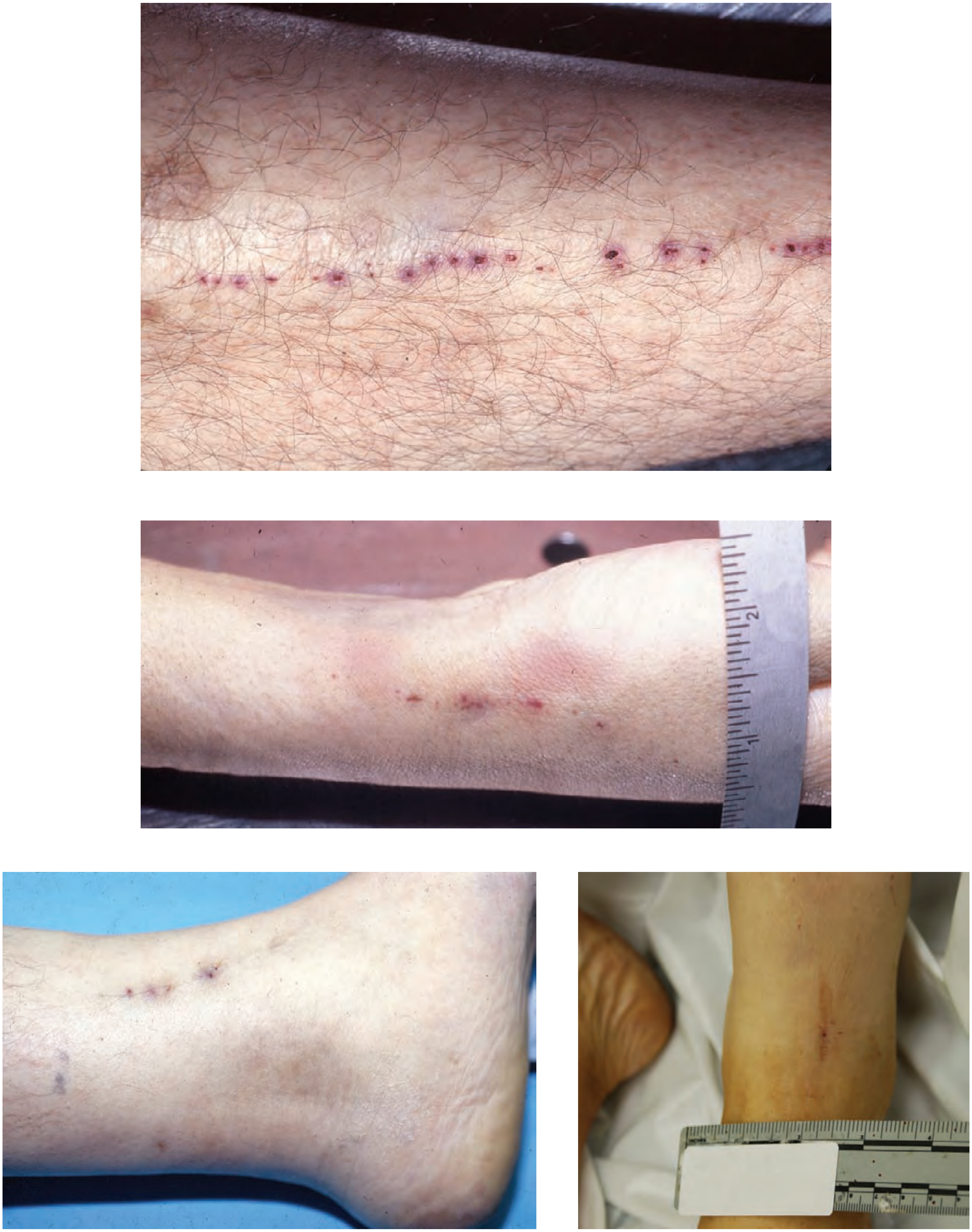

Acute track marks. Multiple fresh needle mark injection sites that traverse the path of underlying veins. Incision of these regions will often reveal underlying hemorrhage. Toxicological testing of underlying injection site tissue may reveal the parent compounds, such as heroine.

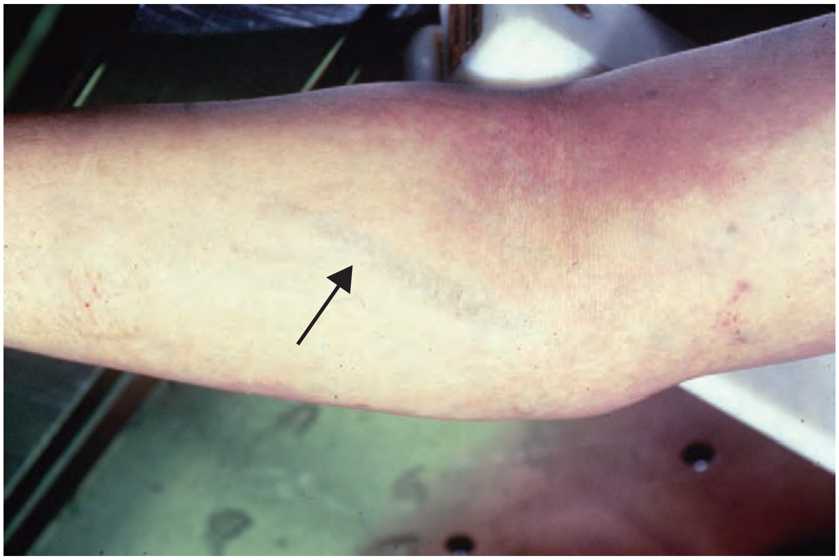

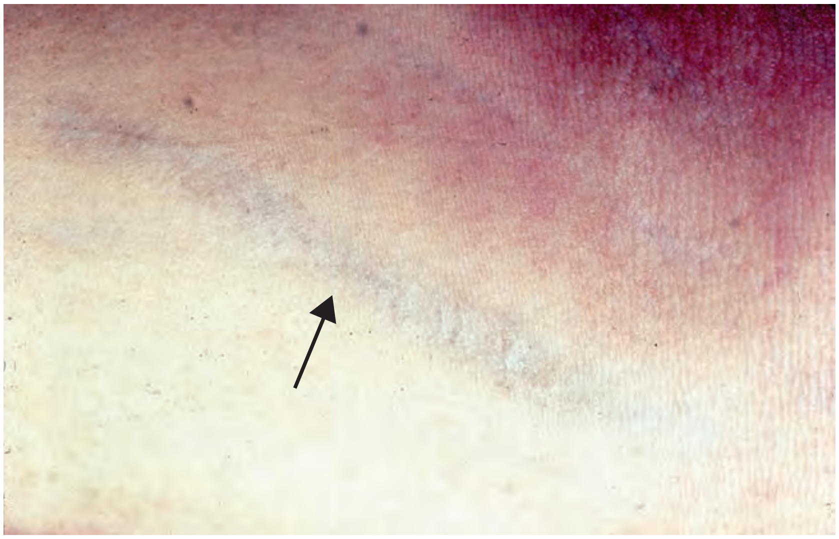

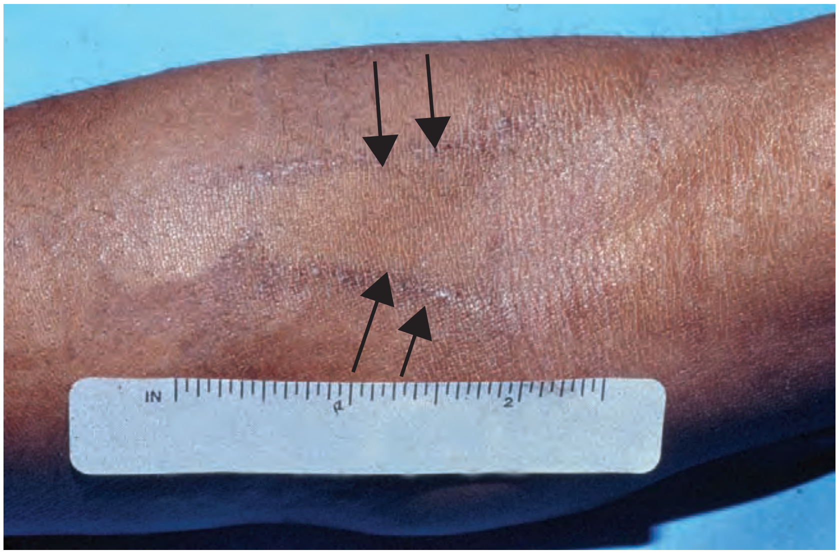

Chronic track marks. Linear scars that traverse the path of underlying veins. Microscopic sections may demonstrate inorganic debris in sub-adjacent soft tissues.

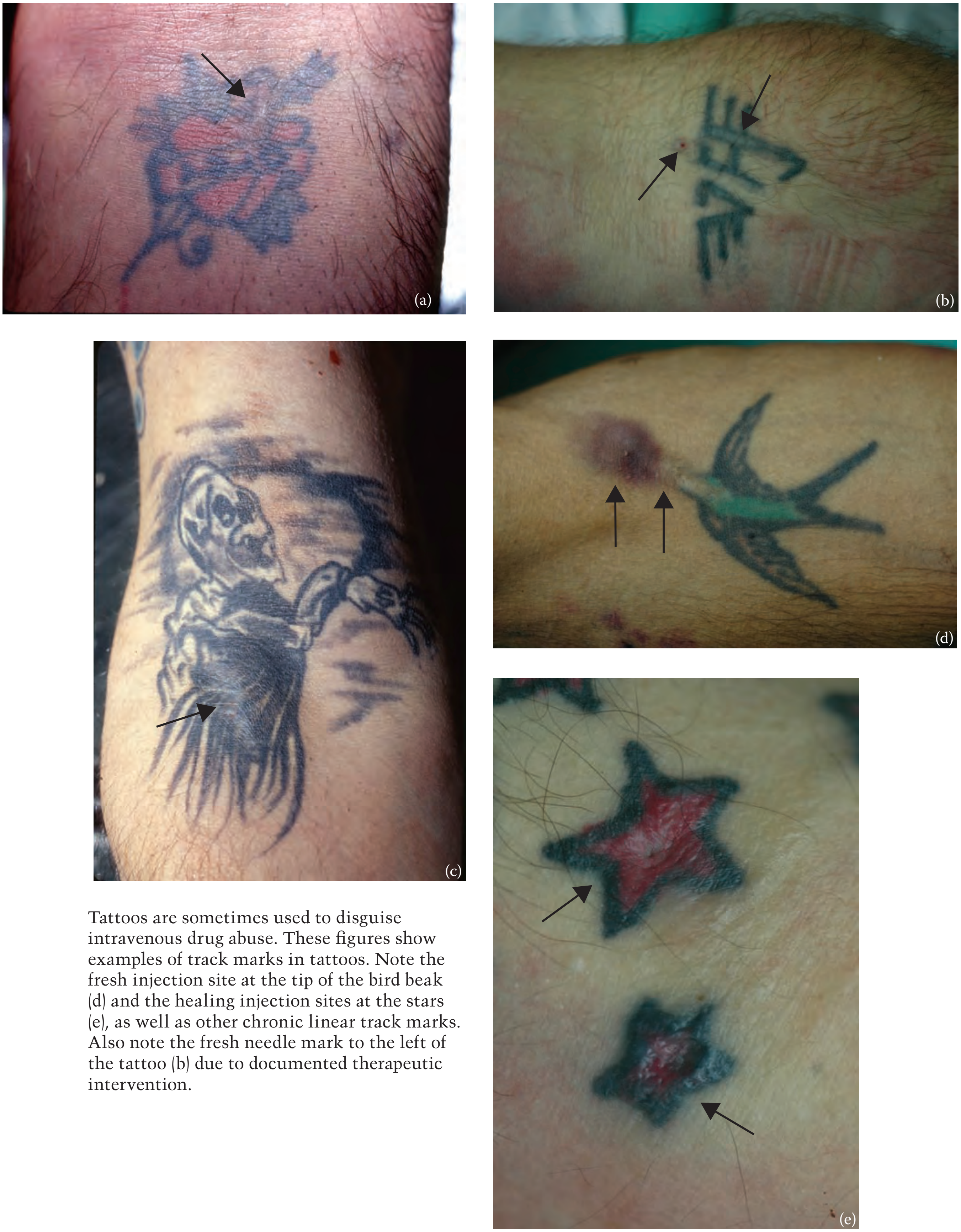

Tattoos are sometimes used to disguise intravenous drug abuse. These figures show examples of track marks in tattoos. Note the fresh injection site at the tip of the bird beak (d) and the healing injection sites at the stars (e), as well as other chronic linear track marks. Also note the fresh needle mark to the left of the tattoo (b) due to documented therapeutic intervention.

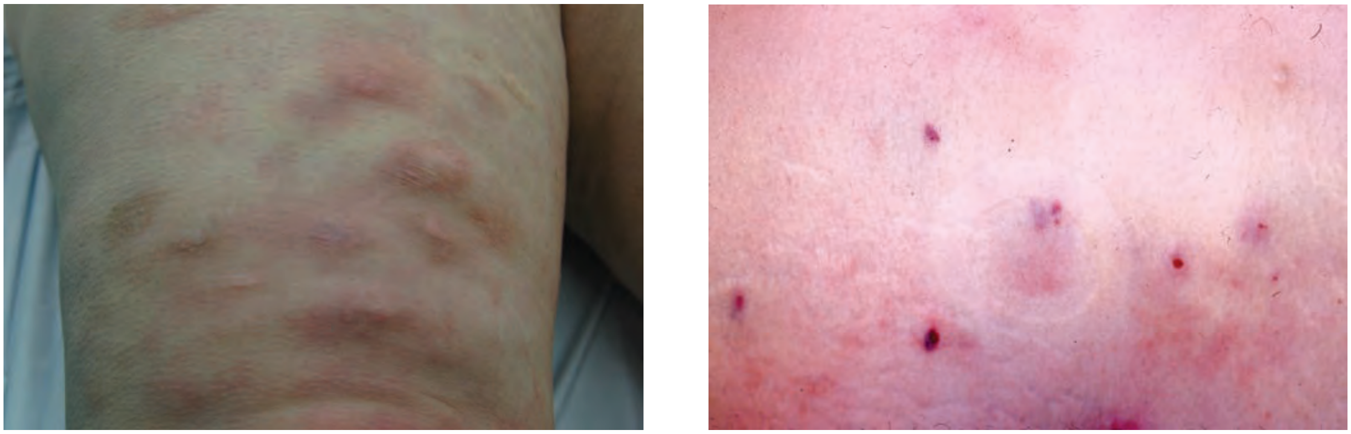

Fresh "skin popping" lesions. When drug abusers exhaust peripheral vein access they may start injecting drugs subcutaneously.

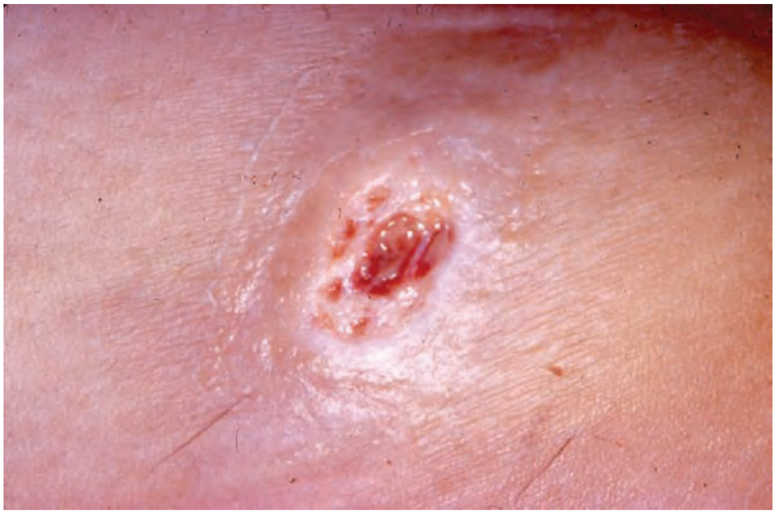

Recent healing infected skin popping lesion with adjacent cellulitis. It is common for these lesions to become infected due to the nonsterile nature of the injection equipment.

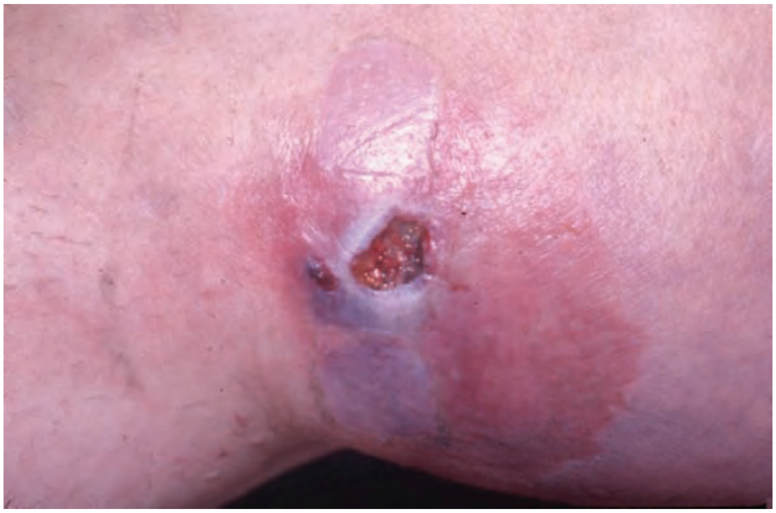

Recent healing infected skin popping lesion.

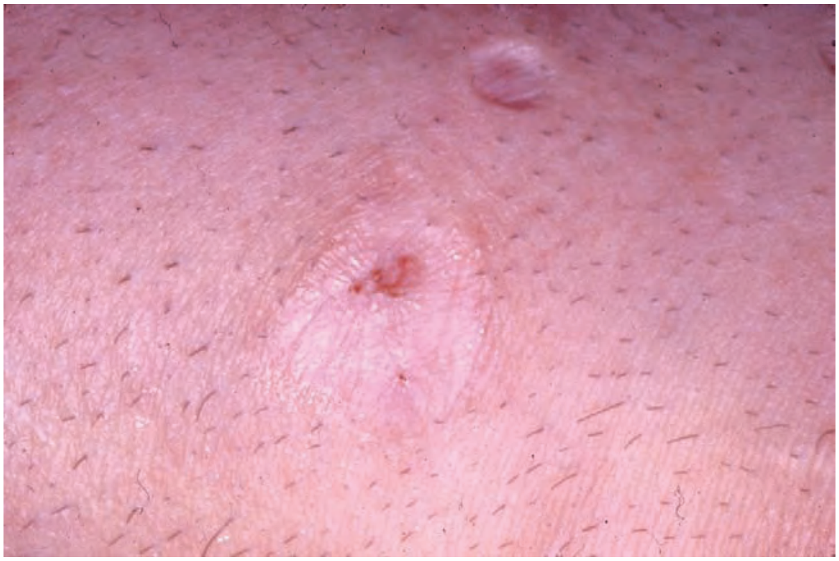

Healed and almost completely healed skin popping lesions. Note the lower wound has almost complete replacement by scar except for the central healing defect.

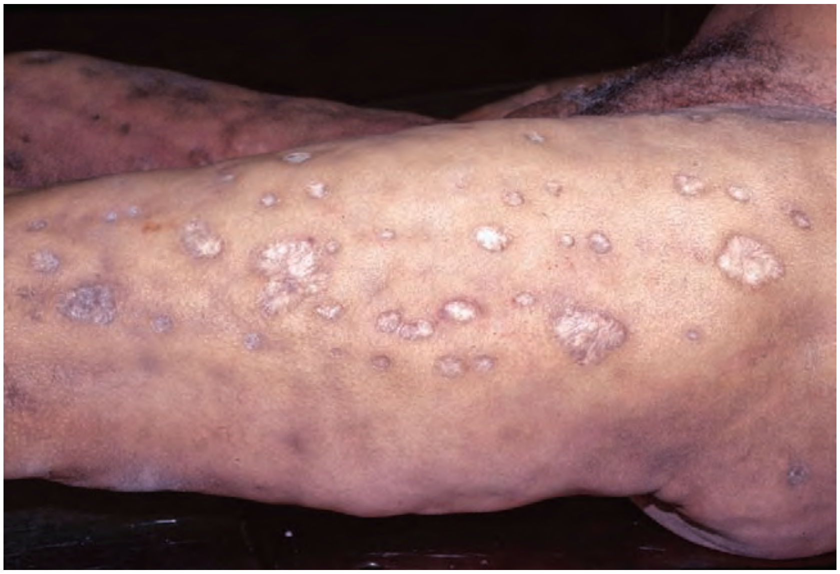

Remote old healed skin popping lesions.

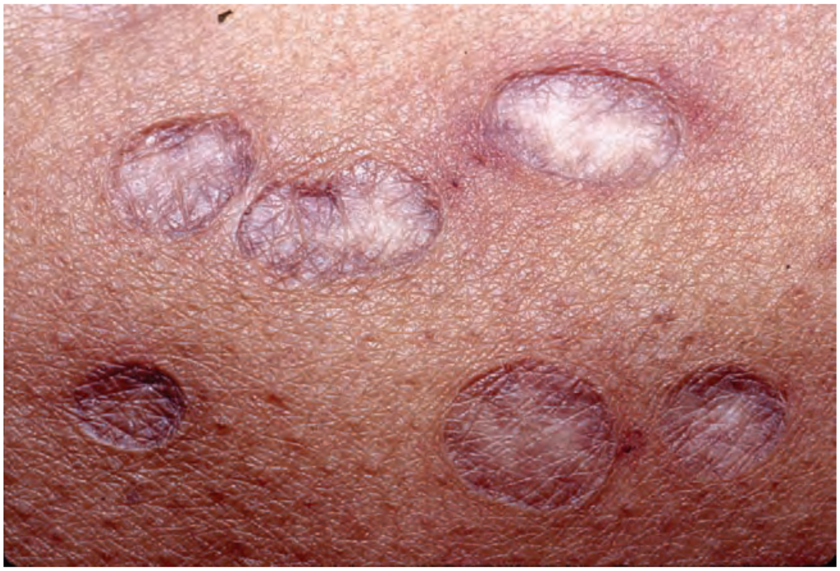

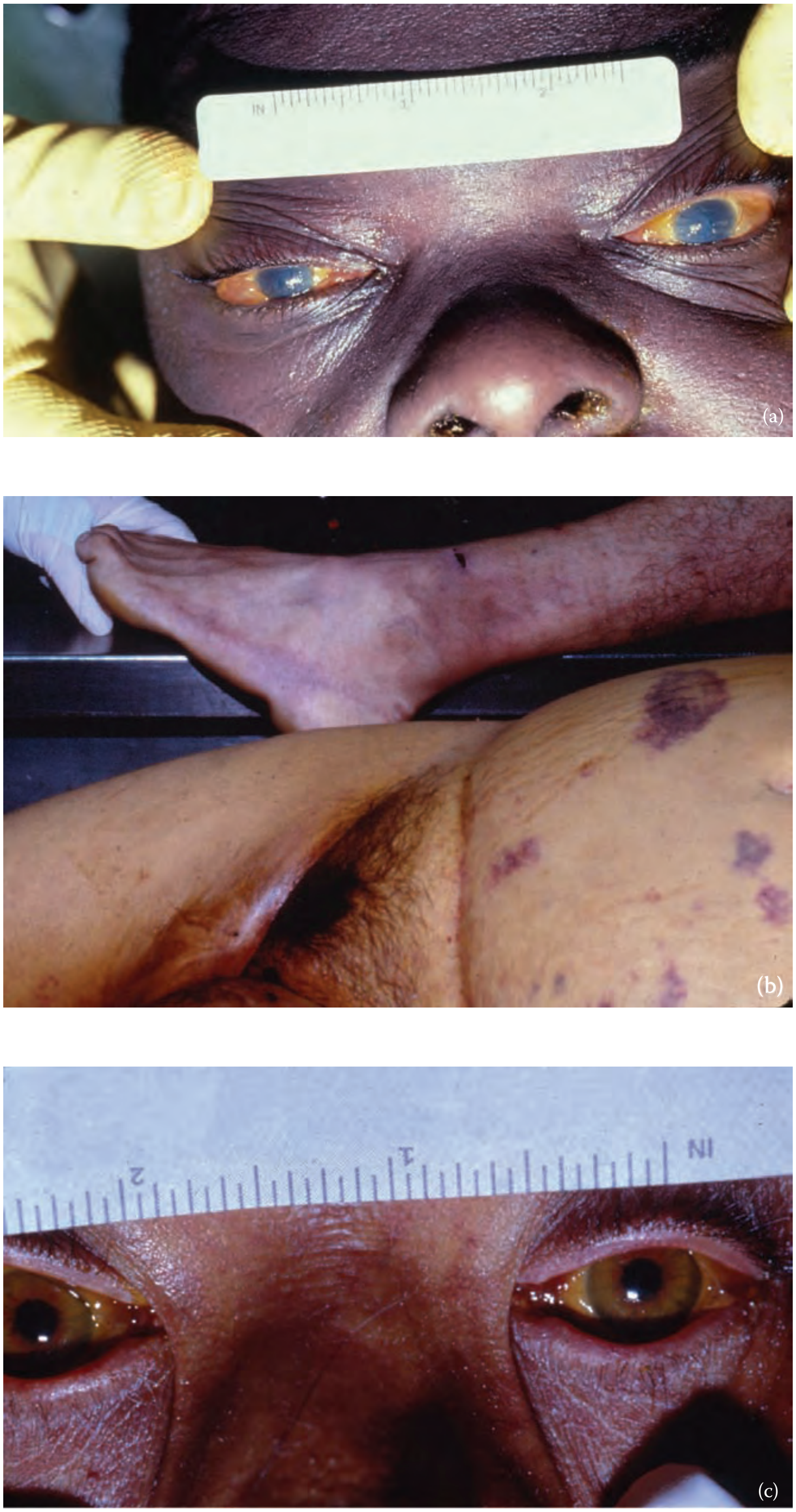

Examples of jaundice following hepatic failure due to hepatic cirrhosis as the result of chronic alcoholism; and viral hepatitis caused by chronic intravenous drug abuse. Note the yellow discoloration in the sclera and skin (a) and (c). Figure (b) demonstrates a comparison to a nonjaundiced individual. Also note the patchy areas of ecchymosis due to coagulopathy associated with the liver disease.

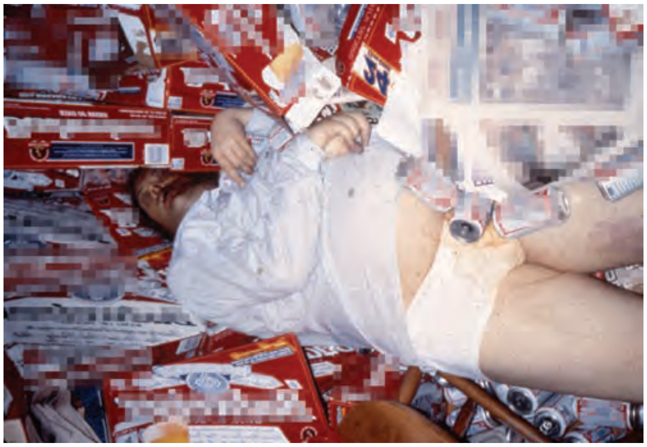

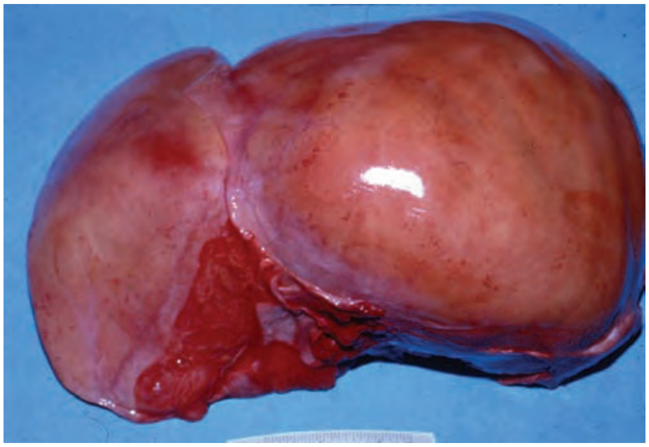

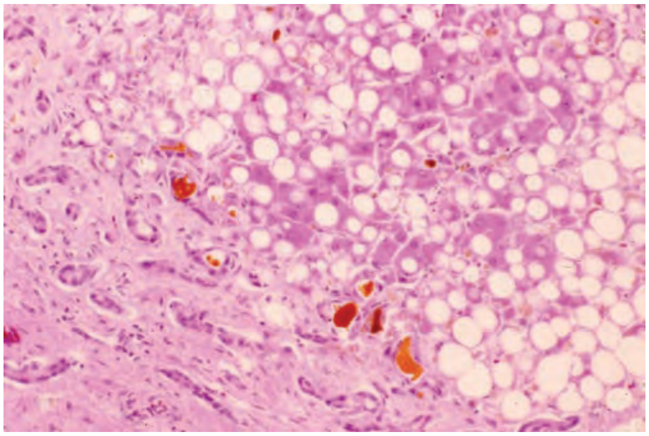

Marked hepatic steatosis due to acute alcoholism.

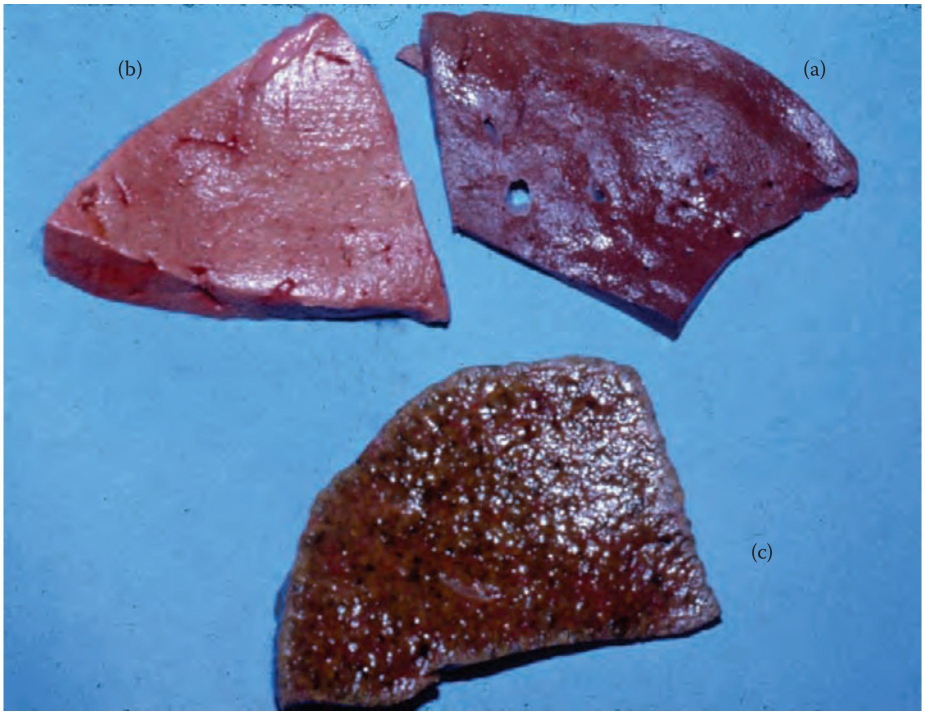

This image demonstrates cut portions of liver. Normal liver is brown (a); the yellow section demonstrates marked steatosis or fatty liver (b). The green section demonstrates micronodular cirrhosis due to chronic alcoholism with inspisated bile (c). The bile imparts this green discoloration.

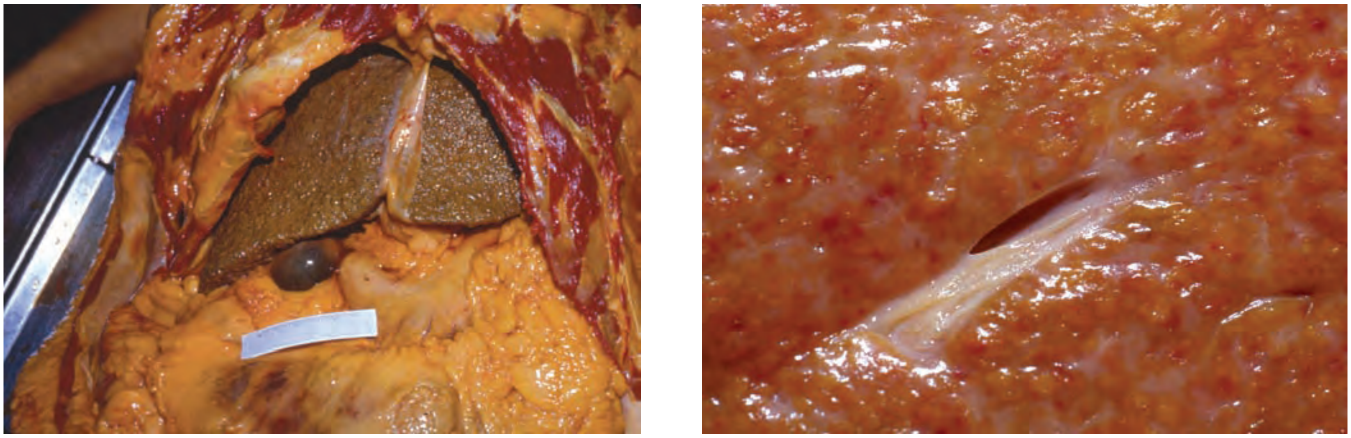

Micronodular hepatic cirrhosis with steatosis due to chronic alcoholism.

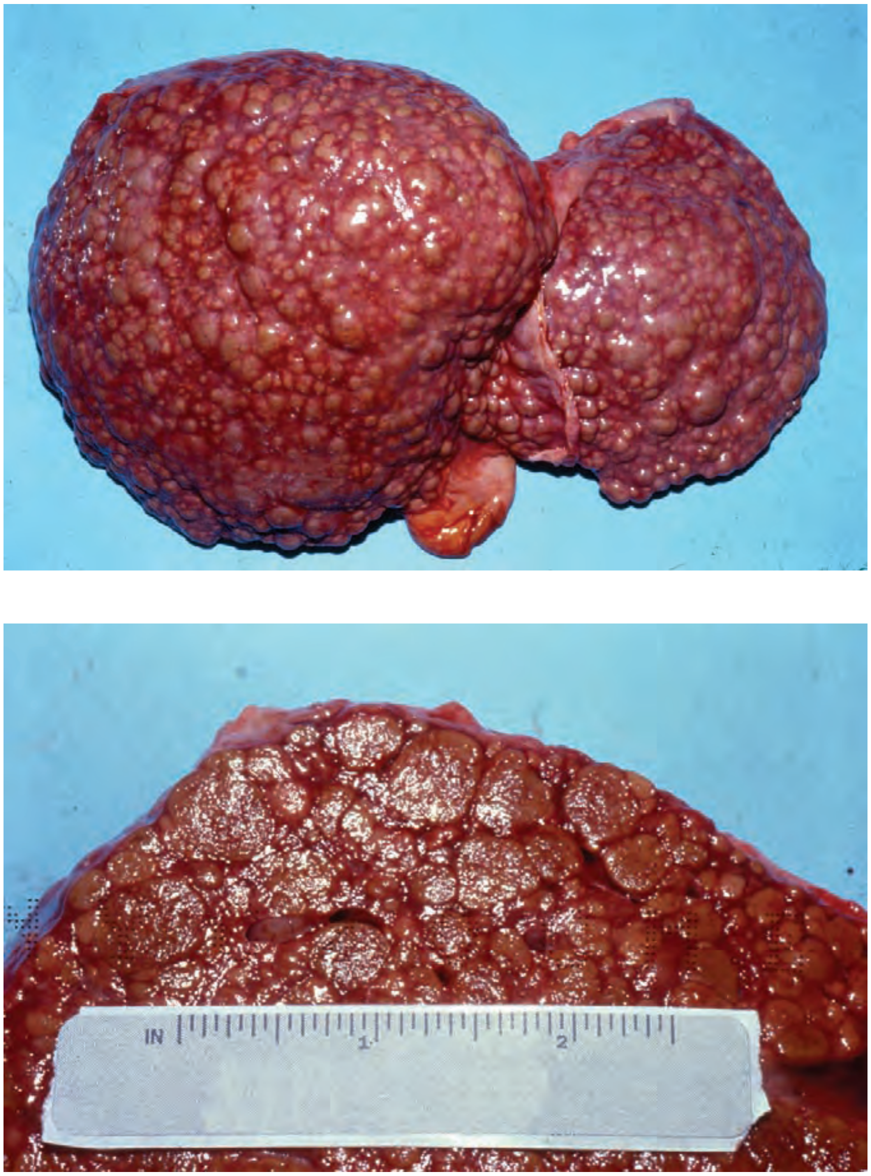

Macronodular hepatic cirrhosis due to hepatitis B infection as a result of chronic intravenous drug abuse.

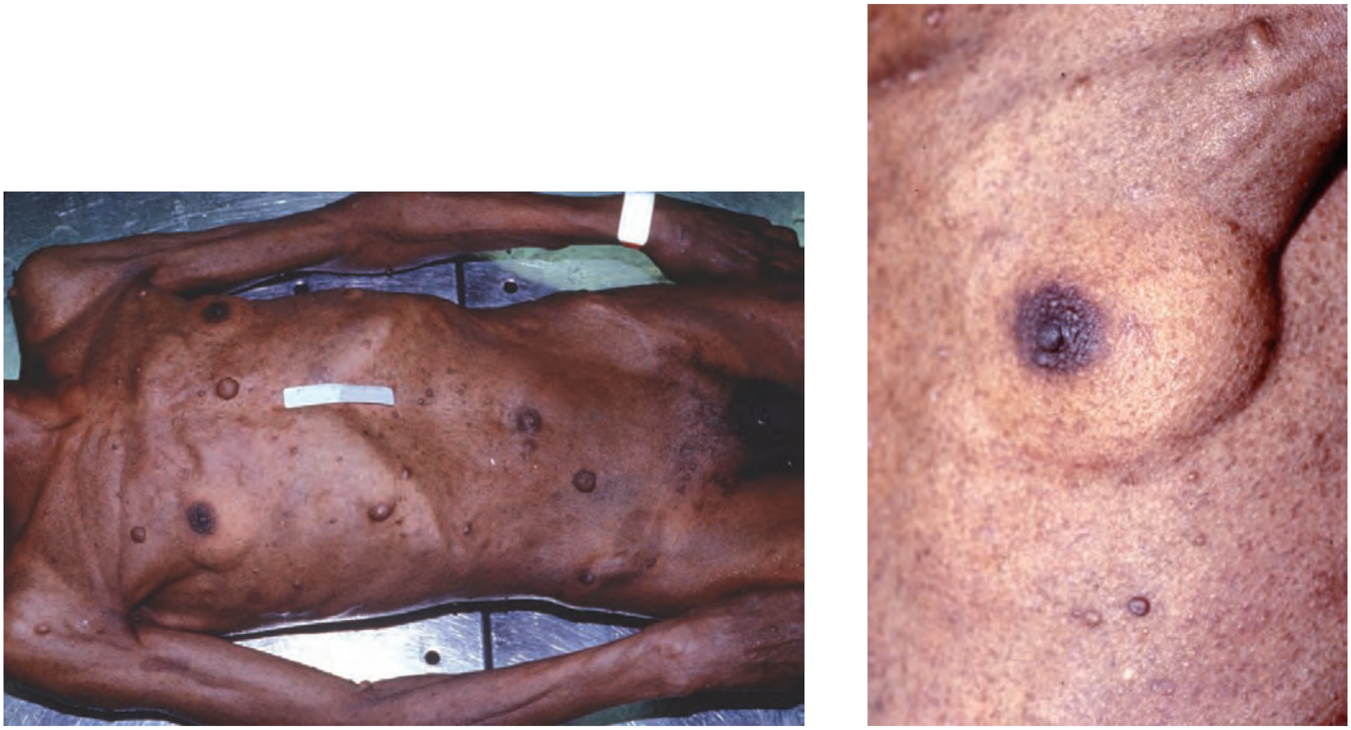

Chronic alcoholic with hepatic cirrhosis and gynecomastia. This individual also had neurofibromatosis, which is demonstrated by the multiple subcutaneous nodules.

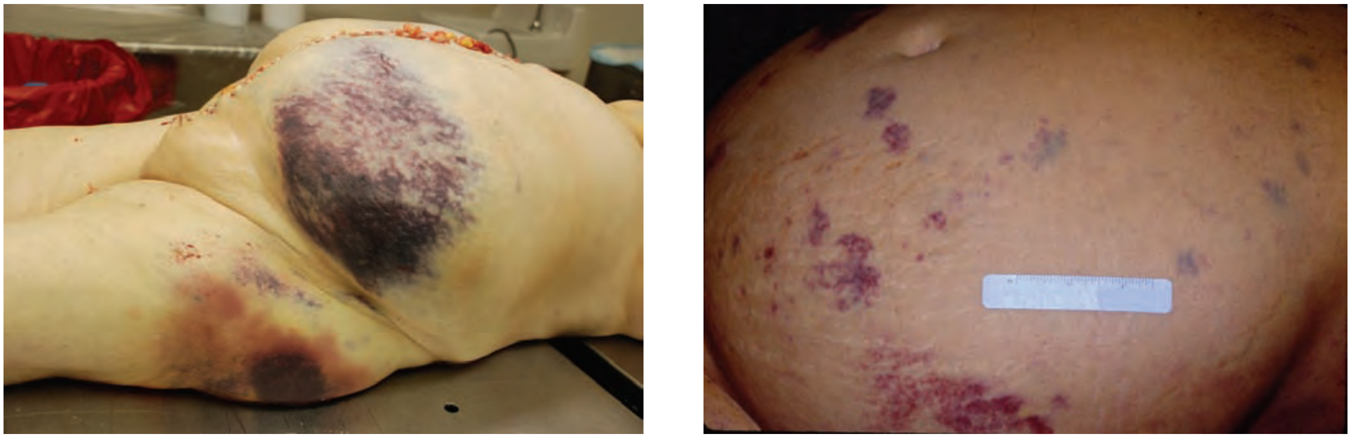

Chronic alcoholic with hepatic cirrhosis and coagulopathy. Note the multiple areas of ecchymosis and contusion caused by minimal blunt force trauma.

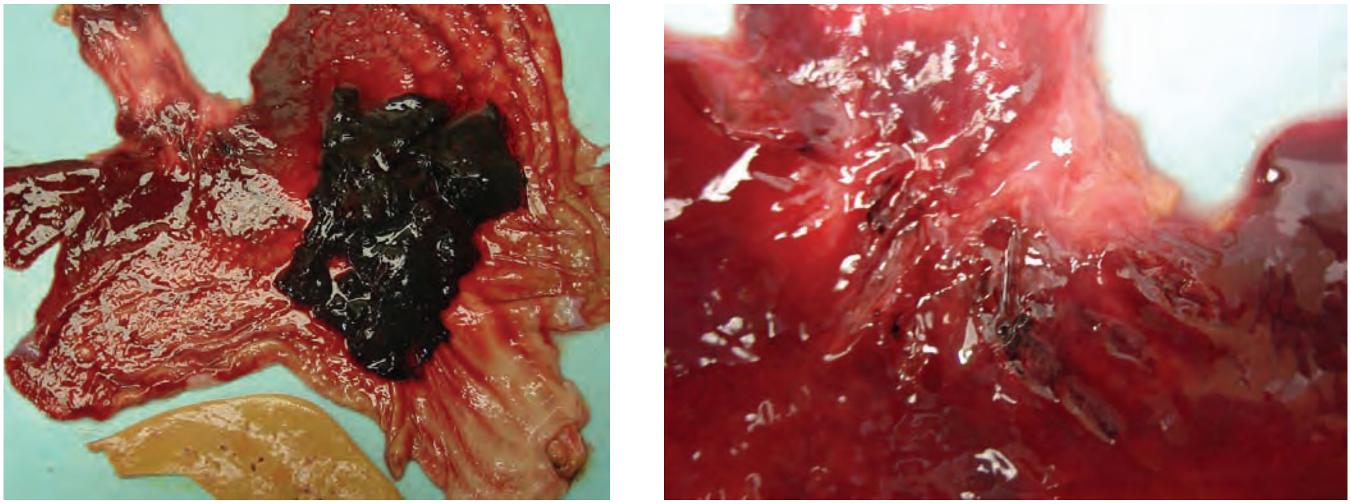

Gastrointestinal hemorrhage due to Mallory-Weiss tears in a chronic alcoholic. Note the laceration of the gastroesophageal junction leading to death from gastrointestinal hemorrhage following multiple episodes of vomiting.

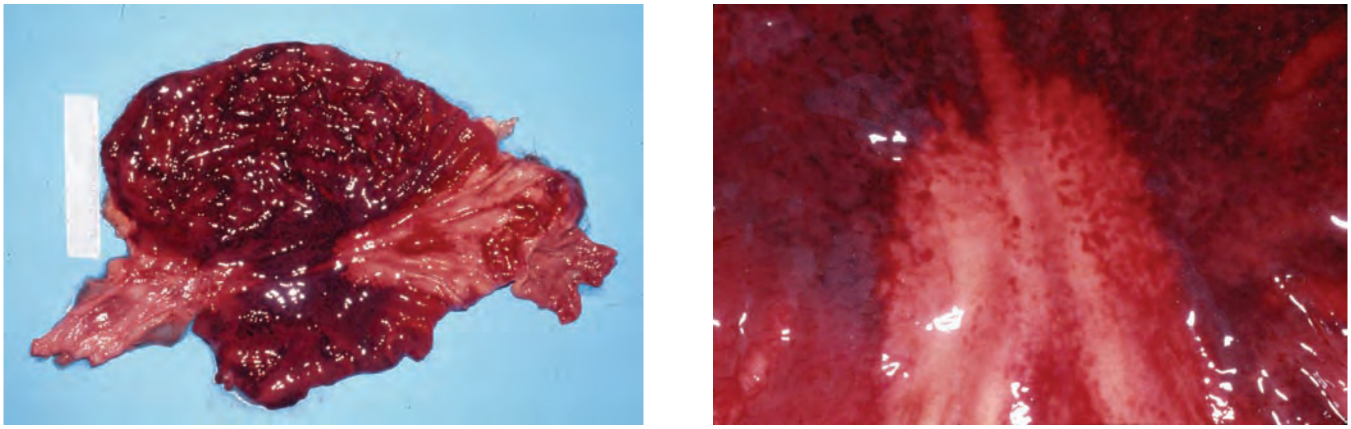

Acute gastritis associated with chronic alcoholism. Note the red discoloration of the mucosa.

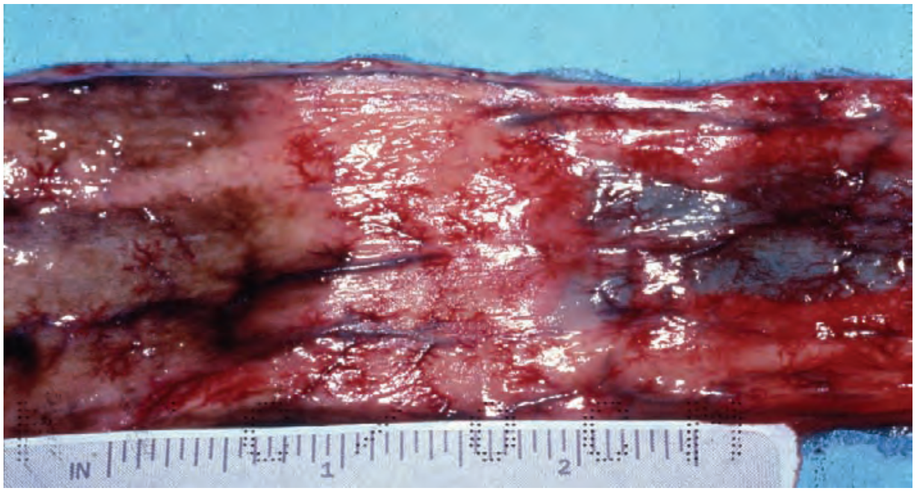

Esophageal varices due to hepatic cirrhosis and increased portal hypertension. Note the dilated red to brown submucosal veins that may rupture and possibly cause exsanguination.

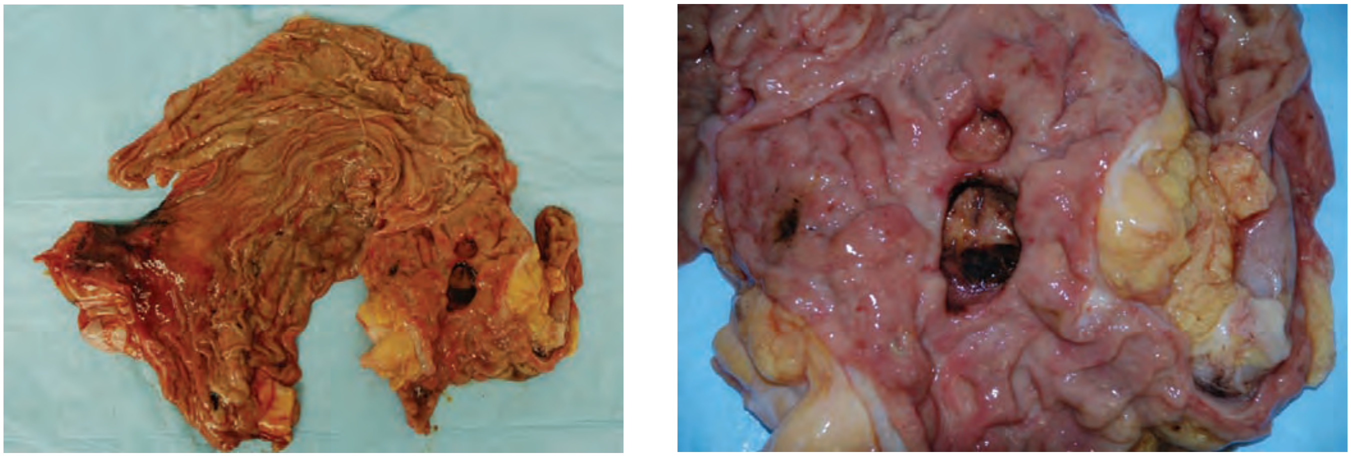

Bleeding peptic ulcers in a chronic alcoholic with gastrointestinal hemorrhage and hepatic cirrhosis.

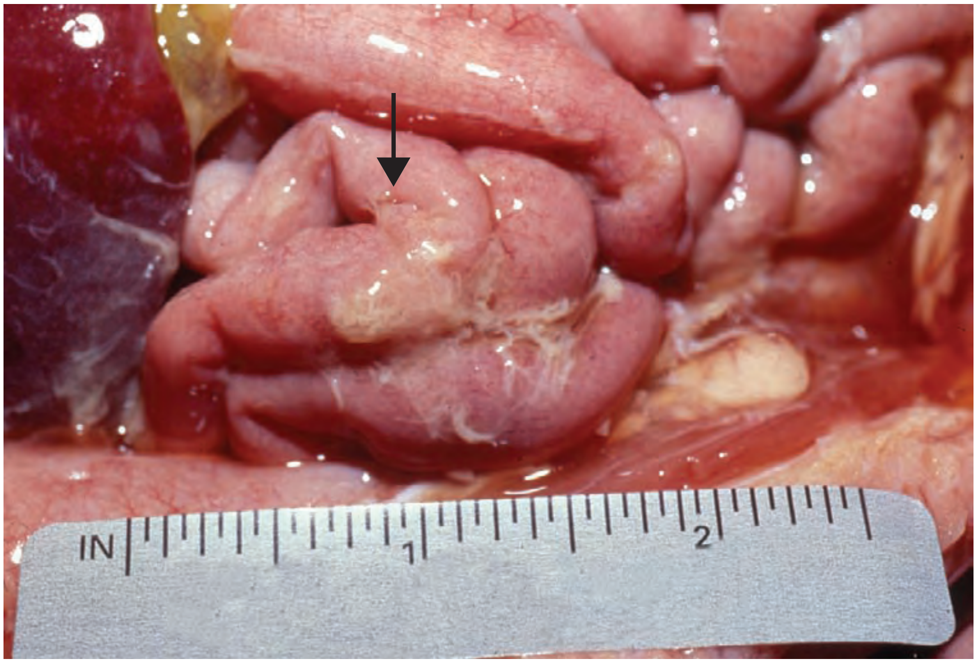

Acute peritonitis in a chronic alcoholic. Note the purulent exudate at the intestinal surface.

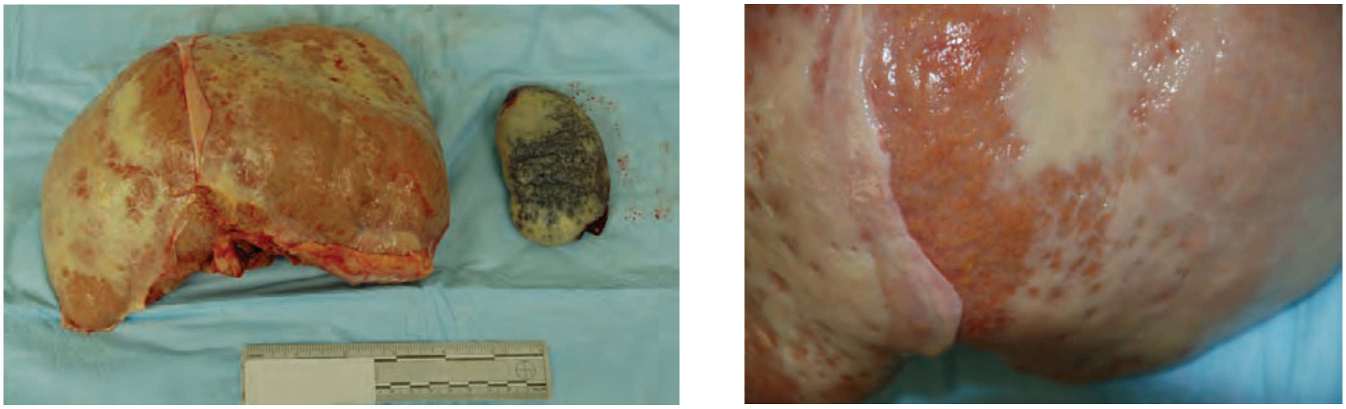

Hyaloserositis. Old healed peritonitis with fibrosis at the surface of the liver and spleen in a chronic alcoholic.

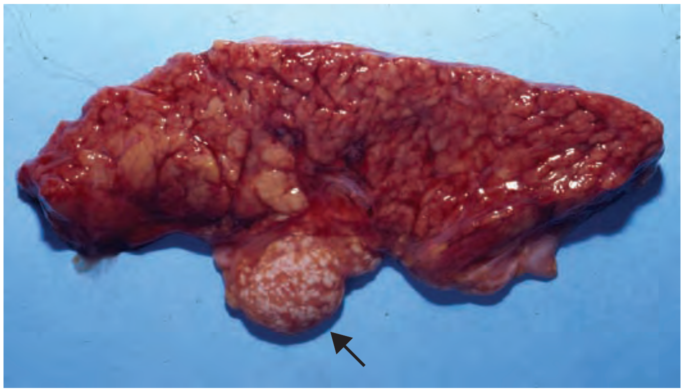

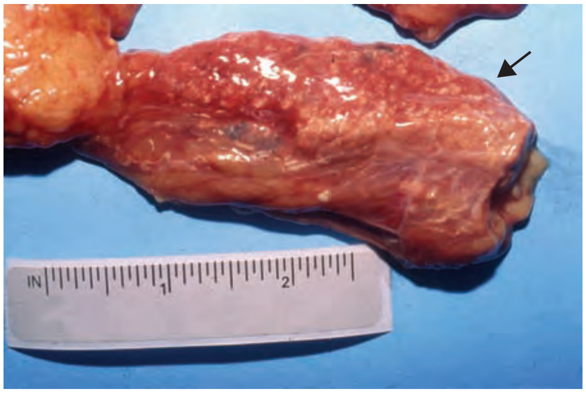

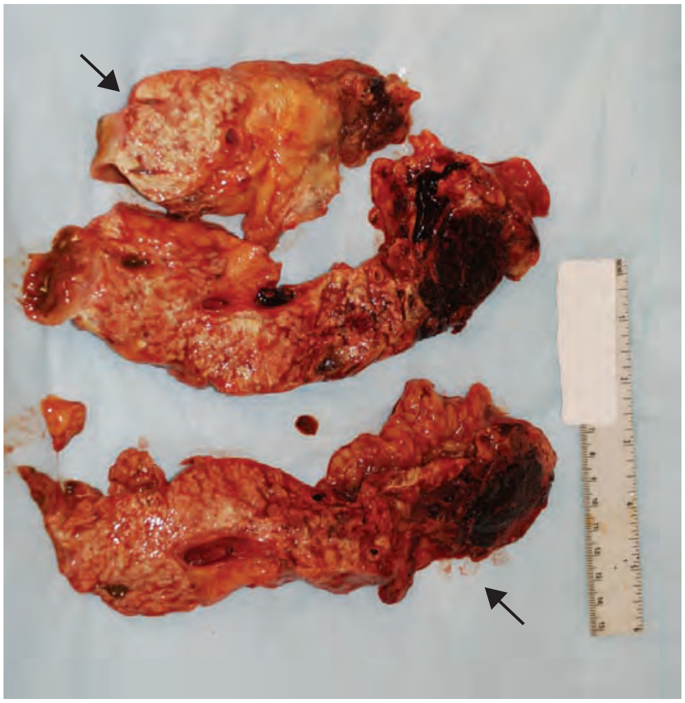

Acute pancreatitis due to chronic alcoholism. Note the white flecks due to fat necrosis and areas of hemorrhage.

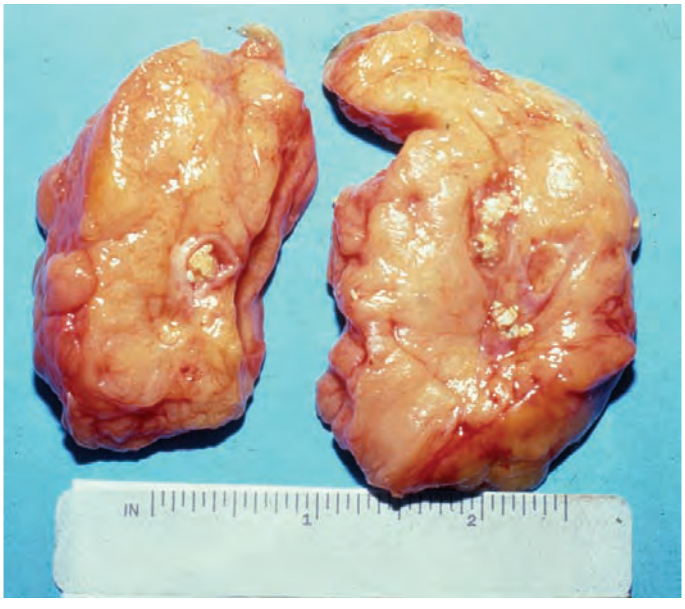

Fatty metamorphosis or atrophy of the pancreas with duct concretions in a chronic alcoholic.

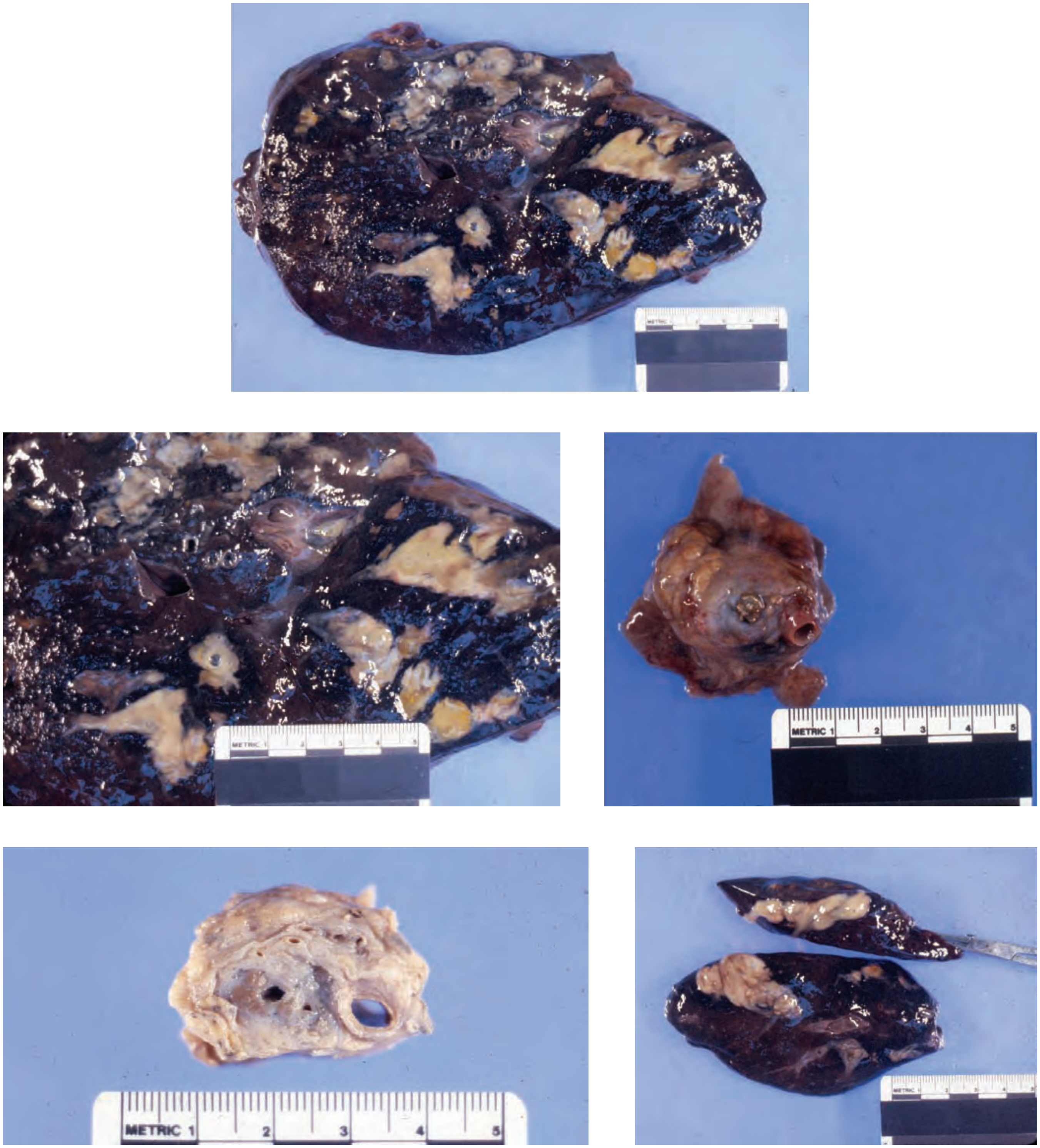

Pyogenic liver abscesses. This decedent had a history of chronic alcoholism with chronic pancreatitis and recent abdominal pain of unknown etiology. This photo depicts multiple geographic, creamy yellow-white collections of purulent material within the liver parenchyma that microscopically proved to be dense collections of neutrophils, karyorrhectic nuclear debris, and masses of faintly basophilic, fine filamentous bacteria. Samples of the abscesses were submitted to the microbiology laboratory but the best that they could do was anaerobic gram negative rods. Candidates included bacteroides, prevotella, and fusobacterium. Actinomyces would have been a consideration but they are gram positive. Pyogenic liver abscesses, while relatively rare, are known complications of chronic alcoholism and pancreatitis.

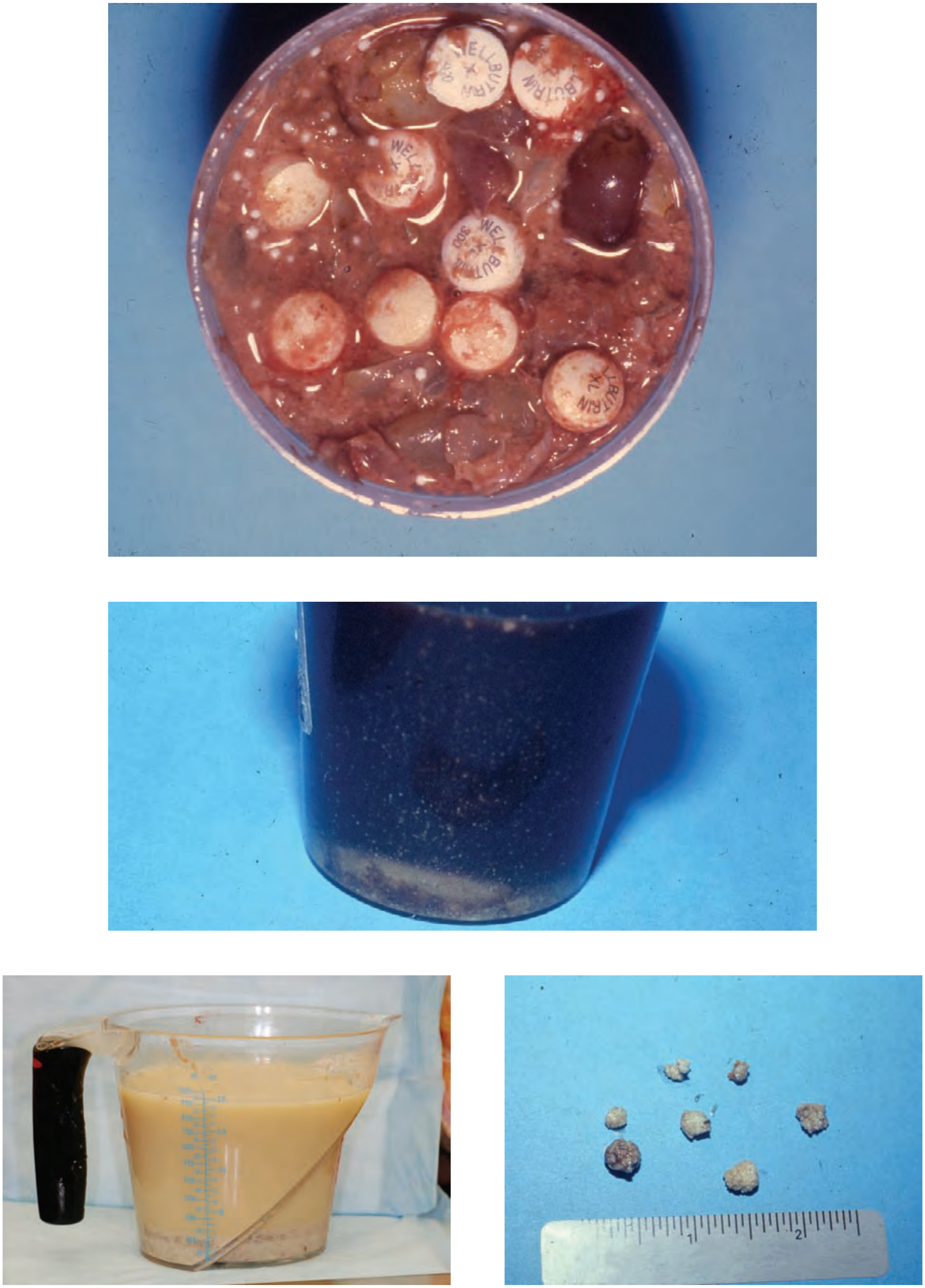

Suicidal overdoses demonstrating intact and fragmented granular pieces of partially digested pills. Note the containers with the granular white flecks at the bottom that indicate pill fragments.

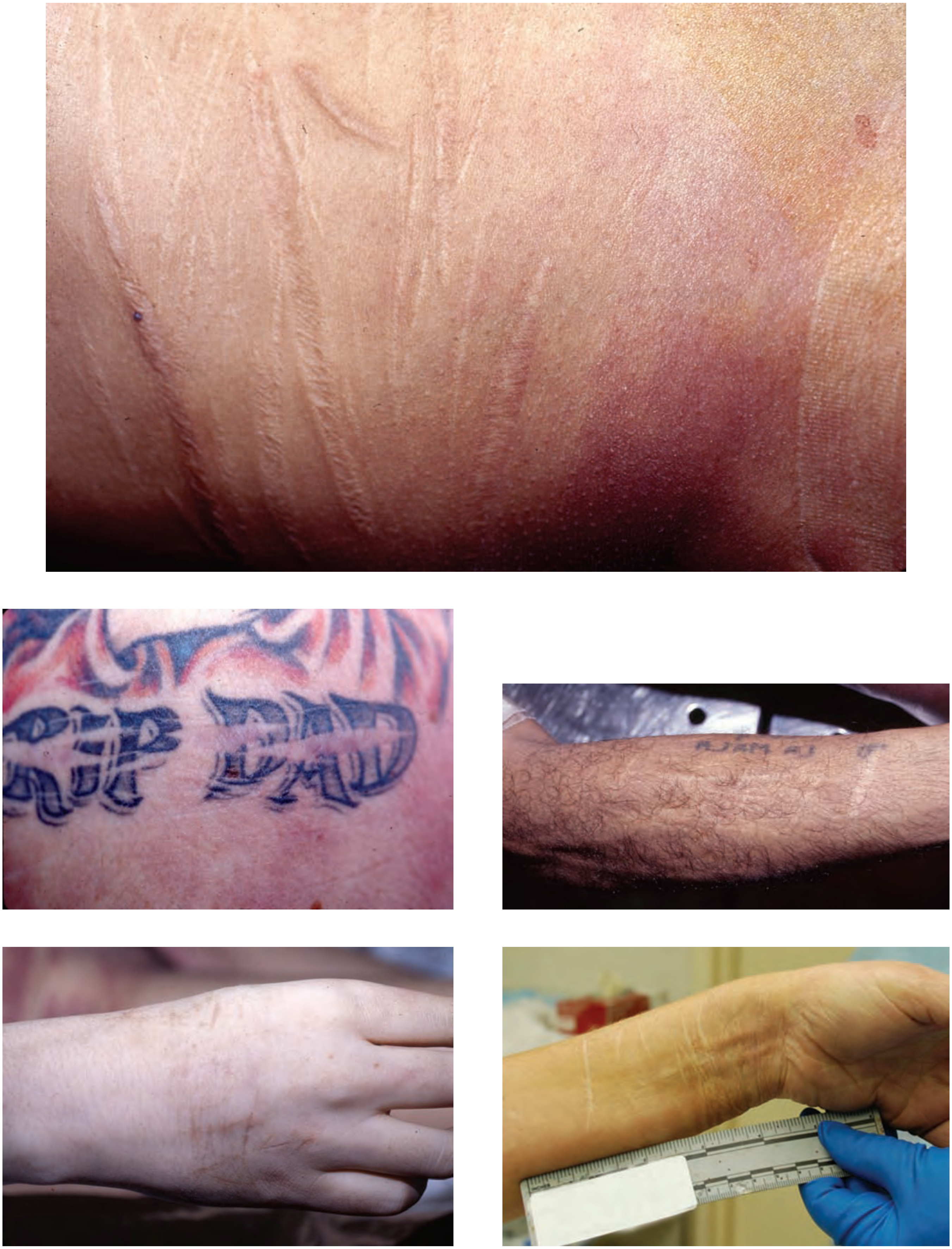

Note the multiple scars demonstrating destructive behavior associated with drug abuse, mental illness, and suicide.

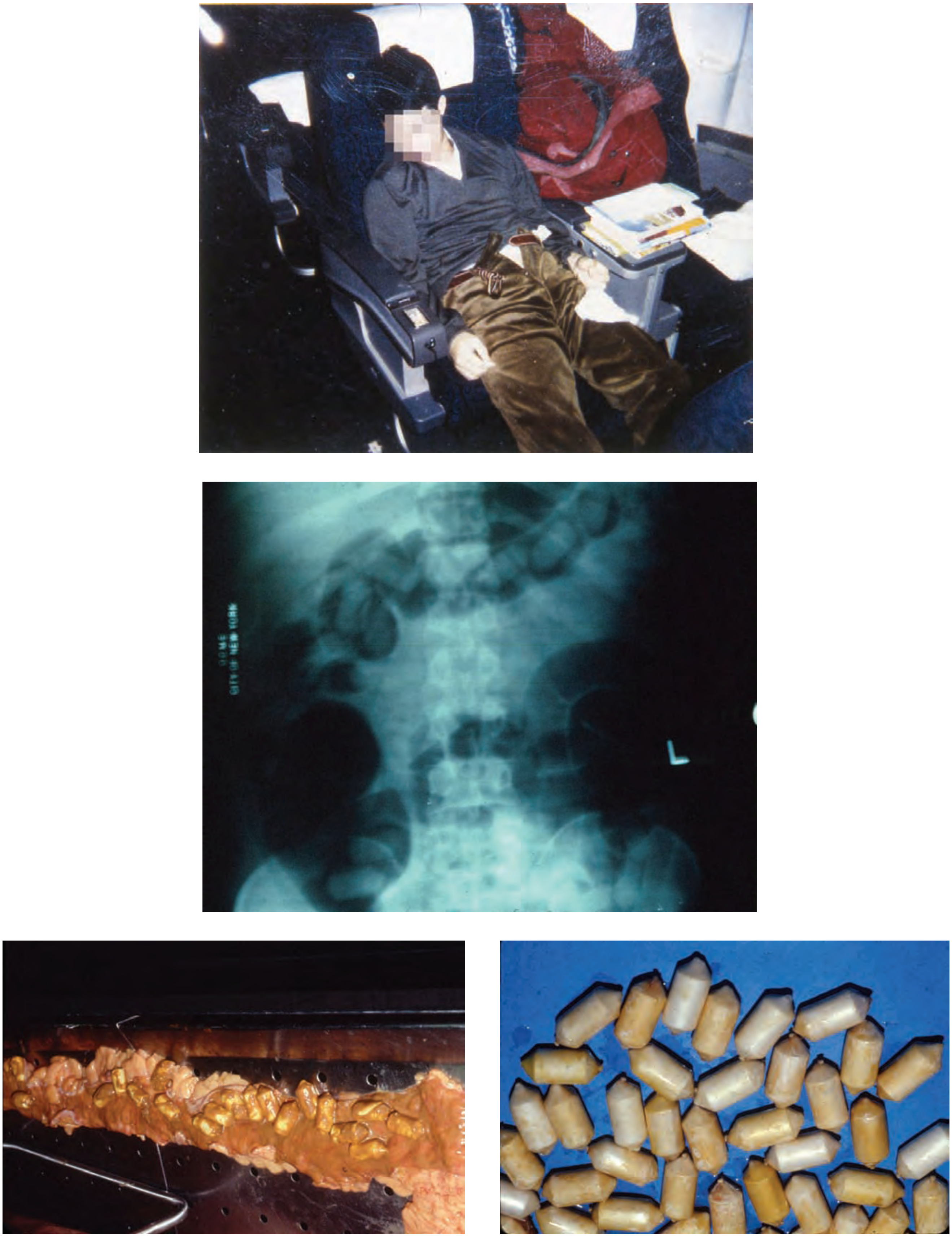

Acting as a “mule,” this person ingested multiple packets of drugs to smuggle them into the country. One of the packets ruptured and the individual died of a drug overdose before the plane landed. Note the typical x-ray findings demonstrating these packets within the gastrointestinal tract.

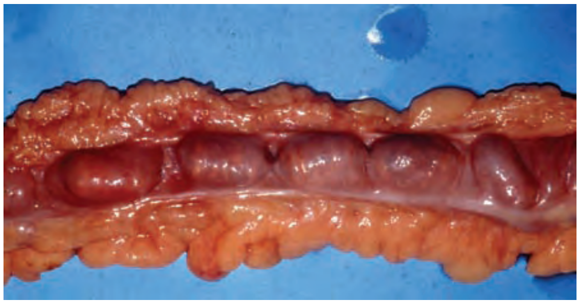

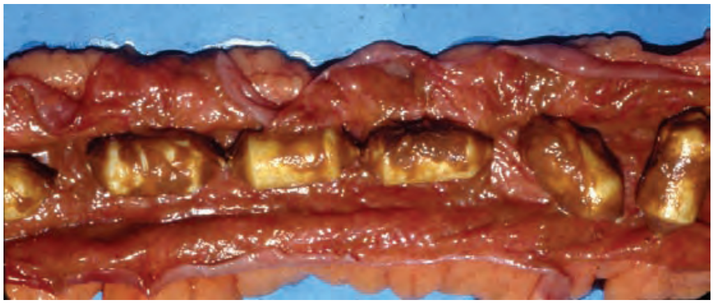

Packets of drugs discovered at autopsy in individuals who died in hotels next to an airport. Note the broken packets demonstrated at the bottom of the last picture.

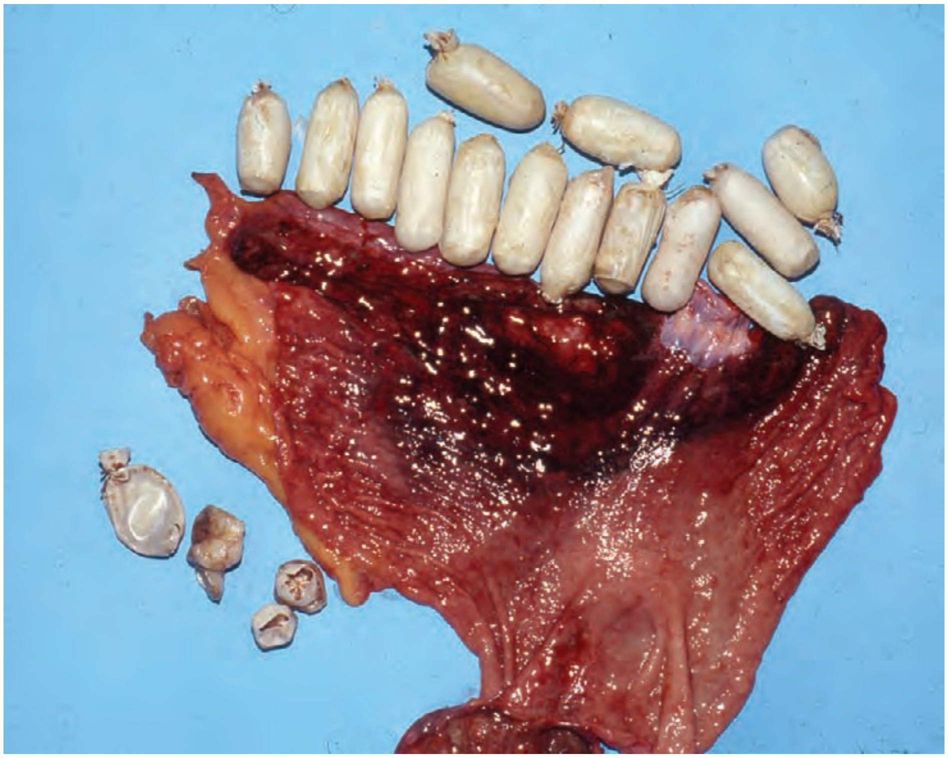

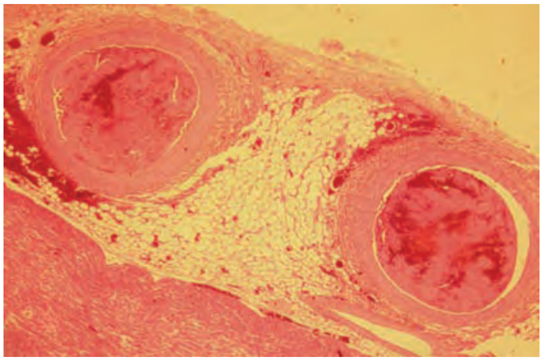

Epicardial vessels with slight atherosclerosis and occluding thrombosis associated with acute cocaine intoxication.

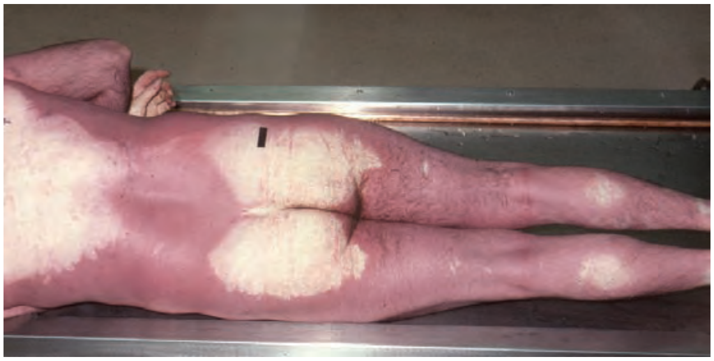

Carbon monoxide poisoning with cherry red lividity.

Postmortem Change and Time of Death

4

CHARLES A. CATANESE, BRUCE LEVY,
AND GERARD CATANESE

Introduction

In most jurisdictions the time of death is legally defined as the time that the person is declared or recognized to be dead. Thus, a decomposed body or skeletonized remains, clearly deceased for significant periods of time, may have an official time of death that is days, weeks, months, or even years after the actual death of the individual. However, being able to estimate the actual time of death (or the postmortem interval) can be critically important in criminal investigations, civil litigation, or settlement of the deceased's estate.

The concept of a "window of death" was developed to help answer these questions. This window can be initially bracketed by the time that the person was last known to be alive and the time he or she was declared dead. Forensic science can then apply a variety of observations and tests in an attempt to narrow that window as much as is scientifically and medically possible. Accurate interpretation of postmortem change is crucial in helping to establish the actual time of death.

Postmortem changes can be subdivided into several categories. These categories include early postmortem changes, decomposition, and skeltonization. Each has characteristics that can overlap with advancing time and are variable depending on environmental conditions and the physiologic state of the body at the time of death. The activity of insects and animals will also create artifacts that can lead to misinterpretation of the postmortem interval. Interpreting postmortem changes is not an exact science, and without a witness, one can only estimate a time period in which the death was most likely to have occurred.

Early Postmortem Changes

Generally considered to occur or evolve within the initial 24 hours after death, these consist of algor mortis, livor mortis, and rigor mortis.

Algor Mortis

The cooling of the body after death. After death the body will gradually cool until it equilibrates with the ambient temperature. Under standard climate controlled conditions, with average humidity and a room temperature of about 72° F, the body generally loses about 2° F per hour for the first 12 hours and than approximately 1° F per hour for the next 12 to 18 hours. If the ambient temperature is greater than the body temperature, algor mortis occurs in reverse and the body temperature rises until it reaches the ambient temperature.

The above calculations apply only at or near room temperature and will become variable as the ambient temperature changes. As a general rule, the rate of temperature change in a body is proportional to the difference between the temperature of the body and its environment. In other words, if the ambient temperature is below 72° F, the rate of heat loss will be increased from the numbers above. The reverse would be true for temperatures greater than 72° F. Moving air will cool faster than still air, and wet bodies will cool faster than dry bodies. Bodies found in stagnant water, flowing water, or those found buried in the ground will cool at different rates from those in the air and from each other. When the body and environment are at the same temperature, they have reached steady state and estimates of time of death based on temperature are limited to a minimum period of time.

Many other variables can influence algor mortis. In general, an elevated body temperature at the time of death will give the appearance of a shorter postmortem interval based on a measurement of the body temperature. A lower body temperature will have the reverse effect. Things that effect body temperature include infection, strokes, seizures, thyroid disease, and many other natural diseases. Whether or not the person was involved in strenuous physical activity prior to death (such as a violent struggle) will increase body temperature. Many other factors, including the age of the person, his or her overall health, and many drugs or medications can also have an effect on the calculation of postmortem interval based on this and other observations.

It is important to make certain there is consistency in how the body temperature is obtained. First, these calculations are based on a core body temperature, meaning either taking a rectal temperature or, more ideally, introducing a thermometer into the abdominal cavity beneath the liver to measure core temperature. Temperature readings on the surface of the body, including axillary, oral, or ear temperature are unacceptable for determining a postmortem interval. The thermometer must be kept in place long enough for the body temperature to equilibrate.

Livor Mortis

The settling of blood under the effects of gravity after death. It can first be appreciated as early as 20 minutes after death in very light-skinned individuals. The color of the livor will deepen to a purple color over the first 8 to 12 hours after death. During this period the livor is typically blanching, meaning that pressure on an area of livor will cause the color to briefly leave the area, resulting in a pale mark. After about 12 hours the livor will become fixed and no longer blanch under pressure. As with algor mortis, temperature will influence the rate at which livor becomes fixed, with increased temperatures shortening the time for livor to become fixed and decreased temperatures lengthening the fixation time. In some cases, livor mortis advances to cause small visible hemorrhages in the skin known as Tardieu spots, which should be distinguished from petechiae.

The medical condition of the person and the cause of the death can also create changes in livor mortis. People who are anemic (low blood count) or sustain an injury resulting in significant loss of blood might have very faint or even absent livor mortis. It can also be difficult to appreciate livor in dark-skinned individuals.

When evaluating livor mortis, it is important to note its location, intensity, whether it is blanching or fixed, and if it is appropriate for the position of the body. If livor is inappropriate for position, you can conclude that the person was moved at some point after the onset of livor. You can also evaluate livor for its color. For example, people who die as a result of carbon monoxide or cyanide poisoning will have a bright “cherry red” color to their livor mortis. People who die in a cold environment might also have a similar red color to their livor.

Rigor Mortis

The stiffening of the body after death. It is the result of a physiochemical process within the muscles of the body that does not cause actual contraction of the muscles. Under typical conditions, rigor will first be appreciated within 30 minutes to 2 hours after death. It will progress to a maximum intensity over the first 12 hours and remain at a maximum until about 24 hours after death. It will then “pass” from the body between 24 and 36 hours after death, after which the body will remain flaccid.

As with algor mortis and livor mortis, many variables will affect the rate of development and passing of rigor mortis. In general, increased environmental or body temperatures will speed the rate of development and passing of rigor mortis. Lower environmental or body temperatures will have the reverse effect. People at the extremes of age (children and the elderly) will have different rates of appreciation of rigor mortis due to the decreased muscle mass in these age groups.

When evaluating rigor mortis, it is important to note its location and intensity and whether it is appropriate for the position of the body. If rigor is inappropriate for position you can conclude that the person was moved at some point after the onset of rigor.

Decomposition

As time advances, decomposition gradually increases. Decomposition can be subdivided into two basic categories, putrefaction and mummification. While they can be seen in isolation, careful observation will typically reveal features of both processes simultaneously.

Putrefaction

Decomposition that occurs due to the actions of bacteria. The bacteria typically break down the body from the inside out, causing many of the changes we associate with a person who has been deceased for a longer period of time. However, when there is penetrating trauma that breaks the surface of the body, bacteria from the environment can gain access and hasten putrefaction. These changes include darkening and slipping of the skin, the production of a foul-smelling gas, with bloating of the body and marbling of blood vessels. A dark colored bloody-appearing purge fluid will come from the nose and mouth and should not be confused with blood related to trauma.

This process is extremely environmentally dependent. As with the early postmortem changes, warmer temperatures accelerate this process and cooler temperatures slow it down, with many of the same factors playing a role. A body placed in a dark dumpster in the hot summer sun for half a day can reach a state of putrefaction equivalent to a body left at room temperature for several days or one in the winter cold for greater than a week. In the case of fire fatalities, charring will preserve the body and decelerate this process. A frozen body that has thawed will putrefy at a markedly accelerated rate due to spaces left by frozen ice crystals.

Mummification

Mummification occurs in a dry environment, typically such dry outside environments as a desert, or in heated indoor environments during the winter months when the relative humidity is low. Initially, there is a darkening and hardening of the skin. This will progress to produce flaking of the surface of the skin that can give extremities the appearance of a log, thus the term "tree barking." As with putrefaction, the skin will begin to split and eventually the skin and soft tissues under the skin are also lost.

Skeletonization

Both putrefaction and mummification eventually lead to skeletonization of the body. The loss of soft tissue progresses at a variable rate, depending on the environmental conditions and access of the body to a variety of animals and insects. The face and ends of the extremities are the first areas where bone is exposed, and the pelvic soft tissues are the last to be lost. While this process is usually measured in months, work performed at the Anthropologic Research Facility at the University of Tennessee in Knoxville has demonstrated that complete skeletonization of a person can occur within days under appropriate conditions.

Autolysis

Autolysis associated with decomposition occurs after cell death and is due to the actions of digestive and catalytic enzymes released from cells in the body. This term, "self-destruction," is advanced in certain organs, especially the intestinal tract and pancreas, due to their rich enzyme content. This is important to note because autolysis can mimic certain disease processes at the time of gross examination. Maceration of stillborns who were dead inside a uterus for days is a type of autolysis associated with moist sterile environments, appears as a red discoloration with skin slippage, and is not to be confused with trauma or disease.

Adipocere

A process that is rarely seen outside of exhumation cases. It is a chemical change that occurs in the fatty tissues of the body exposed to wet conditions. Adipocere takes many months to develop, and once it forms is extremely persistent, stopping the typical loss of soft tissue and skeletonization of the body.

Other Factors in Evaluating Postmortem Interval

When a deceased body has been exposed to insect or animal activity, we can make observations that might assist in the determination of the post-mortem interval. Maggots (the larval stage of blowflies) can be collected from the body and examined by an entomologist. You need to collect two sets of specimens. One set should be kept alive while the other set is killed. The entomologist will determine the exact species of fly from the matured living specimens and the likely time of death from the killed specimens. In persons who have been deceased for a prolonged period of time (months to years), an evaluation of the type and age of plants growing up through the body might also provide useful information.

Examination of the stomach and intestinal contents can also provide clues to help us in the determination of time of death. Our bodies digest food at a fairly predictable rate, depending on the quantity and types of foods consumed. Higher caloric content foods, such as fats, are digested slower than less complex foods. Larger meals remain in the stomach for a longer period of time. The rate of stomach emptying can be accelerated or slowed by natural disease, drugs, alcohol, or emotional stress. The description of the quantity, type, and condition of food in the stomach is part of the routine autopsy. This can be combined with information about that person's recent consumption of food obtained during the investigation to estimate the time from the eating of that meal until the time of death. Even without information about the last meal consumed, the stomach contents might provide clues as to when death occurred based on the types of food in the stomach.

Finally, markers found at the scene of death might provide the best information regarding the date or time of death of the individual. Observations regarding uncollected mail or newspapers, information from the telephone company about the last time the phone was used, dated sales receipts, or a noted change in habitual behavior might assist with determining the date of death. The clothing the person is wearing or the status of lighting in the home might provide clues regarding the time of the day or night that death occurred.

In conclusion, determining the postmortem interval can be difficult. Start with a broad window of death (last-known-alive to found-dead). Then use as many of the above elements as are available to attempt to narrow that window as much as possible. It is important to remember that there is variability of these observations and calculations, so keep an open mind and be willing to reevaluate your opinion based on new information as it becomes available.

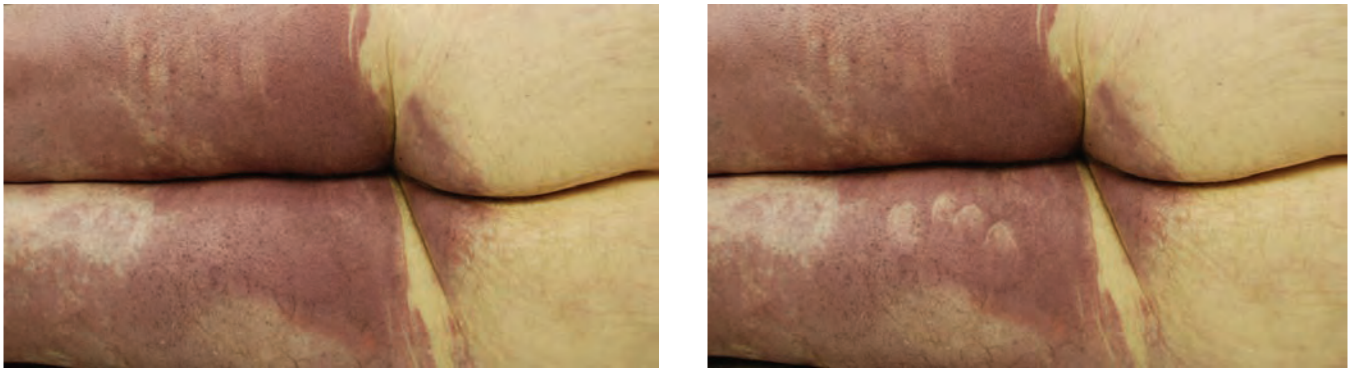

Blanching lividity. Note the finger marks caused by pressing the blood away from the skin surface at the posteromedial aspect of the right thigh.

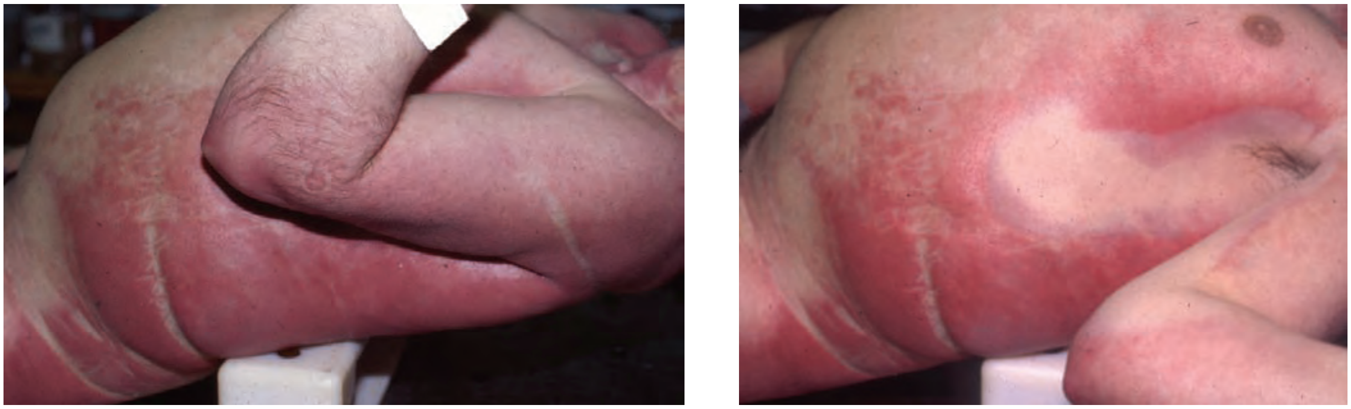

Fixed lividity at the lateral and posterior torso. Note the absence of lividity underlying the pressure region exerted by the arm against the weight of the chest. This child was discovered lying on his left side. This body was stored in a cold refrigerator for many hours prior to being autopsied.

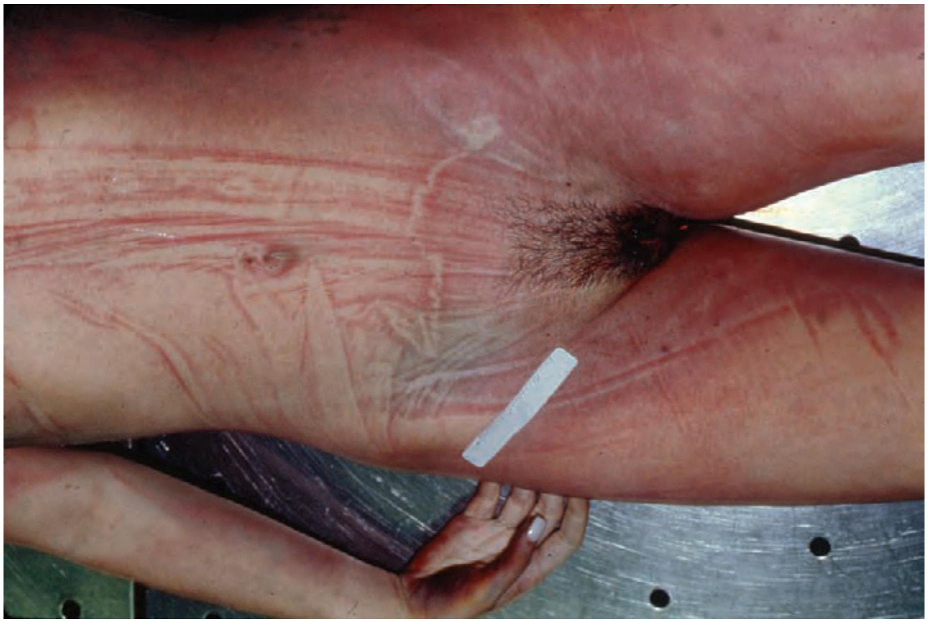

This individual was placed in the morgue on her back the day before this examination. Note the fixed anterior lividity with partial sparing over with pressure points caused by lying on a wrinkled bedspread in her home for many hours prior to being discovered. Also note the early putrefaction consisting of slight green discoloration of the lower right abdomen.

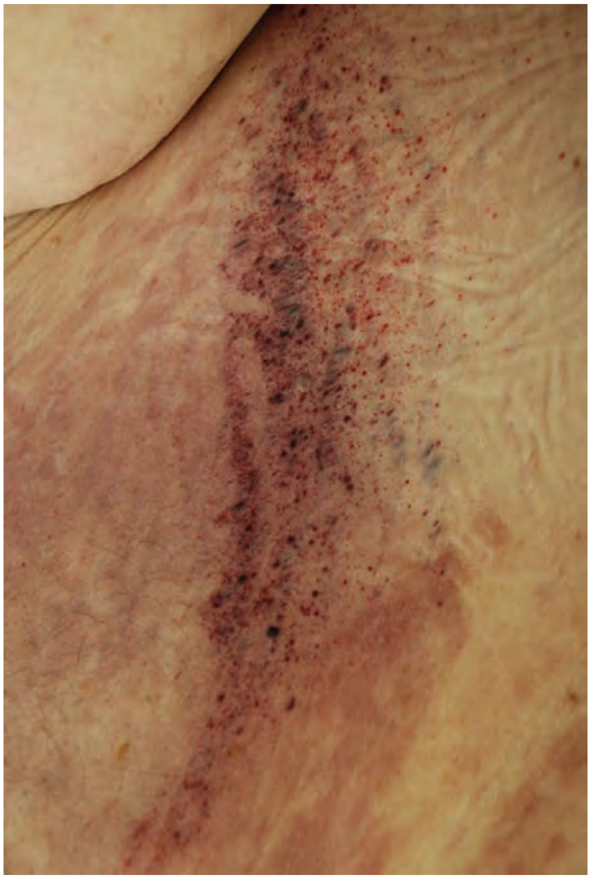

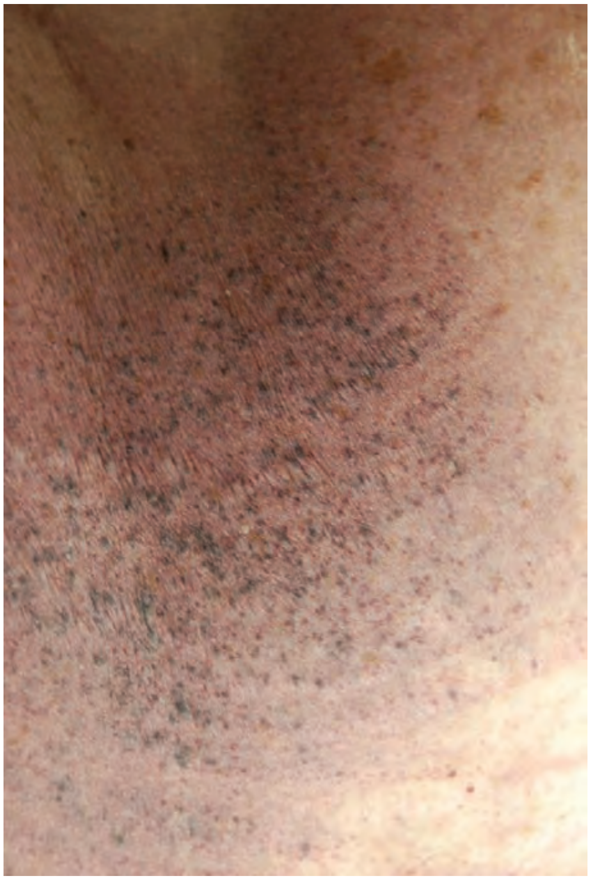

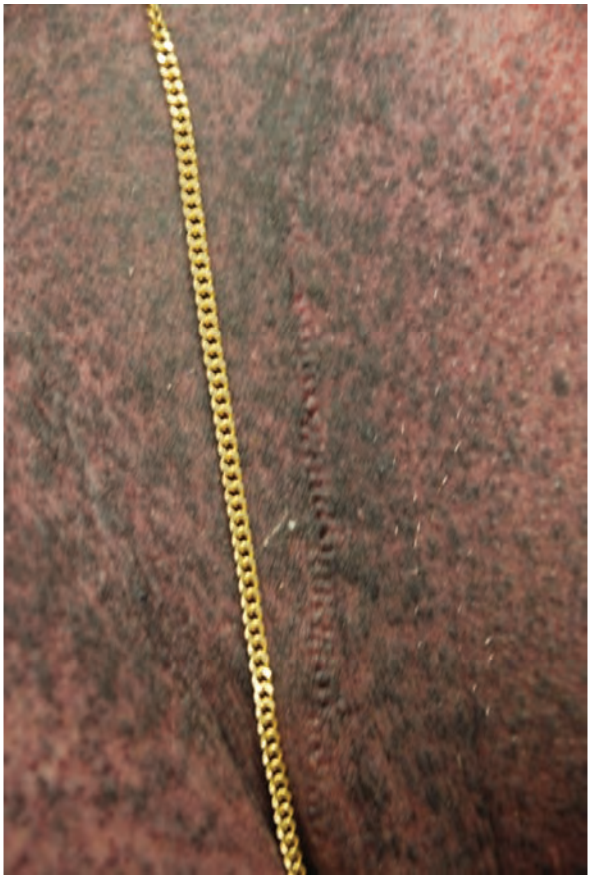

Marked fixed lividity with dark purple spots known as "tardieu spots." Also note the vague chain pattern. It is not unusual for objects such as this to leave postmortem imprints.

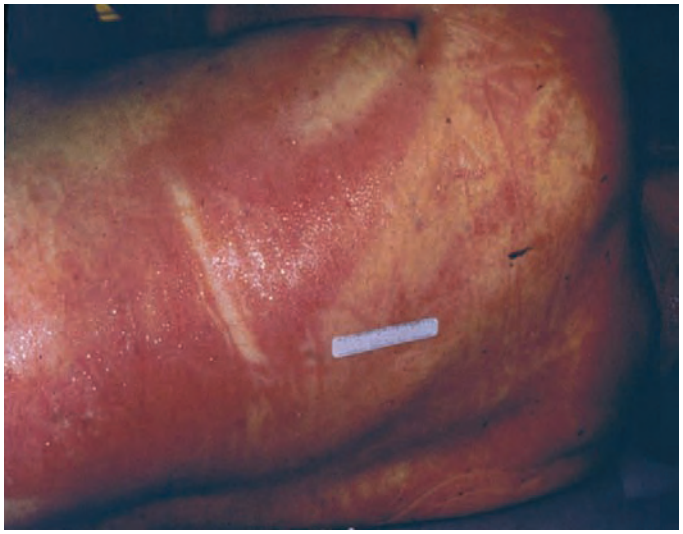

In carbon monoxide poisoning, the color of the livor mortis is more of a bright red instead of the typical red-purple color. This change can also be seen in cases of cyanide poisoning or in decedents who have died in cold environments. Note that the color differences are very subtle and it may not be possible to make a determination from visual inspection alone.

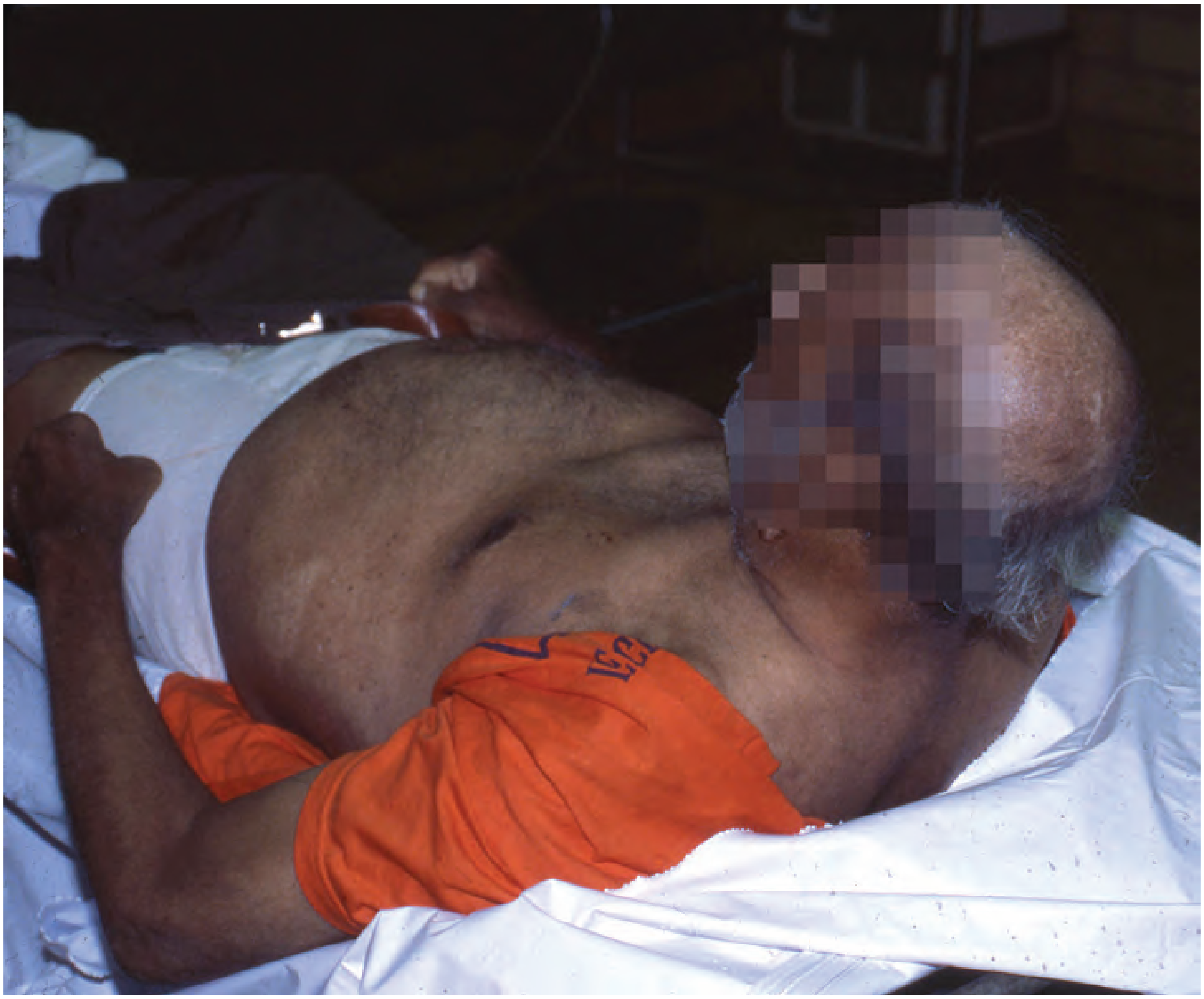

The decedent was found lying on his back with his head elevated on top of several pillows. He was in full rigor mortis at the scene and maintained this position through his arrival into the morgue.

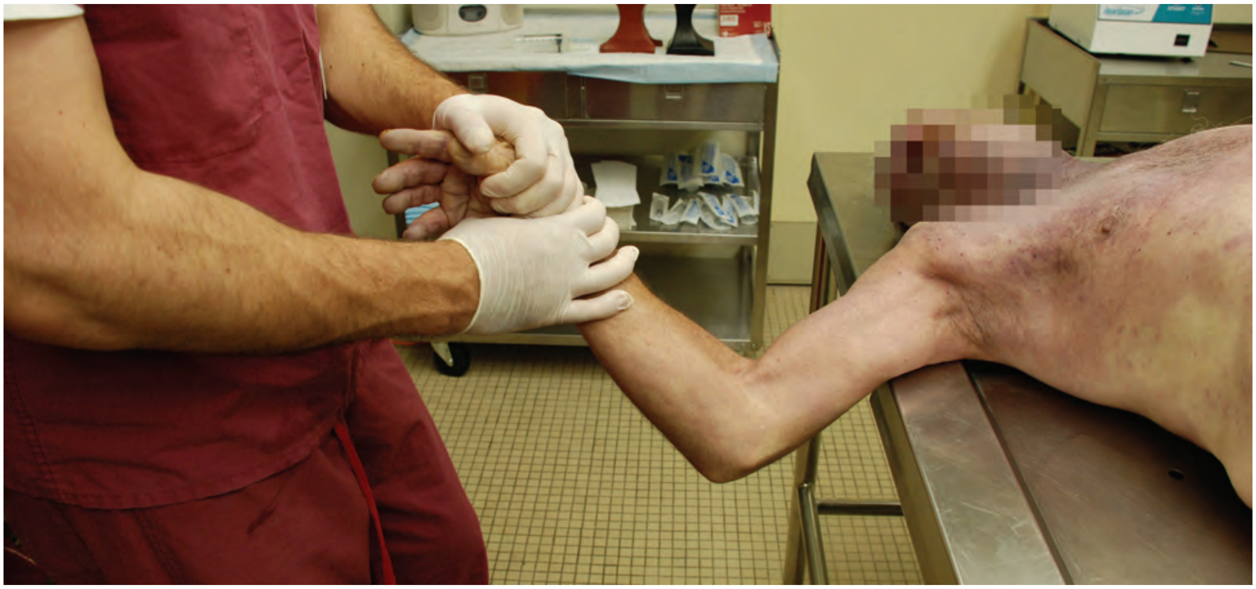

Full rigor mortis. Note the morgue technician pulling tightly on this decedent's arm, which is stuck in a bent position.

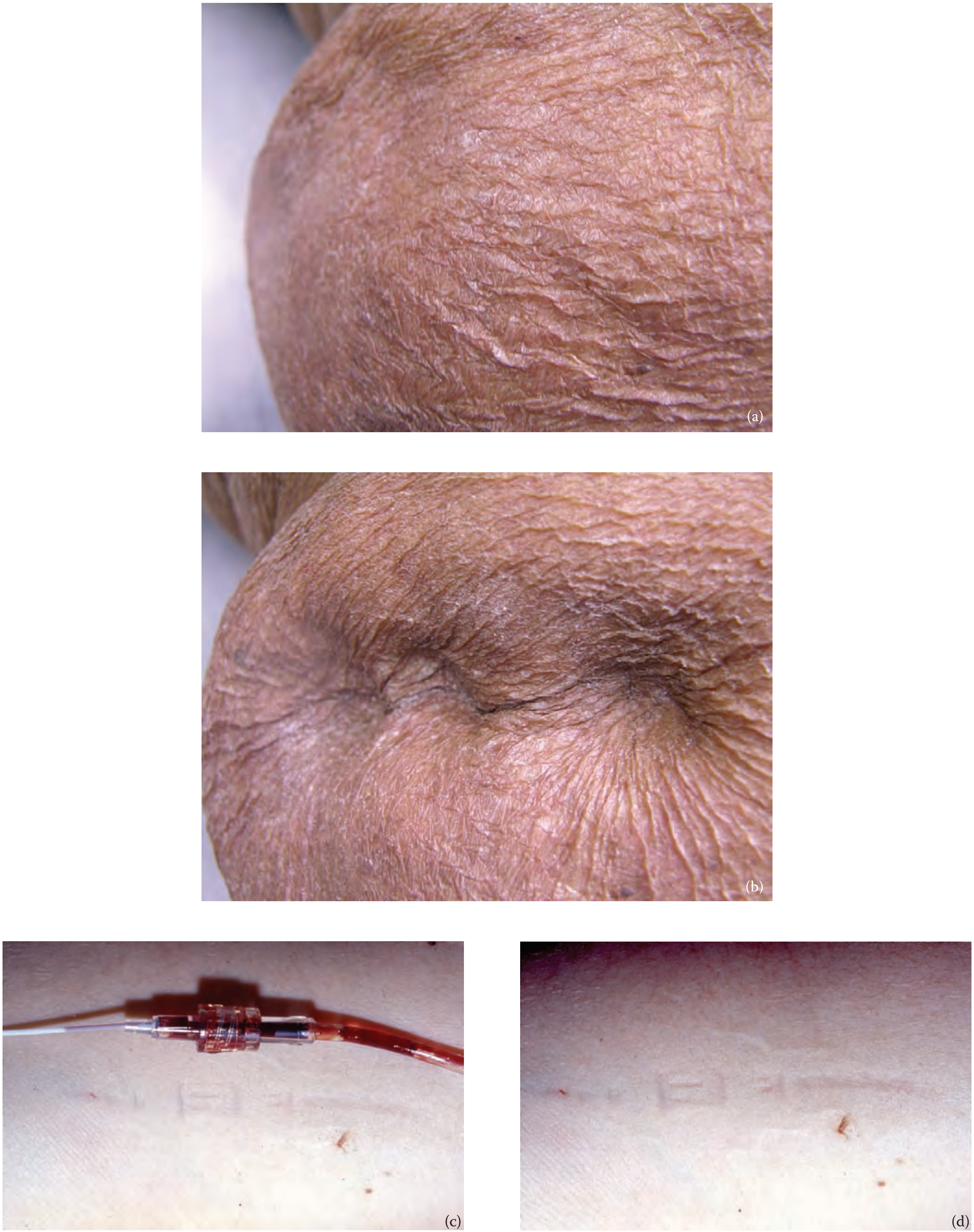

Congealing of fat. Adipose tissue at very cold temperatures will demonstrate increased viscosity. This might be mistaken for poor skin elasticity due to dehydration, which may or may not be present with congealing of fat. Note figure (b) shows indentations that have been retained following pressure exerted by fingers.

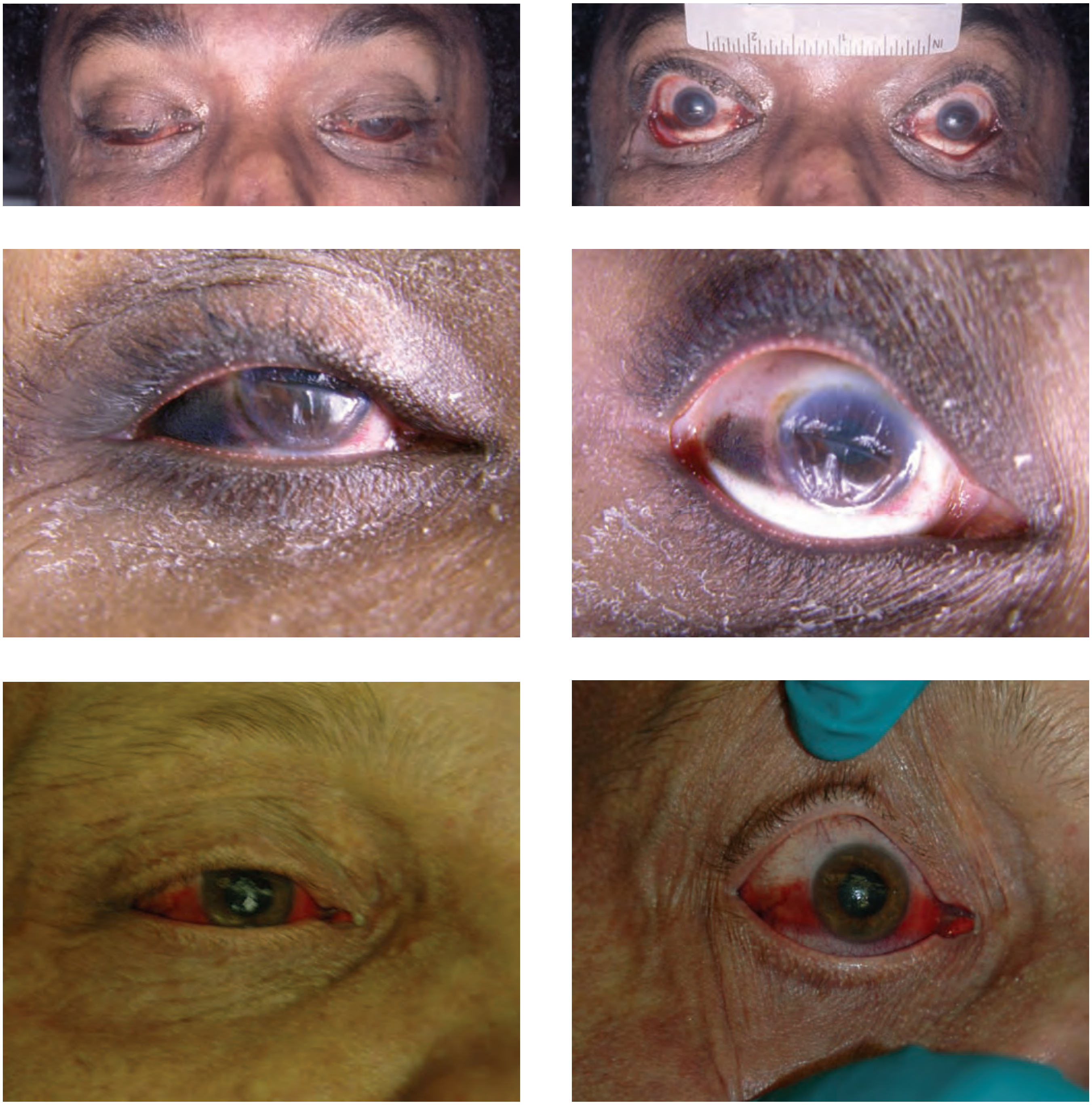

These are examples of "tache noire," which is a dark discoloration of a portion of sclera exposed to air. This is due to drying and is usually brown to red. This postmortem change may be misinterpreted as hemorrhage associated with strangulation.

This drying of the lips and tongue with dark discoloration is due to postmortem drying associated with mucosal exposure to air. This may be misinterpreted as an antemortem finding resulting from ingestion of caustic substances. The process of suicidal hanging caused the individual's tongue to stick out and become dry and dark.

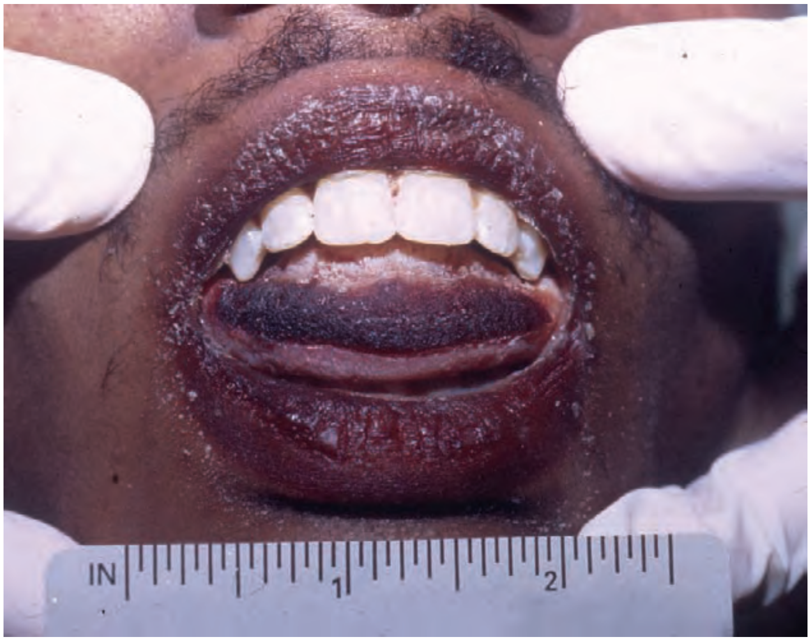

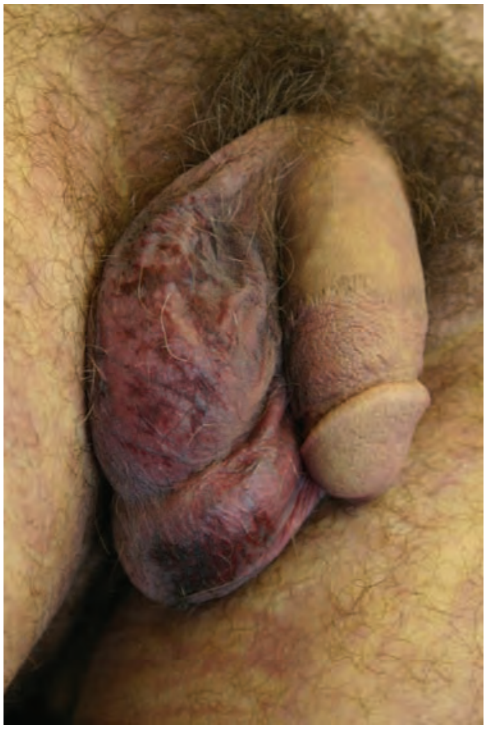

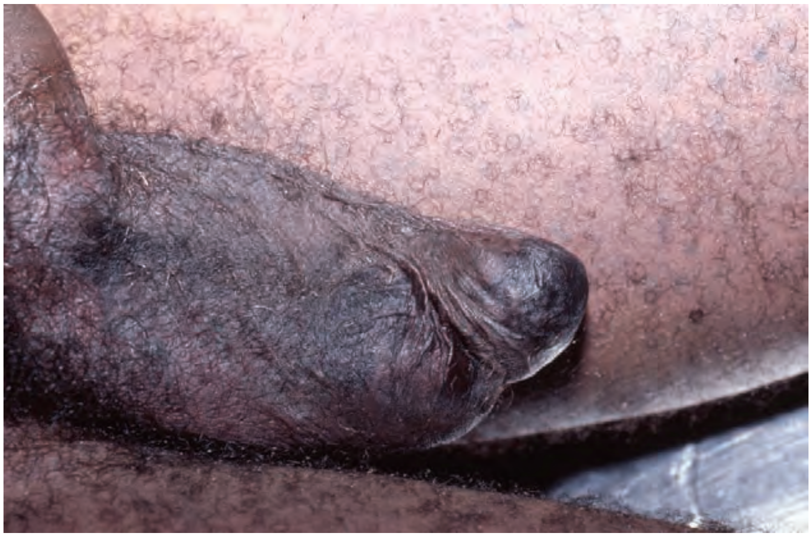

This is also postmortem drying with dark discoloration of the scrotum and shaft of the penis. This may also be mistaken for an antemortem injury such as an abrasion. If there is doubt, one may make an incision to document underlying hemorrhage.

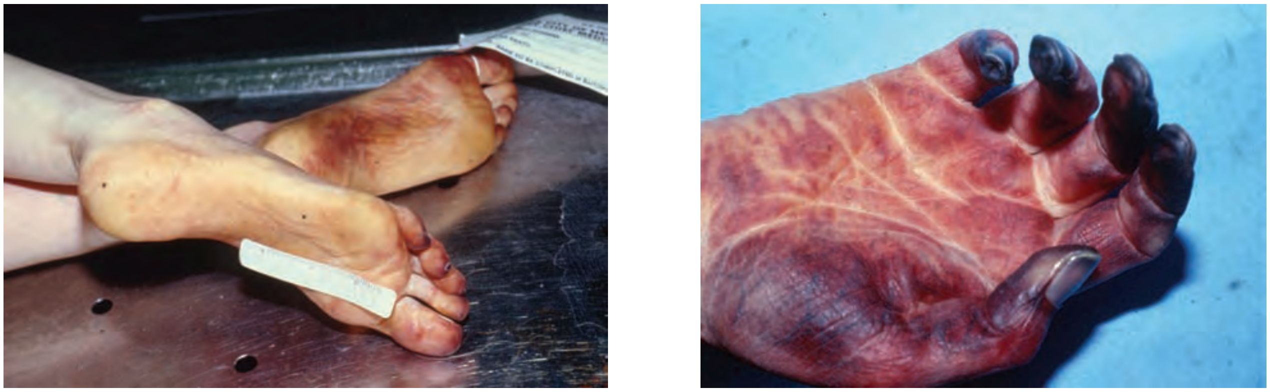

This is early mummification with drying of the hands and feet. Note the dark discoloration of the fingers with indentation from dehydration. Portions of the body with larger surface areas relative to underlying tissue mass will mummify more rapidly.

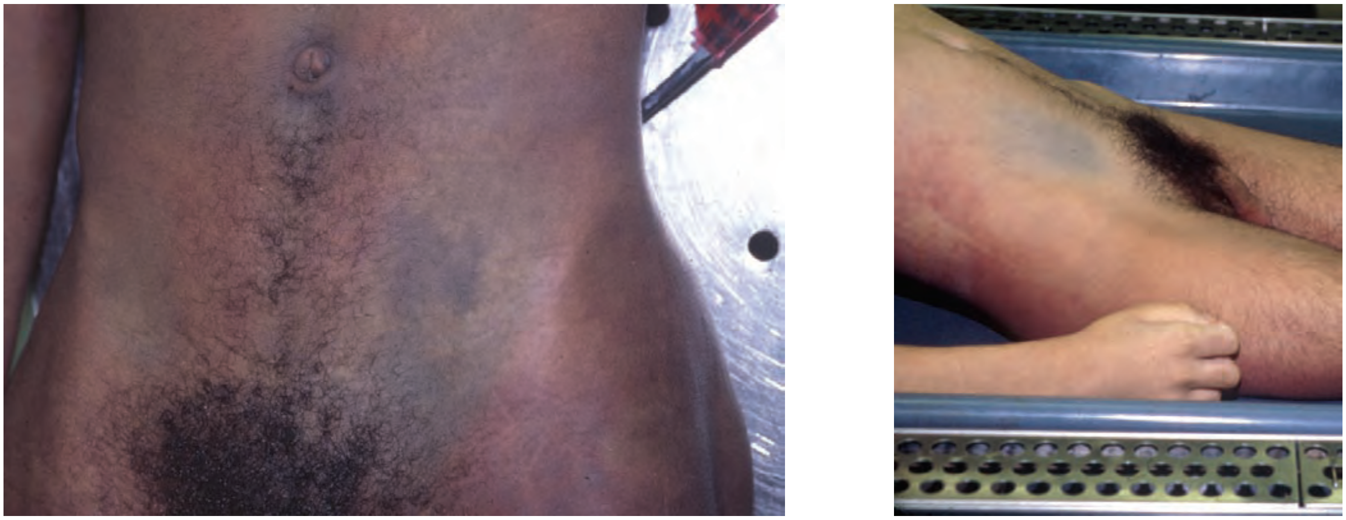

Early putrefactive change with green-brown discoloration of the lower abdomen. The first place for this to occur is typically the lower left abdomen above the cecal region.

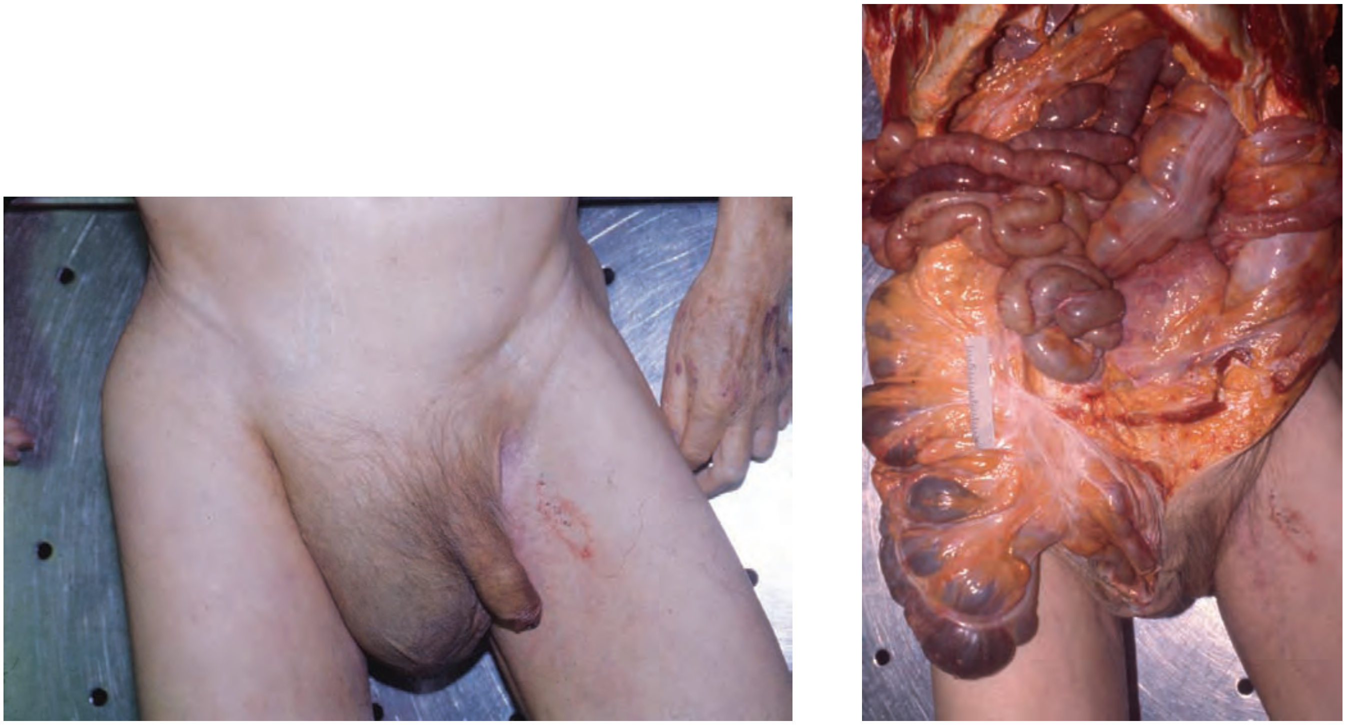

This large hernia demonstrates early decomposition with slight green discoloration of the scrotum due to a portion of the large intestine extending into the scrotal sac.

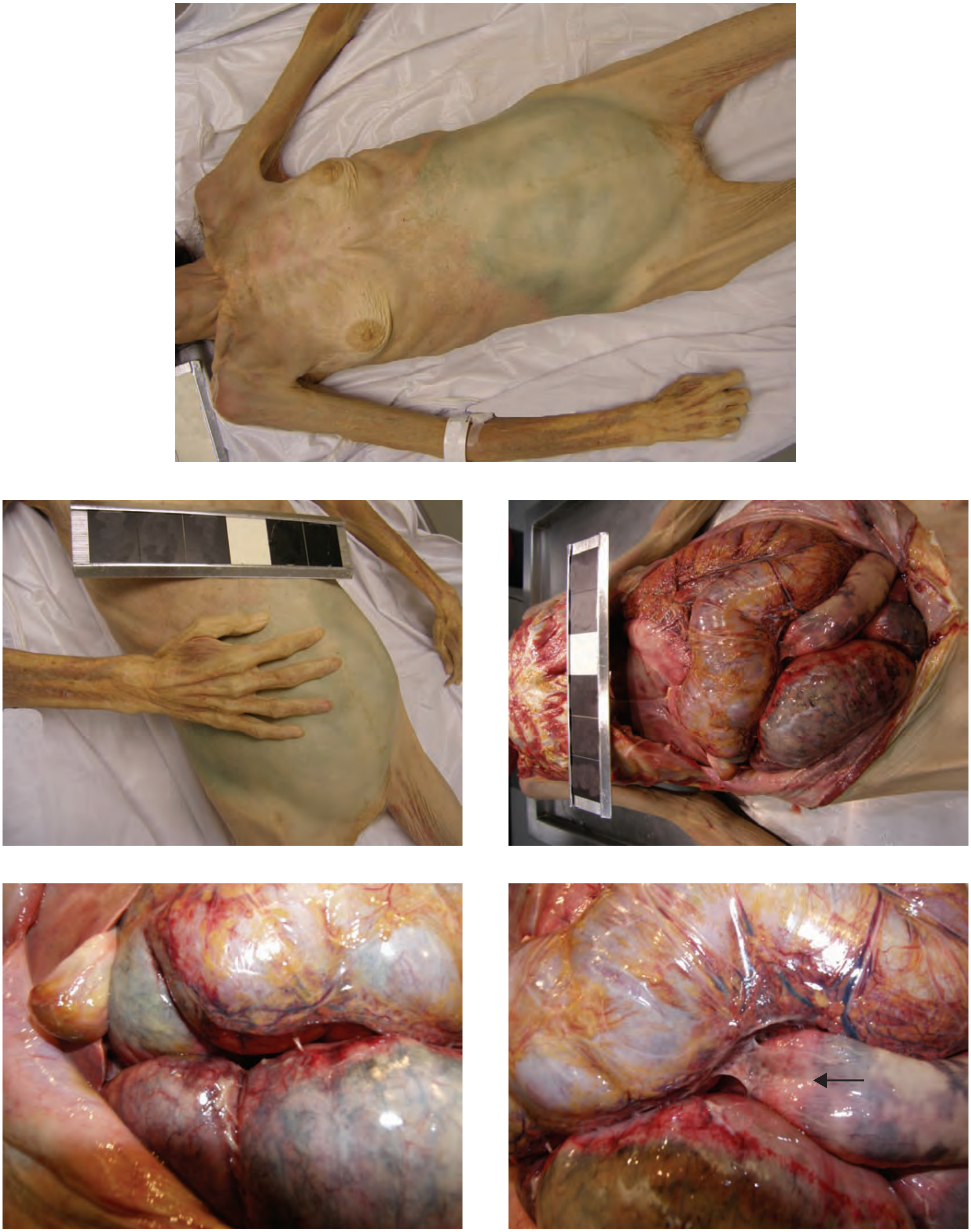

Disproportionate advancing putrefactive change of the abdomen due to acute peritonitis. Note the green discoloration of the abdomen and the absence of green discoloration elsewhere. The presence of infection with bacteria caused this regionally accelerated process. Note the parulent exudate at the surface of the intestines, indicated by the arrow.

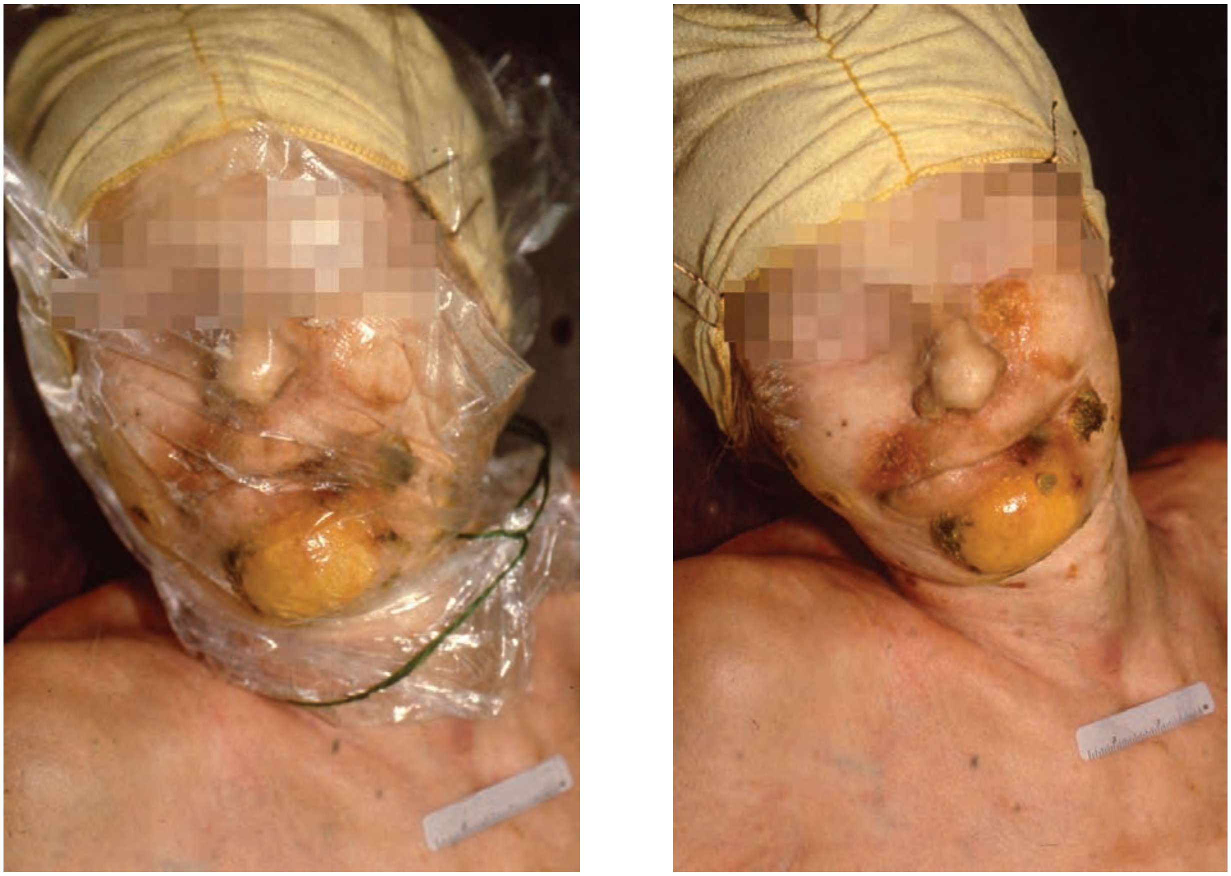

This individual committed suicide by ingesting excess prescription medication and placing a loosely fixed plastic bag over the head. The moisture collecting at the face covered by plastic during breathing helped to create an environment where mold and mildew could form at the chin and face.

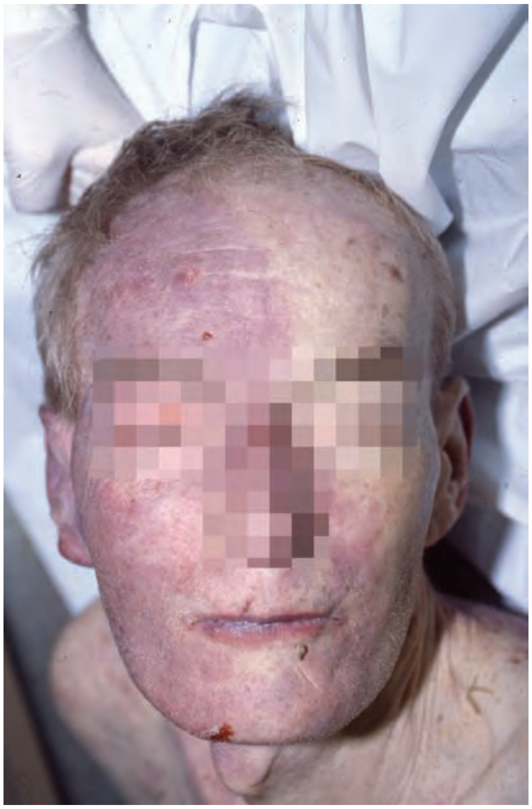

Postmortem hot water burns of the face in this individual, who was found partially submerged in the warm water of a bathtub. He suffered a cardiac event while taking a shower. It requires less heat to cause thermal damage to a dead body than to a living body. Note the sharply demarcated red region indicating his nose and mouth were below the hot water.

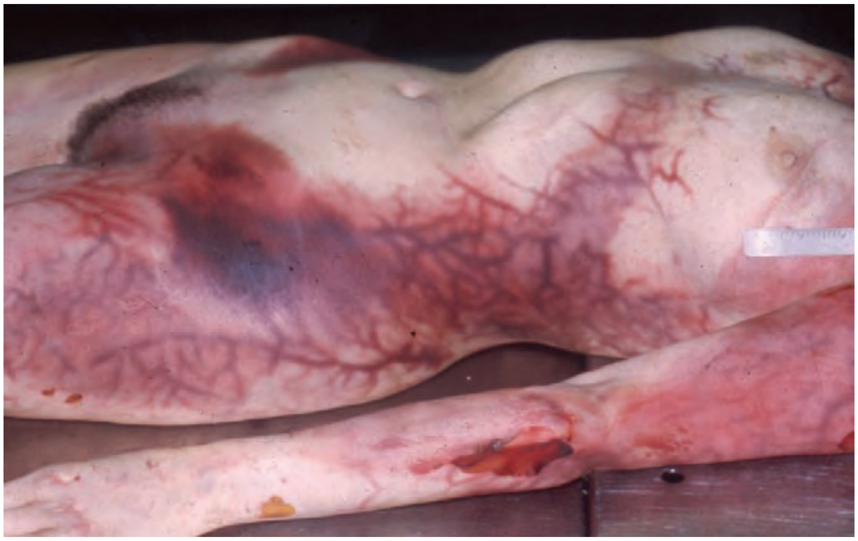

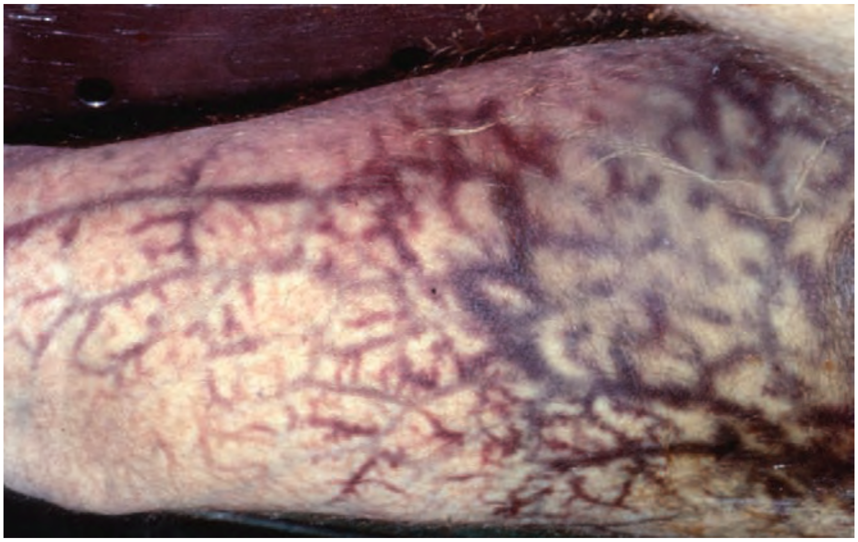

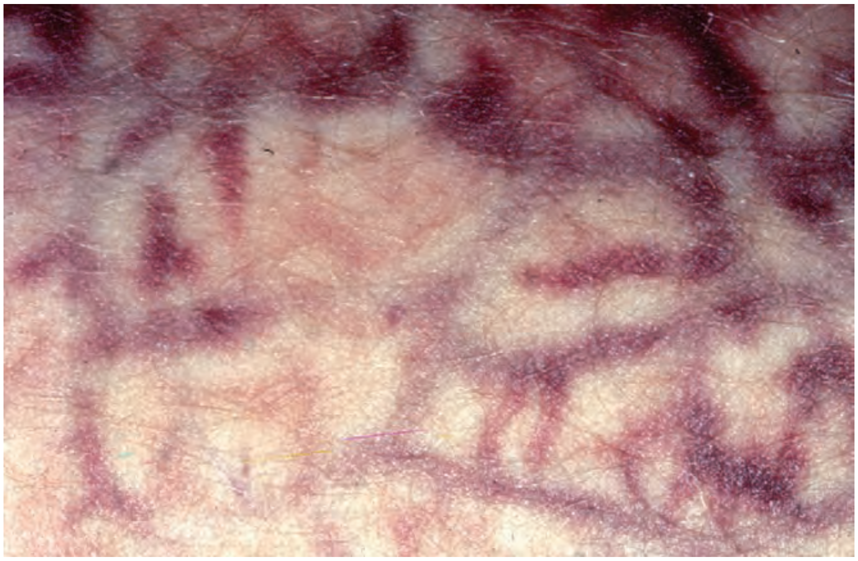

Marbling. This early presentation of putrefaction is predominantly caused by bacteria tracking through the superficial blood vessels causing pigment changes in blood and vessel walls as microorganisms digest the body.

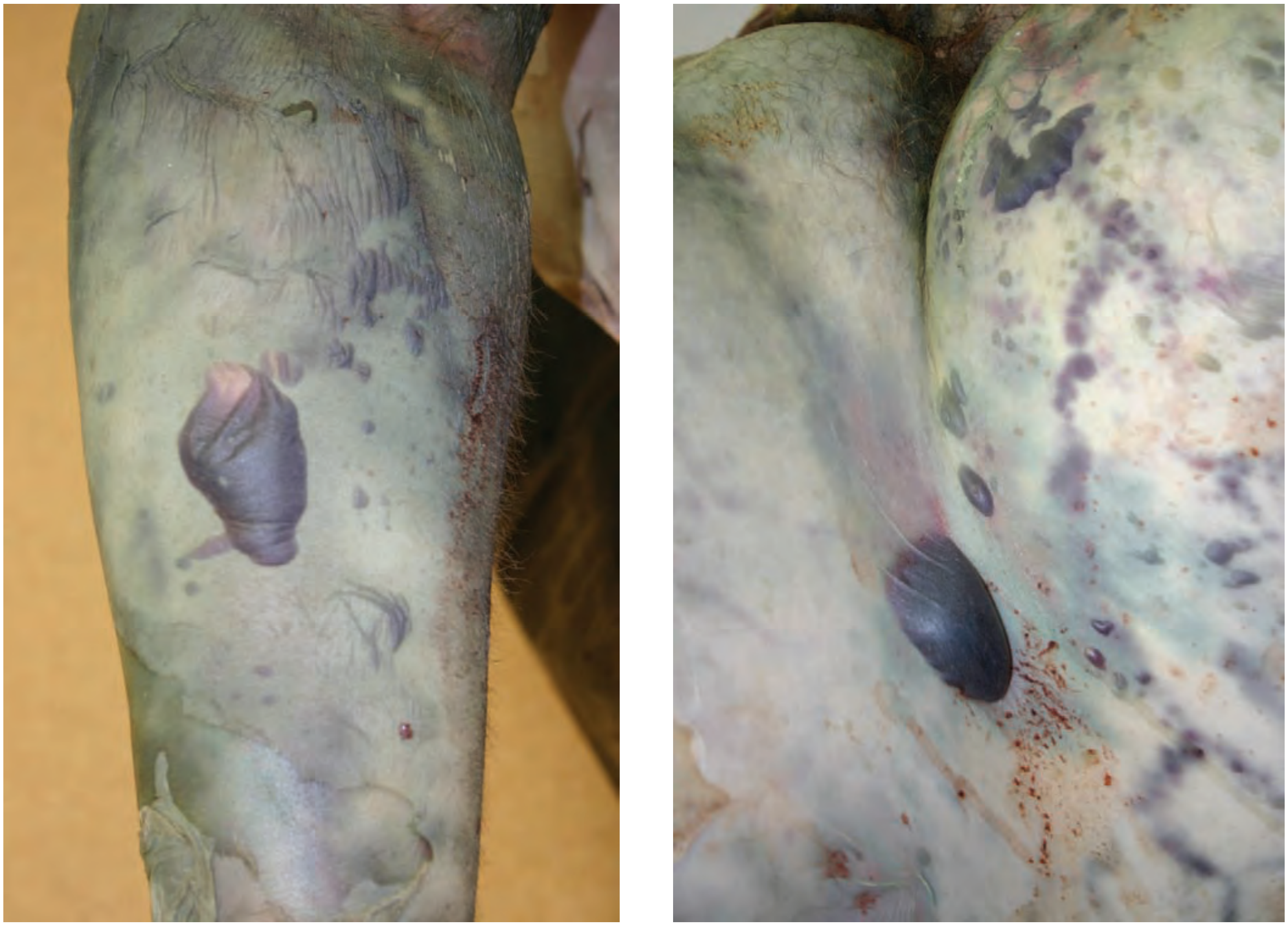

Early to moderate putrefactive change with green-brown fluid-filled blisters. Such blisters may be mistaken for second degree thermal burns or aggressive antemortem bacterial skin infections.

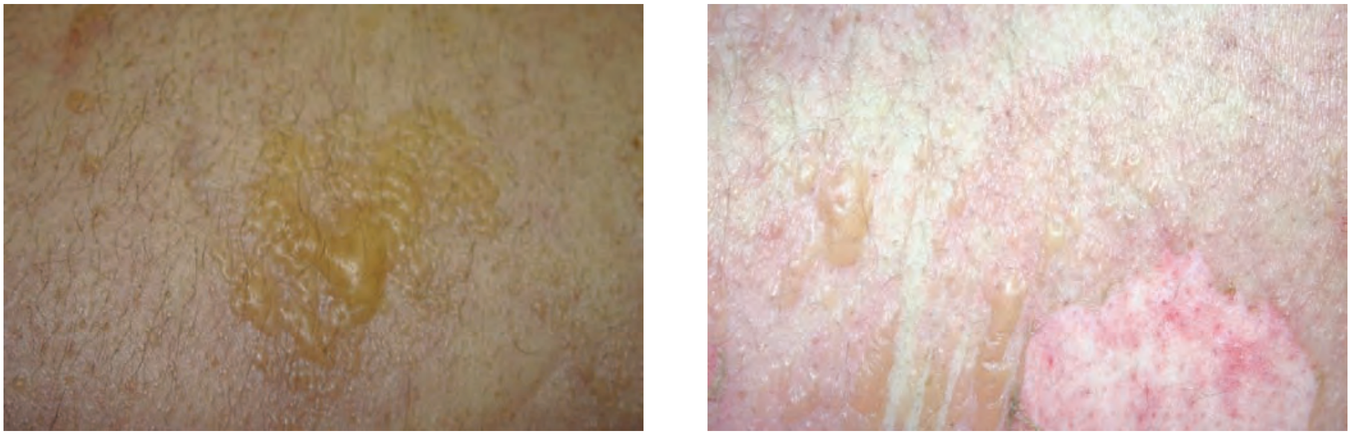

Early putrefactive change with clear yellow fluid-filled blister formation. The rupture of these blisters would appear as skin slippage.

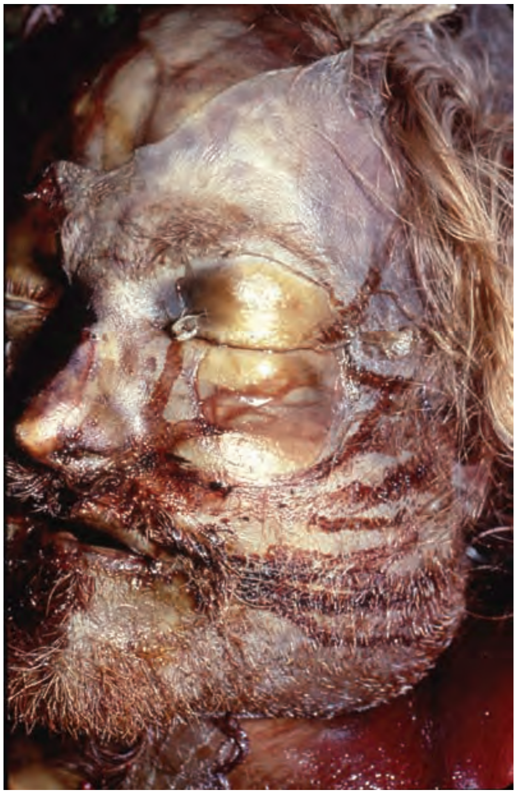

Putrefaction with "purging." Note the red to brown fluid gurgling from the mouth and nose. This can sometimes be mistaken for an upper gastrointestinal hemorrhage.

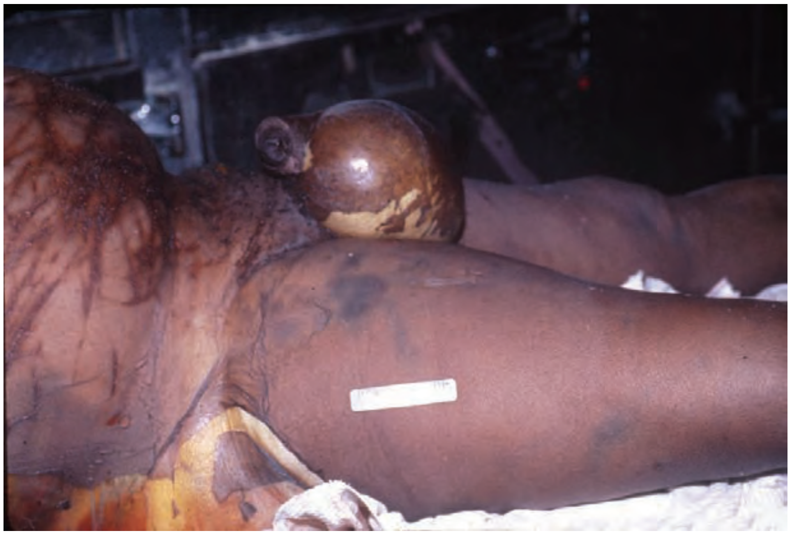

Putrefactive change with bloating and expansion of the scrotum due to gas accumulation.

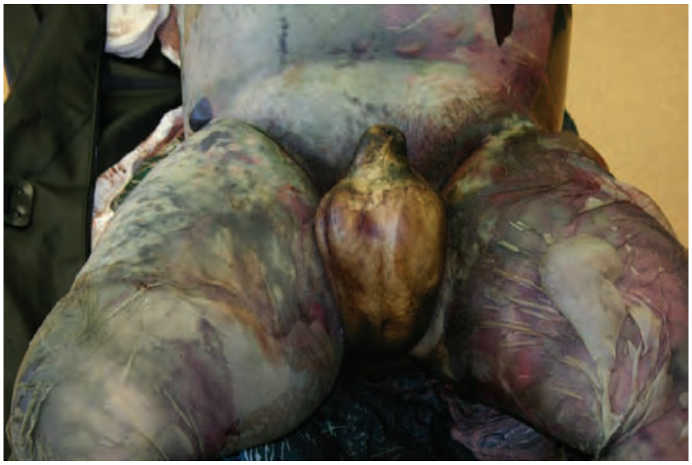

Green-brown-red discoloration with fluid-filled blister formation and skin slippage, scrotal enlargement from gas formation due to metabolism of proliferating microorganisms.

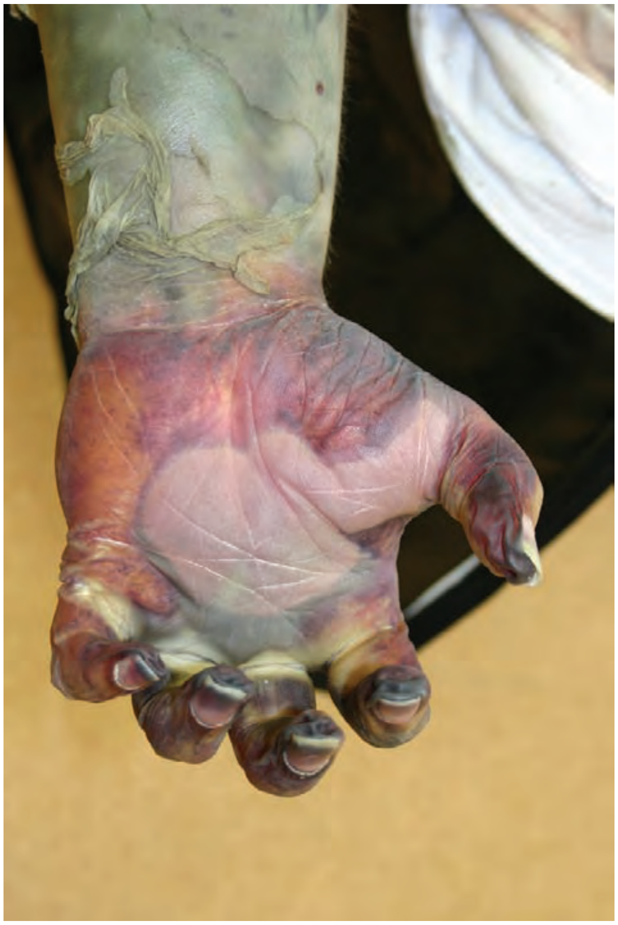

There is mummification of the fingertips with an expanded gas-filled blister and green to brown putrefied fluid in its inferior aspect.

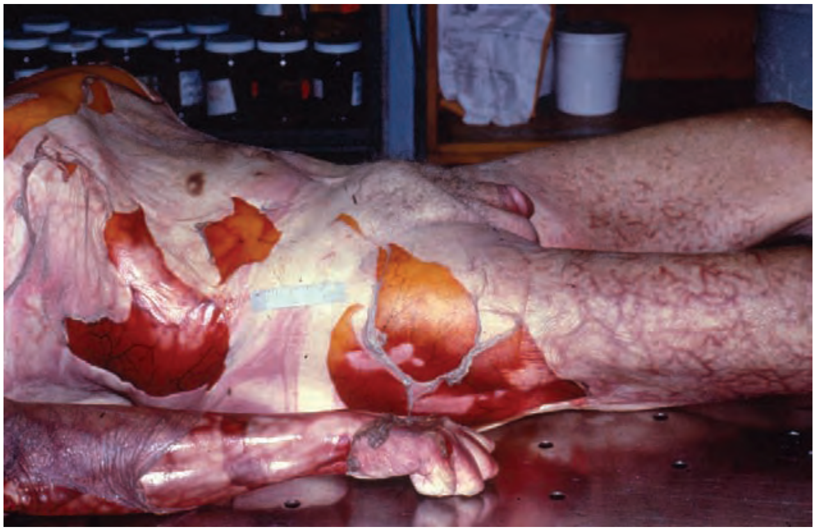

Marbling and skin slippage; early to moderate putrefaction.

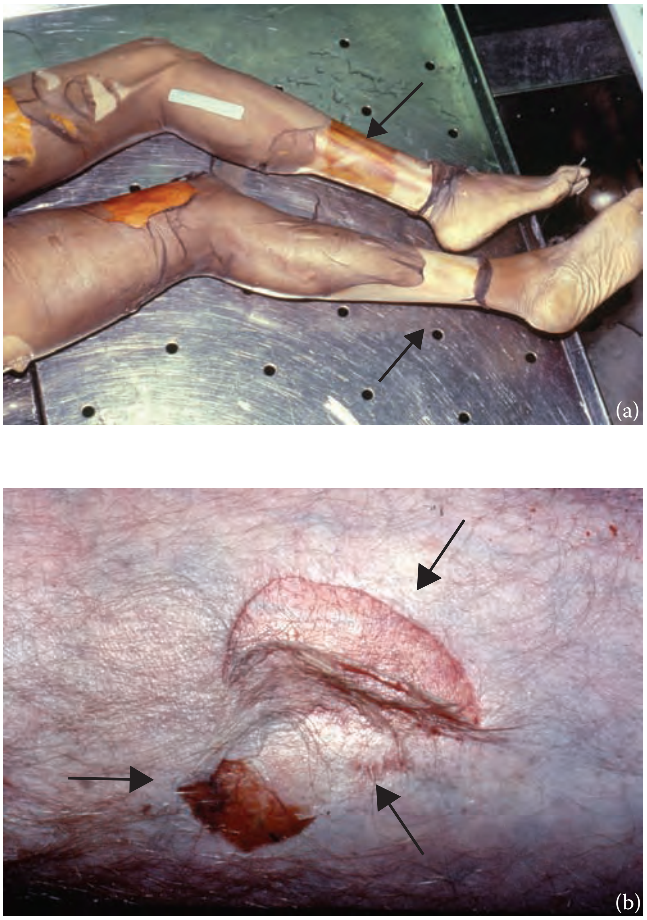

These images, (a) and (b), demonstrate skin slippage that occurred at different times. Note the regions of underlying dermis that are dry and dark, indicating older regions of skin slippage. Also note the other adjacent regions, which are moist, pale, and less dark, indicating shorter duration of underlying dermis exposure to air. This darker discoloration is due to more drying and longer exposure to the air.

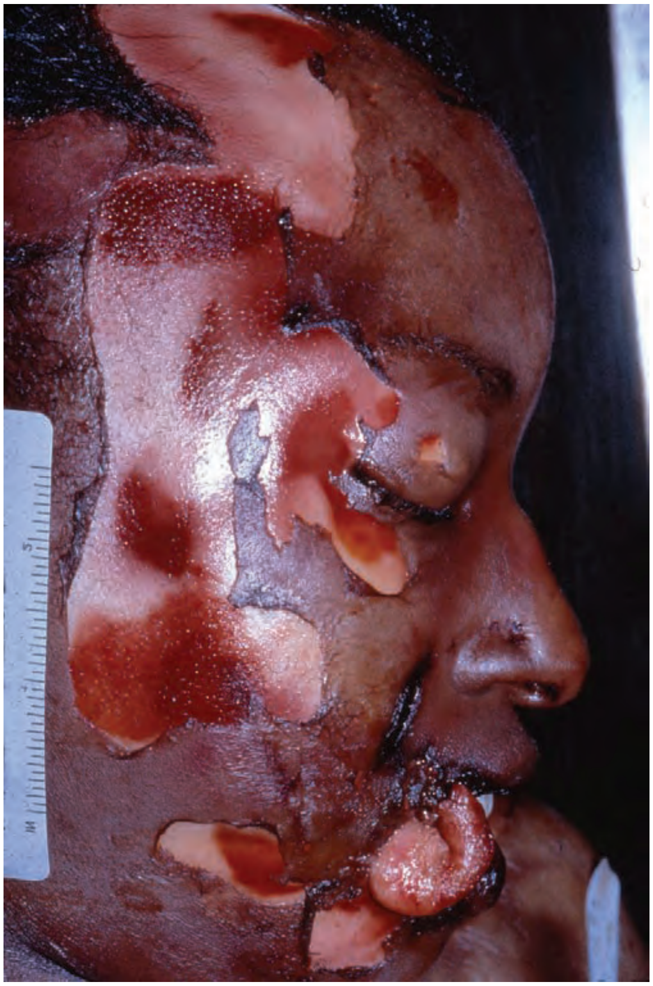

Decomposing homicide victim with multiple decomposing blunt impact injuries. The combination of postmortem change with injuries may make interpretation challenging. Note the red to brown abraded contusions at the individual's face and head, with skin slippage and darker discoloration.

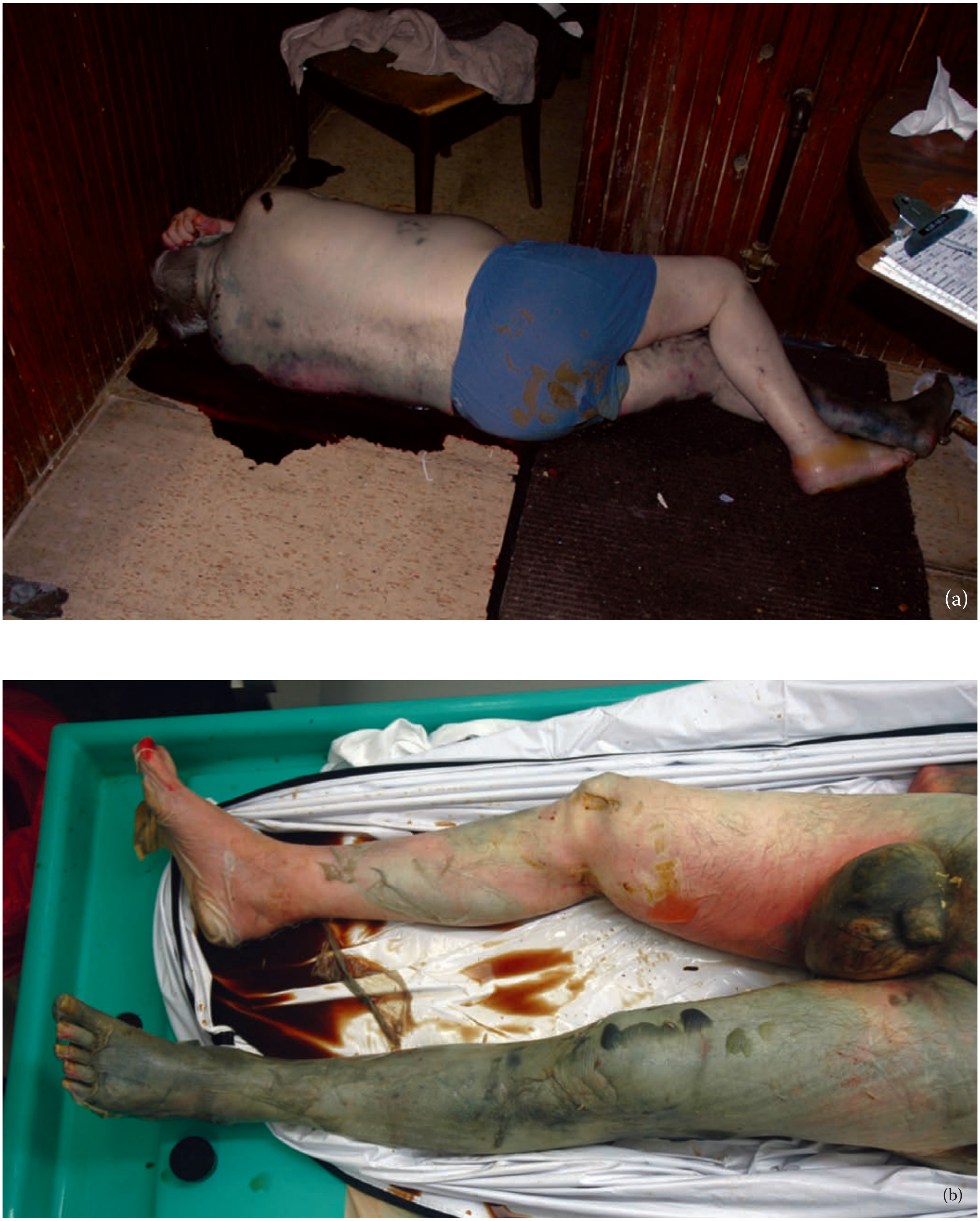

This individual was found lying on his left side with a mild to moderate degree of putrefaction. Note the puddle of purged fluid underneath his upper trunk and head (a). Also note the greater decomposition with green discoloration in the regions of dependent lividity where the body contacted the warm floor. Purged fluid such as this is often misinterpreted as the result of traumatic injury.

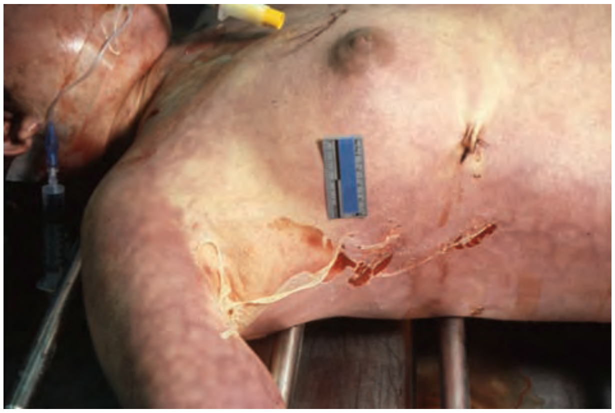

Accelerated putrefaction due to bacterial sepsis at the time of death. This picture depicts the left side of the patient's body with extensive skin slippage as well as green discoloration within the left infraclavicular area. Additionally, an endotracheal tube can be seen protruding from the mouth and taped to the side of her face. Such a depiction might lead one to conclude that overzealous paramedics had worked on a decedent who was decomposing. Marbling is also visible on the lateral and anterior aspects of the left arm. What this case really illustrates is the rapidity with which bacteria, already present within the blood stream at the time of death, can disseminate and propagate throughout the blood system, leading to accelerated postmortem putrefactive change.

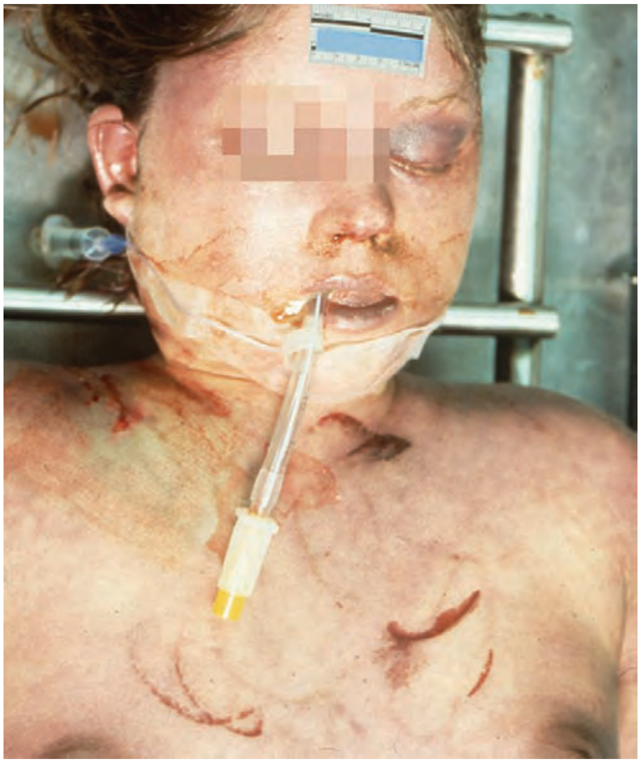

Same case with front view of patient's face, also illustrating green discoloration around right infraclavicular puncture site and defibrillator marks.

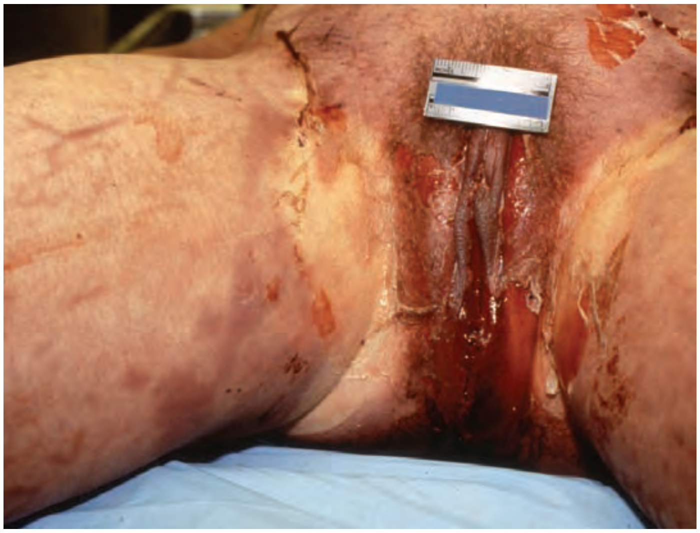

This is the same case, demonstrating skin sloughage and exudation of hemolytic fluid from the external genital region.

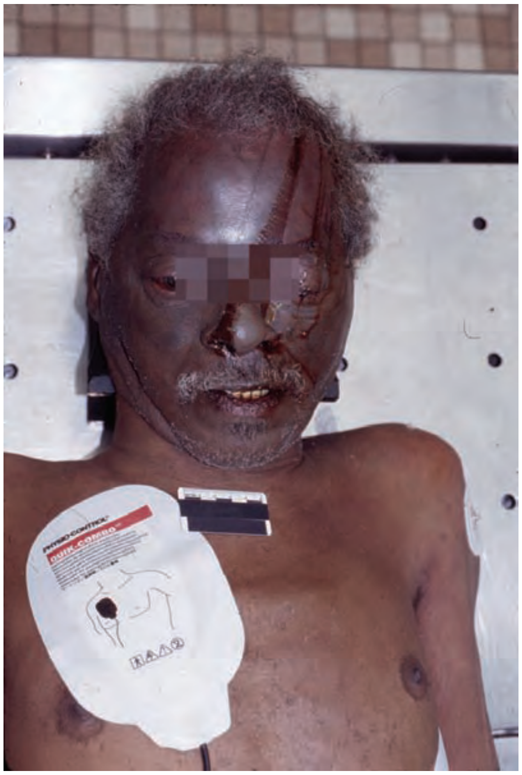

Another case depicting a decedent with dark red-brown discoloration due to postmortem putrefactive change that was accelerated by the probable bacteremic state at the time of death.

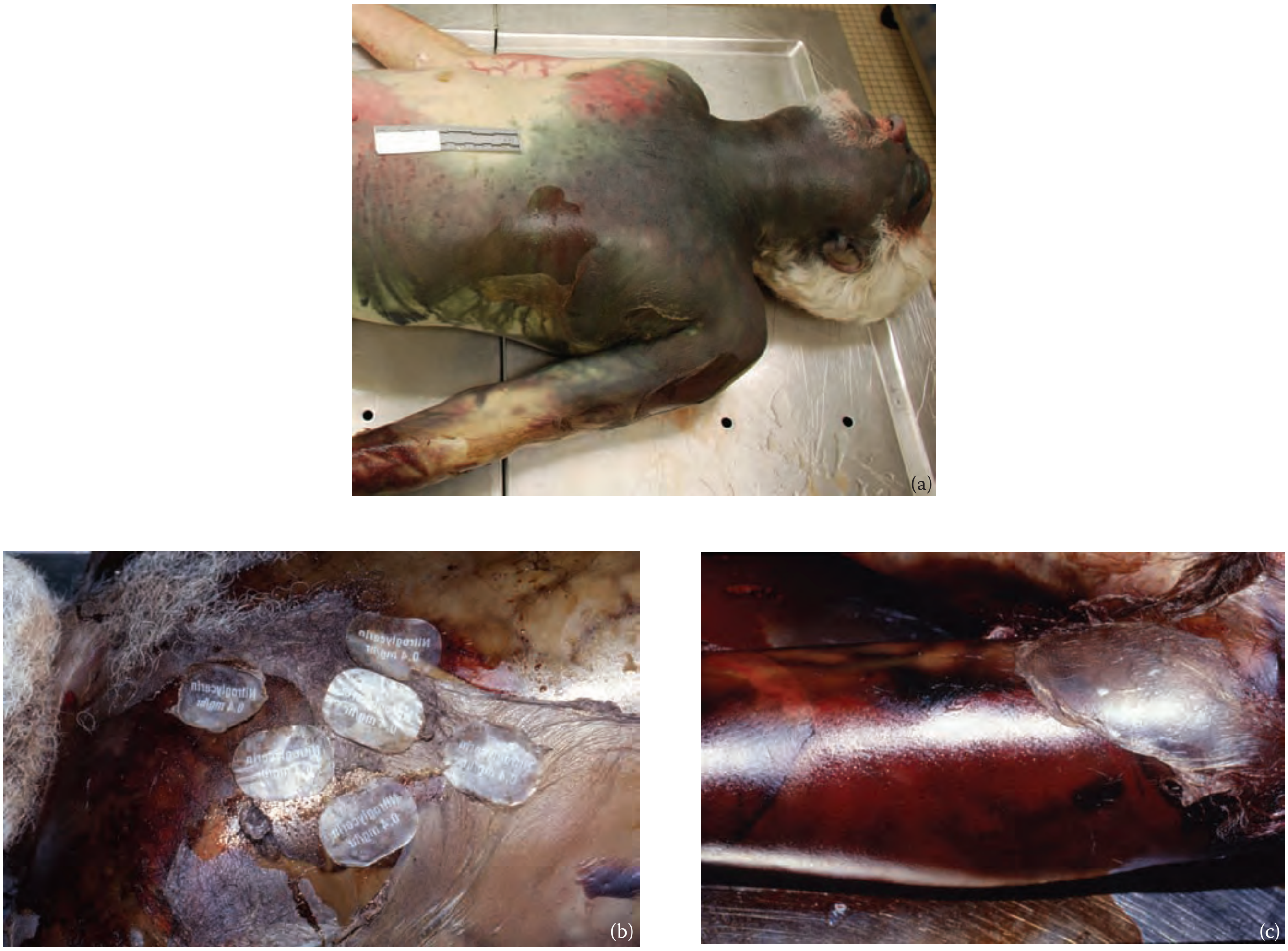

Skin slippage with confluent red to brown discoloration, moderate putrefaction. Note figure (b) shows multiple nitroglycerin patches in an individual with known significant heart disease.

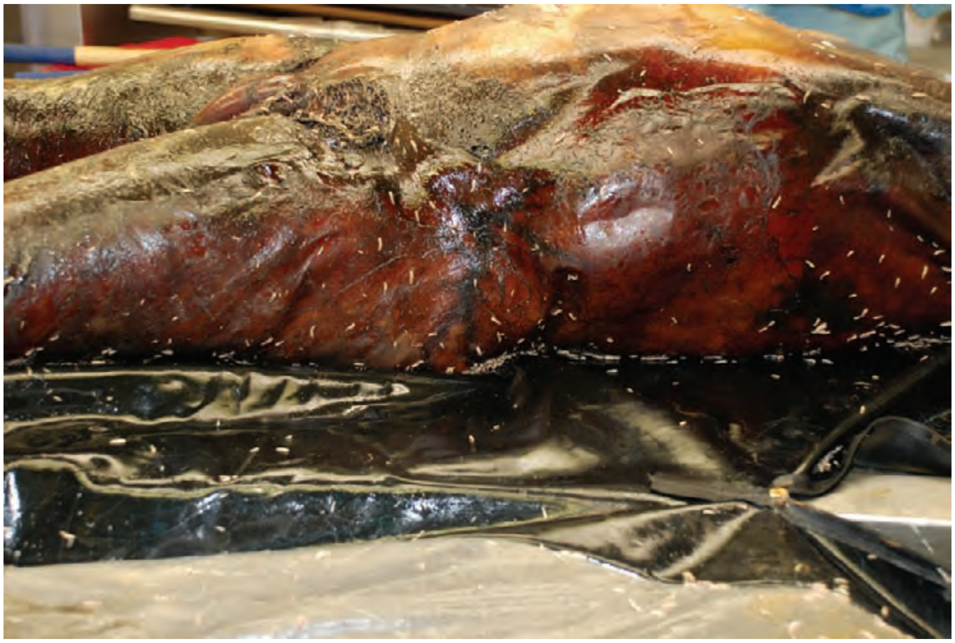

Moderate to marked putrefaction with dark brown discoloration and early mummification.

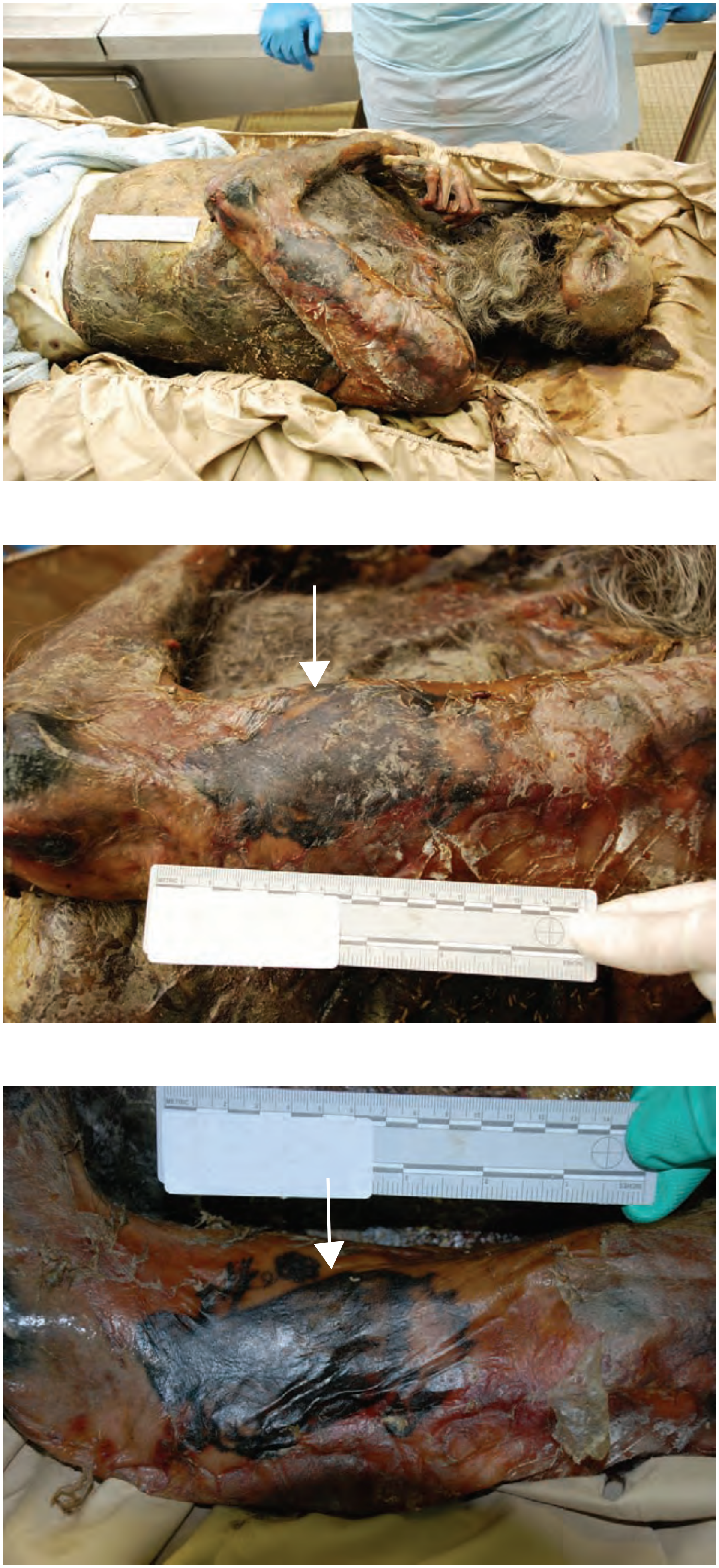

Moderate to marked putrefaction. Identification may sometimes be challenging with advancing decomposition. This individual had a tattoo on his left arm that was initially difficult to view due to putrefactive change. Cleaning of this area with removal of the superficial layers of skin made visualization and thus identification much easier.

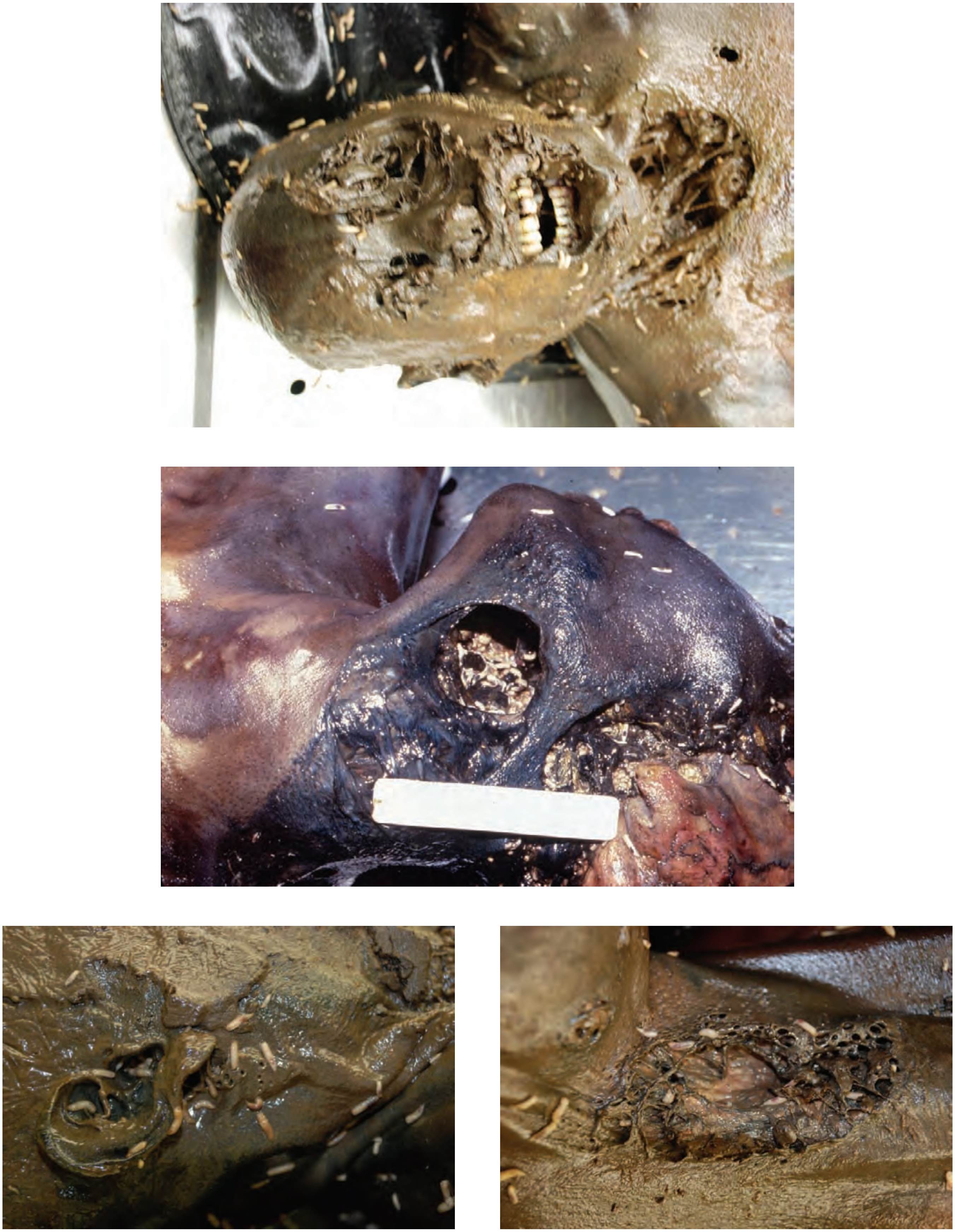

Marked putrefactive change with dark brown discoloration and maggot feeding. Note the small circular perforations caused by maggots tunneling through skin and soft tissue.

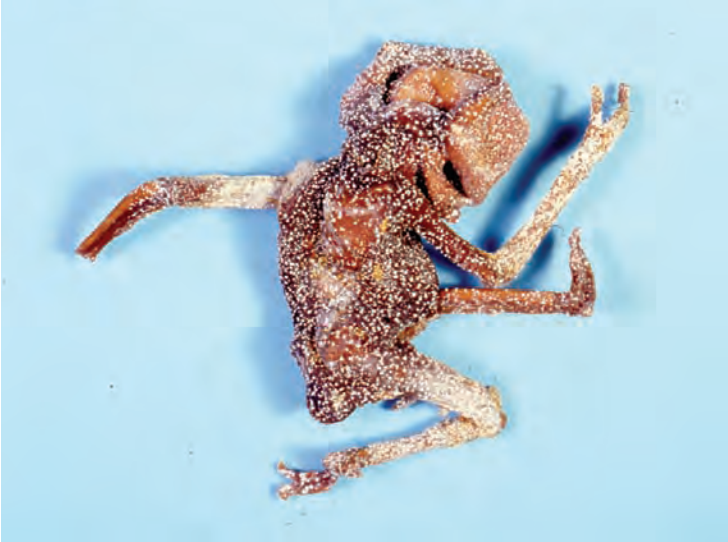

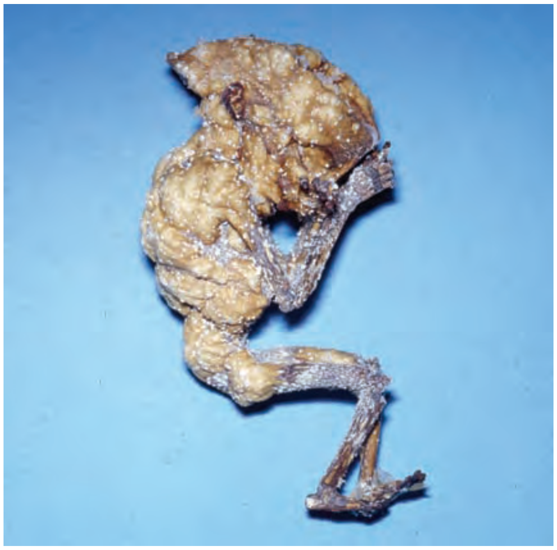

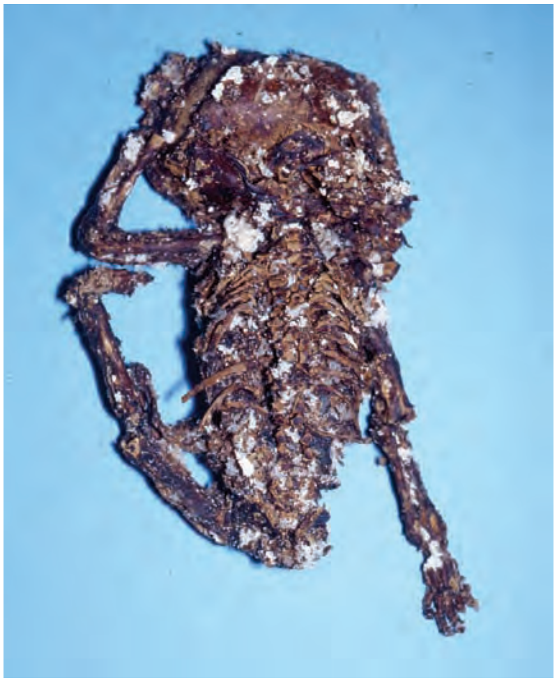

Mummified fetuses retrieved from dried-out formaldehyde containers found within an abandoned building that once housed a doctor's office that closed more than 30 years earlier.

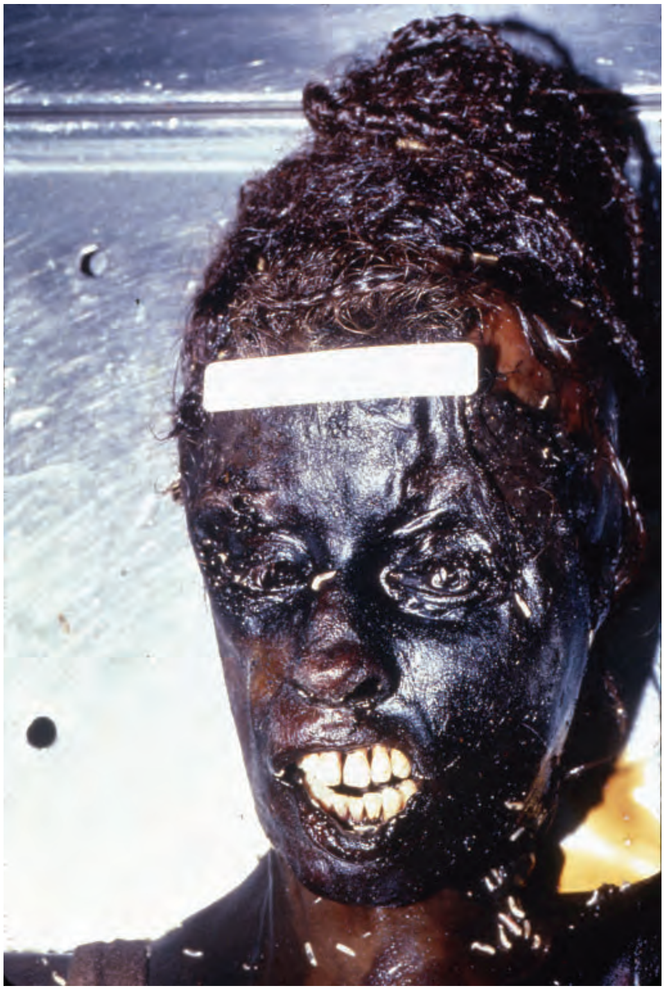

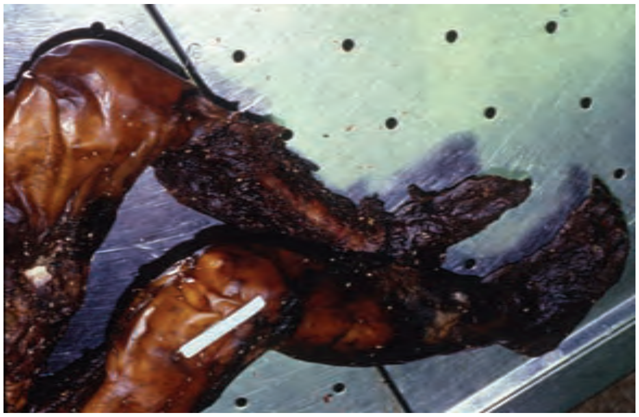

Advanced decomposition with mummification of the entire body. Note the dry dark leathery appearance of the skin.

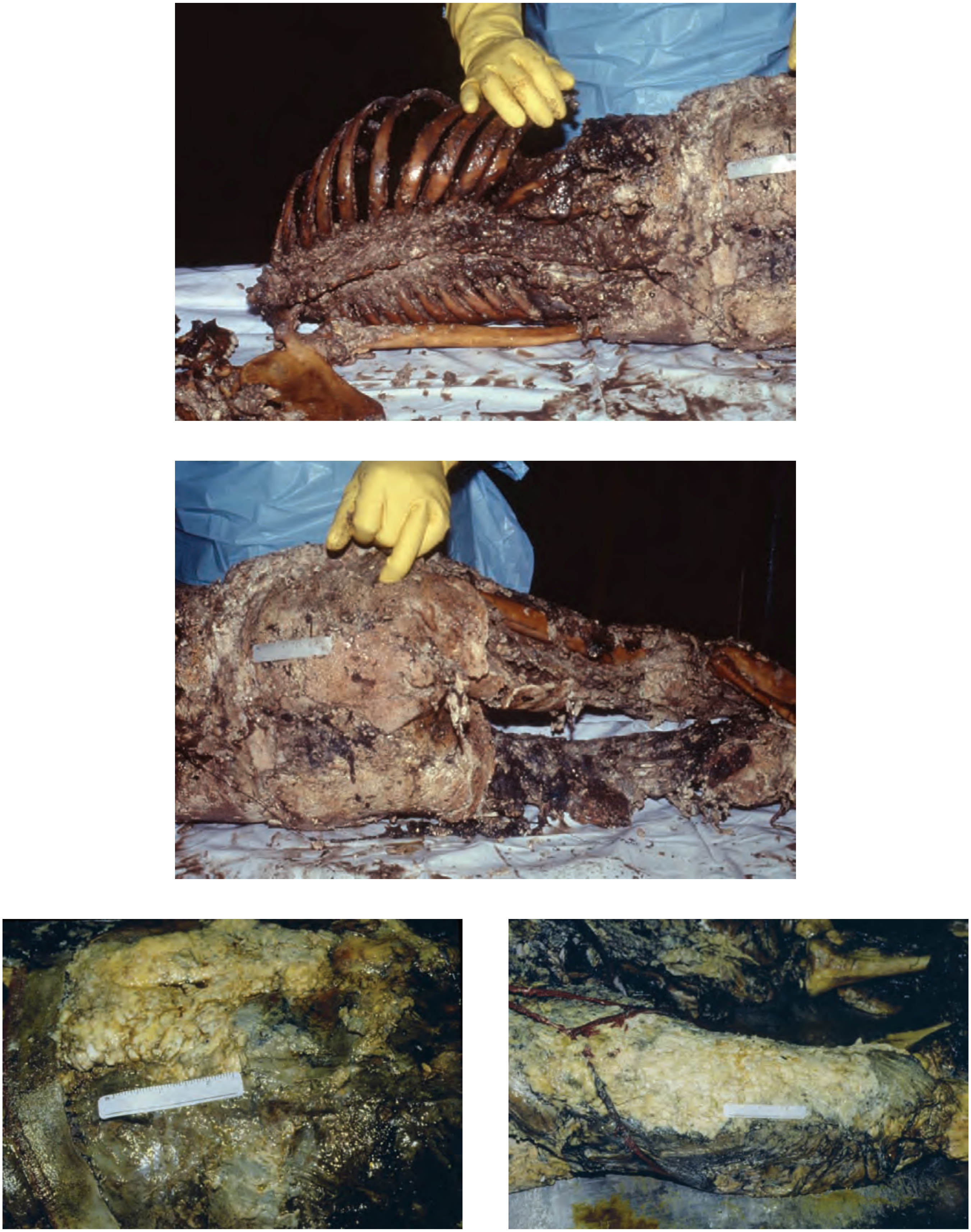

Adipocere is a decomposition process seen with immersions or damp, warm environments. The neutral fats are converted to oleic, palmitic, and stearic acids. Note the white/tan-colored adipocere, which has a waxy feel. In some areas you can see a light sheen of oil on the surface. Once formed, adipocere is resistant to further decomposition.

Skull found tucked under the bottom of a stairwell in the basement of a brownstone in Brooklyn, New York. Note the dried skin at the nose. There was resin and perforation of the cribiform plate. This was the historical remains of an Egyptian mummy skull that was most likely used in religious ceremonies. There were also commingled chicken bones, feathers, and wax.

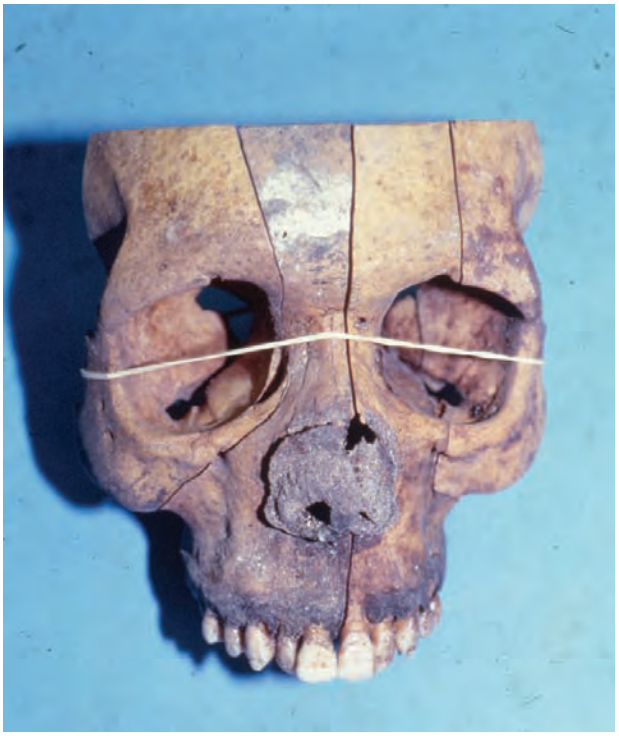

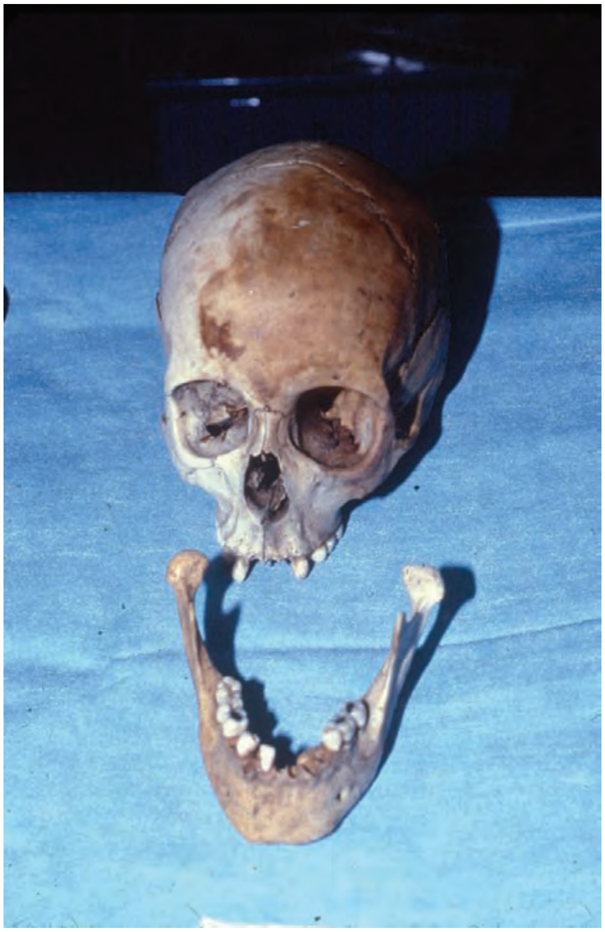

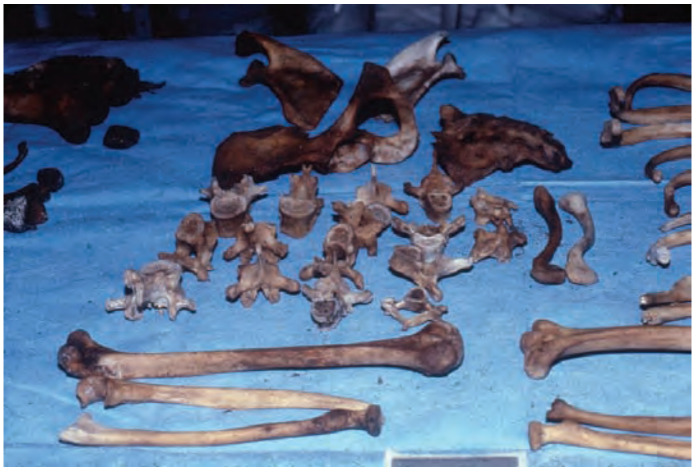

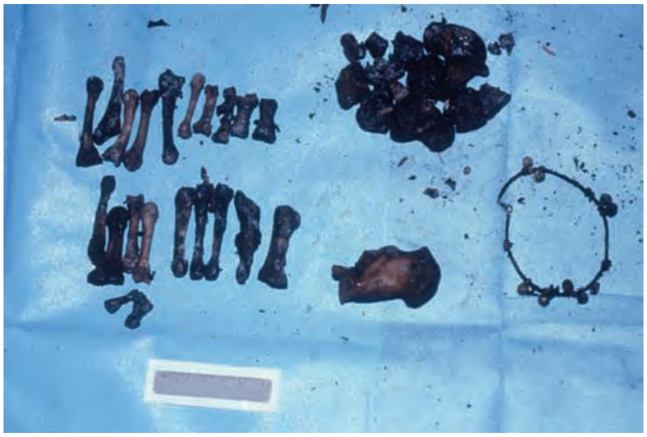

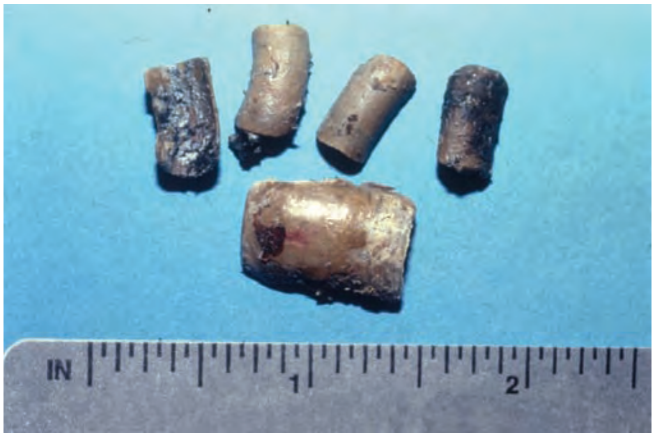

Advanced decomposition with skeletonization from a body dumped in a wooded area approximately 1 year earlier. The last image demonstrates five toenails from inside a shoe.

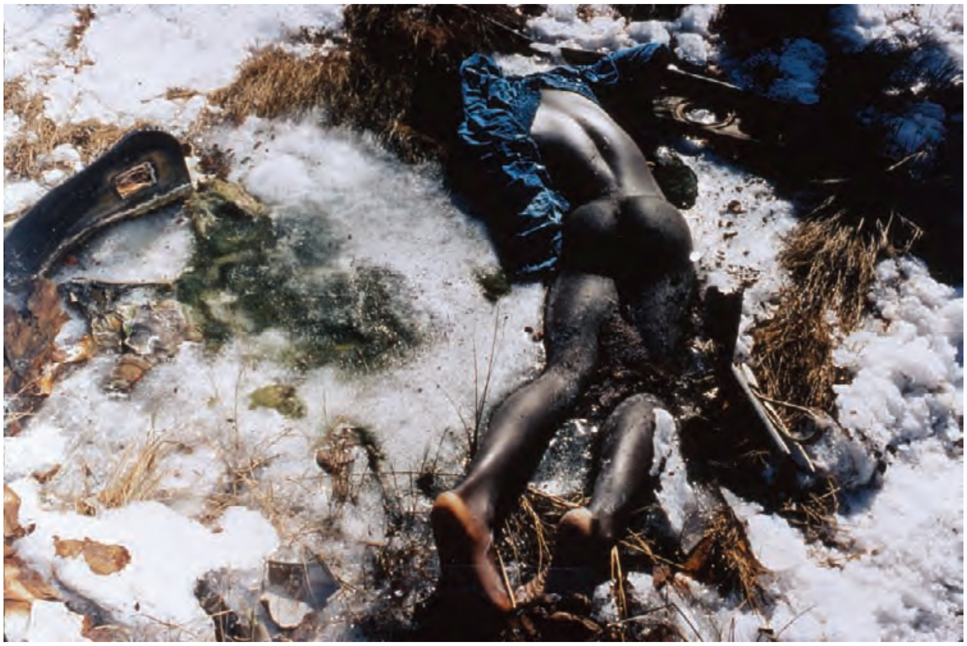

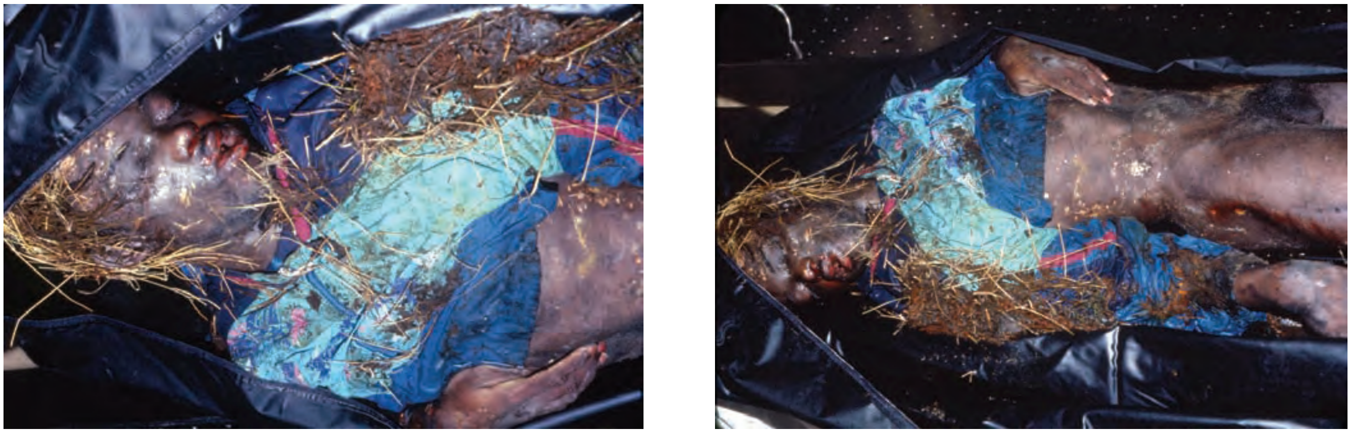

This frozen individual was murdered approximately 1 month prior to being found. His body was in a slight state of putrefaction due to preservation by the freezing winter temperature. Once the body thawed out for autopsy, putrefaction advanced at a markedly accelerated rate.

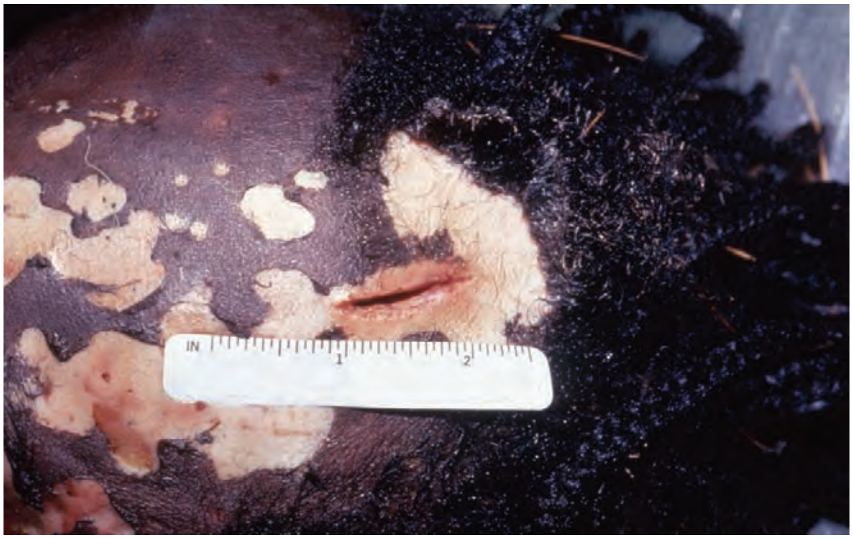

Photograph taken at time of autopsy. Note the advancing putrefaction with skin slippage. Also note the scalp laceration.

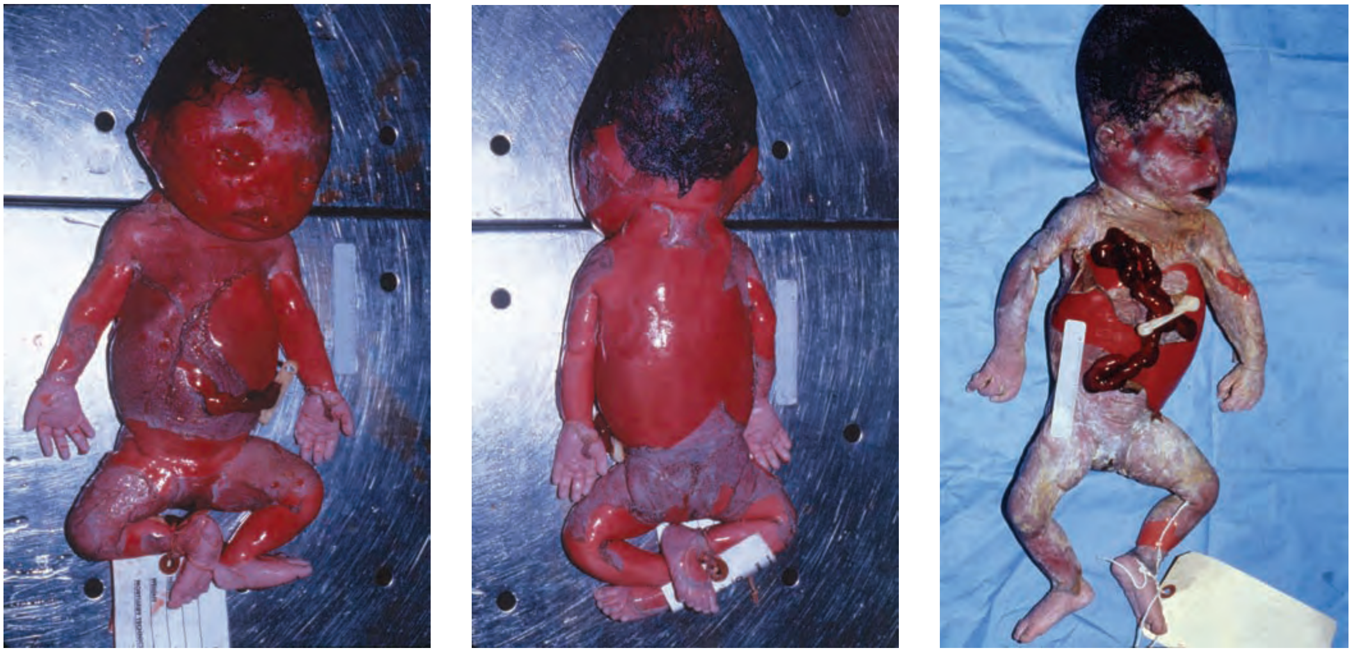

Macerated stillborns with red to brown discoloration. This type of decomposition is autolysis. The womb is normally a sterile environment and there should not be putrefaction unless there is an infection such as chorioamnionitis.

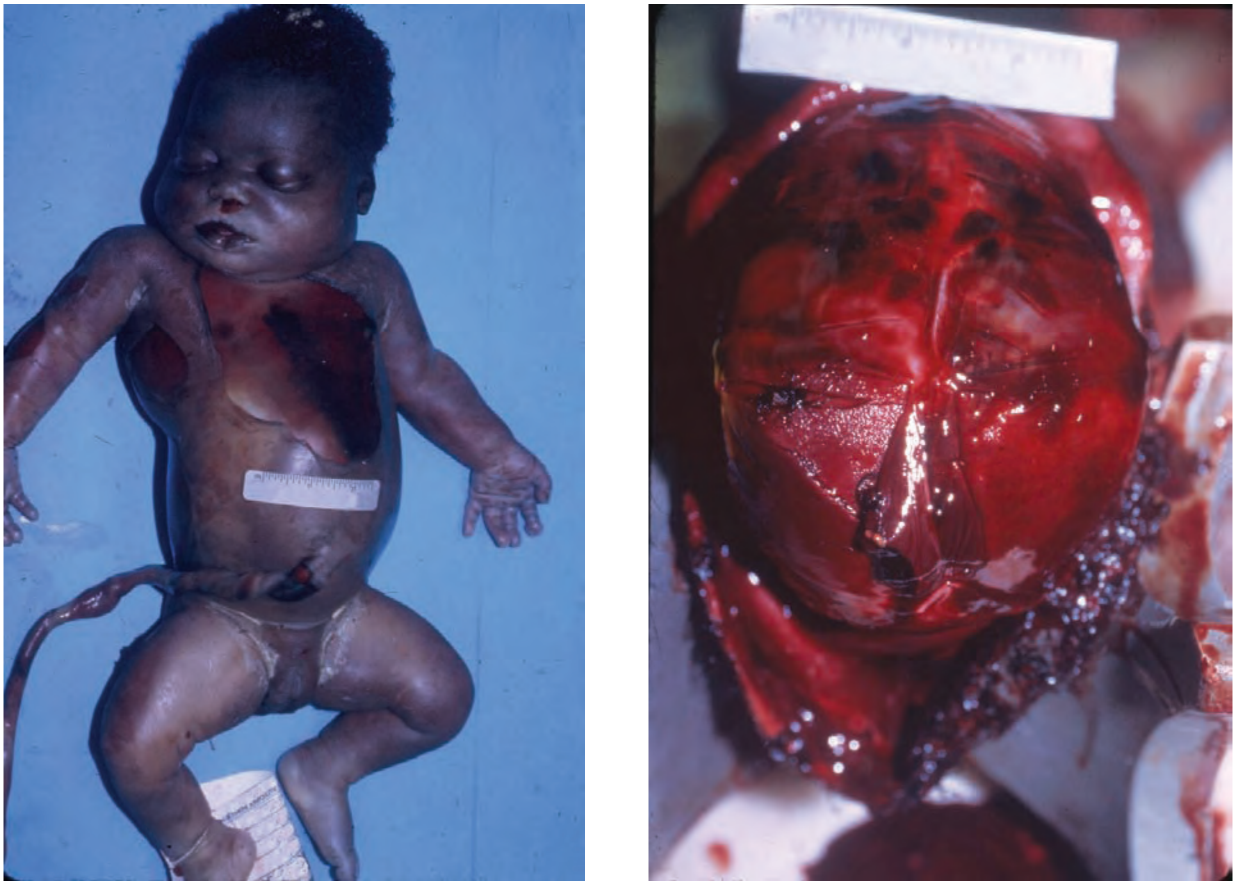

This child was discovered in the back of a garbage truck after being crushed by the trash compactor. The child was commingled with rotting food. Note the green discoloration of the skin due to putrefaction. Note the skull fracture with hemorrhage and dark discoloration due to putrefaction.

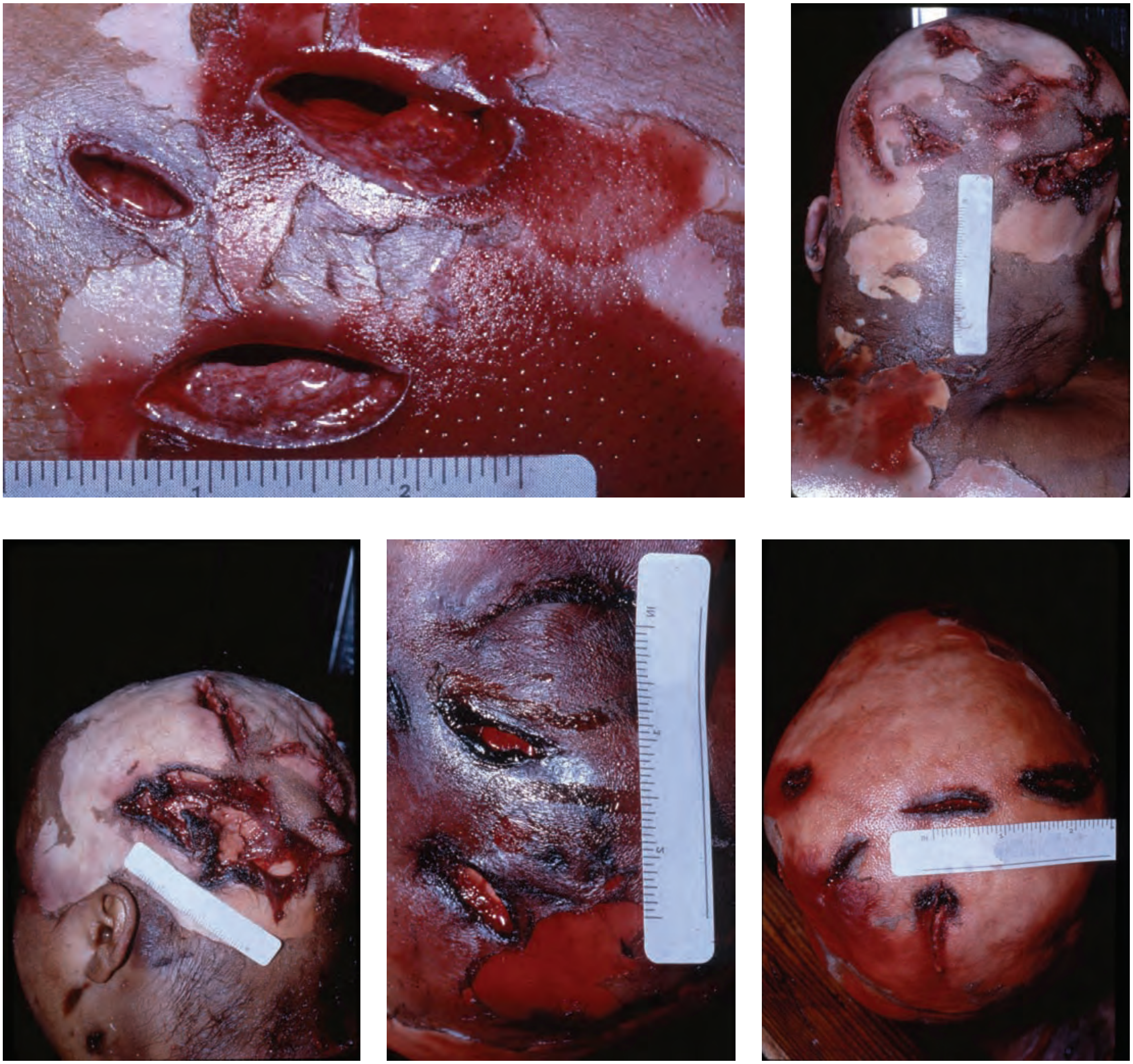

These slides show antemortem injuries that have been obscured by postmortem change. These individuals had multiple stab wounds and lacerations. Note the wound margins are dark and irregular due to drying. This may make wound interpretation challenging and at some point impossible. Also, animals and insects will often more readily feed from injured areas with exposed soft tissue and blood, further obscuring these findings.

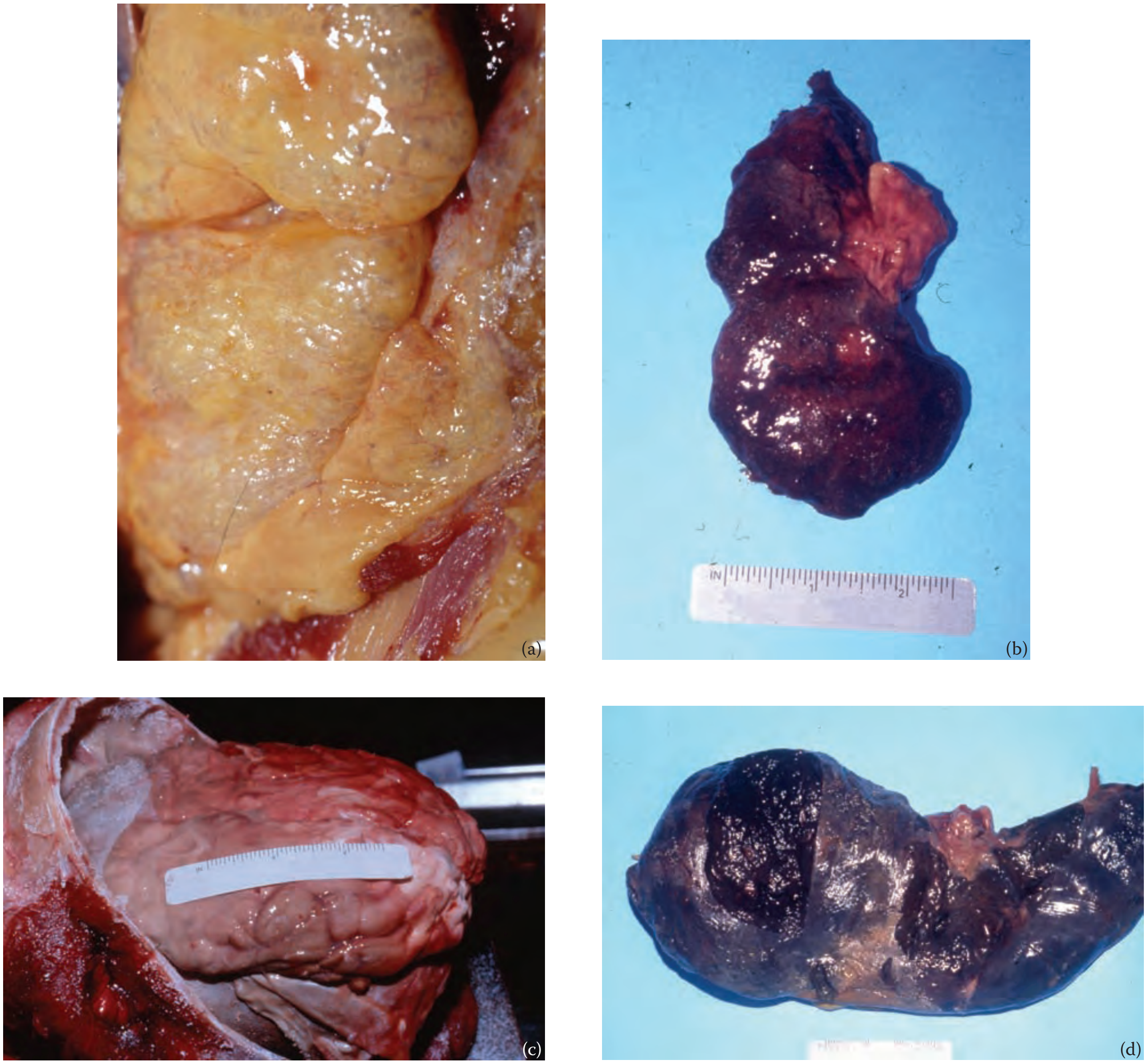

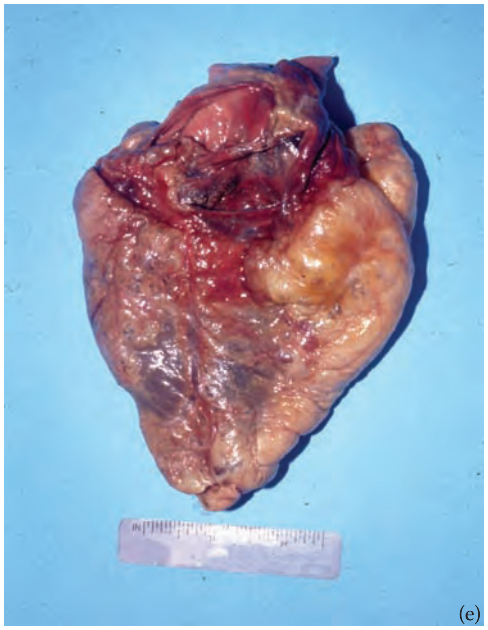

These are examples of moderate putrefactive changes of internal organs with softening, dehydration, shrinkage, and partial loss of architecture. In (a) note the crepitant adipose tissue due to expanding cavities of gas from proliferating microorganism metabolism. Also note in (c) the gray discoloration of the brain with partial loss of architecture. As central nervous system putrefaction advances, the brain will develop into a liquid, oatmeal-like consistency with few or more recognizable structures. Figure (d) is of a putrefying liver and (e) is of a putrefying heart.

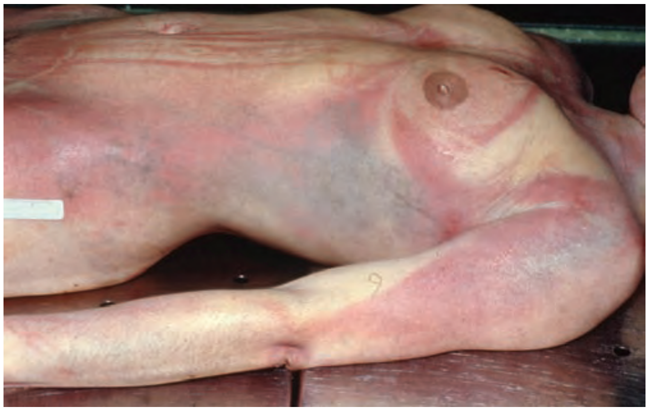

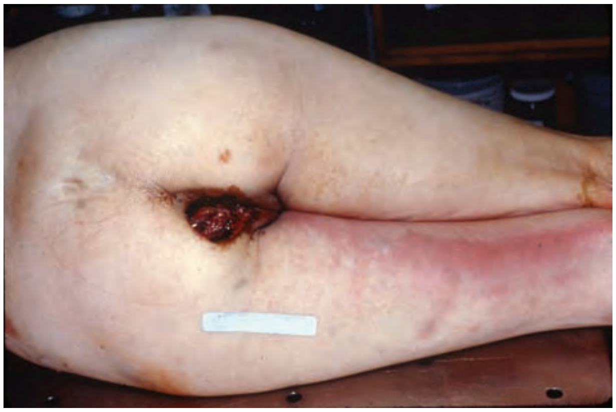

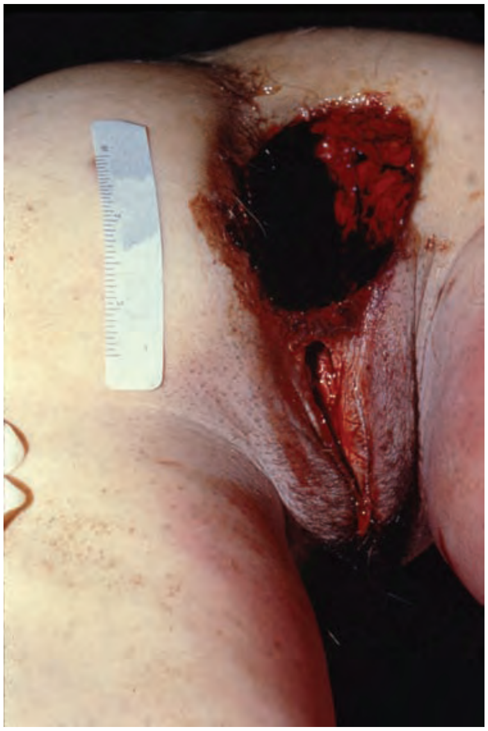

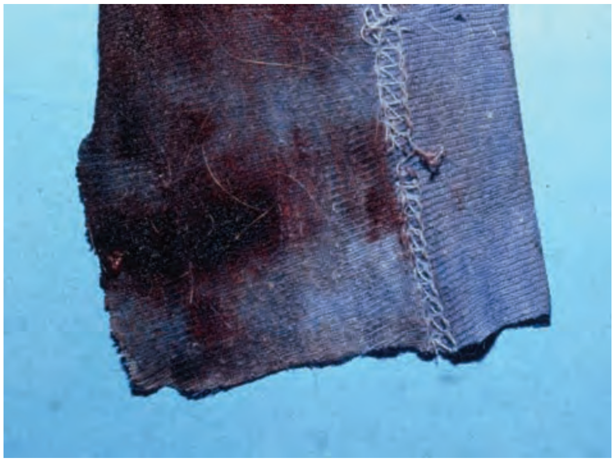

Note the fixed anterior lividity. The decedent was lying prone on a bed with a wrinkled comforter. Postmortem dog feeding occurred around the anus. Autopsy revealed the absence of large portions of intestines. Note that the torn clothing has dog hairs and blood staining. This case was first suspicious for homicidal violence. Large amounts of blood had drained from the body to the bedding, which was further spread about by the dog.

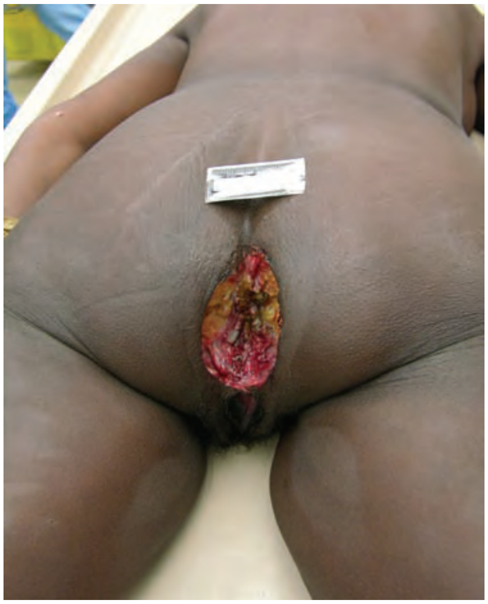

This is another example of postmortem dog feeding, which often initially involves the face, genitals, anus, and areas of wounds.

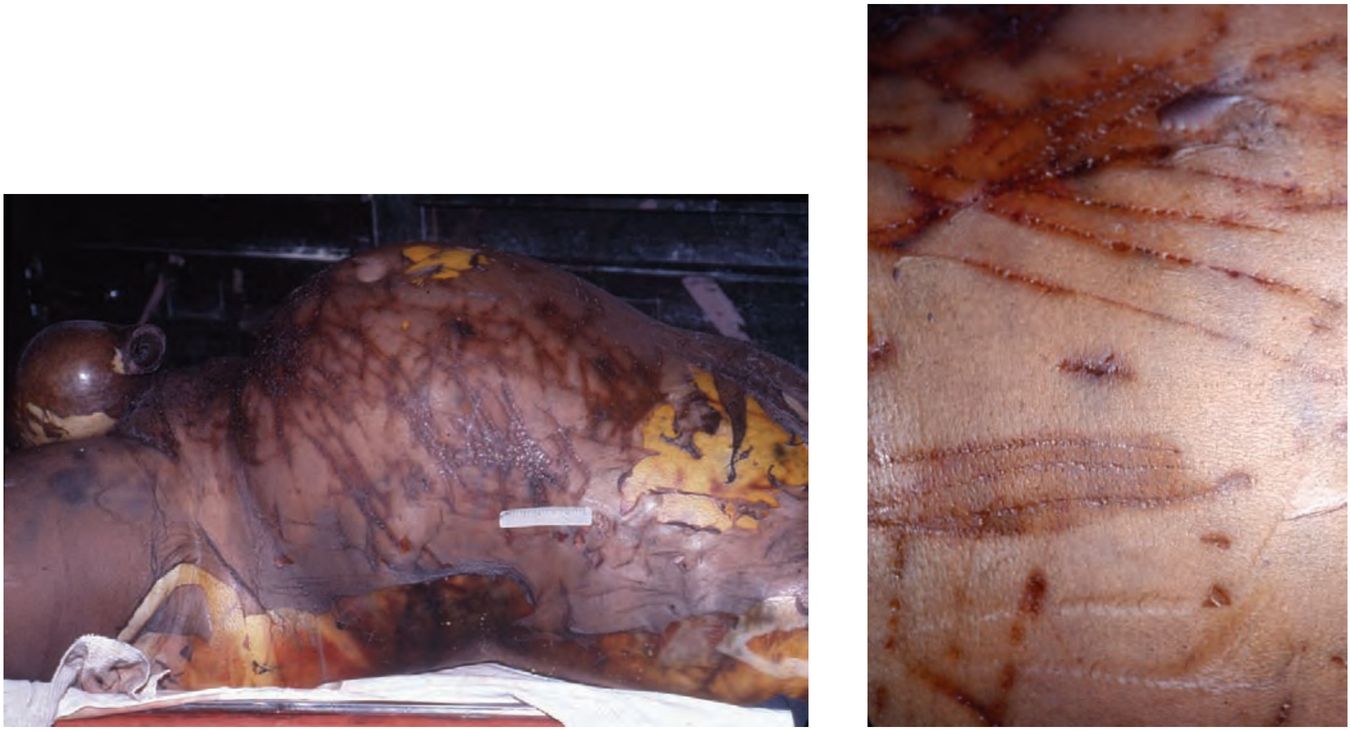

Postmortem animal feeding with dog claw abrasions.

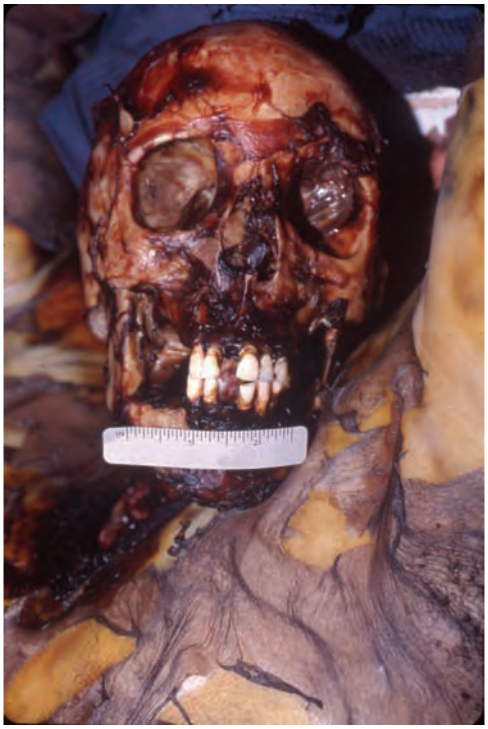

Postmortem feeding by the decedent's dog.

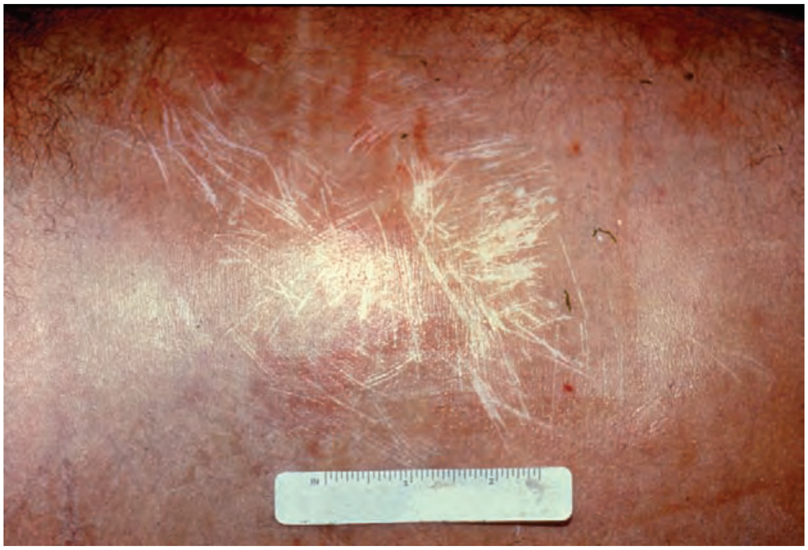

Postmortem anemic superficial linear abrasions from a body retrieved from the Hudson River. This region of the body was noted to be scraping against a rocky surface just prior to retrieval from the water.

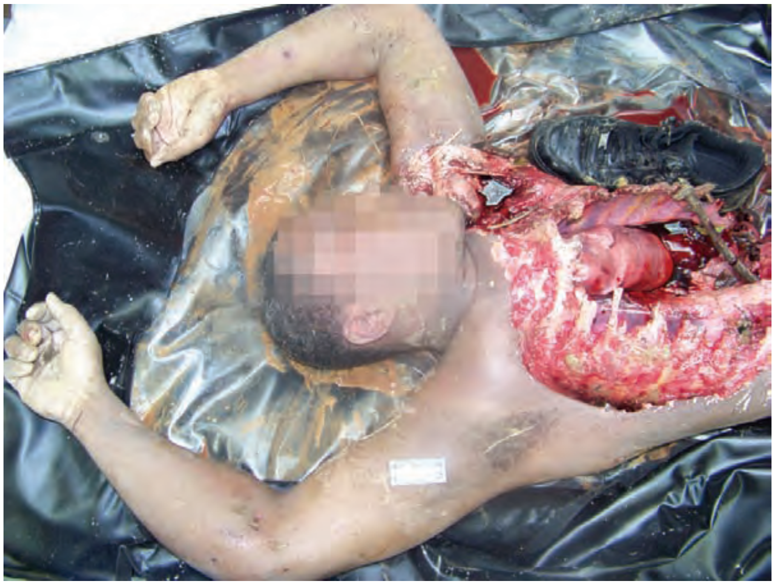

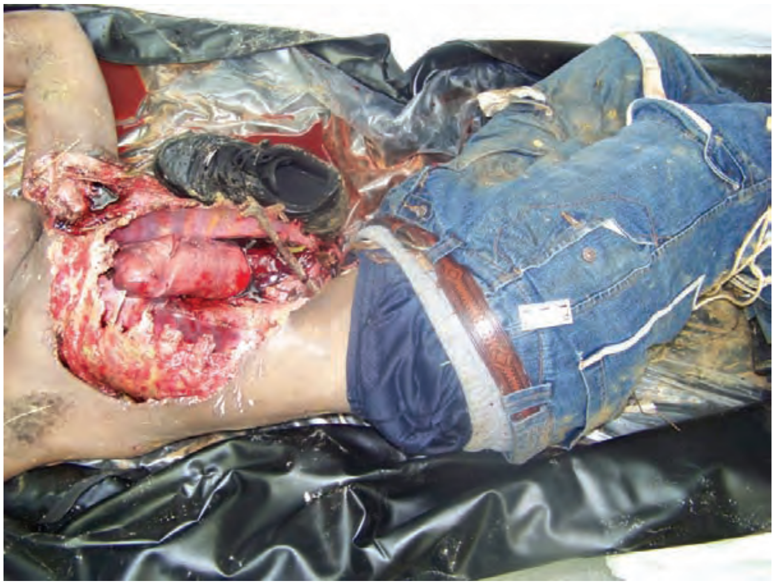

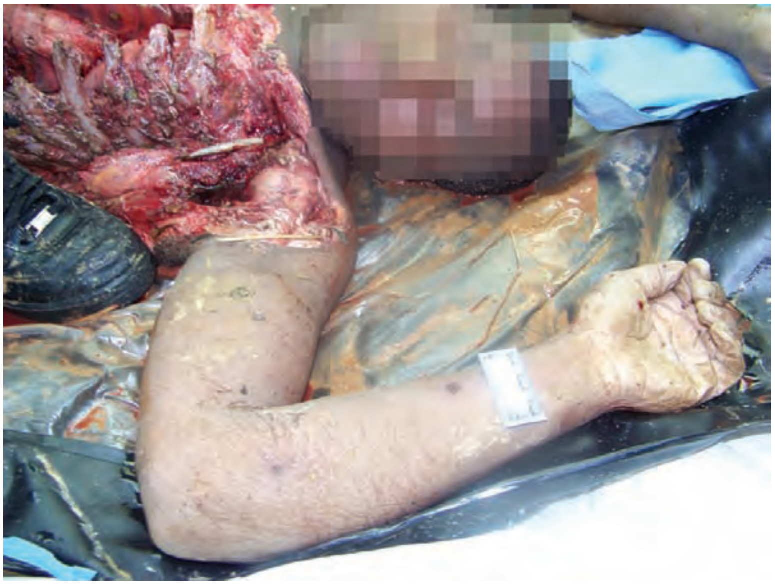

Larger carnivore activity is common in rural and suburban areas. In many cases, there may be activity from several different animals in the same environment. In this case, the damage appears to mostly be from the canine family (likely wild) with evidence of tearing of the skin and ribs. There is no obvious vital reaction, indicating the damage occurred postmortem. In these cases, one must consider whether the postmortem activity has obscured injuries sustained during life.

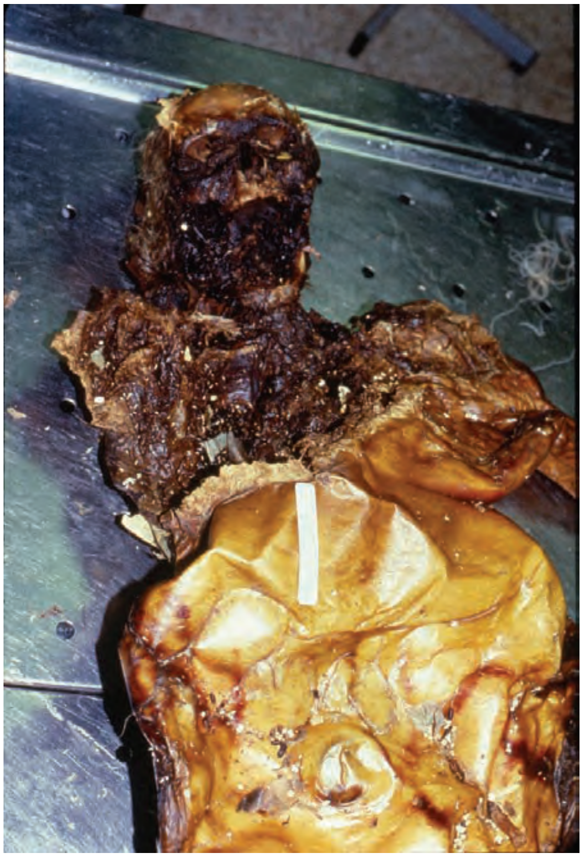

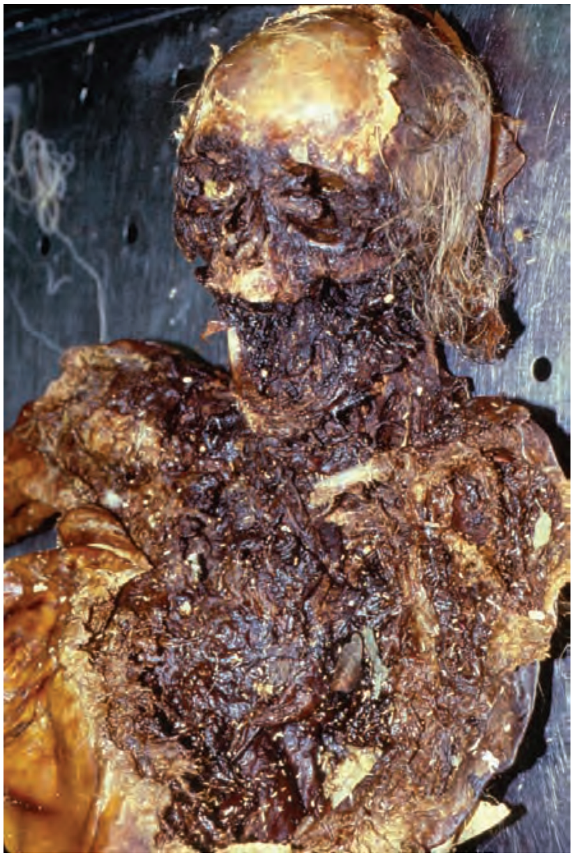

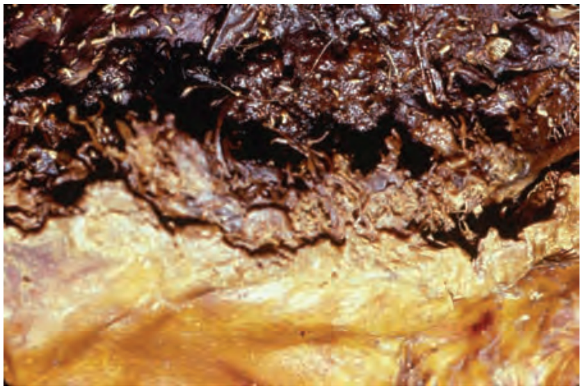

Advanced decomposition with mummification and animal feeding by the decedent's cats. Note the margin with claw marks and a scalloped border from feeding. One arm, a larger portion of the chest and most of the internal organs were absent.

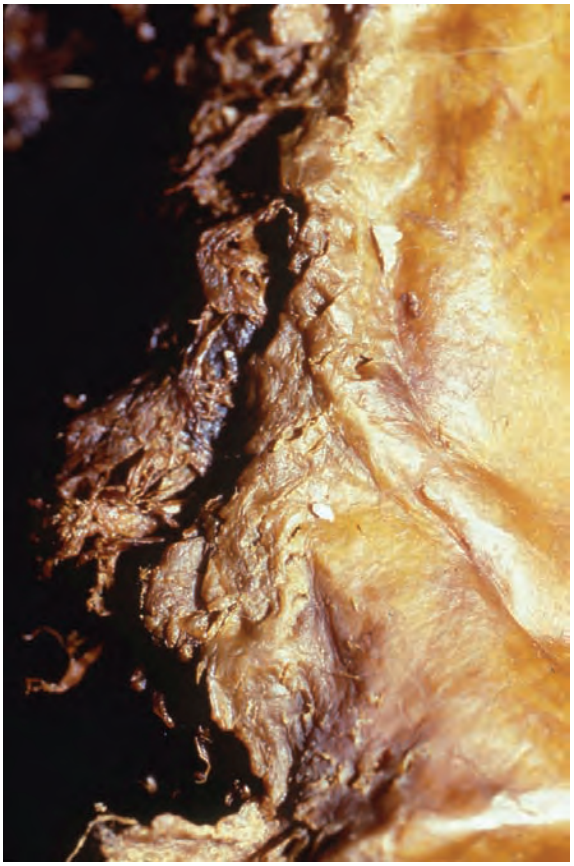

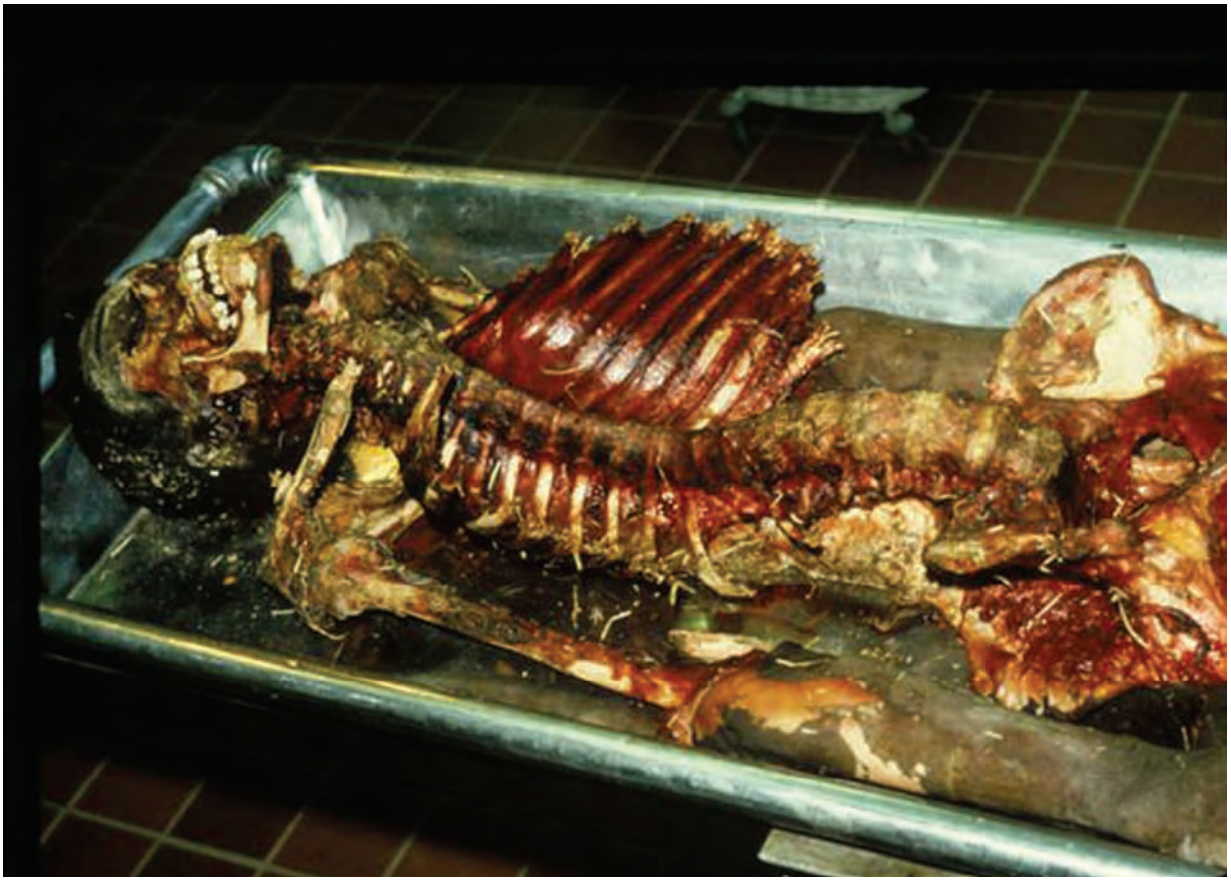

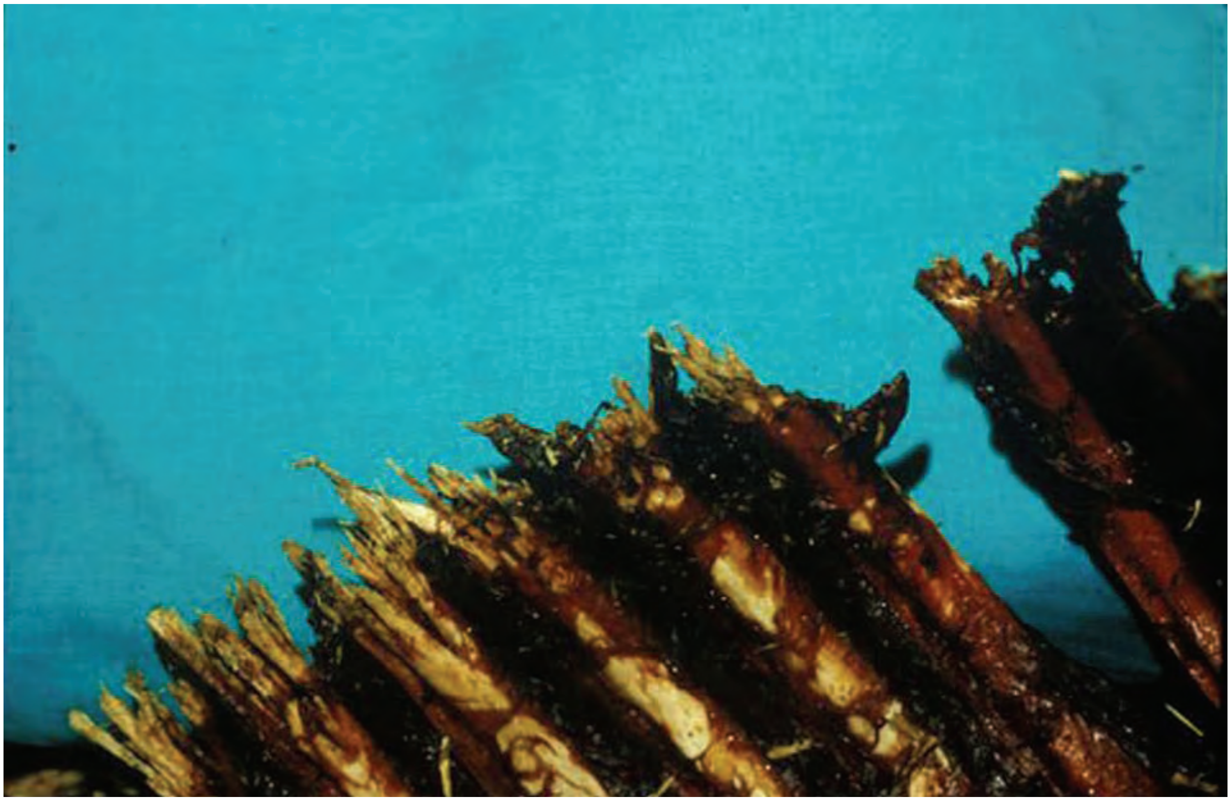

In this case the carnivore activity was caused by domestic pigs.

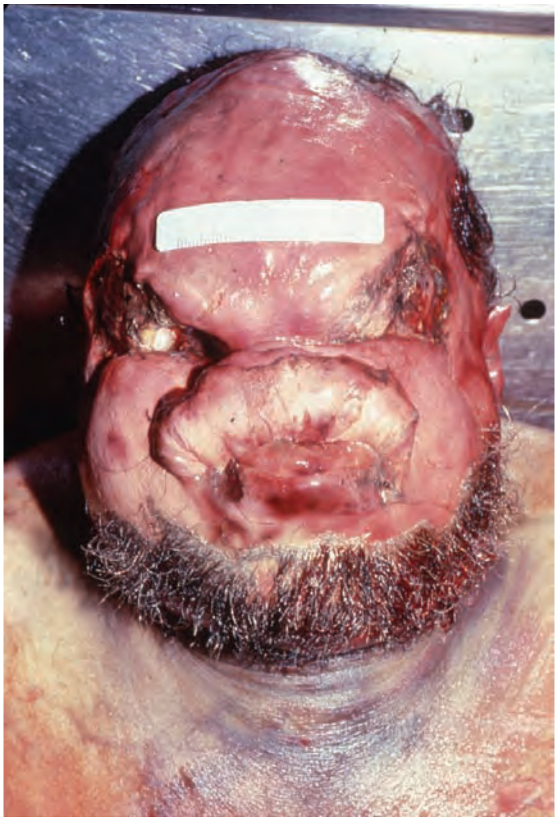

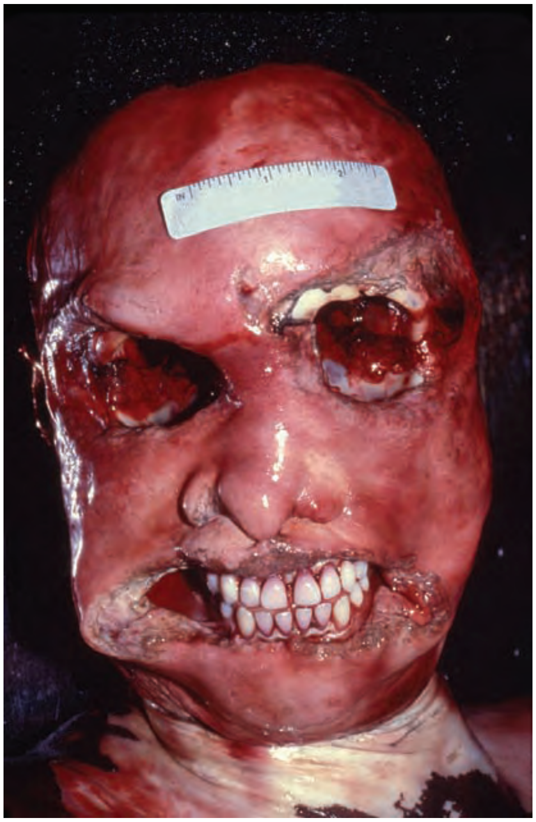

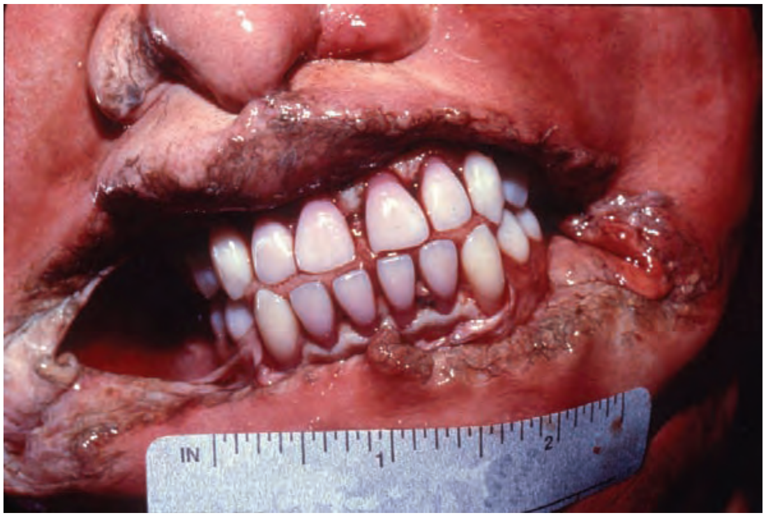

Postmortem feeding from fish and crustaceans found on these bodies retrieved from the ocean.

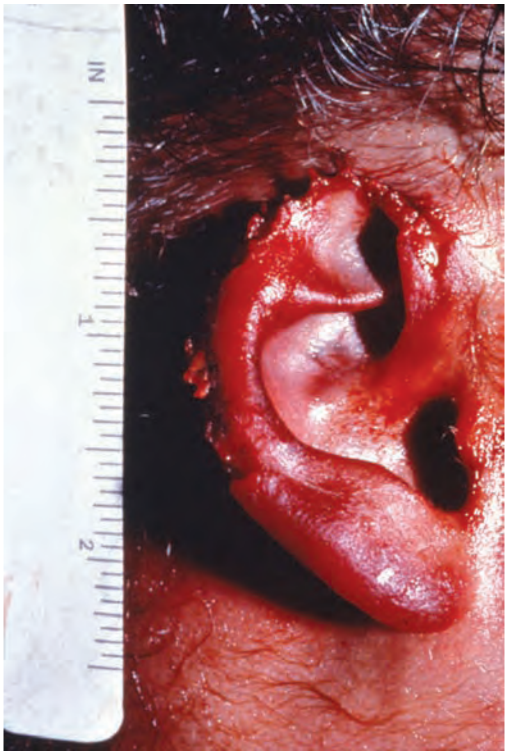

Postmortem feeding of the ear due to mice and rats.

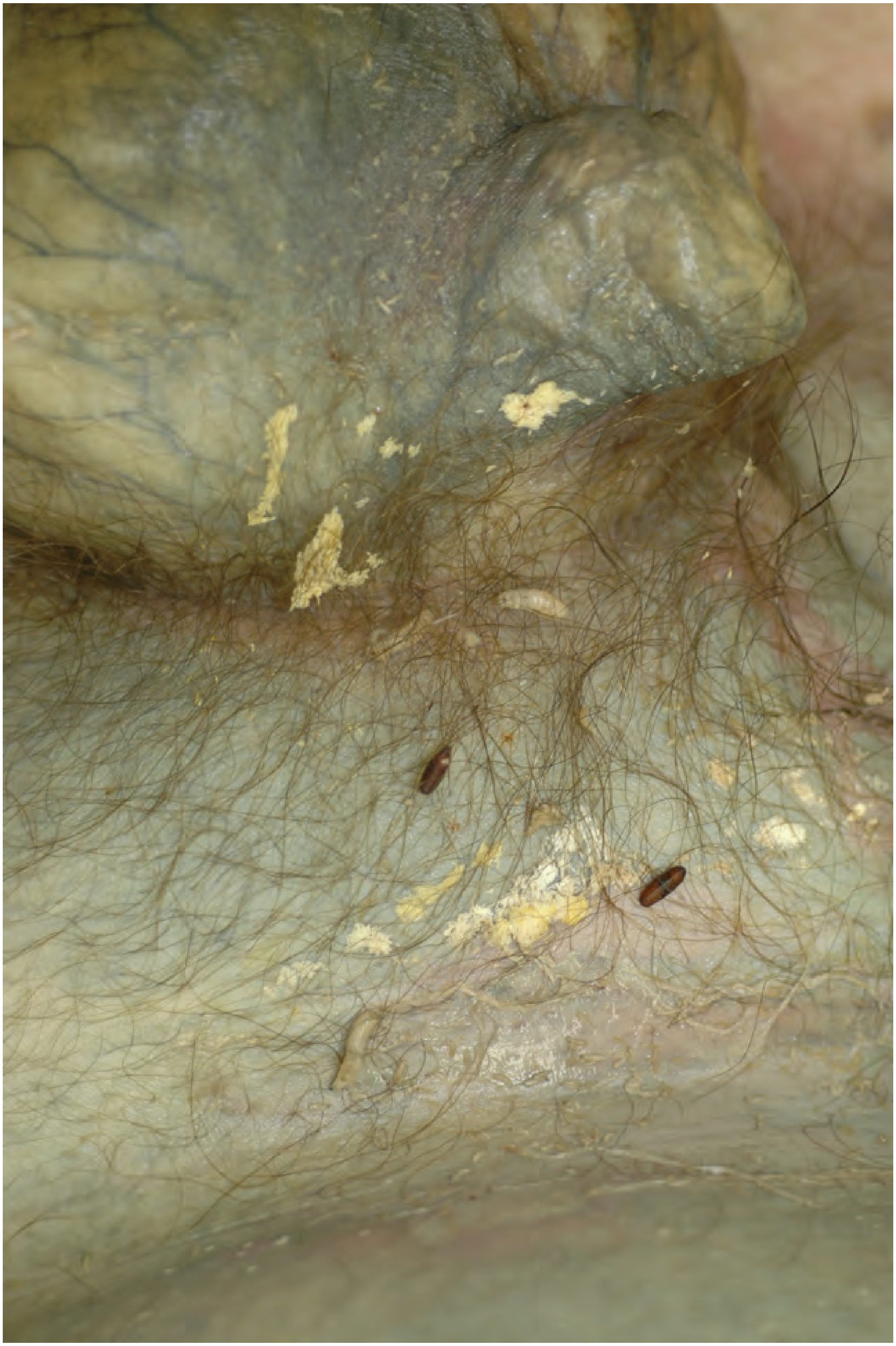

Note the varying stages of blowfly development. Eggs have an appearance similar to grated cheese or sawdust. When the eggs hatch, maggots develop and are shown in this picture in varying sizes as small, white, and wormlike. The dark brown pupa cases from hatching flies are also apparent. This entire cycle is accelerated in hotter temperatures.

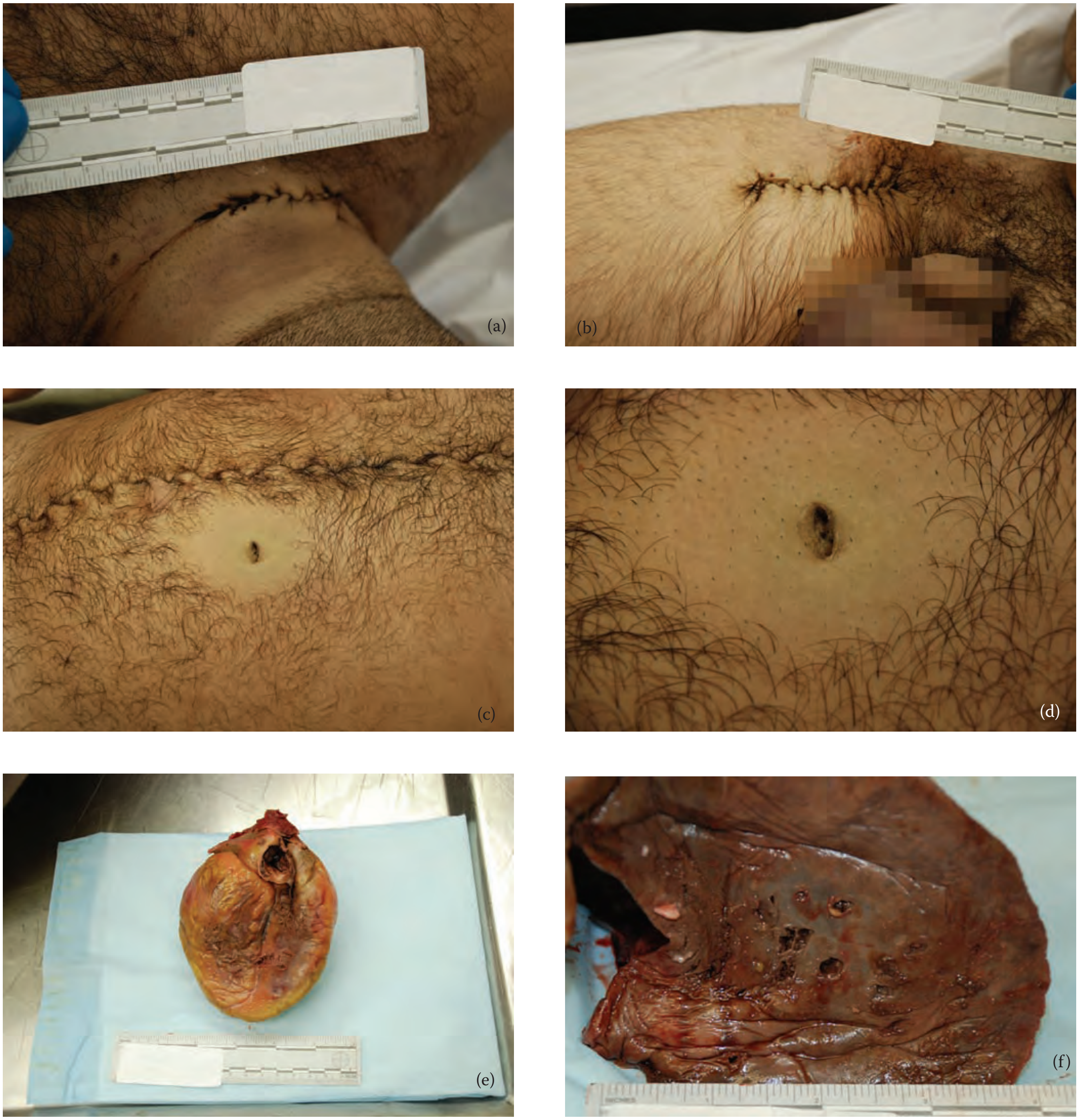

Embalmed body. Figures (a) and (b) demonstrate sutured incisions made by funeral directors to gain access for blood removal with addition of embalming fluid. Figures (c)–(f) demonstrate multiple "trocar" (a metal rod used to infuse embalming fluid) puncture marks through the body's surface and organs. These may be misinterpreted as injuries.

Pediatric Forensic Pathology

5

MICHAEL J. CAPLAN AND CHARLES A. CATANESE

Introduction

Pediatric forensic pathology is the name for the unofficial and poorly defined subspecialty of forensic pathology that focuses on the evaluation of sudden, unexpected, unexplained or traumatic deaths in children, infants, neonates, and even fetuses. By definition, therefore, this discipline requires specialized knowledge of and familiarity with the various stages in growth and development that occur during childhood, infancy, and intrauterine life, along with the differential vulnerabilities of these periods to various insults, and finally, the unique patterns of injury that characterize these different periods.

Perhaps one of the most illustrative examples of this concept is the assessment of nonaccidental or inflicted head trauma in infants and very young children. An understanding of this type of injury would not be possible without being thoroughly acquainted with the following concepts: (1) the elasticity of the infant's scalp, such that it could potentially sustain a serious blunt impact without necessarily manifesting the impact in the form of a subgaleal contusion; (2) the thin, pliable, unilaminar quality that renders the skull less susceptible to fracture but also allows it to transmit forces to the underlying brain more readily; (3) the broad, shallow skull base in infants, which facilitates rotational movement of the brain and lowers the threshold for diffuse axonal injury; (4) the incompletely myelinated infant brain, with a gray- and white-matter water content substantially higher than that of an adult, imparting a consistency of unset gelatin and making the brain more vulnerable to shearing forces; and (5) the top-heavy calvarium and the weak, underdeveloped neck muscles that fail to effectively dampen the oscillations that are initiated when rotational movement of the brain begins. It is the interface of pediatric and forensic pathology that allows the most complete and comprehensive understanding of these concepts to occur; however, very few individuals actually have such expertise in both fields. As a result, forensic pathologists may have to rely on pediatric pathologists and neuropathologists when they are evaluating difficult, complex, or problematic cases involving deaths in the pediatric population.

Of the many scenarios encountered by the pediatric forensic pathologist, none is as enigmatic as the seemingly healthy infant who is found unresponsive after being put to sleep sometime earlier by a parent, sibling, or other caretaker, either in a crib, a child or adult bed, a couch, or some other location, often in a prone (face-down) position, and with soft bedding, pillows, and clothing. Sudden infant death syndrome (SIDS), a term that was coined in the early 1960s to remove the stigma of an unforeseeable and unpreventable death of an infant from that baby's parents, has now come full circle and runs the risk of being used as a "wastebasket" term for any infant death that is not obviously explainable by a disease or injury. The time-honored external and internal autopsy findings of froth (pulmonary edema) at the nares, intrathoracic (thymic, epicardial, or visceral pleural) petechiae, moderately congested organs, and an empty urinary bladder, while characteristic, are in no way specific for or pathognomonic of SIDS, and have been observed in a wide variety of other types of deaths. Forensic pathologists and pediatricians are just beginning to gain some appreciation of the myriad of factors involved in achieving respiratory control, arousal, and autoresuscitation.

Even in this modern age of molecular medicine, SIDS remains largely an unexplored entity. Thus, the forensic pathologist probably makes the greatest contribution in this area by identifying which deaths do not lie within the spectrum of SIDS. To do this effectively demands a carefully conducted medicolegal death investigation, a meticulously performed autopsy, and the use of ancillary studies, including postmortem toxicology, microbiology, chemistry, radiography, and metabolic screening. Only after all of these studies have been completed and integrated with the autopsy, scene, and anamnestic findings can the pathologist make an opinion about the cause and manner of death.

Other situations necessitating specific pediatric forensic expertise include the assessment of sudden unexplained deaths in otherwise healthy children, either during sleep or while awake, at rest or following exertion; sudden deaths in infants or children who have undergone operations for congenital heart disease; unexpected or unexplained deaths during or following diagnostic or therapeutic procedures; complex child abuse cases involving a multiplicity of injuries separated over space and time; deaths from genetic diseases that might mimic child abuse, such as osteogenesis imperfecta; and unexplained deaths following apparent minor trauma. In all of these scenarios, knowledge of the specific disease processes or related therapies is required in order to achieve the most complete understanding of the circumstances that culminated in the infant's or child's death. Other types of deaths affecting infants and children that are similar to those involving adults are typically fatal injuries sustained in motor vehicle crashes, either as occupants or pedestrians, although the considerable disparity in height, weight, and center of gravity often produces patterns of injury in infants and children that are distinctly different from those in adults.

Fatal child abuse presents primarily in two forms: (1) as the culmination of a repetitive pattern of inflicted injury, creating wounds of varying ages, in different parts of the body (what has been coined "the battered child syndrome"); and (2) as an impulse act by a caretaker who, following an act of urinating, defecating, or vomiting by the infant or child, lacks the self-control to prevent anger and frustration from translating into physical violence. The pediatric forensic pathologist must know what studies are necessary to perform in order to document not only what did happen, but to exclude all of the other reasonable possibilities that undoubtedly will be introduced by the defense in the adjudication of the case.

In some ways, the pediatric forensic pathologist often walks a diagnostic tightrope—deciding between SIDS and smothering, between inflicted closed-head injury and spontaneous intracranial hemorrhage, or between preeclampsia and maternal blunt trauma precipitating placental abruption. Toward that end, it is equally important to be aware of the limitations of a particular case and not offer opinions beyond what the circumstances and pathologic evidence allow. In some situations, "undetermined," while unsatisfying, is at the same time the fairest and most reasonable ruling until additional information is obtained that allows a more definitive ruling.

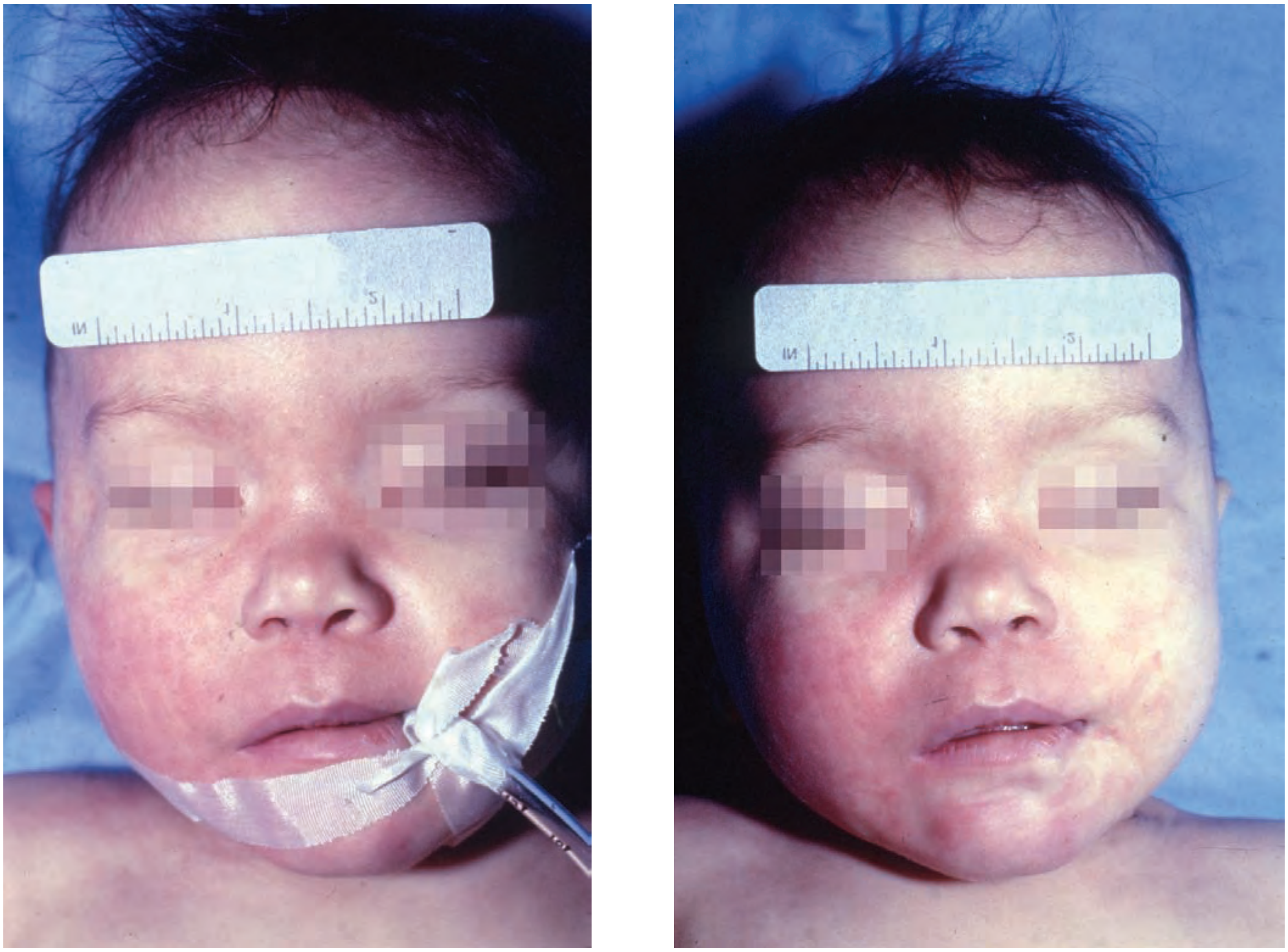

This child was brought to the emergency room and vigorously resuscitated. Note the marks left on the face as a result of endotracheal intubation and application of tape. It is important to document such findings and not confuse them with possible signs of child abuse, such as faint abrasions or contusions due to smothering. It is very important for clinicians to leave all tubes in place for postmortem (autopsy) confirmation of their placement.

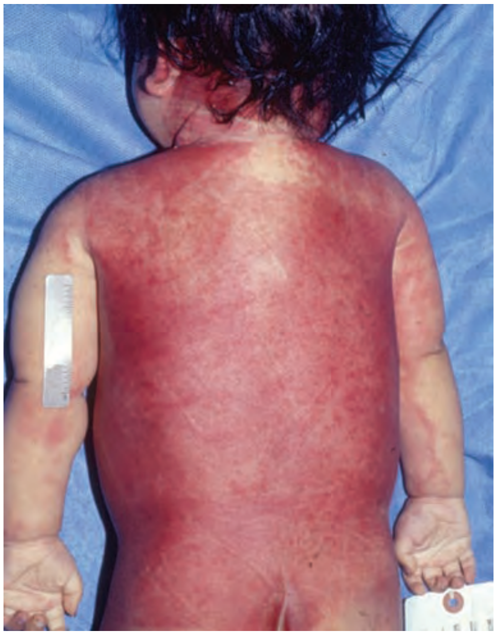

Posterior lividity, which may be mistaken for injuries such as contusions related to child abuse.

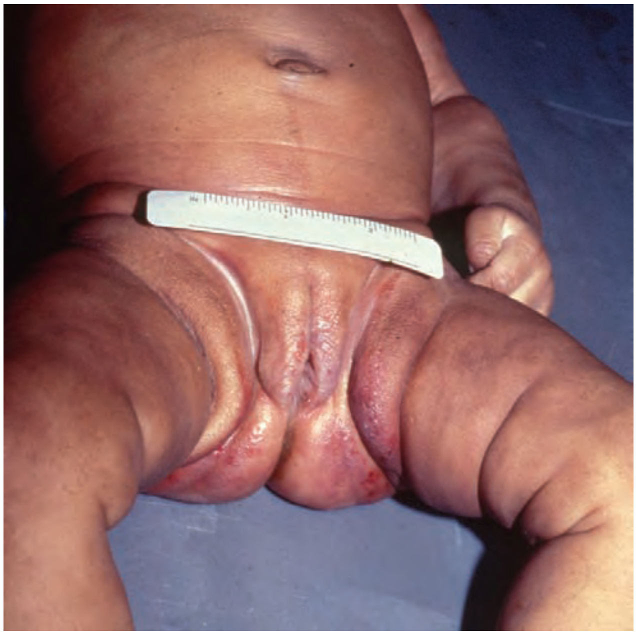

Diaper rash. Note the red cutaneous eruptions. These indicate some degree of neglect.

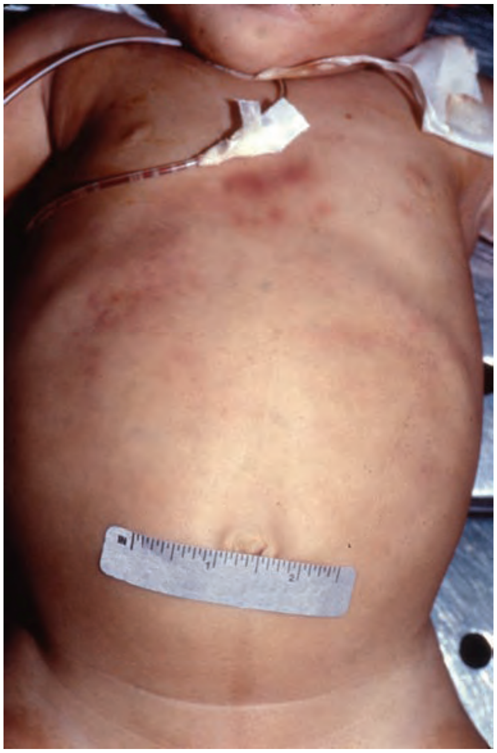

Child abuse case with resuscitative efforts and perimortem sternal contusions due to chest compressions. Note other small contusions at the lateral aspect of the upper abdomen and lower chest. These were inflicted while the child was grabbed and flung about violently during the assault.

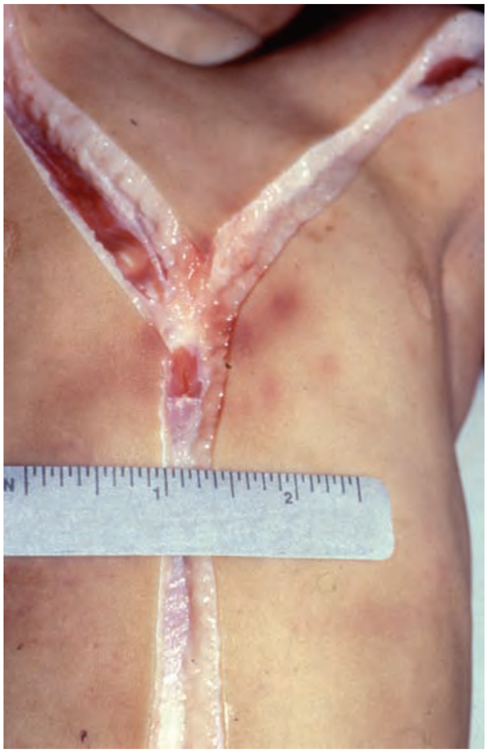

The same case showing perimortem contusions to the middle aspect of the chest due to resuscitation. Note the small amounts of subcutaneous hemorrhage overlying the sternum. Note the other segment of the "Y"-shaped incision is devoid of hemorrhage.

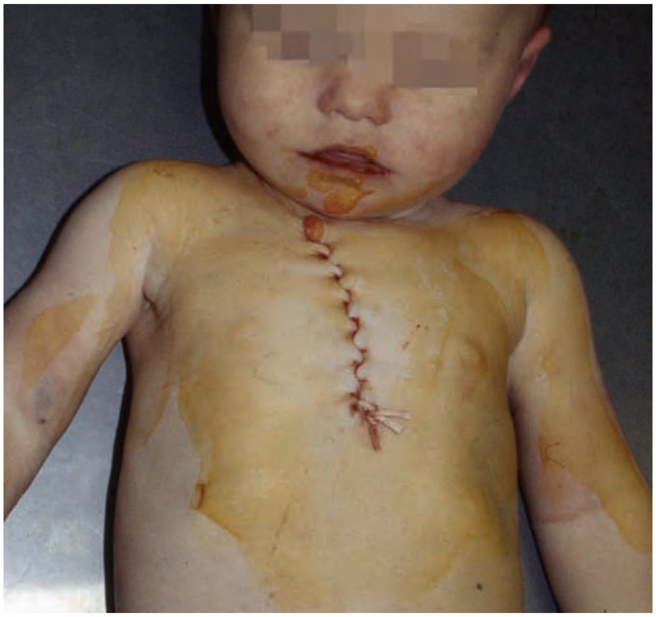

This case involved an infant who was found unresponsive in her crib. The case was referred to the medical examiner's office for autopsy. However, as the external examination was about to begin, this sutured midline incision was encountered. Further inquires revealed that the emergency department physician had authorized postmortem procurement of the infant's heart for heart valves. When this was discovered, the organ procurement agency was contacted, and released the heart to the medical examiner's office. The heart was found to have an anomalous origin of the left coronary artery from the right sinus of Valsalva, a rare yet documented cause of sudden death in both adults and infants. The issue of organ procurement remains a controversial and at times problematic one for the forensic pathologist. Consistently reliable methods for evaluating the coronary ostia have not been established in many tissue procurement agencies for the retrieval of hearts.

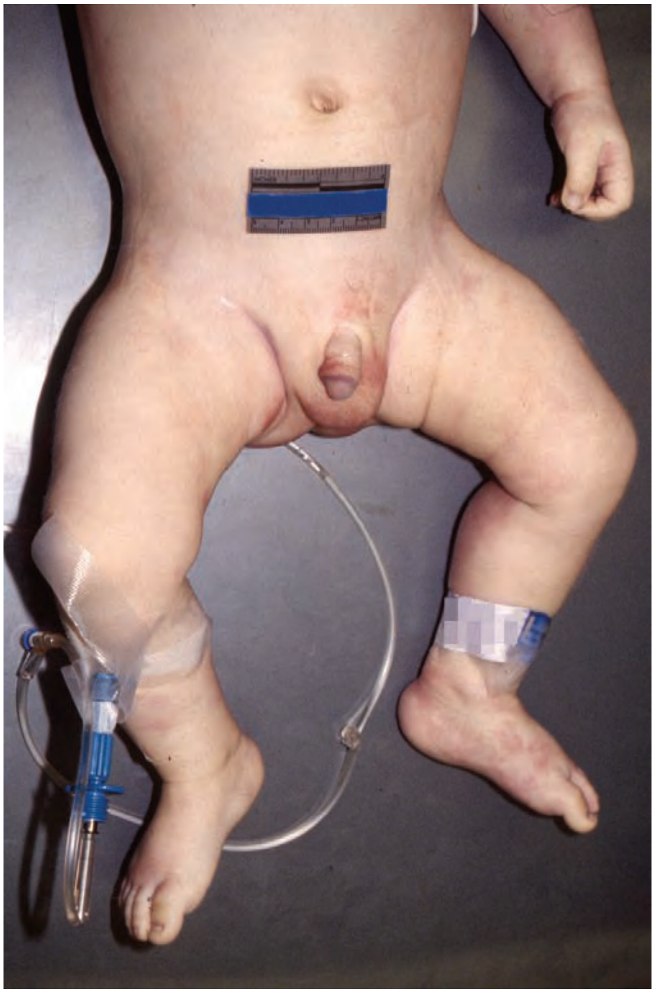

View of lower trunk, external genitalia, and lower extremities depicting a right-sided intraosseous catheter. This is a common mode of resuscitation in infants with extremely small, often collapsed blood vessels.

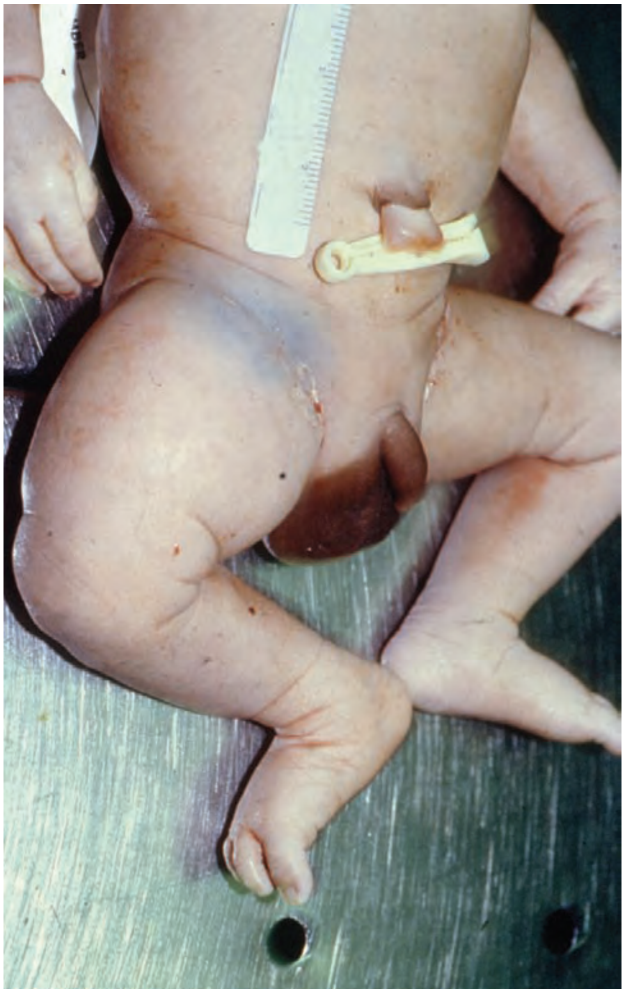

Hemorrhage with swelling of the thigh due to a misplaced femoral line.

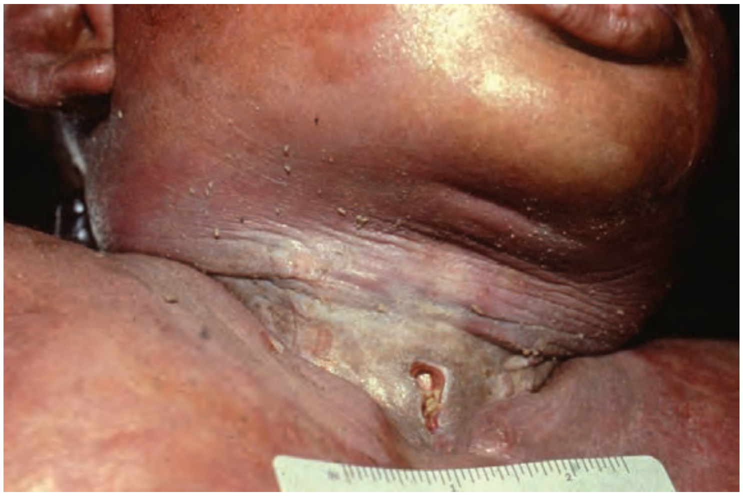

Healing infected tracheotomy incision.

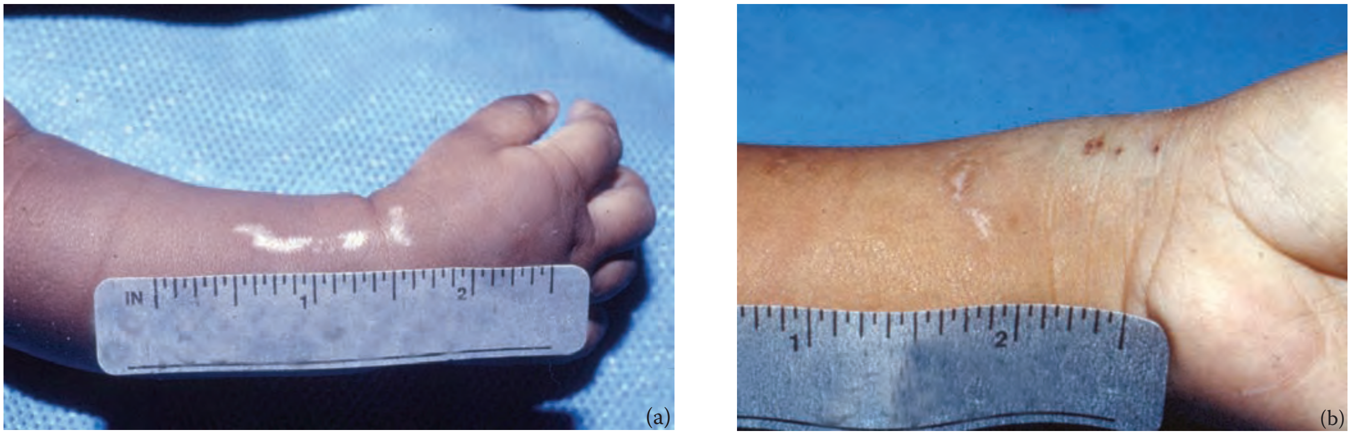

Scars to each wrist due to past therapy. Note image (b) has fresh needle marks as well. Therapeutic procedures should not be mistaken for injuries.

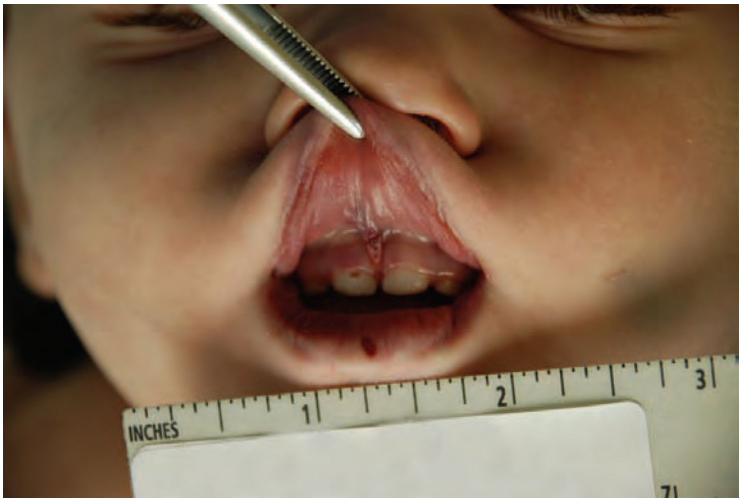

Small tear of the upper frenulum due to vigorous resuscitative efforts by an untrained person. Note the small bruise to the left side of the face, which had been caused by a fall several days earlier. The first phase of resuscitation was captured on a department store video camera when this child became lifeless following a seizure.

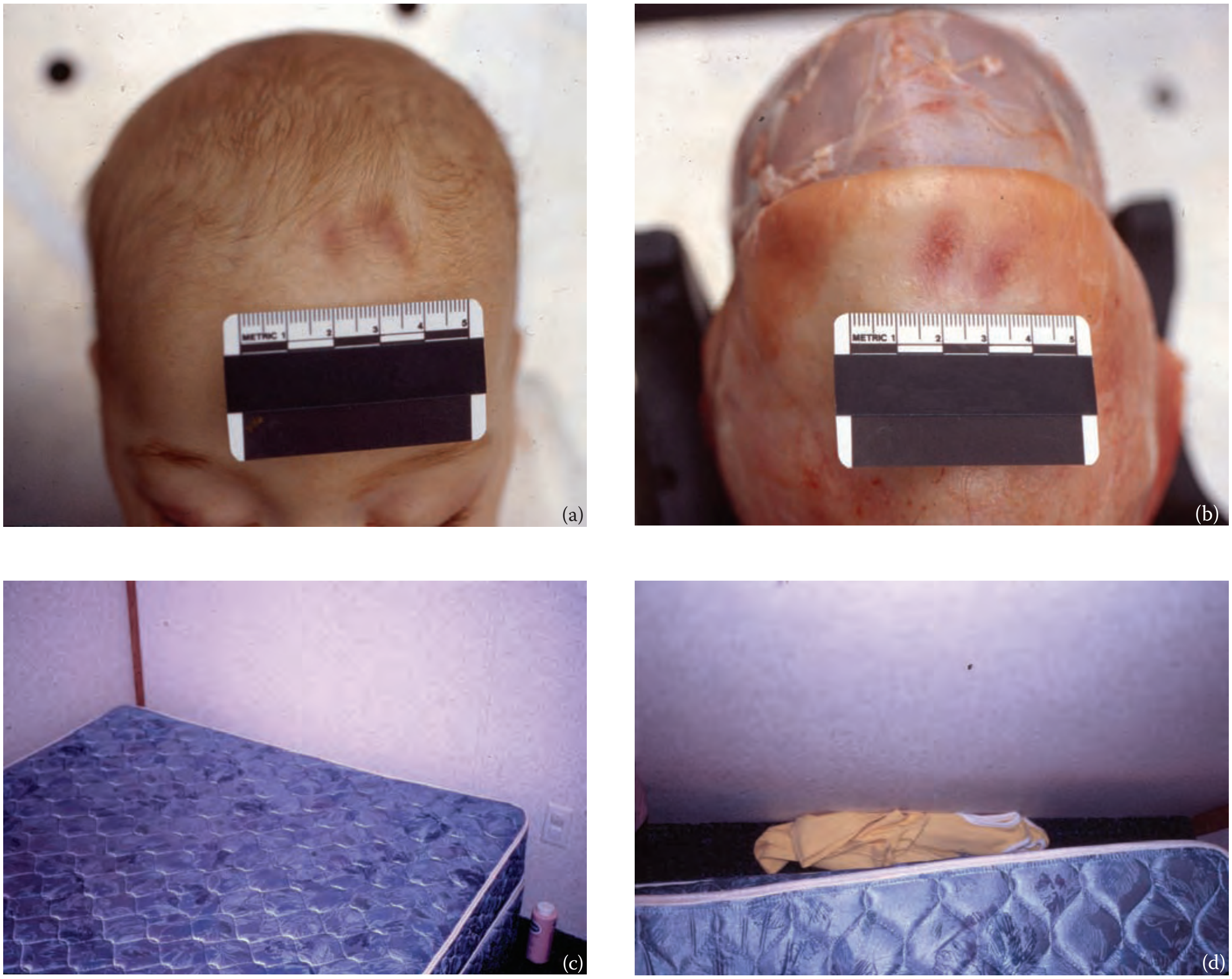

This case involved an infant who suffered an asphyxial death. He was found with his head compressed (wedged) between an adult bed and a wall. He was pulseless, apneic, and asystolic, and despite resuscitative efforts could not be revived; (b) depicts the frontal and temporal regions of the scalp with two ill-defined contusions within the frontal subgaleal region; (c) shows the relationship between the mattress and the wall. The mattress was separated from the wall by at least several inches, allowing the infant to fall into this space and become compressed; (d) is a view looking toward the floor, where the infant was found lying wedged between the wall and the mattress.

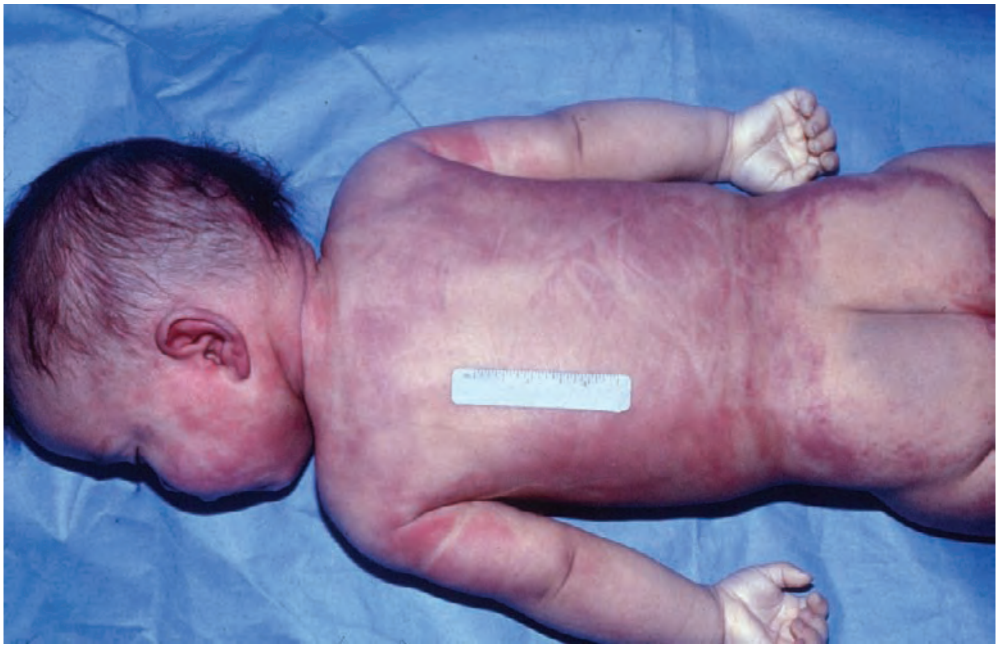

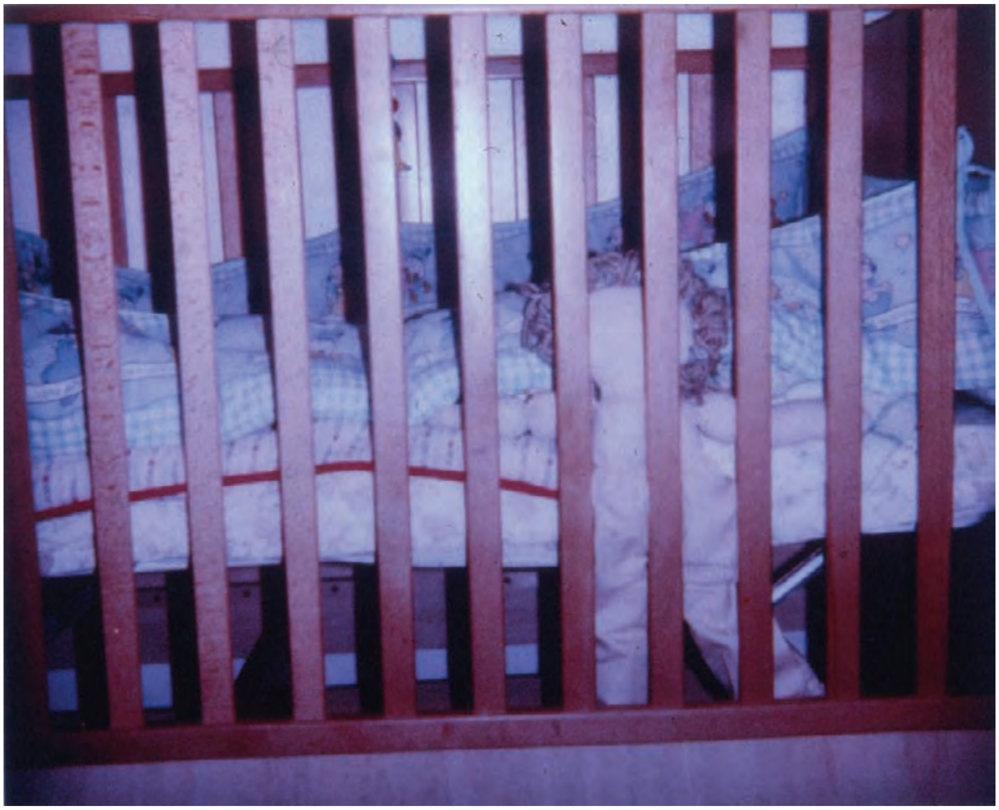

This child was brought to the Emergency Department (ED) by parents who spoke a foreign language. The child was reported to be found dead in a crib. The ED physicians believed that this death could be a result of SIDS but there was no initial scene investigation. Note the posterior patchy red lividity. There is also a faint Mongolian-type birthmark within the left gluteal region. Further scene investigation by the Medical Examiner's office and police department revealed that the child had actually slid between the crib railing and mattress, suffering an asphyxial death. There were no marks on the child's body at autopsy. The mattress was from another crib and was slightly smaller, allowing the child to slide between the railing and mattress. Note the doll demonstrating how the child was found at the scene.

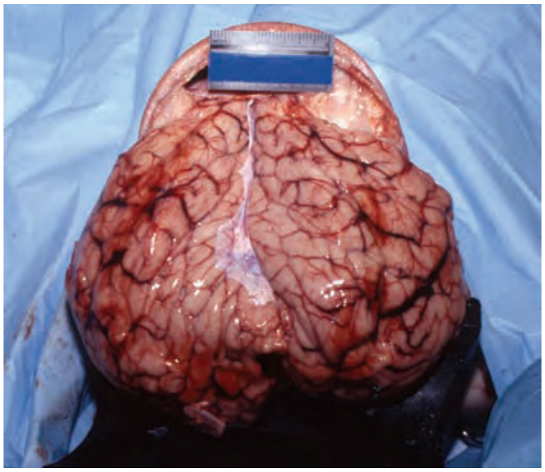

This image demonstrates brain swelling with flattening of gyri and obliteration of sulci due to hypoxic-ischemic brain injury. This 7-month-old female infant was found unresponsive in a car seat that fell off of an adult bed, with her chin tightly opposed to her chest, satisfying the criteria for positional asphyxia. She survived for several days and was ultimately declared brain dead. No other injuries were identified upon autopsy examination.

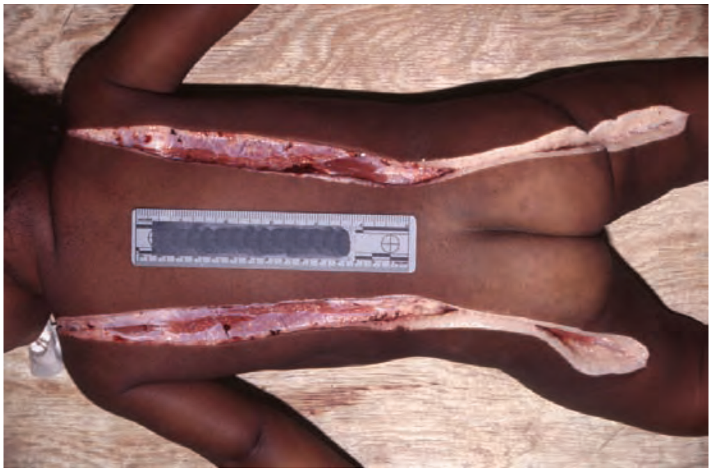

This 9-month-old child with a history of asthma was found unresponsive in her crib. Autopsy demonstrated multiple mucous plugs within the medium-sized airways, along with well-defined pathologic features of chronic asthma. No injuries were found. Death was certified as complications of bronchial asthma. Although unusual in a child this young, asthma can indeed be fatal. These bilateral vertical incisions made at the time of autopsy failed to demonstrate contusions.

This was the body of a completely healthy female infant. Her mother, who had left her unattended for an unknown period of time with her 18-month-old brother, found her submerged in a bathtub. The mother reportedly left them both in the tub without water and claimed the brother had to have turned on the faucets. When police attempted to reconstruct the scene, the older brother was unable to turn the faucets. Although the cause of death in this case was certified as drowning and the manner homicide, the mother was released on probation.

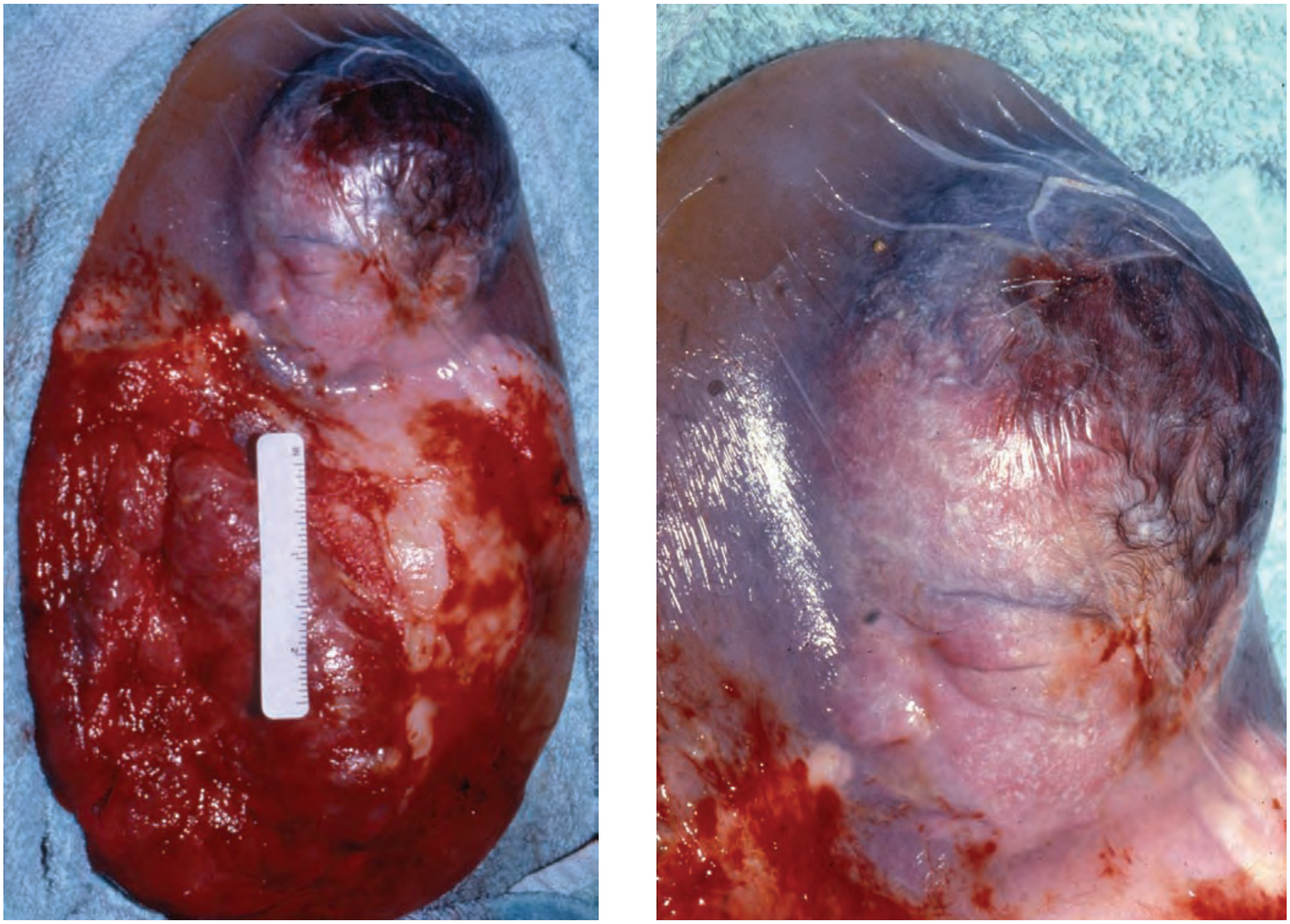

This normally developed neonate was delivered with the placenta enclosed in an intact amniotic sac. The baby died of asphyxia and the manner of death was certified as natural. The child was delivered at home and the young mother did not have medical knowledge to know that the membranes had to be immediately opened following delivery. She noted the child to be moving within this placental unit.

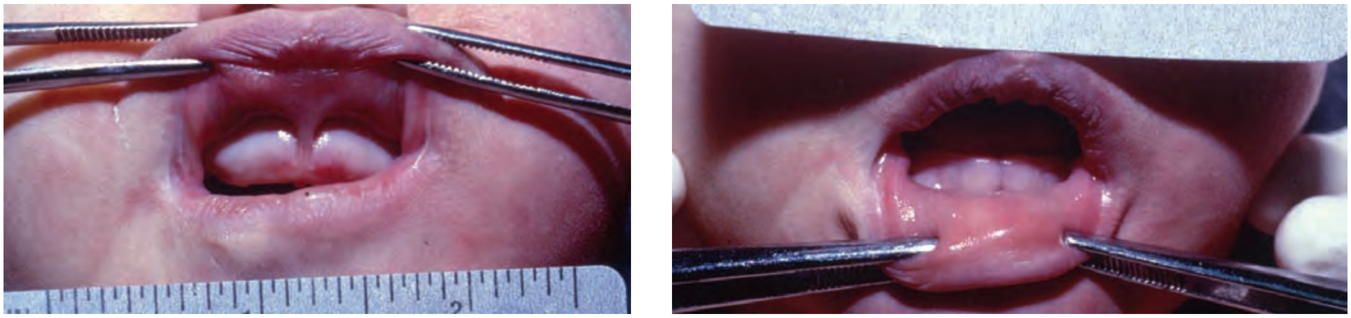

It is important to examine the frenulum and mouth region internally and externally in infants to rule out possible findings associated with smothering. Both of these cases show no injuries.

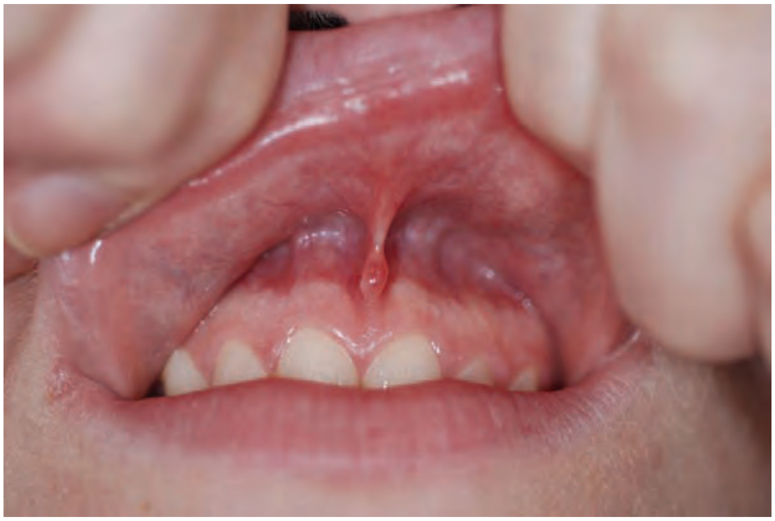

Healed frenulum laceration. This individual fell as a small child and accidentally struck her face on a kitchen cabinet.

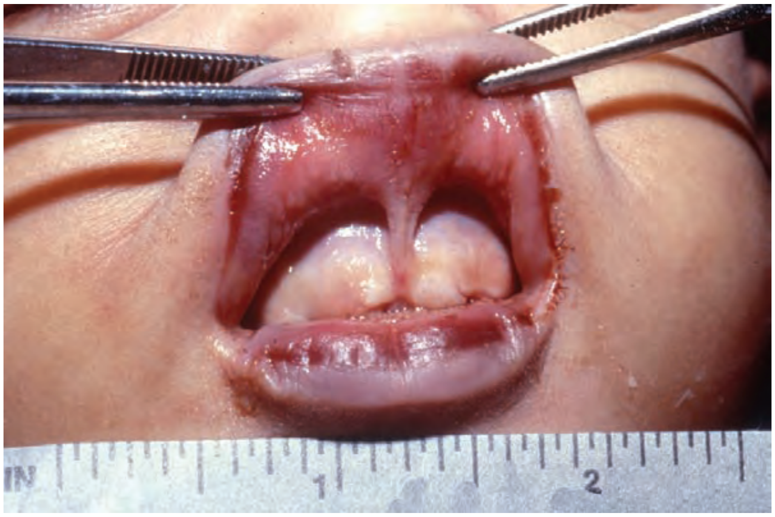

Intact frenulum with postmortem drying of the lips. This does not indicate child abuse.

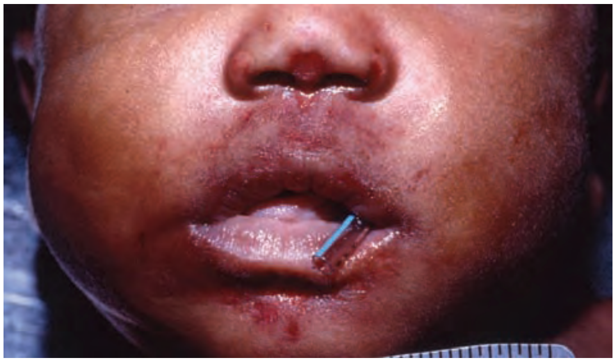

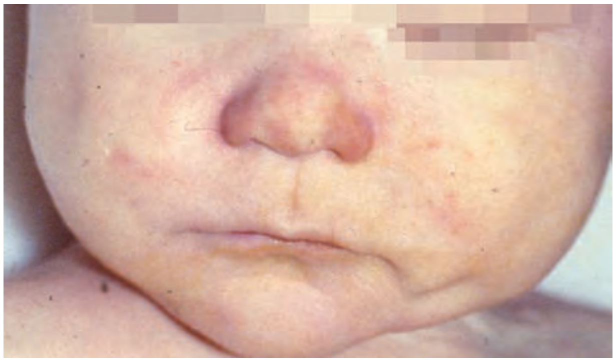

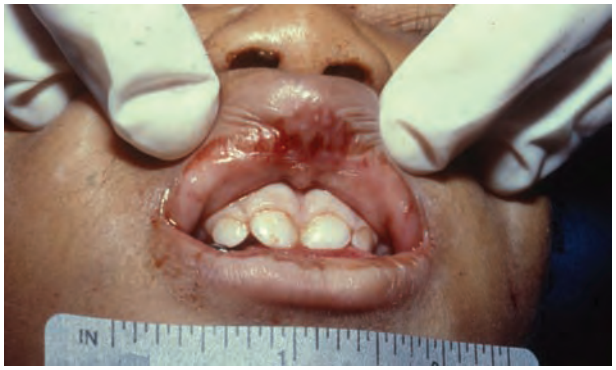

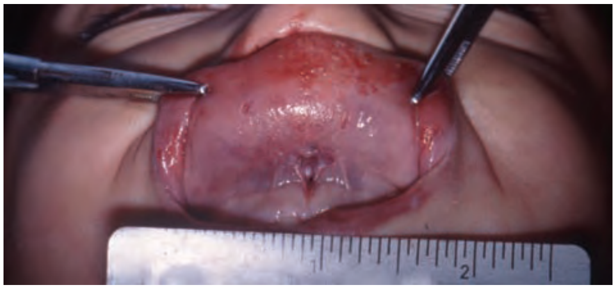

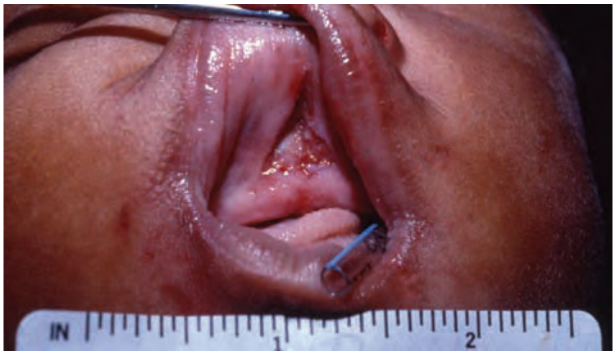

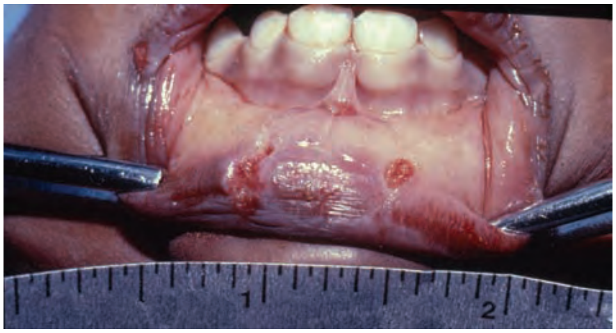

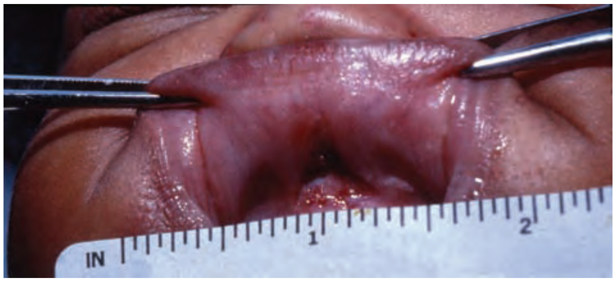

These children have varying degrees of abrasions to their face with trauma to the lips and torn frenula. All are cases of homicidal smothering. It is important to also realize that a homicidal smothering may occur without leaving marks of any kind.

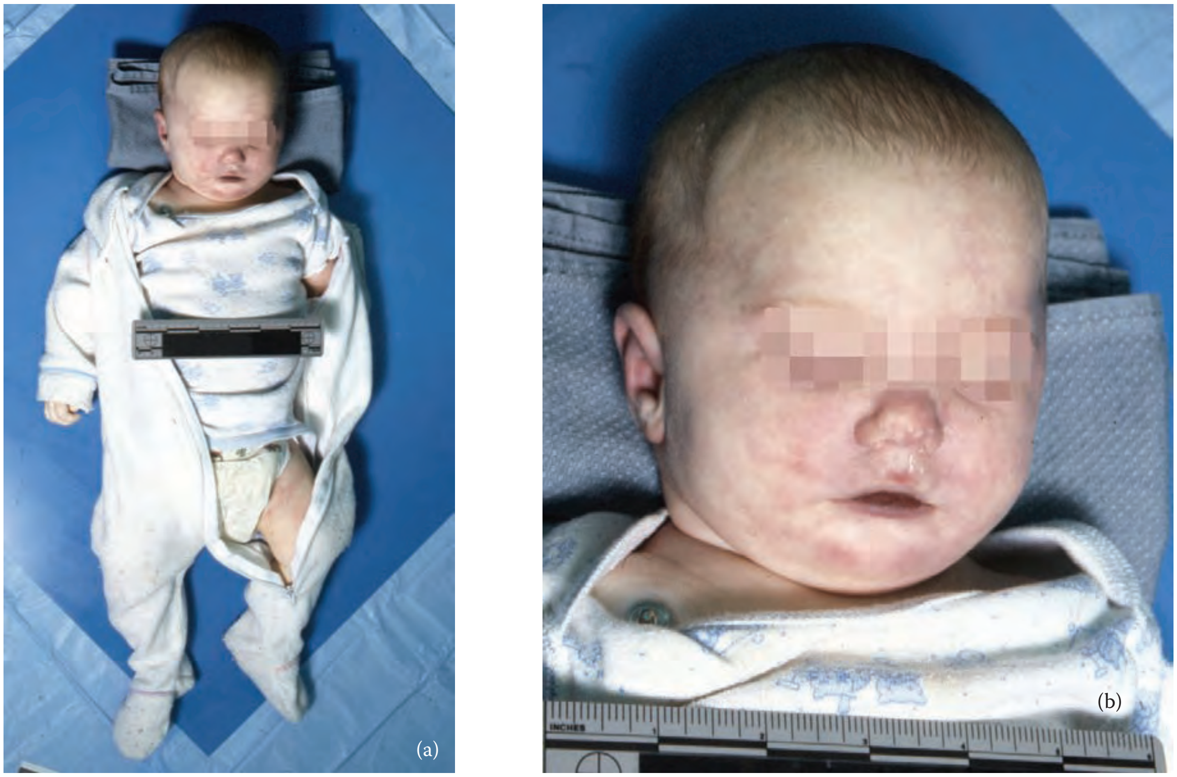

These images demonstrate a case involving a several-month-old infant who was found unresponsive in her crib. She had no known previous illnesses. If possible, it is always important to photograph infants in their original clothing, which allows for more accurate reconstruction of the terminal events (i.e., stains of blood-tinged fluid.) View (a) demonstrates clear fluid emanating from the left nostril with a faintly frothy quality, probably indicative of pulmonary edema, which is a common finding in victims of SIDS, particularly when attempts are made to resuscitate. The elastic nature of infants' lungs and the vigorous resuscitation efforts collectively contribute to the formation of this edema fluid, which is often blood-tinged due to rupture of small capillaries. Oftentimes, the blood-tinged fluid is misinterpreted by police personnel as being suspicious of foul play. It is important for pathologists and emergency medical services (EMS) personnel to educate people with nonmedical backgrounds regarding the distinction between blood-tinged edema fluid and actual clotted blood.

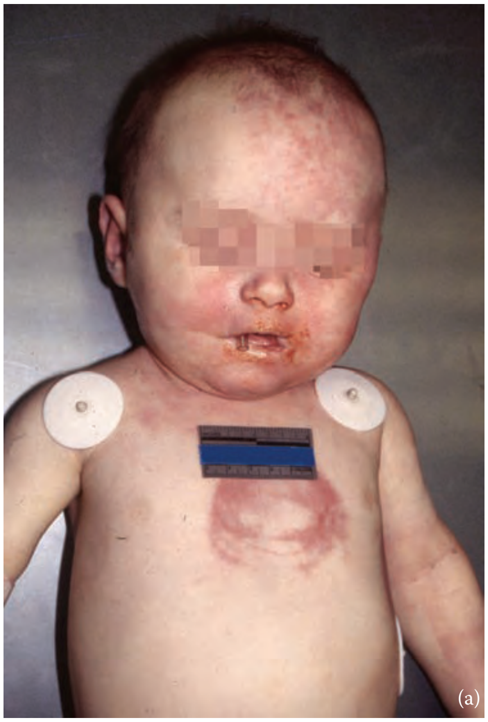

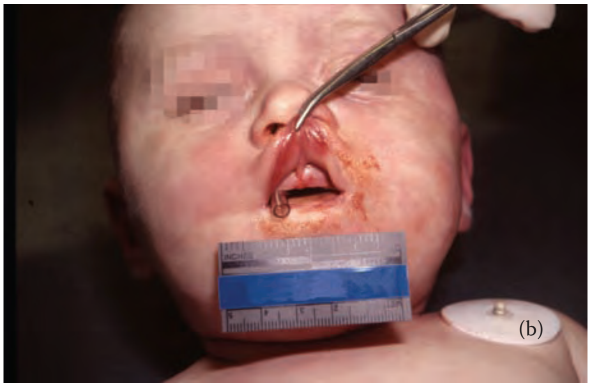

This case involved sudden death in an otherwise healthy infant. Notice the confluent burns and abrasions on his left anterior chest wall, extending to the midline (a). These marks are indicative of defibrillation attempts from resuscitation. Additional therapeutic interventions can be seen, including an endotracheal tube, along with electrocardiogram (ECG) electro pads. Notice also the blotchy red discoloration of the forehead and face (left side slightly greater than right). While one might be tempted to conclude that this is livor mortis, it is important to realize that vigorous resuscitative efforts may also alter the patterns of vascular congestion; in the same case, (b) demonstrates an intact upper frenulum with dried blood-tinged secretion around the left side of the mouth and no trauma.

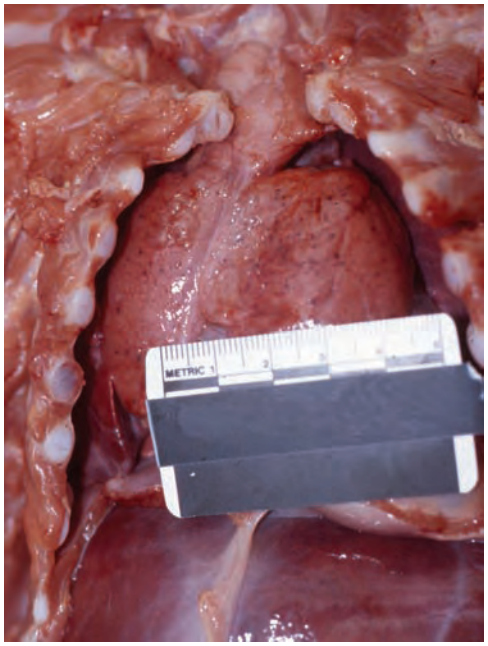

This infant was found unresponsive in his crib and could not be resuscitated. Maternal history was significant only for group B streptococcal infection. This picture depicts prominent petechiae on the anterior capsular surface of the thymus. Petechiae involving the thymic, visceral pleural, and epicardial surfaces are extremely common findings in sudden infant deaths, including deaths that have been ascribed to SIDS (up to ~85 % of SIDS fatalities). The precise mechanism underlying the formation of petechiae remains elusive, but is believed to be related to the negative pressure created by terminal gasping in infants. However, it is well known that these petechiae are not in any way specific or pathognomonic for SIDS and may be found in deaths due to many other causes, including definable natural disease conditions, accidental injuries, and inflicted injuries (homicides). In this case, the petechiae, which appear as pinpoint hemorrhages, are particularly striking.

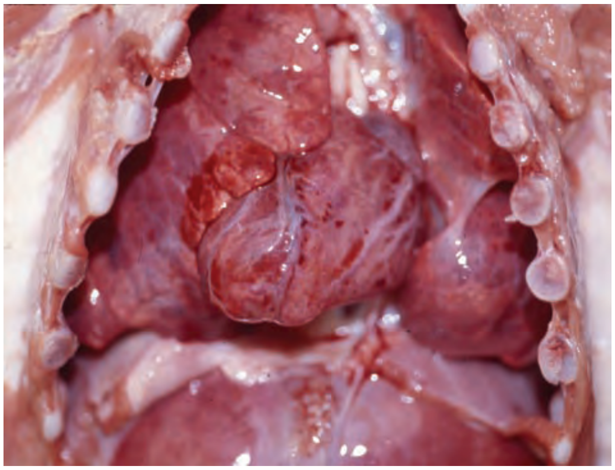

This is a view of the thoracic contents of an infant who was found unresponsive while sharing a bed with his mother while in a supine position. She called 911 and paramedics arrived to find him apneic and pulseless. He could not be resuscitated. Pertinent history included prematurity, maternal group B streptococcal infection, and a recent upper respiratory infection. The only significant findings at autopsy were a moderately cellular mononuclear leptomeningeal exudate consistent with a viral meningoencephalitis. However, there was no significant brain swelling. This fatality occurred in the early to mid-1990s, and the death was ascribed to SIDS. If a similar case were encountered today, bedsharing in an adult bed would probably be listed as a risk factor in part two of the cause of death statement and the manner of death would be undetermined. This picture depicts prominent petechiae on the anterior epicardial surface of the heart.

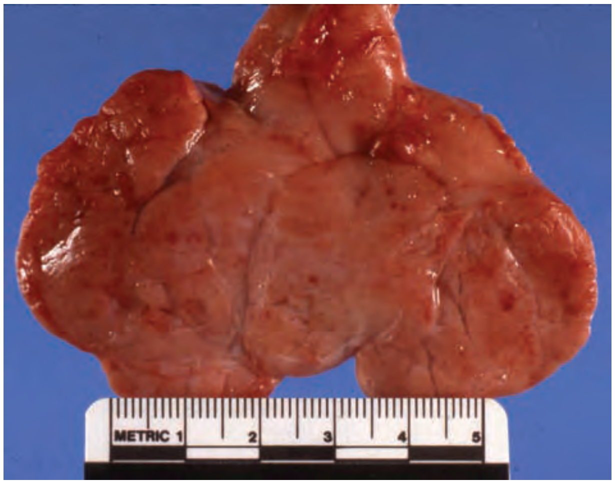

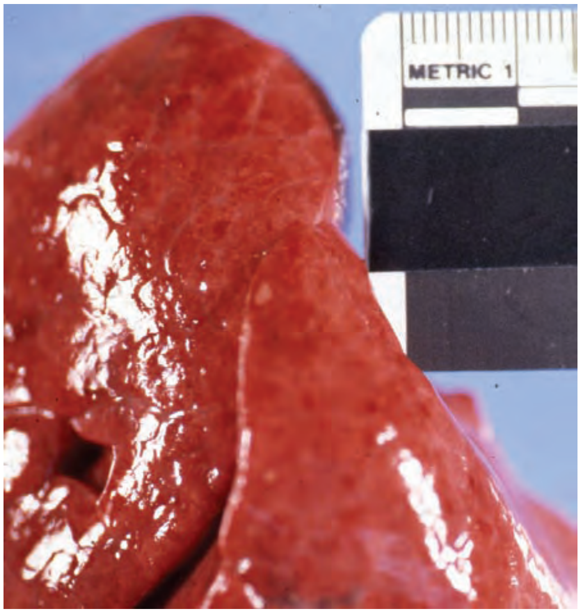

This case involved an infant who was born prematurely at 35 weeks gestation and was found prone in a bassinet with his face between a “covered” adult-sized pillow and the corner of the bassinet. The findings of this case were not sufficient to render a diagnosis of mechanical asphyxia, but asphyxia was of sufficient concern to denote the cause of death as sudden unexplained death in infancy and the manner undetermined. This picture shows sparse, inconspicuous visceral pleural and thymic petechiae, which were the only gross findings at autopsy.

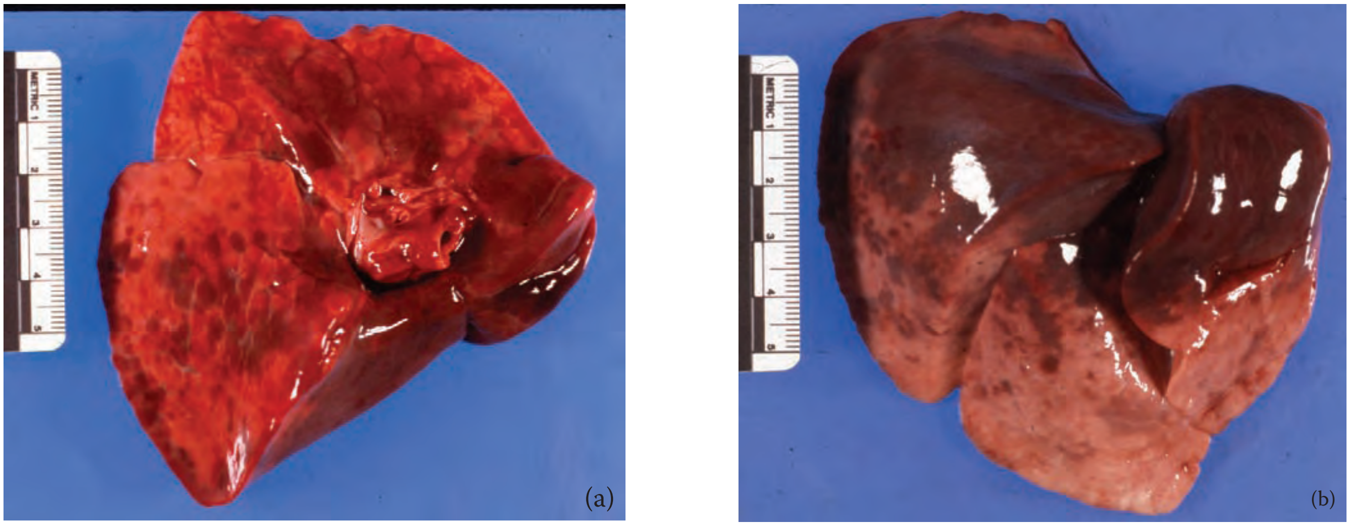

View of the mediastinal surface of the left lung (a) demonstrates confluent subpleural hemorrhages within the lower lobe, likely an artifact of vigorous cardiopulmonary resuscitation (CPR). Note the way the hemorrhages conform to anatomic boundaries, similar to what is observed in aspiration of blood; (b) demonstrates the costal surface of the lung, with blotchy subpleural hemorrhages, some slightly larger than petechiae. Note also the confluent congestion within the posterior aspects of the lung, most likely a result of postmortem hypostasis, or settling of blood due to gravity.

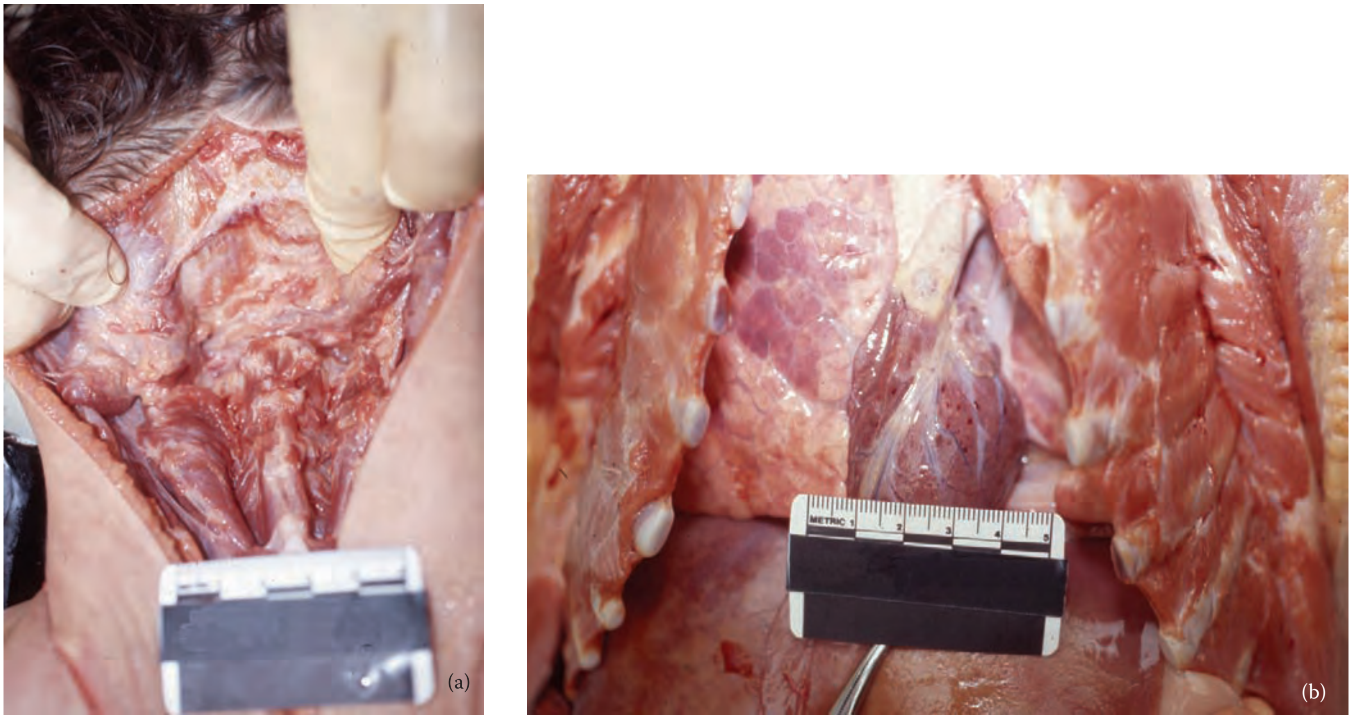

This was a case involving the enigmatic death of a 2-year-old child who was previously completely healthy. She had finished a course of amoxicillin for otitis media 3 days prior to her death. On the day before her death, she apparently fell off a "teeter-totter" and struck her head on the ground, which was a muddy, grassy surface, but she never lost consciousness and experienced no mental status changes. She could not be aroused from an afternoon nap and could not be resuscitated by pediatric life support protocol. The gross autopsy findings were unrevealing, including absence of injuries of scalp, skull, and brain. The only finding in addition to thymic, epicardial, and visceral pleural petechiae (b) was fibromuscular hyperplasia of the atrioventricular (AV) nodal artery, which was discovered by an expert cardiac pathologist who consulted on the case. This picture depicts a posterior neck dissection (a), demonstrating soft tissues and skeletal muscles that are entirely free of injury. In this case, anterior neck dissection was negative for injuries as well; (b) depicts prominent epicardial petechiae on the anterior surface of the heart. This case illustrates that the findings of intrathoracic petechiae are not confined to infants but can also be seen in young children.

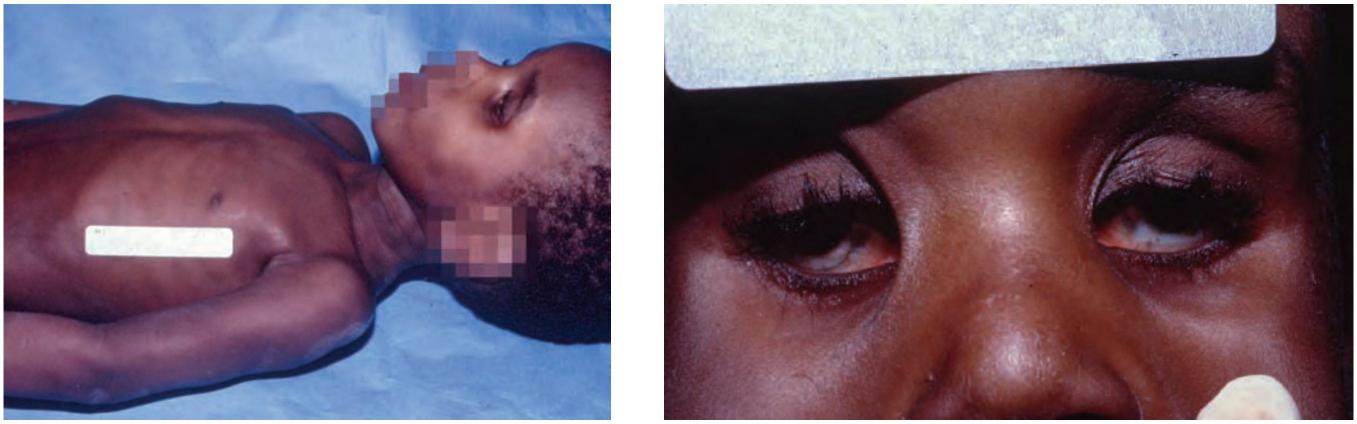

This child died of complications of dehydration following a prolonged viral gastroenteritis. Note the sunken eyes.

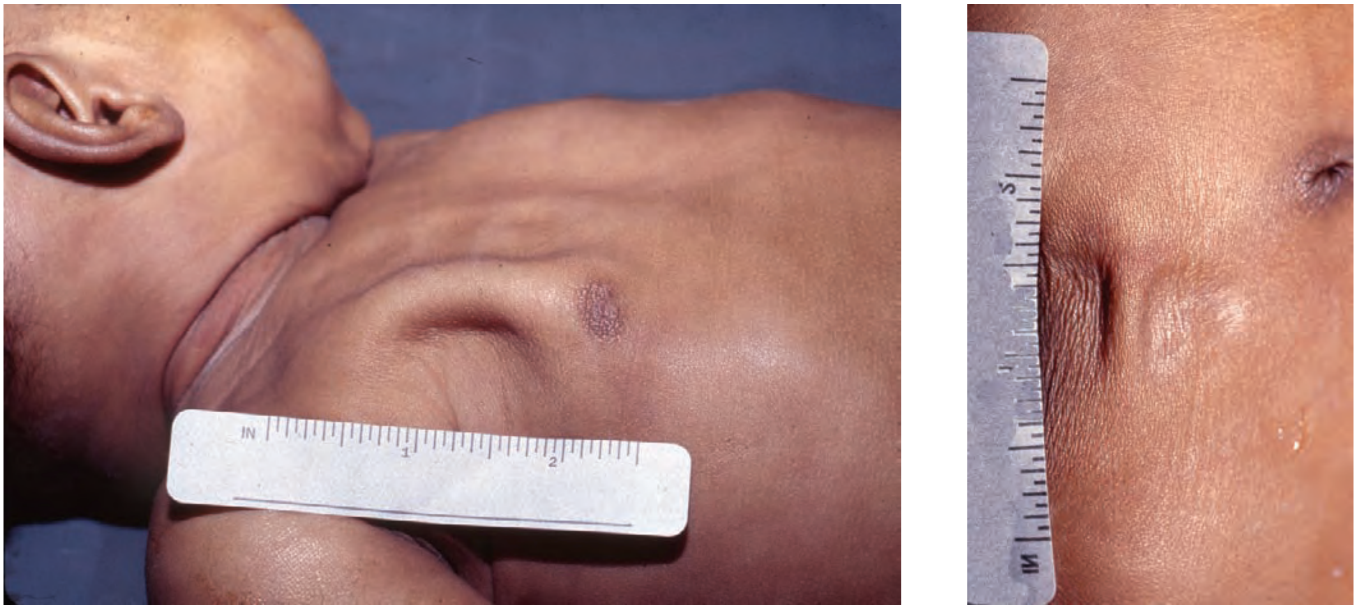

Poor skin elasticity (turgor) is demonstrated in these infants who were refrigerated in the morgue overnight. These infants were shown not to be dehydrated by vitreous analyte determination and antemortem hospital chemistry testing. The poor skin elasticity in these cases is due to the congealing of fat as a result of postmortem refrigeration.

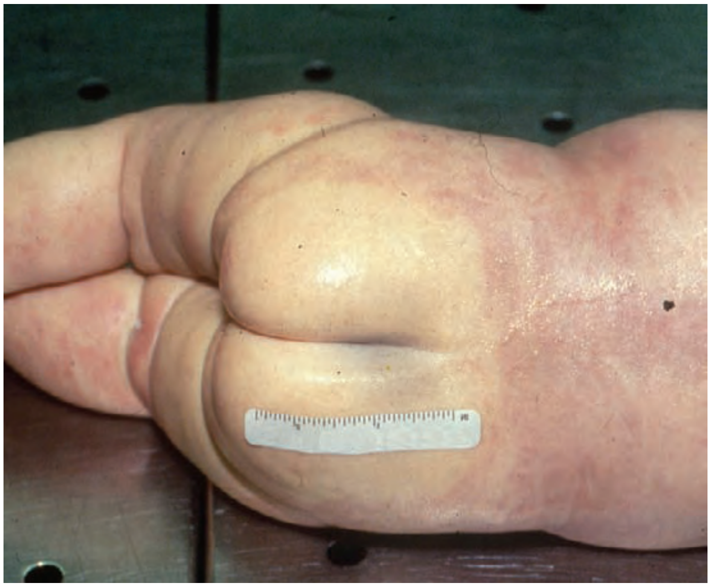

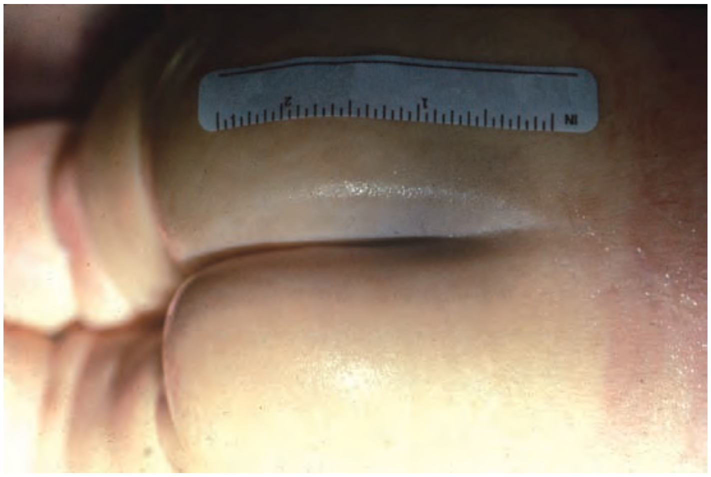

Mongolian spot. This is a birthmark and is sometimes mistaken for a contusion. During the time of autopsy it is easy to differentiate between the two by making an incision and documenting the presence or absence of underlying hemorrhage. The author has had at least one case where a contusion was demonstrated subjacent to a Mongolian spot.

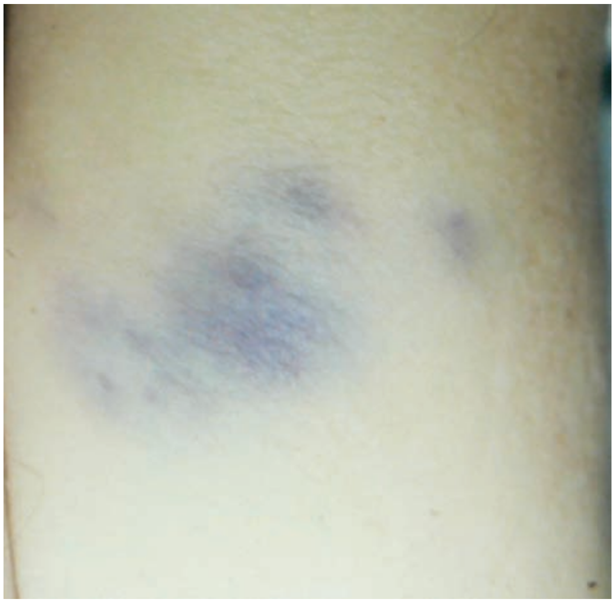

Recent contusion.

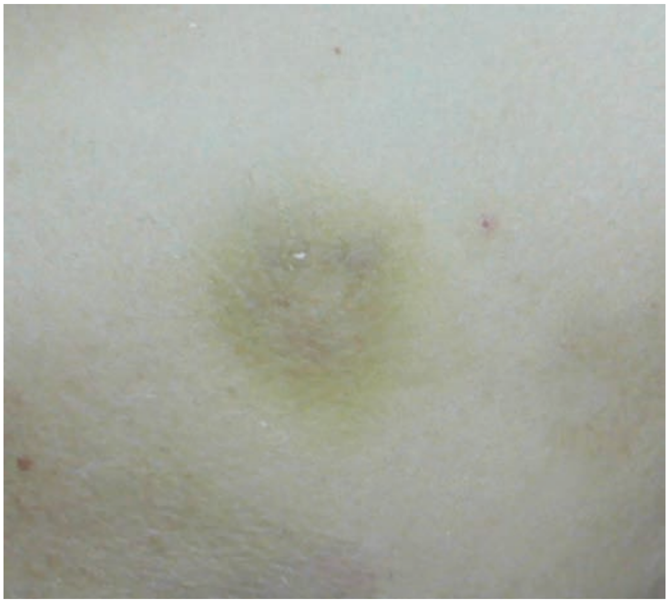

Old contusion.

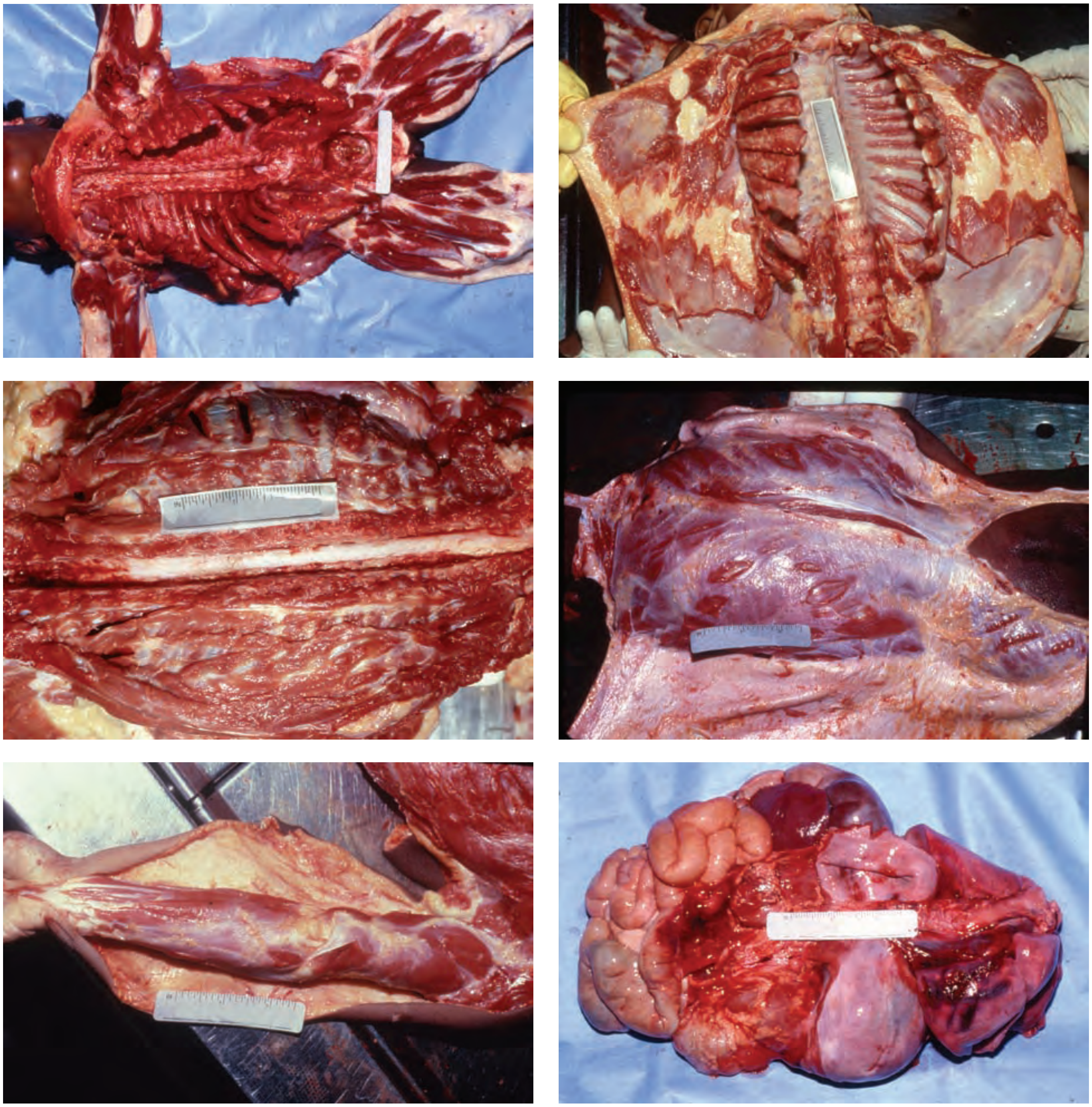

While performing autopsies on suspected child abuse cases it is important to be very thorough and it is recommended to incise the extremities and back to further document deep contusions, particularly in darker-skinned children. When examining the rib cage it is important to strip the pleural surface, separate each rib and evaluate each one separately. It is also good practice to do a posterior neck and back dissection while examining each rib as well.

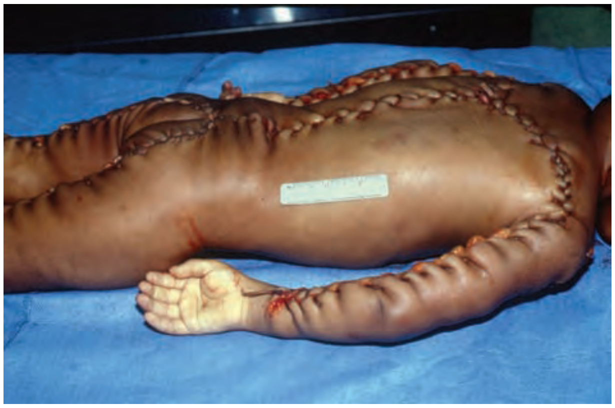

This figure demonstrates how the body is sutured prior to funeral home release.

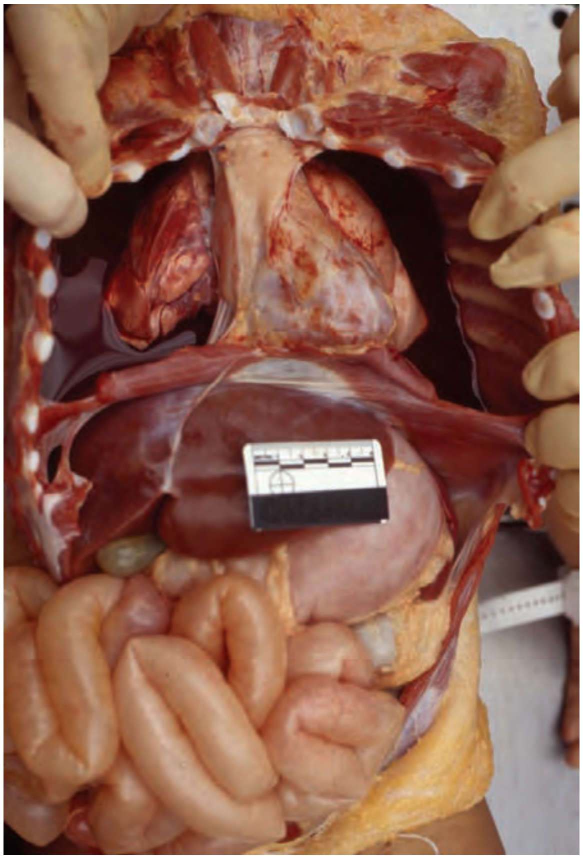

This 2-year-old child was brought to the ED by his mother's boyfriend after he was found unresponsive. The only observed external injury was a small faint contusion to the child's forehead. Autopsy revealed a fracture of the lumbar spine with a transected descending thoracic aorta and an old healed fractured clavicle. This picture demonstrates bilateral hemothoraces.

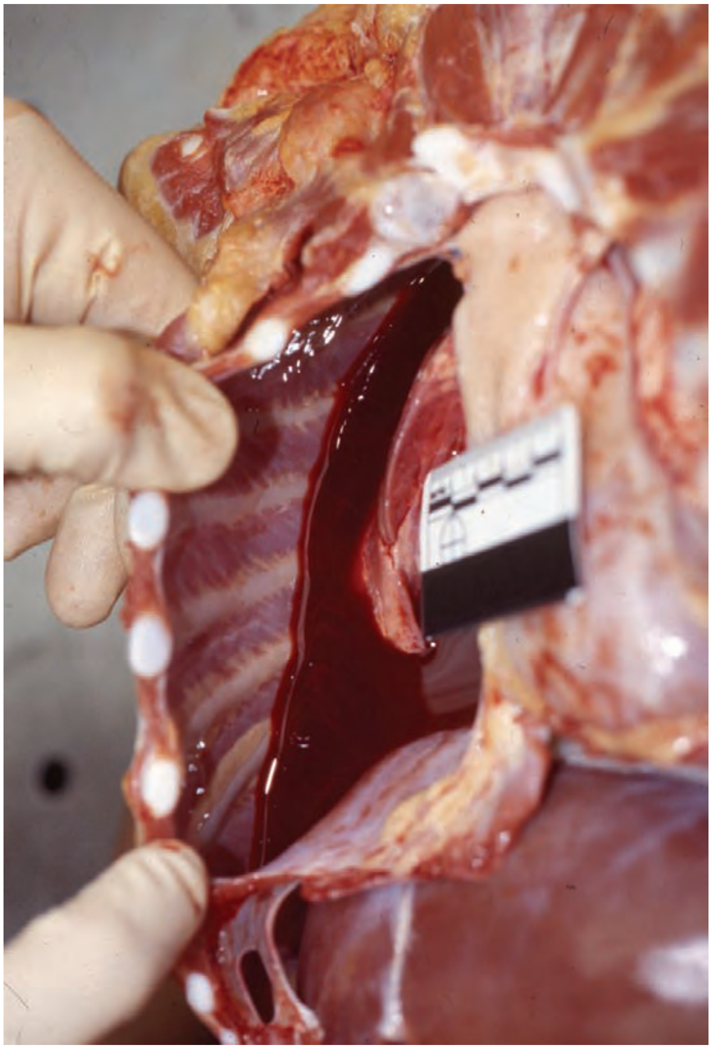

Right hemothorax due to aortic transection.

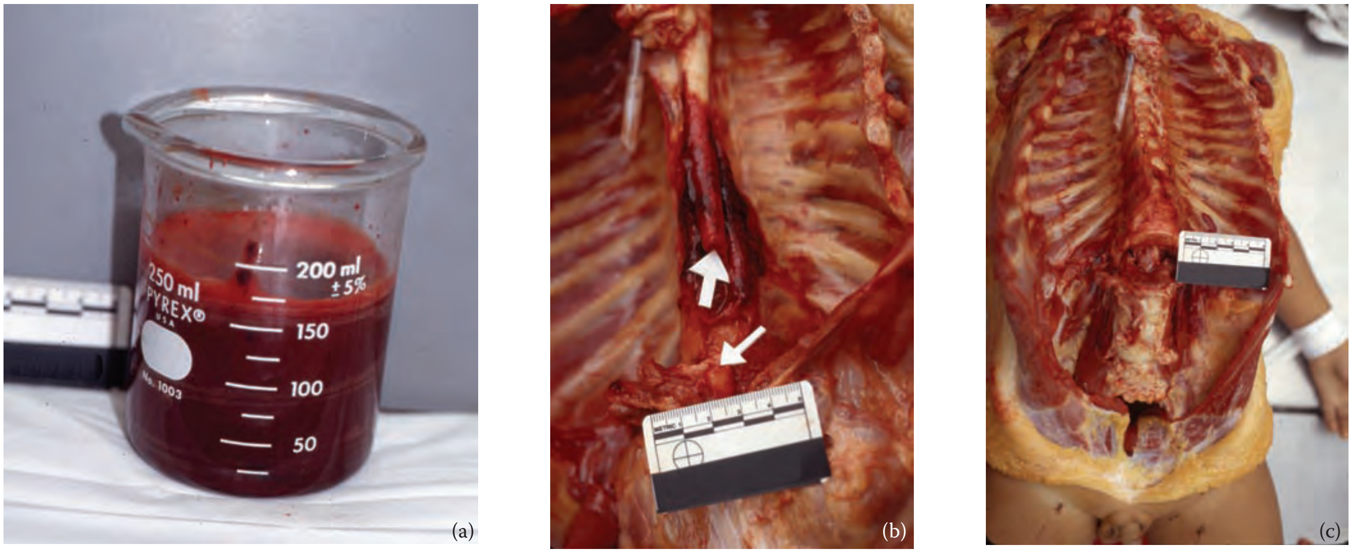

Image (a) demonstrates the volume of blood collected from the bilateral hemothoraces (approximately 165 mL in total). In a case such as this, the mechanisms of death are probably multifactorial, including not only blood loss but also a component of spinal cord or sympathetic nervous system dysfunction; (b) depicts transected free ends of lacerated descending thoracic aorta, marked with arrows; (c) is a panoramic view of lumbar fracture, which resembles a "hinge" at L1-L2.

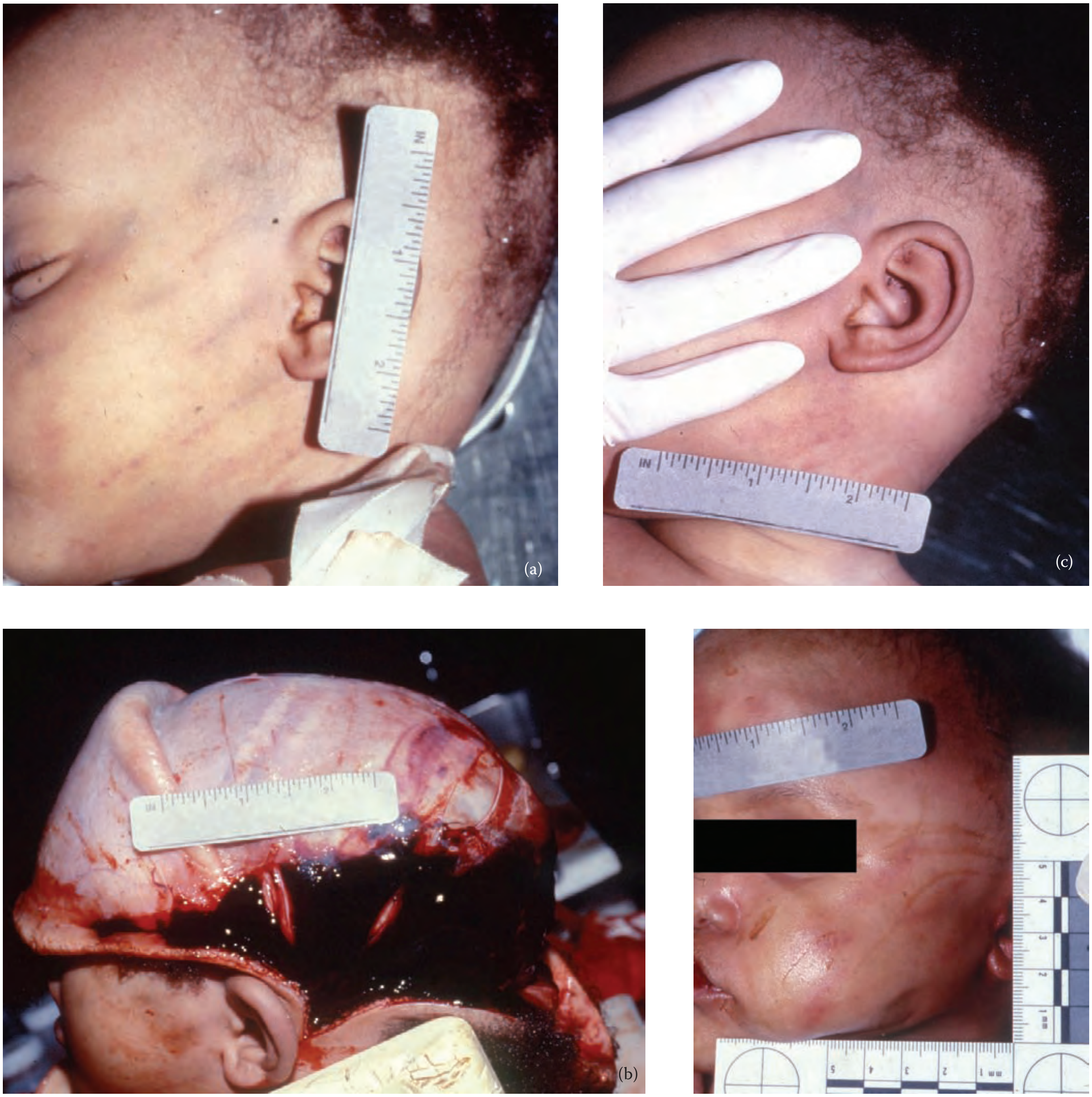

This child died from multiple blunt force injuries. (a) Note the hand pattern with parallel contusions across the child's face. The contusions are more obvious when depicted following the intermastoid incision and scalp reflection (b). Note the excessive dark red blood clot at the inferior portion of the side of the head. Also note that this is not as visible when viewed from the external skin surface. Deep contusions may be difficult to see when viewed only from the external surface of the body; (c) demonstrates the same child approximately 2 days after autopsy. Note that the characteristics of the facial bruises have changed slightly due to the extended postmortem interval.

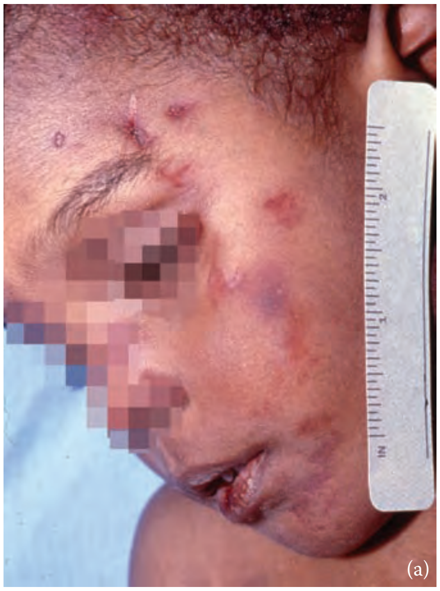

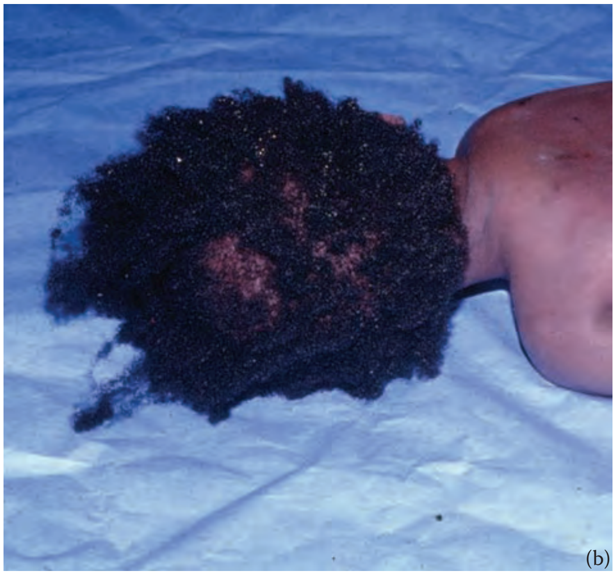

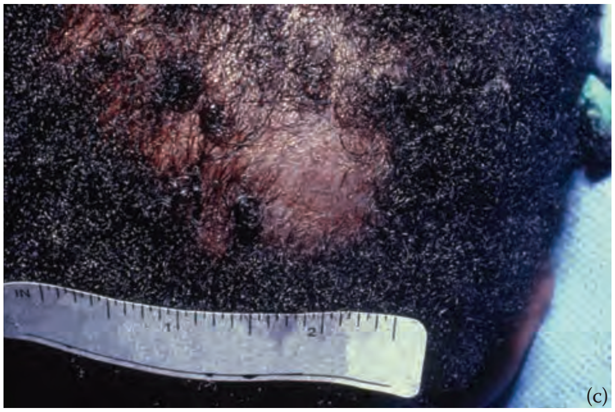

This child sustained multiple blunt force injuries to his head, back, and extremities (a–b). The child's hair was traumatically pulled from his scalp (c). Note the multiple angulated red to brown abraded contusions on the child's back (d), consistent with a belt buckle strike. Also note, following incision of the back and extremities during autopsy (e), the large degree of red bruising within the subcutaneous soft tissues of the back. Note the pale yellow subcutaneous tissue within the posterior compartments of the lower extremities, which are atraumatic and without hemorrhage.

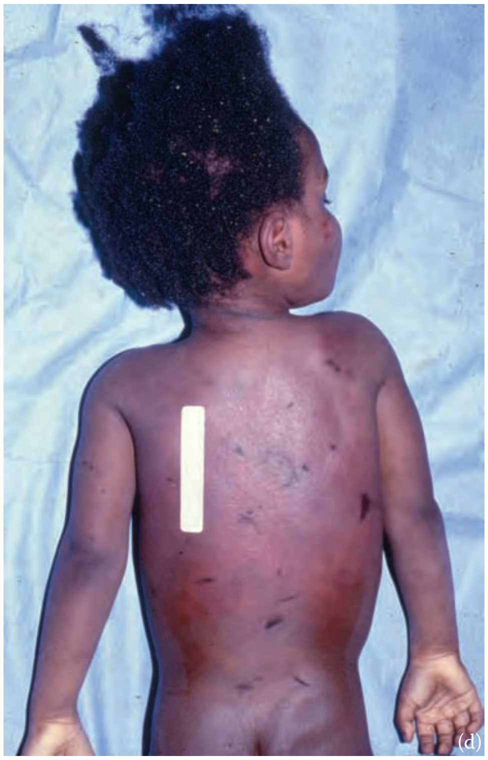

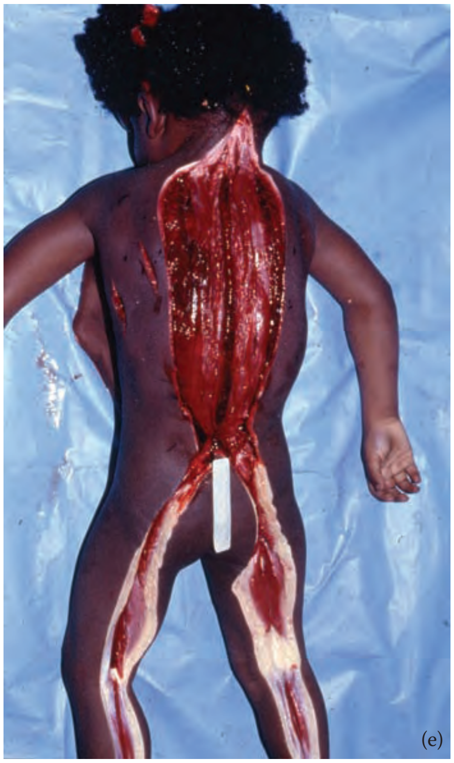

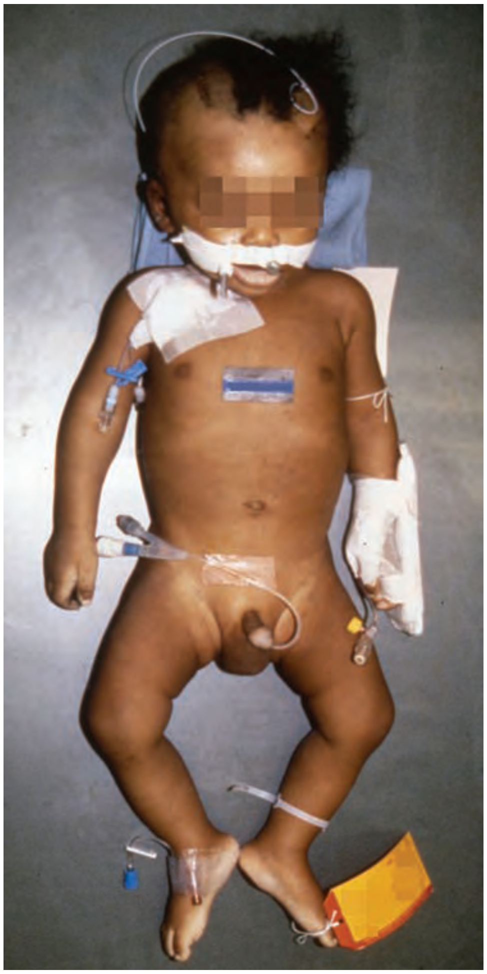

Note the multiple therapeutic interventions documented at autopsy. This documentation is important to help prevent misinterpretation of therapeutic intervention as injuries.

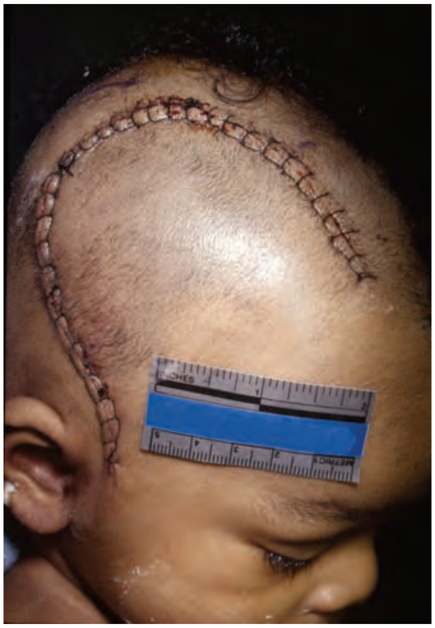

This case demonstrates a sutured right-sided scalp incision as part of a craniotomy for evacuation of a right-sided subdural hematoma.

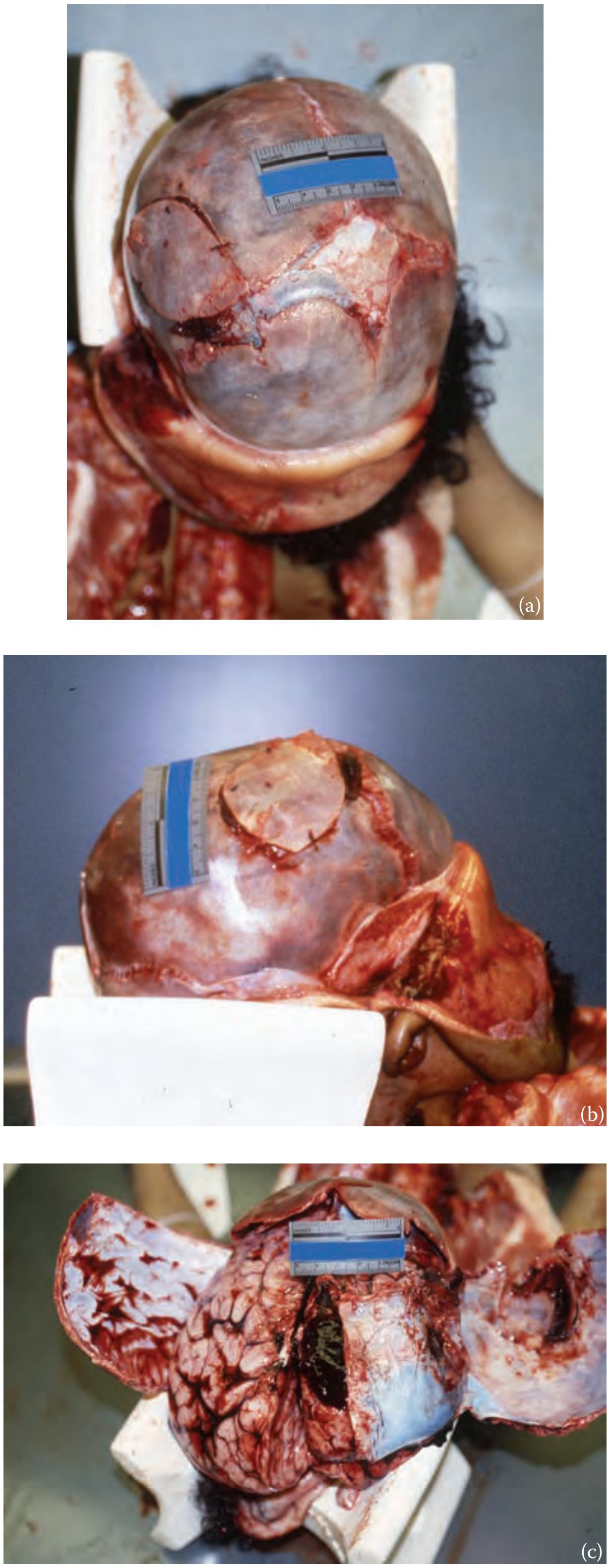

Superior view of the calvarium (a) illustrating marked widening of the coronal suture (diastasis) due to brain swelling, along with a right-sided craniotomy window; (b) right side of face and skull, including craniotomy window. Marked diastasis of the coronal suture is also seen; (c) residual right-sided subdural hemorrhage and marked brain swelling, characterized by complete flattening of gyri and effacement of sulci on the cerebral convexity of the left cerebral hemisphere. The inner surface of the left vertex dural leaf is blood-stained.

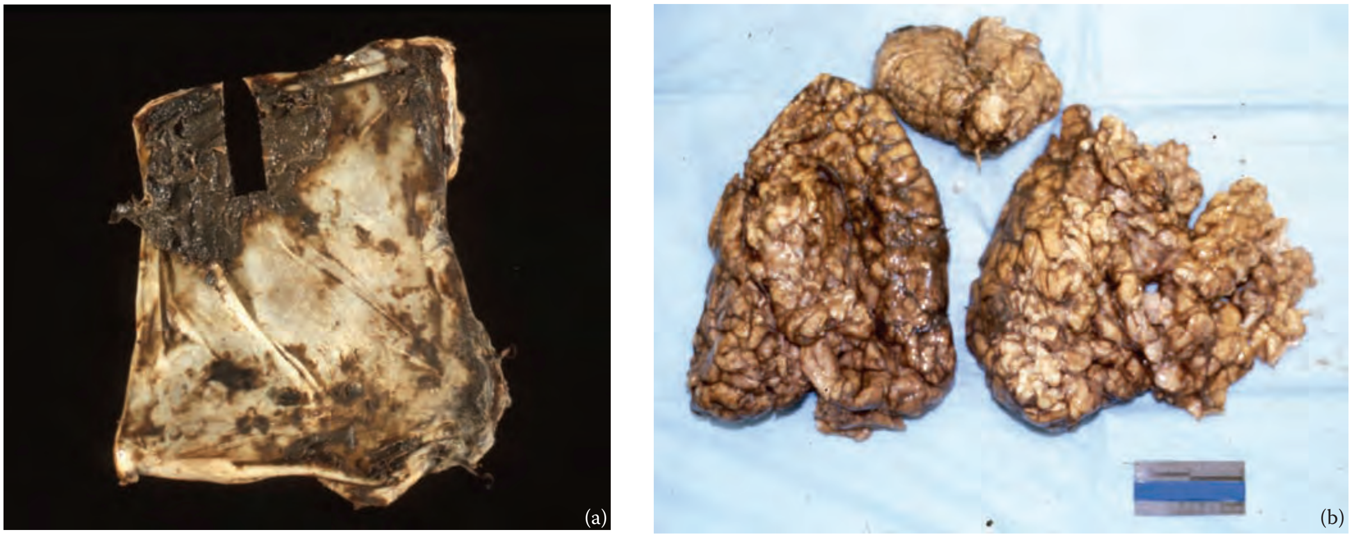

Formaldehyde-fixed portion of dura mater from an approximately 4-month-old infant with a history of tetralogy of Fallot who was found unresponsive by his mother's boyfriend. The boyfriend claimed that he had fallen approximately 3 feet from a couch onto a carpeted floor. Such explanations are often incompatible with the spectrum of injuries observed at autopsy. In this case, the adherent, thin-layered, clotted subdural blood indicates a duration of approximately 3–4 days following the injury, which correlates with the survival period of this infant after being found unresponsive. (b) Illustrates the fixed, markedly friable, fragmented brain attributable to marked hypoxic-ischemic injury, and is therefore a nonspecific finding, as it can be seen within the setting of hypoxic-ischemic brain injury of any cause.

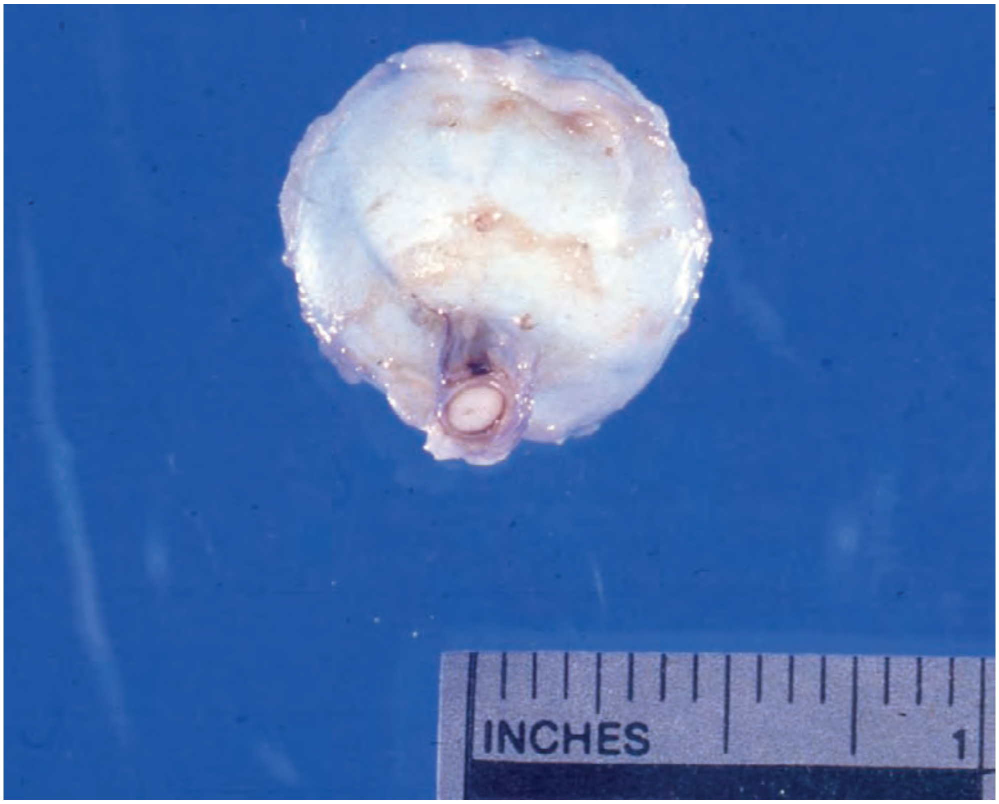

Cross-section of the optic nerve in an infant found to have unilateral subdural hematoma and brain swelling, who underwent evacuation of the subdural hematoma. The infant never regained consciousness and was declared brain dead approximately 3 days following the surgery. Optic nerve sheath hemorrhage is correlated with retinal hemorrhages, although it is not specific for inflicted injury , and may be seen in conditions of sepsis, hemorrhagic diathesis, and other rare natural conditions.

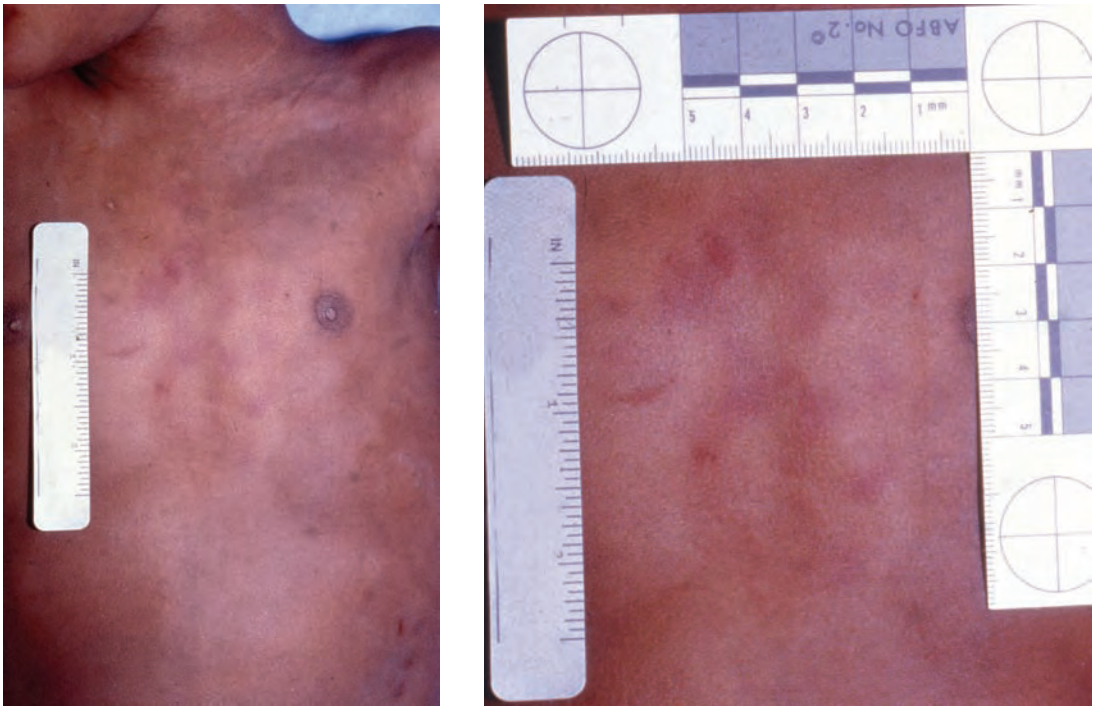

Note the vague pattern contusion of a fist on the anterior chest in this child-abuse fatality. This child died of multiple blunt impact injuries and was discovered dead at the scene in full rigor. No resuscitation was performed.

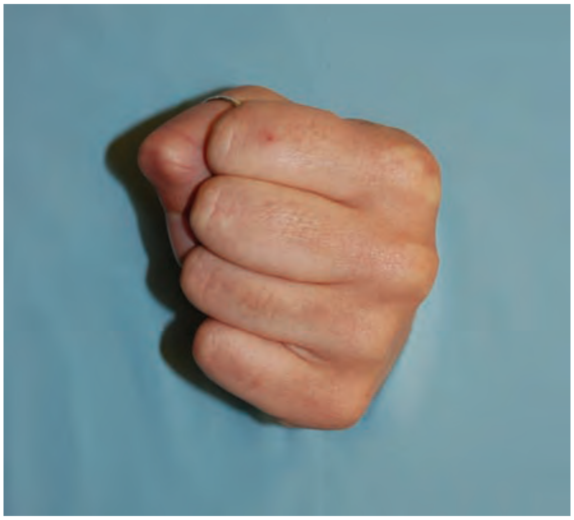

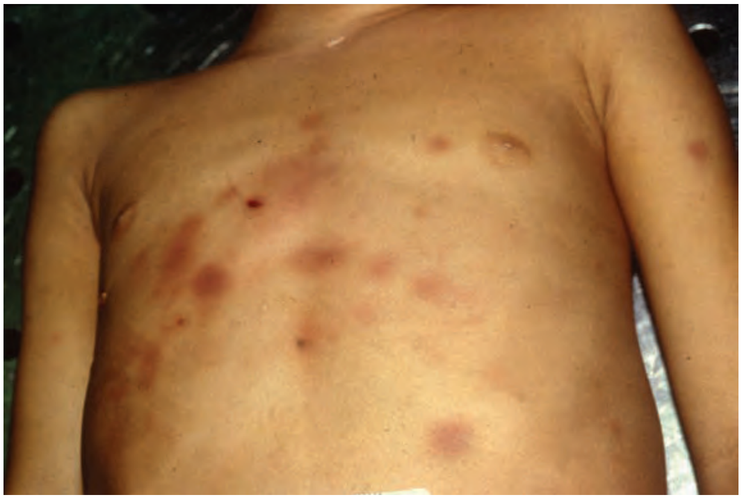

Antemortem contusions associated with child abuse combined with superimposed perimortem contusions associated with resuscitation. The resuscitative injuries overlie the sternum.

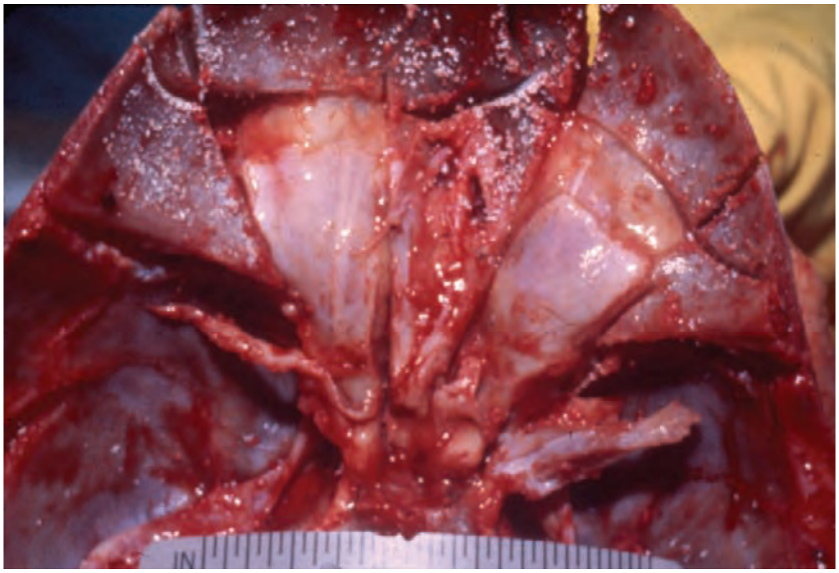

This demonstrates saw cuts through the floor of the anterior cranial fossa (orbital plate of the frontal bone) prior to removal of the eye with its attached optic nerve.

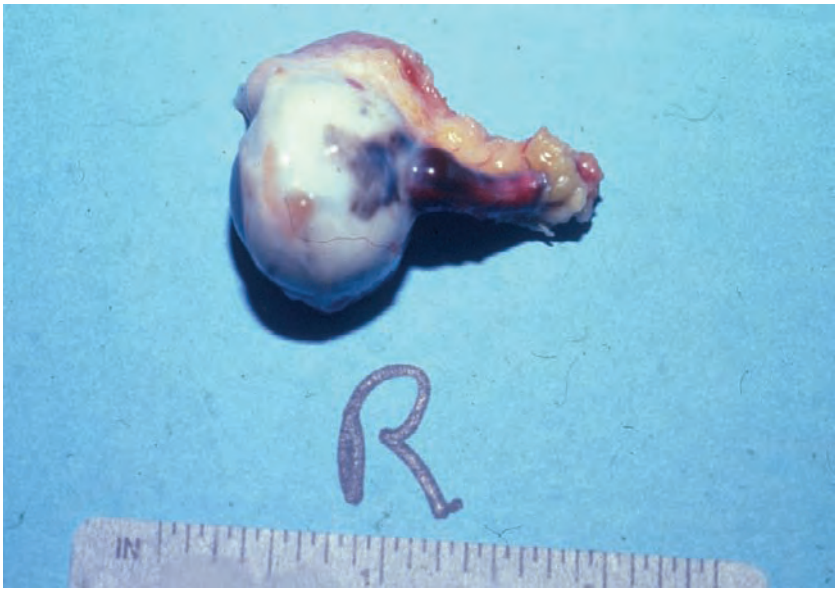

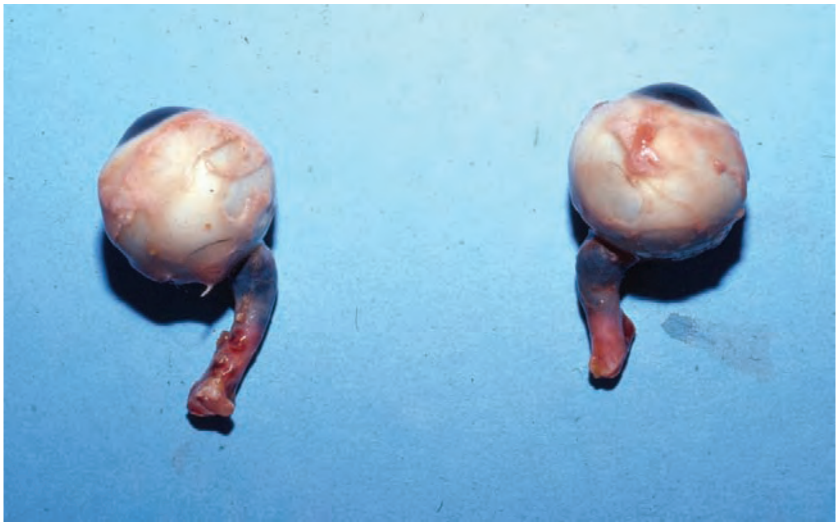

Note the optic nerve sheath hemorrhages. Both of these images demonstrate optic nerve sheath hemorrhages in small children who died from shaken-impact syndrome. Histopathology revealed retinal hemorrhages. There was no evidence of natural disease including coagulopathic syndromes or connective tissue disorders.

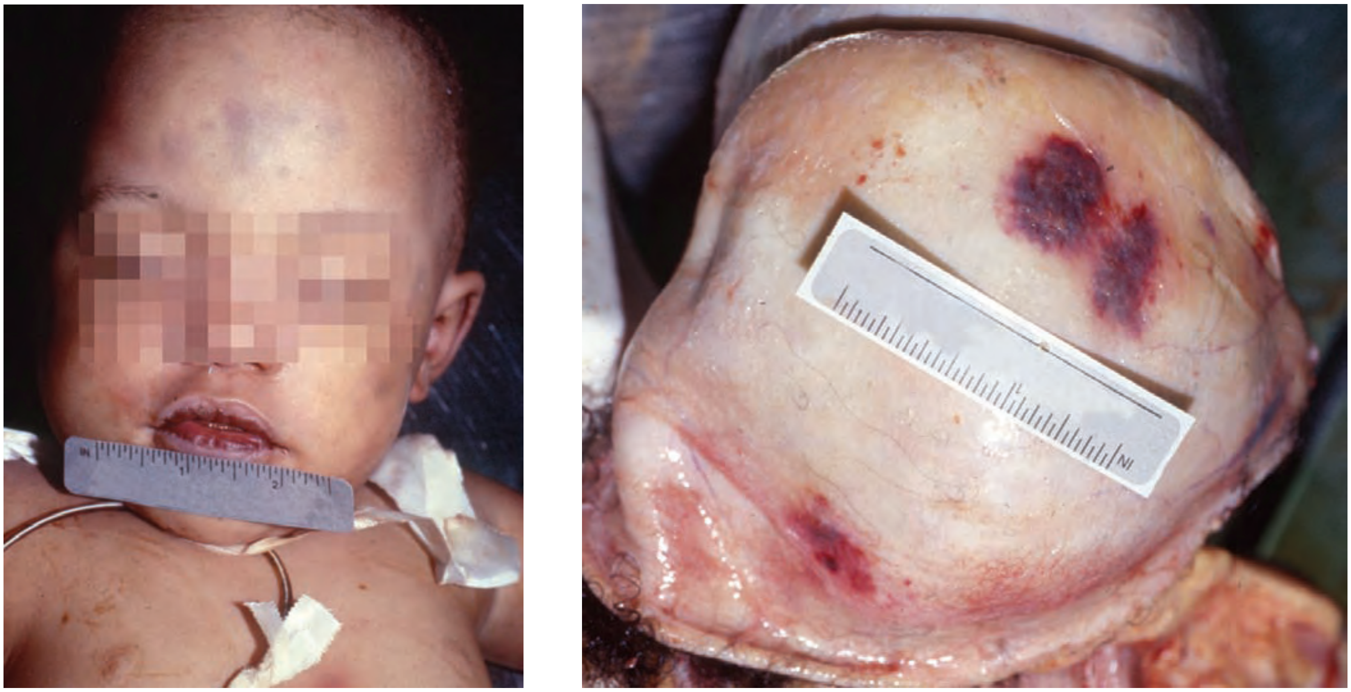

The first image of this child-abuse case demonstrates vague contusions on the child's forehead. Reflection of the scalp shows much more prominent soft tissue (subgaleal) hemorrhage.

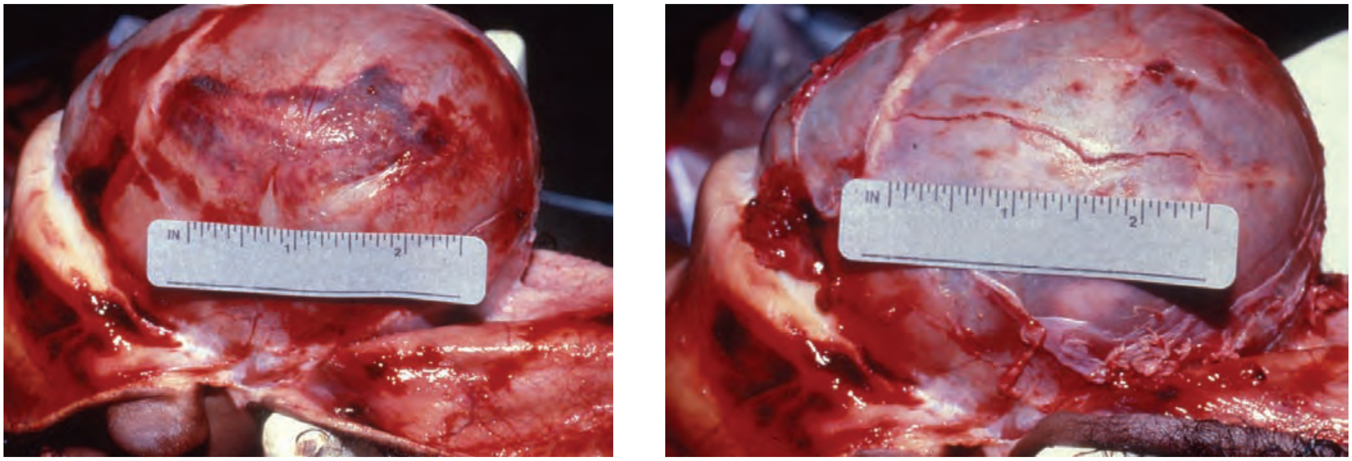

Child abuse case with reflected scalp demonstrating contusions and underlying linear skull fracture.

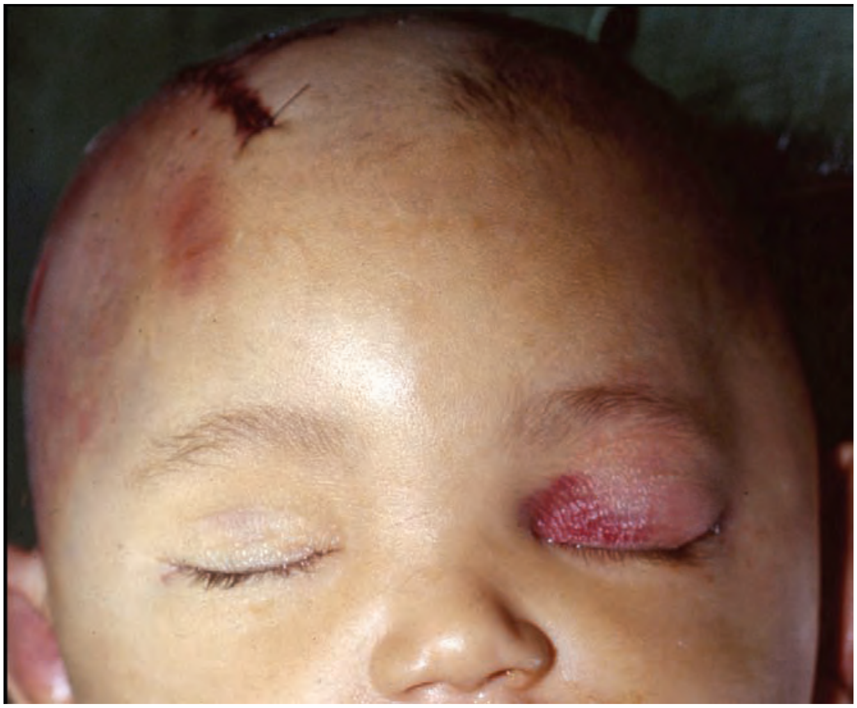

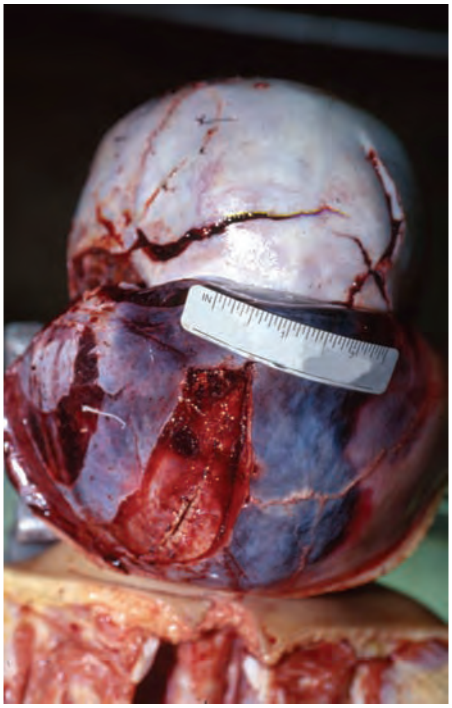

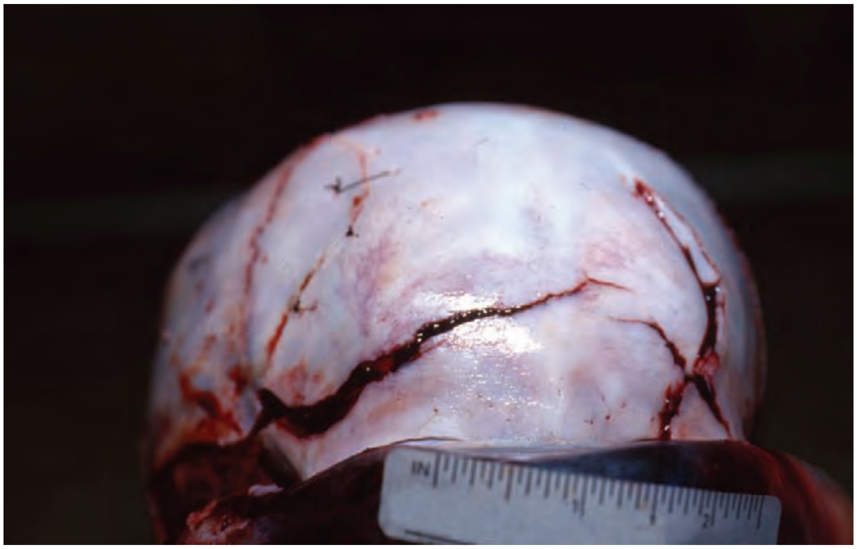

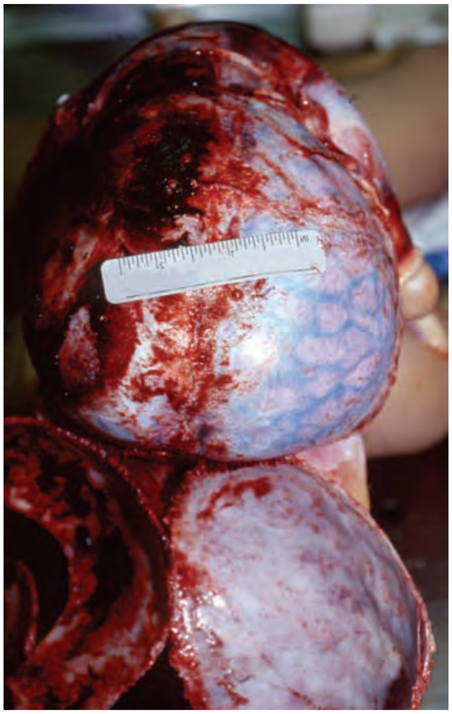

This child was visiting his grandfather with his parents from another country. The grandfather's apartment was not made "child-safe" prior to the visit. The toddler was standing and rocking a large old wooden television back and forth on a stand that was missing several screws. The television toppled over and crushed the child's head. Note the fracture pattern to the skull indicating a crush-type injury. Also note the photograph demonstrating the top of the skull reflected with underlying dura mater still overlying the brain with epidural hemorrhage. The fractures involved the temporal, frontal, and parietal bones and there was severe brain injury with brain swelling and herniation.

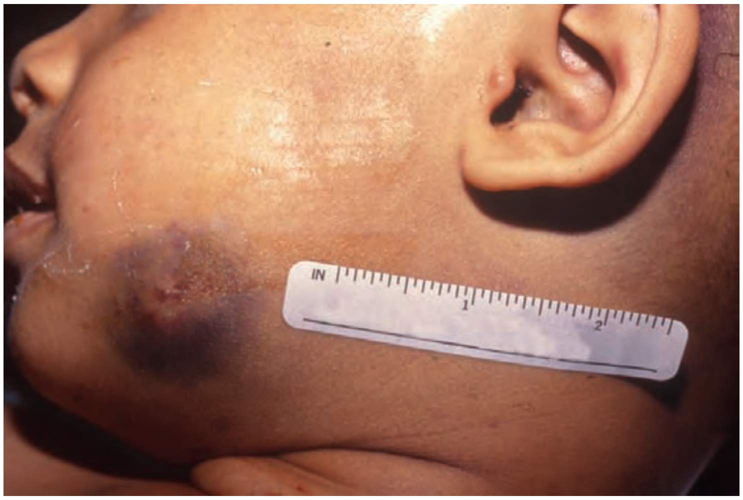

Note the large contusion to the side of the child's face from being punched.

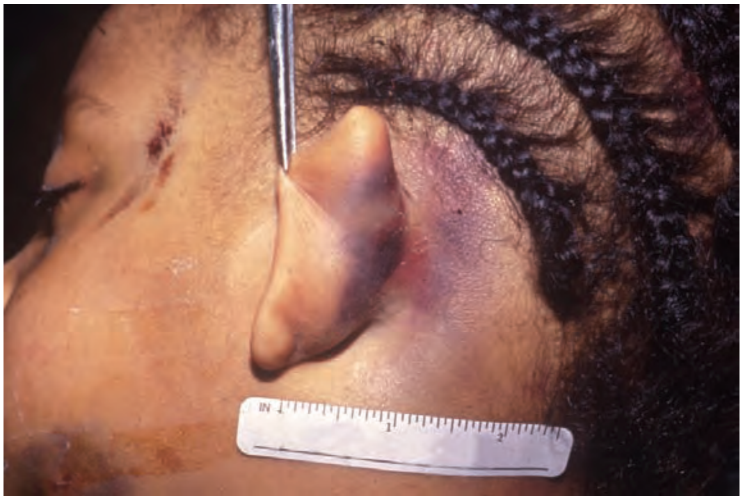

This child was the victim of an assault and suffered a skull fracture with brain injury that caused his death. Note the area of hemorrhage to the mastoid portion of the scalp demonstrating "battle sign."

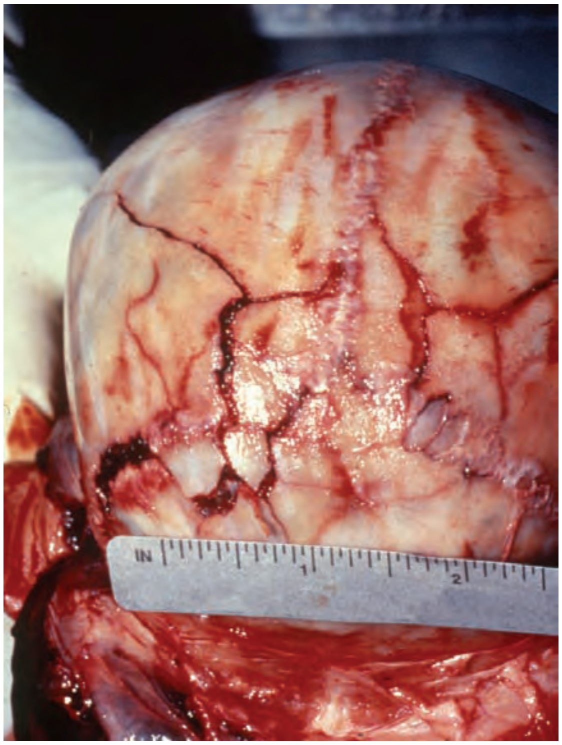

Child abuse case with comminuted skull fracture. This was most likely produced by flinging the child in a downward trajectory, and striking the child's head against a hard countertop. A fracture of this nature would take tremendous force to produce.

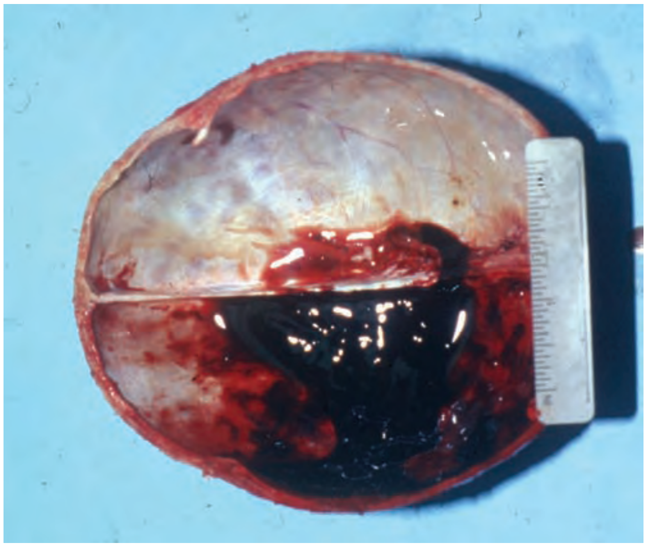

This demonstrates the top portion of the skull (calvarium) removed during autopsy with the dura mater still attached and clotted, partially adherent, subdural hemorrhage. Clotted adherent red-black subdural hemorrhage indicates an injury age of approximately 2–4 days.

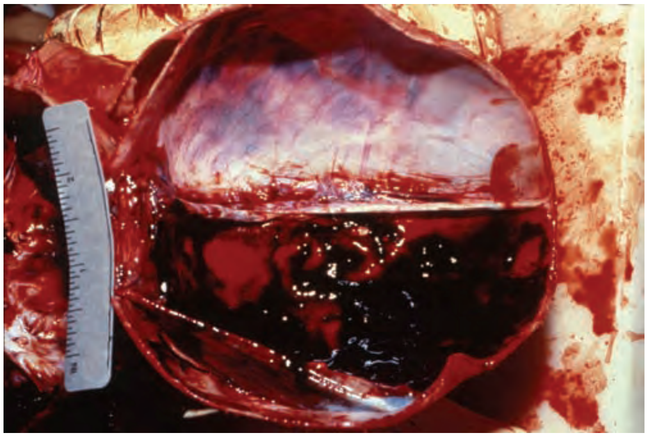

This is the top portion of the skull with adherent clotted subdural hemorrhage involving the left vertex dura (overlying the left cerebral hemisphere) with portions of dura mater separated at the posterior temporal-parietal-occipital regions, adjacent to sites of skull fracture.

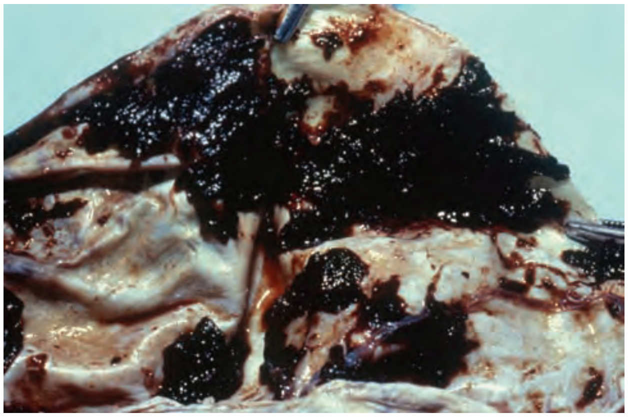

Portion of formaldehyde-fixed dura mater saved for further neuropathologic evaluation. Note the adherent clotted, black-brown subdural hemorrhage.

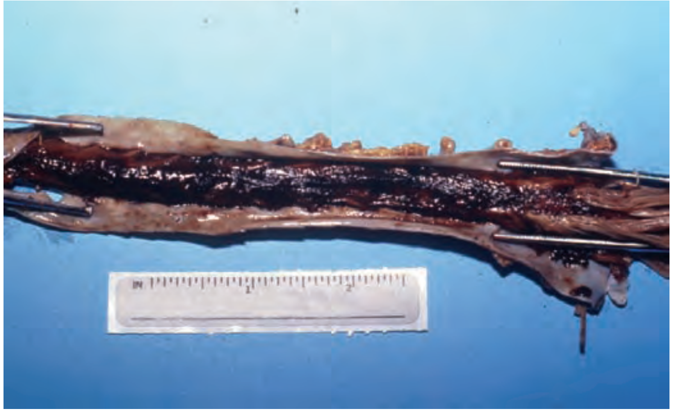

Formaldehyde-fixed portion of spinal cord removed with the dura mater demonstrating subdural hemorrhage. It is important to remove the cervical spinal cord as carefully as possible during the autopsy process in order to avoid potentially troubling postmortem artifact.

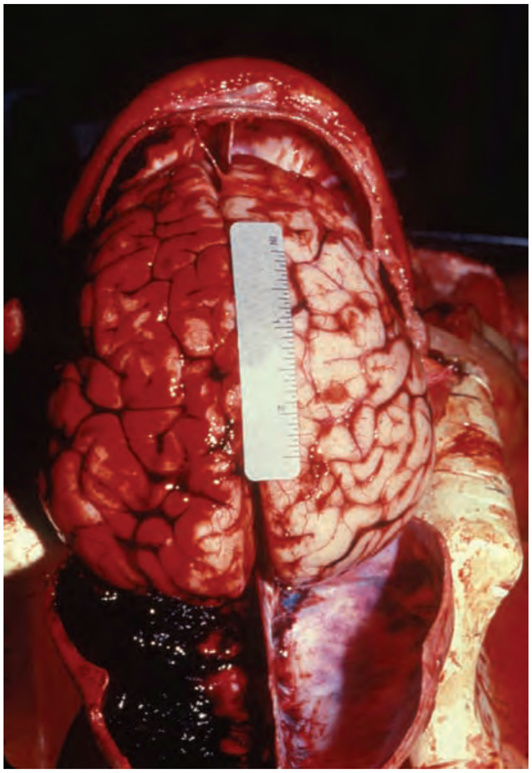

This shows the top portion of the skull (calvarium) being unroofed during the autopsy process for removal of the brain. Note that the dura mater is still adherent to the inner aspect of the skull (endocalvarial surface). There is a large left-sided adherent clotted subdural hemorrhage. There is diffuse cerebral swelling with flattening of the gyral configuration and obliteration of the sulci. There is patchy subarachnoid hemorrhage, which is greatest over the left cerebral hemisphere.

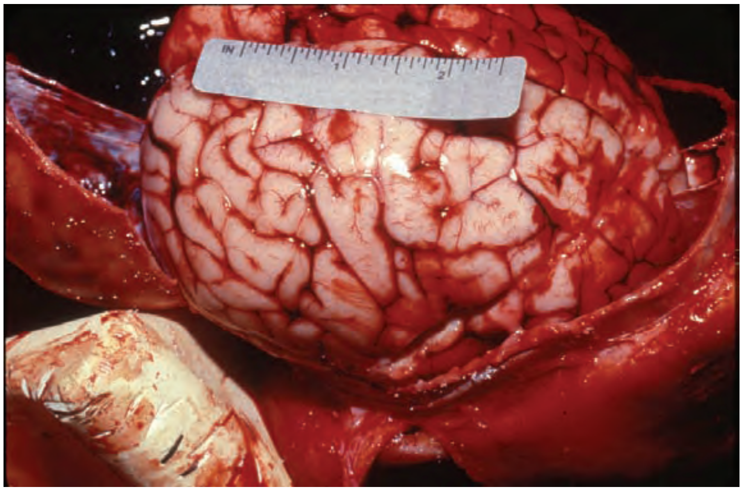

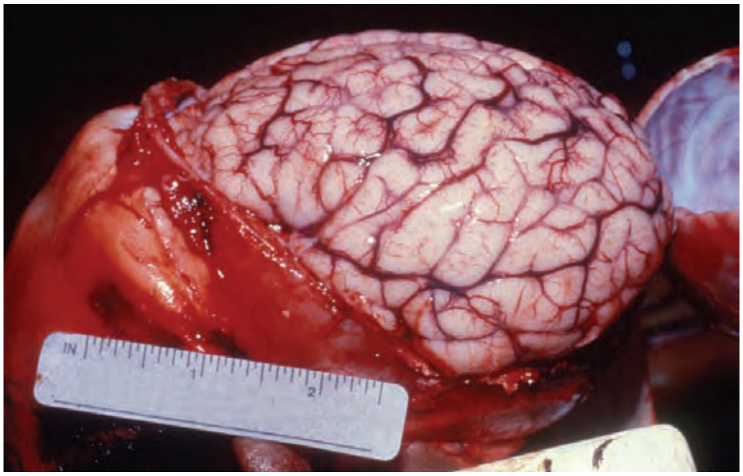

Close-up views of brain swelling with flattening of gyri and obliteration of sulci in a child who survived several days following blunt force trauma due to homicidal violence.

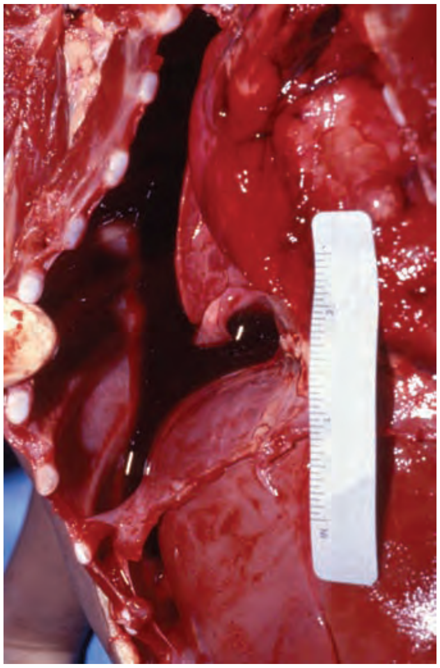

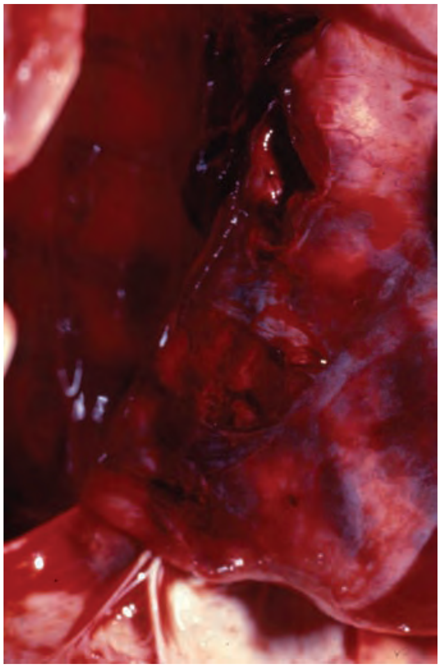

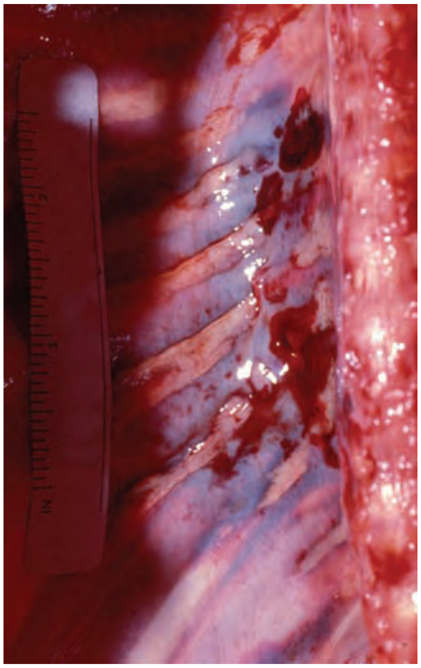

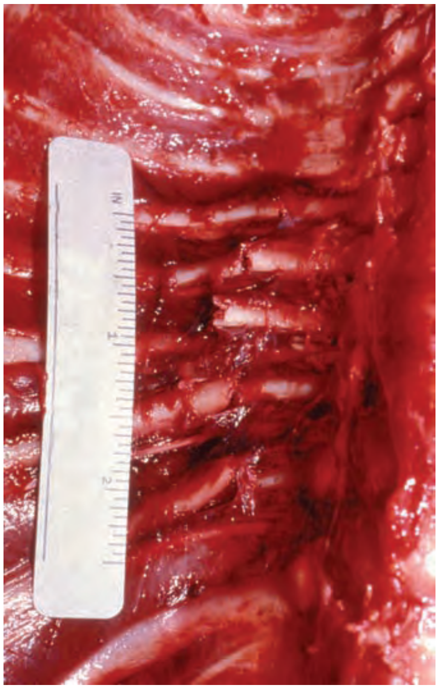

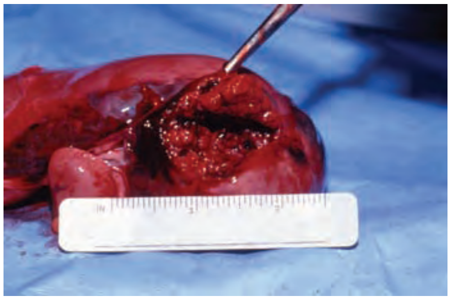

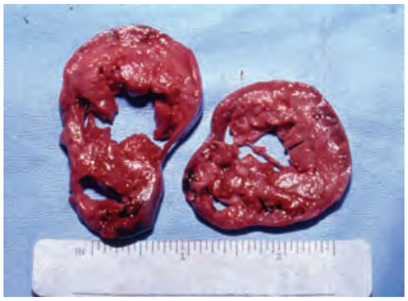

This child sustained multiple posterior rib fractures after being stomped by an adult while lying face-down on the floor. The fractured segments of the ribs penetrated the posterior aspect of the right lung, producing extensive pulmonary and intrathoracic hemorrhage. There were also large lacerations to the posterior aspect of the liver and heart.

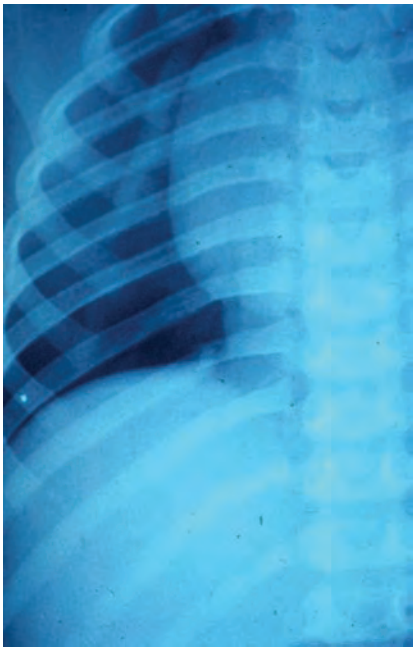

X-ray showing multiple recent rib fractures.

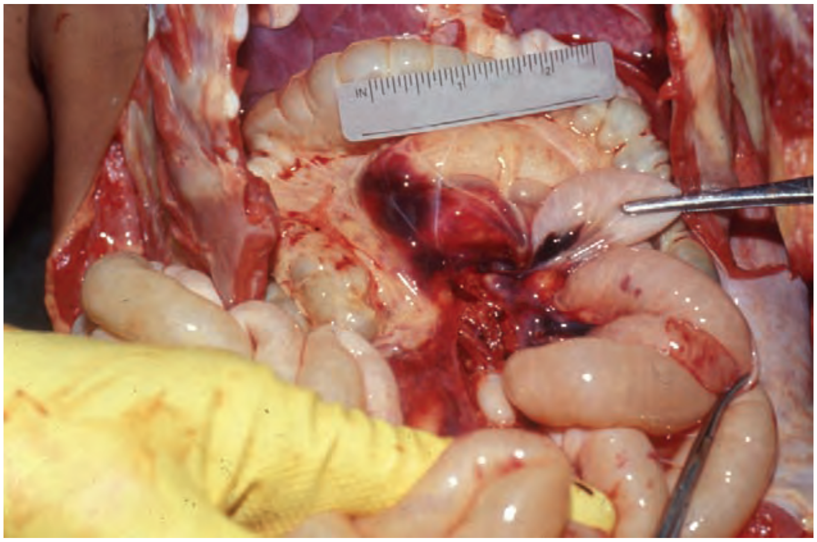

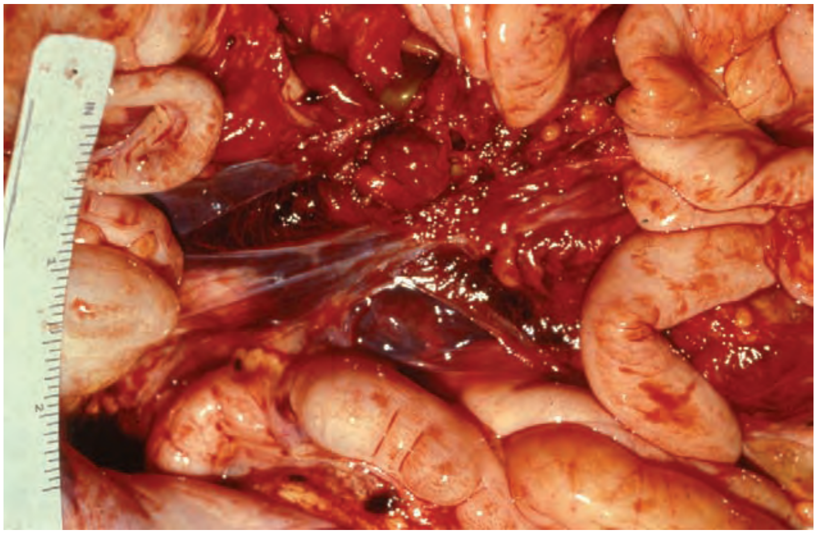

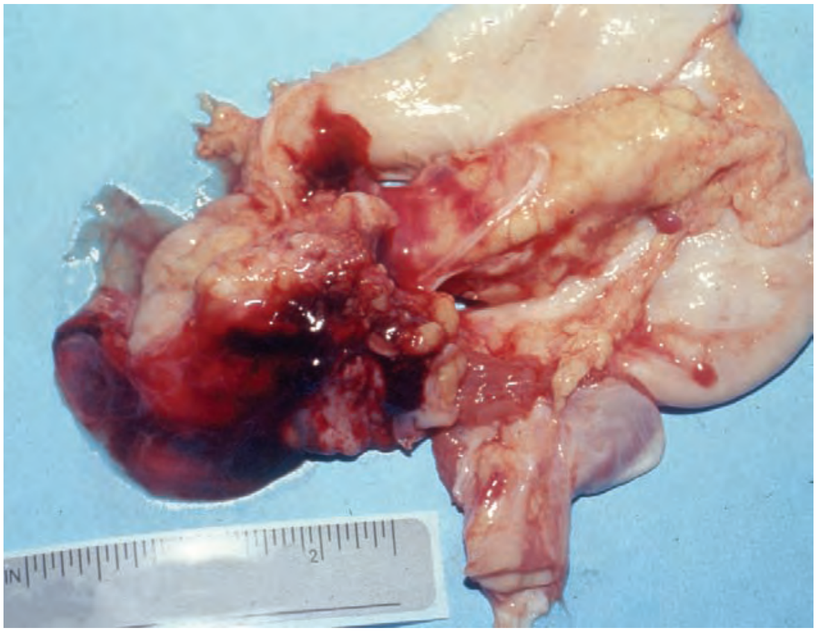

These images demonstrate multiple lacerations involving the mesentery, mesocolon, and peripancreatic soft tissues following blunt abdominal impacts associated with homicidal violence.

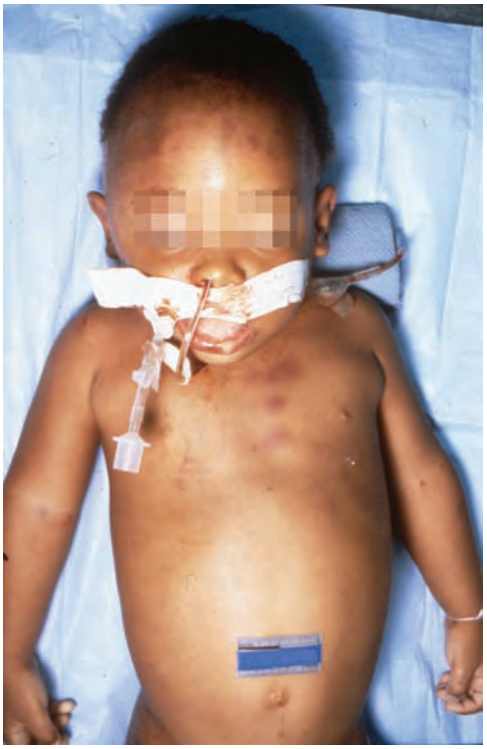

This 2-year-old child was found by a caretaker unresponsive in the bathtub. An autopsy revealed confluent subgaleal contusions, essentially spanning from ear to ear, along with brain swelling, although the skull was intact. Death was attributed to blunt impacts to the head with brain swelling. A component of drowning cannot be ruled out in this case. The defendant, a casual acquaintance of the mother, was convicted of second-degree murder and sentenced. Note the bilateral intraosseous catheters, used while attempting to resuscitate this child even though he was slightly over 2 years old. This image depicts the face and anterior trunk of a child with patchy contusions involving the forehead and the left paramedian aspect of the anterior chest wall. Multiple therapeutic interventions are also present. As stated previously, differentiating between the effects of therapy and antecedent injury can be problematic, and the two processes may coexist.

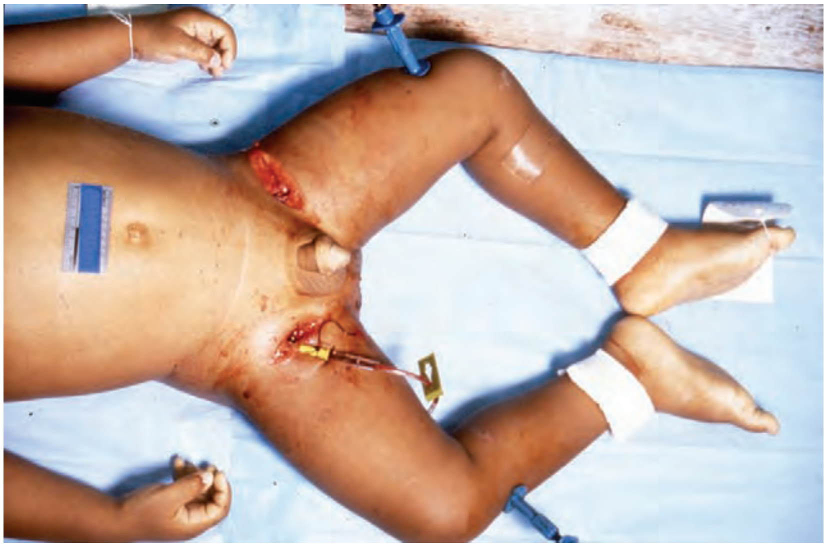

This depicts lower trunk and lower extremities of child, with multiple therapeutic interventions including transverse cutdown sites within both groins, bilateral intraosseous catheters, and a right groin catheter.

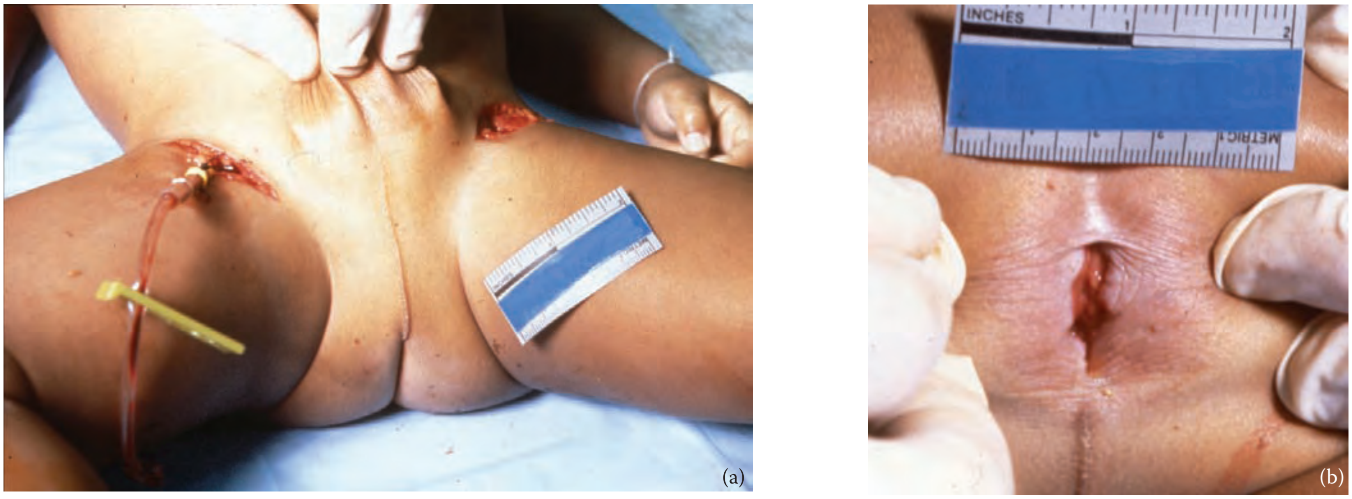

This image shows the perineal area of this child (a). No injuries are detected. (b) depicts the child's anus. No injuries are demonstrated. In cases such as these, it is good practice to document photographically the presence or absence of injuries, as the multiplicity of injuries may be one of the determining factors in the ultimate sentencing. Sometimes external examination may be difficult and mucosal folds may be misinterpreted as lacerations. If there is any question as to the nature of a possible lesion or injury, we recommend removing the region en bloc, opening the anorectal segment and photographing it for documentation.

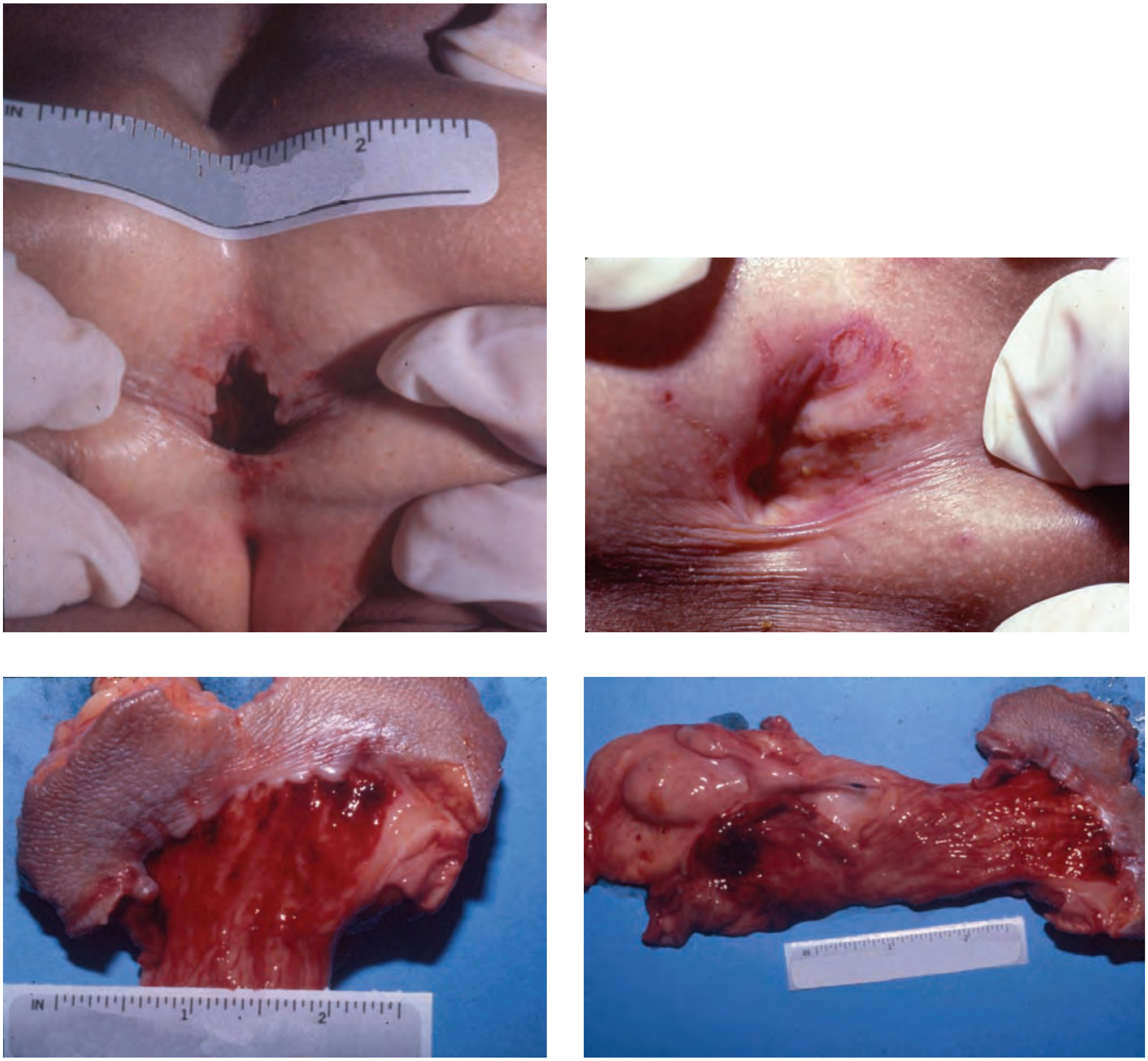

These images demonstrate sexual abuse in a young child. Note the abrasions and lacerations around the anal region. En bloc dissection of this region with the anal-rectal segment of the intestine removed. Note the internal lacerations near the anal border and the intestinal hemorrhage

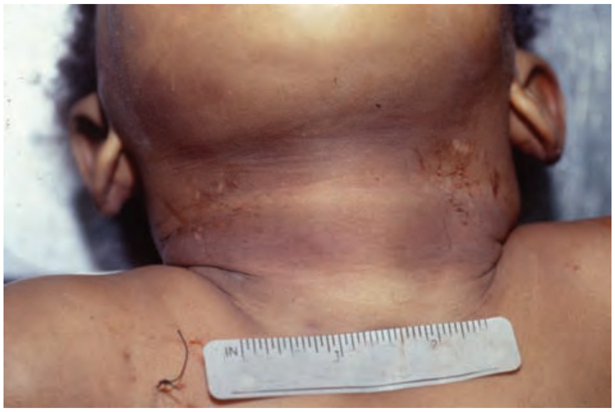

Superficial, healing, criss-crossed, linear abrasions due to rubbing of Velcro jacket patches on each side of this child's neck. This indicates rough handling. This child was later killed by multiple blunt impact injuries during an assault. Note the suture and fresh needle marks at the right chest from a removed subclavian line.

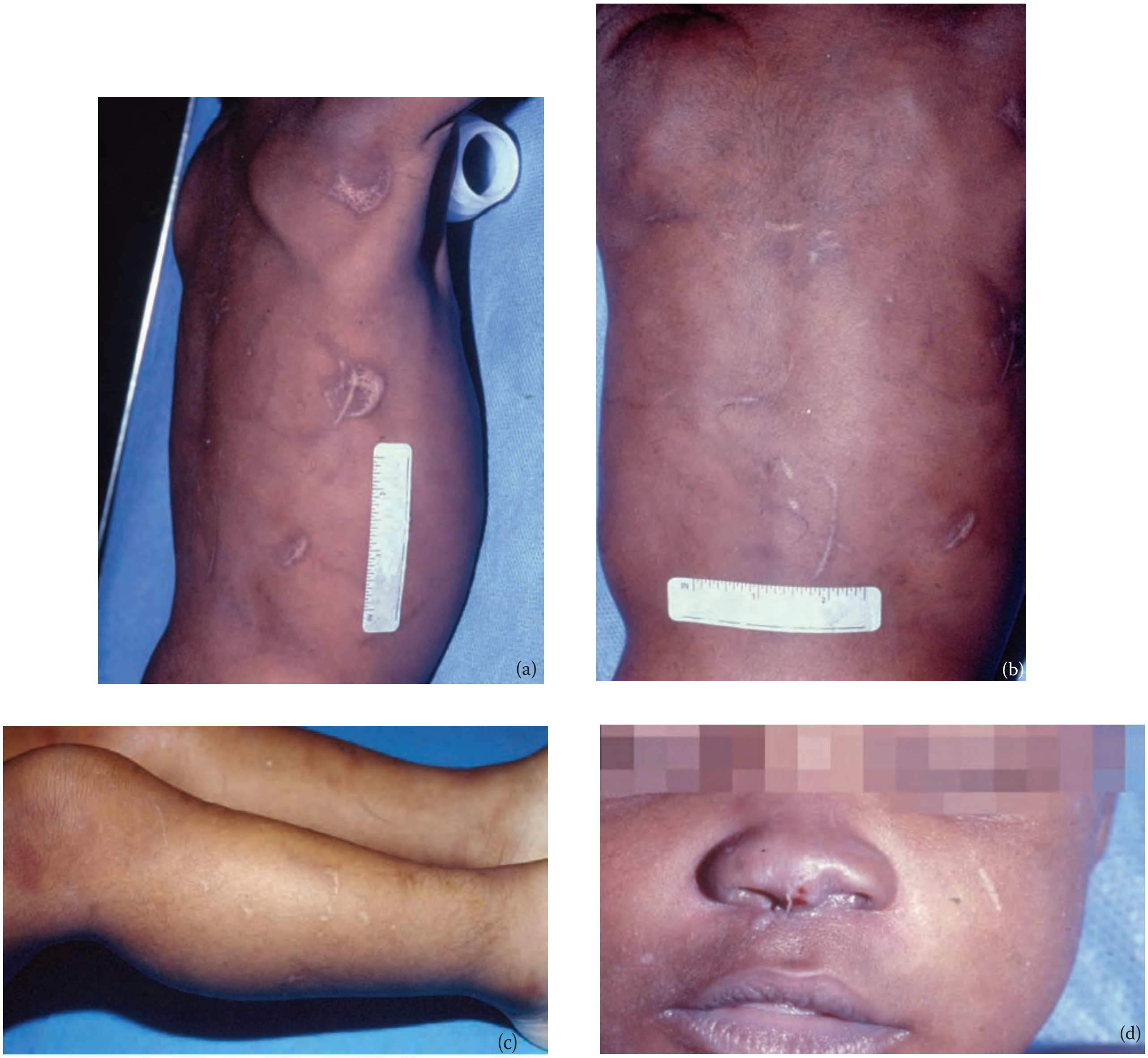

These are multiple scars, many of which were associated with chronic child abuse. When evaluating old injuries, it is important to take into context the child's age and developmental status. A child who is able to run and jump about may fall and sustain blunt-impact injuries. Sometimes it may be difficult, if not impossible, to formulate the nature of such old injuries without history or witnesses. (a) and (b) images are associated with abuse; (c) and (d) represent healed scars from accidental injuries. Note the healing injury at the nose associated with a nasogastric tube and therapy.

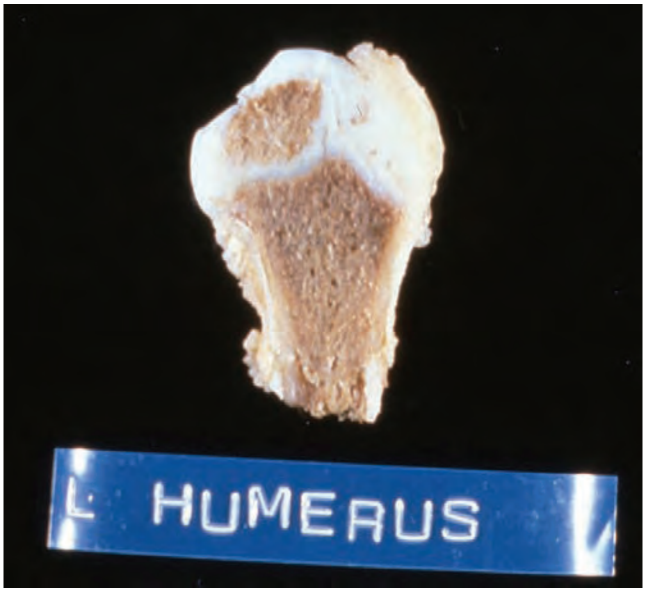

This demonstrates a coronal section of a humerus. Microscopic sections of long bones may be sampled in cases involving a suspicious death of an infant or young child in order to identify transmetaphyseal fractures in situations in which shearing forces have been applied (i.e., violent shaking). Unlike the case in more obvious fractures with extensive visible hemorrhage, these fractures (alternatively referred to as "bucket handle" or "corner" fractures on radiologic plain films, depending upon the view), may sometimes be more readily documented with microscopic examination.

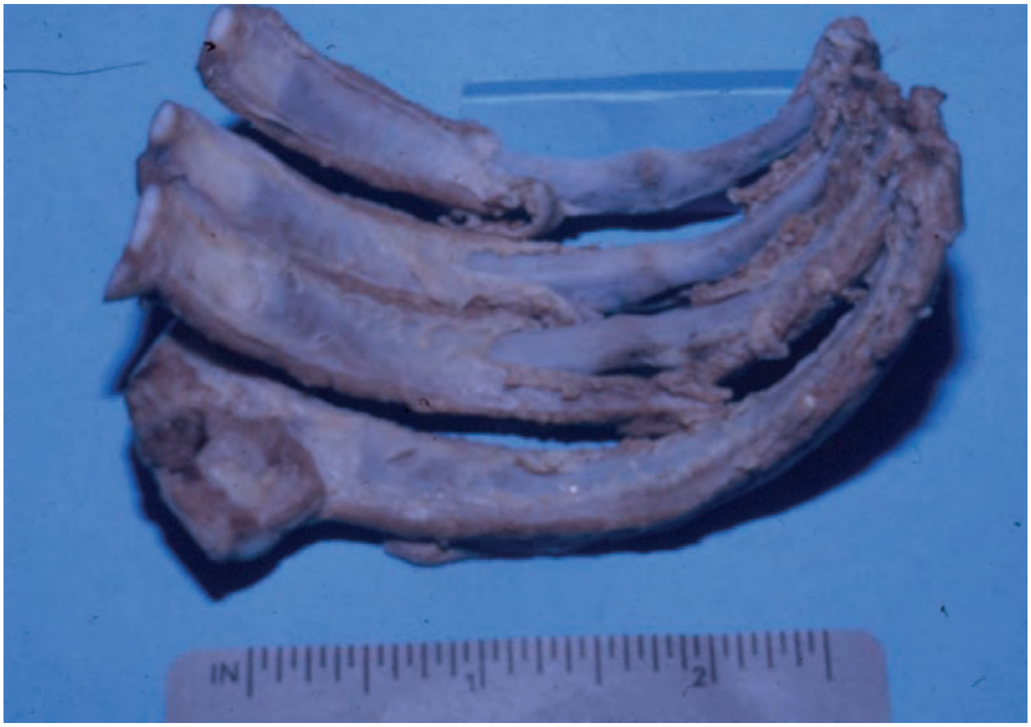

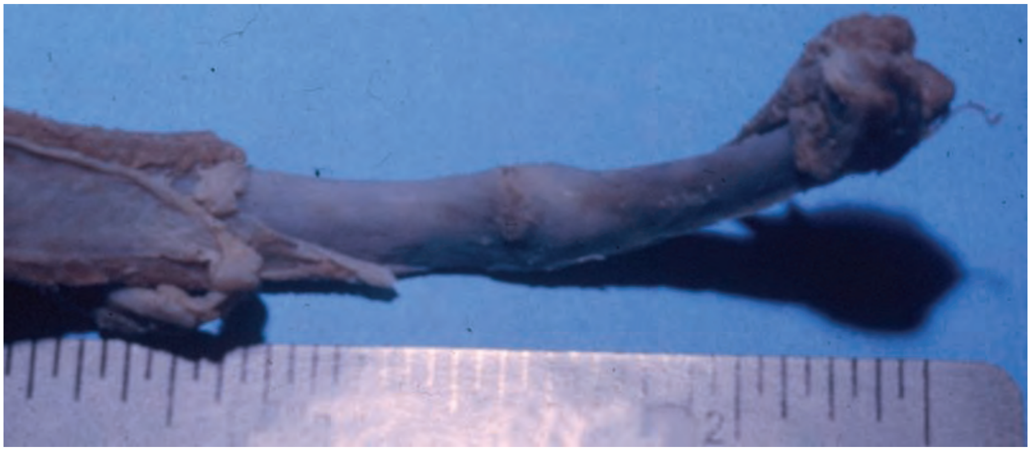

Old healed rib fractures from a case of battered child syndrome.

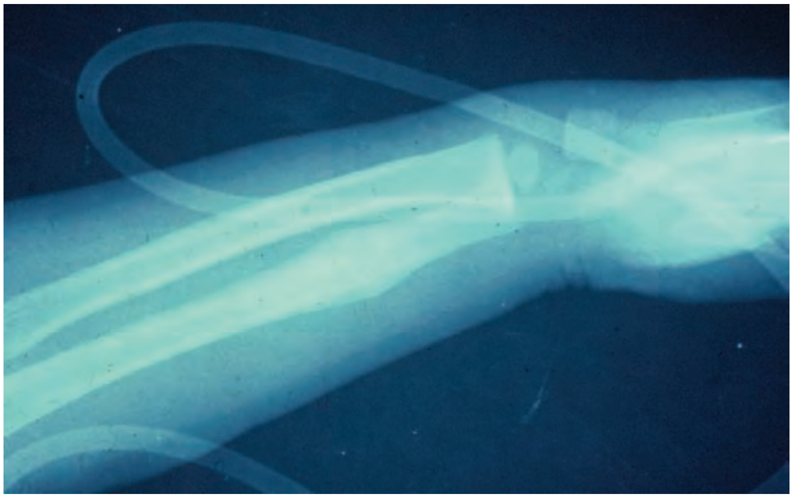

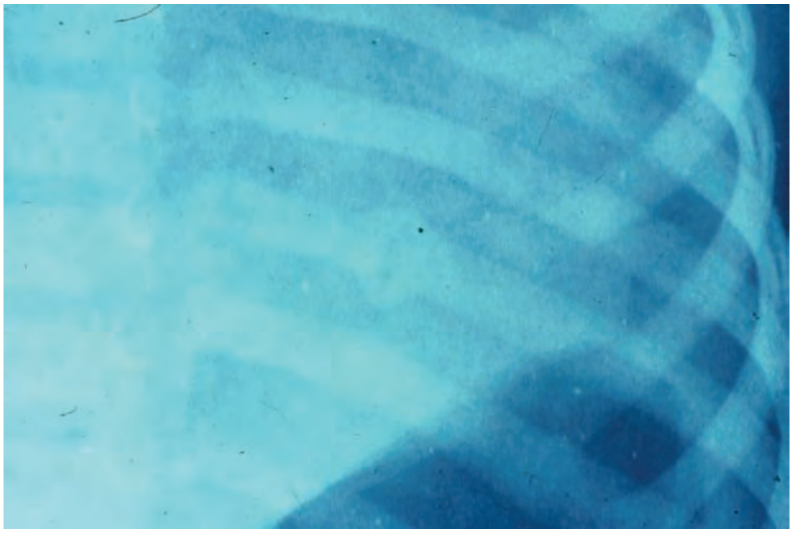

These x-rays demonstrate healing fractures.

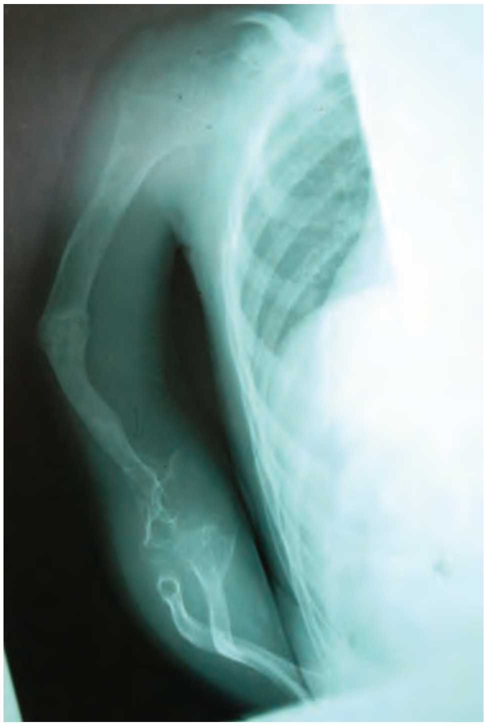

Osteogenesis imperfecta with multiple fractures. This may or may not indicate child abuse. Fractures may occur with very little force.

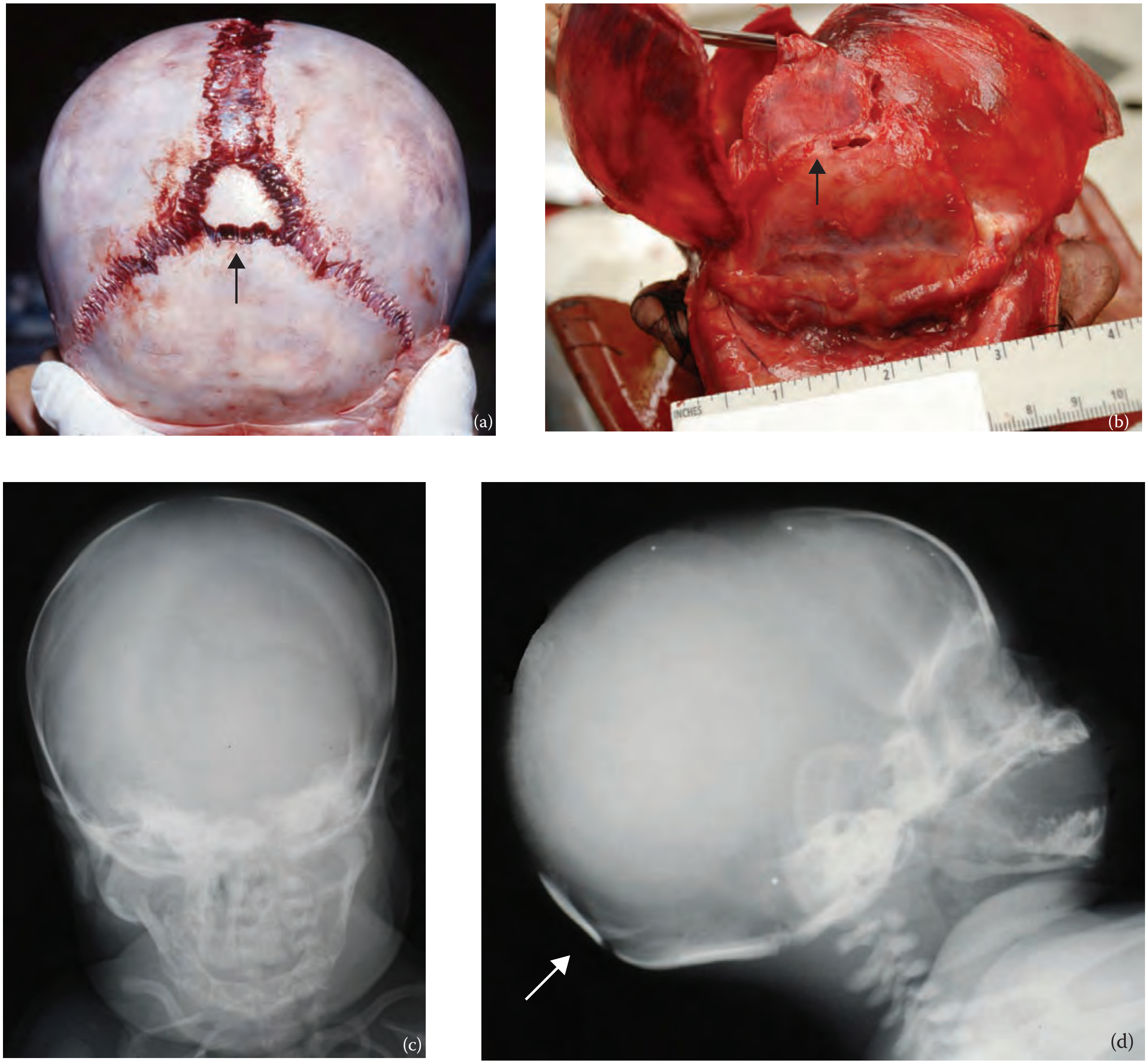

Inca bone skull anomaly. In both of these cases it was mistaken for a skull fracture. The x-rays at the bottom correspond to (b).

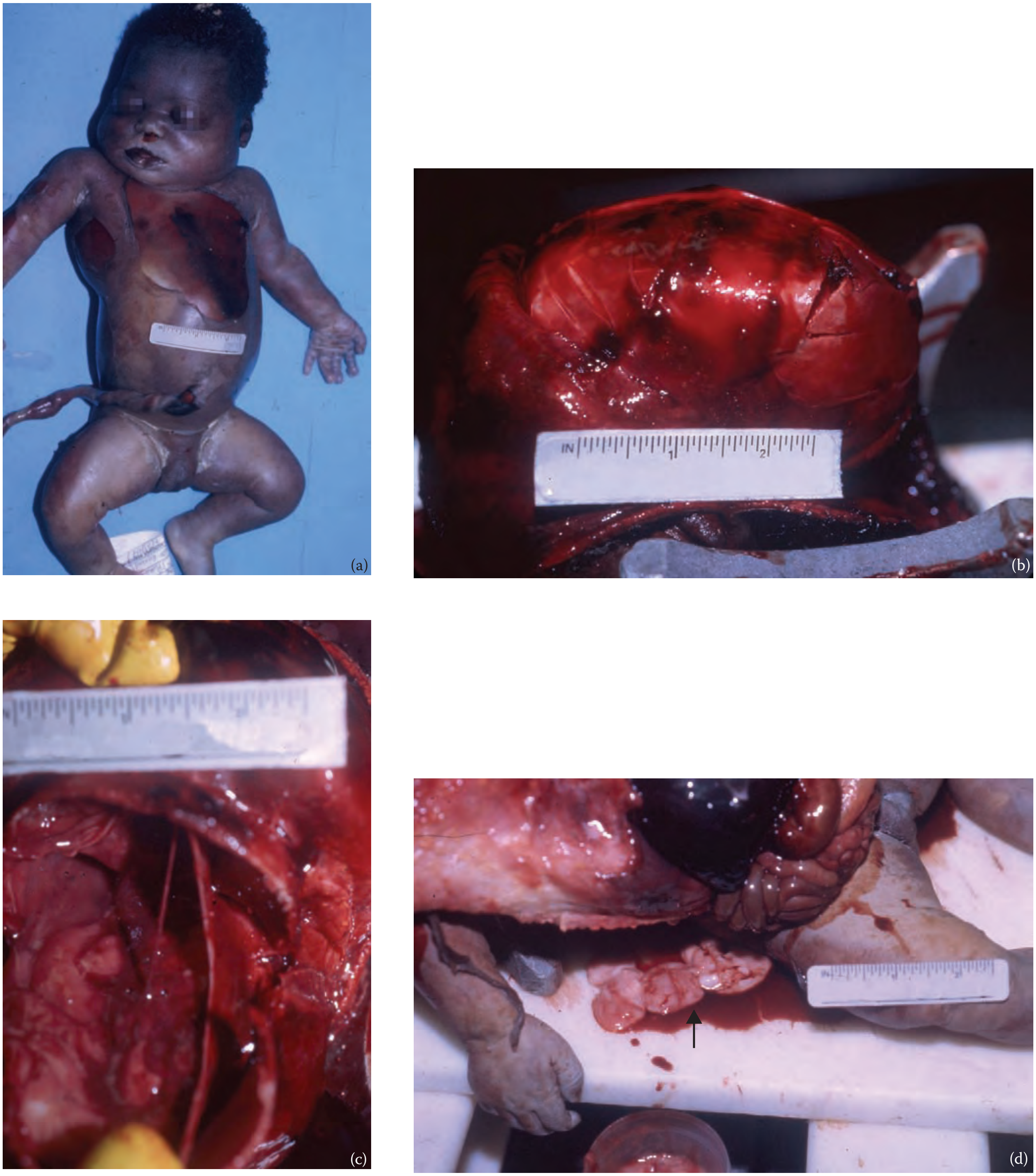

This child was retrieved from the back of a garbage truck partially tied within a torn plastic bag following compression by a trash compactor. Note the confluent hemorrhage to the scalp and the largely empty cranial vault. This child was mildly to moderately decomposed with bloating and separation of tissue planes. The brain material was squeezed from the cranium through the neck following compression of the head by the compactor. When the initial Y-shaped autopsy incision was made to the skin of the trunk, brain material leaked from the chest region (d). The child's body was found mixed with decomposing food material. Note the green discoloration from putrefaction.

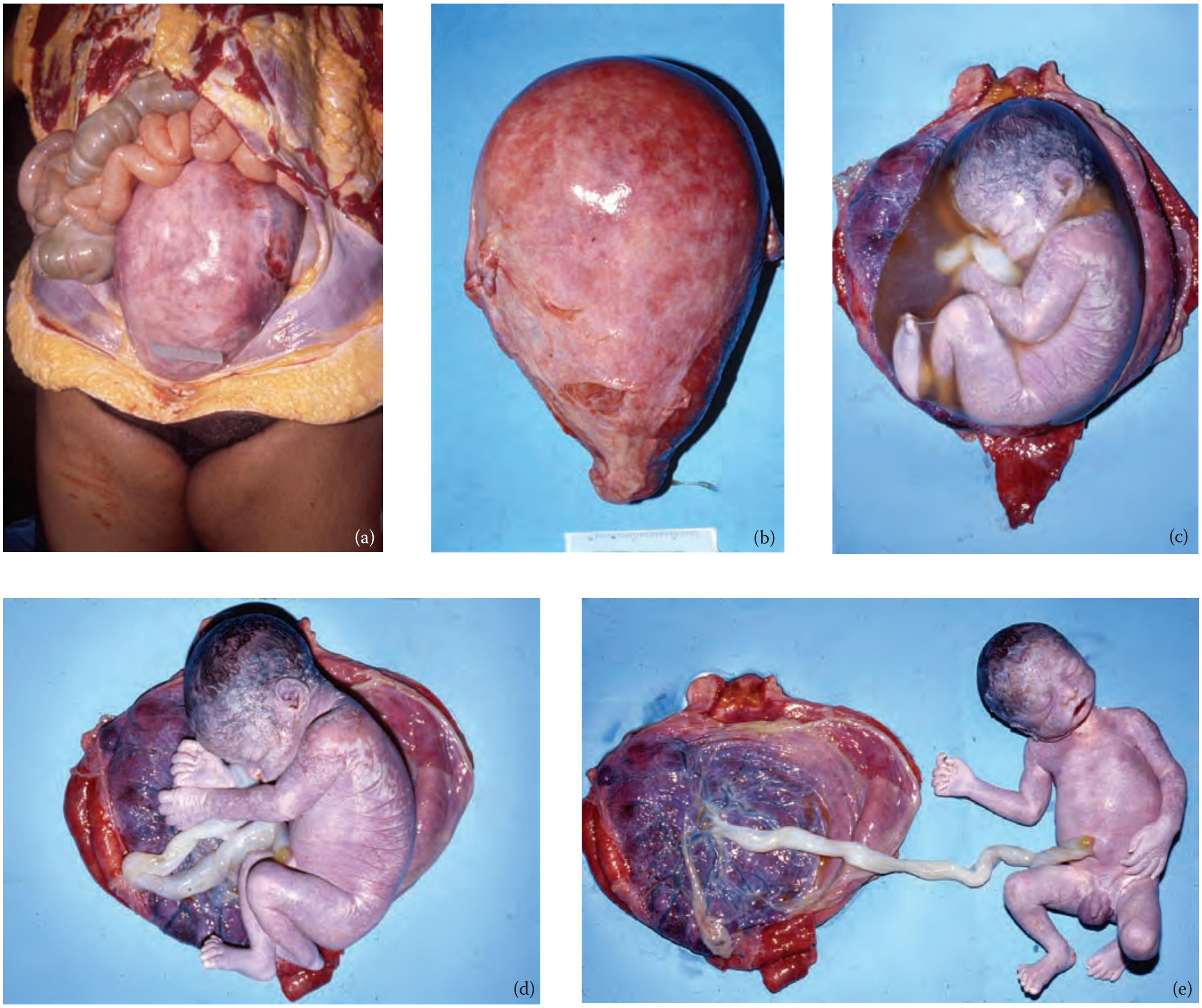

Late-second-trimester pregnancy demonstrating the uterus (a), placenta (b), and fetus (c).

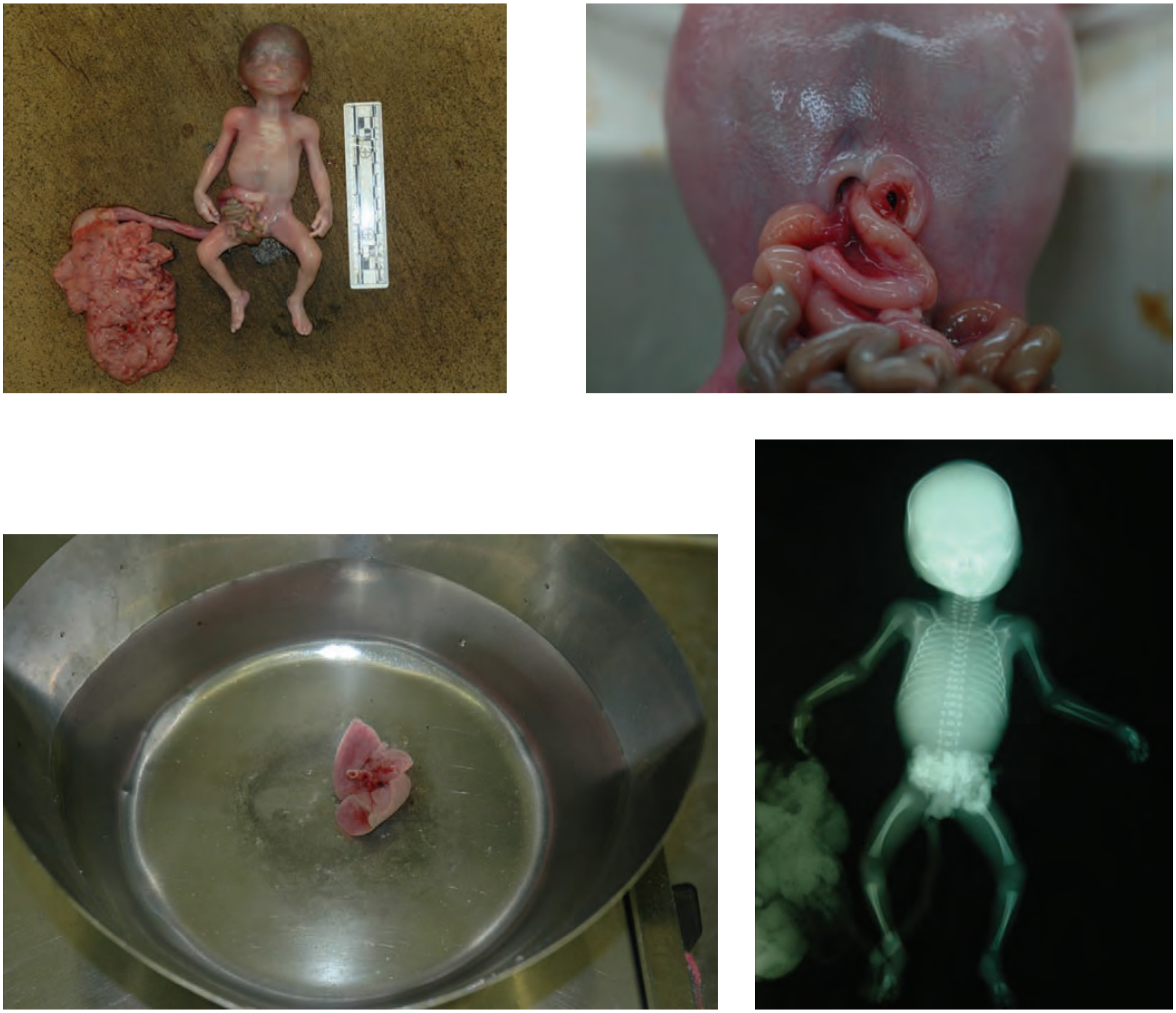

Second trimester stillborn with omphalocele, illegally aborted following administration of drugs given by a nonmedical practitioner. Note the x-ray and the sunken lungs at the bottom of the water container indicating no breaths were taken.

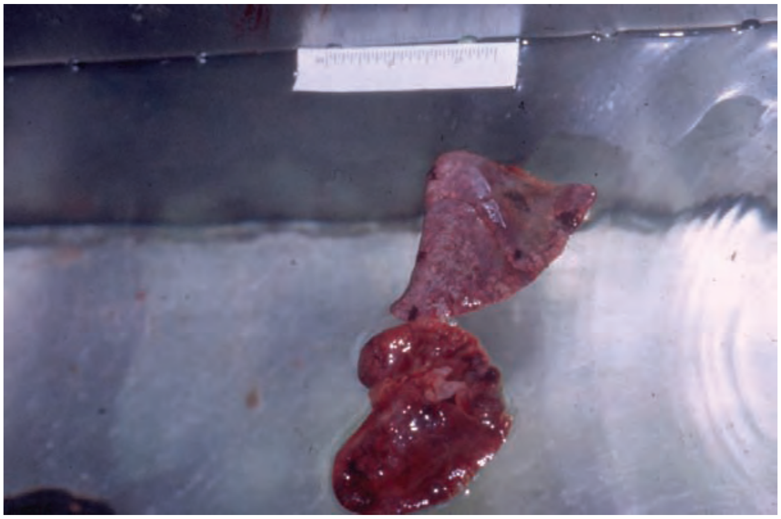

Lungs from a live birth that were expanded with air. Note how they float on top of this water bath. Admittedly, the "flotation test" to confirm independent respiration is far from foolproof and is subject to a number of artifacts and exceptions, thus limiting its utility in resolving the question of live birth versus stillbirth.

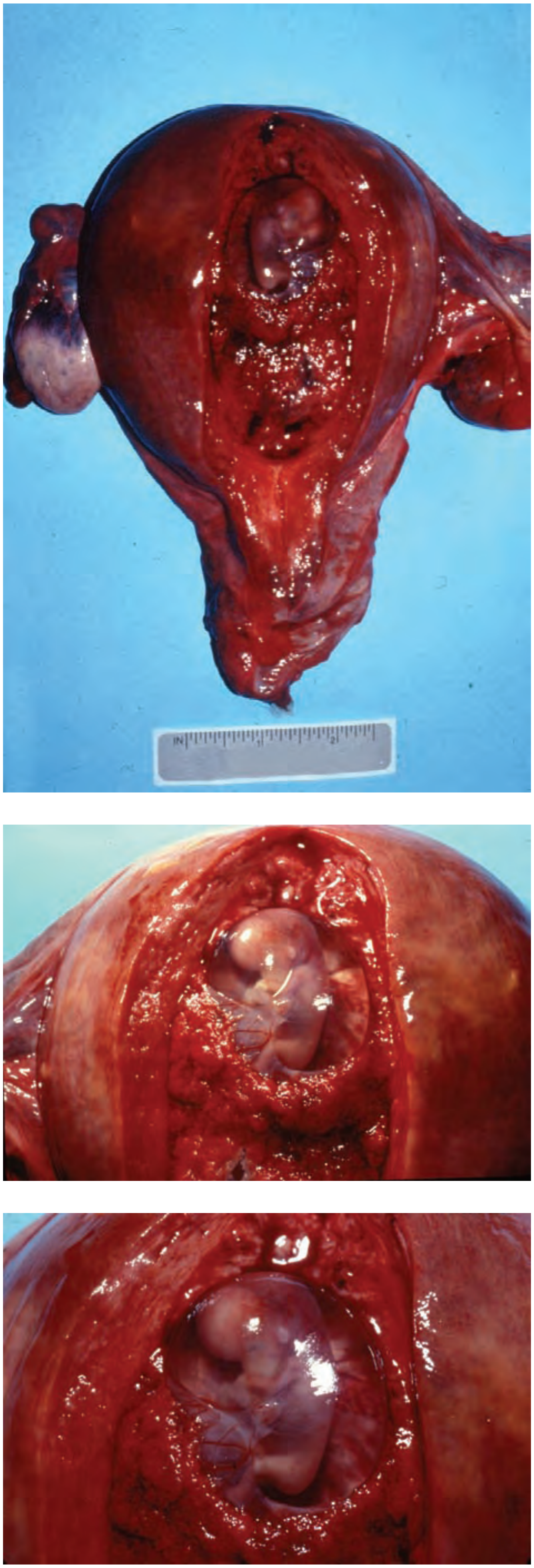

First-trimester pregnancy.

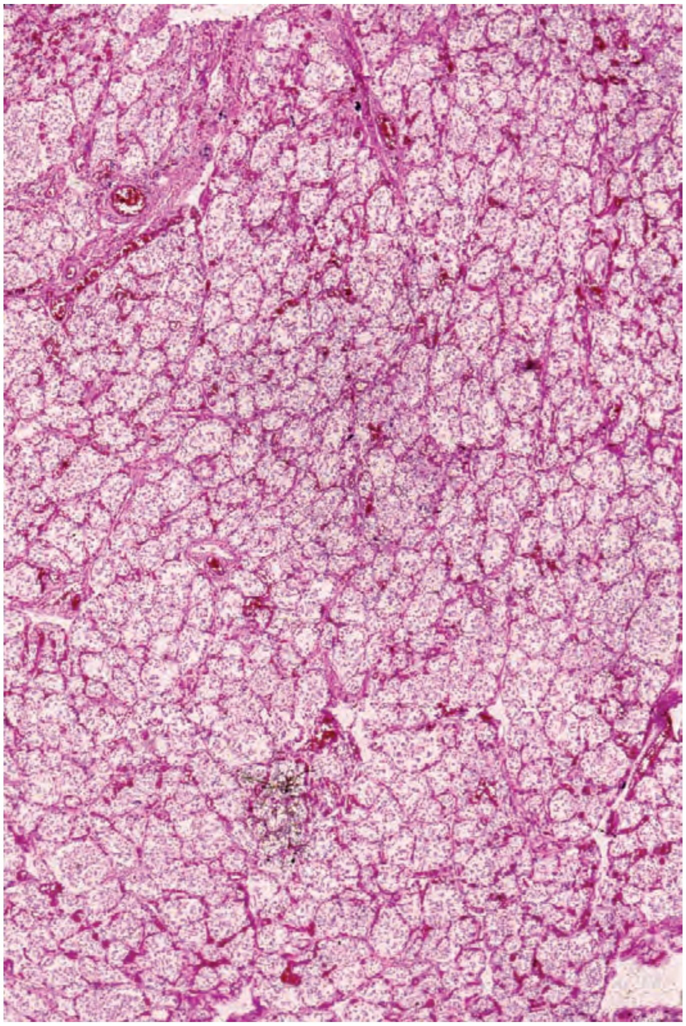

Brown fat.

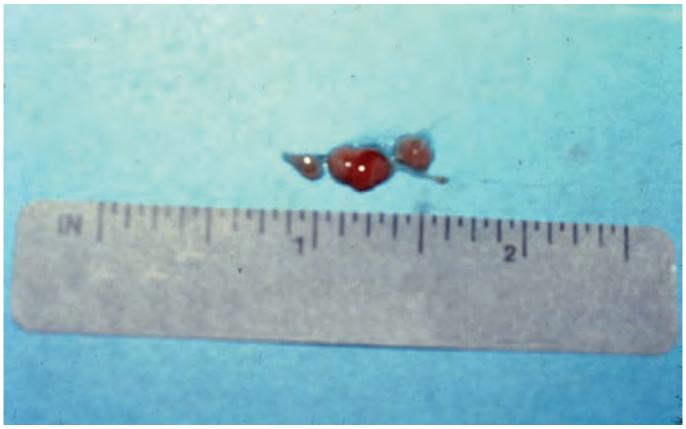

Early first-trimester pregnancy.

Blunt-Force Injuries

6

CHARLES A. CATANESE

Introduction

These are injuries produced when the body is struck with, or strikes, a blunt object. A blunt surface produces injuries by scraping, tearing, shearing, or crushing. This is in contrast to a sharp force injury, which cuts and separates the tissue as it penetrates. What injuries result from trauma are a balance of the amount of force, the area over which it is applied, and the duration of force. In general, the greater the force, smaller the area, or shorter the duration over which the force is applied, the greater the injury will be.

Blunt objects have a relatively large surface area in contrast to sharp objects, where the cutting edge has a relatively small surface area. It takes much less force to penetrate the skin with an ice pick than with the end of a baseball bat. Examples of blunt objects are fists, shoes, pipes, bricks, bats, hammers, roadways, sidewalks, cars, trains, airplanes, walls, etc. Types of blunt-force injuries include *abrasions, contusions* and *lacerations*. These may occur separately but are often present at the same time. As an example, an injury maybe described as an abraded contusion with central laceration. *Fractures* are breaks in the bone as a result of blunt force. *Avulsions* are splits in the soft tissue or soft tissue planes with or without a laceration as a result of shearing forces.

Abrasions

Abrasions are simply scrapes. They are produced as the body contacts a surface and rubs across it with sufficient force. The distance of travel can be very short (falling and scraping your knee) or very long (being dragged under a car). It is often possible to determine the direction of impact. Layers of skin are scraped away and bead up at the margin where the contact to the wound last occurred. Fragments of skin beading up at the inferior wound margin indicate a downward impact. Antemortem abrasions or those that occur during life are typically red to brown and will eventually form a scab with dried blood. Postmortem abrasions are yellow in non-lividity-dependent areas. In general, one needs a beating heart with blood pressure to produce hemorrhage and a red to brown discoloration. A postmortem scrape in a lividity-dependent region will appear red and may be difficult to differentiate from an antemortem injury.

This is important in certain cases. Lawyers will argue different ways to create reasonable doubt. For example, some may argue that an injury occurred during resuscitative efforts in the hospital, or the body was dropped from the stretcher in the morgue. Some other arguments may include: he was dead already when the second car hit him, the other perpetrator shot the body after he was dead, the child had no injuries before entering the hospital, etc.

Abrasions may also change character with increasing time. For example, initial examination of a body retrieved from water may reveal no or much less obvious injuries due to the moisture from the water at the skin's surface. As the body is stored in the morgue overnight and allowed to dry, the abrasions will darken and may become much more apparent.

Contusions

Contusions are simply bruises. These are produced following an impact where the soft tissues and blood vessels, underneath or within the skin, are torn and produce hemorrhage. Grabbing an arm tightly can produce bruising with minimal impact and greater crushing force. One needs a beating heart and blood pressure to produce a contusion. A postmortem impact may accumulate blood due to lividity in a gravity-dependent area. This type of postmortem artifact may be difficult to differentiate from an antemortem injury.

One must exercise great care when dating contusions at the time of autopsy with gross findings alone. One must account for skin color, whether the contusion is deep or superficial, the presence of hematoma, etc. Bruises go through various color changes with advancing time as the body reacts to repair the injury. This depends on the size of the injury, the physiologic state of the individual, including his or her immune and coagulation systems, the vascular efficiency adjacent to the damaged tissue, etc. Color changes range from light blue-red to dark purple then green to yellow-brown as time progresses. Inaccurate interpretations of dating contusions grossly can create significant problems in court. If one sees a variability with color ranges demonstrating some bruises as red-purple and others as yellow-brown, it is reasonable to say that some injuries are older than others. It is much more accurate to use histopathology to date the injuries using established time ranges for advancing inflammation and healing. Random sections of injuries should be taken including the wound margins. The first several hours reveal hemorrhage with no inflammation. This is followed by iron deposition as a result of hemoglobin breakdown. Estimation of injury dating should be given in ranges.

Bruises can also change appearance as the postmortem time interval increases. A contusion will become more obvious as blood settles away from the impact site. The surrounding tissue will become paler and the contusion will be more pronounced, as in a supine body with anterior contusions in full livor. Bleeding associated with a contusion does not settle away from the impact site as lividity forms, because it is spread throughout the soft tissues and cannot drain within the vessels. Also, during vigorous emergency room resuscitation, the head and upper trunk may become congested. As lividity settles to the back of the head, injuries to the face become more prominent. Clinicians may sometimes miss these subtle contusions.

It is also important to realize that deep contusions may not be visible at external exam and may be visualized only after incisions are made. Clinicians should realize this and create a medical record that is as accurate as possible. If one documents that no bruises are present to regions of the body, yet they are found on autopsy, this may potentially weaken the case in court. The autopsy remains the most accurate way to demonstrate these findings.

Lacerations

Lacerations are tears of the skin, soft tissues, or internal organs or vessels as a result of an impact, overstretching, or crushing-type forces. These injuries are characterized by irregular margins, often with a marginal abrasion, with tissue bridging by fibrous strands and small blood vessels. If the laceration is large and gaping, tissue bridging may not be present due to the strands' being pulled apart. Skin lacerations tend to occur more often over hard surfaces, such as the scalp, knees, elbows, etc. The direction of the impact can be determined by the presence of soft tissue undermining. The underlying soft tissue adjacent to the laceration may form a pocket of separation extending in the direction of impact. A downward impact will produce undermining at the inferior aspect. These directions should be described with reference to standard anatomic planes.

Motor Vehicle Injuries

Motor vehicles include any motorized means of carrying or transporting someone. These include trucks, buses, cars, motorcycles, mopeds, snow mobiles, etc. It is always important to be as accurate as possible, including the *type of vehicle* on the death certificate for vital records. The type of vehicle is obviously important when evaluating injuries.

It is important to recognize various patterns that might help *differentiate drivers from passengers.* In high-speed collisions with unrestrained occupants, people may be ejected from the vehicle. If criminal or civil charges are filed, the living driver may indicate that the dead passenger was driving.

One should look for steering wheel impact marks to the chest, seatbelt-related abraded contusions, and pattern injuries associated with impacts to the windshield, dashboard, or vehicle roof. *Front and back windshields* are often made of laminated glass and fracture with elongated curves or splinters. *Side windows* are often made of tempered glass and fracture into small cubes. Seatbelt-patterned injuries or side impact dicing pattern injuries to the right or left side of the head will help formulate opinions about an individual's seating position. Dicing injuries to the left side of the head of an individual found next to a car with a broken left side window is evidence that he or she was the driver. One could also collect DNA samples from these regions and compare it with ejected occupants. Those not wearing seatbelts are much more likely to be ejected from the vehicle, especially in a rollover accident. Those who are ejected may impact other objects (such as a tree or pole) or may sustain crush injuries due to the vehicle's rolling over them.

Pedestrian clothing and impact sites to the body can reveal many clues with evidence about the circumstances of a collision. Questions one should ask include: Was the pedestrian standing or lying in the street? Was the individual run over or run under by the vehicle? Was more than one vehicle involved? Impact sites may reveal different front grill or tire pattern injuries. With standing pedestrian impacts, one should measure the distance between the impact site and the bottom of the foot or shoe the pedestrian was wearing at the time of the incident. It may be necessary to incise this region to visualize it better. It is good practice to include the measuring stick in the picture. This may give insight into whether the vehicle was braking before the impact occurred. As a driver jams on the brakes, the front end of a car will go downward. Fractures may occur to weight-bearing legs. Fracture patterns may be difficult to interpret when the bone is splintered and fractured into many pieces. Generally speaking, as a force passes through a bone, the fracture pattern extends outward from the impact site similar to a skull fracture from an entrance gunshot wound producing internal beveling. With posterior standing impact, one might find stretch marks at the inguinal region opposite the impact site. This is caused by hyperextension of the hip and leg in an anterior direction.

Slow-moving vehicles tend to run over people. Fast-moving vehicles tend to run under people, meaning after being struck, the victim is tossed onto and over the hood. Flaps of skin may be torn away as a tire passes over a body. The clothing may yield significant evidence such as paint fragments. The car may have fragments of blood and hair that can be used for DNA analysis. This can be useful when there are multiple pedestrians and cars involved in collisions. Many pedestrian collisions involve children, who are sometimes impulsive and careless and may run out into oncoming traffic; the elderly, may not be quick enough to get out of the way; or individuals who have psychological histories, such as the homeless, or intoxicated individuals, who may think it's not that dangerous to cross a busy freeway.

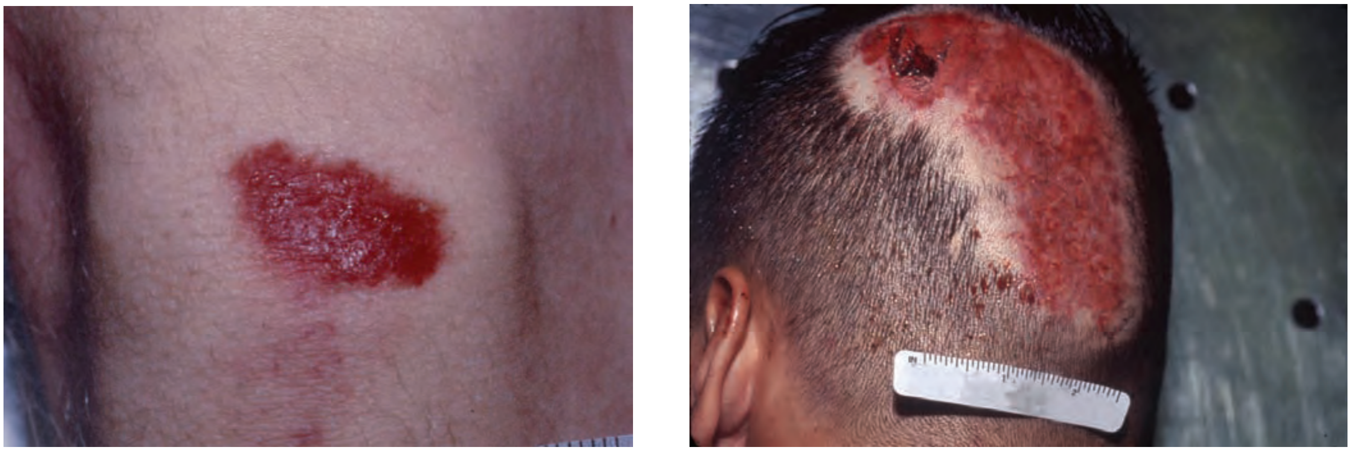

Recent antemortem abrasions.

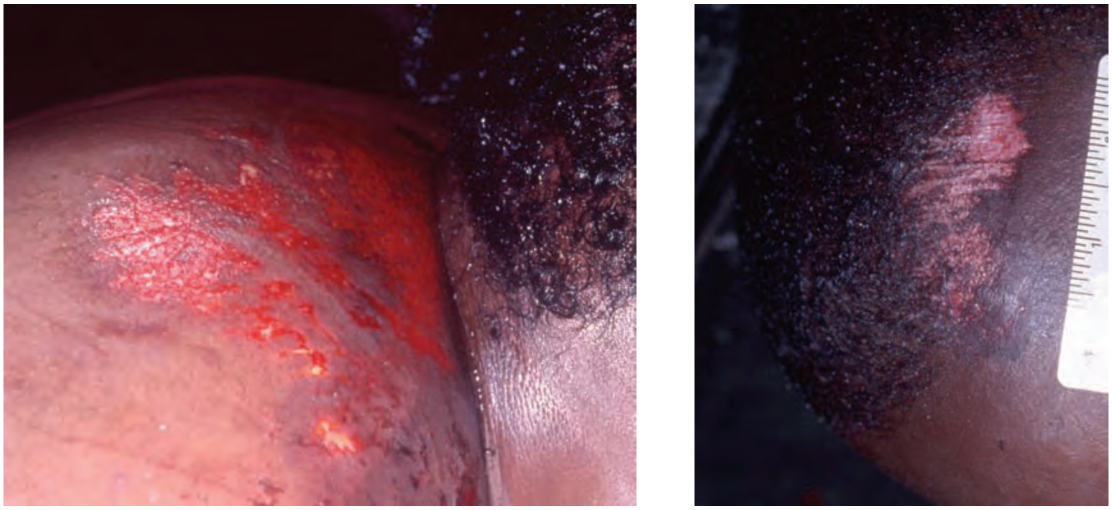

Recent deep antemortem abrasions.

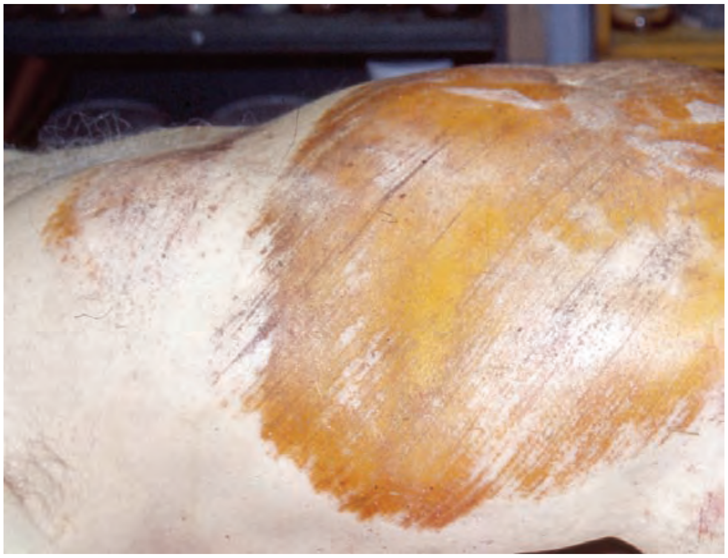

Large yellow anemic postmortem abrasion.

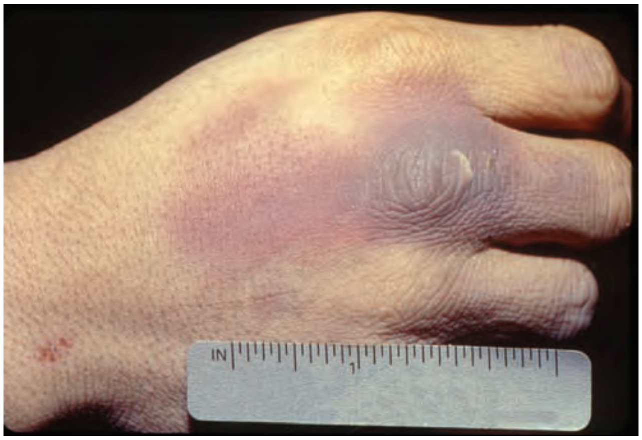

Recent contusion of the left hand associated with a fist fight. This individual punched the attacker before being murdered.

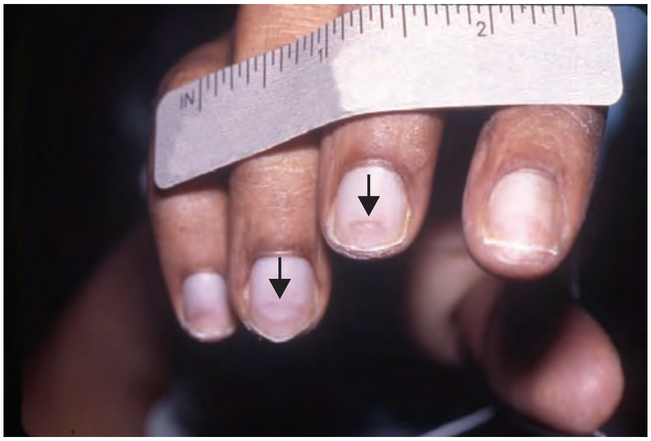

Damaged fingernails with nail bed contusions. Blunt-force injury during struggle.

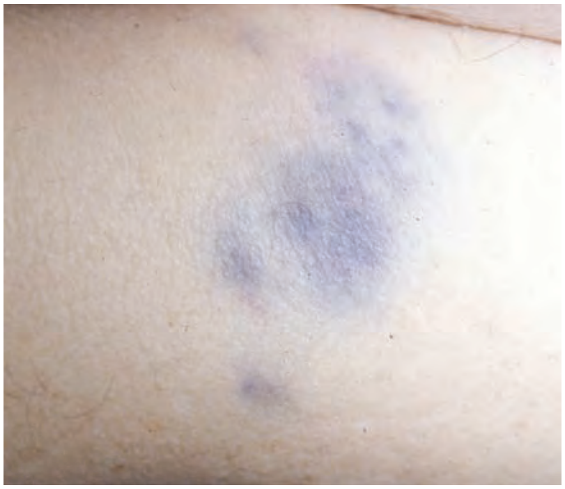

Recent purple contusion of arm.

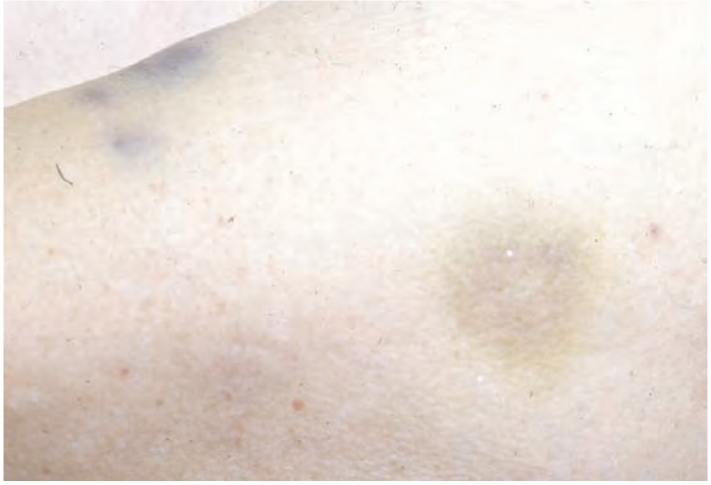

Older yellow-brown contusion.

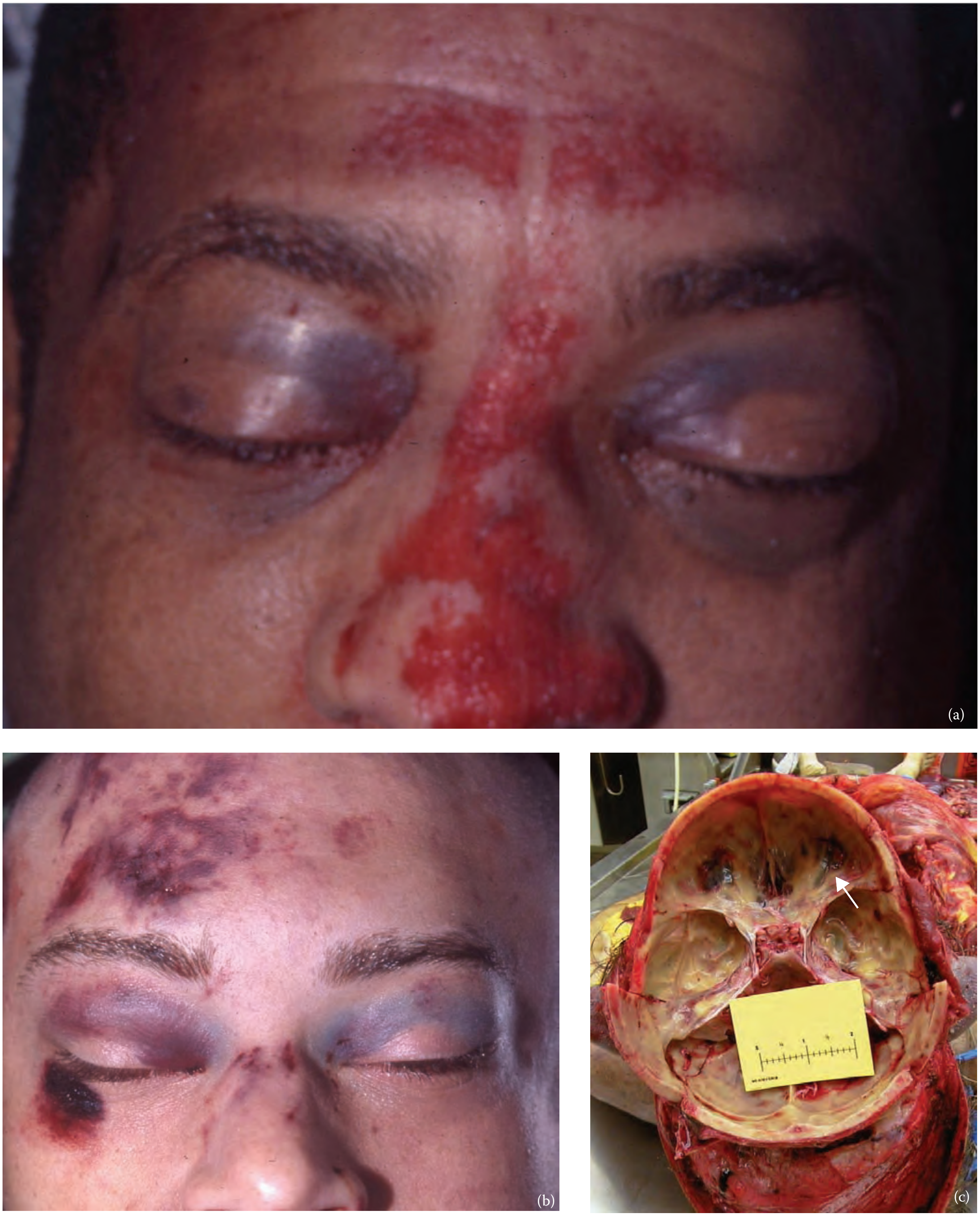

Multiple abrasions and contusions of face. Note the bilateral periorbital ecchymosis associated with fracture of the anterior cranial fossa. This is also known as "raccoon eyes." A similar injury is Battle's sign, hemorrhage behind the ear, also a sign of basilar skull fractures. Figure (c) demonstrates a basilar skull fracture.

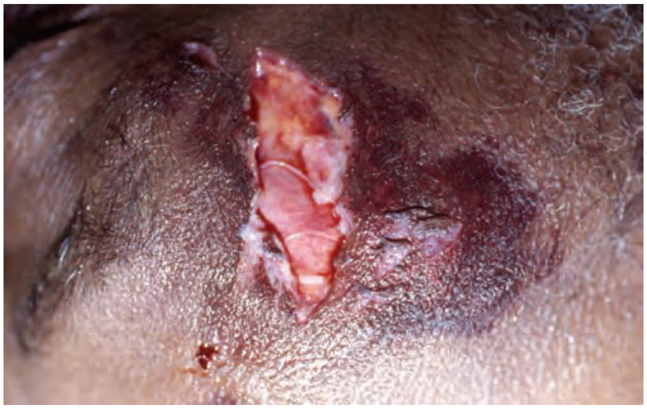

Recent laceration with adjacent abrasion. Note the irregular margin at the point of skin separation due to the skin's ripping or tearing apart.

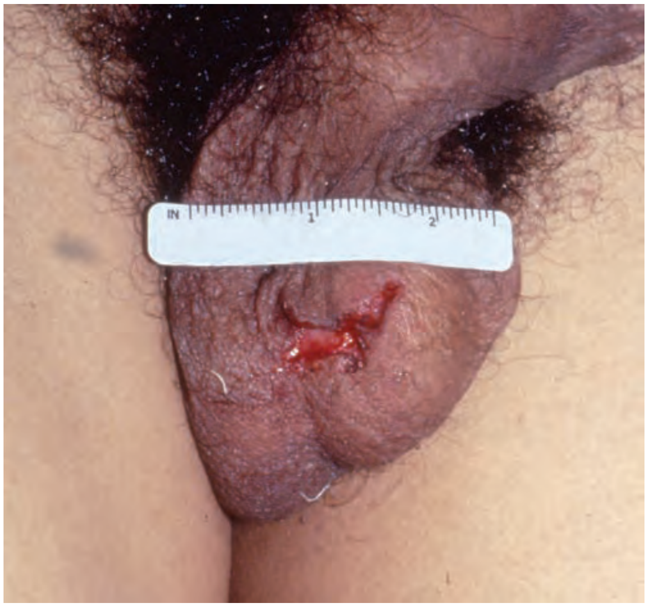

Recent laceration of scrotum.

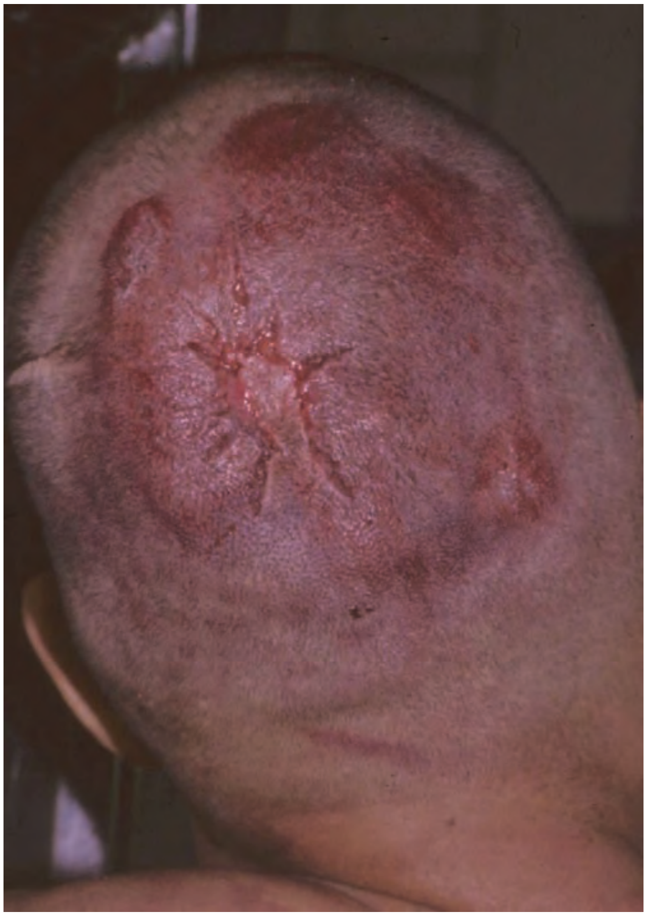

Large recent abraded contusion at the back of the head with a stellate laceration. This individual was intoxicated at a party on a rooftop. He was taking a group photograph and while backing up accidentally fell three stories to the pavement below. He had a comminuted skull fracture.

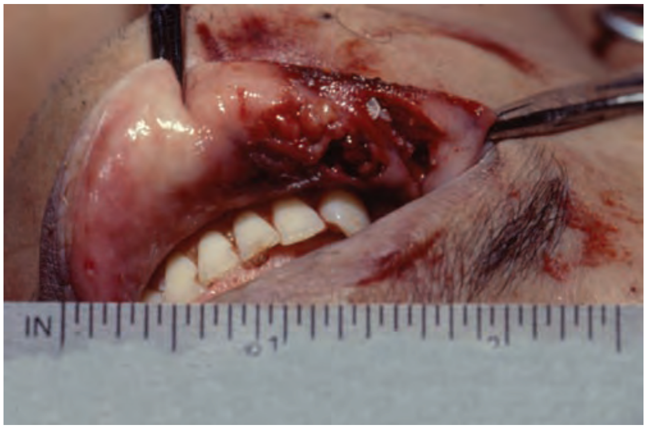

These are multiple recent lacerations and contusions at the inner aspect of the mouth following multiple blunt impacts to this individual's face with a fist.

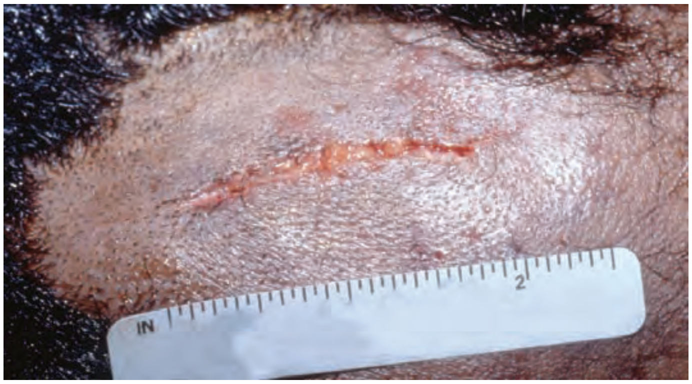

This is a postmortem anemic laceration in an individual who was stabbed to death in an apartment and later thrown out of a window.

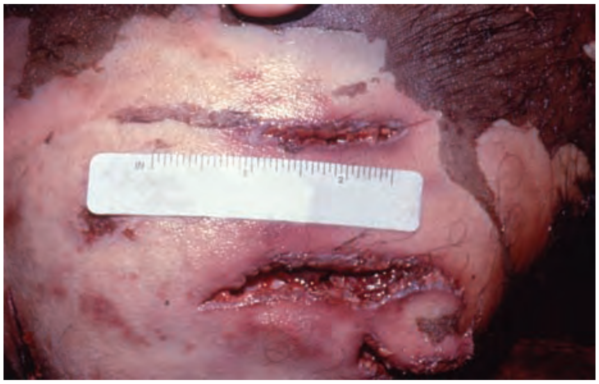

This decomposing body has skin slippage with separation of scalp hair. Note the scalp lacerations with slight to moderate decomposition.

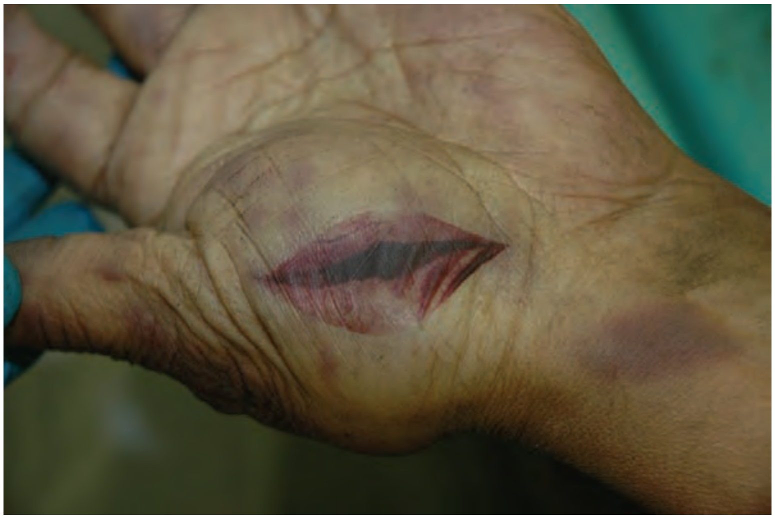

This individual was struck by a train. Note the soft tissue avulsion underneath the intact thick skin.

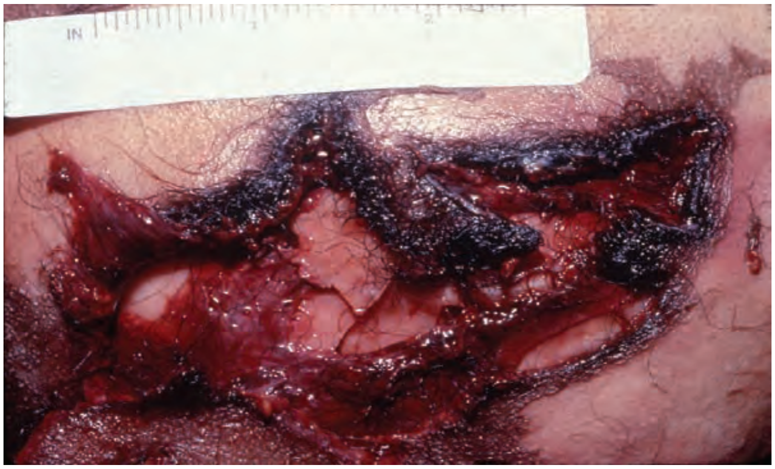

This individual died of multiple homicidal blunt- and sharp-force injuries. These are scalp lacerations in an individual with moderate decomposition. Note the extensive drying of the wound margins with clotted blood and fragments of hair. Also note the separation of the scalp from the skull with sub adjacent fractures.

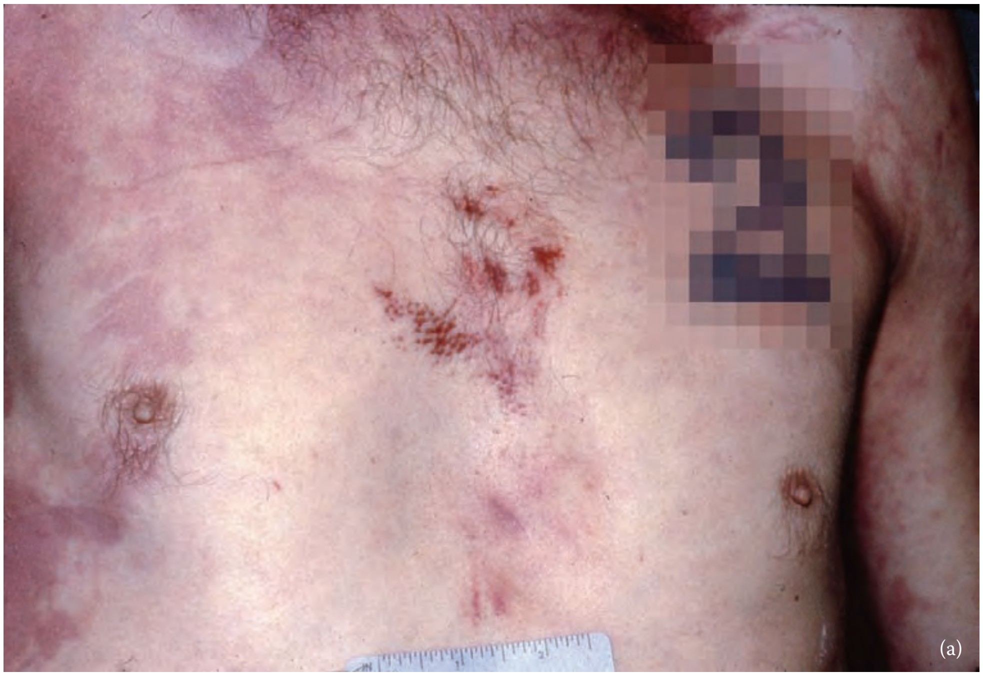

Note the recent abraded contusion to the middle aspect of the chest (a). This individual had multiple layers of clothing and was reportedly stomped on by an individual with heavy boots following assaults with other weapons. The abraded contusion to the middle aspect of his chest forms a rough outline of a boot (b). The multiple layers of clothing prevented a more discernable defined boot pattern. The individual had a fractured sternum and a cardiac contusion (c).

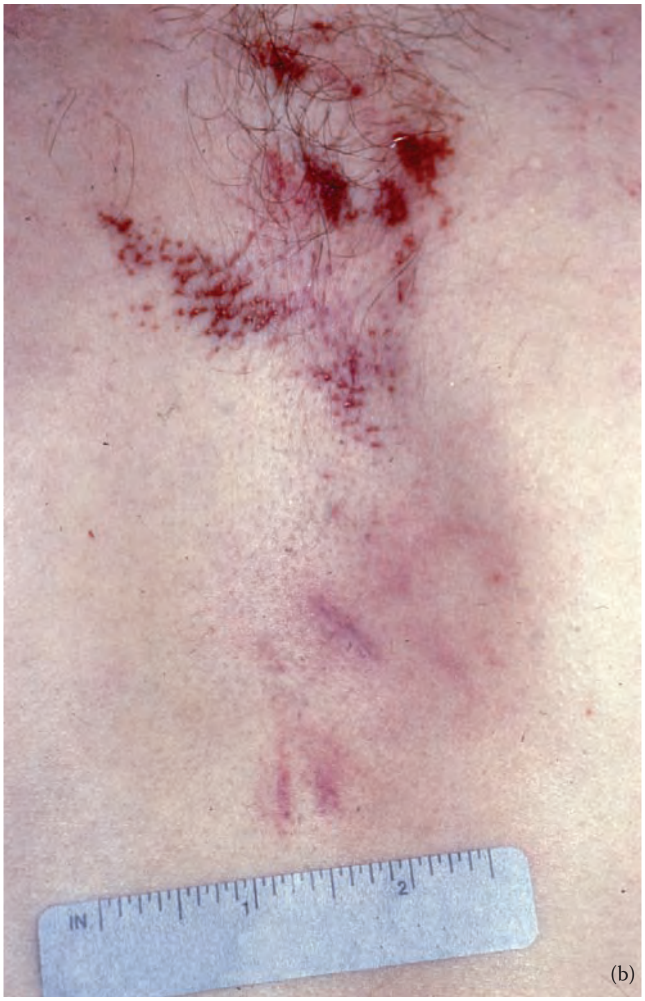

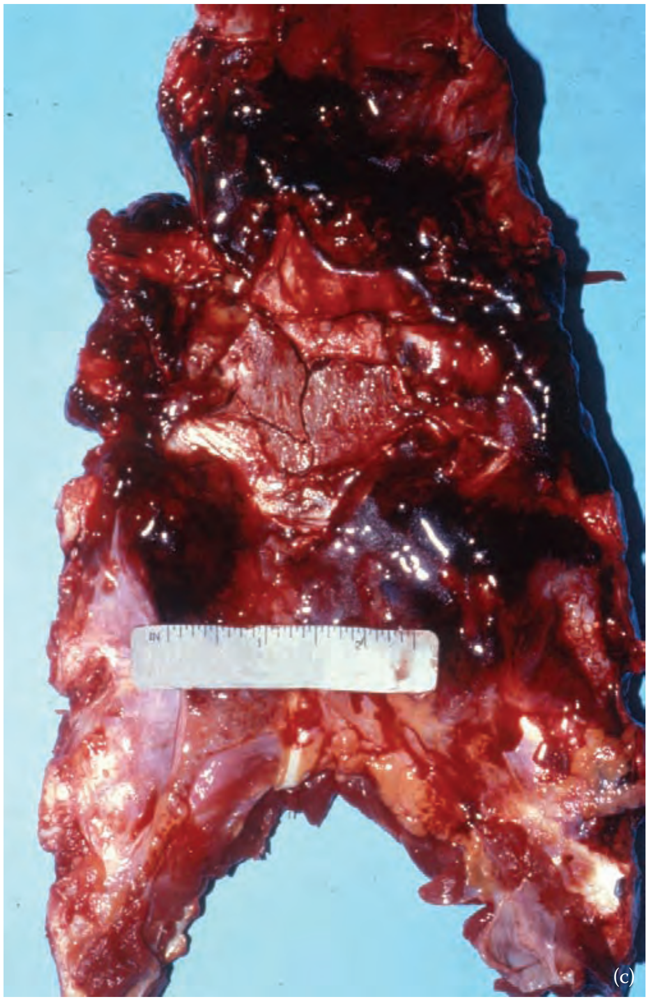

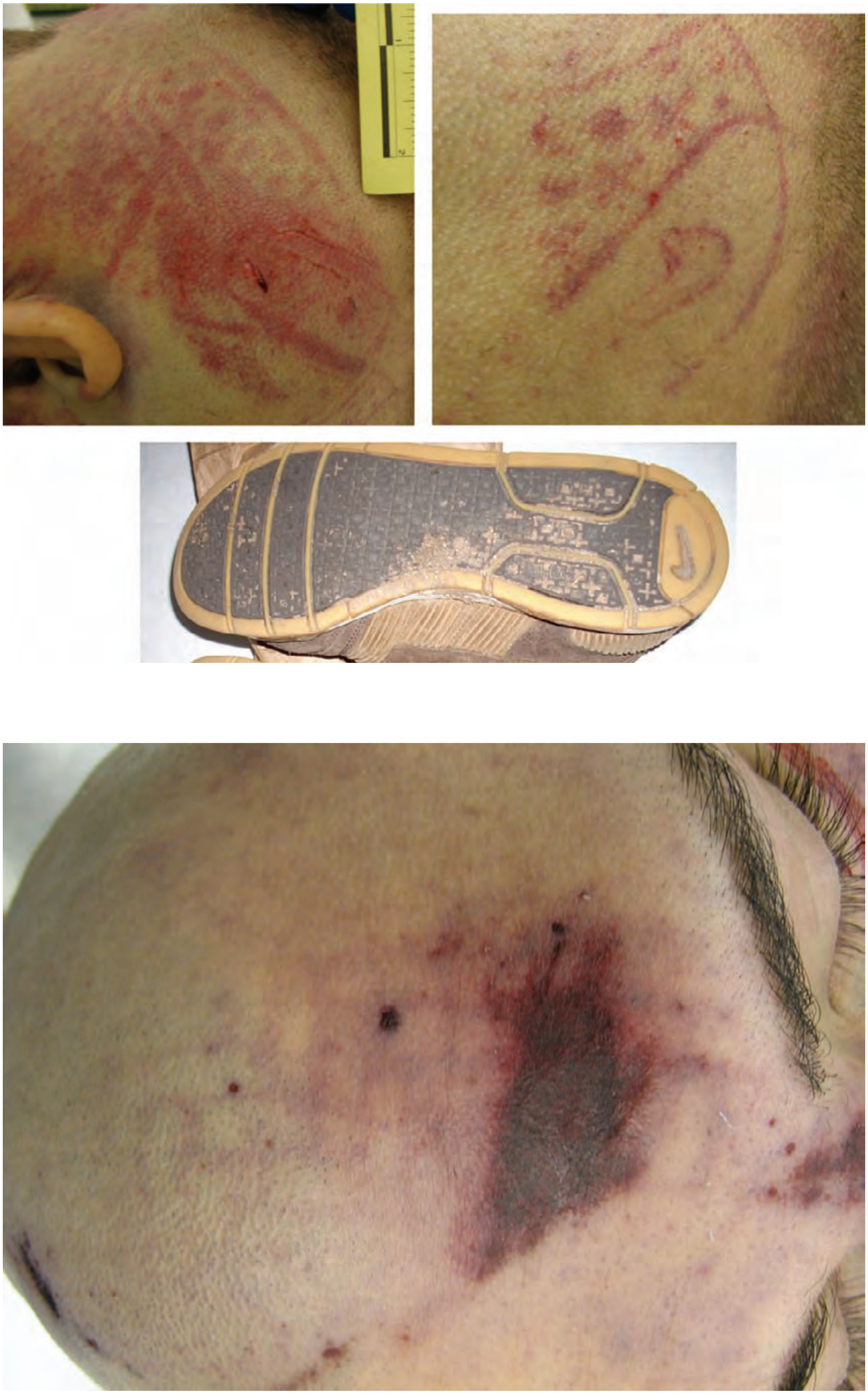

Pattern injuries associated with impacts from the bottom of a shoe while being stomped and kicked.

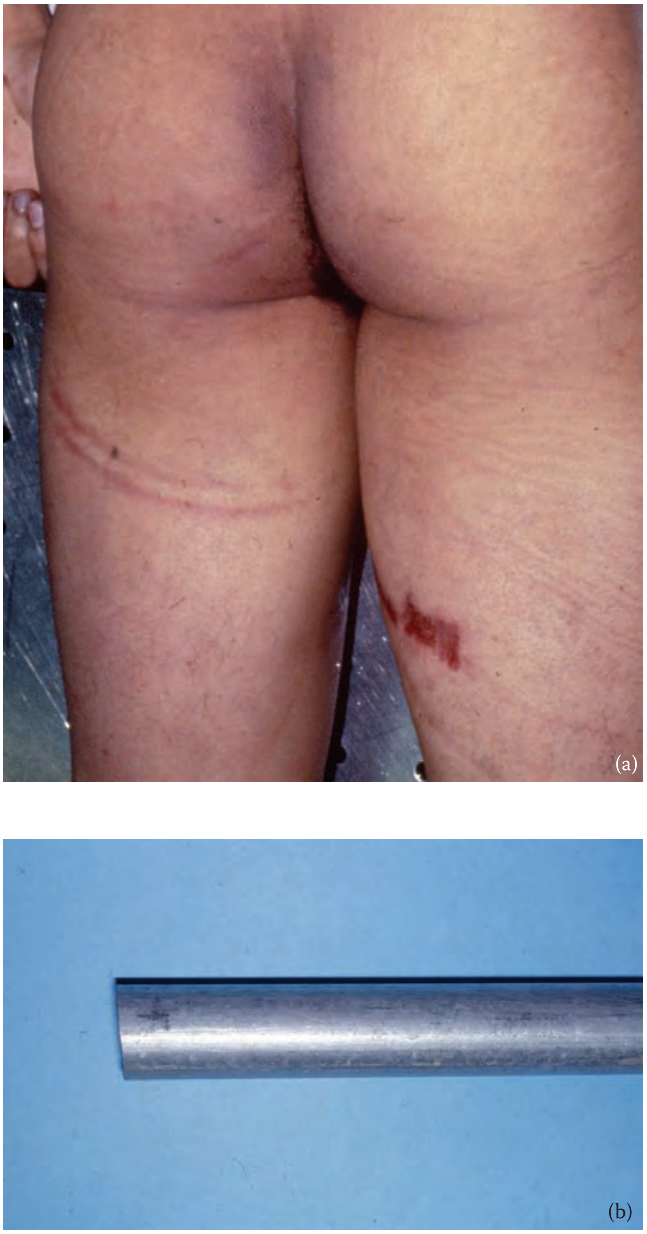

This individual was assaulted with fists and a metal pipe or rod. Note the parallel linear contusions at the posterior aspect of the left thigh (a). Depression of the soft tissues contacting the rod causes stretching at each margin with blood vessel injury and parallel linear bruises.

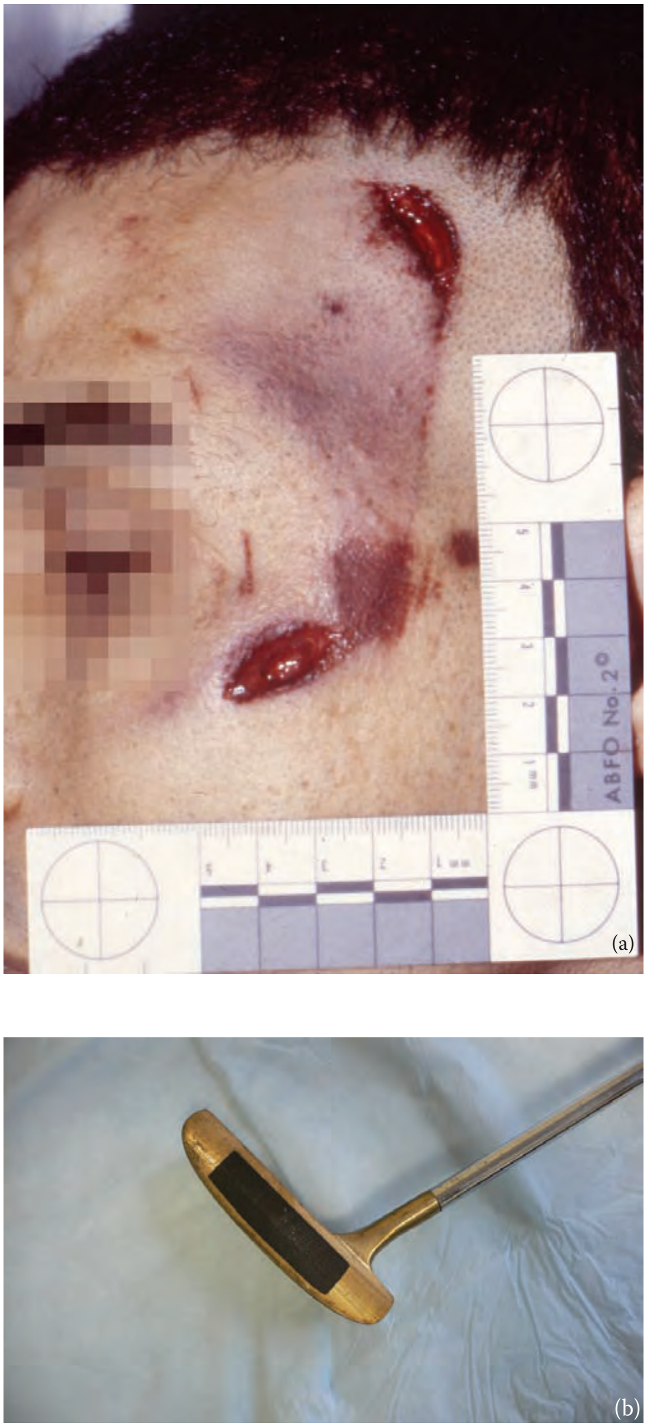

This individual was struck multiple times with a putter golf club. Note the pattern injury at the forehead and face, with the shaft of the club extending at the inferior aspect.

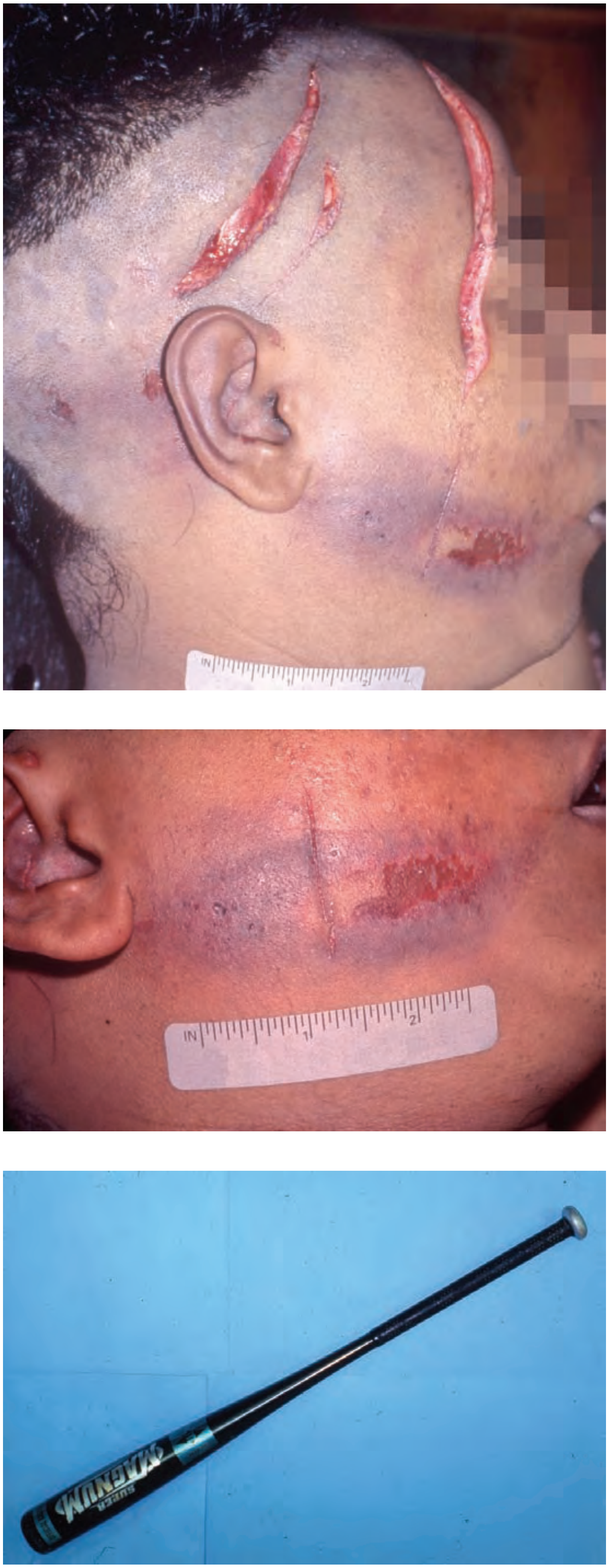

This individual was assaulted by multiple people with bottles and a baseball bat. These abraded contusions at his face and head are characteristic of a baseball bat impact. Note the oval contusion with sparing of the central aspect with overlying abrasion. Also note the abraded central region.

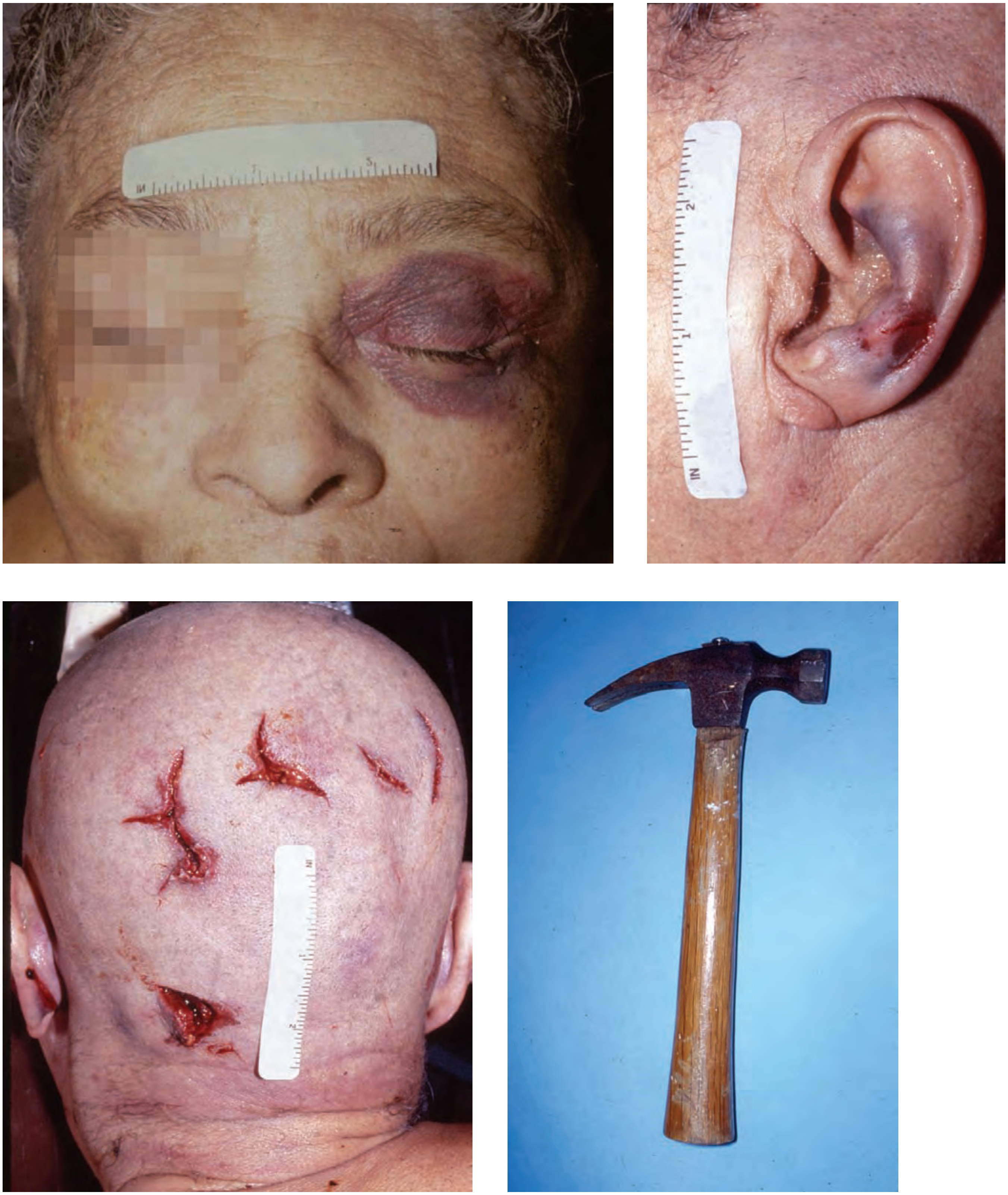

This elderly women had a history of an unsteady gait associated with Parkinson's disease and remote stroke. She was found in her apartment with several impact injuries to her scalp. It was initially thought by investigators that she had fallen several times and possibly suffered a heart attack. Further examination of her scalp revealed more lacerations and impacts that were initially not observed at the scene due to poor lighting and dried blood matted in her scalp hair. These scalp lacerations were produced by being struck with a hammer head.

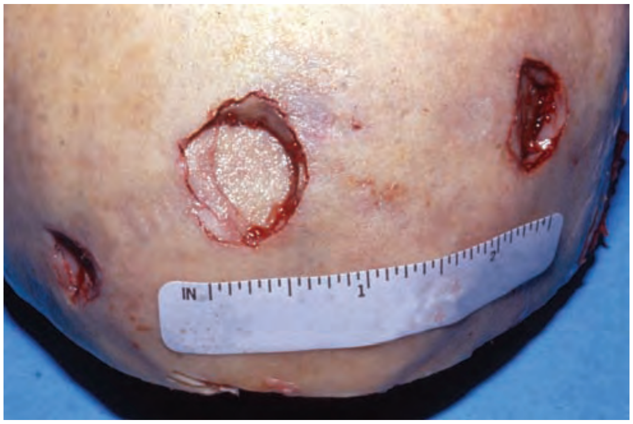

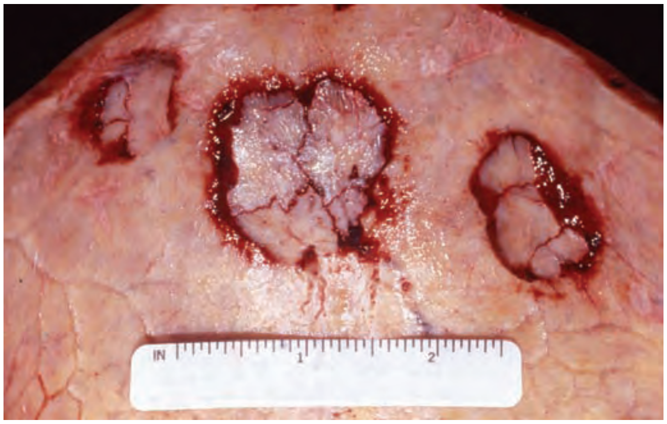

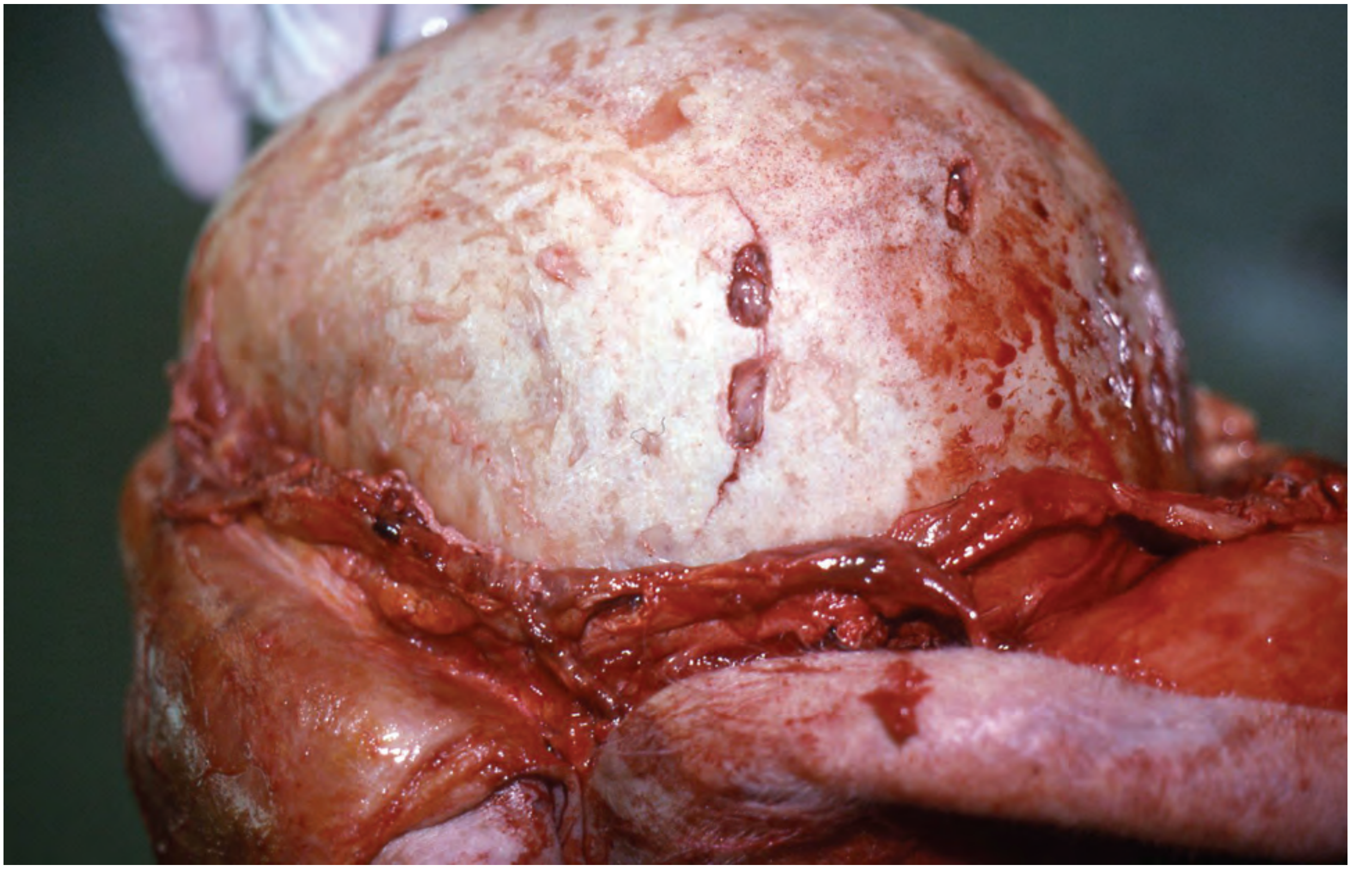

Depressed skull fractures due to multiple impacts with the head and claw of a hammer. Standard household hammer heads have a diameter of 3/4 to 1 inch, and the injuries on the skull tend to reflect this.

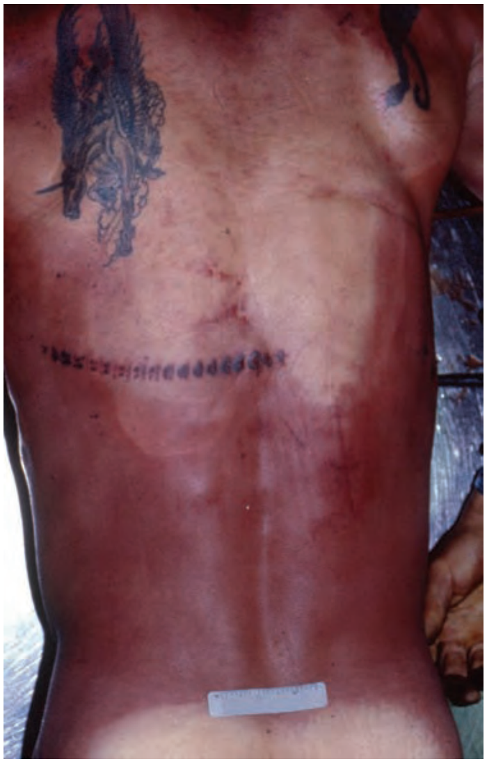

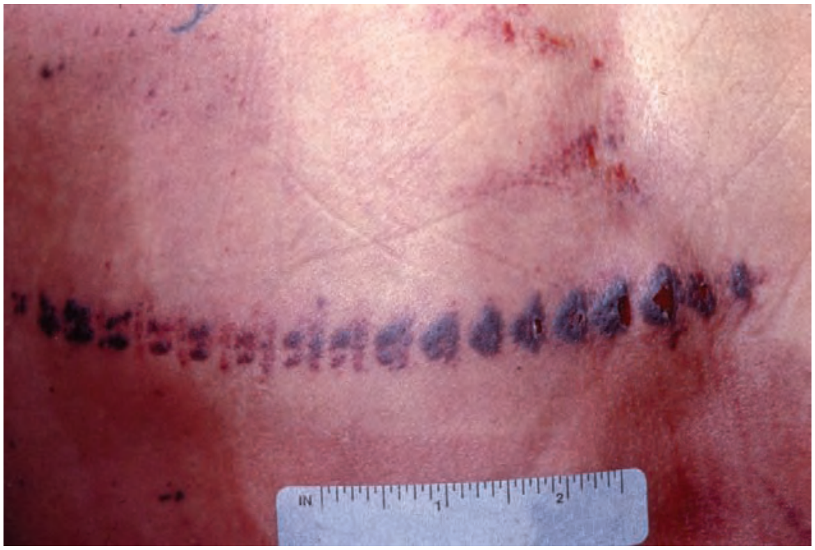

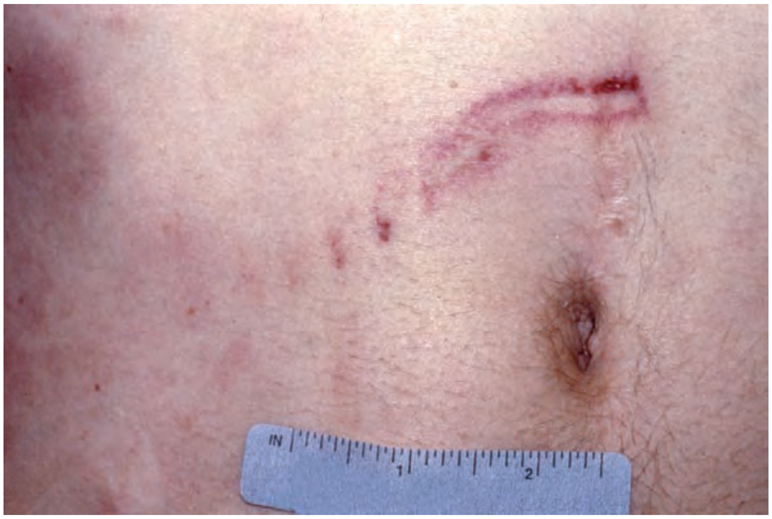

This individual was assaulted by multiple people including one with a bicycle chain. These injuries are characteristic for bicycle chain impacts.

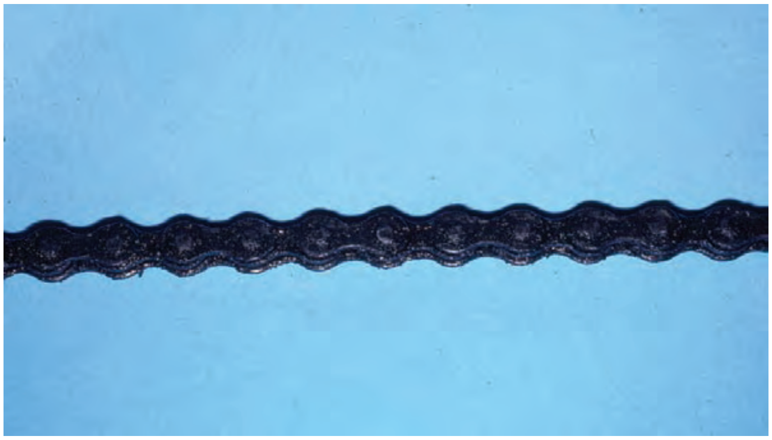

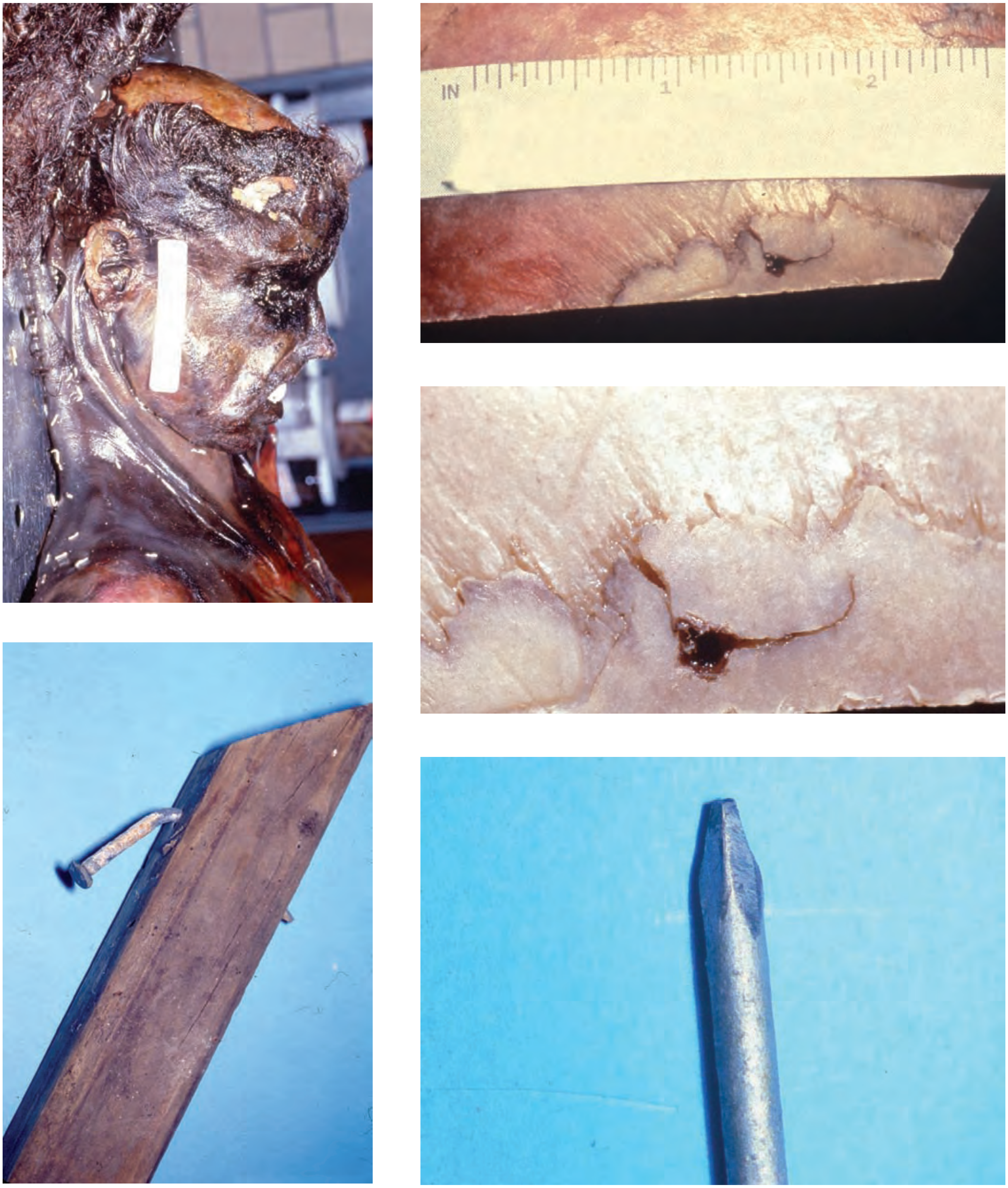

This individual was found naked from the waist down in an abandoned pipe yard. She had multiple defects to her scalp with exposed bone and extensive maggot and animal feeding. Careful inspection of the skull revealed a 1/4″ linear fracture of the superior temporal bone with a 1/16″ roughly square indentation. Further investigation of the scene revealed a board with a nail at the other side of the yard. Further testing confirmed this was the murder weapon.

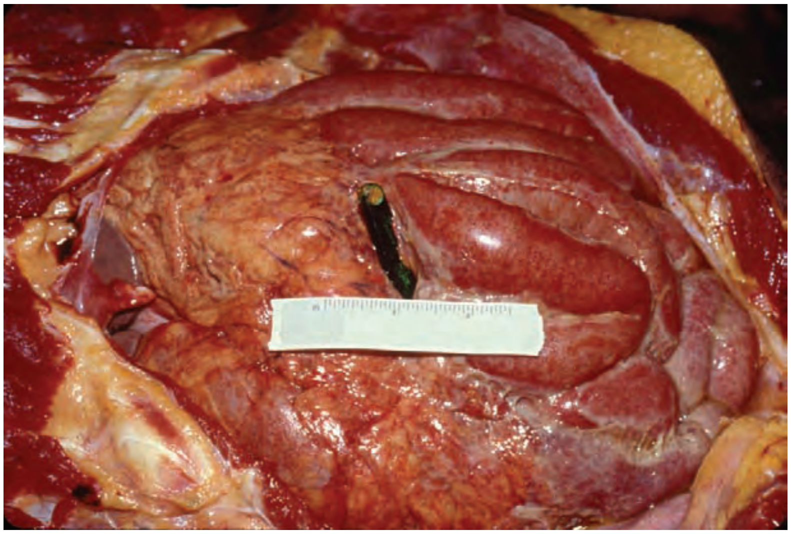

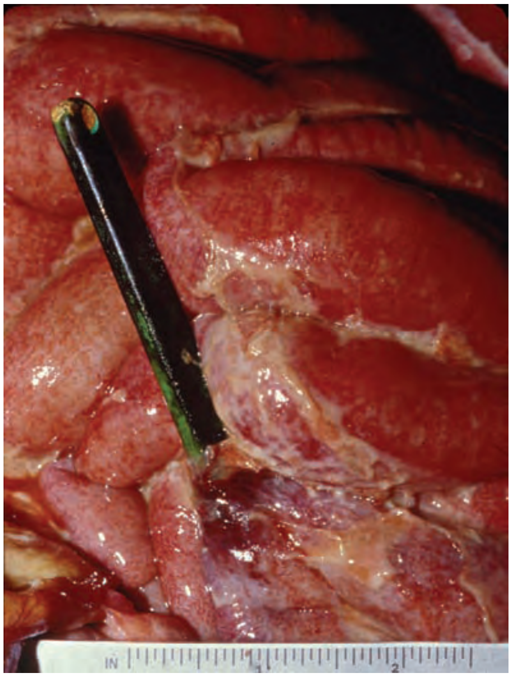

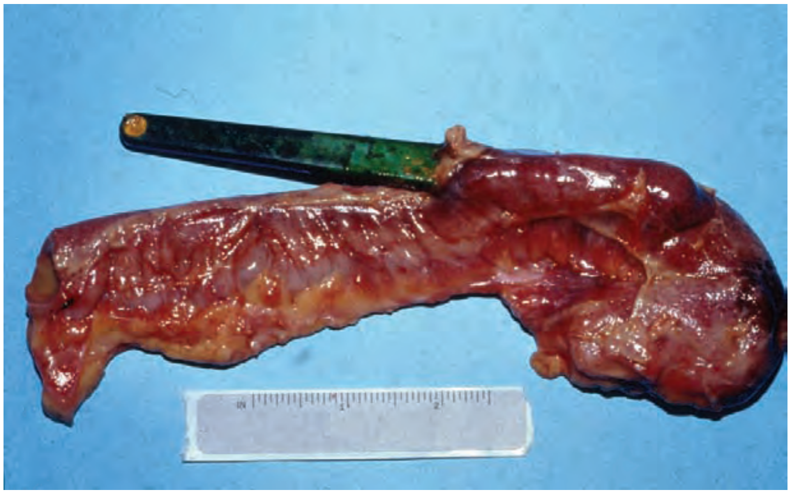

This psych patient had a history of swallowing miscellaneous objects. The blunt side of this toothbrush, in conjunction with peristalsis, eroded through the intestinal wall. This person died of peritonitis with septic complications.

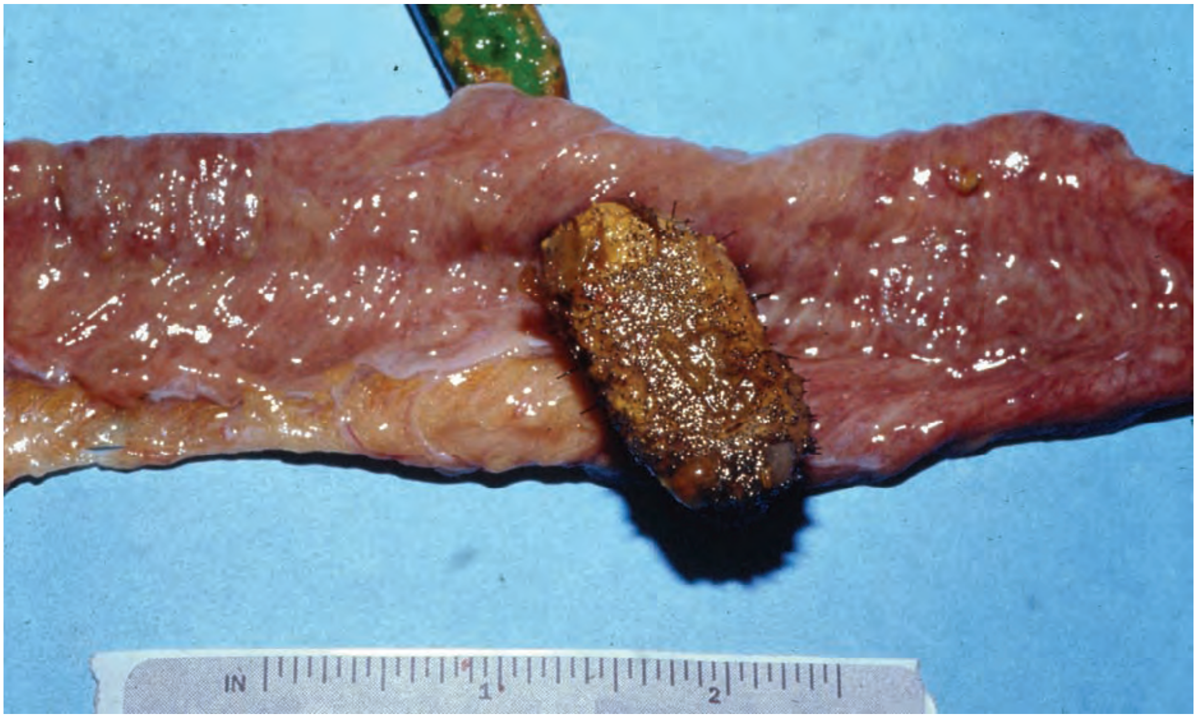

Pierced ear lobe with sociopathic pirate earring.

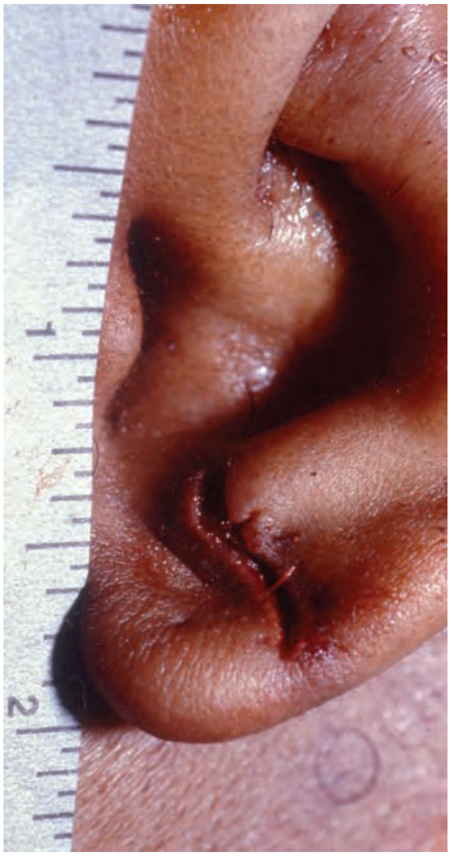

This is a homicide case where the individual had her earring torn from her ear lobe during the assault.

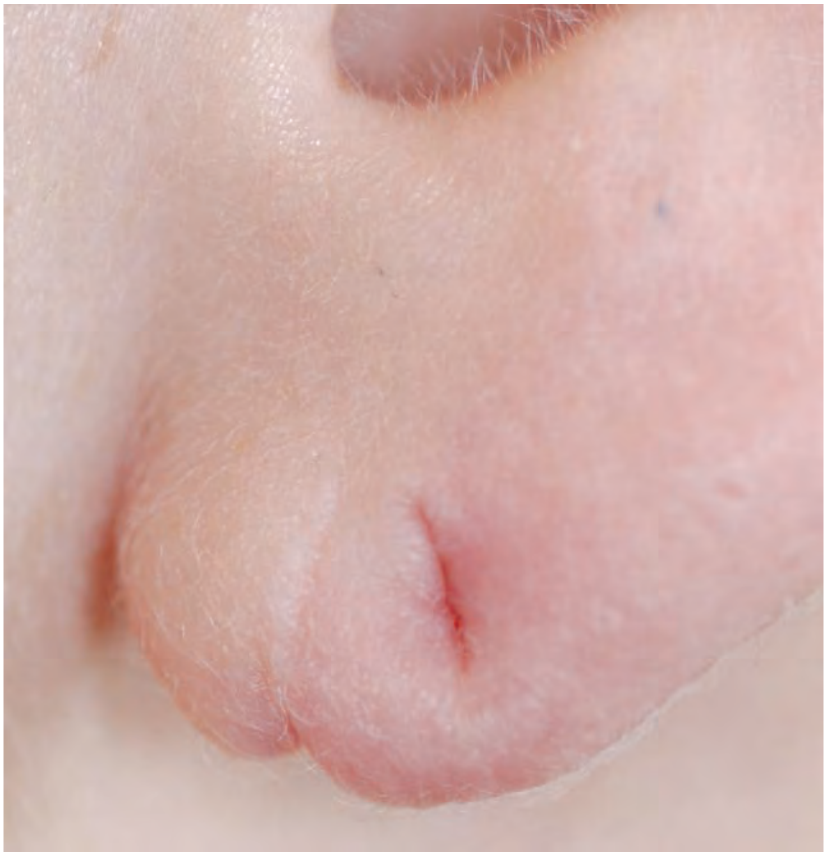

Note the slit elongation of the piercing site from wearing very heavy earrings. Also note the healed linear scar to the left due to a traumatic tearing of the earring from the ear lobe with complete separation and nonplastic-surgical repair.

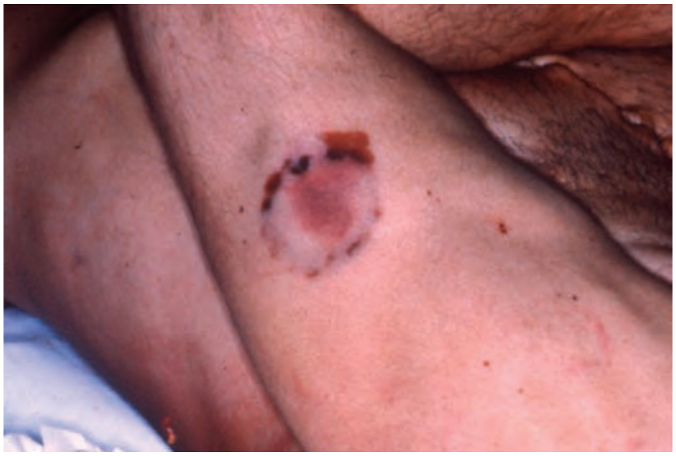

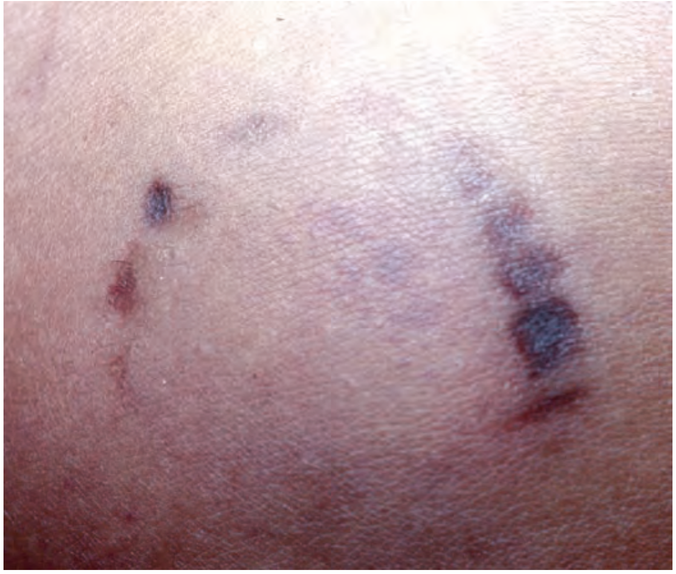

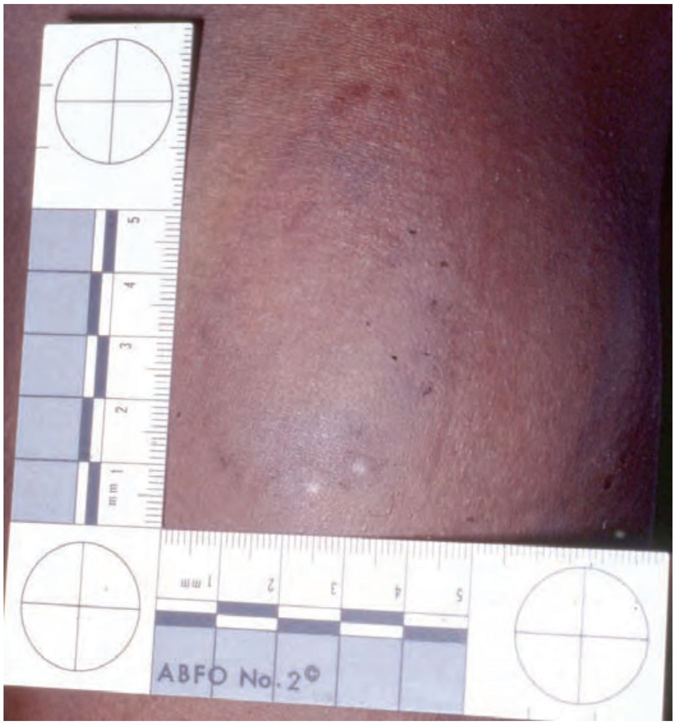

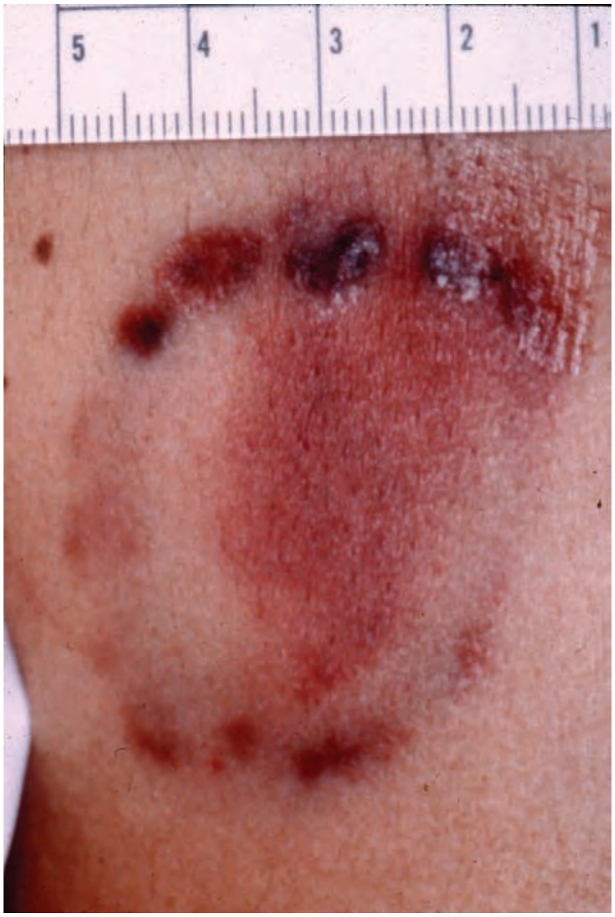

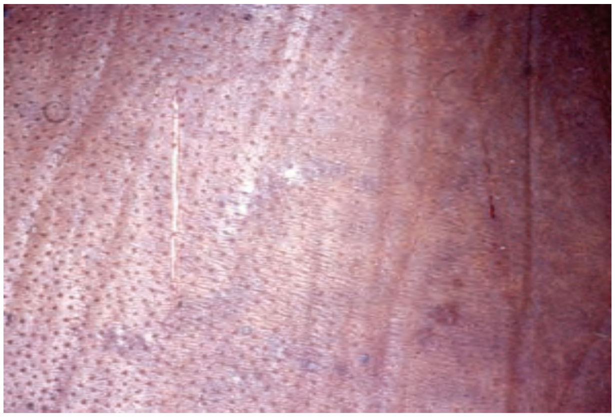

These are various examples of recent bite marks with different placement of the mouth while exerting pressure and different degrees of force applied. Note some patterns are very vague and others are extremely prominent. The typical example of a bite mark reveals a circular pattern with a central region of contusion. It is good practice to consult a forensic dentist as soon as possible whenever a bite mark is suspected. The injury should not be cleaned until swabs are taken to detect oral DNA left behind from the perpetrator. As time goes on with drying and decomposition, the injury may yield less valuable information for dental comparison and DNA analysis.

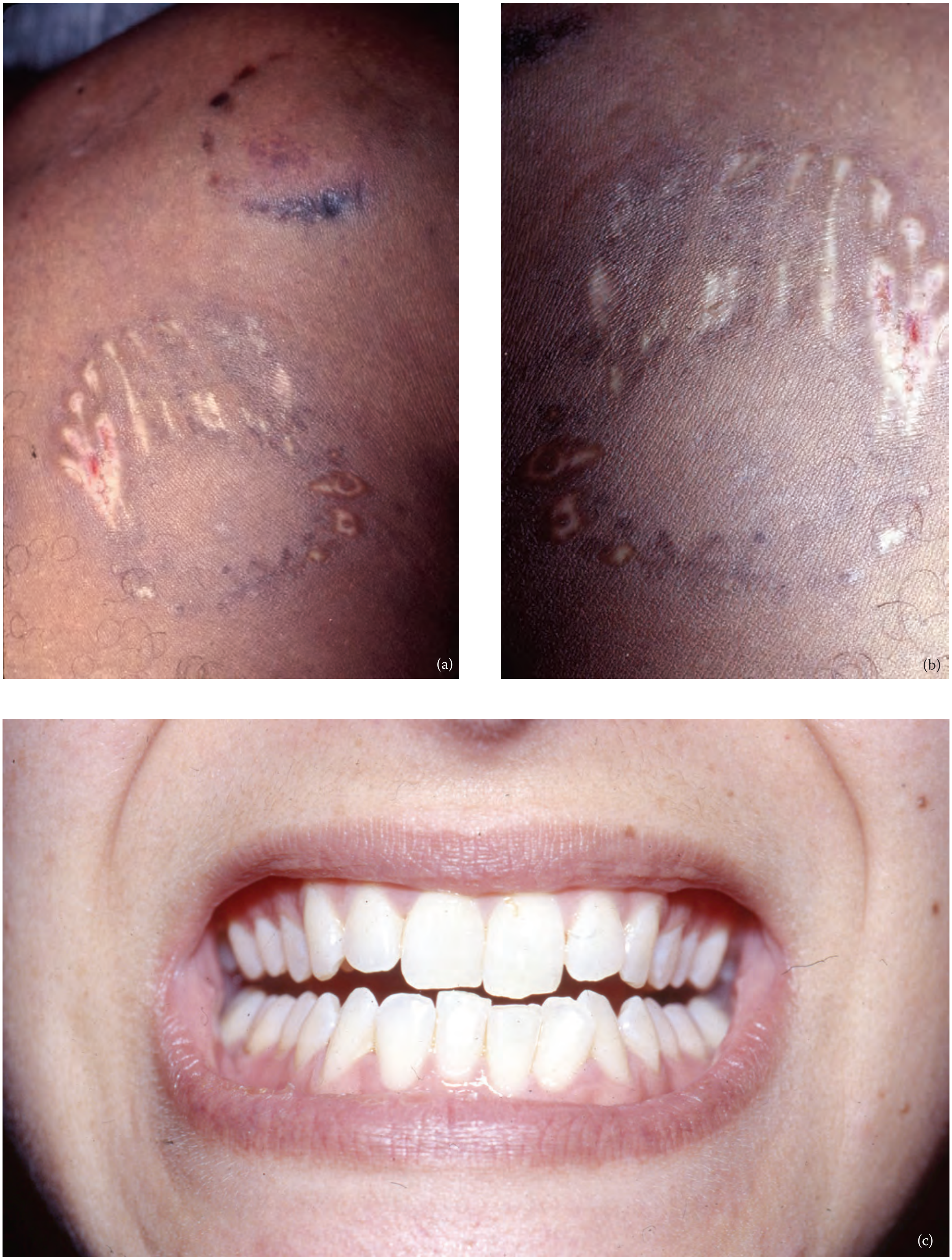

Image (a) demonstrates both a recent and a healed bite mark. The old bite mark is largely healed with hypo-pigmented deep abrasions from teeth dragging across the epidermal surface.

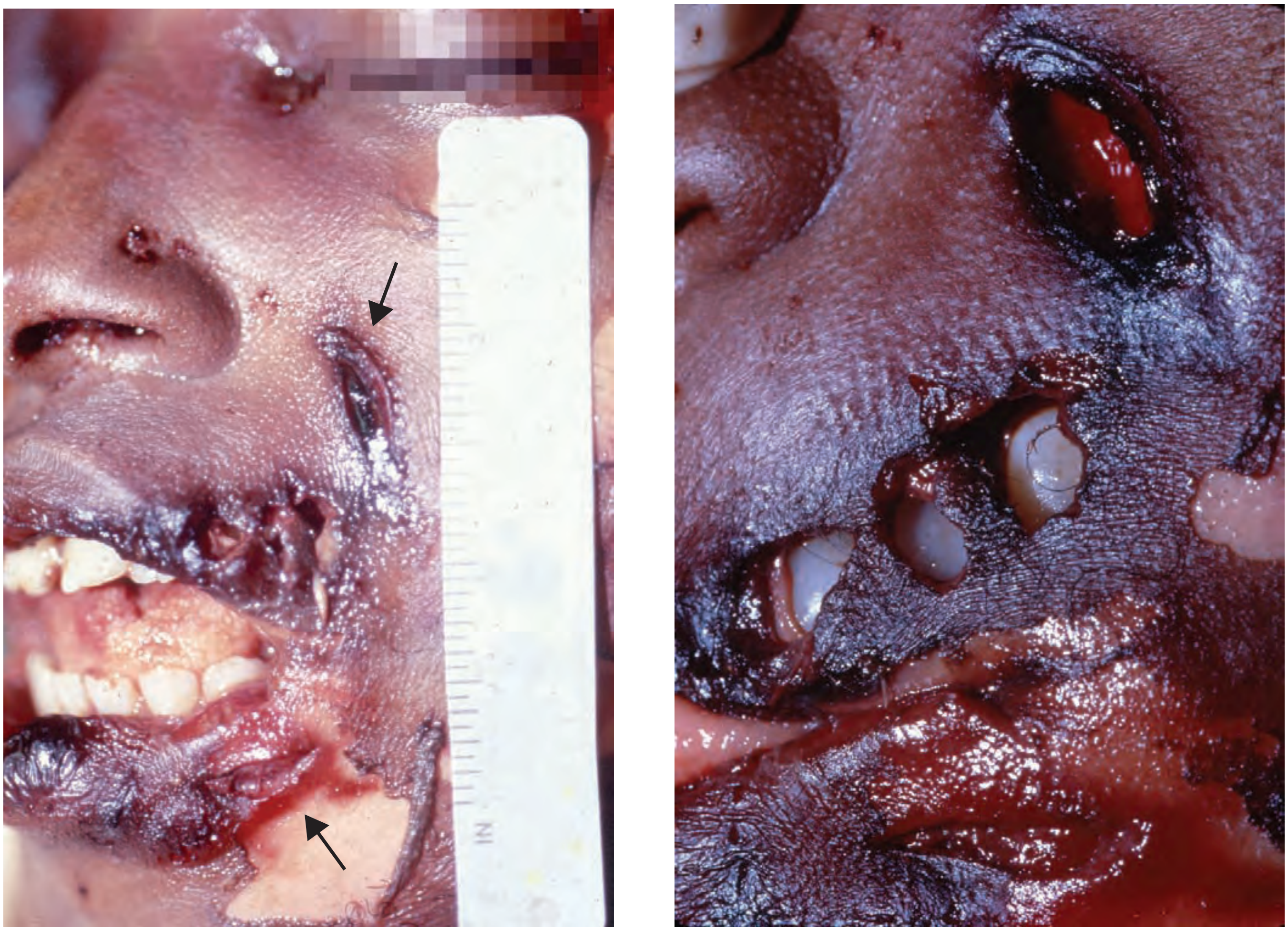

This individual and her roommate were found in their apartment with multiple blunt- and sharp-force injuries. Note the roughly semicircular lacerations at the superior and inferior aspect of the cheek, with the deeper lacerations of the lip revealing exposed underlying teeth. There was a large cylindrical storefront padlock within a tube sock found at the scene. There were multiple other pattern injuries to the decedent's body consistent with these roughly circular impacts.

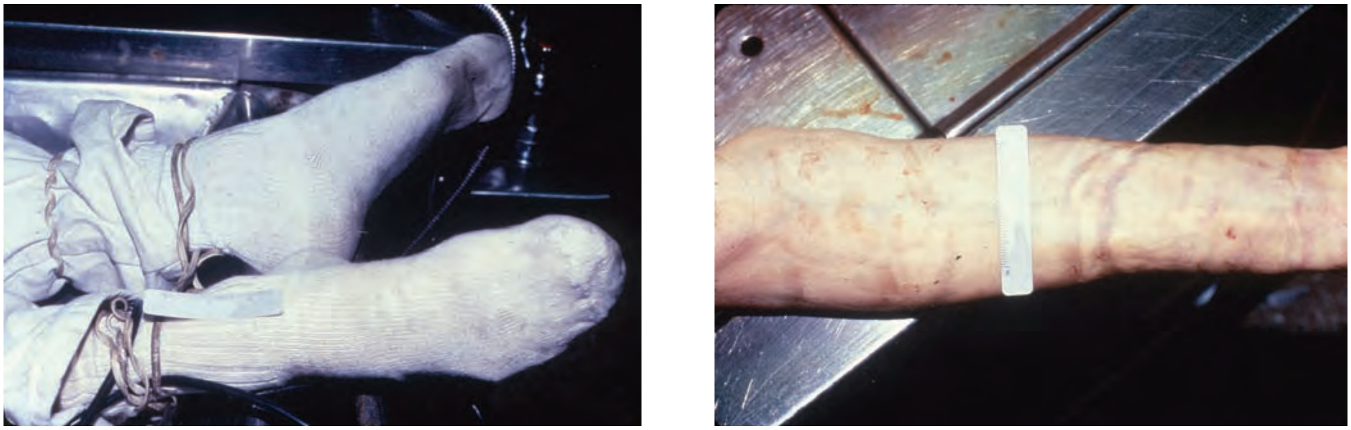

Note the furrow indentations associated with extremities bound by rope and cord prior to the victim's being stabbed to death.

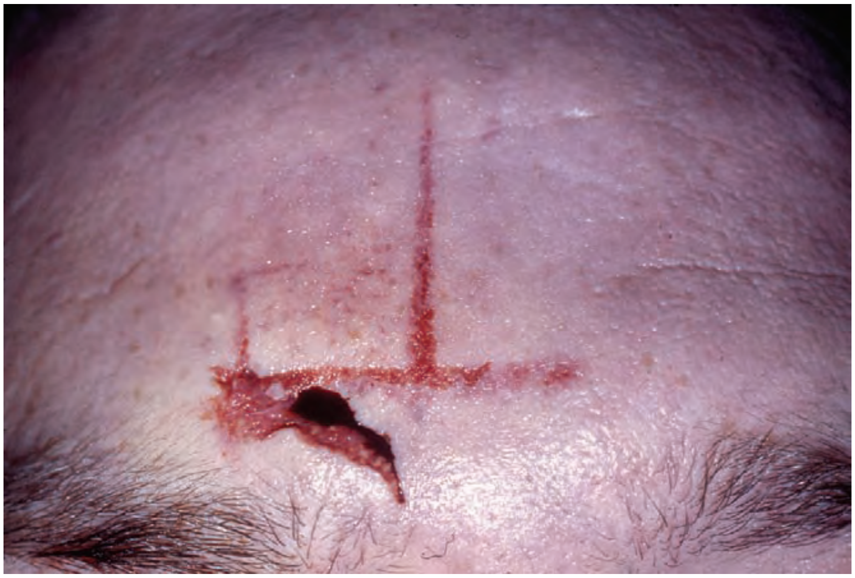

This individual suffered a witnessed fatal cardiac event while standing. While falling to the ground he sustained this pattern injury by striking his head on a radiator.

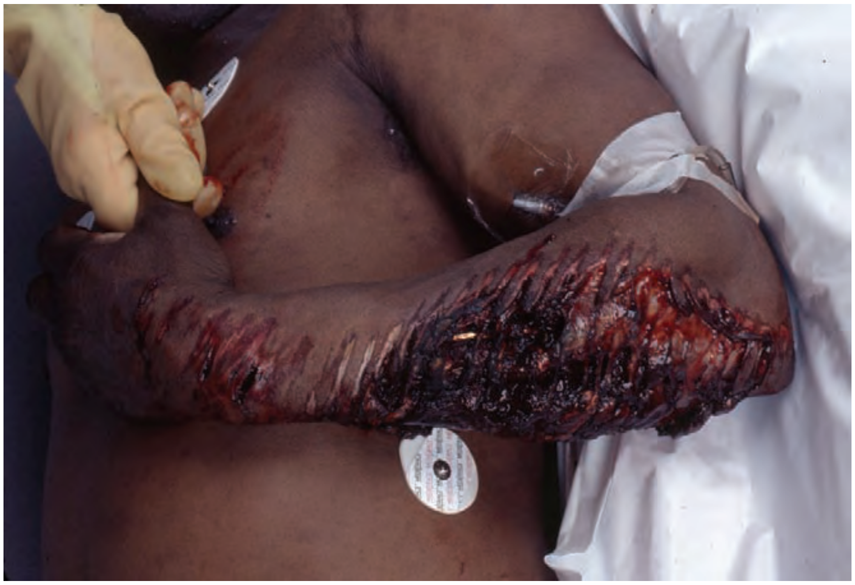

This individual suffered a fatal cardiac event while riding on an escalator. He was found lying at the bottom of the escalator. Note the pattern injury to his arm from the continued rubbing of escalator treads.

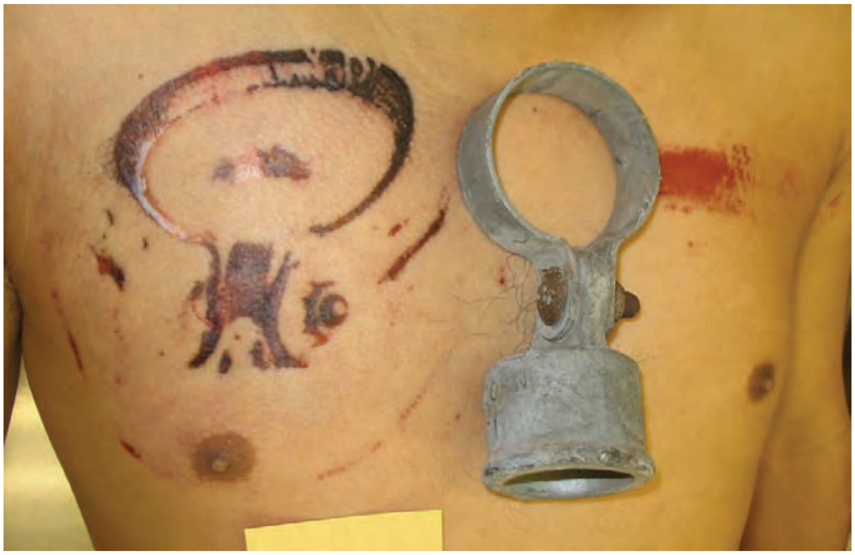

This individual impacted a fence at a high rate of speed, leaving a pattern injury from the top of the fencepost's contacting his chest.

This individual was an unrestrained passenger in a motor vehicle during a head-on collision. His face went through the front windshield. He sustained multiple curvilinear lacerations and sharp-force injuries from broken glass and impact with the car roof. This type of pattern is consistent with a front windshield impact.

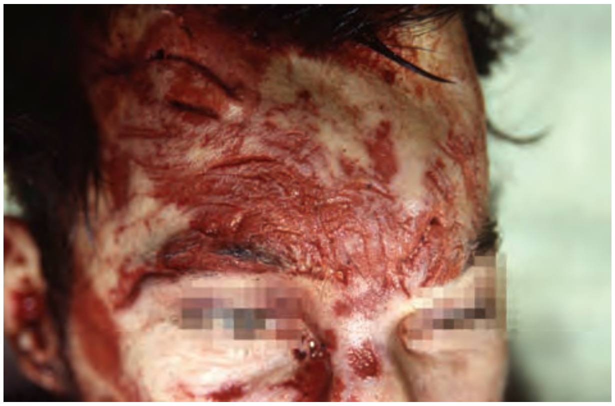

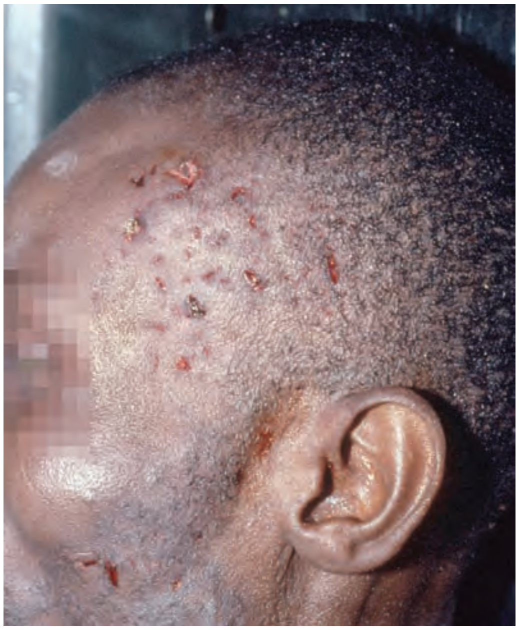

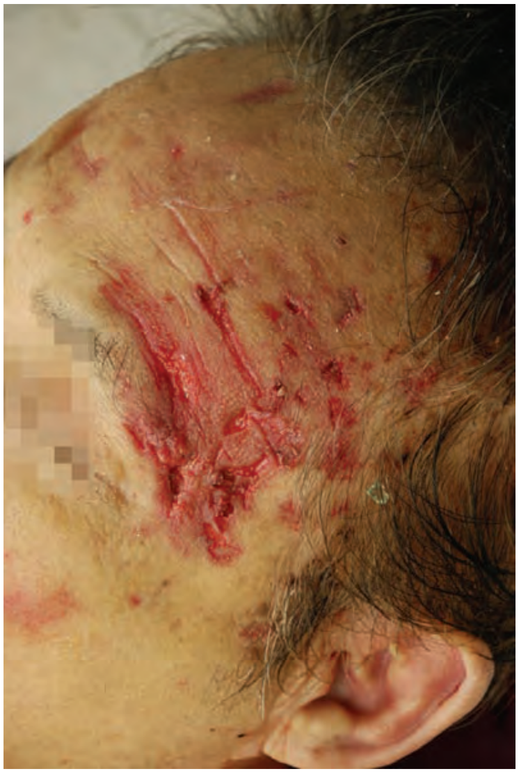

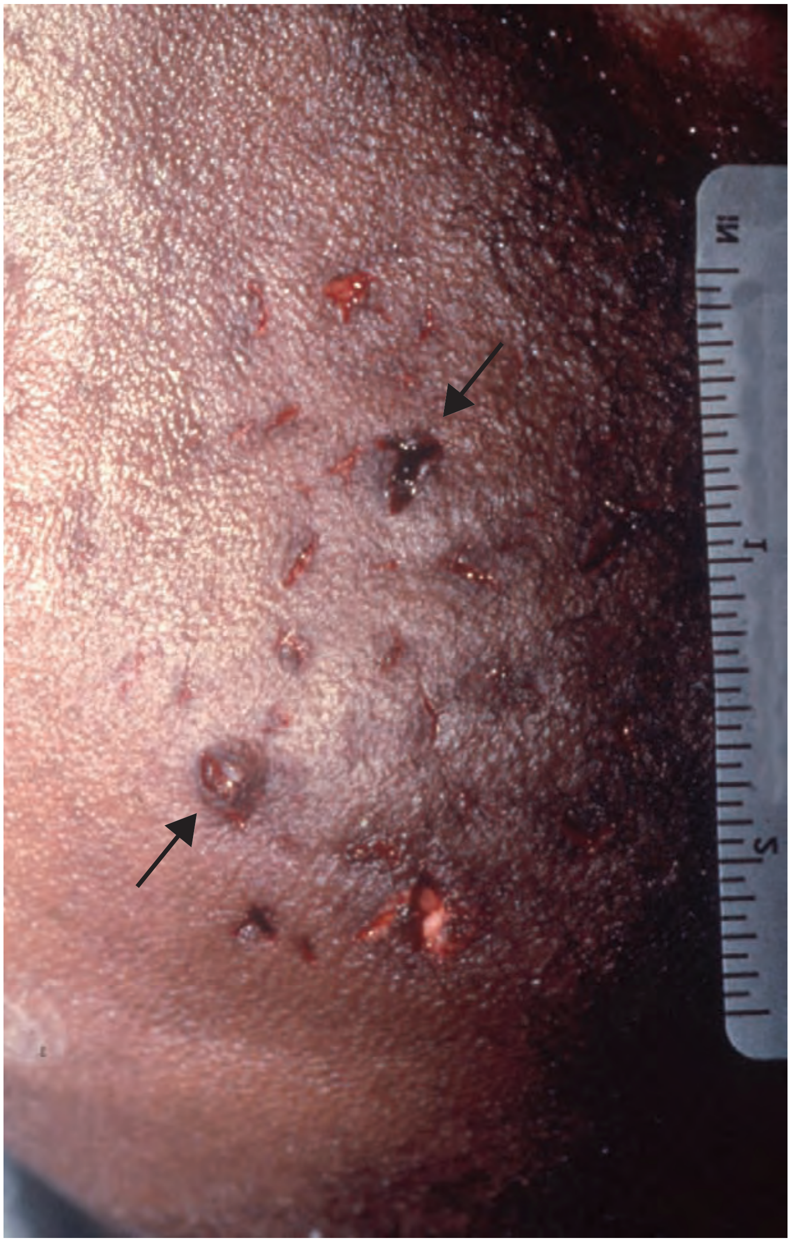

Note the dicing type injuries to the scalp with cubed fragments of glass embedded in the injuries. This type of injury is consistent with an impact and fracture of tempered glass, which is present in many side windows. These individuals were passengers with left-door-window head impact.

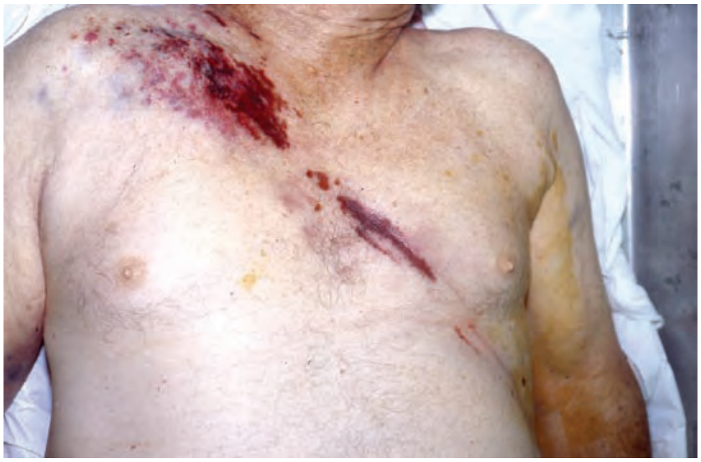

This individual was the passenger in a front seat collision. Note the abraded contusion from his seat belt.

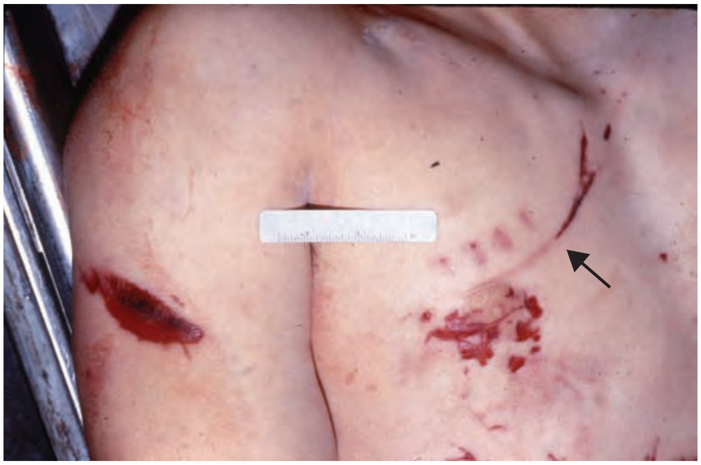

Note the abraded contusion at the chest and arm of this individual from a steering wheel impact. He was the driver of this car in a multiple-fatality vehicular accident.

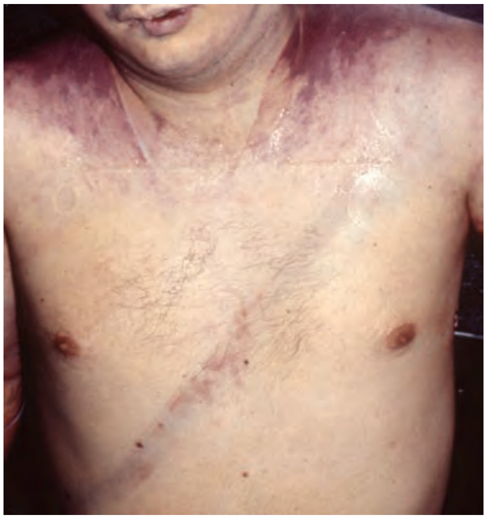

This was the driver of a car in a motor vehicle accident. Note the abraded contusion from the seat belt.

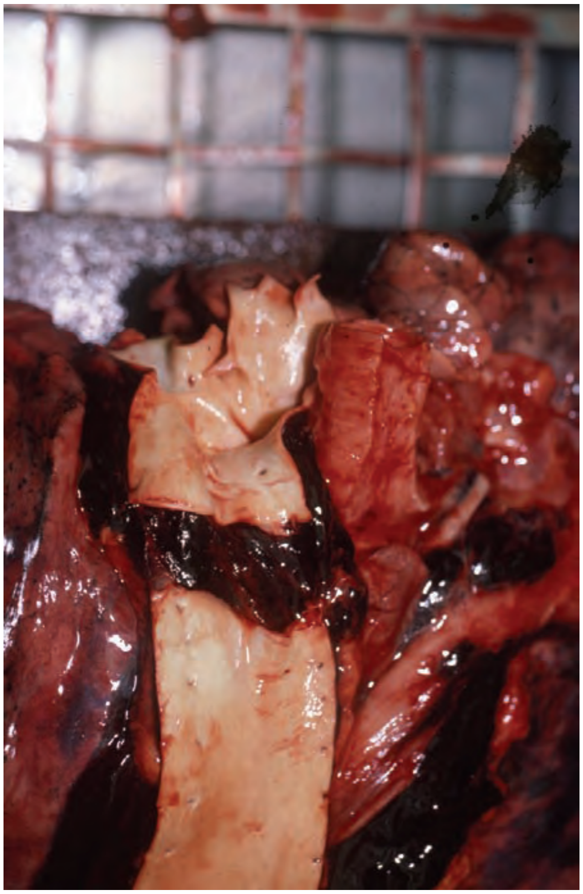

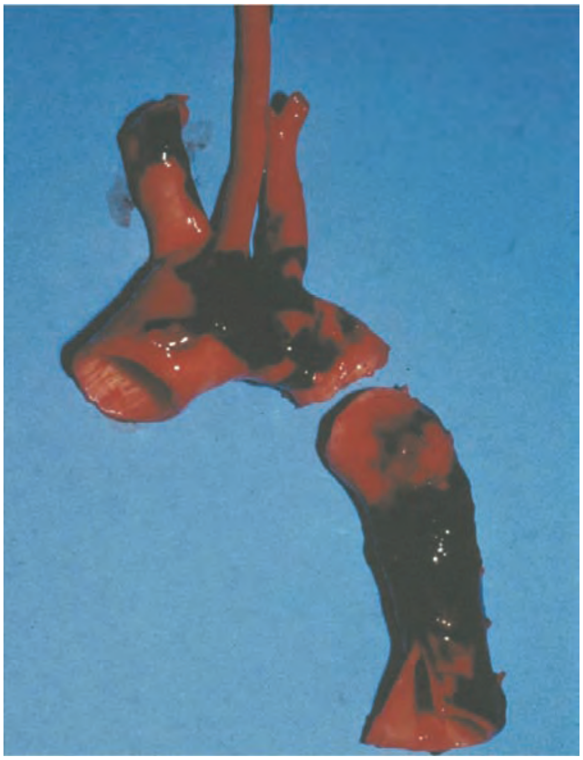

These demonstrate a transection of the aorta in the proximal descending region just distal to the ductus ligatosum. This is a common location of laceration.

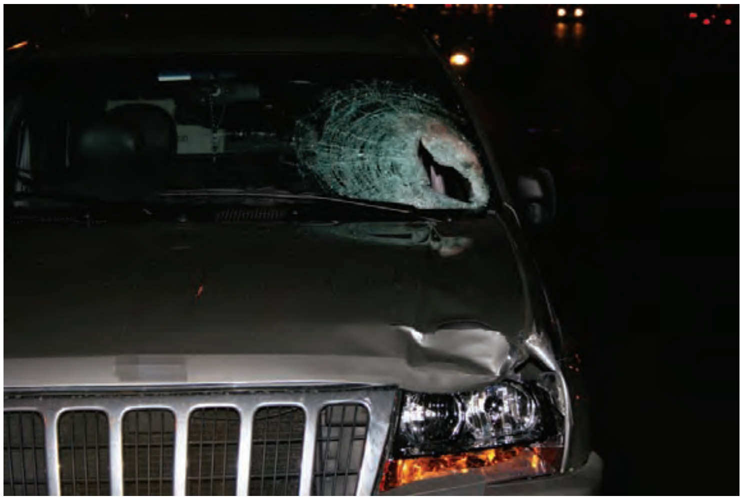

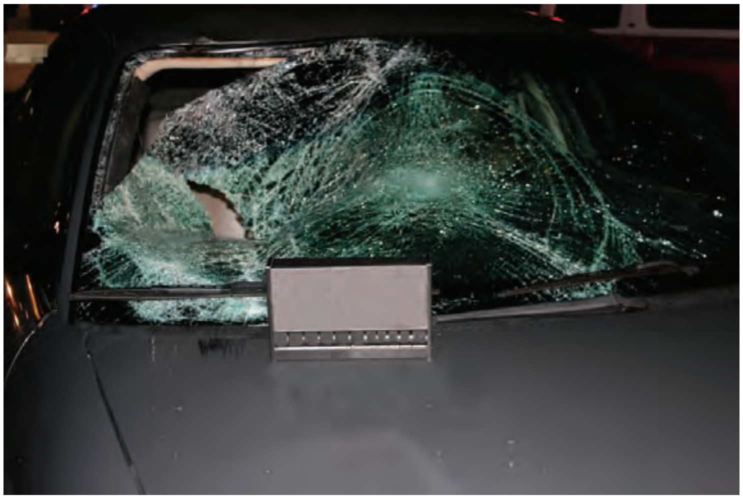

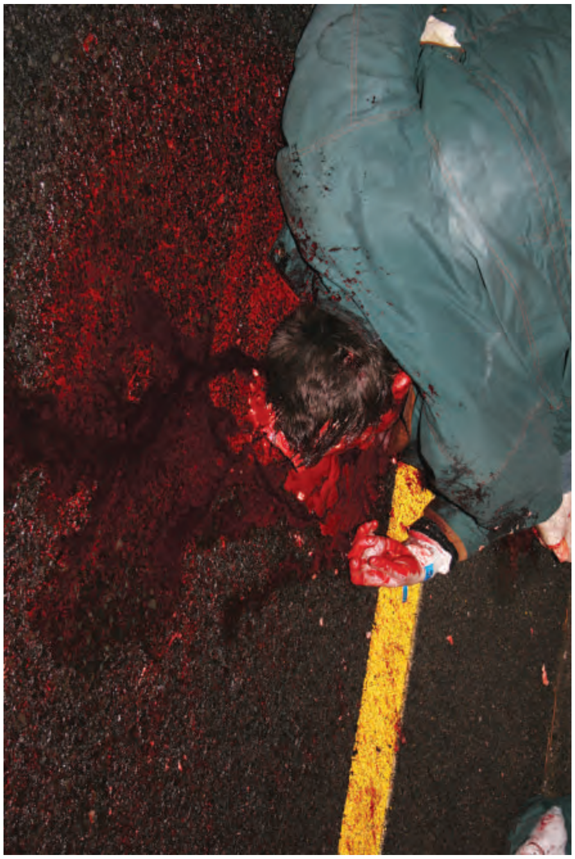

This photograph demonstrates an individual who was struck at a high rate of speed and hit the pavement with great force. Note the exposed comminuted skull fracture with brain material.

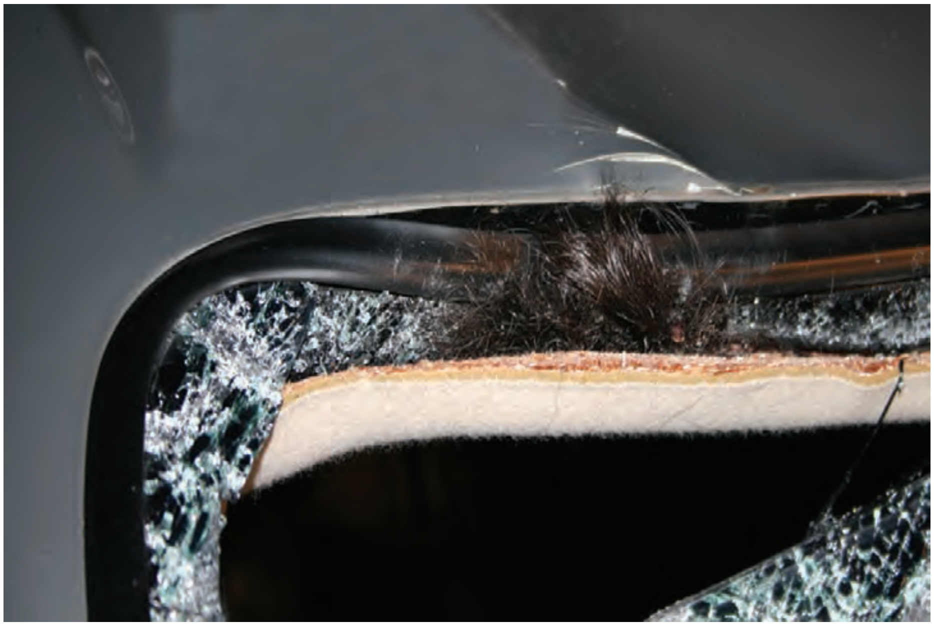

These pictures of a car show an impact from a pedestrian who was struck and run under. Note the fragment of scalp with hair embedded in the top part of the windshield and adjacent car roof. The driver of this car initially fled the scene.

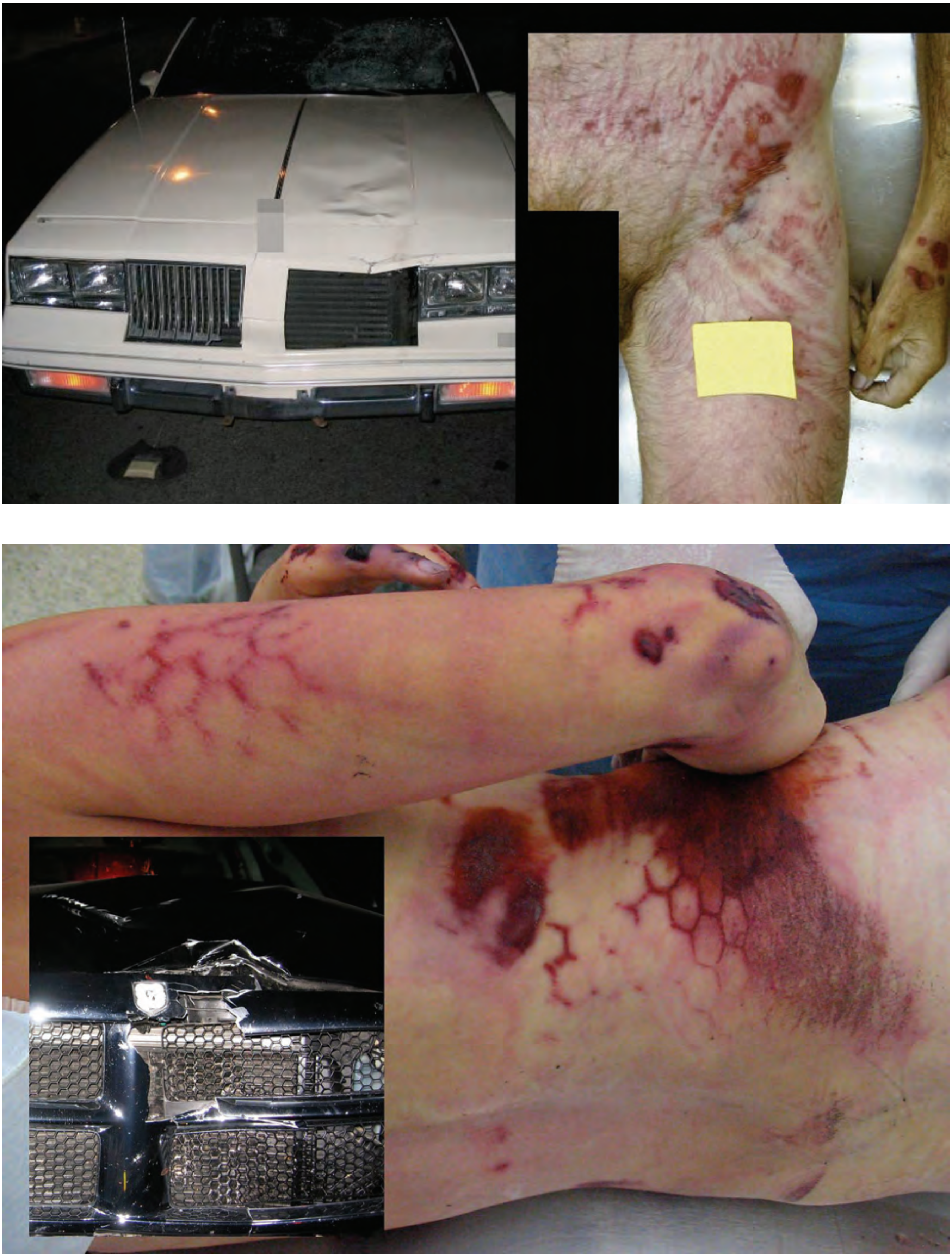

Pedestrians struck by cars with grill pattern injuries.

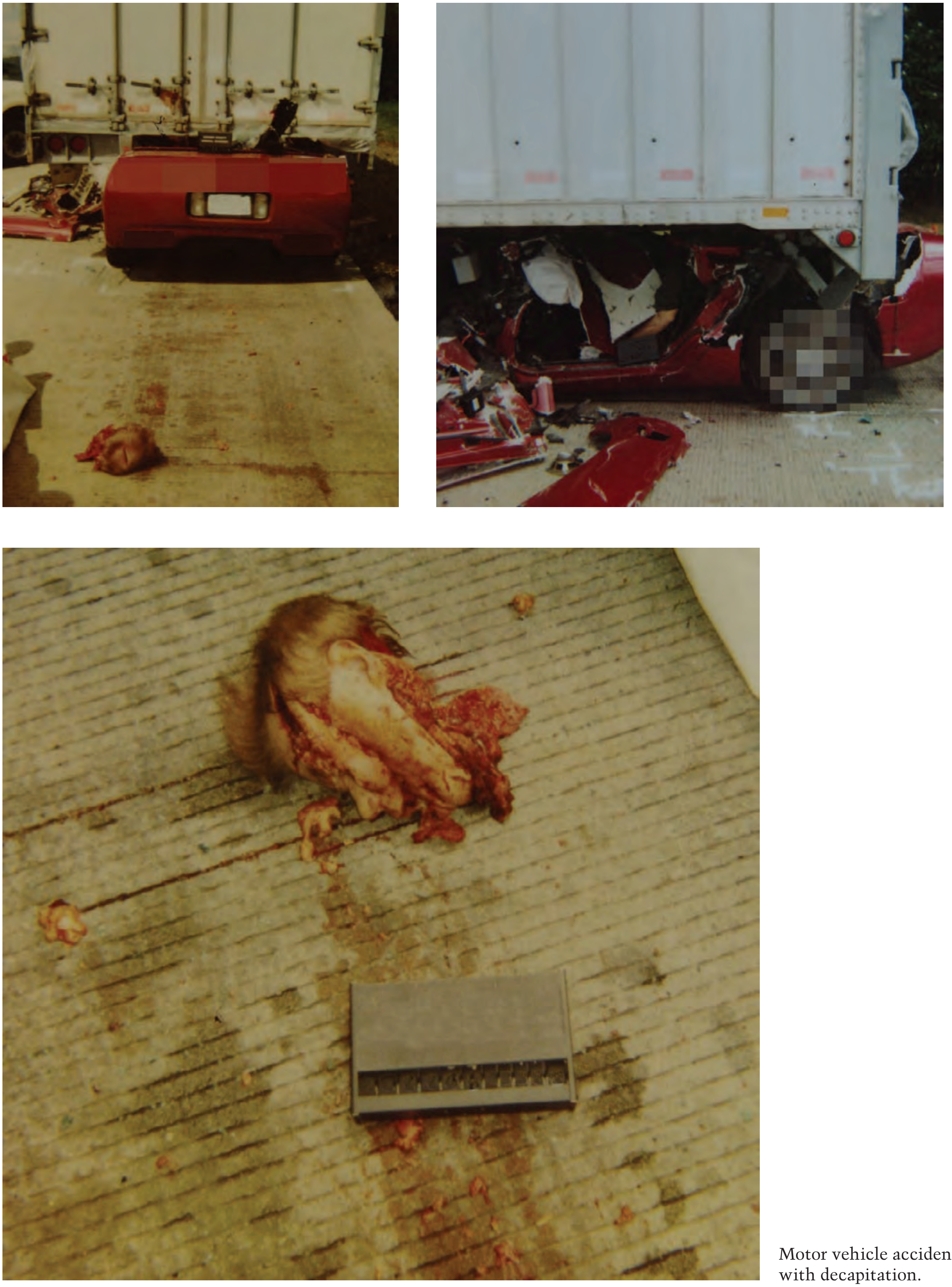

Motor vehicle accident with decapitation.

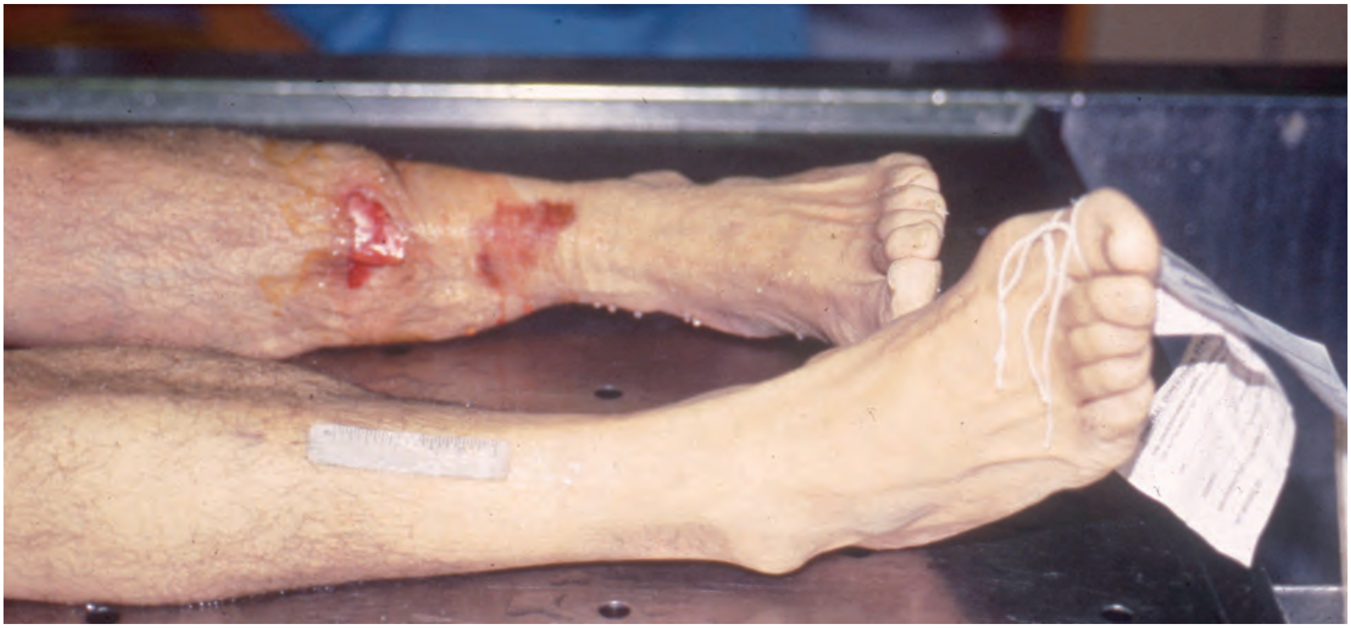

This is another pedestrian who was struck by a car. Note the tibial fracture at the weight-bearing left leg.

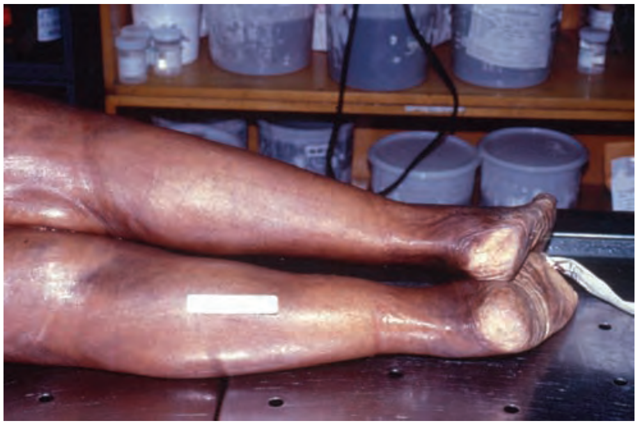

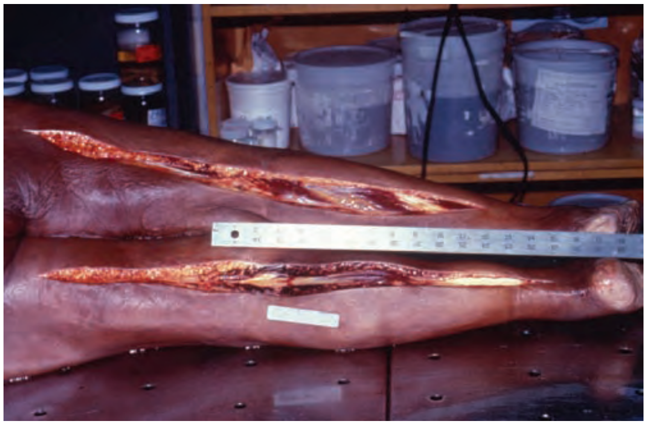

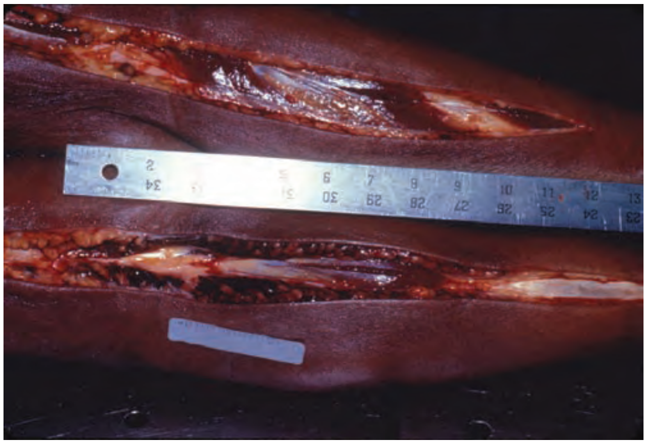

This individual was a pedestrian who was struck from behind by a car. Due to her dark skin, the contusions are not obvious from external examination alone. Incision of the posterior aspects of her leg reveal hemorrhage due to the bumper impact. It is good practice to photograph these impact sites with a ruler to demonstrate the distance from the decedent's heels. This can be matched to a particular car and to whether the driver applied brakes before striking this pedestrian. The front of the motor vehicle will go downward when the braking occurs.

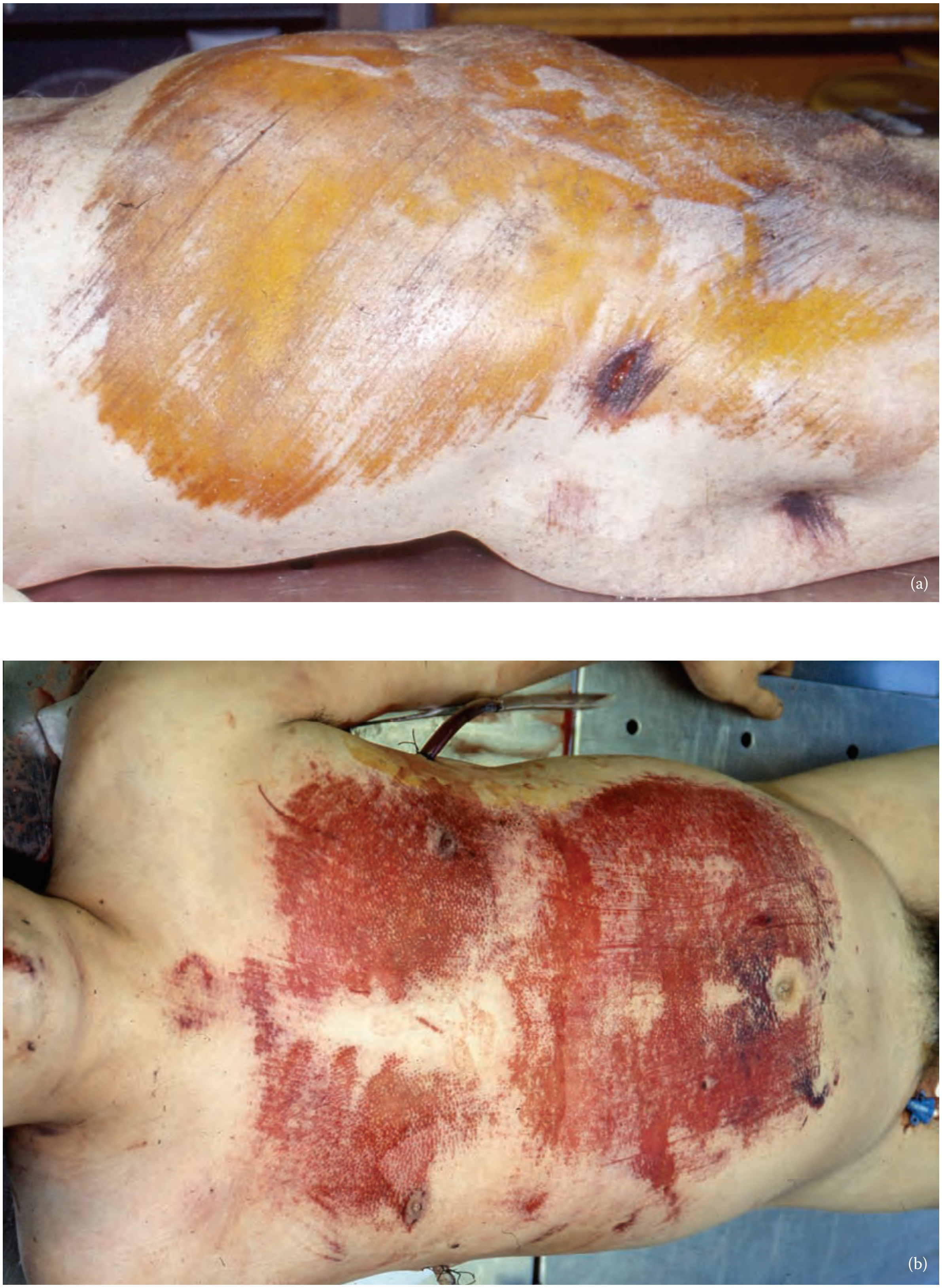

These individuals were both struck by cars and dragged across the pavement. (a) shows yellow anemic abrasions that occurred after the first impact, where the decedent sustained extensive central nervous system injury and a transected aorta. He was thrown into another lane of traffic and dragged by another car. The anemic nature of this injury and yellow discoloration suggests decreased blood perfusion. The injury in (b) shows red to brown discoloration, which is significant for vital reaction in an individual who had an intact beating heart with blood pressure.

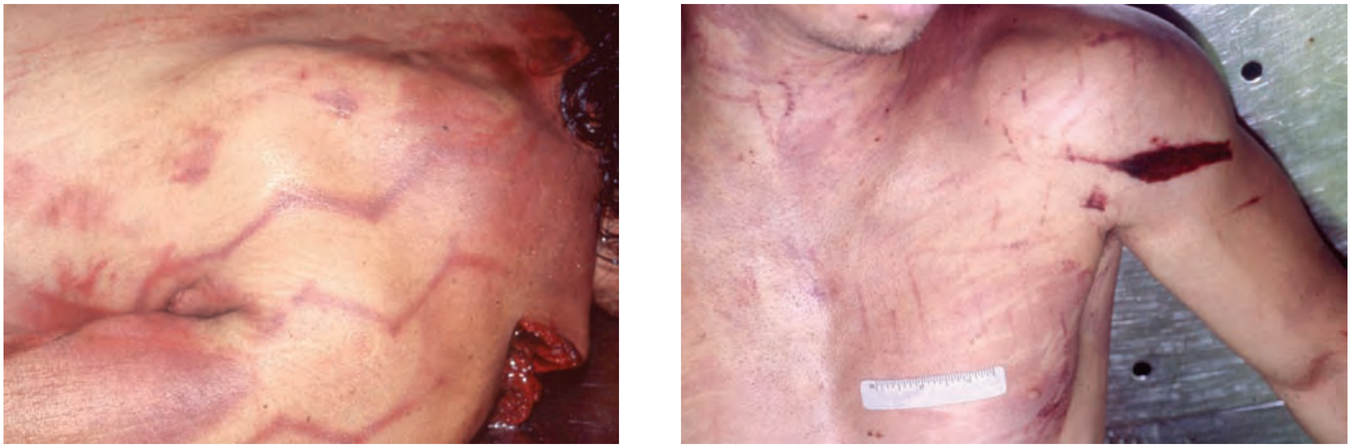

These are multiple examples of tire imprints from people who were run over by cars. In one case, the individual was thrown into another lane of traffic after being struck. The second car denied hitting the individual, but his tire pattern was a perfect match and there was forensic evidence found on the under surface of his car.

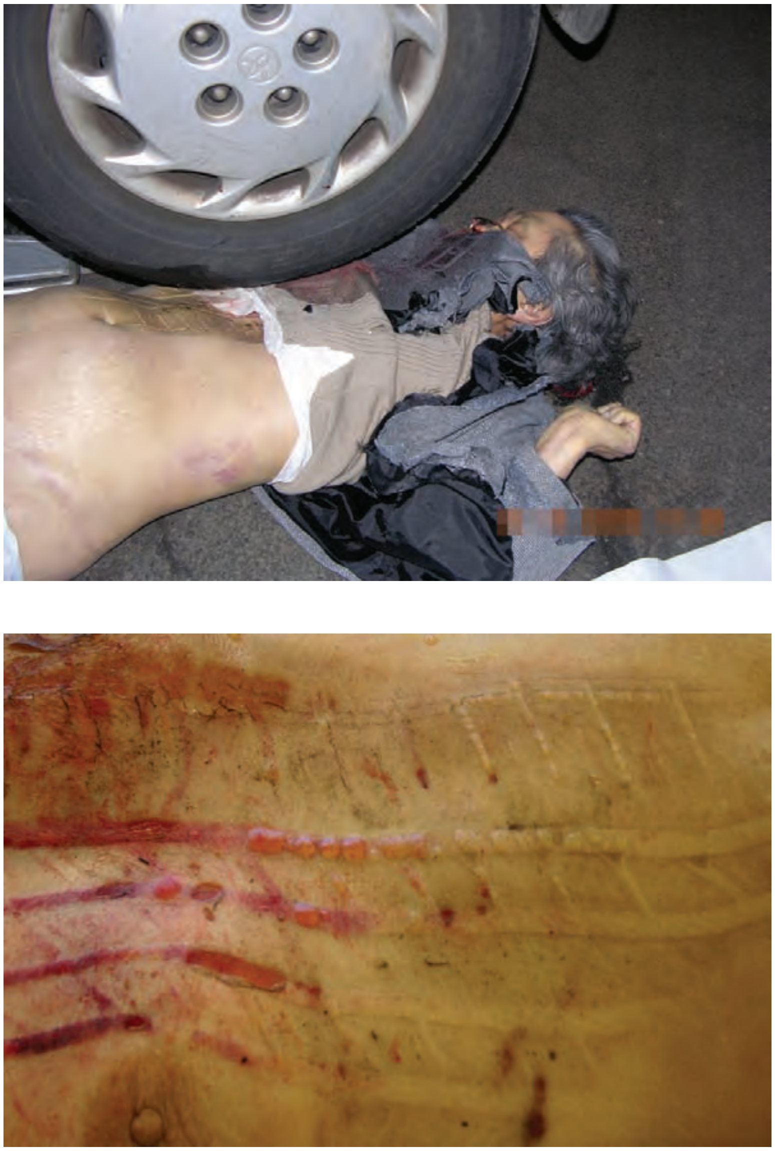

Individual trapped under a car with tire imprint to chest.

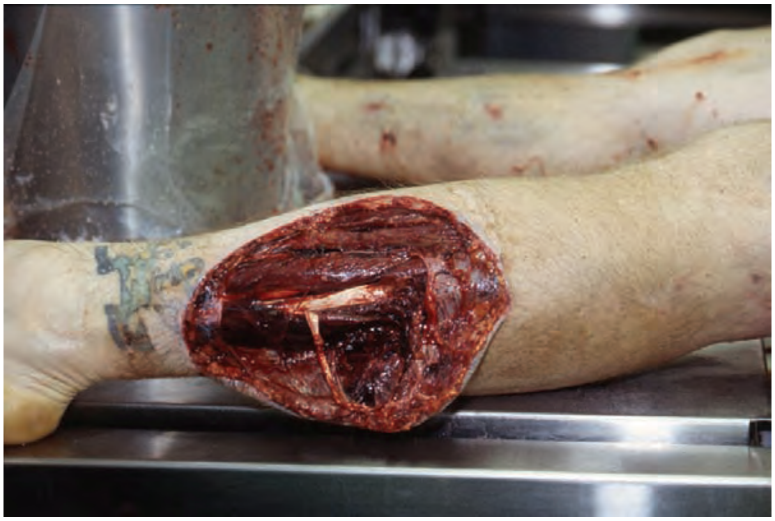

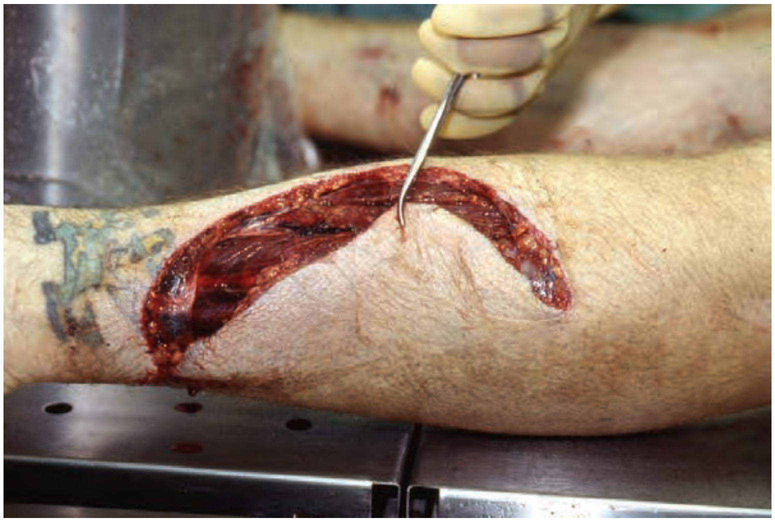

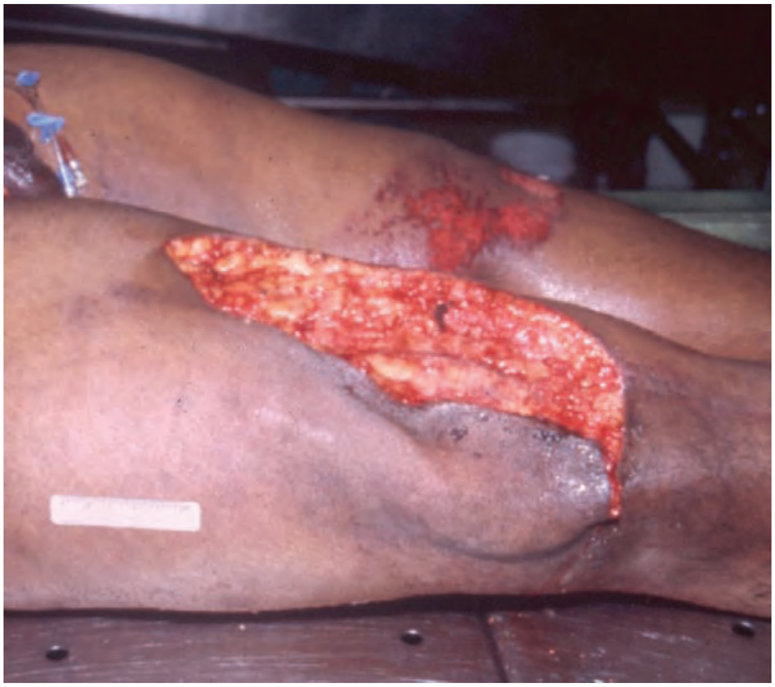

These individuals were run over by a motor vehicle while lying on the ground. Note the flap of skin being torn away from the thigh as the tire rolled over the leg.

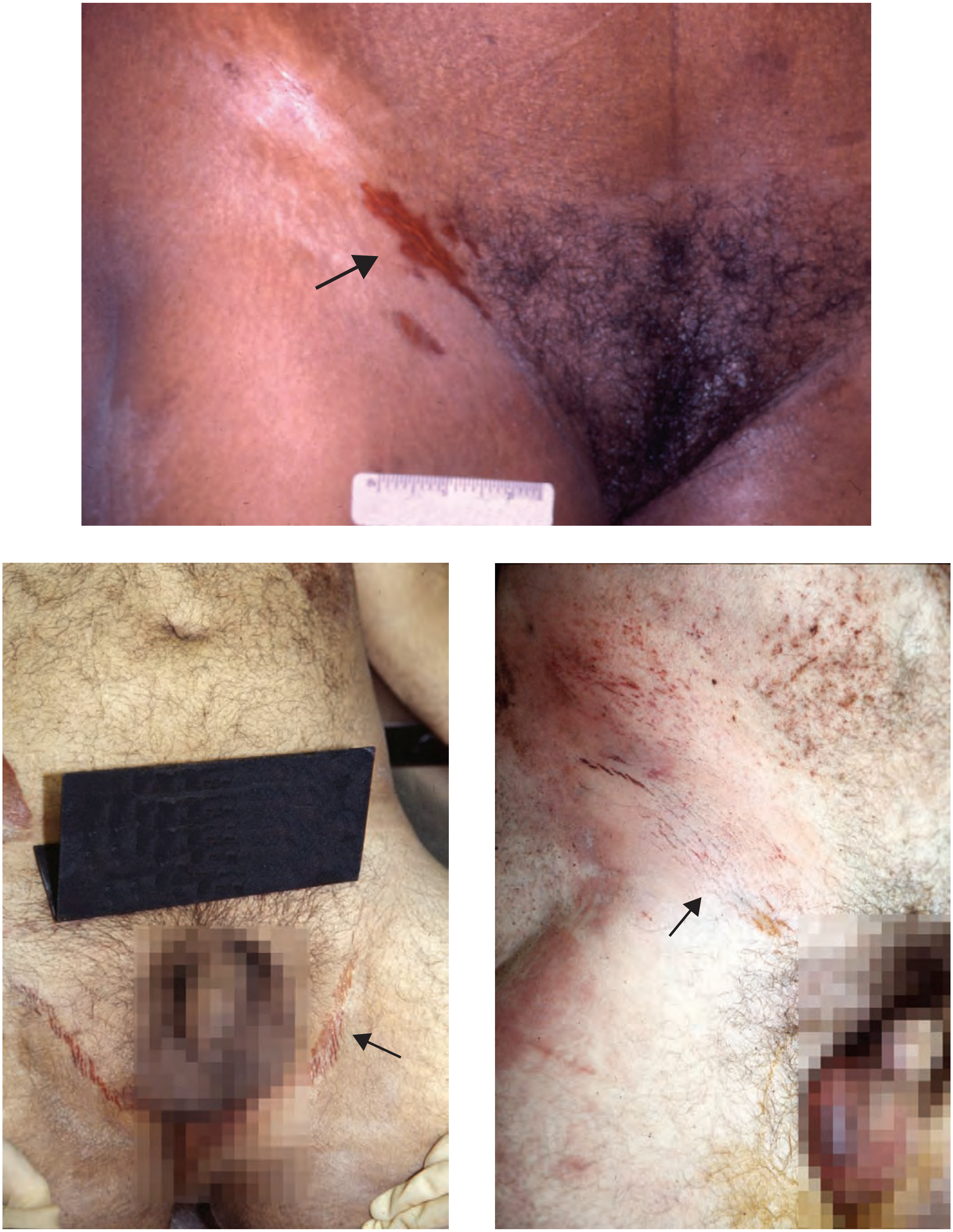

These individuals were struck from behind by a car. Note the inguinal stretch marks caused by hyperextension of the hips and legs at the time of impact.

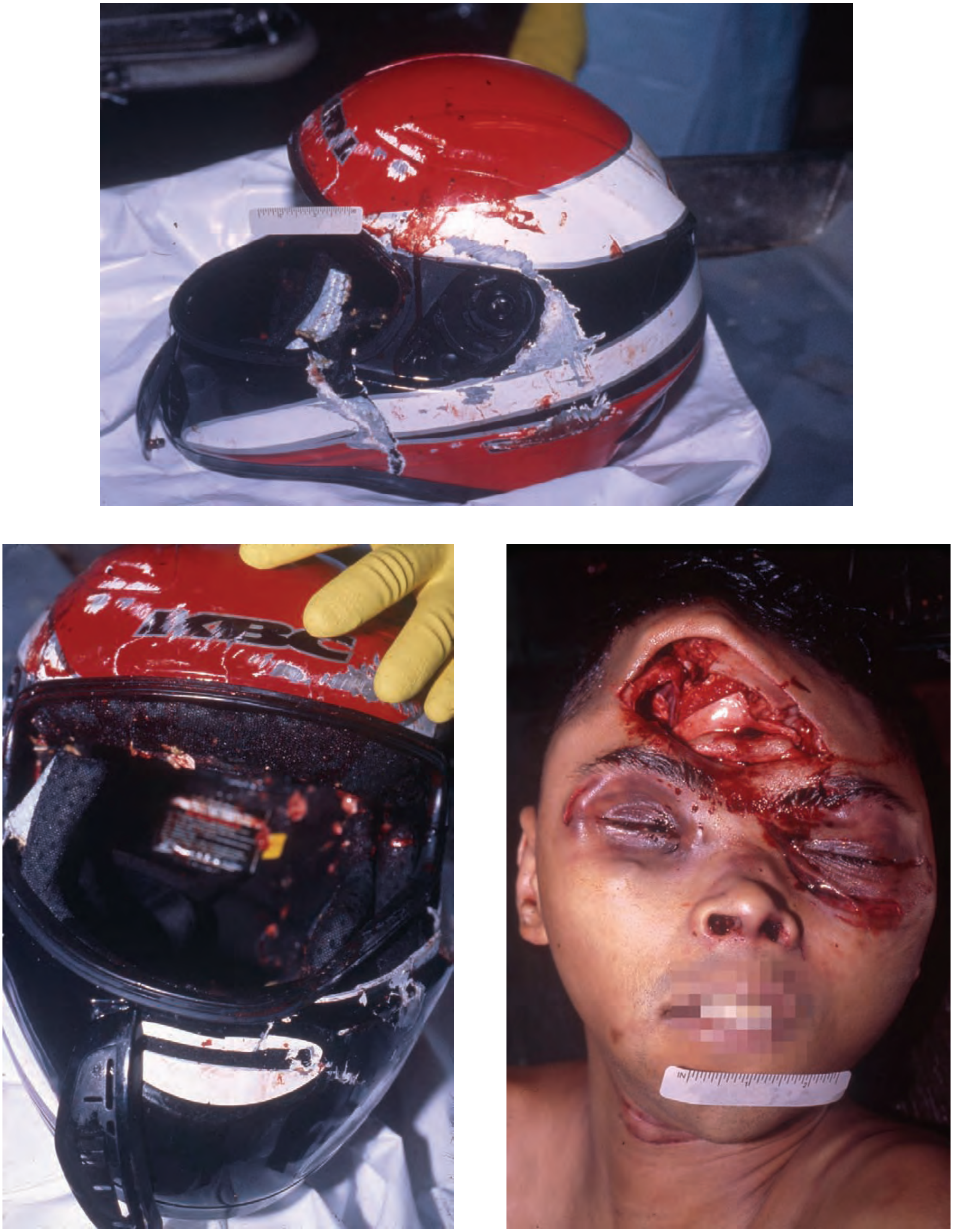

This fractured motorcycle helmet shows blood clot and brain material from an individual who fractured his skull after striking a pole.

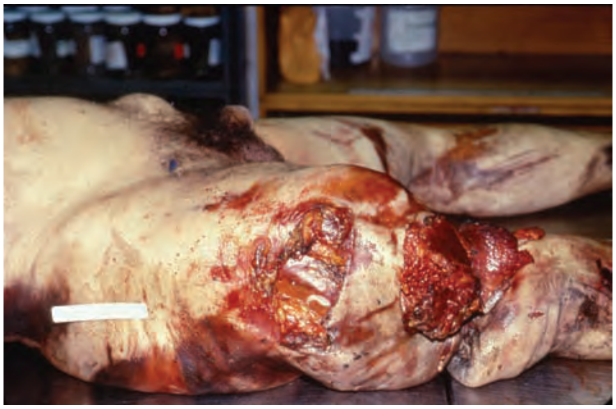

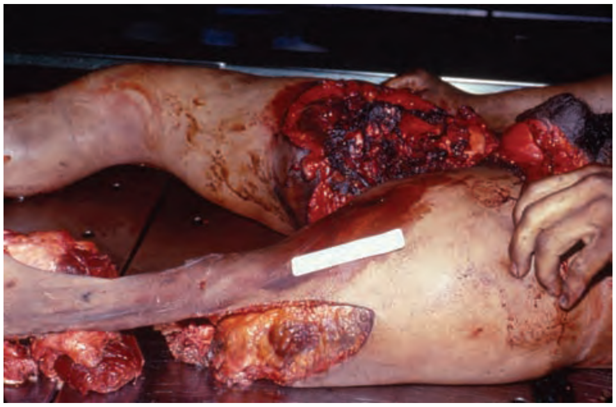

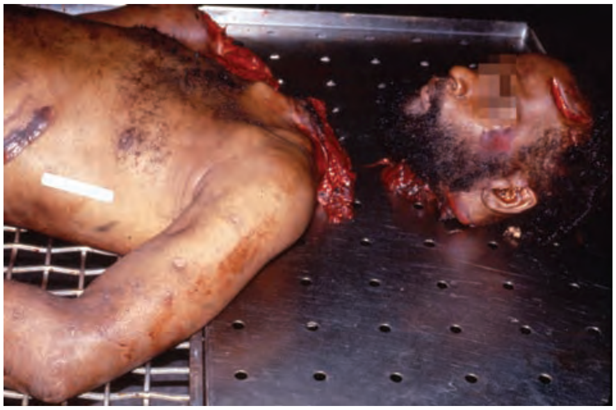

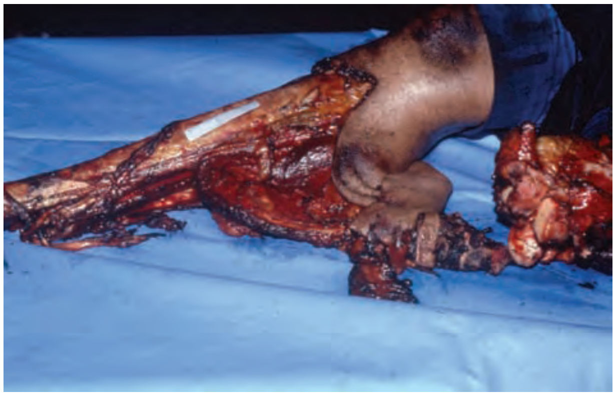

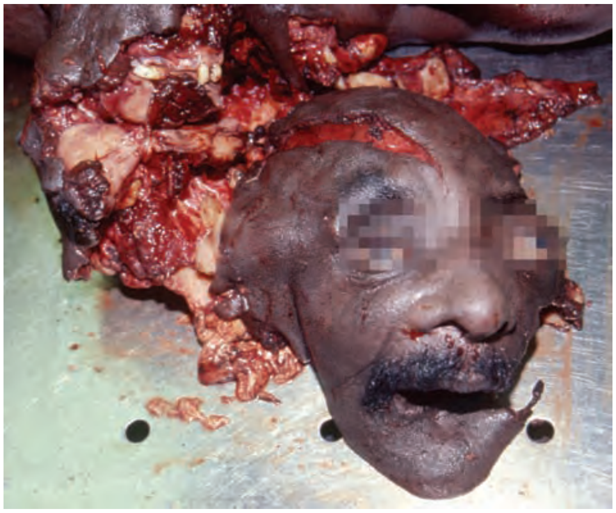

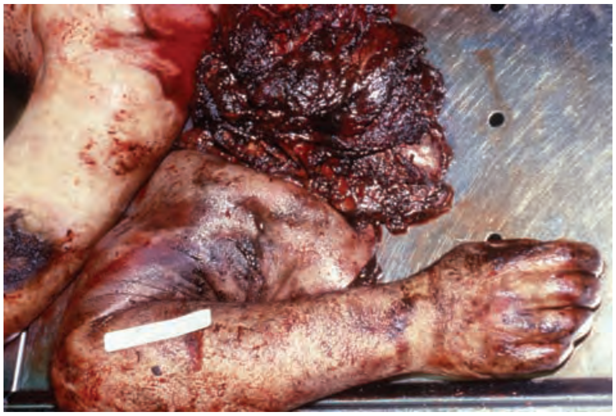

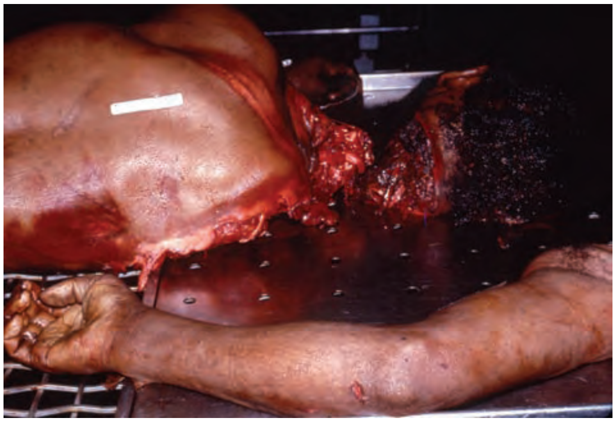

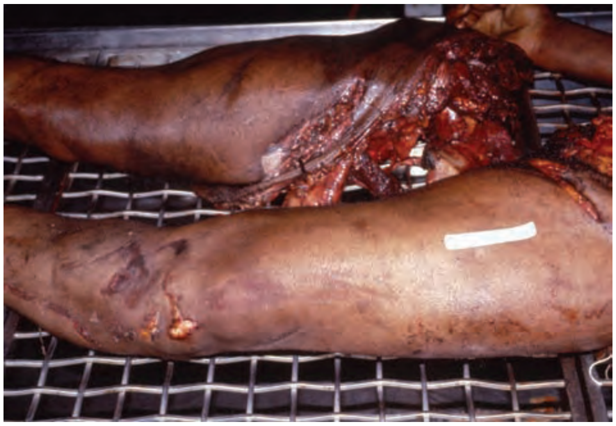

These figures demonstrate various subway fatalities where individuals were run over by trains. Note the separated body portions with extensive crush injury and axle grease from the train wheels.

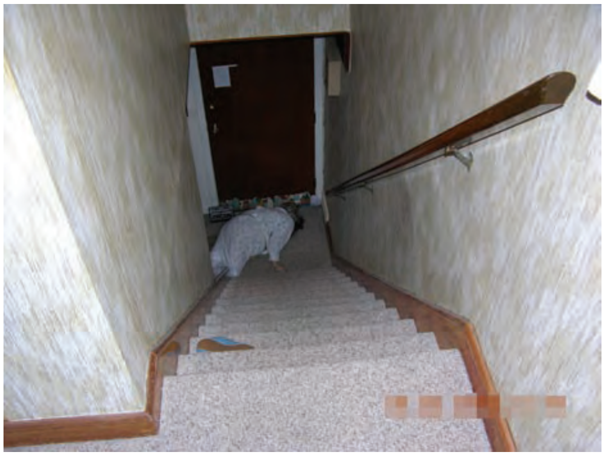

This individual fell from the top of her steps and sustained multiple blunt-impact injuries including a femur fracture and a skull fracture. Note the slipper at the top steps.

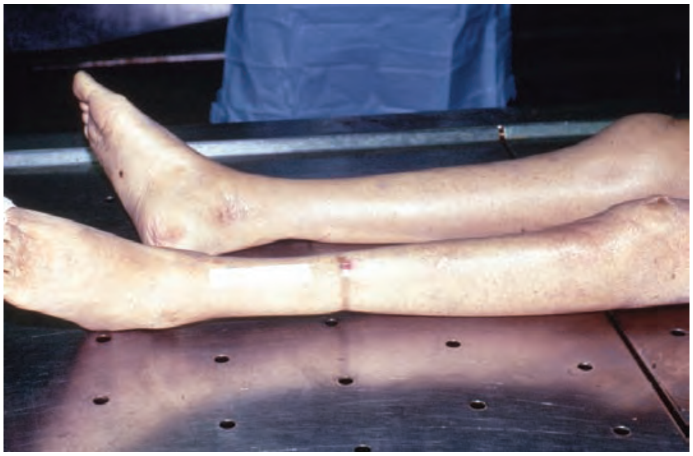

This individual has a femur fracture. Note that the fractured leg is shorter and the foot is laterally rotated.

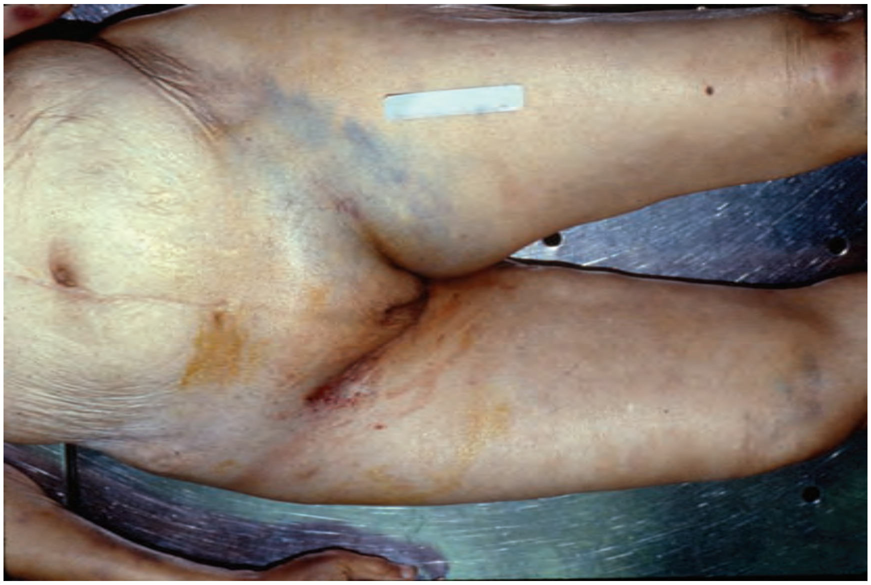

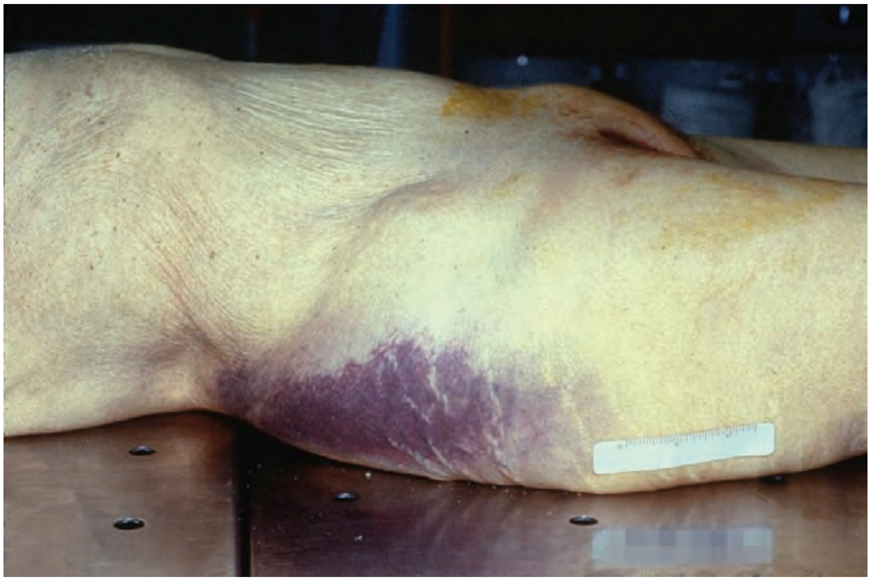

People with significant osteoporosis may fracture bones with little force. These two pictures demonstrate large areas of contusion and ecchymosis following a femur fracture secondary to a standing height fall.

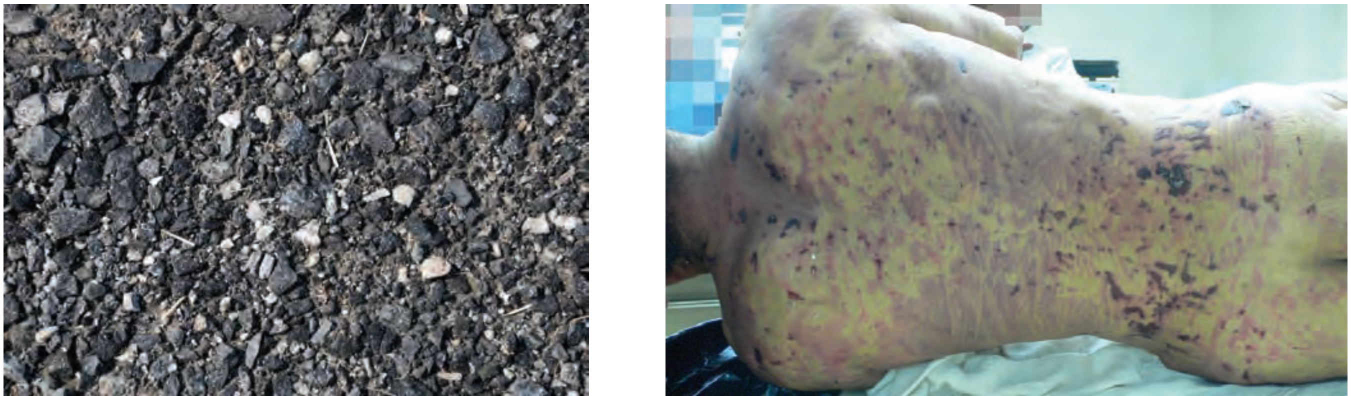

Angulated irregular abrasions of back due to striking a gravel bed with great force.

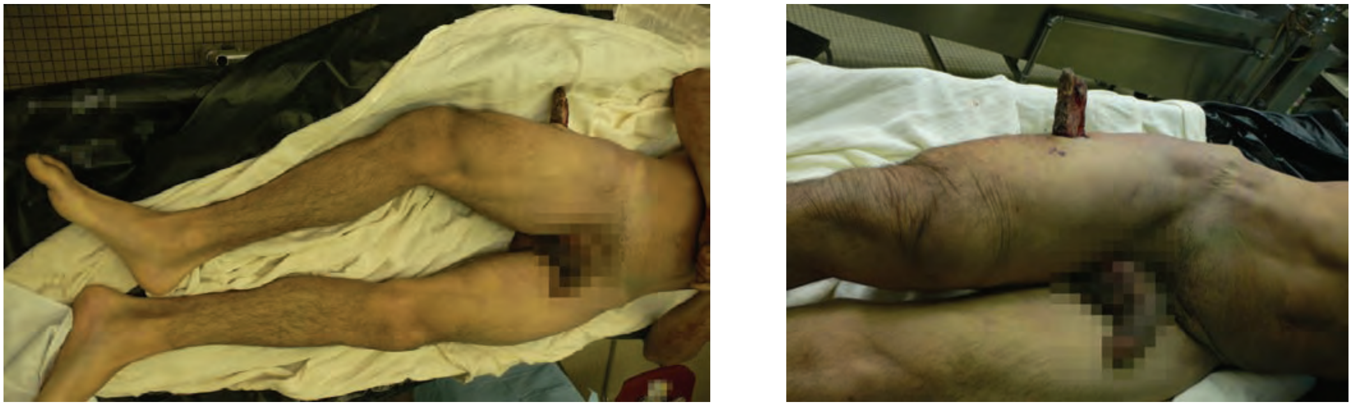

This decedent jumped from a great height and fractured his right femur. Note the fractured bone protruding from the thigh. Also note the shortening of the right leg with lateral rotation.

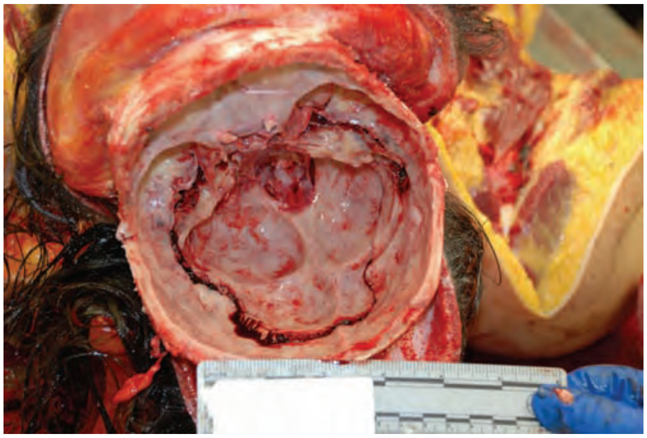

Ring fracture. This usually occurs following a fall with initial impact to the feet or buttocks.

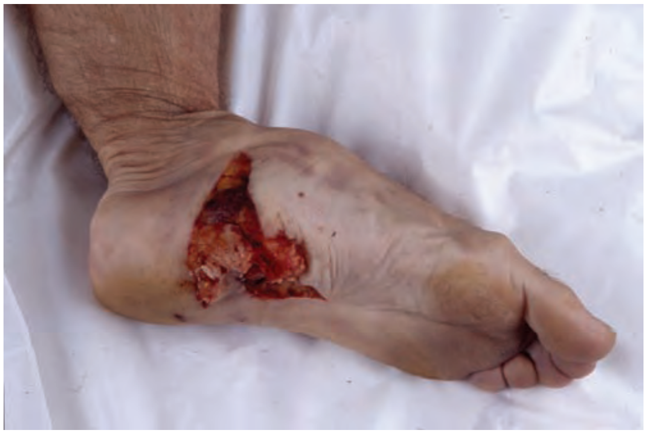

The decedent fell from a great height and landed on his feet. The fractured tibia was forced through the bottom of the foot upon impact.

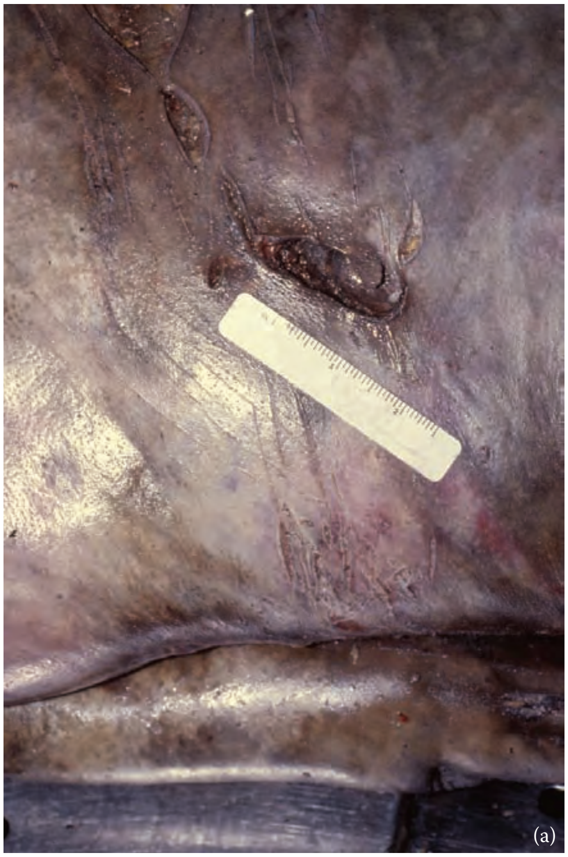

This individual was placed in a garbage dumpster and left to decompose. The trash compactor in the truck was engaged before the body was observed. Note the garbage truck shovel laceration to the back (a). Also note the intestines forced from the anus following compactor compression (b).

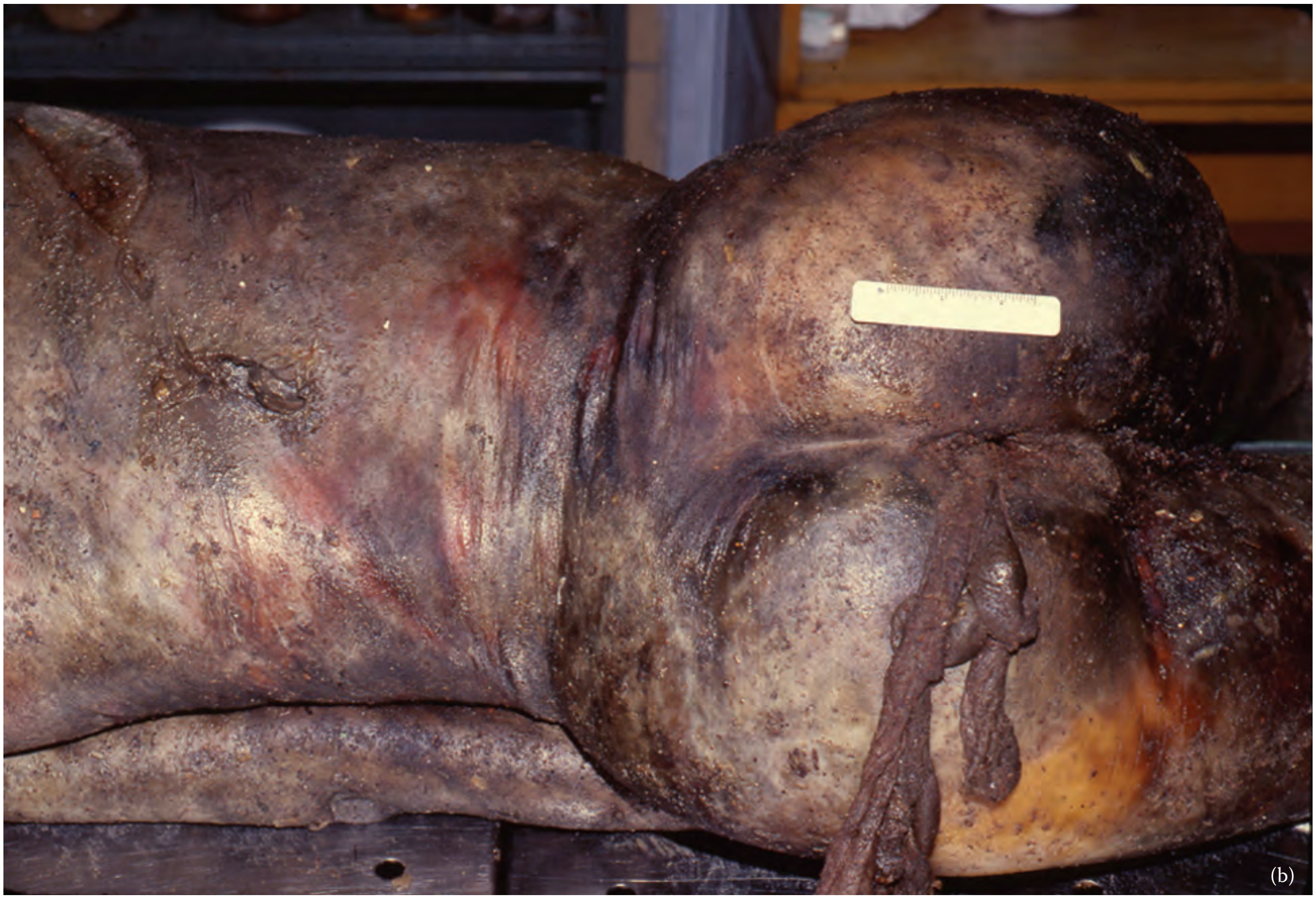

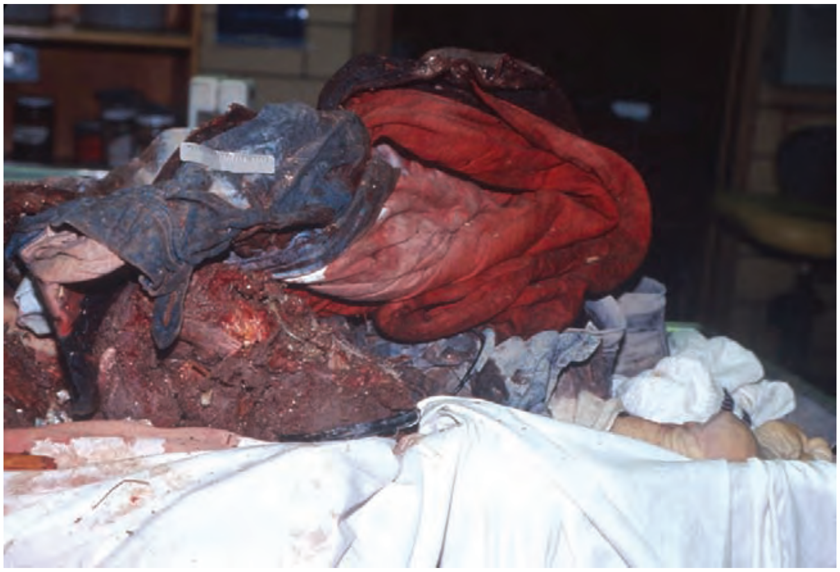

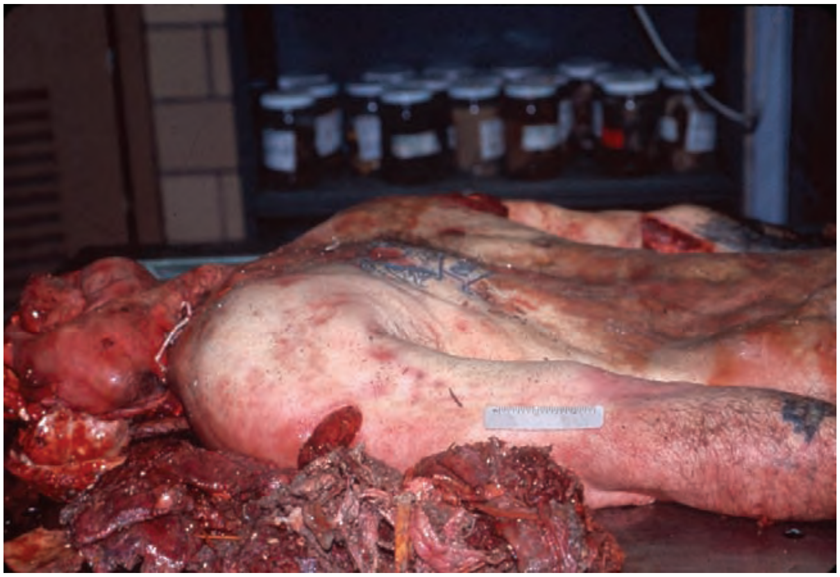

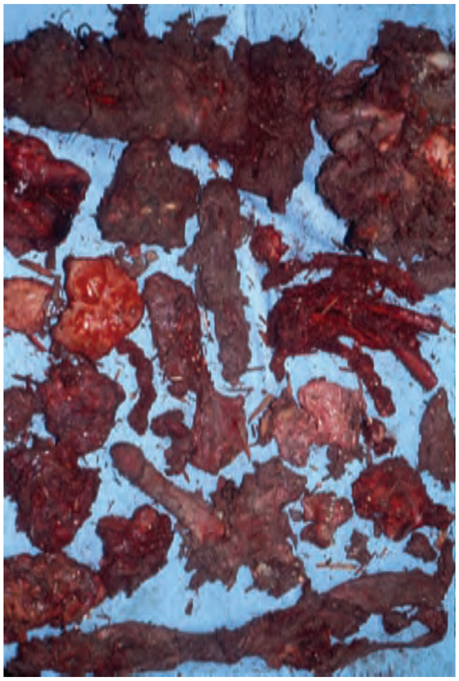

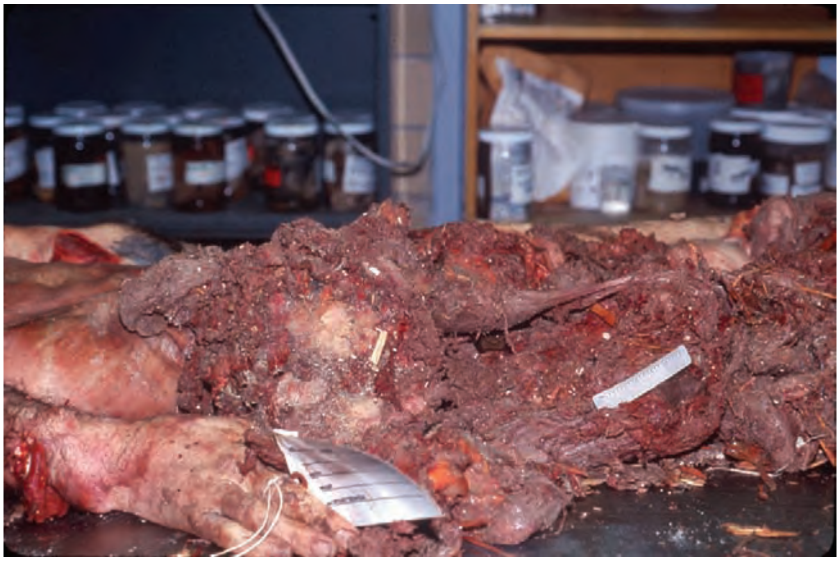

This individual was shot in the head, processed through a trash compactor, and left to decompose. Note the flattening of the body with extensive blunt-force injury and fragmentation.

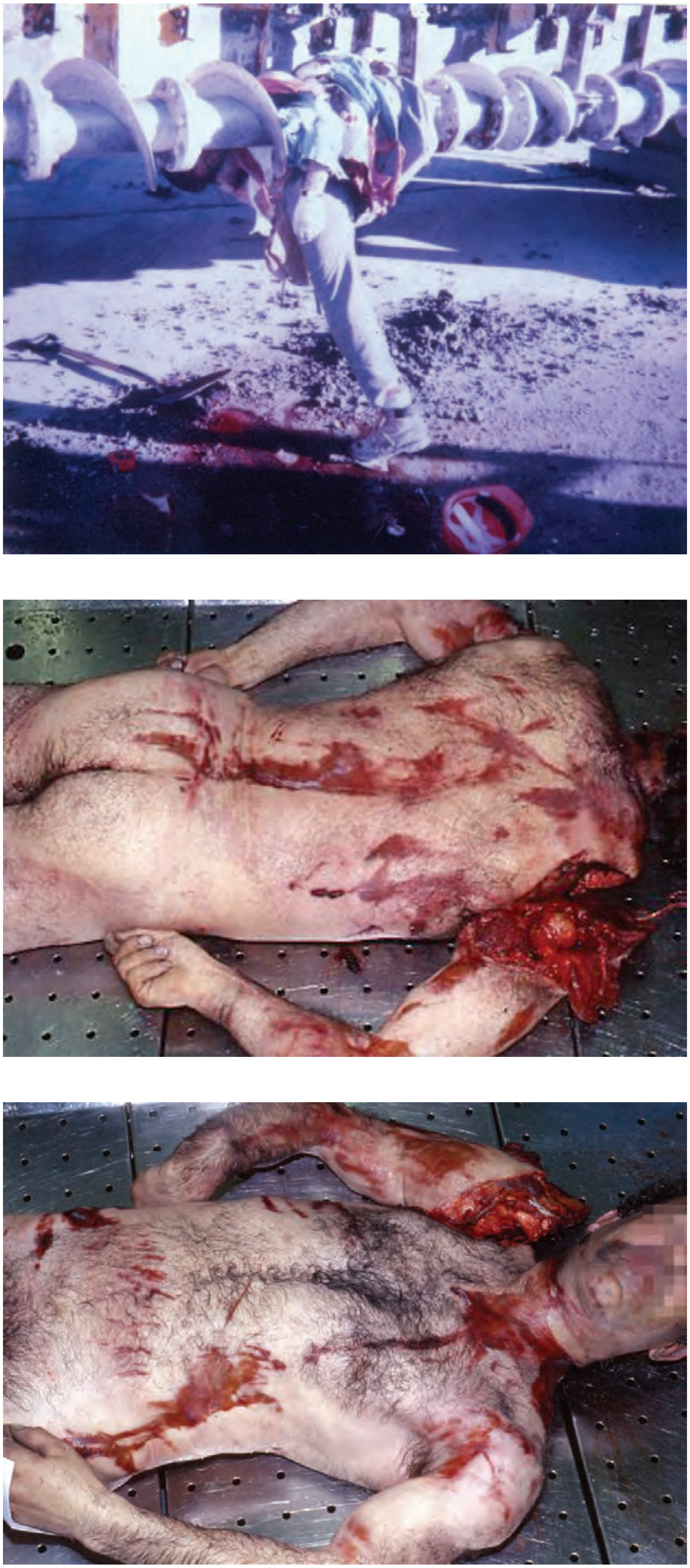

This individual slipped into a terrain leveler and sustained multiple abrasions, with blunt- and sharp-force injuries including extensive fractures and internal lacerations.

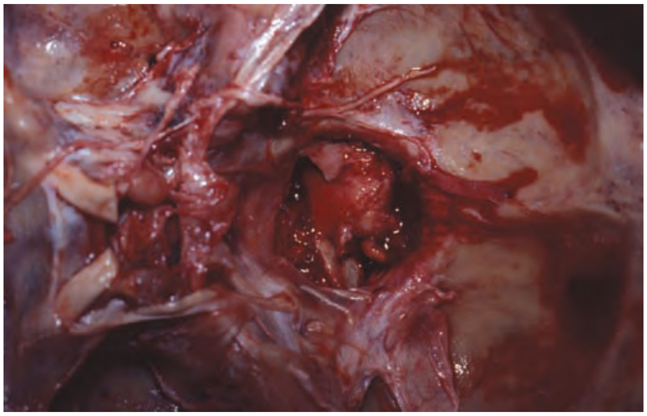

This individual had a severe blunt impact to his face that caused the atlanto-occipital ligaments to stretch and dislocate. This caused brainstem and upper cervical spinal cord injury resulting in asystole and apnea.

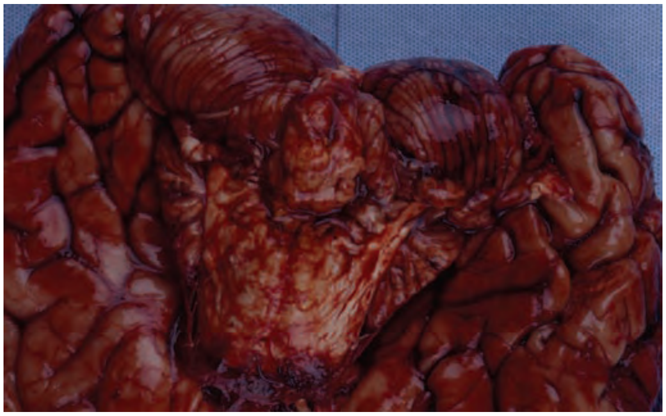

This picture of the base of the brain shows a medullary pontine laceration with extensive subarachnoid hemorrhage.

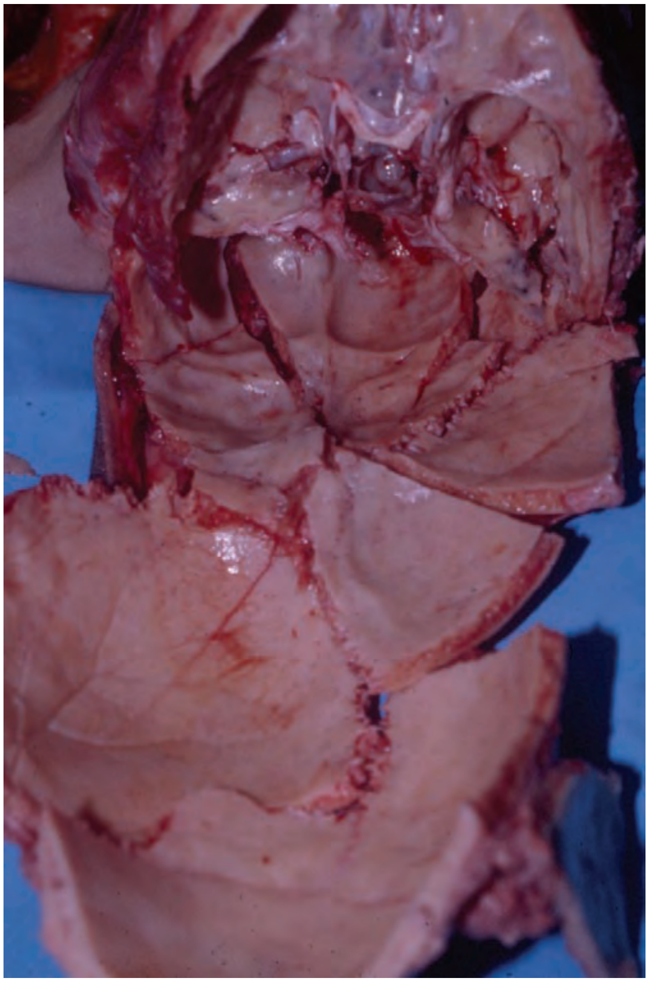

Comminuted skull fracture where the bone is broken into separate pieces.

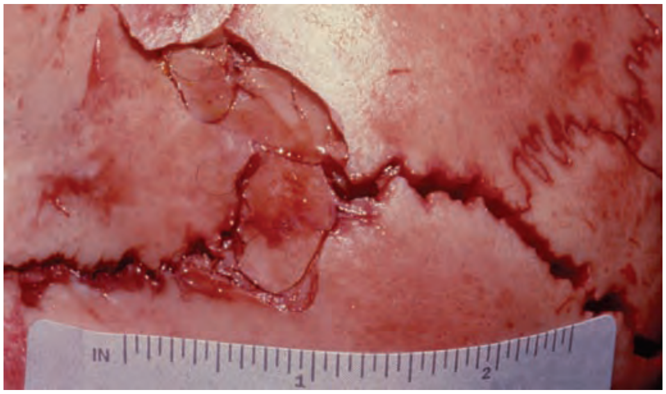

Diastatic skull fracture showing the fracture site passes through a cranial suture.

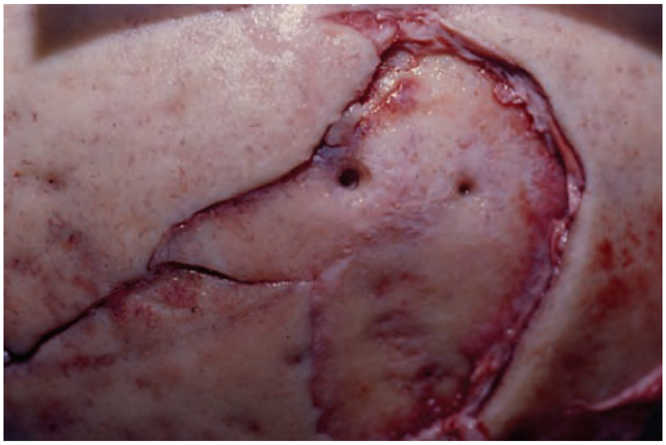

Healing skull fracture.

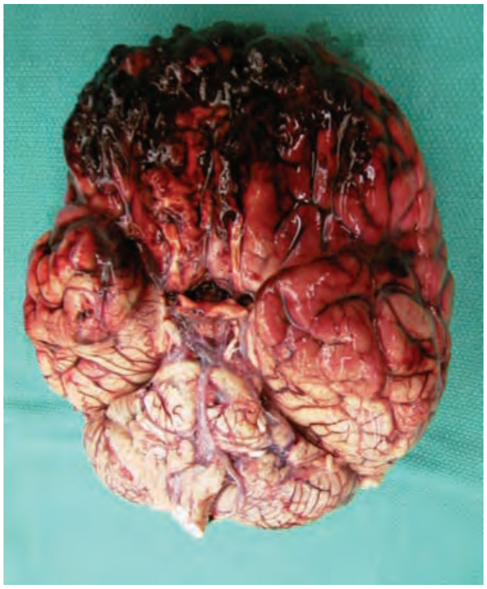

Frontal and temporal contusional hematomas associated with a falling impact to the back of the head in an individual with liver disease.

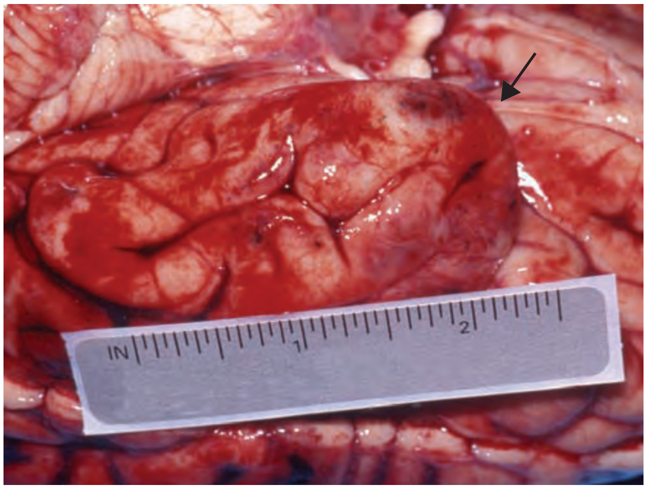

Temporal lobe with multiple splinter-type hemorrhages demonstrating typical cerebrocortical contusions.

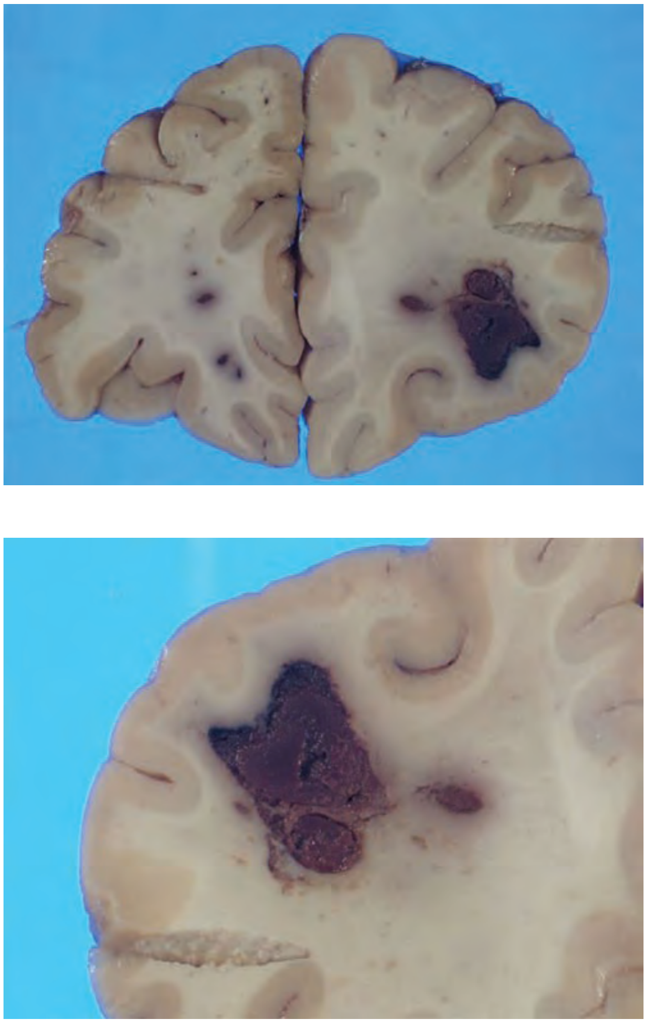

These images demonstrate gliding cortical contusions that typically occur near the white and gray matter junction following blunt-force head trauma.

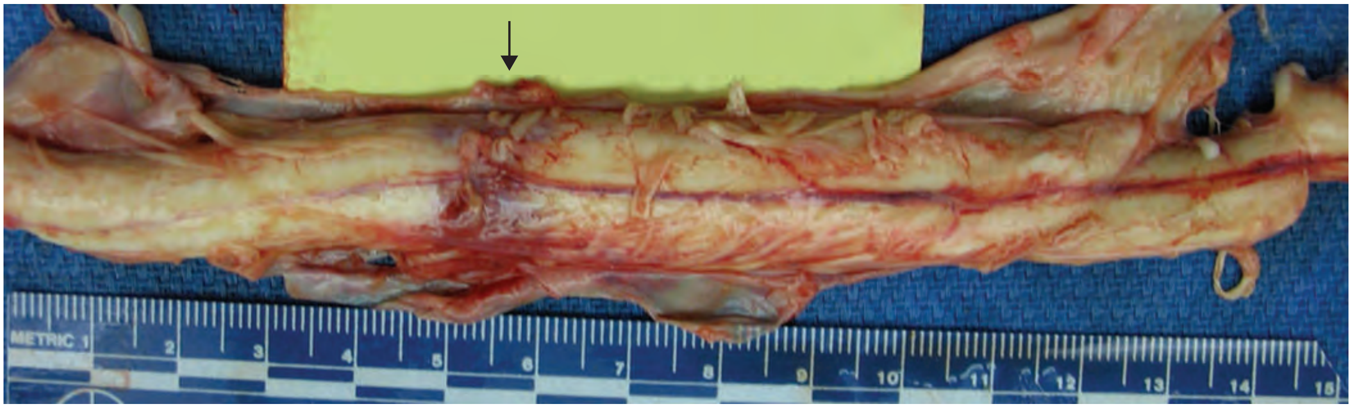

Spinal cord with a region of purple discoloration and softening demonstrating a spinal cord contusion.

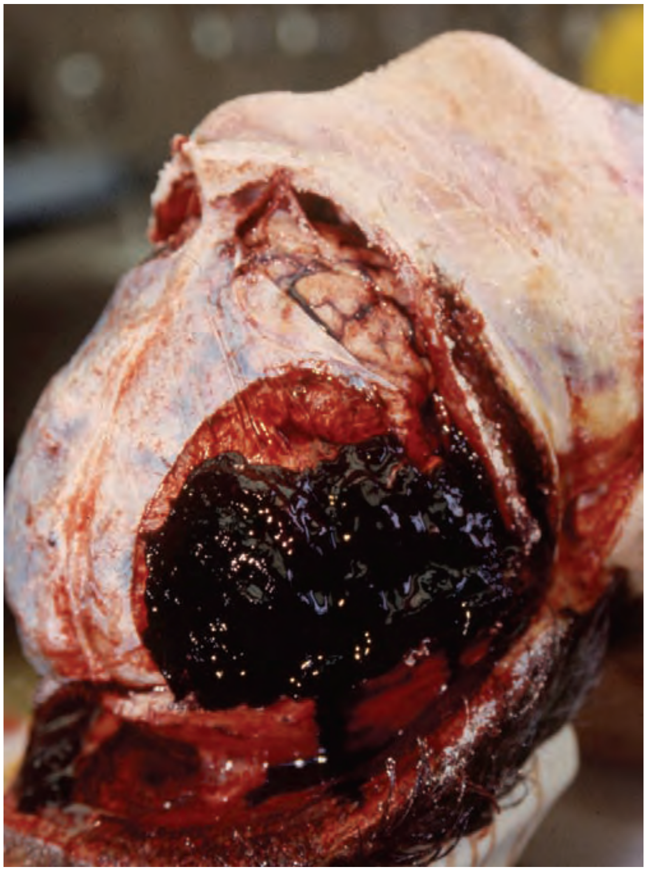

Epidural hemorrhage in an individual with a temporal bone fracture and middle meningeal artery laceration. Note the clotted blood at the surface of the dura mater.

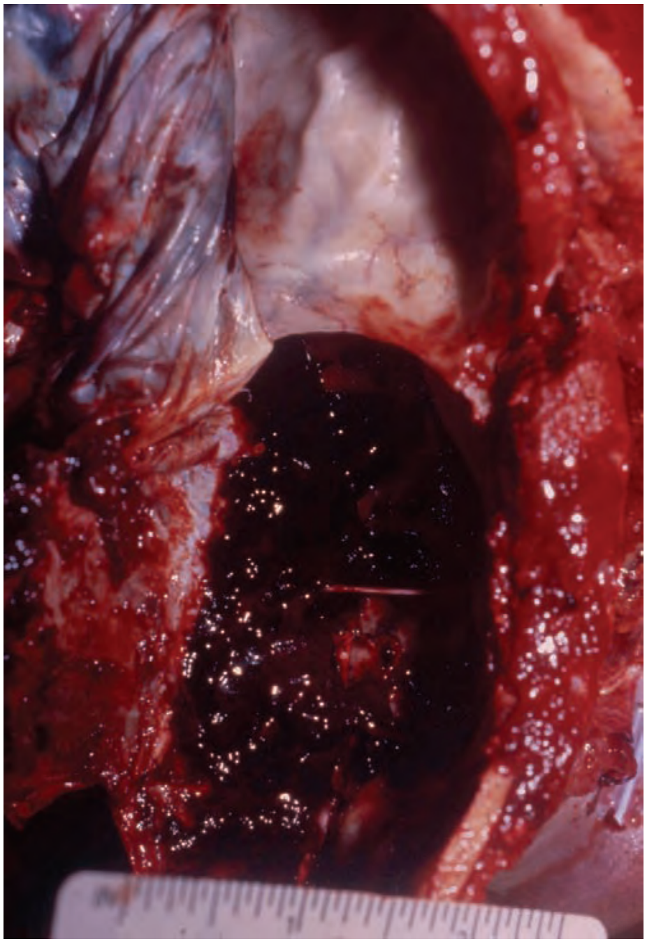

Epidural hemorrhage.

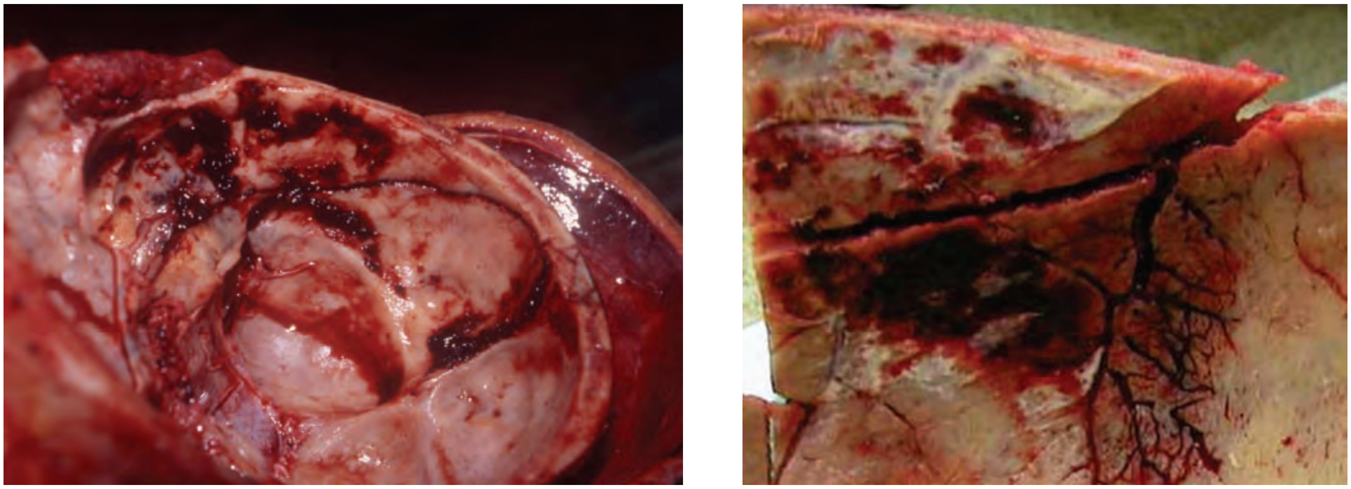

This shows the dura mater stripped from the inner aspect of the calvarium. Note the clotted adherent epidural hemorrhage within the temporal region of the skull.

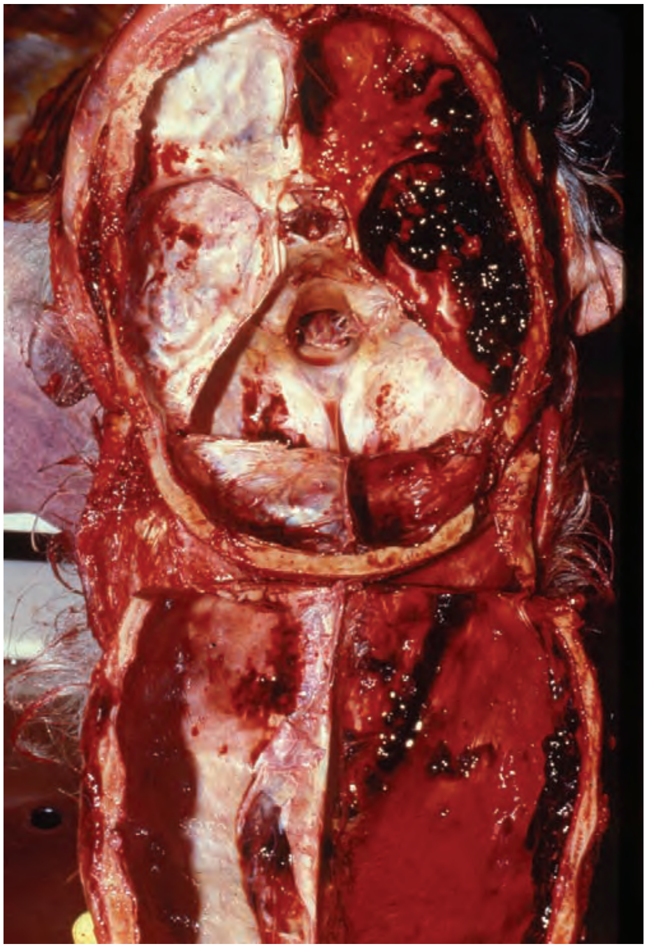

Acute right subdural hemorrhage.

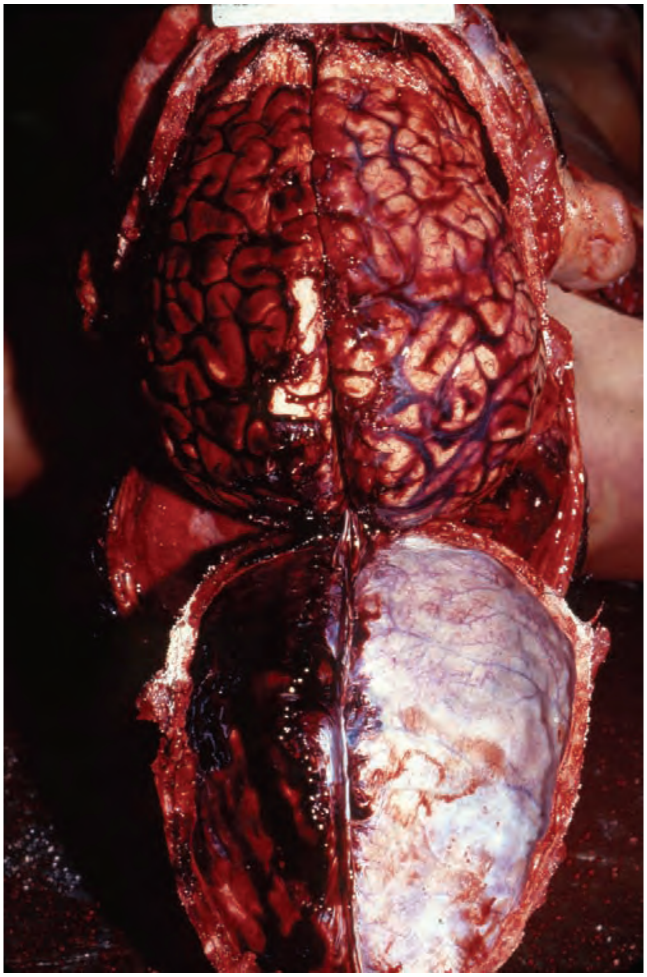

Left subdural hemorrhage that is slightly older with clotted adherent blood.

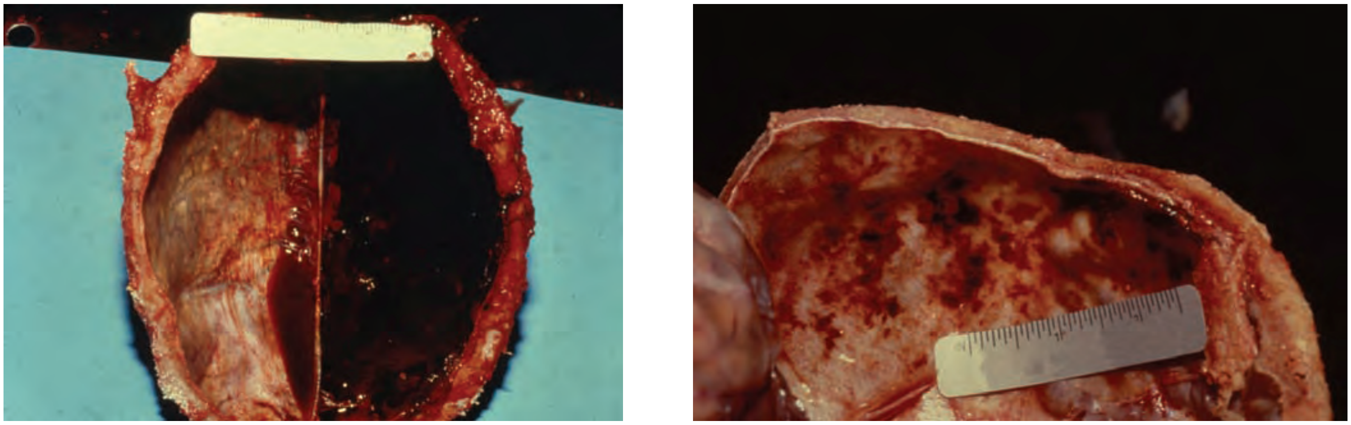

This demonstrates an older, clotted, red, adherent, subdural hemorrhage with portions of rust discoloration and membrane formation.

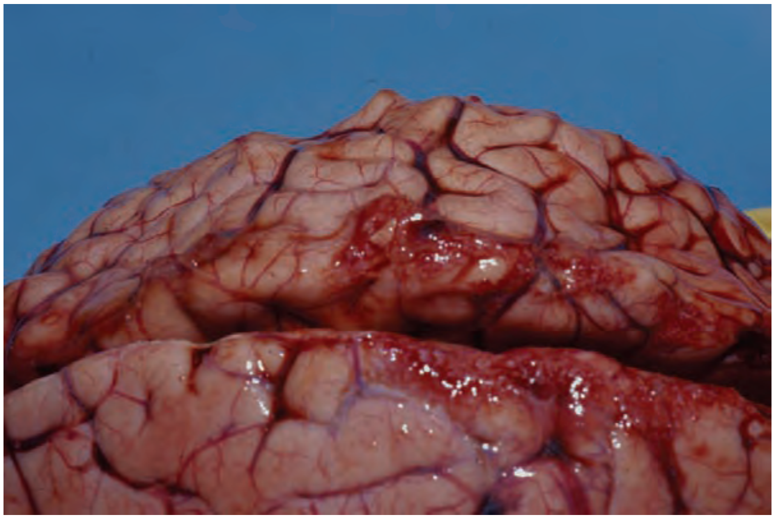

This cerebral cortex demonstrates an irregular flattened surface caused by a chronic subdural hematoma. In contrast, a chronic epidural hematoma generally leaves a flattened and less irregular cerebral cortex deformation.

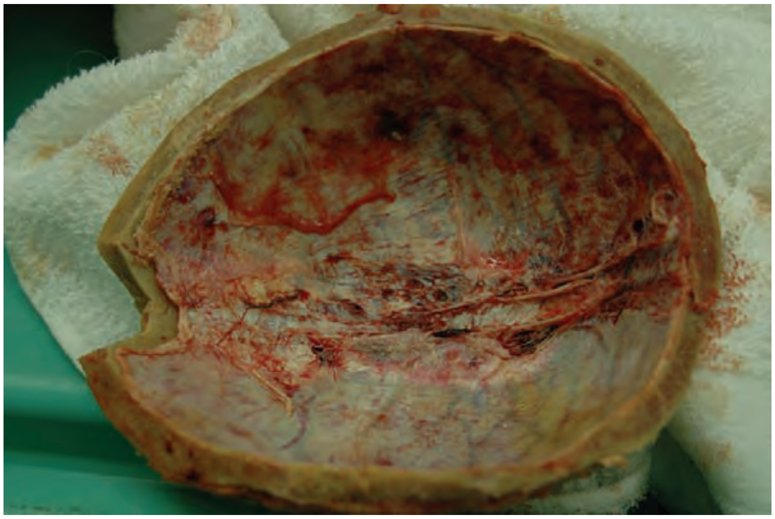

Old subdural hemorrhage with clear to rust-colored membrane formation. Note the membrane separation with beading up away from the midline caused by scraping a scalpel blade along the subdural surface.

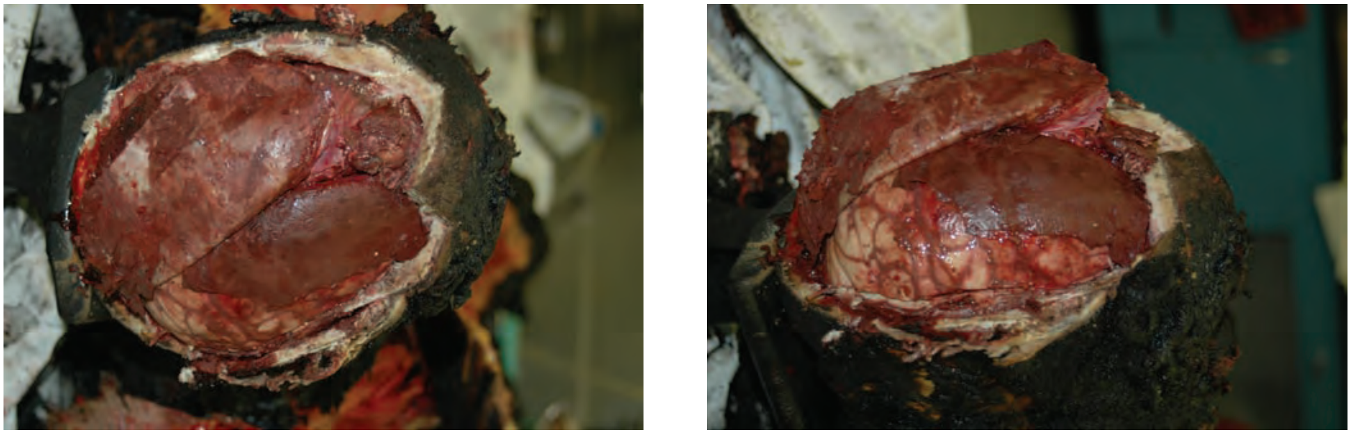

Large antemortem subdural hemorrhage altered by excessive heat due to fire.

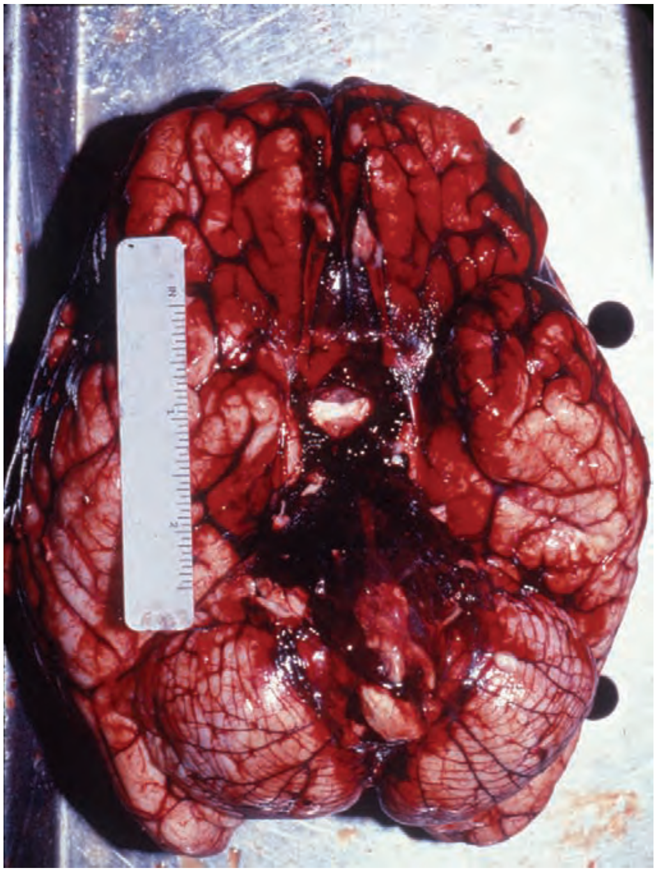

This brain demonstrates multifocal subarachnoid hemorrhage with subarachnoid hematoma at the base of the brain. This hemorrhage occurred following a blunt impact to the face causing hyperextension and rotation of the head with laceration of the right vertebral artery.
Source: Kindleburger D., Gilmore K., Catanese C., Armbrustmacher V. Subarachnoid hemorrhage due to internal carotid artery laceration following blunt impact to face with hyperextension and rotation of neck: A case report and review of the literature. *Journal of Forensic Science* 48:1366–1368, 2003.

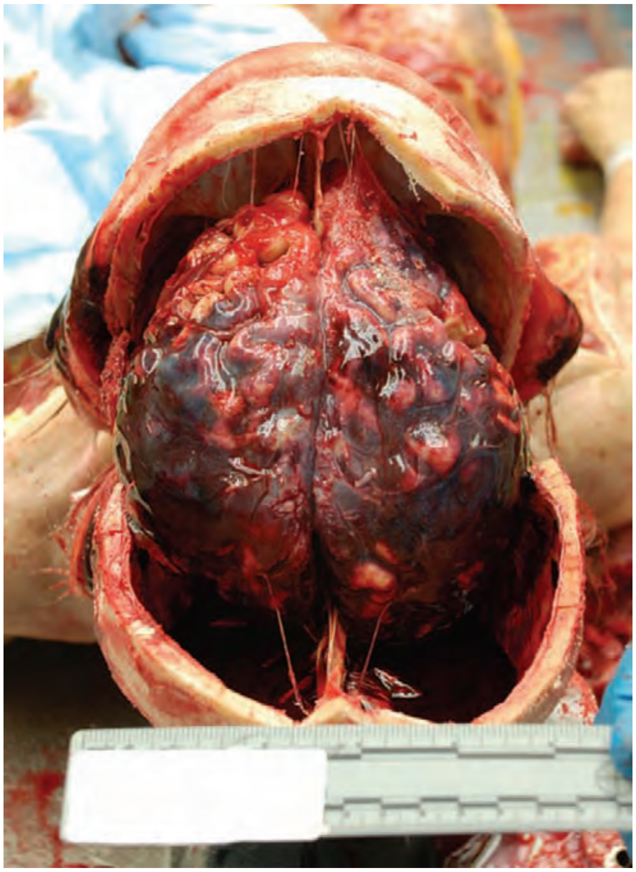

This demonstrates a large subarachnoid hemorrhage in an individual who died several days following a traumatic head injury.

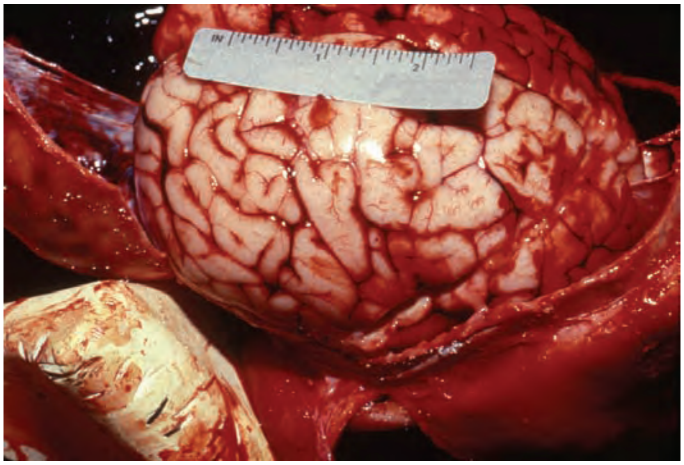

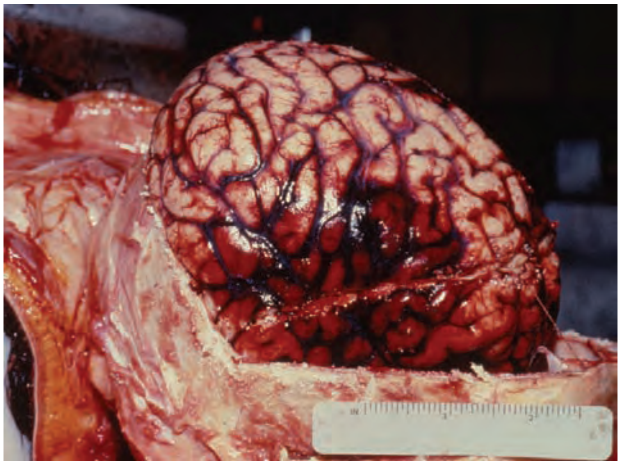

These brains are diffusely swollen with flattened gyral configuration and sulci obliteration. These individuals lived from several hours to several days after the initial insult.

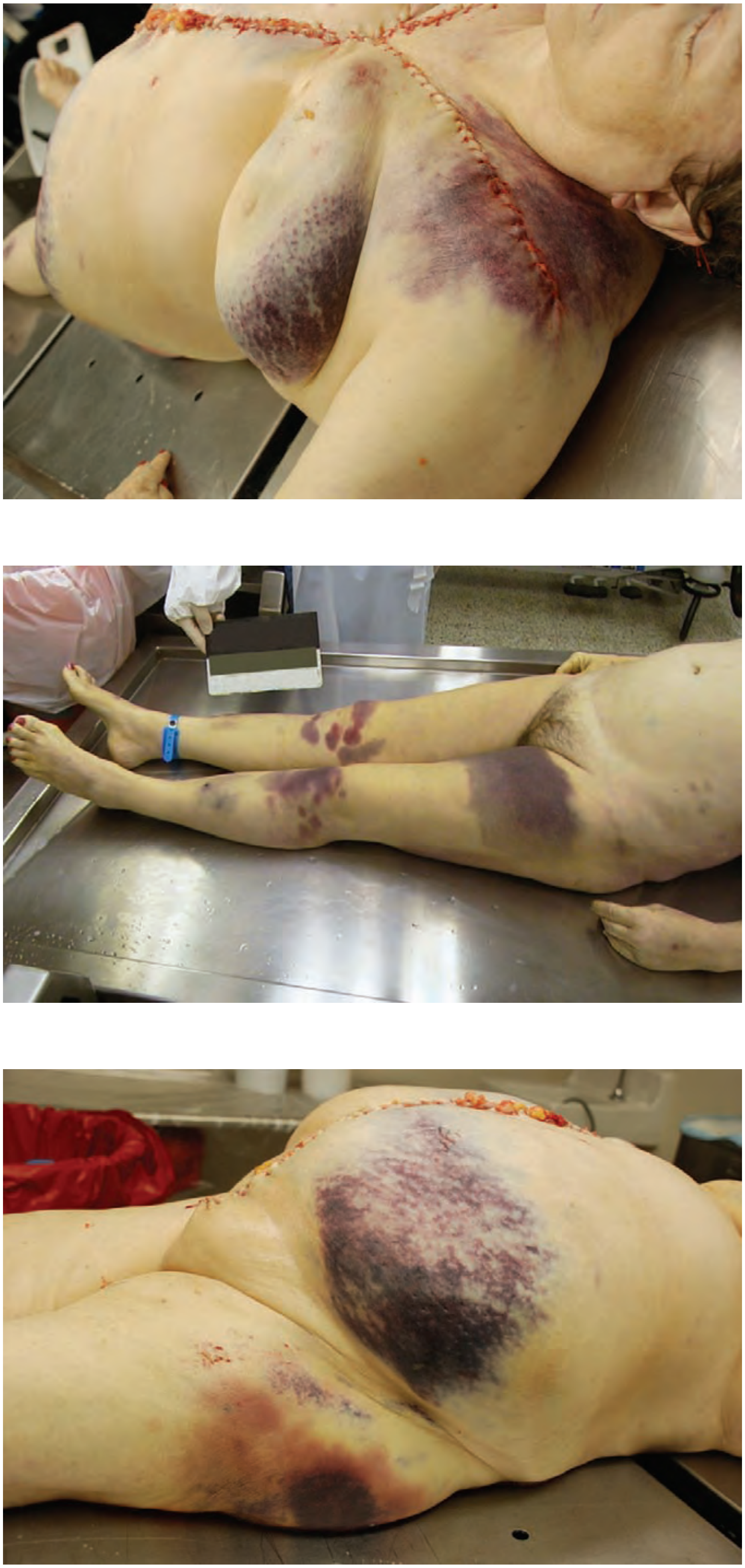

Note the large areas of contusion with ecchymosis following minimal trauma in these individuals with chronic alcoholism, hepatic cirrhosis, and coagulopathy.

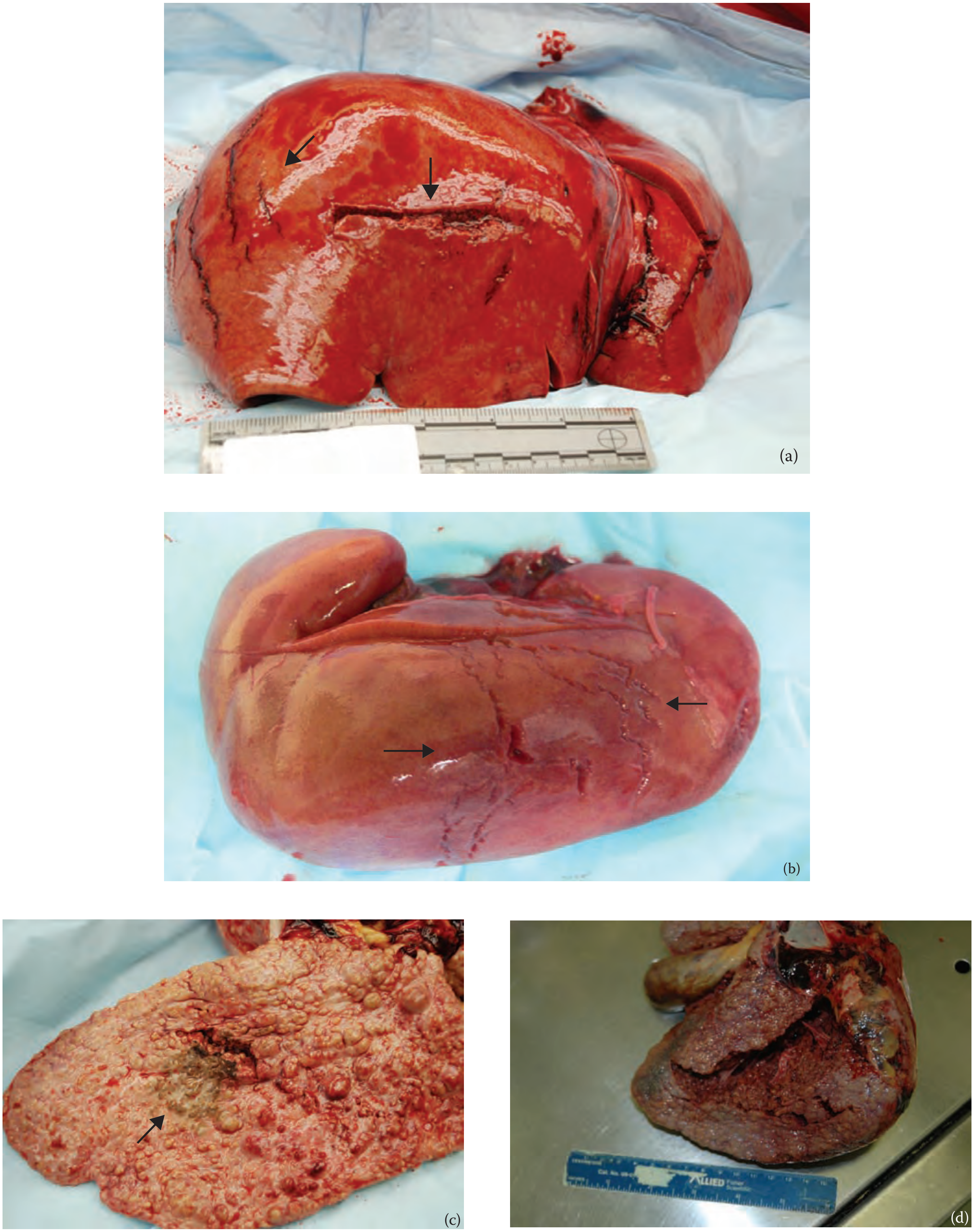

Lacerations of liver and kidney due to blunt-force trauma. Note that (c) and (d) have hepatic cirrhosis, which is less commonly associated with laceration due to fibrosis. A normal liver is the most common organ in the peritoneal cavity to lacerate (a) in association with blunt-force trauma.

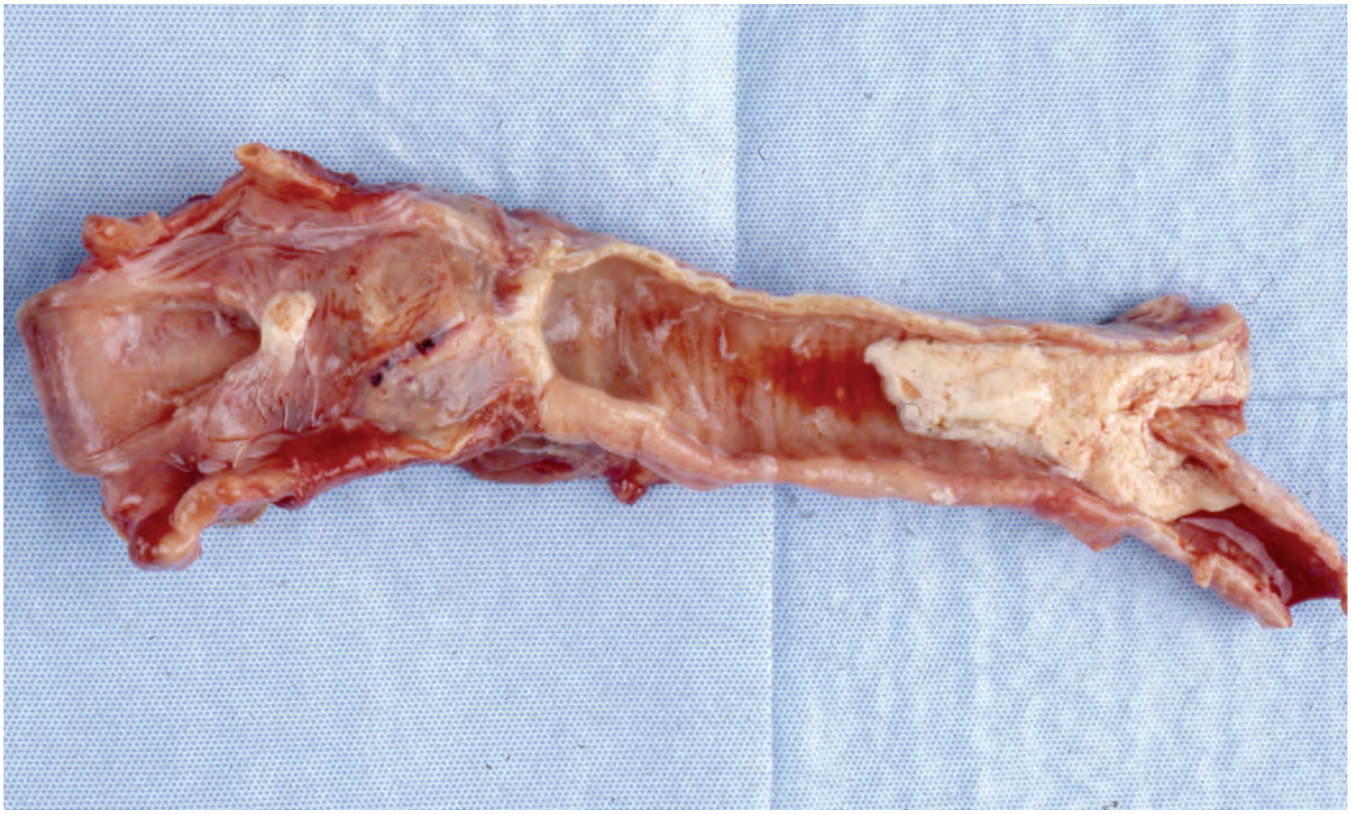

This opened airway demonstrates a large piece of aspirated brain in an individual who was resuscitated following extensive blunt-force head trauma with comminuted skull fracture and communicating lacerations extending into the oral cavity. When Emergency Services arrived, the decedent was found in agonal respirations.

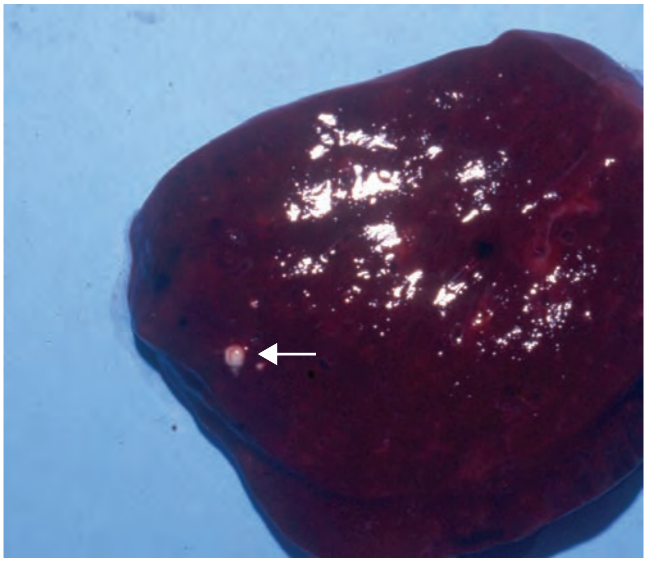

This cut section of lung demonstrates a small white to gray fragment in the parenchyma at the lower lateral aspect. Microscopic examination of this revealed intravascular embolic cerebellar tissue. This individual sustained a comminuted skull fracture with multiple central nervous system lacerations. He was vigorously resuscitated several times over the course of several hours. Brain tissue embolized through damaged blood vessels at the base of the brain.

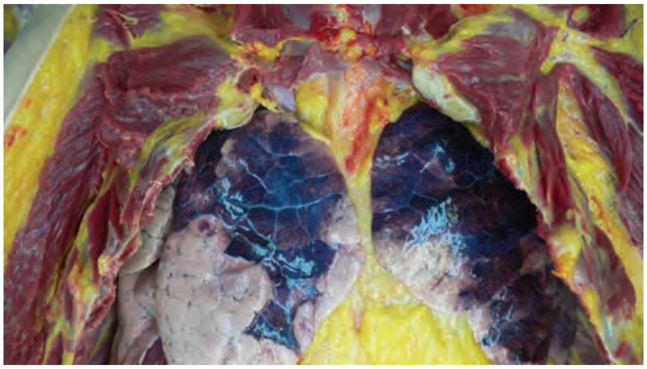

Note the patchy red to purple discoloration indicating aspiration of blood in an individual with a gunshot wound to the head, lacerations extending to the oral pharynx and residual respiratory effort before death.

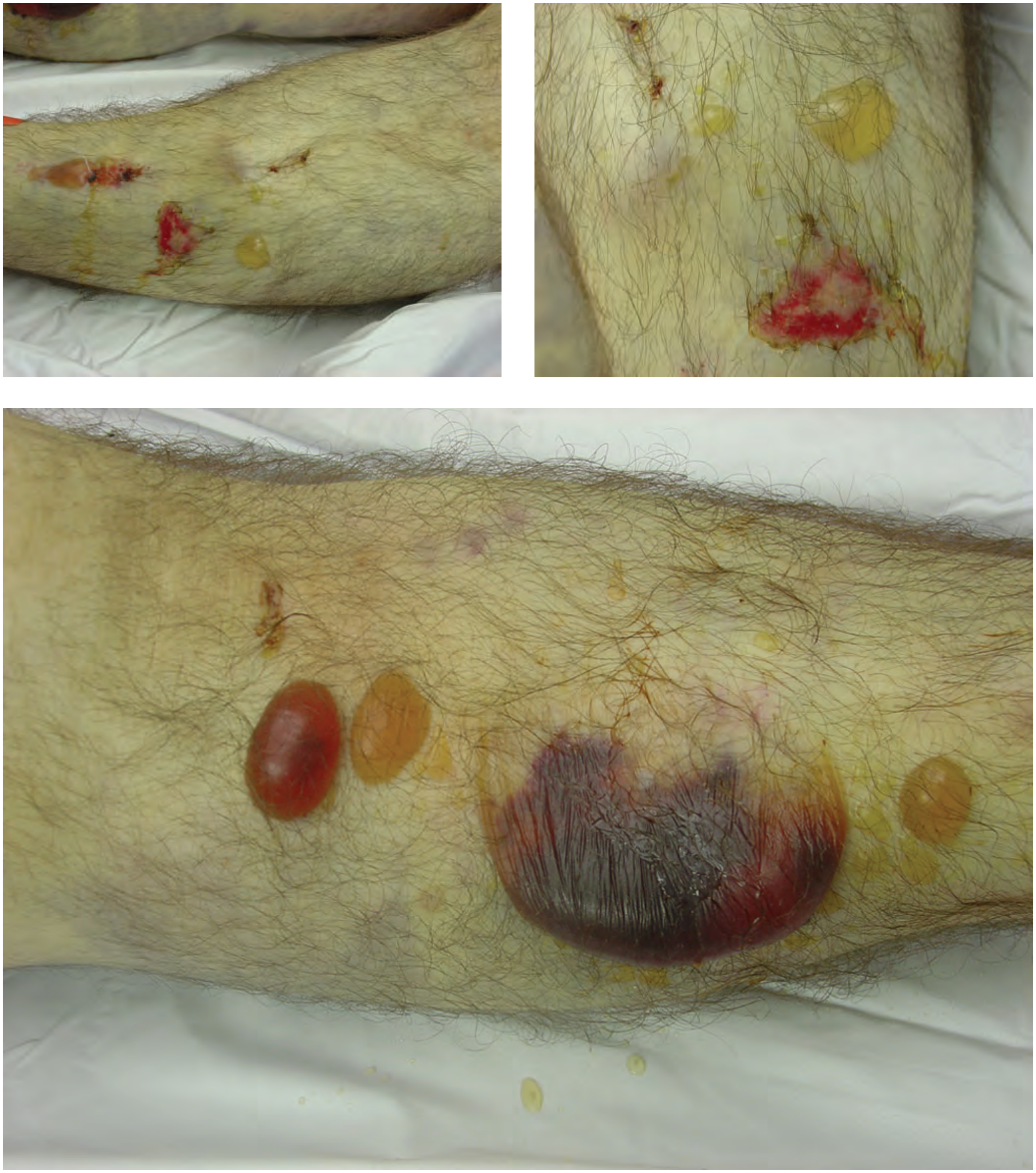

This is another pedestrian who was struck by a car and survived approximately 2 days in the hospital. Note the tibial fracture with adjacent abrasions and broken skin. There is also blister formation confined to this region associated with sepsis following infection associated with this trauma.

Sharp-Force Injuries

7

CHARLES A. CATANESE

Introduction

Sharp-force injuries are defined as injuries produced by an instrument with a thin edge or point. Examples of these instruments include a knife, razor, box cutter, scalpel, sharp-edged piece of metal, broken glass bottle, broken glass window, scissor, ice pick, fork, propeller, screw driver, saw blade, axe, machete, arrow, nail, pick-axe, spiked fence post, meat cleaver, etc.

A *stab wound* is typically made by a knife blade and is defined as having a greater depth of penetration than surface dimension. An *incised wound* is a slicing-type injury where the surface dimension is greater than the depth of penetration.

Accurate, concise, and organized wound documentation is important, as with all other injuries. Each injury should have a documented location on the body, including a description of adjacent abrasions or contusions, wound dimensions, depth of penetration, and direction of penetration into the body. All injured structures should be documented, including the amount of hemorrhage both in the wound track and within body cavities. In cases where there are multiple injuries, it is acceptable to group them with ranges. It is good practice to take overall photographs of the body before and after cleaning; as well as close-up photographs of each wound.

Important aspects concerning interpretation of injury involve pattern recognition. Familiarization with this will allow opinion formulation concerning correlation of a particular instrument to a particular wound. One example involved the arrest of several suspects with different concealed weapons. The police may approach you to render an opinion about what type of weapon produced injuries so they can focus their early investigation. In one actual case, each suspect had a different instrument in his or her pocket, including a flat-edged pocket knife, a box cutter, a screw driver, and a slightly bent serrated table knife. The injuries to the decedent's body consisted of slit-like perforations with multiple adjacent parallel linear abrasions. This pattern injury is consistent with a serrated knife. Many of the images in this chapter are designed to help with pattern recognition.

Location and Direction of Injury

This should be given with reference to a particular body position, usually standard anatomic planes. Each wound should be documented by location on the body's surface, and measured from vertical and horizontal planes of reference. An example of this would be from below the top of the head or above the feet, and to the right or left of the midline. Standard anatomic planes are demonstrated with the body in an upright position with the head tilted slightly upward, the legs together, the arms at the sides, and palms facing forward. The head is superior and the feet inferior, medial is toward the midline and lateral away from the midline. The anterior or front of the body includes the face, chest, and the palms. The posterior part of the body includes the back, buttocks, etc. The direction of the wound into or through the body should be given with reference to three planes when possible, front–back, right–left, and up–down. This is important because it allows one to correlate the injuries to possible assault descriptions and help discredit or substantiate statements.

Wound Dimension

This should be documented separately for each sharp-force injury, unless there are many that can be grouped together and described in ranges. Example: There are twenty 1-inch to 2-inch, by up to 1/4-inch, stab wounds within a 5-inch × 7-inch region at the middle aspect of the right chest, which is centered 13 inches below the top of the head and 4 inches to the right of the midline.

It is good practice to document the injury as it exists on the body and then again when in a relaxed state. The important aspect is to document the wound dimension in ranges that most closely reflect the actual dimensions of the knife blade or instrument. A stab wound can be put into a relaxed state by pressing the surrounding skin toward the wound and releasing the surrounding tension, or by placing tape over the perforation site to approximate the margins. It is also acceptable to cut a square around the surrounding skin to release the tension. The skin and underlying tissues are elastic, with different degrees of tension. This can make the surface dimension of the wound length and width slightly variable depending on location and orientation of the body. Cut marks through bone and cartilage may reflect accurate weapon dimension with tool marks and should be retained in formaldehyde.

This same concept applies for the depth of penetration, as well. Dimensions should be given in a range to account for changes in body position when examined on the autopsy table compared with the body position when the injuries were inflicted. Variables that may change this parameter while the assault is taking place include deep breaths or exhales, flexion, extension, rotation, the force used to inflict the injury, the location on the body including underlying bone or soft tissue, etc.

Adjacent Abrasions and Contusions

These may indicate body contact from the knife handle, lower part of the knife blade, or the knife hilt. This is important information in formulating an opinion as to how much force was needed to produce the injury. If the knife blade penetrates a bone and there are hilt marks adjacent to the perforation site, one can extrapolate that the knife must have been stuck into the body with great force.

Injured Organs or Structures

This information helps to allow interpretation of how functional one might be after an assault. The number and extent of internal injuries give insight into the nature of the assault and perpetrator.

For instance, “I stabbed him in the neck and he continued to chase me so I stabbed him again in the chest and abdomen.” If the neck wound injured the spinal cord, you know the statement is false because the victim would no longer have voluntary movement to part of the body, as with a tendon or peripheral nerve’s being cut. Also, depending on the structures damaged, the rate of blood loss may be quite variable. This would help define how fast the individual would lose blood, and what the individual might be capable of doing after the injury, and for how long. Questions like this may often come up in trial. A transected aorta would incapacitate someone more rapidly than a transected brachial (arm) artery, which would be more rapid than a transected cephalic (superficial) vein. Someone with a stab wound to the heart will often lose consciousness within minutes but still be capable of running away, defending themselves, or continuing an assault, particularly if the heart does not go into a lethal arrhythmia. It is possible for someone to get stabbed in the heart, run several blocks and shoot several people before dying.

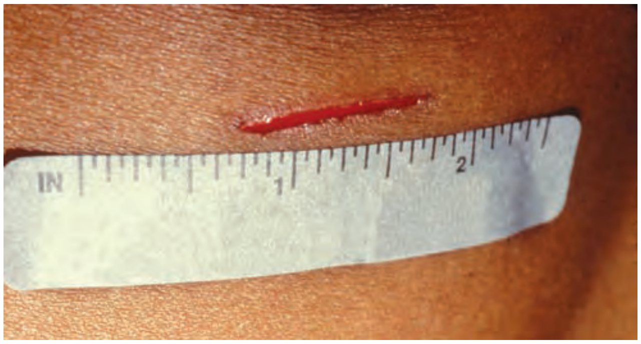
Stab wound consisting of a flat slit-like perforation with minimal surrounding skin tension.

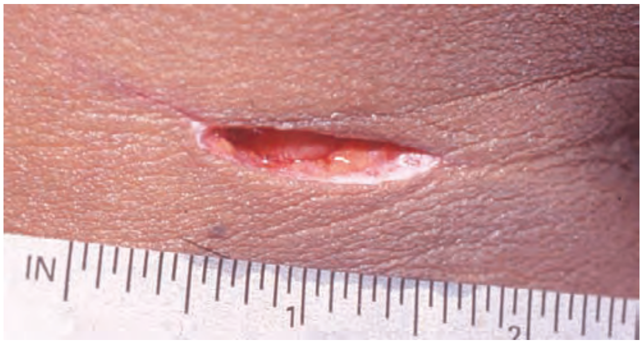
Stab wound with moderate surrounding skin tension. Note the separation of the wound margins.

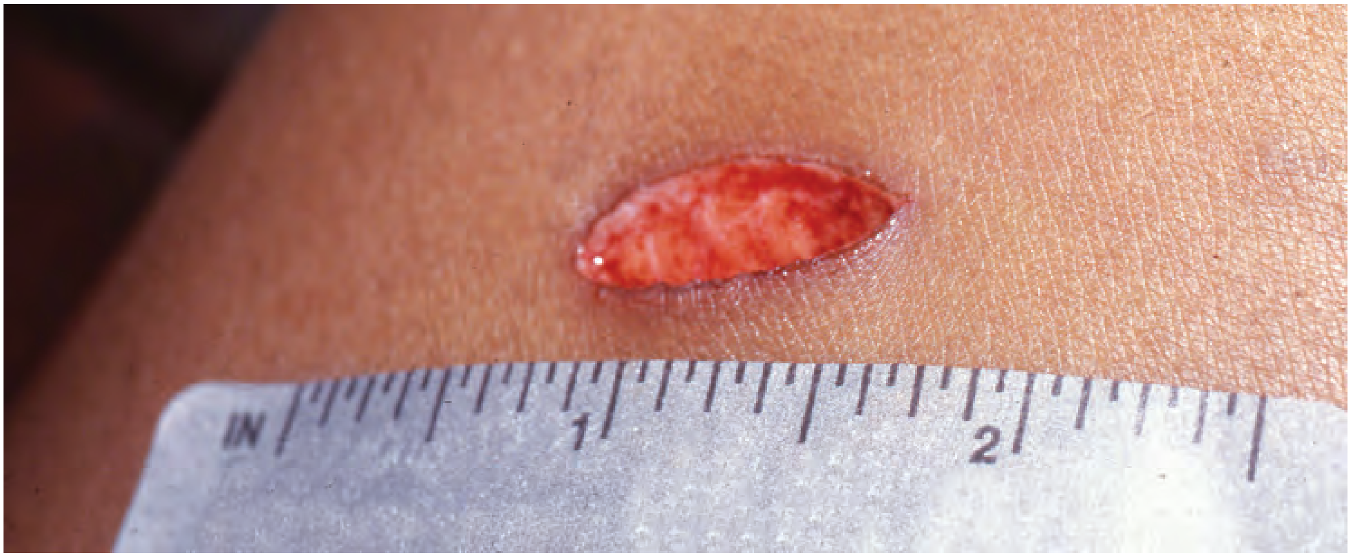
Gaping stab wound consisting of an oval perforation caused by significant surrounding skin tension.

A non-serrated knife blade similar to the one used to inflict these wounds. It is important to take the surrounding skin tension into account when estimating the size and type of weapon used.

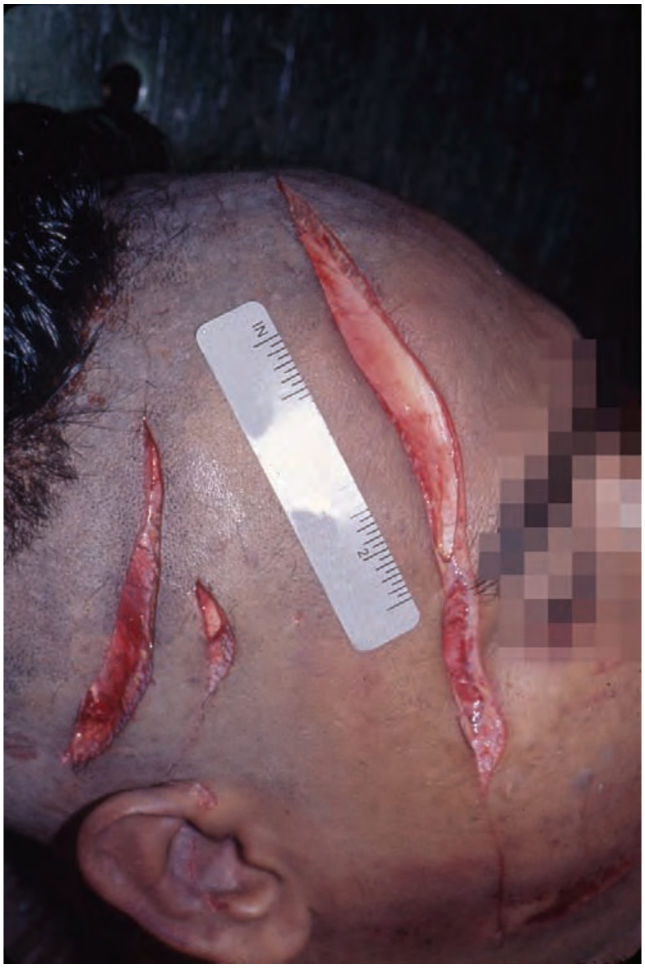

Incised wounds produced by a broken bottle.

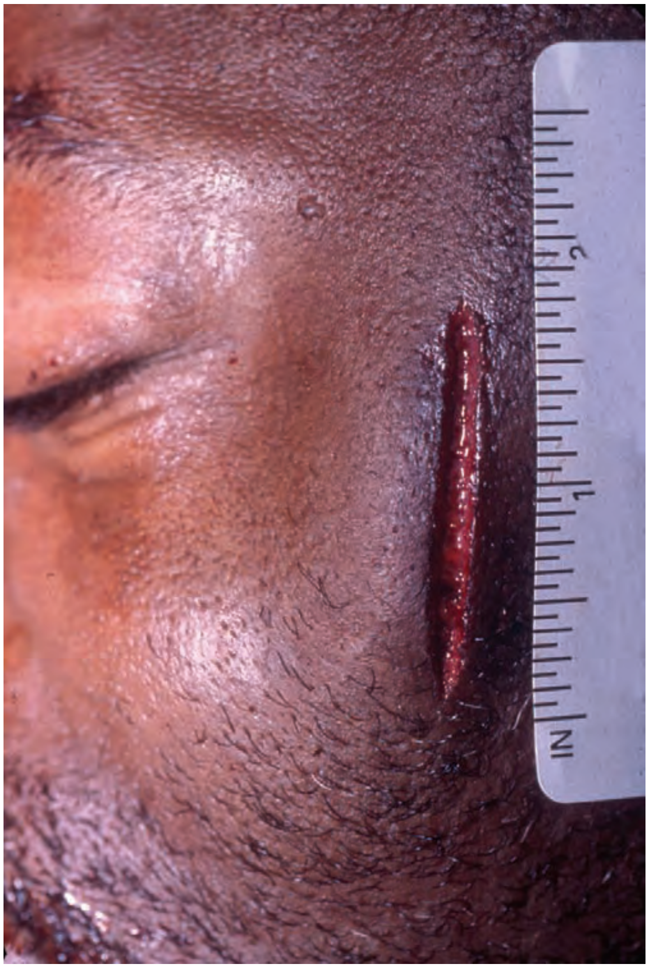

Incised wound to the face produced by a knife.

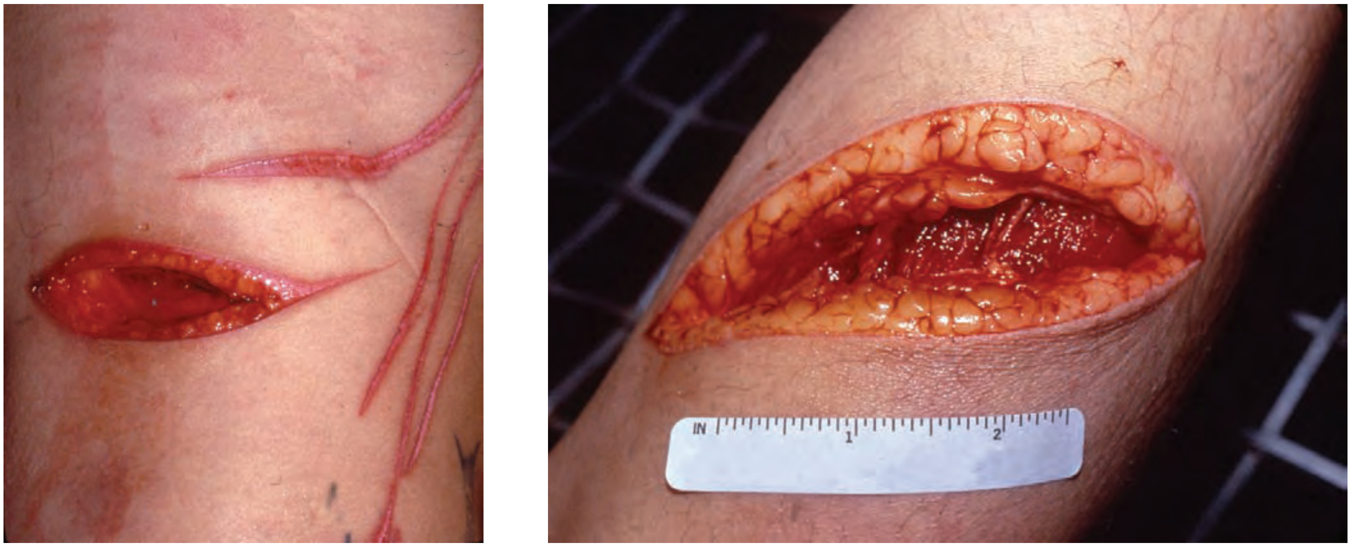

Incised wounds produced by a box cutter.

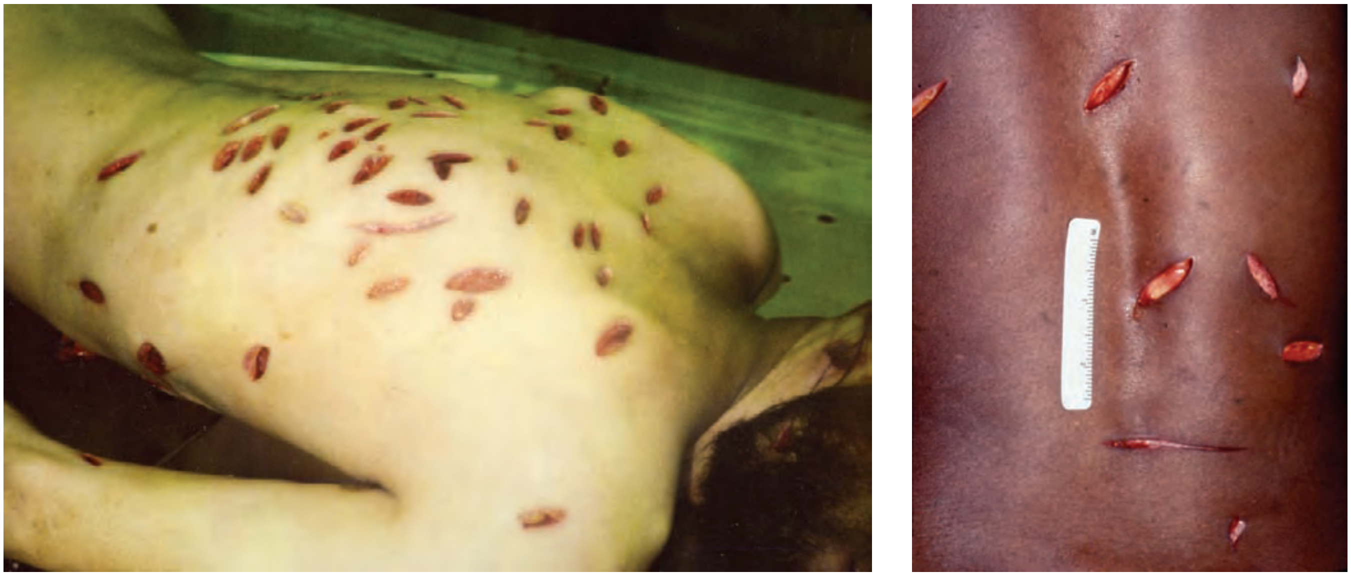

These are multiple stab wounds to the back. Note the varying dimensions and gaping nature due to the varying degrees of skin tension from underlying tissue planes in different locations.

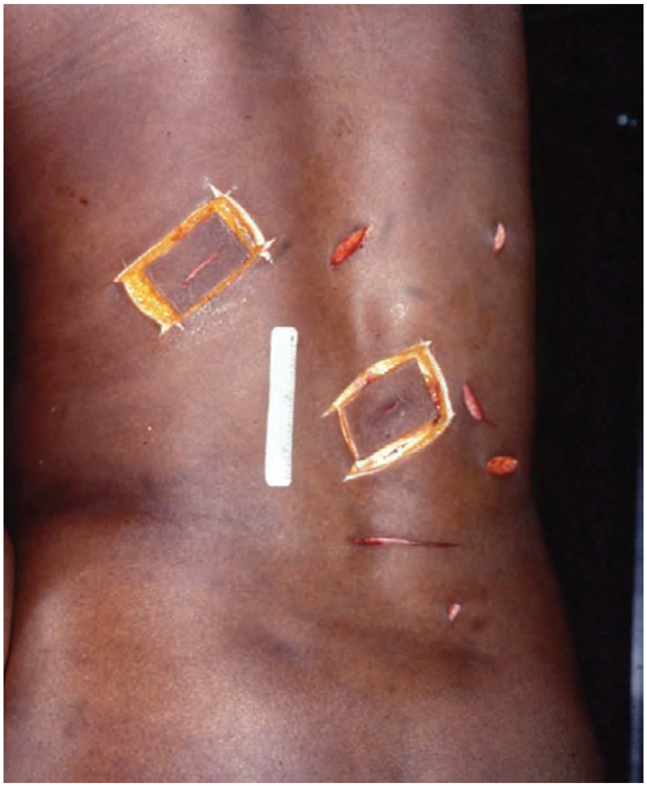

Several of these wounds were chosen based on estimation that they were more reflective of the actual knife blade size. These wounds were then cut into a relaxed state and re-measured to help estimate the actual size of the knife blade.

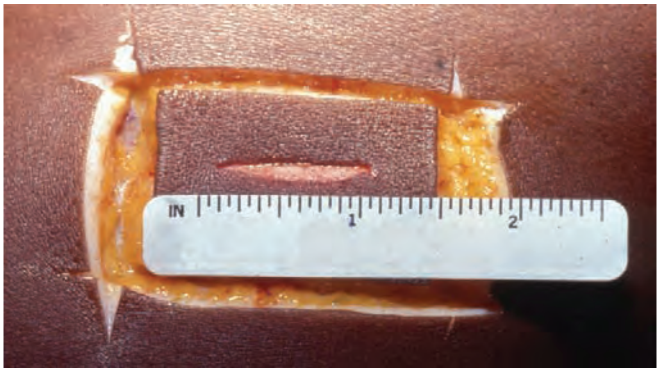

This is a close-up of one of the wounds that were cut into a relaxed state and measured. Note the sharp end at the left aspect and the blunt end at the right aspect.

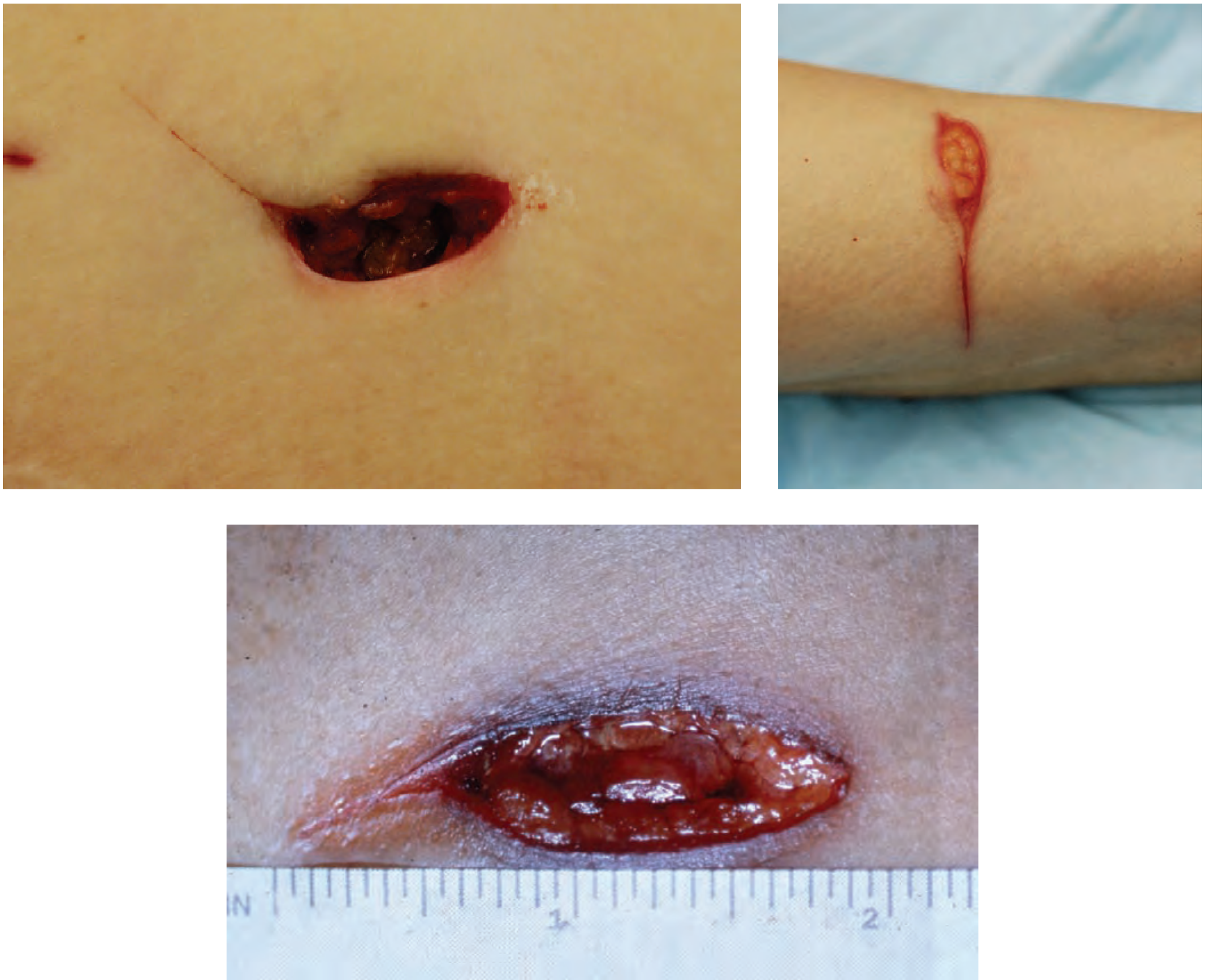

Stab wounds to body with slight drying of the margins and a "linear extension" from one end due to the sharp part of the knife blade being dragged through the wound and across the body surface, changing the dimension of the injury.

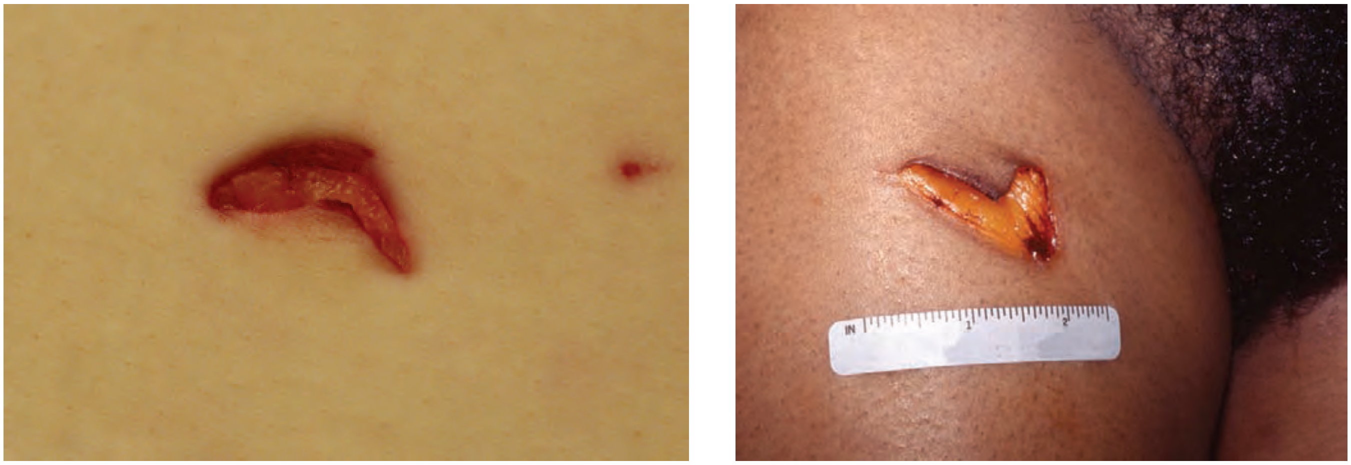

These two images demonstrate a "swallowtail"-type injury where the knife was inserted into the body in one direction, turned and then removed in another direction, producing two angulated cuts. This type of injury may be produced by the movement of the perpetrator, victim, or both.

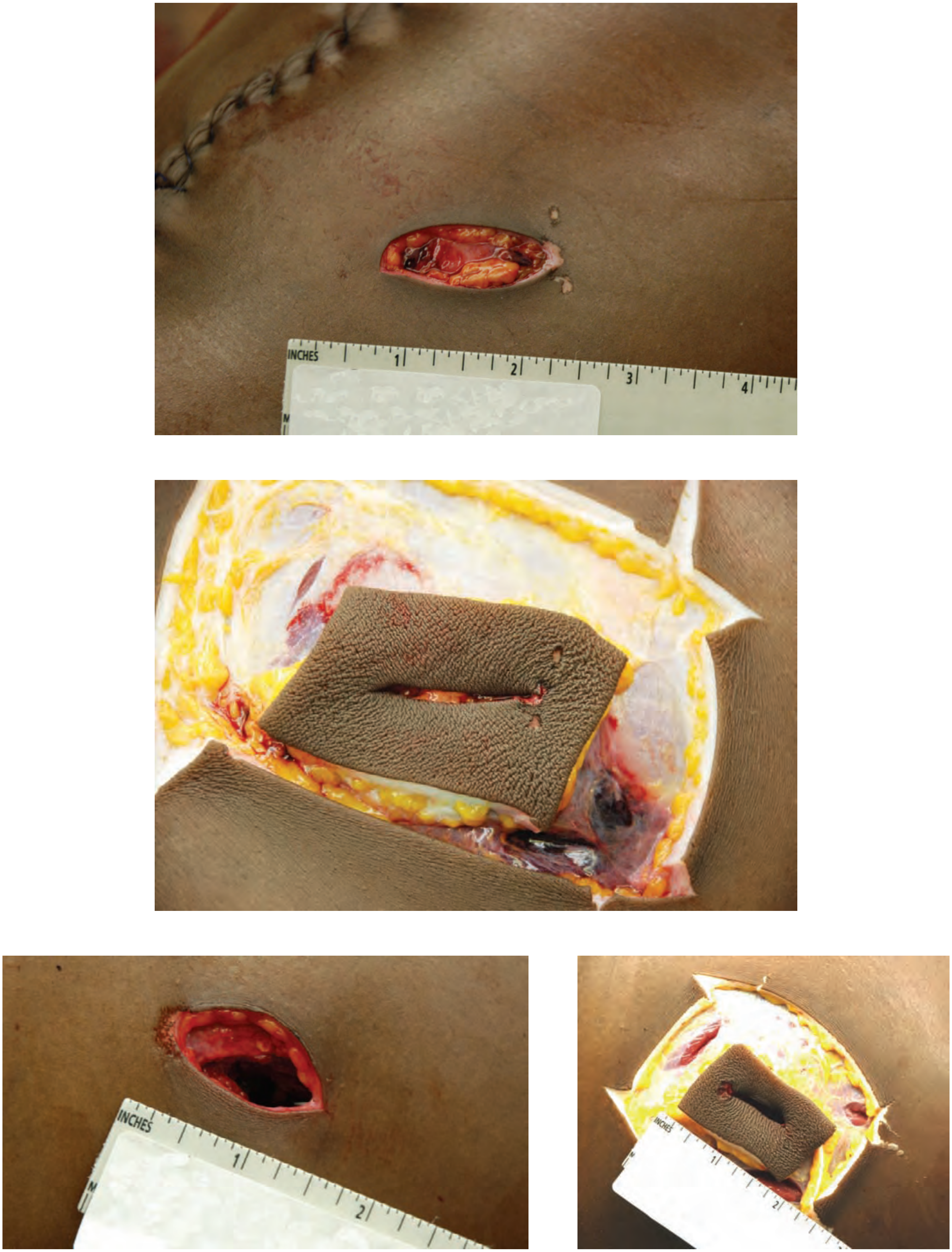

These stab wounds demonstrate adjacent abrasions from the hilt of the knife contacting the skin with great force.

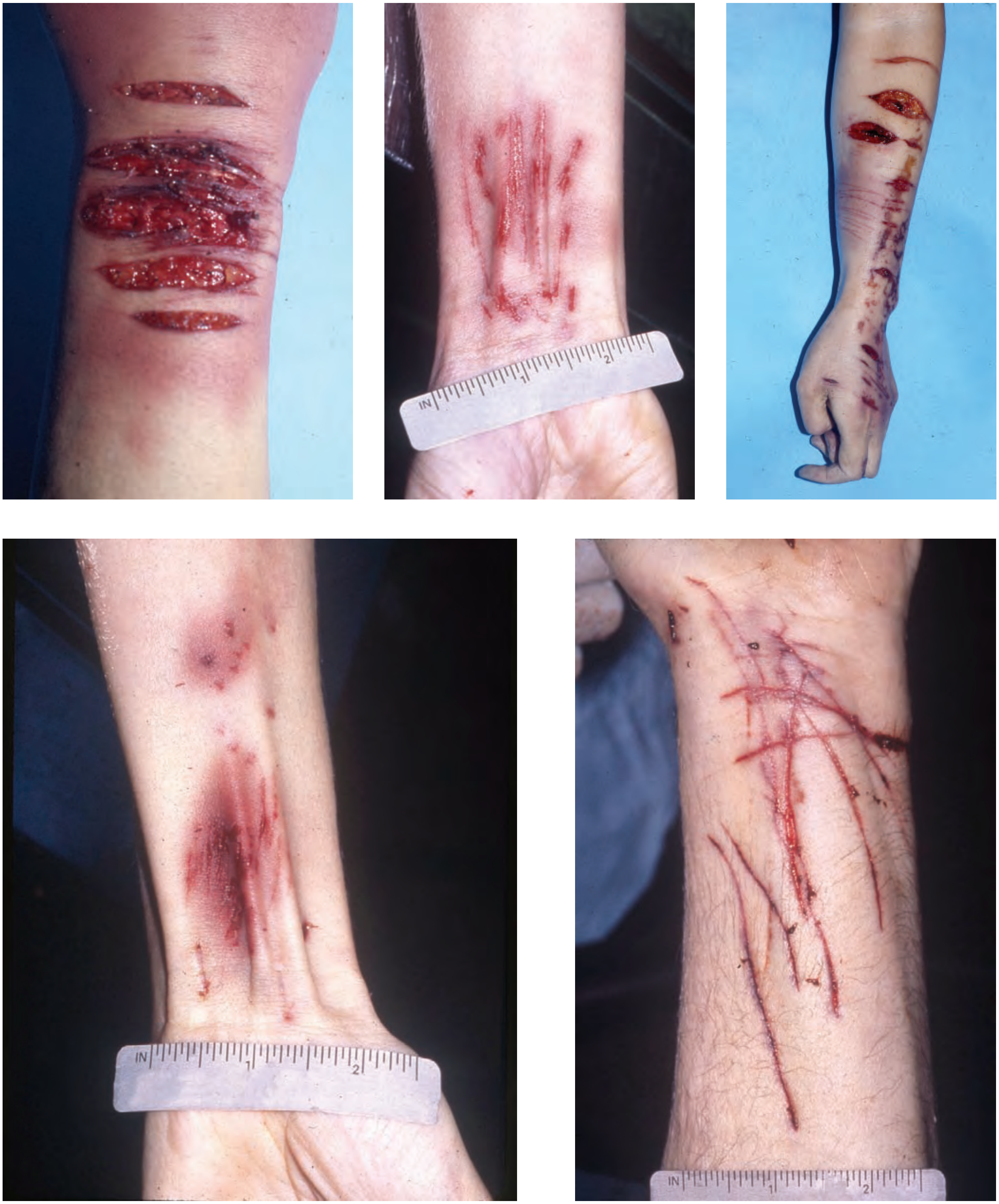

Multiple self-inflected linear abrasions and incised wounds to the arms and wrists demonstrating hesitation marks in individuals who committed suicide with other lethal injuries.

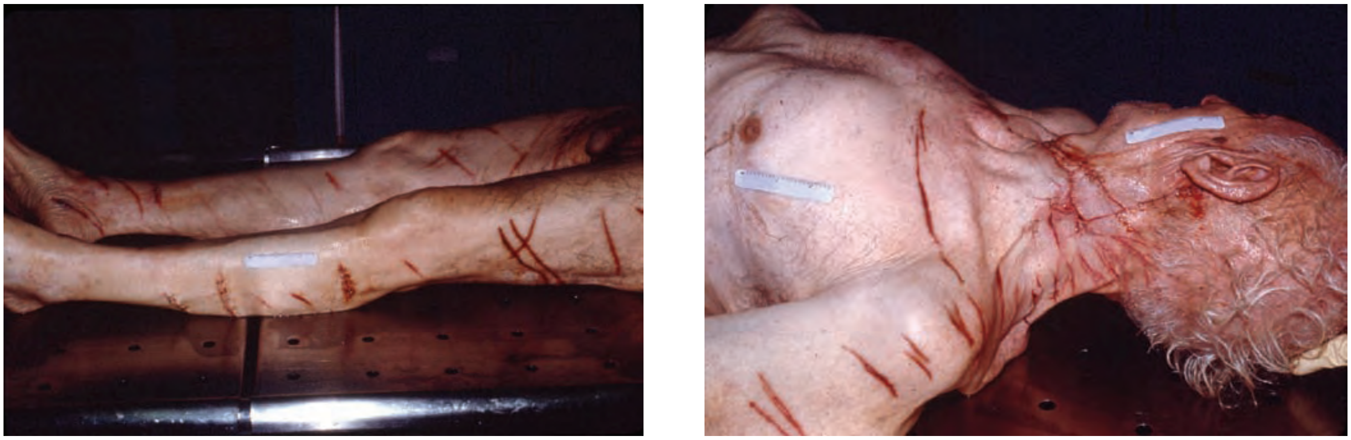

Multiple linear incised wounds in an individual who lived for several hours. He died of a heart attack associated with the stress and blood loss from the self-inflicted injuries. The manner of death in this case was classified as suicide. The cause of death was listed as both multiple incised wounds and heart disease.

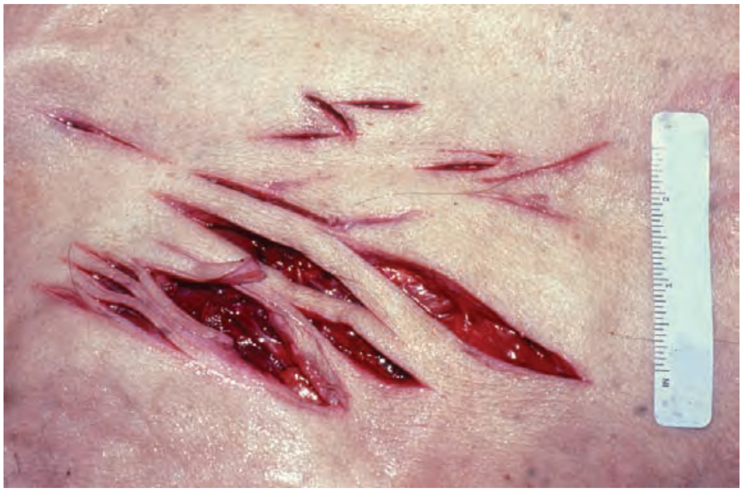

Multiple self-inflicted sharp-force injuries including stab and incised wounds. Note the tightly clustered orientation of these injuries, which is common in suicides.

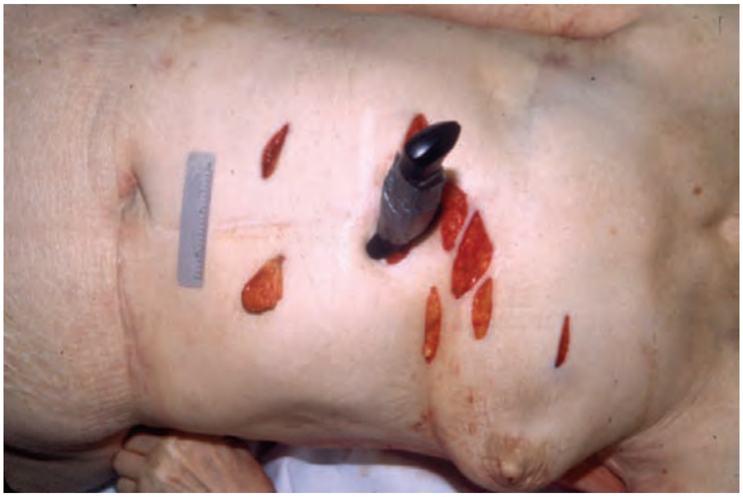

Multiple self-inflicted stab wounds. This individual was found with her hands surrounding the knife handle in a moderate state of rigor mortis.

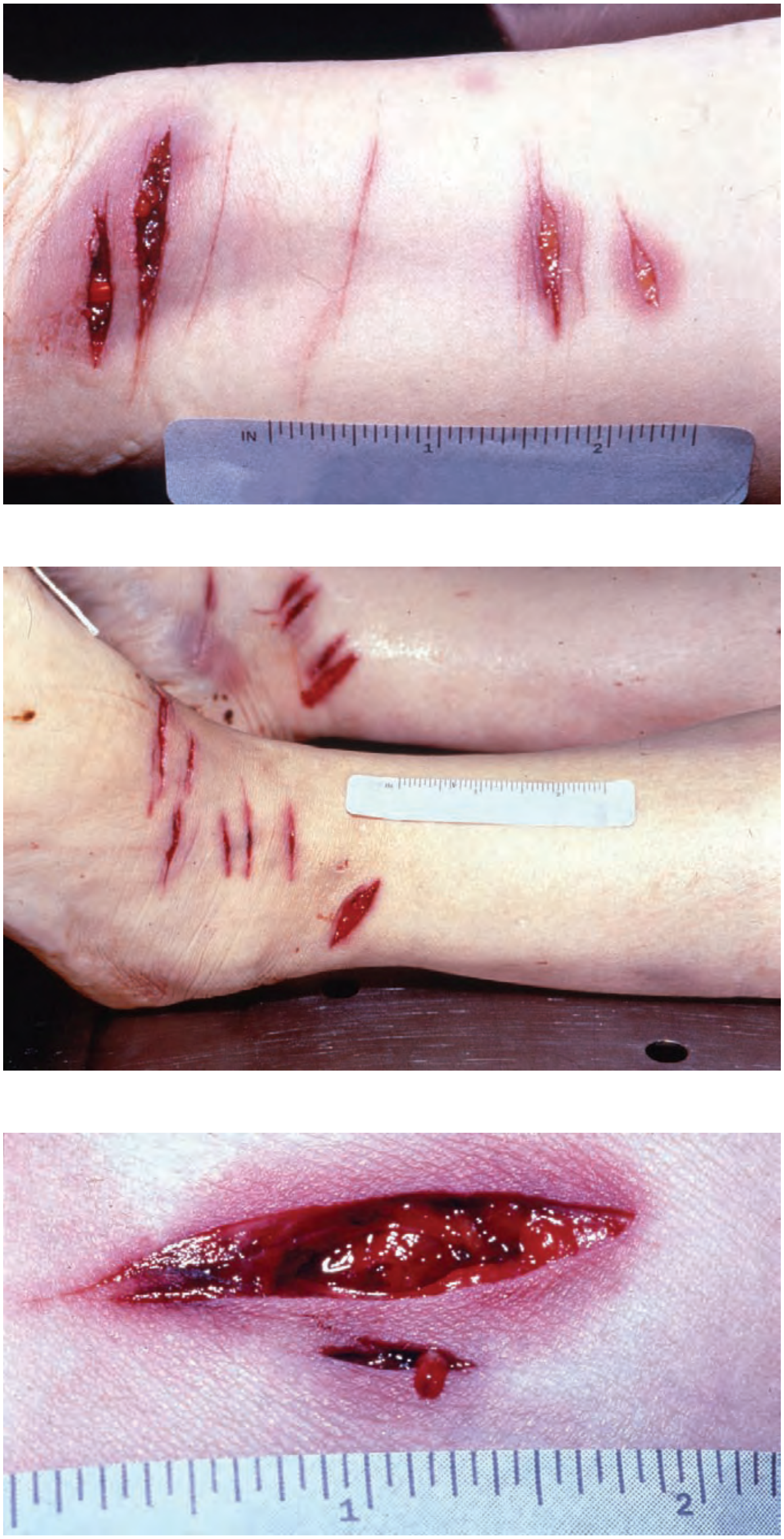

This individual was found with multiple clustered linear abrasions, incised wounds, and stab wounds to the wrists, arms, and ankles. She first attempted suicide by ingesting large quantities of acetaminophen, then inflicted the above sharp-force injuries. Note the erythema around the wounds, indicating vital reaction. There was blood noted all over the apartment. The individual later jumped out a window and died from blunt-impact injuries. The wounds shown are hesitation wounds.

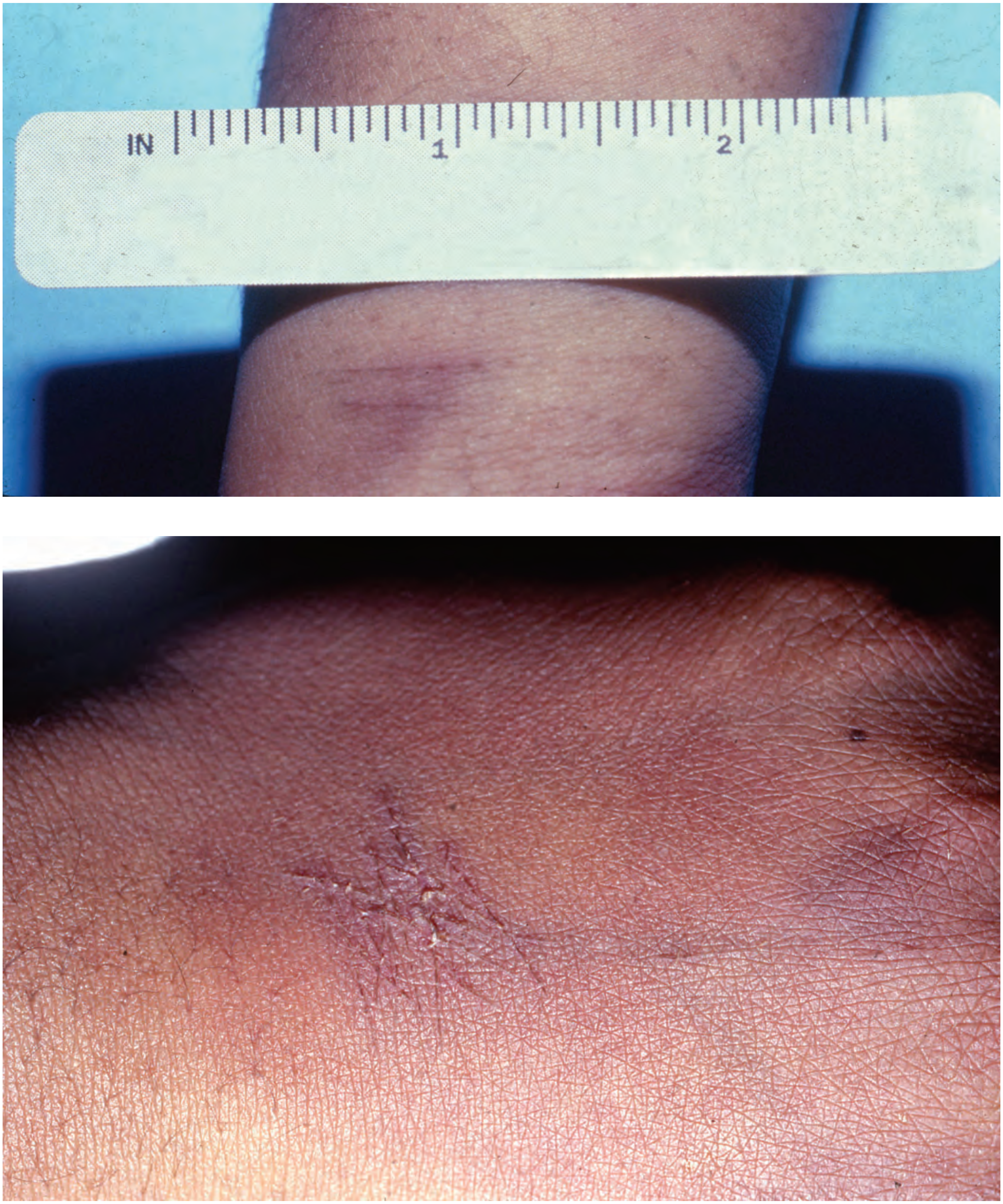

Superficial linear abrasions caused by dragging a knife blade across the wrist and hand prior to committing suicide by hanging.

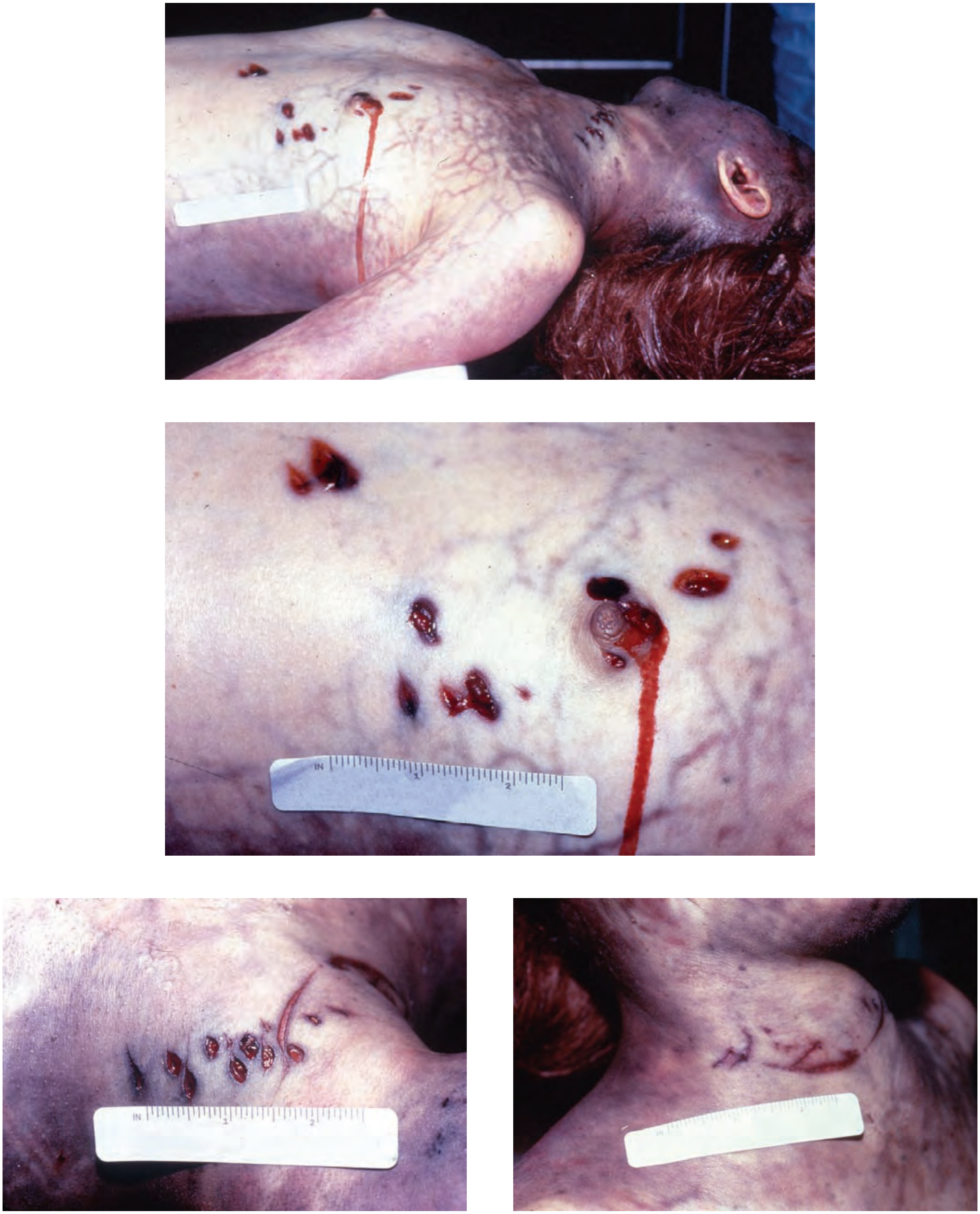

Multiple clustered self-inflicted incised and stab wounds located on each side of the neck and chest in an individual who committed suicide. Most of these stab wounds were very superficial in depth of penetration and can be described as hesitation marks. Three of these stab wounds to the chest entered the thoracic cavity and penetrated the heart and lung, causing death. Note the slight degree of decomposition with wound margin drying.

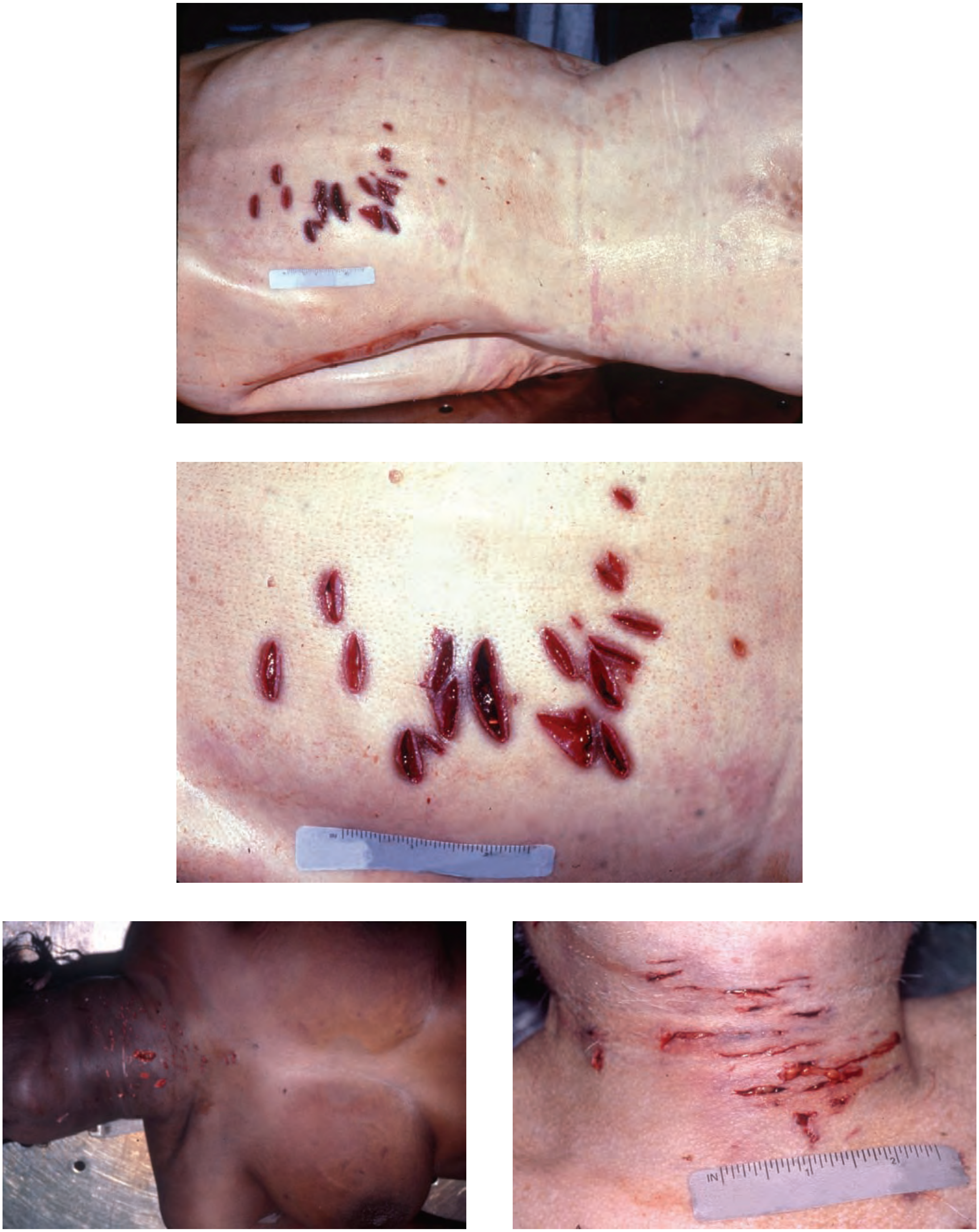

Multiple clustered homicidal stab wounds in individuals who were also bound during a sexual assault. Note the location of the multiple stab wounds to the middle aspect of the individual's back, making it virtually impossible for them to have been self-inflicted.

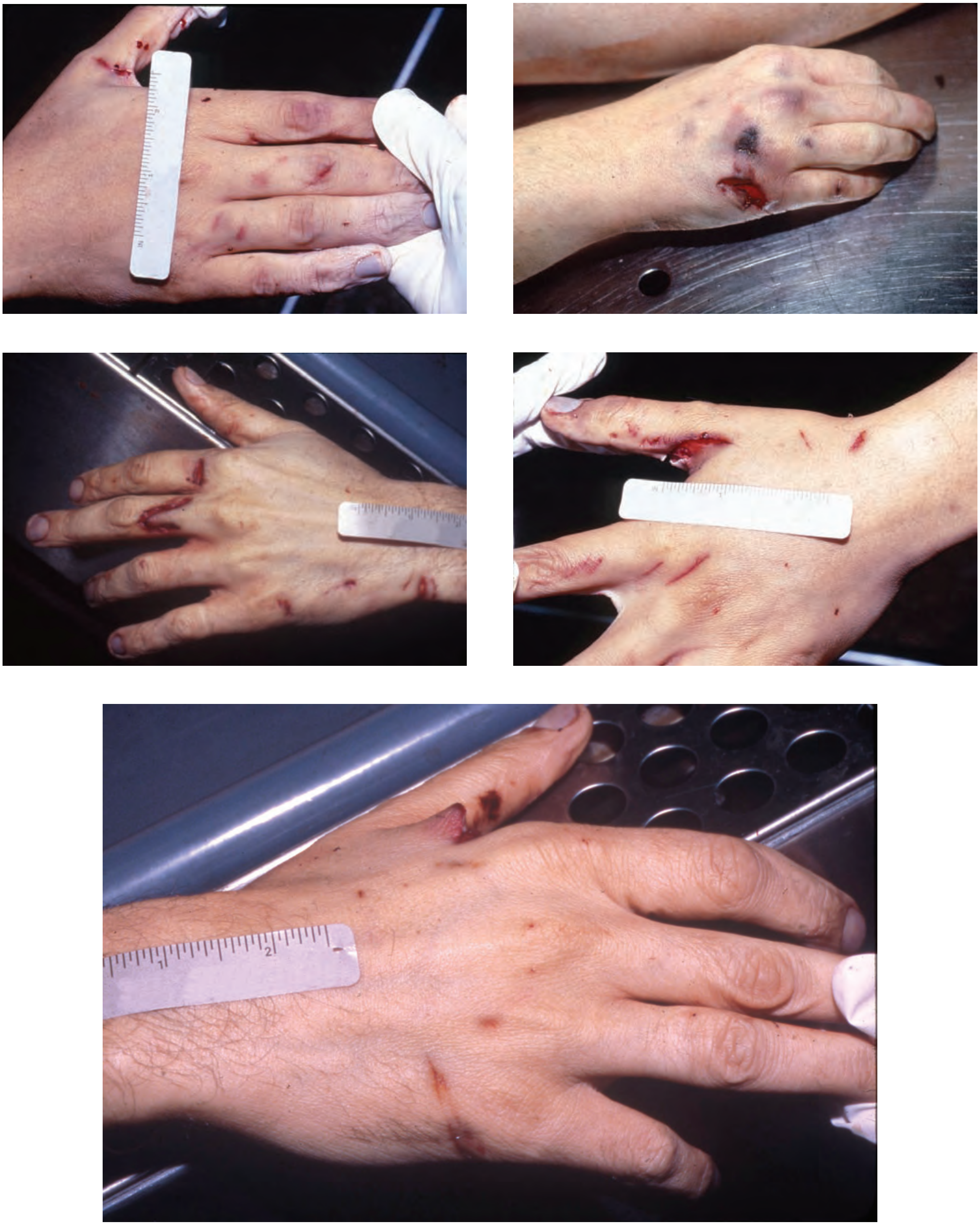

These are multiple examples of defensive-type wounds from individuals who were attempting to ward off an assault from a knife-wielding attacker. Note the multiple linear abrasions, incised wounds, and stab wounds to the hands. Several of these injuries show cut marks between the thumb and fingers, due to the individual's attempting to grab the knife blade.

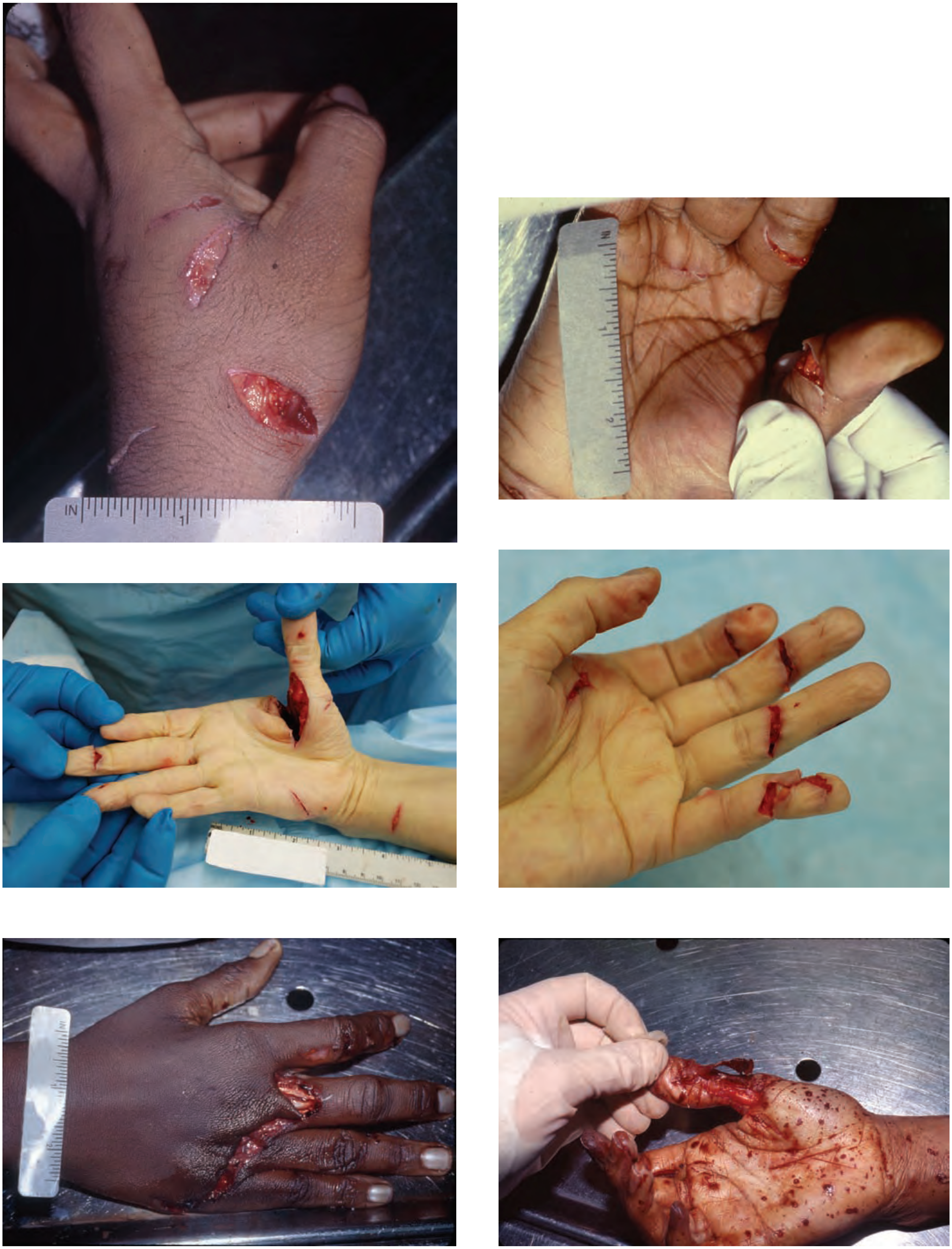

Other examples of wounds consistent with being inflicted while an individual is fending off an attack.

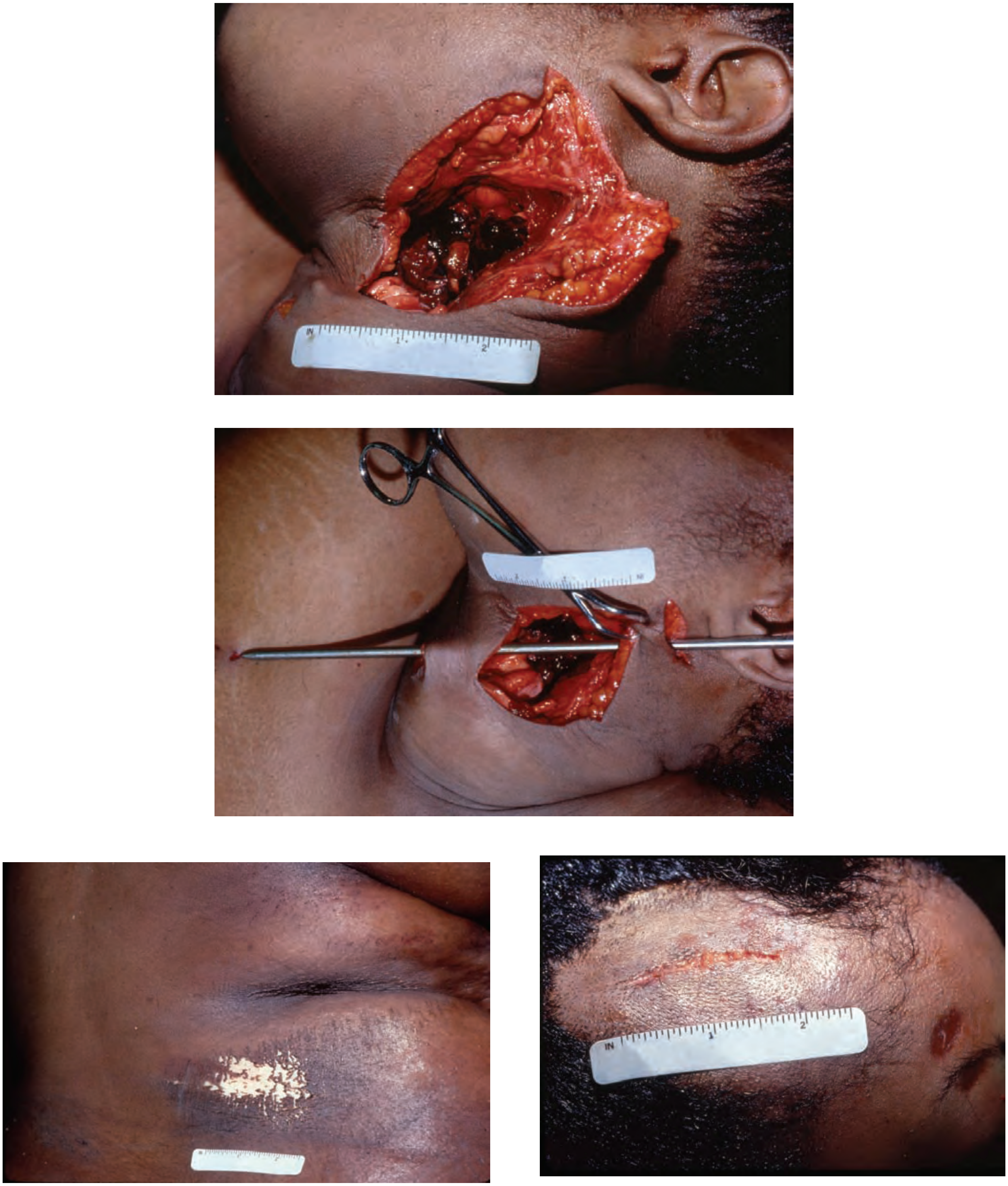

This individual was stabbed in the neck and then thrown from a window. The perpetrator described the injuries as being self-inflicted by the decedent who cut herself and then jumped out a window. The window ledge and outside wall were smeared with blood, which was easily visible from below. During autopsy, approximation of the margins revealed a stab wound just below the left ear from a knife blade that exited through her neck and stuck in her right shoulder. There was also an anemic abrasion to her back and an anemic laceration to the scalp, indicating the decedent had lost much of her blood volume and did not have a beating heart with blood pressure enough to produce hemorrhage from the injuries sustained when striking the ground.

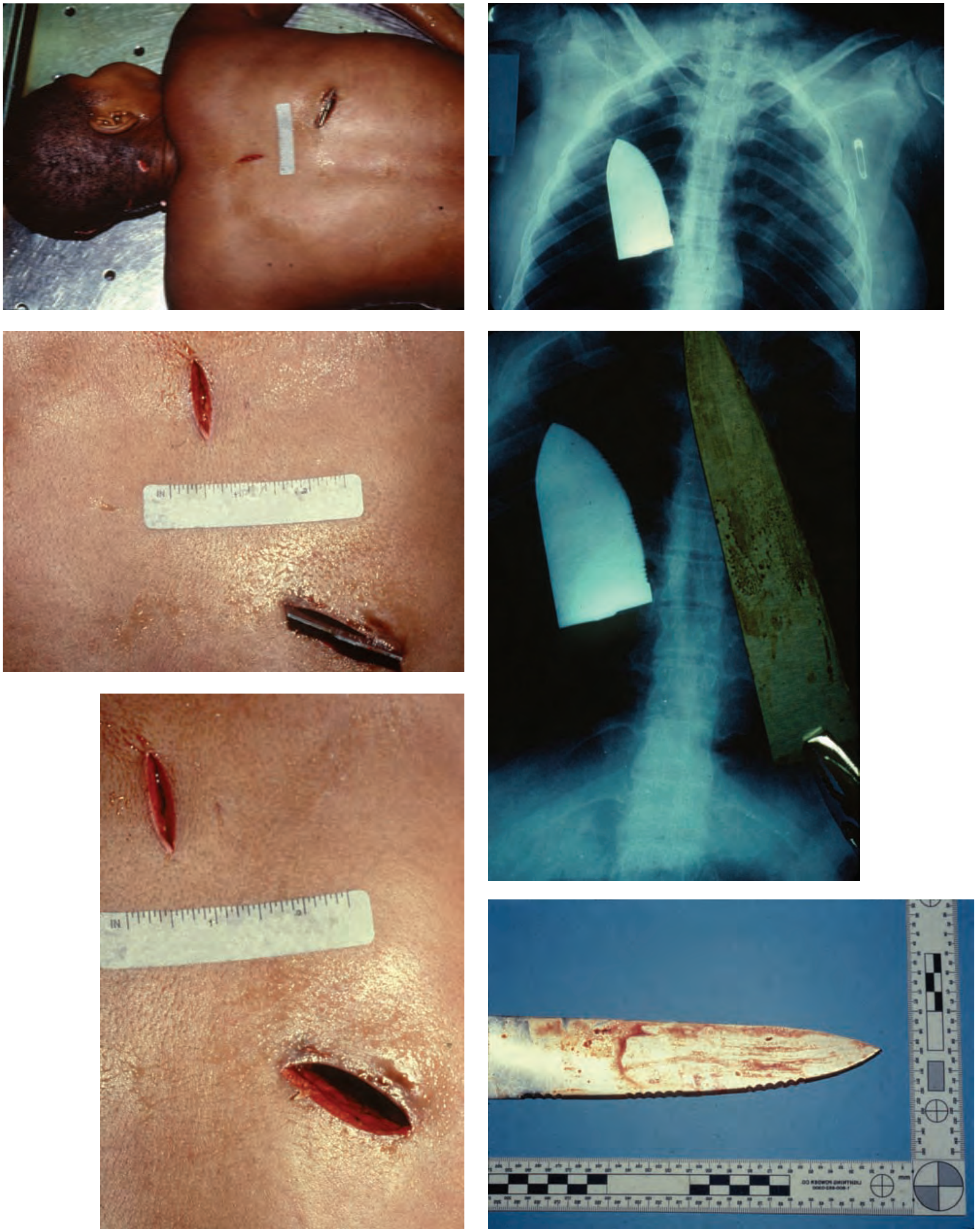

Note the two stab wounds to the left back and the broken knife blade extending from one of the wound tracks. Note the close-up view of the back, demonstrating the knife within the wound track and then with the knife removed. Note the change in the wound dimension after the knife was removed. This demonstrates the elastic nature of skin and the possible challenges of correlating a weapon to a particular wound dimension. Also note the x-ray, which demonstrates the knife blade within the body. Note the dimension of the actual knife blade compared with the dimension of the blade image on the x-ray. This is due to the angle in which the x-rays contacted the knife blade in the thoracic cavity. Weapon.

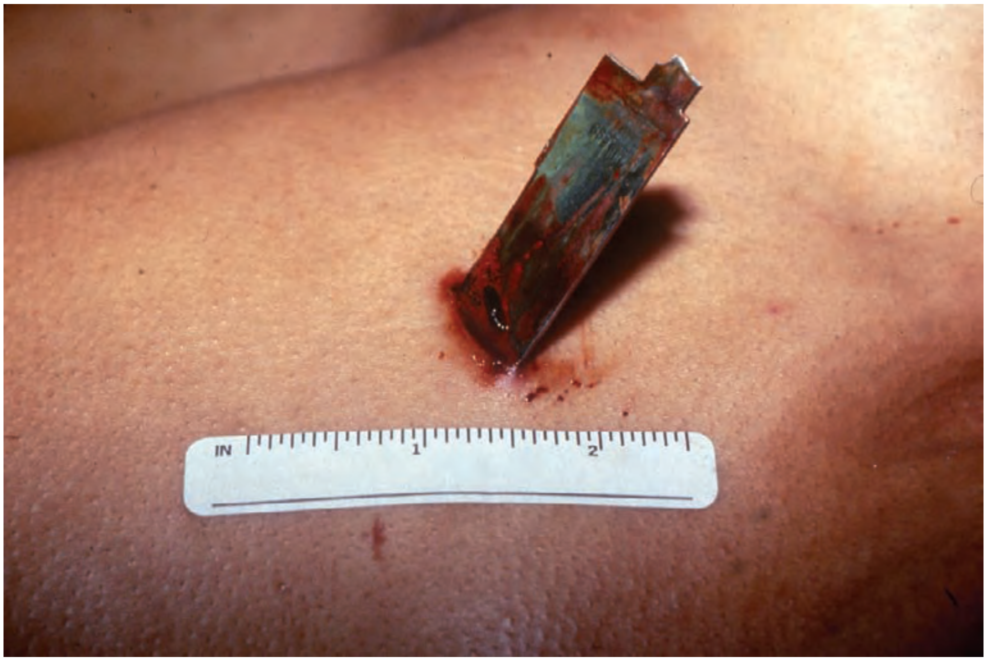

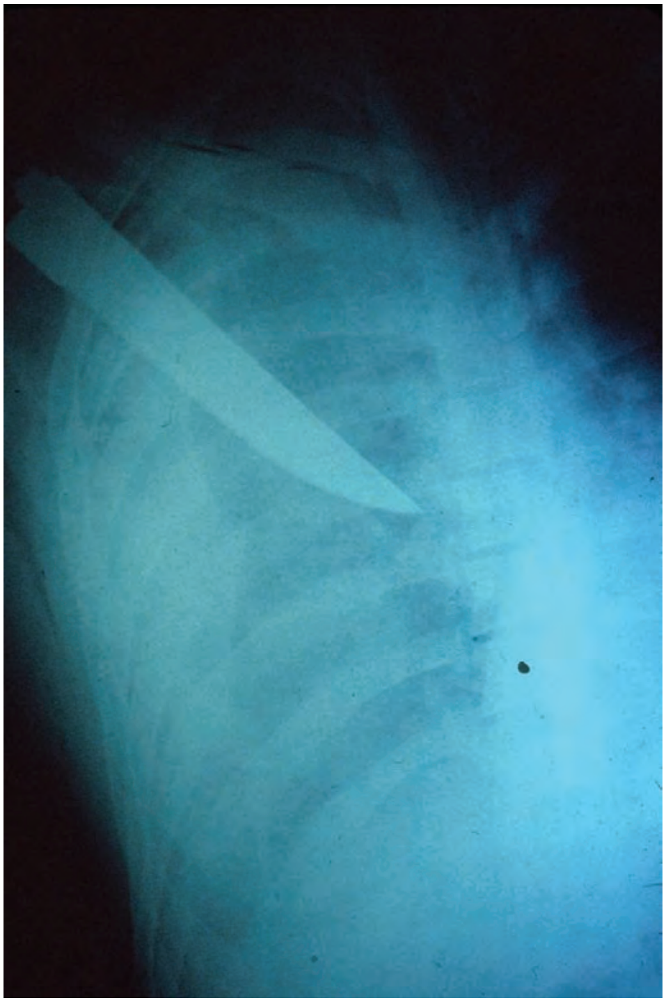

Broken knife blade left in body following a homicidal assault. This knife was stuck into this individual with great force and the blade got stuck in a bone. When the perpetrator attempted to remove it, the handle broke off and the knife blade was left behind.

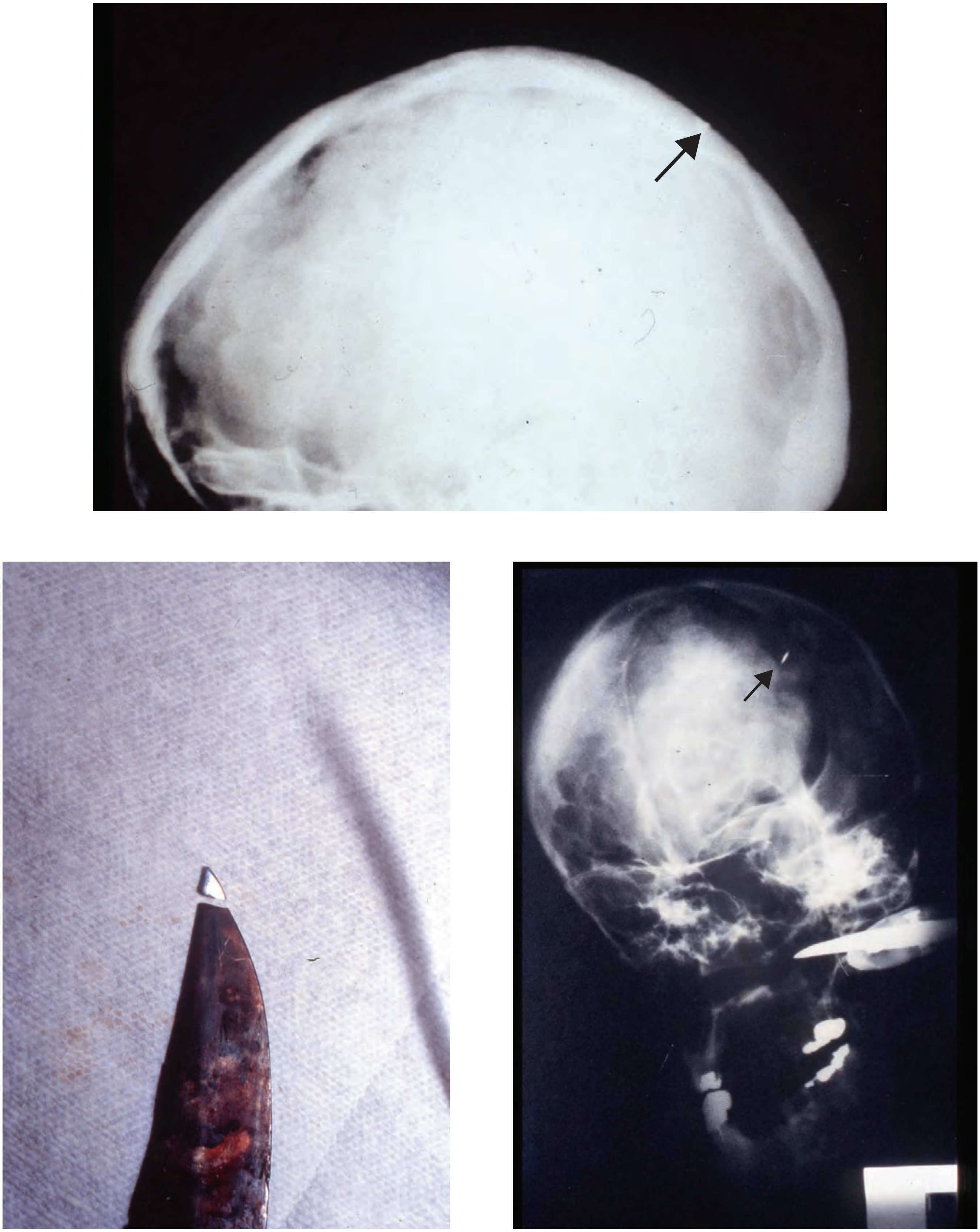

This individual was stabbed more than 75 times by a scorned lover. The tip of the knife blade broke off in the skull during this violent assault. Note the small radio-dense fragment depicted in these skull x-rays. This reinforces the notion that it is good practice to x-ray all sharp-force homicidal violence cases. A fragmented portion of knife blade can later be matched to a suspect's knife.

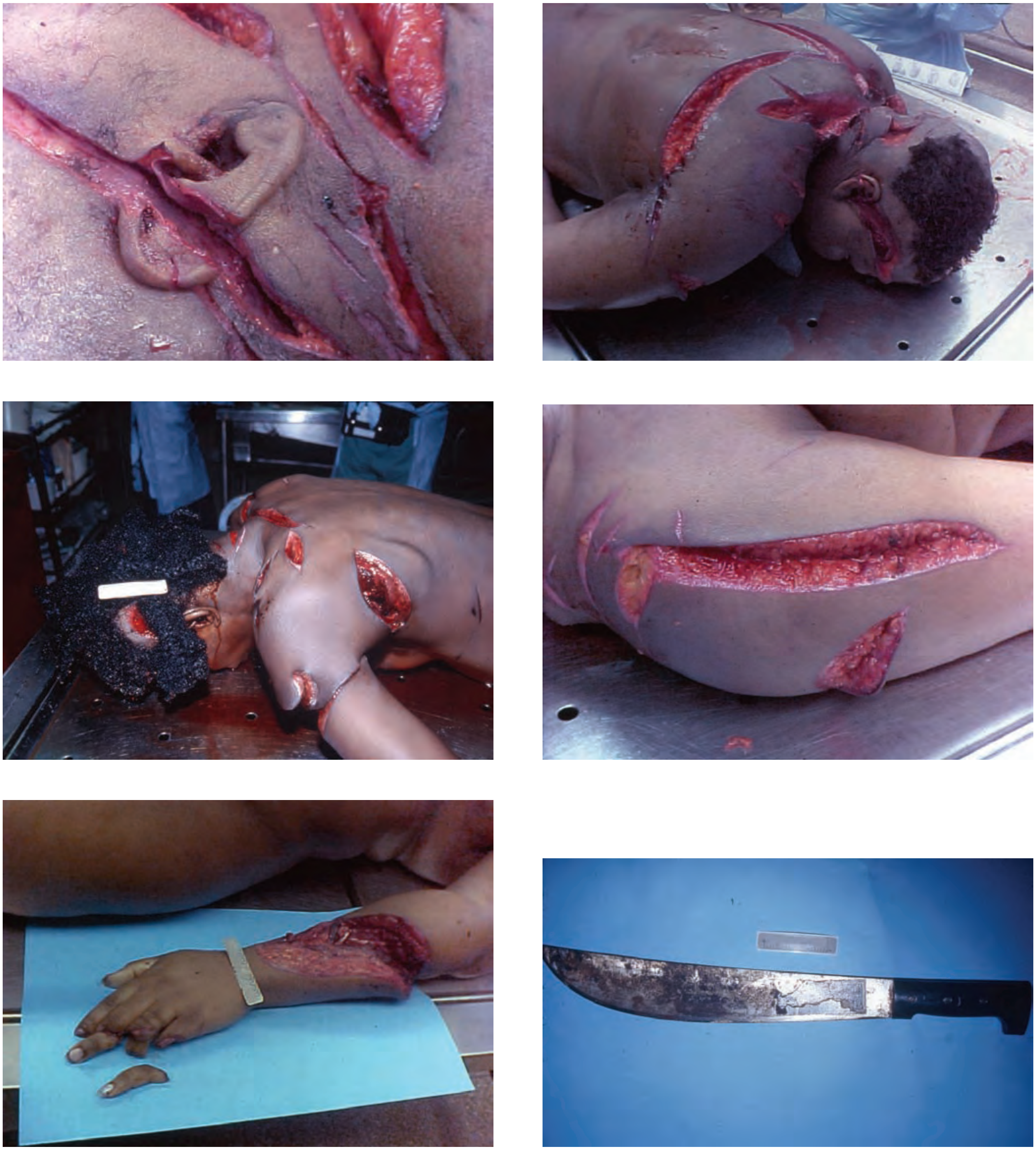

These individuals were attacked multiple times with machetes. Note the long gaping hack marks, several of which cut into and through bone. Most often, this degree of violent rage indicates the victim and the perpetrator had a relationship.

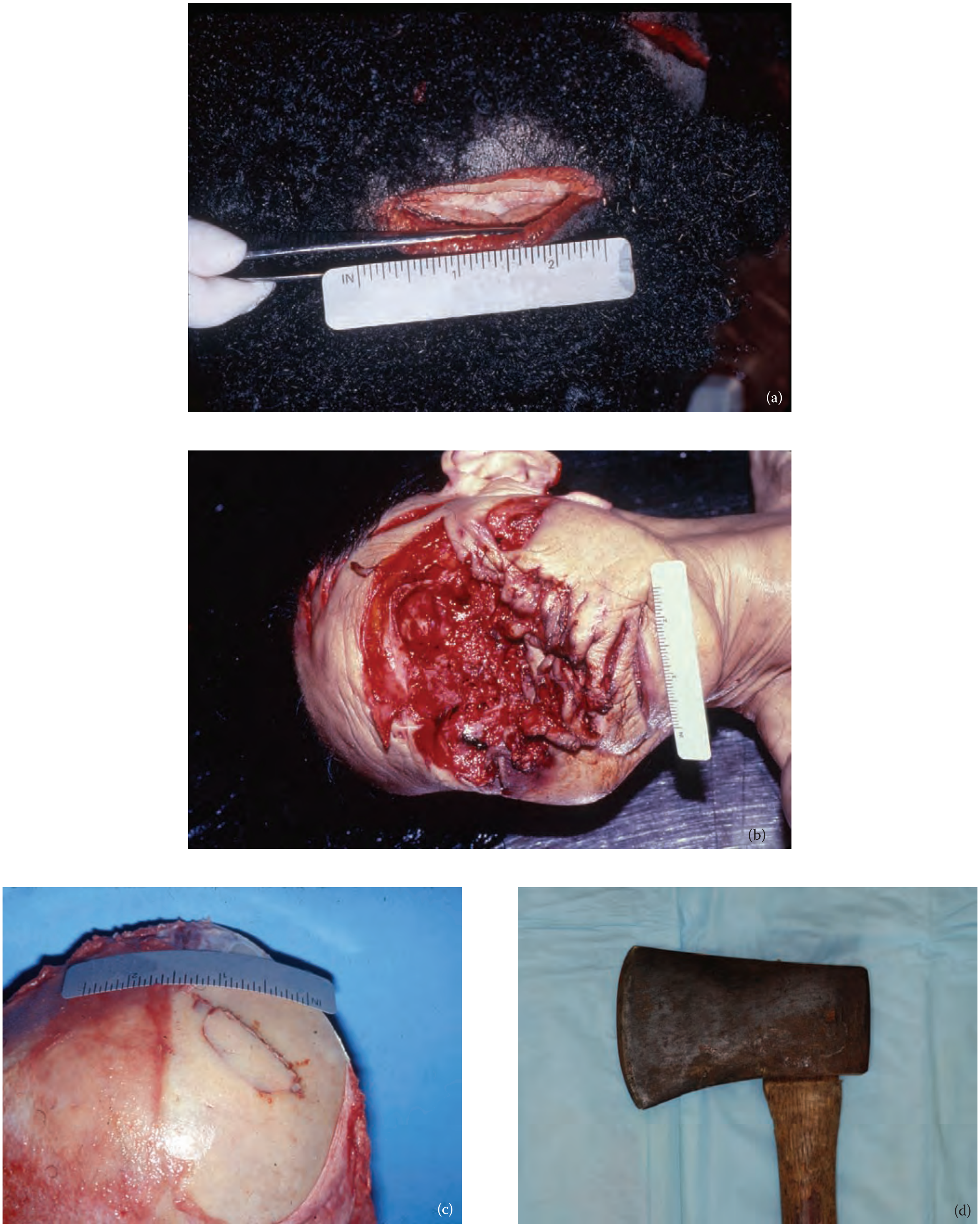

Chopping-type sharp-force injuries produced by an axe. Image (b) shows an individual whose eyes were chopped out by another person with a long history of psychosis. These types of assaults are not common. He reportedly did this because she kept staring at him.

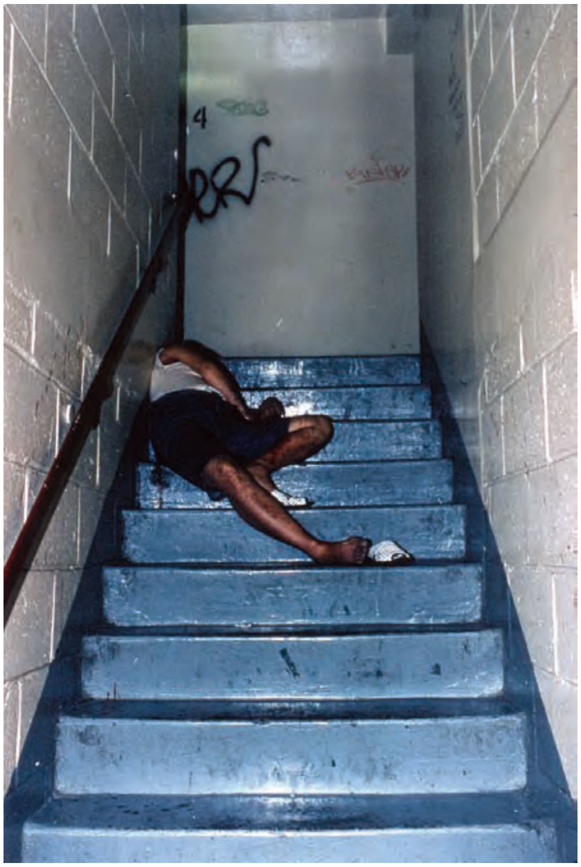

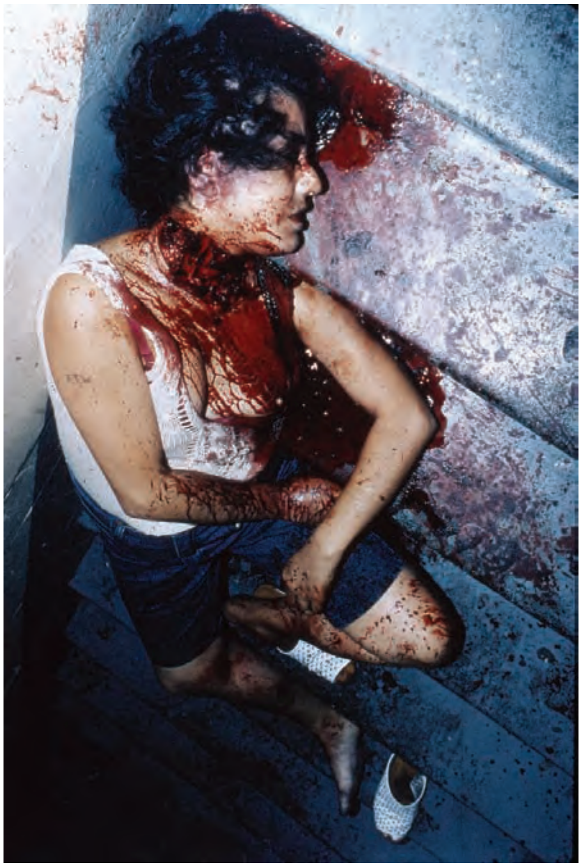

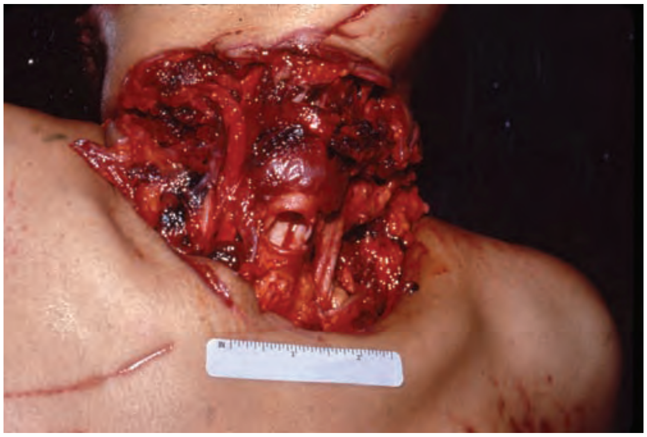

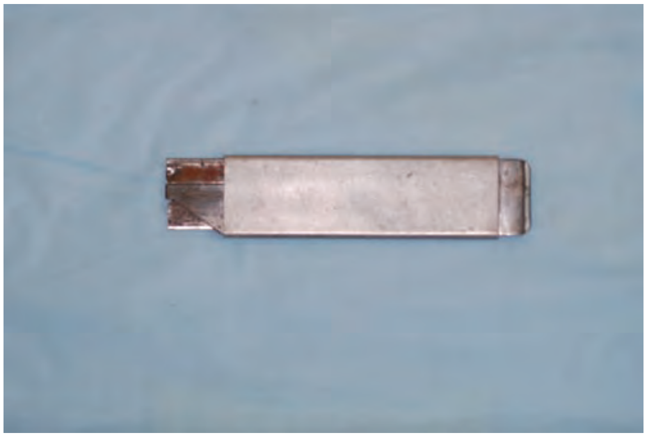

This individual was chased up several flights of stairs in an apartment complex and was cut multiple times with a box cutter similar to the one shown here. Note the blood splatter pattern on the steps. Also note that the individual was almost decapitated. This was reportedly over a dispute concerning gang-related drug dealing territory.

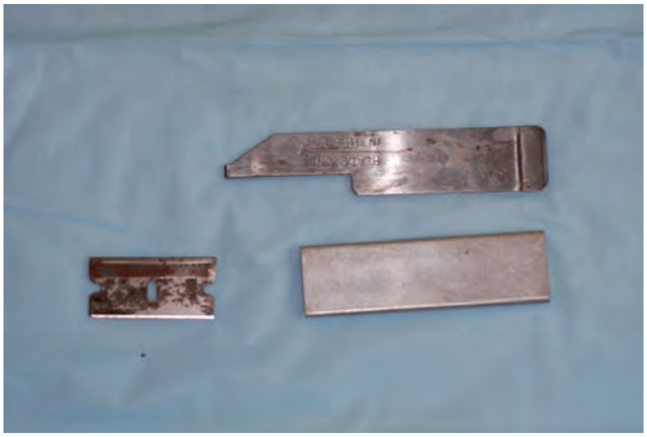

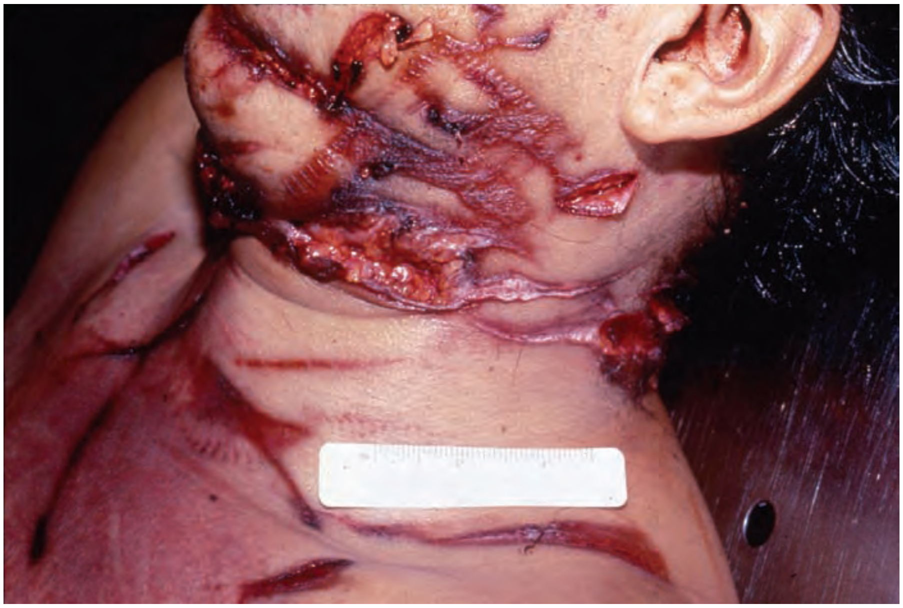

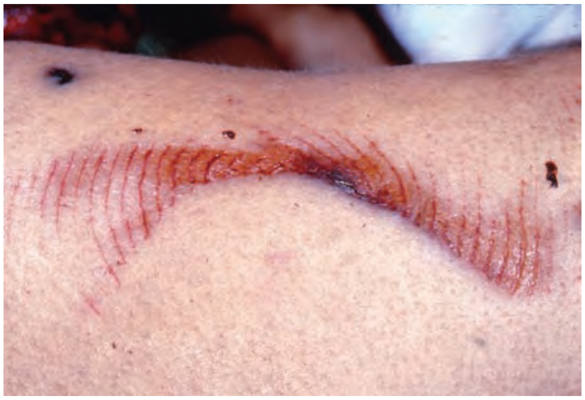

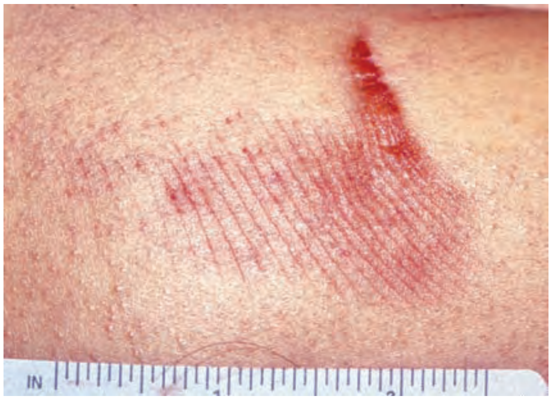

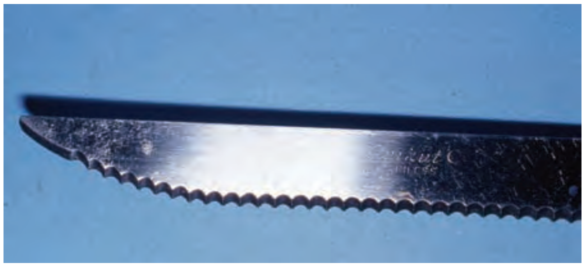

This individual had multiple stab wounds with adjacent abrasions consisting of multiple superficial linear parallel lines. This is a typical pattern found following a serrated knife attack. The absence of these marks does not preclude the use of a serrated knife. If a serrated knife is stuck directly into a body and not dragged across the surface, there will be no parallel linear abrasions.

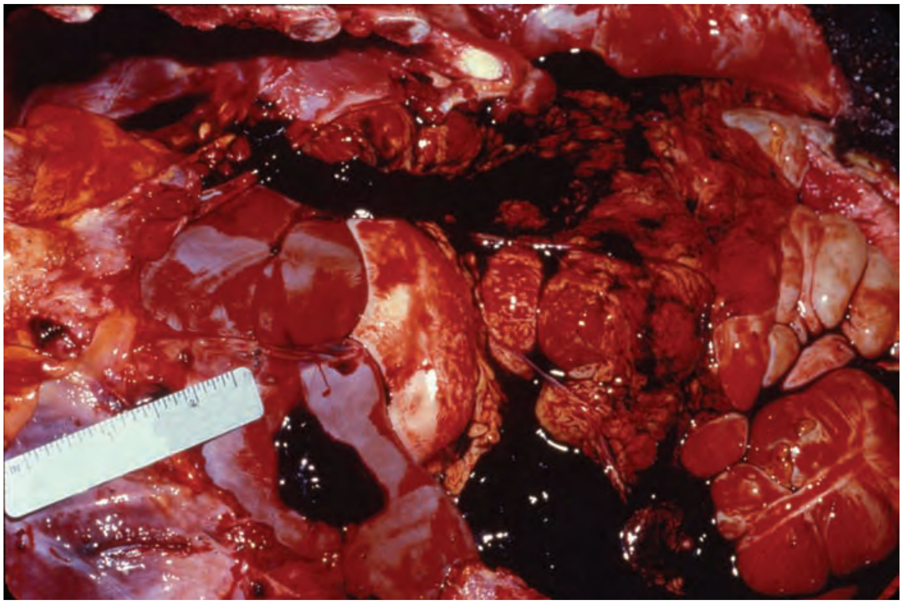

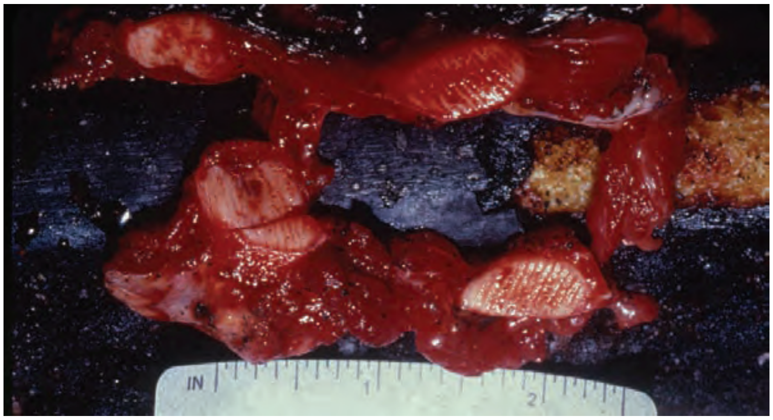

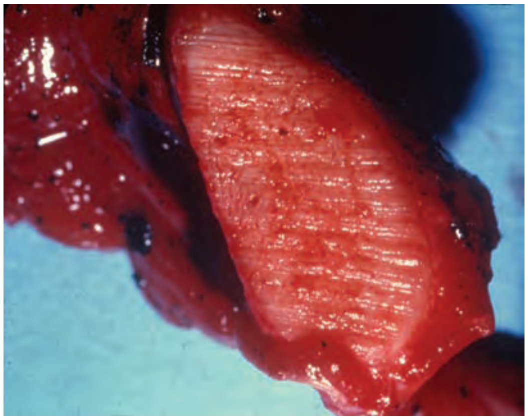

This individual was stabbed multiple times with a Ginsu-type knife. His dwelling was then set on fire and he sustained extensive thermal burns with charring of the skin. Note the costal cartilages with tool markings forming an imprint of the knife blade. Image on lower right demonstrates a knife similar to the one used in the attack.

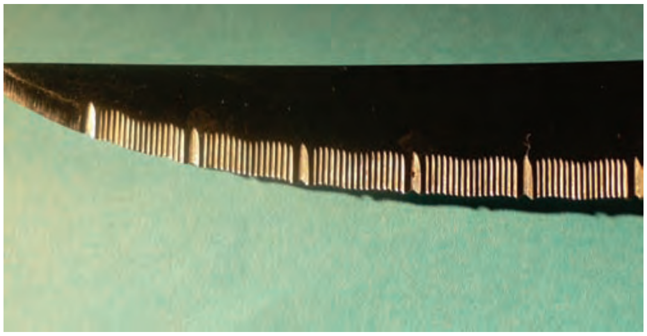

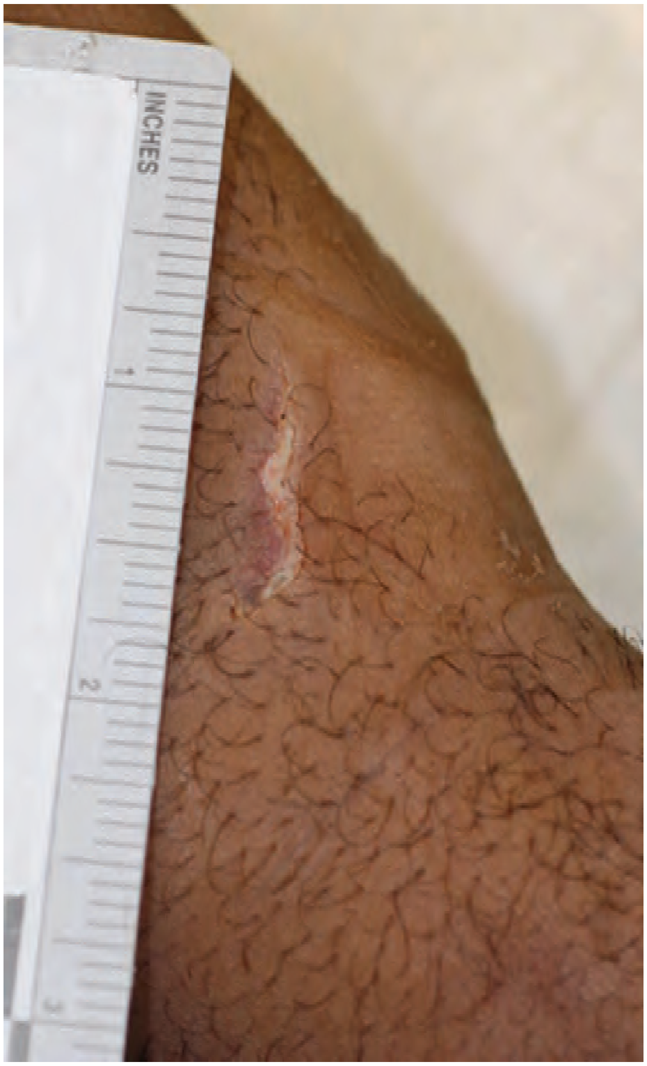

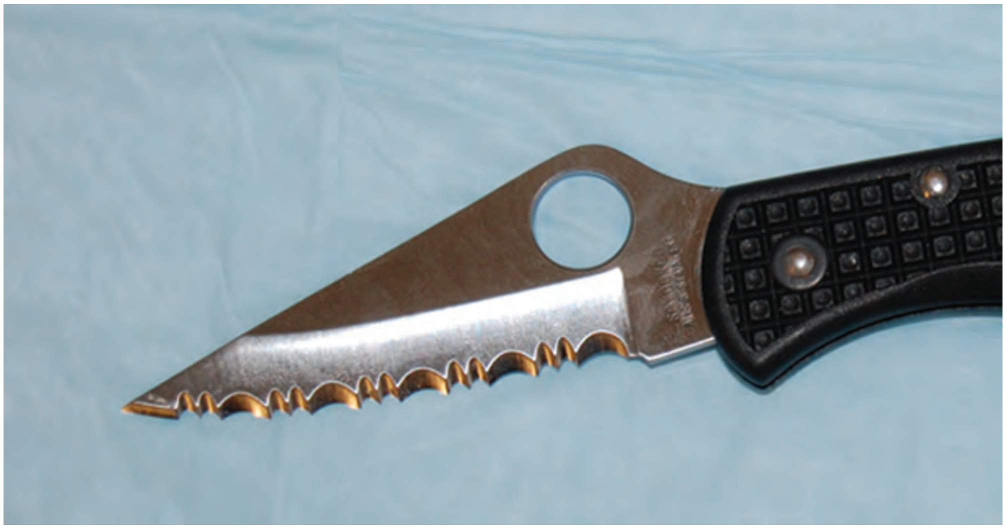

Note this irregular incised wound with curved sharp margins produced by a knife similar to the one depicted in the previous figure.

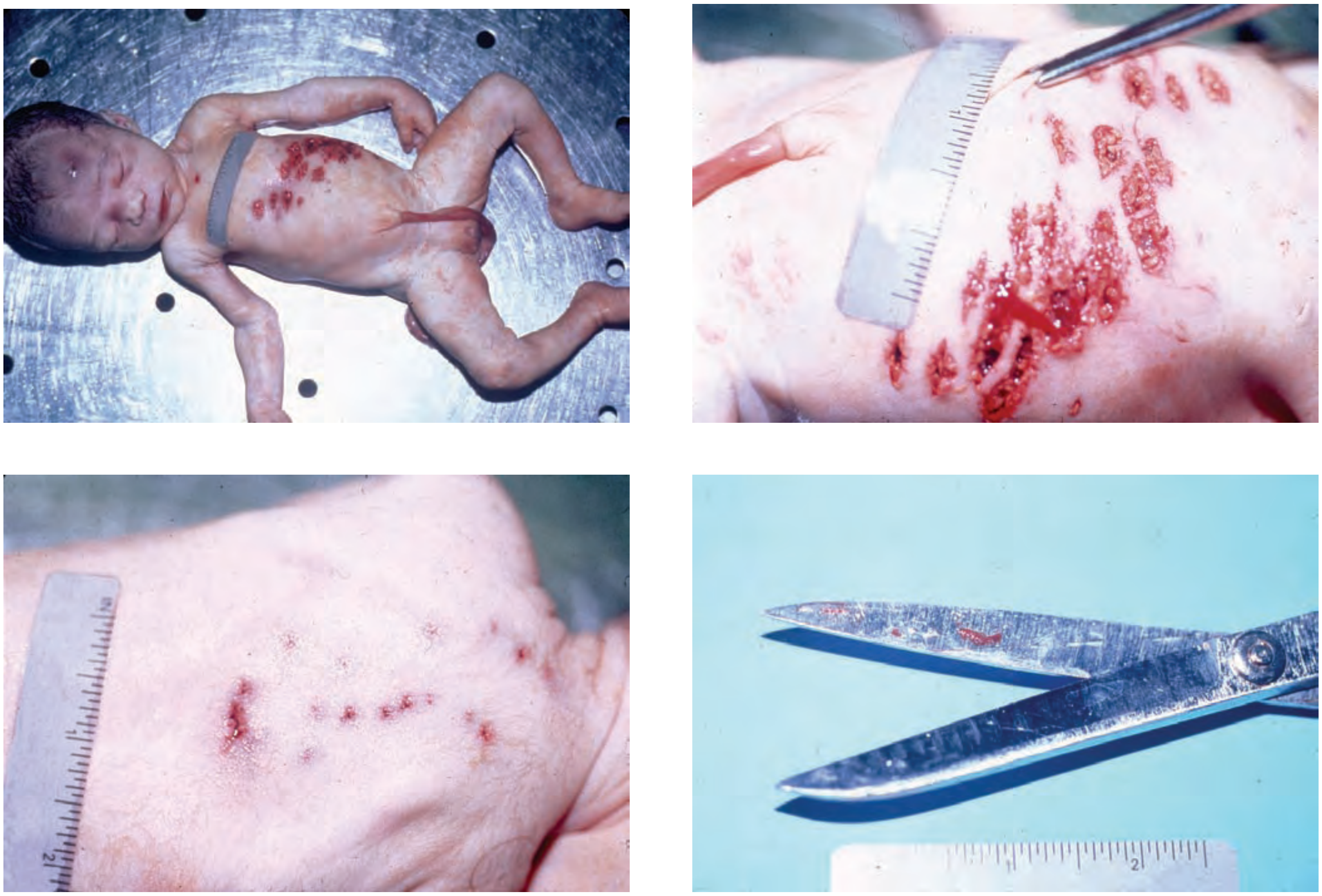

This child was stabbed multiple times with a pair of scissors. Note that the entrance defects to the anterior trunk have a slightly widened angulated character consistent with being inflicted by a pair of scissors. The posterior trunk demonstrates the pointed end of the scissor perforating the entire body.

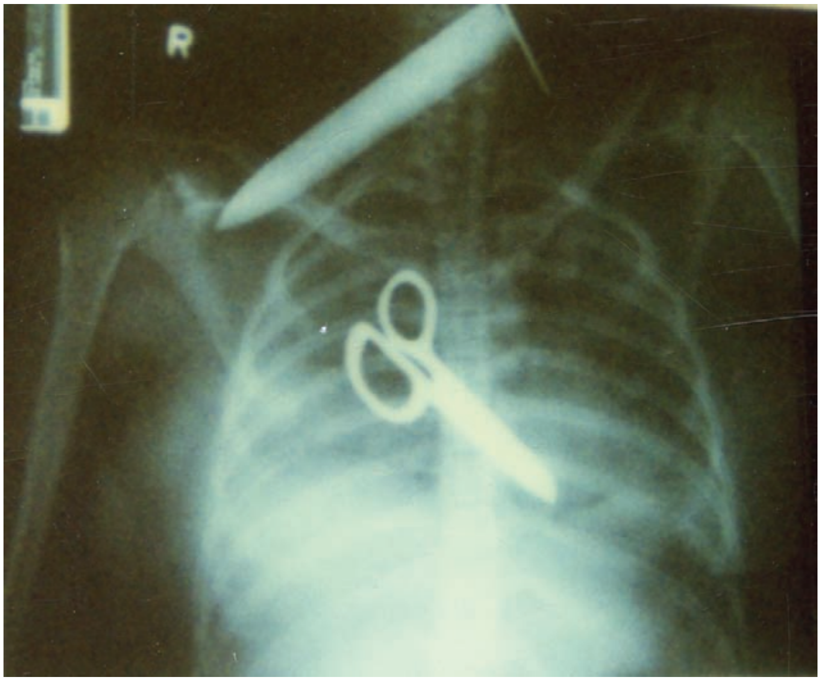

This image shows an x-ray of an adult who was stabbed with a pair of scissors that were left in the wound.

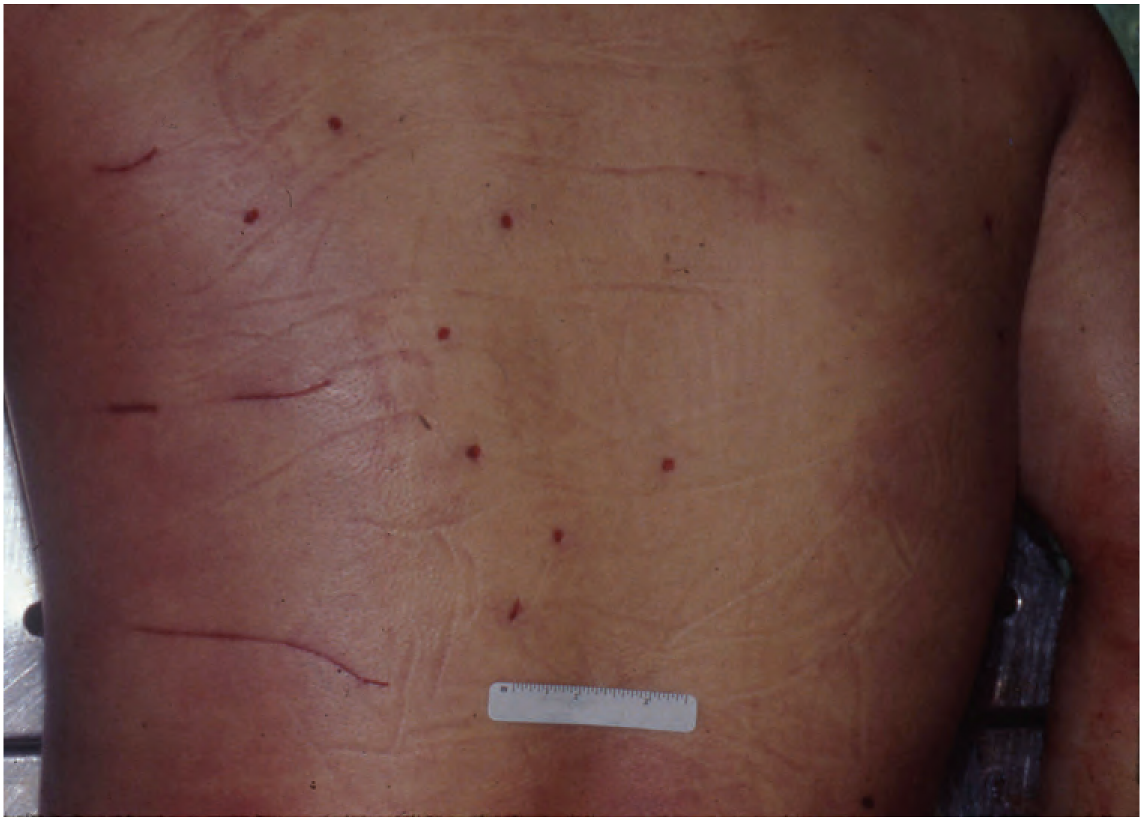

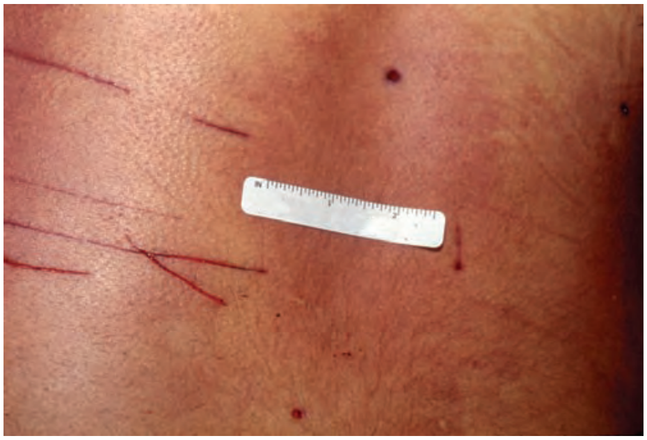

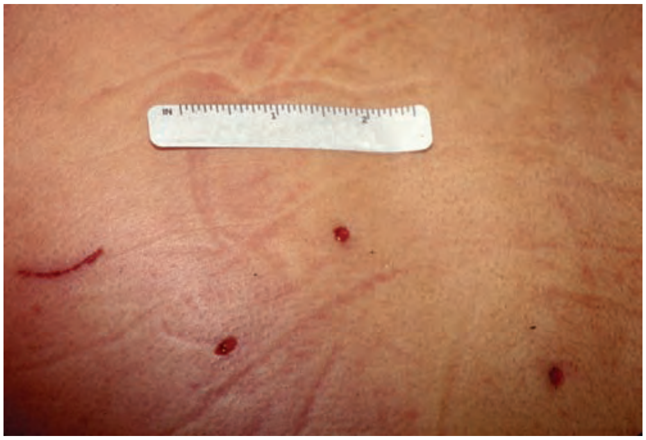

This individual was stabbed more than 75 times with an ice pick. Note the linear abrasions extending in different angles across the back as the tip of the ice pick was dragged across the skin's surface. The small dimension of these perforation wounds would take more time to lose significant blood compared with a typical knife blade. Generally speaking, the larger the wound, the faster the blood flow in a extravascular region. There was evidence of a violent struggle over a long period of time in this case.

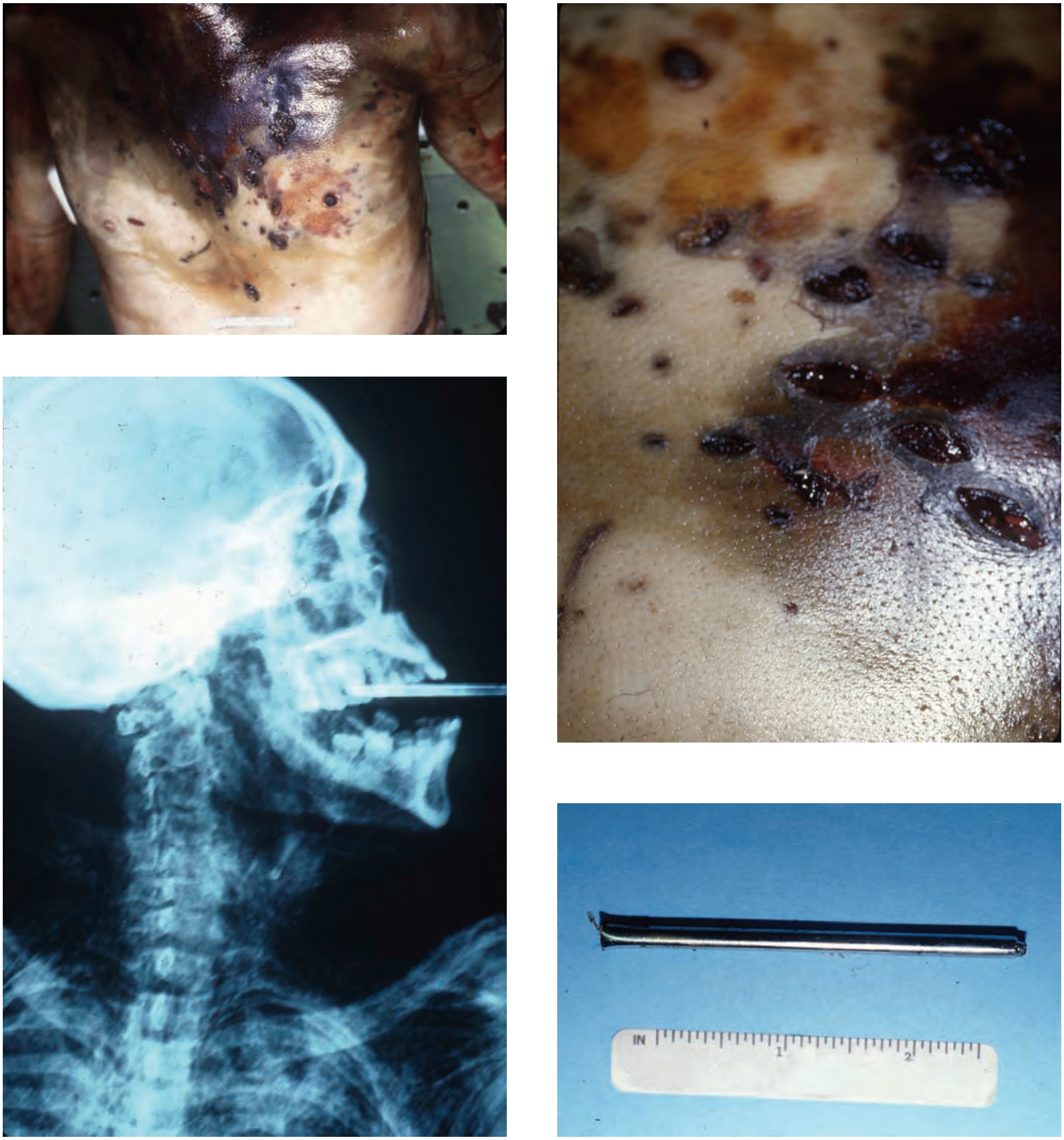

This individual was found within a crack house, in a mild to moderate state of decomposition. He sustained multiple stab wounds and puncture-type wounds to his body that were produced by broken pieces of antenna used as crack pipes. One of these antennas was jammed down his throat. He also sustained multiple blunt-impact injuries with rib fractures and other broken bones from being stomped. Note the drying and slight distortion of the wound margins due to decomposition.

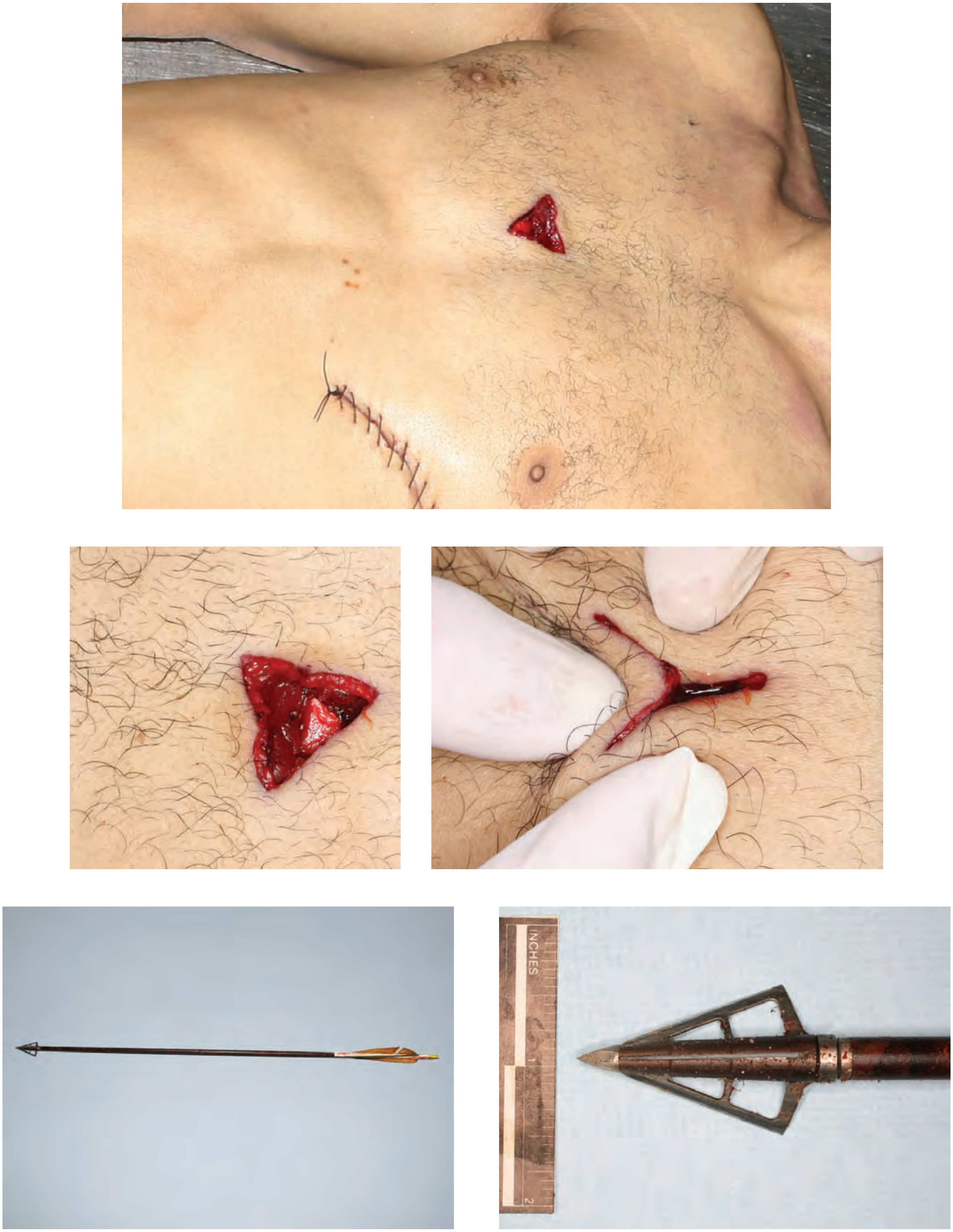

This individual was shot with a hunting arrow. The entrance wound is to the medial aspect of the right chest and the exit wound was to the right back. Note the comparison of the injury to the arrow head.

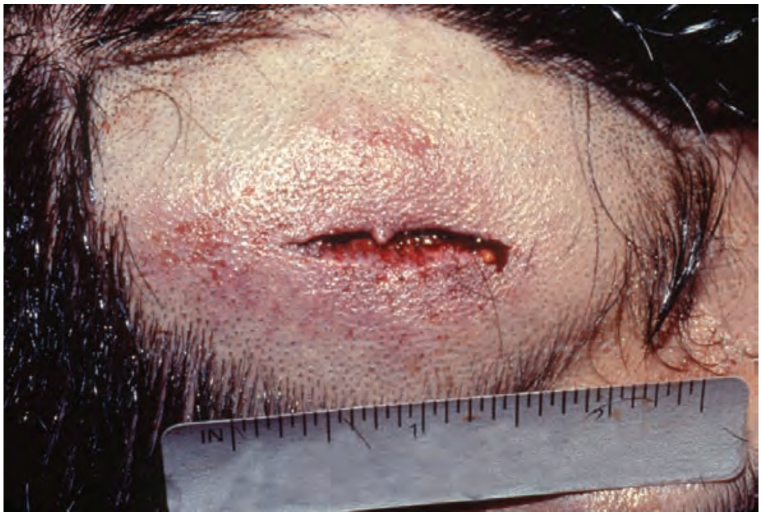

This individual was struck with a bottle that produced a minimal abrasion with a greater underlying contusion and a central laceration. Note the slightly irregular margins at the laceration site due to this blunt impact. The surrounding scalp hair was shaved to demonstrate the injury in greater detail. The presence of the scalp hair at the time this injury was inflicted also served to cushion the impact site.

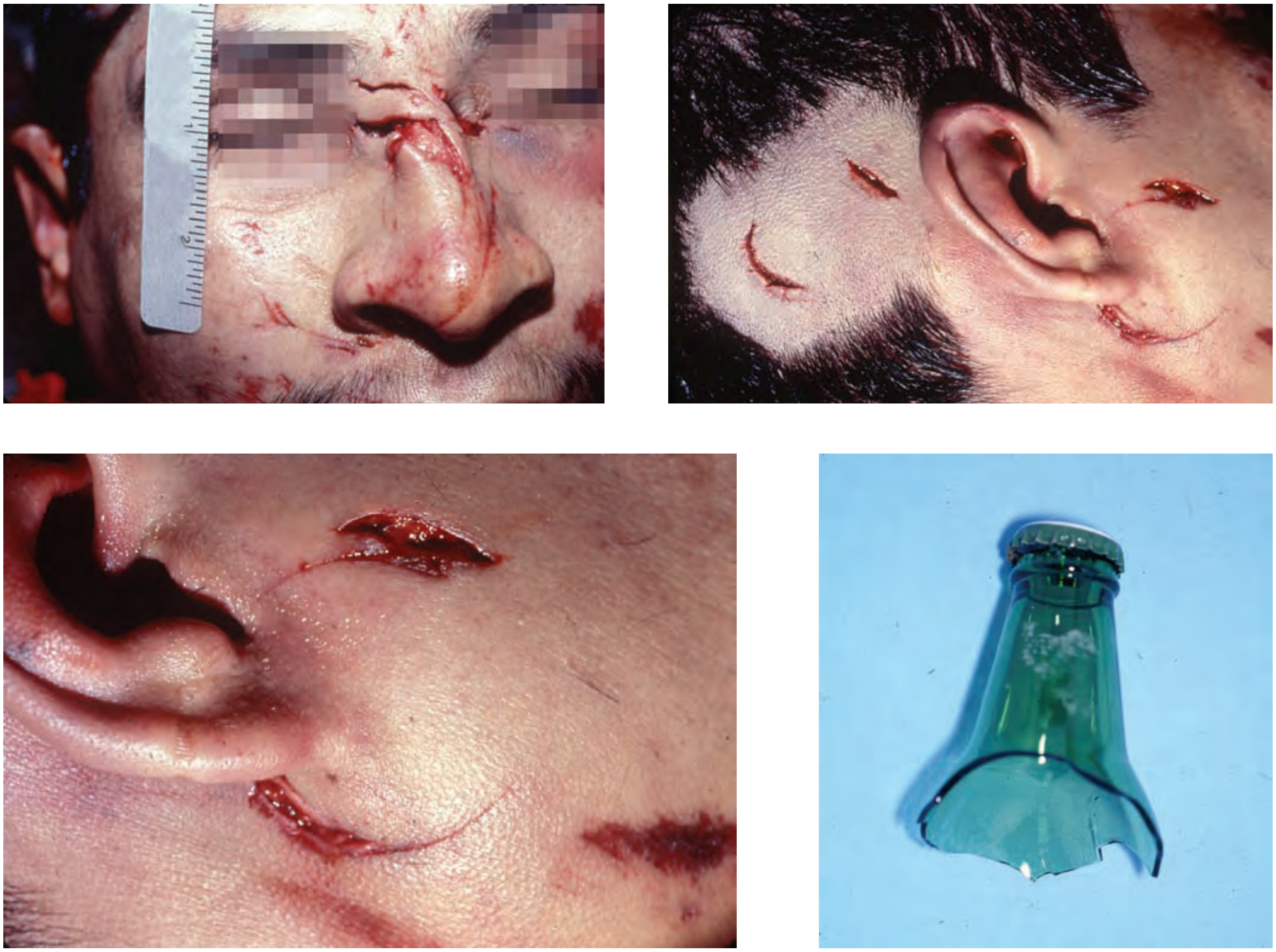

This is the same individual, who was then cut multiple times with the sharp broken glass end of the bottle, producing a partial transection to the carotid artery and exsanguination.

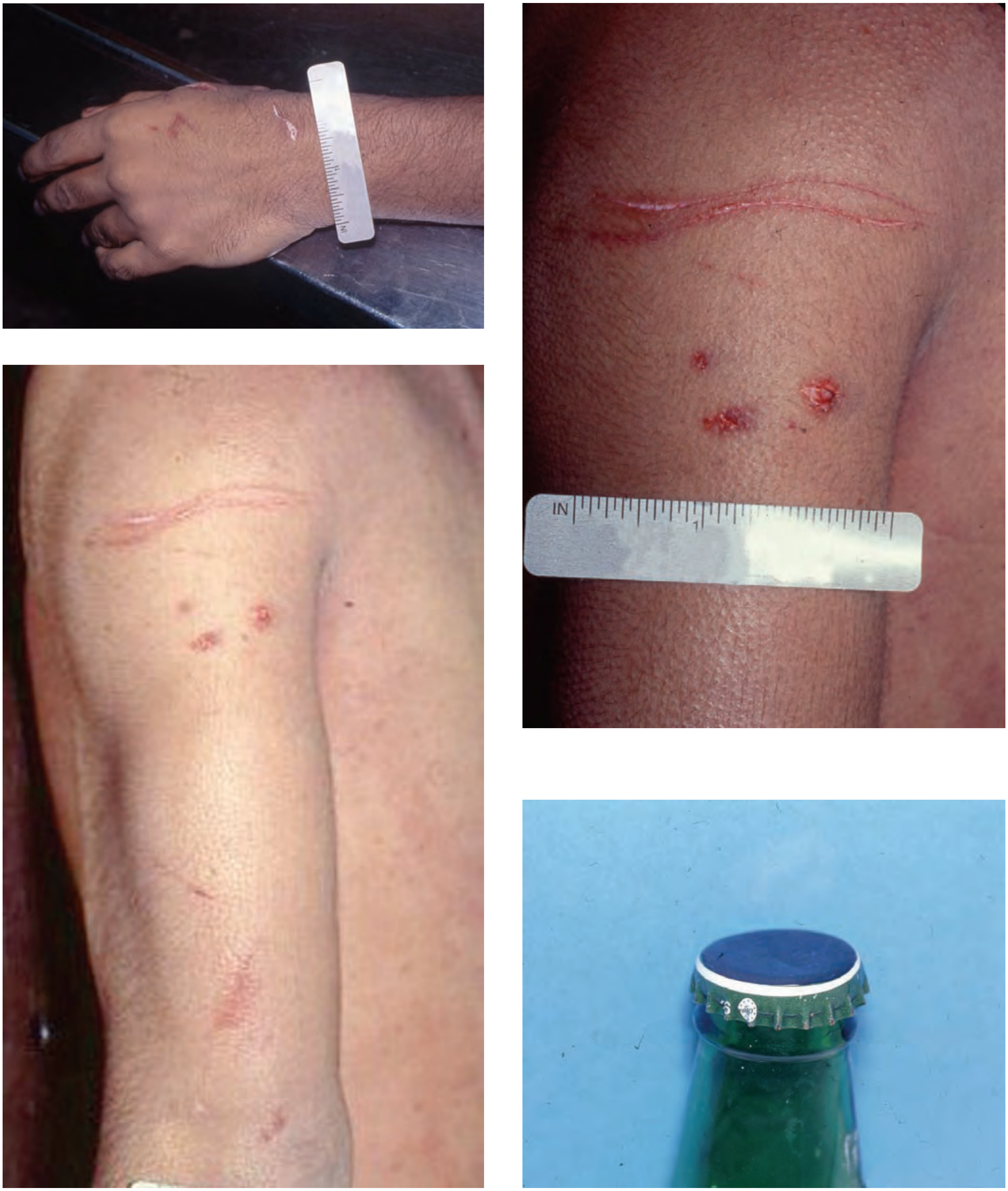

The individual had multiple curvilinear parallel abrasions and other superficial abrasions caused by a bottle top and broken glass scraping across the body surface during the assault. This initial argument began over which rival baseball team was better.

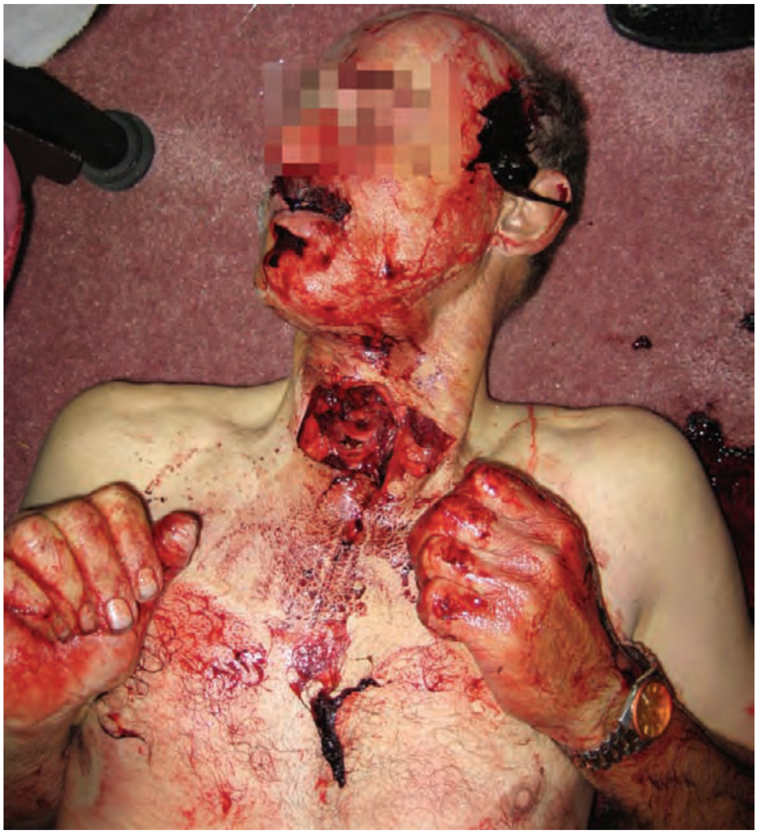

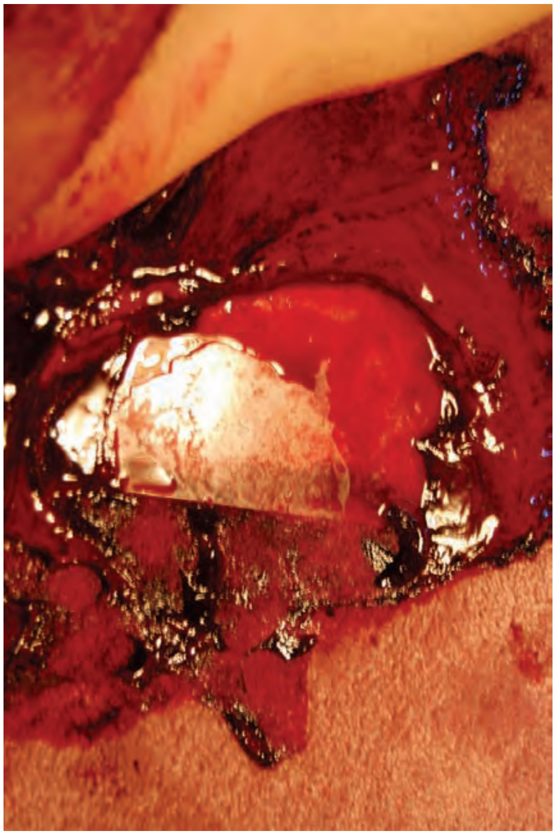

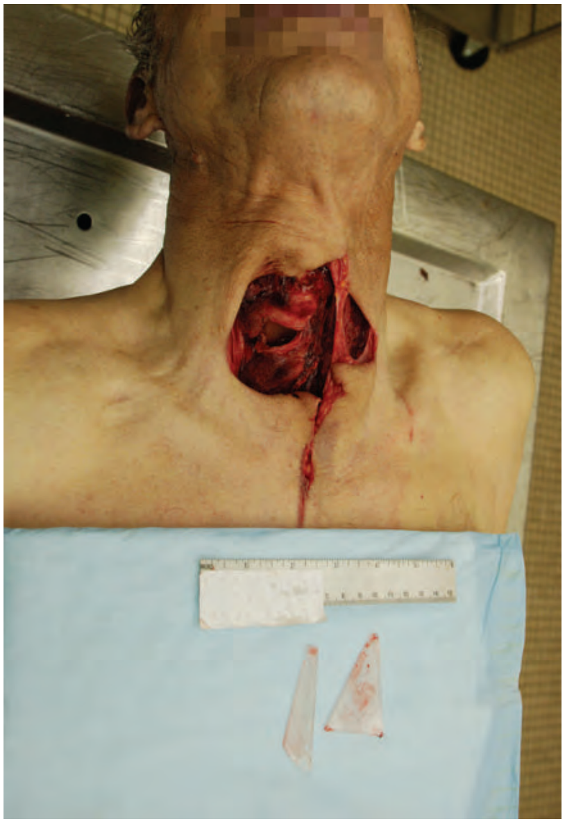

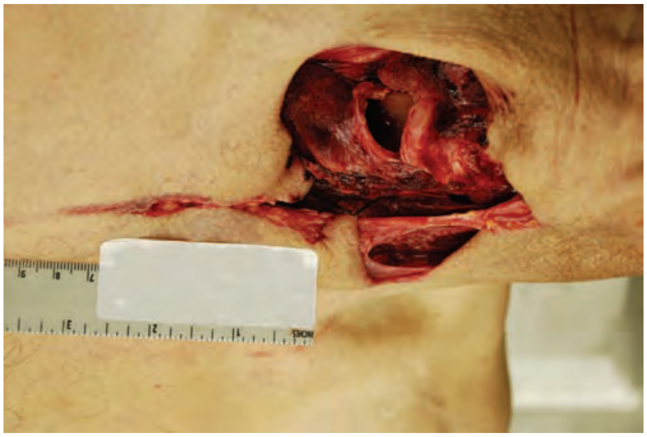

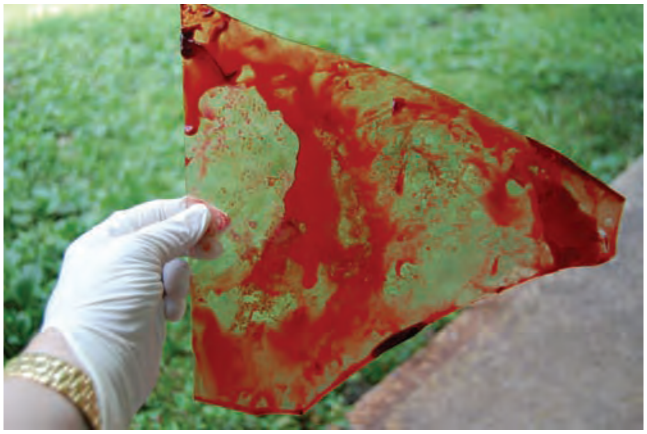

Suicide by multiple sharp-force injuries with broken glass taken from an empty picture frame.

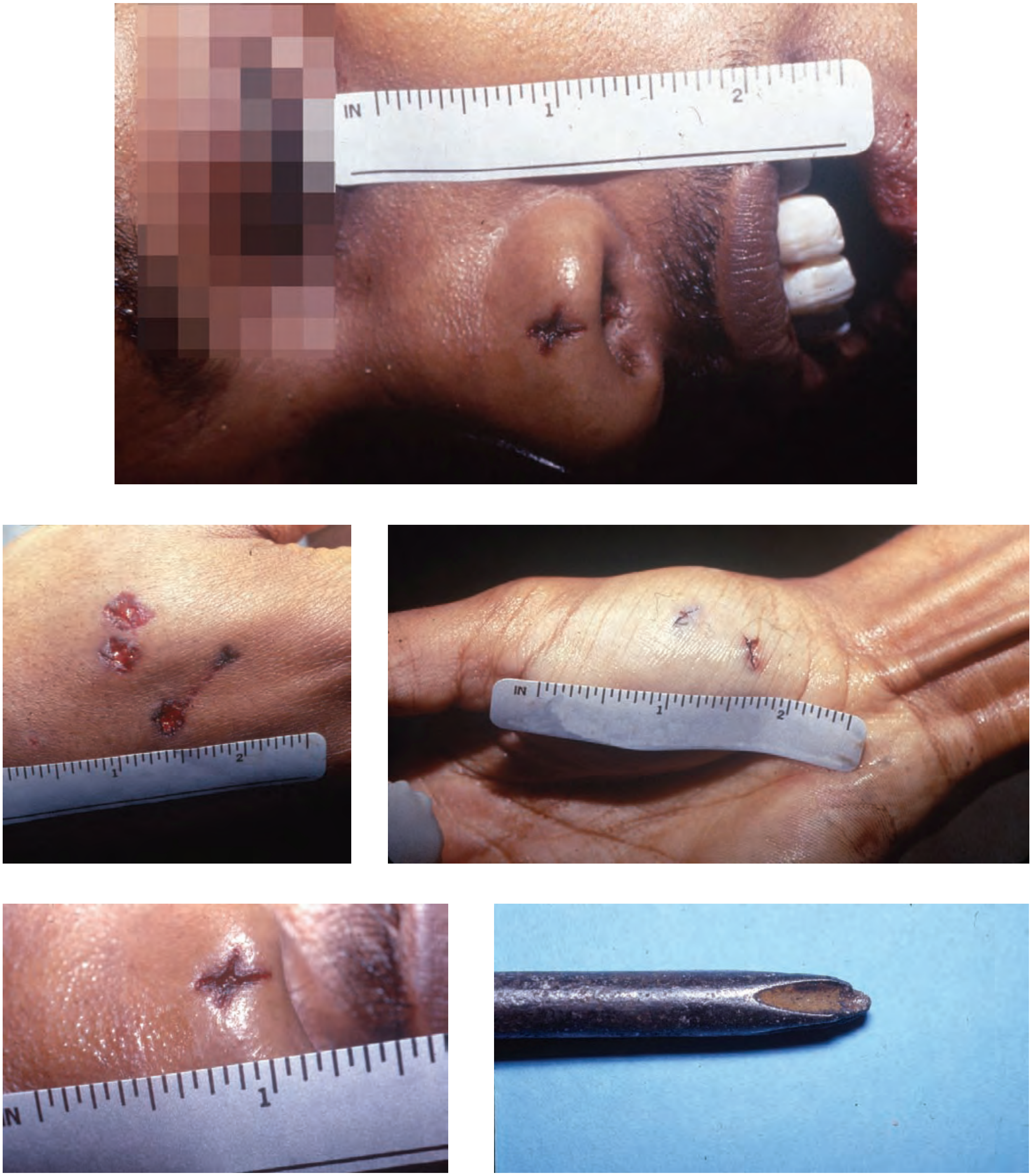

These are multiple examples of puncture wounds caused by a Phillips-head screwdriver, demonstrating the range of presentation of this type of injury.

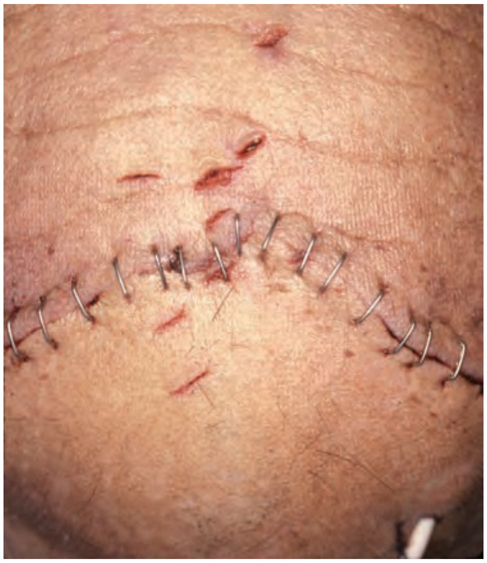

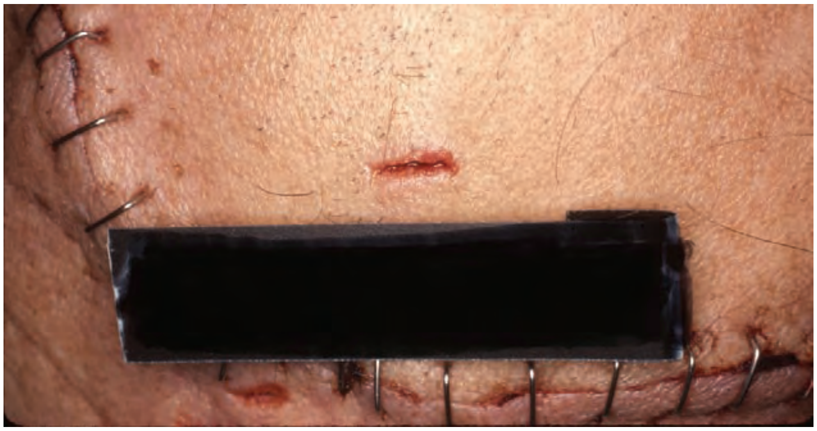

[redacted]

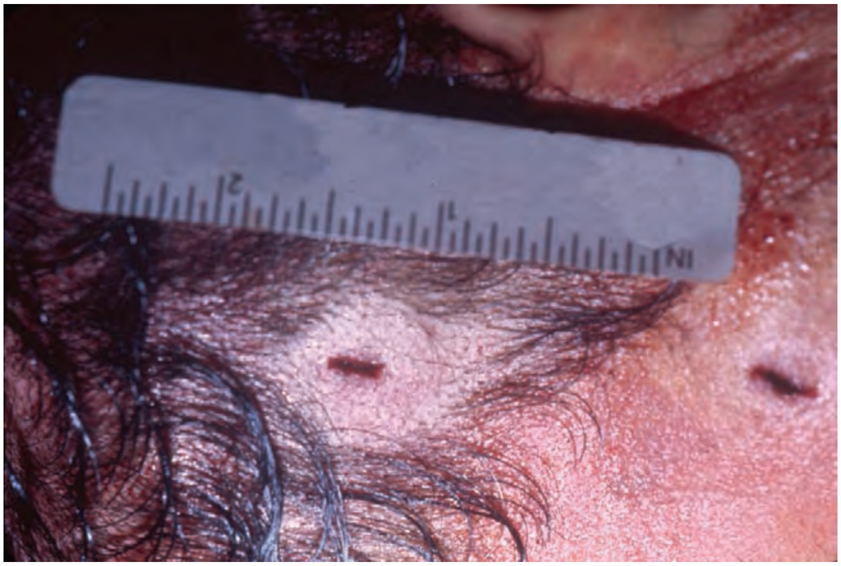

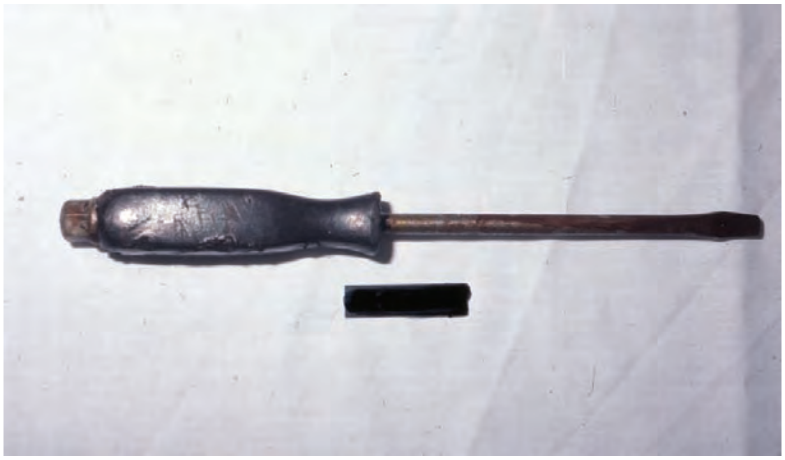

[redacted]

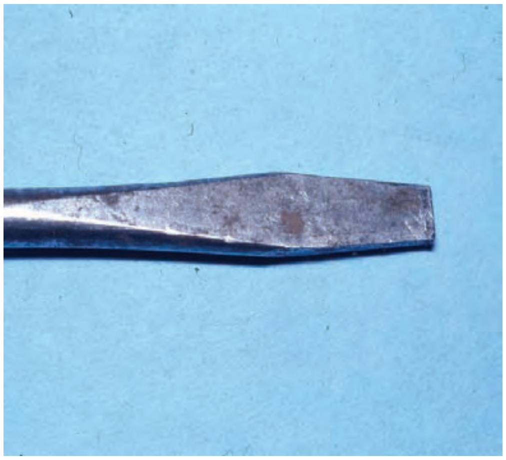

These are multiple examples of flat-head screwdriver puncture wounds.

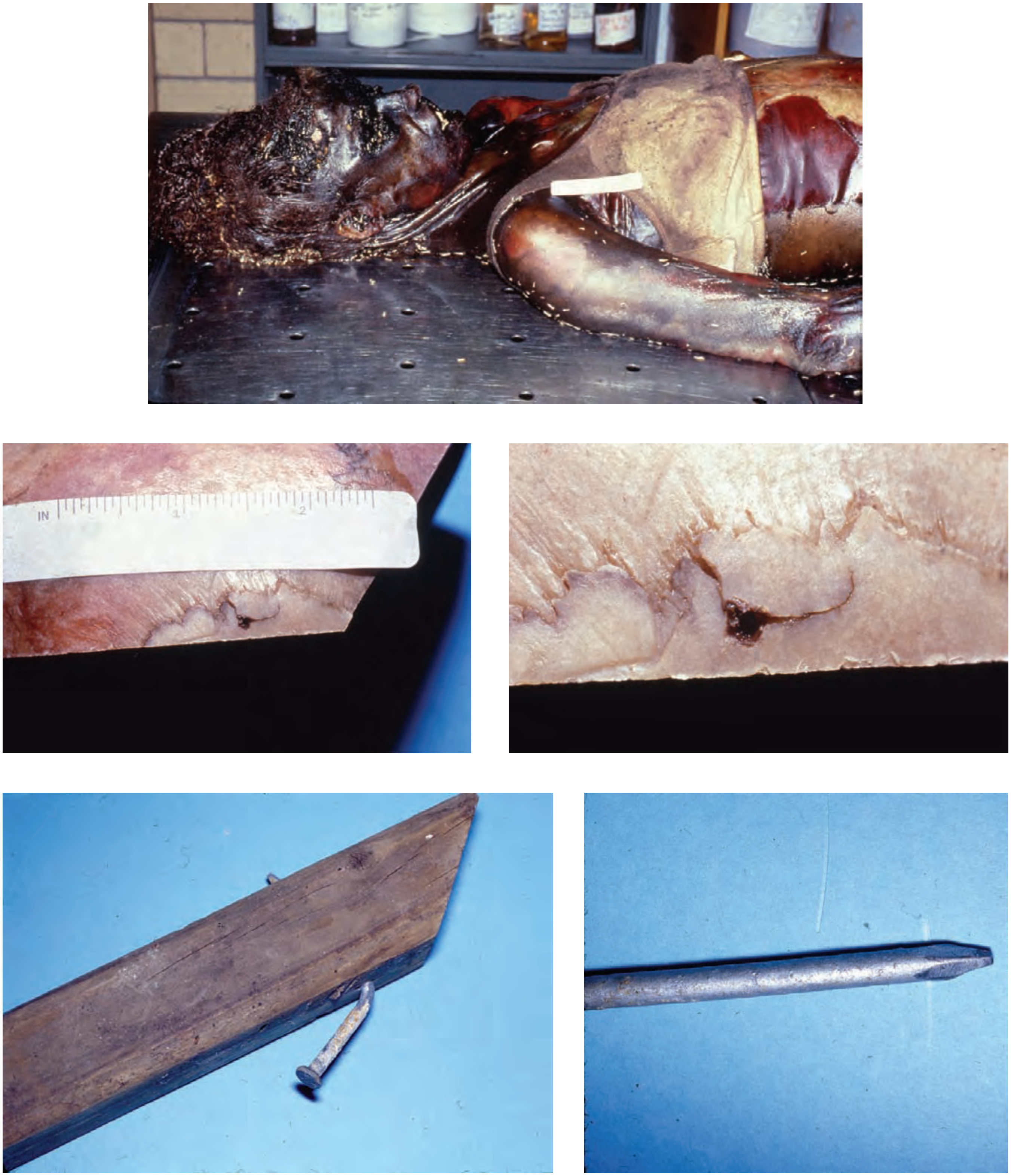

This approximately 30-year-old woman was found naked from the waist down in a construction yard. She had moderate decompositional changes with insect and animal feeding. She had multiple defects to her scalp and face with numerous maggots tracking through the underlying soft tissue. The soft tissues in this region were darkly discolored throughout. It becomes more difficult to interpret soft-tissue injuries as decomposition progresses. Careful examination of the underlying bone in such cases may often yield valuable information. Note the approximately 1/4″ linear skull fracture with a 1/16″ roughly square indentation at the superior right temporal bone. Further examination of the scene revealed a bloodstained board with a nail at the opposite side of the construction site.

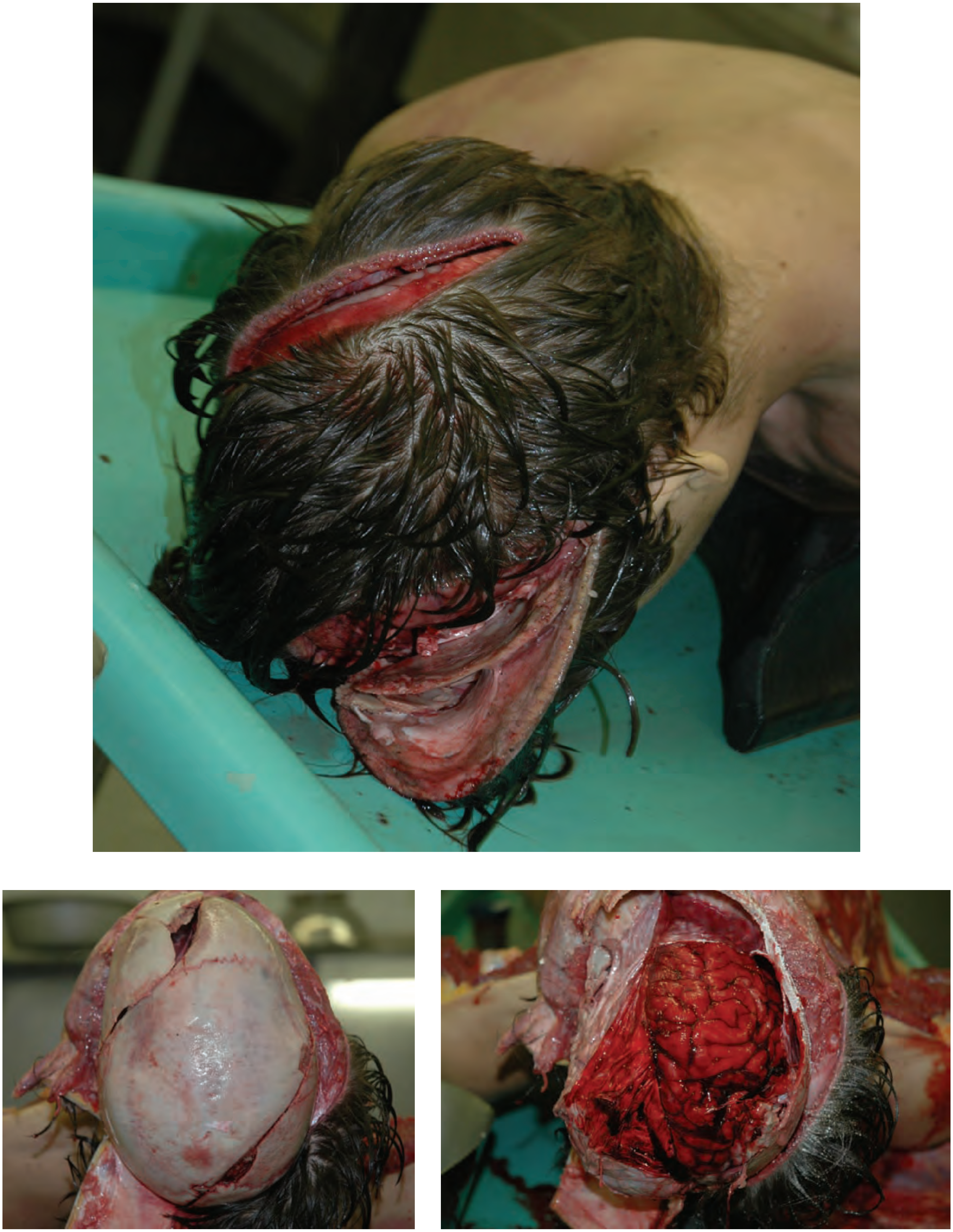

Antemortem boat propeller injuries. Note the presence of wound track hemorrhage and the parallel slicing-type injuries.

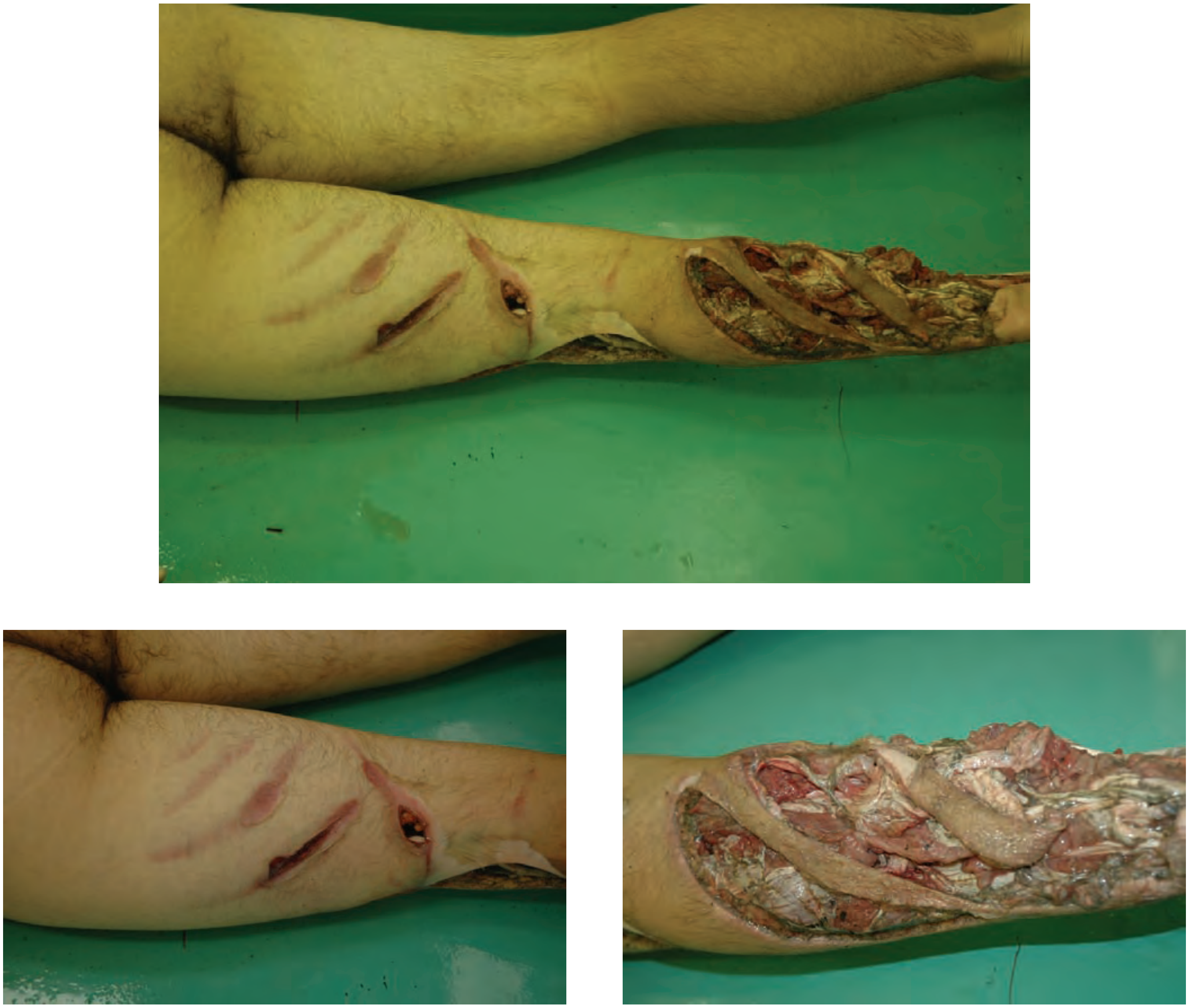

Injuries produced by a boat propeller.

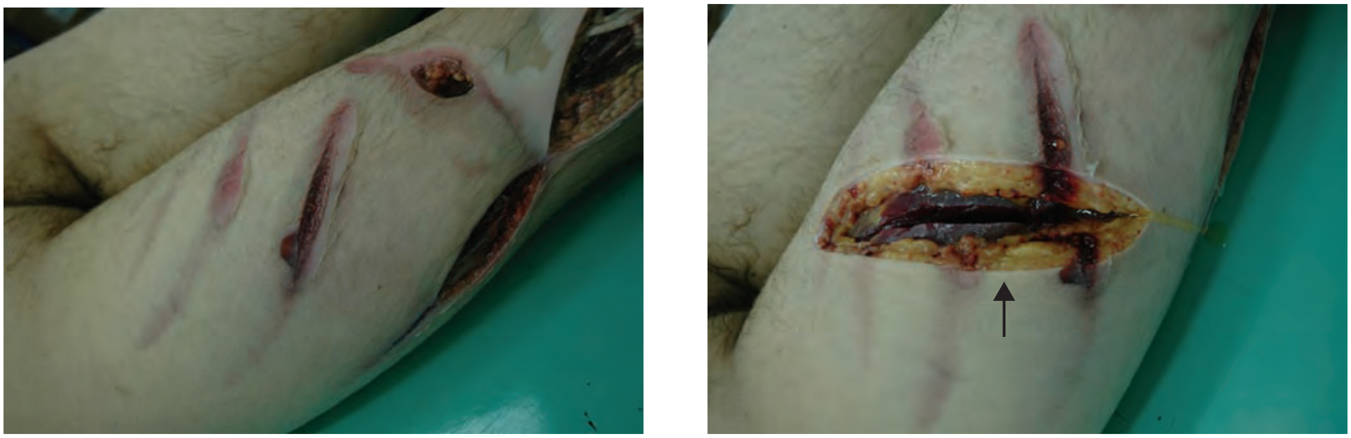

Injuries produced by a boat propeller. Note the hemorrhage at the autopsy incision (lower right), demonstrating the individual was alive when these injuries occurred. Note the range of injury presentation, including superficial abrasion to deep slicing-type wounds with extensive soft-tissue damage.

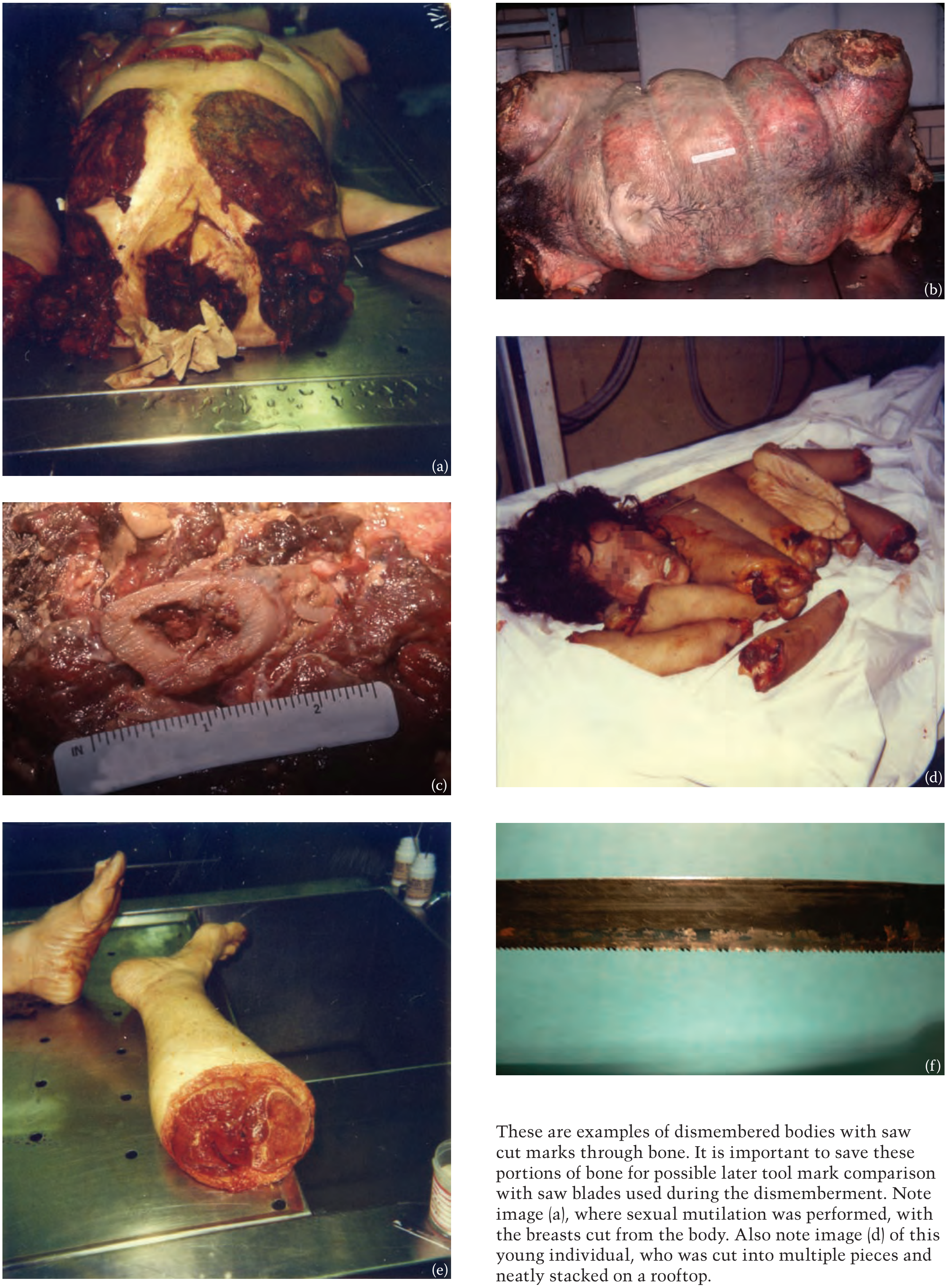

These are examples of dismembered bodies with saw cut marks through bone. It is important to save these portions of bone for possible later tool mark comparison with saw blades used during the dismemberment. Note image (a), where sexual mutilation was performed, with the breasts cut from the body. Also note image (d) of this young individual, who was cut into multiple pieces and neatly stacked on a rooftop.

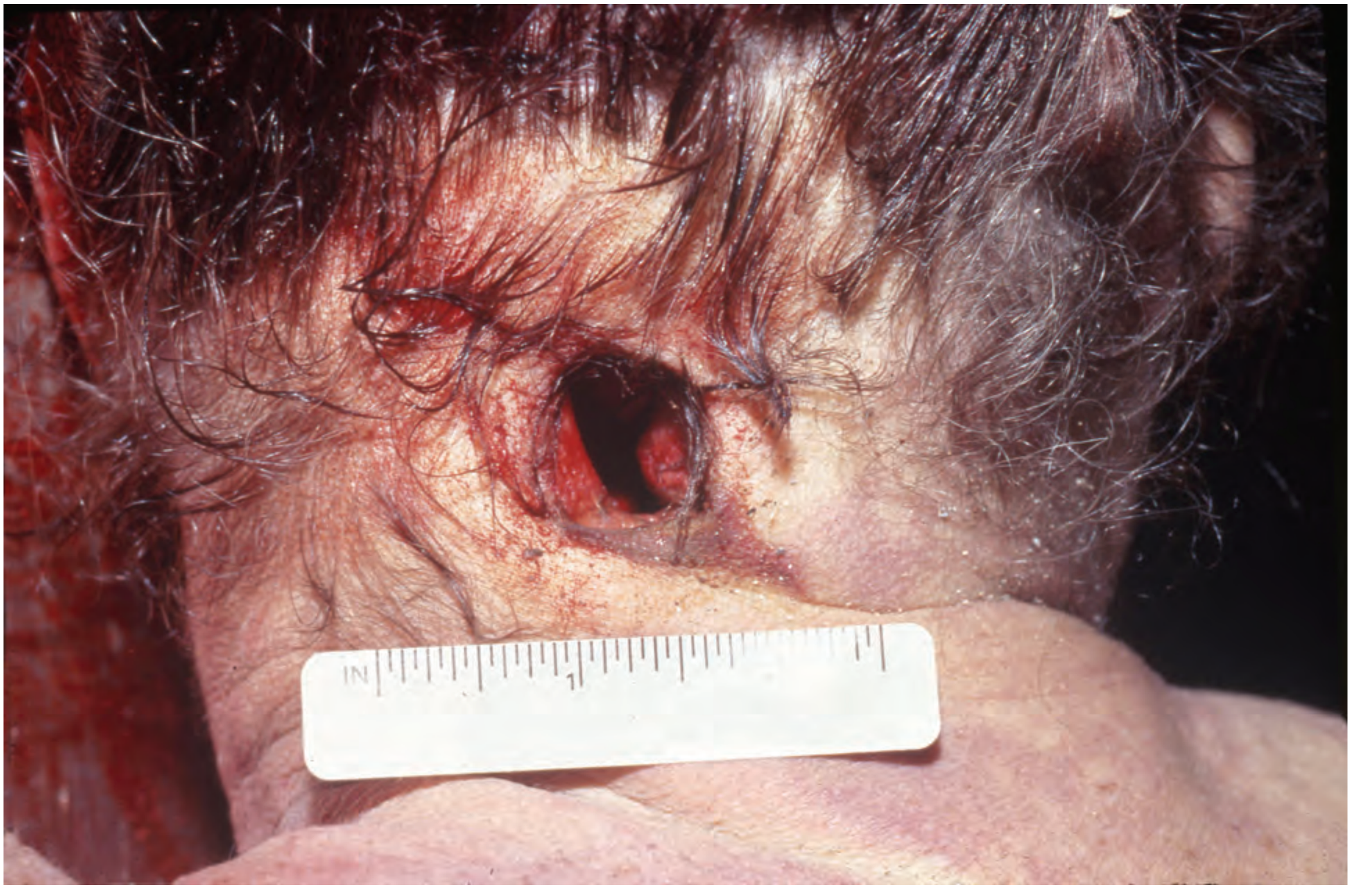

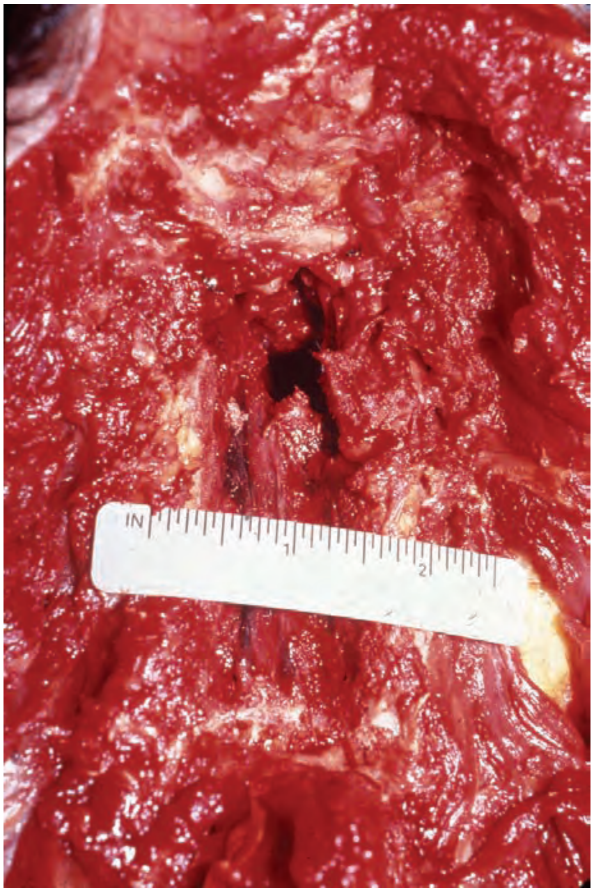

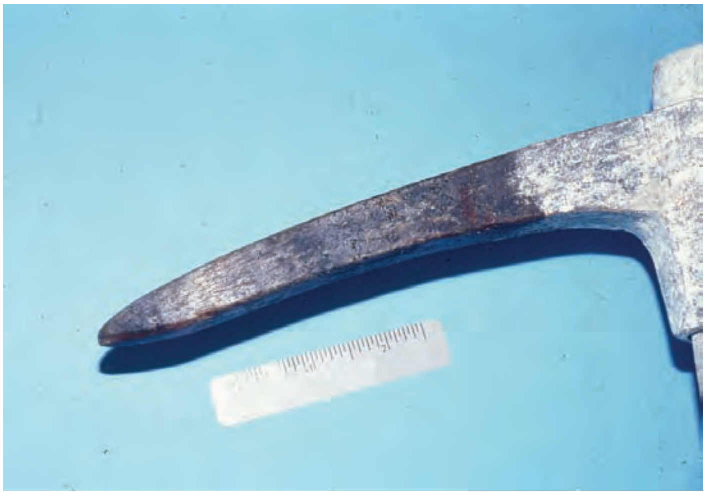

This individual was killed by being struck in the back of the neck with a pickaxe. Note the abraded margin and the perforation matching the roughly squared dimension of the pickaxe. This dispute was reportedly over not receiving back pay after complete work.

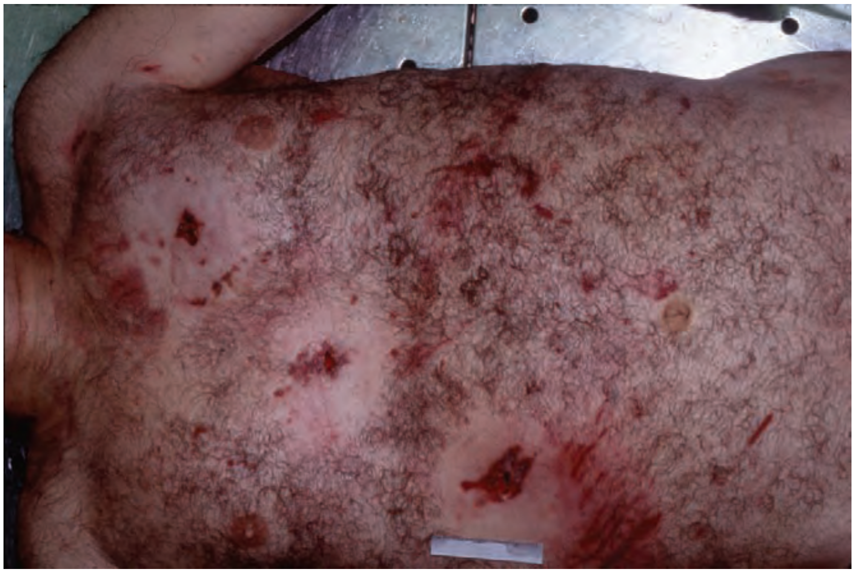

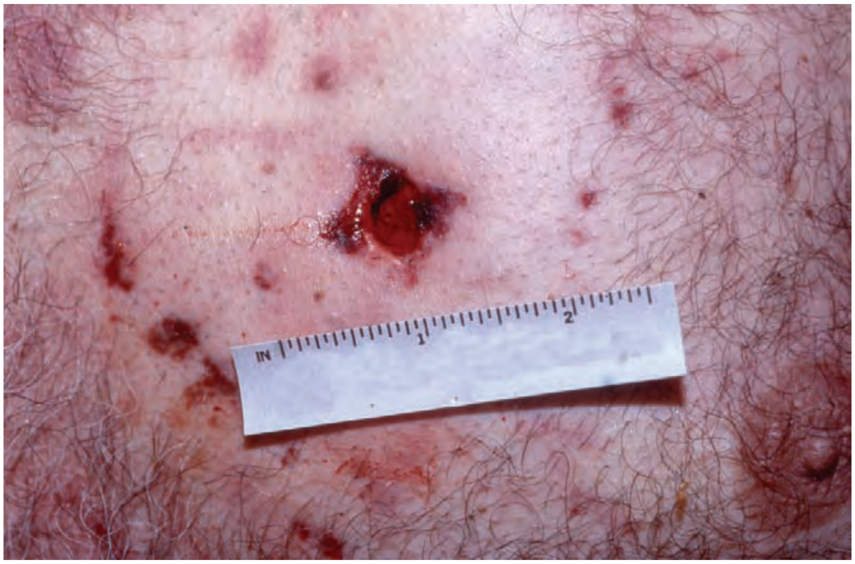

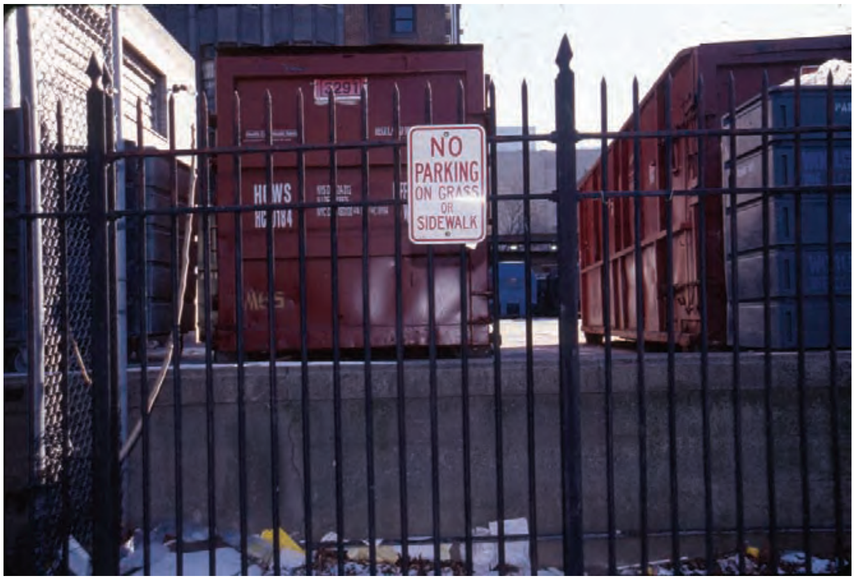

This individual committed suicide by jumping out of a window and landing on top of a spiked metal fence similar to the one shown in image (c). The hair surrounding these injuries was shaved to demonstrate the nature of the wound in greater detail. Note the pointed circular perforation site leading into a square abrasion, which was perfectly consistent with the fence spike that penetrated his chest.

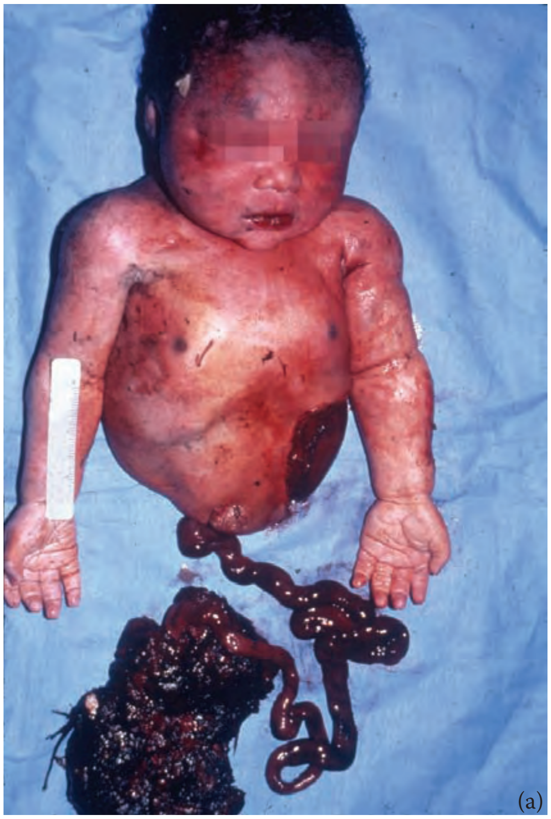

This infant was found partially eaten by dogs in a building courtyard. Note the irregular nature of the wounds to the back (b–d), demonstrating a scalloped border produced by teeth and claws. Other injuries to the child's body were inconsistent with animal feeding and more likely were produced by a sharp instrument such as a knife. Note the sharp wound margins in image (e). Dogs may eat a decomposing body when they are left to starve, or they may kill living individuals, more often without eating them. It has been my experience that it is rare for a domesticated dog to eat an individual unless it is coaxed into it by exposing the hungry dog to initiate feeding activity. The dog should always be examined including the gastrointestinal contents. In one such case, neatly cut strips of soft tissue were found within the dog's stomach that were fed to the dog by the perpetrator to initiate more feeding activity.

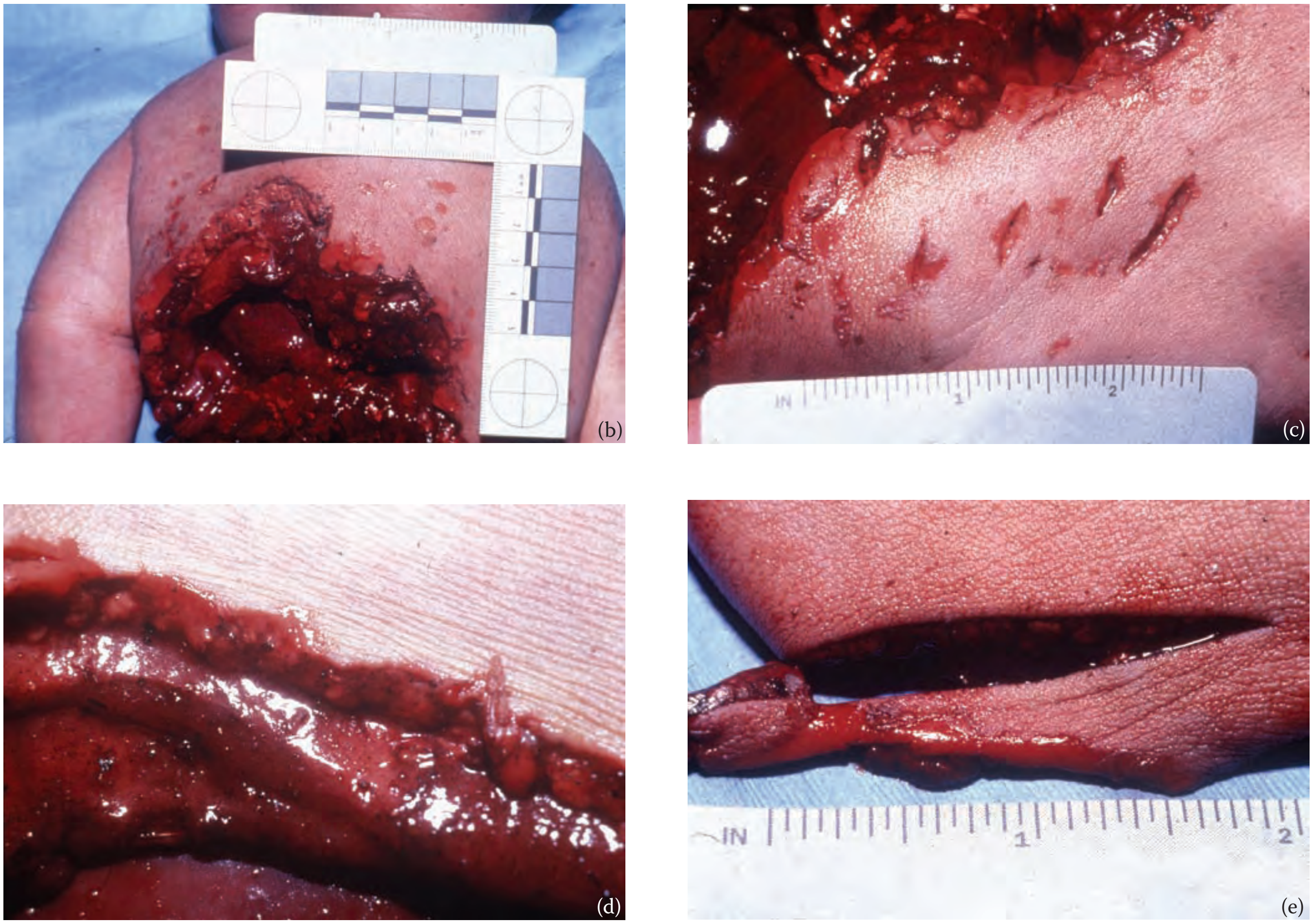

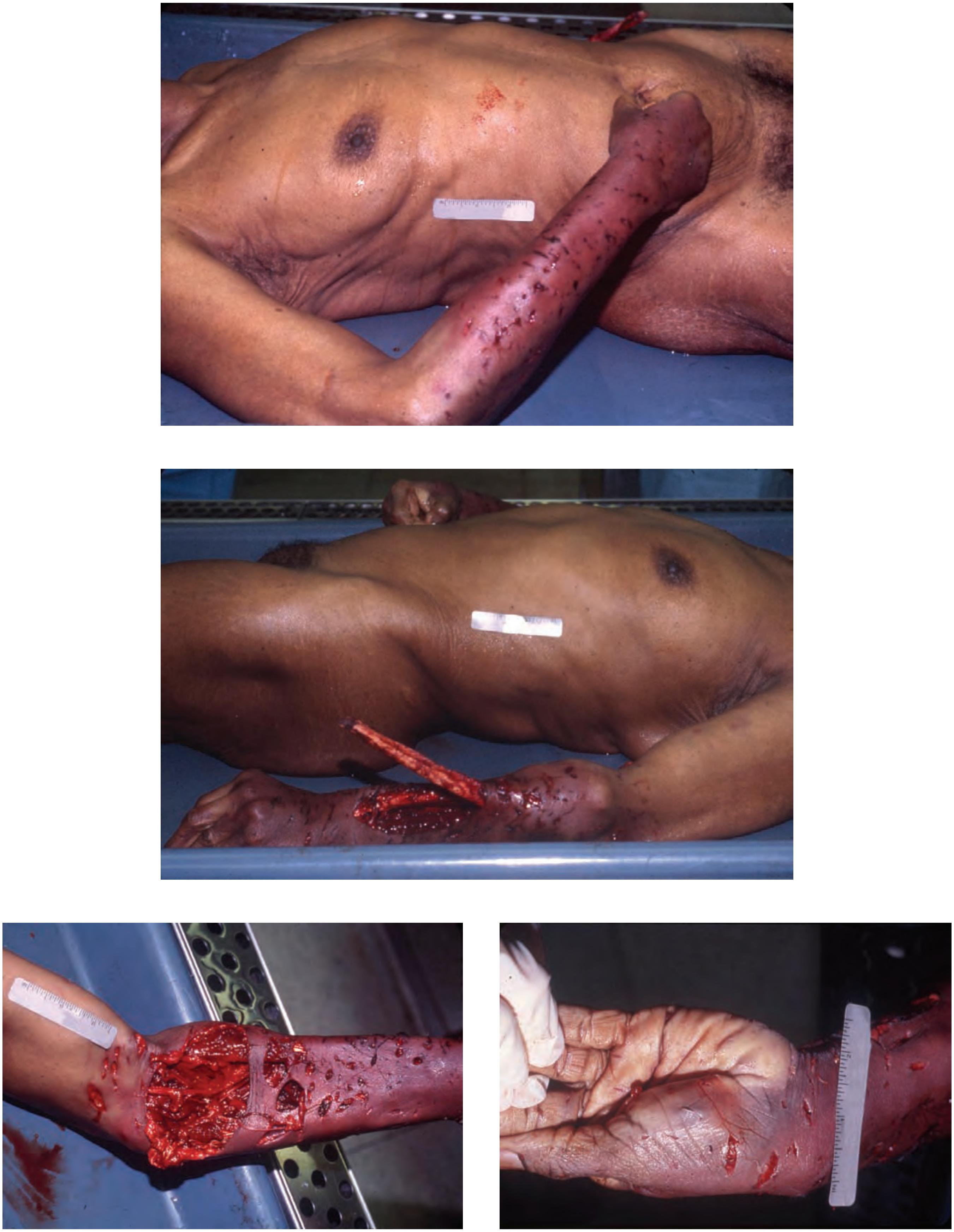

These are multiple puncture marks and lacerations with extensive adjacent hemorrhage due to an antemortem dog attack by a pit bull. This individual died as a result of blood loss.

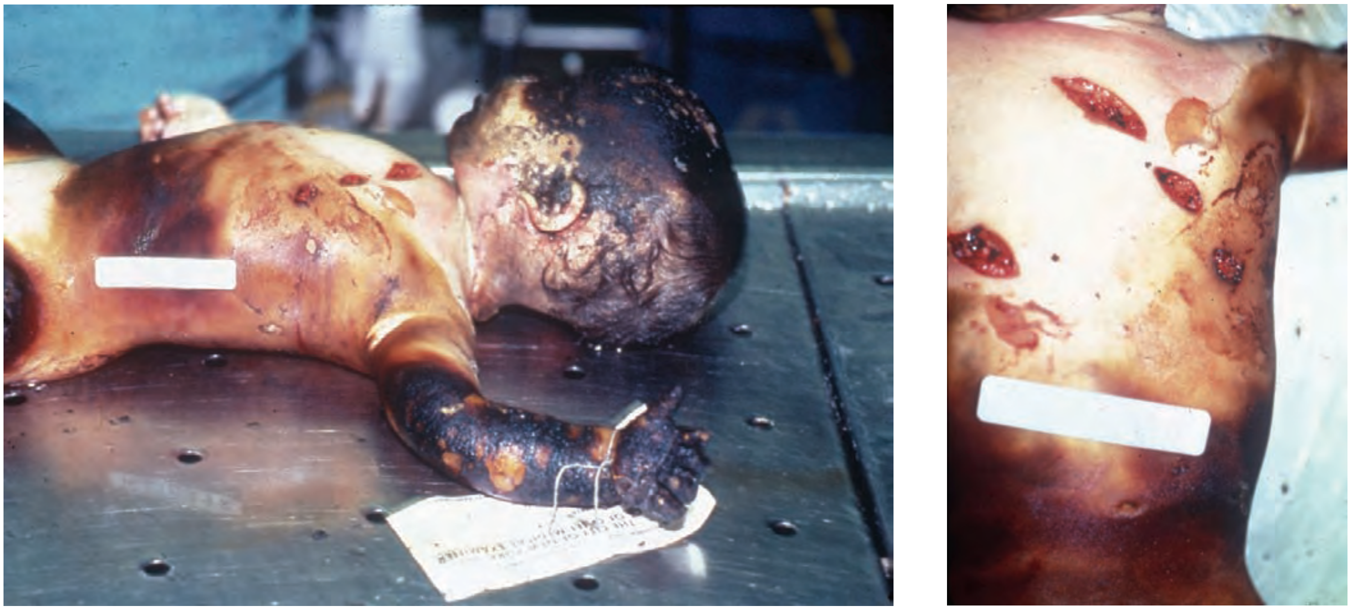

This perpetrator killed his family by inflicting multiple stab wounds and then setting their house on fire. Note the relationship between environmental factors such as heat and the change in wound characteristic. This is demonstrated to a greater degree in the following figures.

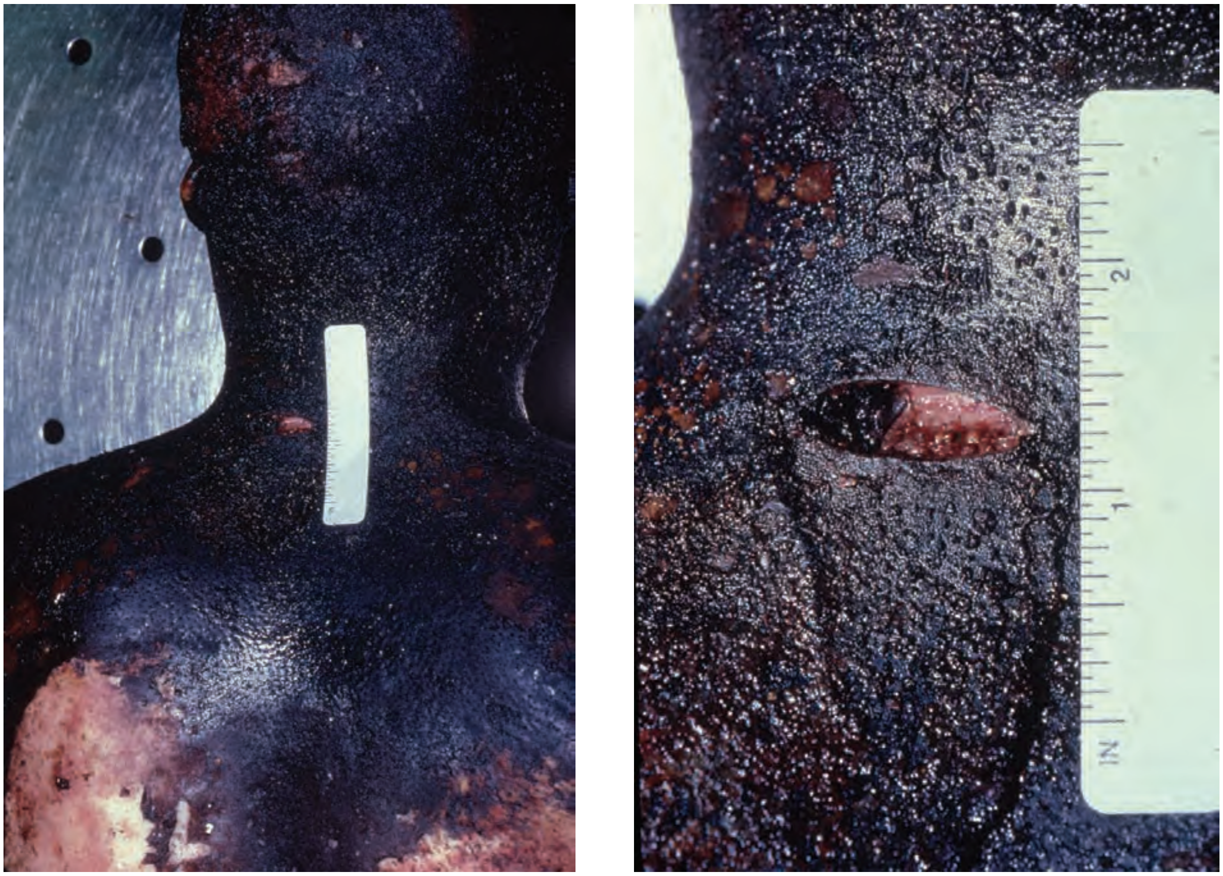

This individual was stabbed multiple times and his apartment was set on fire to conceal evidence of the crime. Note the charring of the skin with the sharply margined defect, which was from a stab wound. Internal examination revealed extensive hemorrhage throughout and surrounding the wound track. Sometimes the surface interpretation of such injuries can be challenging. Thermal damage may cause cracking of the skin, which may be misinterpreted as antemortem sharp- or blunt-force injuries.

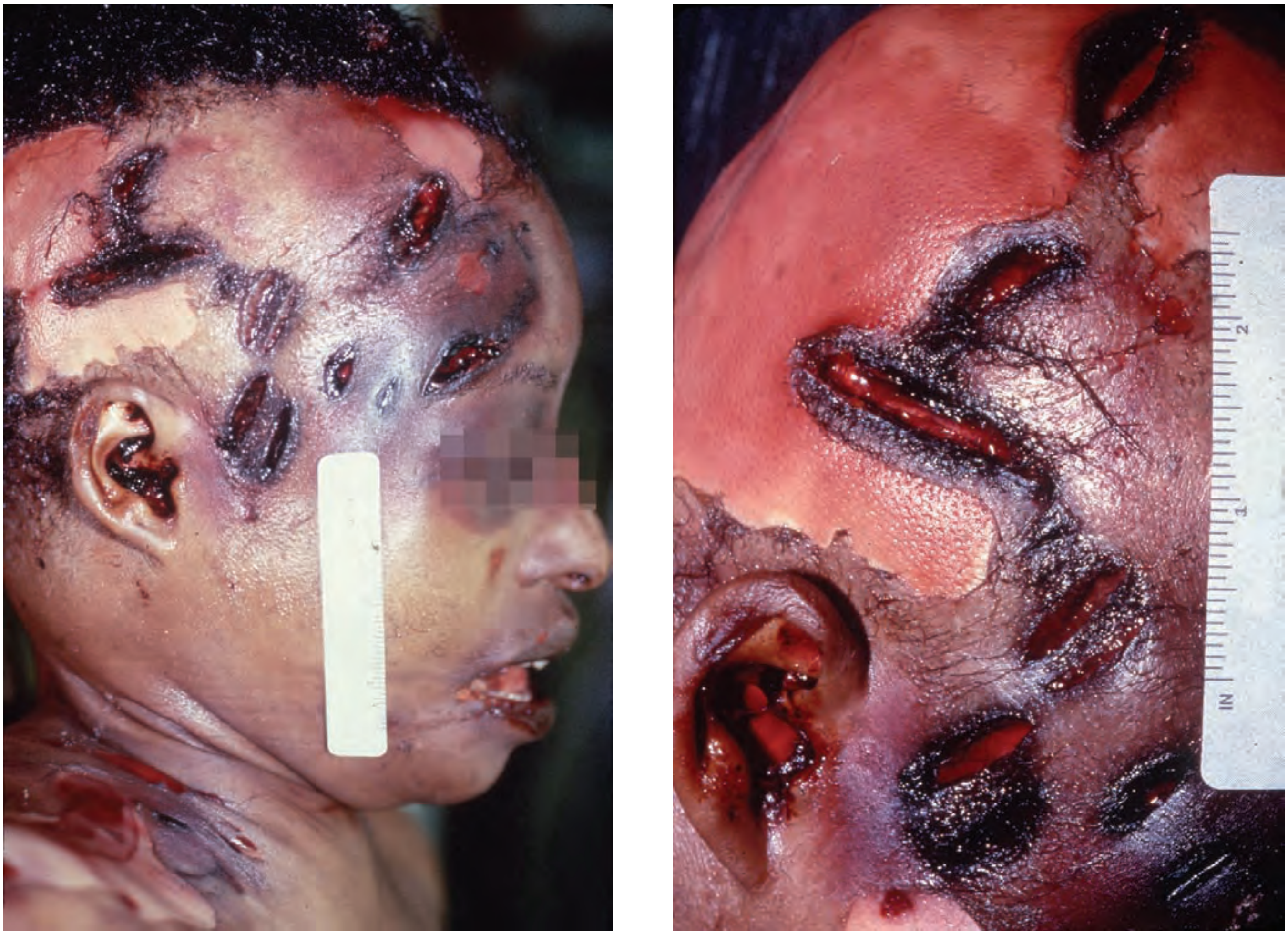

This individual sustained multiple blunt- and sharp-force injuries due to homicidal violence. She was found many hours after the assault during a hot summer month. Note the putrefactive changes with skin slippage and green to brown discoloration. Also note the wound margins are dry and dark. Some of these wounds could obviously be classified as lacerations or stab wounds and others could not be classified due to distortion by decomposition.

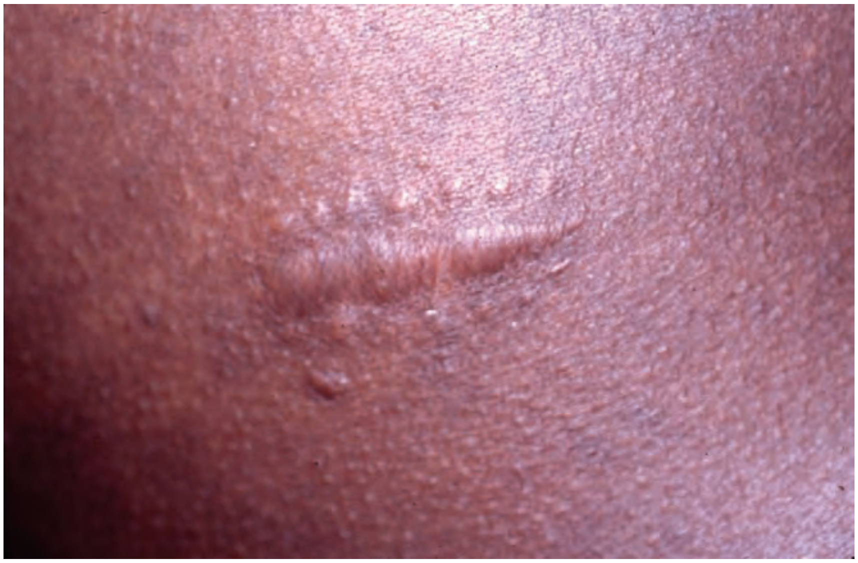

Healed stab wound.

Gunshot Wounds

8

CHARLES A. CATANESE, THOMAS GILSON, AND THOMAS ANDREW

Introduction

The evaluation of gunshot wounds is an area of importance in forensic medicine. Analysis of gunshot wounds should involve:

- Differentiation of wounds of entrance from exit
- Range of fire estimation for entrance wounds
- Determination of trajectory through the body
- Recovery of evidence (e.g. ballistics)

Whenever possible, clothing worn by a gunshot wound victim should also be examined.

Entrance and Exit Differentation

Entrance Gunshot Wounds

Most entrance wounds appear as circular perforations with a margin of abrasion, a collar of abraded skin, and the entrance defect. The margin of abrasion is produced by stretching with eventual tearing of the skin surface as the bullet enters the body. It is the most reliable feature in identifying entrance gunshot wounds. Exit gunshot wounds lack this feature, as the undersurface of the skin is stretched outward when a bullet exits the body. The margin of abrasion may provide useful preliminary information regarding trajectory of a bullet through the body. When the bullet enters the body perpendicular to the surface, the margin of abrasion is symmetric. As the bullet enters the body on an angle, the margin will be elongated on the side where the bullet first contacts the skin. The trajectory is confirmed by subsequent internal examination.

Exit Gunshot Wounds

These types of wounds typically appear as slit-like or irregular perforations without margins of abrasion. An exception to this is a shored or supported exit, which occurs when a bullet exits from the body where the skin is firmly supported. This might occur as a bullet exits when an individual is leaning against a wall, lying on the sidewalk, or even wearing tight-fitting clothing. These wounds generally have a round appearance and may have superficial abrasion around the defect. Unlike the true abrasion margin of an entrance wound, the shored exit abrasion tends to lack any significant depth and is often very irregular.

Atypical Gunshot Wounds

These are entrance wounds characterized by an irregular appearance. They may have irregular or obscure margins of abrasion. The skin adjacent to the entrance wound may show irregular abrasions or lacerations. Sometimes these perforations may appear as irregular tears. Ricochet bullets, or bullets that pass through intermediate targets, often produce atypical entrance wounds. Sometimes the jacket and the slug may separate and produce two separate irregular entrance wounds. Bullets may strike surfaces such as concrete and fragment into pieces, causing multiple irregular defects from impacts of both twisted metal and rock fragments. Fragments from upholstery, chrome, plastic, etc. may be found at the wound surface or even within the wound track. This atypical appearance may also be seen in certain areas of the body where the skin is thicker (palms and soles) or the skin contour is irregular (face). Atypical wounds may be difficult to interpret, particularly in perforating gunshot wounds where the bullet passes through soft tissue only. Fracture characteristics of bone (especially flat bones such as the skull) and lead fragmentation patterns may give great insight into establishing direction of fire.

Range of Fire Estimation

In estimating range of fire, an attempt is made to determine the distance from the end of the barrel of the gun to an entrance wound at the time the weapon is discharged. This is most reliably estimated by observing the wound and trying to re-create an observed pattern by test firing the weapon. Because the weapon is often not available, estimates based on the physical characteristics of the wound are frequently employed.

An understanding of ammunition aids in understanding range of fire estimation. The ammunition placed into a firearm is a cartridge. A cartridge consists of a bullet, which rests atop a casing containing gunpowder. A primer located at the base of the cartridge ignites the powder. The primer is itself ignited by the mechanical action of the hammer of the firearm, when the hammer strikes the base of the cartridge. The bullet is the portion of the cartridge that exits the barrel. It is propelled by the burning gunpowder. In general, there is some residual gunpowder that does not burn. In addition to the bullet, burned and unburned gunpowder exits the barrel when a weapon is discharged. This forms the basis of range of fire and estimation.

Range of fire can be divided into three major categories. These include *close range, intermediate range,* and *distant range.* As noted, range of fire is best estimated by the use of test firings employing the known weapon with similar ammunition. Lacking these ideal conditions, estimates can be made on the basic wound characteristics as described next. Even when test firings can be performed it is worth remembering that the wounding characteristics of skin are not identical to gunpowder residue deposition on a test-firing cloth.

Close Range of Fire

This is characterized by the presence of *soot* on the adjacent skin surface or within the wound track. Soot is burned gunpowder residue with a dark powdery appearance. It is critical to remember that soot can be wiped from the body's surface, so gunshot wounds need to be examined prior to any washing, whenever possible. The presence of soot indicates the end of the barrel of the gun was held within approximately 6 to 8 inches from the body's surface. Different guns may have different ranges. The closer the end of the barrel is to the body, the denser the soot deposition. As this distance increases, the soot deposition becomes sparser. Soot may be present with stippling (see intermediate range of fire), but this is more accurately still considered close range of fire. When the end of the barrel is very close to the body, there may also be searing of the adjacent skin from burning gunpowder and flame that extends from the end of the barrel. Soot can be filtered away through intermediate targets such as clothing, car doors, walls, etc. This pattern becomes more dispersed as the distance becomes greater.

Within the close-range-of-fire category are contact gunshot wounds where the end of the barrel touches the body. These may be further subdivided into *tight contact* and *loose contact.* Some classify wounds where the barrel is very close to the body as *near contact*. The term "close contact" is redundant, potentially confusing, and best avoided.

Tight Contact Gunshot Wounds

These are produced when the barrel of the gun is held tightly against the body's surface. Most of the soot will be deposited within the wound track and may not be apparent until autopsy examination. Small amounts of soot may be present at the wound margins. Depending on the body region, type of gun, type of ammunition, and force with which the gun is pressed against the body, the wound may appear different. There are often abrasions adjacent to the gunshot wound. These are associated with the muzzle of the gun rubbing against the body's surface as the gun is discharged. These may take the form of a semicircular rim or a complete imprint pattern of the entire muzzle. When the gun is held tightly to the body, particularly over a bony surface such as the skull, there may be multiple radiating linear lacerations. These lacerations may be small or large and are more commonly associated with larger-caliber bullets. Approximation of the laceration margins reveals a central circular perforation typical of an entrance wound with abraded margins. Observation of the muscle under a tight contact wound may reveal a pink discoloration owing to the introduction of combustion products into the wound track such as nitrites and especially carbon monoxide binding with myoglobin.

Loose Contact and Near Contact

When a gun is held less tightly against the body, soot may escape more readily to the skin's surface. If the gun is discharged at an angle to the body's surface, the pattern of soot dispersion will extend over a greater surface area in the direction of fire. A partial muzzle imprint may be observed. In near contact wounds, the soot is deposited in a dense, relatively small area around the entrance and searing of the skin may be observed as noted above.

Intermediate Range of Fire

This range of fire is characterized by the presence of *stippling* defects around the entrance wound. These are small, (approximately 1/16 inch) red to brown, punctuate abrasions of the skin's surface. Stippling indicates the end of the barrel of the gun was held approximately 18 to 24 inches away from the body's surface. These defects are produced primarily by fragments of unburned gunpowder that exit the barrel with the bullet. These fragments strike the body with greater force than soot and produce permanent defects. Such defects cannot be wiped away from the body's surface. The appearance of stippling may vary depending on the type of gunpowder used in the cartridge, (e.g., ball, disk, or flake).

The term *stippling* applies to the defects in the body's surface. This pattern will become more dispersed as the distance from the weapon to the body's surface increases. Like soot, stippling can be dampened or completely filtered by clothing or other intermediate targets. Powder residue found on intermediate targets at this range of fire is characterized by small, separate, gunpowder fragments.

Soot or stippling defects may be present on an outstretched hand but not surrounding the entrance wound. For this reason, close examination of the upper extremities in gunshot wound cases may be helpful. Range of fire is estimated in the same way and the distance between the weapon and the entrance wound can be approximated by factoring in the distance from the wound to the upper-extremity findings.

Occasional "pseudostippling" may be seen around an entrance wound. This phenomenon is generally seen when a bullet passes through an intermediate target and fragments of the intermediate target strike the skin's surface (causing abrasions) around the entrance wound. As noted, such entrance defects tend to be irregular and the adjacent stippling defects exhibit wider size variation than true stippling, where the abrading gunpowder tends to produce abrasions of relatively uniform size.

Distant Range of Fire

This range of fire is characterized by the absence of gunpowder residue around an entrance gunshot wound without an intermediate target between the end of the barrel and the wound. This indicates the end of the barrel of the gun was held more than approximately 18 to 24 inches away from the body's surface. Some experts prefer to call this intermediate range, especially when the possibility of an intermediate target cannot be excluded.

Trajectory Determination

The path a bullet takes through the body may have substantial medicolegal significance. In some cases, the trajectory can be easily determined by careful dissection. In other cases, (e.g., multiple gunshot fatalities), full delineation of wound tracks may be harder to accomplish as a result of extensive internal injury. Surgical intervention may also obscure wound tracks in part. It is also worth remembering that a wound sustained in a certain body position may look slightly out of alignment with the body lying supine on the autopsy table. Superficial wounds can pose difficulties in trajectory determination as entry and exit wounds may not be easily identified. These include graze and tangential gunshot wounds.

Graze-Type Gunshot Wounds

These are produced when the bullet contacts the skin superficially at a very narrow angle, producing an elongated superficial oval abrasion. One may determine direction of fire by a semicircular margin of abrasion at the entrance side and a more irregular margin at the exit side.

Tangential Gunshot Wounds

These are produced when the bullet strikes the skin superficially at a narrow angle and creates lacerations of the underlying subcutaneous tissue and overlying skin. Direction of fire may be established by the presence of a semicircular margin of abrasion at the entrance side. Also, the laceration of the skin overlying the wound track produces tears with skin tag formation. These tags will extend outward on angles. The tips of these lacerations, at their most medial aspects (e.g. within the wound track), point toward the side from which the bullet entered the body.

Gunshot Wounds in Bone

Gunshot injuries of bone may be helpful in trajectory determination, especially in the skull and, less frequently, other flat bones. When a bullet passes through the skull, it creates a cone-shaped beveled defect. The entry point has crisp margins and the exit point is larger and represents the wider portion of the cone. Bullets entering the skull produce internal beveling and those exiting the skull produce external beveling. A tangential bullet strike of the skull may cause the bullet to fragment and result in complex beveling if a portion enters the cranial cavity and a portion shears off and remains outside the surface. In such cases, the bony defect created may exhibit internal and external beveling. Such defects are referred to as "keyhole" defects, as their appearance may resemble a keyhole.

High-Velocity Gunshot Wounds

Trajectory delineation in high-velocity gunshot wounds is similar to low-velocity injury, however, the extent of injury is far more significant and exit wounds are often dramatically large. Hunting ammunition used at high velocity will fragment more in the body than low-velocity bullets, creating a characteristic x-ray picture called a "lead snowstorm."

Evidence Recovery

Bullets retained in the body must be recovered at autopsy. Rifled firearms leave characteristic markings on bullets that can be used for identification of a firearm by comparison with projectiles known to have been fired from that weapon. Care must be taken during dissection to avoid or minimize the creation of any additional marks on the bullet that might obscure the rifling marks. This also applies to any identifying marks inscribed place directly on the bullet after recovery. These identifying marks should be inscribed on the base of the bullet. Older bullets should be recovered at the time of autopsy. These may be enclosed in fibrous capsules at the site of lodgment and often have a dull gray appearance as a result of oxidation. Rifling marks may still be present and useful for ballistic comparison.

Shotgun Wounds

Because of differences in construction and ammunition, shotguns merit special consideration. Shotguns can be used to fire a single projectile (slug) or several pellets (shot). Shot size can vary (buckshot versus birdshot). Slugs behave essentially as single projectiles. The shotgun cartridge (or shell), contains some additional components (e.g., wadding, filler material) that might produce injuries that can be used in range of fire estimation. Soot and stippling remain useful in close- and intermediate-range determinations. The wounds produced by shotgun shells containing shot look different from conventional bullet wounds. Soot, stippling, wadding, and other components can be used for range of fire estimation, but the behavior of the pellet cluster over distance provides additional information regarding range of fire. When the shot exits the barrel, it travels initially as a tightly grouped cluster. Striking the skin, the tight cluster produces a round defect. The grouping opens up over distance with the first noticeable change in wound appearance occurring at about 3 feet. A single defect is still observed but the margins now take on a scalloped appearance. With further distance (and more pellet dispersal), wounds now consist of a central defect with scattered satellite defects surrounding them. These satellite defects increase in number with greater distance, until the central defect is lost entirely and pellets maintaining sufficient velocity strike the skin's surface individually. The dispersal pattern of pellets is not predictable once an intermediate surface intervenes. It is not possible to evaluate these wounds for range estimations and, for similar reasons; estimates of range of fire by x-ray examination are unreliable. As with other firearms, test firings are the most reliable means for range of fire approximations.

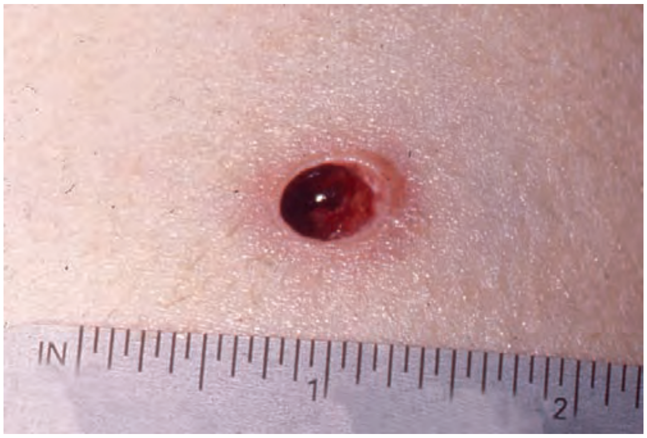

Distant entrance gunshot wound. This bullet struck the body nose end first, roughly perpendicular to the surface. Note the thin and roughly symmetric margin of abrasion surrounding the slightly oval perforation.

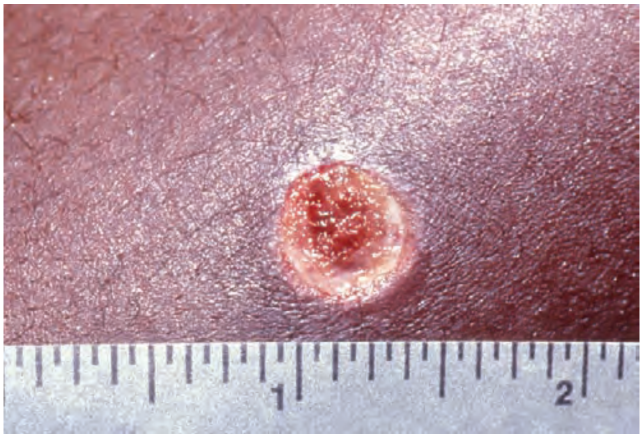

Distant entrance gunshot wound in a darker-skinned individual. Note the slightly asymmetric margin of abrasion, which is greatest at its inferior aspect. This indicates the bullet struck the body nose first, and at a slight upward trajectory. These directions are stated *relative to the body.*

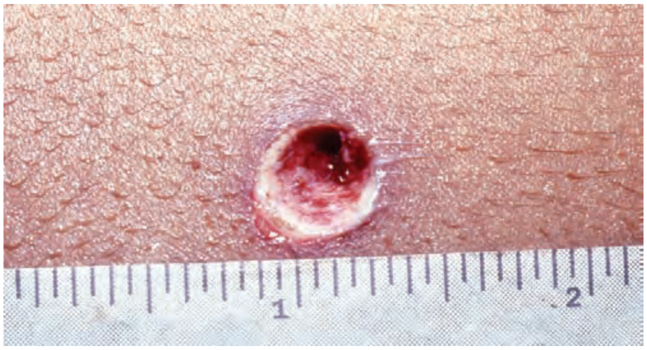

Distant entrance gunshot wound with slightly asymmetric margin of abrasion, which is greatest at the right inferior aspect. This usually indicates the bullet struck the body nose first, almost perpendicular and at a slightly upward and right-to-left trajectory. These directions are stated *relative to the body.*

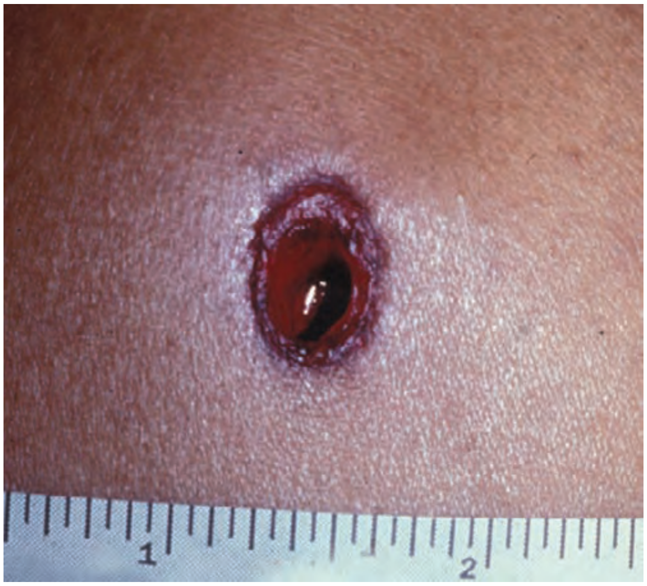

Distant entrance gunshot wound with slightly asymmetric margin of abrasion greatest on its superior and left aspect. This usually indicates that the bullet struck the body nose first, almost perpendicular and at a slightly downward, left-to-right trajectory. These directions are stated *relative to the body.*

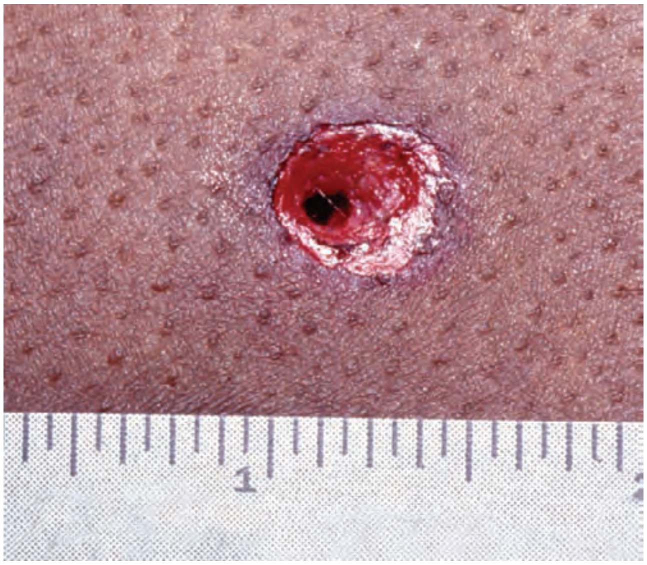

Distant entrance gunshot wound with slightly irregular, widened margin of abrasion, which is greatest at the right lateral aspects. This indicates the bullet struck the body from a slightly inferior and right-to-left trajectory. The slightly irregular nature of the margin of abrasion may sometimes be seen, as the bullet perforates clothing first.

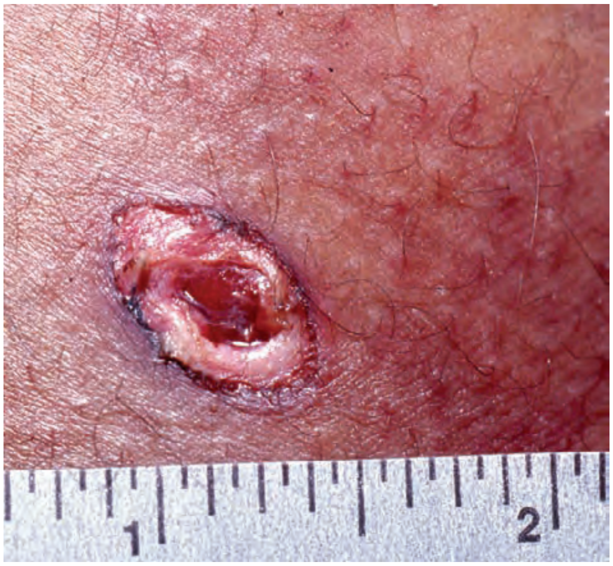

Distant entrance gunshot wound that is slight stretched into an oval due to skin tension. Note the slightly asymmetric margin of abrasion, which is greatest at the 9 o'clock through 12 o'clock position, indicating the bullet struck the body with a slightly downward and left-to-right trajectory.

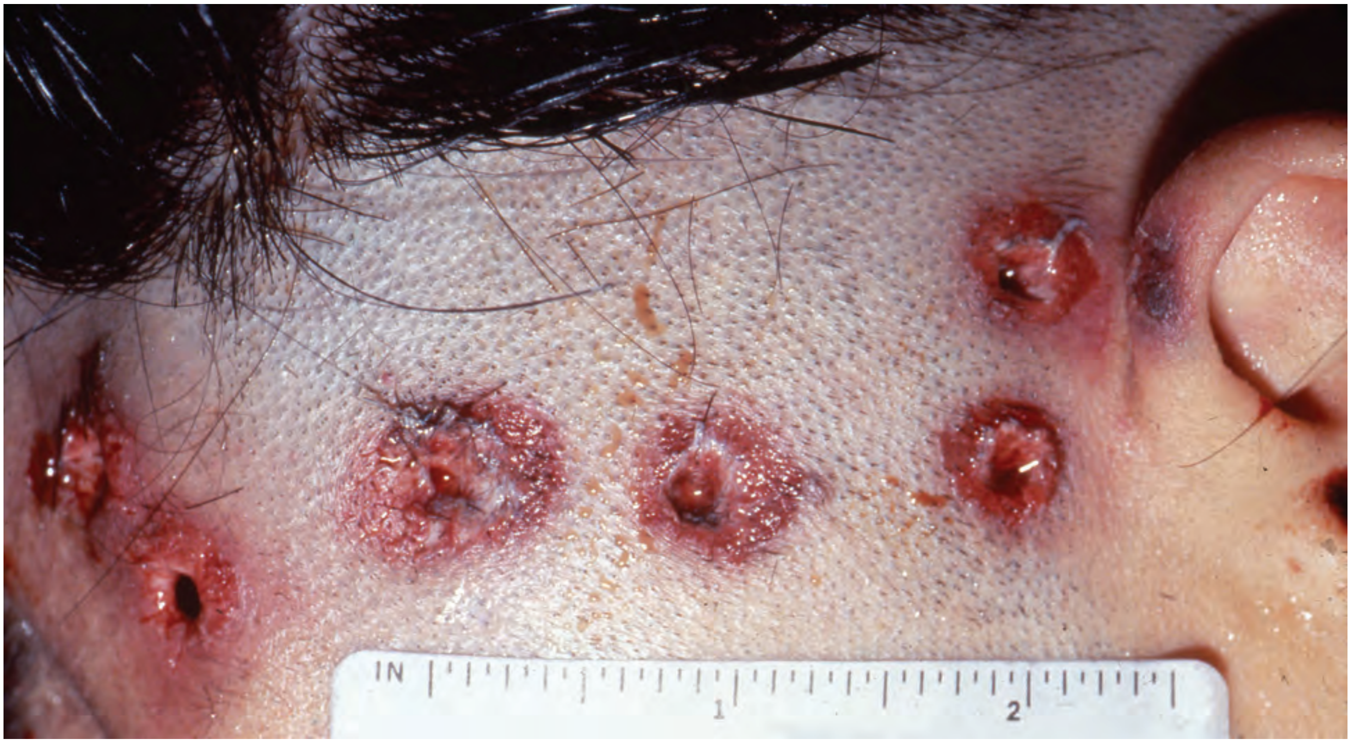

There are six separate clustered entrance gunshot wounds to this individual's head. Note the abrasion to the posterior ear helix caused by a bullet. Shaving the hair is recommended for external examination of such gunshot wounds.

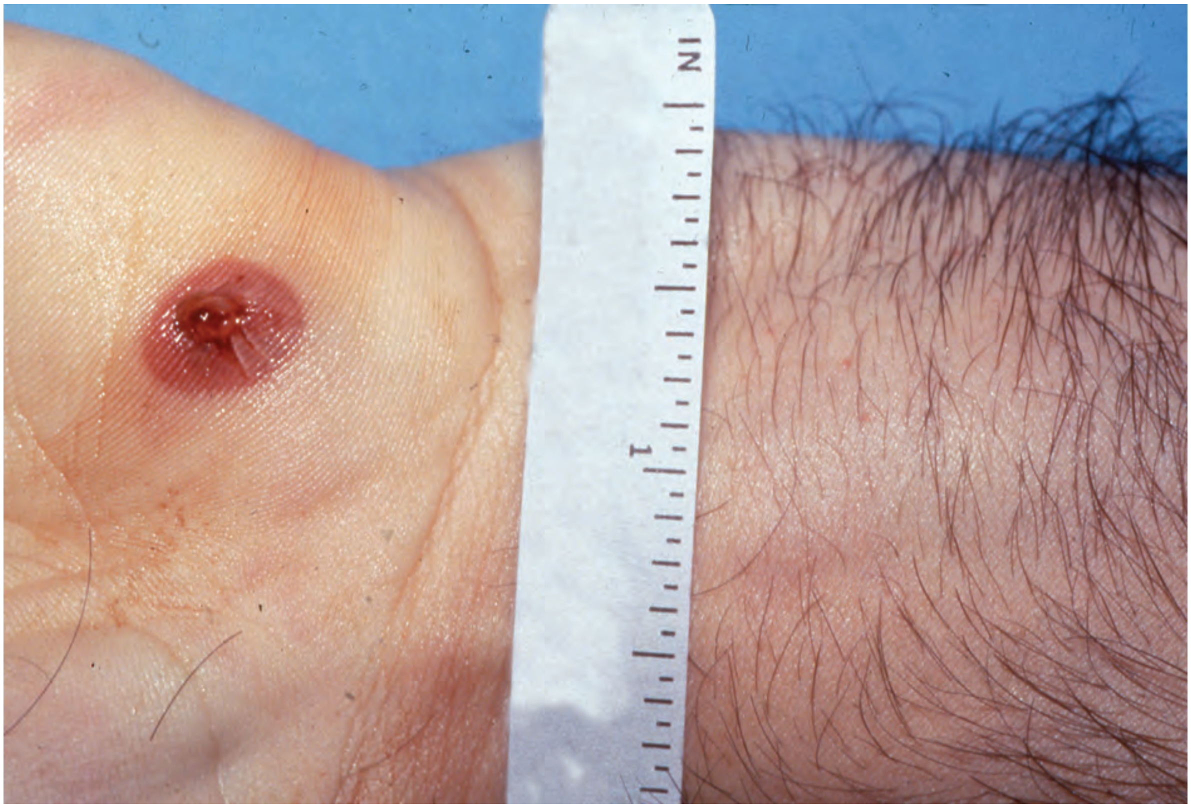

Distant entrance gunshot wound to the palm of the hand through thick skin.

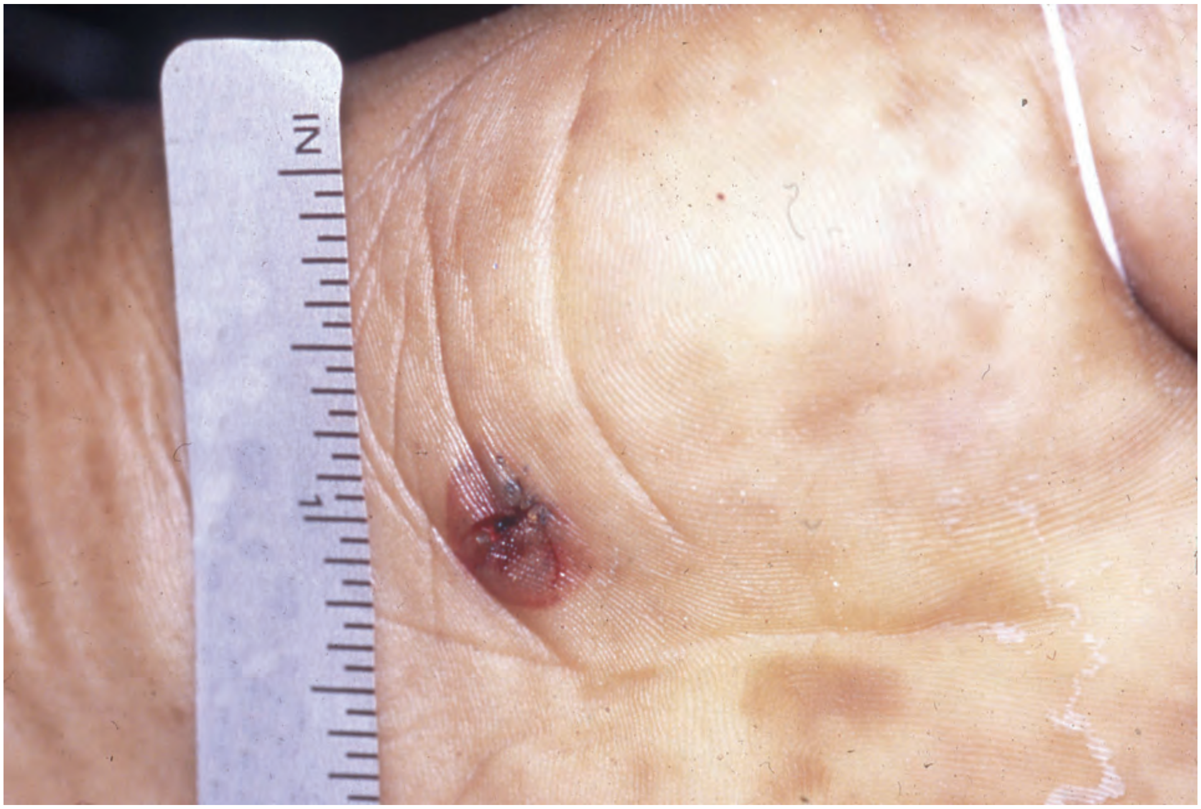

Distant entrance gunshot wound through the thick skin of the sole of the foot. This bullet first perforated the bottom of a shoe and a sock before entering the body.

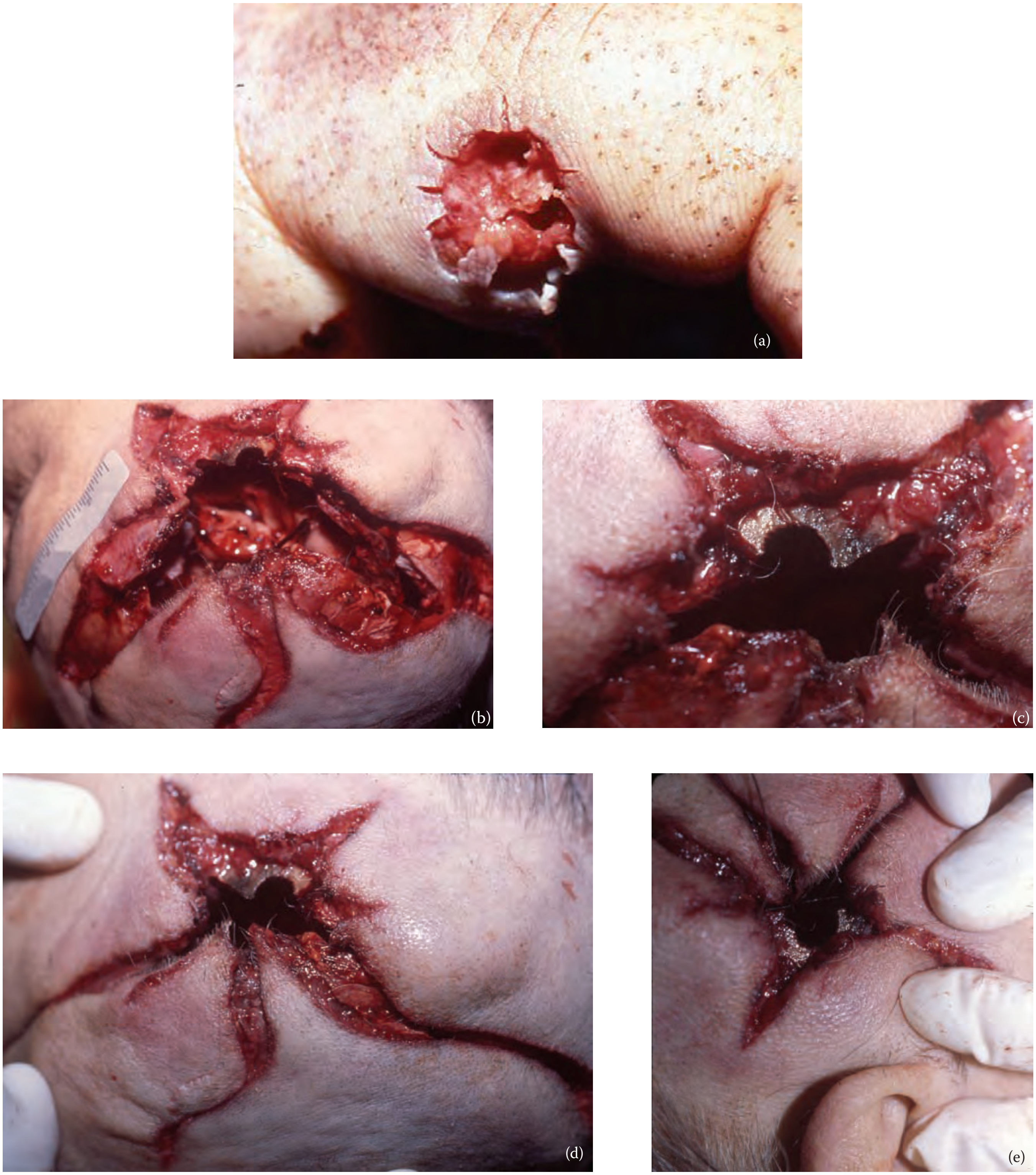

Entrance gunshot wound with surrounding stippling defects indicating intermediate range of fire. This gunshot wound (a) to the thick skin of the palm of the hand reveals typical small radiating lacerations without prominent margin of abrasion. These characteristics are typical for gunshot wounds of the palms of the hand and soles of the feet. These images (b–e) represent the same contact gunshot wound with varying degrees of tension surrounding the wound's surface. Note that with approximation of the margins (d–e) the wound forms a roughly circular perforation with margin of abrasion typical of an entrance gunshot wound. Also note the copious amounts of soot within the wound track and at the underlying bone surface. Such large radiating linear lacerations are usually associated with higher-caliber guns and tight contact of the gun muzzle on the body surface with underlying bone close to the skin. The expanding gases from burning gunpowder forced into the wound track causes such lacerations.

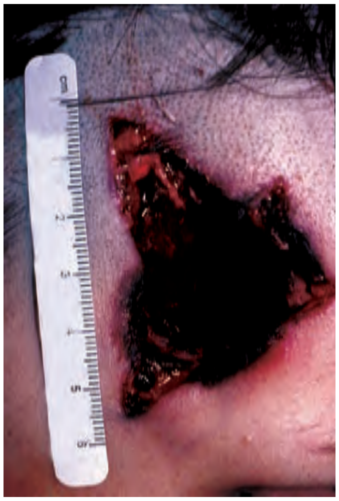

Contact gunshot wound to the scalp with radiating lacerations and soot deposition within the wound track and at the wound margins. These lacerations are caused by burning gunpowder, with expanding gases forced into the wound track and causing the overlying skin to lacerate.

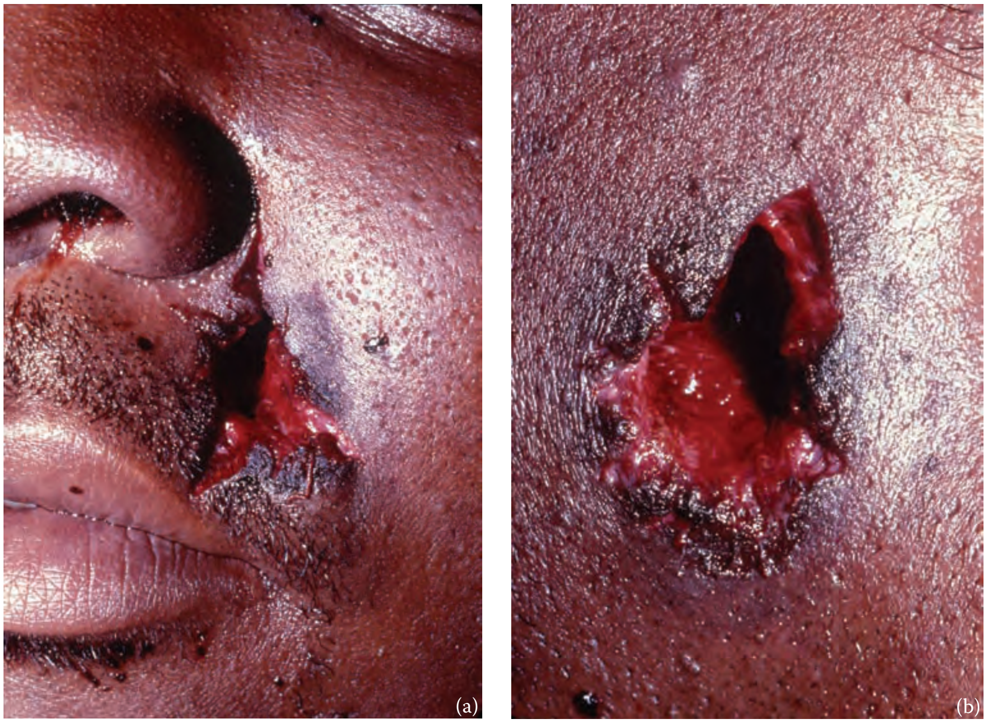

Contact gunshot wound to the face (a). Note the stellate lacerations radiating from the perforation site with soot deposition; (b) a close-up view with a different orientation.

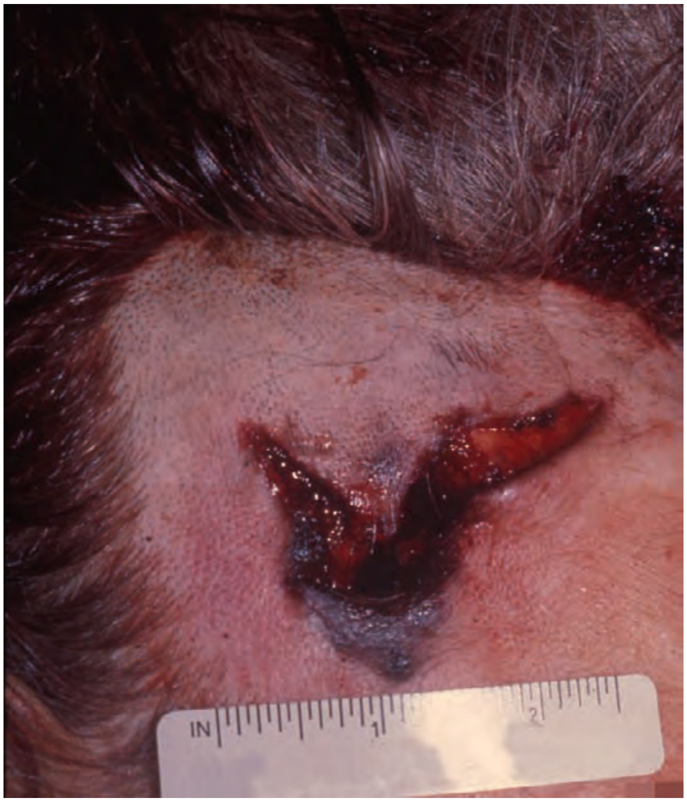

Contact gunshot wound to the right temporal region. Note the larger lacerations extending from the perforation site with adjacent soot deposition.

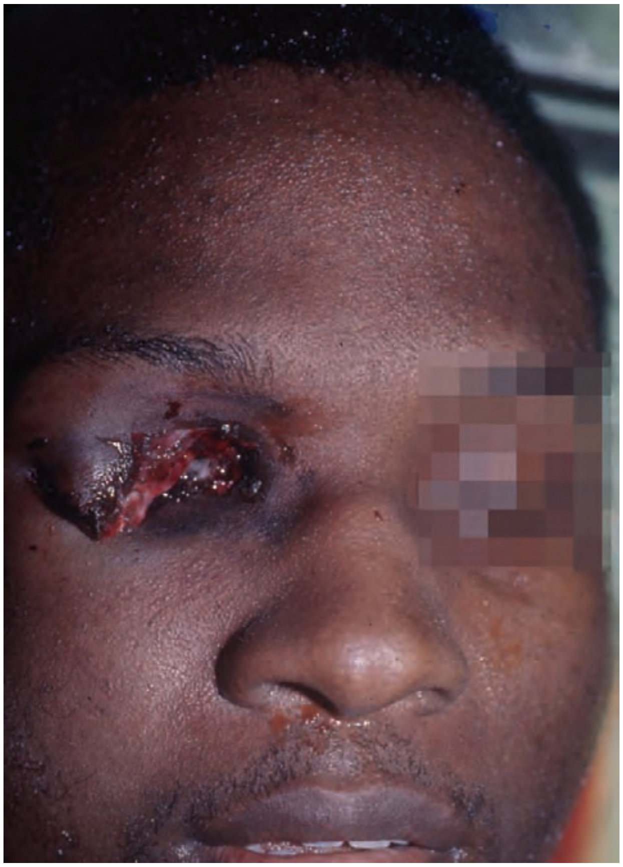

Contact gunshot wound to the eye. Note the lacerations of the eyelid and eye globe with soot deposition at the adjacent skin and within the wound track.

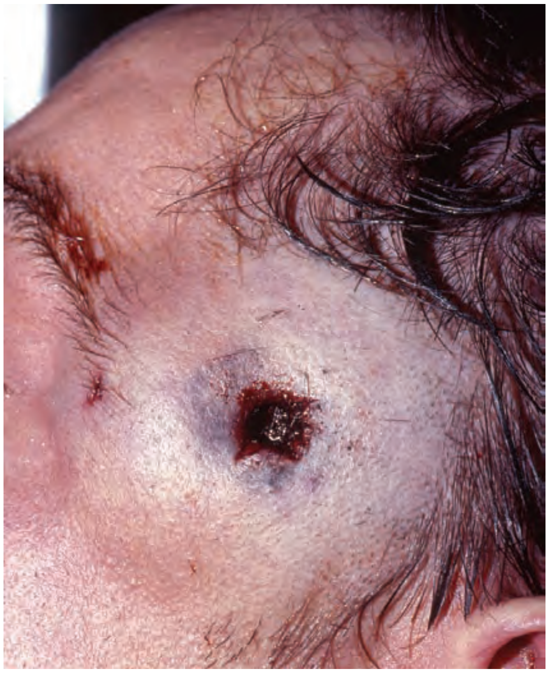

Contact gunshot wound to the temple. Note the small radiating lacerations from the perforation site. Note the presence of soot within the wound track and adjacent skin.

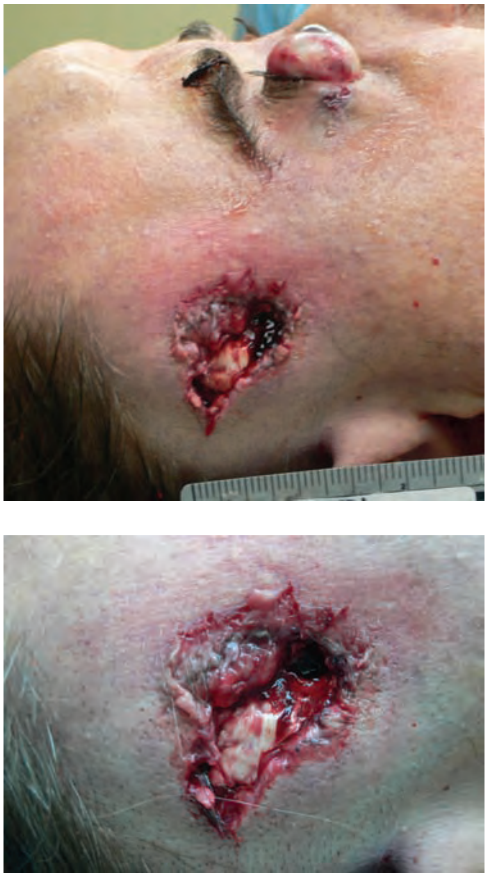

Note this tight contact gunshot wound to the right temporal region with lacerated borders and soot deposition. A .357 Magnum caused a comminuted skull fracture with head deformity and the eye protruding from the socket.

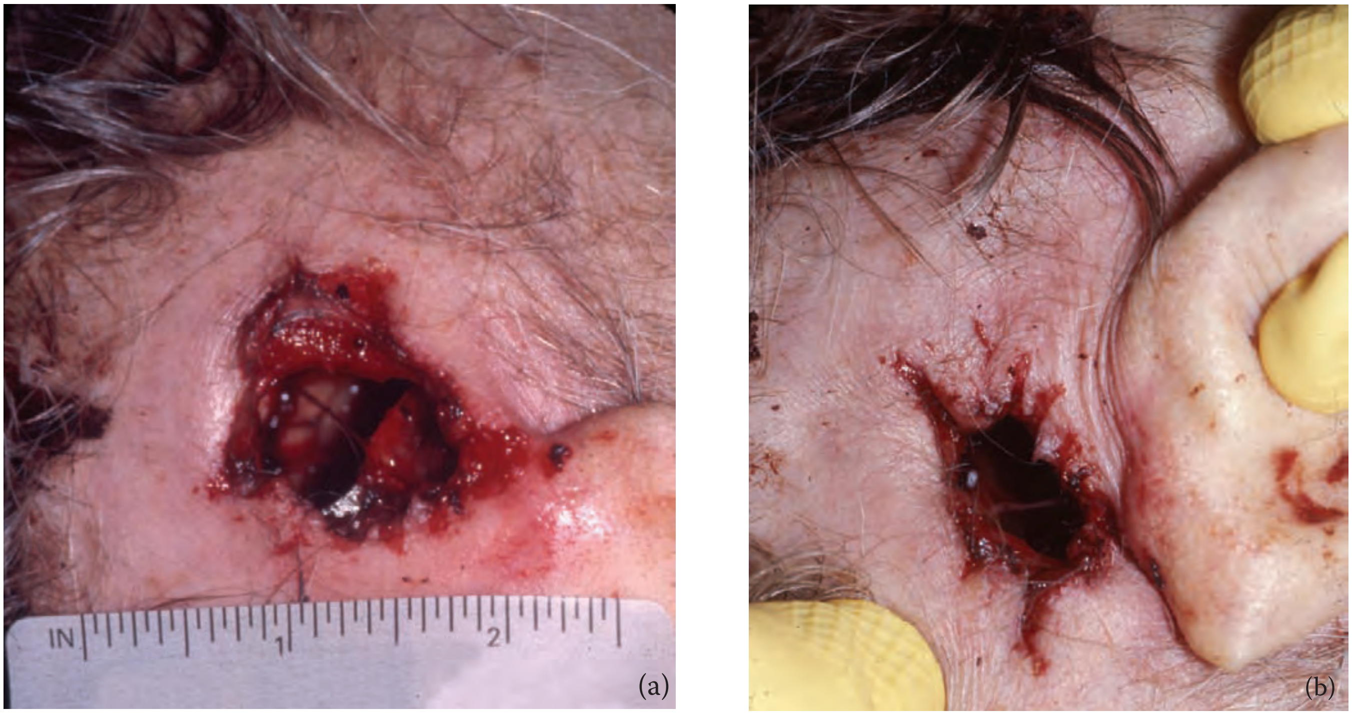

Contact gunshot wound to the scalp. (a) Note the small radiating lacerations with soot deposition. (b) Note the same wound after cleaning. The tension surrounding the wound is relieved by pressing the surrounding skin inward toward the perforation. This will approximate the wound margins and reveal a roughly circular perforation with margin of abrasion characteristic for an entrance gunshot wound. Note that even after cleaning, small amounts of soot are still present at the wound margins. Further cleaning might eliminate this gunpowder residue as well. In many cases, such as this one, it is essential to photograph the body before cleaning.

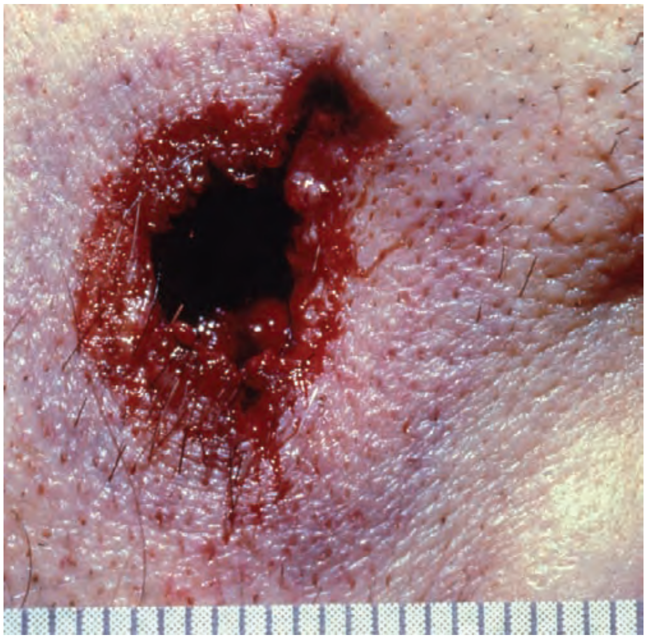

Contact entrance gunshot wound to the temple. Note the abraded imprint of the eyepiece portion of the gun at the superior aspect. The perforation site has a margin of abrasion. There is no obvious soot at the surface of the adjacent skin, but there were copious amounts of soot within the wound track. Note the pink to red discoloration surrounding the perforation due to nitrates and carbon monoxide released from burning gunpowder. These components may sometimes cause this discoloration when reacting with the underlying muscle. In this case, it is more obviously due to the decedent's light skin color.

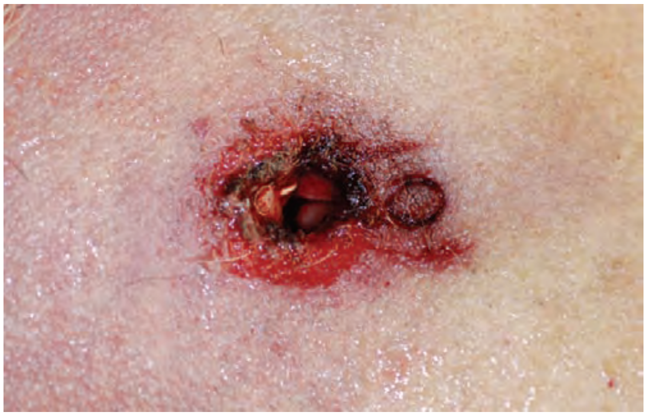

Contact gunshot wound with muzzle imprint.

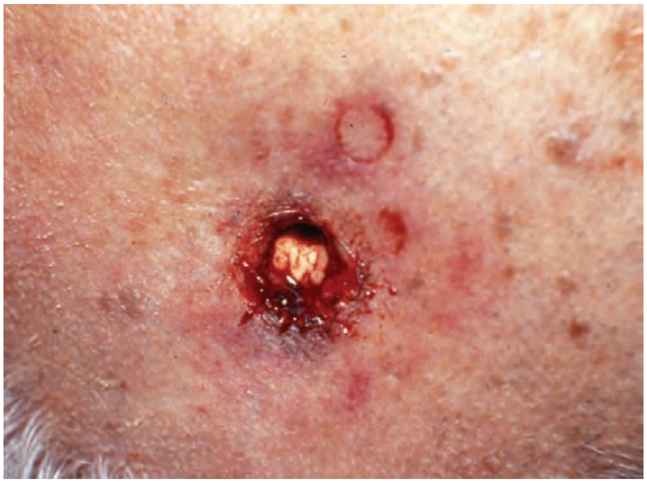

Contact gunshot wound. Note the adjacent abrasion to the perforation site due to contact with a revolver ejector rod when the gun was discharged. There are also small amounts of soot visible at the wound margins and more within the wound track.

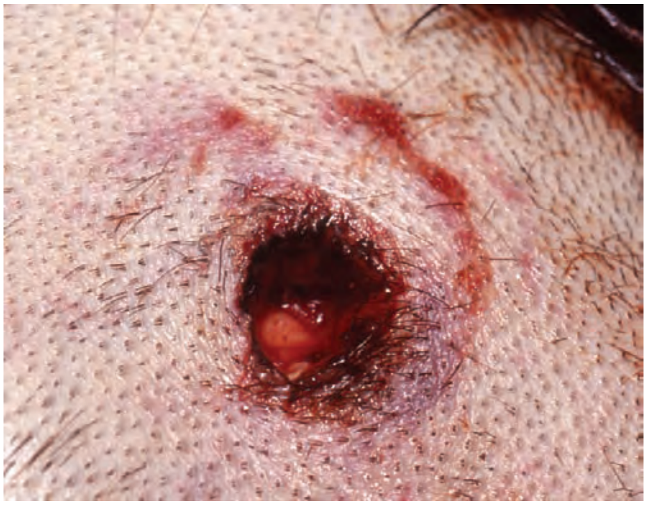

Contact gunshot wound to the scalp with an abrasion at the superior left aspect due to the gun muzzle. There is visible soot within the wound track. Sometimes it may be difficult to visualize the presence of gunpowder residue on thick, dark, scalp hair.

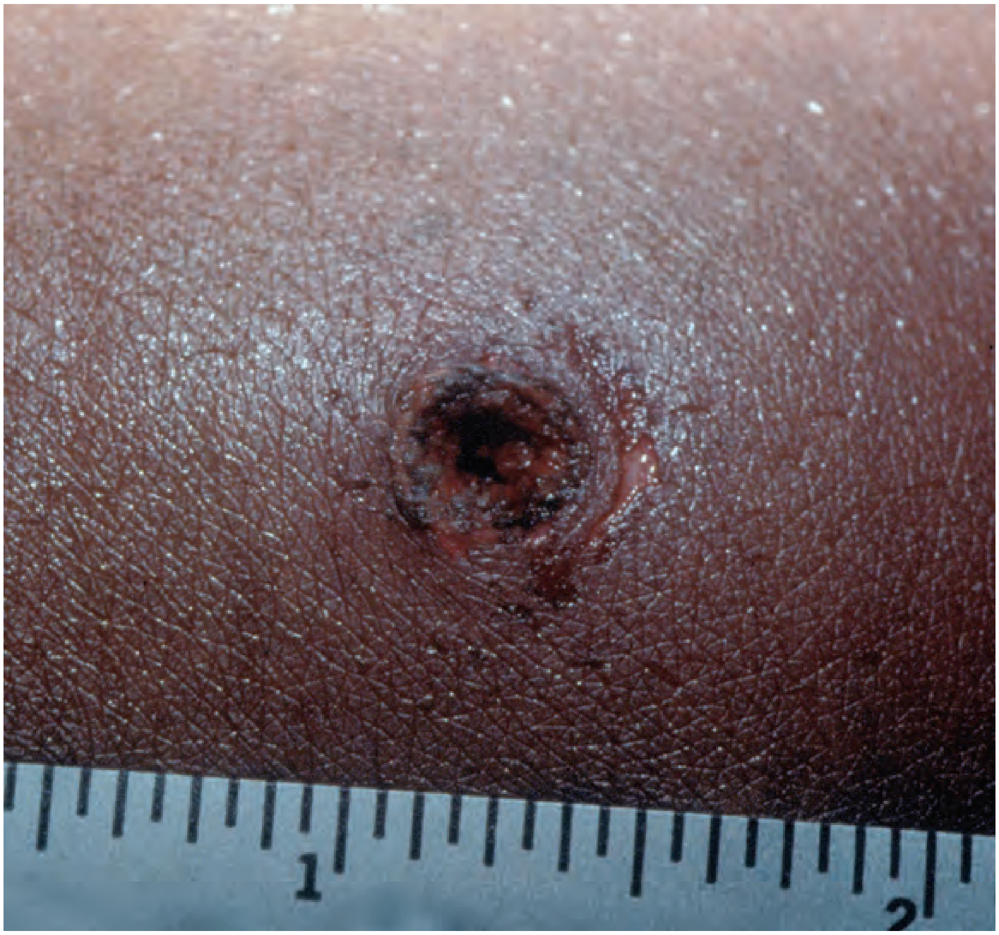

Contact gunshot wound. Note the soot at the margin of the perforation site. Note the surrounding abrasion from a partial muzzle imprint.

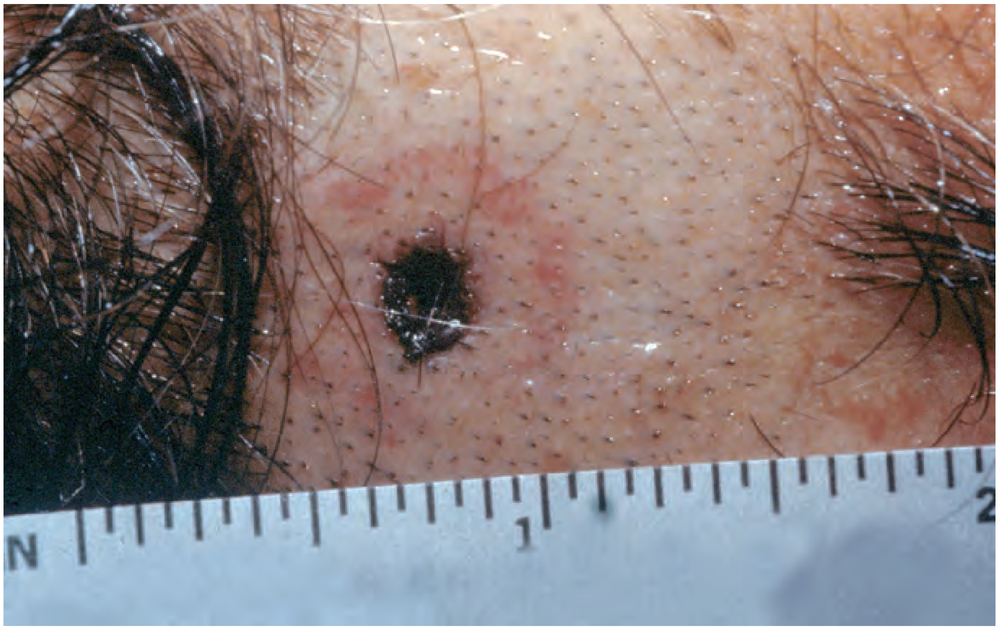

Contact gunshot wound to the right temporal scalp. Note the soot deposition surrounding the margin and the red discoloration of the adjacent skin.

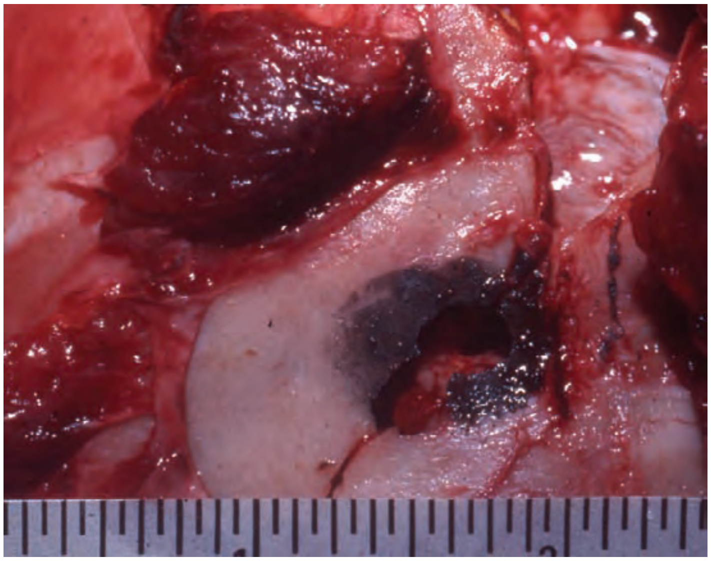

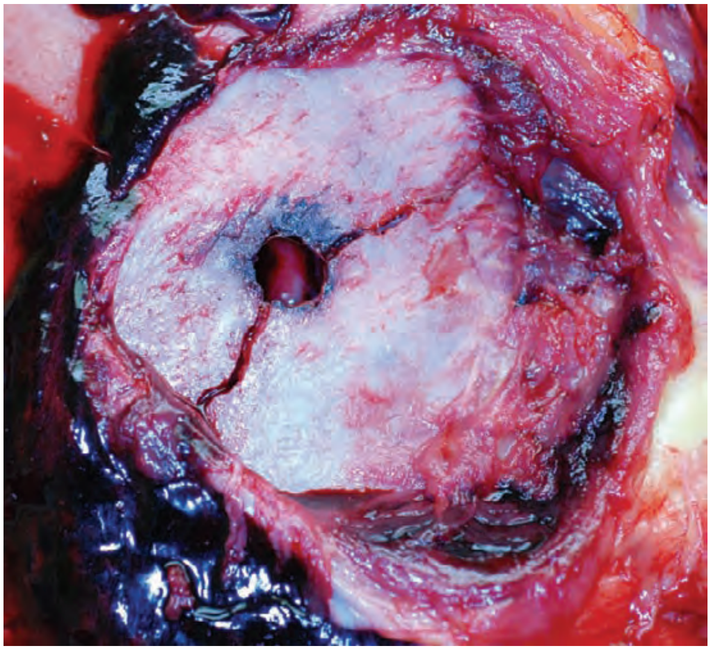

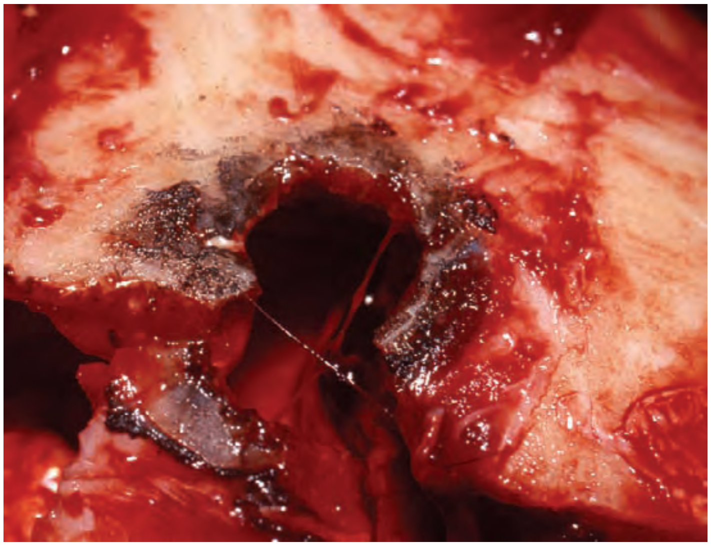

Underlying soot deposition within the wound track on the outer surface of the exposed skull. This is typically seen with a tight contact gunshot wound where soot is forced into the wound track.

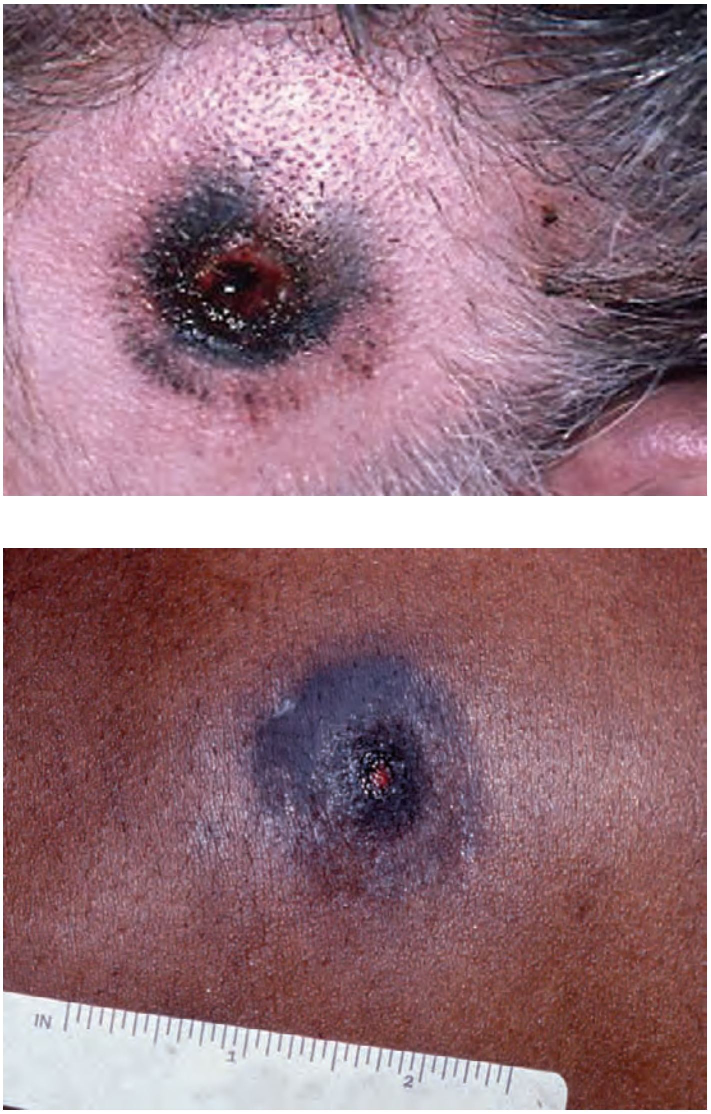

Close-range gunshot wounds with soot deposition.

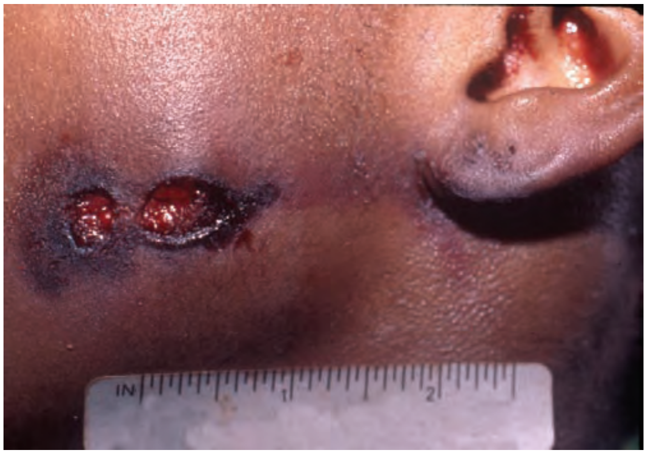

Note the entrance wound is to the left where soot is visible at the adjacent skin.

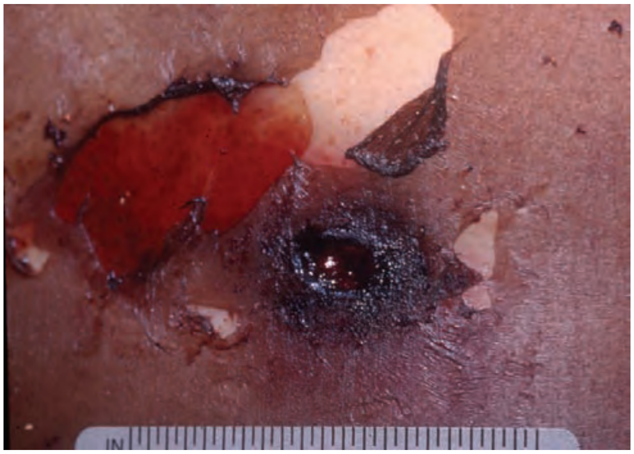

Close-range gunshot wound to a body with early decompositional changes. Note the soot deposition surrounding the perforation site. These early putrefactive changes consist of skin slippage and discoloration.

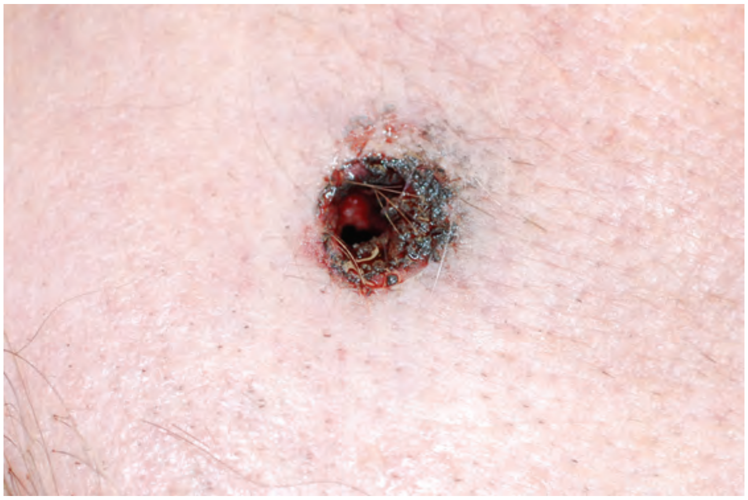

Loose contact with more soot extending to the right side of the perforation.

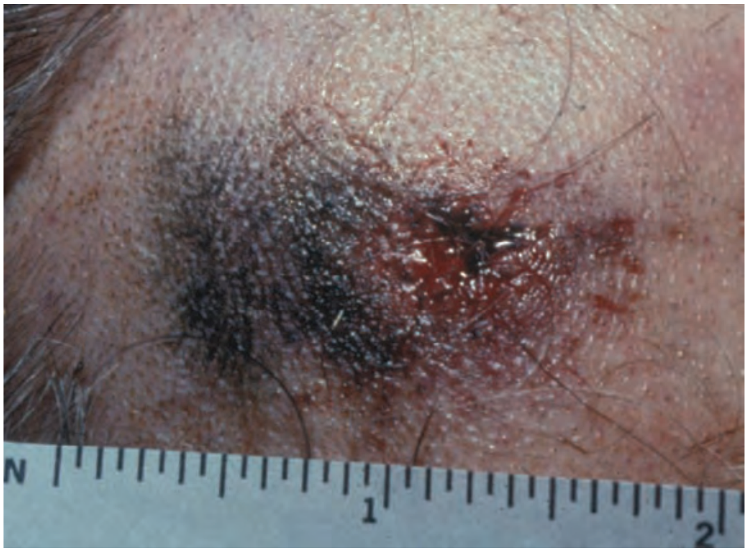

Loose contact with the muzzle more tightly applied at the right aspect. Note the greater soot deposition to the left of the perforation.

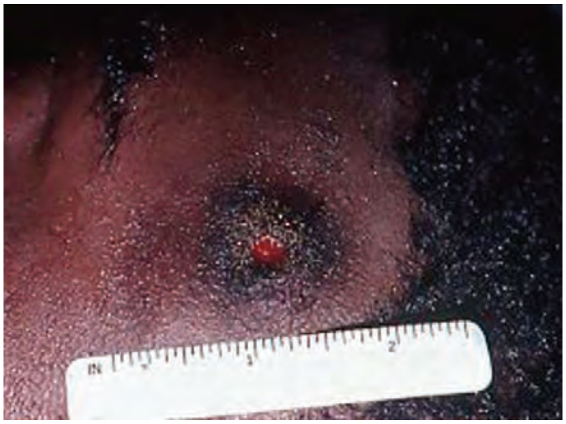

Close-range gunshot wound with soot deposition.

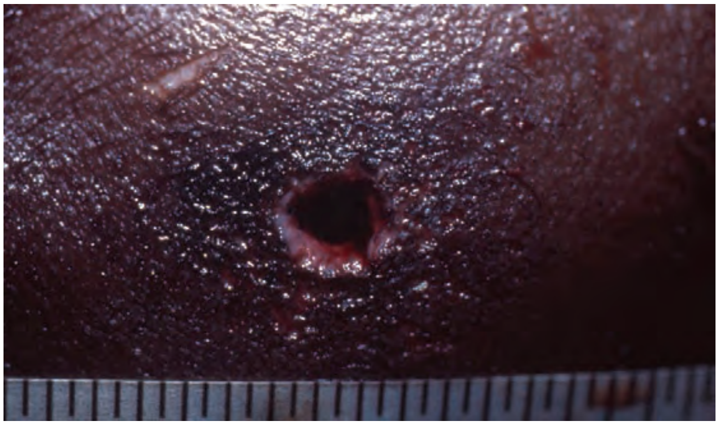

Close-range gunshot wound in a very dark-skinned individual. Note the gunpowder residue surrounding the entrance gunshot wound.

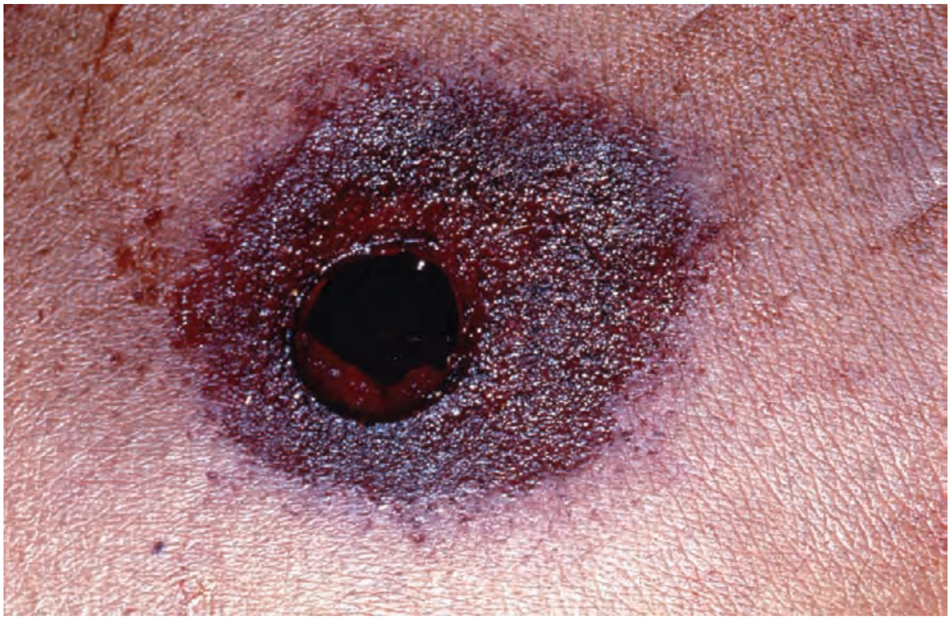

Close-range gunshot wound with a muzzle flare burn and soot encircling the perforation site.

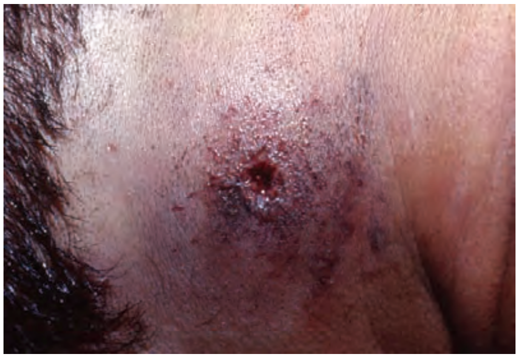

Close-range gunshot wound with sparse soot deposition surrounding the perforation site with adjacent superficial skin erosions due to muzzle flare.

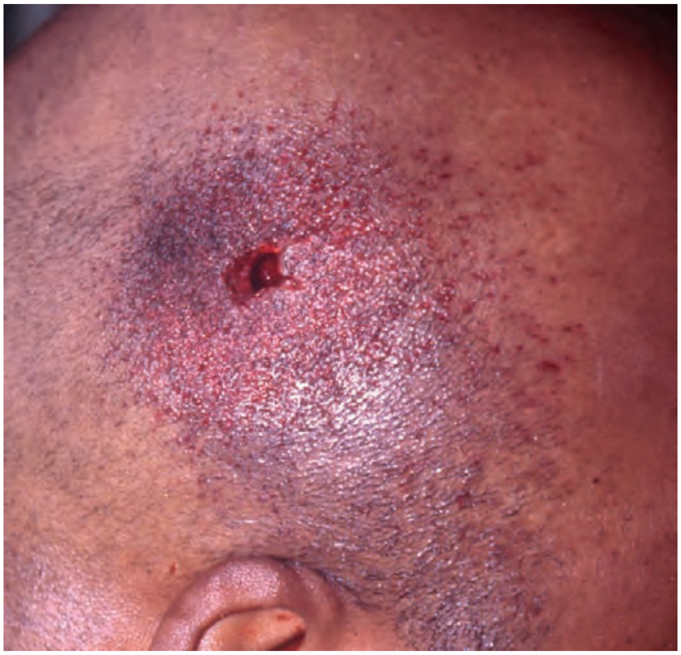

Close-range gunshot wound to the scalp with sparse soot deposition, adjacent burning of the scalp due to muzzle flare, and closely packed stippling defects.

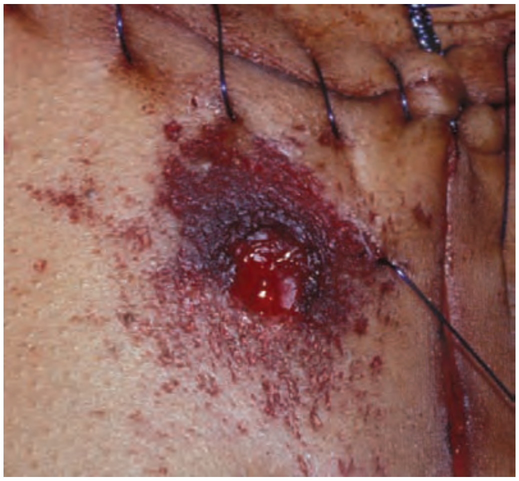

Close-range gunshot wound with soot deposition, superficial skin erosions from muzzle flare with tightly compacted stippling-type defects.

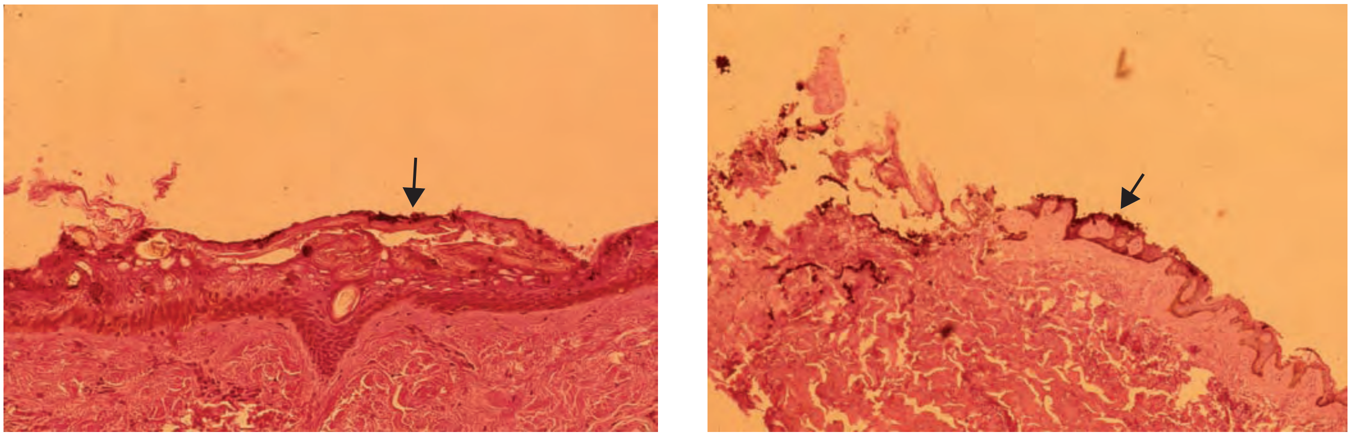

These demonstrate histopathology sections of skin with soot deposition at its surface. The soot appears as a black film on the skin's surface; unburned powder particles may also be visible.

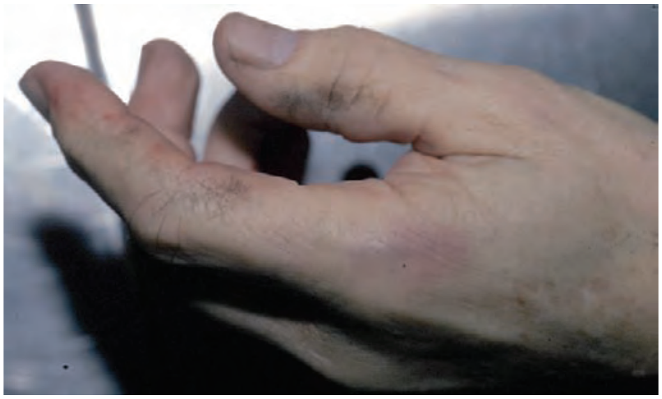

Gunpowder residue at the fingers due to firing a handgun. This individual died from a self-inflicted gunshot wound to the head. The amount of gunpowder residue following discharge of a firearm may be quite variable and sometimes not very obvious.

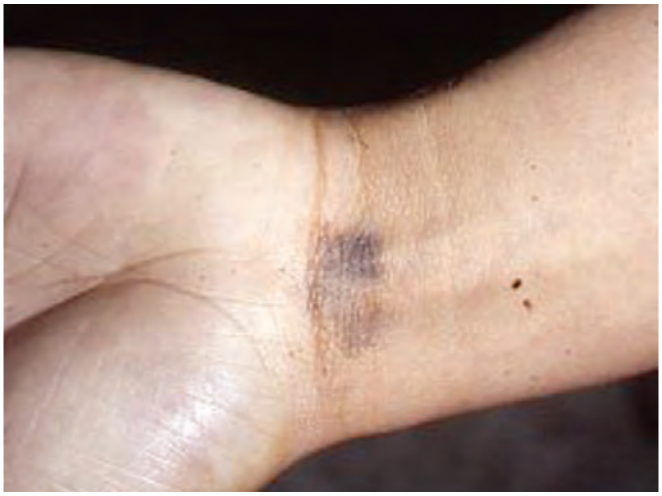

Soot may also be deposited as it exits from the muzzle of the gun or from cylinder gap in revolver type handguns.

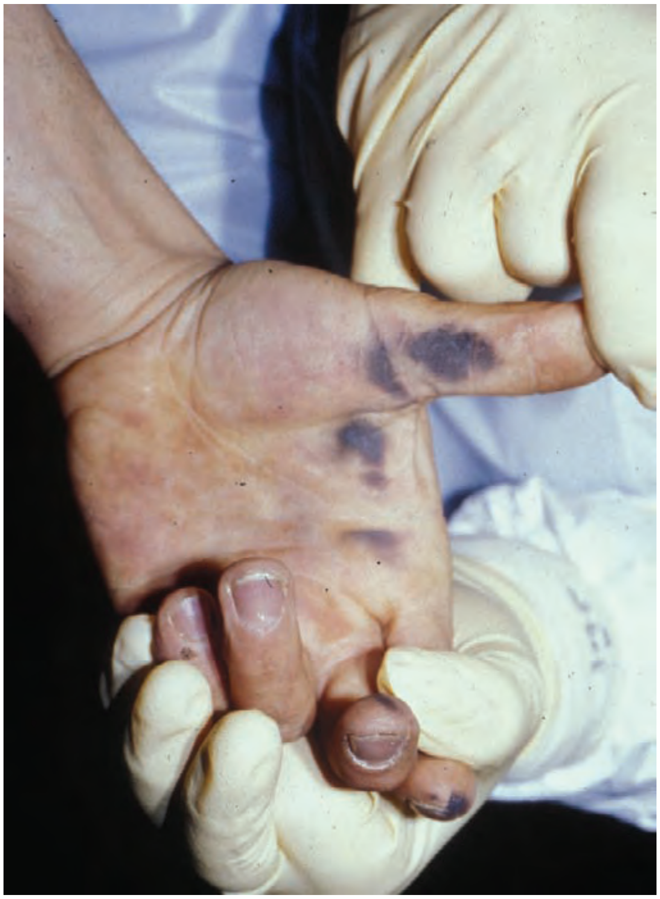

This pattern of soot was caused by holding a revolving cylinder while discharging the weapon.

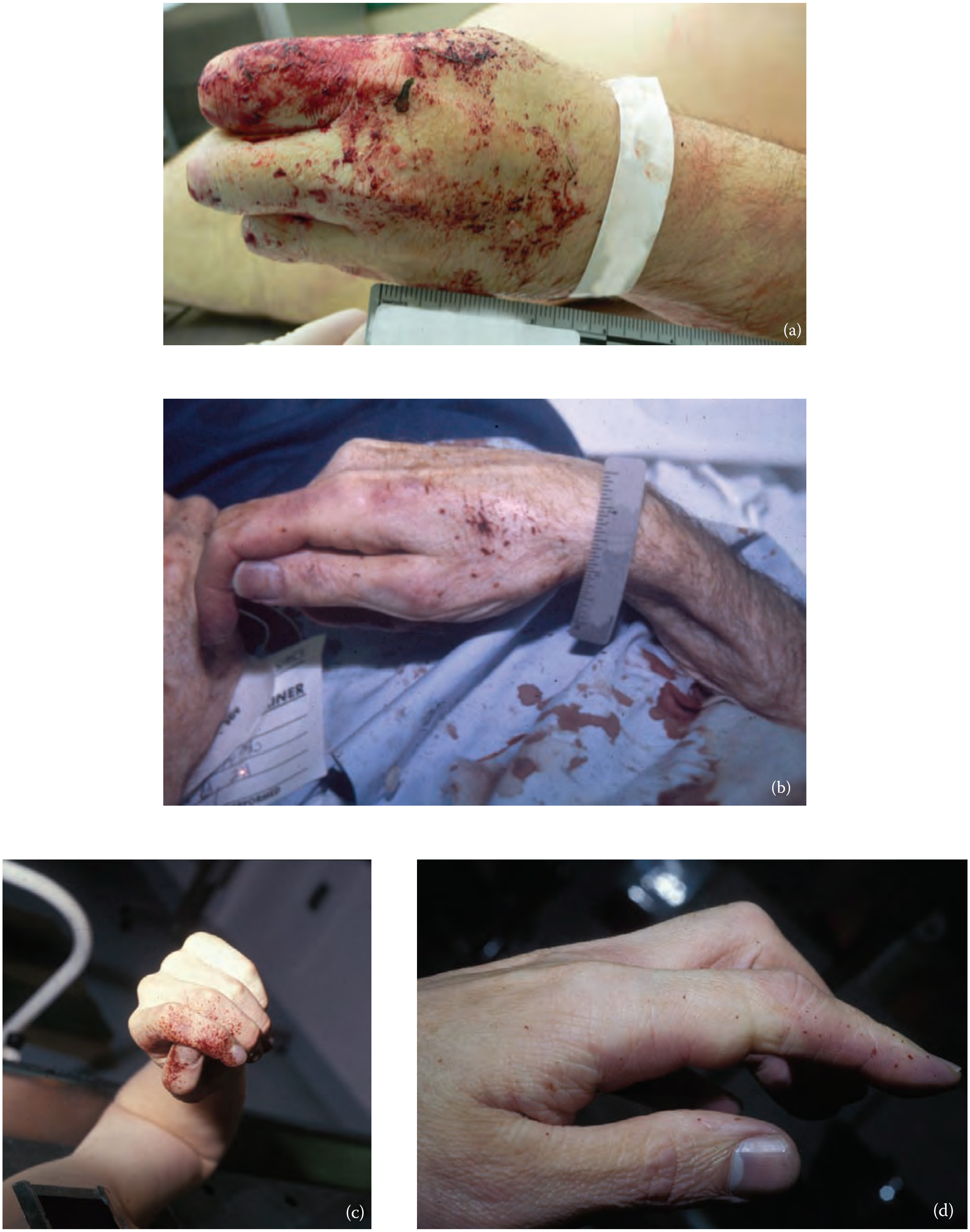

Note the "blowback" blood spatter pattern on the hands of these individuals who died from self-inflicted gunshot wounds to the head. Note that in (a) the spatter pattern is partially obscured by subsequent contact deposit of blood. Also note in (d) the extent of spatter may be minimal.

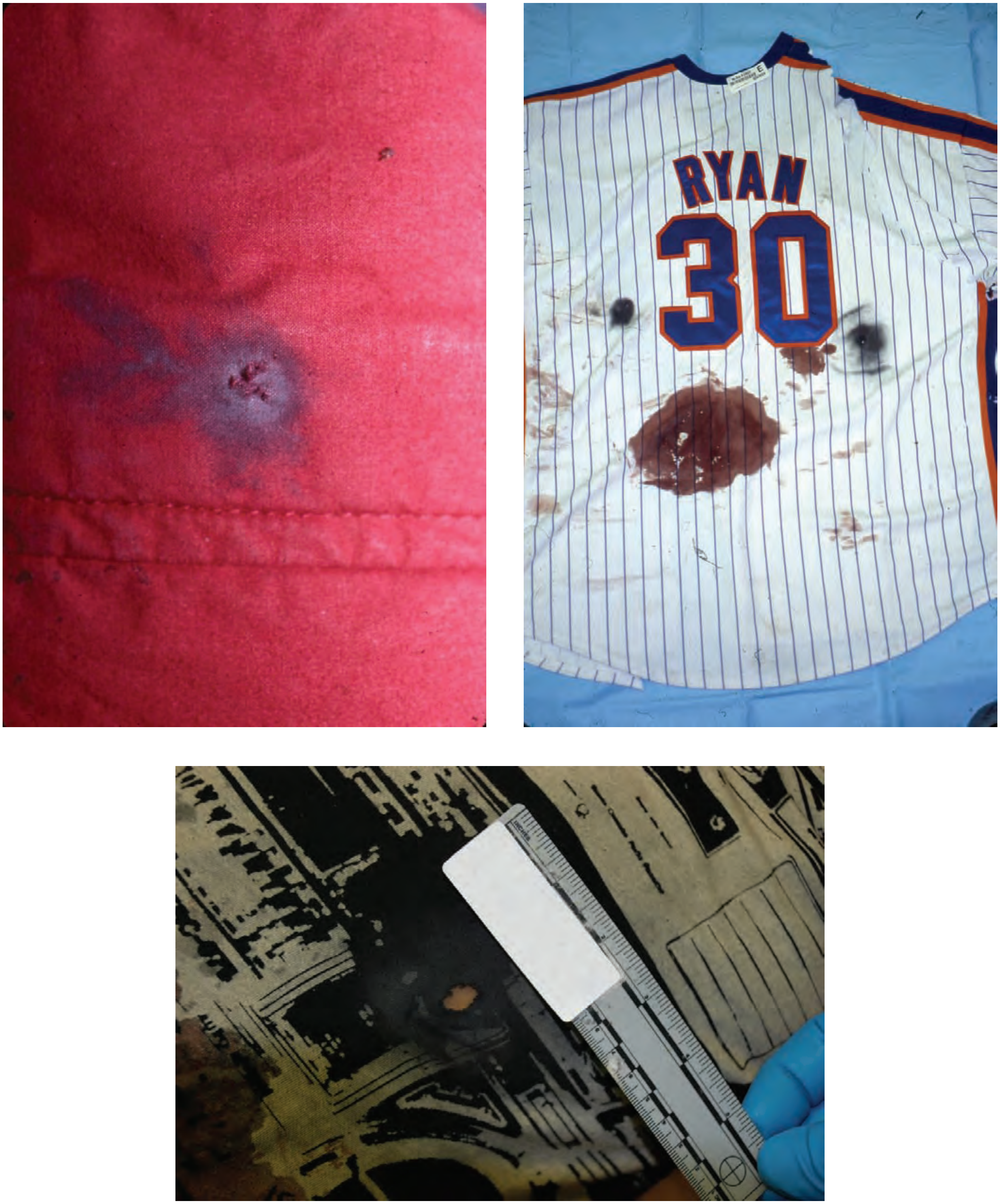

All of these are articles of clothing demonstrate close-range gunshot perforations with soot deposition. It is always important to inspect the decedent's clothing. Multiple layers of heavy clothing may filter gunpowder residue from the body surface. It is also important to distinguish bullet wipe from soot deposition. Bullet wipe is a small, encircling, gray discoloration around the perforation site of the clothing due to lubricants and residue from within the barrel of the gun that adhere to the bullet surface as it passes through the barrel. Unlike soot, bullet wipe is not useful in range of fire estimations.

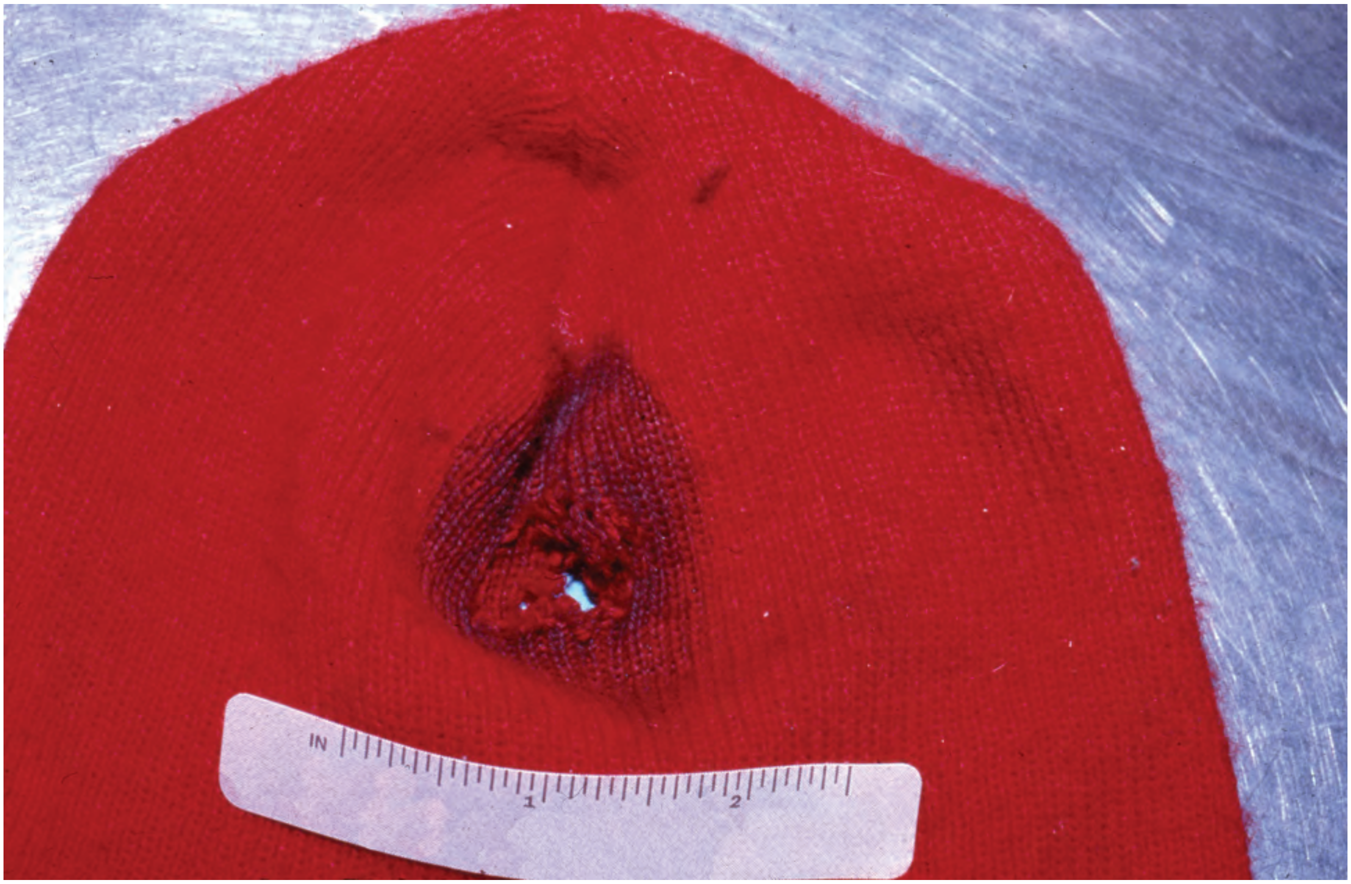

Clothing with soot indicating close range of fire. The individual wearing the hat had long curly black hair and there was no appreciable soot noted at the scalp.

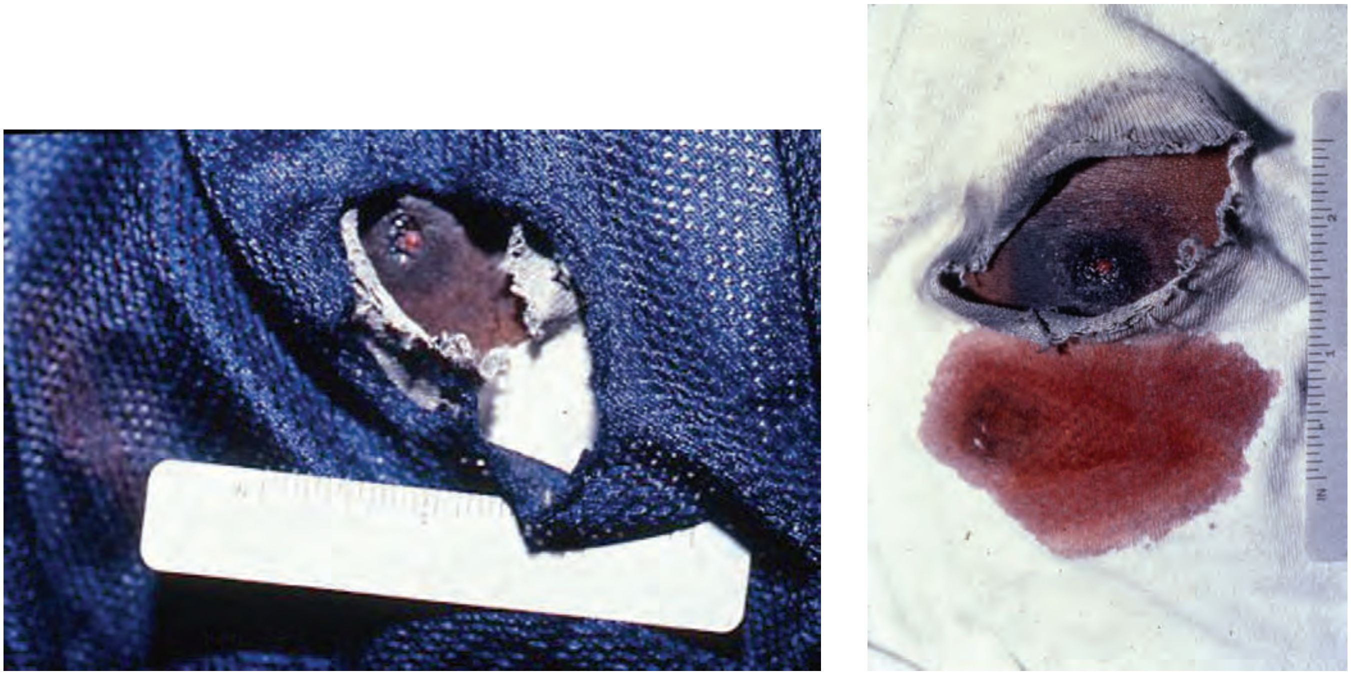

Note the large tear in the clothing from this contact wound and the minimal amount of adjacent external hemorrhage. This is an antemortem injury with more than a liter of blood observed in the thoracic cavity during autopsy.

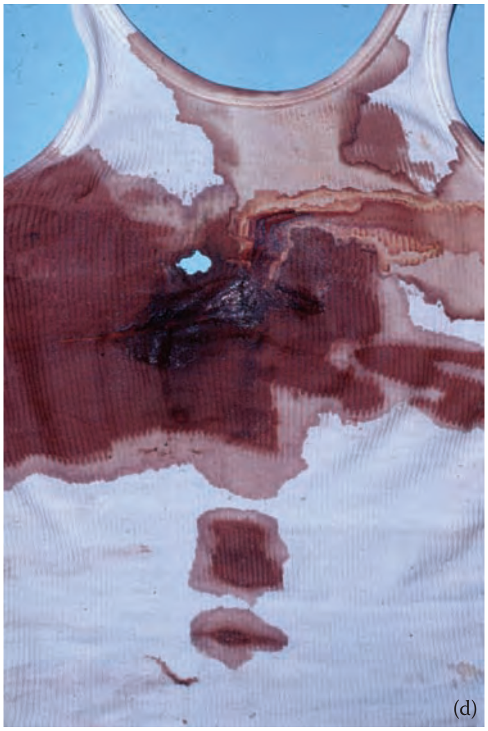

This entrance gunshot defect is surrounded by dark clotted blood. The dark discoloration of blood, particularly on dark clothing, may make if difficult to observe soot. This is a distant gunshot wound (d).

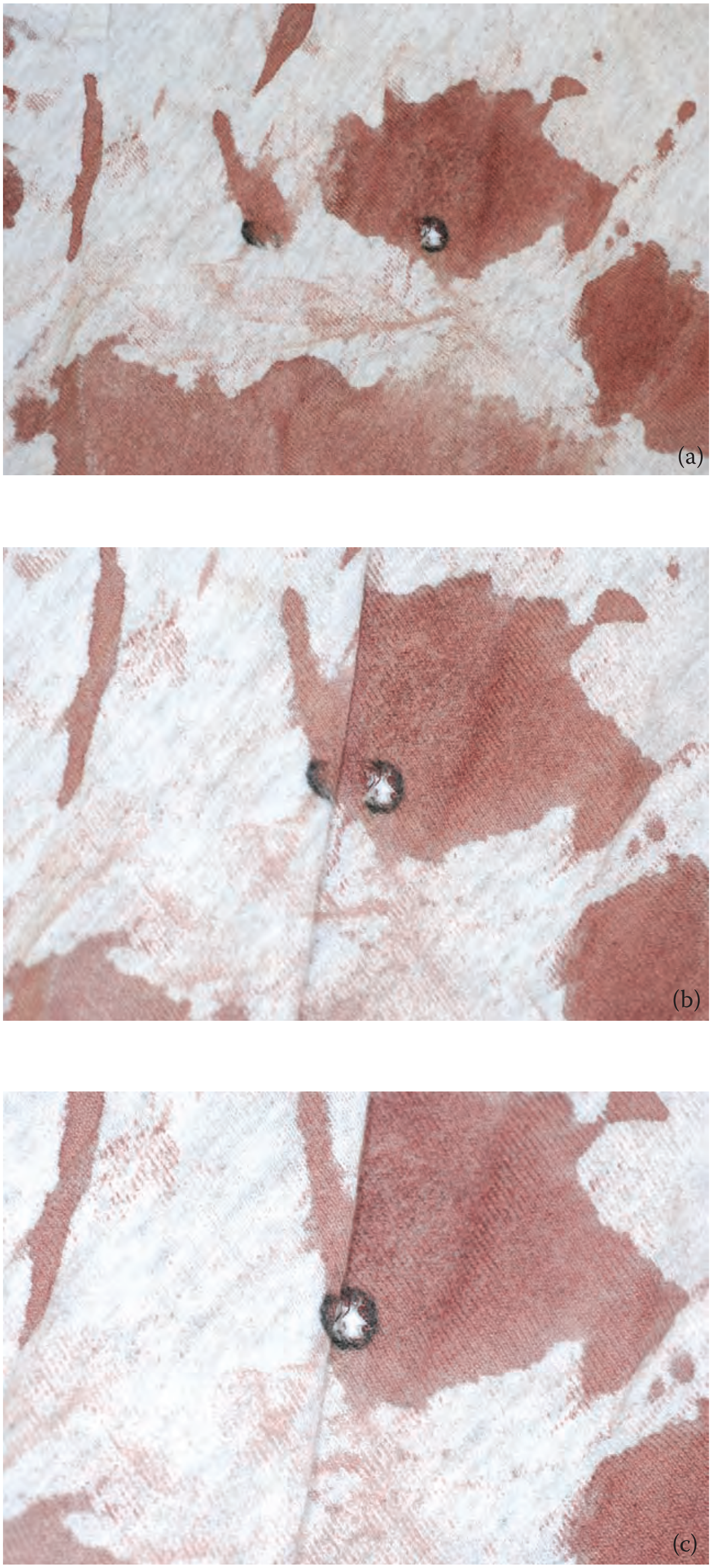

Bullet wipe. This individual had his shirt tucked in, creating folds in the cloth. Note the dark discoloration around the perforation site with the right aspect separated several inches from the site of bullet entry. By refolding the fabric and approximating the margins we can simulate how the individual actually wore his shirt (a–c).

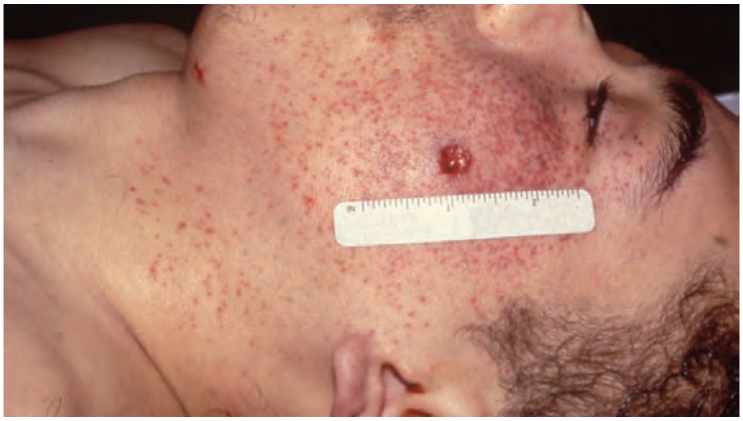

Intermediate range of fire. Tightly packed stippling defects surrounding this entrance gunshot wound to the cheek. There is no apparent soot deposition.

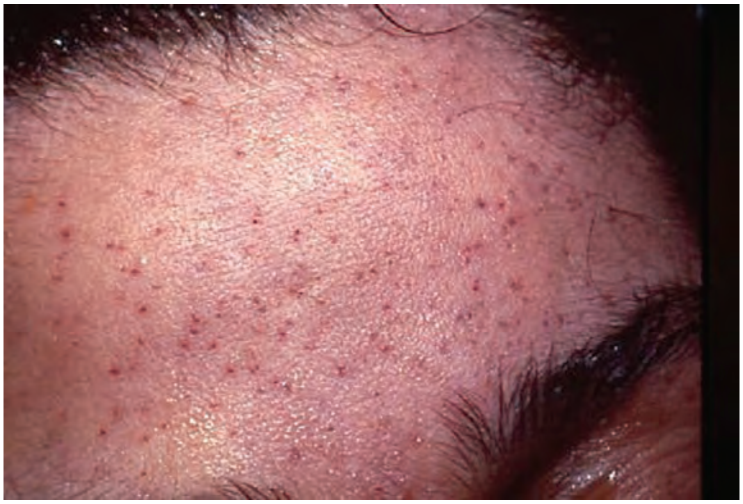

Multiple stippling defects to this individual's forehead indicating intermediate range of fire.

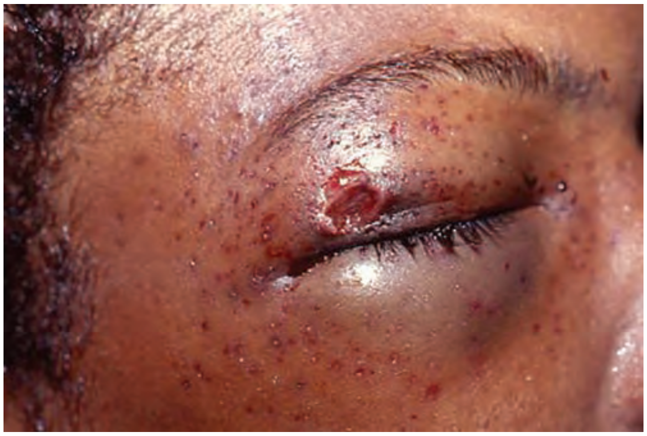

Intermediate range of fire with stippling defects. Stippling defects represent abrasions that, unlike soot, cannot be wiped away.

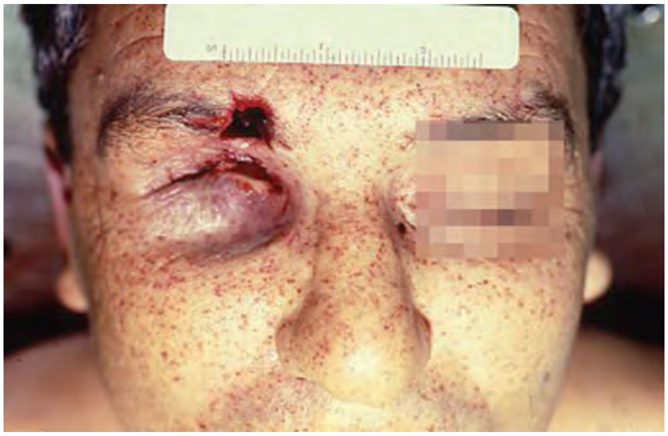

Intermediate range entrance gunshot wound to the eyebrow with stippling defects across the face. Note the irregular nature of the wound due to the location of the gunshot wound through the eyebrow ridge, with the underlying frontal bone closely subadjacent to the skin surface.

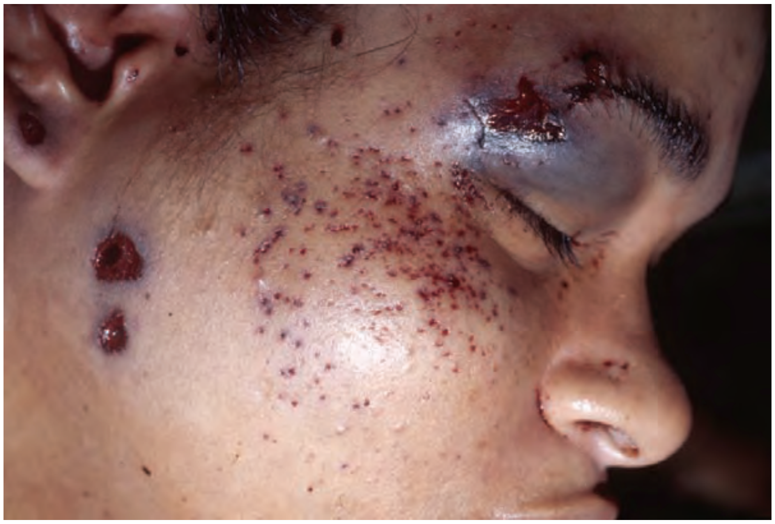

Intermediate range of fire.

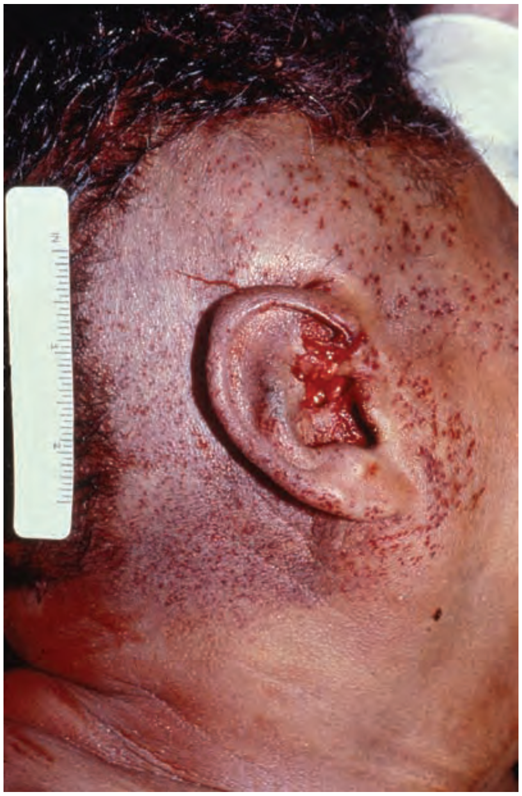

Intermediate-range entrance gunshot wound with stippling defects. The scalp is partially shaved. Note the decreased number of defects in the shaved area as a result of hair dampening the effect. The weapon was a sawed-off 30/30 hunting rifle that was reportedly discharged approximately 15 inches away from the decedent's head.

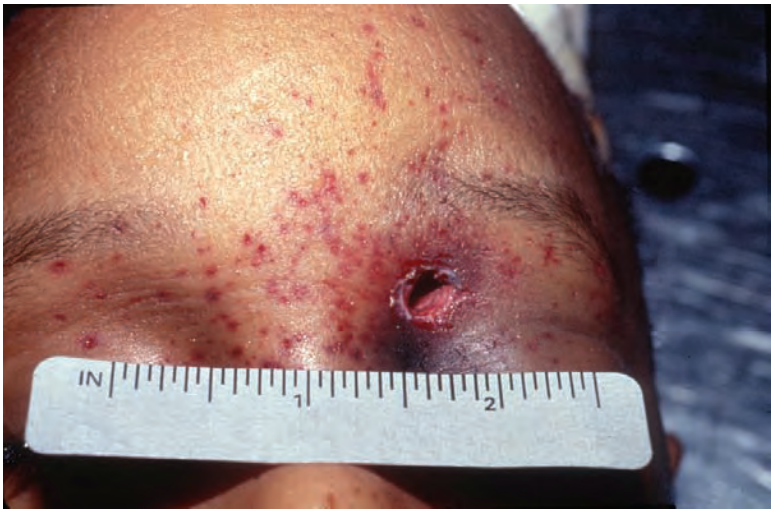

Entrance gunshot wound with stippling defects indicating intermediate range of fire.

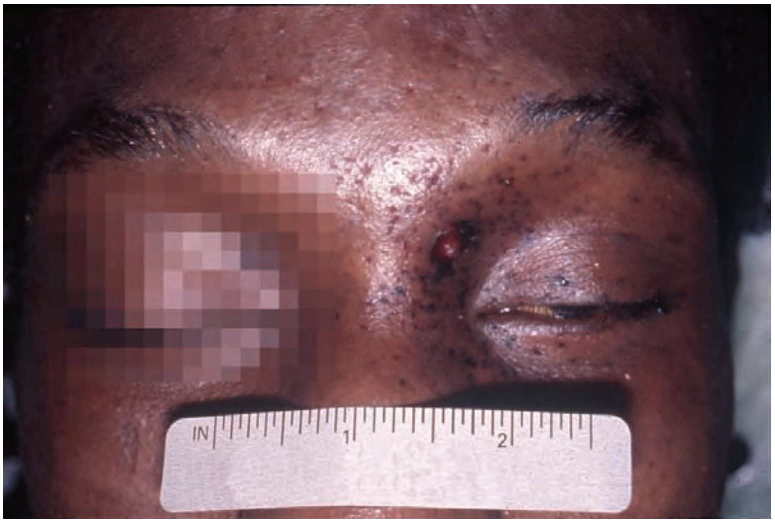

Intermediate-range gunshot wound with stippling in a dark-skinned individual.

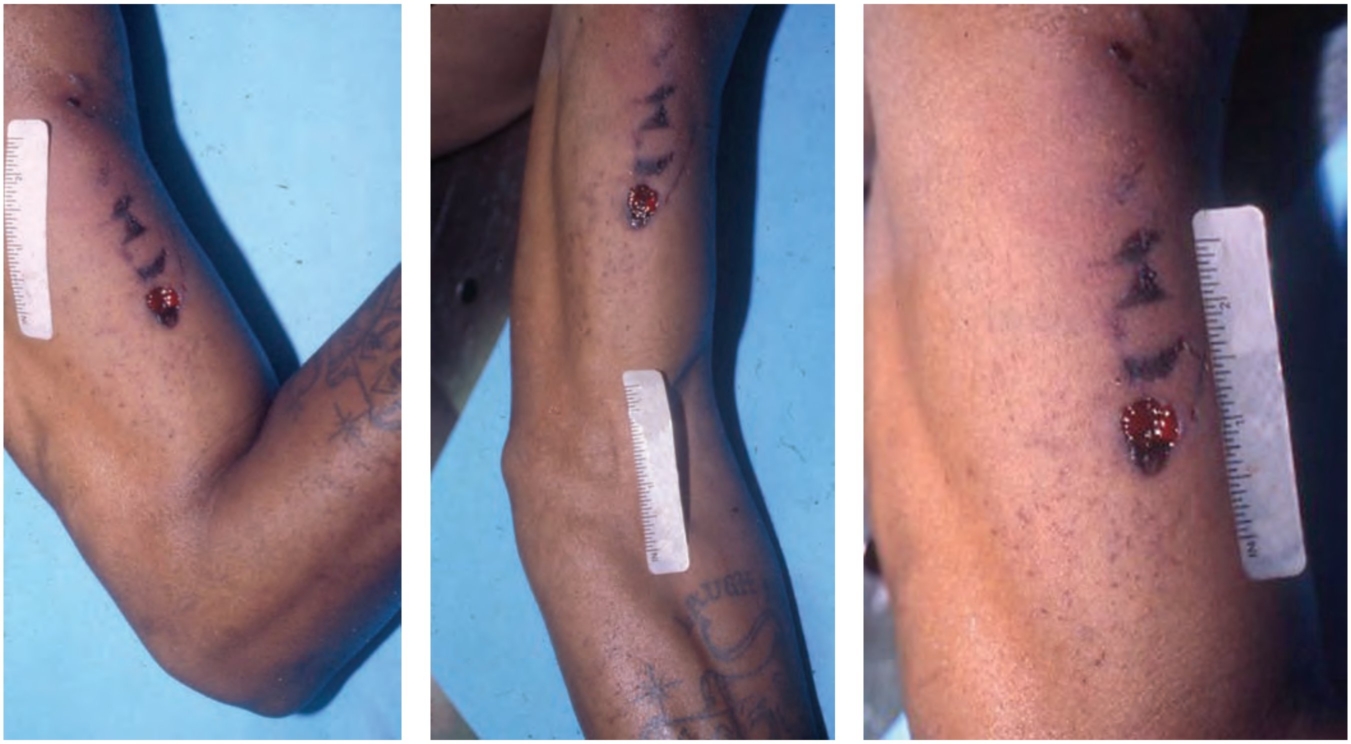

Intermediate-range gunshot wound. Note the sparse stippling defects along the bicep and forearm with sparing of the anticubital fossa region. This indicates the decedent had his arm bent when he was shot.

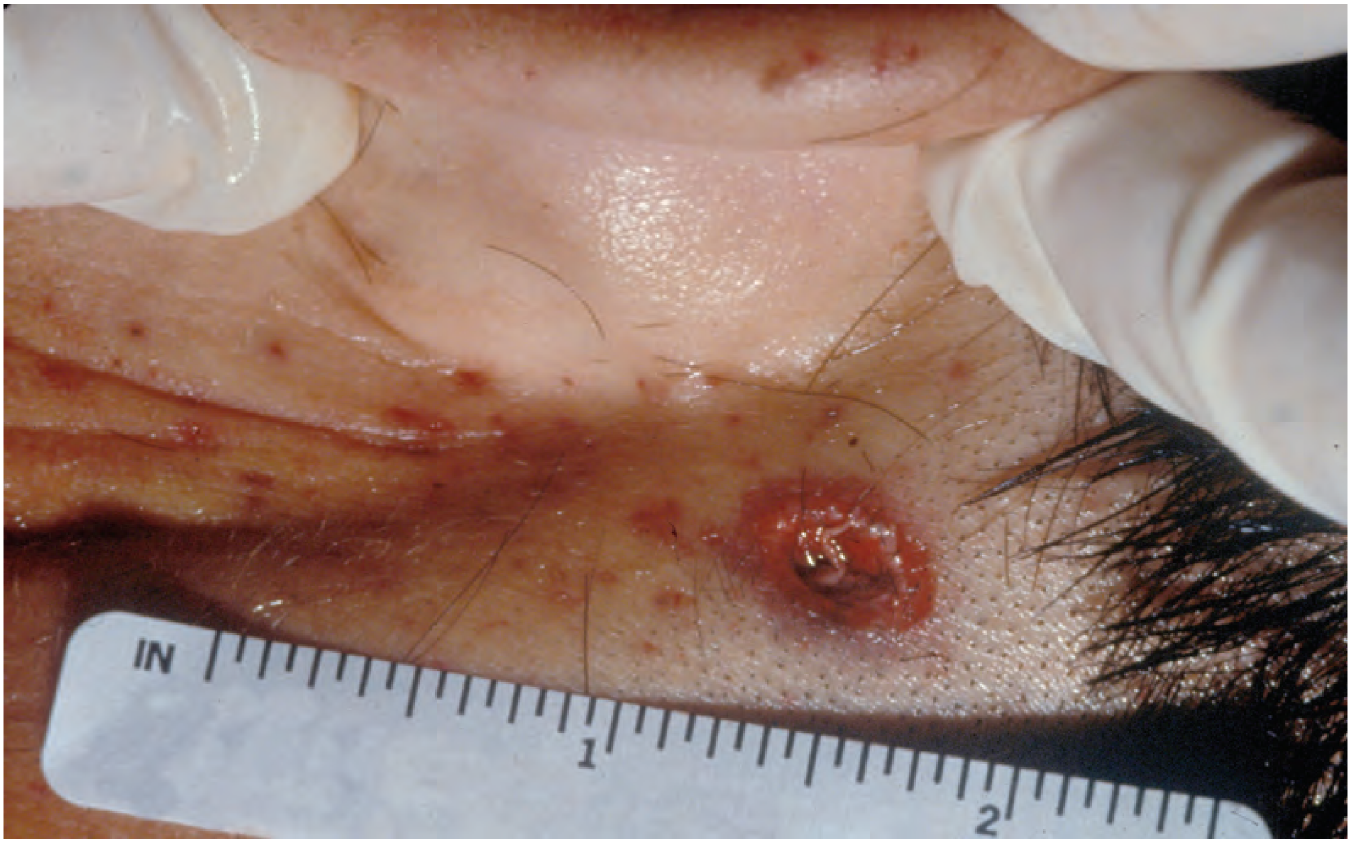

This individual had stippling defects involving the posterior portion of his lateral outer ear and temporal scalp. Note the sparing of the posterior medial ear.

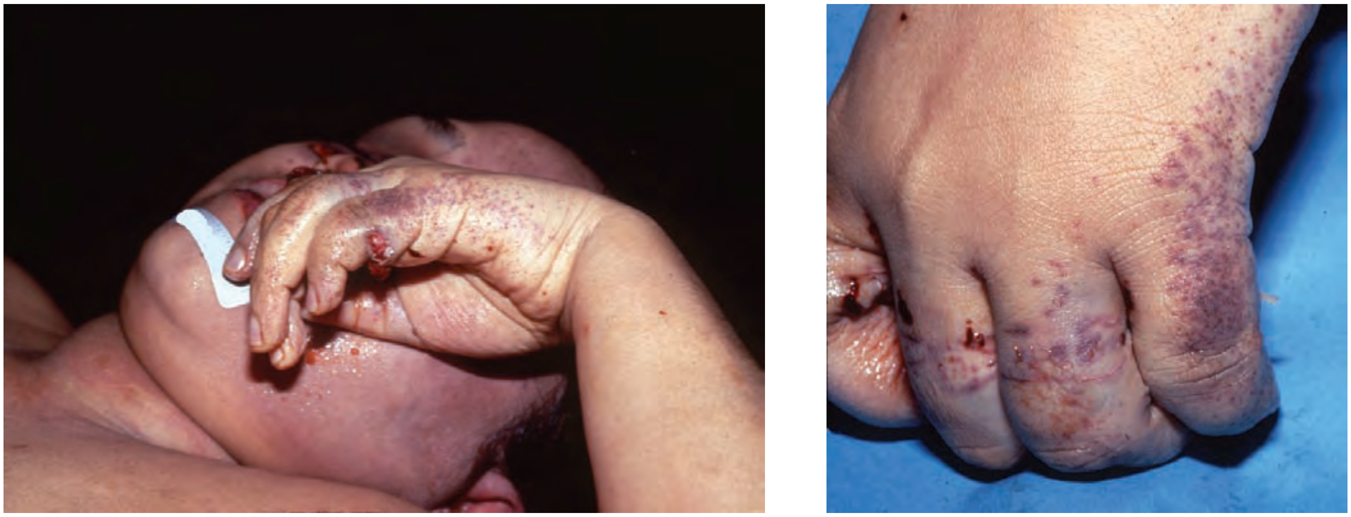

Stippling defects with soot deposition on the side of an outstretched hand, indicating intermediate to close range of fire.

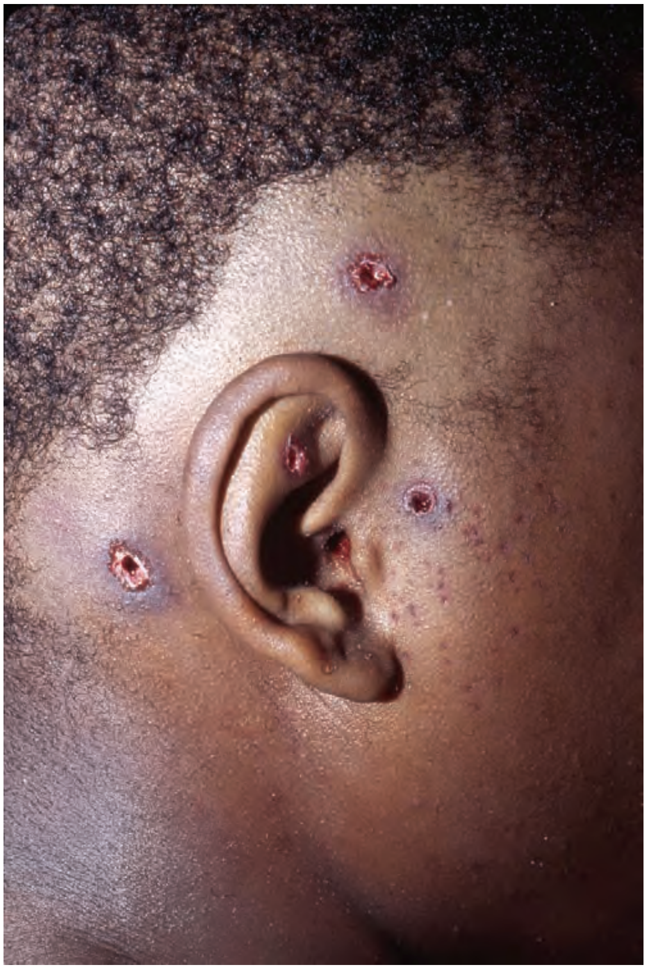

Multiple-entrance gunshot wounds. Note the sparse stippling defects indicating intermediate range of fire at the skin anterior to the ear. The farther away a gun is discharged from the body, the more spread out the stippling defects become.

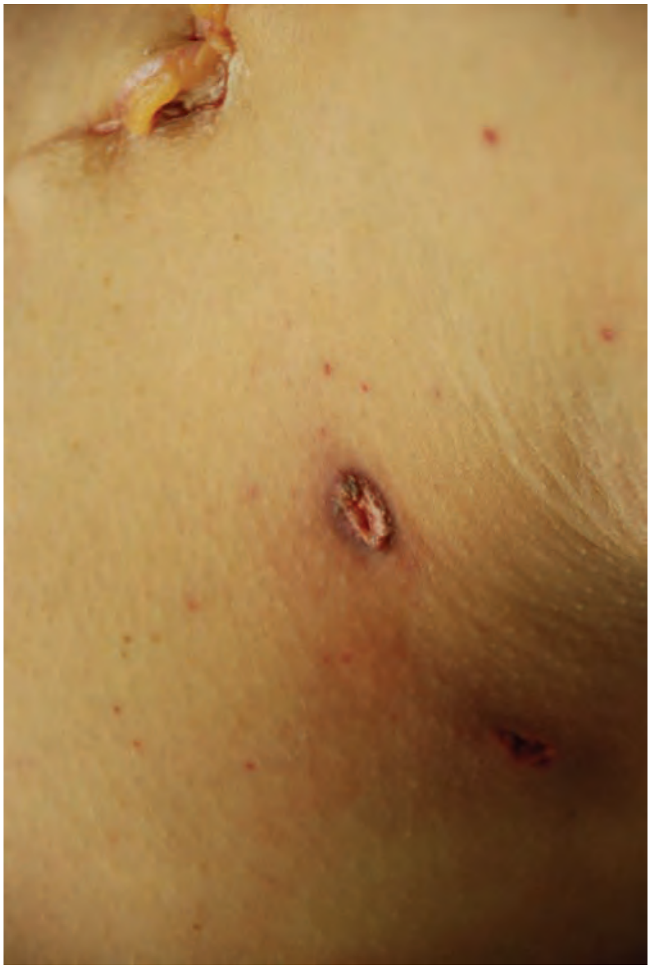

Intermediate range of fire. Note the sparse stippling defect surrounding the bullet perforation site.

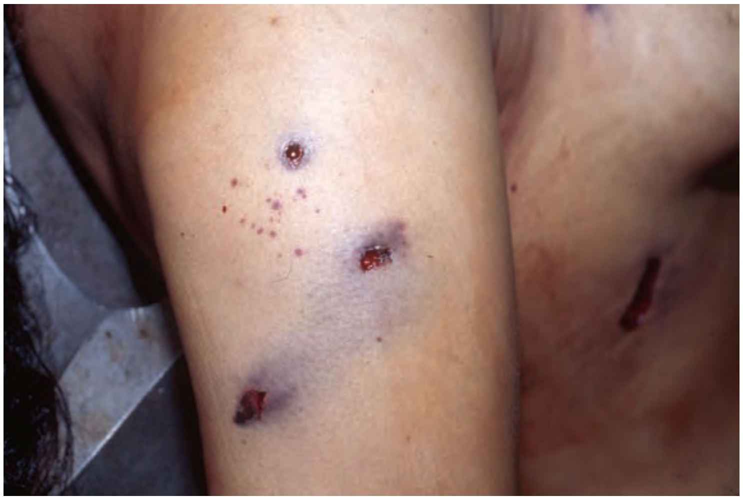

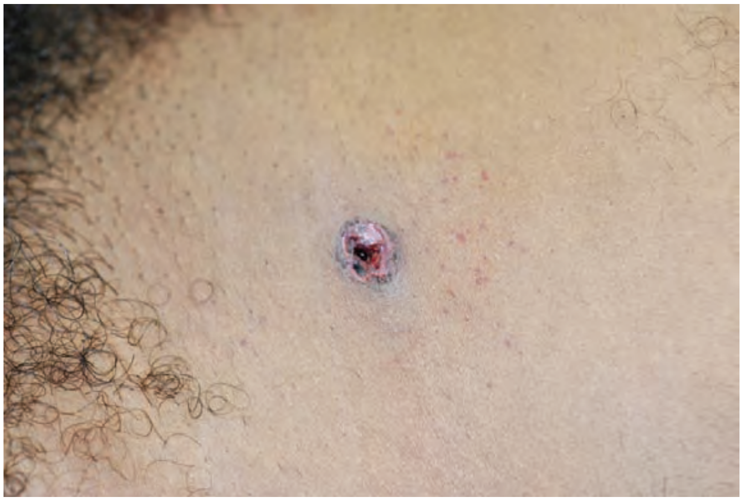

A few sparse stippling defects indicating intermediate range of fire. Note the contusion surrounding the perforating gunshot wound at the arm. This dark discoloration should not be confused with gunpowder residue.

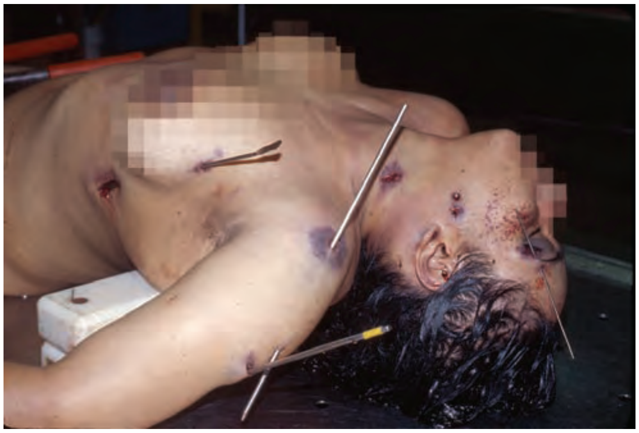

Note the multiple gunshot wounds to this individual in varying directions. She was shot multiple times while jumping in different directions in an attempt to run away. Note the contusion to her left shoulder and the stippling defects to her face. It is important to realize that the direction of fire may vary greatly.

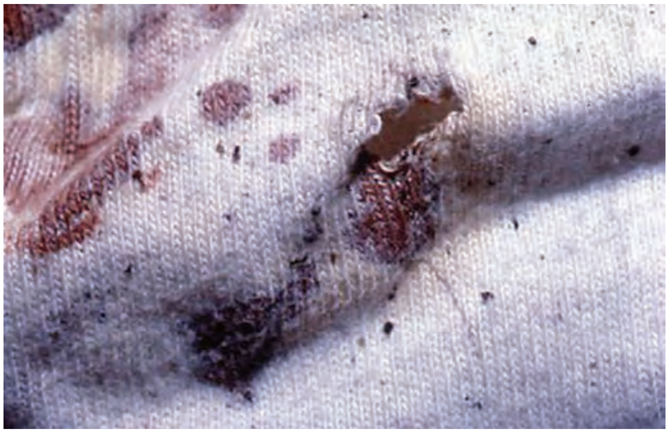

Gunshot wound perforation of clothing with multiple gunpowder residue flecks at the surrounding surface. The underlying entrance gunshot wound still revealed a sparse stippling pattern consistent with intermediate range of fire.

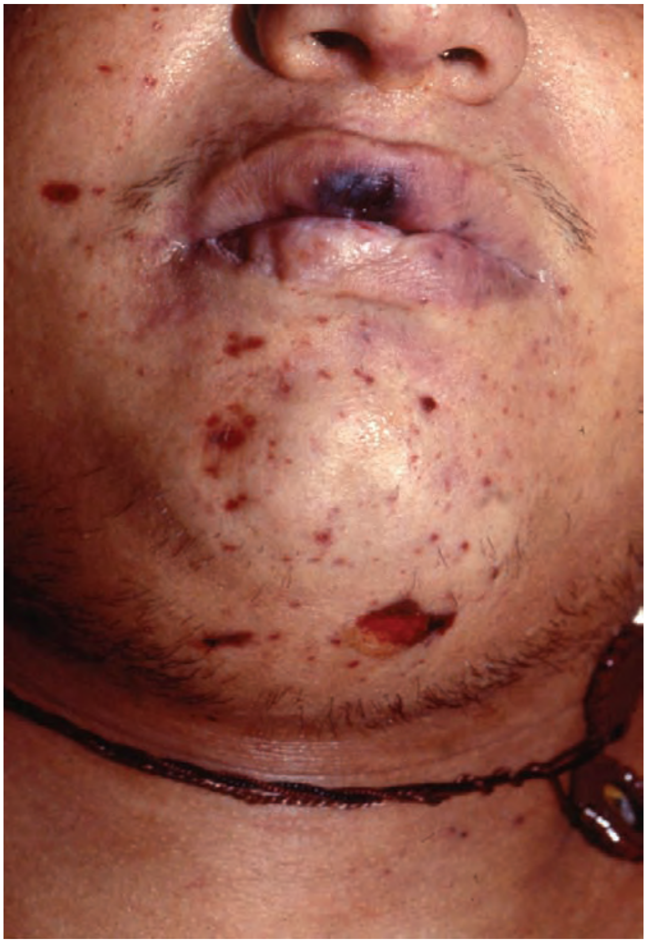

Distant gunshot wound with pseudostippling to this individual's face caused by fragmented debris as the bullet passed through a car window.

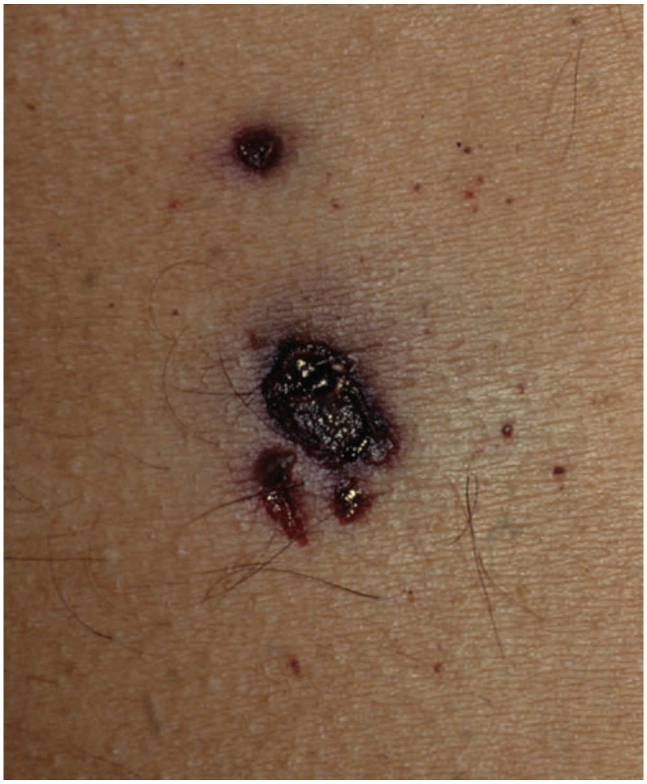

Distant gunshot wound with pseudostippling. Note the large size variation in the abrasions to the gunshot wound. True stippling defects are more uniform in size.

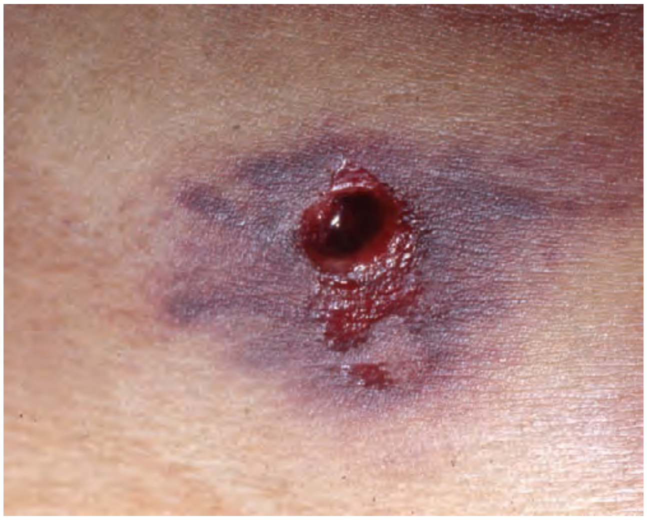

Atypical distant-entrance gunshot wound with surrounding contusion.

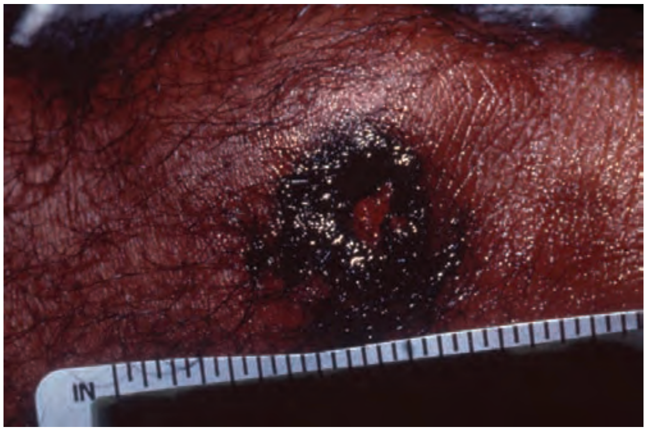

Distant-entrance gunshot wound with surrounding clotted blood. In this case, histopathology of the adjacent skin demonstrated no dark particles consistent with gunpowder residue. It may be challenging to visualize soot when it is mixed with dried blood. Dried clotted blood has a shinier character than gunpowder residue, which has a duller appearance. It may be challenging to interpret range of fire when both dried blood and soot are present simultaneously. If there is any question concerning the presence of soot, we recommend microscopic examination to help clarify this point.

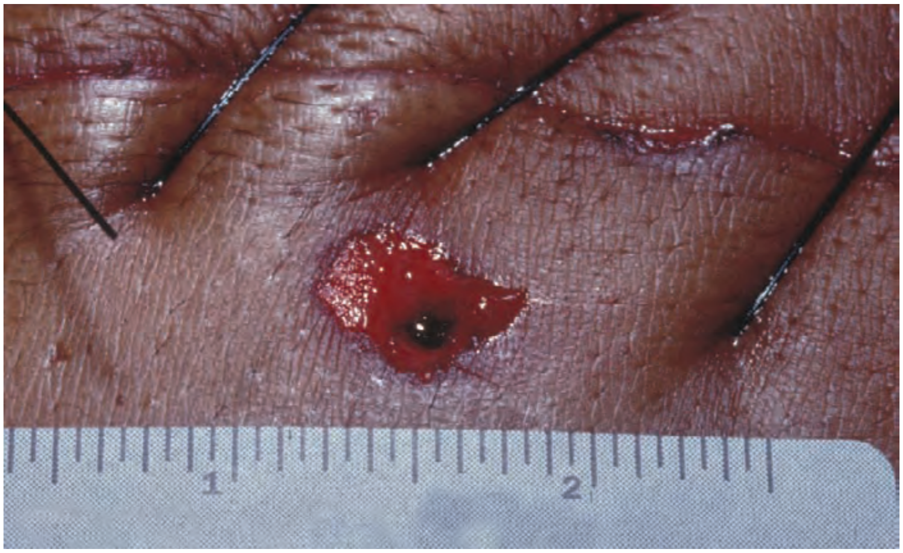

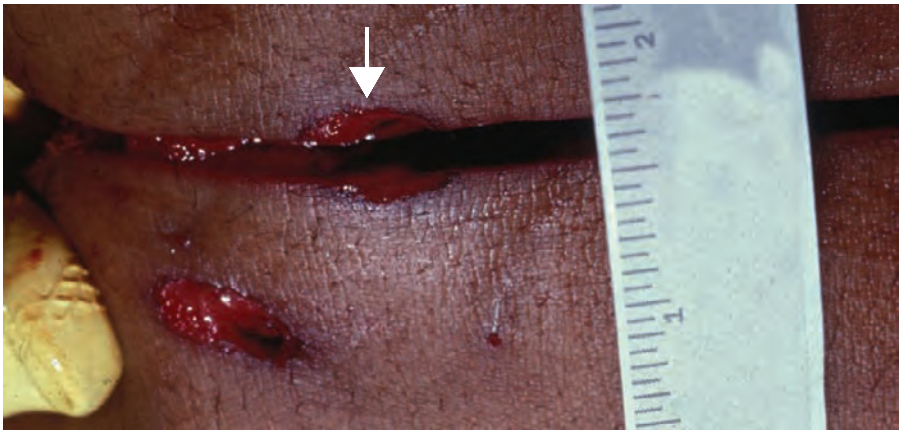

Thoracotomy incision through entrance gunshot wound. One should never cut through the wound during resuscitative efforts. No one would argue the importance of saving the individual's life. Cutting next to the wound will not affect the chances for survival. Cutting through the wound may make wound interpretation very difficult and hinder criminal proceedings. Note one gunshot wound was cut through and hidden in the suture line. The defect becomes more obvious with suture removal.

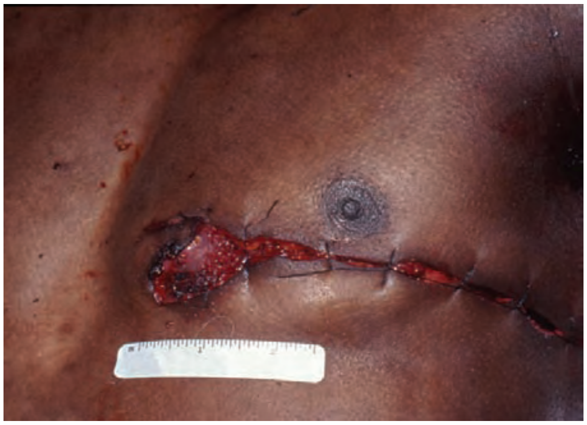

Note the large gunshot wound at the medial aspect of this thoracotomy incision.

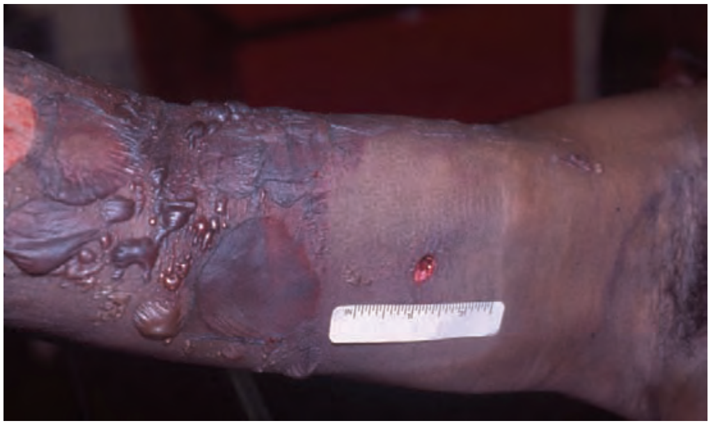

This is a gunshot wound that is several days old. This person died of septic complications. Note the adjacent wound infection distal to the GSW site.

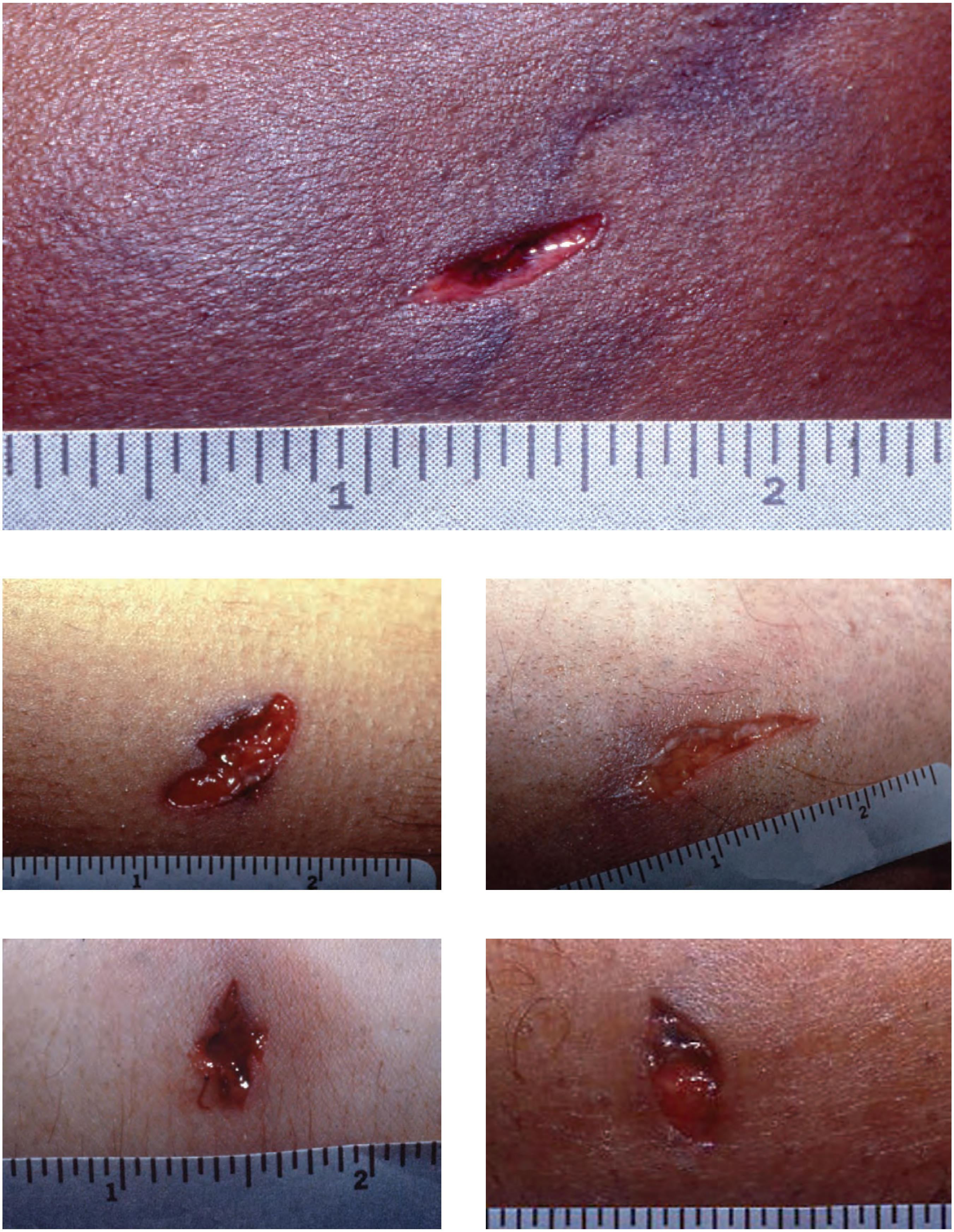

Exit gunshot wounds. Note the typical, often irregular, slit-like defects without margins of abrasion. The appearance is generally more irregular and less round than the entrance wound.

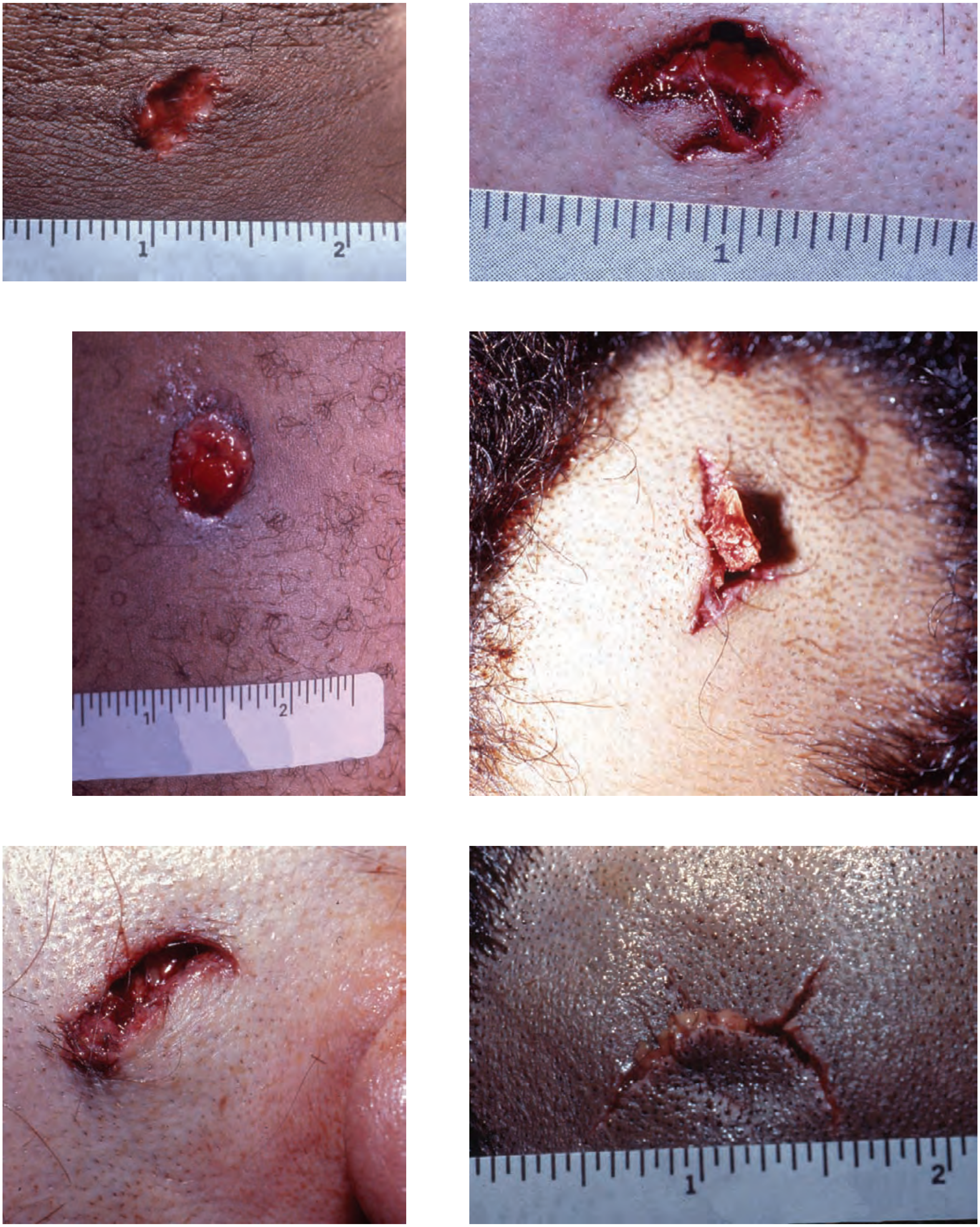

Exit gunshot wounds.

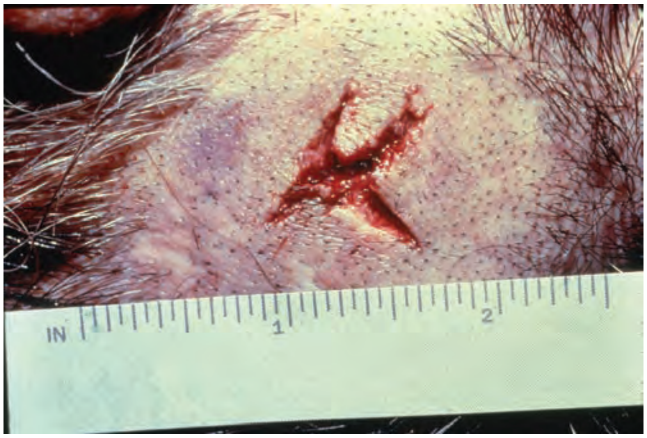

Exit wound with lacerated margins.

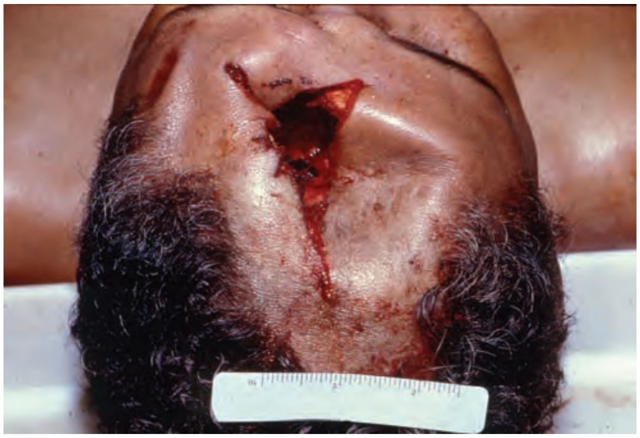

Exit wound from a sawed-off 30/30 hunting rifle.

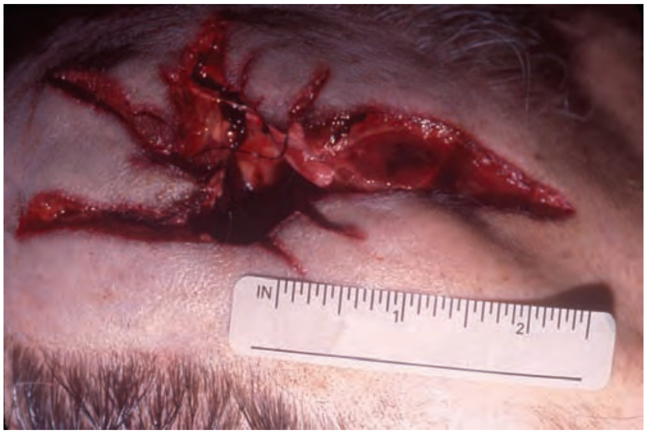

Exit wound from a handgun. It is not typical to get such a large laceration from a medium-caliber bullet that already perforated the skull.

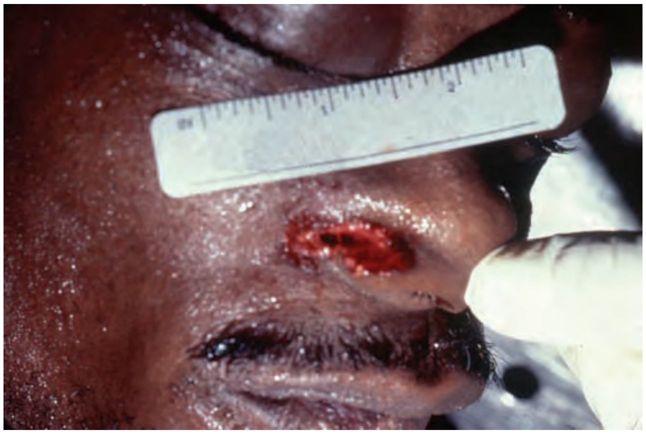

Exit gunshot wound through the right nares.

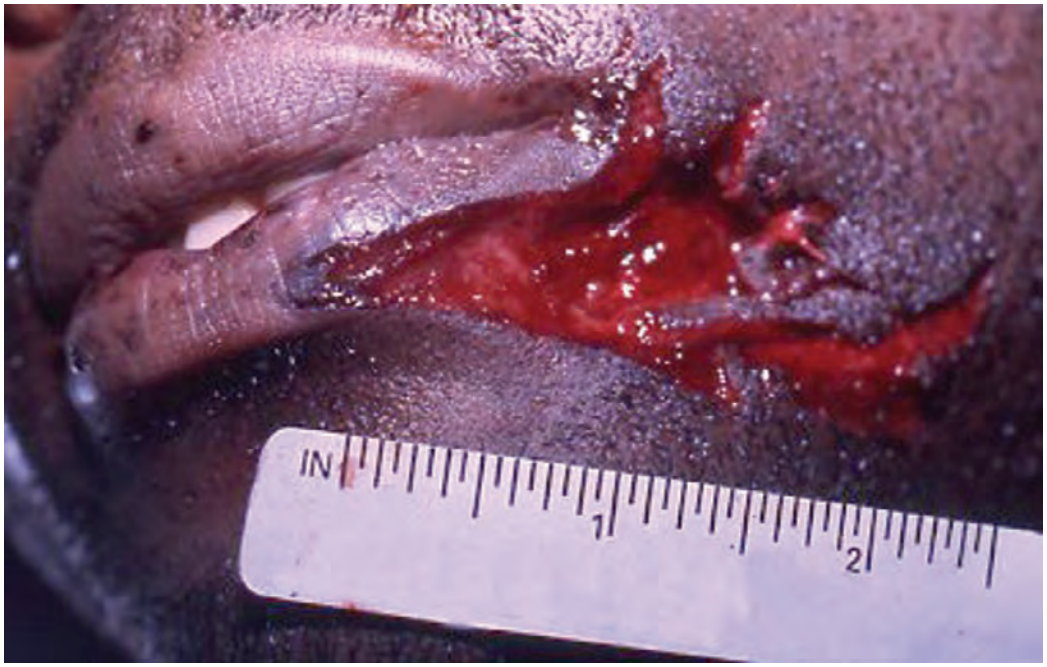

Exit gunshot wound to the lower face caused by a bullet that was markedly deformed by striking underlying bone before exiting.

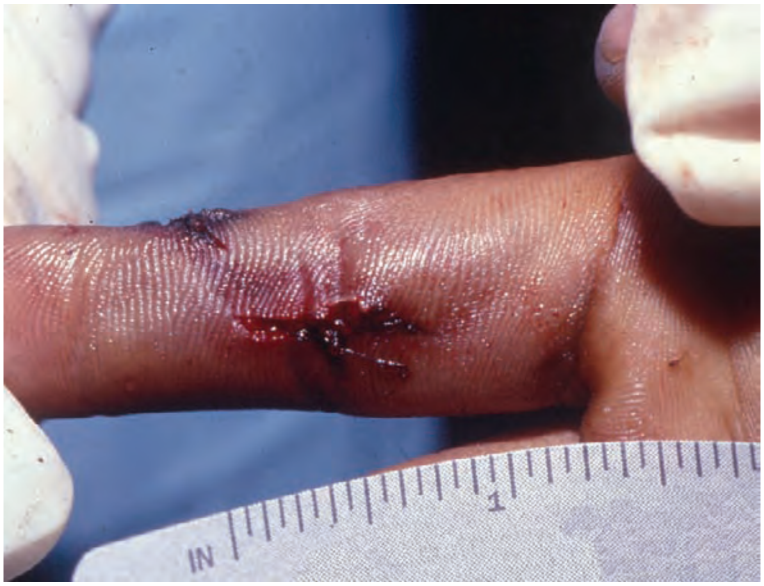

Exit gunshot wound through the finger.

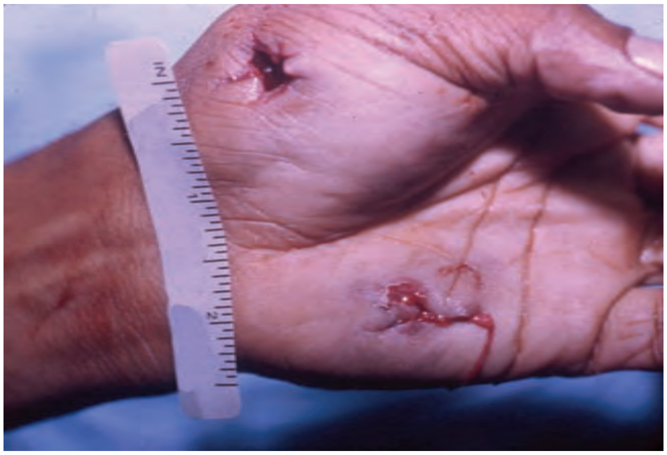

Exit gunshot wounds through the palm of the hand.

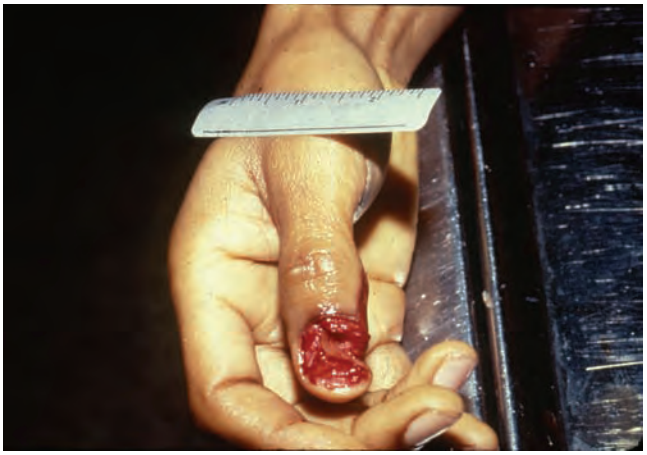

Exit gunshot wound through the thumbnail bed.

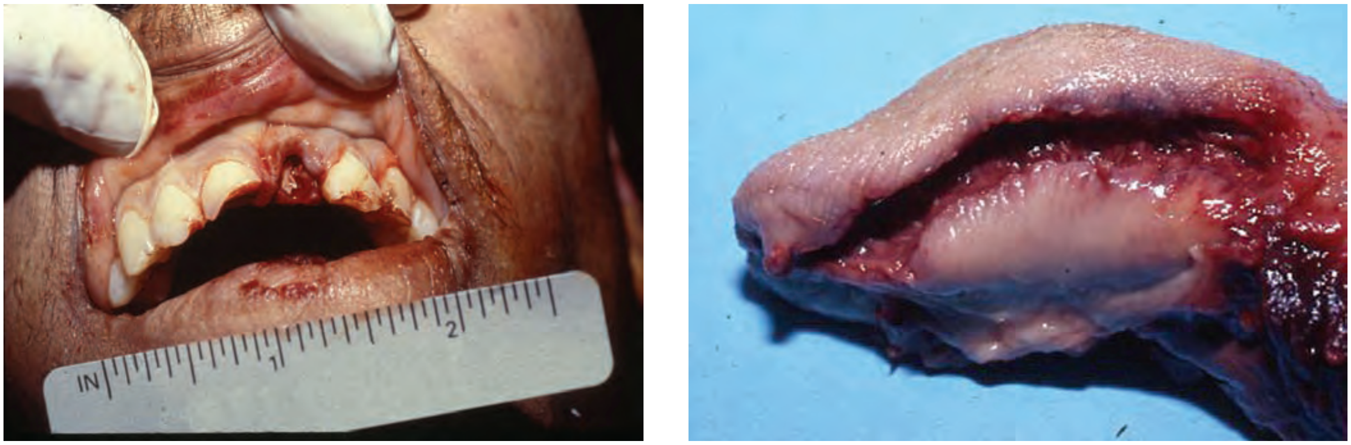

This individual was shot in the lower back while leaning forward. The bullet perforated the structures of the thoracic cavity and neck, producing a grazing-type gunshot wound to the tongue. The bullet then knocked out the individual's front tooth before exiting the body from the open mouth.

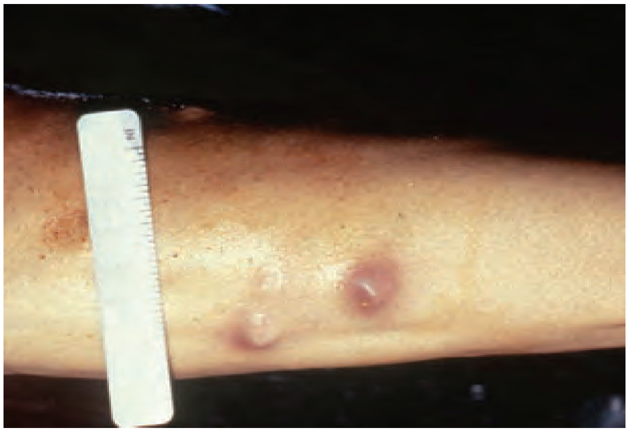

Note the lumps directly underneath the skin's surface. These are caused by bullets just beneath the surface of the skin that did not have enough kinetic energy to exit the body.

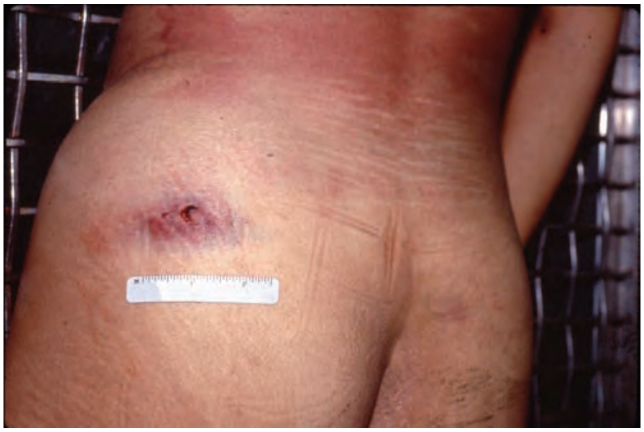

A bullet was retrieved just beneath the surface of the skin within this area of contusion.

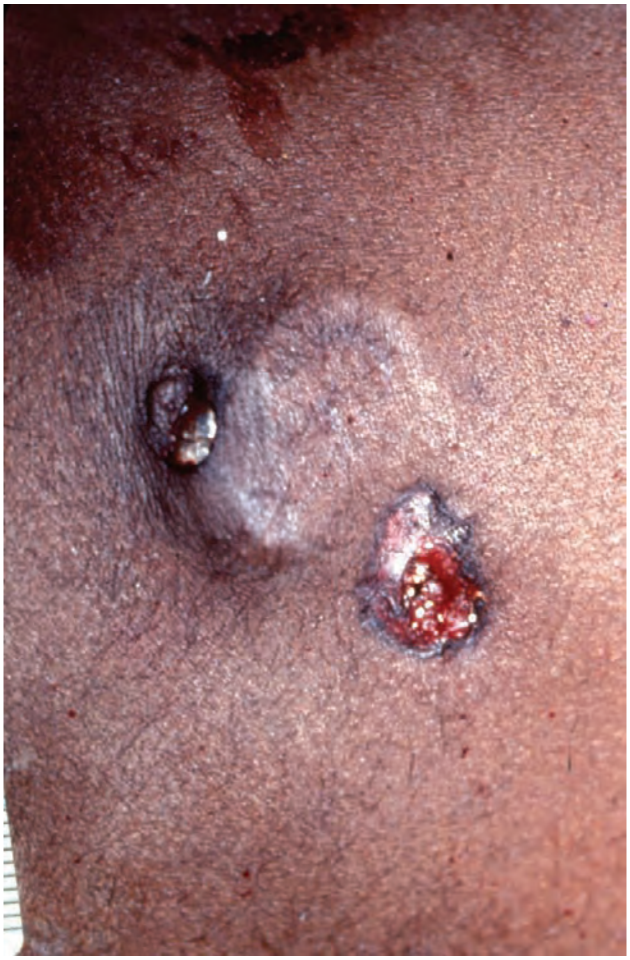

There is a large deformed gray metal slug just underneath the surface of the skin that is partially visible underneath the broken skin.

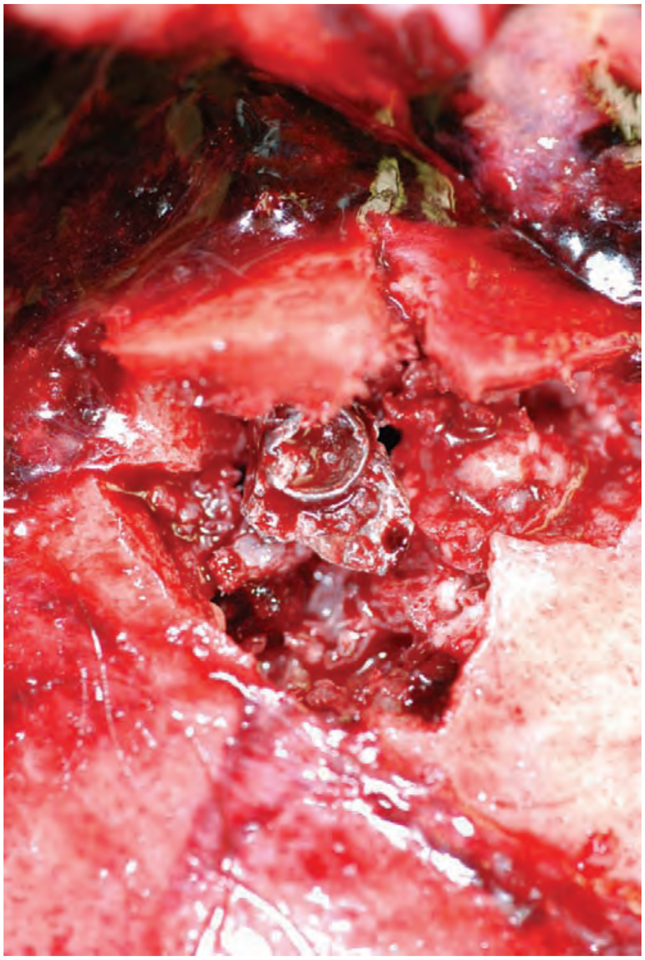

Bullet retrieved from skull fracture directly underneath the skin's surface.

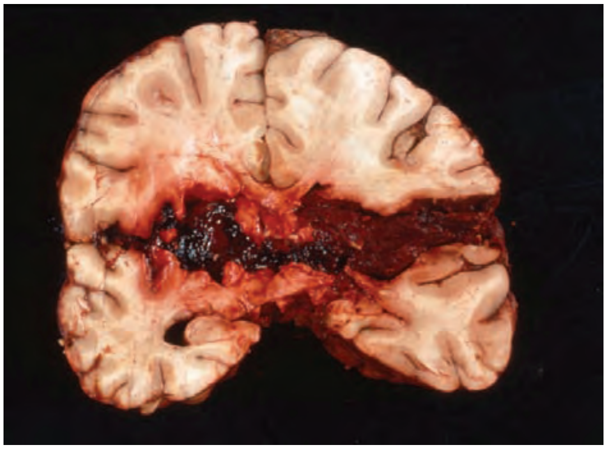

This image demonstrates a hemorrhagic pulpified wound track through the cerebral hemispheres caused by a medium-caliber handgun bullet.

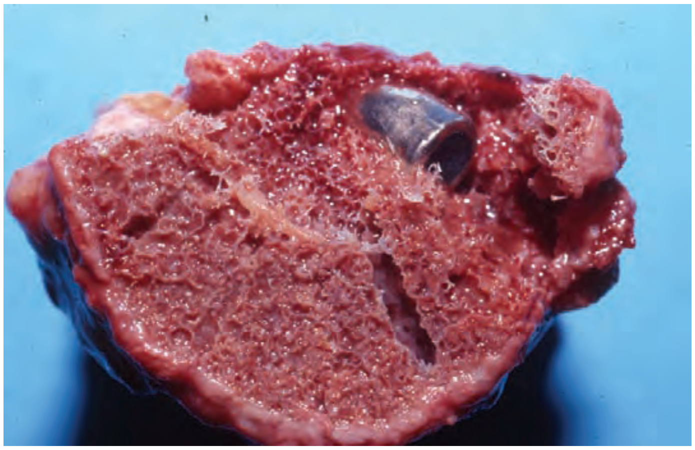

This bullet was retrieved within this portion of thoracic vertebra. It is important to carefully remove such bullets encased in bone, so as not to destroy ballistic markings on the sides of the bullet.

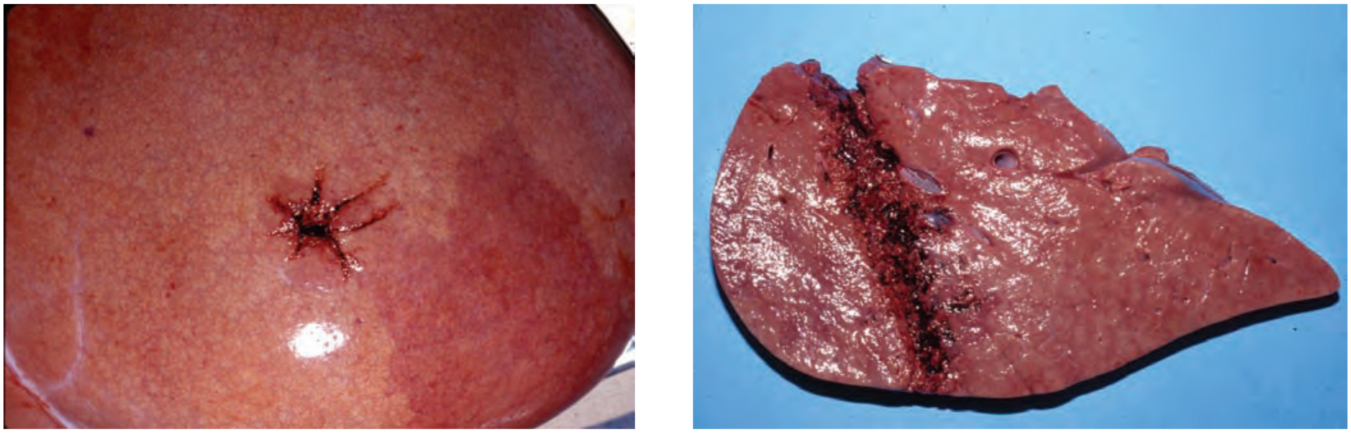

These figures demonstrate a wound track through the liver caused by a medium-caliber handgun.

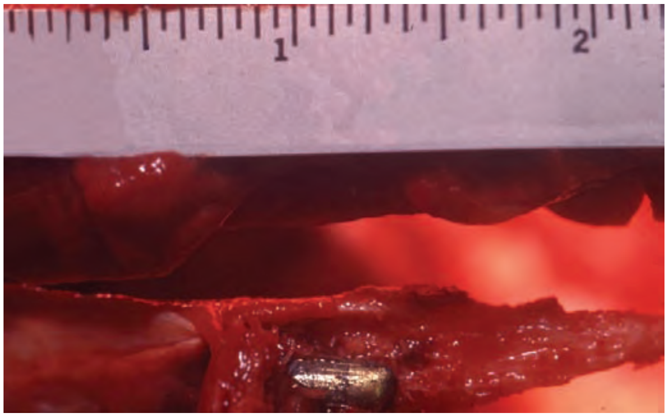

This individual had been shot several weeks earlier with a .22 caliber handgun. One of his ribs fractured while removing the breastplate during autopsy and a bullet was retrieved from within the medullary cavity.

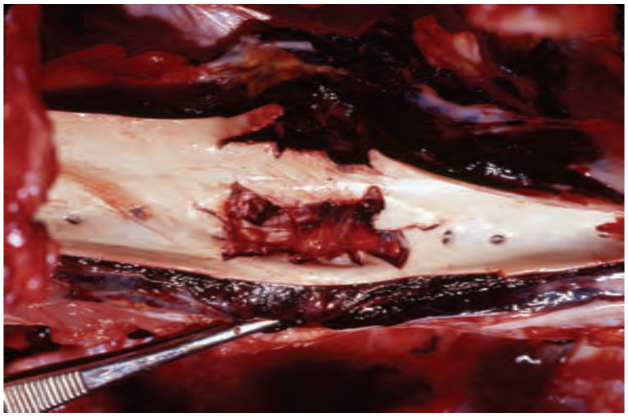

A wound track through the aorta with adjacent surrounding hemorrhage. This wound was caused by a medium-caliber weapon.

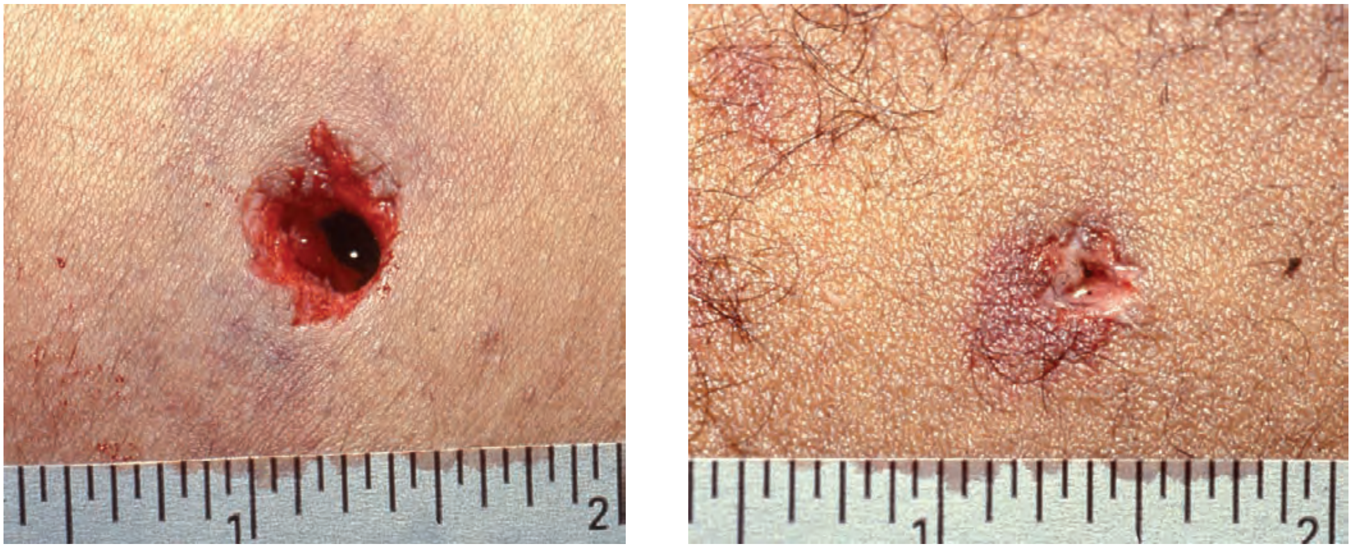

Shored exits. These wounds were produced by bullets exiting the body where the skin's surface is supported (e.g., by a firm surface or tight clothing).

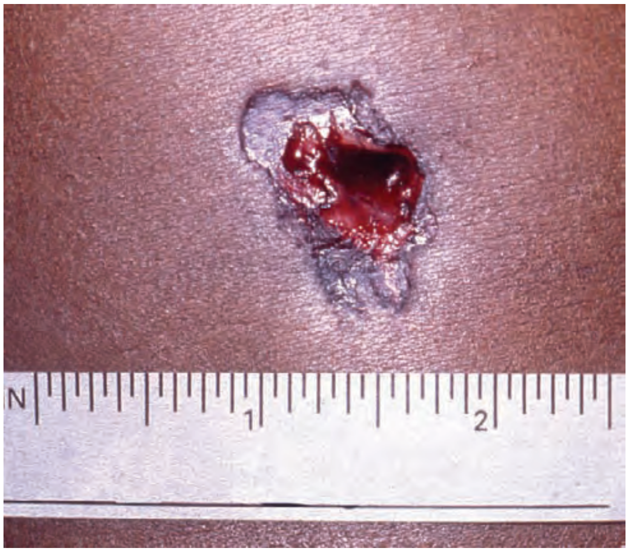

Atypical entrance gunshot wound caused by a bullet that was deformed by passing through an intermediate target. Note the large irregular abrasion caused by the deformed edges of the bullet's striking the skin surface.

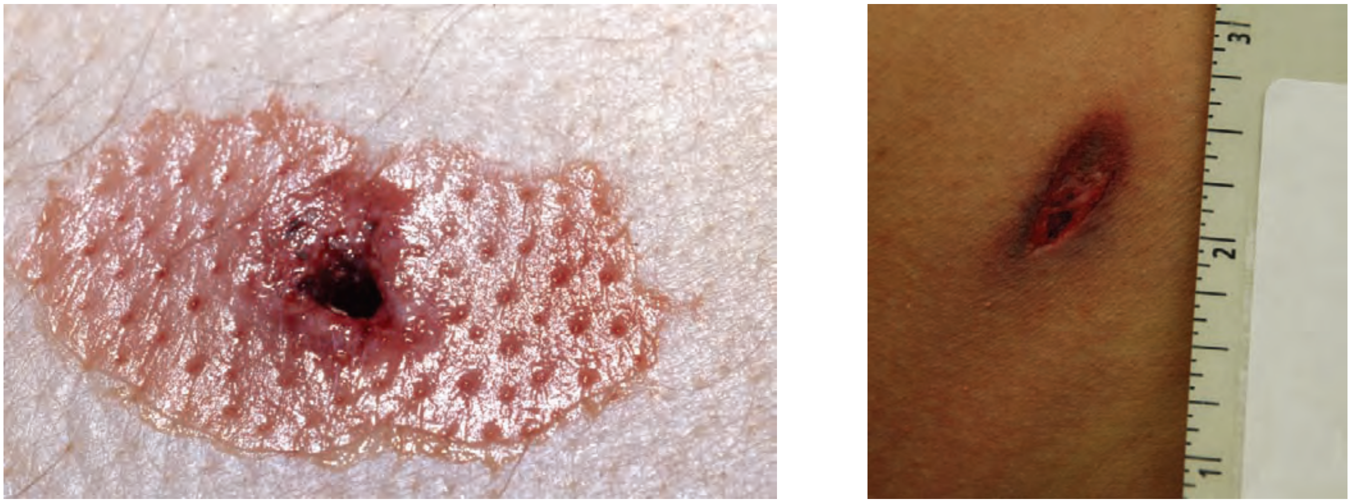

Shored exit wounds.

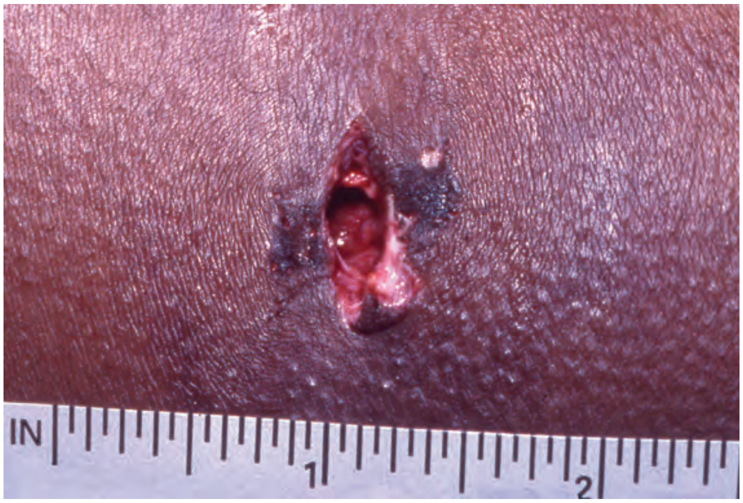

Irregular exit gunshot wound with adjacent abrasion due to its location on the body through a region of folded skin.

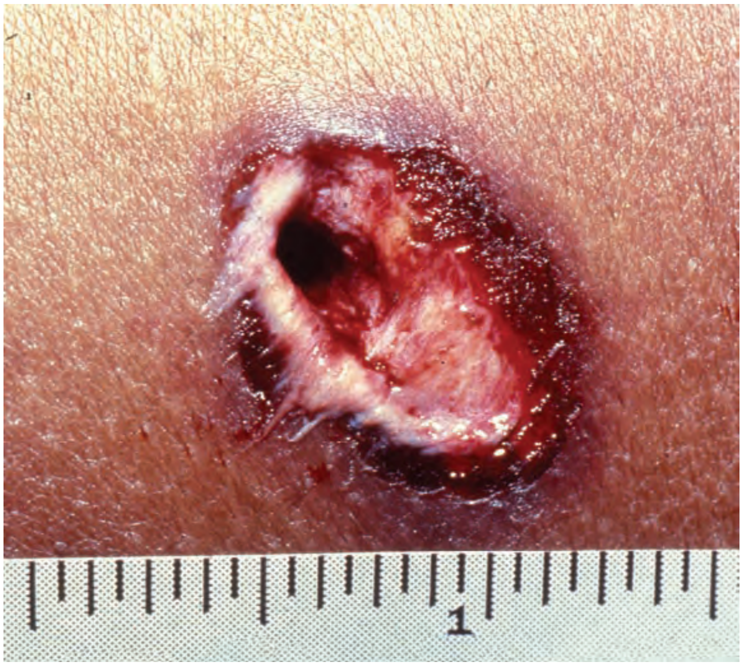

This distant-entrance gunshot wound has slight oval shape with an irregular margin of abrasion that is greatest at its inferior right aspect. This wound was produced by a ricochet bullet that first struck a brick wall. Part of the bullet had an irregular scratched surface. Also, this bullet most likely struck the body sideways.

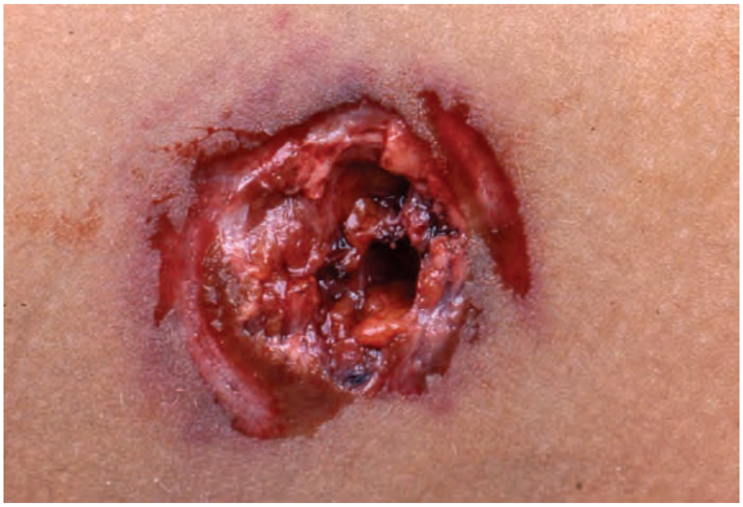

Atypical gunshot wound. This is a re-entry gunshot wound produced from a bullet that was markedly deformed from striking a bone in another part of the body. Note the irregular nature of the perforation with the irregular adjacent abrasions.

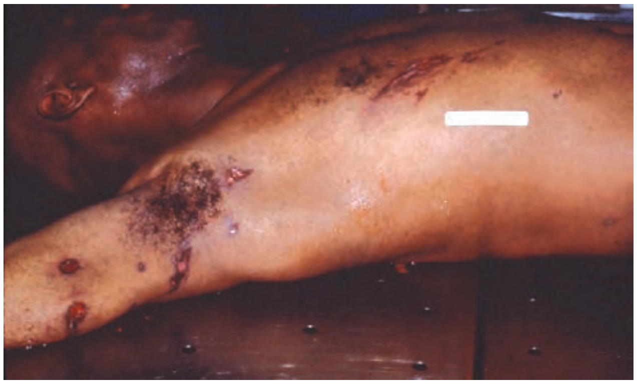

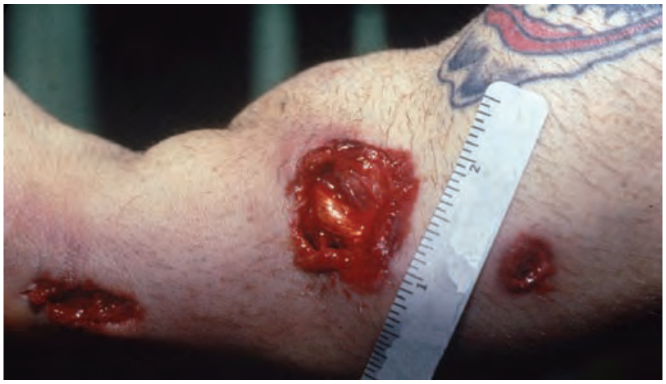

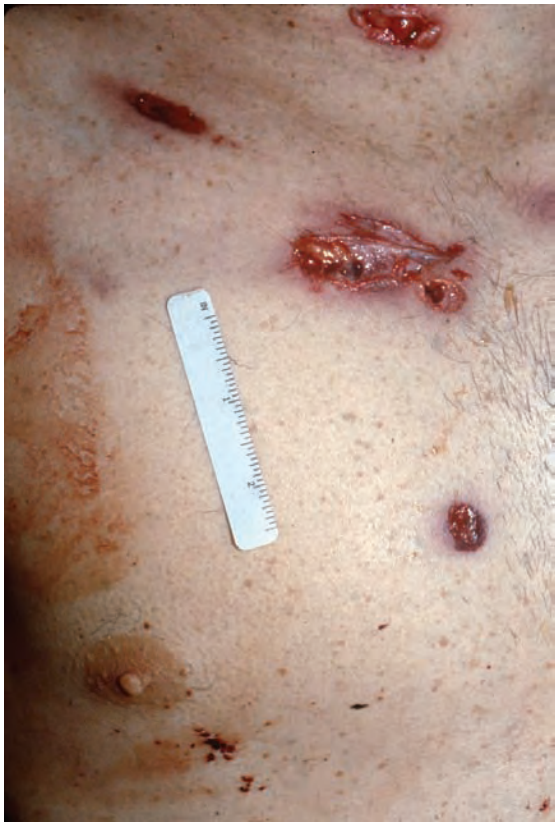

Atypical gunshot wounds from deformed and fragmented bullets that passed through an intermediate target, such as a car door and window, before entering and exiting the body. Note the markedly irregular nature of these injuries.

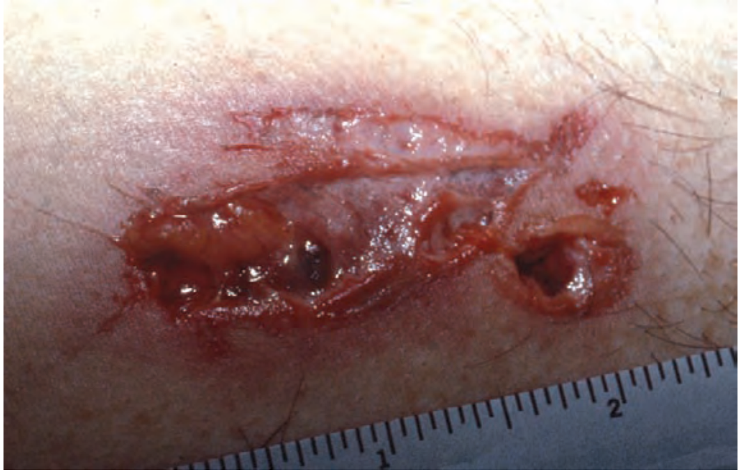

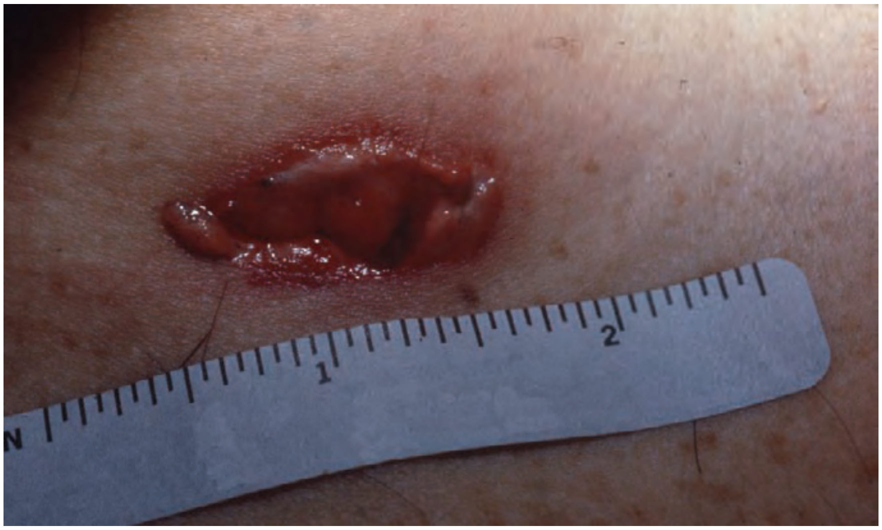

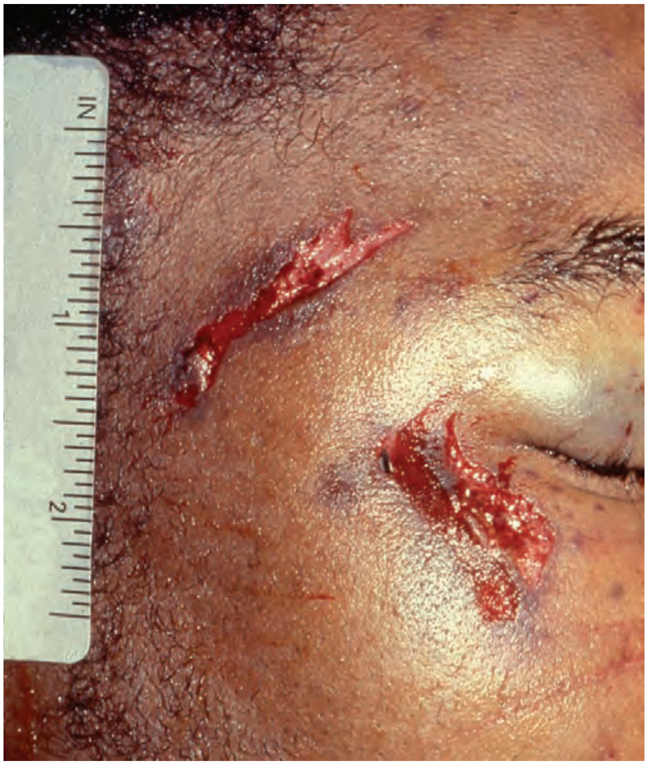

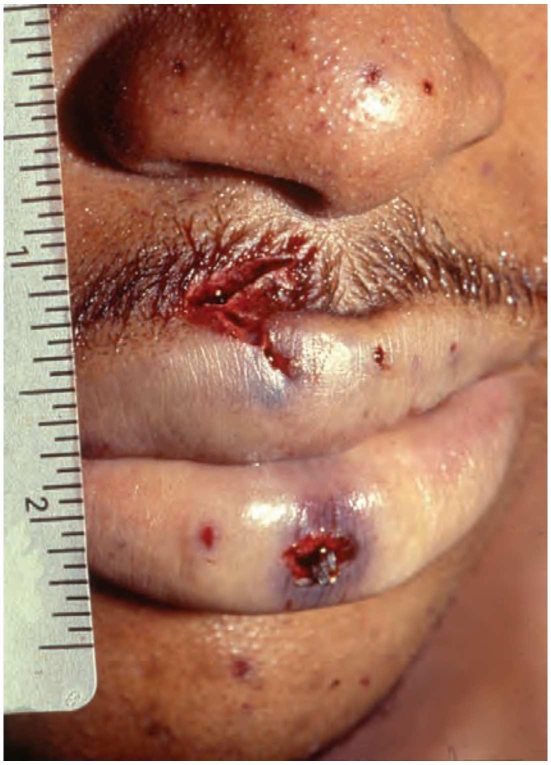

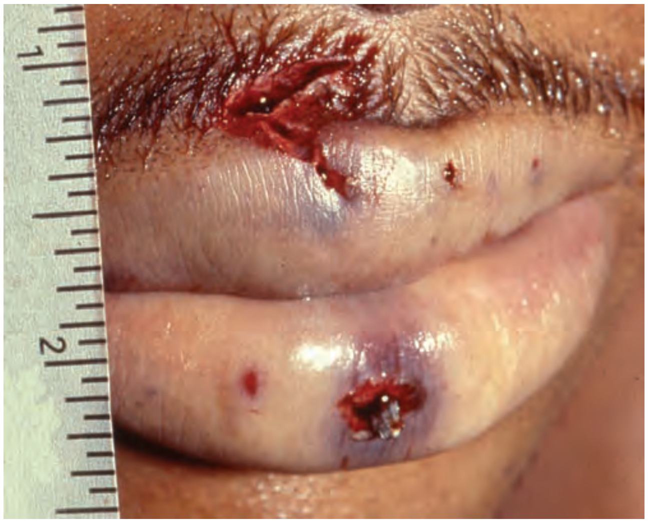

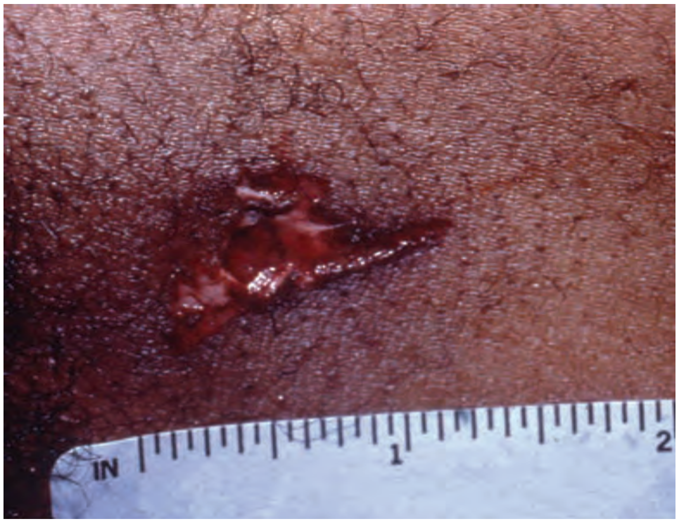

Multiple atypical entrance gunshot wounds produced by fragmented pieces of lead caused by .22 caliber nonjacketed bullets that struck concrete and fragmented before striking an individual who was lying on the sidewalk. Note the irregular nature of these injuries with superficial fragmented pieces of lead observed in several of the wound tracks.

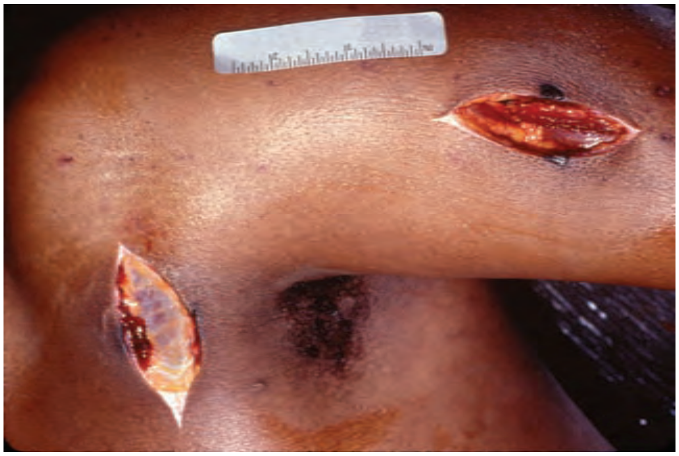

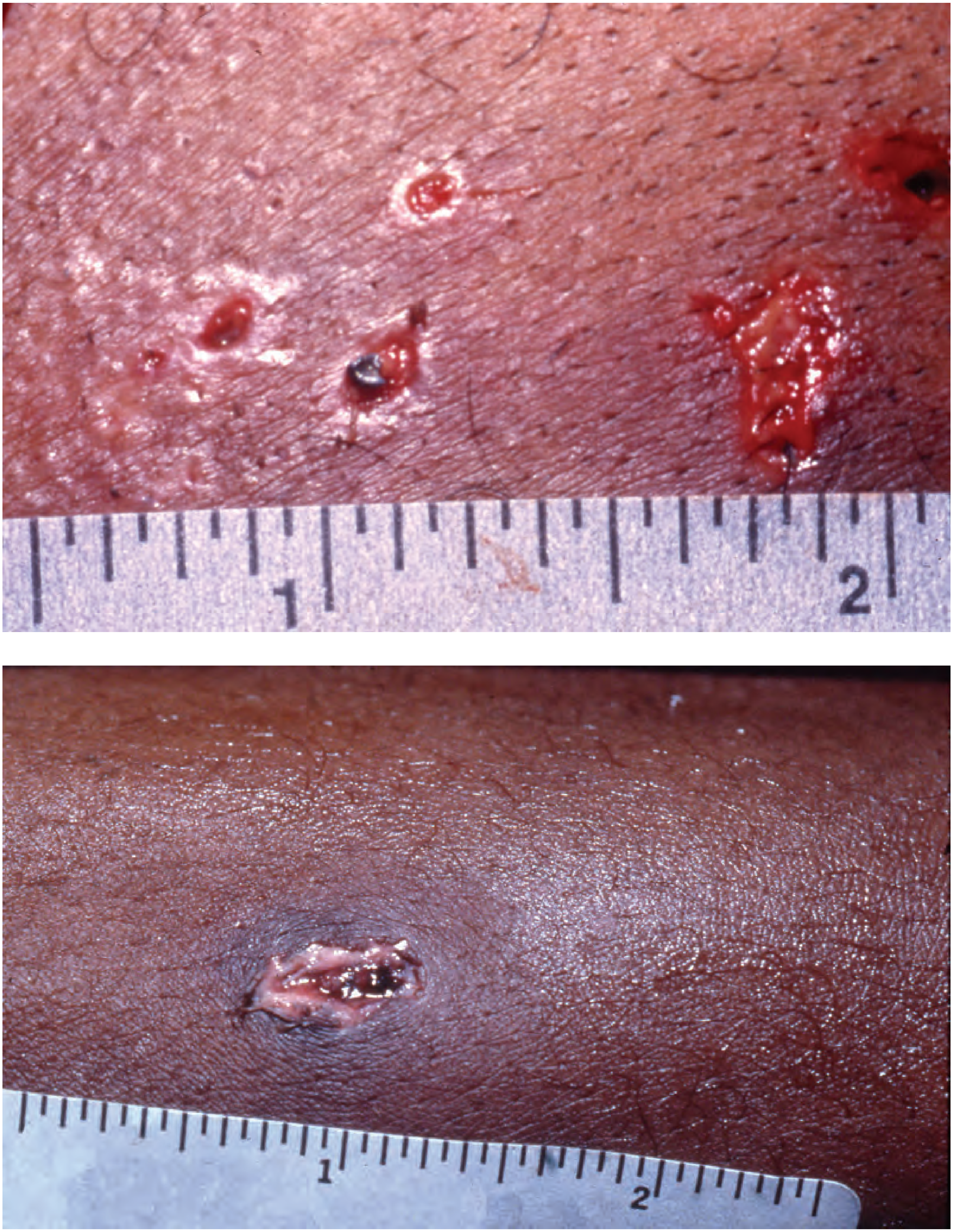

Several atypical injuries produced by fragmented nonjacketed lead bullets. The individual survived in the hospital for days. Note the healing margins and fragments of lead being pushed from the underlying soft tissue.

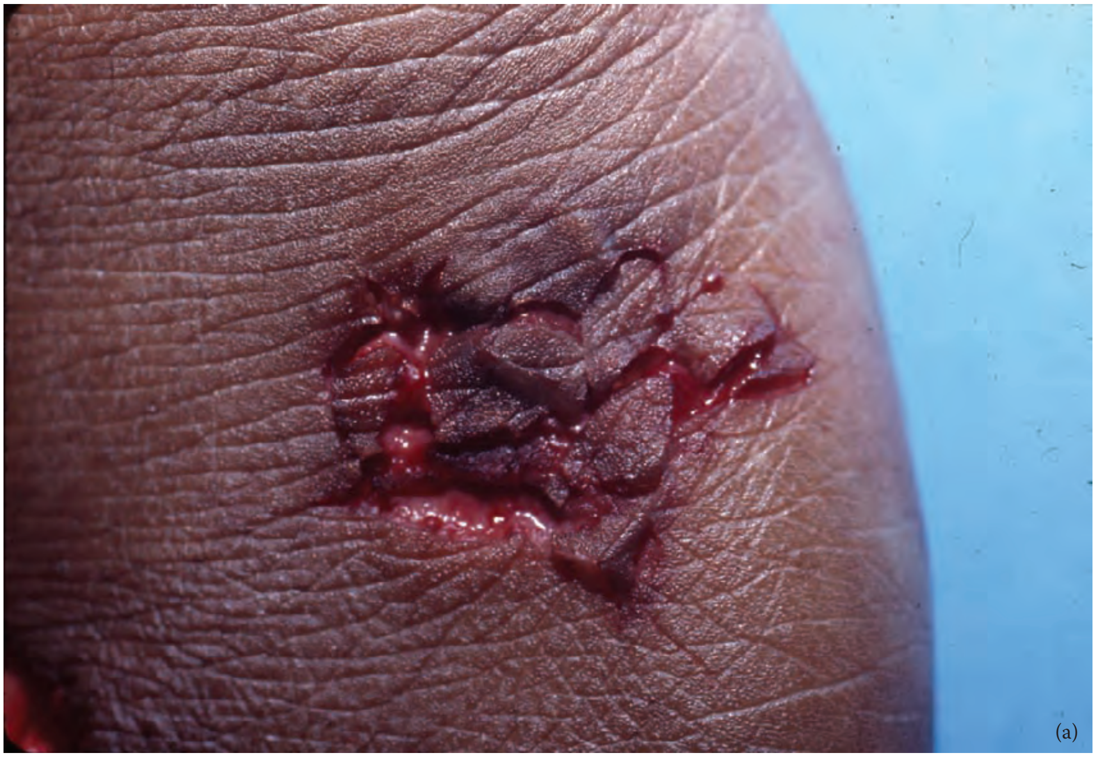

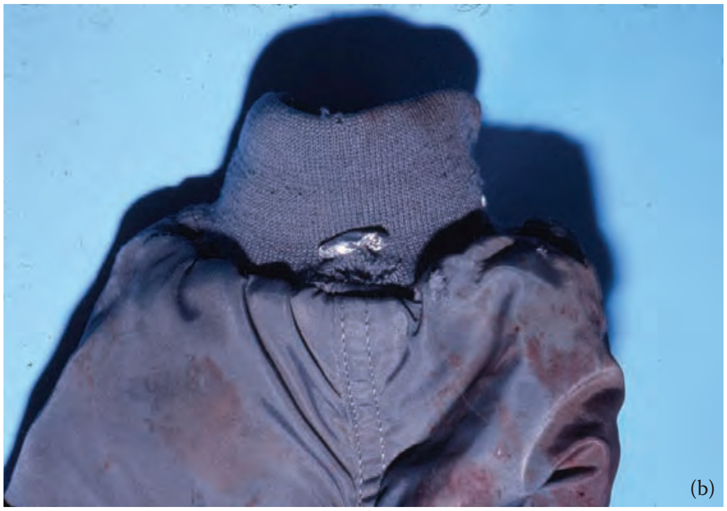

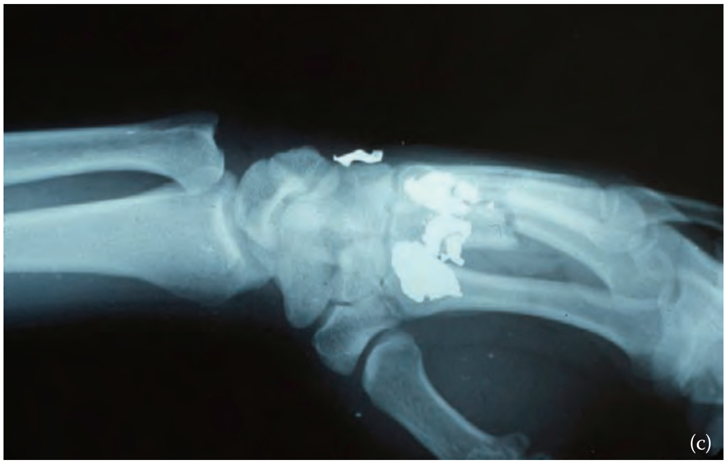

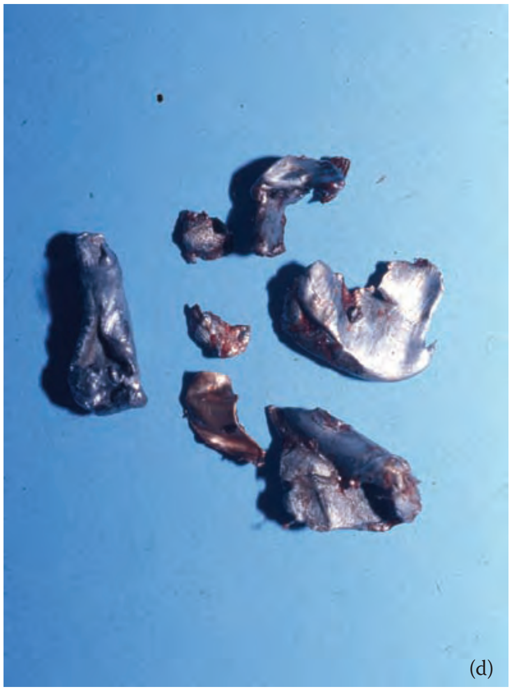

This child was shot multiple times. One of the bullets struck another object, fragmented, and produced this atypical entrance wound to his wrist (a). There were fragments of bullets retrieved from the decedent's jacket corresponding to this location (b). X-ray (c) shows multiple fragments of deformed metal retrieved just underneath the skin surface adjacent to the bones of the hand. The last image (d) shows all the bullet fragments retrieved from this individual's hand and wrist. There were no other wounds to this region of the body.

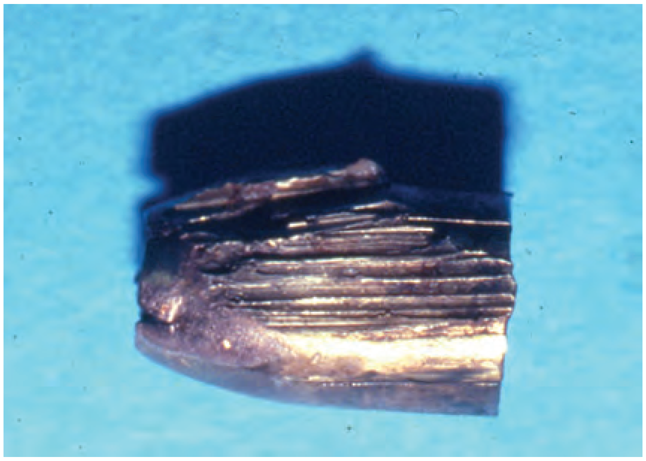

This ricochet bullet was retrieved within the abdomen of an individual with a perforated iliac artery. The bullet perforated the individual's shirt and did not strike bone while passing through the body.

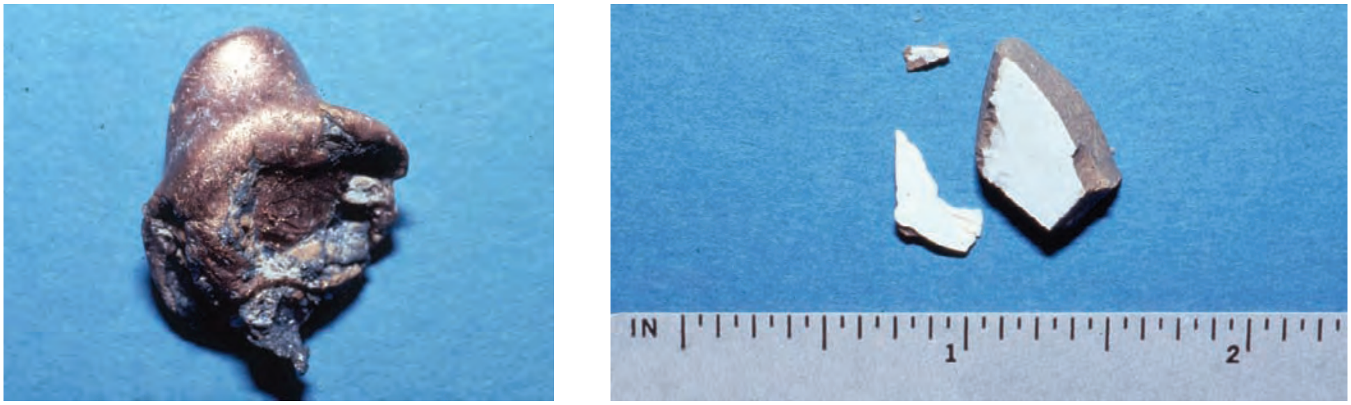

This bullet struck the ground, fragmented portions of tile, and became markedly deformed before ricocheting upward and striking the body of an individual already lying on the ground. The resulting gunshot wound was markedly atypical, producing irregular injuries to the body surface.

This deformed bullet was retrieved from an individual after it ricocheted off an intermediate target.

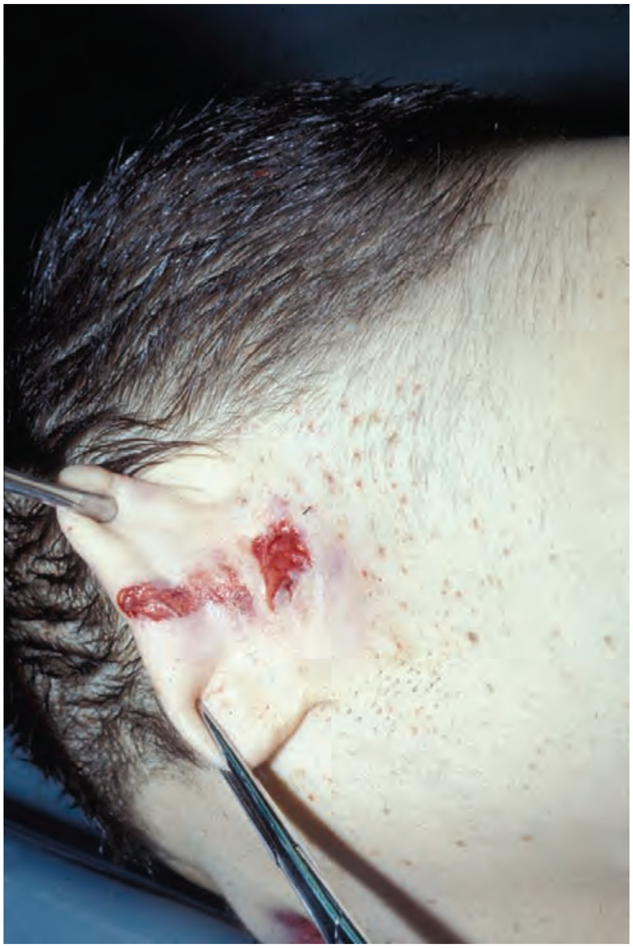

Intermediate-entrance gunshot wound with stippling. Note the irregular natural of the wound due to the location on the body. The bullet entered between the cartilage of the ear at a skin fold between the ear and the scalp. Note the elongated abrasion at the posterior ear due to the bullet's grazing the skin before entering the body.

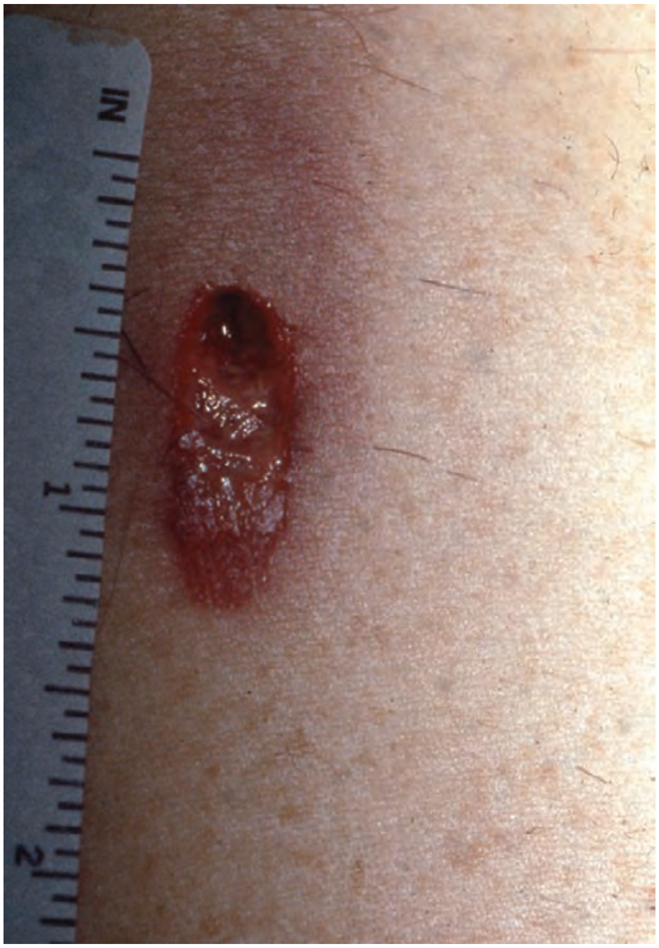

Distant-entrance gunshot wound with an asymmetric margin of abrasion. This indicates that the bullet grazed the body at an upward trajectory before perforating the skin.

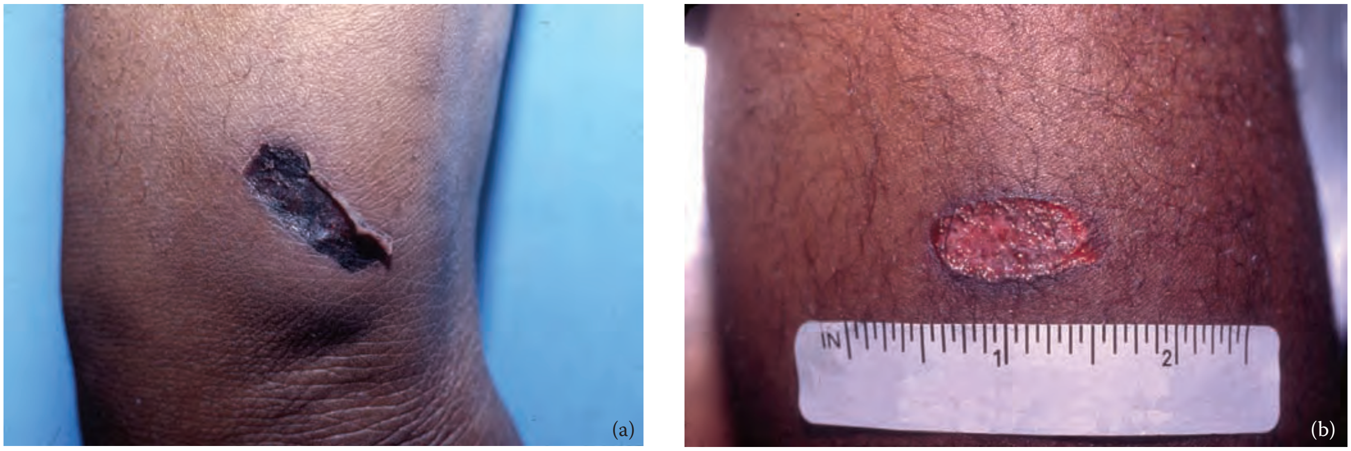

Graze gunshot wounds; (a) is more atypical.

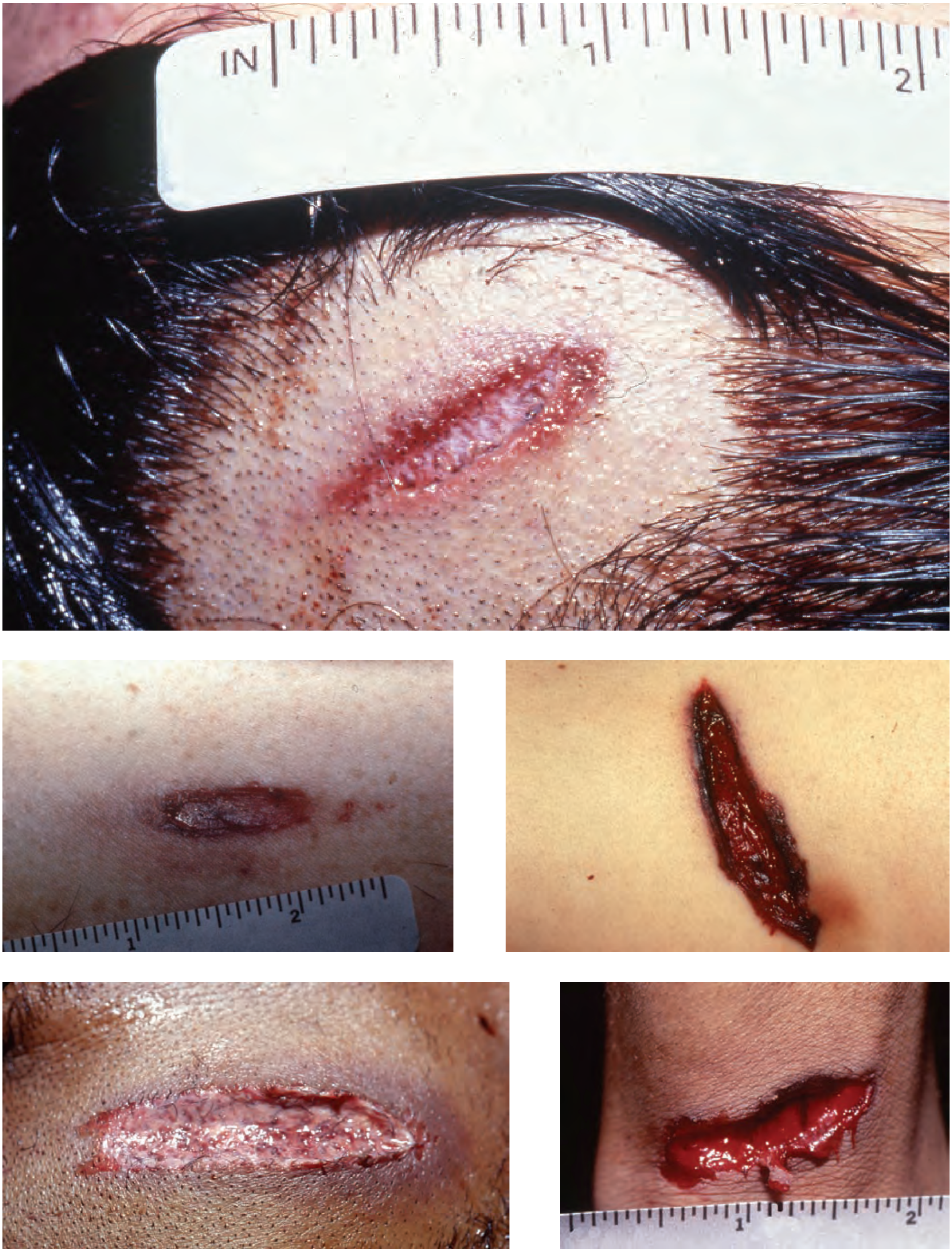

Graze gunshot wounds.

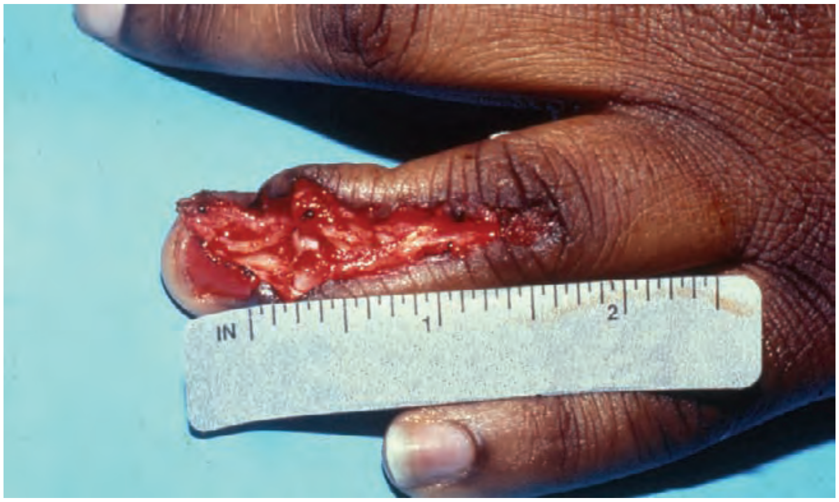

Tangential gunshot wound with a trajectory from the middle to distal aspect of the finger. Note the semicircular entrance defect at the right side of this wound overlying the proximal middle phalange.

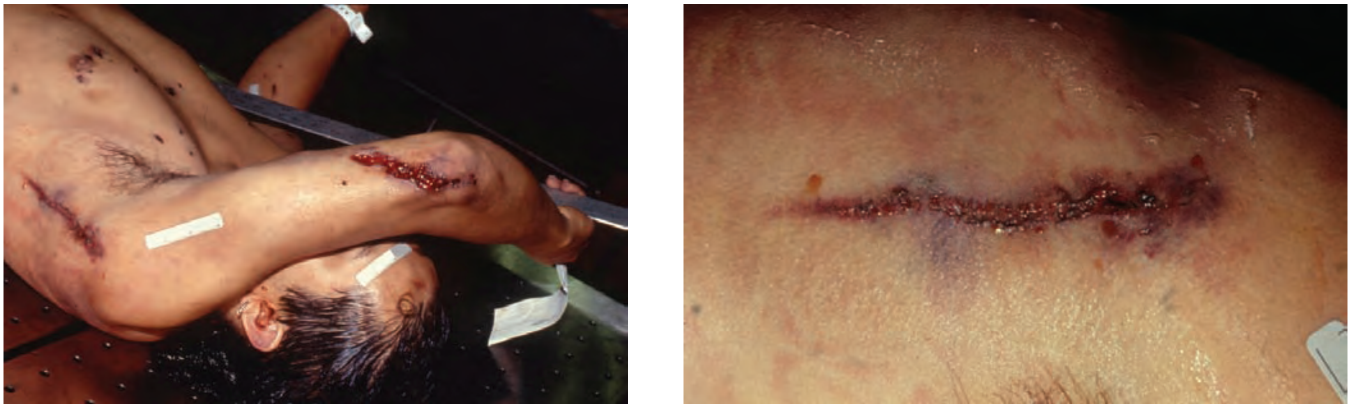

Tangential gunshot wounds to the arm and trunk. Note the skin tag formation at the wound margin produced as the bullet perforated underneath the skin's surface.

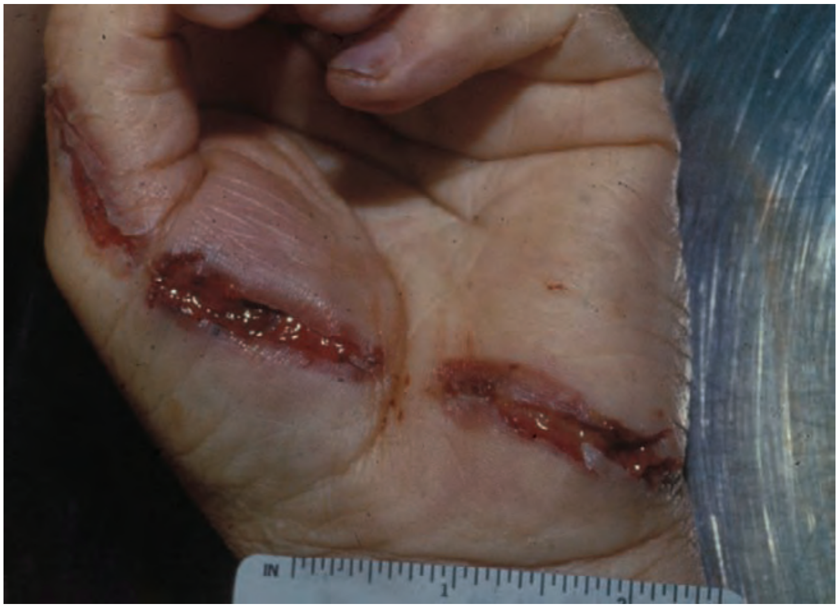

Tangential gunshot wound with skin tag formation. The direction of fire is from left to right.

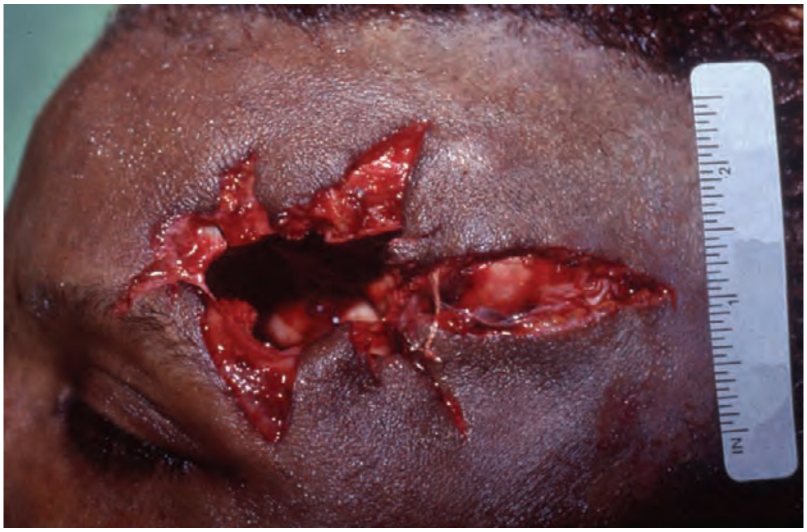

Distant tangential gunshot wound. Tangential gunshot wounds are produced when the bullet strikes the body at a narrow angle, producing skin tag formation. Usually the bullet exits the body leaving an open wound through the skin's surface, connecting the path of entrance and exit perforations. These wounds are deeper than graze gunshot wounds. This wound is associated with bullet fragmentation and partial exit, which may occur when the bullet strikes the body at a surface directly adjacent to underlying bone. This is an example with bullet fragmentation, partial exit, and underlying keyhole deformity of the skull. Note the entrance side of this wound is at the anterior aspect. There is a semicircular margin of abrasion leading into this laceration. The direction of fire is from left to right.

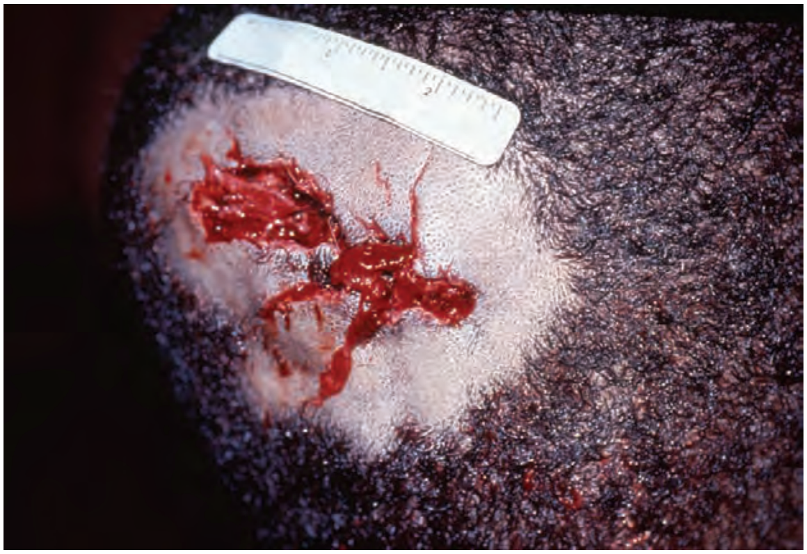

Tangential gunshot wound of the scalp. The direction of fire is from right to left. Note the skin tag formation pointing away from the semicircular entrance site at the right side of the wound.

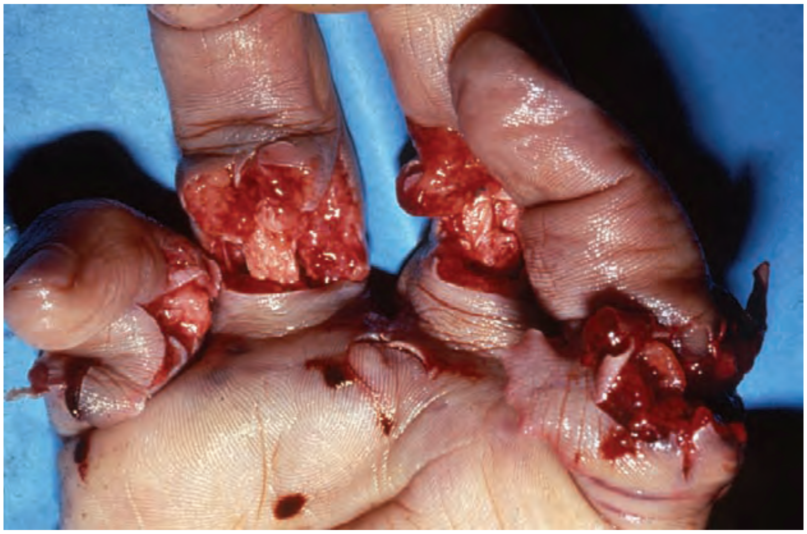

Note the tangential gunshot wound through the palmar surfaces of the fingers.

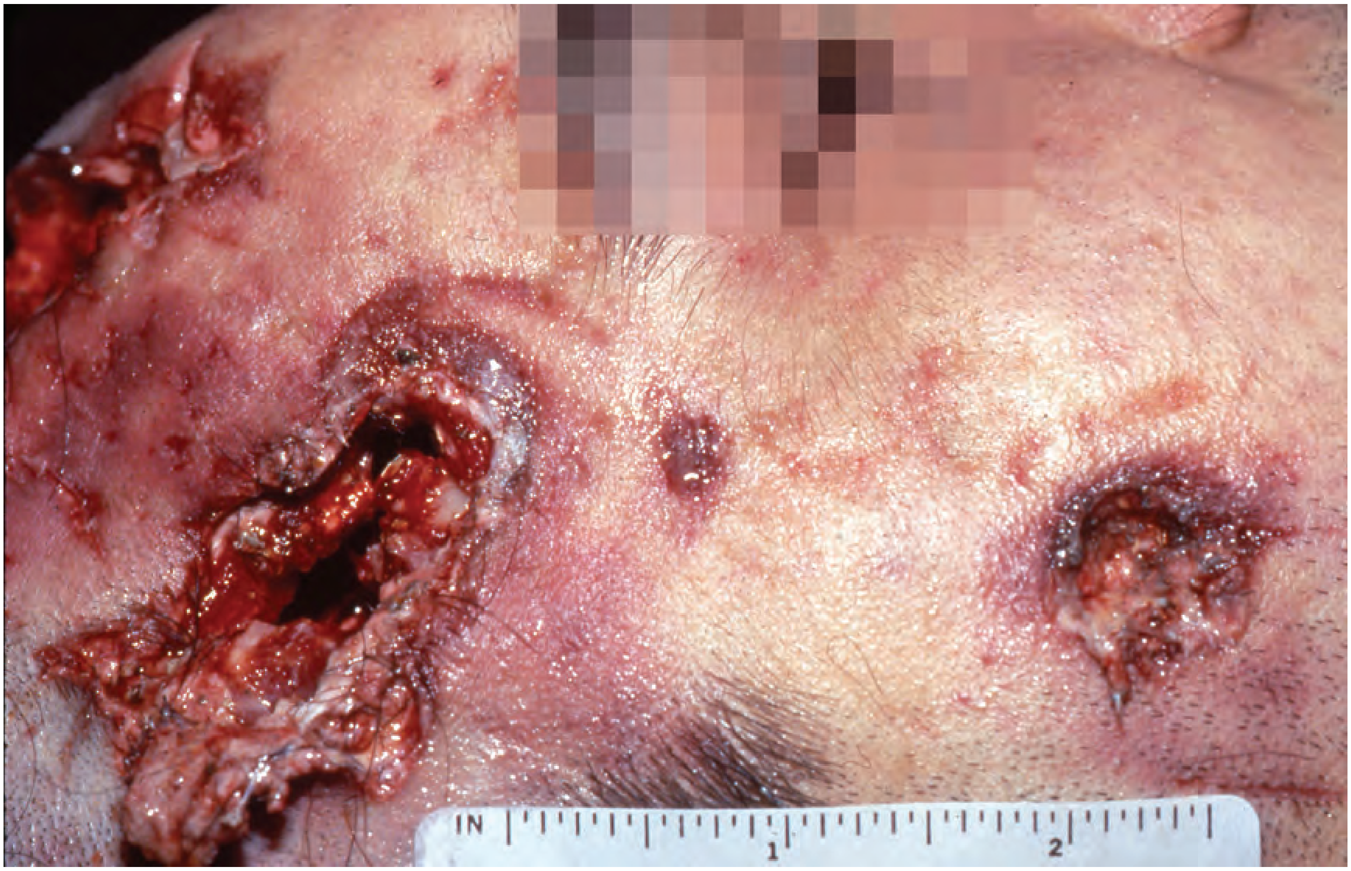

Tangential gunshot wound to the forehead. Part of this bullet entered the cranium and a portion of the bullet exited the body. There was an underlying keyhole deformity to the skull.

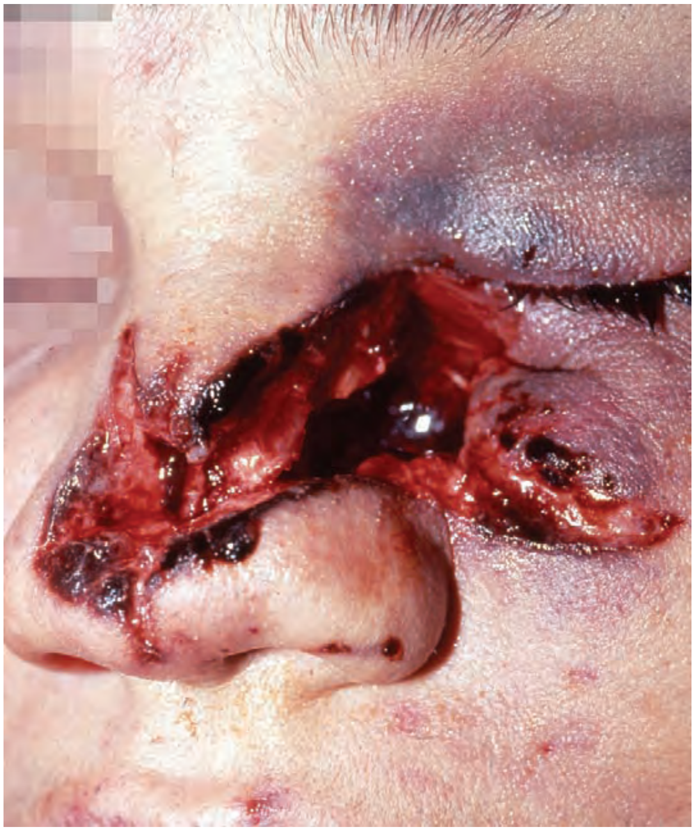

After this bullet perforated the fingers it produced a tangential gunshot wound to the nose before entering the body at the face.

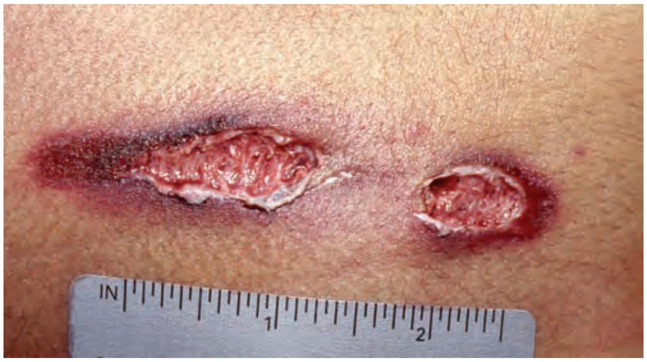

This demonstrates an entrance and exit gunshot wound that passed superficially beneath the skin surface. The direction of fire is right to left. Note that the perforation at the right side has a more uniform oval shape with a more symmetric margin of abrasion. The exit component to this wound is more irregular.

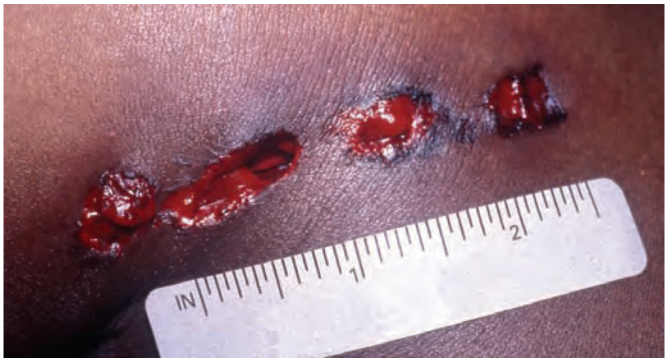

This gunshot wound produced multiple tears and perforations to the surface of the skin as the bullet passed close to the underlying surface. Such injuries often occur in regions of the body where skin folds on itself, such as the inguinal, gluteal, and axillary regions.

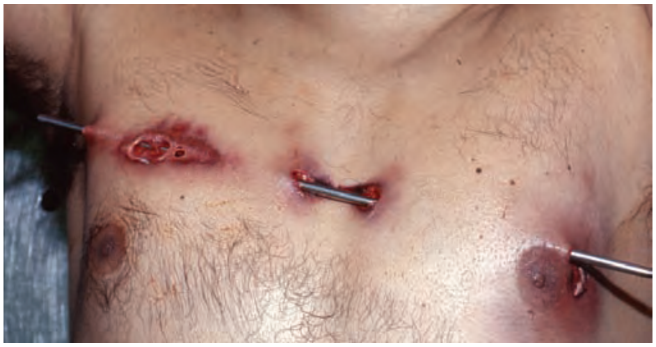

This single gunshot wound passed through the superficial soft tissues of the chest, producing multiple perforations and tears to the surface. Note the irregular nature of the torn skin and irregular abrasions at the exit reentry site.

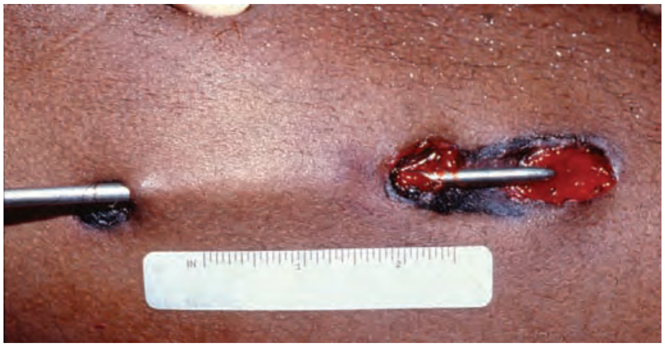

This is another example of an entry–exit–reentry wound where the bullet passed close to the underlying skin surface. Note the larger irregular abrasions connecting the exit and re-entry sites. Also note the dark discoloration due to drying.

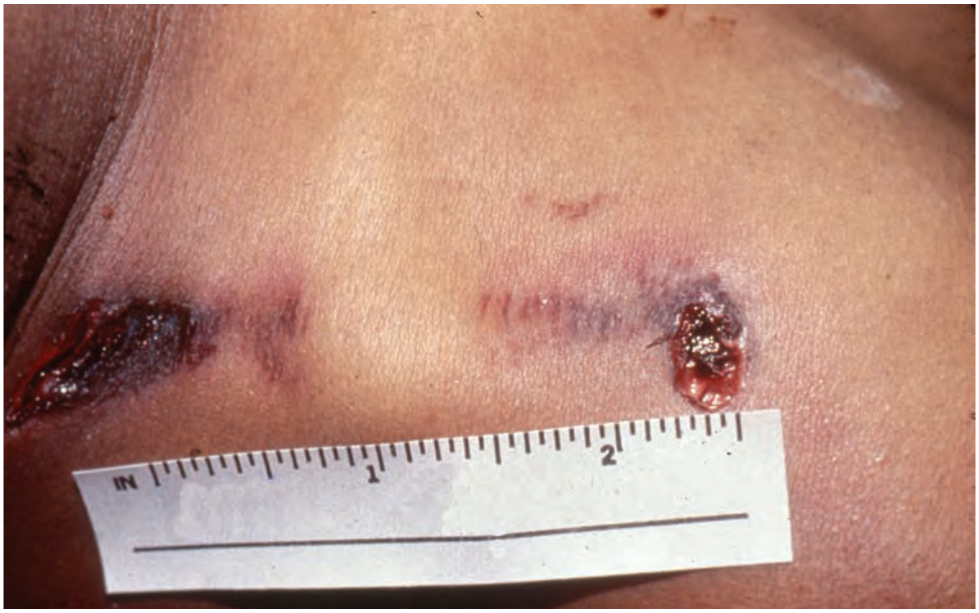

This is an entrance and exit gunshot wound where the bullet passed very close to the under surface of the skin.

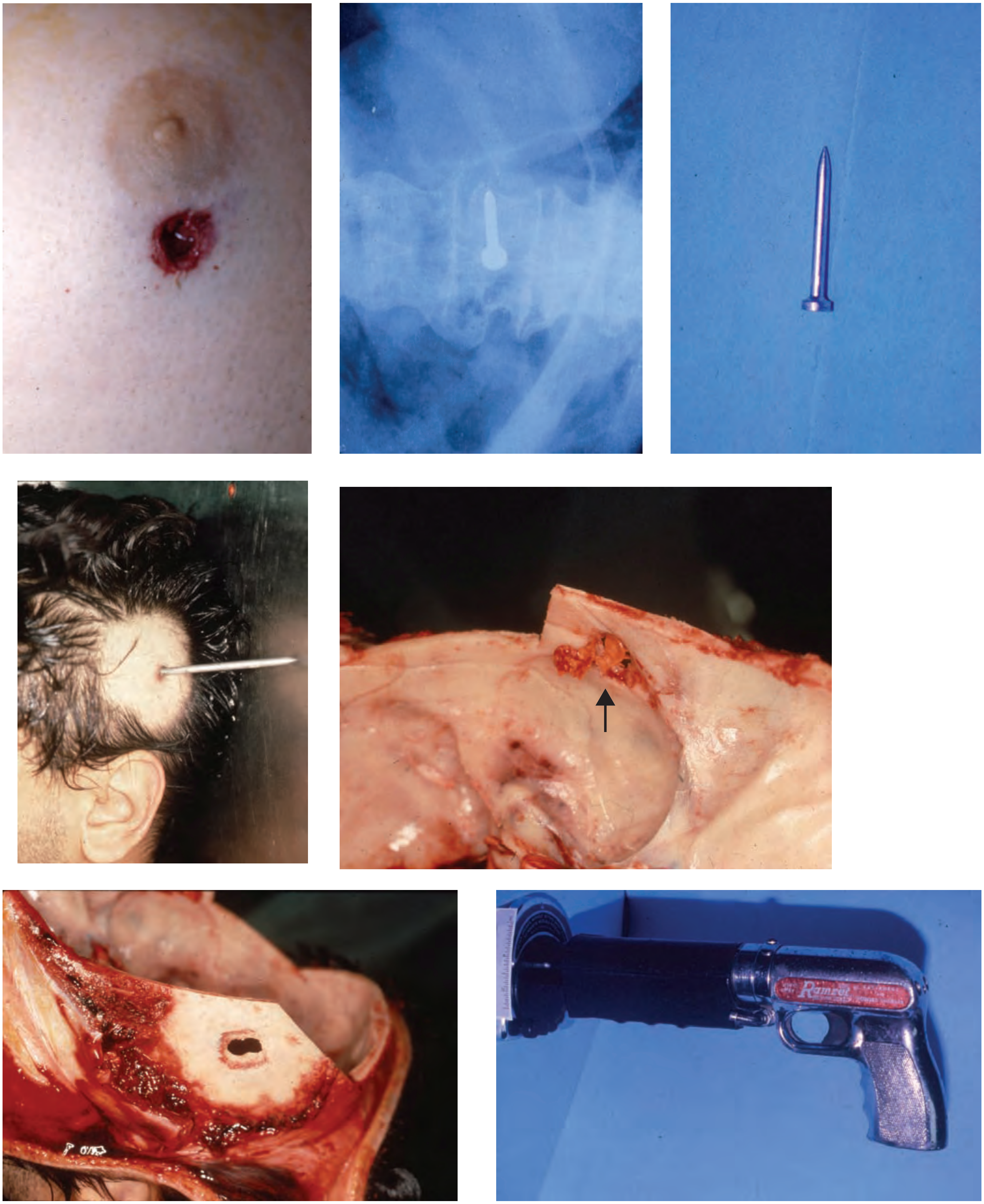

These images show different examples of nailgun wounds. Both of these cases were suicides. Both nailguns were similar to the one shown in the last image, which used gunpowder-loaded cartridges. Both entrance wounds consisted of circular perforations with symmetric margins of abrasion indistinguishable from typical entrance gunshot wounds. Note the orange plastic ejected into the entrance perforation of the temporal skull that is used in some nail guns to hold and steady the nail prior to discharge.

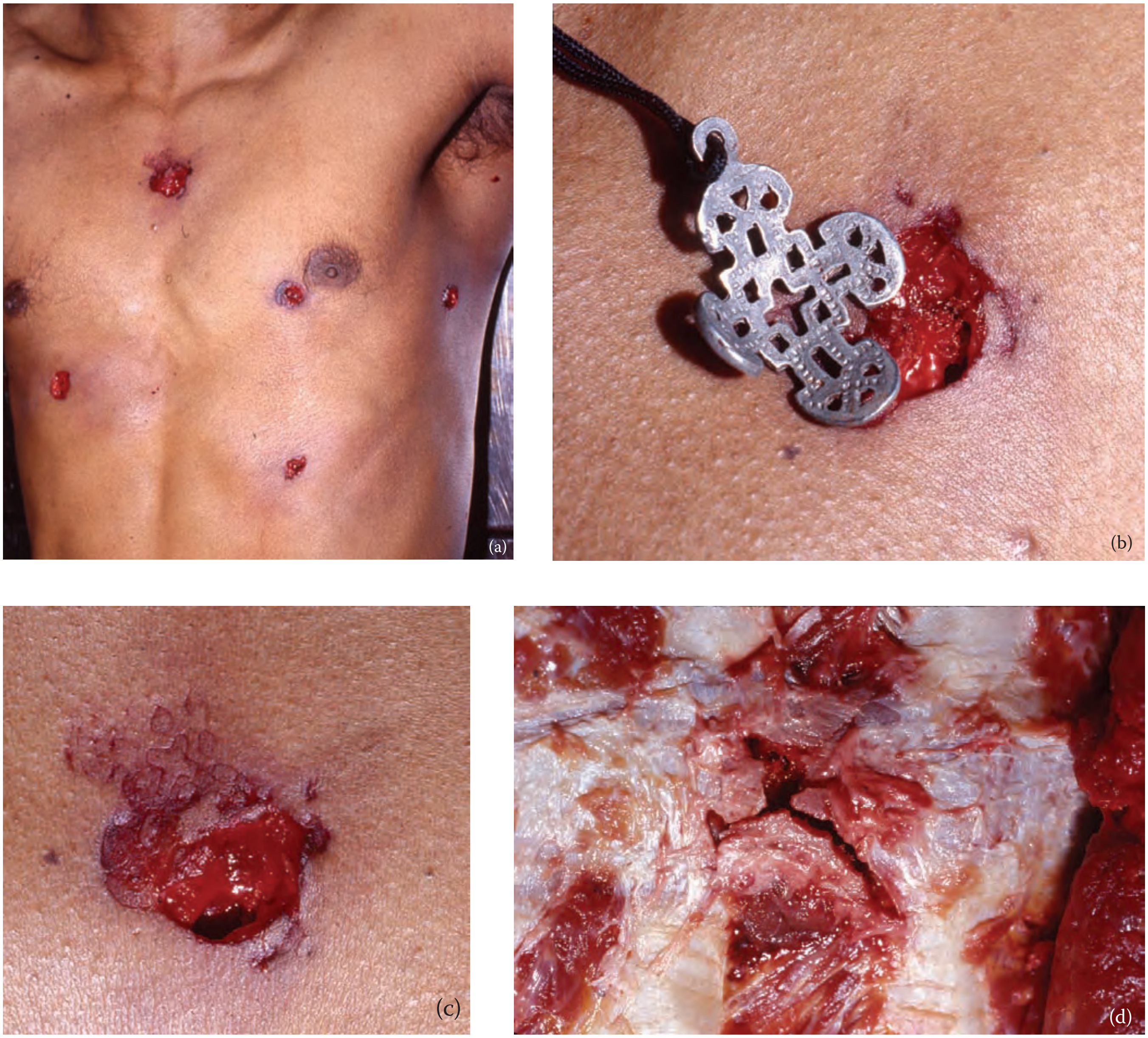

This individual was shot multiple times. Note the exit gunshot wound to the superior aspect of his middle chest (a). He was wearing a medallion on a string that was struck by the bullet as it exited the body, producing an imprint on the skin surface. The chest surface was most likely pressing against another object when the bullet exited. Figure (b) illustrating the medallion actually demonstrates it facing the wrong way. The medallion should be oriented with the inward curvature facing the chest. This was at first erroneously thought to be an entrance gunshot wound. Among other points arguing this to be an exit wound was the sternum fracture with bone splinters pointing in an outward anterior direction(d).

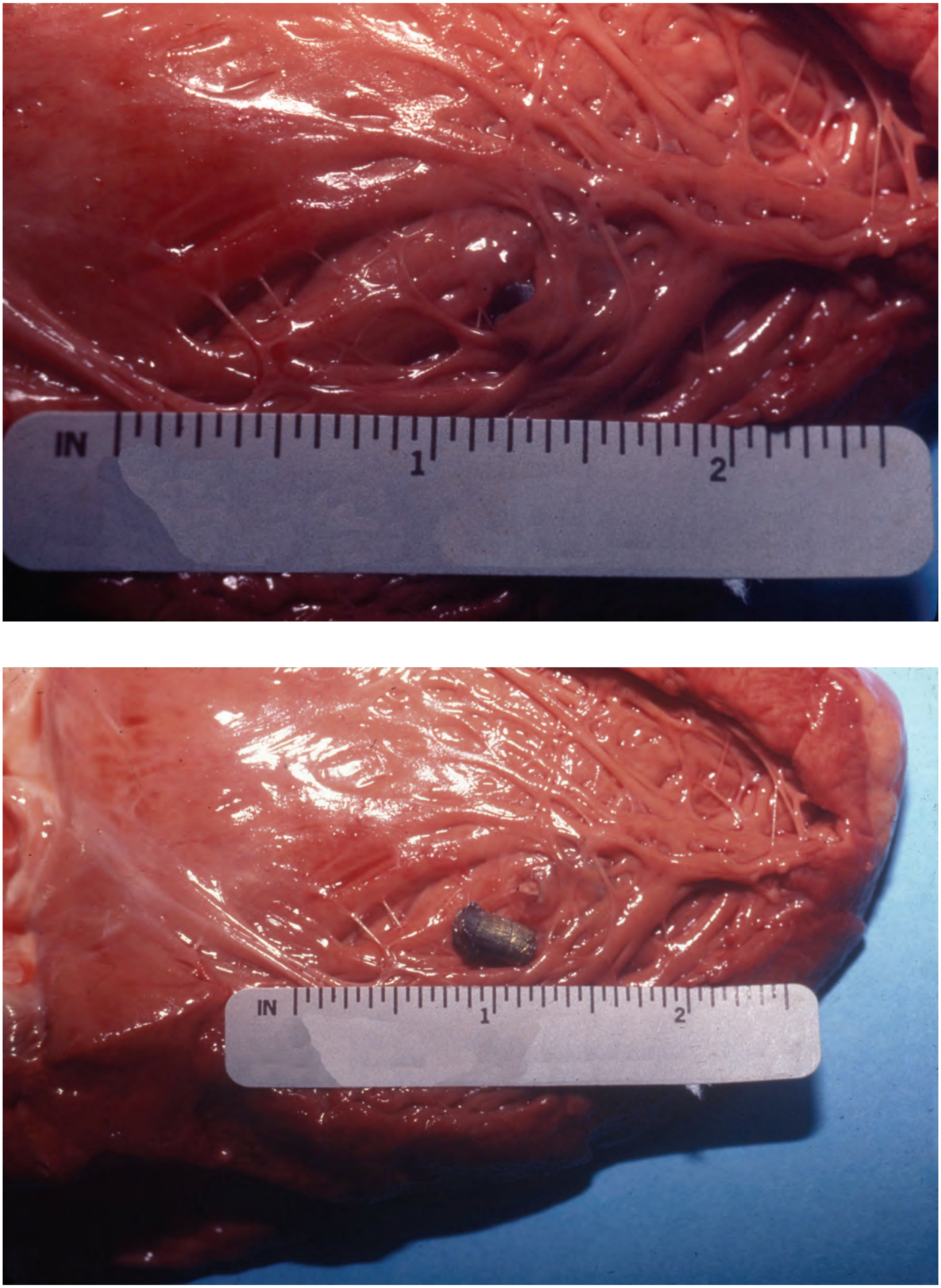

This individual was shot multiple times and survived in the hospital for approximately 2 weeks. One of the bullets entered the lung parenchyma and embolized to the heart, where it got wedged in papillary muscles of the left ventricle. He developed bronchopneumonia and his lung wound reopened, causing hemorrhage and death.

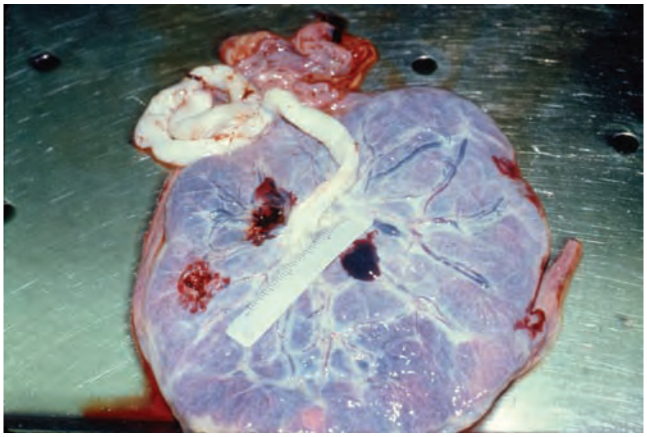

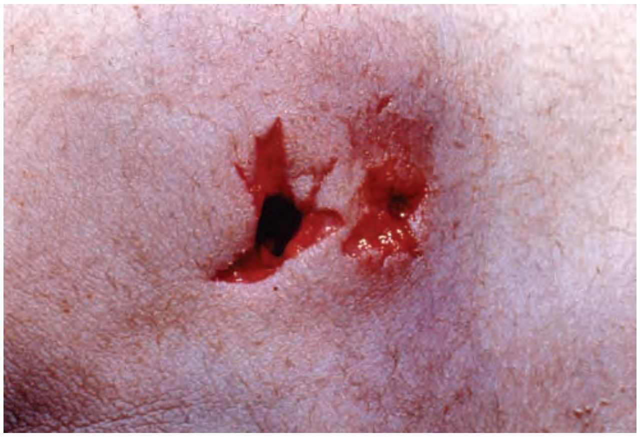

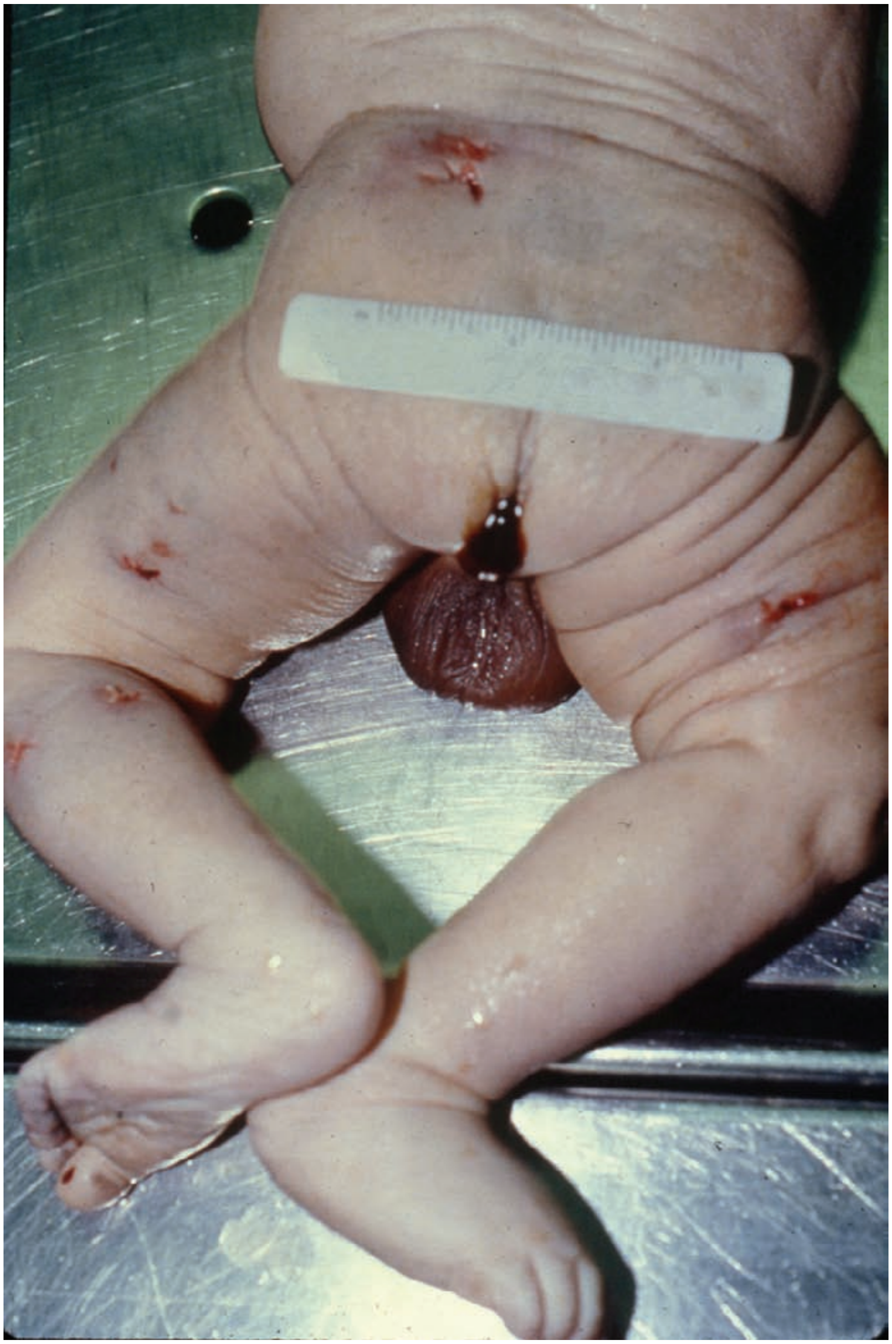

An otherwise healthy, full-term pregnant woman was shot multiple times in the abdomen. There were multiple perforating gunshot wounds to her fetus that were atypical due to the intermediate targets, including the mother, uteroplacental unit, and amniotic fluid. All the fetal wounds appeared as irregular lacerations with irregular abrasions. Direction of fire could not be determined.
Source for upper right and lower figures: Catanese C. and Gilmore K. A case report and brief analysis of fetal gunshot wound characteristics. *Journal of Forensic Science* 47:1067–069; 2002. Reprinted with permission from ASTM International.

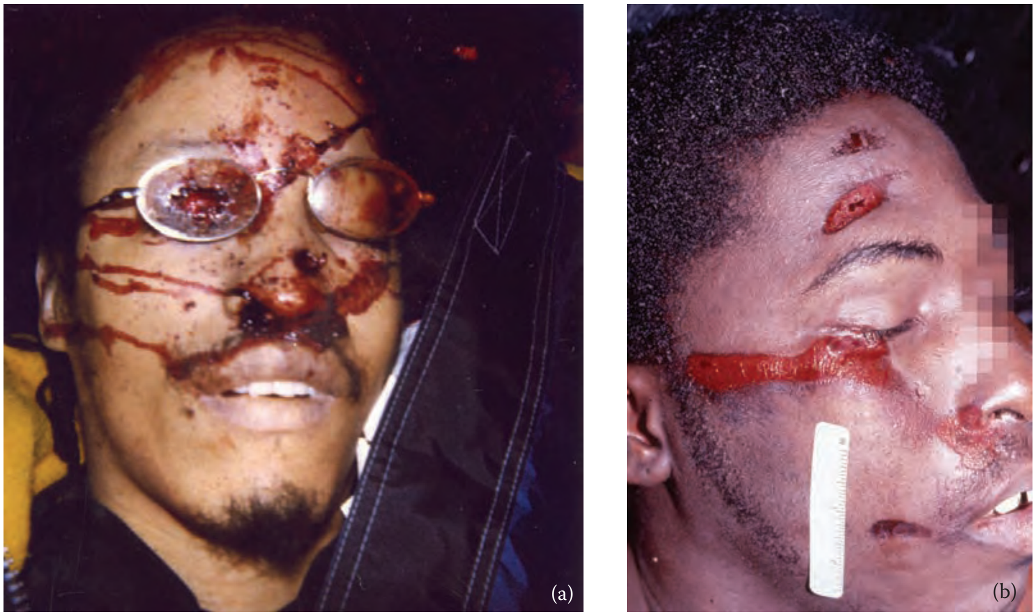

This individual was shot in the face while wearing eyeglasses. Note the irregular nature of the entrance gunshot wound just beneath the eye (a) and the elongated abrasion extending across the cheek to the decedent's ear (b), corresponding to the eyeglass frame.

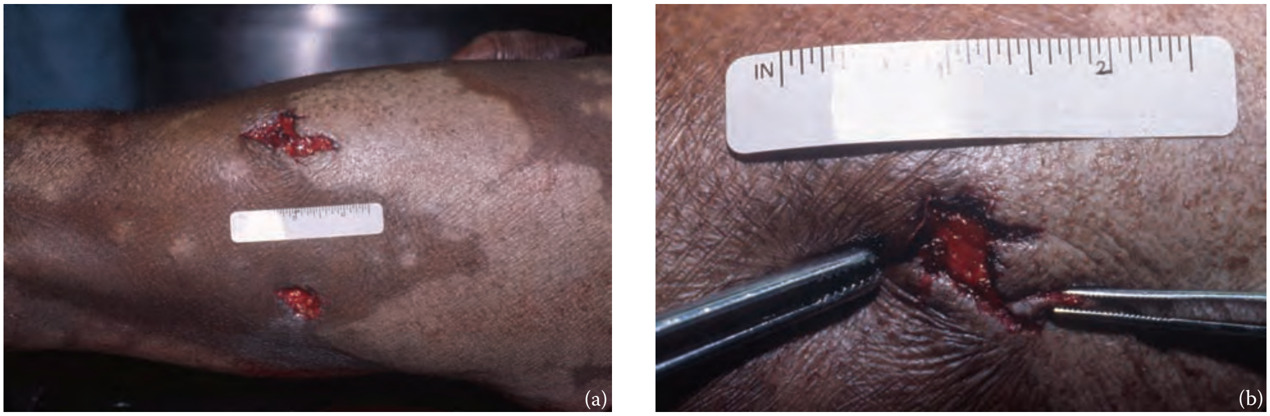

This individual was shot multiple times and had psoriatic skin disease. Note the irregular nature of this entrance gunshot wound through one of the plaques (a). The skin had small radiating lacerations and decreased abrasion similar to what is sometimes seen with gunshot wounds through thick skin found on the palms of the hands or soles of the feet (b).

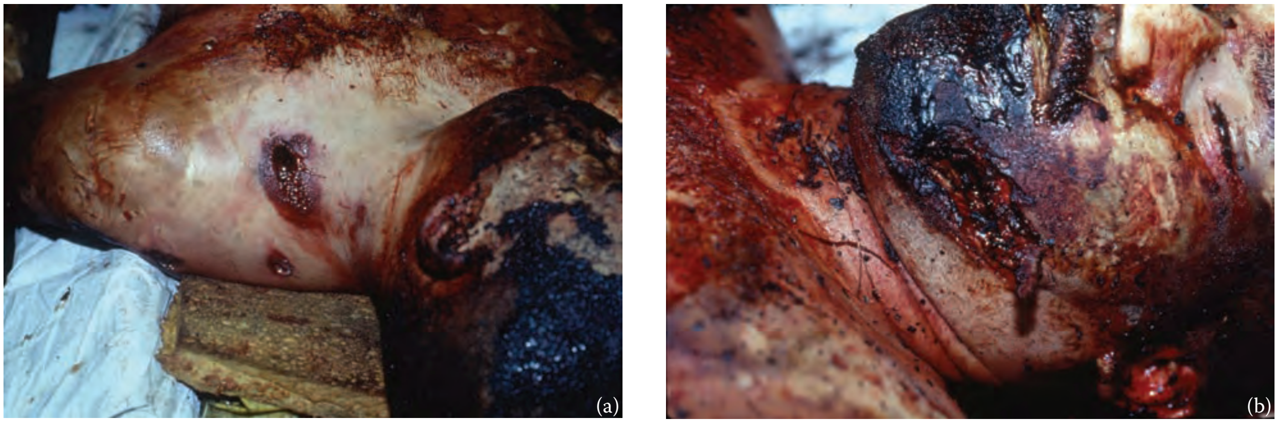

These are gunshot wounds to an individual who was set on fire to destroy evidence after being shot in his residence. His carbon monoxide blood level was negligible. Note the change in character of these injuries due to thermal effect. These thermal burns are postmortem.

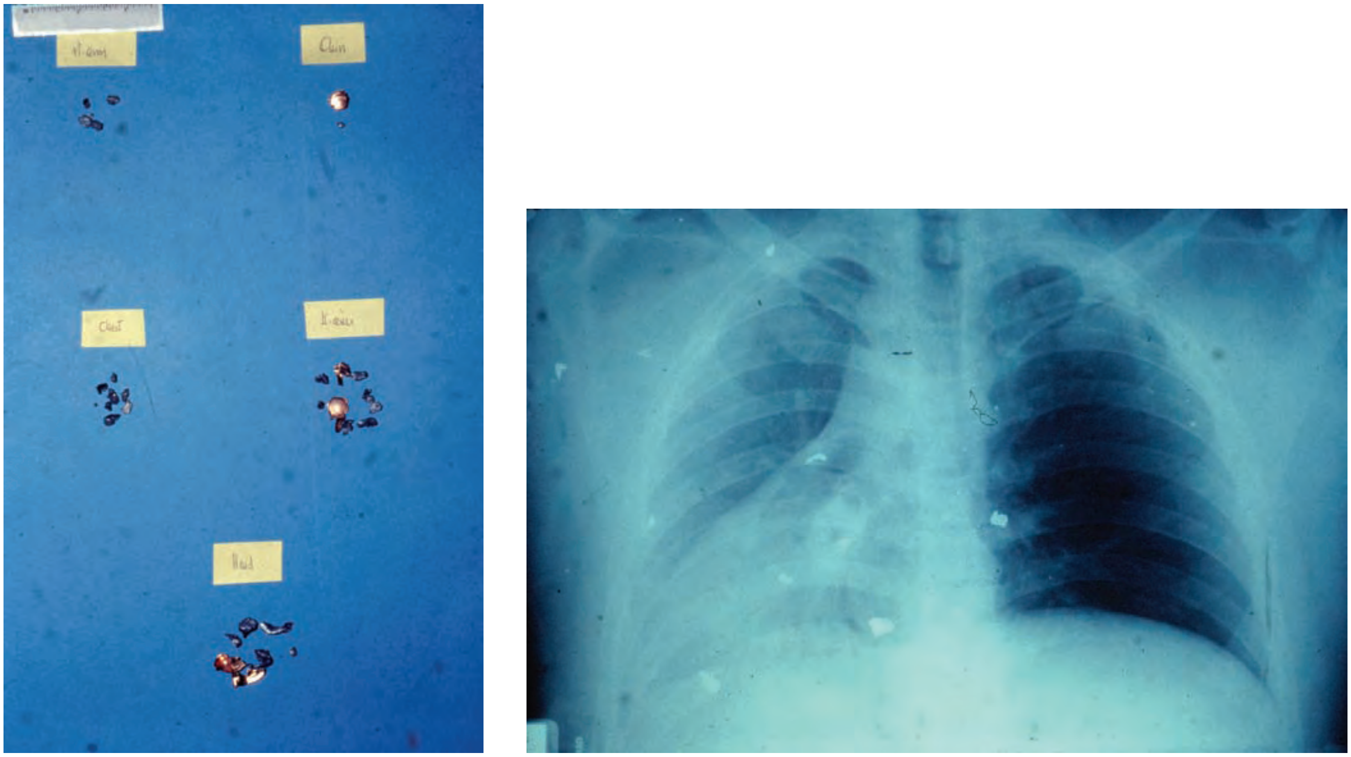

This individual was shot multiple times where the bullet struck bones and fragmented into multiple pieces. It is important to keep track of which bullets are associated with which gunshot wounds. There may be three different shooters and only one of the bullet wounds lethal. Linking the lethal bullet to a particular shooter may have different legal implications.

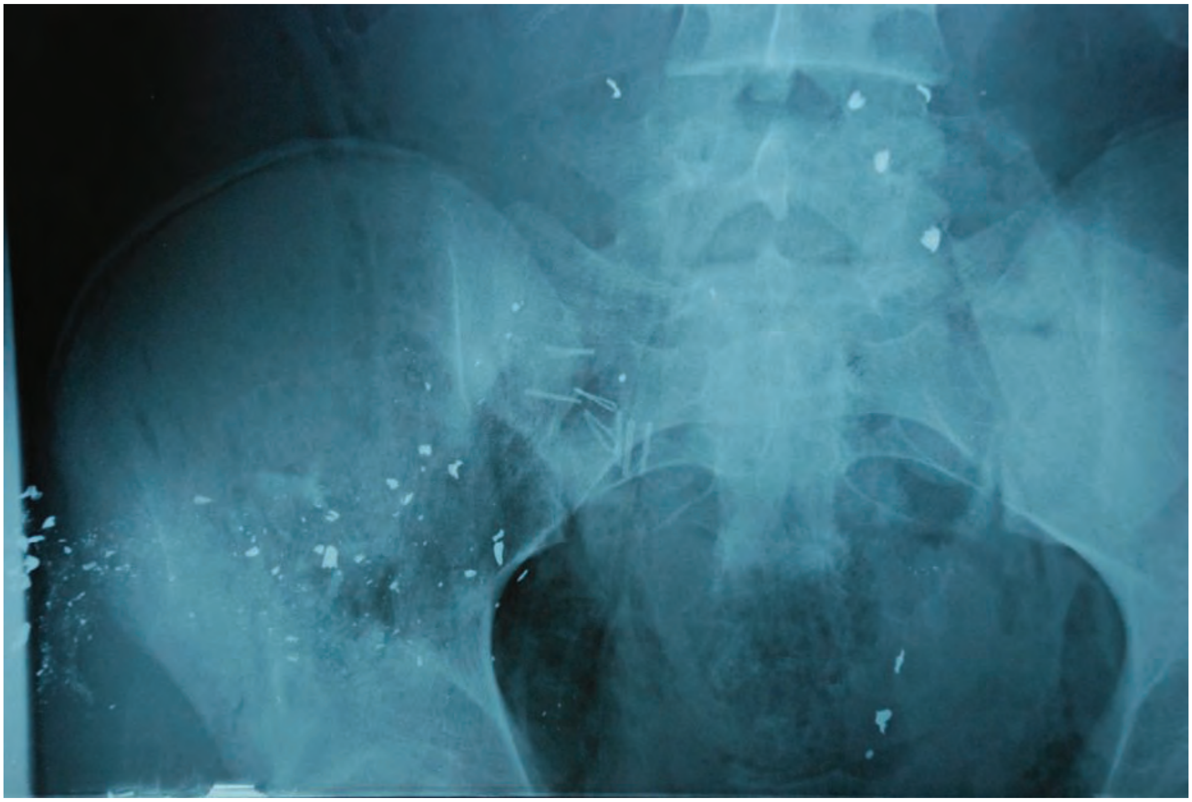

This individual was shot in the abdomen with a hunting rifle. Note the snowstorm effect of bullet lead fragmentation depicted on this x-ray.

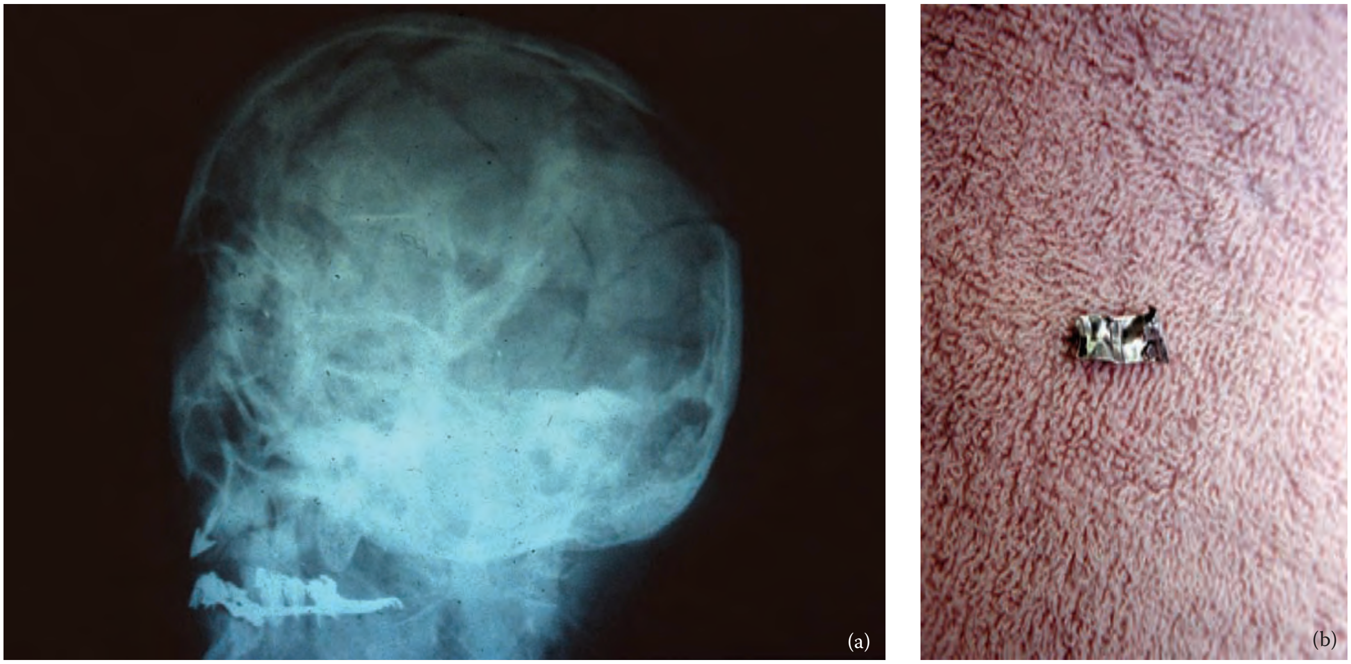

This individual (a) was shot with an aluminum-jacketed bullet. This fragment of aluminum (b) was retrieved within the wound track. Aluminum, being a relatively less dense metal, may not be apparent on x-ray, particularly when it is lying over a dense thick bone.

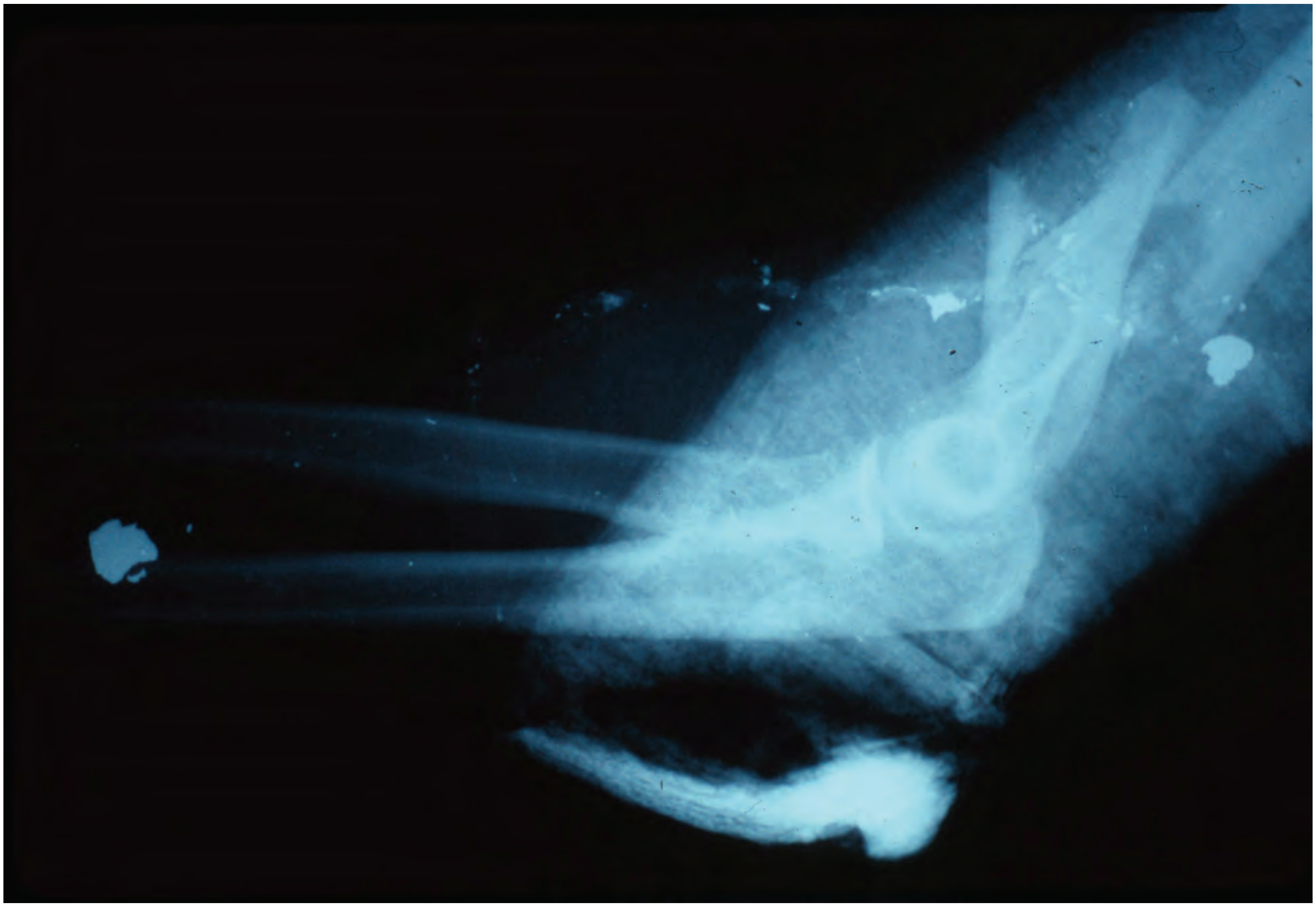

Note from this x-ray that one can tell the direction of travel. The bullet struck the humerus, fragmented into pieces, and then left a trail of metal fragments as it passed through the soft tissues before coming to rest.

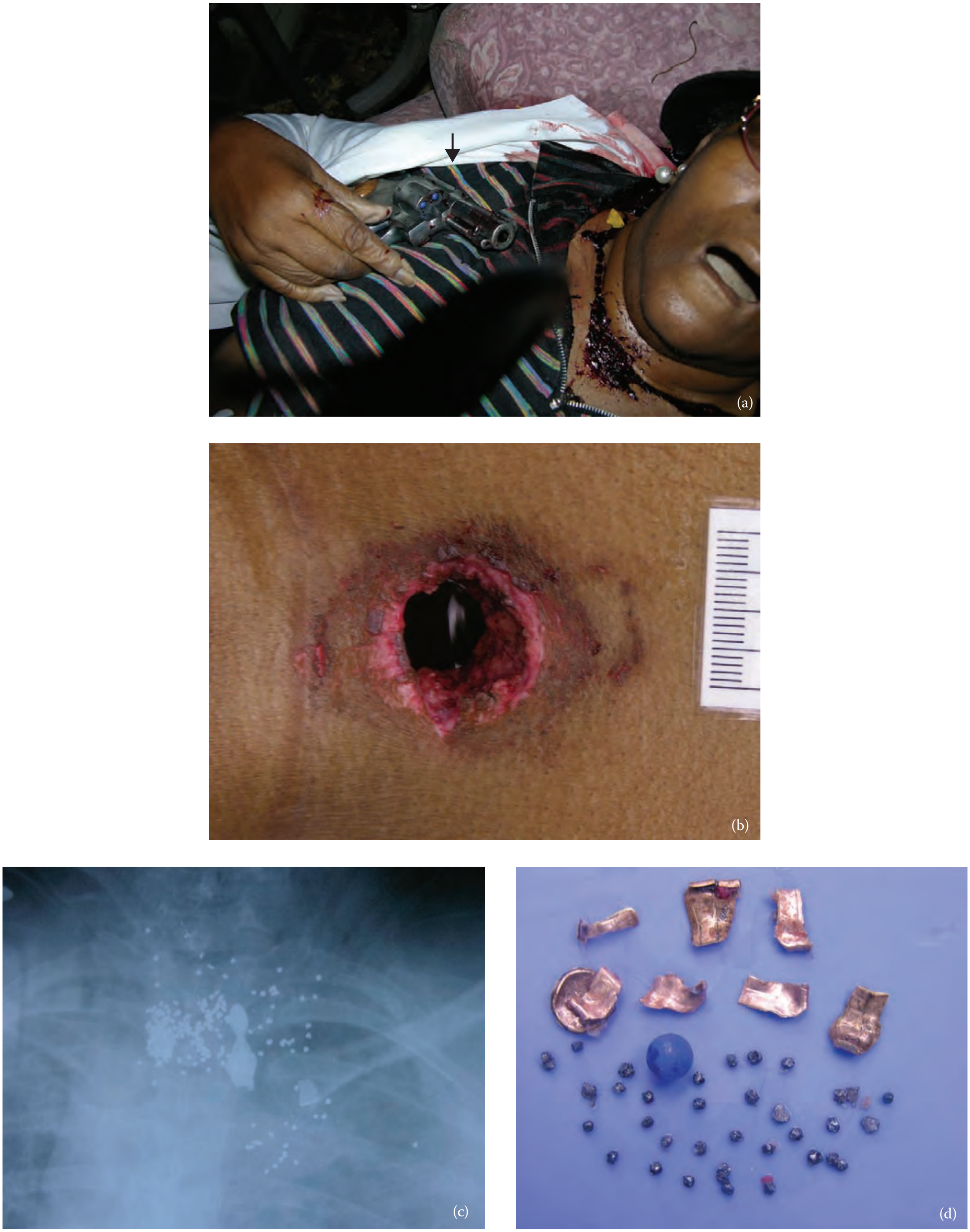

This is a contact gunshot wound to the chest with visible muzzle imprint abrasion (b). Note the blue-tipped Teflon plug Glazer safety ammunition visible within the revolver chamber (a). Note the yellow metal jacket, Teflon plug, and multiple lead pellets seen on x-ray (c), and demonstrated after removal from the body (d).

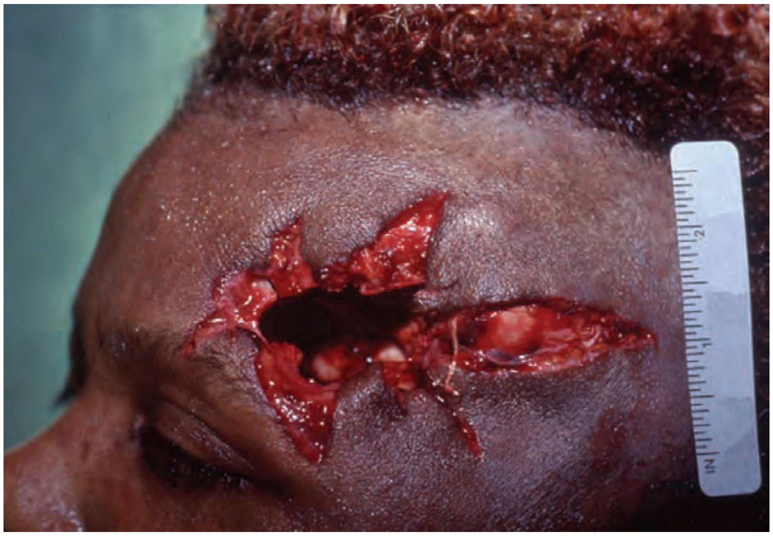

Tangential distant gunshot wound produced by Mag-safe ammunition (prefragmented).

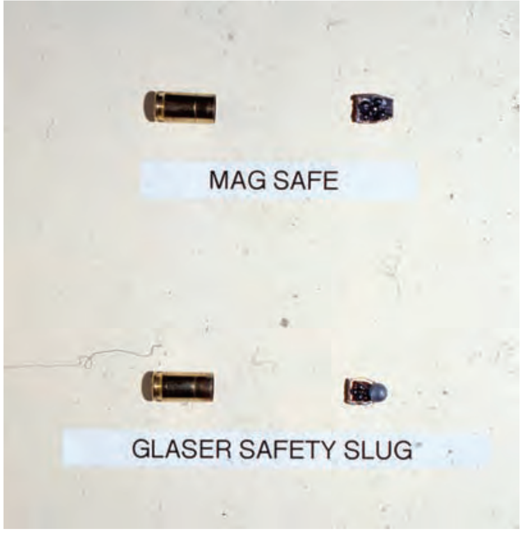

This diagram shows a comparison of the components of Mag-safe ammunition and Glaser safety slug.

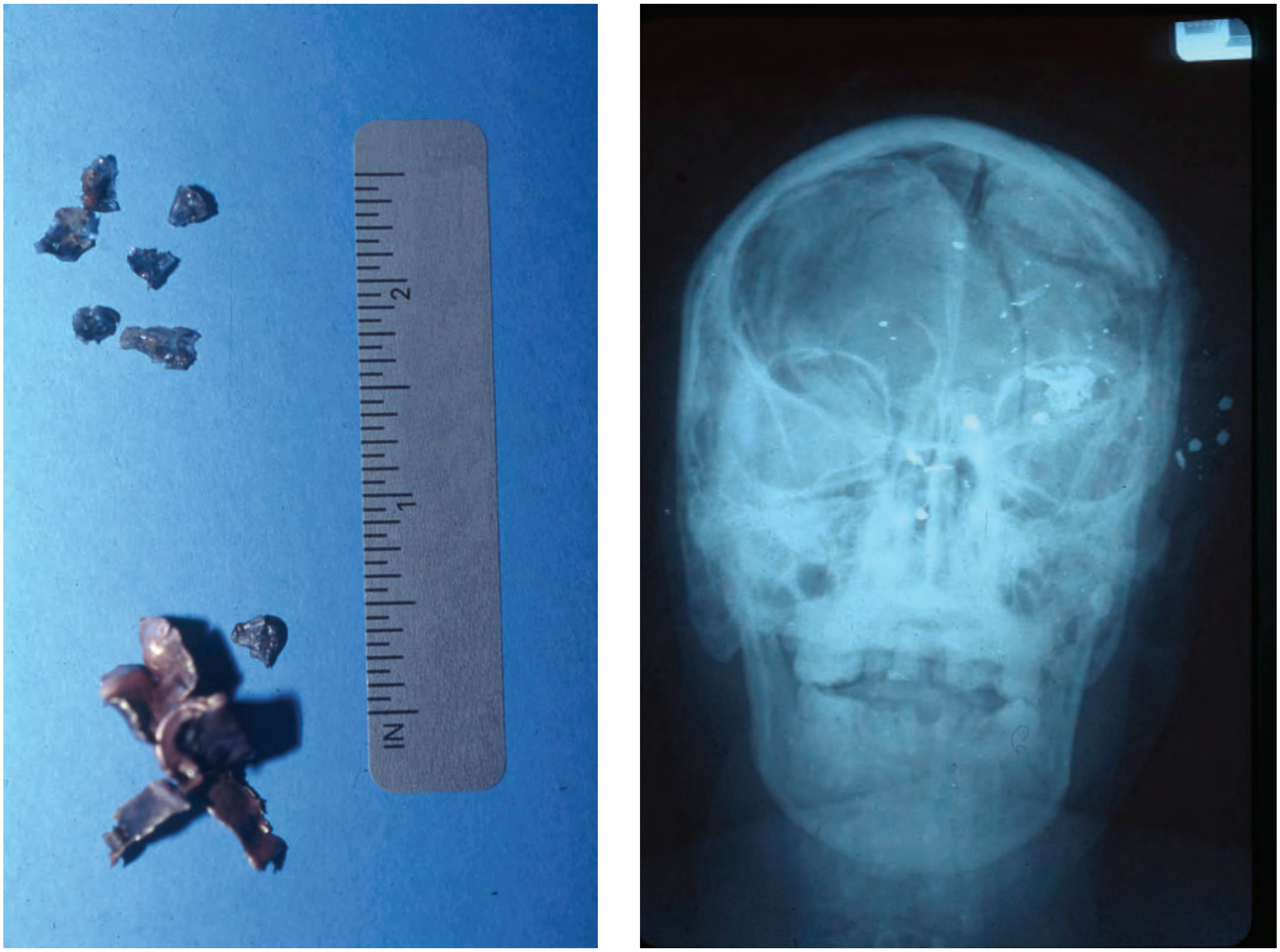

This x-ray is taken from an individual shot with Mag-safe ammunition. Note the metal jacket, gray to blue epoxy fragments, and gray metal pellets.

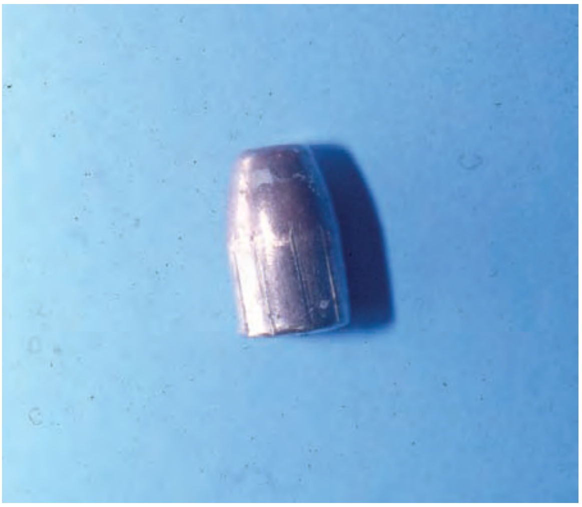

This demonstrates a bullet retrieved from an individual who was shot approximately 1 day prior to autopsy. Note the shiny surface of the bullet, signifying no significant oxidation.

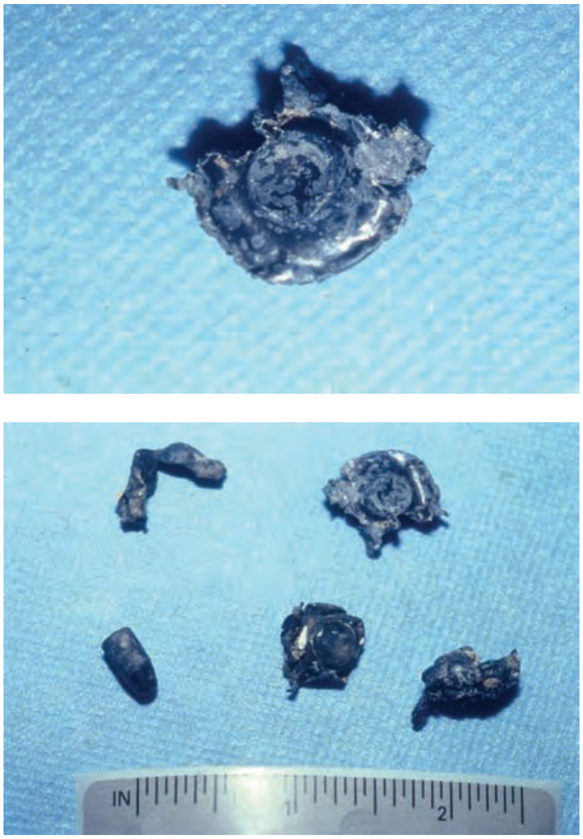

These are multiple weathered fragments of bullet retrieved from underneath and within the body of a partially skeletonized individual who was shot and thrown down a well shaft several years earlier. Note the irregular weathering marks at the surface of the bullets due to erosion.

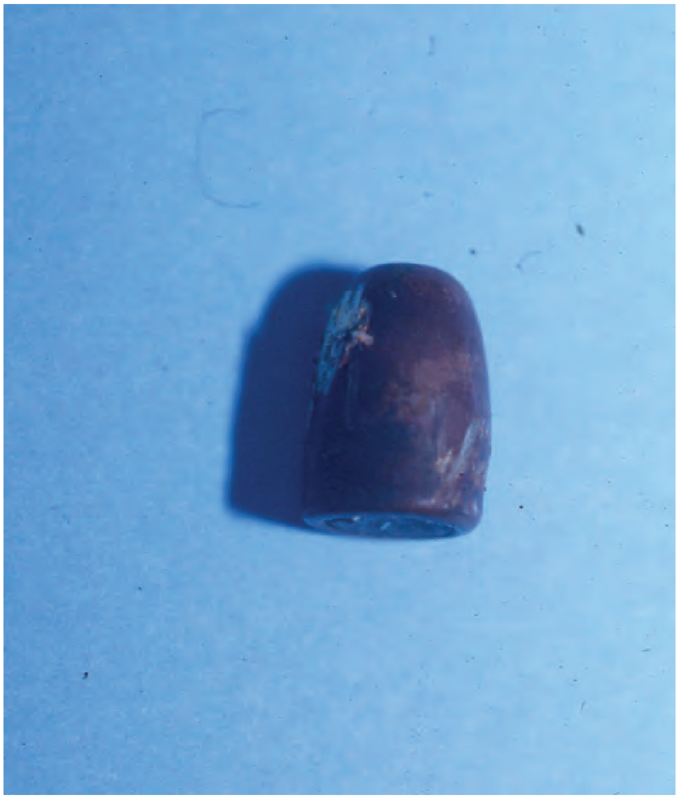

This demonstrates a bullet retrieved from an individual who was shot over a year before. This bullet remained in his body until it was retrieved during autopsy after somebody else shot and killed him. Note the dull oxidized surface on this old bullet.

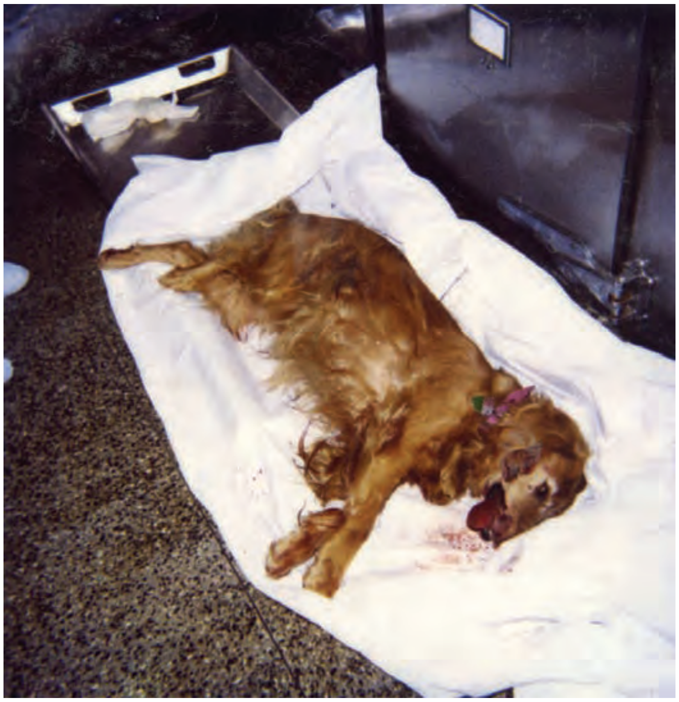

The homicide victim who owned this dog died of multiple perforating gunshot wounds. The dog sustained penetrating gunshot wounds during the attack and was autopsied to retrieve ballistic evidence.

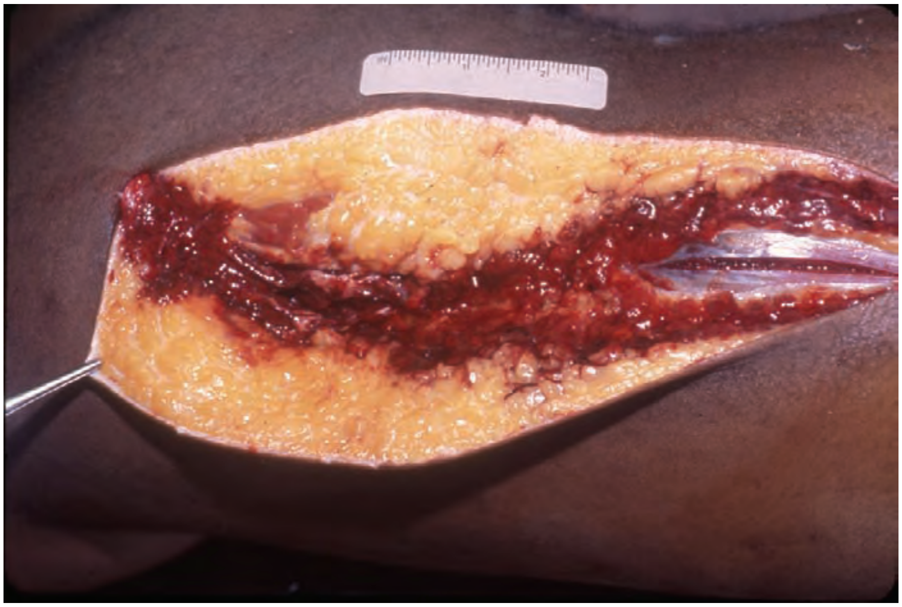

This is a recent gunshot wound to an individual who was shot through the subcutaneous and fatty tissues of the gluteal region and thigh. Note the curvature of the wound track. This demonstrates how wound tracks may change when the body is laid on a flat autopsy table. One must keep this in mind when formulating bullet trajectories with reference to standard anatomic planes. This information may later be used as a reference to help explain possible body positions during the actual shooting. Interpreting wound track directions can become complicated when there are multiple gunshot wounds in proximity, particularly when the individual was shot while curled up in a fetal position and then later examined while spread out on an autopsy table. Directions should be stated with reference to standard anatomic planes.

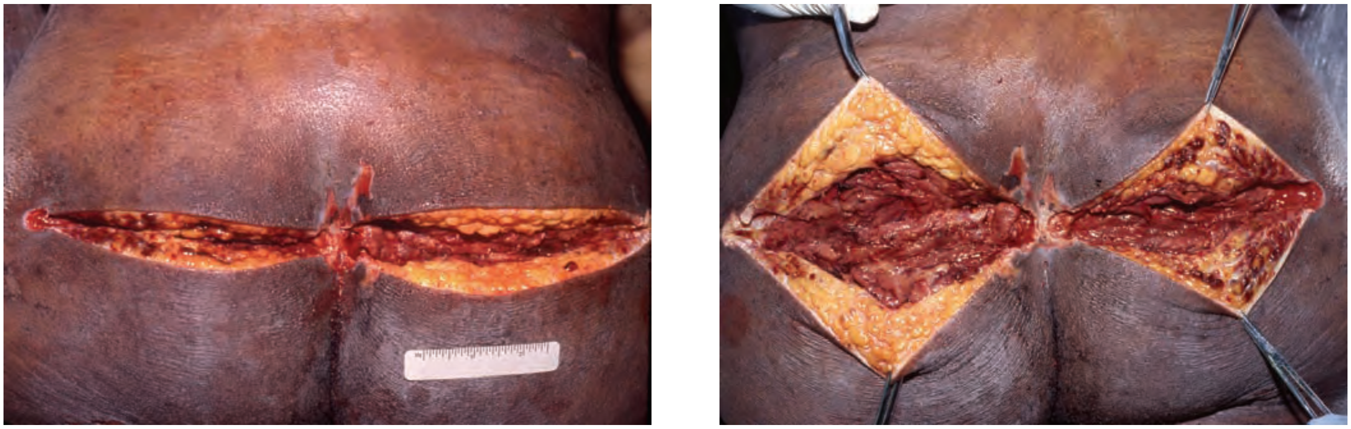

This is a healing wound track through the subcutaneous and fatty tissues of the gluteal region. Note the dull granular nature due to healing and the presence of granulation tissue.

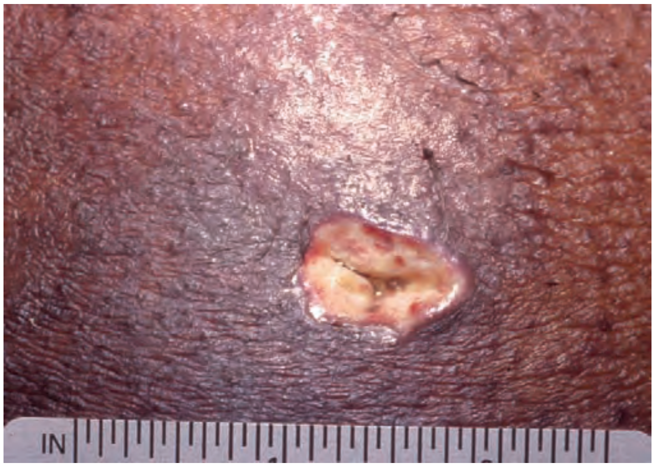

This is a healing gunshot wound that is approximately a week old.

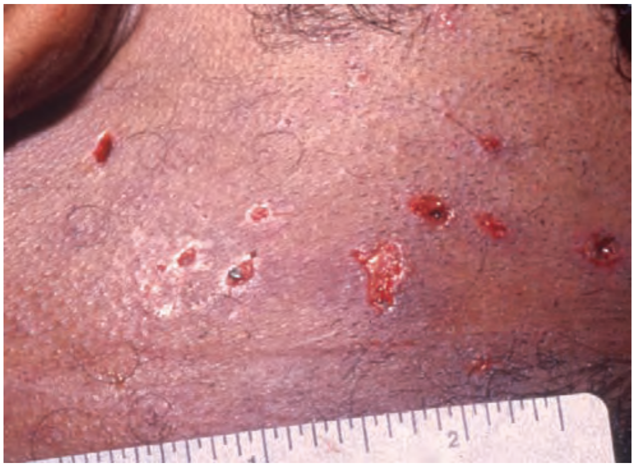

Atypical healing gunshot wounds with fragmented pieces of lead.

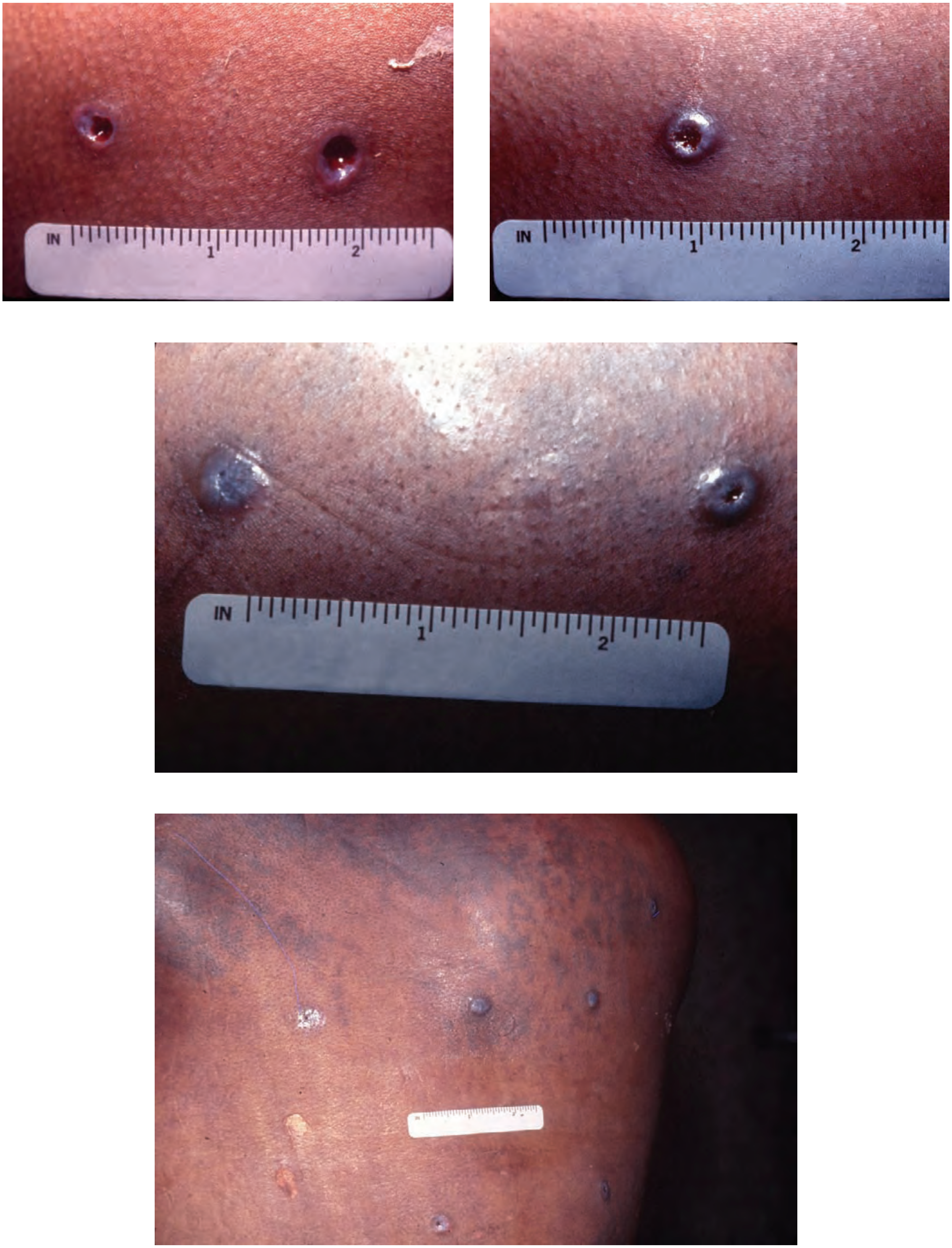

Varying stages of healing gunshot wounds.

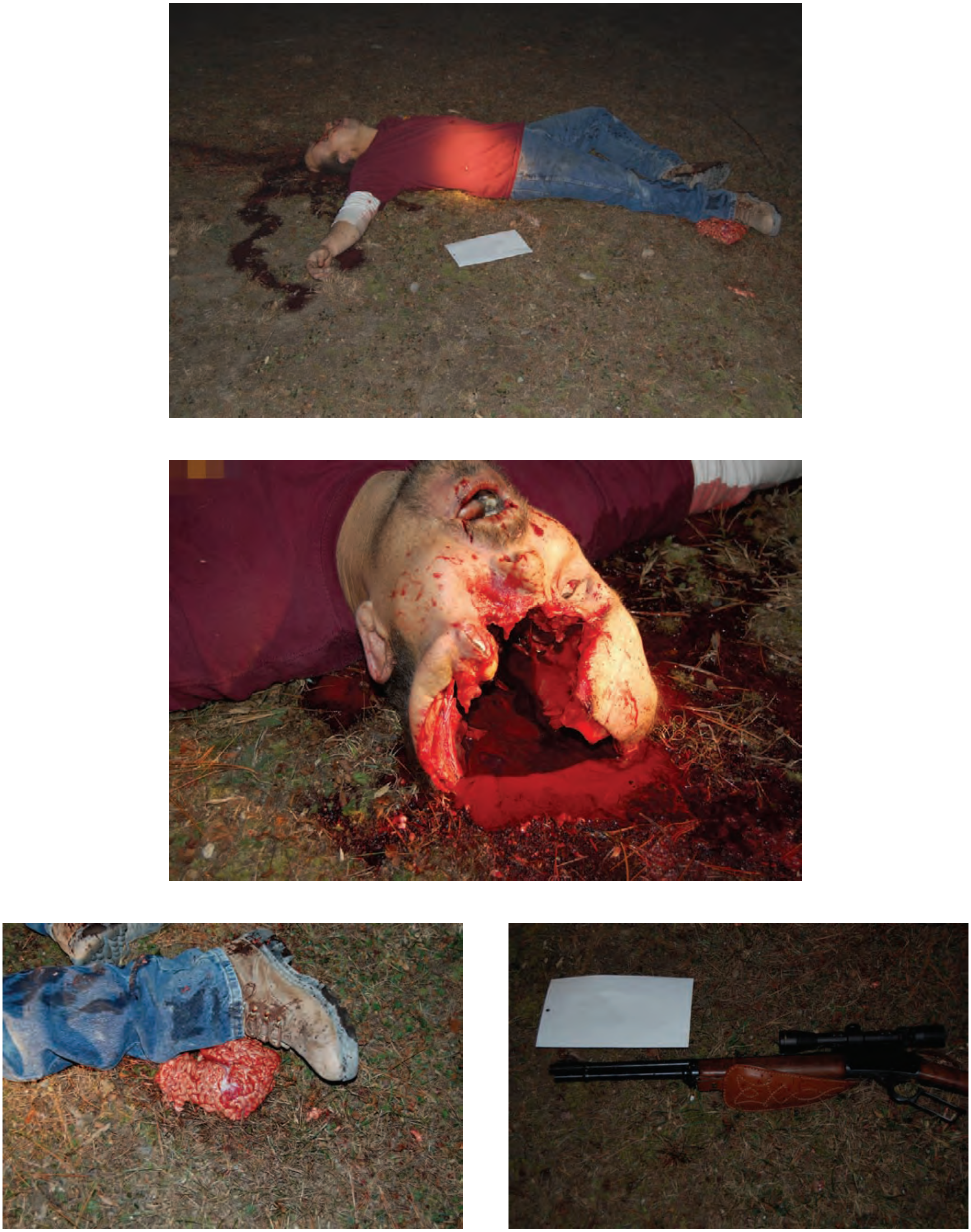

This case involves a suicidal tight-contact intraoral .44-caliber rifle wound with extensive fractures and lacerations to the head. Note the gaping injury at the top of the head with the empty cranial vault. The brain was ejected almost entirely intact due to expanding muzzle gases.

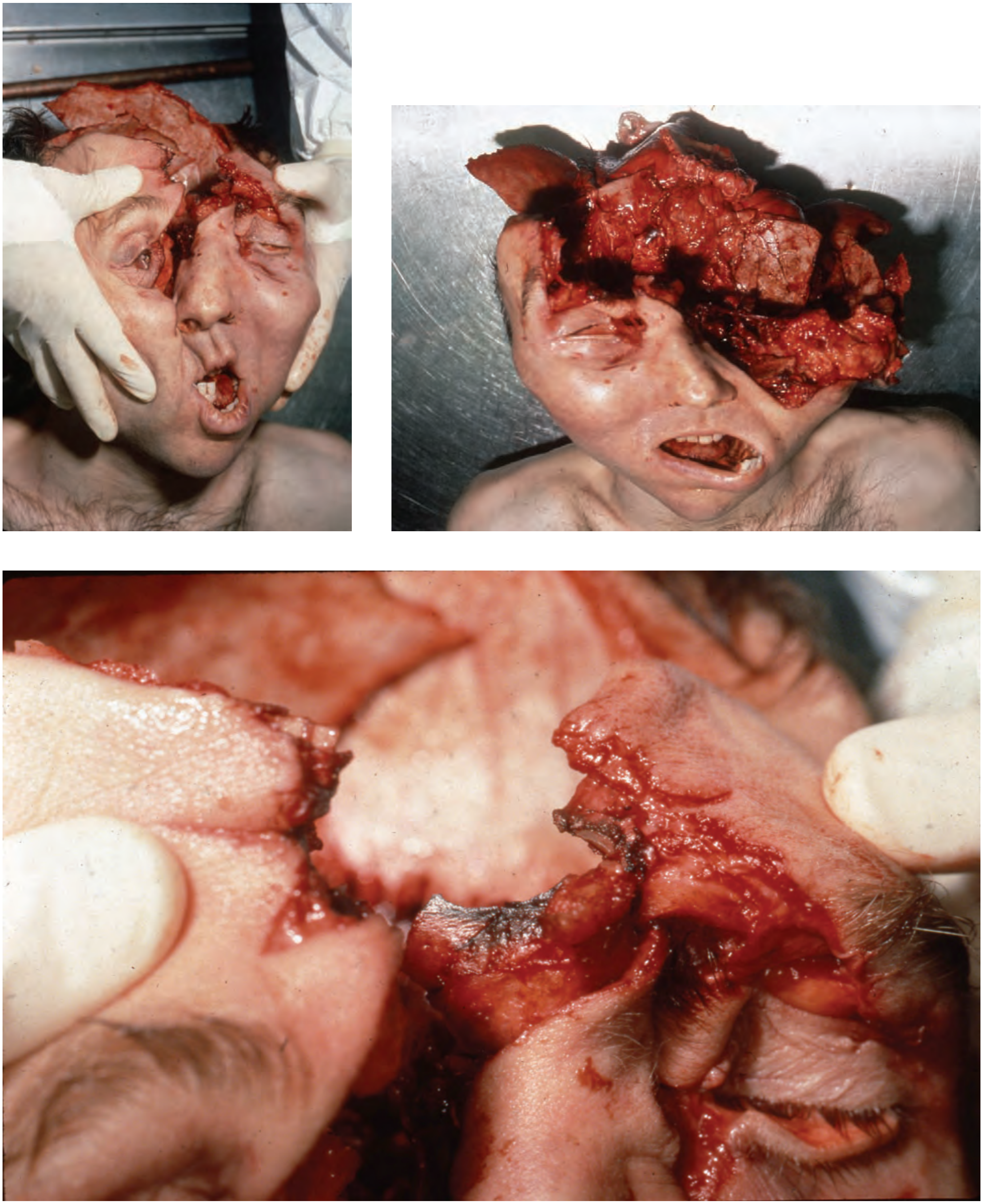

This individual died from a self-inflicted contact 16-gauge buckshot shotgun wound to his forehead. His cerebral hemispheres were ejected from the cranium almost completely intact and were discovered behind the body. Note reconstruction of this wound produces an obvious circular perforation with soot at the individual's forehead. Initially, at the scene investigation, it was thought to have been an intraoral shotgun wound, many of which are associated with stretching and lacerations of the lips and mouth region.

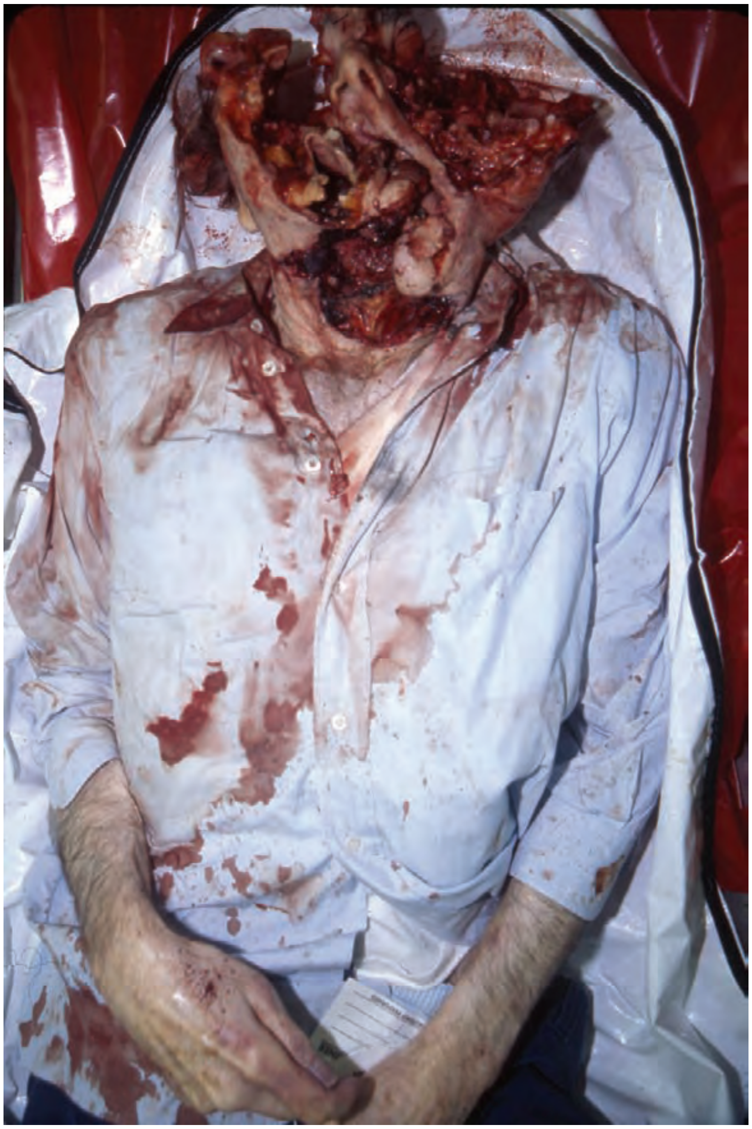

This individual died of a self-inflicted, contact, 12-gauge shotgun wound to the superior aspect of his neck underneath his chin. Note approximation of the wound margins produces a roughly circular perforation with soot. There were extensive lacerations to the face and head. Note the extensive stretch marks at the decedent's face. Note the blood spatter pattern and gunpowder residue on the hand.

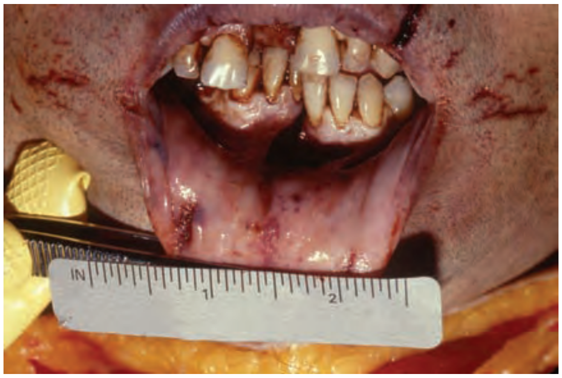

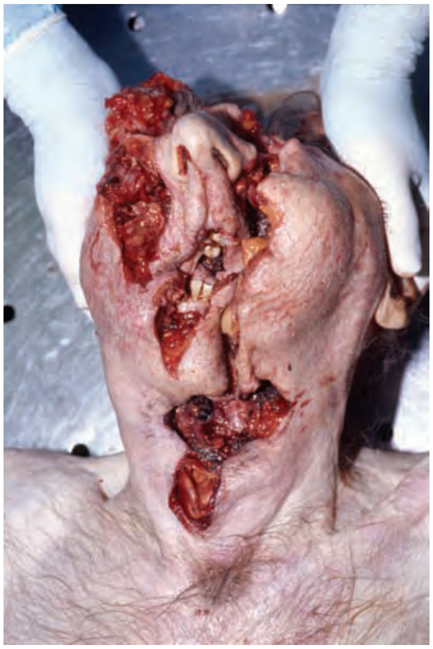

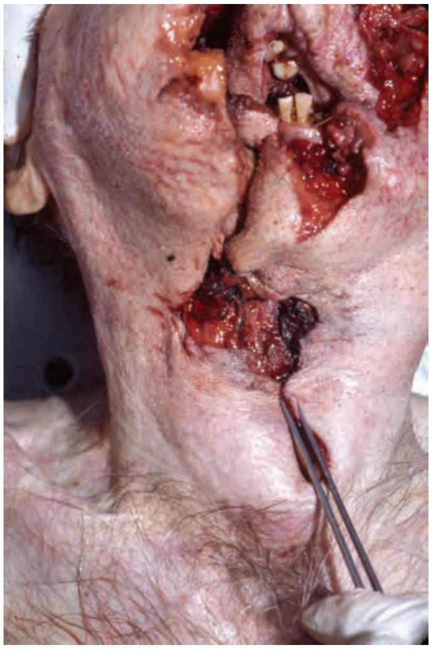

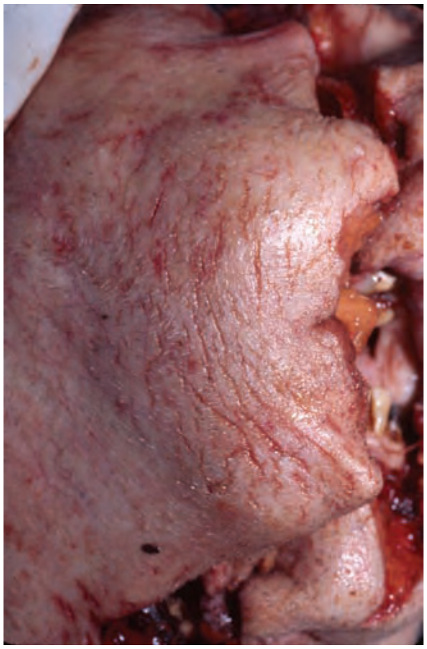

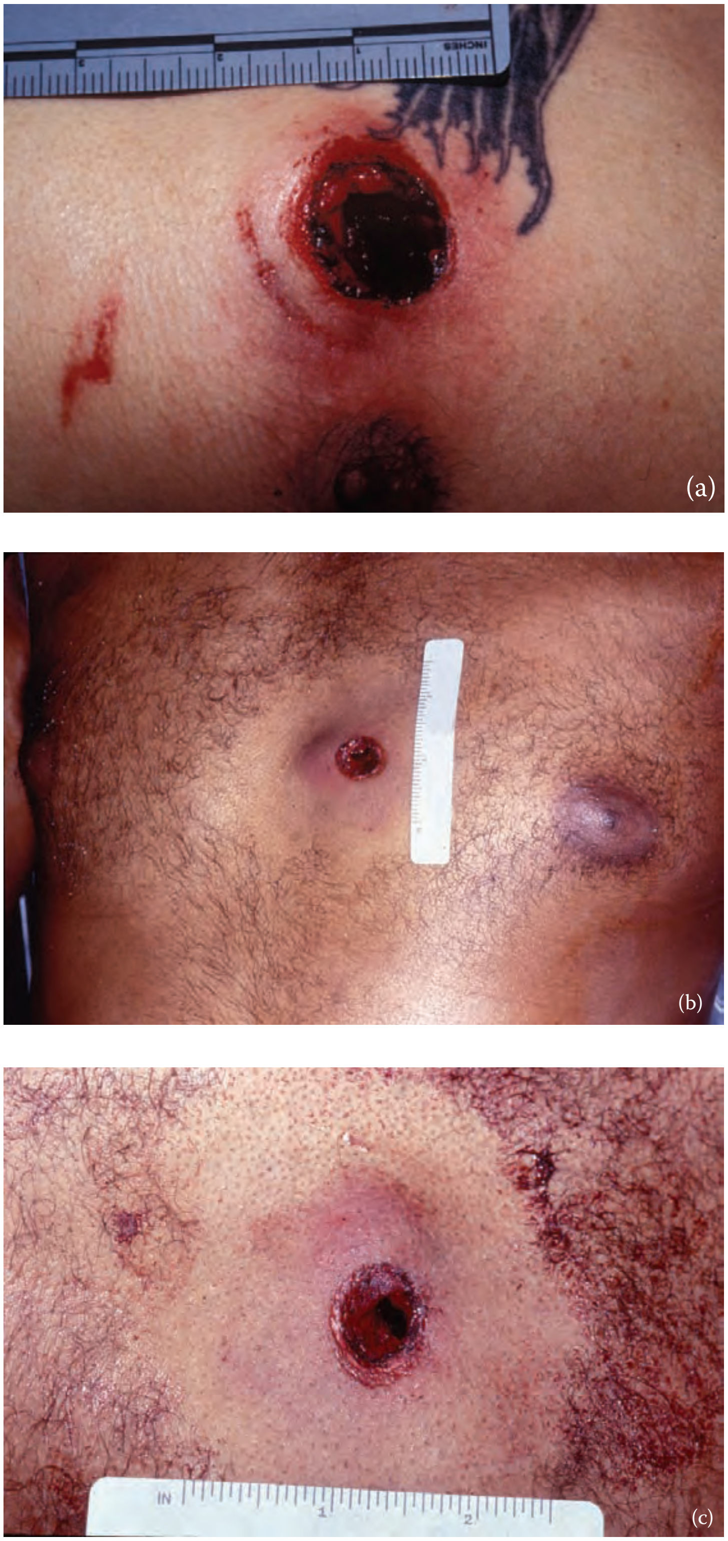

These are contact shotgun wounds to the chest. Note the roughly symmetric margin of abrasion with small amounts of soot. More soot was observed within the wound track. Note in (a) another encircling abraded ring corresponding to where the muzzle contacted the body when the gun was discharged. Note in (b–c) there is an encircling vague pink discoloration from carbon monoxide and nitrites in the burning gunpowder that reacted with the underlying muscle and blood, producing this red discoloration.

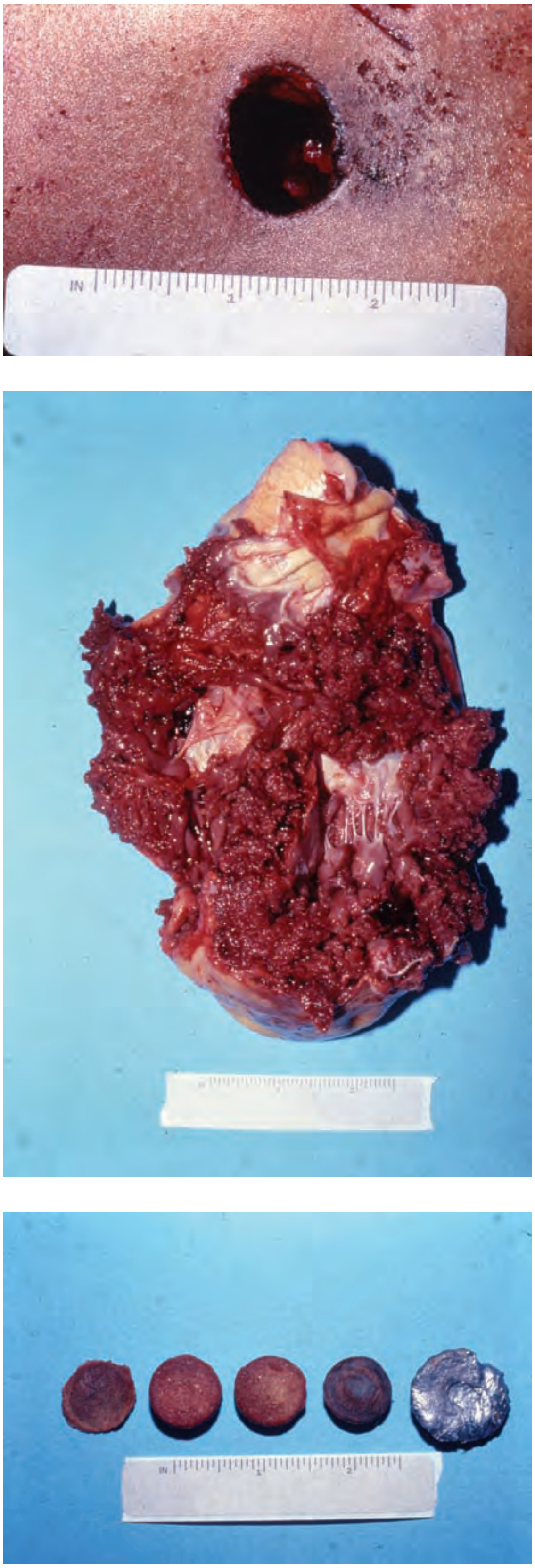

This is another example of a shotgun-slug wound. Note the extensive fragmentation and laceration to the heart. The last image shows the deformed lead shotgun slug with wadding retrieved from the wound track.

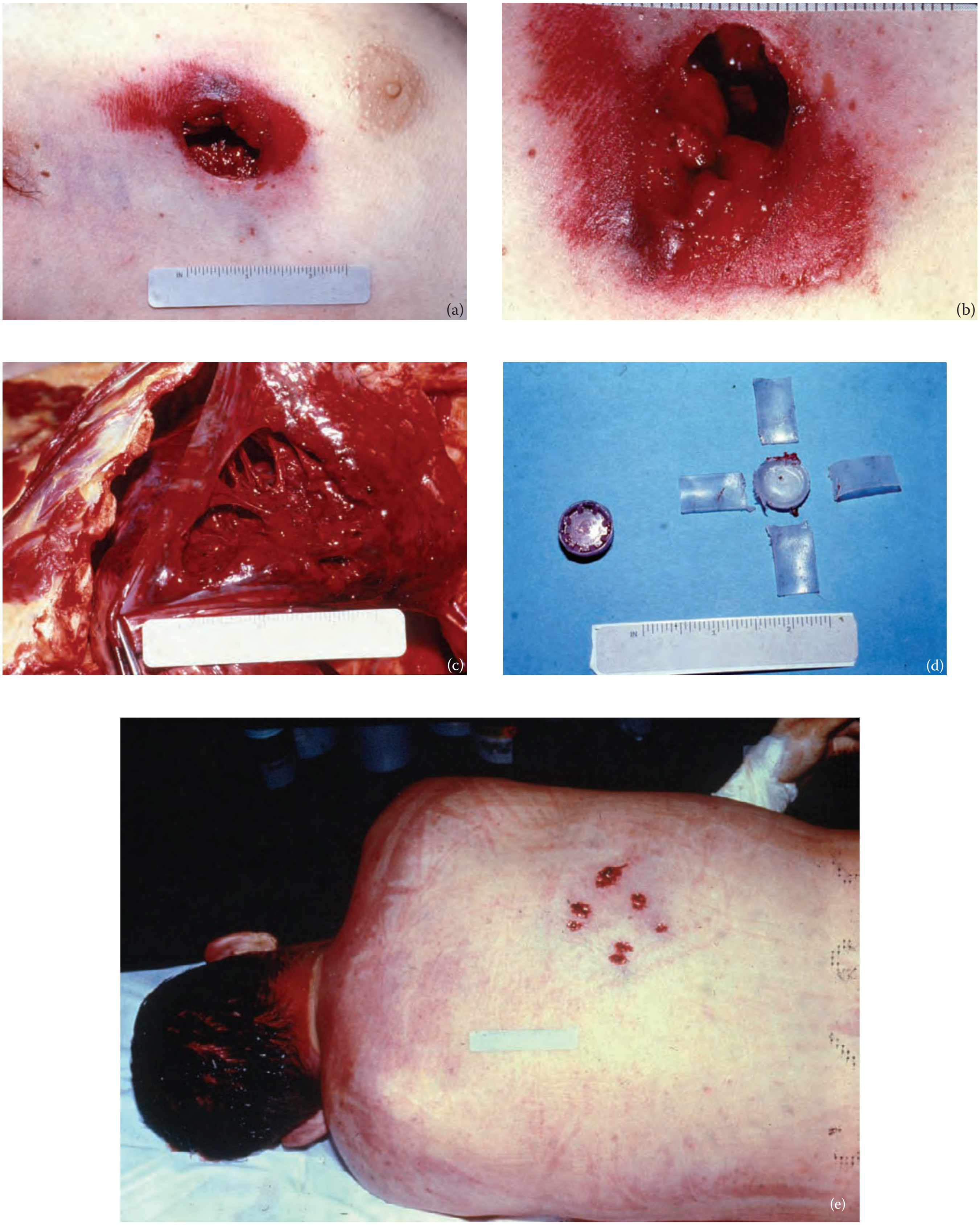

This is a .12-gauge shotgun wound through clothing showing few stippling defects and a larger abrasion adjacent to the entrance defect (a and b). The power piston was retrieved within the wound track (d). Note the abrasions adjacent to the entrance wound due to power piston impact (b). There was extensive fragmentation of the lung (c). Note the multiple exit buckshot-pellet wounds to the decedent's back (e).

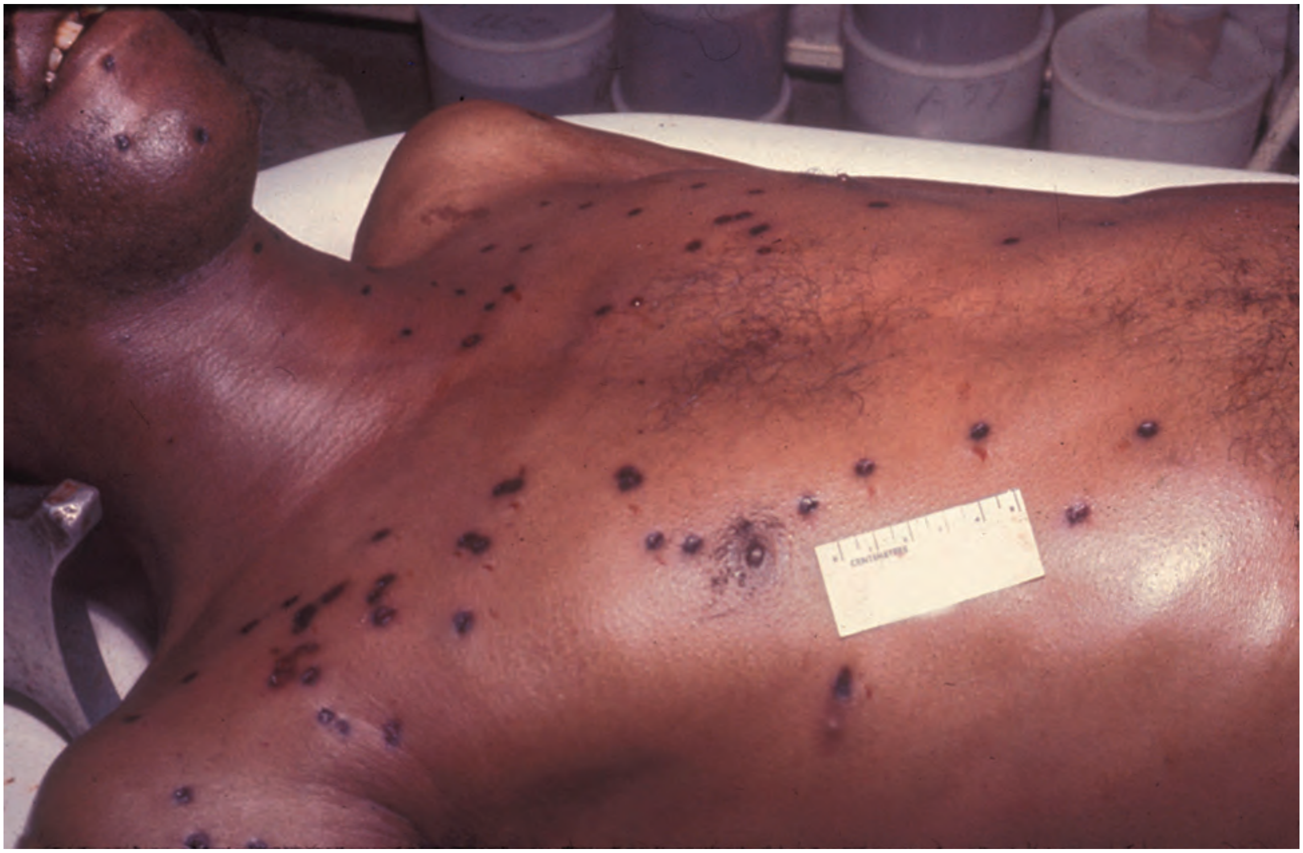
CENTIMETERS

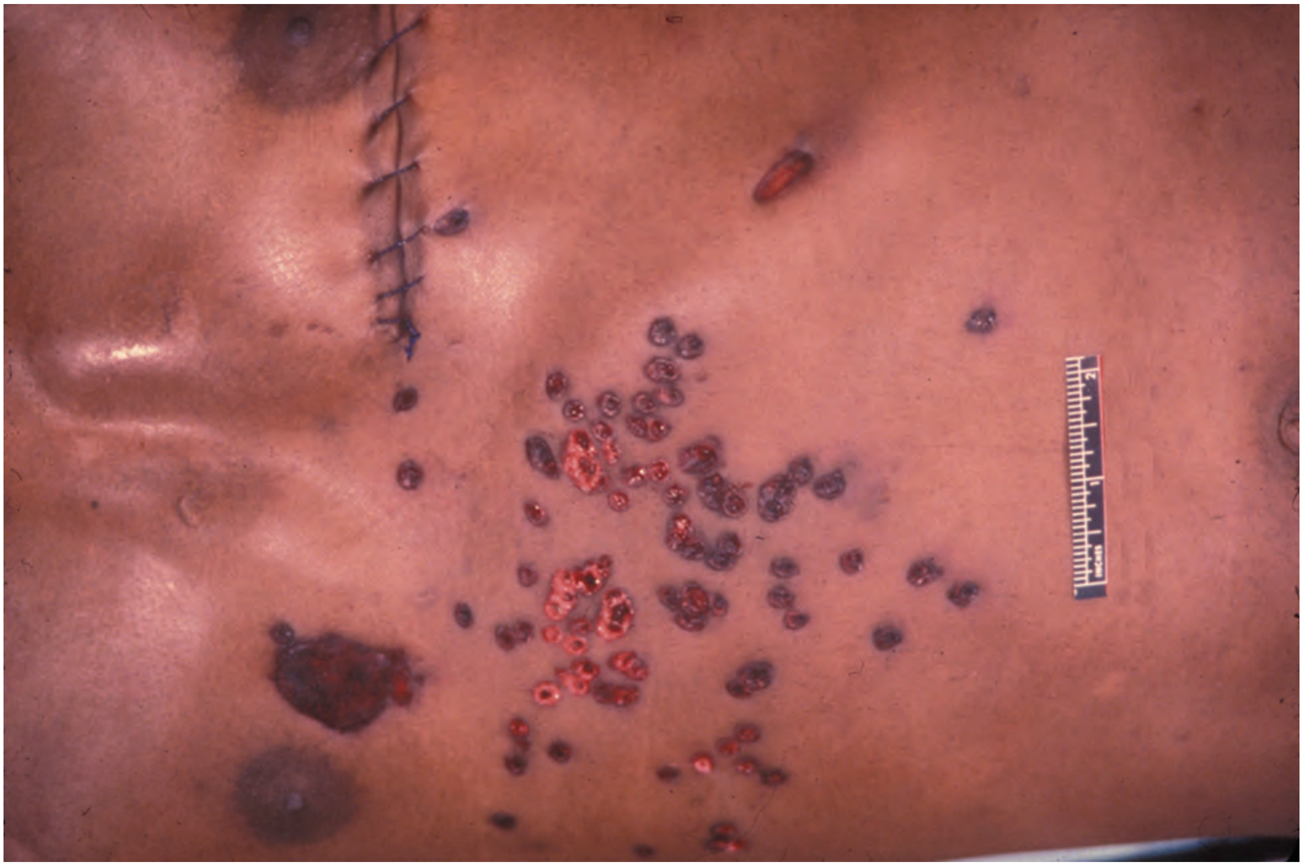

Distant birdshot-shotgun wounds.

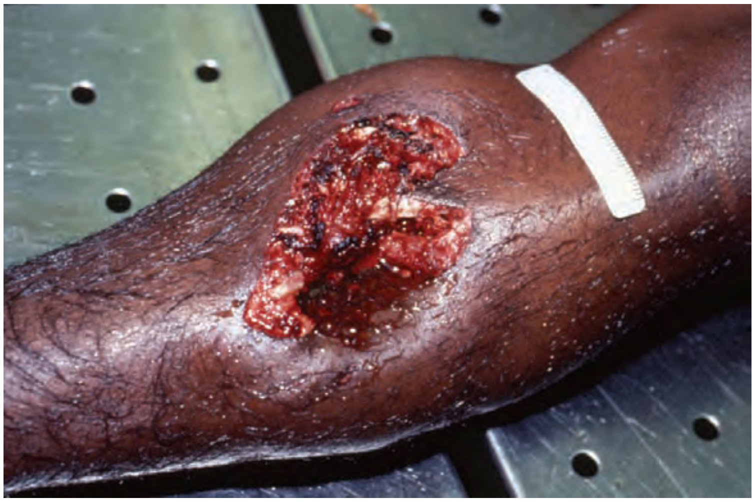

This birdshot-shotgun wound to the forearm is estimated at a range of approximately 3 to 4 feet.

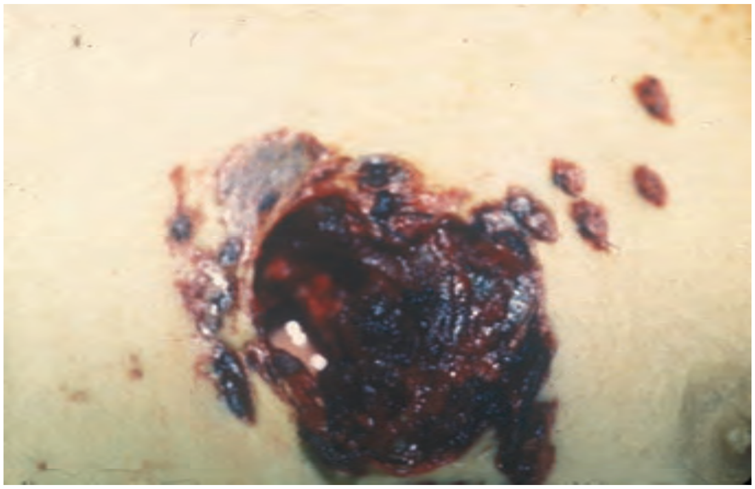

Birdshot shotgun wound with an estimated range of fire of 4 to 5 feet.

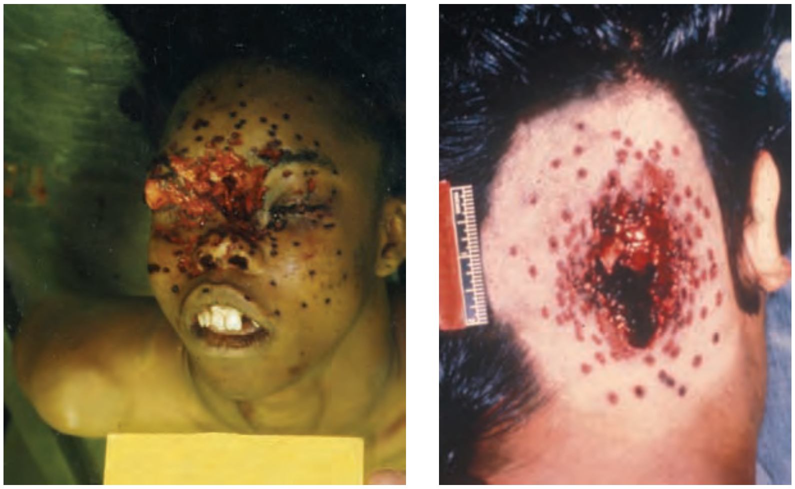

Birdshot shotgun wound with an estimated range of fire of 6 to 8 feet.

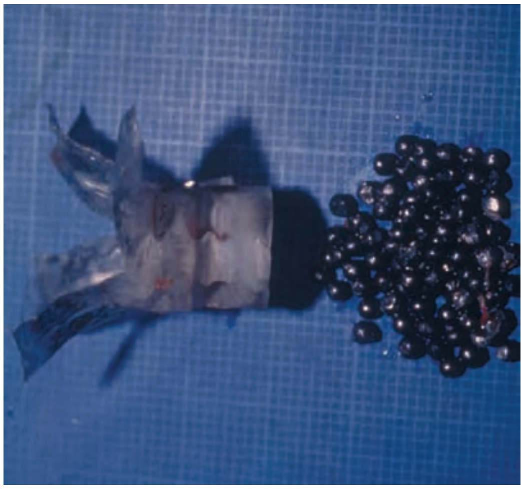

Birdshot with power piston type cup from a shotgun shell.

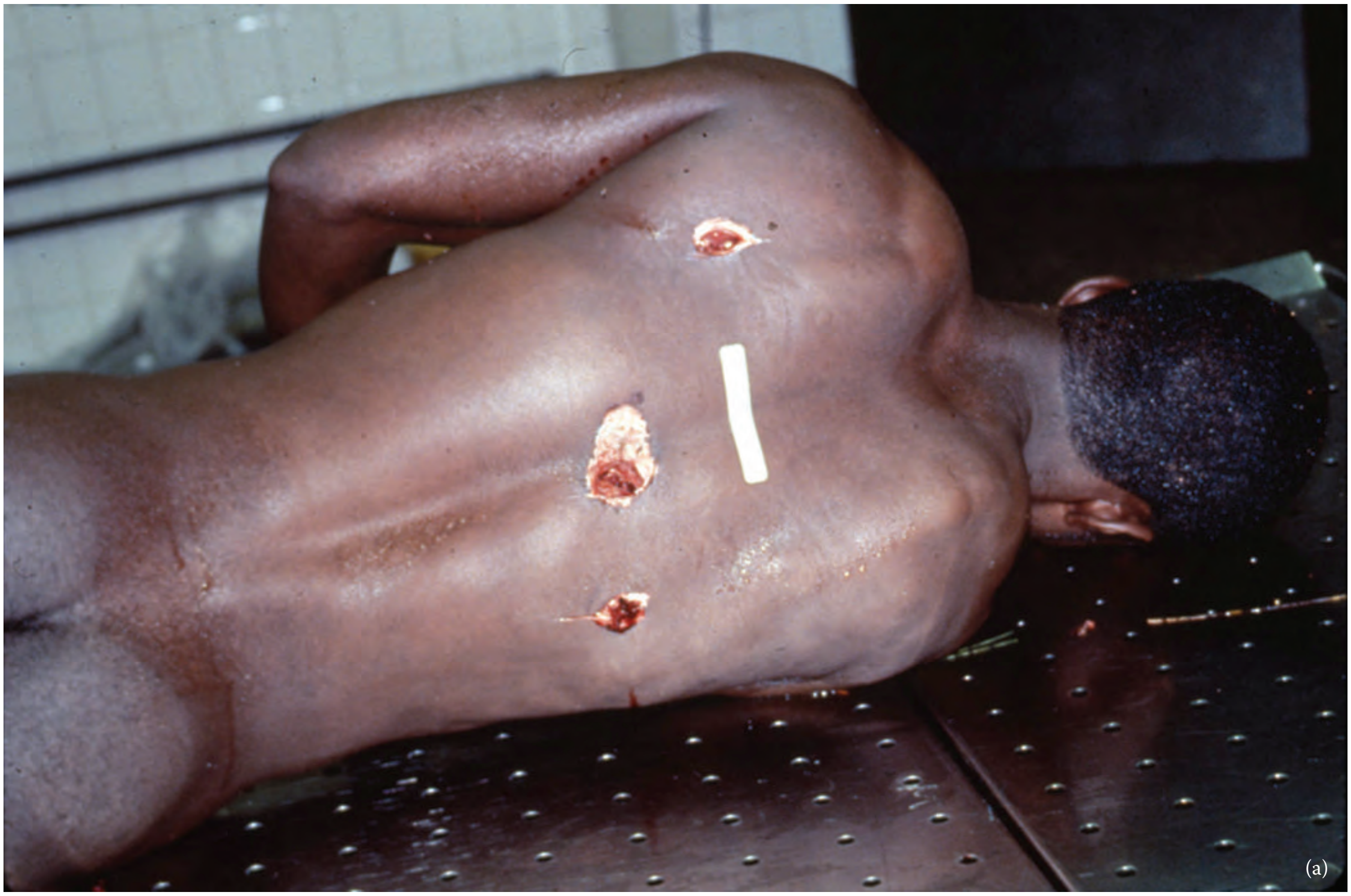

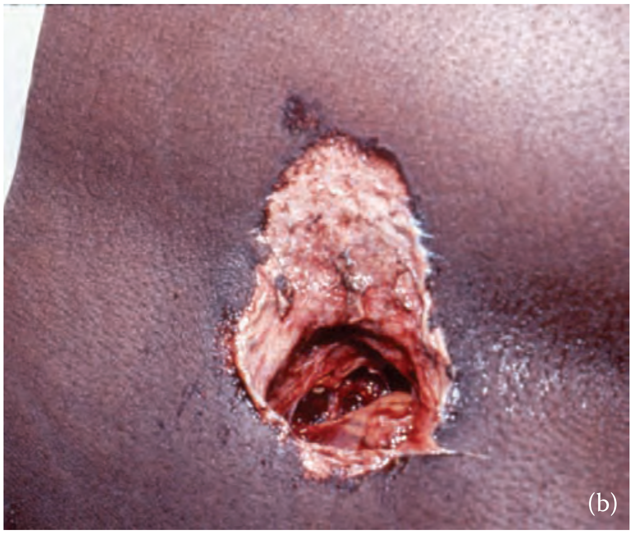

This individual was shot multiple times with shotgun slugs. Note the large circular perforations with margins of abrasions typical for entrance handgun wounds (a). Note the elongated margin of abrasion at the middle wound (b), indicating the slug struck the body on an angle, traveling in a left-to-right direction.

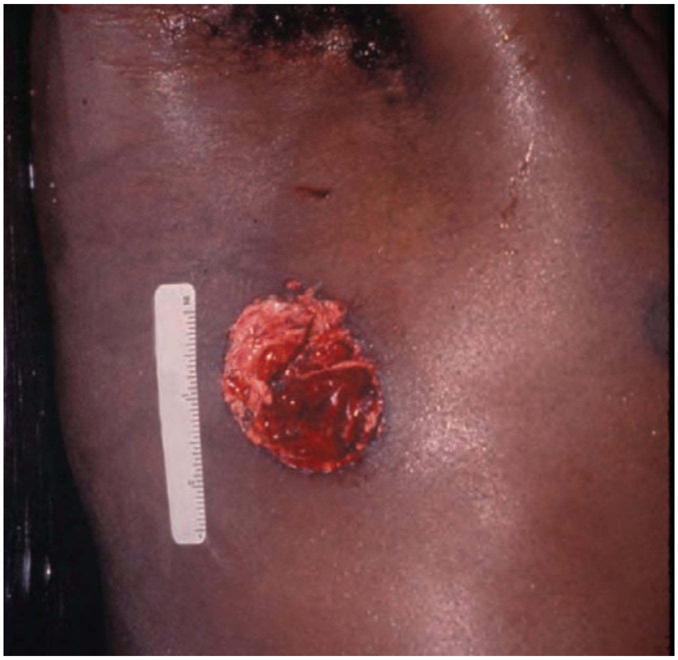

Entrance shotgun wound with an estimated range of fire of 3 to 4 feet based on the scalloped appearance of the wound margins.

Mini-14 Ruger rifle.

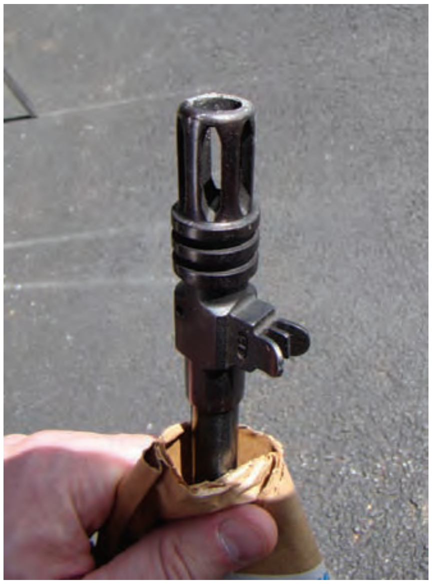

Flash suppressor.

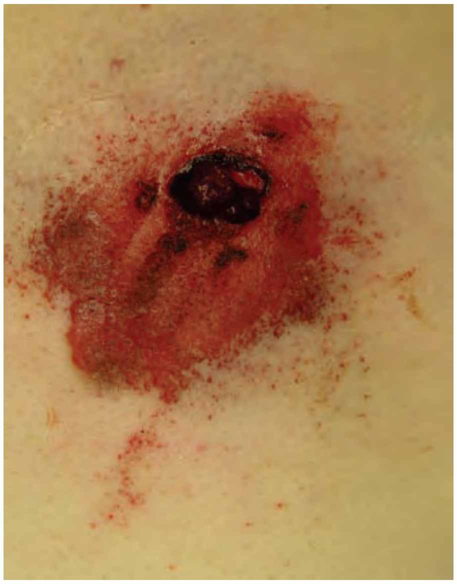

Contact Ruger mini-14 gunshot wound with a .223 bullet. Note the flash suppressor burns with soot deposition.

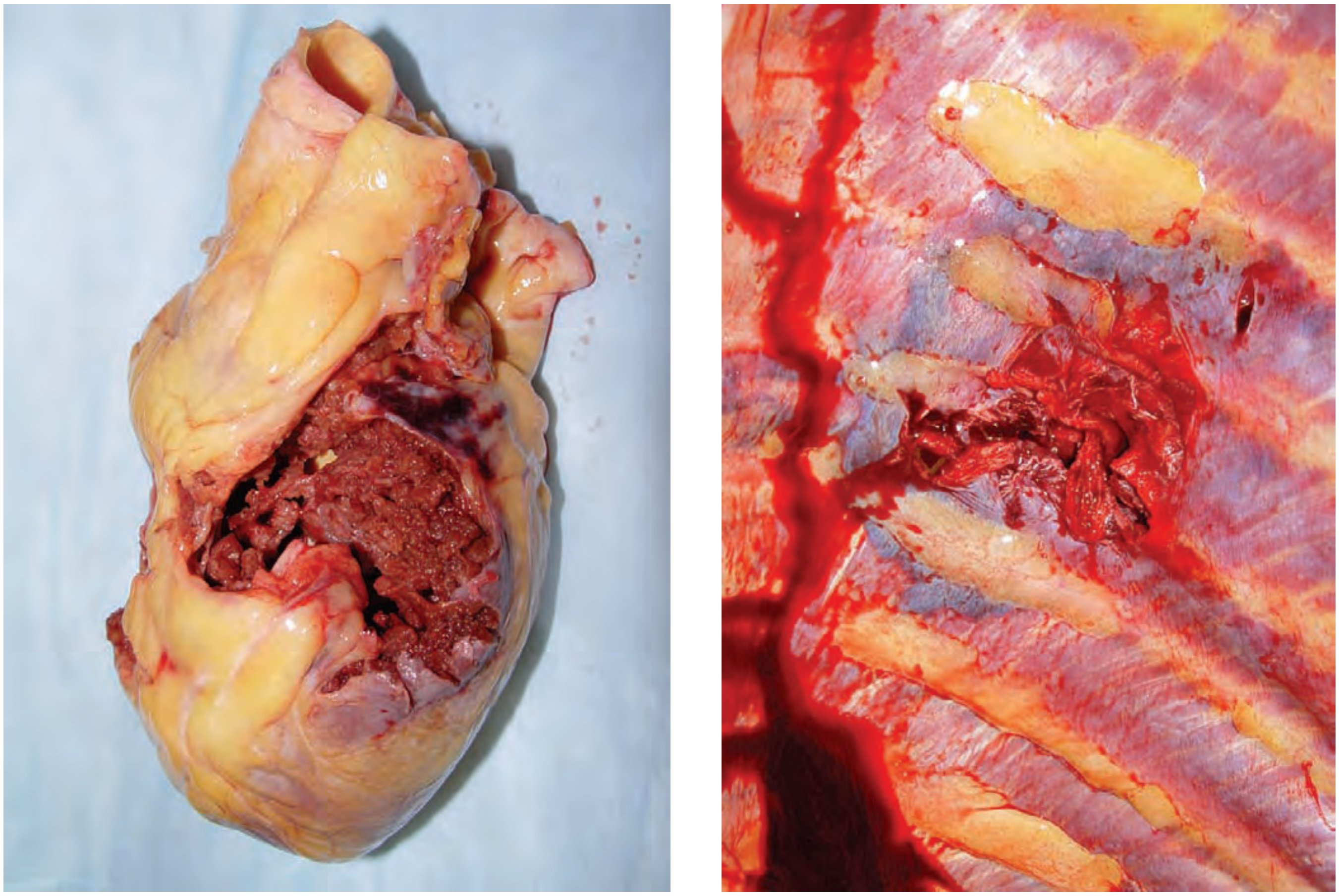

Internal injuries caused by a .223-caliber high-velocity military bullet.

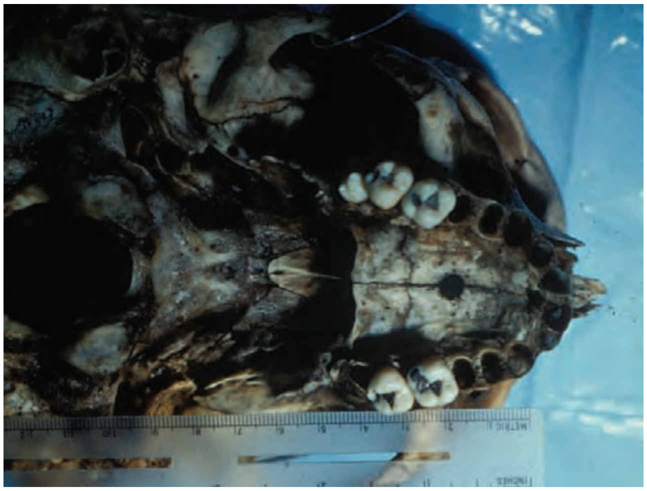

Entrance gunshot wound to the roof of the mouth in the skeletonized remains of this individual with a history of depression. The remains were found with a handgun.

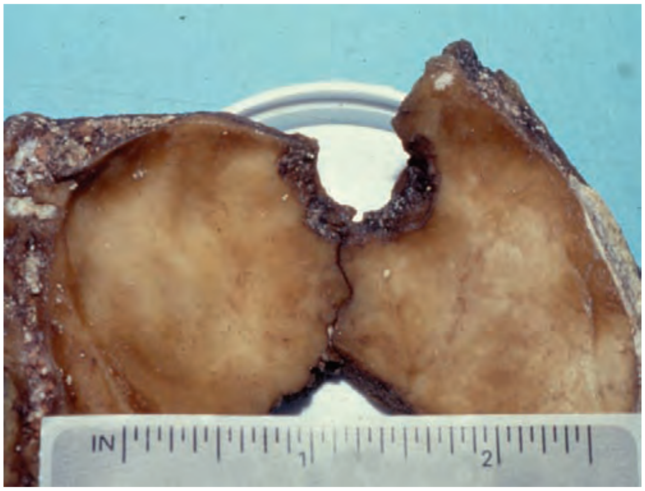

This is the inner aspect of a fractured portion of skull with a roughly circular perforation and internal beveling visible at the inner aspect of the skull indicating this to be an entrance gunshot wound.

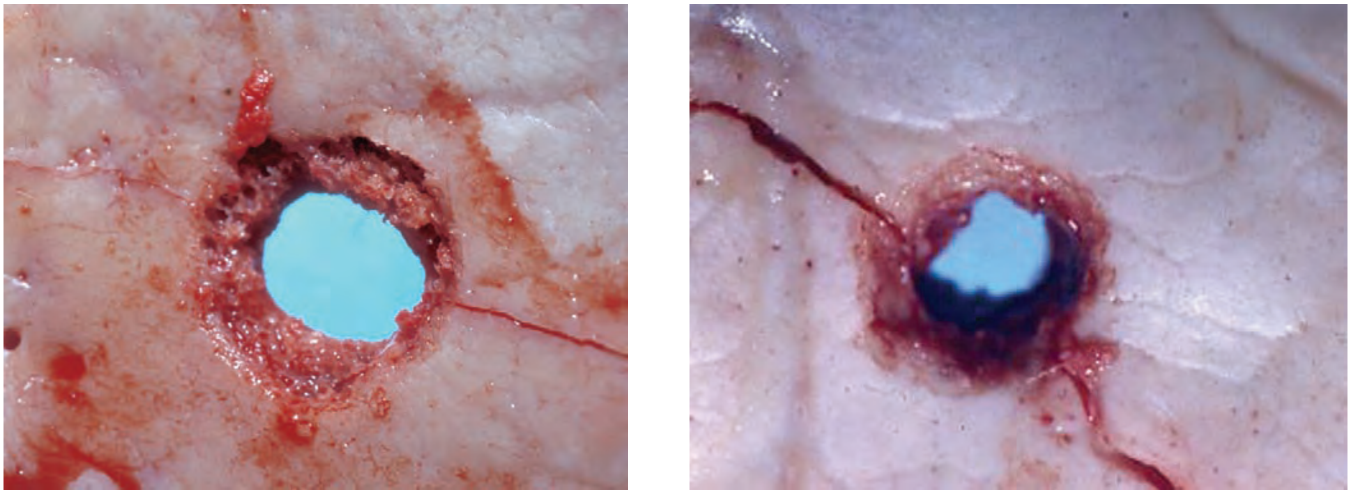

Entrance gunshot wounds to the head demonstrating internal beveling of the inner calvaria.

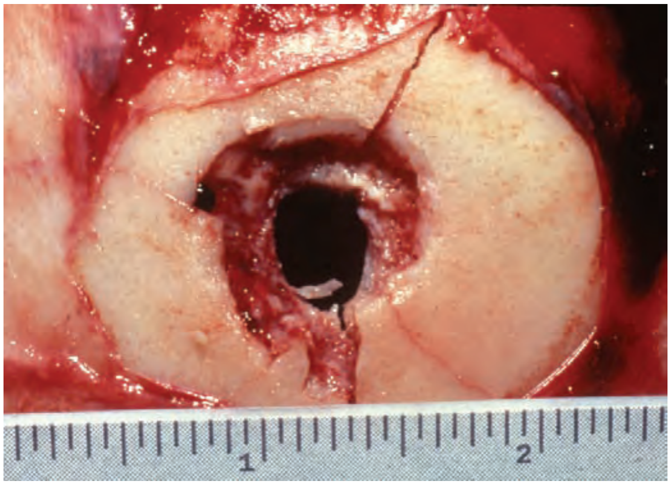

Exit gunshot wound through the skull. This image demonstrates the surface of the skull with external beveling typical for an exit gunshot wound.

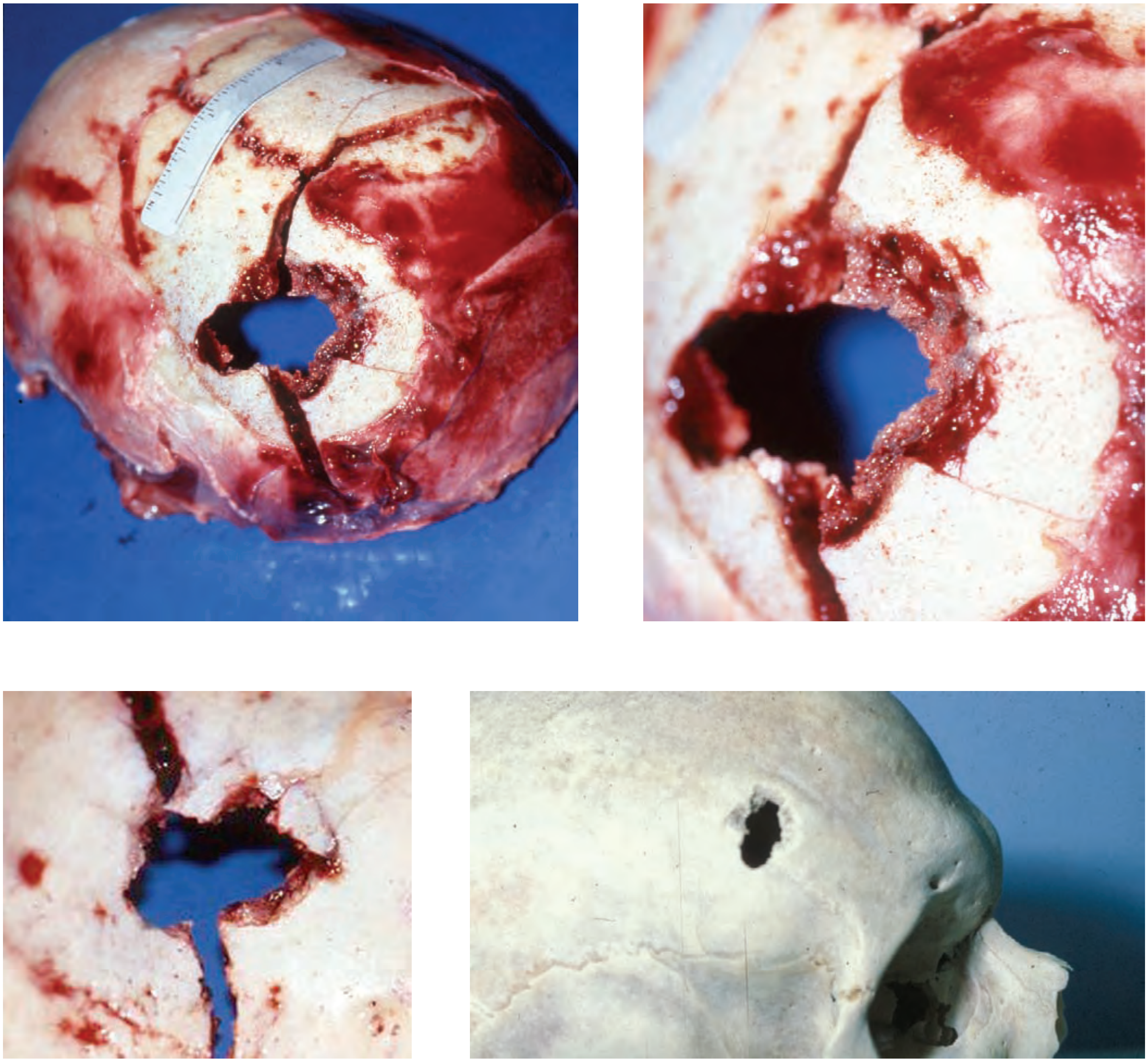

These defects at the external surface of these skulls were produced by an entrance tangential gunshot wound. This demonstrates a keyhole deformity with both internal and external beveling. The bullet typically strikes the bone tangentially, producing internal beveling at the entrance side. Then it will often fragment and partially exit the body. The exit side will have external beveling.

Burns

9

CHARLES A. CATANESE AND GERARD CATANESE

Introduction

Burns may occur following exposure to heat (thermal burns), electricity, chemicals, or radiation.

First-degree burns are the most superficial and involve the epidermis. They appear as a red discoloration of the skin. An example of this would be sunburn without blister formation. *Second-degree burns* have deeper penetration of injury involving the epidermis and dermis causing blister formation. An example of this may occur when one touches a hot pot on a stove and pulls the hand away quickly. *Third-degree burns* are full thickness, involving the epidermis, dermis, and subcutaneous layer, and appear as collapsed blisters with skin sloughing and red-to-brown discoloration. *Fourth-degree burns* have even deeper penetration of damage, often with charring and exposed underlying tissue including bone. The bones may be fractured from intense heat and the internal organs may have a firm, discolored, and shrunken appearance.

Burns following exposure to heat are *thermal burns,* which may occur following contact with hot liquid or fire. Children may be scalded while left unattended in a bath tub. They may inadvertently hit the hot water knob to increase hot water flow. They may jump into a bath tub full of very hot water, scalding their feet. One must exercise great care in evaluating these cases as they may be a result of abuse or neglect. A child with both feet scalded with scarring around a shoe pattern is more likely the result of abuse or neglect. A child who is waiting to take a bath will not usually have shoes on and will most likely step one foot into the tub at a time. The burns will be on one foot but not usually both. Both feet scalded at once may indicate dunking into hot water as punishment.

If an individual dies and is placed in a tub full of hot water after death, the body will develop thermal injury more readily than if a living body was placed in the same water. A living body can counteract heat injury by vasodilation and circulating the heat to the body's core away from the surface, thus providing some protection to the skin. A dead body left in warm water will quickly develop postmortem thermal burns that appear as skin slippage. Decomposition may also present with skin slippage and blister formation. An individual may sustain full thickness burns to greater than half of his or her body and still be conscious without immediate death. Death often occurs later due to electrolyte imbalances or infection.

Also, mortality increases with age. Second- to third-degree thermal burns to half the body's surface would much more likely kill a senior citizen than a child. Second-degree burns are more painful than third- or fourth-degree burns due to less damage to nerve endings.

Antemortem burns may be characterized by fluid-filled-blister formation. To have fluid-filled blisters in a nondecomposing body in a nongravity-dependent area not adjacent to an area with more extensive burns with contracted tissue, one needs a blood pressure and a beating heart. Antemortem blisters also typically have a red base with surrounding erythema. This concept remains controversial in some jurisdictions. Also, a dead body from a house fire will decompose at a much slower rate than a dead body not exposed to smoke and intense heat Smoke and heat serve as preservatives. Low heat will accelerate the putrefication process. Radiant heat in a dry environment will quickly cause tissue to become firm and dehydrated, often with hair still present. This depends on the amount of heat and humidity and the duration of exposure. Also, intense heat may produce postmortem artifacts that may be misinterpreted as antemortem injury, such as an epidural hemorrhage. Epidural hemorrhage may be postmortem and is caused by heat-related contracture of the dura mater forcing blood from adjacent vessels into the epidural space. Subdural hemorrhages do not occur in this manner and are antemortem injuries.

Thermal burns sustained by fire may be accompanied by smoke inhalation. In general, most fatalities from house fires are caused by smoke inhalation. Decedents dying directly in house fires usually have mostly postmortem burns. Fire fatalities that occur outside in an open space often do not have associated significant smoke inhalation because the smoke rises rapidly and is not inhaled. In the case of a flash fire, inhaled super-heated gases damage the upper airways, including the laryngeal mucosa, and causing death from reflexive closure of the airway at the level of the vocal cords with asphyxia and eventual fatal arrhythmia. The effects of smoke inhalation are often reflected by the amount of carbon monoxide present in the blood. This depends on the nature of the burning material.

There are often other significant poisons associated with burning materials that can rapidly contribute to death, such as hydrogen cyanide. A fire victim does not need a lethal level of carbon monoxide to die of smoke inhalation. Carbon monoxide is produced from incomplete combustion of organic fuels. Carbon monoxide, in the absence of smoke, is a colorless, odorless gas that reversibly binds the hemoglobin molecule approximately 200 times greater than oxygen, resulting in hypoxia and possible death. Levels of carbon monoxide that exceed 50% saturation are considered life threatening, but may cause death with levels less than 26% saturation. Carbon monoxide levels of greater than 80% are possible. These levels need to be correlated with the physiologic disease state of the individual's body, including heart disease, for example. Someone with marked coronary-artery atherosclerosis would often require much less carbon monoxide exposure to produce death than a young healthy individual with slight atherosclerosis. Cigarette smokers may reach carbon monoxide levels of 10%. Carbon monoxide levels of 15–30% are associated with dizziness, nausea, and headache. Cherry-red lividity first becomes apparent at levels of 30–35%. The half-life for carboxyhemoglobin elimination in a resting adult at sea level is generally 4 to 5 hours. This may be reduced to 80 minutes following administration of pure oxygen, and may be further reduced to 24 minutes by using oxygen at 3 atmospheres of pressure. Primary elimination of unchanged carbon monoxide occurs by pulmonary excretion. Also, if there is more than one fatality without obvious cause, one should consider carbon monoxide poisoning.

Electrical burns may be due to low- or high-voltage exposure. The electrical current may be direct or alternating in nature. Alternating current is more likely to cause a fatal cardiac arrhythmia than direct current. High voltage is generally defined as greater than 1000 volts for alternating current and greater than 1500 volts for direct current. High-voltage burns are usually associated with extensive obvious injury. Low voltage is generally defined as being less than 1000 volts for alternating current and less than 1500 volts for direct current. Low-voltage burns may present with no visible marks to the body's surface at all. The degree of injury depends on many factors, including the duration of exposure and the amount of heat generated.

Electrical burns may also occur as a result of a lightning strike. Lightning bolts occur with an enormous short-term release of electricity, often producing minimal injuries. Lightning strikes may also produce a fern-like red pattern at the skin's surface. Patterns similar to this may be observed in high-voltage electrocution. These types of electrical discharges present with different autopsy findings. The mechanism of death is usually arrhythmia and more likely to occur if the current passes directly through the heart. Death may also occur due to asphyxia if there is interference with the central nervous system's respiratory centers or paralysis of the chest muscles. To complete an electrical circuit one needs an entrance and exit point for electricity to pass through the body. An otherwise healthy individual may be found lying barefoot and on a damp floor next to a power tool with a frayed electrical cord. This is why adequate scene investigation is crucial. It is also important to keep the electrical device as evidence to be tested, and to prevent any other fatalities. There may or may not be burns to the body's surface at autopsy.

Chemical burns are due to exposures to caustic substances. These burns most often involve injury to the skin or mucosa, leaving red discoloration or sloughing of the superficial layers. More extensive injuries may involve damage to the underlying tissue including bone. This depends on the strength and nature of the caustic substance, which include acids, bases, and other chemicals that can damage the body. Individuals may die acutely following chemical burns from many different mechanisms including hemorrhage, infection, dehydration, or they may die many years following such injuries. For instance, if an individual attempts to commit suicide by ingesting lye 20 years earlier and later develops esophageal cancer as a result of these burns, the manner of death would be suicide. Children may accidently drink caustic substances, leading to gastrointestinal perforation that may lead to adhesion and gastrointestinal obstruction many years later. In this case, the manner of death would be accidental. It is always very important to find out the initial event that starts the ball rolling in the sequence of events that eventually leads to an individual's demise.

Radiation is defined as energy distributed as waves or particles across the electromagnetic spectrum. This includes electric, radio, radar, microwaves, infrared, visible light (lasers), ultraviolet light, x-rays, gamma rays, and cosmic radiation. Waves are characterized as having long wavelengths and low frequencies, whereas particles have short wavelengths and high frequencies. The types of biological effects vary greatly depending on the type of radiation, duration of exposure, and intermediate barriers. Acute exposure to skin may range from erythema to overt necrosis with eventual epidermal atrophy and dermal fibrosis. Biological effects include cataracts, burns to the retina and skin, necrosis, fibrosis, and cancer. Generally speaking proliferating cells are affected more substantially with acute exposure as indicated by damage to the gastrointestinal and hematopoietic systems with increased risks of infection, nausea, vomiting, diarrhea, and hemorrhage. Damage to DNA may eventually lead to many different forms of cancer.

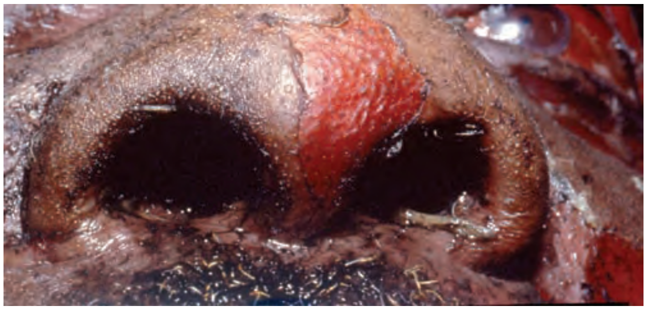

Soot within the nares due to smoke inhalation.

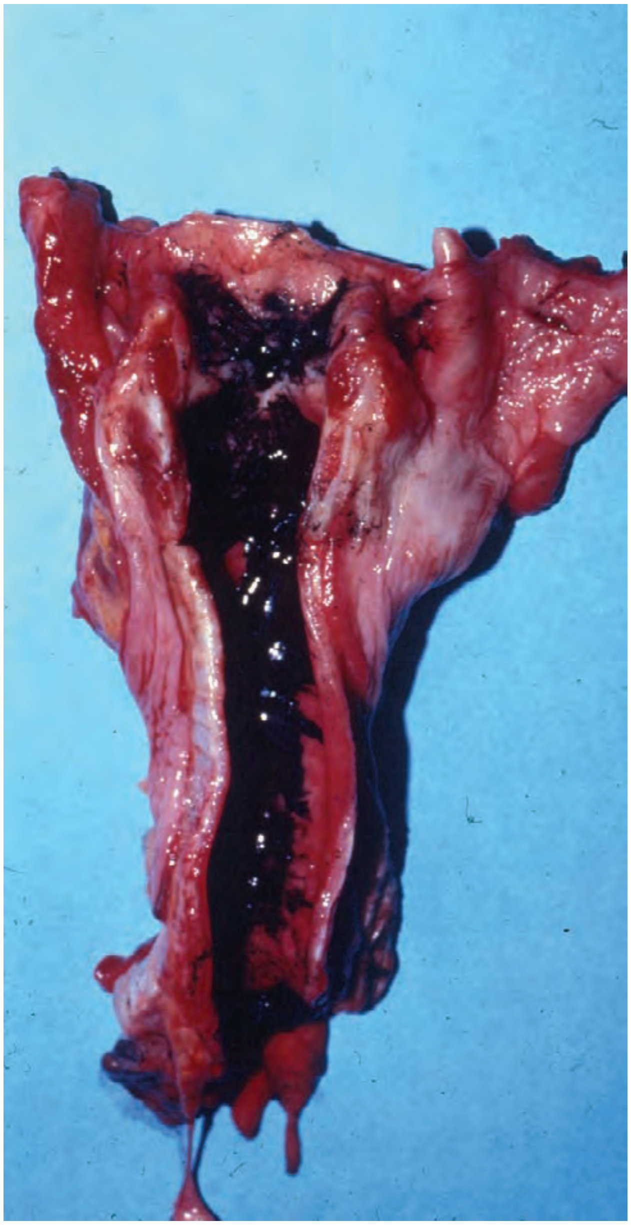

Smoke inhalation with soot deposition on the airway mucosa of the larynx and trachea.

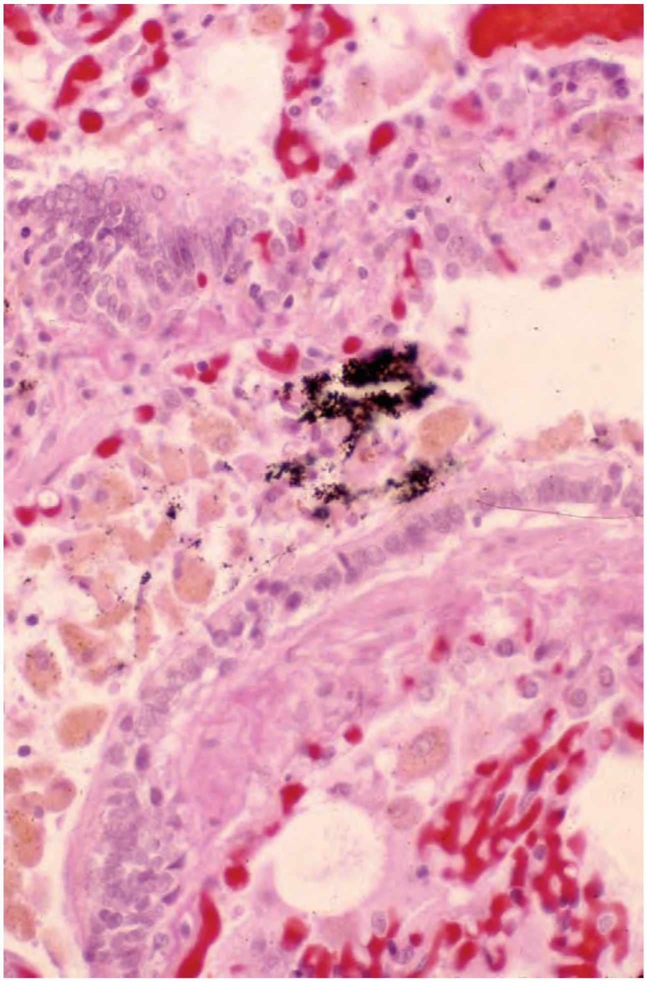

Microscopic view of lung with soot deposition in the distal bronchi.

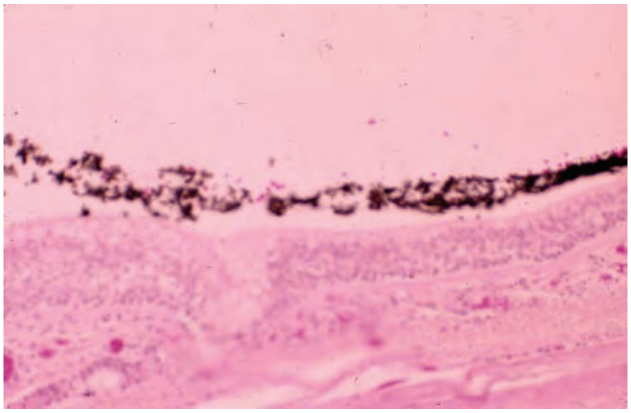

Microscopic view of airway with soot deposition at the surface of respiratory epithelium.

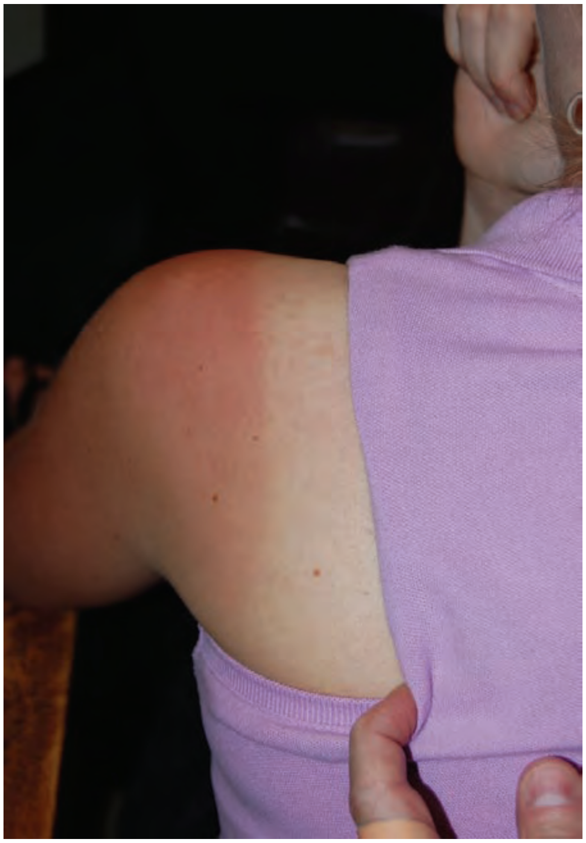

First-degree burn characterized by red discoloration with injury limited to the epidermis.

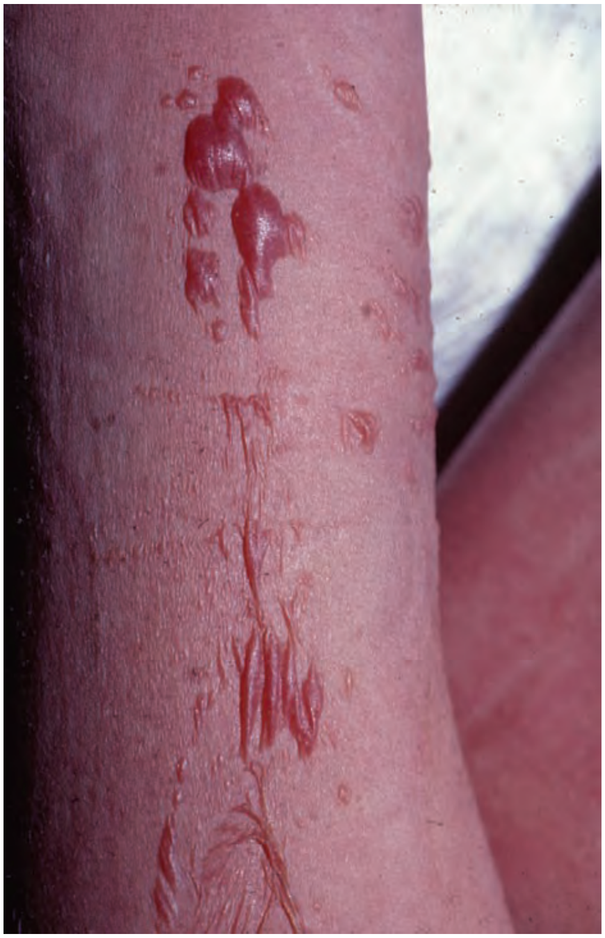

Second-degree thermal burns characterized by fluid-filled-blister formation. Second-degree burns include damage to both the epidermis and cermis and are often more painful than third-degree burns due to less destruction of nerve endings.

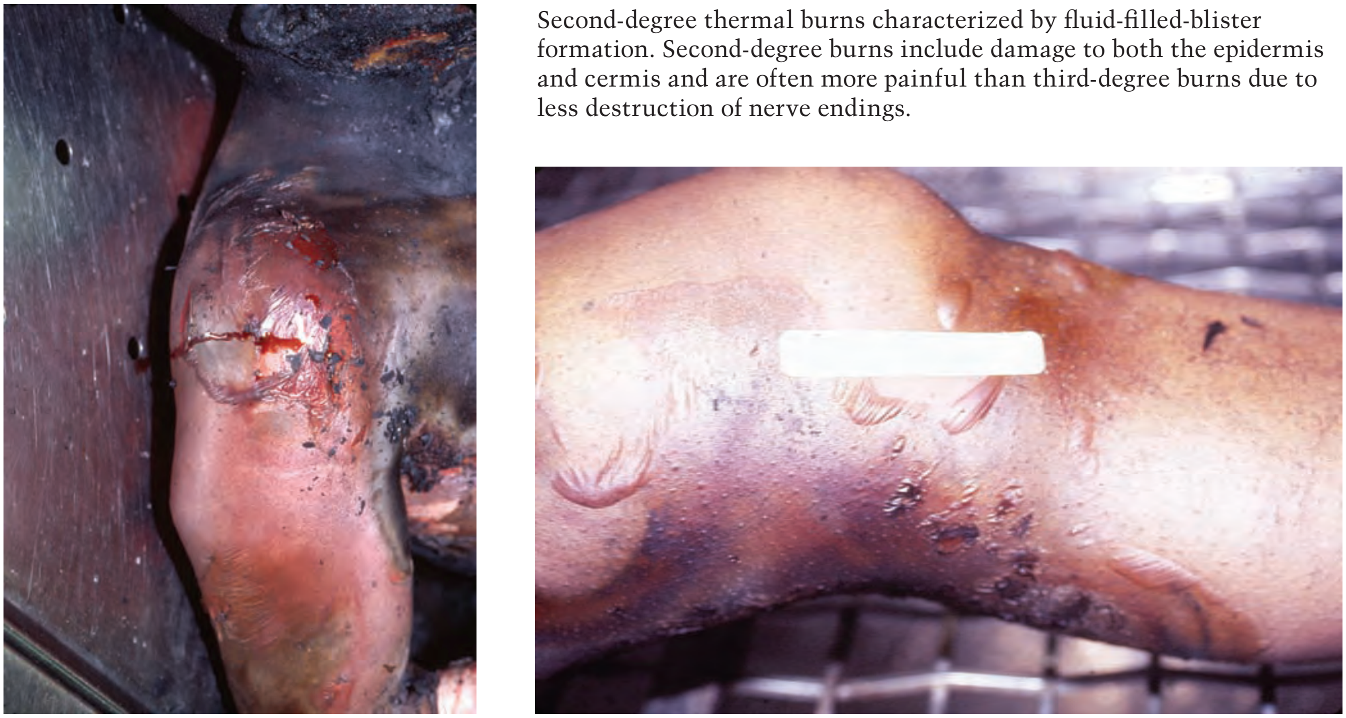

Second- to third-degree thermal burns. These areas demonstrate fluid-filled-blister formations that were interpreted as antemortem. To have fluid-filled-blister formation associated with thermal injury, one usually needs to have a blood pressure. Fluid-filled blisters generally do not occur due to putrefaction in burn victims due to the preservative effects of fire and smoke. Postmortem blister formation may occur in gravity-dependent regions without significant thermal damage or adjacent to regions with soft-tissue heat-related contractures. Also, postmortem blisters lack vital reaction.

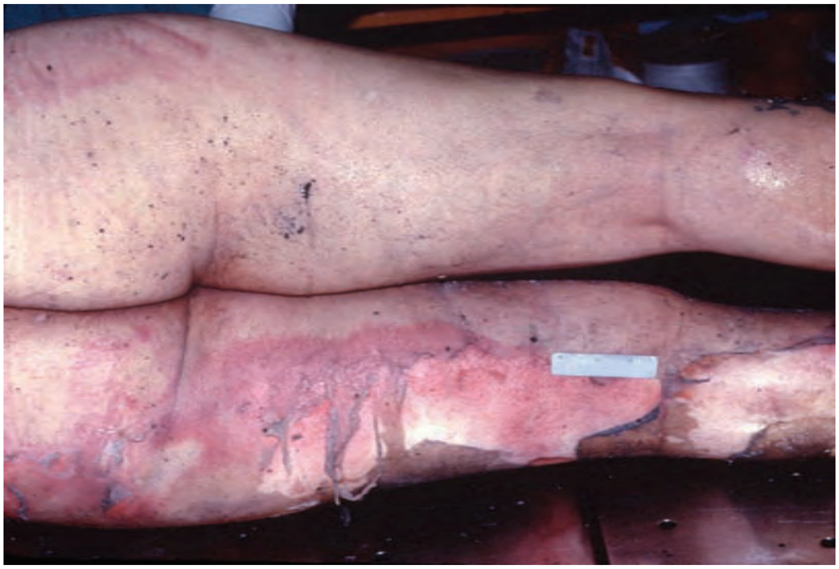

First- to third-degree thermal burns involving the posterior aspect of the left thigh and gluteal region. The right leg is without thermal burns. First-degree thermal burns in this picture are characterized by the red discoloration without blister formation or skin slippage. Note the areas of collapsed blister formation, which are consistent with a postmortem burn. Sometimes it is difficult to interpret antemortem burns if continued heat causes fluid-filled blisters to collapse and fluid to evaporate. Most fire fatalities succumb from smoke inhalation before extensive burns occur.

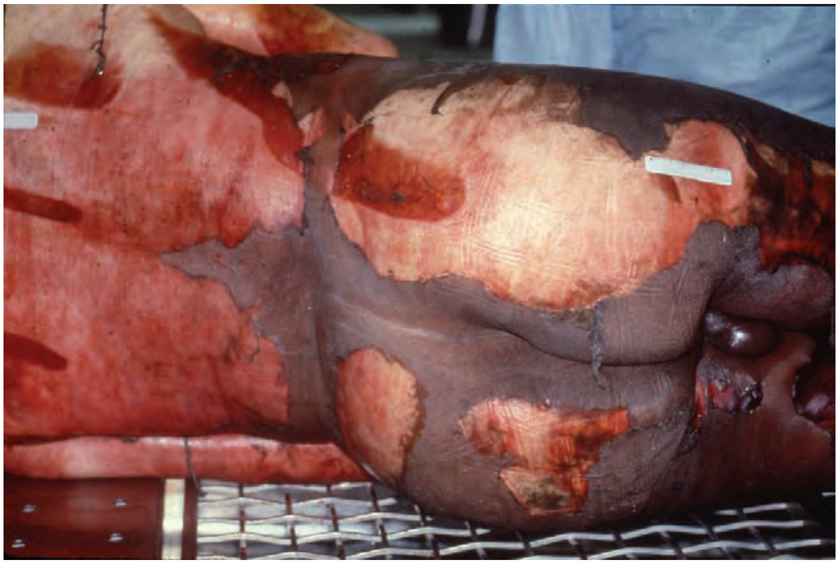

Postmortem second- to third-degree thermal burn with skin slippage and collapsed blister formation.

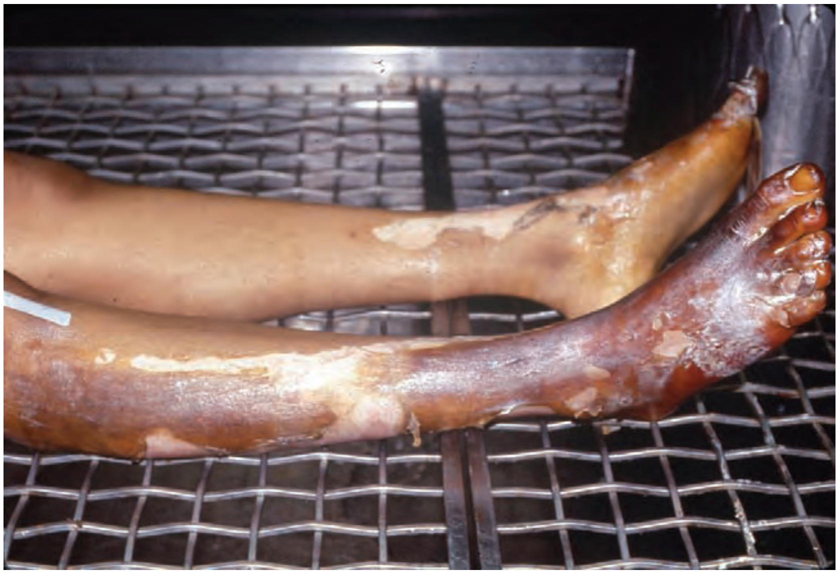

Full-thickness or third-degree burns in an individual who lived hours after being shot in the head and having his or her residence set on fire to destroy evidence. Full thickness refers to involvement of the epidermis, dermis, and subcutaneous layers. These are often less painful than second-degree burns due to more complete damage of nerve endings in the latter.

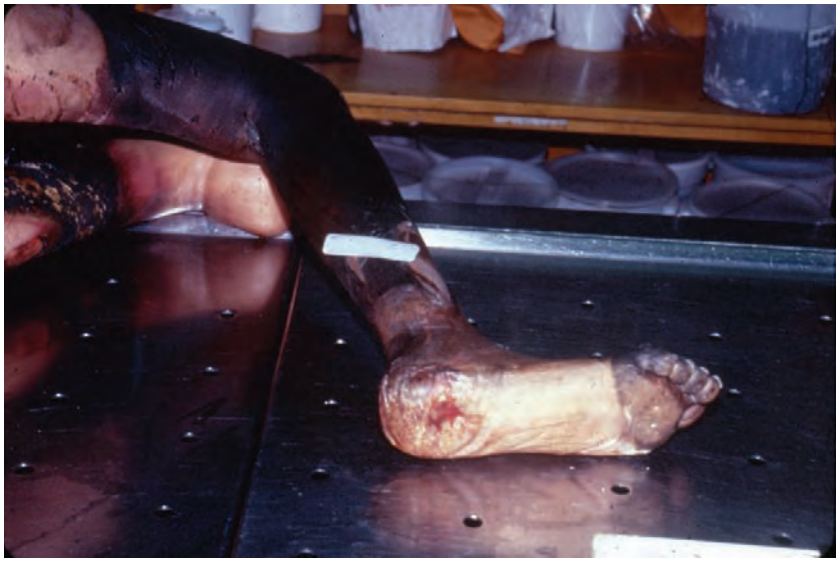

Note the sparing of thermal injury at the bottom of the decedent's foot due to protection by a shoe.

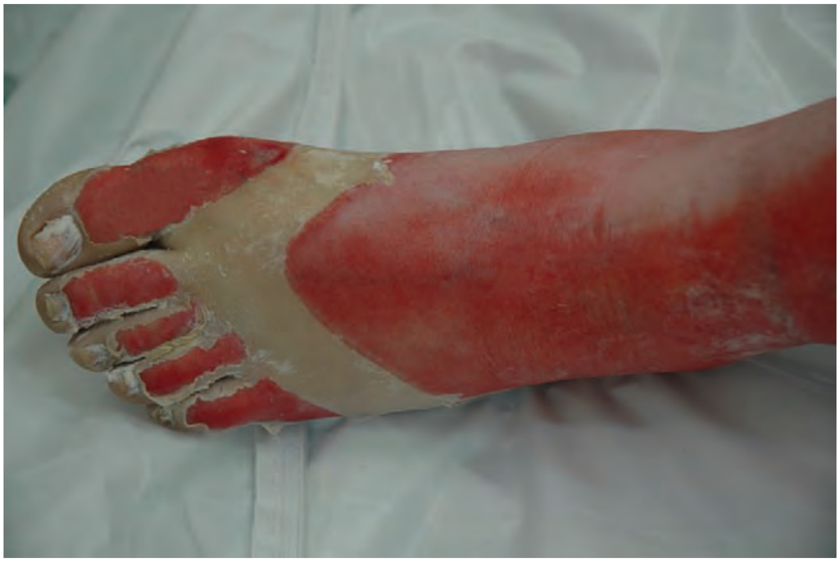

Antemortem third-degree burns. This is self-immolation of an individual who was wearing sandals. Note the lack of thermal injury in the region protected by the shoe strap.

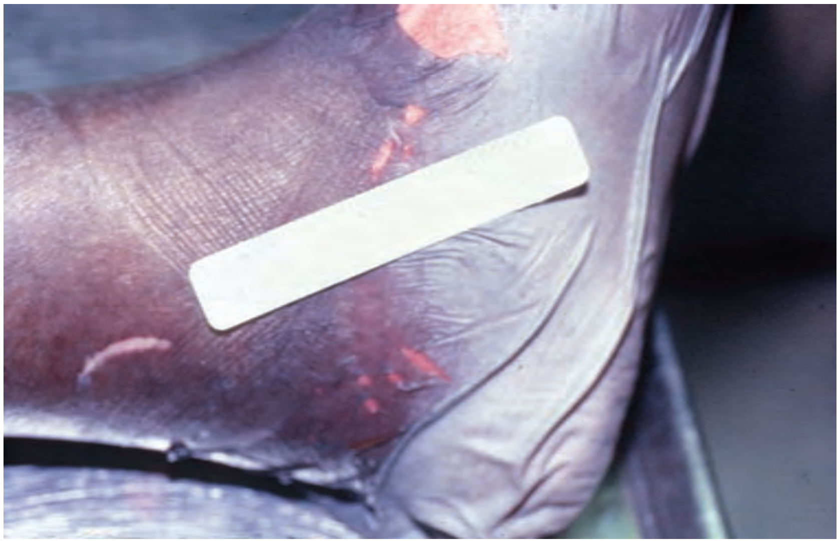

This image demonstrates postmortem second- to third-degree thermal burns to the sole of this foot. Note the wrinkled thick skin demonstrating skin slippage. There was minimal underlying fluid accumulation or blister formation.

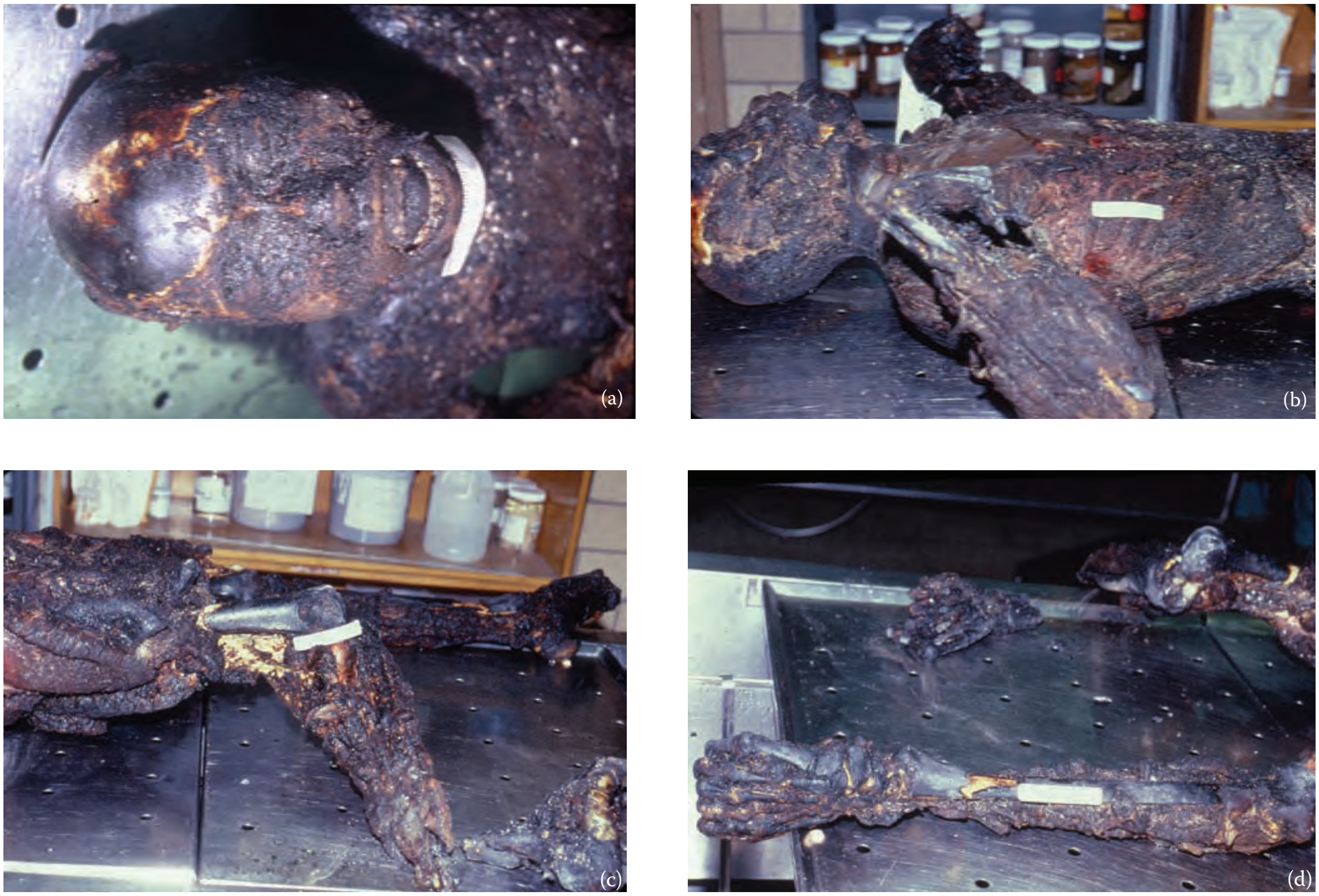

Postmortem fourth-degree thermal burns with extensive charring and exposed muscle and bone. Note the pugilistic stance of the body with arms raised (b). Also note the postmortem skeletal fractures due to extensive heat (c).

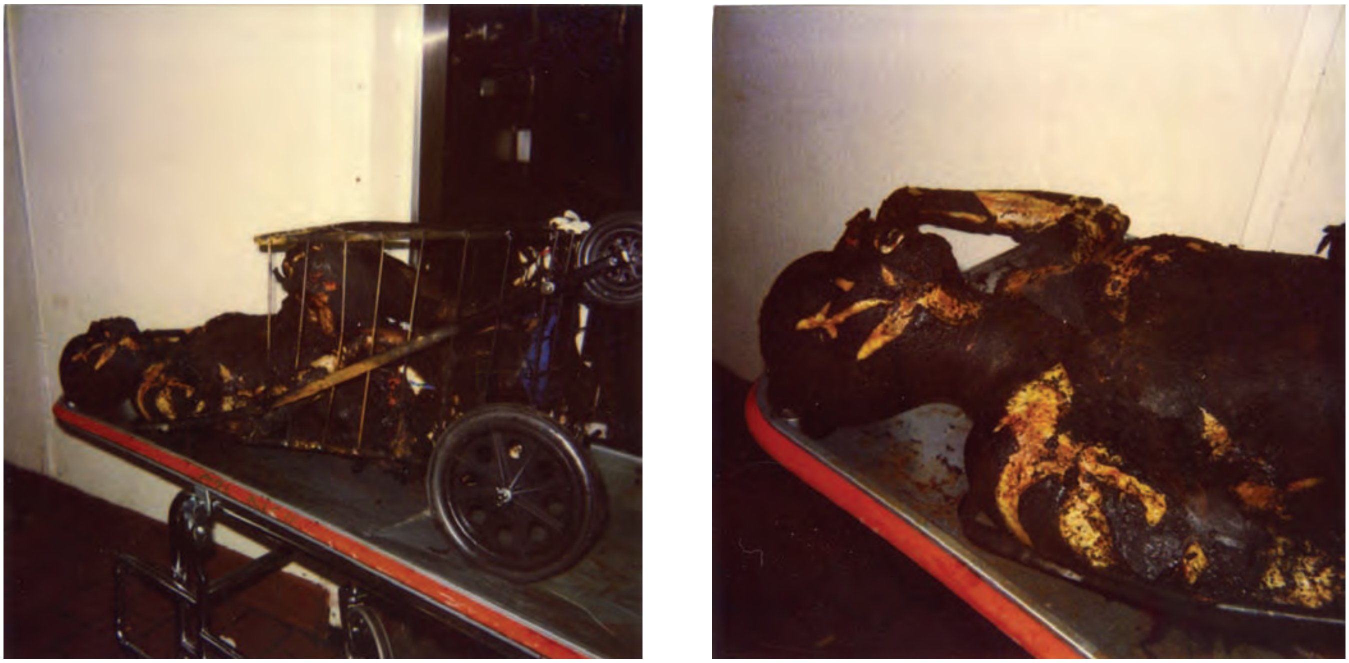

This individual was wheeled out of a building in a shopping cart and set on fire. Note the postmortem cracking of the skin due to heat exposure. This is indicated by the exposed yellow subcutaneous tissue with no hemorrhage. There were extensive fourth-degree thermal burns to the entire body.

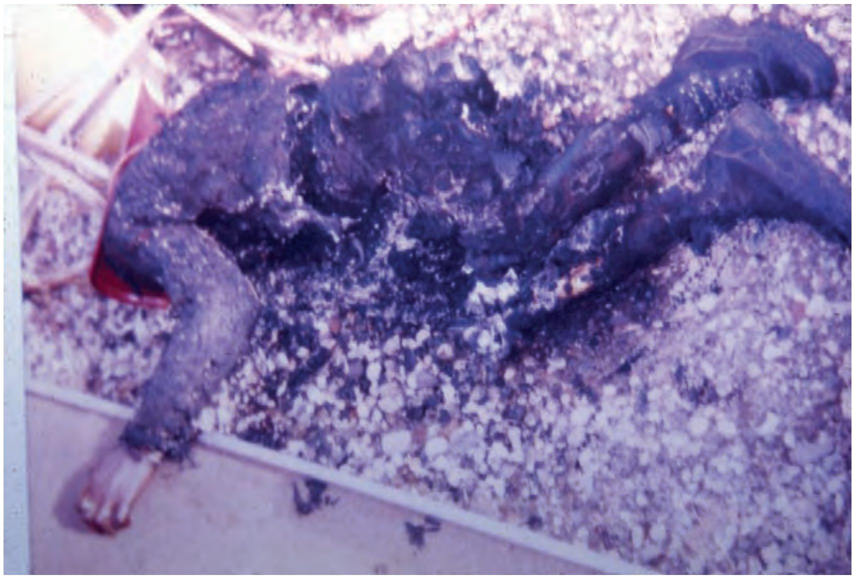

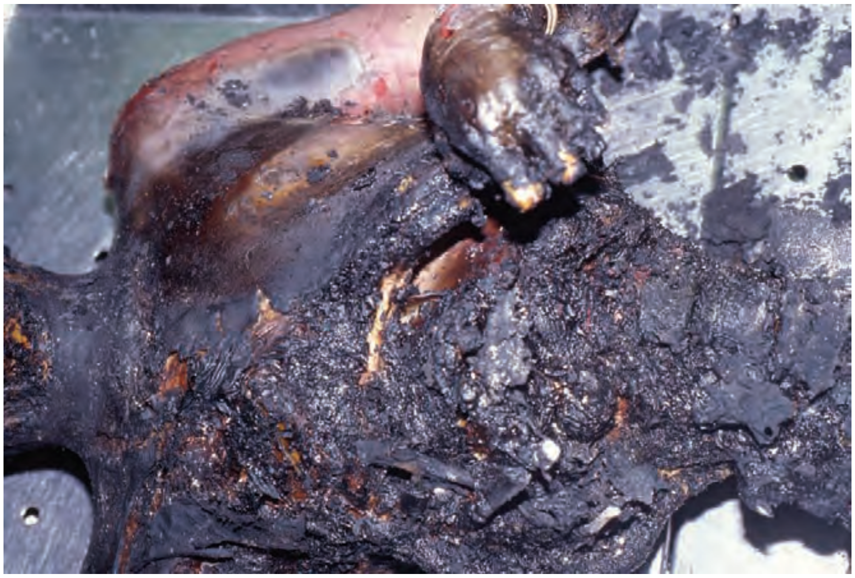

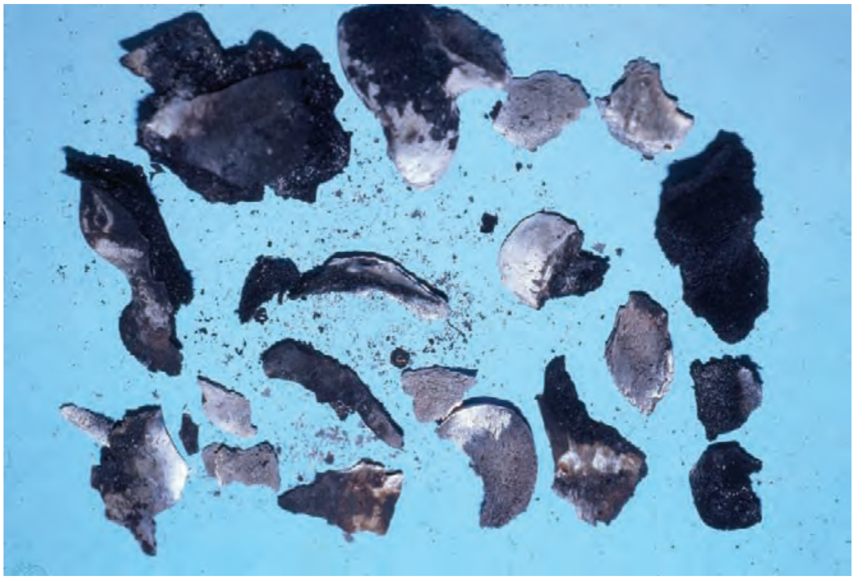

Extensive third- to fourth-degree thermal burns with partial skeletal fragmentation due to the wick effect, which refers to a self-perpetuating, low-intensity flame following ignition of certain materials contacting the body, where the skin is cracked from heat and the underlying fatty tissue is rendered into oil that is absorbed into the charred clothing, producing a wick. This low-level heat can produce extensive destruction to a body over hours.

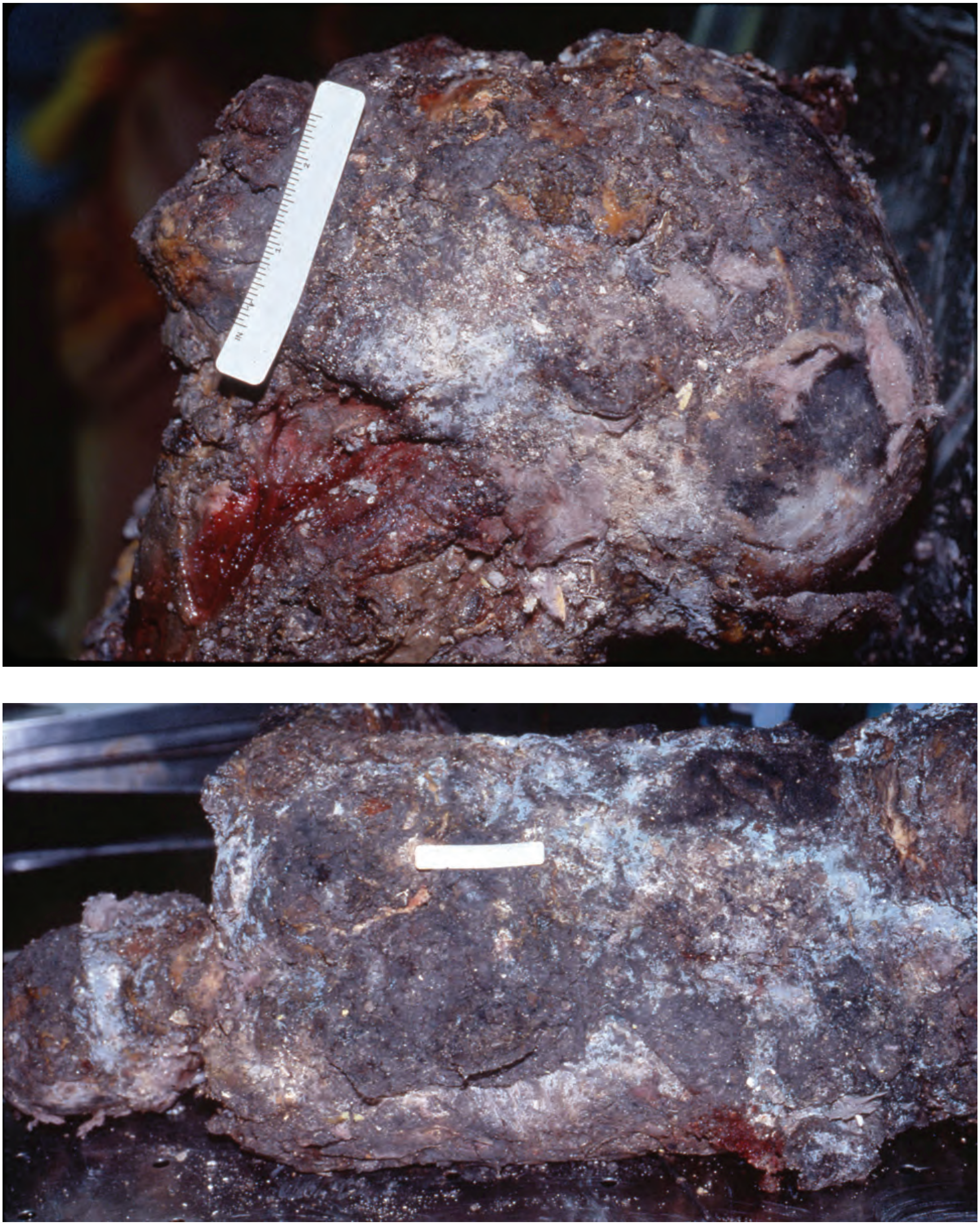

This is a rare finding of decomposition after extensive thermal injury and exposure to smoke. The decedent was involved in a fatal fire and the body was not discovered for several days after being soaked in water following fire extinguishing. Note the microorganisms and mold growing at the body surface shown by the gray-white discoloration. Putrefactive changes were markedly inhibited due to the effects of exposure to smoke and heat.

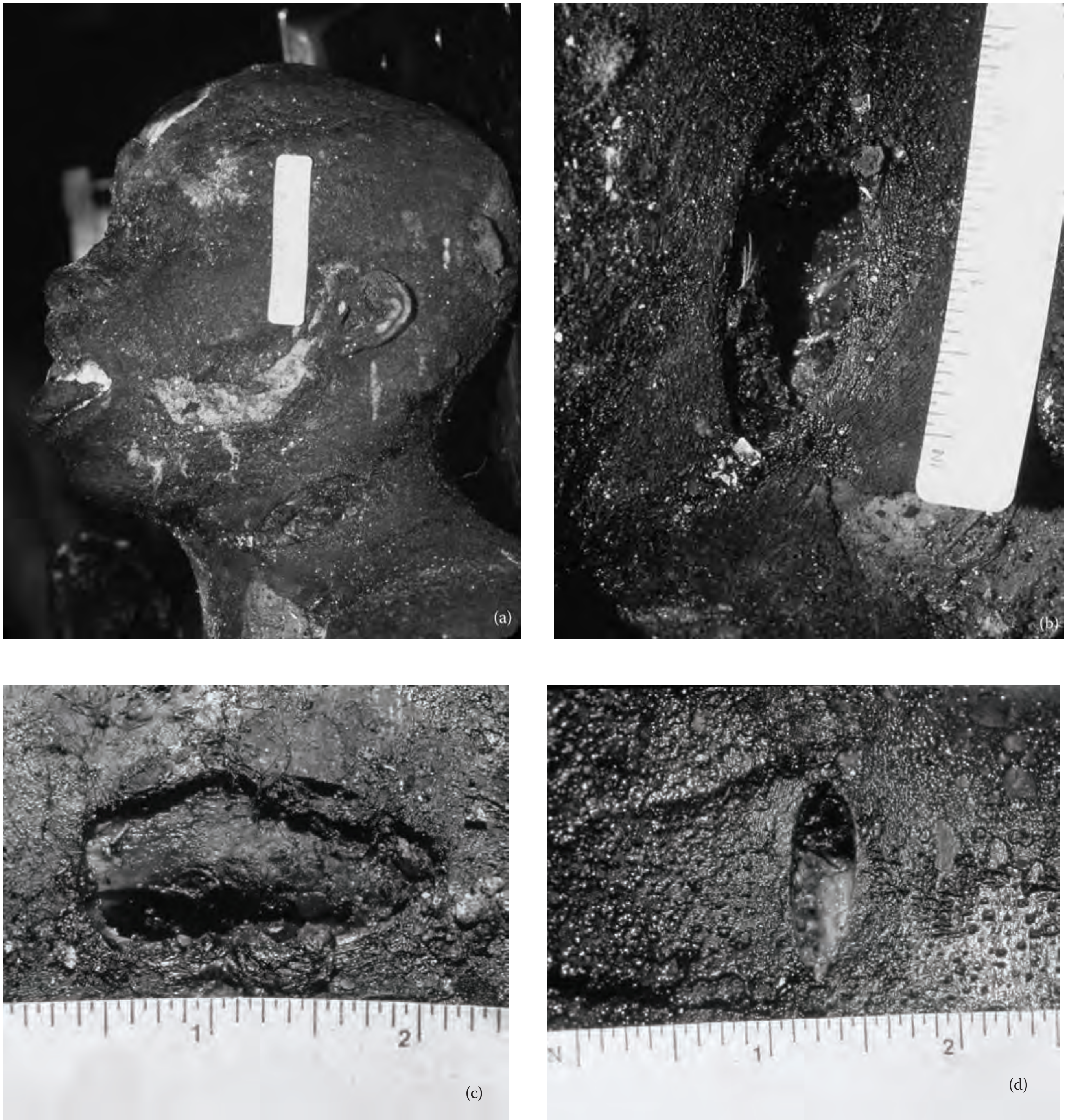

These are all examples of homicides where apartments or houses were set on fire to destroy evidence. Figures (b–d) are examples of antemortem stab wounds that have been altered by postmortem thermal burns. Note that the margins are relatively sharp and do not appear as though the skin has cracked secondary to heat exposure. Upon internal examination, these injuries become much more apparent with hemorrhage and blood accumulation. Note in (a) areas of cracked skin with exposed yellow subcutaneous tissue without hemorrhage, which is indicative of postmortem thermal injuries. Note the wound at the upper aspect of the lateral left neck with hemorrhage due to an antemortem stab wound.

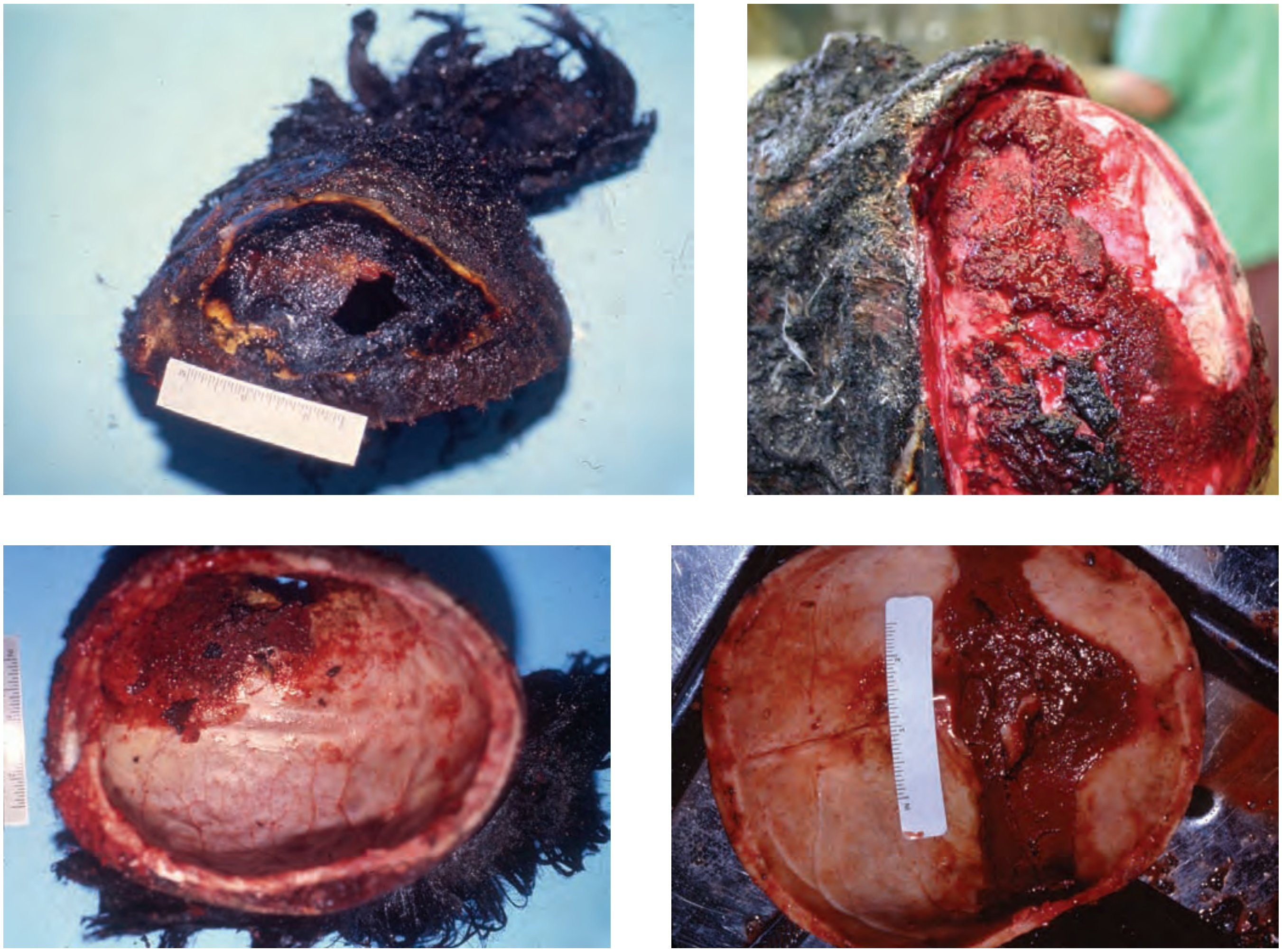

These figures demonstrate postmortem epidural hemorrhage. These can sometimes be misinterpreted as antemortem blunt force trauma. With exposure to flames, the brain and dura mater may contract, and blood may be forced from the small vessels at the inner aspect of the cranial bone and through the dural sinuses, producing epidural blood accumulation that will coagulate with heat.

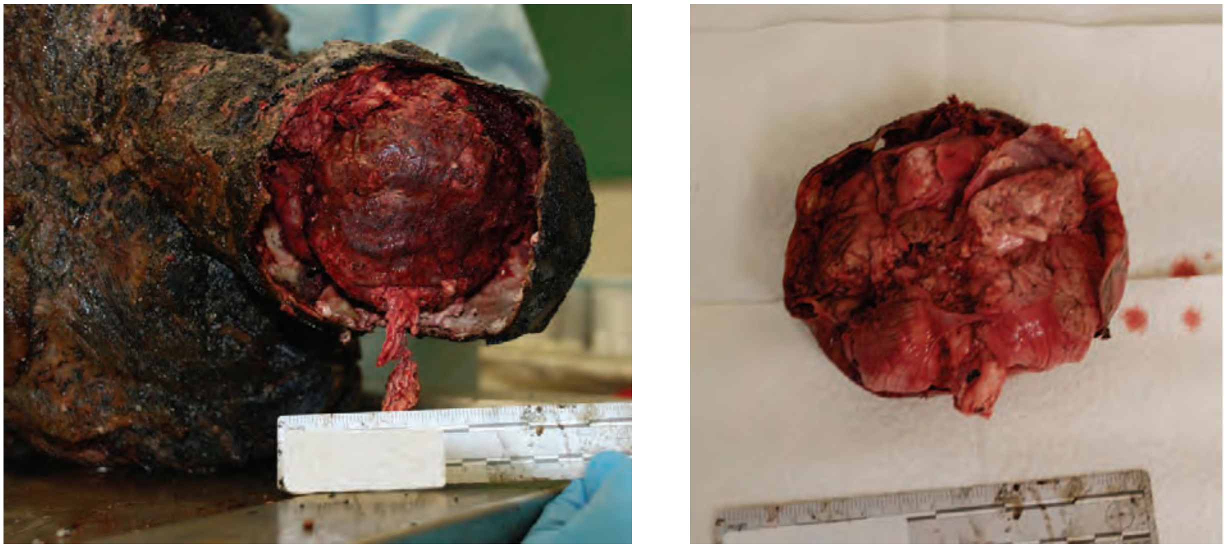

Extensive postmortem thermal injury with brain shrinkage. Organs exposed to extensive heat will decrease in size largely due to dehydration.

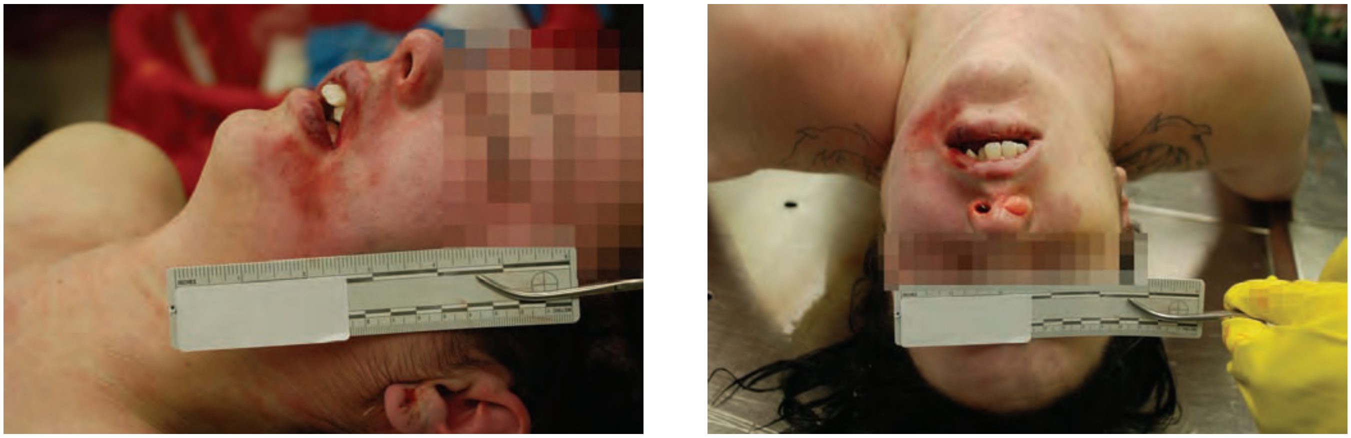

This individual died of an overdose and was found with dried vomitus on her face. Note the red discoloration caused by gastric acids producing burns to her face.

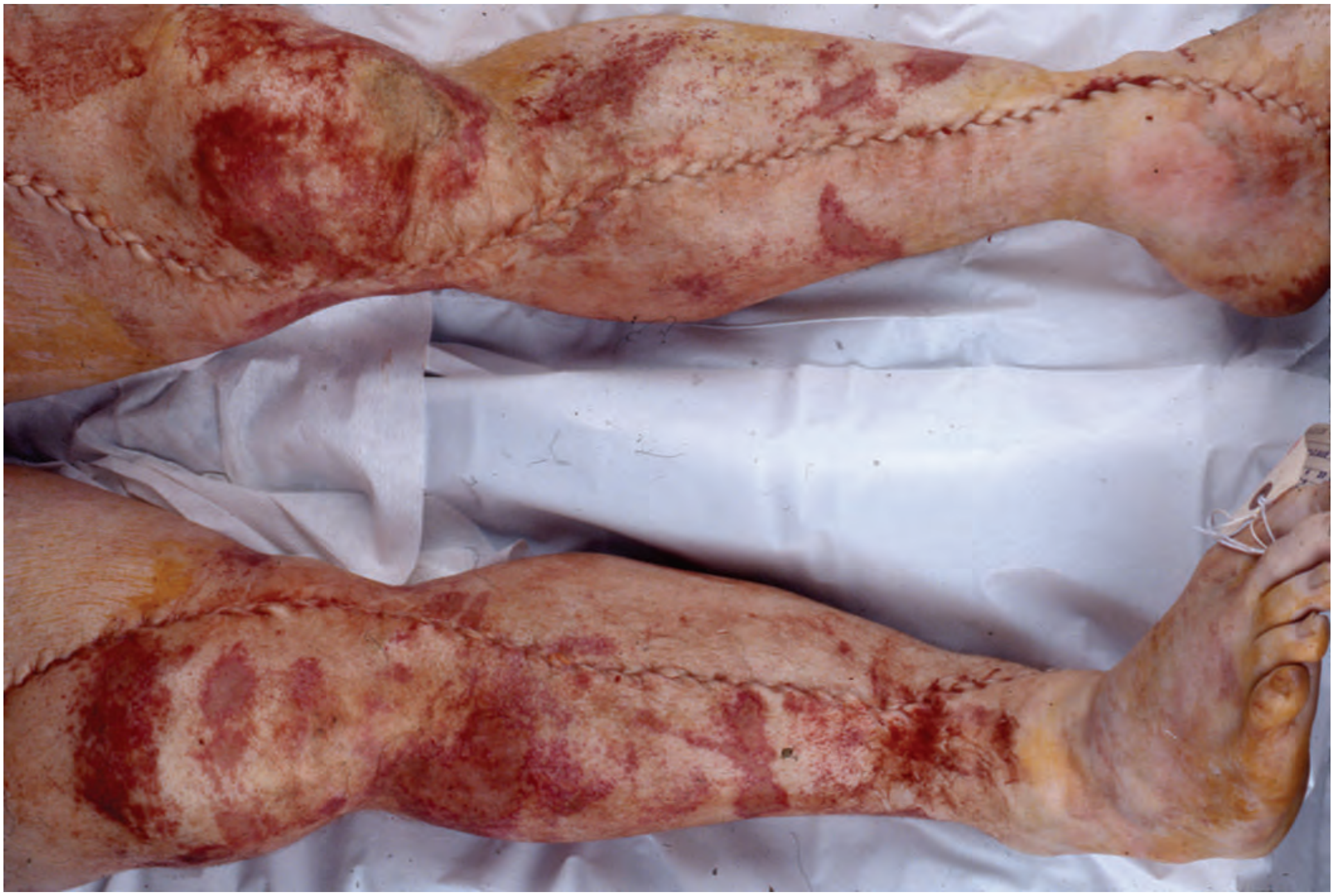

Chemical skin burns caused by spilling sulfuric acid. The sutured linear incisions are due to organ donation with retrieval of bone and soft tissues.

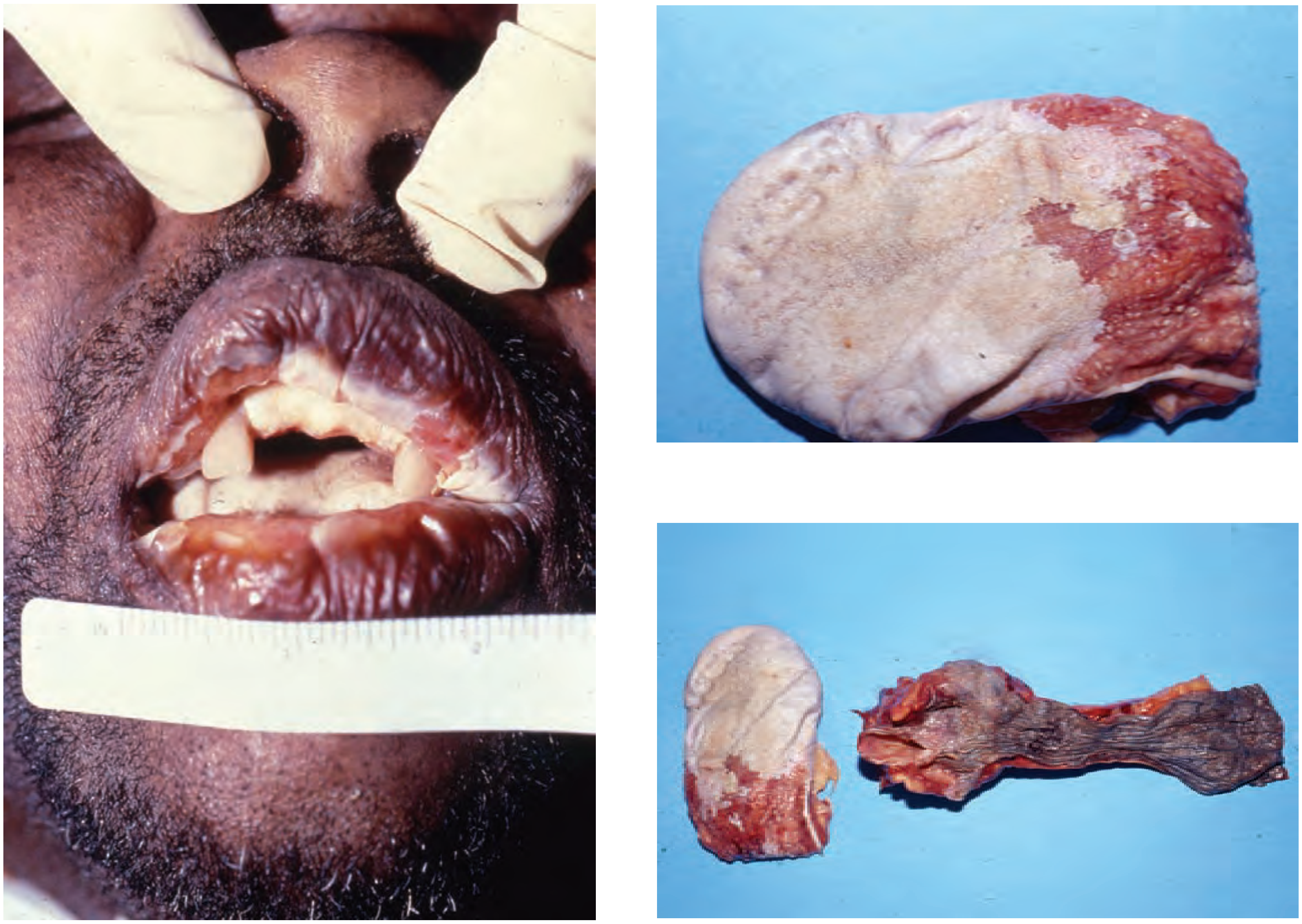

This individual drank a mixture of lye, kerosene and other caustic chemicals. Note the white discoloration from chemical burns at the lips, mouth, tongue, and esophagus.

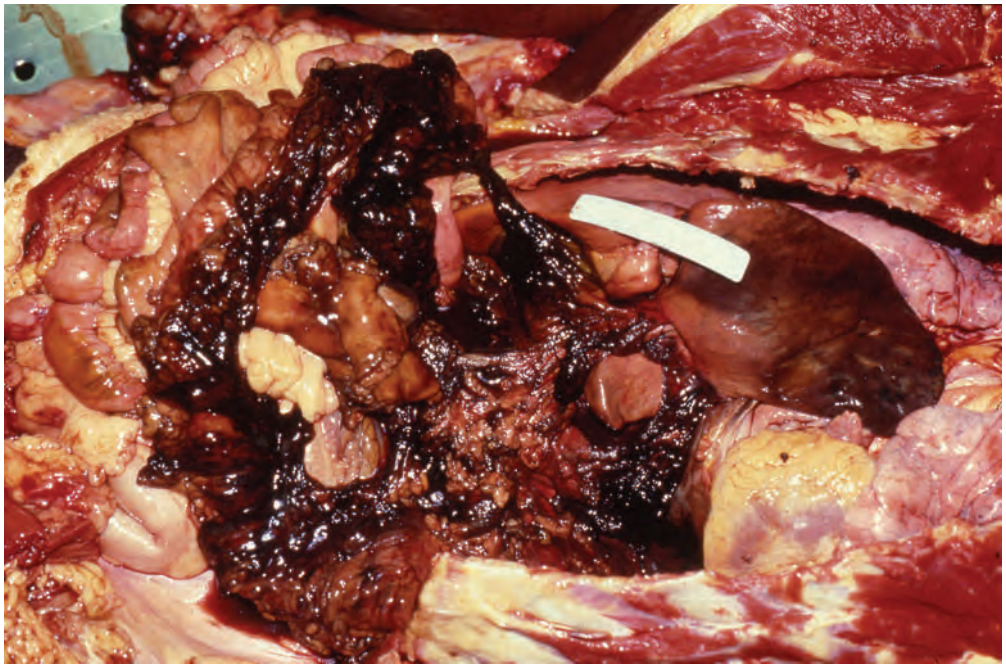

Note the red to brown discoloration in the abdominal cavity following gastric perforation and leakage of the caustic chemicals into the peritoneal cavity.

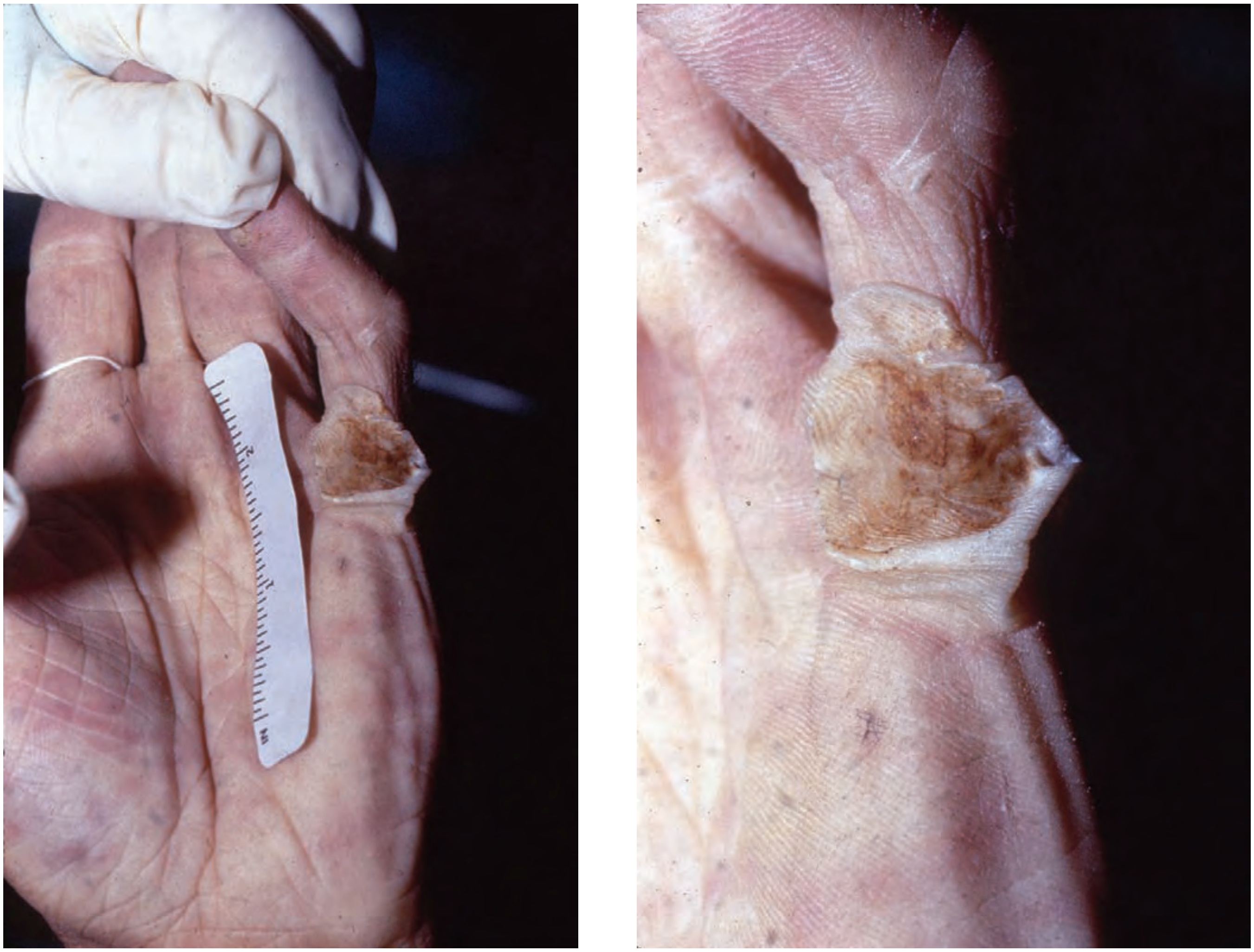

Electrical burn on the hand of this individual who reached into a ceiling and grabbed a live wire. He was standing on an aluminum ladder wearing shorts. His leg contacting the ladder completed the circuit through his heart, producing a fatal arrhythmia.

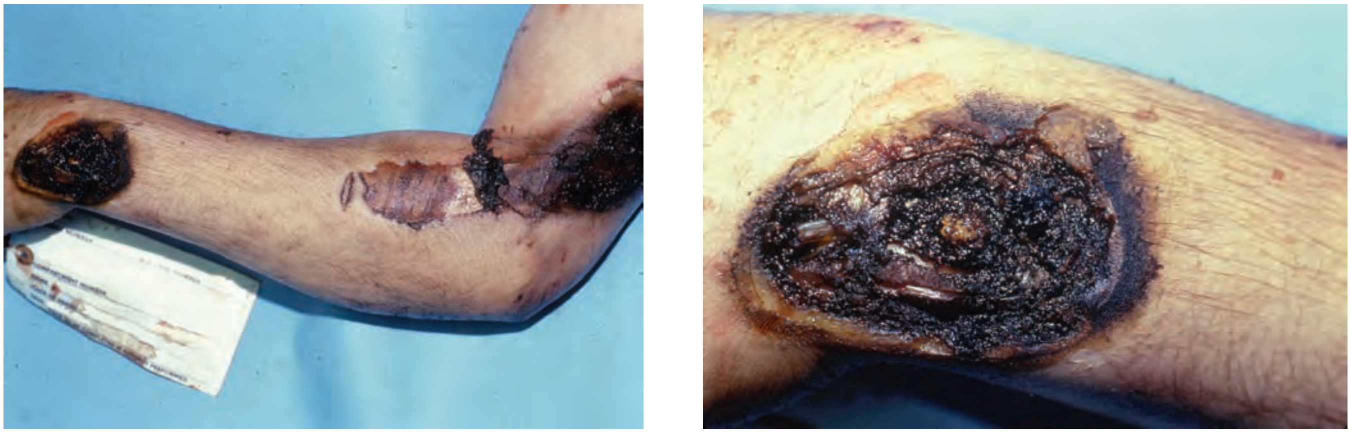

This individual fell from a subway platform onto the third rail and sustained these electrical burns.

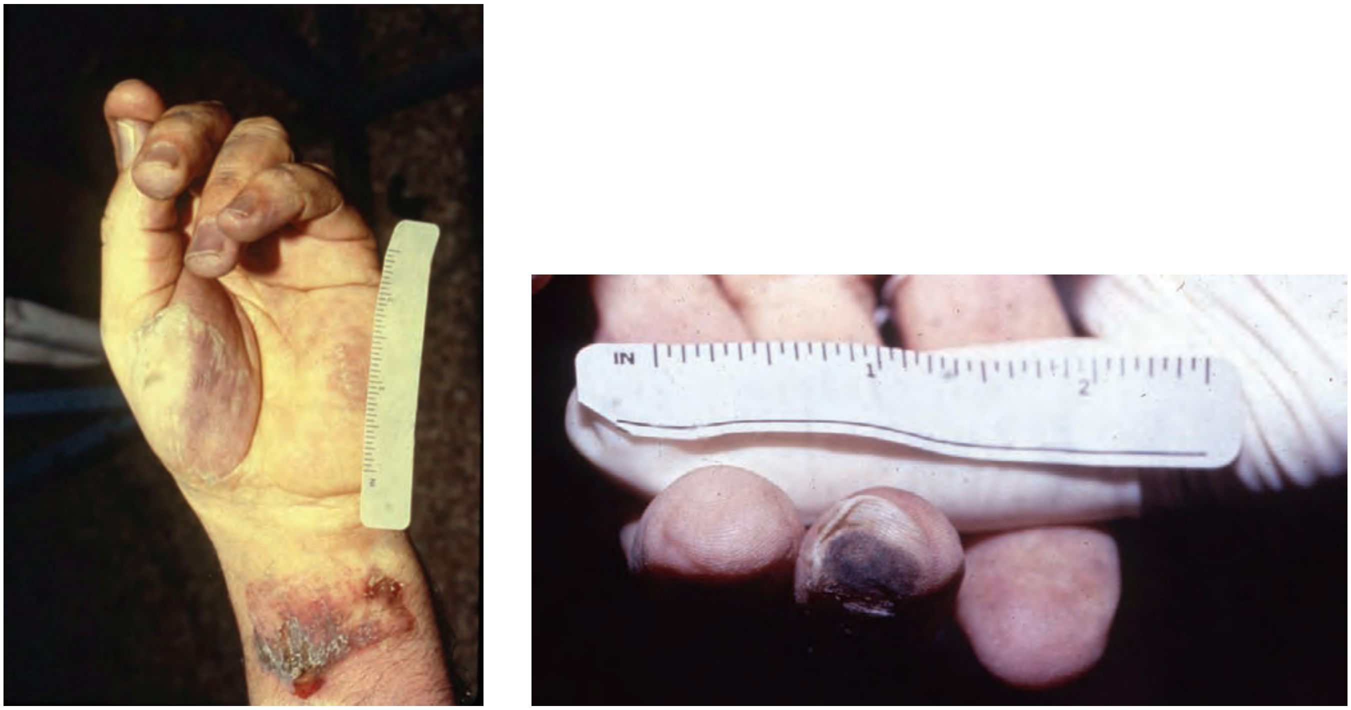

These are examples of second- to third-degree electrical burns caused by inadvertently contacting live wires at construction sites. These individuals died as a result of a fatal cardiac arrhythmia. They fell to the ground lifeless within about 15 to 20 seconds after contact.

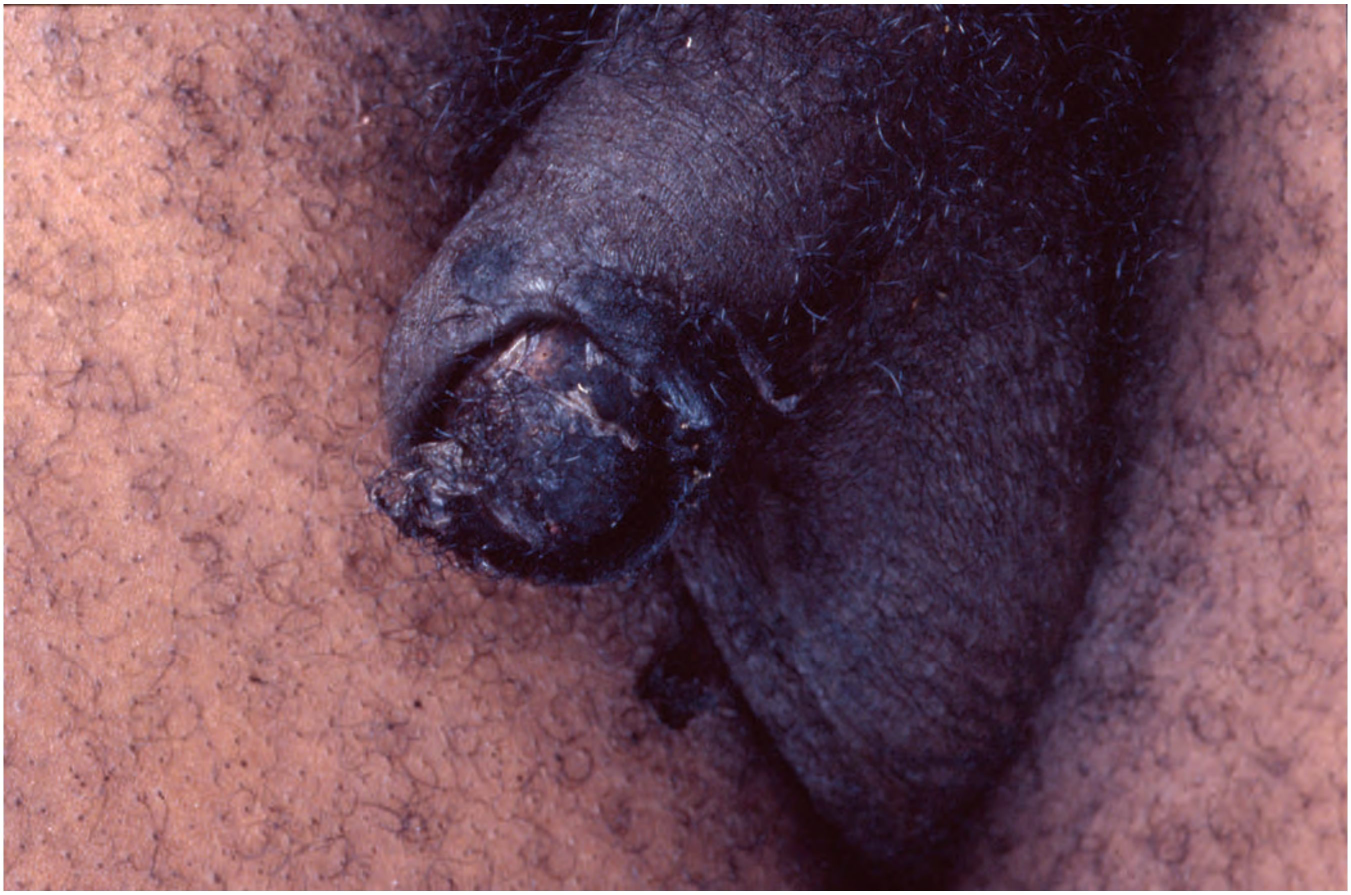

This individual was found lying next to an electrified subway rail with his pants down, and an electrical burn to his penis. He was lying in a puddle of urine and was markedly intoxicated at the time of his death.

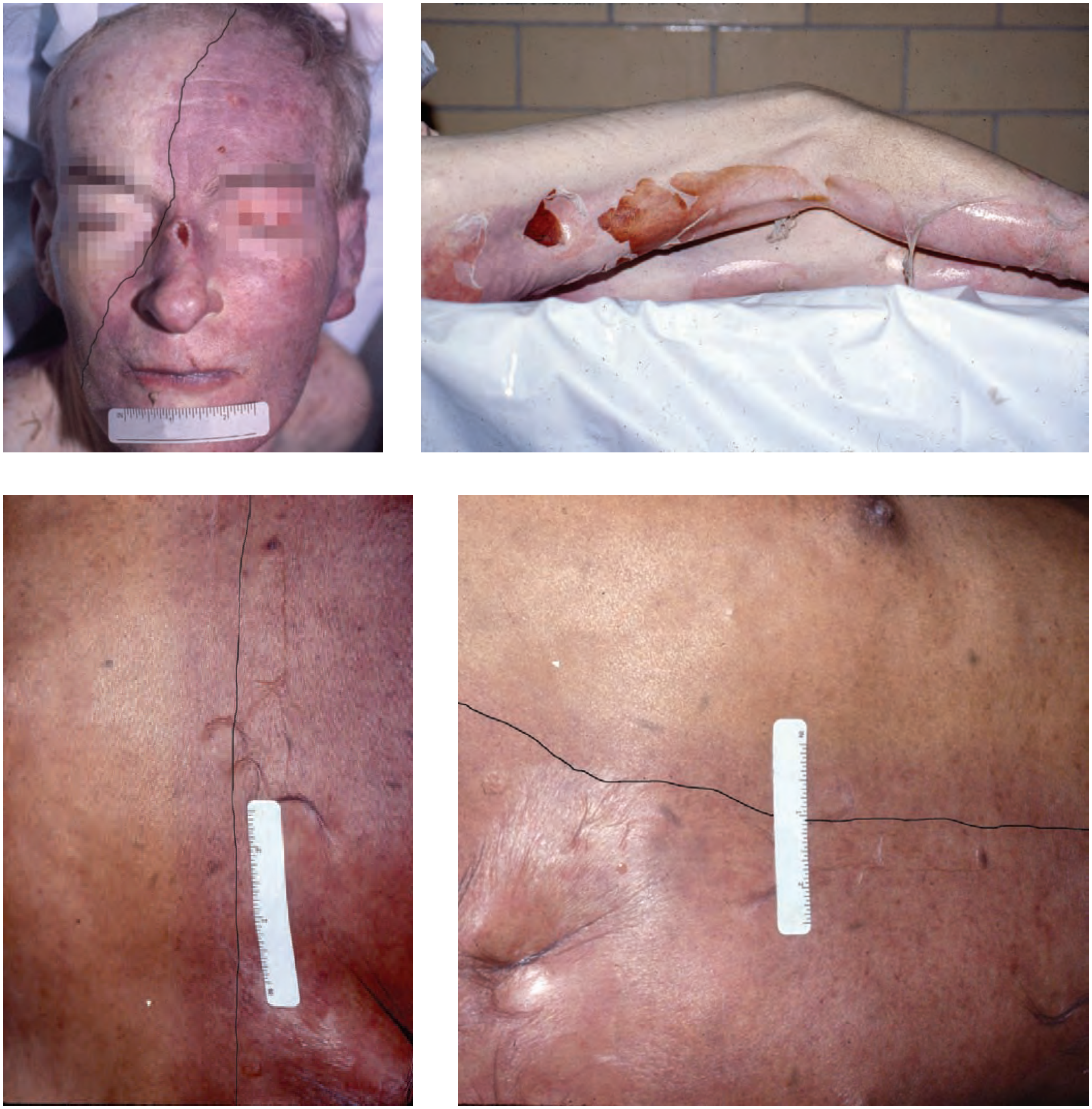

These individuals died of heart disease while taking a bath. These are examples of second- to third-degree postmortem burns due to being submerged in warm to hot water. Postmortem burns occur with less heat than antemortem burns. Note the red discoloration with skin slippage and a sharply demarcated border defining the submerged and unsubmerged areas. To help the viewer distinguish between these regions we placed a line adjacent to this demarcation. Individuals who drown in bathtubs have some contributing factor dictating why they could not keep their head above the water. They maybe neurologically compromised or intoxicated. People with seizures may accidentally drown.

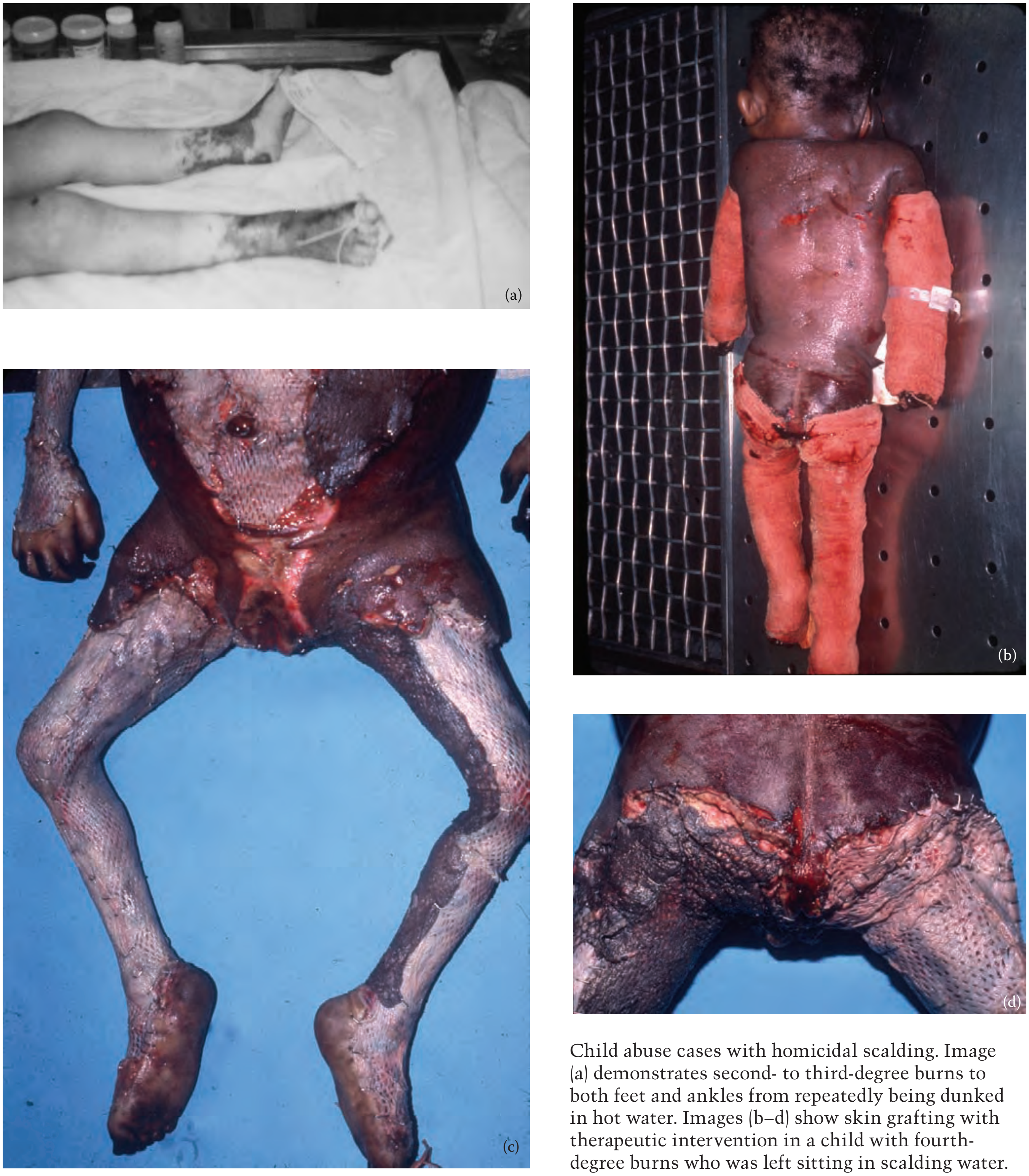

Child abuse cases with homicidal scalding. Image (a) demonstrates second- to third-degree burns to both feet and ankles from repeatedly being dunked in hot water. Images (b–d) show skin grafting with therapeutic intervention in a child with fourth-degree burns who was left sitting in scalding water.

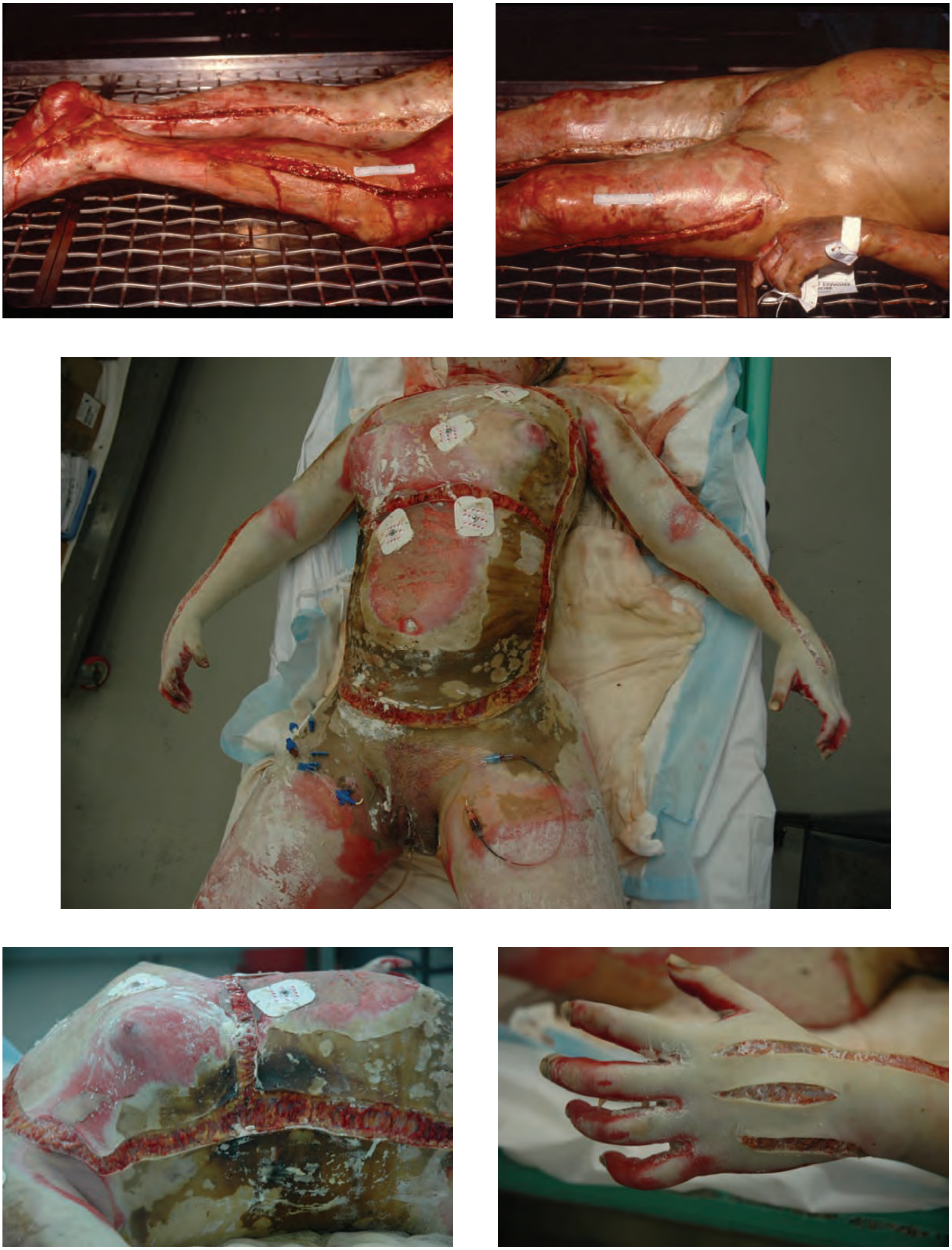

These are fasciotomy incisions made by physicians for medical therapy to relieve pressure in extremities.

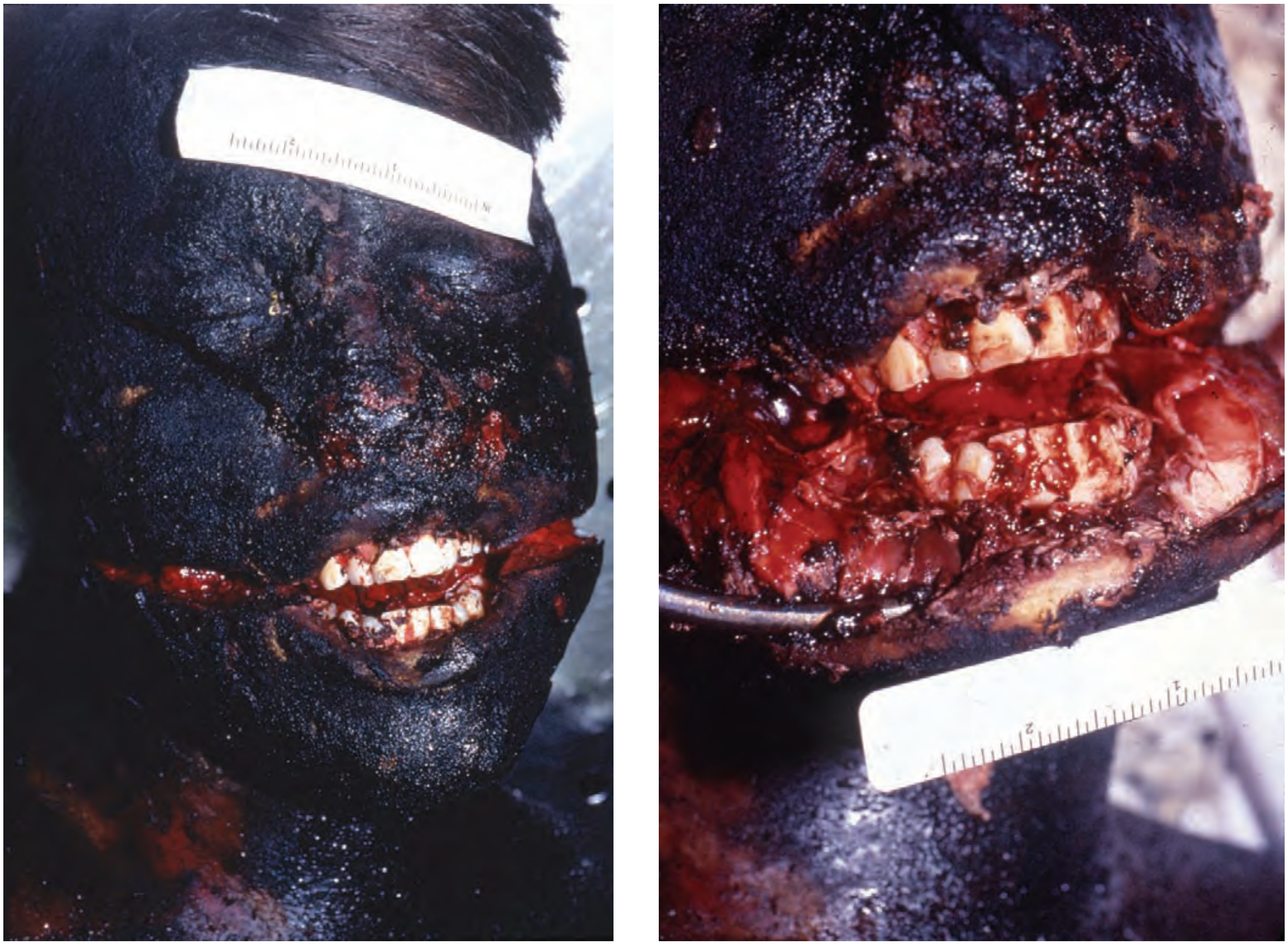

Identifications may be very challenging in fire fatalities. These cases may require dental analysis for identification. Incisions may be made in the face to access the teeth. Funerals in such cases are closed casket due to the severe extensive nature of the injuries.

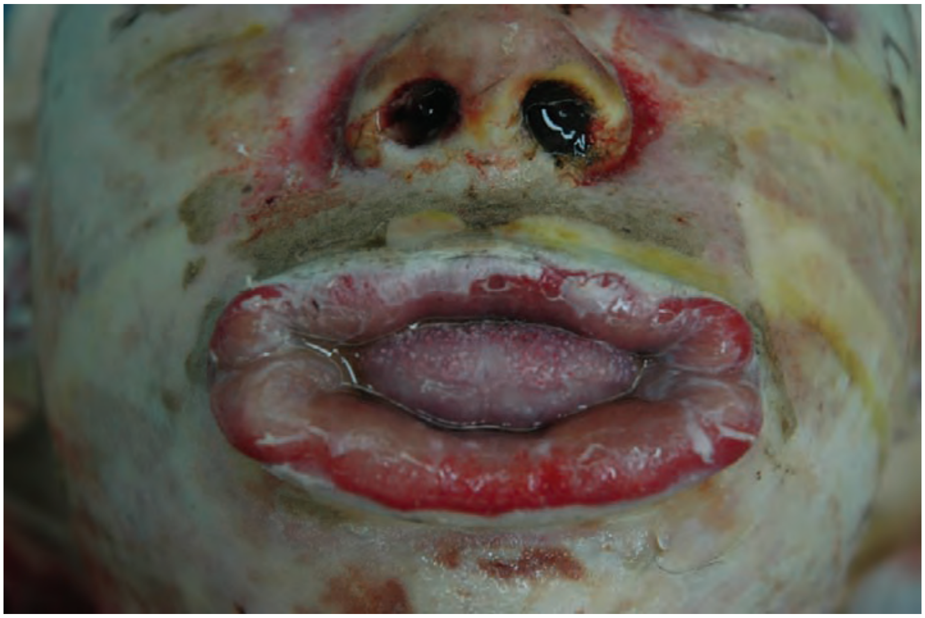

Swelling of the tongue due to inhaling super heated gases in an individual who survived for 1 day.

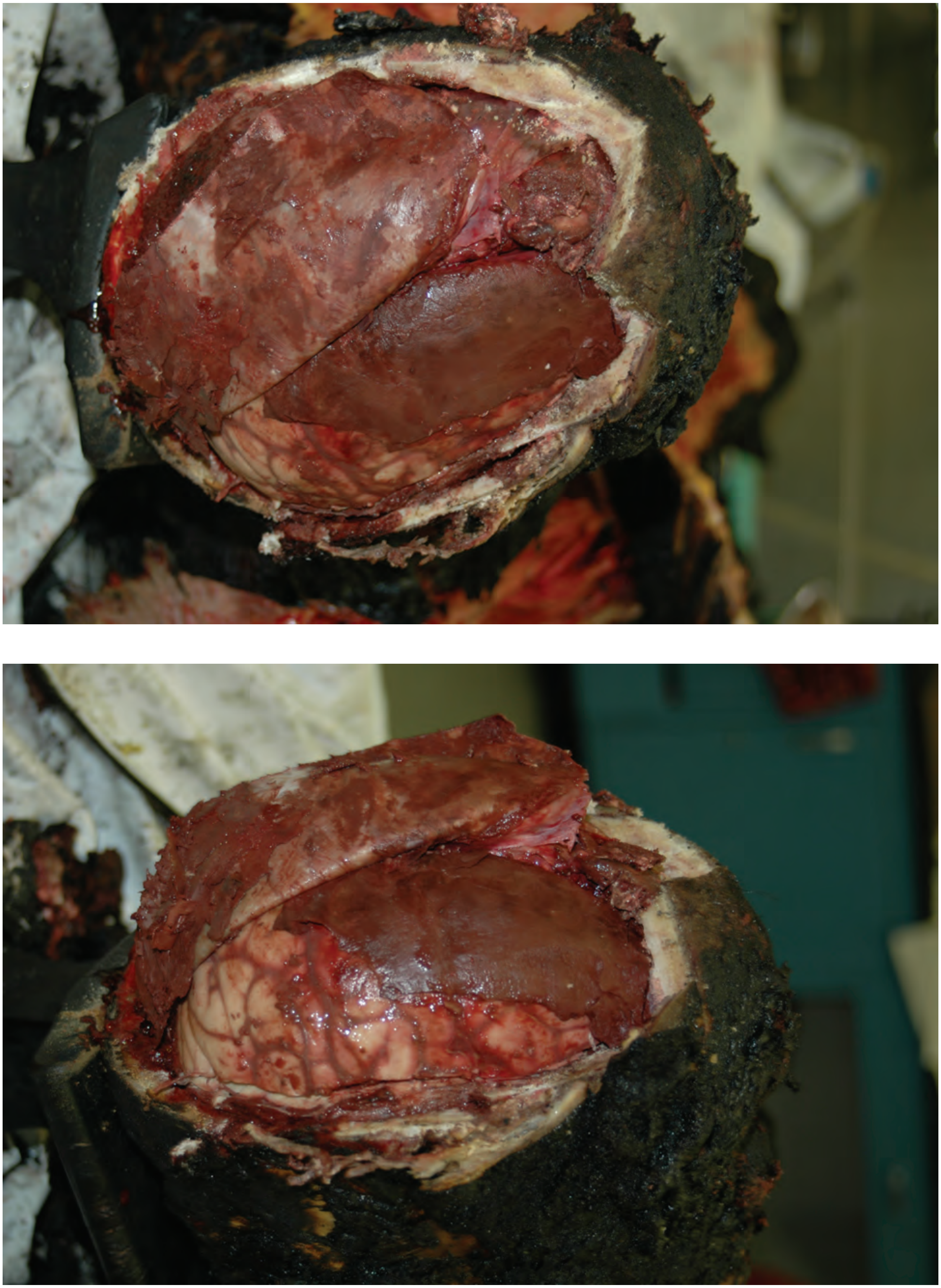

This individual crashed an airplane and sustained extensive blunt force trauma. These images demonstrate an antemortem subdural hematoma altered by extensive postmortem thermal injuries.

Asphyxia

10

CHARLES A. CATANESE AND BARBARA K. BOLLINGER

This is a general term used to describe decreased oxygen uptake or use, together with decreased carbon dioxide elimination.

Airway obstruction may occur by smothering, neck compression, foreign body aspiration, excess secretions or swelling of the airway, etc. *Smothering* is defined as external occlusion of the mouth and/or nose which prevents air exchange. Children may *aspirate foreign bodies* such as peanuts, hotdogs, popcorn, watch batteries, coins, etc. Adults who aspirate food are invariably neurologically compromised or intoxicated. Airway obstruction due to *excess mucus or swelling* as with asthma or anaphylaxis can produce asphyxia. Also various body positions may produce airway obstruction *(positional asphyxia)* as with occupants of motor vehicles trapped after collisions or intoxicated people passing out and sliding into awkward positions that prevent chest expansion and air exchange.

Autopsy findings associated with *smothering* may be very subtle or non-existent. Findings may include abrasions around the nose and/or mouth that *cannot* be explained by other means (i.e., resuscitative efforts). Great force is applied to the mouth and lips which may cause tears to the frenulum of the lip, the mucous membrane that connects the inside of the lip to the corresponding gum. Smothering may occur with the use of hands or by placing an object over a face, such as a pillow.

Depending on the degree of force applied and the structures compressed, there may or may not be petechiae present on the skin of the face, mucous membranes or eyes. Arguably, the greater the disparity in size between the perpetrator and the victim (i.e., adult and child), the less likely there will be a demonstrable injury due to the overpowering relative nature of this type of struggle. It is important that in cases of suspected smothering, experienced police interrogators and medical investigators perform the interviews. In cases where autopsy findings are very subtle, well-documented descriptions of the circumstances with specific details are extremely important.

Chemicals can produce asphyxia. *Inert gases* like methane or carbon dioxide will displace oxygen from the air and produce asphyxia by depletion or replacement of oxygen. Various *poisons* such as carbon monoxide or cyanide interfere with oxygen uptake and utilization, respectively.

Chest compression can produce asphyxia by *preventing air flow* into the lungs.

Neck compression as with *hanging and strangulation* can also produce asphyxia by obstruction of various neck structures including the airway, venous circulation and arterial circulation.

Interpretation of autopsy findings with respect to *hanging vs. strangulation* can be challenging. Each subheading below will describe the presenting classic and most common features, and then elaborate on less common features. It is important to realize that there is overlap between how the two present; depending on how the act is carried out, they may appear very similar. In establishing the manner of death, one should consider all aspects of the case including the past medical history (i.e., depression, end stage cancer, etc.), scene investigation and autopsy findings.

Hanging refers to ligature compression of the neck mitigated by the gravitational forces of the hanging head, causing partial or complete obstruction of the neck structures including blood vessels and the airway.

In a typical non-judicial suicidal hanging an individual places a ligature with a slip knot encircling the superior aspect of his or her neck. He or she secures the other end of the rope to a fixed support and allows the entire or partial body weight to pull downward, occluding the neck structures until loss of consciousness and death. In this case there should be furrow pattern that matches the overlying ligature which forms an inverted "v" mark or indentation, extending upward at the superior aspect of the neck and head. With the entire body weight pulling downward, all of the neck structures are (i.e., arterial, venous, and airway) are usually occluded at the same time and one would not expect to find petechiae in the face or eyes. There are typically no hemorrhages or fractures of the neck structures or other injuries to the body indicating a struggle. The cervical vertebrae are rarely fractured in suicidal hangings. When the body hangs for longer periods of time the furrow indentation becomes more prominent. Individuals cut down shortly after this act may have little or no furrow mark. This depends on the type of ligature used. A wide soft ligature will leave less of a mark than a narrow, more resistant ligature. If the body is left to hang for days, decomposition with stretching may eventually lead to the head being pulled away from the body.

In the case of a judicial hanging, the body is dropped from a height to produce sufficient force to fracture the upper cervical vertebrae resulting in spinal cord injury with cardiac and respiratory satiation.

An individual hanging in a sitting or lying position may have partial occlusion of the neck structures before loss of consciousness ensues. In situations where the body is adjacent to another structure, the individual may partially pull his or her body up and down, causing varying degrees of pressure–release before loss of consciousness. This will produce a similar effect to what is seen in strangulation. In these circumstances, the up and down motion of pressure release will obstruct different neck structures at different intervals. Venous circulation requires the least amount of pressure for occlusion, as compared to the arterial system and the airway. When venous circulation is obstructed without the arterial circulation, the higher pressure arterial blood beats through the capillary beds rupturing small blood vessels producing petechiae. As this process continues the hemorrhage size increases and may become confluent. These are most obvious within the sclera and conjunctivae. These movements may also produce hemorrhages or fractures to the neck structures including the airway cartilages and hyoid bone. These findings are more characteristic for strangulation, but may be seen in hangings.

Though unusual, people have been known to tie their hands behind their backs during the process of hanging themselves. The individual may have tried to complete this act several times in the past but the will to survive overpowered the will to end life. The tied hands are usually loosely tangled and not tightly tied. This will give the individual enough time to prevent himself from stopping the process. There are usually no other signs of a struggle or defensive type injuries.

With a free hanging and total body weight suspension, an otherwise healthy individual, using a slip knot, would be expected to lose consciousness within 15 to 20 seconds, and suffer irreversible brain damage within 4 to 6 minutes.

Hanging is usually suicidal, but may be accidental or homicidal.

Strangulation may be by ligature or manual. A *ligature* is something flexible that can encircle the neck, like a cord, belt or piece of clothing, etc. *Manual* strangulation refers to the use of one's hands leading to compression and blockage of the neck structures.

Manual strangulation is usually characterized by multiple irregular, angulated, abraded contusions around the neck. The marks may be curvilinear, corresponding to fingernail prints. These external marks can be somewhat variable and can range from a few to many.

Ligature strangulation is usually characterized by a horizontal furrow or mark pattern around the neck. The extent of these injuries depends on the type of ligature, how broad and soft it is, and the amount of struggle, etc.

The act of strangulation is often a very physically dominating, often non-premeditated, way of killing somebody. This act takes time and comes with the risk of injury to all those involved. The individual strangled is usually smaller and of weaker strength. Often a sexual component to the assault exists and a rape kit should be performed in all cases of suspected strangulation. Petechiae are usually present in the face and eyes. There are usually hemorrhages in the strap muscles of the neck, and there may be fractures of the laryngeal cartilages and/or hyoid bone. These fractures are more common in older victims because the cartilages are more calcified, brittle and less elastic. Older people may also have osteoporosis. Younger victims or children tend to have more flexible upper airways that often will stretch or collapse rather than fracture. Depending on how great the struggle, the amount of force used, and the type of neck compression, there may or may not be petechiae and/or hemorrhages above the ligature or region of neck compression. The presence of petechiae formation is more likely when there is a pressure–release component associated with a struggle or if less variable force is applied and if the ligature is wide with a large surface area. The presence of petechiae is less common when the force is very strong, consistent, and applied with a small surface area ligature. The latter example is more similar in nature to a hanging. Victims of strangulation often have defensive type injuries including other abrasions and contusions to their body.

It is possible to ligature strangle yourself; it is not possible to manually strangle yourself. If one is able to apply enough force to lose consciousness manually, revival occurs after the pressure is released. Continued force is necessary for death to ensue.

Someone murdered by ligature strangulation may die within a similar time frame as someone hanged; however, the time frame is usually longer. If there are multiple petechiae with hemorrhages and fractures of the neck structures, whether ligature or manual, the time frame may be much longer. This latter example usually indicates pressure–release, pressure–release over several minutes until loss of consciousness, and then continued pressure for several more minutes until death occurs. If an individual is released shortly following loss of consciousness, revival may follow.

Proper autopsy technique dictates that the *brain and visceral organs be removed prior to a layered neck* *dissection* being performed. Photographs of the neck dissection in layers are recommended.

Drowning occurs when water is inhaled, filling up the alveolar spaces and preventing gas exchange. The manner of death will vary depending on how the individual came to be in the water. If the event is not witnessed the manner often remains undetermined. Water in the lungs (pulmonary edema) and the paranasal sinus is often present. The degree of pulmonary edema may vary in a fraction of the cases due to the heart beating after respirations cease. As the heart beats, before eventual asystole, some of the fluid in the lungs will be absorbed. An adult may drown in a big pool or ocean but not in a small pool or bathtub unless *neurologically compromised or intoxicated*. Homicide victims are sometimes placed in a water-filled *bathtub* to wash away evidence. The story given may be, “I found him in the tub, and he must have drowned.” Common sense would dictate that an individual does not need to move very much to get his head above the water.

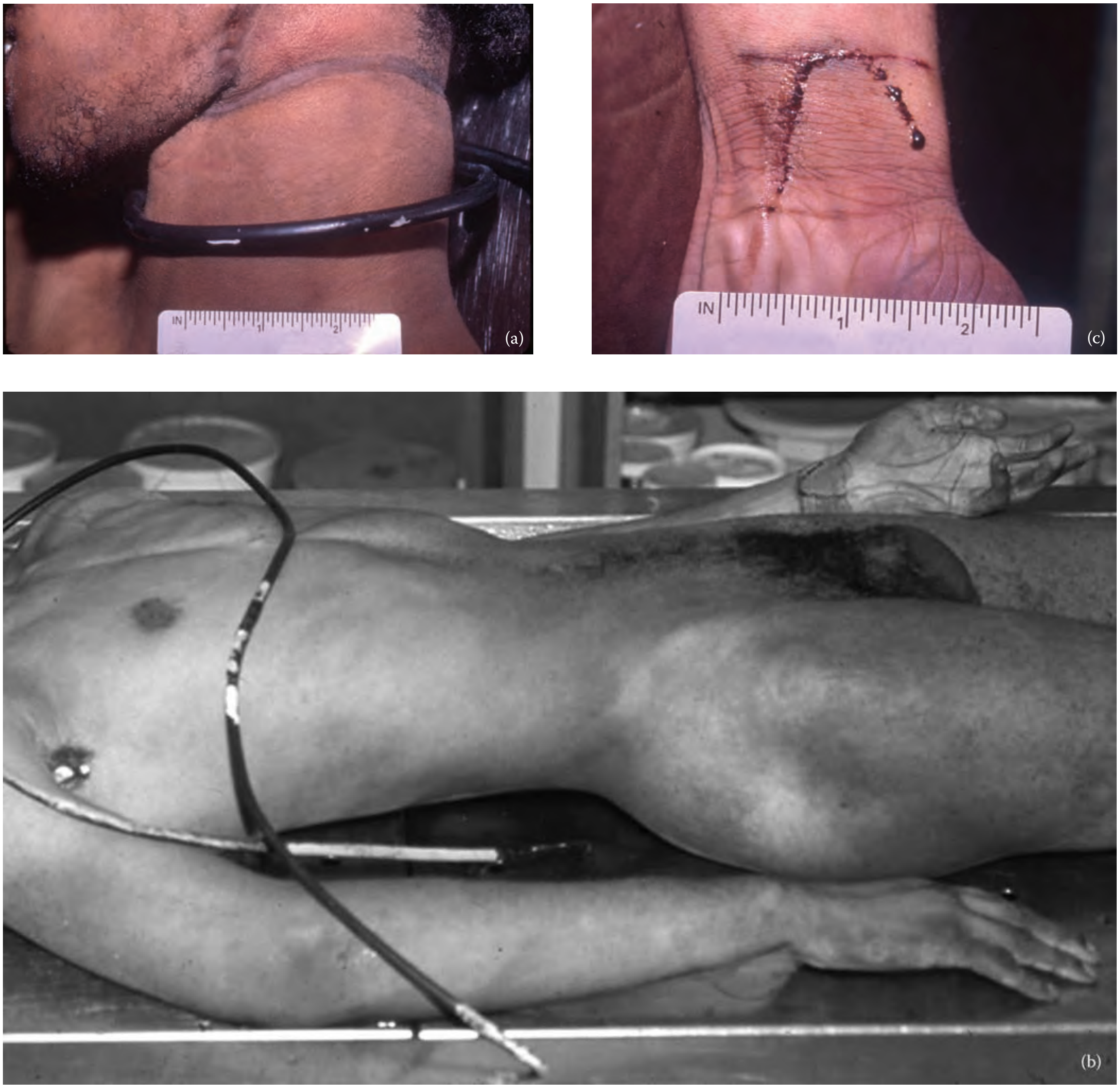

Suicidal hanging. Note the black cord (a) that matches the underlying furrow pattern at the superior aspect of the neck with upward extension at the left side of the face. Also note that the individual's lividity becomes more apparent and is fixed at the level of the upper thigh extending down to the feet (b). This fixed lividity pattern is appropriate for an individual who remained in an upright position for many hours after death. If his lividity was fixed posteriorly and not inferior; this would indicate prior scene alteration. There were hesitation marks at the wrist with blood seeping downward due to gravity (c).

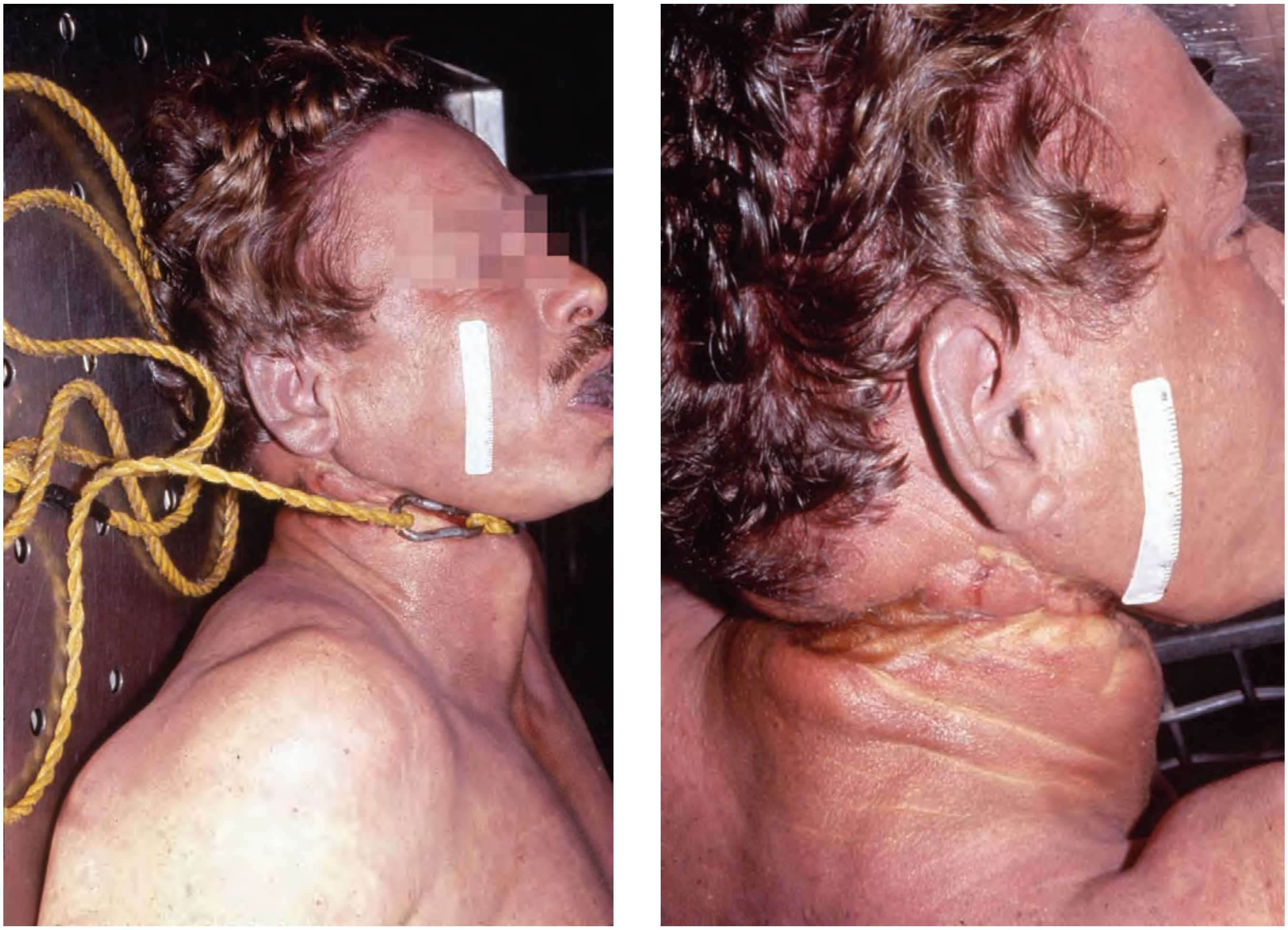

Suicidal hanging with matching rope and furrow pattern circling superior aspect of the neck forming a slight inverted "V" pattern behind the right ear.

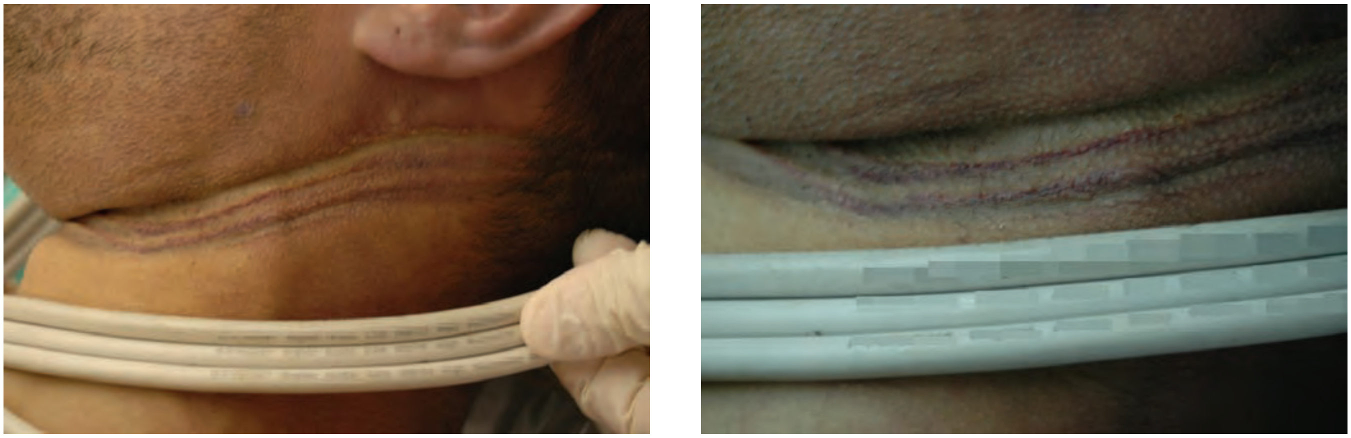

Suicidal hanging with computer cord.

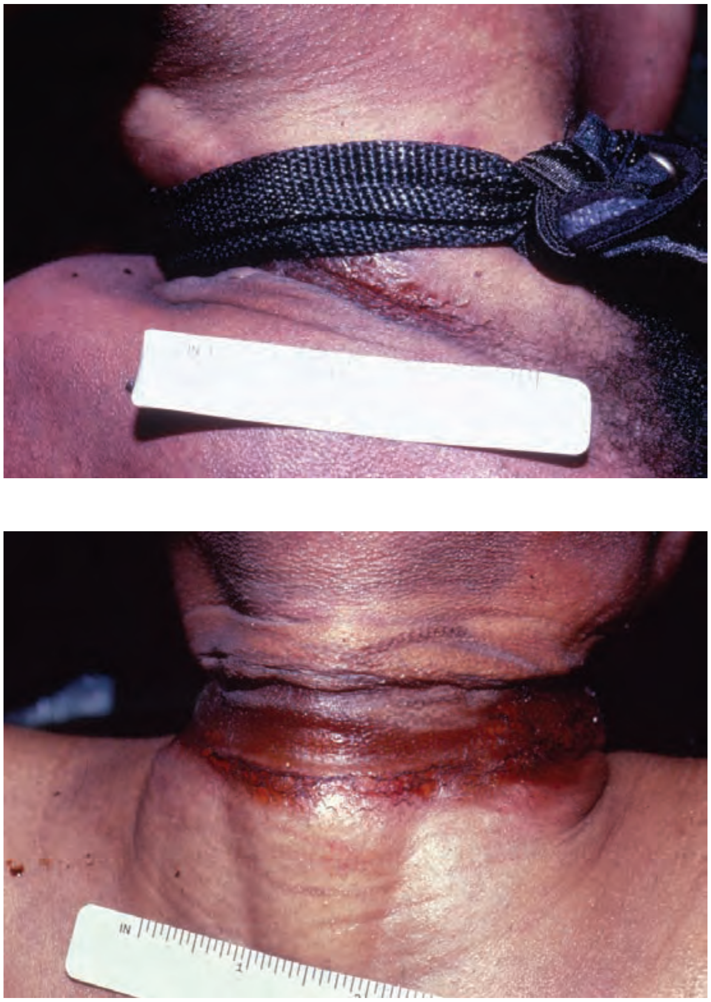

Suicidal hanging with nylon bag strap.

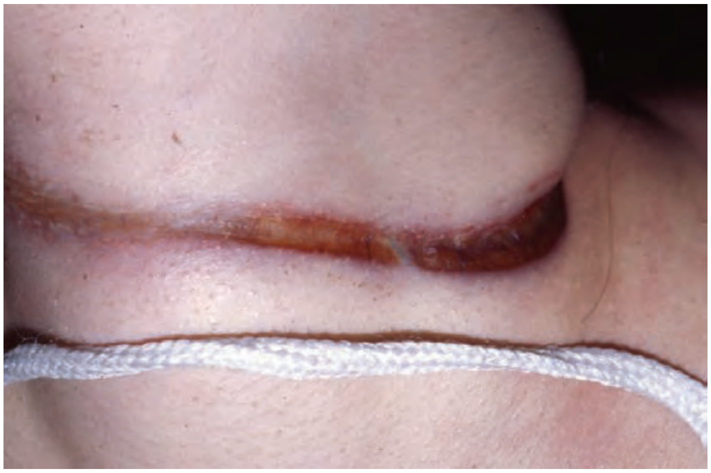

Accidental hanging in an individual with Alzheimer's dementia who enjoyed whirling around in circles while sitting on a desk chair. The person was left unattended for a short period of time by chronic care nursing staff, and got tangled in the cord from a window blind. This furrow pattern matches a blind cord as a ligature.

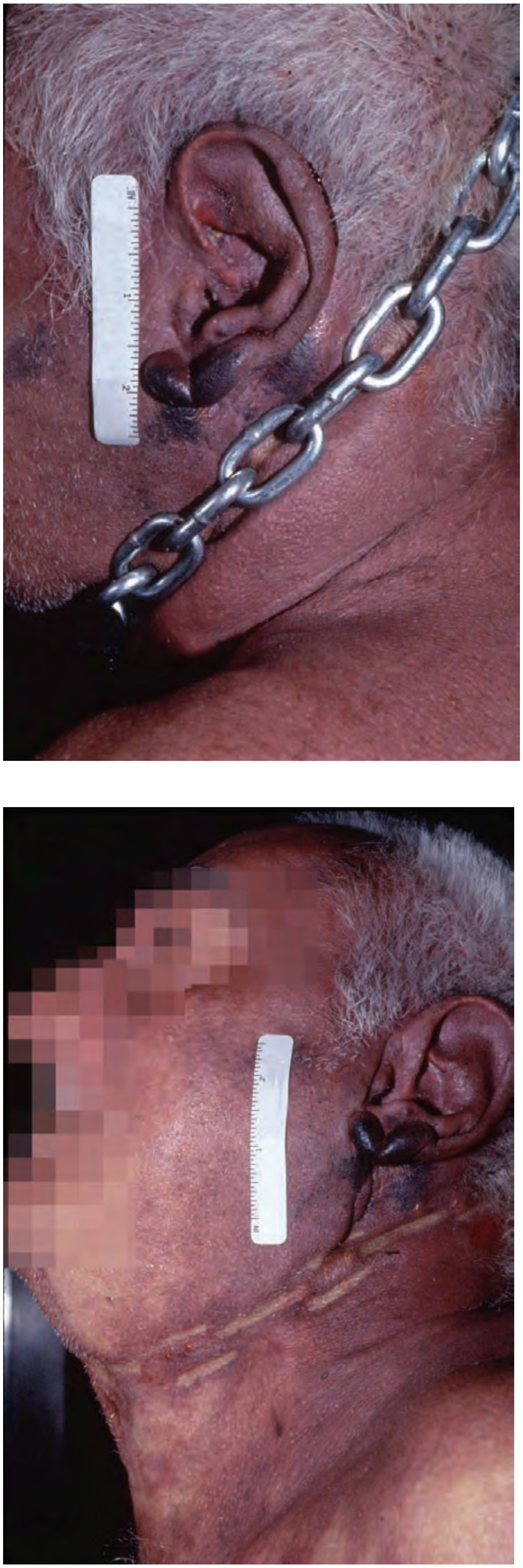

Suicidal hanging with chain.

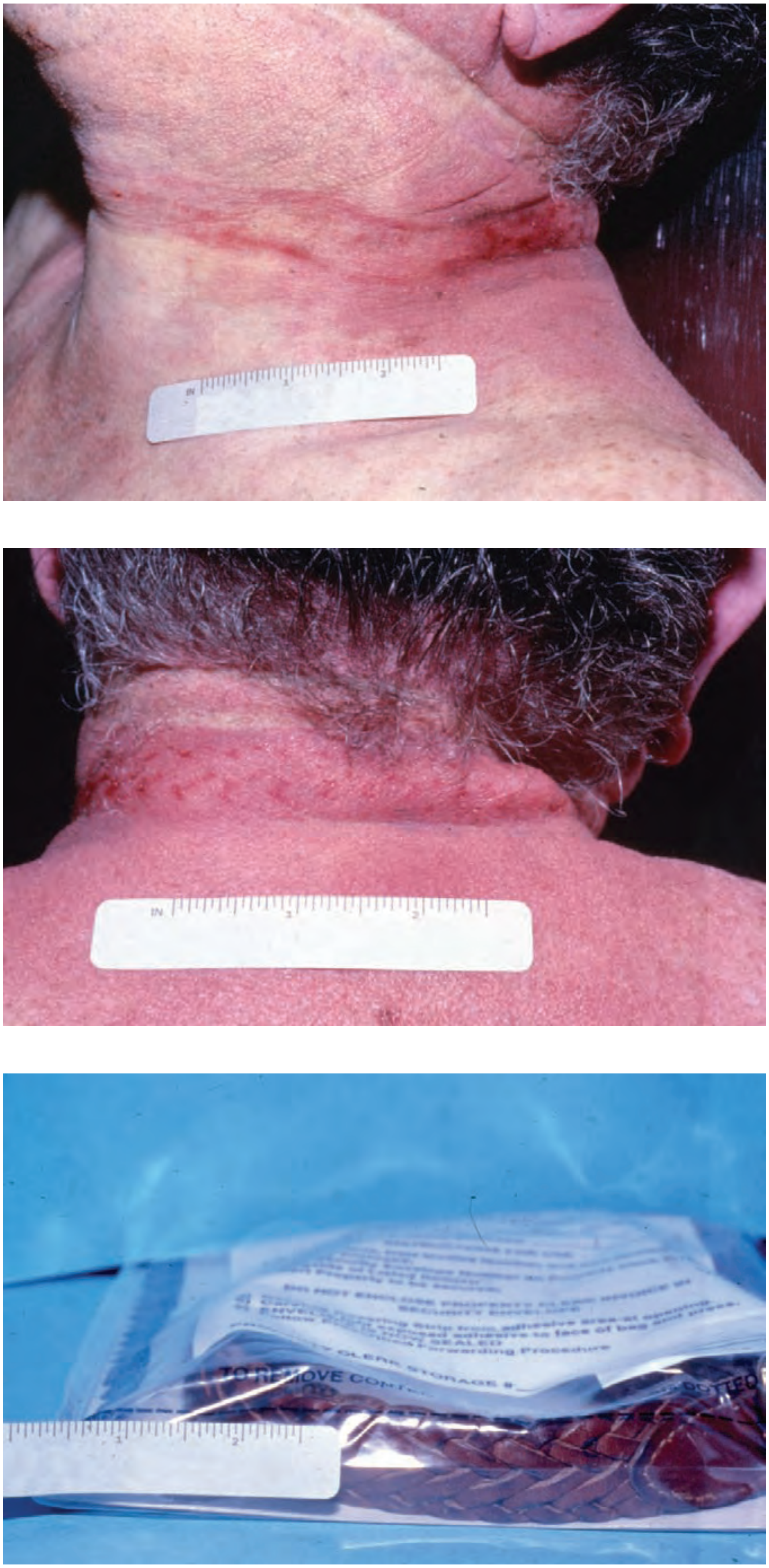

Suicidal hanging by braided belt. Note the matching skin pattern.

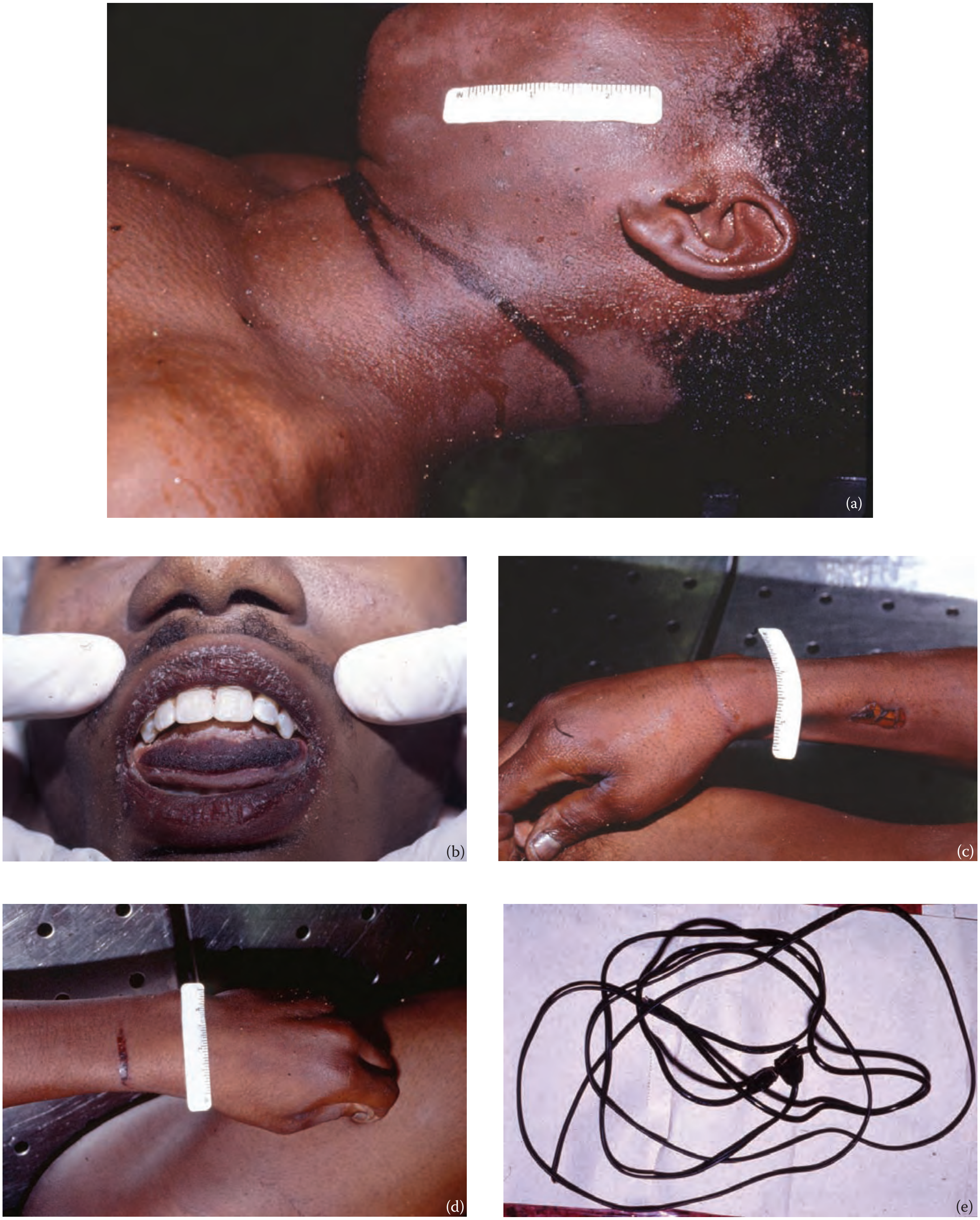

Suicidal hanging with earphone cord (a). This individual also took part of the cord and loosely wrapped his hands behind his back so he would not be able to reach up and prevent the hanging from being successful. This is not very unusual for people to hang themselves and loosely tie their hands in this fashion. Note the force of the ligature caused the tongue to protrude from his mouth (b). Also note the dark postmortem drying of the mucosa exposed to air.

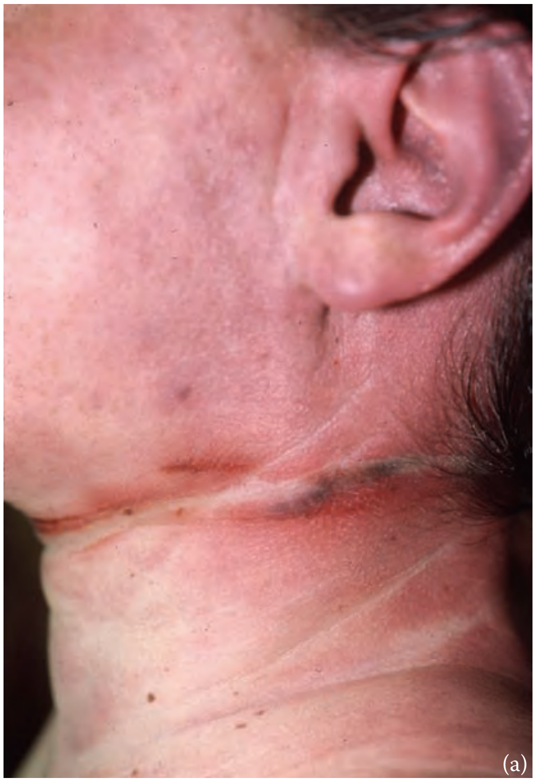

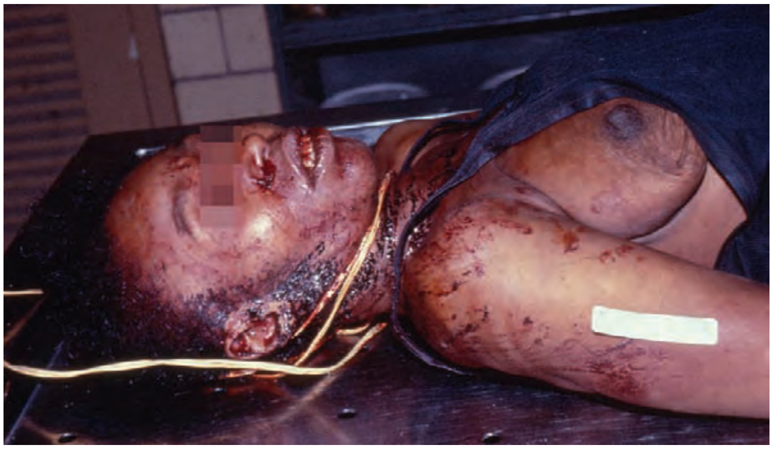

Homicidal ligature strangulation in a prostitute who had large amounts of cocaine and heroin in her system. Her fingers were also noted to be crushed by a pair of pliers during the assault.

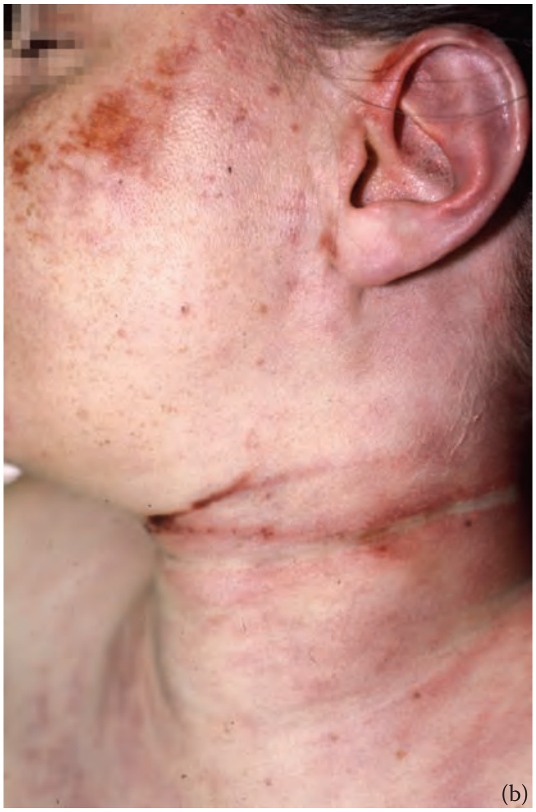

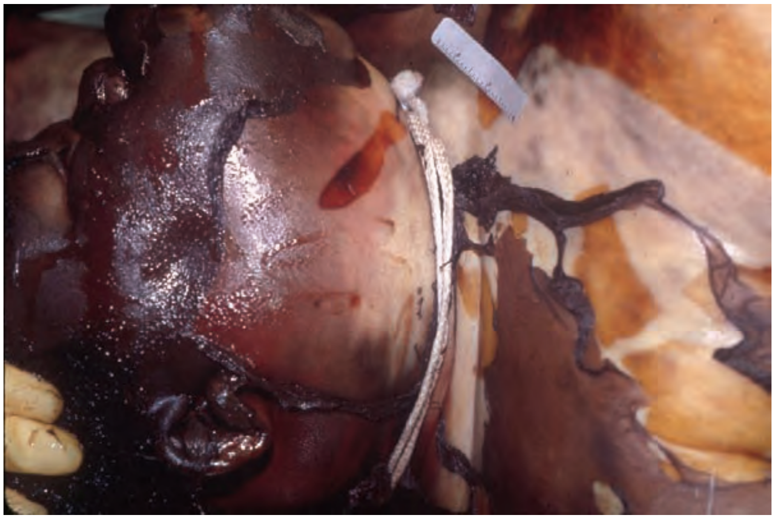

This is a suicidal ligature strangulation in an individual with a longstanding history of depression who was found in a locked secured apartment with a suicide note. Note the cord is tied tightly around his neck. There is also a moderate state of putrefactive change with skin slippage, bloating, and purging.

Homicidal ligature strangulations with horizontal cord marks surrounding the necks. Note the other injuries to the decedent's face in (b), including abrasions. Both of these victims were also sexually assaulted.

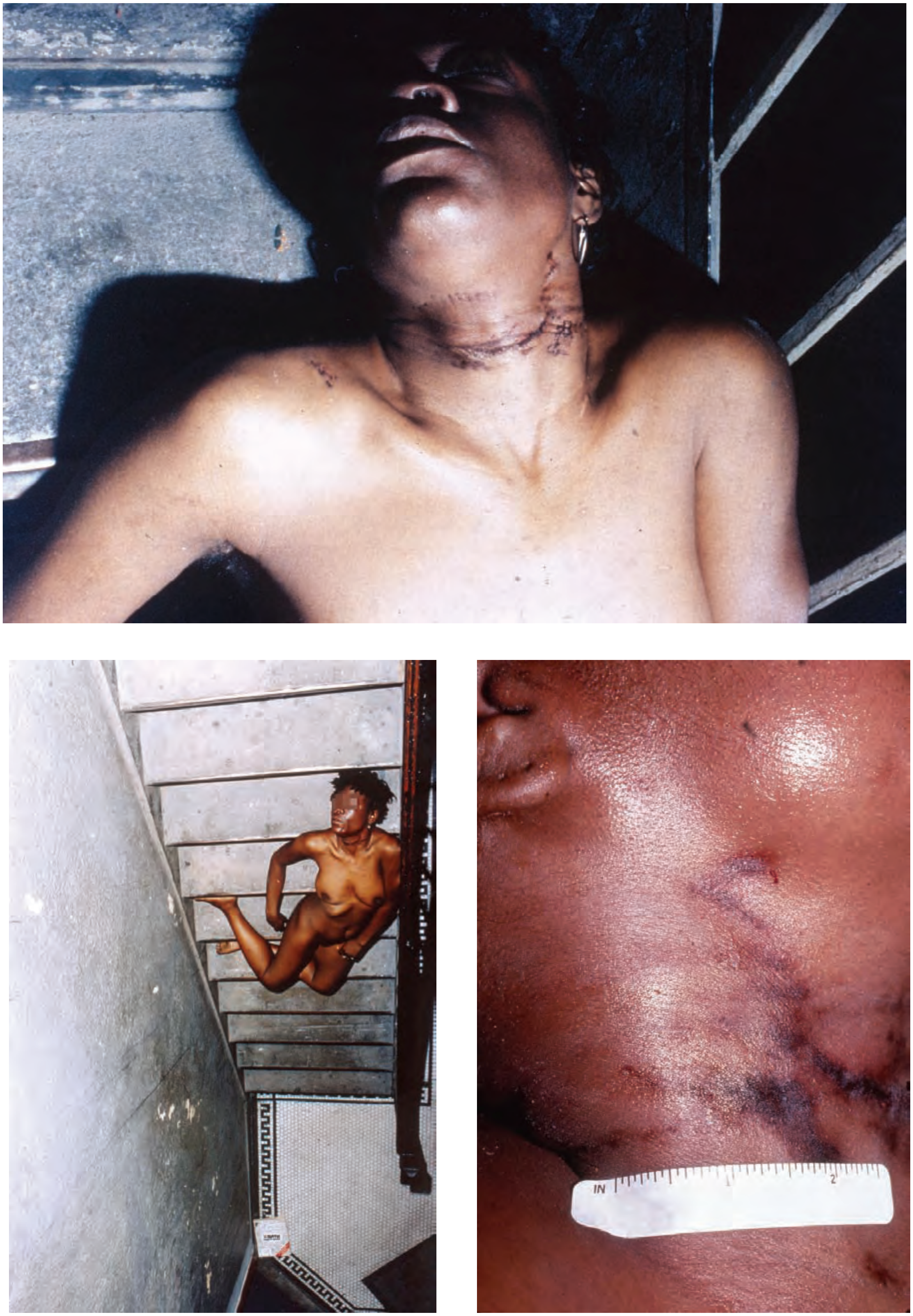

Manual strangulation. Note the multiple irregular abraded contusions surrounding the neck. There were multiple areas of strap muscle hemorrhage and a fractured hyoid bone. She was sexually assaulted and left in a stairwell. There were multiple petechiae with areas of hemorrhage in the sclerae and conjunctivae of each eye. The constellation of these findings are typical for a homicidal manual strangulation.

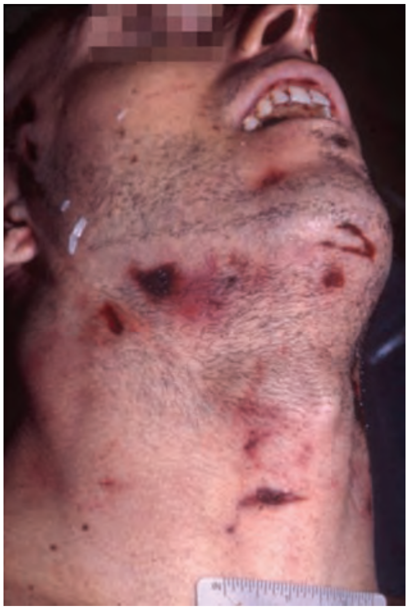

Manual strangulation. Note the abraded contusions to the anterior neck.

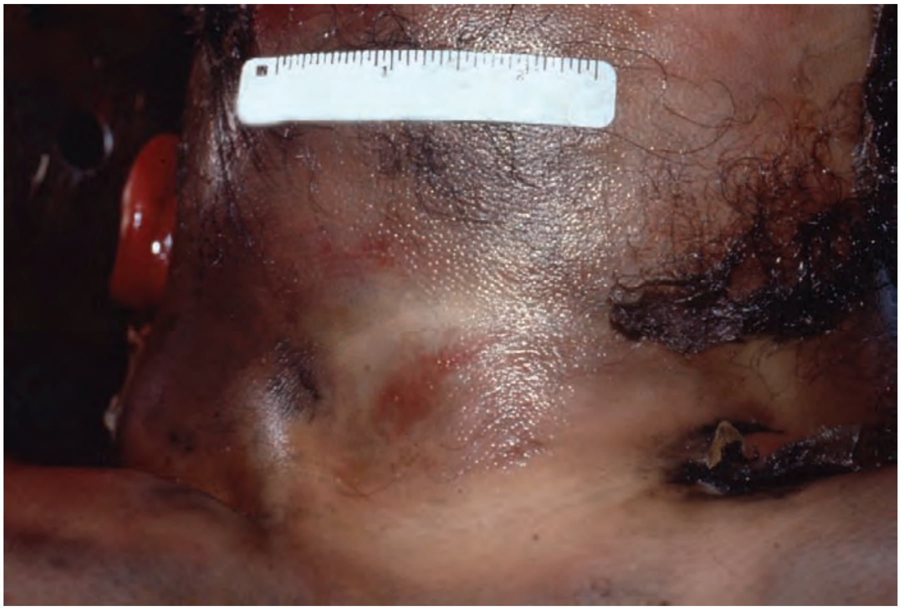

Manual strangulation. There is early decomposition with skin slippage. There are vague abrasions and superficial contusions at the neck. As decomposition progresses it may become more difficult to interpret these findings.

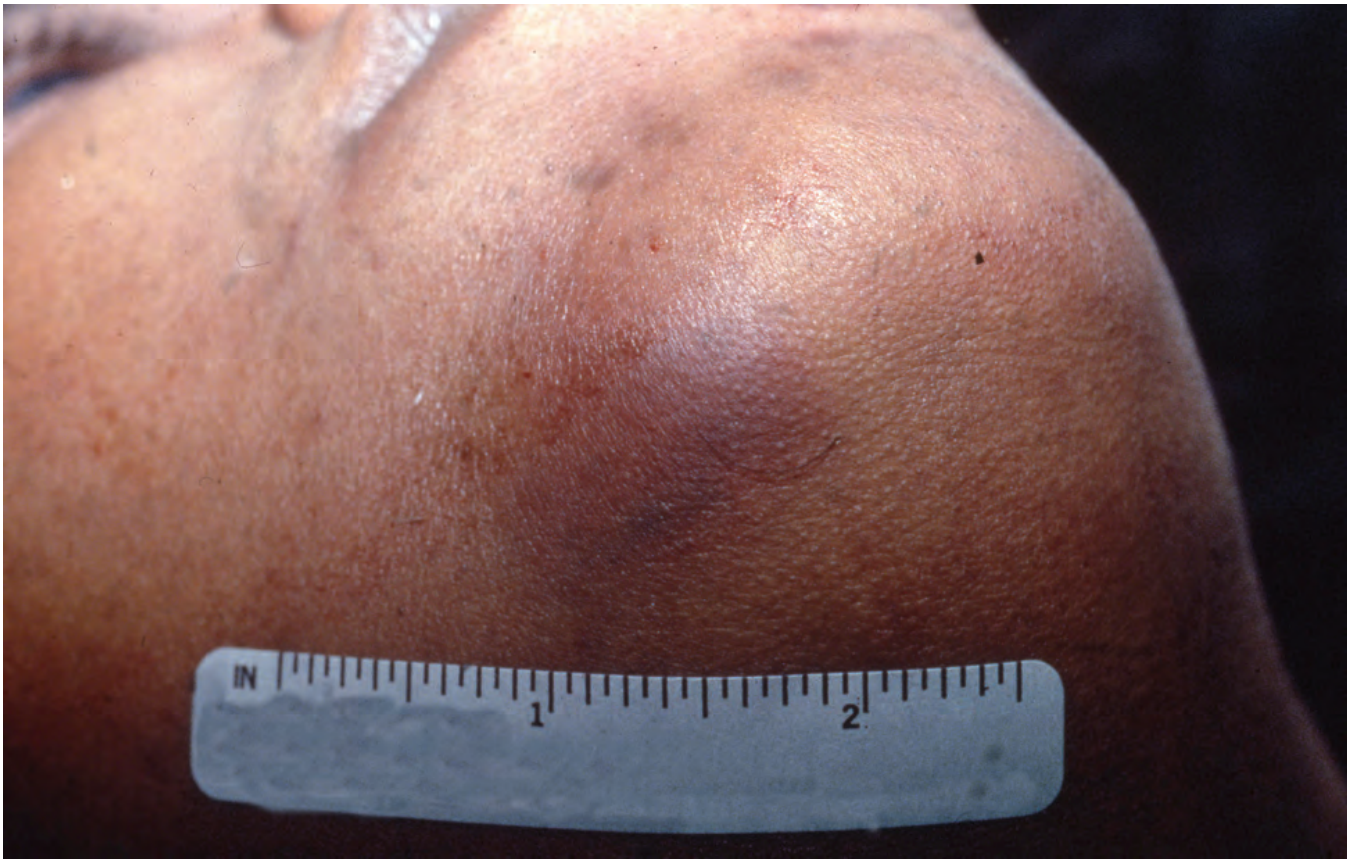

This individual was punched in the face and then yoked from behind during a sexual assault. There were no visible external injuries to the neck except this contusion at the border of the right chin. Careful internal examination after the brain and visceral organs had been removed revealed areas of hemorrhage within the anterior strap muscles and posterior paraspinal muscles. There were also petechiae with hemorrhage of the sclera and conjunctivae. This individual was found face down and the eye hemorrhages were originally thought by some to be associated with postmortem lividity. The perpetrator was caught after bragging about the assault.

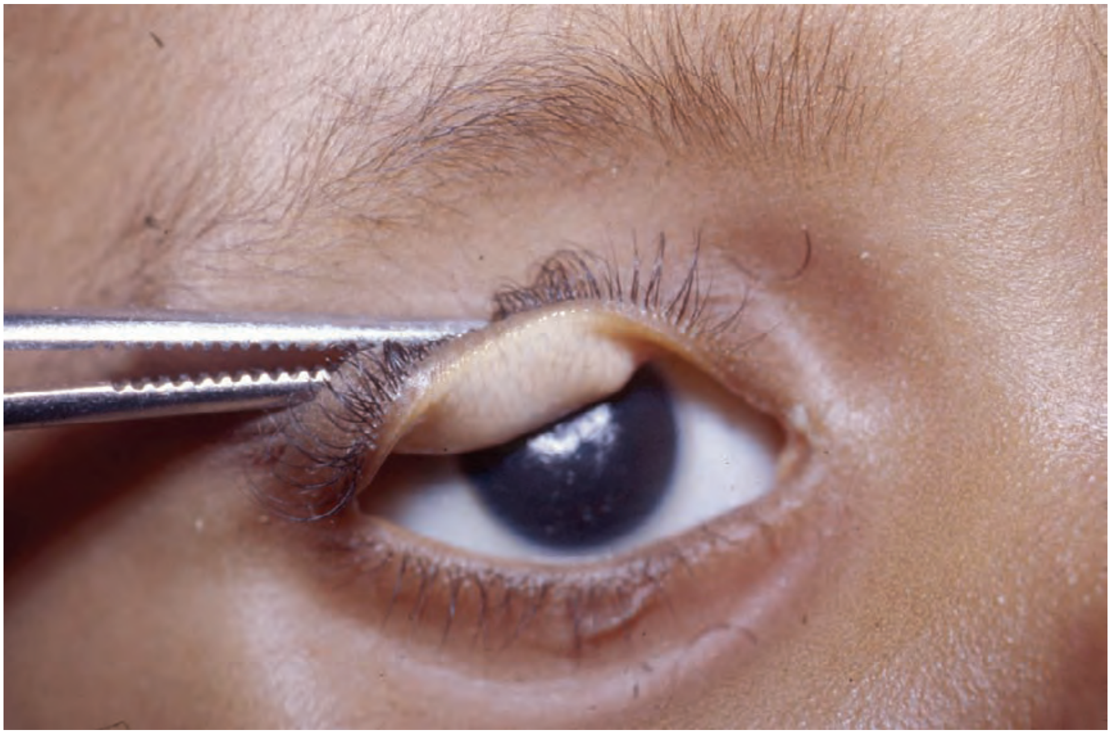

This image demonstrates an eye from a person who hanged him or herself. Note there are no petechiae or scleral/conjunctival hemorrhages. The decedent's weight pulling down on the ligature produced a significant enough force to obstruct the entire blood supply to the head. Therefore, there was no pressure–release mechanism, leading to capillary rupture and hemorrhage. The absence of scleral or conjunctivae hemorrhages is more typical in hanging fatalities.

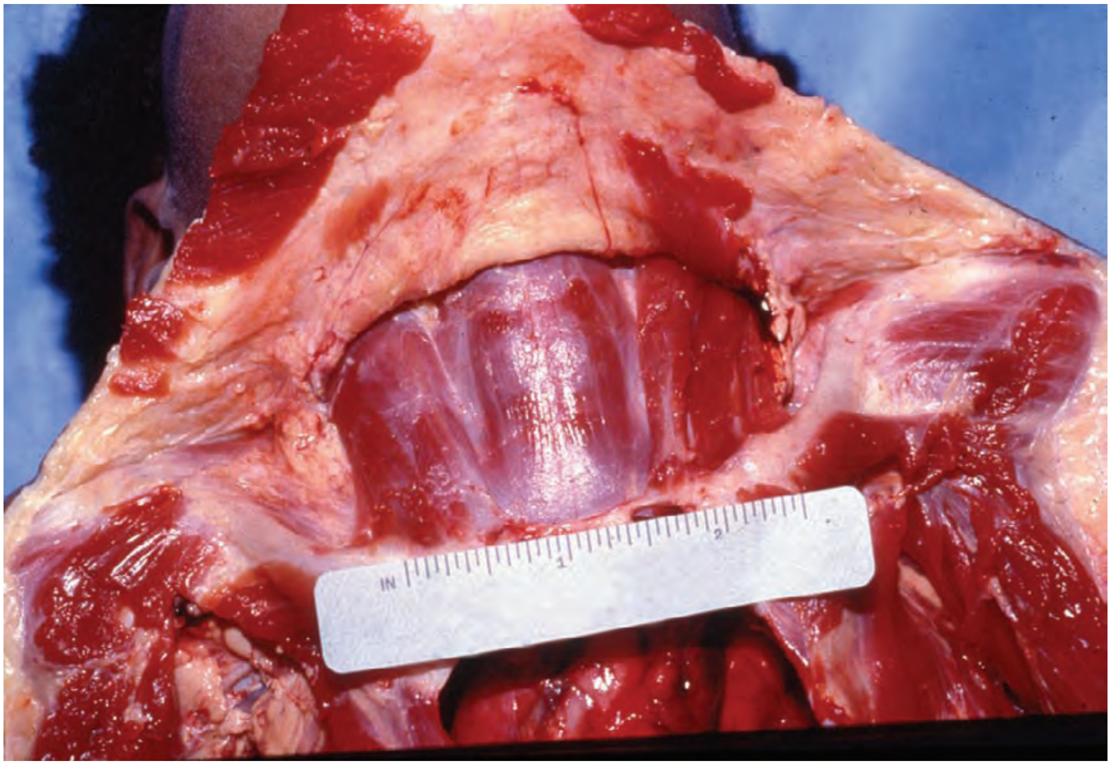

Anterior neck dissection after the visceral organs and brain have been removed. Part of the undersurface of the platysma muscle is visible at the top of this image adjacent to the yellow subcutaneous tissue of the neck. The anterior strap muscles are visible directly above the label and are free of antemortem injury. There were no hemorrhages or fractures of the remaining neck structures. The absence of these injuries is common in hanging fatalities.

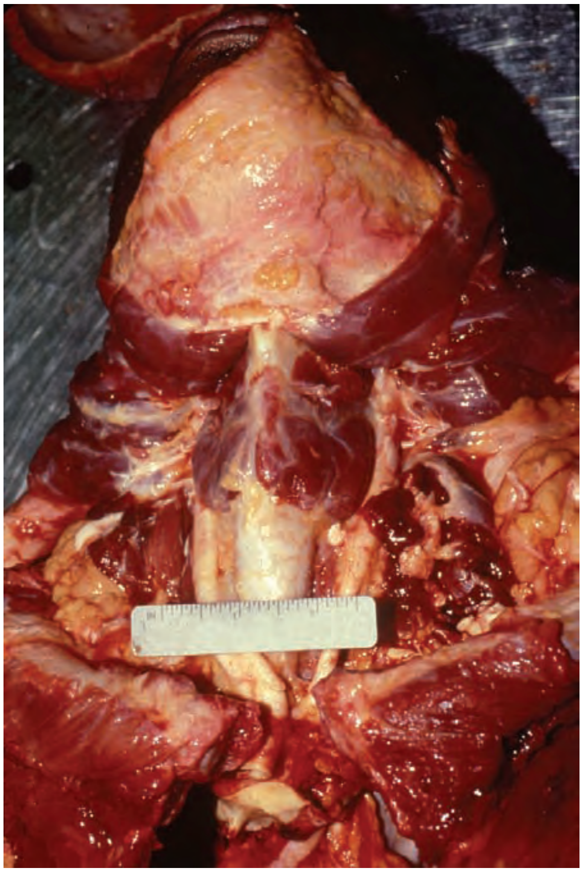

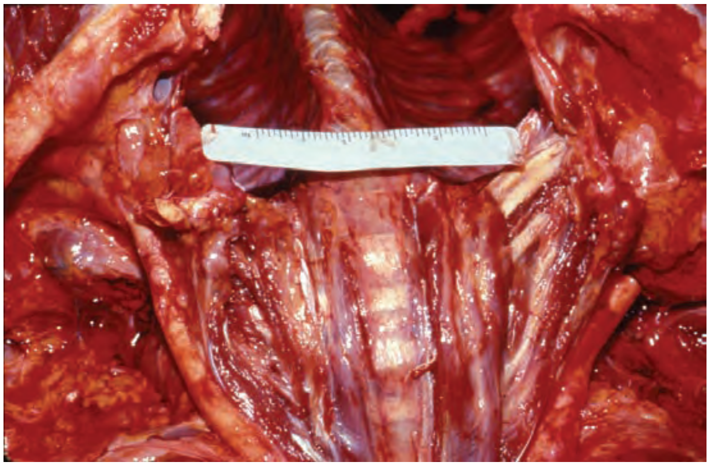

These figures demonstrate anterior neck dissections with no hemorrhages. The absence of hemorrhage is typical for most hangings.

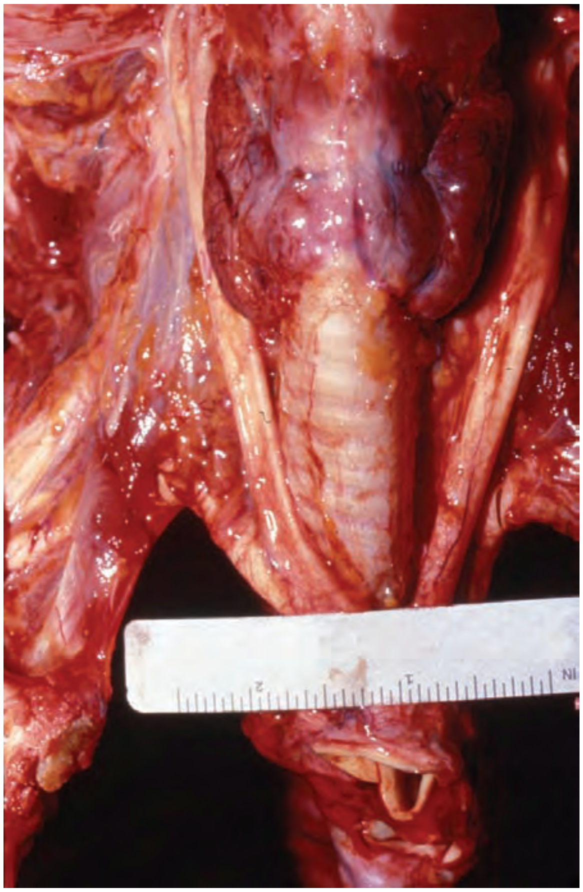

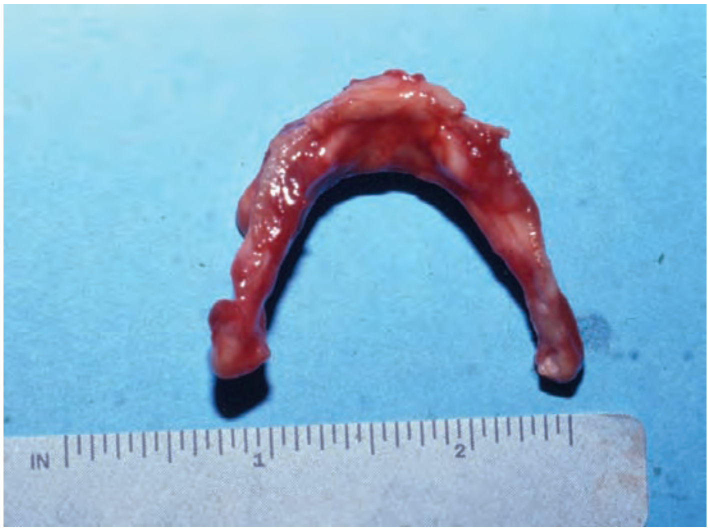

This demonstrates a normal hyoid bone with no fractures or contusions.

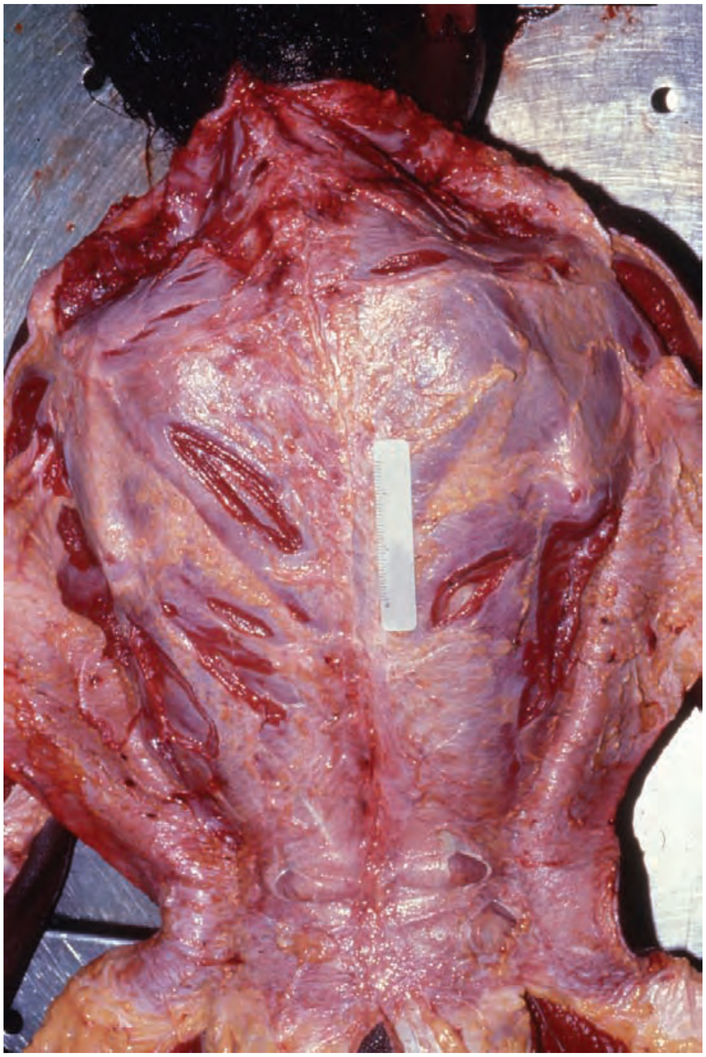

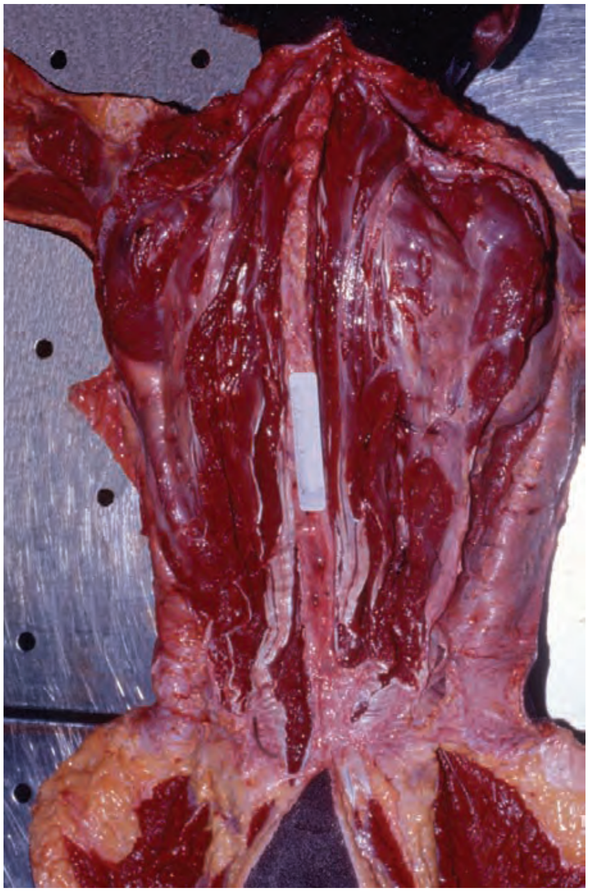

Some cases may require further evaluation including a posterior neck and back dissection. This case involved a 12-year-old African American child who was strangled. Dissection should be done in a layer-by-layer fashion until the surface of bone is exposed. It is important to dissect the arms as well, which may demonstrate contusions from being held during a struggle. There were no hemorrhages found in this case. It is often more difficult to externally visualize contusions in darker-skinned individuals.

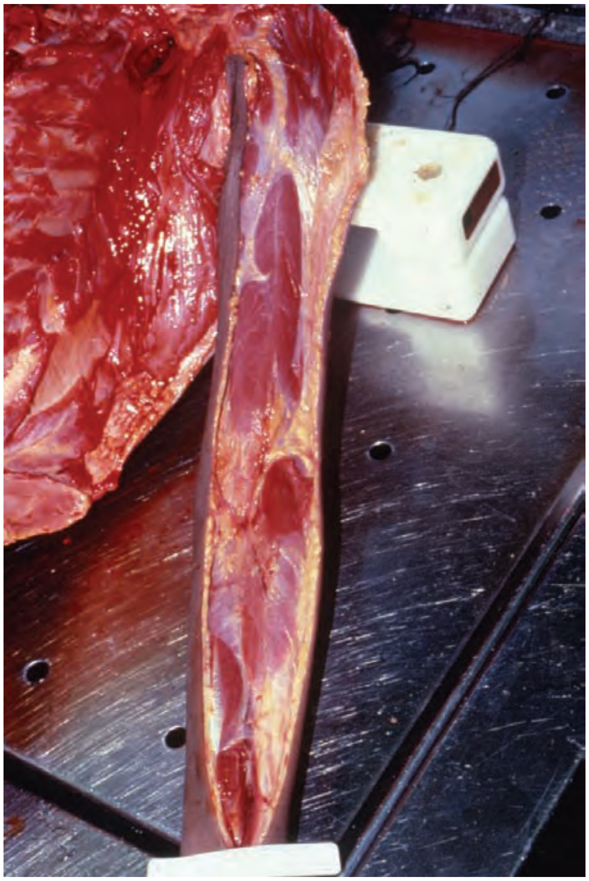

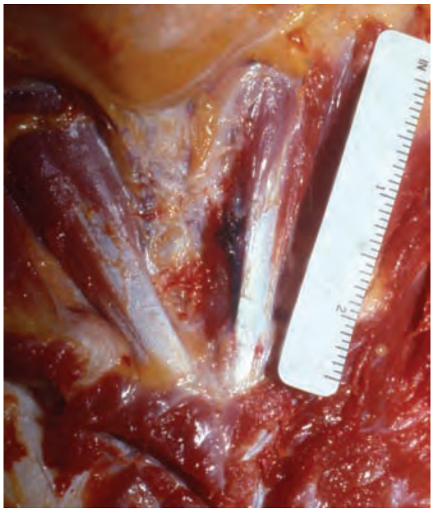

Strangulation. Hemorrhage of the medial left sternocleidomastoid muscle.

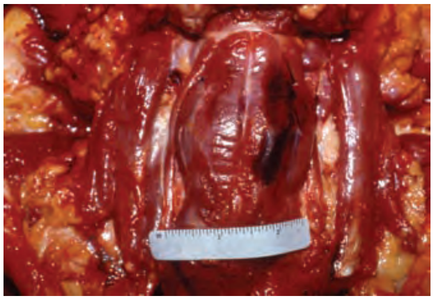

Strangulation. Note the large hemorrhage to the anterior neck structures including the left sternohyoid muscle.

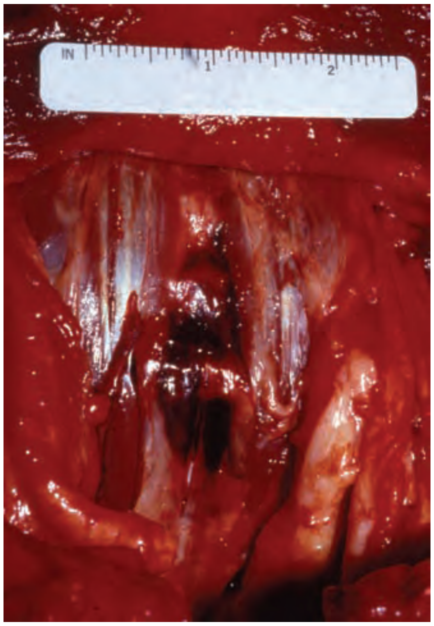

Strangulation. Note the hemorrhage in the anterior neck structures.

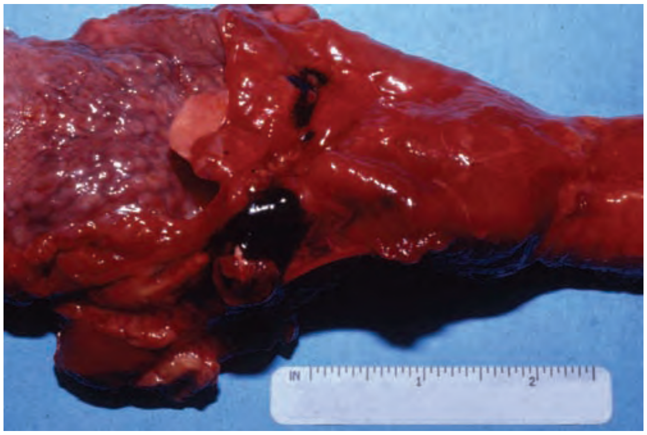

Strangulation. Note the hemorrhage overlying the left superior horn of the thyroid cartilage. The underlying cartilage was fractured.

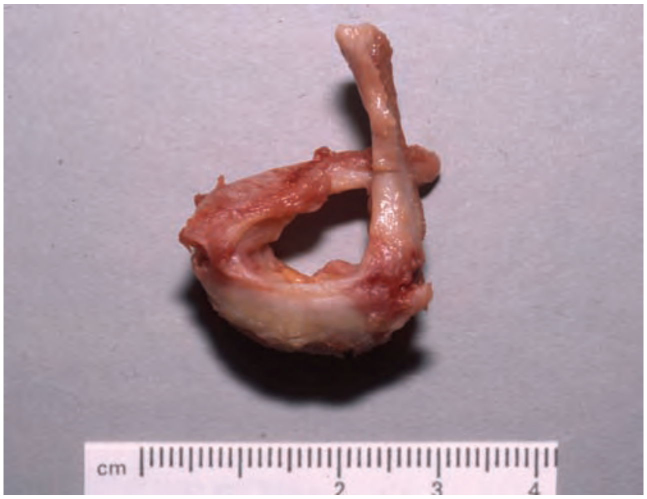

Note the fractures of this hyoid bone with vague hemorrhages in a decedent with slight to moderate decomposition. Fractures of the hyoid bone are often found in association with homicidal strangulation. The presence of a hyoid bone fracture does not indicate that the case must be a strangulation and the absence of fractures to the hyoid bone does not indicate the decedent was not strangled. Fractures can occur as a result of a blunt impact as well.

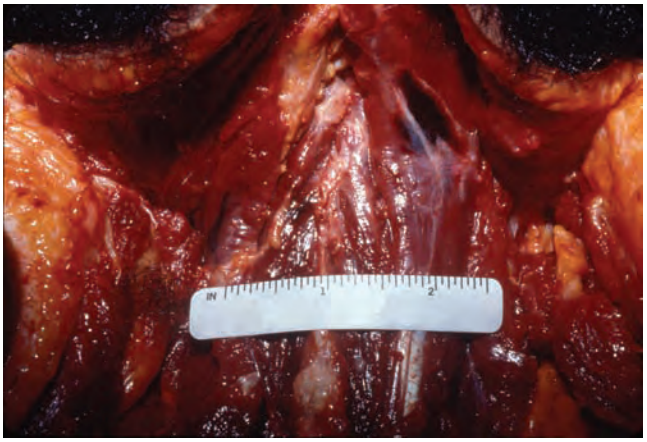

Strangulation. Posterior neck dissection with hemorrhage to the superior aspect of the semispinalis capitis muscle.

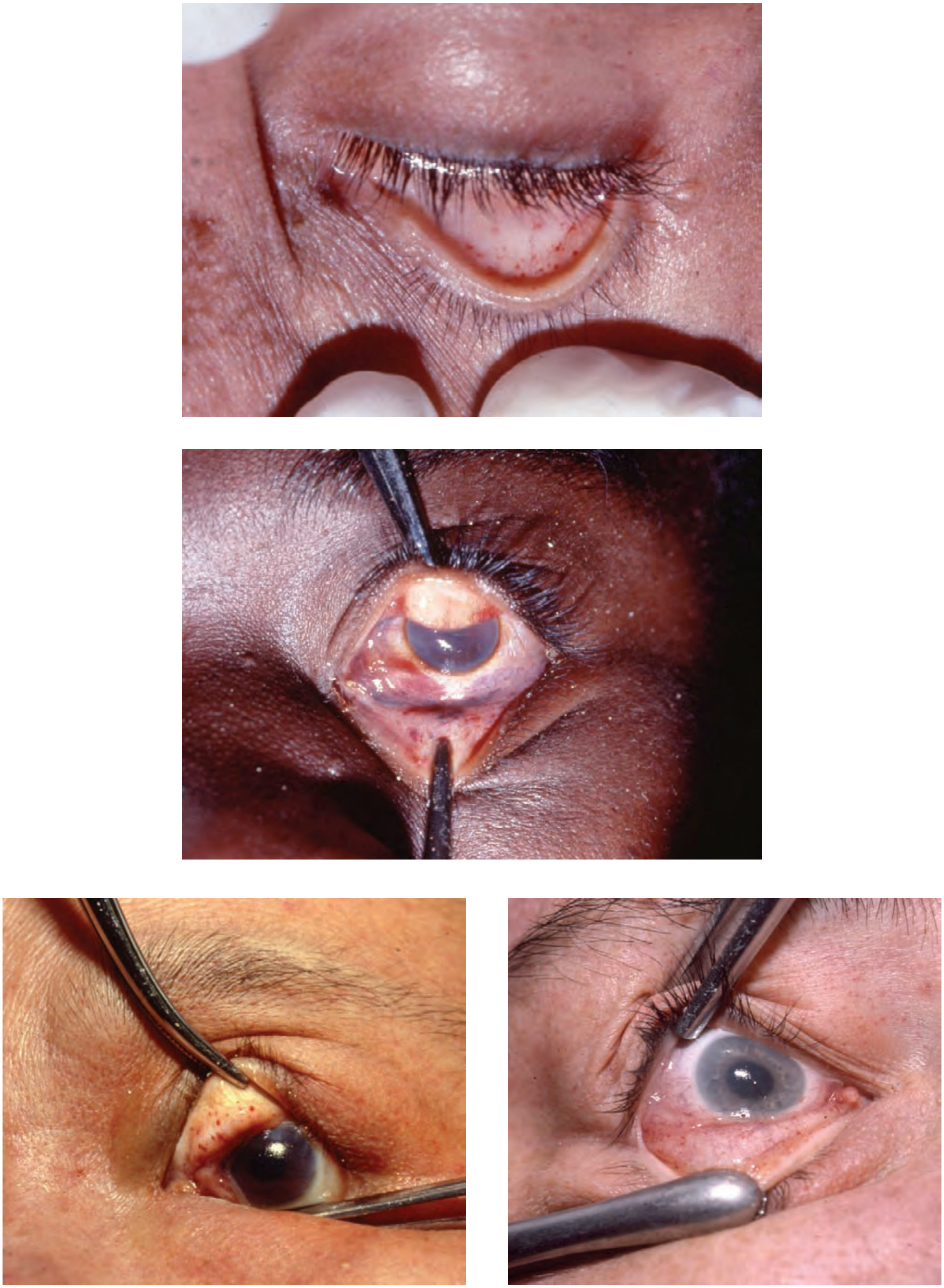

Cases of strangulation demonstrating petechiael hemorrhages of the sclera and conjunctivae.

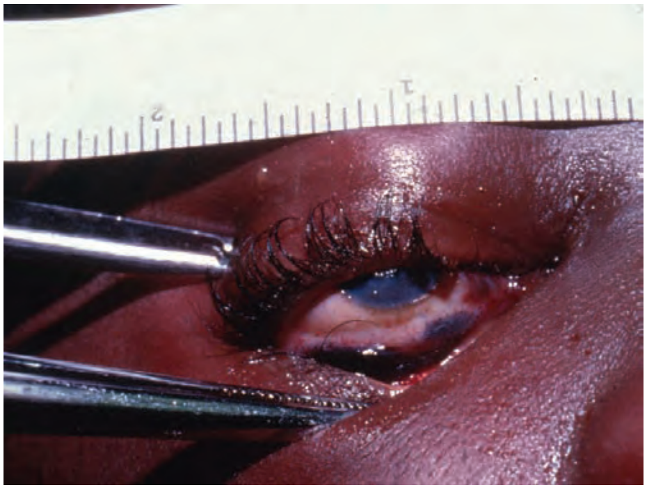

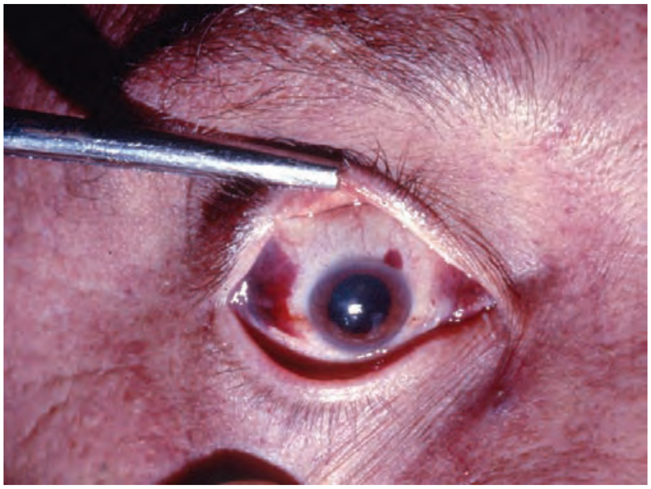

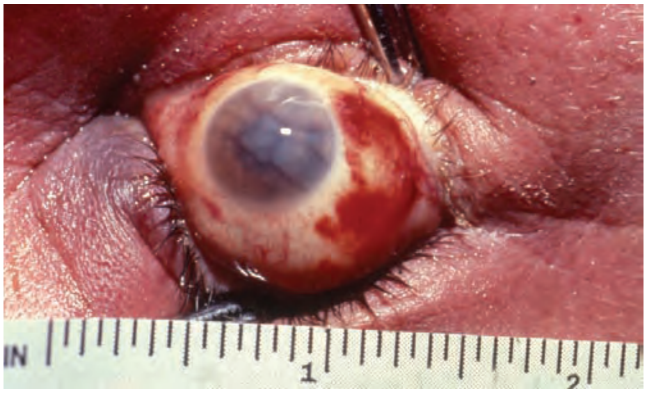

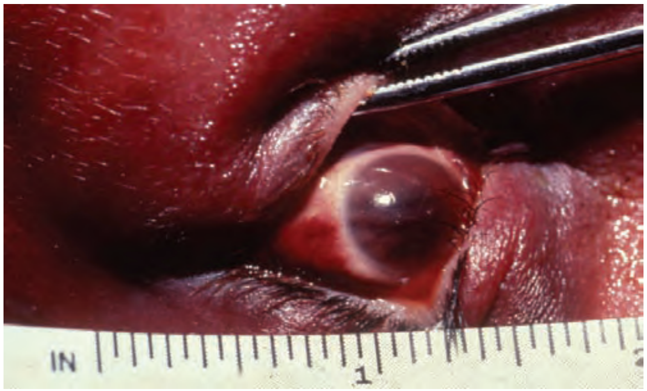

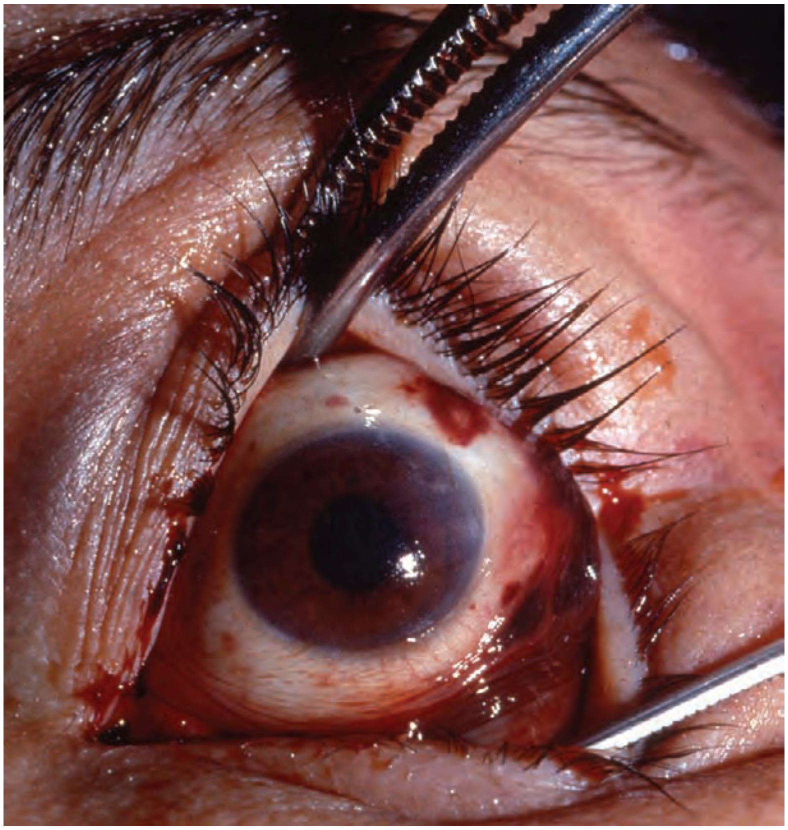

These demonstrate a range of scleral and conjunctival hemorrhages. As varied degrees of neck pressure continue after petechiae formation these hemorrhages progressively become larger and possibly confluent until death ensues.

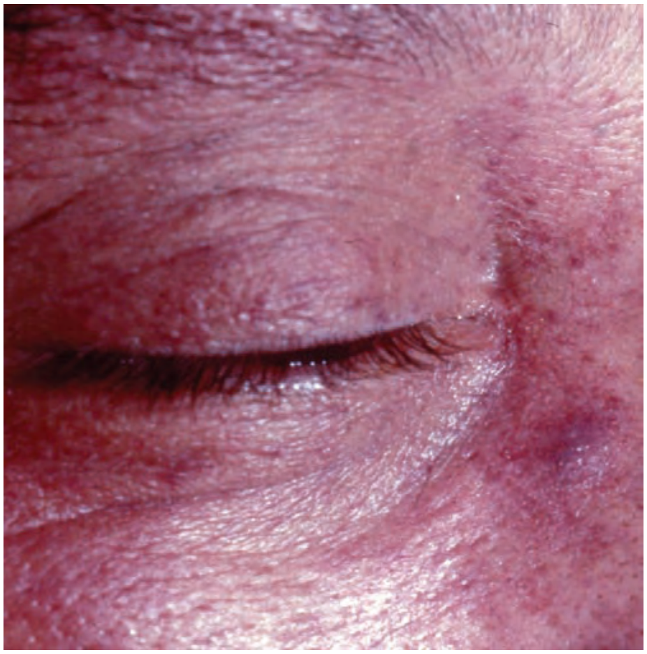

Petechiael hemorrhages of the face.

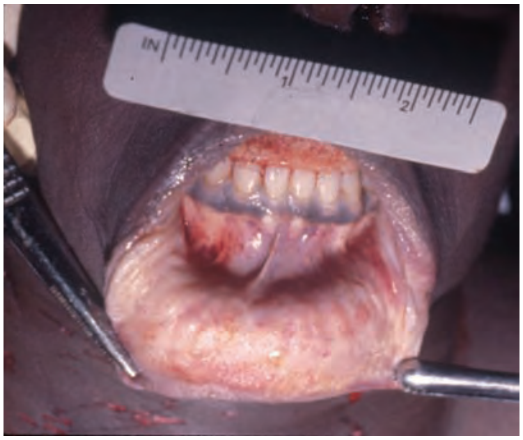

Strangulation. Petechiael hemorrhages within the mucosal surface of the mouth.

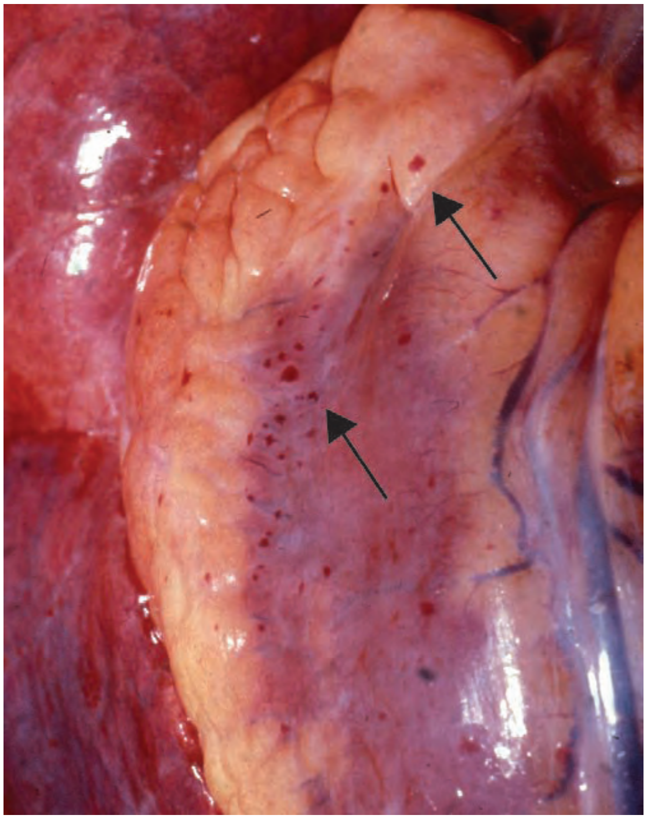

Petechiael hemorrhages at the surface of the heart associated with asphyxia due to chest compression.

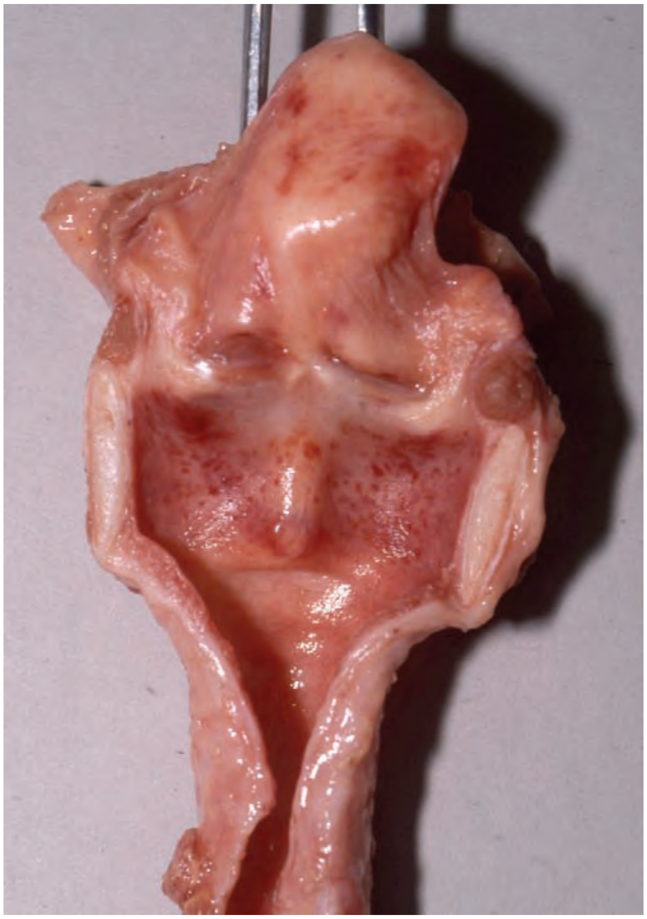

Strangulation. Petechiael hemorrhages within the mucosa of the tracheal and laryngeal cartilage.

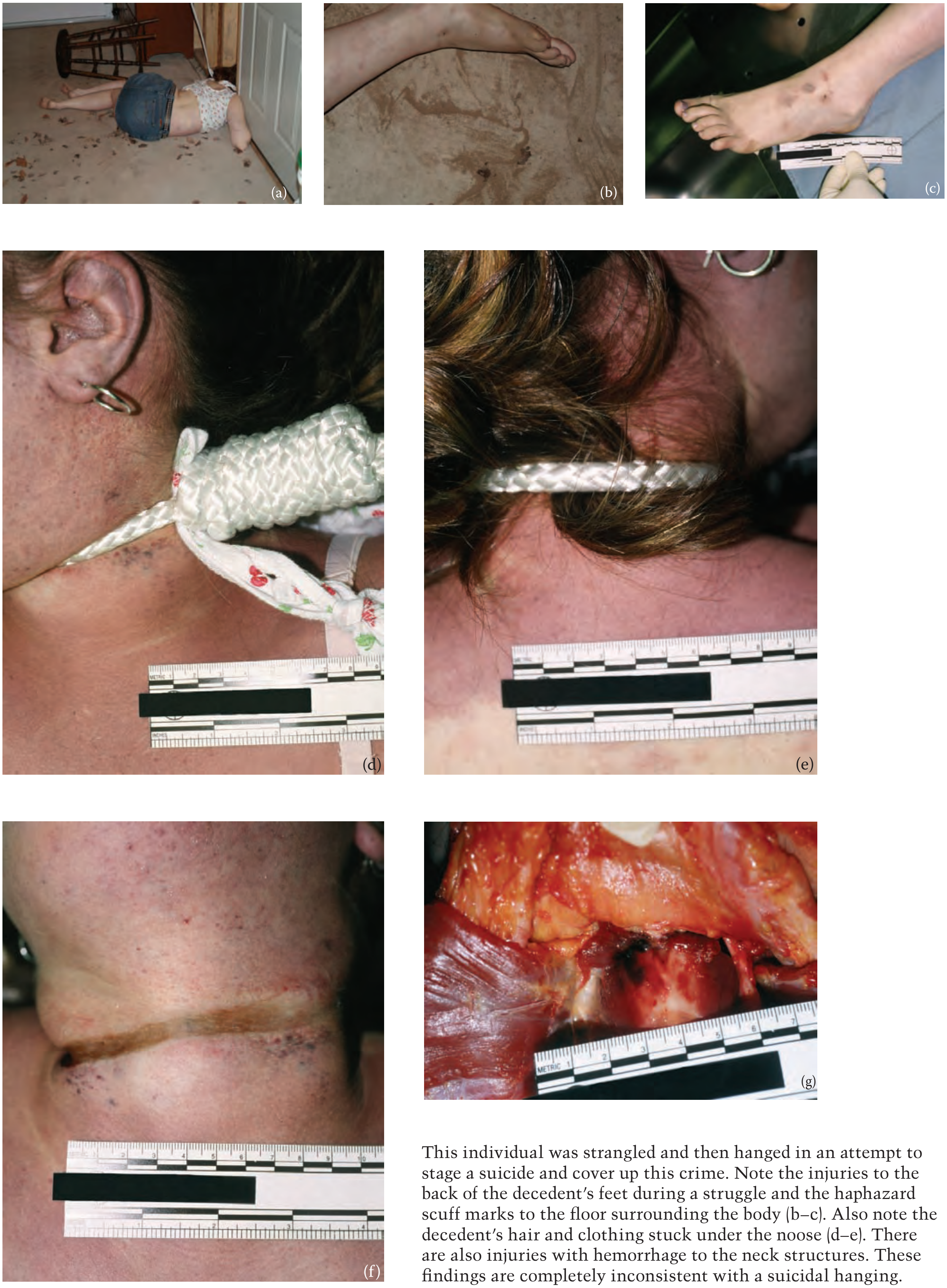
(a)
(b)
[redacted]
(c)
[redacted]
(d)
[redacted]
(e)
[redacted]
(g)
[redacted]
(f)

This individual was strangled and then hanged in an attempt to stage a suicide and cover up this crime. Note the injuries to the back of the decedent's feet during a struggle and the haphazard scuff marks to the floor surrounding the body (b–c). Also note the decedent's hair and clothing stuck under the noose (d–e). There are also injuries with hemorrhage to the neck structures. These findings are completely inconsistent with a suicidal hanging.

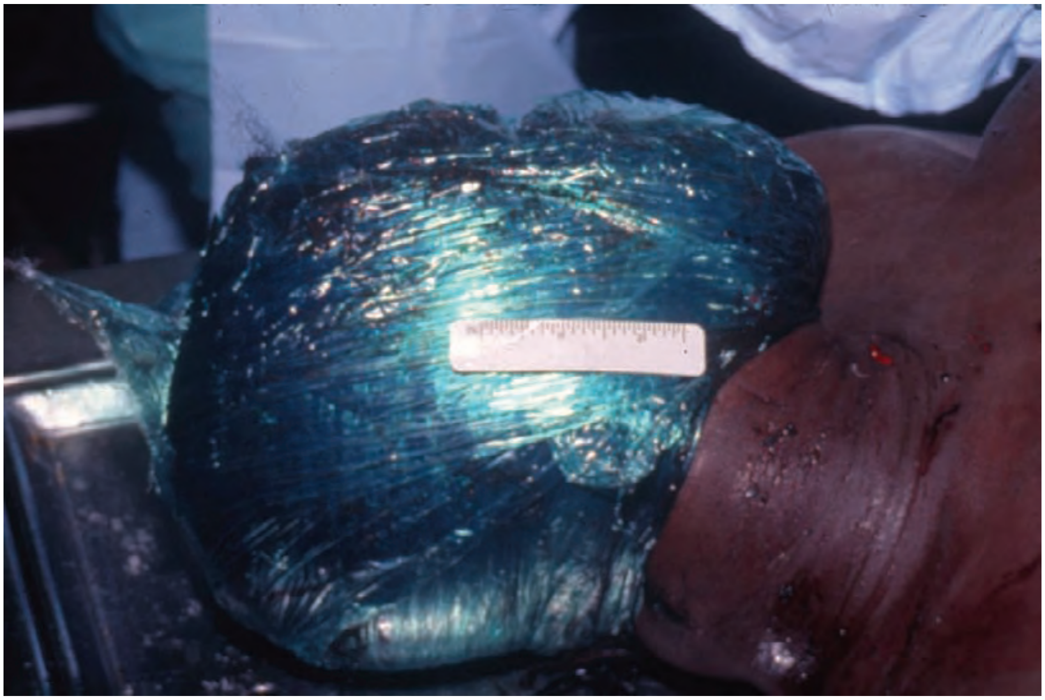

This individual was abducted, bound, and had multiple superficial stab wounds associated with a sexual assault. Her head was wrapped in multiple layers of plastic that covered her mouth and nose.

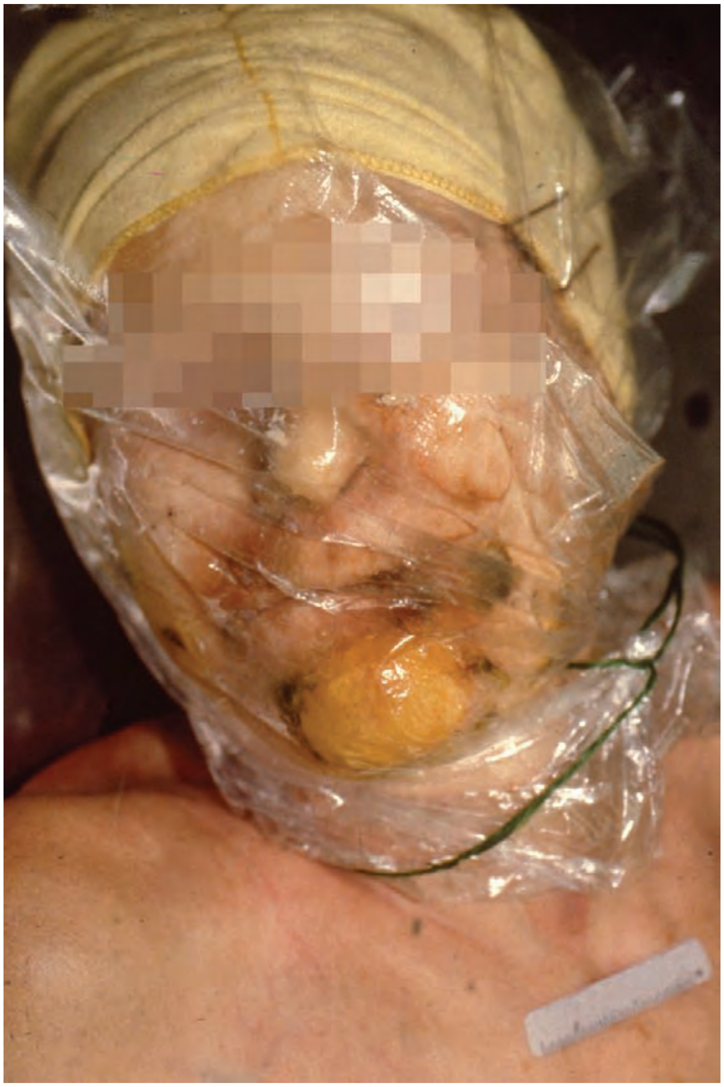

This individual committed suicide by taking multiple pills and tying a plastic bag tightly over her head. It is the author's experience that it is better to pend these cases for toxicology.

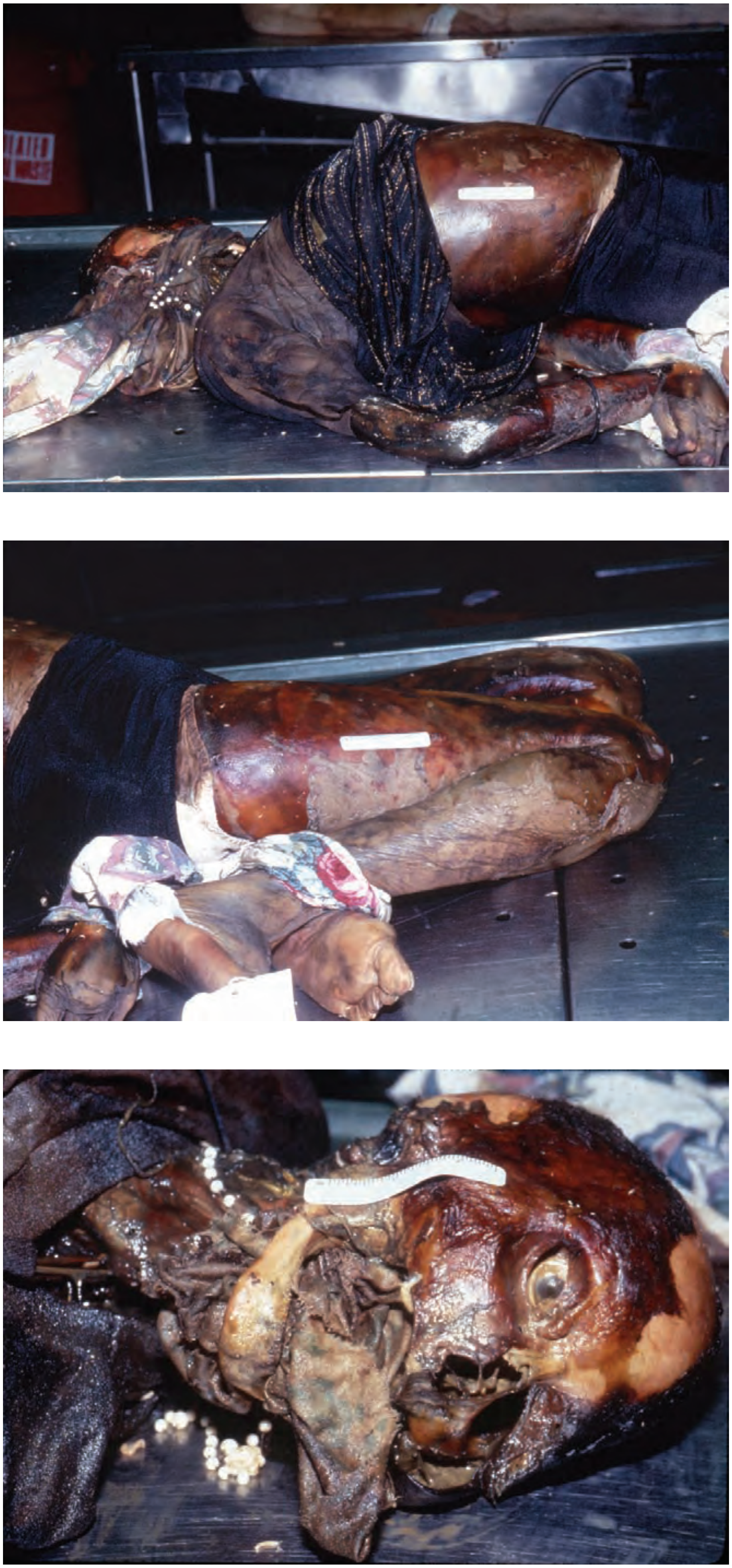

Homicidal asphyxia due to airway obstruction. This individual was found in a moderate to marked state of decomposition within her apartment. She was reportedly dealing drugs and was found tied up with a piece of cloth stuck down her mouth and throat.

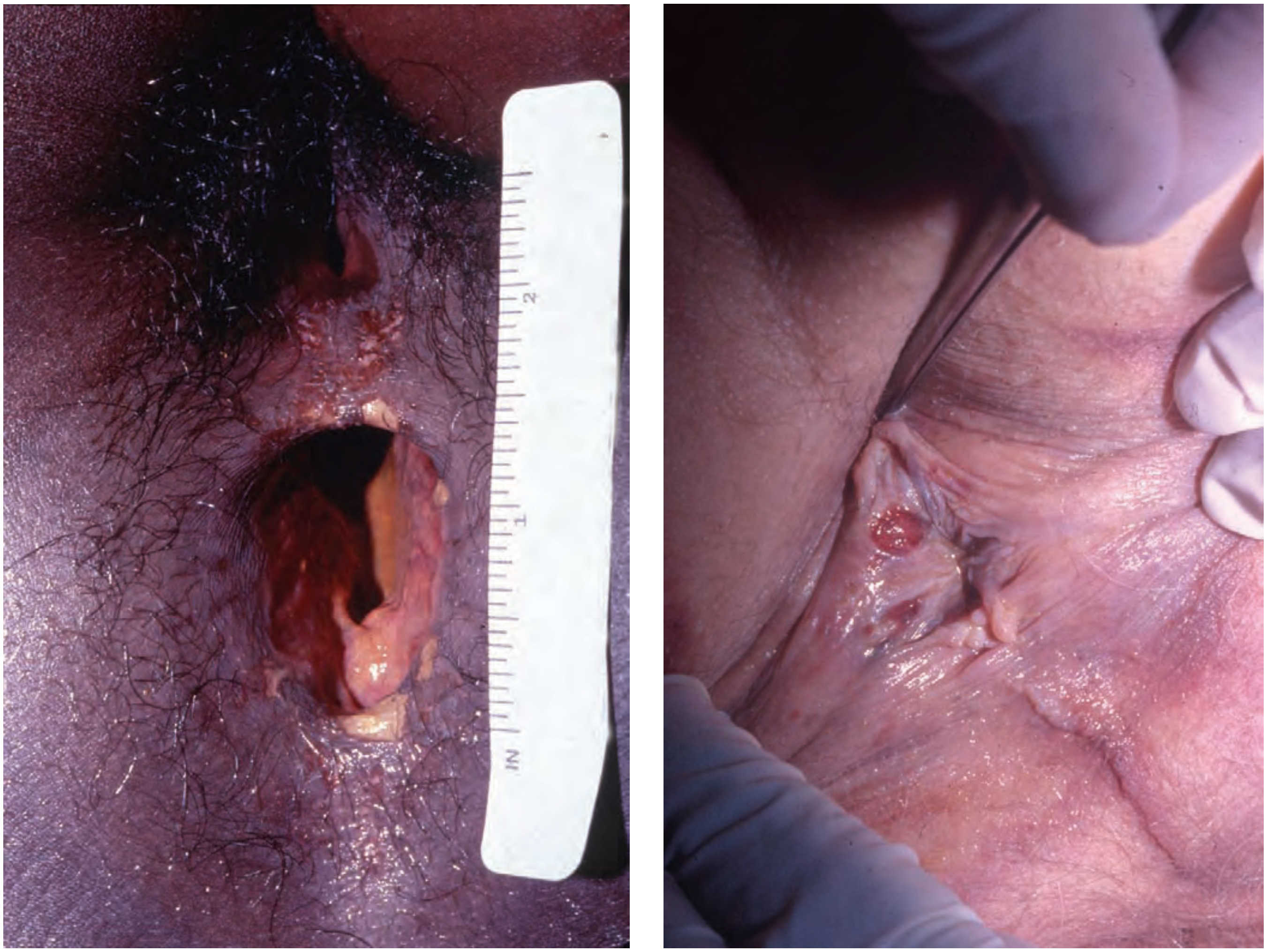

Anal trauma with abrasions and lacerations due to sexual assault during strangulation. One should always assume strangulation victims have been sexually assaulted. A rape kit should always be performed in these cases.

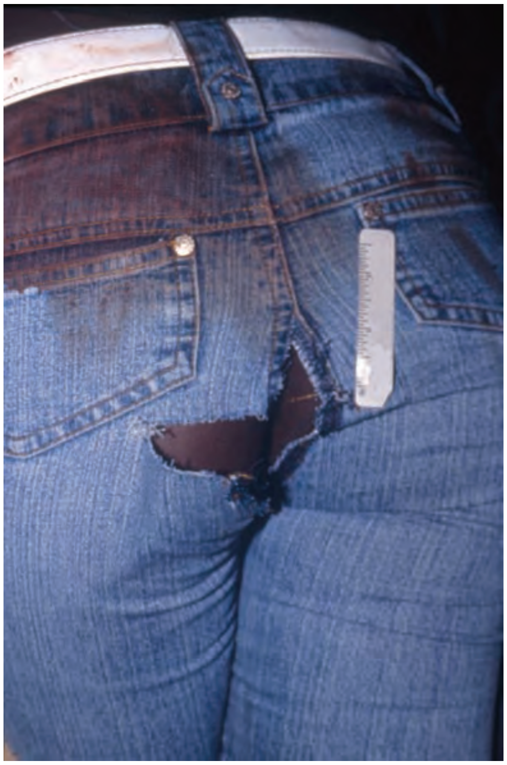

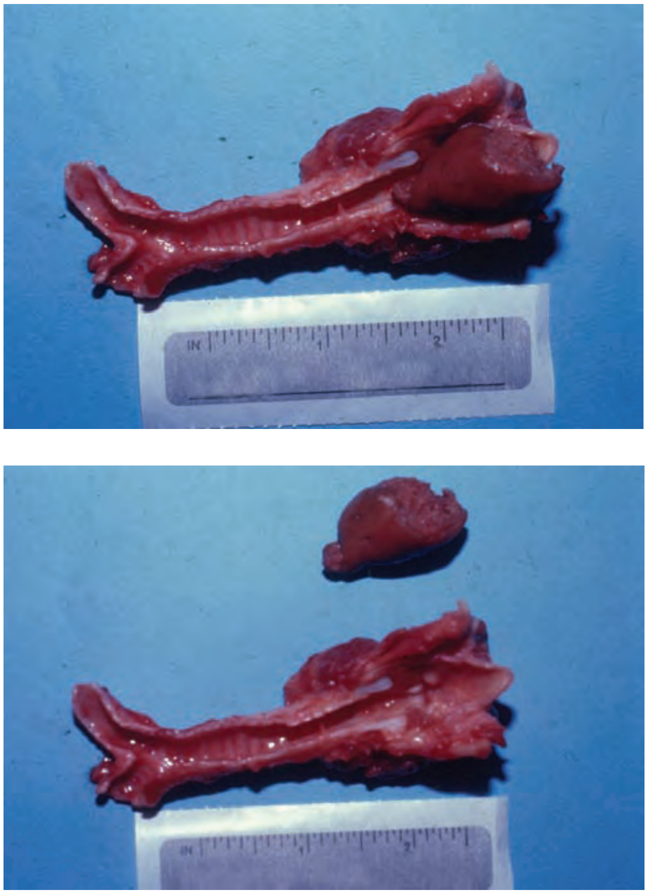

This is an airway from a 2-1/2 year-old who was fed a hot dog by her older sibling. The large size of the hot dog piece caused it to get wedged in her throat, leading to asphyxia.

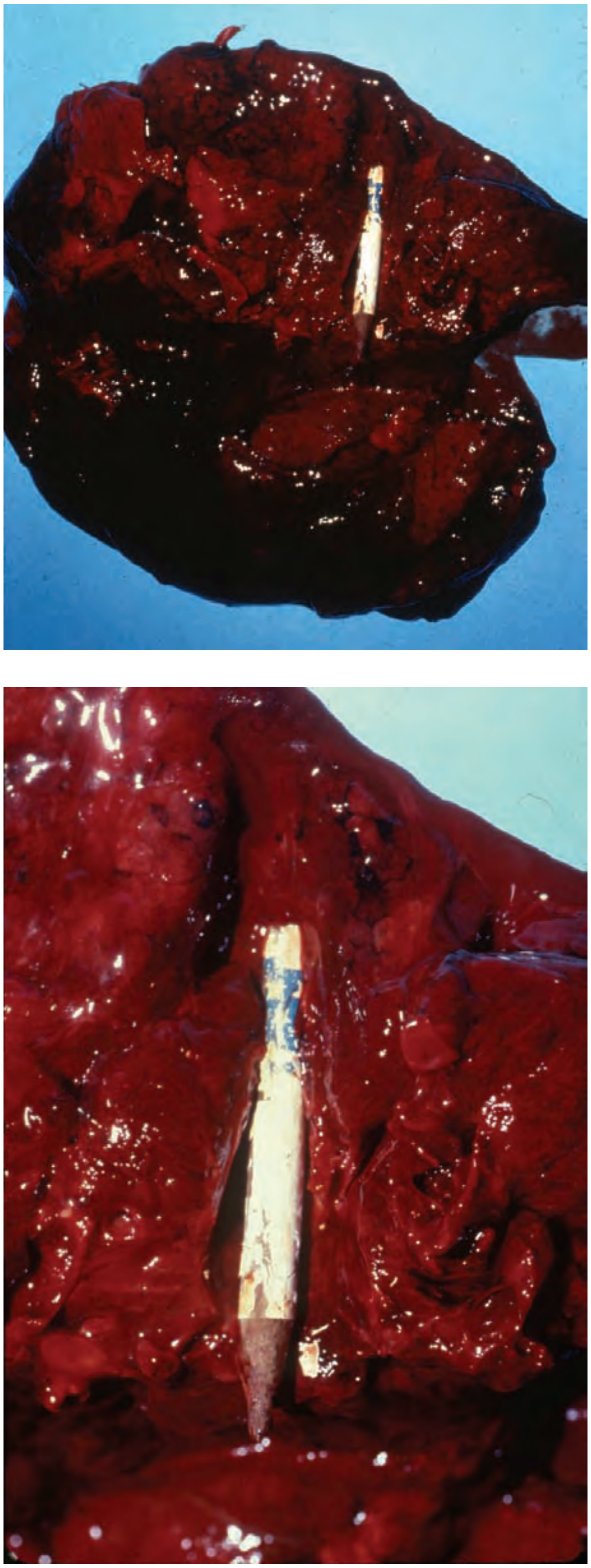

This individual had a pencil in his mouth when he collapsed at work. The pencil was inhaled and wedged in the right mainstem bronchus.

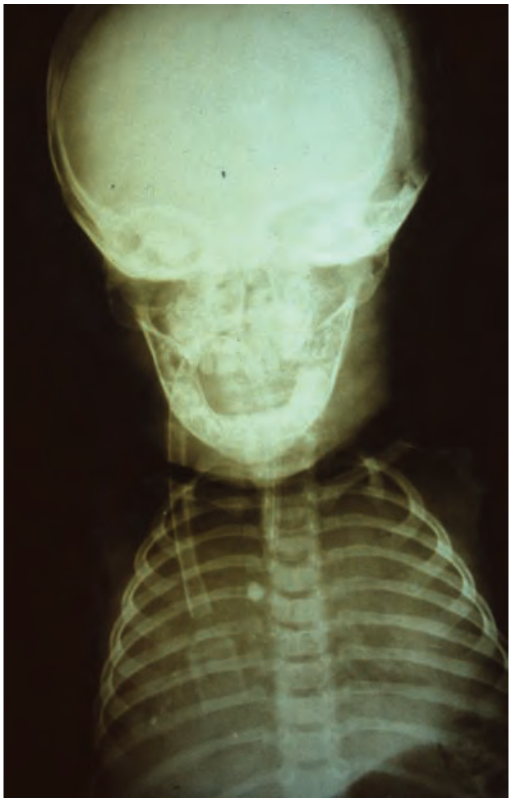

This child was found to have a pebble wedged within her right mainstem bronchus, as evident on this radiograph.

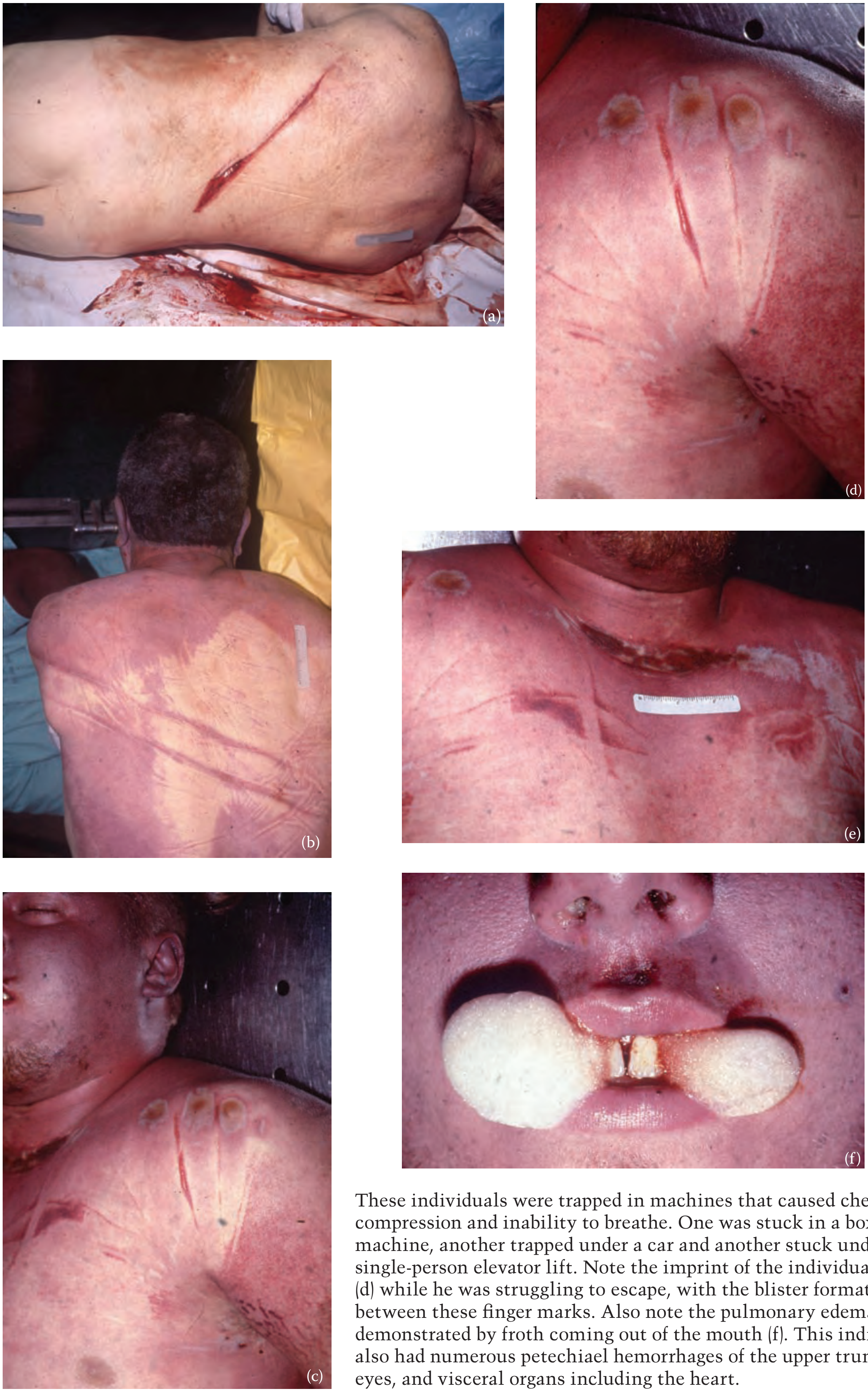

These individuals were trapped in machines that caused chest compression and inability to breathe. One was stuck in a box folding machine, another trapped under a car and another stuck under a single-person elevator lift. Note the imprint of the individual's hand (d) while he was struggling to escape, with the blister formation between these finger marks. Also note the pulmonary edema demonstrated by froth coming out of the mouth (f). This individual also had numerous petechiael hemorrhages of the upper trunk, face, eyes, and visceral organs including the heart.

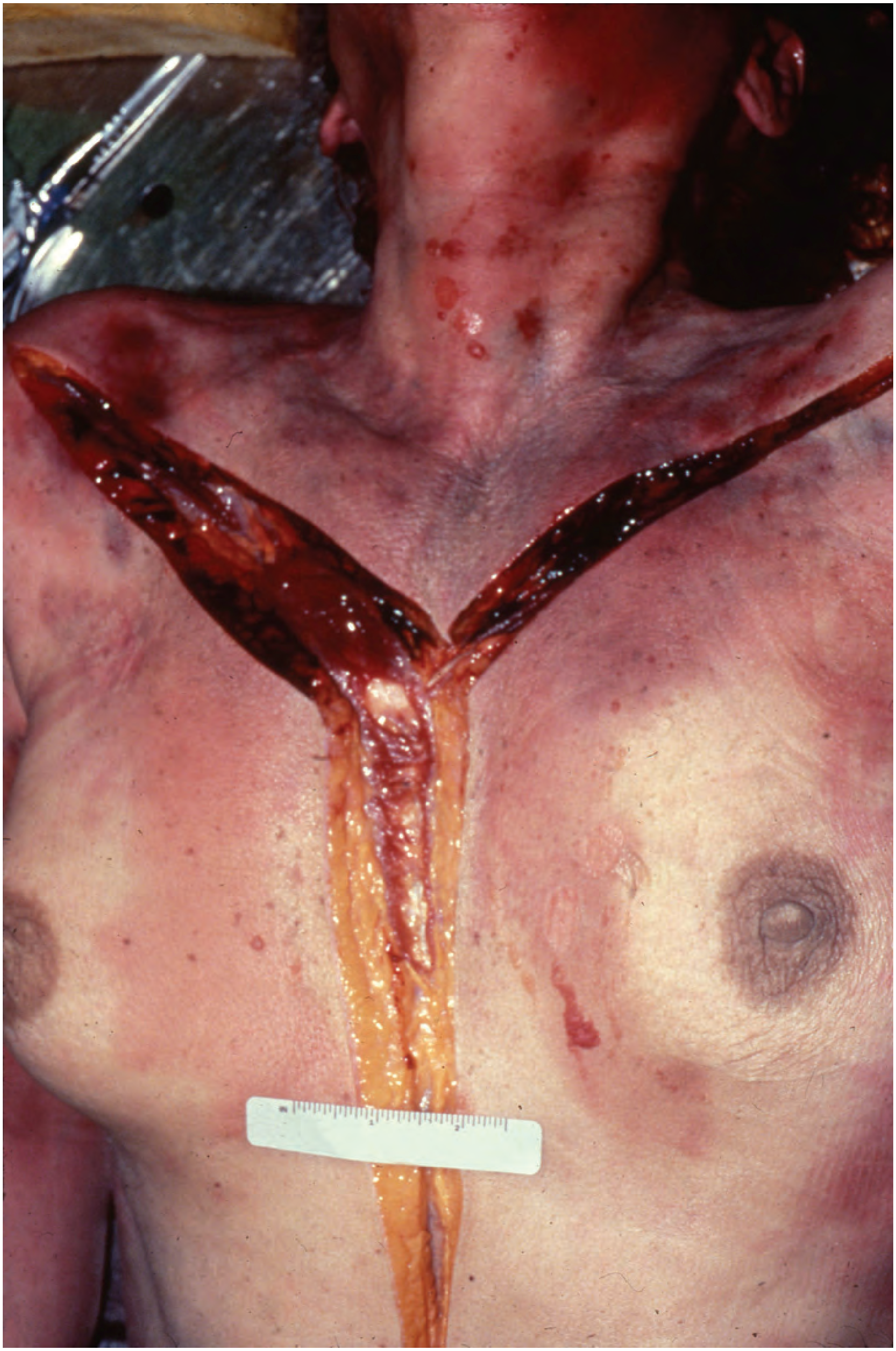

Homicidal asphyxia due to compression of chest and neck. Note the extensive hemorrhage at the superior chest visible at the superior aspect of the Y-shaped incision during autopsy. The lower aspect of this incision is yellow, anemic, and postmortem. This individual was punched, strangled and then the perpetrator sat on her chest during the assault.

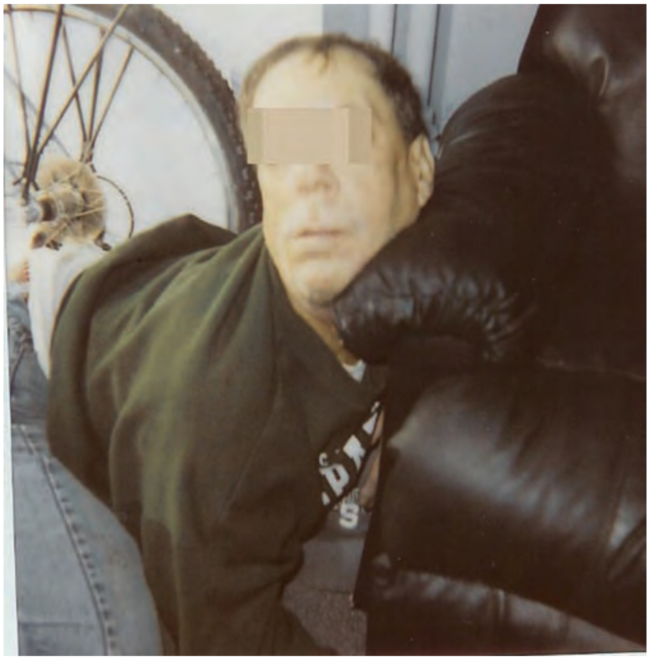

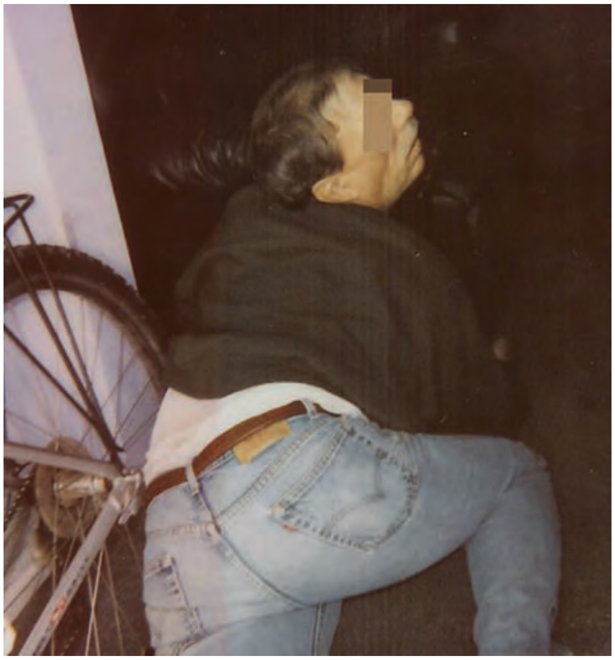

Positional asphyxia. These individuals were markedly intoxicated and passed out in positions that prevented them from breathing, thus obstructing blood circulation.

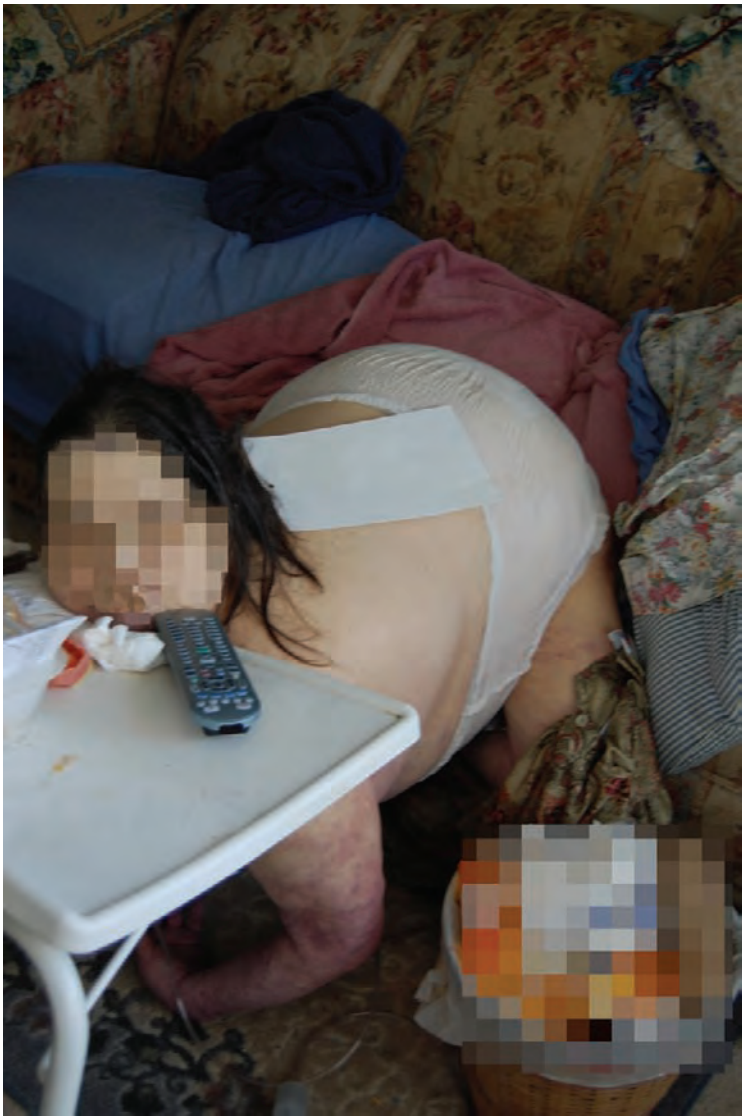

FORENSICS & CRIMINAL JUSTICE

Color Atlas of
Forensic Medicine and Pathology

New York City has the largest medical examiner's office in the United States, and the Brooklyn division is the busiest of the five boroughs. Charles A. Catanese received his Forensic Pathology fellowship training in New York, and then worked full time as a Medical Examiner in the Brooklyn office for more than 10 years. He has personally performed more than 4000 autopsies, including over 400 homicides. Dr. Catanese has worked through several disasters, including TWA Flight 800, AA Flight 587, and more than nine months on the World Trade Center fatalities. He is currently the Chief Medical Examiner of Orange County, New York. Drawing on his wealth of knowledge and experience in solving some of the most difficult cases a forensic examiner could encounter, he assembles hundreds of images from his own work experience to present the ***Color Atlas of Forensic Medicine and Pathology.***

Featuring twice the number of images as any other forensic pathology atlas, the book is filled with high-resolution photos that demonstrate postmortem changes of the human body and the different types of patterns produced in deaths caused by

- Natural causes
- Diagnostic or therapeutic procedures
- Substance abuse
- Poisoning
- Child abuse
- Firearms
- Blunt instruments
- Sharp instruments
- Burns
- Asphyxia

This easy-to-read atlas, created for medical and non-medical personnel, covers basic and advanced forensic concepts that relate to all manners of deaths. The carefully worded, unambiguous text describing each photo and the side-by-side comparisons of similar, yet different, pathologies make this remarkable atlas a powerful teaching tool for all those who must confront and solve the mystery of human demise. A fully searchable DVD version is also available.

4320X

ISBN: 978-1-4200-4320-4
90000
9 781420 043204

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

www.crcpress.com

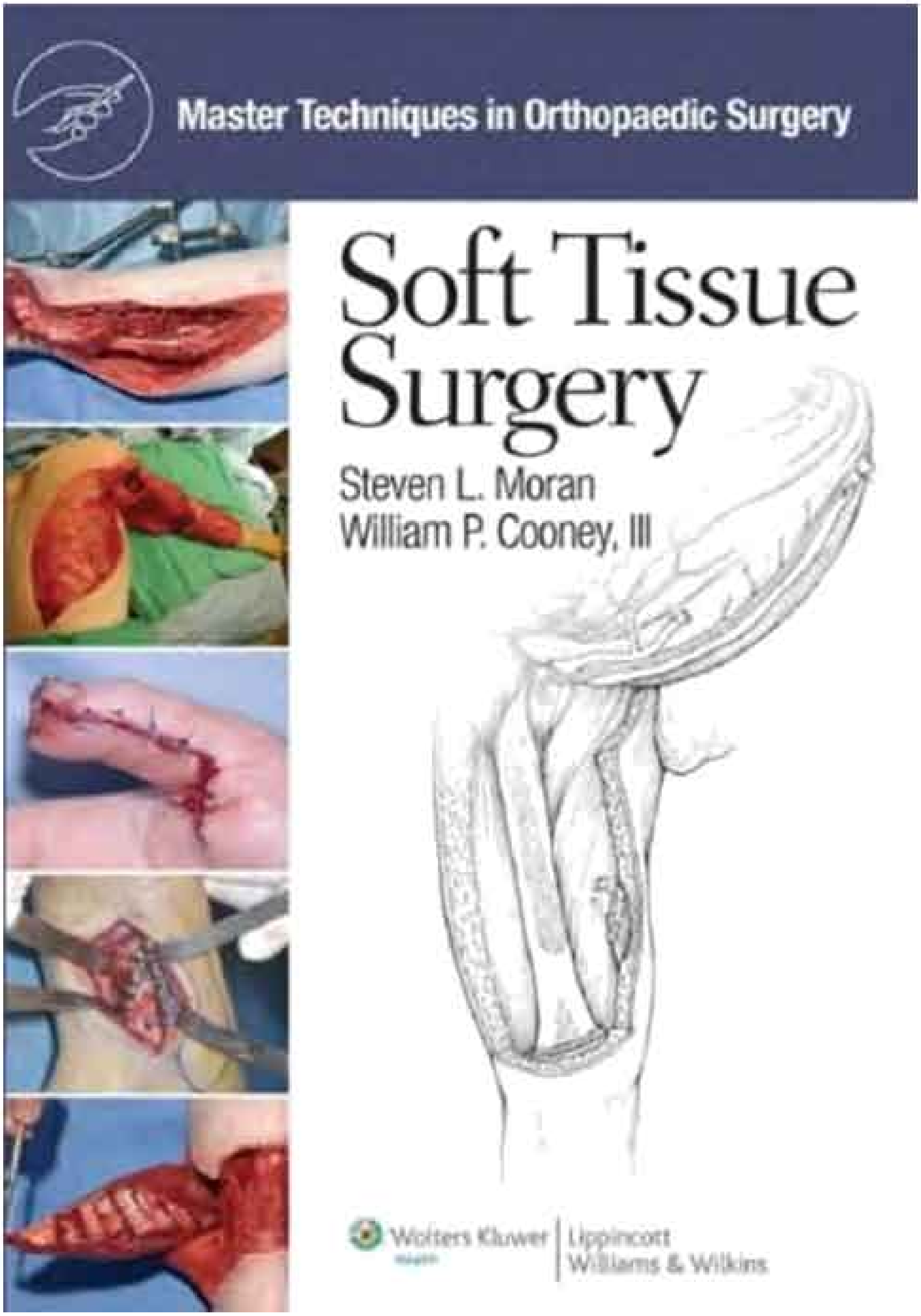

Master Techniques in Orthopaedic Surgery

Soft Tissue Surgery

Steven L. Moran
William P. Cooney, III

Wolters Kluwer Health | Lippincott Williams & Wilkins

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part I - Overview and Foundation > Chapter 1 - Why Wounds Fail to Heal

---

Chapter 1

Why Wounds Fail to Heal

Mark H. Jensen
Steven L. Moran

Surgical wound complications can be a significant problem resulting in prolonged hospital stay, reoperations, and significant morbidity for our patients. Often the surgeon can predict preoperatively which patients will have problems healing wounds or will need additional soft tissue coverage. Surgical planning should not only include the surgical approach but also an assessment of the patient's healing risks. It is our hope that preoperative risk reduction may avoid wound healing complications.

Prevention of wound healing complications starts with an understanding of the healing process. Risk factors can be recognized and modified. Techniques and practices that are proven to prevent wound infections must be employed. This chapter will focus on understanding wound healing and identifying patients at risk for wound complications; we will also attempt to offer means to minimize risk factors for complications through surgical technique and preoperative planning.

Phases of Wound Healing

Surgically induced wounds heal in several stages. The wound passes through phases of coagulation, inflammation, matrix synthesis and deposition, angiogenesis, fibroplasia, epithelialization, contraction, and remodeling. These processes have been grouped into three main stages: inflammation, fibroplasia, and maturation. Interruption in any one of these stages can lead to wound healing complications.

The inflammatory phase of wound healing involves cellular responses to clear the wound of debris and devitalized tissue. Increased capillary permeability and leukocyte infiltration occur secondary to inflammatory mediators and vasoactive substances. Polymorphonuclear cells (PMNs) are the first cell population in the wound followed by mononuclear leukocytes which mature into wound macrophages. Inflammatory cells clean the wound of harmful bacteria and devitalized tissue. Adequate tissue oxygen tension is necessary for the release of oxygen free radicals by neutrophils. Following the initial introduction of PMNs into the wound, lymphocytes enter the wound in great number, clearing the wound of old neutrophils and secreting important cytokines and chemoattractants for fibroblasts. Fibronectin and hyaluronate deposition from fibroblasts in the first 24 to 48 hours provides scaffolding for further fibroblast migration (1,2).

The fibroblast proliferation phase starts within the initial 2 to 3 days as large populations of fibroblasts migrate to the wound. Secretion of a variety of substances necessary for wound healing and includes large quantities of glycosaminoglycans and collagen. Ground substance formed from the four main glycosaminoglycans (hyaluronic acid, chondroitin-4-sulfate, dermatan sulfate, and heparin sulfate) acts as an amorphous gel that is necessary for collagen aggregation. Collagen levels rise for approximately 3 weeks corresponding to increasing tensile strength. After 3 weeks the rate of degradation equals the rate of deposition. Angiogenesis is an important aspect of the fibroblast proliferation phase as it helps to support new cells in the healing wound.

The maturation phase starts around 3 weeks and lasts up to 2 years. It is characterized by collagen remodeling and wound strengthening. Collagen is the principal building block of connective tissue and is found in at least 13 different types. Types I to IV are the most common in the human body. Each has a distinct feature and is found in different levels in many tissues. For example, type III collagen is high in hydroxyproline and low in hydroxylysine. It is commonly found in skin, arteries, bowel wall, and healing wounds. Type I collagen is found in skin, tendon, and bone; is low in hydroxylysine content; and is the most common collagen type, accounting for more than 90% of body collagen. Early wounds are comprised of a majority of type III collagen. As the wound matures, type III collagen is replaced by type I collagen.

Collagen cross-linking improves tensile strength. There is a rapid increase in strength of the wound by 6 weeks as the wound reaches 70% of the strength of normal tissue. The wound then gradually plateaus to 80% of normal strength, but never returns to preinjury levels.

Wound re-epithelialization occurs as adjacent cells migrate through a sequence of mobilization, migration, mitosis, and cellular differentiation of epithelial cells. Wound contraction starts at about 1 week. It is facilitated by the transformation of certain fibroblasts into myofibroblasts containing Î±Ã—smooth muscle actin. These cells adhere to the wound margins as well as each other and effect contraction of the wound. These stages are imperative for proper wound healing as interruption of these processes results in chronic wound complications (1,3).

Risk Factors

The identification of patients at risk for aberrant wound healing allows the surgeon to make appropriate plans for skin closure technique, flap utilization, and postoperative wound management. This will ideally result in modification of risk factors prior to surgery. In cases of chronic diseases or nonmodifiable risk factors, patients must be informed of increased wound healing risks. The following discussion focuses on commonly encountered risk factors with recommendations to ameliorate their effects.

Diabetes

Patients with diabetes are both more likely to undergo surgery and develop perioperative complications than nondiabetic patients. This leads to longer hospital stay with higher health care costs and increased perioperative mortality. Diabetic patients have increased rates of hypertension, cardiac disease, and renal failure. These factors lead to much higher wound healing complications (3). Diabetes inhibits wound healing through many mechanisms. It is a disease affecting small vessels which are critical in supplying nutrients to the healing wound. Elevated glucose levels also affect a myriad of inflammatory systems. Neutrophil adherence, chemotaxis, phagocytosis, and intracellular bactericidal activity are all impaired. Pseudohypoxia develops as a result of altered redox reactions and vascular permeability secondary to hyperglycemia. Furthermore, glucose is a proinflammatory mediator stimulating cytokine production and inhibiting endothelial nitric oxide levels (4). This translates clinically into higher infectious complications. *Tight control of glucose in the perioperative period mitigates the postoperative complications seen in the diabetic patient*. This has been found in both the intensive care setting as well as routine operative cases (3,5).

Management of patients with diabetes starts in the pre-operative period. There is good evidence to suggest that a preoperative hemoglobin A1C of less than 7% drastically decreases postoperative infectious complications (4). Physicians should aggressively improve glucose management using diet, oral hypoglycemic agents, and insulin as needed. Perioperative management should include sliding scale insulin or continuous insulin infusion to maintain glucose levels below 150 mg per deciliter (3). Evidence suggests that improved outcomes are possible with tighter control of glucose levels between 80 and 110 mg per deciliter in critically ill patients (5). After dismissal from the hospital, the patient should continue their preoperative diabetic regiment.

Obesity

Obese patients have a higher rate of wound infections, dehiscence, hematomas, seromas, and pressure ulcers (6Ã—8). This is felt to be due to multiple factors including difficulty with the operation, altered immune response, increased tension with closure, increased dead space, decreased microperfusion, and decreased mobility post operatively (9,10). Patients should be encouraged to lose weight prior to elective operations. Increased levels of physical activity preoperatively translates to an improved postoperative rehabilitation process. Successful weight loss is difficult to achieve in most patients. This has prompted some surgeons to recommend gastric bypass surgery prior to certain operations such as joint replacement. This has been associated with improved outcomes in hip replacement patients with morbid obesity (11).

Smoking

Numerous studies have consistently found that smokers have significantly higher rates of wound healing complications than nonsmokers. This is related to several causes including the vasoactive effect of cigarette smoke through the sympathetic alpha receptors, increased levels of carboxyhemoglobin with a reduction in oxygen-carrying capacity, increased platelet activation leading to microangiopathic thrombosis, increased levels of fibrinogen with decreased fibrinolytic activity, endothelial injury, and increased hemoglobin levels leading to increased blood viscosity. Regardless of the mechanism the effect is significant. Complication rates of 2.5% to 6% in nonsmokers versus 7.5% to 49% in smokers are reported in the literature.

The increased risk of wound complications in smokers is most pronounced in cases where a large area of tissue is undermined. This is most likely related to failure of the dermal and subdermal plexus supplying the resultant skin flap (Fig. 1-1). Poor outcome in smokers prompts many clinicians to postpone elective procedures until the patient has quit smoking for at least 3 to 4 weeks, particularly if the procedure involves large areas of undermining (12,13). Aggressive use of smoking cessation programs should be implemented. Still, other surgeons feel that this would restrict too many patients from receiving necessary surgery (14,15). We prefer to delay surgery for 4 weeks prior to elective surgical procedures where soft tissue coverage may be an issue and advocate early referral to a nicotine dependence unit. If smoking cessation is not possible, the patient is informed of the likelihood of additional flap coverage for wound closure and a higher likelihood of postoperative wound complications.

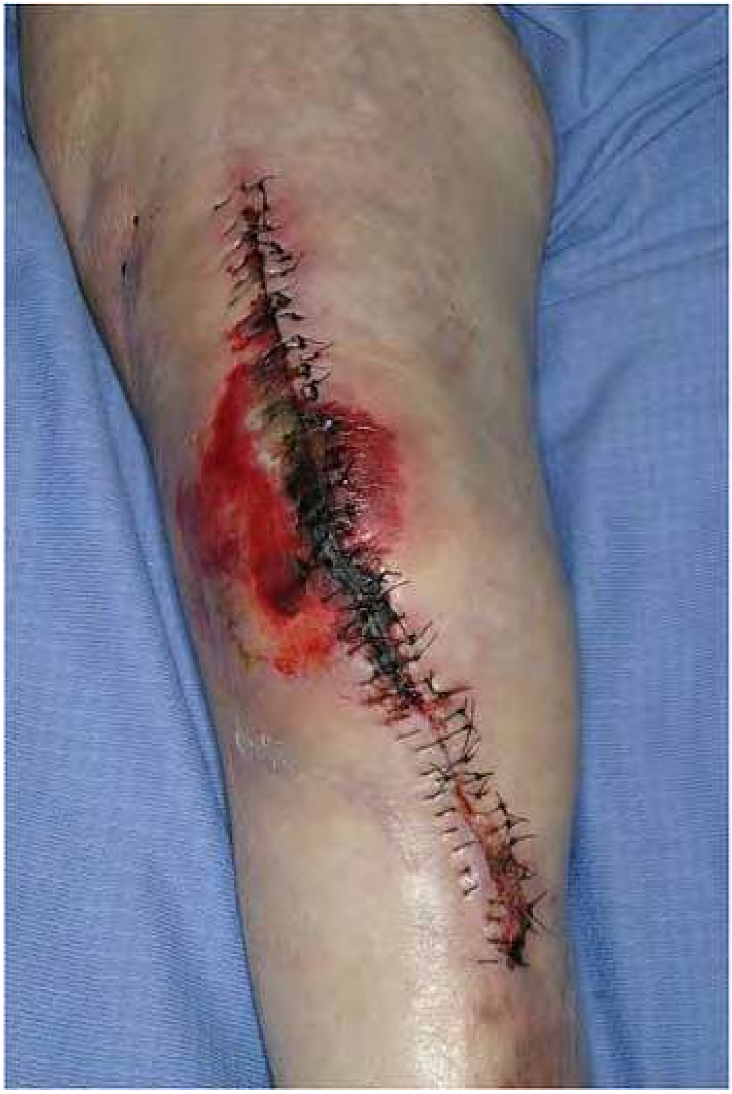

FIGURE 1-1 Following total knee arthroplasty this 76-year-old diabetic smoker developed wound complications necessitating wound debridement and flap coverage.

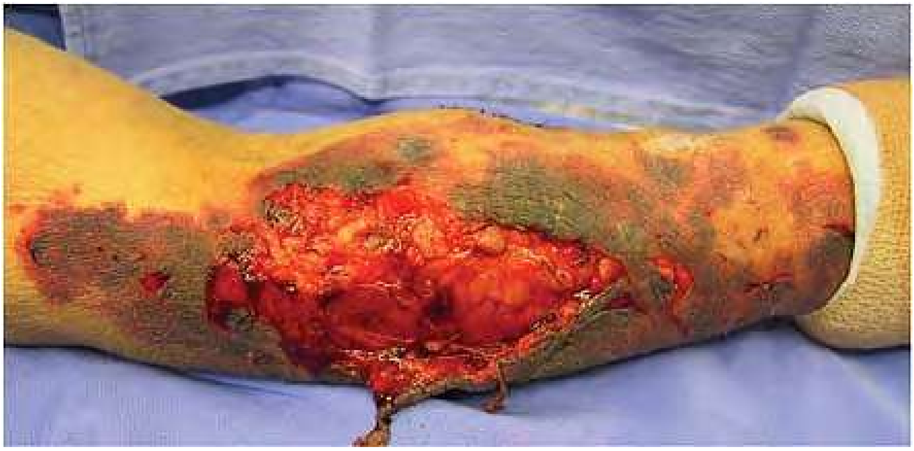

FIGURE 1-2 Chronic steroid use leads to increased skin and vessel fragility. This skin avulsion injury occurred in this 76-year-old steroid dependent man following a fall from a chair.

Immunosuppressive Medications

Patients with intrinsic or acquired immunodeficiencies are at increased risks for wound healing complications. Patients with MHC class-II deficiency have impaired wound healing because of altered T cell immune function (16). Patient populations at risk for wound healing problems due to altered immune function include patients with hereditary, infectious, and iatrogenic immune deficiencies.

Corticosteroids inhibit wound healing by their anti-inflammatory effect. Decreased numbers of inflammatory cells are noted at the wound site; delays in collagen synthesis, fibroblast proliferation, angiogenesis, wound contracture rates, and epithelial migration are also observed (Fig. 1-2). Vitamin A has been shown to counteract the effects of steroids on wound healing in all areas except wound contraction and infection (17). The exact mechanism is not known, but may be related to the TGF-beta, IGF-I, and hydroxyproline content in the tissue (18). Factors that can help improve wound healing in immunocompromised states include prevention of malnutrition, hypoxia, endocrine disorders, anemia, and other metabolic disorders. Dead tissue, foreign bodies, tissue ischemia, and hematoma should be minimized.

Radiation and Chemotherapy

Radiation and chemotherapy are known to cause delays in wound healing. Operations in irradiated fields are particularly problematic due to dense fibrosis and decreased perfusion caused by small vessel injury (Fig. 1-3). Irradiated fields are more susceptible to infection and delayed healing. Postoperative radiation initiated after the initial 3 to 4 weeks of primary wound healing does not seem to have as marked an effect on wound healing, but can lead to contracture, wound break down, and flap necrosis (21). Neoadjuvant chemotherapy can alter wound healing, especially in cases where chemotherapy has led to neutropenia. Surgery would be ideally delayed until full recovery of platelets and leukocytes. Healing seems to proceed normally in patients who receive their chemotherapy 3 to 4 weeks after surgery, as the wound has been allowed to proceed through the first stages of healing (8,22). Certain chemotherapeutic agents, however, can have a negative effect on wound healing far greater than 4 weeks; for example, Avastin (Bevacizumab), which inhibits vascular endothelial growth factor, has an extremely long half life ranging from day 11 to 50 and has been shown to inhibit wound healing when given in the neoadjuvant setting (23). Surgeons should have a good understanding of specific complications associated with individual chemotherapeutic regiments.

Malnutrition

It has been known for many years that malnutrition has deleterious effects on wound healing. Loss of nutrients alters host immunity through decreased T-cell function, phagocytosis, complement, and antibody levels. Certain patient groups are at particular risk of malnutrition. Severe catabolic states can be induced after multi-system trauma, sepsis, and burns. Significant increases in metabolic rate can also follow uncomplicated abdominal surgery (increase by 10%), uncomplicated injuries such as femoral fracture (20%), peritonitis (40%), and fever (10% for every 1Â°C above normothermia). This is particularly alarming considering the high portion of trauma patients who are unable to eat for extended periods leading up to and following surgery (24,25).

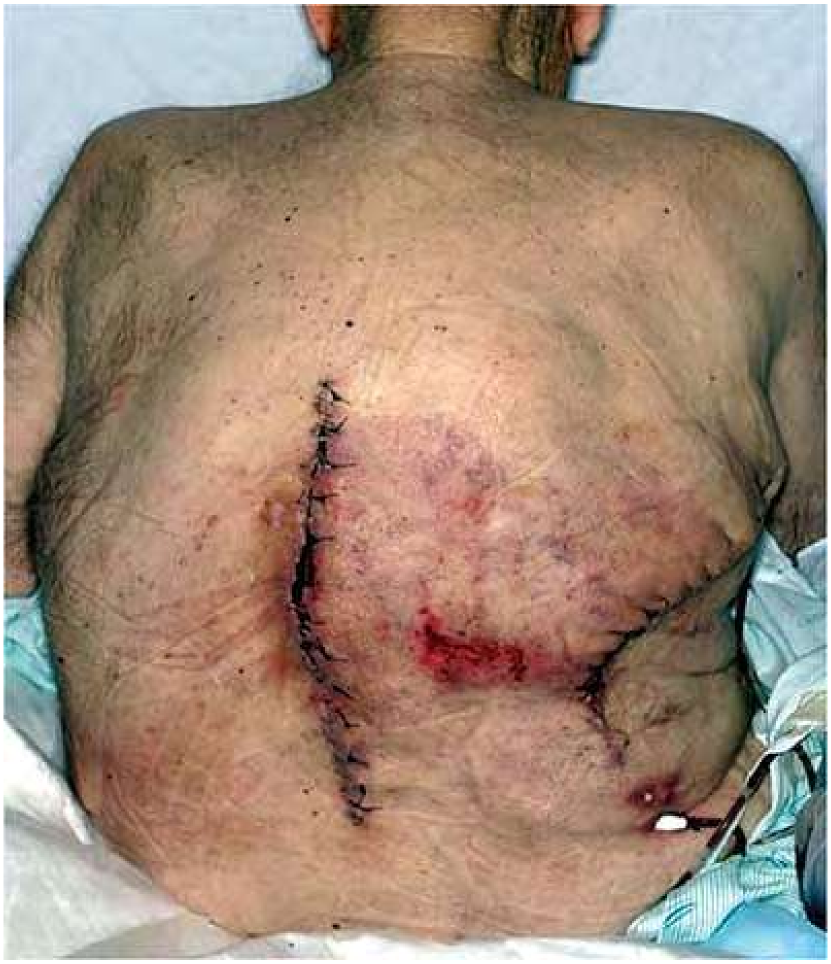

FIGURE 1-3 Signs of trouble in this 80-year-old gentleman who has undergone preoperative radiation with subsequent resection of a spinal tumor and coverage with a latissimus flap. Spinal tumor resection and flap harvest were preformed through separate and parallel incisions. The parallel incision in conjunction with the radiation damage has lead to ischemia within the center portion of skin bridge, which now shows signs of early ischemia with epidermolysis and ulceration.

Assessment of nutritional status becomes an important task for the surgeon. Patients at risk include those with moderate weight loss (10%Ã—20%), severe weight loss (>20%) and serum albumin <3.2 to 2.5 g/dL. Other groups include patients with trauma, burns, gastrointestinal (GI) dysfunction, cancer, fever, and those on chemotherapy. Consideration for preoperative total parenteral nutrition (TPN) should be given to patients who are severely malnourished, however this must be weighted against the risks associated with line infection and liver failure (24). Prealbumin is a laboratory value that can be followed as a surrogate of albumin due to its shorter half life.

Postoperatively patients should be started on a diet as soon as possible. Rapid diet advancement has been shown to decrease postoperative complications and decrease hospital stay (26). In patients who are unable to restart a diet, but have a functioning GI tract, tube feeds should be administered. TPN should be started early on patients who are expected to not tolerate GI feeds. The benefit risk ratio of TPN is equal for patients with normal metabolic demands after 1 week of fasting or inability to take oral intake. Patients with hypermetabolic states can benefit from earlier initiation of TPN (25,26).

Peripheral Vascular Disease

Peripheral artery disease is present in an estimated 0.9% of patients aged 40 to 49 compared with 15% to 29% of patients over the age of 70. Risk factors for the development of PVD include diabetes mellitus, hyperlipidemia, cigarette smoking, and hypertension. Patients at risk for PVD should be screened with blood pressure measurements of the arm and ankle to determine the patient's ankle brachial index; a normal value should produce a ratio >0.9. Symptoms of PVD include claudication, rest pain, atypical leg pain, and ischemic tissue loss. Aggressive lifestyle and risk factor modification is indicated to prevent progression of disease and death from cardiovascular or cerebrovascular events. These patients should undergo cardiac risk evaluation as peri-operative myocardial infarction (MI) and stroke are very common (27).

Wounds in the setting of peripheral vascular disease heal very slowly and can be difficult to manage. Patients with nonhealing wounds in spite of maximal medical management should undergo evaluation for revascularization. Amputation may be considered in the absence of target vessels for revascularization (27). Transcutaneous oxygen measurements in peripheral occlusive disease can be

helpful to determine healing potential. $TcPO_2$ levels greater than 30 mm Hg correlate with a high likelihood of healing whereas levels less than 15 mm Hg rarely heal (28,29).

Infection

Surgical site infections are a major impediment to wound healing and contribute to substantial morbidity and mortality. Postoperative infections increase the hospital length of stay by 7 to 10 days and increase charges by $2,000 to $5,000 (30). Rates of surgical infections vary by type of procedure, health of the patient, and skill of the surgeon. Wound classification corresponds with expected rates of infection as follows: clean (1.3%Ã—2.9%), clean-contaminated (2.4%Ã—7.7%), contaminated (6.4%Ã—15.2%), and dirty (5.1%Ã—40%) (6). Other risk factors for the development of postoperative infections include diabetes, obesity, smoking, steroids, malnutrition, colonization with *S. aureus*, preoperative hospitalization, and patient health.

The most important factors which can be used to prevent surgical site infections are timely use of preoperative antibiotics and operative technique. Other practices that are helpful include preoperative showering with antimicrobial soap, scrubbing and draping the patient with sterile drapes, hand washing, gloving, and the use of sterile gowns, masks, and hats by surgical personnel (31). These practices are aimed at reducing the amount of skin associated bacteria. Pathogens are not eliminated because approximately 20% of bacteria reside in the hair follicles and sweat glands where antiseptics do not reach. Although numerous studies have been aimed at identifying the best skin antiseptic, a systematic review failed to show superiority of one antiseptic over another (32).

Preoperative antibiotics administered within 2 hours of incision correlate strongly with the lowest rate of infections. Choices of antibiotics along with common pathogens are listed in Table 1-1. Endocarditis prophylaxis may need to be added in susceptible patients. In spite of clear benefits with the use of preoperative antibiotics, compliance is not perfect. Institutional policies should be constructed to promoting 100% compliance (33).

Hair removal is commonly performed prior to surgical procedures, however most studies have shown that this is not necessary and is associated with an increased risk of surgical site infections (34). If hair removal is preferred, clippers or depilatory creams are safer than shaving and should be used just prior to skin incision. Razors are associated with microtrauma to the skin and should not be used secondary to increased infection rates (35).

TABLE 1-1. Antimicrobial Prophylaxis for Surgery

Nature of operation	Common pathogens	Recommended Antimicrobials	Adult dosage before surgery[1]
Cardiac	*Staphylococcus aureus,*	cefazolin or	1-2 g IV[2]
	S. epidermis	cefuroxime	1.5 g IV[2]
		OR vancomycin[3]	1 g IV
Gastrointestinal			
Esophageal,	Enteric gram-negative	*High risk[4] only:*	
gastroduodenal	bacilli, gram positive cocci	cefazolin[7]	1-2 g IV
Biliary Tract	Enteric gram-negative bacilli,	*High risk[5] only:*	
	enterococci, clostridia	cefazolin[7]	1-2 g IV

Colorectal	Enteric gram-negative bacilli,	*Oral:* neomycin	
	anaerobes, enterococci	+ erythromycin base[6] OR metronidazole[6]	
		Parenteral:	
		cefoxitin[7]	1-2 g IV
		OR cefazolin	1-2 g IV
		+ metronidazole[7]	0.5 g IV
		OR ampicillin/sulbactam	3 g IV
Appendectomy,	Enteric gram-negative bacilli,	cefoxitin[7]	1-2 g IV
non-perforated[8]	anaerobes, enterococci	OR cefazolin	1-2 g IV
		+ metrodinazole[7]	0.5 g IV
		OR ampicillin/sulbactam[7]	3 g IV

Genitourinary	Enteric gram-negative bacilli, enterococci	*High risk only*[9]*:* ciprofloxacin	500 mg PO or 400 mg IV
Gynecologic and Obstretric			
Vaginal, abdominal	Enteric gram-negative bacilli,	cefoxitin[7] or cefazolin[7]	1-2 g IV
or laparoscopic	anaerobes, Gp B strep,	OR ampicillin/sulbactam[7]	3 g IV
hysterectomy	enterococci		
Cesarean section	same as for hysterectomy	cefazolin[7]	1-2 g IV
			after cord clamping
Abortion	same as for hysterectomy	*First trimester, high risk*[10]*:*	
		aqueous penicillin G	2 mill units IV
		OR doxycycline	300 mg PO[11]

		Second Trimester:	
		cefazolin[7]	1-2 g IV
Head and Neck Surgery			
Incisions through	Anaerobes, enteric gram-	clindamycin	600-900 mg IV
oral or pharyngeal	negative bacilli, *S. aureus*	+ *gentamicin*	1.5 mg/kg IV
mucosa		OR cefazolin	1-2 g IV
Neurosurgery	*S. aureus, S. epidermis*	cefazolin	1-2 g IV
		OR vancomycin[3]	1 g IV
Ophthalmic	*S. epidermis, S. aureus,*	gentamicin, tobramycin,	multiple drops
	streptococci, enteric gram-	ciprofloxacin, gatifloxacin	topically over
	negative bacilli,	levofloxacin, moxifloxacin,	2 to 24 hours

	Pseudomonas spp.	ofloxacin or neomycin-gramicidin-polymyxin B	
		cefazolin	100 mg subconjunctivally
Orthopedic	*S. aureus, S. epidermis*	cefazolin[12]	1-2 g IV
		or cefuroxime[12]	1.5 g IV
		OR vancomycin[3,12]	1 g IV
Thoracic (Non-Cardiac)	*S. aureus, S. epidermis,*	cefazolin or	1-2 g IV
	streptococci, enteric gram-	cefuroxime	1.5 g IV
	negative bacilli	OR vancomycin[3]	1 g IV
Vascular			
Arterial surgery	*S. aureus, S. epidermis,*	*cefazolin*	1-2 g IV
involving a	enteric gram-	OR vancomycin[3]	1 g IV

prosthesis,	negative bacilli		
the abdominal aorta,			
or a groin incision			
Lower extremity	*S. aureus, S. epidermis,*	cefazolin	1-2 g IV
amputation for ischemia	enteric gram-negative bacilli,	OR vancomycin[3]	1 g IV
	clostridia		

Reproduced from Antimicrobial prophylaxis for surgery. *Treat Guide Med Lett* 2006;4(52):83Ã—88, with permission.

1. Parenteral prophylactic antimicrobials can be given as a single IV dose begun 60 minutes or less before the operation. For prolonged operations (>4 hours), or those with major blood loss, additional intraoperative doses should be given at intervals 1Ã—2 times the half-life of the drug for the duration of the procedure in patients with normal renal function. If vancomycin or a fluoroquinolone is used, the infusion should be started 60Ã—120 minutes before the initial incision in order to minimize the possibility of an infusion reaction close to the time of induction of anesthesia and to have adequate tissue levels at the time of incision.

2. Some consultants recommend and additional dose when patients are removed from bypass during open-heart surgery.

3. Vancomycin is used in hospitals in which methicillin-resistant S. aureus and

S. epidermis are a frequent cause of postoperative wound infection, for patients previously colonized with MRSA, or for those who are allergic to penicillins or cephalosporins. Rapid IV administration may cause hypotension, which could be especially dangerous during induction of anesthesia. Even when the drug is given over 60 minutes, hypotension may occur; treatment with diphenhydramine (Benadryl, and others) and further slowing of the infusion rate may be helpful. Some experts would give 15 mg/kg of vancomycin to patients weighing more than 75 kg, up to a maximum of 1.5 g, with a slower infusion rate (90 minutes for 1.5 g). To provide coverage against gram-negative bacteria, most Medical Letter consultants would also include cefazolin or cefuroxime in the prophylaxis regimen for patients not allergic to cephalosporins; ciprofloxacin, levofloxacin, gentamicin, or aztreonam, each one in combination with vancomycin, can be used in patients who cannot tolerate a cephalosporin.

4. Morbid obesity, esophageal obstruction, decreased gastric acidity or gastrointestinal motility.

5. Age >70 years, acute cholecystisis, non-functioning gall bladder, obstructive jaundice or common duct stones.

6. After appropriate diet and catharsis, 1 g of neomycin plus 1 g of erythromycin at 1 PM, 2 PM and 11 PM or 2 g of neomycin plus 2 g of metronidazole at 7 PM and 11 PM the day before an 8 AM operation.

7. For patients allergic to penicillins and cephalosporins, clindamycin with either gentamicin, ciprofloxacin, levofloxacin or aztreonam is a reasonable alternative.

8. For a ruptured viscus, therapy is often continued for about five days. Ruptured viscus in postoperative setting (dehiscence) requires antibacterials to include coverage of nosocomial pathogens.

9. Urine culture positive or unavailable, preoperative catheter, transrectal prostatic biopsy, placement of prosthetic material.

10. Patients with previous pelvic inflammatory disease, previous gonorrhea or multiple sex partners.

11. Divided into 100 mg one hour before the abortion and 200 mg one half hour later.

12. If a tourniquet is to be used in the procedure, the entire dose of antibiotic

must be infused prior to its inflation.

Hypothermia has benefits of decreasing tissue oxygen consumption and is tissue protective in cardiac bypass and organ transplant. However, hypothermia may predispose to infection through cutaneous vasoconstriction. It has been shown that perioperative normothermia is associated with reduced rates of surgical site infections (36).

Several health care improvement initiative studies have shown that surgical site infection surveillance is effective in reducing the infection rates (37). The mechanism is unknown, but many researchers postulate that surveillance is associated with a greater awareness of sterile technique. Operative teams more rapidly recognize problems and find solutions. Surgeons should seek a program at their institution for health care improvement and operative safety. The Institute for Healthcare Improvement (IHI) is a nonprofit organization that leads health care improvement initiatives. They organized the 100,000 Lives Campaign to implement changes in United States hospitals which focused on increased compliance to four areas: appropriate use of antibiotics, removing razors from operative rooms, maintaining glucose control, and ensuring perioperative normothermia (38). Hospitals in compliance with these changes reported a 27% reduction in surgical site infections (7).

Surgical Technique

Numerous studies aimed at discovering risk factors for complications have identified surgeons as independent risk factors. This could be due to numerous factors including clinical judgment, baseline characteristics of patient populations, procedure type, and surgical technique. There is no doubt that meticulous surgical technique can improve outcomes. This is particularly important when dealing with a patient at risk for complications. Excess intraoperative blood loss has been associated with surgical site infections and poor wound healing (39). Blood loss leading to hypotension with subsequent vasoconstriction and tissue hypoxia is one explanation for increased wound healing complications. Techniques aimed at improving sterility, minimizing tissue destruction, maintaining meticulous hemostasis, improving speed and accuracy, minimizing foreign body placement, and attention to closure are expected to reduce complications.

The choice of suture and method of suture placement can also have an impact on wound healing. There is a wide array of suture types. In general, sutures can be divided by life expectancy (permanent, absorbable, fast absorbable) and consistency (braided vs. monofilament). Permanent suture is indicated in situations where suture absorption is not desired (e.g., vasculature anastomosis). Drawbacks include development of suture abscesses and chronic infection. Stitch abscesses are less problematic with absorbable sutures, but may lose strength prior to tissue healing. Braided sutures are strong, hold well, and are easy to tie. However, they are more prone to infection than monofilament. Braided sutures may also cause tissue tearing as in the case of venous repairs. Monofilament sutures are smooth and resist infection, but are less forgiving and require more knots to resist slippage.

Placement of sutures can also contribute to wound compromise. Simple interrupted sutures and vertical matrass sutures produce less wound edge ischemia than running and horizontal matrass sutures when tissue viability is in question. Horizontal matrass sutures placed under excessive tension can further impede blood supply to wound edges resulting in infection, necrosis, or dehiscence. Care must be taken not to incorporate or impinge vasculature pedicles when placing sutures.

Conclusion

Prevention of wound complications starts with the preoperative evaluation where risk factors for wound healing are identified and modified prior to elective procedures. In the traumatic setting and in the established soft tissue deficit, patient factors such as glucose control and nutritional status can be optimized to maximize the patients healing potential. With this chapter as an introduction, the following chapters will address evaluation of traumatic wounds followed by techniques for recruiting or mobilizing surrounding tissue to close large wounds and soft tissue deficits.

References

1. Falanga V. Wound healing and its impairment in the diabetic foot. *Lancet.* 2005;66(9498):1736Ã—1743.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 2

Initial Evaluation and Management of Complex Traumatic Wounds

Patricia L. McKay

George Nanos

The principles in this chapter are forged from our experience with the treatment of high energy contaminated wounds resulting from recent overseas military conflicts. The soft tissue injuries discussed within this chapter encompass a broad spectrum of clinical presentations, both acute and chronic. These treatment protocols reflect the ever-growing body of scientific knowledge and technological advances contributing to the successful treatment of our patients. In the following illustrations, we hope to point out the common mistakes and potential pitfalls of operative and perioperative wound management and provide surgeons with a framework for successful treatment of all types of wounds.

Initial Evaluation

A thorough initial patient evaluation gives the surgeon the information by which to formulate a treatment plan with the best chance for success. A complete and accurate history must include the circumstances leading to the current wound presentation, the mechanism of injury if associated with trauma, underlying medical conditions, current occupational and socioeconomic status, and patient social habits such as smoking which may have detrimental effects on reconstructive efforts (Table 2-1).

In cases of trauma the energy level, mechanism, location, and the time course from injury to presentation are invaluable in predicting prognosis and planning treatment. Medical comorbidities such as underlying cardiopulmonary or peripheral vascular disease, endocrinopathy, neuropathic disease, immunocompromising conditions, psychiatric illness, nutritional deficits, and allergies can negatively impact the success of treatment if not identified and optimized. Medical consultation and comanagement is recommended in these cases. Tetanus status must be addressed in accordance with Centers for Disease Control guidelines.

Failure to fully recognize the scope of the wound and/or injury can have dreadful consequences to the outcome. A detailed physical examination with critical assessment of vital signs, secondary survey, and multi-system examination is required in all cases to evaluate the wound and exclude other medical conditions or associated injury that may take greater priority in treatment.

In multi-system trauma, the Advanced Trauma Life Support (ATLS) approach, beginning with airway, breathing, and circulation (ABC's) is recommended. Traumatologists or critical care specialists should be the principle coordinators of all initial medical care to ensure appropriate global management of the patient.

Critical attention should be directed to vascular status. If capillary refill and/or pulses are abnormal or absent, or if the mechanism introduces suspicion for vascular injury, Doppler testing and/or perfusion studies must be obtained. Revascularization or repair of vessels should be completed prior to, or concurrent with, surgical wound treatment. Special attention should also be paid to the neurological examination, and all deficits clearly documented. In cases of extremity trauma, a high index of suspicion for compartment syndrome is essential. Continued reevaluation is required with adjunctive use of compartment pressure monitoring if the diagnosis is in question, emergent compartment release should be performed if clinically indicated, and prophylactic release should be considered in trauma cases with revascularization procedures.

Table 2-1. Wound Treatment Pitfalls

Inadequate clinical and surgical resources for proper treatment
Failure to recognize and optimize host factors
Failure to recognize and treat vascular compromise
Inadequate debridement
Failure to recognize and treat infection
Wound closure with excessive tension
Prominent bone or hardware

Ancillary studies aid in full comprehension of the clinical situation. Plain radiographs should be routinely obtained to evaluate for associated fracture, foreign body, exostosis, osteomyelitis, soft tissue emphysema, vascular calcifications such as those associated with Diabetes Mellitus, or other factors contributing to or resulting from the overlying wound. CT scans and MRI may provide additional valuable information. Markers of hematologic and immune status, clotting factors, electrolyte and renal function, and adequacy of resuscitation must all be accounted for based on the clinical situation.

Careful documentation of every aspect of the evaluation and treatment must be made, and generous use of medical photography is helpful not only for documentation, but also for preoperative planning purposes.

Preoperative Planning

Once the evaluation is complete, the surgeon can formulate a treatment plan and determine if surgical intervention is required. The most important question to consider is whether the surgeon and the facility possess the requisite experience, capability, equipment, and consultative and ancillary services to render optimum treatment (Fig. 2-1). The surgeon must have the technical ability and anatomic knowledge for the given location of the injury. The treating hospital and operating room must also be able to furnish the necessary instruments, implants, and wound treatment dressings.

When confronted with cases of multi-system trauma, a multidisciplinary team involving traumatologists, intensivists, medical consultants, infectious disease specialists, and wound care support personnel is generally considered to be mandatory. If unable to meet these basic requirements, the clinical situation should be carefully reviewed and the strongest consideration should be given for referral to a higher level of care.

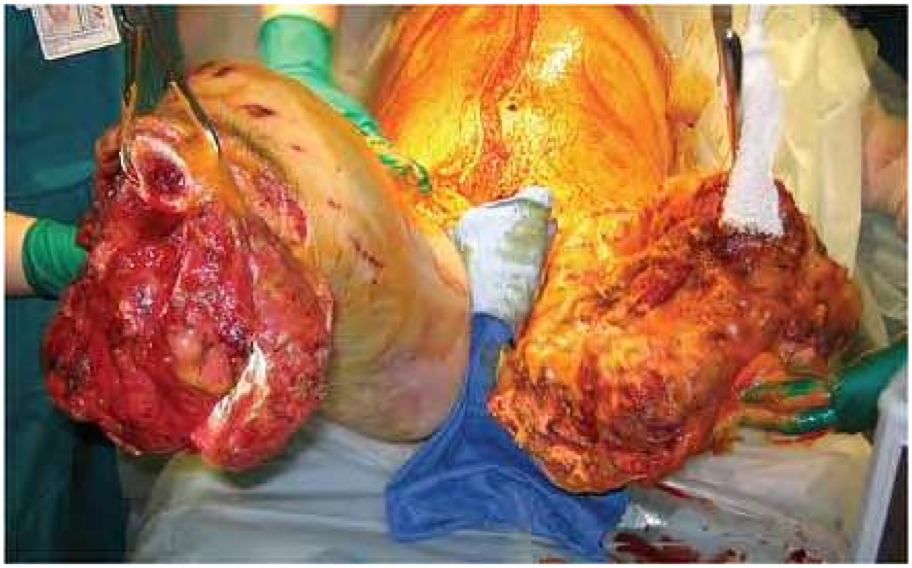

FIGURE 2-1 Injury to multiple extremities and organ systems changes surgical options, increases the risk for systemic problems, and has greater psychological impact. Options for flaps or skin grafts are far more limited. Proper room setup and a hospital team that is prepared to manage comorbidities will facilitate delivery of appropriate care and provide the greatest likelihood of success.

Surgery

Patient positioning varies depending on the site of the wound in question. General or regional anesthesia may be employed, and consideration for peripheral nerve blocks should be considered for perioperative pain control. A sterile or non-sterile tourniquet is indispensable to ensure optimal visualization for meticulous, thorough debridement and protection of vital structures. Appropriate broad-spectrum antibiotics should be given as scheduled throughout the surgical procedure.

Debridement

Successful surgical treatment of wounds begins with the meticulous and complete removal of foreign material, infection, and devitalized tissue to create a healthy wound bed (Fig. 2-2). In chronic wounds, we attempt to create an acute wound to promote healing. In acute injury, wounds must be extended past the zone of injury to ensure complete treatment, and failure to do so significantly limits the effectiveness of debridement. Judicious use of lavage may help remove foreign matter, but care must be taken not to extend the zone of contamination by forcing debris into the surrounding tissue. Use of a tourniquet early in the case is important to best visualize all contaminants and devitalized tissue and avoid injury to vital structures such as nerves and blood vessels. The tourniquet should be released prior to closure or dressing application to confirm removal of all devascularized tissue and excellent hemostasis.

A systematic approach to wound debridement is required and sharp debridement is the cornerstone of this surgical technique. We prefer the centripetal approach working from superficial tissues to deep, from the margins to the center of the wound. Starting at the skin edges, we meticulously work towards the deeper structures within the wound (Fig. 2-3). In general, we prefer excision of all devitalized tissue to a healthy tissue margin instead of a â€œwait and seeâ€• approach to suspect tissue as we feel this limits persistent contamination and infection. All non-viable or suspect tissue is sharply debrided from the wound until a healthy margin of viable tissue achieved. Every effort to preserve nerves and blood vessels crossing the zone of injury is made, and if they are transected these structures are carefully tagged with dyed monofilament suture and documented in the operative records so that they may be more easily visualized during later wound debridements or reconstructive efforts.

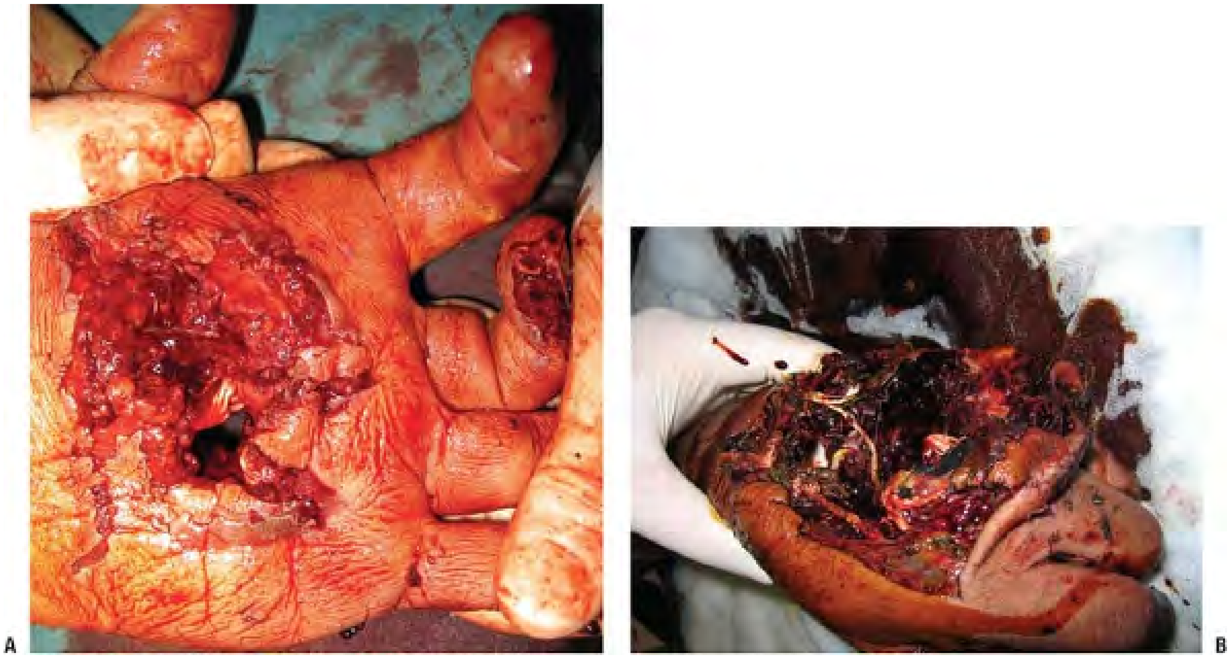

FIGURE 2-2 Palmar (A) and dorsal (B) views of an acute hand injury produced by an improvised explosive device. Jagged skin margins, necrotic tissue, foreign material, and hematoma are visible (photos courtesy of Dr. Dana Covey, CAPT, MC, USN).

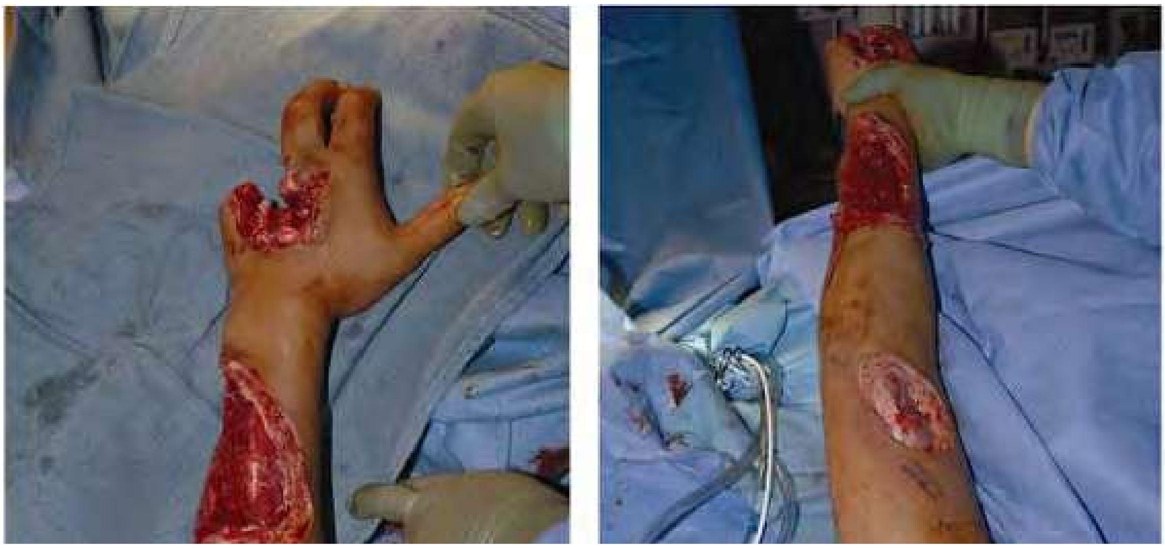

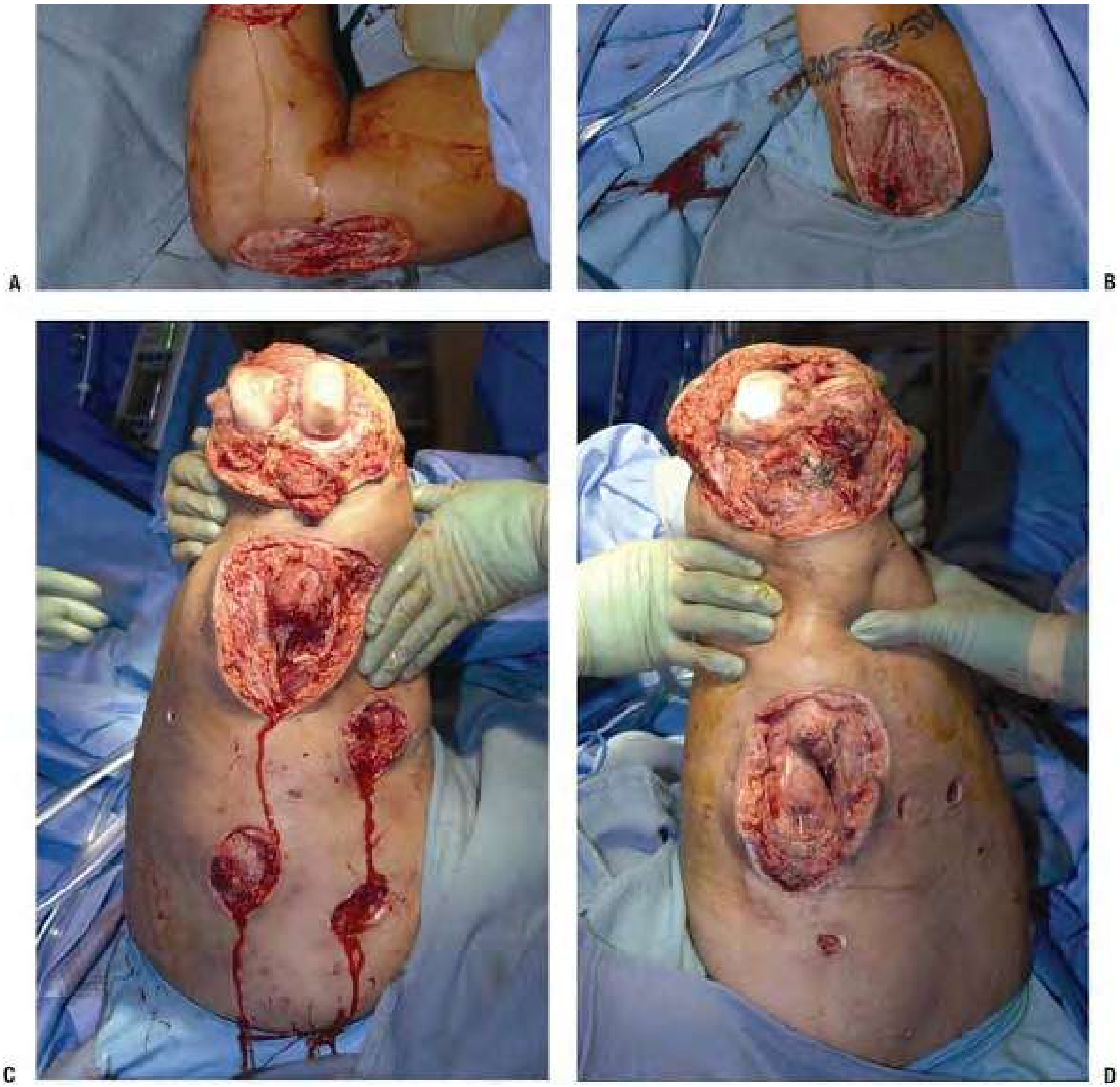

FIGURE 2-3 This patient, injured by an improvised explosive device, sustained left upper (A,B) and bilateral lower extremity injuries (C,D). Excellent serial wound debridement results in smooth margins, healthy skin and muscle ready for skeletal stabilization, and delayed primary wound closure for the lower extremities.

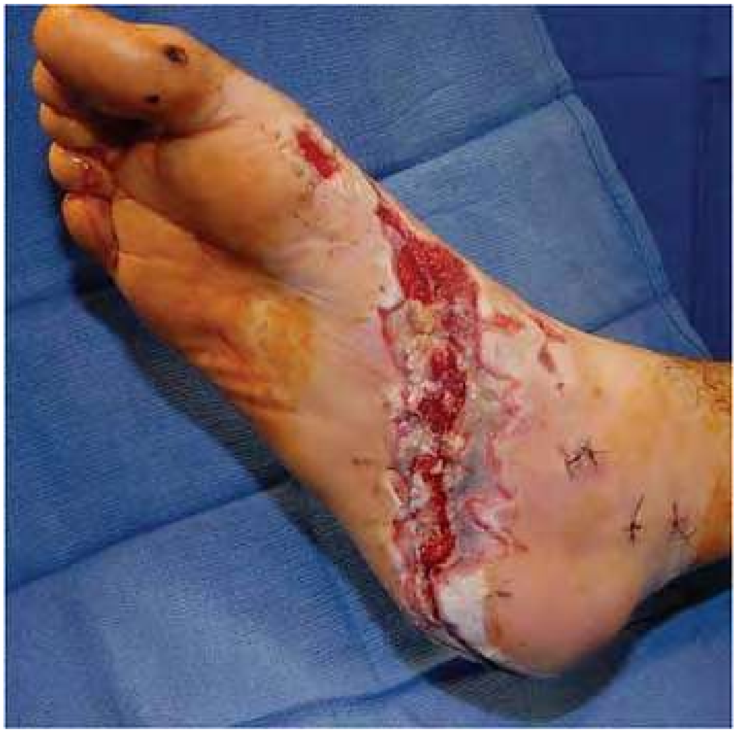

FIGURE 2-4 Plantar wound is not ready for closure or coverage; jagged skin and eschar at the wound edge needs to be sharply debrided. Serial debridements allow the zone of injury to fully declare itself.

Identification of non-viable tissue remains a challenge, and there is no substitute for experience. In general, non-bleeding skin that appears dusky or does not blanch should be excised, creating a smooth wound margin and avoiding the creation of a ragged skin edge that is difficult for subsequent skin grafting or closure. Subcutaneous fat should be soft and yellow; hard, dusky, or gray fat should be excised (Fig. 2-4). It is also important to keep this tissue moist during deeper debridement, as it will easily desiccate. Injured blood vessels or nerves must be carefully assessed for primary or delayed repair or grafting. Smaller sensory nerve branches may not be amenable to salvage, and if so, we like to pull traction on the proximal end, cut sharply, and allow retraction into the soft tissues. If the stump cannot be retracted, we make every effort to bury it in muscle. To the greatest extent possible local soft tissues should be used to cover exposed tendons, nerves, and vessels to prevent desiccation and further injury.

In debridement of muscle, fascia, and tendon, there are several important points to keep in mind. Muscle fascia should be stout, white, and shiny, while non-viable fascia will often appear grey or black, fragile, and stringy. When excising fascia, however, great care should be taken as neurovascular bundles may be in close proximity. Muscle should be red, shiny, of good consistency, contractile, and bleeding (Fig. 2-5). Anything to the contrary should be considered for debridement. Knowledge of anatomy and local blood supply is paramount in this endeavor as overly aggressive debridement within muscle compartments may devascularize previously viable tissue. Tendon debridement must be carefully considered due to potential loss of function. Tendons are also easily desiccated, especially if overlying paratenon or sheath is missing.

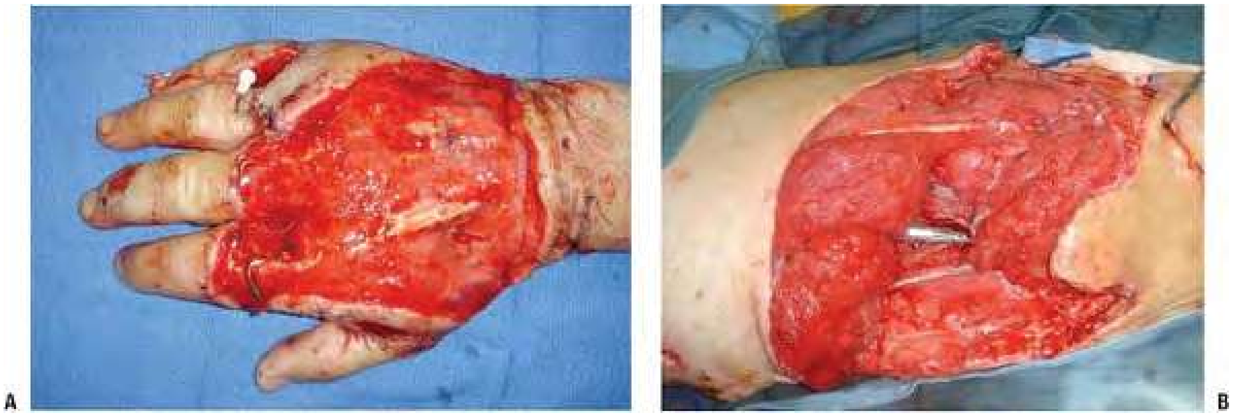

FIGURE 2-5 A: Right hand and (B) left hip open wounds ready for further coverage. Note the healthy, smooth skin margins, clean granulating tissue bed, minimal edema, and skeletal stabilization.

Devascularized bone fragments must be removed from the wound bed, with the exception of substantial articular fragments, which should be retained in an attempt to preserve the articular surface (Fig. 2-6). Curettes, rongeurs, and burs are useful to check for punctated bleeding indicative of healthy bone that should be preserved.

Cultures of any contaminated or osteolytic bone will help guide antibiotic selection.

There are many tools available for mechanical debridement. One tool we find useful for very large wounds is the VersaJetâ„¢ Hydrosurgery System (Smith and Nephew). This device uses negative pressure from a high-speed stream of water across a small aperture at the tip to remove softer tissue and surface debris by suction. Additional lavage is not required when using this device. However, great care must be used around neurovascular structures to avoid injury.

Strict hemostasis is critical to prevent hematoma and limit further infection and morbidity due to blood loss. Suture ligatures and surgical clips should be used for larger vessels, and bovie or bipolar cautery for smaller vessels. We avoid use of braided suture when possible to avoid harboring bacteria.

Judicious use of a tourniquet is helpful to identify and control large bleeding vessels and includes release to assess hemostasis prior to closure, grafting, or dressing application. Adjunctive topical hemostatic agents are available and have been used successfully in some of our most severely war injured patients. Lavage is important for removal of foreign debris and lowering bacterial counts. Reflecting prior published research, we use plain saline for lavage without antibiotics. Based on our experience and the literature, we limit use of pulsatile lavage in the forearm and lower leg, and almost never use it in the hand or foot. Pulsatile lavage can further damage delicate tissues, exacerbating the potential for adhesions and functional loss. When pulsatile lavage is used in the hand and foot, it is usually at the half pressure setting. Usually, bulb irrigation is sufficient when combined with careful debridement.

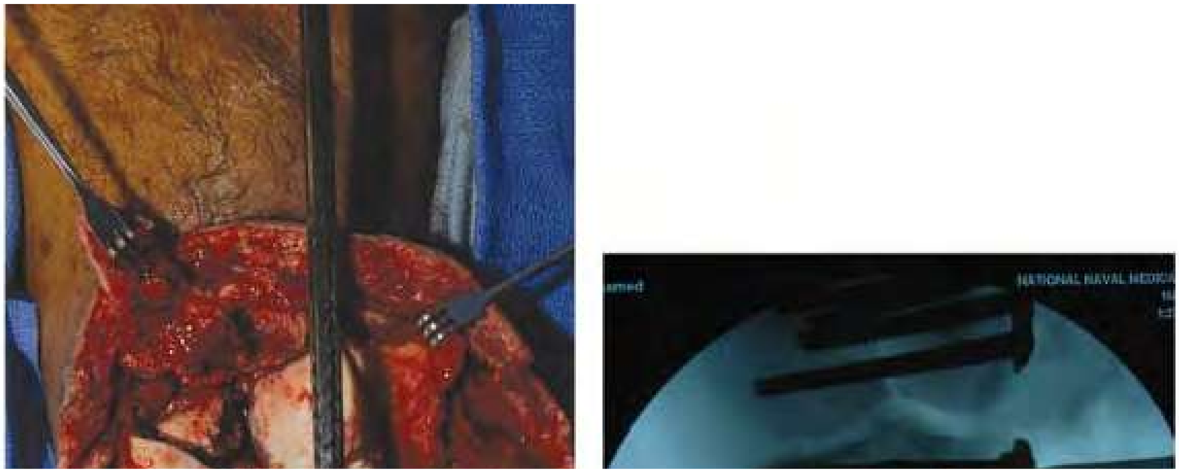

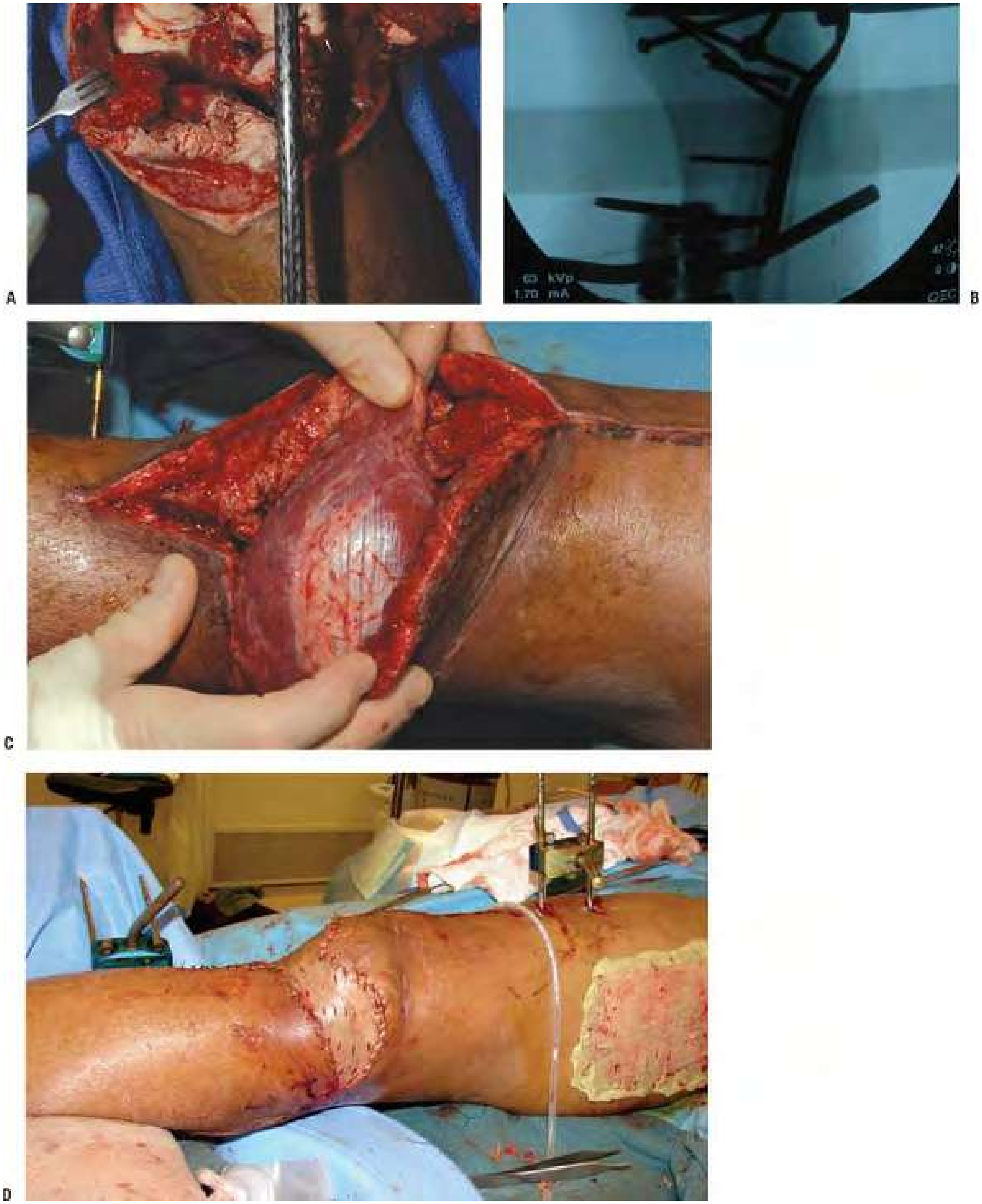

FIGURE 2-6 During serial debridements every effort is made to preserve substantial articular cartilage fragments. This case also illustrates the value of a

team approach, after extensive surgical time to obtain the best possible ORIF of this comminuted intraarticular knee injury (A,B), the plastic surgery service performed a medial gastrocnemius pedicled rotation flap and split thickness skin grafting (C,D).

Temporary Coverage and Void Fillers

Negative pressure dressings are a great advance in the treatment of wounds not amenable to primary closure (Fig. 2-7). The wound VACÂ® (Vacuum Assisted Closure) dressing is commonly employed at our institution. It continues to debride wounds while reducing edema and local bacterial counts and it promotes growth of healthy granulation tissue. It also eliminates the need for multiple daily dressing changes, thereby reducing the patient's discomfort and nursing staff workload. We use portable suction devices supplied by the manufacturer set at the recommended 125 mm Hg intermittent suction. Wall suction can be used if a portable unit is not available, but care must be made to ensure the unit is calibrated to provide a true measurement of suction. We prefer to use less suction in the upper extremity due to the more delicate nature of the soft tissues and have found this to be effective. In the hand, use of the wound VACÂ® is limited to avoid desiccation and injury to vital nerves, blood vessels, and tendons. It can be of benefit when the underlying tissue is robust such as the thenar or hypothenar eminence, or when the more delicate tissues are absent due to the injury and debridements (Fig. 2-8). When neurovascular structures are exposed we employ standard wet to dry dressings. In general, we try to limit exposure of blood vessels, nerves, or tendons to the wound VACÂ® and try to rotate available local tissue to provide coverage prior to placement of the wound VACÂ®. Multiple wounds can be treated with the same suction tube using a foam bridging technique. However, care must be taken to cover the skin with the supplied biofilm or ioban to prevent local skin breakdown.

Adequately eliminating contamination and infection is essential to successful wound treatment. In addition to appropriate broad-spectrum antibiotic use, there are many different options available to provide local infection control that can be tailored to the clinical or surgical situation. Antibiotic bead pouches or fracture spacers have been used effectively to provide local infection control in cases of wounds with associated high-energy fracture patterns. With comminution and bone loss soft tissue space can be maintained for future reconstruction and enhanced mechanical stability provided (Fig. 2-9). In highly resistant bacterial infection, silver impregnated films, colloidal materials, and wound VACÂ® sponges are additional options for the surgeon and have been utilized with great frequency at our institution. We presently reserve the use of silver for refractory cases due to concerns of resistance and the current high cost of this modality. For extremely large wounds with highly resistant bacterial colonization or infection that are not amenable to wound VACÂ® treatment, Sulfamylon (mafenide acetate) or Dakin's soaked wet-to-dry dressings have proven effective and resulted in successful wound closure. Infectious disease specialty assistance is recommended in such cases.

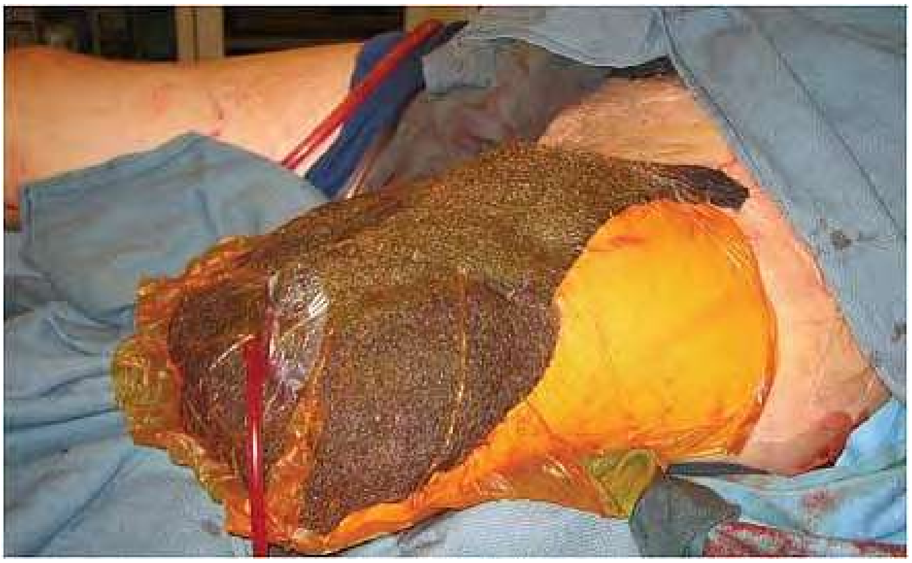

FIGURE 2-7 Use of serial debridements and negative pressure dressings prepare the wound for subsequent flap coverage.

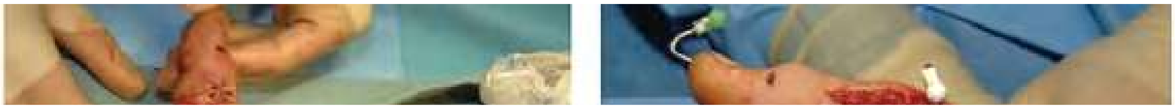

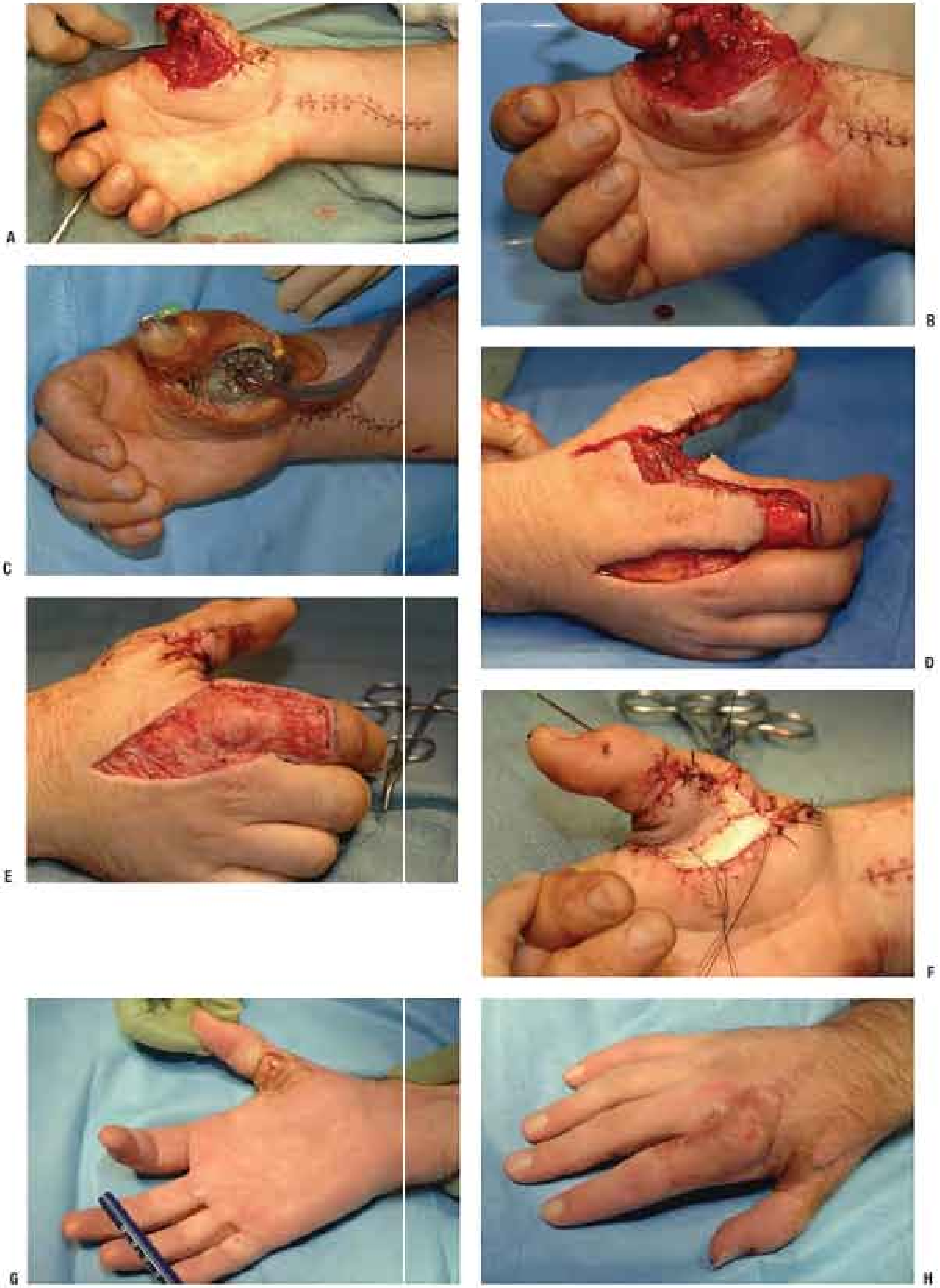
A
B
C
D
E
F
G
H

FIGURE 2-8 This case, of a patient who sustained an isolated injury to the thumb from an AK-47 rifle round, illustrates several principles of wound management: thorough debridement preserving vital structures (A) and provisional skeletal fixation with K-wires (B); C: the use of negative pressure dressings to prepare the wound for coverage. D-H; attainment of a closed healthy soft tissue envelope prior to reconstruction, in this case a local rotation flap and full-thickness skin graft were selected.

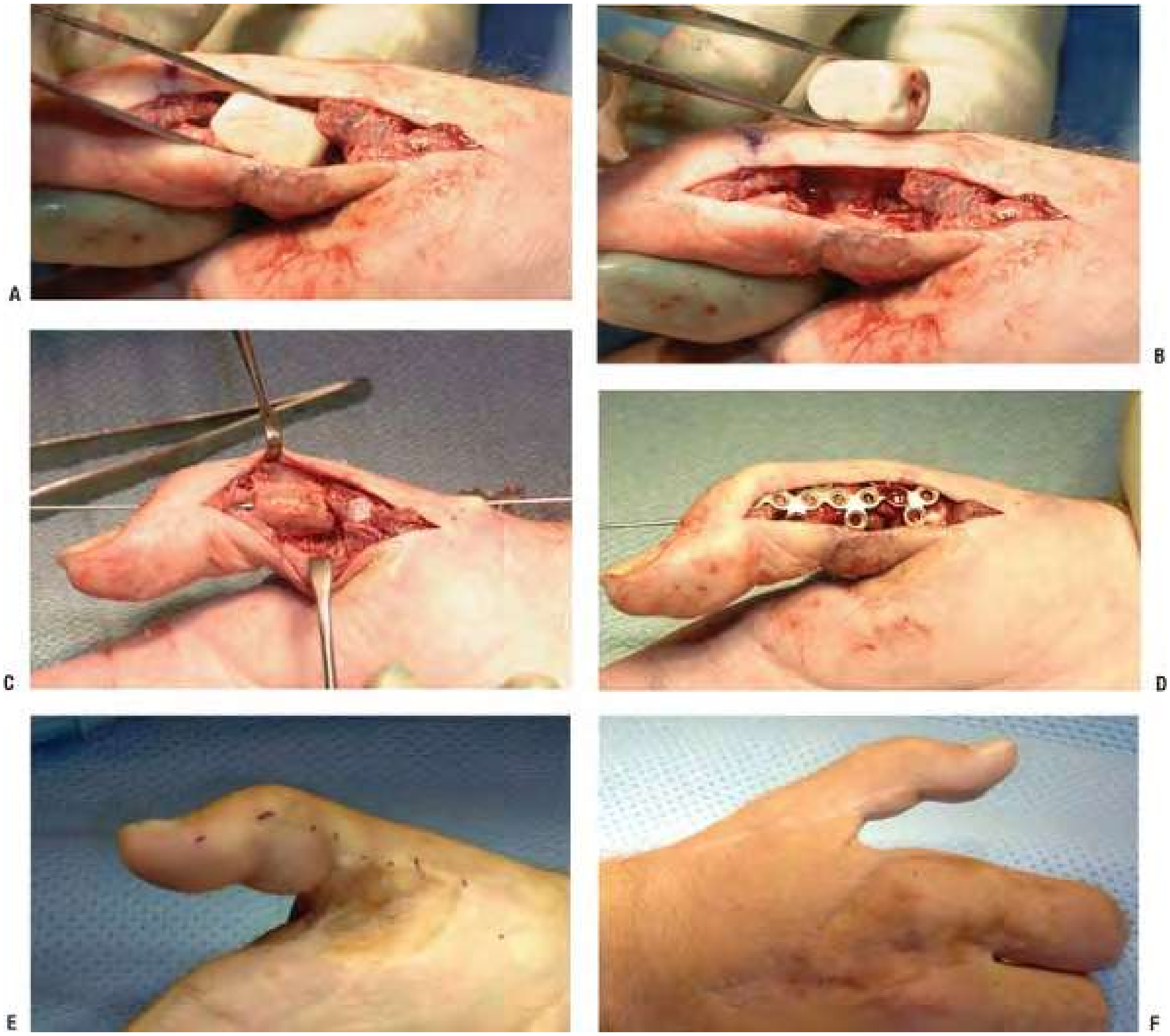

FIGURE 2-9 A,B: In cases of contaminated segmental bone loss (case from Fig. 2-8), an antibiotic impregnated cement spacer can be used to decrease infection risk and maintain space and alignment until the wound is clean enough for soft tissue reconstruction with bone grafting (C,D). E,F: A well healed soft tissue envelope, demonstrated here, is necessary prior to performing nerve and tendon grafting.

Choice of Fixation

When wounds are associated with fractures in the acute setting, provisional stabilization should be attempted to maintain soft tissue space, prevent mechanical agitation of the surrounding tissues, and optimize pain control (Fig. 2-10). In general, external fixators are preferred acutely with conversion to definitive fixation as indicated by the injury. In the setting of blast injuries, large amounts of debris are forced into the wounds with tremendous energy and the level of contamination is typically higher than that seen in most blunt open trauma. Our experience has shown the significant potential for widespread osteomyelitis when intramedullary fixation is selected for blast injured patients and we now prefer definitive treatment with an external fixation device in many cases. For the distal extremity, as well as periarticular fractures, Kirschner wires are indispensable for temporary and sometimes definitive fixation. Definitive fixation that requires significant soft tissue stripping may only compound the injury and should be entertained with extreme caution as the clinical situation dictates.

For highly contaminated wounds, or when there is concern for viability in critical areas or structures, repeat operative debridement should be planned every 24 to 36 hours until a healthy, vascularized soft tissue bed is achieved.

Timing of Final Closure

When is the wound ready to close? There are several elements that must be present to ensure success of wound closure. Certainly, all non-viable or necrotic tissue must be absent from the wound. It cannot be overemphasized that good vascular flow must be present for healing of tissues. In addition, a tension free closure must be achieved or blood flow will be compromised at the wound margins leading to wound breakdown and dehiscence. Acute limb shortening can be considered when significant bone and soft tissue defects are present (Fig. 2-11). Soft tissue expanders can also be a valuable tool in the reconstructive phase (Fig. 2-12) One significant marker of excessive tension is blanching at the wound margins. Suture type, configuration, and binding tightness of the suture must all be considered. Mattress sutures can be tension relieving and provide nice eversion of wound edges, but one must also recognize that inappropriate orientation or overtightening of these sutures, especially in horizontal mattress fashion, can create ischemia at the wound edges. In wounds prone to tension in the closure, we typically choose nylon suture sized appropriately to the wound in vertical mattress fashion to distribute tension evenly. At times, we will also utilize â€œtrauma retentionâ€• type sutures using heavier gauge nylon to improve local tension. When excessive tension remains, it is better to leave the wound open and return to the operating room later when tension free closure can be obtained. In these cases, we prefer a â€œJacob's ladderâ€• type closure with staples and vessel loops, usually over or under a wound VACÂ® sponge or gauze dressings to create gentle traction at the wound edges and promote future closure attempts. While commercial skin traction devices are available (Fig. 2-13), we have not found them to be more effective, and they are certainly more costly.

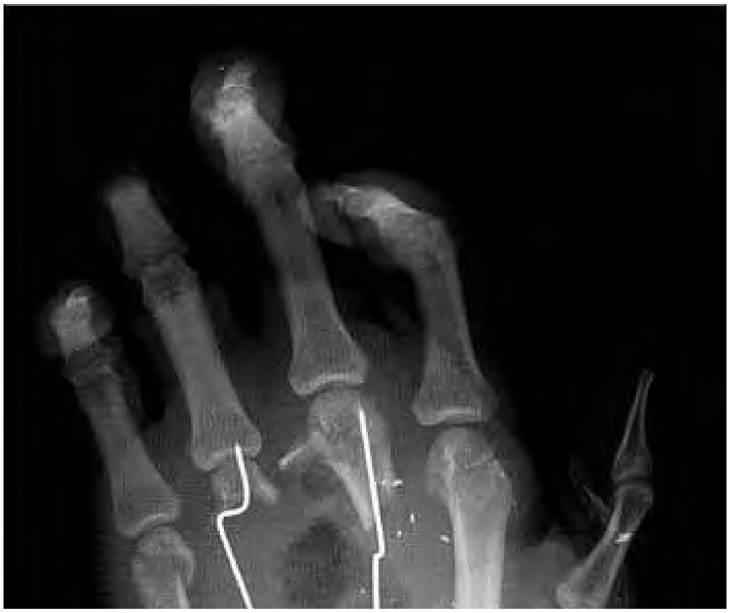

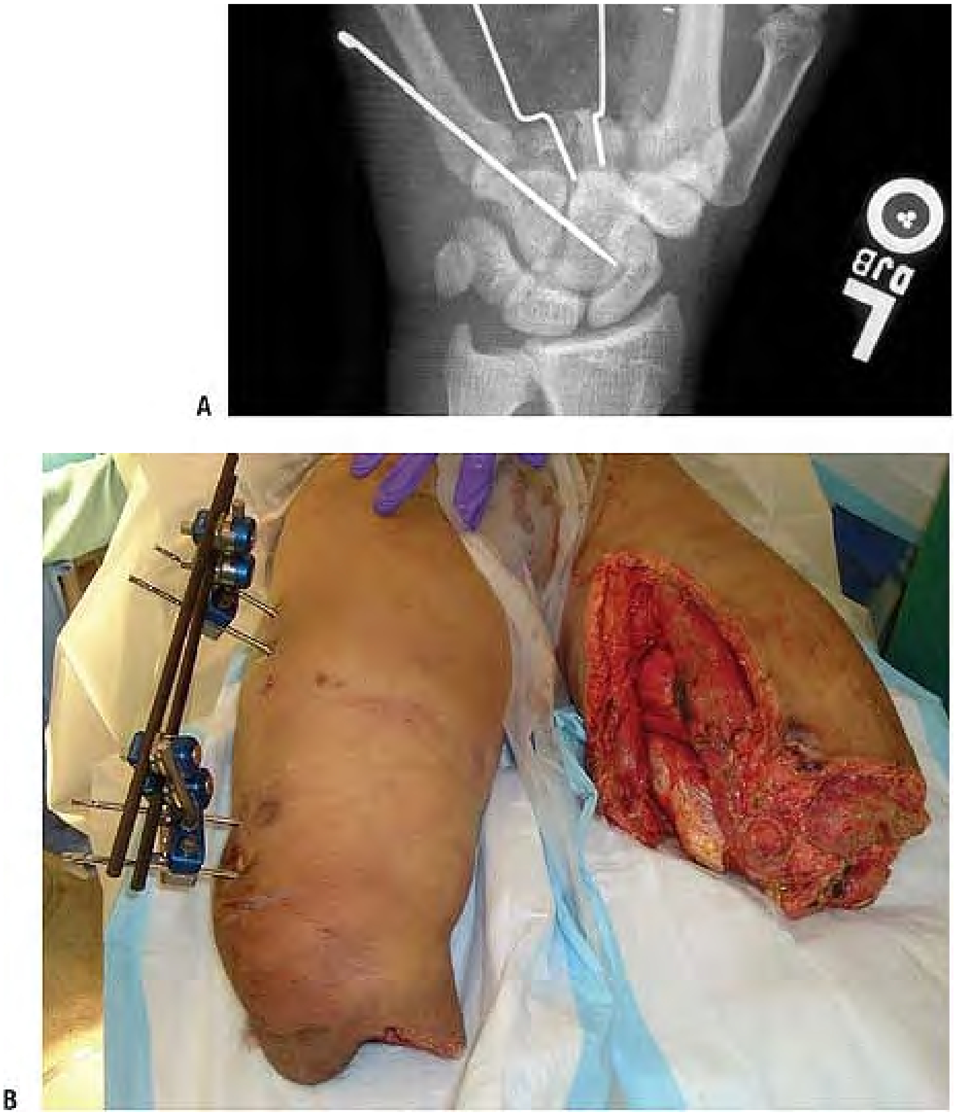

FIGURE 2-10 A,B: Fracture stabilization is required to assist in soft tissue management and to maintain structural relationships for subsequent reconstruction. K-wires and external fixation are especially useful in traumatic, highly contaminated wounds.

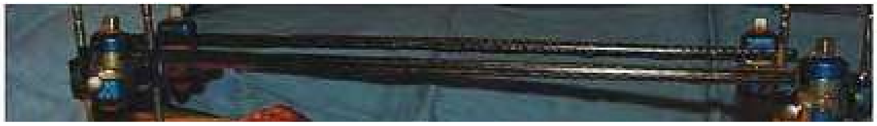

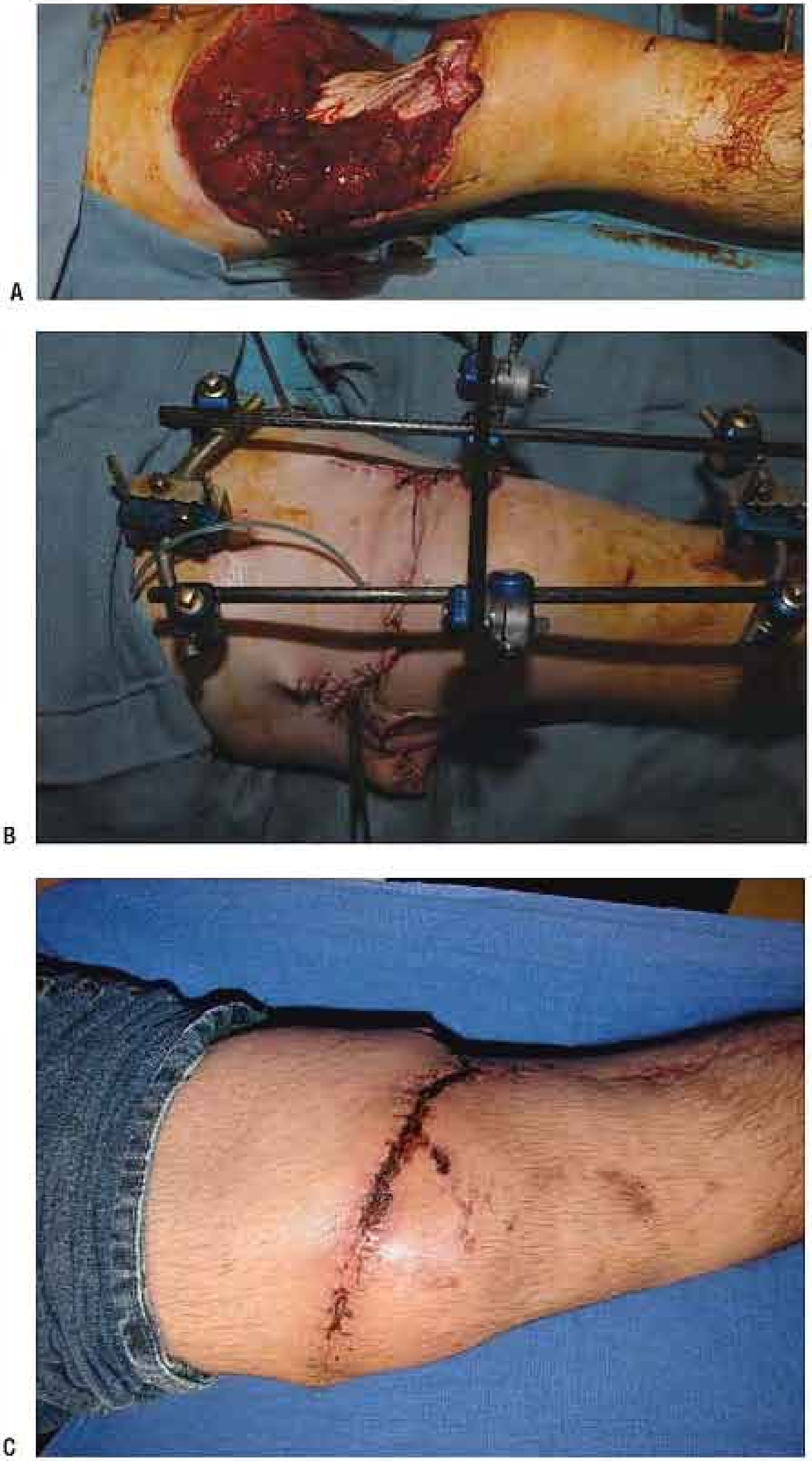
A
B
C

FIGURE 2-11 A-C: Large wound with a segmental femoral defect and associated loss of the anterior thigh soft tissue was managed with acute femoral shortening. Following soft tissue stabilization the femur was subsequently lengthened.

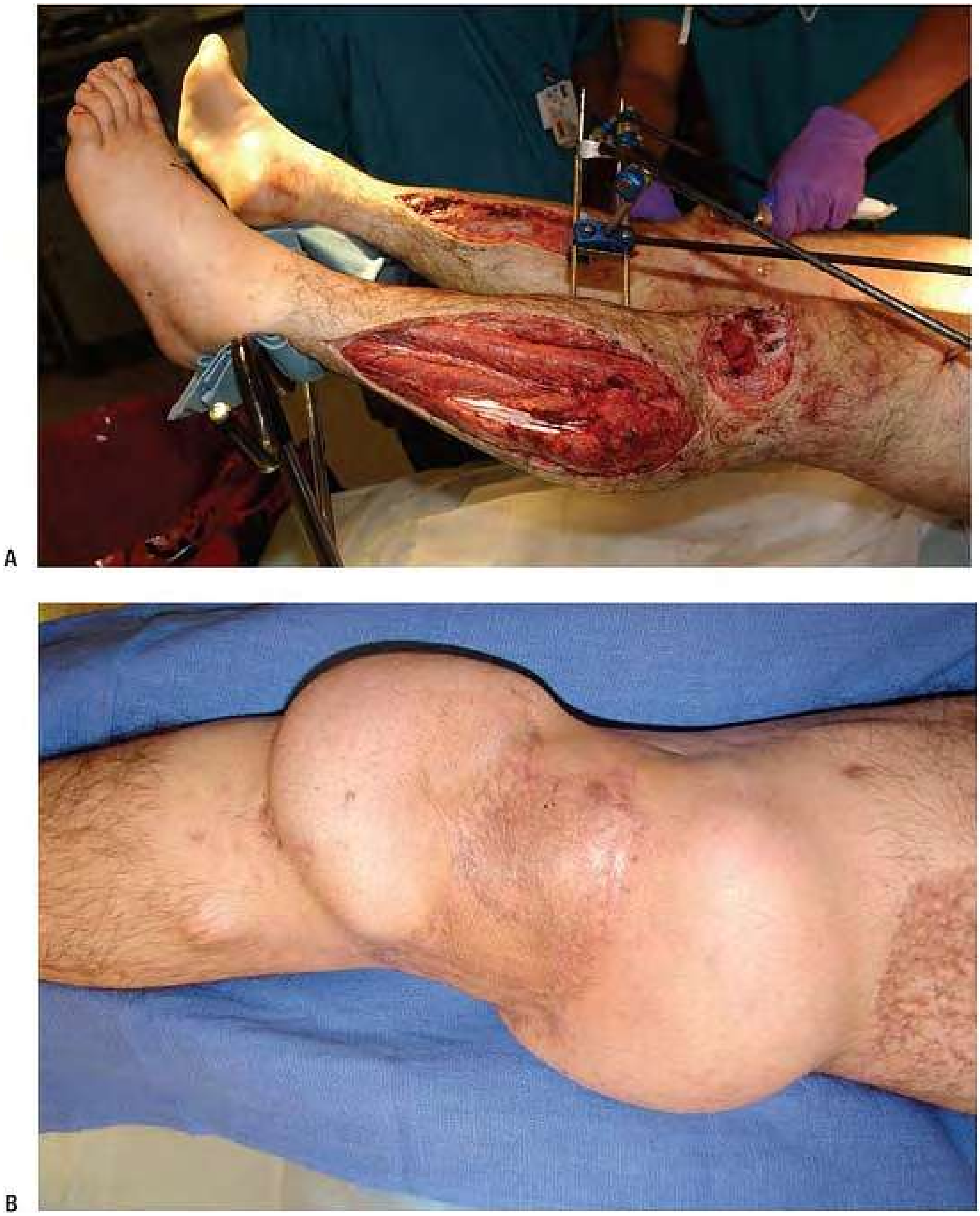

FIGURE 2-12 A,B: This patient required extensor mechanism reconstruction after an open knee injury. Soft tissue expanders were used after the initial fracture care, flap, and skin grafting, in order to create adequate skin for reconstruction and knee motion. It is important to remember that soft tissue scarring as well as

split thickness skin grafting limit skin excursion and pliability. Subsequent surgery may require release of scar or replacement of damaged skin to obtain motion in affected joints.

Absence of infection is also critical to wound closure. We routinely culture infected or suspicious wounds to help guide antibiotic therapy. Previous literature suggests a culture threshold of less than 10^5 bacteria per gram of tissue is required to allow successful wound closure. However, the utility of cultures is highly dependent on laboratory technician experience, and they have not proven to be clinically helpful at our institution. Clinical and laboratory indicators of infection such as fever, elevated white blood cell count, and elevated inflammatory markers such as C- reactive protein and erythrocyte sedimentation rate can help guide decision making. Examination of the wound may provide obvious clues like purulent material, unhealthy appearing sheen to the tissues, and foul smell. However, even in the most experienced hands, this determination can be exceedingly difficult; surgeons must rely on the clinical appearance and laboratory markers (Fig. 2-14). Current research into

other markers or mediators of inflammation and infection is ongoing, but until these efforts bear fruit, this assessment continues to be a significant challenge to the surgeon.

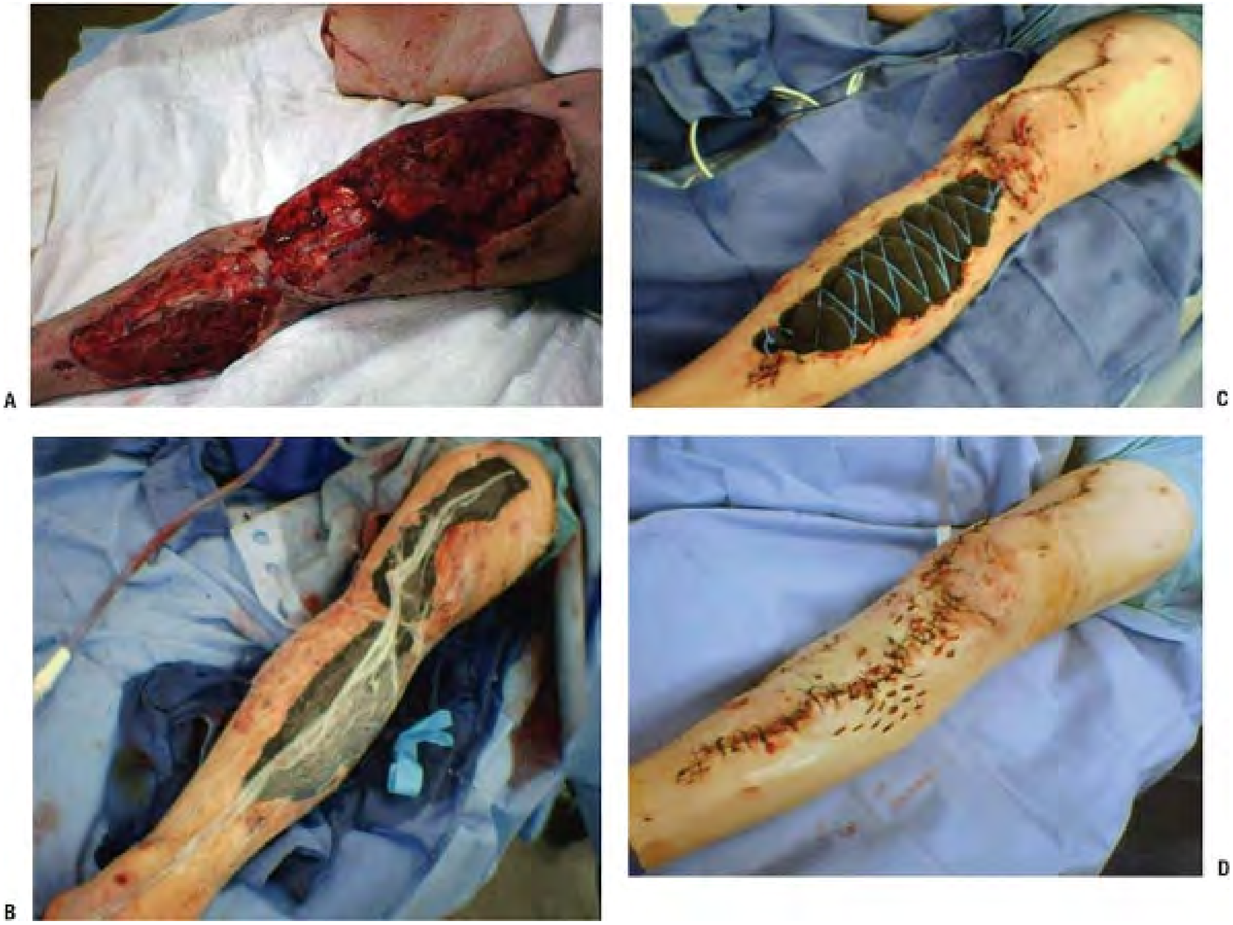

FIGURE 2-13 In some cases tension relieving devices can be used alone or in conjunction with negative pressure dressings to facilitate closure, as in this patient (A) with a large wound involving the anterior aspect of the lower extremity. B,C: Serial debridements and use of a Jacob's ladder with a negative pressure dressing allowed eventual primary wound closure. D: Tension relieving slits in the skin of the lower leg were also used in this case.

Postoperative Management

Postoperative wound care focuses on protecting the healing wound and optimizing medical care to ensure success. Initially, after wound closure or flap coverage, complete soft tissue rest of the involved extremity by splinting or external fixation should be considered (Figs. 2-15 and 2-16). Care should be taken when splinting to ensure functional positions are maintained; the intrinsic plus position for the hand, full extension at the knee, and a plantigrade position for the ankle can facilitate return to function and reduce the need for secondary procedures. Provisional fracture fixation constructs must be checked to ensure integrity of the construct with modifications made as needed to control motion at the fracture site, and ongoing tissue injury.

Elevation of the extremity is critical to reduce local swelling and edema formation. When an external fixation device is present it can be tied to balanced suspension supported by an overhead trapeze. We avoid the use of slings on the lower extremity due to the potential for pressure necrosis.

If an external fixator is not present and a sling is used under the calf, it should always be tied to balanced suspension, not the overhead frame, in this manner. As the patient moves, the pressure on the calf remains constant due to the hanging weights. The skin must be assessed periodically throughout each day (Fig. 2-17).

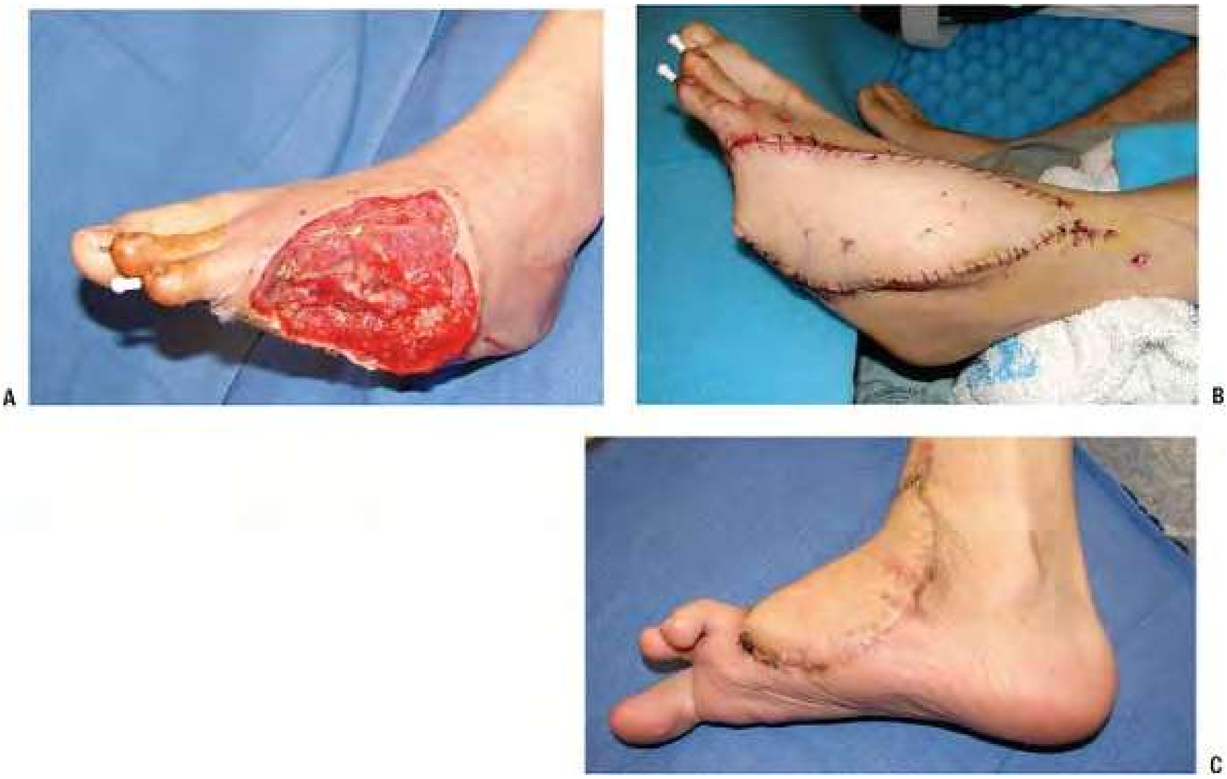

FIGURE 2-14 Aâ€“C: In contrast to Figure 2-4, this foot wound was ready for a lateral thigh free flap as evidenced by the healthy wound bed and margins, minimal residual limb edema, and excellent flap healing.

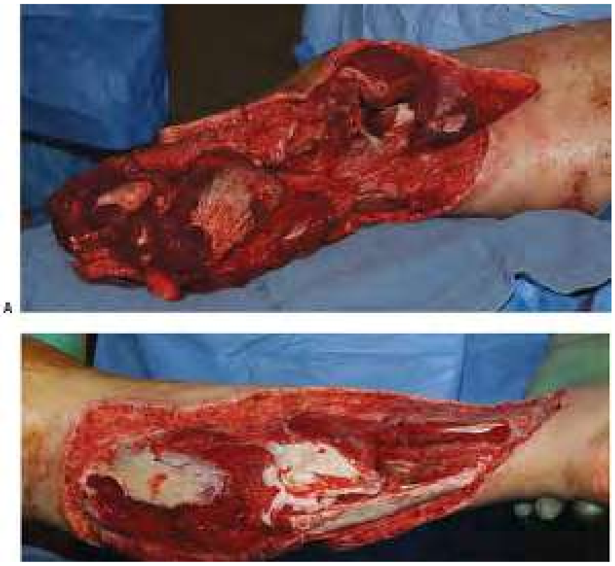

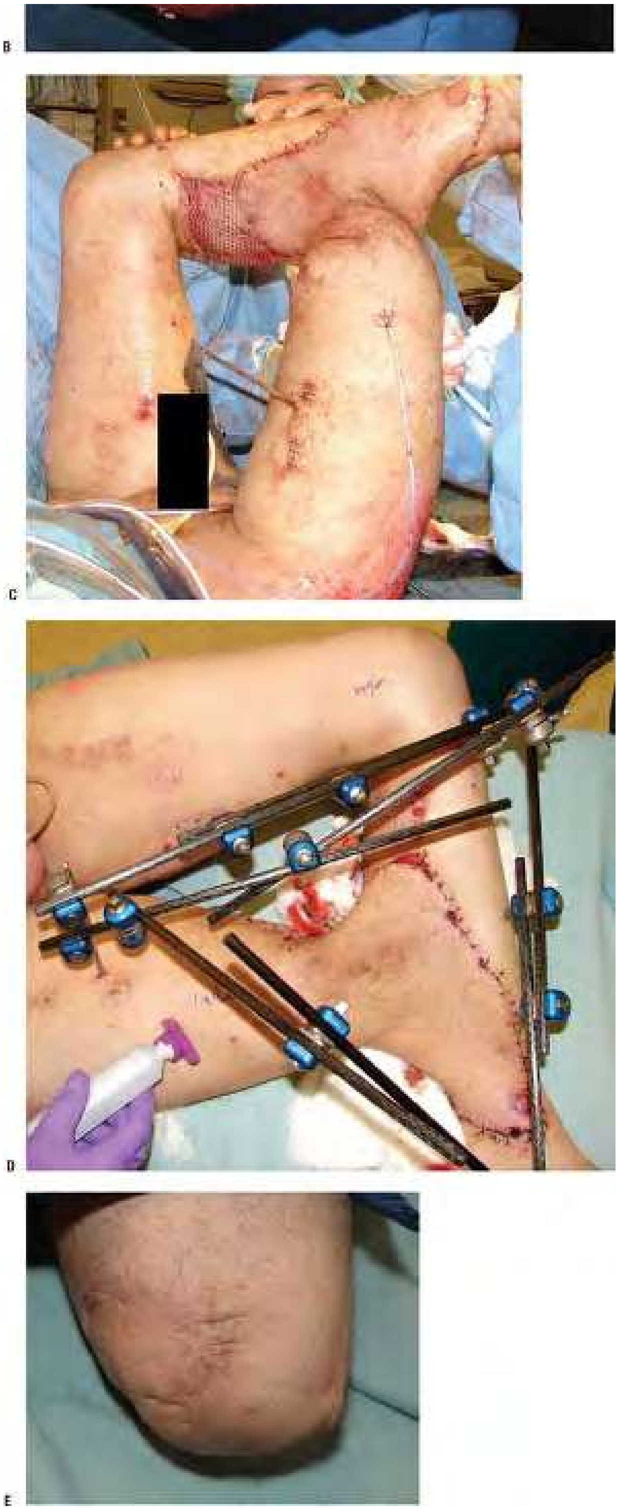

[redacted]

FIGURE 2-15 Patient presented with bilateral lower extremity wounds. His right below knee amputation lacked adequate soft tissue for closure and prosthetic

wear (A); his left lower extremity had a large soft tissue wound (B). The tissue that would have been discarded in a revision amputation to an above knee level was used in a â€œcross-legâ€• flap (C,D) with an excellent clinical result (E). Note the external fixator used to protect the flap after inset (D).

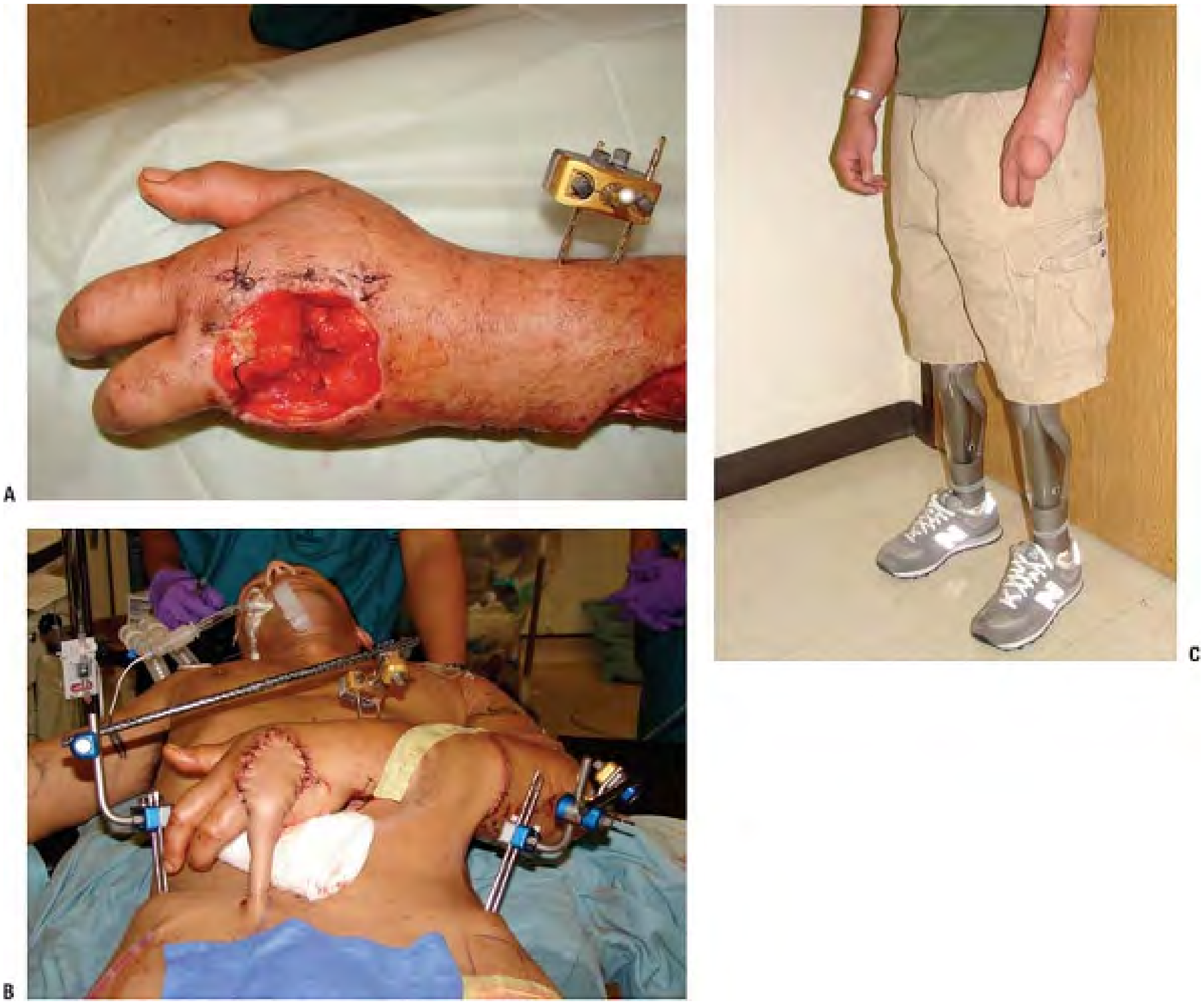

FIGURE 2-16 A: After continued treatment with VAC dressings and ORIF, the left upper extremity wounds from Figure 2-3A,B are ready for definitive coverage. In this case flap selection is limited by the injuries. B: Bilateral groin flaps were used to obtain coverage of both the hand and forearm wounds. To minimize tension on the pedicles, external fixators were used to stabilize the forearm to

the pelvis. C: Healed wounds with durable, healthy coverage.

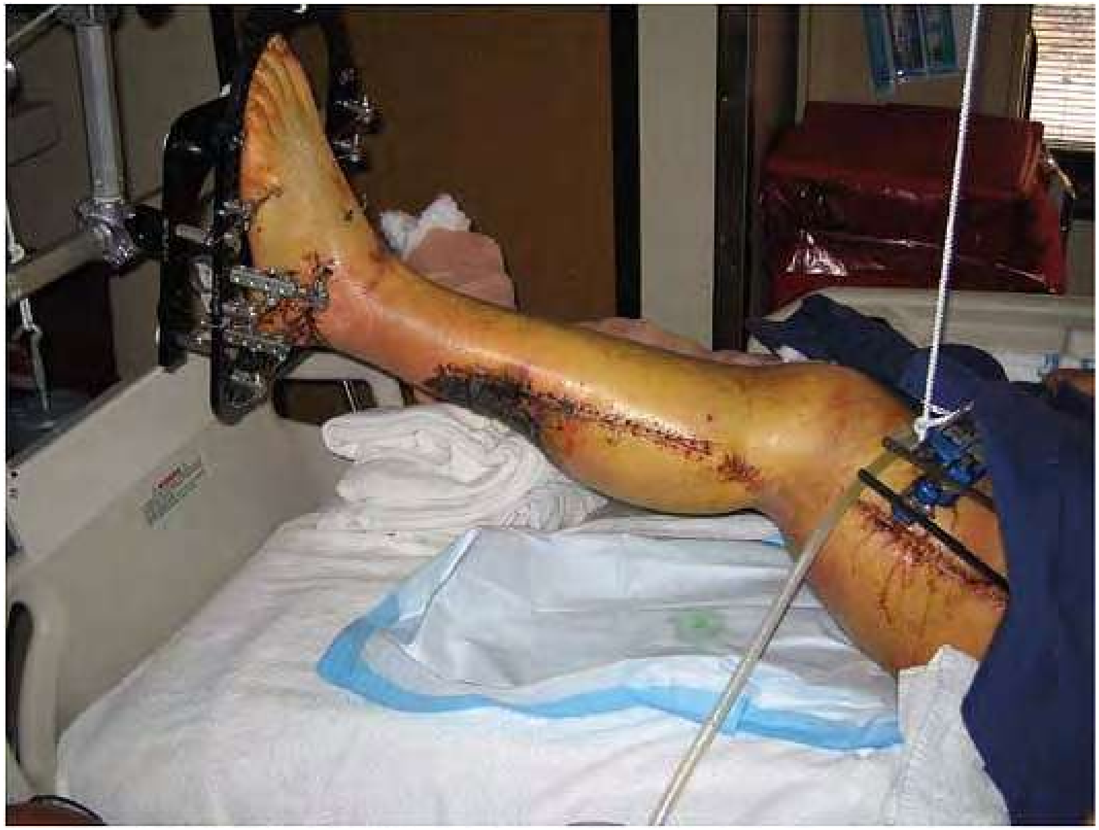

FIGURE 2-17 Elevation is a valuable tool for managing edema. Care must be taken when applying elevation, particularly in the ICU and poly-trauma setting, where patients are often sedated and less mobile. This patient, with multiple bilateral lower extremity fractures eventually received a below knee amputation after development of a full thickness posterior pressure ulcer. The leg was initially elevated by a sheepskin sling tied directly to the overhead bed frame resulting in significant pressure over the posterior calf. Use of a balanced suspension device secured to an external ankle fixator, as demonstrated in the photo, may have prevented this complication.

Mobilization and edema control of the digits should begin as early as possible. Once or twice daily motion begun within seven days of injury can significantly reduce edema and long term stiffness. Burn injuries can be particularly challenging due to pain, contracture, and poor durability of the grafted skin (Fig. 2-18). While consultation with physical and occupational therapists aids in maximizing return of function, a balance between early motion and soft tissue rest must be achieved to ensure proper wound healing. Effective communication with these vital specialists will ultimately benefit both the surgeon and patient.

Patient physiologic factors and comorbidities must continually be reevaluated and optimized in the postoperative period (Figs. 2-19 and 2-20). Markers for infection, hemodynamic status, systemic function, and nutritional status should be reassessed in accordance with the clinical picture. Wound cultures are checked and antibiotic choices are reevaluated in consultation with infectious disease specialists. When using the wound VACâ„¢ system, special attention must be made to the pressure settings and quality of suction. Nursing and ancillary staff education is invaluable. Prominent bone or hardware, especially when combined with inadequate or marginal soft tissue coverage, can lead to wound breakdown and the need for additional surgery (Figs. 2-21 and 2-22).

Pain management is not only humane, but we believe greatly contributes to the success of treatment, especially in cases requiring frequent bedside dressing changes. Reuben et al have shown a reduced incidence of CRPS with aggressive pain management. In many cases, especially trauma, our anesthesia colleagues on the pain service oversee medical management of pain, and provide regional nerve blocks when appropriate. The mental health of the patient must also never be forgotten, especially in cases of traumatic injury, and appropriate consultation with psychiatric specialists and/or a chaplain should be considered to assist in the patient's overall well-being and ability to participate fully in their own recovery and rehabilitation.

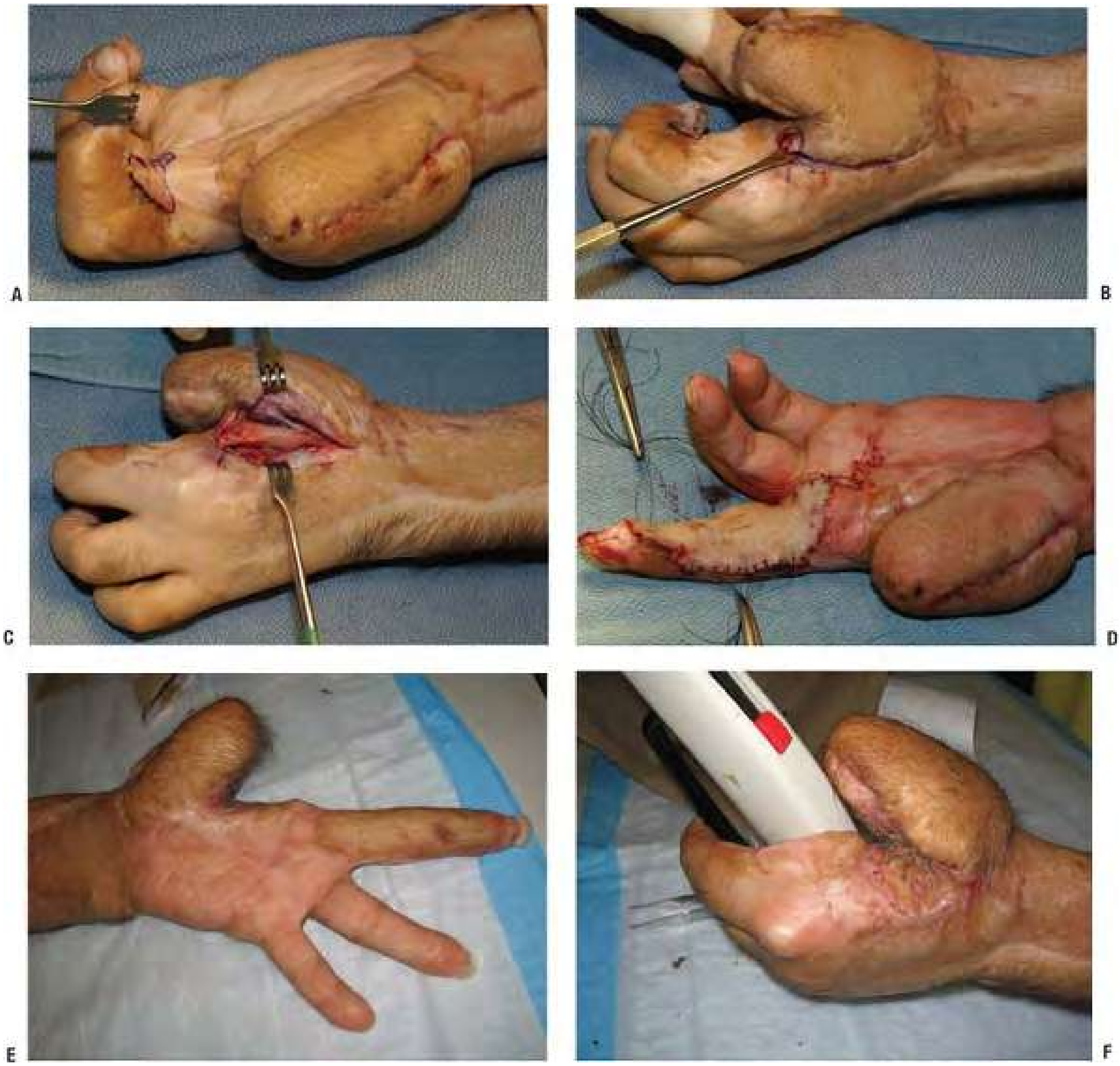

FIGURE 2-18 This patient received severe burns to the upper extremity and chest requiring groin flap to the thumb, and multiple skin grafting procedures complicated by recurrent infections with MRSA. Upon presentation to us he had severe contracture of the long finger (A), erosion of the skin over the residual index metacarpal (B), and essentially no use of the thumb or long finger. Staged reconstruction included thumb metacarpal lengthening, completion of the index ray amputation (C), full thickness skin grafting to the long finger (D), first web-space deepening, thumb carpometacarpal joint arthroplasty, and tendon transfer

to the thumb metacarpal for abduction and opposition. The result: a supple, durable soft tissue envelope and functional hand (E,F). Attention to detail through all phases of care can reduce the need for additional procedures.

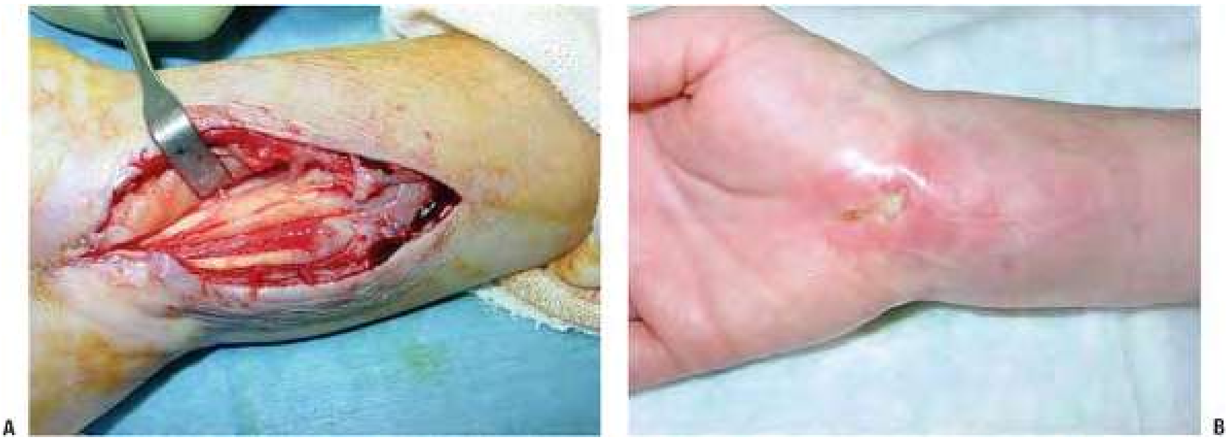

FIGURE 2-19 Underlying medical conditions where immunosuppression is present can lead to wound healing problems even after elective surgery. This rheumatoid arthritis patient underwent routine open carpal tunnel release with subsequent persistent wound drainage. Ultimately, she was diagnosed with Mycobacterium avium. The initial debridement and culture (A) failed to heal fully (B). The gelatinous appearing tenosynovium and indolent course are typical of mycobacterial infections. Successful wound healing followed appropriate antibiotic therapy and delayed primary closure.

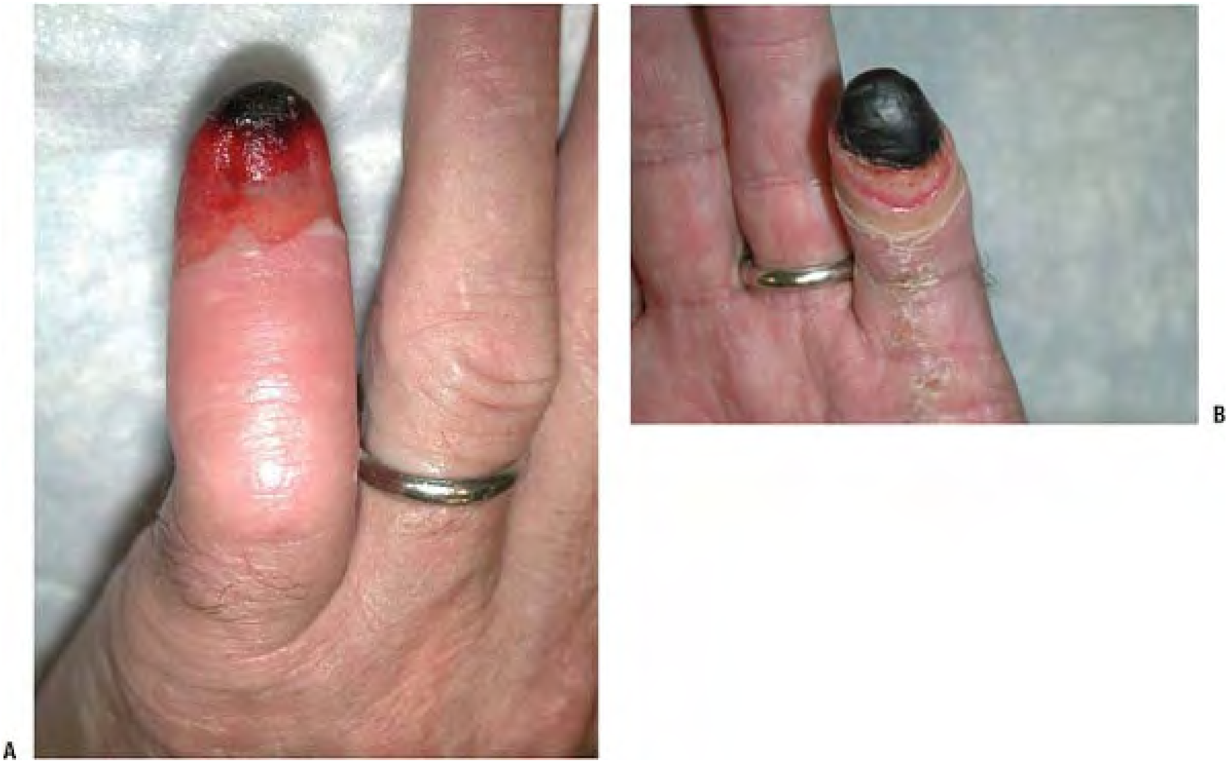

FIGURE 2-20 A,B: Patient had compromised sensation and vascularity following multiple surgeries to the small finger for Dupuytren's disease. The patient sustained a burn to the distal phalanx on the moist heat pads used before therapy. The initial blister and mild erythema were followed by progressive distal tip necrosis; ultimately, the patient requested an amputation of the digit.

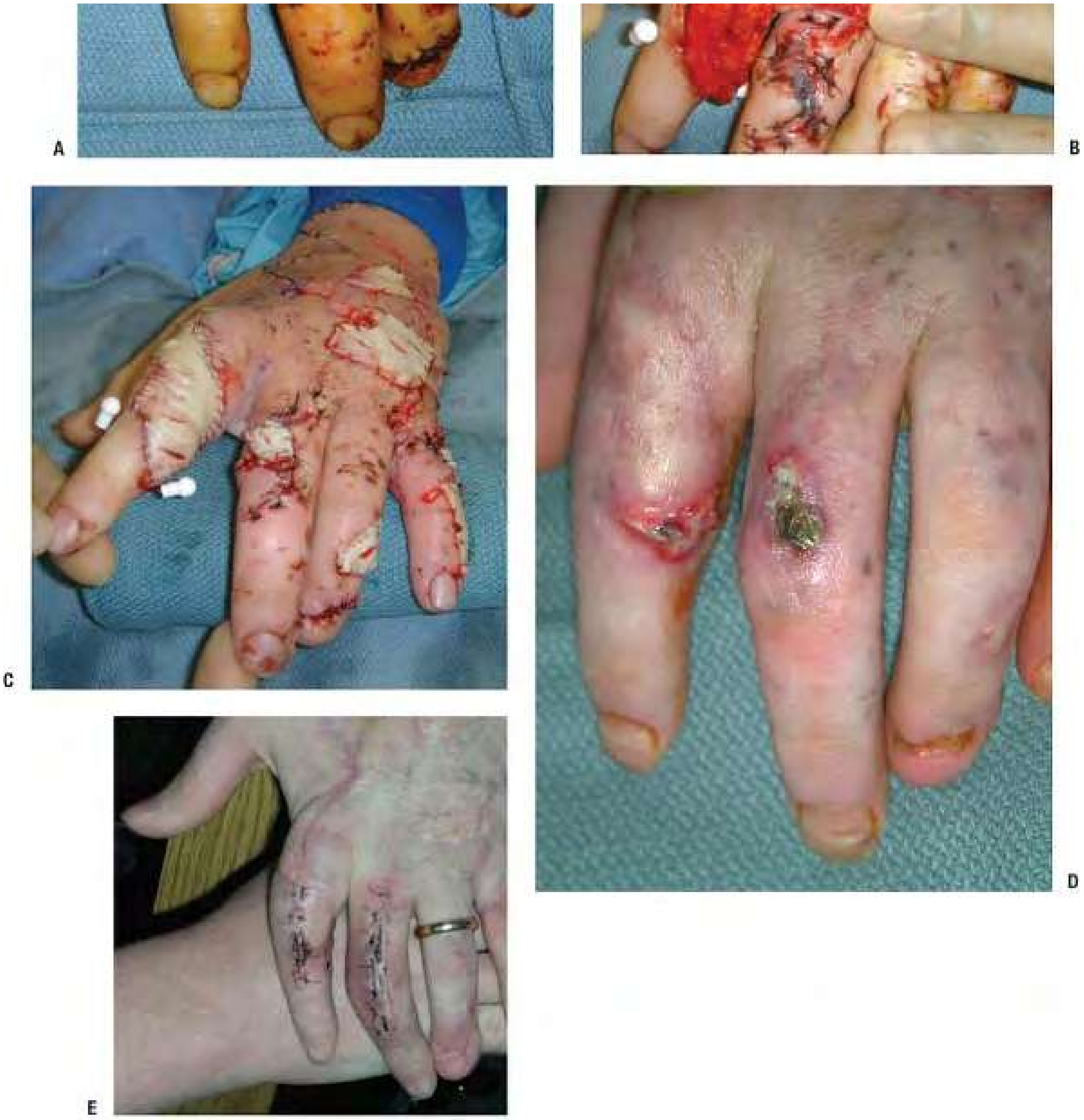

FIGURE 2-21 In this series, the patient's initial wounds were closed primarily but there was inadequate coverage of the PIPJs dorsally (A). A delayed adipofacial flap (B) and overlying full thickness skin grafts taken from his amputated lower extremity (C) were not robust enough coverage over the dorsally applied plates (D). Removal of the dorsal plates and primary closure lead to successful wound healing (E).

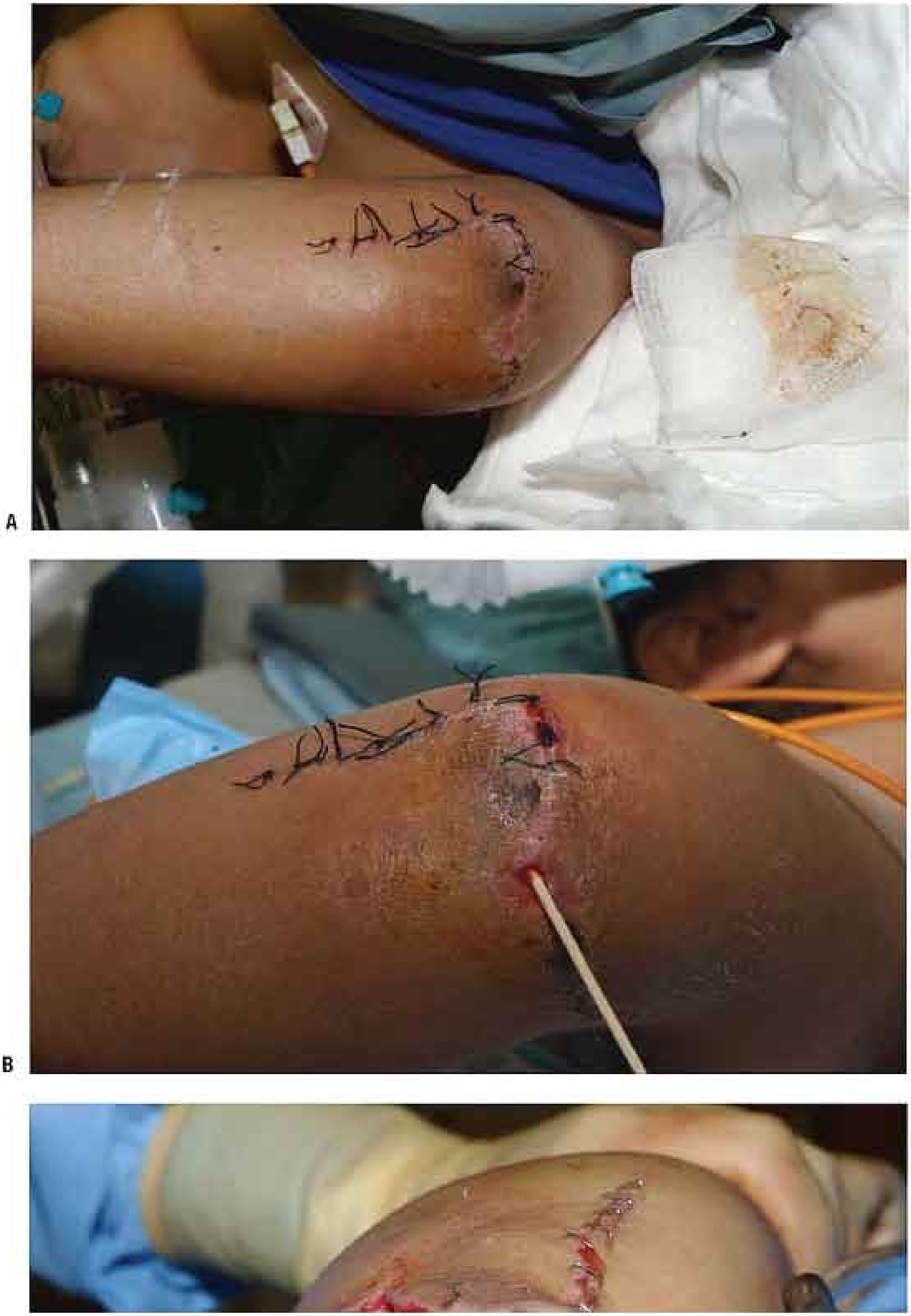
A
B

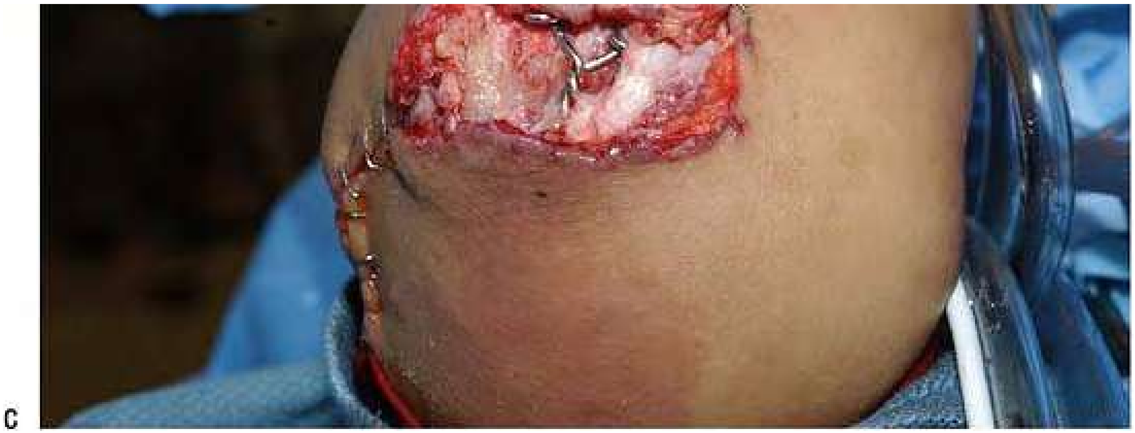

FIGURE 2-22 Patient presented with an innocuous appearing wound and persistent serous drainage 3 weeks after ORIF of an open olecranon fracture overseas (A). Careful inspection showed that the wound communicated with the prominent underlying hardware (B). Irrigation, debridement and IV antibiotics combined with removal of the prominent wire knots (C) and primary closure resulted in uneventful healing of the wound and fracture.

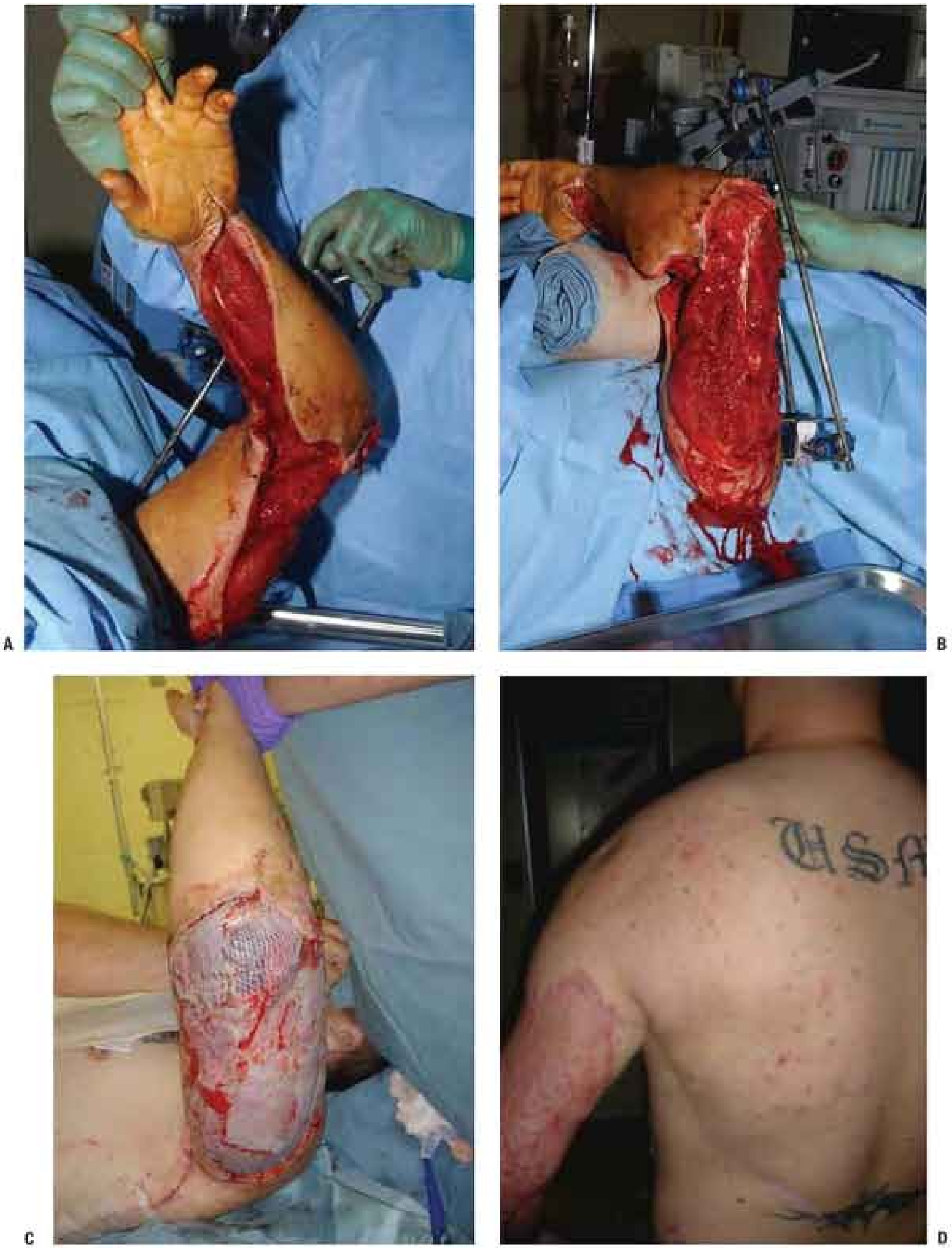
A
B
C
D

FIGURE 2-23 A-D: Patient suffered a blast injury resulting in an open fracture of the elbow, in addition to severe soft tissue injury and a forearm fasciotomy. He was initially managed with a spanning external fixation device and serial debridements followed by ORIF. (ORIF; pedicled Latissimus Dorsi flap and split thickness skin grafting.)

Conclusion

Treatment of wounds can be a challenging enterprise, consuming considerable time, energy, and resources. By combining the basic principles of evaluation and treatment set forth in this chapter, and with meticulous attention to detail, the surgeon will have an excellent starting point for the treatment of complicated soft tissue trauma (Fig. 2-23).

Disclaimer. The views expressed in this chapter are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

Recommended Reading

Anglen JO. Wound irrigation in musculoskeletal injury. *J Am Acad Orthop Surg.* 2001;9:219â€“226.

Argenta LC, Morykwas MJ. Vacuum-assisted closure: a new method for wound control treatment: clinical experience. *Ann Plast Surg.* 1997;38(6):563â€“576.

Attinger CE, Janis JE, Steinberg J, et al. Clinical approach to wounds: debridement and wound bed preparation including the use of dressings and wound-healing adjuvants. *Plast Reconstr Surg.* 2006;117(Suppl.):72â€“109S.

CDC. Preventing tetanus, diphtheria, and pertussis among adults: use of tetanus toxoid, reduced diphtheria toxoid and acellular pertussis vaccines. Recommendations of the Advisory Committee on Immunization Practices (ACIP)

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part I - Overview and Foundation > Chapter 3 - Management of Simple Wounds: Local Flaps, Z-Plasty, and Skin Grafts

---

Chapter 3

Management of Simple Wounds: Local Flaps, Z-Plasty, and Skin Grafts

John B. Hijjawi

Allen T. Bishop

Skin grafting and local flap coverage have remained a common means of covering traumatic wounds. Skin grafts, local flaps, and random flaps are not as simple as direct wound closure nor are they are as â€œelegantâ€• as procedures that are found higher on the reconstructive ladder such as pedicled flaps or free tissue transfer (Fig. 3-1). However, the procedures described in this chapter remain straightforward, reliable, and time-tested and should be an essential component of any surgeon's reconstructive armamentarium.

Skin Grafts

Skin grafts are classified by source and thickness. By far the most common, durable, and successful skin grafts are autografts harvested from the patient's own skin. There are no immunologic issues, since the tissue comes from the patient's own body. Concerns over disease transmission are eliminated and expense is minimal. The only disadvantage is the creation and care of a donor site.

Skin grafts are also available as cadaveric allografts, xenografts (typically porcine skin), and most recently, cultured epithelial grafts. These materials are lifesaving sources of temporary wound coverage in the context of massive burns. Allografts and xenografts do have the disadvantage of immunogenicity and thus impermanence since they are bound to be rejected. However, if the quality of a wound bed is questionable, preserved porcine xenografts may be an excellent option for temporary wound coverage.

Currently available cultured epithelial grafts are expensive, require time to culture, and are not as durable as autografts. Additionally, studies have shown that they are significantly more susceptible to infection than standard autografts.

Skin grafts are also classified based on thickness. The skin is composed of an outer epidermis and a deeper dermis, which is further subdivided into the reticular and papillary dermis (Fig. 3-2). All skin grafts consist of the entire epidermis and a variable amount of dermis. Split-thickness skin grafts contain only a portion of the dermis, whereas full-thickness skin grafts contain the entire dermis and dermal appendages. As a result, full-thickness skin grafts continue to support hair growth following transfer. This needs to be carefully considered when selecting donor sites in situations when a full-thickness skin graft is to be transferred to a conspicuous, previously hairless area.

Since full-thickness skin grafts include all dermal appendages, skin at the donor site will not spontaneously regenerate following harvest; therefore, these donor sites need to be closed primarily or with a split-thickness skin graft. Split-thickness donor sites retain the ability to generate epithelium and will be largely healed within several weeks if cared for properly.

Split-thickness and full-thickness skin grafts have quite different contractile characteristics on harvesting (primary contraction) and after they heal (secondary contraction). Primary contraction is

the initial contraction of a graft when it is harvested. Due to the greater proportion of elastic fibers in a full-thickness skin graft (the dermis is the location of all elastic fibers in the skin), it will undergo more primary contraction than a split-thickness skin graft when harvested. Similarly, a thicker split-thickness skin graft (e.g., 0.018 in) will undergo more primary contraction than a thin split-thickness skin graft. Most skin grafts can be stretched under minimal tension at inset to overcome this contraction, restoring their original size.

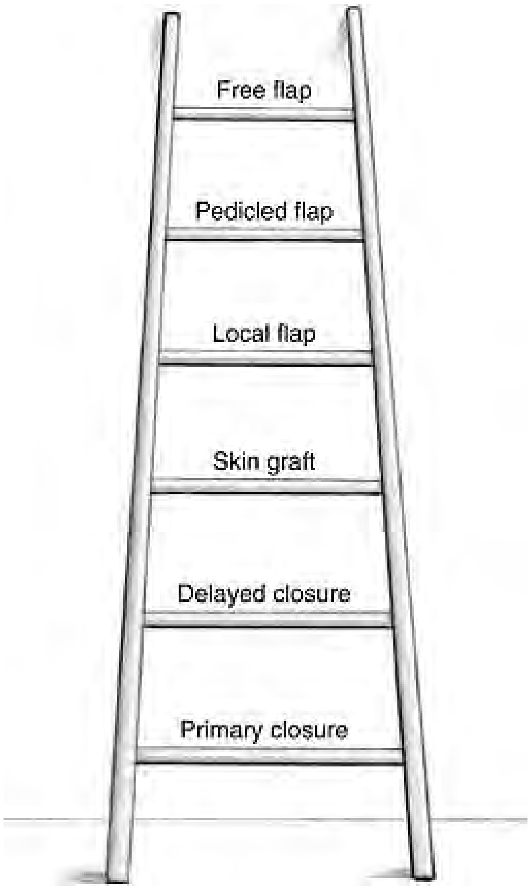

FIGURE 3-1 Reconstructive ladder. Historically, surgeons have closed wounds with the simple procedure first, moving up the rungs of the reconstructive ladder as wounds become larger and more complex.

Conversely, split-thickness skin grafts undergo more secondary contraction than full-thickness grafts. This can be exploited to provide gradual contraction of a wound over the course of several months. An example is a fasciotomy wound that is under too much tension to close primarily within the days following compartment release. A very thin split-thickness skin graft applied to the wound will undergo significantly more secondary contraction than would a full-thickness skin graft, resulting in contraction of the wound itself. After several months, this may result in a wound that is small enough to allow serial excision of the skin graft and primary closure under minimal tension. In comparison, full-thickness skin grafts can be relied on to undergo virtually no secondary contraction in situations where this is not desirable, such as across a joint surface or in a web space.

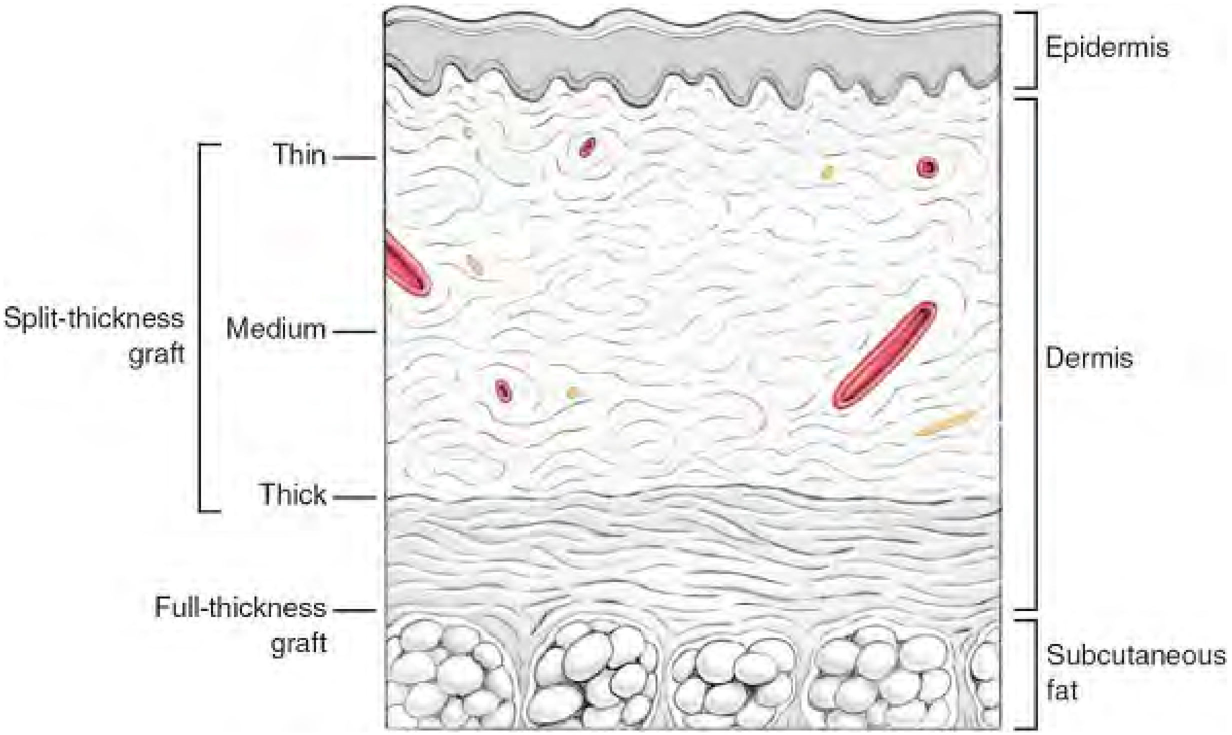

FIGURE 3-2 Skin anatomy.

Initial adhesion of a skin graft is the result of fibrin present between the graft and the recipient bed. Initial survival of a skin graft relies on plasmatic imbibition, which is the process of nutrient diffusion from the recipient site into the skin graft. Later, vascular channels within the graft line up with vascular channels in the recipient bed through the process of inosculation. Finally, long-term graft survival relies on neovascularization, or the process of new blood vessel growth into the skin graft from the recipient bed. Because skin grafts, both split thickness and full thickness, are completely reliant on the recipient bed for survival, the wound bed must be well-vascularized and free of infection to support skin graft take. Split-thickness skin grafts have lower metabolic demands than do full-thickness grafts, and so they donâ€™t require recipient beds with as rich a blood supply. Along the same lines, full-thickness grafts take longer, from 7 to 10 days, to heal. Split-thickness grafts are generally considered healed by 5 days and should be left immobilized and dressed at least that long postoperatively. The time of healing in specific situations depends most on the quality of the recipient wound's blood supply.

Indications/Contraindications

Indications

Skin grafting may be indicated for any defect that cannot be closed primarily and that has a wound bed that can support skin graft take (Table 3-1). Skin grafts survive for the first several days through a process called imbibition. During this stage of skin graft healing, the graft obtains nutrients from the underlying wound bed through a process of diffusion. Wound beds devoid of blood flow, or with little vascularized tissue, will make for poor recipient sites. Exposed structures that will accept a graft include subcutaneous tissue, paratenon, and muscle. Other tissues, such as exposed bone, joint, tendon, and nerve, may be covered temporarily by graft used as a biologic dressing but will not support a graft for permanent coverage. Beds containing tissues of questionable viability, chronic granulation tissue, or frank infection can accept a skin graft but require thorough debridement prior to graft placement. Wounds that contain fewer than 10^5 bacteria per gram of tissue or that allow xenograft adherence within 24 hours allow successful skin grafting.

Contraindications

Skin grafts are contraindicated in areas that are exposed to repetitive trauma or that lie over osseous prominences. Bone, cartilage, and tendons denuded of periosteum, perichondrium, or paratenon cannot be covered with skin grafts as there is inadequate vascular supply to support healing of the skin graft. Controversy exists over whether skin grafts should be placed over bone, cartilage, or tendons with healthy periosteum, perichondrium, or paratenon. It is certainly possible to get skin-graft healing over such structures. However, skin grafts in these situations are rarely optimal for long-term durable coverage. In addition, skin grafting should be avoided in areas that may require secondary surgery for bone or nerve grafting, as adherence to underlying muscle, nerve, and tendon may complicate secondary surgery. All split thickness grafts will undergo some component of contracture over time; thus, if these grafts are placed over large areas of the antecubital fossa, popliteal fossa, or olecranon, there is a risk of limitation in joint motion.

Preoperative Planning

The wound must be debrided and clean prior to attempts at skin grafting. Infection is one of the leading causes of skin graft failure. Since skin grafts are completely dependent on the wound bed they are transplanted to for nutrition, they possess no intrinsic ability to resolve infection. Quantitative wound cultures have been used for many years in some centers to determine the adequacy of a wound's microenvironment for closure. A quantitative culture revealing less than 10^5 bacteria per gram of tissue has been traditionally regarded as an acceptable level of colonization below which a wound can be closed by skin graft or local flap. However, such cultures are highly dependent on the experience of the technician performing them. A careful clinical evaluation and serial sharp debridement of clinically infected or contaminated wounds are advised.

Table 3-1. Wound Analysis

Is there too much tension across the wound for primary closure?
Is there adequate perfusion of the wound bed?
Are vital structures exposed at the base of the wound?
Is the wound infected or contaminated?
Is the wound geometry favorable for closure?

Donor site selection must also be decided before surgery. Considerations when choosing a donor site for split thickness skin grafts include cosmesis, the thickness of skin in the donor site region, and ease of care of the donor site after graft harvest. The most common split thickness skin donor sites include the buttocks, lateral and anterior thighs, and lower abdomen. These sites are relatively easy to conceal, have thick skin resulting in less pigmentation once healed, and are readily accessible in even a bed-bound patient, easing postoperative care.

Full thickness skin grafts are most commonly harvested from the groin, antecubital fossa, volar wrist crease, medial arm, postauricular sulcus, or lower abdomen. These donor sites all exist in areas where closure can be performed within pre-existing skin creases, thus resulting in relatively inconspicuous donor sites. The hypothenar skin or plantar instep offers the unique quality of glabrous skin if needed for graft material. Always keep in mind that any amputated â€œspare partsâ€• can provide a good source of viable skin graft with no added morbidity to the patient.

Surgery

Patient Positioning

The patient's position will depend on the location of the graft to be harvested. Most frequently we harvest split thickness grafts from the upper thigh area, where they may easily be concealed under clothing A supine position with a roll underneath one hip is ideal. The majority of full-thickness grafts are harvested from the hairless skin of the groin crease, though the inner upper arm may be used as well. The patient is positioned supine for such harvests. For glaborous skin grafts, the instep of the foot may be positioned so as to allow ease of harvest. For wounds on the posterior aspect of the lower extremity or trunk, a lateral decubitus position readily exposes both the wound and a lateral thigh donor site. With careful planning it is almost never necessary to reposition a patient after harvesting the skin graft.

Split Thickness Grafts

Power dermatomes are the most common method of harvesting split thickness skin grafts, although for very small split thickness grafts hand-driven Weck blades may be more convenient.

Measure the recipient wound and choose a dermatome guard based on that measurement (Fig. 3-3).

- Set and check the dermatome thickness (usually 0.010 to 0.015 in) with a No. 15 scalpel blade. The thin, beveled edge of the knife is about 0.010 in, whereas the thickest portion of the blade is 0.015 in thick (Fig. 3-4).
- Mark the donor site with a ruler so that you will know where the dermatome needs to â€œtouch downâ€• and â€œlift off.â€• Relying on your memory or estimating how far you will have to drive the dermatome can lead to harvesting too little graft and having to reharvest, creating an unnecessary seam in the graft. In cases of overharvesting, the extra graft can be replaced onto the donor site, but this complicates the donor site dressing.

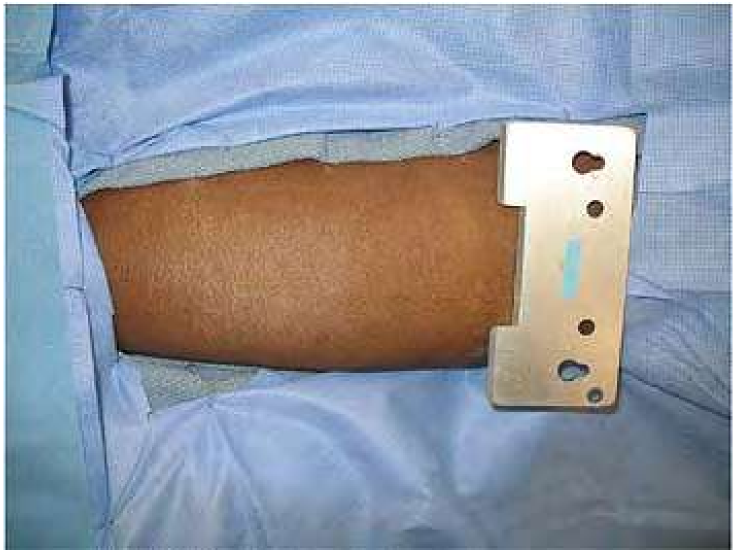

FIGURE 3-3 A dermatome guard of appropriate width is chosen.

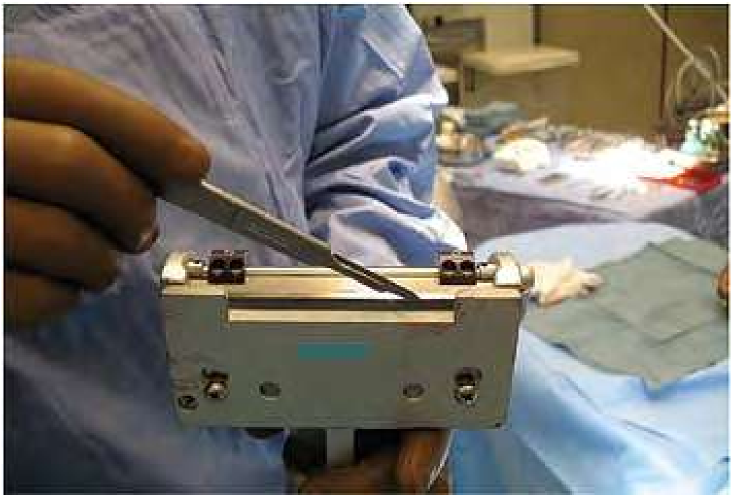

FIGURE 3-4 The thickness setting is double checked.

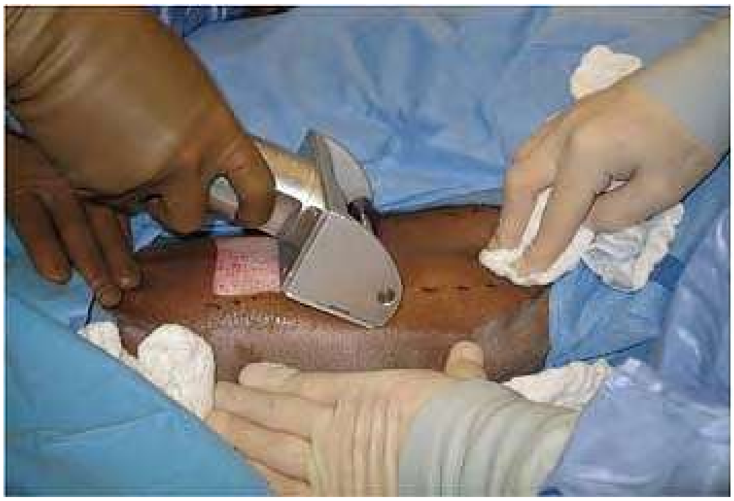

FIGURE 3-5 Countertraction is applied to avoid â€œskippingâ€• along the skin.

- Clean the donor site to remove any sticky material that will cause the dermatome to stick and apply copious amounts of mineral oil to the donor skin and the dermatome.

- Apply countertraction to the skin in front of and behind the dermatome blade. Dermatomes, particularly when fitted with larger guards, function much more effectively on flat surfaces (Fig. 3-5).
- Activate the dermatome before â€œtouching downâ€• on the skin and plan to keep it activated until after â€œlifting offâ€• of the skin. Touch down at 45 degrees to the skin, then slightly flatten the angle between the dermatome and skin, maintaining constant firm pressure on the head of the dermatome.
- Realize that if you need to reset a hand for countertraction, you can stop the blade without lifting off of the skin and reset countertraction. The harvest can then be continued without interrupting the continuous sheet of skin graft.
- Transfer the harvested graft onto a dermal carrier with the dermis side up. This side is shinier, and has less friction when rubbed than the epidermis side. If you become confused as to which side is the dermis side, realize that the skin graft edges will always roll toward the dermis side.
- Either make several small slits in the skin graft with a scalpel (â€œpie crustingâ€•) to allow for drainage of accumulated fluid from the wound bed, or run the graft and dermal carrier through a skin graft mesher (Fig. 3-6). Most typically, grafts are meshed at a 1:1.5 ratio. This is done to allow drainage through the graft, to make grafts more conformable to the underlying wound bed, and to increase the area a graft can cover. It is not necessary, however, and many surgeons avoid it since the meshed appearance will be permanently obvious and will significantly compromise the final cosmesis.
- Fix the graft dermis side down to the wound bed with either staples or absorbable sutures such as 5-0 chromic. Traditional bolster or â€œtie-overâ€• dressings employ silk suture placed circumferentially around the skin graft and left long. They are then tied over mineral oil soaked cotton wrapped in a non-adherent dressing and placed firmly onto the skin graft (Figs. 3-7,3-8,3-9). A very convenient bolster dressing can be fashioned by placing a non-adherent Nterface dressing (Delasco, Council Bluffs, IA) over the skin graft, followed by mineral oil soaked cotton pushed firmly into the wound bed to compress the skin graft. Finally, a Reston (3M, St. Paul, MN) sponge can be cut to conform to

the wound, placed adhesive side up, and stapled to the skin surrounding the wound. This is a very stable bolster construct that resists shearing forces and provides firm compression. For skin grafts placed on an extremity, an Ace wrap can be used to protect this entire dressing.

- Appropriate splints should be applied to immobilize the recipient site until the skin graft is totally healed.

Full Thickness Grafts

Full thickness grafts are harvested after drawing an ellipse, which includes the necessary amount of skin based on the wound measurements. Closure without dogears is best achieved when the length of the ellipse is about four times the width of the ellipse. A scalpel is used to incise the skin just through the dermis. One end of the ellipse is then lifted firmly with a single skin hook or toothed forceps and a fresh scalpel is used to elevate the skin graft just deep to the dermis, taking as little fat as possible with the skin graft. Once harvested, the full-thickness graft can be rolled over a finger or shot glass with the dermis side up, and a curved iris scissors used to remove every bit of residual fat from the dermis.

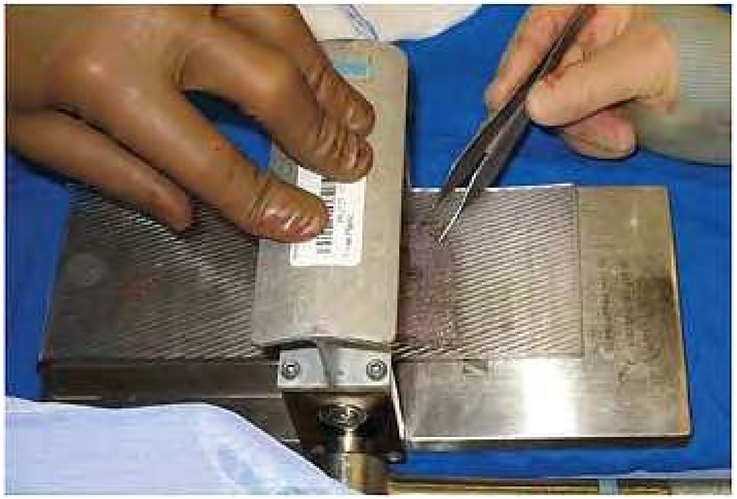

FIGURE 3-6 Meshing the skin graft dermis side up facilitates placement onto the wound.

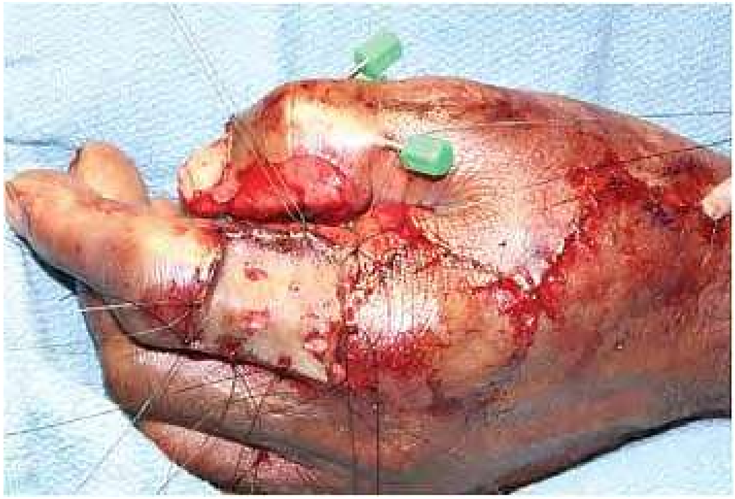

FIGURE 3-7 The bolster or â€œtie-overâ€• dressing begins with placement of long sutures along the edge of the skin graft.

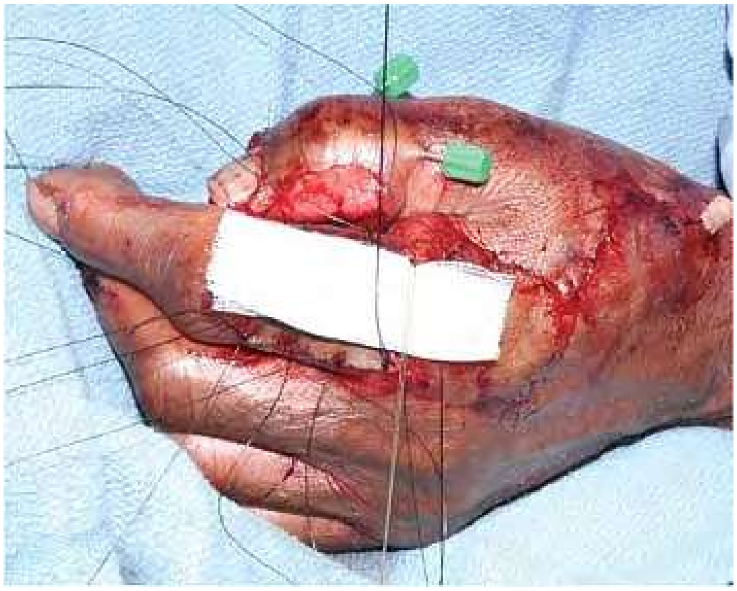

FIGURE 3-8 Cotton balls are wrapped in a non-adherent dressing and compressed onto the skin graft.

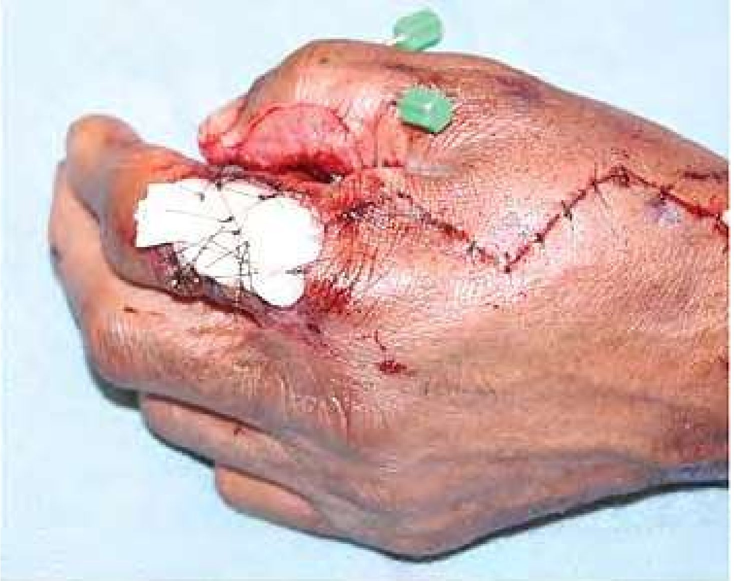

FIGURE 3-9 The completed tie-over.

Full thickness grafts are not typically meshed. Otherwise, they are fixed into position as described for split thickness grafts. They should be left undisturbed for 7 to 10 days. Donor sites are best treated with primary closure after moderate undermining of the wound edge.

Postoperative Management

Split thickness skin grafts should not be disturbed for a minimum of 5 days. The dressing can then be carefully removed to not disturb the healing skin graft, and once daily dressings are begun with an antibacterial ointment and a non-adherent dressing such as Adaptic (Johnson and Johnson, New Brunswick, NJ). Full thickness grafts should remain covered for 7 to 10 days before removing the split or bolster.

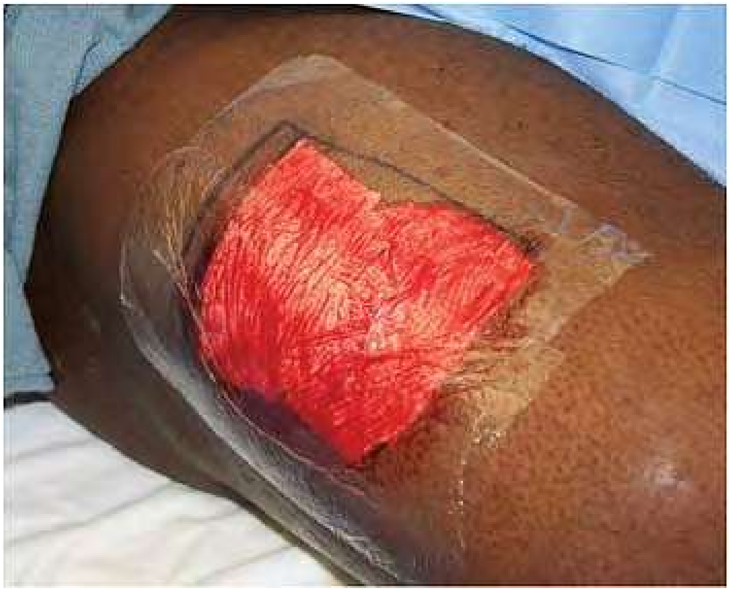

FIGURE 3-10 Donor site dressing.

Donor sites up to 100 cm^2 can be covered with Tegaderm (3M, St. Paul, MN) as long as care is taken to carefully dry the skin surrounding the donor site before applying the dressing. Serum will collect under the Tegaderm after several days, and can be left alone. If the Tegaderm leaks, a patch can be placed over the hole. The benefit of Tegaderm is a completely isolated, moist environment that is virtually painless for the patient (Fig. 3-10). The donor site is typically re-epithelialized by 2 weeks at which point the Tegaderm can be removed and replaced with once daily application of a moisturizing cream and non-adherent dressing. Alternatively, the donor site can be covered with a Xeroform (Sherwood Medical Industries Ltd., Markham, Ontario, Canada), which will dry into an eschar when exposed to air. Once dry this will be painless but is extremely sensitive until the eschar develops.

Rehabilitation

Rehabilitation of the affected extremity can begin once the surgeon is assured of stable graft take. Grafts placed over joints may benefit from 7 to 10 days of immobilization before initiating motion across the joint. For grafts placed away from a joint, normal motion can begin almost immediately if the graft has been securely bolstered to its wound bed.

Complications

Skin graft necrosis or failure is the most common complication following split thickness skin grafting. The formation of fluid under a skin graft, whether hematoma or seroma, is the most common cause of skin graft loss. As noted, the graft is completely reliant on the recipient bed for nutrition and so needs to be in complete contact with the recipient bed to survive. Therefore, precise hemostasis of the wound bed is critical before skin graft application as is firm compression of the graft onto the recipient bed through the use of bolster dressings.

Infection is the next most common cause of skin graft loss. This is best avoided by careful debridement before placement of the graft. Shearing forces can interrupt the formation of vascular connections between the graft and the recipient bed. This will ultimately lead to loss of a graft since healing relies on the formation of genuine vascular connections and is not possible through plasmatic imbibition alone. Therefore, firm immobilization of all skin grafts is critical to their healing.

Finally, an inadequately perfused recipient bed will certainly lead to skin graft loss. Common situations include patients with peripheral vascular disease, previously radiated tissues, and tendon or bone denuded of paratenon or periosteum.

Random Flaps

Random flaps, by definition, have no named or defined blood supply. They are raised in a subdermal plane and so rely on the subdermal vascular plexus of skin for circulation. To ensure adequate circulation, random flaps should be limited to a length no greater than 2.5 times the width of their base, which is the uncut border of the flap. This ratio may be even more limited in poorly perfused extremities.

Indications/Contraindications

Indications

The elasticity and slight redundancy of local tissue are critical to the success of most random flaps. They are based on local tissue so are typically contraindicated for wounds with significant surrounding soft tissue damage such as radiation wounds. The advantage of random flaps in comparison to skin grafts is that they can provide well-vascularized, full-thickness tissue for wound coverage. Far more durable to friction and repeated stress than skin grafts, local flaps are preferred for closure over vital structures such as tendons, nerves, and blood vessels. An added benefit is that they can be used to close wounds over vital structures denuded of paratenon, periosteum, or perineurium, since a well-vascularized wound bed is not required as it is for skin graft survival. Finally, despite the new scars made to raise random flaps, they are ideal in terms of color match, since they come from local skin.

Contraindications

Gross contamination or frank infection is an absolute contraindication to wound closure with local flaps. In fact, while free flaps and pedicled flaps have been shown to introduce enough independent new blood supply to overcome infectious processes like osteomyelitis, local flaps need to be placed over a clean, uncontaminated wound bed. Finally, local flaps are limited to small wounds usually less than 15 to 20 cm^2; wounds exceeding these dimensions should be covered with another method of soft tissue coverage.

Preoperative Planning

As mentioned previously, adequate surgical debridement is a must before any attempt at closure. Tissue culture is a useful adjunct, particularly in cases of significant blunt trauma with devitalized tissue or cases of gross contamination. Often, serial wound debridement every other day for several days is necessary to obtain negative cultures.

It may not be possible to completely resolve a patient's systemic medical issues, but they should be as optimized as much as possible. Tight control of blood glucose levels in patients with diabetes, management of extremity edema, and nutritional status can all be significantly improved in many patients with several days of focused inpatient care.

Surgery

Z-Plasty

Probably the most familiar random flap to any surgeon, the Z-plasty is not actually indicated for the treatment of open wounds. Rather, the main indications for a Z-plasty include lengthening scars, interrupting linear scars with the transposition of unscarred tissue, and disrupting circumferential or constricting scars (Table 3-2). A common indication in hand surgery is to employ a Z-plasty in contracted web spaces in an effort to introduce healthy adjacent tissue relieving web contractures.

Z-plasties rely on limbs of equal length to facilitate closure. The most common Z-plasty design employs 60-degree angles, which theoretically results in a 75% increase in the length of the central limb of the Z-plasty (Fig. 3-11). Although clinically not feasible, a Z-plasty with angles of 90 degrees would result in the greatest theoretical gain of central limb length, approximately 120%.

Executing an effective Z-plasty is not simply a matter of elevating the triangular flaps and transposing them. Mobility and a tension-free closure are greatly facilitated when the tissue at the bases of the triangular flaps is also elevated.

Four Flap Z-Plasty

Most commonly used for first web space contractures, the four flap Z-plasty results in a 150% gain in length of the original central limb, or scar contracture. Essentially, a 120-degree standard Z-plasty is drawn. Each triangular flap is then bisected, resulting in four equivalent 60-degree triangles that are raised and interdigitated as shown (Fig. 3-12).

Table 3-2. Z-Plasty Indications

Lengthening scars
Interrupting linear scars
Disrupting circumferential or constricting scars

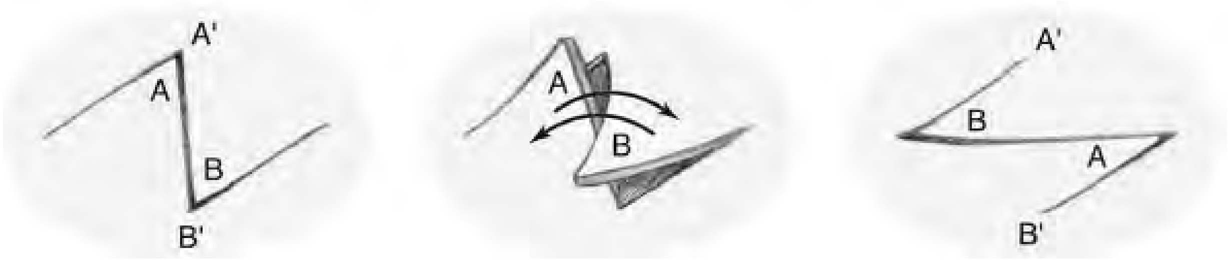

FIGURE 3-11 A basic Z-plasty.

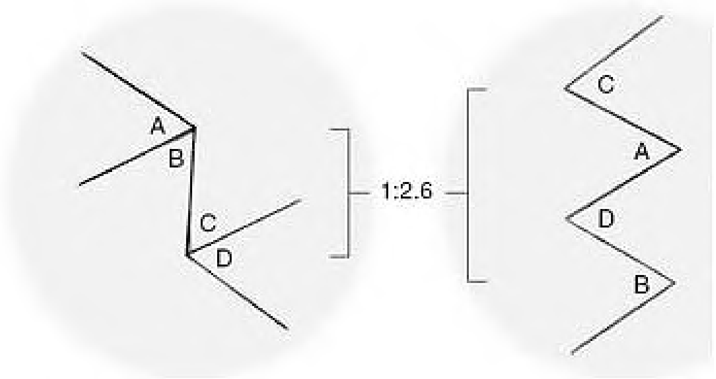

FIGURE 3-12 Four-flap Z-plasty.

Rhomboid Flap

Executing a rhomboid flap begins with converting the defect (even if it is circular) into a rhomboid to better visualize the flap design. A line equivalent in length to the limbs of the rhomboid is then drawn perpendicular to the short axis of the rhomboid (Fig. 3-13). Next, a line (B-C) is drawn at 60 degrees to the A-B line. It will be parallel to one of the limbs of the original rhomboid.

The flap is then elevated. It is also critical to elevate the skin around the base of the flap very liberally to facilitate transposition of the flap and closure of the wound. The flap should not be closed under tension; rather, further undermining of the base of the flap should be executed until the flap closes without tension.

Banner Flap

The banner flap is a type of transposition flap. A pendant or banner of skin is designed with one edge of the banner (near the base of the flap) running tangentially to the wound edge. The flap is elevated and transposed after which any redundant flap can be trimmed. It is important that the original design of the flap places the banner in an area of redundant skin. Since the banner can be designed along any border of the defect, it is helpful for eventual cosmesis if the original banner is designed within a relaxed skin tension line.

Rotational Flap

Rotational flaps are frequently employed on the dorsum of the hand, fingers, and in the scalp to close triangular defects. They are deceptively simple, and poor planning can lead to a large incision with an inadequate flap.

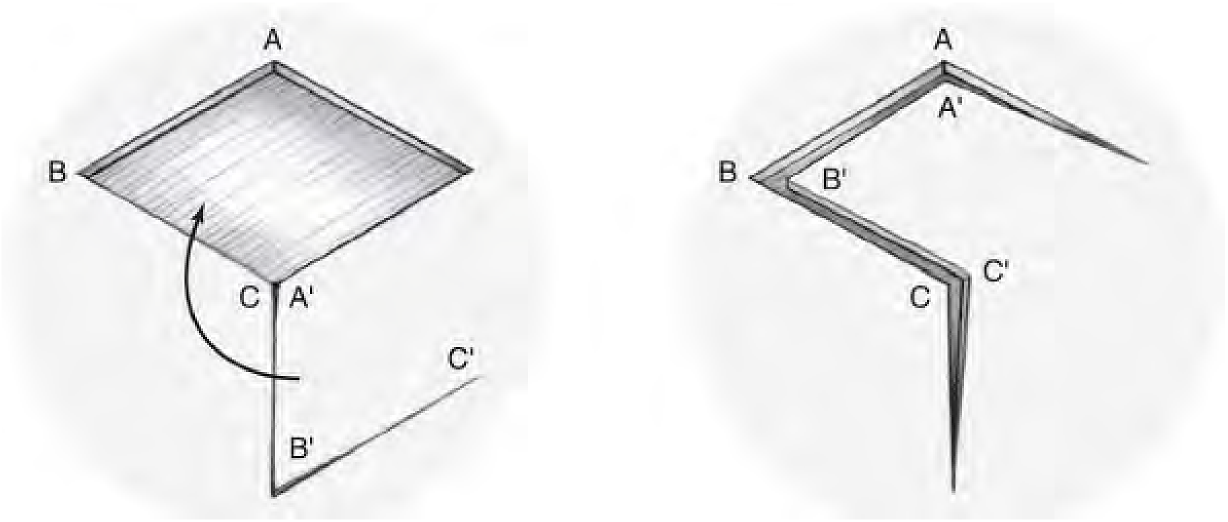

FIGURE 3-13 Rhomboid flap.

Unfortunately, relatively little advancement of the flap edge is possible without making a flap that is approximately four times greater in length than the defect length. Even with an adequately designed flap, tension can develop at a point opposite the pivot point in the base of the flap. This can be overcome by backcutting the base of the flap or excising a Burow's triangle. The pitfall here is the risk of cutting into the base of the flap, which can reduce the circulation of the flap. This added tension in the flap can lead to tip necrosis.

One simple solution for a rotation flap that seems to be under excessive tension is to simply advance the leading edge of the flap into the defect, which results in a new defect at the opposite end of the flap. Since this donor defect should be well vascularized (the flap should be taken from an area of healthy tissue), the donor site can be skin-grafted rather than closed primarily. This strategy, while not elegant, is mush less likely to result in excessive tension on the closure.

V-Y Advancement Flap

V-Y advancement flaps are extremely useful, but their execution differs greatly from transposition flaps. V-Y advancement flaps are not elevated completely, but rather must remain connected to the subcutaneous tissue underlying the flap to maintain their viability. This is because they are incised along all skin borders and so have no connected â€œbaseâ€• through which a dermal or subdermal blood supply can provide circulation.

A V-shaped flap is designed immediately adjacent to the defect with the widest portion of the V equivalent in width to the width of the defect (Fig. 3-14). The two limbs of the V are then gradually tapered so that their length is at least 1.5 times the length of the desired advancement. The distance the edge of the flap can be advanced is somewhat determined by the laxity of the local skin.

Following advancement of the flap edge into the defect, the base of the V is closed in a linear fashion, resulting in a â€œY.â€•

Postoperative Management

On completion of the operation, a bulky non-compressive dressing is applied to the wound, which may be supported with a plaster splint if immobilization is required. A small window can be left to monitor the flap for signs of ischemia or congestion. Sutures remain for an average of 10 to 14 days, depending on the status and tension of the surrounding tissue. Ideally, flaps are inset under minimal tension, allowing the patient to begin gentle range of motion exercises immediately following surgery. Once sutures are removed, the wounds may be kept moist with a petroleum-based product, which will prevent itching during the early postoperative period.

Complications

The most feared postoperative complication following local flap reconstruction is necrosis or partial necrosis of the transferred tissue. This complication occurs because of inadequate blood supply to the flap. Flap ischemia may result from flap closure under excessive tension, hematoma, or infection.

If total flap loss develops, the wound will require closure by another means, either pedicled flap or skin graft. If partial necrosis has occurred, one can debride the necrotic portion of the flap and begin dressing changes to the underlying wound bed. Once the underlying tissues are clean of any infection and necrotic debris, the surgeon can consider allowing the wound to heal through secondary intention or attempt wound closure with an alternative reconstructive method.

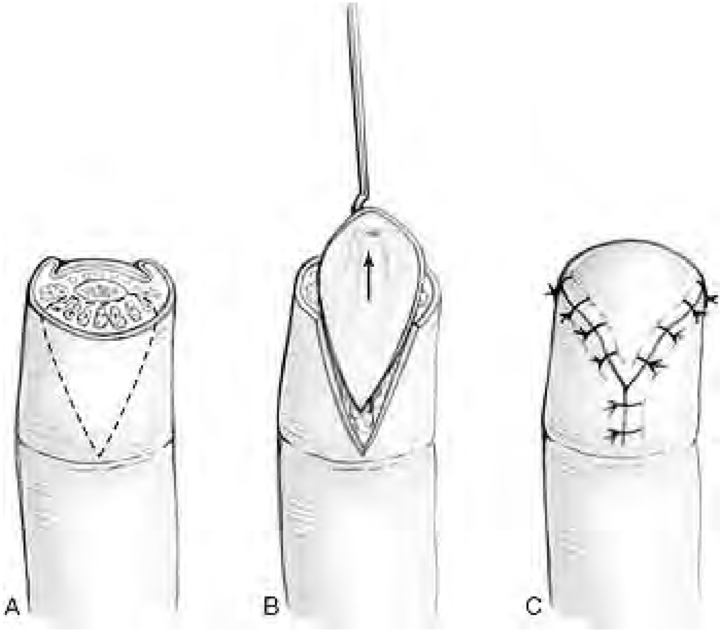

FIGURE 3-14 The Atasoy-Kleinert V-Y advancement flap. Note that the subcutaneous tissue has remained undisturbed to maintain blood supply. (Redrawn after Louis DS, Jebson PLJ, Graham TJ. Amputations. In: Green DP, Hotchkiss RN, Pederson WC, eds. *Operative hand surgery.* New York: Churchill Livingstone; 1999:48-94.)

Pearls and Pitfalls

Pearls

- Use a VAC or Reston bolster to immobilize and compress skin grafts quickly.

- Do not mesh pie-crust skin grafts on hands.
- Exploit secondary contraction of thin split-thickness grafts (0.010â€“0.012) to induce contraction of fasciotomy wounds.
- Employ Tegaderm for small donor sites (100 cm^2) to decrease pain and provide a moist wound-healing environment.
- Splint extremities to immobilize skin grafts.
- Salvage improperly planned local flaps by skin grafting the local flap donor site.

Pitfalls

- Hematoma
- Seroma
- Inadequate immobilization of grafts or flaps while healing
- Meshing grafts on the hand
- Inadequate initial wound debridement

Recommended Reading

Ablove RH, Howell RM. The physiology and technique of skin grafting. *Hand Clin.* 1997;13:163â€“173.

Chao JD, Huang JM, Weidrich TA. Local hand flaps. *J Amer Soc Surg Hand*. 2001;1:25â€“44.

Lin SJ, Hijjawi JB. Skin grafting. In: Lin SJ, Hijjawi JB, eds. *Plastic and reconstructive surgery pearls of wisdom.* New York: McGraw Hill; 2006:435â€“436.

Louis DS, Jebson PLJ, Graham TJ. Amputations. In: Green DP, Hotchkiss RN, Pederson WC, eds. *Operative hand surgery*. New York: Churchill Livingstone;

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 4

Vacuum-Assisted Closure in Extremity Trauma

Anthony J. DeFranzo

Louis C. Argenta

Indications/Contraindications

Wounds of the upper and lower extremities provide a constant challenge for orthopaedic and plastic surgeons. Complex severe injuries involving bone and soft tissue demand major reconstructive procedures. Less severe injuries exposing tendon, bone, and joints also demand innovative surgical approaches. Many patients with injured extremities have been the victim of multiple traumas and may have major associated injuries involving head, chest, and abdomen. Major blood loss, disseminated intravascular coagulopathy (DIC), high intracranial pressure, acute respiratory distress syndrome (ARDS), septicemia, and renal failure frequently complicate the treatment of these patients. Vacuum-assisted closure (VAC) has provided a way to effectively manage these wounds until definitive reconstruction can be performed. Vacuum-assisted closure has also greatly simplified reconstruction in many of these patients. Skin grafts are performed in many cases that would have otherwise required major rotational flaps or free flaps prior to the advent of VAC therapy.

The mechanics of the VAC system are straightforward. The system consists of an open cell polyurethane ether foam sponge sealed by an adhesive drape. All pores in the sponge communicate so that negative pressure applied to the sponge is applied equally and completely to the entire wound surface. The effects of the VAC on the wound are multiple. The application of negative pressure causes the sponge to collapse toward its center. Traction forces are thus applied to the wound perimeter pulling the wound edges together progressively making the wound smaller. The VAC sponge should be cut to fit inside the wound to maximize traction forces on the wound edges. The sponge should not overlap intact skin, as skin maceration may occur. In addition, the VAC removes wound edema, appears to increase circulation and decrease bacterial counts, and significantly increases the rate of granulation tissue formation (1,6).

All wounds are thoroughly debrided prior to VAC placement. No clinical infection, purulence, or suspected osteomyelitis should be present prior to VAC placement. Grossly contaminated wounds are a contraindication for VAC placement. The VAC sponge should not be applied directly over major exposed blood vessels in the extremity status posttrauma, especially if the vessel wall has been damaged or a vessel repair has been performed. A protective interface such as Adapticw may be used. Exposed, repaired, or damaged major vessels should be covered with flaps such as muscle flaps or fasciocutaneous flaps. The VAC may be applied over intact nerves if muscle flap or adequate soft tissue coverage is not possible. A layer of Adaptic may help prevent pain caused by VAC traction especially when VAC sponges are changed. Nerve repairs should also be protected by an interface layer such as Adaptic placed under the VAC sponge. Adequate soft tissue or flap coverage of major peripheral nerves is preferred for final long-term management. The VAC may frequently allow secondary closure of wounds over major peripheral vessels or nerves by pulling together adequate soft tissue as edema is removed. Split-thickness skin grafts placed directly over major exposed vessels or nerves is possible with VAC therapy after granulation tissue occurs, but not preferred.

The VAC sponge may be placed directly on bone if the bone surface bleeds following sharp debridement by an osteotome or low-speed power burr. Occasionally, desiccation of bone may occur with the VAC. If desiccation occurs, further debridement to punctuate bone surface bleeding with the immediate placement of Integraw Dermal Regeneration Template and then VAC placement directly over Integra may be tried. This technique has provided Integra take over bone allowing subsequent skin graft coverage. The Integra must be meshed or holes cut in the outer silicone layer before VAC sponge placement.

Relatively small areas of bone have been successfully treated in this manner. Large areas of exposed bone are better treated with pedicle muscle flaps, pedicle fasciocutaneous flaps, or free flaps. The VAC should not be placed over desiccated bone with questionable viability without punctuate bone surface bleeding.

Preoperative Planning

Wounds should be thoroughly debrided before VAC application. Once the wounds are clean, VAC therapy may be initiated. VAC sponges can initially be applied in the operating room following definitive debridement.

Surgery

To illustrate the application of the VAC device to wounds of the extremities, case studies will be shown to illustrate degloving injuries, gunshot wounds, and a variety of avulsion injuries with exposed tendons, bones, and joints. VAC treatment of traumatic wound complications such as infection and hematoma is also illustrated.

Degloving Injuries

Several degloving injuries to the hand have been treated by the VAC with varying results from 0% to 95% take of replaced degloved skin. The degree of take depends on the condition of the degloved skin and the viability of underlying tissue. The degloved skin is frequently crushed and lacerated. It may be difficult to clinically assess the viability of the degloved avascular skin. Also, roller injuries which frequently cause degloving injuries of the hand may crush as well as deglove the hand. Contusion to the hand intrinsic muscle and crush injury of blood vessels in the hand may also be difficult to evaluate. Progressive loss of viability may occur.

In all hand degloving injuries (N = 6) treated to date at Wake Forest University School of Medicine, fingers were not viable significantly beyond the proximal interphalangeal (PIP) joints and amputations were made at or just distal to that level (4). The thumb has been salvaged at full length in one case. In degloving injuries, most extensor and flexor tendons remain on the hand. If the distal phalanx is avulsed with the skin envelope the flexor digitorum profundus tendon(s) may also be avulsed from the hand and even from the forearm. Extensor peritenon and flexor tendon sheaths are frequently intact, providing vascularized coverage to the tendon.

The clinical approach has been to assess viability of the degloved skin and hand after meticulous debridement and completing the amputations at the appropriate level. The skin is thoroughly defatted, pie-crusted, and placed back on the hand with a few staples and/or sutures. The VAC is then immediately applied. The sponge is cut to fit over the grafted, degloved skin, using one or multiple contiguous pieces of sponge. A â€œhand VACâ€• kit is also commercially available. Controlled suction of 125 mm Hg pressure is applied (Fig. 4-1). This procedure takes very little operative time. If successful, the patient is spared a great deal of overall treatment time and morbidity compared to more complicated methods of reconstruction. Therefore, with a degloving injury, our treatment algorithm is as follows: (a) skin defatting and debridement, (b) wound debrideded, (c) skin applied to wound, (d) VAC applied to wound. The VAC is left in place for 4 to 6 days. If there are areas which have not revascularized, the VAC may be replaced for 2 more days for a total of 6 days. After day 6, further take with VAC therapy does not seem to occur and we have not used the VAC longer than 6 days.

If VAC therapy is totally unsuccessful, no more than 6 days are lost with the attempt at degloved skin replacement before other methods of reconstruction are initiated. The best case to date of VAC replacement of degloved skin achieved a 95% take of avulsed skin with preservation of fingers at or just distal to the PIP joints and preservation of the thumb at full length. Excellent range of motion of the metacarpal phalangeal joints was achieved. Further surgery involved outpatient revision of the amputations of the four amputated fingertips and deepening of the first web space (Fig. 4-2).

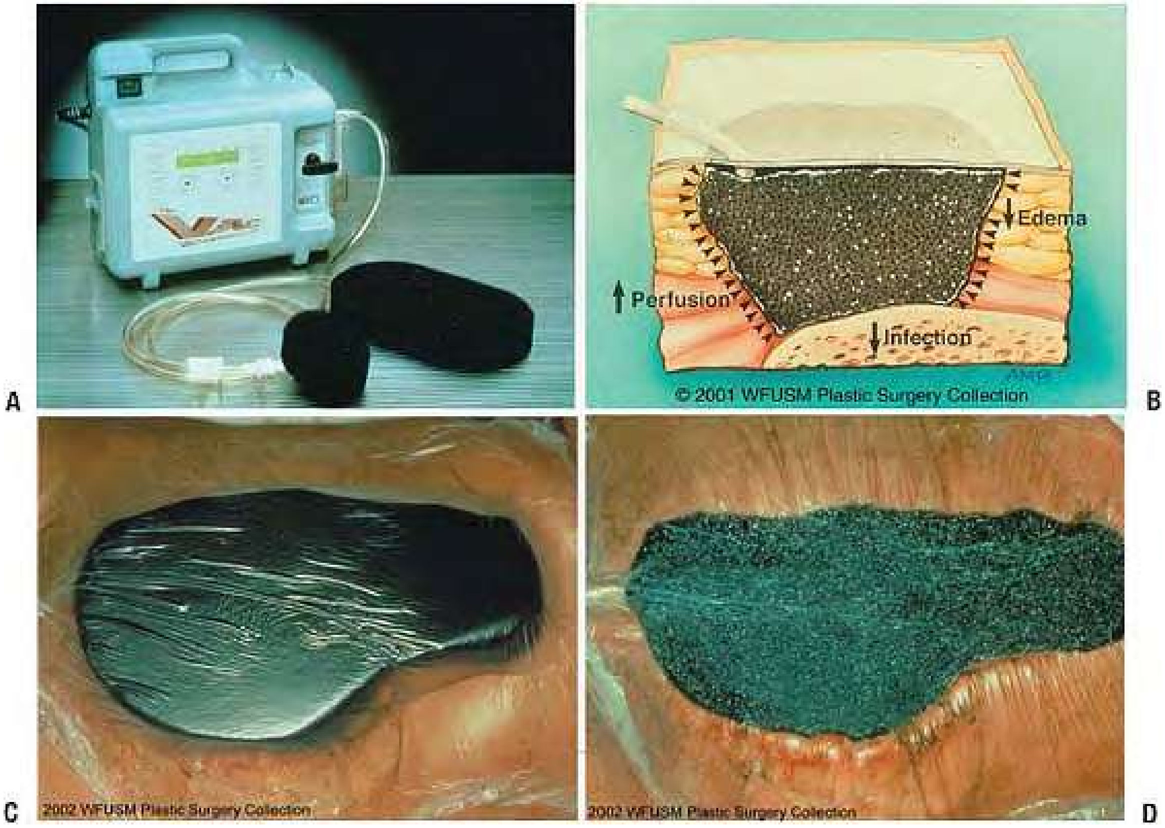
A
B
C
D
VAC
Edema
Perfusion
Infection
© 2001 WFUSM Plastic Surgery Collection
2002 WFUSM Plastic Surgery Collection
2002 WFUSM Plastic Surgery Collection

FIGURE 4-1 Mechanics of the VAC system. A: Open cell polyurethane ether foam dressing. B: A 400â€“600 micron pore diameter. C: Foam dressing cut to fit the wound. D: Wound sealed with adhesive drape. There is 50â€“125 mm Hg controlled suction applied continuously.

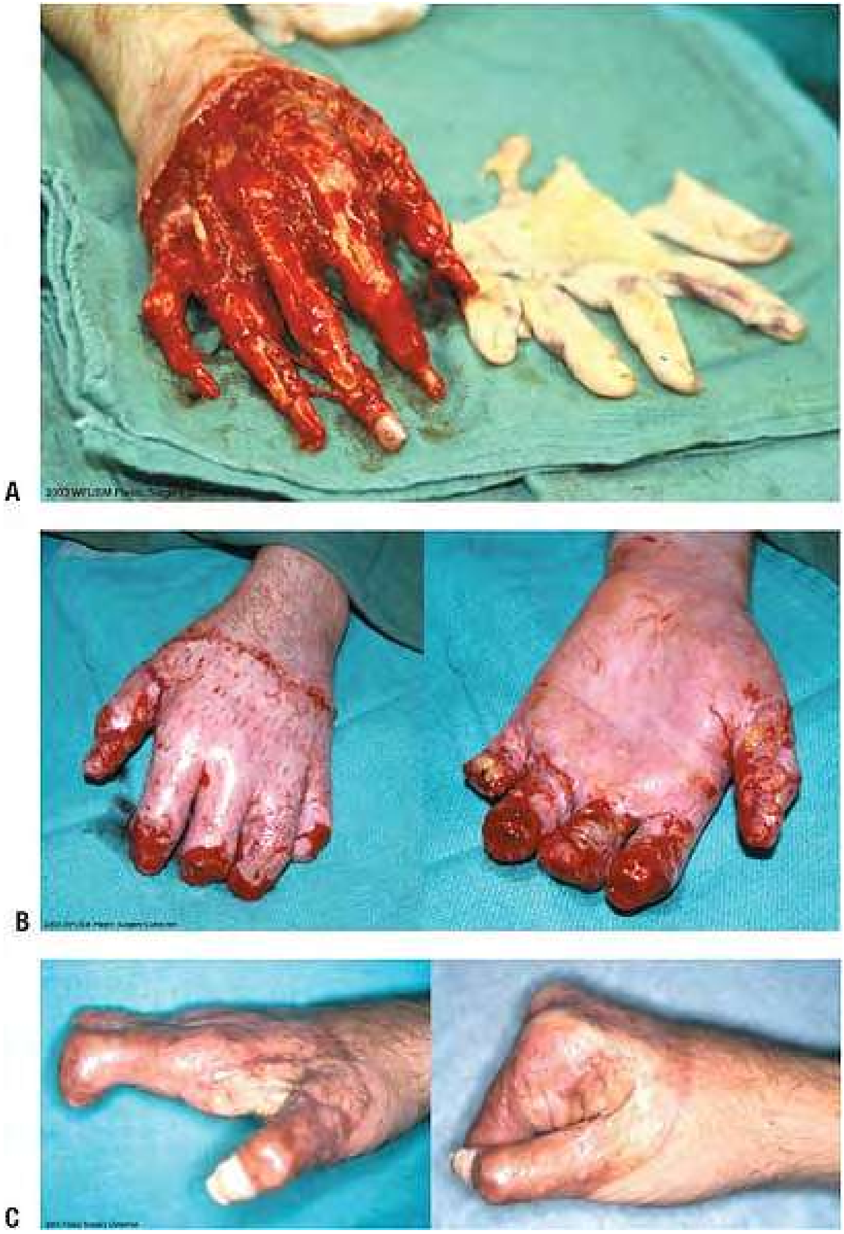

FIGURE 4-2 Degloving injury to the hand. A: Entire dorsum and all fingers avulsed to distal palm. Neurovascular bundles present to or just beyond proximal interphalangeal joints. Extensor paratenon intact; flexor tendon sheaths intact. B: Fingers surgically amputated at or just distal to PIP joints. Skin defatted, pie-crusted, and reapplied. VAC change day 4 with 95% take of reapplied degloved full-thickness skin. C: Final result with good range of motion at metacarpal phalangeal joints.

One degloving injury of the foot has been treated with the VAC. The toes were amputated with the degloved skin at the time of the injury. Metatarsals were shortened surgically so that they were covered by viable soft tissue. The skin was thoroughly defatted, pie-crusted, and returned to the foot. A few staples were used to secure the skin, and the VAC was applied for 4 days. Skin viability was 95%. No revisional surgery has been required. A small area of breakdown on the plantar surface of the heel occurred with ambulation, but healed with simple dressing changes and has not been a recurring problem with proper footwear (Fig. 4-3).

Crush Avulsion Injuries with Exposed Bone, Joints, and/or Tendons

Crush/avulsion injuries are commonly seen after industrial accidents or accidents at home with lawnmowers and farm machinery. Crush/avulsion injuries are also frequently seen with motor vehicle accidents and may be associated with other major critical injuries. Expedient closure of wounds without major reconstructive procedures with long anesthesia times may be beneficial to overall patient

care in the critically injured patient. Injuries to the upper extremity and lower extremity from the knee to the ankle result in exposure of tendons, bones, and joints. The VAC has been used on such patients only after one or more thorough operative debridements of all nonviable tissue. In some of these cases, coverage of tendon, bones, and joints has been facilitated by the use of Integraw.

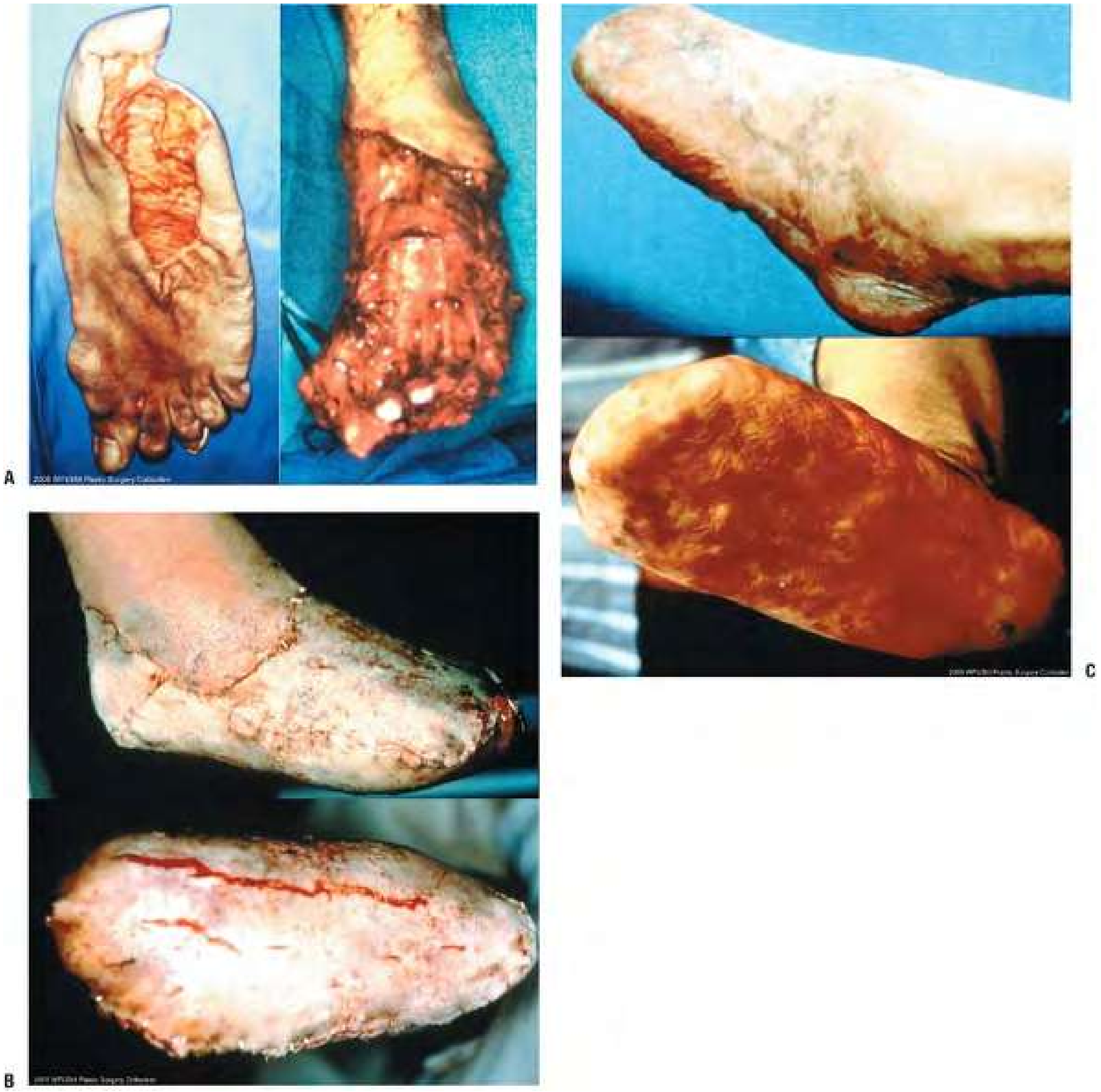

FIGURE 4-3 Degloving injury to the foot. A: Entire dorsum, plantar surface, and skin of posterior heel avulsed. All toes traumatically amputated at metatarsal phalangeal joints. B: Amputations revised at metatarsal phalangeal joints. Avulsed skin defatted, pie-crusted, and replaced. The VAC is removed on day 4 with 95% of replaced skin viable. C: Final result with durable cover with proper footwear at 2 years.

Integra Dermal Regeneration Template is a clear bilayer membrane designed originally for skin replacement in burn patients. It consists of an epidermal layer comprised of a clear thin sheet of silicone (polysiloxane) and a dermal layer comprised of a porous matrix of bovine collagen fibers cross linked with chondroctin-6-sulfate from shark cartilage. The silicone sheet is removed after take of the Integra, and a split-thickness skin graft typically 8 to 12 thousandth of an inch is applied. With VAC therapy over pie-crusted or meshed Integra (noncrushing mesher), take occurs in as little as 6 days. Our practice is to maintain VAC therapy for 6 days undisturbed over Integra. Clinically, the Integra assumes a salmon-pink color when take occurs; but with 6 days of VAC therapy, a more robust red color may be achieved. We now routinely plan to perform split-thickness skin grafts 6 days after VAC placement over Integra.

Integra and split-thickness skin graft provide a pliable durable bilaminar skin reconstruction ideal to cover joint surfaces (3). With VAC therapy, Integra has taken well over a vascularized wound bed, at 6 days in most patients. In addition to simple Integra take, however, viable tendons, bones, and open joints without cover have been bridged for a distance of 1 to 2 cm. At day 6, if structures have not been â€œbridgedâ€• sufficiently, VAC therapy may be continued for a total of 10 to 12 days. Red vascular tissue can be seen spreading transversely through the Integra that is bridging a bone, joint, or tendon.

Tendons and bones can granulate in the abscence of Integra to allow split-thickness skin graft take (2). However, the use of Integra appears to provide a more durable skin cover over bone and over tendon allows tendon glide with appropriate early occupational therapy. Split-thickness skin graft applied over Integra is followed by VAC therapy for another 6 days to achieve take of the split-thickness skin graft to complete the reconstruction. The VAC has been useful in maximizing the take of split-thickness skin grafts over irregular surfaces in many of our cases involving the extremities (7). Whenever possible, range of motion of joints and tendons is begun early at 1 to 2 weeks after successful skin graft take (Figs. 4-4,4-5,4-6,4-7).

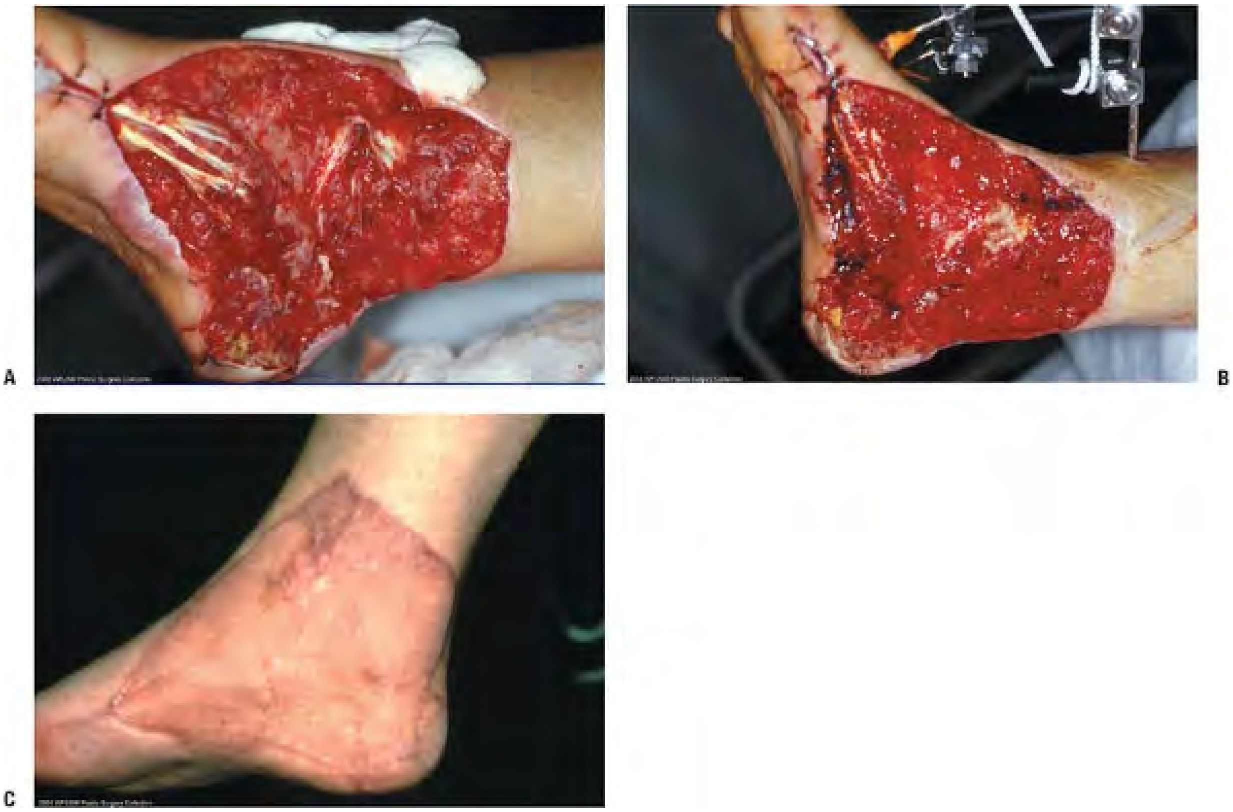

FIGURE 4-4 A 68-year-old status post lawnmower avulsion injury. Smoker with atherosclerosis. A: Exposed medial malleolus, ankle joint, and tendon. B: VAC therapy for 2 weeks with granulation tissue covering exposed tendons and joint. No use of Integra^{W}. C: Split-thickness skin graft stable for 4 years with normal footwear.

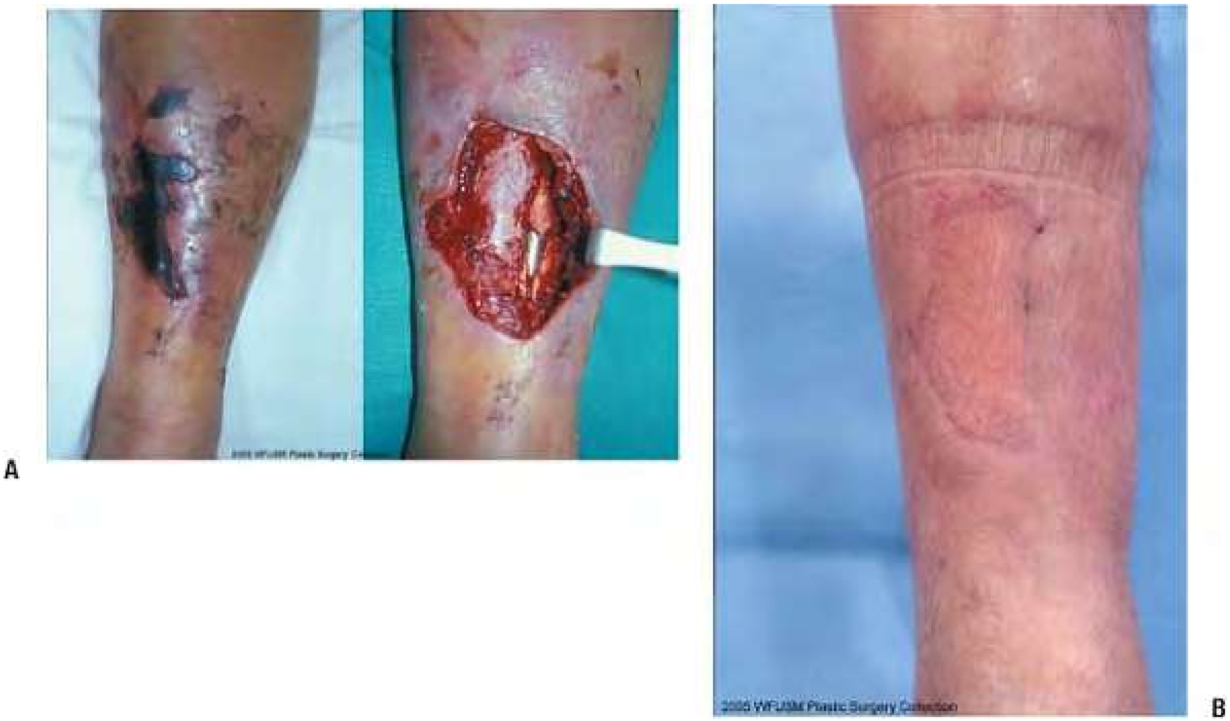
A
2005 WFUSM Plastic Surgery Collection
2005 WFUSM Plastic Surgery Collection
B

FIGURE 4-5 A 67-year-old status post motor vehicle accident with crush of lower extremity and fracture of underlying tibia treated with an intramedullary rod. A: Exposure of rod post status debridement. B: VAC therapy allowed soft tissue to cover bone by both a decrease in wound size and the formation of granulation tissue. Final skin closure was obtained with a split thickness skin graft. Wound stable at 1 year.

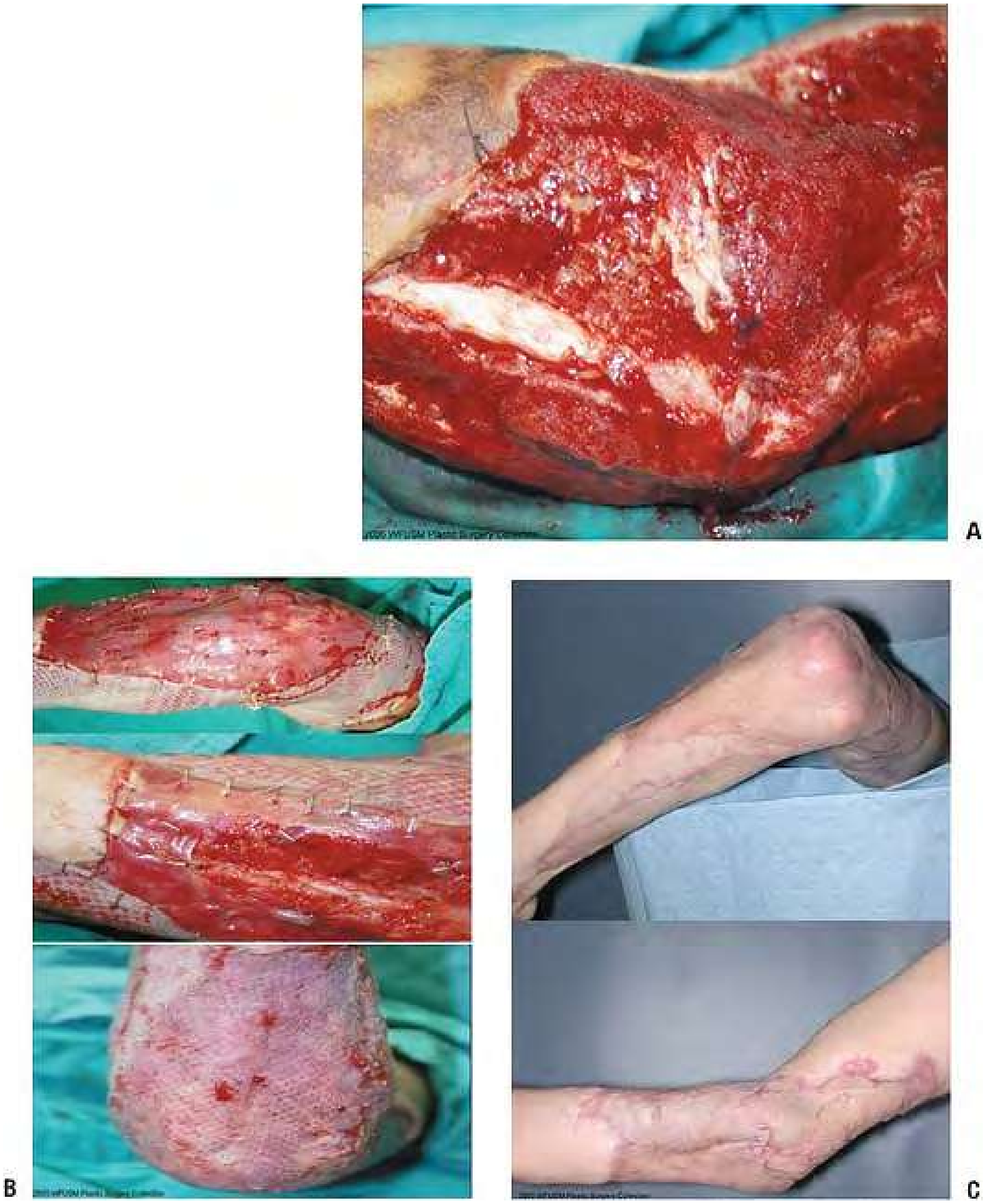

FIGURE 4-6 A 17-year-old with skin avulsion of the left arm status post motor vehicle accident. A: Exposed ulna is debrided to bleeding bone. B: VAC applied for 6 days with formation of granulation tissue. Integraw was then applied and VAC continued for 7 days. Finally split-thickness skin graft was applied. C: Stable cover after 3 years for definitive wound closure.

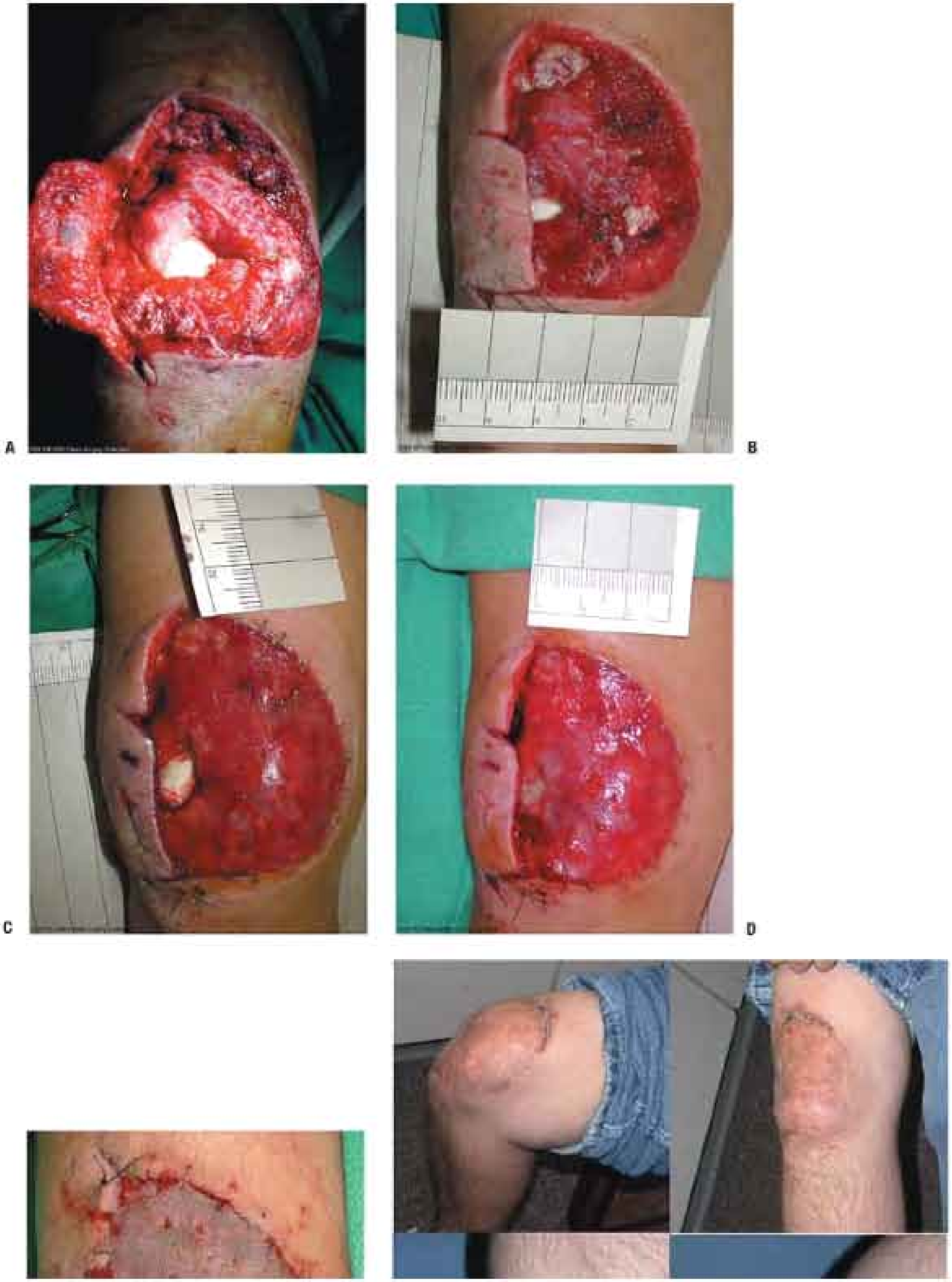
A
B
C
D

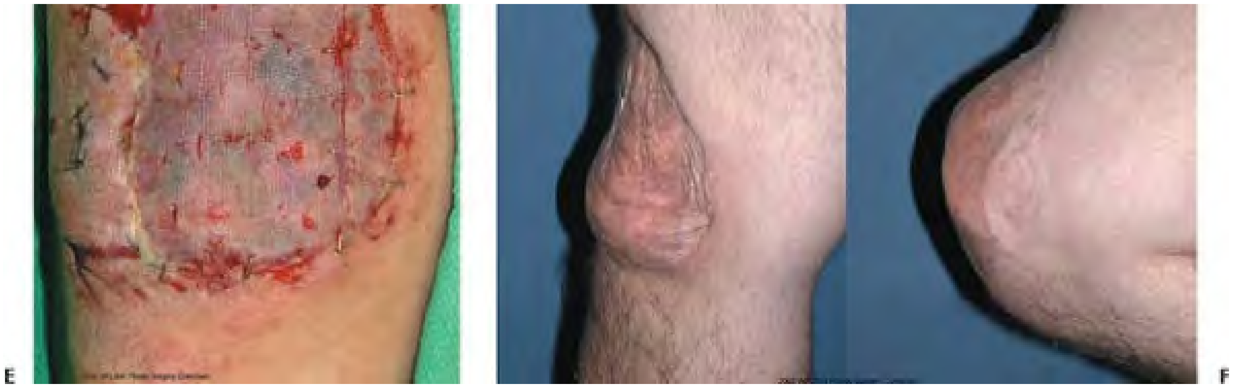

FIGURE 4-7 A 24-year-old male status post motorcycle accident with right brachial plexus injury, right clavicle and right scapula fracture, right occluded axillary artery, right pneumothorax, right open comminuted patella fracture with avulsion of patellar bone and soft tissue with an anterior open knee joint. A: Open knee joint post status debridement of nonviable patella and soft tissue. B: Soft tissue deficit following ligament reconstruction and VAC placement.C: Integraw was placed over the open knee joint. D: Granulation bridging knee joint through Integra. E: A split-thickness skin graft was applied over the Integraw after the Integra showed signs of levascularization. F: Excellent range of motion achieved at 3 months with stable cover at 2.5 years.

Gunshot Wounds

Gunshot wounds are associated with blast effect which causes major damage to bone and surrounding soft tissue. Massive wound edema develops. Reconstruction can become exceedingly difficult. When applied immediately after proper debridement, VAC therapy can remove a large amount of edemic fluid from the wound. Multiple operative debridements may be required to ensure that all nonviable bone and muscle have been removed. As much as 2 L of edema fluid have been removed from wounds of the lower extremity over a 24 hour period (5). Application of VAC can also pull retracted wound edges together and promoted the remaining wound to fill with granulation tissue covering exposed bone, tendon, and hardware (Figs. 4-8 and 4-9).

Vacuum-Assisted Closure Therapy for Posttraumatic

Wound Complications

Posttraumatic wounds are subject to higher hematoma and infection rates due to tissue devascularization and tissue contamination. If a wound dehisces, rapid edema may make wound reclosure impossible. Bone, tendon, and/or plates become exposed and their coverage is essential to a successful outcome. Vacuum-assisted closure therapy has been useful in regaining closure of posttraumatic open wound complications (Fig. 4-10).

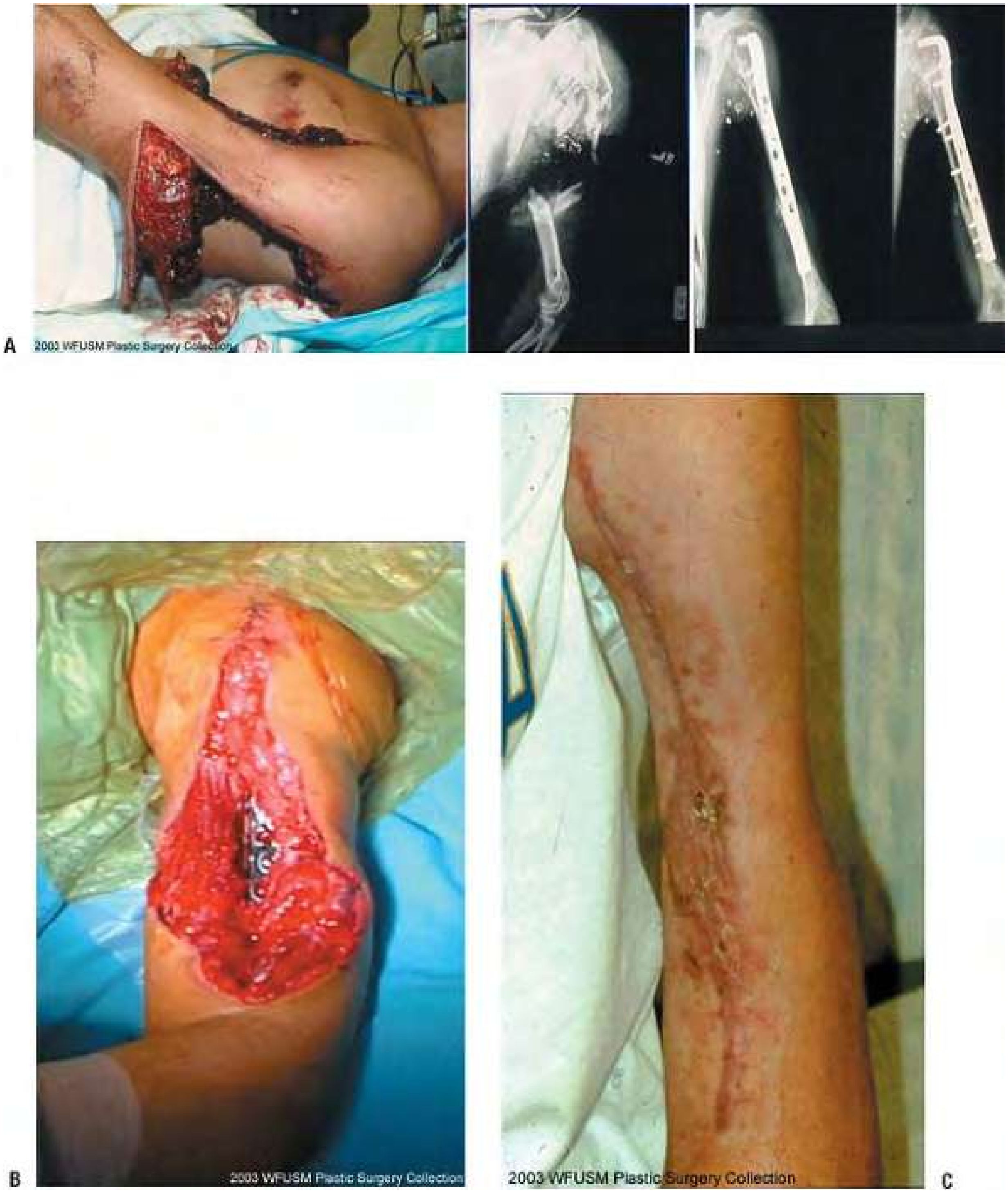
A
2003 WFUSM Plastic Surgery Collection
B
2003 WFUSM Plastic Surgery Collection
2003 WFUSM Plastic Surgery Collection
C

FIGURE 4-8 A 25-year-old with a self inflected gunshot wound to the left arm. A: Debridement and plate fixation of the humerus. B: Exposed plate status post debridement with decreased wound size and granulation tissue covering plate

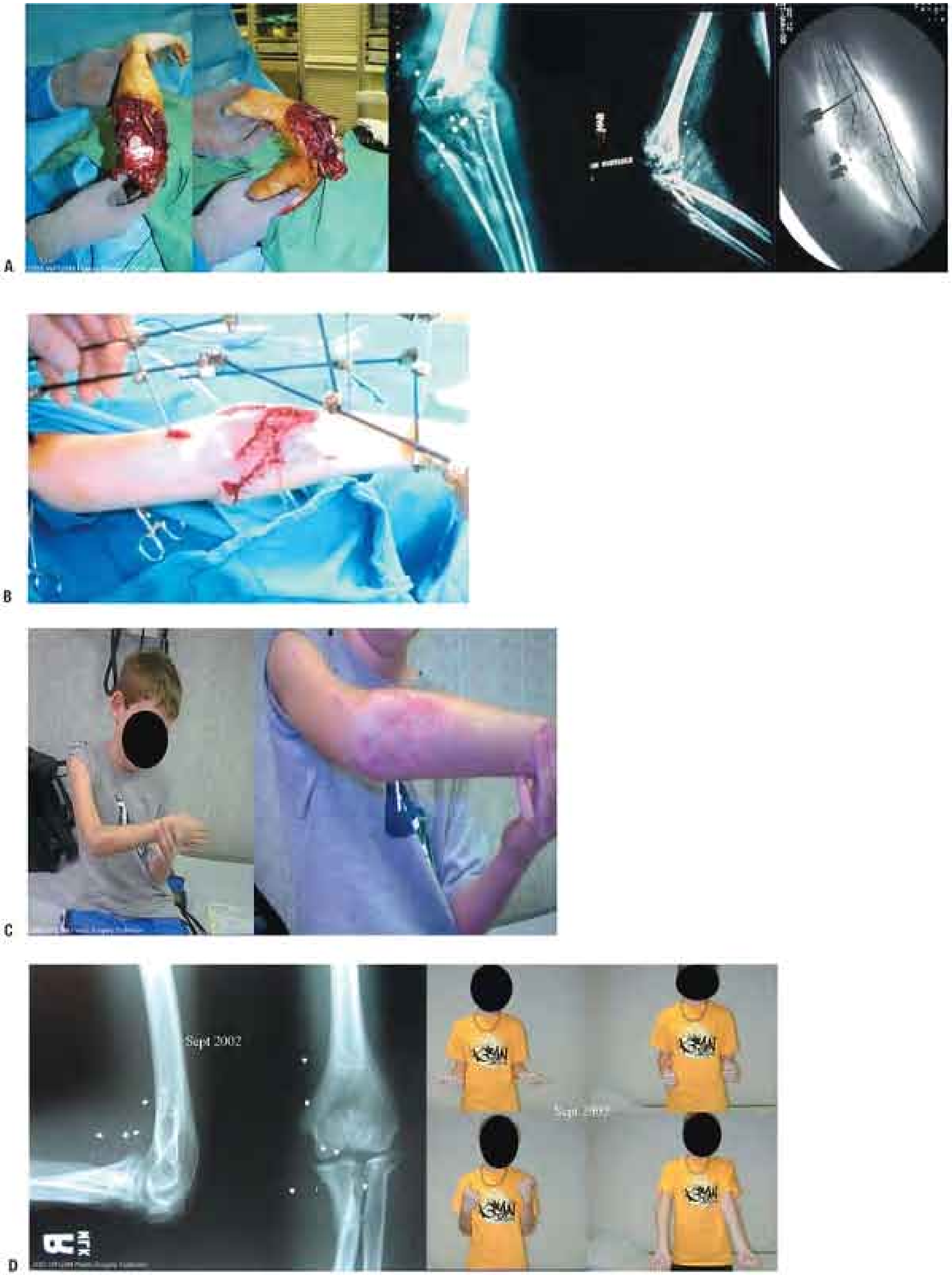
A
B
C
Sept 2002
Sept 2002
D

FIGURE 4-9 A: A 10-year-old boy with a gunshot wound to the right elbow, forearm, and hand extremity well vascularized. VAC placed on day 4 status post debridement of all nonviable tissue. Three operative debridements were required prior to VAC placement. B: Early post-operative course showing a small amount of exposed bone. The soft tissue defect was advanced to closure by VAC therapy. C: Split-thickness skin graft placed 10 days status post injury after granulation tissue covers all exposed bone. D: Wound stable 2 years. Excellent bone healing and range of motion.

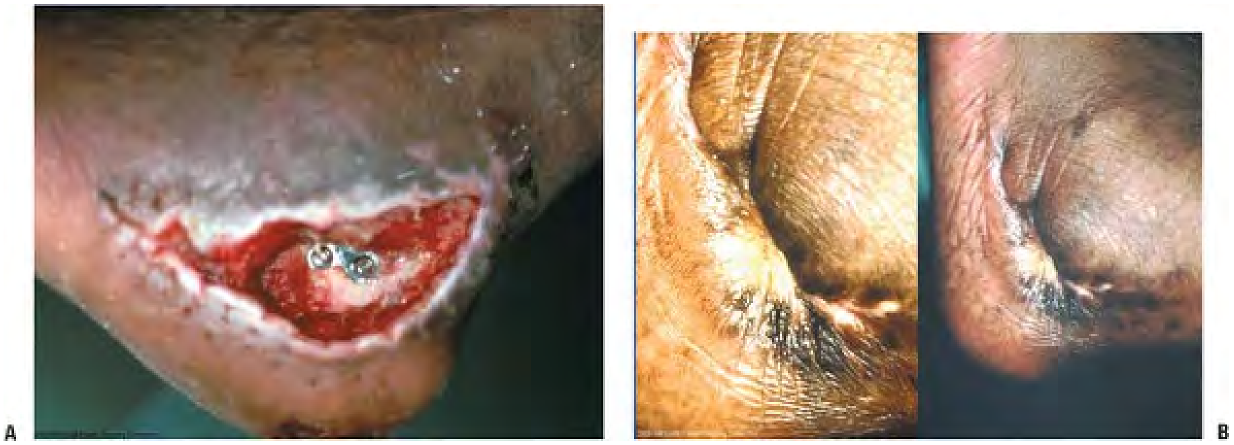

FIGURE 4-10 A 32-year-old status post fractured calcaneus treated by open reduction and plate fixation. A: A post-operative hematoma resulted in wound dehiscence and exposed the plate. B: Four weeks of VAC therapy were used to obtain wound closure (3.5 weeks as an outpatient). No surgical procedure to achieve skin closure was required. Wounds stable for 3 years with normal foot wear.

Postoperative Management

We believe that the VAC sponge ideally should not be left longer than 2 days between changes to maintain proper seal and to avoid infection. Small, clean outpatient wounds may be changed three times a week (Monday, Wednesday, Friday schedule), by visiting home nurses with a 2â€“3 day change interval.

Oral or intravenous narcotics are usually sufficient for VAC changes in adults. Infusion of Xylocaine into the sponge discontinued from suction 15 minutes before VAC change has been beneficial for local analgesia prior to VAC change. Care must be taken to administer doses of Xylocaine usually well below safe standard dose guidelines. A dose of 10 to 20 cc of 1% Xylocaine with epinephrine for a small to medium size sponge would be sufficient. Much larger wounds may not benefit from Xylocaine infusion. Children granulate more quickly than adults and may require a smaller pore size sponge (white sponge) than the standard black sponge. Granulation tissue may grow into the black sponge and cause increased pain or bleeding with the VAC sponge change in children. Every-24-hour VAC changes may occasionally be required with a standard black sponge in children. However, with granulation tissue development, split-thickness skin grafting is possible even sooner in children, and the uncomfortable course of VAC sponge changes is soon over. Children may require a brief general anesthetic or Ketamine for each VAC change in a dressing room setting staffed by an anesthesiologist.

Complications

Complications are experienced occasionally due to mechanical failure of the machine with loss of suction and unfortunate wound deterioration. The wound may become a sealed abscess cavity if suction is not properly maintained. All new machines have alarms that should signal loss of suction. Every-2-day VAC sponge changes ensures that proper machine function and seal are provided. Technical errors such as an improper seal have occurred with inexperienced personnel. In-service training is mandatory for all nursing and physician staff. Sponge overlap of the skin will cause skin maceration. Bleeding may occur with VAC change, especially in children. Bleeding can usually be controlled with pressure. Generalized bleeding from the wound usually means that sufficient granulation has occurred to allow skin grafting.

Osteomyelitis has occurred infrequently if insufficient bone debridement was performed prior to VAC therapy. Granulation tissue unfortunately may quickly cover a bony sequestrum with VAC therapy. Significant drainage from such wounds soon occurs and points to the need for further debridement of the nonviable bony sequestrum. All questionable bone should be aggressively debrided prior to VAC placement so that all remaining bone bleeds well.

A small number of plates have become infected after VAC coverage with granulation tissue. Plate infection jeopardizes bony union and delays skin closure or skin grafting. Loose plates and screws or plates that do not conform tightly to underlying bone are likely to become infected. Relatively small plate exposures have been successfully treated with the VAC. Major plate exposure should not be treated with the VAC. Clinical judgment should dictate that a major plate exposure, especially with poor-quality surrounding tissue, must be treated with some form of flap.

Results

Vacuum-assisted closure has decreased the requirements for complex wound closure in the extremities. Fewer major pedicle flaps and fewer free tissue transfers are required. With vacuum-assisted closure of traumatic extremity wounds, the number of free tissue transfers now required at our institution is approximately 33% when compared to the number prior to VAC therapy. Long operative and anesthesia times are avoided, which is especially beneficial for critically severely injured patients.

References

1. Argenta LC, Morykwas MJ. Vacuum-assisted closure: A new method for wound control and treatment: Clinical experience. *Ann Plast Surg.* 1997;38(6):563â€“577.

2. David L, DeFranzo A, Argenta L, et al. The use of vacuum-assisted closure therapy for the treatment of lower extremity wounds with exposed bone. *Plast Reconstr Surg.* 2001;108(5):1184â€“1191.

3. DeFranzo AJ, Argenta LC, David LR, et al. Treatment of traumatic hand/upper extremity wounds with the VAC wound healing device. *Plast Surg Forum.* 2002;264â€“265.

4. DeFranzo AJ, Marks MW, Argenta LC, et al. Vacuum-assisted closure of

Originally, peripheral nerve injuries were described by Sir Herbert Seddon in 1943 as neurapraxia, axonotmesis, and neurotmesis (1). The classification was later expanded by Sunderland and further defined by Mackinnon to include six degrees of injuries (Table 5-2) (2). First degree (neurapraxia) and second degree (axonotmesis) injuries recover spontaneously, the latter at the classic rate of 1 in/month or 1 to 1.5 mm/day (3). Third degree injuries must regenerate through some amount of scar tissue, thus recovery is variable and less than normal depending on the amount of scar tissue around the nerve. In fourth degree injuries, also known as a neuroma-in-continuity, regeneration is blocked by scar tissue within the fascicles and recovery is usually very poor or does not occur. A fifth degree nerve injury (neurotmesis) is a transection of the nerve, which always will require surgical repair, while the sixth degree nerve injury encompasses a variety of nerve injuries within a single nerve. The difficulty with surgical correction of a sixth degree injury is limiting the repair to the fascicles affected by fourth and fifth degree damage and not damaging the fascicles with the potential for spontaneous recovery (Fig. 5-1).

neurapraxia
Degree I

conduction block resolves
fast/excellent

spontaneously

axonotmesis
Degree II

axonal rupture without interruption of the basal lamina tubes
slow/excellent

Degree III

rupture of both axons and basal
slow/incomplete

lamina tubes, some scar

Degree IV
(neuroma-in-continuity)
complete scar block
none
neurotmesis
Degree V

complete transection
none

Degree VI

combination of I-V +/- normal

fascicles
mixed

Seddon Sunderland Mackinnon Injury Recovery

Table 5-2. Classification of nerve injury

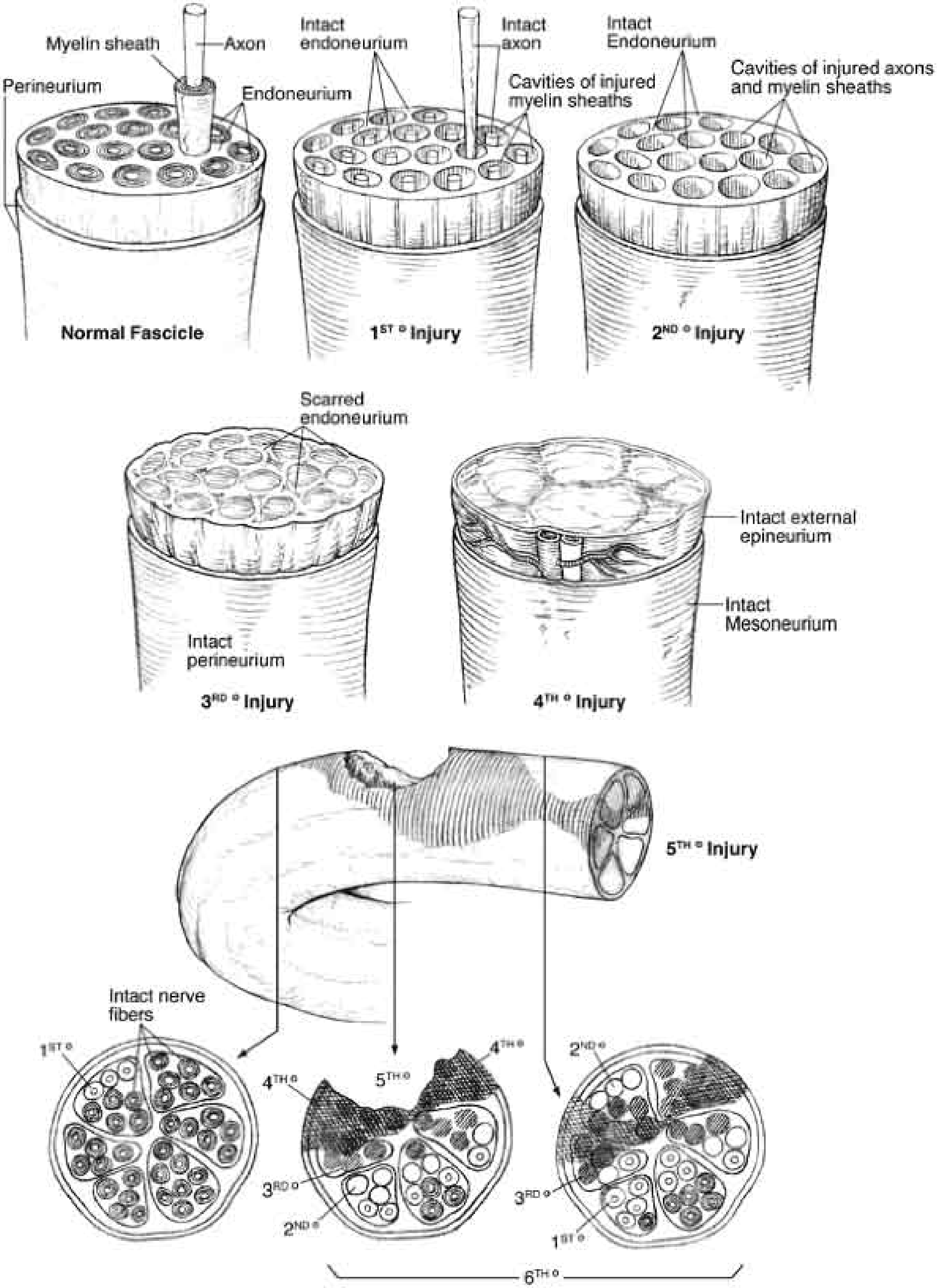
Myelin sheath
Axon
Perineurium
Endoneurium
Normal Fascicle
Intact endoneurium
Intact axon
Cavities of injured myelin sheaths
1ST ° Injury
Intact Endoneurium
Cavities of injured axons and myelin sheaths
2ND ° Injury
Scarred endoneurium
Intact perineurium
3RD ° Injury
Intact external epineurium
Intact Mesoneurium
4TH ° Injury
5TH ° Injury
Intact nerve fibers
1ST °
4TH °
5TH °
4TH °
3RD °
2ND °
2ND °
3RD °
1ST °
6TH °

FIGURE 5-1 Schematic depiction of nerve injuries from first through sixth degree. (By permission of Mayo Foundation for Medical Education and Research. All rights reserved.)

Stretch and Nerve Avulsion Injuries

Traction on a nerve can result in a mild stretch injury to the axon with preservation of the basal lamina. The Schwann cells around both myelinated and unmyelinated fibers may be locally injured and, until the cells are replaced from surrounding Schwann cells, will cause a conduction block. Such injuries are classified as first degree injuries and recover completely. As the pulling forces increase across the nerve, the basal lamina, the endoneurium, and perineurium become injured. As the nerve heals, internal scaring of the nerve may occur in some of the fascicles, leading to incomplete recovery. Grossly, the nerve will appear intact as the epineurium is usually not violated. Occasionally, neurolysis or even an internal neurolysis of the injured nerve may improve recovery as long as there is evidence that the nerve recovery is being hindered by scar surrounding the nerve, rather than internal scar to the fascicles.

Nerves that are stretched beyond the breaking point will avulse, resulting in gross disruption of the epineurium, perineurium, and endoneurium. A neuroma will typically form at the proximal end, and these injuries are treated with excision of the neuroma and primary repair when possible. Often, if the time from injury is greater than 2 weeks, even excessive mobilization of the nerve ends will not be enough to overcome the resulting gap that occurs when the unhealthy neuroma tissue at both ends is resected. A nerve graft is then needed to repair the gap.

In extreme cases, nerves can be avulsed from their insertion into the spinal cord. In the past, these injuries were treated conservatively. Once no further nerve recovery was noted, tendon transfers were used to restore the residual functional deficits (4). More recently, the introduction of nerve transfers allows for rewiring of nonfunctioning nerves by using local uninjured nerves to restore electrical continuity to the deinnervated muscles (5). Tendon transfers may be used to augment the nerve transfers once maximum recovery is achieved (6).

Nerves avulsed at the neuromuscular junction present a different problem. Nerves that are injured just prior to entering the muscle, or shortly thereafter, may still be repaired or grafted in most cases as long as a large enough nerve stump is found. Motor nerves that are avulsed from the muscle bellies are treated by implanting proximal nerve, when available, directly into the muscle with the hope that some of the fibers will find a neuromuscular junction and reinnervate at least part of the muscle. Some studies (7) show as good as M4 motor recovery 1 to 2 years after direct nerve to muscle neurotization; however, experimental studies do not support these findings. Rather, recovery is much less than a nerve coaptation would produce (8).

Crush Injuries

Crush injuries comprise the most common peripheral nerve injury to the extremity. External compression may be complicated by increased internal pressure from hematomas, fractures, and local tissue edema. When minor, this may cause a temporary neurapraxia, but with greater compression the likelihood of permanent injury increases. The most severe consequence of a crush injury is the progression to compartment syndrome. Often an early sign of impending compartment syndrome is a decrease in vibration sensibility (9). Compartment syndrome of the upper extremity and lower extremity are surgical emergencies and are reviewed separately within this text.

Nerve compression injuries may also develop distal to the actual soft tissue trauma. Local edema and inflammation after injury can exacerbate a preexisting condition, such as a mild carpal tunnel turning into an acute event after a distal radius fracture. Occasionally, on a case by case basis, surgical decompression is necessary, even though the majority will resolve spontaneously. Likewise, an anterior cruciate ligament tear of the knee requiring reconstruction or repair may precipitate an acute foot drop postoperatively despite the deep peroneal nerve being usually uninjured. Some surgeons postulated that intraoperative positioning of the limb may be a factor in the late development of palsy (10); however, we believe that local tissue inflammation probably exacerbates a preexisting condition that manifests itself as a postoperative foot drop. When conservative nonoperative measures do not lead to sufficient improvement in nerve function after 2 to 3 months, decompression of the peroneal nerve should be considered (11). Figure 5-

2 shows an algorithm for treatment of closed nerve injuries.

Penetrating Injuries

Blunt penetrating trauma is usually more locally destructive than a sharp injury. Often nearby structures such as blood vessels and tendons are injured in addition to the nerve. However, the size of a sharp laceration, such as with glass or a knife, can mislead the surgeon into underestimating the extent of injury. A seemingly small skin laceration may in fact extend under the surface and result in

a bigger injury than would be expected from the external size of the wound. Exploration is imperative if a nerve palsy is present, as the likelihood that the nerve is partially or completely transected is high. It is recommended to explore these injuries semi-electively within the first 2 weeks. The further from the time of injury, the more likely a nerve graft will be needed to overcome the resulting nerve gap. In the event of a penetrating trauma with an associated vascular injury, immediate exploration is warranted. Often the nerve injury is overlooked and not identified in the face of a more urgent vascular injury. In such cases the functional deficit may be first noticed postoperatively, when it is unclear if the nerve injury is from the inciting event, iatrogenic during the repair of the vascular injury, or secondary to edema or hematoma. While a CT scan or MRI may be helpful to evaluate for the latter, internal scarring of the nerve may not always be seen.

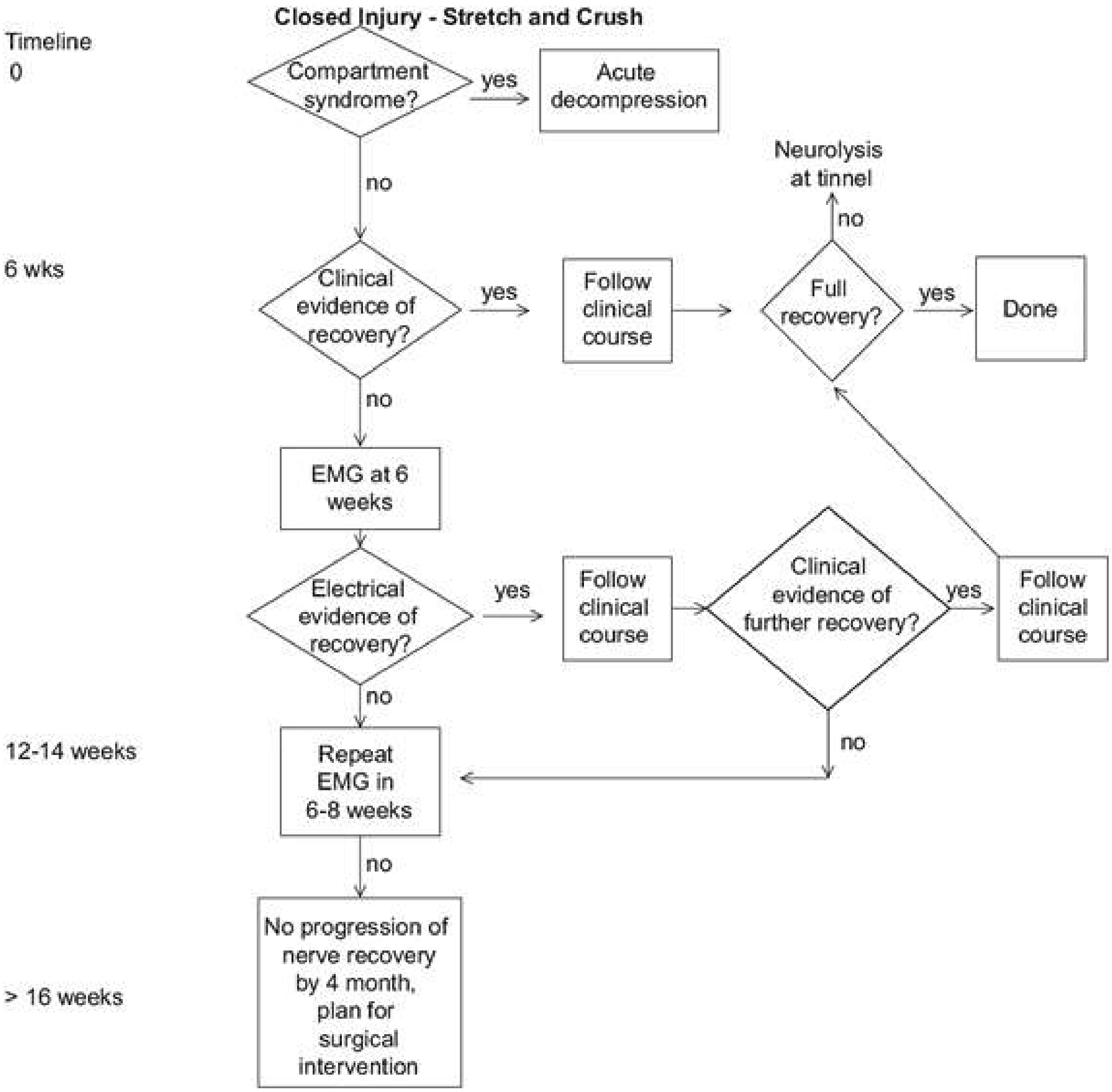
Closed Injury - Stretch and Crush
Timeline
0
Compartment syndrome?
yes
Acute decompression
no
6 wks
Clinical evidence of recovery?
yes
Follow clinical course
Full recovery?
yes
Done
no
Neurolysis at tinnel
no
EMG at 6 weeks
Electrical evidence of recovery?
yes
Follow clinical course
Clinical evidence of further recovery?
yes
Follow clinical course
no
no
12-14 weeks
Repeat EMG in 6-8 weeks
no
> 16 weeks
No progression of nerve recovery by 4 month, plan for surgical intervention

FIGURE 5-2 Algorithm for closed peripheral nerve injuries such as stretch and crush injuries.

Blunt penetrating injuries are initially treated conservatively, similar to closed crush and stretch injuries, because they are may recover spontaneously. The local tissue edema often causes a neurapraxia that resolves; however, those that do not recover after 3 months should be evaluated by electrodiagnostic studies and treated as a traction injury. Figure 5-3 shows an algorithm for managing nerve injuries and the timing of additional studies.

Two specific blunt penetrating injuries deserve special mention: gun shot wounds and electrical injuries. Gun shot injuries present a unique problem since the trajectory of the bullet is unpredictable. The type, caliber, and velocity of the bullet each play a role in tissue destruction (12). The belief that the higher velocity bullet causes more tissue destruction can lead to early massive debridement; however, current recommendations are for judicious debridement and staged exploration, as in crush injuries (13). In fact, a low velocity bullet like a slug that fragments and stays in the soft tissue will more likely lead to infection and local tissue destruction than an Army issue missile that passes through the tissue with relative ease (14).

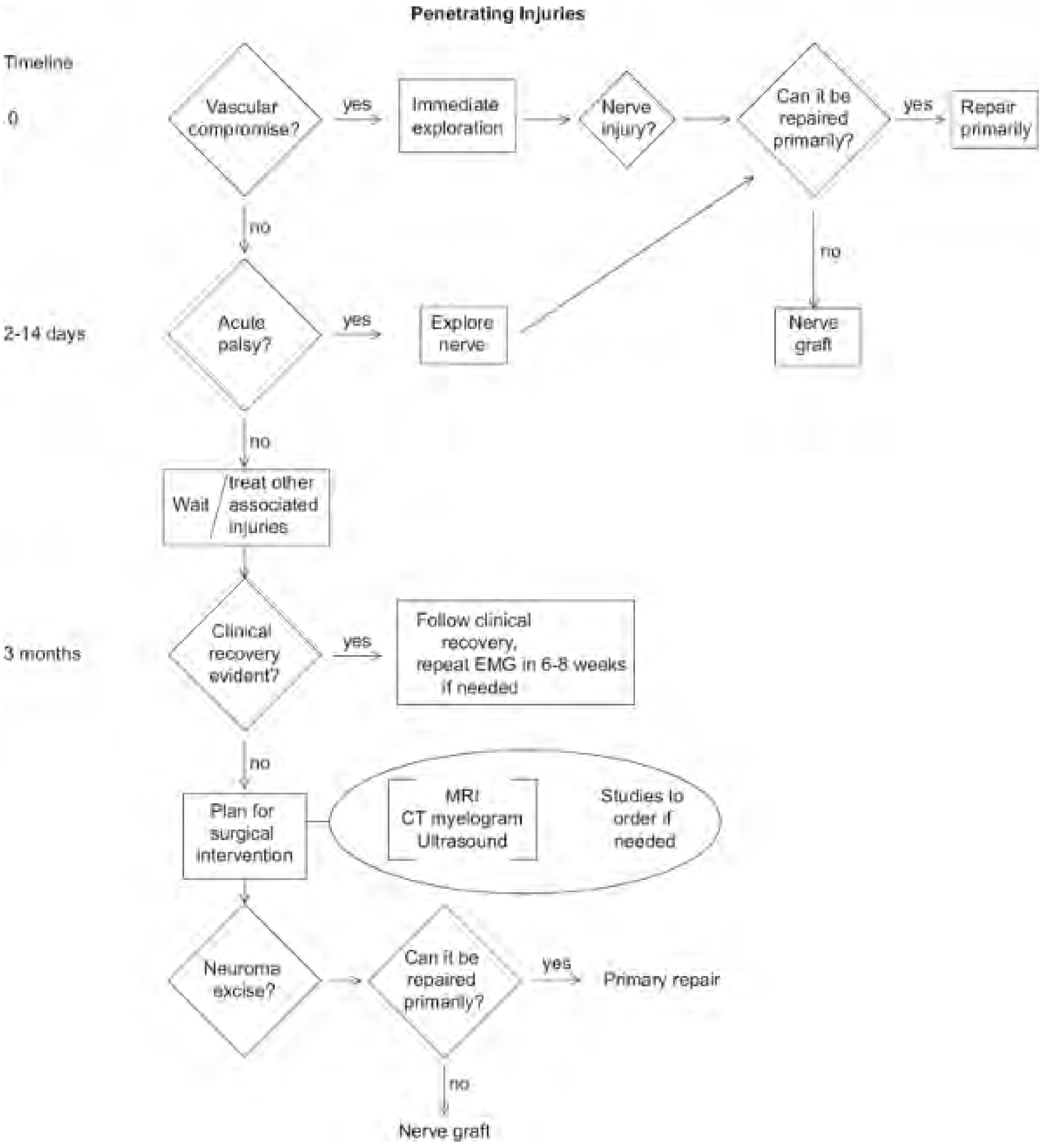
Penetrating Injuries
Timeline
0
Vascular compromise?
yes
Immediate exploration
Nerve injury?
Can it be repaired primarily?
yes
Repair primarily
no
no
2-14 days
Acute palsy?
yes
Explore nerve
Nerve graft
no
Wait / treat other associated injuries
3 months
Clinical recovery evident?
yes
Follow clinical recovery, repeat EMG in 6-8 weeks if needed
no
Plan for surgical intervention
MRI
CT myelogram
Ultrasound
Studies to order if needed
Neuroma excise?
Can it be repaired primarily?
yes
Primary repair
no
Nerve graft

FIGURE 5-3 Algorithm for penetrating injuries.

Electrical injuries that are not severe enough to cause death from heart arrhythmia and respiratory paralysis cause vigorous nerve stimulation which leads to paralysis and vasospasm. Massive muscle contractions due to nerve stimulation or the direct triggering of striated muscles can cause muscle rupture, ligamentous tears, fractures, and joint dislocations (15). In addition, electric current damages tissue from direct thermal heating; the coagulation necrosis is similar to a burn and is managed in the same manner (16).

One long term sequelae of gun shot wounds to the peripheral nervous system known as causalgia (complex regional pain syndrome I) was first described during the Civil War by Silas Weir Mitchell. The patients present with burning pain, paresthesias, skin atrophy, and temperature changes typical of the syndrome. This condition is also seen in electrical injury patients who survive the electrical contact. Axonal injury is usually from the direct thermal injury. Myelin injury is from direct thermal contact as well as from the electrical destruction of the myofibrils in the muscle. Post-neurological symptoms to the peripheral nervous system can vary from neuropathy to reflex sympathetic dystrophy (complex regional pain syndrome II) (17).

Paresthesias are thought to be due from perineurial fibrosis resulting in symptoms of a compressive peripheral neuropathy (18); surgical decompression can sometimes ameliorate the symptoms. Sympathectomy has been used to provide relief in those patients that respond to sympathetic blocks (19). In general nerve injuries associated with gun shot and electrical injuries are treated as closed nerve injuries.

Nerve Injury Associated with Soft Tissue Avulsion Injury

Not to be confused with an avulsion of the nerve itself, a soft tissue avulsion or degloving injury presents a unique problem of coverage. Nerves that are exposed will need coverage to prevent desiccation. The type of soft tissue coverage will depend on the location of the injury and is addressed elsewhere in this book. The coverage can be as simple as replacing missing skin with a skin graft to something more extensive that requires fascia or muscle with skin graft, and ultimately a musculocutaneous or fasciocutaneous flap. If the soft tissue is missing directly over the nerve repair, a flap is needed to cover the nerve repair; this can be either local or a free tissue transfer. The use of vacuum assisted closure dressings has changed our management of complex open fractures and may be used to cover a wound with exposed nerves as well. If the nerves are exposed but uninjured, a protective dressing that keeps the nerves from desiccation may be used, such as a layer of AlloDermw with or without a wound vacuum assisted closure dressing, or one of the various hydrogels on the market that provides enzymatic debridement at the same time as it absorbs excess fluid from the wound. If the nerves are injured and need to be repaired, primarily or with a nerve graft, the definitive repair should be staged so that it is done at the time of the soft tissue.

Nerve Injury Associated with Bony Injury

Nerve injuries resulting from isolated closed fractures are most often due to compression from surrounding tissue edema or are due to, though less frequently, nerve laceration from the fracture ends. Iatrogenic nerve injury may also occur in the course of fracture fixation (20 ,21).

The anatomic positions of the radial, median, and ulnar nerves and their major branches make them vulnerable at several sites as they course the upper extremity, which explains the typical relationship seen between particular nerve injuries and fracture patterns (22). The more common nerve injuries with associated fracture patterns are listed in Table 5-3 and include the distal humeral shaft fractures and radial nerve palsy (23), posterior interosseous nerve injury with Monteggia fracture-dislocations (24), and median nerve and/or radial nerve injuries with supracondylar and medial epicondyle fractures in children (25). The time frame to intervention remains controversial; however, dysfunction lasting more than 3 to 4 months necessitates investigation and possible surgical exploration with neurolysis, transposition, repair, and/or reconstruction with nerve grafts or nerve transfers.

Fracture
    Humerus (mid/distal)
Radial n.
Non-operative if suspect compression

Exploration if suspect laceration
    Supracondylar
PIN

Neurolysis, possible grafting if no recovery after 3 months

Med epicondylar

Radius (prox)

Median n

Neurolysis, possible grafting if no

Radial n.

recovery after 3 months

Radius (mid/distal)

Median n.

Carpal tunnel release for neuropathy

Tibia (prox)

Peroneal n.

Peroneal n. release at fibular head for foot drop

Tibia (distal)

Tibial n.

Tarsal tunnel release for neuropathy

Injury Pattern	Nerve(s) Injured	Treatment

Table 5-3. Typical nerve injury associated with fractures

Shoulder

Axillary

Nerve reconstruction with grafts or

Musculocutaneous

nerve transfer of Ax or MC

Elbow

Ulnar n.

Ulnar nerve transposition, neurolysis of AIN,

AIN
reconstruction with nerve grafts or transfers
Hip
Sciatic n.
Neurolysis, possible graft repair, lower

Tibial n.
extremity nerve transfers

Peroneal n.

Knee
Peroneal n.
Peroneal n release at fibular head, tarsal tunnel

Tibial n.
release, possible graft repair or nerve transfer

Joint affected	Nerve(s) Injured	Treatment if conservative treatment fails

Table 5-4. Typical nerve injury associated with dislocations

For patients with chronic nerve injuries extending beyond 2 years, options for primary repair can be limited. Muscle fibrosis and motor end plate degeneration make attempts at primary muscle reinnervation unsuccessful. Instead, in these situations, tendon transfers can provide improvement in function. In comparison, nerve grafting for sensory nerve recovery is not time dependant and can be performed at any time following trauma.

Dislocations and their associated nerve injuries are listed in Table 5-4 . Ulnar nerve injury at the cubital tunnel and/or anterior interosseous nerve injuries are common in elbow dislocations (26 ,27). Shoulder dislocations have been implicated in upper plexus nerve injuries with the axillary nerve being the most vulnerable to injury (28). In our experience with brachial plexus injuries, a second separate injury at the level of the quadrangular space is also often noted; this â€œdouble crush phenomenonâ€• (two separate nerve injuries occurring along the length of a major peripheral nerve) may account for the poor recovery that is cited in the literature with regard to axillary nerve recovery in these injury patterns (29). The double crush syndrome was originally described in the upper extremity to explain nerve problems resulting from a combination of distal nerve compression at the wrist or elbow in conjunction with proximal cervical root or thoracic outlet pathology (30). Nerves that are traumatized proximally may also become more susceptible to compression at distal sites (31).

In the lower extremity, between 10% and 25% of cases of acetabular fracture and traumatic posterior hip dislocation are associated with sciatic nerve injuries (32 ,33). Tethering of the sciatic nerve at the sciatic notch may exaggerate the traction effect on the nerve in the buttock, thus causing a stretch injury proximally and then a distal compression at the fibular head where the nerve is again tethered. The common peroneal portion of the sciatic nerve is more vulnerable because of its anatomical position and internal architecture (34).

Nerves that are not tethered, but which reside in tight fibro-osseous tunnels (such as the ulnar nerve in the cubital tunnel, the median nerve in the carpal tunnel, and the posterior tibial nerve in the tarsal tunnel), may be affected by a more proximal injury due to tissue edema. For example, the incidence of acute transient median nerve compression syndrome after distal radius fracture is estimated between 12% and 17%, and it occurs regardless of fracture type, the amount of initial displacement, the adequacy of reduction, or the method of operative treatment (35 ,36). In the majority of cases, the neuropathy resolve spontaneously; however, it certain cases, the symptoms persist after the fracture has healed. This leads to the recommendation of performing prophylactic carpal tunnel decompression whenever there is a distal radius fracture. More recently, the consensus has reversed as there is no advantage to the prophylaxis; in fact, evidence suggests there is increased morbidity to acutely decompress the carpal tunnel in the face of mild median nerve compression (37 ,38). In contrast, acute severe carpal tunnel with progressive symptoms despite preliminary fracture reduction *is* an indication for surgical release at time of distal radius fixation. Similar conditions are seen in the lower extremity, such as acute foot drop after knee reconstruction for ligament injury or tarsal tunnel with distal tibia or maleolar fractures (39). In general, if a patient has loss of median nerve sensation following forearm fracture or injury, a carpal tunnel release is universally recommended. By contrast, a patient with foot drop following a knee injury is treated expectantly often for several months. We believe there should be *no* distinction between the upper and lower extremities and recommend peroneal nerve release at the fibular head in the case of acute foot drop, just as we would perform a carpal tunnel release for acute compression of the median nerve at the wrist.

Preoperative Assessment

Emergency evaluation of a nerve injury should include a thorough motor and sensory exam. Grossly, the muscle function of the nerve in question needs to be evaluated. For the hand, two point static and moving discrimination can be performed to determine the decrease of sensation. Alternatively, the quick and easy â€œten testâ€• sensory exam uses the patient's own subjective perception to moving light touch in order to elicit differences in sensation (40). For example, to test for a median nerve injury, both the injured and uninjured index fingers of the patient are touched at the same time over corresponding areas of each finger and the patient is asked if the subjective sensation is the same or different. This technique is particularly useful in children, is extremely sensitive, and requires no instrumentation.

In patients who present with acute motor deficits or palsy, determining if the nerve will recover spontaneously can often be confusing. The mechanism of the injury can often assist in the initial evaluation. Any sharp penetrating injury with no clinical evidence of recovery should be explored. Optimal timing is between 2 to 14 days, as long as the patient is surgically stable. Occasionally, an MRI or CT scan may be useful if there is evidence of a neuroma or neural disruption; however, any injury where there is a high index of suspicion of nerve transection should be explored and repaired. The advantage of repair within the first 2 weeks of injury is that the nerve ends have not retracted and primary repair is often possible.

Closed traction injuries and closed fractures with palsy are the most difficult to assess. Waiting 3 to 4 months for evidence of spontaneous recovery is the gold standard. EMGs at 4 months are advised when no clinical recovery is evident. If there is no evidence of reinnervation occurring (MUPs), surgical exploration and reconstruction will be necessary. Figure 5-2 shows the algorithm for sequence of evaluation and repair. If an initial EMG shows some recovery, a follow up EMG is done 4 to 6 weeks later and correlated with clinical evidence of recovery. Once nerve recovery stops progressing as expected, surgical intervention can be planned based on the EMG results and the clinical exam without further delay.

We always prefer to be involved with the care of the patient early and follow the clinical exam and any additional testing, such as electrodiagnostic testing, ultrasound, MRI, and CT scan in order to minimize delay in recognizing a nerve deficit that will not resolve spontaneously. It is important to be aware that the injury may not be limited to only the site of the trauma, but may affect distal functioning muscle groups and may be impinged at distal sites or proximal sites. The more complex injuries will usually require a team approach, both for the operative reconstruction and for the postoperative management. A pain specialist is paramount to manage the associated chronic pain these patients often have for the more complex injuries. Physical therapy and occupational therapy during the rehabilitative period is essential to prevent joint contracture, to fabricate and adjust protective splints, and to assist in motor and sensory re-education as the recovery process is underway.

Operative Management

For acutely transected nerves primary repair remains the gold standard. Adequate resection of the nerve ends beyond the zone of nerve injury is essential to ensure healing with or without limited intraneural scaring. Large associated soft tissue wounds with varying levels of tissue destruction, like degloving injuries and crush injuries, with comminuted bony fractures will need serial debridement prior to ultimate reconstruction. In such cases primary nerve repair should be delayed until one can adequately assess the complete nerve defect. If the nerve cannot be approximated after debridement, a graft will be necessary. Nerve grafts may be used as long as the wound is clean. If the wound is significantly contaminated it is preferable to tag the ends of the nerves and return to graft when all devitalized tissue has been debrided. We do not recommend acutely repairing or grafting nerve injuries in contaminated wound beds or in crush injuries, where multiple debridements will be required, as postoperative infection, ongoing tissue necrosis, and progressive soft tissue ischemia will all contribute to poor return of function following nerve repair.

Formal intraoperative nerve stimulation is helpful in complex closed traction injuries, especially of the brachial plexus. Proximal injuries of mixed nerves result in complete or partial deficits to several nerves, as opposed to the discrete single nerve injury pattern of distal peripheral nerve injuries. In most cases, an injury distal to the shoulder level will result in one or two discrete nerve injuries and intraoperative electrical stimulation is not necessary. In the case where several segments along the same nerve are injured, or the brachial plexus is injured, electrical stimulation is useful in guiding

intra-operative decision making. Electrical stimulation allows the surgeon to determine which fascicles within the nerve are functional and which ones are not. We limit use of formal intraoperative electrical stimulation to three major situations: (a) in cases where the preoperative EMG results were equivocal; (b) in cases where additional recovery may have occurred between the time of the previous study and the operation; and (c) in situations where the intraoperative mapping of the nerve function can affect the surgical plan. An example of a situation where the intraoperative plan may change is in the case of an incomplete neuroma of the median nerve at the wrist. If the median motor fascicle is intact, the surgeon may try to preserve those fibers and graft the sensory fibers that are injured. Alternatively, if the neuroma is nearly circumferential and the motor fibers poorly conduct to the thenar muscles, the surgeon may opt to completely resect the neuroma and primarily repair or graft the nerve defect.

A handheld 2 mA nerve stimulator will not provide as much information about nerve function as a formal intraoperative electrical stimulation, but is useful in many situations. In the first 72 hours after an acute transaction, the distal half of a motor nerve will contract when stimulated, facilitating the matching of nerves when several are cut, or when the nerve is cut close to the muscle and has begun to branch. The small motor fibers would be otherwise difficult to identify because of their small diameter, which can sometimes be as small as 0.5 to 2 mm in diameter. We find the handheld device a useful aid when performing nerve transfers; it allows us to pick out grossly motor nerve fibers that were uninjured, which will be spliced into nonfunctioning motor nerves to restore animation. For a nerve injury that develops a partial neuroma (as in a 6th degree nerve injury), a nerve stimulator can help determine which nerve fascicles to preserve while resecting the damaged tissue, as long as the fascicles in question are mixed nerves. For injuries to purely or mostly sensory fibers, a handheld nerve stimulator is not useful and formal intraoperative nerve stimulation is the only means to evaluate sensory conduction.

Common Clinical Examples

Primary Repair for Nerve Transection

A common soft tissue injury that orthopaedic surgeons may encounter and easily repair is the laceration of a major and minor peripheral nerve as a result of a sharp penetrating injury. In the mid to distal forearm and wrist, the median nerve is fairly superficial, and thus easily injured from a deep laceration with glass or metal. The median nerve can be found deep to the palmaris tendon in patients that have a palmaris longus tendon, and just ulnar to the flexor carpi radialis tendon. Knowledge of the anatomy as well as the topography of the nerve is important in order to align the motor branches when repairing the nerve. Primary repair is the procedure of choice whenever possible and usually can be performed within the first 2 weeks after injury. Repairs performed later than this often need a nerve graft as the nerve retracts with time. The repair should be performed with as little tension as possible at the repair site in order to minimize scarring that occurs within the nerve. A small amount of tension at the repair site is acceptable and has been shown in some animal studies to stimulate neurotropic growth factors and improve healing (41). A useful tool in determining excessive tension is to bring in proximity the two ends of a nerve with a single 8-0 nylon suture. If the freshened nerve ends reach without the suture pulling through the epineurium, the tension is not excessive.

Mobilizing the nerve is crucial. Often sufficient mobilization of the nerve will eliminate the need for a graft. Because of the inherent springiness of nerve tissue, when a nerve is transected it will recoil, leaving a gap. A nerve can stretch 10% to 15% without compromising its inherent blood supply (42). If, when bringing the nerve ends together, there is still a 5 mm gap between the nerve ends, mobilizing 5 cm of nerve (usually 2.5 cm on either side, but not always) will allow the ends to come together. If a 1 cm nerve gap is present when the two freshened ends of nerve are laid in proximity, 10 cm of total nerve needs to be mobilized in order to bring the two halves together with â€œminimalâ€• tension. This does not mean that any gap should be closed primarily at all costs. If after reasonable mobilization there is a persistent gap, a nerve graft is preferable to forcing a primary repair. It is important to note that if a nerve graft is necessary, the nerve ends should be trimmed generously so that the repair will be outside the zone of injury. When a nerve graft is needed, there is no reason to limit the debridement of the remaining nerve.

External markers such as the native vessel of the nerve on the surface is used along with the different fascicular size and grouping to align the nerve (Fig. 5-4). In the past, it was believed that a perineural repair of grouped fascicles was better than an epineural repair; however, recent studies show that either technique is equally effective (43). We prefer to use a circumferential epineural repair with 9-0 nylon interrupted sutures to minimize the fibrosis seen with perineural repairs (Fig. 5-4B). Care is taken to prevent overlap (Fig. 5-5); we prefer to have the fascicle slightly retracted with respect to the epineurium, rather than bulging out the side or overlapping internally. At the conclusion of the repair, the area under question is put through the full range of movement to verify that the repair will hold up under gentile protected motion postoperatively. Any repair that withstands gentle range of motion is considered a â€œtension-freeâ€• repair, although in fact it may be still under some mild yet acceptable tension. In larger nerves, such as the sciatic and femoral nerves, or in nerves with significant inflammation in the epineurium, we use 8-0 nylon sutures. In contrast, for cutaneous nerves and digital nerves, 10-0 nylon works best. The least number of epineural sutures to approximate and keep the repair intact through a full range of motion is used. After neurorrhaphy, the area is immobilized for 1 to 2 weeks; however, protected active range of motion is allowed and should be initiated immediately. One exception is in flexor tendon lacerations with associated digital nerve injuries. Both are repaired and the patient begins early protected range of motion under the direction of an occupational therapist. While sensibility is somewhat decreased when compared to patients who are immobilized for 3 to 4 weeks, there is no statistical difference of sensibility at 1 year, and the advantages of early range of motion for the rehabilitation of the tendon injury outweighs the negligible delay in sensory return (44).

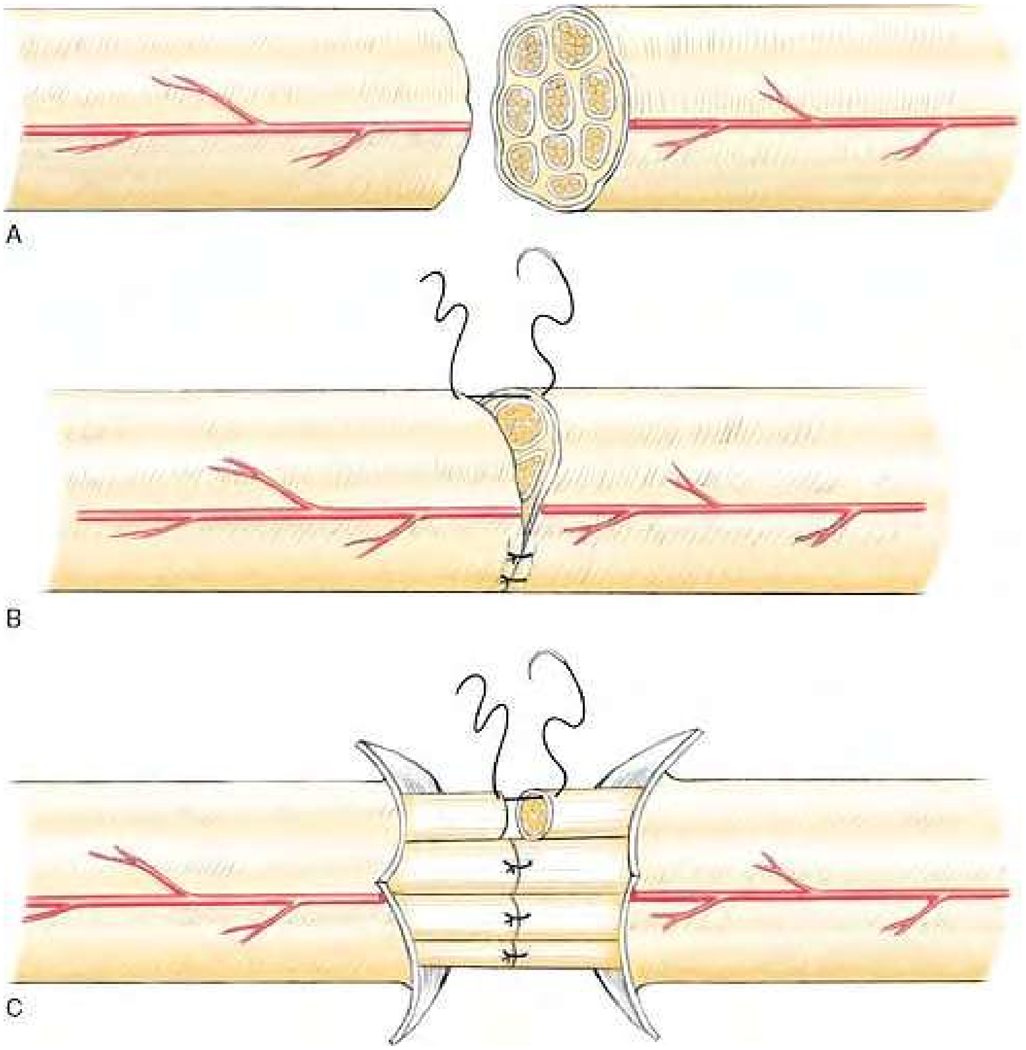

FIGURE 5-4 The two nerve ends are aligned using external markers such as the artery and the grouped fascicle pattern (A) , an epineural (B) or perineural (C) repair may be used, although we prefer the former.

Repair of Nerve Gap with Nerve Graft

While a primary repair is preferred, two neurorrhaphy sites under favorable conditions are better than a single neurorrhaphy under unfavorable conditions (45). In contaminated wounds, such as crush injuries from motor vehicle accidents, the wound bed needs to be cleaned before any reconstruction of the resulting deformities can be performed. If a large defect to both the nerve and the soft tissue is present, nerve repair with grafts in addition to soft tissue coverage will be needed. In areas where local muscle is unavailable, a free tissue transfer may be necessary to cover both bone and nerve repair.

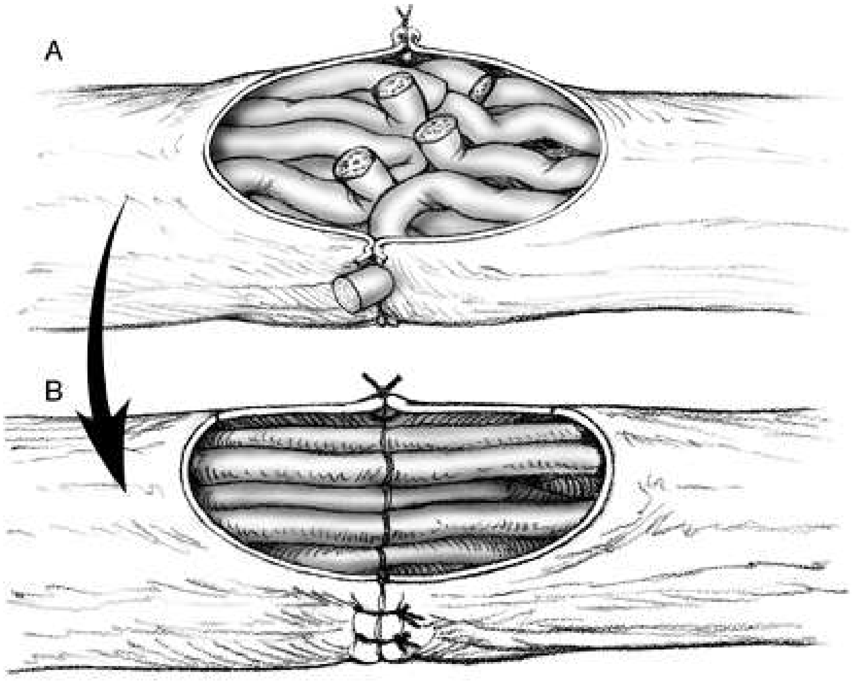

FIGURE 5-5 When aligning the nerve, make sure to trim back the fascicles so that there is no overlap when the epineurium is sutured closed (B) . It is preferable to have a small gap internally than to have fibers overlap and potentially â€œescapeâ€• from the edge or â€œfail to findâ€• the distal half (A) . (By permission of Mayo Foundation for Medical Education and Research. All rights reserved.)

A common situation which may require a nerve graft can be seen in cases of severe knee dislocation with damage to the tibial-peroneal nerve trunk. The stretch injury, if severe enough, will cause internal damage of the nerve. While EMG studies may confirm the suspicion of a third or fourth degree injury, the history and clinical findings are usually enough to warrant surgical exploration with the plan to release the nerve at the very least from the surrounding scar tissue and compression points, and most likely excise the neuroma and graft the resulting nerve defect. If the neuroma is small, a primary repair may be possible; however, the resultant gap following neuroma excision is often not amenable to primary repair as mobilization of the sciatic nerve proximally and the tibial and peroneal nerves distally are limited by nerve branches to the underlying muscles. The resultant 2 to 3 cm gap will need to be repaired with a nerve graft. A sural nerve is the most often used donor nerve. In the case of a tibial-peroneal nerve injury described above, the nerve would be cut in half or thirds and cabled. Once again, 9-0 nylon interrupted sutures are used to repair both sites. The native artery location on the external portion of the nerve as well as the relative size of the fascicles is used to help line up the fascicles. We still prefer to do an epineural repair of the nerve to nerve graft, rather than a perineural repair of the actual fascicles.

Nerve grafts less than 10 cm long work well in vascularized wound beds. The longer the graft, the less optimal the recovery. Some people have suggested using vascularized nerve grafts when the distance is greater than 20 cm (46). However, in general, small caliber nerve grafts, such as sural and medial antibrachial cutaneous nerves, do not need to be vascularized. Large caliber nerves, such as ulnar, do need to be vascularized or the central portion will necrose and then scar, limiting its effectiveness as a graft.

Obtaining a Nerve Graft

The nerve graft serves as a guide for the proximal axon as it regrows toward the distal stump. The sural nerve is by far the most commonly used donor nerve, although other suitable donor nerves include the lateral and medial antibrachial cutaneous nerve (47 ,48) and the terminal portion of the anterior interosseous nerve (AIN) that innervates the pronator quadratus (49). The sural nerve is often chosen as a nerve donor because of its size, length, and relatively minimal donor site and minimal morbidity. Typically, harvest of the sural nerve will result in numbness over the posterior aspect of the leg and the lateral aspect of the foot and malleolus.

With the patient either prone or supine with the leg frog-legged, the sural nerve is identified lateral to the Achilles tendon. The easiest method for harvest of the sural nerve is to use a posterior midline incision. The closure is slightly lengthier and the resulting scar may be cosmetically undesirable. The alternative is to use a stepwise technique, which requires between 4 and 6 separate horizontal incisions along the back of the calf, each approximately 1 cm in length. The resulting scar is cosmetically more favorable and the technique adds little operative time. An endoscopic or a nerve harvester may be used to limit scars. The trick with using the nerve harvester is that contributions from the lateral and anterior sural nerves, which are cutaneous branches from the peroneal and tibial nerves, may be difficult to incorporate into the sural nerve proper and shorten the overall length that is harvested. Approximately 25 to 30 cm of graft may be harvested from a typical adult patient. The next most often used graft is the medial anti-brachial nerve. It is easily found coursing along side the bacillic vein in the upper arm. Up to 25 cm of nerve graft may be obtained.

While nerve grafts are the gold standard for motor nerve gap repairs, the disadvantage is the limited number of donor nerves available. This has led to the development of new techniques for bridging nerve gaps. For nerve gaps less than or equal to 3 cm on small caliber sensory nerves (50), the nerve may be repaired with a conduit, which may be autologous, such as a vein, or synthetic, such as one of the commercially available nerve tubes (Fig. 5-6A). The end of each nerve is trimmed until healthy fascicles are seen. The end of the nerve is inserted into the tube for approximately 2 to 4 mm and the epineurium is sutured to the end of the conduit with 9-0 nylon using a simple suture technique, a horizontal matrass suture technique, or a half buried matrass suture technique. Clinically, for nerve defects longer than 3 cm, in mixed nerves, and in pure motor nerves the recommendation is still a nerve graft. Alternatively, we have used nerve conduits as a covering over nerve repairs that required several cabled grafts into one larger nerve to serve as a temporary protective covering instead of wrapping a spatulated strip of vein (Fig. 5-6Bâ€“D).

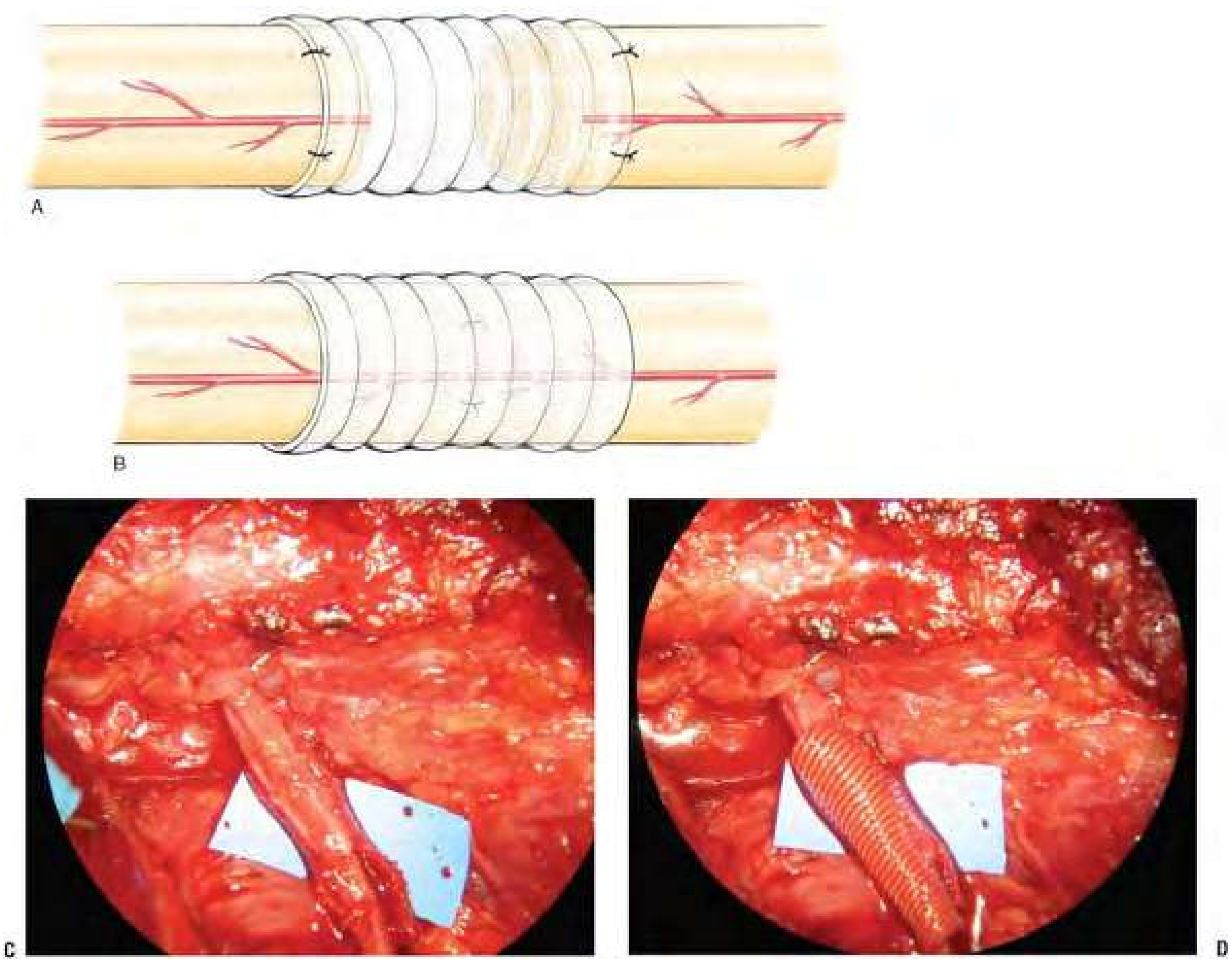

FIGURE 5-6 Neural tube may be used to bridge a gap or as a protective covering after nerve repair. When used to bridge a nerve gap (A) , the epineurium is sutured to the edge of the neural tube. Bâ€“D: When used to cover a neurorrhaphy site, the tube may be cut longitudinally and wrapped around the neurorrhaphy site. A suture is placed around the outside of the conduit to help keep it in place.

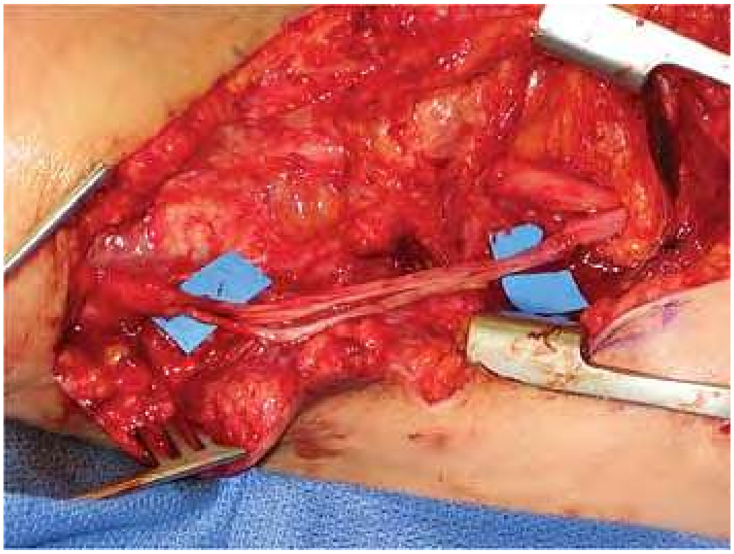

FIGURE 5-7 Three cabled sural nerves, each 2 mm in diameter, were used to bridge the defect in the radial nerve. The diameter of the radial nerve is approximately 6 mm in this location.

Neuroma of Radial Nerve with Humeral Fracture

Considerable controversy remains as to the need for early exploration in closed fracture dislocation injuries with acute nerve palsy. In a recent meta-analysis review of the literature over the last 40 years regarding radial nerve palsy after fracture of the shaft of the humerus, the prevalence of fracture associated nerve palsy was about 12% (51). Over 70% of these compressions resolve spontaneously, therefore the current recommendation is for conservative therapy with closed fractures. If after 3 months there is no clinical recovery, using high-resolution ultrasound to evaluate the injured nerve is less invasive than electrodiagnostic testing and has been used in this specific situation for effectively detecting a neuroma in the presence of a fracture (26 ,52). However, at 3 to 4 months, if spontaneous recovery is to be anticipated, there should be clinical evidence of reinnervation or electrical evidence of regeneration (motor unit potentials). If radial nerve regeneration stops at the level of the Arcade of Frohse, then decompression of the posterior interosseous nerve is indicated. In addition, open fractures, oblique fractures, and fractures that can compress or transect the radial nerve between the fracture fragments should be explored early (53).

If instead of recovery, all preoperative studies indicate a neuroma of the radial nerve in the distal arm, the nerve should be explored with the plan to resect and graft the defect. Rarely can the nerve be mobilized enough to be repaired primarily. Because the nerve and scar surrounding it are usually densely adherent and are similar in color, entering the scarred area directly is not advised. The uninjured proximal and distal ends of the nerve are exposed first. In the case of a partial palsy or radial nerve weakness, a sixth degree nerve injury is most likely. A handheld nerve stimulator can be used to stimulate fibers proximal to the injury. Those fascicles that result in movement of muscles distal to the injury are intact and these intact fibers are marked and protected from further dissection. The injured fibers are resected until normal fascicles are noted from each end. A nerve graft is used similar to before. If more than one third of the radial nerve has been resected, two or more cabled sural nerve grafts may be needed to bridge the gap (Fig. 5-7). The larger sized fascicles should be lined up and grafted directly whenever possible. 9-0 or 8-0 nylon interrupted sutures are used to repair both sites with an epineural or perineurial repair.

Postoperative Management

After any typical nerve repair, whether it is done primarily, with a nerve graft or a synthetic tube, or a primary neurorrhaphy using a nerve transfer, the affected area is splinted or placed in a sling for 2 weeks to minimize movement. The only exception is in digital nerve repairs in conjunction with a tendon repair as noted before. We begin early passive range of motion to prevent tendon adhesion and leave the patient in a dorsal blocking splint to prevent hyperextension and additional strain to the repair site.

In both motor and sensory nerve repairs, an advancing Tinel sign or â€œtinglingâ€• sign seen during the regeneration phase is followed until the patient has restoration of function or sensation. Even before there is clinical evidence of recovery, we begin desensitization of the affected area to assist in sensory reeducation and to prevent hypersensitivity. Formal objective testing, such as Simien-Weinstein testing, and sensory grading is useful for objective data collection, but is not necessarily useful in clinical practice. Motor reeducation begins as soon as the patient has some clinical evidence of muscle contraction.

Occasionally, EMGs are used postoperatively to follow functional recovery. This is useful in cases where a primary repair may not show clinical evidence of recovery and a second stage reconstruction may be planned. In those cases, EMGs are invaluable in answering the question of whether to wait longer in the case of recovery that may not manifest in actual clinical movement yet, versus proceeding with the next phase of reconstruction, if no evidence of recovery is present.

Conclusion

Nerve injuries alone are fairly complex and when the soft tissue and bony structures in the vicinity are injured as well, the complexity increases. Bony trauma with or without soft tissue injury indicates a greater force was expended to cause the injury, thus the associated nerve injuries can be severe and at multiple levels. Avulsion and stretch injuries, crush injuries, and penetrating injuries can occur in combination or alone. The mechanism of injury varies greatly, thus no single solution to all nerve injuries exists. The deficits are often not readily apparent because of concomitant injuries or because the patient can not reliably participate in the examination. The reconstruction of both the nerves and the soft tissue will ultimately be tailored to the specific situation. Any study that can aid in the decision process, such as radiology exams and electrical diagnostic testing, should be ordered in a timely fashion. Additional assistance from the following may be needed: (a) a plastic or orthopaedic surgeon knowledgeable about peripheral nerve surgery and soft tissue coverage if the defect is more complex than initially anticipated or if extensive wound coverage issues are beyond the scope of the treating physician; (b) a pain specialist for management of chronic narcotic and nonnarcotic pain medication and nonoperative therapies such as nerve blocks and nerve stimulators for chronic pain; (c) a psychiatrist or psychologist for depression if warranted in the complex peripheral nerve injuries; and (d) occupational and physical therapists for motor and sensory reeducation, splinting, and joint mobility maintenance.

References

1. Seddon HJ. Three types of nerve injury. *Brain.* 1943;66:237.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 6

Management of Vascular Injuries Following Soft Tissue and Bony Trauma

Gustavo S. Oderich

Timothy M. Sullivan

Accidental injuries affect 2.6 million people in the United States each year. Mechanisms of injury include penetrating, blunt, and iatrogenic trauma. It is estimated that vascular trauma affects 0.2% to 4% of all injured patients. Therefore, approximately 20 to 100,000 patients will sustain a vascular injury. Over 80% of these vascular injuries will be located in extremities. Ninety percent of all vascular injuries are associated with penetrating trauma, of which 70% are due to gunshot injuries, 20% are due to stab wounds, and 10% are due to blunt trauma. Combined vascular and orthopaedic trauma is relatively uncommon, accounting for less than 1% of all cases of traumatic injuries. Vascular injuries can occur because of the superficial location of vessels, their proximity to bones, and their relatively fixed position across joints. Patients sustaining combined injuries are exposed to substantially increased risk of amputation and limb dysfunction. This chapter focuses on the indications, contraindications, preoperative planning, operative approach, and results of repair of the most common combined vascular and orthopaedic injuries.

Indications/Contraindications

Primary repair of arterial and venous injuries is indicated to control bleeding and/or to relieve limb or organ ischemia. Life-threatening injuries should be recognized early during the resuscitation phase and prioritized before proceeding with vascular and orthopaedic repair. Arterial injuries affecting only one of the tibioperoneal or forearm arteries may be treated conservatively provided that there is no evidence of distal ischemia. Nonocclusive arterial injuries (e.g., small intimal flaps or dissections) incidentally found on imaging studies can be safely observed. The benign natural history of these injuries is well documented in several large series; operative repair is required in less than 1% of patients.

Extremity trauma with complex soft tissue, vascular, and skeletal injuries poses one of the most difficult management problems. Patients should be evaluated by a multidisciplinary team. Coordinated interaction of various specialists including a vascular, orthopaedic, and plastic surgeon is of paramount importance to optimize outcome. Injuries of the head, chest, and abdomen may require additional neurosurgical and general trauma consultation. Every effort is made to balance the potential success of an arterial reconstruction and correction of the orthopaedic injury with the overall clinical status of the patient and the potential for complete functional recovery. In general, once life-threatening injuries are stabilized, treatment priorities are as follows:

- Control of bleeding and restoration of arterial inflow.
- Fracture reduction and stabilization.
- Soft tissue coverage.

Vascular repair should be prioritized over definitive orthopaedic repair. Any delay in vascular reconstruction is a gamble and may risk the only opportunity for limb salvage. Patients with stable fractures or dislocations in which minimal manipulation and length discrepancy is anticipated should be treated with immediate definitive arterial revascularization. However, patients with severely comminuted fractures and dislocations, segmental bone loss causing limb discrepancy, or severe soft tissue disruption and contamination should be treated initially with temporary intra-arterial shunts. In these cases, it is wise to delay the definitive vascular reconstruction until wide debridement and initial skeletal repair are accomplished.

Contra-indications for vascular repair include presence of other life-threatening injuries requiring immediate attention or causing hemodynamic instability. Several predictive factors should be taken into consideration when deciding to perform revascularization versus primary amputation (Table 6-1). Overall, approximately 10% to 20% of patients with complex extremity injuries have nonsalvageable limbs and require primary amputation. Assessment of these patients should be individualized. Several factors should be taken into consideration, including the overall clinical status, severity of the arterial, neurologic, and orthopaedic trauma, and expected functional recovery. A primary amputation is considered in cases of dysvascular extremity with complex fractures and extensive soft tissue and nerve damage. Patients with major nerve transections (e.g., tibial nerve transection) and open comminuted tibiofibular fractures with arterial injuries (Gustilo III-C) have very poor functional outcome and high amputation rates; these patients are generally treated with primary amputation. Major nerve transections should be confirmed by direct visualization. Other indication for amputation is prolonged ischemia time (>12 hours) with evidence of a cadaveric extremity (e.g., mottled with absence of motor function, arterial, or venous Doppler signals).

Preoperative Planning

Prolonged ischemia time is the most important factor associated with limb dysfunction or amputation. A high index of suspicion coupled with accurate neurovascular examination is necessary for prompt diagnosis of a vascular injury. Accurate diagnosis of vascular injuries is an essential aspect in the preoperative evaluation of patients with complex extremity trauma. Prompt restoration of arterial blood flow within 6 hours from the time of initial extremity injury is the most critical factor that determines limb salvage and function.

Complete history including the mechanism of trauma, associated injuries, medical history, medications, and allergies should be recorded. Extremity vascular trauma is immediately apparent because of external bleeding, hematoma, or obvious limb ischemia. Physical examination includes inspection of the injured limb for open wounds, obvious deformities, and signs of ischemia. Distal ischemia is manifested by the five Ps: pallor, paresthesia, paralysis, pain, pulselessness, and poikilothermia. A thorough sensory and motor examination and pulse examination should be noted. The presence of hematoma, pulsatile or not, bruit, and thrill must be documented. If distal pulses are diminished or absent, the ankle-brachial index should be determined using a hand-held Doppler device.

Signs of arterial injury have been traditionally classified into â€œhardâ€• and â€œsoftâ€• signs (Table 6-2). These correlate with the presence of a hemodynamically significant arterial lesion. â€œHardâ€• signs include

absence of distal pulses, pulsatile bleeding, expanding hematoma, palpable thrill, and audible bruit. â€œSoftâ€• signs include proximity to the vessel, peripheral nerve deficit, history of moderate hemorrhage, and *diminished* distal pulses.

Table 6-1. Factors Associated with Poor Functional Recovery or Need for Primary Amputation After Combined Vascular and Orthopaedic Extremity Trauma

Transected sciatic or tibial nerve

Transection of two of the three upper extremity nerves

Gustillo III-C orthopaedic injury

Below-knee arterial injury with two of the three arteries injured

Prolonged limb ischemia greater than 12 hours

Multiple comminuted fractures

Extensive soft tissue loss

Crush injury

Severe contamination

Elderly patients or multiple medical comorbidities

Cadaveric limb	Shock or other life-threatening injuries

Table 6-2. Clinical Signs of Vascular Injury

Hard Signs	Soft Signs
Absent distal pulses	Diminished distal pulses or ABI <0.90
Active pulsatile bleeding	Unexplained hypotension or large blood loss at the scene
Expanding hematoma	Small or moderate nonexpanding hematoma
Bruit	Injury in proximity to a major vessel
Thrill	Neurologic injury in proximity to vessel

ABI, ankle-brachial index.

The general recommendation is that patients with hard signs associated with uncomplicated penetrating trauma should undergo immediate operative exploration, without need for arteriography or duplex ultrasound. The indications for arteriography are summarized in Table 6-3. Patients with multiple penetrating injuries should be evaluated with arteriography to determine the exact location and extent of arterial lesions. Arteriography is also advised in patients with combined arterial and orthopaedic trauma, even in the presence of hard signs. While the presence of hard signs predicts major vascular injury in nearly 100% of patients with uncomplicated penetrating trauma, less than 15% of patients with complex blunt injuries will require vascular repair. The inaccuracy of hard signs in this subgroup is explained by a combination of multiple other factors that ultimately lead to diminished pulses, including fractures, soft tissue disruption, compartment syndrome, and extrinsic arterial compression. Therefore, we generally recommend arteriography in all patients with complex skeletal trauma and hard signs of vascular injury.

We generally prefer a one-shot intraoperative arteriography technique. This avoids the 1 to 3 hour time delay required to obtain a formal arteriography in the angiography suite. An antegrade or retrograde approach in the affected limb is used for most cases. The availability of a portable C-arm and fluoroscopy permits localization of bony landmarks, selective catheterization of arterial branches, and endovascular treatment of arterial lesions. However, because fluoroscopy is usually not readily available in the emergency setting, we usually use ultrasound guidance and a micro-puncture set for arterial access. The target artery (e.g., common femoral artery or brachial artery) is accessed with a micropuncture needle (18 gauge) in an antegrade or retrograde fashion. A 0.018 inch guidewire is advanced to allow placement of a sheath. The sheath can be used for contrast injection. Alternatively, the 0.018 inch wire is exchanged into a 0.035 inch wire and a 4 French sheath is advanced. The arterial inflow proximal to the access site should be manually compressed during contrast injection using digital pressure or a tourniquet. A single hand injection of 30 mL of diluted (50:50) iso-osmolar contrast allows adequate visualization of the area of concern and distal runoff vessels.

Other noninvasive imaging modalities are duplex ultrasound, computed tomography angiography (CTA), and magnetic resonance angiography (MRA). Computed tomography angiography is now available in most centers, is quite expeditious, and permits excellent image of the arterial circulation, soft tissue, and bone. This is particularly useful for planning operative approach in patients with centrally located lesions (e.g, aorta, subclavian, or iliac arteries). However, limitations are the relatively large contrast load (150 mL) in a patient who may potentially require additional angiography.

We recommend duplex arterial ultrasound in patients with â€œsoftâ€• signs of vascular injury, including those with an ankle-brachial index of less than 0.90.

Arterial imaging is not required in patients with normal pulses and no other sign of arterial injury. Physical examination excludes significant arterial injuries as reliably as arteriography or surgical exploration. This is also true for patients with posterior knee dislocations, in whom arteriography used to be obtained routinely in the 1980s. Results of contemporary series show that in the absence of hard signs, less than 5% of patients had abnormalities on the arteriography but none required operative treatment or had amputation. On the other hand, the incidence of significant vascular injuries is 70% in the presence of hard signs. In patients with normal pulse examination, nonocclusive lesions include small intimal flaps, dissections, contusions, or small pseudoaneurysms; these lesions should be treated with antiplatelet therapy only.

Table 6-3. Indications for Preoperative or Intraoperative Diagnostic Arteriography in Patients with Combined Extremity Trauma

Multiple penetrating injuries
Unclear location or extent of arterial injuries
Extensive soft tissue injury
Fracture or dislocations with hard signs of vascular injury
Trajectory parallel to artery
Underlying peripheral arterial disease

Surgery

Patient Positioning and General Approach

All vascular injuries can be accessed using the supine anatomic position (Fig. 6-1). Although prone position and a posterior approach may be used for isolated popliteal artery injuries which do not extend into the superficial femoral or tibioperoneal arteries, this approach limits access to the great saphenous vein if harvesting is necessary and may require excessive manipulation of the fractured limb. A generous sterile field should be prepared and draped to allow adequate exposure, proximal and distal control, and options of extra-anatomic reconstruction if indicated. The wise surgeon should always consider the â€œworst case scenarioâ€• (e.g., axillofemoral graft). A noninjured lower extremity should be prepped circumferentially for possible vein harvesting. The ideal conduit is an autologous great or small saphenous vein. Preoperative intravenous antibiotics (e.g., a first-generation cephalosporin) should be administered.

Technique

Repair of vascular injuries can be one of the most challenging aspects of trauma management. Some basic principles of vascular repair are applicable to all vascular injuries (Table 6-4). The operative sequence consists of access, exposure, control, and repair. Initial control of external hemorrhage is usually obtained with simple digital or manual pressure. Use of surgical instruments such as hemostats is not only ineffective but risks iatrogenic injury of adjacent nerves or veins. Manual compression is maintained by a member of the surgical team until definitive proximal and distal control of the injured vessel is obtained. The hematoma should not be entered without first obtaining proximal

and distal control away from the site of injury. Access is gained using the standard exposure techniques described in the following discussion. Vascular clamps should not be applied forcefully and blindly; instead, the artery should be completely dissected, looped with a silastic vessel loop, and clamped under direct vision. One important adjunct is the use of balloon occlusion catheters. This facilitates control in areas of difficult access and avoids excessive dissection. In addition, pressure cuff tourniquets may be used in the extremities to achieve prompt vascular control. Once control is gained, the hematoma is explored with careful attention to avoid injury to adjacent structures. The extent of vascular injury is assessed and clamps are moved closer to the vascular wound. The injured vessel should be debrided and cleaned using tenotomy scissors. The extent of debridement should take into consideration the mechanism of injury, with more extensive debridements for high-velocity gunshot wounds, and blast or crush injuries. One should avoid excising the adventitia of the vessel while dissecting or debriding the artery. Distal and proximal thrombectomy using a Fogarty balloon thromboembolectomy catheter should be performed even in the presence of relatively good prograde or retrograde bleeding. The size of the catheter is generally a 4 or 5 for the iliacs, 4 for the superficial femoral, 3 or 4 for the deep femoral, 2 or 3 for the tibials, and 2 for the pedal and smaller arteries of the forearm. Passage of thromboembolectomy catheters into the deep femoral, tibials, and smaller arteries should be done gently because of risk of arterial rupture, dissection, or intimal injury due to excessive balloon inflation. At least two â€œcleanâ€• passes should be made before considering the artery free of thrombus. Preoperative and intraoperative systemic heparinization (80 mg/kg bolus) should be used unless there is a contraindication (e.g., head injury). An activated clotting time (ACT) above 250 seconds is considered optimal. Liberal regional heparinization should be used with flushes of heparinized saline proximally and distally to prevent propagation of any thrombus.

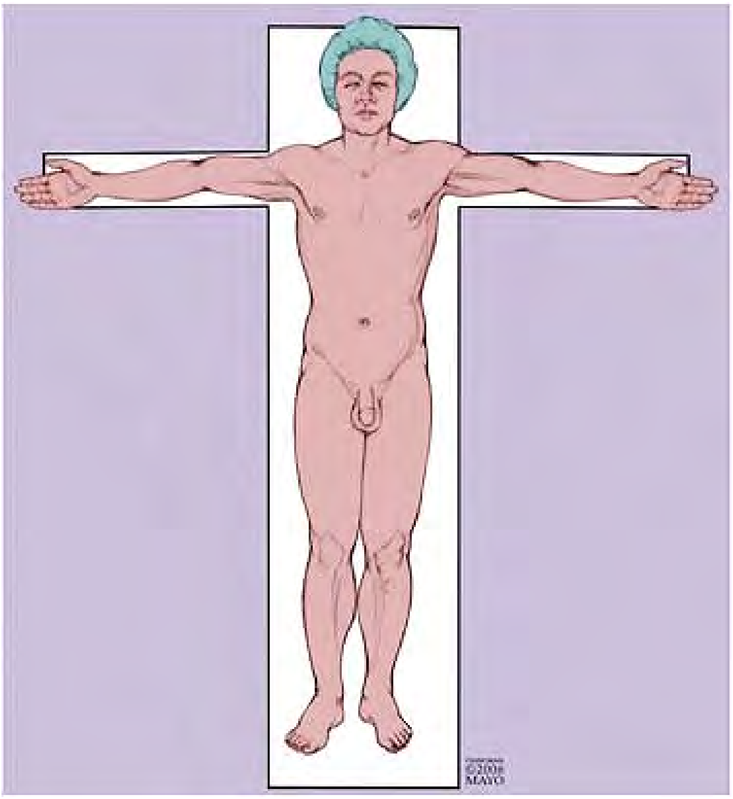

FIGURE 6-1 Patients should be positioned supine with one or both upper extremities abducted. At least one unaffected lower extremity should be circumferentially prepped for vein harvesting if indicated.

Table 6-4. Basic Principles of Repair of Vascular Injuries

Manual pressure for vascular control
Access using standard vascular exposure
Ensure proximal and distal control prior to entering a hematoma
Carefully enter the hematoma, avoiding injury to adjacent structures
Assess the extent of injury and presence of concomitant venous injury
Use systemic heparinization whenever possible
Determine type of vascular repair
Debridement of vessel edges
Proximal and distal Fogarty catheter thromboembolectomy
Instill regional heparin after two â€œcleanâ€• passes
Perform vascular anastomosis in a tensionless fashion
Allow prograde and retrograde bleeding prior to completing the anastomosis
Complete suture line, remove distal and proximal clamp sequentially
Assess distal circulation with pulse examination and hand-held Doppler
Obtain completion arteriography
Consider need for fasciotomy
Ensure adequate soft tissue coverage of vascular reconstruction

Restoration of extremity perfusion does not always require definitive arterial reconstruction. Another important adjunct is a temporary arterial shunt placed in the proximal and distal arterial ends. This promptly controls hemorrhage and re-establishes inflow. Our preference is to use either a short or long Sundt intraluminal arterial shunt (Fig. 6-2). The shunt is secured to the artery using silk suture or a Rummel tourniquet. Although systemic heparinization is preferred, this is not an absolute requirement for shunt placement. The main advantage of temporary shunts is the avoidance of damage to a fresh arterial repair during subsequent orthopaedic manipulations in cases of comminuted, unstable skeletal injuries.

The injured vessels are â€œset upâ€• for the reconstruction using stay sutures and clean white towels. The anastomosis should be performed without any tension using optic magnification and small monofilament suture (e.g., Prolene). In general, a 4 or 5-0 monofilament suture is used for the femoral arteries, and 6 or 7-0 for the popliteal, tibials, and smaller arteries. The type of reconstruction varies and should be tailored to both the patient condition and extent of injury (Fig. 6-3). Simple repair entails a lateral arteriorrhaphy or venorrhaphy. One should avoid excessive narrowing of the lumen and use patch angioplasty if there is any question. A segment of saphenous vein or bovine pericardium can be used for vascular patches. Vessels with injuries involving less than 1.5 cm in length can usually be re-approximated using end-to-end anastomosis. Mobilization of the artery or vein may require ligation of multiple side branches to gain enough length to allow a tensionless anastomosis. The vessel ends should be spatulated to prevent anastomotic narrowing. A running anastomosis is used most cases, but interrupted sutures may be required in smaller vessels or pediatric cases. Before completion of the anastomosis, the proximal and distal clamps are temporarily removed to ensure adequate prograde and retrograde bleeding. Absence of back bleeding indicates thrombus formation or a technical defect and warrants further catheter embolectomy or revision. Approximately 80% of all vascular extremity injuries that occur in association with orthopaedic trauma can be repaired using one of these three techniques: simple closure, end-to-end anastomosis, or patch angioplasty.

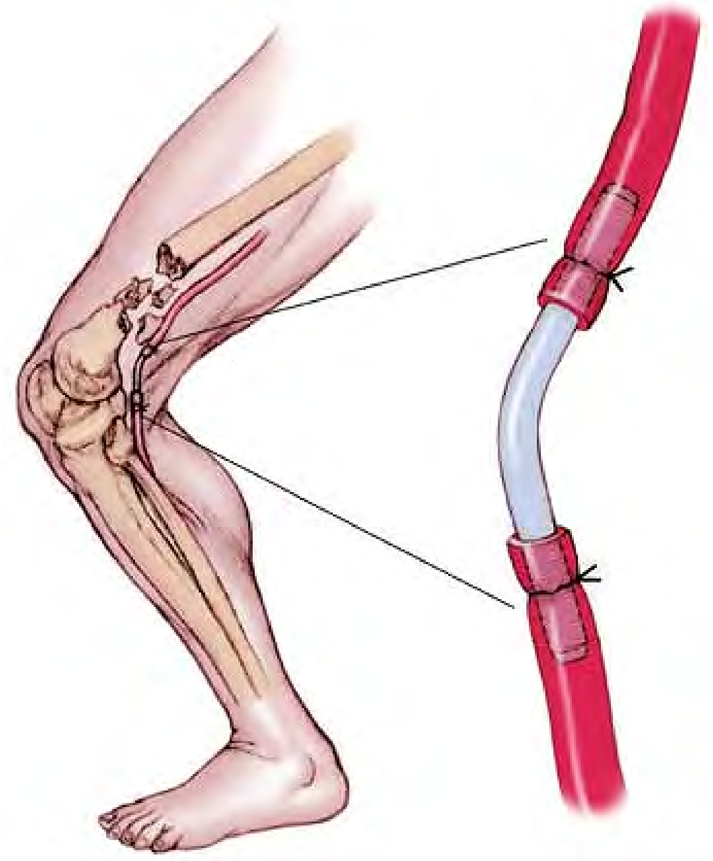

FIGURE 6-2 Intra-arterial Sundt shunt for temporary control and restoration of blood flow.

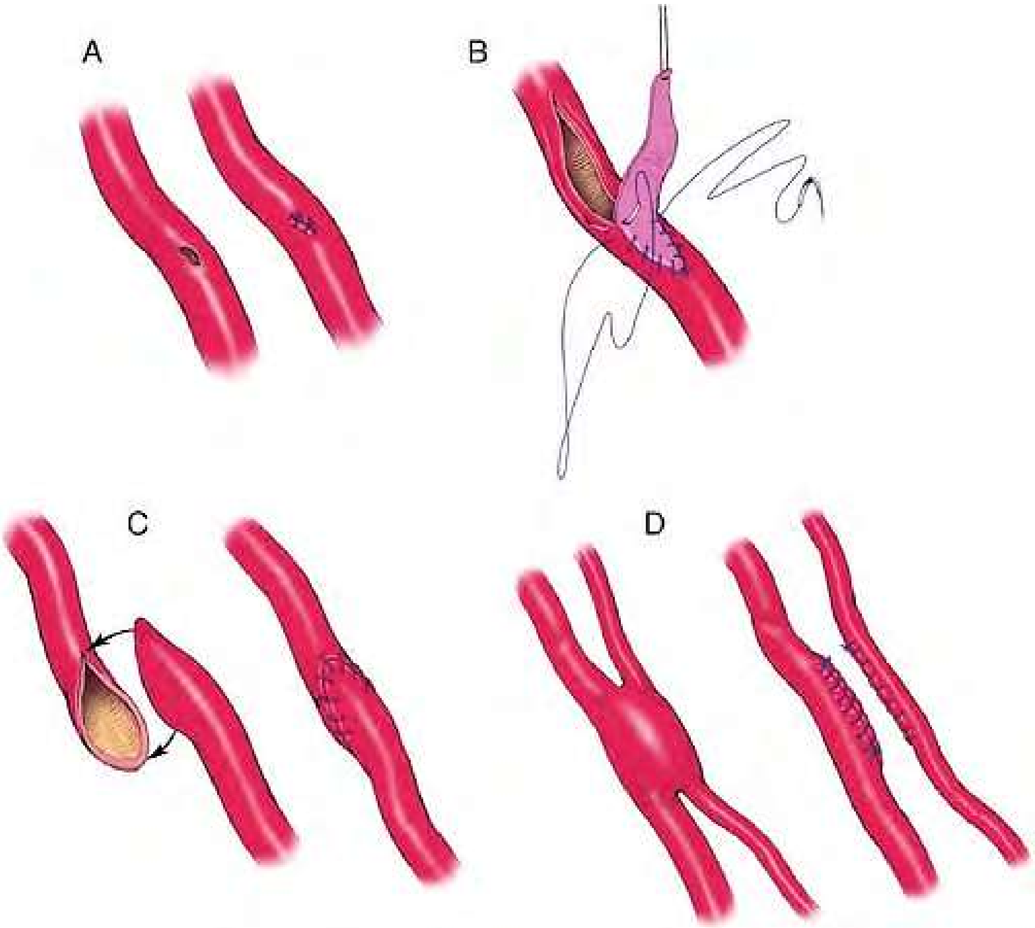

FIGURE 6-3 Types of vascular injuries and repair: simple laceration treated with arteriorrhaphy (A); use of patch angioplasty for longer lesions (B); end-to-end anastomosis (C) and lateral arteriorrhaphy (D).

Insertion of an interposition or bypass graft is required in approximately 20% of cases. Vessels with more extensive injuries of greater than 1.5 cm in length cannot be re-approximated without tension. The preferred conduit in these cases is the great saphenous vein harvested from the contralateral noninjured extremity. The ipsilateral saphenous vein should be avoided due to a high incidence of concomitant deep venous injuries (50%) and postoperative deep venous thrombosis. The great saphenous vein is ideal for vessels 6 mm or smaller. Some size mismatch is acceptable. However, for injuries of larger vessels, a larger conduit should be selected. Spiral or panel vein grafts, or femoral vein grafts, are acceptable options. For arterial injuries of the aorta and major branches, polytetrafluoroethylene (PTFE) or dacron grafts have been used extensively with excellent results. Obviously, contamination of the surgical field represents a limiting factor in these cases. Every attempt should be made to repair both the arterial and venous injury. Although venous repair has limited patency, improved outflow decreases edema and may affect patency of the arterial repair. The traditional recommendation is that all venous injuries involving the popliteal, common femoral, axillary, and portal veins should be repaired because of poor collateral flow. This recommendation has now been extended to all major named veins.

After the vascular reconstruction is complete, the surgeon should assess the adequacy of the repair for any residual stenosis or kinks. The limb should be examined and distal perfusion documented with pulse and hand-held Doppler examination. The goal in a patient without pre-existing vascular disease is to obtain normal distal pulses at the end of the operation. Completion arteriography is important to detect any technical abnormalities that may cause early thrombosis of the repair, as well as to assess the patency of the distal run-off vessels. Finally, the vascular reconstruction should be covered with viable soft tissue. Occasionally, flaps are required to bring vascularized tissue over the reconstruction.

Specific Injuries: Anatomy and Surgical Exposure

Penetrating injuries can affect any arterial segment. Specific patterns of blunt orthopaedic and arterial trauma are well recognized (Table 6-5). The surgical anatomy and exposure of the most common arterial injuries are outlined in the following paragraphs.

Upper Extremity Injuries

Subclavian Artery

The right subclavian artery originates from the innominate artery, and the left subclavian from the aorta (Fig. 6-4). The subclavian artery is divided into three parts in relation to the anterior scalene muscle: proximal, middle, and distal. The proximal subclavian artery gives off the vertebral, internal thoracic, and thyrocervical trunk. The middle subclavian artery contains the costocervical trunk and the dorsal scapular artery. The distal subclavian artery has no branches. A supraclavicular incision one fingerbreadth above the clavicle allows excellent exposure of the middle and distal subclavian artery (Fig. 6-5). For exposure of the proximal subclavian artery, median sternotomy is required for right-sided injuries and left anterolateral thoracotomy for left-sided injuries.

Subclavian artery injuries are uncommon and represent less than 5% of all vascular injuries. Blunt injuries may occur in association with clavicular or first rib fracture. Over 50% of patients present with massive bleeding. Subclavian artery occlusion is well tolerated in 85% of patients because of a rich collateral system. Amputation is rarely required. Arterial repair usually requires end-to-end anastomosis or a small interposition graft using autologous vein. Overall, the mortality rate for subclavian artery injuries is 16%.

Table 6-5. Patterns of Combined Orthopaedic and Vascular Trauma

Orthopaedic Injury	Arterial Injury Location
Supracondylar humeral fracture	Brachial artery
Clavicular fracture	Subclavian and axillary artery
Shoulder dislocation	Axillary artery
First rib fracture	Subclavian artery and aorta

Femoral shaft fracture	Superficial femoral artery and above-knee popliteal artery
Posterior knee dislocation	Popliteal artery
Proximal tibiofibular fracture	Popliteal and tibioperoneal arteries

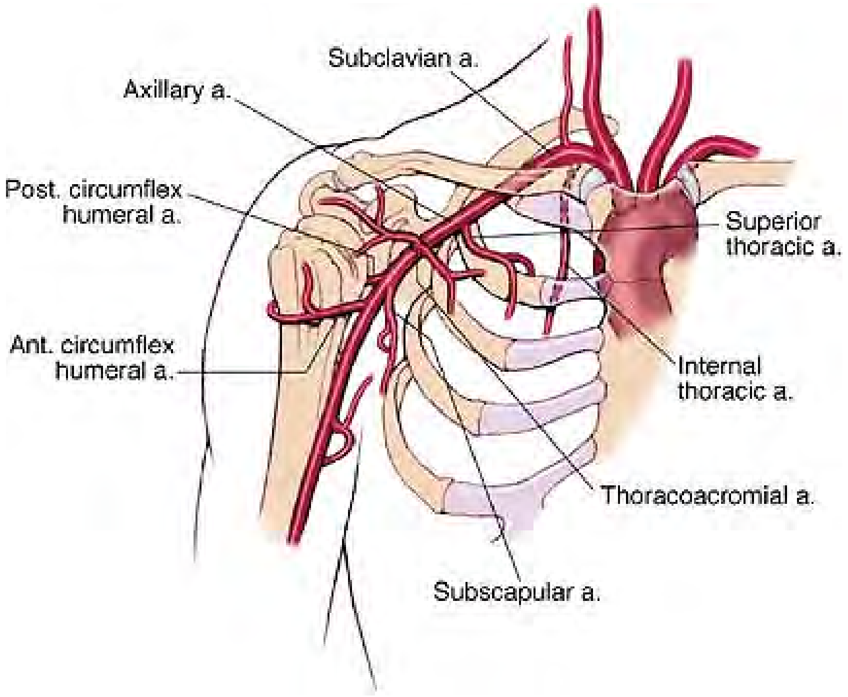
Subclavian a.
Axillary a.
Post. circumflex humeral a.
Superior thoracic a.
Ant. circumflex humeral a.
Internal thoracic a.
Thoracoacromial a.
Subscapular a.

FIGURE 6-4 Surgical anatomy of the subclavian and axillary arteries.

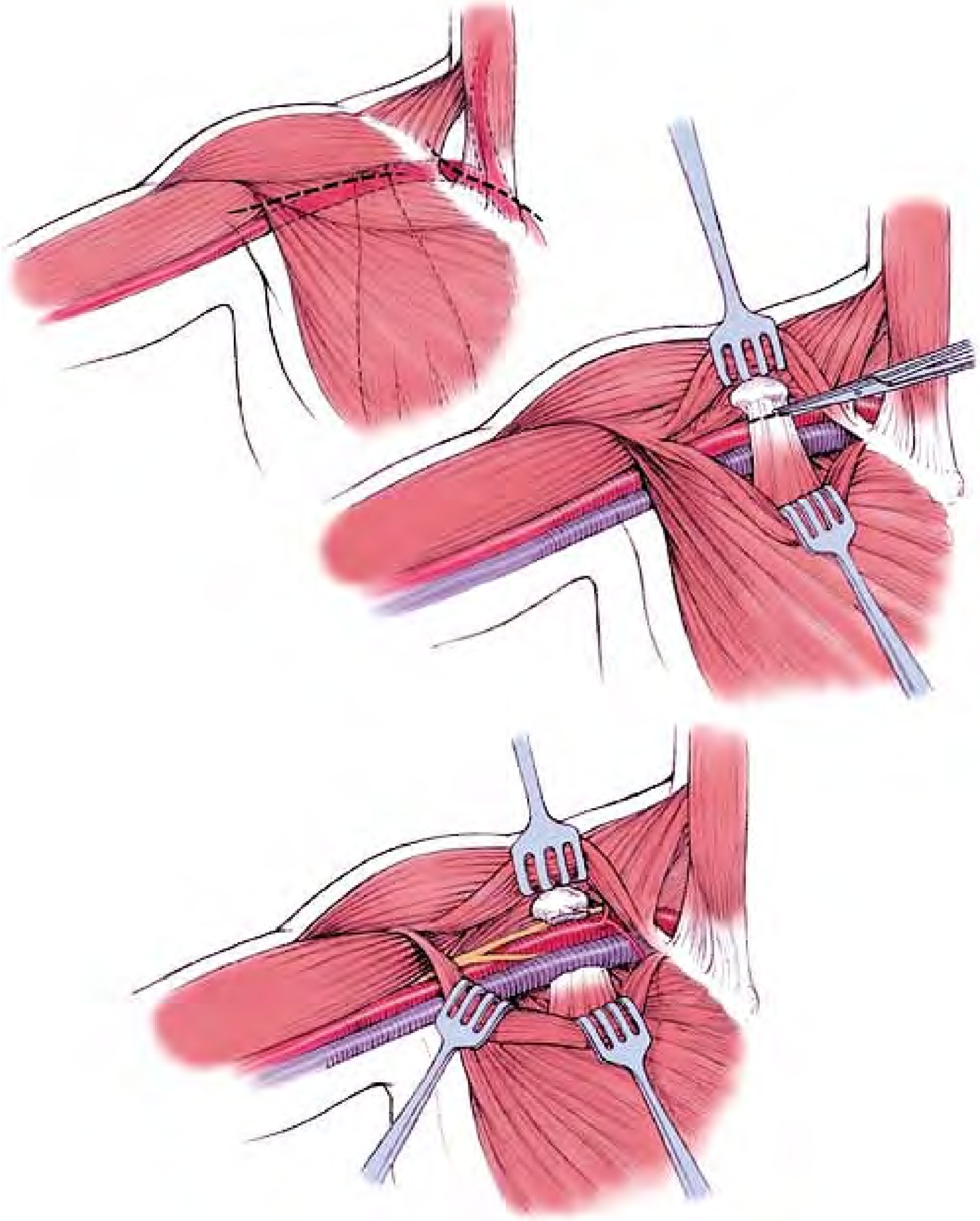

FIGURE 6-5 Preferred surgical approaches for exposure of the subclavian (supracalvicular) and axillary (infraclavicular) arteries.

Axillary Artery

The axillary artery extends from the lateral border of the first rib to the lateral border of the teres major muscle (see Fig. 6-4). The artery is divided into three parts in relation to the pectoralis minor muscle: proximal, middle (beneath), and distal. The first part gives of one branch (supreme thoracic), the second two branches (thoracoacromial and lateral thoracic), and the third three branches (subscapular, anterior and posterior humeral circumflex). The artery lies in close proximity to the axillary vein and to the brachial plexus.

Axillary artery injuries represent 5% to 10% of all arterial injuries. The vast majority (95%) are due to penetrating trauma. Although blunt injuries are rare, these occur in 1% of the patients presenting with either a fracture of the proximal humerus or anterior dislocation of the shoulder. In addition, patients who chronically use crutches may develop stenosis or occlusion from repetitive trauma. Patients with axillary artery injuries often sustain associated nerve trauma.

Exposure of the axillary artery is best achieved using an infraclavicular incision one fingerbreadth bellow the clavicle (see Fig. 6-5). Most injuries can be repaired with end-to-end anastomosis or a small interposition graft. The mortality and amputation rates are exceedingly low for axillary injuries. However, two-thirds of the patients have significant neurologic dysfunction or persistent neuralgia from associated trauma to the brachial plexus.

Brachial Artery

The brachial artery is a continuation of the axillary artery (Fig. 6-6). It starts at the lower edge of the teres major muscle and terminates approximately 2 cm below the antecubital crease, where it bifurcates into the radial and ulnar arteries. A high bifurcation of the brachial artery above the antecubital crease is found in 20% of the population. In the upper and mid-arm, the brachial artery is accompanied by the median nerve laterally and by the ulnar and radial nerves medially. The median nerve crosses anterior to the artery and is located medially at the level of the elbow joint. The brachial artery has three important branches: deep brachial artery, and superior and inferior ulnar collateral arteries.

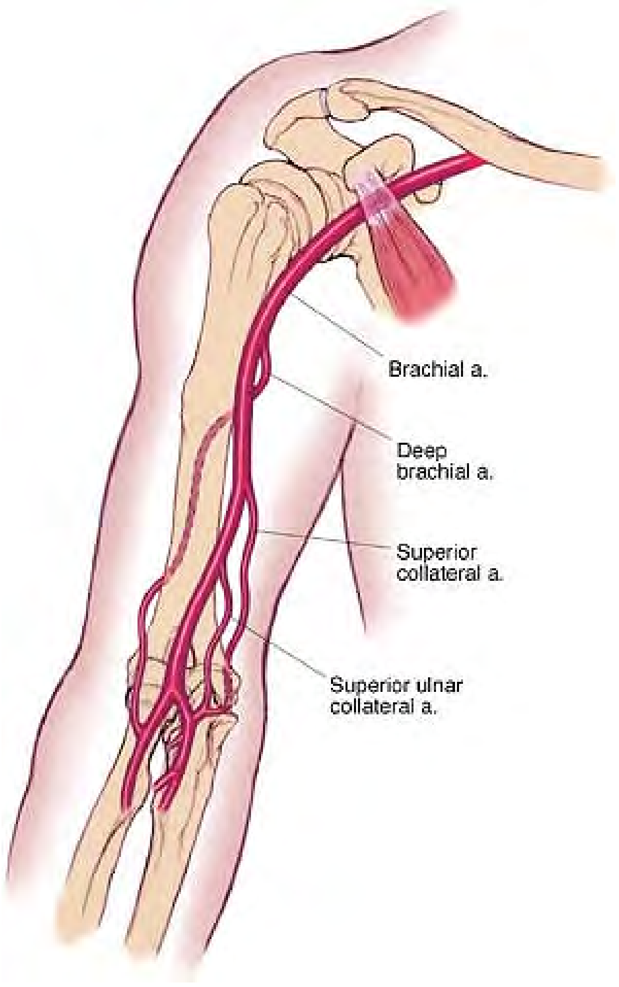

FIGURE 6-6 Surgical anatomy of the brachial artery.

The brachial artery is the most commonly injured artery of the upper extremity, accounting for 15% to 30% of all peripheral vascular injuries. Patients typically present with hand ischemia. Patients with supracondylar humeral fractures or elbow dislocations and hard signs of vascular injury should be further evaluated with arteriography to rule out brachial artery injury. The brachial artery is best exposed using a longitudinal incision along the course of the artery medial to the biceps muscle or an S-shaped incision across the antecubital fossa (Fig. 6-7). Most injuries are repaired with either end-to-end anastomosis or interposition vein graft.

Clinical outcome after repair is better than for injuries of the axillary or subclavian artery because the incidence of nerve injury is significantly less. Amputation or death is rare.

Radial and Ulnar Arteries

The ulnar artery is the larger branch of the brachial artery and gives off the ulnar recurrent arteries and the common interosseus artery (Fig. 6-8). The ulnar artery terminates in the superficial palmar arch. The radial artery gives off the radial recurrent artery, which anastomoses to the deep brachial artery. Distally, the brachial artery gives off a small branch to the superficial palmar arch and terminates in the deep palmar arch.

Injuries to the radial and ulnar artery are also common, but most often result from penetrating trauma. Complete occlusion or transection of either the radial or ulnar artery will often have no adverse effect on the circulation of the hand because of rich collateral circulation. However, patients with marked hematoma in the forearm may develop compartment syndrome requiring fasciotomy. Distal thrombectomy catheters should be handled gently because these arteries are prone to rupture or intimal injury. Repair often requires simple closure, end-to-end anastomosis, or small interposition grafts. Amputation is rare.

Lower Extremity Injuries

Common, Profunda, and Superficial Femoral Arteries

The common femoral artery originates at the level of the inguinal ligament as a continuation of the external iliac artery (Fig. 6-9). The common femoral artery is located adjacent to the femoral nerve (laterally) and common femoral vein (medially). The first branch of the common femoral artery is the superficial circumflex iliac artery, which marks the transition from external iliac to common femoral artery just below the inguinal ligament. The second branch is the superficial epigastric artery. Approximately 5 cm distal to the inguinal ligament the common femoral artery bifurcates into the superficial and deep femoral arteries. The deep (profunda) femoral artery is located in a posterolateral position. This artery gives off the medial and lateral femoral circumflex arteries and four to five perforator branches in the thigh. The superficial femoral artery follows its course underneath the sartorius muscle (Hunter

canal) and terminates in the popliteal artery at the level of the adductor hiatus. The artery is adjacent to the saphenous nerve and femoral vein.

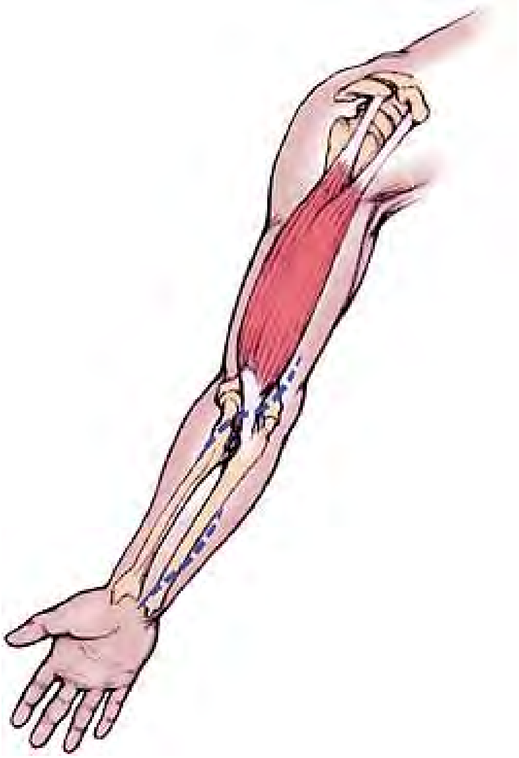

FIGURE 6-7 Preferred surgical approaches for exposure of the distal brachial artery and forearm vessels.

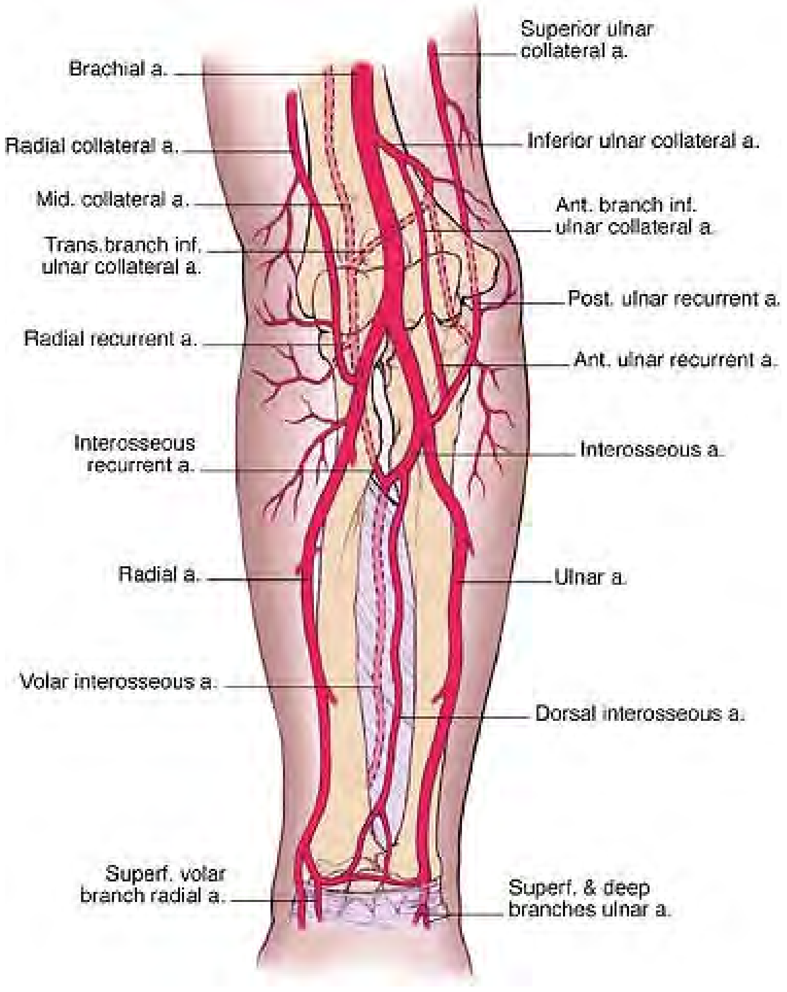
Brachial a.
Superior ulnar collateral a.
Radial collateral a.
Inferior ulnar collateral a.
Mid. collateral a.
Ant. branch inf. ulnar collateral a.
Trans. branch inf. ulnar collateral a.
Post. ulnar recurrent a.
Radial recurrent a.
Ant. ulnar recurrent a.
Interosseous recurrent a.
Interosseous a.
Radial a.
Ulnar a.
Volar interosseous a.
Dorsal interosseous a.
Superf. volar branch radial a.
Superf. & deep branches ulnar a.

FIGURE 6-8 Surgical anatomy of the forearm arteries.

Injury to the femoral vessels accounts for one-third of all peripheral vascular injuries in civilian and military series. The most common mechanism of injury is penetrating trauma. Anterior dislocation of the femoral head is a rare cause of blunt injury. The superficial femoral artery is injured in 5% of the patients presenting with femoral shaft fracture. Bleeding is the most common presentation in cases of penetrating trauma, whereas distal limb ischemia predominates in patients with blunt injury.

Exposure of the common femoral artery and femoral bifurcation is best obtained using a longitudinal incision two fingerbreadths lateral to the pubic tubercle (see Fig. 6-9). The incision extends from the level of the inguinal as far distally as necessary, depending on how extensive the injury is to the proximal deep and superficial femoral arteries. Rarely, a suprainguinal curvilinear incision is required for retroperitoneal exposure of the external iliac artery to achieve proximal control. Exposure of the superficial femoral artery can be achieved through a longitudinal incision along the course of the sartorius muscle. Patients with small, clean injuries of the femoral arteries may be treated with primary closure, end-to-end anastomosis, or patch angioplasty. Longer lesions require an interposition graft using autologous saphenous vein from the contralateral thigh. Long-term patency rates approach 100% for interposition grafts. Lower extremity function is predominantly determined by the extent of skeletal and neurologic injury.

Popliteal and Tibioperoneal Arteries

The popliteal artery is the continuation of the superficial femoral artery and originates at the adductor magnus hiatus (Fig. 6-10). The popliteal artery gives off multiple genicular collateral branches at the above- and below-knee level. The artery bifurcates in 90% of individuals and gives off the anterior tibial artery and the tibioperoneal trunk. The tibioperoneal trunk gives off the posterior tibial artery and the peroneal artery. The anterior tibial artery follows an anterior and lateral course, perforates the interosseous membrane, and is located in the anterior compartment of the leg along with the deep peroneal nerve. The peroneal artery is the middle branch of the three, follows its course in the deep posterior compartment, and bifurcates into the perforating branch to the anterior artery and communicating artery to the posterior tibial artery.

The posterior tibial artery is the most medial branch of the three tibioperoneal arteries and terminates as the common plantar artery below the ankle joint.

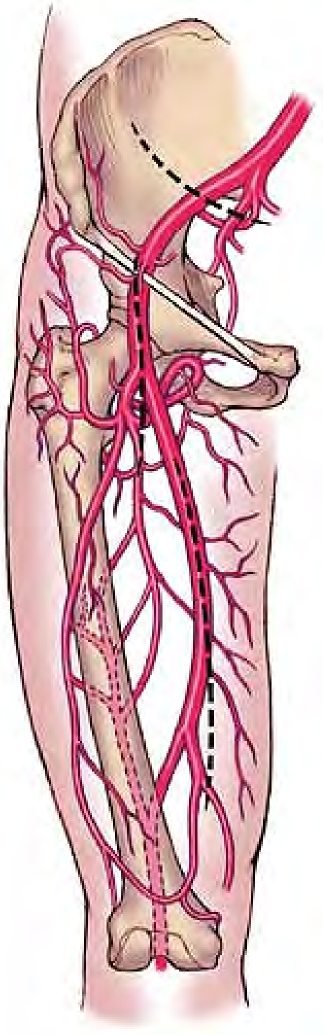

FIGURE 6-9 Surgical anatomy and preferred surgical approaches for exposure of the common, superficial, and deep femoral arteries.

Popliteal artery injuries often result from blunt trauma. Fracture or posterior dislocation of the knee is a known mechanism of injury. Most patients with popliteal artery injury present with distal ischemia because the genicular collaterals are not effective in maintaining adequate distal perfusion. Occlusion of a single tibioperoneal artery is well tolerated and does not require repair.

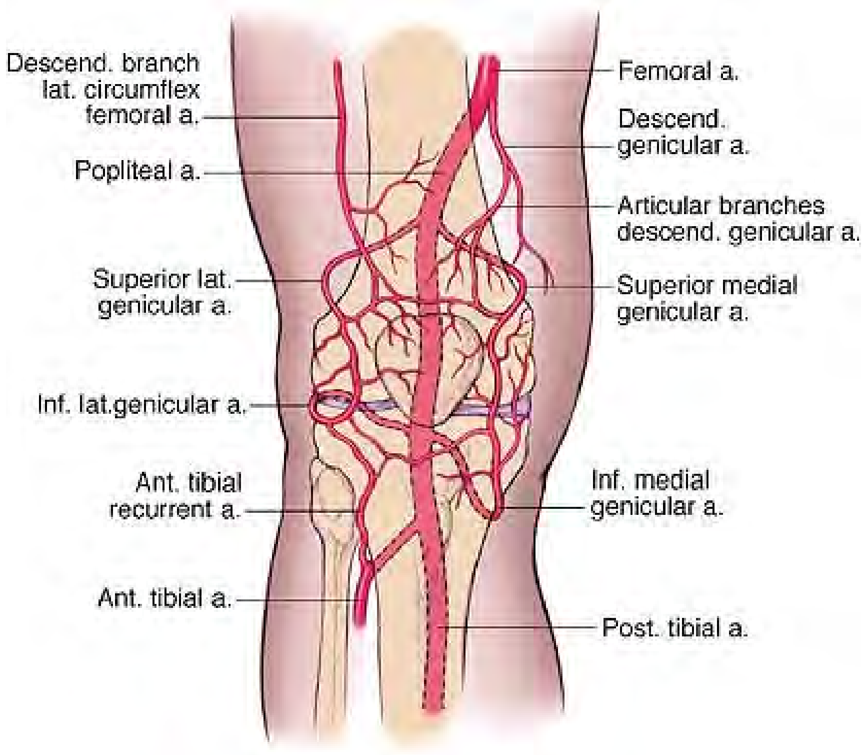
Descend. branch lat. circumflex femoral a.
Popliteal a.
Superior lat. genicular a.
Inf. lat.genicular a.
Ant. tibial recurrent a.
Ant. tibial a.
Femoral a.
Descend. genicular a.
Articular branches descend. genicular a.
Superior medial genicular a.
Inf. medial genicular a.
Post. tibial a.

FIGURE 6-10 Surgical anatomy of the popliteal artery.

Most popliteal and tibioperoneal artery injuries can be exposed using a generous medial approach above and/or below the knee level depending on the extent of injury (Fig. 6-11). The great saphenous vein should be identified and protected. Division of the semimembranous and semitendinous muscles is often required for adequate exposure. The posterior tibial and peroneal arteries are also exposed using a medial incision along the posterior margin of the tibia. The soleus muscle is incised longitudinally allowing exposure of the mid- and distal portions of the posterior tibial and peroneal arteries. Although the origin of the anterior tibial artery can be exposed through a medial incision, an incision in the anterior compartment two fingerbreadths lateral to the tibia is required for exposure of the anterior tibial artery.

Injury to the popliteal and tibioperoneal arteries is associated with significant morbidity. Clinical outcome is affected by the extent of skeletal and neurologic trauma, time of ischemia, and adequacy of runoff through the tibial and peroneal arteries.

Venous Injuries

The management of venous injuries is essentially identical to that for any arterial injury. Although venous repair does not yield the same satisfactory results of arterial repair and there is a higher incidence of early thrombosis, current recommendations are to attempt venous repair of any named major vein whenever possible. Large veins such as the inferior vena cava and iliac veins pose a problem in terms of conduit size mismatch. Options in these cases are use of spiral or paneled vein grafts or externally supported PTFE grafts. Injuries affecting minor veins or patients with other life-threatening injuries should be managed with primary ligation. Major veins can be exposed using the same incisions as for their arterial counterpart. Direct pressure using sponge sticks or pressure cuff tourniquets are excellent means of obtaining control in cases of bleeding. Techniques for repair should be the same as for arterial injuries with the caveat that catheter embolectomy often cannot be used because of competent valves.

Mangled Extremity

One should use clinical judgment in cases of severe trauma with complex soft tissue, skeletal, and vascular injuries. Although there are scoring systems to quantify the degree of injury, cases should be individualized. The mangled extremity severity score (MESS) grades skeletal/soft tissue damage, limb ischemia, shock, and age. However, this scoring system is inaccurate in identifying the irretrievable limb requiring primary amputation. Factors associated with need for amputation after below-knee fractures are injury of greater than three muscle compartments, occlusion of greater than two tibial arteries, and presence of cadaveric changes at initial presentation. When more than two of these predictive factors are present, none of the extremities were salvageable. Primary amputation should be considered early in patients with a dysvascular extremity and extensive soft tissue loss, as the chance of functional recovery is minimal.

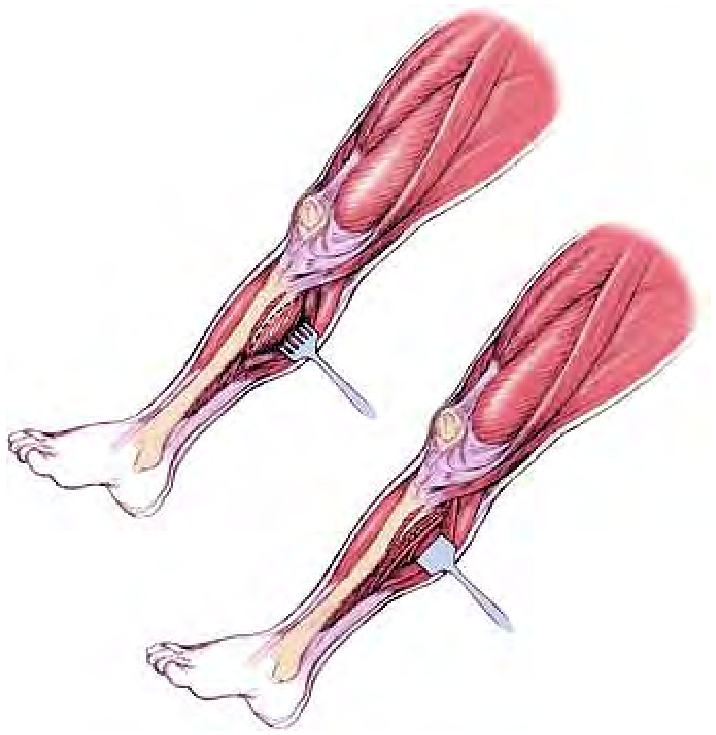

FIGURE 6-11 Preferred medial knee approach for exposure of the below-knee popliteal artery and proximal tibioperoneal vessels.

Postoperative Management

Patients should be continuously evaluated for bleeding, distal limb perfusion, and development of compartment syndrome. Following massive bleeding and coagulopathy, one should aggressively resuscitate the patient and monitor laboratory results every 4 hours (complete blood count, platelets, prothrombin and partial thromboplastin times, and fibrinogen levels). We recommend that laboratory levels are kept within normal limits during the first 48 hours. Parameters include hemoglobin above 10 g/dL, platelet count above 100,000, international normalized ratio of less than 1.2, and partial thromboplastin time of less than 40 seconds. Mild abnormalities contribute to postoperative â€œoozing,â€• large hematoma, and initiate a cascade of coagulopathy that ultimately may lead to bleeding requiring re-exploration. Assessment of vascular pulses should be done routinely with serial examination and hand-held Doppler interrogation. This is repeated every 2 hours for the first 24 hours and every 4 hours thereafter.

The injured extremity should be checked for any signs of infection. Early wound infection, particularly if associated with soft tissue necrosis, is one of the most important determinants of delayed amputation. Aggressive perioperative antibiotic therapy is recommended.

Compartment syndrome is a common occurrence in patients with combined vascular and orthopaedic trauma. Patients with ischemic limbs for more than 4 hours can be expected to have some degree of compartment syndrome. Presentation can be delayed up to 12 to 24 hours after reperfusion of the ischemic limb. Early or prophylactic fasciotomy in this setting may be associated with improved outcome. If fasciotomy is deferred at the time of arterial reconstruction, physical examination and compartment pressure measurements should be repeated frequently. Compartment pressures of higher than 30 cm H^2O strongly indicate fasciotomy.

Complications

The most common complications after vascular reconstructions are bleeding, infection, early thrombosis and development of compartment syndrome. Medical complications involving cardiac, pulmonary, renal, and neurologic systems are also common. Amputation can result because of early technical failure, venous outflow problems, infection, and fracture instability. Deep venous thrombosis is a common occurrence after combined vascular and orthopaedic trauma. Aggressive prophylaxis should be instituted in the early postoperative period. Postoperative duplex ultrasound is indicated for any sign of worsening limb edema or pain. Patients need to be followed regularly because of the potential for late complications. Arterial stenosis or thrombosis, aneurysmal degeneration of vein grafts, and chronic venous insufficiency after primary venous ligation or thrombosis of venous repair can be identified using duplex ultrasound surveillance. We generally re-evaluate patients with combined injuries 6 months after dismissal or earlier if the patient develops symptoms of chronic limb ischemia (e.g., claudication).

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 7

Management of the Soft Tissue with Shoulder Trauma

Scott F. M. Duncan

John W. Sperling

Indications/Contraindications

Shoulder trauma can result in significant soft tissue loss over some of the bony prominences of the shoulder such as the acromion and clavicle. Areas of the scapula can be exposed as well from trauma about the shoulder. Fortunately, the glenohumeral joint itself is covered with a multilayer muscle configuration. If there are concomitant fractures, nerve injuries, or vessel injuries, one must keep in mind not only the pre-existing soft tissue trauma and how to protect the healing of that, but also how to obtain the best exposure to facilitate fixing the concomitant injuries. Open fractures and neurovascular injuries require emergent intervention, or otherwise the extremity may be compromised. In such circumstances, the rescue of the threatened limb takes priority. However, again, one must keep in mind the potential need for soft tissue coverage in these patients. Fortunately, numerous flap options exist about the shoulder for soft tissue coverage (see Chapter 8).

The addition of locking plates, which are fixed angle devices for fixing proximal humerus fractures, has greatly augmented our armamentarium for treating these fractures. In significant injuries, it is not unreasonable to perform a delayed intervention once the soft tissues have been covered or stabilized. Concomitant injuries such as rotator cuff tears can be addressed months down the line, if needed, once the other injuries and soft tissue envelope have finished healing. When the wound bed is grossly purulent, coverage and otherwise operative repair should be delayed until a clean wound can be established. This being said, it is obviously preferable to have all structures fixed prior to any type of soft tissue coverage to minimize injury to the soft tissue flap upon re-operation or re-exploration of the wound.

Preoperative Planning

A thorough examination of the shoulder and the involved extremity as well as a radiographic study of the shoulder are mandatory to recognize soft tissue and bony injuries. Neurovascular injuries may be easily missed if the more distal aspects of the extremity are not examined. Furthermore, radiographic examination must include axillary lateral views to rule out a glenohumeral dislocation. Active flexion, abduction, internal and external rotation should be tested about the shoulder when possible. Active flexion and extension about the elbow, wrist, and fingers will also help delineate more proximal neurovascular compromise. Sensation and motor function of the extremity should be documented. Radial pulses as well as ulnar pulses need to be documented as well as the time for capillary refill.

From a radiographic standpoint in the emergent setting, a computed tomography scan sometimes is required to further delineate the characteristics of bony trauma. Occasionally, an arteriogram may be needed to investigate vascular injuries. The most commonly injured nerve about the shoulder is the axillary nerve, and an examination of deltoid function and shoulder sensation should be performed. When planning the surgical exposure or exposures, one must appreciate what other injuries need to be addressed, if any, and how the incision and exposure will jeopardize any pre-existing skin or muscle trauma. Incisions should be selected that will not compromise the viability of the soft tissue envelope and that once they have healed will not create a tethering-type scar, resulting in loss of shoulder range of motion.

Surgery

Patient Positioning

We prefer to perform surgery about the shoulder with the patient in a beach-chair position. Depending on the injuries to be addressed, supine and lateral decubitus may be used as well. However, in our experience, the beach-chair position has the advantages of allowing manipulation of the arm as needed to facilitate reduction or retraction as well as providing access to both the anterior and posterior aspects of the shoulder. These surgeries are usually done under general anesthesia, but postoperative pain management can be enhanced by the administration of interscalene blocks by the anesthesiologist. The patient is brought into the regular operating room where preparation and draping are carried out in the usual sterile fashion and manner. As with any surgery, atraumatic technique should be adhered to in an effort to lessen skin, muscle, and other soft tissue necrosis. Poorly executed shoulder surgery can result in greater functional loss than would have otherwise occurred had no surgical intervention been attempted. However, as with any surgery, risks are involved and nerve and vessel injury can occur. Excessive scarring and postoperative joint contractures may result despite the best surgical techniques.

The beach-chair position facilitates the use of fluoroscopy during the course of the procedure (Fig. 7-1). This flexibility is sometimes needed when attempting to image complex injuries. Sometimes it is necessary to extend the wound of the injury both proximally and distally to provide visibility to the area of trauma. However, most important is that an adequate view of the surgical field be created to avoid the need to perform surgical repairs in a challenged manner through a small wound. In general, we try to avoid â€œTâ€• extensions of transverse lacerations, but occasionally this may need to be done to gain access to neurovascular structures. Again, the type of shoulder injury will necessitate the surgical approach.

Technique

For proximal humerus fractures, there are two basic approaches. The first is a superior deltoid approach in which a skin incision is made in Langer's lines just lateral to the anterior lateral aspect of the acromion. This approach allows the deltoid to be split from the edge of the acromion distally for approximately 4 to 5 cm. Again, care must be taken to protect the axillary nerve. Of note is that the deltoid origin is not removed, but there is still exposure of the superior aspect of the proximal humerus.

This type of exposure is quite useful for a fixation of greater tuberosity fractures as well as the insertion of intramedullary nailing. In the beach-chair position rotation, flexion, and extension

of the humerus can enhance the exposure of the underlying structures as well as reduction of the fragments.

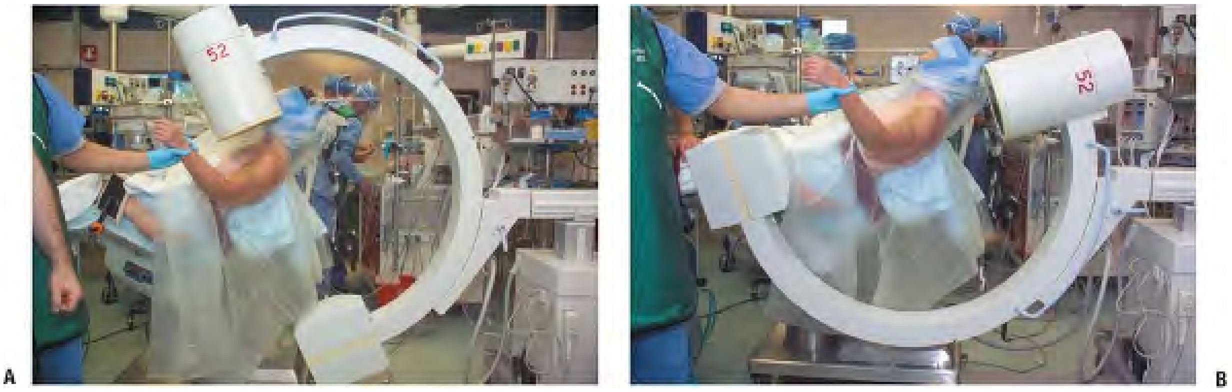

FIGURE 7-1 A,B: The beach-chair position helps facilitate use of intraoperative fluoroscopy.

The second approach is a long deltopectoral approach in which both the deltoid origin and insertion are preserved (Fig. 7-2). The skin incision begins just inferior to the clavicle and extends across the coracoid process and down into the area of insertion of the deltoid on the humerus. The cephalic vein should be preserved and retracted either laterally or medially depending on which is easiest for the surgeon. The cephalic vein is less likely to be injured when taken medially, given that over-vigorous retraction on the cephalic vein and deltoids can result in vein injury. If the vein is accidentally transected or significantly injured, either from the trauma or the surgical approach, ligation can be considered. In the deltopectoral approach, if more exposure is needed to gain access to the proximal aspect of the humerus, the superior part of the pectoralis major tendon insertion can be divided. This may need to be done for multipart fractures. In significant trauma injuries, the pectoralis major tendon may be found to be disrupted and should be repaired when possible using nonabsorbable suture and/or suture anchors.

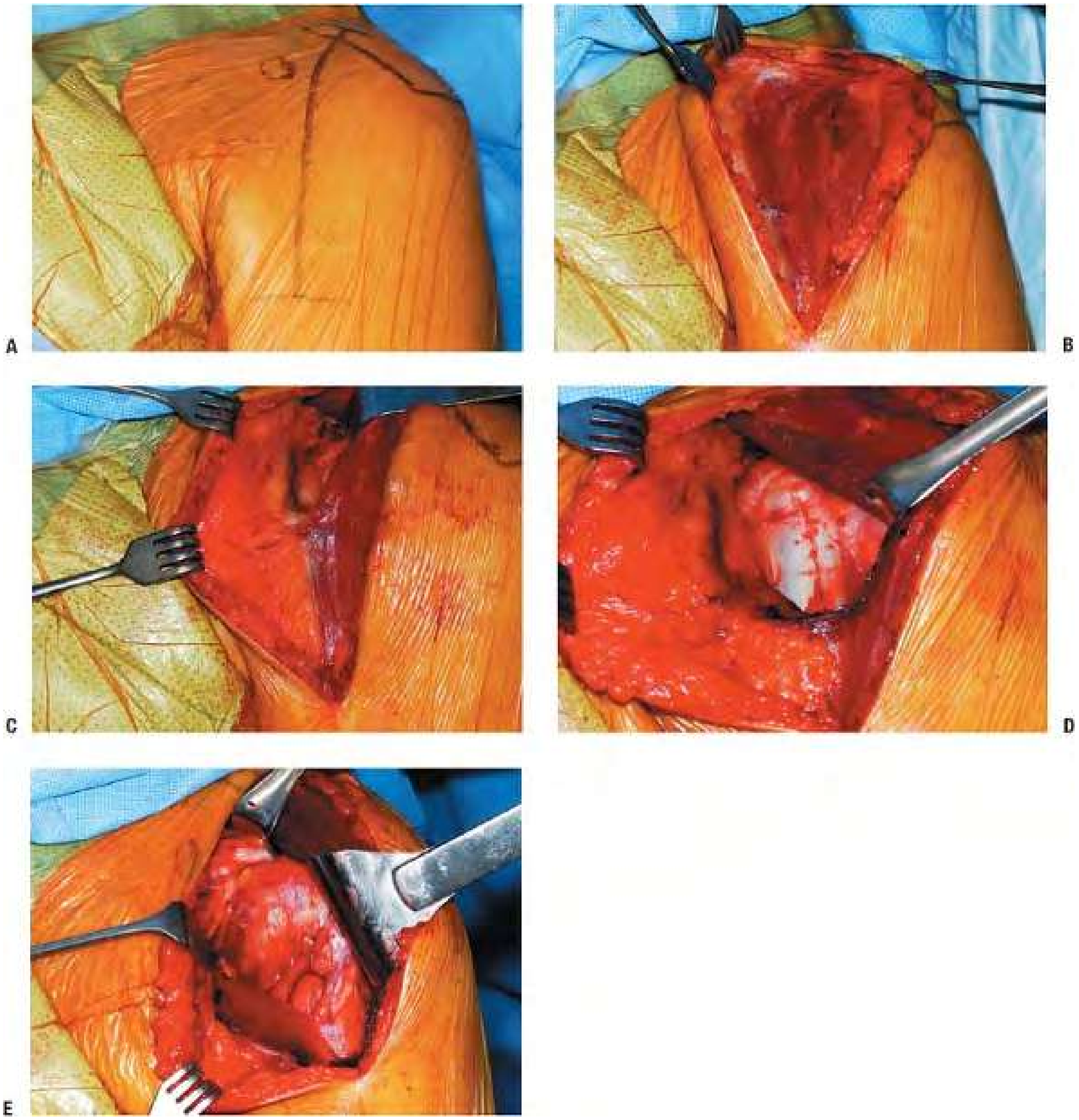

FIGURE 7-2 A: The landmarks of the shoulder are outlined and the incision is marked. B: A triangle of fat is usually present at the proximal interval between the pectoralis and deltoid. C: Typically, the cephalic vein is left within its bed

medially. D: One continues the deltopectoral interval distally. E: One places a retractor medially beneath the conjoint group and one laterally to retract the deltoid.

We do not advocate procedures that split or remove part of the acromion given the significant risk of complications. Methods of internal fixation of proximal humerus fractures go beyond the scope of what this chapter is intended to provide. However, a brief mention of various options will be discussed. As mentioned earlier in the chapter, there is an improved role for open reduction internal fixation of these fractures given the new fixed angled locking plates that have been developed. However, there still remains a role for intramedullary nailing, tension band techniques, and percutaneous pinning of proximal humerus fractures. Most important, though, is that a majority of proximal humerus fractures in our practices are still treated nonoperatively. Shoulder arthroplasty is still used in those cases where there is significant comminution and adequate fixation cannot be achieved by other means. Successful treatment of shoulder injuries requires understanding of their specific biologic and mechanical challenges. This should also encompass knowledge about the natural history and potential complications from the various treatment options.

With the beach-chair position, we take care to secure the head with use of towels to the head supporting apparatus. This can be done with prefabricated devices such as the Skytron attachment. Care needs to be taken to make sure the neck angle is appropriate. Also, when securing the head, care needs to be taken that there are no pressure points on the eyes. With regard to the anterior approach, the exposure of the shoulder joint is useful for drainage of sepsis, open reduction internal fixation of fractures, hemiarthroplasty, and reconstruction of dislocation trauma. Landmarks for this approach are the coracoid process and the deltoid insertion. The incision starts at about the level of the clavicle and goes over the coracoid process through the area where the deltopectoral groove can be palpated down to the anterior aspect of the shoulder. An alternative to this incision is an axillary incision. The shoulder must be abducted and externally rotated to make this type of approach. The downside of this in the trauma setting is that the skin flaps must be undermined extensively to visualize the area of the deltopectoral groove. However, this approach does have a cosmetic advantage. This approach is also difficult in extremely muscular patients.

The deep surgical dissection involves identifying the short head of the biceps and coracobrachialis; these are retracted medially. The overlying fascia must be carefully incised. The long head of the biceps can be used to identify the division between the lesser and greater tuberosities. This can also be a good marker for where to begin subscapularis dissection. The arm should be kept adducted at most times, because bringing the arm into abduction brings the neurovascular structures closer to the operative field. This is especially true if any work is being done around the coracoid. Care must be taken when retracting the coracoid and coracobrachialis, as over-vigorous retraction can cause musculocutaneous paralysis. Care should be taken with the axillary nerve as well. External rotation of the shoulder increases the distance between the subscapularis and the axillary nerve. A leash of vessels on the inferior border of the subscapularis frequently requires cauterization to maintain good visualization in the surgical field.

The anterolateral approach in the trauma setting is most useful for repair stabilization of the long head of the biceps tendon as well as fixation of greater tuberosity fractures. A transverse or longitudinal incision can be used beginning at the anterolateral corner of the acromion and extending lateral to the coracoid process or going over the anterolateral aspect of the shoulder. The superficial dissection involves subcutaneous fat and deltoid fascia. This is incised in line with the skin incision. The split should not extend more than 5 cm down from the acromion as axillary nerve injury is possible. In the case of rotator cuff tears, one may consider taking the deltoid off the anterior aspect of the acromion, which also improves visualization (Fig. 7-3). Avoid inadvertent inferior retraction as well, which can also injure the axillary nerve. Bleeding will obstruct the view and is frequently encountered during this dissection. Try to obtain good hemostasis and cautery of the acromial branch of the coracoacromial artery; it usually must be electrocauterized. Also avoid excessive stripping and detachment of the deltoid. Deep dissection will reveal the rotator cuff under surface of the acromion and greater tuberosity. A stay suture in the fascia of the deltoid can be useful to prevent excessive splitting and injury to the axillary nerve.

The lateral approach provides limited access in a similar fashion to the anterolateral approach. Again, it can be useful for open reduction internal fixation of greater tuberosity fractures, as well as for open reduction internal fixation of neck fractures using a tension band technique. The rotator cuff is easily accessible through this approach. This is a common approach for insertion of intramedullary rods into the humerus. An incision approximately 5 cm long is made from just above the acromion down the lateral aspect of the arm. Deltoid muscle is split in line with its fibers down to a level no more than 5 cm. Most surgeons insert a suture at the inferior aspect of the split to help prevent it from extending any further and potentially injuring the axillary nerve. Deep surgical dissection is continued through the deltoid. The rotator cuff and its bursa in the subacromial space are immediately encountered. Again, the humeral neck, greater tuberosity, rotator cuff, and even the long head of the biceps, to some degree, can be accessed from this approach. Care needs to be taken to avoid injuring the axillary nerve or detaching an excessive amount of deltoid.

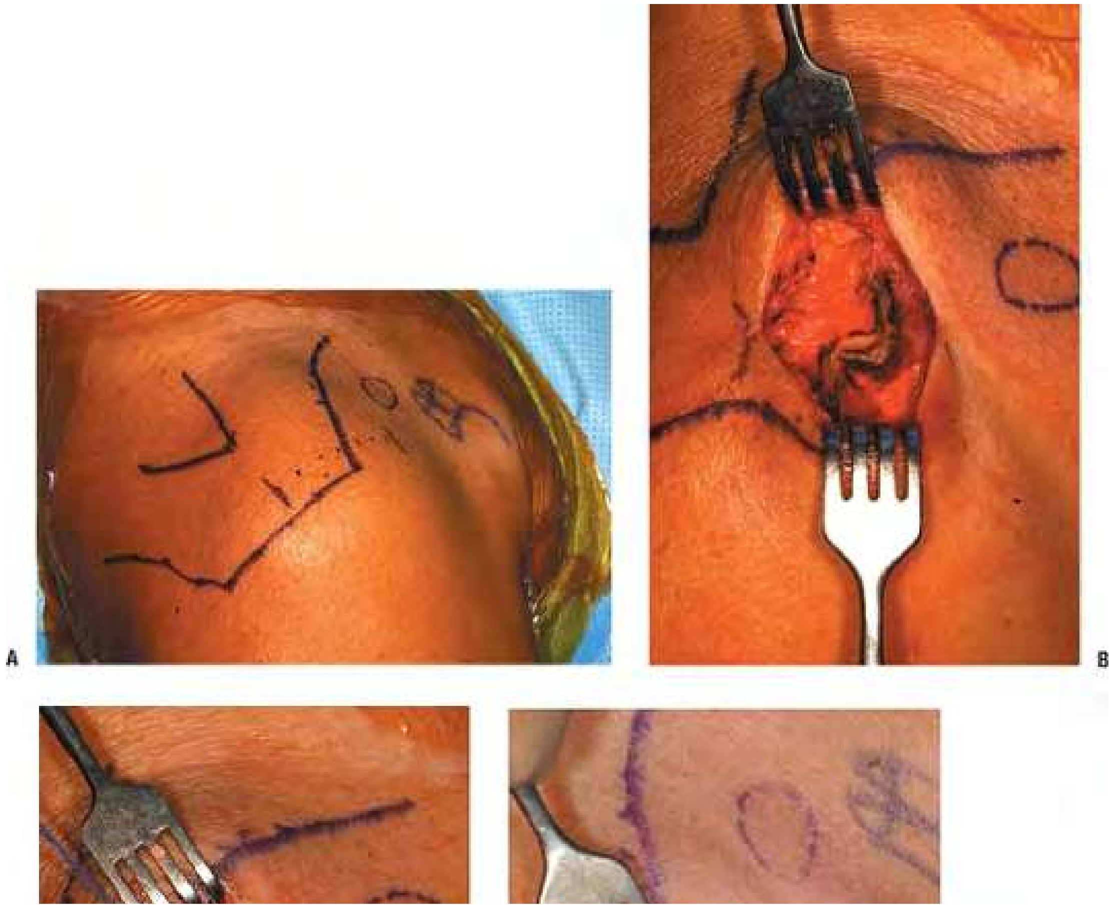

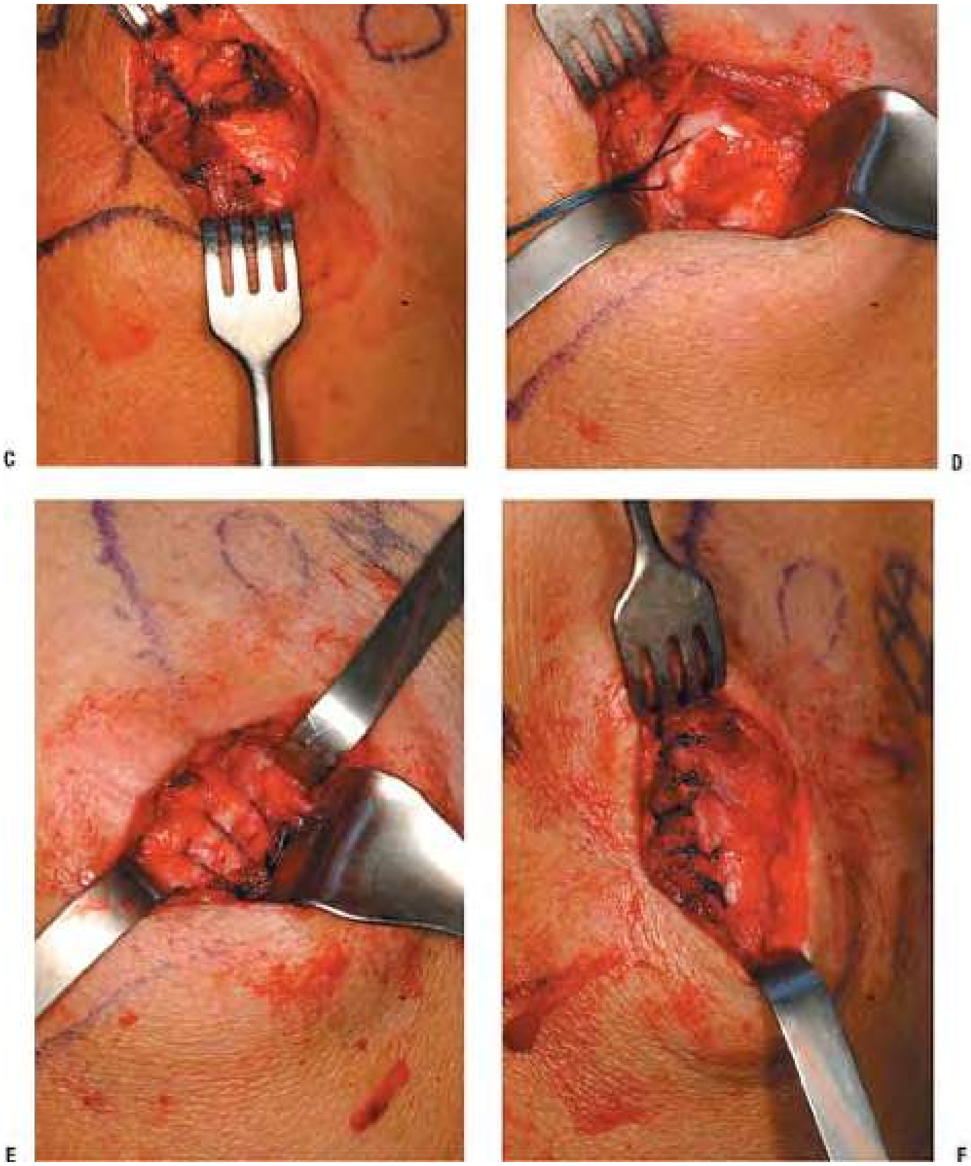

FIGURE 7-3 A: An incision is marked out parallel to the lateral border of the acromion. B: The deltoid is taken off of the acromion as outlined. C: A stitch is placed in the corner of the deltoid to assist in later repair and ensure proper alignment deltoid repair. D: The rotator cuff tear is identified and retention stitches are placed. E,F: A rotator cuff repair is performed.

The posterior approach offers access to the posterior and inferior aspects of the glenohumeral joint. In the trauma scenario, open reduction and internal fixation of the glenoid, scapular body, and sepsis drainage can be provided. Depending on the amount of exposure required, the deltoid can be split rather than detached. The interval between the deltoid muscle and the inner lying infraspinatus muscle is identified next. The internervous interval between the infraspinatus and the teres minor is next identified and is usually developed with blunt finger dissection. Some surgeons prefer to use the interval between the two heads of the infraspinatus. By retracting each of these muscles, the inferior and posterior aspects of the glenoid can be reached.

If the joint needs to be explored, incise it longitudinally close to the scapular edge. Care needs to be taken to protect the axillary nerve as it runs in the quadrangular space beneath the teres minor. The other nerve that can potentially be injured is the suprascapular nerve as it passes around the base of the spine of the scapula and runs from the supraspinatus fossa to the infraspinatus fossa. Damage to this can cause loss of innervation to the supraspinatus and infraspinatus muscles. Excessive retraction of the infraspinatus medially can result in neurapraxia. Finally, the posterior circumflex humeral artery runs with the axillary nerve in the quadrangular space beneath the inferior border of the teres minor; and again, if the dissection is performed too low, the structure can be injured.

Postoperative Management

The shoulder is a complex joint and will rapidly become stiff. Ideally, adequate motion is needed for optimum function of the extremity for the patient. Rehabilitation depends on type of injury, stability of repairs, and amount of soft tissue injury. In most cases, the three-phase system developed by Hughes and Neer will suffice. The first phase incorporates passive-assistive exercises. Active and early resistive exercises are started in the second phase. Finally, the third phase is a program dedicated to stretching and strengthening exercises. In general, these exercises are performed three to

four times per day for 20 to 30 minutes. In the first phase, it is not uncommon for pain medication to be used. Working with a skilled physical therapist can improve patient compliance and outcome.

Phase one is started in the early posttraumatic period. If there is secure fixation of the underlying injuries, passive exercises may be begun 24 to 72 hours after the index procedure. The surgeon must be careful not to forget about the elbow, wrist, and fingers. If these are neglected, permanent stiffness can result in these joints, and the extremity is at risk for complex regional pain syndrome (hand-shoulder syndrome). The first exercise usually consists of the pendulum or Codman's type in which the arm is rotated both outwardly and inwardly in small circles. Next, the supine external rotation with a stick exercise is used. The elbow and distal humerus must be supported with either a towel or sheets. A 20-degree shoulder abduction can facilitate performing this exercise. Gentle pulley exercises can be added to the regimen as well as extension. In general, isometrics are not started until there is evidence of some fracture healing.

Phase two exercises involve active, gentle resistive, and stretching programs. Supine active forward elevation is the first one because the effects of gravity are reduced by making the elevation easier. The patient then can be progressed to doing this exercise in a standing or sitting position. A stick in the unaffected arm can assist the involved arm with this maneuver. TheraBands can be used to strengthen the internal and external rotators as well as the anterior, middle, and posterior aspects of the deltoid. These are usually done in repetitions of 10 to 15 per session. Wall climbing exercises with stretching are also performed. Internal rotation can be one of the most difficult aspects of motion to regain in the shoulder, and the patient can work on assisting the affected extremity with the unaffected one by trying to have the hand crawl up the back of the patient.

Phase three exercises are started after solid evidence of fracture healing. The rubber strips may be replaced by rubber tubing to increase resistance. Continued aggressive stretching is performed. Light weights are usually tolerated at this time. Weights start at 1 lb and are increased in 1-lb increments, usually stopping at 5 lb. If there is significant pain with the weight, then these should be discontinued as strength can generally recover with functional activity.

Complications

Postoperative Complications

Postoperative complications, as previously mentioned, include neurovascular injury and soft tissue loss. In the traumatized shoulder, the usual anatomic landmarks may be distorted or absent. This makes it imperative that surgeons take their time and appropriately identify neurovascular structures and internervous planes when possible. Other factors such as hardware loosening and infection can occur. In the event of excessive soft tissue loss, flap coverage may be required (see Chapter 8). Free flaps such as a latissimus or anterolateral thigh flap can be considered depending on the need for muscle versus subcutaneous and cutaneous layers. With pinning techniques, pin migration can be a problem and has been reported in the literature.

In general, complications can be broken down into vascular injury, brachial plexus injury, chest injury, myositis ossificans, frozen shoulder, avascular necrosis, nonunion and malunion, and complex regional pain syndrome. Fortunately, vascular injuries are infrequent, but they can occur and can result in loss of the extremity. Injury of the axillary artery represents about 6% of all arterial traumas and usually occurs secondary to fractures of the proximal humerus. It can be seen in older individuals with significant atherosclerosis as the vessel may tear secondary to dislocation or blunt trauma. The most common site of injury to the axillary artery is proximal to the take-off of the anterior circumflex artery. It is important to check the radial pulse, but this is no guarantee that arterial injury has not occurred. Other signs to look for include an expanding hematoma, pallor, and paresthesias. Missed arterial injuries can result in gas gangrene, amputation, and compressive neuropathies at the brachial plexus leading to permanent nerve palsies. If three is any suspicion, angiography should be performed immediately.

Brachial Plexus Injuries

Brachial plexus injuries can occur after proximal humerus fractures. This has been reported to have an incidence of 6.1%. Isolated injury to the axillary nerve is not uncommon and has been reported, but any or all components of the brachial plexus can be involved. It is, thus, extremely important to evaluate the patient initially and document any compromise in skin sensation and motor function. If a nerve injury is clinically suspected, it should be carefully followed, and electromyography and

nerve conduction velocity studies should be used to monitor for progression or failure of healing. Nerve injuries that fail to show improvement clinically and electrically may need to be explored.

Chest Injury to the Thoracic Cavity

Chest injury to the thoracic cavity can occur after fractures and trauma to the shoulder. There have been reports of intrathoracic dislocation of the head with surgical neck fractures of the humerus, as well as pneumothorax and hemathorax associated with the trauma that resulted in the proximal humerus fracture.

Frozen Shoulder

Frozen shoulder is extremely common after any type of injury or surgery to the shoulder. Even with implementation of well-organized and carefully monitored physical therapy, adhesive capsulitis can develop. Fortunately, most cases of adhesive capsulitis respond to programs of exercise and stretching. In those that fail to improve after 4 to 6 months, manipulation and lysis of adhesions can be considered. There is a risk of refracture, however. Painful and impinging hardware can also result in shoulder motion limitations; if the patient desires, once fracture union is obtained, the hardware may be removed.

Myositis Ossificans

Myositis ossificans is known to occur after fracture dislocations. It can also occur in head-injured patients. It is more commonly seen when there is a chronic unreduced fracture dislocation. Avascular necrosis of the humeral head can occur with shoulder trauma. This is more common in three- and four-part fractures, but can occur with any type of injury about the shoulder, including soft tissue injuries. Excessive periosteal stripping and excessive soft tissue dissection may also result in this complication. Nonunions and malunions can occur with any type of fracture and these may result from fractures about the shoulder. Their treatment goes beyond the purpose of this chapter.

Complex Regional Pain Syndrome

This entity, formerly known as reflex sympathetic dystrophy and before that as hand-shoulder syndrome, is cited to affect from 10% to 30% of patients who have shoulder trauma that required surgery. Soft tissue and bony injury can result in this pathologic process marked chronic pain. The patients do not necessarily have to have any type of penetrating or open trauma. The treatment of complex regional pain syndrome is early recognition of the problem, sympathetic blocks, Neurontin, and aggressive therapy. Complex regional pain syndrome can occur in any setting with light or aggressive therapy and minimal or extensive trauma. Once the condition is recognized, treatment should be initiated. If there is any question, quantitative sudomotor axon reflex testing may be helpful.

Recommended Reading

Blom S, Dahlback LO. Nerve injuries in dislocations of the shoulder joint and fractures of the neck of the humerus: a clinical and electromyographic study. *Acta Chir Scand*. 1970;136:461â€“466.

Hughes M, Neer CS II. Glenohumeral joint replacement and postoperative rehabilitation. *Phys Ther.* 1975;55:850â€“858.

Linson MA. Axillary artery thrombosis after fracture of the humerus. *J Bone Joint Surg*. 1980;62A:1214â€“1215.

Neer CS II. Displaced proximal humeral fractures. Part I. Classification and evaluation. *J Bone Joint Surg.* 1970;52A:1077â€“1089.

Neer CS II. Displaced proximal humeral fractures: Part 2. Treatment of three-part and four-part displacement. *J Bone Joint Surg*. 1970;52A:1090â€“1103.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 8

Rotational Flaps for Rotator Cuff Repair

Emilie V. Cheung

Scott P. Steinmann

Indications/ Contraindications

Massive soft tissue tears of the rotator cuff pose significant challenges to the shoulder surgeon. Although most tears can be successfully repaired in a primary fashion, some massive tears of the rotator cuff are termed irreparable, either due to their severely contracted state or simply because of their size. Cofield has defined the term massive rotator cuff tear as greater than 5 cm, whereas others have defined it as involvement of at least two tendons.

Chronic tears, commonly seen in older patients, have evidence of attritional changes in both the tendon substance and the muscle fibers. This has been termed fatty degeneration by Gerber and can be best characterized and graded based on its appearance on T1 weighted magnetic resonance imaging (MRI) of the shoulder in the oblique sagittal plane. The presence of fatty degeneration of the musculotendinous unit may compromise the results of even a seemingly successful repair due to poor tissue quality and limited regenerative capacity. The limitations of successful tendon repair are thus a function of the length of time since injury, the degree of tissue contraction, and extent of fatty degeneration.

Management of such tears often begins with conservative management. If this fails, surgical management may include debridement of the cuff tear, which may result in satisfactory pain relief, but will not restore strength or function. Partial repair may be possible in some instances, but this does not restore normal function. Other options include allograft or synthetic tendon augmentation, but these have met with varying success in the literature.

Long-standing rotator cuff tears result in altered kinematics of the glenohumeral joint, and rotator cuff tear arthropathy may eventually develop as a painful end-stage condition. Hemiarthroplasty can be a reasonable option in patients who have an intact coracoacromial arch limiting anterior superior escape of the humeral head. The reverse shoulder prosthesis has recently been shown to be an effective treatment option for patients with irreparable rotator cuff tears and anterior superior instability of the humeral head. At mid-term follow-up, improvements in range of motion and pain relief have been reported in the majority of cases. However, this is generally reserved for select indications in an elderly patient. The long-term results of this prosthesis have yet to be determined.

A patient who has had a previously failed massive rotator cuff repair may be challenging to treat. Tendon transfers are another option for salvaging massive rotator cuff defects. They have the distinct advantage in their ability to restore strength, in addition to providing pain relief. Local tendon transpositions historically have included superior subscapularis, teres minor, and teres major transfers. Deltoid flap reconstruction has also been reported. Our preferred method of tendon transfer for posterior superior rotator cuff tears is the latissimus dorsi transfer.

Gerber et al described the procedure of latissimus dorsi transfer as an option for the treatment of massive posterior superior rotator cuff defects in 1988. This procedure has been shown to restore an active external rotation and forward flexion moment at the shoulder, which are the primary functional deficits for this type of massive rotator cuff tear.

Latissimus transfer in cases of a failed previous rotator cuff tear results in overall improvement in function and pain relief, but is generally inferior to that observed after a primary transfer. A comparative analysis on the results of latissimus dorsi transfer for primary irreparable rotator cuff tears, versus for salvage reconstruction for a failed previous rotator cuff repair was performed by Warner and Parsons.

Patientsâ€™ satisfaction and function were superior in primary latissimus dorsi transfer when compared with patients undergoing salvage reconstruction of failed previous rotator cuff repairs.

Subscapularis tendon ruptures occur less frequently than supraspinatus and infraspinatus ruptures. However, they can occur in isolation, during anterior shoulder dislocation, or an as extension of a massive rotator cuff tear. Our preferred method of tendon transfer for irreparable subscapularis tendon tears is the pectoralis major transfer. Combined latissimus dorsi transfer with pectoralis major transfer may be indicated in massive rotator cuff deficiency involving both the posterior superior cuff and the subscapularis.

The primary indication for latissimus dorsi transfer is the loss of active external rotation due to an irreparable tear of the posterior superior rotator cuff. The patient is dissatisfied with increased functional demands, which are beyond the limitations of conservative treatment.

The primary indication for pectoralis major tendon transfer is irreparable rupture of the subscapularis, characterized by retraction of the musculotendinous unit to the glenoid, and fatty infiltration of the muscle. In a patient with posterosuperior rotator cuff tear with extension of the defect into the subscapularis, isolated latissimus transfer is contraindicated unless there is either concomitant repair of the subscapularis, or transfer of the pectoralis major, to compensate for subscapularis dysfunction.

Preoperative Planning

We begin with the history and physical examination, which are essential in diagnosis and assessment of the nature of the rotator cuff tear. Often, the patient complains of pain or dysfunction with an insidious onset. Other times, the patient complains of fatigue during forward elevation or external rotation of the shoulder. No history of trauma can be recalled and the symptoms have been chronic over months or years, which leads us to believe that the etiology is attritional. Thus, we suspect poor tendon quality and the possibility of an irreparable tear.

The physical examination of the shoulder begins by inspection of both shoulders. Deformity associated with anterior superior escape of the humeral head indicates anterosuperior rotator cuff tear, which is a contraindication for latissimus transfer. Spinati muscle atrophy is noted, and often indicates an irreparable rotator cuff tear is present. A defect of the deltoid is a relative contraindication for this procedure. Manual motor testing of shoulder forward elevation, abduction, external rotation, and internal rotation are recorded.

The *hornblower's sign*, or external rotation lag sign (discrepancy between active and passive external rotation of greater than 30 degrees), is important to recognize. A positive sign is highly suggestive of a massive rotator cuff tear, amenable to latissimus dorsi transfer.

Another physical examination finding that is important to recognize is the *lift-off test* (the inability to actively lift the dorsum of the hand off from a resting position on the lower back), or *belly press test* (the inability to actively maintain the elbow anterior to the midline of the trunk as viewed from the side). An inability to perform these maneuvers is highly suggestive of involvement of the subscapularis in the rotator cuff tear. If a subscapularis tear is present and determined irreparable, then a pectoralis major transfer should be performed in conjunction with the latissimus transfer.

Active and passive range of motion in forward elevation, external rotation, and internal rotation are documented. Loss of passive external rotation and forward elevation is a contraindication to this procedure. Active range of motion is also measured. Patients need to be able to achieve active forward elevation to 90 degrees to be candidates for the procedure. If the patient has pseudoparalysis and minimal forward elevation, the procedure will fail. The main goal of the procedure is to promote active external rotation to eliminate the hornblower's sign and enable the patient to reach the top of the head.

It should be noted that some patients with massive cuff tears are amenable to conservative treatment, and eventually may experience minimal pain and satisfactory function, especially in the elderly population. Latissimus dorsi transfer is more commonly performed in those individuals with higher functional demands.

Imaging studies are useful for assessing the patient who is suspected of having a painful, torn rotator cuff, unresponsive to conservative treatment. Our standard radiographic views include a true anteroposterior (AP) view of the shoulder and an axillary view. The true AP view best demonstrates superior migration of the humeral head in the context of massive rotator cuff tears. The acromiohumeral distance (ACHD) is measured on this view. A normal ACHD value is 10.5 mm. An ACHD value of 7 mm or less suggests an irreparable tear of the infraspinatus and is a strong indicator to consider a latissimus dorsi transfer, in the context of the previously mentioned physical examination findings. The true AP and axillary views demonstrate degenerative changes of the glenohumeral joint, which would be a contraindication to tendon transfer.

Magnetic resonance imaging is commonly performed at our institution in patients with suspected large rotator cuff tears in whom we are considering surgical treatment. It is useful for assessing not only the size of the tear, but the degree of muscle atrophy and the degree of fatty degeneration. Advanced fatty degeneration and a massive tear indicate a low likelihood of successful repair, but may be considered for latissimus dorsi transfer with or without pectoralis major transfer. In many patients with massive rotator cuff tears, we are able to repair the rotator cuff tear arthroscopically. Therefore, in almost all patients who are candidates for a latissimus dorsi transfer, we recommend an attempted arthroscopic repair as the initial surgical procedure.

Latissimus Dorsi Transfer

Surgery

The patient is placed in the lateral decubitus positioning with all bony prominences well-padded. The superior approach for rotator cuff repair in Langer's lines is used. Skin flaps are raised. The deltoid is split in line with its fibers or as an alternative if greater exposure is desired; the deltoid can be reflected from the acromion with a bone chip using an osteotome (Fig. 8-1). A subacromial retractor is placed to visualize the rotator cuff defect. Excise the subacromial bursa for optimal visualization. If the tear is determined to be irreparable, proceed with latissimus dorsi tendon transfer.

A second skin incision is placed along the posterior aspect of axilla, along the lateral border of the latissimus (Fig. 8-2). Skin flaps are created. The latissimus and teres major muscles are visualized (Fig. 8-3). Dissecting distally, identify the latissimus and teres major tendons, which are in intimate contact (Fig. 8-4). The latissimus tendon is very thin as you follow its insertion into the axilla (Fig. 8-5). The teres major tendon can be transferred with the latissimus tendon if the subscapularis is intact and if the patient has extreme weakness. However, the teres major is a bulky muscle with poor excursion. It is difficult to pass both muscles under the acromion, and often the teres major will not reach the superior aspect of the humeral head.

The latissimus tendon should be released directly off of the bone to achieve greatest length. This is best done with electrocautery (Fig. 8-6). Palpation of the radial nerve is helpful at this stage to note its location and lessen the chance of nerve injury. Note that the latissimus tendon is tendinous, without muscular fibers at its humeral attachment site. In contrast, the teres major tendon is very short, and the musculotendinous unit is very muscular at its attachment on the humerus. Internal rotation of the arm helps to bring the insertion site into view and allows for easier dissection. However, when the arm is internally rotated, the axillary and radial nerves are brought closer to the latissimus dorsi insertion, and care must be taken with this exposure and arm positioning (Fig. 8-7).

Locking nonabsorbable sutures are placed into the distal end of the latissimus tendon, with four tails left distally for tying at the end. This step may be repeated for the distal end of the teres major tendon. Ensure the muscles are gliding freely with adequate excursion. The neurovascular pedicle of the latissimus enters anteriorly, about 10 to 15 cm from its musculotendinous junction.

Identify the teres minor, triceps, and deltoid. Pass a Kelly clamp from the superior incision, over teres minor, and underneath deltoid, to exit through the inferior incision, by blunt dissection. The sutures are placed into the clamp, and the latissimus tendon is pulled up into the superior wound, over the superior posterior aspect of the humeral head (Fig. 8-8). The teres major can be passed similarly. A rongeur is used to remove residual soft tissue from the posterior superior aspect of the greater tuberosity, which will receive the transferred tendon ends. Pass the sutures through bone with a free curved needle, and tie over a bony bridge or use suture anchors (Fig. 8-9). A strong repair is required since there will be a lot of tension on the repair. It is often helpful to tie the sutures over a small metal or absorbable washer. Suture the border of subscapularis and rotator interval to the transferred tendon. The goal is to have a strong mechanical repair, and not necessarily a watertight repair.

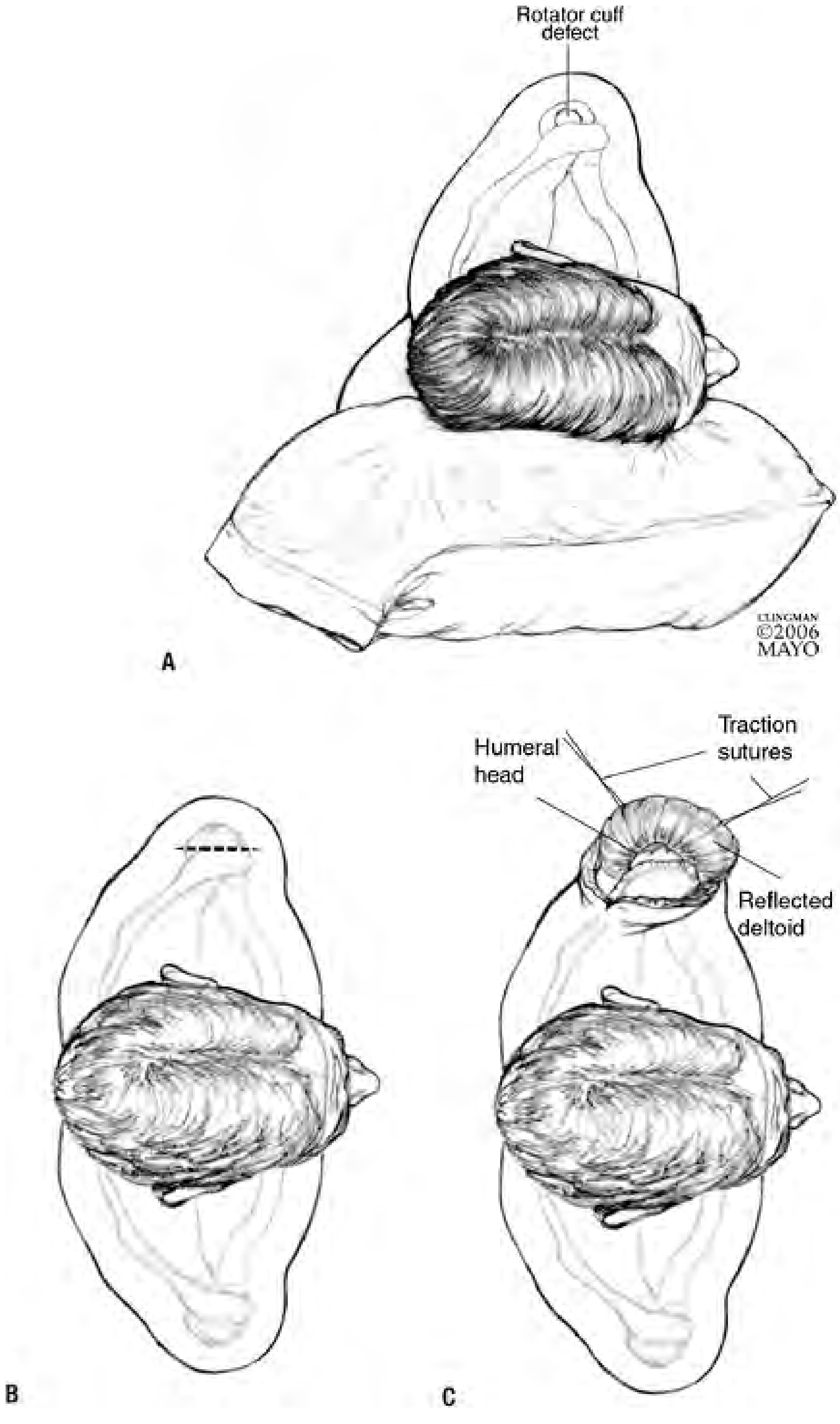
Rotator cuff defect
CLINGMAN ©2006 MAYO
A
Humeral head
Traction sutures
Reflected deltoid
B
C

FIGURE 8-1 A: The patient is placed in the lateral decubitus position with all bony prominences well padded. B: The superior approach for rotator cuff repair

in Langer's lines is used. C: The deltoid is split in line with its fibers, or as an alternative if greater exposure is desired, the deltoid can be reflected from the acromion with a bone chip using an osteotome.

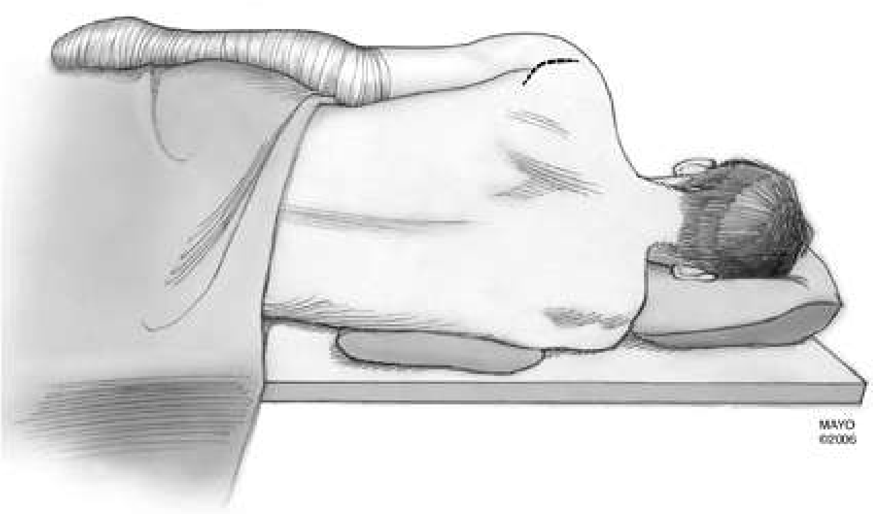

FIGURE 8-2 The second skin incision is placed along the posterior aspect of axilla, along the lateral border of the latissimus.

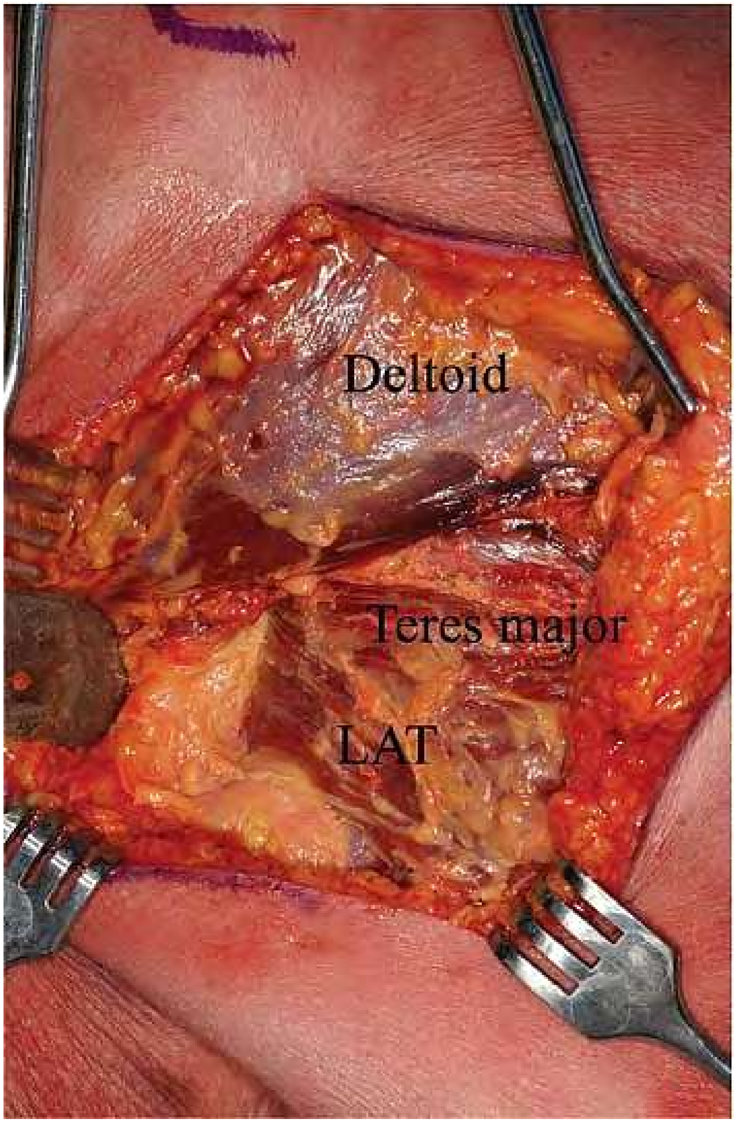
Deltoid
Teres major
LAT

FIGURE 8-3 The latissimus and teres major muscles are visualized inferior to the deltoid muscle.

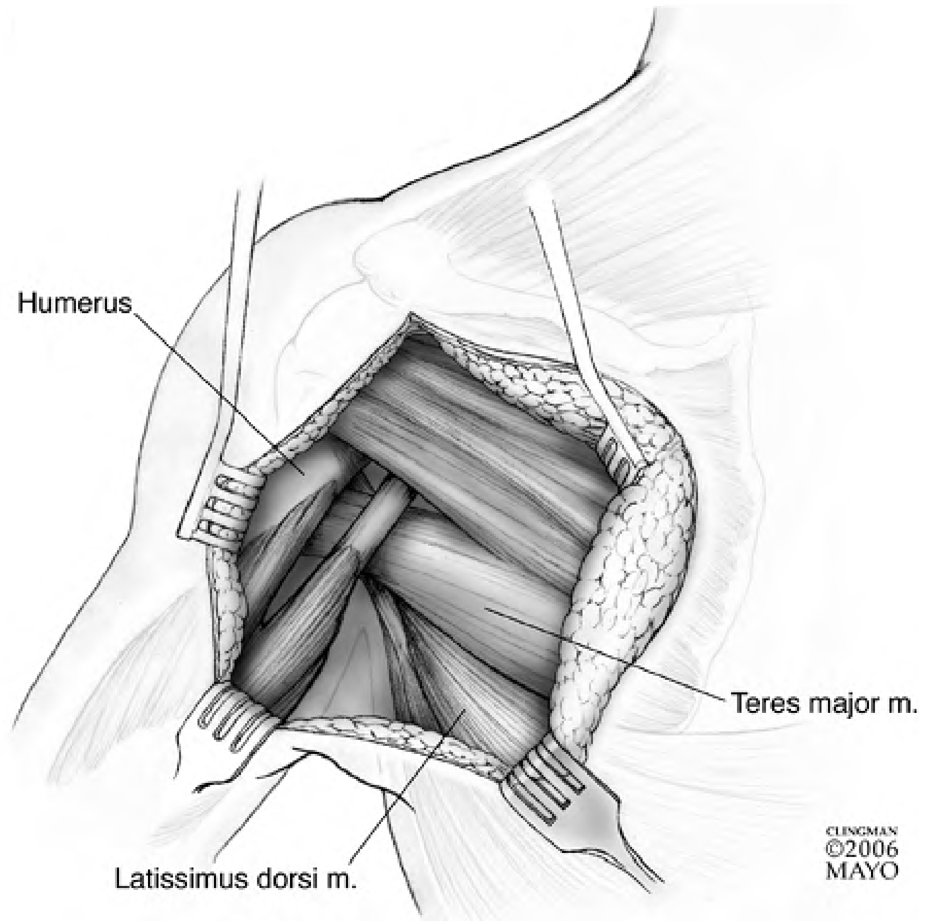
Humerus
Teres major m.
Latissimus dorsi m.
CLINGMAN
©2006
MAYO

FIGURE 8-4 Dissecting distally, identify the latissimus and teres major tendons, which are in intimate contact. Note that the deltoid muscle has been removed for illustrative purposes.

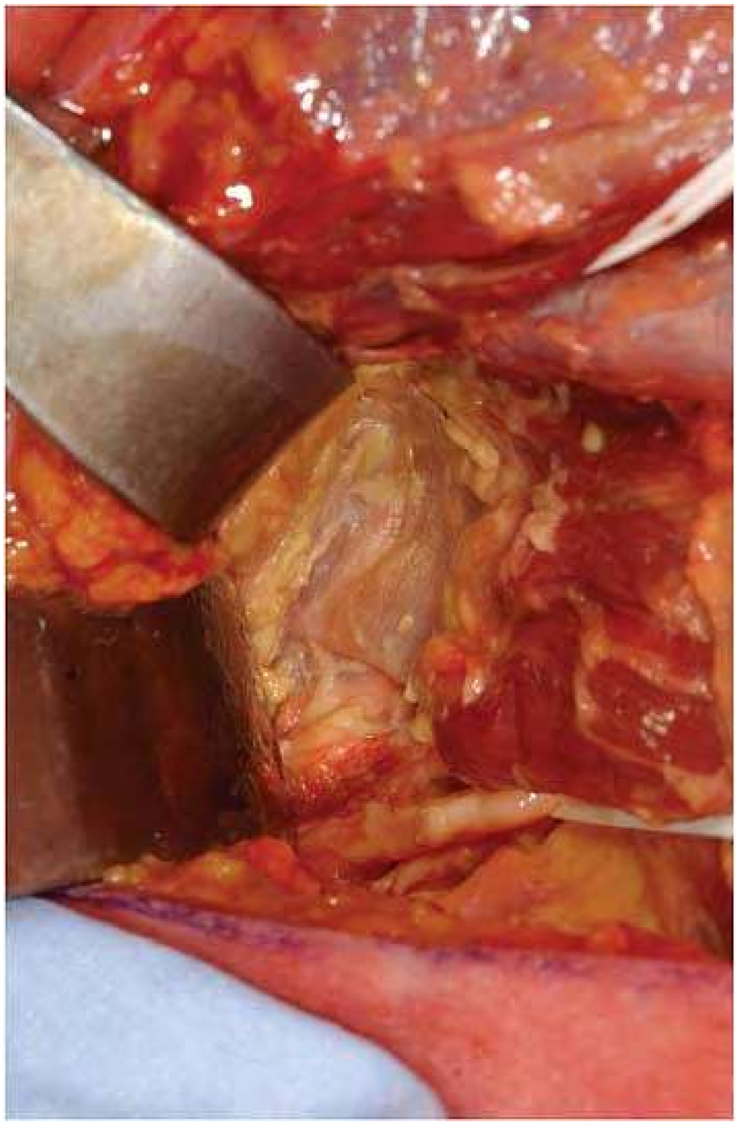

FIGURE 8-5 The latissimus tendon is very thin, as you follow its insertion site into the axilla.

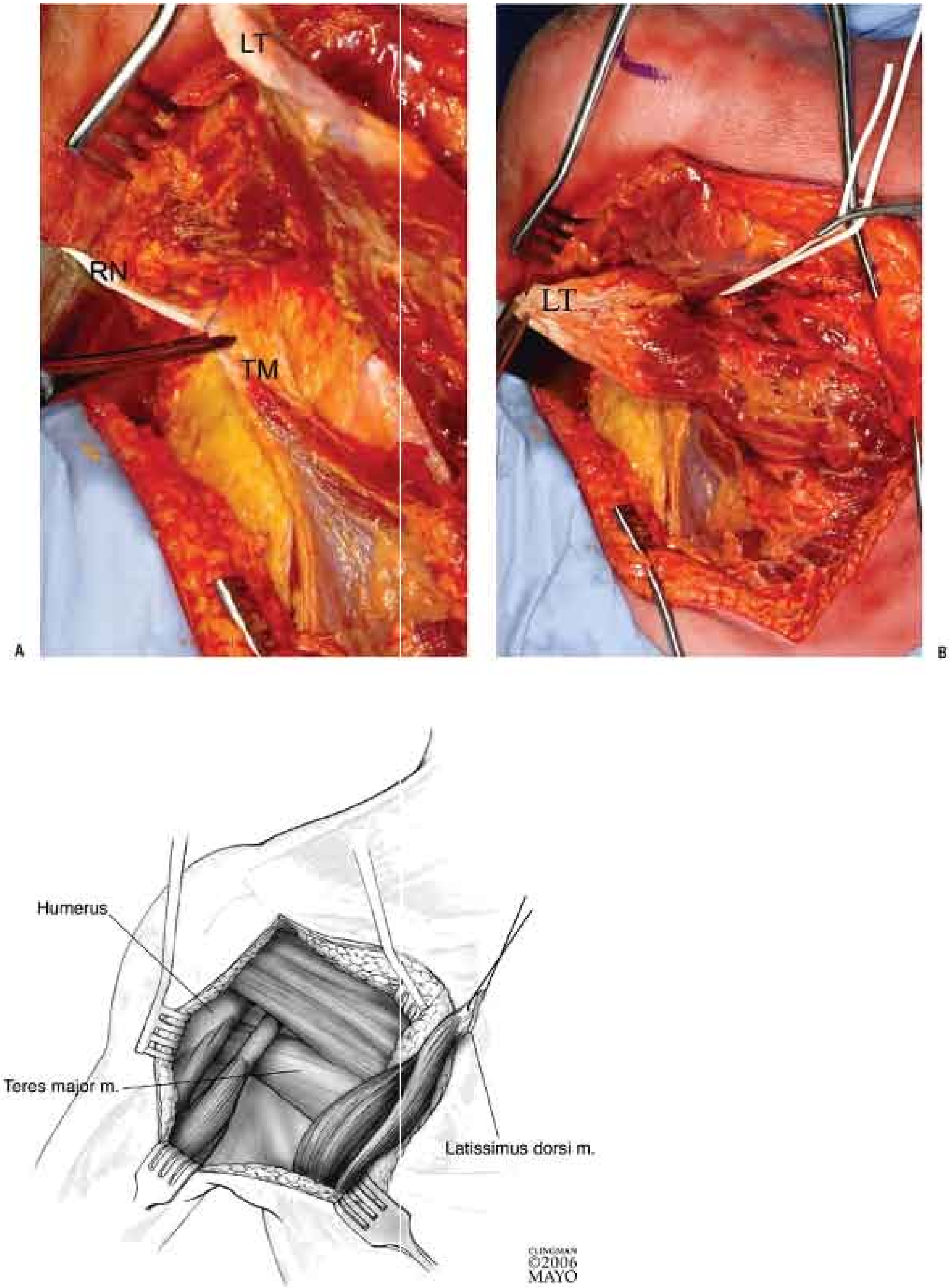
LT
RN
TM
A
LT
B
Humerus
Teres major m.
Latissimus dorsi m.
CLINGMAN
©2006
MAYO
C

FIGURE 8-6 A: The teres major tendon (TM), held with the forceps, can be transferred with the latissimus tendon if the subscapularis is intact and if the patient has extreme weakness. Note the radial nerve is identified in this photograph with the white vessel loop. B: The latissimus tendon should be released directly off of the bone to achieve greatest length. C: The latissimus tendon has been released. Note that the deltoid muscle has been removed for illustrative purposes.

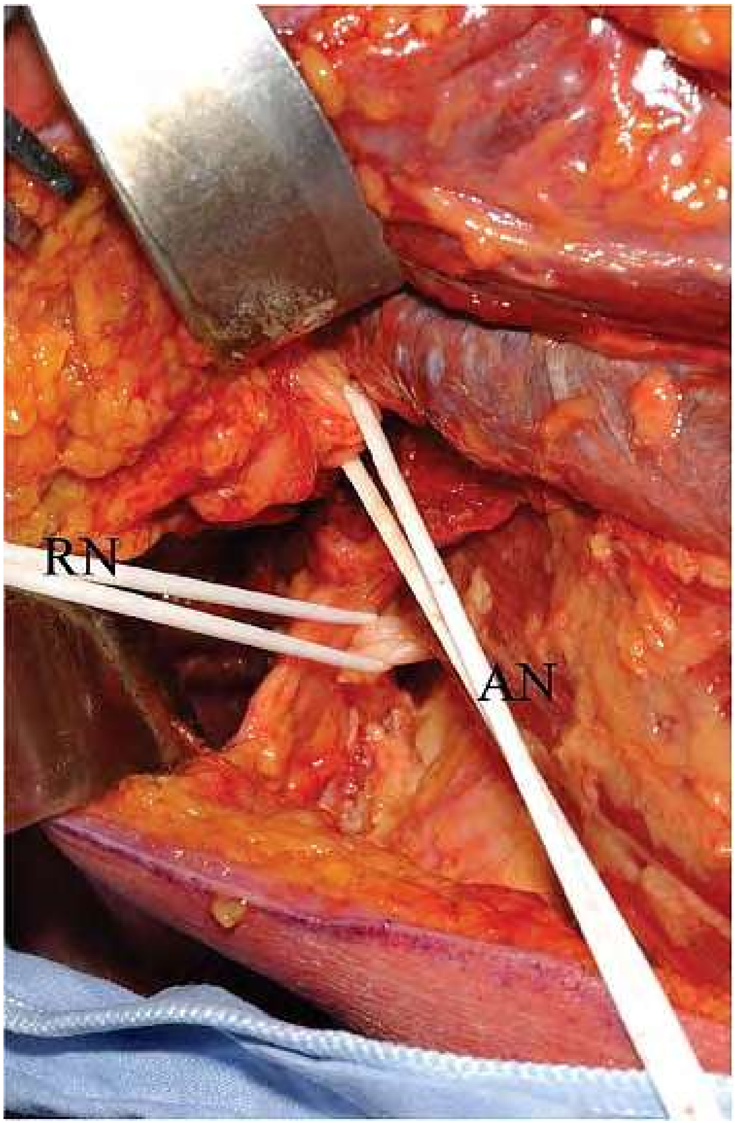

FIGURE 8-7 The axillary and radial nerves have been identified for illustrative purposes.

The deltoid with its bony attachment is repaired back to the acromion with nonabsorbable sutures. The wound is closed in layers over a drain.

Postoperative Management

The arm is placed into a pillow sling for 6 weeks to remove tension from the repair. During this period, passive flexion and external rotation are performed with the arm in the brace. Internal rotation is prohibited to prevent stress on the tendon repair. Specific training with a physical therapist is helpful to retrain the transferred musculotendinous unit to contract during active external rotation and forward flexion of the shoulder. The ability to gain voluntary control of the transferred latissimus is variable, and in many instances probably functions as an interposition. Many patients will be able to actively recruit the muscle during external rotation and forward flexion after 6 months.

Complications

Complications after latissimus dorsi transfer have been rare. Rupture of the transferred tendon has been reported. Ensuring that the tension of the repair is not due to under-excessive tension, protection of the repair with the splint, and avoiding active internal rotation for 6 weeks should prevent this complication. Refixation of the transferred tendon may be attempted.

Temporary axillary and radial nerve palsy has also been reported and perhaps avoided by careful surgical dissection and avoiding exuberant retraction of the deltoid fibers during exposure.

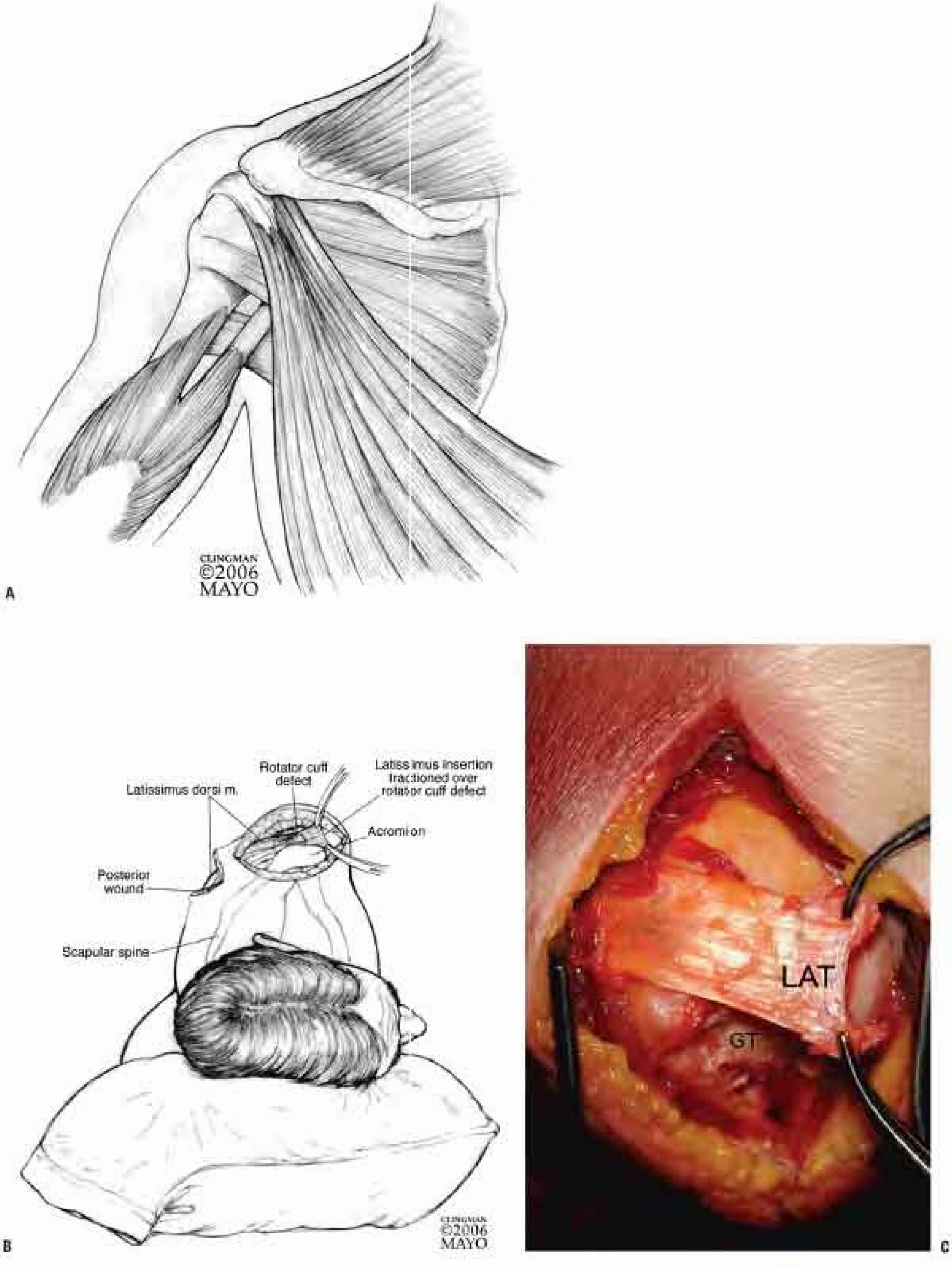
A
CLINGMAN
©2006
MAYO
Rotator cuff defect
Latissimus insertion tractioned over rotator cuff defect
Latissimus dorsi m.
Acromion
Posterior wound
Scapular spine
B
CLINGMAN
©2006
MAYO
LAT
GT
C

FIGURE 8-8 Aâ€“C: The latissimus tendon has been passed under the posterior deltoid, up into the superior wound, and over the superior posterior aspect of the humeral head (HH). The teres major can be passed similarly. Note that the deltoid muscle has been removed for illustrative purposes.

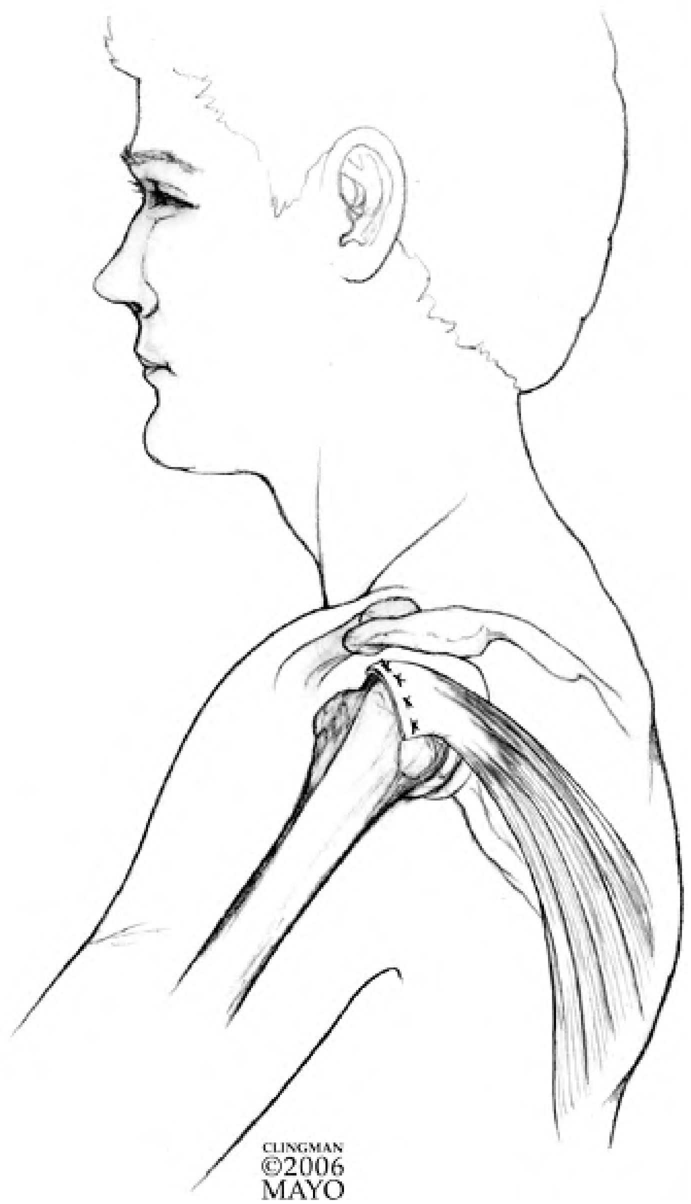

FIGURE 8-9 The tendon is tied over a bony bridge at the greater tuberosity.

Pectoralis Major Tendon Transfer

Surgery

Standard beach-chair positioning is used for this surgery. A deltopectoral approach is made, and skin flaps are created (Fig. 8-10). The torn subscapularis musculotendinous unit is mobilized by releasing adhesions from the base of the coracoid, the brachial plexus, and the subscapularis fossa. A repair of the subscapularis is attempted by placement of modified Mason-Allen sutures in the remaining tendon and fascial tissue using nonabsorbable suture. The repair to the lesser tuberosity should be made with the arm in neutral rotation, if possible.

If repair is not possible, a pectoralis transfer is performed. Through the deltopectoral approach, the conjoint tendon, the tendon of the pectoralis major, and the anterior surface of the humeral head are exposed (Fig. 8-11). The tendon of the pectoralis major is exposed over its full length at the humerus, and the superior one-half to two-thirds of the tendon (depending on the size of the defect) is detached from the humerus (Fig. 8-12). The clavicular portion of the pectoralis major is taken for the transfer, and is bluntly dissected over a length of about 10 cm from the sternal portion of the pectoralis major, which remains intact. Modified Mason-Allen sutures are placed in the tendon using nonabsorbable suture (see Fig. 8-12).

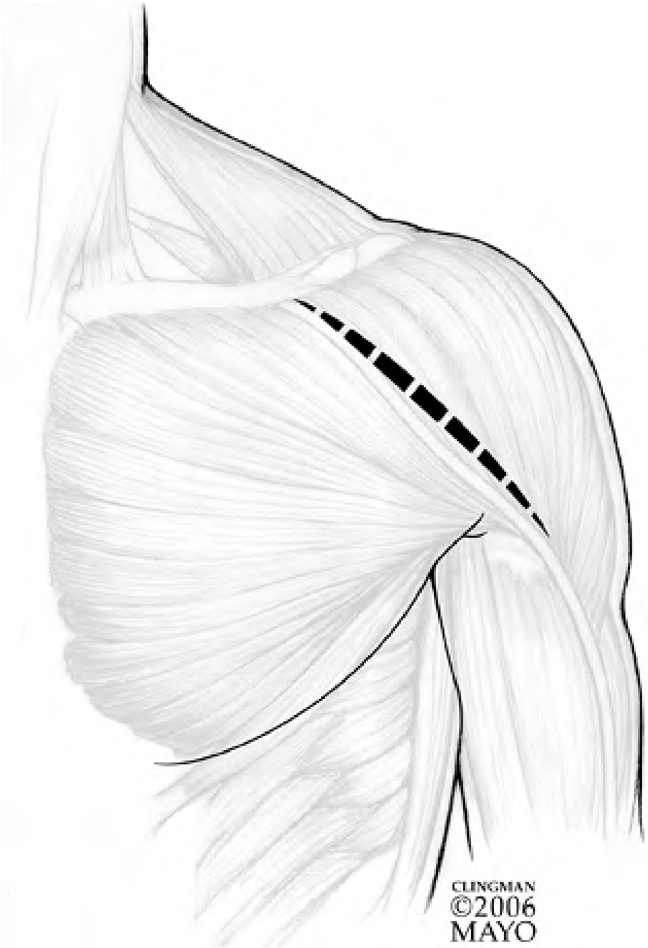
CLINGMAN
©2006
MAYO

FIGURE 8-10 A deltopectoral incision is made.

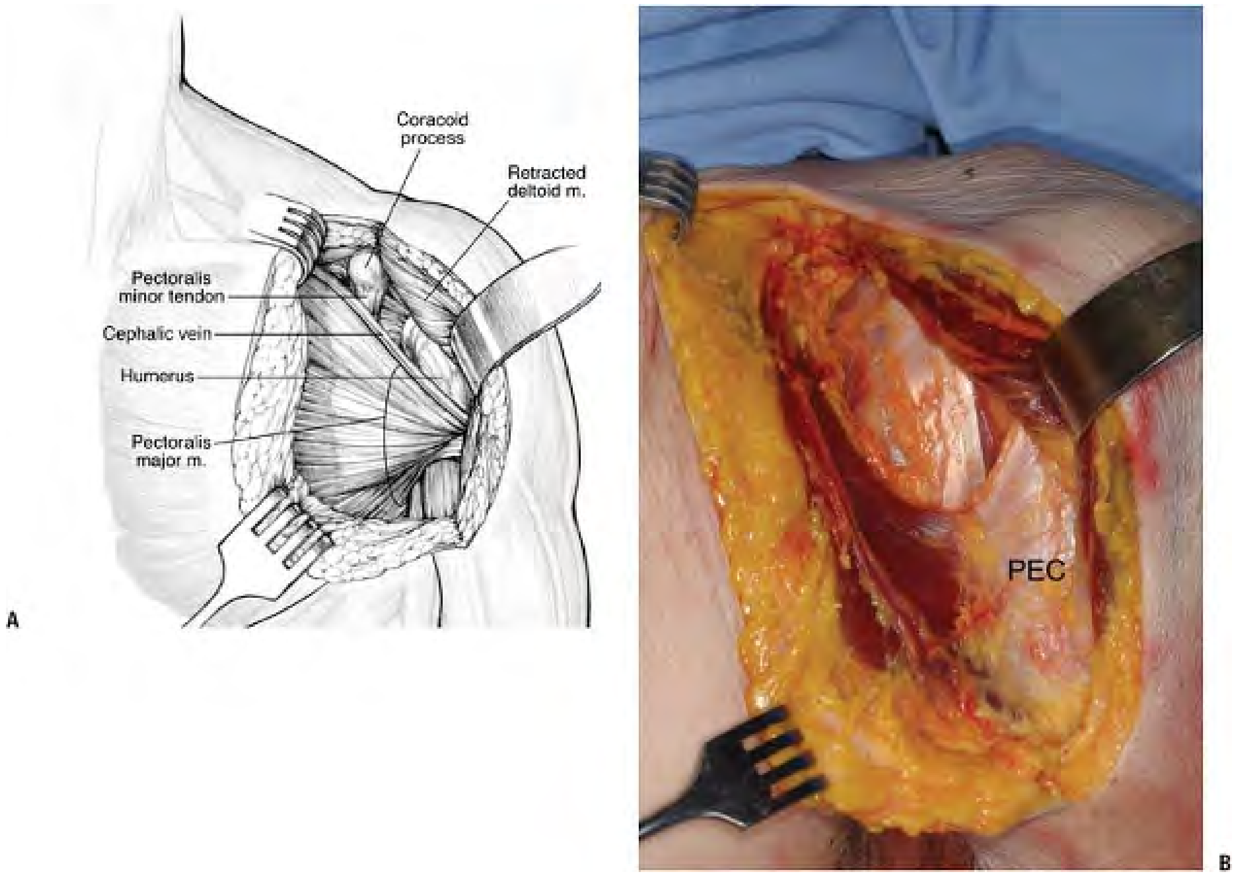
Coracoid process
Retracted deltoid m.
Pectoralis minor tendon
Cephalic vein
Humerus
Pectoralis major m.
A
PEC
B

FIGURE 8-11 A,B: Through the deltopectoral approach, the conjoint tendon, the tendon of the pectoralis major, and the anterior surface of the humeral head are exposed.

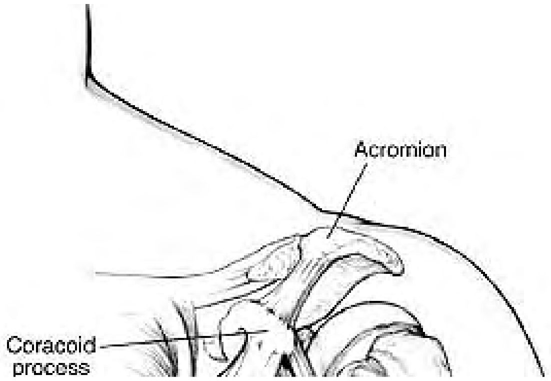
Acromion
Coracoid process

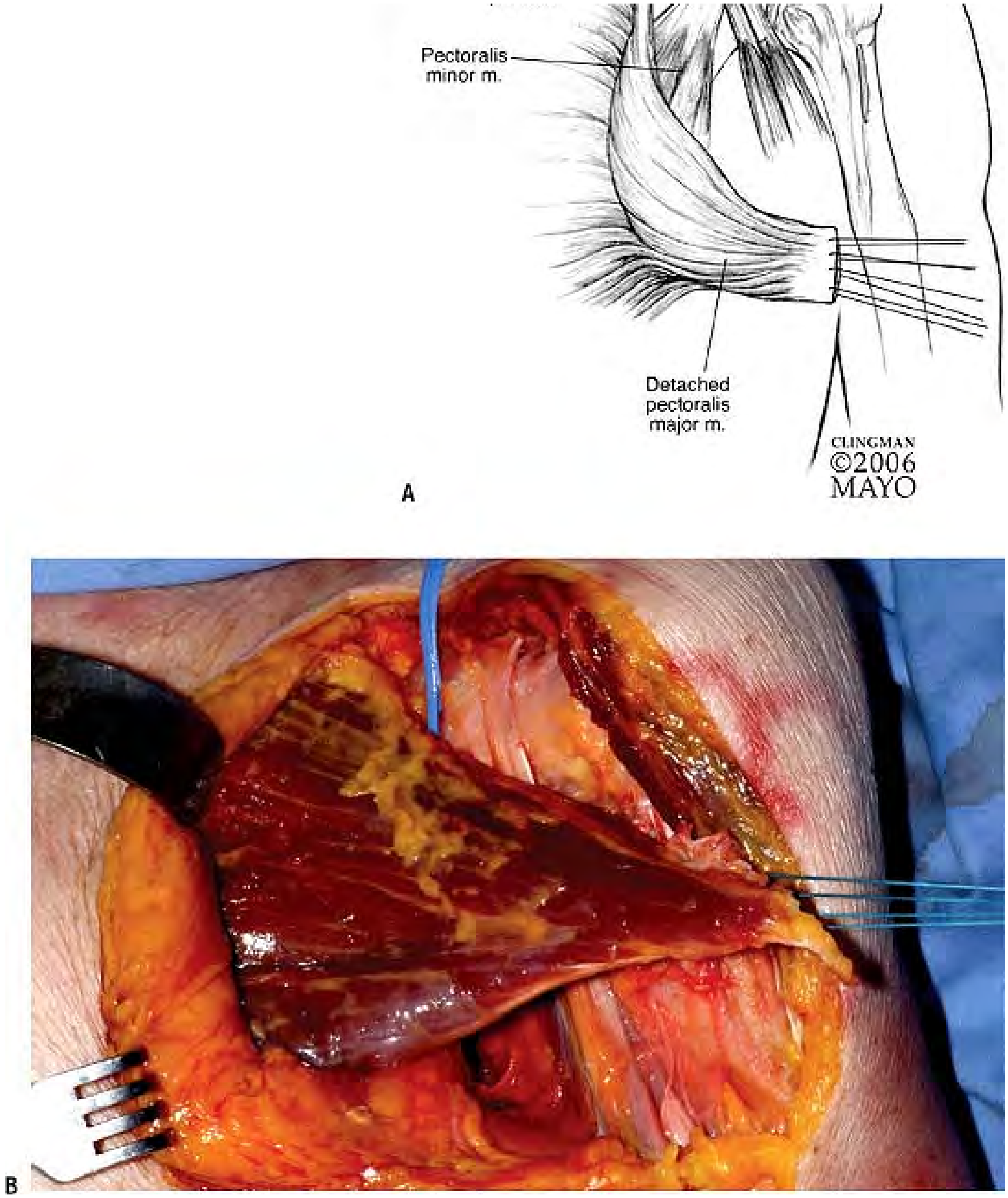
Pectoralis minor m.
Detached pectoralis major m.
CLINGMAN
©2006
MAYO
A
B

FIGURE 8-12 A: The tendon of the pectoralis major is exposed over its full length at the humerus, and the superior one-half to two-thirds of the tendon (depending on the size of the defect) is detached from the humerus. B: Locking sutures have been placed into the pectoralis major tendon.

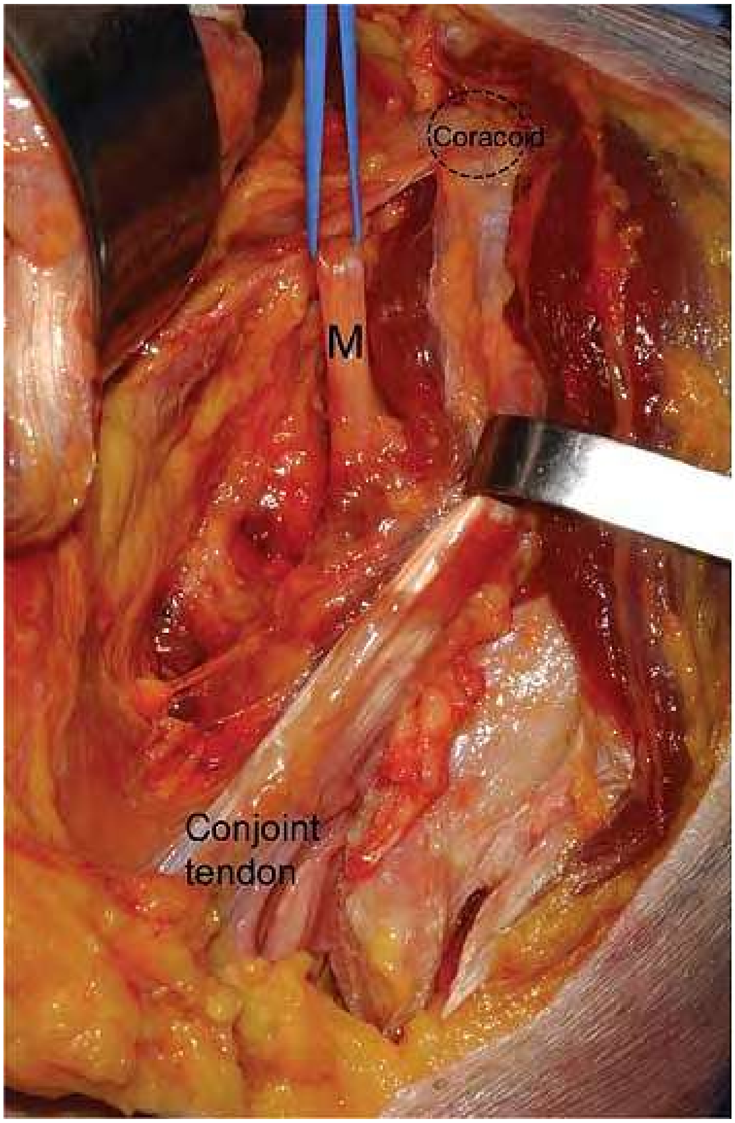

FIGURE 8-13 The musculocutaneous nerve has been identified with the blue vessel loop, between the conjoined tendon and the pectoralis major, about 5 cm distal to the coracoid.

The musculocutaneous nerve is identified by palpation, in the interval between the conjoined tendon and the pectoralis minor (Fig. 8-13). A reinsertion site for the pectoralis major tendon at the medial aspect of the lesser tuberosity is prepared with a rongeur. Using curved forceps, the pectoralis major muscle is advanced behind the conjoined tendon in front of the musculocutaneous nerve, and the tendon is attached to the lesser tuberosity using transosseous repair, with the arm in neutral rotation (Fig. 8-14). The musculocutaneous nerve is palpated within the space between muscles to confirm that there is no tension on the nerve. If tension is found, the size of the muscle belly should be reduced. The supraspinatus is sutured to the proximal border of the pectoralis major tendon to close the rotator interval, if possible. In most patients, the tendon should be long enough to permit at least 30 degrees of external rotation of the arm after the tendon has been sutured to the lesser tuberosity. The conjoined tendon has been noted to arch forward slightly.

A proximal biceps tenodesis or release may be performed, depending on the presence of tendinopathy. Acromioplasty is not performed if the patient has an isolated subscapularis tear, or if the patient has a high likelihood of postoperative anterosuperior subluxation.

Postoperative Management

A sling is worn for 6 weeks postoperatively, and passive range-of-motion exercises are performed within the safe range of motion found intraoperatively starting on postoperative day 1. External rotation to neutral is allowed. Active range-of-motion exercises in all directions, including external rotation, are begun at 6 weeks postoperatively. Full loading is allowed at 12 weeks postoperatively.

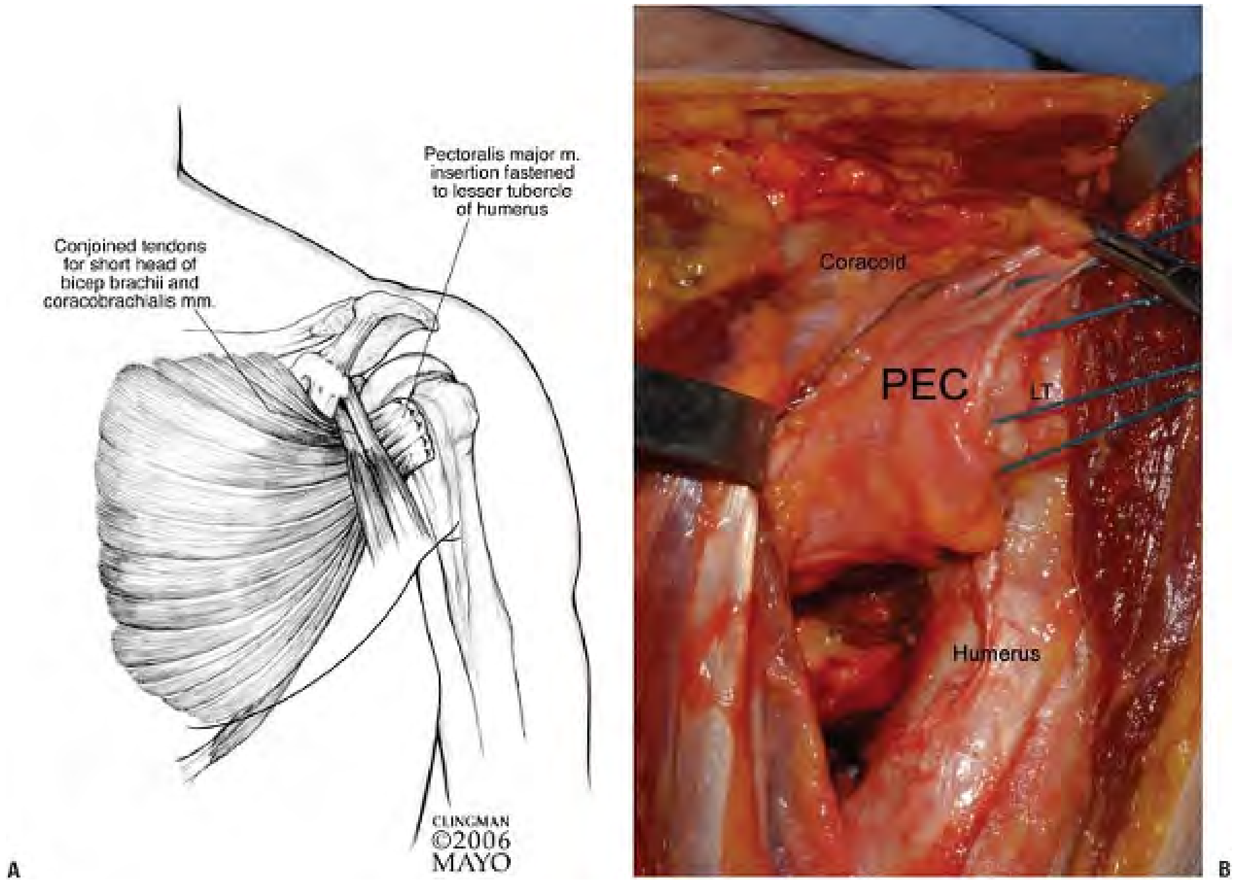
Conjoined tendons for short head of bicep brachii and coracobrachialis mm.
Pectoralis major m. insertion fastened to lesser tubercle of humerus
CLINGMAN ©2006 MAYO
A
Coracoid
PEC
LT
Humerus
B

FIGURE 8-14 A: The pectoralis major muscle is advanced behind the conjoined tendon in front of the musculocutaneous nerve. For illustrative purposes, the inferior half of the pectoralis tendon is not shown. B: The tendon is attached to the lesser tuberosity using transosseous repair, with the arm in neutral rotation. For illustrative purposes, the inferior half of the pectoralis tendon is not shown.

Complications

Avulsion of the transferred pectoralis major tendon has been reported as a potential complication by Jost et al. This can be diagnosed clinically and by MRI, and has been treated with refixation of the transfer. Postoperative infection is another rare complication that should be treated with open irrigation, debridement, and a 6-week course of antibiotics. In very thin patients, a slightly more prominent anterior bulging

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 9

Surgical Exposure of the Elbow Following Bony and Soft Tissue Trauma

Emilie V. Cheung

Nathan A. Hoekzema

Scott P. Steinmann

Indications/Contraindications

The soft tissue surrounding the elbow is thin and pliable, yet durable enough to withstand constant flexion and extension. Optimal functional recovery in traumatic elbow injuries requires early and, in many cases, immediate motion. Skin slough, delayed healing, and dehiscence of skin flaps following treatment of elbow fractures can adversely affect elbow rehabilitation, delaying motion and increasing the risks of infection and nonunion. Properly planned incisions can allow for excellent exposure of the elbow while maximizing skin flap vascularity, minimizing the risks of postoperative wound complications.

Surgical exposures of the elbow typically involve a â€˜universalâ€™ posterior incision. This allows circumferential access by the creation of full-thickness skin flaps medially and laterally as needed. This allows for all intraoperative possibilities and avoids the creation of skin bridges if further surgery is needed in the future. Alternatively, separate medial and lateral skin incisions may be performed for isolated medial and lateral exposure. Prior surgical procedures should be carefully noted preoperatively.

Previous incisions may need to be incorporated in order to avoid creating narrow skin bridges which are susceptible to skin necrosis and potential wound complications.

Surgery

Patient Positioning

In our practice, the patient is routinely placed in the supine position with a small stack of towels placed underneath the ipsilateral scapula, and the arm draped across the chest after sterile preparation. The operating table should be slightly tilted away from the surgeon to help with visualization and exposure. This positioning is optimal for fixation of distal humeral fractures, radial head fractures, and medial or lateral ligament reconstruction. For medial ligament or coronoid fracture fixation, the supine position is used, but a sterile Mayo stand is utilized to position the arm more accurately for medial elbow exposure. Intraoperative fluoroscopy, if indicated, is placed on the ipsilateral side, and the elbow may be brought out laterally from the chest for the fluoroscope intermittently during the procedure. A sterile tourniquet is usually used in all elbow approaches. This allows for ease of removal if more proximal exposure of the humerus is needed. Hip rests or a bean bag may provide additional stability.

Posterior Approaches

Indications

- Triceps repair
- Combined medial and lateral approach
- Distal humerus ORIF
- Total Elbow Arthroplasty

Contraindications

Prior medial or lateral incisions.

Incision

The bony landmarks, including the olecranon process and the subcutaneous border of the proximal ulna, are marked (Figs. 9-1,9-2,9-3,9-4,9-5,9-6). The incision starts about 5 centimeters proximal to the olecranon process centered on the triceps tendon. It is then taken distally to either the lateral or medial side of the olecrenon, according to surgeon preference, and finishes distally following the subcutaneous border of the ulna. Full-thickness skin flaps are then developed, kept as thick as possible with the deep plane being the triceps fascia and epitenon proximally, and the forearm fascia and ulnar periosteum distally. If most of the surgical procedure will be on the medial side of the elbow, then the posterior incision should be placed to that side. Likewise, if a radial head fracture is being exposed, then the posterior incision should be made on the lateral side of the olecranon. If one chooses to err on the medial side of the olecranon, the incision should not be placed directly over the cubital tunnel, to avoid injuring the ulnar nerve.

Deep Dissection

The ulnar nerve is most easily found proximally between the medial intermuscular septum and the medial head of the triceps muscle. If the nerve is going to be transposed, it should be freed from proximal to distal in order to minimize damage to the nerve branches. Articular branches usually need to be sacrificed, but the first motor branch should be identified, mobilized,

and preserved. The leading edge, about 1 cm, of the distal medial intermuscular septum is removed to prevent tethering of the nerve after being anteriorly transposed. The nerve is then placed into a subcutaneous pocket anterior to the medial epicondyle. A suture is placed in the subcutaneous tissue and secured to the fascia to create a sling and prevent posterior subluxation of the nerve. If the nerve is not going to be transposed, care should be taken to avoid destabilizing the nerve in the cubital tunnel by leaving the soft tissue constraints of Osborne's fascia intact.

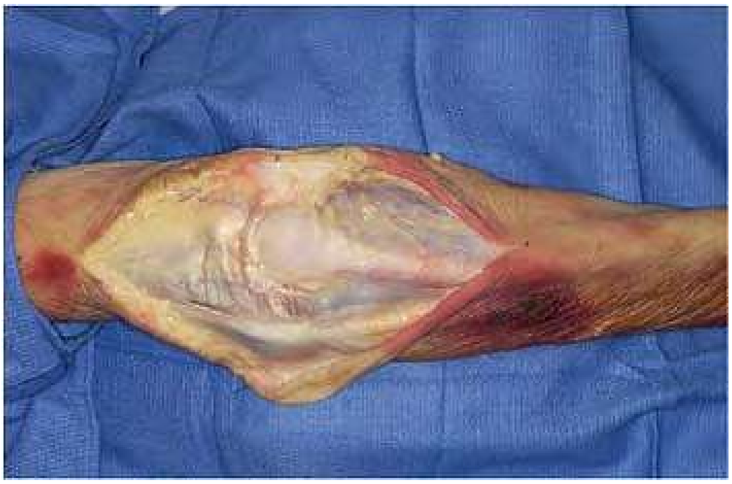

FIGURE 9-1 The skin is incised posteriorly with full thickness flaps raised as needed.

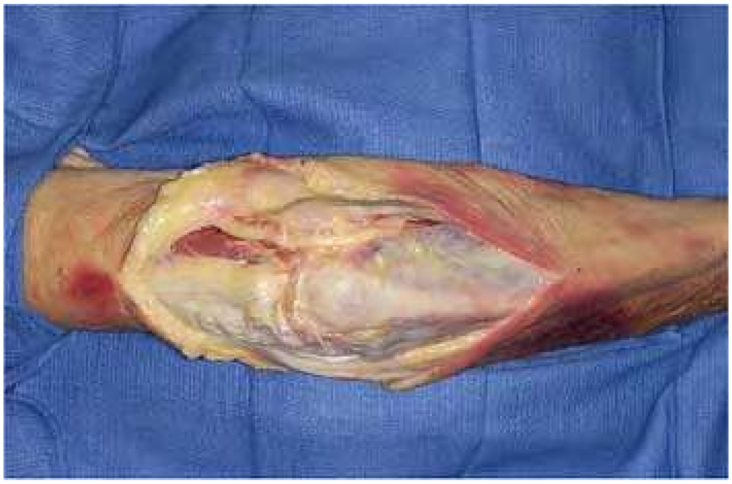

FIGURE 9-2 The ulnar nerve is carefully released from the cubital tunnel.

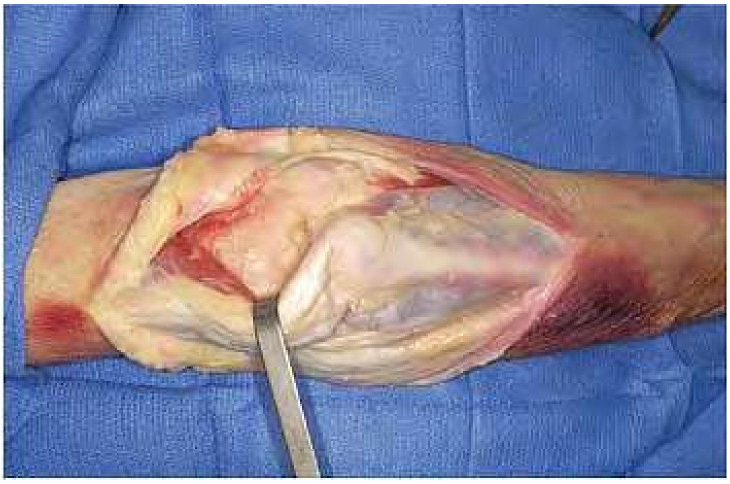

FIGURE 9-3 The medial and lateral sides of the triceps muscle are released in order to fully visualize the posterior fossa and the ulnohumeral articulation.

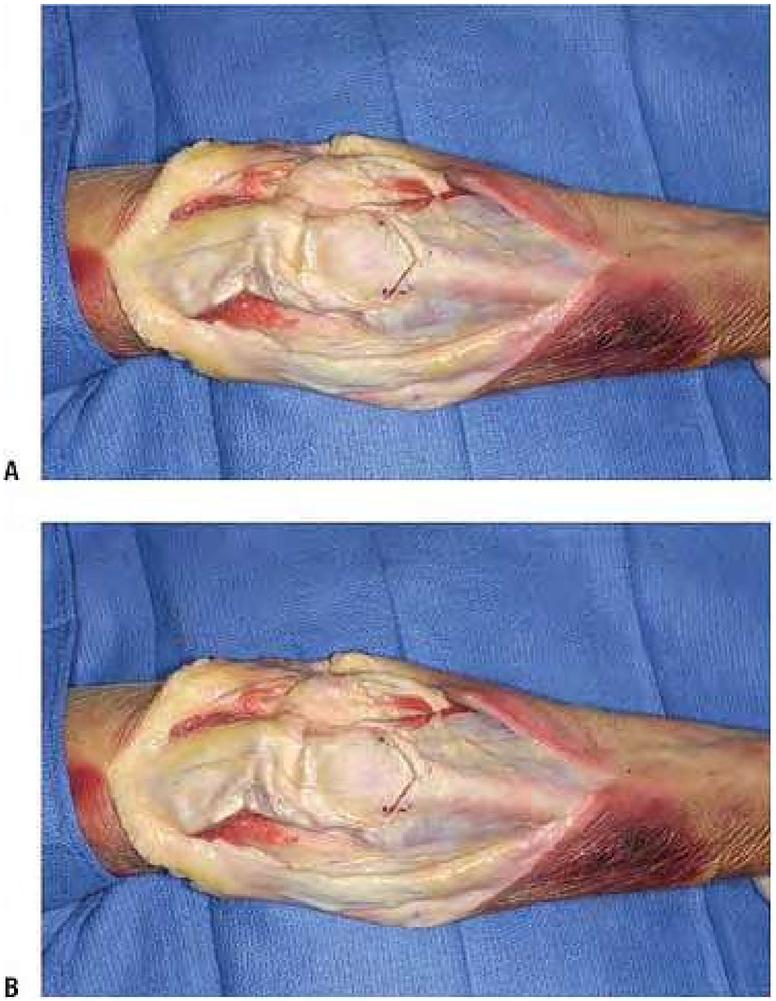

FIGURE 9-4 A,B: The olecranon osteotomy is best located at the small area of olecranon that is devoid of cartilage. A lap sponge may be threaded between the humerus and ulna to avoid inadvertent cartilage damage. The osteotomy is started with a reciprocating saw and then finished with an osteotome. Hardware may be predrilled to facilitate final osteotomy repair.

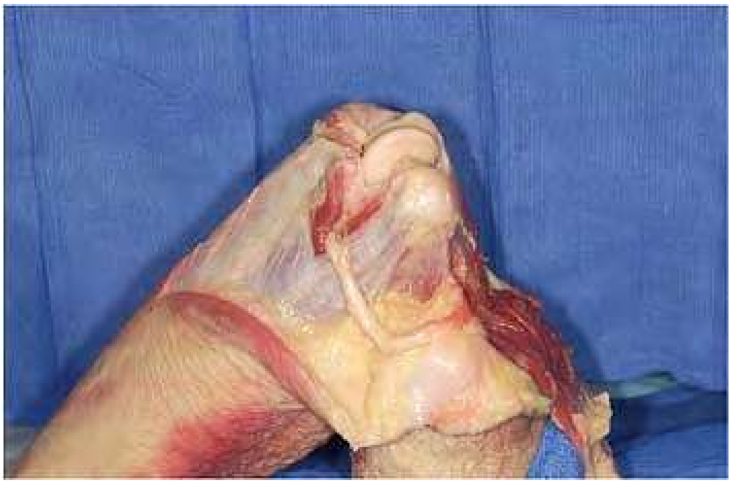

FIGURE 9-5 The osteotomied olecranon is retracted with the triceps tendon demonstrating near full visualization of the articular spool.

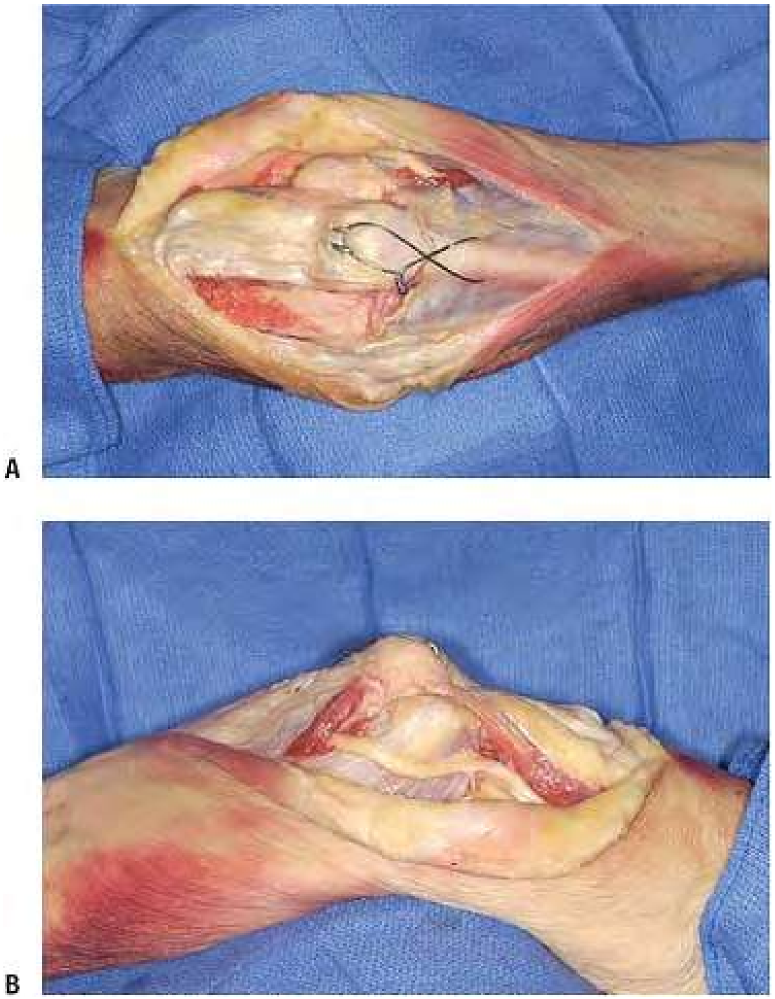

FIGURE 9-6 A,B: The completed tension band repair. The bent K-wires will then be sunk beneath the triceps tendon and the tendon repaired over them to minimize pin backout.

Olecranon Osteotomy

An olecranon osteotomy may be performed to gain full visualization of the articular surface of the distal humerus.

After identification and protection of the ulnar nerve, a capsulotomy is made at both the medial and lateral sides of the olecranon at the apex of the greater sigmoid notch to locate the olecranon â€œbare area.â€• This is the area of the olecranon where the articular cartilage narrows and is the optimal location for the osteotomy. A lap sponge may be threaded under the olecrenon to protect the distal humeral articular cartilage from inadvertent damage. The osteotomy is created in a Chevron configuration, with the apex usually pointed distally using an oscillating saw. The saw is used for one-third of the bone cut, the second third is made with an osteotome, and the last third is cracked or fractured open with an osteotome to create the completed osteotomy. Once the osteotomy has been made, the anconeus needs to be taken off the lateral side of the olecranon fragment. Transection of the anconeus will denervate it. To avoid this problem, one technique is an anconeus flap transolecranon approach (AFT) which involves detaching the anconeus as a flap distally to proximally until the osteotomy site is reached. The osteotomy is then performed and the olecranon fragment and the anconeus are retracted proximally together to expose the distal humerus. Alternatively, the anconeus may be elevated from the proximal fragment of the olecranon osteotomy and elbow region with its distal attachment on the ulnar shaft left intact. The anconeus is then retracted laterally while the proximal fragment of the olecranon osteotomy is retracted proximally, which allows ample visualization of the distal humerus. The osteotomy site has been reported to heal reliably with few complications.

It is advantageous to place the hardware on the proposed olecranon fragment (if using a cannulated screw or Kirschner wires) prior to making the osteotomy so that the tract of the hardware is created anatomically, which will facilitate optimal alignment of the osteotomy site at the end of the procedure. We use either two Kirschner wires or a 6.5 or 7.6 mm cannulated screw and washer, with an 18 gauge tension band wire. The drill hole for placement of the wire is placed at an equal distance from the olecranon osteotomy as the distance of the osteotomy site to the tip of the olecranon. The osteotomy should be repaired by advancing the K-wires into the anterior ulnar cortex distal to the coronoid. After the K-wires have reached the anterior cortex, they are backed out about 5 mm and bent 180 degrees and tapped back until buried under the triceps tendon. The triceps may then be sutured over the wires to discourage backing out. Alternatively, a precontoured olecranon plate may be utilized for fixation.

Bryan-Morrey triceps reflecting approach

After identification and protection of the ulnar nerve, a periosteal elevator is used to dissect the triceps muscle, including the medial and lateral margins of the triceps muscle, from the posterior humeral cortex. The triceps tendon is sharply dissected through Sharpey's fibers directly off of the olecranon starting medially and extending laterally. It is helpful to use a Beaver blade scalpel to sharply dissect the triceps tendon off of bone. The fascial overlying Kocher's interval is identified and longitudinally split. The triceps tendon and anconeus should be reflected medially to laterally, ending at Kocher's interval. The triceps may be removed with a thin wafer of bone as well, but this is not our standard practice. At the end of the procedure, the triceps tendon is repaired back to the olecranon using two transosseous drill holes placed in a cruciate configuration and one additional drill hole placed in a transverse orientation with nonabsorbable suture. The repair should be protected postoperatively by avoiding active elbow extension against resistance for a minimum of 6 weeks.

Van Goerder approach

After identification and protection of the ulnar nerve, the triceps tendon is identified several centimeters proximal to its insertion on the olecranon and a Chevron-shaped transection of the tendon is performed with the apex proximal. This is repaired at the end of the procedure with nonabsorbable suture. The repair should be protected postoperatively by avoiding active elbow extension against resistance, and avoiding passive stretching in positions of terminal elbow flexion for a minimum of 6 weeks.

Trap approach

The trans-anconeus pedicle flap (TRAP) approach requires a longer skin incision distally along the subcutaneous border of the ulna. The anconeus muscle is identified along the lateral aspect of the subcutaneous border of the ulna, and released by sub-periosteal dissection from its insertion on the ulna. The muscle is released in its entirety distally to proximally. It is reflected proximally to afford visualization of the distal humerus.

Medial Approach

Indications

- Ulnar nerve pathology

- Capsular release for stiffness
- Need to preserve the lateral ulno-humeral ligamentous complex

Contraindications

- Need for access to radial head or lateral ligaments

Incision

The medial exposure can use either the posterior or medial-posterior skin incision (Figs. 9-7 and 9-8). Once again, the skin incision can be tailored to work with need for exposure of other structures or existent scars from prior exposures.

Identification of the ulnar nerve is necessary for an adequate and safe medial exposure. The nerve should be identified proximally and subsequent dissection carried distally. In revision surgery and cases with a prior transposition, it is especially helpful to identify the nerve in a normal area prior to the dissection from scarred tissue. Once the nerve is safely identified, the medial supracondylar ridge is palpated along with the overlaying medial intermuscular septum. The medial antebrachial cutaneous nerve will be found on the fascia anterior to the septum. This should be protected to avoid postoperative neuroma formation.

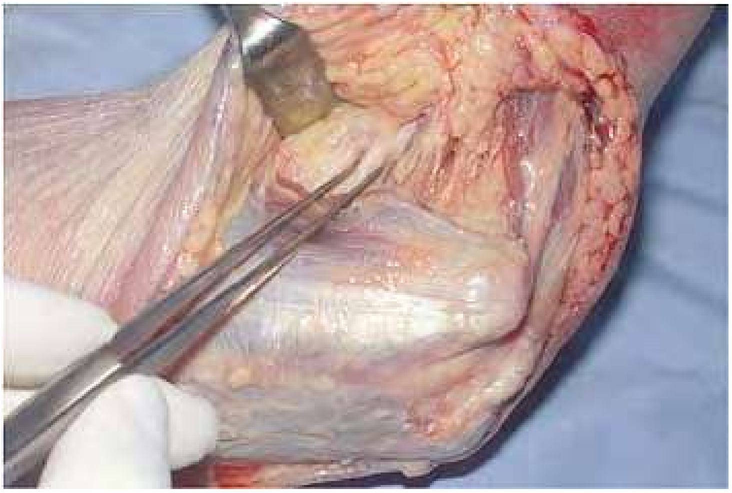

FIGURE 9-7 A skin incision is made and the subcutaneous fat is elevated from the forearm fascia. Care is taken to protect the medial antebrachial cutaneous nerve that lies in the subcutaneous adipose tissue.

The medial intermuscular septum is identified along with the medial supracondylar ridge. The brachial fascia is incised along the anterior aspect of the septum and the flexor-pronator group is released from the supracondylar ridge. Distally the flexor group is split leaving the posterior aspect of the flexor carpi ulnaris in place. The muscle group can be elevated off the anterior capsule and extend all the way across to the lateral aspect of the joint. A cuff of tissue may be left on the ridge, so that the muscle group can be repaired at the end of the procedure. The dissection is subperiosteal, deep to the brachialis so that the brachial artery and median nerve are protected.

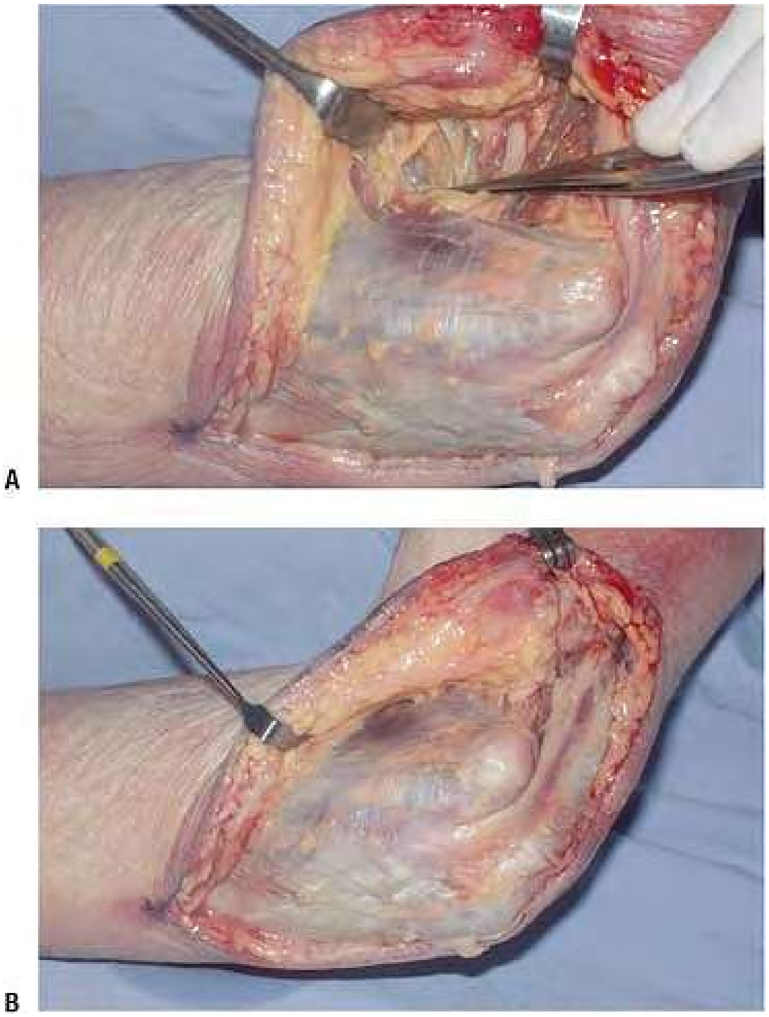

FIGURE 9-8 A,B: The median nerve is found lateral and deep to the flexor pronator group with the lateral antebrachial cutaneous nerve even more lateral.

Medial Coronoid Approach

Indications

- Fixation of coronoid fractures

Incision

The medial coronoid can be easily exposed through the floor of the cubital tunnel (Figs. 9-9,9-10,9-11,9-12,9-13). The posterior or medial approach is used to expose the ulnar nerve. The nerve is gently dissected and may be anteriorly transposed as previously described. It may be secured to a fasciocutaneous sling during fracture fixation to avoid excessive manipulation and inadvertent traction during the procedure. The nerve may be left in place if there are no concerns for neuropathy and as long as it is not destabilized. The two heads of the flexor carpi ulnaris (FCU) are split. The anterior half is retracted anteriorly, and the posterior half is retracted posteriorly in order to expose the coronoid.

The anterior band of the medial collateral ligament (MCL) is usually attached to a large anteromedial coronoid fragment. Care should be taken not to detatch the ligament from this fragment. If the capsular attachments are intact, the surgeon can judge fracture reduction based on realignment of the metaphyseal fracture fragments. The dissection of the flexor carpi ulnaris muscle fibers from the medial collateral ligament should start distally and the muscle should be brought proximally to avoid damaging the ligament and potentially destabilizing the elbow. The insertion of the MCL on the sublime tubercle should be evident. The coronoid will be in the deep portion of the wound, anterior to the ligament.

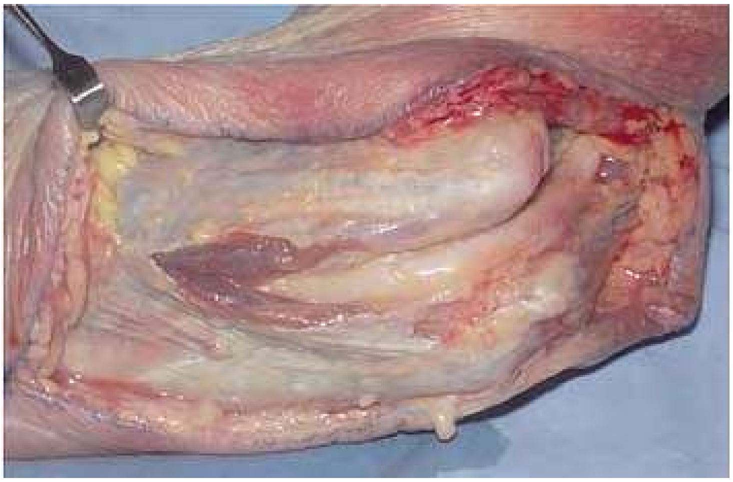

FIGURE 9-9 The ulnar nerve is released from the cubital tunnel in the standard fashion, taking care to split the FCU between the two heads. In this photograph, Osborne's ligament overlying the cubital tunnel has been released.

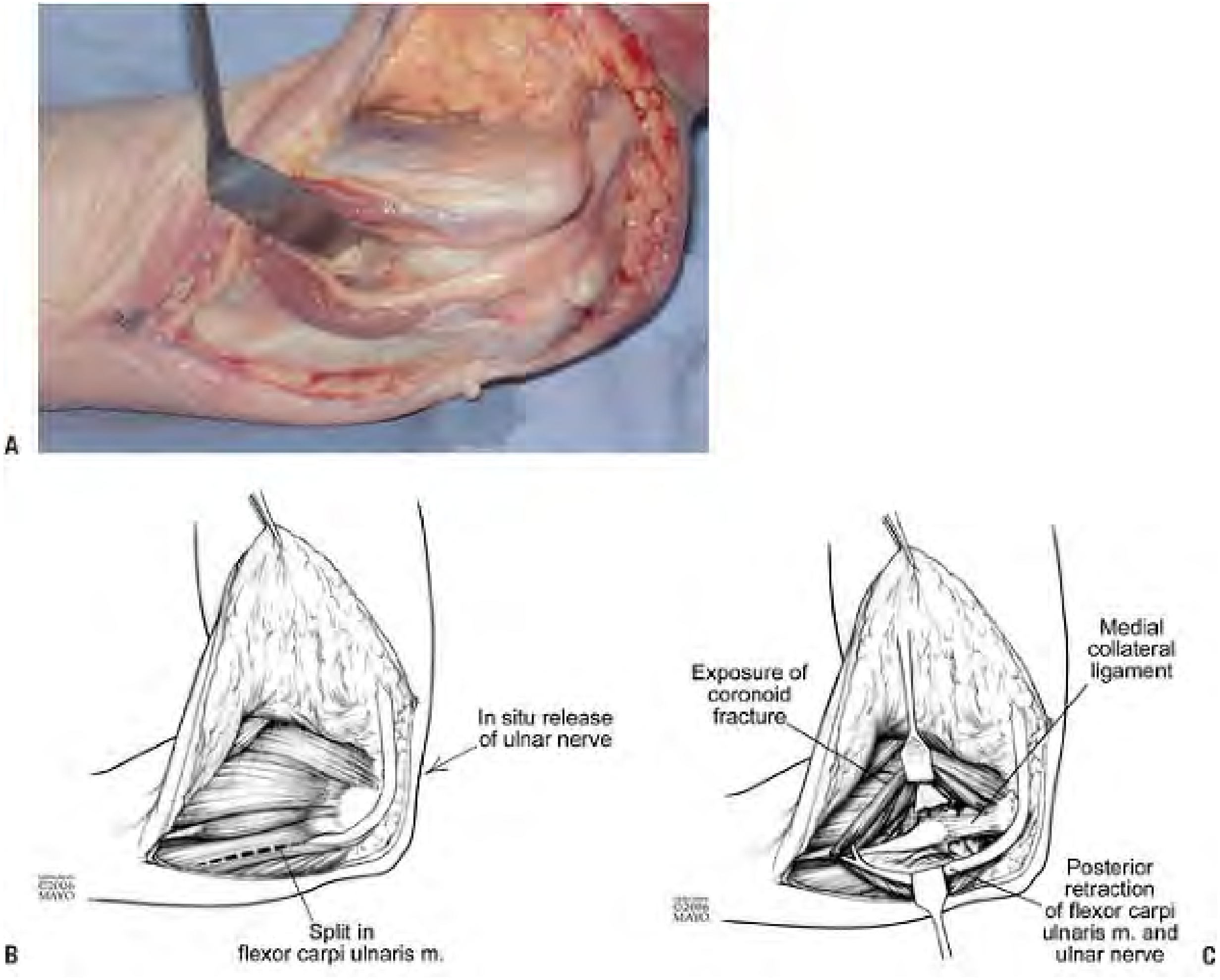
A
In situ release of ulnar nerve
Split in flexor carpi ulnaris m.
©2006 MAYO
B
Exposure of coronoid fracture
Medial collateral ligament
Posterior retraction of flexor carpi ulnaris m. and ulnar nerve
©2006 MAYO
C

FIGURE 9-10 Aâ€“C: The nerve is gently retracted posteriorly and the ulnar head of the FCU is carefully dissected off of the medial collateral ligament (MCL), exposing the coronoid.

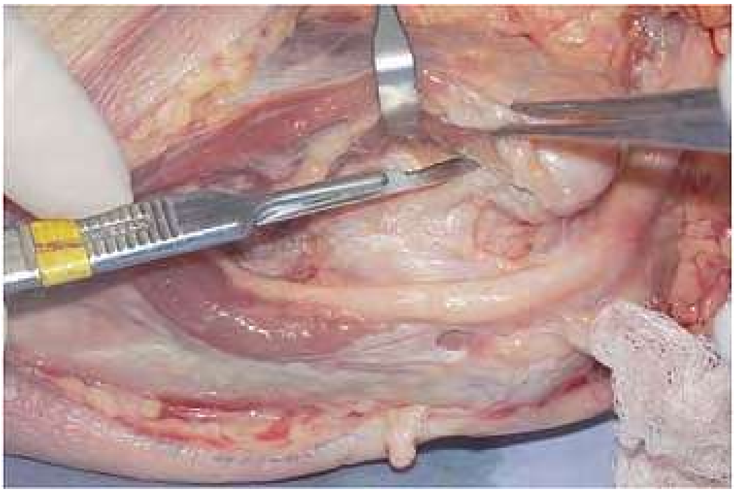

FIGURE 9-11 The medial collateral ligament is nearly fully visualized by peeling the humeral head of the FCU laterally and superiorly.

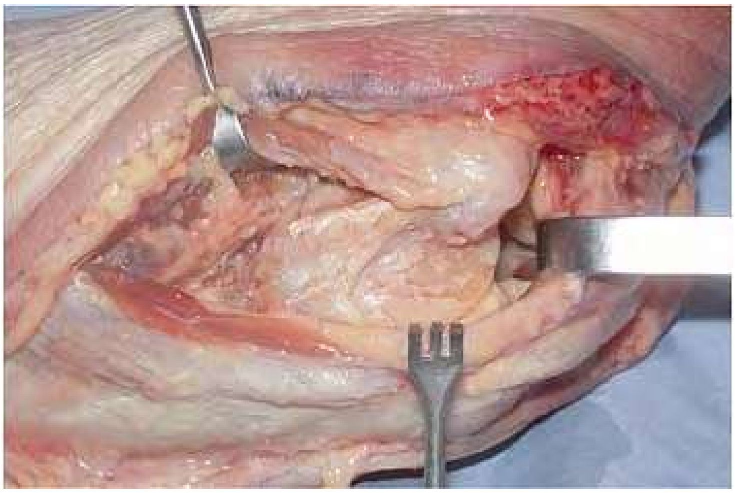

FIGURE 9-12 Utilizing the full extent of this exposure affords access to the coronoid, the medial collateral ligament, and even limited access to the posterior fossa.

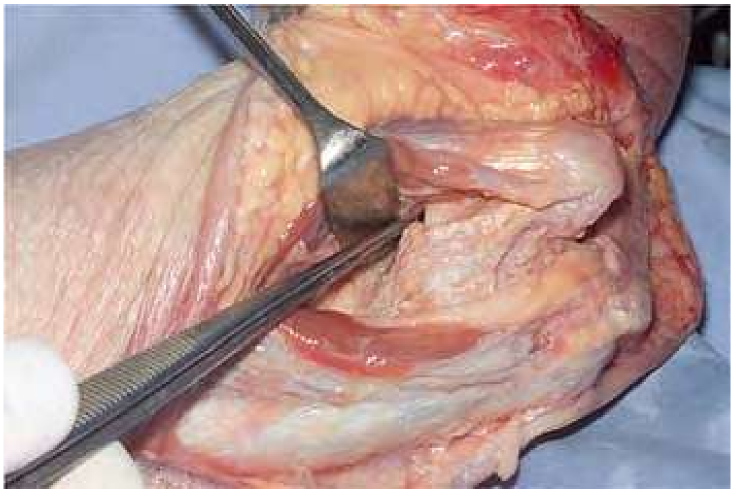

FIGURE 9-13 Entire coronoid may be seen. A full view of the MCL can be achieved including the medial epicondyle as well as the sublime tubercle.

Alternatively, if there is a large coronoid fragment, the FCU may be reflected anteriorly using subperiosteal dissection from the proximal ulna including the flexor-pronator mass proximally, as described by Taylor and Scham. Care must be taken to protect the ulnar nerve, which may need to be transposed when this approach is used.

Lateral Approaches

Indications

- Radial head ORIF, resection, or replacement
- Lateral epicondylitis debridement
- Repair of lateral ulnocollateral ligament
- Release of contractures

Contraindications

- The need to approach medial pathology
- Ulnar nerve involvement
- Medial humeral condyle fractures

Incision

The lateral approach to the elbow has become a standard means to gain access to the elbow joint (Figs. 9-14,9-15,9-16,9-17,9-18,9-19,9-20,9-21,9-22). The deep lateral approach was initially described by Kocher in 1911 and has been subsequently modified by Cohen, Hastings, and Morrey in its extensile exposure to preserve the lateral collateral annular ligament complex.

There are several different variations of the lateral exposure. The direct lateral, the LCL preserving, Kocher's approach, and the Mayo Modification have all been described. The general principle and anatomic intervals, however, are the same and are described here.

This approach to the radial head splits the lateral annular ligament complex and stays anterior to the lateral ulnar collateral ligament. The skin incision is placed either as a posterior incision with a large full-thickness lateral flap, or the incision is placed along the lateral epicondyle over the radial head to the lateral aspect of the ulna. Skin flaps are created over the antebrachial fascia.

Kocher's interval

The interval between the extensor carpi ulnaris (ECU) and the anconeus is identified. The fascia is incised from the lateral epicondyle distally following the junction of the ECU and anconeus. Care is taken to elevate the ECU anteriorly and the anconeus posteriorly. The capsule is incised along the anterior border of the lateral ulnocollateral ligament, about 1 cm above the crista supinatorius. Care should be taken during this dissection in order to preserve the lateral collateral ligamentous complex, and avoid destabilizing the elbow.

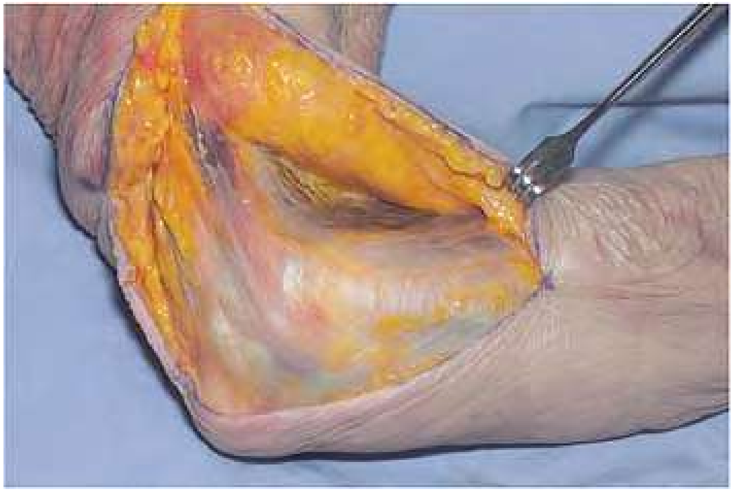

FIGURE 9-14 A lateral or posterior skin incision may be used. The skin and subcutaneous tissues are carefully removed from the investing fascia. Special attention should be taken to avoid damage to the lateral antebrachial cutaneous nerve, which will travel within the fat at the distal aspect of this incision.

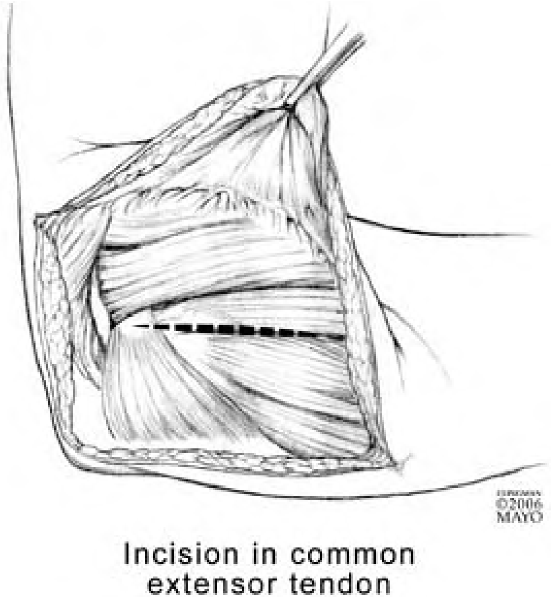

FIGURE 9-15 An alternative to the Kocher's approach is to split the common extensor group at the equator of the radiocapitellar joint. This decreases the likelihood of disrupting the lateral collateral complex.

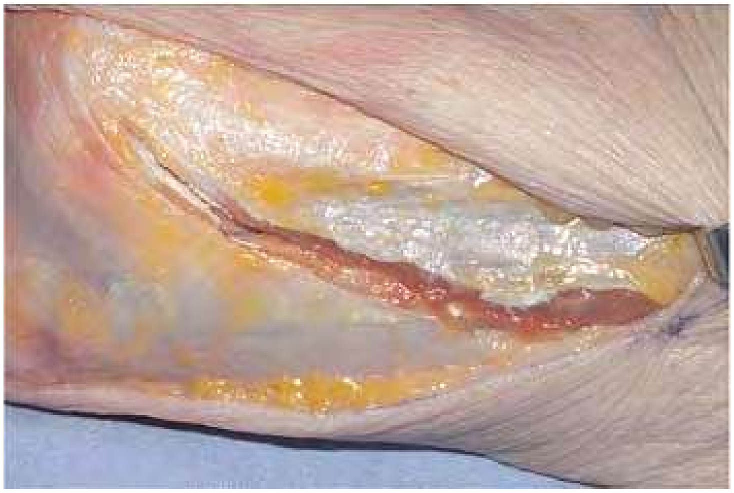

FIGURE 9-16 A traditional Kocher incision is made between the anconeus and the extensor carpi ulnaris (ECU). Care is taken to develop this interval between the lateral collateral ligament complex (LCL) and the ECU.

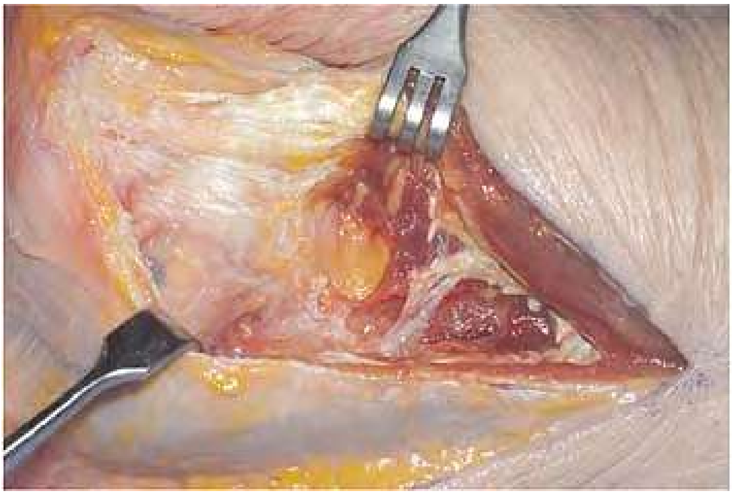

FIGURE 9-17 The extensor origin is dissected from the LCL.

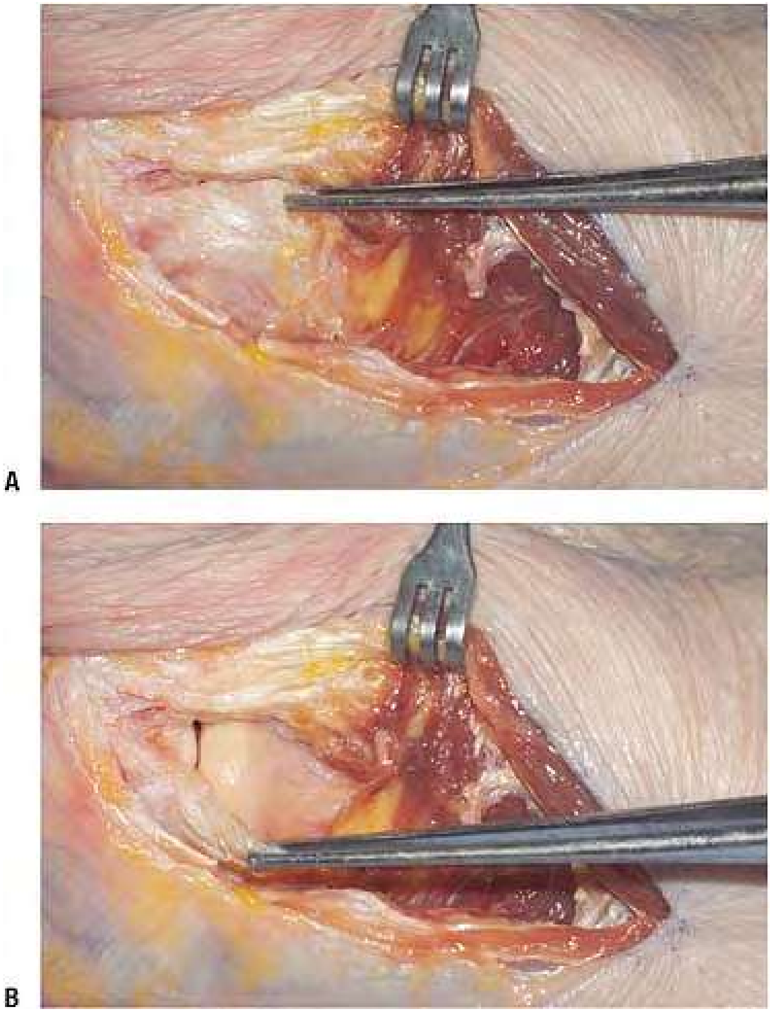

FIGURE 9-18 A,B: The capsule is incised anterior to the equator of the radial head in order to preserve the LCL complex.

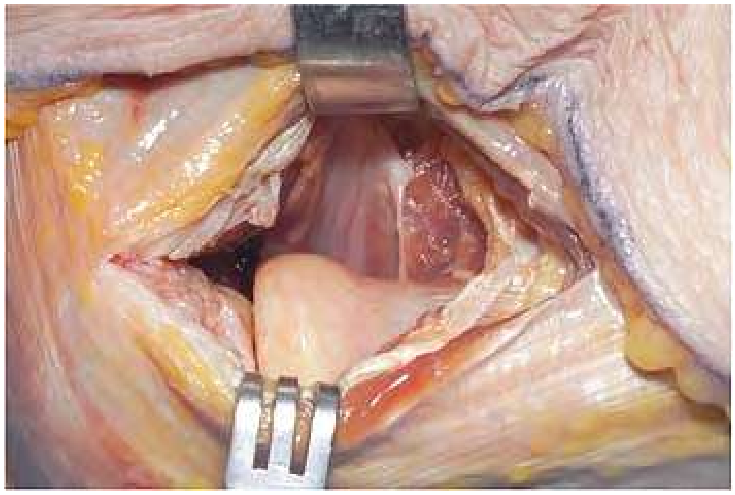

FIGURE 9-19 Excellent exposure of the radial head and neck is achieved. When retracting the anterior structures, pronating the forearm will reduce tension on the posterior interosseous nerve.

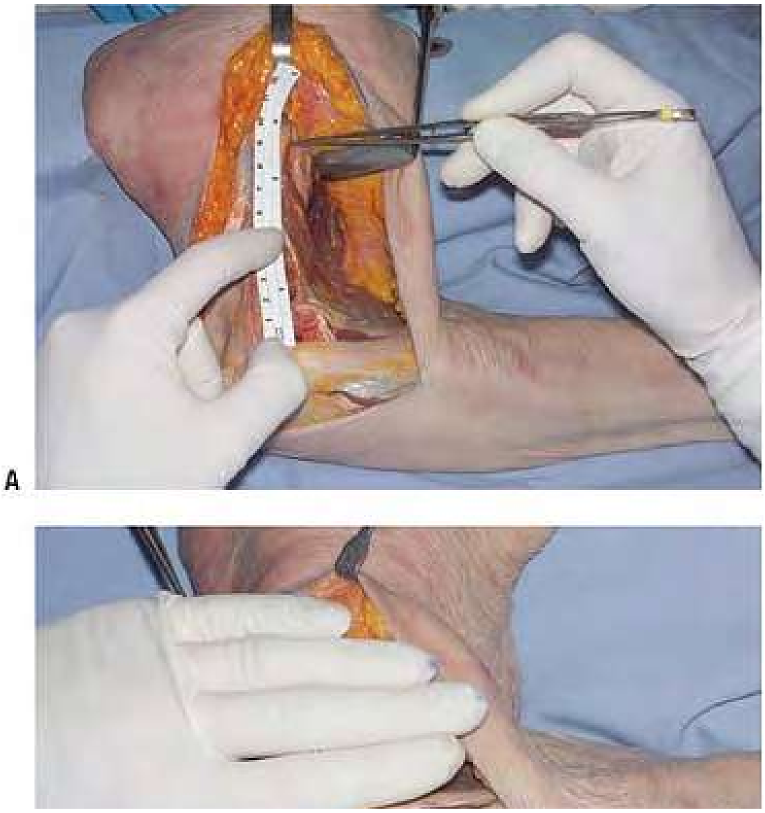

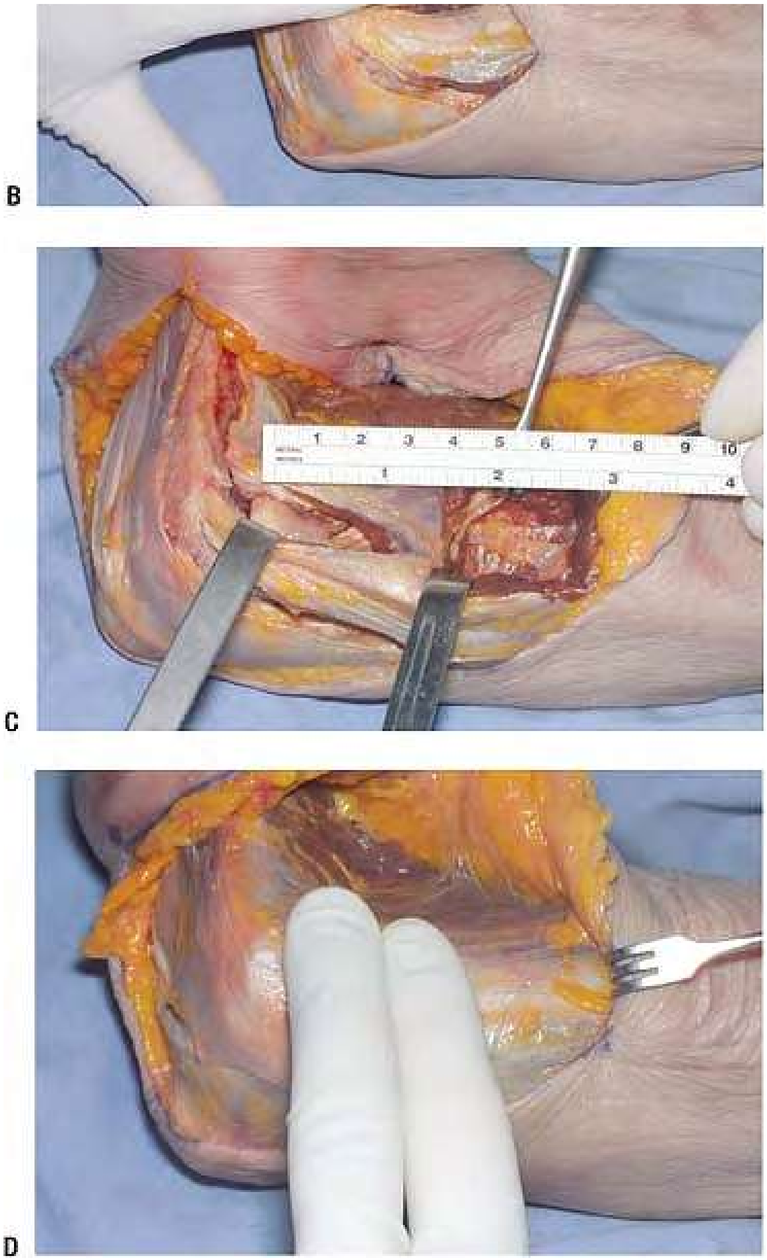

FIGURE 9-20 Aâ€“D: The radial nerve can be found proximally as close as 4 fingerbreadths (9 cm) proximally from the lateral epicondyle. The safe zone for the PIN is within 2 fingerbreadths (4 cm) distal of the radial head.

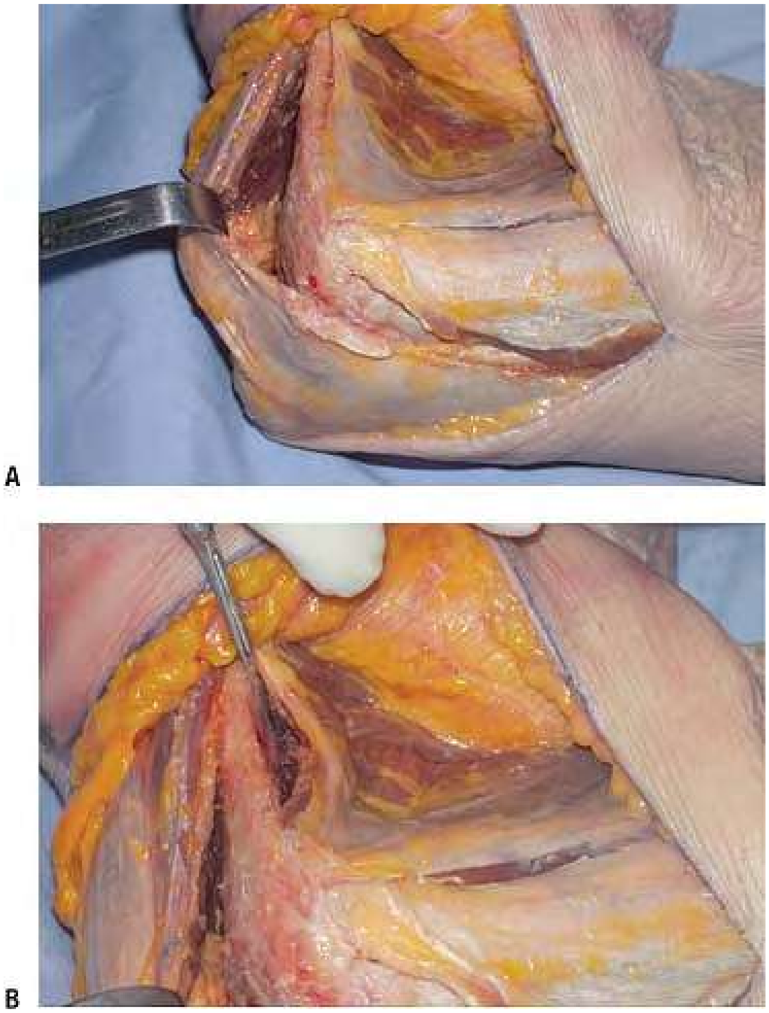

FIGURE 9-21 A,B: The lateral approach can be extended by continuing the dissection proximally on both the anterior and posterior side of the lateral humerus. Once again, one must be mindful of the LCL origin at all times.

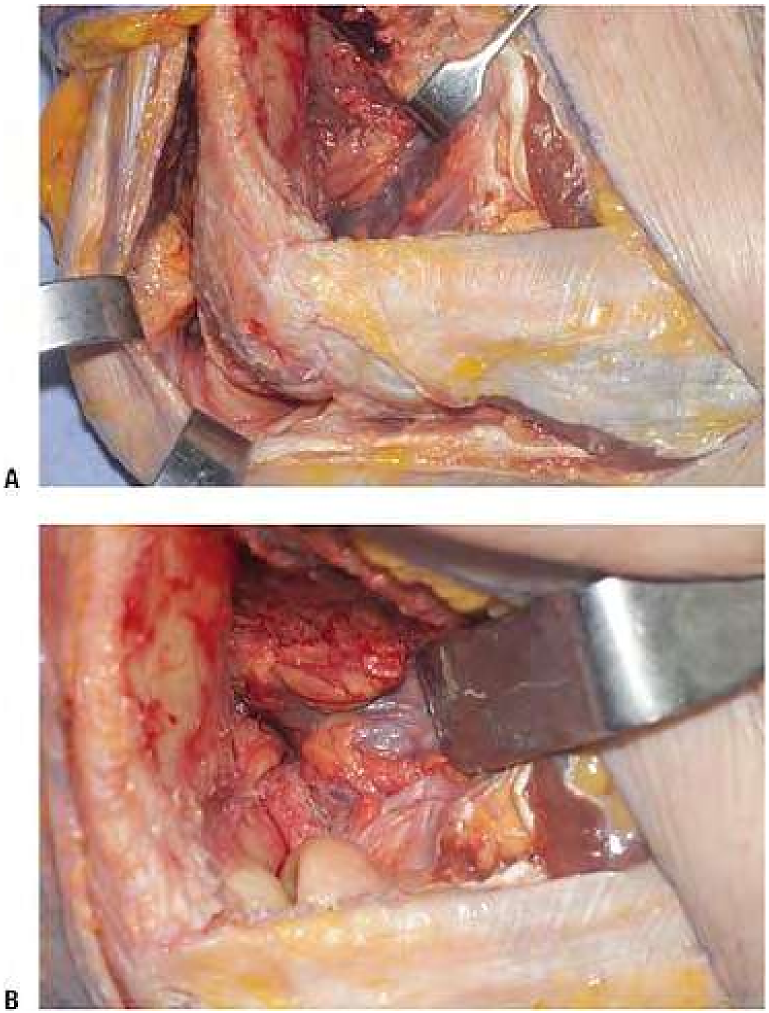

FIGURE 9-22 A,B: The brachioradialis is elevated anteriorly allowing for full visualization, even of the coronoid.

This exposure can then be extended both proximally and distally as in the case of a lateral ulnar collateral ligament reconstruction. The proximal dissection is achieved by elevating the common extensor tendon off the supracondylar ridge. The radial nerve is usually found proximally as close as 4 finger breadths (8 cm) above the lateral epicondyle, or on average (10 cm) in men from the articular surface. Distally, the posterior interosseus nerve will be found in the area of the radial neck about 2 finger breadths (4 cm) distal to the radiocapitellar joint.

Postoperative Management

Elbow rehabilitation is determined by fracture stability and concomitant ligamentous injury.

Complications

Flap necrosis, dehiscence, and wound infection are complications of any surgical incision surrounding the elbow. If the skin has been significantly traumatized prior to elbow exposure, skin slough and partial flap loss should be suspected and early wound examination should be performed to prevent wound infection. Many times partial skin loss may be treated with local wound care and dressing changes; however, if underlying vital structures are exposed, alternative means of wound closure must be considered. For exposed hardware, prosthesis, or nerve we favor aggressive wound debridement and early coverage with either the anconeus, a pedicled radial forearm fasciocutaneous flap, or latissimus flap.

Recommended Reading

Cohen MS, Hastings H. Posttraumatic contracture of the elbow. Operative release using a lateral collateral ligament sparing approach. *J Bone Joint Surg Br.* 1998;80(5):805â€“812.

Hotchkiss RN: Elbow contracture. In: Green DP, Hotchkiss RN, Pederson WC, eds. *Green's operative hand surgery*. Philadelphia: Churchill-Livingstone, 1999:667â€“682.

Oâ€™Driscoll SW. The triceps-reflecting anconeus pedicle (TRAP) approach for distal humeral fractures and nonunions. *Orthop Clin North Am*. 2000;31(1):91â€“101.

Ring D, Gulotta L, Chin K, et al. Olecranon osteotomy for exposure of fractures and nonunions of the distal humerus. *J Orthop Trauma*. 2004;18(7):446â€“449.

Ring D. Fractures of the coronoid process of the ulna. *J Hand Surg Am*.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 10

Radial Forearm Flap for Elbow Coverage

Kodi K. Azari

W. P. Andrew Lee

Indications/Contraindications

Soft tissue defects involving the posterior aspect of the elbow are not uncommon and can be challenging to manage. Tissue defects can be from trauma, burns, post-oncologic resection, pressure ulcers, extravasation injury, chronic bursitis, or chronic infection (4,7,15). In addition, elbow prosthetic devices can be exposed with devastating consequences. The goals of treating posterior elbow soft tissue defects is to provide wound closure, decrease the risk of infection, decrease edema, and allow the initiation of early rehabilitation (2,12). The soft tissue reconstruction must be aesthetically acceptable, durable, and elastic enough to allow for the constant unhindered movement of skin over the olecranon with elbow flexion and extension (7).

Many soft tissue reconstructive options exist and must be tailored to the needs of the patient's wound characteristics. Superficial wounds can be addressed by primary wound closure. Wounds with exposed â€œwhite structuresâ€• such as tendons, neurovascular structures, bone, and joint will necessitate flap coverage (9). Available flap options include regional muscle and musculocutaneous flaps, distant staged pedicle flaps, and microvascular free tissue transfer (8). Although these flaps are useful for elbow soft tissue coverage, they can carry significant morbidity. For example, regional muscle and musculocutaneous flaps necessitate the harvesting of a functional muscle; distant pedicle flaps (such as the groin or thoracoepigastric flap) require the binding of the extremity to the flank for several weeks with significant discomfort and ensuing stiffness; and free flaps introduce the added complexity of microsurgery and prolonged surgical time (4,8). Because of the liabilities of the previously mentioned flaps, local fasciocutaneous flaps have gained popularity (10), of which the proximally based radial forearm flap is the recognized workhorse for elbow soft tissue coverage (8).

The radial forearm flap is composed of the skin, subcutaneous fat, and fascia that, if needed, can be elevated to include the entire volar surface of the forearm (13) (see Fig. 10-2). This is a reliable flap with a rich arterial supply from the septocutaneous perforator branches of the radial artery (see Fig. 10-3). The deep venous drainage for the flap is provided by the paired venae comitantes that run parallel with the radial artery and the superficial venous drainage is by branches of the cephalic vein. The radial forearm flap is versatile because it may be transposed either in the radial or ulnar direction and can be made sensate by encompassing the cutaneous nerves of the forearm. Furthermore, this flap can provide stable soft tissue coverage while still preserving elbow range of motion.

Contraindications to the use of the radial forearm flap include any injury to the radial artery, severe forearm soft tissue injuries, and inadequate collateral flow to the hand and thumb. A possible relative contraindication is recent cannulation of the superficial venous system of the upper extremity, as this may result in venous thrombus formation in the flap and subsequent venous congestion (5).

Preoperative Planning

Preparation of the Wound Bed

In cases of complex elbow wounds, it is extremely important to gain control over the wound before definitive soft tissue reconstruction. Osseous injuries and dislocations need to be stabilized with appropriate internal or external fixation and ligament reconstruction. When soft tissue injuries are present, it is mandatory to perform serial debridement of devitalized tissues until only viable tissues remains. Infections require appropriate debridement and culture-specific antibiotic coverage. It should be noted that meeting the above requisites as quickly as possible will allow for earlier definitive vascularized wound coverage with radial forearm flap and thus afford a more successful functional outcome.

Assessment of the Arterial Vascularity of the Hand

It is imperative to assess the arterial vascularity of the hand by performing a modified Allen's test before performing the radial forearm flap. A modified Allen's test is performed by occluding the radial artery at the wrist crease and evaluating the perfusion of the digits using a hand-held pencil Doppler. In particular, one must ensure that there are Doppler signals to the thumb digital arteries once the radial artery is occluded. If the modified Allen's test shows evidence of vascular insufficiency, then one must be prepared to reconstruct the arterial tree using a saphenous vein graft or choose a soft tissue reconstruction technique other than the proximally based radial forearm flap.

Surgery

Patient Positioning

For optimal position, patients are placed supine on the operating room table. The entire upper extremity from axilla to hand is prepped and draped in standard surgical fashion, and a sterile proximal arm tourniquet applied.

Technique

The course of the radial artery is marked by drawing a line from the center of the anticubital fossa to the radial border of the proximal wrist crease (where the radial artery pulse is palpable) (Fig. 10-1). A skin island that is the shape and slightly larger than the periolecranon defect is designed and centered along the central axis of the radial artery (Fig. 10-2). The limit of the flap width is the radial and ulnar borders of the volar forearm and can extend from the antibrachial fossa to a few centimeters proximal to the distal wrist crease. The exact position of the flap is determined by a using a nonstretchable template with the antibrachial fossa as the pivot point of the vascular leash (Fig. 10-3).

Usually, the distal incision is made first and the radial artery and accompanying venae comitantes that reside between the flexor carpii radialis (FCR) and brachioradialis (BR) tendons are isolated and divided. The cephalic vein, which lies radial to the brachioradialis tendon, is also isolated and divided.

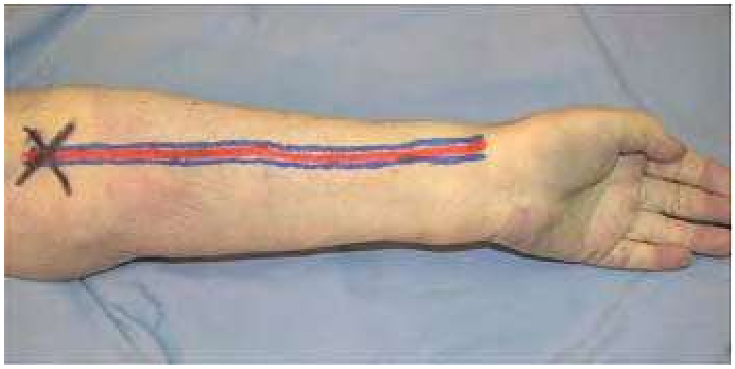

FIGURE 10-1 The course of the radial artery and venae comitantes is marked by drawing a line from the center of the anticubital fossa to the radial border of the proximal wrist crease.

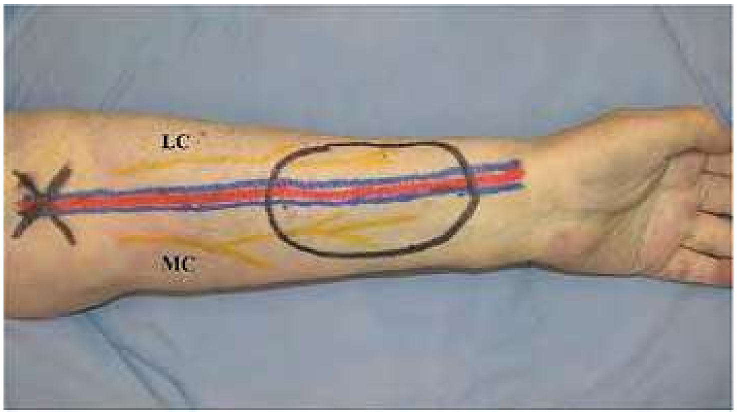

FIGURE 10-2 A skin island that is the shape and slightly larger than the periolecranon defect is designed and centered along the central axis of the radial

artery. The medial cutaneous (MC) or lateral cutaneous (LC) nerves of the forearm can be included in the skin paddle to form an innervated neurosensory flap.

The proximal incision of the flap is made and the radial artery is identified between the FCR and BR. Care should be taken not to injure the cephalic vein (or the medial or lateral cutaneous nerves if the radial forearm flap is designed as an innervated neurosensory flap) (see Fig. 10-2). Next, the ulnar incision is made and carried to the tissue plane deep to the deep forearm fascia. This subfascial plane is developed and the flap is raised from ulnar to radial off the underlying flexor carpii ulnaris, flexor digitorum sublimus, palmaris longus, and flexor carpii radialis muscles. It is imperative to preserve the paratenon of the tendons to allow skin graft coverage of the flap donor site. The dissection is carried ulnarly until the radial border of the FCR. At this important juncture the intermuscular septum carrying the septocutaneous branches of the radial artery is identified and carefully preserved (Fig. 10-4).

The radial border incision is made and the subfascial dissection is carried ulnarly. The brachioradialis muscle is retracted laterally allowing for the identification of the radial artery, intermuscular septum, and superficial radial sensory nerve. In the upper border of the flap, the cephalic vein is located and integrated in the flap for added venous drainage. The flap is next elevated and the radial artery, venae comitantes, cephalic vein, and medial or lateral cutaneous nerves (for an innervated sensory flap) are dissected proximally to an appropriate pivot point that allows for a tension free flap transfer (Fig. 10-5).

The tourniquet is released and the vascularity of flap is ascertained. In cases where large flaps are required, the perfusion of the peripheral flap skin can be evaluated by the administration of intravenous fluorescein. Initially a small test dose of 100 mg is given, and the patient's vital signs are closely monitored for 20 minutes. If there is no evidence of hemodynamic instability, then the full dose of 10 mg/kg is administered and the skin paddle evaluated 10 minutes later under ultraviolet light (Wood's lamp). The extreme areas of the flap that appear dark and do not fluoresce are sharply trimmed.

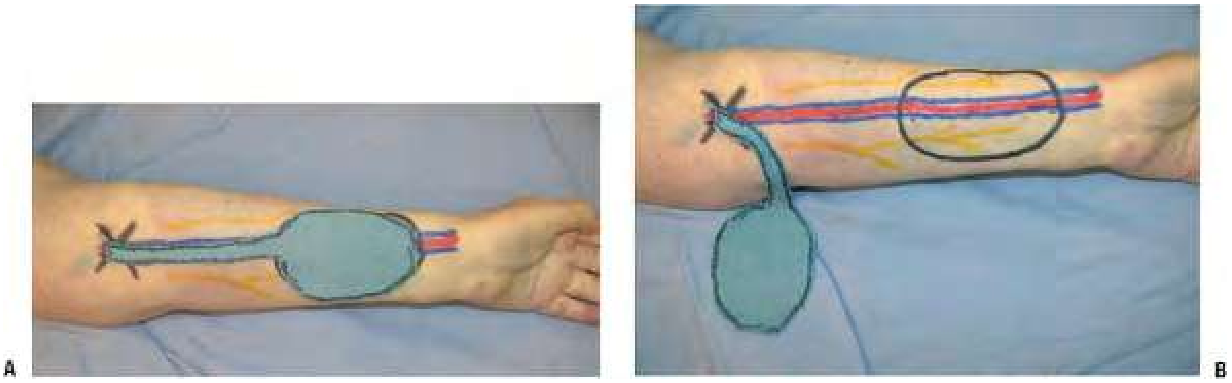

FIGURE 10-3 A,B: The exact position of the skin paddle is established by a using a nonstretchable template with the antibrachial fossa as the pivot point of the vascular leash.

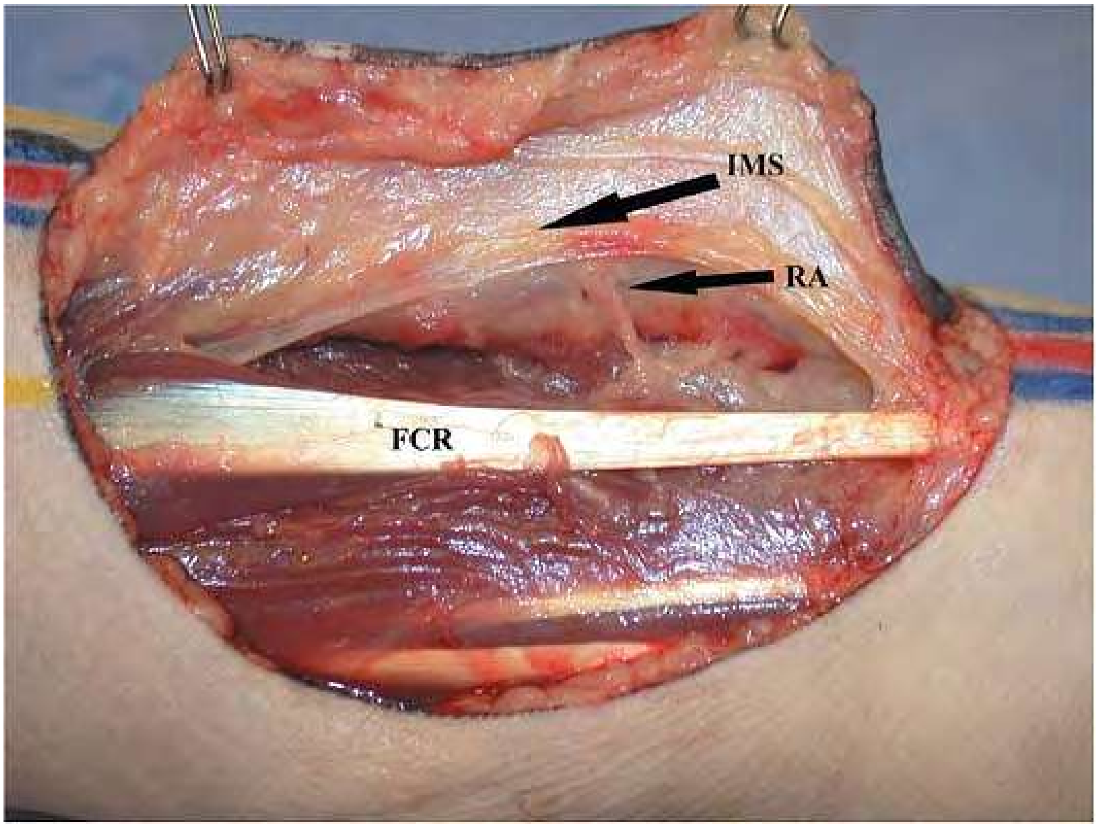

FIGURE 10-4 View from an ulnar to radial subfascual dissection demonstrating the radial artery (*RA*), intermuscular septum (*IMS*) through which septocutaneous perforator branches of the radial artery traverse, and flexor carpii radialis tendon (*FCR*).

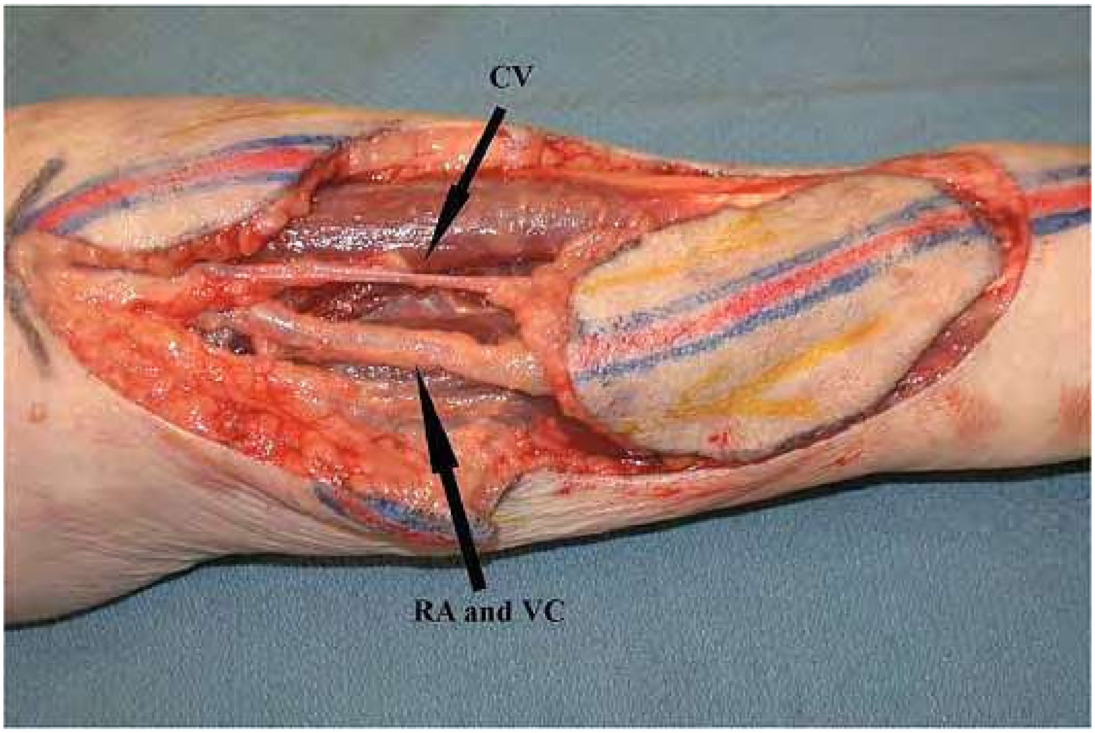

FIGURE 10-5 This figure demonstrates a flap that has been dissected proximally to an appropriate pivot point that will allow a tension free flap transfer. Note the vascular pedicle containing the radial artery (*RA*) and venae comitantes (*VC*). The cephalic vein (*CV*) was included in this flap for venous drainage augmentation.

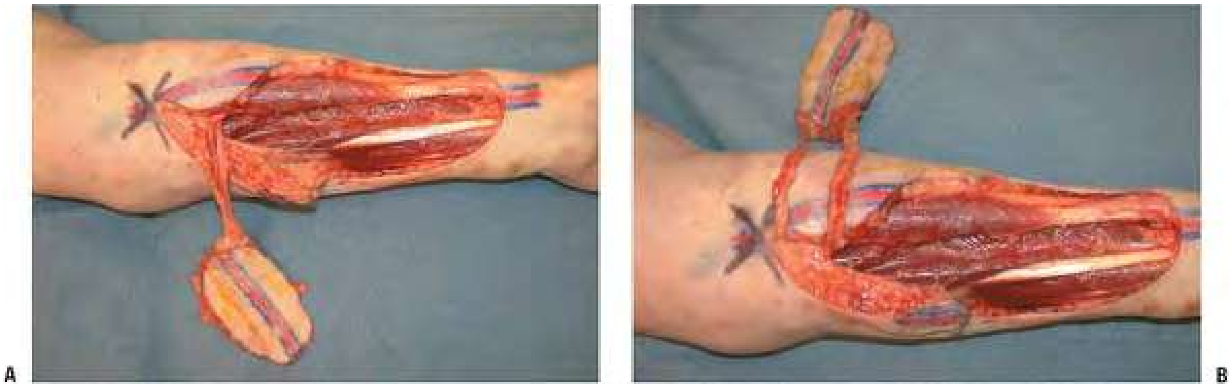

FIGURE 10-6 Once the vascular pedicle is dissected to an appropriate pivot point, the flap can be transposed to the posterior elbow region from either an ulnar (A) or radial (B) direction.

When certain that the flap is well perfused, the flap is transferred from either an ulnar or radial direction through a subcutaneous tunnel to the elbow and inset with interrupted suture (Fig. 10-6). After obtaining meticulous hemostasis, we routinely attempt to cover the exposed tendons with local flexor tendon muscle bellies and soft tissues. This is followed by unmeshed split-thickness skin graft coverage of the entire donor site (Fig. 10-7).

A case of a proximally based neurosensory radial forearm flap for soft tissue reconstruction of the periolecranon region is illustrated in Figure 10-8.

Postoperative Management

Immobilization and elevation are absolutely critical in the postoperative management of elbow reconstruction with the radial forearm flap. A nonconstricting and bulky above-elbow splint (including fingers) is applied with a window cut out for accessibility and inspection of the flap. The elbow is held at 90 degrees, and the digits are placed in the position of safety. The digits are immobilized to prevent shear forces on the forearm donor site skin graft by tendon excursion. To help decrease postoperative edema and the possibility of the patient exerting pressure on the flap while asleep, the arm can be hung from an intravenous fluid pole. For the first 3 postoperative days, the flap is evaluated on a frequent (every 4 to 8 hr) basis for capillary refill and the absence of venous congestion. The splint and the donor site bolster are removed on postoperative day 5 when capillary inosculation of the skin graft has occurred. Thereafter, a soft dressing is applied and the skin graft and flap suture lines are kept moist with bacitracin ointment. Digital and elbow range of motion are gently begun and progressively increased.

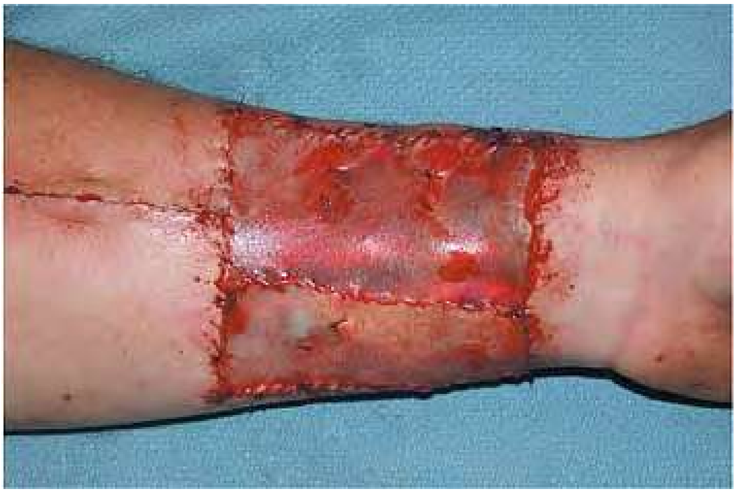

FIGURE 10-7 The forearm donor site is covered with an unmeshed split-thickness skin graft.

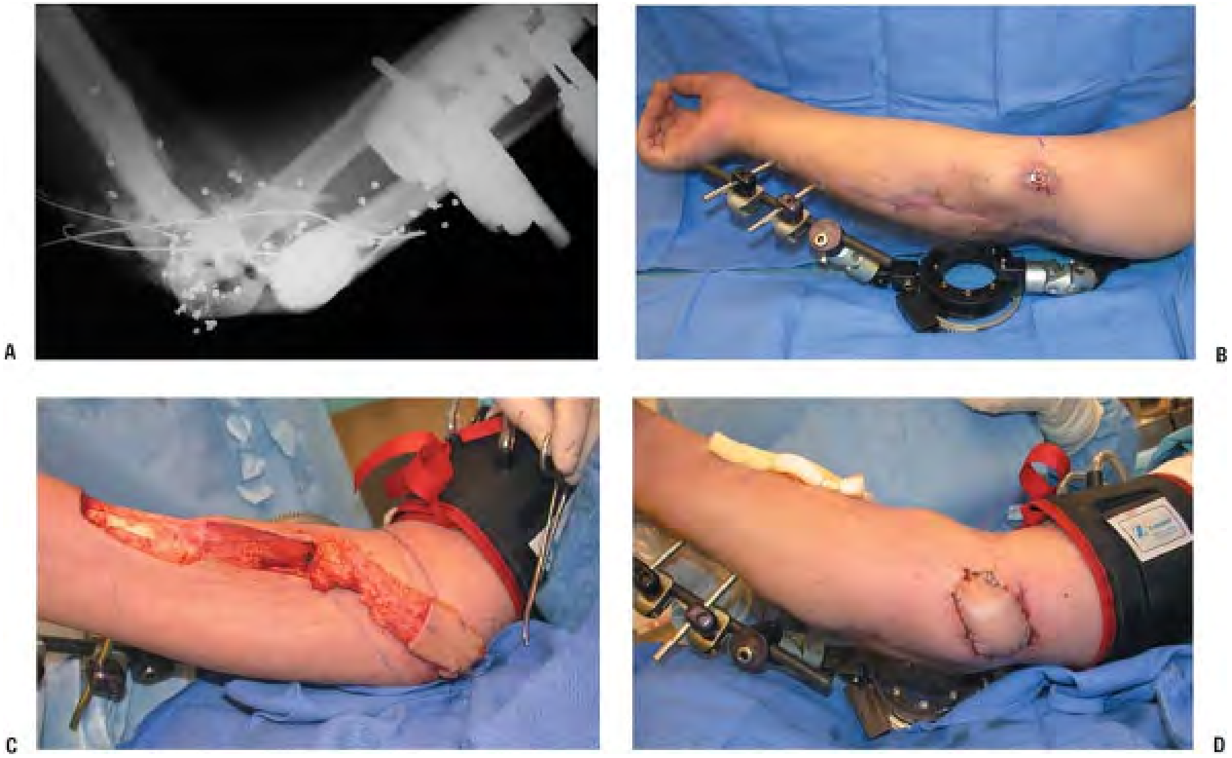

FIGURE 10-8 A 32 year old male with significant injury to the elbow from a shotgun assault. Following appropriate bone stabilization with external and internal fixation (A) he developed wound dehiscence and hardware exposure (B). An innervated neurosensory radial forearm flap was designed incorporating the medial cutaneous nerve of the forearm (C) and inset through a subcutaneous tunnel (D).

Results

Case reports of soft tissue reconstructions about the elbow region can be found in the medical literature. Hallock (6) described two cases of chemotherapy extravasation and exposure of hardware with successful healing following radial forearm flap. Thornton (13) reported a case of olecranon osteoradionecrosis treated with the radial forearm flap resulting in early elbow range of motion and complete healing. Small and Miller (11) present three cases in elderly patients to illustrate the utility of the radial forearm flap. One patient had an established infection with radial head and lateral epicondyle ischaemia. Despite aggressive treatment this patient died 32 days postoperatively from sepsis. Another patient with a crush injury developed an ankylosed elbow that was free of pain and infection. Meland (7) reported successful pedicled radial forearm flap coverage for three patients with recalcitrant about the elbow defects in whom one or more previous flaps had failed. Tizian (15) reported the use of proximally based radial forearm flap in 14 patients. All patients showed complete healing of the periolecranon region and forearm donor site with unrestricted elbow range of motion. In a recent comparative review by Chondry and colleagues, the radical forearm flap provided excellent coverage of moderate size defects of the elbow. The flap was associated with fewer complications when compared to the pedicled latissimus dorsi flap (1).

Complications

A major, yet rarely encountered complication of the radial forearm flap is ischemia of the hand. A meticulous preoperative Doppler examination of the hand using the modified Allen's test should identify patients who are at risk. However, if arterial insufficiency of the hand is encountered once the flap is elevated, the radial artery needs to be reconstructed using an interposed vein graft. A great saphenous vein graft anastomosed end-to-side proximally and end-to-end distally is the authors preferred method.

Other major complications include flap ischemia and congestion. In both of these instances careful evaluation of the flap is mandatory to ensure that the subcutaneous tunnel is adequate, the skin paddle is not under excessive tension or external pressure, and there is no undue stretching or twisting of the vascular pedicle.

Donor site problems are one of the most common complications of the radial forearm flap. These complications can include hypertrophic scarring, poor skin graft take, and tendon loss (3,14). As mentioned previously, it is imperative to preserve the flexor tendon paratenon and attempt to cover exposed tendons such as the FCR with local muscle bellies and soft tissue advancement. If there is inadequate skin graft take and tendon exposure, local dressing changes with saline-dampened gauze are initiated to prevent tendon desiccation until the wound healing by secondary intention is complete. In rare cases, secondary flaps may be required to cover exposed tendons in the forearm. Other reported donor site complications can include injury to the radial sensory nerve or hand swelling and stiffness (14).

Conclusion

Soft tissue defects of the posterior elbow are not uncommon and can often be difficult to reconstruct. Although a multitude of reconstruction options are available, each patient needs to be individualized. The advantages of the pedicled radial forearm fasciocutaneous flap for the management of elbow defects include ipsilateral donor site and scar, relatively simple dissection, long vascular pedicle that allows a wide arc of rotation, ability to make into a neurosensory flap by including forearm cutaneous nerves, no requirement for microsurgical expertise, ease of postoperative limb positioning, and highly reliable and vascularized tissue that is an excellent option even in the setting of osteomyelitis (7).

Recommended Reading

Mathes S, Nahai F. *Reconstructive Surgery: Principles, Anatomy, and Technique.* New York: Churchill Livingstone, 1997.

References

1. Chondry AH, Moran SL, Li S, et al. Soft tissue coverage of the elbow: an outcome analysis and reconstructive algorithm. *Plast Reconstr Surg*. 2007;119:1852.

2. Davalbhakta AV, Niranjan NS. Fasciocutaneous flaps based on fascial feeding vessels for defects in the periolecranon area. *Br J Plast Surg.* 1999;52:60.

3. Fenton OM, Roberts JO. Improving the donor site of the radial forearm flap. *Br J Plast Surg.* 1985;38:504.

4. Frost-Arner L, Bjorgell O. Local perforator flap for reconstruction of deep tissue defects in the elbow area. *Ann Plast Surg.* 2003;50:491.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 11

Pedicled and Free Latissimus Flap for Elbow and Forearm Coverage

Kenji Kawamura

Kevin C. Chung

Indications/Contraindications

The latissimus dorsi (LD) has proven to be a reliable muscle in the coverage of soft tissue defects about the shoulder and elbow. Familiarity with this muscle and its application in cases of trauma is essential for all upper extremity surgeons. The major arterial inflow and venous outflow for the LD flap is based on the thoracodorsal artery and venae comitantes. The average diameter of the thoracodorsal artery is 2.5 mm, and that of venae comitantes is 3.0 mm. The average length of maximum vascular pedicle is 12 cm, which can be obtained by dissecting the thoracodorsal vessels toward the proximal axillary artery and vein.

The indication for the use of the LD flap to the elbow and forearm is to cover a large skin and soft tissue defect that cannot be managed with local flaps. In cases in which vital muscle structures were damaged, the LD flap can provide coverage of soft tissue defects, as well as functional muscle transfer to restore of elbow, wrist, and finger motion. Unstable soft tissue coverage over elbow fracture or recalcitrant infection around the elbow may also be amenable to the LD flap because this flap provides well-vascularized tissue that can seal dead spaces and increase blood flow to the local environment.

The pedicled LD flap can cover the forearm up to approximately 8 cm distal to the olecranon when transposed anteriorly, and about 6 cm distal to the olecranon when transposed posteriorly. Soft tissue defects over 8 cm distal from the olecranon are not suitable for the pedicled LD flap transfer, and if this muscle is to be used, the free LD flap transfer should be considered.

Contraindications for flap use include previous injury to the muscle or pedicle, such as in cases of previous thoracotomy, axillary arterial injury, or in some cases, axillary lymphadenectomy. Breast cancer surgery may injure the nerve or arterial supply to the LD muscle, rendering it fibrotic and inadequate for transfer. In such patients, palpation of a contracting LD muscle usually verifies an uninjured nerve and vascular pedicle.

The LD muscle functions as an expendable adductor, extender, and internal rotator of the arm. These functions are essential for activities of daily living for patients with contralateral shoulder girdle paralysis or extremity paralysis from spinal cord injury. One must be careful in evaluating the impact of removing the LD muscle in these situations. In general, one should not use the LD muscle for soft tissue coverage in these patients because of functional requirements.

Preoperative Planning

The LD flap can be harvested as not only a muscle flap but also a musculocutaneous flap. Muscle flap covered with skin graft is less bulky and can seal deep defects, however, musculocutaneous flap gives better aesthetic reconstruction because the skin paddle can conform to the skin texture of the upper arm. When a LD musculocutaneous flap is applied, it is recommended that the skin paddle is designed several centimeters superior to the muscle origin at the iliac crest to avoid the risk of skin paddle necrosis because the blood supply of the subcutaneous tissue near the origin is inconsistent. When greater flap reach is required, a portion of the skin paddle may be designed over the inferior margin of the muscle. This technique can facilitate a greater distal coverage of the forearm when the pedicled LD musculocutaneous flap is rotated. The entire skin paddle can survive if enough musculocutaneous perforators are included in the proximal portion of the skin paddle design.

The arc of transposition of the LD flap depends on the location of the wound. When defects are present in the posterior aspect of the elbow and forearm, posterior arc transposition of the pedicled LD flap is recommended. Similarly, when defects are present in the anterior aspect, anterior arc transposition is recommended.

It is important to assess the strength of the LD muscle preoperatively when the LD flap is used for simultaneous soft tissue coverage and functional restoration. When a soft tissue defect is caused by a trauma, particularly associated with a motorcycle accident, the LD muscle may be paralyzed by brachial plexus injury. If the paralyzed LD muscle is planned to be used as a functional transfer, reinnervation of the LD muscle by another motor nerve, for example, intercostal nerve or accessory nerve, can be performed during operative procedures. LD muscle previously paralyzed by neurologic disorders can only be used for soft tissue coverage and should not be used for functional muscle transfer.

Preoperative angiography of the injured upper extremity will help plan for free LD flap transfer. The free flap transfer may require a more proximal vessel dissection, vein grafting, and end-to-side anastomoses in cases with a potential disruption of the recipient vessels.

For aesthetic consideration at the donor site, it is important to ask for the patient's clothing preference preoperatively for planning the skin incision. For example, a transverse incision along the bra line is recommended to conceal the donor site scar for a woman, but an oblique incision may be suitable for a patient who likes to wear backless dresses. Furthermore, the endoscopic harvesting technique may be preferred for children, those who are prone to hypertrophic scars, and for other aesthetic considerations.

Surgery

The patient, under general anesthesia, is placed in the lateral decubitus position. The injured upper limb and ipsilateral hemithorax caudal to the iliac crest are prepped and draped. The injured upper limb is supported on a sterile Mayo stand with a 90-degree abduction position of the shoulder (Fig. 11-1A). The lateral buttock and thigh are also prepared for skin grafting if necessary. A beanbag or axillary roll with padding is placed on the opposite axilla to protect the dependent shoulder in the decubitus position, and all other pressure points are also carefully padded.

In cases of traumatic defects, the wounds are again thoroughly debrided before the flap procedure is performed. When a LD musculocutaneous flap is applied, the skin paddle that includes the perforators arising from the branch of the thoracodorsal artery is designed on the LD muscle. There is a longitudinal row of perforators 2 to 5 centimeters from the lateral border of the LD muscle, which can be easily identified with a Doppler blood flowmeter. The largest skin paddle that can be moved safely is 20 Ã— 15 cm, however, primary closure is only possible with skin paddles 8 cm in width.

The procedure is performed with Ã—2.5 loupe magnification. The incision usually begins at the posterior aspect of the axilla angled along the anterolateral border of the latissimus dorsi muscle and extends inferiorly to the iliac crest. An aesthetic incision in a woman is designed in a transverse incision along the bra line. The LD muscle is dissected medially to the paraspinous muscle and caudally to the lumbosacral fascia. The origin of the muscle is detached from the thoracic and lumbar vertebra, posterior ribs, and lumbosacral fascia. The flap is elevated superiorly toward the axilla to separate the undersurface of the muscle from the chest wall. Perforators from the intercostal and lumbar arteries are clipped. The neurovascular pedicle, which includes the thoracodorsal artery, venae comitantes, and nerve, is identified more proximally on the undersurface of the muscle. The thoracodorsal

vessels are carefully dissected toward the proximal axillary artery. The circumflex scapular vessels and the serratus anterior and teres major branches are ligated to allow maximal pedicle length. Leaving the thoracodorsal nerve intact may be considered when the flap is transferred on the pedicle, even though the flap is not used as a functional transfer. Retaining innervation minimizes postoperative muscle atrophy and will maintain muscle thickness and durability. This consideration is particularly important for distal elbow coverage when the thin muscle near the origin of the LD muscle at the iliac crest maintains its bulk.

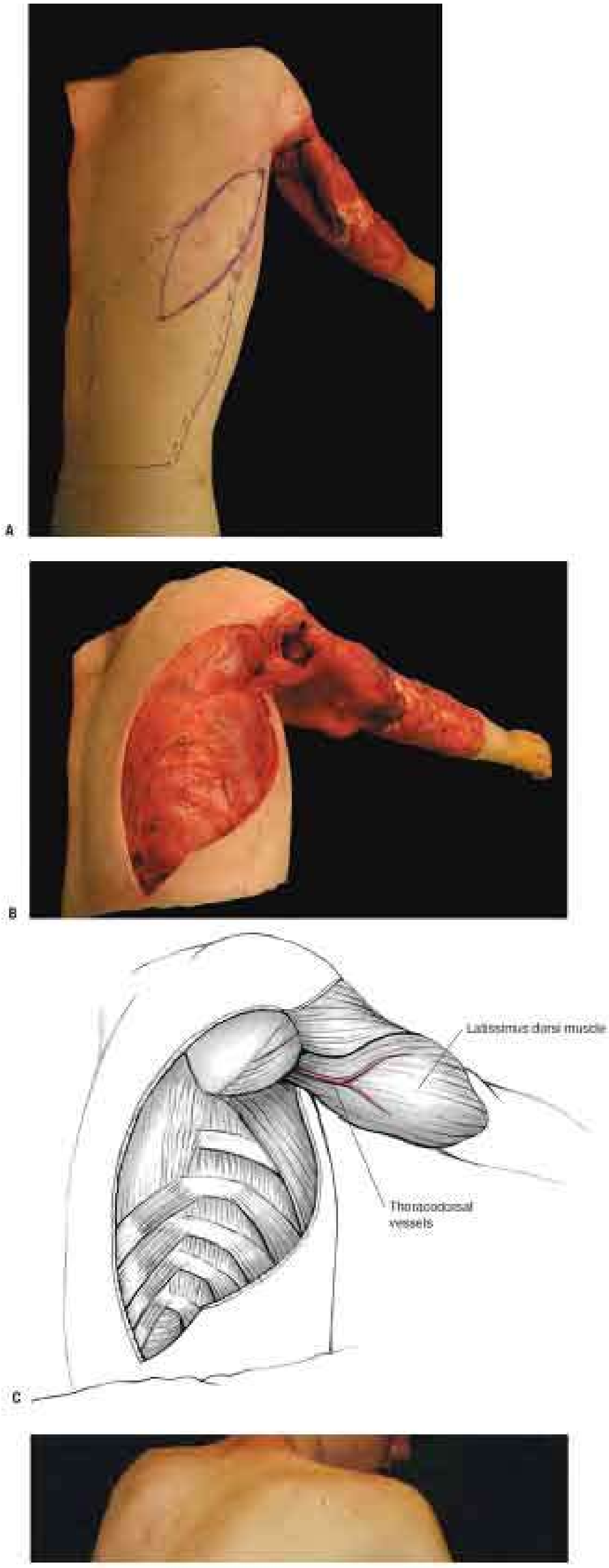
A
B
Latissimus dorsi muscle
Thoracodorsal vessels
C

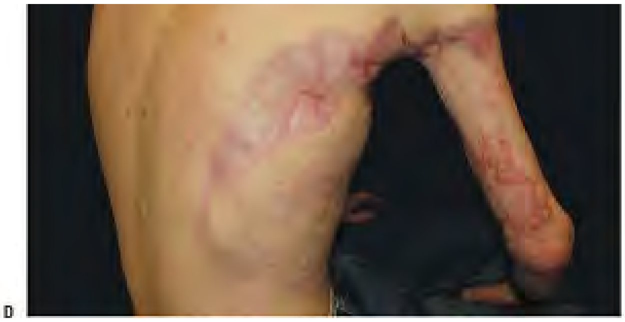

FIGURE 11-1 A: The patient was placed in the lateral decubitus position with the injured upper limb on a sterile Mayo stand with a 90-degree abduction position of the shoulder. B: A pedicled latissimus dorsi muscle flap was harvested and rotated posteriorly to the upper arm. A skin paddle was excised for contouring the flap on the arm. It is difficult to design a skin paddle to fit the wound geometry and it is simpler to just use the muscle only and cover with skin graft. C: Schematic drawing of a pedicled latissimus dorsi muscle transfer based on Figure 11-1B. D: Complete wound healing at 3 months postoperatively.

After the dissection of the neurovascular pedicle has been completed, the insertion of the LD muscle at the humerus is divided. Additional distal coverage with the pedicled LD flap can be obtained

by releasing its insertion from the intertubercular grooves of the humerus. When the LD flap is transferred as a free flap, vascular pedicle is divided at the juncture with the axillary artery and vein to obtain the maximum pedicle length. The thoracodorsal nerve is also divided at its proximal point when the flap is used as a functional transfer. When the LD flap is transposed to the elbow and forearm defects, the muscle can be positioned along its path by incising intervening skin or can pass under a subcutaneous tunnel (see Fig. 11-1B,C). The subcutaneous tunnel may be useful in situations when the defect size is relatively small and the transposed small muscle can be tailored to pass easily through the tunnel. But for large muscle, the subcutaneous tunnel may compress the vascular pedicle and the intervening skin should be incised to allow a tension-free passage.

The muscle is sutured to the subcutaneous tissue in the proximal portion of the wound to prevent tension on the pedicle. For functional muscle transfer, muscle tension is critical. For elbow flexor reconstruction, the length of the transferred muscle is adjusted so that the elbow remains at 100 degrees flexion and the forearm is held in complete supination after both ends are sutured. By extending the muscle with strips of fascia lata, finger flexion or extension can also be restored.

Split-thickness skin graft is applied as needed for the transposed muscle or over the donor site (see Fig. 11-1D). Two silastic drains are placed at the donor site, and one is placed below the transposed muscle. Before closure of the donor site, fibrin sealant may be used to reduce seroma formation over the back.

When a free LD flap transfer is used for coverage of a forearm defect (Fig. 11-2A), dissection of recipient vessels should be performed before harvesting the flap to confirm the necessary length of the pedicle. A large LD muscle flap is suitable to cover a forearm defect because the flat configuration of the LD muscle allows it to be wrapped around the forearm. The LD muscle is a versatile muscle for coverage of a complex wound whereby the geometry of the wound requires two separate flaps. Splitting of the LD muscle is helpful in this situation. The thoracodorsal vessels usually bifurcate into medial and lateral branches just after entering the LD muscle. The LD muscle can be split longitudinally into halves based on these two branches. Splitting the LD muscle provides two flaps to coverage different wounds on the forearm (see Figs. 11-2B,C).

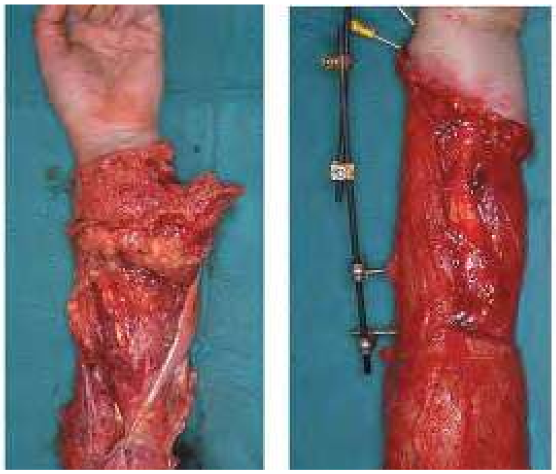

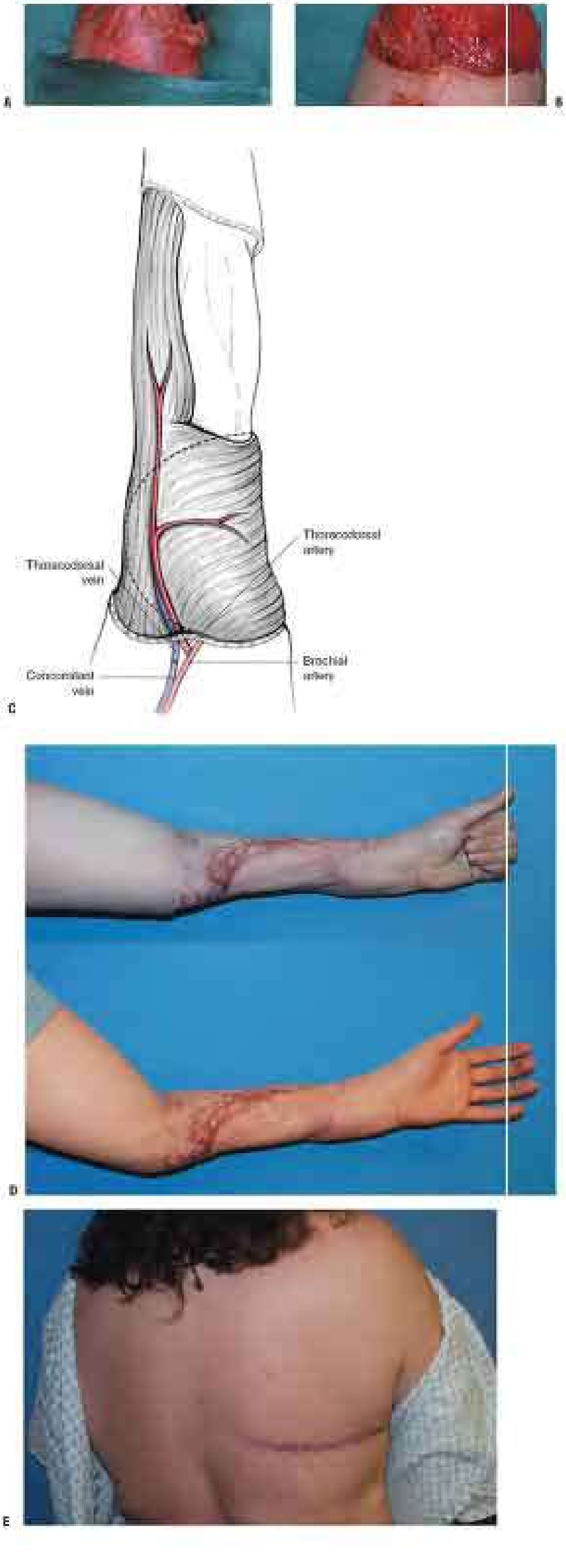
A
B
Thoracodorsal artery
Thoracodorsal vein
Concomitant vein
Brachial artery
C
D
E

FIGURE 11-2 A: Crush injury of the forearm. B: The latissimus dorsi muscle was split longitudinally into halves by identifying the arterial pedicles. One half was used to cover the soft tissue defect over the posterior elbow, and the other half was used to cover the exposed bone at the wrist and exposed tendons. The thoracodorsal artery was anastomosed to the brachial artery using an end-to-side technique, and the thoracodorsal vein was anastomosed to the brachial concomitant vein with an end-to-end technique. C: Schematic drawing of a free latissimus dorsi muscle transfer based on Figure 11-2B. D: Complete wound healing at 4 months postoperatively. There are no limitations of finger motions. E: Donor site scar resulting from an aesthetic transverse incision along the bra line.

Postoperative Management

During the first 48 hours after the operation, flap circulation is closely monitored by direct observation of the flap, which includes color, temperature, capillary refilling time, and bleeding of the flap. A Doppler blood flowmeter is useful to examine the arterial flow into the flap. If there is concern about the arterial inflow or venous outflow, the pedicle should be reexamined and the wound reexplored. Kinking and tension on the pedicle must be relieved to prevent total necrosis of the muscle. Postoperative hematoma can also compromise venous outflow.

The drain at the donor site must remain in place for 2 to 3 weeks to prevent seroma formation. Application of fibrin sealant in the donor site may reduce seroma formation, but the drains should remain in place until drainage is minimal, usually less than 25 mL/day. Compression of the donor site by thoracic elastic bandages may be effective to enhance sealing of the wound. When a seroma is diagnosed, needle aspiration is performed. Small seromas (<100 mL) may resorb spontaneously, but a large seroma needs serial aspirations.

The elbow is immobilized at 90 degrees of flexion, and the shoulder is adducted for 3 weeks after the operation to avoid wound dehiscence and tension on the pedicle. During this period, finger and wrist exercises are actively practiced (see Fig. 11-

2D). After splint removal, elbow and shoulder range of motion exercises are started. The goal of physiotherapy is to recover the full range of motion of the elbow and shoulder.

Complications

Donor site seroma or hematoma is the most commonly encountered complication at the donor site following harvest of the LD flap. Seromas can be relieved with frequent needle aspirations, but hematomas must be drained immediately to avoid secondary donor site complications relating to wound dehiscence or skin necrosis.

Scarring over the donor site is inevitable, but can be easily covered by clothing. Planning the skin incision and endoscopic harvesting technique can reduce back scar problems (see Fig. 11-2E).

Total flap necrosis is a rare complication with the LD flap method. However, partial flap necrosis at the distal portion may occur because of inconsistent blood supply to the lower third of the LD muscle. Bleeding at the distal end of the LD flap should be checked when the flap is elevated. Kinking and tension on the pedicle can cause disturbance of flap circulation and must be recognized immediately. It is also important to note potential ischemia of the transferred muscle caused by subcutaneous tunnel compression.

Power and endurance of shoulder extension and adduction may be weakened in the absence of the LD muscle. The weakness is not noticeable in most patients and will recover within a few months after removal of the LD muscle because the synergistic shoulder muscles such as teres major, pectoralis major, and subscapularis substitute the function of the LD muscle.

Hyperabduction of the arm during the operation may cause iatrogenic brachial plexus injury, which is reversible within several weeks. Care must be taken to avoid prolonged shoulder abduction during the operative procedure.

Lumbar hernia after the latissimus dorsi flap is an extremely uncommon complication and it may be misdiagnosed as lumbar seromas. Lumbar hernia can be avoided by preserving of the fascia underlying the distal latissimus dorsi aponeurosis.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part II - Management of Soft Tissues Within the Upper Extremity > Chapter 12 - Fasciotomies for Forearm and Hand Compartment Syndrome

---

Chapter 12

Fasciotomies for Forearm and Hand Compartment Syndrome

Jeffrey B. Friedrich

Alexander Y. Shin

Indications/Contraindications

Indications

The chief indication for forearm and/or hand fasciotomies is essentially any suspicion that the patient has a forearm or hand compartment syndrome. The diagnosis of upper extremity compartment syndrome is made with a combination of clinical examination and objective diagnostic measurements. The clinical examination findings for upper extremity compartment syndrome traditionally have been taught as the five â€œP'sâ€• which include pain in the forearm and hand, increased pain with passive extensive extension of the fingers, pallor, pulselessness, and paresthesias. However, one must note that not all of these criteria need be present to determine that a patient has compartment syndrome. In fact, pulselessness can be a very late or nonexistent finding in the patient who has a compartment syndrome. Therefore, reliance on these five P's should also be balanced with other clinical examination findings as well as objective diagnostic findings. In addition to the criteria mentioned, the forearm and hand suffering from compartment syndrome will typically be very tense and swollen and can often have signs of trauma from the inciting injury, if that is the cause of the compartment syndrome. Skin changes such as blistering and a â€œshineâ€• due to significant swelling are common findings as well (Fig. 12-1).

Compartment syndrome can be caused by extrinsic pressure such as casts or splints, and this possibility should be factored into the diagnostic algorithm. The first step in evaluation with a patient who is in a cast or splint is removal or loosening of the immobilization device. Other scenarios in which the physician should suspect the development of compartment syndrome include crush injuries, supracondylar humerus fractures, two-bone forearm fractures, and any situation where there has been a period of ischemia followed by reperfusion, such as brachial arterial injury, upper arm replants, or a comatose patient who has been found with the arm in compression.

Objective data are obtained to help corroborate the diagnosis of compartment syndrome, usually in the form of intracompartmental pressure measurements. These can be taken by a variety of techniques including saline infusion pressure transducers, slit or wick catheters, or even arterial pressure monitors. There are newer diagnostic tools to diagnose compartment syndrome including near-infrared spectroscopy, electromyography, myotonometry, and laser Doppler; however, these modalities are, at this point, experimental and can only be used as correlative measures at best.

If compartment syndrome is suspected, it is essential that intervention be implemented quickly to lessen injury to the underlying muscles and nerves. Increased pressure on muscle can lead to irreversible ischemia and eventual necrosis. Late sequeli of untreated compartment syndrome of the forearm produces contractures of the volar forearm commonly known as Volkmann's ischemic contracture.

In general, prompt administration of surgical decompression of the forearm compartment will adequately treat the forearm compartment syndrome and leave a patient with minimal residual dysfunction.

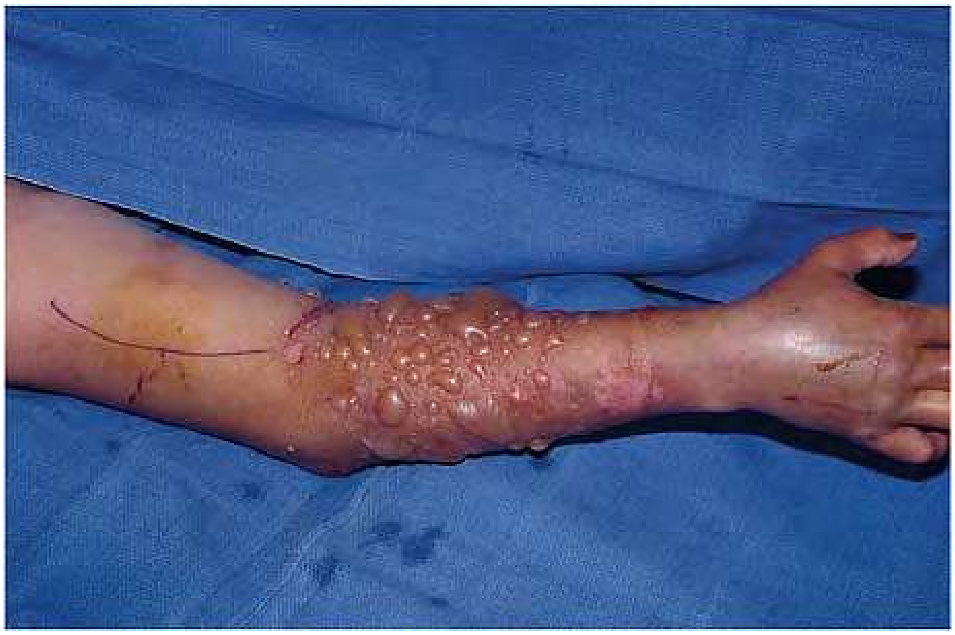

FIGURE 12-1 Patient with forearm compartment syndrome demonstrating typical cutaneous findings including tense swelling and blistering.

Contraindications

While one can argue that hemodynamic instability would be a relative contraindication to upper extremity fasciotomies, it is recommended that this not be viewed as an absolute contraindication, and that once the patient is stabilized, the surgical team proceeds with forearm and/or hand fasciotomies at the earliest available time. In the worst-case scenario, fasciotomies can be performed at the bedside in the intensive care unit. In general, upper extremity compartment syndrome is an emergency and there should be no absolute contraindications to this procedure.

Preoperative Planning

When planning for surgical intervention for upper extremity compartment syndrome, the previously mentioned diagnostic modalities should be used to document increased forearm or hand pressures. In addition, if it is thought that the patient has developed a compartment syndrome in the setting of external compression including splints or casts, these devices should be loosened or removed immediately to help diagnose the cause of the patient's symptoms as well as provide some measure of relief from the compressive insult. Often, the removal or loosening of a cast can provide adequate treatment; but even if the patient does show resolution of symptoms, practitioners are encouraged to remain vigilant and continue to closely monitor the patient's extremity until there has been a total resolution in symptoms. Other planning should include early notification of anesthesia and operating room staff and the availability of blood for transfusion if it is indicated during the surgical procedure.

In some cases coagulopathy can contribute to or be the sole cause of compartment syndrome. Excessive bleeding following fracture reduction or continued bleeding following surgery can lead to increased compartmental pressures. If a bleeding dyathesis is suspected, appropriate steps should be taken to correct the deficiency before operative intervention.

Surgery

Patient Positioning

Once anesthetized, the patient should be positioned in a way that affords adequate access to the entire affected upper extremity for both surgeon and assistant. This includes turning the operating table at a right angle to the anesthesia provider and placing the affected upper extremity on an arm board that is attached to the operating room table. Good lighting is essential, especially because the viability of the muscles in the forearm and hand needs to be assessed at the time of fasciotomy. While pneumatic tourniquets are frequently used for upper extremity surgery, they should not be used for a forearm fasciotomy because, as stated previously, the perfusion of the affected forearm musculature must be assessed at the time of fasciotomy.

Technique for Fasciotomy of Forearm

The classic teaching of forearm fasciotomies includes two incisions, one on the volar aspect and one on the dorsal aspect of the forearm. In general, both volar forearm compartments (superficial and deep) are decompressed via the volar incision while the dorsal compartment and the mobile wad can be adequately decompressed via the dorsal incision. A variety of incisions on the volar forearm has been proposed and has even been studied in cadavers. Ulnar sided incisions have been shown to be the safest approach to both superficial and deep volar forearm compartments, providing adequate visualization while avoiding injury to the radial artery and median nerve. Typically, the volar forearm incision extends from the antecubital fossa down to the wrist. The most common design is that of a sigmoid incision. The incision is marked beginning on the medial aspect of the elbow just anterior to the medial epicondyle, curving radially along the mid-forearm, curving back to the distal ulnar side of the forearm, and finally coming along the distal wrist crease to approximately the level of the carpal tunnel (Fig. 12-2). The purpose of this design is to, in theory, offer protection for the median nerve once decompression has been performed. The radially-based flap in the distal forearm will provide this median nerve protection even with marked edema and gaping of the wound. Once the skin incision is marked, it is incised with a scalpel. The dissection is further carried down through the subcutaneous tissues with either electrocautery or scissor dissection. Once the antebrachial fascia of the volar forearm is visualized, the fascia is incised with curved tenotomy-type scissors. The fascial division should be extended proximally and distally until the entire superficial volar forearm compartment is decompressed. Once the fascia is open, the viability of the musculature must to be assessed. If compartment syndrome is detected early and fasciotomy is employed in a rapid manner, there will typically not be any visible muscle necrosis. If, after observing the musculature for several minutes there is no reperfusion, or there is distinct necrosis of any of the forearm muscles, consideration should be given to muscle debridement at that time.

At this point only the superficial volar forearm compartment has been decompressed and the operating surgeon must then decompress and observe the musculature of the deep volar compartment as this is the area most frequently affected during a forearm compartment syndrome (Fig. 12-3). The deep volar compartment is approached through the interval between the flexor digitorum superficialis (FDS) and the flexor carpi ulnaris (FCU). The FDS muscle belly is retracted radially and the FCU muscle belly is retracted ulnarly; this allows visualization of the deep compartment and protects the median nerve and ulnar neurovascular bundle. The deep volar fascia is opened along the entire length of the flexor digitorum profundus (FDP) and flexor pollicis longus (FPL) muscle bellies. Once the fascia is incised, the muscles will quite often appear edematous and will herniate from their respective compartments. The viability of these muscles must also be assessed at the time of the procedure. Consideration should be given to excision of any distinctly necrotic muscle at the time of fasciotomy, as this can become a nidus for infection if left in situ.

Some authors believe that decompression of the volar forearm compartments will secondarily lead to adequate decompression of the dorsal forearm compartment as well as the mobile wad. However, if one is planning to take this approach it is strongly advised that compartmental pressure readings of the dorsal forearm compartment as well as the mobile wad be taken intraoperatively to confirm

that decompression of the volar side of the forearm has indeed led to adequate decompression of the dorsal forearm.

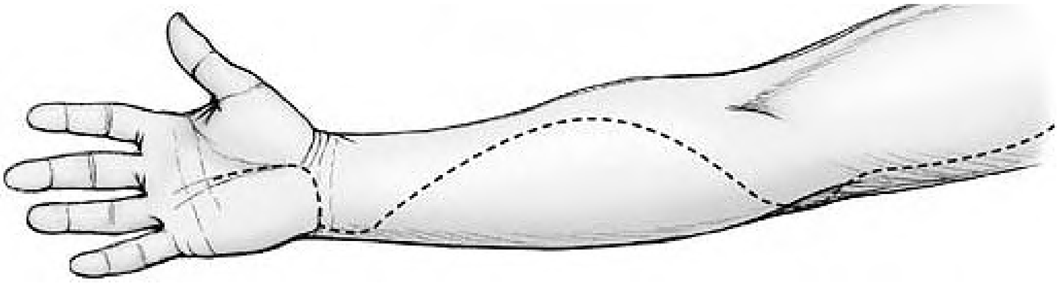

FIGURE 12-2 Incision pattern for fasciotomies of the forearm and carpal tunnel. (Adapted from Gulgonen A. Compartment syndrome. In: Green, et al, eds. *Green's operative hand surgery.* 5th ed. Philadelphia: Elsevier; 2005. With permission.)

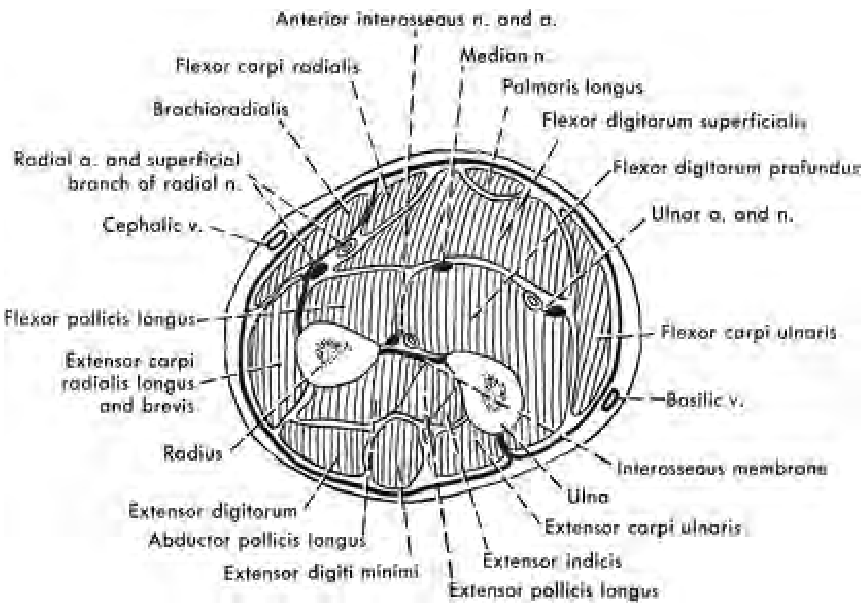
Anterior interosseous n. and a.
Median n.
Flexor carpi radialis
Palmaris longus
Brachioradialis
Flexor digitorum superficialis
Radial a. and superficial branch of radial n.
Flexor digitorum profundus
Ulnar a. and n.
Cephalic v.
Flexor pollicis longus
Flexor carpi ulnaris
Extensor carpi radialis longus and brevis
Basilic v.
Radius
Interosseous membrane
Ulna
Extensor digitorum
Extensor carpi ulnaris
Abductor pollicis longus
Extensor indicis
Extensor digiti minimi
Extensor pollicis longus

FIGURE 12-3 Cross-sectional anatomy of the mid-forearm. Note the deep position of the flexor pollicis longus and flexor digitorum profundus adjacent to the radius, ulna, and interosseus membrane. This deep position against the rigid skeleton makes these muscles particularly vulnerable to damage due to compartment syndrome (used by permission of Mayo Foundation).

After completing the volar forearm decompression, the dorsal side is now approached. The incision of the dorsal forearm is in linear fashion over the midline of the forearm dorsum. Again, this is incised with a scalpel and carried down through the subcutaneous tissue with either electrocautery or scissor dissection. The dorsal forearm fascia is incised typically with tenotomy scissors along the length of the extensor muscle bellies both proximally and distally. Once the musculature of the dorsal forearm compartment has been assessed, one must then dissect in the subcutaneous plane in the radial direction to access the mobile wad. It is here that the fascia over the brachioradialis and the radial wrist extensors (extensor carpi radialis brevis and longus) is incised with scissors and, again, the musculature is assessed.

Technique for Fasciotomy of the Hand

The hand has seven compartments: the thenar, hypothenar, carpal tunnel, and four interosseus compartments. Fortunately, all seven compartments can be decompressed with only four incisions. The carpal tunnel is opened in continuity with the forearm fasciotomy. The forearm incision is carried transversely within the wrist crease to the division of the thenar and hypothenar mounds. Here the incision is carried distally to the distal point of the transverse carpal ligament (see Fig. 12-2). After dividing the skin and subcutaneous tissues, the transverse carpal ligament is divided with scissors or a scalpel, all while continually protecting the median nerve.

The thenar and hypothenar compartments are decompressed next. These incisions are made at the border of the glabrous and hair-bearing skin of the thenar and hypothenar eminences, respectively, and are approximately 3 to 4 cm long. Once the skin and fat is opened, the fascias of these compartments are also opened with spreading scissor dissection.

Finally, the four interosseus compartments can be opened with two incisions on the dorsum of the hand. These incisions are centered over the index and ring finger metacarpal bones, and are 3 to 4 cm long. After making the index metacarpal incision, scissors are used to spread radially to open the first interosseus compartment and ulnarly to open the second interosseus space. Similarly, one enters the ring metacarpal incision to spread radially into the third interosseus compartment and ulnarly into the fourth interosseus compartment (Fig. 12-4).

At this point all of the compartments of the forearm and hand have been adequately decompressed. Primary closure is usually not possible, and not advised. One should consider delayed primary closure of the forearm and hand wounds following resolution of the patient's compartment syndrome, stabilization of the patient, and assurance that there is no underlying muscle necrosis or signs of infection. Delayed primary closure can be undertaken from 1 to 7 days following a fasciotomy, depending on the health of the patient and status of the underlying tissue. Temporary retention sutures or running elastic sutures (the â€œJacob's ladderâ€• or â€œRoman sandalâ€•) can be placed in the wound to provide some skin approximation (Fig. 12-5), but it is not advised that full closure be

attempted at this time as this can essentially recreate a compartment syndrome due to a tight skin envelope. Another acceptable method of wound temporization is application of a negative pressure dressing, which has the dual benefit of wound fluid evacuation and prevention of further skin edge retraction.

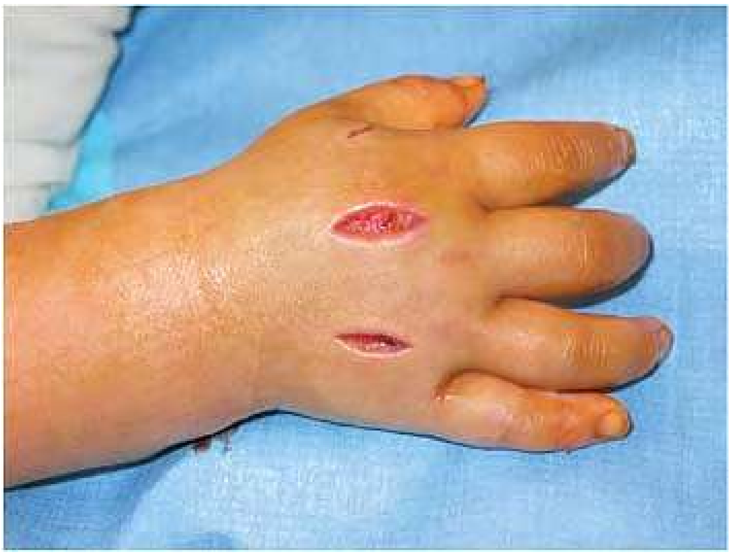

FIGURE 12-4 Two dorsal hand incisions placed over the index and ring metacarpals for decompression of the interosseus musculature of the hand.

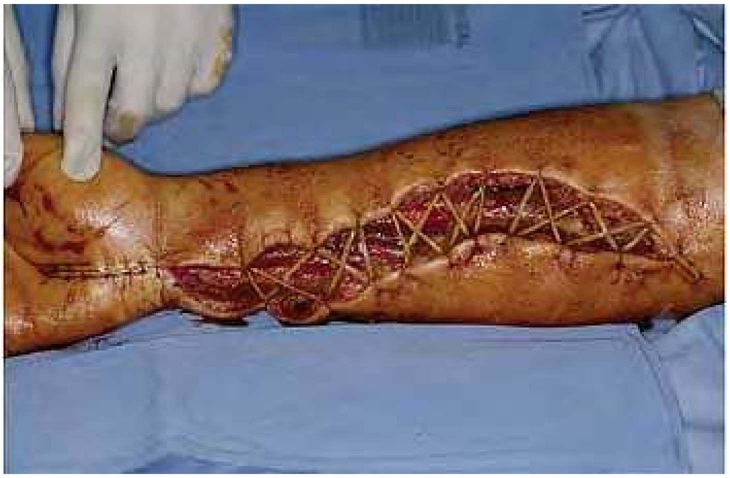

FIGURE 12-5 Partial coaptation of forearm fasciotomy wound edges using the â€œRoman sandalâ€• method. Vessel loops or rubber bands are weaved across the wound to bring the skin edges closer while allowing for some expansion due to postoperative edema.

Following initial fascial release, repeat examination under anesthesia is recommended at 24 to 48 hours; this allows the surgeon to verify that all necrotic muscle has been removed. If extremity edema has subsided, the skin may be closed with delayed primary closure. If there is any concern that closure is â€œtightâ€• or that closure could recreate a compartment syndrome, a â€œJacob's ladderâ€• or negative pressure dressing is reapplied to the wound and the patient is scheduled for surgical re-evaluation in 24 to 48 hours. Delayed primary closure of both palmar and dorsal wounds may be impractical in cases where the antecedent trauma has been significant. In such cases, skin grafting may be carried out over exposed muscle bellies to expedite wound coverage and patient rehabilitation (Fig. 12-6).

Postoperative Management

Postoperatively, the patient's forearm wounds should be dressed with nonadherent gauze. We most commonly use gauze impregnated with either petroleum or antibiotics to prevent wound desiccation. The patient should also be placed in a plaster or thermoplastic splint that covers the entire length of the forearm and places the hand in an intrinsic-plus position. One must take great care with the placement of this splint or cast as it can itself cause a compartment syndrome and lead to further muscle damage despite adequate decompression of the forearm and hand compartments.

In the several days between fasciotomy and wound closure, adequate wound care should be ensured. This is performed with frequent dressing changes and either whirlpool or pulse lavage type irrigation. At this time it is also wise to instruct the patient in hand motion exercises. The chief reason for these exercises is to ensure that the patient's range of motion remains relatively stable and that tendon gliding continues, especially with the flexor tendons at the wrist level.

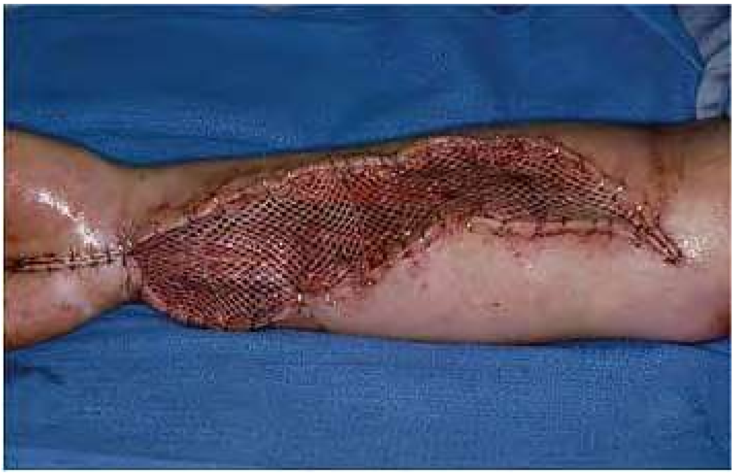

FIGURE 12-6 Forearm fasciotomy wound immediately following skin grafting.

Complications

As with any surgery, bleeding and hematoma are possible, especially in a forearm fasciotomy in which no tourniquet is used intraoperatively. Any open wound is prone to infection, and good wound care should be instituted to prevent this problem. Finally, the most dreaded complication of forearm fasciotomy is continued muscle necrosis and eventual fibrosis despite forearm fasciotomies. In general, this will likely not occur with adequate decompression. As stated previously, the most likely reason for continued muscle necrosis following this operation is inadequate decompression of the forearm and hand compartments.

Results

In general, adequate decompression of the compartments of the forearm and hand will allow resolution of the compartment syndrome and will lead to an outcome in which the patient has no or minimal residual dysfunction, although this general scenario is not without exceptions. Most frequently, the cause of muscle necrosis and subsequent fibrosis following forearm fasciotomy is due to incomplete fasciotomy. Therefore, care must be taken intraoperatively to fully extend the fasciotomy proximally and distally, thereby decompressing the full length of all the

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part II - Management of Soft Tissues Within the Upper Extremity > Chapter 13 - Soft Tissue Interposition Flaps in the Management of Heterotopic Ossification and Proximal Radioulnar Synostosis

---

Chapter 13

Soft Tissue Interposition Flaps in the Management of Heterotopic Ossification and Proximal Radioulnar Synostosis

Douglas P. Hanel

Seth D. Dodds

Indications/Contraindications

Heterotopic ossification of the elbow with loss of motion can be severely debilitating. The functional arc of elbow motion spans from 30 to 130 degrees of flexion and from 50 degrees of pronation to 50 degrees of supination. Limitations to this functional arc significantly impair the ability to perform activities of daily living. Cases of elbow ankylosis or radioulnar synostosis further minimize use of the affected extremity. Once heterotopic ossification has developed and constricts motion, it is nearly impossible to regain the lost motion with conservative measures, such as physical therapy, dynamic splinting, radiation therapy, or medication.

Surgical resection of heterotopic bone about the elbow should be considered in patients who present with an unacceptable loss of flexion/extension or pronation/supination. Excision is also warranted in cases of neurovascular impingement caused by ectopic bone. In cases of proximal radioulnar synostosis or even radial head excision, interposition materials can be used to cover exposed bone surfaces.

While options for soft tissue interposition include silicone sheeting, fat graft, and free adipofascial flaps, pedicled myofascial flaps and allograft fascia lata have become increasingly popular. The pedicled anconeus myofascial flap is an ideal choice in those cases approached posterolaterally, whereas the pedicled brachioradialis muscle flap passed through the interosseous membrane is best suited for those cases approached from an anterior exposure. The soft tissue interposition acts as a barrier to the formation of recurrent heterotopic calcification and allegedly decreases pain with pronation and supination. It is suggested, though unproven, that pedicled graft tissue has greater potential for sustained viability when judged against nonvascularized tissue transfers, such as subcutaneous fat or an adipose fascial graft. When properly fixed to the underlying bone, these pedicled flaps can sustain aggressive postoperative range-of-motion exercises that might dislodge less robust tissue such as fat alone or adipose-fascial grafts. We prefer the anconeus muscle pedicle flap when available and tensor fascia lata allograft when the anconeus is not available or not large enough.

Historically, contraindications to excision of heterotopic bone with or without soft tissue interposition flaps included immature ossification and an unreliable soft tissue envelope. In the past decade there have been a number of reports documenting the efficacy of early excision of heterotopic bone. It is our experience that patients suffer less soft tissue contracture and have superior function with early release of a stiff elbow. The recurrence of heterotopic bone in the posttraumatic setting has not been shown to be predicated on the timing of the excision. Outcomes of surgical excision, however, will be threatened by a poor soft tissue envelope. Once the posttraumatic or post-burn wounds have healed, soft tissue swelling has abated, and nerve recovery has plateaued, patients may safely undergo elbow contracture release.

Contraindications to pedicled soft tissue interposition flaps depend on the specific muscle selected. For an anconeus muscle interposition, previous traumatic or surgical disruption of its primary vascular supply (the medial collateral artery from the profunda brachii) jeopardizes the viability of the raised muscle flap. The brachioradialis â€œwrap aroundâ€• flap should not be raised if there is a nonfunctioning biceps brachii or brachialis, as the brachioradialis provides assistance with elbow flexion as well as supination when the forearm is fully pronated. Assuming that the arm and forearm musculature is intact, using the anconeus or brachioradialis as a pedicled interposition flap causes little functional loss.

Preoperative Planning

Before operative release and soft tissue interposition, patients must be carefully evaluated. The history should focus on the primary complaint. Patients must verbalize appropriate frustration with their disability from elbow stiffness to warrant release. It is also imperative that patients demonstrate the willingness and capacity for intensive rehabilitation. If the presenting complaint is predominantly pain, then contracture release will be futile. A focused surgical history needs to be elicited. Previous injuries and surgeries of the involved extremity offer critical information about the status of osseous and cartilaginous structures as well as the elbow's soft tissue envelope. Operative reports from previous surgeries help understand the integrity and location of possibly transposed neurovascular structures.

Elbow range of motion, stability, pain, and functional ability are assessed. The Mayo Elbow Performance Score serves as a summary of these findings. In addition, the examination includes assessment of forearm rotation and wrist and hand function. If there is physical evidence of nerve dysfunction, electrodiagnostic studies should be obtained to confirm the location of compression and to act as a baseline of nerve function. All of our patients have had previous surgeries, and as such, a careful assessment of incisions about the elbow with regard to the palpable and radiographic location of heterotopic ossification is essential. The integrity of the skin and subcutaneous tissues should be evaluated. Lingering soft tissue swelling, edema, or erythema may all point to additional diagnoses to be contended with before deciding on heterotopic bone excision.

Standard radiographs of the elbow are obtained with oblique views to improve visualization of ectopic ossification. We routinely perform computed tomography (CT) on cases of elbow heterotopic ossification where the congruity of the articular surface is in question and in all cases with proximal radioulnar synostosis. An axial CT scan with coronal and sagittal reformatted images is currently the most helpful method to visualize the location and extent of a bony bridge between the radius and ulna (Figs. 13-1 and 13-2). Vascular study of the elbow and proximal forearm should be considered if there are concerns about the integrity of the regional blood supply, especially in cases of pedicled soft tissue interposition. Unless there are specific historical or physical findings suggesting potential or indolent infection, we limit blood studies to those required for a prolonged general anesthetic.

The timing for intervention is somewhat nebulous and is certainly directed by physician bias. We do not believe there is sufficient scientific evidence to suggest that nuclear medicine scans or blood alkaline phosphatase levels are beneficial in the assessment or timing of surgical intervention. In reviewing the literature and comparing our experience, it would appear that waiting for fracture healing and ectopic bone maturation, defined as well-delineated borders, is the most commonly used parameters. This would suggest that operative intervention be carried out 4 to 12 months after injury.

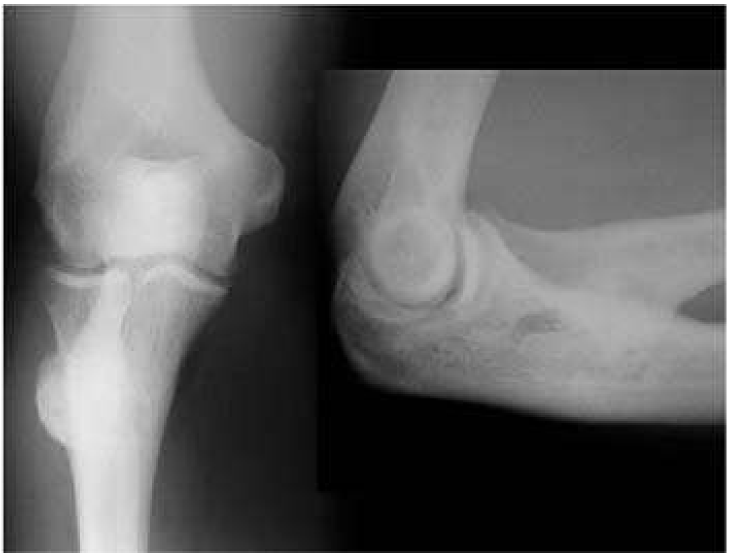

FIGURE 13-1 The anteroposterior and lateral elbow radiographs of a 42-year-old power lifter 8 weeks after a single incision repair of distal biceps tendon rupture. The forearm is ankylosed in mid pronation-supination.

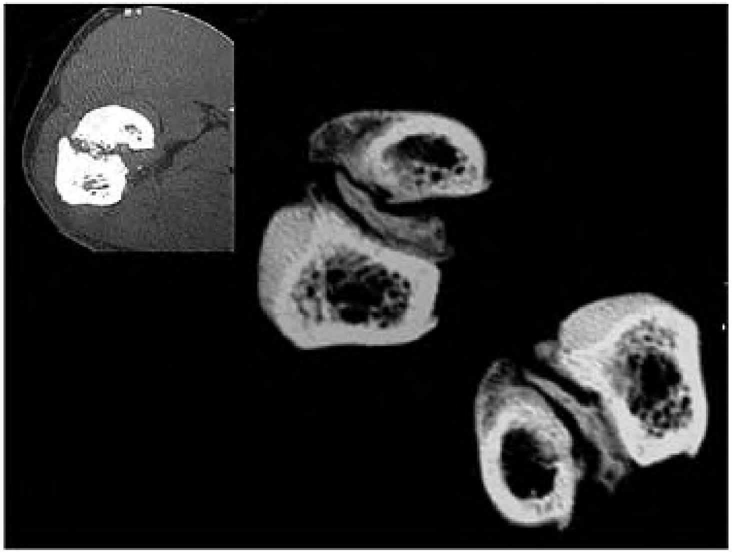

FIGURE 13-2 Coronal CT image reveals the extent of heterotopic bone involvement in the region of the bicipital tuberosity.

Excision of heterotopic ossification and soft tissue flap interposition does not require a great deal of special equipment. We have listed a few items that facilitate operative intervention and postoperative rehabilitation in Table 13-1. In addition to the tools listed in the table, we have the following items readily available: a hinged fixator, radial head prosthetics, allograft tendon for ligament reconstruction, allograft fascia lata if local interposition material is not available, and a total elbow arthroplasty set when all else fails and the patient is the appropriate age.

Surgery

Patient Positioning

Operative approach and patient positioning are determined by previous incisions, location of heterotopic bone or synostosis, and shoulder mobility. While many surgeons prefer to operate on the elbow with the affected extremity positioned across the patient's supine chest, we prefer to position

the patient with the affected extremity abducted onto a radiolucent table. The extremity is draped free up to the clavicle. If the shoulder is mobile, external rotation will present the medial and anterior aspects of the elbow while internal rotation presents the lateral and posterior elbow. The flexibility to effortlessly alternate between lateral and medial approaches facilitates complete excision of heterotopic bone about the elbow.

Table 13-1. Soft Tissue Interposition Flaps for Elbow Heterotopic Ossification: Operative Equipment

Standard Equipment	Optional Equipment
Sterile tourniquet	Fluoroscopy
Retractors, thyroid type (long and narrow)	Hardware removal instruments
Lamina spreader	Total elbow arthroplasty
Vessel loops	Lidocaine infusion pump
Rongeurs (including Kerrison and pituitary)	Continuous passive motion machine
Osteotomes and curettes	Postoperative radiation therapy
Suture anchors	Static progressive vs. dynamic splinting
Hinged elbow distractor	
Tensor fascia lata allograft	
Closed suction drain	

Technique

Exposure

We do not use tourniquets for these cases. They are a detriment to tissue mobilization, mask small arterial bleeding, and lead to venous engorgement during prolonged cases. The initial increase in bleeding at the time of incision is mitigated by injecting the proposed incision with 0.25% bupivacaine and 1/200,000 dilute epinephrine solution. The incision is delayed 7 to 10 minutes to allow the epinephrine to affect local capillaries.

It is our preference to use a posterior skin incision for elbow release surgery. This window allows a â€œglobalâ€• approach to the elbow and can be used to access the medial, lateral, and anterior sides of the joint. The skin incision is typically straight, passing 2 cm medial or lateral to the tip of the olecranon, but may be curvilinear to incorporate previously placed posteromedial or posterolateral surgical scars. Depending on the required heterotopic bone excision or hardware removal, the incision may extend from the proximal arm to the distal forearm. This dissection avoids injury to both the medial and the lateral brachial cutaneous nerves, preserving sensibility to the proximal forearm.

When employing this approach, it is critical to create thick soft tissue flaps. The skin incision is carried down to and includes the triceps fascia proximally and the extensor fascia of the forearm distally. This effectively creates robust fasciocutaneous flaps that can be elevated circumferentially about the elbow. The extent of elevation is dictated by the location of the joint involvement. If there is medial joint involvement, as determined by CT scan, or ulnar nerve symptoms, we address these first. If there is no involvement we go directly to the lateral elbow and proximal forearm. The dissection is carried to the medial intermuscular septum, and the ulnar nerve is identified as it passes from the anterior to the posterior compartment approximately 8 to 10 cm proximal to the medial epicondyle. Even if the ulnar nerve has been â€œtransposed anteriorlyâ€• in previous procedures, we believe that it is critical to identify the entire medial intermuscular septum, and follow it to the humerus (Fig. 13-3). In cases in which the medial intermuscular septum has been excised in part or in whole, the dissection follows the medial border of the triceps until the humerus is encountered. When the ulnar nerve is found, it should be mobilized from the cubital tunnel, preferably with a small cuff of medial triceps to protect the nerve's vascular supply. The nerve is followed into the forearm until disappeared deeply between the heads of the flexor carpi ulnaris. Any tight scar or fascial bands crossing the nerve in this dissection are divided. The medial intermuscular septum is removed in its entirety. By following the ulnar nerve proximally in the arm, the median nerve and its accompanying brachial artery can be located along the anterior margin of the intermuscular septum. Alternatively, in arms that are not densely scarred the median nerve and brachial artery can be identified in the distal arm superior to the leading edge of the pronator teres origin as it runs medial to the substance of the biceps and brachialis muscles and just beneath the fibers of the bicipital aponeurosis. Excision of heterotopic bone involving the medial posterior and anterior elbow joint is conducted and described in greater detail in the next section.

Next, the lateral side of the elbow and involvement of the proximal forearm are addressed. In cases of anterolateral heterotopic ossification, the radial nerve is identified at mid-arm and followed distally. This is done by elevating the lateral flap in the same tissue plane as the medial dissection, between the muscular investing fascia and the muscle belly of the triceps. When the lateral intermuscular septum is encountered, the radial nerve should be identified and protected. Identifying the radial nerve in a scarred bed can be daunting. We therefore use the following strategies. First, inspect the undersurface of the flap; frequently the posterior antebrachial cutaneous branches of the radial nerve to arm and forearm are visible. These branches can be followed proximally into the lateral intermuscular septum where they are found to take off from the radial nerve proper (Fig. 13-4). If this landmark is not readily available, our second approach is to dissect the distal portion of the flap toward the lateral epicondyle. On reaching the lateral intermuscular septum, the dissection is directed cephalad. The triceps muscle belly is freed from the posterior aspect of the intermuscular septum. Small vessels and nerves seen entering the triceps muscle should be followed proximally; they will lead to the radial nerve proper. If these markers fail to lead to the nerve, cautiously proceed along the posterior intermuscular septum. Somewhere between 6 and 10 cm proximal to the tip of the lateral epicondyle, the nerve will be encountered passing along the spiral groove to pass through the lateral intermuscular septum to enter the anterior aspect of the arm. Once the radial nerve is identified, the lateral intermuscular septum is removed and the nerve followed distally. At the level of the elbow, the nerve is easily followed into the internervous plane between the proximal aspect of the brachioradialis and the distal aspect of the brachialis. Simple blunt dissection between these two muscles just superior to the joint line will reveal the radial nerve before it dives under the supinator more distally.

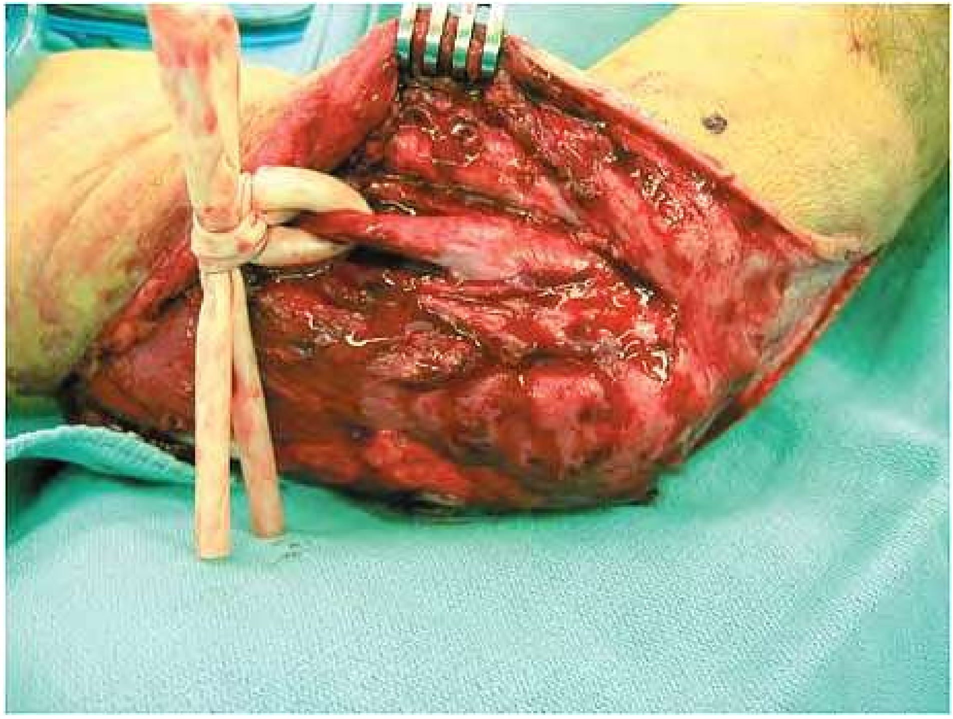

FIGURE 13-3 This previously â€œtransposed ulnar nerveâ€• sits directly on the medial epicondyle. It is essential to identify and protect the ulnar nerve before proceeding with capsular and heterotopic bone resection.

Excision of Heterotopic Bone

The excision of heterotopic bone and the scarred joint capsule follows. Anterior heterotopic bone frequently resides in the distal aspect of the brachialis muscle. It can extend medially and laterally encasing the collateral ligaments or even neighboring neurovascular structures. If a medial approach is chosen, entrance to the joint capsule and distal aspect of the brachialis can be achieved by exploiting the internervous plane between the ulnar innervated flexor carpi ulnaris and the median innervated palmaris longus, flexor carpi radialis, and pronator teres. Alternatively, the flexor-pronator muscle mass can be sharply elevated off the medial epicondyle, carefully preserving the underlying anterior band of the medial collateral ligament. In severely contracted elbows, this exposure affords uncompromised visualization of the anterior capsule and crossing neurovascular bundle. This elevated myofascial flap can be reattached to the medial epicondyle with multiple suture anchors or sutured down to the epicondyle through bone tunnels. If the flexor-pronator mass is released, consideration can be given to submuscular transposition of the ulnar nerve.

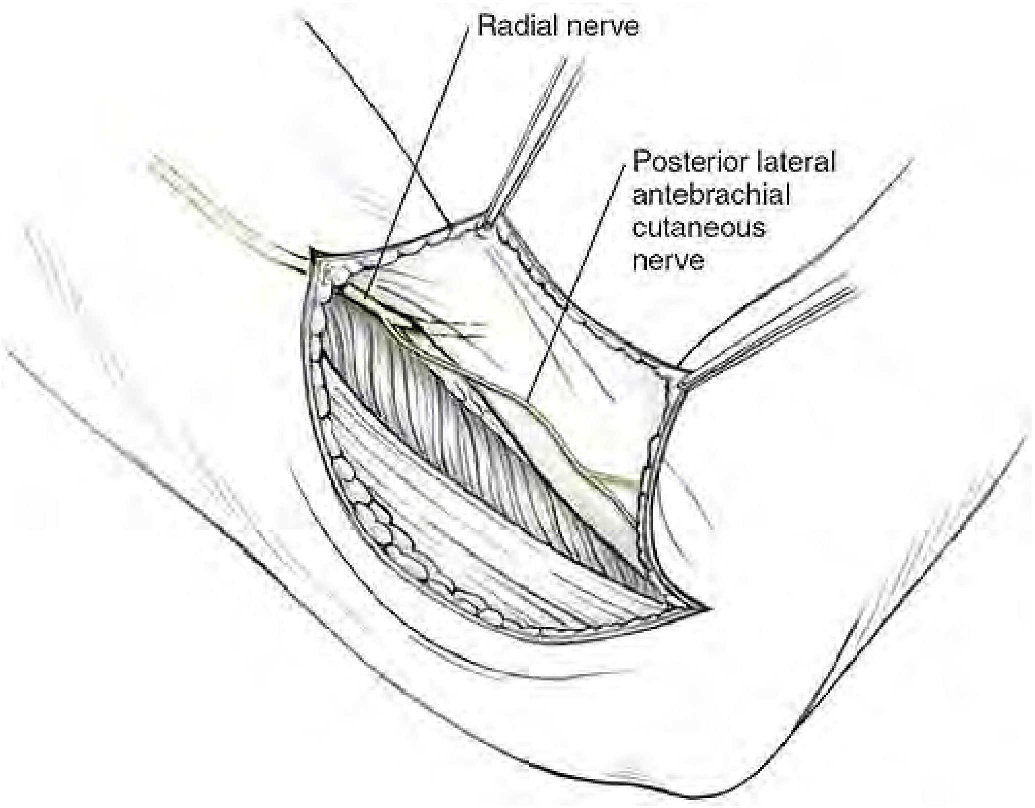

FIGURE 13-4 The posterior antebrachial cutaneous nerve is identified distally and followed proximally into the lateral intermuscular septum. This cutaneous nerve leads to the radial nerve proper. The radial nerve proper can then be followed from into the anterior compartment of the arm.

The anterolateral aspect of the elbow is approached next. If there is complete elbow flexion and extension, the dissection is directed toward the posterolateral forearm. In cases in which the anterior joint needs to be exposed, one of three intervals is used (Fig. 13-5). The first approach, and in our experience the most frequently used, elevates the proximal most portion of the muscle taking origin from the medial epicondyle, usually the brachioradialis and a portion of the extensor carpi radialis longus. The dissection is carried medially, sweeping the brachialis from the front of the humerus. The interval between this muscle and the anterior joint capsule is developed. Heterotopic bone, if encountered, is dissected with blunt-tipped elevators and left attached to the anterior joint structures. If the relationship between these structures and the radial nerve is doubtful, the nerve is again identified in the distal arm and followed into the area of joint dissection. Under direct visualization, a blunt right-angled retractor is placed in the interval between the anterior joint dissection and the more superficially located muscle and nerve. The dissection stops distally when the coronoid process is encountered or the heterotopic bone becomes confluent the forearm bones. The entire anterior joint capsule and the heterotopic bone are removed. The coronoid fossa is cleared of soft tissues. Two other approaches to the anterolateral elbow consist of developing the interval between the anconeus and extensor carpi ulnaris, or the interval between the extensor carpi radialis brevis origin and the extensor digitorum communis. Both dissections allow easy access to the joint capsule overlying the radial head and neck. The capsular, radial, and annular portion of the lateral collateral ligament complex can be reflected or excised to expose the joint. The ulnar portion of the lateral collateral ligament, if not encased in ectopic bone, should be preserved (Fig. 13-6). More often than not, elevation of the lateral elbow complex from the ulna it is necessary. The ligamentous attachments will be reconstructed after the ectopic bone has been excised. If visualization of the anterior joint capsule is insufficient with these intermuscular approaches, the origin of the brachioradialis and extensor carpi radialis longus is elevated as described previously.

The posterior elbow joint is approached from the lateral side by elevating the triceps muscle from the lateral column of the distal humerus and developing the interval between triceps muscle and the posterior joint capsule. The triceps muscle insertion onto the olecranon is preserved. The posterior capsule, the contents of the olecranon fossa, and any bony impediments to elbow motion arising from theses posterior structures are debrided.

Proximal Radioulnar Synostosis Takedown

In cases in which pathology is limited to the proximal radioulnar joint in the vicinity of the bicipital tuberosity, the synostosis may be approached by sweeping the entire anconeus and ulnar origin of the extensor carpi ulnaris complex off of the posterolateral edge of the proximal ulna (Fig. 13-7Aâ€“C). The dissection along the proximal ulna is carried along the radial edge to the base of the synostosis and followed along the interosseous space until the radius is encountered. If the forearm can be pronated, there should be no danger to the posterior interosseous nerve. A pronated forearm allows the posterior interosseous nerve to lay anteriorly and medially protected in the substance of the supinator muscle. However, a supinated forearm does just the opposite. In cases where the forearm is ankylosed in supination, consideration should be given to locating the posterior interosseous nerve before the posterior dissection. This usually requires an anterior incision immediately adjacent to the ulnar border of the brachioradialis. The radial nerve is found in the interval between the brachioradialis and brachialis muscle and followed distally until it is seen to divide into the posterior interosseous nerve (PIN) and the superficial branch of the radial nerve. The PIN is followed distally into the supinator muscle. The arcade of Frohse is divided along with fibrosed muscle and scarred tissue that could lead to potential compression injury during the dissection. The posterior dissection is then resumed.

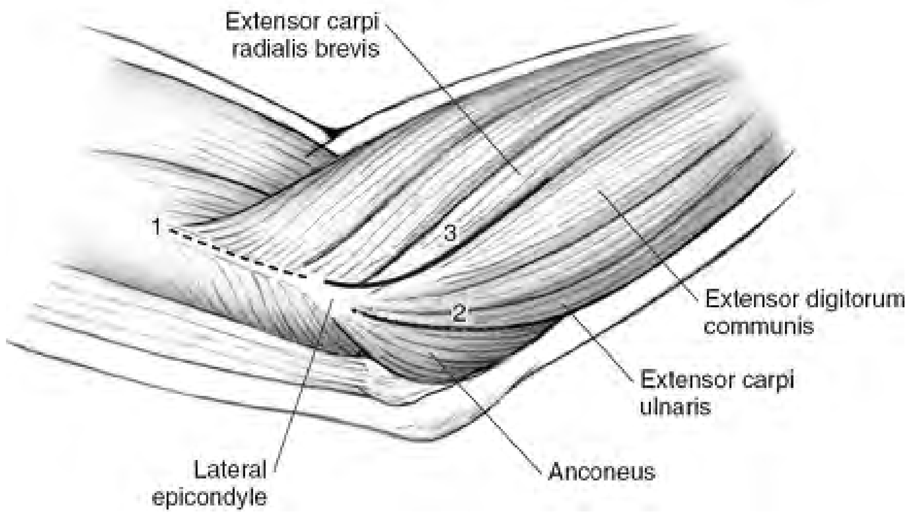

FIGURE 13-5 Three intervals used to approach the anterior lateral elbow. (1) Partial elevation of the muscles taking origin from the lateral epicondyle, the

brachioradialis, and cephalad portion of the extensor carpi ulnaris. (2) Anconeus and extensor carpi ulnaris. (3) Extensor carpi radialis brevis and extensor digitorum communis.

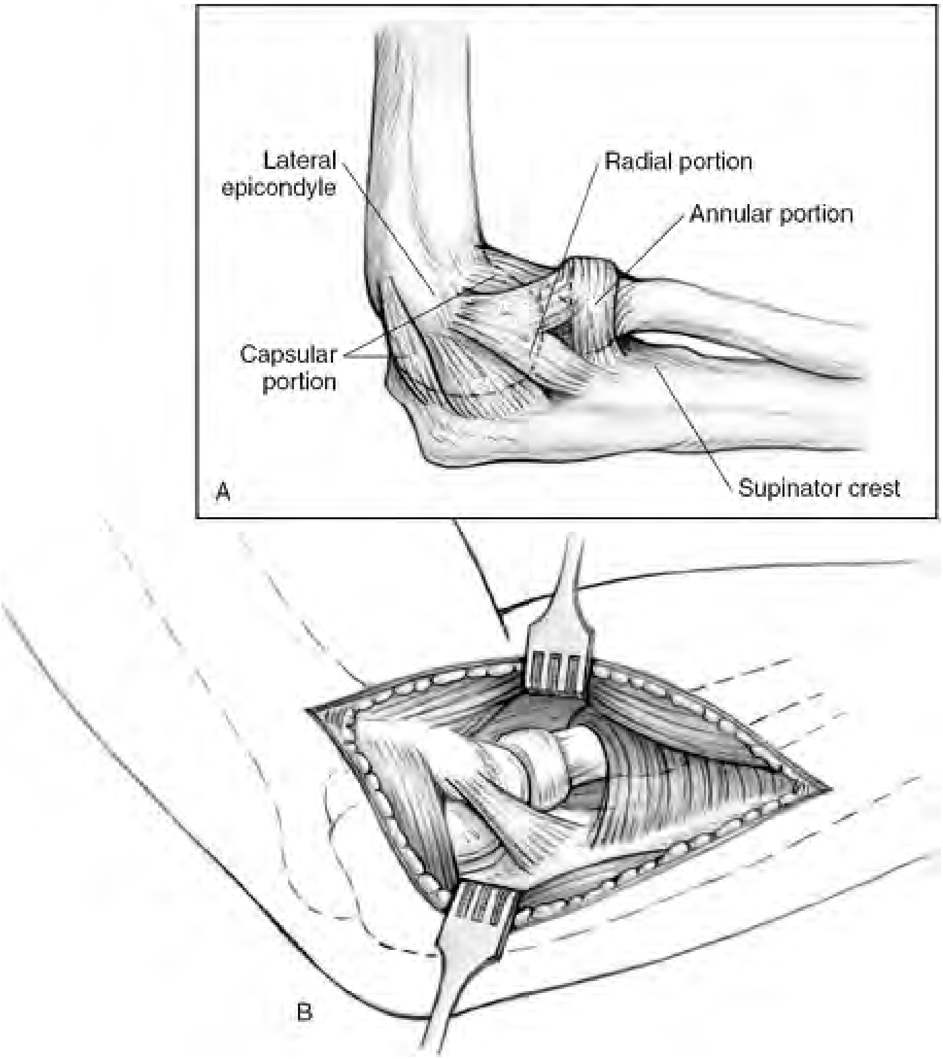
Lateral epicondyle
Radial portion
Annular portion
Capsular portion
Supinator crest
A
B

FIGURE 13-6 A: Lateral collateral ligament complex of the elbow consisting of the capsular, radial, and annular portions (*stippled*) removed to expose the lateral elbow articulations. B: The stippled portion of Figure 13-6A has been removed, exposing the radial head and neck, and the distal humerus. Preserving the ulnar portion of lateral collateral ligament (*shaded area*), when possible,

preserves elbow stability.

Once the anconeus is elevated, tissue retraction is maintained by broad right-angled retractors or broad Homan retractors. Narrow lever arm type retractors are avoided. Although anecdotal, it is thought that narrow retractors place excessive pressure on adjacent neurovascular structures, especially the PIN.

Synostosis resection proceeds using a combination of small rongeurs, osteotomes, and Kerrison rongeurs. Small lamina spreaders help retract the proximal radius and proximal ulna, opening up the space between. Once the synostosis has been released and there is freedom between the two bones, the lamina spreader offers greater direct visualization of remaining ectopic bone. The synostosis can now be excised down to the native cortices of the proximal radius and proximal ulna. It is imperative that the exposed bony surfaces be debrided to a smooth surface. Sharp edges combined with forearm rotation will penetrate and disrupt interposition flaps.

Intraoperative radiography is used to demonstrate the thoroughness of the ectopic bone excision. On completion of the resection, the surgical field is irrigated with 3 L of normal saline. Bleeding

points in soft tissues are coagulated, and bleeding points in exposed bone are sealed with bone wax. If the medullary canal of either the radius or ulna is breached, the hole is packed with allograft bone graft soaked in thrombin. This stops the marrow content from flowing into the operative field (see Fig. 13-7D,E).

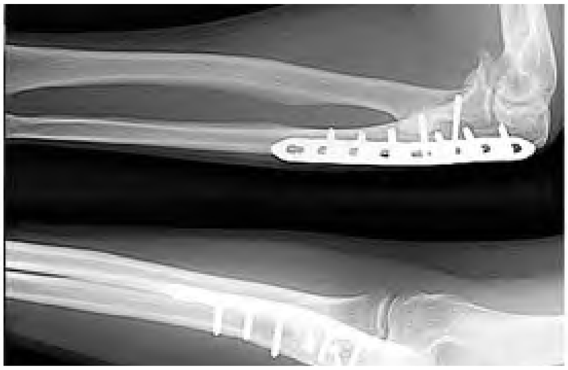

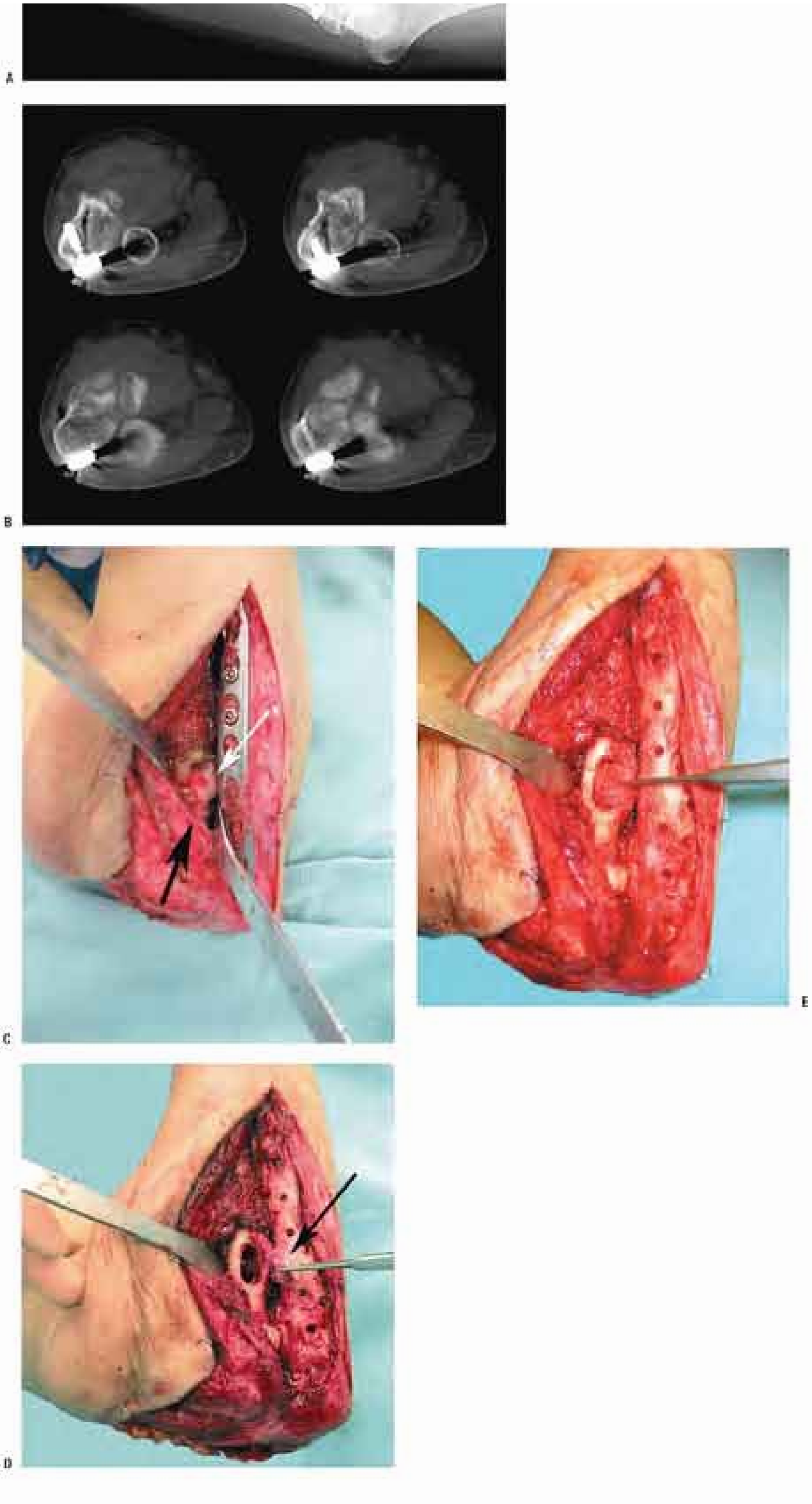
A
B
C
E
D

FIGURE 13-7 A: This patient developed a radioulnar synostosis after treatment of a Monteggia fracture. B: The CT scans demonstrate the location of the synostosis adjacent to the bicipital tuberosity. C: The entire anconeus and extensor carpi ulnaris complex have been elevated from the ulna (*dark arrow*), exposing the bridge of heterotopic bone between the radius and ulna (*light arrow*). D: The synostosis encased the bicipital tuberosity and tendon. The ectopic bone was removed, and the tendon left attached to a block of heterotopic bone (*dark arrow*). The resection exposed the medullary canal of the radius. The fixation plate was removed. E: The medullary canal is packed with allograft bone chips to minimize extrusion of marrow content. The biceps tendon and attached bone block are pushed into the medullary canal and secured with a stout nonabsorbable suture (not shown).

Interposition Material

Even with meticulous technique, the proximity of the proximal radius and ulna to each other makes these structures prone to recurrent synostosis. Some type of interposition material is a useful preventative adjunct. Three techniques are discussed in the following section. Each of these procedures is preceded with copious irrigation of the synostosis site and diligent hemostasis as described previously.

Pedicled Anconeus Myofascial Flap Interposition

The anconeus myofascial flap receives its blood supply from the collateral circulation about the elbow, predominantly the medial collateral artery (MCA) branch of the profunda brachii artery and venae commitantes (Fig. 13-8). The muscle is elevated from the posterior lateral border of the ulna, usually in a distal to proximal direction. The distal edge starts roughly at the junction of the proximal and mid one-third of the forearm.

A stout septum separates this muscle from the flexor carpi ulnaris. Identification of this septum and the anconeus itself may be difficult if previous surgery or initial trauma has injured the muscle. In these cases, an alternative material may be required.

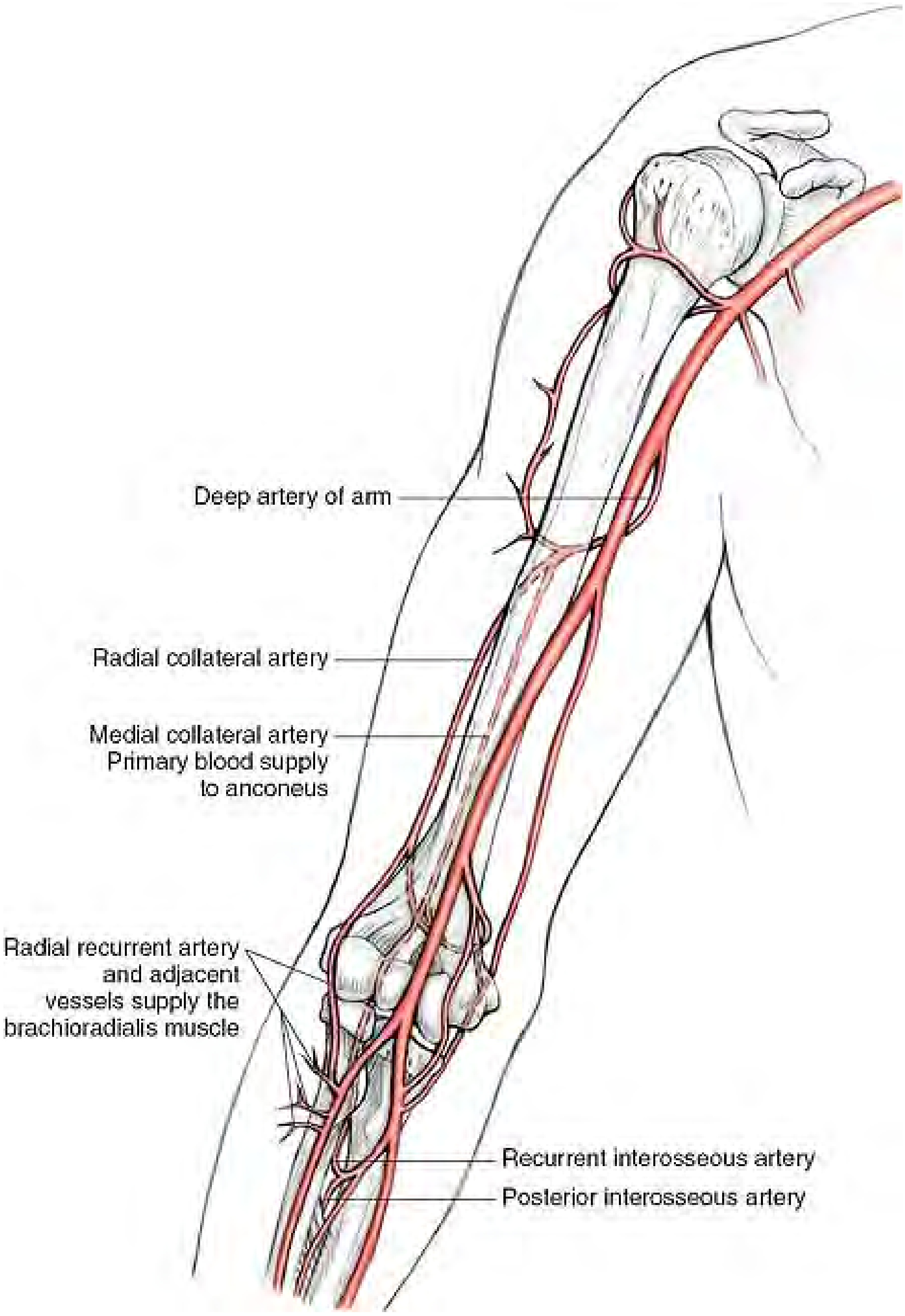
Deep artery of arm
Radial collateral artery
Medial collateral artery Primary blood supply to anconeus
Radial recurrent artery and adjacent vessels supply the brachioradialis muscle
Recurrent interosseous artery
Posterior interosseous artery

FIGURE 13-8 The arterial anatomy of the distal arm. The medial collateral artery (MCA), a branch of the profunda brachii, is the primary blood supply to the anconeus. The radial recurrent artery (RRA), a branch of the radial artery, and

small arterial branches within 3 cm of the RRA, provide the primary blood supply to the brachioradialis muscle.

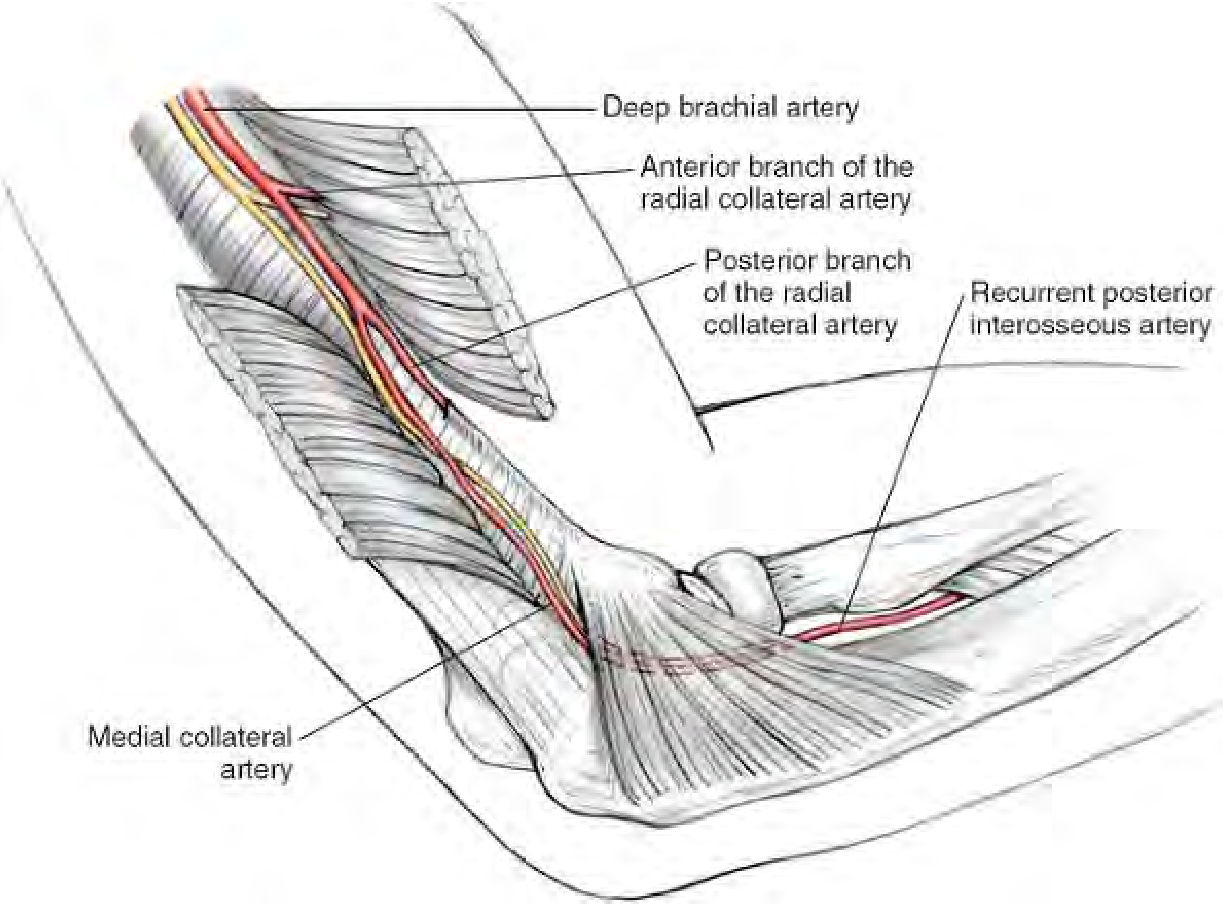
Deep brachial artery
Anterior branch of the radial collateral artery
Posterior branch of the radial collateral artery
Recurrent posterior interosseous artery
Medial collateral artery

FIGURE 13-9 In addition to the medial collateral artery (MCA), the anconeus receives blood supply distally from the recurrent posterior interosseous artery (RPIA). These vessels (the MCA and RPIA) frequently form an anastomosis on the deep surface of the anconeus. The RPIA is frequently sacrificed in the process of elevating this flap. (From Schmidt CC, Kohut GN, Greenberg JA, Kann SE, Idler RS, Kiefhaber TR. The anconeus muscle flap: its anatomy and clinical application. *J Hand Surg Am.* 1999;24(2):359-369. With permission.)

Elevation of the flap is carried up to the lateral epicondyle until sufficient freedom exists to permit mobilization of the tissue's leading edge to the distal extent of the synostosis. By protecting the proximal origin and the lateral fascial attachments of the triceps, the medial collateral artery of the elbow can be preserved. The anconeus also receives blood supply distally from the recurrent posterior interosseous artery (RPIA). These vessels (the MCA and RPIA) frequently form an anatomosis on the deep surface of the anconeus (Fig. 13-9). The RPIA is usually sacrificed in the process of removing the synostosis or in elevating this flap.

The distal end of the anconeus flap is attached to the ulnar edge of the biceps tuberosity while the forearm is placed in full pronation. Sutures alone or suture anchors placed into the proximal radius or biceps tendon secure the anconeus flap with mattress or Bunnel locking stitches. Now, as the forearm is supinated, the anconeus will be drawn into the proximal radioulnar space where the synostosis previously layâ€”creating a thin, pedicled interposition flap of viable muscle and fascia (Fig. 13-10).

The utility of the anconeus myofascial interposition flap has recently been expanded by Morrey and Schneeberger to include soft tissue interposition between the capitellum and the proximal radius in cases of radial head excision or failed radial head arthroplasty (Fig. 13-11). Three interposition options are described: type I, interposition at the radiocapitellar joint; type II, the muscle is brought posterior to the intact fibers of the ulnar portion of the lateral ulnar collateral ligament and interposed between radiocapitellar and proximal radioulnar joint; and type III, proximal radioulnar wrap. In each type of interposition, the muscle is secured with stout suture placed through drill holes.

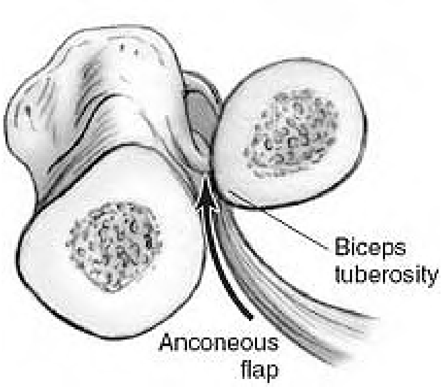
Biceps tuberosity
Anconeous flap

FIGURE 13-10 With the forearm placed in full pronation, the distal end of the anconeus flap is attached to the ulnar edge of the biceps tuberosity (BT) with a suture anchor or direct stitch into the biceps tendon. As the forearm is supinated (*arrow*), the anconeus will be drawn into the proximal radioulnar space.

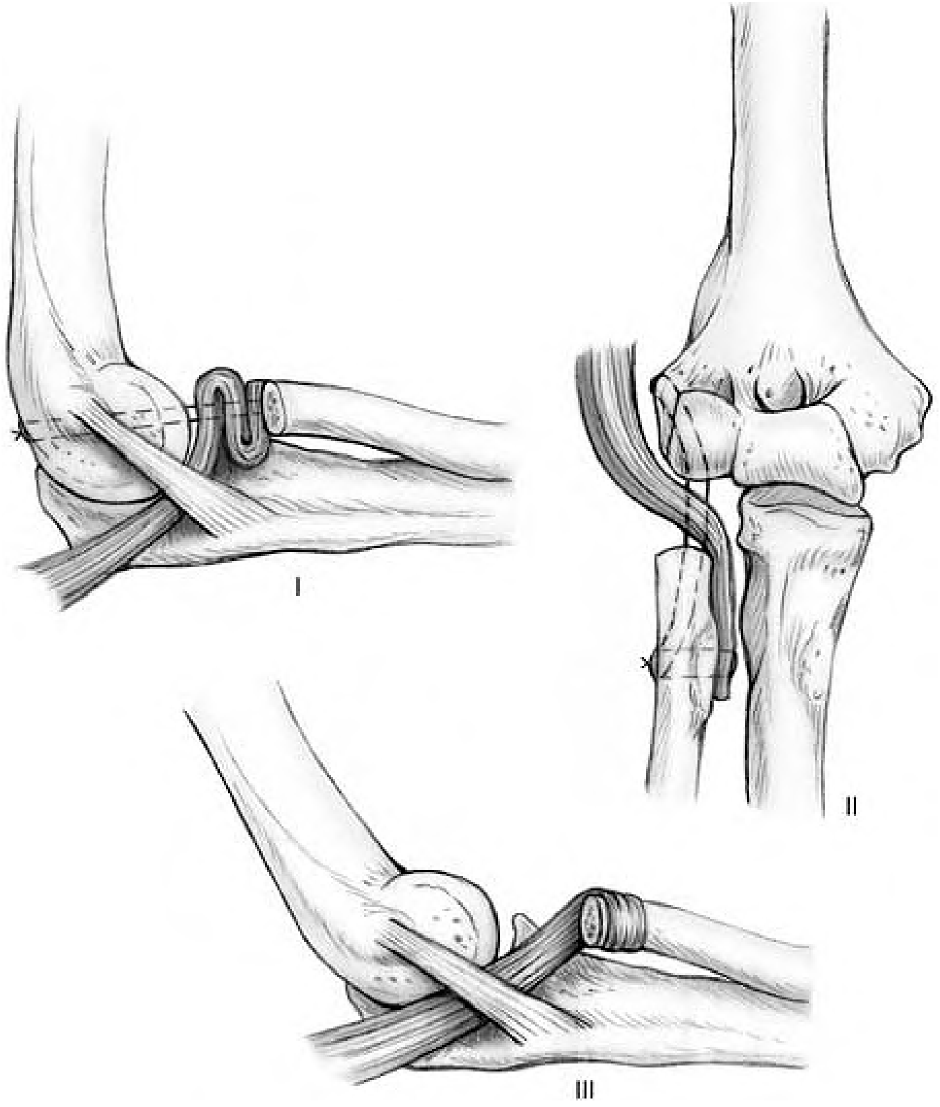

FIGURE 13-11 Three interposition options for the anconeus flap after radial head resection: type I, radiocapitellar joint: type II, radiocapitellar and proximal radioulnar joint; and type III, proximal radioulnar wrap. (From Morrey BF, Schneeberger AG. Anconeus arthroplasty: a new technique for reconstruction of the radiocapitellar and/or proximal radioulnar joint. *J Bone Joint Surg Am.* 2002;84-A(11):1960â€“1969. With permission.)

Distally based anconeus myofascial flaps composed of anconeus and some neighboring extensor carpi ulnaris have been described. This flap depends on an intact RPIA for its blood supply, a small vessel that is easily injured during the initial fracture fixation or during resection of a synostosis. Distally based anconeus flaps may have application in the management of soft tissue defects not associated with elbow fractures such as burns or skin avulsion but should be used with caution in the setting of synostosis resection.

The proximally based anconeus pedicle flap can also be used to provide soft tissue coverage of small defects about the traumatized elbow. Three areas where this muscle will reach is the lateral epicondyle, the posterior surface of the olecranon, and the distal radial aspect of the triceps insertion into the olecranon. The muscle is elevated, rotated over the defect, and covered with split thickness skin grafts.

Pedicled Brachioradialis Myofascial Flap Interposition

The brachioradialis interposition flap is useful when resection of a radioulnar synostosis is performed through a combined anterior and posterior approach. The posterior incision and approach is the same as described previously: the anconeus-extensor carpi ulnaris interval is developed, the supinator sharply elevated from the ulna, and the synostosis excised. The anterior approach is conducted through a longitudinal incision made along the ulnar border of the brachioradialis, from the elbow flexion crease to the distal forearm. The superficial branch of the radial nerve is located in the distal forearm and carefully dissected from the undersurface of the brachioradialis. This nerve is followed proximally into the cephalad region of antecubital fossa where it is found to branch from the radial nerve. At this level the radial nerve branches to the brachioradialis, and extensor carpi radialis muscle may be encountered

and is protected. Most important, the posterior interosseous nerve branch is identified and followed into the substance of the supinator muscle. The stout fibers on the leading edge of the supinator, the arcade of Frohse, are a potential source for compression during the remainder of the procedure and are divided. Simultaneously with the dissection of the radial nerve, the radial artery and its branches to the lateral elbow are encountered and protected. The largest of these branches, the radial recurrent artery, is found crossing from ulnar to radial, just distal to the biceps tendon (see Fig. 13-8). Care is taken to preserve not only the radial recurrent artery, but also the branches of the radial artery just distal to the radial recurrent artery. While the proximal aspect of the brachioradialis muscle belly is vascularized by the radial recurrent artery, the distal aspect of the muscle receives its blood supply from the arborization of the radial artery approximately 3 cm distal to the radial recurrent artery. In this dissection the median nerve and the ulnar artery are shielded by the medially retracted flexor-pronator muscle mass on the ulnar aspect of the proximal forearm.

Once these pertinent neurovascular structures are dissected free and protected, attention can be turned to the synostosis excision. Typically, the heterotrophic bone is found within or deep to the supinator muscle as it extends between the proximal radius and ulna. This anterior approach is helpful to visualize and debride the full extent of bridging bone in complex or revision cases of synostosis resection. After the synostosis has been taken down and the exposed bone edges covered with bone wax, the pedicled brachioradialis myofascial interposition flap is raised (see Fig. 13-12).

The tendon of the brachioradialis is divided and the muscle raised from distal to proximal. The previously identified neurovascular structures are left intact. Although the muscle can be released proximally from its origin on the lateral epicondyle and lateral intermuscular septum, this is usually not necessary; the muscle typically provides enough length for interposition between the proximal radius and ulna.

The muscle is interposed between the radius and ulna from anterior to posterior and secured in one of two ways. The first is to secure muscle to the posterior surface of the fully supinated radius using suture anchors or drill holes. When the forearm is pronated, the muscle is advanced into the interosseous space. The second method, introduced by Diego Fernandez, wraps the muscle around the proximal radius, deep to the superficial branch of the radial nerve, and is sutured to itself with locking sutures. Passing the stitches through the muscle fascia as well as through some of the tendonous fibers optimizes the fixation of this interposition flap to the bone.

In addition to being used for an interposition material, the brachioradialis can be used as a myofasciocutaneous flap to provide coverage for medium-sized elbow defects. The skin over the brachioradialis, centered on the radial recurrent artery, may be harvested up to 2 to 3 cm in width and 6 to 10 cm in length from proximal to distal. A narrow skin paddle permits primary closure of the donor site. Alternatively, the muscle alone can be elevated, inset, and split thickness skin graft is applied over it. The brachioradialis flap can be safely rotated to reach soft tissue defects involving the antecubital fossa, the volar one-third of the forearm, and the posterolateral aspect of the elbow between the lateral epicondyle and the olecranon.

Tensor Fascia Lata Interposition

There are certainly cases of heterotopic ossification and radioulnar synostosis in which transposing a local muscle flap may be difficult or impossible. Most frequently these are cases in which the anconeus was injured during the initial traumatic event and where the extensive dissection of the brachioradialis muscle and additional scars are avoidable. In such circumstances, an alternative approach is to use tensor fascia lata (TFL) for interposition. Autograft TFL has the advantage of biocompatibility but the marked disadvantage of donor site morbidity including superficial nerve injury, additional site of scarring, and asymmetry of the thigh secondary to muscle herniation. Given the potential complications and the reality that the interposition graft (whether autograft or allograft) is not living tissue after harvest, we have abandoned the use of autograft TFL in favor of allograft.

The technique is straightforward. Following synostosis exposure and resection, the tensor fascia lata allograft is shaped to cover the exposed bone. A graft 10 to 12 cm long and 4 to 5 cm wide is required in most cases. The tensor fascia lata graft is wrapped around the most accessible bone, usually the ulna, and sutured into place using absorbable sutures (see Fig. 13-13). With a complete wrap, the graft can simply be sutured to itself or neighboring soft tissues to maintain its position. Another technique is to suture the graft to the exposed surface of the ulna, interpose it between the radius and ulna, securing to the exposed surface of the fully pronated radius with suture anchors. As the forearm supinates, the graft follows the radius and provides an effective barrier to reformation of the synostosis. The TFL, once secured to the underlying bone, creates an intact, but not water tight, sleeve around the proximal aspect of one of the forearm bones. The goals of this type of fascial interposition are twofold: to provide painless forearm rotation and to act as a barrier to future bridging heterotopic bone.

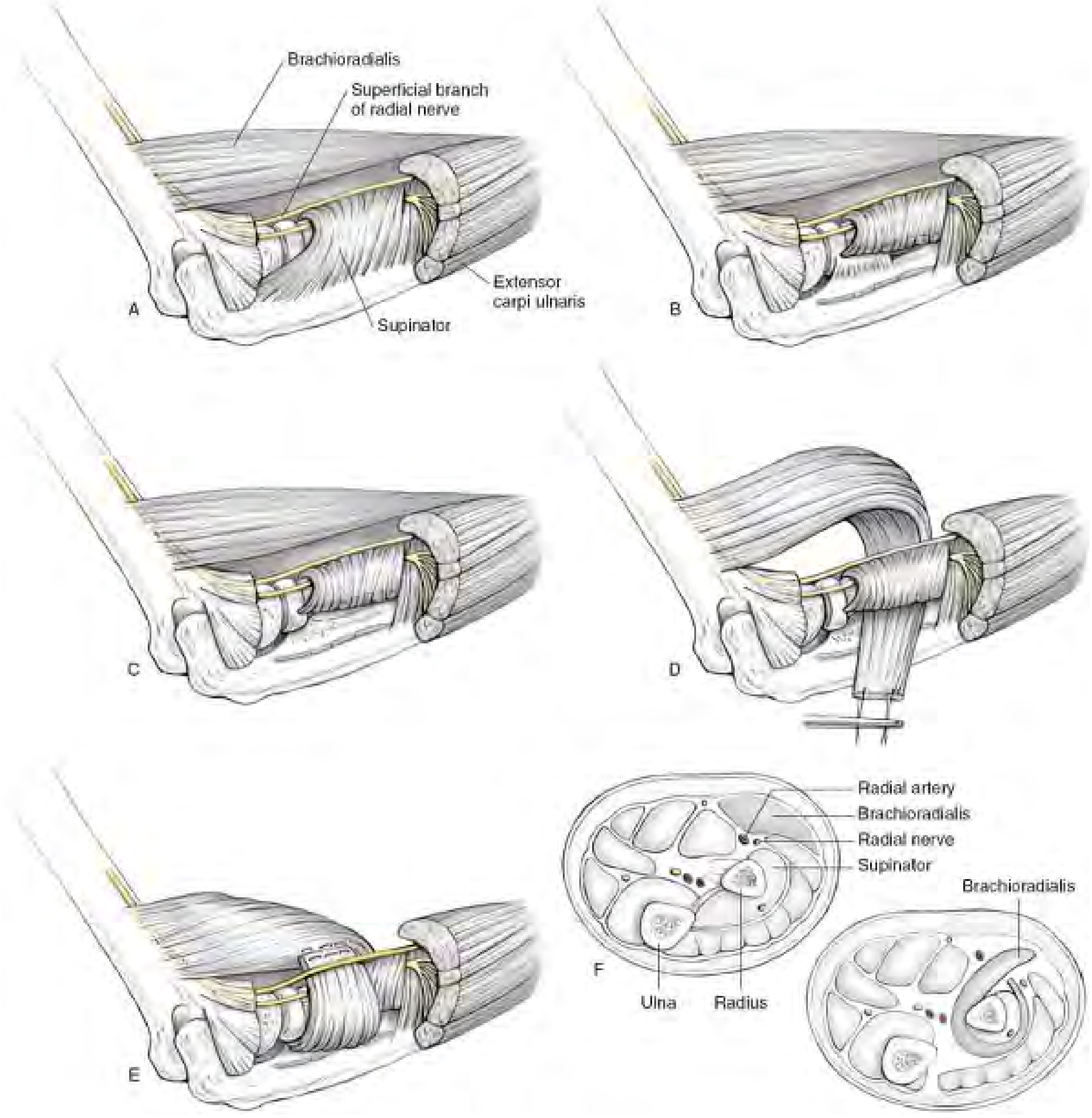
Brachioradialis
Superficial branch of radial nerve
Extensor carpi ulnaris
Supinator
A
B
C
D
E
Radial artery
Brachioradialis
Radial nerve
Supinator
Brachioradialis
F
Ulna
Radius

FIGURE 13-12 A: Posterior lateral approach (Kocher). Through a separate incision, the interval between the anconeus and extensor carpi ulnaris is developed. B: The supinator is sharply elevated off of the ulna exposing the radioulnar synostosis. C: The synostosis is removed and if necessary the interosseous ligament-membrane is excised. D: A separate anterior incision runs just ulnar to the brachioradialis muscle. The tendon is divided distally, the superficial branch of the radial nerve is separated from the under belly of the

muscle, and the vascular bundles entering the muscle in the proximal forearm are preserved. The muscle is passed through the interosseous space created by the resection. E: The tail of the muscle is brought deep to the radial nerve and sutured to the proximal muscle belly. F: Cross-section of the forearm. Note the muscle passes radial to the radial artery in the mid forearm, distal to the biceps tendon and deep to the superficial branch of the radial nerve. (From Fernandez DL, Joneschild E. â€œWrap aroundâ€• pedicled muscle flaps for the treatment of recurrent forearm synostosis. *Tech Hand Up Extrem Surg.* 2004;8(2):102â€“109. With permission.)

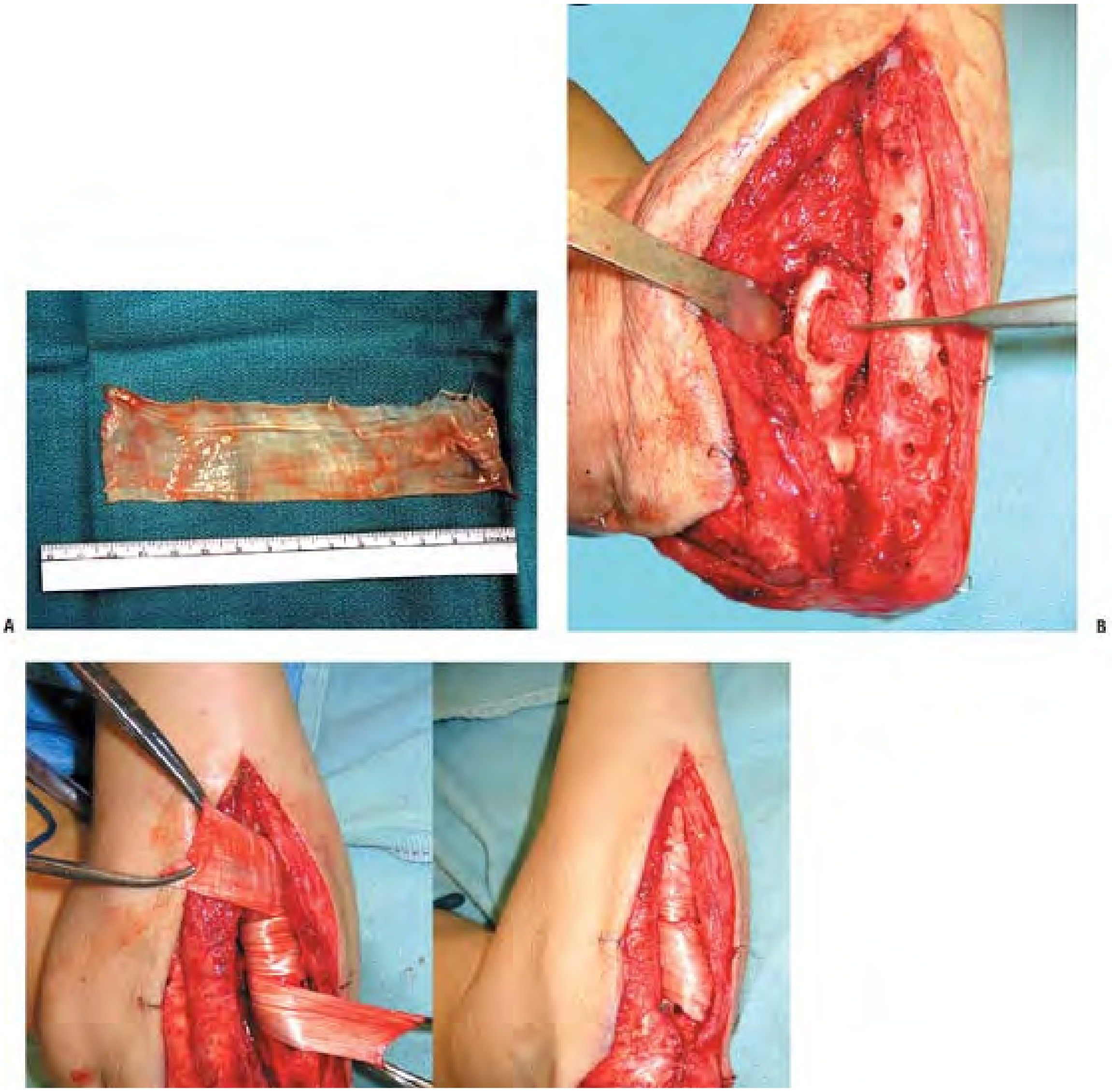

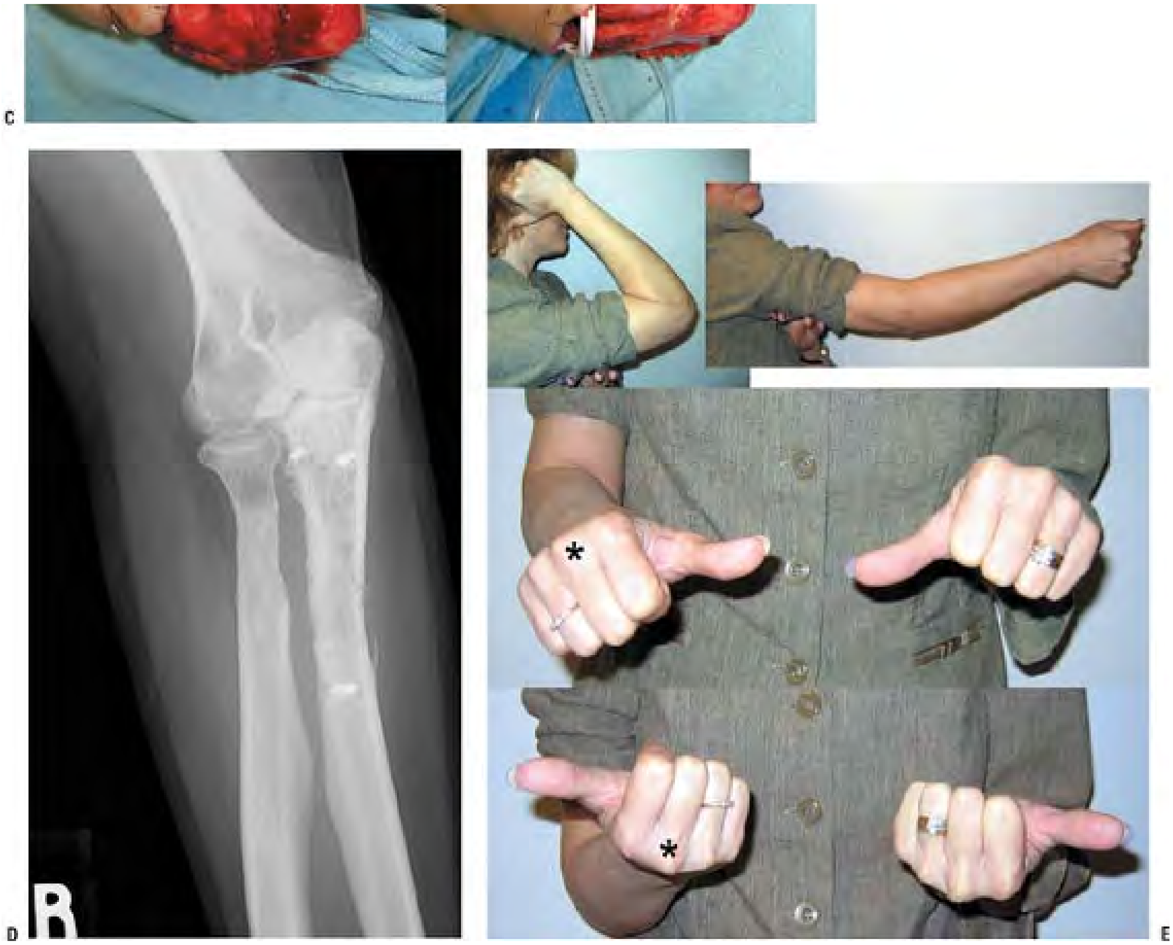

FIGURE 13-13 A: This 4 Ã— 14 cm tensor fascia lata allograft will be used as an interposition material. B: This is the forearm of the patient shown in Figure 13-7. After synostosis resection, the rough surfaces of the radius and the ulna are prone to reformation of the synostosis. C: The tensor fascia lata allograft is wrapped around the ulna and secured with stout absorbable sutures. D: Radiographs of the elbow 6 months postoperatively. E: Range of motion 6 months postoperatively.

Closure

After heterotopic bone excision from an elbow, medial or lateral instability frequently arises. If the collateral ligaments have been excised, they should be reconstructed. Allograft tendon provides an alternative to harvesting autogenous palmaris longus or plantaris for these patients. Using either form of collagen, a tendon reconstruction can be performed with a combination of suture anchors and bone tunnels. Care should be given to reconstructing the anterior band of the medial collateral ligament and the ulnar band of the lateral collateral ligament, as these ligaments have been shown to be instrumental in the maintenance of elbow stability. If stability cannot be achieved with ligamentous reconstruction, a dynamic external fixator or dynamic traction fixator should be applied.

Closed suction drains anterior and posterior to the elbow joint are an integral part of minimizing postoperative swelling and hematoma formation. It is not necessary to place a drain in the site of a synostosis takedown, but it should be placed deep to the fascial layer. Before closing the deep soft tissues, a lidocaine or bupivacaine continuous infusion pump may be placed to help maximize postoperative pain control (in cases in which axillary catheters are not used). The catheter for a bupivacaine infusion pump should be placed in a soft tissue layer or location separate from that of the closed suction drains. Divisions in the intermuscular septi do not need to be sutured, but muscle attachments to the distal humerus or proximal ulna should be repaired. For origins and insertions of muscle about the medial and lateral elbow, suture anchors provide sufficient stability for the bony reattachment. If the medial, flexor-pronator muscle mass has been elevated, consideration can be made to submuscular transposition of the ulnar nerve. Otherwise, a simple subcutaneous transposition maintained with a small strip of flexor-pronator fascia frequently will suffice. The fascial layer may be repaired with absorbable suture in a running or interrupted fashion. Subcutaneous and skin closure follows. The skin closure should not be under tension, nor should it be so tight as to prevent egress of drainage when early motion is initiated.

Postoperative Management

Surgical outcomes of heterotopic bone excision and interposition soft tissue flaps weigh heavily on the patient's postoperative rehabilitation. The affected upper extremity is immobilized in a soft dressing to allow immediate motion. An axillary block catheter placed by the anesthesia team can provide substantial pain relief in the immediate postoperative setting. When such a catheter cannot be placed, we have used continuous bupivacaine infusion pumps to supplement intravenous narcotic pain control.

With adequate analgesia, continuous passive motion of the elbow can be performed in flexion and extension. It is our preference to use a bedside continuous passive motion (CPM) machine for the elbow rather than the more mobile units that provide motion during ambulation. The bedside units tend to maintain a constant position, capitalizing on leverage to impart motion at the patient's elbow. We have not identified a reliable method for continuous passive forearm motion and instead rely on intermittent sessions wherein the extremity is removed from the CPM machine, the elbow flexed 90 degrees, and forearm motion addressed. Resting splints are worn while the patient sleeps at night. The position of the splint is dictated by the particular problems addressed with surgery. For patients with elbow flexion and extension problems, the limb is immobilized in maximum extension and supination. For those patients whose primary problem is forearm motion, the limb is immobilized with elbow flexed 90 degrees and the forearm in maximum supination. Patients are usually discharged on the fourth day postoperatively, and an aggressive home and outpatient therapy program is instituted.

Elbow and forearm specific exercises include active, active-assisted range of motion as well as gentle passive assisted range of motion. To minimize the â€œtrick maneuverâ€• of radiocarpal and intercarpal rotation to augment apparent pronation and supination, the wrist is secured in an immobilizer and a solid object such as a common household tool, a hammer, is grasped tightly in the hand. Doing this directs all pronation and supination efforts to the forearm. When not exercising, the patient should use a static progressive splint, especially at night. The splint is applied an hour before sleep, and the pain associated with initial application should have dissipated enough that sleep will not be interrupted. Dynamic rubber band or spring driven splints are never used. Splinting and exercises are rarely carried out past 6 months postoperatively, as there is usually no change in the range of motion past this date. Anecdotally, however, patients frequently report that even though their range of motion does not change after 6 months, their strength and ease of motion improves up to 2 years postoperatively.

Radiation treatment of the resected bed of heterotopic ossification or synostosis has been shown to be successful in preventing recurrence of ectopic bone. However, no prospective randomized trial has shown this modality to be better than resection alone. As this procedure comes with some inherent risk of wound breakdown, neuritis, and lymphedema, and a remote risk of sarcoma, it should be discussed with the patient preoperatively. We limit radiation therapy to a single dose of 700 cGy within 36 hours of surgery and only in those patients in whom there was ankylosis of the proximal radioulnar joint. All patients are given ketorolac 30 mg IV daily for the 4 days; beyond that, non-steroidal anti-inflammatory medicines are not used.

Complications

While the release of a stiff or ankylosed elbow can provide dramatic improvement in a patient's upper extremity function, the procedure is not without its complications. It is a high-risk operation which requires strong appreciation for the anatomy of the elbow and traversing neurovascular structures. Nerve injury from retraction or surgical dissection is possible and can be minimized with frequent assessment of nerve location as well as limiting the time and rigorousness of retraction on a nerve. Loss of fixation of the interposition material can be avoided by careful suture fixation into muscle as well as fascia.

Certainly, the most frequent complication of this procedure is recurrence of the heterotopic bone. In many cases new calcifications arise on postoperative radiographs well into the rehabilitation period. If the physical examination demonstrates no signs of motion loss, these ectopic calcifications are clinically insignificant and donâ€™t require routine followup. Some motion loss from the gains made intraoperatively can be expected. But a majority of patients who actively participate in their daily motion exercises will maintain a functional arc of motion and achieve an optimal outcome.

Recommended Reading

Alonso-Llames M. Bilaterotricipital approach to the elbow. Its application in the osteosynthesis of supracondylar fractures of the humerus in children. *Acta Orthop Scand.* 1972;43(6):479â€“490.

Bell SN, Benger D. Management of radioulnar synostosis with mobilization, anconeus interposition, and a forearm rotation assist splint. *J Shoulder Elbow Surg*. 1999;8(6):621â€“624.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 14

Principles of Hand Incisions

Marco Rizzo

William P. Cooney III

Indications/Contraindications

Because of the intricate balance of function in the hand and wrist, incisions in this area are unique. In an effort to ensure healing and maintain function, it is important to plan the incisions appropriately. The fingers have well-established approaches that allow for good visualization of deeper structures without compromising healing. Likewise, the hand and wrist are best approached while following basic principles. Poorly placed incisions result in restrictive scarring (Fig. 14-1), which may lead to diminished motion and function. In addition, the limited skin redundancy on the palmar surface of the hand may make delayed closure of edematous wounds difficult. Sound planning of incisions about the hand and wrist will help to avoid some of these difficulties. The purpose of this chapter is to review accepted incision designs for various aspects of the hand and wrist.

Preoperative Planning

Volar Approach to the Fingers

There are several commonly used approaches to the volar aspect of the fingers and thumb. The traditional Bruner incision is the approach that is most common (1). It is a zigzag incision that angles at the flexion creases of the metacarpophalangeal (MP), proximal interphalangeal (PIP), and distal interphalangeal (DIP) joints (Fig. 14-2). As the incision progresses proximally, the flexion creases of the palm are also points of direction changes. The principle is based on the fact that longitudinal incisions across these creases can generate excessive scarring and limit extension. For optimal healing of the flaps, it is best to keep the angles close to 90 degrees. Angles more acute than 60 degrees carry a higher incidence of skin necrosis.

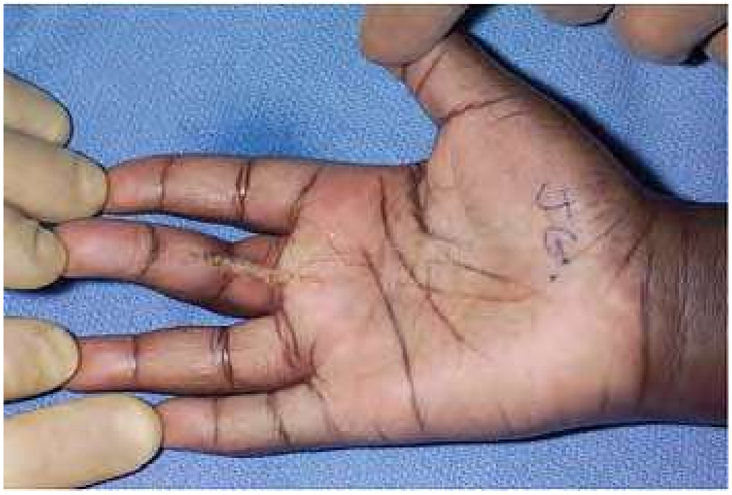

FIGURE 14-1 Example of restrictive scar resulting from straight line incision placed over the flexor surface of the finger.

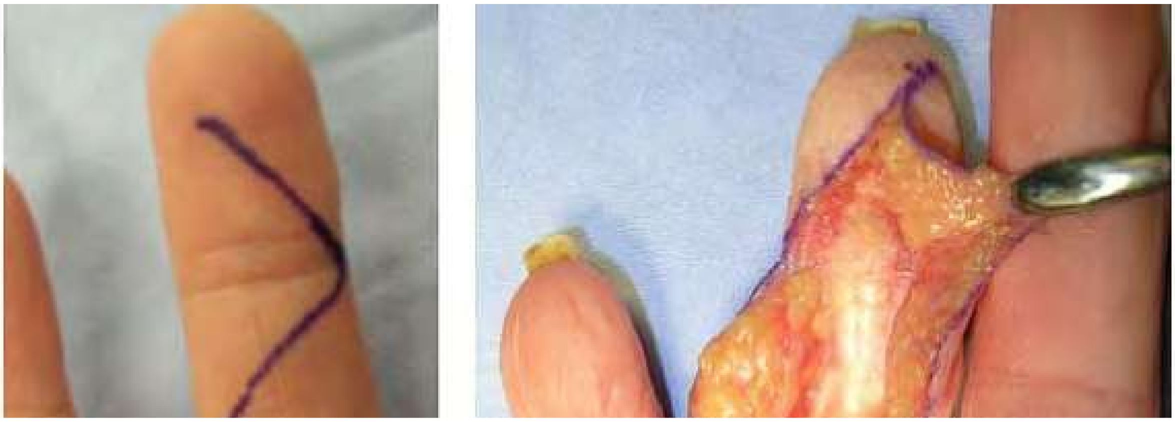

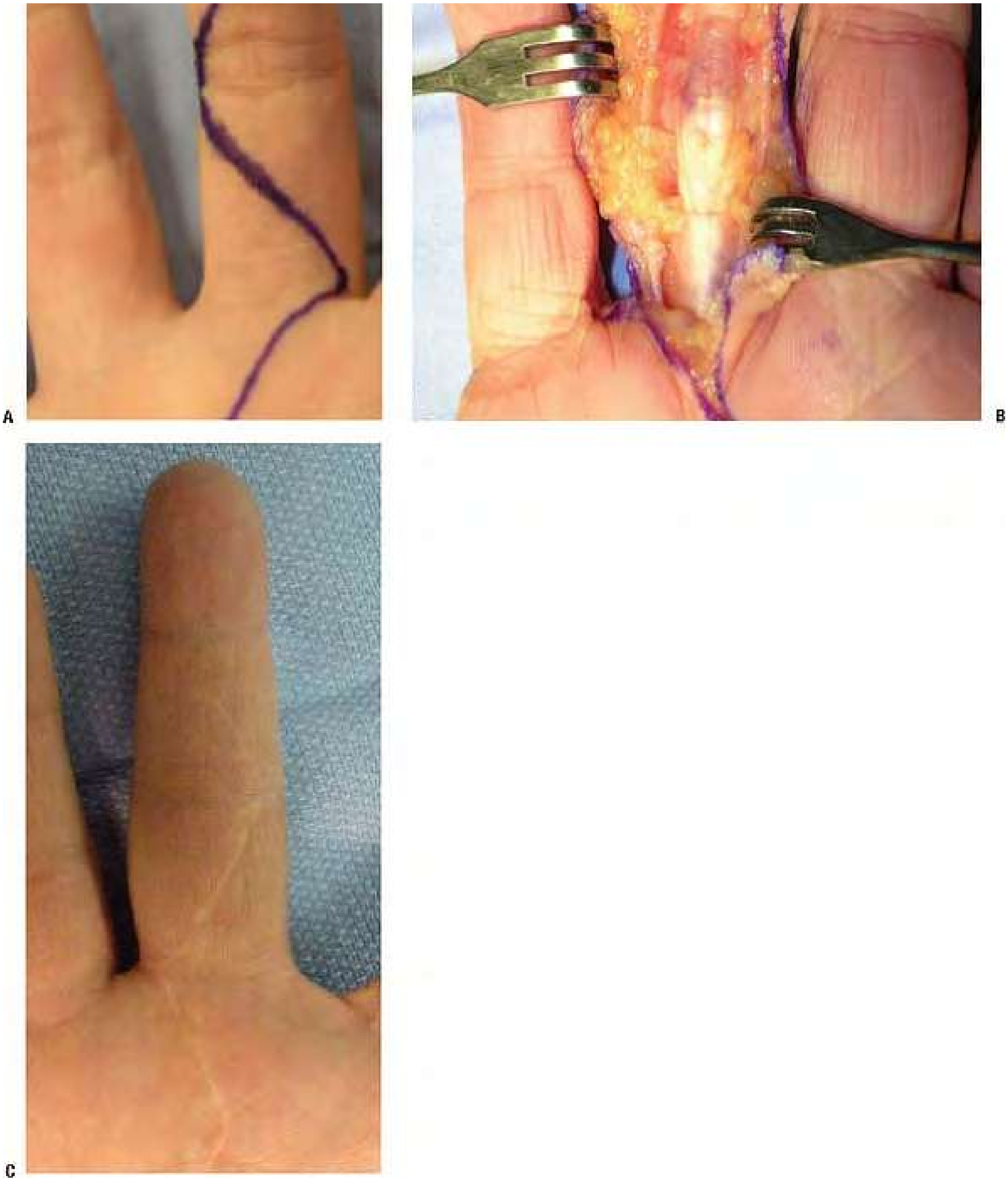

FIGURE 14-2 A: Bruner incision approach to the volar aspect of the fingers. The most commonly preferred method of crossing flexion creases is with these undulating angled incisions. B: Excellent exposure to the volar aspect of the finger can be achieved. C: These scars typically heal nicely.

An alternative to the Bruner incision in the finger is the mid-lateral approach (Figure 3A). It has the advantage of being cosmetically more appealing. The incision is based about the lateral (or medial) side of the finger and the skin is elevated to provide exposure of the underlying tissues. A good way to ensure that the surgeon is not directly over the neurovascular bundles is to connect the incisions at the apex of the flexion creases (Fig. 14-3B). This is also a preferred approach in cases of replantation surgery. Although cosmetically appealing, the challenges of this technique include elevating the skin and ensuring its viability.

In cases of flexion deformity of the fingers such as Dupuytren's contracture, a straight midline incision with z-plasties is an alternative approach to the traditional Bruner incision (Fig. 14-4). This will allow for the elongation of the incision as the flexion deformity is corrected. Angles of 60 degrees help preserve viability at the apex of the skin flap.

Incisions over the distal pulp of the finger, such as is necessary in felons, are best made directly over the area of swelling and induration. These wounds can be loosely closed or allowed to heal via secondary intention. If it appears that the incision may need to be extended proximal to the distal interphalangeal joint, a Bruner incision may be preferred for better exposure.

Dorsal Incisions About the Fingers

The dorsal skin of the hand lacks glaborous connections and has more redundancy, due to the requirements of joint flexion. Because of this redundancy, longitudinal incisions are acceptable. A longitudinal incision allows (Fig. 14-5) for protection of the venous and lymphatic drainage, while giving good visualization of the extensor mechanism (2). Lazy S or curvilinear incisions may also be designed to avoid crossing extension creases (see Fig. 14-5). Incisions over the distal interphalangeal joint can be longitudinal as well, but dissection can be limited due to the nail bed and fold distally. An alternative is the â€œTâ€• or â€œHâ€•-shaped incision that allows for visualization of the extensor mechanism and distal interphalangeal joint (Fig. 14-6).

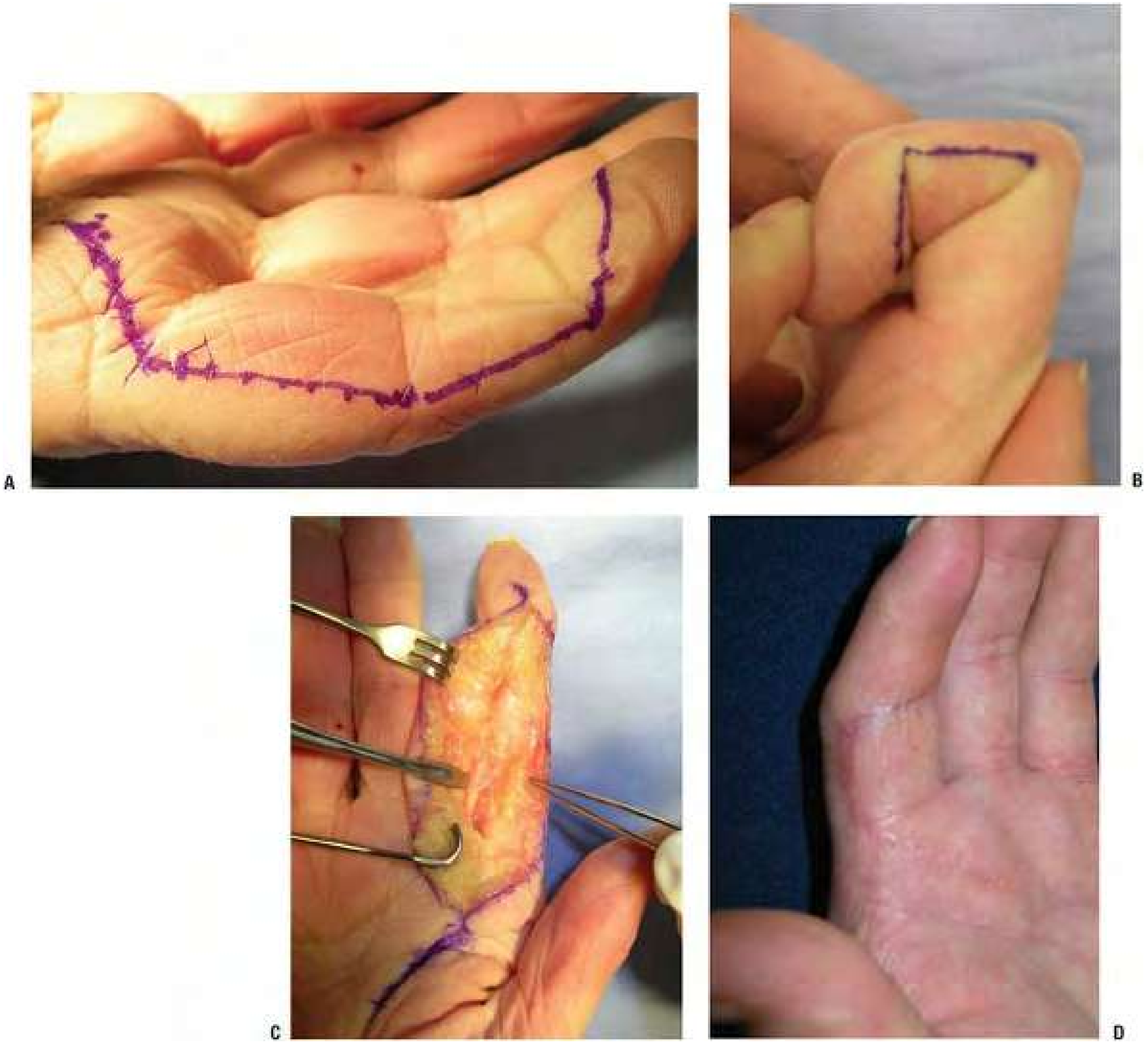

FIGURE 14-3 A: The mid-axial or lateral approach to the fingers. This technique can be a more aesthetic alternative to the Bruner incision. B: Care must be taken to connect the incision at the apex of the flexion creases in order to avoid direct dissection over the underlying neurovascular structures and prevent excessive volar scarring. C: Although it requires more dissection, excellent exposure can be achieved. D: An example of a healed mid-lateral incision.

Volar Incisions About the Hand

Transverse incisions or incisions that parallel the flexion creases offer the best cosmetic results. However, these incisions are perpendicular to the underlying vessels, nerves, and tendons. Therefore, care must be taken when performing these approaches. Incisions for trigger finger releases are a good example (Fig. 14-7). In particular, transverse incisions, in line with the metacarpophalangeal joint flexion crease, at the base of the thumb for A1 pulley release have been found to produce better aesthetic results and result in less restrictive scarring. Traditional longitudinal Bruner-type incisions are commonly used and are useful when more exposure is necessary. Again, you want to follow the basic premise of changing directions at the creases to minimize scarring. Wounds or scars in the web spaces are at risk of developing functionally limiting scars. Depending on the patient, these can sometimes necessitate revision z-plasties and web-space deepening.

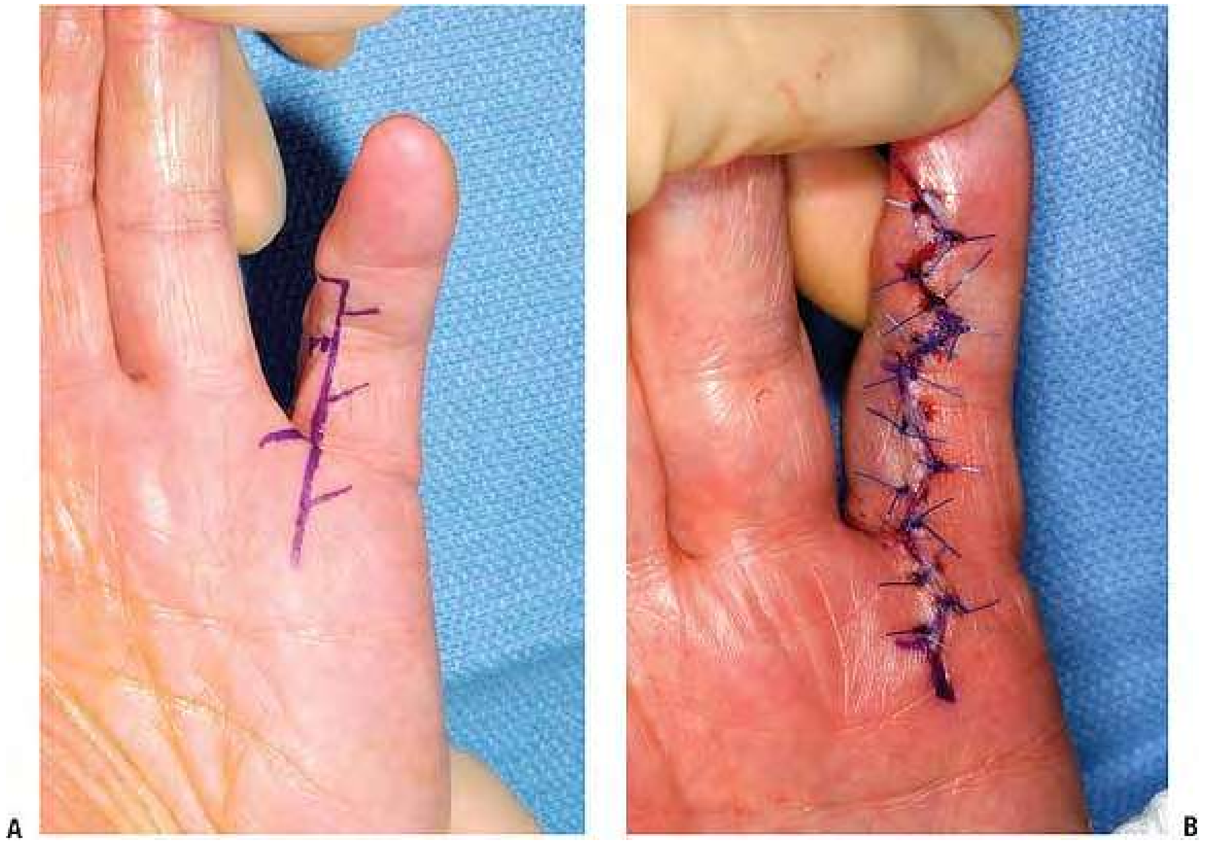

FIGURE 14-4 A: A longitudinal z-plasty approach to the volar aspect of the finger can be useful in cases of pre-existing flexion contracture, such as in Dupuytren's contracture. B: Wound closure is easily achieved with little or no tension with the finger in full extension.

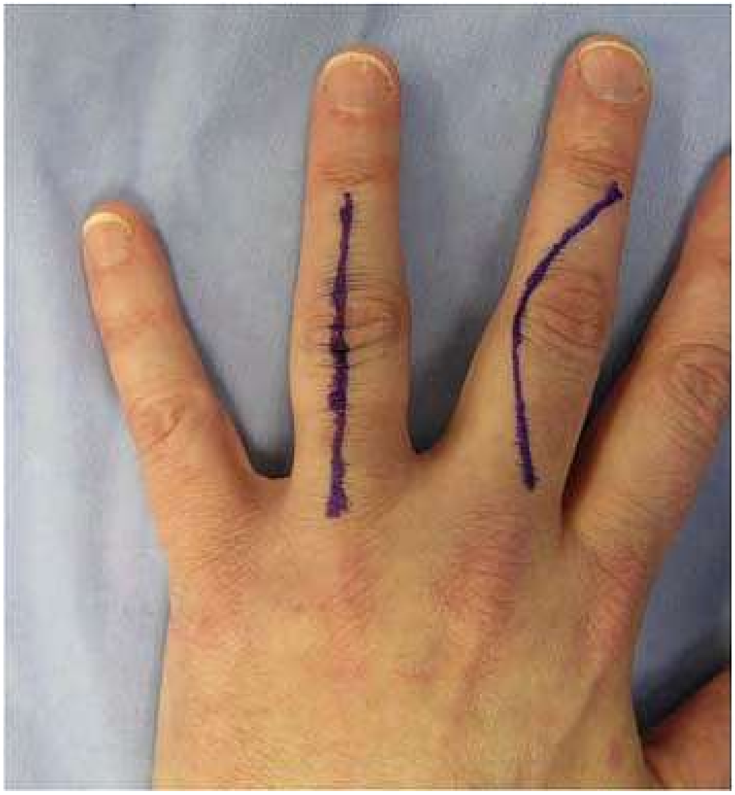

FIGURE 14-5 Two common approaches to the dorsum of the fingers: a straight longitudinal and a more curved longitudinal, which avoids the extension creases.

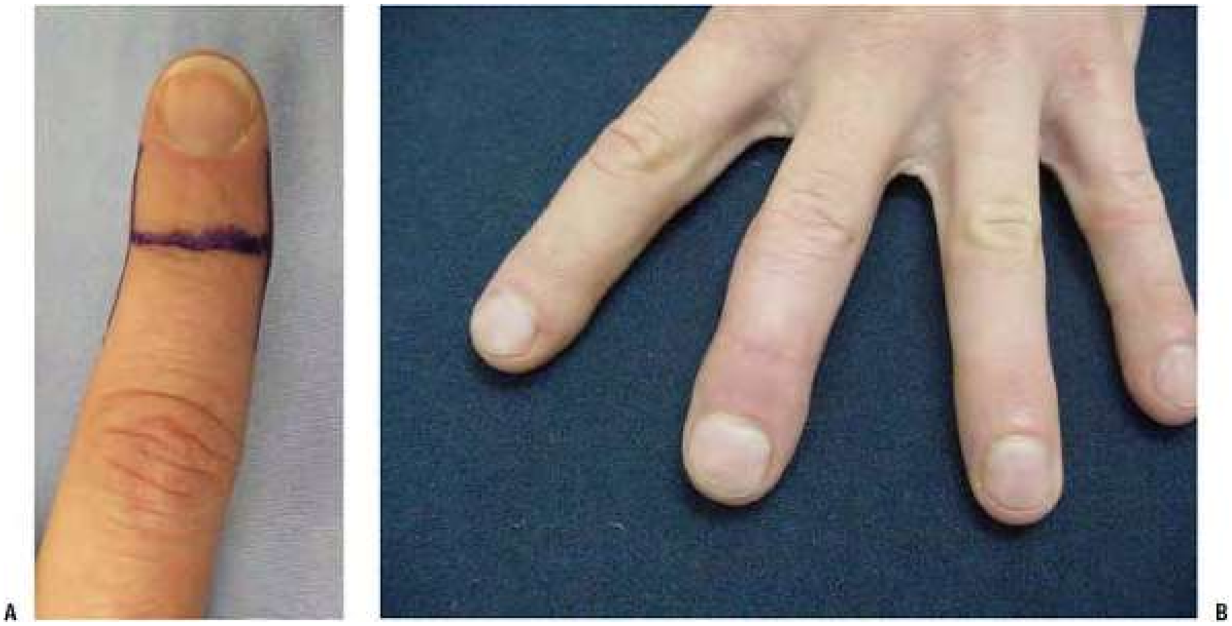

FIGURE 14-6 A: An â€œHâ€•-shaped incision over the dorsal aspect of the distal interphalangeal (DIP) joint allows for excellent exposure of the dorsal distal finger. B: The wounds typically heal very nicely.

In exposing the carpal tunnel, a longitudinal incision, preferably in or in line with the palmar crease, is commonly used. If the incision needs to be extended proximally, Bruner-type zigzag incisions can be made between the distal and proximal wrist flexion creases (Fig. 14-8). These angled incisions are the most common and useful method of crossing flexion creases in the hand and wrist. Care should be taken to stay ulnar to the palmaris longus tendon to avoid injury to the palmar sensory branch of the median nerve.

Dorsal Incisions About the Hand

Longitudinal incision can be successfully used over the dorsum of the hand without the risk of restrictive scarring. Transverse and curvilinear incisions can also be used over the dorsum of the hand if desired. The skin over the dorsum of the hand is mobile and scarring is generally less of a functional problem as the flexor tendons are the dominant force and will minimize the cosmetic and/or functional deficit. Approaches to the carpometacarpal joint can be made longitudinally between the abductor pollicis longus and extensor pollicis brevis tendons, or they can be more radial (along the

plane of the intersection of the dorsal and volar skin) and extended proximally transversely along the wrist flexion crease, forming a â€œhockey stickâ€• appearance.

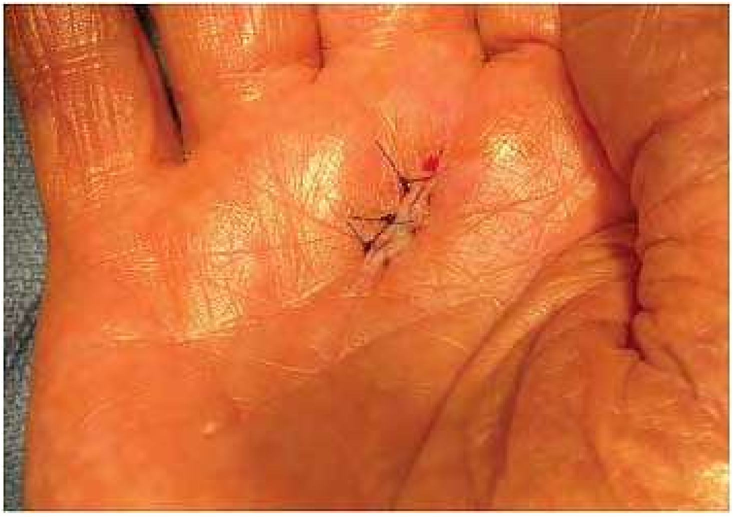

FIGURE 14-7 An incision for trigger finger approach using the palm flexion crease. These can result in cosmetically appealing scars with little or no residual deficit. In cases where greater exposure is necessary, a longitudinal Bruner-type incision is preferred.

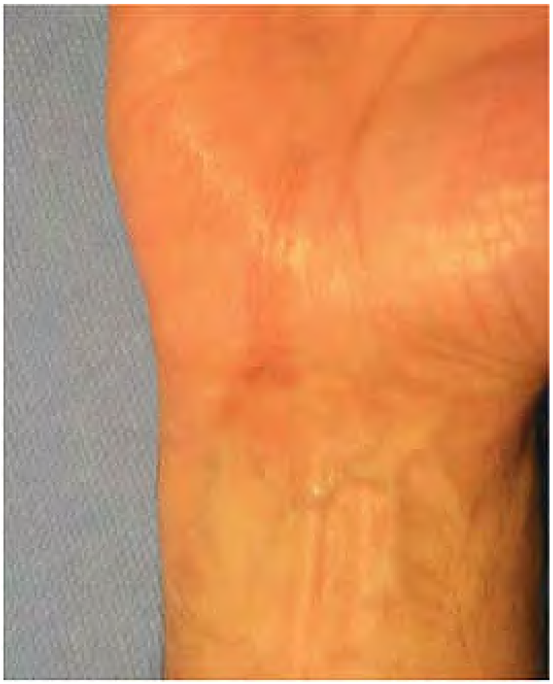

FIGURE 14-8 An extensile carpal tunnel incision. Angulation at the wrist flexion crease will minimize scarring of the wound. By keeping the incision ulnar to the palmaris longus tendon (*dotted line*), injury to the palmar sensory branch of the median nerve can be avoided.

Incisions About the Wrist

Dorsal approaches to the wrist can be longitudinal, transverse, or curvilinear. The tissues on the dorsal aspect of the wrist are redundant and dorsal scarring is not typically a problem. These can include incisions to expose the base of the thumb. Volar incisions to the wrist can be along flexion creases or longitudinal provided the incision changes direction at the flexion creases in a modified Bruner fashion.

Incisions About the Forearm

As we extend proximally, more traditional incisions can be used. If it is a fairly short incision, such as in open treatment of distal radius fractures, a longitudinal incision will suffice (Fig. 14-9). If the incision is extended distally, a Bruner angulation is made at the wrist flexion crease. In cases where extensive exposure is necessary, such as is required in compartment syndrome release, an undulating longitudinal incision works well. If the elbow must be crossed, this is usually done in a lazy S fashion passing to the medial aspect of the anticubital fossa.

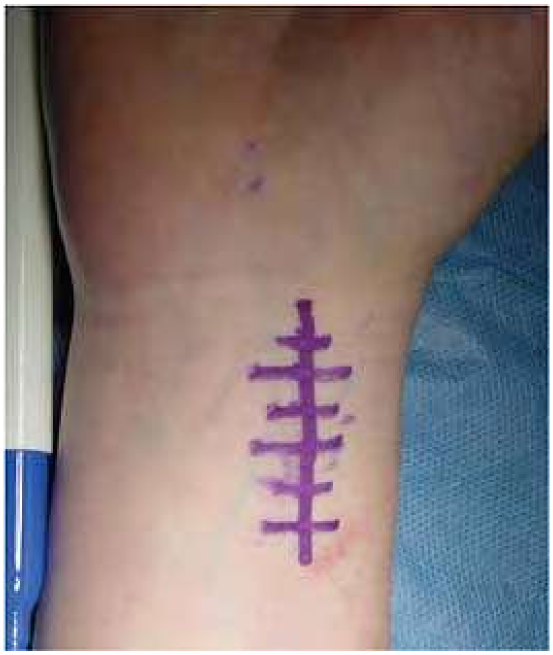

FIGURE 14-9 A longitudinal approach to the distal radius in a case of open reduction and internal fixation of a distal radius fracture. Angulation of the incision can be performed if the incision needs to be extended distally.

Healing by Secondary Intention

In cases such as infection and Dupuytren's contracture, it may be advantageous to leave the wounds at least partially open. These wounds can heal nicely by secondary intention. Modalities such as whirlpool and wound vacuum assisted closure can be helpful in accelerating wound healing as well as minimizing risk of infection. Disadvantages include time to healing and the fact that some patients find the healing process cosmetically unappealing. Unfortunately, functional limitations such as joint contractures can result as rehabilitation is delayed during the wound healing process. However, with good postoperative wound care, this can be an effective way to manage some of the more difficult cases or cases where it is necessary to leave the wounds open (3).

Surgery

Patient Positioning

Typically, the patient is in a supine position and a hand table is used. A tourniquet is placed in the proximal portion of the arm. We prefer the pressure to be inflated to approximately 250 mm Hg for adults and 200 to 225 mm Hg for children (generally 70â€“100 mm Hg above the patient's systolic pressure). If the surgery is isolated to the distal aspect of the finger, a finger tourniquet can be placed. In these cases, we routinely also place the upper arm tourniquet which can be used if need be. The entire extremity is prepped. Perioperative antibiotics are administered when indicated.

Technique

Loupe magnification should be used to help avoid injury to important underlying structures. The incision is typically made with a 15-blade knife, taking care to stay perpendicular to the skin and not skiving. The knife can be used to cut the dermis and epidermis. Generally, dissection should proceed from known to unknown or normal to abnormal tissues. Electrocautery, with needle-point cautery, can be used for small veins or vessels. We prefer the use of bipolar cautery. Drains should be liberally used, especially in contaminated wounds or wounds with compromised hemostasis. Options include a small silastic, penrose, or hemovac drain. The drain can be placed without suture so that it can be easily removed postoperatively to not disturb the overlying dressing. Otherwise, drains can be sutured for added protection if the surgeon wants it kept it in for a longer period of time. It is our preference to deflate the tourniquet before wound closure to evaluate hand perfusion as well as hemostasis prior to wound closure. After ensuring adequate hemostasis and viability of the hand, the tourniquet may be reinflated during wound closure. The wounds can be closed in a single layer with a nonabsorbable interrupted 4-0 or 5-0 suture such as nylon or prolene. Either horizontal or vertical mattress sutures can be used. Vertical mattresses are better for everting the skin edges, whereas horizontal mattress sutures allow for less overall sutures. Increased ischemia of the soft tissues can occur with horizontal mattress closure. In young children, an absorbable suture such as catgut or chromic is preferred.

Postoperative Management

A thoughtfully applied hand dressing is imperative. Depending on the injury, some period of immobilization is generally recommended. Proper immobilization will allow for the skin and soft tissues to heal with minimal stretch and tension. However, this needs to be balanced with the risk of developing joint contractures. In cases of fracture care, immobilization time is prolonged compared with that for soft tissues. The dressing should be mildly compressive and supportive (Fig. 14-10). Surgeries involving the distal aspects of the digits may be adequately dressed with a tube gauze type of dressing. More proximal reconstructive surgeries will require varying periods of immobilization of the wrist and hand in the position of function. Specifically, with the wrist in mild extension, metacarpophalangeal joints flexed and interphalangeal joints extended. Elective soft tissue procedures such as carpal tunnel releases can be immobilized with the wrist in mild extension and the fingers free to move immediately. Although this is our preference, we acknowledge these can be successfully treated with simply a soft dressing. Drains are discontinued when the wound is dry.

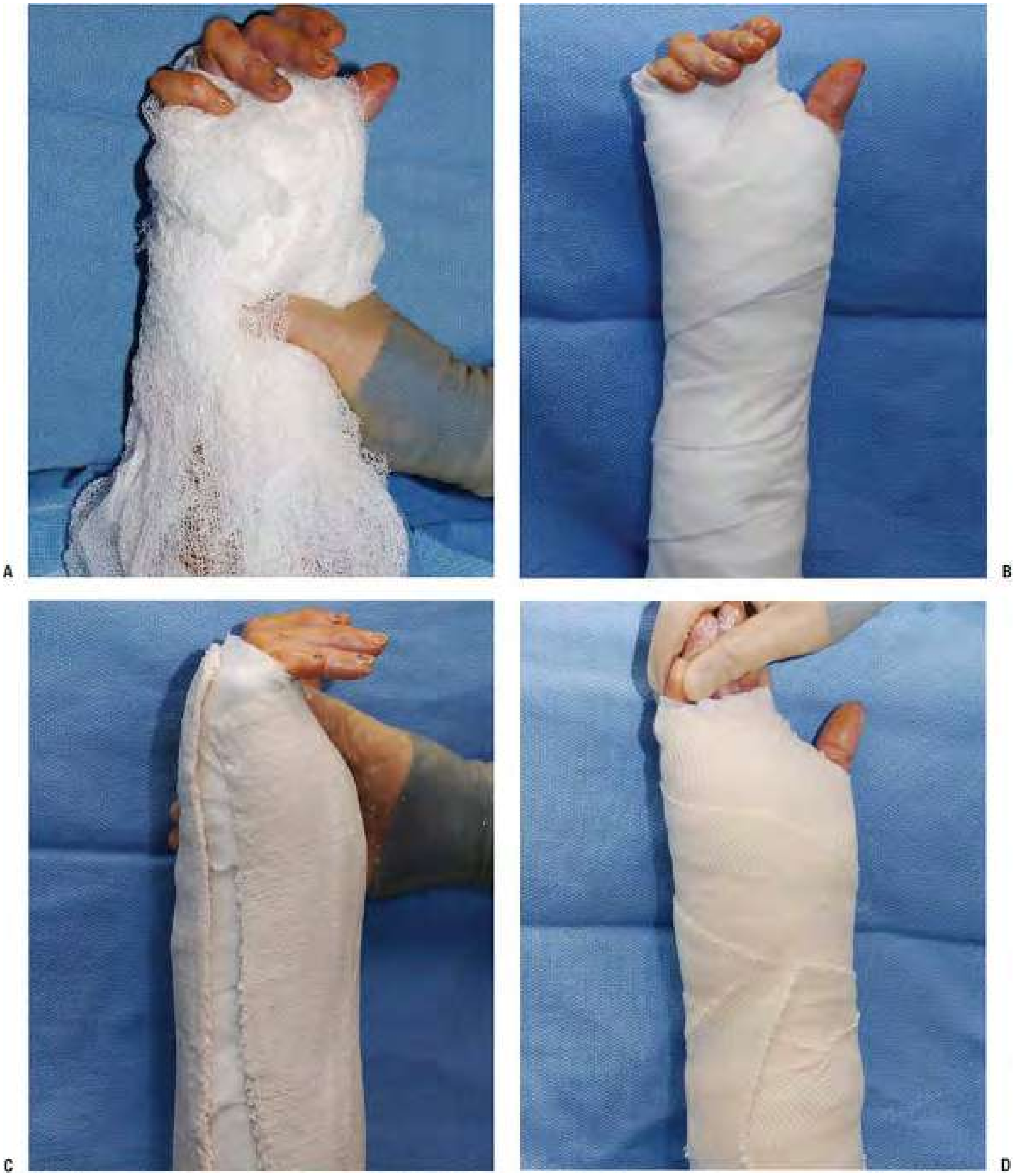

FIGURE 14-10 An example of a compressive dressing in a patient with rheumatoid arthritis who underwent metacarpophalangeal joint arthroplasty. A: Following application of xeroform on the incision, fluffs are placed on the volar, dorsal, radial, and ulnar aspects of the wrist and hand. Gauze is lightly placed within the

web spaces of the digits. B: Abundant cast padding is applied followed by (C) plaster. In the immediate postoperative period, we prefer a splint as it will allow for swelling and will compress as the swelling improves. D: A bias wrap completes the dressing. In this example, the MP joints need to be immobilized in extension and neutral deviation. If finger motion is not contraindicated, care should be taken not to have the dressing restrict finger motion.

Complications

The most common adverse result of poorly planned incisions is scar contracture (see Fig. 14-1), which can lead to cosmetic and/or functional deficits. More difficult are skin necrosis and subsequent wound healing problems. These can lead to problems such as infection, excessive scar formation, and diminished function. Other less common complications related to inadequate or suboptimal incisions include injury to nerves, arteries, or tendons. However, with good preoperative planning and technique, these complications are usually avoidable.

References

1. Bruner JM. The zigzag volar digital incision for flexor tendon surgery. *Plast Reconstr Surg.* 1967;40:571.

2. Graham WP. Incisions, amputations, and skin grafting in the hand. *Orthop Clin North Am.* 1970;1:227.

3. McCash CR. The open palm technique in Dupuytren's Contracture. *Br J Plast Surg.* 1964;17:271.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 15

Lateral Arm Flap for Hand and Wrist Coverage

Michael Sauerbier

Goetz A. Giessler

The first description of the lateral arm flap was by Song and coworkers. They used it primarily as a free microvascular flap for small and medium-sized defects in reconstructive procedures of the head and neck. A detailed study by Katsaros et al later popularized it as an extremely versatile, easy to harvest septocutaneous flap with relatively low donor site morbidity. The vascular anatomy is reliable, and the dissection is rather quick and can be performed under regional anesthesia for reconstruction of hand and wrist defects. To date, the flap is used for a wide spectrum of defects all over the body.

The shortcomings of the original lateral arm flap were the flap's bulkiness and a pedicle of moderate length when skin flaps were centered over the middle aspect of the lateral arm. Recent anatomic studies have led to significant modifications in flap design; the skin paddle is now centered over the lower third of the upper arm extending to the lateral epicondyle (extended lateral arm flap). This produces a thin flap with a long pedicle length. The flap may also be used in a reversed fashion (reversed lateral arm flap) or as a pedicled antegrade V-Y flap (extreme lateral arm flap). A pure fascial flap has also been described. The flap may be designed to include vascularized nerve, tendon, and bone. This versatility has made the lateral arm flap a â€œworkhorseâ€• for elective, urgent, and emergency reconstructions of the hand and wrist.

Indications/Contraindications

Indications

The lateral arm flap has a wide range of applications for closure of small and medium-sized defects in hand and wrist surgery. The classic fasciocutaneous or fascial flap serves well for reconstruction of the dorsal aspect of the hand or the thumb index web space. Patients with a thin subcutaneous fat layer are the best candidates for the flap because it is soft and easily pliable for complex three-dimensional defects.

For cases of palmar reconstruction, we prefer to use a fascial flap, as the cutaneous lateral arm flap may be too thick. This flap is then covered with a full thickness graft and made sensate by anastomising the posterior cutaneous forearm nerve to the palmar branch of the median nerve. Defects with loss of extensor tendons or metacarpal bone can be closed using a composite lateral arm flap using a distal humeral cortical segment and/or a vascularized triceps tendon graft. For all these indications, the flap has to be used as a free microvascular flap. The pedicle is long enough (see following) for flap positioning around the hand or wrist in many cases. The main alternatives to the lateral arm flap are the (osteo-) cutaneous scapular/parascapular, radial forearm, enterolateral thigh flap, and dorsalis pedis flaps. The specific features of each are provided in Table 15-1 .

Contraindications

The free lateral arm flap represents a microsurgical procedure of considerable length. Patients with poor general health who need an elective procedure should be optimized preoperatively; otherwise,

other treatment options should be considered. Previous trauma or surgical procedures in this area, such as plate osteosyntheses of the humerus, probably have destroyed the perforators and are a contraindication. Patients with considerable body fat provide a bulky lateral arm skin flap which is unsuitable for most hand and wrist reconstruction. Using the fasciocutaneous lateral arm flap together with a skin graft can be an option in these cases. The relatively conspicuous scar of the lateral arm flap poses a relative contraindication in female patients. Whenever the donor site cannot be closed primarily, other flap options should be thought of (see previously). Hair transfer may pose a problem in head and neck reconstructions. Likewise, male patients may prove a challenge with hair transfer in hand and wrist reconstructions.

Max. skin island size (cm)
8 Ã— 25
10 Ã— 20
7 Ã— 25
15 Ã— 28
5 Ã— 8
Max. pedicle length (cm)
11
15
15
10
8
Fascia only
Yes
Yes
â€”
â€”
â€”
Tendon component
Lateral triceps tendon
Palmaris longus, half of the radial carpal flexor
Fascia lata
â€”
Short toe extensors
Muscle component
â€”
â€”
â€”

Latissimus dorsi, anterior serratus
No
Bone component
Distal humerus
Distal radius
â€”
Lateral scapular border
2nd metatarsal
Nerve component
Posterior cutaneous forearm nerve
Lateral antebrachial cutaneous nerve
Lateral femoral cutaneous nerve
â€”
Superficial peroneal nerve
Patient positioning
Supine or lateral prone
Supine or lateral prone
Supine
Lateral prone
Supine

	Lateral Arm Flap	Radial Forearm Flap	Anterolateral Thigh Flap	Scapular Parascapular Flap	Dorsalis Pedis Flap

Table 15-1. Comparison of the Lateral Arm Flap to Possible Alternatives for Simple or Composite Reconstructions at the Hand and Wrist

Anatomy

The skin on the upper lateral arm receives its perfusion by septocutaneous perforators in the lateral intermuscular septum inferior to the tip of the deltoid muscle. They originate from two main branches in the middle of the upper arm between the acromion and the lateral epicondyle originating from the profunda brachii artery. Those main branches are called the anterior and posterior radial collateral arteries (ARCA and PRCA), with the latter having a reliable longitudinal network along the septum to distal arteries. The PRCA is the nourishing artery for the lateral arm flap as it runs close the humerus in the septum between triceps brachii and the brachial muscle. The ARCA branches off between the brachial and brachioradial muscles and is ligated during flap elevation. In rare cases (about 6%), the PRCA is found to be duplicated. On its way distally to the plexus around the elbow, the PRCA gives off three to five septocutaneous perforators serving as feeder vessels for the lateral arm fascia or fasciocutaneous flap. Those surfacing perforators are easily visible during flap dissection. After reaching the subcutaneous plexus, the branches run both anteriorly and posteriorly. In this region, the PRCA anastomoses with the interosseus recurrent artery (between the lateral epicondyle and the olecranon), the recurrent radial artery (anterior to the lateral epicondyle directly above the periosteum), and the inferior ulnar collateral artery to this plexus. These multiple arterial connections allow for distal extension of the skin paddle, which can be harvested up to 12 cm distal to the epicondyle. These anastomoses also allow for the use of the flap in a reverse fashion using one of the perforators from this periarticular plexus.

The venous drainage of the lateral arm flap is by veins of the PRCA together with a rich network of subcutaneous veins including the cephalic vein, which may be included in the anterior part of larger flaps. The cephalic vein does not need to be included in smaller flaps necessarily, but can provide adequate additional venous drainage in larger or composite flaps.

Depending on flap positioning, the overall pedicle length can reach up to 11 cm in the authorsâ€™ experience, with the artery having a diameter of 2 to 2.5 mm at its proximal origin from the brachial artery.

Bone, tendon, and nerve may be included within the lateral arm flap and used for reconstruction of composite defects. A strip of vascularized bone may be harvested from the distal, epicondylar section of the lateral humerus. Detailed studies have shown one to four nutrient vessels from the PRCA to the bone. These vessels enter the lateral humeral aspect 2 to 7 cm proximal to the lateral epicondyle. Fifty percent of the triceps tendon may also be harvested with the flap, resulting in a maximum tendon graft of 10 Ã— 2 cm. Gosain et al demonstrated by injection studies that the blood supply to this strip is not via the lateral intermuscular septum, but rather through the triceps muscle. Thus, a cuff of muscle must be included to have a vascularized tendon segment; alternatively, an avascular triceps tendon autograft may be dissected with the flap if one does not wish to violate the triceps muscle. Finally, the posterior cutaneous forearm nerve can be included within the flap allowing for the possibility of a sensate free flap. Harvest of this nerve will result in a strip of distal forearm numbness postoperatively.

Preoperative Planning

Doppler examination can be used for centering the skin paddle over the vascular pedicle. The flap can be harvested in regional anesthesia using a sterile tourniquet (250â€“300 mm Hg). For a transplantation to the contralateral arm, general anesthesia is adequate. Loupe magnification for flap dissection is strongly recommended (2â€“4.5 times). The standard lateral arm flap is centered over a virtual axis between the tip of the deltoid muscle and the easily palpable prominence of the lateral epicondyle of the humerus (Fig. 15-1). When marking this axis in patients who are obese or with very mobile skin, the surgeon has to support redundant tissue of the posterolateral upper arm, as gravity might pull the tissues posteriorly, endangering pedicle inclusion. When flap design is continued beyond the lateral epicondyle, the central axis of the flap is continued distally from the lateral epicondyle in the direction of the distal radioulnar joint on the wrist dorsum. Harvesting the flap from the distal humerus and dorsal forearm produces a thinner flap than flaps raised at the midhumeral level. The reversed or extreme lateral arm flaps have their major indications as a pedicled flap for defect closure around the elbow and proximal forearm.

In planning a composite lateral arm flap, a precise idea about the future positioning of the several components at the recipient site is mandatory. If an osteocutaneous flap is dissected, it should be centered at the distal upper arm, as the receivable bone graft has a maximum length of about 10 cm from the lateral epicondyle proximally. It should not be wider than 25% of the humeral circumference, which is sufficient for most indications at the hand and wrist.

The maximal cutaneous flap area supported by the PRCA on the upper arm varies in size from about 5 Ã— 10 cm to 8 Ã— 20 cm. Primary closure is ideal but this is dependent on skin laxity, patient habitus, and general tissue conditions. Preoperative expansion is described in the literature with the expander being placed subfascially and medially to the lateral brachial septum; but this is certainly reserved for rare indications. Flaps less than 6 cm in width can usually be closed primarily.

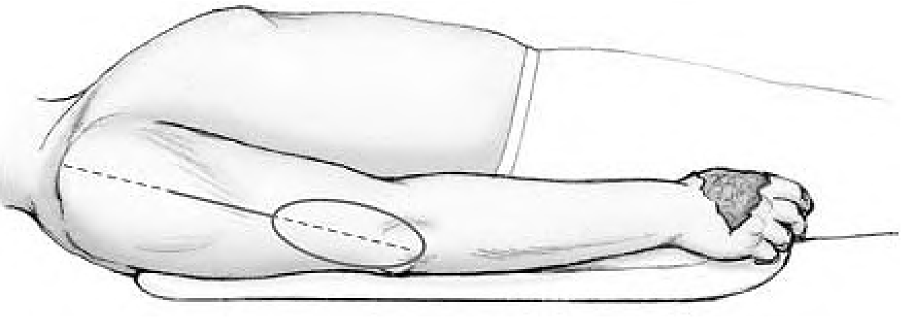

FIGURE 15-1 Planning the axis of a standard lateral arm flap.

Surgery

Patient Positioning

The patient is placed in the supine or lateral supine position with the extremity on a hand table and the arm draped fully mobile. The tip of deltoid muscle, lateral epicondyle, and dorsal aspect of the radioulnar joint are marked and connected as the flap axis (see Preoperative Planning).

Technique

Dissection is started under tourniquet control. Skin and deep fascia are dissected from the muscles together and the fascia is tacked to the skin by 4-0 resorbable sutures to prevent shear forces on the skin perforaters. In an extended or distal free lateral arm flap, it is likewise important to include the deep fascia together with a small strip of intermuscular fascia. Over the epicondylar prominence, the periosteum should be taken en bloc with the fascia to ensure including the connections to the recurrent radial artery. Retracting the lateral head of the triceps posteriorly, the septum, its attachments to the humerus, the PRCA, and the septocutaneous perforators are visible. Once the posterior exposure is complete, the anterior incision is made and the lateral intermuscular septum is exposed from the anterior. While doing this, the surgeon often encounters the smaller, more oblique oriented septum between the brachial and the brachioradial muscle, which contains the ARCA. The identification of the PRCA perforators is done by carefully flipping the flap over along the axis to look at the lateral septum from both sides, after which all smaller branches from the ARCA are ligated. If the cephalic vein runs through the skin island, it may be dissected out more proximally to have an â€œemergency run-offâ€• and included in the flap, or left in situ. However, it is not obligatory for a successful lateral arm flap. Now the flap is very mobile and only attached to the septum (Fig. 15-2). A self-retaining retractor now can be inserted. In a fasciocutaneous or purely fascial flap, the periosteum is sharply incised bilaterally to the PRCA and sharply taken off the humerus with a periosteal elevator, as the PRCA often runs very close to it. This implies ligating smaller nourishing branches to the bone in an area of 2 to 7 cm proximal to the lateral epicondyle. The vascular pedicle and the cutaneous nerve then are developed proximally under the deltoid to gain enough pedicle length, which should match with the recipient site situation at this time. Dissecting the cutaneous nerves from the radial nerve has to be done with utmost care to prevent uncontrolled bleeding and hematoma around the nerve. Finally, the tourniquet is released, precise bipolar coagulation is performed, and the flap perfusion is checked. The pedicle is ligated only after proper dissection of the recipient vessels and shortly before microsurgical anastomosis to save ischemia time.

If an osteocutaneous composite lateral arm flap is dissected, the periosteum is not stripped off the bone, but only incised and an oscillating saw used for corticotomy (see Preoperative Planning). The release of the bone segment is carefully finished with a sharp osteotome.

The tendon harvest of a triceps tendon requires identification of the feeding branches from the PRCA into the lateral triceps head. A cuff of muscle is included with the tendon for proper perfusion (see Anatomy).

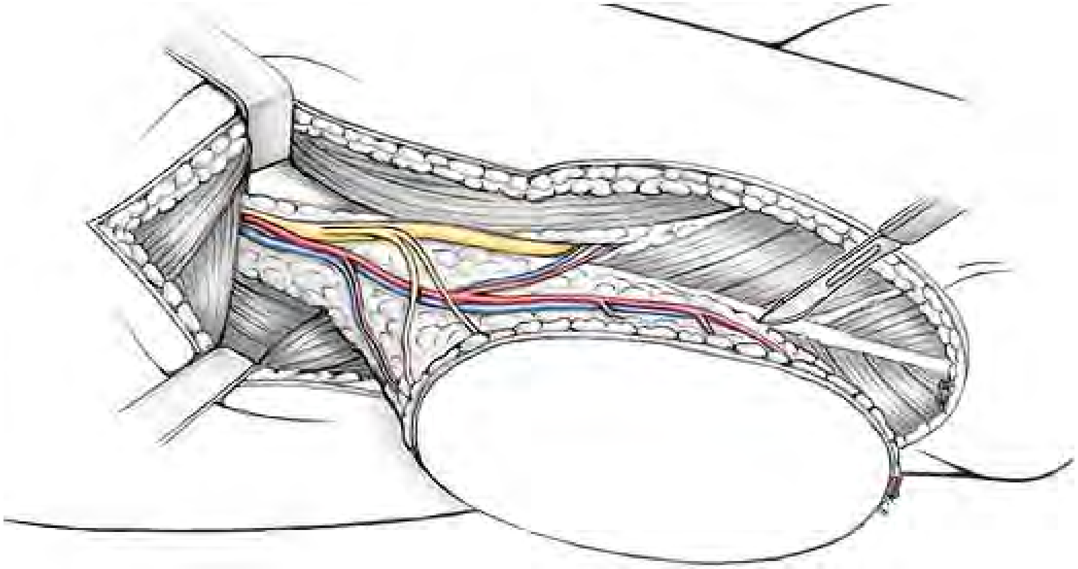

FIGURE 15-2 After the flap is isolated on the septum, the pedicle is followed proximally quite easily.

If the recipient site is prepared properly, the pedicle can be ligated with vessel clips or suture ligation. A suction drain is placed before a double-layered skin closure.

Donor Site Morbidity

Preoperatively, the upper lateral arm can be pinched to check if primary wound closure is possible after removal of the planned skin island. In flaps wider than 8 to 9 cm, primary wound closure is impossible; a skin graft must be applied, and a secondary scar correction or excision after 6 months might be necessary. In the literature, aesthetically unsatisfactory results are reported in about 27% of patients, as even the linear scar is quite visible or can show considerable hypertrophy or pigmentation changes or cause an hourglass-deformity of the upper arm. Early use of silicone sheeting, silicone gel appliance, or intralesional corticoid application should be considered. The sometimes unsightly scar is one of the reasons for restricting the use of the lateral arm flap in female patients.

Postoperative Management

At the donor site, early motion of the limb is encouraged, whereas shear forces to the scar should be avoided. Early scar therapy such as with silicone gel, sheet application, or compressive garments has proved to be successful. If an osteocutaneous flap was harvested, the patient should refrain from heavy lifting until 3 months postoperatively. Protective splints might be considered in noncompliant patients. In the case of a triceps tendon segment removal, full weight bearing (e.g., pushups) should not be allowed before 12 weeks postoperatively. If flap healing at the hand and wrist is undisturbed, physiotherapy is begun on the second postoperative day with constant monitoring of flap perfusion. If tendons were reconstructed, special rehabilitation protocols may then be initiated. Compressive gloves are useful after removal of the suture material. Flap thinning or secondary procedures should not be performed until 6 months postoperatively.

Complications

Specific complications of the lateral arm flap are rare and primarily due to poor planning. As it is harvested most often as a fasciocutaneous flap, no attempt should be made to close the fascia postoperatively, as this might lead to a muscular compartment syndrome. Even adapting the fascial edges at the proximal and distal donor site should be avoided, as this might lead to herniation of the triceps through the gap, which is less of a case if only the skin is closed in a two-layered fashion. Hematoma formation can also lead to considerable tension in that area and has to be evacuated surgically. In case of a fracture of the humerus after harvesting a bone graft, open or closed reposition and osteosynthesis should be encouraged to prevent prolonged immobilization of the elbow joint.

Complications with respect to microsurgical routines are no different than those of other free flaps. Extreme pin-cushioning and volume discrepancies due to secondary overall weight gain make flap debulking procedures necessary (e.g., aspiration lipectomy), whereas hair growth can be reduced with laser therapy.

Disturbances in forearm sensitivity (58.7%), lateral epicondylar pain (19.4%), and a hypersensitive donor site (17%) were quoted as long-term complications. To minimize this sensory forearm disturbance, the posterior cutaneous forearm nerve can be dissected carefully out of the flap. Impaired range of motion in the elbow is reported after dissecting a composite flap with a triceps tendon graft, but is a rare complication. Primary or secondary fracture of the humerus following elevation of an osteofasciocutaneous lateral arm flap can be avoided with careful technique (oscillating saw) and removal of only 25% of the lateral cortical circumference. If a neuroma of the posterior cutaneous forearm nerve develops despite division close to the main radial nerve, it should be reexplored, resected, or buried in muscle or bone.

Illustrative Cases

Case 1

An 18-year-old man injured his left wrist with a contaminated butcher knife. Tendons and neurovascular structures were found to be intact. Treatment is demonstrated in Figure 15-3 .

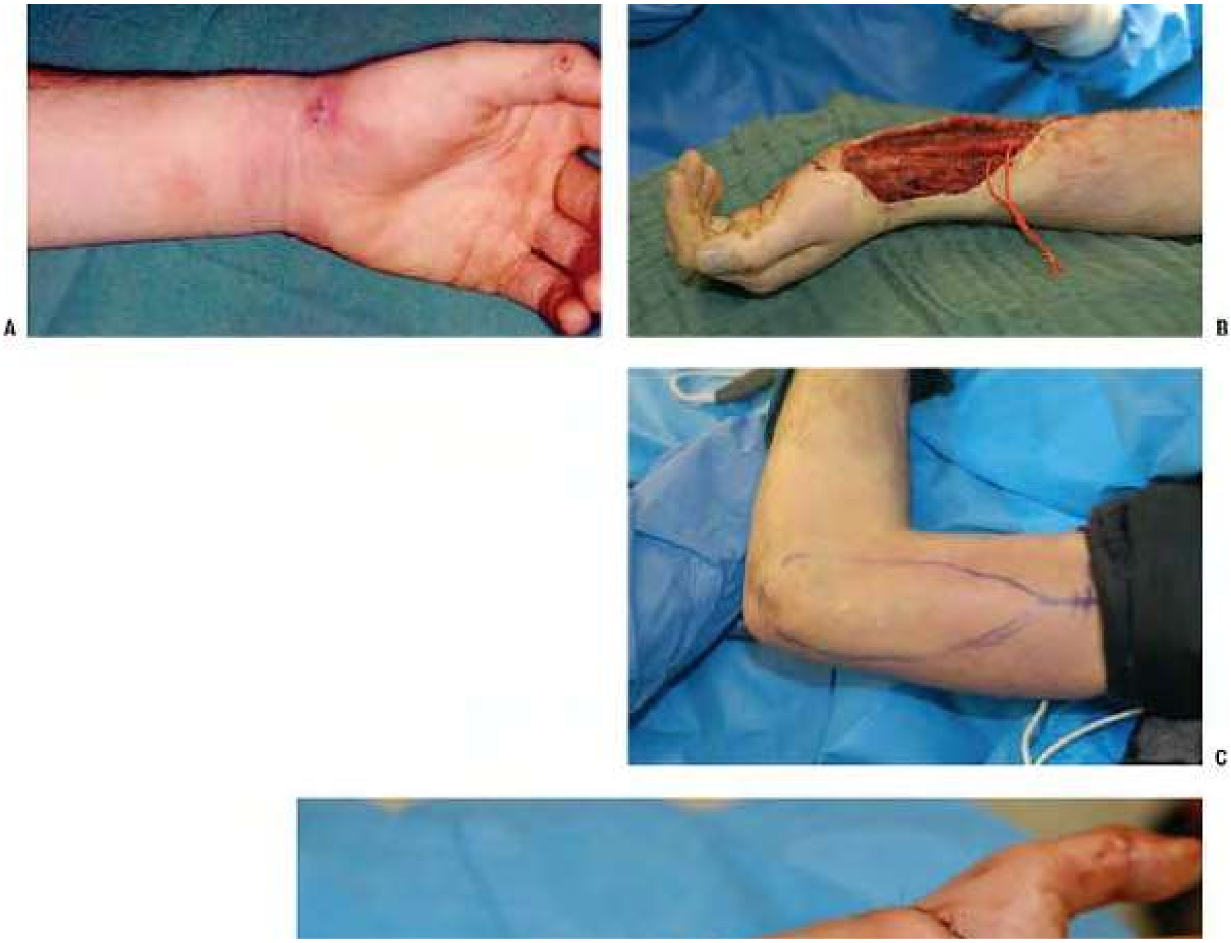

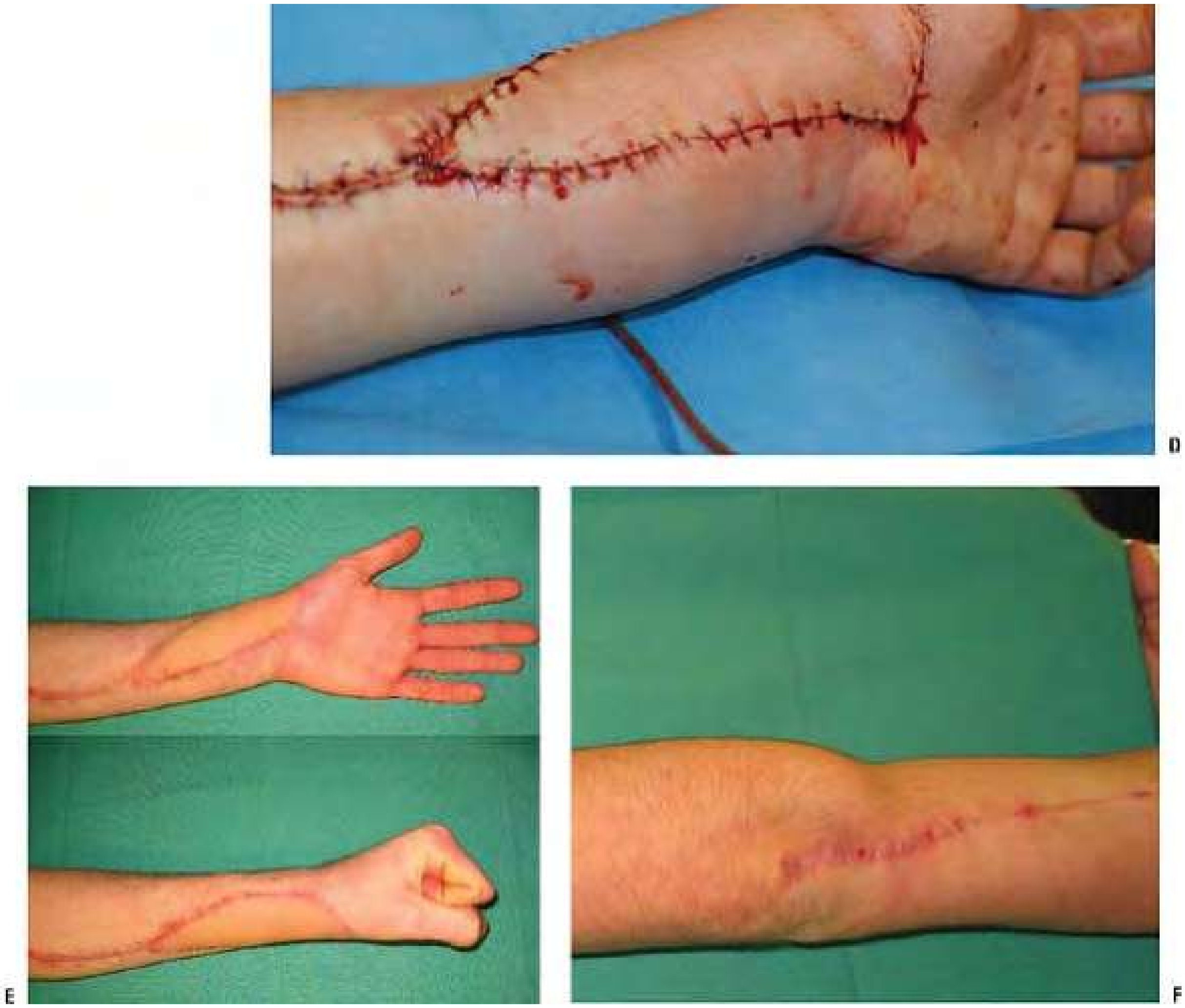

FIGURE 15-3 The wound was irrigated and sutured in a small regional hospital. B: Two days later the patient developed necrotizing fasciitis and was transferred to our hospital, where the patient underwent two serial debridements with the initiation of antibiotic therapy. The defect now had a size of 12 Ã— 6 cm. The flexor tendons and median nerve were exposed. C: A distal, thin lateral arm flap was outlined from the same extremity. D: The flap was connected end-to-side to the radial artery and end-to-end to the cephalic vein and sutured into the defect. E: One year postoperative with full range of motion. F: Donor site 1 year postoperative. Note the slightly widened scar around the lateral epicondyle.

Case 2

A 70-year-old woman mutilated her left wrist in a suicide attempt and was found 3 days later in her apartment. At the time of presentation, the patient had a large soft-tissue defect in addition to desiccation of the palmar structures. Treatment is demonstrated in Figure 15-4 .

Case 3

A combine accident in a 21-year-old man resulted in a rupture of the second extensor digitorum communis, extensor indicis proprius, extensor carpi radialis longus and brevis and extensor pollicis longus tendons, three main branches of the superficial radial nerve, the radial artery and the dorsal wrist capsule, and ligaments and dorsal corticalis of the carpal bones. The whole wound was extremely contaminated with organic debris. Treatment is demonstrated in Figure 15-5 .

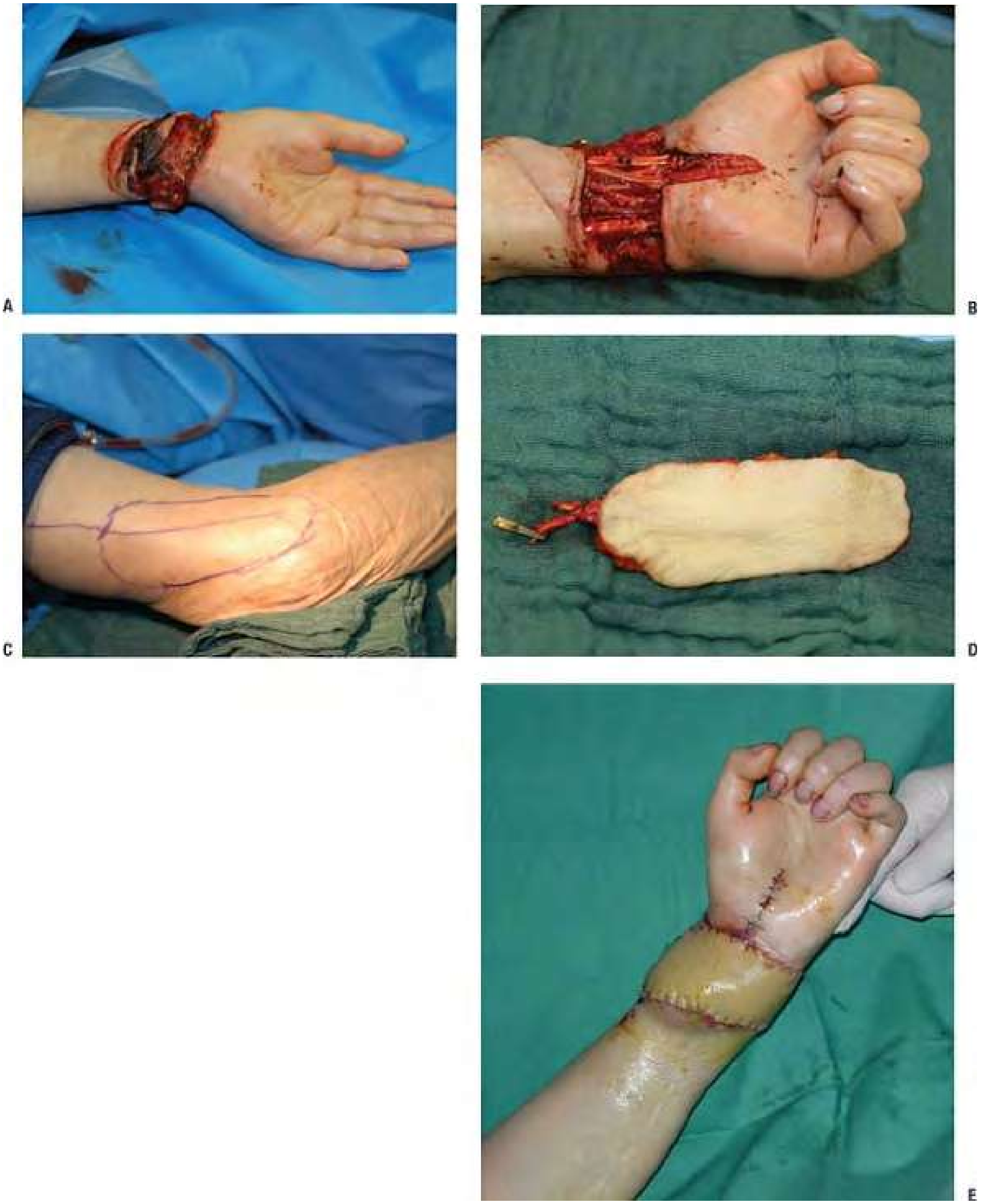

FIGURE 15-4 A: Intraoperative exploration showed traumatic division of the ulnar artery, median nerve, all superficial and deep flexor tendons, flexor pollicis longus, and flexor carpi radialis and ulnaris tendons. B: After a detailed debridement, all

functional structures were reconstructed except the superficial flexor tendons, which were resected. Three sural nerve grafts (each 4.5 cm) were interposed; the ulnar artery was reconstructed with a vein graft (6.5 cm). C: A fasciocutaneous, thin lateral arm flap was outlined on the ipsilateral distal upper arm. D: The harvested flap had a size of 8 Ã— 4 cm and was anastomosed end-to-end to the radial artery and the cephalic vein. E: Result 10 days after the operation. The patient started physiotherapy already on the second postoperative day.

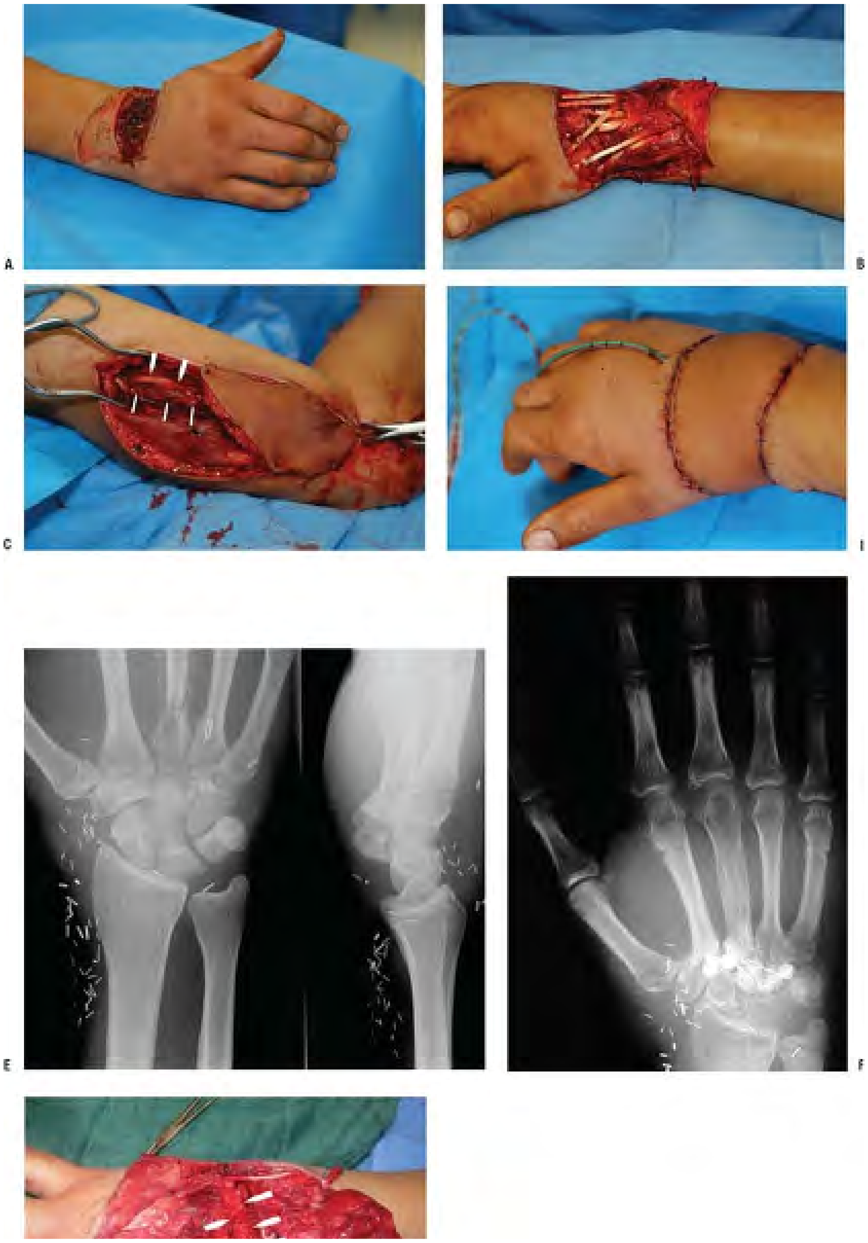

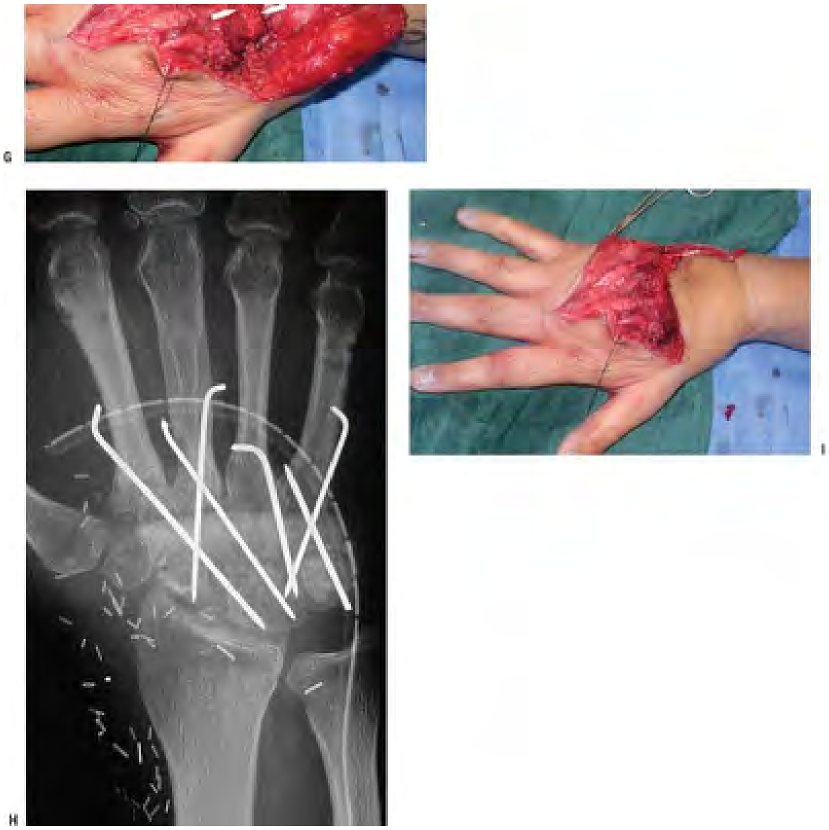

FIGURE 15-5 A: A thorough debridement and irrigation was performed as a first stage. B: After debridement, all tendons were sutured and the wound was closed with a free fasciocutaneous ipsilateral lateral arm flap. C: The pedicle is marked with thin white arrows; the radial nerve is indicated by thick white arrows. D: It was connected end to end to the radial artery and the cephalic vein, respectively. E: Unfortunately, a persisting wound infection in the carpus made additional debridements necessary. Antibiotic beads were set into the space and an external fixator was placed. F: Finally, the distal carpal row and the bases of the second and third metacarpal bones had to be removed. G: A non-vascularized iliac crest bone graft was fixed with K-wires for a carpometacarpal arthrodesis of the second to fifth ray (white arrows). H: Postoperative radiograph. I: Due to the infection-related elongated immobilization of the hand, the extensor tendons were severely impaired in their motion, even under the flap, and had to be partially resected and tenolysed. Silicon rods were placed under the flap for a secondary extensor tendon replacement. The flap remained viable and pliable throughout all procedures.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part III - Management of Soft Tissues Within the Hand and Wrist > Chapter 16 - Posterior Interosseous Artery Island Flap for Dorsal Hand Coverage

---

Chapter 16

Posterior Interosseous Artery Island Flap for Dorsal Hand Coverage

Eduardo A. Zancolli

The posterior interosseous flap is an island pedicled fasciocutaneous flap from the dorsal aspect of the forearm supplied by the cutaneous branches of the posterior interosseous artery. The posterior interosseous artery is located in the septum between the extensor carpi ulnaris and the extensor indicis propius. This flap has proven to be an excellent coverage option for the dorsum of the hand and the first interdigital web. The dorsal cutaneous area of the forearm supplied by the posterior interosseous artery was initially described by Manchot in 1889 (8) and later by Salmon in 1936 (12).

The vascular anatomy of the flap was described along with 20 initial cases (17 cases to cover the first web and three cases to cover the dorsum of the hand). The dorsal cutaneous area of the forearm irrigated by the posterior interosseous artery was initially described by Manchot in 1889 (8) and later by Salmon in 1936 (12).

This island flap was presented for the first time by the author at the VIth European Hand Surgery Course in Ãœmea (Sweden) (13). In this opportunity the vascular anatomy of the flap and an experience on 20 initial cases were presented (17 cases to cover the first web and three cases to cover the dorsum of the hand). Hand outs distributed during the course were published by the author in 1993 (18). The procedure was also presented in the same year in the XI Congress of the Argentine Society for Surgery of the Hand (awarded with the â€œJ. Goyena Prizeâ€•), in November 1985, Buenos Aires, Argentina (14). Other papers on the procedure were published by us in 1986 (15,16), 1988 (17) and 1993 (1,18). The vascular anatomy of the posterior interosseous island flap was presented by Penteado et al in 1986 (11), and other studies on the posterior interosseous flap have been published by Masquelet et al (9), Costa et al (3,4), Gilbert et al (7), Buchler et al (2), Dap et al (5), Mazzer et al (10), and Goubier et al (6). These studies have further supported the efficacy of the this flap.

Indications/Contraindications

The posterior interosseous artery island flap is capable of transporting the skin of the distal two-thirds of the forearm. The flap's pedicle length determines its coverage possibilities.

It is particularly useful for covering extensive defects on the back of the hand as far distally as the dorsal aspect of the proximal phalanges of the fingers and for obtaining a complete coverage of the first web after the release of severe adduction contractures of the thumb. Buchler et al (2) noted that the flap could reach the dorsum of the proximal interphalangeal joints. It can be indicated in different types of pathologies such as burn sequelae, wounds on the dorsum of the hand, tumors, and congenital deformities.

Contraindications for flap use include any injury to the posterior interosseous artery or deep laceration or crush wounds to the posterior forearm in the area of the flap pedicle. The flap is not recommended in patients with diabetes, in whom small vessels may be diseased. The flap has been used successfully in smokers, but smoking certainly carries a higher risk of flap complications.

Preoperative Planning

Successful flap harvest is predicated on a thorough knowledge of the flap anatomy. Normally the anterior and posterior interosseous arteries in their course through the forearm are united through two main anastomoses, one proximal, at the level of the distal border of the supinator muscle, and one distal at the most distal part of the interosseous space (2 cm proximally of the distal radioulnar joint) (Fig. 16-1). The island posterior interosseous island flap is supplied by the reverse arterial blood flow through the distal anastomosis once the proximal anastomosis has been ligated. Venous drainage is produced by the venae comitantes of the posterior interosseous artery.

The angiosomal territory of the posterior interosseous artery was studied in a series of 80 cadaveric forearms in 1993 (1). Ink injections performed through catheter placed in the distal part of the anterior interosseous artery stained the distal two-thirds of the posterior forearm skin through the reverse flow through the distal anastomosis. The proximal third of the forearm skin remained unstained even when larger amounts of ink were injected (see Fig. 16-5). Ink injections through the catheter placed in the proximal part of the posterior interosseous artery stained the proximal two-thirds of the posterior forearm through the direct flow through the proximal cutaneous branch (1).

Other anatomic conclusions from this study can be summarized as follows: the posterior interosseous artery usually branches from the common interosseous artery in the proximal third of the forearm in 90% of cases, it can however be a direct branch of the ulnar artery in 10% of cases (1). It pierces the interosseous membrane about 6 cm distal to the lateral epicondyle of the humerus (14,15,16,17,18) forming its origin or proximal anastomosis between the anterior and posterior interosseous arteries. It enters into the posterior compartment of the forearm below the distal edge of the supinator muscle, located at the union of the proximal with the distal two-thirds of the dorsal aspect of the forearm (60 mm distal to the lateral epicondyle). At its entrance into the posterior compartment of the forearm, the posterior interosseous artery gives off branches to the recurrent interosseous artery that anastomoses at the elbow with a descending branch from the superior profunda, with the posterior ulnar recurrent and the anastomotica magna. In its ascending course the artery runs between the lateral condyle and the olecranon (Figs. 16-1 and 16-2).

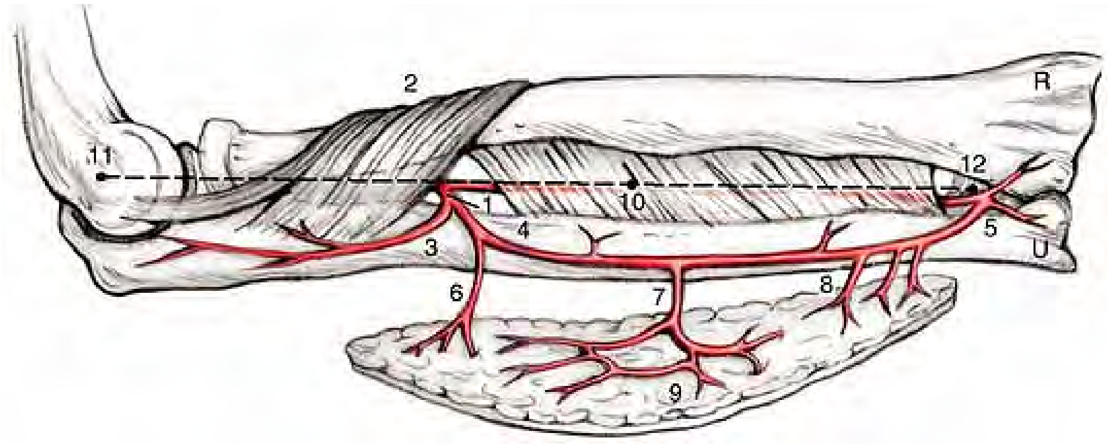

FIGURE 16-1 Anatomy of the posterior interosseous artery. (*1*) Proximal end of the posterior interosseous artery at the distal border of the supinator muscle. (*2*) The artery is emerging as a division of the common interosseous artery. (*3*) Recurrent interosseous artery. (*4*) Posterior interosseous artery following the longitudinal line X-Xâ€™, between the lateral epycondile (*11*) and the distal radio-ulnar joint (*12*). (*5*) Distal anastomosis between both interosseous arteries (2 cm proximal to the distal radio-ulnar joint). Cutaneous branches from the posterior intererosseous artery that irrigte the two distal thirds of the posterior skin of the forearm. Proximal (*9*), middle (*7*), and distal (*8*) branches. (*10*) Middle of the forearm where the middle cutaneous branch is located 1â€“2 cm distal to this point.

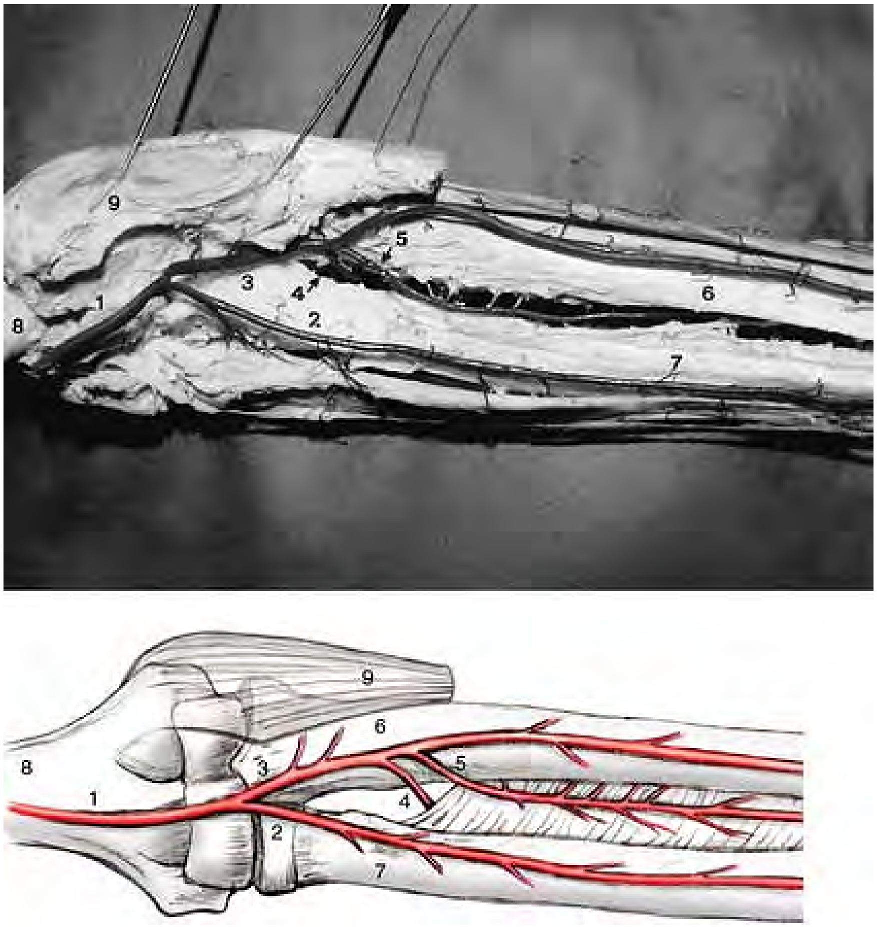

FIGURE 16-2 Cadaver specimen dissected to show the proximal origin of the posterior interosseous artery. Volar view of the proximal part of the forearm and the elbow. (*1*) Brachial artery. (*2*) Radial artery. (*3*) Ulnar artery showing the separated origin of the posterior interosseous artery (*4*) and the anterior interosseous artery (*5*). (*6*) Ulna. (*7*) Radius. (*8*) Humerus. (*9*) Medial epicondyle muscles.

From the proximal anastomosis, the posterior interosseous artery follows a line between the lateral epicondyle and the distal radio-ulnar joint. In its course it can be divided into three parts: proximal, middle, and distal (17).

In the proximal part, the artery runs deep to the abductor pollicis longus muscle and is covered by the extensor digiti minimi and the extensor carpi ulnaris muscles and in close relation with the posterior interosseous nerve and a large venous plexus. In the middle part, it becomes superficial, just beneath the superficial antebrachial fascia in the middle third of the forearm, running between the extensor digiti minimi and the extensor carpi ulnaris muscles. In this part, the artery reduces its diameter in 90% of the cases and is usually found to be between 0.3 and 0.6 mm in diameter (1) (see Fig. 16-4).

In the distal part, the posterior interosseous artery joins with the distal end of the anterior interosseous artery, forming the distal anastomosis between the two vessels (13,15). This anastomosis is located 2 cm proximal to the distal radio-ulnar joint and very close to the periosteum of the ulnar metaphysis (Figs. 16-1 and 16-3). This anastomosis has been present in all our cadaver specimens and operative cases and represents the point of rotation of the vascular pedicle of the posterior interosseous flap. At the distal anastomosis the posterior interosseous artery enlarges between 0.9 and 1.1 mm (1).

During its course, the posterior interosseous artery gives four to six cutaneous branches which run through the septum between the extensor digiti minimi and the extensor carpi ulnaris muscles to reach the dorsal skin of the forearm, of these the principal cutaneous branches are the proximal and middle branches.

The proximal cutaneous branch emerges from the proximal part of the artery, forming a large branch that irrigates the skin over the upper third of the dorsal aspect of the forearm. It is consistently present but it has a variable origin. Of our 80 specimens, it was found as a branch of the recurrent interosseous artery in 28, a branch of the common interosseous artery in 22, and as a branch of the posterior interosseous artery in 30. This variation indicates that the proximal cutaneous branch cannot be preserved to supply the flap in all cases.

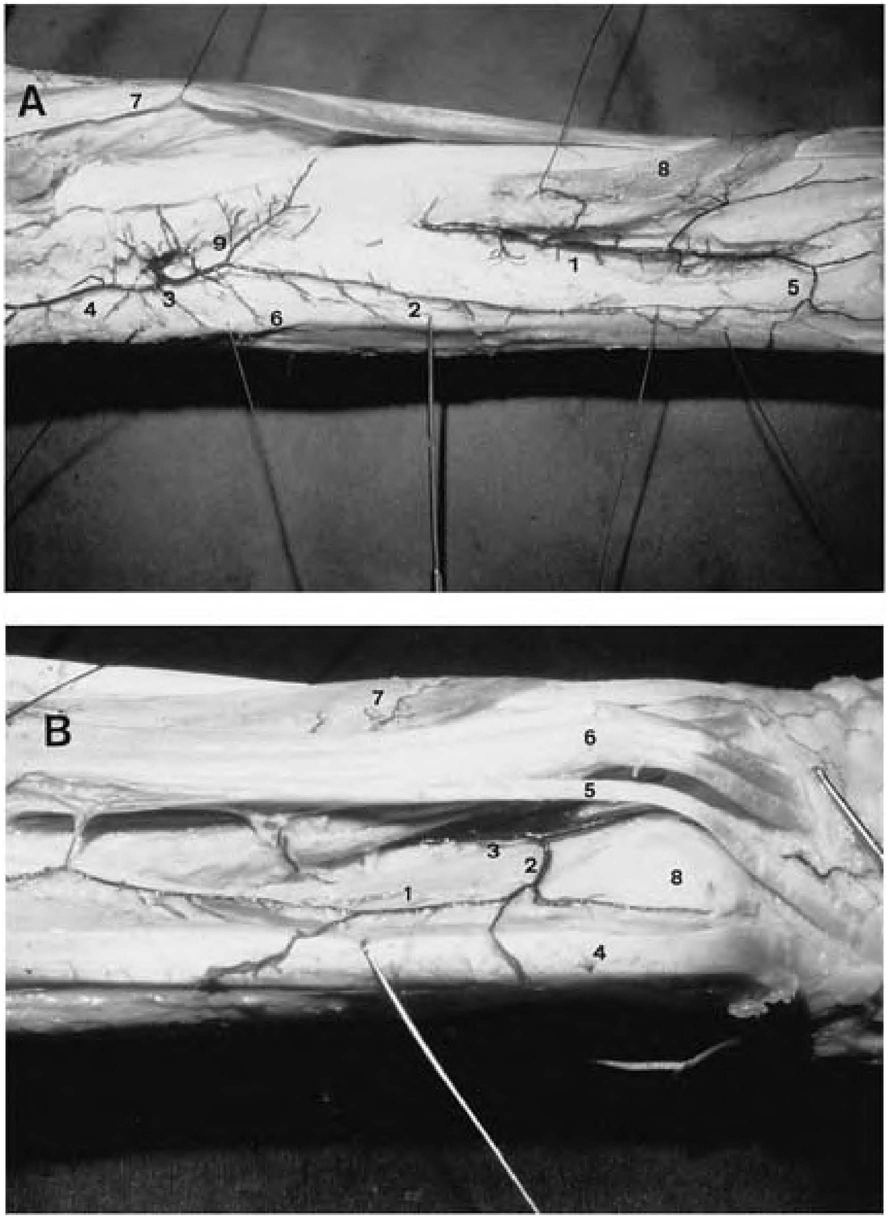

FIGURE 16-3 A: Cadaver specimen injected with latex (right forearm). Dissection of both interosseous arteries. (*1*) Anterior interosseous artery. (*2*) Posterior interosseous artery piercing the interosseous membrane (*3*) and showing the emergence of the recurrent interosseous artery (*4*). The cutaneous branches have been eliminated. (*7*) Distal anastomosis between both interosseous arteries located very close to the periostum of the distal metaphysis of the ulna. (*8*) Extensor carpi ulnaris. (*9*) Extensor carpi radialis brevis muscle. (*10*) Extensor pollicis brevis and abductor pollicis longus

muscles. (*11*) Large branch of the posterior interosseous artery to irrigate common extensor muscle. B: Posterior interosseous artery showing its proximal anastomosis (*2*) with the anterior interosseous artery (*3*). Right forearm. Several branches to the dorsal muscles of the forearm are seen. (*4*) Extensor carpi ulnaris tendon. (*5*) Extensor digiti quinti tendon. (*6*) Extensor communis digitorum tendons. (*7*) Extensor pollicis brevis and abductor pollicis longus muscles. (*8*) Distal head of the ulna.

The middle cutaneous branch originates in the middle third of the forearm where the posterior interosseous artery becomes superficial. It represents a large branch located at 1 to 2 cm distal to the middle of the forearm. It was consistently present in our cadaveric investigations and clinical cases.

No cutaneous branches were found between the described proximal and middle branches. In our cadaveric observations, one or two large interconnecting venous perforators were consistently found running together with the middle cutaneous arterial branch, uniting the superficial to the deep venous system (the venae comitants of the posterior interosseous artery) (1).

At the distal third of the posterior forearm, the posterior interosseous artery may also contribute six to eight cutaneous branches to the skin of variable diameter (1).

The posterior interosseous artery also gives several branches to the muscles of the posterior compartment of the forearm (see Fig. 16-10E) and contributes some blood supply to the periosteum of the radius and ulna.

In only two cases the continuity of the posterior interosseous artery was absent at the level of the middle of the forearm: one in an anatomic specimen and one in a clinical case. This uncommon anatomic finding is in accordance with other reports in the medical literature. Thus, Penteado et al. (11) found absence of the posterior interosseous artery in the forearm in 4 of 70 specimens. In their same series, the distal anastomosis was absent in only one case. Buchler and Frey (2) found the posterior interosseous artery missing at the middle of the forearm in 2 of 36 cases. Despite this, angiograms and Doppler examination are not part of our standard preoperative protocol.

Surgery

The surgical procedure is performed under general anesthesia. Pneumatic tourniquet is employed. Exsanguination is obtained by simple elevation of the upper limb for 30 seconds. Loop magnification (4Ã—3) permits a better identification of the vascular pedicles and their cutaneous branches. The patient is placed in the supine position with the forearm placed in pronation over the inferior-lateral thoracic wall with the elbow flexed in 90 degrees (Fig. 16-6). This position facilitates flap preparation and raising. The tourniquet is insufflated (250 mmHg) after the flap has been designed. The operation consists into two main steps: flap design and flap elevation.

Flap Design

Flap design (Fig. 16-7A) is strictly related to the shape and size of the recipient area. A longitudinal line (X-Xâ€™) is drawn from the lateral humeral epicondyle to the distal radioulnar joint. This line represents the course of the posterior interosseous artery in the posterior forearm. A point *A* is marked 1 cm distal to the middle of line. This point corresponds to the emergence of the middle cutaneous branch of the posterior interosseous artery and to the medial interconnecting venous perforator. The flap is planned to be irrigated by this middle cutaneous pedicle, so the point A must be always included in the flap. A second point *B* is marked at the distal end of the line X-Xâ€™, 2 cm proximal to the distal radio-ulnar joint. This point corresponds to the distal anastomosis between both interosseous arteries and is the point of rotation of the pedicle flap.

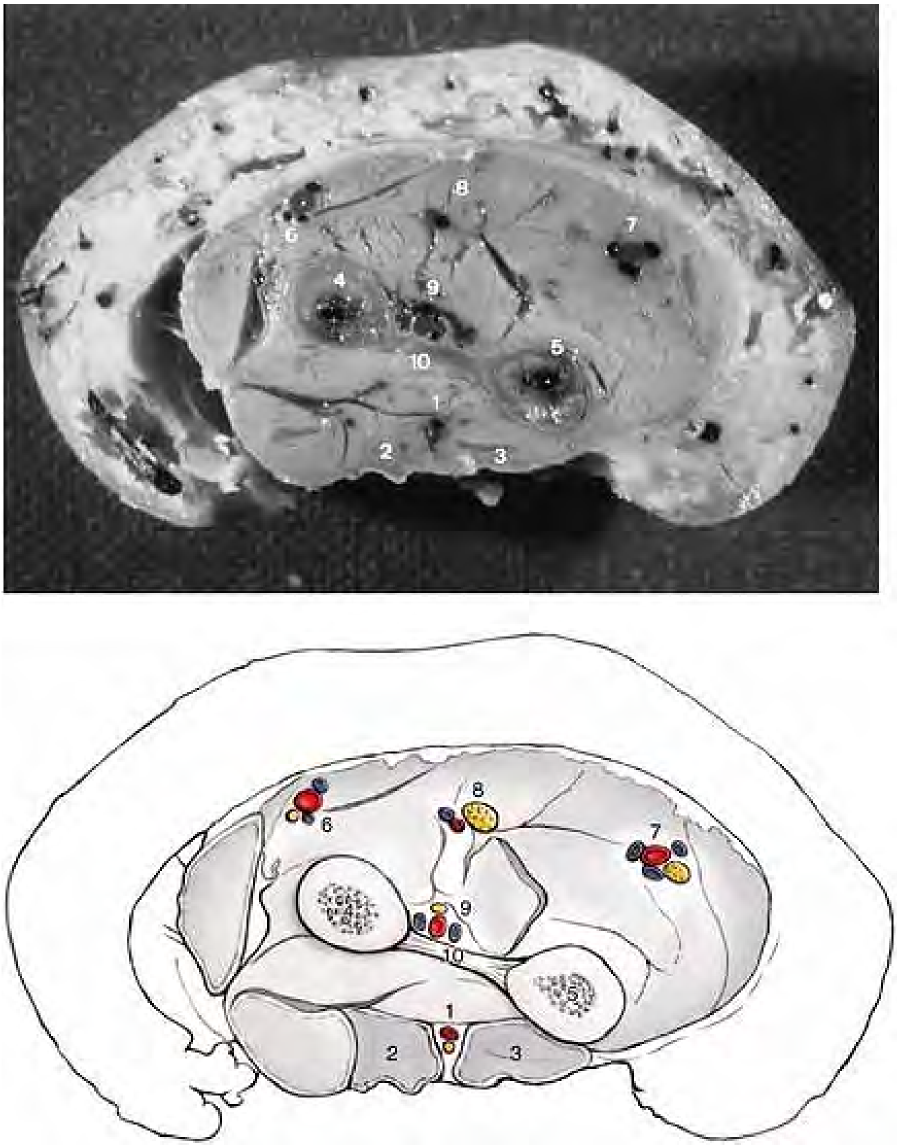

FIGURE 16-4 Transverse section of the forearm at its middle third. (*1*) Posterior interosseous arteryâ€"in its proximal courseâ€"running deep between the extensor digiti quinti (*2*) and the extensor carpi ulnaris (*3*) muscles. (*4*) Radius. (*5*) Ulna. Radial artery (*6*) and ulnar artery (*7*) both with their venae. Median nerve and median nerve artery (*8*). Anterior interosseous artery and its venae (*9*). Interosseous membrane (*10*).

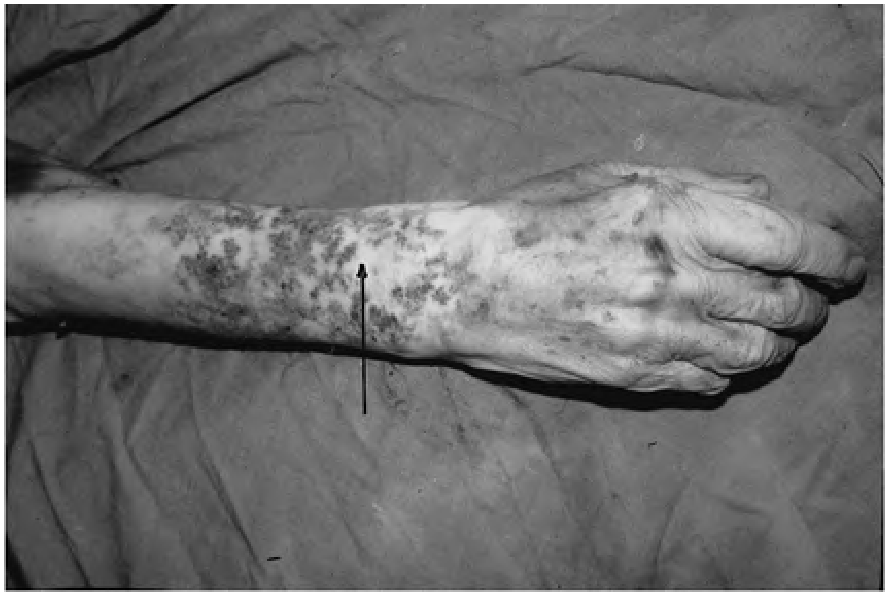

FIGURE 16-5 Fresh cadaver injected with 20 mL of black ink through a catheter placed in the distal anterior interosseous artery. The distal and middle thirds of the dorsal forearm are stained.

The proximal limit of the flap can be safely placed to a point 6 cm distal to the lateral epicondyle (point *C*). The flap may be extended distally up to the wrist joint. In this case distal cutaneous branches can be included. It should be remarked that the medial vascular cutaneous pedicle is capable of irrigating a flap as large as the complete width of the posterior forearm. The center of the flap always corresponds to the line X-Xâ€™.

It was observed in our clinical cases that there is no venous insufficiency in the posterior interosseous reverse forearm flap if it is raised with the large middle venous interconnecting perforator. It ensures sufficient drainage of the subcutaneous tissue and skin into the venae comitants of the posterior interosseous artery (1).

A flap of approximately 3 to 4 cm in width will allow for direct closure of the donor area of the forearm leaving an inconspicuous scar (14,15,16) (Figs. 16-8 and 16-10). Wider flaps will need free skin graft to cover the dorsum of the forearm and may lead to a poorer aesthetic result (Fig. 16-9).

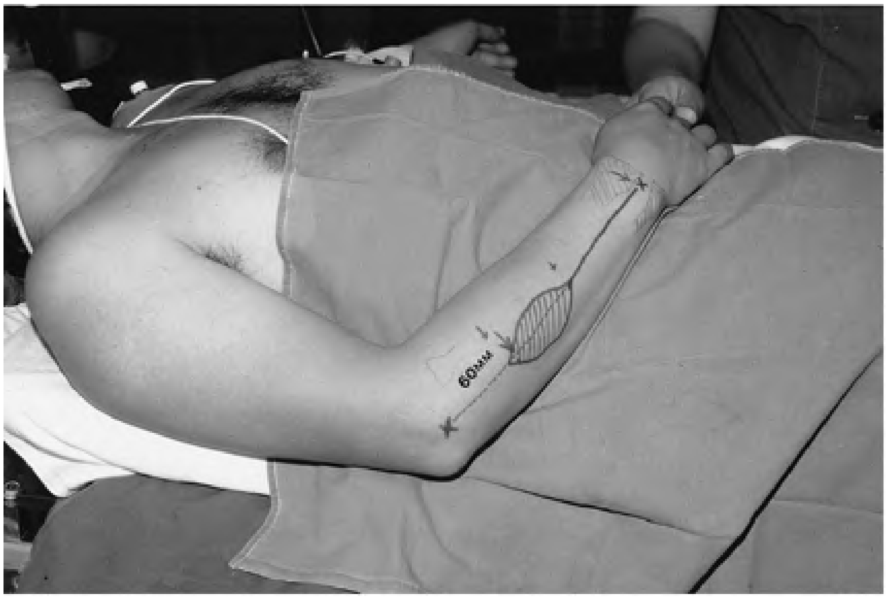

FIGURE 16-6 Position of the upper limb during surgery. In this patient a posterior interosseous artery island flap 4 cm wide has been designed to cover the first interdigital web. The flap is located at the level of the middle cutaneous branch of the posterior interosseous artery. Details on the flap design are completed in Figure 16-7.

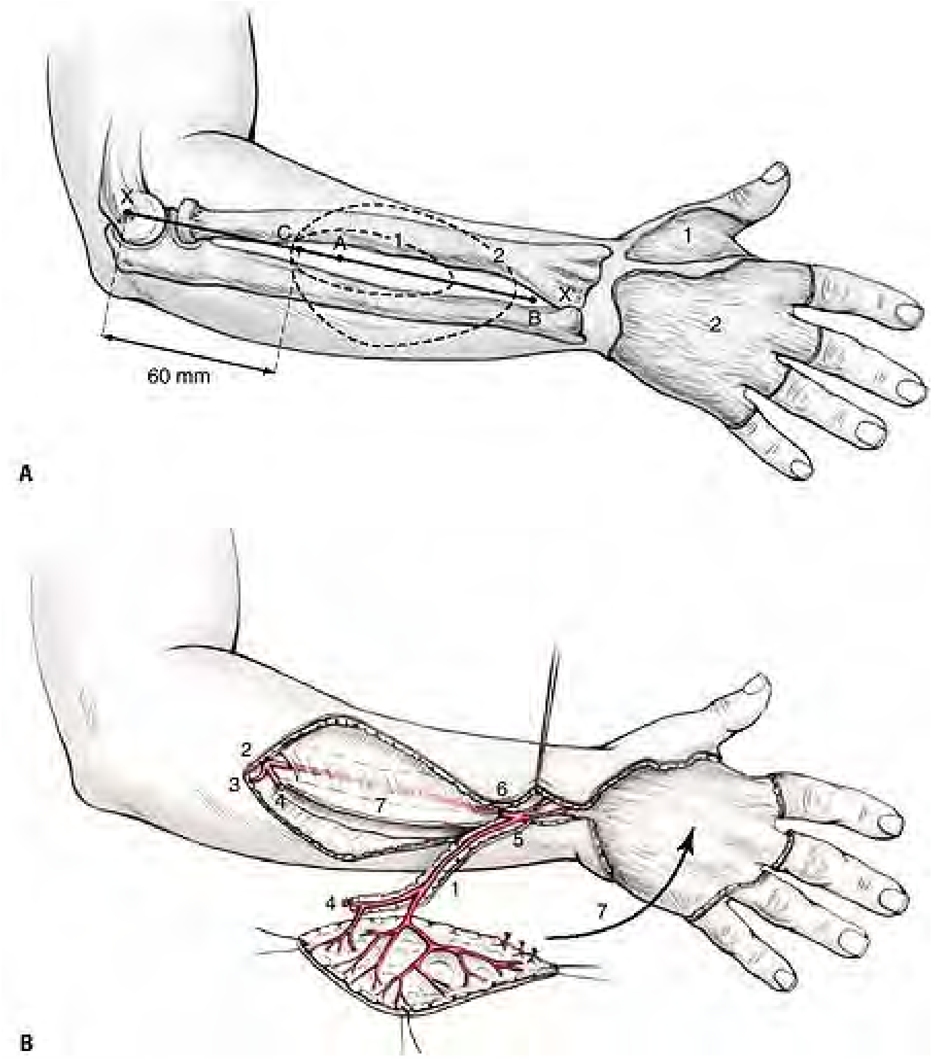

FIGURE 16-7 A: Design of the flap. The location of the island flap is in the line X-Xâ€™, between the lateral epicondyle (X) and the distal radio-ulnar joint (Xâ€™). Three points are marked in the skin. Point A corresponds to the location of the middle cutaneous branch of the posterior interosseous artery. Point B is located 2 cm proximal to the distal radio-ulnar point and corresponds with the location of the distal anastomosis; and point C is located at the distal edge of the supinator muscle where the posterior interosseous artery pierces the interosseous membrane. This point is 60 mm distal to the lateral epicondyle. The width, length, and shape of the island flap is in accordance with the defect to be covered: 1 is representing the shape, size, and location of a flap to cover the first interdigital web, and 2 represents a

flap to cover the dorsal aspect of the hand. B: Raising of the island flap. The flap has been raised with the posterior interosseous artery and its venae comitantes (*1*). In this case the middle and the proximal cutaneous branches are irrigating the flap. The intermuscular septum and a strip of fascia are raised with the posterior interosseous artery. (*2*) Common interosseous artery. (*3*) Recurrent interosseous artery. (*4*) The posterior interosseous artery has been ligated at its emergence (proximal anastomosis between both interosseous arteries). (*5*) Distal anastomosis is preserved (point of rotation of the flap). (*6*) Anterior interosseous artery. (*7*) In this case the flap is rotated to cover the dorsum of the hand. The donor area is covered with a free skin graft (see Figs. 16-8 and 16-9).

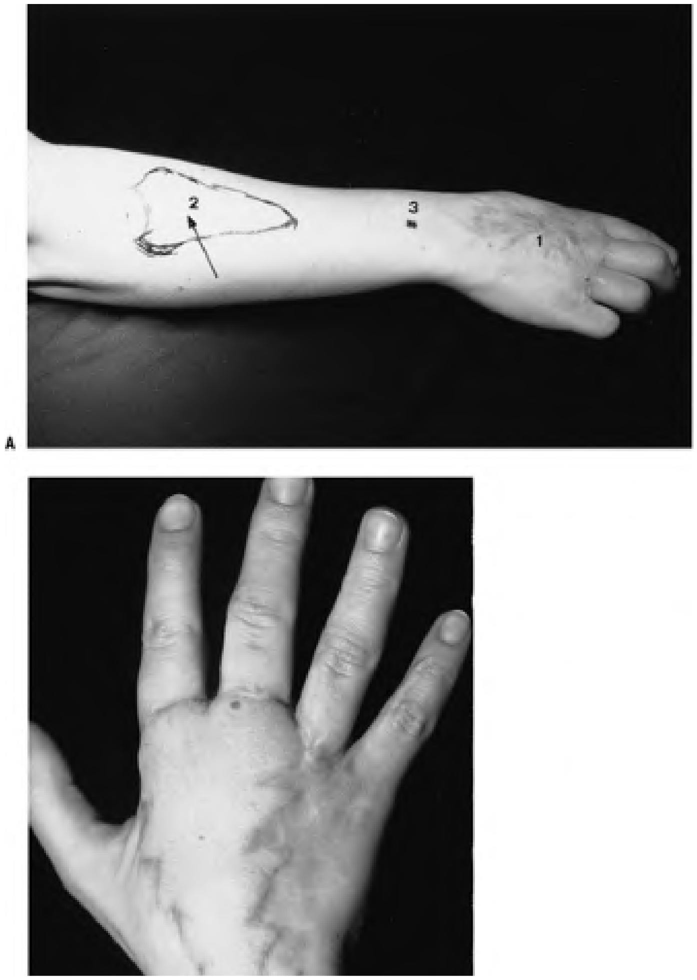

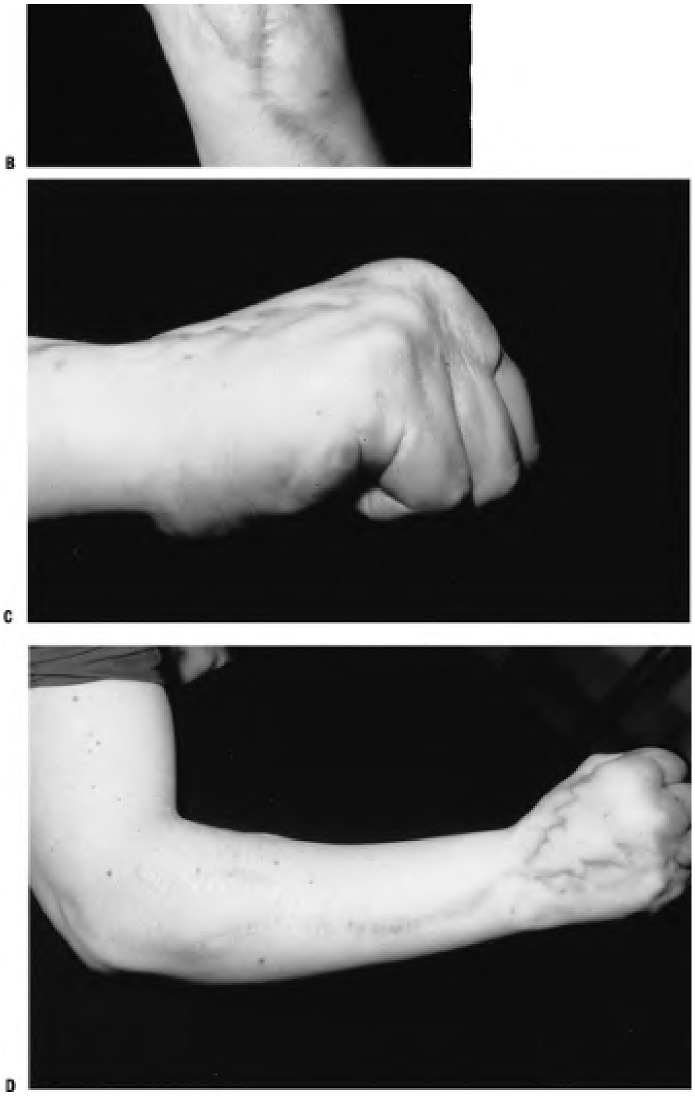

FIGURE 16-8 A: Sequela of an injury in a 16-year-old girl on the dorsum of the hand (*1*) with laceration of the long extensor tendons of the index and middle fingers and stiffness of their metacarpophalangeal joints. A posterior interosseous artery island flap was designed (*2*) with the shape of the area to be covered on the hand. Middle and proximal pedicles were preserved to irrigate the flap. Excision of the injured extensor tendons and metacarpophalangeal capsulectomies were indicated simultaneously with the island flap. The fingers were immobilized in flexion during the postoperative period. Location of the distal anastomosis (*3*). B: Finger extension obtained after a secondary surgical stage with Z plasties at the borders of the flap and tendon transfer from the flexor superficialis tendon of the ring finger, divided

into two strips, to the base of the middle and index fingers proximal phalanges. The interdigital webs were reconstructed with the flap. C: Digital flexion obtained. D: The donor area of the forearm was initially closed.

Flap Elevation

Flap elevation (see Fig. 16-7B) begins at its radial side. The dissection is carried out between the subcutaneous tissue and the superficial fascia. The incision is continued in the direction of the wrist, and the proximal and medial cutaneous branches are easily identified. The interconnecting venous perforator that accompanies the medial cutaneous branch is clearly visualized. The proximal cutaneous branch can be coagulated. Our clinical cases have shown that the flap can be safely raised with the middle cutaneous branch. The posterior interosseous artery is sectioned at its proximal origin proximal to the medial cutaneous branch. The posterior interosseous artery with its venae comitants are raised with the intermuscular septum located between the extensor carpi ulnaris and the extensor indicis propius and a very thin strip of fascia covering the muscles. Elevation is completed after sectioning the skin at the ulnar side of the flap. Finally, the flap is turned through 180 degrees to cover the recipient area: to the dorsum of the hand, or to a released first web space, or to both. Now the pneumatic tourniquet is released to observe flap circulation.

At the end of the procedure, the hand is immobilized in a neutral position of the wrist or with some dorsiflexion. In the case of coverage of the back of the hand, where the extensor tendons have been

excised and a metacarpophalangeal capsulotomy has been performed, the fingers are immobilized in flexion to prevent the recurrence of metacarpophalangeal stiffness in extension. In these cases tendon grafts can be indicated during a second surgical stage (see Figs. 16-8 and 16-9).

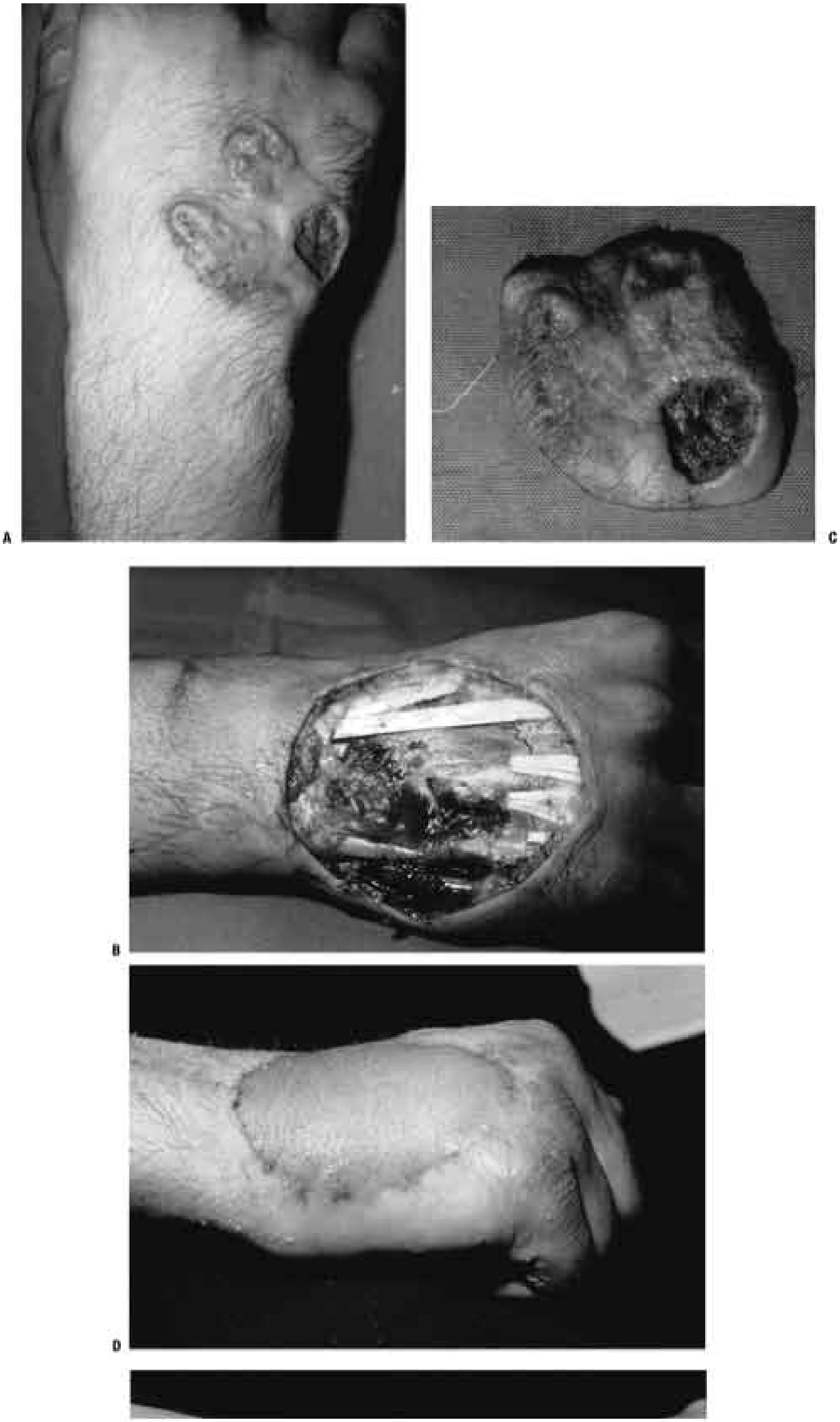
A
C
B
D

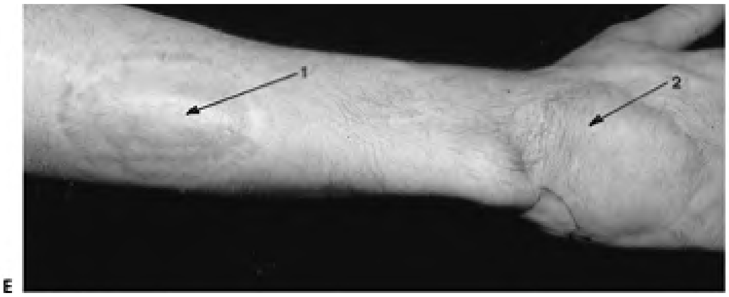

FIGURE 16-9 A: An epitheloid sarcoma with invasion of skin on the dorsal aspect of the forearm in a 20-year-old man. B: An ample excision of the affected skin, including the extensor tendons of the two last fingers, was indicated. C: Piece of the lesion. D: Immediate postoperative period after a posterior interosseous artery island flap. In a secondary surgical stage, the extensor tendons were reconstructed. E: Result after 2 years following the initial operation showing the free skin graft in the donor area (*1*) and the posterior interosseous artery island flap in the recipient area (*2*). Complete function of the fingers was obtained.

Postoperative Management

The hand and wrist are immobilized in a long plaster splint for 2 weeks following the surgery. A large window is left in the dressing to monitor the flap for signs of ischemia or venous congestion. No blood thinning agents are used during the postoperative period. Following removal of the splint, the sutures are removed and the patient may begin postoperative therapy.

Results

Our series of 80 cases (47 males and 33 females) was published in 1993 (1). In this series, 22 flaps were indicated for the coverage of skin defects in the dorsum of the hand, 37 for the reconstruction of the first interdigital web, 5 to cover the dorsum of the hand and first web space, 15 for the volar wrist, and 4 for the palm of the hand. In three cases extensor tendon grafts to extend the fingers were indicated in a second surgical stage. Seventy-six flaps were successful and four were lost. Two of these were clearly due to twisting of the pedicle. The flap provides a large amount of skin for hand reconstruction without interfering with lymphatic drainage, venous drainage, or the integrity of the principal vessels of the volar side of the forearm. It furnishes good-quality skin to permit secondary surgical reconstructions of tendons or the skeleton of the hand.

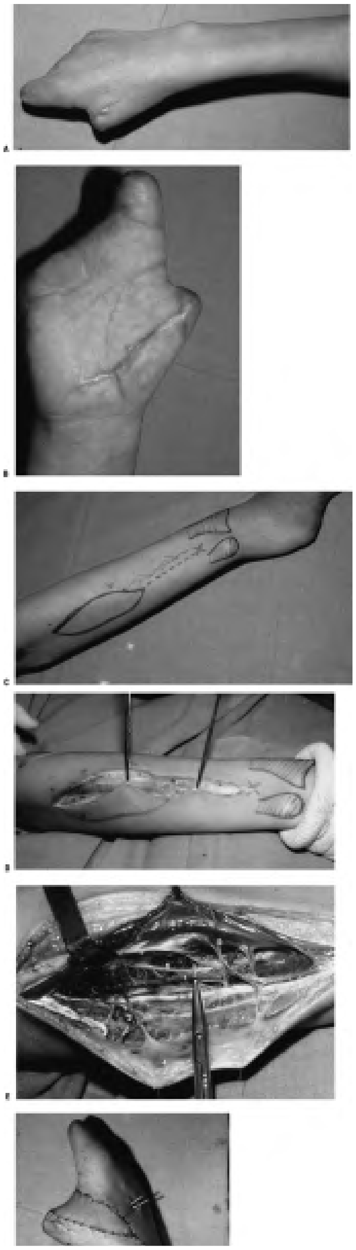

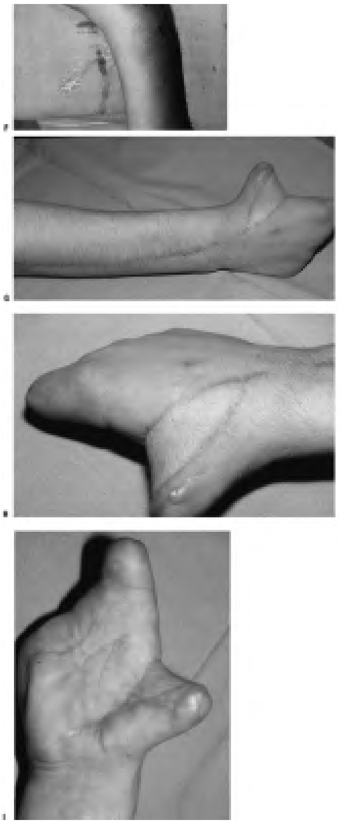

FIGURE 16-10 A,B: Sequela of a trauma with amputation of the digits and adduction contracture of the first intermetacarpal space in a 18-year-old man. C: Design of the posterior interosseous artery island flap of 3 cm width. The middle cutaneous branch was preserved to nourish the island flap. D: The initial skin incision is shown. The radial side of the flap is initially elevated from its radial side. E: Anatomy of the dissected posterior interosseous artery (*1*) with its middle (*2*) and distal (*3*) cutaneous branches to the dorsal skin of the forearm (*4*). Several arterial branches to the dorsal muscles of the forearm (*5*) and the posterior interosseous nerve (*6*) are shown. The extensor carpi ulnaris muscle (*7*) and the extensor muscles to the fingers (*8*) are separated. The hand is at the right of the picture. F: A few days after the release of the first intermetacarpal space and the transport of the island flap. Two KW are maintaining the separation of the first two metacarpals. G: Direct closure of the donor area. H,I: Result obtained.

Complications

As mentioned previously, flap loss is the most serious complication; this is most commonly due to inadvertent twisting of the pedicle or injury to the veins during dissection resulting in flap congestion. Injury to the radial nerve has also been reported with temporary paralysis to the extensor carpi ulnaris and extensor digiti minimi muscles following flap harvest. Such occurrences are rare.

References

1. Angrigiani C, Grilli D, Dominikow D, et al. Posterior interosseous reverse forearm flap: experience with 80 consecutive cases. *Plast Reconst Surg.* 1993;92(2):285â€“293.

2. Buchler U, Frey HP. Retrograde posterior interosseous flap. *J Hand Surg.* 1991;16A(2):283â€“292.

3. Costa H, Soutar DS. The distally based island posterior interosseous flap. *Br J Plast Surg.* 1988;41:221â€“227.

4. Costa H, Smith R, McGrouther DA. Thumb reconstruction by the posterior interosseous flap. *Br J Plast Surg.* 1988;41:228â€“233.

5. Dap F, Dantel G, Voche P, et al. The posterior interosseous flap in primary repair of hand injuries. *J Hand Surg.* 1993;18-B(4):437â€“445.

6. Goubier JN, RomaÃ±a C, Masquelet AC. Le Lambeaux interosseoux posterieur chez lâ€™enfant. *Chirurgie de la Main.* 2002;21:102â€“106.

7. Gilbert A, Masquelet AC, Hentz VR. Pedicle flaps of the upper limb. In: Dunitz M, ed. *Vascular anatomy: surgical technique and current indications.* 1992.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 17

Fillet Flaps in Cases of Mutilating Trauma

Bradley Medling
Michael W. Neumeister

Fillet flaps can be a very important reconstructive option in many cases of hand and lower limb trauma. Fillet flaps are, by definition, axial flaps that may provide skin, muscle, fascia, and bone. They can be harvested either as a pedicled flap or as a free flap. Careful a planning and execution of the flaps can allow for reconstruction of local defects with a flap of a similar tissue type. This technique takes advantage of spare parts that would otherwise be discarded, avoiding associated morbidity of a separate flap harvest.

Fillet flaps are most commonly employed after mutilating trauma to the hand. The concept of spare parts surgery should be broadly applied to include any tissue that would otherwise be discarded; these flaps may include bone, nerve, tendon, joints, or the whole digit. Segments of the injured part can be harvested with their blood supply intact, and thereby be used as a vascularized tissue graft or flap.

Frequently, a traumatized extremity is left without soft tissue coverage, leaving exposed vital structures such as bone, joint, tendon, or neurovascular bundles. Vascularized coverage can be salvaged from attached parts of the mangled hand that are otherwise rendered dysfunctional or from amputated parts where a flap is designed over a preserved vascular bundle.

The key component to the safe harvest of a fillet flap is an understanding of the vascular supply to tissue. Since the entire upper extremity and lower extremity can be used in a fillet fashion, the blood supply to the arm and leg must be understood. In the upper extremity, the axillary artery terminates at the lateral border of the pectoralis minor muscle becoming a brachial artery to a point just distal to the antecubital fossa. The brachial artery runs in the medial upper arm in relative proximity to the ulnar and median nerve. At the level of the elbow, the brachial artery is lateral to the median nerve. The brachial artery bifurcates into the radial and ulnar arteries, respectively, under the superficial flexor muscles of the forearm. The radial artery travels under the flexor carpi radialis muscle exiting more superficial in the forearm just lateral to the tendon of the flexor carpi radialis. The ulnar artery travels with the ulnar nerve under the muscle belly of the flexor carpi ulnaris becoming more superficial in the distal forearm. The ulnar artery travels through Guyon's canal and bifurcates into a superficial and deep branch, each of which communicates with the radial artery component of the superficial and deep palmar arches. The radial artery bifurcates at the wrist into a superficial palmar and deep dorsal branch. The superficial palmar branch courses around the base of the thenar eminence, and the larger deep branch travels under the abductor pollicis longus, extensor pollicis brevis, and the extensor pollicis longus, through the anatomical snuffbox. It then courses between the first and second metacarpals, through the adductor space to form the radial component of the superficial and deep palmar arches. The common digital vessels arise from the superficial palmar arch and travel to the level of the metacarpophalangeal joint, where these vessels now bifurcate to become the digital vessels proper (Fig. 17-1).

In the lower extremity, the superficial femoral and the profunda femoral vessels may be used as arterial inflow for fillet flaps of the thigh. The superficial femoral vessels continue distally to form the popliteal artery, which bifurcates in the popliteal fossa to the anterior tibial and posterior tibial arteries. The posterior tibial artery subsequently bifurcates to become the posterior tibial artery and the peroneal artery. The anterior tibial artery travels in the anterior compartment above the interosseous membrane, becoming more superficial as it approaches the extensor retinaculum of the ankle. At this point, this vessel becomes the dorsalis pedis artery, which travels between the first and second metatarsal to the foot. The superficial arcuate branch provides arterial inflow to the dorsum of the foot, while the deep branch pierces between the first and second metatarsal to form the plantar vascular arch. The common digital vessels arise from this arch, and bifurcate at the metatarsal-phalangeal joints to become the digital vessels to the toes. The peroneal artery travels beside the medial aspect of the fibula and terminates as the lateral calcaneal artery behind the lateral malleolus. The poster tibial artery travels in the poster compartment of the lower leg and exits superficial behind the medial malleolus and subsequently bifurcates into the medial and lateral plantar vessels (Fig. 17-2).

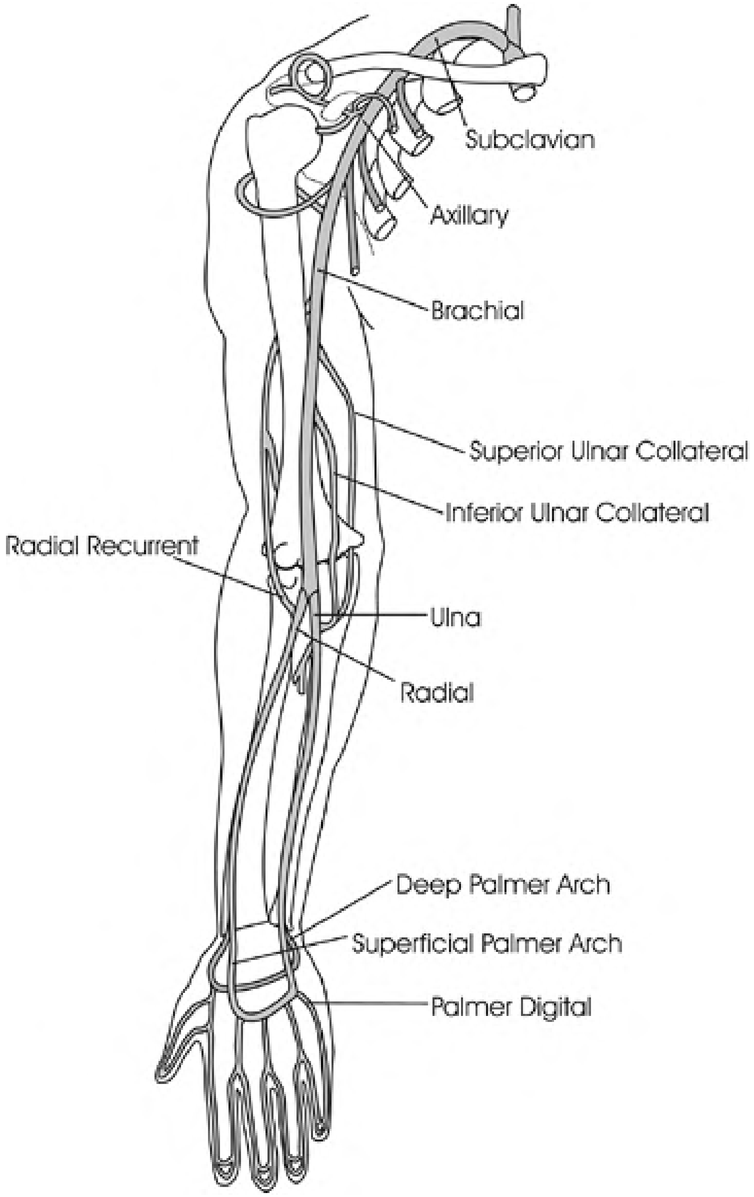
Subclavian
Axillary
Brachial
Superior Ulnar Collateral
Inferior Ulnar Collateral
Radial Recurrent
Ulna
Radial
Deep Palmer Arch
Superficial Palmer Arch
Palmer Digital

FIGURE 17-1 Upper extremity vascular anatomy. Axial pattern flaps can be designed from known vasculature and their smaller branches.

Indications/Contraindications

Before harvesting a fillet flap for soft tissue coverage, the surgeon should identify the ultimate goal of the immediate and delayed reconstructive procedures. Decisions may be less complex for some lower extremity trauma, where discarding the mutilating limb may be in the best interest of the patient from a functional point of view. Prostheses may be much more functional than attempts at salvage where stability and sensation are compromised. Conversely, mangled hands can have significant function if pinch and grasp can be restored. Reconstructive efforts may be turned toward

salvaging fingers or the thumb, rather than using their tissue as spare parts for soft tissue coverage for the remaining injured hand. The ability to restore some sensate function is a key component of the hand's role in activities of daily living. This separates the overall treatment protocols and goals of upper extremity reconstruction from that of lower extremity reconstruction (Fig. 17-3). The reconstruction process is dictated by the complexity and level of injury. Nerve regeneration, tendon and bone reconstruction, and secondary toe to hand procedures are alternatives to amputation of digits or hands and permit acceptable restoration of function. In both upper and lower extremity, however, there are opportunities that arise to use tissue that would otherwise be discarded, for definitive soft tissue closure.

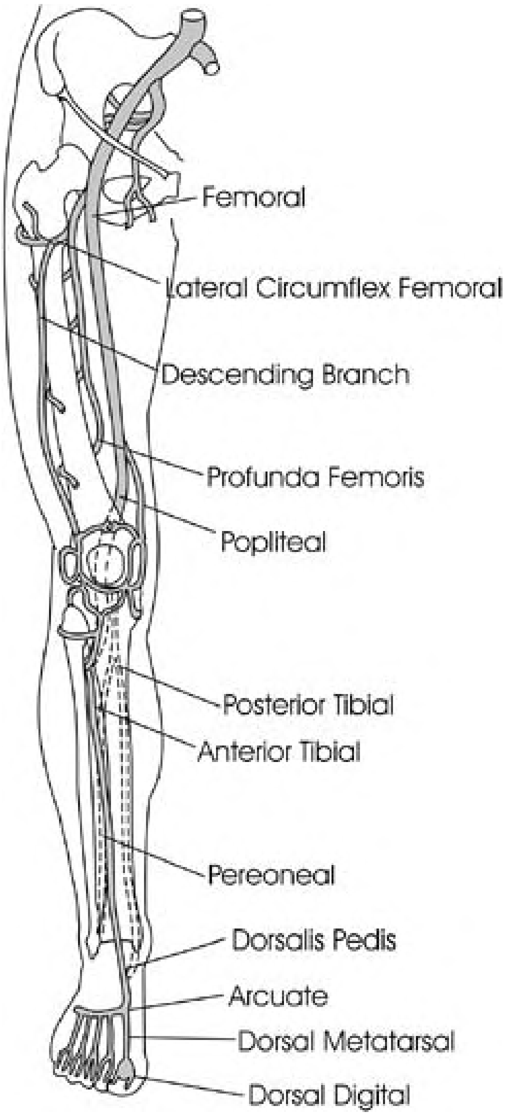
Femoral
Lateral Circumflex Femoral
Descending Branch
Profunda Femoris
Popliteal
Posterior Tibial
Anterior Tibial
Pereoneal
Dorsalis Pedis
Arcuate
Dorsal Metatarsal
Dorsal Digital

FIGURE 17-2 Lower extremity vascular anatomy.

The availability of adequate tissue often depends on the mechanism of injury to a limb. Clean guillotine lacerations or amputations, with minimal tissue loss, offer the best results at restoration of function. Fortunately, most of these injuries do not mandate that need of spare parts or fillet flaps due to the lack of soft tissue deficits. More traumatic injuries, such as crush and avulsion injuries, or electrical burns, require greater debridement and have more tissue loss. Consequently, these injuries frequently take advantage of spare parts and fillet flap surgery for coverage.

Pedicle or free tissue fillet flaps can be used for proximal amputations that require stable coverage to better preserve function, such as elbow flexion or knee mobility. This technique is usually indicated for amputations with significant proximal tissue loss and where there is a contraindication for replantation. The remaining distal tissue is filleted, based on one neurovascular bundle, and transferred to cover the proximal exposed structures as a free tissue transfer. If there is an intact neurovascular bundle, but with segmental intervening loss of tissue in the limb, the fillet flap can be harvested as a pedicle flap, leaving the vessels and nerves intact. Indications for fillet flaps are included in Table 17-1.

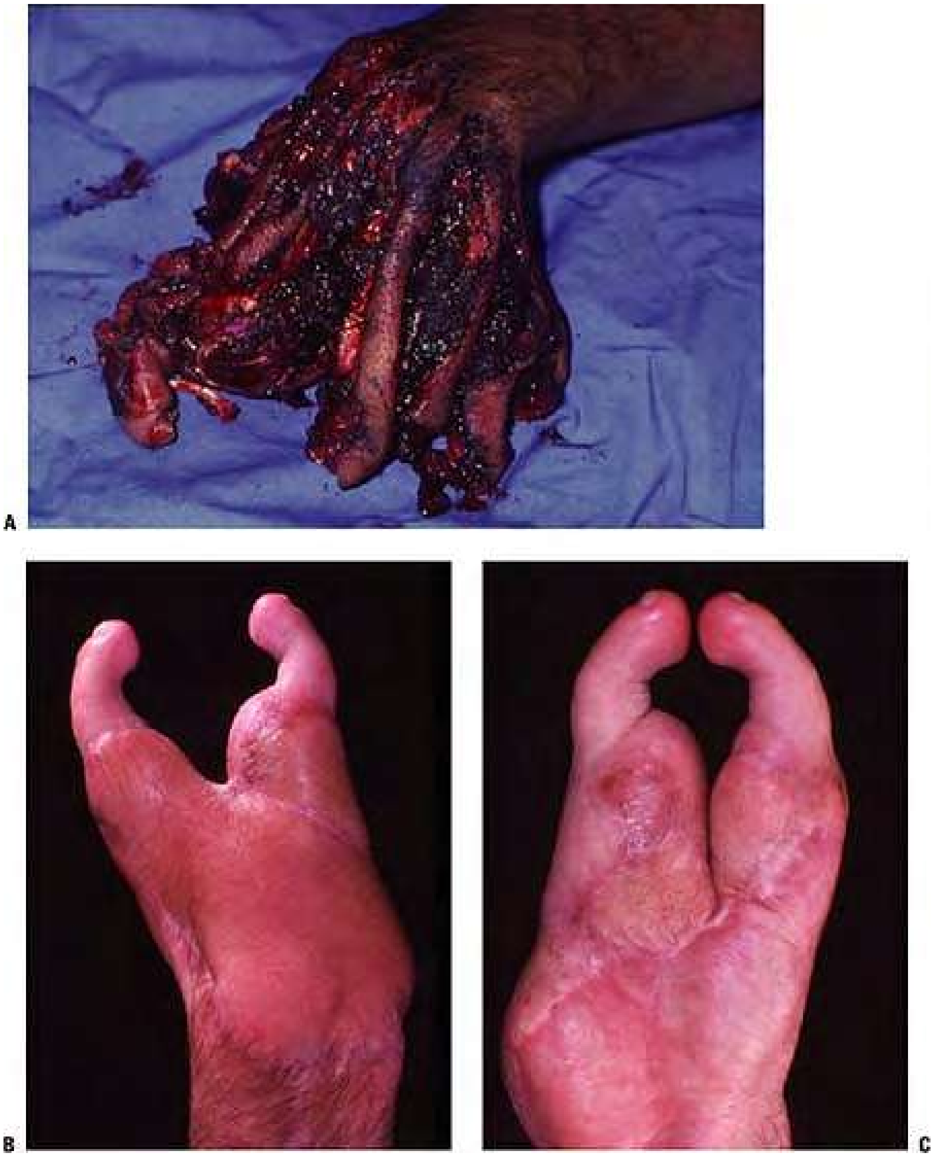

FIGURE 17-3 A: Mutilated hand from corn picker injury with complete loss of all fingers. B,C: Restoration of pinch from two toe-to-hand transfers.

Table 17-1. Indications for Fillet Flap

Nonreplantable extremity with adjacent tissue loss from trauma	Availability of tissue from discarded, amputated, or nonfunctioning limb or digit
Nonsalvageable digit with proximal or adjacent digit/hand defect	Segmental loss tumor extirpation or trauma

Preoperative Planning

The preoperative planning should take into consideration patient stability, premorbid health, a history of smoking, limb function (i.e., preservation of elbow flexion), available local tissue, prostheses application, and patient occupation. Patient selection is very important when contemplating the use of fillet flaps for soft tissue coverage of limb or trunk defects. In acute, mutilating trauma, the patient must be stabilized initially as survival should be the first goal. Following the full advanced trauma life support (ATLS) workup, with primary and secondary surveys, limb salvage is planned to optimize function.

The initial surgery should entail an aggressive debridement and irrigation to make sure all tissue is viable. Many times, a second and third surgery are required to allow tissue to demarcate and to avoid infection. Before using fillet tissues, the surgeon should plan to transfer based on the blood supply to the spare tissue. The fillet flaps can either be transferred as a free flap or a pedicle flap. Digital fillet flaps are neurovascular flaps usually based on the digital arterial and nerve supply to the finger (Fig. 17-4). The surgeon will plan to use the digit that has been, in part, rendered dysfunctional as a result of the proximal trauma but where viability is still present distally through an intact neurovascular bundle. In the elective cases, angiography may help to identify intact vascular anatomy before transfer. Angiography is rarely indicated in the acute traumatic case as the surgeon can visualize the vessels and observe the vascularity (bleeding) of the distal tissues.

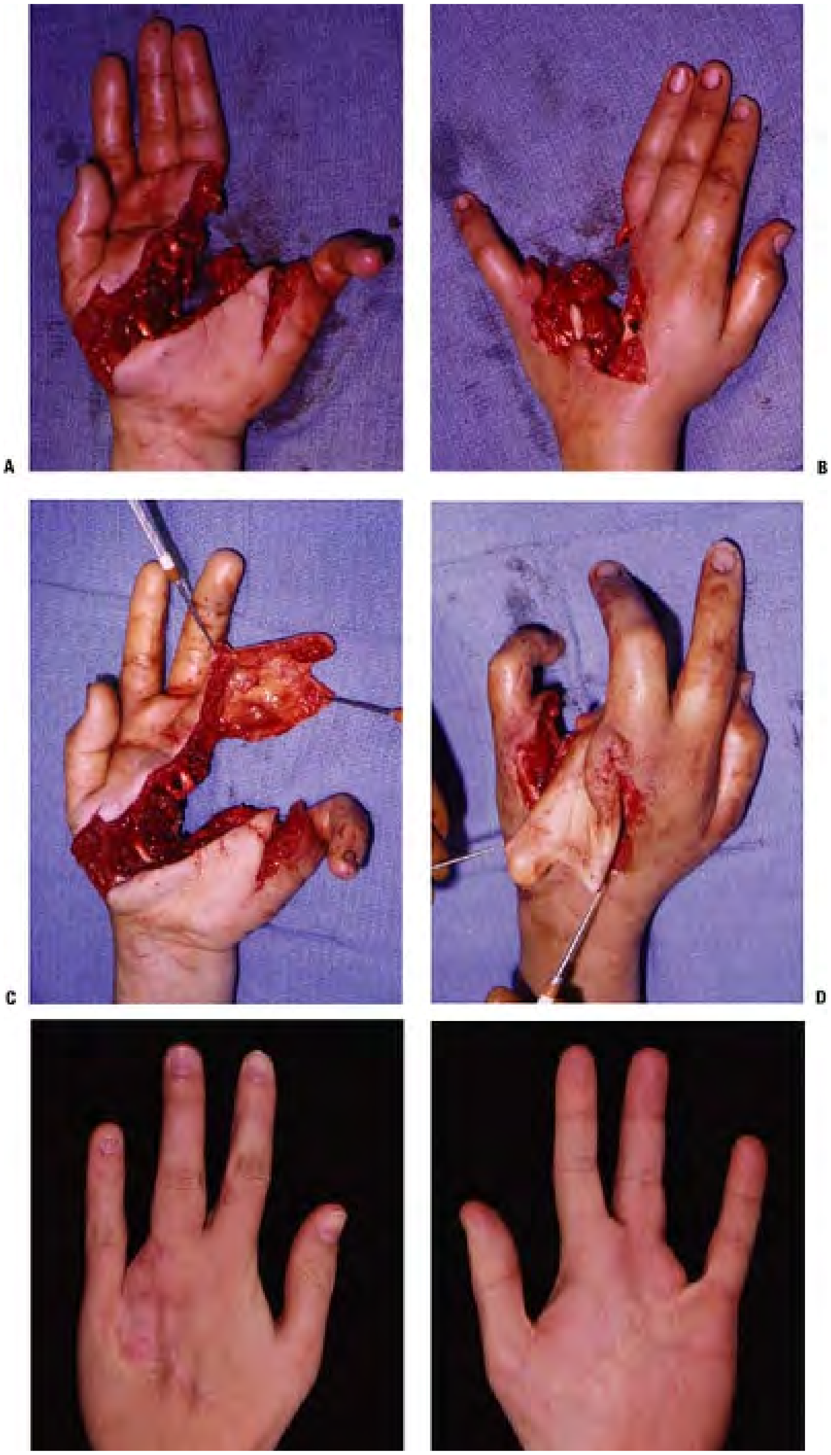

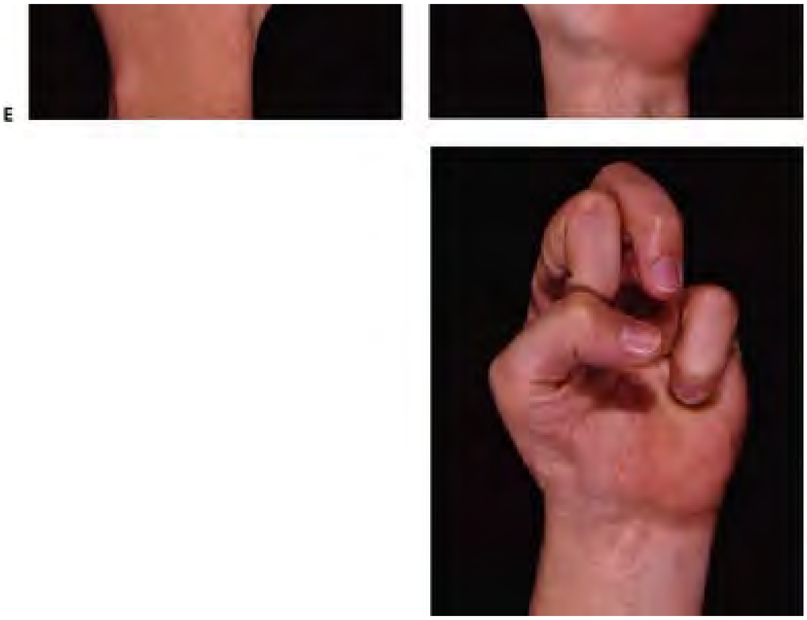

FIGURE 17-4 Aâ€“D: Tablesaw to the right hand with intrusion through the web space between the small and ring finger in an ulnar to radial direction. Complex repair of flexor/extensor tendons, neurovascular repair of the arch and common digital vessels was undertaken. The ring finger is filleted based on its radial neurovascular bundle. Resection of the head of the fourth metacarpal was performed secondary to the level of comminution. Long-term follow-up reveals good prehension. Eâ€“G: Tablesaw to the right hand with intrusion through the web space between the small and ring finger in an ulnar to radial direction. Complex repair of flexor/extensor tendons, neurovascular repair of the arch and common digital vessels was undertaken. The ring finger is filleted based on its radial neurovascular bundle. Resection of the head of the fourth metacarpal was performed secondary to the level of comminution. Long-term follow-up reveals good prehension.

Surgery

Patient Positioning

The patient should be positioned appropriately to allow dissection of any digit, arm, or leg based on the blood supply of that part. The patient is usually in the supine position for most upper extremity and lower extremity flaps. An exception to this is the use of a fillet flap for recurrent, recalcitrant pressure sores on the ischial and para-sacral areas, for which the patient should be placed in the prone position.

Technique

The assessment of vascular integrity should be performed before inflation of the tourniquet. Although a pulse-lavage system can be employed, a simple spray bottle allows adequate force to propel smaller particulate debris from the wound (Fig. 17-5). Larger debris should be mechanically debrided with sharp dissection. The preservation of questionable tissue is acceptable if further assessment is planned in 24 to 48 hours for further debridements. Temporary coverage with dressings, xenograft, or allograft can be used in the interim to prevent desiccation in the interim (Fig. 17-6).

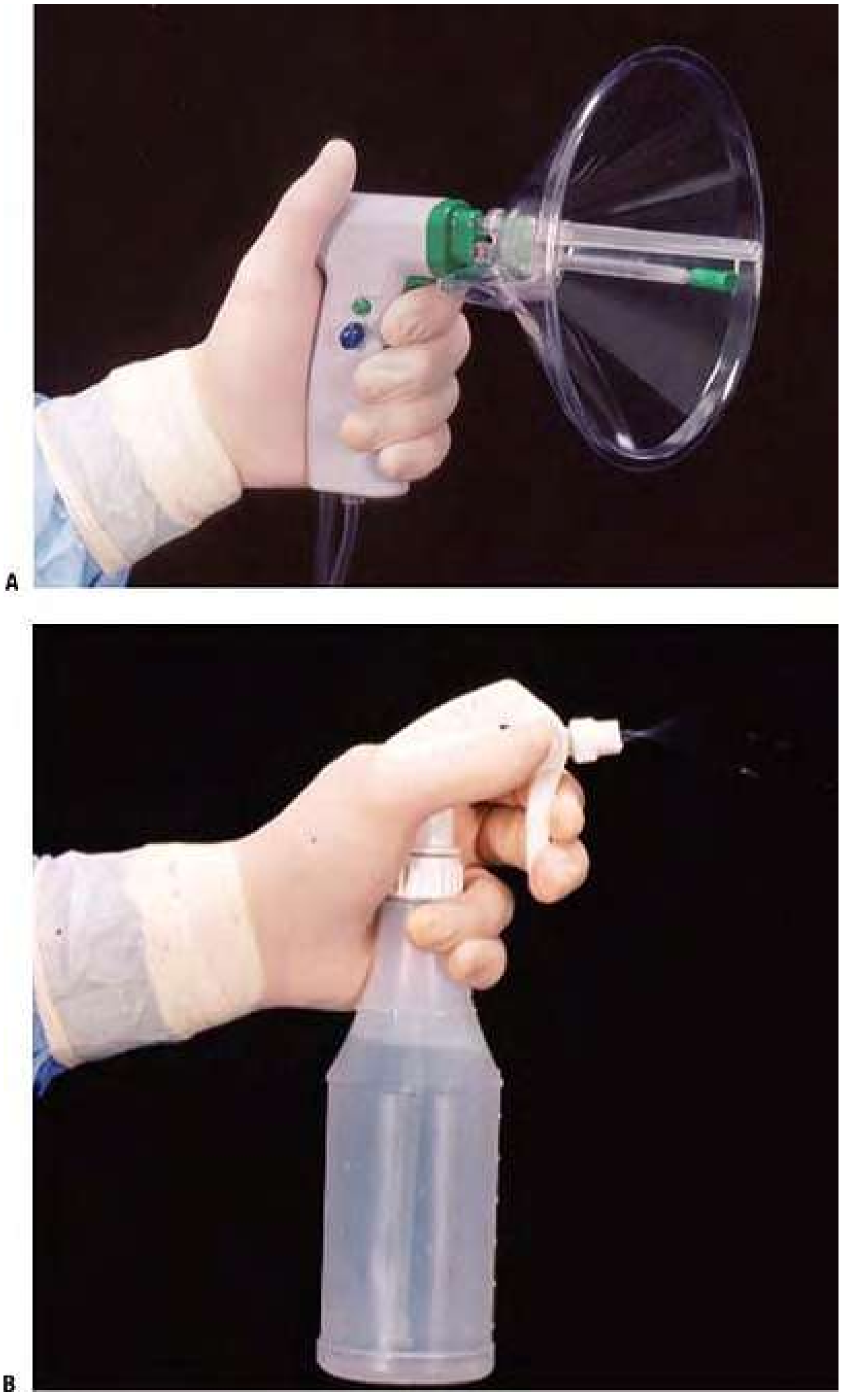

FIGURE 17-5 A: Pulse lavage system. B: Simple spray bottle.

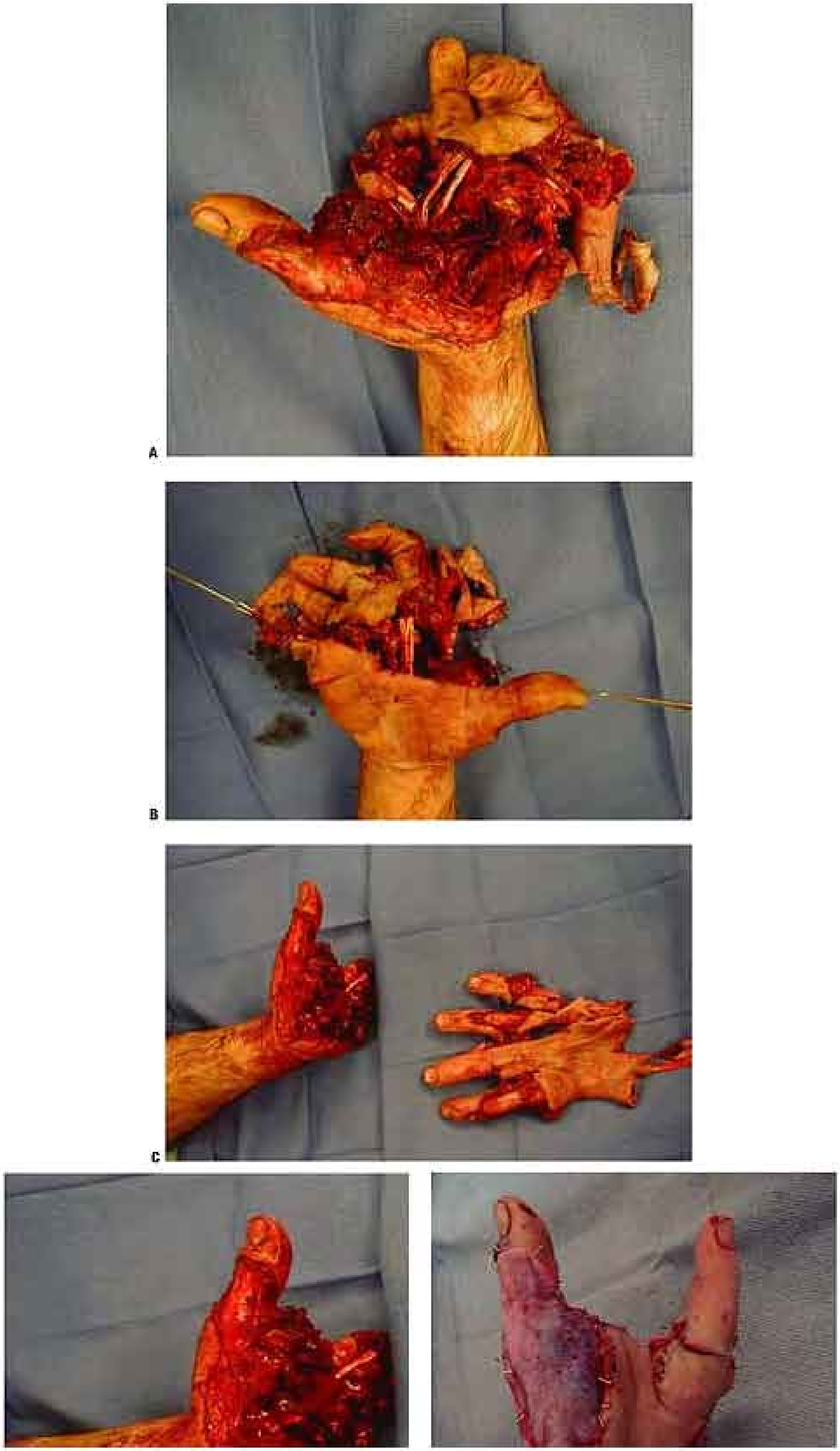
A
B
C

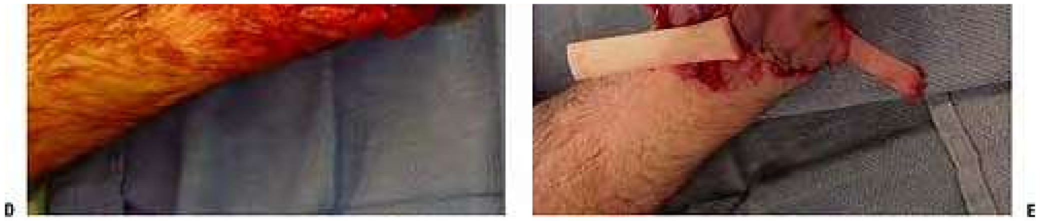

FIGURE 17-6 Aâ€“C: Mutilating injury necessitating temporary coverage with allograft secondary to the size of the wound, and possible need for further debridement. Dâ€“E: Mutilating injury necessitating temporary coverage with allograft secondary to the size of the wound, and possible need for further debridement.

Digital Fillet Flaps of the Hand

The hand provides an ideal setting for the use of fillet flaps for soft tissue coverage of mutilating injuries. A single finger can provide all the components required for composite reconstruction of the hand. Digital fillet flaps can be designed to cover both palmar and dorsal hand defects (Fig. 17-7). The use of adjacent digits that are rendered dysfunctional from the injury can provide coverage with glaborous skin and normal sensibility (Fig. 17-8). Similarly, microsurgical techniques can be used to move the filleted digit to an area of the hand not immediately adjacent to the flap. Idler describes the use of a filleted digit as a free flap to transfer to cover proximal hand wounds. An osteocutaneous fillet flap was described by Gainor in which the index finger with the bone intact was transferred to reconstruct the thumb. A composite, neurovascular island flap can be used in a similar fashion incorporating the distal phalanx, nail bed, and skin for sensate reconstruction of a hemi-thumb defect (Fig. 17-9).

The vascularity of the digit to be a fillet is determined by gross inspection and the use of a hand-held Doppler. The arc of rotation depends on the fillet flap's vascular pedicle. The fillet flap must reach the primary defect based on the pivot point of a pedicle. Al-Qattan described lengthening a fillet finger flap by incising the skin at the finger. He then recommends basing the flap on either the radial or ulnar vessel, which will allow one to unfold of the flap and increase its functional pedicle length. Preoperative planning using a surgical sponge to measure the arc of rotation and the subsequent length of the flap required is a helpful tool in determining the utility of the proposed flap. Incomplete filling of the defect will result in the need for skin grafting some other form of soft tissue coverage. Split or full thickness skin grafts can be harvested from remaining spare parts. Excess tissue within the fillet flap can be used to fill in contour defects. In such cases, the distal portion of the fillet flap can be de-epithelialized and folded beneath the remaining flap to augment soft tissue bulk.

Fillet Flaps of the Legs and Arms

Fillet flaps of the legs and arms are used to cover proximal amputation stumps to prevent further shortening of the limb or to add length around the elbow or shoulder. In upper extremity trauma, these fillet flaps are based on the radial or ulnar arteries. The flap is usually a free tissue transfer because there is often an intervening defect in the vessels. The brachial artery and venae commitantes act as the recipient vessels around the elbow. In lower extremity trauma, the fillet flap can be based on either the anterior tibial or the posterior tibial vessels. Incisions in the distal amputated part are made on the opposite side of the limb from where the blood supply is present. This keeps the blood supply to the skin, subcutaneous tissue, and muscles based axially with the medial and lateral tissue equal distances from the primary vessels. Care must be taken during the removal of the bony and muscular components of the flap not to injure the neurovascular bundles. The flap is trimmed appropriately to fit the recipient defect. The fillet flaps can be harvested as a pedicle flap or as a free tissue transfer (Fig. 17-10).

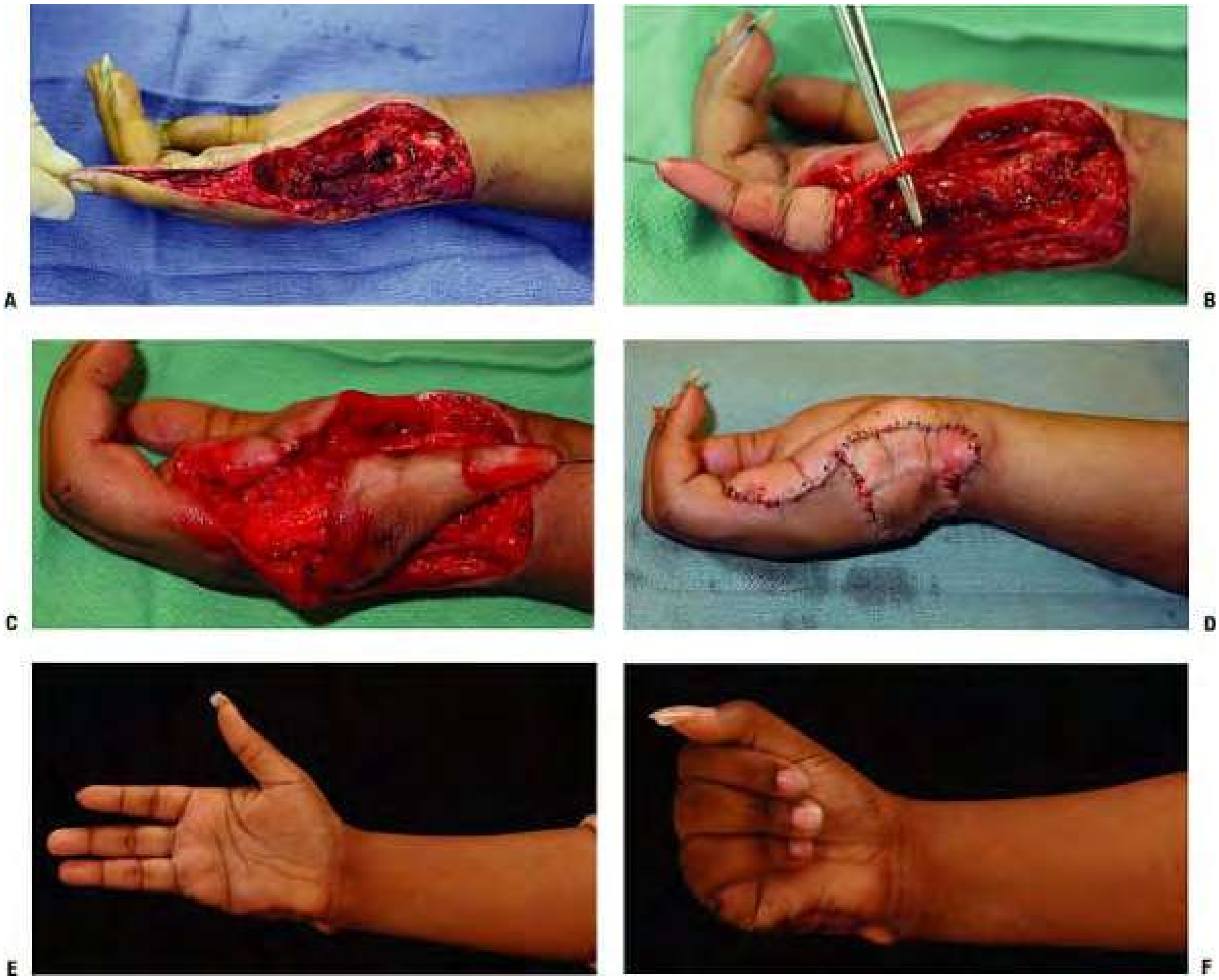

FIGURE 17-7 Fourteen-year-old patient with hypothenar defect. Aâ€“D: The small finger is filleted based on its radial neurovascular bundle for coverage. This technique provides stable, sensate closure of the defect. E,F: Three-month follow-up reveals stable coverage and good hand function.

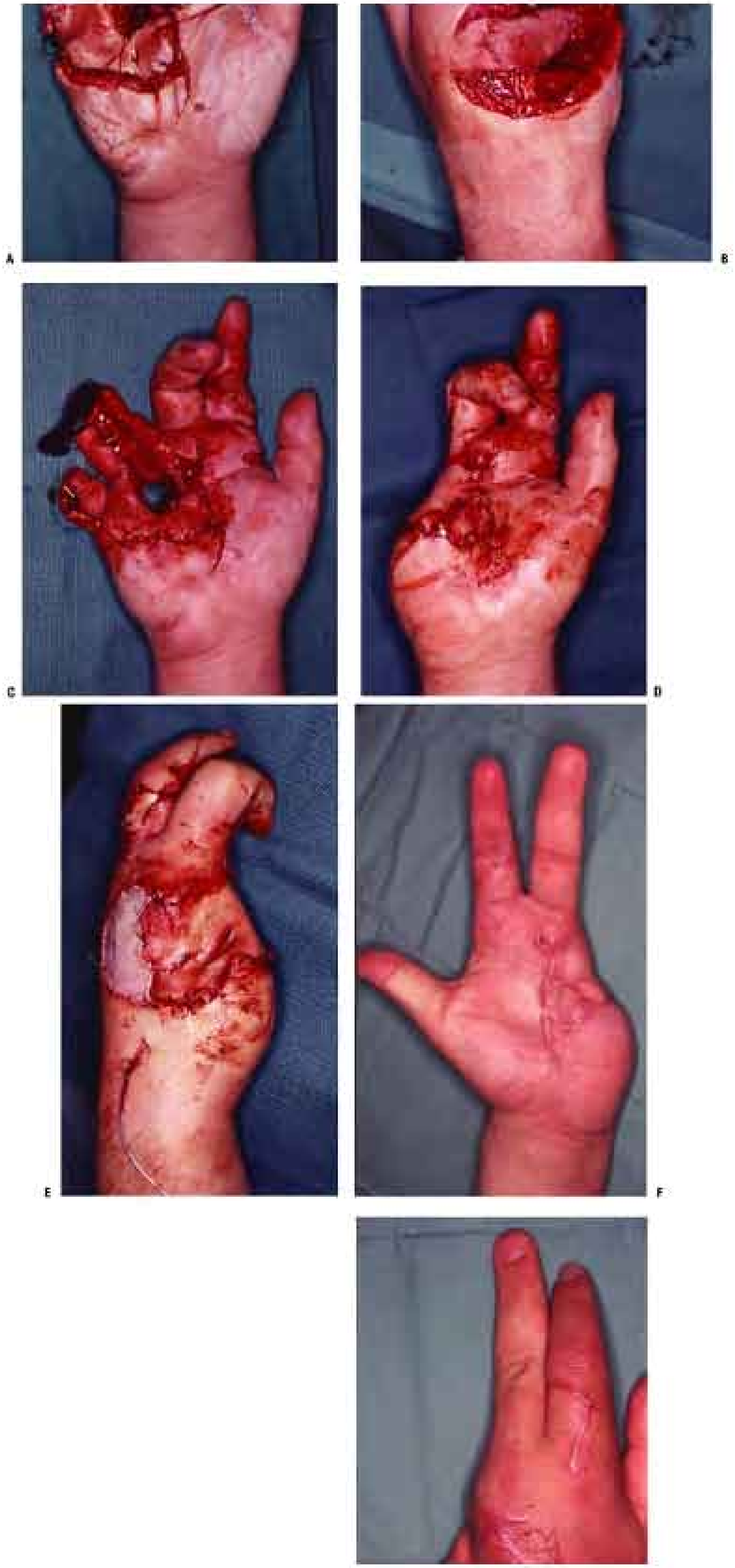
A
B
C
D
E
F

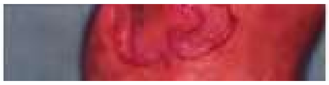

FIGURE 17-8 Aâ€“D: Punch press injury to ulnar aspect of left hand. Ring and small fingers filleted for coverage. A small skin graft is placed dorsally beyond the reach of the tip of the flap. Long-term postoperative follow-up reveals stable coverage. Eâ€“G: Punch press injury to ulnar aspect of left hand. Ring and small fingers filleted for coverage. A small skin graft is placed dorsally beyond the reach of the tip of the flap. Long-term postoperative follow-up reveals stable coverage.

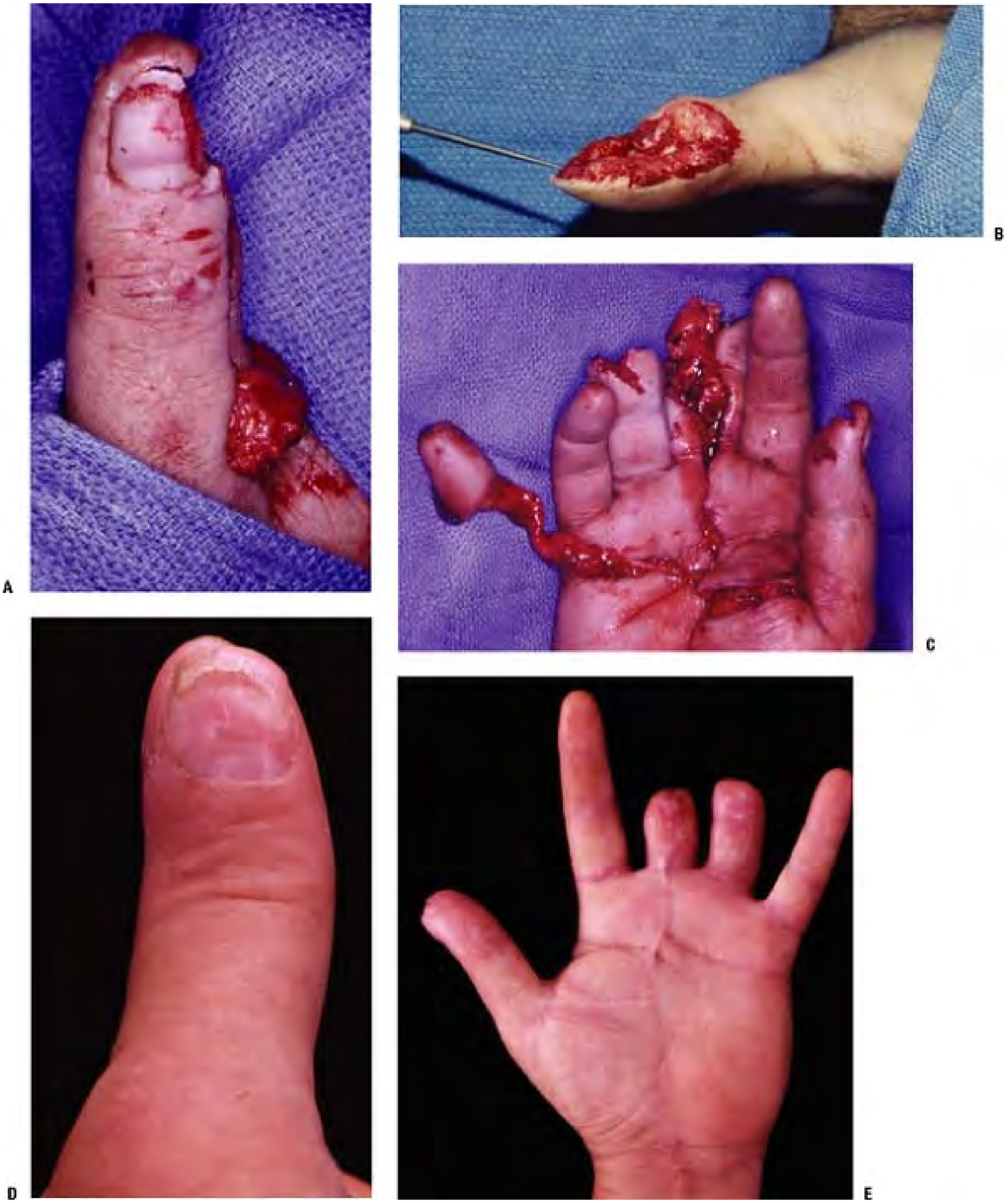

FIGURE 17-9 A,B: Table saw injury to ulnar aspect of a thumb. Câ€“E: Additional injury to the middle and ring fingers. Neurovascular island flap harvested from the middle finger transferred for composite coverage of the thumb.

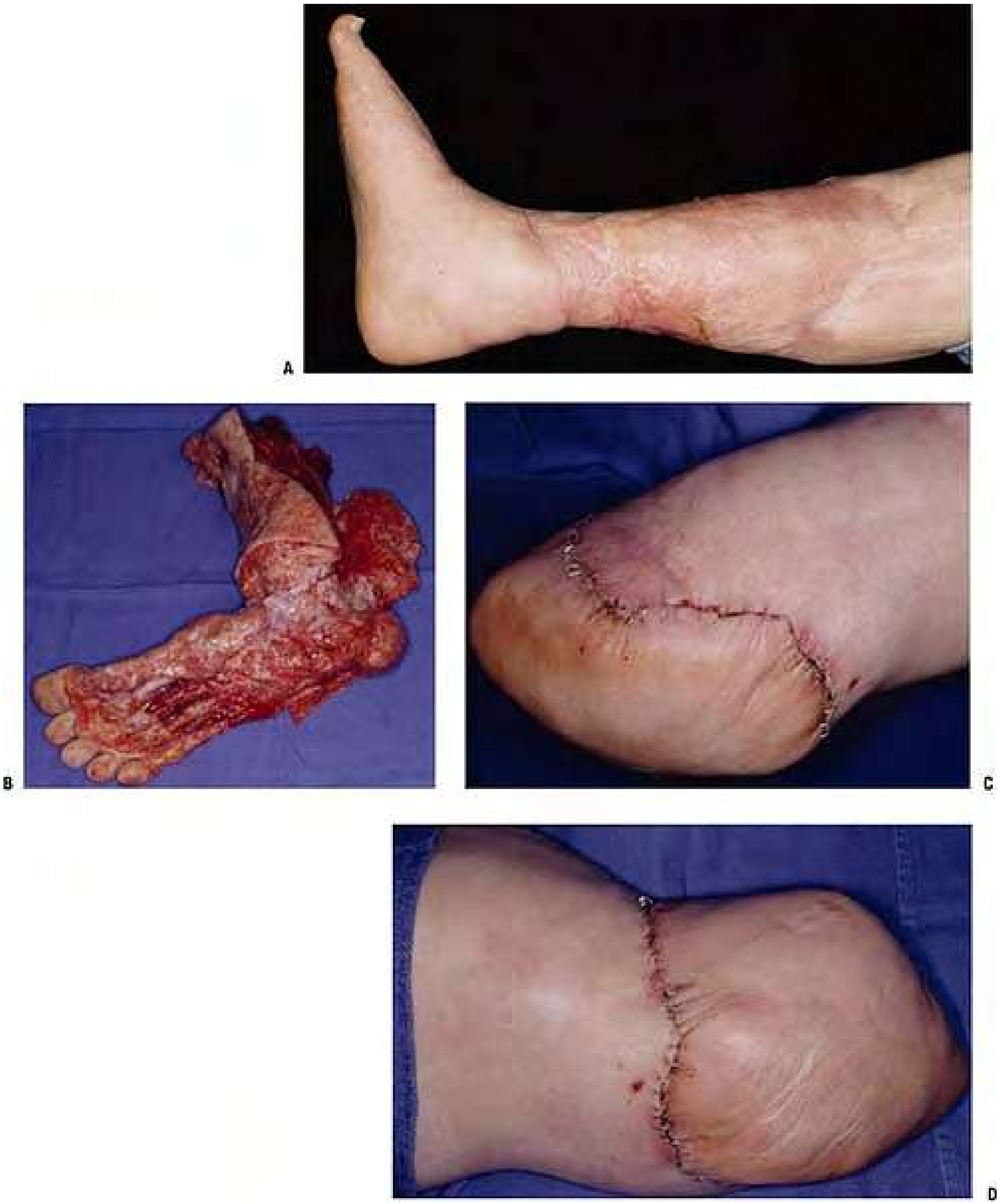

FIGURE 17-10 Lower extremity reconstruction after amputation. The sole of the foot is filleted and transferred proximally to cover the proximal stump after amputation.

Postoperative Care

Postoperative care depends on the area reconstructed. For upper extremity defects, bulky noncompressive dressings are used to protect lacerated or damaged digits. A window is left within the dressing to monitor the vascular status of the fillet flap. If the spare part has been transferred as a free flap, then flap monitoring is performed hourly for the first 24 to 48 hours to ensure there are no signs of arterial ischemia or venous congestion. Signs of either problem will necessitate anastomotic re-exploration to prevent flap failure.

If extensive skin grafting has been performed, motion within the digit or extremity is usually delayed for 5 to 7 days to allow for skin graft take. If a bolster dressing has been placed or if skin grafts have been secured with the use of vacuum-assisted-closure (VAC), motion may be initiated sooner.

Lower extremity wounds are also treated initially with bulky noncompressive dressings and immobilization with the application of a posterior splint. Leg elevation is essential to limit swelling and minimize the risk of venous congestion within the flap. At 1 week postoperatively, compressive wraps may be initiated to the flap and reconstructive area to help with the resolution of postoperative edema. Compressive wrapping may also aid in long-term flap contours.

Results/Complications

Flap loss is obviously the most devastating complication following fillet flap reconstruction. In such cases, alternative means of wound coverage will be necessary to cover resultant defects. Kuntscher et al. present their results from pedicle fillet flaps of fingers for dorsal, palmer, and adjacent digit defects secondary to trauma, burn wounds, Dupuytren's disease, malignant tumor resection, and diabetic gangrene. In their series of 30 patients, the fillet flap provided a stable coverage option with normal sensibility compared to the unaffected digits. The overall complication rate was 18%, with a flap loss, revision surgery, and infection rate of 7.5%.

Recommended Reading

al Qattan MM. Lengthening of the finger fillet flap to cover dorsal wrist defects. *J*

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part III - Management of Soft Tissues Within the Hand and Wrist > Chapter 18 - The Use of Free Flaps in Upper Extremity Reconstruction/ Anterolateral Thigh Flap

---

Chapter 18

The Use of Free Flaps in Upper Extremity Reconstruction/ Anterolateral Thigh Flap

Lawrence Lin
Samir Mardini
Steven L. Moran

Indications/Contraindications

Soft tissue deficiencies within the upper extremity are common following trauma, burns, infection, and tumor extirpation. The coverage of such defects can usually be accomplished with the use of pedicled flaps or local rotational flaps. However, when defects are very large or encompass multiple structures including nerve, bone, or muscle, the use of composite free tissue transfer provides a reliable and single stage means of reconstructing complex defects.

The benefits of free tissue transfer within the upper extremity include the transfer of additional vascularized tissue to the injured area, the ability to carry vascularized nerve, bone, skin, and muscle to the injured area in one procedure, and the avoidance of any additional functional deficits to the injured limb which may be incurred with the use of a local or pedicled flap. Free flaps are not tethered at one end, as is the cases for pedicled flaps, and this allows for more freedom in flap positioning and insetting. More recent fasciocutaneous and perforator flaps also allow for primary closure of donor sites with minimal sacrifice of donor site muscle. With current microsurgical techniques, free flap loss rates are between 1% and 4% for cases requiring elective free tissue reconstruction. Finally, the upper extremity is particularly suited for free tissue transfer as the majority of recipient blood vessels utilized for anastomosis are located close to the skin, and are of relatively large caliber.

Major indications for free tissue transfer are: (a) primary coverage of large traumatic wounds with exposed bone, joint, and tendons or hardware, (b) coverage of complex composite defects requiring bone and soft tissue replacement, (c) coverage of soft tissue deficits resulting from release of contractures or scarring from previous trauma, and (d) significant burns.

There are few absolute contraindications for free flap transfer and in many cases free tissue transfer may be the only option for upper limb salvage following significant soft tissue loss. Despite this, relative contraindications to free tissue transfer include a history of a hypercoagulable state, history of recent upper extremity DVT, and evidence of ongoing infection with the traumatic defect. Other contraindications would include inadequate recipient vessels for flap anastomosis. Disregarding technical error, the status of the recipient vessel used for flap anastomosis may play the greatest role in flap failure; recipient vessels within the zone of injury are prone to postoperative and intraoperative thrombosis. Recipient vessels for microvascular transfer should ideally be located out of the zone of injury, radiation, or infection. In rare cases, arterial venous fistulas may be created proximally within the upper extremity or axilla using the cephalic or saphenous vein. These fistulas can be brought into the zone of injury and divided to provide adequate inflow and outflow for free tissue transfer.

Specific Indication for the Anterolateral Thigh Flap

There are many choices for free flap coverage of the upper extremity. The scapular, parascapular, latissimus dorsi myocutaneous, and lateral arm flap have long been favorites of surgeons for reconstruction of traumatic upper extremity injuries. Many of these flaps will be described in later chapters of this text. If joints are to be crossed, fasciocutaneous flaps are much preferred as muscle flaps can undergo atrophy and restrict flexion and extension across joints or fingers (Fig. 18-1).

Classic cutaneous free flaps, such as the radial forearm flap, lateral arm flap, and scapular flap, have limitations in size, donor site morbidity, and overall thickness. Musculocutaneous flaps such as the latissimus dorsi and rectus abdominus flaps result in functional loss and donor site morbidity including, particularly in the abdomen, potential hernia formation. In addition, in coverage of joint surfaces, muscle flaps tend to undergo fibrosis and atrophy over time, which may limit muscle excursion, particularly when placed over the elbow or dorsum of the hand. Muscle is still indicated for those circumstances involving osteomyelitis or significant soft tissue contamination. More recently, chimeric flaps have been harvested to include both muscle and a large component of skin, providing the ideal coverage for many complex defects in the upper extremity.

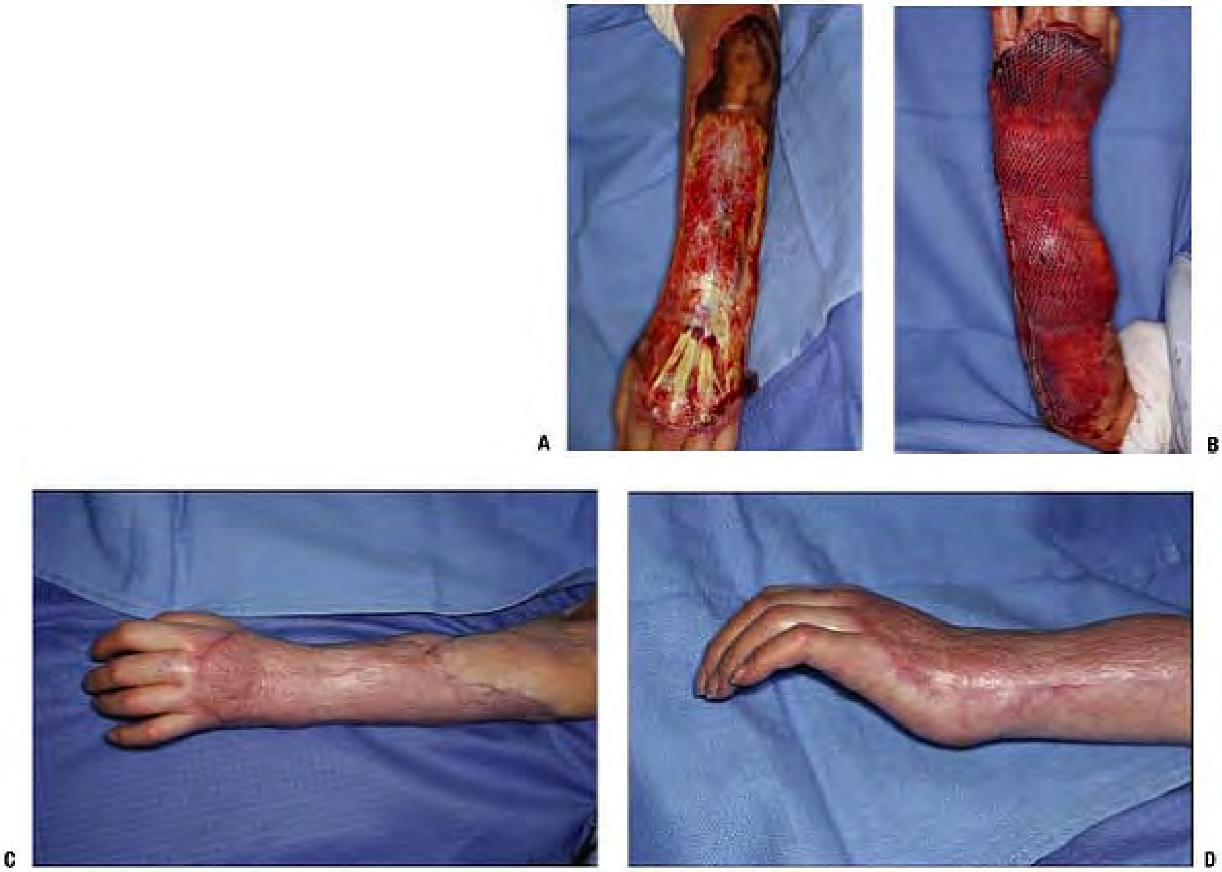

FIGURE 18-1 A: An extensive iv infiltrate to back of hand and forearm resulted in

full thickness skin loss over the majority of the hand and forearm in this 56 year old woman. B: Soft tissue coverage was obtained with the use of a rectus abdominus free flap. C,D: Despite successful soft tissue coverage, atrophy and fibrosis within the muscle over time has lead to limitations in wrist motion resulting in decreased wrist flexion.

In recent years, the anterolateral thigh flap has become the major flap in reconstructive microsurgery, including head and neck defects and extremity wounds. It has replaced many other flaps. The skin overlying the anterior thigh region has relatively constant anatomy with the descending branch of the lateral femoral circumflex artery (LFCA), giving rise to either pure muscular (musculocutaneous) or subcutaneous (septocutaneous) perforators that supply the area. Based on their experience with over 1,500 anterolateral thigh flaps for various anatomic defects including the upper extremity, Chen et al determined that 12% were based on direct septocutaneous perforators, and 88% were based on musculocutaneous perforators. Variations in perforator anatomy can exist, which include absence of skin perforator, perforators which are too small for elevation, a perforator artery which does not run with the vein, and perforator arteries that have no accompanying vein. These anatomical variations are rare, accounting for 2% of cases, however they need to be noted by the surgeon. As proposed by Chen, an algorithm for managing anatomical variations begins with attempting to identify a more proximal perforator in the upper thigh, usually arising from the transverse branch of the LFCA, and harvesting the flap based on this perforator. Alternatively, an anteromedial thigh flap may be raised or the vastus lateralis may be taken as a musculocutaneous flap. Finally, exploration can be performed on the contralateral side as the anatomy may be different.

The anterolateral thigh flap serves as the ideal flap for upper extremity coverage due to its many considerable advantages. The flap provides a long pedicle (up to 16 cm) with suitable vessel diameters; the arterial pedicle can measure up to 2.5 mm and the two venae comitantes can measure up to 3 mm in diameter. It is also a versatile flap with the ability to incorporate different tissue components with large amounts of skin, as the flap can be harvested as a cutaneous, fasciocutaneous, or musculocutaneous flap with vastus lateralis. In addition, based on the supply of the LFCA system, a chimeric flap incorporating the rectus femoris or tensor fascia lata can be raised to cover extensive, complex defects. The flap may be harvested as a sensate flap by including the lateral femoral cutaneous nerve or as a flow-through flap in cases of significant arterial trauma. Inclusion of thigh fascia with the flap allows its use as an interposition graft for tendon reconstruction. The thickness of the flap may be debulked primarily, optimizing the match of donor tissue for the upper extremity. Ordinary skin flaps can sometimes produce bulkiness with poor aesthetics. Thick skin paddles, such as with parascapular flaps, may interfere with motor function and flexion of the metacarpal phalangeal joints or inter-phalangeal joints. Flaps as thin as 3 mm to 5 mm have been harvested for tendon coverage. The donor site results in minimal morbidity with most sites able to be closed primarily, resulting in a linear scar and absence of any long term leg dysfunction. Lastly, its anatomic location allows for a two-team approach for flap elevation and recipient site preparation, saving considerable operative time.

Anterolateral Thigh Flap for Upper Extremity Reconstruction

Preoperative Planning

Preoperative requirements for flap consideration begin with the preparation of a clean wound bed. Radical debridement of all necrotic tissue is the most important component of a successful reconstruction. Tissue considered to be of marginal viability should be debrided early rather than performing multiple dressing changes or utilizing vacuum-assisted therapy in the hopes of rescuing traumatized tissue; such measures can lead to delayed definitive surgical reconstruction, perpetuate the inflammatory component of wound healing, perpetuate distal edema, and result in hand and limb stiffness. If the surgeon can guarantee a clean wound bed, free of any necrotic material, immediate flap coverage may be attempted in cases of acute trauma. We have found that most high energy traumatic injuries and agricultural accidents require at least one to two surgical debridements prior to definitive wound closure. Wound debridements in these cases are performed in conjunction with wound cultures for bacteria and fungal species. The ideal timing for upper limb free tissue reconstruction has been debated within the literature but should be within 72 to 96 hours of injury.

The upper extremity is evaluated for any evidence of concomitant bony or neurovascular injury. A careful vascular evaluation is also performed and if there is any question as to the status of the inflow vessels, an angiogram or a CT angiogram may be obtained to verify inflow. Our preference is to perform the majority of arterial anastomosis in end-to-side fashion, while the veins are anastomosed in an end-to-end fashion. If consideration is being given to performing an arterial anastomosis in end-to-end fashion to either the radial or ulnar artery, the surgeon must verify a patent palmar

arch with an Allen's test or Doppler examination prior to surgery. The donor leg for the anterior lateral thigh flap should be free of concomitant soft tissue trauma. If the patient has a history of lower extremity arterial atherosclerotic disease or diabetes, examination of the lower extremity is warranted, to verify that there is an intact profundus femoral artery which gives rise to the lateral femoral circumflex system.

Patient positioning

The patient is positioned in the supine position for harvest of the anterior lateral thigh flap. The injured arm is positioned on a standard hand table. Recipient site preparation is aided with the use of an upper extremity tourniquet.

Flap harvest is usually performed under general anesthesia, though spinal/epidural block for flap elevation and axillary block of the affected extremity can be undertaken when general anesthesia is contraindicated.

Flap Elevation Surgery

General principles and steps in elevation for perforator flap include:

- Doppler mapping of the perforator,
- Design of the flap for operation,
- Identification of the perforators leading to the main pedicle,
- Intra-muscular dissection of the perforator with preservation of the nerves,
- Elevation of the flap,
- Thinning of the flap, and

- Transfer of the flap to cover the defect.

The anterolateral thigh flap is based on the lateral femoral circumflex vessel, particularly the descending branch, which courses inferiorly along the intermuscular septum giving rise to subcutaneous and/or intramuscular perforators that penetrate the fascia to supply the skin and subcutaneous tissue overlying the anterolateral thigh. A preoperative Doppler examination of the lower leg is performed to identify these perforators. A line is drawn from the anterior superior iliac spine to the lateral margin of the patella. At the midpoint of this line is a reliable perforator identifiable in almost 90% of cases (Fig. 18-2). An additional perforator can be identified in 80% of cases within the upper third of the line, again using a hand held Doppler. The skin paddle, incorporating both perforators if possible, is designed in an elliptical fashion based on the dimensions of the defect in the upper extremity.

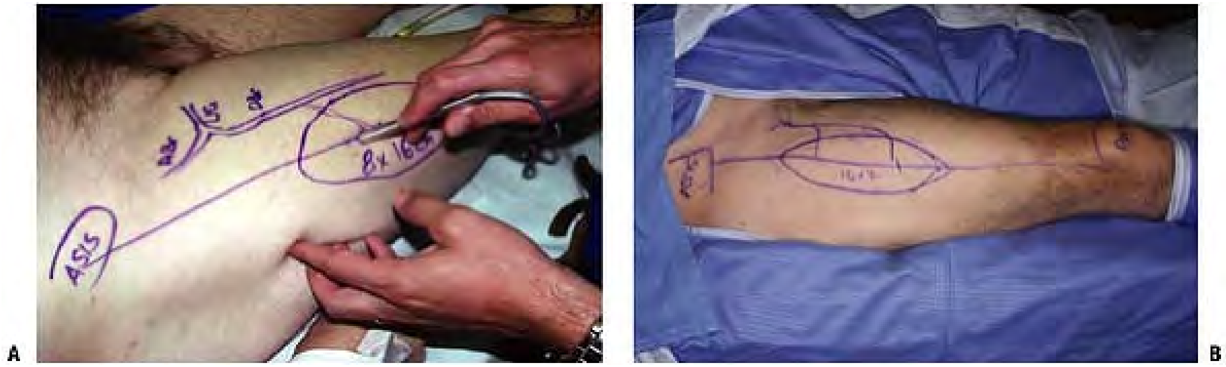

FIGURE 18-2 Prior to flap elevation the arterial perforators to the ALT flap are identified on the leg. A line is drawn from the anterior superior iliac spine (ASIS) to the lateral margin of the patella. In 90% of cases a major perforator can be identified within 3 cm of the middle portion of this line. A second major perforator is usually identified at the upper third of this line. A: The sites for the perforators are confirmed with a handheld Doppler probe and marked with a skin scribe. B: The flap is then designed around the perforators in an elliptical fashion.

Dissection begins medially by dissecting down to the rectus femoris fascia. The fascia of the rectus femoris muscle is divided and dissection proceeds in a lateral direction until the intermuscular septum separating the rectus femoris and vastus lateralis muscles is encountered. Medial retraction of the rectus femoris muscle exposes the entire septum, allowing for visualization of the descending branch of the lateral femoral circumflex artery in addition to its branch going to the rectus femoris itself (Fig. 18-3). At this point there may be 1â€“3 major perforators exiting from the descending branch and passing through the muscle of the vastus lateralis as musculocutaneous perforators or passing directly through the septum as septocutaneous perforators to the skin. Usually, one large perforator is sufficient to supply a large skin island, and allows for easy thinning of the flap. However, more perforators may be included if the skin island is very large, and flaps as large as 15 Ã— 35 cm have been reported. If septocutaneous perforators are present, one may simply proceed to the posterior incision, including the portion of the tensor fascia lata, and elevate the flap. If intramuscular perforators are encountered then dissection must proceed as one would for a true perforator flap, in which case the artery needs to be carefully dissected free from the surrounding muscle. A cuff of muscle may be included in the flap should the upper extremity wound require additional bulk or if the wound contains exposed bone or significant dead space (see Figure 18-3D).

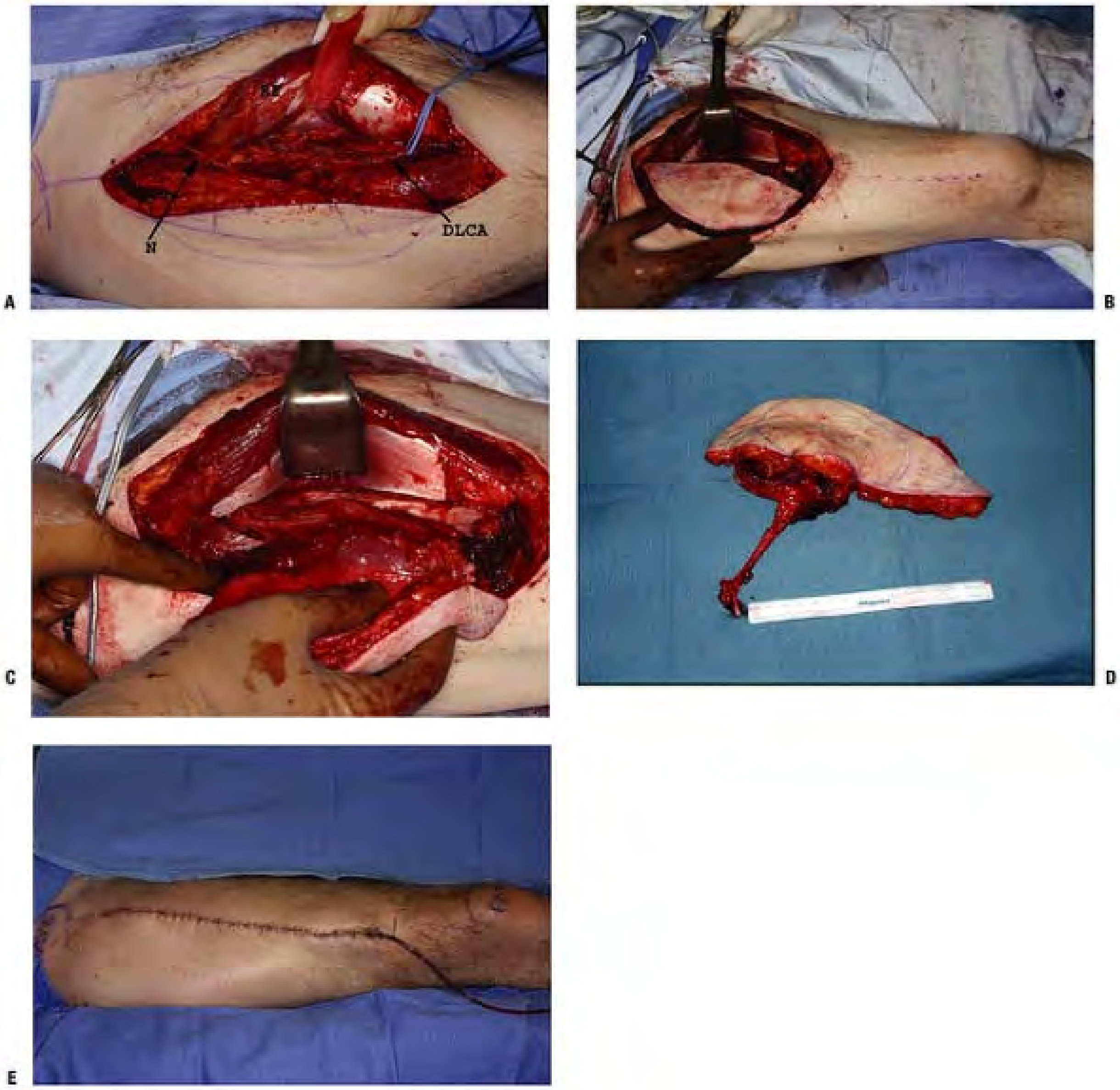

FIGURE 18-3 A: After the flap has been marked the medial incision is made first and the rectus femoris muscle is identified. The rectus femoris (RF) muscle is retracted medially to reveal the descending branch of the lateral femoral circumflex vessels (DCLA). A branch of the lateral femoral cutaneous nerve (n) can usually be identified running superior to the vastus lateralis muscle B: Once the flap perforators are identified, dissection may begin at the lateral margin of the flap. C: In this photograph 2 perforators of the lateral femoral circumflex vessels are seen passing into the vastus lateralis. These vessels may be dissected

through the muscle into the fascia of the overlying skin flap or a small cuff of muscle may be preserved around the vessels. D: The flap is now ready for transfer to the arm. The muscle cuff can be seen beneath the skin paddle. The pedicle length and large diameter allow for easy insetting and microvascular anastomosis. E: The donor site may be closed primarily over a closed suction drain.

Dissection proceeds along the descending branch of the LFCA separating venae comitantes from the artery. The pedicle can be dissected back to the main trunk of the lateral femoral circumflex artery, or divided at the descending branch. If one traces the vessels back to the origin, the motor nerve to the vastus lateralis will need to be dissected free from the arterial pedicle. This may be a tedious dissection as the nerve may pass through and around the venous pedicle. Dissection of the vessels back to the origin results in very large caliber vessels of up to 3 mm in some cases.

Preparation of the arterial recipient site is then performed. If the radial artery is going to be used, a Henry approach is usually carried out. For approach to the brachial artery a curved lazy-S incision is performed over the antecubital fossa and exposure of the brachial artery at the level of the medial forearm is performed. The flap is then transferred to the upper extremity. The tourniquet is then released in order to verify adequate arterial inflow. Once adequate arterial inflow has been verified, the anterolateral thigh flap may be divided from the lower extremity and transferred to the upper extremity. The anastomosis is then performed in end-to-side fashion using 8-0 nylon. Venous coupling may also be used to expedite the microsurgical procedure (Fig. 18-4).

The donor site is approximated primarily over closed suction drainage when flap width is less than 8 cm, or a skin graft may be used for larger width flaps. For cases involving intramuscular perforator dissection, the muscle edges are reapproximated to preserve maximal quadriceps function.

In certain cases, such as coverage of the hand, a thin flap may be required and primary defatting of the flap can be undertaken. The flap is usually thinned by excising the deep fat consisting of wide, flat fat lobules up to the junction of the superficial fat, made up of smaller, round lobules. Defatting before ligation of the pedicle allows for monitoring of flap perfusion during the thinning process. The flap may be thinned up to 3 mm without compromise to the blood supply. A more conservative approach towards primary flap thinning should be undertaken until adequate experience is gained as flap debulking can be performed safely and easily as a secondary procedure (Fig. 18-5). Innervated flaps can be achieved by including the lateral femoral cutaneous nerve in the proximal portion of the flap. The nerve arises from the deep fascia approximately 10 cm caudad to the anterior superior iliac spine and divides into two or three branches. As previously mentioned, when added bulk is required to cover a complex wound, the anterolateral thigh flap can be harvested as a chimeric flap along with the vastus lateralis or rectus femoris muscles, the tensor fascia lata, or other skin flaps in the thigh based on the supply of lateral femoral circumflex system.

Postoperative Management

Monitoring

Flap monitoring is of paramount importance following microvascular surgery. Postoperative thrombosis of either the artery or vein can be salvaged in greater than 50% of patients if detected early. Intraoperatively, a stitch is placed over the perforator on the skin paddle to facilitate postoperative monitoring with a Doppler probe. Alternatively, an implantable Doppler probe can be placed around the vein or artery intraoperatively to allow for continuous monitoring of the anastomosis. The Doppler signal over the marked skin paddle is checked hourly with the use of a hand held Doppler while the patient remains in the Intensive Care Unit. The highest incidence of postoperative arterial thrombosis is within the first 24 hours, with the incidence of venous thrombosis occurring most frequently within the first three days. Donor site drain output is recorded and drains are kept in place until daily output is less than 30 cc. The patient is allowed to ambulate on postoperative day 2. Postoperative complications include partial to total flap loss, temporary weakness in the lower extremity which usually resolves within the first two weeks, and sensory deficit at the donor site if there has been any stretching or injury to the lateral femoral cutaneous nerve. In cases of venous congestion, exploration of the anastomosis is necessary. If prolonged venous thrombosis is experienced, the use of either TPA or leeches can sometimes be used to salvage the flap.

The use of postoperative anticoagulation is debated. Our patients are given one baby aspirin per day, and are continued on subcutaneous heparin while they are in the hospital. Formal anticoagulation with heparin is discouraged and has been linked to an increase risk of hematoma formation and flap loss. For those situations where arterial inflow has been difficult to establish, dextran at 25 mL/hour may be used in adults.

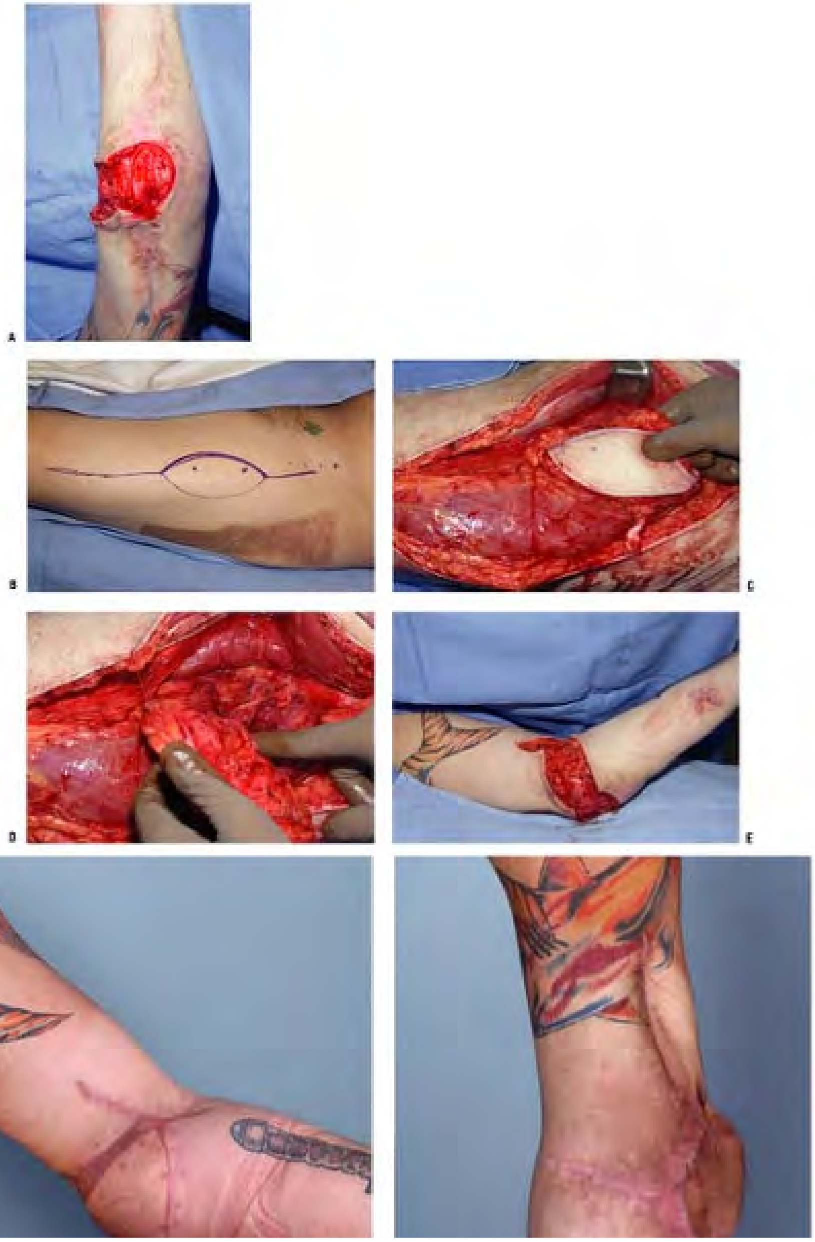

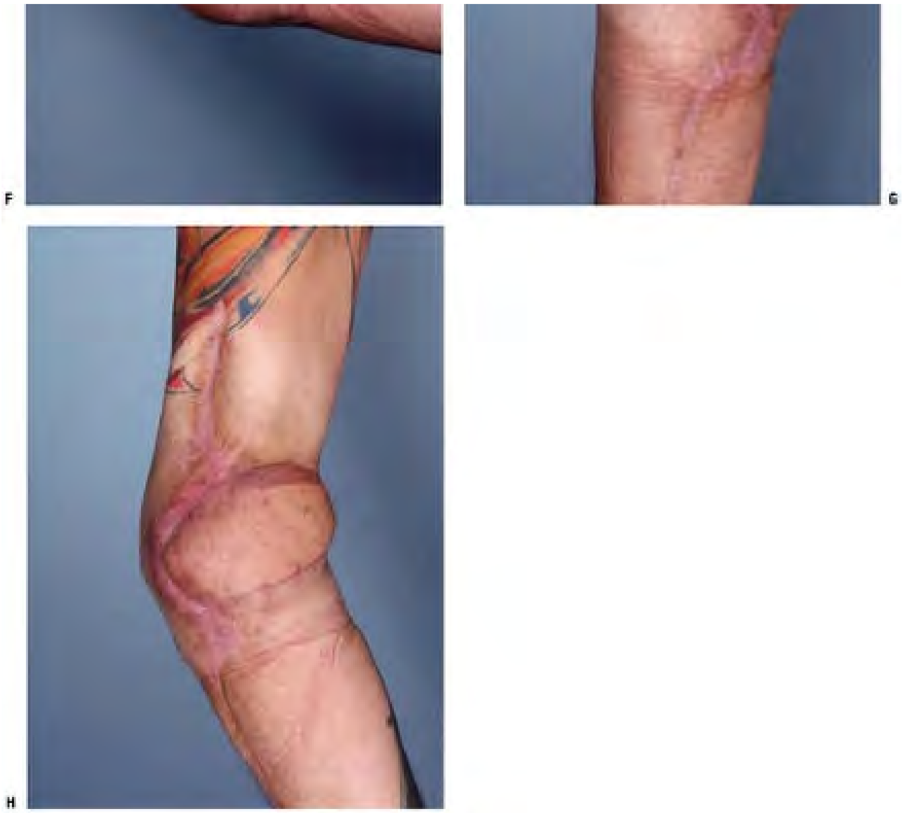

FIGURE 18-4 A: 34 year old male, status post liver transplantation, with a chronic left posterior elbow wound. The multiple posterior scares are evidence of previous attempts to close this defect. B: The skin incision for an ipsilateral ALT flap is marked on the thigh after identification of the underlying perforators. C: The flap is separated from surrounding tissue. D: The lateral femoral circumflex vessels can be seen above the surgeon's index finger entering the flap. E: In preparation for flap transfer the brachial artery is exposed at the level of the elbow. Fâ€“H: The result at 3 months postop with well healed wound, with good contour and no signs of infection.

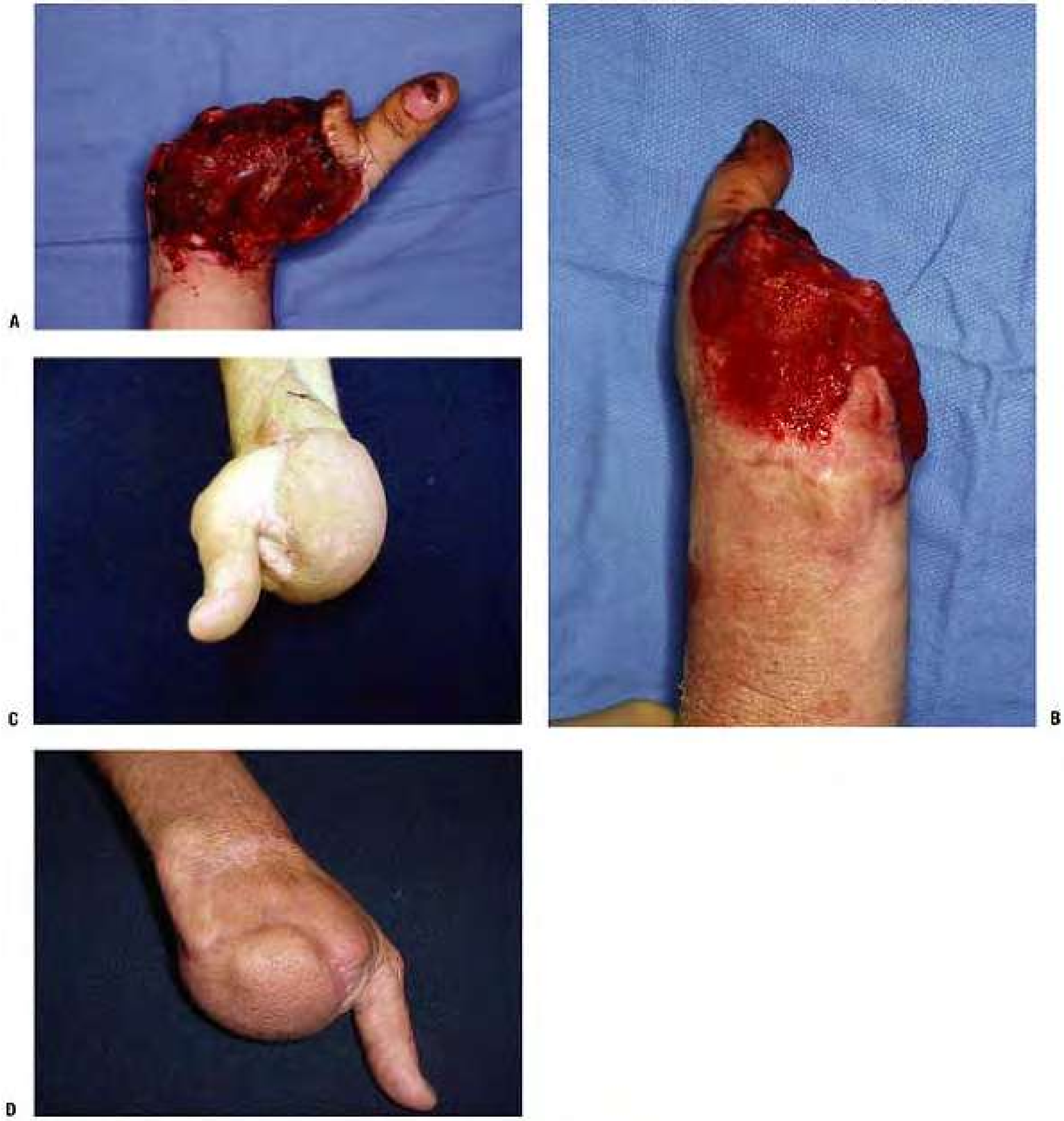

FIGURE 18-5 Aâ€“D: This 49 year old man was involved in a farming accident which resulted in the amputation of the index thru small fingers at the level of the metacarpal bases. Soft tissue coverage was obtained with the use of a thinned ALT flap. Six months following reconstruction, the ALT flap has provided adequate contour and durability allowing patient to return to farming activities. The patient refused further surgery for finger reconstuction (C,D).

Rehabilitation

Underlying injuries usually dictate when motion may begin. We usually wait 5 days for the anastomosis to mature and for the postoperative swelling to subside. At this point if the underlying bone injury is stabilized, early mobilization is initiated. A light ace wrap is applied to the flap beginning on day five to help aid in resolution of post-operative swelling. Re-elevation of the flap for additional bone grafting or tendon grafting can be done as soon as 4 to 6 weeks. Defatting or thinning of the flap is usually delayed for 6 to 9 months after the original surgery.

Complications

The greatest complication of free flap surgery is flap loss due to arterial or venous thrombosis. The best means of preventing these problems is adequate preoperative planning. Repairs within the zone of injury, technical errors including inadequate visualization of vessels, and tension or kinking at the anastomosis site can all lead to flap failure. Ongoing infection can also result in partial flap loss. For the anterolateral thigh flap, twisting of the pedicle can occur easily due to the absence of muscle to keep its proper orientation. Marking the anterior wall of the pedicle and lifting of the flap in the air to allow the pedicle to lie in its natural orientation prior to final inset may prevent pedicle twisting. It is also important to avoid compression of the flap.

Results

Good success has been reported with the use of anterolateral thigh flaps in the upper extremity. Flap failure rates have been noted to be as low as 2%. Additional thinning procedures may be required in patients with continued bulk. Microvascular surgery and free tissue transfer within the upper extremity is facilitated by large recipient vessels, and a relatively shallow operating field.

Conclusion

With proper planning, the anterolateral thigh flap is capable of managing many complex injuries within the upper extremity. It is now often our first choice for coverage of upper extremity defects. The advantages of the ALT flap are: relatively consistent anatomy; ease in dissection; long pedicle length; ability to achieve a thin, pliable flap; ability to provide sensate coverage through the lateral femoral cutaneous nerve; no sacrifice of major artery of the lower limb; and versatility in flap construct which allows one to harvest not only muscle, but also functional muscle such as the rectus femoris. The donor site can be closed primarily.

Recommended Reading

Breidenbach WC. Emergency free tissue transfer for reconstruction of acute upper extremity wounds. *Clin Plast Surg.* 1989;16:505â€“514.

Celik N, Wei FC, et al. Technique and strategy in anteriorlateral thigh perforator flap surgery, based on an analysis of 15 complete and partial failures in 439 cases. *Plast Reconstr Surg.* 2002;109:2211â€“2216.

Chen HC, Tang YB, et al. Reconstruction of the hand and upper limb with free flaps based on musculocutaneous perforators. *Microsurgery.* 2004.

Chen SH, Wei FC, et al. Emergency free-flap transfer for reconstruction of acute complex extremity wounds. *Plast Reconstr Surg.* 1992;89:882â€“888.

Derderian CA, Olivier WA, et al. Microvascular free-tissue transfer for traumatic defects of the upper-extremity: a 25 year experience. *J Reconstr Microsurg.* 2003;19:455â€“462.

Godina M. Early microsurgical reconstruction of complex trauma of the extremities. *Plast Reconstr Surg*. 1986;78:285â€“292.

Hamdi M, Van Landuyt K, et al. A clinical experience with perforator flaps in the coverage of extensive defects of the upper extremity. *Plast Reconstr Surg*. 2004;113: 1175â€“1183.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 19

Groin Flap Coverage of the Hand and Wrist

Jeffrey B. Friedrich

Nicholas B. Vedder

Indications/Contraindications

The pedicled groin flap is an extremely versatile and reliable flap that was initially described in 1972 by MacGregor and Jackson. It enjoys a revered place in the field of reconstructive surgery because it was one of the first axial-pattern flaps to be described and applied in humans. After several decades, many surgical techniques, including pedicled flaps, have been supplanted by new methods. However, despite the advent of microsurgery, the pedicled groin flap continues to be a venerable technique with a variety of applications.

The groin flap has been used most commonly for upper extremity reconstruction, and is indicated for soft tissue defects in which there is destruction or devitalization of soft tissue on the hand, wrist, and distal forearm. The groin flap is not designed to supplant skin-grafting on wounds that will accept a graft; rather, it is to be used when the injury or resection leaves exposed structures that will not accept a skin graft, such as tendon or bone. It has been described for soft tissue reconstruction of a number of traumatic etiologies including crush injuries, avulsions, gunshots, blast wounds, and burns. Additionally, the flap is warranted for soft tissue defects following elective resection such as contractures, hypertrophic scarring, or tumor, and has even been described as soft tissue coverage following recurrent carpal tunnel syndrome.

In the setting of trauma, the groin flap is most often used for delayed primary coverage of upper extremity wounds. Several groups have employed the flap for very early coverage (within 24â€“48 hours from injury); however, the complication rate has been found to be high. Therefore, unless a clean and healthy wound bed can be ensured initially, it is generally prudent to thoroughly debride the wound, observe it for a few days, and then apply the groin flap only after a clean, healthy wound surface is ensured.

As noted above, the groin flap will provide reliable soft tissue coverage to practically any portion of the hand as well as the distal forearm. It is particularly well-suited to the dorsal surface of the hand. The flap can certainly be used to provide coverage for the palmar surface of the hand; however, the flap can be bulky, and does not at all approximate the native, adherent, glabrous skin of the palm. This can leave the palm with a bulky flap that is prone to shearing motion when weight is borne on the palmar surface of the hand, such as seen with pushing, gripping, tool use, or twisting motions. For this reason, it is wise to consider other thinner flaps such as the reversed radial forearm flap or free fasciocutaneous flaps such as the anterolateral thigh flap for most palmar reconstructions.

A brief mention must be made of the groin flap for use as a free tissue transfer. The free groin flap was the first free flap described and can be used for many of the same indications for which other free flaps are currently used. Its reliability is somewhat limited because of the variable arterial pedicle and the short length of the pedicle. Furthermore, even in experienced hands, the free groin flap failure rate is notably higher than it is for other free flaps, thus limiting its utility.

Another chief indication for use of the groin flap is in the setting of thumb reconstruction. Often with total or near-total thumb loss, it is necessary to use free digit transfers such as the great or second toe. Unfortunately, traumatic thumb amputations can leave the patient with insufficient soft tissue of the thenar region. In these cases, the groin flap is an excellent choice to provide a soft-tissue â€œbaseâ€• on to which the digit transfer can be placed at a later date.

The contraindications for the use of the pedicled groin flap are few. In general, the flap should not be implemented in patients who are deemed to be uncooperative, or who would otherwise not be able to tolerate a 2- or 3-week interval in which their affected hand is immobilized and attached to the groin. In the past, it was felt that elderly patients were not good candidates for this flap because of the risk of shoulder and elbow stiffness. However, a recent study has shown that elderly patients are able to tolerate the procedure with no more complications (including upper extremity stiffness) than the average young patient.

As with any flap, active tobacco use is a concern with the groin flap. Specifically, the flap can be subject to marginal necrosis in active smokers. The surgeon must balance the need for complex reconstruction with the potential morbidity that may be incurred by patients who smoke.

Previous groin surgery may or may not be a contraindication to the use of this flap. In general, inguinal hernia repair does not preclude the use of the groin flap, nor do other scars in the general vicinity such as abdominoplasty or Pfannenstiel incisions. However, a previous inguinal lymph node dissection would render this flap tenuous, or outright unsafe. The potential disruption of superficial vessels in the groin is too high to risk use of a flap based on these vessels. Additionally, patients who have had groin lymph node dissections have quite often had external beam radiation to the same region, further rendering flap dissection in this area a dangerous proposition.

Preoperative Planning

When using the groin flap in the acute posttraumatic setting, wound preparation is of paramount importance. The surgeon must perform any hand or forearm repairs that are necessary in the acute phase. This includes fracture reduction and fixation, as well as tendon, vessel, and nerve repair. Wound debridement must also be complete so as to minimize the infection risk once the site is covered with a flap. Wound debridement and cleansing may need to be done several times in the operating room prior to flap coverage, and several days of dressing changes may be necessary to ensure that the wound bed is as healthy as possible. This is especially true with crush and avulsion injuries.

In the posttraumatic setting, timing of coverage with a groin flap has become an important issue. In general, most published reports advocate use of the flap in a delayed-primary fashion, specifically between 48 hours and 7 days following injury. Others have attempted to employ the flap in a more acute setting (24â€“48 hours following injury); however, the consensus seems to be that early timing leads to an unacceptably high rate of infection and dehiscence, and is thus not advocated. In the end, timing of flap implementation should be based on the surgeon's best clinical judgment about the mechanism of injury, and the status of the wound bed.

In the elective setting, the preoperative preparation for a groin flap is somewhat less crucial. The patient must understand that the interval between flap elevation and flap division will entail a moderate amount of disability, and will require a significant amount of diligence in regard to their self-care. Additionally, it is a good idea for the patient to make prior arrangements to have a friend or family member at home who will be able to assist them with various activities of daily living.

Surgery

The groin flap coverage surgery must be done with a general anesthetic. Additionally, it is recommended that a bladder catheter be placed as the operation can be in excess of 3 hours. If the patient is not already being administered intravenous antibiotics for prophylaxis of a traumatic wound, one dose of antibiotics should be administered 30 minutes before skin incision.

Patient Positioning

Once anesthetized, care should be taken with patient positioning. Although an ipsilateral groin flap is most commonly used, a contralateral flap is sometimes acceptable if the ipsilateral side is for some reason unavailable. Regardless of which side is chosen, a small towel roll or â€œbumpâ€• should be placed under the hip of the side that will undergo groin flap dissection to allow extension of the flap dissection back to the flank. The upper extremity to be reconstructed should be placed on a hand table for initial wound preparation. Additionally, one should leave enough room on the side of the operating table to be able to comfortably lay the upper extremity beside the torso during groin flap inset. The other necessary step before initiating the operation is placement of a pneumatic tourniquet on the upper extremity to be reconstructed.

Technique

The operation is begun by preparation of the recipient bed. If the flap is being employed for traumatic wound coverage, a final thorough debridement is done at this time. If it is being done for scarring or other elective soft tissue coverage, the creation of the recipient wound to be covered is done in its entirety before moving to the groin.

In preparation for flap elevation, measurements are taken to determine the required dimensions of the groin flap. This can be done in several ways. One can simply use a ruler to determine the largest dimensions of the wound, or a template can be fashioned. The paper wrapper from surgical glove packaging makes a nice template material and is readily available. This can be pressed onto the wound so that the wound fluid makes an imprint in the paper. Then the template can be made simply by cutting around the outline of the defect. Alternatively, one can use a piece of the elastic wrap used to exsanguinate the hand (Fig. 19-1). The template is then transferred to the groin. In general, the more ulnar aspects of the wound will be covered by the distal end of the flap.

When designing the flap, one must keep in mind the arterial anatomy of the flap. Specifically, the flap is supplied by the superficial circumflex iliac artery and vein (Fig. 19-2). Numerous studies have shown the artery to consistently extend to the level of the anterior superior iliac spine (ASIS) before any significant caliber diminishment. In other words, the flap is an axial pattern flap medial to the ASIS, and a random-pattern flap lateral to this point. The area lateral to the ASIS can be used as part of the flap design; however, its effectiveness is limited by its random nature. Therefore, the operator should generally design a flap that, at the portion lateral to the ASIS, has a width:length ratio between 1:1 to 1:1.5 (although the medial portion needs only be wide enough to include the axial vessels). It is possible to design a flap whose distal extent is more posterior than this limit, but a delay procedure should then be employed for the flap. A delay procedure involves incision, but not elevation

of the desired flap. The incised flap is then left in situ for a week so that the circulation at the distal end of the flap can be augmented during that time. Following that interval, the flap can then be elevated and inset in the upper extremity wound.

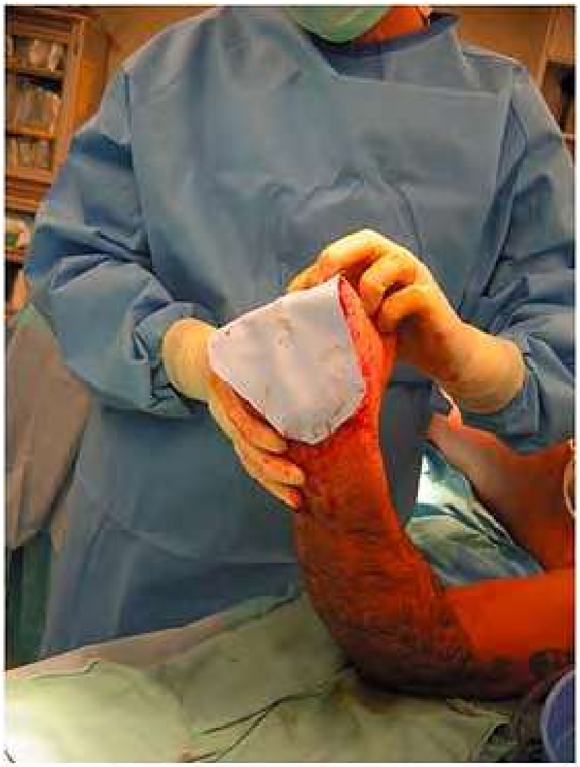

FIGURE 19-1 Use of a piece of the elastic exsanguination wrap to create a template of the defect to be covered with a groin flap.

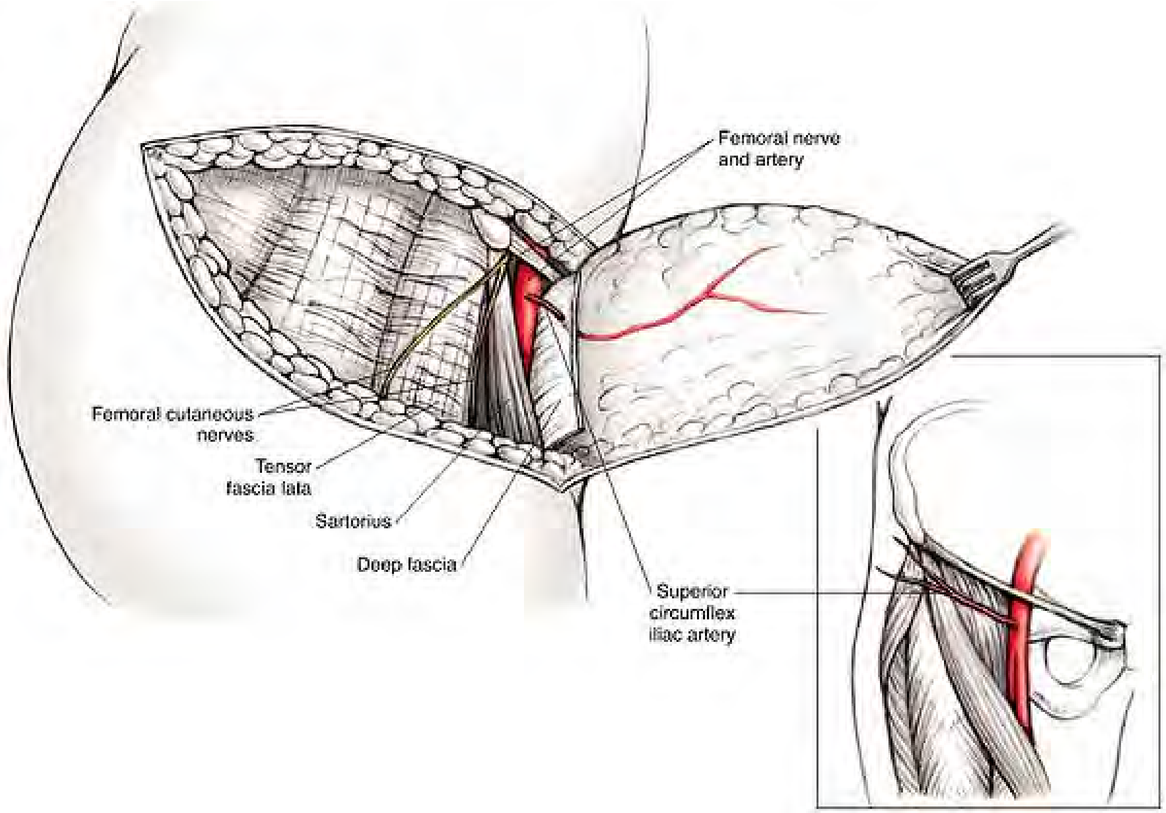
Femoral nerve and artery
Femoral cutaneous nerves
Tensor fascia lata
Sartorius
Deep fascia
Superior circumflex iliac artery

FIGURE 19-2 Detail of groin flap anatomy (right-sided), showing the pedicle (superficial circumflex iliac artery), inclusion of the superficial sartorius fascia with the flap, and the location of the lateral femoral cutaneous nerve branches as they emerge from under the inguinal ligament. (Adapted from Brown EZ Jr, Pederson WC. Skin grafts and skin flaps. In: *Green's operative hand surgery.* 5th ed. Philadelphia: Elsevier/Churchill Livingstone; 2005:1692-1702, with permission.)

When beginning the overall design of the flap, it is best to keep in mind the recommendations of Chuang, who published a set of groin flap design rules based on experience amassed via the use of over 200 groin flaps. The mainstay of these guidelines is the â€œtwo fingerbreadthsâ€• rule (the two fingerbreadths being at the level of the index and long finger distal interphalangeal joint of the operator) (Fig. 19-3). The first step is to estimate the location of the superficial circumflex iliac artery, done by marking a point two fingerbreadths below the junction between the inguinal ligament (which runs along a line between the anterior superior iliac spine and the pubic tubercle) and the femoral artery. Once that artery location is estimated, its true path should be verified by Doppler ultrasound. The upper flap border is marked two fingerbreadths above the inguinal ligament (parallel to the artery), while the lower flap border is a line two fingerbreadths below and parallel to the artery. Also, as will be further discussed following, the superficial branch of the superficial circumflex iliac artery (SCIA) emerges from the deep fascia two fingerbreadths medial to the ASIS, then continues laterally in a subcutaneous plane.

It should be noted that if using the two fingerbreadths rule for flap design, a relatively slender flap will result medially. In reality, when the flap is centered over the SCIA, the cephalad-caudad width of the flap lateral to the ASIS can be quite substantial. However, most authors agree that flap widths up to 10 cm enable primary closure of the groin wound site, while those that are wider stand a good

chance of requiring closure with a split-thickness skin graft because of inability to achieve primary closure. It is usually possible to determine how wide a flap can safely be harvested and closed primarily by pinching the skin and subcutaneous tissue together. Flexing the operating table at the hip to relieve some of the tension can facilitate closure.

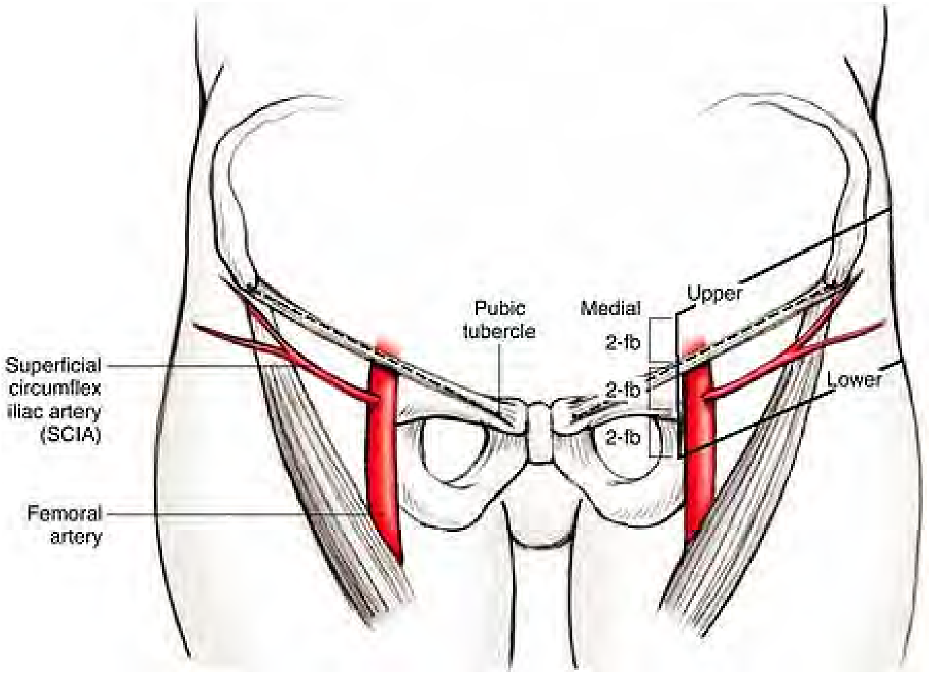

FIGURE 19-3 Drawing of the application of the two-fingerbreadth rule. The *31* on the left of the patient is the location of the inguinal ligament. The *31* immediately below that is the location of the superficial circumflex iliac artery. Note that the upper border of the flap is two fingerbreadths above the inguinal ligament, and the lower border is two fingerbreadths below the vascular pedicle (SCIA). (Adapted from Chuang DC, Colony LH, Chen HC, Wei FC. Groin flap design and versatility. *31* 1989;84(1):100â€“107, with permission.)

Once the marks are drawn, the incisions of the flap are made. The incision is carried down to the level of the external oblique fascia on the abdomen, and the muscle fascia on the leg. Then the flap dissection is begun at the lateral end in the plane between the subcutaneous fat and the muscle fascia of the thigh (below the inguinal ligament); and between the subcutaneous fat and the external oblique muscle fascia (above the inguinal ligament). The flap is most easily and safely raised with curved Metzenbaum scissors, although low-powered electrocautery can be used as well. The dissection proceeds in this fashion until reaching the lateral border of the sartorius muscle. On reaching the sartorius, the fascia over this muscle is incised at its lateral border. The sartorius fascia is then elevated from the muscle. Above the inguinal ligament, the dissection plane remains immediately superficial to the external oblique fascia. At the medial border of the sartorius muscle, the SCIA divides into a superficial and a deep branch. By elevating the sartorius fascia, the superficial branch (which supplies the lateral extent of the groin flap) is preserved with the flap. When the medial sartorius fascia is reached, and the deep branch of the SCIA encountered, it is divided and ligated with either vascular clips or sutures. This deep branch of the SCIA is not necessary for flap survival, and it will only serve to tether the flap if more flap length is desired (Fig. 19-4).

During the flap dissection, one must remain cognizant of the lateral femoral cutaneous nerve and its course in the vicinity of the flap. The nerve emerges from under the inguinal ligament at the medial edge of the sartorius muscle (see Fig. 19-2). The nerve branches run deep to the SCIA; therefore, care should be taken to preserve the nerve branches in this area. If nerve branches are found in the sartorius muscle fascia, efforts must be made to dissect free and preserve them. Despite attempts to preserve the nerve branches, patients should be told before the procedure that nerve damage (both transient and permanent) is possible and can lead to numbness of the lateral thigh.

The endpoint of flap dissection is generally determined by the acquisition of enough flap length to be easily inset in the defect of the hand or forearm. Some references caution against proceeding more medially than the lateral border of the sartorius muscle; however, we have found that a longer flap can be raised if proper precautions are taken. That said, the absolute anatomic endpoint of flap dissection is at the lateral border of the femoral artery.

While this chapter chiefly concentrates on the use of the pedicled groin flap, a brief mention of the free groin flap will be made. The dissection of the flap proceeds in an identical manner to that described previously. However, one of the over-arching goals of free flap harvest is to obtain donor

vessels that are as long as safely possible. With the free groin flap, this involves dividing the SCIA very close to the femoral artery, and dividing the donor vein very close to the femoral vein. This must be done carefully as vascular clips or sutures placed too close to the femoral vessels can serve to narrow the lumen of the vessel. When harvesting the groin flap for free tissue transfer, one anatomic caveat must be kept in mind: in a minority of persons (estimated at 30%), the SCIA emerges from a common vessel that bifurcates into the SCIA and the superficial inferior epigastric artery (SEIA). When this is found to be the case, the common trunk is dissected back to the femoral artery and used as the arterial donor. Both the SCIA and SEIA will reliably supply the groin skin; so when they emerge separately, the larger of the two vessels is chosen as the arterial donor. Usually, the larger vessel is the SCIA.

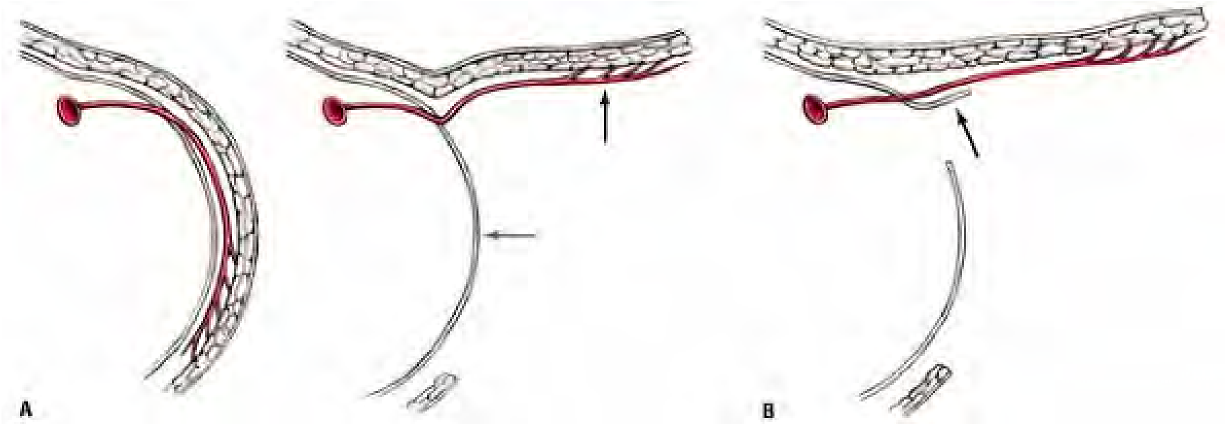

FIGURE 19-4 A: Schematic representation of vascular kinking that is possible if the sartorius fascia is not elevated with the groin flap. The illustration shows an axial view of the lower extremity. The *31* points to the superficial circumflex iliac artery (SCIA); the Î points to the fascia of the sartorius muscle. B: Schematic representation of the relatively straight course of the superficial branch of the SCIA following elevation of the sartorius fascia with the groin flap. The *31* points to the now-released sartorius fascia. (Adapted from Brown EZ Jr, Pederson WC. Skin grafts and skin flaps. In: *Green's operative hand surgery.* 5th ed. Philadelphia: Elsevier/Churchill Livingstone; 2005:1692â€“1702, with permission.)

Several steps must be taken before inset of the flap into the upper extremity. First, one must close the donor defect, because closure will be much more difficult if it is done after flap inset. This closure should optimally be a three-layered one, with 0-0 or 2-0 braided absorbable suture in the superficial fascia (Scarpa's fascia on the abdominal wall to superficial fascia on the lower extremity), 2-0 or 3-0 braided absorbable suture in the dermis, and 3-0 or 4-0 nonabsorbable monofilament or staples in the skin. This closure is done over a suction drain, usually brought out through a separate stab incision. The closure is facilitated by flexion of the operating table at the hip, and use of towel clips or staples to temporarily approximate the skin while the sutures are being placed (Fig. 19-5). It is also important for the patient's bed to remain flexed at the hip in the immediate postoperative period to minimize tension on the suture line.

Next, the flap base should be â€œtubedâ€• before inset. Tubing of the flap diminishes the amount of exposed flap undersurface, thereby helping with postoperative wound care. The tubing is done simply by suturing the upper and lower sides of the flap together such that it forms a cylinder (Fig. 19-6). The closure of the tube can be done with a single layer of nonabsorbable monofilament sutures. One must be cautious about tubing the flap, as this can cause vascular compromise. A simple way to deal with a congested flap is to release a few sutures at the base (since this is the thickest part of the flap). In a seminal article on groin flap design, Schlenker illustrates another method to increase the circumference of the tube while simultaneously positioning the flap for inset depending on whether the hand defect is dorsal or volar. This is done by incising either the cephalad or caudad side of the flap longer than the other, then performing either a clockwise or counterclockwise closure of the donor defect. This then positions the portion of the flap that will be inset into a more optimal orientation for coverage (Fig. 19-7).

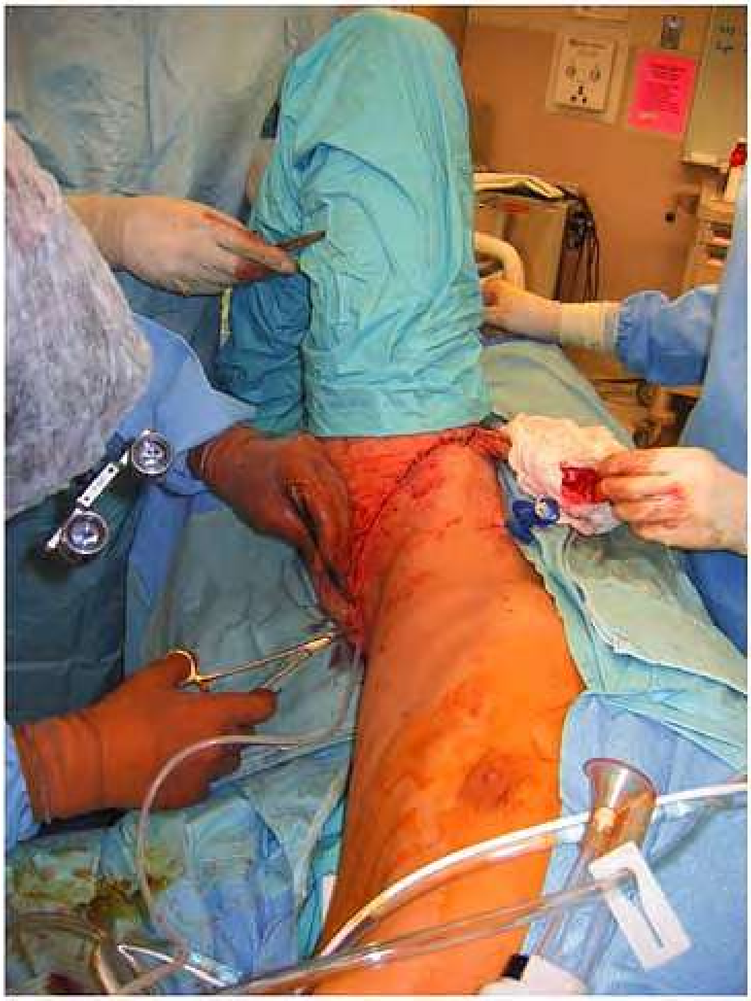

FIGURE 19-5 Flexion of the hip to facilitate groin flap donor site closure.

The other task to be done before flap inset is to thin the distal part of the flap that was originally lateral to the ASIS. This is most easily performed with curved Mayo or Metzenbaum scissors, trimming the fat down to the subdermal level. This will facilitate flap inset, and will reduce the amount of defatting that needs to be done in the secondary setting.

Once elevated, the inset of the groin flap is relatively simple. The affected hand is moved into the groin, and the flap is then inset into the recipient site, usually over a closed-suction drain. This is typically performed with a two-layered closure, specifically 3-0 or 4-0 braided absorbable suture in the dermis, and 3-0 or 4-0 nonabsorbable monofilament suture in the skin (Fig. 19-8).

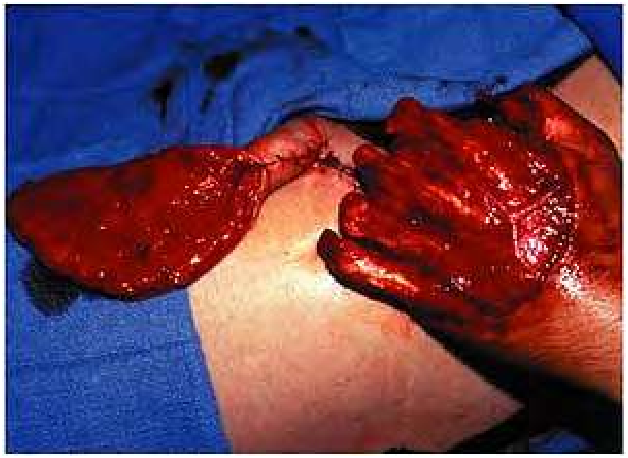

FIGURE 19-6 Groin flap following â€œtubingâ€• of the proximal portion of the flap.

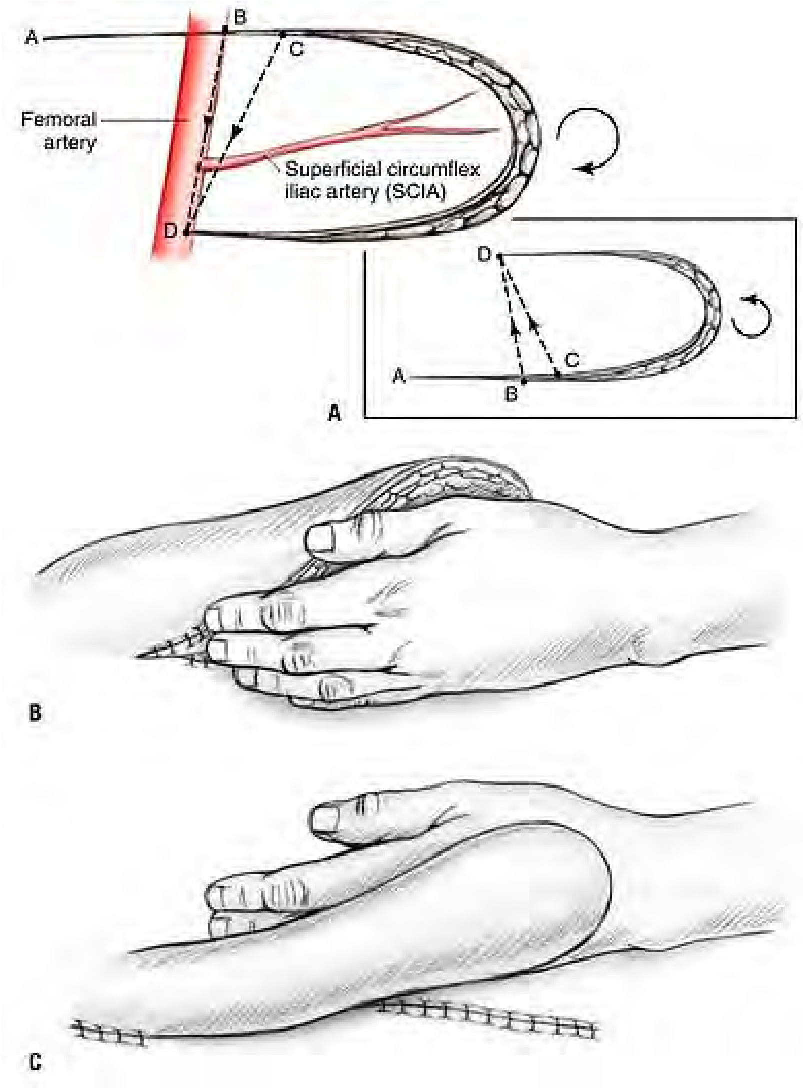

FIGURE 19-7 A: Schlenker's modification of the groin flap incisions. By making the top incision longer, and closing points B and C to point D, clockwise rotation is facilitated. Similarly, by making the lower incision longer, counterclockwise rotation is facilitated. B: By creating a flap that is rotated clockwise, and therefore facing downward, closure of a palmar defect is made easier. C: Similarly, counterclockwise rotation of the flap causes the raw surface to face upward, allowing easier closure of a dorsal hand defect. (Adapted from Schlenker JD. Important considerations in the design and construction of groin flaps. *31* 1980;5(5):353â€“357, with permission.)

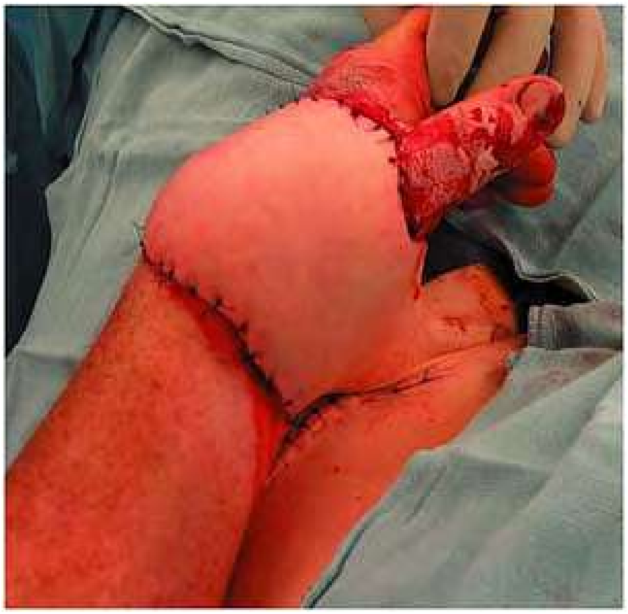

FIGURE 19-8 Completion of groin flap inset on the hand.

Care must be taken with postoperative dressings and positioning. The wounds on the hand can simply be covered with antibiotic ointment. We find that it is best to cover the small exposed portions of the flap (at the base) with antibiotic (bismuth)-impregnated gauze. The closed groin wound can be dressed with a gauze dressing or covered with antibiotic ointment. Cast padding or soft towels should be placed in the affected axilla and under the hand as these areas will become moist with the arm at the side. A pillow should be placed underneath the elbow to give the arm support, and to ensure that the flap itself is just the slightest bit extended anterior and lateral to the body. It is important to avoid kinking of the flap since this will easily cause venous congestion. Finally, the donor-side leg should have a pillow placed under the knee to relieve tension on the groin flap closure. Once the patient is awake, it is important to educate them in positioning of the flap to avoid kinking.

Flap Division and Final Inset

In general, the pedicle of the groin flap is divided between 2 and 3 weeks after flap elevation. Wei's group and others have done numerous studies using ischemic preconditioning of the flap to allow flap division at an earlier time (7â€“10 days). However, we have found that, with proper patient selection, the groin flap will be tolerated for 3 weeks just as easily as it would be tolerated for 1 week. Furthermore, we have not found any complications associated with the longer attachment times.

Flap division proceeds as follows: one should test the vascularity of the flap by firmly occluding the tubed pedicle with thumb and forefinger or by using a penrose drain clamped around the pedicle, then observing capillary refill of the flap. If there is *any* question about the perfusion of the flap, the flap division should be delayed. A hand table should be set up before the start of the case. An arm tourniquet is not necessary for this procedure, and in fact, it is discouraged because the vascularity of the flap needs to be assessed during this operation. The arm is prepared from at least the elbow down, and the entire groin area is prepared simultaneously. The flap is marked where it is thought that it will cover the remaining wound, taking care to make this marking generous. The flap is then incised with a scalpel and carried through with electrocautery. The arm is then placed on the hand table. Sometimes, there is a residual flap â€œstump,â€• which is excised to the point that the remaining groin wound can be easily closed without a large dog-ear deformity. If the stump of the SCIA is large, this should be ligated with a suture or vascular clips. The remaining groin wound is closed in a fashion similar to that of the original groin wound closure.

Finally, the remainder of the groin flap is inset to the hand. Some very judicious defatting of the open end of the flap can be undertaken at this time. This can be done with scissors, but it should not be overly aggressive. Both the donor and recipient sites should be well debrided and cleansed before closure. The closure should be loose to allow drainage, as these will be contaminated wounds. The wound is sutured with 4-0 nylon in the skin (Fig. 19-9). A soft dressing should then be applied to the

hand, usually with moist gauze in the layer next to the skin as this will encourage drainage from these contaminated wounds. We find it helpful to take advantage of the general anesthesia by manipulating the patient's elbow and shoulder before his or her awakening.

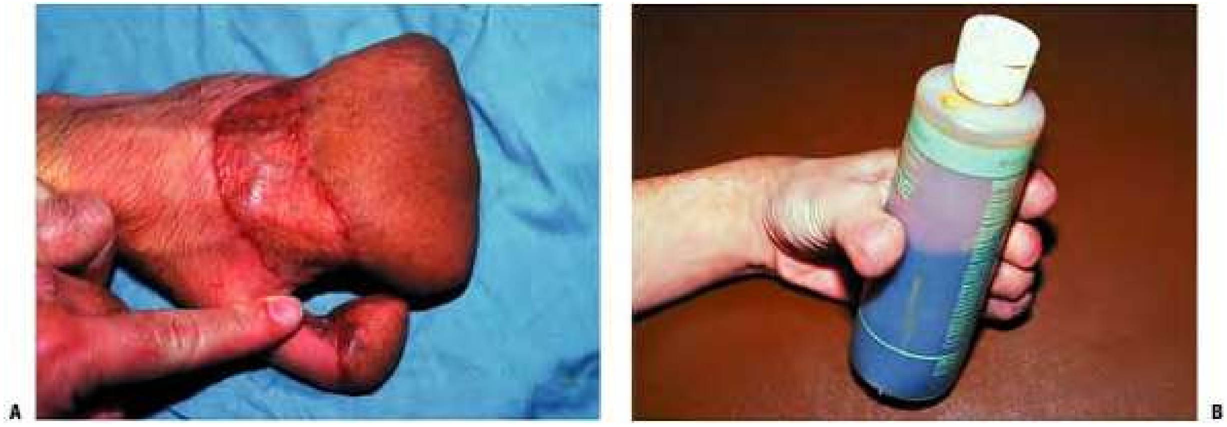

FIGURE 19-9 A: Photograph of same patient in Figure 19-6 after division of groin flap pedicle. The hand has a â€œmittenâ€• appearance after covering the finger stumps with the groin flap. B: Long-term results seen in patient from Figures 19-5 and 19-8B. The groin flap has been divided into individual flaps for each finger, and they have each been further contoured.

Postoperative Management

After the first stage of the groin flap procedure, most patients will need a significant amount of assistance from physical and/or occupational/hand therapists. The involvement of the therapist is essential to help with the transition to home life for the interval in which the hand is attached to the groin. For the purposes of practicality and cost of care, it is beneficial to send the patient home a few days after the initial stage of the procedure, rather than keeping the patient in an inpatient setting for 2 to 3 weeks. The therapist will assist the patient with normal self-care activities such as dressing (including any modifications of clothing made necessary by the attachment of the hand to the groin), hygiene, and other daily activities.

The chief purpose of the hand therapist is to ensure joint mobility to the degree that it is possible during the pedicled stage of the flap. This includes range of motion exercises with the involved shoulder and elbow, as well as the affected hand. In our experience, the groin flap is most often used to cover hand and wrist defects. Therefore, if the patientsâ€™ other hand injuries permit, we encourage them to work on range of motion exercises with the hand, most commonly in the interphalangeal and metacarpophalangeal joints.

Postoperative wound care is relatively simple after the first stage of the surgery. The chief aim is to keep any flap surfaces that are exposed (i.e., any raw surfaces that are not covered by tubing of the flap) moist and clean. The application of antibiotic (bismuth)-impregnated gauze or antibiotic ointment serves this purpose quite well. Patients are encouraged to wash the wounds daily with soap and water and even to shower. Suture removal depends on the surgeon's preference: they can be removed in 10 to 14 days, or this can be done at the time of flap division 2 to 3 weeks after the initial procedure.

After the conclusion of the flap division procedure 2 to 3 weeks after flap elevation, physiotherapy continues to be an important part of the patientsâ€™ care. Range of motion exercises should be continued, especially in the affected elbow and shoulder. These joints can become quite stiff between the two procedures; however, this stiffness is most often transient if physical therapy is maintained. Finally, aggressive rehabilitation of the injured hand is a long-term endeavor.

Much later during the patient's care, debulking, revision, and functional improvement of the reconstruction can be taken into consideration. In many patients, the groin flap can be relatively bulky, potentially giving the reconstructed hand a â€œloaf of breadâ€• appearance. Once the flap has completely healed, this problem is relatively easily treated by partial opening of the flap followed by direct lipectomy in the subcutaneous plane, or by using suction lipectomy with a small (3 mm) suction cannula.

Complications

In the intraoperative setting, a potential complication is construction of a flap that is too wide in the cephalocaudal dimension, thereby preventing primary closure of the donor site. This is not a grave complication, but it does require closure of the donor site with a split-thickness skin graft, which obviously creates another donor site. Additionally, the appearance of the groin wound covered by a skin graft is less desirable than that of a groin wound that has been closed primarily.

The most dreaded complication of the use of the groin flap is flap loss, either partial or total. Fortunately, the pedicled groin flap is quite hardy, making flap loss unusual if care is taken with flap design and tubing. Most series have found flap loss rates in pedicled groin flaps around 1%, and these are most often partial, rather than total flap losses. Not surprisingly, the rate of flap loss when using the groin flap as a free tissue transfer is higher than that of pedicled groin flaps. This can be attributed to the relatively increased difficulty of free groin flap dissection, and to the nature of free flaps in general. Cooper analyzed a series of 130 patients who had undergone free groin flaps for a variety of indications. The series had seven partial and nine total flap losses, and a total flap failure rate of 8.5%. This is slightly higher than most published series of free tissue transfers, where the general free flap failure rate is around 5%. Cooper's series corroborates these numbers in that they experienced an overall free tissue transfer failure rate (using flaps other than the free groin flap) of 4.2%, which is less than half of the failure rate of the free groin flap in their hands.

A complication that is of low morbidity, but can be somewhat distressing to the patient is lateral thigh numbness due to lateral femoral cutaneous nerve injury. Many of these injuries are transient, and sensation will return within several weeks. More distressing to the patient is permanent numbness of the lateral thigh. The true incidence of permanent nerve injury is difficult to gauge due to the variable amount reported in the literature (1%â€“50%). This can be prevented by careful dissection around the area of the ASIS during flap elevation, taking care to identify and protect the adjacent cutaneous nerves.

Joint stiffness of the elbow and shoulder, though usually transient, is almost a guarantee with the use of the pedicled groin flap. Joint stiffness was thought to be more pronounced and problematic in elderly (over 55 years) patients; however, Buchman published a small series of elderly patients who underwent groin flaps in which they contend that the severity of joint stiffness in this population is not any worse than that of a younger population. As mentioned in the previous section, aggressive physiotherapy is crucial to ensure that any joint stiffness is merely transient.

As with any procedure that involves a period of immobilization, deep venous thrombosis is a concern when dealing with patients who have undergone a groin flap reconstruction. Because groin flap patients are at least initially confined to bed, it is recommended that these patients all be administered low-dose subcutaneous heparin or low-molecular-weight heparin as prophylaxis against deep venous thromboses.

Finally, there are a number of complications that are possible with just about every flap surgery, including the groin flap. Some groups have reported significant seromas in the groin donor site, but these are usually adequately treated with needle aspiration or the opening of a small part of the donor incision. Infection is certainly a possibility, especially with raw wound surfaces for a number of days. However, these are surprisingly uncommon, and are usually sufficiently treated with intravenous or oral antibiotics, along with limited wound opening.

Recommended Reading

Arner M, Moller K. Morbidity of the pedicled groin flap. A retrospective study of 44 cases. *Scand J Plast Reconstr Surg Hand Surg.* 1994;28(2):143â€“146.

Barillo DJ, Arabitg R, Cancio LC, et al. Distant pedicle flaps for soft tissue coverage of severely burned hands: an old idea revisited. *Burns.* 2001;27(6):613â€“619.

Buchman SJ, Eglseder WA Jr, Robertson BC. Pedicled groin flaps for upper-extremity reconstruction in the elderly: a report of 4 cases. *Arch Phys Med Rehabil.* 2002;83(6):850â€“854.

Cheng MH, Chen HC, Wei FC, et al. Combined ischemic preconditioning and laser Doppler measurement for early division of pedicled groin flap. *J Trauma.* 1999;47(1):89â€“95.

Cheng MH, Chen HC, Wei FC, et al. Devices for ischemic preconditioning of the pedicled groin flap. *J Trauma.* 2000;48(3):552â€“557.

Chow JA, Bilos ZJ, Hui P, et al. The groin flap in reparative surgery of the hand. *Plast Reconstr Surg.* 1986;77(3):421â€“426.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 20

Cross-Finger Flaps for Digital Soft Tissue Reconstruction

Bassem T. Elhassan

Alexander Y. Shin

Injuries of the fingertips are very common occurrences secondary to both domestic and industrial accidents and are among the most common hand injuries encountered in the emergency room and the hand practice. To function properly, the fingertip requires adequate padding with durable, sensate skin coverage. As such, any reconstructive procedure of the fingertip must include the provision of durable skin coverage, preservation of adequate sensation, maintenance of finger length, prevention of proximal joint stiffness, and restoration of cosmesis when possible.

Several techniques have been described to address tissue loss of the fingertips. These techniques include direct wound approximation, healing by secondary intention (i.e., allowing the wound to granulate), split and full thickness skin grafting, cross-finger and reverse cross (deepithelialized) finger flaps, adipofascial turnover flaps, revision amputation, volar V-Y advancement flap, Kutler's bilateral V-Y advancement, vascularized island flaps (homodigital flaps), and partial toe transfers (1,2,3,4,5,6). Each technique has its own indications and contraindications depending on the size of soft tissue loss and the concomitant injuries (exposed bone, tendon, nerve, or vessels).

Cross-finger flaps are local flaps that can reliably provide padding and durable cover for fingertip injuries and have been in use for over 50 years (7,8). The anatomy, indications and contraindications, technique, and results of treatment of the cross-finger flap and the reverse (deepithelized) cross-finger flap for coverage of fingertip injuries will be described.

Anatomy

Cross-Finger Flap

The cross-finger flap is a random local (regional) flap that is usually raised from the dorsum of the donor finger at the level of the proximal or middle phalanx. The blood supply of this flap is from the many small vessels of the subdermal and subcutaneous plexus. These plexi are fed by two major blood supplies. The first is the dorsal digital arteries that originate from the dorsal metacarpal arteries, which in turn originate from the dorsal carpal arterial arch. These branches mostly feed the plexus covering the skin and subcutaneous tissues of the dorsum of the proximal phalanx. The second blood supply is the proper digital arteries, which feed the plexus of vessels covering the skin and subcutaneous of the dorsum of the middle and distal phalanges.

When used as an innervated flap, the cross-finger flap derives its innervation from the sensory innervation of the dorsum of the hand (9,10). The superficial branch of the radial nerve and the dorsal sensory branch of the ulnar nerve innervate the dorsum of the hand up to the level of the proximal one-third of the dorsal skin of the proximal phalanx. Distal to this point, most of the innervation comes from the proper digital nerves through dorsal and distal branches (Figs. 20-1 and 20-2).

Reverse Cross-Finger Flap

The flap used for reverse cross-finger flap consists of the same tissues used for the standard cross-finger flap but without the covering skin, and thus has also been called a deepithelialized cross-finger flap. The subdermal plexus becomes the main blood supply to this flap, fed by branches of the dorsal digital arteries and the proper digital arteries as previously described.

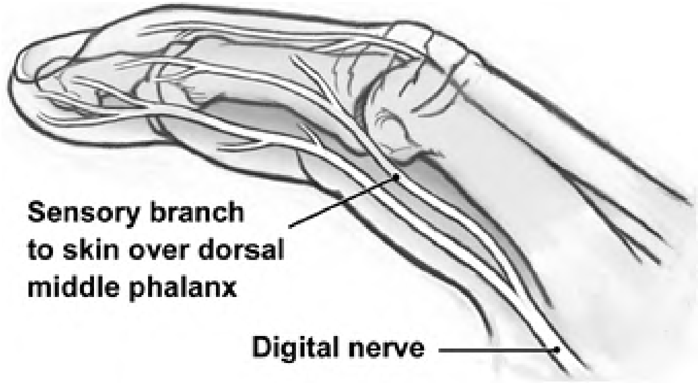
Sensory branch
to skin over dorsal
middle phalanx
Digital nerve

FIGURE 20-1 Innervation to the dorsum of the middle and distal phalanges from the proper digital nerves through dorsal and distal branches. (Reproduced with permission of the Mayo Foundation, 2005.)

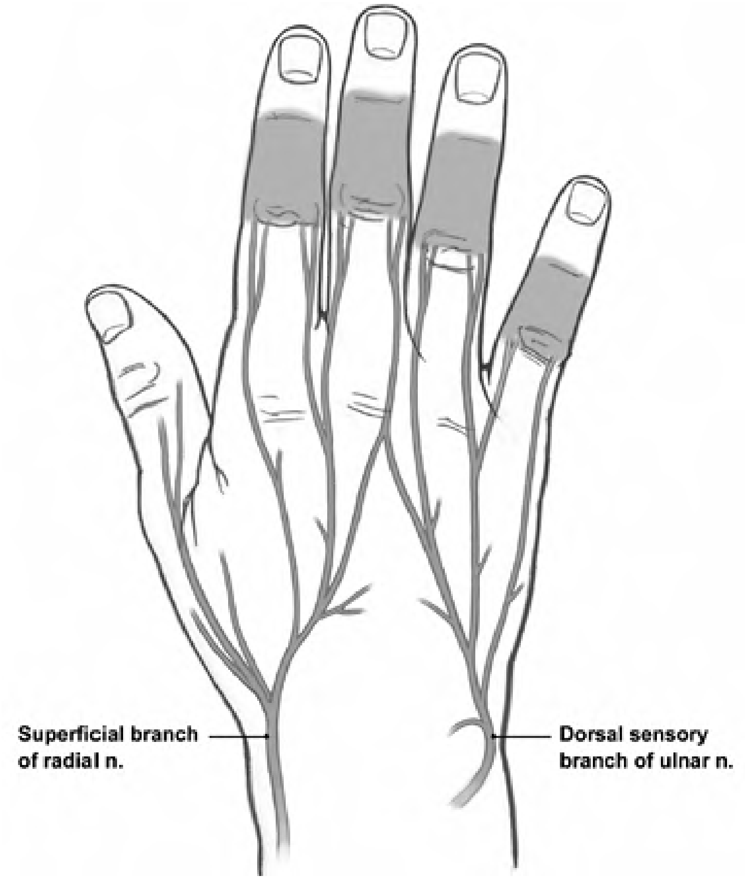

FIGURE 20-2 Innervation to the dorsum of the hand and dorsum of the proximal fingers from branches of the radial and ulnar nerves. (Reproduced with permission of the Mayo Foundation, 2005.)

Indications/Contraindications

Cross-Finger Flap

A cross-finger flap can be used any age group. In the pediatric patient, its use is reported to be successful as early as the first year of life (11). Some authors have recommended alternative tissue coverage in patients older than 50 years because of the risk of development of persistent postoperative stiffness at the proximal interphalangeal (IP) joint (12,13).

The main indication for this flap coverage is a volar fingertip wound with a major loss of skin and subcutaneous tissue with exposed bones and/or tendons. It can also be used for volar defects of the middle phalanx.

As a random flap, the base is longer than the limbs of the flap. Thus, the entire dorsal skin of the finger can be raised and used. However, if the flap includes skin from the dorsal creases of the IP joints, the resultant dorsal scar may lead to stiffness and possible flexion contracture. It is recommended that the flap be raised between the IP creases. The flap's size therefore varies from patient to patient and depends on the size of the finger.

Contraindications include (14,15,16):

- Multiple injuries to the hand involving the potential donor finger
- Vasospastic conditions such as Raynaud disease and Berger disease
- Preexisting disabling problem such as Dupuytren contracture

Advanced age, diabetes mellitus, and rheumatoid arthritis are considered relative contraindications.

Reverse Cross-Finger Flap

A reverse cross-finger flap is indicated in adults with defects on the dorsum of the finger with major loss of skin and subcutaneous tissue and exposed bones and tendons or a nail bed defect that cannot be covered by other techniques (17).

The best donor site is the dorsum of the middle and proximal phalanges. The skin over the distal interphalangeal (DIP) and proximal interphalangeal (PIP) joints should be avoided because the subcutaneous tissue over these areas is thin and there is higher risk of developing contractures.

The size and contraindications of this flap are similar to those of the cross-finger flap.

Surgery

Cross-Finger Flap

The donor finger is adjacent to the injured finger. Typically, the donor finger is chosen based on how the injured finger approximates to the middle phalanx of the donor finger. A majority of these flaps can be performed as outpatient surgery under regional anesthesia. Preoperative radiographs of the injured hand should be evaluated for concomitant injuries, which should be treated prior to commencement of the cross-finger flap procedure. Prophylactic antibiotic is given per the discretion of the surgeon.

Under brachial or digital tourniquet control, the injured digit is debrided and irrigated and the skin edges freshened (Fig. 20-3).

The donor and recipient fingers are placed in a position of comfort, before planning the location and the shape of the donor area. To ensure adequate soft tissue coverage, the size of the defect can be outlined on a piece of glove paper allowing 20% excess for the flap pedicle (Fig. 20-4). The outlined site is drawn on the dorsum of the adjacent donor digit. The three sides of the flap are raised as a full thickness graft on the donor finger down to but not including the paratenon (Fig. 20-5A). It is essential that the paratenon layer not be violated, as the full thickness skin graft, which is placed over the donor area, may not adhere. Meticulous hemostasis is obtained. If mobility of the raised flap is limited, or if there is kinking of the flap when the recipient site is covered, further dissection of the flap is performed by incising carefully Cleland's ligament, while protecting the neurovascular bundle that lies volar. Once elevated, the tourniquet is deflated and the vascularity of the flap and the paratenon is evaluated. The donor defect is covered with a full thickness skin graft obtained form the anticubital fossa, upper brachium, or groin (Fig. 20-5B). The flap is sutured in place on the finger defect with nonabsorbable 5-0 suture in adults or absorbable suture in children (Figs. 20-6 through 20-9). A compressive dressing of the surgeon's choice is applied over the skin graft, followed by a bulky hand dressing with appropriate splints. The wound is kept elevated, with encouragement of movement of the uninjured digits. The dressing can be lightened at 1 week postoperatively, whereas the sutures are removed between 10 and 14 days. Between 2 and 3 weeks postoperatively, the patient is returned to the operating room, where the flap is divided, contoured, and inset (Figs. 20-10 through 20-12).

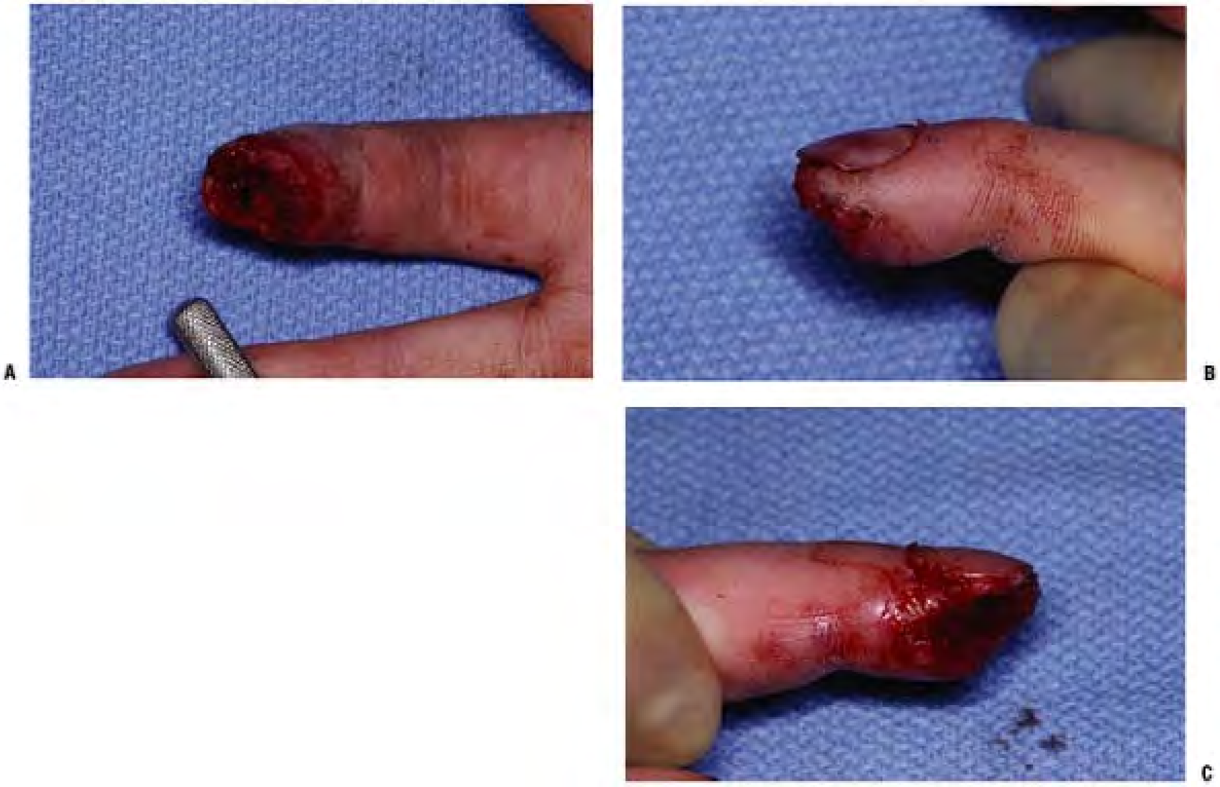

FIGURE 20-3 A-C: Front and side views of an oblique volar fingertip amputation.

Reverse Cross-Finger Flap

The reverse cross-finger flap is useful for dorsal tissue loss, unlike the standard cross-finger flap, which is indicated for volar wounds. The preparations for this flap are similar to those of the standard cross-finger flap. The differences include the hinge of the flap and the donor tissue. A thin full thickness skin flap is raised from the dorsum of the donor finger at the level of the middle phalanx, preserving the subcutaneous layer.

The hinge of the skin flap is opposite to the primary defect. This is in contradistinction to the standard cross-finger flap, where the hinge is adjacent to the site of the defect (Figs. 20-13 and 20-14). Attention should be paid to keep the dissection at the level of the dermis, below the hair follicles, and above the layer of subcutaneous veins (Fig. 20-15). The subcutaneous tissue is raised, keeping the paratenon intact, with the hinge of the subcutaneous tissue adjacent to the defect. The subcutaneous flap is inset into the defect and the skin is placed back over the donor site (Fig. 20-16). A thin full thickness skin graft is harvested and set on the flap (Fig. 20-17). Division of the flap occurs between 2 and 3 weeks postoperatively (Fig. 20-18). A clinical example of a reverse cross-finger flap is illustrated in Figure 20-19.

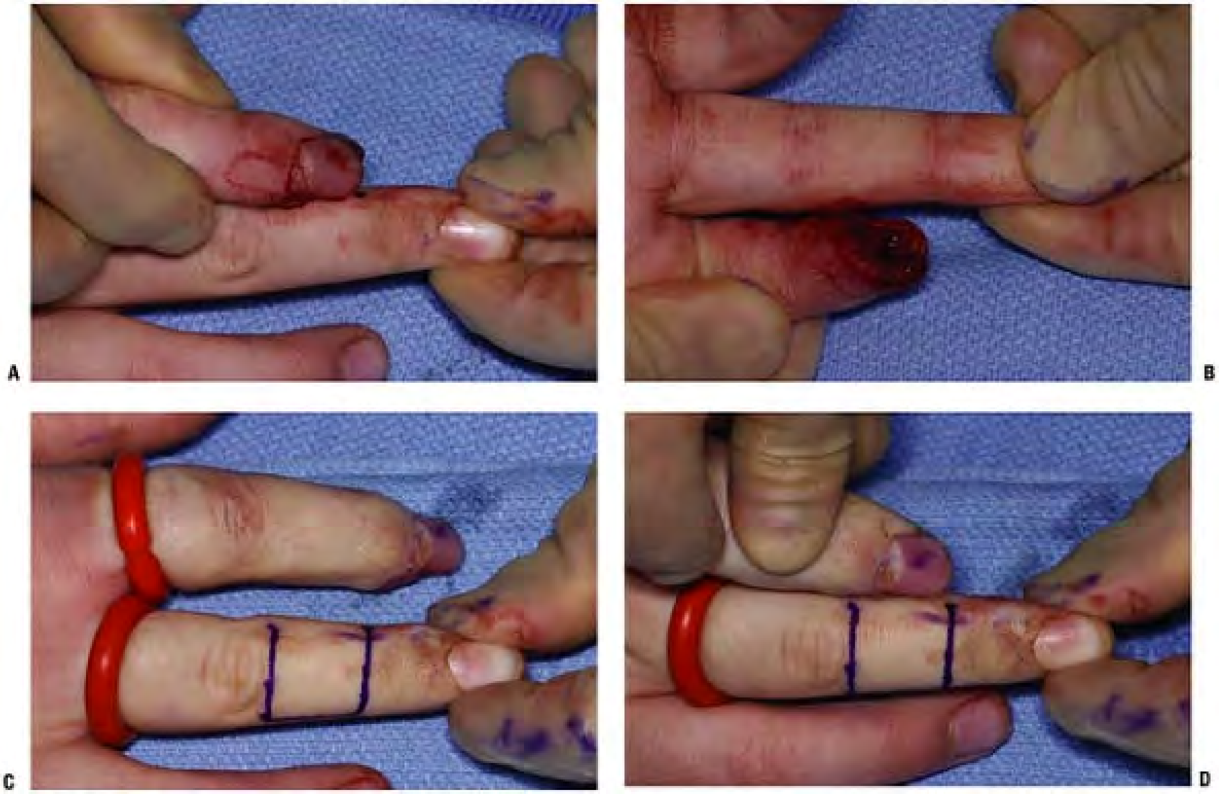

FIGURE 20-4 A-D: Intraoperative pictures showing the steps of preparation for cross-finger flap coverage of a ring fingertip defect. The injured finger is placed at the level of the planned flap from the dorsum of the middle finger at the level of the middle phalanx. The flap should be 20% larger than the actual defect to minimize tension at the time of closure.

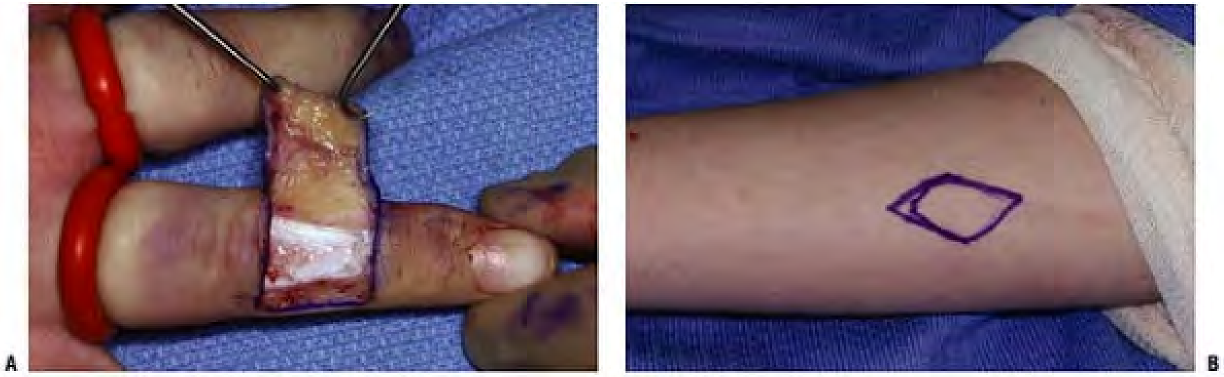

FIGURE 20-5 A: Elevation of the cross-finger flap, keeping the paratenon intact. B: Drawing the site for the full-thickness skin graft harvesting from the forearm of the involved finger. (Reproduced with permission of the Mayo Foundation, 2005.)

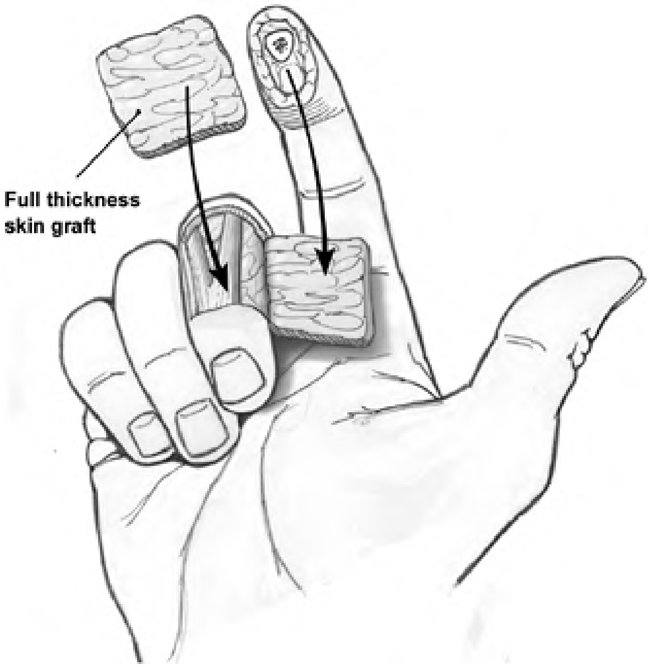

FIGURE 20-6 Illustration demonstrating where the cross-finger flap from the

dorsum of the middle finger is to be inset onto the amputated index tip. The full thickness graft is placed to cover the defect over the dorsum of the donor area. (Reproduced with permission of the Mayo Foundation, 2005.)

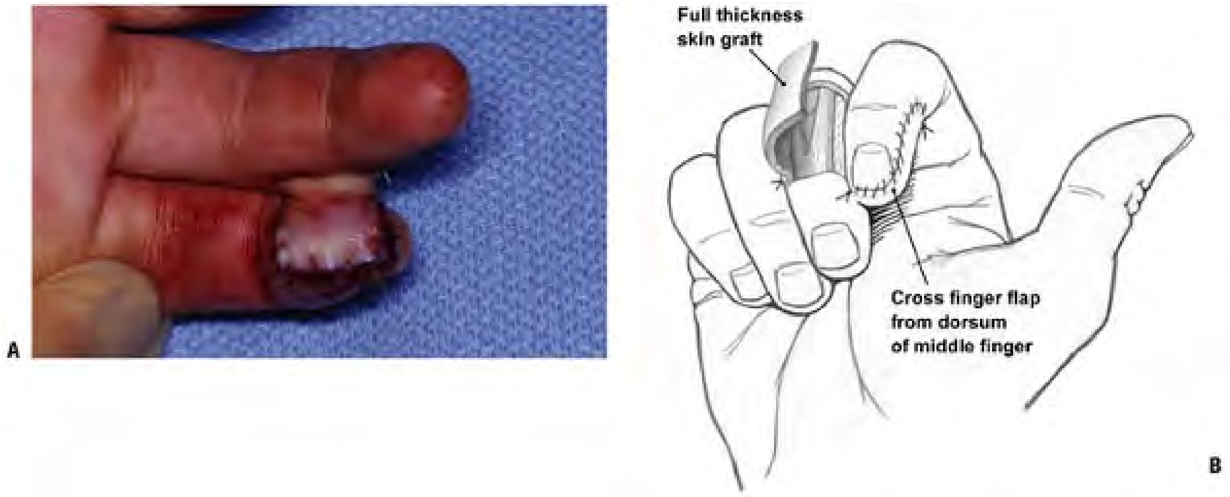
A
Full thickness skin graft
Cross finger flap from dorsum of middle finger
B

FIGURE 20-7 A: A clinical example of a cross-finger flap taken from the dorsum of the middle finger to cover the ring finger. B: Illustration demonstrating insetting of the flap. (Reproduced with permission of the Mayo Foundation, 2005.)

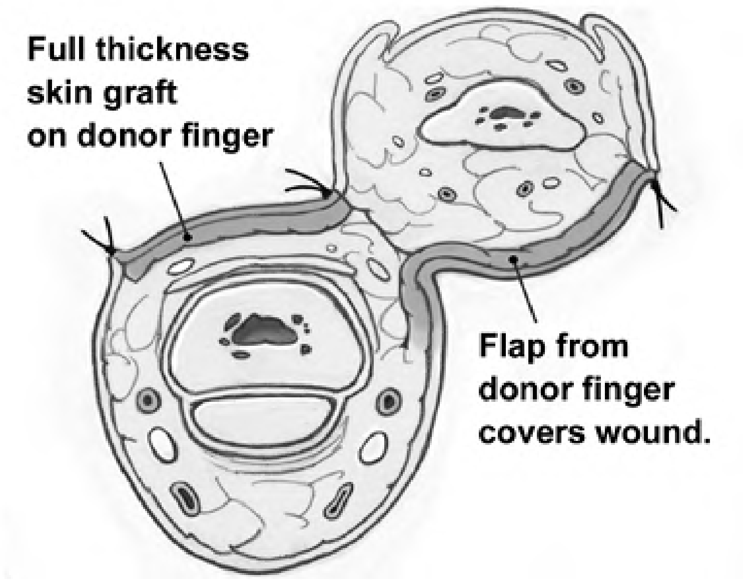
Full thickness
skin graft
on donor finger
Flap from
donor finger
covers wound.

FIGURE 20-8 Illustration demonstrating cross section through donor and recipient sites. (Reproduced with permission of the Mayo Foundation, 2005.)

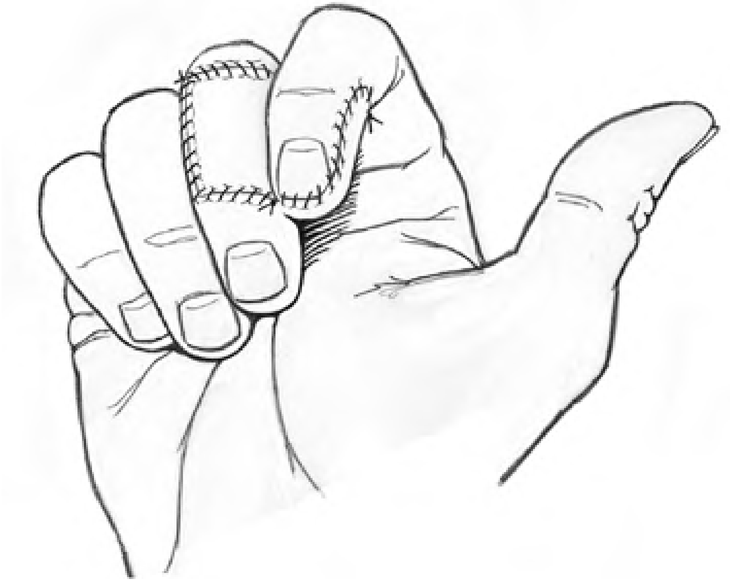

FIGURE 20-9 Illustration of the position of the finger after insetting of flap and

skin grafting of donor defect. (Reproduced with permission of the Mayo Foundation, 2005.)

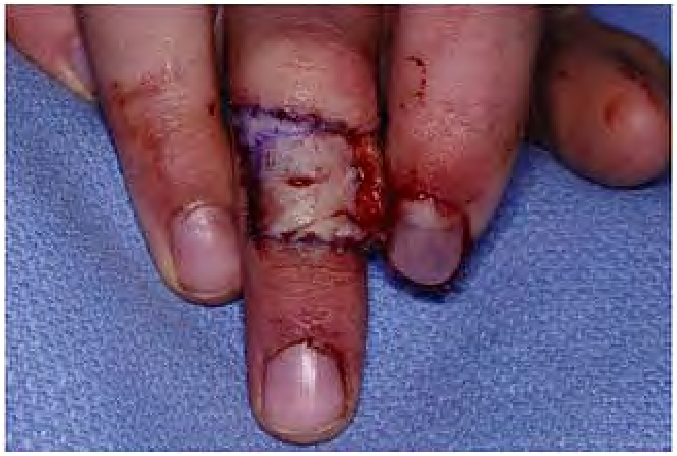

FIGURE 20-10 A full thickness graft from the forearm is used to cover the defect over the donor site.

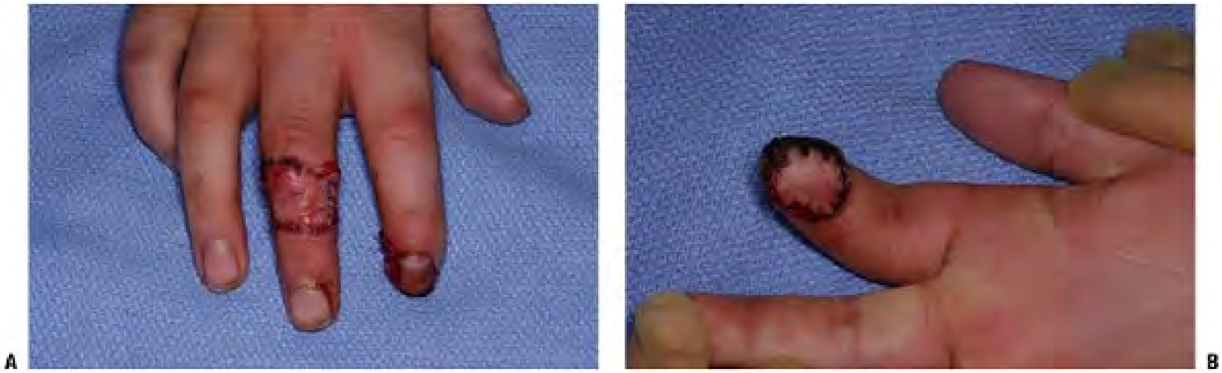

FIGURE 20-11 A,B: Division of the flap at 3 weeks postoperatively, with contouring and insetting of the flap.

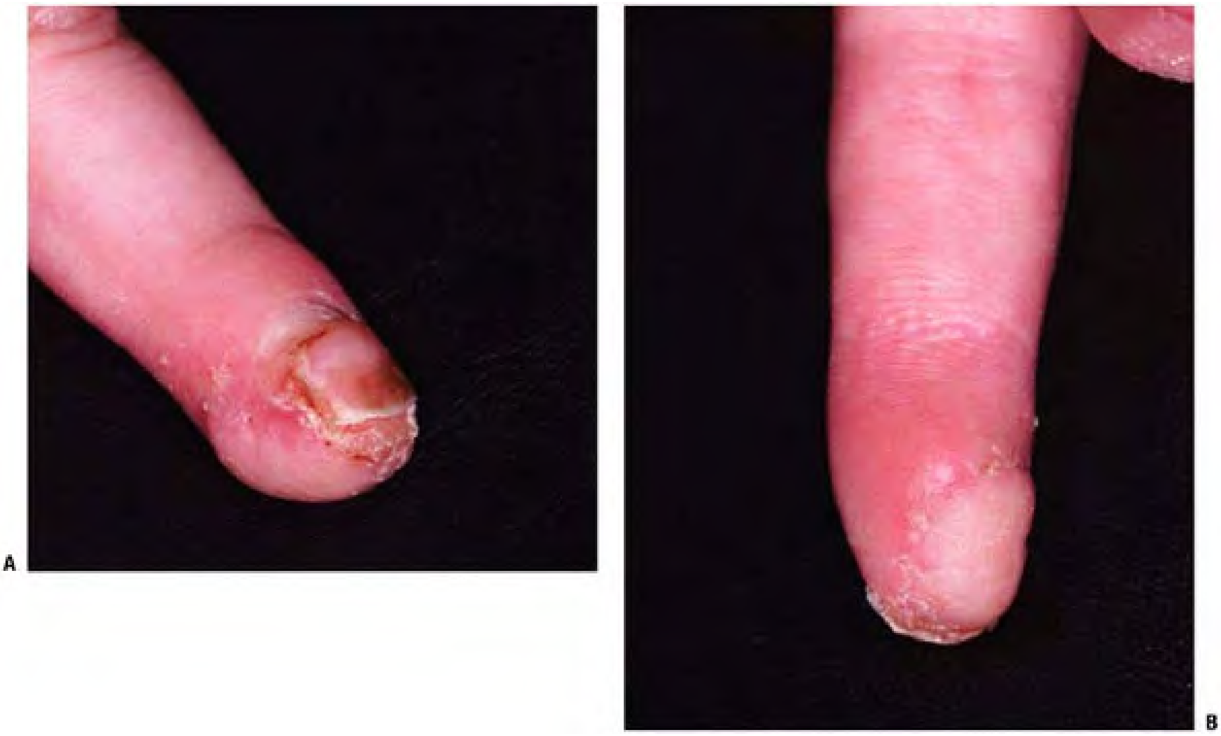

FIGURE 20-12 A,B: The appearance of the finger 3 months after surgery.

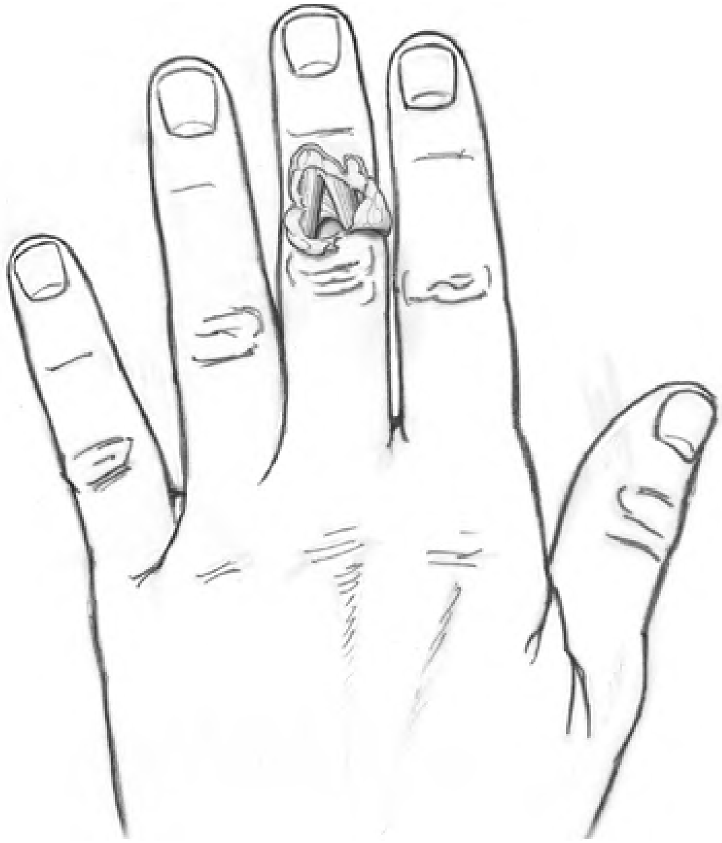

FIGURE 20-13 A defect on the dorsum of middle finger at the level of the middle phalanx with exposure of the extensor tendon denuded of its paratenon is an ideal indication for a reverse cross-finger flap. (Reproduced with permission of the Mayo Foundation, 2005.)

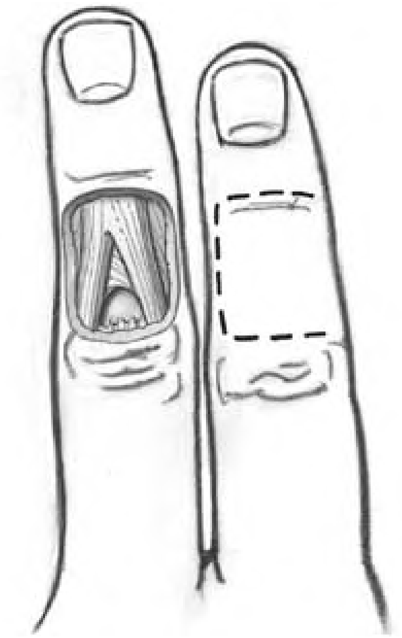

FIGURE 20-14 After the wound is debrided to fresh edges, the reverse cross-finger flap can be designed. (Reproduced with permission of the Mayo Foundation, 2005.)

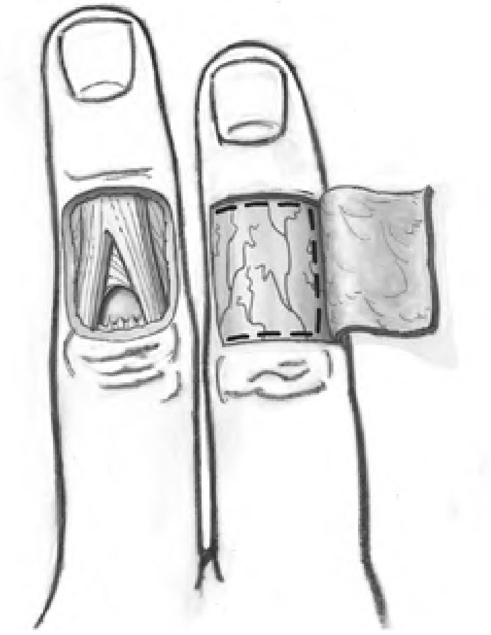

FIGURE 20-15 A thin full thickness skin flap is raised from the dorsum of the donor finger at the level of the middle phalanx, keeping the subcutaneous fat flap intact. The *31* indicate the site of the planned reverse cross-finger flap, which is divided to the level of the paratenon. (Reproduced with permission of the Mayo Foundation, 2005.)

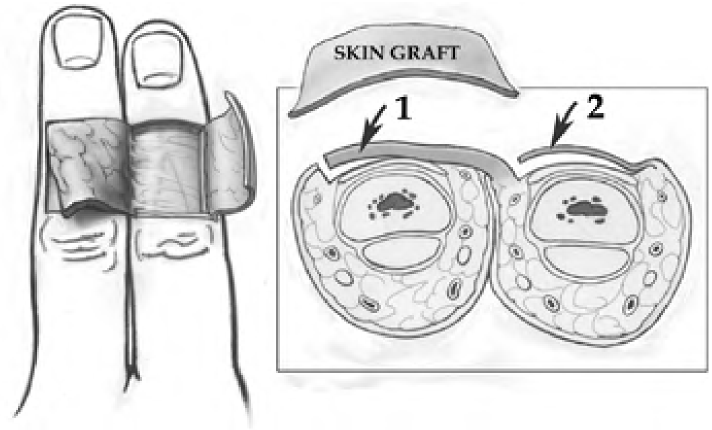

FIGURE 20-16 The hinge of the skin flap is opposite to the primary defect. The drawing shows how the reversed fat flap covers the dorsal finger skin defect (*31*). The raised thin thickness skin graft is re-placed again into its native site (*31*). Then, a thin thickness skin graft is placed over the reversed cross-finger flap. (Reproduced with permission of the Mayo Foundation, 2005.)

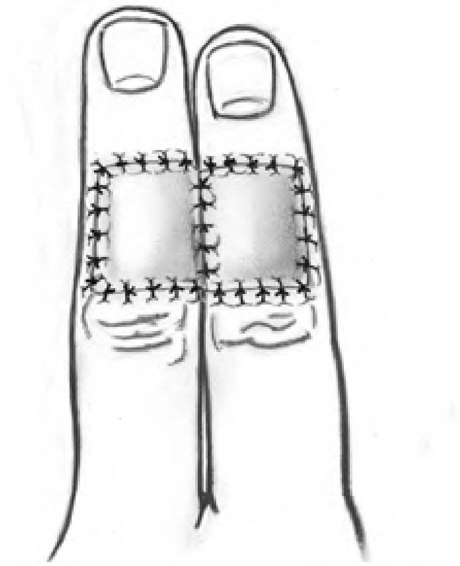

FIGURE 20-17 A drawing showing the appearance of the fingers after the after insetting of the reverse flap and placement of the skin graft. (Reproduced with permission of the Mayo Foundation 2005).

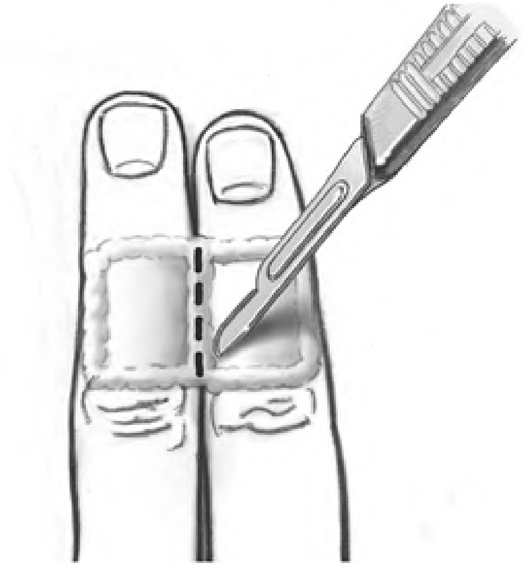

FIGURE 20-18 Division of the reverse cross-finger flap 2â€“3 weeks after the procedure. (Reproduced with permission of the Mayo Foundation 2005).

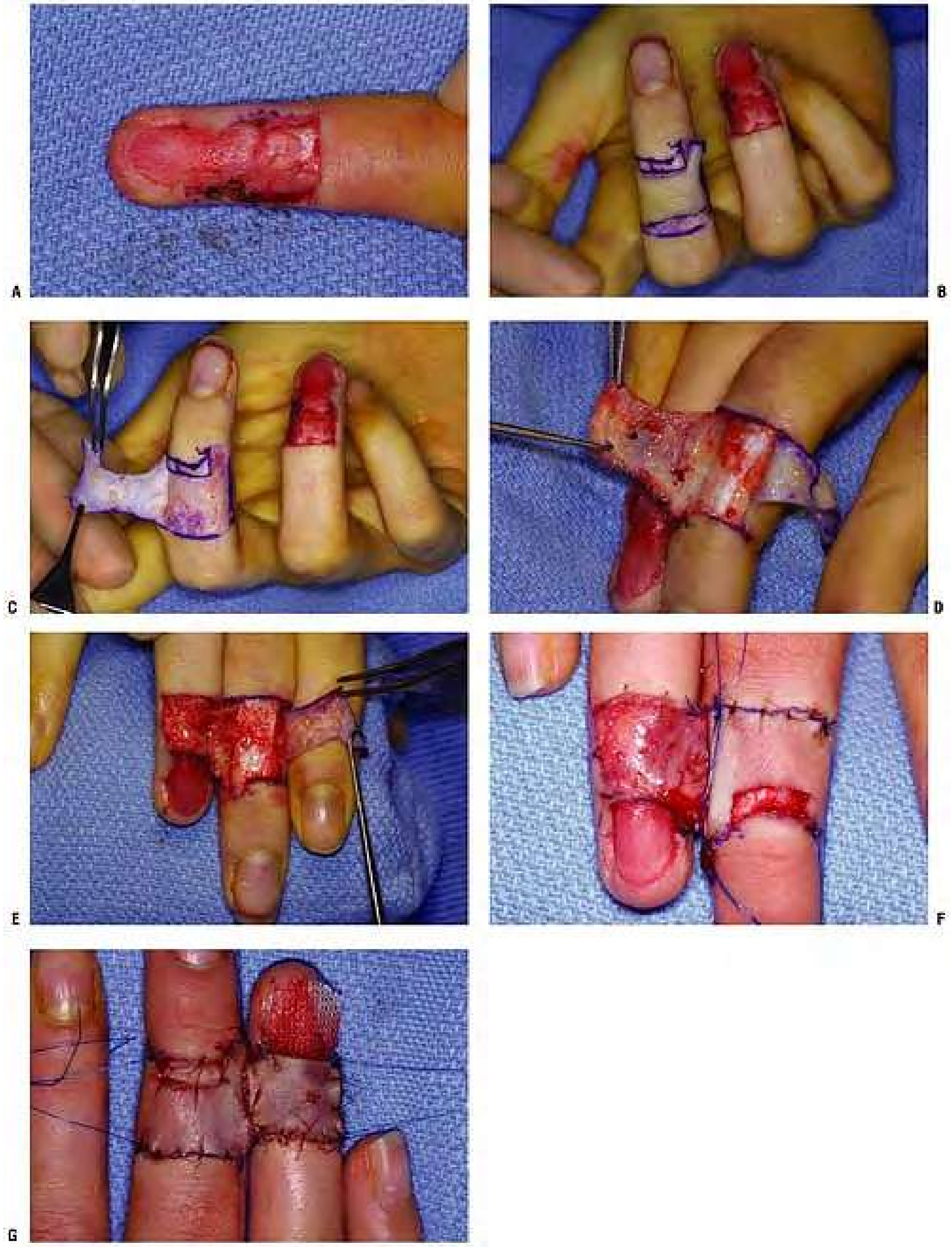

FIGURE 20-19 A-G: Clinical example of a reverse cross-finger flap, illustrating flap harvest and insetting. *31*

Postoperative Management and Results

After flap division, supervised finger range of motion commences, which includes combined active motion and passive stretching to eliminate stiffness and regain lost motion (see Fig. 20-9).

Postoperative infection, wound problems, and loss of flap are unusual (15).

The appearance of the skin graft and defect over the donor site are minimized or eliminated with the use of full thickness skin graft. Color mismatch between the dorsum and the pulp of the fingers is minimal, except in individuals with dark-colored skin.

The most common complications encountered with the reverse cross-finger flap include cold sensitivity and decreased sensation in the new transferred skin. In one report reviewing several follow-up studies of finger pulp loss coverage, 30% to 50% of patients developed cold sensitivity and 30% developed altered sensation regardless of the technique used (18). In another study of 54 patients followed over a 5-year period, 92% of the patients were satisfied with the result. However, 53% suffered cold sensitivity. All fingers developed protective sensation, but none had normal sensation or normal sweating (19). The altered sensory recovery varies by age. More than 90% of patients younger than 12 years had 6 mm or less two-point discrimination, compared to 40% of patients older than 40 years (20).

Atasoy (17) reported on four patients who had this reverse cross-finger flap done for nail bed defects (2 patients) with dorsal digital skin avulsions and/or extensor tendon injuries (2 patients). All patients had satisfactory results at the time of final follow-up. The patients with nail bed injury had satisfactory nail growth, and all patients had satisfactory extensor tendon function and good coverage of the defect. A potential complication of the reverse cross-finger flap is creation of an epithelial cyst. Atasoy recommended not grafting the reverse flap in case of large nail bed defect with an intact germinal matrix, in order not to preclude the nail from growing. However, if the there is a complete avulsion of the germinal matrix, nail bed, and skin, with exposure of bone and tendon and little possibility for the nail to return, the reversed flap should be skin grafted.

In another study, Al-Qattan compared the results of deepithelialized cross-finger flap reconstruction (31 patients) versus adipofascial turnover flaps (42 patients) for reconstruction of small complex dorsal digital defects (21). All flaps survived with no infection or hematoma. Patients who underwent reverse cross-finger flap had the following complications: flap dehiscence in one patient, stiffness of the donor finger in five patients, inclusion cyst in one patient, and significant skin graft loss in two patients.

Six patients who underwent adipofascial turnover flaps developed epidermolysis of the donor site, which resolved spontaneously. Al-Qattan recommended using the latter flap in children to avoid using of general anesthesia during flap division, and in older patients to avoid stiffness, and in multiple defects of adjacent border digits.

Conclusions

The cross-finger and reverse cross-finger flaps are reliable and dependable methods to cover volar and dorsal fingertip soft tissue defects. If carefully performed in properly selected patients, they are versatile flaps that provide a superior means of reconstruction for the injured finger with loss of significant tissue. The cross-finger flap has been used in a wide variety of patients of all ages with rewarding overall patient satisfaction.

References

1. Atasoy E, Iokimidis E, Kasdan ML, et al. Reconstruction of the amputated finger tip with a triangular volar flap. *J Bone Joint Surg.* 1970;52-A:921â€“926.

2. Fisher RH. The Kutler method of repair of finger tip amputation. *J Bone Joint Surg.* 1967;49-A: 317â€“321.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 21

Heterodigital Arterialized Flap

Shian Chao Tay

Lam Chuan Teoh

Indications/Contraindications

Hand trauma that results in exposed bone, tendon, joint, or neurovascular structures requires some type of flap coverage. Defects located around joints and web spaces may also require flap coverage to avoid contractures that could occur following split thickness grafting. The location and size of these defects can often preclude the use of small rotation or advancement flaps. Free flap coverage is an option, but the presence of surrounding infection or a wide zone of trauma may make the recipient vesselsâ€™ dissection difficult. In such situations, a heterodigital arterialized (HTA) flap may be the best option.

The HTA flap is raised from the lateral side of a donor finger together with the digital artery and a dorsal digital vein. Unlike Littler's neurovascular island flap, or Hueston's extended neurovascular island flap, the heterodigital arterialized flap's main function is to provide non-sensory reconstruction of skin defects in the hand or fingers. Thus, this flap is never harvested with the finger pulp or the digital nerve of the donor finger. This is an important feature that serves to reduce morbidity to the donor finger. The inclusion of the digital dorsal vein into the flap improves venous drainage of the flap and reduces the incidence of venous congestion, which is a well-documented complication of the classic Littler flap.

The heterodigital arterialized flap is a thin flap. In this aspect it is ideal for reconstruction in the hand and fingers as it provides near like-to-like reconstruction. The other advantages of this flap is that being regional, it obviates the need to prepare a separate surgical site and allows for almost immediate motion of the reconstructed finger. The reconstruction can be performed as a single-stage procedure (as compared to the cross-finger flap or groin flap).

In our experience, the heterodigital arterialized flap has been used to reconstruct volar, lateral, or dorsal defects in the fingers proximal to the distal interphalangeal joint, web spaces, palm, dorsum of the hand, and the thumb. The average flap dimensions are 4.1 cm (range, 1.5 to 5.5 cm) in the longitudinal axis and 2.1 cm (range, 1.0 to 3.5 cm) in the transverse axis. The HTA flap is thus a useful option to consider when faced with the problem of a relatively large defect in the hand or finger that requires non-sensory flap reconstruction. The HTA flap has been used for reconstruction of defects following infection, trauma, post-replantation of digits, electrical burns, chemical burns, and high-pressure injection injuries.

In terms of flap mobility, the reach of the heterodigital arterialized flap is limited to coverage of defects in the adjacent fingers or thumb or adjacent parts of the hand. This is principally due to the combined limitations in the reach of the dorsal vein and the digital artery of the flap. In situations where a greater reach is required, it is possible to divide the dorsal vein of the flap over the dorsum of the hand and reanastomose it in a region of healthy tissue after flap transfer. With this maneuver, the flap should have the ability to reach defects two to three fingers away, much like the traditional reach of Littler's neurovascular island flap, when pivoting at the common digital artery's offshoot at the superficial palmer arterial arch.

The HTA flap is absolutely contraindicated if there is only one functioning digital artery in the donor finger and when the vascular viability of any other adjacent fingers is threatened by flap harvesting.

It is also relatively contraindicated if the functional prognosis of the finger to be reconstructed is judged to be so poor that it might be better to amputate it. However, we recognize that social or cultural practices may still dictate reconstruction for the purpose of cosmesis or the maintenance of the â€œwholeâ€•

self. Conversely, if there is a rare situation wherein a need to perform a digital amputation exists in a hand containing a defect requiring flap reconstruction, the HTA flap can be used as a fillet flap from the amputated finger for defect reconstruction. In such a situation, the HTA flap is very useful in spare parts surgery as there is no concern for donor finger morbidity. If the defect occurs beyond the reach of the flap or in the contralateral hand, a free fillet HTA flap can be created from the finger to be amputated provided the conditions at the recipient finger allows for a free flap.

Preoperative Planning

All practical options available for reconstruction should be considered by the surgeon before deciding on the HTA flap. Risks of heterodigital donor finger morbidity should be carefully weighed against benefits of reconstruction. Often, the location, size, and nature of the defect will preclude other reconstructive options, leaving the HTA flap as the most viable option with the highest flap survival rate.

The choice of donor finger is often dictated by the location of the defect and is often an adjacent finger. The outside borders of the index and little fingers are not used as donor sites to preserve hand cosmesis and to maintain native skin for protection and sensation. This requirement is not absolute; however, we prefer the use of the middle finger, ring finger, and the ulnar border of the index finger. Theoretically, the radial border of the little finger can be used as a donor site but this is rare as the little finger is small, providing little skin, and the remaining ulnar digital artery is often absent or vestigial.

A digital Allen's test must be performed on the intended donor finger to ensure that both digital arteries are sufficiently patent. In addition, the Allen's test should also be performed on the finger that is adjacent to the donor site as this would dictate if the digital artery mobilization can proceed proximal to the point of bifurcation of the common digital artery. More often than not, a longer reach is required and ligation and division of the neighboring branch of the digital artery to the adjacent finger would need to be performed to create an arterial pedicle of sufficient length.

If the donor site is just next to the defect in the adjacent finger, also known as the contiguous or near side of the adjacent finger (e.g., defect on ulnar side of index finger with donor site on radial side of middle finger), the pivot point can be at the bifurcation of the common digital artery, as a short pedicle would be sufficient. This pivot point can be approximately landmarked in the intermetacarpal space at the level of the distal palmer crease. The distance from this pivot point to the proximal edge of the defect can then be measured to determine the reach needed by the flap and hence the location of the proximal margin of the flap. If the flap is not coming from the contiguous side of the adjacent finger, it is likely that a more proximal pivot point would be required to ensure sufficient reach of the flap. The most proximal point would be the superficial palmer arterial arch, which can be approximated to the level of the proximal palmar crease in the relevant intermetacarpal space. In such situations, as mentioned previously, it will be necessary to ligate and divide the neighboring digital artery of the common digital artery that supplies the adjacent finger to maximize the reach of the flap.

Experience has shown that the reach of the flap increases by 10% once the digital artery is adequately mobilized. However, in the interests of safety and ensuring minimum tension in the pedicle, a 1:1 ratio should still be maintained for the length of the arterial pedicle during preoperative planning of flap surgery.

The timing for reconstruction depends on the pathology. As far as possible, early reconstruction (within 1 week) is preferred as it minimizes overall hand stiffness from prolonged disuse. For traumatic conditions, the reconstruction is ideally performed within 1 week. For post-replantation surgery with residual skin defects, the resurfacing is performed within 10 to 15 days when re-endothelialization of the vascular anastomoses is completed. Infection cases are ideally resurfaced within 4 to 7 days when the surrounding cellulitis and edema have subsided. In order to achieve this time line in infective cases, infection reversal should be achieved within one to two formal definitive excisional debridement surgeries in conjunction with appropriate intravenous antibiotics.

The surgery can be performed under general or regional anesthesia with sedation. The patient's general condition and coagulation profile should be optimized as per any reconstructive surgery. The patient should be consented and counseled as to the surgery, the postoperative course, and the duration of stay following surgery.

Surgery

Patient Positioning

The patient should be in the supine position, and care should be taken to protect bony prominences with suitable pressure-relieving bolsters. The upper limb with the hand requiring reconstruction should be abducted to not more than 80 degrees at the shoulder to prevent excessive stretch to the nerves in the axilla. The elbow should be flexed slightly so that the forearm can be placed transversely on a stable hand table. A pneumatic tourniquet should be applied to the most proximal part of the upper arm, and the upper limb should be cleaned and draped as far proximal until the distal edge of the tourniquet. If the surgery is expected to last more than 2 hours, urinary catheterization should be considered.

Technique

The surgery is performed under tourniquet control, which should be inflated with the upper limb elevated but not exsanguinated. This is to facilitate identification and mobilization of the digital artery and the digital dorsal vein. The defect should be thoroughly debrided and irrigated. A template of the defect is made. At this point, the digital Allen's test can be repeated to confirm the adequacy of both digital arteries in the donor finger, and also in the adjacent finger if the flap is not taken from the contiguous side of the finger adjacent to the defect.

Using the template created, the appropriate dimension of the flap is then transferred onto the donor site. The flap should be centered on the lateral or dorsolateral side of the finger, depending on the width of flap required, to ensure the inclusion of the digital artery and a dominant digital dorsal vein. The maximum width of the flap should not exceed the mid-palmer line and the mid-dorsal line (a width of approximately 3 cm). The maximum length of the flap is from the base of the finger to the distal interphalangeal joint crease (usually 4 to 5 cm depending on the size of the finger). If necessary, the flap length may be extended 0.5 cm distal to the distal interphalangeal joint crease. Next, a palmar Z incision is drawn, biased to the respective side of the finger to access the proximal part of the digital artery. This incision is carried into the palm to the level of the common digital artery. The location of the dominant dorsal digital vein is then verified and marked to ensure that it will drain the flap. Once this is completed, the flap is ready to be harvested.

Heterodigital arterialized flap harvesting is performed with the aid of loupe magnification. The flap is first raised from its palmar margin (Fig. 21-1). If the palmar margin of the flap crosses any digital palmer flexion creases, a mini-z cut can be made at these creases, to break up the resultant scar line. Superficial palmer veins can be carefully cauterized with bipolar electrocautery and divided. As one proceeds deeper, the dissection follows the digital artery from proximal to distal starting from the bifurcation of the common digital artery. The angle of dissection toward the neurovascular bundle is an oblique plane skiving dorsal to the digital nerve, which is more central and superficial compared to the digital artery. As far as possible, the nerve should be left undisturbed in its bed with a collar of fat kept around it. Besides ensuring that the nerve is not traumatized, the collar of fat also provides additional cushioning and enhances the take of the full-thickness skin graft at the donor site. Ideally, a collar of fat or soft tissue should also be kept around the digital artery to preserve the venae comitantes. Fortunately, in the HTA flap, this is not as critical, as the dorsal vein would provide the main conduit for venous drainage.

At this point, an intraoperative occlusion test can be performed using non-traumatic microsurgical vascular clamps to determine the integrity of the contralateral digital artery in the donor finger. If mobilization of the digital artery to the level of the superficial palmer arch is necessary, a similar occlusion test should be performed on the bifurcated digital artery supplying the adjacent non-donor finger to ensure that it also has a healthy contralateral digital artery.

As mobilization of the digital artery progresses, the transverse palmer arches of the digital artery will be encountered. These are rather large and short. Sufficient room is developed and the arches can be carefully cauterized with bipolar electrocautery before division. Care should be taken to ensure that the bipolar electrocautery is applied at least 2 mm from the parent artery to prevent compromising flow in the main digital artery.

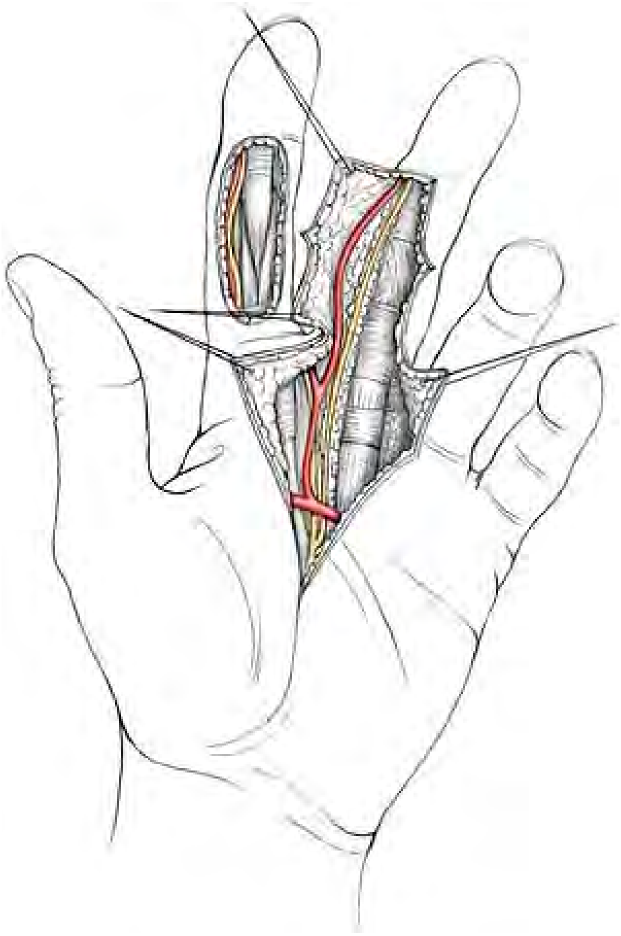

FIGURE 21-1 Completed palmar dissection of the HTA flap with the digital nerve left undisturbed in its bed with a collar of fat around it. Proximal dissection of the arterial pedicle is facilitated by a palmar Z incision into the palm.

Dorsal dissection is performed next with proximal to distal dissection and mobilization of the dominant dorsal digital vein to ensure inclusion of the vein in the flap. The flap is dissected free from the extensor tendons in the plane superficial to the paratenon. Care should be taken not to damage the dorsal skin branches of the digital artery supplying the flap. At the distal margin of the flap, the distal end of the dominant dorsal vein is ligated and divided, and the same is performed for the digital artery on the palmer side.

At this stage, the flap should be detached from distal to proximal. The adherent fibrous septa and Cleland's ligaments are divided. At the base of the digit, some dissection may be necessary to complete the division of subcutaneous tissues attached to the flap. The natatory ligament at the web space may also be divided to prevent kinking of the dorsal vein and lengthen the reach of the flap.

A generous subcutaneous tunnel is created between the donor and the recipient site. Tunneling is performed from the proximal edge of the recipient site into the palm to reach the pivot point of the flap. A hemostat forceps is used to gently enlarge the tunnel for easy flap delivery. The distal end of the flap is carefully grasped and gently maneuvered into the tunnel. A recommended method of flap delivery is to completely wrap the flap with tulle gras dressing and deliver the flap by grasping only on the tulle gras. This serves to prevent direct grasping of the flap and also protects the flap and pedicles from being avulsed if the delivery in the tunnel is inadvertently rough or jerky. As the flap traverses the tunnel, care is taken to ensure that the digital artery and the dorsal vein do not become twisted or kinked. The flap should not be rotated more than a half turn. If there is any uncertainty, the flap should be re-delivered to ensure its correct orientation.

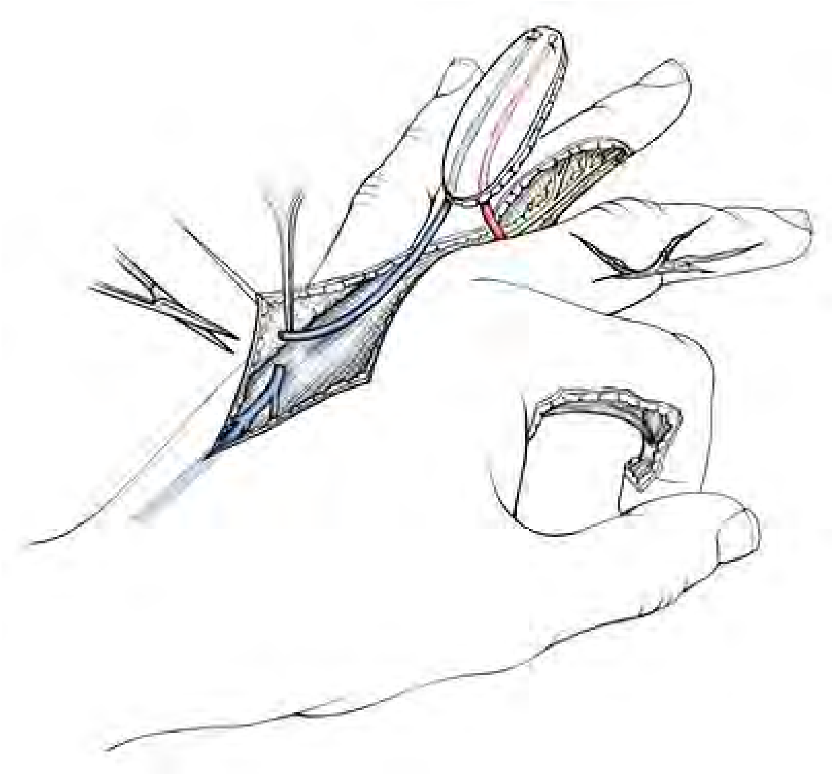

FIGURE 21-2 Completed dorsal and palmar dissection of the HTA flap with the flap completely islanded from the donor ring finger. The flap is now to be transferred. In this case, the defect requiring resurfacing is on the volar side of the index finger. The middle finger has a significant injury and is not suitable as a donor. The bi-pedicled nature of the flap will obstruct the transfer through the palmar subcutaneous tunnel. Thus, the venous pedicle of the flap has been divided over the dorsum of the hand to partially free the flap and convert it into a mono-pedicled flap.

When a longer reach of the flap is required, and the bipedicled nature of the flap is restricting the transfer, the dominant dorsal vein may be divided before the transfer (Fig. 21-2) and anastomosed to the original vein or another vein over the dorsum of the hand in an area free of adversity, after the flap has been transferred (Fig. 21-3). When this procedure is performed, it is called the â€œpartially free HTA flap.â€•

After flap transfer, the flap is provisionally held in place with a few fine sutures.

The tourniquet is then deflated to check for flap perfusion and to secure hemostasis. If flap perfusion does not return after 5 minutes or if there is flap congestion, some sutures may have to be removed. If this fails to improve flap color, kinks in the vascular pedicle may have to be located and relieved. If the vascular pedicle suffered mild inadvertent trauma during surgery, appropriate flap resuscitation measures such as warming the patient, warm saline soaks to the artery, and immersion of the arterial pedicle in 2% lidocaine or papavrine may have to be performed before adequate flap perfusion is restored.

Once adequate flap perfusion is restored, the flap is then fully inset with fine sutures, usually nonabsorbable monofilament 5-0 sutures, spaced not less than 3 to 5 mm apart. Care should be taken during the insetting to ensure that both the dorsal vein and the digital artery are not compressed or traumatized. The donor defect is resurfaced with a full-thickness skin graft. This can be taken from the medial proximal forearm, the anterior elbow, or groin if necessary. As early postoperative active rehabilitation is required, it is highly recommended to secure the skin graft with a cotton wool bolster and a meticulous tie-over dressing to prevent skin graft loss during therapy. The rest of the surgical wounds are closed with fine sutures with or without non-suction drains. A bulky, non-constrictive dressing is applied with a window left open for flap monitoring. Healthy flap perfusion should be confirmed before leaving the operating room.

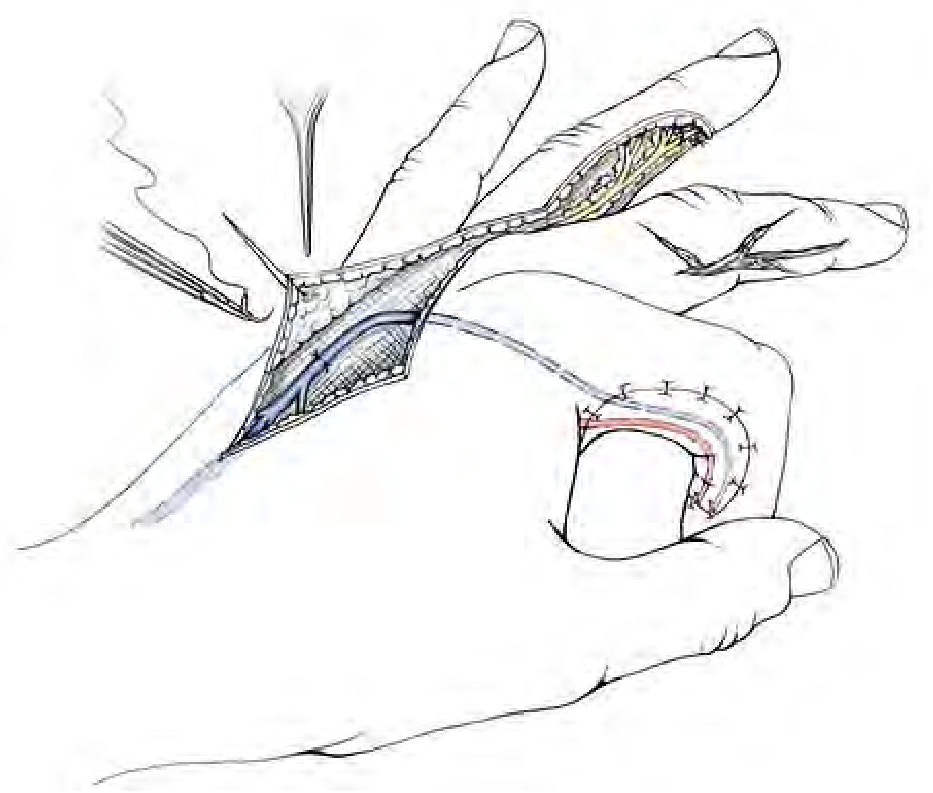

FIGURE 21-3 The partially free HTA flap after successful transfer to the defect site on the index finger. The venous pedicle of the flap has been tunnelled back for microsurgical repair to the original draining vein over the dorsum of the hand.

Results

A total of 40 HTA flaps were performed between 1991 and 2001. Thirteen of these flaps involved division and repair of the dominant dorsal digital vein to extend the reach of the flap. Total active range of motion of the donor finger was excellent in 82.5%, good in 15%, and fair in 2.5% (1 patient) according to Strickland and Glogovac's criteria for flexor tendon surgery (see Figs. 21-4, 21-5, and 21-12). The fair result was due to the development of reflex sympathetic dystrophy and associated flexion contracture. Donor finger two-point discrimination was 3 to 5 mm, except in one case (6 mm). None of the donor fingers had hypersensitivity, symptomatic neuromas, or suffered from cold intolerance.

Flap survival was 100% with no cases of flap congestion or ischaemia documented. All flaps healed primarily and provided supple coverage of the defects. Full-thickness skin graft take was similarly successful at 100% in the donor fingers. Although a contour concavity is present at the time of removal of the cotton wool bolus on the fifth postoperative day, the concavity usually fills out within 3 months after surgery. None of the cases required any further surgery to the flap such as insetting or defatting.

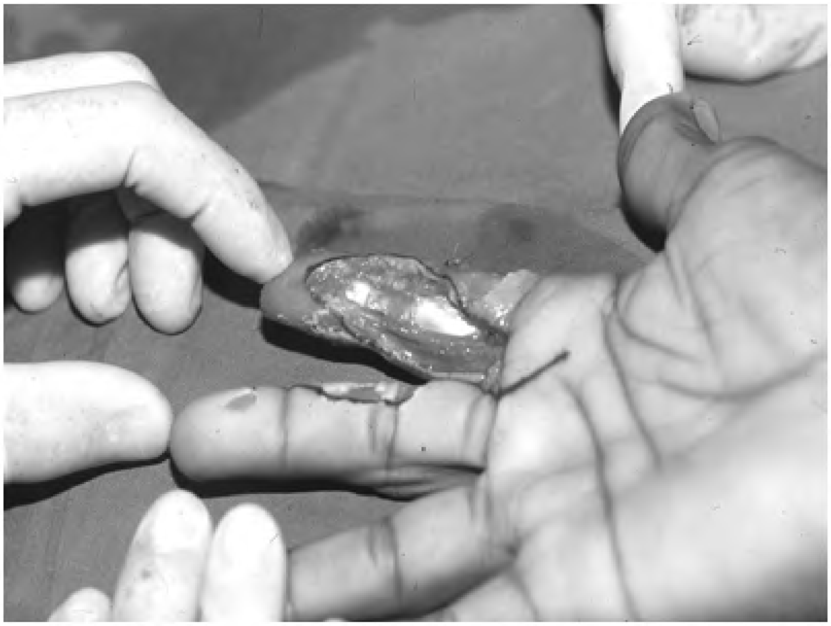

FIGURE 21-4 A sizeable defect (5 Ã— 2 cm) over the volar-ulnar aspect of the right index finger exposing bare flexor tendons and the neurovascular bundles following trauma. The HTA flap would come from the radial side of the middle finger. Thus, it would be a contiguous transfer of the HTA flap. There is no need to divide the venous pedicle to partially free the HTA flap in such a situation. (From Teoh LC, Tay SC, Yong FC, Tan SH, Khoo DBA. Heterodigital arterialized flaps for large finger wounds: results and indications. *31*2003;111(6):1905â€“1913. With permission.)

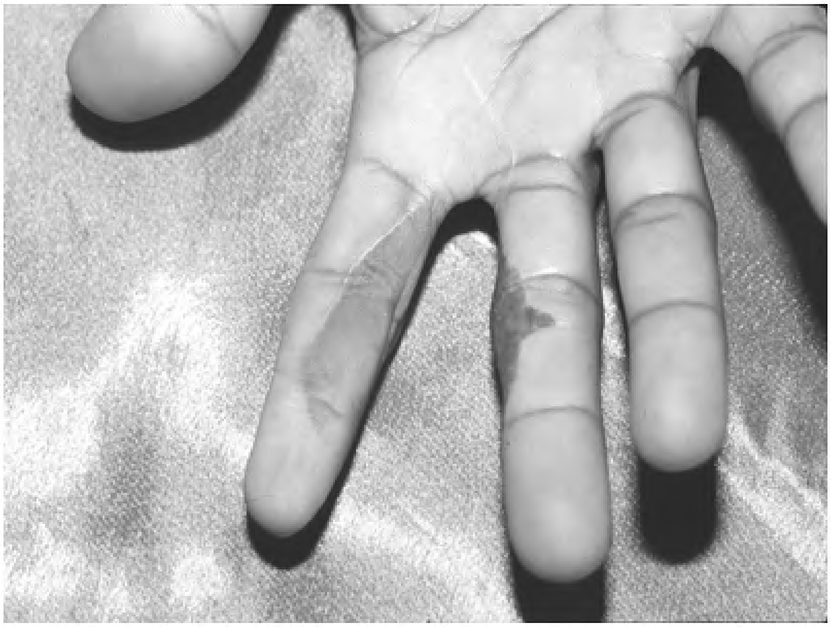

FIGURE 21-5 Cosmetic result in this dark-skinned individual at 3 months postoperatively. Patient has full extension of the donor middle finger and has excellent primary healing of the HTA flap on his index finger. (From Teoh LC, Tay SC, Yong FC, Tan SH, Khoo DBA. Heterodigital arterialized flaps for large finger wounds: results and indications. *31*2003;111(6):1905â€“1913. With permission.)

We evaluated the total active motion in 22 fingers which had undergone reconstruction with the HTA flap. According to Strickland and Glogovac's criteria, 45.4% achieved excellent and good results, with the remainder achieving fair to poor results. It should be noted that the poorer outcomes in the reconstructed fingers are due to severe insults resulting in multiple tissue involvement with concomitant damage to joints, tendons, and/or ligaments and not due to contractures of the HTA flap.

Postoperative Management

The first 24 hours are the most critical for the flap. The flap should be monitored frequently with the hand nursed in an elevated position during this time and the patient kept nil per oral (NPO). The ambient temperature can be raised to a level appropriate for the season at least for the first night depending on the surgeon's preference. Drains, if used, should be removed within the first 24 to 48 hours. Rehabilitation should commence on the second postoperative day but can be delayed if circumstances dictate. An early and active rehabilitation program is crucial in minimizing heterodigital donor finger morbidity. Active and passive range of motion exercises are instituted for both the recipient and donor digits. Regional anesthesia in the form of infusion blocks can be maintained during the postoperative period to assist with pain control during rehabilitation. The use of a cotton bolster ensures that the full-thickness skin graft is firmly secured so as to allow progressive hand therapy within the first week. Interval splinting of the donor digit between exercises may be necessary to prevent flexion contractures.

Complications

Heterodigital donor finger morbidity is the main concern. As mentioned previously, the preservation of the digital nerve and the finger pulp on the donor finger is critical toward preserving sensory function in the donor finger. In addition, the meticulous dissection of the digital artery away from the digital nerve is crucial in minimizing inadvertent trauma to the digital nerve, which can affect sensory function in the finger pulp and produce sensitive mini-neuromas along the digital nerve. Together with careful resurfacing of the donor site using full-thickness skin grafting and institution of early active finger mobilization, stiffness of the donor and reconstructed finger can be significantly reduced.

Mild loss of total active motion can be expected in about 10% to 15% of cases. Reflex sympathetic dystrophy with flexion contracture is another serious complication that can occur. Flap failure, either partial or complete, is always a potential complication, but it has not occurred in our series. In the five cases of free HTA flap, we did not experience any vascular complications and all the free flaps survived completely. In dark-skinned races, the darker dorsal skin may be a poor cosmetic match for defects on the volar side of the hand (Fig. 21-5).

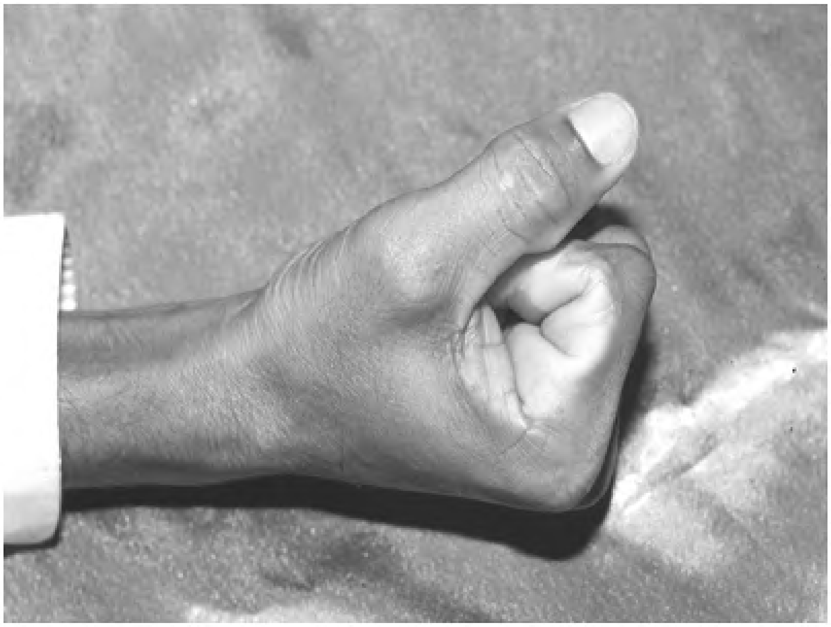

FIGURE 21-6 Excellent result and hand function with full finger flexion in both donor and reconstructed fingers. (From Teoh LC, Tay SC, Yong FC, Tan SH, Khoo DBA. Heterodigital arterialized flaps for large finger wounds: results and indications. *31*2003;111(6): 1905â€“1913.

Variants of the Heterodigital Arterialized Flap

Besides the free fillet flap, other variants of the HTA flap include the cross-finger HTA flap, which is particularly suited for coverage of dorsal defects distal to the proximal interphalangeal joint which cannot be reached by the reverse dorsal intermetacarpal flap or the Quaba flap. Another indication for the cross-finger HTA flap is in situations where proximal dissection of the digital artery in the hand is not advisable due to previous trauma or infection. The cross-finger HTA flap only requires dissection of the venous and arterial pedicle up to the level of the base of the finger and will allow transfer to any surface on the adjacent finger (Fig. 21-6). This is accessed via mid-lateral incisions raising small dorsal and volar soft tissue flaps proximal to the donor site. Care should be taken not to extend these incisions into the web space or the lateral walls of the web commissure. A similar mid-lateral incision is performed proximal to the defect on the injured finger. The flap is directly transferred across the interdigital space and onto the defect (Fig. 21-7). This avoids the need to tunnel the pedicles and flap through the distal part of the hand and also allows the flap to reach a defect on any surface of the finger distal to the proximal interphalangeal joint. The dorsal and volar soft tissue flaps of the mid-lateral incisions are then sutured together to create a soft tissue pouch that will protect the flap pedicle and prevent separation of the digits. Division of the cross-finger version of the HTA flap is performed at 3 weeks. In the interim, active rehabilitation of the joined fingers should be started to prevent finger stiffness.

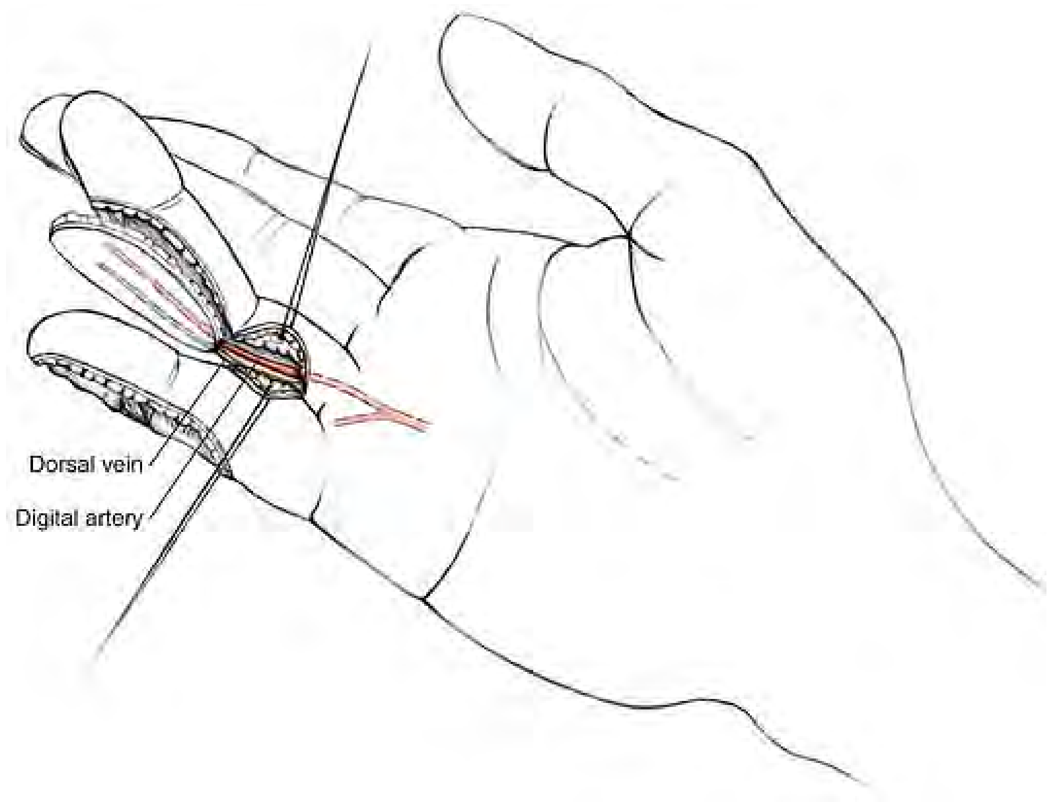

FIGURE 21-7 Cross-finger HTA flap from the ulnar side of the ring finger for transfer to a defect on the ulnar side of the distal little finger. Note that minimal pedicle length is required for this direct cross-finger transfer and

there is no need to perform any pedicle dissection in the palm.

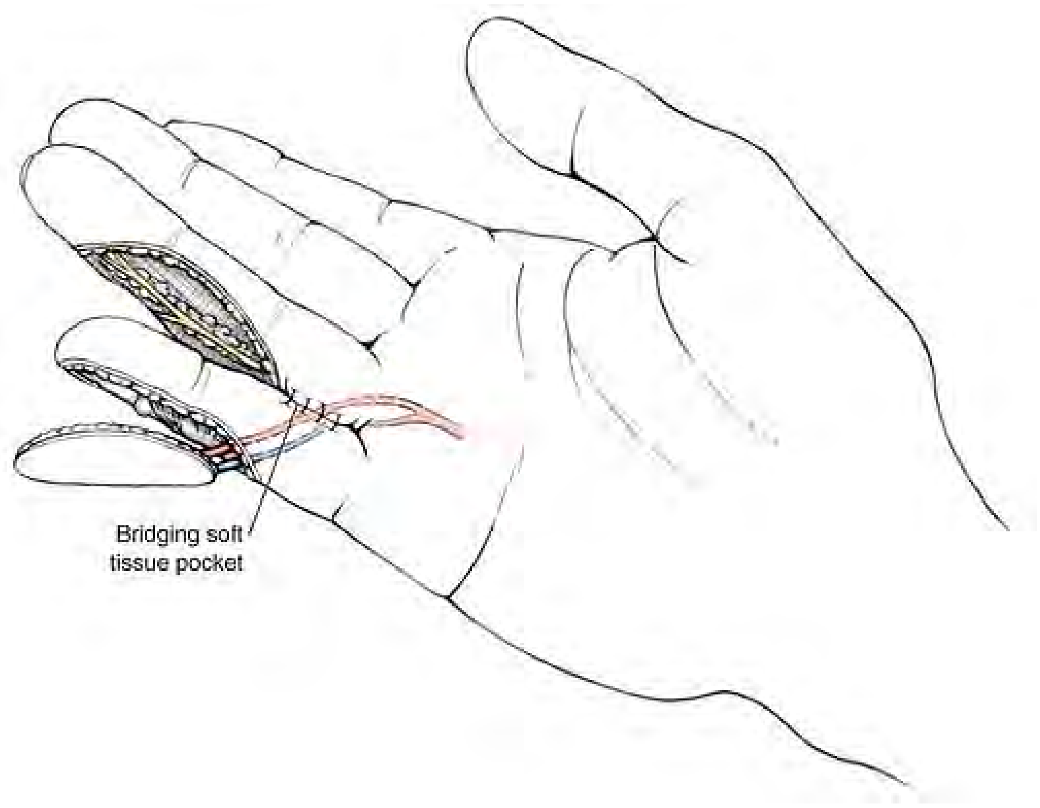

FIGURE 21-8 Cross-finger HTA flap after transfer to the defect site. The flap has been tunnelled through volar tissue of the little finger. The vascular pedicles of the flap are now protected by a soft tissue skin bridge that has been created between the base of the ring and little finger. A full-thickness skin graft with a cotton wool bolus with tie over will be applied to the donor site. The soft tissue skin bridge, together with the HTA flap vascular pedicles, will be divided at 3 weeks postoperatively when the cross-finger HTA flap has established its own vascular supply in the reconstructed little finger.

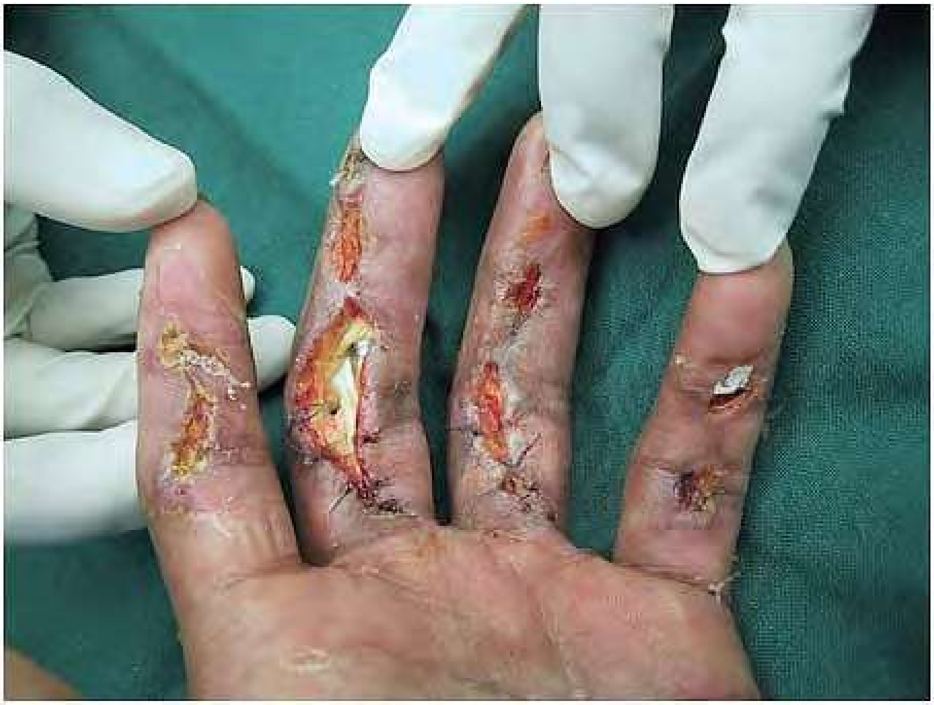

FIGURE 21-9 A patient referred from another hospital with a sizeable chronic non-healing defect on his left middle finger with concomitant injuries to all the other fingers.

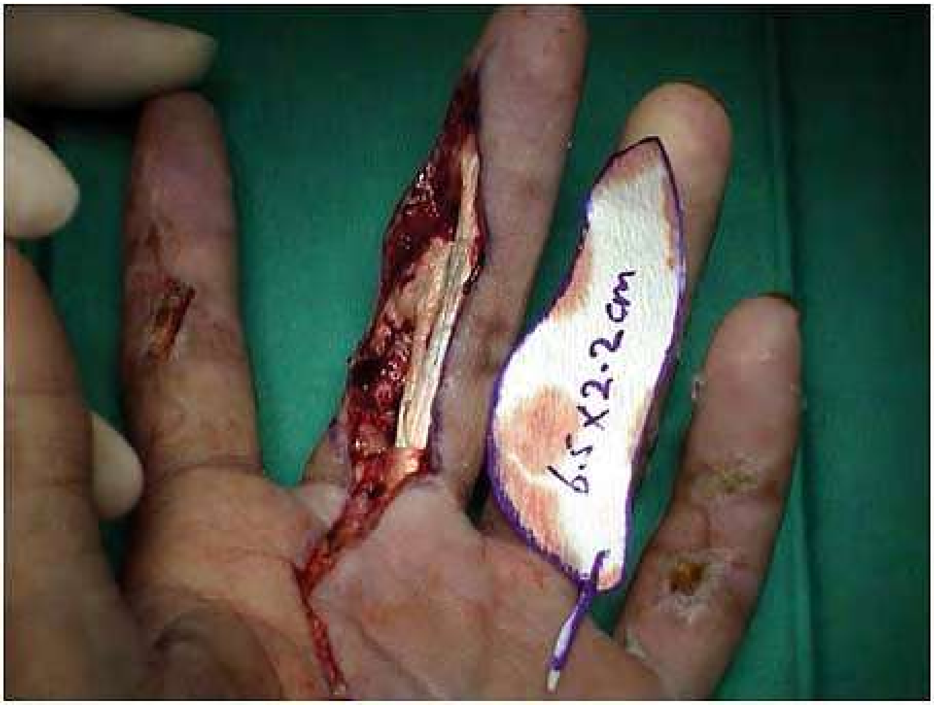

FIGURE 21-10 After debridement, the true defect size is apparent (6.5 Ã— 2.2 cm) with exposed bare flexor tendons. The patient will be undergoing free HTA flap procedure with the flap being harvested from the ulnar side of the right middle finger.

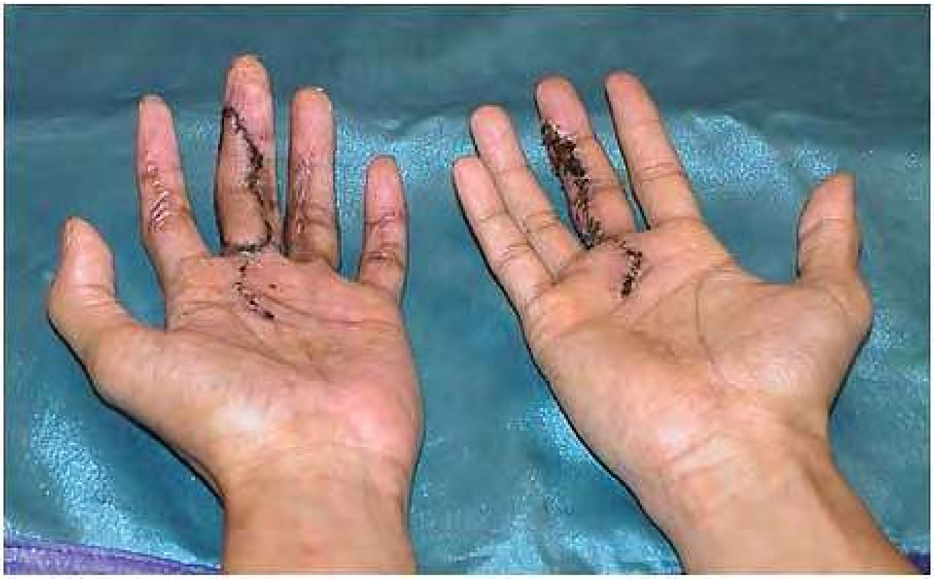

FIGURE 21-11 Two weeks post free HTA flap with 100% survival of the free HTA flap with excellent primary healing in the reconstructed left middle finger. The full-thickness skin graft is also healing well on the donor site (ulnar side of the right middle finger).

The final variant is the free HTA flap. This free flap should only be used by hand microsurgeons already well experienced with the HTA flap. In occasional situations, a sizeable defect exists in a hand in which there is concomitant or prior injuries to the other fingers, rendering them unsuitable as donors of the HTA flap. In such situations, if the defect is amenable for a free flap procedure and other free flaps are not available, the HTA flap may be harvested as a completely free flap from the uninjured contralateral hand. In such situations, the free HTA flap is taken from the ulnar side of the middle finger. The indication for the free HTA flap is certainly unique and rare. First performed in 1991, we have since found cause to only have performed five such flaps. In one case, it was used to electively resurface a sizeable defect on the small finger following a successful four-finger replantation. In another case, it was used to resurface a non-healing chronic defect on the middle finger following an injury which also involved the other three fingers (Figs. 21-8 and 21-9). In two cases, there were acute concomitant injuries to the other fingers. In the last case, the free HTA flap was used as a free flow-through arterial flap, not only to reconstruct the skin defect, but also to reconstruct a digital artery segment defect in a finger that had arterial insufficiency following an injury that also involved all the fingers. The excellent size matching of the digital arteries certainly contributed significantly to the success of the procedure. In all five cases, except for the mild cosmetic defect of the full thickness skin graft at the donor site, all the donor fingers retained full and complete range of active motion and preservation of normal pulp sensation (Figs. 21-10 and 21-11). An important reason why the free HTA flap might have such an excellent result in the donor fingers might be due to the fact that there were no other injuries in the donor hand, which greatly facilitated active rehabilitation.

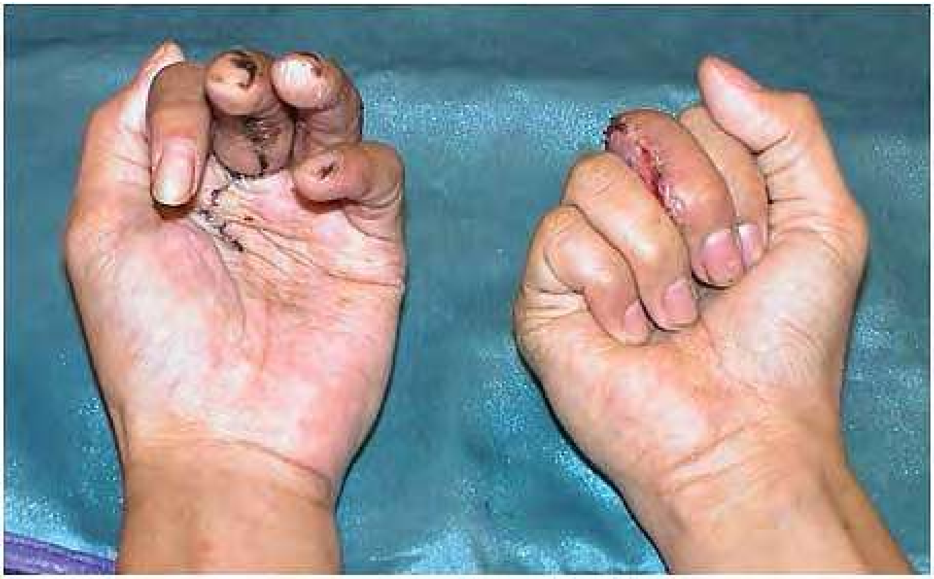

FIGURE 21-12 Full flexion of the right middle donor finger demonstrated. The stiffness that is apparent in the left hand is due to prolonged immobilization and disuse as a result of the chronic non-healing wound. With successful and robust coverage of the defect in the left hand with excellent primary healing, aggressive rehabilitation of the hand can be started to alleviate the stiffness.

Recommended Reading

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 22

Homodigital Island Flap

Bradon J. Wilhelmi

Damon Cooney

Single-stage reconstruction of the finger pulp has become possible with a better understanding of hand anatomy and the advent of new surgical techniques. Single-stage reconstruction eliminates the risk of a stiff finger resulting from immobilization and the necessity of a second operation that is required with traditional regional flap reconstruction. The options for one-stage reconstruction of pulp defects include heterodigital and homodigital neurovascular island flaps. The heterodigital neurovascular island flap has the disadvantage of late or no development of sensory reorientation and paresthesia in the donor finger. Harvesting the flap from the injured finger, as is the case in a homodigital island flap, avoids unnecessary surgery to an adjacent uninjured finger as well as potentially enhancing rehabilitation by minimizing additional hand trauma. Many homodigital neurovascular flaps based on one or two volar digital pedicles have been described which can provide the fingertip with sensate coverage. The palmar advancement flap based on bilateral neurovascular pedicles has been used for reconstruction of the thumb and fingers. The interphalangeal (IP) joint is usually needed to be kept in flexion to obtain adequate advancement of the flap, which increases the complication risk for flexion contracture of the distal joint with these advancement flaps. Also, the use of the advancement flap increases the risk for dorsal skin necrosis if the dorsal branches from the proper digital artery are severed for distal advancement of the flap. The volar V-Y flap is still a suitable flap for small fingertip defects less than 1 cm when there is sufficient skin distal to the distal interphalangeal (DIP) crease to base the flap.

Indications/Contraindications

The homodigital neurovascular flap has been described with elevation of the dorsal finger skin to repair pulp deficits which cannot be closed with the volar V-Y advancement flap. Anatomic studies have demonstrated that a dorsal arterial branch from the proper digital artery exists over the middle phalanx which runs obliquely, dorsally, and distally over the distal interphalangeal crease. The dorsal branch at the middle phalanx collateralizes with the dorsal blood supply of the finger from another dorsal branch from the proper digital artery at the level of the distal phalanx, forming a dorsal arterial arcade over the DIP joint. Also, a dorsal nerve branch can be found leaving the proper digital nerve to innervate the dorsal finger skin proximal to the nail fold. The dorsal homodigital neurovascular island flap can be based on one or both of these arteries and this dorsal nerve. When the flap is harvested on the distal-most branches, it can be called the distal dorsal homodigital artery (DDHDA) flap. This DDHDA flap can be used to close defects of 2 Ã— 2.5 cm just proximal to the tip and hyponychium (Fig. 22-1). Larger defects in this region can be reconstructed by a flap based on the dorsal branch at the middle phalanx, called the middle dorsal homodigital neurovascular island flap, or MDHDA flap. This MDHDA flap can be used to close defect 2 Ã— 4 cm in size (Fig. 22-2). Alternatively, when the loss is very distal and an antegrade homodigital neurovascular island flap will not reach the defect, a reverse homodigital island flap (RHDA flap) can be used. The RHDA flap is harvested from the lateral base of the finger and based on proximal dorsal branches of the proper digital artery (Fig. 22-3). Dorsal branches of the proper digital nerve can be used to innervate this skin paddle, coapting this nerve to the transected nerve at the tip. These homodigital flaps are contraindicated when there is significant injury involving the pedicle or the skin on which the flap is based over the dorsum of the finger. Moreover, these homodigital flaps should not be used when

a skin bridge cannot be preserved for venous drainage to the distal finger. If the harvest of the flap would result in the exposure of already injured extensor tendon, devoid of paratenon, then the flap should be avoided. Crush and electrical injuries to the dorsal finger are also relative contraindications for the homodigital flap.

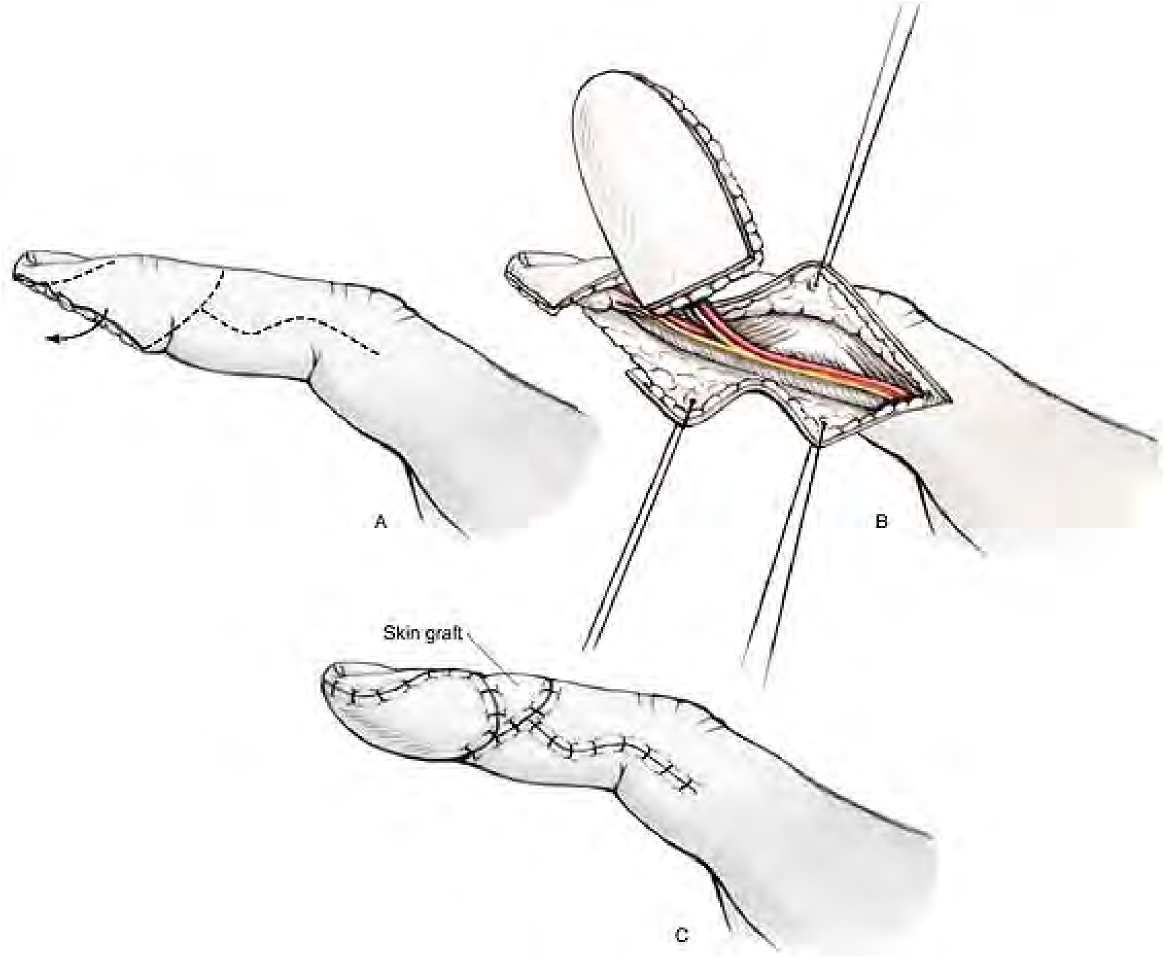

FIGURE 22-1 A: The design of the incision and distal dorsal homodigital neurovascular island (DDHD) flap. The axis of the flap curves obliquely and dorsally toward the contralateral side. B: The flap is elevated with its neurovascular pedicle and transposed obliquely into the defect and the typical wound suitable for a DDHDA flap with an intact tip and hypochium. C: The flap inset and grafted donor defect.

Preoperative Planning

The reconstruction of fingertip defects with homodigital island flaps should be performed in the operating room. The flap should be harvested with the use of a tourniquet to facilitate identification of anatomic structures. A microscope should be available should the digital artery be injured and require repair, or if the nerve has to be coapted. Amputated remnants should be saved. These amputated parts can be used as a source for a skin graft for the tip defect if the flap cannot be used or even as the donor defect of the flap. The risks of the procedure should be explained to the patient and family. Patients need to understand that if the flap fails, they may require an amputation, as well as other risks such as loss of extension, cold intolerance, swelling, and loss of some sensation.

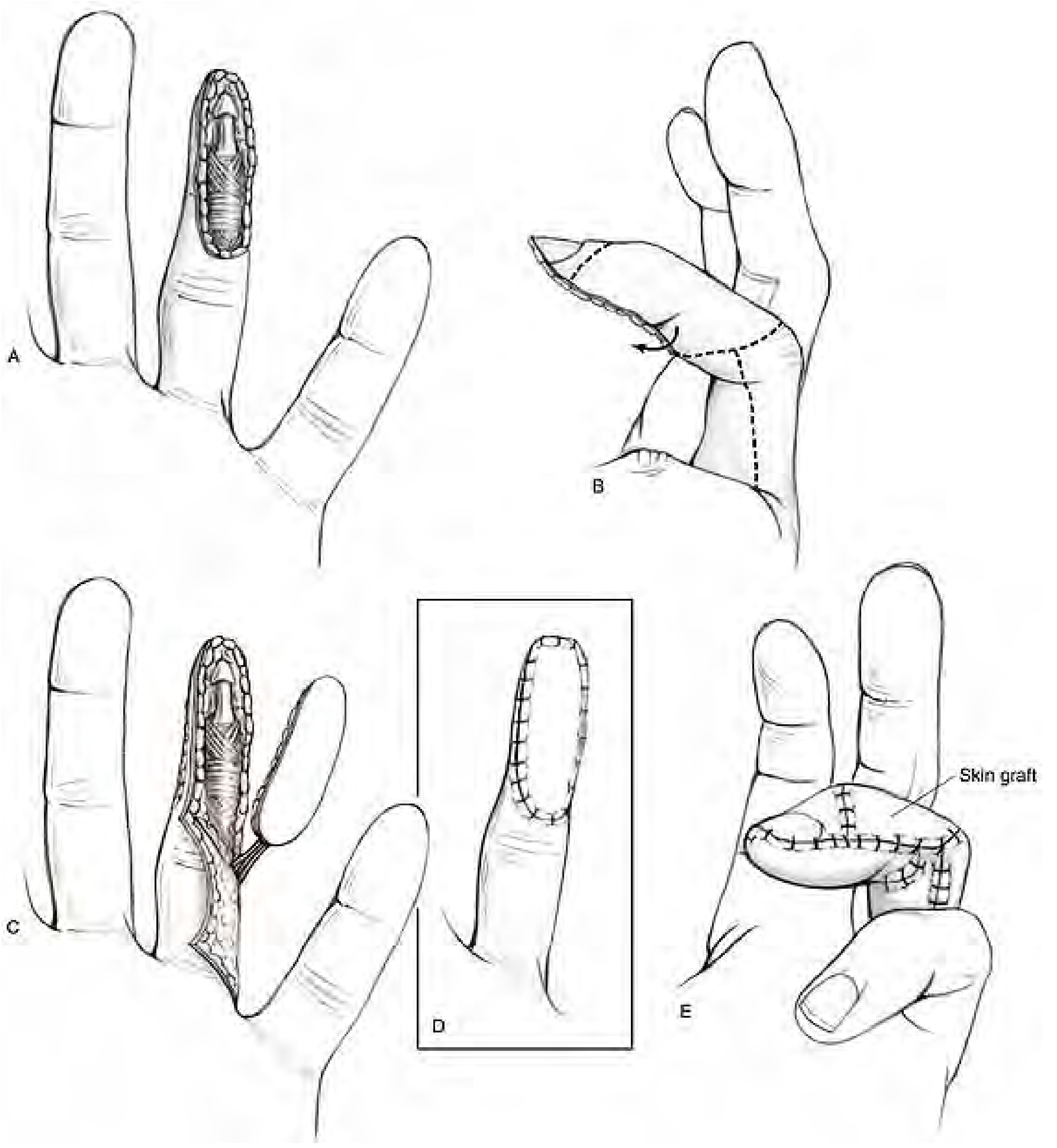

FIGURE 22-2 A,B: The rectangular flap design for the middle dorsal homodigital neurovascular island flap. C: The proximal pedicle dissection. D,E: The insetting of the flap and the grafted donor.

Surgery

Patient Positioning

The homodigital flap reconstruction of a finger defect can be performed in the operating room with the patient in the supine position with the involved hand on an arm board and an upper arm tourniquet. The arm is prepared up to the tourniquet. The procedure can be performed with a regional block, such as a bier block or axillary block, or under general anesthesia.

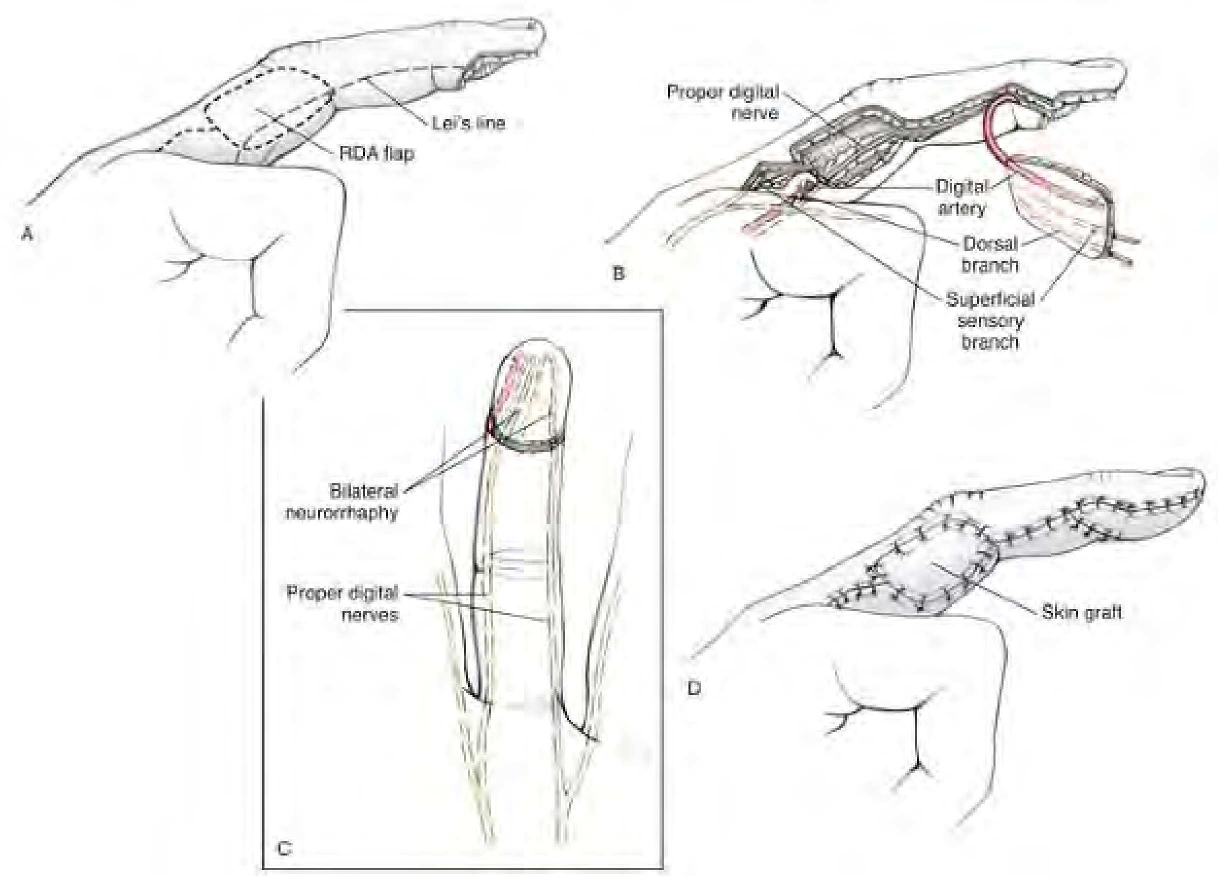
Lei's line
RDA flap
A
Proper digital nerve
Digital artery
Dorsal branch
Superficial sensory branch
B
Bilateral neurorrhaphy
Proper digital nerves
C
Skin graft
D

FIGURE 22-3 The required size of RDHDA flap is marked on the dorsolateral aspect of the involved proximal phalanx. A: The flap is based on the digital artery which topographically courses under the dorsoventral skin junction line. B: The vascular pedicle is ligated at its proximal end. Both the dorsal branch from the proper digital nerve and the superficial sensory branch from the corresponding radial or ulnar nerve are sectioned proximally, leaving 1 cm nerve tails attached

to the flap. C:The flap is transposed to the recipient wound. The attached sensory nerves of the flap are microanastomosed with both ulnar and radial digital nerves at the recipient wound. D: The flap donor site is resurfaced with full thickness skin graft. The pivot area of the vascular pedicle is also covered with a piece of skin to eliminate the pressure completely. (From *31*1990;46B:484. With permission.)

Technique for Distal Dorsal Homodigital Artery (DDHDA) Flap

A template of the defect is made and is used to design the skin paddle of the flap. The flap is outlined on either the ulnar-dorsal side or the radial-dorsal side of the involved digit, depending on the wound location. When either digital artery can be used, the non-opposition side of the digit is preferred. The flap design is based either over the distal phalanx and DIP joint (defect <2.5 cm) or the middle and distal phalanx, depending on the size of the defect. The distal margin of the flap should be maintained 4 mm proximal to the nail fold to avoid germinal matrix and resultant nail abnormality. Based on the width of the pulp defect, the DDHDA flap proximal incision line is marked obliquely from the volar edge of the pulp defect toward the DIP crease, reaching proximally to the dorsum of the middle phalanx of the finger. The flap may include a portion of volar tissue of the terminal phalanx if any residual pulp tissue remains intact. The extension for the distal and proximal incision lines on the dorsum of the digit is determined by the length of the designed flap. The length

of the flap should be longer than the defect to ensure adequate coverage of the flap pedicle (see Fig. 22-1). The neurovascular bundle is approached first via a zig-zag midlateral incision extending at least into the proximal interphalangeal crease. The digital pedicle is dissected proximal to distal with inclusion of surrounding fatty tissue to preserve the venae commitante with the proper digital artery to ensure venous drainage of the flap. This dissection should be performed with loupe magnification. In the proximal edge of the flap, particular attention is paid to include sufficient soft tissue between the pedicle and the flap to ensure adequate blood supply to the flap because the dorsal feeder arteries coming off the proper digital artery in the middle phalanx of the finger are usually divided during dissection. The flap is elevated in the plane above the paratenon from the dorsum of the finger toward the pedicle, then to the volar edge (see Fig. 22-1). Once the flap elevation is complete, the flap is advanced obliquely to fill the pulp defect. The donor site on the dorsum of the finger is covered with a full thickness graft taken from the hypothenar eminence or upper forearm.

Technique Middle Homodigital Neurovascular Island (MDHDA) Flap

When a larger skin paddle (>2.5 cm and extending into the middle phalanx) is needed, the MDHDA flap is preferred. The MDHDA flap is designed as a rectangular island, with the axis of the flap parallel to the length of the volar skin defect. The rest of the procedure proceeds as described previously, except the proximal incision line is extended obliquely toward the proximal interphalangeal crease, and the dorsal digital branches at the middle phalanx are included (see Fig. 22-2).

Technique for Reverse Digital Island Flap

For very distal level injuries, when the arc of pedicle length might not reach for the DDHDA and MDHDA flaps, the reverse homodigital island flap can be used. The reverse digital island flap is designed on the dorsolateral side of the involved proximal phalanx according to the size and shape of the pulp defect (Fig. 22-3). The dissection begins proximally at the proximal digital neurovascular bundle at the base of the finger (Fig. 22-4). The dorsal nerve and artery branches from the proper digital nerve and artery are identified while preserving a thick tuft of soft tissue from the proper digital artery to the skin paddle. The skin paddle is elevated in the dorsal to volar direction. The skin over the neurovascular bundle is incised in zig-zag fashion. The proper neurovascular bundle is dissected in the proximal to distal direction, preserving the soft tissue around the proper digital artery to include the venae comitante to avoid venous insufficiency of the flap. Ideally, the proper digital nerve to the finger is preserved. Collateral flow from the contralateral digital artery to the pedicle proper digital artery should be preserved. The flap is inset and the dorsal nerve branch is coapted to the digital nerve in the defect. Alternatively, the proper digital nerve can be included with the pedicle and repaired to the contralateral digital nerve in the defect.

However, this sacrifice of the proper digital nerve renders the finger insensate on the pedicle side, which can be avoided by using a dorsal nerve branch. The donor site at the base of the finger can usually be closed primarily, if a full thickness graft is not preferred.

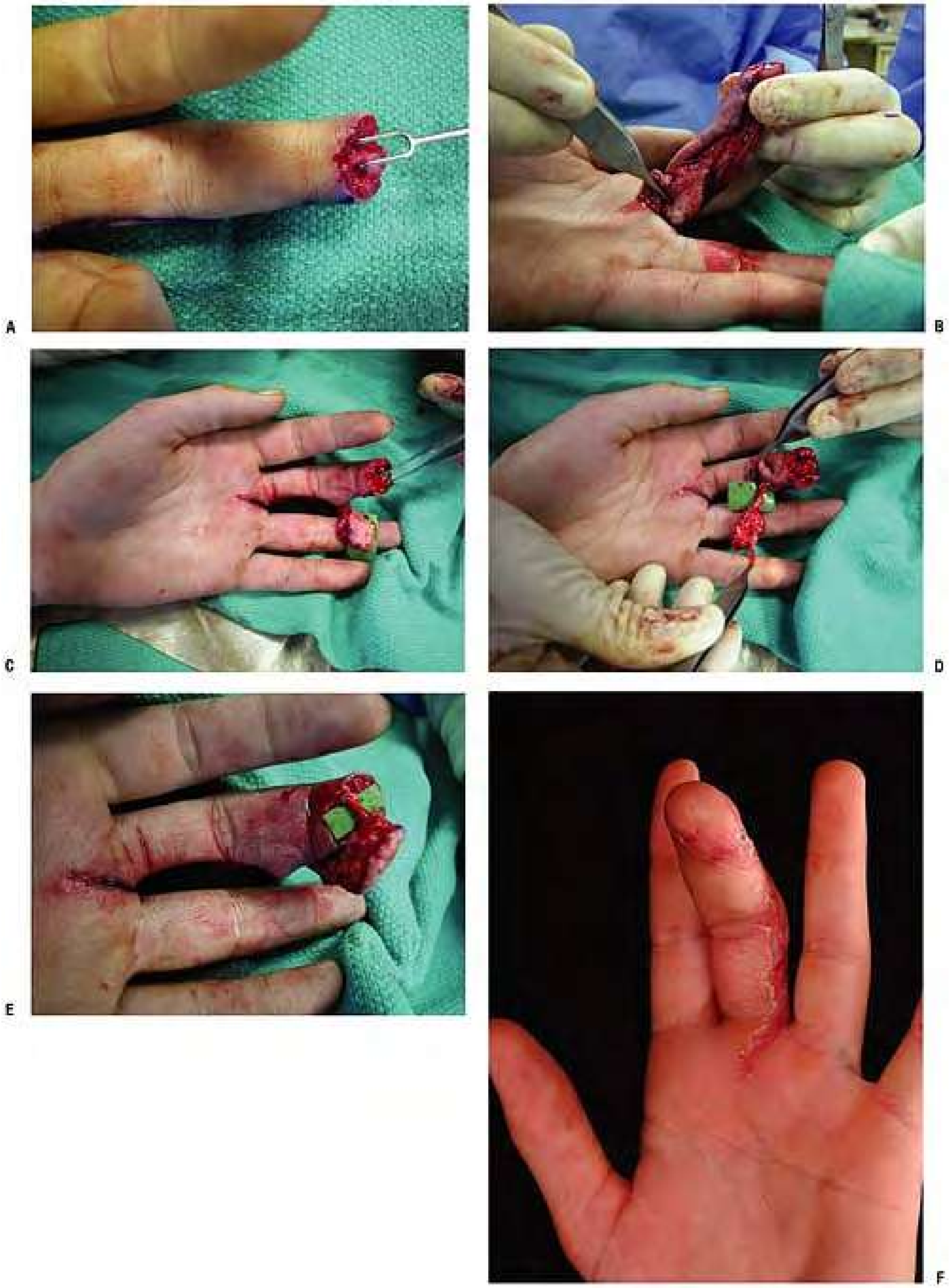

FIGURE 22-4 A: Distal tip amputation of the long finger with exposure of underlying bone. In such a case, further shortening of the finger to allow for primary coverage would lead to a loss of the profundus tendon insertion and decrease finger function. B: A reverse homodigital island flap is designed on the ulnar aspect of the long finger. C: The digital artery is then dissected just distal to the proximal interphalangeal crease. D: The flap is ready for transfer. Fibrofatty tissue has been preserved around the vascular pedicle to prevent injury to the venous comitantes. The forceps hold a dorsal sensory nerve branch which has been included in the flap. E: The flap is rotated 180 degrees allowing the dorsal sensory nerve to be coapted to the injured radial digital nerve. F: The early postoperative appearance with excellent contour and well-healed donor site.

Results

Several reports with these flaps have demonstrated that the flaps have reliable arterial and venous flow. Moreover, these studies have shown good return of sensation in the 5â€“9 mm range of 2-point discrimination. Some of the reviews have demonstrated a mild loss of extension complicating the skin grafted donor site on the dorsum of the injured fingers.

Postoperative Management

After the homodigital island artery flap, splint immobilization is unnecessary. Early mobilization can be encouraged which will lower the risk for a stiff finger postoperatively. If the RDHDA flap is used, usually the donor can be closed primarily. If a skin graft is required at the base of the finger, a splint is used postoperatively for 1 week.

Complications

Postoperative complications from the homodigital flap include loss of extension, loss of sensation, cold intolerance, and flap loss, most likely from venous compromise. The risk of loss of extension can be minimized with early motion. Fingers that have been reconstructed with innervated homodigital flaps can achieve 5 mm 2-point discrimination, but the grafted donor site can be less sensate. There is no increased risk for cold intolerance in patients reconstructed with homodigital flaps over other options. The risk for venous compromise can be lessened by harvesting the pedicle to include soft tissue and venae comitante with the digital artery.

Recommended Reading

Atasoy E, Ioakimidis E, Kasdan ML, et al. Reconstruction of the amputated finger tip with a triangular volar flap: a new surgical procedure. *J Bone Joint Surg.* 1970;52A:921â€“926.

Flint MH, Harrison SH. A local neurovascular flap to repair loss of the digital pulp. *Br J Plast Surg.* 1965;18:156â€“163.

Foucher G, Smith D, Pempinello C, et al. Homodigital neurovascular island flaps for digital pulp loss. *J Hand Surg.* 1989;14B:204â€“208.

Joshi BB. A local dorsolateral island flap for restoration of sensation after avulsion injury of fingertip pulp. *Plast Recontr Surg.* 1974;54:175â€“182.

Lai CS, Lin SD, Chou CK, et al. Innervated reverse digital artery flap through bilateral neurorrhaphy for pulp defects. *Br J Plast Surg.* 1993;46:483â€“488.

Lucas GL. The pattern of venous drainage of the digits. *J Hand Surg.* 1984;9A:448â€“450.

Mober E. Aspects of sensation in reconstructive surgery of the upper extremity. *J Bone Joint Surg.* 1964;46A;817â€“825.

Oâ€™Brien B. Neurovascular island pedicle flaps for terminal amputations and

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part IV - Soft Tissue Coverage of the Fingers Following Trauma > Chapter 23 - First Dorsal Metacarpal Artery Island Flap

---

Chapter 23

First Dorsal Metacarpal Artery Island Flap

GÃ¼nter Germann
Katrin Palm-BrÃ¶king

Soft tissue defects in the hand, and particularly in the thumb, frequently present difficult reconstructive problems because of the restricted availability of local tissue. Although local skin flaps have been employed and time-proven over decades, the philosophy of immediate wound closure of complex defects with exposure of tendon, bone, or joints has stimulated interest in anatomic research and the development of refined reconstructive techniques (10).

Traditional flaps have many disadvantages such as two-stage operations (i.e., cross-finger flap [17 ,24]), tedious dissections with considerable donor-site morbidity and loss of discriminative power (i.e., Littler flap [27 ,28]), or limited arc of rotation and mobility (i.e., transposition flaps [7] and the Moberg flap [1 ,8]). On the other hand, more recent microsurgical reconstructions such as the free pulp flap (7) require a significant amount of time and a familiarity with microsurgical technique. The first dorsal metacarpal artery island flap overcomes the disadvantages of traditional hand flaps and can provide a moderate-sized skin paddle on a long consistent vascular leash, allowing one-stage reconstruction for thumb and dorsal hand wounds without the need for microsurgical anastomosis.

The first dorsal metacarpal artery island flap was originally described by Hilgenfeldt in 1950 (16) and subsequently refined by Paneva-Holevic in 1968 and finally described as a pure island flap (â€œkite flapâ€•) by Foucher and Braun in 1979 (8). Many variations of the flap have since been described and established in daily clinical use based on the consistent anatomy of the dorsal metacarpal arterial arcade (3 ,4 ,5 ,6 ,7 ,9 ,10 ,11 ,12 ,19 ,20). The majority of the flaps in recently published reports are raised from the dorsal aspect of the hand with only a few being found on the palmar surface (2 ,13 ,14).

The dorsum of the hand is supplied by the network of dorsal metacarpal arteries which are fed from the main forearm vessels: the radial, ulnar, and interosseus arteries. The first dorsal metacarpal artery (FDMA) consistently arises from the radial artery or the princeps pollicis artery and runs distally to the fascia of the interosseus muscle, frequently embedded in a fascial pocket. In most cases the first dorsal metacarpal artery divides in the middle of the second metacarpal into three terminal branches (18 ,22 ,24 ,30). The radial branch of the first dorsal metacarpal artery goes to the thumb, while the intermediate branch runs to the first web space. Distally the ulnar branch usually terminates at the level of the metacarpo-phalangeal joint and then arborizes into the dorsal skin of the index finger after giving off a perforating branch at the level of the metacarpal neck.

The exact location of the artery with respect to the first dorsal interosseous muscle may vary. Kuhlmann and de Frenne found the artery to run superficial to the muscle in 75% of specimens, while the FDMA was found within the muscle in 15% of specimens; in 10% of specimens the artery ran within and on top of the muscle during its course. Dautel and Merle identified a superficial course of the FDMA in 36% of specimens, a supra-fascial course in 23% of specimens, and a deep, intramuscular course in 56%. Despite its exact location all authors agree on the consistency of a FDMA (11 ,13 ,18 ,22 ,24 ,30) (Fig. 23-1).

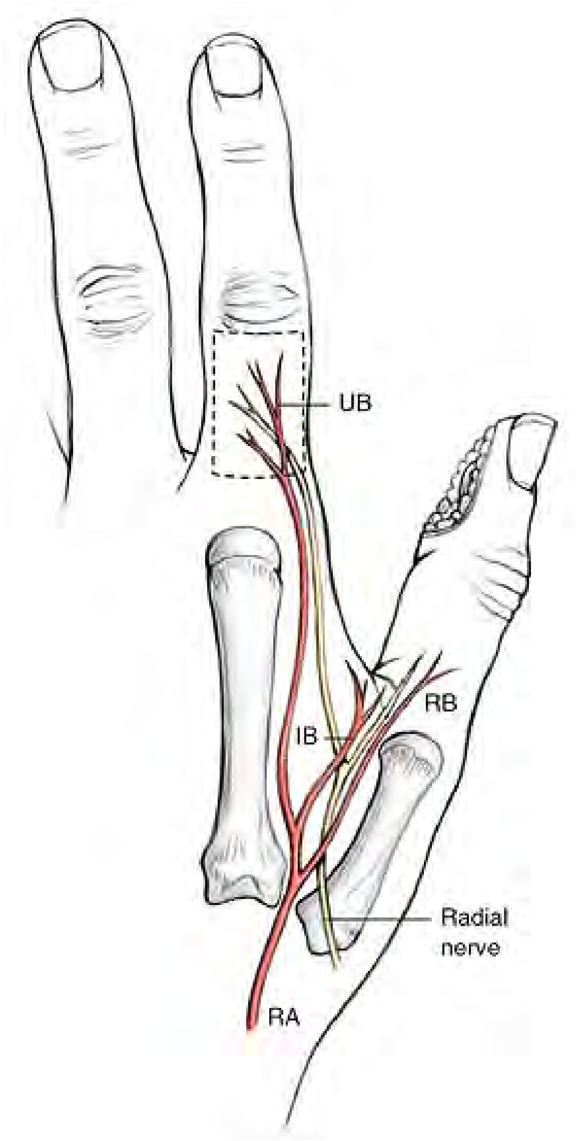

FIGURE 23-1 The three branches of the radial artery (*31*). The radial branch of the first dorsal metacarpal artery (*31*) goes to the thumb, the intermediate branch (*31*) to the first web space, and the ulnar branch (*31*) of the first dorsal metacarpal artery runs to the index finger.

Overall, the flap provides excellent sensibility, pliability, and stability for a variety of hand defects. We feel the first dorsal metacarpal island flap is one of the workhorses for reconstructive hand surgery (22 ,24 ,25 ,26 ,27).

Indications/Contraindications

Indications

Indications for the first dorsal metacarpal artery flap (â€œkite flapâ€•) are based on the following criteria:

- Etiology, location, size, and condition of the defect
- Availability of the flap (i.e., previous injuries that may compromise the vascular pedicle)
- Suitability of the flap for the defect (e.g., size, arc of rotation)
- Inclusion of other tissue components (i.e., tendon, bony segment)
- Patient's wishes (donor site morbidity, aesthetic appearance)
- Surgeon's familiarity with the technique

Based on these decision-making criteria, the flap is indicated for primary or secondary reconstructions of:

- The dorsal aspect of the thumb
- Restoration of sensibility with pulp reconstruction in the thumb
- Skin defects at the dorsal aspect of the hand within the arc of rotation of the antegrade kite flap
- Palmar defects of the index finger
- Defects of the web space

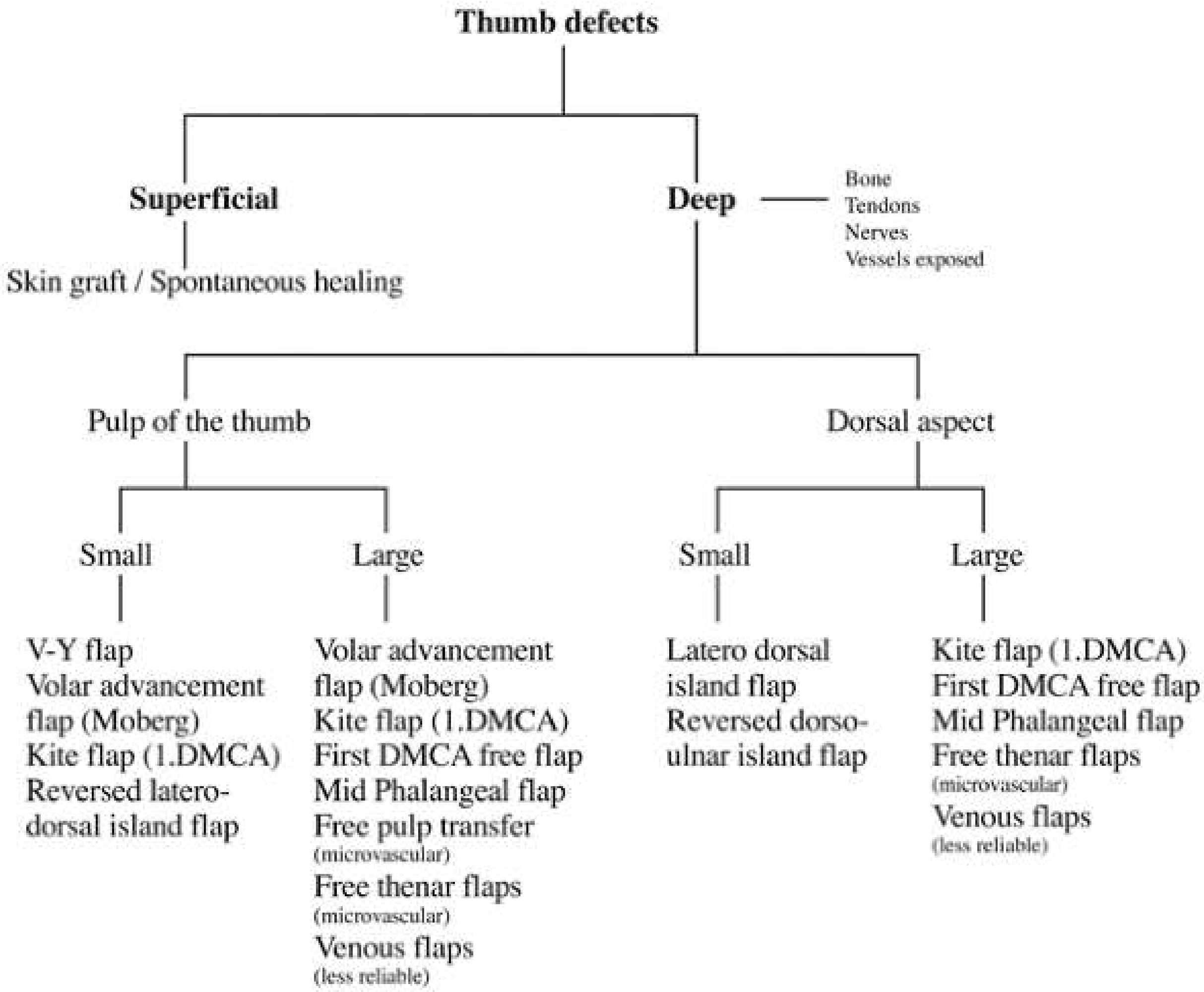
Thumb defects
Superficial
Skin graft / Spontaneous healing
Deep
Bone
Tendons
Nerves
Vessels exposed
Pulp of the thumb
Dorsal aspect
Small
Large
Small
Large
V-Y flap
Volar advancement flap (Moberg)
Kite flap (1.DMCA)
Reversed latero-dorsal island flap
Volar advancement flap (Moberg)
Kite flap (1.DMCA)
First DMCA free flap
Mid Phalangeal flap
Free pulp transfer (microvascular)
Free thenar flaps (microvascular)
Venous flaps (less reliable)
Latero dorsal island flap
Reversed dorso-ulnar island flap
Kite flap (1.DMCA)
First DMCA free flap
Mid Phalangeal flap
Free thenar flaps (microvascular)
Venous flaps (less reliable)

FIGURE 23-2 Decision-making algorithm for flap selection in treatment of thumb defects.

- Defects around the wrist
- Defects at the base of the thumb or the middle finger (â€œretrograde kite flapâ€•)
- Complex flaps including various tissue components
- Microvascular flap transfer based on the vascular pedicle in the snuff box

The list of indications demonstrates the versatility of the flap: however, reconstruction of the thumb takes precedence as the main indication in most centers.

Following the pioneer work of TrÃ¤nkle et al, the flap has replaced other sensate neurovascular island flaps for restoration of sensibility in the thumb (10 ,13 ,26 ,29) due to the dissection technique, the discriminative power that can be achieved, and the acceptable donor site morbidity. Based on this study, the flap is now widely used for this indication (Fig. 23-2).

Contraindications

Contraindications for flap use include acute trauma or a history of significant trauma to the donor site (15). In addition, injury mechanisms such as crush avulsions, wringer injuries, high-energy trauma (e.g., blast injuries or gun shot wounds) may limit the reliability of this flap if they have disrupted the course of the FDMA (10 ,13). If the tissue viability of the donor site cannot be estimated properly in the first hours of the trauma, wound closure should be delayed or a different flap should be used for wound coverage.

Burn injuries of the dorsum of the hand do not present a contraindication for the use of a first dorsal metacarpal artery flap in all cases. The dorsal metacarpal artery system is not damaged by deep partial-thickness burns that are excised and grafted. In 80% of patients with full-thickness burns, the dorsal metacarpal artery system is still intact. The potential for elevating a first dorsal metacarpal artery flap is therefore preserved after burn excision and grafting (11). Nevertheless, preoperative Doppler ultrasound examination is recommended before flap elevation.

Preoperative Planning

As in any surgical procedure, careful preoperative assessment and planning are key factors to success and the avoidance of complications or failure. Decision making follows an algorithmic approach in which the defect must be assessed for:

- Size
- Etiology
- Characteristics

The decision for flap selection is based on the criteria outlined previously (see indications). The patient should be thoroughly informed about the therapeutic options, and his or her expectations and wishes should also be discussed. Special attention should be paid to donor site morbidity including numbness over the radial aspect of the dorsum of the hand and the implications of a full-thickness skin graft to the donor site. The postoperative management and potential complications such as flap failure or impaired mobility of the fingers, impaired sensibility, and other possible undesirable results should be discussed with the patient. Donor-site morbidity is an important issue that should also be discussed with the patient before surgery. The functional and aesthetic morbidity of the donor site is an issue often neglected in the choice of a flap, and it must be borne in mind that the donor area of a first dorsal metacarpal artery island flap is in a very exposed position.

Clinical examination should exclude prior injuries in the area of the vascular pedicle. In most cases the artery is palpable, but Doppler ultrasound examination before flap elevation is mandatory (13) since the course of the artery is variable (see previous discussion). The course of the artery should be marked on the skin preoperatively after Doppler examination (13 ,19 ,26 ,27).

Surgery

The defect size is templated to the dorsum of the index finger. The course of the artery is marked on the dorsum of the hand with the aid of a hand-held Doppler ultrasound device. Surgery is performed using either an axillary bloc, or general or regional anesthesia. The use of a pneumatic tourniquet and loupe magnification are necessary for the dissection. The flap is harvested from the dorsum of the index finger and includes the first dorsal metacarpal artery, a branch of the superficial radial nerve, and at least one subcutaneous vein (Figure 23-3).

Dissection is performed starting with the vascular pedicle including a large subcutaneous vein which has first to be identified and mobilized. Elevation of the flap starts radially by incising the interosseus fascia at the most radial edge of the muscle. The fascia is then peeled off the muscle towards the second metacarpal. Thereby, a wide pedicle including the subcutaneous vein, a branch of the superficial radial nerve, and the vascular pedicle is created (24). A vessel loop is used to mark this pedicle. The pedicle is traced back to its origin in the snuff-box. There is no need to isolate the vascular pedicle from the interosseus fascia, since this may injure the delicate vessels. Transillumination will verify the inclusion of the artery and the small accompanying veins in the pedicle.

Dissection now proceeds from the distal border of the flap. The flap is incised and dissected off of the paratenon of the extensor apparatus toward the proximal border of the flap. Care must be taken to preserve the paratenon in order to secure graft-take at the donor site. A critical point in the dissection is the radial aspect of the extensor hood of the metaphalangeal joint where the vessels enter the subcutaneous network of the flap. It is recommended to include a small strip of extensor hood to secure the vessel entrance. This area can be easily repaired without any functional deficit by a 4.0 polydioxanone (PDS) suture.

Once the flap is raised, the tourniquet is deflated to assess the blood supply of the flap. Occasionally, the flap may first appear pale but usually begins to turn pink within a few minutes (26 ,27). The flap may be passed through a subcutaneous tunnel into the defect (21 ,24 ,27). If the skin is too tight, the tunnel is opened and the pedicle is skin grafted. As a sensory island flap, the first dorsal metacarpal artery island flap allows a wide arc of rotation due to its pedicle length of up to 7 cm (24).

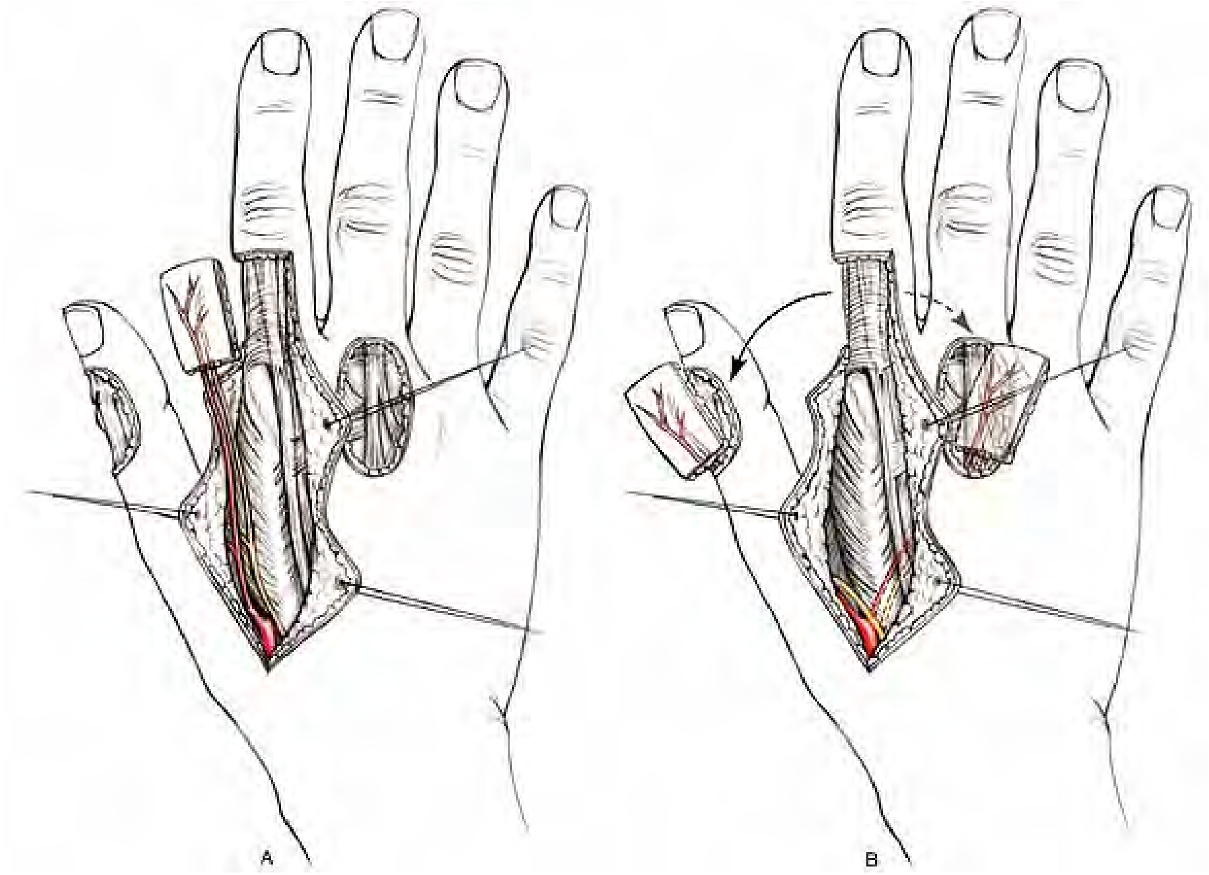

FIGURE 23-3 Schematic drawing of the dissected (A) first dorsal metacarpal artery flap and the first dorsal metacarpal artery flap with its pedicle passed distally under the skin tunnel (B) . (From Germann G, et al. *31* Berlin: Springer-Verlag; 2000. With permission.)

The donor defect of the index finger is covered with a full-thickness skin graft (19). A split-thickness skin graft may also be used; however a full-thickness skin graft will provide a better color match (23). Full-thickness grafting will provide an aesthetically acceptable appearance at the donor site within a shorter period of time, although the long-term results using a splitâ€“thickness graft are comparable (19) (Figs. 23-4 and 23-5).

Extensive experience with this flap has lead to some refinements. To avoid any tension on the pedicle, the subcutaneous tunnel has to be opened frequently. We now raise the flap constantly with a â€œcutaneous tailâ€• extending from the proximal border of the flap to the radial aspect of the second metacarpal. This extension allows opening of the tunnel and direct skin-skin closure after flap transfer. This technique does not only provide a more stable skin closure, but also a better aesthetic appearance (see Fig. 23-4C,F). In addition, segments of the second metacarpal or the extensor indicis proprius tendon can be included in the flap for more complex reconstruction.

The first dorsal metacarpal artery flap can also be raised as a microvascular free flap (10 ,12 ,13). The flap is harvested using the same technique as the pedicled first dorsal metacarpal artery flap described previously. Dissection of the FDMA back to its origin from the radial artery provides an artery of adequate diameter for microvascular anastomosis. A long pedicle length facilitates microvascular reconstruction outside the area of trauma; inclusion of the cutaneous nerve permits reconstruction of sensibility. We believe that dissection of this flap is easier and faster than a toe and first web-space flap (19) (Fig. 23-6).

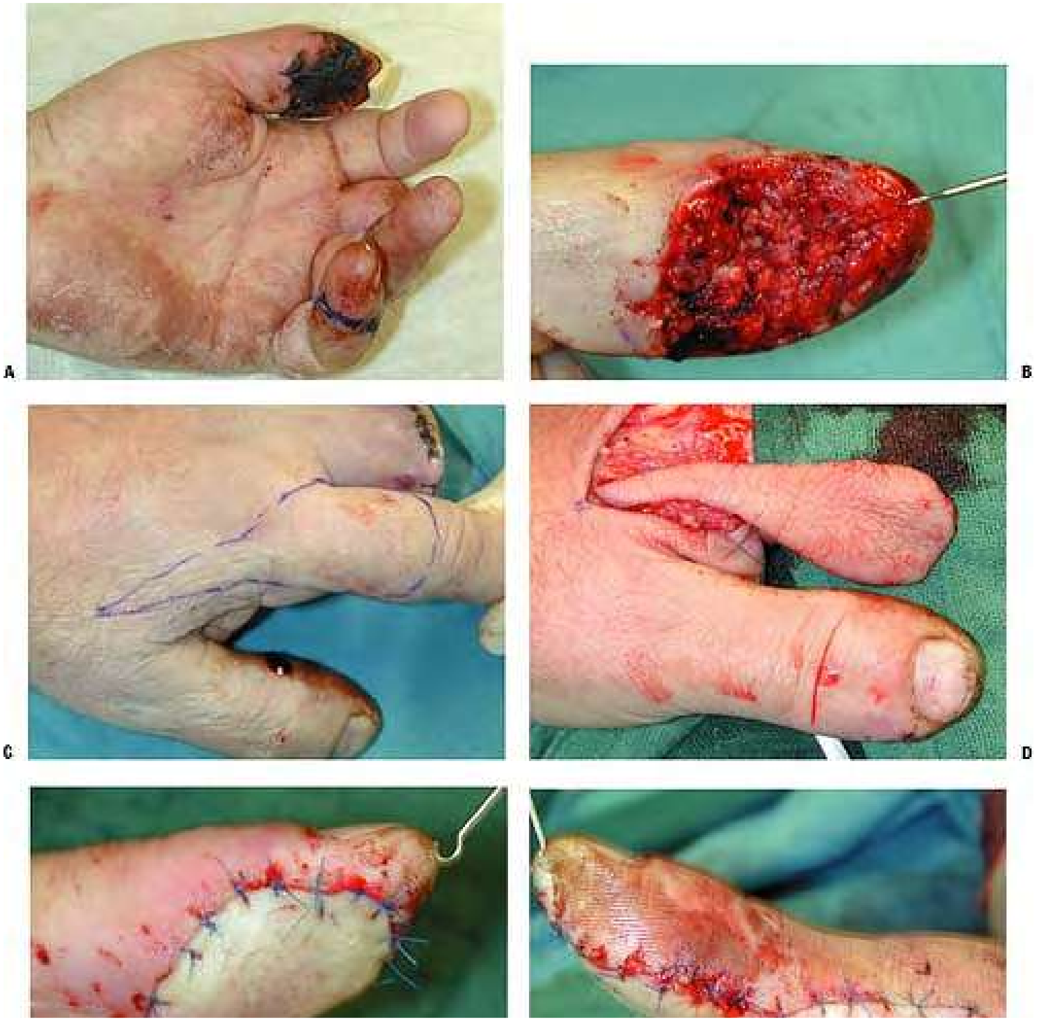

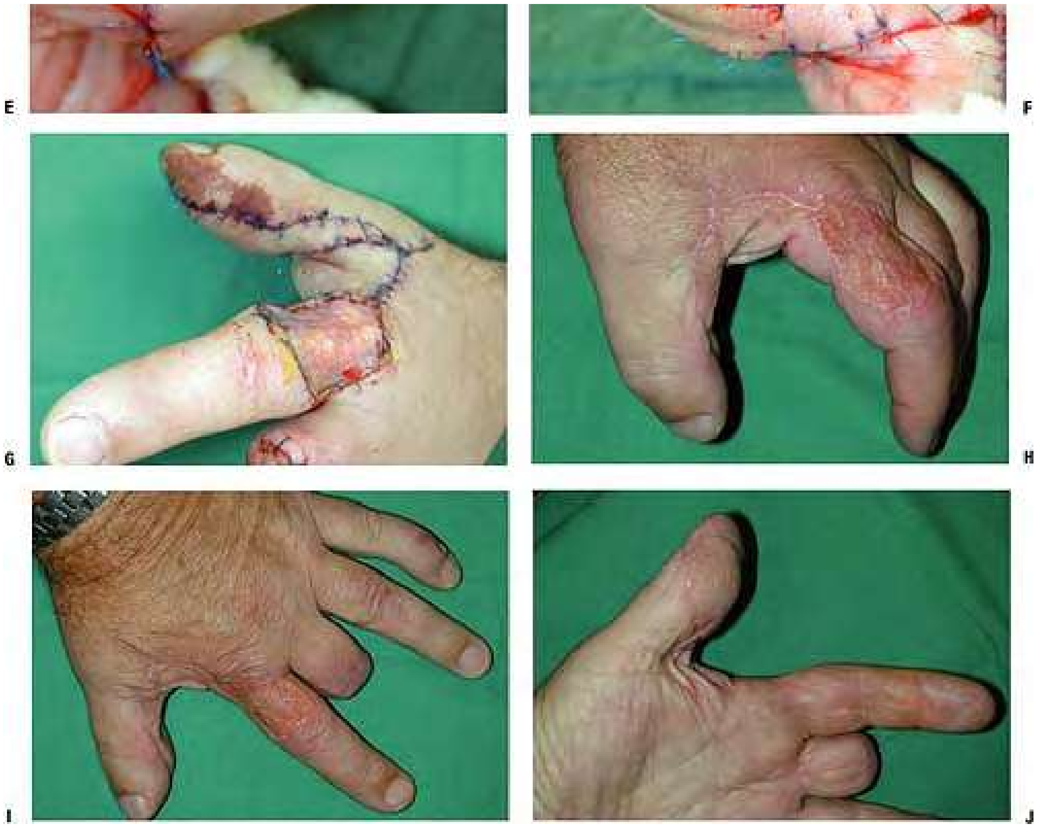

FIGURE 23-4 A: A 63-year-old man sustained a mutilating saw injury of the hand with avulsion of the palmar aspect of the distal phalanx of the left thumb (additional injuries in the middle and small fingers). Six days after the injury: complete necrosis of the pulp of the thumb. B: Intraoperative view after radical debridement. C: Flap design: the first dorsal metacarpal artery island flap is outlined over the dorsal aspect of the proximal phalanx of the index finger. D: The first dorsal metacarpal artery island flap is elevated. Note the â€œcutaneous tailâ€• of the flap that allows primary skin-skin closure. E,F: After tunneling the flap into the defect, primary skin-skin closure is achieved. E,F: After tunneling the flap into the defect, primary skin-skin closure is achieved. G: Reconstruction of the donor site with skin graft. Hâ€“J: Aesthetic and functional outcome 2 years postoperatively. Full extension of thumb and index finger. The patient describes cold intolerance of the donor site. Two PD of the flap: 6.5 mm; 2 PD of the donor site: 6 mm.

Postoperative Management

The hand may be immobilized in a bulky dressing following surgery. A palmar plaster splint may be used to immobilize any concomitant fractures. A window is left in the dressing for evaluation of the flap. Care is taken not to apply the dressing too tight for fear of causing arterial or venous compromise. Forty-eight to 72 hours after surgery, the flap's vascularity is ensured, and (13) physical therapy may be initiated if underlying fracture fixation is stabile. The index finger metacarpalphalangeal joint is usually immobilized for 5 to 7 days to ensure take of the skin graft over the donor site, but the proximal and distal interphalangeal joints may be mobilized as soon as possible to avoid postoperative stiffness. Occupational therapy is also initiated for sensory re-education of the flap. Initially, sensation in the flap will be perceived as coming from the dorsum of the index finger, or a dual sensation phenomenon may result. Cortical reorientation of sensation is possible with a structured sensory re-education program. The first dorsal metacarpal artery island flap allows immediate restoration of sensibility even in older patients, among whom neuronal co-adaptation of a pedicled or a free flap yields poorer results than in younger people (26 ,27).

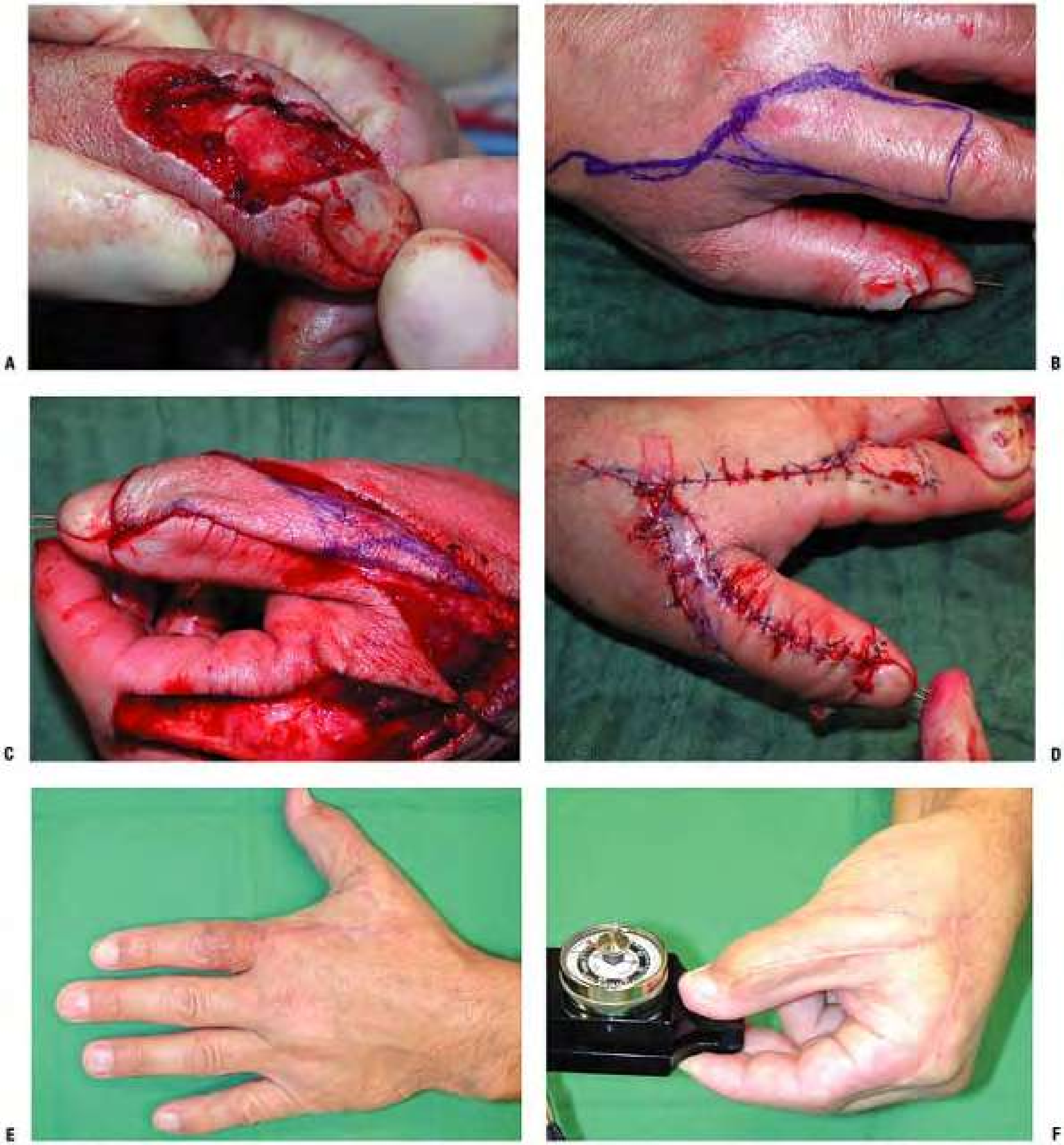

FIGURE 23-5 A: A 44-year-old man suffered a work-related accident with a chain saw, resulting in a skin defect including a defect of the extensor pollicis longus tendon at its insertion and exposure of the interphalangeal joint. B: Design of the first dorsal metacarpal artery island flap on the dorsal of the index finger. C: The first dorsal metacarpal artery island flap prior to insertion into the defect. D: Result after insetting of the first dorsal metacarpal island flap. The flap shows perfect capillary refill. The donor site has been diminished in size by mobilization of the surrounding skin. Aesthetic

(E) and functional (F) outcome 2 years postoperatively. Excellent hand function and aesthetic appearance have been achieved. The patient returned to his previous employment without limitations.

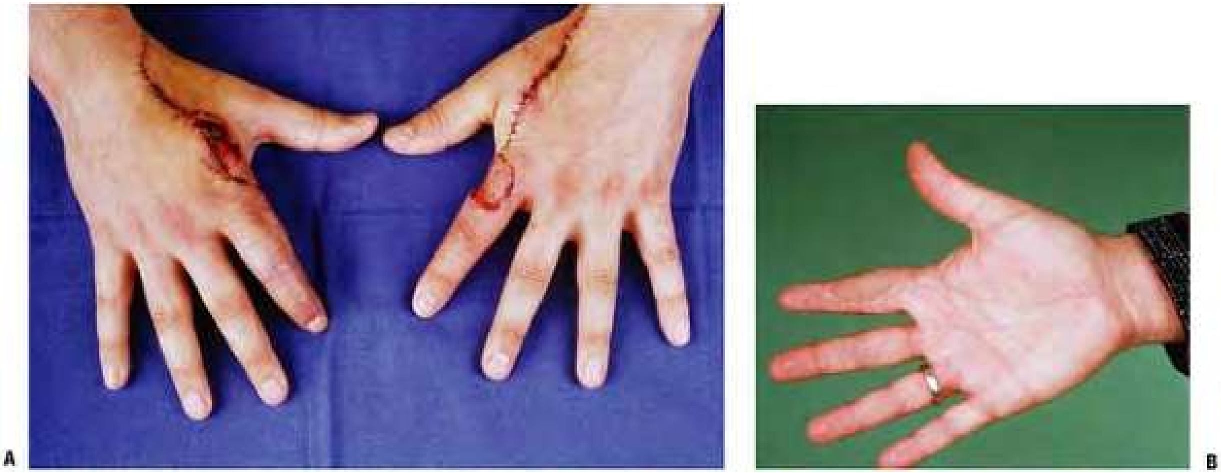

FIGURE 23-6 A 22-year-old carpenter sustained an occupational high pressure-injection injury of the palmar aspect of the proximal interphalangeal joint of the index finger of the non-dominant right hand. A soft-tissue defect resulted, which included a large portion of the palmar aspect of the finger and exposure of both flexor tendons and neurovascular bundles. This was successfully reconstructed with a free kite flap (A) . The flap healed uneventfully with excellent aesthetic and functional result. Follow-up after 40 months (B) demonstrated a static two-point discrimination of 15 mm. Grip strength values were excellent. Time off work was 11 weeks.

Complications

Although the first dorsal metacarpal artery island flap has proven to be very reliable, various complications can occur. These complications include planning errors, complications during the operation, and postoperative difficulties. To avoid an inadequate arc of rotation or an inadequate flap size, the surgeon can add 10% to 15% more length to the pedicle, and the flap can be planned and harvested 10% to 15% larger in size (10 ,13).

Care should be taken to exert only gentle traction when passing the flap through the subcutaneous tunnel to close the thumb pulp defect. Excessive tension can cause a partial compression of the vessels, and marginal necrosis could result. Whereas vascular insufficiency on the arterial side is frequently the result of tunnelling, a short pedicle or tight wound closure can compromise the venous outflow. In case of postoperative vascular insufficiency, sutures can be released and leech therapy may be employed.

Donor-site morbidity of dorsal metacarpal artery island flaps is low compared with axial homodigital island flaps. Neuromas are rare, and the patients report less hypersensitivity than in homodigital island flaps (13). However, one of the most important donor-site problems of the first dorsal metacarpal artery island flap is a diminished protective sensation over the transplanted skin graft, and patients may complain of cold intolerance. Hypertrophic scars present a minor donor-site problem. Functional limitation because of the donor site is negligible, and the range of motion of the donor finger reaches approximately 95% of the opposite index finger (26 ,27).

Recommended Reading

Adani R, Busa R, Castagnetti C, Bathia A, Caroli A. Homodigital neurovascular island flaps with â€œdirect flowâ€• vascularization. *Ann Plast Surg.* 1997;38(1):36â€“40.

Adani R, Squarzina PB, Castagnetti C, Lagana A, Pancaldi G, Caroli A. A comparative study of the heterodigital neurovascular island flap in thumb reconstruction, with and without nerve reconnection. *J Hand Surg.* 1994;19(5):552â€“559.

Dellon AL. Sensory recovery in replanted digits and transplanted toes. *J Reconstr Microsurg.* 1986;2:123.

Early MJ, Milner RH. Dorsal metacarpal flaps. *Br J Plast Surg.* 1987;40:333â€“341.

Gebhard B, Meissl G. An extended first dorsal metacarpal artery neurovascular island flap. *J Hand Surg.* 1995;20(4):529â€“531.

Grossman JAI, Robotti EB. The use of split-thickness hypothenar grafts for coverage

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part IV - Soft Tissue Coverage of the Fingers Following Trauma > Chapter 24 - Thumb Coverage

---

Chapter 24

Thumb Coverage

David Elliot

The primary goals of thumb reconstruction are to restore length and sensation. A thumb should be of adequate length to allow for pinch to the index and middle fingertips and to participate in span grasp with all five digital tips. Adequate length is also essential for developing power grip. The surgeon aims to provide adequate coverage of the deeper, vital structures with soft tissue and skin of good quality. Sensibility of the tip is of particular importance, being crucial to fine pinch. A thumb â€œpostâ€• with no or little movement distal to the basal joint, and/or poor tip shape and sensation, can still work to aid in hand function but will not allow for finesse of function.

In attempting to obtain optimum reconstruction one should avoid shortening the thumb unless absolutely necessary. We encourage distal replantation or composite graft replacement, and also look to homodigital flap reconstruction whenever possible. Care should be spent in nail bed preservation and repair. Any nail greater in length than one third of its normal length should be retained, at least at primary surgery. Reconstruction of the thumb defects with the use of skin grafts from elsewhere on the body can create patches of dissimilar color, texture, and thickness and a noticeably poor cosmetic result. Finally, with respect to mobility and rehabilitation, the time-honored plastic surgical techniques of skin cross-finger, thenar, groin, and cross-arm flaps should be avoided, if possible, as they tether the injured thumb to another part of the body, thus preventing adequate therapy, which may promote edema and stiffness.

Homodigital Flap Reconstruction

Homodigital flap reconstruction involves rearrangement of the soft tissues of the injured digit to achieve healing without seeking tissue for reconstruction from outside that digit. This concept of reconstruction has definite advantages for thumb reconstruction. In particular, it reconstructs â€œlike with likeâ€• and avoids the creation of further scarring and morbidity elsewhere on the hand or body. While homodigital reconstruction is advantageous in these respects, the availability of donor tissues within the thumb is obviously limited.

Perhaps the single most important surgical pearl to remember when using homodigital reconstruction is that the thumb and fingers have an astonishing ability to close skin defects of considerable size by a combination of wound contraction and re-epithelialization (Fig. 24-1). Thus, if a portion of the wound can heal under moist antiseptic dressings through secondary intention, the surgeon may use the homodigital flaps to cover only those portions of the wound which contain exposed vital structures or to improve soft tissue coverage over bone. The flap does not have to cover the entire defect, only the essential components of the defect.

Preoperative Planning

Preoperative planning should include AP and lateral radiographs in addition to a thorough hand exam. Concomitant arterial and nerve injuries can often be identified prior to exploration within the operating room. Many thumb injuries will occur in conjunction with other hand injuries and revascularization of injured digits takes precedence over coverage issues. Wounds should be debrided of all necrotic tissue and free of infection before embarking on soft tissue coverage.

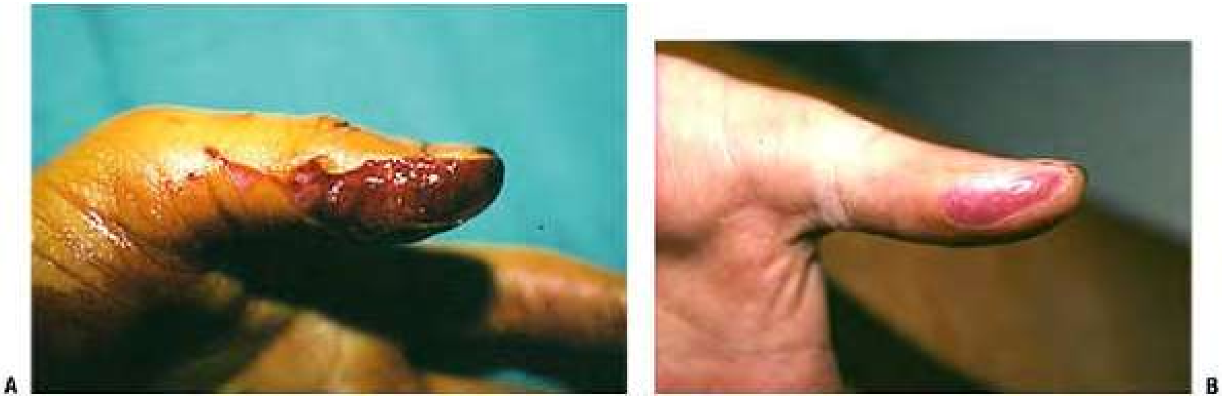

FIGURE 24-1 A: Industrial scalpel wound of the thumb tip. B: Same digital surface after healing by re-epithelialization under moist antiseptic dressings for 3 weeks. (Reproduced from Elliot D. Specific flaps for the thumb. *Tech Hand Up Extrem Surg.* 2004;8(4):198â€“211, with permission.)

Patient Positioning

Patients are positioned supine on the operating room table. General or regional anaesthesia may be used during the surgical procedures. In some cases, thumb or wrist blocks may be used when performing smaller flaps or for debridement procedures. An upper arm tourniquet is placed prior to prepping and draping the patient. Surgery is always performed under tourniquet control to aid in visualization of vital structures.

The following sections present our preferred techniques for thumb reconstruction based on location of the original defect.

Palmar Defects

Neurovascular Tranquilli-Leali or Atasoy-Kleinert Flap

Indications/Contraindications

Historically, the original Tranquilli-Leali or Atasoy-Kleinert flap has been described to cover partial amputation defects of the distal phalanx; however, these flaps do not work well on the thumb because of the inflexibility of the subcutaneous soft tissues. Laterally based single pedicle flaps, vascularized in the same way by the small vessels beyond the trifurcations of the digital arteries, also move poorly. The *neurovascular* Tranquilli-Leali *or* Atasoy-Kleinert flap is designed to be much larger and is islanded on both neurovascular pedicles bilaterally (Fig. 24-2). The flap is designed to extend to, or across, the IP joint crease proximally. This flap works well for minor transverse or oblique injuries of the distal phalanx with exposure of bone.

This flap can be used for stump reconstruction of any length of amputated thumb and is more useful than the original Tranquilli-Leali flap; however, it moves less freely than on a finger, partly because of the fibrous nature of the subcutaneous tissues of the thumb and partly because of what has been described as the â€œvertical dimensionâ€• of the thumb, that is its palmar-dorsal width at the tip (1). On fingers, it can be used to reconstruct defects with a palmar slope of up to 30 degrees; on the thumb it can only reconstruct defects with bone exposure which are dorsally facing, transverse, or palmar facing with less than 10 degrees of slope (Fig. 24-3).

Surgical Technique

The incisions of the V cross the interphalangeal joint crease at an angle and thus do not cause contractures (see Figs. 24-2 and 24-3). When designing the flap, one takes the V incisions out almost to the lateral nail folds distally. Having made the flap wide, the leading edge of the flap after advancement is wider than the original thumb tip. Unless the lateral corners of the flap are excised, this results in a spatulate end to the digit. Cutting off the lateral corners and allowing the resulting raw edges and tip to epithelialize not only narrows the digital tip but also rounds it to achieve a good appearance. The flap is designed as a â€œVâ€• at its proximal extremity and was conceived to close proximally as a â€œYâ€• after the flap has moved distally. Mostly, the proximal donor defect is left open to close under dressings as primary closure of the vertical limb of the â€œYâ€• tightens the thumb too much proximally over the vascular pedicles of the flaps and compromises vascular inflow into the flap.

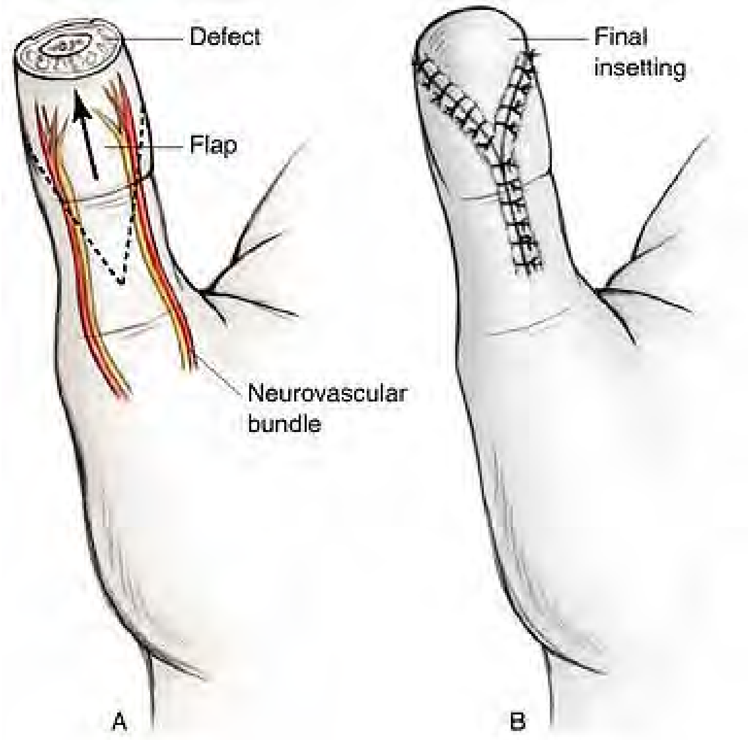
Defect
Flap
Neurovascular bundle
A
Final insetting
B

FIGURE 24-2 A,B: Illustration of the neurovascular Tranquilli-Leali flap. The flap extends proximally beyond the IP crease and includes both the neurovascular bundles.

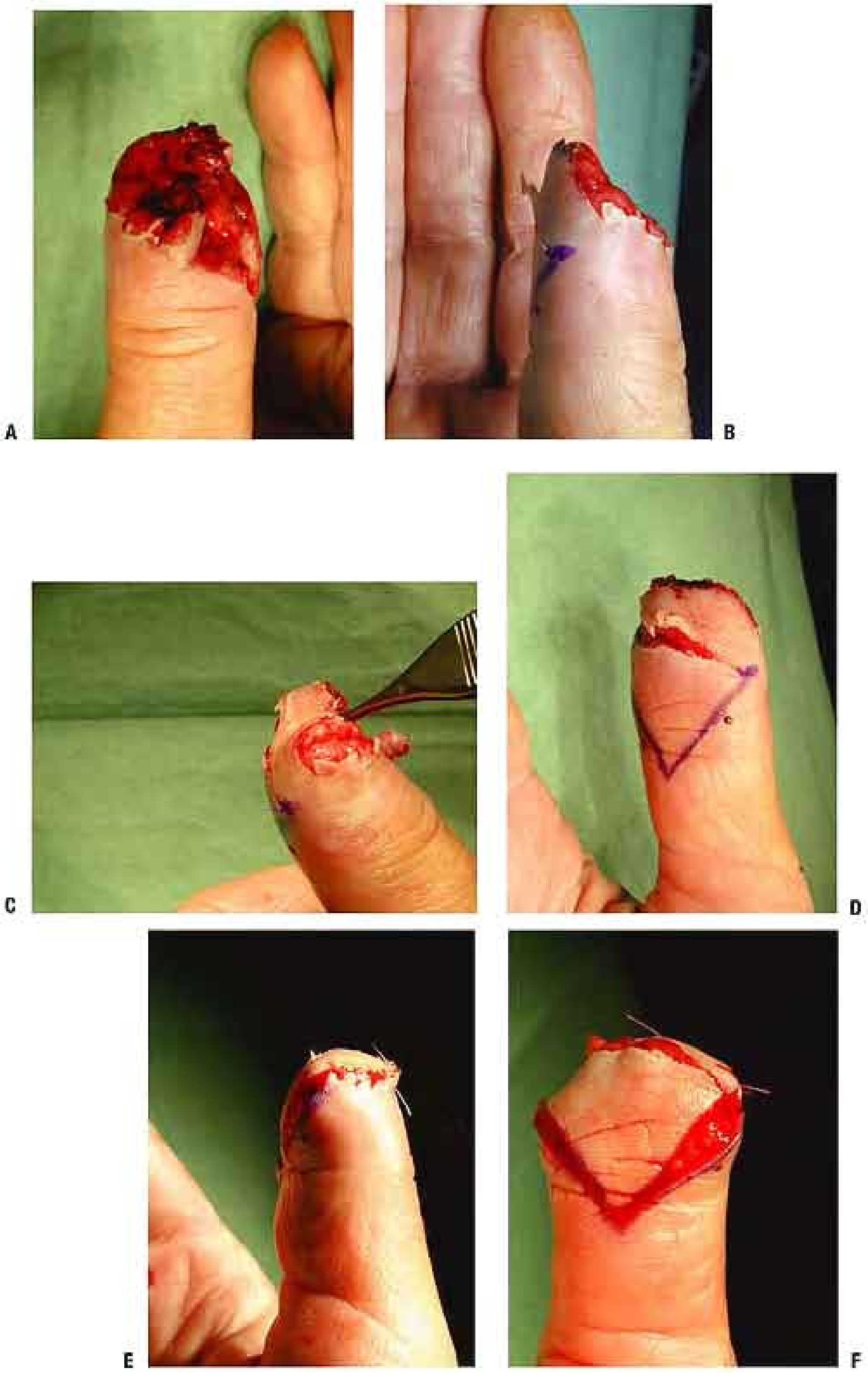
A
B
C
D
E
F

FIGURE 24-3 A,B: Dorsal sloping crush amputation of the right thumb tip of a 70-year-old man. C:Simple folding back of the palmar tissues after excision of the nail germinal matrix to suture the skin to the proximal nail fold is not possible because of the bulk of the pulp tissue. D: A neurovascular Tranquilli-Leali flap designed to allow easier folding back of the palmar tissues and shaping of the thumb tip. The transverse skin split just beyond the â€œVâ€• was part of the injury, but only skin deep, so not affecting the vascularity of the tissues distal to it. E,F: The reconstructed shape after trimming the lateral points of the advanced flap to round the thumb tip. (Reproduced from Elliot D. Specific flaps for the thumb. *Tech Hand Up Extrem Surg.* 2004;8(4):198â€“211, with permission.)

Moberg Flaps

Indications/contraindications

With greater losses of palmar thumb tissue, more sloping palmar oblique injuries, and when the whole distal pulp of the thumb has been avulsed, the neurovascular Tranquilli-Leali flap is too small and cannot advance sufficiently. For larger thumb defects confined to the distal phalanx we favor the use of a modified Moberg flap. The Moberg flap, as described by Moberg in 1964, is an advancement flap based on both neurovascular bundles of the thumb (2). The advantages of this flap are that it returns sensation and glabrous skin to the defect site. The downside to the original flap design was the IP joint flexion contracture necessary to achieve suture of the flap distally to the nail. Our preferred modification of the Moberg flap avoids the IP joint flexion necessary with the original to advance the flap to the tip of the digit. This is, essentially, Oâ€™Brien's modification of the Moberg flap, but uses a â€œVâ€• tail proximally instead of a skin graft at the base of the thumb (Fig. 24-4) (3). This V-Y Moberg flap achieves the same excellent results as the original Moberg flap in terms of sensibility of the thumb tip, but without the restrictions of interphalangeal joint movement associated with the original flap or the need for skin graft of the Oâ€™Brien modification.

There is no danger of dorsal skin loss in raising Moberg-type flaps on the thumb, as the dorsal skin of the thumb has a separate blood supply. Because of the safety of the V-Y Moberg flap, the speed and ease with which it can be dissected, and the excellent sensibility of the thumb tip which can be achieved with it, this technique remains our favorite thumb tip reconstruction if the defect is limited to the distal phalangeal segment of the thumb.

Technique

The flap is created laterally along the mid-lateral lines on each side of the thumb and with the tip of the proximal â€œVâ€• well back on the thenar eminence, at least as far proximally as a line drawn proximally from the ulnar border of the middle finger. This creates a large flap which incorporates the more lax subcutaneous tissues of the thenar eminence. Smaller flaps with the proximal â€œVâ€• at the base of the thumb move less easily and less far distally. The wider and longer flap avoids any need for addition of techniques to accommodate tightness at the base of the thumb. The use of this flap is considerably extended in reconstructing palmar oblique defects if the pulp distally which is denuded of skin cover remains as the leading edge of the flap, with suture of subcutaneous tissue, not skin, to the nail distally then epithelialization under dressings (Fig. 24-4).

Other Flaps

Other flaps have an infrequent use in reconstruction of the tip and palmar surface of the thumb but can be useful.

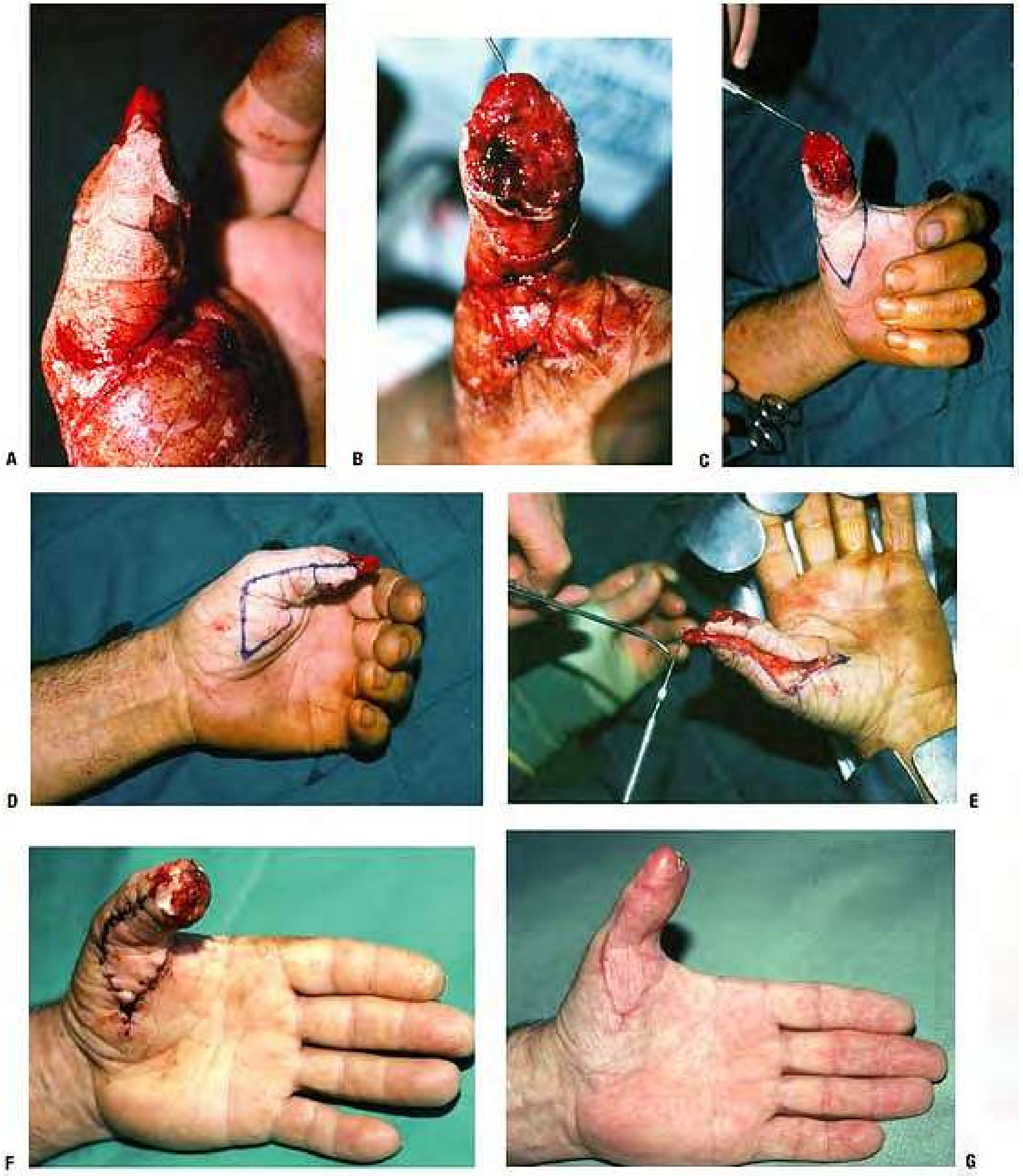

FIGURE 24-4 A,B: Preoperative views of the thumb of a 47-year-old man with a typical crush avulsion of the distal pulp. C,D: Markings of a V-Y modification of the Moberg flap shown preoperatively. E: Intra-operative view of the fully mobilized flap. F: Immediate postoperative view with the flap advanced to provide pulp cover of the distal bone. G: Final result after epithelialization of the tip under moist antiseptic dressings. (Reproduced from Elliot D, Yii NW.

Homodigital reconstruction of the digitsâ€”the perspective of one unit. *Handchir Mikrochir Plast Chir.* 2001;33(1):7â€“19, with permission.)

Lateral Pulp Flap

We devised the Lateral Pulp Flap for losses of the radial border of the tip of the index finger with exposure of bone but it is also useful on either border of the thumb tip (Fig. 24-5). This flap exploits the excess of pulp in the digital tip.

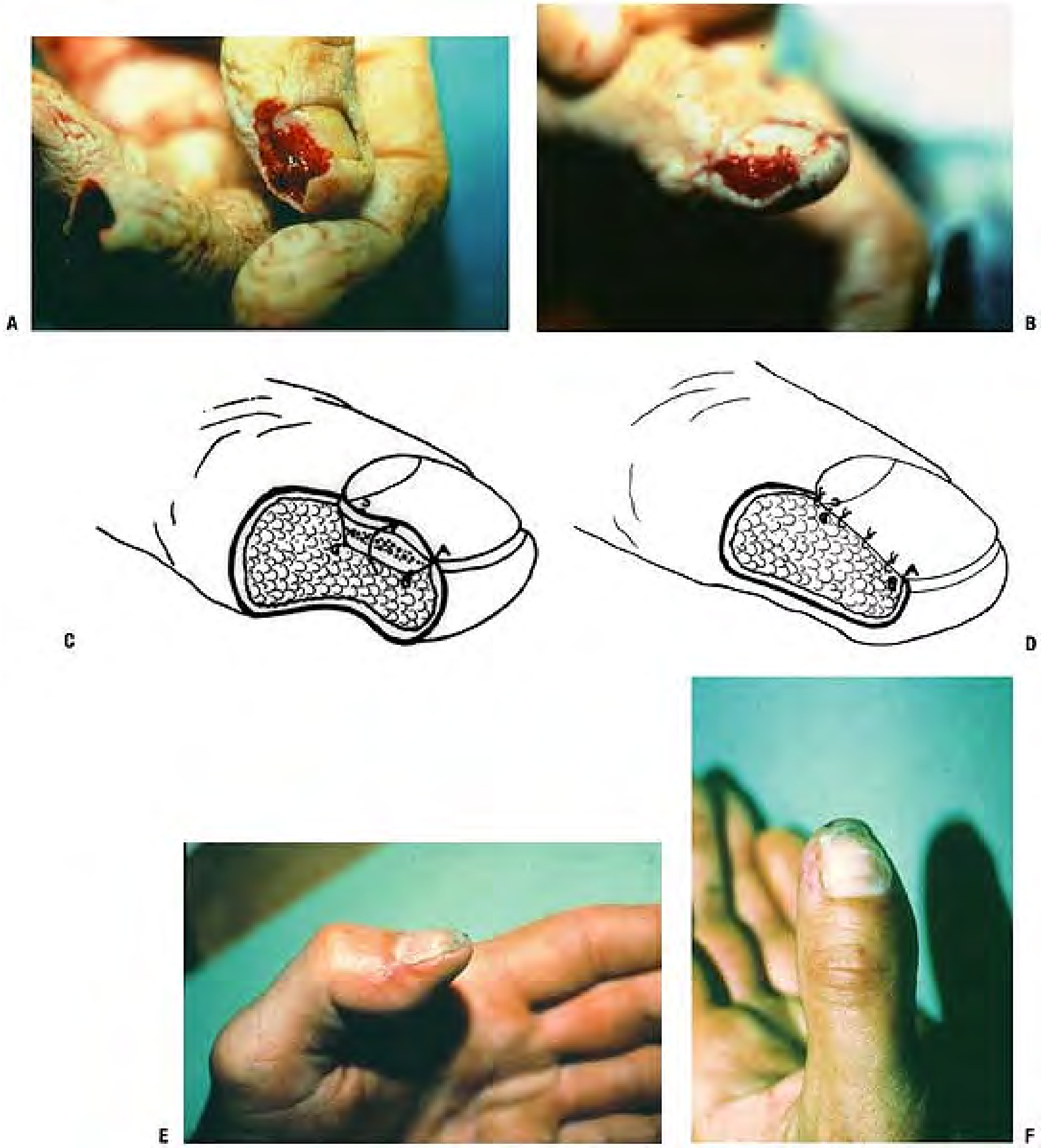

FIGURE 24-5 A: A slicing defect with loss of the radial lateral pulp, lateral nail fold, and lateral one-third of the nail of the thumb in a 50-year-old man. B: Lateral pulp transposed to cover the distal phalanx. C,D: Diagrams to show the original injury and transposition of the pulp to cover the bone of the distal phalanx. E,F: Late views of the reconstruction, 1 year after the injury. (Parts C and D reproduced from Elliot D, Jigjinni VS. The lateral pulp flap. *J Hand Surg Br.* 1993;18(4):423â€“426, with permission.)

The flap is raised by opening the tip of the digit with a fish mouth incision close to the distal nailbed and freeing the pulp attachments to the bone. The pulp then moves laterally and is lifted over the bone and sutured to the edge of the nail. The deep edge of the pulpâ€”not the superficial edgeâ€”is brought up to the nailbed to cover the bone. The pulp is then epithelialized under moist antiseptic dressings. This reconstruction creates a digital tip which is sensate but has no lateral nail fold. This seems to cause no functional problems and is less obvious than most lateral nail fold reconstructions, as these are usually too bulky.

Side-to-Side Homodigital Switch Flaps

Side-to-side homodigital switch flaps, which reconstruct or re-innervate one side of a digital tip at the expense of the other side, can be useful to resurface areas of pulp sensibility in the hand which are critical for pinch activity after localized loss of tissue or irrevocable digital nerve injury. Replacement of the radial side of the index or middle fingertips by a vascularized composite transfer from the ulnar side of the same digit is well established, both as a simple transposition and as an island flap. Use of the radial pulp of the thumb tip to reconstruct and/or reinnervate the ulnar side is logical, as the ulnar portion of the thumb tip is most involved in majority of gross and fine pinch (Fig. 24-6).

Littler Flaps

Indications/Contraindications

The classic Littler flap, as described by Littler in 1956, was designed for thumb tip and palmar thumb resurfacing. Unlike the flaps described above, the Littler flap is a heterodigital flap, which means a donor site defect will be created on the middle or ring finger. The flap utilizes tissue from either the ulnar border of the ring or middle finger; this tissue is transferred on its neurovascular bundle, thus allowing for the return of sensation to the thumb (Fig. 24-7). This differs from the heterodigital flaps described elsewhere in this text, which are designed to spare the digital nerves. This flap is easy to dissect and transfer. However, the flap can considerably downgrade the donor finger function and restoration of adequate two-point sensation after transfer has been a problem and suffers problems of achieving â€œthumb-tipâ€• sensibility. To improve thumb- tip sensation, and avoid patient's perception of the thumb-tip as still coming from the donor finger, division of the donor finger nerves and reconnection of the nerves in the flap to those of the recipient thumb has been recommended. Nerve disconnection and reconnection extends its usefulness in younger patients but, like all neurorrhaphies, gives less satisfactory tip innervation in the middle-aged and elderly patients.

Direct contraindications for this flap include injury to the neurovascular pedicle to the donor finger. Laceration of the superficial arterial arch or common vascular bundle to the donor finger could potentially lead to compromise of the flap or donor finger following transfer. Very distal tip injuries may also be difficult to cover with this flap.

Complications

Specific complications to be noted with the use of the Littler flap include the risk of venous congestion, specifically if the nerve is dissected or divided in preparation for neurorrhaphy at the level of the thumb. In such cases leech therapy may be required for resolution of venous congestion. Poor return of thumb sensation may also occur or the inability of the patient to recorticate, in which case the sensation within the flap is always perceived by the patient as coming from the donor finger donor site. Either sensory problem may necessitate a secondary nerve graft or nerve repair, but two-point discrimination may well remain <12 mm within the flap.

Reconstruction of the Dorsal Surface of the Thumb

The Hatchet Flaps

Indications/Contraindications

For the more common, smaller defects of the dorsum of the thumb, homodigital techniques are very useful. They are generally simpler and quicker to execute and avoid injury of adjacent fingers. For these small defects the Hatchet Flap provides an easy reconstructive option. Hatchet Flaps are also very useful in replantation, as they allow one to advance the dorsal veins of the proximal part of the thumb to the replant and achieve direct vein anastomoses under good skin cover without recourse to either vein grafts or a separate skin cover procedure. These flaps can be used as far distally as the proximal nail fold (Fig. 24-8).

Surgical Technique

The flap is designed as a random dorsal cutaneous flap (Fig. 24-8). A curvilinear incision is made over the radial (Fig. 24-9) or ulnar (Fig. 24-8) aspect of the thumb. A broad base is preserved on the opposite side of the digit. As on the palmar surface when utilizing a Moberg flap, a V-Y tail can be added to these flaps to avoid the need to skin graft the donor site proximally (Fig. 24-9) (4). On this surface of the digits, it is necessary to close the proximal V as a Y to protect the extensor paratenon from desiccation.

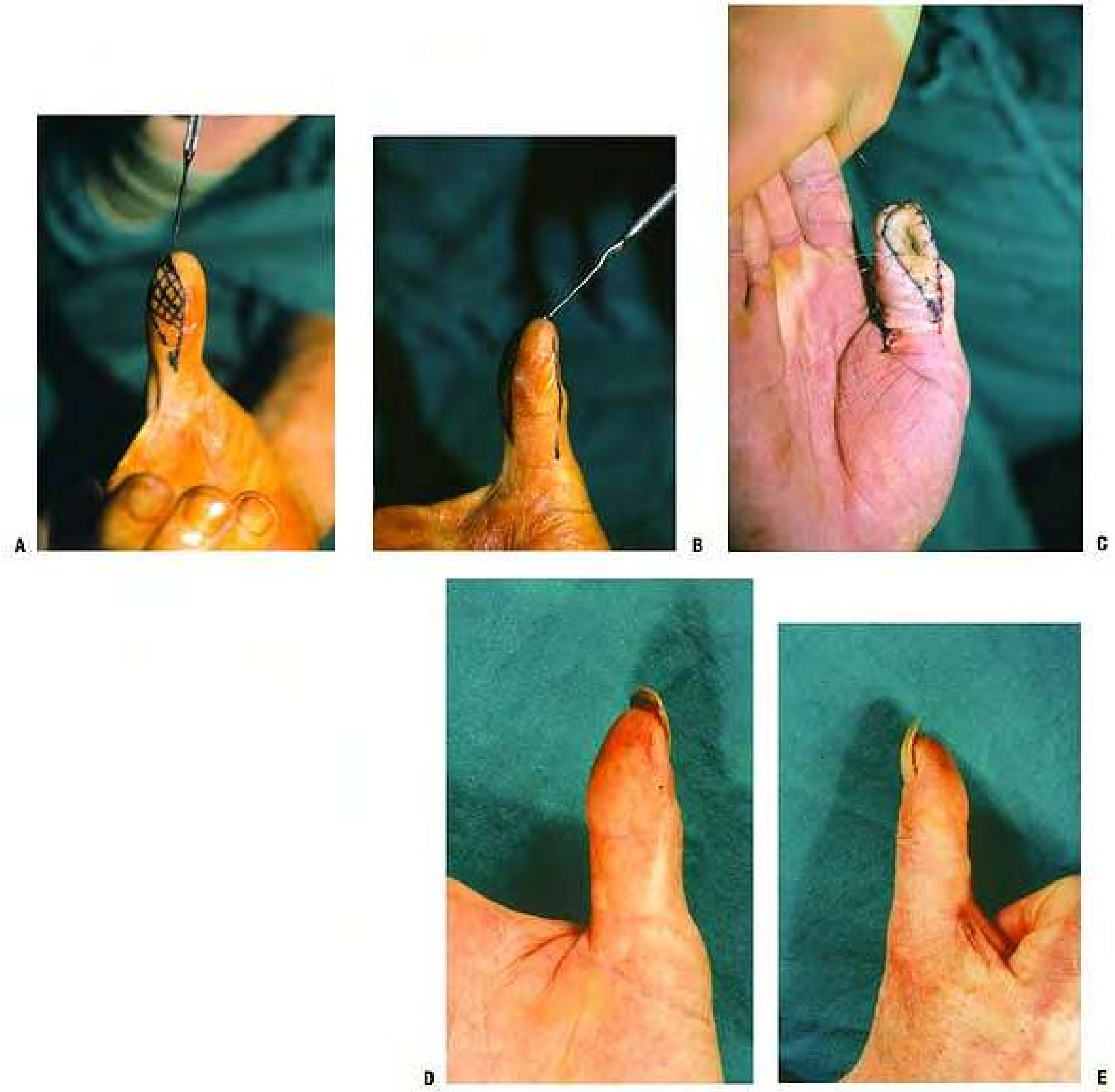

FIGURE 24-6 A: Thumb of a patient with irreparable damage to the ulnar digital nerve. A triangle of the denervated ulnar digital nerve territory is shaded, prior to excision down to bone. B: The same thumb showing the pre-operative marking of the radial switch flap. C: The thumb at surgery, after transposition of the flap and with the graft applied to the donor defect. D: Late view of the grafted donor site. E: Late view of the switch flap. (Reproduced from Elliot D, Southgate CM, Staiano JJ. A homodigital switch flap to restore sensation to the ulnar border of the thumb tip. *J Hand Surg Br.* 2003; 28(5):409â€“413, with permission)

Results

In a review of 1,077 dorsal wounds of all digits, which we treated over a 6-year period, 154 digits required flap reconstruction (5). The dorsal V-Y flap accounted for 42% of the flaps used and was the most common skin flap used on this surface of the digits, reflecting the relative incidence of defects of this size and shape.

Brunelli Flap

Surgical Technique

This homodigital flap is a useful adjunct for reconstructing smaller defects of the dorsum of the thumb. In 1993, Brunelli and his colleagues in Paris described a very easily and quickly raised flap based on the dorsoulnar artery of the thumb which is useful for reconstruction of dorsal defects of the tip of the thumb involving all or parts of the nail complex (Fig. 24-10) (6). The artery may be verified preoperatively with a handheld Doppler probe, but is very consistent running on the ulnar aspect of the extensor tendons. The flap is raised with a skin tail to reduce the risk of compression of the pedicle just proximal and lateral to the nail fold, which is the tightest part of its course. The flap may be safely elevated to the proximal margin of the interphalangeal joint skin crease. The donor site may be closed with a skin graft or advancement flap from the dorsum of the hand (Fig. 24-10).

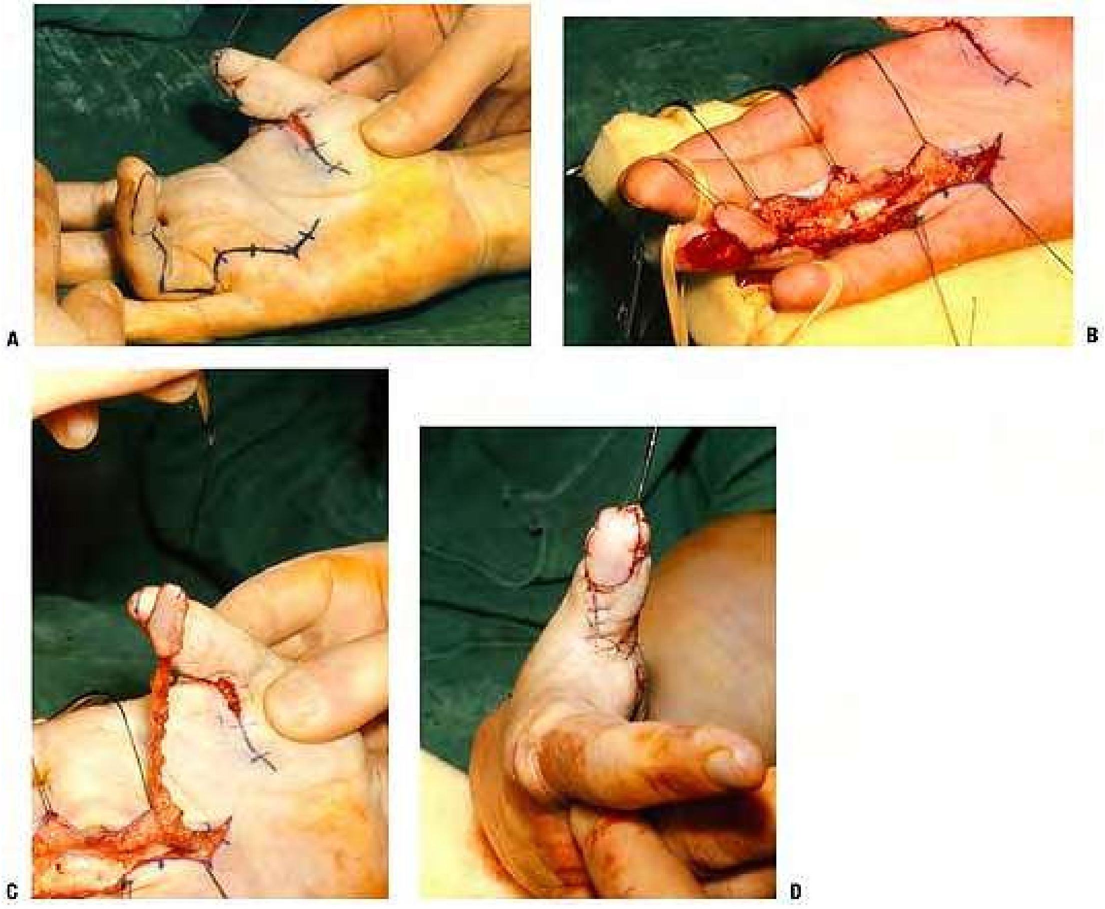

FIGURE 24-7 A: Intra-operative view showing dissection of damaged and denervated skin on the ulnar aspect of the thumb tip and the markings of a Littler flap on the ulnar side of the ring finger. B: Dissection of the pedicle of the flap. C: Demonstration of the transfer. D: Final view of the flap after tunnelling across the palm and insetting.

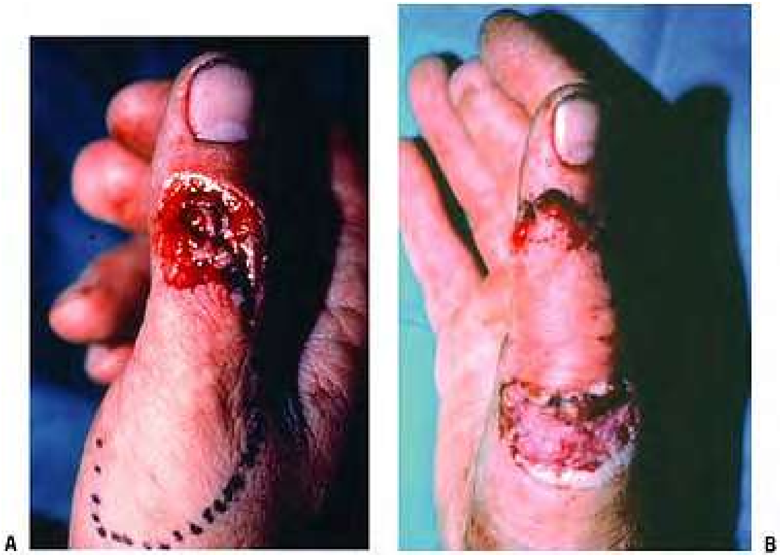

FIGURE 24-8 A: Preoperative view of a typical â€œsmallâ€• skin defect on the dorsum of the thumb. A simple dorsal hatchet flap is marked. B: Closure of the defect with this flap, with split skin graft reconstruction of the donor defect. (Reproduced from Elliot D. Specific flaps for the thumb. *Tech Hand Up Extrem Surg.* 2004;8(4):198â€“211, with permission.)

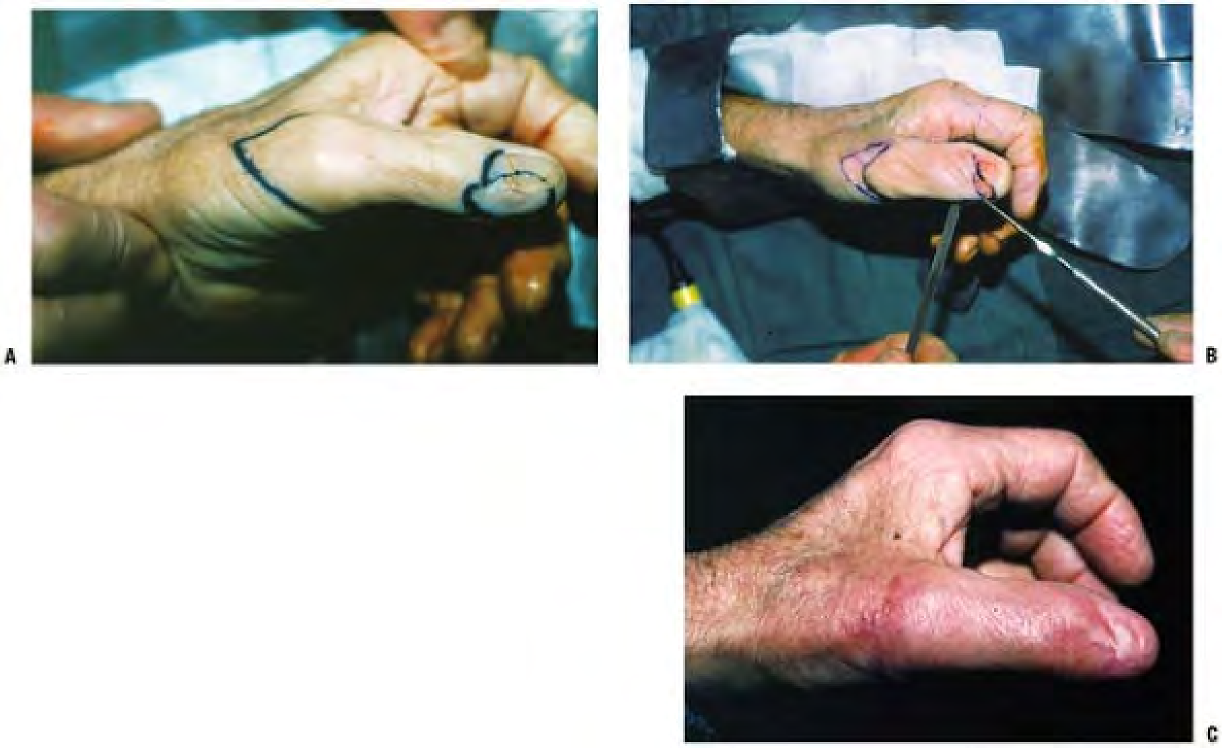

FIGURE 24-9 A: Preoperative view of the thumb of a 73-year-old man with a squamous cell carcinoma of the radial lateral fold of the nail. The thumb is marked with a dorsal V-Y hatchet flap for reconstruction. B: Intra-operative view of the flap being advanced. C: Late postoperative view. (Reproduced from Elliot D. Specific flaps for the thumb. *Tech Hand Up Extrem Surg.* 2004;8(4):198â€“211, with permission.)

This flap can be innervated for reconstruction of the thumb tip but, like the other flaps of dorsal origin, carries skin with poor innervation. It has more recently been redescribed carrying a segment of vascularized bone from the first metacarpal with it and may prove a very useful means of placing vascularized bone at the tip of the thumb without microsurgical transfer from the foot (7).

Reconstruction of Large Dorsal Defects

A very few techniques of reconstruction suffice to cover almost all full-thickness soft tissue losses on the dorsum of the thumb. Complete loss of the dorsal surface cover can be replaced using the first dorsal metacarpal artery or with a reverse posterior interosseous artery flap; both flaps are described in detail elsewhere in the text. The latter is particularly useful if the adjacent dorsal tissues of the hand have been injured and are not available as a donor site. It can also carry vascularized bone (and tendon) to the thumb.

The posterior interosseous vessel is small and the dissection relatively difficult, requiring microsurgical instrumentation and expertise. However, this flap is mostly safe and does not involve sacrifice of a major blood vessel of the forearm. For this reason, it is preferred to the use of the radial forearm flap for larger defects of the dorsum of the thumb, although the latter is useful for complete wrap-around of the first ray. When used to reconstruct the dorsum of the thumb, the reverse posterior interosseous artery flap donor site can almost always be closed primarily and then heals with a fine, narrow scar which may be almost invisible in the hair of a male forearm.

Postoperative Management

We remove all dressings and splints from reconstructed hand injuries on the day after surgery to allow washing and dressing of the raw wounds and verification of flap viability. Although this sometimes has to be done with care to avoid tension on flap pedicles, it allows the hand to be bathed in warm water. This both helps general cleaning and debridement of the wounds and is a comfortable environment in which to encourage very early hand mobilization. The wounds are dressed as lightly as possible while maintaining a moist antiseptic presence around any areas without skin which will heal through secondary intention. Suture lines that are not oozing are not covered. Sometimes homodigital and more distant reconstructions require that one or more joints be prevented from moving in a certain direction to take tension off the pedicle. For example, with all of the advancement flaps on the palmar surface, full extension will pull on the flap pedicles, so the thumb is splinted in slight flexion. Conversely, a slightly extended wrist position will ease tension on the pedicle of reverse posterior interosseous artery flaps. Protection of pedicles and flaps is continued during bathing and dressing changes by appropriate positioning and guarding with the other hand and between dressing changes by use of thermoplastic splints. Bathing and dressing changes are carried out twice a day with the patients learning to do this themselves as soon as possible.

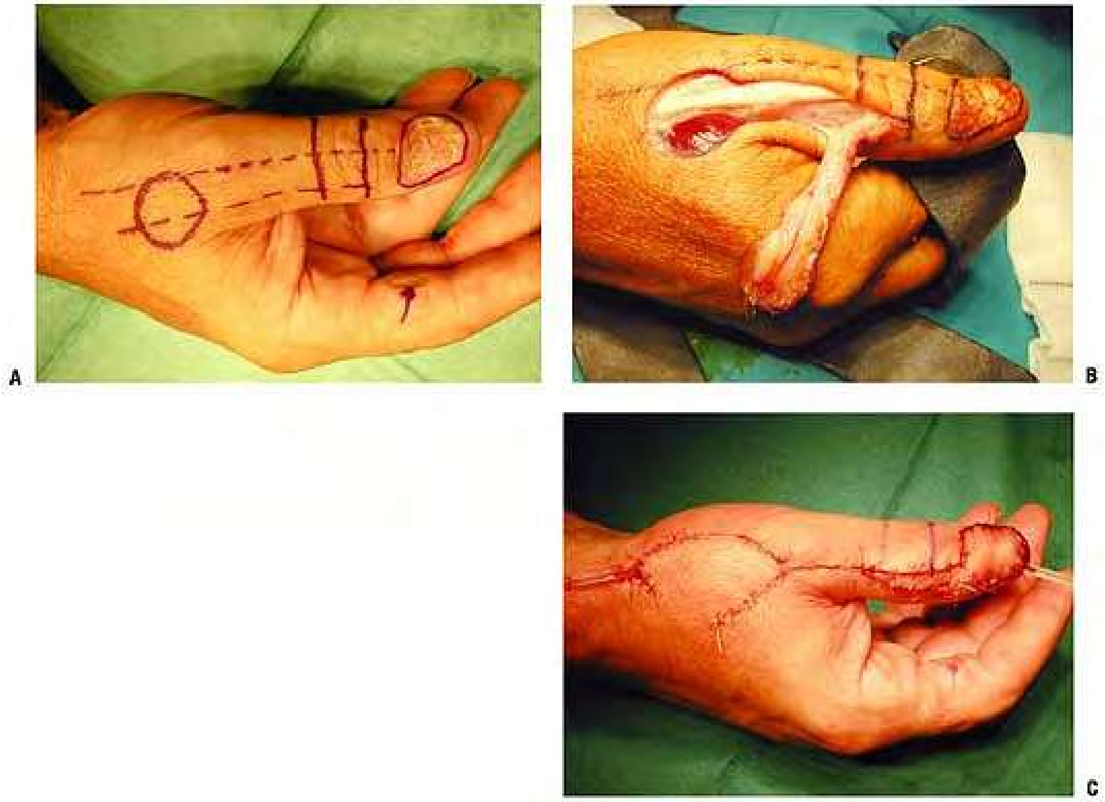

FIGURE 24-10 A: A thumb with a badly damaged nail before excision for biopsy. The Brunelli flap and its pedicle have been marked. B: The flap has been raised and shows the artery on which it is based on its undersurface. C: The flap inset after total excision of the nail complex. The flap has been taken with a â€œtailâ€• to ease pressure on the pedicle beyond its turning point. The donor site has been closed with a dorsal V-Y hatchet flap. (Reproduced with permission from Elliot D. Homodigital reconstruction of the digits: The perspective of one unit. *Indian J Plast Surg.* 2003;36:106â€“119.)

Rehabilitation

The use of homodigital flaps allows for early mobilization of the reconstructed thumb.

Motion will be restricted by concomitant tendon, nerve, and bony injuries; however with stabile skeletal fixation, most thumbs can be mobilized using limited motion and passive motion protocols under the supervision of a hand therapist within the first 72 hours following surgery. Whenever possible, we avoid techniques of bandaging, splinting, and skeletal fixation which will completely immobilize the reconstructed thumb. Over and above the early mobilization under water during dressing changes, mobilization of as much of the thumb and the rest of the hand as possible is continued from this early stage by exercises three or four times daily with the splint in place, albeit sometimes with loosening of some straps if this is possible without loss of the protection of flap pedicles. The splint provides a solid resting position for the thumb at night and gives the patient confidence that their thumb will not suffer from jarring by unexpected contact with external forces. Splints are generally used for 1 to 2 weeks. However, this period may lengthen to 4 or 5 weeks, if required to protect structures repaired under the flap. Between activities, the hand is elevated. This regimen of elevation and enthusiastic exercising is intended to eliminate edema from the hand as quickly as possible to minimize fibrin restriction of joint and extensor tendon movements.

Complications

Major complications include flap loss, partial flap loss, infection, and iatrogenic injury to the digital arteries or nerves. Partial flap loss is often best managed by allowing the wound to heal through secondary intention. Total flap loss, while infrequent with these homodigital flaps, will necessitate the use of a heterodigital island flap such as the first dorsal metacarpal artery island flap, or a larger pedicled forearm flap, such as the posterior interosseous flap or radial forearm flap. Obviously such flaps will create further injury to the arm but are reliable means of providing soft tissue coverage. An additional downside to the use of heterodigital flaps is that they are often insensate or require a neurorrhaphy to re-establish sensation. Return of sensation in these situations may be dependent on the age of the patient. In addition, heterodigital flaps will create an injury to another digit, which may further compromise hand function.

Results

The techniques described here have proved adequate to reconstruct most of the thumb defects we have seen in our practice over the last 12 years. The principle used in the smaller cases, particularly in reconstructing the tip of the thumb, has been to capitalize on the enormous healing capacity of digital skin and assist this, when necessary, by simple local flap reconstructions. These provide the best sensation possible and the best cosmetic profile to the thumb tip. For larger defects we prefer to use homodigital techniques whenever possible, although there are upper limits to the size of defects which can be reconstructed by these. For any defect, our choice of flaps is, in part, determined by the site, size, and shape of the defect and, in part, by flap reliability and ease of use. The homodigital flaps described above have proved particularly versatile and reliable in the hands of surgeons of varying levels of experience within our unit, although requiring some familiarity with microsurgical technique and instrumentation. They are quick to perform, which can be an advantage following lengthy bony or soft tissue reconstructions. They allow early and independent mobilization of the thumb and the use of local tissue respects the cosmetic principle of reconstruction of â€œlike with like.â€• Donor site morbidity is also limited to the already injured part. When used for fingertip reconstruction, advancement homodigital flaps cannot achieve the sensibility of the original, but are durable and intrinsically more likely to restore good sensory function in the digital tips than most of the reconstructive alternatives. For larger defects, the Littler transfer, the dorsum of the index ray, and, more recently, the reverse posterior interosseous artery flap are our local work horses and are discussed in detail in subsequent chapters.

References

1. Gaul JS. A palmar-hinged flap for reconstruction of traumatic thumb defects. *J Hand Surg Am.* 1987;12:415â€“421.

2. Moberg E. Aspects of sensation in reconstructive surgery of the upper extremity. *J Bone Joint Surg Am.* 1964;46:817â€“825.

3. Oâ€™Brien B. Neurovascular island pedicle flaps for terminal amputations and

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 25

Soft Tissue Management Following Traumatic Injury to the Femur

Thomas F. Higgins

Open injuries of the femur and thigh are common. The need for major soft tissue coverage in this area in the setting of trauma is uncommon due to the generous surrounding muscular envelope. Soft tissue coverage of the femur can be a concern in certain situations such as power takeoff injuries or pedestrians struck by motor vehicles; however, the majority of these defects may be covered with skin grafts, and the need for vascularized free tissue transfer is unusual. Despite this, there are several components to soft tissue management in this area which may be used during exposure to the proximal and distal portions of the femur. Simple alterations in established techniques can lead to improved postoperative contour, improved soft tissue coverage of hardware, and decreased rates of postoperative infection and stiffness following elective and emergent surgical procedures.

This chapter covers four major areas of concern with regard to soft tissue management in the thigh:

- The approach to the proximal femoral shaft and intertochanteric zone
- The approach for the repair of supracondylar and intra-articular distal

femur fractures

- Exposure for the treatment of open femur fractures
- The management of post-traumatic arthrofibrosis of the knee with the use of a Judet quadricepsplasty

Approach to the Lateral Proximal Femur

Indications/Contraindications

Fractures of the proximal femur, whether femoral neck or intertrochanteric, constitute the majority of all femur fractures (9). The soft tissue management issues in this portion of the proximal femur are the relative dearth of coverage over the greater trochanter and the area of the femur just distal to the vastus lateralis. Frequently, plate and screw implants and laterally based orthopaedic constructs will be prominent and symptomatic, particularly in thin patients. An approach that maximizes soft tissue coverage over the lateral aspect of the proximal femur and restores the most normal anatomy will be described. Rather than a direct lateral approach cutting through the vastus lateralis, which has been frequently advocated for this operation, a simple modification provides a more anatomic dissection and a friendlier soft tissue reconstruction.

Patient Positioning

The patient is positioned as desired for ultimate reduction and fixation goals.

Surgery

A direct lateral approach begins 1 cm proximal to the vastus ridge and extends as far distally as necessary for osteosynthesis. On the coronal plane, the incision should be 1 cm posterior to the midcoronal point of the femur (Fig. 25-1). Dissection is taken down to the iliotibial band, and this is divided in line with its fibers (Figs. 25-2 and 25-3).

A gentle sweeping motion with a sponge will clear trochanteric bursa or areolar soft tissue that lies in this layer. Particularly fibrotic bursal tissue may need to be excised with scissors (Fig. 25-4). This will reveal quite clearly the proximal extent of the vastus lateralis and its insertion at the vastus ridge and the confluence with the distal extent of the vastus medialis tendon. Rather than directly incising the vastus lateralis in the midcoronal point of the femur, a J-shaped incision is performed. Leaving several millimeters of the vastus lateralis tendon still attached to the vastus ridge, the tendon is incised from its anterior extent posteriorly until connecting with the lateral intermuscular septum (Figs. 25-5 and 25-6). The dissection is then taken through the most posterior fibers of the vastus lateralis extending from proximal to distal, just anterior to the lateral intermuscular septum. The

vastus tendon is tagged with a heavy braided nonabsorbable suture before being detached from the vastus ridge (see Fig. 25-5). As the dissection extends greater than 5 cm distal to the vastus ridge, the first perforating branch off the profunda femoris will be encountered penetrating the lateral intermuscular septum. These perforators should be ligated or cauterized. A gentle sweeping motion with an elevator from posterior to anterior along the proximal femur will allow reflection of the vastus lateralis in an anterior direction (Fig. 25-7). The periosteum and deepest soft tissues should not be stripped off the femur. For ideal visualization at this point, a Bennett retractor is placed anterior to the femur, and an assistant holds this forward (Fig. 25-8). Osteosynthesis proceeds as planned.

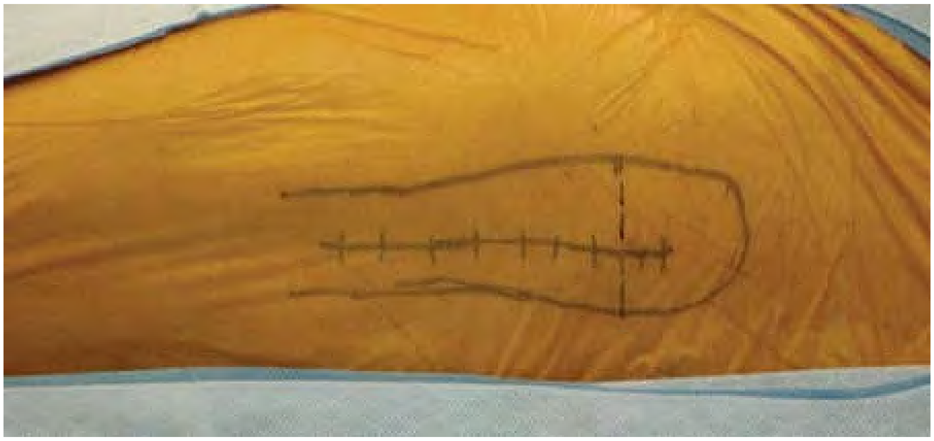

FIGURE 25-1 Incision drawn over the greater trochanter. *Dotted transverse line* represents the vastus ridge.

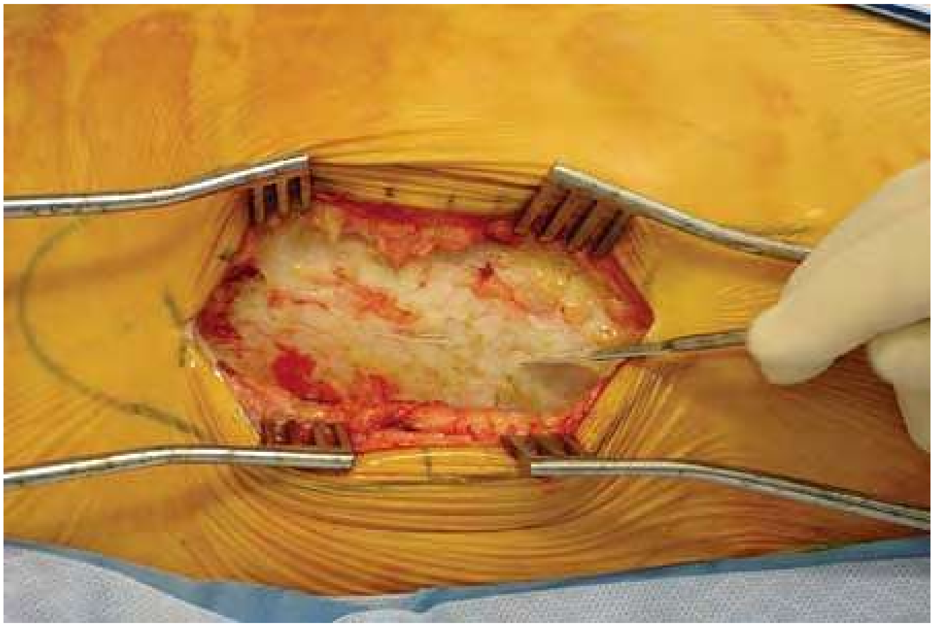

FIGURE 25-2 Iliotibial band is incised in line with its fibers.

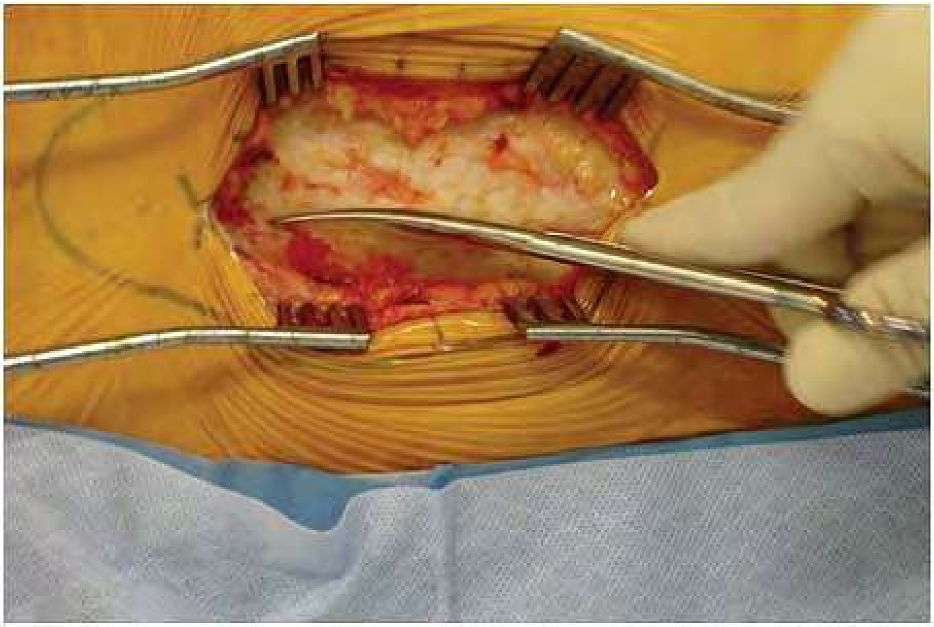

FIGURE 25-3 Iliotibial band is divided proximal to distal with curved Mayo scissors.

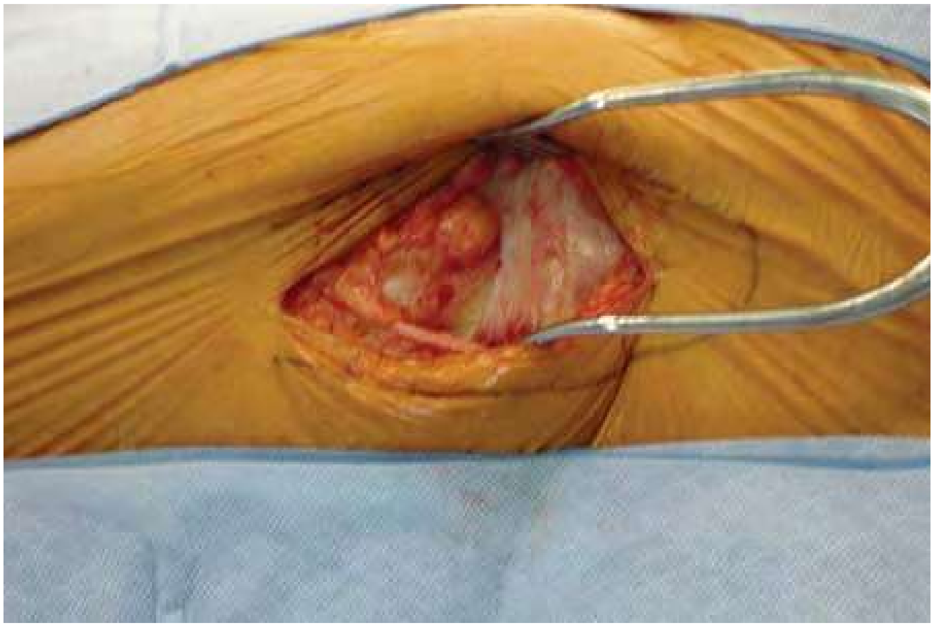

FIGURE 25-4 Fibrotic trochanteric bursa may need to be excised.

The Bennett retractors make for easy retraction and permit reduction and fixation. At the conclusion of the procedure, Bennett retractors are simply removed, and gravity allows the vastus to fall back into place, padding and directly covering the implant (Fig. 25-9). The most proximal origin of the vastus lateralis is then repaired back to its tendinous origin with a heavy braided nonabsorbable suture (Figs. 25-10 and 25-11). The posterior aspect of the vastus lateralis does not need to be repaired. The iliotibial band is repaired with interrupted nonabsorable suture, and the skin may be closed according to surgeon preference (Figs. 25-12 and 25-13).

This approach is most helpful for the placement of a dynamic hip screw, blade plate, or locking proximal femoral plate. It offers less scarring, as the muscle belly has not been interrupted. This should also be a more functional repair, as less damage presumably has been done to the vastus origin by dividing it and repairing it, rather than by directly insulting it with a midgastric approach.

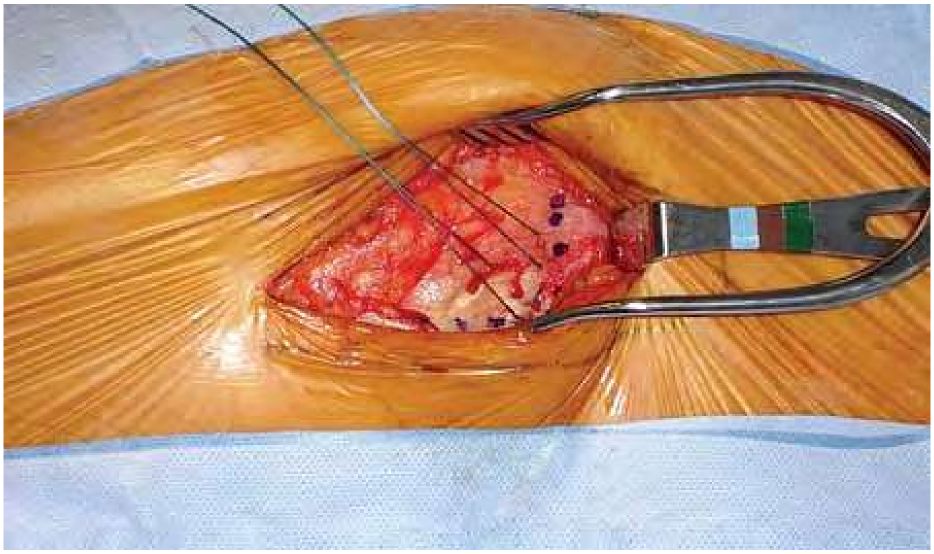

FIGURE 25-5 Proximal tendon of vastus lateralis is to be incised in a posteriorly based â€œJ.â€•

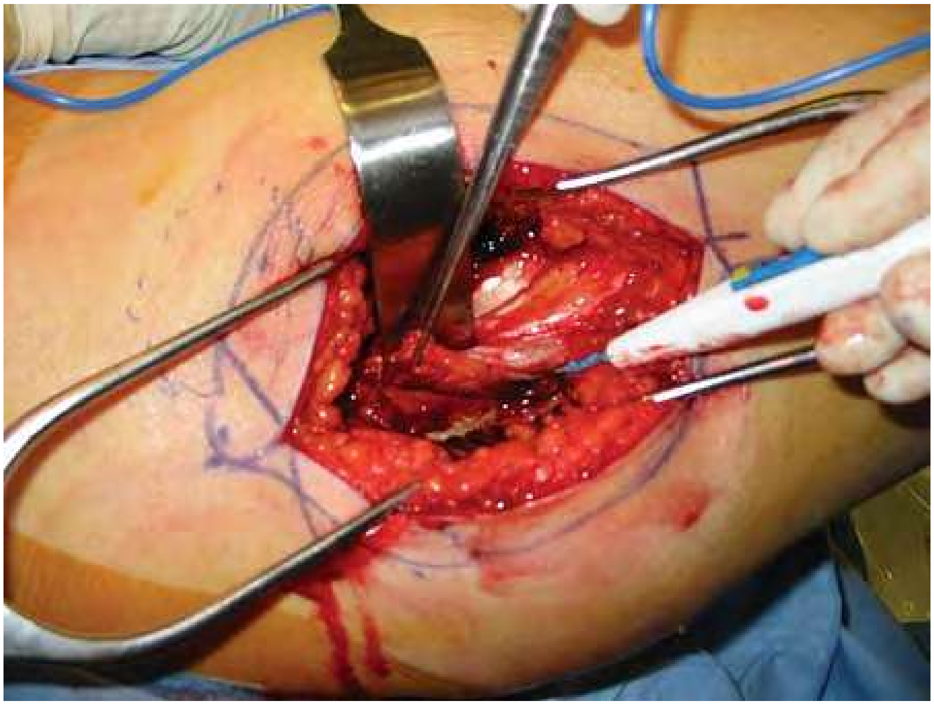

FIGURE 25-6 Vastus lateralis origin is elevated in along the â€œJ.â€•

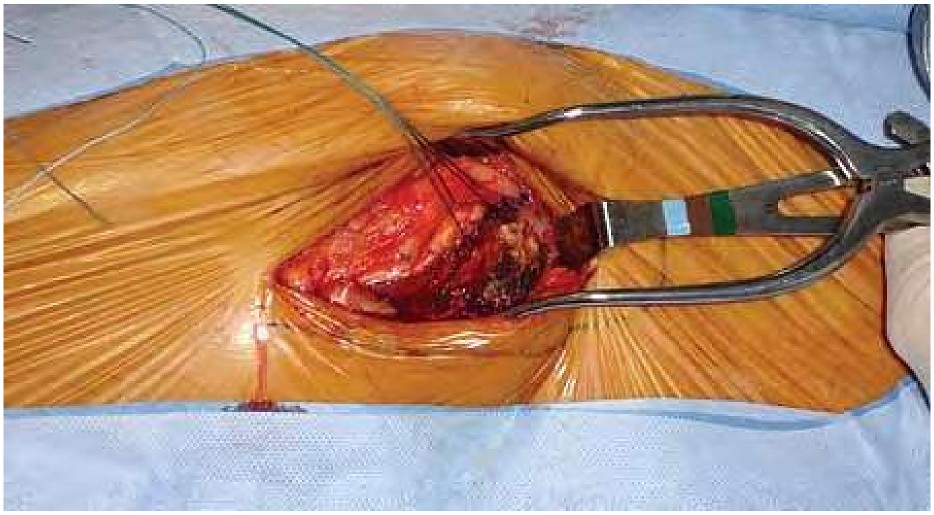

FIGURE 25-7 Lateralis is retracted anteriorly with tagging sutures.

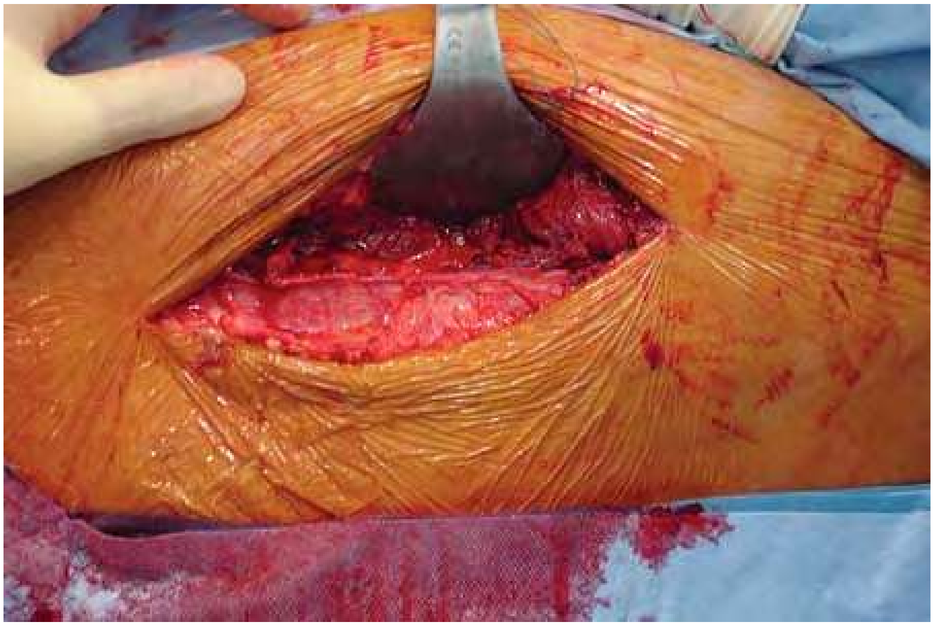

FIGURE 25-8 Bennett retractor is used to reflect the vastus anteriorly.

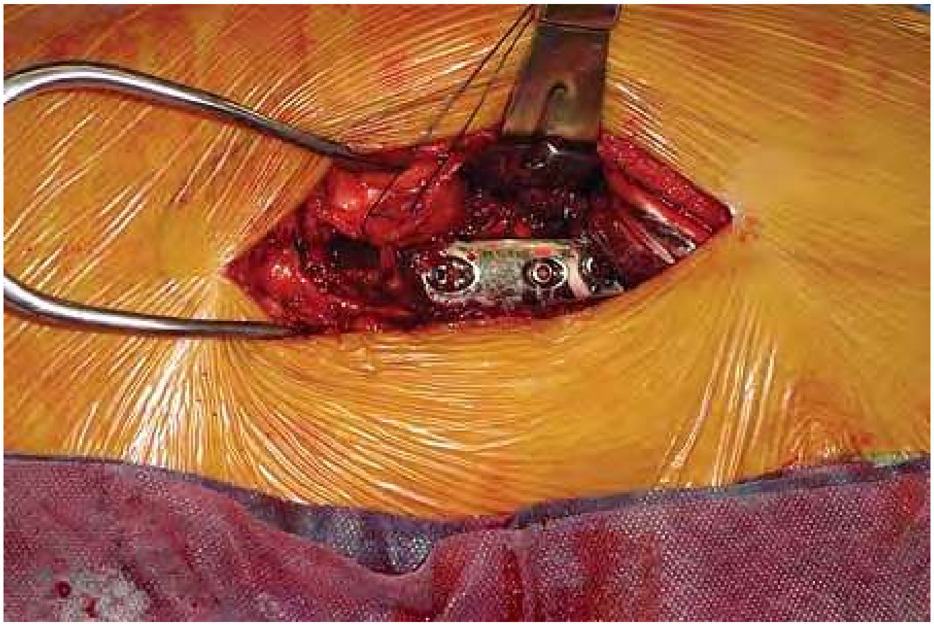

FIGURE 25-9 Osteosynthesis hardware, after placement, is covered by vastus.

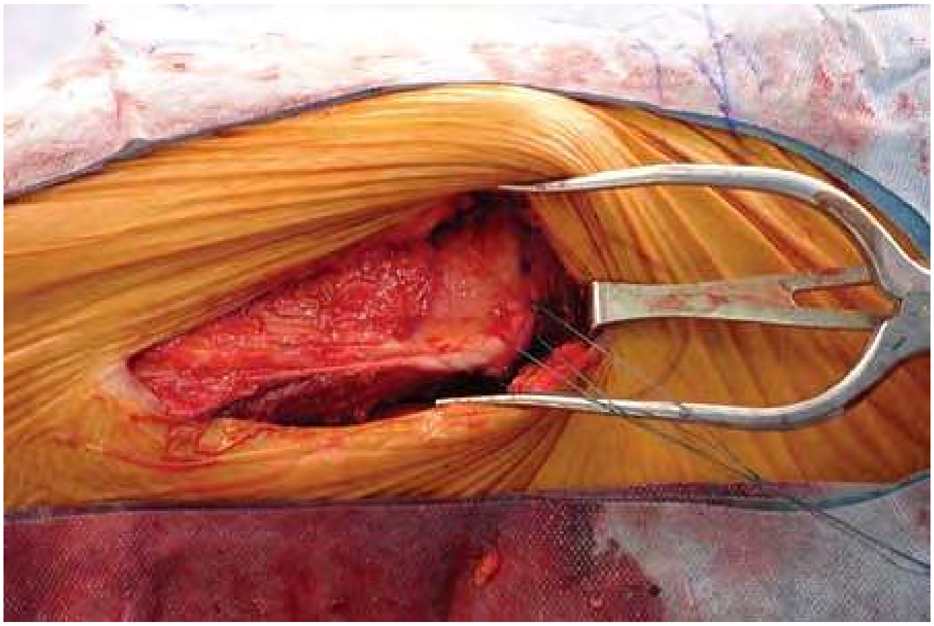

FIGURE 25-10 Vastus lateralis pulled back to its origin at the vastus ridge.

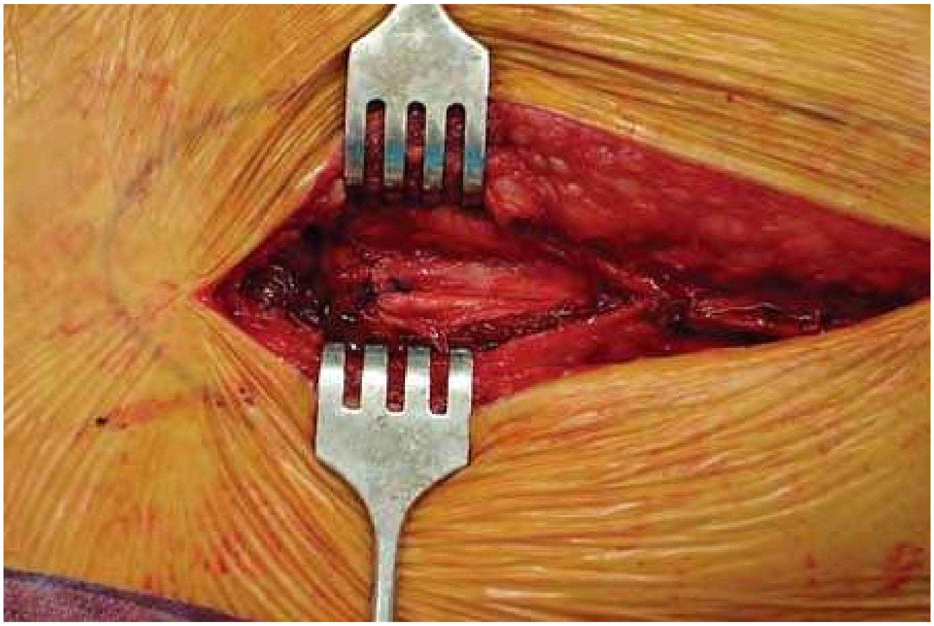

FIGURE 25-11 Through the closing iliotibial band, repair of the vastus origin may be visualized.

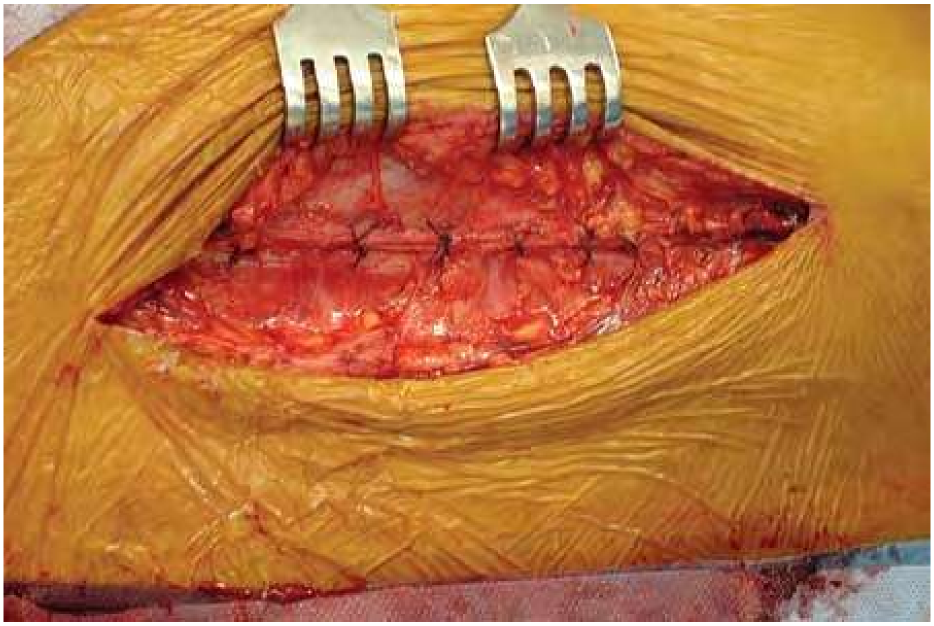

FIGURE 25-12 Iliotibial band is repaired with interrupted nonabsorbable suture.

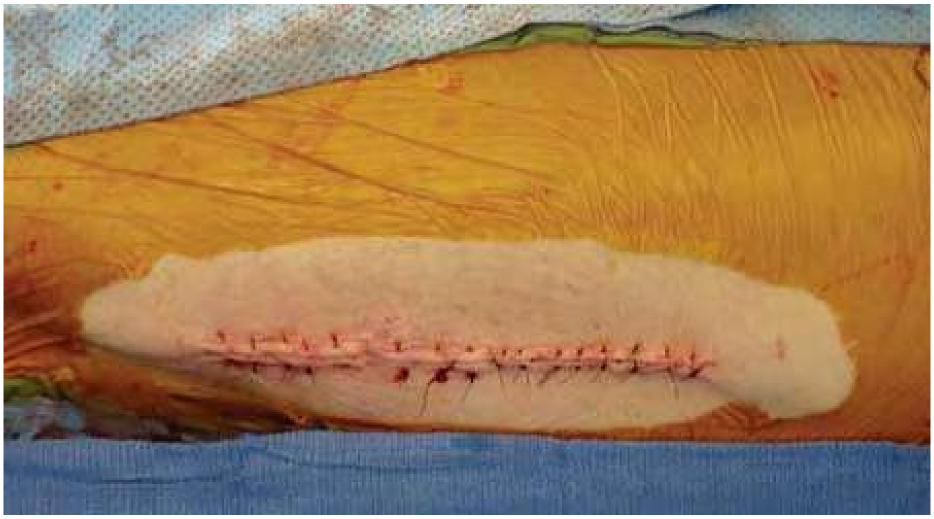

FIGURE 25-13 Skin closure at completion of intertrochanteric fracture osteosynthesis.

Postoperative Management

In the elderly population with intertrochanteric hip fractures, patients are generally allowed to bear weight as tolerated to facilitate mobilization (8). The first 2 weeks postoperatively focus on range of motion of the hip and knee, and gait training with assistive devices as necessary. At 2 weeks postoperatively, strengthening of the abductors, hip flexors, and quadriceps is added to the regimen.

In younger patients, or fractures with a subtrochanteric component, weight bearing may be protected initially, and advanced as desirable by the treating surgeon.

Complications

Complications of operations through this approach are generally related to the osteosynthesis, and these may be avoided somewhat with careful attention to achieving reduction and adherence to correct placement of implants.

Approach to the Distal Femur and Knee

Indications/Contraindications

Multiple variations on parapatellar and lateral parapatellar approaches have been described. Similar to the muscle sparing approach described for the proximal femur, the â€œswashbucklerâ€• approach described by Starr et al at Parkland Hospital (14) offers excellent visualization of the joint for anatomic fixation of interarticular fracture while preserving the vastus and quadriceps with no intramuscular dissection.

This approach offers the advantage of not interfering with or dividing the extensor mechanism while gaining adequate visualization of the joint space. With proximal and distal retraction, the intervening metaphyseal soft tissues may be largely undisturbed to enhance healing of comminuted metaphyseal bone.

Surgery

Patient Positioning

The patient is positioned supine. The lower extremity is prepared and draped free. A bump may be used under the knee to facilitate exposure and/or reduction.

Technique

The skin incision starts at the insertion of the patellar ligament and extends directly cephalad to the apex of the patella. It then extends lateral and proximal toward the lateral intermuscular septum (Fig. 25-14). A full thickness skin flap is elevated off the extensor fascia. The lateral margin of the patella and patellar ligament are identified distally. Proximally, the lateral margin of the vastus lateralis is identified.

The knee capsule is incised immediately along the lateral aspect of the patellar ligament (Fig. 25-15). Care must be taken not to cut the anterior horn of the lateral meniscus. Dissection is taken directly along the lateral margin of the patella, leaving a small cuff of tissue for eventual repair. At the superolateral shoulder of the patella, the deep dissection extends laterally toward the lateral intermuscular septum. The vastus lateralis is thus spared and is elevated anteriorly off the lateral intermuscular septum. As dissection nears the femur, perforating branches of the profunda femoris artery must be identified and ligated. These will be located every 3 to 4 cm approximately 1 cm lateral to the femoral cortex.

Right angle retractors may be placed at the level of the quadriceps tendon for medial excursion of the quadriceps and patella (Fig. 25-16). Bennett retractors may be placed over the femoral diaphysis. These

retractors in combination will allow adequate visualization of the femur for reconstruction. If possible, no retractors at all should be placed medially at the level of the metaphysis, as most techniques demand preservation of the medial soft tissues so as to not interfere with healing of metaphyseal comminution. Adequate visualization may easily be achieved with retractors placed proximal and distal to this area.

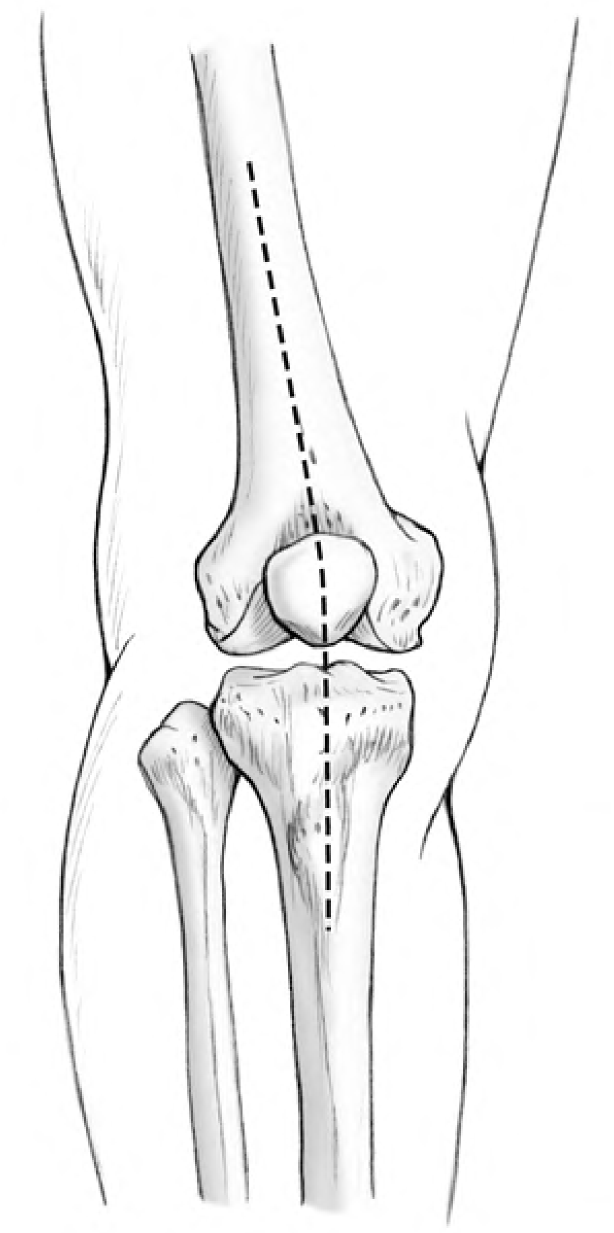

FIGURE 25-14 The skin incision for the â€œswashbucklerâ€• approach to the distal femur. (Redrawn from Starr AJ, Jones AL, Reinert CM. The â€œswashbucklerâ€•: a modified anterior approach for fractures of the distal femur. *J Orthop Trauma.* 1999;13(2):138â€“140. With permission.)

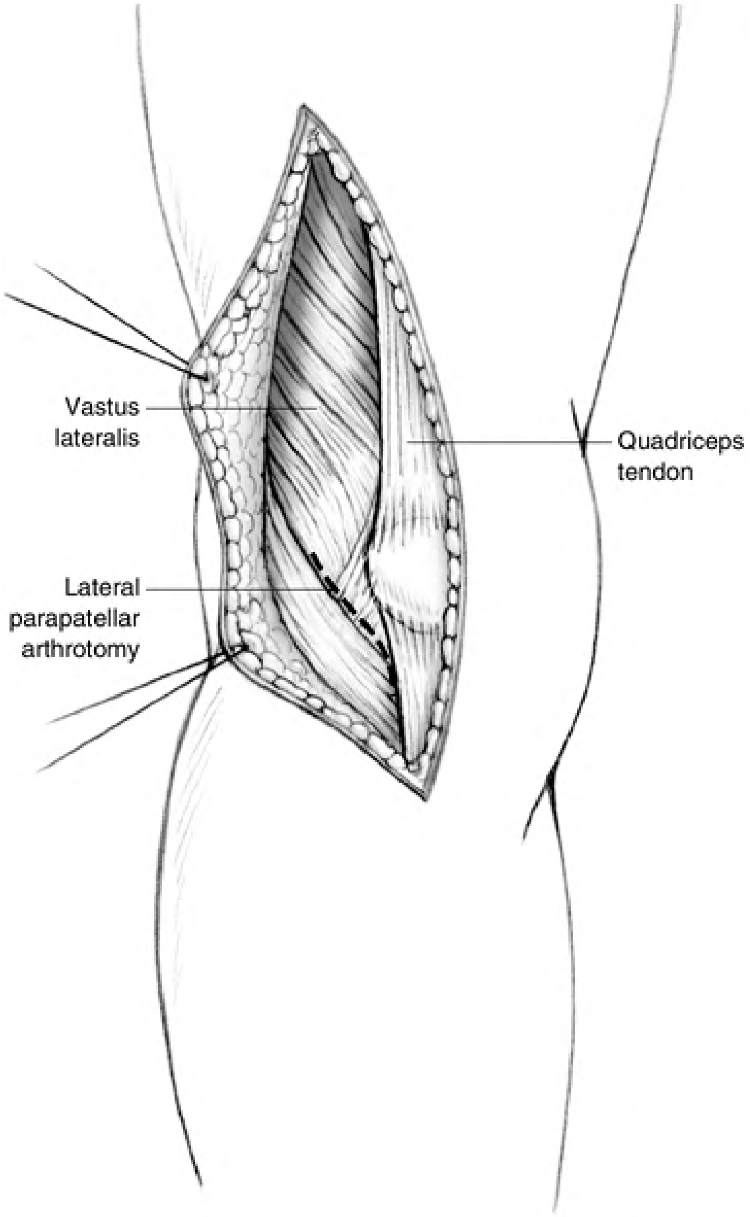
Vastus lateralis
Quadriceps tendon
Lateral parapatellar arthrotomy

FIGURE 25-15 The lateral parapatellar arthrotomy extends to the lateral margin of the vastus lateralis. (Redrawn from Starr AJ, Jones AL, Reinert CM. The â€œswashbucklerâ€•: a modified anterior approach for fractures of the distal femur. *J Orthop Trauma.* 1999;13(2):138â€“140. With permission.)

For full visualization of the femoral condyles, the patella may be everted. This may be facilitated by elevating a small proximal portion of the most lateral insertion of the patellar ligament off the tibia. Osteosynthesis is then performed.

Closure is achieved with repair of the lateral parapatellar retinaculum and capsule lateral to the patellar ligament. A drain may be placed in the knee and run along the lateral intermuscular septum, exiting the skin proximally. Subcutaneous tissue and skin are repaired according to surgeon preference.

Postoperative Management

Assuming stable osteosynthesis has been achieved, immediate range of motion is initiated. A physical therapist may perform passive supervised range of motion, 0 to 90 degrees, with no strengthening and no active quadriceps. Alternatively, a continuous passive motion machine may be used. Quadriceps strengthening begins at 6 weeks postoperatively, and weight bearing is advanced according to surgeon judgment.

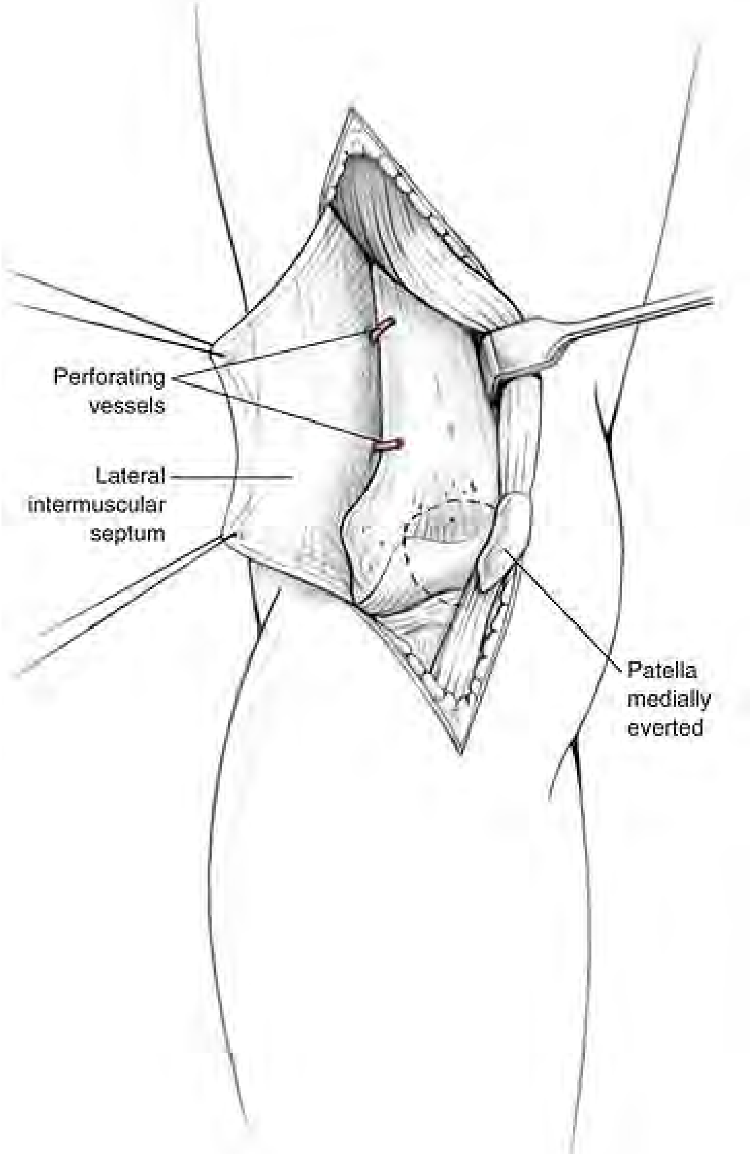
Perforating vessels
Lateral intermuscular septum
Patella medially everted

FIGURE 25-16 With quad retracted and patella everted, visualization of the distal femur is excellent. Starr AJ, Jones AL, Reinert CM. The â€œswashbucklerâ€•: a modified anterior approach for fractures of the distal femur. *J Orthop Trauma.* 1999;13(2):138â€“140. With permission.)

Complications

Potential complications of complex distal femur fractures include infection, nonunion, and arthrofibrosis. Judet quadricepsplasty for loss of knee motion will be addressed later in this chapter.

Surgical Exposure of Open Femoral Fractures

Indications/Contraindications

Given the wide girth of soft tissues around the femur, an open fracture by definition implies a high-energy injury. There are compelling data to support early irrigation and debridement of open fractures for the prevention of infection (3,4). Small lacerations or puncture wounds are often the tell-tail signs of an open fracture and should not be overlooked; these relatively small surface injuries often disguise a significant deeper soft tissue insult (13). Open fractures of the distal femur in particular will present with a very small opening of the skin where the femoral shaft has pistoned distally to penetrate the vastus medialis obliquus or the vastus lateralis and the skin (Figs. 25-17 and 25-18).

Frequently, the treatment of open intra-articular distal femur fractures will necessitate an aggressive irrigation and debridement at the time of presentation and transarticular spanning external fixation for definitive stabilization. Even in the hemodynamically unstable trauma patient, early stabilization of the femur is mandatory. Prolonged maintenance of a patient in Hare traction is not advisable given the attendant complications of skin breakdown and sciatic nerve palsy, and so spanning external fixation is recommended over prolonged traction.

The initial management of these fractures should not include the placement of a pulsatile irrigation catheter through the small traumatic wound. This may succeed only in further embedding debris

in an open fracture site and will certainly not provide an adequate debridement of the injured soft tissues that would ultimately be responsible for the development of a deep infection. Adequate debridement of an open femoral fracture demands wound extension (5).

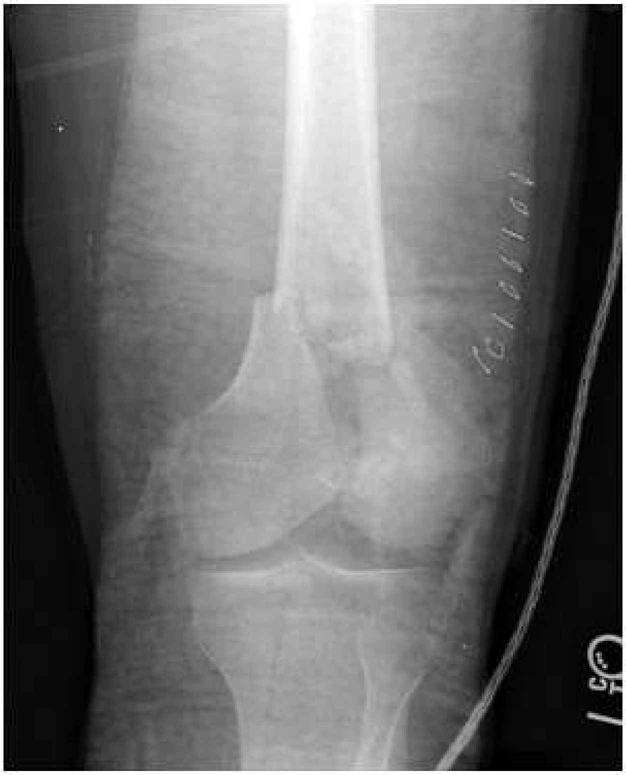

FIGURE 25-17 Skin staples demonstrate the length of the inadequate incision used to debride this open intracondylar distal femur fracture.

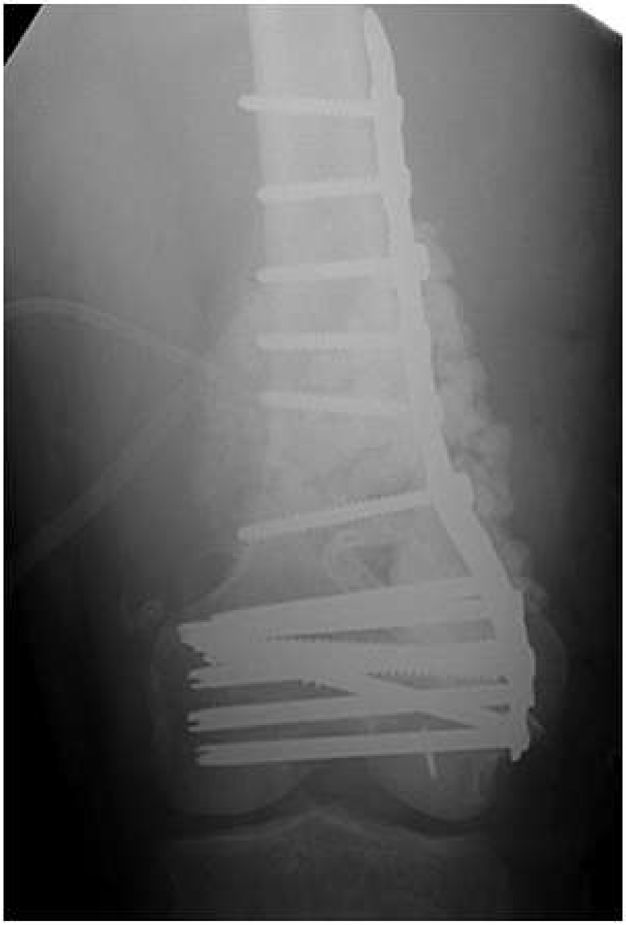

FIGURE 25-18 This inadequate index debridement leads to a long and complicated course of treatment, including a deep infection and the placement of antibiotic beads.

Preoperative Planning

Initial debridement entails some insight by the initially treating surgeon into the definitive fixation plan. Frequently, the entire approach for the eventual repair of an open femur fracture may be performed at the time of debridement, particularly if this will include plating of a distal femur fracture. Open fractures of the femoral diaphysis may still be treated quite successfully with intramedullary nailing after wound extension and debridement (13). Regardless of eventual fixation plan, proximal and distal extension of soft tissue wounds is always necessary.

Surgery

Extension of the skin laceration is performed both proximally and distally down to the investing facsia of the thigh, which is always torn due to the pre-existing trauma. Fracture exposure will often require intramuscular dissection. Specific landmarks or structures to avoid will vary depending on location of the open wounds. Extensive dissection through muscle may be required, but most of this muscle has been traumatized and may not be viable. It is exactly this tissue that must be removed for the adequate debridement of the open femoral fracture and the prevention of a potential deep infection.

If, following aggressive irrigation and debridement, it is determined that a transarticular spanning external fixation is going to be used for definitive or temporary stabilization, consideration must be given to the underlying soft tissues. In applying joint spanning external fixation, one should be careful to place the femoral pins proximal enough that the eventual internal fixation is not communicating with the pin sites if at all possible. Femoral pins may be introduced directly anterior through the quadriceps, anterolateral, or directly lateral. Tibial pins are generally introduced anteromedially. Spanning external fixation may be maintained for several weeks while waiting for physiologic or soft tissue stabilization. Nowotarski et al. demonstrated in non-spanning external fixation that patients

could safely undergo one-stage conversion to internal fixation without significant infectious problems (11). However, waiting much longer than 3 weeks may create problems in having to take down interval callus formation and makes the articular reconstruction more difficult.

Postoperative Management

In an effort to stabilize the soft tissue envelope, initiation of range of motion may be delayed after open fracture of the distal femur. If an open distal femur fracture has been spanned with an external fixator, subsequent debridements may be necessary, and the exact timing of definitive internal fixation is at the surgeon's discretion.

If a thorough debridement and immediate intramedullary nailing of an open diaphyseal fracture has been completed, range of motion and weight bearing may be initiated right away.

Complications

The prompt and aggressive debridement of open fractures is an attempt to avoid the most obvious complication, osteomyelitis. Further complications from open fractures include knee stiffness and non-union. In cases with a segmental bone loss, or a high degree of metaphyseal comminution, patients may be informed at the time of definitive fixation that an elective bone grafting procedure will be performed 4 to 6 weeks later. The author prefers to use proximal tibial cancellous autograft, as it is biologically active, is a low morbidity harvest site when compared to iliac crest, and limits the affected area of the patient to a zone already involved in the injury. For large segmental defects, a cement spacer may be placed at the time of index fixation, with the plan to exploit the â€œbiologic membraneâ€• formed around this spacer to promote healing at the time of subsequent bone grafting (12).

Judet Quadricepsplasty

Indications/Contraindications

Frequently, patients who have suffered comminuted intra-articular injuries of the distal femur or supracondylar femur will suffer from poor range of motion secondary to post-traumatic scarring and contracture. Various methods have been proposed for the treatment of this loss of flexion at the knee. Arthroscopic lysis, closed manipulation, and quadriceps tendon lengthening have all been advocated (15). For particularly stubborn cases of post-traumatic arthrofibrosis and quadriceps contracture leading to loss of knee flexion, the author finds the Judet quadricepsplasty to be the most effective way to regain functional motion of the knee (1,2,6,7,10,16).

This approach recognizes that, in addition to the skeletal injury apparent on the radiographs, there was great soft tissue damage sustained at the time of distal femoral fracture. It is this damage to the rectus and vastus musculature that is at least in part responsible for the subsequent loss of range of motion. Arthrofibrosis and capsular contracture also contribute to post-traumatic stiffness and are addressed by this operation.

Judet attributed the stiffness of the knee to four principle anatomic reasons. First was adhesion of the subquadricipital pouch with tethering of the patella. Second was retraction of the parapatellar fibrous elements and intra-articular adhesions. Third was adherence of the vastus intermedius muscle of the femur. Fourth was degenerative fibrosis of the quadriceps. The surgical procedure is based on a sequential release of these various parts of the contracture. Range of motion is to be assessed after each stage, and the operation proceeds until the desired results are achieved or maximum range of motion is achieved after completion of all steps.

Preoperative Planning

An epidural catheter is placed preoperatively. General and epidural anesthetic are used intraoperatively, but the regional anesthesia will help with patient compliance and tolerance with postoperative continuous passive motion.

Surgery

Range of motion is assessed on administration of the general anesthetic (Fig. 25-19). Two incisions are generally used (Fig. 25-20). The first is a medial parapatellar approach. This allows access to the medial aspect of the patellar tendon, the suprapatellar pouch medially, and the medial gutter. The medial retinaculum is released, and adhesions of the medial joint and suprapatellar pouch are excised.

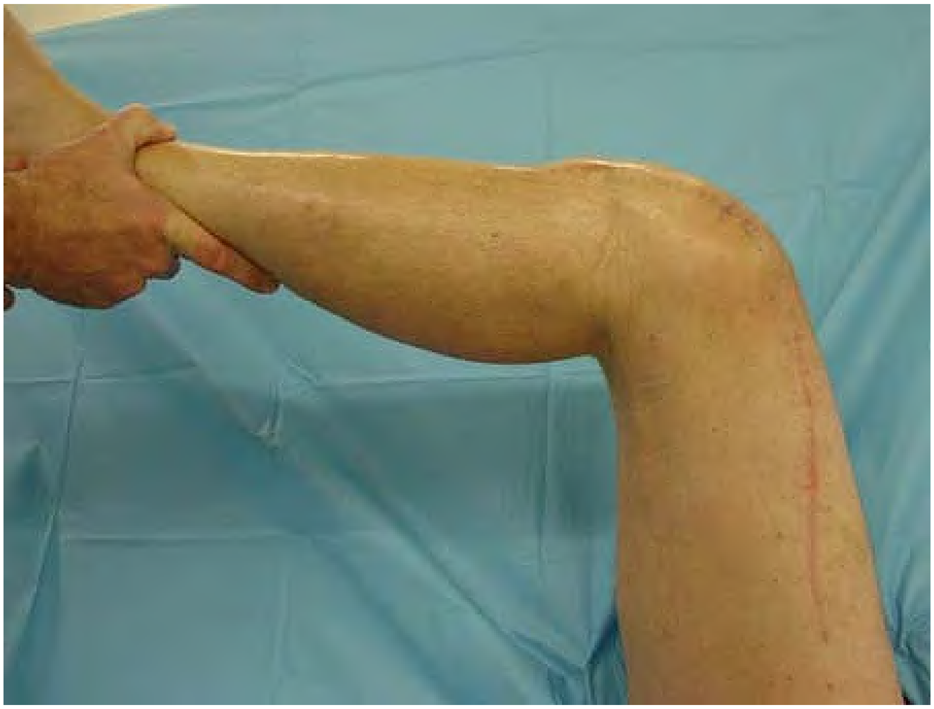

FIGURE 25-19 Range of motion is assessed preoperatively.

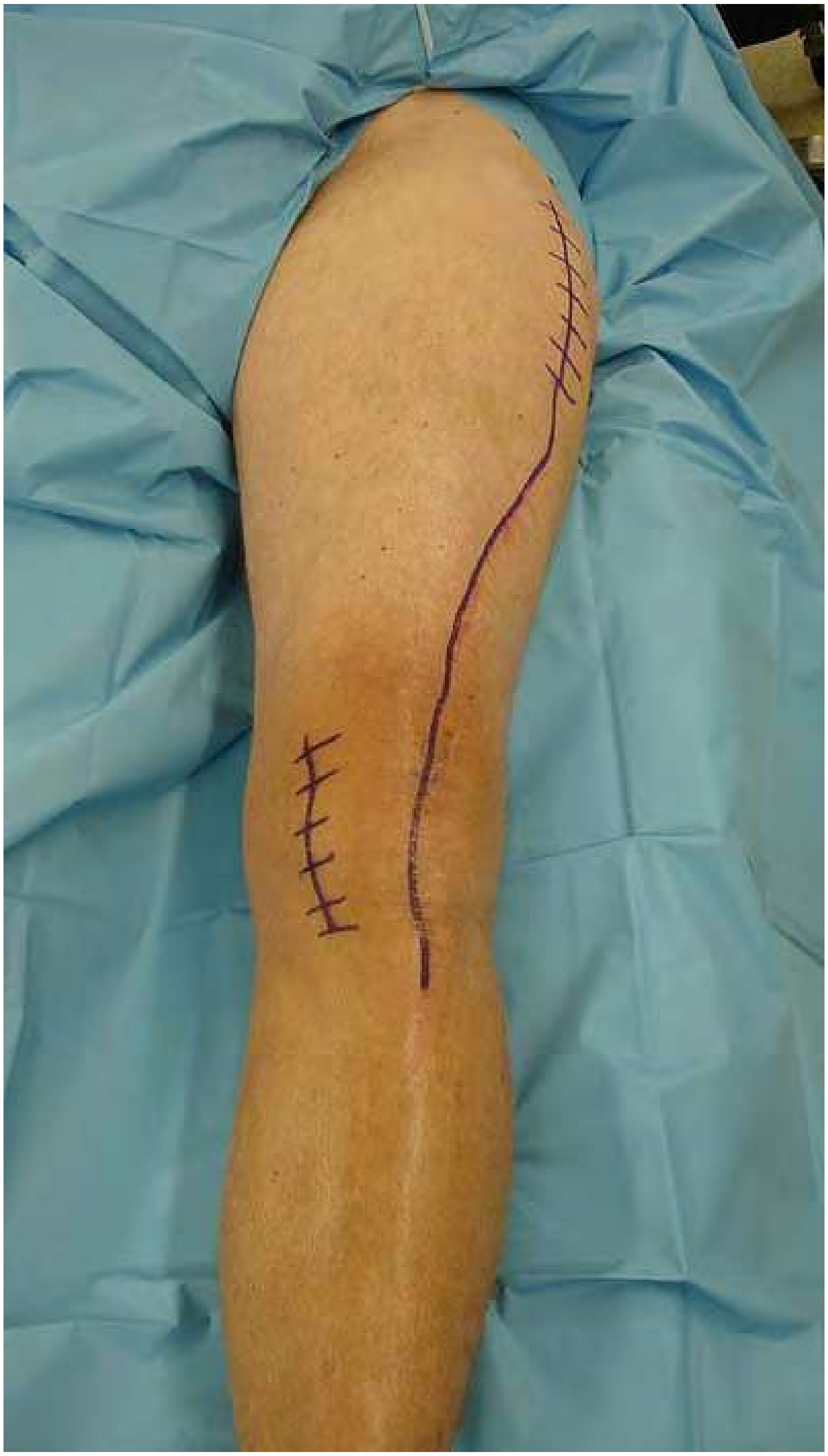

FIGURE 25-20 Two incisions planned preoperatively for Judet quadricepsplasty.

The second incision often follows the incision made from the operative repair of the fracture (Fig. 25-21). This will usually follow the lateral aspect of the patella ligament through the lateral patellar retinaculum and proximal up along the lateral intramuscular septum to the level of the greater trochanter. In the distal segment the parapatellar and lateral retinacular tissues are released and intra-articular adhesions are excised. The patella is freed until it may be easily lifted off the femoral condyles (Fig. 25-22). Scar tissue along the medial and lateral aspects of the patellar ligament, particularly distally, would likewise need to be excised. At this point, range of motion is checked, and the quadriceps may be manipulated at each subsequent step. The vastus lateralis is elevated from the linea aspera, and perforating vessels are ligated as they are encountered. The vastus intermedius is then lifted extraperiosteally from the front and side of the femur (Fig. 25-23). Much of the intermedius will be replaced by scar and fibrotic tissue, and much of this may be excised without significant functional impact (Fig. 25-24). Frequently, one will find heterotopic bone formation within the muscle bellies, and this may be excised along with fibrotic muscle tissue in an effort to regain length. Once again, the knee is manipulated and range assessed.

Next, the vastus lateralis is released from its femoral origin superolaterally, and the rectus femoris may be released from its origin on the anterior capsule (Figs. 25-25 and 25-26). This allows distal excursion of much of the quadriceps mechanism in an effort to regain range of motion at the knee. The knee is manipulated for a final time and final range of motion is measured (Fig. 25-27). Only the skin is closed, and a drain is used (Fig. 25-28). Given the somewhat sanguinous nature of the dissection along the well-perfused margins of scar tissue, postoperative hematoma must be guarded against, and drain use is encouraged.

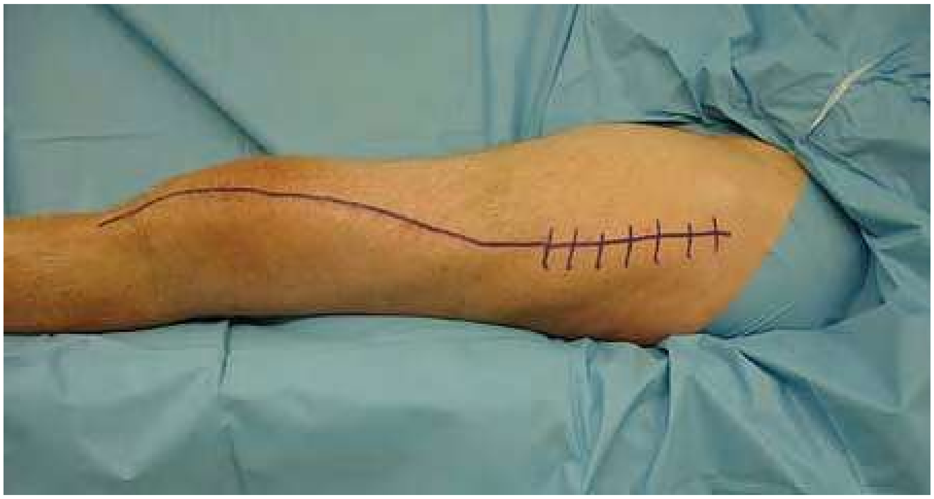

FIGURE 25-21 Lateral incision planned as extension (*dotted line*) off existing scar (*solid line*).

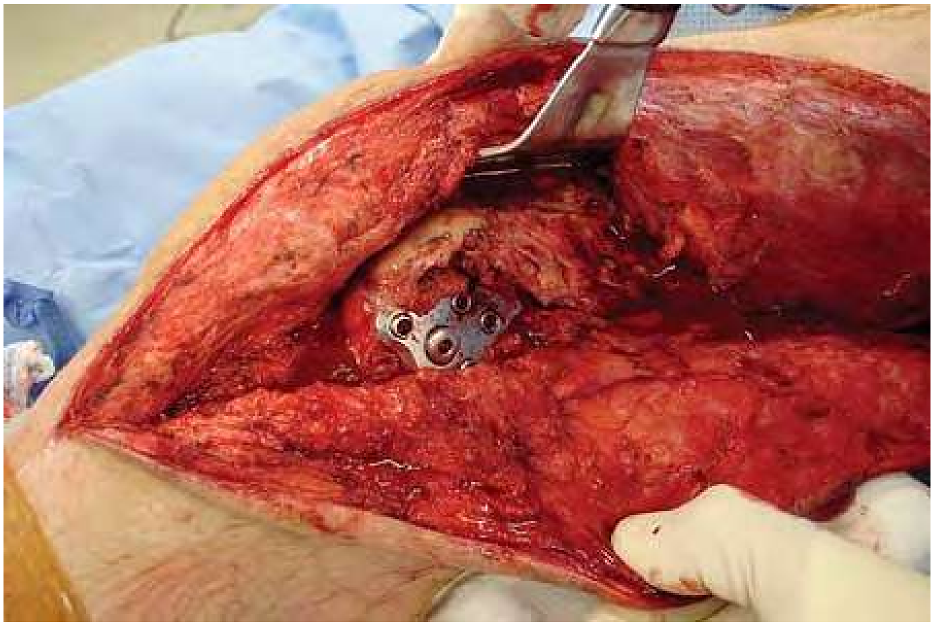

FIGURE 25-22 Supracondylar zone shows dense fibrotic scar tissue.

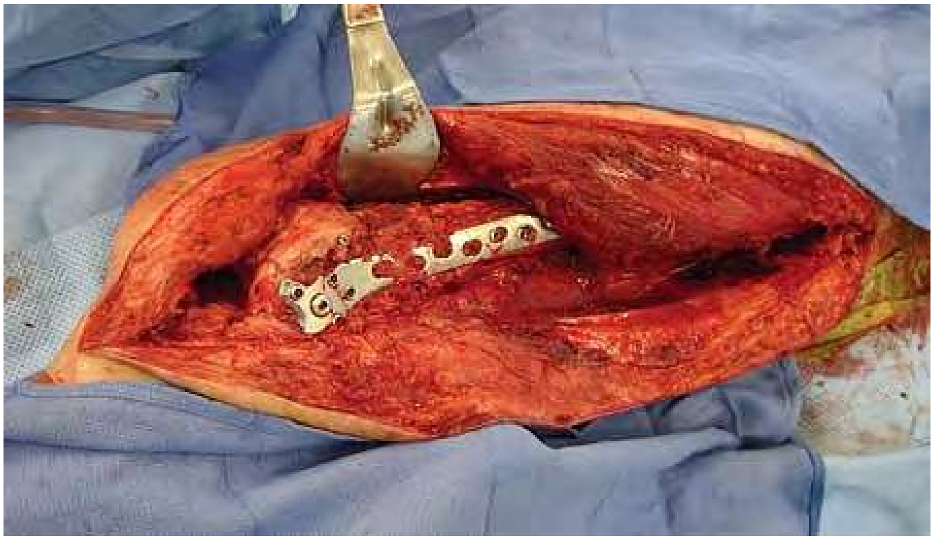

FIGURE 25-23 Scar tissue release mobilizes the extensor mechanism off the anterior femur.

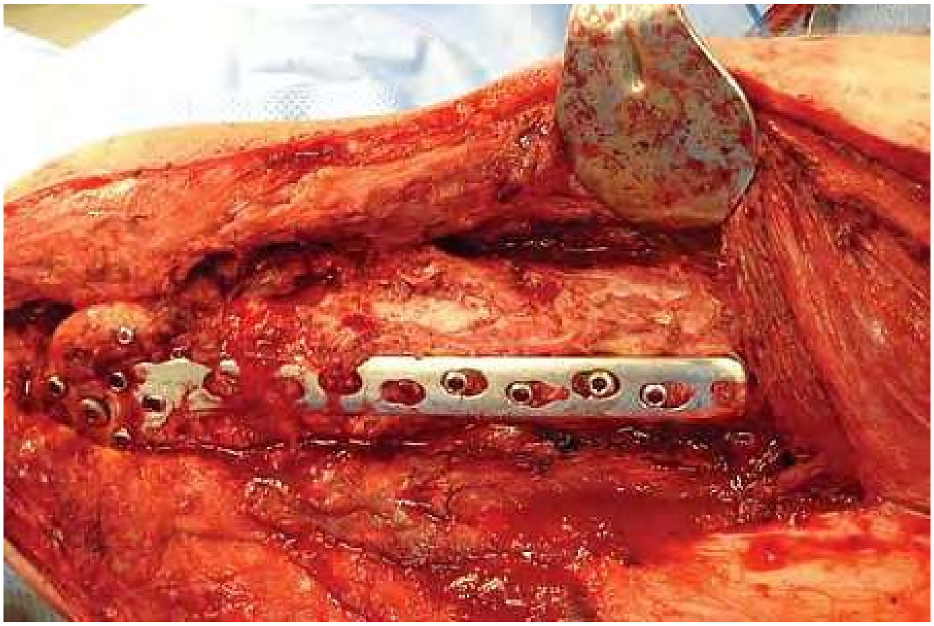

FIGURE 25-24 Heterotopic bone may scar the quadriceps to the femur and must be excised.

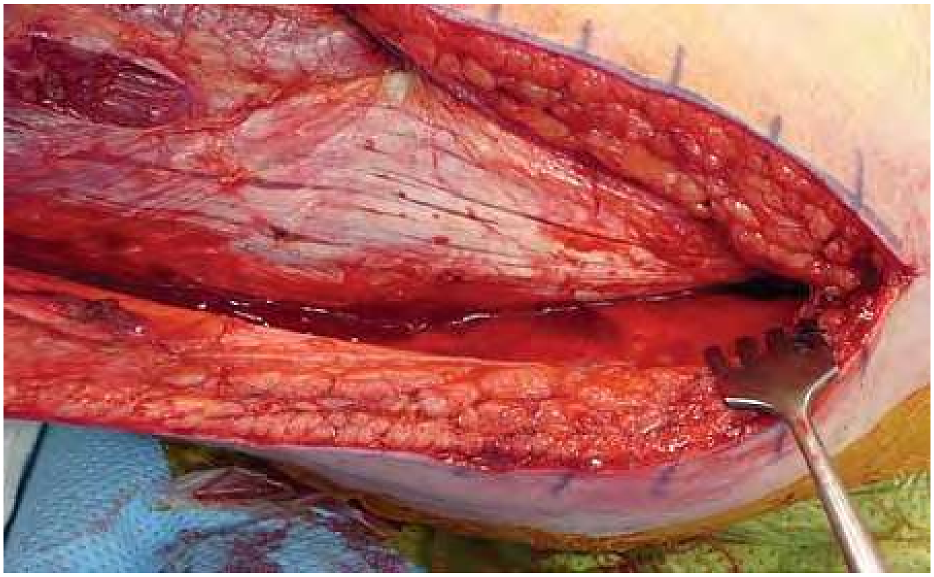

FIGURE 25-25 Disection may be extended up to the level of the vastus lateralis origin and the rectus origin anteriorly.

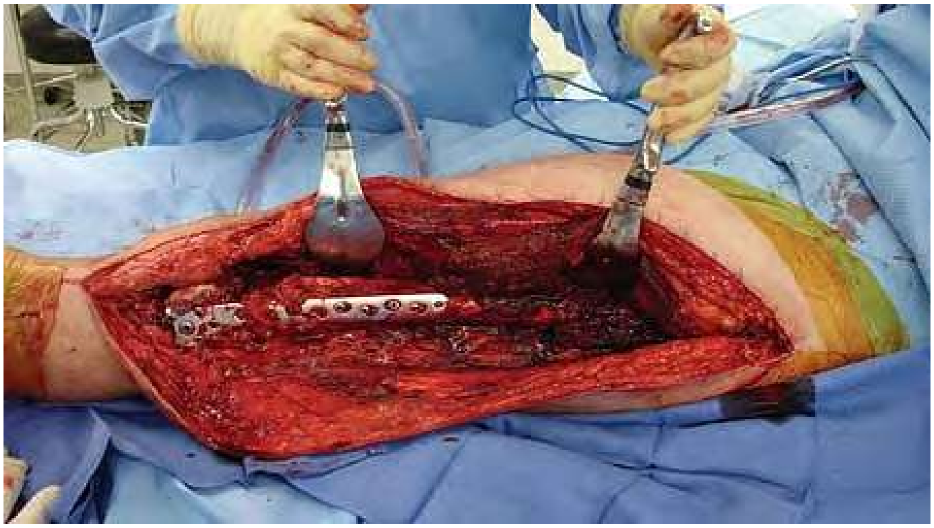

FIGURE 25-26 With release of the rectus proximally, ultimate excursion of the extensor mechanism may be achieved.

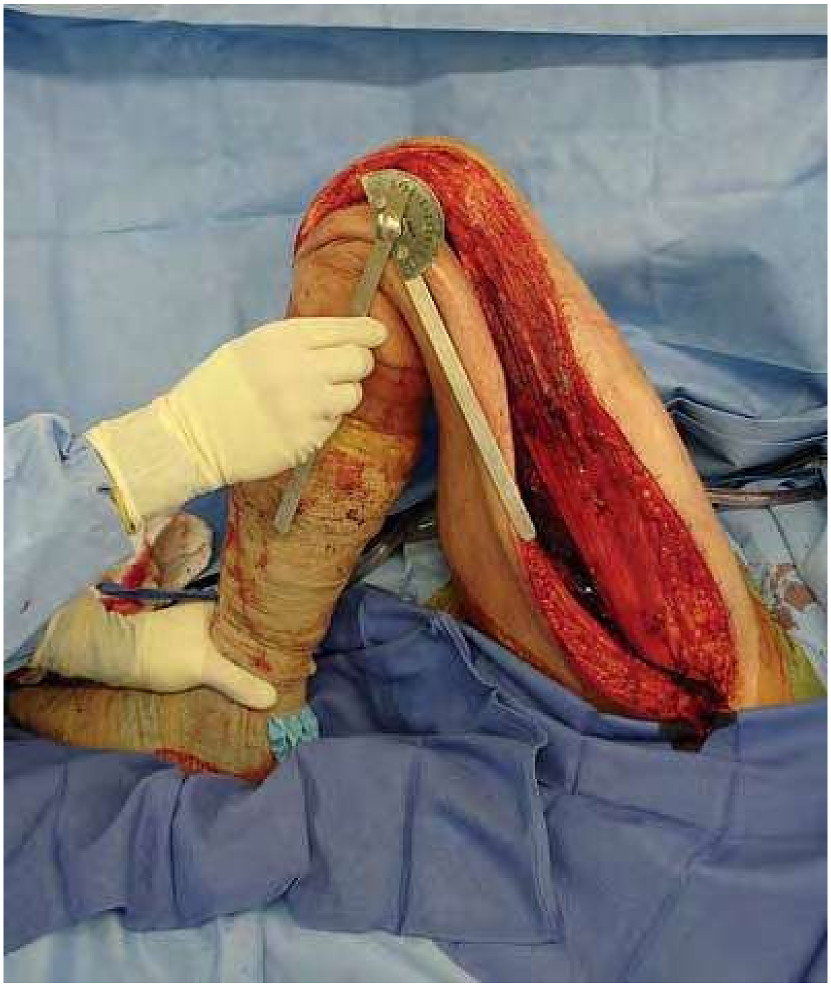

FIGURE 25-27 Ultimate range of motion should be assessed before and after closure.

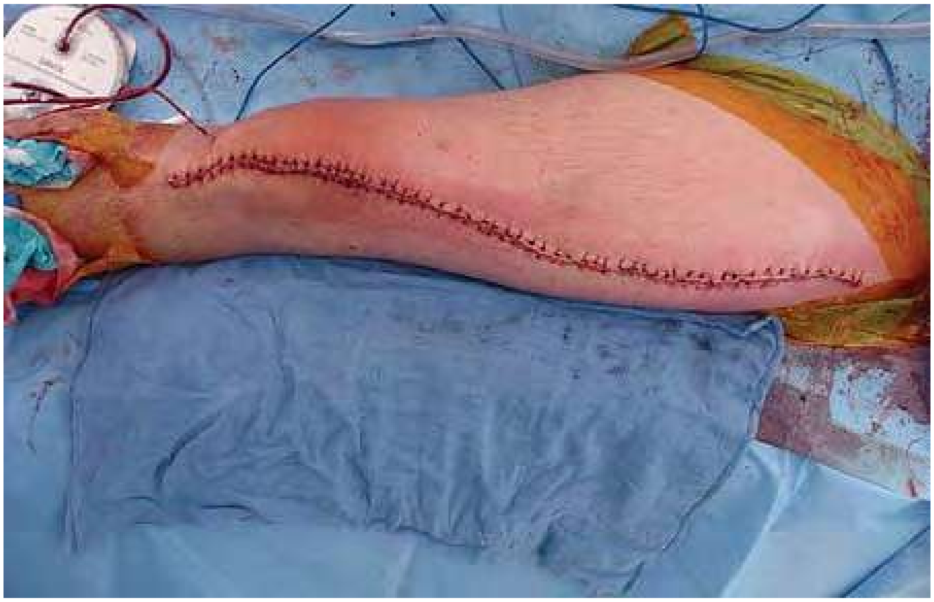

FIGURE 25-28 Final skin closure of the lateral incision.

Postoperative Management

The knee may be immobilized in a position of flexion so that, on awakening from anesthetic, the patient may have some idea of what the range of motion may be achieved (Fig. 25-29). Epidural anesthetic is generally used, and continuous passive motion is initiated after the recovery room. The drain is removed in approximately 48 hours. Due to the bleeding which may be encountered in dissection of scar and muscle, we have found that blood transfusion may be necessary before discharge.

No postoperative immobilization is typically used at the time of discharge, and patients may need crutches for ambulation, as they frequently will have a functional extensor lag postoperatively. This extensor lag resolves over time.

Complications/Results

The most common immediate complication following this procedure is hematoma formation. This complication can be minimized with intraoperative hemostasis and with the routine use of closed suction drains in all patients. Secondary complications include extensor lag, which will resolve in most cases with postoperative therapy. Recurrent stiffness and arthrofibrosis can also complicate postoperative results; however, the results of this operation in the literature have been encouraging (1,2,6,7,10,16), with total gains in range of motion between 55 to 69 degrees in the two largest series.

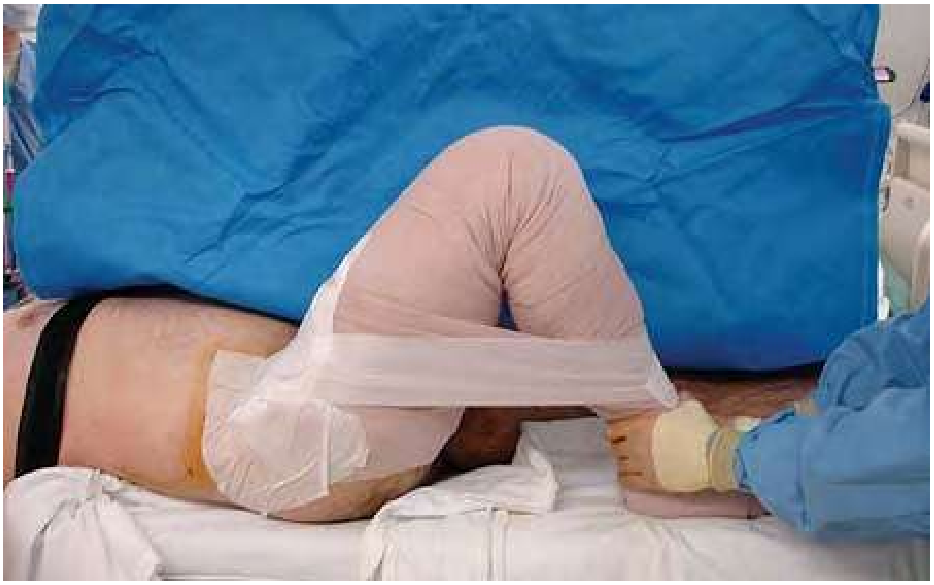

FIGURE 25-29 Knee is immobilized in flexion postoperatively, and continuous passive motion is initiated after the recovery room.

Summary

The soft tissue around the femur and thigh segment is adequate enough to sustain extensive trauma without requiring free tissue transfer. Some of the techniques listed in this chapter may facilitate the correct handling of the soft tissues, help to prevent infection in the setting of open fracture, or help patients to regain motion after a comminuted supracondylar injury.

References

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 26

Management of Soft Tissue Defects Surrounding the Knee and Tibia: The Gastrocnemius Muscle Flap

Steven Myerthall

David J. Jacofsky

Steven L. Moran

Defects surrounding the knee and patellar region are common and may result from trauma, tumor, or infection. One of the most important goals in the management of open wounds surrounding the knee is stable soft tissue coverage. Soft tissue coverage with muscle can seal the joint and bone from ongoing contamination and help to revascularize underlying structures, while preventing the development of late infection and nonunion. Options for soft tissue coverage are determined by the size of the wound, location, and surrounding zone of injury. The gastrocnemius muscle flap has historically been used to cover defects surrounding the knee, lower thigh, and patellar region. Its consistent vascular anatomy and superficial location have made it a workhorse for coverage of defects in this area.

The gastrocnemius is the most superficial muscle of the posterior calf. The muscle has two heads arising from the medial and lateral condyles at the femur, and the adjacent capsule of the knee; these muscles then insert into the calcaneal tendon (Fig. 26-1).

The two muscle heads, medial and lateral, unite at the level of the fibular head. Proceeding distally the muscles join with the tendon of the soleus at the mid-leg to form the Achilles tendon. The medial sural artery is the dominant vascular pedicle for the medial gastrocnemius muscle, whereas the lateral sural artery supplies the lateral head. The origin of the arteries is approximately 4 cm above the head of the fibula and both originate from the popliteal artery. Innervation to the muscles originates from the tibial nerve and lies posterior to the vascular pedicle as it enters the muscle.

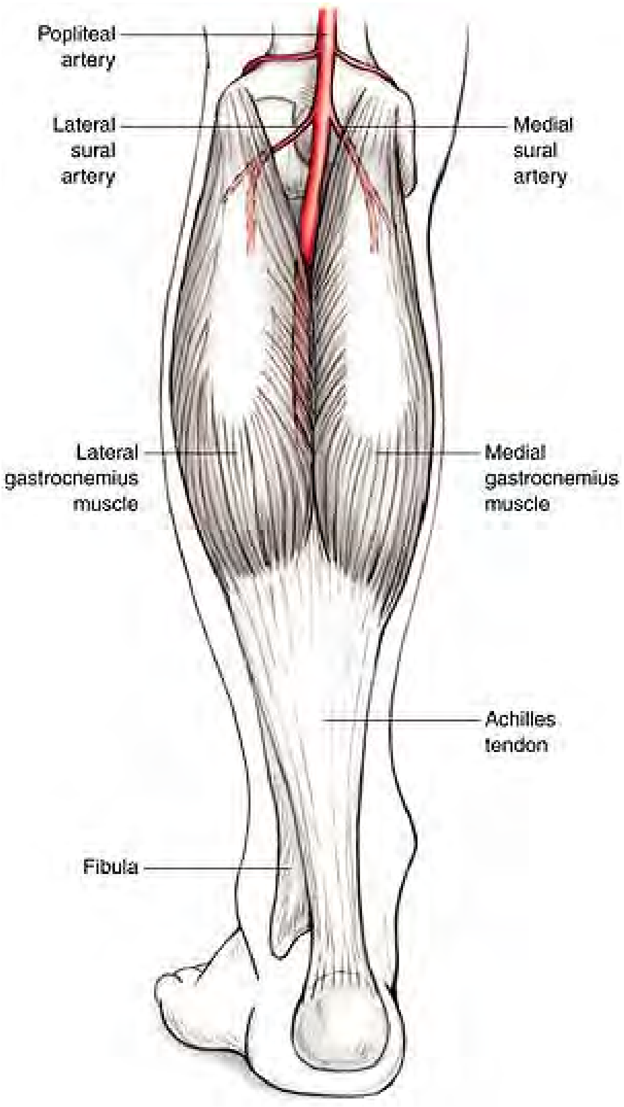
Popliteal artery
Lateral sural artery
Medial sural artery
Lateral gastrocnemius muscle
Medial gastrocnemius muscle
Achilles tendon
Fibula

FIGURE 26-1 The medial and lateral gastrocnemius muscles as seen from the posterior approach.

Indications/Contraindications

Both the medial and lateral gastrocnemius muscles may be used independently or in conjunction with each other. The medial gastrocnemius will cover the inferior thigh, knee, and proximal tibia. The medial head of the muscle is used most frequently as a proximally based flap due to its larger size in comparison to the lateral head of the muscle. The lateral head may also be used alone or in combination with the medial head for coverage of large tibial defects or for lateral distal thigh wounds. For defects at the level of the mid-portion of the tibia, the gastrocnemius muscle may not provide adequate coverage and the soleus muscle is preferred for coverage of middle third defects. For defects involving the distal third of the tibia and ankle, free tissue transfer is usually required. The tendinous inferior margin of the gastrocnemius muscle may be used to augment the repair of an injured suprapatellar tendon.

Contraindications to the use of the gastrocnemius muscle flap include active infection and/or significant disruption of the soft tissue and/or vascular pedicle. Additional contraindications for flap use include any procedure or injury which may have traumatized or injured the sural artery, such as a previous repair of a popliteal arterial laceration or repair of popliteal aneurysm. Occasionally, severe compartment syndromes may render the muscle fibriotic and useable for transfer, but due to its superficial location and high take off of the sural artery, this is unusual. Direct trauma to the posterior calf may result in muscle destruction and should be ruled out prior to proceeding with flap elevation. Radiation to the knee following tumor extirpation may also compromise the vascular pedicle,

increasing the chance of total or partial muscle necrosis following transfer. The muscle should not be used if the soleus and contralateral gastrocnemius muscle are no longer functional, as this will create postoperative difficulty with plantar flexion. A history of a recent deep venous thrombosis within the involved extremity is a relative contraindication for the use of the flap, although we have used the flap successfully in such situations.

Distally based gastrocnemius muscle flaps have been described for coverage of middle third and lower leg defects. Vascular supply to the muscle in these cases is supplied through crossing anastomotic arterial connections between the medial and lateral gastrocnemius muscle bellies extending across the midline raphe. Use of the flap in this manner is not common, as more reliable methods of middle third and lower leg coverage are available and include the soleus muscle flap, the sural artery flap, and free flap coverage.

Preoperative Planning

The preoperative management of severe open fractures involves stabilization of the patient, tetanus prophylaxis, and broad-spectrum antibiotics. Trauma often compromises the availability of the skin and soft tissue surrounding the knee. Nonviable tissue requires aggressive debridement. Serial debridements may be needed if tissue viability is uncertain. Once the wound bed is clean, flap coverage may proceed, ideally within 72 to 96 hours following initial injury. Preoperative physical examination should focus on determining the function of the remaining muscles of the lower leg to determine if sacrifice of the gastrocnemius could result in a significant loss of plantar flexor following muscle transfer. A vascular examination should also be performed to assess for the patency of the popliteal artery. In those patients with a history of peripheral vascular disease or a history of popliteal arterial trauma, an angiogram can confirm patency of the sural artery prior to surgery.

Surgery

Patient Positioning

Operating room setup is standard for lower extremity coverage. A leg tourniquet may facilitate dissection, especially in cases of lateral gastrocnemius harvest where the perineal nerve must be identified and protected during flap elevation.

The muscle may be harvested with the patient in prone or supine position. If the patient is prone, a stockinette or mid-posterior S-shaped incision is made and significantly facilitates exposure of the muscle's origin posteriorly over the femoral condyle. The posterior approach is made with an incision starting 5 cm above the popliteal crease and extending down to the distal end of the muscle belly. More commonly, however, for anterior traumatic defects surrounding the patellar and tibial region, the patient is positioned in the supine position with the leg internally or externally rotated to facilitate exposure of the medial or lateral heads, respectively. Lateral decubitus positioning is also an option for lateral gastrocnemius muscle elevation.

The patient may receive either general or spinal anesthesia. In additional, depending on associated injuries, the ipsilateral or contralateral thigh should be prepared for a skin graft donor site. After sterile preparation, the entire extremity is draped and fully exposed. Appropriate preoperative antibiotics should be given 30 to 60 minutes prior to incision.

Technique

Medial Gastrocnemius Muscle Flap

A longitudinal incision is made where the separation of the bellies of the gastrocnemius and soleus muscle can be palpated. This should parallel the medial border of the tibia and can be curved proximally in a posterior direction to facilitate exposure and dissection of the proximal portion of the muscle. The incision should extend from the level of the tibial plateau to 10 cm above the medial malleolus. If the flap is to be tunneled into position on the anteromedial aspect of the leg, a skin bridge of at least 7 cm must be maintained to prevent skin necrosis. For more anterior defects, or preexisting peripatellar defects, a curved incision can be extended both proximally and distally from the defect site to a line 4 cm medial to the edge of the tibia (Fig. 26-2A,B).

During initial subcutaneous dissection, the saphenous vein is identified and preserved (Fig. 26-2C). The underlying deep investing fascia of the leg is then opened to expose the plane between the gastrocnemius and soleus muscles (Fig. 26-2D,E). In most patients, the plantaris tendon is located in this interval and should be left intact. This plane is easily opened with blunt finger dissection. The plane between the external fascia of the muscle and the overlying skin should also be opened using blunt dissection. Large, perforating vessel can be identified running from the muscle to the overlying skin. These vessels serve as the main blood supply to the skin paddle of a gastrocnemius fasciocutaneous flap. These vessels should be ligated in the gastrocnemius muscle flap dissection. The

sural nerve and lesser saphenous vein can be identified running between the medial and lateral heads of the gastrocnemius muscle and should be preserved.

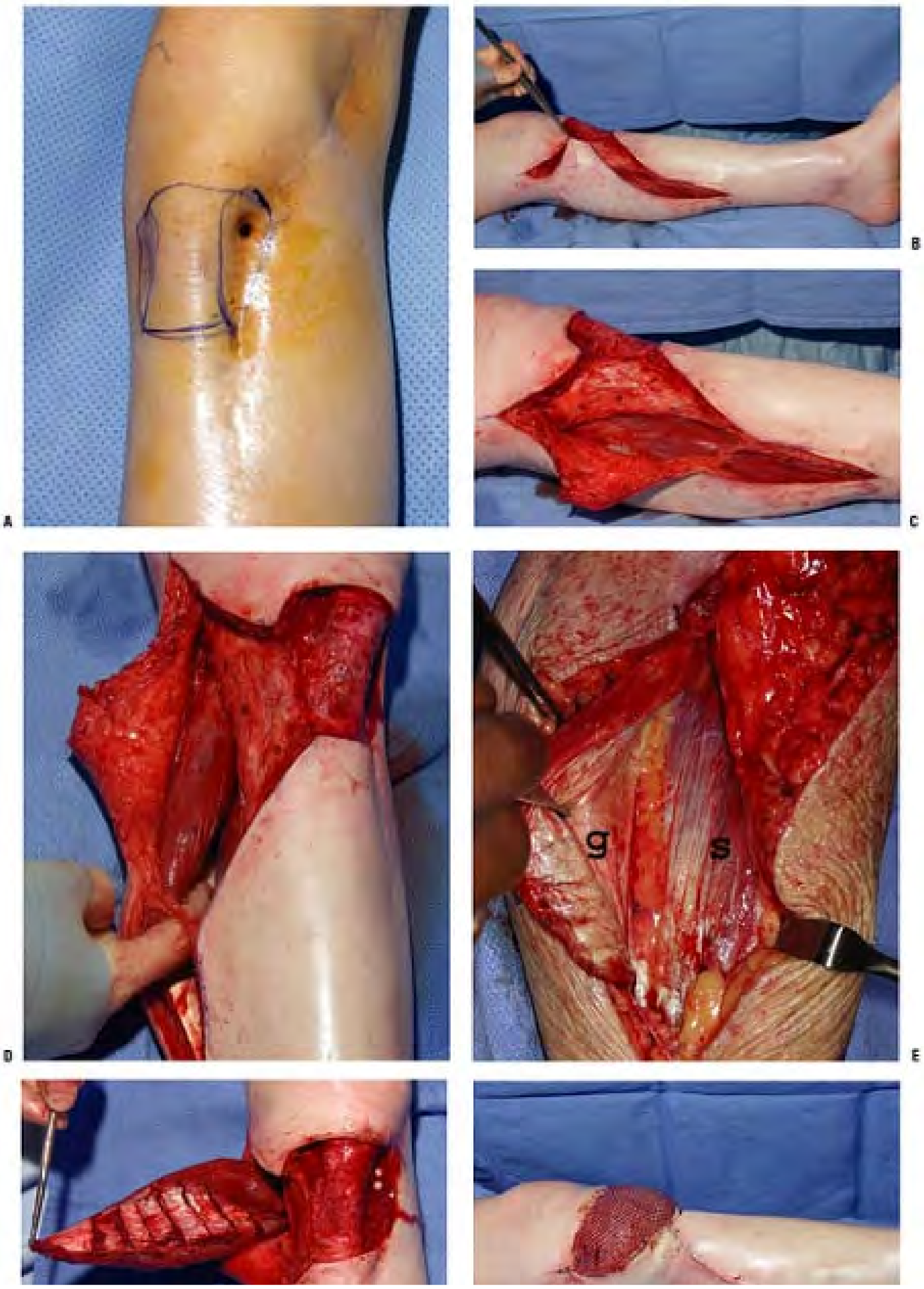

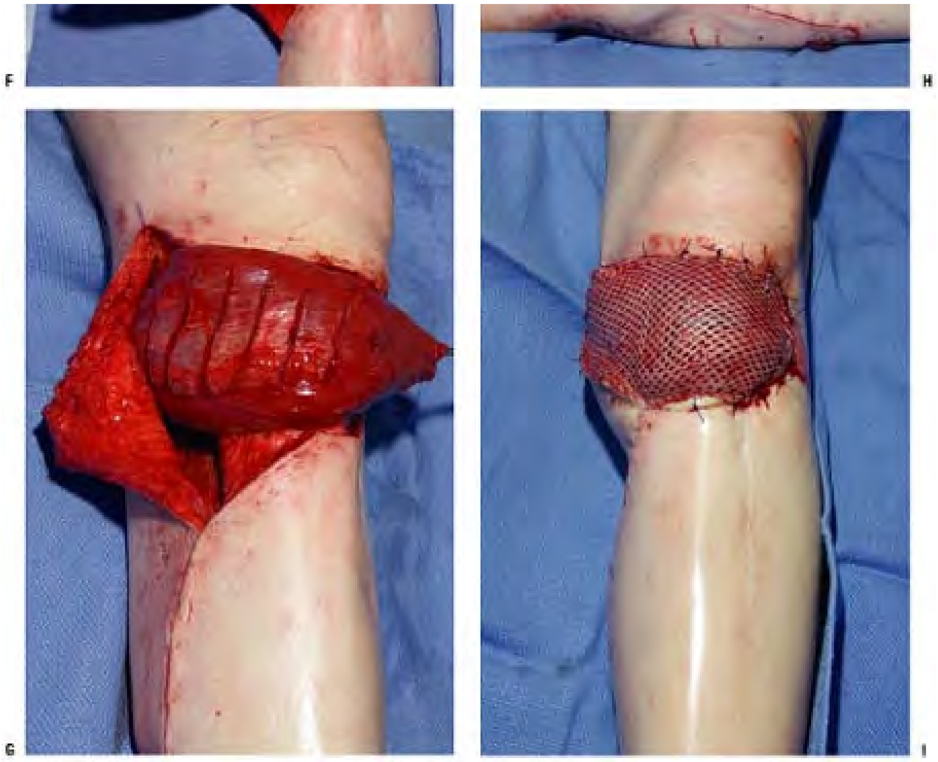

FIGURE 26-2 A: A 53-year-old patient with longstanding osteomyelitis of the tibia requiring debridement and soft tissue coverage. B: The skin surrounding the defect is excised. The bone is debrided and filled with antibiotic beads. The skin defect is incorporated into the skin incision for exposure of the gastrocnemius muscle. The skin incision is curved posterior approximately 4 cm beyond the medial border of the tibia. C: This incision gives excellent exposure of the posterior compartment. Note that the saphenous vein has been preserved at the inferior extent of the skin incision. D,E: The deep investing fascia of the leg is divided and the plane between the gastrocnemius and the soleus muscle may be identified using finger dissection. F,G: Once the muscle is mobilized, the anterior and posterior fascia may be transversely scored to allow for easier insetting. H,I: Once the muscle is inset, it is covered with a split thickness skin graft. The leg incision is closed over a suction drain. g, gastrocnemius; s, soleus.

Dissection is then continued medially between the plane of the gastrocnemius and the soleus until the midline raphe of the gastrocnemius muscle is identified. Often this raphe will be identified by decasations of the muscle fibers distally. More proximally, the two heads may be more easily separated and identified if the raphe is not clearly visualized distally. Once the limits of the medial muscle have been clearly identified, the inferior margin of the muscle is divided from the Achilles tendon with a 2-cm cuff of tendon, which facilitates insetting of the muscle. This cuff of tendon is also very valuable for reconstruction of the terminal extensor mechanism of the quadriceps muscle.

The sural nerve will be identified within the gastrocnemius raphe and should be preserved. Using cautery, the raphe can be divided and the gastrocnemius can be separated from the Achilles tendon. At this point, the muscle may be pedicled and mobilized for coverage of the upper one-third of the lower leg. If the knee or distal thigh requires coverage, the origin of the muscle will require release. For release from the femoral condyle, posterior extension of the skin incision significantly facilitates visualization of the muscle's vascular pedicle. Skeletonization of the sural vessels from the deep surface of the muscle will also allow further mobilization of the muscle. The muscle can be transferred with either its deep or superficial surfaces exposed. Cross-hatching or removal of the fascia will allow for further mobility and advancement of the muscle without impairing blood supply (Fig. 26-2F,G).

The fascia is then inset into the defect using horizontal mattress sutures, pulling the muscle beneath the overlying skin on the far side of the defect. If there is excessive tension, sutures may be mattressed through the skin and tied over bolster dressings. Prior to inset, a 10-mm flat or round drain can be placed under the flap; a second drain should be placed at the donor site prior to closure.

A meshed skin graft should then be obtained and placed on the transposed muscle flap, followed by a bolster dressing. This dressing should include a nonadherent layer followed by a bolster with gentle compression maintained on the graft. A VAC sponge may also be applied to facilitate skin graft take to the underlying muscle (Fig. 26-2H,I). The remainder of the wound should be covered with sterile dressings and the leg placed in a knee immobilizer for support. The skin graft donor site may be dressed with a single-layer transparent dressing.

Lateral Gastrocnemius Muscle

The lateral gastrocnemius muscle will cover the inferior thigh, knee, and proximal tibia. For harvesting of the lateral gastrocnemius the skin incision is made at the posterior midline of the lower calf extending to the lateral popliteal fossa. The proximal mark is placed 2 or 3 cm posterior to the fibula. Once again, an adequate skin bridge must be maintained between the traumatic defect and the incision to avoid skin bridge necrosis. Alternatively, the existing skin defect may be extended proximally and distally over the posterior margin of the fibula to allow for exposure of the muscle (Fig. 26-3).

Dissection begins 5 to 6 fingerbreadths below the fibular head so as not to encounter the peroneal nerve. Once the plane between the soleus and gastrocnemius is clearly identified, superficial dissection can begin between the gastrocnemius and the investing fascia of the lower leg. The common peroneal nerve must be identified at the level of the neck of the fibula and protected throughout the rest of the dissection. Since the common peroneal nerve passes between the lateral head of the gastrocnemius and the tendon of the biceps femoris, particular attention should be paid to identify the nerve during this portion of the dissection. The muscle is then separated as described for the medial gastrocnemius muscle flap, and may then be pedicled into the overlying defect site (Fig. 26-4).

The flap is inset, and the donor site closed as previously described for the medial gastrocnemius flap.

Postoperative Management

The patient should be maintained on bed rest for 5 to 7 days. The bolster dressing may be removed on postoperative day 5 to ensure adequate skin graft â€œtake.â€• Following skin graft take, Xeroform dressing changes are performed once to twice a day until the skin graft has matured, at which point the patient is encouraged to keep the skin graft lubricated with a petroleum-based lotion to prevent desiccation of the newly grafted skin. If immediate postoperative motion is essential, the origin of the gastrocnemius muscle should be divided to minimize forces across the flap in the early postoperative period. A VAC sponge may also aid in securing skin grafts if immediate motion is essential.

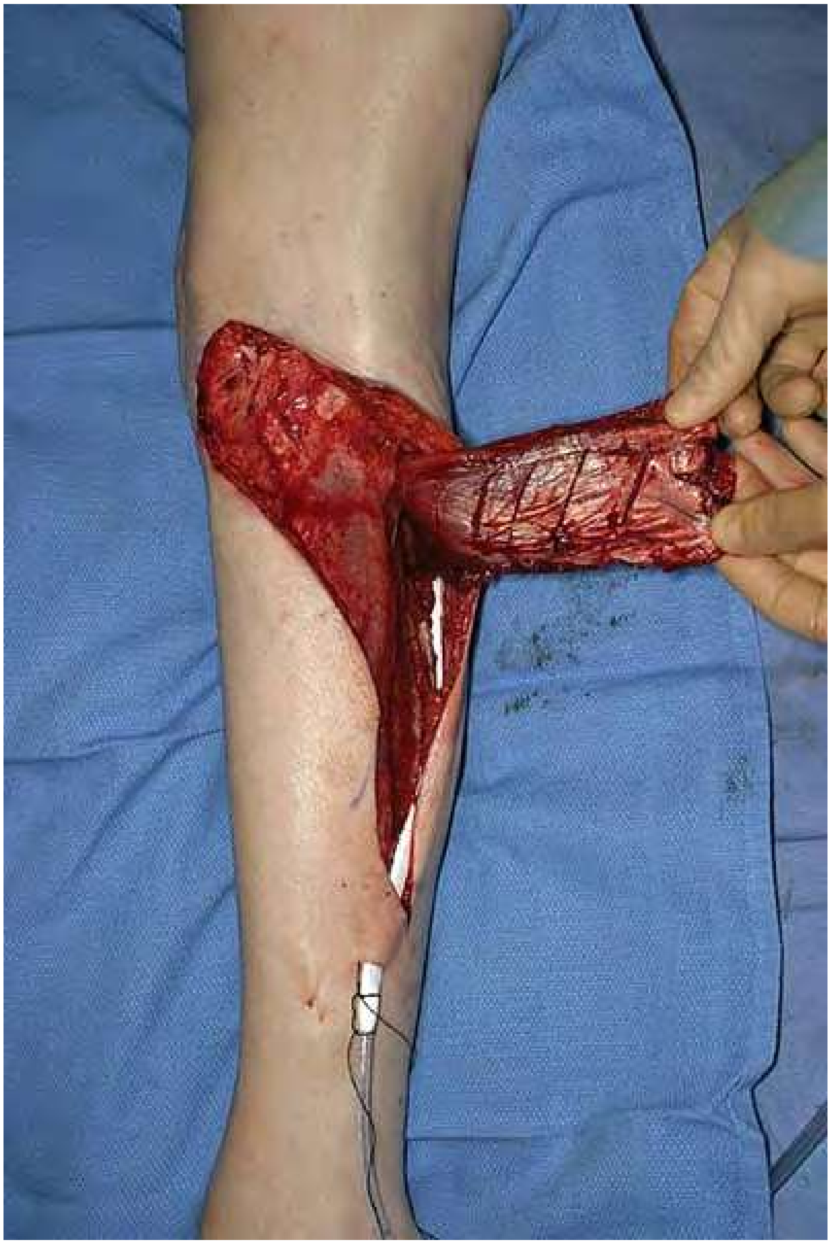

FIGURE 26-3 The incision for exposure of the lateral gastrocnemius muscle flap is made 2â€“3 cm posterior to the fibula.

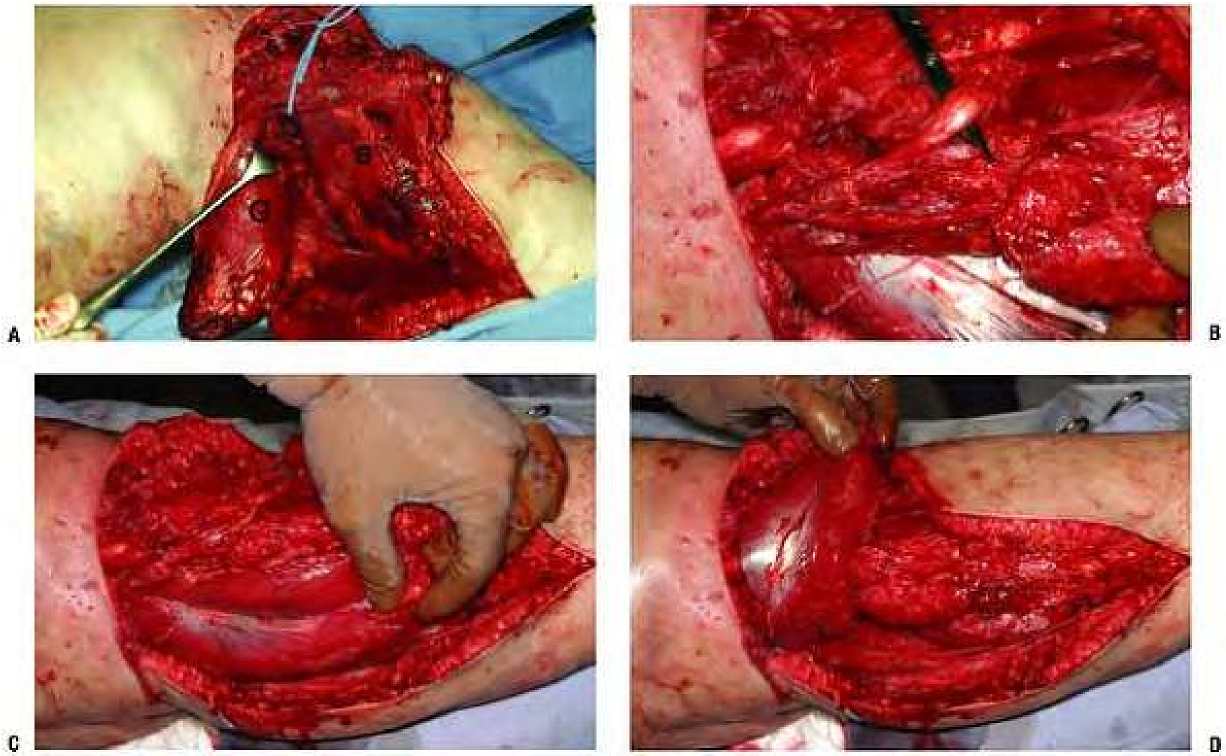

FIGURE 26-4 A: A traumatic lateral tibial and patellar defect in 28-year-old woman. The lateral gastrocnemius muscle has been harvested with a posterior and inferior incision extended from the original defect site. Note that the common perineal nerve has been identified and marked with blue vessel loop. The soleus muscle belly (*s*) can be seen below the gastrocnemius muscle (*g*). B: The common perineal nerve (seen here running above the scissors) must be identified prior to pedicling the flap into the defect site to ensure it is not inadvertently compressed or injured during flap insetting. C,D: Once the lateral gastrocnemius muscle is isolated from midline raphe, it can easily reach the lateral aspect of the peri-patellar region.

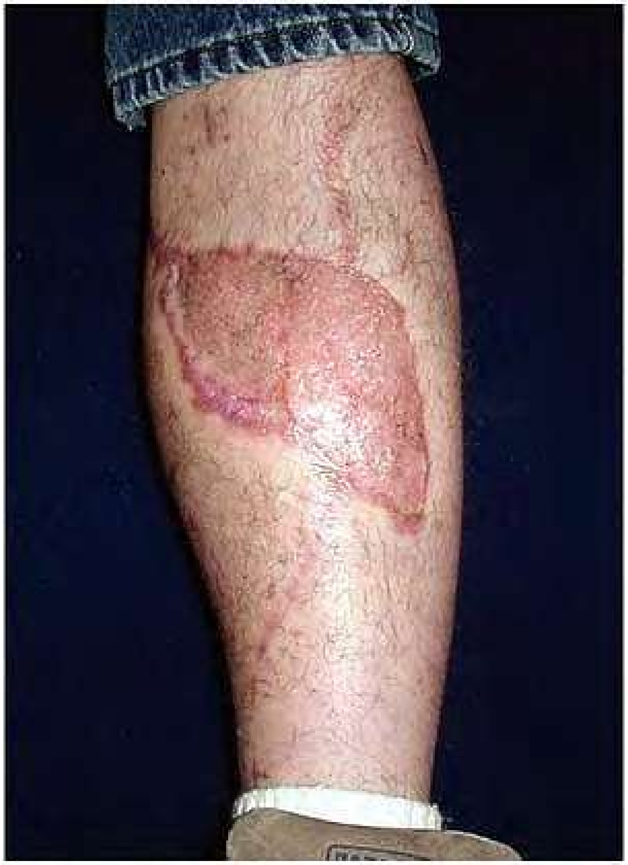

FIGURE 26-5 A 2-year follow-up of a 35-year-old man who underwent medial gastrocnemius muscle coverage for exposed tibia following a motor vehicle accident. Significant improvement in contour is obtained with long-term use of a compression wrap or compression stocking.

The drains are removed once drainage is less than 30 mL/day. Gentle range of motion should then begin at 10 to 14 days, increasing the motion by 5 or 10 degrees per day. Ambulation can begin on the postoperative day 10, with knee flexion limited with the use of a knee immobilizer. If the flap is inset under tension or if the healing environment is less than ideal, the knee immobilizer can be maintained for an additional week to prevent undo tension on the margins of the muscle flap.

Postoperative swelling is minimized with the use of elastic wraps or supportive stockings. Compression stockings worn long-term can facilitate flap contouring. Muscle flaps undergo a process of atrophy over 6 to 12 months; patient concerns regarding scaring or flap contour should not be addressed surgically until 12 to 24 months following the initial reconstructive surgery (Fig. 26-5).

Complications

Complications following the use of the gastrocnemius muscle flap include bleeding, infection, nerve injury, and flap loss. Bleeding (hematoma) and infection are most common, though rare if meticulous hemostasis is achieved at closure and aggressive debridement of the wound is performed prior to muscle flap transfer. Suction drains placed within the donor site at the time of closure will significantly minimize the risk of postoperative hematoma formation.

Partial flap necrosis or complete flap loss is usually associated with technical error, including injury to the vascular pedicle and/or excessive tension with secondary flap ischemia. If flap ischemia is suspected intraoperatively, the vascular pedicle should be examined for signs of kinking or injury to the sural artery. Small areas of skin graft loss may be seen postoperatively, but can be conservatively managed with topical ointments and wound care without the need for further surgical intervention.

Nerve injury is uncommon with meticulous dissection and nerve preservation. Injury to the sural nerve can occur during separation of the medial and lateral heads of the gastrocnemius muscle. Injury to the common perineal nerve may occur during elevation of the lateral gastrocnemius flap. Nerve injury is best avoided by clearly visualizing the nerve prior to flap transfer. The saphenous

nerve may be ligated inadvertently with the saphenous vein during exposure of the posterior compartment; attempts should be made at nerve identification and preservation during flap dissection.

The gastrocnemius muscle provides plantar flexion of the foot and flexion of the knee joint. Removal of the muscle does not result in significant morbidity, although some loss of plantar flexion may lessen jumping and leaping ability. Either or both heads may be used with little functional deficit as long as the soleus muscle is preserved and uninjured.

Recommended Reading

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 27

Revision and Infected Total Knee Arthroplasty

Henry D. Clarke
William J. Casey III
Mark J. Spangehl

Indications/Contraindications

Soft tissue management in revision and infected total knee arthroplasty (TKA) is an extremely important factor in determining the success or failure of the procedure. Without optimal wound healing, complication rates increase and the ultimate benefit of the operation may be diminished. Despite the critical role played by the soft tissues surrounding the knee, very little scientific data exist to help guide the practicing orthopaedic surgeon in this area. Rather, most of the recommendations and guidelines are based on extrapolations from other surgical fields and anecdotal experience from experts in knee arthroplasty. In this chapter, important factors that should be considered in the preoperative, intraoperative, and postoperative settings are reviewed. In addition, surgical techniques that may minimize potential soft tissue problems and help manage complications when they do occur are also presented. The soft tissues about the knee include skin and subcutaneous coverage, as well as the extensor mechanism; therefore, in this chapter management of both of these important components is addressed.

The very circumstances that necessitate revision TKA in the settings of either infection or aseptic failure require that every patient be considered at risk for soft tissue complications. Therefore, every patient in whom revision TKA is being considered should be thoroughly evaluated preoperatively. This allows all necessary plans to be addressed in order to meet individual needs. Indeed, absolute indications and contradictions are difficult to define: rather, all factors should be considered. While any patient who undergoes revision TKA should be considered at risk for soft tissue complications, some patients will have greater risks. A healthy patient with just one prior knee surgery, with aseptic loosening, is at less risk than a patient with diabetes and peripheral vascular disease who has a chronically infected total knee replacement. However, each patient should be carefully assessed. Certainly, once the risks have been evaluated, a decision to proceed with revision TKA surgery should be based on the individual circumstances. In each case, the current symptoms, including pain and disability, presence or absence of prosthetic infection, and age and activity demands of the patient, must be evaluated in the context of the potential risks. The potential increase in symptoms, time frame for deterioration, and additional surgical problems that may be encountered by delaying surgery should also be considered. Absolute contraindications for revision TKA include irreversible medical comorbidities that raise the risk of perioperative mortality to unacceptable levels, an avascular extremity where revascularization options have been exhausted, uncontrolled sepsis, and neurologic injuries with no motor function of the extremity. Relative contraindications include failed prior soft tissue flaps, massive bone loss, recurrent prosthetic infection, extensor mechanism disruption, and unstable medical comorbidities that require optimization. When the potential risks are too great, or when other rare circumstances are encountered, such as life-threatening sepsis or an avascular limb, alternative procedures including permanent resection arthroplasty, knee arthrodesis, or amputation may be required.

Preoperative Planning

Preoperative planning is critical for minimizing soft tissue problems about the knee. Both systemic factors and the characteristics of the knee and leg should be considered before revision knee surgery.

Systemic Factors

Patient-specific medical comorbidities should be assessed and treated before surgery (Table 27-1). While optimization of these conditions appears to reduce the risk of developing wound complications after revision TKA, the increased risk of wound problems in many cases is multi-factorial and does not return to baseline levels even when medical management is optimal. For example, the increased risks of delayed wound healing and infection noted in patients with diabetes appears to be caused by both the hyperglycemia, which has been shown to inhibit collagen synthesis, disrupt fibroblast proliferation, slow capillary in-growth, and deficiencies in polymorphonuclear neutrophil cell function. Furthermore, the increased risk of soft tissue complications in patients with rheumatoid arthritis is likely due to numerous factors including the long-term sequelae of corticosteroid use, which regulates macrophage function, reduces collagen synthesis, and delays vascular in-growth, as well as direct causes of the disease itself such as skin atrophy, decreased albumin, and vasculitis.

Another important risk factor in the pre- and postoperative periods is the use of tobacco products. Nicotine and its metabolites cause vasoconstriction, which interferes with micro-circulation; however, the effects on wound healing are more profound than from vasoconstriction alone. Other constituents of tobacco smoke, such as carbon monoxide, reduce the oxygen carrying capabilities of hemoglobin, which reduces tissue oxygenation. Nicotine also appears to have a direct effect on fibroblast and immune function. The optimal time for smoking cessation appears to be 4 to 8 weeks preoperatively, but even a week of abstinence appears to reduce the risk of complications. Abstinence in the entire postoperative period is also critical. For elective procedures, all patients who smoke and have other risk factors for wound healing must stop smoking before the procedure. In some cases, professional services must be used to achieve this end. The use of nicotine delivery substitutes are controversial, as these do not eliminate the patient's use of nicotine but do eliminate multiple other toxic substances that are inhaled in tobacco smoke, such as carbon monoxide and hydrogen cyanide. Therefore, while not an optimal solution, in certain circumstances these substitutes may be preferable to reduce the overall risk.

Table 27-1. Complicating Comorbidities

Condition	Evaluation	Management
Diabetes	Blood glucose	Optimize diet and medication.
Malnutrition	Total lymphocyte count < 1,500 mm^3 Serum albumin < 3.4 g/dL	Nutritional supplements before surgery
Inflammatory arthritis	Review disease modifying medications	Hold medications, if possible, until wound healing has occurred.
Tobacco smoking	Serum kotinine level if poor abstinence/compliance	Avoid tobacco use 1 month before surgery.
Anemia	Hgb/Hct	Correct before surgery with erythropoietin injections.
Hypoxia	SaO^2	Optimize pulmonary function.
Peripheral vascular disease	Peripheral pulses, Doppler ultrasound, arteriogram	Revascularization prior to revision TKA

Obesity	BMI	Avoid significant weight loss in the immediate pre- or postoperative periods.
Cancer	History	Avoid elective surgery, if possible, while on chemotherapeutic agents.
Chronic corticosteroid use	History	> 2 years duration appears to increase risks of wound complications

BMI, body mass index; TKA, total knee arthroplasty.

The Knee

The local vascular anatomy of the knee can be unforgiving; in comparison to the hip, where wound necrosis is rarely a problem. Most of the perfusion to the skin is derived from perforator vessels that originate below the level of the deep fascia. Little subcutaneous communication occurs; rather, the perforating vessels connect at the level of the sub dermal plexus. Thus, when raising skin flaps about the knee, the dissection should be below the level of the deep fascia, rather than in the sub dermal plane to avoid damaging this superficial plexus. While the peri-patellar plexus formed by the medial and lateral, superior and inferior genicular arteries has been well described, the majority of the inflow to the skin appears to be derived from the medial side of the knee in the distribution of the saphenous and descending genicular arteries. Therefore, laterally based flaps should be avoided and any prior laterally biased vertical incision should be carefully considered when selecting the optimal incision. On the lateral side, the perforators from the superior and inferior genicular arteries are relatively more important, and it has been reported that use of a lateral patellar release with disruption of the superior genicular artery is associated with decreased tissue oxygenation and increased lateral wound problems. Some general guidelines for selection and placement of skin incisions follow.

A single, longitudinal anterior midline incision provides the most extensile exposure and is preferred for all revision and infected total knee replacement cases. Prior incisions about the knee should be carefully evaluated. Factors that are believed to be important, although little scientific data exist to support these suppositions, include age, orientation, length, and placement in relation to other or intended incisions. A single, transverse incision can be crossed by a new perpendicular anterior incision with relatively little concern. When a single, prior longitudinal anterior incision exists, this incision should be used. If placement of this prior incision is not directly midline, the proximal and distal ends of the incision may be extended back toward the midline to reduce tension on the wound. In these cases, subcutaneous dissection should be minimized; it is critical to maintain full-thickness flaps if the subcutaneous tissues must be mobilized to allow adequate exposure of the extensor mechanism. To preserve the sub dermal plexus, these flaps should be raised at the level of the deep fascia. If the prior incision is located far from the midline, and would necessitate creation of a large laterally based subcutaneous flap, one may consider using a new incision. In these circumstances, skin bridges greater than 3 to 5 cm should be maintained between the new and old incisions. When multiple incisions, or a single anterior incision located well away from the midline, are encountered, selection of the optimal incision should be carefully considered. In some circumstances where numerous prior incisions exist, or where the skin and subcutaneous tissues have been severely damaged by prior trauma, radiation exposure, infection, or are atrophic, due to systemic conditions, preoperative soft tissue management techniques that are described in following sections, such as soft tissue expansion or soft tissue flaps, may be required. Again, little definitive data exist to guide the surgeon when these interventions are absolutely required; rather, much of this assessment is subjective.

Surgery

Soft Tissue Procedures

Sham Incision

Technique

The sham incision or delay procedure is reviewed primarily for its historical significance as an early attempt to manage patients considered at risk for wound healing problems after TKA. Approximately 10 to 14 days before the planned knee procedure, the intended skin incision is made and extended to the level of the extensor mechanism. Next, the skin is undermined on both sides of the incision to expose the extensor mechanism, as required for the upcoming knee replacement, and then the incision is closed and observed. If the wound heals without complications, the intended knee procedure is subsequently performed. If skin necrosis occurs, a soft tissue flap is performed before knee replacement. The main advantage in these cases is that the knee joint has not been violated, and there is no prosthesis at risk for infection. Furthermore, increased collateral flow caused by the incision due to the delay phenomenon was felt to reduce the risk of subsequent wound

healing problems. More contemporary soft tissue expansion techniques have, for the most part, eliminated the use of the sham incision.

Results

There is little except anecdotal evidence to support the use of this technique. Rothaus has detailed the outcomes in a small group of 12 patients, with multiple prior incisions, in whom this technique was used. In all 12 patients, the incisions healed and TKA was successfully performed without wound healing problems.

Soft Tissue Expanders

Technique

The concept of tissue expansion is not new and has been successfully used in a variety of soft tissue reconstructive procedures throughout the body. Despite the challenges posed by the unforgiving nature of the vascular supply around the knee, the use of tissue expanders is a relatively new concept with only a limited number of large patient series.

Tissue expansion has three important benefits that aid wound healing in the overlying soft tissues. First, and probably most important, it stimulates neovascularization of the overlying soft tissue; this improves the capillary in-flow that directly promotes healing. Second, it physically expands the overlying skin and subcutaneous tissue allowing larger areas, due to correction of malalignment or limb lengthening, to be covered. This also reduces the tension on the wound. This expansion is not simply a stretching and thinning of the overlying soft tissue, which would leave the flap less resilient, but actually a hypertrophy of the dermis and epidermis due to fibroblast stimulation, leading to an increase in the quantity of tissue available. Third, a fibrous capsule develops around the expander, and this thick, robust, vascular tissue is very useful as a â€œpseudo-fascialâ€• layer to close over the joint and is particularly helpful in the area of the proximal medial tibia where the periosteum and soft tissue layer can be thin.

The technique for soft tissue expansion over the knee is begun approximately 6 to 8 weeks prior to the intended TKA (Fig. 27-1). Typically two to four 200 to 300 mL expanders are placed in pockets deep to the subcutaneous layer, just above the fascia (see Fig. 27-1C). These pockets are strategically placed based on the placement of prior incisions and characteristics of the overlying subcutaneous tissue and skin. The first step involves infiltrating the deep subcutaneous layer with a dilute solution composed of 1000 mL Ringer's lactate, 50 mL 1% lidocaine, and 1 mL 1:1000 epinephrine to create a pocket. Typically, injection of 250 to 300 mL of this solution is sufficient to produce the hydro-dissection that separates the skin and subcutaneous tissue from the underlying fascia to create a pocket. Also, use of the local anesthetic helps with postoperative pain relief. Next, through a short incision placed in one of the prior incisions, scissors are used to bluntly define and enlarge this plane within the deep subcutaneous tissue to create a pocket (see Fig. 27-1B). The un-inflated expander is then inserted through the same incision (see Fig. 27-1D). The access port to the expander is then tunneled through the subcutaneous tissue to an easily accessible site. The port should be accessed prior to skin closure and saline injected to ensure that the port is functional and that the expander was not damaged during the insertion. After this initial expansion, the incision is closed and a sterile dressing applied. Patients are typically placed in a knee immobilizer for a week and allowed to weight bear as tolerated. Most patients are admitted for an overnight stay and treated with 24 hours of intravenous antibiotics. Subsequently each week, about 10% to 15% of the volume of the expander is infused with the patient as an outpatient, via the access port. If at any time the overlying skin blanches and doesnâ€™t recover after a few minutes of observation, or if the patient experiences significant pain, saline must be removed until the problem is alleviated.

At the time of the intended TKA, the expanders are extracted through the incision used for the procedure (see Fig. 27-1F). The subcutaneous tissue and skin flaps should be protected during the TKA. At the end of the procedure, a superficial drain is placed in each individual expander pocket. These drains are removed individually once output is less than 10 mL per 8 hours or 30 mL per 24 hours. This protocol appears to reduce the risk of subcutaneous hematoma formation. Early in the experience of tissue expansion about the knee when subcutaneous drains were not routinely employed, hematomas occurred more frequently and occasionally had to be drained to reduce tension on the overlying tissue. During wound closure, the expansion process may have created excessive amounts of soft tissue that need to be excised. In these cases, the edge of the flap, especially focusing on old widened scars, can be trimmed or removed. However, the tension on the wound should not be increased by resecting too much surplus tissue.

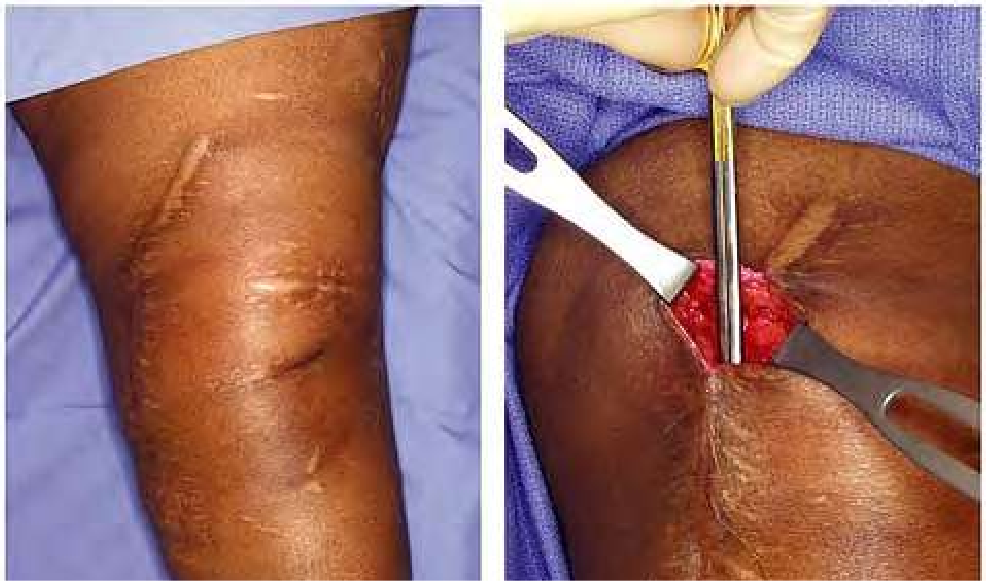

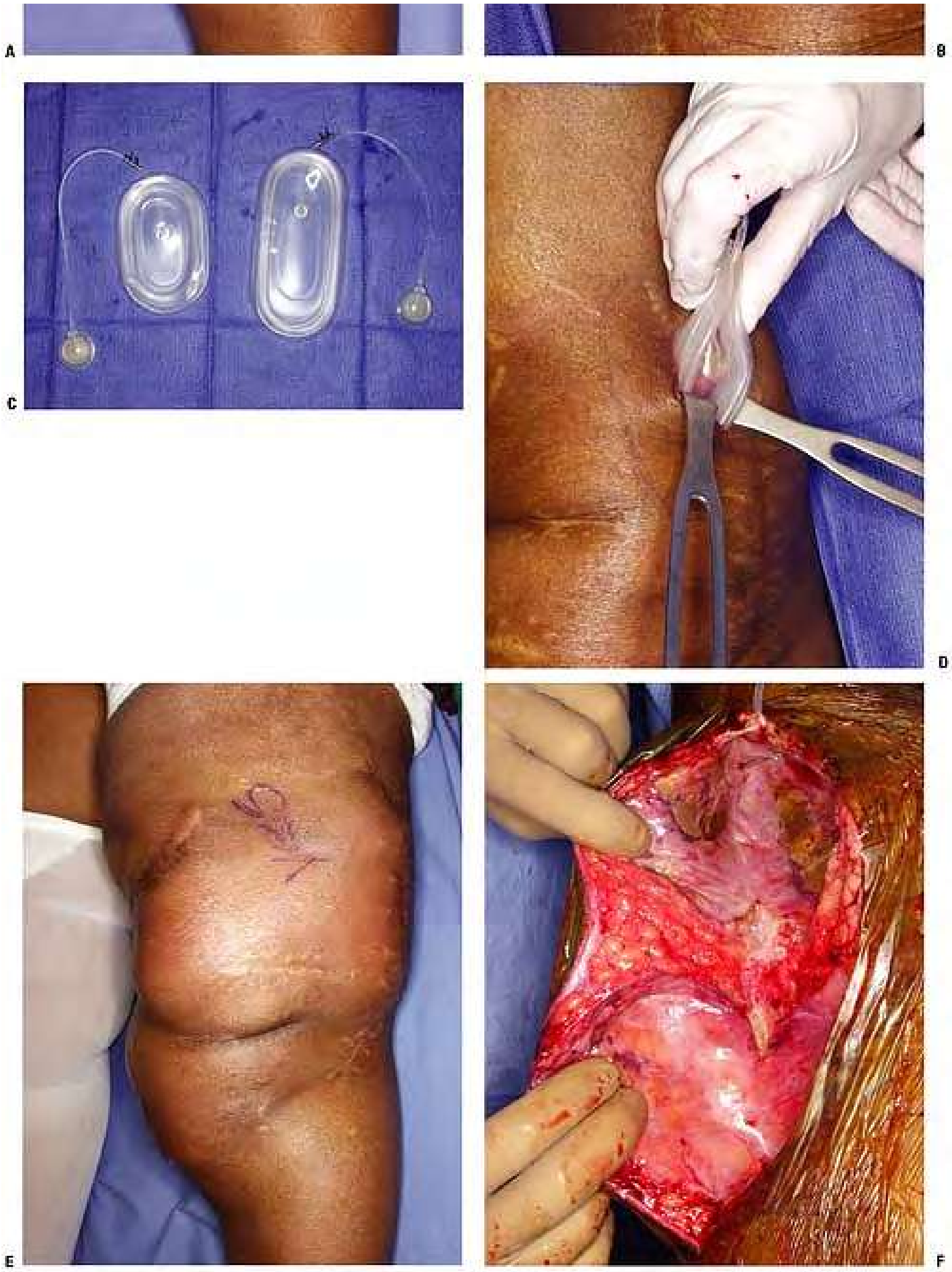

FIGURE 27-1 A: Knee with multiple prior vertical and horizontal incisions at risk for poor postoperative wound healing after TKA.B: Creation of subcutaneous pocket for soft tissue expander.C: Saline expanders demonstrating access portals for infusing saline. D: Insertion of the deflated saline soft tissue expander. E: Knee after tissue expansion process is

complete. F: Fibrous, vascular membrane that surrounds the expander cavity after removal of the expander at the time of TKA.

Results

Few large series of tissue expansion in association with TKA have been published, with most reports presenting only a handful of cases. The morbidity associated with the use of tissue expanders has been previously reported and includes infection, hematoma, expander rupture or deflation, and skin necrosis. Two significant papers have reported very favorable results. In an initial study, no significant complications occurred during the expansion process in a small group of 10 patients. In a second larger group of 29 knees, six (21%) minor wound complications occurred during expansion, including mild erythema or skin blistering. These complications were successfully managed by reducing the volume in the expander, and delaying further expansion until the problem resolved. In this same group, one major complication occurred after the insertion of the soft tissue expanders; this involved full-thickness skin necrosis in a patient with a history of radiation to the anterior knee. The knee replacement was subsequently not performed, as the patient declined to accept a prophylactic muscle flap, choosing to endure her arthritic symptoms.

Following the knee replacement, 5 (18%) of 29 patients experienced minor complications; three knees developed persistent drainage, and two knees developed subcutaneous hematomas. The persistent drainage resolved after wound compression and immobilization was initiated. The two subcutaneous hematomas required surgical evacuation, and it was these problems that prompted the use of subcutaneous drains in the expander pockets.

Soft Tissue Flap Coverage Procedures

Soft tissue flap procedures about the knee have been used both prophylactically and for salvage of wound complications about the knee. Although a large number of different techniques have been described, including simple skin grafts, random or axial pattern skin flaps, fasciocutaneous flaps, and rotational or free muscle flaps, the most reliable and most frequently used tissues about the knee include medial gastrocnemius muscle flaps, and free latissimus dorsi or rectus abdominus flaps. In addition to these techniques, the development of musculocutaneous perforator flaps during the past decade has provided another alternative for coverage about the knee. The principles involved in these three types of techniques are described following in ascending order of the extent of the area that can be adequately managed.

Musculocutaneous Perforator Flaps

The development of musculocutaneous perforator flaps in plastic surgery occurred during the 1990s as an extension of the generally very favorable experience with musculocutaneous flaps during the 1980s. Perforator flaps were based on the principle that neither the underlying fascial plexus, nor muscle, was required if the vessel that perforated through the muscle was carefully dissected. The major advantage of musculocutaneous perforator flaps versus musculocutaneous flaps is that the functional and cosmetic morbidity due to transfer of the underlying muscle could be avoided by transferring only the overlying skin.

Although perforator flaps can provide additional vascularized tissue in an area that may be deficient or fibrotic, their use in the setting of infection is controversial; and in these cases, traditional muscle flaps are preferred. Their use is best reserved for aseptic cases with a deficient or scarred soft tissue envelope.

Technique

The wound is first debrided to remove all necrotic tissue (Fig. 27-2A,B). Next, elevation of the perforator flap requires careful microsurgical dissection of the musculocutaneous perforator vessels (Fig. 27-2C). Due to anatomical variability of these perforating vessels, Doppler ultrasound is used to locate the perforator before elevation of the flap. Once isolated, the perforator flap can be rotated about the vascular pedicle or transferred as a free flap (Fig. 27-2D). The donor site can be managed with either primary closure or split-thickness skin grafts depending on the specific perforator flap used (Fig. 27-2E). A number of perforator flaps have proven quite reliable for managing soft tissue problems in the lower extremity; these include flaps based on the muscle perforators of the vastus lateralis, tensor fascia lata, sartorius, gracilis, and medial gastrocnemius. Unfortunately, due to the volume of tissue required in many cases associated with failed TKA, these muscle perforator flaps often would not provide enough soft tissue to cover the defect; in these circumstances, one of the alternative techniques must be considered.

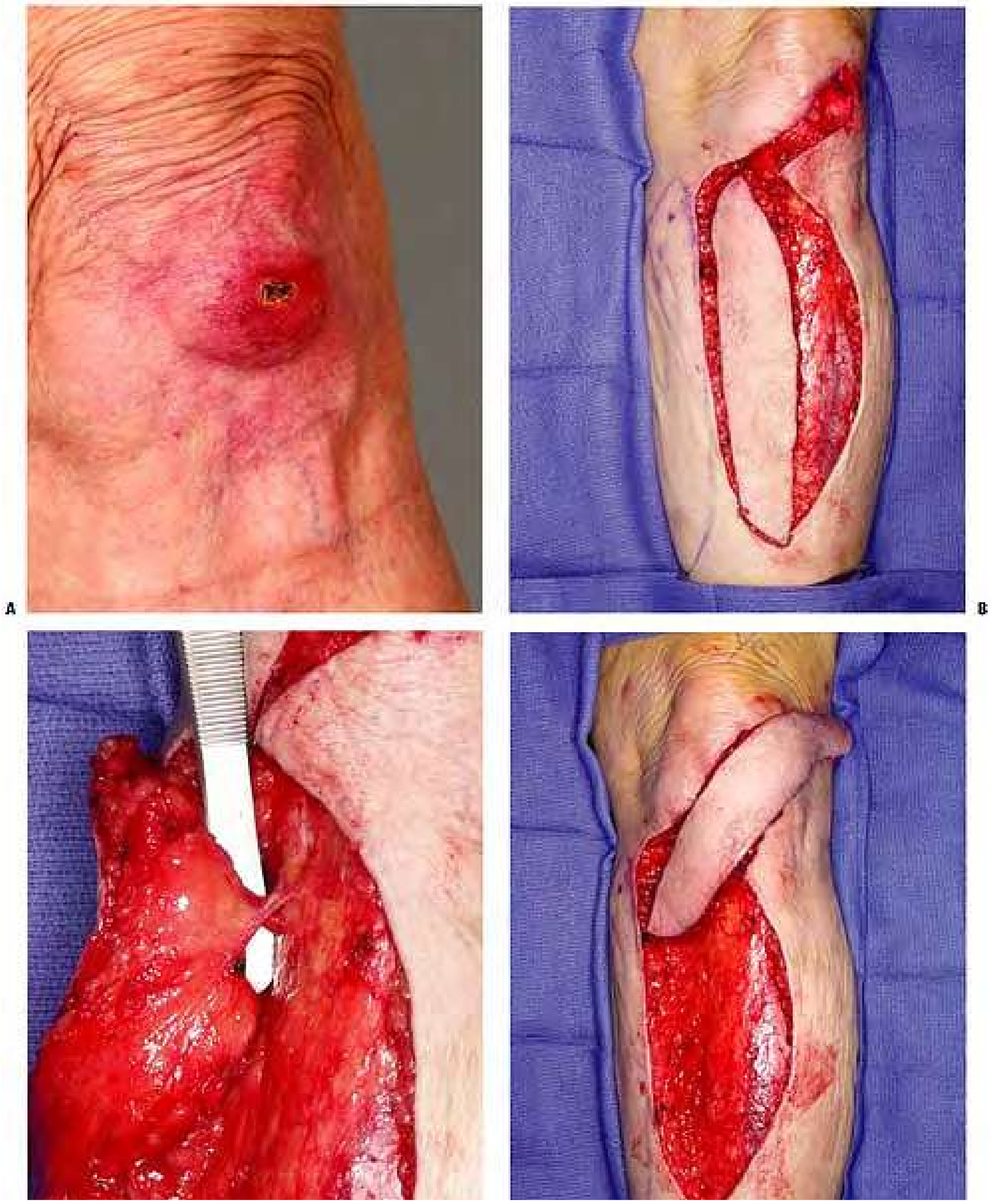

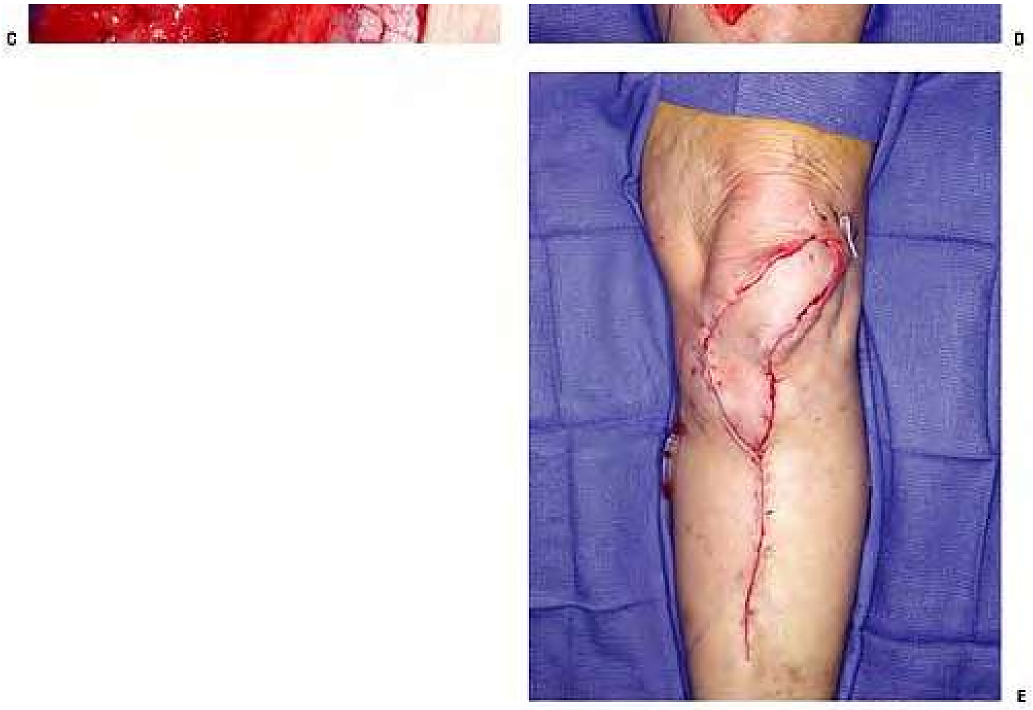

FIGURE 27-2 A: Pre-patellar bursa with central necrosis and recurrent infection that failed numerous attempts at local wound care. B: A perforator flap has been harvested after identifying the vessel, and the pre-patellar area has been debrided. C: Perforator vessel to flap. D: The flap is rotated 180 degrees on its pedicle and transferred to the area of the defect. E: Harvest site incisions have been closed primarily and the flap sutured in place.

Medial Gastrocnemius Flap

Technique

Although both the lateral and medial heads of the gastrocnemius muscle can be used to provide soft tissue coverage about the knee, both the quantity of tissue provided by the lateral gastrocnemius, and its shorter vascular pedicle, restrict its use. The lateral head is generally 3 to 4 cm shorter than the medial head, and its transfer can place the common peroneal nerve at risk for injury as the muscle is passed across the proximal fibula. Therefore, the medial gastrocnemius rotational flap has been the preferred option for treating wound problems in association with total knee replacement (Figs. 27-3 and 27-4). First, the wound is debrided to healthy margins. Next, the medial

gastrocnemius muscle flap is developed through a posteromedial incision, separating the muscle belly away from the deep tissue superficially, and the underlying soleus muscle along its deep surface. The plantaris tendon provides an excellent landmark. The medial head is separated from the lateral head along its decussation, which does place the sural nerve at risk for injury. The vascular pedicle is based proximally on the sural artery, which is the first branch of the popliteal artery. The muscle is then divided distally at the musculotendinous junction and folded proximally. Next, the flap is tunneled under a medial skin bridge to cover the defect, where it is sutured in place (see Fig. 27-4). If added length is needed, radial scoring of the fascia along the undersurface of the muscle can be performed at 1-cm intervals to increase its arc of rotation. Once transferred, the muscle flap is covered with a split-thickness skin graft that can be performed either primarily, or at a later date; in most circumstances, the donor site can be closed primarily. Alternatively, the flap may be harvested as a musculocutaneous flap, where the entire soft tissue flap including the skin, subcutaneous tissue, fascia, and muscle is rotated about the vascular pedicle. In these cases the donor site must be skin-grafted (see Chapter 29).

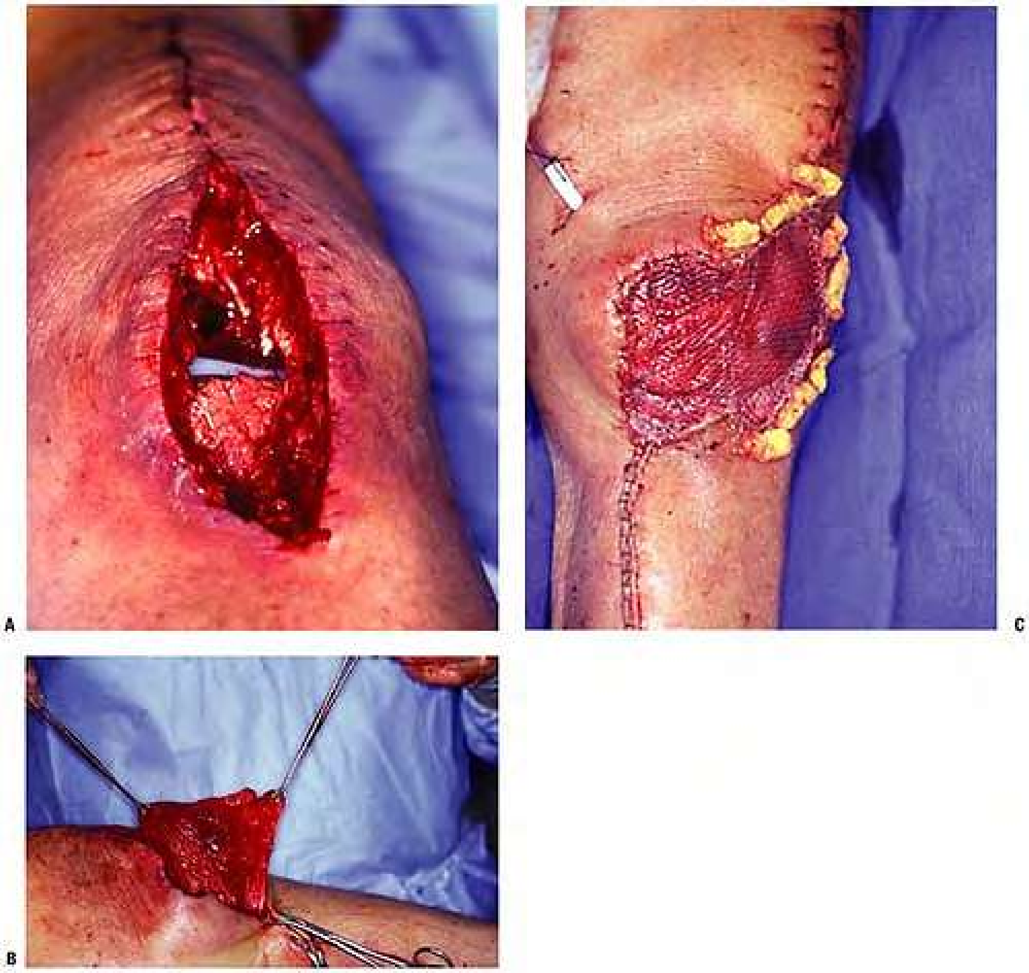

FIGURE 27-3 A: Debrided wound after full-thickness necrosis over proximal tibia after TKA. B: The medial gastrocnemius is harvested through a posteromedial incision of the calf and then tunneled under a skin bridge to the debrided area. C: The muscle flap is covered with a split-thickness skin graft.

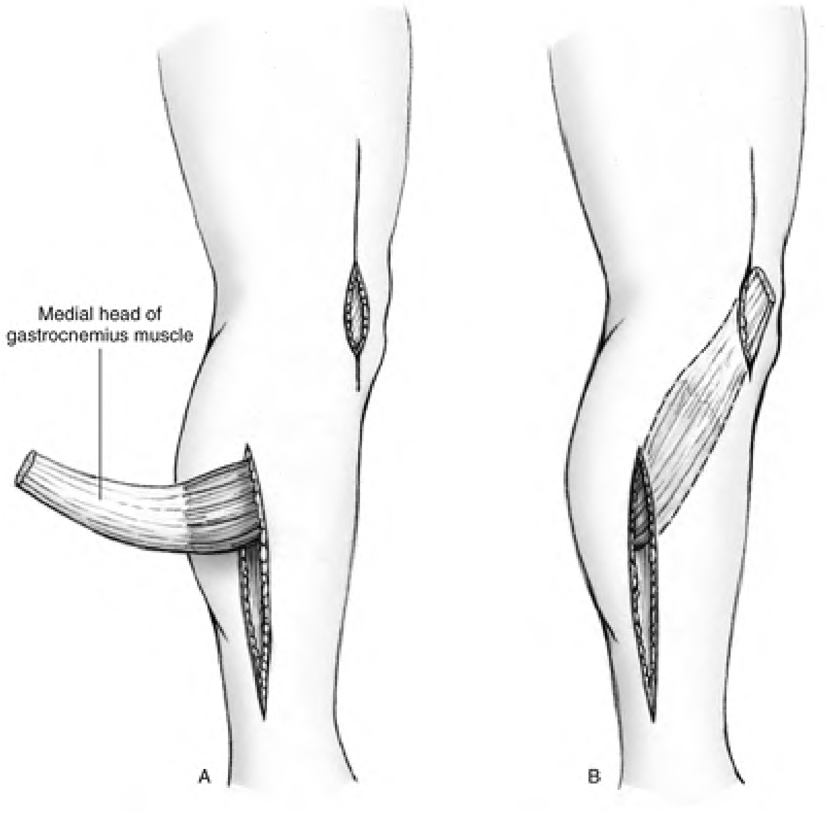

FIGURE 27-4 A: The medial head of the gastrocnemius muscle is harvested through a medial calf incision. B: The muscle is then rotated and delivered a skin bridge into the defect.

Results

In association with infected or exposed total knee prostheses where wound breakdown has occurred, use of medial gastrocnemius flaps has been quite successful, considering the complexity of the problems. In one series, Gerwin et al reported their experience from the Hospital for Special Surgery; successful salvage of an infected, or exposed total knee prosthesis with a medial gastrocnemius flap was achieved in 10 of 12 patients. Similar results have been reported by McPherson et al, with successful reimplantation in 20 of 21 patients who had chronically infected total knee replacements and compromised soft tissues about the knee.

Utilization of the gastrocnemius muscle for knee coverage results in little functional morbidity, with no significant deficits at walking speed, and only mild deficits as demand increases.

Free Latissimus Dorsi and Rectus Abdominus Muscle Flaps

In cases where the medial gastrocnemius flap has been previously used and failed, or when the extent of the soft tissue necrosis is large, alternative flaps must be considered. The free transfer of a latissimus dorsi or rectus

abdominus muscle flap has been reliable in these difficult circumstances. The muscle is harvested with its vascular pedicle and the donor site primarily closed. The recipient site is debrided to healthy wound edges and then the free muscle is sutured into the defect after the vascular anastomosis is complete. The vascular anastomosis is typically performed to the popliteal vessels.

Postoperative Management of Soft Tissue Flap Coverage Procedures

Immobilization with elevation for 7 to 10 days to reduce venous congestion is common to all flap coverage procedures. Partial weight bearing and progressive range of motion are then begun on a case by case basis. Factors to be considered include any wound drainage, marginal necrosis, venous congestion, and success of the overlying skin graft. If the flap transfer is done prior to the TKA as a prophylactic measure, then the total joint replacement may be performed after the flap has matured, which is generally a minimum of 8 to 12 weeks.

Complications of Soft Tissue Flap Coverage Procedures

While successful retention or reimplantation of the knee prosthesis has been reported, the functional results associated with these so-called successful outcomes falls far short of knee replacement in association with uncomplicated wound healing. In addition to poor functional results, other complications associated with all types of flaps can include recurrent infection or recurrent wound problems, due to marginal necrosis or complete loss of the flap, as well as skin graft problems at the donor or graft sites. Weakness caused by transfer of the donor muscle can also occur. Due to loss of the medial gastrocnemius, patients have reduced plantar flexion strength of the involved ankle; however, given the extent of the associated knee problems, the functional limitations due to loss of the medial gastrocnemius may be relatively minor. Cosmesis can also be an issue as the flaps may be quite bulky; however, this concern must be considered relative to the presenting problem. Occasionally, debulking or thinning of a mature flap will be undertaken where atrophy of the muscle was less than anticipated.

Surgical Exposure

In addition to the specialized plastic surgery techniques that may be required in revision TKA, the orthopaedic surgeon must also be aware of the techniques that allow adequate surgical exposure to be obtained during the knee procedure in a manner that minimizes risks to the skin and extensor mechanism. If an intraoperative extensor mechanism disruption occurs during revision TKA, or if a preoperative extensor mechanism failure is known to exist, the surgeon must also be aware of the techniques used to address these problems. Simple repair, especially when associated with chronic extensor mechanism disruption, has a very poor track record, and the alternative reconstructive options are described following.

General principles to optimize wound healing and avoid complications include meticulous soft tissue handling in every case. Prolonged local tension from self-retaining retractors or from vigorous stretching due to attempts to perform mini-incision surgery should be avoided; in addition, undermining along the margins of the incision should be minimized. Intraoperatively, the importance of optimizing component positioning to avoid the need for a lateral patellar release that is associated with increased wound problems is also important. The use of a tourniquet is controversial, but evidence suggests that postoperative tissue oxygenation is reduced in patients in whom a tourniquet is used. Certainly, in patients with significant peripheral vascular disease, especially those in whom either a bypass has already been performed, or the lateral radiograph shows extensive calcification of the popliteal vessels, consideration should be made to perform the knee replacement without a tourniquet. In addition to the previously noted principles, numerous techniques exist for facilitating surgical exposure and relieving tension on the extensor mechanism to reduce the risk of iatrogenic tendon rupture or patellar fracture. These techniques, along with the results and complications that are specific to each, are reviewed following. All of these procedures, except where noted, are performed during revision TKA with the patient positioned in the supine position. The specific techniques are described generally progressing from less extensile to more extensile.

Medial Parapatellar Arthrotomy and Proximal Tibial Peel

The medial parapatellar arthrotomy allows the optimal exposure in revision TKA and is versatile as it is compatible with each of the more extensile exposures detailed in subsequent sections. The arthrotomy is begun at the proximal end of the quadriceps tendon approximately 6 to 8 cm proximal to the superior pole of the patella (Fig. 27-5). The arthrotomy is extended distally approximately 3 to 5 mm lateral to the medial border of tendon. At the superior pole of the patella, the arthrotomy is curved around the medial

border of the patella and then continued distally along the medial edge of the patellar tendon. Next, the periosteum of the medial tibia is elevated sharply beginning at the medial border of the arthrotomy at the level of the joint. Working medially and distally, the entire medial periosteum is elevated approximately 5 to 7 cm distally. A periosteal elevator can be used once the flap of periosteum has been raised. At this stage the knee is flexed and the tibia is gradually externally rotated while the sub-periosteal elevation is continued all the way to the posteromedial corner. In many cases where this tissue plane was not violated in the primary replacement, the semi-membraneosus insertion on the posteromedial corner will be well defined, and this expansile insertion should be released in the sub-periosteal plane. As the dissection is continued to the posterolateral corner, and the tibia is externally rotated, the tibia will sublux from underneath the femur. In many cases, this will allow access to the modular tibial polyethylene. Once exposed, the tibial insert can be removed; this maneuver relaxes the flexion and extension gaps and usually improves the exposure. This exposure should be incorporated into every revision and provides the foundation for obtaining adequate visualization and access in order to safely remove the existing components and reconstruct the knee.

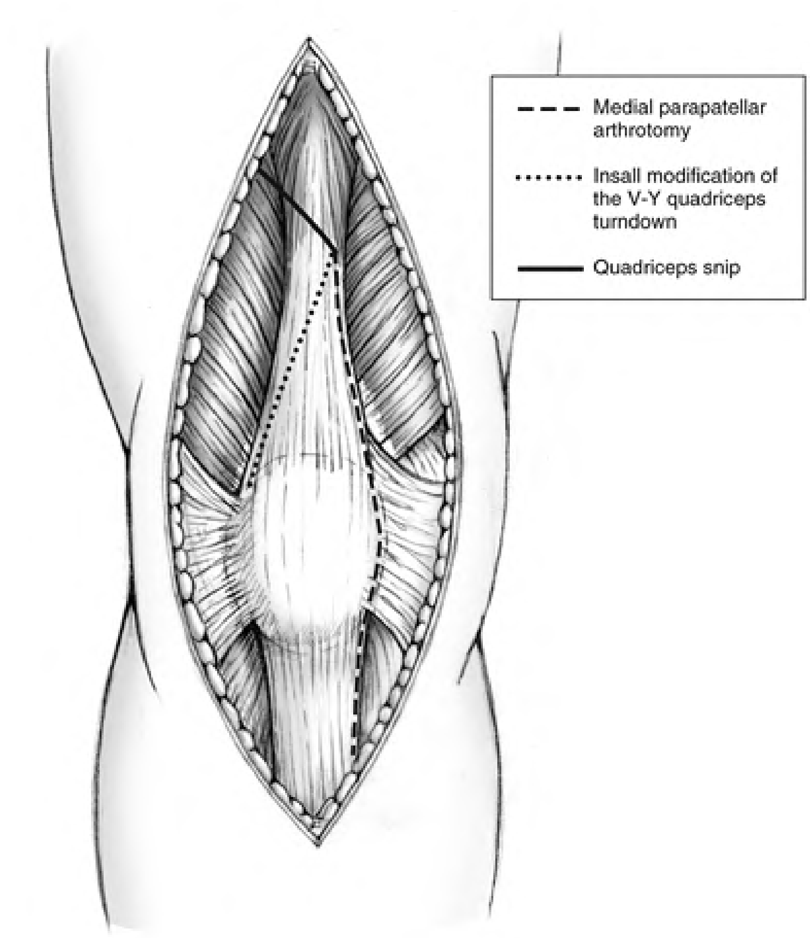
Medial parapatellar arthrotomy
Insall modification of the V-Y quadriceps turndown
Quadriceps snip

FIGURE 27-5 The medial arthrotomy (*dashed line*) is the standard exposure for revision TKA. This can be extended with a quadriceps snip (*solid line*) if the extensor mechanism is tight. The Insall modification of the V-Y quadriceps turndown (*dotted line*) allows extensile exposure in the stiff knee but is associated with postoperative extensor lag.

Quadriceps Snip

Technique

The quadriceps snip, originally promoted by Dr. John Insall, facilitates the exposure of stiff knees in the primary and revision setting. This technique is a proximal extension of a medial parapatellar arthrotomy (see Fig. 27-5). Beginning at the proximal apex of the standard arthrotomy, the incision is extended proximally and laterally at a 45-degree angle into the fibers of the vastus lateralis. This extension not only relieves the tension on the extensor mechanism in a stiff knee that can aid in exposure of the joint, but also allows the patella to be everted more easily during patellar preparation if this is required. The arthrotomy and this proximal extension are closed in routine fashion once the arthroplasty has been performed. In distinction to many of the alternative techniques for optimizing exposure in difficult cases, there is no need to modify or restrict postoperative rehabilitation. Indeed, patients in whom the quadriceps snip has been used may participate in standard TKA rehabilitation protocols. While this technique allows adequate exposure in most circumstances, in particularly stiff or difficult cases, additional measures may be required. In these extreme cases, a tibial tubercle osteotomy can be used even after a quadriceps snip has already been performed.

Results

Insall's clinical experience in 16 patients who had bilateral TKA with a quadriceps snip on only on one side has been reported. In these patients, no differences in quadriceps strength were observed postoperatively between the two sides. In a similar study, no difference was identified between patients who had undergone a quadriceps snip, versus those in whom a standard medial parapatellar arthrotomy had been performed.

Tibial Tubercle Osteotomy

This technique, popularized by Whiteside, facilitates exposure in even very stiff knees. In addition, the tibial tubercle osteotomy can allow access to the tibial canal that may be helpful in revision or infected TKA when well-fixed stemmed, tibial components must be removed. This technique involves first incising the medial periosteum along the medial border of the patella tendon for approximately 6 to 7 cm distal to the joint line. The osteotomy is then performed with an oscillating saw from medial to lateral, creating a wedge of bone that is approximately 2 cm wide and 6 to 8 cm long. The osteotomized wedge should taper from a thickness of about 1 cm proximally, to the level of the anterior cortex distally. This minimizes the stress riser in the anterior tibia, especially at the distal end. In addition, at the proximal end, a step cut should be created about 1 cm distal to the joint line. This step acts as a buttress that helps to resist proximal migration of the osteotomized wedge after it has been reattached at the end of the case. Proximal migration is also resisted by the lateral periosteum and musculature that should remain attached to the osteotomized piece. During the revision knee arthroplasty, the lateral soft tissues act as a hinge that allows the osteotomized segment to be everted and displaced laterally. At the time of closure, three to four 18-gauge wires are used for fixation. The wires are passed through drill holes in the medial tibia and then brought up through the lateral aspect of the tubercle shingle. They are then brought distally and medially over the tibial crest and tightened on the medial side of the tibia. Tightening the wires pulls the osteotomy distally. The arthrotomy is then closed in a standard fashion. In most cases, we favor bypassing the tibial tubercle with a stemmed component. If a cemented stem is used, it is helpful to remember to pass the wires before insertion of cement.

Results

Advantages of the tibial tubercle osteotomy versus the V-Y quadriceps turndown include a lower incidence of extensor lag and quadriceps weakness. Whiteside reported his experience with this technique in 136 patients; only two patients experienced a residual extensor lag. Another advantage of this technique is that if rigid fixation is obtained, full weight bearing and unrestricted range of motion are allowed postoperatively. Reported complications of tibial tubercle osteotomy that are not noted with the quadriceps turndown include proximal migration of the osteotomized tubercle, patellar tendon disruption, and tibial shaft fractures. Despite these unique problems, we favor the use of tubercle osteotomy when adequate exposure cannot be obtained with a standard arthrotomy and quadriceps snip.

V-Y Quadriceps Turndown

Technique

The V-Y quadriceps turndown was originally described by Coonse and Adams and subsequently modified by Insall. In both techniques an inverted V-shaped flap incorporating the quadriceps tendon is created. Insall's modification described creating a second incision beginning at the apex of a standard medial parapatellar arthrotomy that is extended distally and laterally at a 45-degree angle along the tendinous portion of the vastus lateralis (see Fig. 27-5). In addition to allowing excellent exposure, the inverted V of tissue may be lengthened at the time of closure, forming an inverted Y shape.

Results

Although this technique allows excellent exposure and releases excessive tension on the extensor mechanism during surgery, patients may develop an extension lag postoperatively. Trousdale et al reported the Mayo Clinic experience with this technique. Patients who had undergone bilateral TKA with a V-Y quadriceps turndown in one knee, and a standard medial parapatellar arthrotomy in the contralateral knee, underwent postoperative strength testing. In these patients, the results did not show any statistical differences between the strength of the two legs. However, 5 of 14 patients with the V-Y quadriceps turndown had a persistent extensor lag. Furthermore, use of the V-Y quadriceps turndown requires alterations in the postoperative rehabilitation protocol that includes

restricted range of motion, use of a brace, and partial weight bearing for about 6 weeks. Due to the high rate of lag, and the increased recovery time associated with the use of the V-Y turndown, we favor the use of the tibial tubercle osteotomy when a quadriceps snip does not provide adequate exposure.

Femoral Peel

Technique

Windsor and Insall described the femoral peel as a useful technique for exposing an ankylosed knee. Initially, the proximal medial tibia is exposed subperiosteally, preserving the origin of the medial collateral ligament on the femur. Then, the distal femur is skeletonized by subperiosteal dissection, and the soft tissue sleeve is preserved medially and laterally.

Results

Although the soft tissue sleeve, including the medial and lateral stabilizing structures, is preserved, joint stability may be compromised by the extensive dissection; in these cases, a constrained knee prosthesis implant may be required.

Quadriceps Tendon Rupture

Technique

Fortunately, the incidence of quadriceps tendon rupture after TKA is very low with approximately 1 case per 1,000 reported, as this problem has proven to be difficult to manage. In distinction to the native knee, where primary repair of the tendon to its attachment on the superior pole of the patella with nonabsorbable sutures passed through vertical tunnels in the patella has proven quite successful, the same results have not been achieved in the setting of TKA. Therefore, the preferred technique includes a repair of the tendon with augmentation of the repair using autologous hamstring tissue or a synthetic mesh. First, the distal end of the tendon is resected back to healthy tissue. Then, two No. 5 nonabsorbable sutures are woven through the avulsed tendon in a Krackow-type technique and passed through vertical tunnels in the patella using a suture passer or straight needle. These sutures are then tied with the knee completely extended. An autologous semitendinosus or gracilis graft, or a synthetic surgical mesh, is then used to augment the repair. Postoperatively, the patient is immobilized in full extension for 6 to 8 weeks; motion is then increased by 30-degree increments at 2-week intervals, with the goal of 90 degrees by 3 months postoperatively. During this time, the leg is protected in a hinged brace. Patients with postoperative partial quadriceps tendon rupture after TKA can be adequately managed by the same nonoperative protocol as described for those patients with complete tears who have undergone surgical reconstruction.

Results

Patients with partial quadriceps tears after TKA treated nonoperatively, as described previously, have generally done well, with all seven patients in one recent large series achieving good results. In distinction, nonoperative treatment for patients with complete quadriceps tears has been poor, with patients generally requiring drop lock knee braces to ambulate. Furthermore, patients with complete tears treated with simple repair alone also had poor results. Six of 10 patients in this group had unsatisfactory outcomes, including four reruptures. Other complications included knee recurvatum and instability, and deep infection. These generally dismal results prompted the changes in surgical techniques noted previously.

Patellar Tendon Disruption

Technique

In association with TKA, postoperative patellar tendon rupture is an uncommon but potentially catastrophic complication that presents a difficult reconstructive problem. Nonoperative management is associated with loss of extensor power in the involved extremity, usually requiring the use of a drop lock brace or walker. Surgical intervention has also been associated with poor results when simple repair has been attempted. More recently, successful salvage of a functional extremity has been made possible with allograft reconstruction using an entire extensor mechanism. The technique is performed through a midline incision when possible. A midline arthrotomy passing over the medial patella allows the native patella to be shelled out from the retinaculum. If the prior patella has fragmented and there are numerous thin fragments of bone embedded in the soft tissue sleeve, these may be left in situ to prevent causing extensive damage to the retinaculum. Distally, over the proximal tibia the native tendon remnants and retinacular tissue should be elevated sub-periosteally, both medially and laterally, for a distance of 7 to 10 cm from the joint line to expose the entire tubercle area. This exposure creates two sleeves of tissue that will be closed over the allograft at the end of the procedure. If revision TKA needs to be simultaneously performed, the procedure

is then performed at this stage. It is important to ensure that component positioning is optimal to minimize stress and shear forces on the graft.

Preparation of the host site begins with creating a trough about 6 cm long and just under 2 cm wide and deep with a burr, saw, or osteotome close to the location of the native tubercle. If possible, a bridge about 1.5 to 2 cm high should be left between the tibial surface proximally and the proximal end of the trough, although in some revision TKA cases this may not be possible when the bone is deficient. The proximal end of the trough should have an oblique cut that creates a small overhang that will lock in the bone block of the graft and resist proximal migration due to the pull from the quadriceps.

While the host site is being prepared, a second team prepares the allograft. The fresh frozen, extensor allograft must include a tibial bone block that is at least 6 to 8cm long and 2 cm wide and deep, the entire patellar tendon, the patella, and at least 5 to 6 cm of the quadriceps tendon. Distally, a block about 6 cm long and 2 cm wide and deep is prepared. The prepared block should be slightly larger than the trough in the host tibia to facilitate a solid press fit. The proximal end of the block should also have an oblique cut that is directed proximal and posteriorly that will help lock in the block, as previously noted (Fig. 27-6). In the proximal part of the allograft, two No. 5 nonabsorbable braided sutures are stitched through the medial and lateral portions of the quadriceps tendon using a Krackow technique. The free ends of the suture should exit from the proximal end of the quadriceps tendon.

Insertion of the graft begins by weaving two additional No. 5 nonabsorbable sutures into the distal end of the native medial and lateral quadriceps tendon remnants using a Krackow technique. Next, the bone block is gently press fit and tamped into the trough. The proximal portion of the bone block should first be wedged under the oblique step in the native bone, and then the distal end gently impacted to ensure the best fit. If the bone block needs to be trimmed at this stage, either a burr or bone rongeur can be used. Once the block is tapped in place, the block is fixed with two bicortical screws that should be countersunk. Certainly, use of screws creates stress risers in the graft and theoretically increases risk of graft resorption or weakening; however, loss of fixation anecdotally appears to be reduced. Alternatively, two or three 18-gauge wires can be used, as is customary with a standard tibial tubercle osteotomy. However, in these allograft cases, the surgeon must ensure that the proximal lock between the block and trough that is produced by the oblique cuts is solid, or proximal migration may occur. Lack of the native lateral soft tissue attachments that help resist proximal migration of the block with a standard tubercle osteotomy, as well as longer time to union, are perhaps reasons why the risk of migration may be higher in these allograft cases.

After the block is secure, the sutures that were previously woven through the allograft quadriceps tendon are pulled proximally by an assistant. While under maximum tension, these sutures are passed under, and then up through the native quadriceps tendon remnants. The native tissues are simultaneously pulled distally using the other Krackow-type sutures that were previously woven through the medial and lateral sleeves (Fig. 27-7). While the knee is held in full extension and maximal proximal and distal forces are applied, the sutures that are attached to the allograft quadriceps tendon are tied. Additional No. 5 sutures are then placed into the native and allograft tendon to stitch the graft in place. The native medial and lateral tissue sleeves are then closed over the allograft in a â€œpants over vestâ€• technique using a combination of 0 and No. 1 sutures.

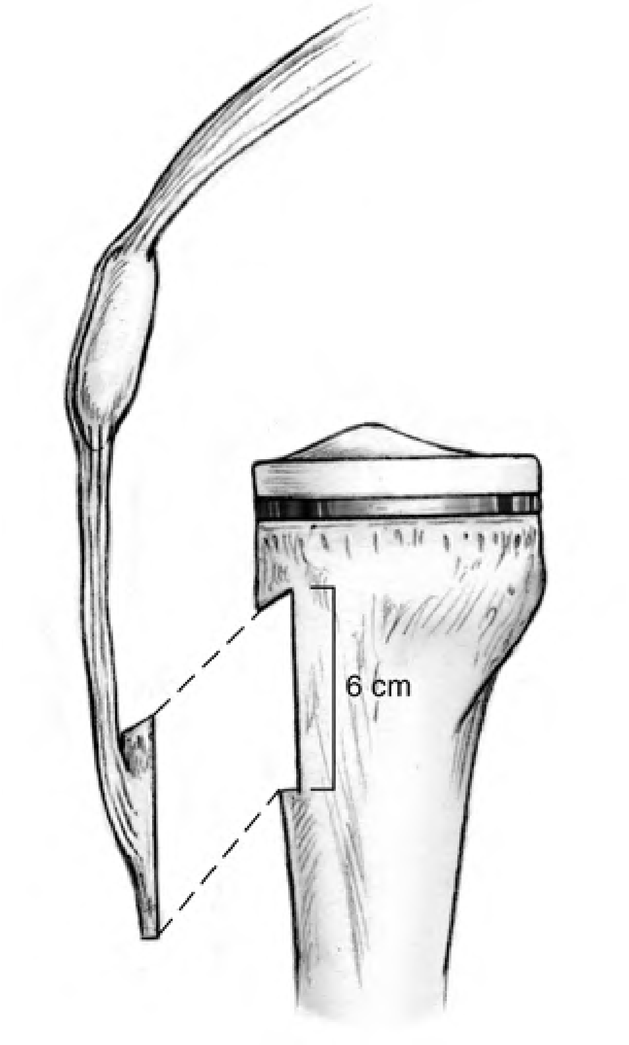

FIGURE 27-6 The trough in the host tibia should be made with a proximal overhang underneath which the oblique cut of the extensor allograft bone block can be wedged.

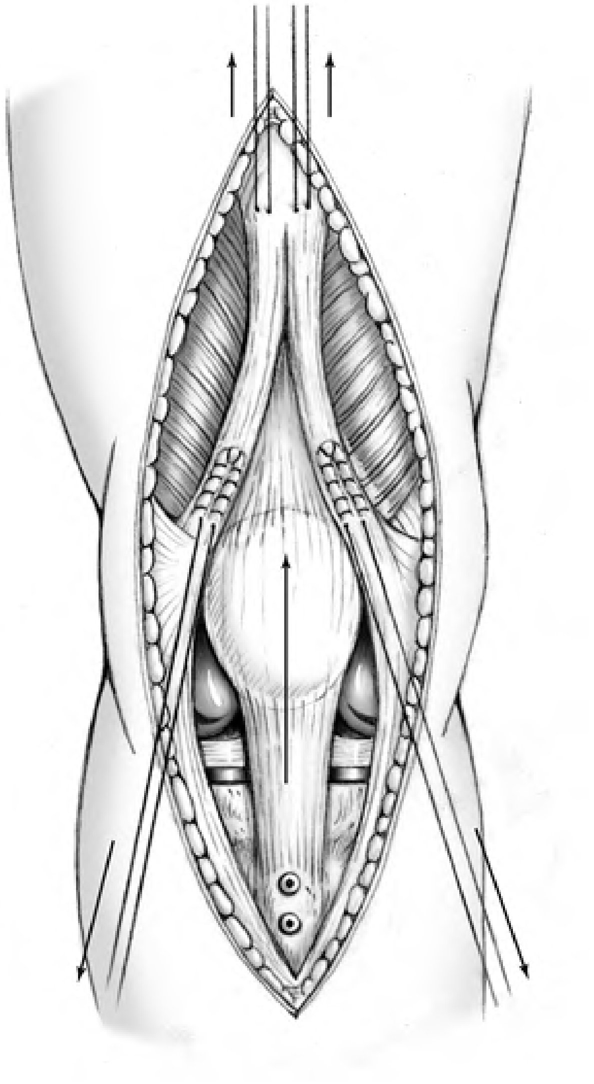

FIGURE 27-7 The extensor allograft should be tensioned and sutures tied with maximal proximal and distal directed forces applied to the allograft and native quadriceps, respectively.

Postoperatively, these patients are immobilized in full extension in a well-fitting brace or cylinder cast for 6 to 8 weeks. Factors including the patient compliance, shape of the leg, and wound care issues must be considered in selecting the optimal method of immobilization. Patients may ambulate with partial weight bearing during this period. After 6 to 8 weeks, flexion is advanced by 30 degrees every 2 weeks, with the goal of 0 to 90 by 3 months postoperatively. Typically, a hinged knee brace with flexion stops is used to protect the patient during this period and the brace is locked in full extension during ambulation.

Results

The results of simple repair of patellar tendon ruptures after TKA have been poor with high failure rates. Furthermore, early attempts at patellar tendon reconstruction after TKA, using extensor allografts with techniques that tensioned the graft in varying degrees of flexion, have also been associated with high rates of extensor lag. Other complications of all techniques include infection and wound-healing problems. However, knee stiffness has not been a significant concern, and changes in the intraoperative technique, described previously, that emphasize tensioning the graft in

maximal flexion, in conjunction with prolonged immobilization in extension postoperatively, have been associated with better results. Burnett et al reported clinical failures, with an average lag of 59 degrees, in all seven patients in whom the graft was minimally tensioned. In the subsequent 13 patients in whom the graft was maximally tensioned, all were successes with a mean of only 4 degrees lag.

Postoperative Management

A number of factors have been reported to influence postoperative wound healing after TKA. Many of these factors include the same systemic conditions that need to be optimized preoperatively. Similarly, postoperatively, it is imperative that significant pulmonary disease and anemia, in addition to tobacco smoking, are aggressively managed to maintain optimal tissue oxygenation. Furthermore, nutritional requirements must be met, and management of diabetes should maintain blood glucose levels within tight control. When possible, immune-modifying medications should be held until primary wound healing has occurred. In addition to these factors, other factors specific to the postoperative period should be considered, especially in patients with other significant risks. Tight dressings should be avoided as these may compromise local capillary flow. Use of a continuous passive motion machine for extended periods should be avoided in any patient at risk for wound-healing problems, as flexion beyond 40 degrees is known to reduce the tissue oxygenation along the lateral wound, and greater than 60 degrees also compromises medial tissue oxygenation. Indeed, immobilization should be instituted in patients that develop problem wounds. Development of a significant postoperative hematoma may also reduce tissue oxygenation; in cases where wound-healing problems occur in conjunction with a hematoma, surgical drainage should be carefully considered.

When wound-healing problems occur, the problem should be carefully followed and aggressively managed to avoid secondary bacterial seeding of the joint. Small areas limited to 1 to 2 mm of marginal superficial skin necrosis that involves short segments of the wound (<1 to 2 cm) may be observed and treated with local wound care in the absence of infection. In these cases, immobilization should be used and activity minimized until the margins of the wound have declared themselves. Wound breakdown or full-thickness necrosis should be debrided early, and one of the previously described soft tissue coverage procedures should be used.

Prolonged wound drainage, without wound breakdown, should also be aggressively managed. Significant serous drainage beyond 3 or 4 days should be managed with a compressive (but not tight) dressing, immobilization, and bed rest. Failure to respond within 48 hours should prompt surgical drainage if a seroma is suspected. In cases where there is significant bloody drainage, initial treatment is the same as for serous drainage. However, in this second group of patients, persistent bloody drainage after TKA is suggestive of capsular dehiscence. Failure to respond to non-surgical treatment should prompt a return to the operating room for an evacuation of the hematoma and closure of the arthrotomy. Certainly, substantial drainage beyond a week from surgery should cause concern and requires careful observation and early intervention if quick resolution doesnâ€™t occur.

Summary

In order for successful revision TKA to be accomplished, early wound healing is required. Numerous factors in the preoperative, intraoperative, and postoperative periods can influence this process. While many of these factors are beyond the control of the orthopaedic surgeon, many are not. Careful preoperative evaluation helps identify and manage potential problems to minimize the risks. Furthermore, the orthopaedic surgeon needs to be aware of the plastic and orthopaedic surgery

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 28

The Pedicled Soleus Muscle Flap for Coverage of the Middle and Distal Third of the Tibia

Salvatore C. Lettieri
Steven L. Moran

Soft tissue defects involving the middle and lower third of the leg may occur following trauma, tumor extirpation, and osteomyelitis. Anteriorly the skin and subcutaneous tissue overriding the middle third and lower third of the tibia are thin and exposed bone and tendon may result from soft tissue injury or open fractures. Historically, middle third defects have been covered with the pedicled soleus muscle flap, while attempts have been made more recently to extend to soleus flap to cover defects of the lower third of the leg.

The soleus muscle, according to the Mathes and Nahai schema, is a type II muscle, containing dominant pedicles from the popliteal, peroneal, and posterior tibial arteries and minor segmental pedicles from the posterior tibial artery. The muscle lies in the superficial posterior compartment extending the entire length of the lower leg. The soleus originates from the posterior surface of the tibia, the interosseous membrane, and the proximal third of the fibula. The muscle runs deep to the gastrocnemius muscle in the upper third of the leg (Fig. 28-1). In the middle third of the leg, the muscle joins with the gastrocnemius muscle and is adherent to the calcaneal tendon. The soleus is a bipennate muscle with the medial and lateral muscle bellies each receiving an independent neurovascular supply; this allows the lateral and medial portions to be mobilized independently while preserving some function within the remaining soleus muscle. The medial head originates from the tibia and receives the majority of its blood supply from the posterior tibial artery. The lateral head originates from the fibula and receives the majority of its blood supply from the perineal artery, although 16% of muscles may be nourished entirely by the posterior tibial artery. The lateral and medial heads are fused proximally while a septum divides the muscle distally. This septum is an extension of the calcaneal tendon and soleus tendon. Dividing the muscle longitudinally at the level of the septum allows for the elevation of the medial and lateral hemi soleus flaps.

In the distal one third of the muscle, the soleus receives segmental arterial perforators from the posterior tibial artery (Fig. 28-2). These distal perforators may be absent in up to 26% of patients; in these cases distal perfusion to the muscle is provided by axial blood flow from more proximal perforators. The diameter and position of these distal perforators is variable but, if present and of large enough caliber, these perforators can allow for a portion of the muscle to be harvested in a reverse fashion. The muscle may then be rotated 90 to 180 degrees, based on its distal perforators, allowing coverage of the lower third of the tibia and ankle region.

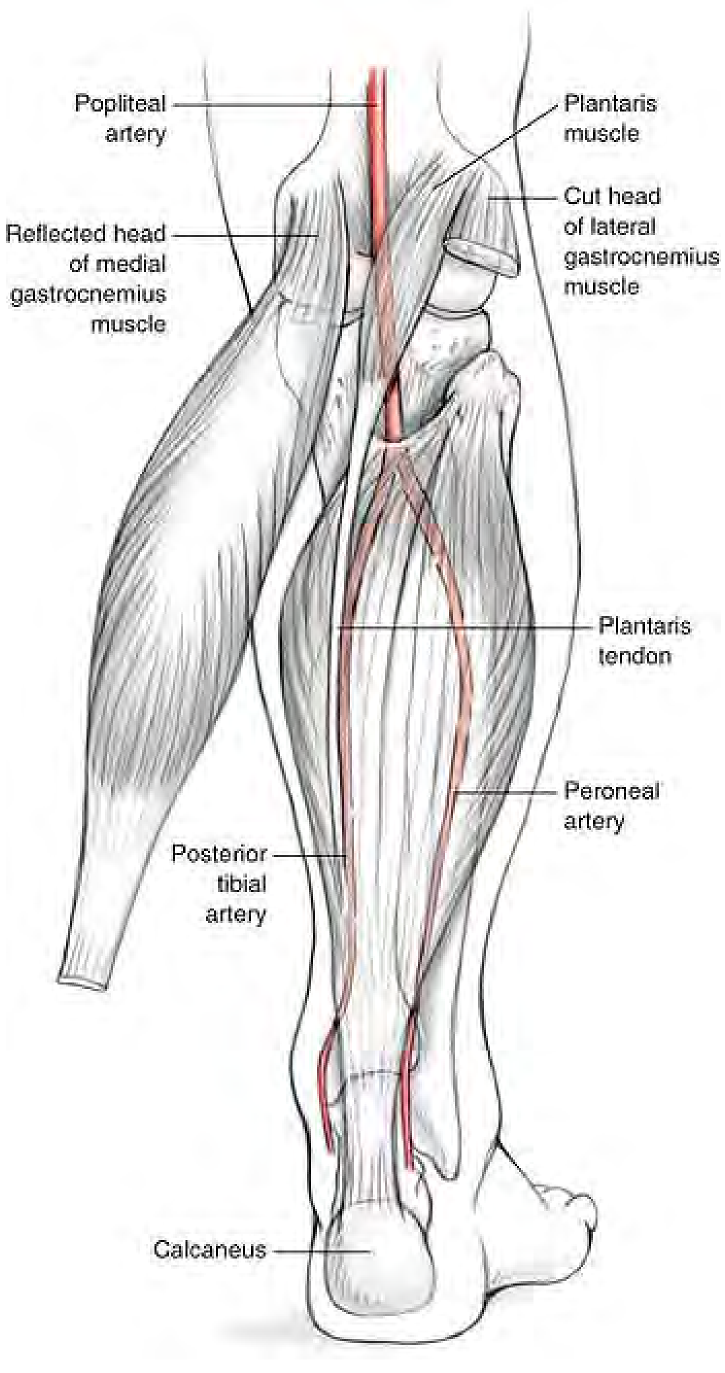
Popliteal artery
Plantaris muscle
Cut head of lateral gastrocnemius muscle
Reflected head of medial gastrocnemius muscle
Plantaris tendon
Peroneal artery
Posterior tibial artery
Calcaneus

FIGURE 28-1 Schematic anatomy of the posterior leg showing the soleus located deep to the gastrocnemius muscle bellies. The plantaris tendon helps define the plane between the soleus and the gastrocnemius muscle bellies. The blood supply to the soleus is from the popliteal, posterior tibial, and perineal arteries.

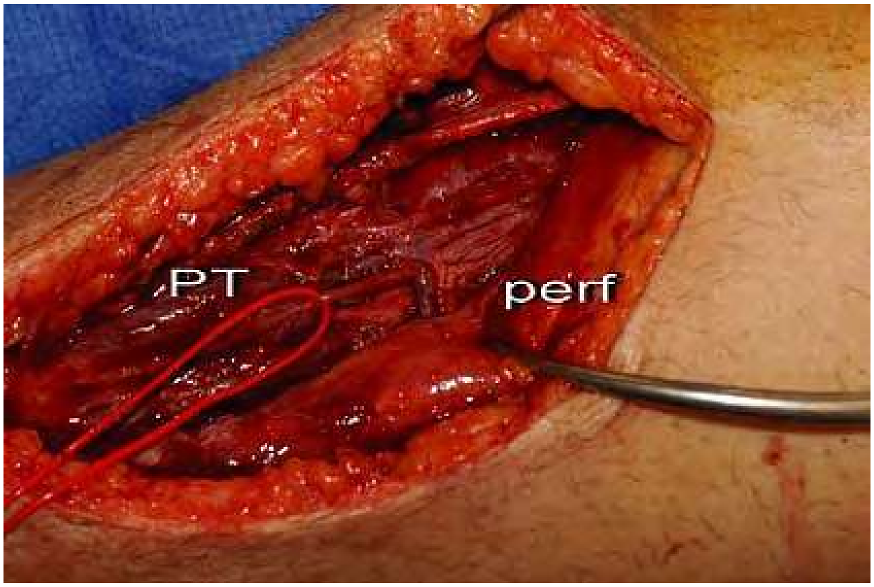

FIGURE 28-2 The posterior tibial artery (PT) as it approaches the ankle gives off several arterial perforators (perfs) to the distal soles muscle. If these perforators are of adequate caliber the muscle can be dissected free and rotated as a distally based flap.

The soleus functions to stabilize the ankle and assist in plantar flexion. The muscle works synergistically with the gastrocnemius and tibialis posterior muscle to provide plantar flexion, while the flexor hallucis longus, flexor digitorum longus, and tibialis posterior all help to provide ankle stability and resist dorsiflexion.

Indications/ Contraindications

The primary indication for the use of the soleus muscle flap is coverage of soft tissue defects in the middle third of the leg. Distal third defects have historically been covered with a free flap; however for small defects and in patients with significant comorbid disease, which would prohibit prolonged anesthesia times, a reverse soleus muscle flap may be considered as an alternative to free tissue transfer.

There are several factors which may prohibit the successful transfer of the soleus muscle and these are:

- Size of the defect
- Status of the muscle
- Status of surrounding tissue and bone
- Size and location of existing perforators

Size of the Defect

The soleus muscle has a limited surface area and a limited arc of rotation (Fig. 28-3). Large defects occupying the majority of the middle third and lower third of the leg are best covered with free tissue transfer. In addition, the distal aspect of the muscle can be unreliable if it must be stretched or inset under significant tension. In such cases, alternative methods of closure should be considered or the soleus may be used in conjunction with another flap.

The soleus can be used in conjunction with the medial or lateral gastrocnemius muscles for larger defects spanning the upper aspect of the lower leg, but this will compromise remaining plantar flexion (Fig. 28-4). When the defect is so large as to require more than just the soleus and medial gastrocnemius flaps, a free flap should strongly be considered as a means of soft tissue coverage and preservation of remaining posterior compartment function. Defects which are to be covered with a reversed soleus muscle flap should be less than 50 cm^2 while the standard soleus flap can cover most defects under 75 cm^2.

Status of the Muscle

Because the soleus muscle is closely adherent to the deep posterior surface of the interosseous membrane, tibia, and fibula, it can often be significantly traumatized following comminuted fractures of the tibia and fibula. During initial wound evaluation and debridement the muscle can often be inspected through the soft tissue defect. If the muscle is significantly lacerated by fracture fragments or contains a significant amount of intramuscular hematoma, it is most prudent to use another flap for soft tissue coverage. In addition, any associated injury to the popliteal, peroneal, or posterior tibial arteries can adversely affect the survival of the soleus muscle.

Status of Surrounding Tissue and Bone

Preexisting damage to the surrounding skin and deeper tissue of the middle third of the leg are also a relative contraindication to soleus muscle flap use. A history of previous radiation therapy, previous surgery, or penetrating trauma to the area surrounding the middle third of the leg should alert the surgeon to potential problems with the use of this flap. The soleus muscle has been shown to provide a source of collateral arterial flow between the posterior tibial and perineal arterial systems. In patients with vascular occlusions of the perineal or posterior tibial arteries, use of the soleus muscle may further compromise limb vascularity.

Size and Location of Existing Perforators

Indications for use of the soleus for distal third defects are based on the work of Pu and include (a) defect size less than 50 cm, (b) defects located over the anterior or medial portion of the distal tibia, (c) the presence of large perforators in distal 4 cm to 5 cm of muscle and (d) a soleus muscle which is nontraumatized on initial exploration (4,5,6). If these factors are met a reversed hemi soleus flap may be attempted for lower third defects, otherwise a free flap is chosen for distal third defects. In addition, patients who have a smoking history, peripheral vascular disease, and history of lower leg radiation therapy should be considered high risk for a reversed soleus muscle flap.

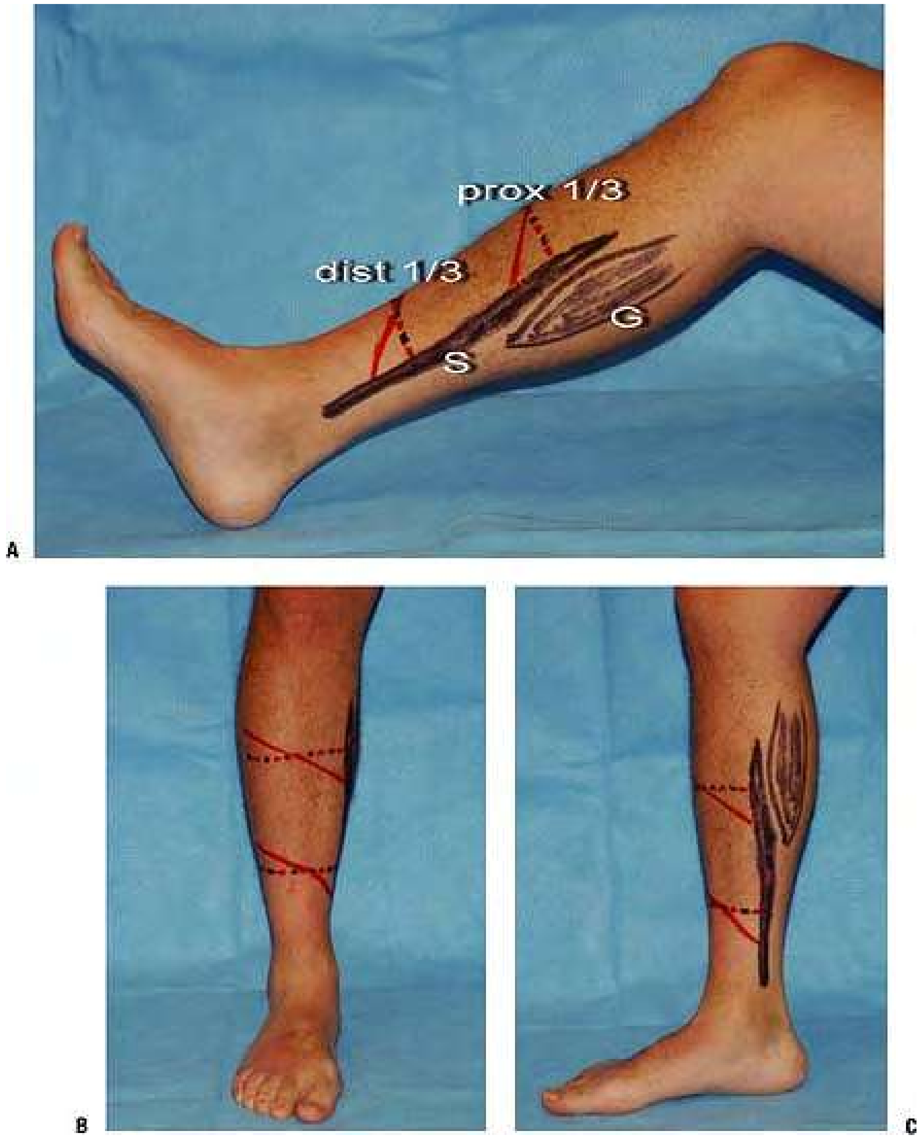

FIGURE 28-3 A-C: The soleus muscle (S) runs anterior to the gastrocnemius muscle (G) and to the area just proximal to the medial malleolus. The arc of rotation of the medially based soleus muscle is limited due to its deep origin and broad proximal attachments. This makes coverage of more laterally and inferiorly based middle third defects difficult. The solid red line shows the actual safe arc of rotation for the medially based soleus muscle flap, with the dashed line depicting the superior and inferior margins of the â€œmiddle thirdâ€• of the leg.

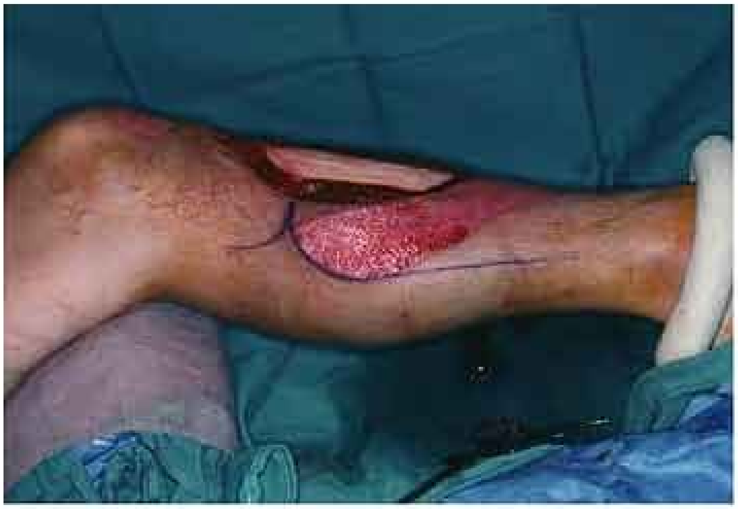
A

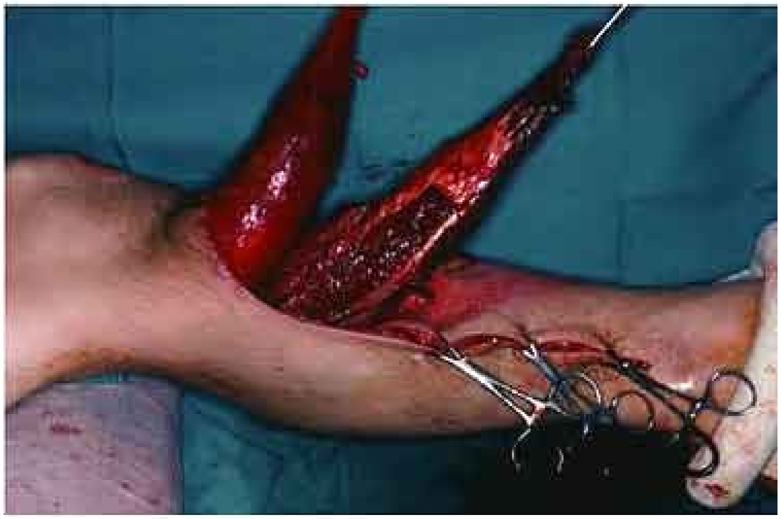
B

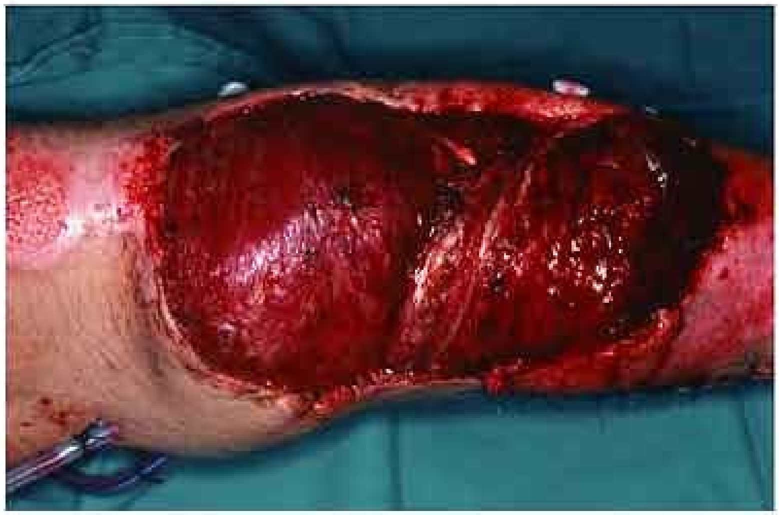
C

FIGURE 28-4 A-C: This large defect is easily covered by elevating both the medial gastrocnemius muscle and the soleus muscle.

Preoperative Planning

Close inspection of the lower leg is necessary prior to surgery. Even though the soleus occupies the same compartment as the gastrocnemius, it can be more significantly damaged in open fractures due to its adherence to the tibia and interosseous membrane. If there is significant ecchymosis or swelling of the posterior compartment, in conjunction with significant displacement of the tibia and fibula on lateral radiographs, one must assume significant damage to the soleus muscle. In this case, free tissue transfer would provide a better alternative for coverage of the middle third of the tibia. If the patient has had previous surgeries, such as fasciotomies for trauma, this may also preclude reliable rotation of the flap, since the level of the fasciotomies and also the location of the injuries may have injured the underlying flap (Fig. 28-5).

A noninvasive vascular exam should be performed on the lower leg, verifying patency of the posterior, peroneal, and anterior tibial arteries. If the patient has a history of significant peripheral vascular disease or long standing diabetes, a CT angiogram may verify patency of the posterior and perineal arteries prior to flap transfer.

Surgery

Patient Positioning and Surgical Preparation

The patient is placed in the supine position for anterior defects, the lateral decubitus position for lateral defects, and prone for posterior defect coverage. As part of the surgical preparation, the entire lower extremity is prepped and draped in the usual fashion. For anterior midtibial defects, a sterile â€œbumpâ€• is placed beneath the distal thigh region to allow for slight external rotation of the leg and bending of the knee; this can facilitate identification of the muscle. The muscle may be harvested with the use of spinal or general anesthetic. A tourniquet on the upper leg allows for a relatively bloodless field during muscle dissection.

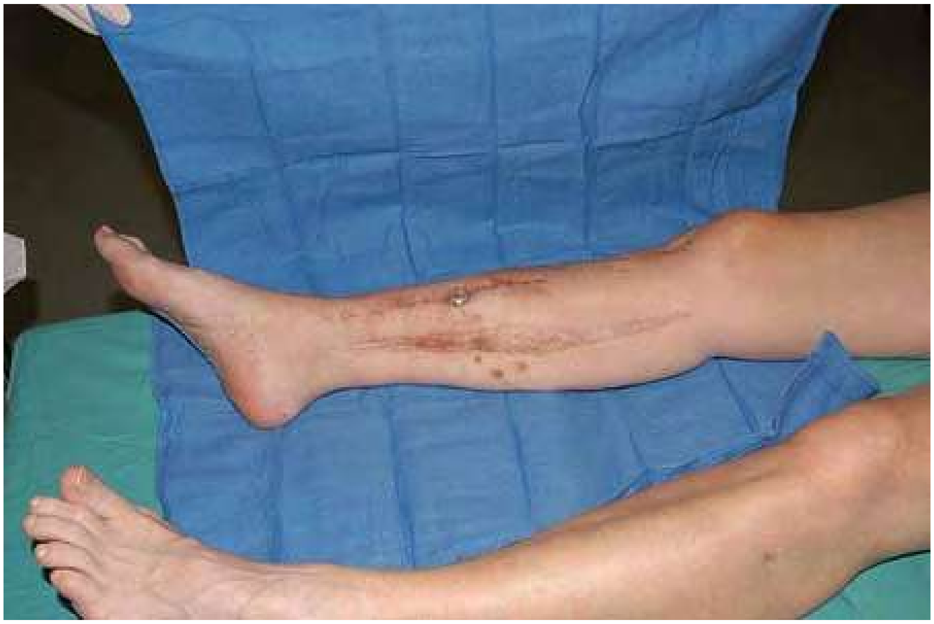

FIGURE 28-5 This patient had lower leg fasciotomies performed at the time of the original tibial injury. The patient now presents with osteomyelitis and exposed hardware over the middle third of the tibia. The defect is located within the arc of coverage of the medial soleus muscle flap; however, there is significant scarring and injury to the superficial posterior compartment secondary to the fasciotomies and previous split thickness skin grafting. Because of concerns of the reliability of the soleus muscle in this situation, this patient underwent coverage of exposed hardware with a free tissue transfer.

Technique

Medial and proximally based soleus flap

The soleus is harvested through an incision which runs from the upper third of the leg to just above the medial malleolus. The incision is made 2 cm to 3 cm posterior to the medial palpable margin of the tibia. The defect site is incorporated into this incision (Fig. 28-6). Skin bridges are avoided. Dissection is carried down to the investing fascia of the superficial posterior compartment. The saphenous vein is identified and preserved during the dissection. The deep fascia is then opened longitudinally to expose the posterior compartment.

The muscle itself is most easily identified proximally deep to the gastrocnemius muscle. The plane between the two muscles is relatively avascular with the exception of some small perforators which may be ligated. The plantaris runs in the plane between the two muscles and can be used as a landmark in cases of severe trauma where hematoma may obscure the tissue planes.

The soleus fuses with the calcaneal tendon in the middle third of the leg. Here, sharp dissection is required to separate the soleus from the common calcaneal tendon as it extends towards the heel.

This dissection is aided by placing medial traction on the edge of the Achilles tendon with the use of several clamps (Fig. 28-7). In the distal aspect, the muscle readily separates from the calcaneal tendon. Dissection from the calcaneal tendon is carried to the midline raphe for hemisoleus elevation. The midline septum does not extend the entire length of the muscle but should be used as guide for hemisoleus elevation. Once the soleus is elevated from the calcaneal tendon, the space between the gastrocnemius muscle and the soleus muscle is easily entered and there is a transverse junction point which is easily cut with scissors or a knife. This will free the entire posterior aspect of the soleus muscle.

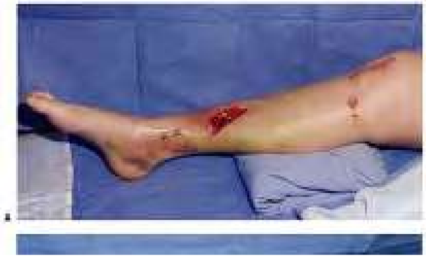

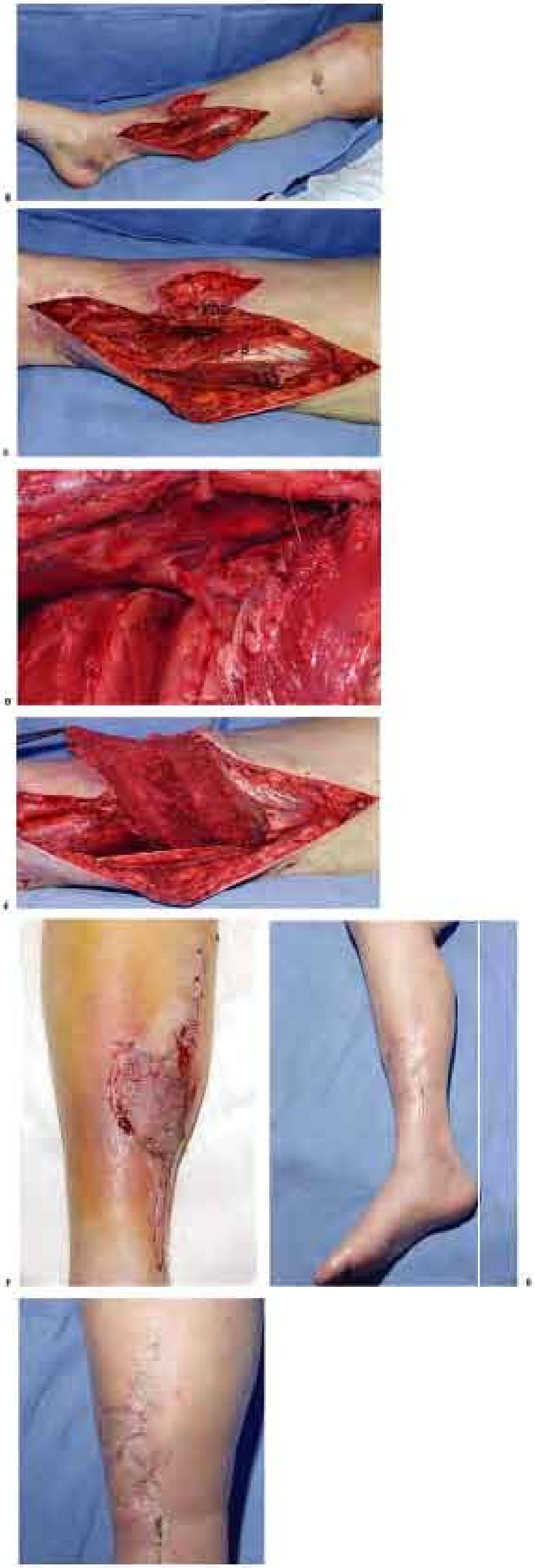
B
C
D
E
F
G

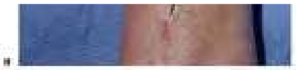

FIGURE 28-6 A: A middle third defect in the right leg of a 66 year old diabetic woman following a Gustilo type IIIB fracture of the tibia. The fracture has been stabilized with an intramedullary rod and the wound has been covered with an antibiotic beads pouch following initial debridement at the time of fracture fixation. B: An incision is made slightly posterior to the medial tibial margin and then carried inferiorly. The open wound is incorporated into the incision to avoid creating a skin bridge. C: The soleus muscle (S) has been dissected free of the calcaneal tendon (C) and separated from the gastrocnemius muscle (G) and flexor digitorum longus muscle (FDL). D: Large perforators from the posterior tibial artery can be seen entering the muscle. E: The muscle is now divided at its distal attachment and pedicled to cover the open defect. Because the soleus is elevated off the common calcaneal tendon the superficial surface of the muscle is devoid of fascia. The muscle is â€œfanned outâ€• and inset with half buried absorbable sutures. F: The donor incision is closed up to the rotated muscle and the exposed muscle is covered with a split thickness skin graft. G,H: At 2 months the patient has a well healed wound with excellent contour.

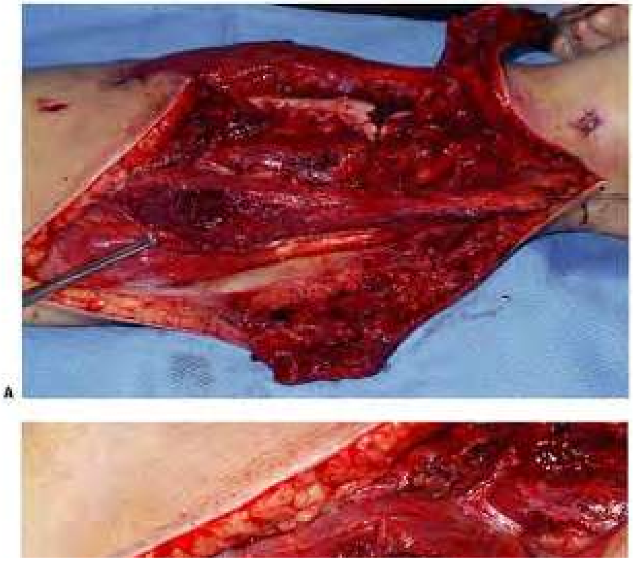

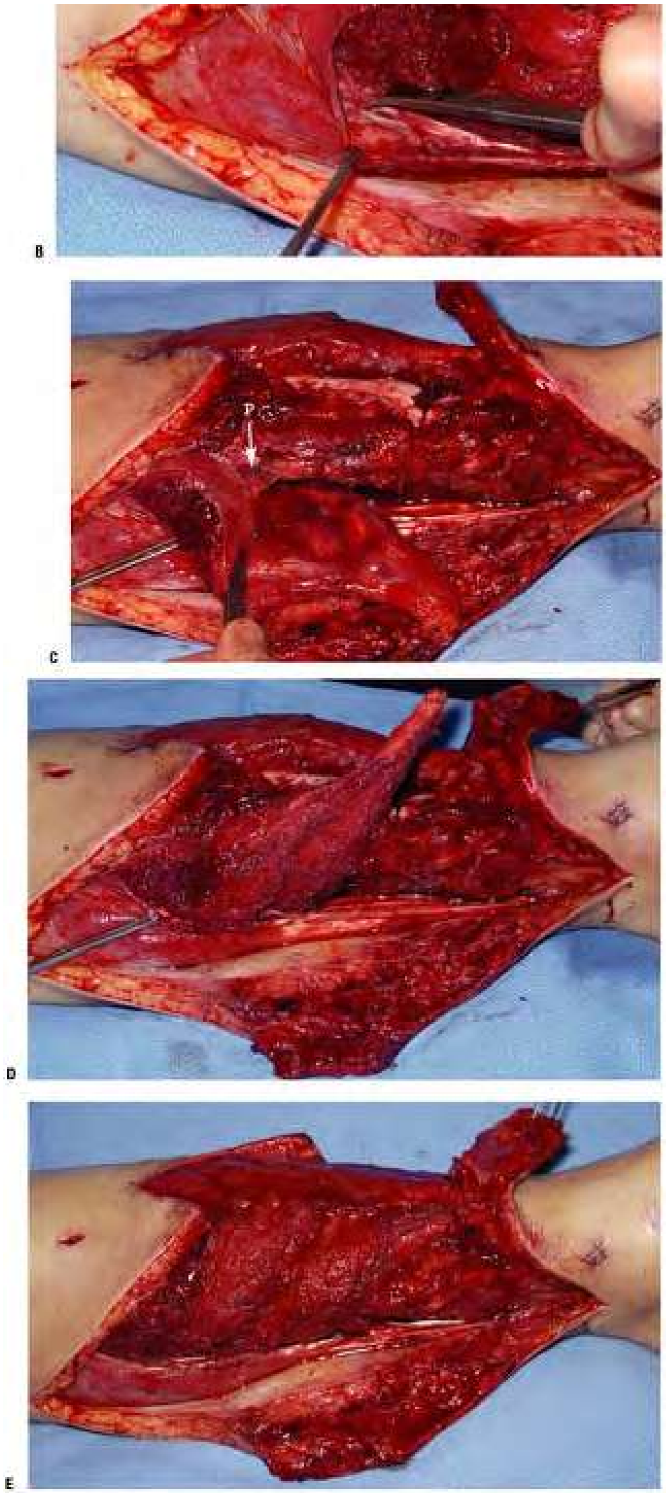

FIGURE 28-7 A: Picture of a middle to lower third Gustilo IIIB injury following a motor vehicle accident in a 58 year old man. The soleus has been separated

from the flexor digitorum longus as well as the gastrocnemius muscle. The muscle has just been divided distally and a stay suture has been placed in the distal aspect of the muscle. B: A Carroll elevator is used to facilitate dissection of the soleus muscle off the common calcaneal tendon as medial traction is applied to the edge of the tendon with Alice clamps. C: The muscle is dissected proximally until enough length is available for defect coverage. Large arterial perforators from the posterior tibial artery are preserved to perfuse the flap (p). D,E: The flap is then pedicled into position and covered with a split thickness skin graft.

The posterior tibial artery is identified in the plane between the soleus and flexor digitorum longus. Inferiorly a finger can be passed beneath the soleus muscle but superficial to the posterior tibial vessels. A cautery is then used to separate these medial attachments of the soleus to the tibia. Once the medial aspect of the muscle is mobilized, the space between the soleus and the deep posterior compartments is readily opened with blunt dissection and the lateral aspect of the muscle is identified.

Once the posterior tibial artery and nerve are identified and the superficial and deep attachments of the muscle have been mobilized, the muscle may be divided distally. Distal perforators from the posterior tibial artery and perineal vessels are then sequentially ligated to allow for mobilization of the muscle. There can be multiple small diameter perforators off the posterior tibial artery; these vessels should be clipped or ligated and not cauterized, as thermal injury may be propagated to the posterior tibial vessels.

Once the muscle has been completely elevated, it is rotated into position to cover the defect. Generally, the fascia on the deep surface is left attached to the muscle, but there will be no fascia on the superficial surface. Multiple half buried absorbable sutures can be used to â€œfan outâ€• the muscle and inset it into the defect site. The distal incision is closed over a drain up to the point of muscle rotation. The muscle should not be tunneled as this can contribute to distal venous congestion within the flap (Figs. 28-6 and 28-7).

The muscle flap is then immediately covered with a meshed skin graft taken from the upper thigh. Vaseline impregnated gauze or a Xeroform dressing are used to cover the skin graft. The leg is then placed in a large Robert-Jones type dressing with the addition of a posterior splint. A window is left in the dressing overlying the muscle so flap checks may be performed while the patient is recovering on the ward.

Lateral and proximally based soleus flap

Though the arc of rotation for this muscle is limited, the lateral approach can be advantageous for some lateral middle third defects. The incision is made just inferior to the lateral border of the fibula. The deep fascia is incised just below the fibula and the plane between the gastrocnemius muscle and soleus are created with blunt dissection. Distally the soleus is again sharply dissected from the calcaneal tendon. Deep proximal dissection involves separating the soleus from its tough attachments to the fibula. During proximal dissection one must be cognizant of the common perineal nerve running close to the fibular head. Once elevated, the muscle is inset as previously described.

Reverse soleus muscle flap

The reverse soleus muscle flap modification (Fig. 28-8) is used to cover small distal defects over the medial anterior aspect of the tibia or medial superior aspect of the medial malleolus (Fig. 28-9). The success of this operation is predicated on the presence of adequate caliber distal perforators from the posterior tibial artery. If during the surgical procedure these perforators are injured or are of insufficient quality the procedure must be abandoned and the defect should be covered with a free tissue transfer. A preoperative angiogram has been recommended by some authors to verify the position of the distal perforator prior to surgery.

Dissection is performed under tourniquet control. The same incision is made as for the proximally based soleus flap (Fig. 28-9). The incision is 2 cm medial to the medial border of the tibia. The existing wound is incorporated into the incision. The medial portion of the soleus is identified as described above and is separated from the gastrocnemius muscle, calcaneal tendon, and the flexor digitorum longus muscle. The posterior tibial artery is identified and the distal perforators are examined. As many distal perforators should be preserved as possible but the authors have had success with the preservation of one or two perforators alone if they are of adequate caliber (vein greater than 1.5 mm and artery of 1 mm or greater). Once the perforatus are determined to be of adequate size, the soleus is then split at the level of the central raphe and divided at the junction of the proximal and middle third. The muscle is then divided longitudinally using the cautery or scissors until the inferior perforator is reached. The muscle is then rotated 90 degrees to 180 degrees. Additional arterial pedicle dissection may be required to prevent vessel kinking. The flap is then inset into the defect with half buried mattress sutures and covered with a split thickness skin graft (Fig. 28-9).

Postoperative Management

The patient should be maintained on bed rest in a posterior splint or knee immobilizer for one week. The bolster dressing may be removed on postoperative day 5 to ensure adequate skin graft â€œtake.â€• Until this time the flap ay be monitored through a window in the postoperative dressing. If the skin graft is adherent to the underlying muscle, the patient may begin to mobilize and bear weight as tolerated, barring any underlying fractures. Gentle range of motion should then begin, increasing the motion by 5 or 10 degrees per day. Ambulation can begin on the 10th post-surgical day. If the flap is inset under tension or if the healing environment is less than ideal, the knee immobilizer can be maintained for an additional week to prevent undue tension on the margins of the muscle flap. The drains are removed once drainage is less than 30 cc a day.

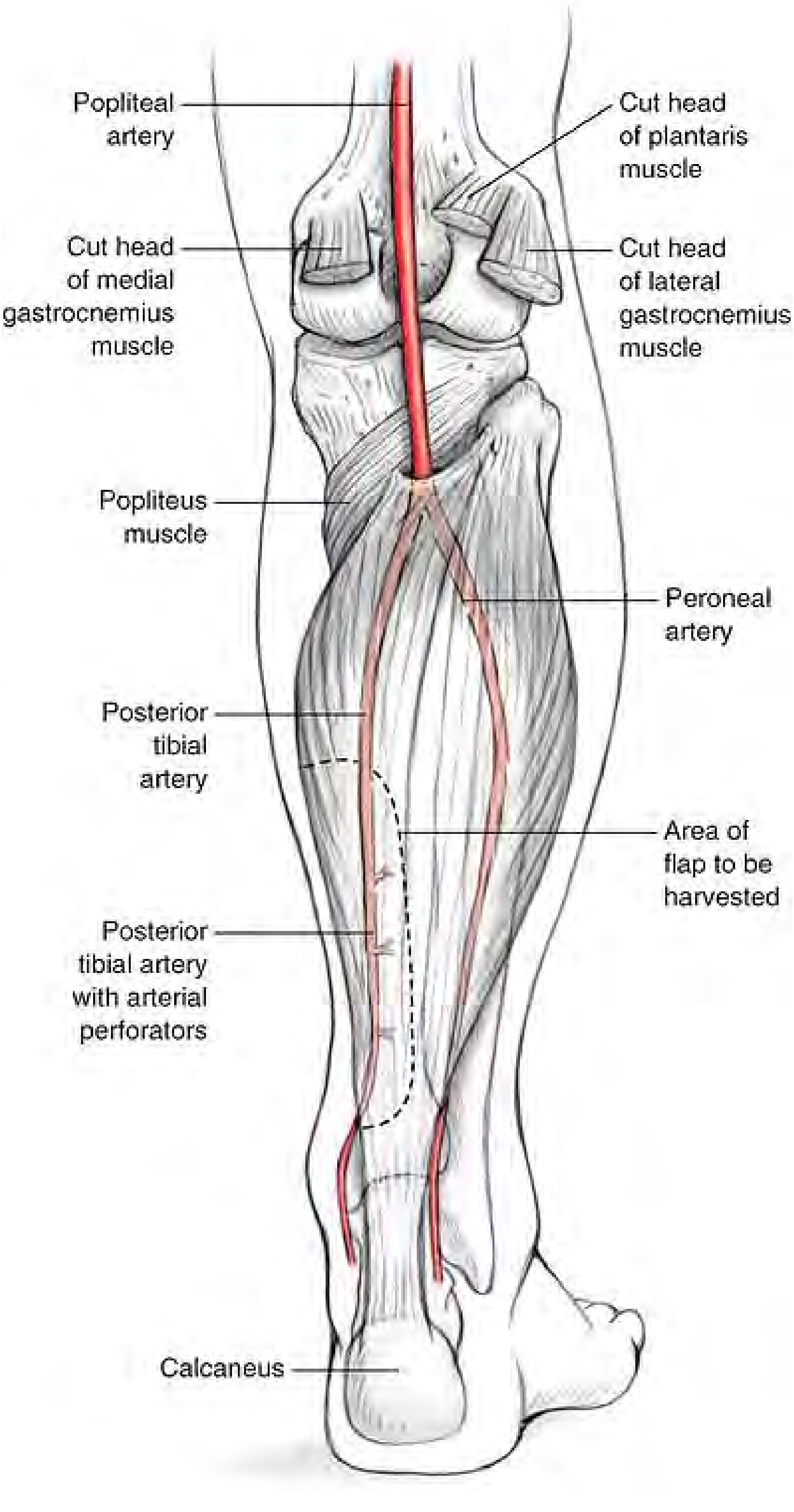
Popliteal artery
Cut head of plantaris muscle
Cut head of medial gastrocnemius muscle
Cut head of lateral gastrocnemius muscle
Popliteus muscle
Peroneal artery
Posterior tibial artery
Area of flap to be harvested
Posterior tibial artery with arterial perforators
Calcaneus

FIGURE 28-8 Schematic drawing of the distally based reverse soleus flap. The

flap's arterial supply is from perforators found at the distal portion of the posterior tibial vessels. If present these vessels may be used to supply a strip of soleus muscle which is cut from the medial margin of the soleus muscle.

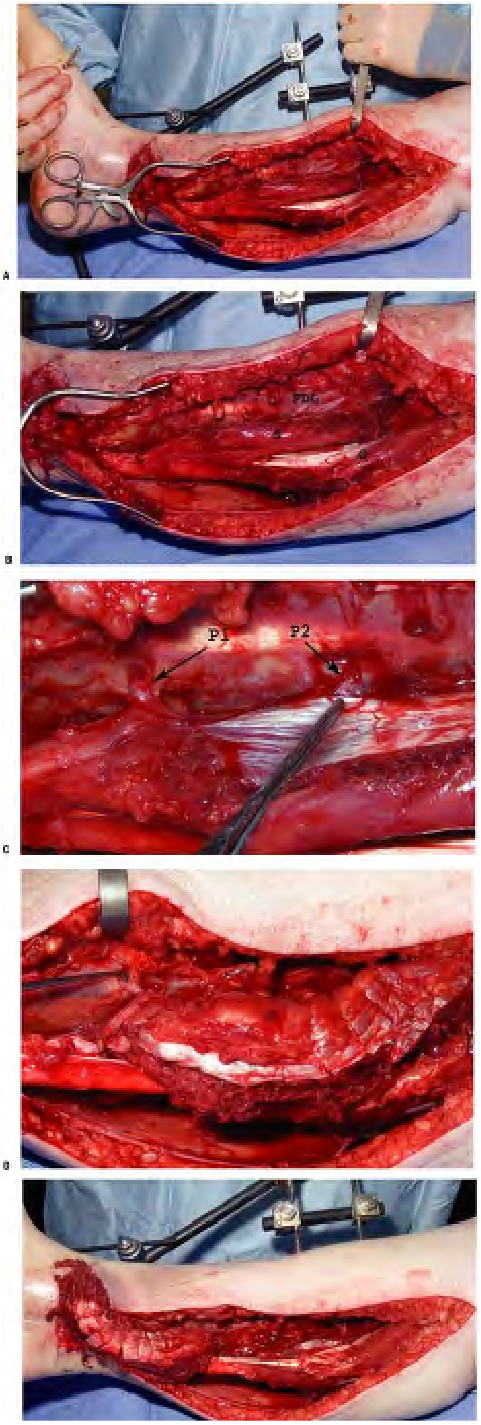

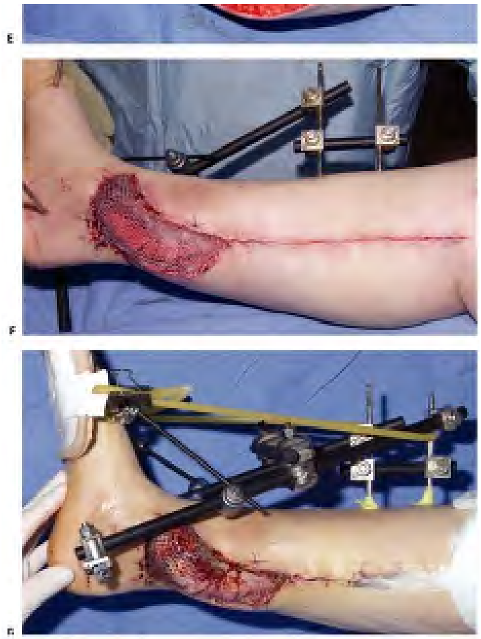

FIGURE 28-9 A: A 61 year old female with a distal defect overlying the superior medial margin of the malleolus. An incision was made at the medial posterior margin of the tibia. B: The gastrocnemius (G), soleus (S) flexor digitorum longus (FDL) and calcaneal tendon (C) are identified. C: Exploration of the posterior tibial vessels, in preparation for free tissue transfer, revealed that the patient had two (p1 and p2) large distal perforators to the soleus muscle. Because of this the patient was felt to be an excellent candidate for a distally based soleus muscle flap. D: A medial strip of soleus is then elevated from the common calcaneal tendon, gastrocnemius muscle and flexor digitorum longus muscle. The nerve hook points to the preserved distal perforator. E: The muscle is then rotated 180 degrees to cover the distal defect. Care is taken to dissect the proximal perforator back to its origin on the tibial artery to prevent kinking of the artery or vein. F,G: The muscle is covered with a split thickness skin graft and the remaining portion of the external fixator is attached to provide needed stabilization.

Once the original dressing is removed from the skin graft, dressings may be changed daily. This should include the application of topical antibiotic ointment and nonadherent gauze followed by a lightly compressive wrap to minimize edema.

Dependent leg position is limited to 30 minutes an hour for the first 2 weeks to minimize edema. On the third postoperative week the patient is fitted for compressive stockings which help the patient continue to manage lower extremity edema and help in the flap contouring.

Rehabilitation

Knee and ankle motion may begin once the skin graft is adherent to the underlying muscle bed. Weight bearing status is determined by the stability of the underlying fractures.

Results

In a study by Hallock of 29 soleus flaps, 24 of 29 flaps were used for coverage of high energy impact defects. All soleus muscle flaps in this study were based on a proximal pedicle. Complication rates were low (13.8%) and there were no cases of total flap loss (1,2). Similar results were reported by Pu, who found no cases of total flap loss when the flap was based on a proximal pedicle for coverage of middle third tibial defects (3,4,7,8).

Complications

Major complications include total and partial flap necrosis. Total flap loss can occur from injury to the vessels at the time of dissection, use of a flap which has been significantly injured with the surrounding bony trauma, or use of the flap in a situation where the posterior tibial vessels and peroneal vessels were compromised or injured. In such cases a free tissue transfer is often needed for leg salvage.

Partial necrosis most often occurs at the distal most margin of the flap. If the partial flap necrosis results in exposure of vital structures, another flap will be required for coverage; most commonly a free flap is used but local fasciocutaneous advancement flaps may be used to cover smaller defects. If the bone is completely covered and the defect is just along the periphery, the wound may be debrided and treated with dressing changes. The remaining wound can then heal through secondary intention.

Skin graft loss may occur due to infection or sheering. Such cases are treated with dressing changes until the underlying muscle bed appears capable of accepting another skin graft. Other minor complications include hematoma in the donor bed and injury to the tibial nerve or the posterior tibial vessels during flap dissection. Both complications may be avoided with meticulous hemostasis and clear identification of the anatomical landmarks prior to muscle division. Limitations in plantar flexion are minimized if the muscle is taken in isolation.

Recommended Reading

Beck JB, Stile F, Lineaweaver W. Reconsidering the soleus muscle flap for coverage of wounds of the distal third of the leg. *Ann Plast Surg*. 2003;50;631â€“635.

Kauffman CA, Lahoda LU, Cederna PS, et al. Use of the soleus muscle flaps for coverage of distal third tibial defects. *J Reconstr Micro.* 2004;20:593â€“597.

References

1. Bos GD, Buehler MJ. Lower Extremity local flaps. *J Am Acad Orthop Surg.* 1994;2:342â€“351
2. Hallock GG. Getting the most from the soleus muscle. *Ann Plast Surg.* 1996;36;139â€“146.
3. Mathes S, Nahai F. Reconstructive Surgery: Principles, anatomy and technique, 1st ed. New York: Churchill Livingstone; 1997:1473â€“1487.
4. Pu LLQ. Soft tissue reconstruction of an open tibial wound in the distal third of the leg. *Ann Plast Surg.* 2007;58:78â€“83.
5. Pu LLQ. The reversed medial hemi soleus muscle flap and its role in reconstruction of an open tibial wound in the lower third of the leg. *Ann Plast Surg.* 2006;56:59â€“64.
6. Pu LLQ. Medial hemi soleus muscle flap: a reliable flap for soft tissue reconstruction of the middle third tibial wound. *Int. Surg.* 2006;91:194â€“200.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 29

The Sural Artery Flap

Michael Sauerbier

Thomas Kremer

Indications/Contraindications

Soft tissue defects of the lower one-third of the leg and the calcaneal region remain crucial issues. This area is easily susceptible to trauma, and defects of this area are commonly experienced in orthopaedic and trauma surgery.

This region is characterized by tightness and limited mobility of the skin and, frequently, by poor circulation. Thus, chronic and diabetic ulcers, pressure sores, unstable scars, or chronic infection following Achilles tendon rupture or trauma surgery often occur in the distal parts of the leg. Furthermore, exposure of viable structures such as neurovascular bundles, tendons, bone, or osteosynthesis material is a common clinical occurrence due to the superficial course of these structures in the lower one-third of the leg.

In these patients, vascularized soft tissue reconstruction is necessary if further soft tissue damage, with subsequent infection and a potential risk of osteitis, is to be prevented. Relatively few procedures for cutaneous coverage have demonstrated real effectiveness and an acceptable morbidity. In recent years, free microvascular tissue transplantation has become a reliable option among patients suffering from acute or chronic defects in the lower one-third of the leg. Free flaps such as the gracilis flap or the latissimus dorsi flap, as well as fasciocutaneous flaps (e.g., anterolateral thigh flap, parascapular flap, lateral arm flap, and free groin flap) are frequently performed for limb salvage. Despite the fact that free microvascular transplantation possesses the disadvantages of extensive surgery, sophisticated equipment, general anesthesia, and high costs, this procedure often remains the only option for reconstruction of large defects in this region for limb salvage (Fig. 29-1).

For the coverage of smaller defects, alternative options such as pedicled flaps (e.g., flexor digitorum communis flap, medial plantar flap, [hemi-] soleus muscle flap, abductor hallucis flap, peroneal brevis flap, and lateral supramalleolar flap) have been developed (Table 29-1). In 1992, Masquelet et al. published their experimental work involving skin island flaps supplied by the vascular axis of sensitive superficial nerves. The distally based sural artery flap, which is perfused by reverse flow through the anastomosis between the superficial sural artery and the lowermost perforator of the peroneal artery, forms part of this group. One of the advantages of this thin fasciocutaneous flap is that it permits skin coverage with ideal contouring. Other advantages are that it can be performed quickly, the soft tissue coverage is durable, postoperative discomfort and donor site morbidity are minimal, and it is unnecessary to sacrifice major vessels. This flap has been used for the successful coverage of defects of the posterior and inferior surface of the heel, the Achilles tendon, the middle and distal one-third of the leg, and the dorsum of the foot and the lateralâ€”as well as medialâ€”malleolus (Tables 29-2 and 29-3; Figs. 29-2,29-3,29-4).

The sural artery flap is insensate and is not an option in cases in which sensitivity is an issue. Furthermore, the sacrifice of the sural nerve results in hyposensitivity of the lateral border of the foot.

Because perfusion of large sural artery flaps (> 9Ã— 12 cm) is unreliable, it is not deemed suitable for coverage of extensive defects of the lower one-third of the leg. Moreover, higher complication rates have to be anticipated in patients with comorbidities such as peripheral artery disease, diabetes mellitus, and venous insufficiency. The sural artery flap is contraindicated in patients with destruction of the vascular pedicle or the lowermost perforator of the peroneal artery (see Table 29-2).

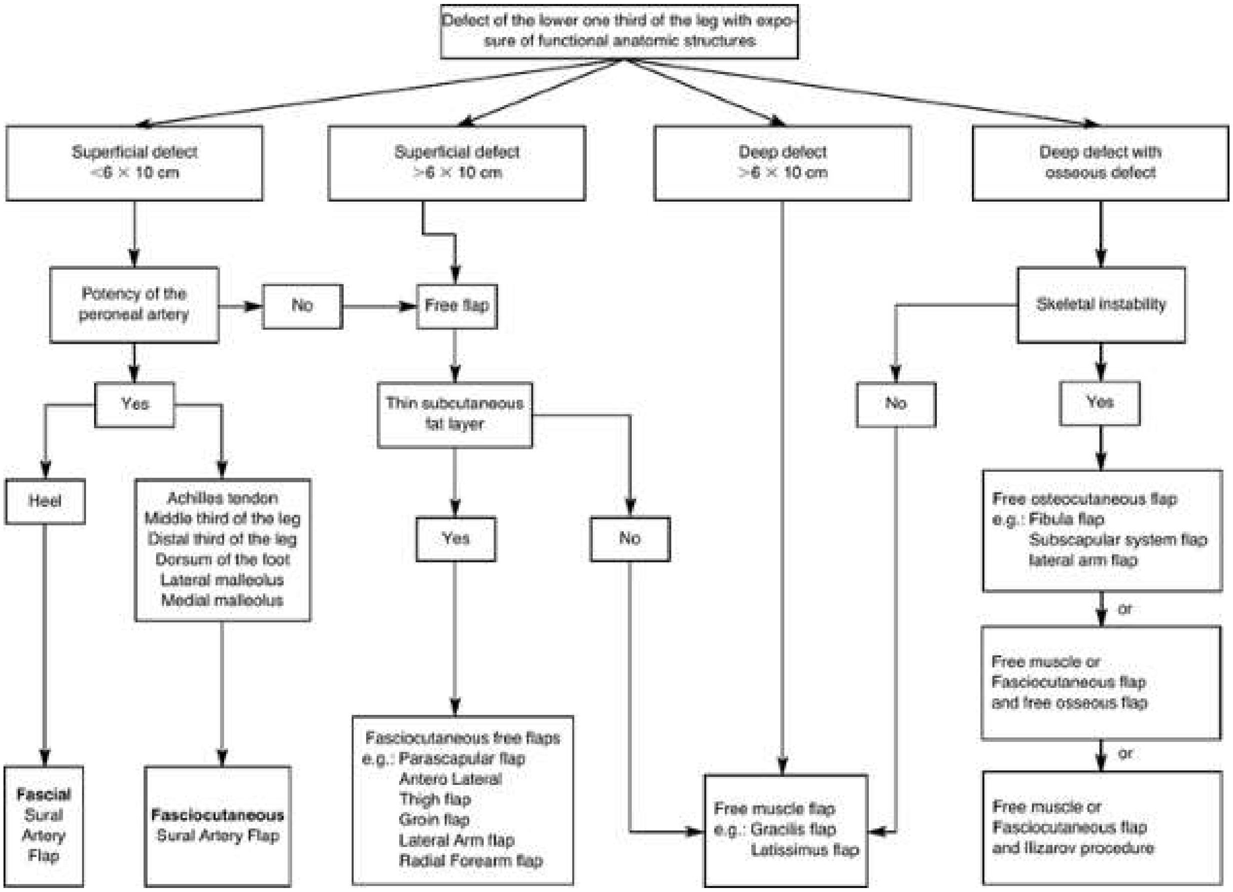
Defect of the lower one third of the leg with exposure of functional anatomic structures
Superficial defect <6 × 10 cm
Superficial defect >6 × 10 cm
Deep defect >6 × 10 cm
Deep defect with osseous defect
Potency of the peroneal artery
No
Free flap
Skeletal instability
Yes
Thin subcutaneous fat layer
No
Yes
Heel
Achilles tendon
Middle third of the leg
Distal third of the leg
Dorsum of the foot
Lateral malleolus
Medial malleolus
Yes
No
Free osteocutaneous flap
e.g.: Fibula flap
Subscapular system flap
lateral arm flap
or
Free muscle or
Fasciocutaneous flap
and free osseous flap
or
Fascial Sural Artery Flap
Fasciocutaneous Sural Artery Flap
Fasciocutaneous free flaps
e.g.: Parascapular flap
Antero Lateral
Thigh flap
Groin flap
Lateral Arm flap
Radial Forearm flap
Free muscle flap
e.g.: Gracilis flap
Latissimus flap
Free muscle or
Fasciocutaneous flap
and Ilizarov procedure

FIGURE 29-1 Treatment algorithm. Small superficial defects of the lower one-third of the leg can be treated by a sural artery flap if the defect size is smaller than 6 Ã— 10 cm. Larger superficial defects can be reconstructed by fasciocutaneous free flaps, whereas deep defects, especially in combination with osseous, defects should be reconstructed by muscle flaps. Defects in the weight bearing area of the heel should be reconstructed by fascial sural artery flaps to prevent pathologic movement of the subcutaneous layer. For alternative pedicled flaps to the lower limb, see Table 29-1.

Table 29-1. Flaps and Their Indications According to Location

Flap	Indications
Soleus flap	Middle third of the leg
Flexor digitorum communis flap	Distal third of the leg
Flexor hallucis longus flap	Distal third of the leg
Peroneus flap	Achilles tendon
Lateral supramalleolar flap	Dorsum of the foot/heel (non-weight bearing)
Dorsalis pedis flap	Foot/distal third of the leg
Extensor digitorum brevis flap	Premalleolar area/dorsum of the foot and toes
Medial plantar flap	Heel (weight bearing)

Table 29-2. Advantages and Disadvantages of the Sural Artery Flap

Advantages	Disadvantages
Constant anatomy	Increased morbidity in patients with comorbidities
Thin flap with ideal contouring	Sacrifice of the sural nerve
Quick procedure	Susceptible to venous congestion
Durable soft tissue coverage	Insensate flap
Low donor site morbidity	Split thickness skin grafts for donor site closure in large flaps (>4 cm)
Acceptable donor site scar in small flaps (<4 cm)	
No sacrifice of major vessels	

Table 29-3. Indications for Sural Artery Flap Reconstruction

- Healthy patients with no comorbidities (diabetes mellitus, peripheral artery disease, venous insufficiency)
- Skin and soft tissue defects up to 9 Ã— 12 cm:
  - Chronic and diabetic ulcers
  - Pressure sores
  - Chronic infection
  - Traumatic defects
  - Unstable scars
  - Exposure of viable structures (neurovascular bundles, tendons, bone, osteosynthesis material)
- One of the following locations:
  - Middle and lower third of the leg
  - Posterior surface of the heel
  - Inferior surface of the heel
  - Achilles tendon
  - Dorsum of the foot
  - Lateral malleolus
  - Medial malleolus

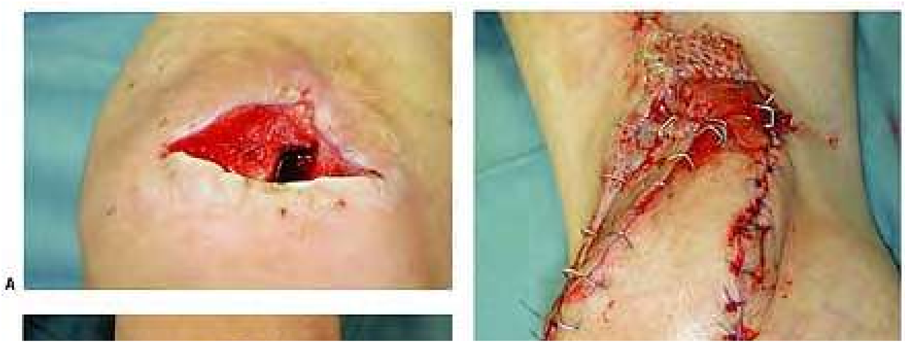

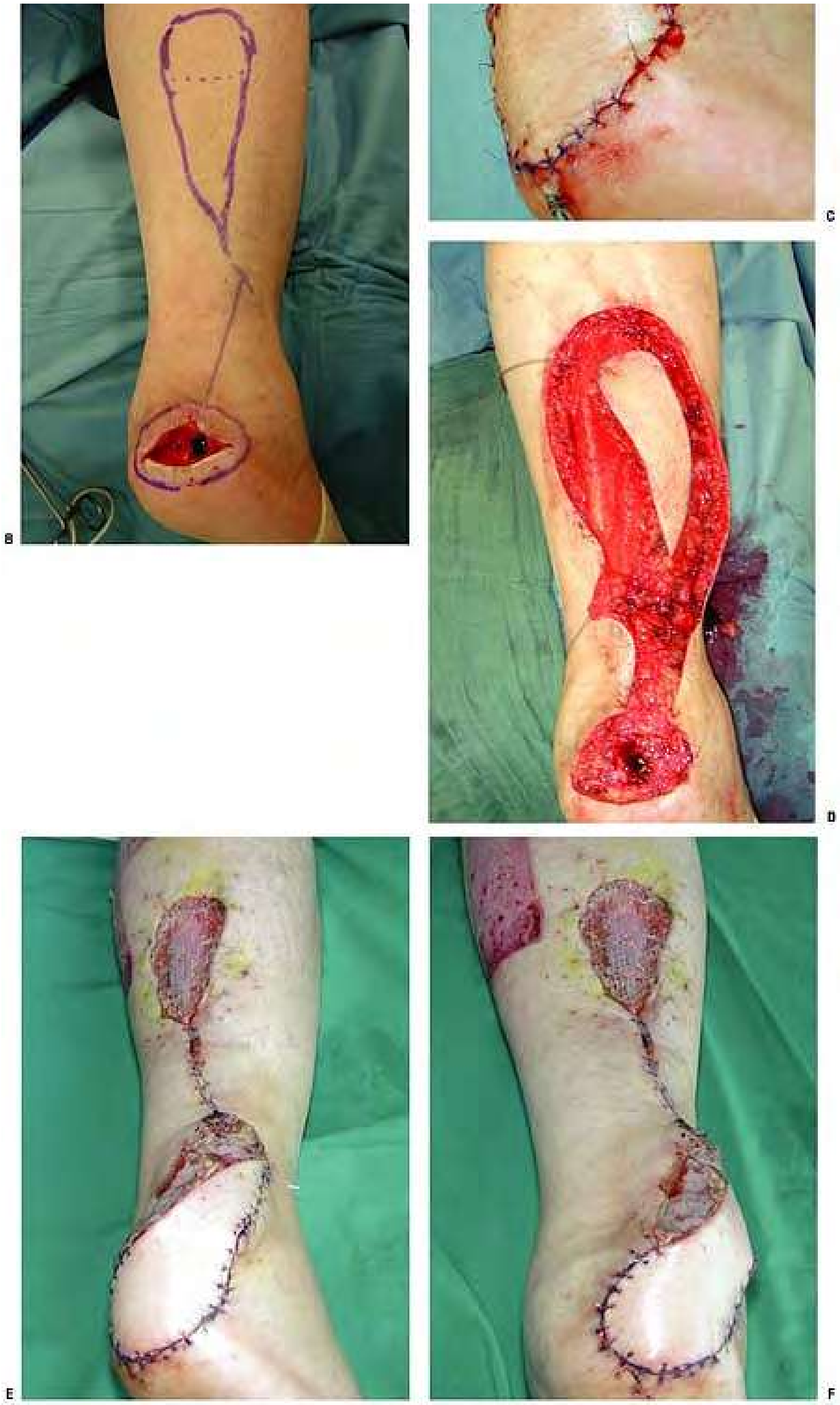

FIGURE 29-2 Male patient suffering from a combined soft tissue and osseous

defect in the heel (A). After debridement, soft tissue reconstruction was planned by a sural artery flap (B). Flap transposition was performed to the defect (C) after flap harvest (D). Split thickness skin grafting of the donor site was necessary. Postoperative course and wound healing were uneventful (E,F).

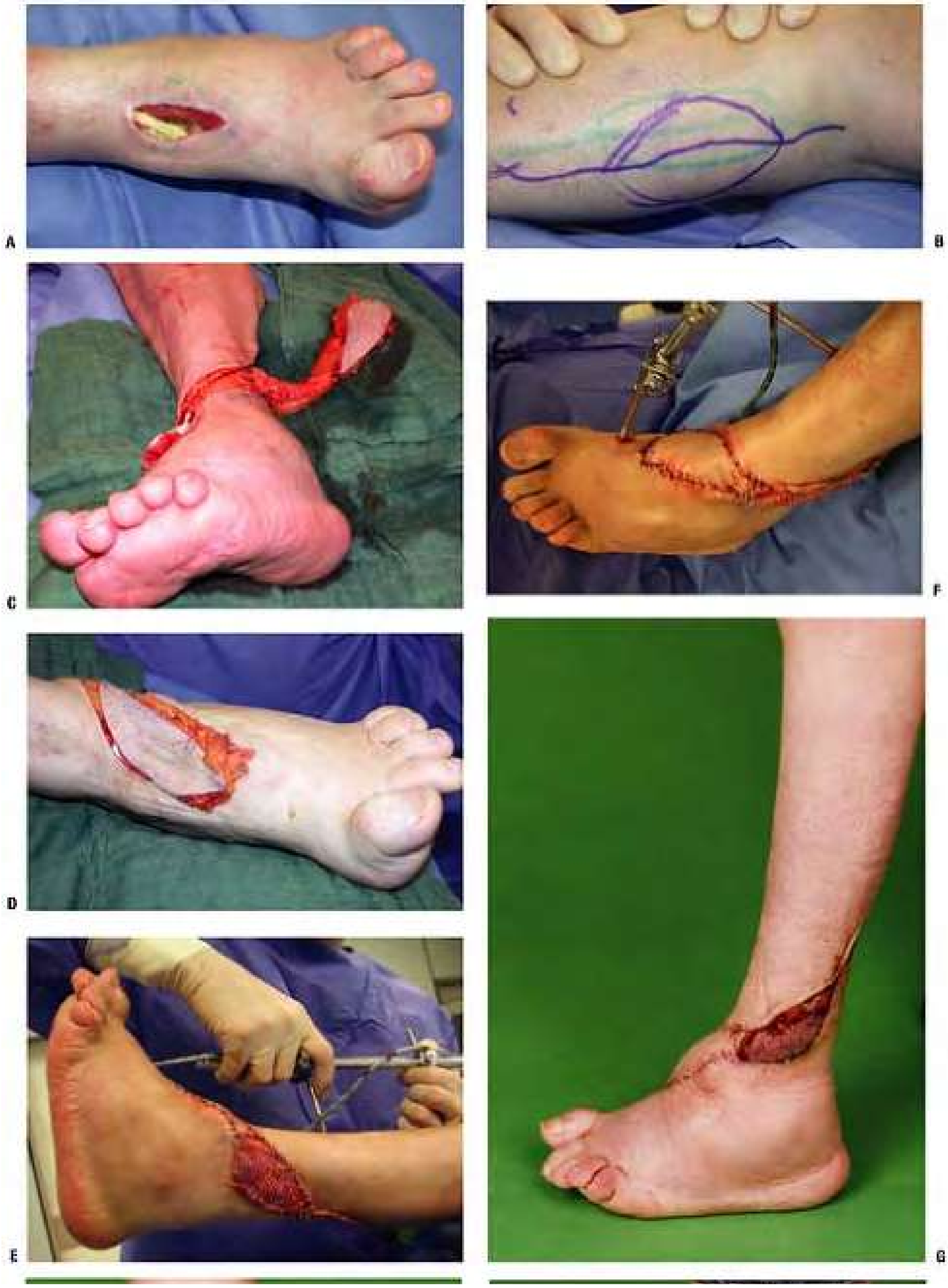

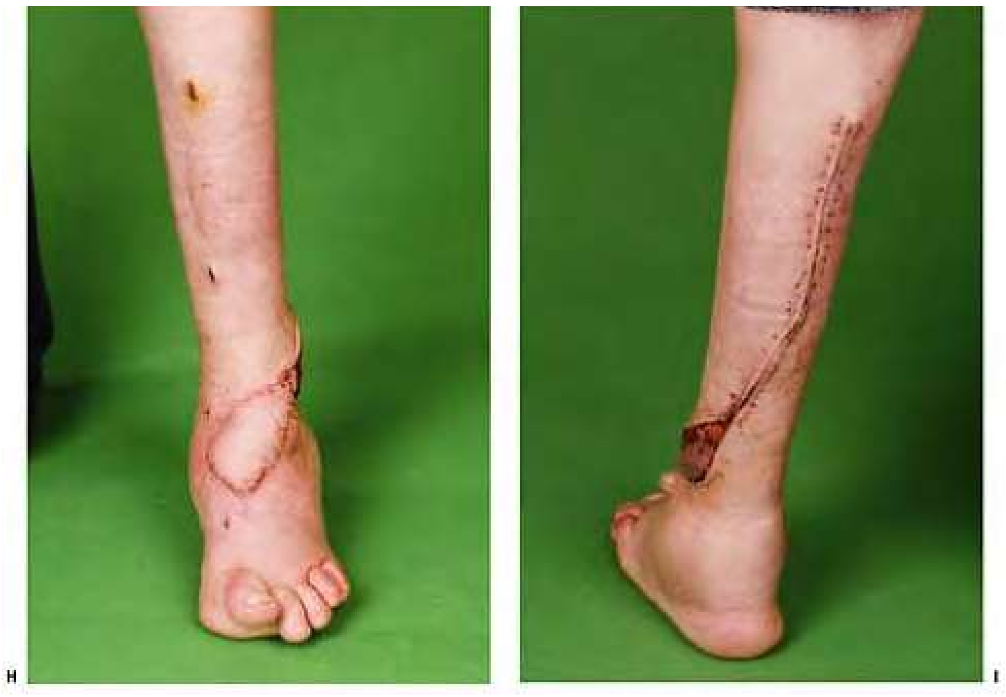

FIGURE 29-3 A 34-year-old man with a chronic ulcer of the dorsum of the foot. A sural artery flap was planned to reconstruct the defect with exposed tendons (A,B). After complete flap harvest, the skin island is transposed to the defect (C,D). The skin over the pedicle is incised and split thickness skin grafted to reduce pressure on the nutrient vessels (E,F). The functional and aesthetic results are excellent 2 weeks post surgery (G) The functional and aesthetic results are excellent 2 weeks post surgery (H,I)..

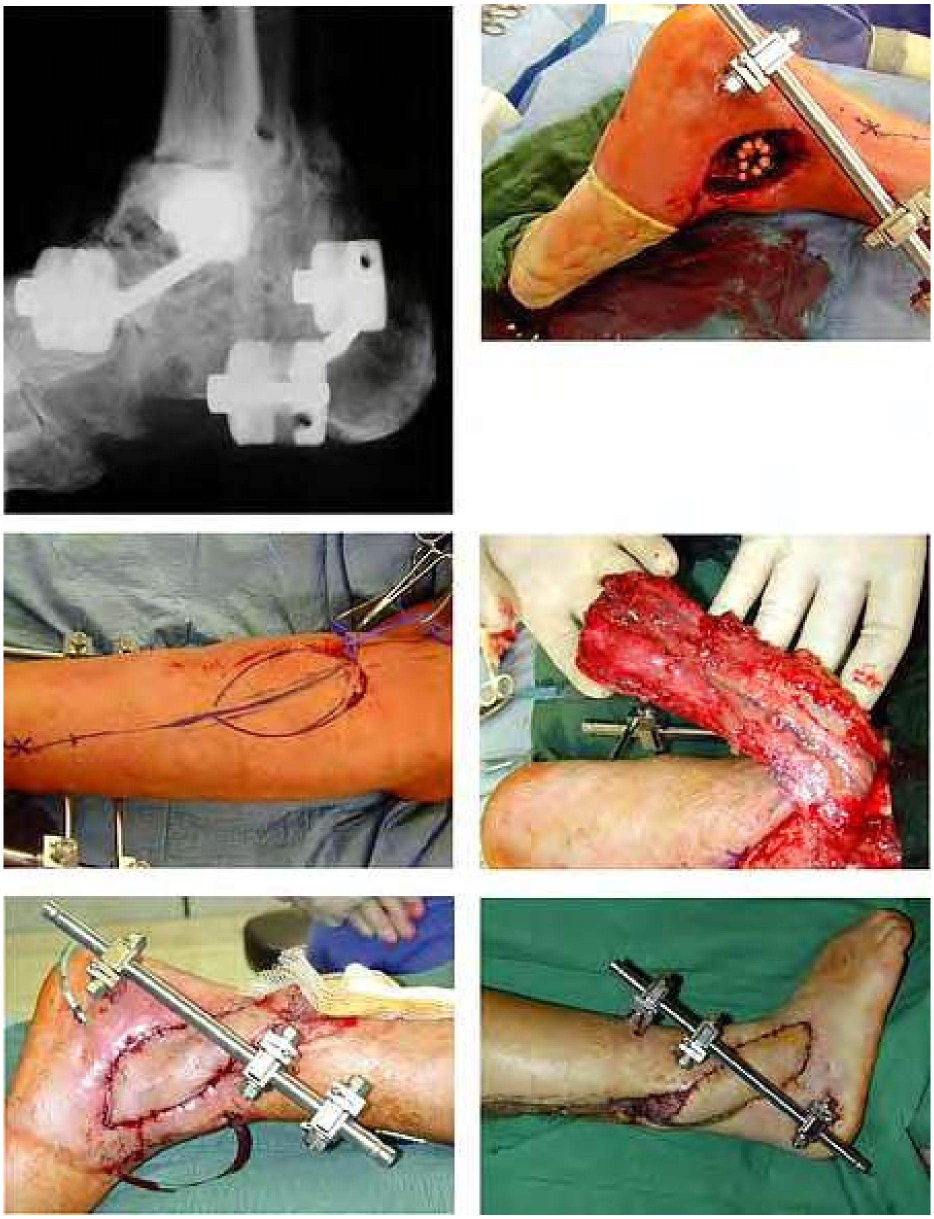

FIGURE 29-4 A 36-year-old woman with osteitis (above left). After debridement and application of an antibiotic chain (above right). Defect reconstruction was planned by a sural artery flap (upper left). After complete flap harvest, the pedicle is visible through the cural fascia, which has to be included in the flap (upper right). After transposition to the defect, wound healing was uneventful (below).

Preoperative Planning

A thorough clinical and radiologic examination of the donor and the recipient sites is mandatory, and it is first necessary to estimate the size of the recipient site defect. There are no studies regarding maximum flap dimensions with regard to safety, but small and moderate-sized defects can usually be covered satisfactorily. Defects up to 9 Ã— 12 cm have been successfully reconstructed using sural artery flaps. Radical debridement may result in an enlargement of the defect, and this should be taken into account in the preoperative planning.

An examination of the patient's peripheral function should be performed to determine the total extent of the injury. Depending on the size and location of the defect, various functional structures may be exposed or destroyed. Changes in the normal resting position of the foot and toes will help identify the loss of continuity of lower extremity tendons, and traditional functional tests will confirm the loss of active range of motion. A careful examination of sensitivity and muscle function helps to evaluate the extent of the defect in terms of destruction of sensor or motor nerve components. Additionally, peripheral perfusion should be ascertained by simple palpation of pulses or by Doppler probe. If in doubt, an angiogram should be performed. If any of these structures have been destroyed, additional reconstruction procedures must be planned. In patients suffering from chronic or infected wounds or diabetic gangrene, additional osteitis or osteomyelitis must be considered and conventional radiographs and/or magnetic resonance imaging (MRI) scans can help to prove the existence or extent of altered osseous structures.

Table 29-4. Preoperative Planning for Sural Artery Flap Reconstruction

Clinical examination
  Defect size (after debridement)
  Peripheral function (to determine destruction of functional structures [nerves, tendons, blood vessels])
MRI or conventional radiographs of the recipient site (fractures, extension of osseous or soft tissue defects, osteitis)
Examination of patency of the lesser sapheneous vein and peroneal artery
  Doppler ultrasound probe in healthy patients
  Conventional angiogram or MRI-angiogramm in comorbid patients
Detailed information of the patient
  Sacrifice of the sural nerve (hyposensitivity of the lateral border of the foot)
  Neuroma formation
  Split thickness skin grafting of the donor site

The sural artery flap pedicle should be examined preoperatively. In all patients the patency of the peroneal artery and the lesser saphenous vein is examined by palpation of pulses and Doppler probe. Additionally, the lowermost perforators between the peroneal artery and the flap pedicle can be examined 4 to 7 cm proximal to the tip of the lateral malleolus. In patients with comorbid conditions such as diabetes mellitus and peripheral vessel disease, flap perfusion may be particularly unreliable. In these patients, a conventional angiogram or an MRI-angiogram can be helpful in preoperative planning procedures.

Finally, the surgeon should provide the patient with a detailed account of the surgical procedure. Information regarding the postoperative function, the possibility of persisting complaints, and donor site morbidity should be given. In particular, the loss of sensitivity of the lateral border of the foot and the possibility of neuroma formation as well as the potential necessity of donor site closure by split thickness skin grafting should be discussed (Table 29-4).

Surgery

Anatomy

The reverse flow sural artery flap is based on the vascular network along the sural nerve. The nutritient artery is the superficial sural artery (SSA), which arises from the popliteal artery proximal to the medial and lateral sural arteries in about 65% of cases, directly from the medial (20%) or lateral (8%) sural artery, or sometimes from the common stem of origin of the two. The artery courses posteriorly for 2 to 3 cm before joining the medial sural nerve descending between the two heads of the gastrocnemius muscle. The medial sural nerve anastomoses with the communicating branch of the lateral sural nerve to become the sural nerve. Both the medial sural nerve and the SSA pierce the crural fascia at the junction of the proximal and middle thirds of the leg to become subcutaneous. The SSA courses alongside the sural nerve to the lateral malleolus in 65% of all patients; in 35% of cases the artery fades into a vascular net at the distal one-third of the leg. Regardless of the termination, the SSA has a constant distal anastomosis from the lateral malleolar arteries, which in turn arises from the peroneal artery. The distally based sural artery flap is dependent on this anastomosis, and its perfusion is reverse flow through the SSA and its cutaneous branches at the distal two-thirds of the leg.

The venous drainage of the sural artery flap is dependent on small commitant veins and mainly on the short saphenous vein which accompanies the sural nerve at the distal two-thirds of the leg. Theoretically, a reverse flow in this vein is impossible due to of the presence of valves; however, clinical findings disprove this theory. Duplex scans of the small saphenous vein generally show a continuous or phasic reverse flow postoperatively. Several theories exist regarding these findings. Some authors stress the existence of bridges between the short saphenous vein and the commitant veins, thus bypassing the valves. Moreover, denervation of the short saphenous vein due to the surgical procedure and increased venous pressure due to the altered flow are thought to be responsible for venous dilatation, which renders the valves insufficient.

Overall, the sural artery flap exhibits a constant anatomy with reliable arterial perfusion and venous drainage.

Surgical Procedure

Surgery is performed in general anesthesia; however, in elderly patients or because of comorbidities, epidural or spinal anesthesia is also possible. Depending on the defect location, the patient is placed in prone position or in a ventral or lateral decubitus position.

The recipient site is first prepared. In patients with chronic wounds, infection, diabetic gangrene, or wound colonization, a radical debridement is mandatory. Analogous to the principles of oncologic resections, a complete resection of the entire infected or altered tissue must be performed. An insufficiently radical debridement renders the reconstructive attempts unsuccessful; therefore, compromises are unacceptable. If, after debridement, the recipient defect is too large for a sural artery flap reconstruction, a temporary wound closure (e.g., vacuum-assisted closure [VAC] therapy) should be performed and other therapeutic options such as free flaps should be planned. For this reason, it can be useful to include the possibility of free flap reconstruction in the initial preoperative patient briefing so that a one-step reconstruction can be carried out.

The authors prefer to perform flap harvest under pneumatic tourniquet (500 mm Hg) to make visualization of the anatomy easier. In addition, the surgeon should use 2.5 to 4 times magnification loupes.

As mentioned previously, the short saphenous vein and the arterial perforators of the peroneal artery on the lateral aspect of the distal leg have to be located using a Doppler ultrasound probe. Following debridement, the defect size is measured and a skin island with adequate dimensions is planned with its center along the line of the short saphenous vein. The proximal limit of the flap should not exceed a boundary of 20 cm proximal to the lateral malleolus, due to the fact that the portion of the flap that exceeds the suprafascial (subcutaneous) portion of the nerve behaves like a random extension, and thus its reliability is unsure. The skin island is planned in the middle or distal third of the leg, according to the pedicle length required (Fig. 29-5). Because of its thicker subcutaneous tissue, the skin island should be designed to be slightly larger than the recipient defect to facilitate skin-to-skin closure of the defect. The maximum size of the skin island should not exceed 9 Ã— 12 cm. The pivot point is marked 5 cm above the tip of the lateral malleolus to include the branches that anastomose with the peroneal artery. The skin incision starts distally along the line of the pedicle to localize the vein. If the skin island is not centered over the short saphenous vein, correction of the flap position is still possible. When the pedicle has been located, the dissection proceeds in the proximal extreme of the flap; again, the short saphenous vein and the sural nerve are identified suprafascially. At this point, the short saphenous vein, the suralis nerve, and the SSA are ligated, divided, and included into the flap. The proximal stump of the nerve is coagulated and buried in the surrounding gastrocnemius musculature to avoid a troublesome neuroma over the donor area. Some authors report raising the flap and sparing the sural nerve to prevent neuroma formation and foot hyposensitivity. In our opinion this strategy is too risky for the vascular supply of the flap and is unnecessary because the morbidity of harvesting the flap with the sural nerve is minimal. In the literature, several series show no significant alteration in foot sensitivity, and in most cases, patients exhibit improved sensitivity over time.

Dissection proceeds distally at the plane below the deep fascia, so that inclusion of the pedicle in the flap is easily accomplished and the nutrient anastomosis of the flap can be preserved. Two to three constant and direct perforators from the gastrocnemius muscle to the vascular axis of the sural nerve are identified and ligated with ligature clips, because electrocoagulation of these perforators increases the risk of pedicle destruction. The fascial and subcutaneous pedicle, which includes the SSA, the sural nerve, and the short saphenous vein, should be at least 3 cm in width. The paratenon of the sural triceps muscle should be carefully preserved, especially when split thickness skin grafting of the donor site defect is necessary. The flap procedure remains simple if care is taken to avoid dissection too close to the pedicle and if dissection is only extended to the demarcated limit of 5 cm above the tip of the lateral malleolus. During elevation of the flap, near this inferior limit of dissection, efforts should be made to preserve the large perforating septocutaneous vessels.

When flap harvest is completed, transposition of the skin island to the recipient defect can be performed. Under normal conditions (e.g., a sufficient amount of elastic skin), the flap can be transposed through a subcutaneous tunnel. However, when there is any sign of tight skin or venous congestion, a decompression of the pedicle must be performed. The risk of pedicle compression increases with edema formation, surrounding scars, or skin induration and/or thick and inelastic skin. In patients with this risk profile, two approaches should be considered. First, skin extensions (the â€œcutaneous tailâ€•) overlying the pedicle can facilitate defect closure and reduction of pressure on the pedicle after transposition. Second, the bridge over the subcutaneous tunnel can be incised and a split thickness skin graft can be used to cover the pedicle. In the literature, the possibility of using the â€œdistally based sural artery interpolation flapâ€• is described. This involves transposition of the flap to the defect without tunneling, and the pedicle is exteriorized and skin grafted. In a second procedure, after randomization, the pedicle is removed. The authors have no experience using this approach, but there appears to be no distinct advantage of this approach over other options. Indeed, it has the disadvantage of a two-step procedure.

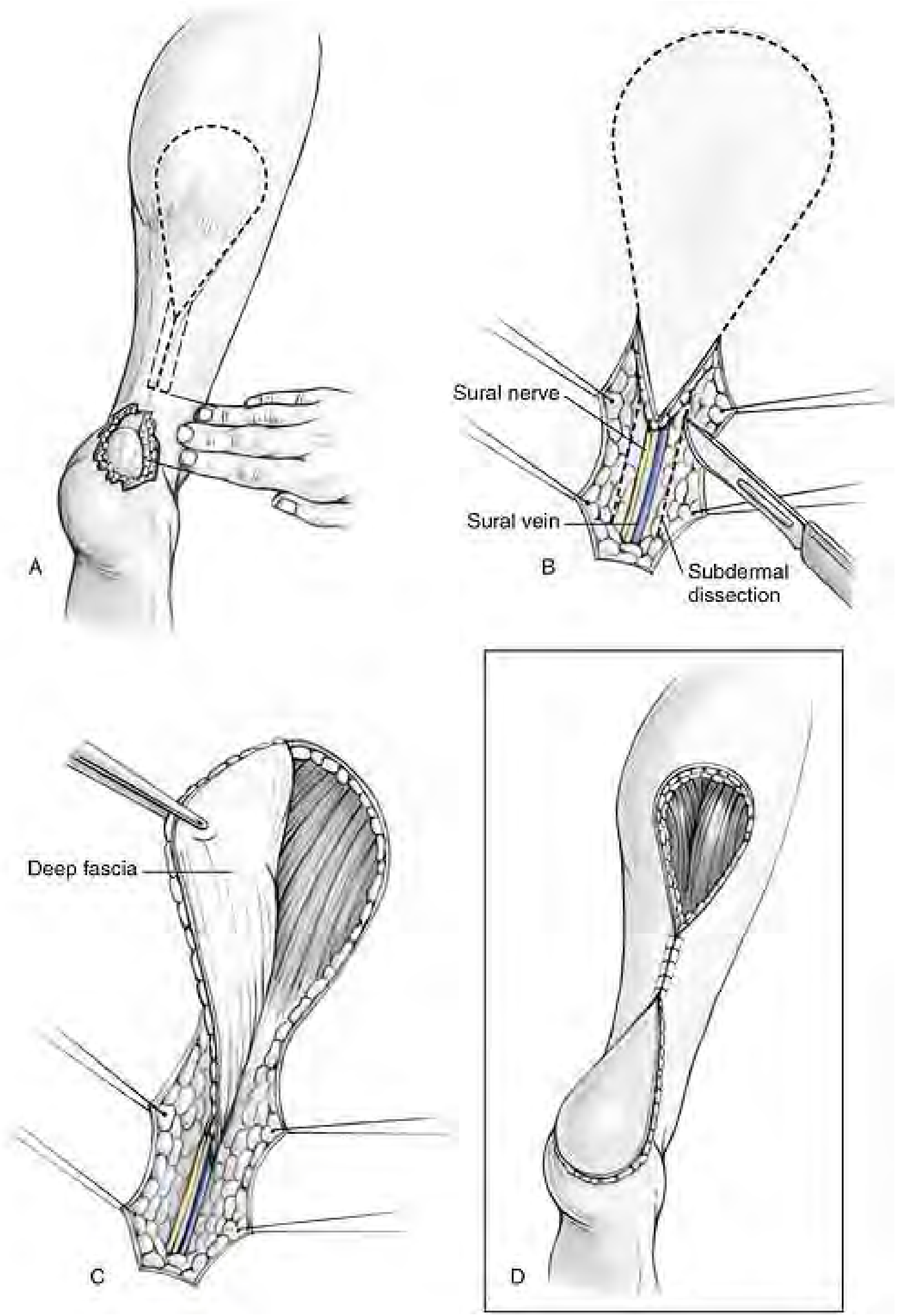
A
Sural nerve
Sural vein
B
Subdermal dissection
Deep fascia
C
D

FIGURE 29-5 The skin island has to be planned according to the size of the defect with its center along the line of the short saphenous vein. The pivotal point is three fingers breadth to the tip of the lateral malleolus (A). The skin incision

starts distally along the line of the pedicle to localize the vein. The subcutaneous pedicle should be at least 3 cm wide (B; 1â€”sural vein and nerve; 2â€”subdermal dissection). Dissection proceeds along the plane below the deep fascia; thus, inclusion of the pedicle in the flap is easily possible and the nutrient anastomosis can be preserved (C: 3â€”deep fascia included). After complete flap elevation, the flap is transposed to the recipient site (D).

If signs of insufficient perfusion or venous drainage occur after flap transposition and if decompression of the pedicle is unsuccessful in resolving this problem, a suralis flap delay procedure is possible. In this procedure, the flap is replaced into its donor site bed, a powder-free glove is placed between the elevated fascia and the gastrocnemius muscle, and transposition is performed 2 to 4 days later.

In some cases, the reverse flow sural artery flap can be modified. Hence, muscle tissue can be included in the flap when the perforating vessels from the gastrocnemius muscle to the vascular axis

of the sural nerve (see previous discussion) are preserved. Flap harvest is similar to the procedure described previously. The only difference is that a fragment of the gastrocnemius muscle is taken from the lower part of the muscle without separating the fascia from the muscle fragment. Another option is to elevate the flap as a fascial flap without a skin island; this is the preferred approach of some authors in patients with heel defects. In these cases, the flap is covered by split thickness skin grafting. Furthermore, some authors describe the possibility of reconstructing defects distant from the sural artery flap donor site by performing a free microvascular sural artery flap.

Donor site closure is possible primarily when a skin island of less than 4 cm in width is harvested. In these cases, the donor site is acceptable in terms of function as well as aesthetics. Larger donor site defects have to be closed by split thickness skin grafting. In these cases, the acceptability of the donor site scar is questionable, particularly among obese patients and females.

Postoperative Management

The most important postoperative issue is prevention of compression of the vascular pedicle. This can be achieved either by an adequately elevated position of the leg and/or by the use of conventional splints with a gap over the flap; however, the authors prefer using an external fixation device. External fixation incorporates the treatment of concomitant fractures and prevention of an equinus, as well as elevated positioning of the leg.

The administration of anticoagulants in the postoperative period after pedicled flap reconstruction remains controversial. To the authorsâ€™ knowledge, there is no scientific evidence of beneficial effects of an anticoagulation therapy in terms of flap survival or thrombosis rates. Moreover, one could postulate that anticoagulants increase perioperative morbidity due to bleeding and hematoma. Nevertheless, in our center, as in many other centers, we usually use an postoperative regimen of hydroxyethyl starch (HAES 10%; 500 mL in 24 hours for 5 days) and heparin or low-molecular-weight heparin (Clexane). Moreover, all patients are treated with antibiotics (cephalosporin) perioperatively. Antibiotic treatment is prolonged if the patient shows an increased risk of postoperative wound infection or persistent infection.

Surveillance of flap perfusion in terms of arterial as well as venous flow must be performed regularly during the first postoperative days. In the authorsâ€™ department, the capillary refill is tested every hour for the first 48 hours. Thereafter, the intervals between testing are increased.

A revision procedure should be performed if any signs of poor arterial perfusion or venous congestion are identified. New positioning of the vascular pedicle or decompression of the nutrient vessels may facilitate flap salvage. Moreover, the flap may be laid back in the donor site bed if these options fail.

The patient has complete bed rest for a minimum of 5 days. After this period compression bandages are applied to the treated leg and mobilization begins. The regimen begins with hanging down the leg for 5 minutes. The tolerance of the flap to this treatment (as well as accompanying injuries) determines the next stages of mobilization. If signs of venous congestion or poor perfusion are recognized, the mobilization regimen is prolonged. In most cases, the patient is completely mobilized after 10 days. The further use of a compression garment is obligatory for a period of at least 1 year after discharge to improve contouring of the flap and avoid hypertrophic scarring. Secondary corrections at the recipient as well as the donor site should not be performed until at least 6 months post surgery.

Complications

Partial or complete necrosis of the sural artery flap remains the most important issue concerning the success or failure of this reconstructive method. Necrosis rates between 5% and 36% have been described in the literature. A detailed analysis of the study cohorts helps to explain these different findings. When patients with comorbidities are excluded, the patient's age alone has been found to be a significant risk factor for flap necrosis. Furthermore, systemic diseases such as diabetes mellitus, peripheral arterial disease, and venous insufficiency are considered to be negative predictors for flap survival. Therefore, necrosis rates should always be interpreted in relation to the risk profile of the individual patient.

However, comorbidity with any of the above-mentioned diseases does not present an absolute contraindication to this approach, especially in view of the limited alternatives. First, it should be

kept in mind that the flap does still have a chance of success; second, the sural artery flap often represents the last chance of limb salvage in multi-morbid patients who are not suitable candidates for a free flap transfer. Some authors suggest that in cases of partial and even complete necrosis, the sural artery flap can serve as a valuable biologic dressing, so that the reconstructive surgeon sometimes experiences a well-vascularized granulating wound after debridement of the necrotic flap that allows successful split thickness skin transplantation. It seems that vascularization improves under the flap, which facilitates a skin graft on a wound bed that initially would not have been possible.

Several technical guidelines should be followed to prevent perioperative morbidity, particularly among patients with an increased risk of flap necrosis. The first of these guidelines is the use of the Doppler ultrasound in preoperative planning. Then, as described previously, the sural nerve should be included in the flap and the surgeon should pay attention to the subcutaneous layer. It can cause the flap to be bulky at the recipient site and, therefore, increases the risk of pedicle compression. The solution is to plan the flap slightly larger than the defect to facilitate skin-to-skin closure or to perform extensive mobilization of the surrounding skin, although this in turn can cause wound healing disturbances. We recommend the use of an external fixation device to prevent pedicle compression and to facilitate postoperative care, especially in less compliant patients. Tunneling of the vascular pedicle should only be performed if elastic skin surrounds the defect; otherwise, the pedicle should be grafted or the skin paddle should be planned in a teardrop shape to facilitate pedicle coverage (see previously). Furthermore, it should be taken into consideration that the more proximal the donor site is planned, the greater the risk of necrosis.

Further complications described in the literature are not specific to sural artery flaps but relate to the underlying defect or the surgical procedure, respectively. Wound infection or persistent infection are common findings as well as hematoma, delayed healing, or persistent osteitis (which itself is usually due to incomplete debridement). As described previously, inelastic skin and induration are familiar findings especially in patients with comorbidities. Therefore, the skin surrounding the flap margins or the pedicle may require skin grafting for wound closure. Edema formation is common, especially in patients with venous insufficiency. In these cases, application of compression bandages may help improve local trophicity.

Donor site morbidity is generally low. The most common findings are neuroma of the sural nerve and scarring. In these patients, secondary corrections should be performed. Neuromas have to be resected, and the nerve stump has to be buried in the surrounding musculature.

Recommended Reading

Almeida MF, da Costa PR, Okawa RY. Reverse-flow island sural flap. *Plast Reconstr Surg.* 2002;109:583-591.

Arnez ZM, Kersnic M, Smith RW, et al. Free lateral arm osteocutaneous neurosensory flap for thumb reconstruction. *J Hand Surg [Br].* 1991;16:395â€“399.

Baumeister SP, Spierer R, Erdmann D, et al. A realistic complication analysis of 70 sural artery flaps in a multimorbid patient group. *Plast Reconstr Surg.* 2003;112:129â€“140.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 30

Fasciotomies of the Lower Extremity

Christopher J. Salgado

Amir A. Jamali

Jason Nascone

Indications/Contraindications

Compartment syndrome is the most common indication for fasciotomy. Compartment syndrome is a clinical condition with elevated tissue pressure within a closed anatomic compartment. Muscles are contained within an osseofascial compartment that has a limited capacity to expand. The causes of compartment syndrome can be divided into two major categories: decreased compartment size and increased compartment volume. Decreases in compartment size can be due to extrinsic factors such as tight dressings or casts or due to intrinsic causes such as bleeding into a compartment after injury or a postoperative coagulopathy. Increased compartment volume can occur at the macroscopic or microscopic level. Bleeding or iatrogenic infiltration of intravenous fluid into a closed compartment are both common causes of compartment syndrome. At the microscopic level, compartment volume can be increased in proportion to either increased capillary permeability and/or capillary pressure. Conditions associated with tissue damage such as burns, ischemia/reperfusion, and trauma can all lead to increased capillary permeability. Increased capillary pressure is the underlying cause of compartment syndrome due to venous obstruction or exercise.

The underlying pathologic condition leading to compartment syndrome is an elevated tissue pressure which leads to decreased arteriolar perfusion. At this point shunting occurs, bypassing the capillary circulation which then worsens the tissue ischemia, and in turn, increases the capillary permeability and interstitial tissue pressure. This vicious cycle can quickly lead to permanent tissue damage if not treated expediently. The tissues most at risk in compartment syndrome are the nerves and muscles. If untreated, compartment syndrome can lead to Volkmann's contracture, a permanent paralysis of muscles in the compartment with scarring in a shortened position leading to the term â€œcontracture.â€• Assessment of an injured lower extremity must include a thorough evaluation of factors, which can contribute either directly or indirectly to compartment syndrome. A list of such factors is given in Table 30-1. The treatment of compartment syndrome is the correction of the underlying pathologic state and the performance of a fasciotomy. Fasciotomy is the incision of fascial compartments in order to expand the size of the compartment and to restore tissue perfusion to the contents of the compartment.

Preoperative Planning

Preoperative planning of a fasciotomy is based on an accurate and timely diagnosis of compartment syndrome. The most important data guiding this decision is the clinical examination. In the setting of an awake, unsedated patient, the diagnosis can commonly be made based on clinical grounds. The classical clinical signs of compartment syndrome are the six P's: pain, pressure, paresthesia, paralysis, pallor, and pulselessness. Note that significant muscle damage occurs prior to the onset of pallor, pulselessness, and paralysis, and that these are late findings of a missed compartment syndrome. The most sensitive clinical sign of compartment syndrome is pain. This is often described as being out of proportion to the injury. In the author's experience, the variability of pain threshold among patients can make the assessment of â€œexpected level of painâ€• somewhat arbitrary. The presence of a greater than expected level of pain does not indicate definitive compartment syndrome and the absence of pain does not rule out the diagnosis, particularly in the setting of a possible neurologic injury. The use of sedatives and analgesics, a history of central nervous system trauma, and the possibility of peripheral nerve injury highlight the need for a high index of suspicion and the need for early intracompartmental pressure measurements. Passive stretching of muscles within the compartment leads to elevated pressures and increased pain, another valuable tool in establishing a clinical diagnosis. Other signs of compartment syndrome include paresthesias and paralysis. Paresthesias are due to ischemic damage to the peripheral nerves running in the compartment. Decreased 2-point discrimination is the most consistent early finding. Correlation has also been reported between diminished vibration sense (256 cycles per sec) and increasing compartment pressure. On deep palpation, a firm wooden feeling is a specific sign when present. Bullae may also be observed. In later stages, the paresthesias can progress to complete anesthesia in the distribution of the peripheral nerve.

Table 30-1. Compartment Syndrome Risk Factors

History	Injury	Treatment
Crush injury	Open and closed fractures	Fluid administration
Entrapment	Arterial injury	Tourniquets
Ischemia	Venous injury	Positioning
Shock/hypotension	Gunshot wounds to extremity	MAST
Overdose/unconsciousness	Coagulopathy	Arthroscopy pumps
Tight ski boots	Shock	Jet lavage
Coumadin	Deep vein thrombosis	Revascularization

Weightlifting/overuse	Burns	Vein ligation
Knee arthroscopy	Muscle tear	Fracture tables
	Snake envenomation	Tight wound closures
	Ruptured Baker's cyst	Constrictive dressings
		Regional anesthesia

MAST, Military anti-shock trousers

Paralysis can be due to muscle ischemia, nerve ischemia, direct injury to these structures, or secondary to pain inhibition. Pulselessness is uncommon in an isolated compartment syndrome and heralds a probable vascular injury. Laboratory testing revealing a creatine kinase (CK) of 1,000 to 5,000 U/mL or higher or the presence of myoglobinuria may alert the physician to the occurrence of compartment syndrome. When the clinical picture is borderline, compartment pressure measurements must be performed as soon as possible.

A number of techniques have been employed to determine compartment pressures, including variations in size and needle design. At our institution all compartment measurements are performed with a side-port needle attached to a commercially available pressure monitor (Stryker, Kalamazoo, MI) or a standard arterial line pressure transduction line (Fig. 30-1). The Stryker pressure tonometer is widely used, and pressure measurements from the Stryker device are within 5 mm Hg of the slit catheter for 95% of all readings. Measurements with a standard 18-gauge needle are not accurate and are not recommended. Pressure measurements should be performed within all compartments and at multiple sites.

The compartment pressure data can be viewed in isolation or in relation to the patient's diastolic blood pressure. Although there is no absolute minimum compartment pressure value, most current literature indicates that the Î”P value from measured compartment pressure to diastolic blood pressure is a more valuable guide in performance of a fasciotomy. Studies suggest that the ischemic threshold of muscle is a perfusion pressure of at least 20 mmHg between the compartment pressure and the diastolic pressure. The Î”P is a direct measure of the pressure gradient between diastolic blood pressure and the tissue pressure within the compartment, indicating the presence of shunting. At our institution, a Î”P of 30 mmHg combined with increased palpable pressure is a strong indication and a Î”P of 20 mmHg an absolute indication for fasciotomy.

Therapy is begun for the treatment of compartment syndrome while preparations are made for actual surgical decompression. The affected limb(s) are placed at the level of the heart. Elevation is contraindicated because it decreases arterial inflow and narrows the arterial-venous pressure gradient and thus worsens the ischemia. If a cast is on the affected extremity, releasing one side of the plaster cast can reduce compartment pressure by 30%; bi-valving can produce an additional 35% reduction; and cutting the cast padding may further decrease compartmental pressure by 10% to 20%. In cases of snake envenomation, administration of antivenom may reverse a developing compartment syndrome. Hypoperfusion may be corrected with crystalloid and blood products, and mannitol may reduce compartment pressures and lessen reperfusion injury.

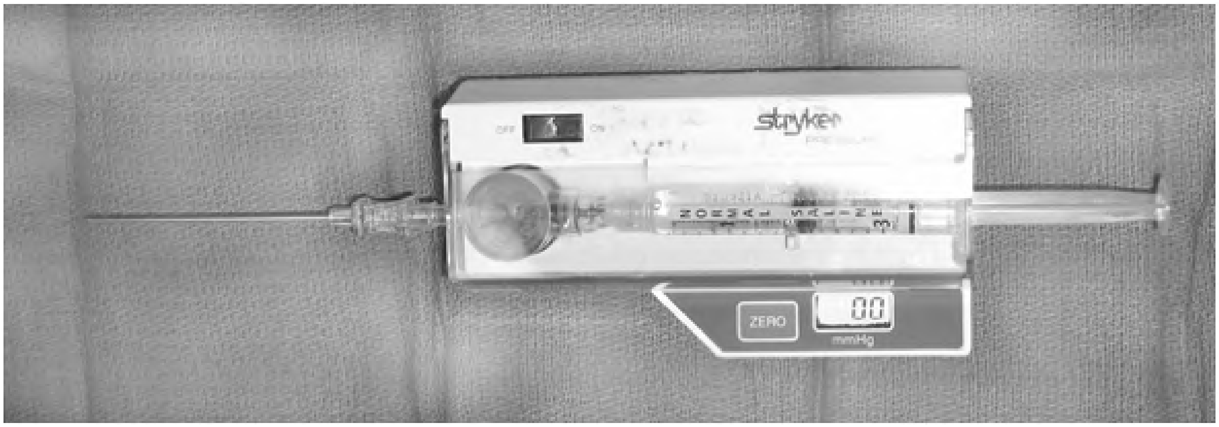

FIGURE 30-1 Stryker pressure monitor (Stryker, Kalamazoo, MI) with side-port needle.

Hyperbaric oxygen (HBO) is a valuable adjunct in the treatment of compartment syndrome. It promotes hyperoxic vasoconstriction, which reduces swelling and edema and improves local blood flow and oxygenation. It also increases tissue oxygen tensions and improves the survival of marginally viable tissue. The best results are obtained when therapy is started early after fasciotomy. Twice-daily treatments at 2.0 atmospheres absolute (ata) to 2.5 ata for 90 to 120 minutes for 5 to 7 days, with frequent examinations of the affected area, may be beneficial. This may be more practical in centers familiar with the use of hyperbaric oxygen therapy.

Vacuum-assisted closure devices (VAC) have a number of advantages in the treatment of post-fasciotomy wounds. They reduce interstitial edema and provide a one-way flow of exudate from the wound. Additionally, this therapy increases granulation tissue formation and may lead to an earlier ability to close or skin graft the wound.

Fasciotomy of the Thigh

Thigh and gluteal compartment syndromes are uncommon and may often go unrecognized. The pathophysiology and the principles of diagnosis and treatment, however, are the same as those for other compartment syndromes. Gluteal compartment syndromes are often associated with substance abuse and a prolonged period of unconsciousness or recumbency and can occur in the absence of any obvious trauma. As a result of the large muscle mass involved, systemic manifestations of a crush syndrome are usually present. Altered mental status and metabolic abnormalities may distract from the primary problem, resulting in delayed diagnosis and treatment. The proximity of the sciatic nerve can result in compression-induced neuropathy.

Clinical Findings

Thigh compartment syndrome is rare because of the large volume required to cause a pathologic increase in the interstitial pressure. It may occur in the setting of high-energy thigh trauma such as femur fractures with an associated crush component. These patients often have pain and swelling after fixation, which may confound the diagnosis. In addition, these trauma patients are often obtunded and require substantial fluid resuscitation, increasing the risk of compartment issues. The fascial compartments in the thigh blend anatomically with muscles of the hip, potentially allowing extravasation of blood outside these compartments. Anticoagulation can be a major risk factor leading to bleeding into the thigh compartments and the development of a compartment syndrome.

Patient Positioning

The approach to thigh compartments may be medial or lateral depending on the area of injury or suspected hematoma. The thigh should be prepared from the iliac crest to the knee joint with the patient in either the lateral decubitis position or supine.

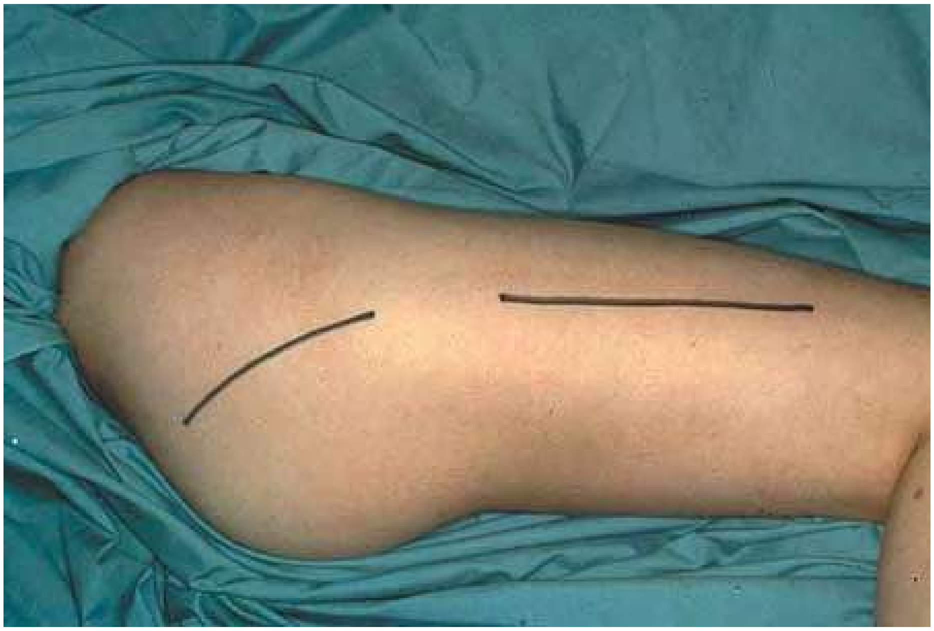

FIGURE 30-2 Skin markings for gluteal and thigh (anterior and posterior) compartment release.

Technique

For lateral and posterior compartment syndromes, the skin and subcutaneous tissues are incised beginning just distal to the intertrochanteric line and extending to the lateral epicondyle of the femur to expose the iliotibial band or fascia lata (Fig. 30-2). The iliotibial band is incised for the length of the incision. The vastus lateralis muscle is reflected superiorly and medially to expose the lateral intermuscular septum, which is incised for the length of the incision, thus freeing the posterior compartment. Caution is required to control the perforating branches of the descending branch of the lateral femoral circumflex artery traversing the lateral intermuscular septum (vessels which supply the anterolateral thigh skin), since these may retract and bleed during this portion of the decompression (Fig. 30-3). After the anterior and posterior compartments have been released, measure the pressure of the medial compartment. If elevated, then the compartment can be approached through a separate medial incision. The incision is carried along the course of the saphenous vein. Reflect the sartorius muscle superiorly, and incise the medial intermuscular septum. Intramuscular hematomas may require release through gentle muscle splitting. The wounds are packed open and a large bulky dressing applied or, alternatively, a vacuum-assisted closure device is applied.

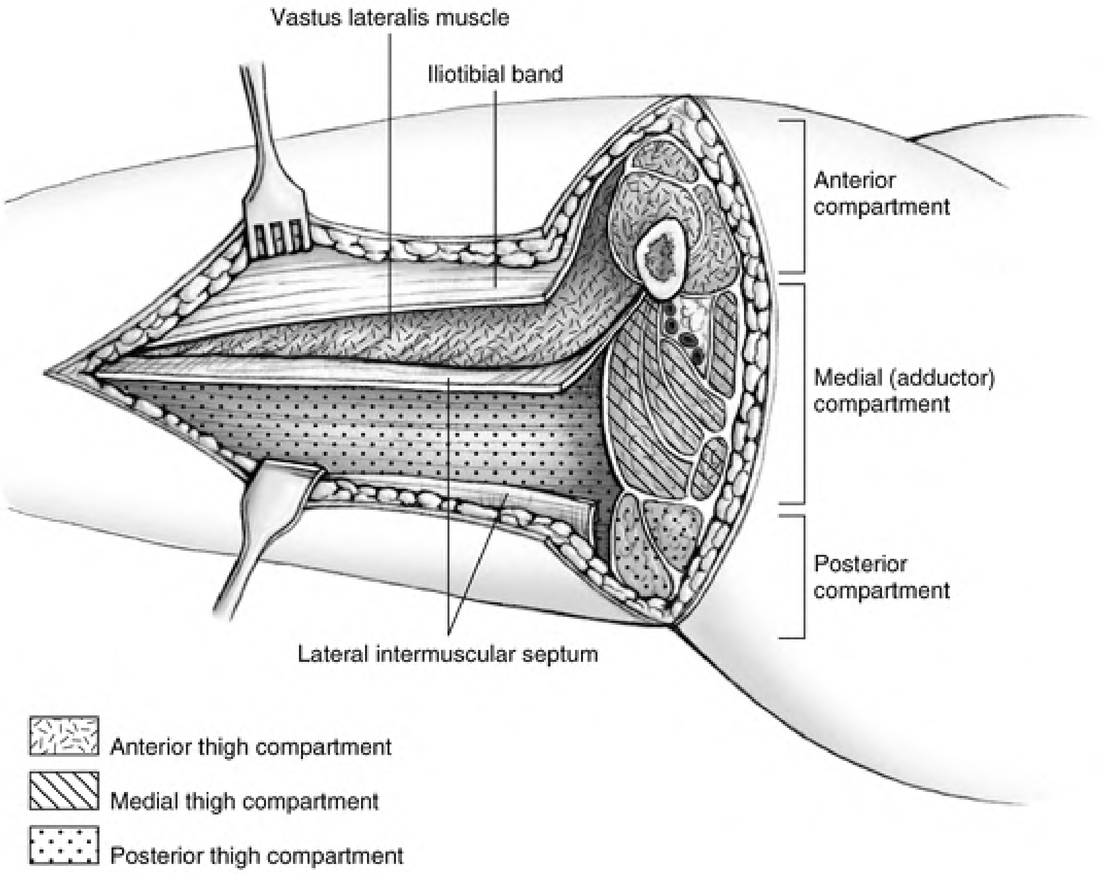
Vastus lateralis muscle
Iliotibial band
Anterior compartment
Medial (adductor) compartment
Posterior compartment
Lateral intermuscular septum
Anterior thigh compartment
Medial thigh compartment
Posterior thigh compartment

FIGURE 30-3 Schematic lateral view showing decompression of anterior compartment performed by incising the fascia latae longitudinally. The vastus lateralis is retracted medially to expose the lateral intermuscular septum, which is incised to decompress the posterior compartment.

Two to three days later, the patient is returned to the operating room for debridement of any nonviable tissue. If there is no evidence of necrotic tissue, the skin is either loosely closed, packed once again for closure at a later date, or again covered with a vacuum-assisted closure device. Often a medial thigh fasciotomy is not needed once a lateral release is performed.

Results

Because this diagnosis is not always obvious, the surgeon must maintain a high index of suspicion. Early treatment by operative compartment release follows anatomic tracts and produces good results.

Fasciotomy of the Leg

The framework of the lower leg is composed of two long bones, the fibula and tibia, which are arranged in parallel and connected along their length by a fibrous membrane termed the interosseous membrane. These three structures together divide the leg into two anatomic sections, the anterior and posterior compartments. The anterior compartment is further divided into anterior and lateral by a thick anterior intermuscular septum. The lateral leg compartment is separated from the posterior compartment by the posterior intermuscular septum.

The anterior compartment has four muscles, the (a) extensor digitorum longus, (b) extensor hallucis longus, (c) peroneus tertius, and (d) the tibialis anterior. These muscles are supplied by the anterior tibial vessels and are innervated by the deep peroneal nerve, all traveling deep to the muscles along the interosseous membrane. The posterior compartment is divided into superficial and deep compartments by a thin fascia termed the transverse intermuscular septum. Three muscles are located in the superficial compartment: the gastrocnemius, soleus, and plantaris. The gastrocnemius and soleus join together at midcalf to form the Achilles tendon, which inserts into the calcaneal bone. The plantaris is a thin, small muscle, the tendon of which medially follows the bigger Achilles tendon to insert into the calcaneal bone. All of these muscles flex the foot in a plantar direction (the gastrocnemius also flexes the knee) with slight inversion. The muscles are vascularized by branches from the popliteal and posterior tibial artery and innervated by branches of the tibialis nerve from the popliteal fossa. The deep posterior leg compartment contains four muscles: the popliteus, flexor digitorum longus, flexor hallucis longus, and tibialis posterior. The lateral compartment contains two muscles, the peroneus longus and brevis. Their action consists of extension and eversion of the foot. They are vascularized by vessels from the peroneal artery and innervated by the superficial peroneal nerve (Fig. 30-4).

The posterior tibial artery, after branching from the popliteal artery, descends posterior to the tibia, within the deep posterior compartment. Distally in the lower third of the leg it is more superficial, covered only by the skin and superficial fascia and parallel to the medial border of the Achilles tendon. The peroneal artery originates from the posterior tibial artery and runs laterally and downward along the posterior fibula providing four to six segmental, circular arterial branches around the fibula, nourishing the bone, periosteum, and surrounding muscles. The anterior tibial artery commences at the bifurcation of the popliteal artery passing forward between the tibialis posterior muscle and through the interosseous membrane to the deep aspect of the anterior leg compartment. After perforation of the interosseous membrane, the artery descends adjacent and along the membrane, gradually approaching the extensor retinaculum of the ankle. At the bend of the ankle joint, it becomes superficial and known as the dorsalis pedis artery.

Patient Positioning

In cases of isolated compartment syndrome of the leg, the patient is positioned in the supine position on a standard operating table. A general anesthetic is employed in most cases. A tourniquet is applied to the thigh and isolated with broad tape or a plastic isolation dressing. The leg is prepared and draped in the standard fashion. A stockinette or isolation sheet is applied to the foot to maintain the sterile field. The tourniquet is not inflated unless active arterial bleeding is encountered.

Technique

The technique for fasciotomy of the leg can be performed using either a one incision or two incision technique (Table 30-2). The two incision technique is the gold standard. Regardless of the technique,

the skin incisions must be of adequate length to decompress all affected compartments. The performance of percutaneous fasciotomies is not recommended, particularly with compartments deformed by soft tissue edema, since this technique can be fraught with complications such as iatrogenic nerve injury. In addition, the skin itself has been shown to exert a constrictive effect on the muscle compartments in the leg.

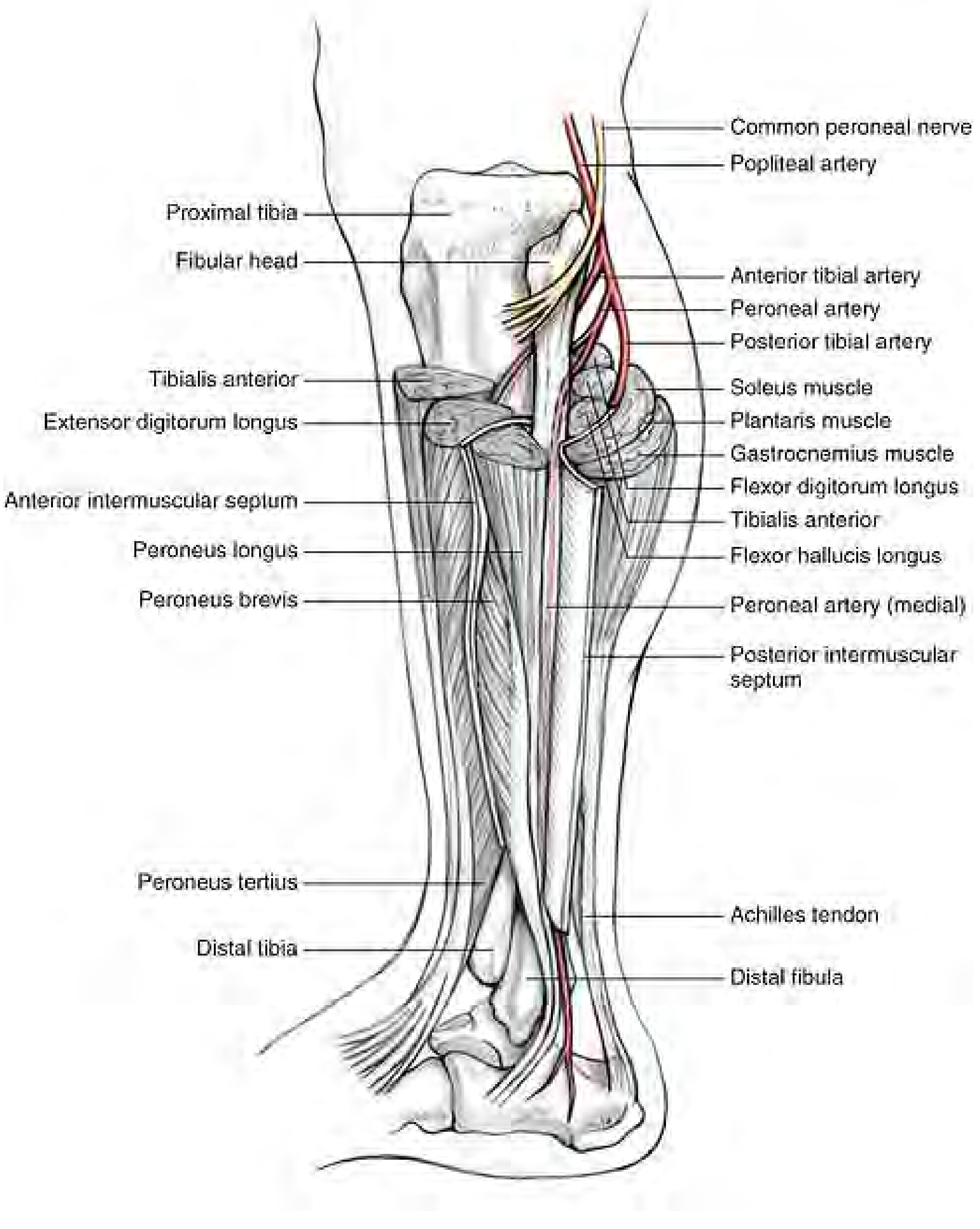
Common peroneal nerve
Popliteal artery
Proximal tibia
Fibular head
Anterior tibial artery
Peroneal artery
Posterior tibial artery
Tibialis anterior
Soleus muscle
Extensor digitorum longus
Plantaris muscle
Gastrocnemius muscle
Flexor digitorum longus
Anterior intermuscular septum
Tibialis anterior
Peroneus longus
Flexor hallucis longus
Peroneus brevis
Peroneal artery (medial)
Posterior intermuscular septum
Peroneus tertius
Achilles tendon
Distal tibia
Distal fibula

FIGURE 30-4 Lateral leg.

One Incision Fasciotomy

The planned incision is marked in line with the fibula extending to 5 cm short of either end of the fibula along the anterolateral leg. The initial step is the identification of the lateral intermuscular septum separating the lateral and anterior compartments. Make a transverse incision to expose this septum and to identify the superficial peroneal nerve just deep to the septum. Separate fasciotomies of the compartments are performed with Metzenbaum scissors. The anterior compartment is released proximally by aiming for the patella and distally by aiming for the center of the ankle in line with the tibialis anterior. Then, perform a longitudinal fasciotomy of the lateral compartment in line with the fibular shaft (Fig. 30-5). Extreme care is taken in the distal aspect of the lateral compartment at the junction of the middle and distal thirds, where the superficial peroneal nerve emerges from the lateral compartment. Direct the scissors toward the posterior lateral malleolus to stay posterior to the superficial peroneal nerve. In the case of acute compartment syndrome, we avoid subcutaneous fasciotomies. The use of the â€œsliding scissorâ€• technique decreases the ability to perform a controlled fasciotomy and should be avoided. Once the anterior and lateral compartments have been decompressed, the fibula is identified after posterior undermining of the skin for release of the deep posterior compartments. The lateral compartment musculature is elevated off the fibula, demonstrating the posterior intermuscular septum. A longitudinal incision of this septum exposes the superficial posterior compartment. Posterior retraction of the soleus and gastrocnemius muscles then exposes the deep posterior compartment for its decompression. We have found that surgical exposure of the posterior aspect of the fibula is critical in the avoidance of bleeding from the perforating branches of the peroneal artery. Release of all four compartments through one incision, however, is not commonly performed at our institutions, and we recommend a two incision fasciotomy.

Table 30-2. Pearls for Fasciotomy of the Leg

- Check intracompartmental pressures at multiple levels within each compartment.

- Threshold for fasciotomy is intracompartmental pressure within 20 mm Hg of the diastolic blood pressure.
- Beware of compartment syndrome secondary to intraoperative positioning of uninjured extremities.
- Pulselessness is NOT a common finding in compartment syndrome unless a concurrent vascular injury exists.
- Fasciotomy should not be performed more than 12 hours after a compartment syndrome is established.
- Compartment syndrome can occur with a late onset 2 to 4 days after the underlying event.
- Compartment pressure measurements with a standard 18-gauge needle are consistently higher than those obtained with specialized needles with a side-port needle or slit catheter.
- Compartment pressures can vary among several different points in each compartment at risk.
- In any patient at risk, baseline compartment pressures are necessary in the case of a suspicion of compartment syndrome in the future.

Two Incision Feg Fasciotomy (Gold Standard)

- Mark both incisions before making the first surgical incision.
- Beware of the superficial peroneal nerve as it emerges anteriorly within the lateral compartment at the junction of the middle and distal one-third of the fibula.
- Beware of the use of vessel loops or other elastic forms of skin tension applied at the time of fasciotomy, as these may lead to recurrence of the compartment syndrome if excessively tight.
- Close medial wound before the lateral wound to avoid need for soft tissue coverage.

One Incision Leg Fasciotomy

- Obtain complete visualization of the release of the deep posterior compartment, as extensive bleeding can occur secondary to perforating vessels on the posterior aspect of the fibula.
- Extend incision to within 5 cm of either end of the fibula on the anterolateral leg.
- Beware of the superficial peroneal nerve at the junction of the middle and distal one-third of the leg.

Two Incision fasciotomy

The two incision fasciotomy is the gold standard treatment for compartment syndrome of the leg, particularly in the hands of surgeons with limited previous experience and limited assistance in the operating room. The positioning and preparation are identical to that of the one incision fasciotomy. It is crucial to mark the medial and lateral skin incisions before making the incision to ensure an adequate skin bridge is maintained. After one single incision is made, the

skin envelope will retract in the opposite direction resulting in a narrow skin bridge anteriorly, and marking these incisions will avoid this complication. After performance of the anterior and lateral fasciotomies, a separate incision is made along the posteromedial leg, 1 to 2 cm posterior to the tibia, measuring at least 15 cm in length. The greater saphenous vein and saphenous nerve are identified and retracted anteriorly. The fasciotomy is extended as far as possible proximally and distally to the level of the medial malleolus. The soleus is then released from the posteromedial tibia with a concurrent release of the deep posterior compartment (Fig. 30-6). A common pitfall is to not adequately release the soleus muscle insertion thereby adequately releasing the deep posterior compartment. Care should be exercised when releasing the deep compartment ligating or cauterizing branches traversing the soleus muscle. The posterior tibial neurovascular bundle is just deep to this transverse intermuscular septum, which separates the deep from posterior compartments and is therefore in close proximity to the release. The skin incisions can also be slightly staggered, with the lateral more proximal (the anterior and lateral compartments are largely tendonous at their distal extent) and the medial incision made more distally. Finally, the fascia and skin are then re-evaluated for adequacy of release.

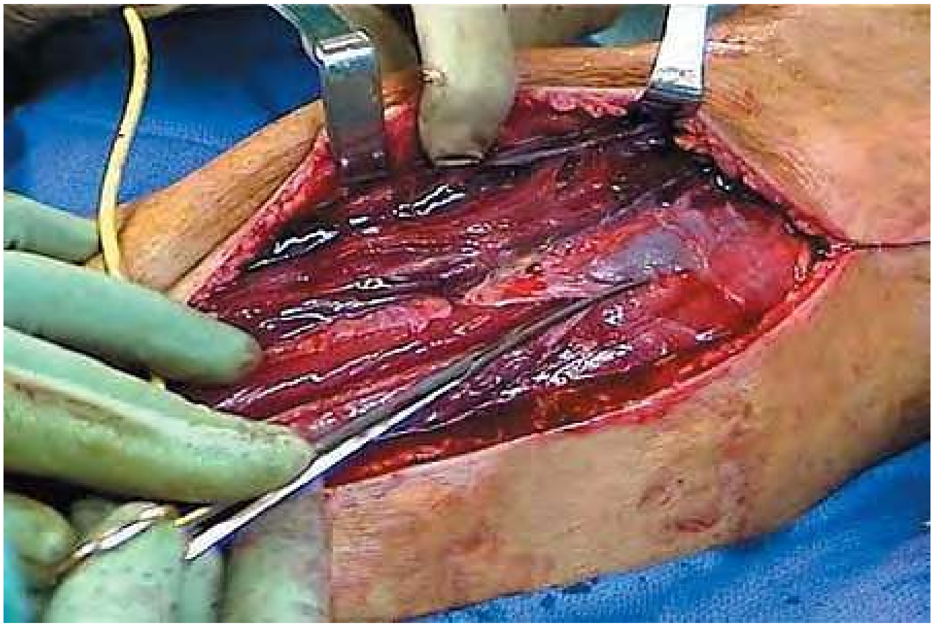

FIGURE 30-5 Intraoperative view after anterior and lateral compartment decompression. Note the intermuscular septum separating the anterior and lateral compartments.

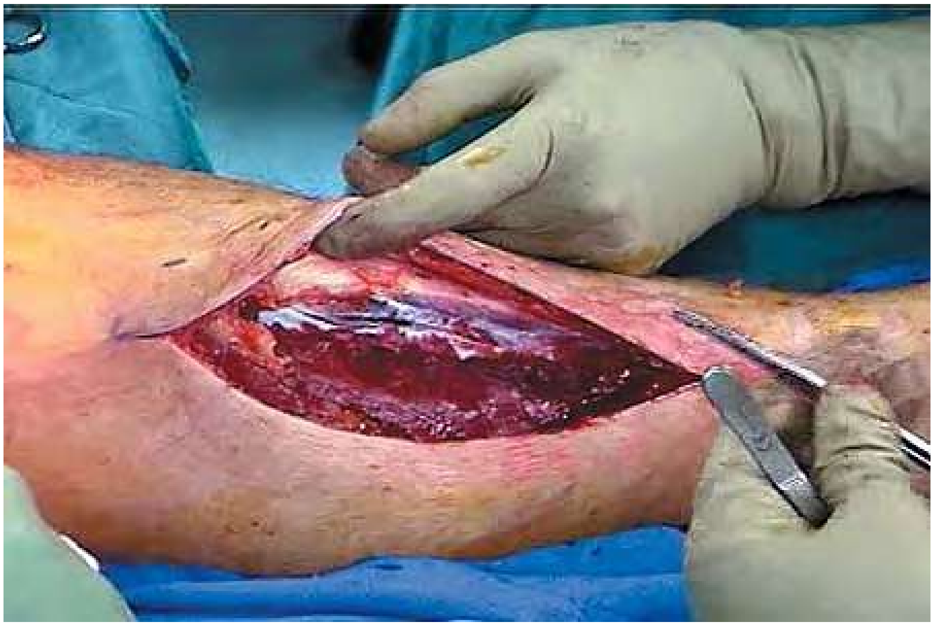

FIGURE 30-6 Intraoperative view after release of superficial posterior compartment and before release of deep posterior compartment.

After either type of fasciotomy, compartment pressures are again checked using a sterile side-port needle attached to the non-sterile pressure monitor held by an unscrubbed assistant or, alternatively, using an arterial pressure monitor setup. The wounds are then copiously irrigated with crystalloid. Devitalized tissue if present is debrided. Vessel loops can be applied to the skin edges to prevent marked skin retraction but must be used judiciously to avoid a recurrence of the compartment syndrome. At our institution, we routinely employ a vacuum-assisted closure device over all open fasciotomy wounds to maintain a one-way flow of extravasated fluid, to encourage the formation of granulation tissue, and potentially to minimize the area needed for later skin grafting. The patient is returned to the operating room at 2 to 3 days for a repeat irrigation and debridement and partial closure. At or around postoperative day 7, the wound is again treated with irrigation and debridement. At that time, definitive closure is performed with additional skin coverage obtained with a split-thickness skin graft. We usually strive to close the medial wound first to provide bony coverage and avoid the need for soft tissue coverage, whereas the lateral wound may be easily skin grafted if appropriate. It is imperative to maintain the foot in a neutral position to avoid equinus contracture either with external fixation or external splinting.

Results

The primary goal of fasciotomy is to prevent permanent nerve and muscle damage leading to Volkmann's contracture. The results of fasciotomy can be analyzed based on two endpoints: muscle and nerve function, and wound- or incision-related complications. A number of studies have demonstrated that the majority of patients have normal leg function if the fasciotomy is performed within 12 hours of the onset of compartment syndrome. In our experience, prompt recognition and early treatment of the compartment syndrome leads to minimal long-term complications. Despite early and aggressive fasciotomy, however, nearly 20% of patients may have persistent motor deficits at 1 year follow-up. Wound complications associated with fasciotomy include numbness and persistent ulceration at the fasciotomy site. The risk of sensory changes has been reported to be as high as 70% in some reports. The use of a VAC dressing may contribute to a lower risk of hematoma and edema and an expeditious granulation of fasciotomy wounds that are not able to be closed. Little or no return of function can be expected when diagnosis and treatment are delayed. Tendon transfers and foot stabilization may be indicated as late treatment; but in most patients, enough scarring and contracture eventually develop in the anterior musculature to prevent foot drop. A foot drop brace (ankle foot orthosis) is indicated for the first few months until fibrosis occurs. Some patients experience persistent gastrocnemius and soleus muscle weakness thought to be due to the loss of the supporting compartment fascia.

Fasciotomy of the Foot

Anatomically, the foot consists of well-demarcated osseofascial spaces that subdivide the foot into discrete compartments. These compartments are filled with muscles, nerves, and tendons and are lined by a tight membrane (the fascia). There are four clinically relevant compartments known as the medial, central, lateral, and interosseous (Fig. 30-7). Other anatomic compartments of the foot may be identified with dyes or injection studies but are not clinically relevant. Muscles within the medial compartment are the abductor hallucis and flexor hallucis brevis, and within the central (calcaneal) compartment lay the flexor digitorum brevis, quadratus plantae, and adductor hallucis muscle. There are four dorsal and plantar interosseous muscles between the first and fifth metatarsals, and these comprise the interosseus (intrinsic) compartment. The lateral compartment houses the abductor digiti minimi and flexor digiti minimi brevis muscle.

Compartment syndrome in the foot is commonly due to severe local trauma after fairly significant industrial, agricultural, and motor vehicle accidents in which crushing of the foot occurs. The trauma leads to bleeding from injured bones or muscles and avulsed vasculature. When the bleeding and inflammation of the muscle becomes significant, it will exceed the capacity of blood flow in and out of the small compartments.

Clinical Findings

Tense tissue bulging may be the most reliable symptom in compartment syndrome of the foot, and in the presence of this massive swelling pulses are usually not palpable. Increased pain on passive dorsiflexion of metatarsophalangeal joints is another key finding that indicates myoneural ischemia in the foot intrinsic muscles. It is imperative to maintain a high clinical suspicion based on the severity of the traumatic incident. Compartment syndromes of the foot are often associated with compartment syndromes of the deep posterior compartment.

Patient Positioning

The patient position is supine to allow easy access to the dorsum and medial aspect of the foot. A tourniquet may be used, but is not insufflated unless there is active arterial bleeding. The knee should be included in the preparation to enable better mobility of the limb during the procedure.

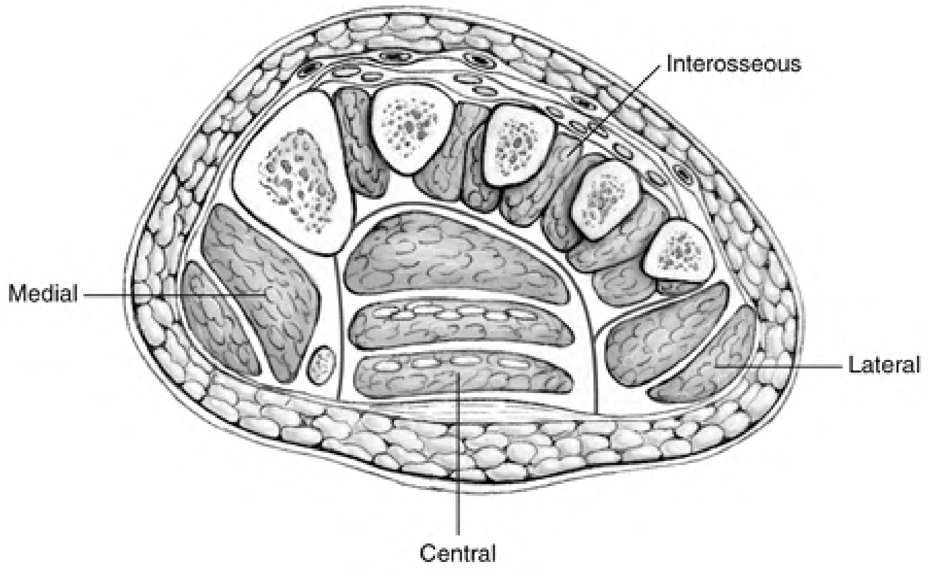

FIGURE 30-7 Schematic coronal section of right foot through base of metatarsals. Medial, central, lateral, and interosseus compartments are shown.

Technique

The same principles that apply to fasciotomy of the thigh and leg apply to the foot. Appropriate treatment for a suspected compartment of the foot is immediate fasciotomy. Debridement of marginal tissue at the time of initial fasciotomy is not advised because it is very difficult to determine muscle viability and contractility in the foot. Once the muscle is decompressed, much of it may recover after compartmental release.

Medial Approach

Effective decompression of all four compartments can be accomplished through a medial longitudinal (Henry) approach or through an additional two parallel dorsal incisions along the length of the second and fourth metatarsal bones (more common). The medial approach can be used to decompress the medial and central compartments as well as the remaining foot compartments (lateral and interosseous). The incision is made 3 cm from the sole of the foot and extends from a point below the medial malleolus to the proximal aspect of the first metatarsal (Fig. 30-8). The posterior tibial neurovascular bundle is identified and preserved. This may be very difficult in a massively swollen foot. The fascia overlying the abductor hallucis and flexor hallucis brevis is released. Dissection is continued adjacent and deep to the first metatarsal toward the medial intermuscular septum, separating the medial compartment from the central compartment, which is opened longitudinally. The lateral plantar neurovascular bundle is found between the flexor digitorum brevis and quadratus plantae muscles and preserved. Downward retraction of the flexor digitorum brevis along with lateral dissection will allow access to the lateral intermuscular septum separating the central compartment from the lateral compartment. This septum is divided to release the lateral compartment. Blunt dissection dorsally via the central compartment will release the interosseous compartment. Scissors or sharp instruments are not used during this portion of the dissection, since it is essentially blind to avoid injury to the neurovascular bundle.

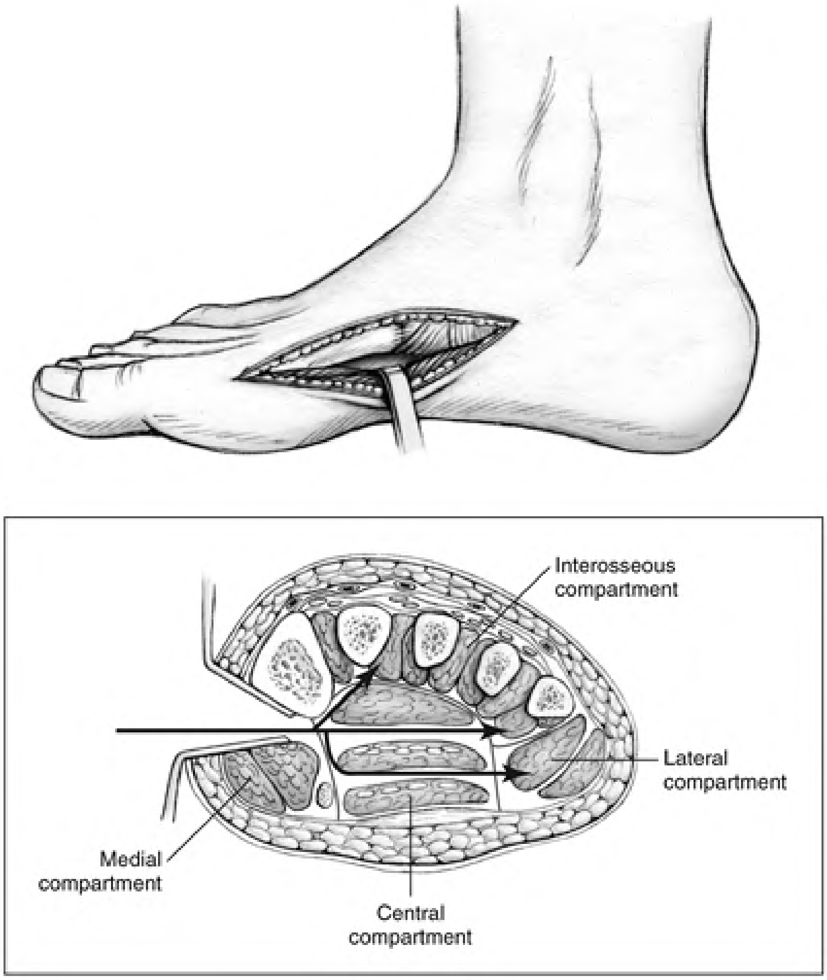
Interosseous compartment
Lateral compartment
Medial compartment
Central compartment

FIGURE 30-8 Medial approach for four-compartment release of foot compartment syndrome.

Dorsal Approach

This approach is more commonly used when there are concomitant metatarsal or Lisfranc fractures. Two parallel dorsal incisions are centered just medial to the second metatarsal and lateral to the fourth metatarsal shafts, maximizing the intervening skin bridge. The dorsal veins and the subcutaneous tissues are elevated laterally and medially to expose the respective interosseous musculature. Injury to the sensory nerves and extensor tendons is avoided. Caution should be used when making the incision between first and second metatarsal to avoid iatrogenic injury to the dorsalis pedis artery. The superficial fascia is incised longitudinally, and the interosseous muscles are elevated off the metatarsals. The first dorsal and plantar interossei are stripped from the medial aspect of the second metatarsal shaft, which is then retracted medially, and the fascia of the central and medial compartment is released longitudinally deep within the inner space. The interosseous musculature is decompressed by releasing the fascia between the second and third metatarsals also through this medial incision. The lateral incision is used to decompress the interosseous muscles between the third and fourth, and fourth and fifth metatarsals, in addition to allowing access to release the central and lateral compartments (Fig. 30-9). More commonly, a separate medial incision as described previously is used to release the medial (adductor) compartment, since the medial approach only to all four compartments carries an increased risk of damage to the neurovascular bundles in the plantar aspect of the foot. These fasciotomy incisions are sometimes used for fracture fixation.

Results

Controversy exists in what results with a missed compartment syndrome of the foot. Predominately patients end up with clawing of the toes that needs to be corrected at a later date. The argument against the release is that it complicates the soft tissue envelope, limits possible incisions for reconstruction, and carries the risk of iatrogenic neurovascular injury.

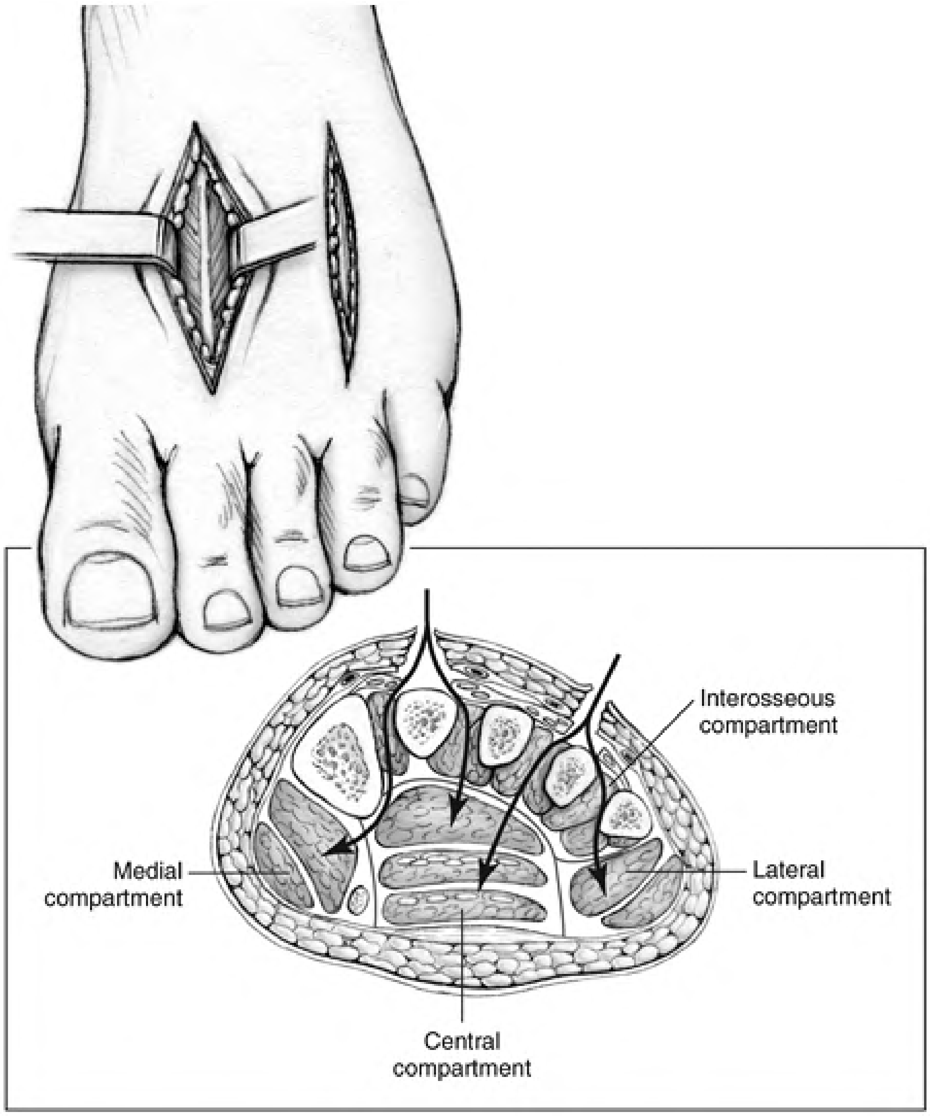
Interosseous compartment
Medial compartment
Lateral compartment
Central compartment

FIGURE 30-9 Dorsal approach for compartment release of foot compartment syndrome.

Treatment of a foot compartment syndrome with â€œbenign neglectâ€• is not advised, and early decompression as in the thigh and leg is recommended. A delay in diagnosis is a potentially devastating occurrence. Early reduction of dislocations and some fractures facilitates reduction in edema in the foot and decreases tissue breakdown due to pressure necrosis. Desire to obtain an early reduction must be weighed heavily against whether the soft tissue envelope of the injured foot will tolerate additional incisions. The fasciotomy incisions may be used to facilitate reduction and fracture fixation, but caution is warranted to avoid excessive undermining of the soft tissues. Delayed fixation is feasible, but usually not before 7 to14 days post injury, due to soft tissue edema. In addition, liberal use of external fixation, particularly in the setting of a mangled foot, is advised for stabilization and edema control.

Postoperative Management

Fasciotomy sites are dressed with a wound vacuum-assisted closure sponge, and wet to dry dressing changes with normal saline solution or silver sulfadiazine depending on the degree of contamination. The patient is returned to the operating room several days later to attempt closure. When muscle necrosis ensues, the patient is brought to surgery earlier for debridement. Wound closure should not be performed until all necrotic tissue is debrided. Direct closure can be attempted when the wound approximates without tension. When the wound edges will not oppose easily, either the wound is treated conservatively with dressing changes or a vacuum-assisted closure device. The patient is brought back to the operating room at a later date for definitive closure or skin grafting. Increased intramuscular pressure can occur with closure of fasciotomy wounds secondarily and, therefore, skin grafting of the wounds is often a safer option. Hypesthesia and painful dysesthesia resulting from compartment syndrome typically resolve slowly with time. This type of neuropathic pain can by treated with diphenylhydantoin (Dilantin), gabapentin (Neurontin), or carbamazepine (Tegretol).

Complications

Although a fasciotomy incision does result in patient morbidity, the incompletely released compartment, delayed diagnosis, or unrecognized compartment syndrome has substantially higher risks. Systemic complications including acute renal failure, sepsis, and acute respiratory distress syndrome (ARDS) have been reported in some cases. Most fatalities are due to prolonged intensive care admissions with sepsis and multisystem organ failure. If fasciotomy is done within 12 hours after the onset of compartment syndrome, the prognosis is typically good. Despite early intervention, approximately 1% to 10% of all cases of compartment syndrome develop Volkmann's contracture. Little or no return of function can be expected when the diagnosis and treatment are delayed. In no instance was benefit from fasciotomy reported after 2 or 3 days. When the procedure is done late, severe infections have been shown to develop in the necrotic tissues of many patients, frequently leading to amputations. In these situations, clinically evident by complete absence of demonstrable muscle function in any segment of the involved limb, the extremity should be splinted to maintain a functional position as muscle fibrosis and contracture develop.

Acknowledgments

The authors express their gratitude to the Department of Orthopaedics at the University of Maryland Shock Trauma and the Surgery Audio Visual Department at Cooper University Hospital (Paul Rogers) for their assistance and efforts in the preparation of this chapter.

Recommended Reading

Azar FM, Pickering RM. Traumatic disorders. In: Canale ST, ed. *Campbell's operative orthopaedics*. 10th ed. St. Louis: Mosby; 2003:1405â€“1411.

Boody AR, Wongworawat MD. Accuracy in the measurement of compartment pressures: a comparison of three commonly used devices. *J Bone Joint Surg Am.* 2005;87:2415â€“2422.

Cohen MS, Garfin SR, Hargens AR, Mubarak SJ. Acute compartment syndrome. Effect of dermotomy on fascial decompression in the leg. *J Bone Joint Surg Br.* 1991;73:287â€“290.

Finkelstein JA, Hunter GA, Hu RW. Lower limb compartment syndrome: course after delayed fasciotomy. *J Trauma.* 1996;40:342â€“344.

Fitzgerald AM, Gaston P, Wilson Y, Quaba A, McQueen MM. Long-term sequelae of fasciotomy wounds. *Br J Plast Surg.* 2000;53:690â€“693.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 31

Amputation and Stump Management

Norman S. Turner

Thomas C. Shives

Indications/Contradictions

Amputation is one of the oldest surgical procedures. Early amputations consisted of severing the extremity, and hemostasis was obtained by dipping the stump in hot oil. Techniques have dramatically improved and most of the advances have occurred during war time. Prosthetic technology now allows for amputees to run, jump, ski, swim, and be involved in competitive sports.

The indications for above and below knee amputation include life-threatening infections, malignant tumors, burns, extensive frost bite, congenital anomalies, ischemic pain, osteomyelitis, extensive trauma (including a tibial nerve ilaceration or unreconstructable vascular injury) and chronic pain. The most common indications for a below knee or above knee amputation are complications of diabetes. Functionally, the patients with below knee amputation are able to walk with prosthesis, with most patients walking within 3 months after surgery. Amputation should not be viewed as a limb salvage failure, but as a reconstructive procedure to improve function.

There are few contraindications to amputations; however, a contraindication to a below knee amputation is a non-ambulatory patient. A non-ambulating patient with a below knee amputation is at high risk for developing a flexion contracture, which can result in increased pressure on the stump and cause ulceration. Therefore, when a patient is wheelchair bound and is not a candidate for prosthetic fitting, an above knee amputation should be considered.

Preoperative Planning

It is imperative that these patients be evaluated preoperatively to determine the vascular status of the limb. The majority of patients undergoing a below knee amputation or above the knee amputation have diabetes and have some component of peripheral vascular disease. Preoperative noninvasive vascular studies, including ankle brachial indexes, are important to determine the level of amputation. An ankle brachial index is determined by measuring the ankle systolic pressure and dividing it by the brachial systolic pressure using Doppler detection of the pulses. The severity of the arterial disease is related to decreased value of the ankle/brachial index (ABI), and a value of less than .5 is considered abnormal in people with diabetes. Also, noninvasive vascular studies using transcutaneous oxygen tension measurement (TcPo2) are useful in assisting with amputation levels. Amputations are likely to heal if the TcPo2 measurements are greater then 40 mm Hg. Patients with TcPo2 values less than 20 mm Hg are at higher risk for not healing and should be evaluated with further vascular testing and possibly an angiogram before surgical intervention. A vascular surgery consult is almost always indicated before performing an amputation. With the advances in distal bypass surgery and invasive radiologic procedures, certain patients can be successfully treated with limb salvage after vascular reconstruction.

Imaging studies are important in determining the underlying pathology. Imaging of the tibia or femur is important if there is a question regarding extension of tumor or infection into the tissues or bone adjacent to the intended level of amputation or if there is a prosthetic device such as a total knee arthroplasty or internal fixation device in place, which may alter the surgical procedure.

If amputation is contemplated, optimizing the patient's medical condition before surgery is recommended. Literature has shown that patients with a serum albumin less than 3.5 g/dL or true lymphocyte count less than 1,500 cells per mL are at high risk for wound-healing difficulties.

Determination of the amputation level is important for both healing and function. The more distal the amputation level, the less energy required to ambulate. In elderly patients, a more proximal amputation may not allow for ambulation secondary to energy requirements. If a patient has good cognitive function, balance, and strength, then the most distal level with a realistic chance of healing should be attempted.

Surgery

In order for the below knee amputation to be performed correctly, proper attention to detail is important to improve the quality of the result. Gentle handling of the soft tissues, especially in diabetic patients, is important to minimize wound complications. The level of the amputation is determined by the extent of the infection, tumor, or the level that would provide optimal function with prosthesis. In general, the patient is positioned in the supine position for above or below knee amputations.

Full-thickness flaps should be used to minimize skin edge necrosis. Meticulous hemostasis and use of a drain is imperative to decrease the risk of a hematoma. The nerves should be divided sharply under tension to minimize the risk of a symptomatic neuroma. Also the bone ends are rasped until smooth to prevent bony prominences.

An open amputation is performed in patients with grossly contaminated wounds or in patients with extensive infection. These patients will require further surgeries to optimize the soft tissues around the stump, and then a definitive closure can be performed once this is accomplished.

Technique

Below Knee Amputation

Below knee amputation is the most commonly performed lower extremity amputation. A long posterior flap is used and brought anteriorly to cover the distal stump of the tibia, which should be 8.5 to 12.5 cm in length (Fig. 31-1). The flaps, if planned properly, will have minimal redundant skin in the corners, or â€œdog ears.â€• This will provide a good prosthetic fit.

The patient is placed supine on the operating room table. A nonsterile tourniquet is used and the leg is prepared and draped in the usual fashion. A skin marker is used to plan the flaps (see Fig. 30-1), and the flaps are drawn so that the posterior flap begins two-thirds of the way posterior to the anterior aspect of the tibia and then extends distally and posteriorly so that the distance will be long enough to cover the tibia (Fig. 31-2). A tourniquet can be used at the discretion of the surgeon. The incision is then made through the skin and subcutaneous tissues down to the fascia. The subcutaneous nerves including the saphenous and sural nerve can be identified and divided under tension. The fascia is then incised. The anterior compartment musculature is cut with a cautery (Fig. 31-3) down to the deep peroneal nerve, which is identified and cut under tension (Fig. 31-4), and the anterior tibial artery is identified and tied with silk suture (Fig. 31-5). The superficial peroneal nerve is identified and cut under tension. The periosteum is reflected off the tibia (Fig. 31-6), and the tibia is cut 1 cm proximal from the skin incision (Fig. 31-7). A segment of the fibula is then resected 1 cm proximal to the tibial bone cut (Figs. 31-8 through 31-10). Traction is applied, and an amputation knife is used to perform the remaining portion of the amputation (Fig. 31-11). Dissection is carried deep until blood from the posterior tibial artery and vein is identified, and then the cut is beveled distally until the fascia is cut (Fig. 31-12). The anterior aspect of the tibia is beveled (Fig. 31-13). The remaining edges are rasped until smooth. The tibial nerve is identified and transected under traction (Fig. 31-14). The tibial artery and vein are identified and tied with silk sutures (Fig. 31-15). The wound is then copiously irrigated. If a tourniquet is used, it is deflated at this time. Hemostasis is obtained. The wound is then closed in layers over a drain with sutures in the fascia (Fig. 31-16). The subcutaneous layer is closed with monofilament suture, and the skin is then closed with nylon sutures in a vertical mattress fashion (Fig. 31-17). A sterile dressing is applied, and then a compressive Robert Jones dressing is applied with plaster in full extension (Fig. 31-18).

Above Knee Amputation

Above knee amputation is the second most frequently performed lower extremity amputation. Stump length is important for the lever arm control of the prosthesis. Too long of a femoral stump, however, can lead to difficulty with fitting the prosthetic knee joint. Therefore, the bone cut should be 10 to 12 cm proximal to distal femoral articular surface.

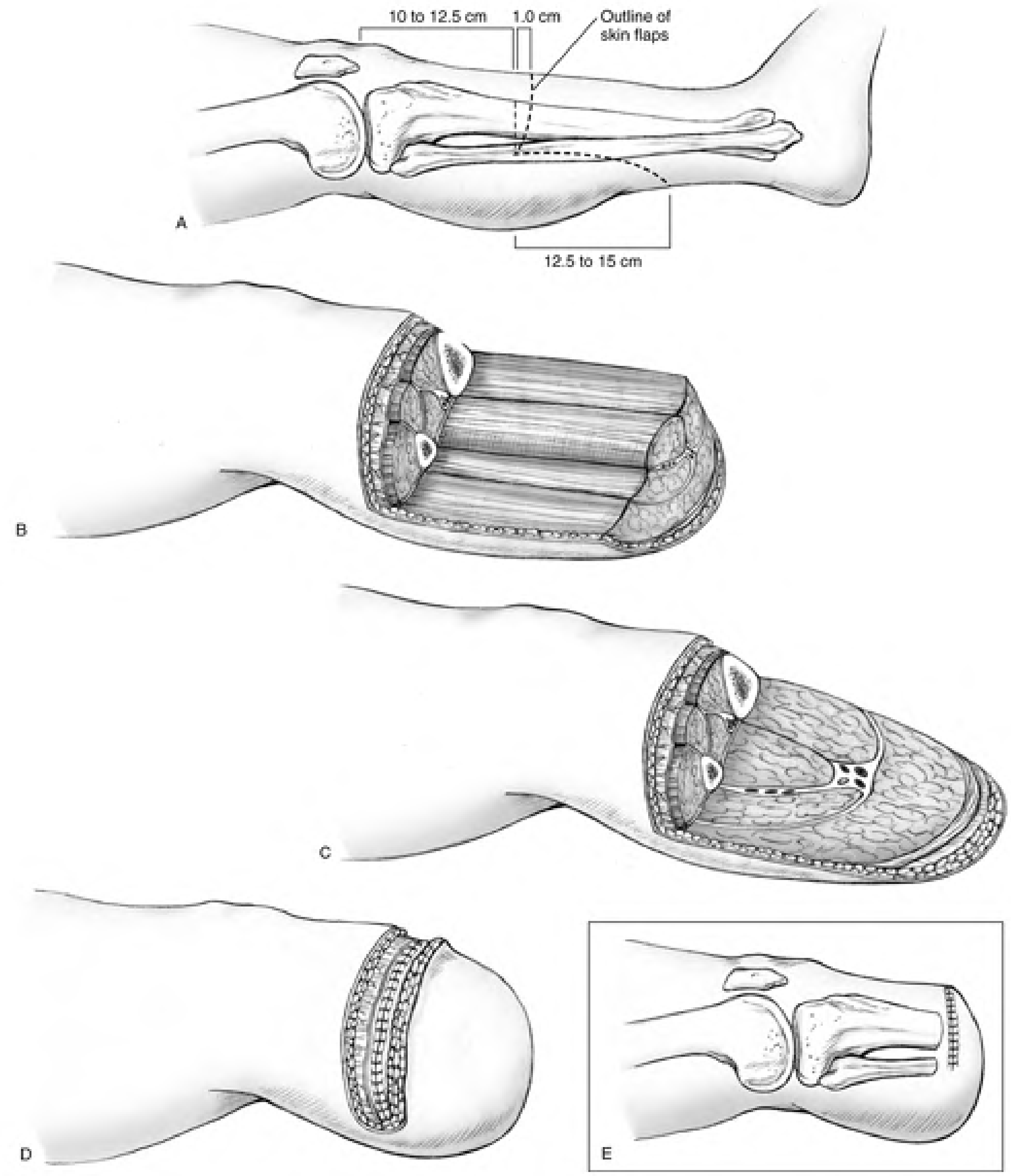
10 to 12.5 cm
1.0 cm
Outline of skin flaps
A
12.5 to 15 cm
B
C
D
E

FIGURE 31-1 Below knee amputation. A: Planning of short anterior and long posterior skin flaps. B: Amputation of distal leg. C: Tailoring of posterior muscle to form flaps. D: Closure of flap to deep fascia. E: Closure of skin flaps. (Redrawn after Burgess EM, Zettl JH. Amputations below the knee. Artif Limbs. 1969;13:1. With permission.)

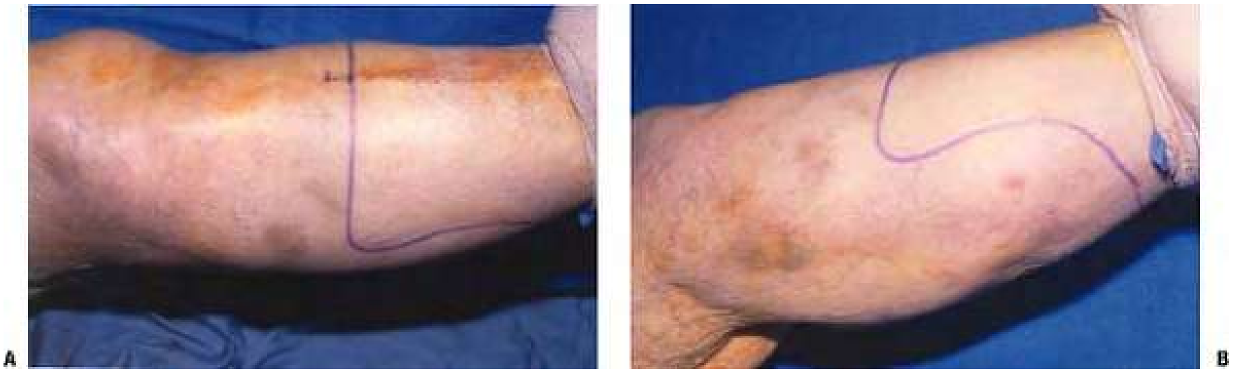

FIGURE 31-2 A,B: The long posterior flap is planned.

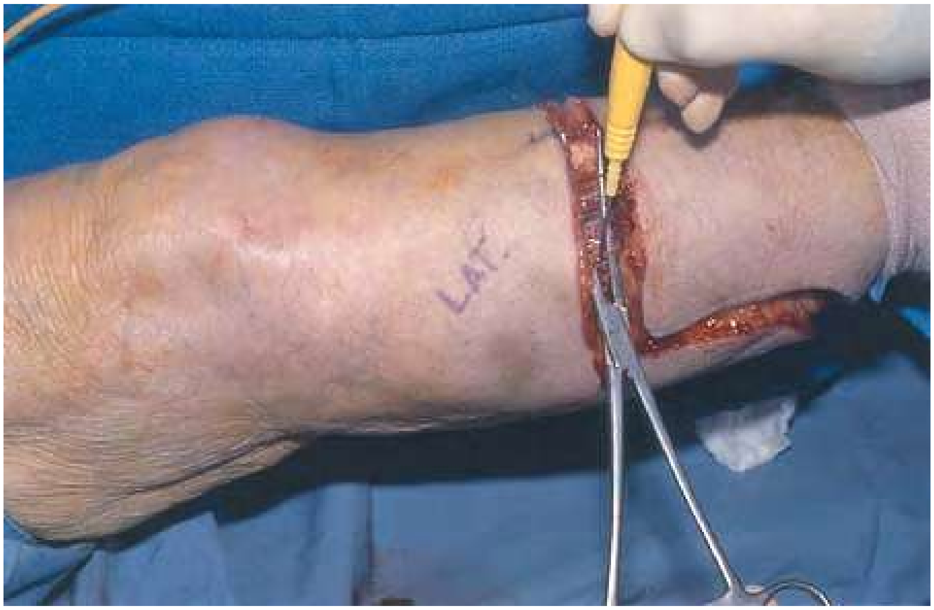

FIGURE 31-3 The anterior compartment muscle is dissected with cautery.

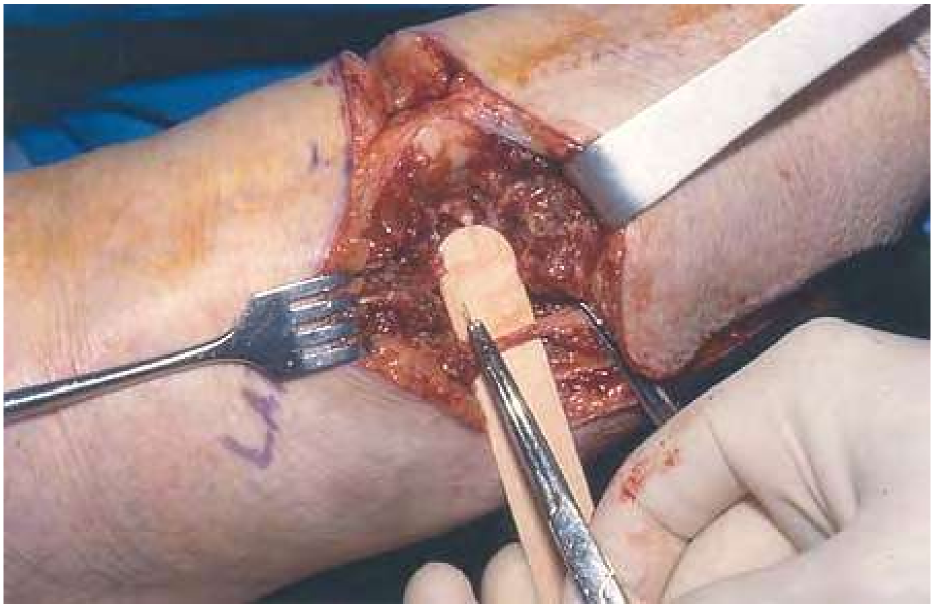

FIGURE 31-4 The deep peroneal nerve is identified and sharply transected.

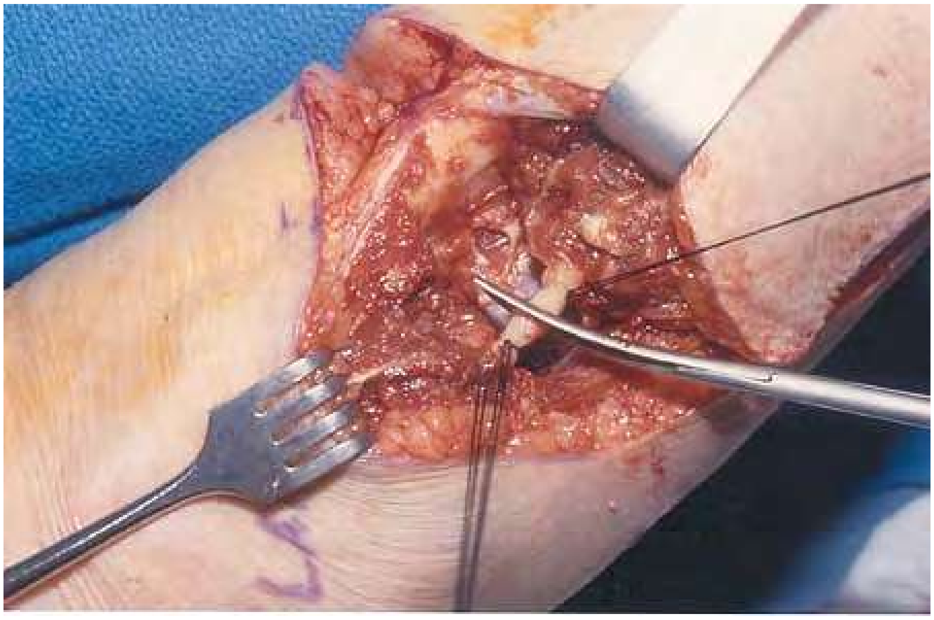

FIGURE 31-5 The anterior tibial artery and vein are ligated.

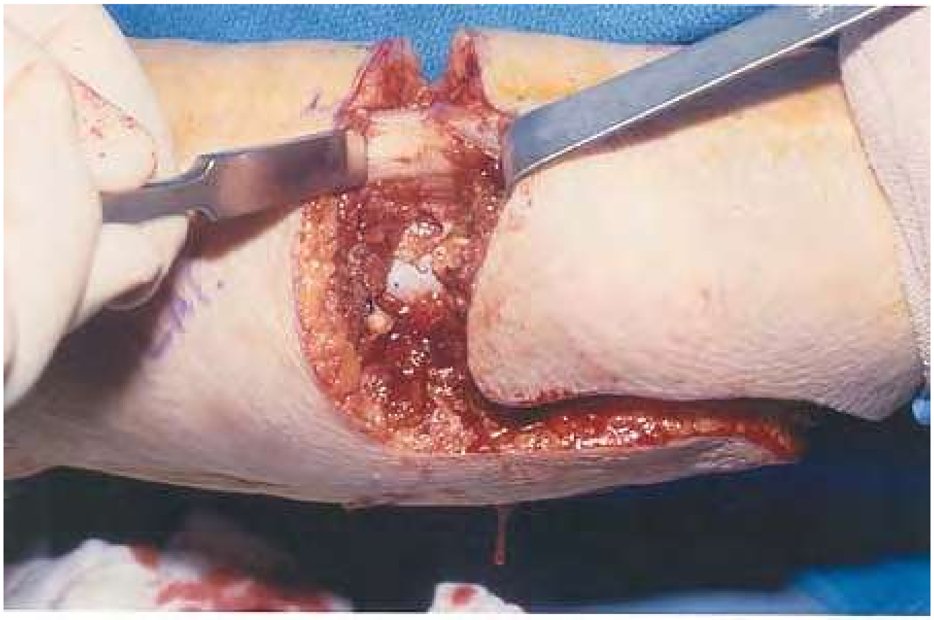

FIGURE 31-6 The periosteum is reflected off the tibia.

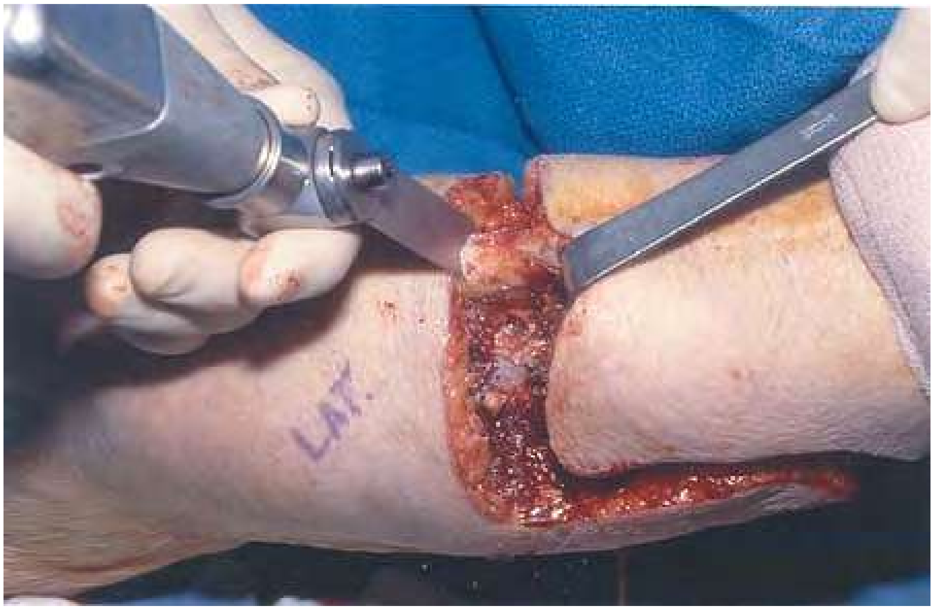

FIGURE 31-7 The tibia is osteotomized 1 cm proximal to skin incision.

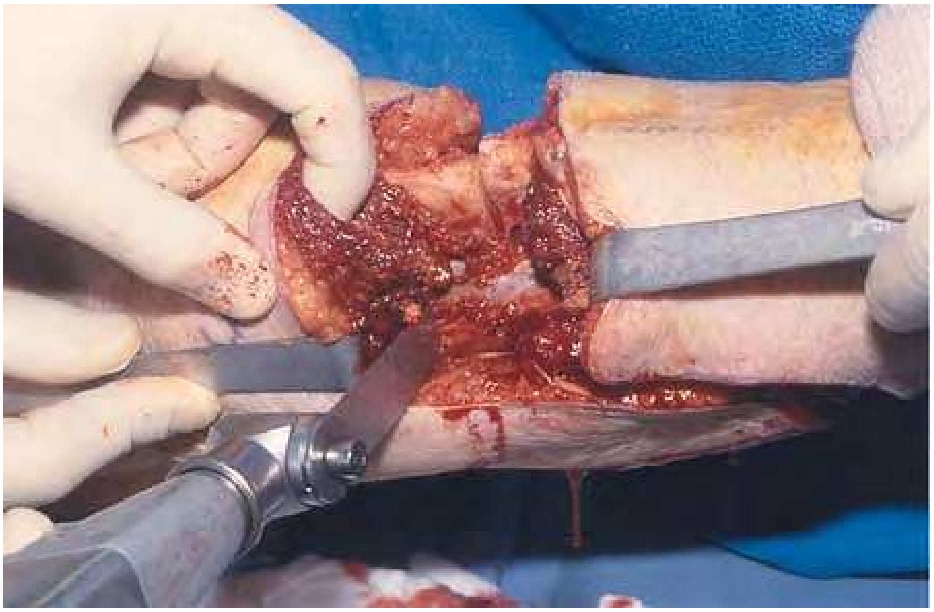

FIGURE 31-8 The fibula is identified and osteotomized 1 cm proximal to tibia osteotomy.

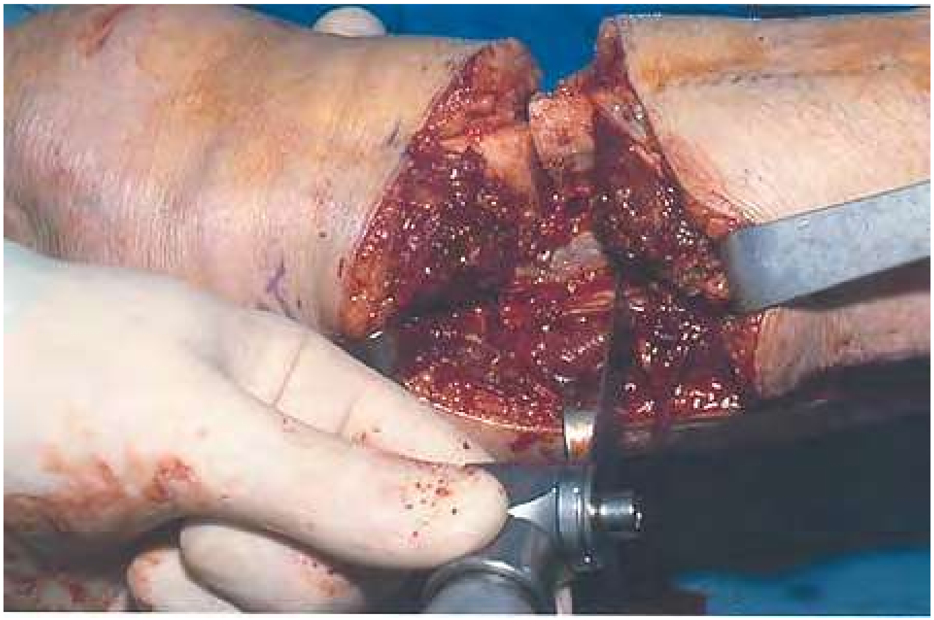

FIGURE 31-9 A second fibula osteotomy is performed 2 to 3 cm distal.

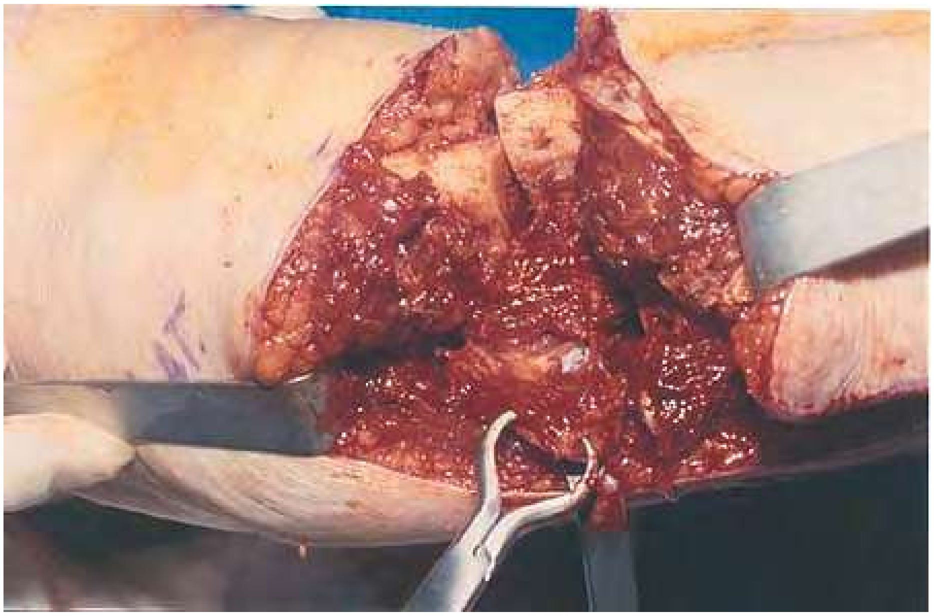

FIGURE 31-10 The segment of the fibula is removed.

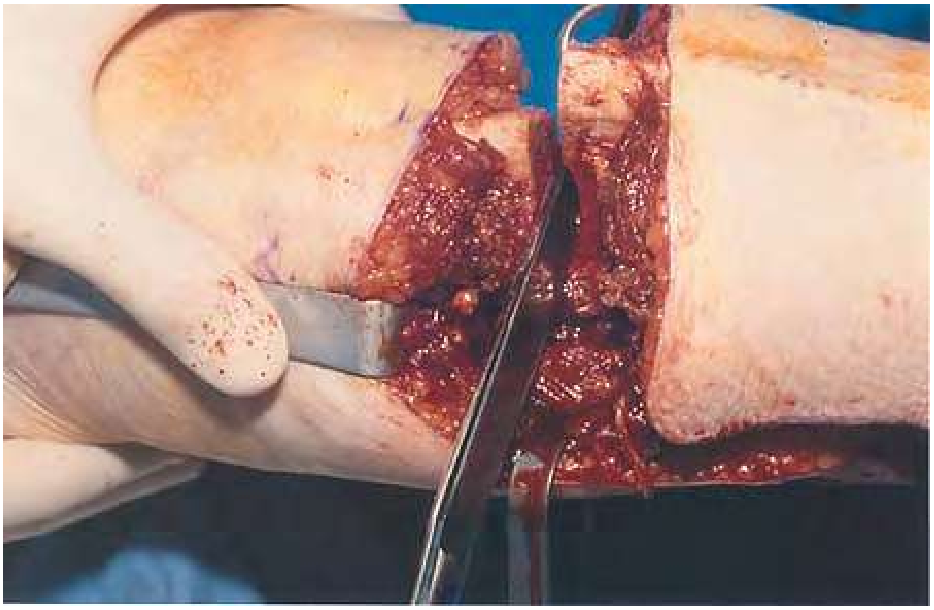

FIGURE 31-11 The amputation knife is used to sharply dissect the posterior compartment muscle.

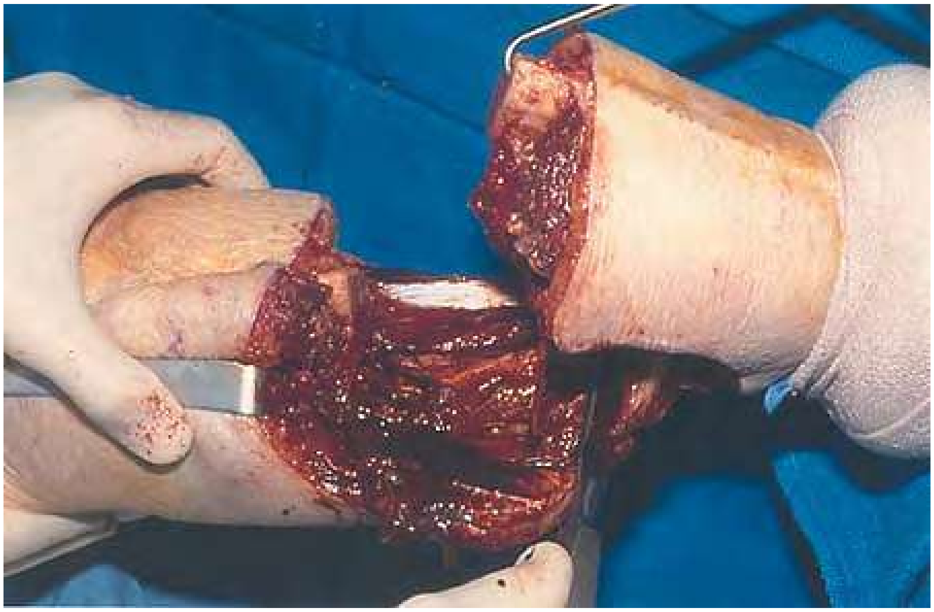

FIGURE 31-12 The posterior compartment muscle is beveled to minimize bulk.

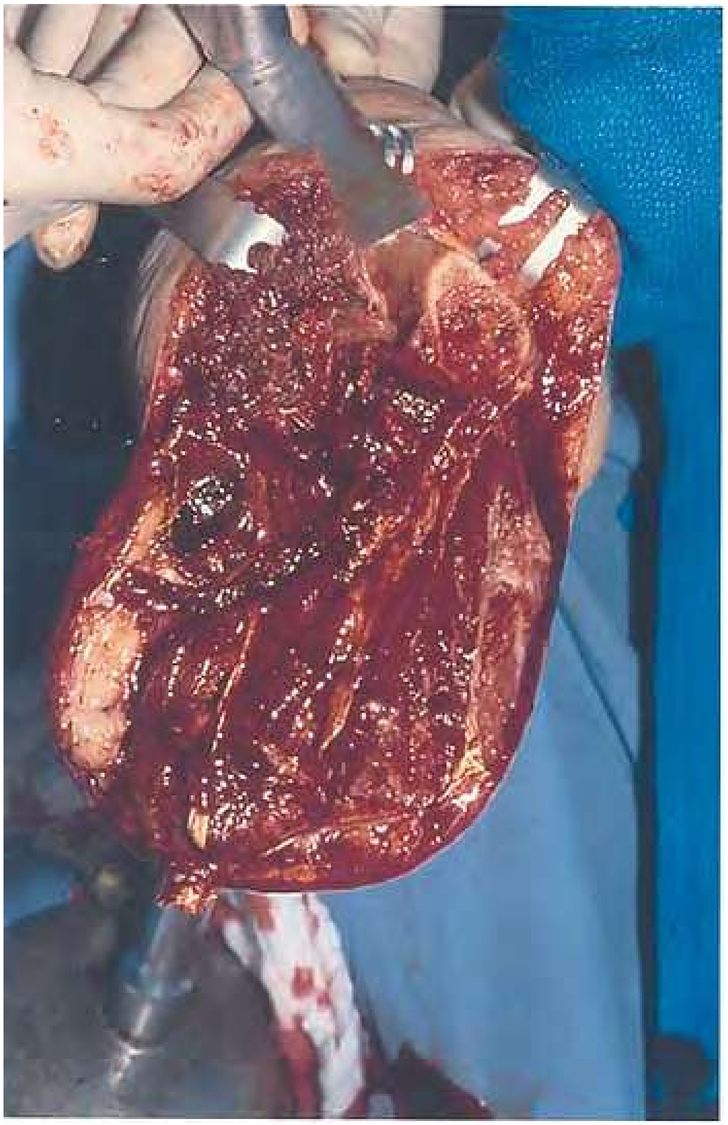

FIGURE 31-13 The saw is used to bevel the anterior tibia.

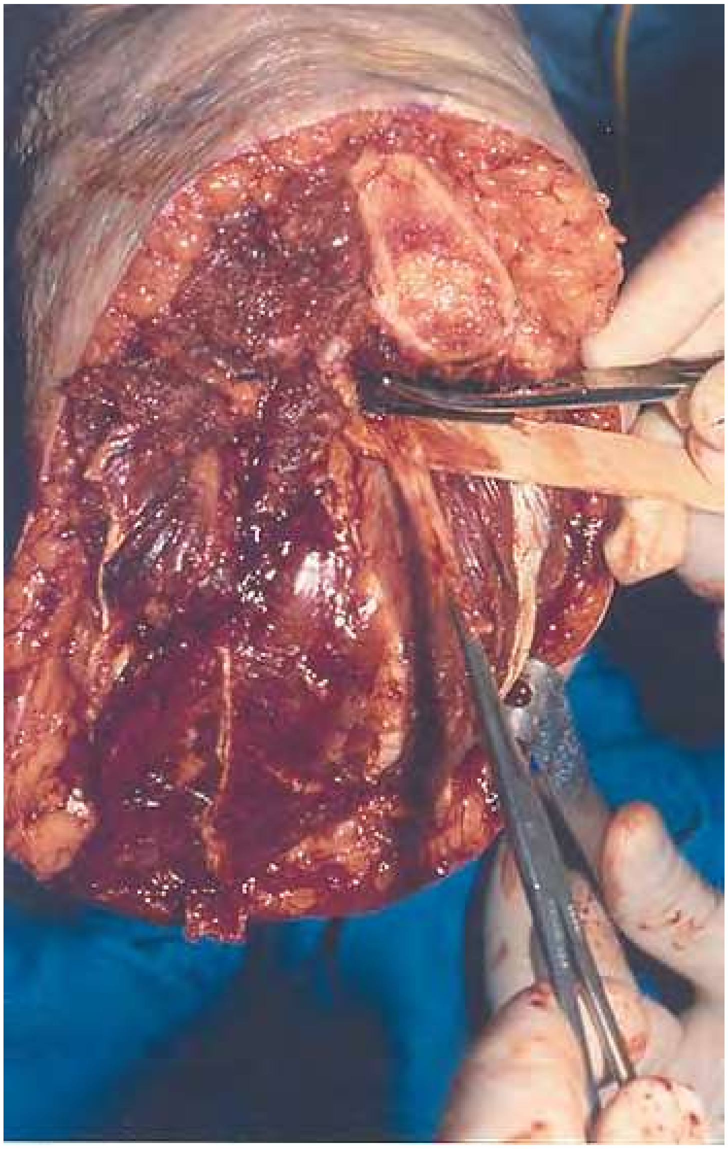

FIGURE 31-14 The tibial nerve is identified and sharply removed.

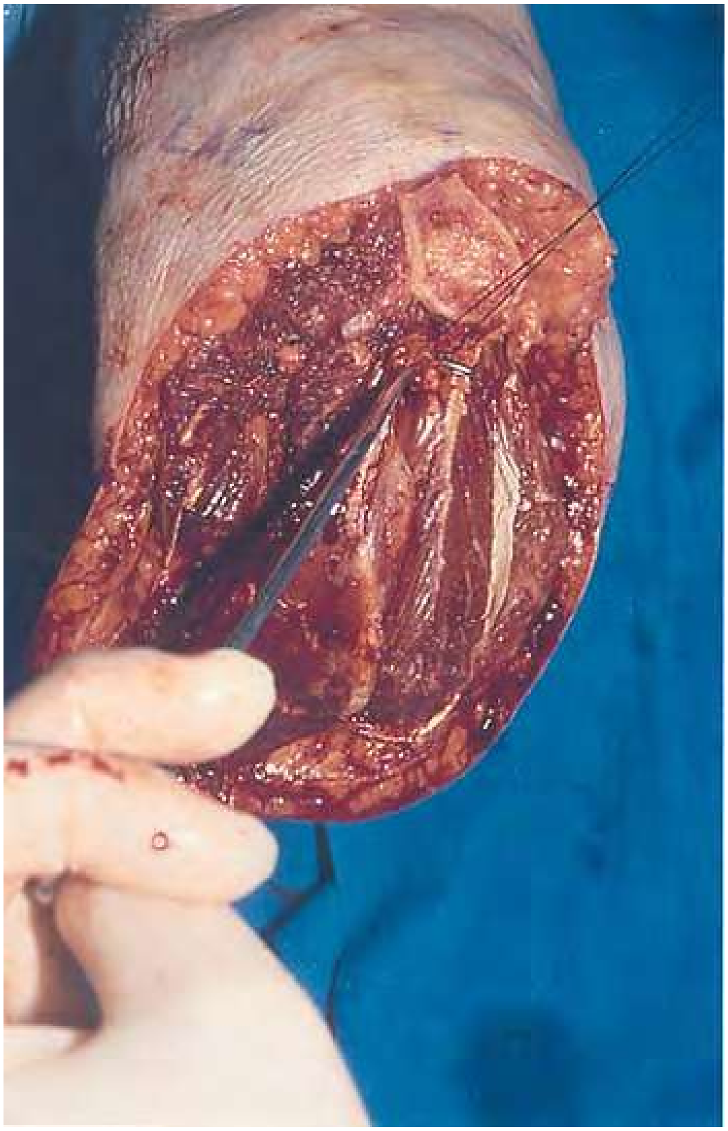

FIGURE 31-15 The tibial artery and vein are ligated.

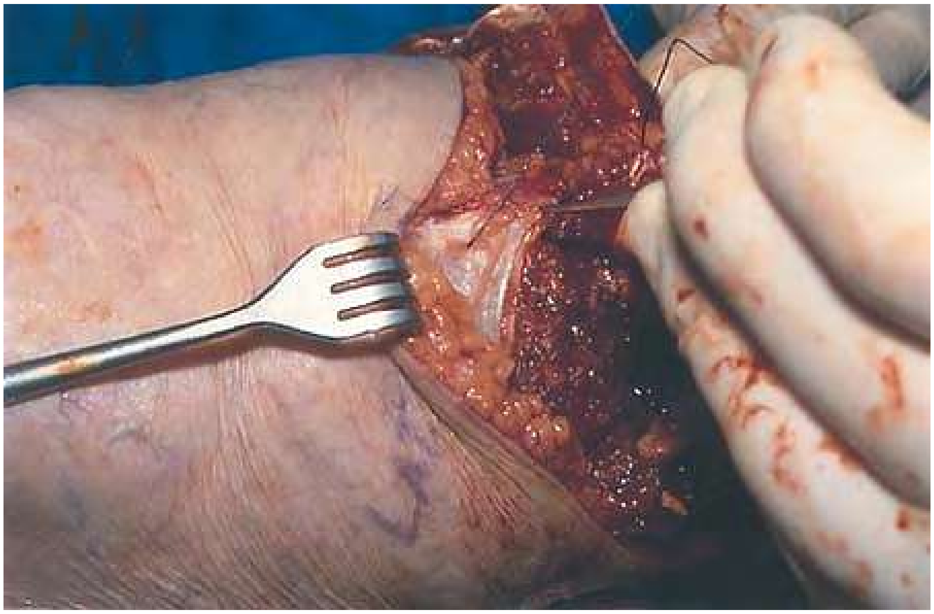

FIGURE 31-16 The fascia is closed with interrupted suture.

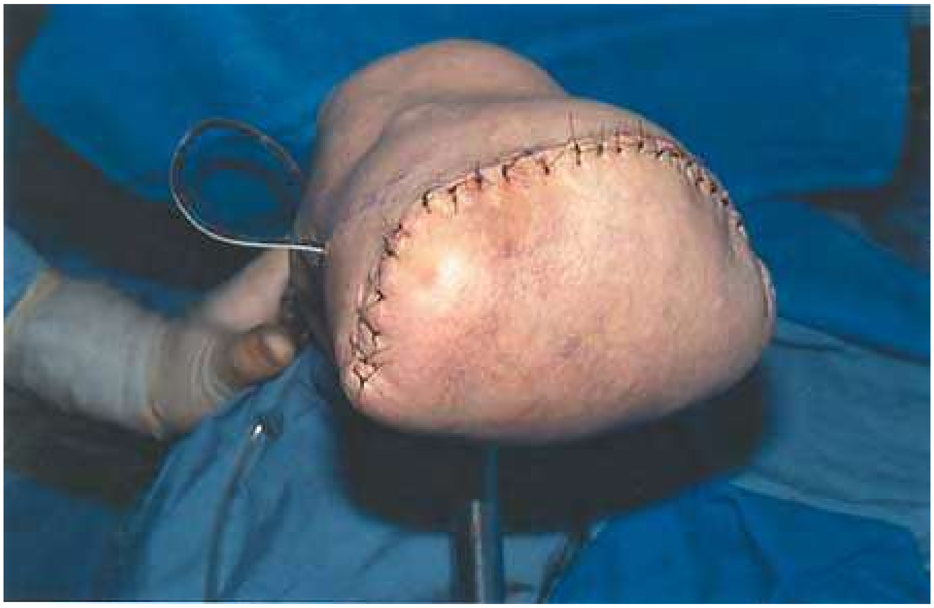

FIGURE 31-17 The skin is closed with interrupted vertical mattress suture.

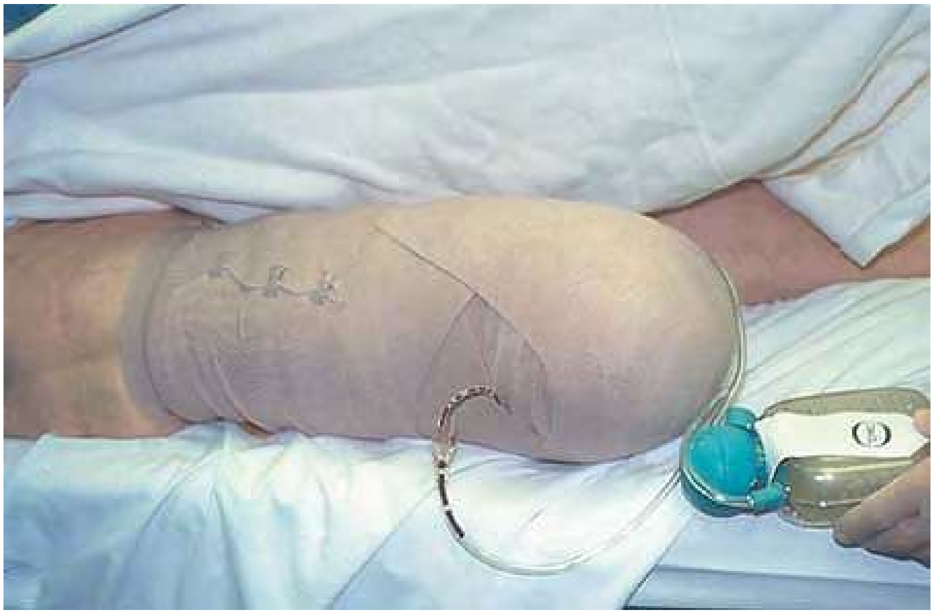

FIGURE 31-18 A compressive bulky dressing with plaster is applied with the knee in full extension.

For non-ambulatory patients with ischemic disease, above knee amputations are performed with equal anterior and posterior flaps, and myodesis is not performed to prevent further vascular compromise. For patients with adequate vascular supply and the potential to ambulate, a myodesis is performed. Above knee amputation is performed with the patient in a supine position. Skin flaps are marked with a long medial flap and a shorter lateral flap (Fig. 31-19). A sterile tourniquet can be used at the discretion of the surgeon. Dissection is carried through the skin and subcutaneous tissues and down to the muscle. The muscles are then identified. The quadriceps is detached proximal to the patella, retaining some of its tendinous portion (Fig. 31-20). The vastus medialis is reflected off of the intermuscular septum, and the adductor magnus (Fig. 31-21) is detached from the adductor tubercle by sharp dissection and reflected medially, exposing the femoral shaft. The vessels are identified (Fig. 31-22) at the level of Hunter's canal, and the artery and vein are ligated. The femur is then exposed proximally 12 to 14 cm above the condylar level and is cut with an oscillating saw approximately 10 to 12 cm above the joint line. The remaining edges are smoothed with a saw or rasp (Fig. 31-23). Small drill holes through the remaining distal femoral cortex are made with a 2.5-mm drill (Fig. 31-24). The adductor magnus tendon is then sutured with nonabsorbable suture to the lateral femur through the drill holes (Fig. 31-25). The femur is held in maximum adduction while this is being sutured to the bone as a myodesis (Fig. 31-26). The quadriceps is then brought over the bone and anchored to the posterior aspect of the femur through the drill holes (Fig. 31-27). The hip is in extension when this is done to try to minimize hip flexion. The fascia lata is then sutured to the medial fascia. Subcutaneous tissue is closed with monofilament suture, and the skin is closed with nylon suture (Fig. 31-28). Sterile dressing is applied, as well as a compressive wrap to minimize swelling.

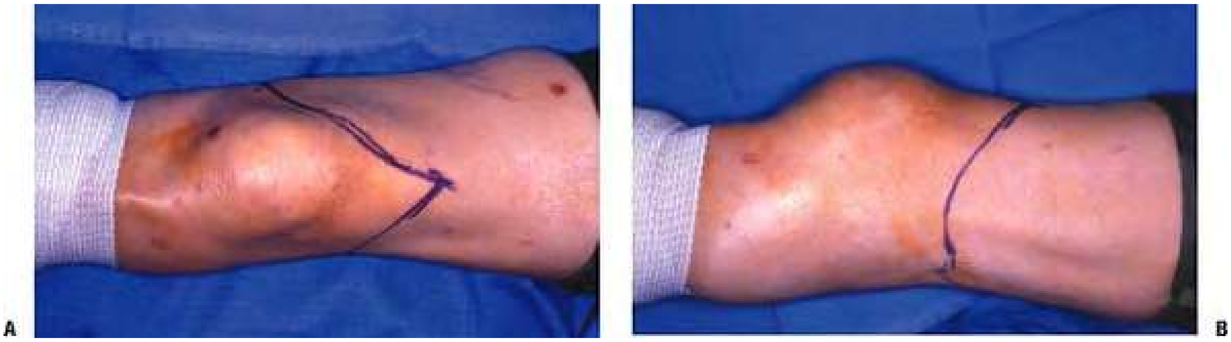

FIGURE 31-19 A,B: Skin flaps are marked with a long medial and shorter lateral flap.

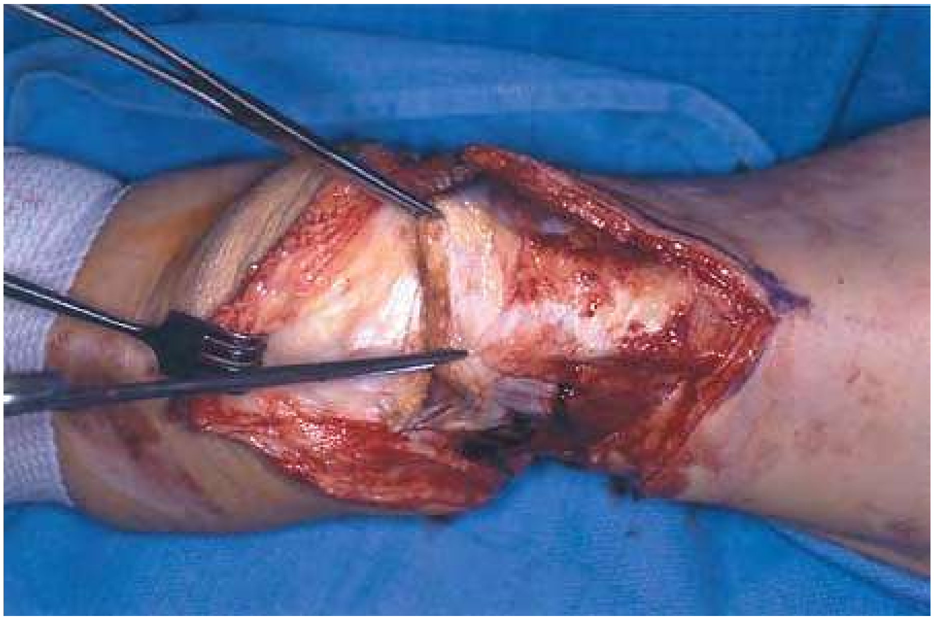

FIGURE 31-20 The quadriceps muscle is detached proximally to the patella, preserving some of the tendinous insertion.

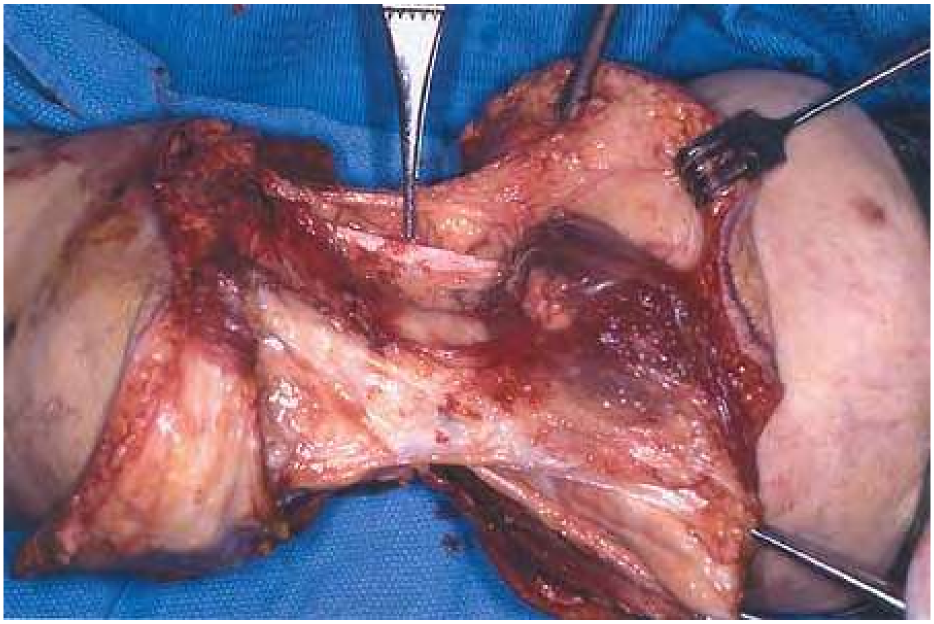

FIGURE 31-21 The adductor magnus is identified and reflected off the femur.

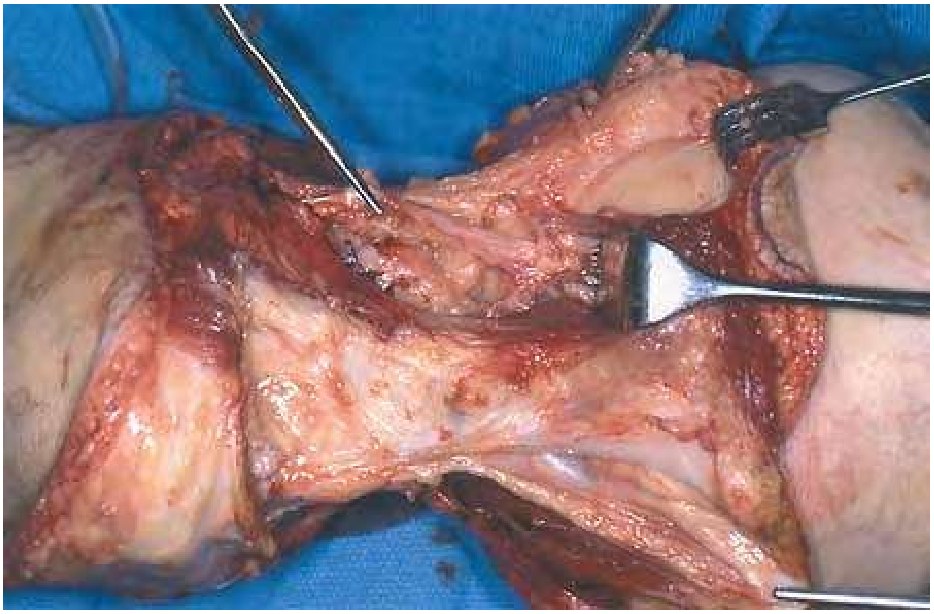

FIGURE 31-22 The femoral vessels at the level of Hunter's canal are identified and ligated.

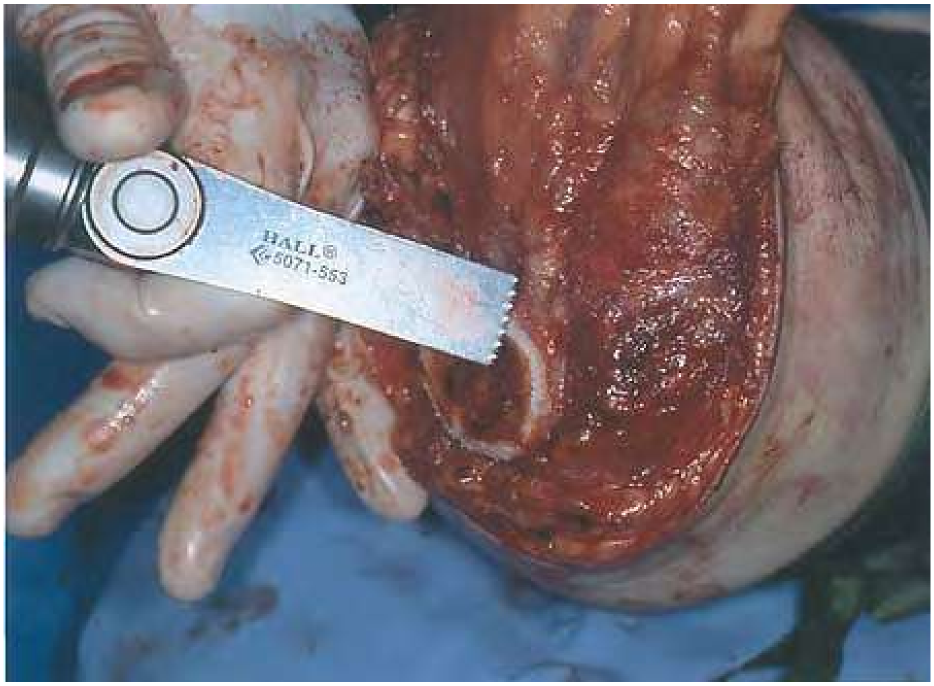

FIGURE 31-23 The femur is osteotomized and the end is rasped until smooth.

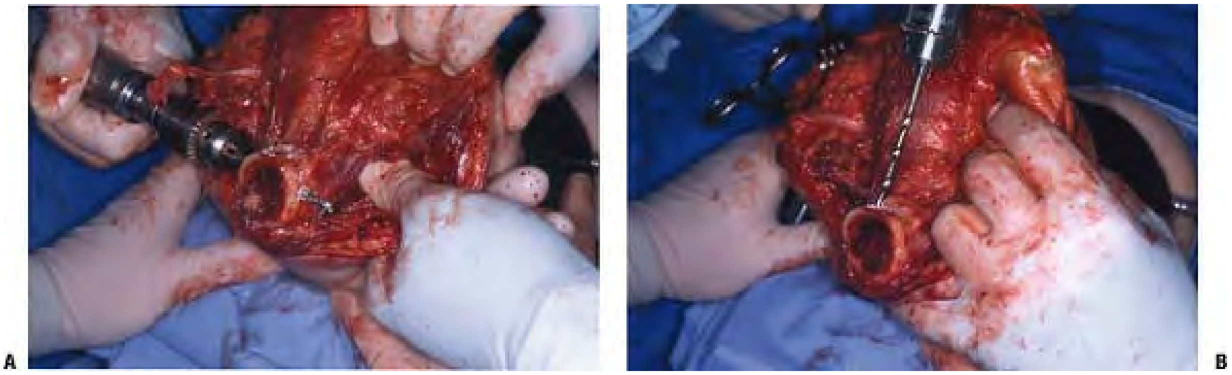

FIGURE 31-24 A,B: Four drill holes are made in the femoral stump for the myodesis.

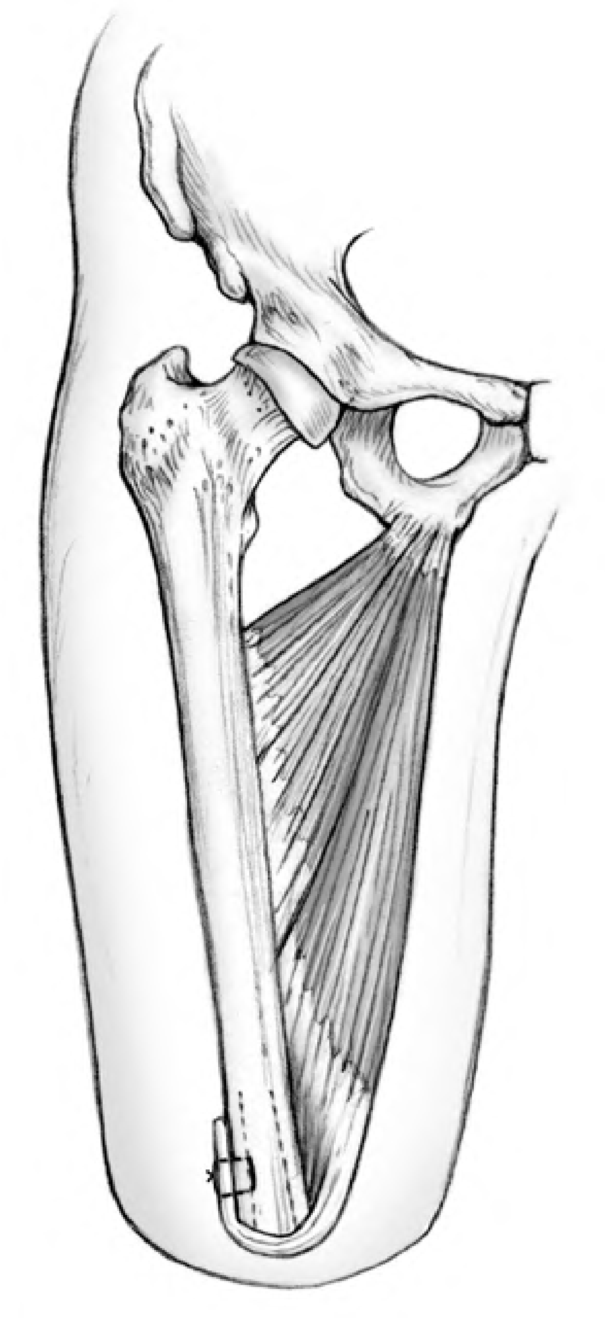

FIGURE 31-25 Attachment of adductor magnus to lateral femur. (Redrawn from Gottschalk F. Transfemoral amputations. In: Bowker JH, Michael JW, eds. Atlas of limb prosthetics: surgical, prosthetic, and rehabilitation principles. 2nd ed. St

Louis: Mosby; 1992.)

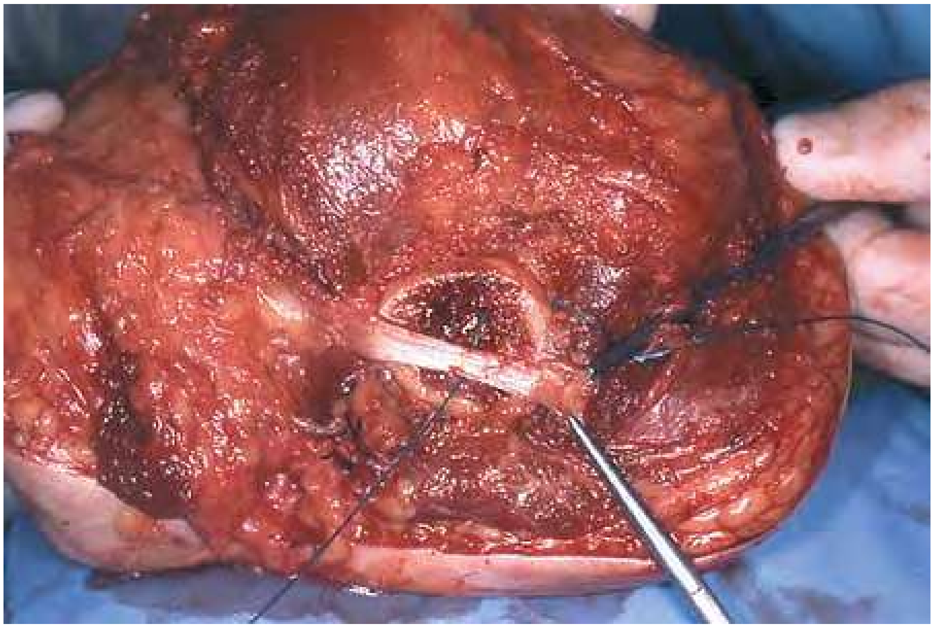

FIGURE 31-26 The adductor magnus tendon is sutured to the femur through the drill holes.

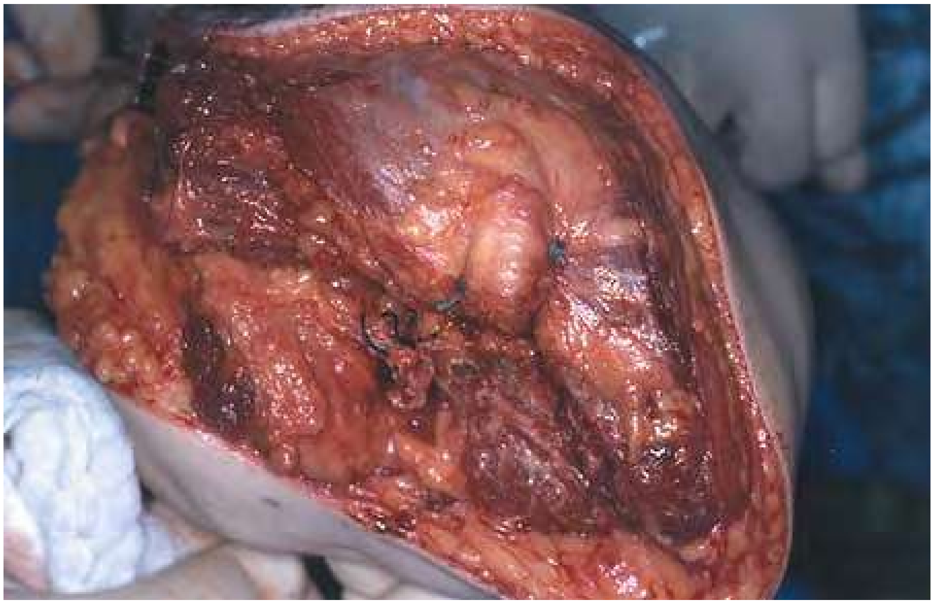

FIGURE 31-27 The quadriceps tendon is brought over the femoral stump and sutured through the drill holes.

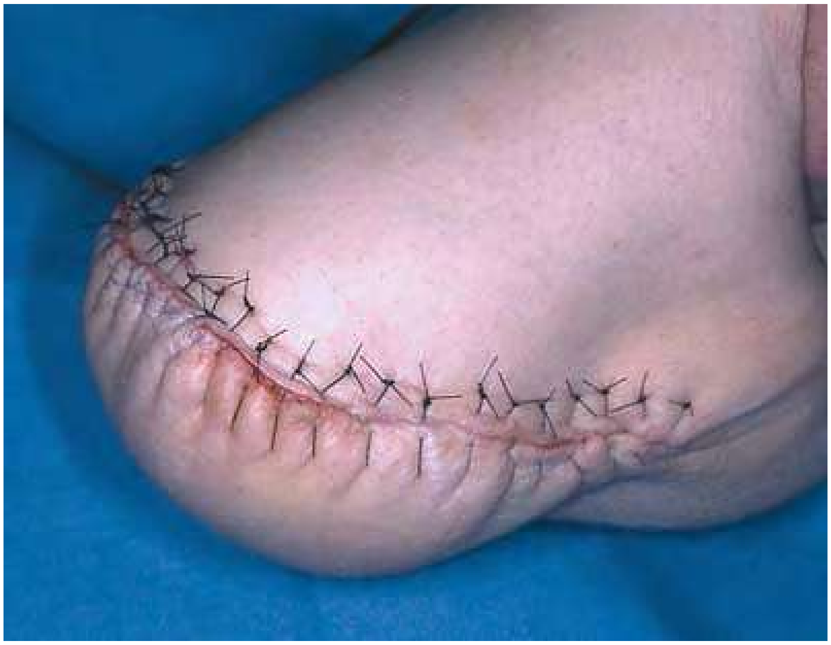

FIGURE 31-28 The skin is closed with interrupted vertical mattress suture.

Postoperative Management

For both a below knee and an above knee amputation, a compressive dressing is used to minimize swelling. The patient is kept in bed for 24 hours. The suction drain is kept in for at least 24 hours or until the drainage has decreased to minimal output. For a below knee amputation, a compressive Robert Jones dressing is removed approximately 48 hours after the surgery. If the swelling is well controlled and the wounds are in good condition, a pilon cast can be applied with a temporary prosthetic foot. The patient is restricted to minimal weight bearing but is allowed to be up and ambulating with gait aids. The pilon cast is removed at 2 weeks after surgery and another one can be placed. At 4 weeks after surgery, the pilon cast is removed completely. Sutures are removed at 2 to 4 weeks depending on the healing of the wound. A stump protector is used between 4 to 6 weeks. Shrinker socks are used to shrink the stump, and once the stump volumes have stabilized, a permanent prosthesis can be fitted.

Above knee amputations are initially treated with a compressive dressing and a stump protector. At approximately 2 to 3 weeks after surgery, if the wounds are healed, the sutures are removed. These patients also use shrinker socks as well as a protector. Once their stump volume has stabilized, they can be fitted for their prosthesis.

Complications/Results

The most common complications of amputation surgery are soft tissue or wound healing problems. These can lead to superficial infections that can ultimately cause deep infection, and possibly progress to osteomyelitis. Such complications may result in more proximal amputations.

The treatment of wound complication initially begins with dressing changes, and frequently these patients will heal with local wound care. Occasionally, a vacuum-assisted closure device can be helpful in increasing the granulation tissue to improve the healing in a more timely fashion. Long-term complications include difficulty with

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part VI - Soft Tissue Management Around the Foot and Ankle > Chapter 32 - Soft Tissue Management of Ankle Fractures and Use of the Gracilis Muscle Flap

Chapter 32

Soft Tissue Management of Ankle Fractures and Use of the Gracilis Muscle Flap

S. Andrew Sems

Matthew DeOrio

Steven L. Moran

Soft tissue management of injuries around the distal tibia and ankle region will often dictate and drive the timing and methods of definitive fixation. Injuries in this region vary from low energy ankle fractures to high energy tibial pilon fractures. The soft tissue injuries often reflect the amount of energy that was involved in creating the bony injury. Regardless of the radiographic appearance of the fractures, certain initial treatment principles hold true for both low and high energy injuries. Early fracture reduction to restore limb alignment, rotation, and appropriate limb length combined with fracture immobilization will allow sooner resolution of soft tissue swelling and impairment that might otherwise prevent or delay early internal fixation.

Ankle Fractures

Indications/Contraindications

Operative treatment is generally recommended for all unstable ankle fractures in which the talus is unable to be maintained in a position beneath the tibial plafond by closed reduction. Nonoperative management is appropriate for stable ankle fractures in which there is no lateral subluxation of the talus within the ankle mortise.

Comorbidities

Open reduction and internal fixation can generally be safely performed on most patients regardless of associated comorbidities. In patients with severe peripheral vascular disease, preoperative transcutaneous pressure oximetry measurements should be obtained to assess the likelihood of the patient healing the surgical wounds. TcPO2s less than 30 mm Hg have a higher risk of wound failure and fracture nonunion. Involving Vascular Medicine to assist with maximizing lower extremity perfusion is reasonable if vascular disease is a concern. The implications of insulin dependent diabetes mellitus and peripheral neuropathy should also be considered when determining operative versus nonoperative treatment but are not contraindications to surgery. Postoperative soft tissue management should be modified in patients with high risks of developing wound complications. Longer periods of cast immobilization and protected weight bearing may be indicated for patients with these comorbidities.

Nonoperative management of fractures in the nonambulatory patient is reasonable provided the fracture does not cause deformity which would result in subsequent skin breakdown. Modified techniques of internal fixation in patients who are nonambulatory may be utilized in order to maintain the anatomy around the ankle joint to prevent soft tissue compromise.

Preoperative Planning

Plain radiographs should be obtained in all patients with unstable ankle fractures prior to surgical intervention. Anteroposterior (AP), mortise, and lateral views are necessary to assess the fracture orientation. Ankle fractures tend to include the medial malleolus, lateral malleolus, and occasionally, the posterior malleolus. Recognition of the three-dimensional plane in which each fracture occurs is important so that surgical incisions are placed appropriately and the correct internal fixation is selected. While computed tomography is utilized more often for pilon type tibial fractures, it may be used if there is not a clear understanding of the fracture pattern.

Surgical approach selection for treatment of ankle fractures depends on the location of the fractures and the quality of the surrounding tissues. Fractures of the medial malleolus are typically approached with an anteromedial incision overlying the medial malleolus. This incision can be adjusted anteriorly or posteriorly depending on the size and orientation of the medial malleolar fragment. Fractures which include both the anterior and posterior caliculus may require slight posterior adjustment of the incision while fractures that only involve the anterior caliculus may be approached through a more anteromedial incision. Fibular fractures are typically treated with a direct lateral approach, although a posterolateral approach is useful for fibula fractures which occur in the coronal plane or in cases of associated posterior malleolus fragments that may be addressed through this same incision.

The timing of surgery is dictated by the state of the soft tissue envelope. For ankle fractures, immediate fixation within the first 24 to 48 hours is feasible so long as the swelling will not compromise wound healing. Patients who present to the clinic or emergency department with ankle fractures within 48 hours from injury can often be treated with immediate open reduction and internal fixation prior to the onset of the maximal amount of soft tissue swelling. Surgery should be delayed when soft issue is so edematous that the surgeon is unable to create skin wrinkles by gently pinching the skin over both the medial and lateral malleolus. Open fractures require immediate surgical intervention for debridement of all contaminated soft tissues. Antibiotic cement bead pouches or Vacuum Assisted Closure (VAC) devices may be used until definitive fixation or final wound closure. External fixation may be necessary to restore length, alignment, and rotation of both the soft tissues and the bone prior to definitive internal fixation.

Ankle fractures should ideally be treated within 3 weeks of injury prior to early callus formation, which can create significant difficulty in fracture reduction. Patients who present in the first week following an ankle fracture and whose soft tissues are not amenable to immediate surgical fixation should be treated with a Robert Jones type compressive dressing with plaster immobilization. These patients should be encouraged to maintain strict elevation of their limb at all times. They are then seen back in

the clinic or preoperative area approximately 7 to 10 days after their injury for soft tissue evaluation prior to surgical intervention. This dressing should be taken down and the skin should be evaluated prior to any type of anesthetic administration. In patients who are admitted to the hospital, sequential compression boots can be applied to the foot prior to the application of the Robert Jones dressing. These pneumatic boots are usually well tolerated so long as they do not extend up past the ankle.

Surgery

Medial/Lateral Approaches

Patient Positioning

The patient is positioned supine on a radiolucent table with a thigh high pneumatic tourniquet and a small bump (towel) placed under the ipsilateral hip to prevent excessive external rotation of the limb. The bump placed underneath the hip area should be adjusted so that the foot points vertically when in its resting position. A foam block or bump of towels can be utilized underneath the leg and ankle area so as to elevate the operative limb from the nonoperative limb. This facilitates obtaining lateral radiographs without manipulation of the operative leg. The intraoperative fluoroscopy unit is brought in from the contralateral side, and it is positioned perpendicular to the long axis of the patient.

Prophylactic antibiotics are utilized for internal fixation of ankle fractures. They should be administered prior to inflation of the tourniquet. Once the tourniquet is inflated, it should be maintained as long as necessary, but not for more than 2 hours. If it is impossible to complete the procedure in the allotted 2 hours, the tourniquet should be deflated and hemostasis should be obtained, and the procedure should be completed without the aid of a tourniquet. If small vessel bleeding continues to be an issue, consider moving the operative bed into a Trendelenburg position and asking the anesthesiologist to decrease the systolic blood pressure within safe limits.

Technique

Anatomic Landmarksâ€”Medial Approach

Palpate the medial malleolus to define the location of the anterior aspect of the anterior caliculus, the posterior aspect of the posterior caliculus, as well as the fracture location. The leg should be palpated to determine the longitudinal axis of the tibia.

Incision and Surgical Approach

The incision is made along the midlateral axis of the distal tibia. Beyond the tip of the medial malleolus the incision may be slightly curved anteriorly (Fig. 32-1). Be aware that excessive curvature of the incision may create difficulties in placing internal fixation in the medial malleolar fragment. Dissection should be carried through the skin and subcutaneous and hemostasis obtained as this is performed (Fig. 32-2). Care should be taken to preserve and protect the saphenous vein and nerve throughout the case. In this location, the saphenous vein is often very subcutaneous and an aggressive skin incision may lacerate the vein. With the saphenous vein and nerve protected, the medial malleolar fragment can be retracted distally with sharp bone hooks to allow inspection of the ankle joint (Fig. 32-3). A 15 blade, used to elevate the periosteum along the fracture edges, can be utilized to ensure that the fracture is reduced in an anatomic position. Preservation of the soft tissue and minimal periosteal stripping is recommended. Following internal fixation, the subcutaneous tissue can be closed using a 2-0 absorbable suture, and the skin closed with a nonabsorbable monofilament suture (Fig. 32-4).

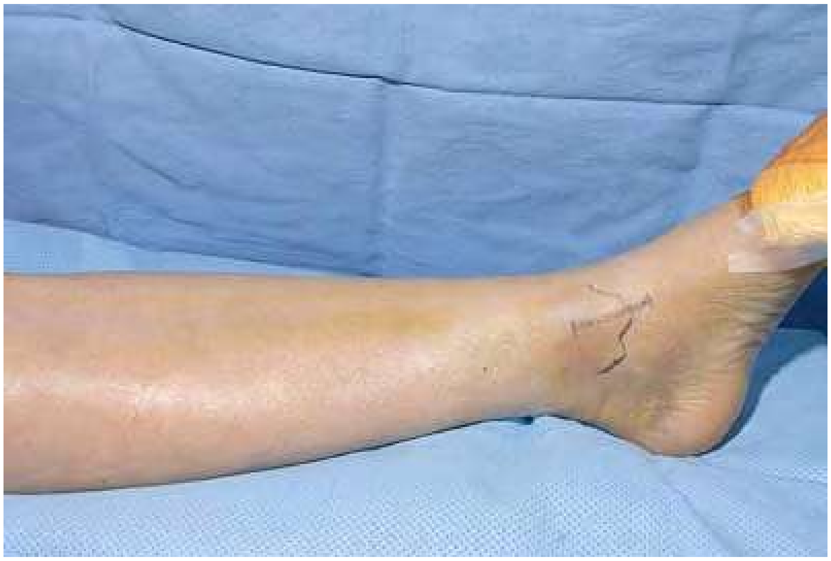

FIGURE 32-1 The incision is made along the mid-lateral axis of the distal tibia, curving anteriorly beyond the tip of the medial malleolus.

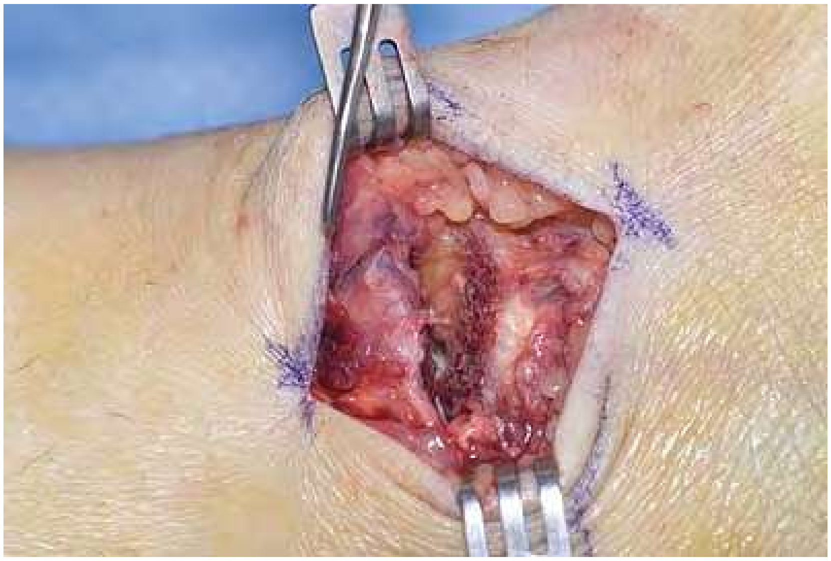

FIGURE 32-2 Exposure of the medial malleolar fracture line prior to elevation of the periosteum along the fracture edges.

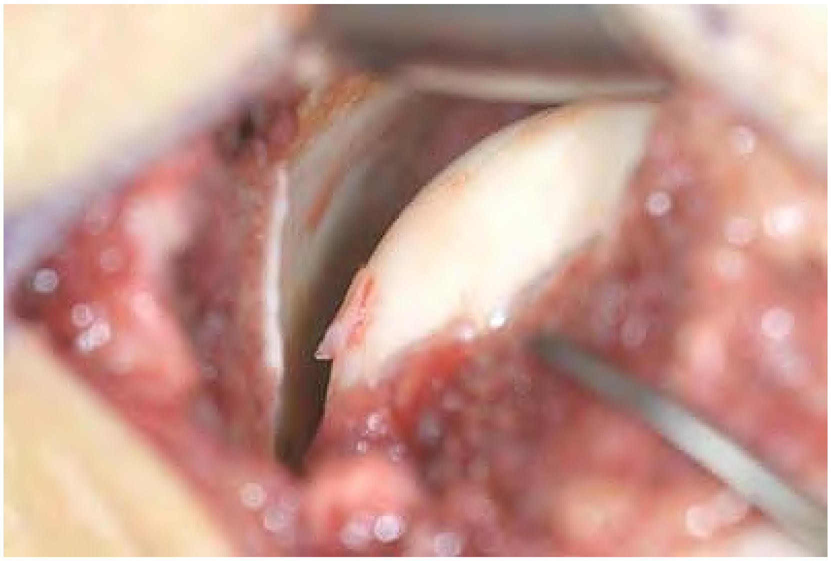

FIGURE 32-3 Retraction of the medial malleolus allows exposure of the talus and ankle joint.

Lateral Approachâ€”Lateral Malleolar Fixation

The direct lateral approach to the fibula allows placement of fixation along the lateral aspect of the fibula. This incision is positioned in line with the longitudinal axis of the fibula and may be curved anterior distally to allow access for reduction of a Chaput-Tillaux fragment (Fig. 32-5). If plate fixation along the posterior aspect of the fibula is desired, such as in a situation when an anti-glide plate is used, rather than committing excessive soft tissue stripping via a direct lateral approach, the incision can be adjusted posteriorly. The superficial peroneal nerve may cross the surgical approach in a subcutaneous location; therefore the skin incision should go no deeper than the skin (Fig. 32-6). Dissection through the subcutaneous tissues should be performed using combinations of sharp and blunt dissection with care to identify and protect the superficial peroneal nerve if encountered. Periosteal elevation at the fracture edges with a 15 blade is recommended to ensure an anatomic reduction of the fibular fracture.

For long spiral oblique fractures in which lag screw only fixation of the fibula is to be used, this incision can be altered with an apex anterior curve along the midportion of the incision. This â€œwaveâ€• in the incision will allow anterior retraction of the soft tissues and appropriate directional placement of lag screws across the fracture site. The same concerns remain with regards to protection of the superficial peroneal nerve when this modification is performed.

Following internal fixation, closure is performed with interrupted 2-0 absorbable sutures in the subcutaneous tissue followed by nonabsorbable monofilament sutures in the skin (Fig. 32-7).

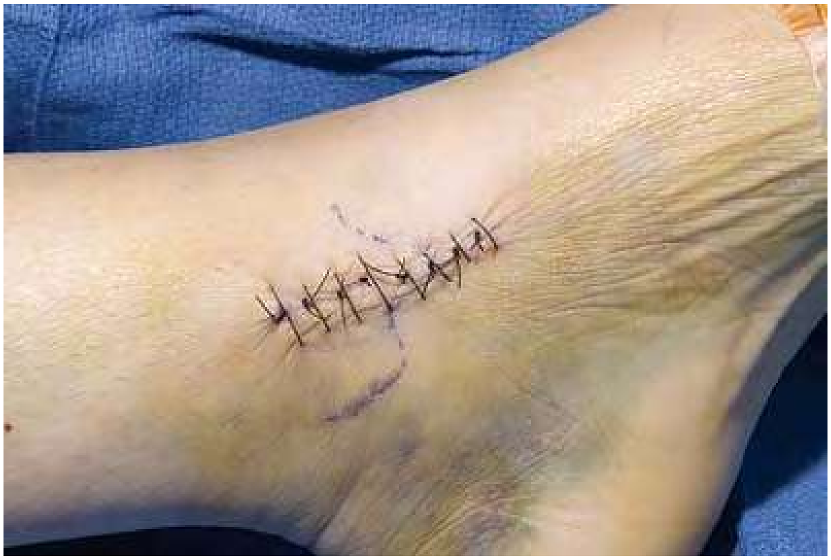

FIGURE 32-4 Skin closure is performed using a nonabsorbable monofilament suture.

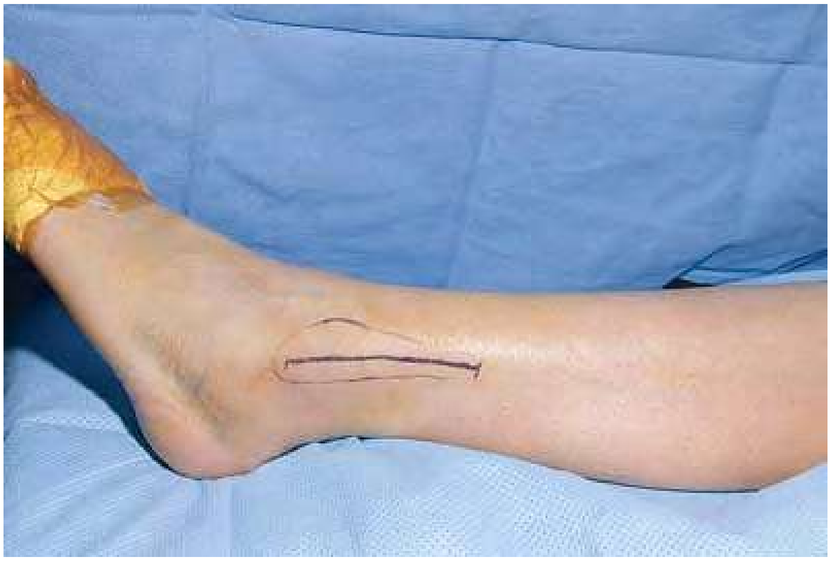

FIGURE 32-5 The incision is positioned in line with the longitudinal axis of the fibula, curving anteriorly at the distal end to allow access for reduction of a Chaput-Tillaux fragment.

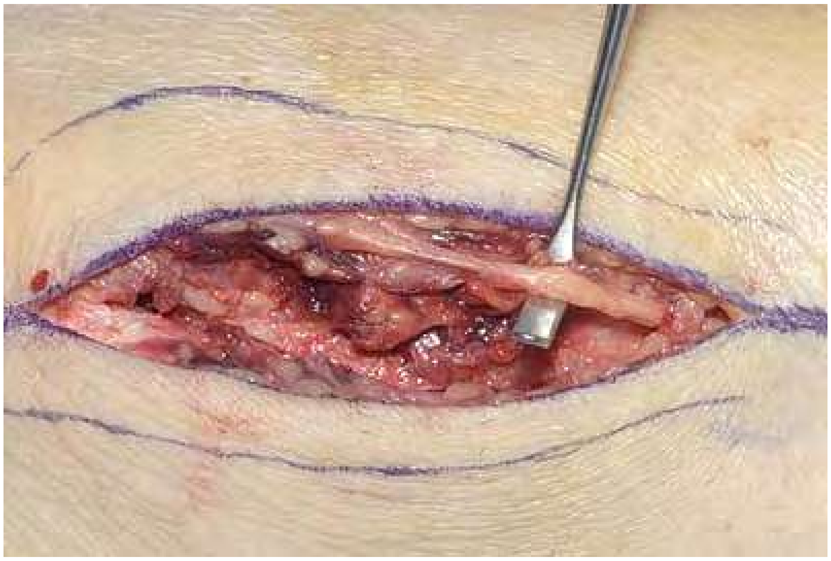

FIGURE 32-6 The superficial peroneal nerve may cross the incision, and care should be taken to identify and protect this structure.

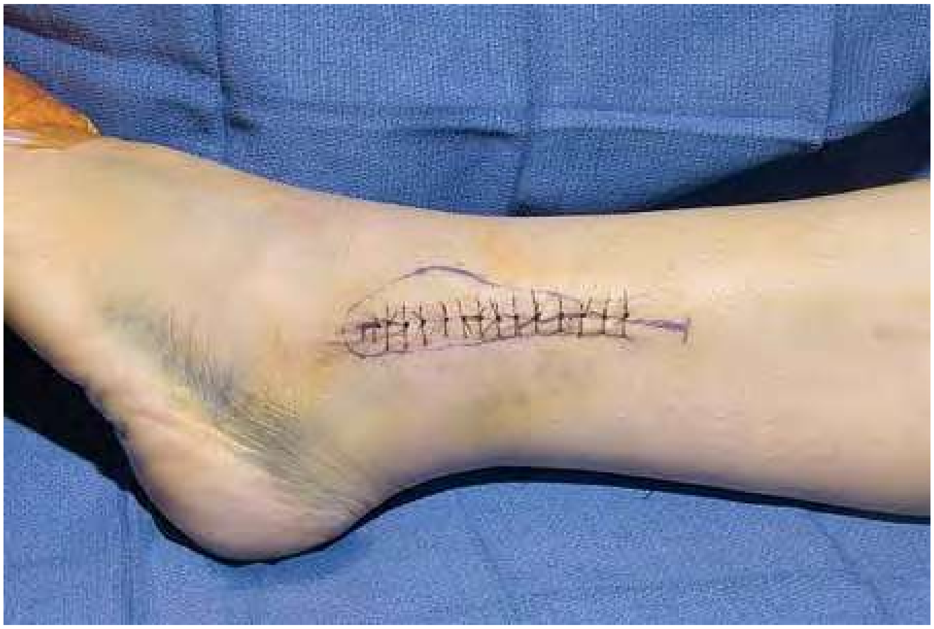

FIGURE 32-7 Skin closure is performed using a nonabsorbable monofilament suture.

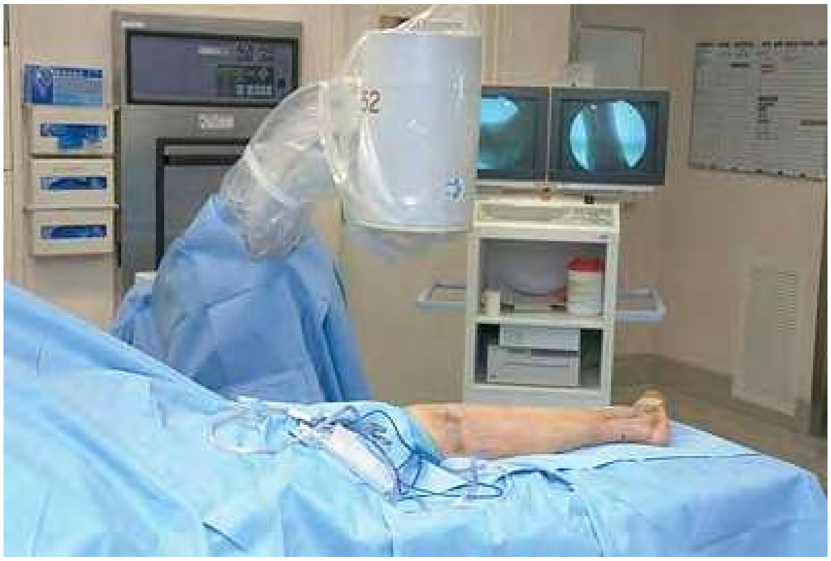

FIGURE 32-8 Intraoperative fluoroscopy can be utilized in either the lateral or prone position, and image quality should be confirmed prior to incision.

Posterolateral Approachâ€”Posterior Malleolus Fractures

Patient Positioning

Fractures which require fixation of the posterior malleolus can be approached through a posterolateral incision. Controversy exists regarding the size of the fragment, but internal fixation is typically recommended for fractures which include 20% to 30% of the articular surface. This approach requires the patient to be in either a lateral or prone position (Fig. 32-8). The prone position allows direct visualization and better stabilization of the limb during the approach and is therefore preferred. The patient should be placed on a well-padded radiolucent table with care to be taken to prevent hyperextension of the shoulders and neck. If the arms are placed in an abducted position, care should be taken to avoid direct compression of the ulnar nerve at the cubital tunnel. A thigh high tourniquet is utilized and can be placed prior to prone positioning. Once the patient is in the prone position, they are translated towards the foot of the bed so that the hindfoot is slightly hanging off the end of the bed to aid with reduction and fixation. The C-arm fluoroscopy unit is placed perpendicular to the longitudinal axis of the patient and intraoperative films can be used on the patient's uninjured side to guarantee symmetry following reduction.

Technique

Palpation of the fibula and Achilles tendon is possible due to the subcutaneous location of these structures. The incision is made midway between the posterior border of the lateral malleolus and the lateral border of the Achilles tendon and will need to extend to the tip of the fibula distally and as far proximally as necessary to obtain visualization of the fibular fracture and the posterior malleolar fracture (Fig. 32-9). The short saphenous vein and sural nerve are located immediately behind the lateral malleolus and the incision is placed posterior and medial to these structures. The deep fascia of the leg is split in line with the incision and the peroneal retinaculum is incised to release the peroneus longus and brevis (Fig. 32-10). The peroneal tendons and muscles are retracted laterally and the flexor hallucis longus is elevated from its origin on the fibula and retracted medially (Fig. 32-11). This allows access to the posterior aspect of the tibia for reduction and fixation (Fig. 32-12). The peroneal tendons are retracted medial or lateral depending on the location of the fracture. For fibular fractures which require a more proximal exposure, dissection to the lateral aspect of the peroneal tendons and muscles is necessary.

Following internal fixation of the posterior malleolus fragment and/or lateral malleolar fracture through the posterolateral approach, the peroneal retinaculum is repaired utilizing a 2-0 absorbable suture (Fig. 32-13). If possible, the deep fascia of the leg is also reapproximated using the same absorbable suture. Subcutaneous tissues are closed with a 2-0 absorbable suture and the skin is closed in interrupted fashion using a 3-0 or 4-0 nylon suture (Fig. 32-14).

Postoperative Management

Following internal fixation, the patient is placed in a well-padded Robert Jones splint to provide further postoperative immobilization (Fig. 32-15). The splint is applied with the knee flexed to allow dorsiflexion of the ankle to a neutral position and avoidance of equinus positioning. This splint is maintained for the first 10 days to allow for postoperative swelling. The patient is then converted to either a short-leg cast or a fracture boot following this initial Robert Jones dressing. The sutures are removed 2 to 3 weeks from the date of surgery, however they should remain in place if there is any concern of delayed wound healing. Weight bearing restrictions are tailored to the individual fracture pattern and in patients with excellent bony quality and excellent healing potential weight bearing may be started at 6 weeks. In patients with multiple comorbidities, poor bone quality, or concerns for delayed fracture healing, weight bearing may be limited for up to 3 months post-operatively.

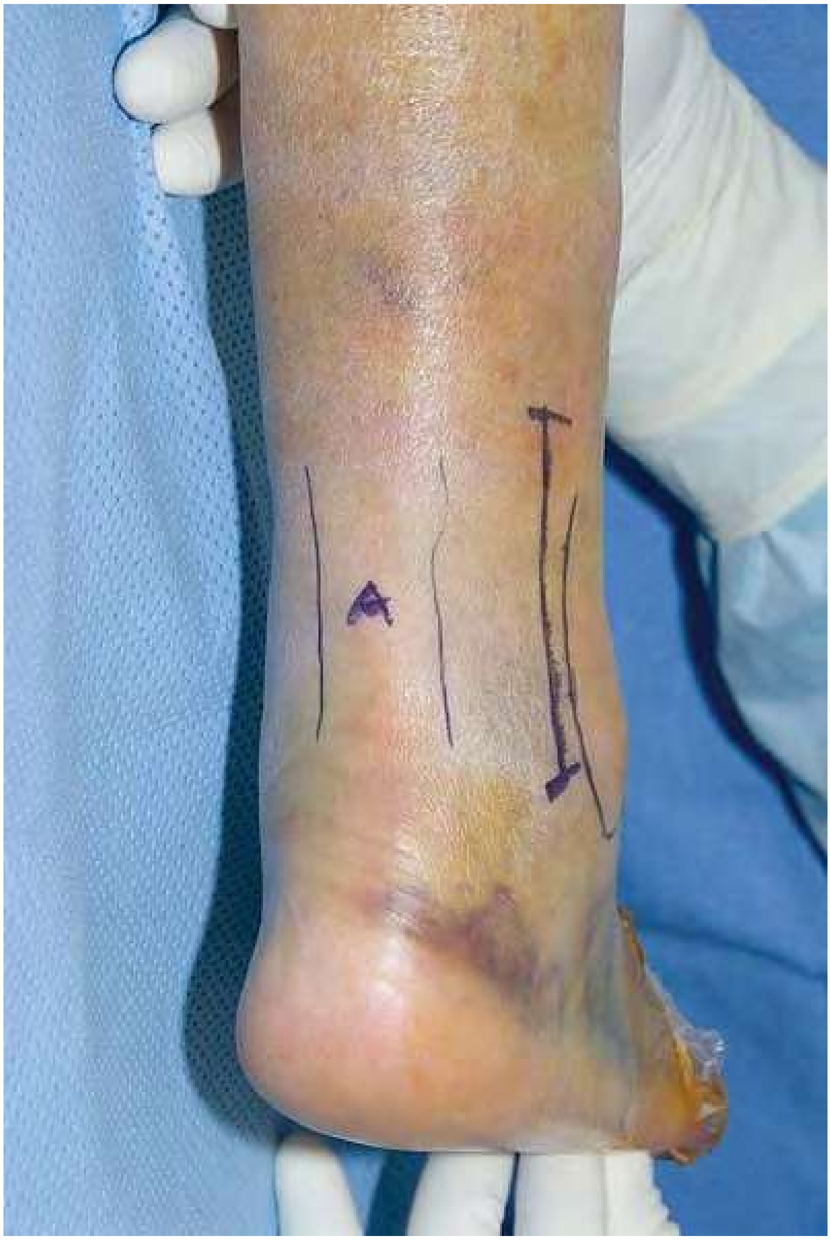

FIGURE 32-9 The incision is made midway between the posterior border of the lateral malleolus and the lateral border of the Achilles tendon.

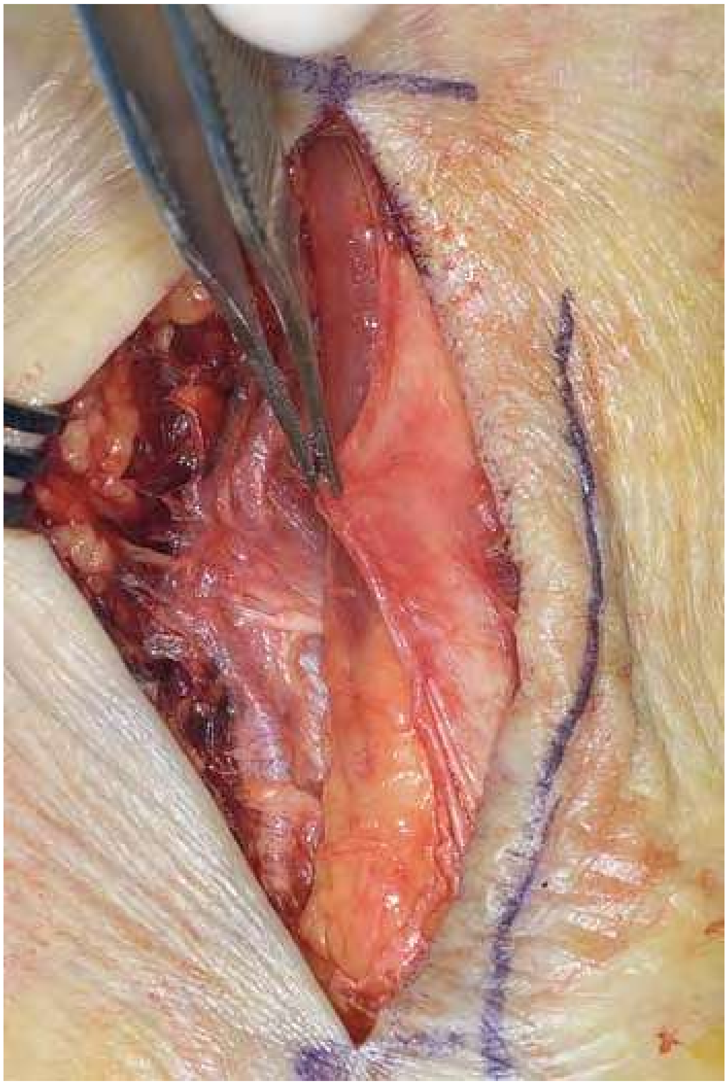

FIGURE 32-10 The deep fascia of the leg is split in line with the incision and the peroneal retinaculum is incised to release the peroneus longus and brevis.

Complications

Postoperative wound dehiscence

With appropriate soft tissue management and timing of internal fixation of ankle fractures, postoperative wound dehiscence should be a relatively infrequent complication. Due to the relatively subcutaneous location of the internal fixation, any wound complication involving full thickness skin dehiscence or necrosis should be considered to communicate with the hardware and appropriate aggressive surgical intervention involving debridement and dressing changes should be performed.

Patients with partial skin necrosis or minor wound dehiscence that is located away from the hardware, particularly on the medial side, can be treated with local wound care including wet-to-dry dressing changes. If this is noted prior to suture removal, the sutures should be kept in place during these dressing changes, particularly near the portion of the wound adjacent to the dehiscence. Surrounding cellulitis should be treated with appropriate antibiotics.

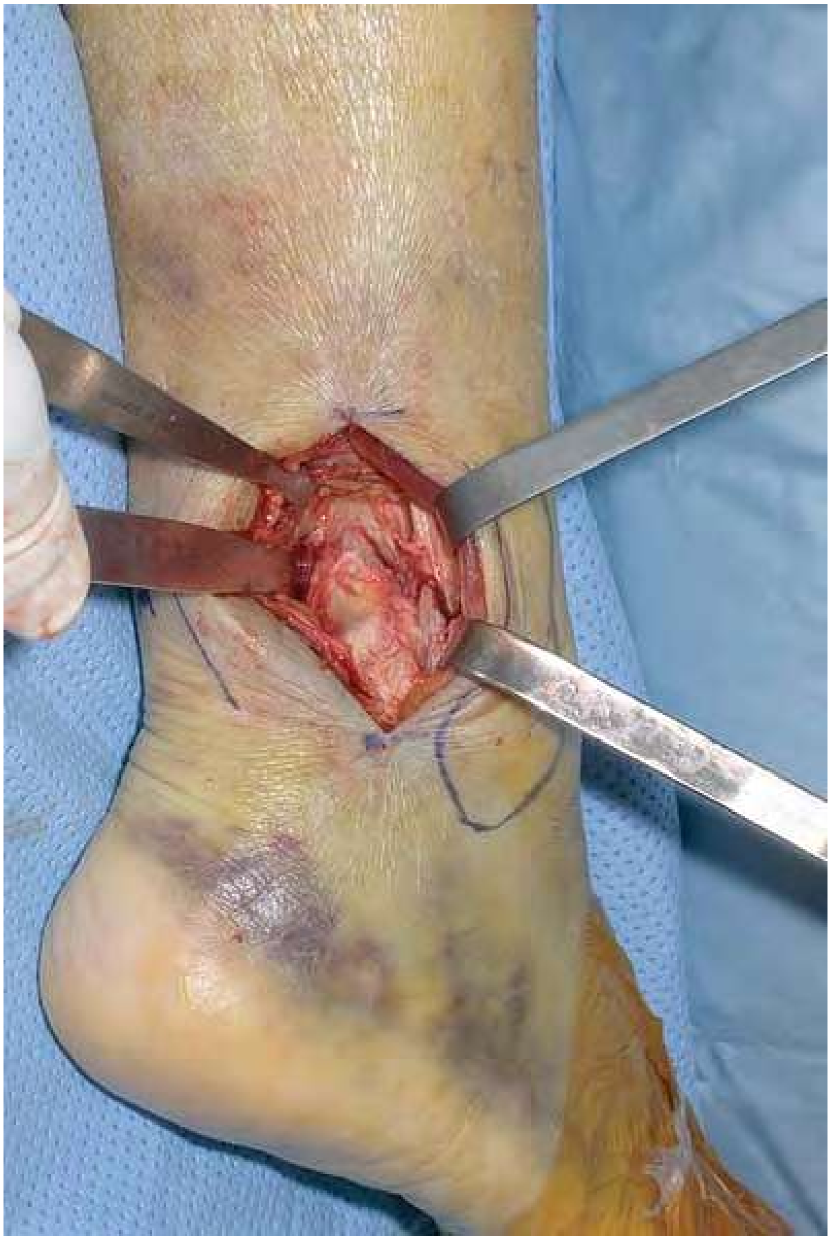

FIGURE 32-11 The peroneal tendons and muscles are retracted laterally and the flexor hallucis longus is elevated from its origin on the fibula and retracted medially.

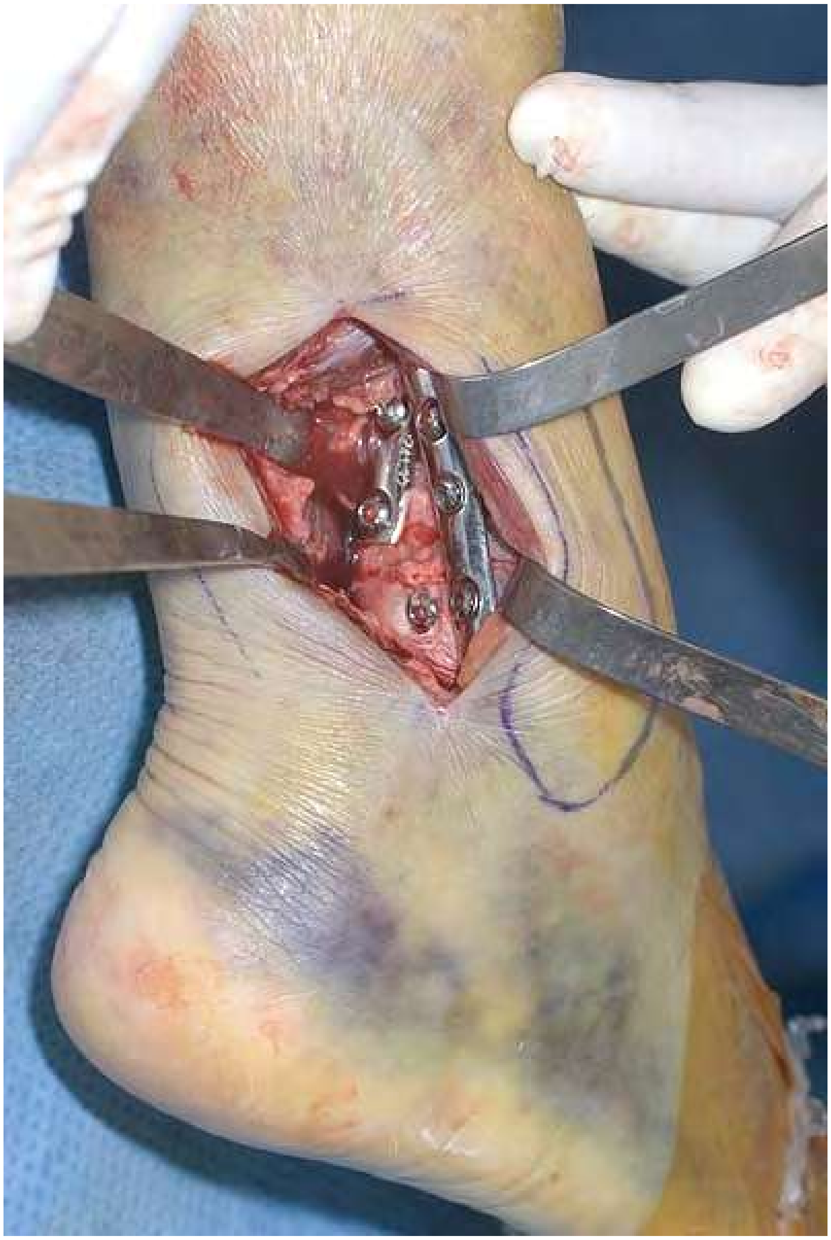

FIGURE 32-12 Following exposure, the reduction and fixation of the posterior malleolar fragment may be completed.

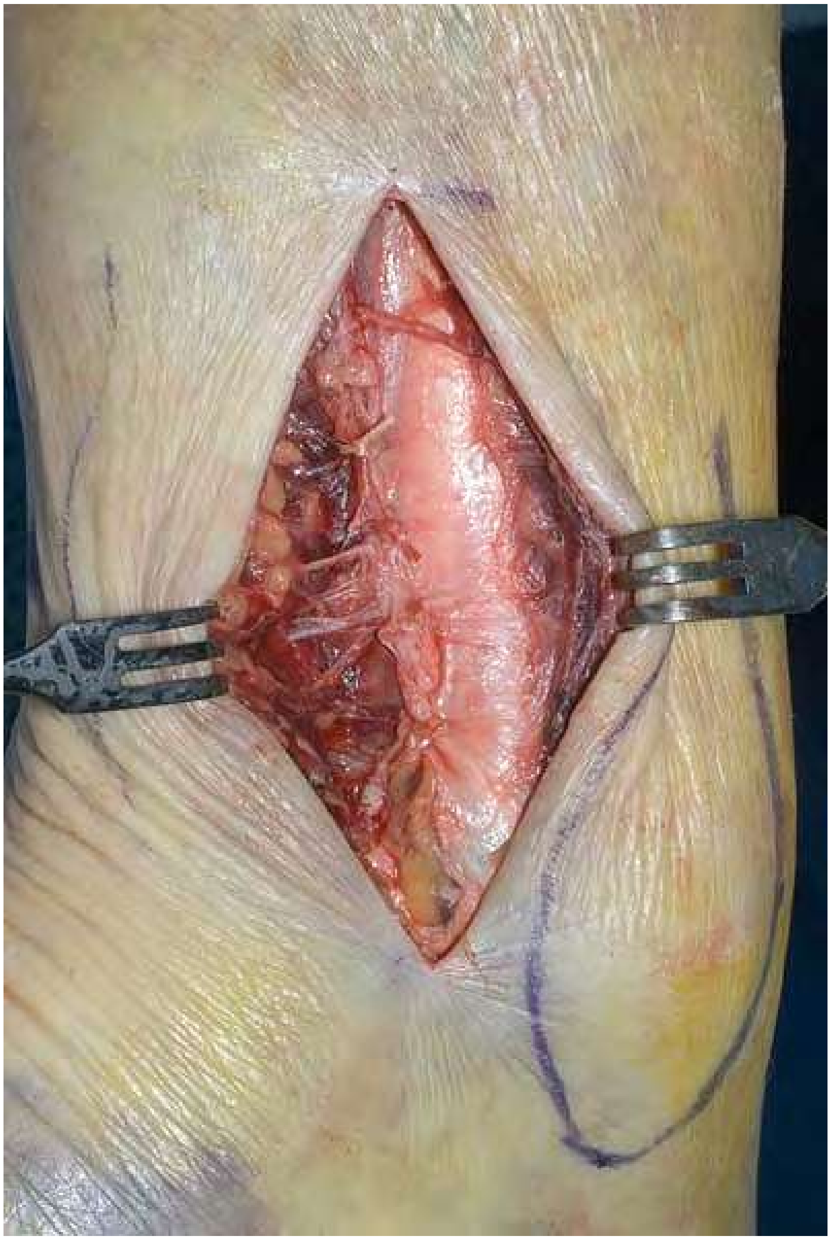

FIGURE 32-13 The peroneal retinaculum is repaired using an absorbable suture.

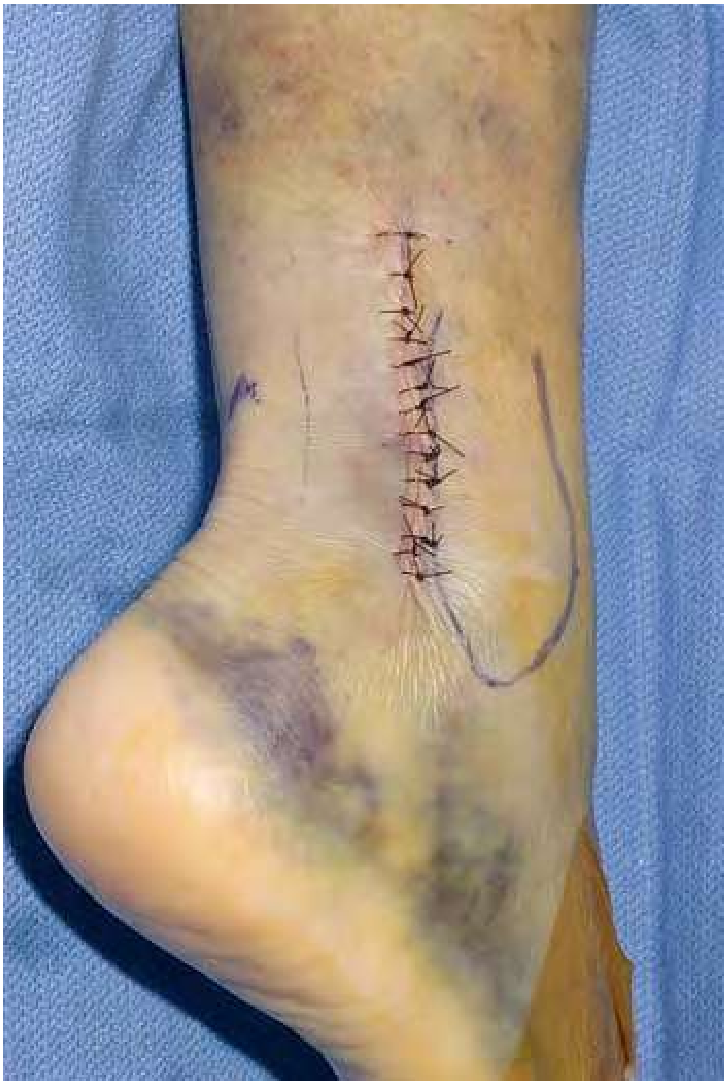

FIGURE 32-14 The skin is closed using a nonabsorbable monofilament suture.

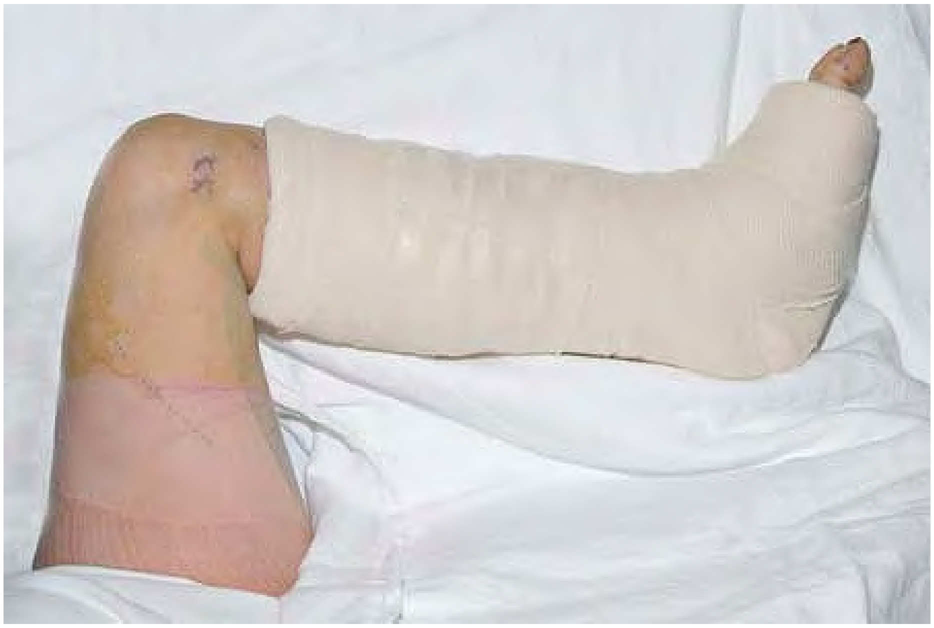

FIGURE 32-15 A well-padded Robert Jones splint is placed to provide further postoperative immobilization.

Tibial Pilon Fractures

The â€œpersonalityâ€• of tibial pilon fractures is much different from that of ankle fractures. Tibial pilon fractures usually represent a much higher energy injury with more involvement of the surrounding soft tissues. While immediate open reduction and internal fixation of some ankle fractures is feasible, the same is not true for pilon fractures. The skin around the distal tibia is not very tolerant of excessive swelling with early operative fixation and therefore allowing sufficient time for solution of swelling is necessary. Initial series of tibial pilon fractures treated with immediate open reduction and internal fixation had significantly higher rates of postoperative wound complications compared with later series in which a staged protocol was utilized. Initial management of pilon fractures is with limited internal fixation of the fibula and spanning external fixation to gain appropriate limb alignment and length. This will provide appropriate soft tissue stabilization to allow for resolution of the swelling that will inevitably occur following tibial pilon fractures.

Preoperative Planning

Initial evaluation of tibial pilon fractures consists of AP and lateral x-rays of the tibia as well as AP, lateral, and mortise radiographs of the ankle. When a pilon fracture has an associated fibula fracture, there is often shortening of the limb. After appropriate soft tissue evaluation, management consists of initial open reduction and internal fixation of the fibular fracture through a lateral or posterolateral approach in addition to application of an external fixator from the tibia to the calcaneus. This

external fixator should be positioned so that the limb length is restored and the tibiotalar joint is reduced. Distraction of the joint may be necessary to prevent â€œcartilage necrosisâ€• if the articular surface is comminuted and irregular. Preoperative planning is important because the incision for fibular fixation should be tailored to the anticipated approach for the tibial internal fixation. A posterolateral approach to the fibula may be utilized when a future anterolateral approach to the tibia is planned. When an anteromedial approach to the tibial pilon fracture is anticipated, a direct lateral approach to the fibula is appropriate. By utilizing a posterolateral approach to the fibula, an appropriate 5 or 6 mm skin bridge can be preserved between this approach and an anterolateral approach to the tibia. A direct lateral approach to the fibula should not be combined with an anterolateral approach to the tibia because the proximity of the incisions and resultant narrow skin bridge may result in skin necrosis between the incisions.

Once careful preoperative planning has determined the initial approaches for fibular stabilization, the patient should be positioned on a radiolucent table. A thigh high pneumatic tourniquet may be utilized during the initial fibular reduction and stabilization. Fixation of the fibula is generally performed prior to application of the external fixator. In the acute setting, fibular reduction is usually easily obtainable as the soft tissues have not yet contracted. In situations when difficulty is encountered gaining fibular length, the external fixator may be applied before fibular reduction. Careful construction of the external fixator will allow access to the fibula during internal fixation.

If the initial stabilization is performed within 24 hours, there is frequently minimal to moderate soft tissue swelling and fibular fixation can be performed. In the polytrauma setting, when fibular fixation may not be performed within 48 hours, careful soft tissue evaluation should be made before planning internal fixation. If the skin does not wrinkle when performing the pinch test, then fibular fixation should be delayed; however, spanning external fixation from the tibial to the calcaneus should not be delayed and this should be performed at the first possible time following injury. The external fixator improves length, angulation, and rotation of the fracture, but more importantly, it restores the appropriate soft tissue tension and can prevent further soft tissue compromise and development of fracture blisters. Tibial pin placement should be proximal enough that anticipated tibial plate fixation will not be contaminated by the pin tracts. A calcaneal transfixion pin is frequently utilized, however pins may also be placed in the talar neck or metatarsals to provide further support to the foot to prevent equinus positioning of the foot.

Following initial spanning external fixation, the patient's lower extremity should be elevated at all times to decrease soft tissue swelling. Deep venous thrombosis prophylaxis should be utilized for these now relatively immobile patients who have joint spanning external fixators in place. Repeat clinical evaluation should be performed on a weekly basis to monitor progression of a soft tissue injury. Once the swelling has subsided to the point that soft tissue wrinkles are attainable over the planned surgical incisions, internal fixation may be performed. There is frequently a 2 to 3 week interval between initial injury and resolution of soft tissue swelling. Once the soft tissues are amenable to surgical intervention, a duplex ultrasound screening examination should be performed on the injured extremity prior to removal of the external fixator and planned internal fixation. Avoidance of tourniquet should be considered in patients who have developed deep venous thrombosis in the injured extremity.

Once the soft tissue swelling has subsided, careful preoperative planning is required to determine appropriate patient positioning, operative approach, and implant choice for internal fixation. Planning consists of reviewing imaging studies, which should include a post reduction CT scan once limb length has been restored with the use of the external fixator and/or fibular fixation. Operative approaches should be selected which allow access to the fracture segment that requires anatomic reduction. Pilon fractures with an associated medial malleolar fracture or coronal split which extends to the fibula are often best approached through a medial or anteromedial approach. Pilon fractures with significant articular comminution, particularly in the anterolateral area, can be approached through an anterolateral approach. Fractures with significant posterior comminution are best approached through a posterolateral approach as previously described for ankle fracture fixation. Two incisions are often necessary and preoperative planning is required to ensure that there is a minimum 6 cm skin bridge between the incisions to prevent skin necrosis.

Posterolateral Approach to the Distal Tibia and Fibula

The patient is positioned in a prone position on a radiolucent operative table. The patient should be translated towards the foot of the bed so that the midfoot and toes are hanging off the bed so the foot

is free. A well-padded pneumatic tourniquet may be placed around the thigh and should be placed prior to placing the patient in a prone position. The landmarks of the lateral malleolus and Achilles tendon are identified by palpation. A longitudinal skin incision is made half-way between the lateral border over the Achilles tendon and the posterior border of the lateral malleolus. The incision should extend distally until it reaches the level of the tip of the lateral malleolus. The deep fascia of the leg is incised in line with the incision, and the peroneal musculature and tendons are identified and retracted laterally. The peroneal retinaculum is incised distally to allow lateral retraction of these tendons. The flexor hallucis longus is elevated from its origin on the fibula and retracted medially to gain access to the posterior aspect of the tibia. Once reduction and fixation have been completed, the peroneal retinaculum and deep fascia of the leg are closed utilizing an absorbable suture. The subcutaneous tissue is closed with a 2-0 resorbable suture and the skin is then closed utilizing either 3-0 or 4-0 nonabsorbable monofilament sutures such as nylon.

Anterolateral Approach to the Distal Tibia

Patient Positioning

The patient is positioned supine on a radiolucent operating table and a well-padded pneumatic tourniquet is placed around the thigh. A bump is placed under the operative hip so that the foot rests vertically in the neutral position. The intraoperative C-arm is positioned on the uninjured side perpendicular to the long axis of the patient. A foam block or several folded towels can be utilized to elevate the operative limb higher than the nonoperative limb to provide a clear image with fluoroscopy. This also minimizes frequent manipulation of the leg that could potentially compromise the reduction.

Surgical Technique

For the anterolateral approach, a longitudinal skin incision is made in line with the fourth metatarsal. The incision is extended proximally as far as necessary to expose the tibia for application of a plate of the appropriate length and distally will end over the talus. The superficial peroneal nerve should be protected and retracted laterally in the wound. The subcutaneous nature of the superficial peroneal nerve requires careful skin dissection and blunt dissection once the skin has been incised. The fascia and extensor retinaculum are incised in line with the skin incisions with care taken to avoid undermining the skin flaps. The extensor digitorum longus is retracted medially and the peroneal muscles are retracted laterally. During medial dissection, avoid injury to the deep peroneal nerve and anterior tibial artery as they cross the ankle joint.

Following reduction and fixation through the anterolateral approach, the superior and inferior extensor retinaculum is repaired utilizing absorbable suture. Once the retinaculum and fascia have been reapproximated, the wound is closed with 2-0 absorbable subcutaneous sutures followed by interrupted vertical mattress nylon sutures in the skin.

Anteromedial Approach to the Distal Tibia

Preoperative Planning

The patient is positioned supine on a radiolucent operating room table and a thigh high pneumatic tourniquet is used. A foam block or set of folded towels is put under the operative leg to elevate it higher than the nonoperative leg for lateral intraoperative fluoroscopic imaging. The foot should rest in a slightly externally rotated position; therefore; placement of bumps underneath the ipsilateral hip is unnecessary. Preoperative templating to select the appropriate implant and length prevents miscalculation and unnecessary delays in the operating room.

Technique

The skin incision parallels the tibialis anterior tendons and extends proximal enough to allow application of an appropriate length of plate. Distally, the incision curves slightly medially over the ankle joint towards the talonavicular joint. The extensor retinaculum over the tibialis anterior is incised in line with the incision, and the tibialis anterior is retracted laterally. Dissection should continue beneath the tibialis anterior on the anterior aspect of the tibia, elevating the hallucis longus and extensor digitorum tendons which are retracted laterally with the tibialis anterior. The dissection may also be carried medially towards the medial malleolus to expose fractures that extend into this region. Following open reduction and internal fixation, repair the extensor retinaculum over the tibialis anterior utilizing an absorbable size 0 suture. The subcutaneous tissue is closed using an 2-0 resorbable suture, and the skin is then closed using an interrupted vertical mattress nylon suture.

Postoperative Management

Following sterile dressing application, a bulky Robert Jones type dressing utilizing both a posterior splint, as well as a stirrup-type splint, provide added stability. The patient is hospitalized until adequate pain control has been achieved. Careful postoperative evaluation should be performed to monitor for development of possible compartment syndromes in the leg following internal fixation of pilon fractures. The patient is maintained in the initial postoperative dressing for the first 7 to 10 days, which can then be exchanged for a short leg cast for further protection of both the bony and soft tissue elements. Sutures may be removed approximately 2 to 3 weeks postoperatively once the wound has healed. Depending on the bone quality and the quality of fixation, the patient is transitioned into a fracture boot or cast when rigid fixation is a concern. Once the soft tissues have healed, the patient should be encouraged to remove the walking boot multiple times per day to begin range of motion of the ankle. Weight bearing is not permitted until radiographic evidence of healing is seen, usually at 12 weeks from surgery. Initial weight bearing begins in a walking boot and once the patient is comfortable in the boot, they wean into a regular shoe over a several week period. After progression of weight bearing, radiographs should be obtained in 3 to 4 weeks to assure there is no displacement of the fractures.

Complications

Delayed Wound Healing/Wound Dehiscence

Soft tissue complications can be minimized when an appropriate protocol of external fixation followed by delayed internal fixation is respected. In the event that delayed wound healing with dehiscence occurs, sutures should be left in place if it is a superficial skin dehiscence. Often, these can be treated often with dressing changes and occasionally oral antibiotics if there is surrounding cellulitis. In the event that there is deep or full thickness necrosis of either the posterolateral or anterolateral approaches, careful clinical inspection should be performed to determine whether the infection communicates with the deep hardware. Due to the relatively subcutaneous nature of the anteromedial approach, full thickness break-down should be treated as a deep infection with formal irrigation, debridement, and antibiotic treatment. Full thickness dehiscence of the anterolateral or posterolateral exposures should be inspected to determine if there is communication with the deep hardware. In the event that the subcutaneous and retinacular layers appear intact, appropriate treatment methods include local debridement and dressing changes. Any infection that communicates deep with the hardware needs to be treated aggressively with formal surgical debridement and postoperative dressing changes to obtain closure. Alternatively, VAC devices may be used in open wounds to help gain granulation tissue prior to an impending soft tissue coverage operation. With any type of wound dehiscence, range of motion exercises should be halted until resolution of the wound complication. While a cast boot is convenient for inspection of the wounds, a formal cast application with window cut-outs is more effective at immobilizing the limb and allows resolution of the soft tissue complications. In the event of early deep infections, repeat serial irrigation and debridements are necessary with intravenous antibiotic treatment to obtain suppression of the infection until union has occurred. If the infection persists following fracture union, then the hardware may be removed and further soft tissue coverage provided as necessary.

Gracilis Free Muscle Flap for Extensive Tissue Loss Following Ankle Injuries

Despite the best attempts at atraumatic tissue handling during the surgical exposure soft tissue defects surrounding the ankle may still occur. Whether due to initial soft tissue loss, or due to tissue devitalization following fixation, soft tissue deficits surrounding the ankle often require urgent coverage to avoid the complications of hardware infection, nonunion, and osteomyelitis.

For soft tissue closure of ankle defects there are many options, including the reversed soleus flap, the sural artery flap, and free latissimus dorsi muscle flap. Our preference however, in these situations, is the use of the gracilis free muscle flap. This flap provides a great deal of versatility, contours well to the ankle, and produces minimal donor site morbidity.

Indications/Contraindications

Any defect surrounding the ankle can be considered for flap coverage. The flap is limited in terms of surface area, and is not well suited for circumferential coverage of the ankle and Achilles tendon

region. Also the pedicle length can be limited if a wide zone of injury requires the use of recipient vessels higher than the midcalf.

Patient factors which may increase the risk of flap failure include atherosclerosis, smoking status, and renal failure but these are not considered absolute contraindications to surgery. Absence of adequate recipient vessels is the only absolute contraindication for flap transfer.

Preoperative Planning

Often following significant lower extremity trauma a CT angio-gram is obtained to evaluate the arterial inflow to the ankle. In severe ankle trauma the posterior or anterior tibial arteries may be occluded or transected, limiting recipient vessel options: either vessel may be used as a recipient vessel for successful flap transfer. If an angiogram is not to be performed prior to surgery a careful assessment of the posterior and anterior tibial vessels must be performed with the use of a hand held Doppler probe.

Surgery

Patient Positioning

The patient is placed supine on the operating room table with the leg in mild abduction and the knee slightly bent. Depending on the extent of the concomitant trauma either the ipsilateral or contralateral gracilis may be used. If possible we prefer to harvest the gracilis from the same leg as the original ankle trauma.

Technique

The gracilis receives its blood supply from the medial femoral circumflex artery which originates from the profunda femoral artery. The major pedicle can be identified approximately 8 to 10 cm inferior to the pubic tubercle. The flap also has a minor arterial pedicle which enters the muscle at the level of the mid thigh. This artery originates from the superficial femoral artery. The muscle receives its nervous innervation from the anterior branch of the obturator nerve. This branch of the obturator nerve can be harvested with the muscle if there are requirements for functional muscle transfer (Fig. 32-16).

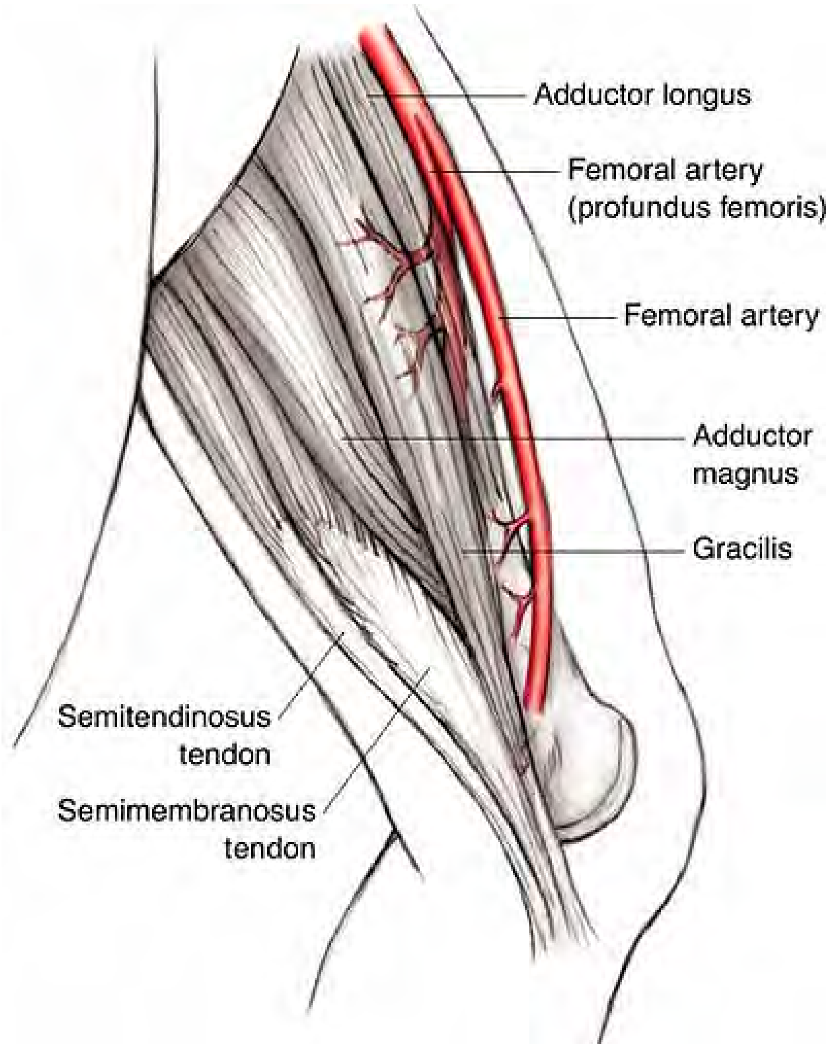
Adductor longus
Femoral artery
(profundus femoris)
Femoral artery
Adductor
magnus
Gracilis
Semitendinosus
tendon
Semimembranosus
tendon

FIGURE 32-16 The gracilis muscle lies in the medial thigh. Its major pedicle is a branch of the medial femoral cutaneous artery which is a branch of the profunda femoris artery. The major pedicle can be identified entering the muscle 8 to 10 cm inferior to the pubic tubercle.

The muscle is exposed through a medial thigh incision. The incision is made 3 cm posterior to a line connecting the pubic tubercle and the medial condyle of the femur. A single long incision provides the most expedient means of identification of the gracilis muscles; however the resultant scar can be unsightly. Alternatively the muscle may be harvested through two separate smaller incisions; a 10 to 12 cm incision may be created over the area of the vascular pedicle and a second 5 cm incision can be created at the level of the medial femoral condyle which allows for division of the muscle insertion (Fig. 32-17).

Proximally the gracilis muscle lies between the adductor longus medially and the semitendinous muscle inferiorly. It lies superficial to the adductor magnus. The pedicle lies deep to the adductor longus but runs along the superficial surface of the adductor magnus.

Once the skin incision has been made the muscle is carefully identified inferiorly at the level of the knee. Inferiorly the muscle lies posterior to the sartorius muscle and anterior to the insertion of the semimembranous and semitendinous muscles (Fig. 32-18). The gracilis can be confused with the sartorius muscle. It may be differentiated from the sartorius and semimembranous tendon by looking for the musculotendinous portion of the gracilis. At the level of the medial femoral condyle the gracilis consists of both muscle and terminal tendon, the semimembranosus is entirely composed of tendon and the sartorius is entirely muscle. Once the gracilis is identified it may be rapidly separated from the surrounding tissue with blunt dissection to the level of the minor pedicle. The minor pedicle is not divided until the major pedicle is clearly visualized (Fig. 32-19).

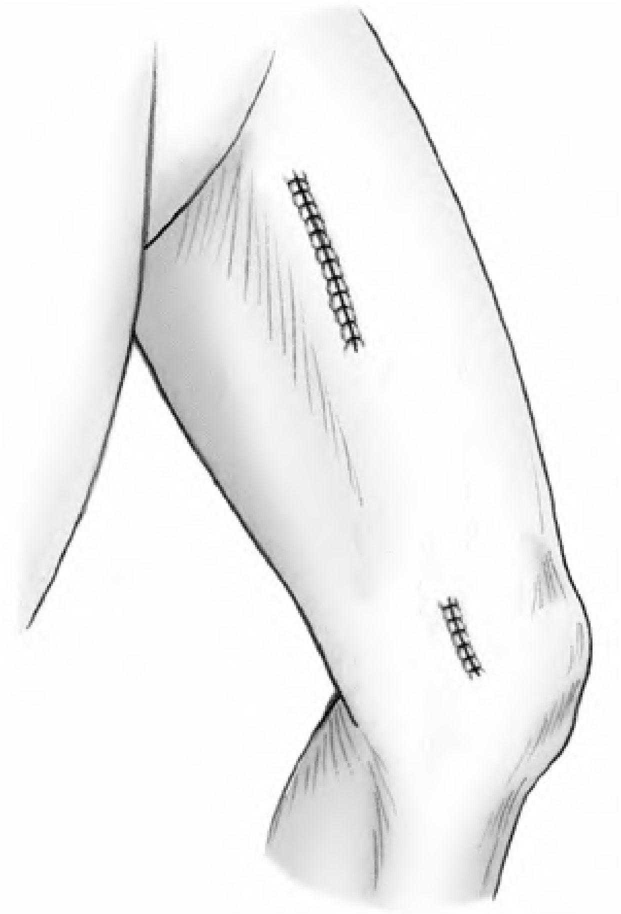

FIGURE 32-17 The muscle may be harvested through a long medial incision, or alternatively through two smaller incisions; a 10 to 12 cm incision may be created over the area of the vascular pedicle and a second 5 cm incision can be created at the level of the medial femoral condyle which allows for division of the muscle insertion.

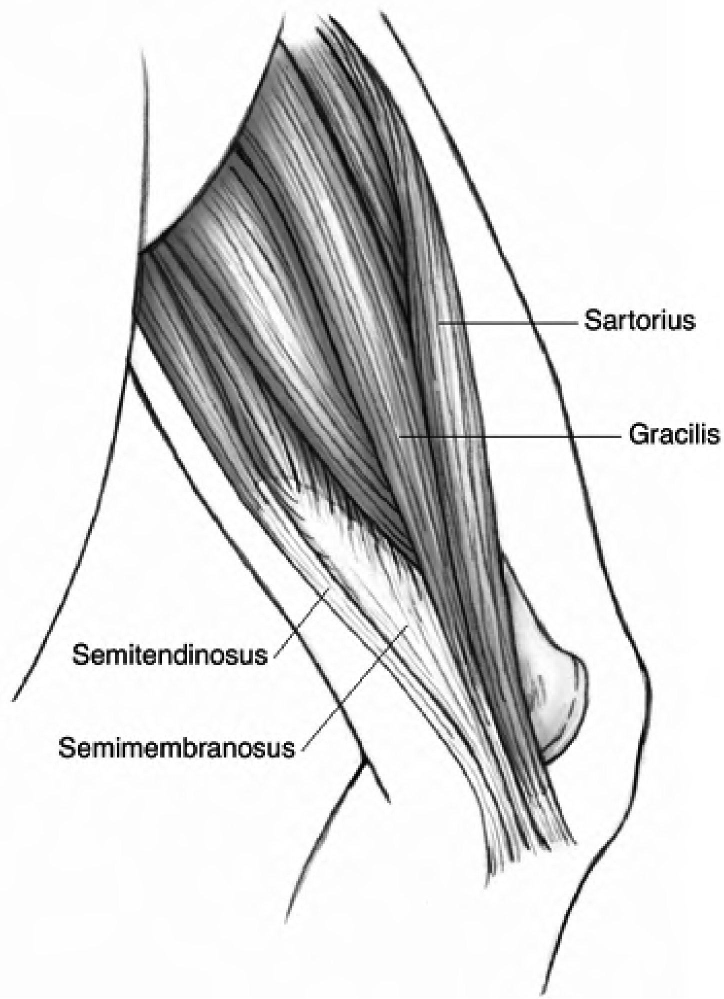

FIGURE 32-18 Inferiorly the muscle lies posterior to the sartorius muscle and anterior to the insertion of the semimembranous and semitendinous muscles. The gracilis can be confused with the sartorius muscle. It may be differentiated from the sartorius and semimembranous tendon by looking for the musculotendinous portion of the gracilis. At the level of the medial femoral condyle the gracilis consists of both muscle and terminal tendon, the semimembranosus is entirely composed of tendon, and the sartorius is entirely muscle.

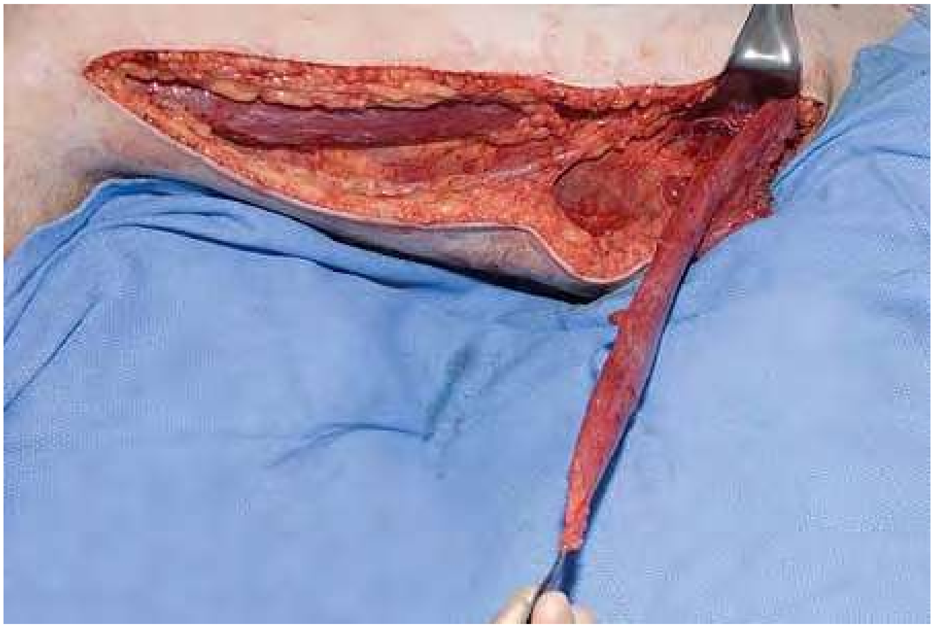

FIGURE 32-19 The muscle may be rapidly elevated back to the major pedicle. The pedicle can be seen inferior to the soft tissue retractor. The saphenous vein has been preserved as it crosses the incision midway down the thigh.

The major pedicle is identified running at the proximal lateral margin of the muscle approximately 8 to 10 cm from the pubic tubercle. The adductor longus is retracted laterally to expose the major pedicle. The fascia of the adductor magnus must be divided to allow for mobilization of the pedicle significantly. The pedicle is traced back to its origin at the profunda femoris vessels. Multiple perforating branches to the adductor muscles must be divided to gain exposure to the profundus artery, thus maximizing pedicle length of up to 6 cm. Dissection and visualization is aided with the use of lighted retractors; alternatively the adductor longus may be mobilized and retracted medially allowing visualization of the pedicle origin deep to the muscle (Fig. 32-20).

Once the major pedicle is identified and determined to be adequate for microvascular anastomosis the secondary pedicle is divided. The origin of the muscle and branch of the obturator nerve is now divided. The muscle is left to perfuse on its major pedicle until the recipient vessels have been prepared for microvascular anastomosis.

Prior to muscle transfer a final definitive debridement of the defect site is performed. The recipient vessels are then exposed away from the original zone of injury (Fig. 32-21). Donor vessel preparation

is performed with the use of high powered operative loop magnification or the microscope to verify that the vessels are adequate for microvascular anastomosis. The anterior tibial artery is used as a recipient vessel for lateral malleolar defects while the posterior tibial artery is preferentially used for medial defects (Fig. 32-22). End to side anastomosis to the artery is preferred while an end to end anastomosis to the venous comitantes is performed in most settings. If two veins are present within the arterial pedicle both of the venous comitantes to the gracilis muscle are anastomosed to minimize the chances of postoperative venous insufficiency.

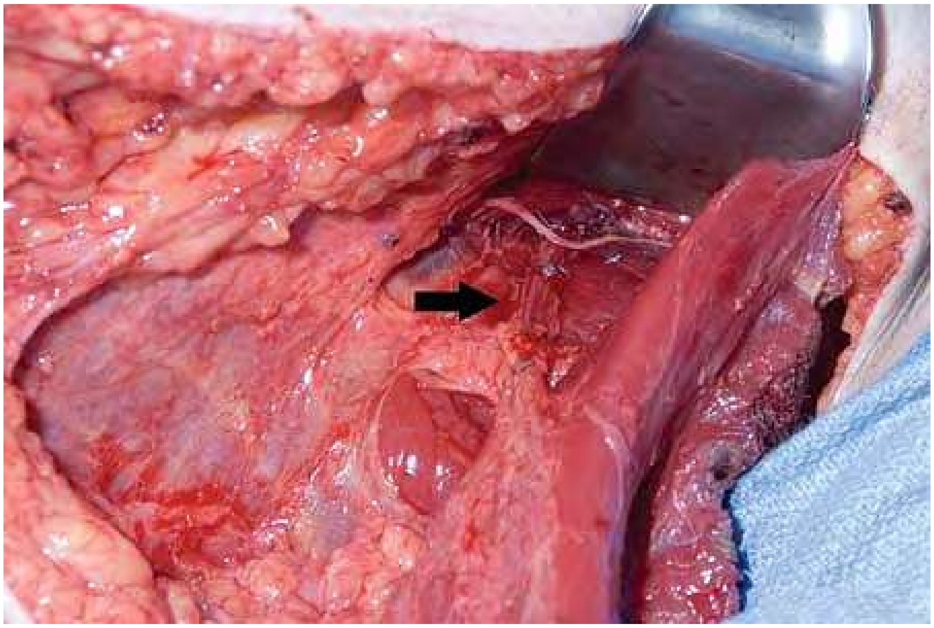

FIGURE 32-20 The pedicle (*black arrow*) is dissected back to its origin at the profunda femoris. This allows for a significant improvement in pedicle length. A branch of the obturator nerve can be seen running on top of pedicle and the

adductor magnus muscle.

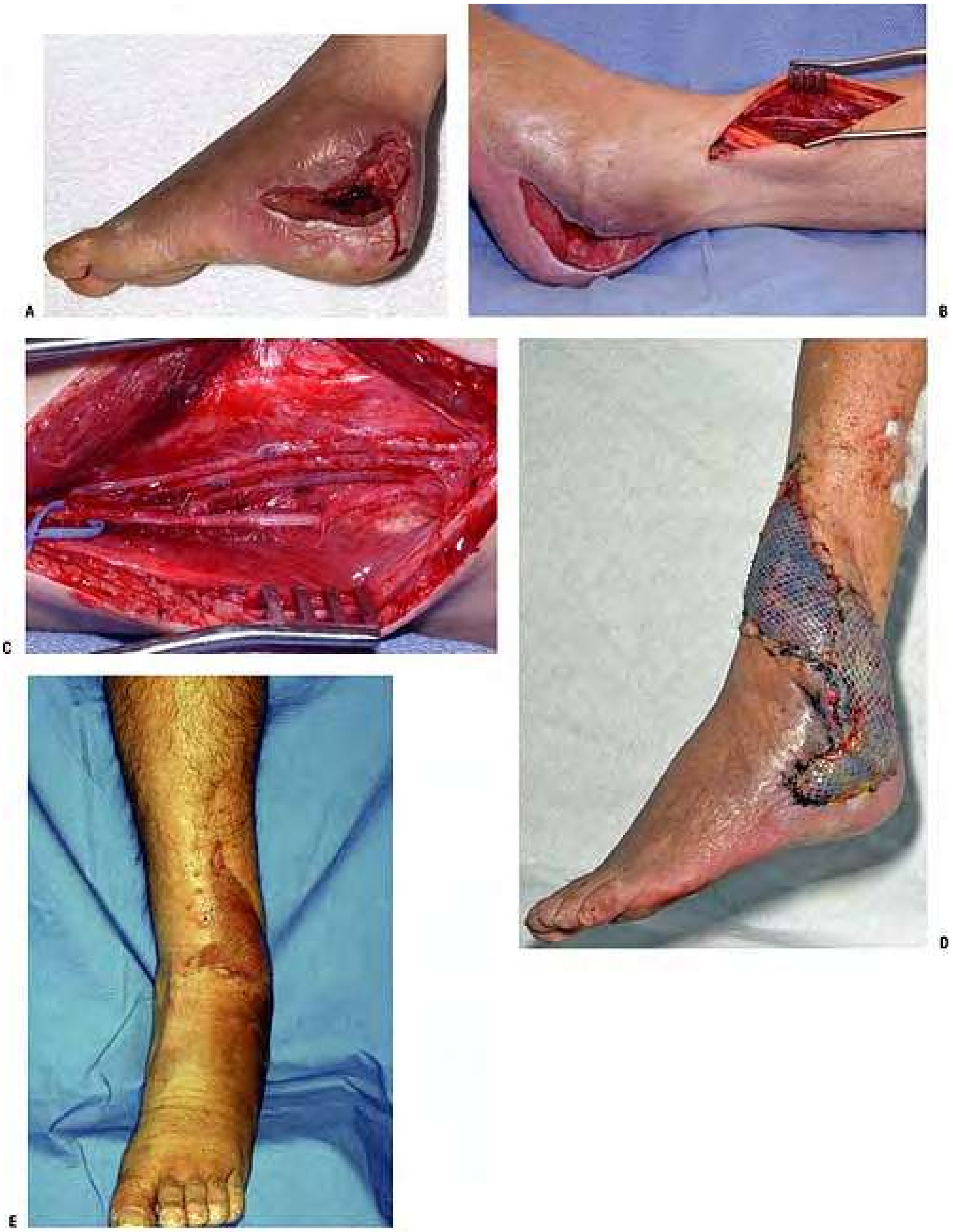

FIGURE 32-21 A: A lateral malleolar defect has developed in a 48-year-old

gentleman following stabilization of a complex malleolar and calcaneal fracture. B: The anterior tibial vessels are exposed away from the zone of injury in preparation for free tissue transfer. C: The anterior vessels are identified just lateral and deep to the anterior tibialis muscle. The perineal nerve is surrounded with a vessel loop and gently freed from the underlying vessels. D: The flap at one week shows good take of the skin graft. E: Four month view shows excellent contouring of the flap with stable wound coverage allowing for normal shoe wear.

Once the anastomosis is complete the muscle is allowed to reperfuse for 20 minutes as the donor site is closed. An implantable Doppler probe is placed circumferentially around one of the two veins to allow for evaluation of anastomotic patency postoperatively. Donor site closure is performed with deep dermal sutures and an absorbable subcutaneous monofilament suture. A closed suction drain is also used to prevent postoperative seroma or hematoma formation. Flap insetting is then performed. Muscle insetting is facilitated with the use of half buried absorbable sutures placed within the muscle and pulled beneath the native skin. The epimysium of the muscle is often excised to allow for expansion of the muscle's surface area. This allows for improved contouring over the malleolus. At the completion of insetting the muscle is covered with a meshed split thickness skin graft. The flap is covered with a Xeroform and the patient's leg is then loosely wrapped in sterile cotton and placed into a posterior splint for postoperative comfort. A large window is created in the dressing for post-operative flap monitoring. A single suture is placed superficially within the muscle over the pedicle so that anastomotic patency can be evaluated postoperatively with the use of a hand held Doppler device.

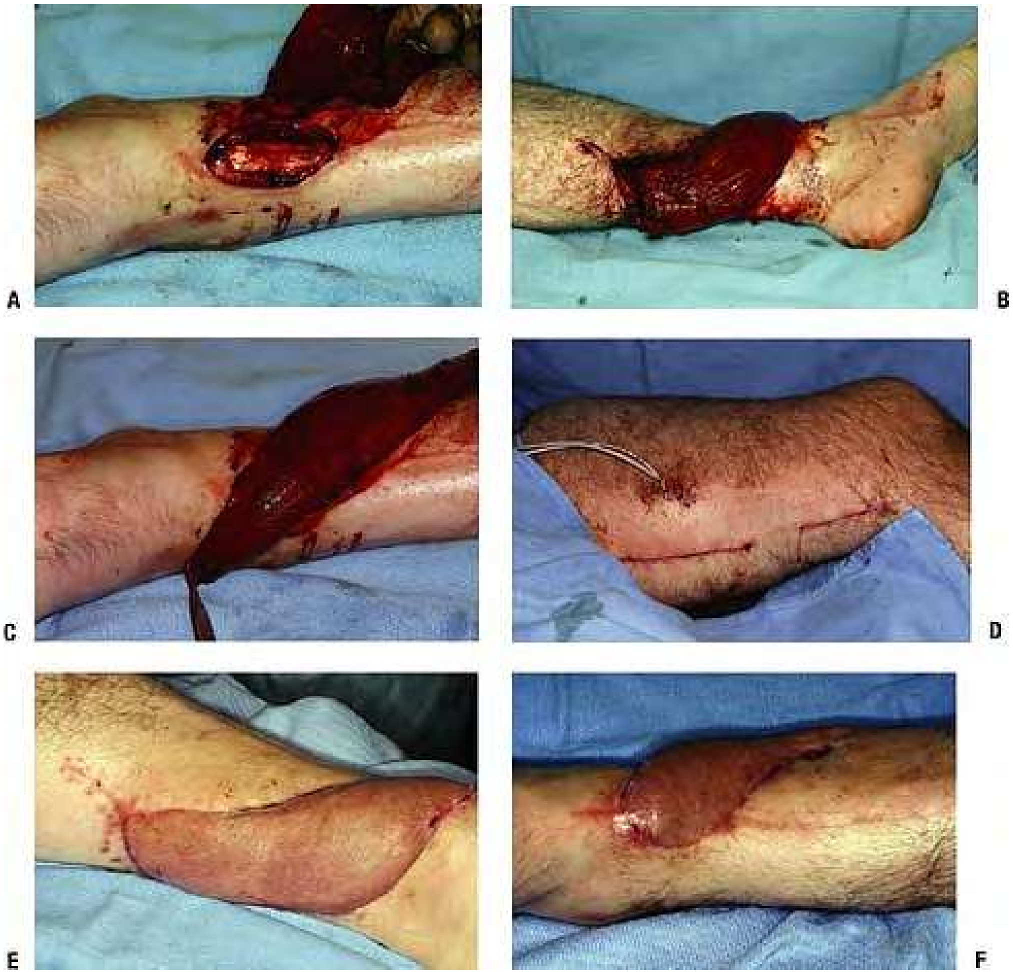

FIGURE 32-22 A: A 32-year-old gentleman presents with a defect of the lower margin of the tibia which developed following treatment of a malleolar fracture. B,C: In this case the gracilis muscle is anastomosed to the posterior tibial vessels and brought anteriorly to cover the bone and soft tissue defect. D: The muscle was harvested through two limited incisions over the medial thigh. The inferior incision is used for division of the muscle insertion while the superior incision is used for dissection of the vascular pedicle and division of the muscle origin. E,F: The wound is stable at 6 months with no further drainage or signs of ongoing infection.

Postoperative Management

Flap monitoring may be performed with either the use of an implantable Doppler probe, or with the use of a hand held Doppler probe at the bedside. Evaluation of perfusion is performed every hour for the first 24 hours by a skilled microsurgical nursing staff who are familiar knows the signs of arterial insufficiency and venous thrombosis.

On postoperative day 4 or 5 the entire dressing is removed and skin graft survival is assessed. At this point the patient may be placed into a loose Robert-Jones dressing or a removal posterior splint. The donor site drain is removed once output is less than 30 cc a day.

Complications/Results

The most devastating complication following any free tissue transfer include partial flap loss and total flap loss. If arterial insufficiency or venous insufficiency occurs postoperatively, immediate re-exploration is recommended. At the time of surgical exploration thrombosis may be identified at the anastomotic site. In such cases early revision of the anastomosis will often allow for flap salvage. If the flap is unsalvageable and flap loss occurs, an additional flap will be required. In such cases a secondary free flap, pedicled flap, or cross-leg flap can be considered. In cases of partial flap loss, occasionally the flap may be advanced distally to cover critical structures. Alternatively small areas of distal necrosis may be managed with dressing changes, skin grafting, or pedicle flap coverage.

In an examination of 50 acute traumatic and post-traumatic wounds, Redett and colleagues found the gracilis muscle effective for covering defects up to 165 cm^2. Successful free tissue transfer was performed in 93% of patients, and limb salvage was possible in 96%. Studies by Hallock have shown that the vascular pedicle is of adequate length for reconstruction of defects extending over the calcaneus and lateral malleolus (Fig. 32-21). For defects which exceed the limits of a gracilis transfer a latissimus free muscle flap can be considered. We have found the gracilis to contour nicely over time allowing for normal shoe wear following reconstruction.

Recommended Reading

Blauth M, Bastian L, Krettek C, et al. Surgical options for the treatment of

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 33

Calcaneal Fractures/Talar Neck Fractures

Michael P. Clare

Roy W. Sanders

Fractures of the calcaneus and talus are among the most complex and challenging of fractures for the orthopaedic surgeon to effectively manage, in that both are three-dimensionally unique structures with highly specialized biomechanical function and a distinctly limited surrounding soft tissue envelope. Fractures of the calcaneus and talus are generally the result of high-energy trauma, such as a motor vehicle accident or a fall from a height, and as such, the severity of fracture displacement and the extent of soft tissue disruption are directly proportional to the amount of force and energy absorbed by the limb.

Fractures of the Calcaneus

Indications/Contraindications

General Considerations

Operative treatment is generally indicated for displaced intra-articular fractures involving the posterior facet as demonstrated on computed tomography (CT) scanning. Nonoperative treatment is best reserved for non- or minimally-displaced extra-articular calcaneal fractures, and truly non-displaced intra-articular fractures as determined on CT scan.

Comorbidities

While nicotine use/dependence is not a contraindication to operative treatment, all patients who are smokers are counseled at length as to the associated risks and encouraged to discontinue tobacco use. We consider heavy smoking (â‰¥2 packs per day) as a relative contraindication to surgery. Specific contraindications for operative treatment include fractures in patients with insulin-dependent diabetes mellitus, severe peripheral vascular disease, or other major medical comorbidities precluding surgery, as well as fractures in elderly patients who are minimal (household) ambulators. Because many older patients are healthy and active well into their 70s, chronologic age itself is not necessarily a contraindication to surgical treatment.

Delayed Treatment by Necessity

Operative treatment may also be contraindicated if initial evaluation has been delayed beyond 3 or 4 weeks from the date of injury, or in certain situations in which injury severity prohibits early surgical intervention, including fractures associated with severe fracture blisters or prolonged edema; fractures with large open wounds; and fractures in patients with life-threatening injuries. Operative treatment following a prolonged delay (>4 weeks) from initial injury to definitive treatment is complicated by the fact that early consolidation of the fracture has occurred, making the fracture fragments increasingly difficult to separate to obtain an adequate reduction, and the articular cartilage may delaminate away from the underlying subchondral bone. In these instances, delayed treatment by necessity is used whereby the fracture is allowed to heal and is later managed as a calcaneal malunion following resolution of the prohibitive factors.

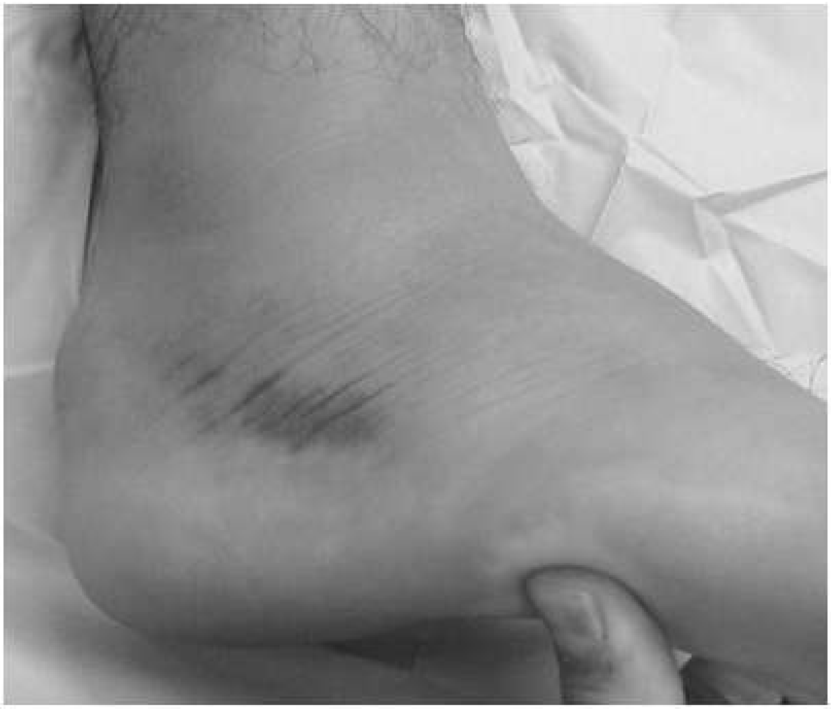

FIGURE 33-1 Preoperative â€œwrinkleâ€• test. Note presence of skin creases indicating that surgery may now be safely undertaken.

Preoperative Planning

In completing preparations for surgery, the surgeon should thoroughly review the plain radiographs and CT scans to gain a preliminary understanding as to the â€œpersonalityâ€• of the fracture pattern, which then allows for planning of patient position and surgical approach, as well as anticipation of specific technical steps in obtaining fracture reduction, and the necessary implants for definitive stabilization. While multiple surgical approaches have historically been described, the general consensus currently favors the extensile lateral approach for most displaced intra-articular calcaneal fractures. Percutaneous reduction techniques, such as the Essex-Lopresti maneuver, are particularly ideal for certain tongue-type fracture (Sanders type II-C) patterns.

Surgery is ideally completed within the first 3 weeks of injury before early fracture consolidation; however, surgery should not be attempted until the associated soft tissue swelling has sufficiently dissipated as demonstrated by a positive wrinkle test. The test involves assessment of the lateral calcaneal skin with passive dorsiflexion and eversion of the injured foot. A positive test is confirmed by the presence of skin wrinkling without residual pitting edema, such that surgical intervention may be safely undertaken (Fig. 33-1). A variety of modalities may used to decrease swelling in the affected extremity. We prefer initial elevation and immobilization in a bulky Jones dressing and supportive splint, with later conversion to an elastic compression stocking and fracture boot locked in neutral flexion as the acute swelling begins to dissipate in the ensuing few days.

Extensile Lateral Approach

Surgery

Patient Positioning/General Considerations

The procedure requires use of a radiolucent table and a standard C-arm. For isolated injuries, we prefer placing the patient in the lateral decubitus position on a beanbag. The lower extremities are positioned in a scissor configuration, whereby the operative (â€œupâ€•) limb is flexed at the knee and angles toward the posterior, distal corner of the operating table, and the nonoperative (â€œdownâ€•) limb is extended at the knee and positioned away from the surgical field to facilitate intraoperative fluoroscopy. Protective padding is placed beneath the contralateral limb for protection of the peroneal nerve, and an operating â€œplatformâ€• is created with blankets and foam padding to elevate the operative limb (Fig. 33-2). The prone position may alternatively be used in the event of bilateral injuries.

The patient is given prophylactic preoperative antibiotics, and a pneumatic thigh tourniquet is used. The procedure should be completed within 120 to 130 minutes of tourniquet time, so as to minimize wound complications. If the procedure extends beyond that time, the tourniquet should be released and the remainder of the procedure performed without it. In allocating tourniquet time, the surgical approach should be completed within 20 minutes, allowing up to 60 minutes for fracture reduction, 20 minutes for implant placement, and 20 minutes for wound closure.

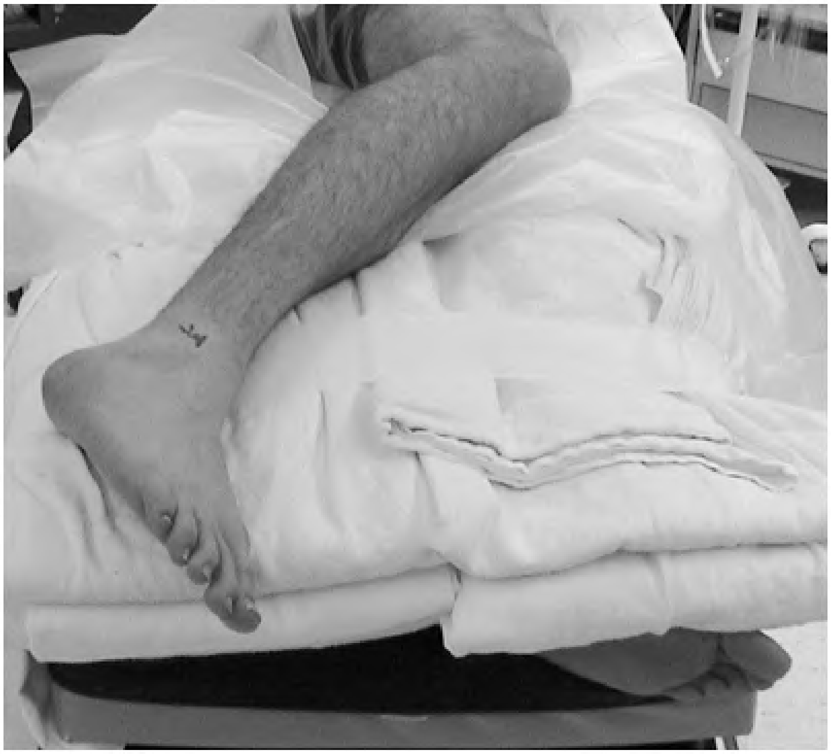

FIGURE 33-2 Lateral decubitus position. Note scissor-like limb configuration to facilitate intraoperative fluoroscopy.

Technique

Anatomic Landmarks

Wound complications following surgical management of calcaneal fractures remain a major source of morbidity with these injuries. The soft tissues overlying the lateral hindfoot receive blood supply from a confluence of three arterial branches: the lateral calcaneal artery, the lateral malleolar artery, and the lateral tarsal artery. Borrelli and Lashgari determined that the majority of the full-thickness flap with an extensile lateral approach is supplied by the lateral calcaneal artery, typically a branch of the peroneal artery. At the level of lateral malleolus, the artery courses parallel to the Achilles tendon and lies approximately 11 to 15 mm anterior to the terminal Achilles tendon and its insertion. Thus, strict attention to detail with respect to placement of the incision and gentle handling of the soft tissues is of paramount importance.

Incision and Surgical Approach

The extensile lateral incision is then outlined on the skin with a marking pencil (Fig. 33-3). The incision begins approximately 2 cm proximal to the tip of the lateral malleolus, just lateral to the Achilles tendon extending toward the plantar foot. In this manner, the vertical limb of the incision will course posterior to the sural nerve and the lateral calcaneal artery, thereby avoiding devascularization of the lateral calcaneal flap. The horizontal limb of the incision is drawn along the junction of the skin of the lateral foot and heel pad, the demarcation of which can be identified by compressing the heel. We prefer to substitute a gentle curve where these two lines intersect to form a right angle, primarily to avoid necrosis of the apical skin. The terminal portion of the horizontal limb includes a gentle curve anteriorly along the skin creases, extending over the calcaneocuboid articulation.

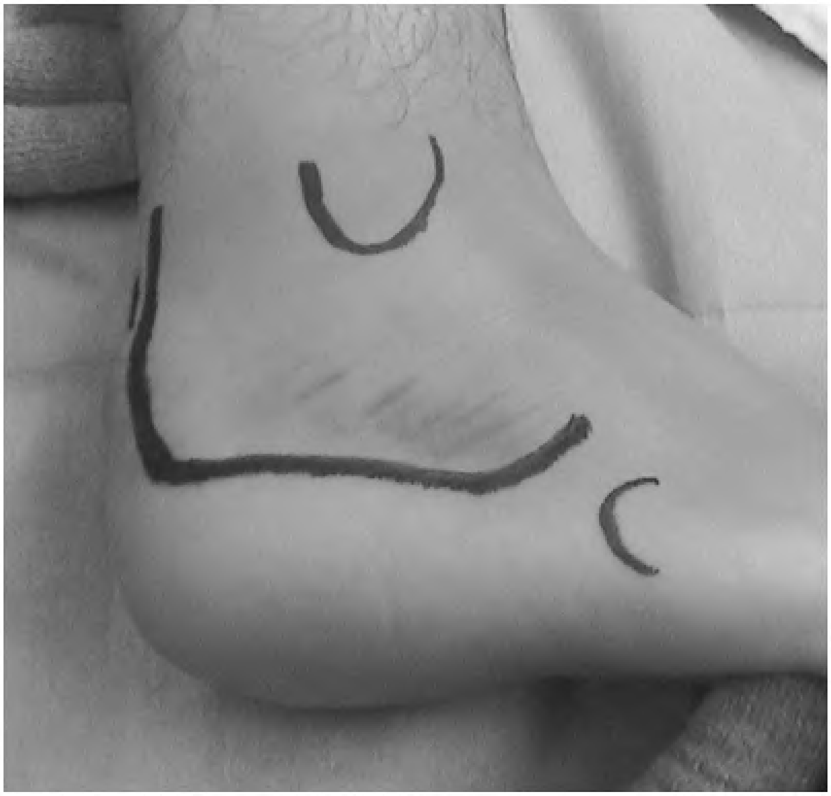

FIGURE 33-3 Planned extensile lateral incision.

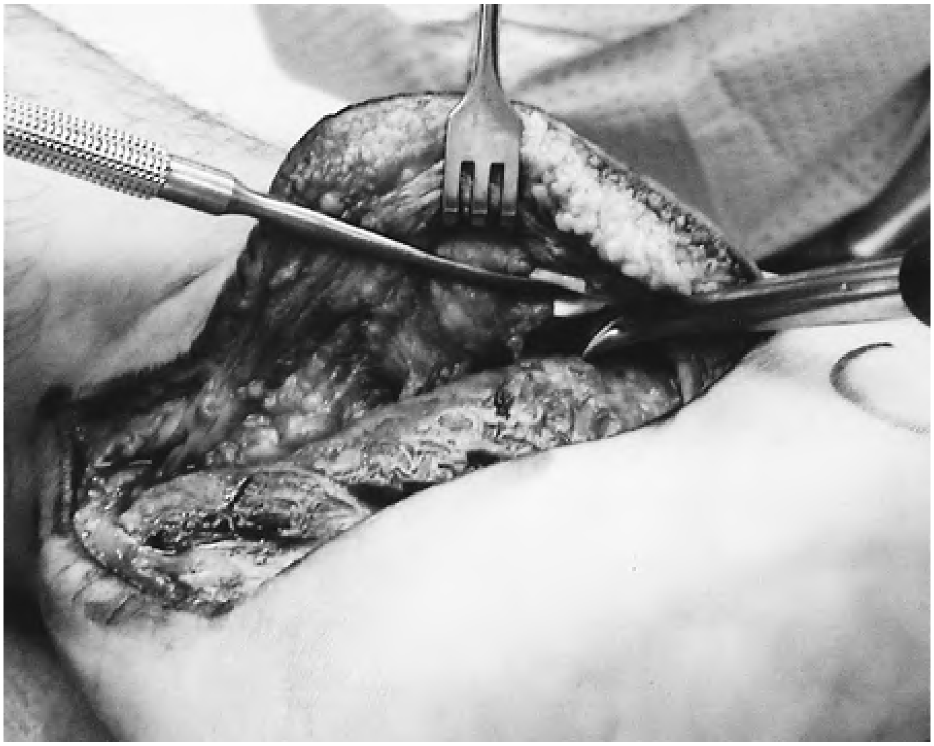

FIGURE 33-4 Full-thickness subperiosteal flap. Note peroneal tendons contained within the flap.

With a sterile bolster placed beneath the medial ankle, the incision begins at the proximal portion of the vertical limb, becoming full-thickness at the level of the calcaneal tuberosity; dissection is specifically taken from â€œskin to boneâ€• at this level while avoiding any beveling of the skin. Scalpel pressure is again lessened beyond the apical curve, roughly at the midpoint of the horizontal limb, at which point a layered incision is again developed.

A full-thickness subperiosteal flap is then developed starting at the apex. Any use of retractors should be avoided until a sizeable flap has been raised, so as to avoid separation of the skin from the subcutaneous tissues and periosteum. The calcaneofibular ligament and inferior peroneal retinaculum are released sharply, thus exposing the peroneal tendons. Both tendons are identified and released from the peroneal tubercle through the cartilaginous pulley, and further gently mobilized distally with a periosteal elevator, thereby exposing the anterolateral calcaneus and calcaneocuboid joint. In this manner, the peroneal tendons and sural nerve are contained entirely within the sub-periosteal flap (Fig. 33-4).

Deep dissection continues anteriorly to the sinus tarsi and anterior process, and posteriorly to the superior-most portion of the calcaneal tuberosity for â€œwindow visualizationâ€• of the posterior facet, so as to prevent rotational malalignment of the posterior facet articular surface in the sagittal plane. Three 1.6-mm Kirschner wires are then placed for retraction of the subperiosteal flap using the â€œno touchâ€• technique. In this technique, one wire is placed into the distal fibula as the peroneal tendons are slightly subluxed anteriorly; a second wire is placed in the talar neck; a third wire is placed in the cuboid as the peroneal tendons are levered away from the anterolateral surface of the calcaneus with a periosteal elevator. Thus, each Kirschner wire retracts its respective portion of the peroneal tendons and subperiosteal flap (Fig. 33-5). A small Bennett-type retractor may additionally be used at the distal margin of the sinus tarsi for further exposure of the anterolateral calcaneus.

Assessment of the Peroneal Tendons

Following fracture reduction, definitive stabilization, and final fluoroscopic images, the wound is copiously irrigated. As the previously placed Kirschner wires are manually removed, the peroneal tendons should easily reduce into the peroneal groove at the posterior border of the lateral malleolus. A Freer elevator is introduced into the tendon sheath, advanced proximally to the level of the lateral malleolus, and levered anteriorly to assess the stability of the peroneal tendon sheath and superior peroneal retinaculum. If the tendon sheath is detached from the lateral malleolus and therefore incompetent, the elevator will easily advance anterior to the lateral malleolus, indicating that a retinacular repair is required. A 3-cm incision is then made along the posterior margin of the lateral malleolus, exposing the tendon sheath and retinaculum. With the peroneal tendons held reduced in the peroneal groove, one to two suture anchors are placed in the lateral malleolus to secure the detached tendon sheath and retinaculum. Tendon stability is then reassessed with a Freer elevator in the same manner.

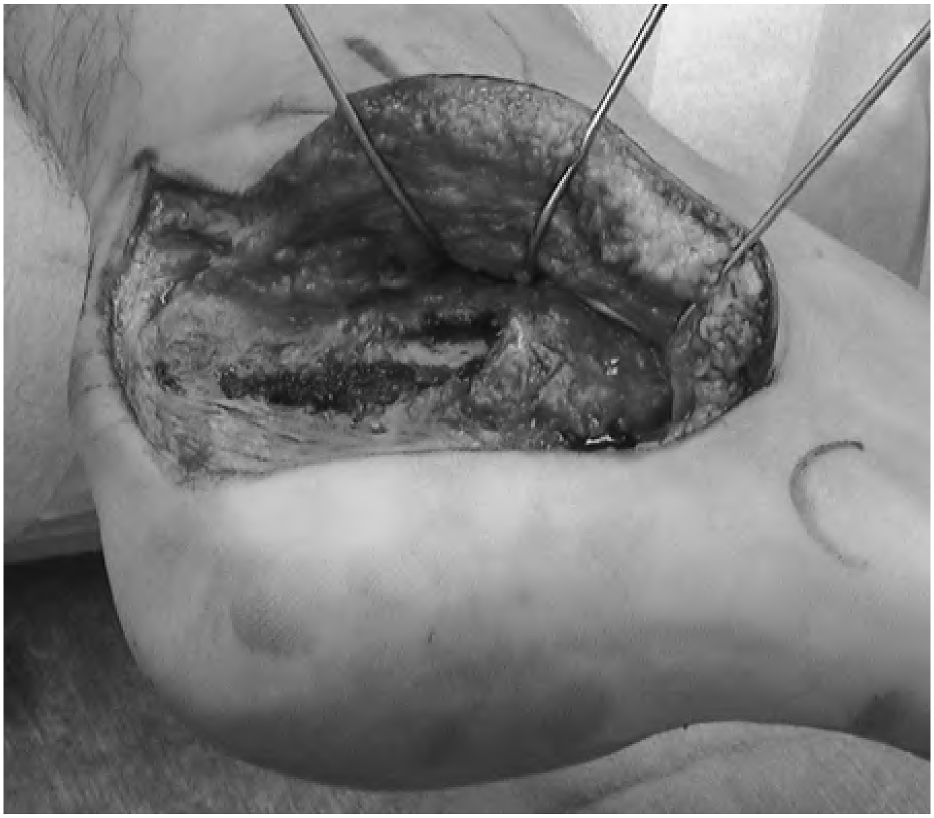

FIGURE 33-5 â€œNo touchâ€• technique: K-wire placement for retraction of the full-thickness flap.

Closure Technique

A deep drain is placed exiting proximally in line with the vertical limb of the incision. Deep No. 0 absorbable sutures are then passed in interrupted, inverted fashion starting with the apex of the incision. Sutures are placed thereafter at the proximal and distal ends of the incision, and progressing toward the apex of the incision, while attempting to advance the flap toward the apex. The suture ends are temporarily clamped until all sutures have been passed (Fig. 33-6). The sutures are then hand tied in sequential fashion, starting at the ends proximally and distally, progressing toward the apex of the incision to minimize tension at the apex.

AllgÃ¶wer-Donati Suture

The skin layer is closed with 3-0 monofilament suture using the modified AllgÃ¶er-Donati technique, again starting at the ends and working toward the apex (Fig. 33-7). The suture technique is a modified vertical mattress stitch, whereby the far end passes subcutaneous to the skin edge to minimize tension on the skin margin. Alternatively, the sutures may be passed in modified horizontal mattress fashion. In the context of an extensile lateral approach, the knots are placed along the periphery of the incision, avoiding violation of the skin margin of the subperiosteal flap.

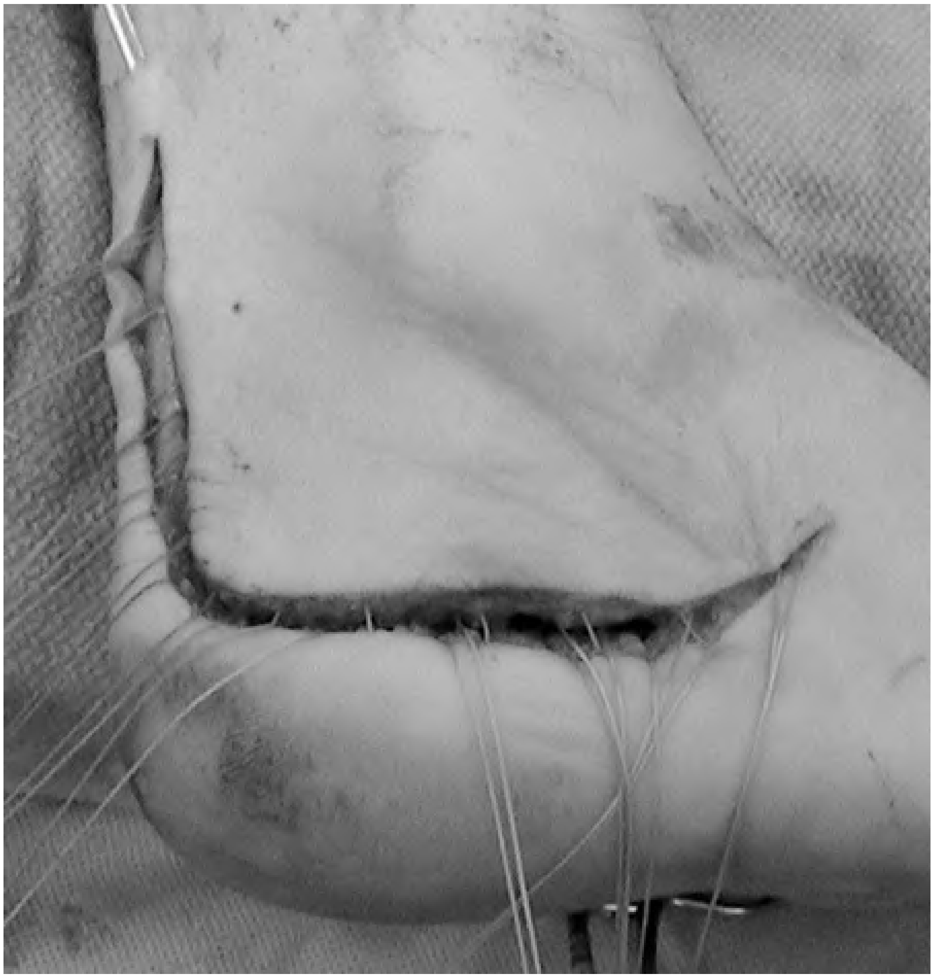

FIGURE 33-6 Deep closure. Note placement of all sutures prior to sequential tying (different patient).

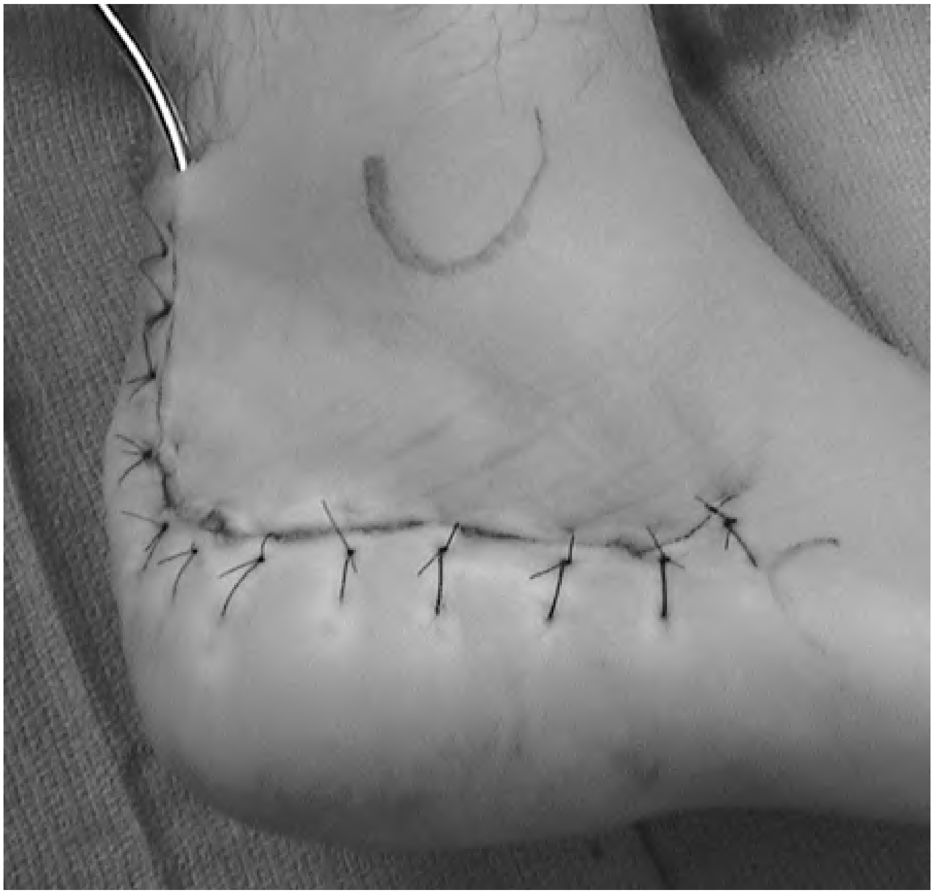

FIGURE 33-7 Skin closure using modified AllgÃ¶er-Donati technique.

The suture is placed initially into the wound edge of the near end, and the needle is then reversed in the needle driver. The suture is introduced into the subcuticular edge of the far end and extended through the subcutaneous tissue of the near end, exiting approximately 1 cm from, and in line with, the initial entry point. A single reverse throw is passed to maintain the skin bridge between the entry and exit point, and the knot is slightly tensioned to approximate the skin edges. A second reverse knot is passed to allow the knot to slide and gently evert the wound edges, followed by a forward knot to lock the stitch in place (Fig. 33-8). Following completion of wound closure, the tourniquet is deflated and sterile dressings are applied, followed by a Jones-type bulky cotton dressing and Weber splint.

Postoperative Management

The patient remains in the hospital overnight for pain control, and is converted to a short-leg non-weight bearing cast prior to discharge. The patient is converted back into an elastic compression stocking and fracture boot at 2 to 3 weeks postoperatively, and subtalar range of motion exercises are initiated. The sutures are removed once the incision is fully sealed and completely dry, typically at 3 to 4 weeks; however, they should not be removed until the wound is fully healed. We prefer that the patient sleep in the fracture boot at night until weight bearing is begun to prevent an equines contracture. Weight bearing is not permitted until 10 to 12 weeks postoperatively, at which point the fracture should be radiographically healed. The patient is then gradually transitioned to a regular shoe, and activities are advanced. In our experience, the patient should be able to return to a moderately active job at approximately 4 to 6 months postoperatively.

Complications

Delayed Wound Healing/Wound Dehiscence

Delayed wound healing or wound dehiscence is the most common complication following surgical management of a calcaneal fracture, and may occur in up to 25% of cases. Risk factors for wound complications include smoking, diabetes mellitus, open fractures, high body mass index, and a single layered wound closure. While the extensile lateral incision approximates relatively easily at the time of initial closure, wound separation may later occurâ€”typically at the apex of the incision, and even up to 4 weeks postoperatively (Fig. 33-9). The vast majority of wounds, however, will ultimately heal; deep infection and osteomyelitis develop in only 1% to 4% of closed fractures.

In the event of a wound dehiscence, all range of motion activities are discontinued to prevent further dehiscence of the wound. We prefer a fairly aggressive approach to the wound, with a management regimen of daily whirlpool treatments, damp-to-dry dressing changes, and oral antibiotics. Other granulation-promoting wound agents may also be beneficial in this instance. Alternatively, the limb may be immobilized in a short leg cast, with a window overlying the wound for access for dressing changes. These treatment regimens will typically prove successful, so long as the wound necrosis is limited to partial thickness of the skin layer. Range of motion exercises are re-initiated once the wound seals and remains dry with the patient off antibiotics.

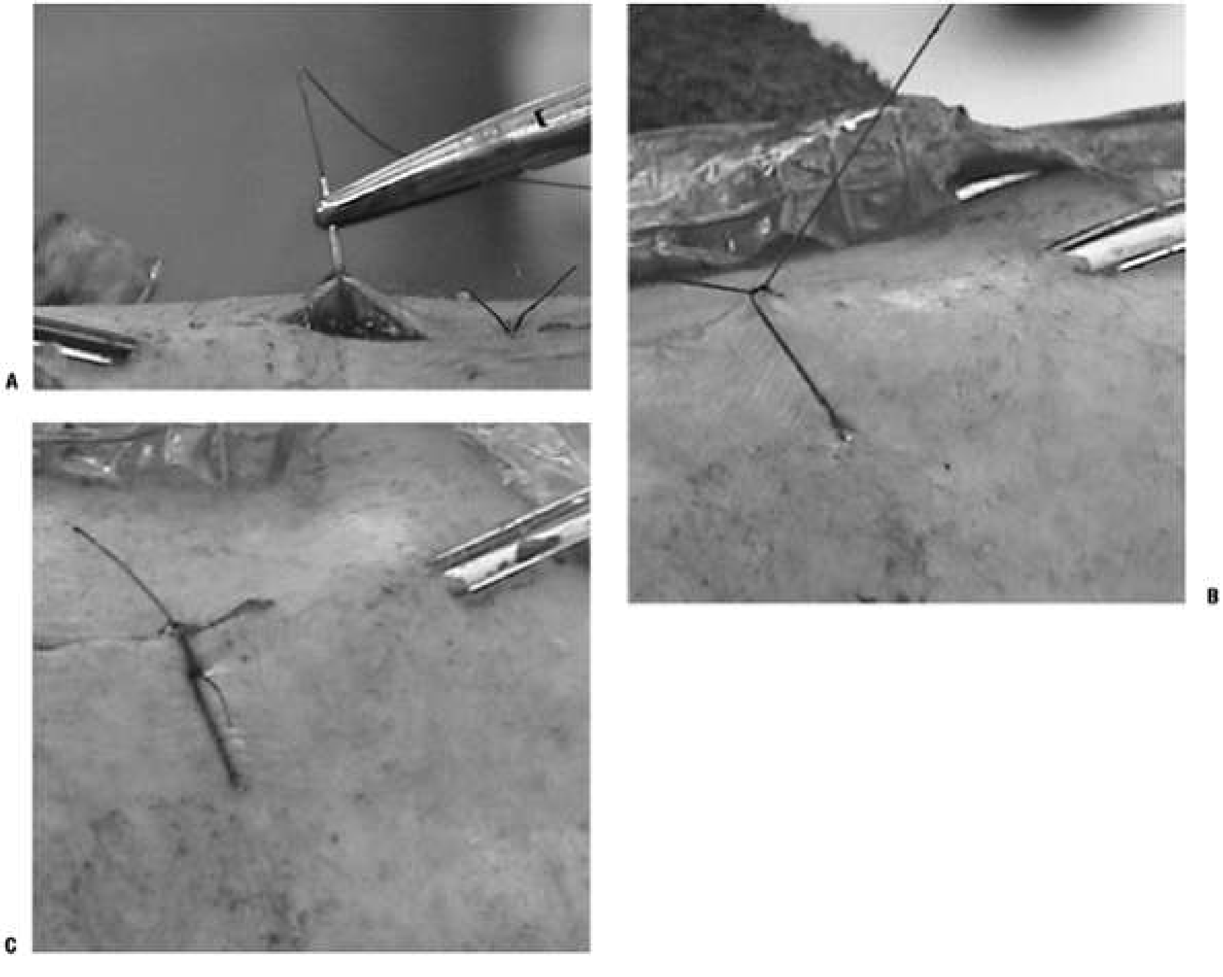

FIGURE 33-8 AllgÃ¶wer-Donati technique. Note reverse throw (B) to maintain skin bridge between suture ends.

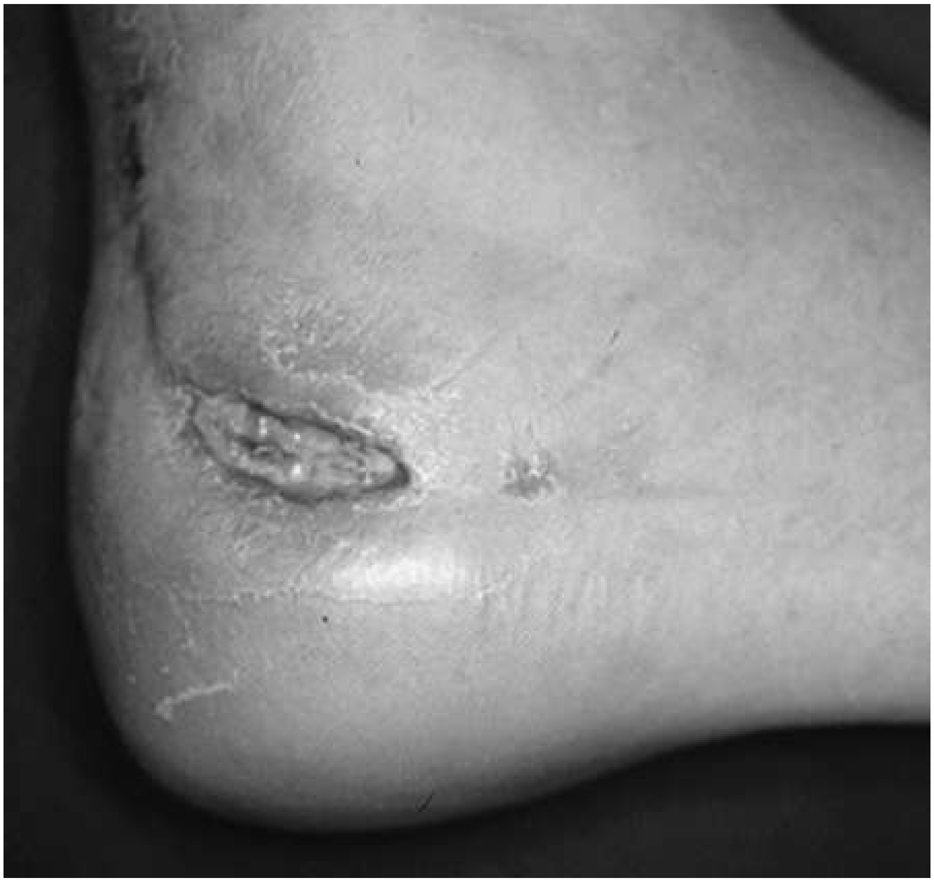

FIGURE 33-9 Apical wound dehiscence at 4 weeks postoperatively.

We prefer use of a negative-pressure device (Vacuum-Assisted ClosureÂ®, KCI, Inc., San Antonio, TX) to promote healing in the event of a recalcitrant wound; if all other treatment methods fail, a low-profile fasciocutaneous flap such as a lateral arm flap may be required for wound coverage.

Peroneal Tendon Adhesions

Peroneal tendon adhesions and scarring may develop in up to 18% of cases, either from the extensile lateral exposure itself or from prominent screwheads adjacent to the tendons, particularly surrounding the anterior process of the calcaneus. Nonoperative treatment includes tendon massage, stretching, strengthening, and other local modalities. Peroneal tenolysis or removal of the symptomatic hardware may be required in refractory cases.

Cutaneous Nerve Injury

The most common neurologic complication with surgical treatment of calcaneal fractures is iatrogenic injury to a sensory cutaneous nerve, particularly the sural nerve. Sural nerve injury occurs in up to 15% of cases and ranges from a stretch neuropraxia, which can be transient or permanent, to complete laceration of the nerve. Clinically, the patient may experience decreased or complete loss of sensation in the lateral hindfoot, or perhaps even a painful neuroma. Initial treatment includes gabapentin or amitriptyline, shoe modifications or soft accommodative inserts, and physical therapy modalities. In the event of a painful neuroma refractory to these measures, surgical neurolysis and resection may be considered, including burial of the proximal stump into deep tissues.

Fractures of the Talar Neck

Indications/Contraindications

Operative treatment is generally indicated for all displaced fractures of the talar neck, and these fractures have traditionally been considered a surgical emergency, due to the high incidence of osteonecrosis associated with displaced fractures. More recent studies have indicated that a delay in surgical management beyond 6 to 8 hours from the time of injury does not necessarily increase the risk of osteonecrosis, such that definitive stabilization within the first 24 hours following injury is now gaining acceptance as standard of care.

Severely displaced fractures or fracture-dislocations, however, may produce sufficient tension and pressure on the surrounding skin acutely to impair local circulation and result in skin necrosis and slough, potentially leading to a catastrophic deep infection and/or osteomyelitis. Thus, a timely attempt at closed reduction of the involved fracture fragment or dislocation is of paramount importance in minimizing soft tissue complications associated with these injuries, particularly if a delay in definitive stabilization is anticipated.

Nonoperative management is reserved for truly non-displaced fractures as confirmed on CT scan. Because of the high-energy nature of these injuries, displaced talar neck fractures occur most commonly in young adults. Thus in the majority of cases, the primary contraindication to surgery is the presence of severe life-threatening injuries where the patient is too medically unstable to tolerate surgery. In this instance, however, an attempt at closed reduction or manipulation of any dislocated fragments should be made in the emergency room to minimize the risk of skin necrosis.

Preoperative Planning

The surgeon should thoroughly review the plain radiographs and CT scans to gain an initial understanding of the fracture pattern, which then allows for appropriate preoperative planning, including: the surgical approach, anticipation of specific technical steps in obtaining fracture reduction, and the necessary implants for definitive stabilization. While we commonly use up to eight different surgical approaches in the management of talar fractures in general, we prefer dual anteromedial and anterolateral approaches for talar neck fractures.

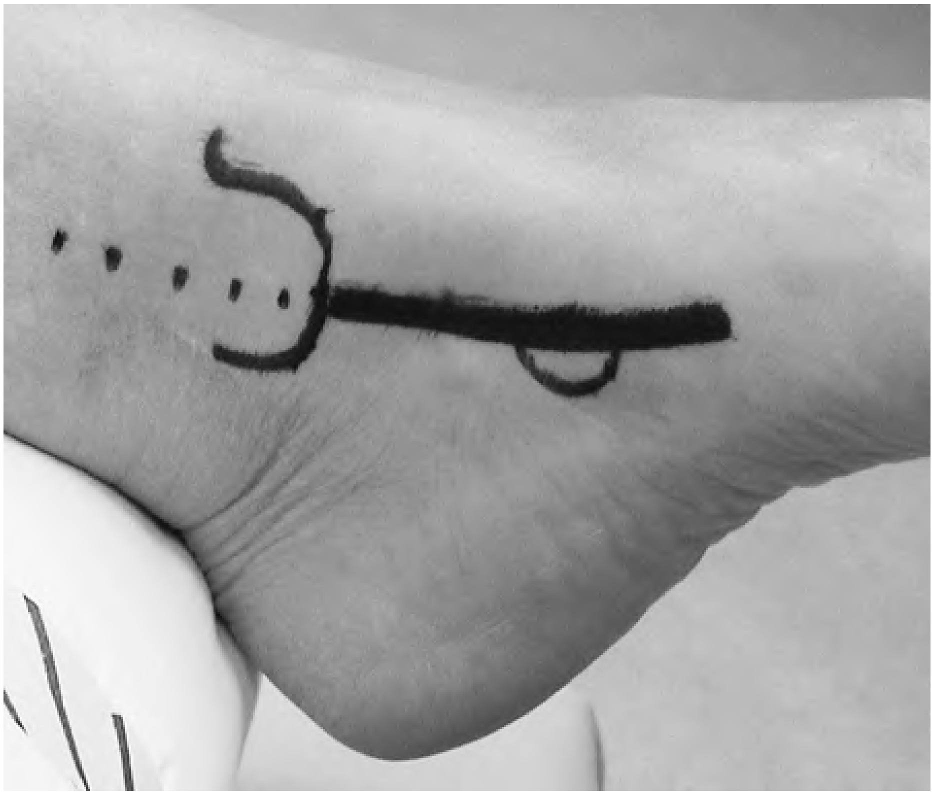

FIGURE 33-10 Planned anteromedial incision.

Surgery

Patient Positioning/General Considerations

Dual Anteromedial and Anterolateral Approaches

The procedure requires use of a radiolucent table and a standard C-arm. For isolated injuries, we prefer placing a non-sterile bolster beneath the ipsilateral hip and pelvis to allow sufficient limb exposure both medially and laterally. Protective padding is placed beneath the contralateral limb for protection of the peroneal nerve, and the contralateral limb is secured to the operating table to allow for axial plane rotation as needed for surgical exposure. The patient is given prophylactic preoperative antibiotics, and a pneumatic thigh tourniquet is employed. The procedure should be completed within 120 to 130 minutes of tourniquet time to minimize wound complications.

Technique

Incision and Surgical Approach

The dual anteromedial and anterolateral incisions are outlined on the skin with a marking pencil. Medially, the talar neck is isolated approximately a thumb-breadth distal to the anterior tip of the medial malleolus, or midway between the medial malleolus and navicular tubercle. Thus, the anteromedial incision extends from the tip of the medial malleolus in line with the medial column of the foot to a point approximately 1 cm beyond the navicular tubercle (Fig. 33-10). In this manner, the deep dissection will course through the â€œsoft spotâ€• between the anterior and posterior tibial tendons, and posterior to the saphenous nerve and vein. The approach is potentially extensile, as it can be extended proximally to allow for a medial malleolar osteotomy for talar neck fractures extending into the talar body or posteromedial process, as well as distally for access to the entire medial column of the foot as necessary (Fig. 33-11).

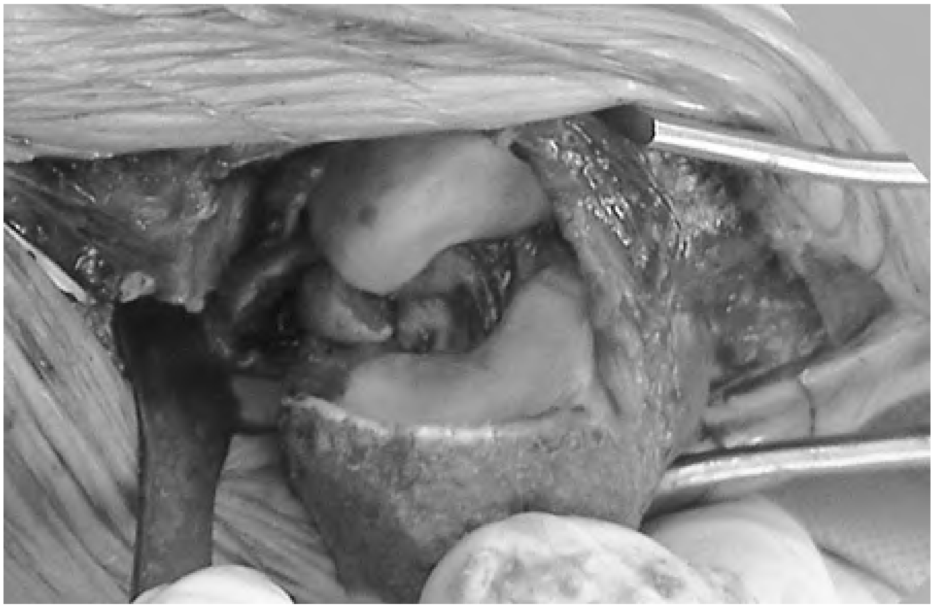

FIGURE 33-11 Medial malleolar osteotomy: anteromedial approach extended proximally for exposure of talar body (different patient).

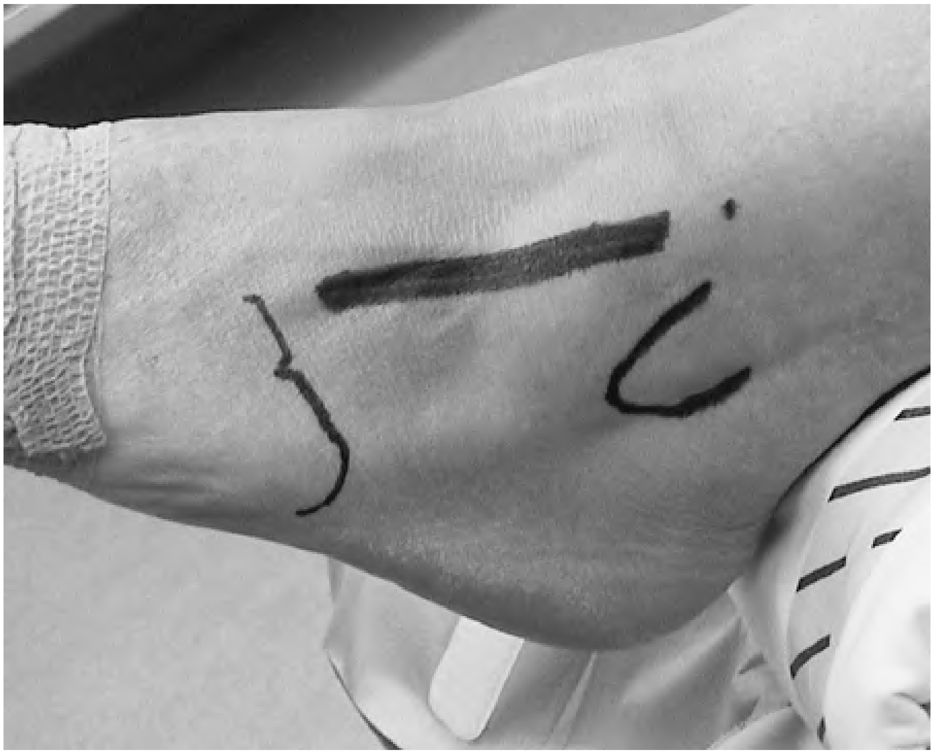

FIGURE 33-12 Planned anterolateral incision.

Laterally, the talar neck is found immediately dorsal to the sinus tarsi in line with the extensor digitorum longus and peroneus tertius tendons. The anterolateral incision actually consists of the BÃ¶hler approach, coursing from the anterolateral corner of the ankle joint in line with the extensor digitorum longus and peroneus tertius tendons toward the base of the fourth metatarsal (Fig. 33-12). This approach is also considered extensile, as it can be extended proximally to allow for exposure of the lateral talar dome, with or without a lateral malleolar osteotomy, and distally for exposure of the entire lateral column of the foot as necessary.

We prefer completing the anterolateral approach first because the majority of the comminution is typically found medially; thus, the most accurate initial indication of the extent of fracture displacement or rotational malalignment is found laterally. With sterile bolsters placed beneath the knee and ankle, the anterolateral approach is initiated, and superficial dissection continues to the extensor retinaculum and tendon sheath of the extensor digitorum longus and peroneus tertius tendons. Care is taken to avoid violation of the superficial peroneal nerve proximally, although the proximal portion of the incision rarely extends proximal enough to visualize the nerve before it begins coursing medially toward the first dorsal web space.

The tendon sheath is incised at the lateral margin of the tendons, and deep dissection is continued to the deep capsule of the ankle and subtalar joints proximally, and the extensor digitorum brevis muscle distally. The extensor brevis muscle is then traced to its origin beneath the tendons working dorsally, and subsequently reflected plantarly, thereby exposing the lateral capsule of the talonavicular joint and distal portion of the talar neck. The deep capsules of the ankle and subtalar joints are then released in line with and including the talonavicular joint capsule extending dorsally and plantarly, thereby completing a full-thickness flap (Fig. 33-13). In this manner, the foot may be adducted to expose the lateral portion of the talar head, thereby facilitating eventual placement of screws, which may be countersunk within the talar head. We make a conscious effort to limit the extent of subperiosteal dissection into the sinus tarsi, exposing only what is necessary to obtain an anatomic reduction, to minimize further compromise of the already precarious blood supply to the talar body.

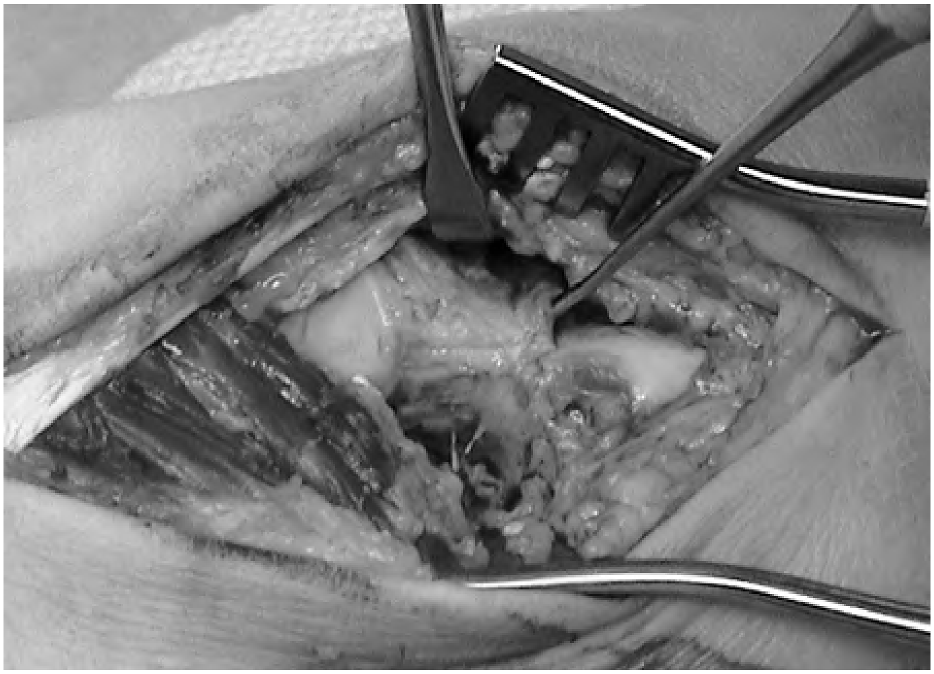

FIGURE 33-13 Anterolateral approach. Note full-thickness flap with simultaneous exposure of ankle, subtalar, and talonavicular joints.

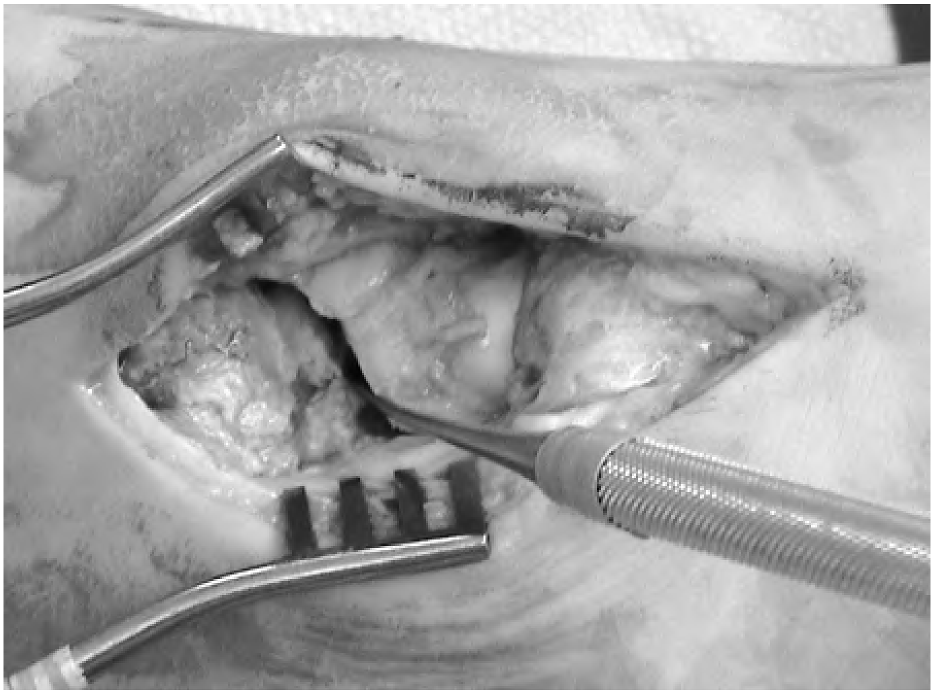

FIGURE 33-14 Anteromedial approach. Note full-thickness flap with simultaneous exposure of ankle, subtalar, and talonavicular joints.

The anteromedial approach is then initiated, and superficial dissection continues to the extensor retinaculum and deep capsules of the ankle and subtalar joints proximally, and dorsal margin of the posterior tibial tendon sheath distally. The extensor retinaculum and deep capsules are longitudinally incised, continuing along the dorsal edge of the posterior tibial tendon through the underlying talo-navicular joint capsule and spring ligament. Care is taken to avoid violation of the deltoid ligament fibers at the proximal margin of the incision.

The talonavicular joint capsule is then elevated in subperiosteal fashion off of the navicular tubercle, extending roughly to the midpoint of the navicular dorsally. The dorsal-most portion of the posterior tibial tendon insertion may also be reflected plantarly to ease soft tissue tension as necessary (Fig. 33-14). In this manner, the foot may be abducted to expose the medial portion of the talar head, again facilitating eventual screw placement. We again limit the extent of subperiosteal dissection along the undersurface of the talar neck medially in an attempt to preserve the vascular anastamoses extending into the tarsal canal. At the completion of the anterolateral and anteromedial approaches, the fracture patterns traversing the dorsal portion of talar neck should be easily visualized.

Closure Technique

Following fracture reduction, definitive stabilization, and final fluoroscopic imaging, the wound is copiously irrigated. The deep capsular layers are closed medially and laterally with interrupted No. 0 absorbable sutures placed in figure-of-eight fashion. Laterally, the extensor digitorum brevis muscle is gently approximated distally with interrupted 2-0 absorbable sutures in similar fashion, as is the extensor retinaculum and extensor digitorum longus and peroneus tertius tendon sheath more proximally. Medially, the extensor retinaculum and posterior tibial tendon sheath are closed in identical fashion with interrupted 2-0 absorbable sutures. The subcutaneous and subcuticular layers are closed with inverted, interrupted 2-0 absorbable sutures. The tourniquet is then deflated, and the skin layers are approximated with interrupted 3-0 monofilament suture, again using the modified AllgÃ¶wer-Donati technique. Sterile dressings are applied, followed by a Jones-type bulky cotton dressing and Weber splint with the ankle in neutral dorsiflexion-plantar flexion.

Postoperative Management

The patient remains in the hospital overnight for pain control, and is converted to a short-leg non-weight bearing cast before discharge. The patient is converted back into an elastic compression stocking and fracture boot at 2 to 3 weeks postoperatively, and ankle and subtalar range of motion exercises are initiated. The sutures are removed once the incision is fully sealed and completely dry, typically at 3 to 4 weeks; however, they should not be removed until the wound is fully healed. We prefer that the patient sleep in the fracture boot at night until weight bearing is begun to prevent an equines contracture.

Weight bearing is not permitted until 10 to 12 weeks postoperatively, and the patient is gradually transitioned to a regular shoe, and activities are advanced. Postoperative radiographs are carefully

scrutinized beginning at 6 weeks following surgery for the presence of a Hawkins sign

(Fig. 33-15), suggesting revascularization of the talar body. The absence of a Hawkins sign at that time, however, does not necessarily indicate osteonecrosis, as revascularization of the talar body may not occur for up to 2 years following surgery.

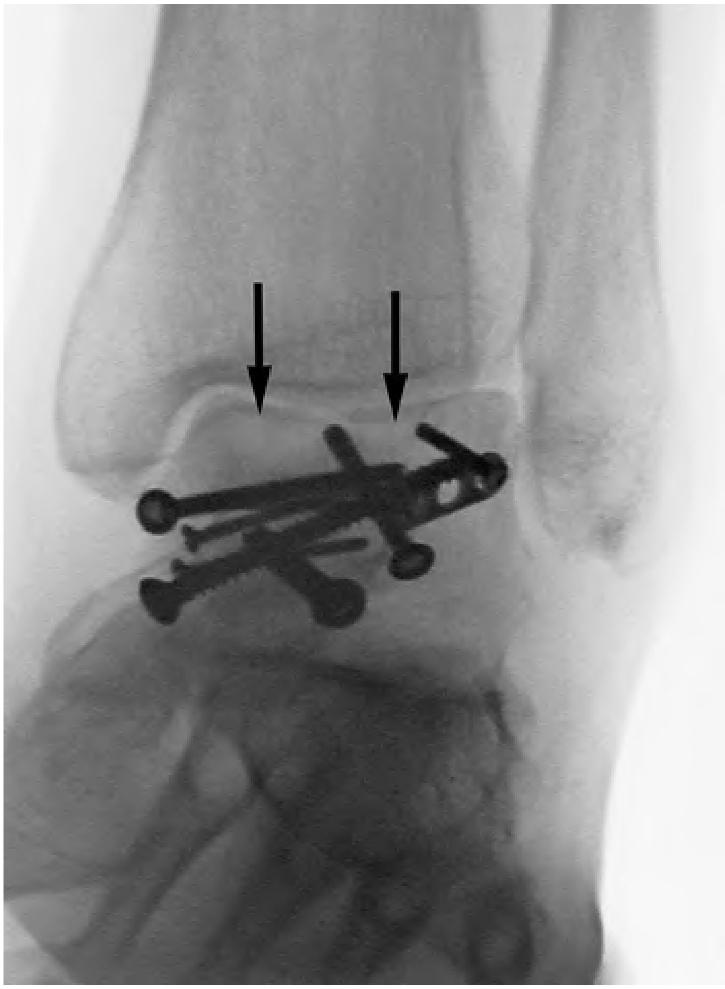

FIGURE 33-15 Hawkins sign: anteroposterior radiograph of the ankle demonstrating subchondral atrophy (arrows) at 8 weeks postoperatively following open reduction and internal fixation of a displaced talar neck fracture.

Complications

Delayed Wound Healing/Wound Dehiscence/Infection

Delayed wound healing or wound dehiscence following open reduction and internal fixation occurs in a relatively small percentage of patients, ranging from 0% to 4%. In the event of a wound dehiscence, all range of motion activities are discontinued to prevent further dehiscence of the wound. We prefer a fairly aggressive approach to the wound, with management regimen of daily whirlpool treatments, damp-to-dry dressing changes, and oral antibiotics. Other granulation-promoting wound agents, or use of a negative-pressure device may also be beneficial in this instance. Alternatively, the limb may be immobilized in a short leg cast, with a window overlying the wound for access for dressing changes. These treatment regimens typically are successful, as long as the wound necrosis is limited to partial thickness of the skin layer. Range of motion exercises are re-initiated once the wound seals and remains dry with the patient off antibiotics.

Deep infection and osteomyelitis rates are similarly low, occurring in up to 5% of closed fractures. Open injuries, however, are associated with markedly higher rates of deep infection and osteomyelitis, up to 38% in some series. Deep infection and osteomyelitis may also result from skin necrosis and slough from an unreduced closed fracture or fracture-dislocation due to excessive tension on the surrounding soft tissue envelope. Management of a deep infection includes serial surgical debridements, with local or flap coverage as needed, and long-term intravenous antibiotics. In the event of osteomyelitis, partial or complete talectomy with staged salvage arthrodesis is usually required.

Osteonecrosis

Osteonecrosis of the talar body is frequent complication following a displaced talar neck fracture, owing to the inherently tenuous blood supply. Historically, osteonecrosis rates

for displaced fractures (Hawkins type II, III, and IV fractures) treated with closed reduction and casting or pinning have ranged from 70% to 100%. Because of these factors, treatment of a displaced talar neck fracture has traditionally been considered a surgical emergency. Despite modern advances in internal fixation techniques and implants, osteonecrosis still develops in up to 30% to 50% of displaced fractures. It is well established in the literature that osteonecrosis is most related to the extent of initial fracture displacement; recent reports have suggested that the timing of

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

> Table of Contents > Part VI - Soft Tissue Management Around the Foot and Ankle > Chapter 34 - Flap Coverage for the Foot

---

Chapter 34

Flap Coverage for the Foot

L. Scott Levin

Alessio Baccarani

The goals of soft tissue reconstruction for the foot and ankle region are satisfactory wound coverage and restoration of function. Ancillary considerations include acceptable appearance and minimal donor site morbidity. For soft tissue coverage alone, muscle and axial fasciocutaneous flaps remain primary choices in the lower extremity. Random pattern cutaneous flaps and musculocutaneous flaps usually have more limited applications, but should be considered. Free flaps are generally the soft tissue coverage of choice for most extensive defects of the foot and of the lower third of the leg. Amputations and fillet flaps always represent a fourth possibile option when the limb cannot be preserved in its entirety. Finally, procedures such as osteotomies and/or ostectomies for the production of soft tissue â€œgainâ€• and resultant coverage of the defect are becoming more common.

Indications

Each anatomic region of the foot has certain characteristics that will influence selection of the flap to be transferred for reconstruction. The foot has special requirements for shoeing and ambulation. The reconstructive ladder for injury to the foot is based on whether there is a fracture, what part of the foot is exposed, and whether the area is weight bearing or nonâ€“weight bearing. The ankle and the dorsum of the foot require thin, pliable soft tissue coverage for exposed tendons, bones, or joints. The plantar skin is thick and heavily keratinized, designed to resist high stress, and anchored to underlying bones and ligaments by thick fibrous connective tissue.

Topographically, the forefoot includes the dorsal areas of the metatarsals and toes. The plantar aspect includes the metatarsal heads and the instep. The hindfoot can be divided into the plantar aspect, instep, and lateral aspect of the calcaneus. The ankle can be divided into the area of the Achilles tendon and the anterior aspect of the tibiotalar joint. In the forefoot, the dorsum and the area over the toes are primarily skin and subcutaneous tissue, making the exposure of tendons and joints more probable with high-energy injury. The plantar forefoot is prone to avulsion because of the vertically oriented septa that bridge from the plantar fascia to the dermal elements of the skin. The heel pad is a very unique structure that contains cushion-like shock-absorbing chambers of fat that are not easily replaced if loss is due to such an avulsion injury.

Achilles Tendon Area

The Achilles tendon area is characterized by thin skin with little or no subcutaneous layer. Among local flaps, we consider the sural fasciocutaneous flap one of the best options for covering this area. Other flaps options include lateral supramalleolar and lateral calcaneal artery.

Ankle and Foot Dorsum

Most shallow wounds on the foot dorsum may be safely closed with a split-thickness skin graft. If necessary, exposed extensor tendons may be resected and the skin graft may be applied to the underlying periosteum, providing a simple and quick solution to the clinical problem. For management of larger wounds with bone or tendon exposure, flaps may be required. The most common local flaps used for coverage of this area are abductor hallucisâ€“abductor digiti minimi muscle flaps, extensor digitorum brevis muscle flap, lateral supramalleolar flap, and sural fasciocutaneous flap.

Plantar Forefoot

Local flaps play a major role in the management of deep wounds of the distal third of the foot. Severe injury or infection to a single toe may be best managed by toe or ray amputation, and subsequent closure by means of a plantar or of a dorsal skin flap. If additional skin is required, an adjacent toe may be filleted and trasposed for closure. Ray amputation may be necessary if the metatarsal bone is infected, injured, or devascularized. Metatarsal head ulceration is the most frequent lesion occurring in this area, especially in patients presenting with peripheral neuropathy and associated arthropathy. Many local flaps have been described to treat plantar forefoot defects. The most commonly used are neurovascular island flap (Moberg's flap), toe fillet flap, V-Y plantar flap, and suprafascial flaps medially or laterally based.

Transmetatarsal amputation provides a functional option when three or more rays have been seriously damaged, expecially in post-traumatic, ischemic, or neuropathic patients. No prosthetic or orthotic device would be necessary, and the patient may be able to wear normal shoes. Achilles tendon lenghtening should be performed in conjunction with transmetatarsal amputation to avoid equinus deformity and stump ulceration.

Plantar Midfoot

The midfoot is defined as the region between the midshaft of the metatarsals and the proximal tarsal row. It comprises the medial non-weight bearing arch as well as the more lateral weight bearing area. Small wounds in this region may be reconstructed with a variety of reconstructive options. Split-thickness skin grafts may provide adequate coverage if the transverse arch of the foot has been maintained, thus allowing the midfoot to remain a largely nonâ€“weight-bearing region. Local flap options for reconstruction of defects in this area include neurovascular island flap, V-Y advancement flap, and medially or laterally based suprafascial flaps.

Wounds larger than 4 to 6 cm generally require either free flap reconstruction or midfoot amputation to achieve a stable coverage. Clearly amputation would represent the second option if foot salvage is not indicated or possible with free tissue transfer. The two most common forms of midfoot amputation are the Lisfranc amputation and the Chopart amputation. The Lisfranc amputation is the amputation at the tarso-metatarsal joint and is associated with a high rate of equinovarus deformity. The Chopart procedure requires an intertarsal resection, just distal to the cuboid and navicular bone. Both types of midfoot amputations will affect the patient's ability to dorsiflect and evert the residual limb because of disruption of the insertions of the peroneal and tibialis anterior tendons.

Plantar Hindfoot

Hindfoot soft tissue repair is the most challenging to the reconstructive surgeon. Reconstruction should provide durable soft tissue for safe weight-bearing, while permitting a normal ankle motion. The dualism of form and function represents a mandatory principle to be considered in the management of wounds in this area. Damage to the neurovascular and tendinous structures beneath the flexor retinaculum is an event that may impair permanently the patient's gait, mandating a below-knee amputation.

Many local flaps have been described to restore healing in the hindfoot area. The most common are intrinsic muscle flaps (abductor hallucis AH, flexor digitorum brevis FDB, abductor digiti minimi ADM); medial plantar artery flap; heel pad flaps; and sural fasciocutaneous flap. With the only exception being the sural artery fasciocutaneous flap, all the regional flaps listed require antegrade blood flow through the posterior tibial artery and its branches. Those procedures are therefore often not possible in patients affected by peripheral vascular disease.

Large hindfoot defects greater than 5 to 6 cm in patients devoid of the posterior tibial vessels should be considered for microsurgical reconstruction. Many neurosensory flaps have been described for microvascular transplantation in an effort to provide sensation to the plantar reconstruction. Lateral arm and deltoid represent two useful donor sites because of their reliable neurovascular anatomy, the ability to recover protective sensation, and because of their thickness; however, a correlation between the presence of flap sensation and the success of hindfoot reconstruction has never been established. Microvascular transplantation of muscle with skin graft coverage can provide succesful outcomes if the surgeon remembers to remove the underlying bone prominences, educates the patient on proper shoe wear, and performs frequent follow-up. We believe the use of muscle to obliterate dead space and aid in delivering antibiotics to the region is important for successful

outcomes in cases of osteomyelitis. Because of its proximity to the wound margin, the posterior tibial artery is usually the preferred recipient vessel in hindfoot reconstruction. For hindfoot coverage, it is important to avoid scarring around the posterior tibial nerve and around the posterior tibial tendon, which may become exposed or trapped during the healing process.

Table 34-1. Surgical Reconstructive Options

Foot Region	Locoregional Flaps	Free Flaps	Amputation and Ancillary Procedures
Achilles tendon region	Sural fasciocutaneous Lateral calcaneal artery	Fasciocutaneous	Syme Below knee
Ankle and dorsum	Intrinsic muscle flaps (AH, ADM, EDB) Sural fasciocutaneous Lateral supramalleolar	Fasciocutaneous Myocutaneous	Syme Below knee
Plantar forefoot	Toe fillet flap Neurovascular island flap V-Y advancement Suprafascial flaps	Fasciocutaneous Myocutaneous Muscle	Ray Transmetatarsal
Plantar midfoot	Neurovascular island flap	Fasciocutaneous Myocutaneous	Lisfranc Chopart

	V-Y advancement Suprafascial flaps	Muscle	
Plantar hindfoot	Intrinsic muscle flaps (AH, FDB, ADM) Medial plantar artery flap Sural fasciocutaneous Suprafascial flaps (heel pad flaps)	Fasciocutaneous Myocutaneous Muscle	Syme Calcanectomy

Syme's amputation has a well-established role in the management of complex hindfoot deformities, especially in diabetic patients. The procedure involves the use of the heel pad as a soft tissue cover over the distal end of the residual tibia and fibula. Table 34-1 summarizes the main surgical reconstructive options for each region of the foot.

Preoperative Planning

Evaluation of the patient with soft tissue injury should include determination of the time of injury, mechanism, energy absorption, fracture configuration, systemic injuries, damage to the soft tissue envelope, vascularity of the extremity, sensibility, ultimate ability to salvage the foot (which is both functional and sensate), and underlying medical conditions of the patient. The principles of evaluation of orthopaedic trauma are the same for any basic medical evaluation. These principles apply whether in the outpatient clinic, emergency room, or trauma unit. An evaluation of the perfusion of the traumatized limb is of paramount importance, and if vascular (arterial) injury is suspected, a vascular surgery or microsurgery consultation should be obtained. Compartment syndrome should be considered and ruled out in any injured extremity, particularly after crush injuries. A general motor examination including the active and passive range of motion as well as a detailed sensory examination should be performed. A nerve deficit may be secondary to a spinal cord injury, nerve laceration, compartment syndrome, traction injury, or entrapment between bony fragments. The radiologic evaluation starts with standard plain radiographic examination. Computed tomography (CT) is indicated in complex foot injuries and may give valuable information regarding soft tissue damage as well.

The wound should be inspected once, and the wound pattern and contamination noted. The next inspection of the wound should then be in the operating room under sterile conditions. Repetitive examination of open wounds in the emergency room has led to higher rates of wound infections and osteomyelitis and should be avoided. In cases of open fractures in polytrauma patients, workup of other injuries may take several hours, not to mention the need for emergent lifesaving visceral surgery that may precede definitive care for open fractures. Prophylactic antibiotics are administered and given on a regular basis until definitive wound debridement and fracture stabilization can be performed.

In summary, with any lower extremity reconstruction, three basic principles are of great importance and should therefore be carefully optimized before undertaking any reconstructive effort:

- Evaluation of underlying skeletal architecture, stabilization, and management of associated orthopaedic injuries
- Adequate wound preparation, which includes full debridement and control of any local infection prior to coverage
- Overall assessment of the patient's suitability for reconstruction and rehabilitation, including the opportunity of restoring some degree of protective sensation to the limb

Recipient Vessels

If free tissue transfer is required for wound coverage, then the last point of consideration prior to surgery should be the selection of recipient vessels for microvascular transfer as this will influence patient positioning within the operating room. A general agreement on which vessels to use has not yet been reached. Conflicting data have been reported on the survival and outcome of the transferred flaps, depending on the vessel used or the location of anastomosis proximal or distal to the zone of injury. For example, the anterior tibial vessels may be preferred for their easy accessibility, whereas the posterior tibial vessels are strongly advocated by others due to their larger diameter.

The most important factors influencing the site of recipient vessel are the site of the injury and the vascular status of the lower extremity; it is best to ensure adequate arterial inflow and adequate venous outflow before surgery. Intraoperatively, it is imparative that the anastomosis be performed outside the zone of injury. The type of flap used, method, and type of microvascular anastomosis represent less important factors in determining the recipient vessels.

Surgery

Locoregional Flaps

As discussed in the Indications section, locoregional flaps play a very important role into the management of foot wound coverage.

Toe Fillet Flap

Anatomy

The flap is based on the medial and lateral neurovascular bundles of the toe to be amputated.

Flap Design

The toe adjacent to the wound is outlined. This island flap is better disssected with the patient in supine position under tourniquet control.

Technique

The flap is elevated beginning distally, off the distal phalanges and flexor tendons. The medial and lateral bundles are identified in the associated web spaces. A connecting incision to the wound is made. The toe is thus disarticulated at the metatarsophalangeal joint, and the dorsal skin is used for donor site closure. The flap is then rotated to the defect, ensuring a safe placement of the neurovascular structures (Fig. 34-1).

Neurovascular Island Flap

Anatomy

The flap is based on the great toe neurovascular bundle on the fibular side.

Flap Design

The flap is designed on the fibular side, centerd over the area of the neurovascular bundle. These flaps may cover wounds up to 2 to 3 cm in diameter. The use of a tourniquet facilitates a safe dissection. The donor site often requires a skin graft closure.

Technique

The flap is elevated on the lateral plantar aspect of the great toe at the level of the phalangeal periosteum. The vascular bundle can be proximally dissected in the web space to allow a longer arc of rotation. An incision is made to connect the web space to the wound, and the flap is thus transposed to the defect.

V-Y Advancement Plantar Flaps

Anatomy

These flaps are based on vertical perforating vessels throughout the plantar aspect of the foot.

Flap Design

Many V-Y advancement flaps may be designed with different orientation on the plantar foot. The use of a tourniqut facilitates a safer dissection.

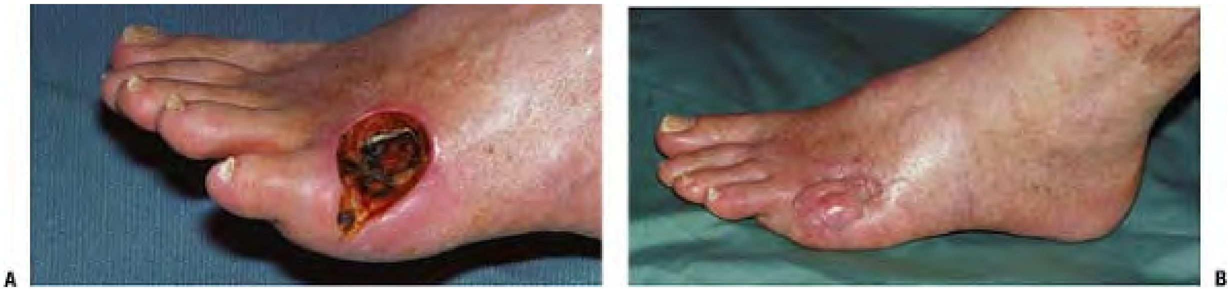

FIGURE 34-1 Dorsal forefoot diabetic ulcer covered with fifth toe fillet flap shown before (A) and after (B) surgery.

Technique

The skin ajacent to the defect is incised in a V fashion (Fig. 34-2). The plantar fascia about the circumference of the flap must be incised as well. Septal attachments to the underlying metatarsal may be divided to provide further advacement. Two V-Y opposing flaps may be combined in the management of a larger wound.

Suprafascial Flaps

Anatomy

Medially or laterally based flaps of plantar skin and fat may be advanced, rotated, or transposed to cover plantar defects. Although popular in the 1970's as random flaps, they have been largely supplanted by the other techniques described. Their vascularization is based on cutaneous branches from the medial or lateral plantar arteries (Fig. 34-3).

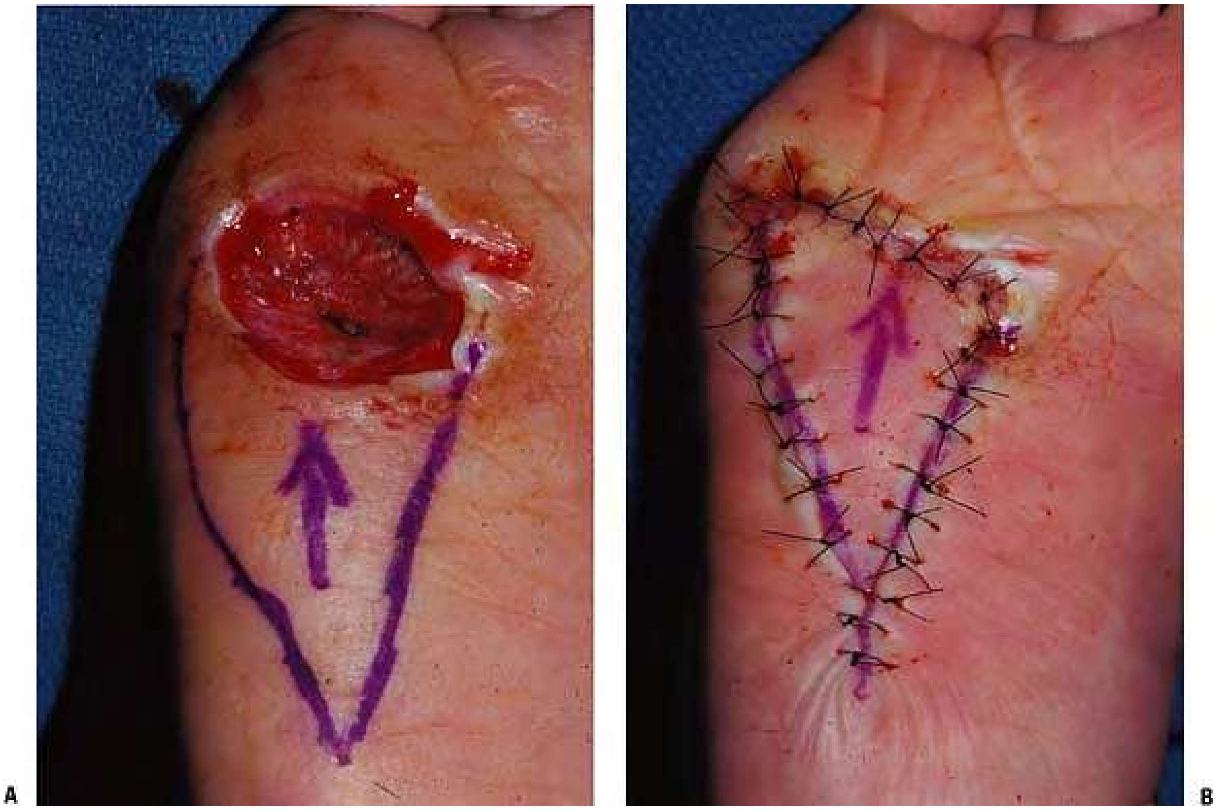

FIGURE 34-2 Plantar forefoot pressure sore in a myelodysplastic patient treated with a V-Y advancement flap shown before (A) and after (B) surgery.

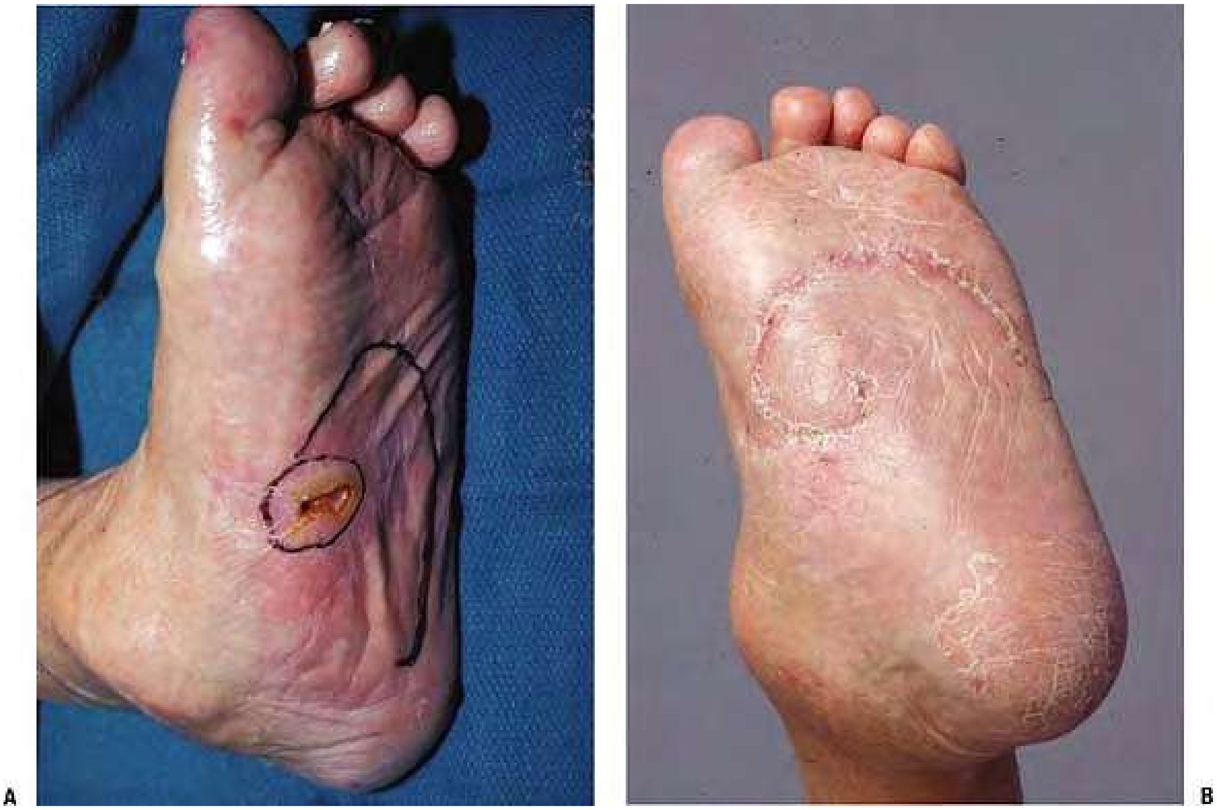

FIGURE 34-3 Plantar midfoot diabetic ulcer covered with a suprafascial rotation flap medially based shown before (A) and after (B) surgery.

Flap Design

The design of a suprafascial flap varies largely according to location of the defect. The donor site often requires a skin graft for coverage.

Technique

Medially based flaps are raised by incising laterally and elevating the subcutaneous tissues off the abductor digiti minimi muscle and plantar fascia from lateral to medial. Branches from the medial plantar nerve and artery should be preserved as they emerge from the cleft between the plantar fascia and the abductor hallucis muscle.

For the laterally based flap, the dissection is similar but requires the sacrifice of the branches from the medial plantar artery, to allow rotation or transposition.

If the flap is designed on the heel pad to cover a small wound over the weight-bearing and posterior heel, a calcanectomy may be associated to remove bone prominences and to simplify the closure.

Intrinsic Muscle Flaps: Abductor Hallucis, Flexor Digitorum Brevis, Abductor Digiti Minimi, and Extensor Digitorum Brevis

Anatomy

The abductor hallucis muscle (AH) is vascularized proximally by branches off the medial plantar artery. The flexor hallucis brevis (FHB) is also vascularized by analogue branches. The flexor digitorum brevis (FDB) is the largest of the foot muscles, and is very useful for heel pad reconstructions. It is vascularized by branches off both the medial and lateral plantar arteries, the lateral usually being dominant.

The abductor digiti minimi (ADM) is the smallest of the muscle flaps that can be harvested. It is innervated and vascularized by the lateral plantar neurovascular bundle through branches entering its proximal portion.

The extensor digitorum brevis (EDB) is a dorsal muscle. It is vascularized by the lateral tarsal artery, which is a branch off the dorsalis pedis artery, at the level of the distal edge of the extensor retinaculum.

Flaps Design

The AH muscle with or without the medial head of the FHB and the ADM can be used to close small proximal dorsal foot wounds, but it is better qualified in the management of plantar hindfoot defects. The use of an ADM flap alone is not recommended because of its small dimensions and its short arc of rotation.

The FDB is especially qualified in heel pad reconstruction.

The EDB muscle is usually transposed proximally to cover the ankle, the dorsal foot, and the malleoli provided that the anterior tibial artery has antegrade blood flow.

Technique

When harvesting the AH, a medial foot incision is placed on the nonâ€“weight-bearing surface. The tendon is devided distally, and the muscle is separated from the medial head of the FHB, if the latter component is not required to enhance the flap dimensions. If an increase in the arc of rotation is needed, the medial plantar artery is ligated and divided distal to the branches to the abductor, and a more proximal dissection of the medial plantar artery to its origin can be accomplished.

When harvesting the FDB, a midline plantar foot incision is used to expose the muscle. The skin is elevated laterally and medially, and detached from the plantar fascia. The fascia is usually elevated with the muscle to add bulk. The four tendons are divided distally and the muscle is turned on itself, after detaching it from the quadratus plantae. The lateral plantar artery may be ligated after it passes beneath the muscle, if further mobilization is required to reach the defect. The division of the origin of the AH also increases the arc of rotation of this flap.

When harvesting the ADM, a lateral foot incision is made onto the nonâ€“weight-bearing skin. The muscle is detached from the fifth metatarsal and the tendinous insertion is divided, allowing posterior rotation of the flap. Further rotation is obtained ligating the lateral plantar artery distal to the branches to the muscle and dissecting the pedicle proximally after dividing the FDB and the AH.

When harvesting the EDB, a curvilinear incision is made on the dorsum of the foot, in continuity to the defect to be repaired. The entire dorsalis pedis pedicle is divided distally to the origin of the lateral tarsal vessels to provide the needed arc of rotation for muscle transposition. The long extensors are then dissected off the underlying short extensors muscle slips. The dissection proceeds proximally ligating the medial tarsal branches. The lateral tarsal vessels are elevated with the muscle while the origin and the tendinous extensions of the muscle are divided. The muscle may thus be rotated to the defect.

Lateral Supramalleolar Flap

Anatomy

This flap is vascularized by a perforating branch off the peroneal artery as it pierces the interosseus membrane 5 cm proximal to the tip of the lateral malleolus. Cutanous vessels then course upward, anterior to the fibula, and anastomose with the vascular network that accompanies the superficial peroneal nerve.

Flap Design

The flap is qualified in the coverage of defects over the lateral malleolus and anterior ankle. The flap should be distally based, with the pedicle centered onto the perforator artery. Flap width includes the tissue between the fibula and tibia. The length should be 6 to 8 cm or more according to the defect. Often the flap is harvested only in its fascial component; then turned over to the defect in a book page fashion; then skin grafted. If designed in this way, the donor site can be closed directly.

Technique

The skin incision is made so that skin flaps may be elvated off the underlying fascia. The fascia is then incised anteriorly and reflected until the perforating branch is visualized. Branches of the superficial peroneal nerve are divided to allow elevation. The posterior margin is eventually incised and released from the septum between the anterior and lateral compartment of the leg.

Medial Plantar Artery Flap

Anatomy

This is a true neurosensory type A fasciocutaneous flap supplied by cutaneous fascicles from the medial and lateral plantar nerves. It is vascularized by the medial plantar artery and its vena comitans. The flap provides an invaluable amount of specialized skin that configures its chief value in heel reconstruction, because of its nonshearing, well-padded, adherent qualities.

Flap Design

Patency of anterior and posterior tibial arteries should be assessed prior to surgery. The presence of a Charcot deformity with midfoot collapse contraindicates the flap harvest. In this deformity, the instep area should remain covered by specialized plantar skin. For heel reconstruction, the flap is designed as an island centered on the medial plantar artery. It has to be outlined 2 to

3 cm proximal to the metatarsal heads, distal to the heel and medial to the lateral midsole weight bearing area. The donor site is usually closed with a split-thickness skin graft.

Technique

The flap is incised distally, exposing the medial plantar artery and nerves. After ligation of the artery, the dissection continues proximally beneath the plantar aponeurosis, including the medial plantar artery and the neurovascular bundles to the overlying fascia and skin. This requires an intraneural dissection of the midsole cutaneous branches from those fascicles supplying three and one-half digits. Often, the second common digital nerve is included with the branches to the flap because of difficulties encountered in separation. The flap is thus elevated in a distal to proximal direction in the plane between the plantar fascia and the first layer of muscles. Fascial comunications to the clefts between the underlying muscles (AH, FDB, and ADM) are cut.

Fascicles from the lateral plantar nerve may also be included in the flap following an intraneural dissection. The medial plantar artery and the fascicles from the medial plantar nerve are traced proximally to the AH muscle, which may be cut if a longer pedicle is required.

Sural Artery Fasciocutaneous Flap

Anatomy

This type A fasciocutaneous flap is innervated by the medial sural cutaneous nerve (S1-2). Its dominant vascular supply is a direct cutaneous sural artery branch that arises in the distal popliteal fossa between the two heads of the gastrocnemius muscle, and minor muscolocutaneous perforators from the gastrocnemius muscle. In most patients, the arterial supply will not be an identifiable vessel, but a â€œvascular networkâ€• that also anastomoses with the peroneal artery. The most relevant of these connections is located approximately 5 cm cephalad to the lateral malleolus. The lesser saphenous vein and its branches provide venous drainage.

Flap Design

The flap is centered between the popliteal fossa and the midposterior leg with a width up to 12 cm, but the length can be extended 20 cm to the Achilles tendon. The donor site is usually closed with a split-thickness skin graft.

Technique

The flap is raised from distal to proximal in the plane beneath the deep fascia and the above gastrocnemius muscle. The sural nerve and the lesser saphenous vein are divided distally and elevated with the flap. The pedicle is carefully dissected proximally, leaving abundant fascio-subcutaneous tissues around the neurovascular structures. It can be dissected up to 7 to 9 cm from the lateral malleolus, according to the location of the defect to be reached. After flap rotation to the defect, the skin over the pedicle is usually not sutured to avoid compression, and a skin graft is usually applied to provide coverage. Flap delay procedures and/or venous supercharging should be evaluated to avoid congestion, which commonly complicates the postoperative course (Fig. 34-4).

Free Flaps

Latissimus Dorsi

This flap is based on the thoracodorsal artery as the major pedicle and on branches of the intercostals and lumbar arteries as secondary segmental branches. The pedicle length is 8 to 10 cm. The latissimus is innervated by the thoracodorsal nerve, which is a direct branch of the brachial plexus and enters the muscle 10 cm from the apex of the axilla. This flap's consistent anatomy and long vascular pedicle make it a common flap choice for larger defects of the foot and Achilles tendon region. For more details on flap dissection, see Chapter 11, Pedicled and Free Latissimus Flap for Elbow and Forearm Coverage. Some of the technical problems associated with this flap for lower extremity reconstruction include difficulty in positioning the patient to allow for flap elevation and simultaneous recipient site preparation. If the anterior tibial system is to be used as a recipient vessel, the ipsilateral latissimus is usually harvested; if the posterior tibial system is to be used, the contralateral latissimus should be harvested to allow for simultaneous flap elevation and recipient site exposure. In addition, in obese patients the musculocutaneous flap may be excessively thick for smooth contour over the foot (Fig. 34-5).

Rectus Abdominis

The rectus abdominis can be harvested with the patient in a supine position. This vertically oriented muscle extends between the costal margin and the pubic region and is enclosed by the anterior and posterior rectus sheats. It is a type 3 muscle (two dominant pedicles) based on the superior epigastric artery and vein and inferior epigastric artery and vein. The pedicle length is 5 to 7 cm superiorly and 8 to 10 cm inferiorly.

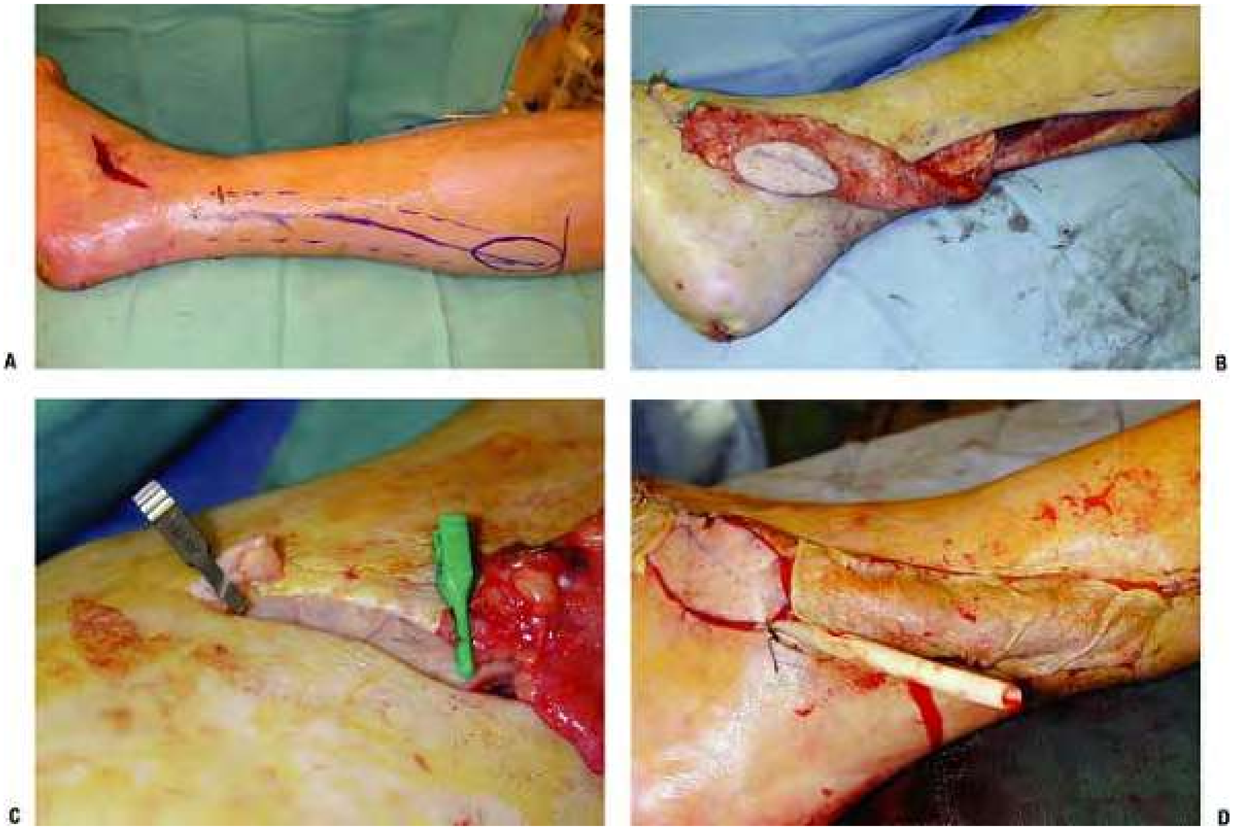

FIGURE 34-4 Wound dehiscence post triple ankle arthrodesis covered with a distally based sural flap. A: Preoperative view and flap design. B: The flap is transposed to the defect. C: The flap will be turbo-charged to improve the venous outflow (lesser saphenous vein to superficial dorsal vein). D: The pedicles are protected with split-thickness skin grafts.

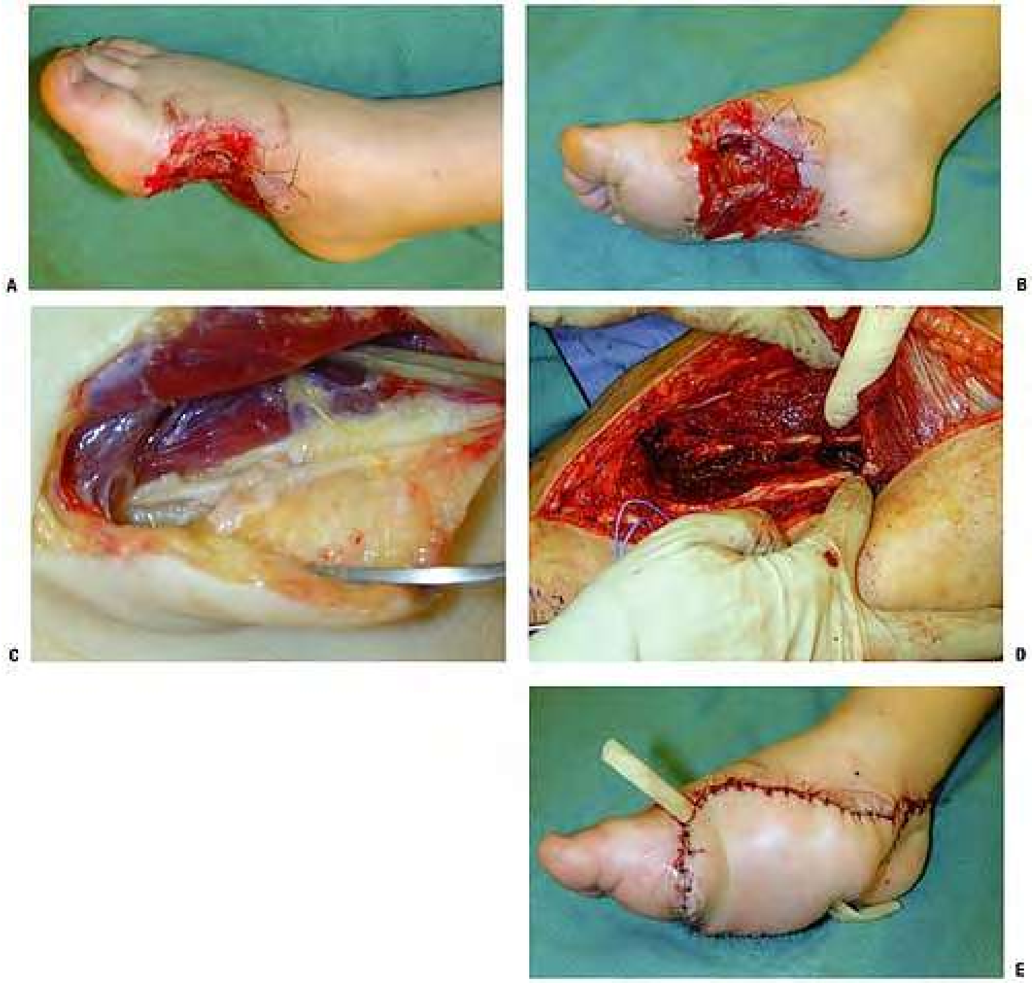

FIGURE 34-5 Medial plantar midfoot traumatic defect in a pediatric patient (lawnmower injury). The defect is reconstructed with a free latissimus dorsi flap shown before (A) and after (B) surgery. Medial plantar midfoot traumatic defect in a pediatric patient (lawnmower injury). The defect is reconstructed with a free latissimus dorsi flapshown before (A) and after (B) surgery. C,D: Latissimus dorsi flap harvesting. E: The flap is revascularized and sutured to the defect.

Each of the dominant pedicles supplies just over one-half of the muscle. There is an anastomosis between these vessels that is usually sufficient to support the nondominant half if one of the two pedicles is ligated. Because of the larger size and easier dissection of the inferior epigastric vessel, it is usually used for free tissue transfer.

The motor innervation is supplied by segmental motor nerves from the seventh through twelfth intercostal nerves that enter the deep surface of the muscle at its middle to lateral aspects. The lateral cutaneous nerves from the seventh through twelfth intercostal nerves provide sensation to the skin territory of the rectus abdominis muscle. The size of the muscle is up to 25 Ã— 6 cm^2. The skin territory that can be harvested is 21 Ã— 14 cm^2 and is based on musculocutaneous perforatore (Fig. 34-6).

Gracilis

The gracilis is a smaller transplant and is useful for defects requiring less bulk than the latissimus or rectus. The gracilis muscle is a type 2 muscle (with a dominant pedicle and several minor pedicles). It is a thin, flat muscle that lies between the adductor longus and sartorius muscle anteriorly and the semimembranosus posteriorly. The dominant pedicle is the ascending branch of medial circumflex femoral artery and venae comitantes. The length of the pedicle is 6 cm and the diameter of the artery is 1.6 mm. The minor pedicles are one or two branches of the superficial femoral artery and venae comitantes. Their length is 2 cm and their diameter is 0.5 mm.

Motor innervation is via the anterior branch of the obturator nerve, which is located between the abductor longus and magnus muscles, and it usually enters the muscle above the level of the dominant vascular pedicle. The anterior femoral cutaneous nerve (L2â€“3) provides sensory innervation to the majority of the anterior medial thigh.

This muscle functions as a thigh adductor. The presence of the adductor longus and magnus makes it an expendable muscle.

The size of the muscle is 6 Ã— 24 cm^2. The skin territory is 16 Ã— 18 cm^2, but the skin over the distal half of the muscle is not reliable when the flap is based on its dominant vascular pedicle with division of the minor vascular pedicles. In obese patients, the musculocutaneous flap may be too bulky, necessitating use of a skin graft placed on the muscle (Fig. 34-7) (see Chapter 32).

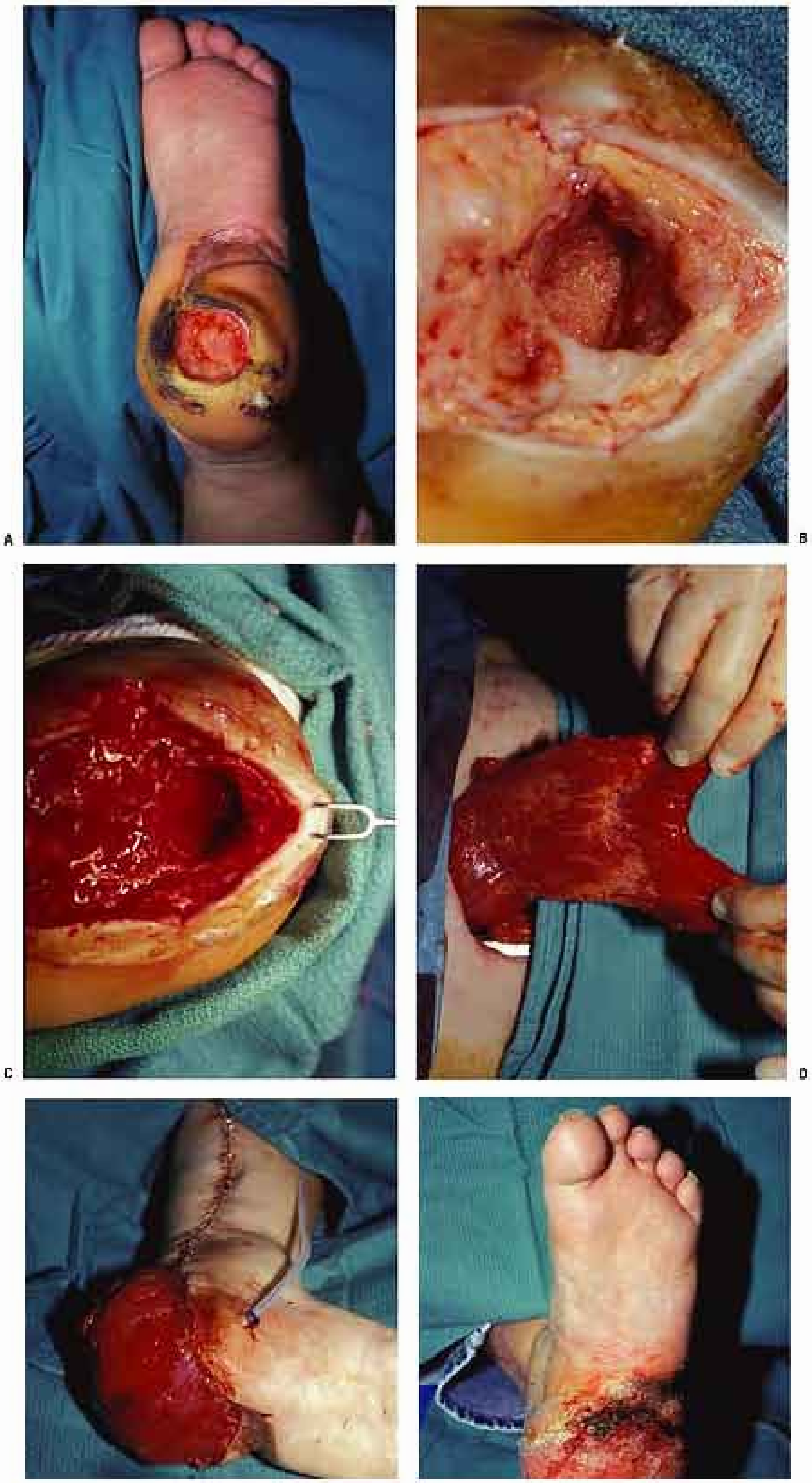
A
B
C
D

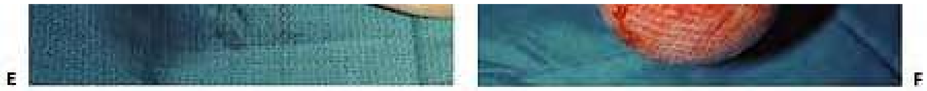

FIGURE 34-6 Calcaneal osteomyelitis on a diabetic foot. The defect was reconstructed with a free rectus abdominis flap. A,B: Preoperative view. C: After debridement. D: The rectus muscle is harvested from the abdomen. E: The flap is tailored to the defect and revascularized. F: Late postoperative view showing calcaneal salvage.

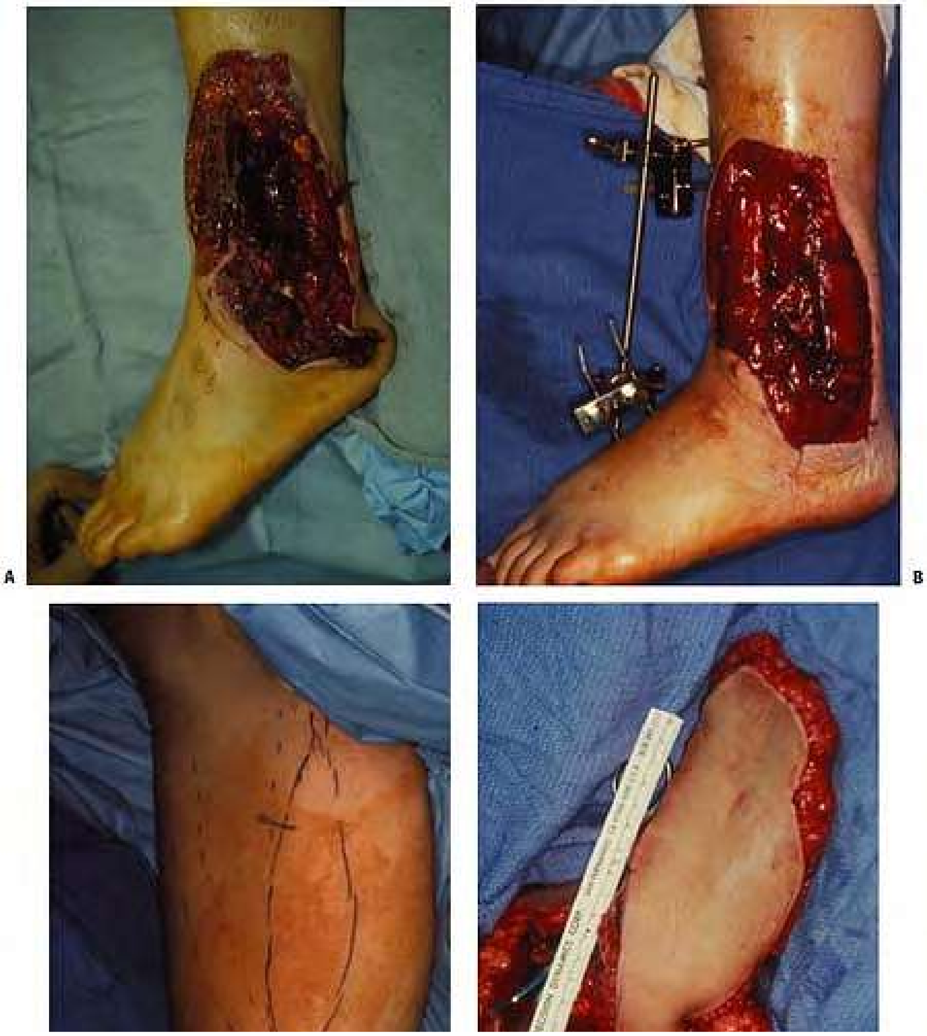

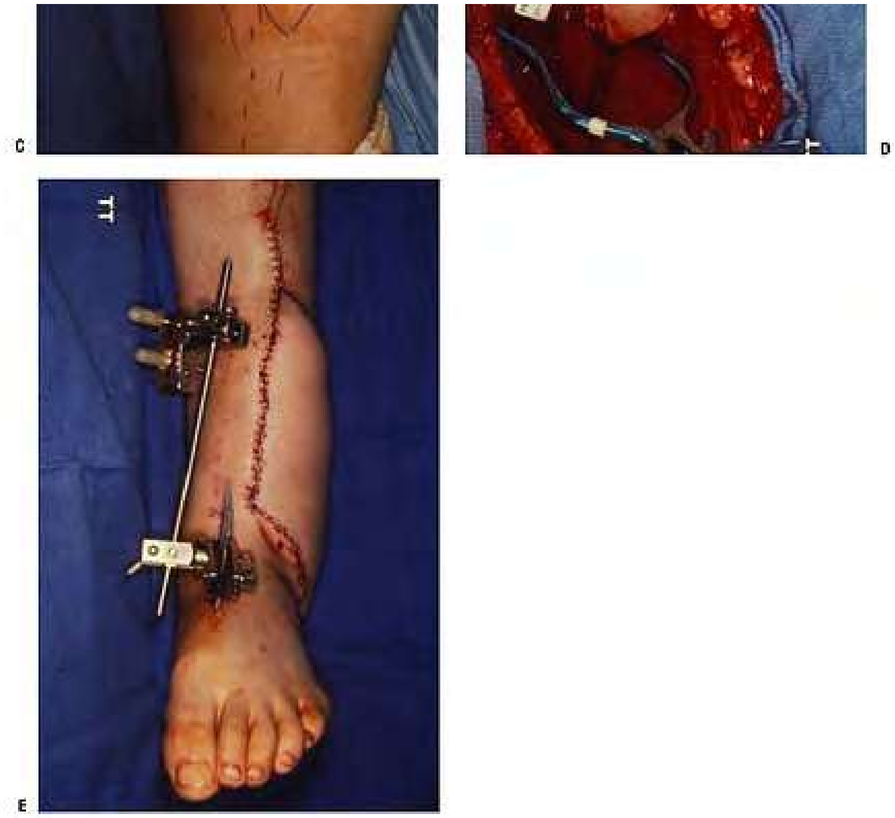

FIGURE 34-7 Gunshot wound to the distal leg-hindfoot. The defect is reconstructed with a free gracilis flap. A: Significant soft-tissue and bony defects are present. B: Debridement and external fixation. C: A myocutaneous free gracilis flap is planned. D: The free flap is harvested. E: The flap is inset to the defect and revascularized.

Radial Forearm Flap

This is a thin, well-vascularized fasciocutaneous flap on the ventral aspect of the forearm that was widely used in China before it was popularized in the Western literature. The flap is based on the radial artery, which can achieve a 20-cm pedicle and has a diameter of 2.5 mm. This length of the pedicle facilitates the microsurgical anastomosis out of the zone of injury. The venous drainage is through the venae comitantes of the radial artery, but the flap can include the cephalic vein, the basilic vein, or both. The flap can contain the lateral antebrachial cutaneous nerve or the medial antebrachial cutaneous nerve and then serve as a neurosensory flap. The size of the flap can be 10 Ã— 40 cm^2.

A portion of the radius can be included as a vascularized bone with this flap. The advantages of this flap are a long pedicle and potential sensory innervation. The quality of the bone from the radius is mainly cortical and not of any substantial volume. Including the bone in the radial forearm flap may lead to stress fracture of the donor radius. Preliminary tissue expansion will increase the flap dimensions, and more importantly, it will allow direct closure of the donor defect (Fig. 34-8).

Scapular and Parascapular Flap

The scapular flap remains the workhorse of skin flaps. It is a thin, usually hairless, skin flap from the posterior chest and can be de-epithelialized and used as subcutaneous fascial flap, pedicled or free.

The flap is perfused by the cutaneous branches of the circumflex scapular artery (CSA) and drained by its venae comitantes. The CSA is the main branch of the subscapular artery and the main blood supply to the scapula, the muscles that attach to the scapula, and the overlying skin. The length of the pedicle is 5 cm and the diameter of the artery is 2.5 mm. The vascular pattern of this territory makes it possible to raise multiple skin flaps on a single vascular pedicle or to harvest the lateral border of the scapula as an osteocutaneous flap for a complex reconstruction.

The cutaneous territory can be 20 Ã— 7 cm^2 and can be divided in two componentsâ€”a horizontal territory (horizontal scapular flap) and a vertical territory (parascapular flap)â€”based on the branches of the CSA after the vessel courses through the triangular space. Preliminary expansion of the territory of the scapular flap will increase the flap dimensions and permit direct donor-site closure.

This flap can be combined with other flaps based on subscapular blood supply and may greatly facilitate certain complex reconstructions. These include the latissimus dorsi and serratus anterior flaps, which can supply additional skin, muscle, and bone (rib) if necessary. The primary indication for the scapular flap is a defect requiring a relatively thin, large cutaneous flap. These kinds of defects are often found in the foot. The osteoseptocutaneous free scapular flap reconstruction has been described in the lower extremity (Fig. 34-9).

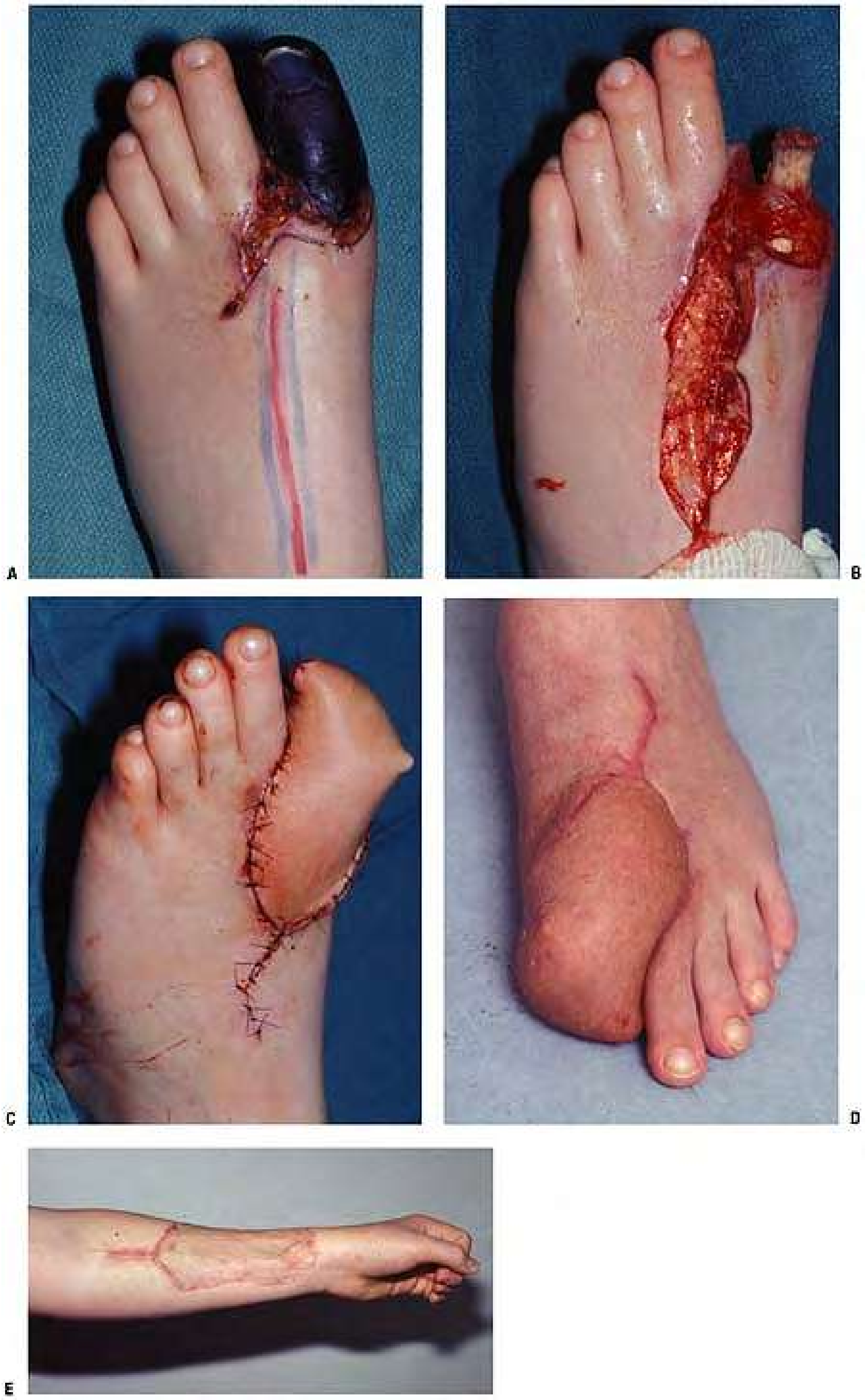

FIGURE 34-8 Post-traumatic left great toe necrosis. After debridement the soft tissue defect is restored with a free radial forearm flap. A: Preoperative view. B: After debridement. C: Radial forearm flap is tailored to cover the exposed phalanx. D,E: Four weeks postoperative showing toe salvage and flap donor site.

Postoperative Management

The success of a reconstructive foot surgery importantly relies on well-planned multidisciplinary postoperative management and rehabilitation. A nonâ€“weight-bearing regimen for at least 3 weeks is mandatory for every patient carrying plantar sutures. Elevation is required for 3 to 8 days after a local flap or a skin graft, and for 2 to 3 weeks after a free flap. Heparin regimen is advisable while the patient is in the aforementioned bed rest phase. Elastic wraps may be useful in controling edema once the patient's limb is permitted in a dependent position. Clinical findings should dictate the use of antibiotics.

The Lâ€™Enard splint (Fig. 34-10) is a useful tool to provide immobilization of the foot and ankle, and in keeping the posterior heel off the bed. It is also useful when the patient begins nonâ€“weight-bearing ambulation with crutches. Local care of all weight bearing surfaces is also recommended after suture removal, together with a frequent multidisciplinary follow-up that involves the plastic surgeon, the orthopaedic surgeon, the podiatrist and, if required, the prosthetist.

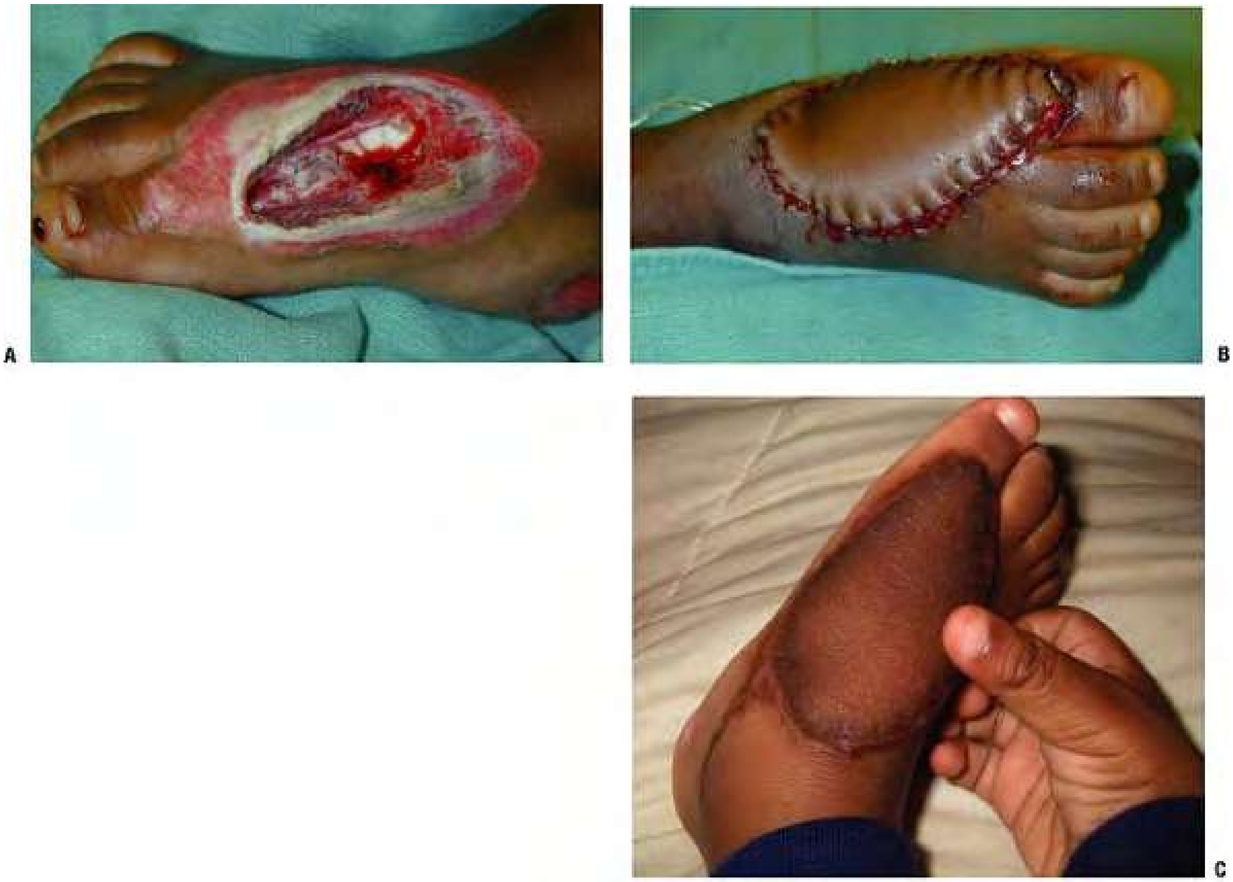

FIGURE 34-9 Dorsal foot avulsion injury in a pediatric patient (go-cart trauma). The dorsal aspect of the foot is reconstructed with a free scapular-parascapular flap. A: After debridement. B: Free scapular-parascapular flap is inset to the defect and revascularized. C: Two months postoperative view.

Postoperative care of free tissue transfer patients requires the patients to be adequately hydrated. Maintenance of proper body temperature and hematocrit is also important. Routine heparinization and anticoagulation are not used.

Flaps are usually monitored for a minimum of 5 days with a laser Doppler in addition to clinical observation. While the immediate postoperative period of 24 to 48 hours is critical, there have been occasional late failures; thus, laser Doppler monitoring should be continued for 4 or 5 days.

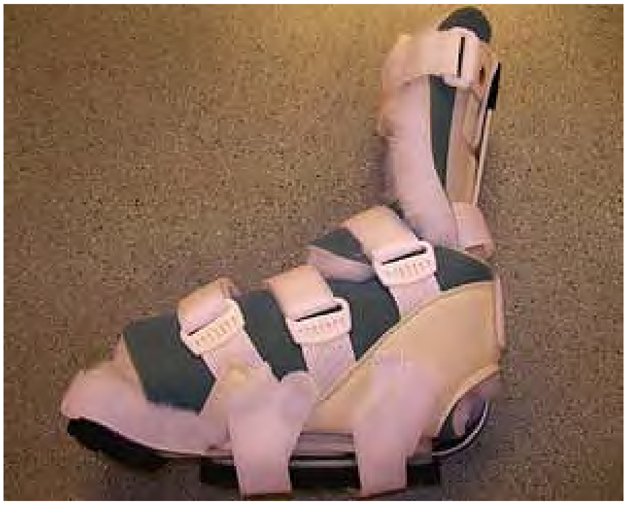

FIGURE 34-10 Lâ€™Enard splint.

Complications

Any flap failure requires a new detailed evaluation of the patient's local and general conditions before undertaking a new reconstruction. If a locoregional flap fails, and no other local option is available, the next step on the reconstructive ladder would be a free flap.

The success of free tissue transfer should be on the order of 95% to 99%. Acute complications usually occur in the first 48 hours and include venous thrombosis, arterial thrombosis, hematoma, hemorrhage, and excessive flap edema. Arterial insufficiency can be recognized by decreased capillary refill, pallor, reduced temperature, and the absence of bleeding after pinprick. This complication can be caused by arterial spasm, vessel plaque, torsion of the pedicle, pressure on the flap, technical error with injury to the pedicle, a flap harvested that is too large for its blood supply, or small vessel disease (due to smoking or diabetes). Management of arterial compromise requires prompt surgical intervention to restore the blood flow. Pharmacologic intervention at the time of exploration includes use of vasodilators, calcium channel blockers, and systemic anticoagulants for flap salvage presenting with arterial insufficiency. Ultimately, if these pharmacologic agents do not relieve spasm at the level of the arterial inflow, the anastomosis should be redone to rule out intra-arterial thrombus.

Venous outflow obstruction can be suspected when the flap has a violaceous color, brisk capillary refill, normal or elevated temperature, and production of dark blood after pinprick. Venous insufficiency can occur due to torsion of the pedicle, flap edema, hematoma, or tight closure of the tissue over the pedicle. The venous outflow obstruction can result in extravasation of red blood cells, endothelial breakdown, microvascular collapse, thrombosis in the microcirculation, and finally flap death. Given the irreversible nature of the microcirculatory changes in venous congestion that occur even after short periods of time, the surgeon must recognize venous compromise as early as possible.

These complications can occur alone or in any combination. The clinical observation and the monitoring of the patient (such as with laser Doppler) should alert the surgeon who has to decide between conservative and operative intervention. Conservative treatment may include drainage of the hematoma at the bedside with release of a few sutures to decrease pressure. In cases of venous congestion, leeches may be helpful if insufficient venous outflow cannot be established despite a patent venous anastomosis. The leeches inject a salivary component (hirudin) that inhibits both platelet aggregation and the coagulation cascade. The flap is decongested initially as the leech extracts blood and is further decongested as the bite wound oozes after the leech detaches.

The donor site should be given the same attention as the recipient site during the postoperative period. Complications of the donor site include hematoma, seroma, sensory nerve dysfunction, and scar formation.

Occasionally free flaps, despite early return to the operating room for vascular compromise, do fail. Options for management include the performance of a second free tissue transfer, noting the technical or physiologic details that led to initial failure. Most of the time, free tissue transfers that fail are due to technical errors in judgment, whether they be flap harvest, compromise of the pedicle during the harvest, improper microvascular technique during anastomosis, improper insetting resulting in increased tissue tension and edema, or postoperative motion of the extremity resulting in pedicle avulsion. The next decision made by the operating surgeon as to the management of this patient is based on several factors. If a patient required a free flap in the first place, a second free flap should be considered. If a decision is made not to redo the flap, it could be left in place using the Crane principle to see if underlying granulation will be sufficient such that skin grafting can be performed once the necrotic flap is removed.

The Crane principle can be applied to cases where a local flap or free tissue transfer that necrotizes in part or totally acts as a biologic dressing or eschar over a wound bed. If there is no infection, the eschar can be left on the wound bed to see if some healing in the form of granulation occurs underneath it. Ultimately, the eschar is removed and the granulation bed skin grafted, obviating another free tissue transfer. If wound observation shows that such a bed is not produced, then a second flap must be considered.

It is usually our preference not to follow this course, as the flap can become a source of sepsis and further compromise local tissues. Necrotic nonviable flaps should be removed and a temporary wound dressing such as an antibiotic bead pouch or wound vacuum-assisted closure (VAC) should be used. Occasionally when flaps fail in a severely compromised extremity, consideration should be given to amputation. If a second free flap is considered, errors that lead to flap compromise need to

be recognized and avoided. It may be prudent to obtain an arteriogram, evaluate the coagulation profile, and research other issues that might have led to failure.

Recommended Reading

Arnold PG, Yugueros P, Hanssen AD. Muscle flaps in osteomyelitis of the lower extremity: a 20-year account. *Plast Reconstr Surg*. 1999;104:107.

Attinger CE, Ducic I, Zelen C. The use of local muscle flaps in foot and ankle reconstruction. *Clin Podiatr Med Surg*. 2000;17(4):681.

Colen L, Uroskie T. Foot reconstruction. In: Mathes SJ, ed. *Plastic surgery.* 2nd ed. Vol.6. Philadelphia: Elsevier; 2006.

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 35

Soft Tissue Coverage of the Pelvis and Sacrum: Hemipelvectomy and Pedicled Flap Coverage

Alex Senchenkov
Robert Esther
Franklin H. Sim
Steven L. Moran

Extensive defects of the pelvis and sacrum can result from tumors, ablation, and severe trauma. Due to limited amounts and relative immobility of the pelvic soft tissue, these defects may pose a serious reconstructive challenge. Until the late 1970s, the majority of large pelvic tumors were treated with external hemipelvectomy. Advances in imaging, chemotherapy, and radiation therapy, as well as improvements in resection and reconstructive techniques, have greatly reduced the need for radical lower extremity amputations, allowing limb preservation in a majority of cases.

Historically, buttock tumors were not amenable to a classic hemipelvectomy and just a few decades ago were considered unresectable. Likewise, extensive buttock defects inflicted by trauma, infection, or end-stage pressure ulcers in paraplegics could not be effectively reconstructed. Secondary intention healing frequently resulted in protracted hospital course, extensive scarring, contractures, and unstable soft tissue coverage. Many of these patients were bound to years of ongoing wound care and immobility.

External hemipelvectomy denotes removal of the hemipelvis with affected lower extremity by disarticulation of the pubic symphysis and the sacroiliac joint. Because external hemipelvectomy resulted in major functional impairment, limb-sparing procedures removing part or all innominate bone with preservation of the extremity have been advocated. These pelvic resections are referred to as internal hemipelvectomies.

Large, composite pelvic defects associated with internal hemipelvectomies are more challenging to reconstruct than the soft tissue defect typically created in external hemipelvectomy patients for two main reasons. First, following removal of bony hemipelvis in external hemipelvectomy, a large amount of soft tissue of the buttock or proximal thigh becomes available for reconstruction. Second, a decrease of the pelvic volume obliterates the dead space.

Sacral resections are performed as a part of extended external hemipelvectomy for musculoskeletal sarcomas and, as such, reconstructed as a part of hemipelvectomy closure. Isolated sacral defects result from composite pelvic resections for locally advanced anal and rectal malignancies or tumors intrinsic to the sacrum such as sacral chordomas and sarcomas.

Indications/Contraindications

When embarking on treatment of pelvic sarcomas, three important questions should be borne in mind.

- Is this patient operable, i.e., can the individual medically withstand a major oncologic resection?
- Is this tumor resectable, i.e., can this patient be rendered disease-free surgically?
- Can the residual defect or deformity from the proposed resection be reconstructed in a functionally satisfactorily manner with stable soft tissue

coverage?

The answers to these questions have to be determined during preoperative evaluation by the surgical oncologist, reconstructive surgeon, and anesthesiologist. Resection of the tumor with negative margins is the only reliable means of obtaining a cure in cases if tumor.

Internal hemipelvectomy is indicated in cases of localized tumor where margin negative resection of the tumor is possible with preservation of the lower extremity. If clean margins cannot be achieved, external hemipelvectomy should be performed. Main indications for external hemipelvectomy are: large tumors involving multiple compartments unresponsive to neoadjuvant therapies, contamination of compartments from pathologic fracture, or failed previous resection, a nonviable extremity. Nononcologic external hemipelvectomy may be performed in the cases of uncontrolled pelvic osteomyelitis, traumatic hemorrhage, and failed aorto-femoral revascularizations (Table 35-1).

Wound complication rates following hemipelvectomy are notoriously high and have been reported to range from 20% to 80%. Proper technical execution of the procedure and the use of well-designed skin and muscle flaps can minimize postoperative wound morbidity.

Although infrequently, pelvic and sacral resections are performed en bloc with pelvic visceral structures for locally advanced rectal and gynecologic malignancies eroding or invading the skeletal pelvis. When such pelvic resection involves removal of a part of the pelvis or sacrum, it is referred to as *composite resection*. Any type of external hemipelvectomy performed in continuity with

visceral structures is known as *compound hemipelvectomy*. Due to the aggressive nature of these tumors, the disease has to be limited to the pelvis and extensive imaging is required to select the patients that can benefit from these extensive operations.

Table 35-1. Basic Tumor Flap Principles

- Safe oncologic margins are the primary requirement
- Reconstruction does not take precedence over adequate, safe resection
- Adequate soft tissue coverage of bony reconstruction/prosthesis and neurovascular structures
- Healed surgical wound
- Durable
- Minimal donor morbidity
- Appropriate function, contour

Primary sacral tumors such as chordomas and sarcomas are relatively uncommon. The majority of these tumors are low-grade malignancies. They infrequently metastasize and therefore local control becomes important. Reconstruction of these defects with flaps facilitates optimal postoperative wound healing.

Preoperative Planning

Prior to surgical resection, patients should undergo local and systemic staging studies. Musculoskeletal malignancies have a propensity to pulmonary spread. Therefore, a chest CT is mandatory to screen for systemic disease. An MRI (and plain radiographs for primary bone tumors) is sufficient for gauging the local extent of disease and response to treatment. A CT of the pelvis is often useful to complement the MRI as this area is difficult to image. Surgical planning relies on MRI images taken before and after neoadjuvant therapies. Pretreatment MRI images may be helpful in distinguishing radiation-induced reactive changes from actual tumor tissue.

Most patients with high grade bone malignancies will undergo some form of neoadjuvant treatment, including chemotherapy and/or radiation therapy prior to tumor resection. Typically, primary bone sarcomas such as osteosarcoma are treated with several cycles of preoperative chemotherapy, surgery, and then several additional cycles of chemotherapy. Radiation therapy also has an established role in treatment of soft tissue sarcomas.

The treatment team must choose between pre- and postoperative radiation therapy. Both approaches have advantages and shortcomings. Preoperative treatment requires a smaller area of treatment, creation of a fibrous rind around the tumor, and often causes tumor shrinkage, leading to an improved ability to obtain wide margins without sacrificing vital structures. Preoperative radiation's disadvantages include a higher rate of wound problems and less viable tumors available for pathologic examination. Postoperative radiation has the advantage of earlier surgery, viable tumors for pathologic study, and fewer wound complications. Treatment volumes however are increased and there is a delay in administering treatment to allow time for adequate healing of operative wounds. We prefer preoperative radiation for pelvic and retroperitoneal sarcomas.

Brachytherapy requires proper reconstruction planning so that flaps do not interfere with catheter placement (Fig. 35-1). Afterloading catheters should be evenly spaced and sutured in place to the tumor bed with fast absorbable sutures to prevent their displacement during postoperative therapy. Alternatively, VAC dressing can be used as a temporary coverage of brachytherapy catheters, followed by delayed primary reconstruction of the defect after completion of brachytherapy. Intraoperative radiation therapy (IORT) is another means of augmenting a preoperative radiation therapy regimen, allowing for directed treatment at close intraoperative margins.

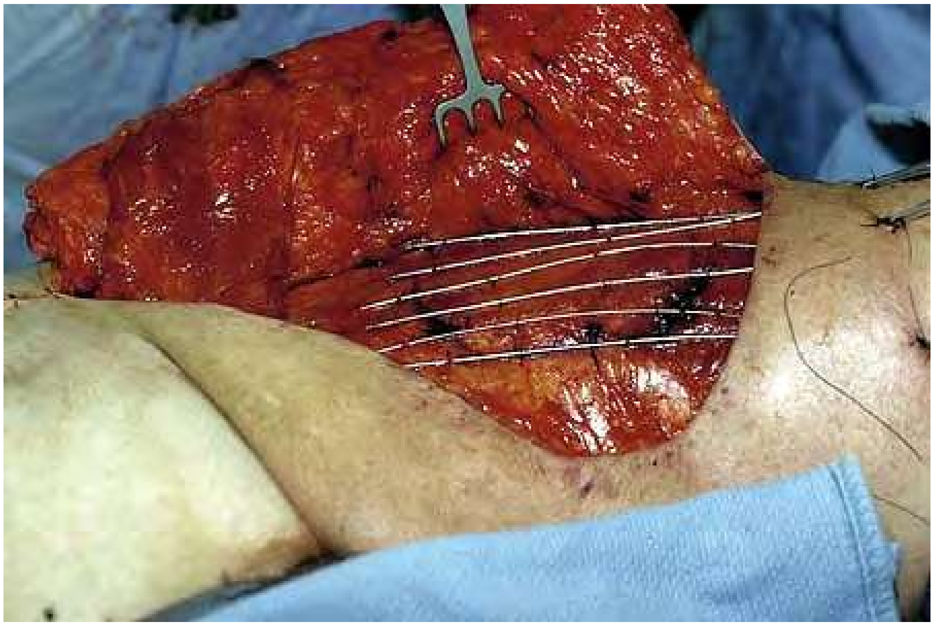

FIGURE 35-1 Placement of afterloading brachytherapy catheters under inferiorly based TRAM flap in treatment of recurrent sarcoma of the thigh.

Table 35-2. Principles of Pelvic Reconstruction

- Routine use of preoperatively placed ureteral stents aids in the identification of ureters intraoperatively.
- Patients should receive both antibiotic and mechanical bowel prep.
- Standard hemipelvectomy flaps provide adequate soft tissue coverage, and there is no difference in postoperative wound complications related to flap design.
- Abdominal and pelvic wall reconstruction is not necessary if the muscle with its investing fascia is a component of an external hemipelvectomy flap, but the abdominal wall should be

reconstructed in cases of internal hemipelvectomy.

- Omentum should be interposed between the intestine and hemipelvectomy flap whenever possible to wall off intra-abdominal contents in the event of hemipelvectomy flap necrosis.
- Precise surgical technique, thorough hemostasis throughout the entire hemipelvectomy wound, wide drainage with multiple closed suction drains, dÃ©bridement of all nonviable tissue off hemipelvectomy flap, and meticulous layered closure of the operative wound are the key to successful postoperative healing.
- Hemipelvectomy wound complications are common and related to the extensive nature of the procedure and the level of vascular ligation. Infected wounds have to be aggressively dÃ©brided until control of the wound is achieved.
- Secondary intention healing with wet-to-moist dressing changes and VACÂ® therapy is the most common approach to complicated hemipelvectomy wounds.
- Tertiary closure is reserved for the patients with healthy granulating wounds and extensive defects who are otherwise good operative candidates.
- Local tissue rearrangement by advancement of the skin flaps is the most common form of tertiary intention closure.
- Contralateral inferiorly based rectus abdominis muscle or musculocutaneous flap is the flap of choice for tertiary closure of large hemipelvectomy wounds. To preserve contralateral rectus abdominis muscle, contralateral ostomies should be avoided by careful preoperative planning and communication between different surgical specialties involved in this procedures.
- Hemipelvectomy reconstruction with a microvascular fillet flap obtained by ex-vivo anatomic exploration of an amputated extremity is an underutilized reconstructive option and should be considered in the cases of paucity or poor quality of local tissues and vascular ligation above the bifurcation of common iliac vessels that is plagued with the high rate of flap necrosis rate.

Preoperatively, the patients with large pelvic tumor undergo mechanical bowel preparation and intravenous antibiotic coverage. Ostomy sites must be preoperatively marked in accordance with anticipated flap use because inappropriate colostomy or ileostomy placement may burn an important reconstructive bridge and prevent rectus abdominis flap elevation. Involvement of several surgical services such as urological, colorectal, vascular, spine, and plastic surgery is common. The patient is positioned on a bean bag prior to induction of general anesthesia. Large bore intravenous access is established in an event of rapid blood loss. We liberally use ureteral stents that facilitate intraoperative identification of the ureters. After placement of the stents and Foley catheter, the patient is placed in the â€œsloppyâ€• lateral decubitus position and is secured with the bean bag. This position is preferred for internal or external hemipelvectomy because it permits a wide skin preparation and an easy access to the abdomen, buttock, and perineal regions. If additional procedures on spine, sacrum, or rectum need to be performed, intraoperative repositioning of the patient will be required (Table 35-2).

Low sacral resection can also be performed in the â€œsloppyâ€• lateral decubitus position (abdominolateral sacral portion) or a full lateral position with the hip and knee joints in 90-degree flexion. When combined abdominal exploration may be required to deal with the intrapelvic anterior component of the tumor, we start the abdominal portion of the operation supine and later reposition the patient for the posterior, sacral stage of the procedure. Plastic surgeon performs an initial marking and flap dissection as dictated by an anticipated defect.

Surgery

External Hemipelvectomy Reconstruction

Pelvic reconstruction following external hemipelvectomy is principally accomplished with three pedicled flap designs: posterior, long anterior, and total thigh fillet flaps. The vast majority of hemipelvectomy defects can be closed with these flaps which constitute the *first choice for hemipelvectomy flap reconstructions*. If these standard hemipelvectomy flaps are unusable due to very proximal vascular ligation, causing flap ischemia, division of the flap origin during tumor resection or previous procedure, or extensive radiation damage, then alternative flaps must be used for coverage. These *second-line reconstructive options* include contralateral inferiorly based vertical rectus musculocutaneous (VRAM) flaps, microvascular lower extremity fillet flaps, or standard free flaps depending on the defect configuration. Likewise, the second-line reconstructive options are useful for closure of hemipelvectomy wounds in the setting of postoperative wound complications (Table 35-3).

Table 35-3. Purpose of Soft Tissue Flaps in Tumor Surgery

- Covers exposed neurovascular structures
- Coverage for endoprostheses or allografts
- Addresses functional deficits (neurotized flaps)
- Obliterates dead space
- Enhances healing of irradiated wounds

Posterior Hemipelvectomy Flap

The classic hemipelvectomy technique relies on pelvic exploration, ligation of the common iliac vessels, division of the pelvic rim by disarticulation of the pubic symphysis and the sacroiliac joint, and creation of the posterior fasciocutaneous flap to achieve soft tissue closure. It was initially recommended that gluteal muscles be left with the specimen. This fasciocutaneous hemipelvectomy flap was based on relatively poor random blood supply due to ipsilateral ligation of the common iliac vessels and was further compromised by removal of the gluteal muscles that greatly increased wound complication rates.

Three modifications of this classic technique aimed to decrease high wound complication rates:

- Incorporation of gluteal muscles in the hemipelvectomy flap
- Whenever oncologically appropriate, ligation at the level of external iliac vessels with reservation of the internal iliac vessels to improve the flap blood supply
- Limited resection of the bony pelvis that allows preservation of the sacral perforators

With these modifications, the posterior hemipelvectomy flap is designed as a musculocutaneous flap based on the superior and inferior gluteal vessels (Fig. 35-2). Preservation of the gluteal muscle decreases posterior flap necrosis rates and makes the construction of a long, viable posterior flap that would reach up to or above the level of umbilicus possible. Impact of the level of vascular ligation on hemipelvectomy wound outcomes has been a point of controversy. Several reports from Karakousis et al suggested that the level of vascular ligation does not affect the posterior hemipelvectomy flap viability and the rate of postoperative wound complications. These authors believed that there was an adequate blood supply of the gluteal muscle through small arterial branches along its sacral origin, which was sufficient to sustain the viability of the flap unless resection of the edge of the sacrum was oncologically necessary. In our experience, we found 2.7-fold higher rates of posterior hemipelvectomy flap necrosis in the patients that had ligation at the level of common iliac vessels. This finding was independent from sacral resection performed during extended hemipelvectomy in some of these patients (1).

Long Anterior and Total Thigh Fillet Hemipelvectomy Flaps

One of the major limitations of the posterior flap external hemipelvectomy is its inability to deal with the advanced tumors of the buttock and posterior pelvis in an oncologically sound manner. In 1953, Bowden et al described utilization of the skin of the femoral triangle based on the preserved segment of the superficial femoral artery for closure of the hemipelvectomy performed for the sarcoma of the buttock. However, it was the critical need for soft tissue reconstruction of the advanced decubiti and the infection of bony pelvis in paraplegic patients that led to increased utilization of the soft tissue obtained from high amputations. The total thigh flap was proposed by Georgiade et al as a last-resort reconstructive option for such patients in the 1950s and subsequently gained wide-spread use. This principle was subsequently applied for coverage of the hemipelvectomy defects whereby a musculocutaneous flap of the anterior thigh compartment was elevated based on the superficial femoral artery. The technique was further refined by Sugarbaker et al, who also demonstrated that the anterior flap can be used as a sensate island flap based on the superficial femoral vessels and saphenous nerve.

Standard long anterior flap hemipelvectomy includes the bulk of quadriceps femoris muscle (Fig. 35-3). A total thigh fillet flap utilizing the majority of the thigh musculature can also be designed as a variation of the anterior hemipelvectomy flap technique (Fig. 35-4). Anterior hemipelvectomy flap is an axial pattern musculocutaneous flap based on the branches of femoral vessels, including lateral and medial circumflex arteries which arise from branches from the profundus femoris artery. The latter perforate the adductor magnus muscle to the posterior and lateral compartments of the thigh and play a role in supporting a total thigh fillet flap.

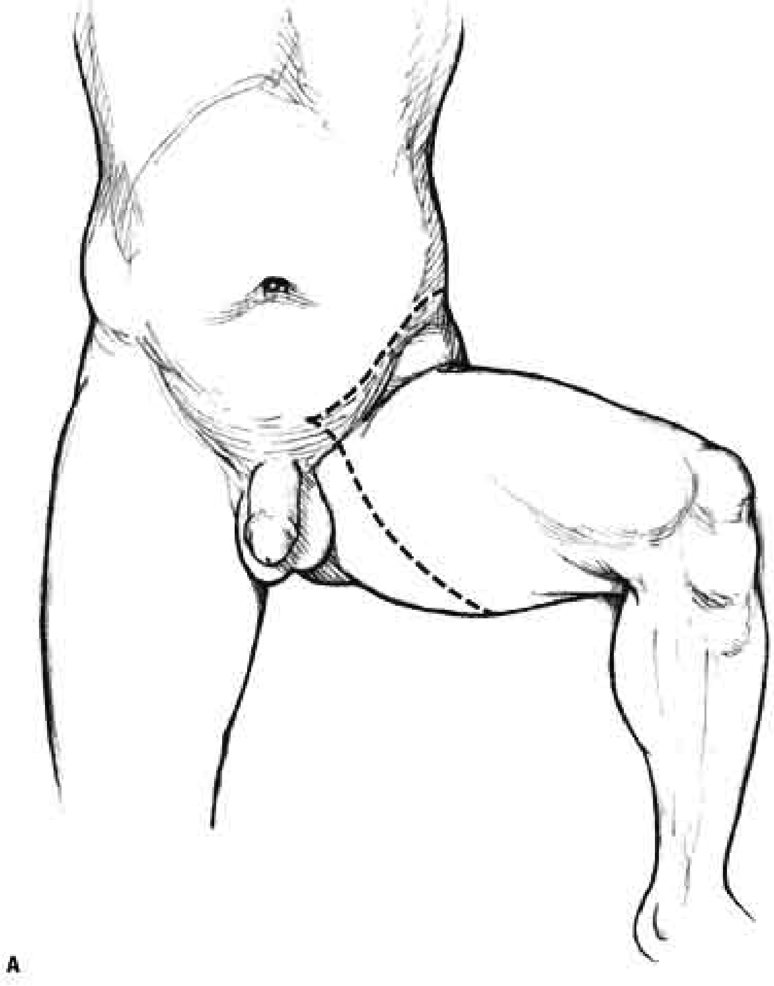
A

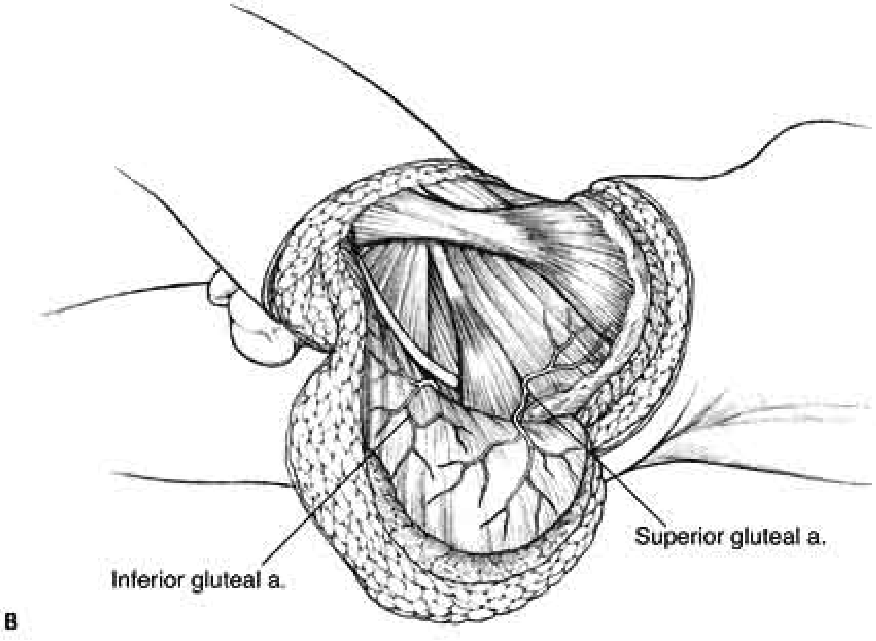
Superior gluteal a.
Inferior gluteal a.
B

FIGURE 35-2 A: Skin markings of external hemipelvectomy with utilization of the posterior flap. B: Musculocutaneous design of the posterior hemipelvectomy flap based on the superior and inferior gluteal vessels.

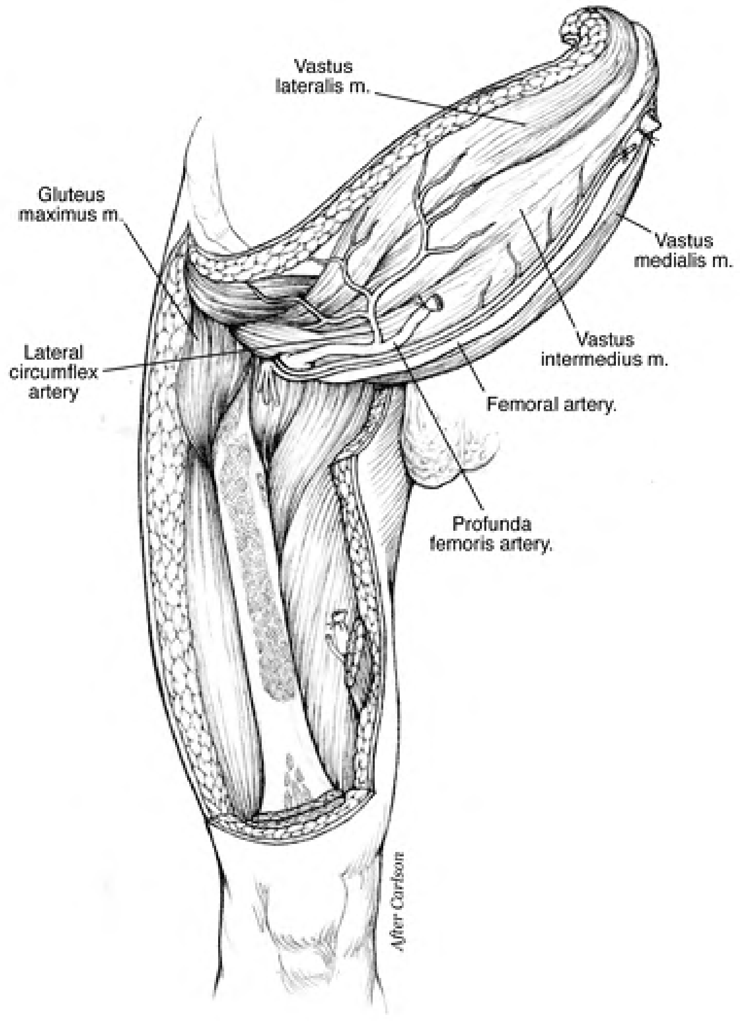
Vastus lateralis m.
Gluteus maximus m.
Vastus medialis m.
Lateral circumflex artery
Vastus intermedius m.
Femoral artery.
Profunda femoris artery.
After Carlson

FIGURE 35-3 Long anterior hemipelvectomy flap based on the branches of profunda femoris and superficial femoris vessels.

The skin of the anterior thigh down to the knee is innervated by the lateral and anterior femoral cutaneous nerves. These nerves can be preserved to provide sensory flap coverage of the hemipelvectomy defect. Both anterior flap and total thigh fillet flap hemipelvectomy provide well-vascularized and sensate immediate coverage of the hemipelvectomy defect. An effort should be made to preserve innervation of this flap by protecting sensory nerves during the dissection (Fig. 35-5). Both long anterior and total thigh fillet hemipelvectomy flaps are sufficient to cover even very extensive hemipelvectomy defects (as well as spinopelvic resections) with tissues that have rich axial-pattern blood supply.

Second-Line Hemipelvectomy Reconstructions

Standard hemipelvectomy flaps either posteriorly or anteriorly based provide reliable reconstruction in a vast majority of the cases because removal of the hemipelvis creates relative soft tissue excess for three reasons:

- Reduction of pelvic volume decreases requirement for the size of soft tissue envelop.
- Skeletal resection eliminates the issue of the dead space.
- Tissues or the proximal part of the amputated lower extremity can be used for hemipelvectomy defect reconstruction.

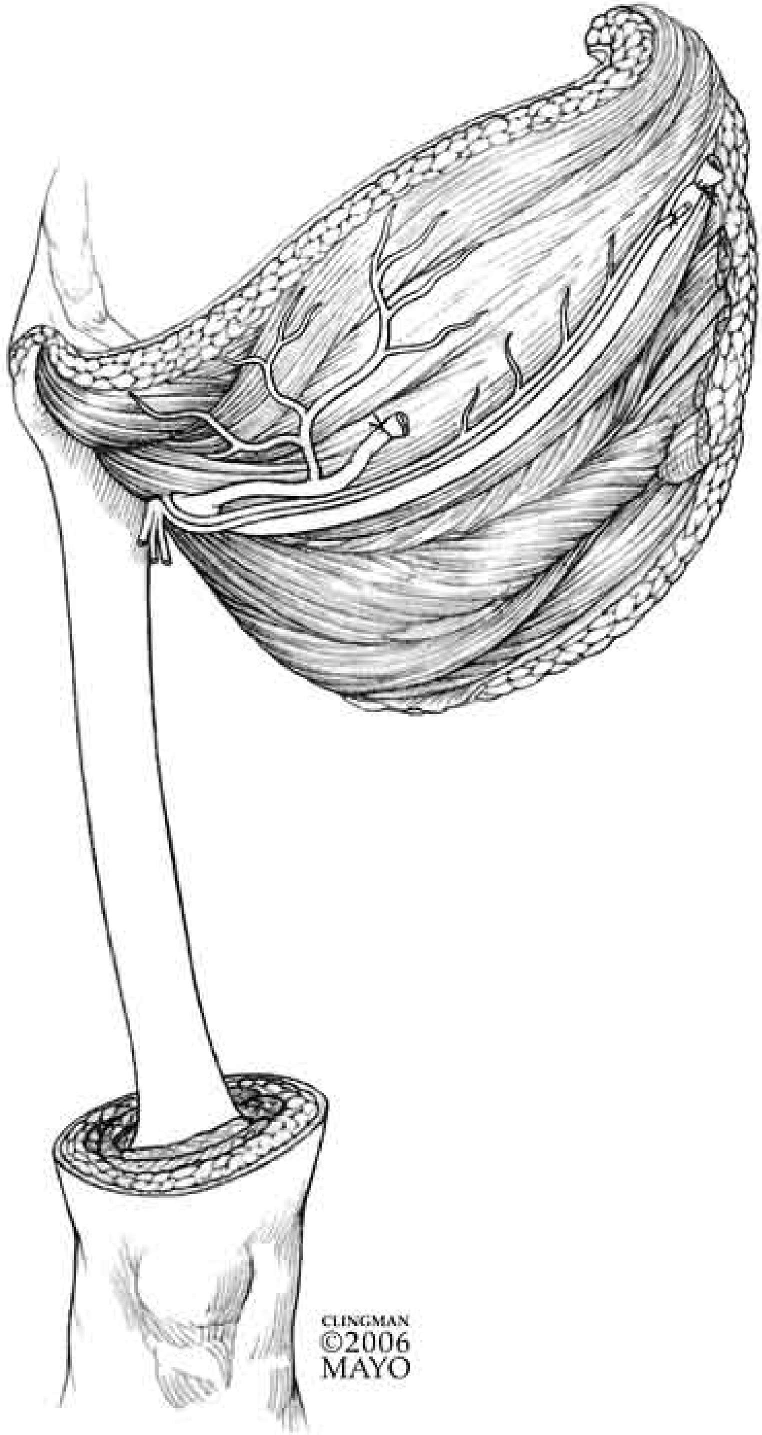
CLINGMAN
©2006
MAYO

FIGURE 35-4 Total thigh fillet flap utilizes the majority of the thigh musculature.

For these reasons, external hemipelvectomy itself is the ultimate solution to soft tissue coverage problems. The difficulty arises when common iliac vessels have to be ligated for oncologic reasons, which precludes creation of the anterior hemipelvectomy flap, and, at the same time, tissues of the buttock are not suitable for posterior flap design due to tumor involvement or sequela of previous operations or radiation therapy. In the past, such patients were considered unresectable. With the advent of microvascular tissue transfer however, a suitable and tumor-free block of tissue can be recovered by ex-vivo anatomic exploration of an amputated extremity and transferred as a free flap to achieve coverage of the hemipelvectomy defect.

Internal Hemipelvectomy Reconstruction

Internal hemipelvectomy involves total or partial removal of the innominate bone with preservation of the lower extremity. This operation provides good local tumor control and acceptable functional outcome. Several types of internal hemipelvectomy as proposed by Enneking and Dunham exist based on the part of innominate bone resected: Type 1â€”ileum; Type 2â€”periacetabular region; Type 3â€”pubic bone; and, described by some, Type 4â€”ileum and sacral ala (Fig. 35-6). This classification provides the basis for the surgical oncologic approaches to malignant tumors of the nonvisceral pelvis.

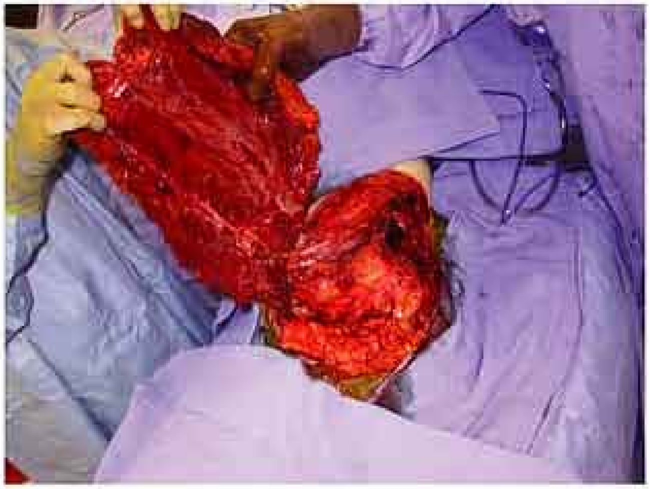

A

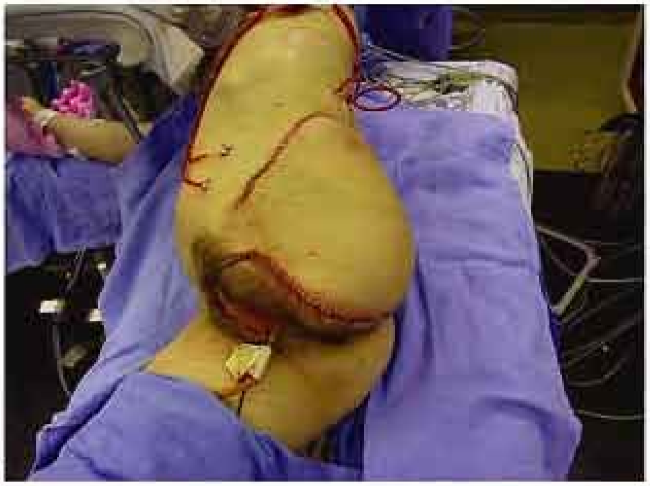

B

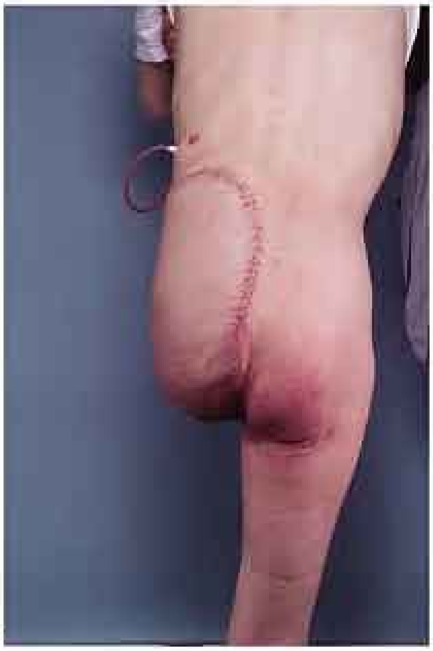

C

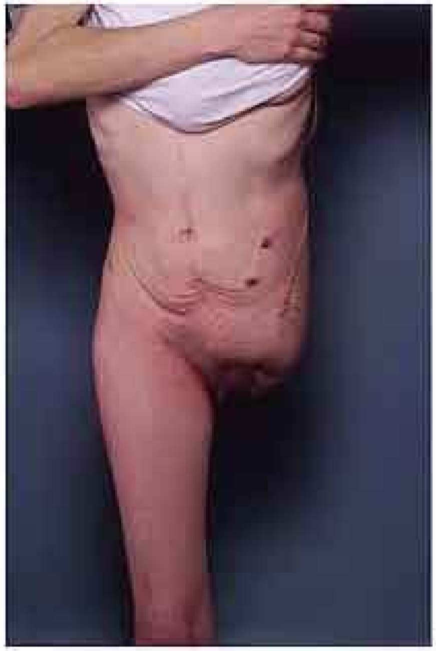

D

FIGURE 35-5 A: Total thigh fillet flap provides abundance of well vascularized soft tissue. B: The flap provides excellent contour and adequate posterior reach. C,D: Postoperative result following extended hemipelvectomy for the tumor of the buttock.

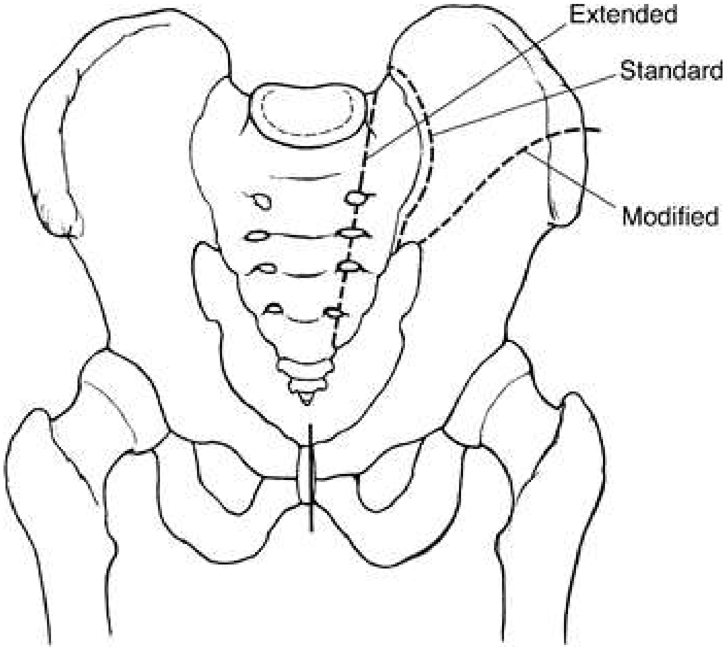

FIGURE 35-6 Schematic drawing of the pelvis showing osteotomy lines for standard modified and extended hemipelvectomies.

Introduction of internal hemipelvectomies increases requirements for soft tissue reconstruction since the pelvic volume is not reduced and lower extremity cannot be filleted. Additionally, a large amount of the dead space is created and often requires obliteration with a flap, especially in the settings of preoperative radiation. Several regional pedicled flap options for filling these defects exist (Table 35-4). Muscle flaps and de-epithelialized musculocutaneous flaps can be used effectively for dead space obliteration. De-epithelialized flaps create more bulk and are used for filling larger defects. Because fat is a part of the skin paddle, the flap retains volume long term. In addition to dead space obliteration, these flaps bring blood supply into relatively ischemic radiated tissues of the operative site and thus enhance stable coverage and promote wound healing. Although flap demands for bony sarcomas without soft tissue extension are not high, however, in the cases of soft tissue sarcomas or soft tissue involvement from the bony sarcomas, flap reconstruction becomes critical for the coverage of exposed skeletal structures. Musculocutaneous, muscle pedicled, or free flaps can be used. When the skin is not part of such a flap, a split-thickness skin graft can be used for coverage of exposed muscle of the flap.

Sacral Reconstruction

Primary sacral tumors are often low-grade lesions such as chordoma. They do not metastasize, but have a tenacious local course. Chordomas and locally-advanced rectal cancers are the two most commonly encountered sacral tumors. Sacral surgery presents a major challenge due to regional anatomic complexity, technical difficulty in obtaining clear margins, functional impairment, and an often prolonged postoperative course due to poor wound healing. Sacrectomy is a procedure that is uncommon outside of specialized cancer centers. Although distal sacral resections are safely performed through the posterior approach, larger tumors, tumors of rectal origin, and resections proximal to S3 require an initial abdominal exploration to dissect visceral and neurovascular structures, perform formal visceral resection when required, assure hemostasis, and complete anterior sacrectomy dissection. Sacrectomy is completed through the posterior approach.

Soft tissue reconstruction becomes important in preventing postoperative wound complication following sacrectomy. Two flaps are commonly utilized: pedicled omentum and rectus abdominis muscle/musculocutaneous flaps. Omentum is preferred for smaller defects (the defects *smaller than the surgeon's fist*), and the rectus flap is used for larger ones. Inferiorly based vertical rectus abdominis musculocutaneous (VRAM) flap is the most commonly used reconstructive technique.

The flap accomplishes two main objectives: it obliterates large dead space and provides tension-free closure of the sacrectomy wound. In the cases of preoperative radiation, well-vascularized tissue of the flap enhances local circulation and further promotes wound healing. Implementation of this practice in our institution led to a marked decrease in sacrectomy wound complication rates. Careful preoperative planning and communication between different services is critical in executing these operations. VRAM flap as a reconstructive option must always be borne in mind in the light of previous abdominal incisions as well as ostomies, drains, and feeding tube placements. The VRAM flap represents a very important surgical technique in oncologic pelvic reconstruction.

If soft tissue defect is too large for sole VRAM coverage, fasciocutaneous V-to-Y advancement flap provides additional recruitment of the local tissues. If aforementioned techniques are not sufficient to achieve an immediate coverage of sacrectomy-buttockectomy defect, microvascular tissue

transfer should be performed with a free flap of appropriate dimensions. Finally, anterior flap external hemipelvectomy is reserved as a last resort reconstructive operation.

Table 35-4. Flap Options for Internal Hemipelvectomy Defect Reconstruction

- Rectus abdominis muscle based are inferiorly based ipsilateral or contralateral:
    - Vertical rectus abdominis musculocutaneous (VRAM) flap
    - Transverse rectus abdominis musculocutaneous (TRAM) flap
    - Rectus abdominis muscle flap with or without skin graft
- Rectus femoris muscle flap
- Latissimus dorsi muscle or musculocutaneous flap
- Vastus lateralis muscle flap
- Tensor fascia lata muscle or musculocutaneous flap
- Anterolateral thigh flap

Technical Elements of Pelvic Flap Surgery

Omentum Pedicled Flap

Omentum is readily available during laparotomy although it can also be harvested laparoscopically. Its blood supply is based on the right or left gastroepiploic artery. Significant mobilization and extent of reach can be gained by basing the flap on left gastroepiploic vessels, dividing the short gastric vessels along the greater curvature of the stomach, and further dividing middle omental artery to release a long vascular pedicle (Fig. 35-7). The latter maneuver releases omentum to its fullest length, which may be important for the reconstruction of the larger pelvic defects in the event that the VRAM flap is unavailable.

Rectus Abdominus Muscle and Musculocutaneous Flaps

The rectus abdominus muscle can be elevated alone or with a skin paddle. The skin paddle may be oriented transversely or vertically.

The transverse rectus abdominis musculocutaneous (TRAM) or vertical rectus abdominis musculocutaneous (VRAM) flaps can be utilized as either pedicled or microvascular free flaps. The rectus abdominus muscle has a dual blood supply from the superior and inferior epigastric vessels. For the purposes of pelvic soft tissue reconstruction, the rectus abdominus muscle flaps are based on the inferior epigastric vessels when pedicled to cover the pelvis or sacrum.

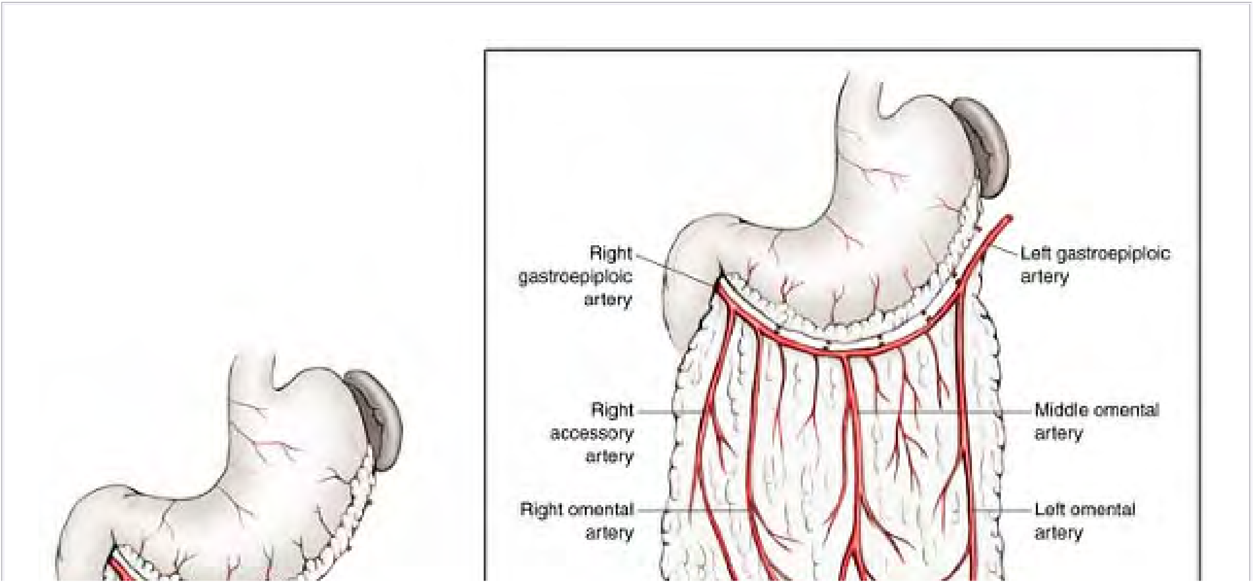

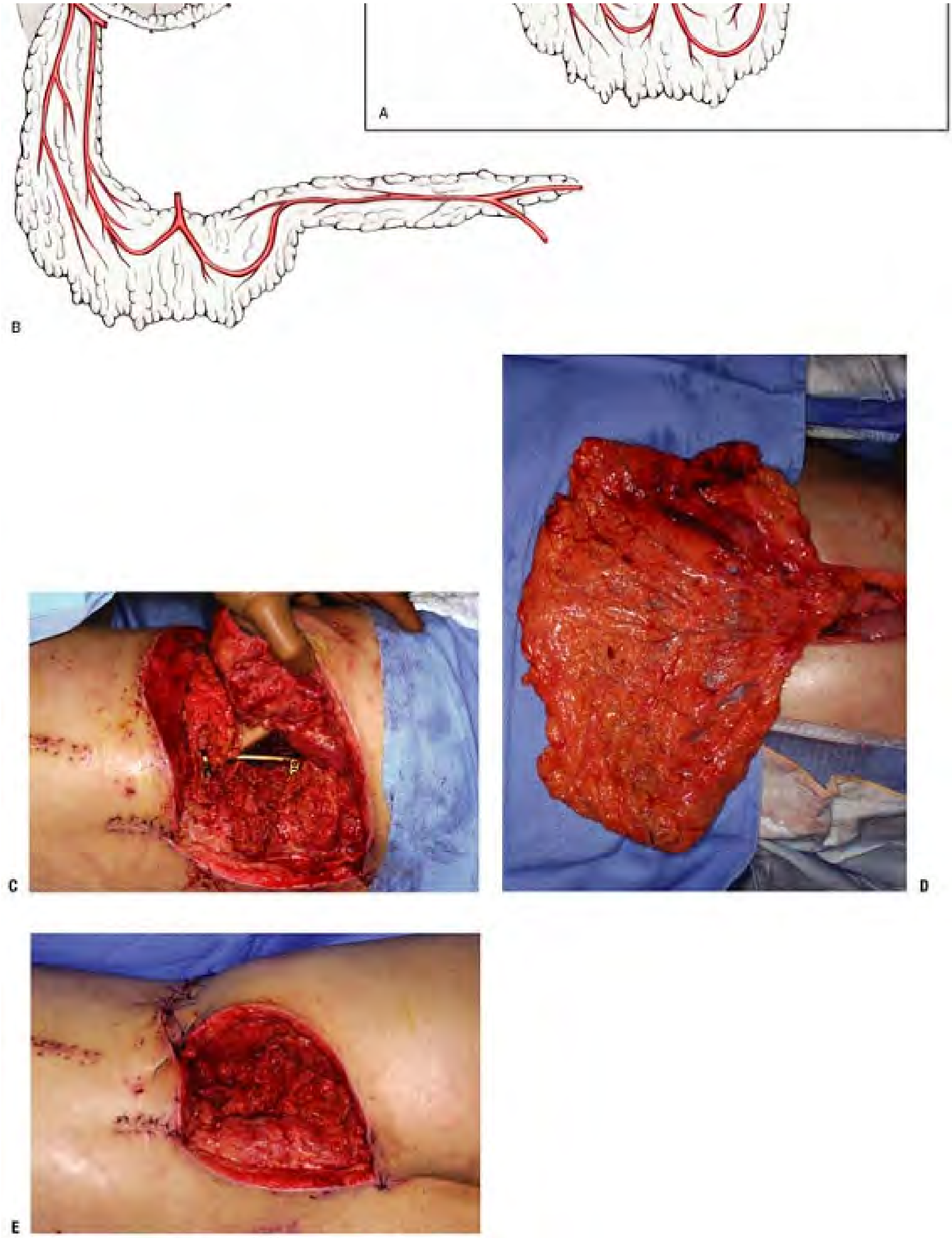

FIGURE 35-7 A: Omentum can be used for coverage in difficult situations following pelvic resection and soft tissue deficit. While the omentum lacks

significant structural strength it provides a well vascularized bed which can support skin grafting. The division of the left gastroepiploic arch allows the omentum to be mobilized off the greater curvature of the stomach. The right gastroepiploic vessels are preserved to supply the flap. B: The omentum may be lengthened by dividing the omental arcade along the lines within the illustration, creating a long vascularized pedicle. Câ€“E: Clinical example of a 16-year-old woman with exposed spinal hardware (C) following sacrectomy. The omentum was harvested through a midline anterior incision (D), lengthened as shown in (B) and then passed through the back to provide coverage for the spinal hardware.

Several key elements of preoperative planning are important in preparation for this procedure. A detailed history of previous intra-abdominal surgery needs to be obtained, because the inferior epigastric vessels may be divided during operations such as appendectomy, inguinal hernia repair, C-section, and colostomy creation (Fig. 35-8). The pedicle may also be damaged by radiation therapy. If there is any question about the integrity of the vascular pedicle, duplex evaluation should be performed.

Flap elevation is performed prior to abdominal exploration. We open the rectus sheath close to midline and dissect it off the rectus abdominis muscle (Fig. 35-8Aâ€“C). In this part of procedure, care should be taken to dissect inscriptions due to their proximity to the underlying vessels that run on the undersurface of the muscle. Once the flap has been dissected for coverage of the sacrum and posterior defects, the flap is placed in a plastic bag or wrapped with towels and placed over the anterior

portion of the lumbar spine and sacrum. The abdominal portion of the procedure is then completed and the abdomen is closed. If a large fascial defect remains following VRAM harvest, this is reconstructed with synthetic mesh or allograft fascia. Following abdominal closure, the patient is placed in the prone jackknife position for the sacral portion of the operation, and the flap is easily visualized after removal of the specimen (Fig. 35-8Dâ€“F). Inset should ideally be performed with no tension to prevent sacral closure breakdown when the patients starts to sit.

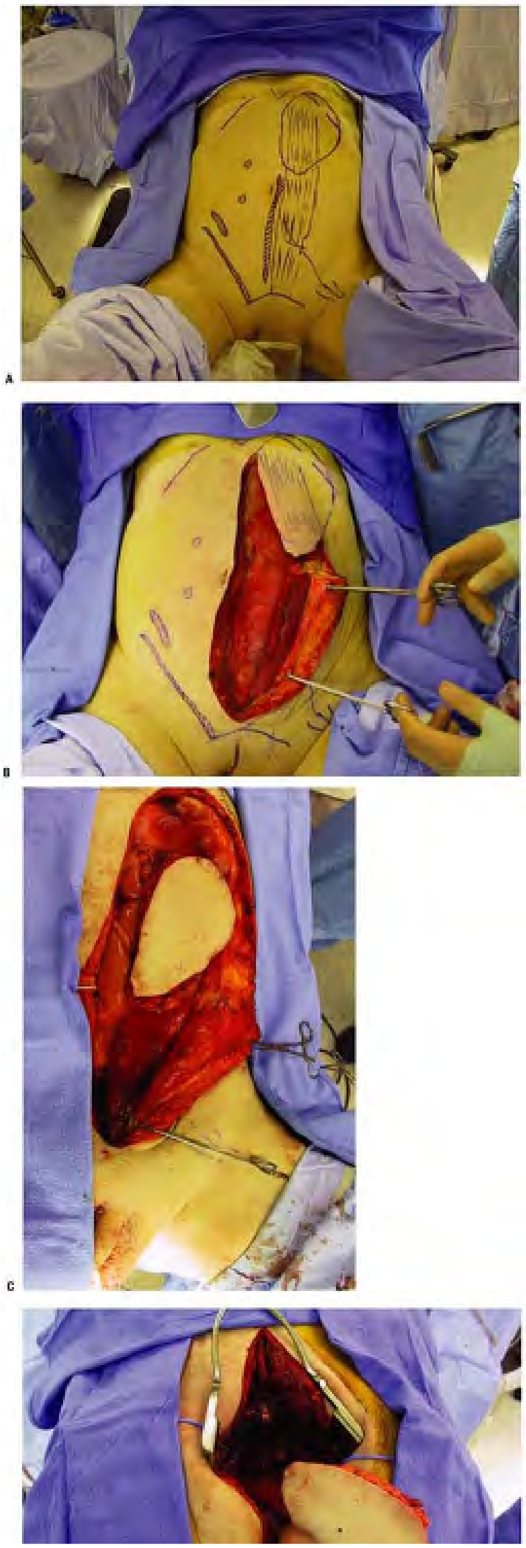

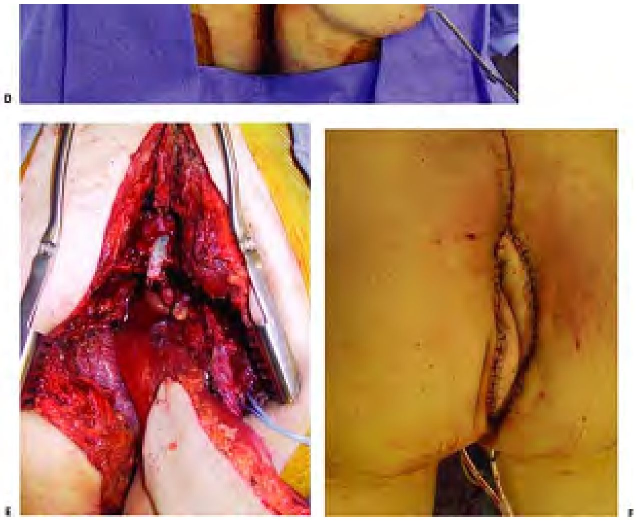

FIGURE 35-8 A: VRAM flap is planned for reconstruction of the defect from *en bloc* abdomino-perineal resection with sacrectomy. Skin paddle is designed to overlie rectus abdominus muscle. B: The rectus sheath is open and the rectus abdominis muscle is exposed. The anterior fascia is preserved with the skin paddle. C: The flap elevation is complete. The rectus abdominus muscle has been divided superiorly from its insertion into the ribs. The inferior epigastric vessels have been mobilized to prevent kinking as the flap is passed into the abdomen. D: The flap is placed into the abdomen and the anterior incision is then closed. The patient is then placed into the prone position and the sacrectomy is completed. At this point the VRAM flap is retrieved from the pelvis. E: Prior to its inset the muscle and vascular pedicle are examined to ensure there is no evidence of kinking or twisting. Note exposed dural sack. F: VRAM inset is complete with generous amount of skin remains to accommodate for sitting.

Other Flap Options

In many instances, the safest approach to patients with pelvic sarcomas is a laparotomy and, when it is performed, omentum becomes a valuable reconstructive material. Omentum, if present, should always be interposed between hemipelvectomy flap and intra-abdominal viscera. This provides an additional protective layer in the event of hemipelvectomy wound problems. Omental flaps can be effectively used for dead space obliteration.

Although omentum and VRAM are truly the workhorses of pelvic reconstruction, several other flaps may become important as second line reconstructive options. Rectus femoris, tensor fascia lata, and anterolateral thigh flap can be used as pedicled flaps. Latissimus dorsi and anterolateral thigh flaps can be transferred as microvascular flaps depending on soft tissue requirements and availability of recipient vessels.

V-to-Y Advancement Gluteal Flaps

This versatile reconstructive technique can be performed in escalating complexity and is usually tried in the following sequence:

- Unilateral fasciocutaneous
- Bilateral fasciocutaneous
- Unilateral musculocutaneous in conjunction with contralateral fasciocutaneous
- Bilateral musculocutaneous
- VRAM and bilateral fascio- or musculocutaneous advancement flaps (Fig. 35-9)

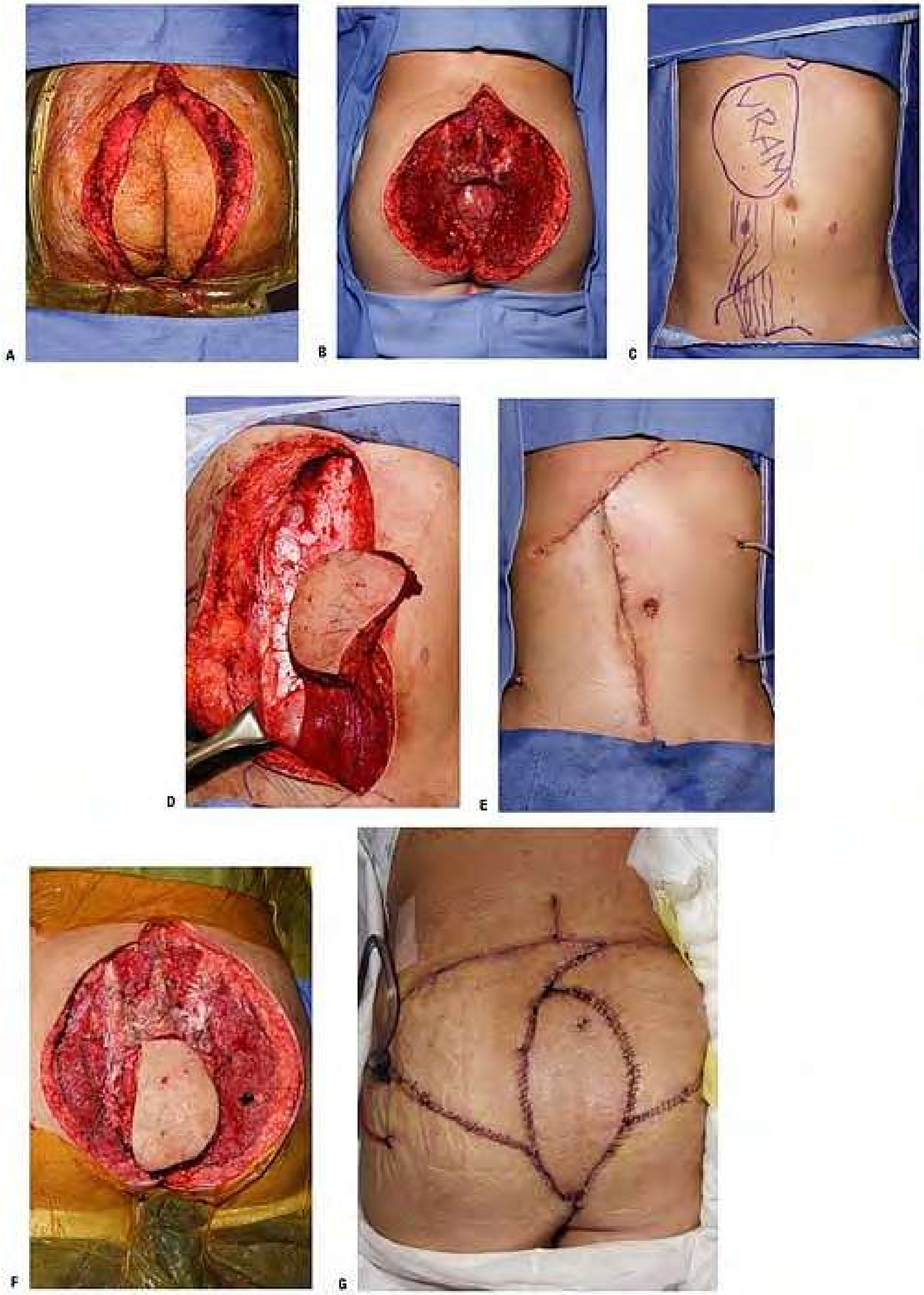

FIGURE 35-9 A,B: Combination of inferiorly based pedicled VRAM flap and bilateral fasciocutaneous V-to-Y advancement flaps for closure of a large sacral defect after wide local excision of the sacral sarcoma. Câ€“E:

Inferiorly based VRAM is elevated with large skin island based over superior portion of rectus abdominus muscle. The abdomen is closed with the aid of extensive undermining. The umbilicus is repositioned beneath the skin flaps to lie in the midline. F: VRAM flap allows for soft tissue coverage over sacrectomy defect and allows for reconstruction of pelvic floor; however large soft tissue defect remains at margins of sacrectomy incision. G: Remaining portion of sacral wound is closed with use of bilateral fasciocutaneous flaps. Flaps are elevated with fascia only to avoid injury to gluteus muscles to maximize post-operative ambulatory ability.

The surgeon must be careful in utilizing this flap in an ambulatory patient as it may affect functional performance. It is safer to reserve musculocutaneous gluteal flaps for paraplegic individuals. For large buttock defects, lateral donor sites of the V-to-Y advancement flaps sometimes may need to be temporarily covered with Vacuum-Assisted Closure (VACÂ®) dressings to be closed later when the edema subsides, skin is grafted, or allowed to heal by secondary intention.

In the face of insufficient amounts of local tissues, the donor part of the V-to-Y advancement flap can be skin grafted or covered with a wound VAC to allow healing by secondary intention.

Abdominal Wall Reconstruction in Pelvic Surgery

Integrity of the abdominal wall has to be maintained to prevent postoperative hernias. For a standard external hemipelvectomy defect without extensive resection of the abdominal wall, in the patients with musculocutaneous design of hemipelvectomy flap, no specific reconstruction of the abdominal wall is necessary other than meticulous layered closure of the hemipelvectomy wound. We did not observe postoperative hemipelvectomy hernias under these circumstances because sturdy fascia of the anterior compartment of the thigh or gluteus muscle along with muscular bulk of the flap effectively withstands pressure of the intra-abdominal viscera. Conversely, abdominal wall reconstruction becomes important for internal hemipelvectomy defects because frequently in these patients there is a very clear area of weakness in the lower abdominal wall. This situation is also aggravated by higher functional level of the internal hemipelvectomy patients.

Reconstructive efforts are guided by the extent of the abdominal wall deficit and degree of operative contamination. Time-tested surgery principles of ventral hernia repair apply. Hernia repair should be tension free with liberal use of prosthetic mesh materials. Well-incorporated prosthetic mesh is superior to nonneurotized autologous options because it secures the dimensions of the abdominal wall and prevents postoperative bulge and recurrent hernia formation. The downside of prosthetic mesh reconstruction is its propensity to infection in the presence of intraoperative contamination. Although prosthetic mesh can be used in clean-contaminated cases, most surgeons are reluctant to use nonabsorbable mesh in this setting and prefer either absorbable mesh such as Vicryl or a biologic substitute such as AlloDermÂ®, processed cadaveric human acellular dermis (Life Cell Corporation, Branchburg, NJ, USA) or SurgisisÂ®, a product obtained by the processing of porcine small intestine sub-mucosa (Cook Surgical, Bloomington, IN, USA). VicrylÂ® mesh is a temporary abdominal reconstruction solution that permits a conversion of the contaminated situation into a clean one at the time of the second procedure. The latter aims for definitive reconstruction, but requires utilization of costly biologic materials and postoperative formation of abdominal wall hernia or bulge continues to be a problem.

Postoperative Management

Postoperatively, patients usually require an ICU admission after major pelvic resections or hemipelvectomies. They are commonly kept on mechanical ventilation overnight until intravascular volumes are stabilized and gas exchange is adequate. To optimize postoperative recovery as well as systemic and flap perfusion the patient needs to be kept well hydrated. Adequate urine output needs to be maintained. The patient with pelvic or sacral flap is usually kept on a Clinitron bedÂ® to avoid pressure injury to the flap. Recovery of bowel function, ostomy care, DVT prophylaxis, and wound care is similar to that for a general surgery patient undergoing abdominal procedures.

Physical and occupational therapy services are involved early in the care of these patients. Even when the patients are on bed rest, range of motion exercises on nonoperated extremities and activities of daily living should be started. When it is safe, from a surgical perspective, to resume ambulation, the patients are evaluated for short-term rehabilitation placement.

Results

Our experience shows that the outcome of external hemipelvectomy wounds was not dependent on hemipelvectomy flap selection. Reliability of its blood supply determines the success of the reconstruction. Historically, classic hemipelvectomy flaps that are based on the random circulation are exceedingly prone to necrosis. Preservation of the gluteal muscle as a part of the flap decreased, but did not solve, the problem of hemipelvectomy flap necrosis. We also observed that preservation of the branches of the internal iliac vessels providing direct blood supply to the gluteus maximus muscle further decreased hemipelvectomy flap necrosis rates. Hemipelvectomy wound outcomes therefore can be improved by increased utilization of microvascular fillet flap transfer that provides superb blood flow to the transferred tissue, not comparable with the periphery of a random hemipelvectomy following common iliac vessel ligation.

Complications

Wound complications such as surgical site infection and skin flap necrosis are the most common postoperative complications in musculoskeletal oncology. There are multiple reasons for the high rate of wound complications following major resections of musculoskeletal malignancies. Despite the fact that the vast majority of these procedures are clean cases with only occasional visceral resection that is applicable for pelvic tumors, wound complication rates are well beyond what one would expect. In modern musculoskeletal oncology practice, many sarcoma patients undergo neoadjuvant (preoperative) radiation and chemotherapy. Soft tissue complication rates approach 40% to 50 % in this group of patients. Moreover, even without neoadjuvant therapy, the duration and operative extent of the procedures significantly influence rates of postoperative wound complications. As noted earlier, wound complications that delay the administration of postoperative adjuvant therapies may have serious oncologic consequences. Rapid, uncomplicated wound healing is essential for patients who require postoperative adjuvant therapy, highlighting the need for early and close collaboration with plastic surgeons.

A hematoma can occur after any oncologic surgery, ranging from incisional biopsy to extensive surgery. All tumor resections should utilize at least one suction drain, although larger resections may require several drains placed deep and in more superficial tissues. The pressure from the collection of blood products in a wound can compromise soft tissue reconstructions and may become infected. Drain sites and tracks are contaminated and usually are resected at the time of definitive tumor removal. For this reason, drains should exit the skin in line with the longitudinal incision used for biopsy. It is important to remember that drain placement can be as important as where one makes an incision. Drains should come out in line and close to an incision.

Hematomas are especially problematic following incisional biopsy, potentially compromising future resection and successful local control of a tumor. A hematoma can spread neoplastic cells beyond the immediate area and even into surrounding compartments if a transverse incision is used. Rigorous attention to hemostasis is therefore essential. Compressive dressings bolstered with plaster can also help minimize hematomas in the immediate postoperative period.

Seromas can occur following tumor resections, again potentially compromising soft tissue reconstructions and wound closures. Most surgeons therefore take a conservative approach to removing drains, allowing output to decrease (to less than 30 mL/day) and remain low before removal. Drains usually remain in place for a minimum of several days, but often are left in for much longer in large or irradiated wound beds. As with other facets of patient care, good communication with plastic surgery is important regarding the timing of drain removal. It is recommended that patients going home with suction drains remain on oral antibiotics.

Oncology patients often have a compromised nutritional status, making them prone to wound infections. Moreover, these patients are subject to the immunosuppressive actions of chemotherapy agents and the local wound effects from radiation. As many oncologic reconstructions rely on allografts, endoprostheses, or combinations of the two, surgeons should approach wound infections extremely aggressively. This approach entails a low threshold for surgical dÃ©bridement. As with all surgical procedures, orthopedic oncology patients should receive pre- and perioperative intravenous antibiotics as well as postoperative coverage when appropriate.

Sacrectomy is prone to wound complications that are related to the extent of the procedure, intraoperative contamination, use of preoperative radiation therapy, local tissue ischemia, hematoma, positional pressure, dead space, and tight closure. In the setting of preoperative radiation, wound complication rates are as high as 50%. These factors can be minimized by meticulous surgical technique, obliteration of the dead space with a flap, and postoperative use of a Clinitron bedÂ®.

Over the past 20 years, 160 external hemipelvectomies were performed in our institution. External hemipelvectomy has been associated with high morbidity but low mortality (5% to 7%). Overall, 54% of patients had at least one complication. Hemipelvectomy wound morbidity was the most common postoperative complication. Thirty-nine percent of patients experienced wound infection and 26% had hemipelvectomy flap necrosis. Wound complications were managed with serial dÃ©bridement until control of the wound was achieved; however, this may result in a sizable defect.

In a delayed reconstruction setting, the amputated extremity is no longer available for tissue procurement and the contralateral pedicled VRAM flap becomes critical in closure of such defects. This flap provides a superb reconstruction and in fact is the flap of choice for postoperative hemipelvectomy wounds. In a very rare circumstance, when VRAM is unavailable and the hemipelvectomy wound is so large it cannot be closed by local tissue rearrangement or a skin graft, a free flap such as a contralateral anterolateral thigh flap or latissimus dorsi may be required. One has to consider a paucity of recipient vessels and the need for vein grafts or saphenous arteriovenous loop that significantly increase the risk of flap failure.

Oncologic outcomes of musculoskeletal tumors are largely dependent on tumor pathology. Limb salvage is possible in the vast majority of sarcoma patients. In those patients who still require proximal amputations such as external hemipelvectomy, stable soft tissue coverage is almost uniformly achieved. Despite postoperative wound complications related to the operative extent of the resection, intraoperative contamination, and preoperative radiation therapy, the vast majority of the patients heal the surgical wounds. Both successful wound healing and high rates of limb salvage became possible due to advances in reconstructive surgery over the past three decades. In our practice, we emphasize early involvement of a plastic and reconstructive surgeon in care of a sarcoma patient.

Pearls and Pitfalls

- Plastic surgery consultation early
- Early, durable reconstructions are important for functional recovery and also prevent delays in administration of postoperative chemotherapy and radiation therapy
- All surgeons should be present in the operating room at the start of a combined case

- Donâ€™t use an Esmarchâ€”exsanguinate, if necessary, by elevating extremity and digital arterial compression
- Use compressive Robert Jones dressings when feasible, even after soft tissue resections

References

1. Senchenkov A, Moran SL, Petty PM, et al. Predictors of complications and outcomes of external hemipelvectomy wound: account of 160 consecutive cases. *Am Surg Oncol.* 2008;15(1):355â€“363. Epub 2007, Oct 23.

Recommended Reading

Abramson DL. Single-stage, multimodality treatment of soft-tissue sarcoma of the extremity. *Ann Plast Surg.* 1997;39:454â€“460.

Aflatoon K, Manoso MW, Deune EG, et al. Brachytherapy tubes and free tissue

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 36

Management of Major Upper Limb Amputation

Tsu-Min Tsai

Huey Y. Tien

Indications/Contraindications

Replanting the upper limb above wrist level involves a large quantity of muscle and, consequently, is regarded as major replantation. Because muscle does not tolerate anoxia well, the success of arm replantation depends on the effects of ischemia, and warm ischemic time is much more detrimental than cold ischemic time. The tissue damage caused by 1 hour of warm ischemia is equal to that caused by 20 hours of cold ischemia. When warm ischemic time is more than 6 hours, the success rate is markedly decreased and the complication rate is significantly increased. When arm amputation is incomplete, there might be a venous blood regurgitation that causes capillary refill to be present. In this situation, health care personnel may be unaware that the injured limb is devascularized, and therefore fail to keep the injured limb in a cool environment. This oversight further lengthens warm ischemic time and jeopardizes the survival of the injured limb. Furthermore, one should remember that on-table time to revascularization is often warm ischemia time, as well. Daigle and Kleinert reported that the average warm/total ischemia times were 4.8/14.8 hours in failed replantation, which is significantly higher compared with the average of 1.1/7.5 hours in successful replantation.

As a rule, the more proximal the amputation, the poorer the prognosis. Although some debate its efficacy in cases of amputation proximal to the mid forearm, the upper limb should be replanted if at all possible, both for function and cosmesis. However, there are some conditions that contraindicate arm replantation, including:

- Significant associated injury
- Extensive injury to the affected limb or to the amputated part
- Severe medical illness

Significant associated injury to major organs prolongs the patient's time in surgery and, therefore, increases blood loss and the need for transfusion. Limb replantation is clearly contraindicated if it may jeopardize the life of the patient. Injury factors that require significant consideration are:

- Avulsion versus guillotine injury
- Ischemia time elapsed
- Quantity of tissue loss
- Presence of multiple levels of injury
- Degree of wound contamination

Chronic illness also may complicate, and even contraindicate, limb replantation. For example, vessel anastomosis may be more difficult among patients with diabetes, who have a higher prevalence of arteriosclerosis. Diabetes, autoimmune disorders, and prolonged steroid use increase patientsâ€™ susceptibility to infection. Serious illnesses, including organ failure, preclude replantation surgery.

In addition to severity of injury to the limb and chronic medical illness, one also needs to take patient factors into consideration. Although there is no absolute cut-off age for replantation, the prognosis for a favorable functional outcome decreases with the increase of age. Children, as a rule, tend to be excellent candidates for major upper limb replantation. They have an advantage over adults in tissue regeneration generally, and especially in nerve regeneration. Patient factors in addition to age include the patient's occupation, hobbies, and wishes.

Preoperative Planning

Of course, always treat any life-threatening associated injury first. As soon as possible, control bleeding on the amputated stump with pressure dressings. Do not use any clamp instrument to try to stop bleeding because this increases the chance of damaging the vital structures (e.g., nerve) next to the bleeding vessels. The cut arterial end, however, can be tied to stop bleeding. Preserve the amputated part in a plastic bag, and cool it in ice slurry to just above freezing, usually 4Â°C. Then arrange rapid transportation to the replantation center.

In cases of incomplete amputation, wrap the limb snugly with gauze and elastic bandage so that bleeding stops, but not so tightly that blood flow is cut off to healthy tissue. Then, splint the injured area and place a regular ice pack on it. Special attention should be paid for any sign of hypovolemia.

While the patient is prepared for surgery or during transportation, the amputated part can be perfused with organ preservation solution, for example, Tsai's solution or UW solution, as shown in Table 36-1. Perfusion may have several benefits over simple immersion for large tissue parts, including the physical benefits of more rapidly cooling deep tissues and flushing stagnant blood. If the ischemic time is more than 6 hours, and especially if the amputation is incomplete, the amputated part also can be perfused with 1 unit of heparinized arterial blood obtained from the patient (Fig. 36-1). When the amputated part is brought to the operating room for cleaning and initial debridement, continue to keep it in a cool condition at all times.

Surgery

The patient is taken to the operating room and placed on the operating table in the supine position. Because arm replantation surgery may take as long as 16 to 18 hours, general anesthesia is practically

always required. The patient is given a dose of intravenous antibiotics 30 minutes before surgery. The patient's injured limb is prepared and draped in standard sterile fashion. An upper-arm tourniquet is inflated to 100 mm Hg above the patient's systolic pressure to minimize bleeding. If necessary, the contralateral lower limb is also prepared and draped for possible skin, nerve, and/or vein grafting; it may even be used to provide a free flap.

Table 36-1. Solutions Used in Arm Replantation

Solution	Content	Concentration	Preparation
Tsai's solution	Injectable lactated Ringer's, heparin, preservative-free Lidocaine		Lactated Ringer's 150 mL + heparin 3,000 units + 1% Lidocaine 30 mL
Papaverine	Papaverine	1.5 mg/mL	Papaverine 1 mL (30 mg) + normal saline 19 mL
LMD/heparin drip	LMD, heparin	10 units/mL	LMD 500 mL + heparin 5,000 units

LMD, low-molecular-weight dextran.

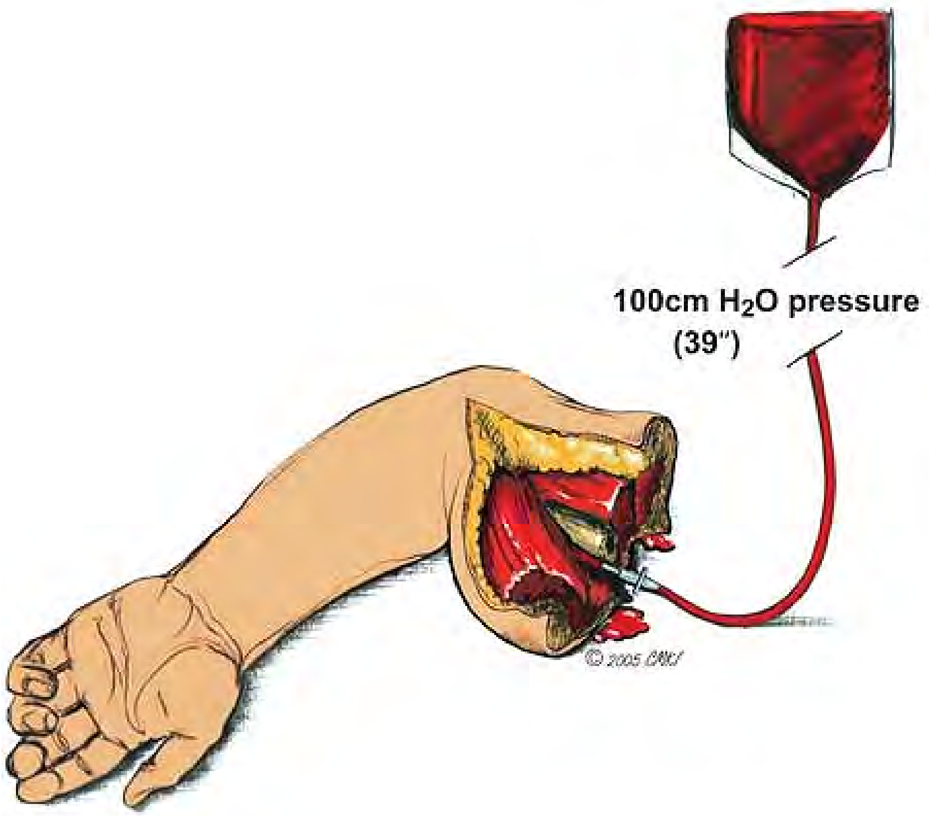

FIGURE 36-1 Perfusion of an amputated limb with heparinized blood. The blood bag is hung 100 cm above the amputated limb and the infusion flows by gravity.

The order for arm replantation is quite different from finger replantation. Before proceeding with replantation, lactated Ringer's solution or plasma is used to flush organ preservation solution from the limb. The recommended order of surgical procedures is as follows:

- Arterial shunt or perfusion to shorten anoxia time
- Thorough debridement
- Bone fixation
- Artery and vein repair, with or without vein graft

- Repair of nerves
- Repair of muscle and tendon
- Skin grafting or flap for coverage
- If indicated, fasciotomy

The artery from the amputated limb is cannulated with a standard vascular shunt to establish the connection between the proximal and distal artery and to intermittently perfuse the devascularized limb. A segment of regular intravenous infusion tube can also be used if no vascular shunt device is available. During perfusion, if the patient becomes hypotensive, temporarily clamp the perfusion. After 20 minutes, the venous shunt can be performed to minimize blood loss.

Thorough debridement is essential for a successful replantation. The injury may occur in an environment such as a farm, an automobile accident, or water sport in which the wound would be severely contaminated. Any crushed, grossly contaminated tissue must be removed. Muscles distal to the amputation level that have lost their contractility must be debrided. Effective debridement converts a dirty, crushed wound to a clean and guillotine-like wound. Residual dead tissue provides a substrate for bacterial growth after major limb replantation. The resulting gas gangrene or septicemia from postoperative infection may be life threatening. In addition, bone infection may result in non-union, delayed functional recovery, and secondary amputation.

Bone shortening is almost always the rule; usually a shortening of at least 2 to 3 cm is required. Bone shortening not only provides more healthy bony structure for an improved chance of bone

healing, but also helps to release tension on the soft tissue and eliminates the need for vein grafting, nerve grafting, or tendon grafting.

Either internal or external bone fixation can be used, depending on the surgeon's preference. Internal fixation can be chosen for more definitive treatment, especially if there is minimal concern for infection. Plates and screws provide better stability and are most commonly used. Kirschner wires can be used when the amputation is around wrist level. Although intramedullary nailing is an option, it is rarely used. External fixation should be considered when there is a high risk of infection or when internal fixation is contraindicated. External fixation can be used for temporary fracture stabilization until soft tissue is healed. When the limb is viable, internal fixation can be implemented. Occasionally, however, external fixation can be a problem when free tissue transfer is required for soft tissue coverage. If this is the case, external fixation should be switched to internal fixation at the time of the free tissue transfer procedure or should be avoided from the beginning, if at all possible.

One should repair as many arteries and veins as possible, although there is no magic number for optimal results. In general, at least one artery and two vein anastomoses are required. The two-to-one ratio also holds when greater numbers of anastomoses are performed: Repair two veins for each arterial repair. The stump of the vessel on each end is gently debrided to remove crushed or severely contused vessel. Special attention is paid in cases of avulsion to make sure there is no hidden intima injury that might cause delayed thrombosis and, eventually, failure of replantation. The adventitia at the stump is then carefully trimmed to provide better exposure for anastomosis. A syringe with a small angiocath filled with Tsai's solution (150 mL lactated Ringer's + 3,000 units heparin + 30 mL 1% preservative-free Lidocaine) is used to flush the vessel lumen to remove any blood clot or debris (see Table 36-1). Vascular anastomosis is performed under a microscope with 8-0 or 9-0 nonabsorbable suture in interrupted fashion. For a more proximal amputation, for example, at the upper arm, anastomosis can be achieved with 6-0 or 7-0 suture. Clamps are removed after each anastomosis is completed, and the replanted part is observed for return of circulation. In the past, we would give a bolus dose of heparin (3,000 units) followed by a low-molecular-weight dextran (LMD)/heparin drip (500 mL LMD + 5,000 units heparin at 20 mL/hr) for anticoagulation. Nowadays, after the bolus of heparin, we use a low fractionated heparin (e.g., enoxaparin sodium injection or Lovenox 30 mg given subcutaneously) due to its efficacy and simplicity. If there is any sign of vessel spasm, the replanted part should be kept warm with warm normal-salineâ€“moistened gauze or abdominal pads. Occasionally, papaverine (1.5 mg/mL) can be used to reverse vasospasm (see Table 36-1).

If the gap of the arterial defect is too big for primary anastomosis without increased tension, arterial reconstruction with a vein graft should be considered. The donor site of the graft can be a superficial vein from the contralateral arm or a saphenous vein from the lower extremity. The vein graft should be reversed routinely to prevent the valve effect. Venous anastomosis is performed in the same manner as arterial repair, but doesnâ€™t require reversal if a vein graft is used. As many veins as possible should be repaired.

Careful monitoring is necessary to avoid a systemic effect from revascularization (i.e., hyperkalemia, increased creatinine, and myoglobinuria). In complete amputation, the first 100 to 200 mL of venous blood, carrying with it toxic metabolites, is allowed to bleed out without going back to systemic circulation. For incomplete amputation, especially when part of the venous system is intact, the repaired artery is intermittently clamped to allow adjustment of systemic circulation.

All major nerves should be repaired, including the radial, median, and ulnar nerves and their branches. Epineurial repair with interrupted 8-0 or 9-0 nonabsorbable suture under a microscope is recommended. If a nerve graft is required, it can be obtained from the lateral or medial antebrachial cutaneous nerve or the sural nerve, depending on the length and size of the nerve graft desired.

Different techniques of tendon suturing can be chosen, according to the surgeon's preference. Strong tendon repair with early protective range of motion provides the best chance of functional recovery. If muscle and tendon are severely damaged, a tendon transfer is performed.

The authors recommend that fasciotomy be performed in anticipation of postischemic swelling. In the forearm, both the volar and dorsal compartments must be opened (Fig. 36-2). The incision on the volar distal forearm can be made along the ulnar side of the distal forearm to minimize tendon exposure. In the hand, all nine compartments (thenar, hypothenar, three volar, and four dorsal interosseous) are opened. This can be achieved with four incisions: one each on the side for the thenar and hypothenar eminences, and two dorsal incisions along second and fourth metacarpal bones in order to approach all interosseous compartments (Fig. 36-3).

Soft tissue coverage for exposed vital structures becomes easier after bone shortening and, in some cases, no extra skin is required. In most cases, primary wound closure can be accomplished with skin

grafting only. Occasionally, especially in complicated amputations, additional soft tissue coverage is required. This can be achieved by forming a pedicled or free vascular flap. It can be done either in an immediate, emergent setting or in a semi-emergent setting (i.e., a few days after initial surgery).

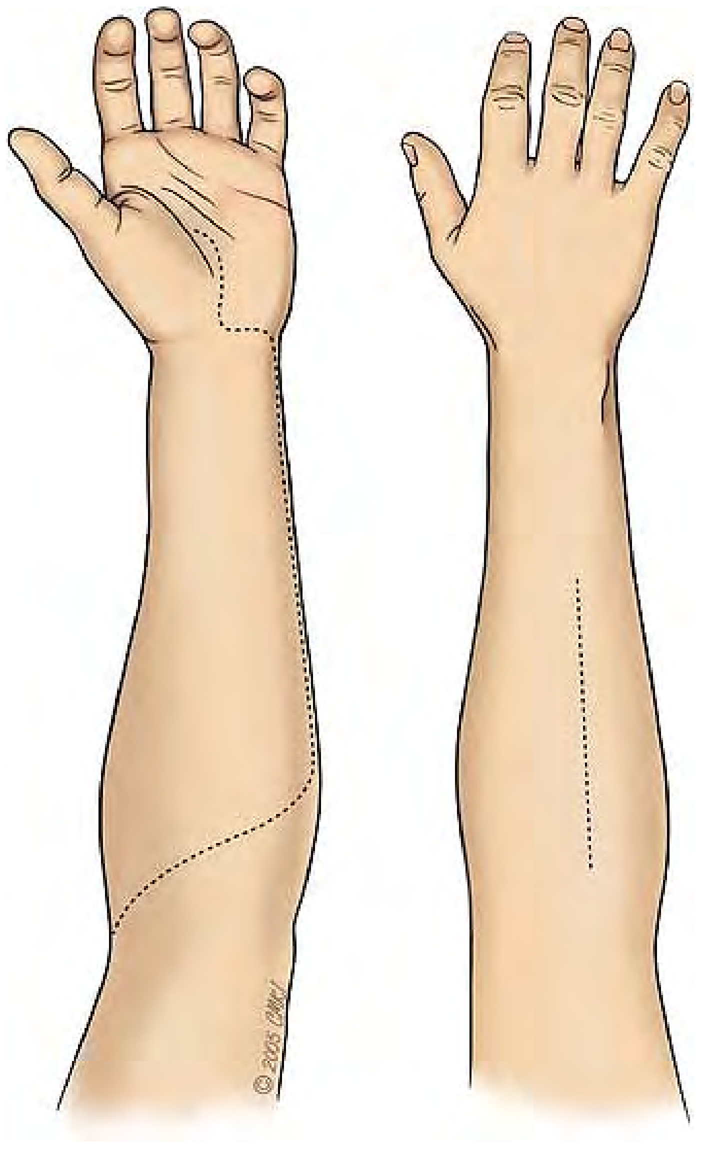

FIGURE 36-2 Volar distal forearm incision along the dorsal side.

Postoperative Management

The patient is admitted and is given nothing by mouth for at least 12 hours in case it becomes necessary to return to the operating room. The replanted limb is elevated to heart level with an intravenous pole or pillows and is monitored at least every hour for color, temperature, blanching, and capillary refill. Among them, temperature is the most sensitive parameter. A temperature drop of more than 2Â°C represents an early sign of vascular compromise. It is also important to make sure that the capillary refill is the result of arterial flow, not venous congestion. The patient is instructed to refrain from smoking and caffeine products for 3 weeks after surgery because they may decrease circulation to revascularized tissue. Dressing change is avoided in the first 3 days to minimize the chance of vasospasm.

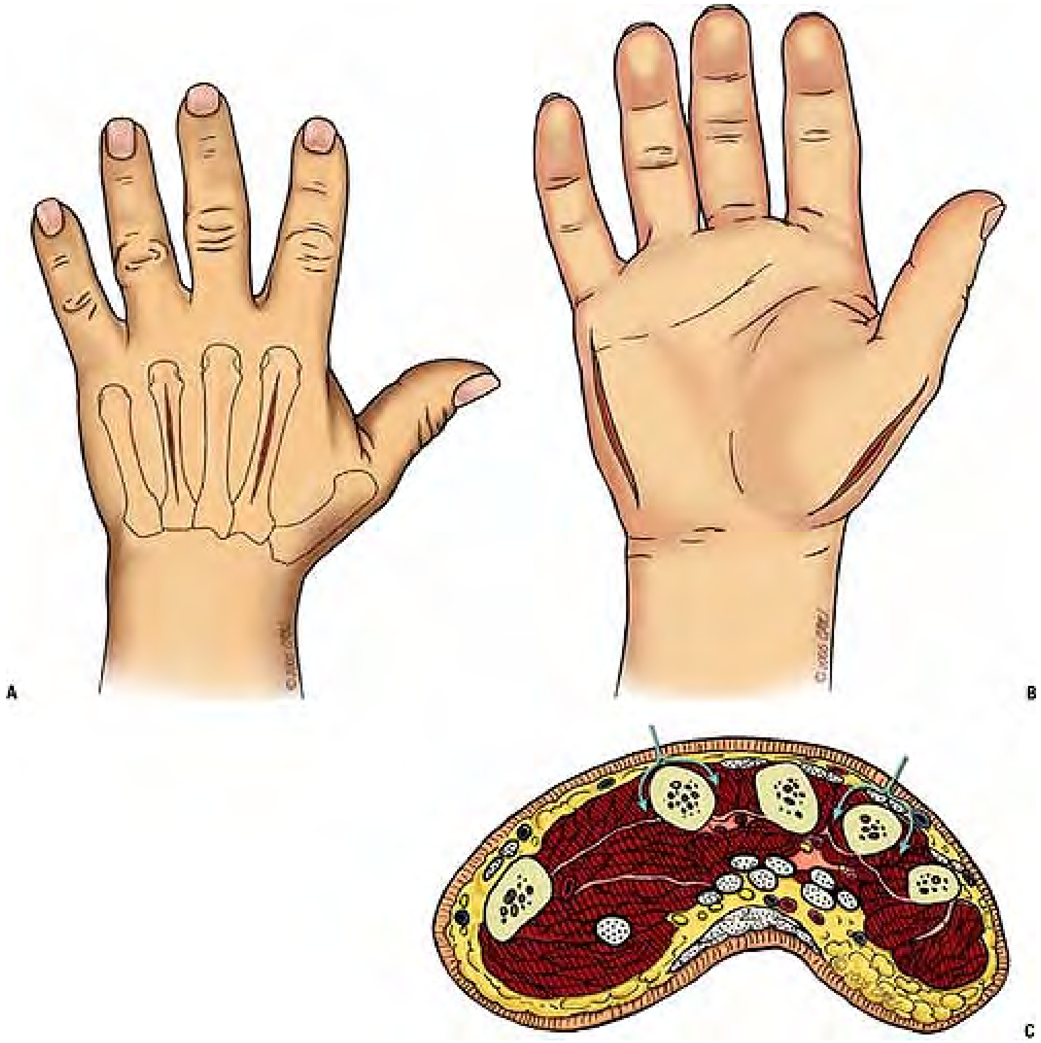

FIGURE 36-3 A: Dorsal incisions along the 2nd and 4th metacarpal bones. B: Incisions on the thenar and hypothenar eminences. C: Interosseous compartments.

By tradition, heparin and LMD are routinely given intraoperatively and postoperatively. However, a new trend is to give one dose of heparin when revascularization is completed or, sometimes, not giving any heparin at all. In these cases, patients are given only aspirin postoperatively. Heparin and LMD are given only when there is any doubt about the efficacy of circulation. Prostaglandin E_1 is also reported to provide some benefit because of its effects of general vasodilation, leukocyte adhesion inhibition, and platelet aggregation inhibition.

The dose given is 40 to 120 Âµg per day for 3 to 10 days.

Hyperbaric oxygen (HBO) treatment is very useful when the replanted limb shows marginal circulation. According to Nylander et al, HBO treatment significantly reduces phosphorylase activity, a sensitive marker for muscle damage, in the post-ischemia phase. Hyperbaric oxygen treatment is given at 2.5 atm of absolute pressure for 45 minutes, and usually three treatments are necessary.

In the weeks and months that follow replantation, additional surgeries may be required to improve the function of the replanted limb. These procedures may include debridement, tenolysis, and improvements in coverage.

Complications

Postoperative complications include persistent ischemia from the no-flow phenomenon, hyperkalemia/metabolic acidosis, myoglobinuria/renal failure, tissue necrosis, and wound infection. Any of these complications may either jeopardize the survivability of the replanted limb or endanger the life of the patient. The no-flow phenomenon is mainly due to long ischemic time and is often irreversible. Although free-radical scavengers (e.g., allopurinol) can be used to treat this condition, the result is often unsatisfactory.

During surgery, frequent test measurements should be used to monitor for the development of hyperkalemia and metabolic acidosis. High alertness and close monitoring of serum potassium during surgery, especially right after the vascular clamp is released, are key. The vessel should be intermittently clamped in the first couple of minutes of revascularization to allow time for adjustment. It's also important to maintain adequate hydration and avoid high potassium-content IV fluid. If it does develop, hyperkalemia/metabolic acidosis can be treated with alkalization and IV insulin. According to Waikakul et al, serum potassium concentration in the amputated segment was the best objective predictor of replantation success. When it is higher than 6.5 mmol/l 30 minutes after reperfusion, replantation should be avoided. A high systemic venous serum potassium concentration was also found before clinical signs of the reperfusion syndrome were seen.

Myoglobinuria and renal failure can be prevented with adequate hydration, alkalinization of urine, osmotic diuretics (mannitol), and free-radical scavengers. Frequent urine myoglobin monitoring is required until two consecutive results are within normal range. Tissue necrosis and infection are mainly due to inadequate debridement; therefore, adequate debridement is essential.

During the last 10 years, cases of myoglobinuria or renal failure are rarely seen following tissue perfusion or shunt anoxia time. Compared with minor replantation, patients with major replantation require longer hospital stays, more surgical procedures, and the functional results are not always satisfactory. The process also has a high psychiatric impact to both patient and family members.

Ultimately, the most important measure of outcome is the recovery of hand function. If there is no recovery of hand function, forearm amputation and a switch to below-elbow prosthesis can be considered. In one study from our institute, 8 of 34 patients with apparently successful major replantation underwent secondary amputation. Among them, four amputations were done because of poor function, one because of dysfunctional pain, and one because of poor cosmesis. The other two were part of planned staged procedures to preserve the length of the stump or joint. In general, a replanted limb tends to have a better functional outcome compared with a prosthesis. A replanted limb typically has better performance in gross prehensile activities and allows pronation and supination that prostheses do not offer.

In summary, arm replantation remains a major challenge to surgeons. The surgeon needs to consider the safety of the patient, to preserve limb viability, and to reconstruct the limb function. The final goal of a successful replantation is to achieve adequate function without pain. In addition, cosmesis must not be ignored, and secondary amputation sometimes is required. According to our experience, ideal conditions do not guarantee ideal results, although more distal amputation and shorter ischemic time predict a more favorable outcome.

Case Studies

A 45-year-old man sustained a blast injury resulting in total amputation at proximal upper arm level. Total ischemic time was 11 hours before revascularization was accomplished (Fig. 36-4).

A 17-year-old boy was involved in a mining accident and sustained bilateral forearm amputations (Fig. 36-5).

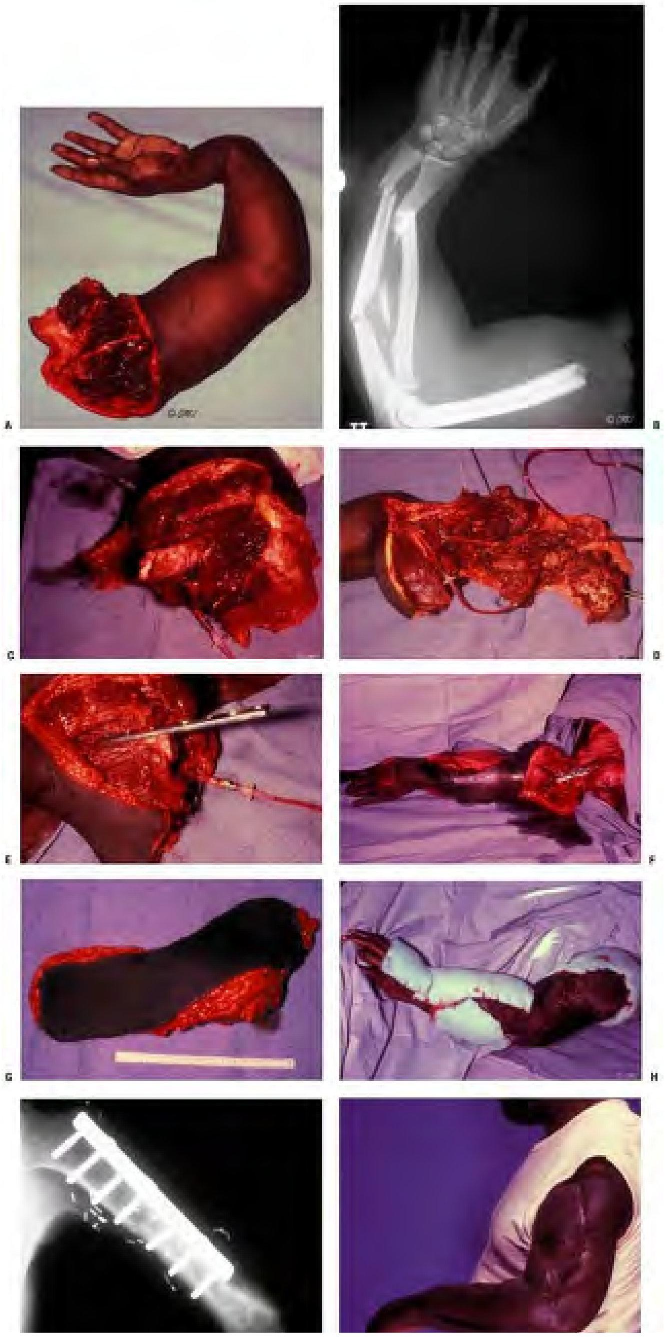

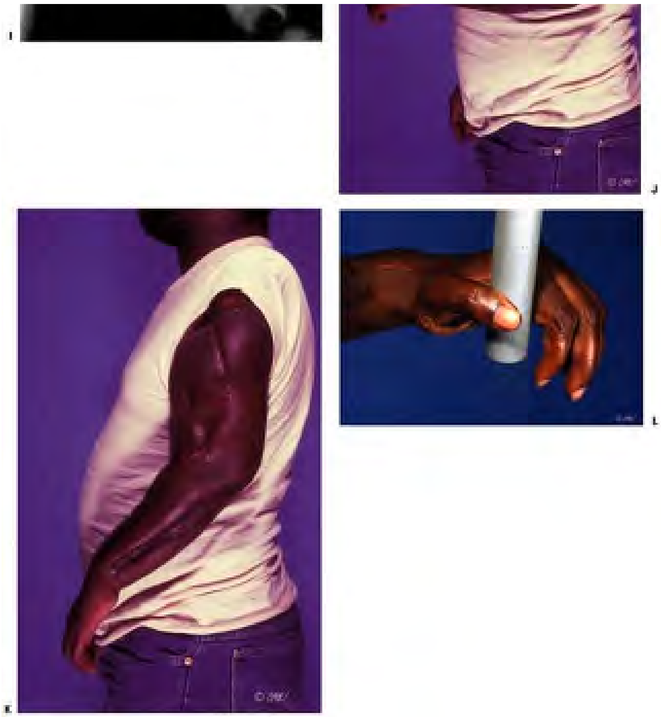

FIGURE 36-4 A: Amputated part demonstrated gross deformity at forearm level. B: Radiographic examination revealed amputation at proximal humerus level with segmental forearm fractures. C: Proximal part showed soft tissue defect with large area of bony exposure. D: Amputated part was perfused with 1 unit of heparinized arterial blood 2 hours before revascularization. E: While patient was getting ready for surgery, amputated limb was debrided and internal fixation partially readied for replantation. F: Completion of internal fixation, revascularization, and fasciotomy. G: Free tensor fascia lata flap for soft tissue coverage. H: Completion of initial surgery with free flap and skin grafting. I: Radiographic examination demonstrated healed humerus with plate in place. J: Final result after multiple surgical procedures including pectoralis major muscle transfer for elbow flexion. K: Final result of elbow extension. L: Final result of hand gripping. The patient has some hand gripping function but has very little intrinsic muscle function. Sensory recovery is protective only.

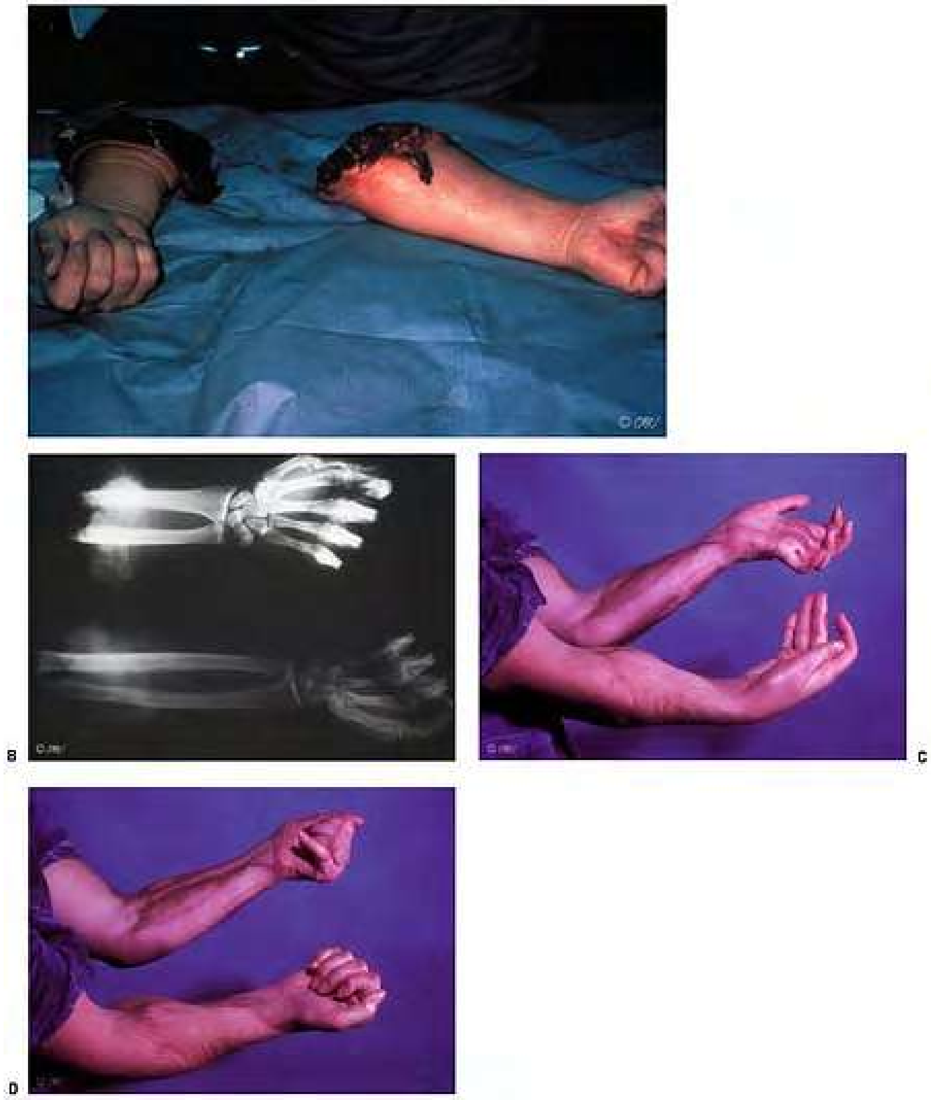

FIGURE 36-5 A: Right forearm amputated at mid-forearm level and left forearm amputated at proximal forearm level. B: Radiograph of amputated parts. C: Finger extension at final follow-up. D: Finger flexion at final follow-up.

Recommended Reading

Bandyk DF. Vascular injury associated with extremity trauma. *Clin Orthop Relat*

Editors: Moran, Steven L.; Cooney, William P.
Title: *Master Techniques in Orthopaedic Surgery: Soft Tissue Surgery, 1st Edition*

Copyright Â©2009 Lippincott Williams & Wilkins

Chapter 37

Replantation of Digits

Samir Mardini

Kuang-Te Chen

Fu-Chan Wei

Indications/Contraindications

Replantation is defined as reattachment and revascularization of a completely amputated part, whereas revascularization is defined as restoration of arterial inflow or venous outflow to or from a part that is incompletely amputated. In revascularization cases, it is of utmost importance that the true severity of the injury is not underestimated due to the grossly intact image of the injured finger. The introduction and improvements in microsurgical instrumentation and techniques have lead to successful revascularization and replantation of amputated parts as distal as the finger tip and pulp. Overall, the results of finger replantation are encouraging and have provided many amputation victims with extremely rewarding results.

Amputations of upper extremity digits are devastating for the patient who is initially stunned by the events of the trauma. Therefore, care is taken to discuss the situation in a gentle and thoughtful manner, considering that the patient has just lost a part of the upper extremity, and that this loss might be permanent if the reconstructive effort is not indicated or is unsuccessful. Generally accepted indications for finger replantation include amputations of the thumb in almost every situation, a single digit with the level of injury distal to the superficialis insertion, multiple digits, and any amputation in children (Table 37-1). The patient's desires have a strong influence on the approach.

Success in replantation is based on the functional outcomes as judged by objective parameters such as strength, range of motion, and sensory function, and not on the survival of the replanted or revascularized parts alone; therefore, the potential outcome must be considered for each individual patient. The potential for a long operative procedure, immediate or late failure, the potential requirement of regional or distant tissue flaps to complete the reconstruction, and the possible need for multiple procedures even after survival of the replants should be included in the formula for decision making and discussed with the patient and his or her family. The location, mechanism of injury, number digits involved, level of injury, condition of the amputated part, condition of the hand, length of ischemia time, the patient's general health, age, and associated injuries, and the patient's desires dictate the course of action. One advantage of replanting any digit is prevention of neuroma formation. A thorough knowledge of the anatomy and adequate exposure of structures aids in attaining optimal results (Figs. 37-1 and 37-2).

The thumb is the most important single digit, and its replantation is the most rewarding since provision of a post for opposition, even with lack of motion at the metacarpophalangeal and interphalangeal joints, is significant. In avulsion cases, the defects of the vessels and nerves may be segmental and the use of vein grafts for the arteries and veins, and nerve grafts or vein grafts for the nerves, may be required. The long nerve stumps of the amputated thumb can be anastomosed to the dorsal sensory nerves in the first web space, to a transposed nerve from the index finger, or to the more proximally located nerve stumps using nerve grafts. Despite the fact that sensory return may only be protective in nature, the replantation of thumb avulsion injuries usually results in significant improvement in function.

Patients presenting with multiple digit amputations are candidates for replantation, and efforts are made to replant all fingers with potential for good functional outcomes. The surgeon uses his or her judgment in deciding the location where an amputated part should be replanted. Replanting a digit onto the stump of the thumb is the first priority, and the other amputated digits should be replanted with the idea that a better grip is achieved when the long, ring, and small fingers are replanted without gaps between the digits, and better fine pinch is achieved when the index and long finger are replanted. Efforts are made to replant each amputated digit to its natural position; however, less severely damaged amputated digits may be better placed onto a stump that is less damaged and is located in a more functionally significant position. For example, in a case of multiple digit amputations with involvement of both the thumb and index finger, in which the amputated thumb has sustained irreparable damage, the index finger is replanted to the stump of the thumb. After the thumb position has been filled, the replantation should begin with the one most appropriate for the patient's daily activities and functional needs at work.

Table 37-1. Indications and Contraindications for Finger Replantation

Indications	Controversial Indications
Amputations of the thumb at any level Amputation of multiple digits Any amputation in children Single-digit injury in zone I	Single-digit amputation at a level proximal to the insertion of the flexor digitorum superficialis tendon Ring avulsion injuries Severe contamination
Contraindications	
Amputated parts that are severely crushed or damaged Multiple-level amputations Significant associated trauma and/or medical conditions	

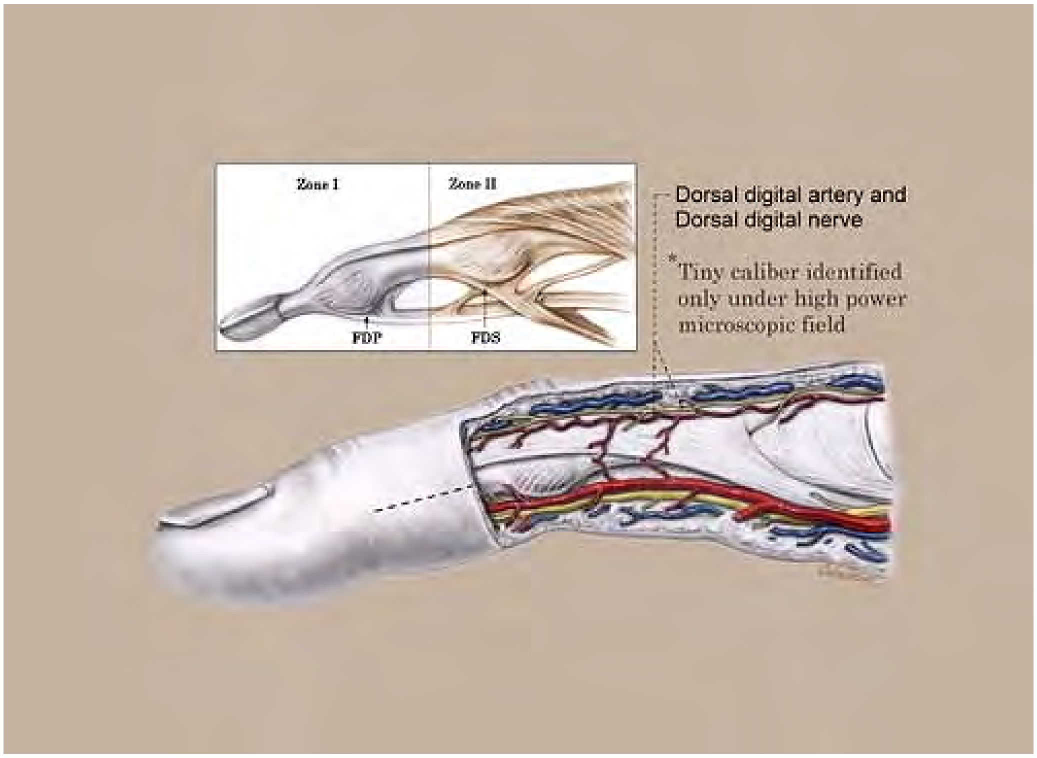
Zone I
Zone II
FDP
FDS
Dorsal digital artery and
Dorsal digital nerve
*Tiny caliber identified
only under high power
microscopic field

FIGURE 37-1 Lateral view of a finger demonstrating the location of arteries, veins, and nerves relevant to finger replantation. Also demonstrated are zones I and II as well as the level at which incisions are made for exploration of the neurovascular bundle.

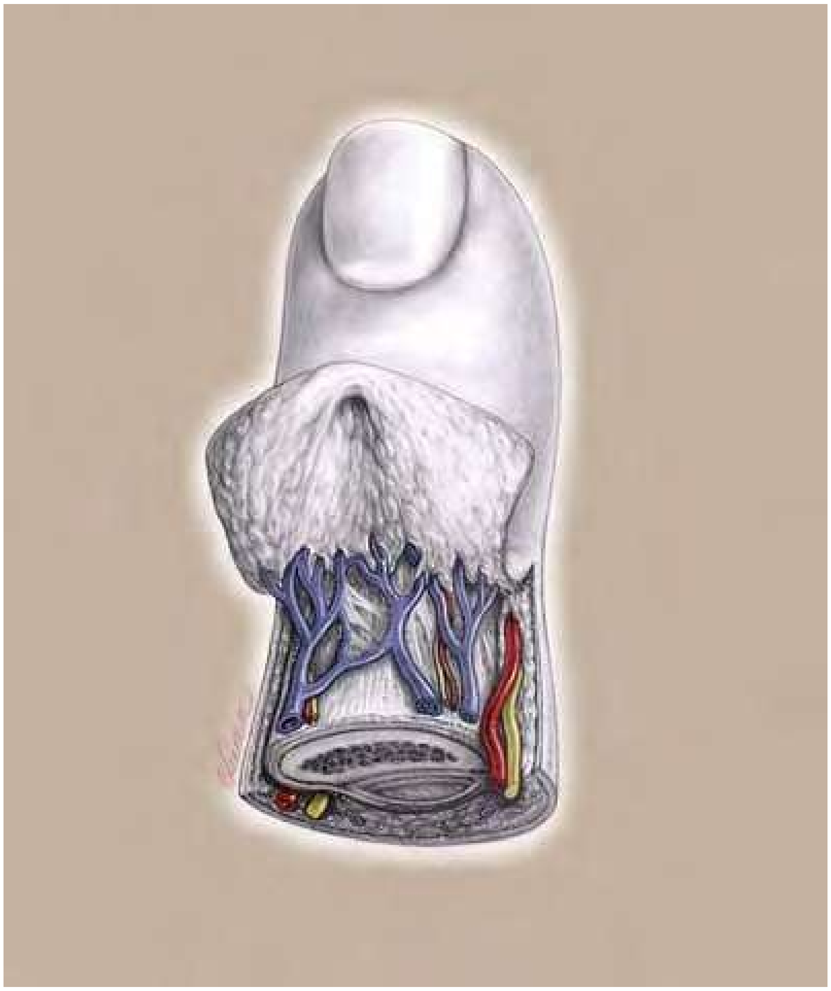

FIGURE 37-2 Mid-axial incisions and exposure of structures in the amputated part.

Single-digit amputations distal to the insertion of the flexor digitorum superficialis (FDS) tendon (zone I) yield good outcomes following replantation, with little need for secondary surgery, even if the distal interphalangeal (DIP) joint requires fusion. In some instances, venous outflow cannot be reconstituted and postoperative leech therapy or heparinization and purposeful exsanguination can achieve digit survival. Four millimeters of dorsal skin proximal to the nail fold must be present for adequate vein size to be repaired. Tsai, McCabe, and

Maki described their technique of finger tip replantation, which used leech therapy and other methods of decongesting the finger when venous congestion developed postoperatively. These authors noted a 69% survival rate, a 75% rate of return of some form of two-point discrimination, and an average range of motion of the DIP joint of 56 degrees. Replantation of a single-digit amputation at a level proximal to the insertion of the flexor superficialis tendon and distal to the first annular pulley (zone II) often results in stiffness of the proximal and DIP joints and limited motion of the finger, with the replanted finger interfering with movement of the other fingers. Therefore, single-digit replantation may be only suitable for well-motivated individuals who are willing to undergo appropriate postoperative physical rehabilitation for a maximal functional recovery. Unless they are highly motivated, concerned about cosmetic outcome, or are children, patients are not usually replanted in this situation, as they often require secondary surgery and end up with disappointing outcomes.

Ring avulsion amputations present with segmental injuries of structures such as arteries, veins, and nerves; therefore, vein and nerve grafts are often necessary to bridge the gaps after debridement of unhealthy appearing neurovascular tissue. Otherwise, transposition of the neurovascular bundle from an adjacent digit (Fig. 37-3) or the use of arteriovenous shunting for inflow to the replanted digit can also yield successful outcomes. Most surgeons consider ring avulsion amputation patients to be poor candidates for replantation; however, replantation of an avulsed digit still can be attempted if the patient insists on replantation and he or she understands the likeliness of a suboptimal outcome.

The mechanism of injury plays an important role in decision making. Sharp, incisive injuries predict better functional outcomes than crush or avulsion injuries.

In elderly patients and patients with compromised medical conditions, the decision is based on the balance between the risk of undergoing the surgery, the functional outcome anticipated, and the significance of the replantation to the patient. Life-threatening injuries and significant medical conditions that preclude long anesthesia times are contraindications to digit replantation. Microsurgical replantation in children or infants is indicated and can yield good functional outcomes. In elderly patients, the surgeon should take into consideration the potential systemic insult from the anesthesia and operation, the arteriosclerosis of vessels, and the poorer potential of functional recovery.

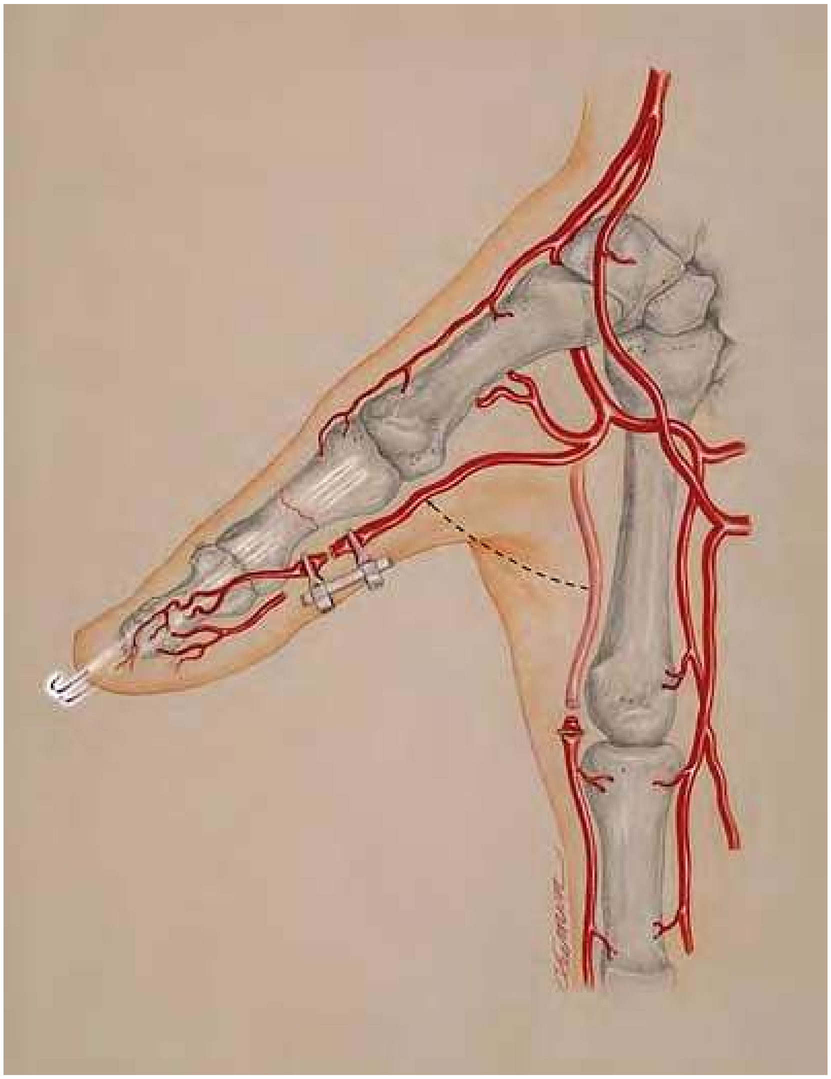

FIGURE 37-3 Transposition of an artery from an adjacent finger.

Warm ischemia times of less than 8 to 12 hours and cold ischemia times of less than 30 hours, measured from the time of amputation to the time of restoration of blood flow, are tolerated in digit replantation. Survival of replanted digits has been reported after as long as 42 hours of warm ischemia and 94 hours of cold ischemia. Immediate and proper cooling of the amputated part lengthens tissue survival.

A preoperative history of cigarette smoking does not seem to confer as negative an effect on the replantation as postoperative smoking. It has been shown to have significant negative effects on blood flow for up to 3 months. If the patient is unwilling to stop cigarette smoking, this could be considered a contraindication.

Amputated digits that are not commonly replanted are ones that are severely crushed or damaged, severely contaminated, have undergone multiple-level amputations, or are those of patients with multiple associated traumas or severe medical problems.

Preoperative Planning

A team approach is essential to the success of replantation. The transport team and the emergency room physicians should place the amputated part in appropriate dressings and arrange efficient transport to a center that has a replantation team. This team includes microvascular surgeons, anesthesiologists, operating room personnel and facilities, a microscope, and postoperative monitoring capabilities.

Minimal time should be wasted in obtaining unnecessary examinations. Many of the studies can be performed on the patient while the operating room is being prepared and, subsequently, the amputated finger is being prepared. If the operating room is ready and the patient is not, the finger can be taken to the operating room and prepared for replantation.

Obtaining a history and physical examination is necessary to aid in determining indications for surgery and avoiding postoperative complications and disappointing outcomes. Current events related to the injury are detailed, including the time and mechanism of injury, the location and conditions where the trauma took place, and the way the amputated digits or stumps were handled prior to admission. Information regarding the patient's general condition, including other injuries that occurred in relation to the current trauma and the patient's medical and surgical history, are obtained. The general examination of the patient is performed to exclude other major injuries. The amputated digit and the stump are examined in the emergency room. This examination is performed without digital blocks, which may cause injury to structures essential for successful replantation.

The amputated part is carefully preserved in a cool condition to minimize damage from ischemia. The amputated part is wrapped in gauze moistened with Ringer's lactate or saline solution, and placed in a plastic bag or specimen cup, which is then placed on ice. The amputated part is never placed directly on or immersed in ice or ice water, as this can cause freezing and permanent damage to the tissues.

Radiographs of the amputated digit and hand are taken to evaluate the condition of the bones and the presence of foreign bodies. The patient is kept in a warm environment, provided with adequate hydration, tetanus prophylaxis, and broad-spectrum antibiotics that are appropriate for the mechanism of injury and condition of the wound.

A discussion with the patient and the patient's family should include the risks of anesthesia, possibility of long operating times, potential need for blood transfusions, requirement for postoperative hospitalization, possibility of further surgery to salvage the finger in case of vascular compromise, the likely need for secondary procedures such as tenolysis, and the definite need for extensive physical rehabilitation postoperatively to achieve optimal outcomes. Patients are informed of the possibility that revision of the stump and closure of the wound might be performed; that vein grafts, nerve grafts, skin grafts, local flaps, regional flaps, or free flaps might be necessary to complete the reconstruction; and that bone shortening is often necessary in the primary reconstruction. The patient is made aware that there is a chance of immediate or late failure of the reconstructive effort and that other options for rehabilitation, such as prosthetic devices, might need to be resorted to in the future. Finally, if the patient is a smoker, he or she is told that smoking can affect the outcome and that postoperative smoking for up to 3 months can be detrimental to the final outcome of the surgery. The patient is given clear and realistic expectations regarding the expected outcomes.

Appropriate consultations with medical specialists are obtained. If a psychiatric disorder is suspected, a psychiatric consultation is obtained which in some circumstances might affect the decision as to whether to proceed with the replantation.

If an operating room is available, the amputated part is taken there first, examined under magnification, and set up for replantation while the patient is being prepared and anesthetized. Replantation is performed as soon as possible to shorten the ischemic time.

Surgery

Patient Positioning

The patient is placed in a supine position with the arm abducted and placed over an arm table. A tourniquet is placed high up on the arm and on one of the lower extremities (in case vein, nerve, or skin grafts are required). Appropriate warming devices are placed to maintain normal body temperatures throughout the procedure.

General anesthesia is recommended, as replanting each digit requires between 2 to 4 hours to complete, and this usually becomes uncomfortable for the patient when an axillary block is performed. A brachial plexus block can be used to decrease the requirement for pain medication and decrease the incidence of vascular spasm in the operating room and postoperatively. Some authors use a brachial plexus block in addition to sedation in adults and older children with single-digit amputations in one extremity who are thought to be non-anxious and may tolerate 2 to 5 hours of immobility.

Preparation and cleansing of the stump and amputated part are performed using a povidone-iodine solution followed by copious irrigation with normal saline. A tourniquet is placed on the involved arm and one leg, and both are prepared and draped.

Technique

A two-team approach is used whenever possible, with one team working on the amputated part and the other working on the hand. The operative sequence for digit replantation varies slightly among surgeons and depends on the individual situation and ischemia time; however, our general routine is (1) identification of structures, (2) debridement, (3) bone shortening and fixation (Fig. 37-4), (4) repair of the extensor tendon, (5) repair of flexor tendons, (6) anastomosis of the artery, (7) coaptation of nerves, (8) anastomosis of veins, and (9) closure of the wounds.

Identification of Structures

During the dissection, the amputated part can be kept cool by placing it over a cooled apparatus. Bilateral mid-axial incisions are made, and the volar skin flap and dorsal skin flap are elevated, folded back, and sutured in position (see Fig. 37-2). If one team is performing the surgery, the exploration begins by identifying the neurovascular bundles and other structures in the amputated part using an operating microscope. High magnification minimizes inadvertent injury to important structures and decreases the chances of vasospasm. The digital arteries, one or two dorsal veins, and the digital nerves are identified and tagged with 6-0 sutures. The extensor and flexor tendons are located, trimmed, and tagged. If there is no artery available in the amputated part for anastomosis as is sometimes the case with ring avulsion amputations, the options for replantation become more limited. The use of an afferent arteriovenous shunt for the revascularization of unfavorable amputations of fingertips and distal phalanges may be an option (Fig. 37-5), or the replantation effort is halted and a revision and closure of the stump is performed (12,28â€“30).

Debridement

All non-viable tissue is excised while being careful not to cause inadvertent injury to important structures including the pulleys, particularly the A2 and A4 pulleys.

Bone Shortening and Fixation

Without excessive periosteal stripping, the bone is trimmed to provide surfaces that allow for good bone contact during fixation. Bone shortening also allows for direct vessel or nerve repair without tension, minimizes the need for using vein grafts and nerve grafts, and permits a more relaxed skin closure as postoperative edema can create compression of the vessels. Preservation of 5 to 10 mm of bone near a joint helps increase the range of motion that can be achieved. In traction avulsion amputations, the bone may require further shortening (19,36). Amputations through a joint are commonly treated with arthrodesis at the time of replantation. If this is not performed in the initial setting, eventual use of a silastic arthroplasty can solve some problems related to joint stiffness but is insufficient when the collateral ligaments are not present or do not provide adequate stability.

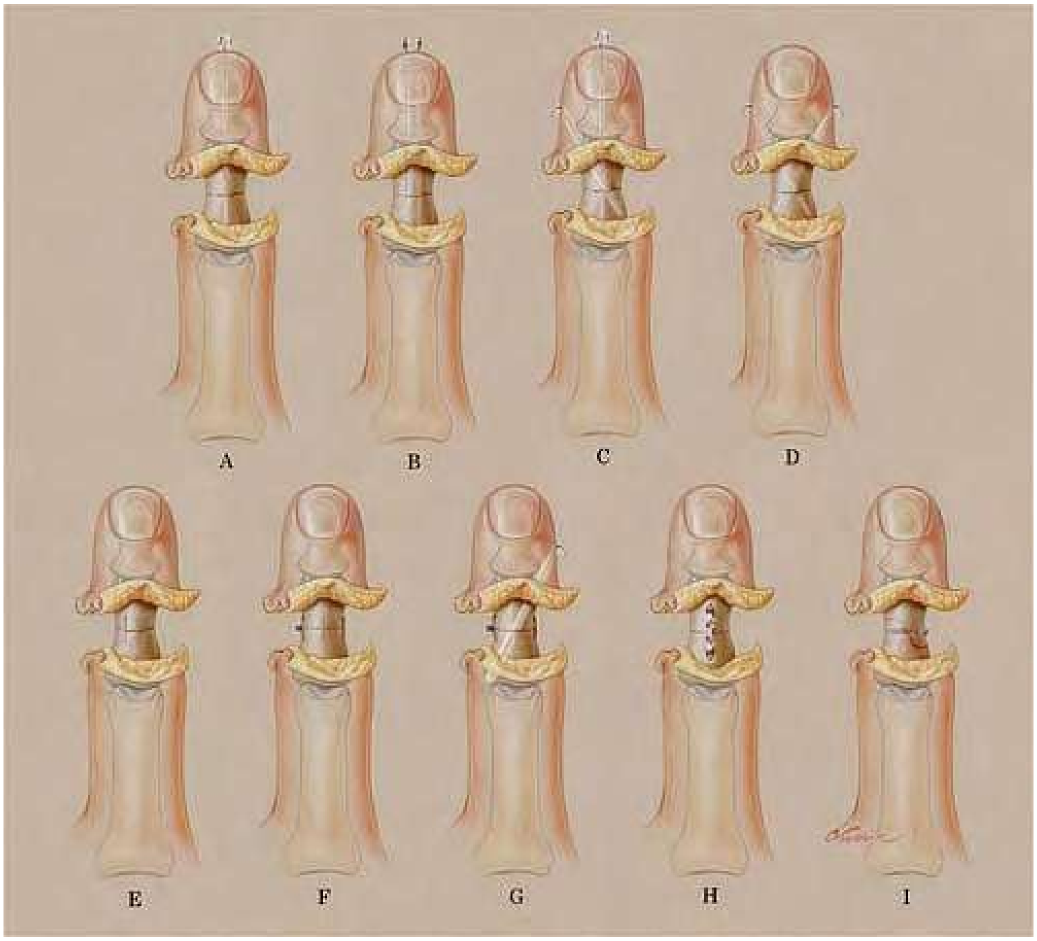

FIGURE 37-4 Different methods used for bone fixation. (*A*) One longitudinal Kirschner pin, (*B*) Two parallel Kirschner pins, (*C*) one longitudinal and one oblique Kirschner pins, (*D*) two oblique Kirschner pins, (*E*) two parallel interosseous wires, (*F*) two perpendicular interosseous wires, (*G*) one interosseous wire and one Kirschner pin, (*H*) plate fixation, (*I*) a screw for oblique fractures.

Bone fixation can be performed using crossed Kirschner wires (K-wires), two longitudinal K-wires, a single intramedullary K-wire, interosseous wiring, intramedullary screws, or external fixation devices (see Fig. 37-4). The type of fixation used depends on the stability of the fragments, the desire for early mobility, the compliance of the patient, and the preference of the surgeon. Plate fixation may be beneficial for fractures at the mid-shaft level as it allows for early mobility; however, plate fixation requires more extensive soft tissue dissection and periosteal stripping, and requires more time to perform. Longitudinally placed intramedullary or crossed K-wires that do not cross non-involved joints are the most commonly used methods. For replantations distal to the DIP, one K-wire may suffice. The authors prefer the use of interosseous wires when the replantation is performed at the level of the proximal part of the proximal or middle phalanx. Interosseous wiring provides adequate stabilization for early mobility and allows for adjustment of the alignment on an outpatient basis if malalignment is noted within the first week postoperatively.

Repair of the Extensor Tendon

The extensor tendon is repaired using 4-0 medium-lasting sutures placed as horizontal mattresses or in a figure-of-eight fashion. Excessive and unnecessary manipulation, pinching, and grasping of tendons are avoided. In amputations at the proximal phalangeal level, the disrupted lateral band is repaired to achieve optimal extension function of the interphalangeal joints.

Repair of Flexor Tendons

The flexor tendons are repaired by applying a modified Kessler technique using 4-0 nonabsorbable sutures followed by a running circumferential suture using a 6-0 monofilament nonabsorbable suture. Both the FDS and flexor digitorum profundus (FDP) are repaired meticulously. In zone II, only the FDP tendon is repaired to minimize subsequent tendon adhesion.

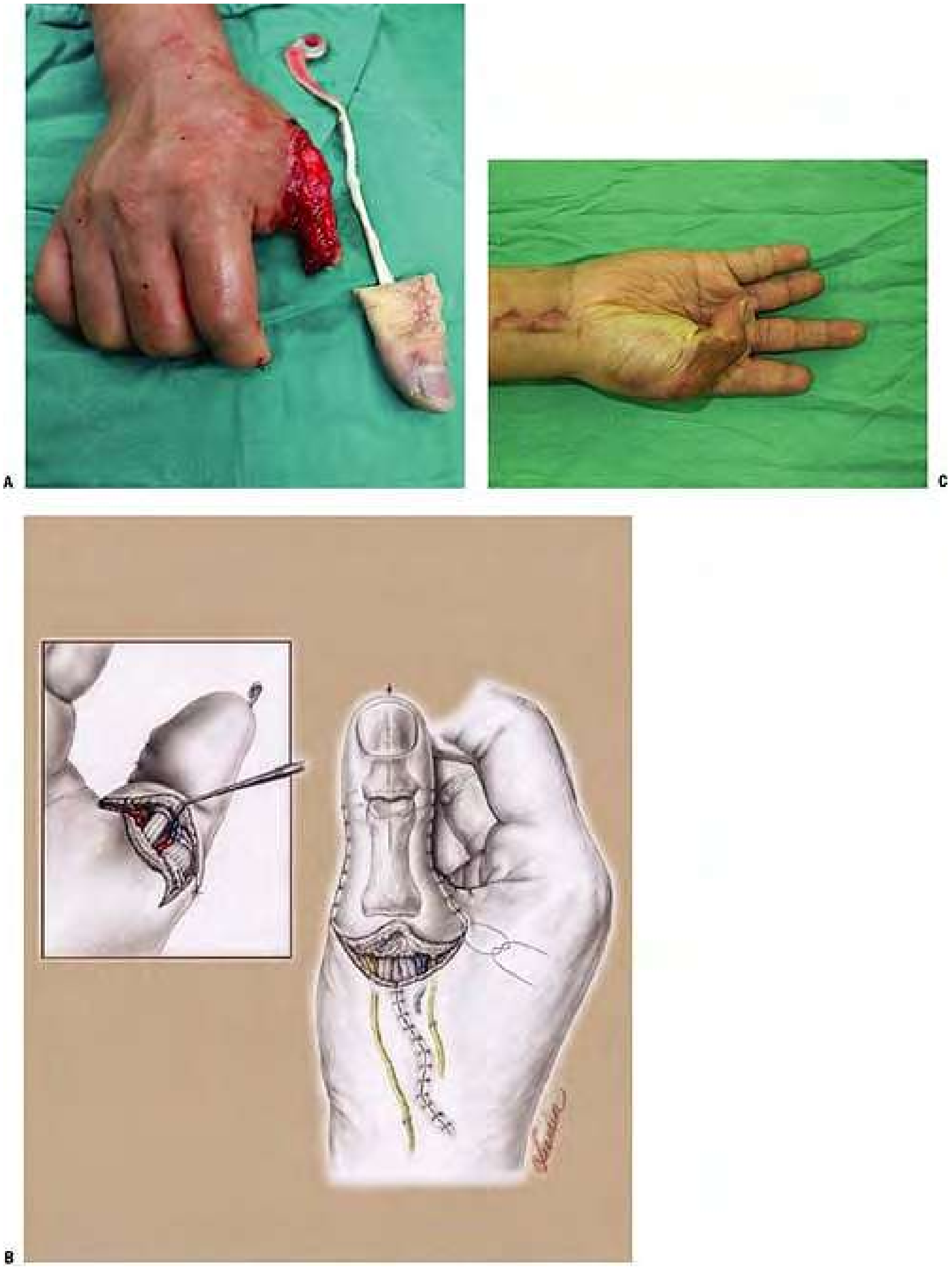

FIGURE 37-5 A: A 28-year-old man suffered from an avulsion amputation of his right thumb. The thumb was disarticulated at the interphalangeal joint and the skin envelope was avulsed proximally near the level of the metacarpophalangeal joint. Both digital arteries were avulsed from the distal amputee around the interphalangeal joint. The digital nerves were avulsed proximally near the level

of wrist. The flexor pollicis longus tendon was avulsed at the level of the tenomuscular junction. B: Illustration demonstrating intraoperative details. The recipient artery in the hand was anastomosed to one of the volar veins to create an arteriovenous fistula. One dorsal vein was used as the outflow of the finger (a vein graft was necessary to bridge the gap). The bone was fixed using one longitudinal Kirschner pin through the interphalangeal joint. Both nerves were coapted to the superficial branches of the superficial radial nerve at the level of the anatomic snuffbox. C: Six months after replantation of the thumb using the arteriovenous technique.

Anastomosis of the Artery

The artery on both the amputated part and the proximal stump are examined under microscopic visualization and are cut at a level that causes no apparent damage to the vessel wall. The presence of bruising or thrombus, or the appearance of a coiled artery, indicates the need to trim the vessel further. The presence of thrombus within the lumen, or separation of the vessel wall layers, indicates damage and requires the vessel to be trimmed further. The outflow from the proximal stump of the artery is checked and should be spurting before it is considered adequate for revascularization. Hydration of the patient and maintenance of a systolic blood pressure over 100 mm Hg are important to achieving this purpose. Adventitia stripping often aids in relieving and preventing vessel spasm. Vein grafts from the volar aspect of the forearm or the dorsal foot are used as conduits whenever the gap is too large to complete the anastomoses without tension.

Revascularization can be provided by mobilizing a healthy undamaged artery from the adjacent digit (see Fig. 37-3). This technique provides inflow from a vessel that is unaffected by trauma, and obviates the need for vein grafting with its inherent risks. This can be performed in conjunction with transfer of the digital nerve in cases of thumb replantation. Vein grafts, nevertheless, are reliable, readily available, and avoid the need for incisions and scars in adjacent non-injured fingers.

When the digital artery is not available on the amputated part, afferent arteriovenous shunting can be used to perfuse the digit. Fukui and Yabe used this technique and found that a higher success rate was found when the inflow was from the artery of the stump to the volar veins rather than the dorsal veins. Anatomic studies by Moss and coworkers revealed that the direction of valves in the oblique communicating veins between the dorsal and volar veins favor flow from the volar veins to the dorsal veins. This technique can be successful in replantation of amputations as high as the interphalangeal joint (see Fig. 37-5).

If after release of the clamps the replant does not pink up, a waiting period up to 30 minutes is suggested. The vessels and the anastomosis are carefully inspected, adventitial stripping is performed, and the patency test is performed. Side branches, both proximal and distal to the anastomosis, can be tested to see if blood flow exists. Redo of the vascular anastomosis may become necessary because of either inadequate debridement of injured vessels or thrombosis at the anastomotic site due to inadequate microsurgical technique. Injection of high concentration heparin solution and/or thrombolytic agents is performed if all other factors are optimized and reperfusion seems to be inadequate.

Coaptation of Nerves

The nerve repair is performed by approximating the epineurium of both ends of the digital nerves using 10-0 or 11-0 sutures. Nerve grafts are used liberally to avoid any tension on the repair. Depending on the severity of the injury and outlook for wound coverage, secondary nerve grafting may be a better option to avoid exposure of this nonvascularized graft which may act as a source of infection and may result in poor function. The medial antebrachial cutaneous nerve and the sural nerve provide good sources of nerve fascicles for grafting purposes. In multiple-digit amputations, nerve grafts can be harvested from parts that are not planned for replantation. In avulsion cases of the thumb, sensory nerves from the avulsed finger can be attached to the dorsal sensory nerves of the first web space to restore protective sensibility.

Anastomosis of the Veins

The veins of the amputated part are inspected carefully after releasing the arterial clamps, and the ones with the most bleeding are anastomosed to the recipient veins in the hand. Dorsal veins are preferred for anastomosis. After separating the dorsal veins from the skin for a short distance and ligating a few branches, the length is usually adequate to complete the anastomoses primarily.

At least two veins are anastomosed whenever possible. Matsuda and coworkers evaluated the correlation between the number of anastomosed veins and survival rate in finger replantation. Their conclusions were that the essential amount of venous anastomoses is one in zone II (lunula to distal interphalangeal joint), two in zone III, and one in zone IV (proximal phalangeal region).

In situations where perfusion of the flap is good and there are no available dorsal veins for anastomosis, volar veins may be used. If there are no available veins or no outflow from any veins after a long waiting period, a last resort is to inspect the contralateral digital artery. If retrograde flow is present from the contralateral artery, that artery may be anastomosed to a recipient vein which can serve as an outflow vessel for the finger.

When replanting finger tip amputations and no sizable vein is present, other options include removal of the nail to keep continuous bleeding from the raw nail bed, postoperative periodic digital massage, and/or use of medical leeches. All these methods are performed with or without systemically heparinizing the patient.

Vein gaps can be bridged with vein grafts from the volar forearm or dorsal foot. If soft tissue is also lacking in the area where a vein graft is needed, a venous flap can be used which will provide a vein graft and a soft tissue cover.

Skin Closure

Meticulous hemostasis is obtained prior to closure of the wound. Under microscopic visualization, free ends of disrupted veins and arteries are carefully ligated. If only one artery is used for a finger, the contralateral digital artery is ligated to avoid postoperative bleeding. The skin is loosely approximated with a few interrupted sutures. The vessels are checked to make sure there is no compression or kinking with placement or manipulation of the hand and finger in different positions. Although a skin graft can be placed directly over the pedicle, rotation of a local skin flap is preferred and the defect from the rotation is skin grafted. Antibiotic ointment, non-adherent dressings, and wet gauze are placed over the incisions. No significant dressings are placed between the fingers to avoid potential compression of the digital vessels. Loosely placed circumferential dressings cover the hand.

Additional Procedures

Carpal tunnel release is performed for patients who have undergone avulsion amputations. The transverse metacarpal ligament is released, and the carpal tunnel is explored. In cases of severe avulsion amputations of the digits, forearm fasciotomies may be necessary.

Results

Survival of replanted fingers ranges from 70% to 95% in both children and adults, with many patients achieving extremely satisfying results (Figs. 37-6 and 37-7). Survival rates are reported to be lower in avulsion and crush injury cases. Re-exploration due to vascular compromise is necessary in fewer than 20% of replantations and carries a salvage rate of approximately 70%.

Replantations of single or multiple digits when the amputation is distal to the sublimis insertion are highly satisfactory, both in terms of function and appearance, even without motion at the DIP. Secondary surgeries in these cases are rarely necessary.

A replanted thumb results in satisfactory outcomes in most patients, regardless of the type of injury or level of amputation (see Fig. 37-6). Even if a replanted thumb has a limited range of motion, it still plays an important role as a post for opposition. Schlenker et al reviewed their experience with 64 thumb replantations and found that 50% of the patients had two-point discrimination of less than 10 mm and an average active range of motion after replantation of 35% of normal for the interphalangeal joint and 29% of normal for the metacarpophalangeal joint. Most patients were able to return to work after approximately 7 months.

Sensory reinnervation is extremely important for the fingers to achieve optimal function. Glickman and Mackinnon reviewed the recovery of sensibility in 367 fingers and 87 thumbs that were replanted successfully. Average static two-point discrimination (S2PD) was 9.3 mm in clean thumbs versus 12.1 mm in crush/avulsion thumb replantations. Average S2PD was 8 mm in incisive injuries and 15 mm in crush/avulsion finger replantations. Mean S2PD was 11 mm in thumb replantations and 12 mm in finger replantations. Sixty-one percent of the thumbs and 54% of the fingers regained S2PD that was useful. They found that the factors that influenced sensibility were the age of the patient, the level and mechanism of injury, digital blood flow, cold intolerance, and postoperative sensory re-education. It was concluded that recovery of sensibility in the replanted digit is comparable to simple nerve repair and to nerve grafting techniques.

Cold intolerance is a significant problem following replantation. The exact mechanism of cold intolerance following digital nerve injury is not clear but may have a neurogenic or vascular origin. The more recovery of sensibility, the less cold intolerance seems to occur. Cold intolerance usually improves after approximately 2 years, but can persist for many more years.

Postoperative Management

The patient is transported to a unit equipped with a warming device and highly trained nursing staff capable of performing hourly checks and recognizing signs of vascular compromise (Table 37-2). Every check includes an inspection of finger color, capillary refill, skin Turgor, and temperature. Pinprick testing and evaluation of vascular flow by hand-held Doppler ultrasonography are performed when the finger displays clinical signs of circulatory compromise. Temperature monitoring of replanted digits has proved a reliable indicator of vascular status, as replanted digits are located in a terminal site with a terminal vascular loop that will not be influenced to a significant degree by heat exchange from the underlying bed. In general, a drop in temperature greater than 2Â°C from that of the control digit, or below 30Â°C, indicates the possibility of circulatory compromise.

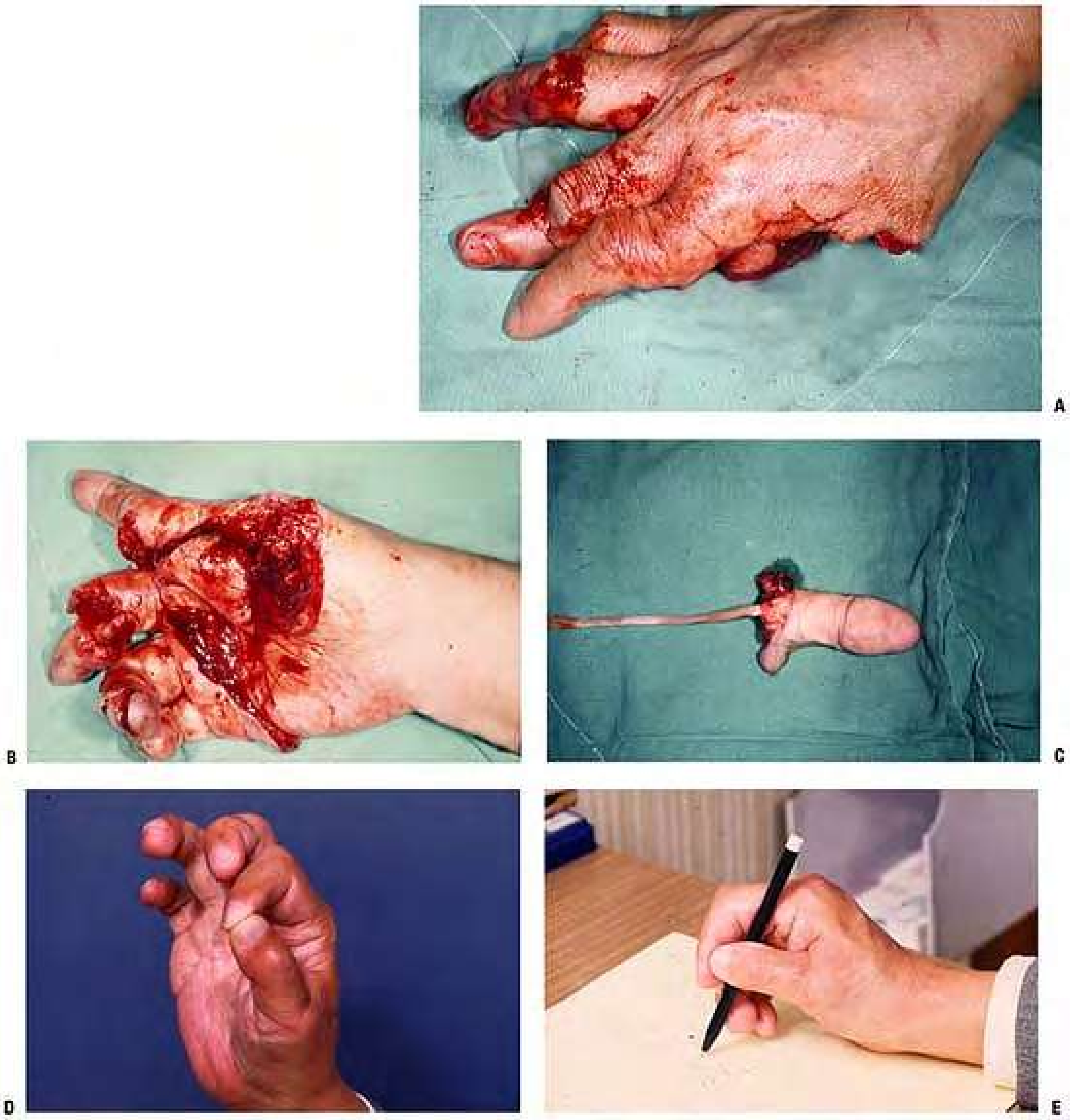

FIGURE 37-6 A: A 57-year-old man sustained a crush injury of his right hand by a rolling machine. B: The thumb was completely amputated and the index and middle fingers were incompletely amputated at the level of the proximal phalanx and middle phalanx, respectively. C: The thumb was avulsed at the level of the proximal phalanx and the flexor pollicis longus tendon was torn from its tenomuscular junction. D: Nine months follow-up demonstrating adequate positioning of the replants and good strength. E: At 9 months follow-up the patient is able to grasp a pen.

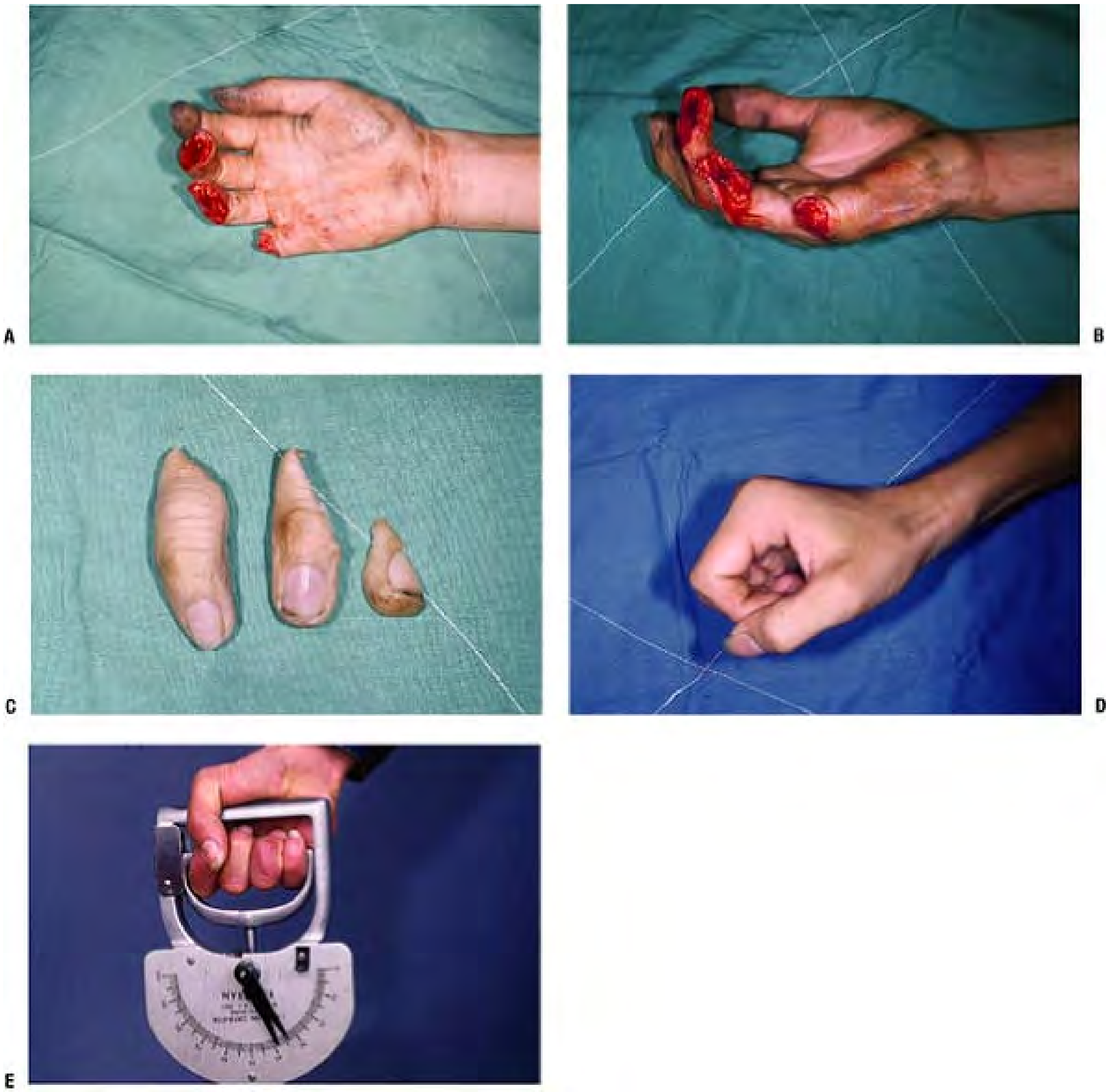

FIGURE 37-7 A: An 18 year-old man suffered from a sharp cutting injury by a machine saw. Volar view demonstrating the level of amputation of the three ulnar digits. B: This lateral view demonstrates the level of amputation of the long, ring, and little fingers. C: Dorsal view of the three amputated digits. D: At 9 months follow-up the patient has a good cascade and is able to make a fist. E: At 9 months follow-up, the patient is able to grip with good strength.

Table 37-2. Signs of Arterial and Venous Compromise in Finger Replantation

	Arterial Compromise	Venous Compromise
Finger color	Mottled, bluish, or pale	Bluish, cyanotic, or dusky
Capillary refill	Sluggish (>2 sec)	Brisker than normal
Tissue turgor	Flat, decreased turgor	Tense, increased turgor
Temperature	Cool (>2Â° difference compared to control)	Cool (>2Â° difference compared to control)
Pinprick test	Scant amount of dark blood or serum	Rapid bleeding of dark blood
Doppler signals	Absence of pulsatile arterial signals	Absence of continuous or venous augment signals

A pulse oximeter placed distally on the replanted digit can be used to monitor changes in oxygenation. Laser Doppler flowmetry and photoplethysmography are used at some centers, but the cost of the instruments and the need for interpretation of the data have not proven its benefit over traditional methods such as clinical monitoring and temperature monitoring.

Intensive monitoring of replanted digits is undertaken for 3 to 5 days post replantation. Although thrombosis of vessels can occur up to 2 weeks following the replantation procedure, it is rare; and when it does occur, the chance of salvage is lower. This may be partly due to the delay in diagnosis in late thrombosis cases. In Betancourt's series of 71 replanted digits, the highest risk of critical thrombosis after replantation occurred in the first 3 days after surgery.

Patientsâ€™ activities are monitored closely for the first 3 to 5 days to ensure that injury to the replanted part does not occur. The arm is elevated above heart level to decrease swelling, pain, and venous congestion. The level of arm elevation should not be excessive as to decrease arterial pressure in the hand and digits. Urine output and electrolytes are monitored to ensure adequate hydration to avoid vasoconstriction. Intraoperative placement of a continuous axillary brachial plexus block catheter can be a source of postoperative pain relief and can provide a chemical sympathectomy which may be applied for 3 to 5 days postoperatively.

The use of anticoagulants is controversial; however, heparin is often used in replantation cases. The authors favor the use of low-molecular-weight dextran starting in the operating room at the time of completion of the first vascular anastomosis. Systemic heparinization is used in replantation of crushed or avulsed digits or in cases where intraoperative vascular thrombosis occurs and requires revision of an anastomosis. Some authors recommend the use of aspirin at 325 mg per day, with the first dose administered prior to the replantation.

A warming lamp placed in the region of the replant and warming the patient's room may have some benefit. All patients are advised to avoid caffeine intake and cigarette smoking for at least 2 to 3 months following the surgery.

Complications

Immediate complications include partial or complete loss of the replanted digit due to circulatory compromise, infection, and bleeding. The presence of infection can lead to vascular thrombosis. This can be partially avoided by performing an adequate debridement and administering proper antibiotics based on the level and type of contamination. Skin necrosis can also be seen in the early phase due to inadequate debridement or compromise of blood supply to a part of the replanted digit. This scenario requires debridement and closure with local skin flaps, skin grafts, or transfer of more distant flaps.

Vascular Compromise Post Replantation

Following the detection of a change in one of the parameters used to assess finger or thumb perfusion, the decision is made on whether to proceed with exploration. It is our practice to have an aggressive approach to re-exploration and attempts at salvage. If arterial occlusion is suspected, the patient is immediately taken back to the operating room for re-exploration. If venous compromise is suspected, several maneuvers may be attempted prior to re-exploration: the dressing and some of the sutures are removed, the limb is elevated, and the patient is placed on a systemic heparin drip. If the digit does not show signs of improvement within 20 to 30 minutes, it is highly likely that the veins draining the replanted finger have undergone thrombosis. In this situation, the decision is made by the surgeon whether to take the patient back to the operating room or pursue a nonsurgical approach. For venous outflow obstruction in distal replantation, management by using medical leech therapy or continuous bleeding can be performed with high success rates. The nail plate is removed, a stab incision in the paraungual area is made, and a heparinized saline drip is placed over the incision site to maintain external bleeding. The smaller the replanted digit, the higher the salvage rate using these methods. In Han's series of 144 finger tip replantations with venous outflow through the use of this technique, the salvage rate was 70% (101 of 144 digits). The average period of the salvage procedure was 7.6 days. The duration of bleeding was different for different types of injuries: guillotine injuries required 5.9 days, crush injuries 8.2 days, and avulsion injuries 8.0 days.

Medical leeches, *Hirudo medicinalis*, have been reported to be effective for the treatment of venous outflow obstruction. The leech secretes hirudin, a potent anticoagulant, which acts locally without significant systemic effects. The therapeutic period required is approximately 3 to 5 days. Significant blood loss and the risk of infection (7%â€“20%) are the main disadvantages of this modality. The organism involved in infections from leeches is *Aeromonas hydrophilia*, and a third-generation cephalosporin is prescribed as a prophylactic method to prevent infections.

On exploration, the artery and vein are inspected. Occasionally, a hematoma is the cause of problem; however, a venous or arterial occlusion is often found. The thrombus is evacuated and arterial inflow or venous outflow are re-established. Vein grafts are used to reach healthy vessels when necessary. The steps are similar to the previously described procedure, including the occasional need to infuse high-concentration heparin solution and/or thrombolytic agents.

Late Complications and Secondary Procedures

The mechanism and level of injury, ischemia time, quality of the original repair, the postoperative rehabilitation program, and the effort put forth by the patient as well as the healing process all determine the final outcome. Depending on the combination of these factors, the resultant quality of finger function and the patient's lifestyle and motivation will determine whether the patient will seek secondary procedures to improve outcome. The majority of patients undergoing replantation (approximately 30% to 40% of patients who underwent finger replantation and up to 90% of patients with injuries in zone II) will require one or more secondary procedures. The physical therapist and surgeon follow the patient closely after the replantation and provide an aggressive and thorough physical rehabilitation program. Once the improvement has plateaued, usually at 6 months post-replantation, an evaluation of finger function and overall hand function will determine the need for secondary procedures. The patient's desires and level of function are of utmost importance in determining the need for surgery.

Frequent sequelae include tendon adhesions and joint stiffness. Avoiding transfixation of non-involved joints and performing a meticulous tendon repair decreases the incidence of these complications. Tenolysis as a secondary procedure is required in approximately 50% of finger replantations and up to 90% in patients with the level of amputation in zone II. Stiff proximal interphalangeal (PIP) and DIP joints are sometimes acceptable if the metacarpophalangeal (MCP) joint is mobile and allows for adequate grip. This is particularly true for the more ulnar digits. If the patient is able to use the thumb, index finger, and long finger to pinch objects adequately, he or she may be satisfied with the result and elect not to undergo further surgery. During the procedure of flexor tenolysis and/or tendon grafting, care is taken not to injure the repaired neurovascular bundle during the dissection. A detailed knowledge of the previous operative procedure can aid in avoiding injury to repaired structures. Pulleys are preserved, particularly the A2 and A4 pulleys. The tendon is exposed and inspected. If tendon rupture has occurred, tendon grafts with or without tendon rods are necessary in some patients. Ruptured or atrophied tendons are debrided. If the tendon is of poor quality, particularly in zone II injuries, reconstruction of the pulleys and insertion of a tendon rod followed later by tendon reconstruction may be the procedure of choice. Tendon grafts can be used to reconstruct the tendons if the quality of tendon appears good and the pulleys are intact. Pulley reconstruction and tenolysis are usually not performed in the same procedure. Extensor tenolysis is performed through dorsal longitudinal incisions and usually yields good results. Intrinsic reconstruction and dorsal arthrolysis are worthwhile when active and passive flexion are present. Arthroplasty can yield excellent results and is indicated for some MP and PIP joints. Arthroplasty can be performed along with other procedures such as tenolysis, capsulotomies, and capsulorrhaphies. Arthroplasty is performed through the same volar incisions that are used for the flexor tenolysis, or dorsal incisions are made. Vascularized joint transfers are indicated for replacing an MP or PIP joint, a flaccid thumb, or a fused MP or PIP of the index or other digits when the global hand function is good. The second or third toe metatarsophalangeal, PIP, or DIP joints are used from the toes.

Certain procedures are not performed simultaneously, such as tenolysis with procedures that require finger or hand immobilization (i.e., nerve repair, osteotomies, joint fusions, vessel repair and bone grafts). Some procedures can be performed together, such as tenolyses and arthroplasties. Malunion or nonunion can occur in up to 20% of replanted digits and seems to occur more commonly with replantations fixed with K-wires. Neurolysis, nerve repairs, and nerve grafts are indicated when the Tinel sign fails to cross the site of repair for 3 months. Early repair is advocated. Nerve grafts are used if the nerve injury is on the same side as the vascular repair to avoid injury to

the vascular structures. It is important to find a healthy proximal stump for anastomosis. If not found, nerves from the dorsal radial sensory branches or the same or adjacent digits may be transferred.

Soft tissue coverage such as pulp transfers for thumb or opposable digits are sometimes necessary. Local flaps or free tissue transfer is sometimes necessary to cover open wounds or defects after contracture release of the flexor or dorsal side or the webspace. Perforator flaps and the lateral arm flap are good options for this.

Cold intolerance can be found in the majority of patients and most often resolves with time. Delayed circulatory compromise may occur up to 2 to 3 weeks post replantation, and often is caused by cigarette smoking or infection. Severe infections after digit replantation are rare and are usually related to inadequate debridement during the initial surgery or poor perfusion to the digit. Severe complications are unusual in digit replantation. Compression syndromes at the forearm or wrist level can occur and are usually associated with severe avulsion injuries. Fasciotomy and carpal tunnel release may be required in those cases. Web space contracture can occur as a result of thenar muscle contracture and/or soft tissue contracture in the web space. If the initial replantation is successful, late problems are usually related to suboptimal function due to poor motion and can be corrected with secondary procedures. Other secondary procedures that may be required are re-amputation, rotational osteotomies, and bone grafts.

Recommended Reading

Axelrod TS, Buchler U. Severe complex injuries to the upper extremity: revascularization and replantation. *J Hand Surg Am.* 1991;16(4):574â€“584.

Baek SM, Kim SS. Successful digital replantation after 42 hours of warm ischemia. *J Reconstr Microsurg.* 1992;8(6):455â€“458.

Betancourt FM, Mah ET, McCabe SJ. Timing of critical thrombosis after replantation surgery of the digits. *J Reconstr Microsurg.* 1998;14(5):313â€“316.

Buncke GM, Buncke HJ, Kind GM, Buntic R. Replantation. In: Achauer BM, Eriksson E, Guyuron B, Coleman JJ, Russell RC, Vander Kolk CA, eds. *Plastic surgery indications, operations, and outcomes*. St. Louis: Mosby; 2000:2131â€“2147.

Buncke HJ, Alpert BS, Johnson-Giebink R. Digital replantation. *Surg Clin North Am.* 1981;61(2):383â€“394.

Buncke HJ, Whitney TM, Hill MK. Bony fixation in replantation. In: Buncke HJ, ed. *Microsurgery: transplantation-replantation. An atlas text.* Philadelphia: Lea &

Latha Ganti
Editor

Atlas of Emergency Medicine Procedures

Springer

Arterial Cannulation (Radial and Femoral)

1

Jeffrey Kile, Katrina John, and Amish Aghera

Arterial cannulation is frequently performed in the care of critically ill patients for purposes of both serial arterial blood gas sampling and continuous intra-arterial blood pressure monitoring. It also provides arterial access for less common procedures, including thrombolysis, embolization, angiography, and infusion of vasoactive drugs. This chapter discusses cannulation of the radial and femoral arteries—the two most common sites for indwelling arterial catheter placement.

1.1 Indications

- Continuous monitoring of blood pressure in acute illness or major surgery
- Serial sampling of arterial blood during resuscitation
- Inability to use noninvasive blood pressure monitoring (*e.g.*, burns, morbid obesity)
- Continuous infusion of vasoactive inotropes (*e.g.*, phentolamine for reversal of local anesthesia)
- Angiography
- Embolization

1.2 Contraindications

- Absolute
  - Circulatory compromise in the extremity
  - Third-degree burns of the extremity
  - Raynaud's syndrome
  - Thromboangiitis obliterans (Buerger's disease)
- Relative
  - Recent surgery in the extremity
  - Local skin infection
  - Abnormal coagulation
  - Insufficient collateral circulation
  - First- or second-degree burns of the extremity
  - Arteriosclerosis

1.3 Materials and Medications

- Radial artery cannulation: standard over-the-needle catheter assembly (Fig. 1.1)
  - Antimicrobial solution and swabs
  - Sterile gloves
  - Local anesthetic (1–2 % lidocaine without epinephrine)
  - Blunt needle
  - 25- or 27-gauge needle
  - Two 5 mL syringes
  - 4″×4″ gauze sponges
  - Standard over-the-needle catheter assembly
- Additional materials required for radial artery cannulation: over-the-needle catheter assembly with integrated guide wire
  - Over-the-needle catheter assembly with integrated guide wire
- Additional materials required for femoral artery cannulation: The Seldinger Technique
  - Introducer needle
  - Guide wire
  - Scalpel
  - Dilator
  - Arterial catheter

J. Kile, MBBS, PhD, MPH (✉) • K. John, MBBS
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: jeffrey.kile@gmail.com; trenjohn@me.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_1

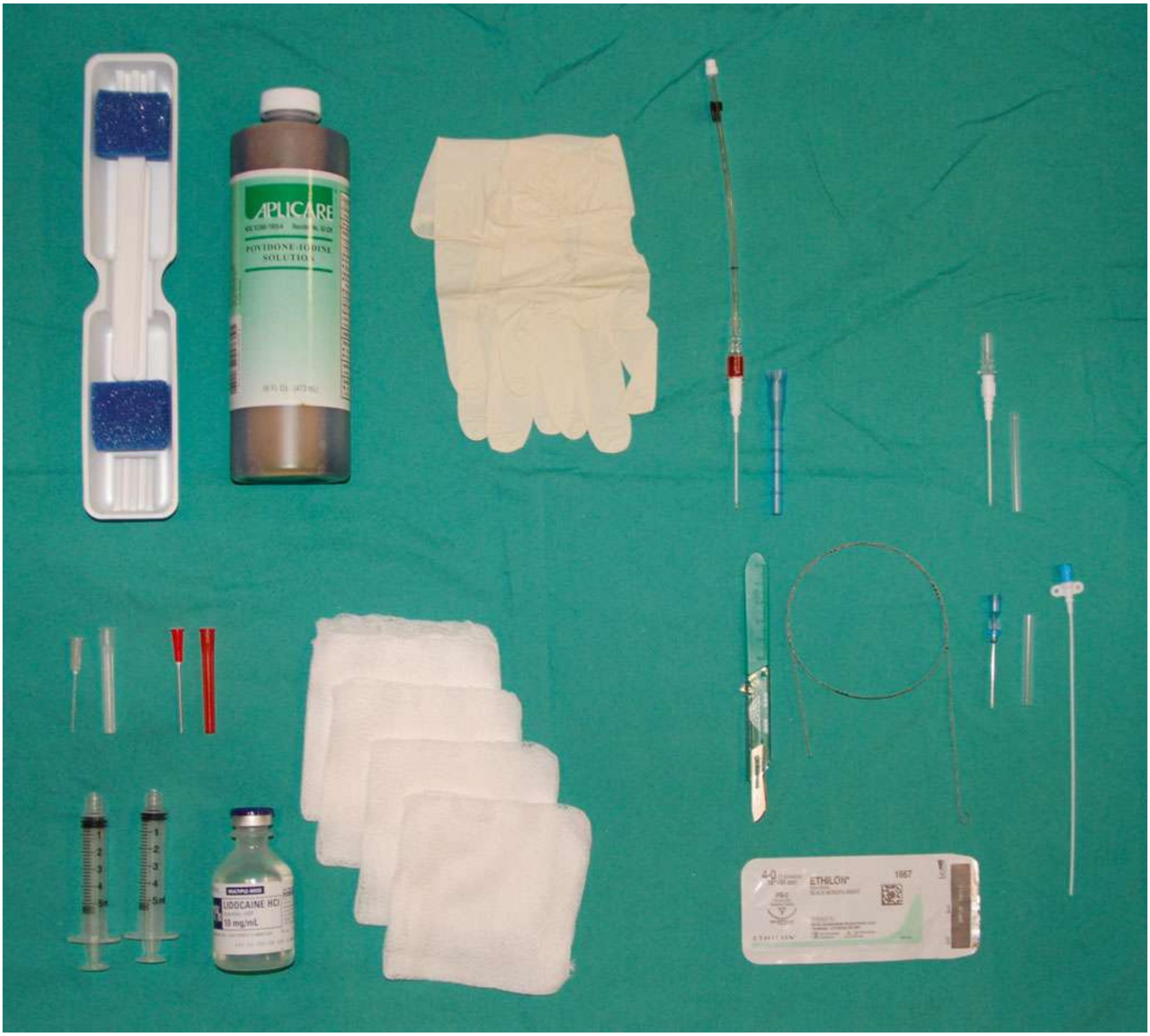

Fig. 1.1 Materials and medications

1.3.1 Procedure: Radial Artery Cannulation—Standard Over-the-Needle Catheter Assembly

1. Ensure adequate collateral flow in the selected extremity using the Allen test (see details below).
2. Immobilize the extremity by dorsiflexing the wrist to approximately 60° over a small towel roll and taping the base of the fingers to an arm board or other flat, fixed surface (Fig. 1.2).
    - Rotation of the wrist may shift the artery from its normal anatomical position, complicating cannulation.
3. Locate vessel by palpation of arterial pulse using the second and third fingers of the gloved nondominant hand.
4. Sterilize overlying skin with antimicrobial solution.
5. Inject local anesthetic to raise a small (0.5 cm) wheal using 25- or 27-gauge needle, and direct needle through wheal to infiltrate skin superficial to the artery with additional local anesthetic.
    - Infiltration of the subcutaneous tissue with local anesthetic may also reduce vessel spasm during arterial puncture.
    - Injection of local anesthetic into the vessel may precipitate arrhythmia, so draw back on the plunger prior to infiltration to ensure the tip of the needle is not inside the vessel.
    - Injection of excessive anesthetic when raising a wheal may obscure palpation of the pulse.
6. Ensure proper function of needle-cannula assembly by checking that cannula advances smoothly over needle.
7. Connect a 5 mL syringe with the plunger removed to the over-the-needle catheter assembly.
    - Attachment of syringe improves control during cannulation.
8. Hold syringe connected to needle-cannula assembly like a pen with needle bevel facing upward.
9. Directing the needle at a 30° angle to the skin, puncture the skin through the anesthetic wheal immediately overlying the palpated artery, and advance needle slowly until tip enters arterial lumen, which is confirmed by visible arterial blood flow (“flashback”) into the needle hub and syringe.
    - Avoid self-puncture by maintaining adequate distance between needle tip and index finger (Fig. 1.3).
10. Reduce the angle between needle and skin (by lowering the needle) and advance an additional 2 mm to ensure catheter tip (which sits approximately 2 mm behind the needle tip) has entered the lumen.
    - Advancing the needle too far (or failing to reduce angle between needle and skin) once initial flashback is visualized may result in piercing the back side (or “double puncture”) of the artery wall, in which case visible blood flow will cease; if this occurs, slowly withdraw needle several millimeters until pulsatile blood flow reappears.
11. Stabilize position of introducer needle and advance catheter alone into artery over needle until hub of catheter is in contact with the skin; blood flow from catheter hub at this point indicates successful cannulation of the artery.
    - If difficulty is encountered at this step, catheter hub may be rotated slightly to facilitate advancement.
12. Remove the needle without dislodging catheter from artery.
13. Manually apply pressure to proximal aspect of artery to occlude blood flow from the catheter.
14. Attach desired extension tubing, injection cap, and stopcock to the catheter hub.
15. Secure the catheter hub to the skin using silk (2.0) or nylon (4.0) sutures as follows. Take a 0.5 cm bite of skin under the catheter hub with the suture needle, tie several knots in the suture without pinching the skin, then tie a second set of knots around the hub of the catheter firmly. If the catheter assembly contains an integrated suture wing for fixation, take a 0.5 cm bite of skin under suture wing with the suture needle, thread suture through wing perforation, and secure wing against the skin with several knots. If suture wing has two perforations, repeat this process to secure other half of wing to skin (Fig. 1.4).
16. Cover the catheter with an appropriate self-adhesive sterile dressing.
    - A small bead of antibiotic ointment applied to the puncture site prior to dressing reduces the likelihood of cutaneous wound infection.
17. Secure the tubing connected to the catheter with gauze and adhesive tape or other sterile dressing.
18. Ensure all connections extending from catheter are tight and well secured, as accidental disconnection may result in rapid exsanguination.

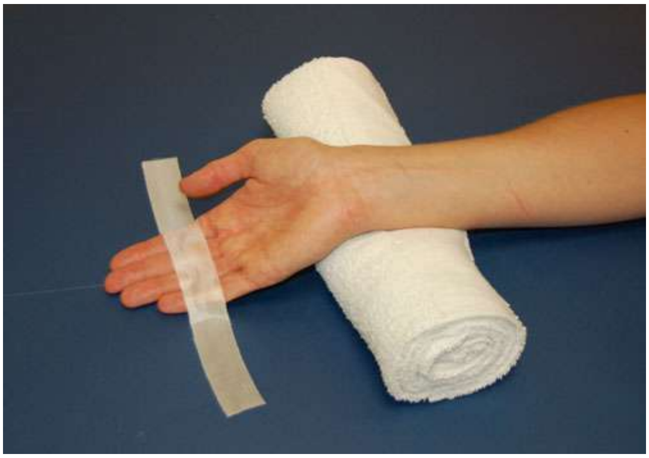

Fig. 1.2 Correct position of wrist prior to cannulation

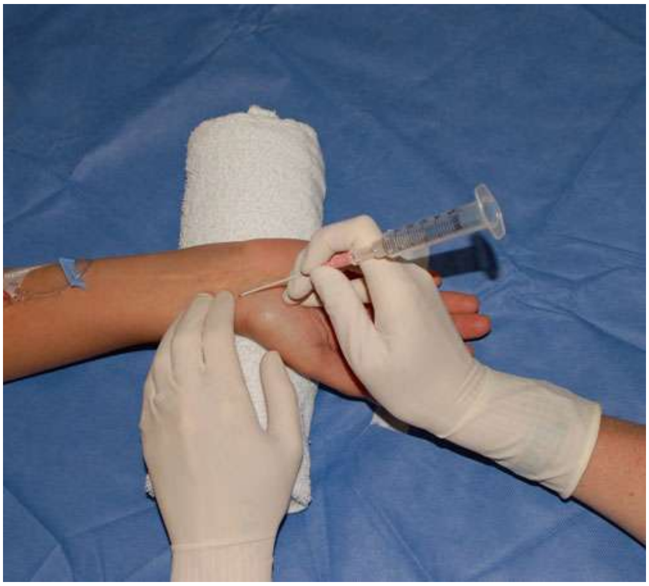

Fig. 1.3 Puncture of radial artery with standard over-the-needle catheter assembly

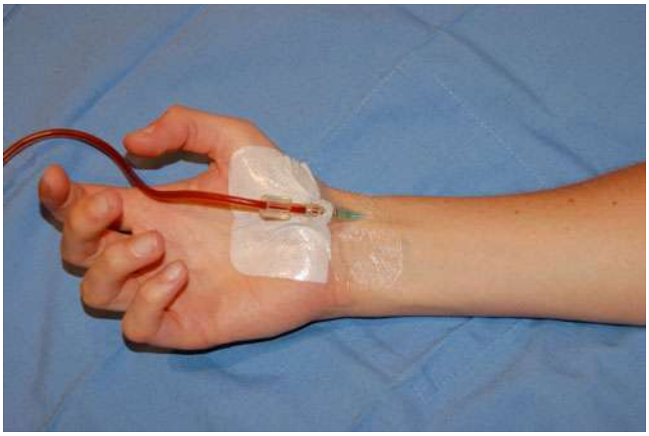

Fig. 1.4 Radial arterial catheter secured to wrist

1.3.2 Procedure: Radial Artery Cannulation—Over-the-Needle Catheter Assembly with Integrated Guide Wire (Arrow and Other Brands)

Perform steps 1–6 from the "Radial Artery Cannulation: Standard Over-the-Needle Catheter Assembly," above, and then proceed with the steps below:

1. Remove protective cap from needle-cannula assembly and ensure proper function by sliding actuation lever along extension tubing to advance and retract guide wire through needle.
2. Retract guide wire as far back as possible using actuation lever to maximize visibility arterial blood flow ("flashback") within introducer hub.

Perform steps 8–10 from the "Radial Artery Cannulation: Standard Over-the-Needle Catheter Assembly," above, and then proceed with the steps below (Fig. 1.5):

1. Hold needle stationery and slowly slide actuating lever forward to feed guide wire as far as possible into artery.
   - If resistance is met while feeding the guide wire, discontinue sliding actuating lever and withdraw entire unit from artery to prevent damage to guide wire or vessel wall
2. Advance entire assembly 1–2 mm further into vessel to ensure catheter tip (which sits approximately 2 mm behind the needle tip) has entered the lumen.
3. Stabilize clear introducer hub in position and advance catheter forward into artery over guide wire until hub of catheter is in contact with the skin.
   - If difficulty is encountered at this step, catheter hub may be rotated slightly to facilitate advancement.
4. Stabilize catheter in position and withdraw introducer needle, guide wire, and feed tube as a single unit; blood flow from catheter hub at this point indicates successful cannulation of the artery.

Perform steps 13–18 from the "Radial Artery Cannulation: Standard Over-the-Needle Catheter Assembly," above.

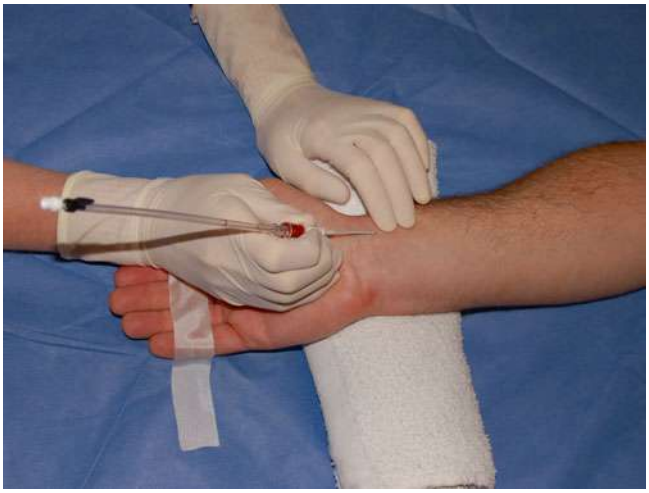

Fig. 1.5 Puncture of radial artery using over-the-needle catheter assembly with integrated guide wire

1.3.3 Procedure: Radial Artery Cannulation—The Allen Test

1. Occlude both radial and ulnar arteries of one extremity with digital pressure at wrist.
2. Instruct patient to repeatedly clench the fist tightly to exsanguinate the hand while occlusion of the arteries is maintained.
3. Without releasing digital pressure on arteries, instruct patient to extend fingers and observe palmar surface to confirm blanching of skin.
4. Release pressure on ulnar artery only and observe palmar surface for reperfusion (Fig. 1.6).
   - If reperfusion of the hand does not occur within 5–10 s, ulnar arterial blood flow may be compromised and radial artery cannulation should not be attempted. If reperfusion is brisk, repeat the test releasing pressure on radial artery only and observing palmar surface for reperfusion. If the return of rubor takes longer than 5–10 s, radial artery puncture should not be performed.

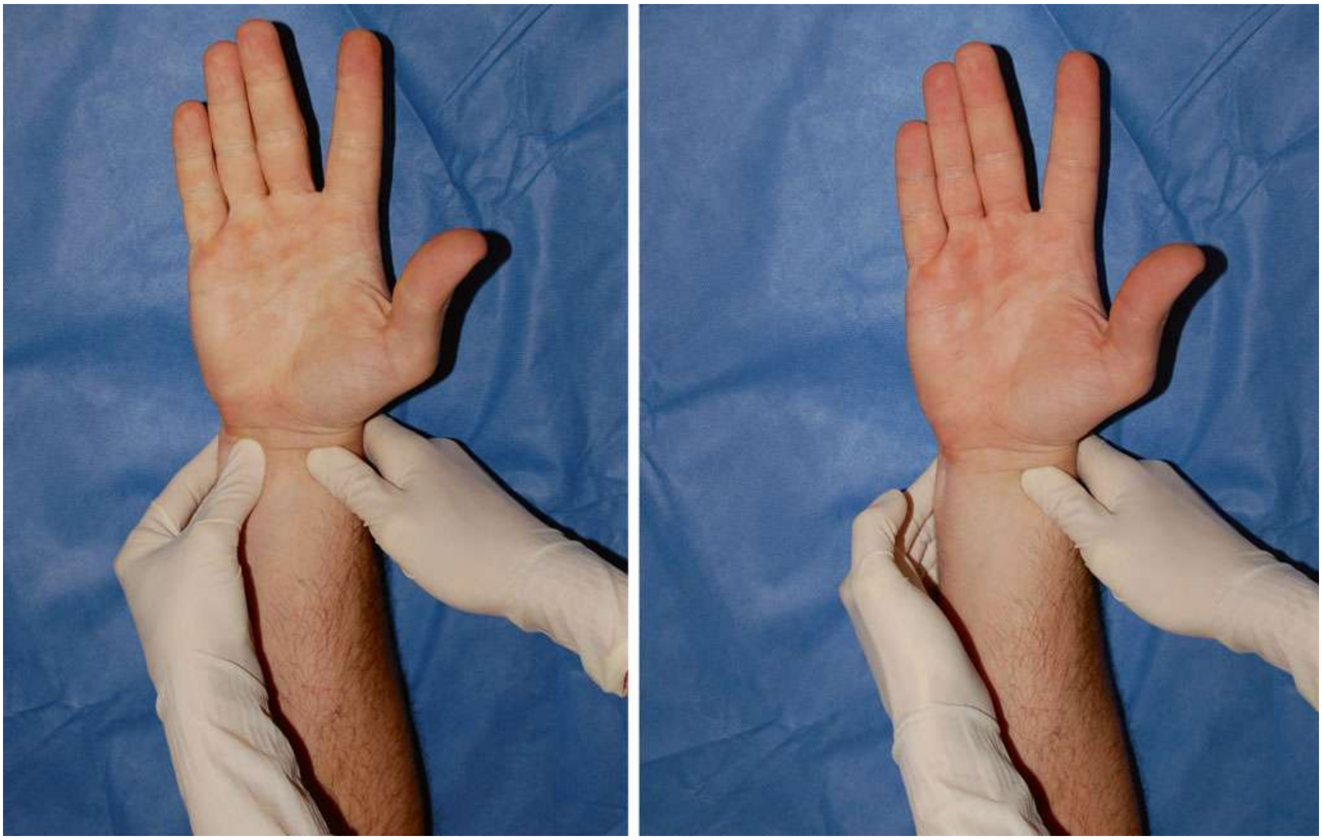

Fig. 1.6 The Allen Test

1.3.4 Procedure: Femoral Artery Cannulation—The Seldinger Technique

1. Place the patient in the supine position with the inguinal region adequately exposed.
2. Palpate the femoral pulse, located at midpoint between pubic symphysis and anterior superior iliac spine, using the second and third fingers of the gloved nondominant hand.

Perform steps 4 and 5 from the "Radial Artery Cannulation: Standard Over-the-Needle Catheter Assembly," above, and then proceed with the steps below (Fig. 1.7):

1. Attach 5 mL syringe to an introducing needle of bore sufficient to accommodate guide wire.
2. Hold syringe connected to introducing needle like a pen with needle bevel facing upward.
3. Directing the needle at a 45° angle to the skin in a cephalic direction, puncture the skin through the anesthetic wheal immediately overlying the palpated artery just distal to the inguinal ligament and advance needle slowly toward palpated artery until tip enters arterial lumen, which is confirmed by visible arterial blood flow ("flashback") into the needle hub and syringe.
    - Avoid self-puncture by maintaining adequate distance between needle tip and index finger.
    - Care must be taken to avoid trauma to the femoral nerve and vein bordering the femoral artery.
4. Hold needle stationery and remove syringe, taking care not to displace the intraluminal position of the needle tip.
    - Advancing the needle too far once initial flashback is visualized may result in piercing the back side (or "double puncture") of the artery wall, in which case visible blood flow will cease; if this occurs, slowly withdraw needle several millimeters until pulsatile blood flow reappears.
5. Occlude needle hub temporarily with gloved finger to prevent unnecessary blood loss and air embolism.
6. Thread blunt end of flexible guide wire smoothly into needle and gently into artery until at least one-quarter of guide wire is intravascular (Fig. 1.8).
    - If resistance is met while threading guide wire, remove wire from needle, reattach syringe, and aspirate blood to confirm continued intraluminal needle tip placement; if resistance is met while *removing* guide wire from needle, remove guide wire and needle from artery *as a single unit* to prevent shearing the guide wire off inside the vessel.
7. Holding the wire securely in place, remove the introducing needle.
8. Using a scalpel, make a small incision (approximately the width of the catheter to be used) through the dermis at the insertion site of the guide wire.
    - Avoid severing the guide wire by facing the sharp edge of the scalpel *away* from the guide wire.
9. While stabilizing the guide wire at its insertion site, thread the dilator over the free end of the guide wire until it is approximately one inch from the skin.
10. Grasp the free end of the guide wire protruding from the tail end of the dilator.
    - If it does not protrude from the tail end of the dilator, the guide wire must be removed sufficiently from the artery to be securely grasped; it must protrude visibly from the tail end of the dilator throughout the subsequent process of threading the dilator into the artery.
11. Holding the dilator firmly near its tip, thread the dilator over the wire into the skin with a back-and-forth twisting motion until it reaches the artery.
    - Only the skin tract should be dilated; dilation of the artery may result in excessive arterial injury and/or hemorrhage.
12. Holding the wire securely in place, remove the dilator.
13. While stabilizing the guide wire at its insertion site, thread the catheter over the free end of the guide wire until it nears the skin.
14. Grasping the guide wire where it protrudes from the tail end of the catheter, thread the catheter into the skin to its appropriate insertion length.
15. While stabilizing the catheter at its insertion site, slowly remove the guide wire.
    - If resistance is met while removing guide wire, remove guide wire and catheter from artery as a single unit to prevent shearing the guide wire off inside the vessel.
16. Secure the catheter to the skin using silk (2.0) or nylon (4.0) sutures. Take a 0.5 cm bite of skin with the suture needle. If the catheter assembly contains integrated "wings" for fixation, thread suture through the perforated wings and secure catheter against the skin with several knots. If no fixation device is included, tie several knots in the suture without pinching the skin, leaving both ends of the suture long. Using the loose ends of the suture, tie a second set of knots around the hub of the catheter, firmly, but without constricting its lumen.

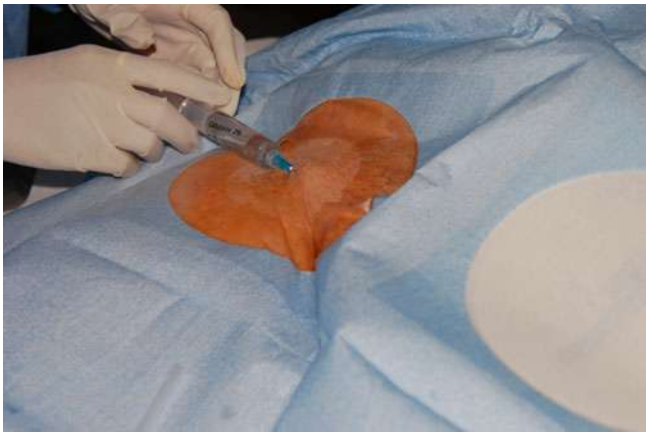

Fig. 1.7 Anesthetic injection over the femoral artery

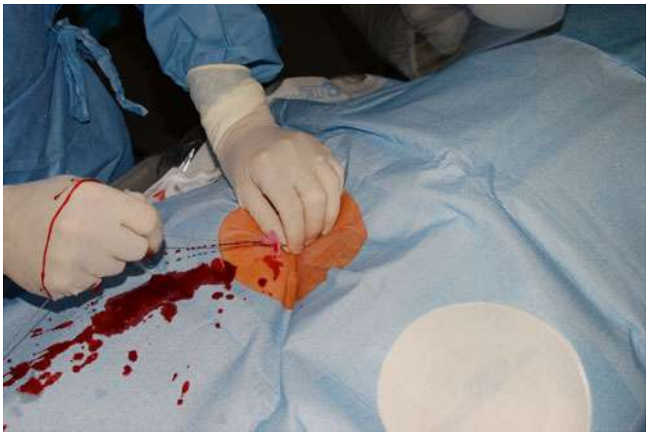

Fig. 1.8 Insertion of guide wire into femoral artery

1.4 Complications

- Hemorrhage
- Hematoma (at puncture site)
- Infection (at insertion site or systemic)
- Thrombosis
- Arteriovenous fistula
- Pseudoaneurysm formation
- Exsanguination (secondary to dislodgement of catheter)
- Cerebrovascular accident (CVA; secondary to air embolism)

1.5 Pearls and Pitfalls

- The shorter and stiffer the plastic tubing connected to the arterial cannula for blood pressure monitoring, the higher its frequency response and the accuracy of measurements.
- Use of an ultrasound probe can facilitate artery location and vessel cannulation.
- Puncture of the femoral artery proximal to the inguinal ligament, or distal to its bifurcation into superficial femoral and deep femoral arteries, may cause massive hemorrhage due to poor vessel compressibility in these regions; the artery should therefore be cannulated just distal to the inguinal ligament, where it is easily compressible against the femoral head if necessary.
- If difficulty is encountered when advancing an over-the-needle catheter into the artery, attach a 10 mL syringe containing 5 mL of sterile normal saline to the catheter hub, aspirate 1 or 2 mL of blood to confirm catheter tip placement within the vessel lumen, and then advance the catheter while gently injecting the saline-blood mixture; the jet of fluid momentarily dilates the lumen, aiding advancement of the catheter.
- An alternative approach to the over-the-needle catheter that will not fully advance is the use of a guide wire. After intraluminal placement of the cannula tip is confirmed by blood return, a guide wire is gently inserted through the catheter into the artery. The cannula is then passed along the guide wire until fully advanced. The guide wire employed must have a blunt, flexible tip to minimize the possibility of vessel wall trauma.
- Two potential consequences of arterial cannulation are vessel obstruction secondary to intravascular thrombosis and hemorrhage (the latter being the most common complication). Choice of puncture site is therefore essential. Due in part to their generous collateral blood flow, as well as ease of compressibility, the radial and femoral are the two most commonly cannulated arteries.
- Repeated puncture following unsuccessful cannulation increases the risk of arterial obstruction secondary to vessel wall damage and thrombosis.

- Double puncture of the cannulated artery by inadvertent over-insertion of the needle has not been shown to increase complications despite the additional trauma to the vessel walls.

Selected Reading

Anderson JS. Arterial cannulation: how to do it. Br J Hosp Med. 1997;57:497.

Gilchrist IC. Reducing collateral damage of the radial artery from catheterization. Catheter Cardiovasc Interv. 2010;76:677–8.

Lemaster CH, Agrawal AT, Hou P, Schuur JD. Systematic review of emergency department central venous and arterial catheter infection. Int J Emerg Med. 2010;3:409–23.

Mitchell JD, Welsby IJ. Techniques of arterial access. Surgery. 2004;22:3–4.

Wilson SR, Grunstein I, Hirvela ER, Price DD. Ultrasound-guided radial artery catheterization and the modified Allen's test. J Emerg Med. 2010;38:354–8.

Ultrasound-Guided Peripheral Intravenous Access

2

Coben Thorn and L. Connor Nickels

2.1 Indications

- Difficulty placing peripheral intravenous (PIV) using traditional methods of direct visualization and palpation
- To reduce needle sticks in a hypercoagulable patient

2.2 Contraindications

- Patient needs emergent central venous access

2.3 Materials and Medications

- Ultrasound machine
- High-frequency (5–8 MHZ) linear probe (Fig. 2.1)
- Ultrasound gel
- Minimum 1.5-in. needle length
- IV setup
- Skilled operator

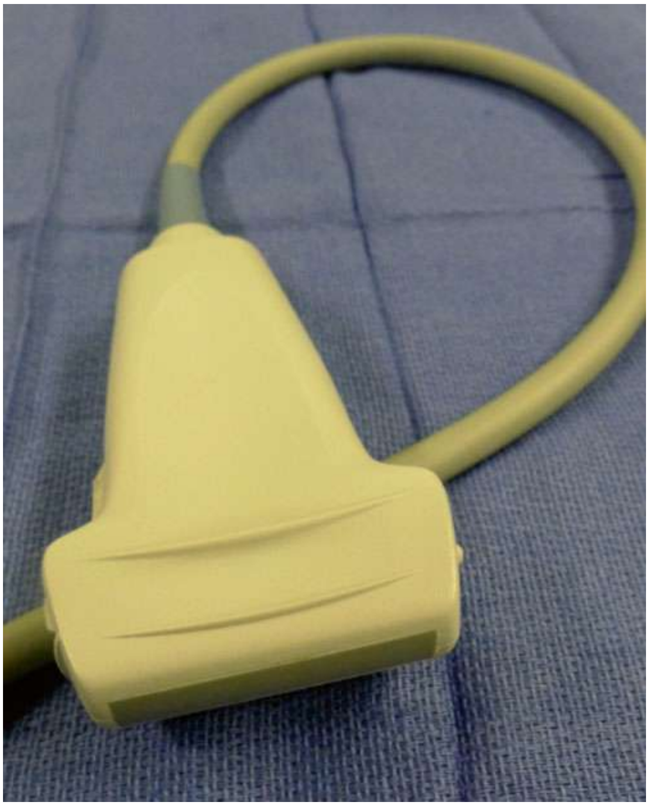

Fig. 2.1 High-frequency (5–8 MHZ) linear probe

C. Thorn, MD
Department of Emergency Medicine, Bon Secours St. Francis Health System, Greenville, SC, USA
e-mail: cobenthorn@gmail.com

L.C. Nickels, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_2

2.4 Procedure

1. Scan the selected area to identify a target vessel for PIV cannulation.
    - Basilic (runs on the medial side of upper arm) (Fig. 2.2) and cephalic (runs on the lateral side of the upper arm) veins are good superficial veins that are generally not seen without ultrasound. The deep brachial vein is also an option; however, this is a deep vein that runs with the brachial artery, and there is higher chance for complications (Fig. 2.3).
2. Now the site should be prepared for IV insertion. It should be cleaned. An appropriate gauge long needle should be selected, and IV setup should be conveniently located for when access is obtained.
3. The ultrasound probe should be placed in the transverse plane (Fig. 2.4) to best visualize surrounding structures and vein. Alternatively, the probe can be placed longitudinally (Fig. 2.5) for better visualization of needle depth and slope.
    - Most practitioners seem to prefer the transverse approach.
4. The concept of the Pythagorean Theorem is used for accuracy. The needle should be inserted at a 45° angle to the skin and at an equal distance back from the probe as the approximate depth of the vessel vertically. The depth is given on the screen, usually at the bottom in centimeters. As soon as the needle has penetrated the skin the needle tip should be located by fanning the probe toward the needle until it is identified. The needle should then be advanced slowly always keeping the needle tip in view. Once directly on top of the vein, it should tent with pressure and then the needle should be inserted into the vein. Follow the needle in the vein as far as possible while keeping the tip of the needle in the center of the vein to make sure the catheter is securely in the vein and does not infiltrate.

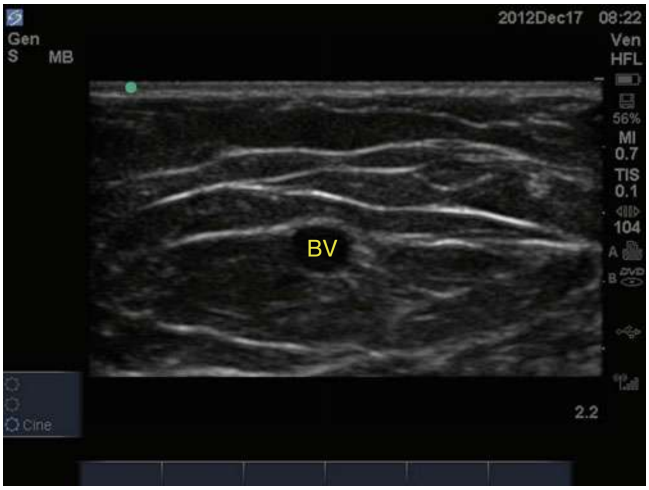

Fig. 2.2 Basilic vein (BV) located medially when scanning proximally from the antecubital fossa

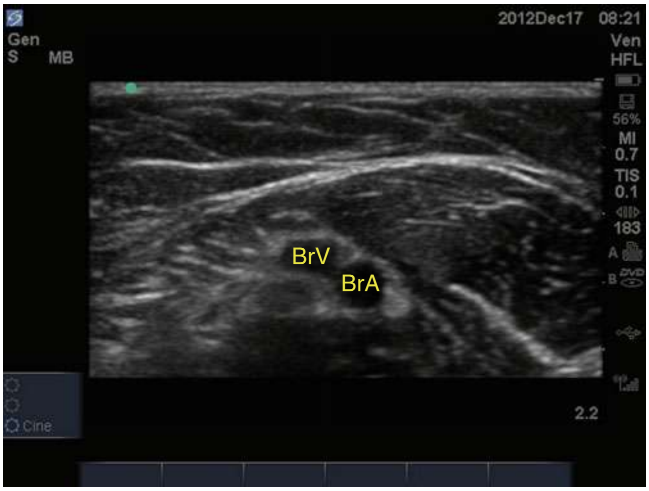

Fig. 2.3 Brachial artery (BrA) and vein (BrV). The less round, slightly compressed, anechoic structure on the left is the vein. The very circular, not compressed, anechoic structure to the right is the artery

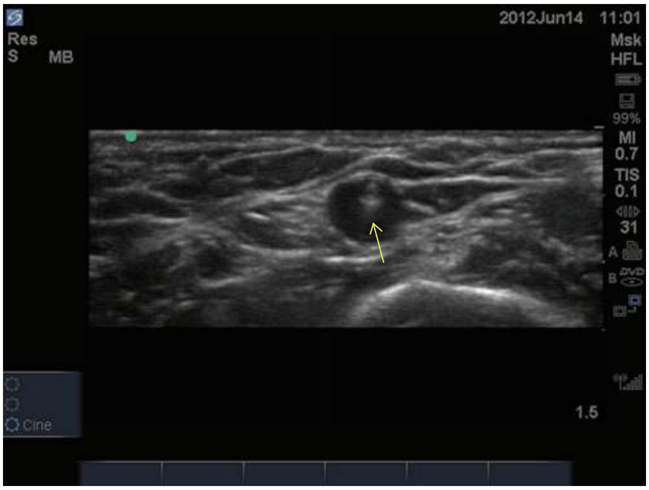

Fig. 2.4 Needle tip in vein seen in a transverse orientation

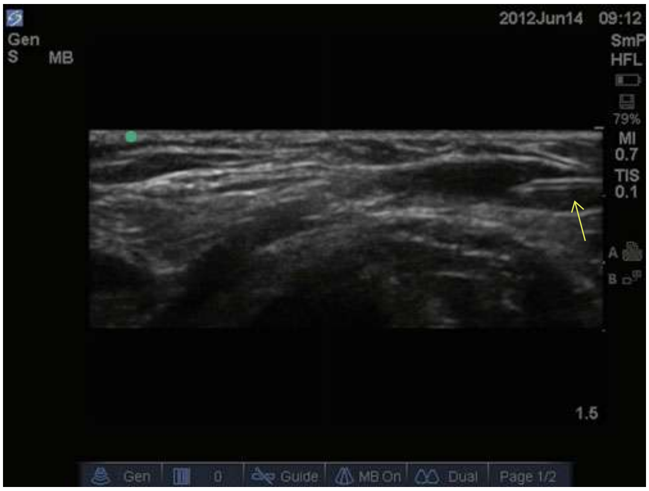

Fig. 2.5 Needle tip in vein seen in a longitudinal orientation

2.5 Complications

Inadvertent puncture of artery. Veins should be thin walled, compressible, and have no pulsations.

2.6 Pearls and Pitfalls

- Mistaking the midshaft of the needle for the needle tip. If this occurs, the needle tip is actually deeper than expected. The ultrasound machine will plot a hyperechoic "dot" on the screen for the needle tip, as long as it crosses the ultrasound beam at any point. This same "dot" will appear whether the tip is directly centered under the beam or *any* segment of the needle shaft is intersecting the beam. This can be visually deceiving and makes this procedure difficult to grasp.
- Very slow movements of the needle and the probe are important for keeping the needle tip in view. Once the needle tip is identified, the probe should be fanned forward (away from operator) just slightly and then the needle advanced until the needle tip comes into view again. This is repeated until the needle is securely moved further into the vein.

Selected Reading

Brannam L, Blaivas M, Lyon M, et al. Emergency nurses' utilization of ultrasound guidance for placement of peripheral intravenous lines in difficult-access patients. Acad Emerg Med. 2004;11:1361–3.
Constantino TG, Parikh AK, Satz WA, et al. Ultrasonography-guided peripheral intravenous access versus traditional approaches in patients with difficult intravenous access. Ann Emerg Med. 2005;46(5):456–61.
Dargin JM, Rebholz CM, Lowenstein RA, et al. Ultrasonography-guided peripheral intravenous catheter survival in ED patients with difficult access. Am J Emerg Med. 2010;28(1):1–7.
Ma JO, Mateer JR, Blaivas M. Emerg ultrasound. 2nd ed. New York: McGraw-Hill Companies, Inc.; 2008.
Saul T, Rivera M, Lewiss R. Ultrasound image quality. ACEP News. 2011;4:24–5.

Central Venous Line Placement: Internal Jugular Vein, Subclavian Vein, and Femoral Vein

3

Kevin D. Ergle, Zachary B. Kramer, Jason Jones, and Rohit Pravin Patel

3.1 Indications

- Volume replacement
- Emergent venous access
- Administration of caustic medications: vasopressors, calcium chloride, hypertonic saline, high dose of potassium
- Dialysis catheter placement (hemodialysis)
- Nutritional support (total parenteral nutrition)
- Long-term antibiotics
- Chemotherapy
- Plasmapheresis
- Frequent or persistent blood draws or intravenous therapy when unable to establish peripheral access due to edema or other causes
- Jugular and subclavian: Central venous pressure monitoring, transvenous pacing wire introduction, pulmonary artery catheterization

3.2 Contraindications

- Absolute
  - Infection at site of insertion
  - Distorted anatomy/landmarks (prior surgery, radiation, or history of thrombus in the specified vein)
  - Subclavian only: Trauma to the ipsilateral clavicle, anterior proximal rib, subclavian or superior vena cava vessels
- Relative
  - Morbid obesity
  - COPD
  - Children less than 2 years (higher complication rates)
    - Coagulopathy (although ultrasound-guided internal jugular can be done in this situation)
  - Agitated or moving patient
  - Jugular only: Trauma to the ipsilateral clavicle, anterior proximal rib, subclavian or superior vena cava vessels
  - Jugular and subclavian: Inability to tolerate potential pneumothorax of the ipsilateral thoracic cage
    - Pneumothorax or hemothorax of the contralateral thorax
    - Patients receiving ventilatory support with high-end expiratory pressures (temporarily reduce the pressures)
  - Femoral only: Intra-abdominal (or retroperitoneal) hemorrhage

3.3 Materials and Medications

- Central venous catheter tray or bundle: single/double/triple/quadruple lumen, dialysis catheter, large-bore introducer (for transvenous pacing or pulmonary artery catheter kit)
- Sterile gloves
- Sterile drapes or towels
- Sterile gown
- Hat/hair cap and mask with eye protection
- Antiseptic solution with skin swabs (e.g., chlorhexidine)
- Sterile saline flushes (one 30 mL syringe or three 10 mL syringes)
- Lidocaine 1 %

K.D. Ergle, MD
Department of Internal Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: kevin.ergle@medicine.ufl.edu

Z.B. Kramer, MD
Emergency Medicine Resident, University of Florida College of Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: zkramer@ufl.edu

J. Jones, MD • R.P. Patel, MD (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: jasonjones@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_3

- Sterile gauze
- No. 11 blade scalpel
- Dressing (sterile waterproof transparent dressing or sterile 4×4 gauze with tape)
- Sterile biopatch
- Suture material with needle driver if needed
- Transducing line (optional)
- Sterile probe cover (if using ultrasound guidance)

3.4 Procedures

3.4.1 Internal Jugular Vein Access Procedure

1. Obtain informed consent if not emergently indicated procedure.
2. Obtain supplies and prepare the room, ensuring that all supplies are within operator reach prior to placing gown and commencing the procedure. Include a sterile ultrasound sheath on the sterile field if ultrasound is being used.
3. Raise bed to a comfortable height for the operator.
4. Place patient with head facing away from side of central line site (if using ultrasound, other positions may be preferred). Place patient in 15–20° Trendelenburg position to help fill the upper central veins and reduce the risk of air embolism.
5. Identify the anatomy. Palpate triangle made by the clavicle and sternal and clavicular heads of the sternocleidomastoid (SCM) muscle to identify the location of the internal jugular vein (Fig. 3.1). If using ultrasound guidance, identify optimal anatomical arrangement.
6. Wash your hands and wear sterile attire using aseptic technique, including cap, mask, gown, and sterile gloves.
7. Prepare the site from the clavicle to the ear and across the trachea with antiseptic solution. Allow the antiseptic (chlorhexidine or iodine) to fully dry.
8. Drape the site and patient with sterile towels and drapes included in most CVL bundles. Make sure to cover the whole area and bed.
9. Cover the ultrasound probe with a sterile sheath. This can be done solo or by holding the sterile ultrasound sheath and having an unsterile assistant hold the probe so that the probe can be covered by the sheath.
10. Prepare the kit by checking the guide wire and flushing the tubing and lines with saline included in the kit.
11. With a 25-gauge needle, use 1 % lidocaine to anesthetize the skin at the apex of the triangle made by the SCM and clavicle. Aspirate to make sure the operator is not in a vessel and make a superficial wheel for the insertion site.
12. Preferred method is with ultrasound guidance (see steps 13–17 and Sect. 3.6 for description of ultrasound guidance). If performing without ultrasound, palpate the carotid artery and insert needle lateral to the artery at the apex of the triangle formed by the SCM, aiming toward the ipsilateral nipple at an angle 30–45° above the horizontal plane (Fig. 3.1). Once blood returned, go to step 18.
13. Place sterile ultrasound gel over the insertion site. Use the ultrasound to identify vessel anatomy including internal jugular vein and carotid artery. Use the ultrasound probe to compress the vein, which is compressible as opposed to the carotid artery, which is not compressible (Fig. 3.2).
14. Prepare the insertion needle and syringe (if long and short needles are available, a short needle may be used to reduce posterior vein perforation) and prime the syringe by pulling back on the plunger prior to making the puncture.
15. Use the ultrasound probe to re-identify the patient's anatomy.
16. Ultrasound can be used in short axis or long axis (Fig. 3.3). Short axis is easier for novice operators due to increased ability to see the artery and vein but has a higher risk of posterior perforation if the needle tip is not visualized well. Once short axis of the vein is found, turning the probe 90° clockwise allows the operator to see the vein in long axis. The needle is better visualized in this view but technically more difficult and has less chance to penetrate the posterior wall. In patients with short necks, it may be difficult to obtain long-axis view and needle insertion in the limited space.
17. Insert the needle using the ultrasound guidance with dynamic approach preferred (see Sect. 3.6 for specifics). Make sure to aspirate while inserting the needle to identify when the venous access is obtained. The needle tip should be visualized through the whole process.
    - If using the static approach (see Sect. 3.6), insert needle lateral to carotid pulsation as this is where the vein anatomically is located. Standard method is to insert the needle as far back as the depth the vessel is visualized (e.g., if the vein is visualized 2 cm below skin surface, the needle should be inserted 2 cm behind the probe at a 45° angle).
    - If inserting the needle ~3 cm does not achieve access, gently withdraw the needle toward the surface of the skin while aspirating. Avoid withdrawing the needle completely from the skin. If needed, redirect the needle and advance until blood is aspirated. Cannulation of the vein often takes place while withdrawing the needle.
18. Hold the needle steady with your nondominant hand and remove the syringe, careful not to advance or withdraw the needle. You can place the base of your hand on the patient's chest to make your hand more stable during

this part of the procedure. Occlude the hub of the needle to prevent air embolus.

19. You may verify that you are in the vein by transducing pressure with a fluid column. The fluid should flow easily into the vein.
    - If the aspirated blood is pulsatile and moves up the column, withdraw the needle completely and apply pressure for 10–20 min while taking the patient out of Trendelenburg position (if nonemergent procedure).
20. Once it is verified that you are in the vein, insert the J-tip of the guide wire into the needle hub and advance into the vein. The J-tip can be straightened with a pinching motion (Fig. 3.4). Always keep one hand on the guide wire until it is removed from the patient. Monitor for arrhythmias as the guide wire is advanced toward the right atrium.
    - If the guide wire does not flow easily, remove the guide wire and reattach the syringe, checking for blood flow.
    - If arrhythmia occurs, slowly withdraw the guide wire until the patient's native rhythm returns.
    - Alternatively, the catheter/syringe found in most kits can be used as a bridge to guide wire placement. For the author, this has improved success when there is difficulty in wire placement. Use the same steps above with the catheter (Fig. 3.5) and when you have return of blood, advance the angiocath into the vein followed by insertion of the guide wire through the angiocath. This is especially useful in moving/agitated patients, patients who have collapsible veins due to hypovolemia, and patients who have abnormal anatomy and may have veins that take an abnormal angle shortly past the needle tip.
21. Remove the needle over the guide wire, making sure to always keep control of the guide wire.
22. Make an incision contiguous with the guide wire using a straight (No. 11) blade with the scalpel blade facing upward (away from the wire).
23. Advance the dilator over the guide wire in a twisting motion, keeping control of the guide wire.
    - The dilator only needs to go slightly beyond the anticipated depth of the patient's jugular vein. Do not advance the entire length of the dilator.
24. Withdraw the dilator and hold pressure over the wound site.
25. Advance the catheter over the guide wire while keeping control of the guide wire.
26. With the catheter inserted 10–12 in. from the skin insertion site, retract the guide wire until it comes out of the distal port. Maintain control of the guide wire and advance the catheter to the appropriate length. Usually catheters are inserted 15–16 cm from the right side and 18–20 cm from the left side (Fig. 3.6).
27. Flush each port of the catheter and check aspiration. If difficulty with aspiration or flushing, concern is raised for catheter malposition. The operator can change the depth slightly or twist the catheter and recheck.
28. At this time, an antibiotic ointment or biopatch may be applied to skin around the intersection with the lumen of the catheter. This step is based on local institutional guidelines.
29. Suture the line in place.
30. Enclose CVL site with sterile waterproof transparent dressing.
31. Confirm placement using chest X-ray (CXR). The tip of the catheter should be in the lower third of the superior vena cava (SVC) at the insertion of the SVC into the right atrium.

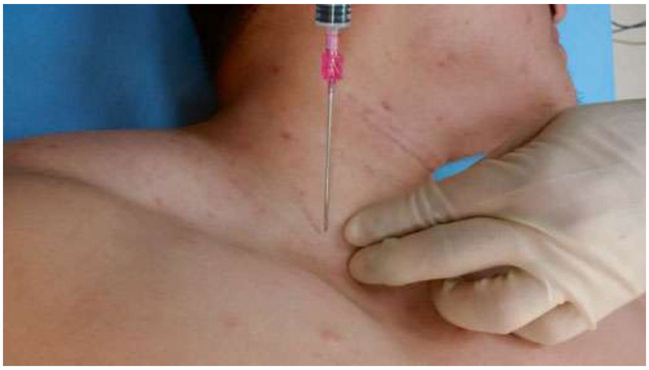

Fig. 3.1 Internal jugular blind approach; this would be the same location for probe placement if doing ultrasound guided

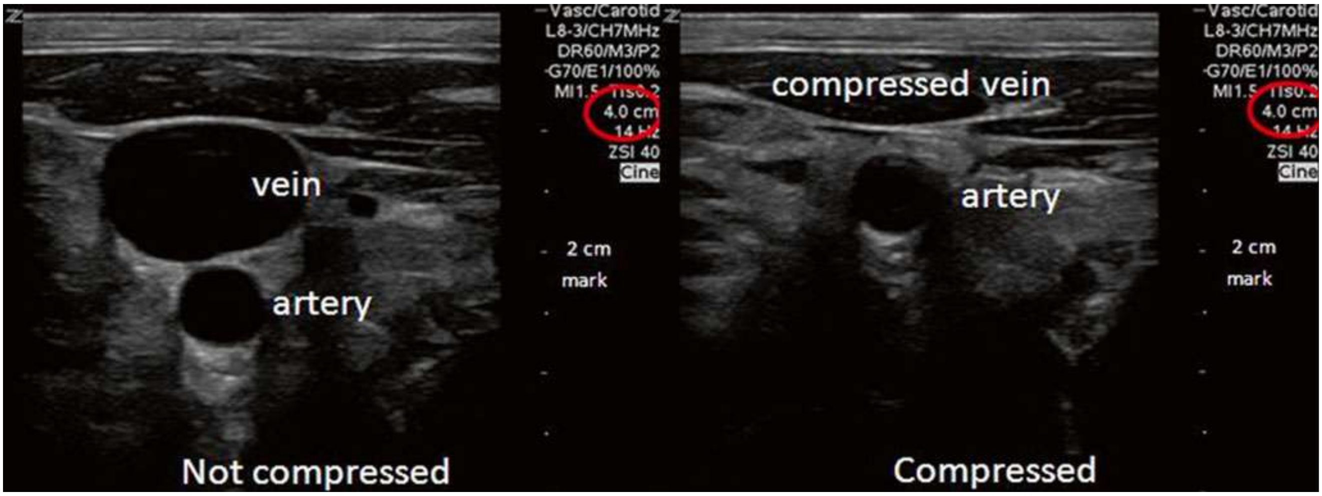
Vasc/Carotid
L8-3/CH7MHz
DR60/M3/P2
G70/E1/100%
MI1.5 TIs0.2
4.0 cm
14 Hz
ZSI 40
Cine
2 cm
mark
vein
artery
Not compressed
compressed vein
artery
Compressed

Fig. 3.2 Ultrasound showing internal jugular vein and artery with and without compression

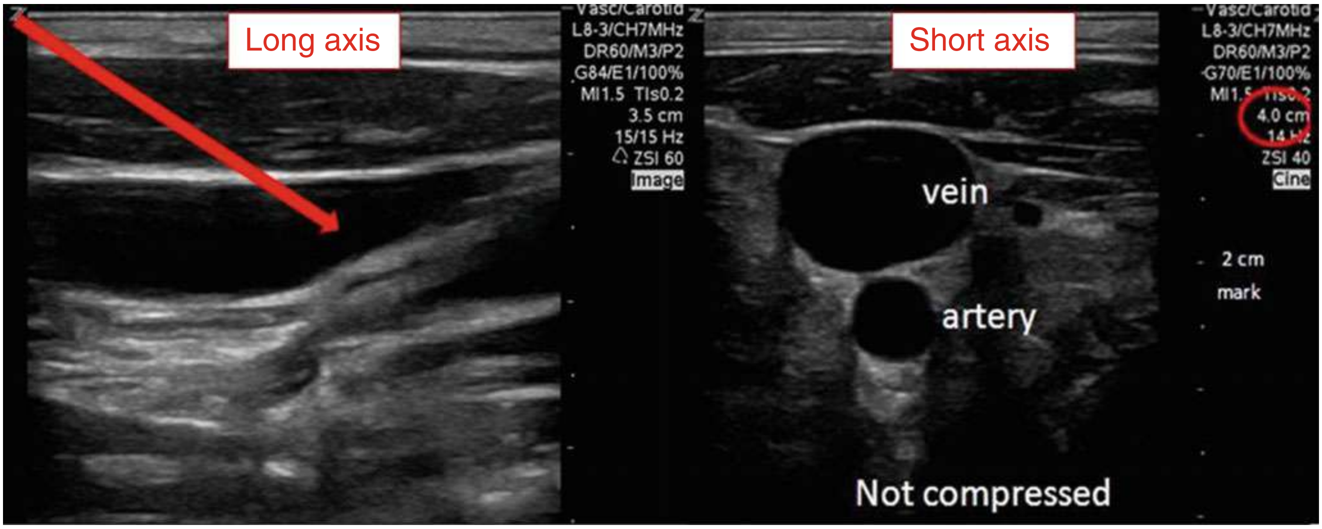
Long axis
Vasc/Carotid
L8-3/CH7MHz
DR60/M3/P2
G84/E1/100%
MI1.5 TIs0.2
3.5 cm
15/15 Hz
ZSI 60
Image
Short axis
Vasc/Carotid
L8-3/CH7MHz
DR60/M3/P2
G70/E1/100%
MI1.5 TIs0.2
4.0 cm
14 Hz
ZSI 40
Cine
2 cm
mark
vein
artery
Not compressed

Fig. 3.3 Long- and short-axis views of the internal jugular vein

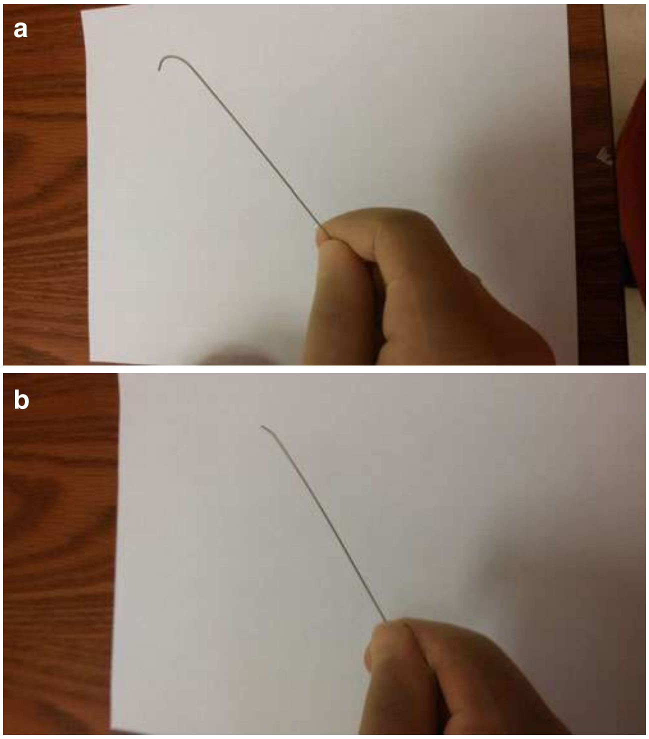

Fig. 3.4 (**a**, **b**) J-tip straightening using pinch/stretch method

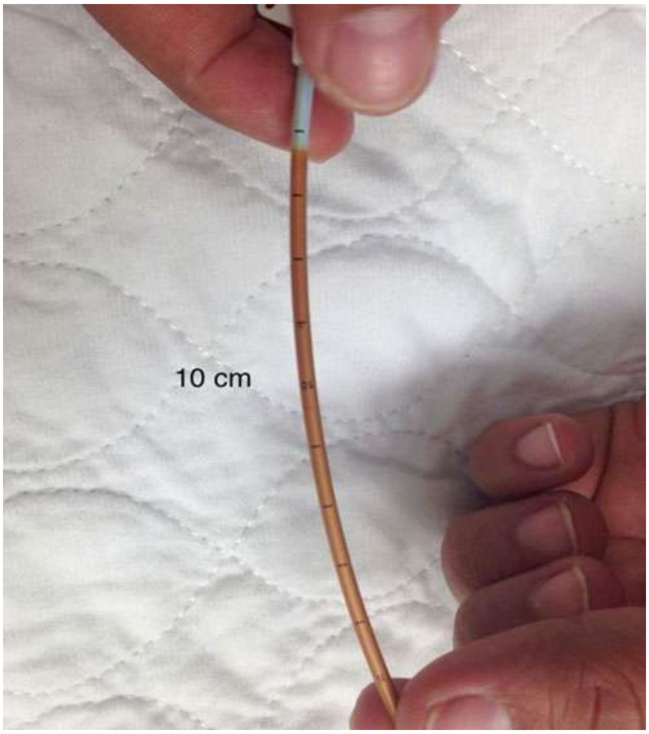

Fig. 3.6 Length: marking seen on typical central venous catheters; number indicates distance in centimeters from distal tip

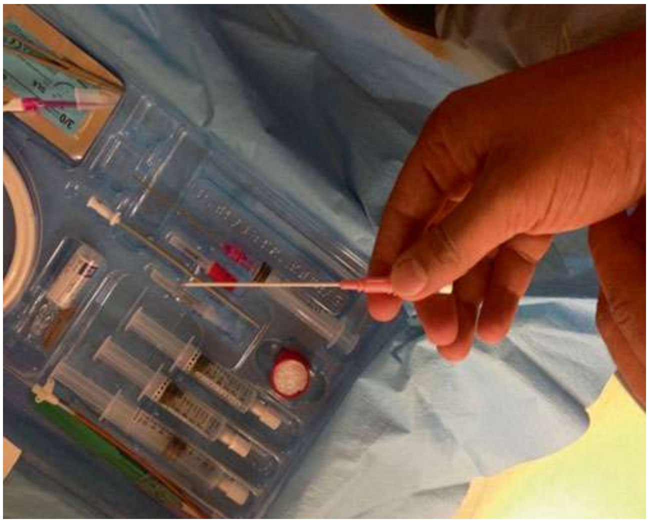

Fig. 3.5 Angiocath that can be used in difficult to cannulate/wire patients

3.4.2 Subclavian Vein Access Procedure

1. Obtain informed consent if not emergently indicated.
2. Raise bed to comfortable height for the operator.
3. Place patient in supine position and position so patient's head is at the top of the bed.
4. Place patient in 15–20° Trendelenburg position (if tolerated) to reduce risk of air embolism. Studies show this will also increase the size of the subclavian vein. Do NOT place towel between shoulder blades (arch shoulder back) as this has been shown to decrease vein diameter and affect reliability of accessibility. Keep shoulders at anatomical location (forward).
5. Prep area chosen from the anterior neck, clavicle, and upper chest (above nipple line) with chlorhexidine prep or iodine.
6. Open kit and place close to operator's dominant hand to allow for easy access. Diameter of catheter/kit used based on clinical situation:
   - Introducer or large bore if requiring large volumes of resuscitation
   - Triple lumen catheters for vasopressors
   - Introducers if anticipating pulmonary artery catheter or venous pacer
7. Operator should prepare with all aseptic techniques (e.g., hand washing) and maximal barrier precautions (e.g., sterile gowns, sterile gloves, caps, masks covering both mouth and nose, eye protection, and full-body patient drapes).
8. Once sterile and able to touch the inside of the CVL kit, the operator may want to retract the curved J-tip wire into the plastic loop sheath for easy directing into the introducer needle. The operator should also uncap all distal lumens and flush all ports with 3–5 cc of the sterile NS syringes to ensure no defects in the lumen of the catheter. Close all ports except the distal tip port (usually marked with words "distal tip") with the slide clamp.
9. Prep area chosen (right or left side) from the anterior neck, clavicle, and upper chest (above nipple line) with sterile chlorhexidine prep (this is the second cleaning).
10. Place full-body drape over patient with opening over selected side where needle will be inserted.
11. Needle insertion site options (Fig. 3.7):
    - One centimeter inferior to the junction of the middle and medial third of the clavicle
    - Just lateral to the midclavicular line, with the needle perpendicular along the inferior lateral clavicle
    - One fingerbreadth lateral to the angle of the clavicle
12. Anesthetize needle insertion site with 5–10 mL of 1 % lidocaine superficially (make sure to pull back on needle syringe to ensure operator is not in the vein or artery).
    - Never place equipment on a patient.
13. Prepare the needle and syringe by placing the long needle on the syringe. Make sure to break seal of syringe by pulling back on the plunger of the syringe prior to making incision with needle.
14. Turn patient's head to opposite of CVL placement and retract ipsilateral shoulder down to improve clavicle-vein relationship. The retraction of the arm can be done a few steps earlier and can be held in position using a person or tape/restraints.
15. Direct the insertion needle toward sternal notch in the coronal plane at an angle no greater than 10–15° while gently withdrawing the plunger of the syringe. Keep bevel of the needle facing up and in line with the numbers on the syringe until operator enters skin, then face bevel caudally to facilitate smooth progression of the guide wire down the vein toward the right atrium.
16. It helps to place nondominant hand (not holding the needle) on the sternal notch so operator can feel where sternal notch is and direct needle in that direction (Fig. 3.7).
17. NEVER increase the angle of the needle greater than 15° as pneumothorax may ensue.
18. Advance the needle under and along the inferior border of the clavicle, making sure the needle is virtually horizontal to the chest wall. Aim medially in the direction of the suprasternal notch, attempting to first aim for the clavicle then "walk" the needle below the clavicle.
19. Once under the clavicle, continue to advance the needle in a plane almost parallel to the skin approximately 2–3 cm until venous blood is freely aspirated into the syringe.
20. When venous blood is freely aspirated, disconnect the syringe from the needle, and immediately occlude the lumen to prevent air embolism and insert the guide wire. If the vein is difficult to locate, remove the introducer needle, flush it clean of clots, and try again. Change insertion sites after three unsuccessful passes with the introducer needle.
21. At this point, the hand holding the needle should be "set in stone." Use the patient's chest wall as a base to keep needle completely still as to not inadvertently advance or retract needle out of the vein.
22. Insert the guide wire through the needle into the vein with the J-tip directed caudally to improve successful placement into the subclavian vein.
    - Beware a return of red pulsatile blood. If this occurs, the wire is in an artery.
    - Beware aspirating air bubbles through the probing introducer needle. This indicates a pneumothorax.
23. Advance the wire until it is mostly in the vein or until arrythmia is seen on the cardiac monitor. Then, retract the wire 3–4 cm.
24. If the wire does not pass easily, remove the wire, reattach the syringe, and confirm that the needle is still in the

lumen of the vein before reattempting. The J-tip can be straightened with a pinching motion (Fig. 3.4).

- Alternatively, the catheter/syringe found in most kits can be used as a bridge to guide wire placement. For the author, this has improved success when there is difficulty in wire placement. Use the same steps above with the catheter (Fig. 3.5) and when you have return of blood, advance the angiocath into the vein followed by insertion of the guide wire through the angiocath. This is especially useful in moving/agitated patients, patients who have collapsible veins due to hypovolemia, and patients who have abnormal anatomy and may have veins that take an abnormal angle shortly past the needle tip.

25. Use the tip of the scalpel to make a small incision just against the needle to enlarge the catheter entry site for the dilator and catheter.
26. Holding the wire in place, withdraw the introducer needle and place in needle holder.
27. Thread the dilator over the wire and into the vein with a firm and gentle twisting motion while maintaining constant control of the wire. If a large-bore introducer is placed, the dilator/introducer goes in one step, after the introducer is inserted, hold the wire in place and remove the dilator.
28. If operator is having difficulty threading the dilator, the skin incision with the scalpel may have been too superficial or small. It may help to enlarge this incision to avoid having the dilator get caught on superficial skin or connective tissue.
29. It is helpful to have sterile gauze handy to apply pressure with the hand not holding the wire as the vein will now bleed profusely from around the wire secondary to dilation.
30. Thread the catheter until it is close to the skin insertion site. Then pull back on the guide wire until it shows outside of the distal port. Grasp the wire outside of the distal port and thread the catheter while holding onto the guide wire. Usually catheters are inserted 15–16 cm from the right side and 18–20 cm from the left side (Fig. 3.6)
31. Hold the catheter in place and remove the wire. After the wire is removed, occlude the open lumen.
32. Attach sterile saline syringe to the hub and aspirate blood. Take needed samples and then flush the line with saline and recap. Repeat this step with all lumens.
33. Place biopatch on skin around the intersection with the lumen of the catheter.
34. Suture the line in place.
35. Enclose CVL site with sterile waterproof transparent dressing.
36. Confirm placement using CXR. Tip of catheter should be in the lower third of the SVC at the insertion of SVC into right atrium (tip at right bronchiotracheal angle or up to 2.5 cm below bronchiotracheal angle).
    - Alternatively, ultrasound can be used for subclavian line access but only a few limited studies have confirmed this as to date so will not describe in detail (see below for typical ultrasound technique used). See references for more information.

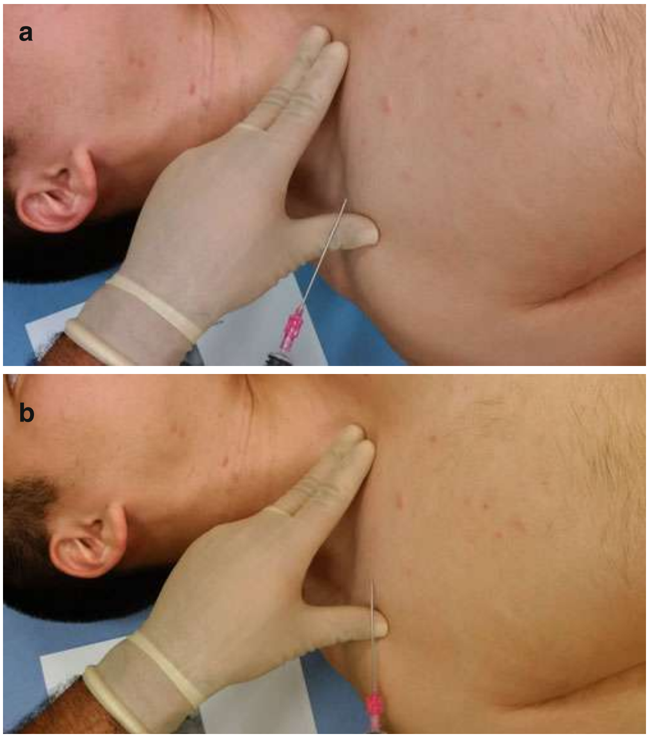

Fig. 3.7 Subclavian vein approach: wrong (**a**) and correct (**b**) angles to take when making skin puncture

3.4.2.1 Subclavian Vein Pearls and Pitfalls

- Inadequate landmark identification: operator should always palpate for landmarks and check anatomy prior to starting the procedure.
- Improper insertion position.
- Insertion of needle through periosteum.
  - Operator should NOT increase angle of needle to avoid the clavicle bone (this can cause a pneumothorax).
  - Operator should press on needle with downward pressure on chest wall to allow needle to maneuver under the clavicle without changing the angle of insertions of the needle.
- Taking too shallow a trajectory with needle.
- Aiming the needle too cephalad (aim for sternoclavicular junction).
- Failure to keep needle in place for wire passage: hand holding the needle should be planted on patient's chest for stabilization.

3.4.3 Femoral Vein Access Procedure

1. Palpate the patient's femoral artery below the inguinal ligament. This is usually found halfway between the anterior superior iliac spine (ASIS) and the midline of the symphysis pubis.
2. Trim overlying hair as necessary.
3. If ultrasound-guided approach is desired, use the linear probe (same as internal jugular) to detect the femoral vein at this location. The femoral vein will be easily compressible, while the femoral artery will be less compressible and pulsatile (Fig. 3.2).
4. Wash hands and use sterile technique to apply iodine or chlorhexidine solution (various forms available).
5. Open your femoral CVL kit and don cap, mask, sterile gown, and sterile gloves. Nonsterile assistants should wear a cap, mask, sterile gown, and sterile gloves. Flush all ports of your CVL kit with saline flushes and check for leaks or malfunction of catheter.
6. Under sterile technique, apply the drape over the area of insertion, and have an assistant extend the drape the length of the bed. Reapply sterile iodine or chlorhexidine at the site.
7. Anesthetize the skin overlying the femoral vein with lidocaine.
8. If ultrasound-guided approach is desired, have an assistant hold up the vascular ultrasound probe. Place your gloved hand through a sterile ultrasound sleeve and grasp the top of the ultrasound probe. Without breaking sterile technique, pull and invert the sterile ultrasound sleeve over the probe and cable. See Sect. 3.6 for detailed ultrasound-guided cannulation technique.
9. Insert the needle at a 45° angle, bevel down, directed superiorly, 1 cm medial to the palpable femoral artery pulse. Once the needle has broken the skin, aspirate by applying a small amount of continuous traction on the plunger of the attached syringe.
10. Advance smoothly and slowly until blood appears in the syringe. Stop once blood is aspirated. If femoral vein is not cannulated, withdraw your needle until just beneath the skin and redirect.
11. Hold the hub of the needle with thumb and forefinger to immobilize in place. Remove the syringe carefully. If blood appears arterial or pulsatile, remove the needle and hold pressure for 5–10 min.
    - Use the palm of your hand on the thigh to stabilize your hand. Not having your hand stable is a common mistake leading to needle movement out of the vein.
12. If blood appears venous (dark color, emerges as a continuous trickle, or transduced), cannulate the needle with the guide wire. Maintain a two-finger grip on the guide wire at all times. Advance the guide wire until approximately 15 cm remains.
    - Must keep handle of guide wire at ALL times and can be done through proper technique.
13. If guide wire does not advance easily, remove guide wire and reposition needle until blood aspirates easily. The J-tip can be straightened using a pinching motion if needed (Fig. 3.4)
    - Alternatively, the catheter/syringe found in most kits can be used as a bridge to guide wire placement. For the author, this has improved success when there is difficulty in wire placement. Use the same steps above with the catheter (Fig. 3.5) and when you have return of blood, advance the angiocath into the vein followed by insertion of the guide wire through the angiocath. This is especially useful in moving/agitated patients, patients who have collapsible veins due to hypovolemia, and patients who have abnormal anatomy and may have veins that take an abnormal angle shortly past the needle tip.
14. Using your scalpel, make a single 1/2 cm stab incision at the site of needle insertion to assist with dilator placement.
    - You can do the stab incision with or without the needle in place, but from experience, most novices have difficulty locating the correct stab location due to small amount of blood accumulation when the needle is taken out.
15. Remove your needle, carefully leaving guide wire in place. Apply dilator over guide wire and advance into the body with gentle pressure and a twisting motion in the same plane that you used to direct the needle.
16. Holding pressure at the insertion site with sterile gauze 4×4 pads, remove the dilator from the guide wire,

leaving the guide wire in place. Insert the central venous catheter over the guide wire until it fits snugly against the skin. Use the markings on the catheter to determine proper length placed (usually with femoral site you can "hub" the catheter) (Fig. 3.6).

17. Using a sterile saline flush, ensure that each lumen of the CVL draws blood easily and flushes easily. Carefully note any that do not and consider repositioning as needed. You can attempt moving catheter in or out a few centimeters or rotating the catheter and recheck. Apply caps to each open lumen of the CVL.
    - Remember to cover the introducer port if not using immediately. If you do not, it poses an infection and air embolism risk to the patient.
18. Suture the line in place.
19. Place an antibiotic biopatch or similar antimicrobial dressing.
20. Enclose CVL site with sterile waterproof transparent dressing.

3.4.3.1 Femoral Vein Pearls and Pitfalls

- Femoral central venous lines cannot accurately transduce central venous pressures.
- Asking the patient to perform a Valsalva maneuver has been shown to increase the width of the femoral vein by 1/3.
- The mnemonic NAVEL (Nerve Artery Vein Empty space Lymphatics) assists in remembering the order of femoral structures from lateral to medial.
- Traditionally, femoral venous lines were thought to have higher rates of infection than subclavian or internal jugular lines, but more recent analyses are challenging this belief.
  - Obesity is a more important risk factor for infection in femoral sites.

3.5 Complications

3.5.1 Jugular and Subclavian Complications

- Pneumothroax/Hemothorax
  - Prevention: Remove patient from ventilator before advancing the needle, choose the right side rather than left, and avoid multiple attempts when possible.
  - Management: Check postprocedure x-ray; if pneumothorax, arrange for thoracostomy depending on the size of the hemo-/pneumothorax.
- Catheter embolization
  - Prevention: Never withdraw a catheter past a needle bevel as this might shear off the catheter.
  - Management: X-ray the patient and contact specialist who can remove the embolized catheter.
- Arterial puncture: Hold compression if this occurs.
- Hematoma: Usually requires monitoring only.
- Thrombosis: This complication may lead to pulmonary embolism.
- Local site or systemic infection: Using maximal sterile precautions has been shown to greatly decrease rate of infection.
- Air embolism
  - May be caused by negative intrathoracic pressure when inspiration by the patient drawing air into an open line hub
  - Prevention: Be sure the line hubs are always occluded; placing the patient in the Trendelenburg position lowers the risk.
  - Management: The patient should be placed in Trendelenburg position with a left lateral decubitus tilt, which may prevent the movement of air into the right ventricle and onward into the left side of the heart. One hundred percent oxygen should be administered to speed the resorption of the air. If a catheter is located in the heart, aspiration of the air should be attempted.
- Dysrhythmias: Due to cardiac irritation by the wire or catheter tip. This can usually be terminated by simply withdrawing the line into the superior vena cava. One should always place a central venous catheter with cardiac monitoring.
- Lost guide wire: If the operator is not careful about maintaining control of the guide wire, it may be lost into the vein. This requires retrieval by interventional radiology or surgery and is an emergency.
- Catheter tip too deep: Check the postprocedure chest radiograph and pull the line back if the tip disappears into the cardiac silhouette.
- Catheter in the wrong vessel: Check the postprocedure chest radiograph for this complication; remove catheter and try again.
- Arterial puncture (subclavian only): The subclavian artery cannot be compressed; so, the subclavian approach should be avoided in anticoagulated patients.

3.5.2 Femoral Complications

- Arterial puncture: The femoral artery site can be compressed, so if punctured hold pressure.
- Hematoma: Usually requires monitoring only.
- Thrombosis: This complication may lead to pulmonary embolism.
- Catheter embolization.
  - Prevention: Never withdraw a catheter past a needle bevel which might shear off the catheter.
  - Management: X-ray the patient and contact specialist who can remove the embolized catheter.

- Lost guide wire: If the operator is not careful about maintaining control of the guide wire, it may be lost into the vein. This requires retrieval by interventional radiology or surgery and is an emergency.
- Local site or systemic infection: Using maximal sterile precautions has been shown to greatly decrease rate of infection.

3.6 Ultrasound Guided Cannulation: Tips for Each Approach

- Venous anatomy is best visualized using high-frequency (5–10 MHz) linear probe. Higher frequencies generate less penetration but better resolution.
- You can use the ultrasound to identify the location of the vessel prior to the procedure and utilize external landmarks during the procedure itself (static technique), or you can use the ultrasound to visualize cannulation of the vessel during the procedure (dynamic technique).
- Static view is advantageous in that the ultrasound transducer is not needed during the sterile portion of the procedure, but it does not allow for direct visualization of cannulation and guidance during the procedure.
- Dynamic view (preferred) allows for direct visualization during the procedure but requires more technique and requires use of transducer during the sterile portion of the procedure.
- The dynamic technique can be used in either a short-axis view, where a cross-sectional view of the vessel and needle is used, or a long-axis view, where a longitudinal view of vessel and needle is used (Fig. 3.3).
- The long-axis view allows for full visualization of the needle throughout the procedure and allows for better visualization and adjustment of needle depth. It is more difficult for lateral changes in positioning and tends to be more difficult technically.
  - Key in this view is that once a good section of vein is obtained, do not move probe to visualize the needle; move the needle into the ultrasound view by slightly adjusting trajectory.
- The short-axis view allows for lateral changes in position but is not as good at visualizing depth throughout the procedure, as visualization of the needle is in cross-sectional imaging. Perforation of the posterior wall is more common in this view.
- When using the short-axis view, remember to position the ultrasound probe such that the field of the ultrasound intersects the vessel (internal jugular, subclavian, femoral) at the anticipated site of insertion of the needle into the vein. Remember that the needle is only visualized as it intersects the plane of the ultrasound.
- When using the long-axis view, make sure to visualize the vessel with the ultrasound such that you can see the greatest diameter of the vessel along the entire length of the ultrasound probe. Keep the ultrasound steady during the procedure, and insert the needle at an angle at the lateral edge of the ultrasound probe. Using this technique, one can visualize the entire length of the needle.

3.7 Removing a Central Line

1. Place patient in supine or Trendelenburg position (for femoral removal can help decrease bleeding).
2. Remove suturing and dressing.
3. Jugular and subclavian: Have patient exhale and pull the line during the exhalation.
   - Exhalation increases intrathoracic pressure as compared to atmospheric pressure, thereby reducing the risk of air thromboembolism.
4. Hold pressure for approximately 1 min to stop bleeding.
5. Dress with a sterile dressing.
6. If central line-related infection is suspected, cut off the tip with sterile scissors and send for culture.

Selected Reading

Internal Jugular Vein Access

McGee DC, Gould MK. Preventing complications of central venous catheterization. N Engl J Med. 2003a;348(12):1123–33.

Mimoz O, et al. Chlorhexidine-based antiseptic solution vs alcohol-based providone-iodine for central venous catheter care. Arch Intern Med. 2007;167(19):2066–72.

Noble V, et al. Manual of emergency and critical care ultrasound. Cambridge: Cambridge University Press; 2007. p. 196–204.

Parry G. Trendelenburg position, head elevation and a midline position optimize right internal jugular vein diameter. Can J Anaesth. 2004;51(4):379.

Vesely T. Central venous catheter tip position: a continuing controversy. J Vasc Interv Radiol. 2003;14(5):527.

Subclavian Vein Access

Elliott TS, Faroqui MH, Armstrong RF, Hanson GC. Guidelines for good practice in central venous catheterization. Hospital Infection Society and the Research Unit of the Royal College of Physicians. J Hosp Infect. 1994;28(3):163–76.

Fortune JB, Feustel P. Effect of patient position on size and location of the subclavian vein for percutaneous puncture. Arch Surg. 2003;138(9):996–1000; discussion 1001.

Fragou M, et al. Real time ultrasound-guided subclavian vein cannulation versus the landmark method in critical care patients: a prospective randomized study. Crit Care Med. 2011;39(7):1–6.

Kilbourne MJ, Bochicchio GV, Scalea T, Xiao Y. Avoiding common technical errors in subclavian central venous catheter placement. J Am Coll Surg. 2009;208(1):104–9.

McGee DC, Gould MK. Preventing complications of central venous catheterization. N Engl J Med. 2003b;348(12):1123–33.

Femoral Vein Access

Dailey RH. Femoral vein cannulation: a review. J Emerg Med. 1985;2:367–72.

Lim T, Ryu H-G, et al. Effect of the bevel direction of puncture needle on success rate and complications during internal jugular vein catheterization. Crit Care Med. 2012;40(2):491–4.

McGee DC, Gould MK. Preventing complications of central venous catheterization. N Engl J Med. 2003c;348:1123–33.

Marik PE, Flemmer M, Harrison W. The risk of catheter-related bloodstream infection with femoral venous catheters as compared to subclavian and internal jugular venous catheters: a systemic review of the literature and meta-analysis. Crit Care Med. 2012;40:2479–85.

Swanson RS, Uhlig PN, Gross PL, et al. Emergency intravenous access through the femoral vein. Ann Emerg Med. 1984;13:244–7.

Pulmonary Artery Catheter

4

Rohit Pravin Patel and Marie-Carmelle Elie

4.1 Indications

- Prevention or treatment of multiorgan failure in high-risk patients
- Preoperative and postoperative management in high-risk patients with cardiac, pulmonary, or renal dysfunction
- Patients with anticipated large fluid shifts (sepsis, bleeding, burns, cirrhosis)
- Oliguria or hypotension not relieved by fluids
- Suspected cardiac event leading to shock
- For continuous SVO2 (central venous oxygenation) monitoring in shock
- To differentiate shock states
- For monitoring cardiac output in patients requiring high-positive end-expiratory pressure (>14 cm H20)
- Monitoring and management of complicated myocardial dysfunction or cardiogenic shock
- Congestive heart failure with poor response to afterload reduction and diuretic therapy
- Suspected tamponade or contusion from blunt chest injury
- Pulmonary hypertension with myocardial dysfunction
- Diagnosis of primary pulmonary hypertension
- Aspiration of air emboli
- Direct pulmonary artery administration of thrombolytic therapy

4.2 Contraindications

- Tricuspid or pulmonary valve mechanical prosthesis
- Right heart mass (thrombus or tumor)
- Tricuspid or pulmonary valve endocarditis
- Recurrent sepsis (catheter could serve as nidus for infection)
- Hypercoagulopathy (catheter could serve as site for thrombus formation)
- Patient known sensitivity to heparin (catheters with heparin coating)
- Electrocardiographic (ECG) monitoring encouraged in conditions of complete left bundle branch block (risk of complete heart block increased), Wolfe–Parkinson–White syndrome, and Ebstein's malformation (risk of tachyarrhythmias)

4.3 Materials

- Pulmonary artery or Swan–Ganz catheter
- Percutaneous sheath introducer and contamination shield
- Compatible cardiac output computer for measuring cardiac output by the bolus thermodilution method
- Injectate temperature sensing probe (bolus thermodilution method)
- Connecting cables
- Sterile flush system and pressure transducers
- Bedside ECG and pressure monitor system
- Appropriate ECG "slave" cables

4.3.1 Catheter Preparation

- Avoid forceful wiping or stretching of catheter to avoid injury to the thermistor wire circuitry; wiping the heparin coat may cause removal of the coating.
- In vivo calibration is required if in vitro calibration is not done; refer to the monitor operator's manual for detailed calibration instructions.
- Connect catheter's injectate and pressure-monitoring lumens to the flush system and pressure transducers; ensure all lines are free of air.
- Connect the thermistor to the monitor and confirm no fault messages appear.

R.P. Patel, MD • M.-C. Elie, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: rohitpatel@ufl.edu; elie@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_4

4.4 Procedure

1. Maintain sterile precautions including sterile cap, mask, gown, gloves.
2. Place a central venous line introducer and verify placement of introducer with chest radiograph; it is also acceptable to obtain the chest radiograph after the insertion of the pulmonary artery catheter if no complication was suspected with the central venous line introducer.
3. Cleanse the skin and introducer thoroughly with chlorhexidine.
4. Have assistant open pulmonary artery catheter kit in sterile fashion.
5. Remove pulmonary artery catheter from kit and have assistant hook up all the ports to the transducers and make sure readings are accurate as the catheter is being manipulated.
6. Gently lift the distal portion of the catheter up from the silicone gripper; do not pull the balloon through the gripper to avoid damage.
7. Have assistant check the proximal and distal ports for patency by flushing with sterile saline. Also have the assistant check the patency of the balloon with the syringe provided in the kit (Fig. 4.1). Check for major asymmetry and for leaks (optional) by submerging in sterile saline or water. Deflate balloon prior to insertion. Carefully wave the distal catheter segment up and down to confirm electrical continuity by observing a pressure tracing on monitor. Ensure proper readings, no information is sometimes better than wrong information. Make sure each port transduces appropriately *prior to* insertion.
8. Familiarize yourself with the catheter line markings. Each thin line indicates 10 cm from the tip and thick line indicates 50 cm from the tip. These are used in combination to indicate length from tip (Fig. 4.2).
9. Place the sterile plastic sleeve (lock side toward patient, Fig. 4.3) over the catheter after flushing all ports to further protect the catheter during manipulation.
10. The distal end of the catheter is inserted into the introducer hub of the central venous line and threaded to the superior vena cava. The catheter must be placed at least 30 cm into the introducer for the balloon to clear the distal end of the introducer prior to inflation. At *no point* should the catheter be withdrawn with the balloon inflated; ensure the assistant has deflated the balloon prior to withdrawal. The balloon assists in directing the catheter through the vascular system using the directional blood flow.
11. At 20 cm, the balloon should be inflated and catheter advanced through right atrium, past the tricuspid valve into the right ventricle, then past the pulmonary valve to the pulmonary artery. The waveform and pressure readings can guide you through the various locations (Fig. 4.4).
12. Once in the pulmonary artery, the catheter should be carefully and slowly advanced to wedge position. The balloon can be deflated and pulmonary artery tracings should reappear. If a wedge is obtained with less than the maximum recommended volume, the catheter should be withdrawn to a position where full inflation volume produces a wedge tracing. Avoid prolonged times when obtaining wedge pressure (2 respiratory cycles or 10–15 s), especially in patients with pulmonary hypertension.
13. General guidelines for distance necessary at various points include: right atrium 20–25 cm, right ventricle 30–35 cm, and pulmonary artery 40–45 cm; catheter usually wedges at 50–55 cm. These are dependent on the starting location you are using to advance the catheter (subclavian, internal jugular, femoral).
14. Once the catheter is in correct position, it should be locked into place with the plastic sleeve tip onto the hub of the introducer.
15. Correct placement is confirmed with chest radiograph (Fig. 4.5).

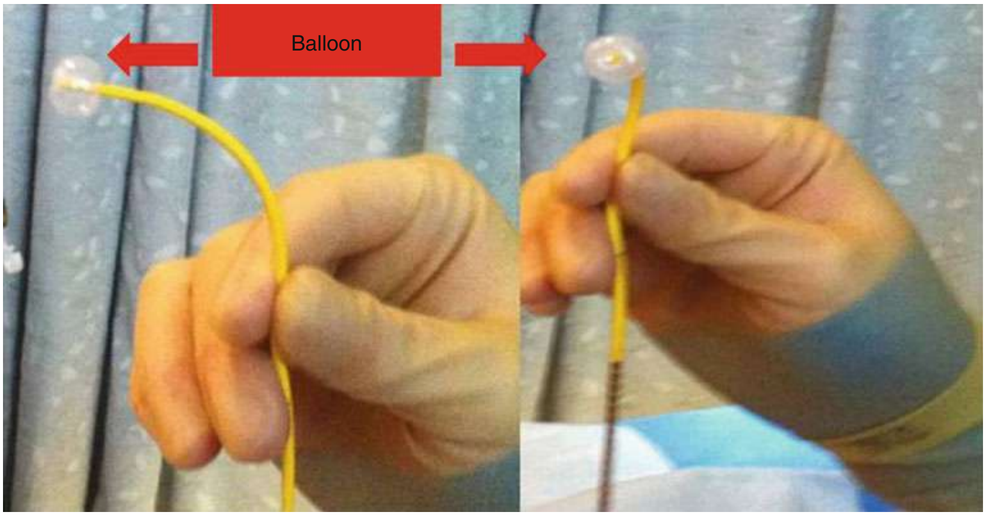

Fig. 4.1 Balloon inflation prior to insertion for evaluation of patency or leaks

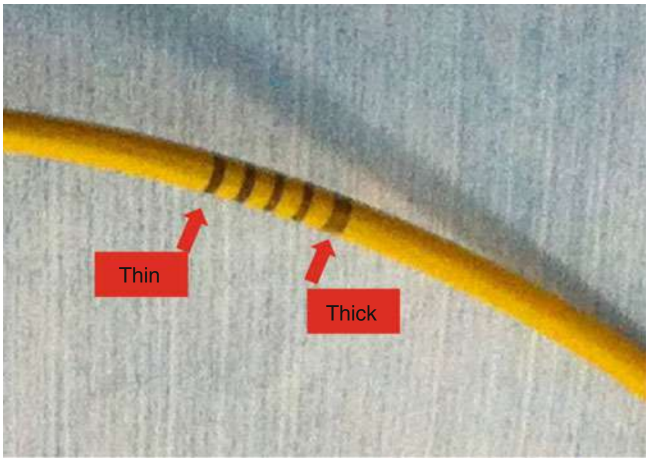

Fig. 4.2 Thick and thin markings found on the catheter representing length from distal tip

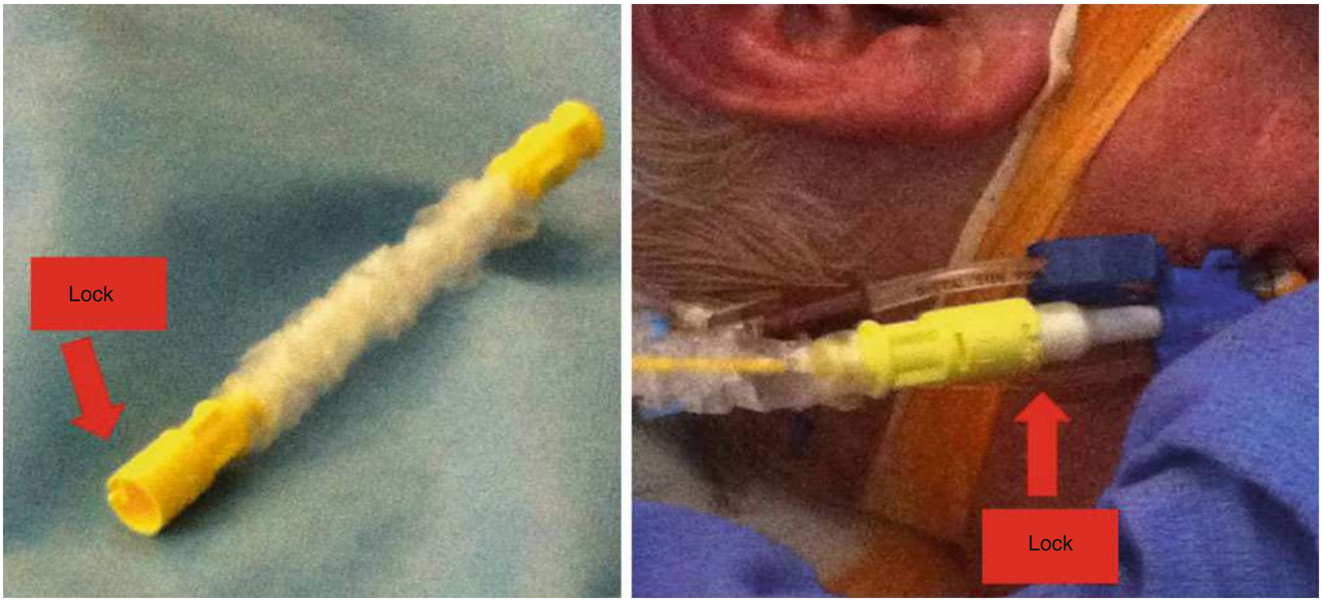

Fig. 4.3 Lock position on catheter for stabilization of catheter

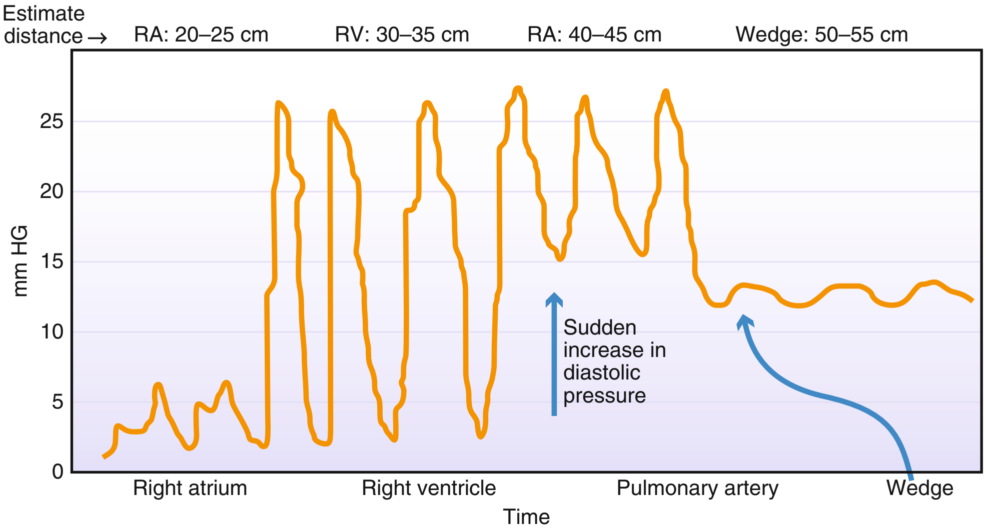

Fig. 4.4 Typical waveform seen at specific locations in the heart with associated estimated distances from catheter tip

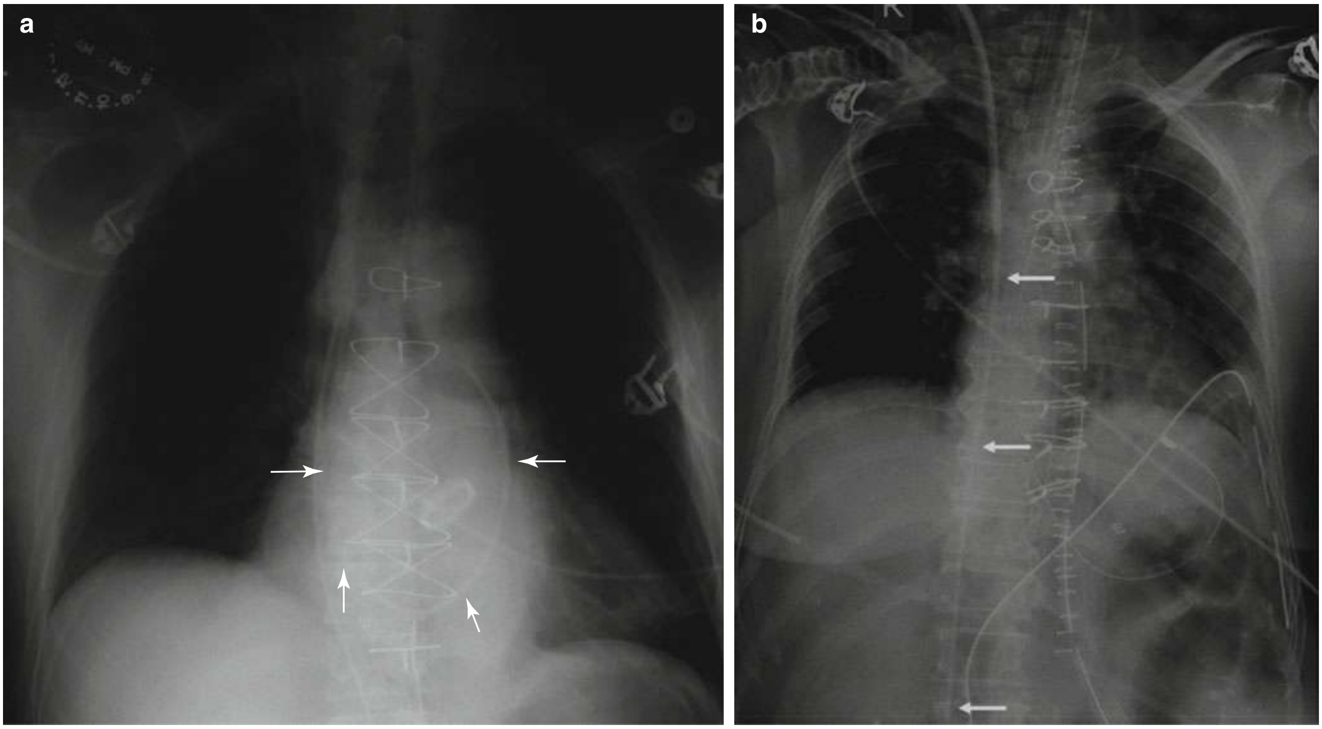

Fig. 4.5 Chest radiograph demonstrating correct (**a**) and incorrect (**b**) positioning of pulmonary artery catheter (*arrows*) (**a** Reproduced with permission from McGraw-Hill: Stead LG, et al. *First Aid for the Radiology Clerkship*. New York: McGraw-Hill; 2009; **b** Reproduced with permission from Wolters Kluwer: Jain SN. A pictorial essay: radiology of lines and tubes in the intensive care unit. *Indian J Radiol Imaging*. 2011;21(3):182–190)

4.5 Complications

- Arrhythmias: most are premature ventricular contractions that are self-limiting and resolve with advancement into pulmonary artery or withdrawal into atrium.
- Right bundle branch block: usually transient after positioning catheter into the pulmonary artery; if has already left bundle branch block may lead to complete heart block; should have temporary pacing equipment on standby.
- Knotting in the right ventricle (RV): risk increased in those with dilated cardiac chambers; a persistent RV tracing (15 cm beyond the point where initial RV tracing was observed) should alert you to this possibility.
- Pulmonary artery rupture: age >60 year, anticoagulation therapy, and presence of pulmonary hypertension increase risk of rupture; hemoptysis shortly after placement is indicative and management includes lateral decubitus positioning (bleeding side down), intubation with double lumen tube, and increasing positive end-expiratory pressure (PEEP).
- Infection.
- Pulmonary infarction: due to unintentional migration of distal tip.

4.6 Pearls and Pitfalls

- If the catheter requires stiffening during insertion, slowly perfuse the catheter with 5–10 mL of cold sterile solution as the catheter is advanced through a peripheral vessel.
- The incidence of complications increases significantly with periods of use longer than 72 h, so assess the need for the catheter on daily basis.
- Anticipate spontaneous catheter tip migration toward periphery of pulmonary bed; if a wedge tracing is observed when balloon is deflated, pull the catheter back.

Selected Reading

1. Edward Lifesciences. Pulmonary artery catheter Instruction manual. 2009. http://ht.edwards.com/resourcegallery/products/swanganz/pdfs/invasivehdmphysprincbook.pdf.
2. Edward Lifesciences. Pulmonary artery catheter Instruction manual. 2009. http://www.edwards.com/products/pacatheters/Pages/ThermodilutionCatheter.aspx.
3. Leatherman JW, Marini JJ. Clinical use of the pulmonary artery catheter. In: Hall JB, Schmidt GA, Wook LDH, editors. Principles of critical care. 3rd ed. New York: McGraw-Hill; 2005. p. 146–50.
4. Moran SE, Pei KY, Yu M. Hemodynamic monitoring: arterial and pulmonary artery catheters. In: Gabrielli A, Layon AJ, Yu M, editors. Civetta, Taylor, and Kirby's critical care. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2009.

Noninvasive Cardiac Monitoring: The Edwards Vigileo System

5

Dawood G. Dalaly and Rohit Pravin Patel

5.1 Indications

Cardiac output monitoring is indicated when trying to determine fluid responsiveness in patients. It assists in directing and assessing results of resuscitative efforts to ensure appropriate tissue perfusion. Although most catheters are systemically invasive, tools like the Vigileo (Edwards Lifesciences; Irvine, CA) are excellent noninvasive devices for determining values such as the stroke volume, stroke volume variation, stroke volume index, cardiac output, and cardiac index (Fig. 5.1).

D.G. Dalaly, DO
Department of Surgery,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: ddalaly@ufl.edu

R.P. Patel, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: drpratik@gmail.com

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_5

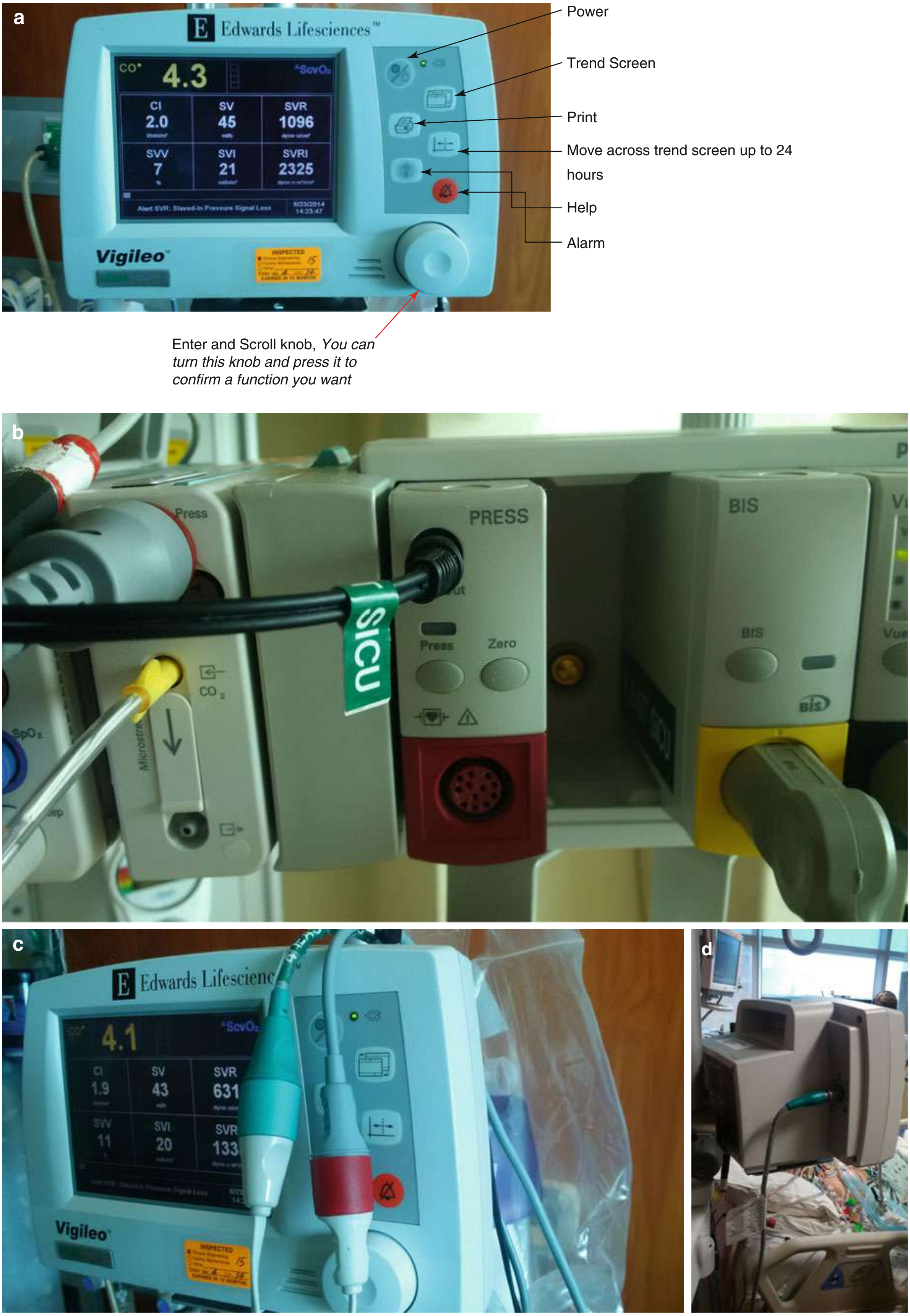
a
Edwards Lifesciences
CO 4.3
ScvO2
CI 2.0
SV 45
SVR 1096
SVV 7
SVI 21
SVRI 2325
Vigileo
Power
Trend Screen
Print
Move across trend screen up to 24 hours
Help
Alarm
Enter and Scroll knob, *You can turn this knob and press it to confirm a function you want*
b
Press
PRESS
SICU
Zero
BIS
CO2
c
Edwards Lifesciences
CO 4.1
CI 1.9
SV 43
SVR 631
SVV 11
SVI 20
Vigileo
d

Fig. 5.1 Typical monitoring devices and connecting devices from monitor to patient: (**a**) FloTrac monitor, (**b**) FloTrac red port connects to the heart monitor, (**c**) the FloTrac system connects to the arterial line on one end and the pressure bag/monitors on the other; monitor connections are color coded, (**d**) the green port connects to the green port from the FloTrac

5.2 Contraindications

There are no contraindications to noninvasive monitoring of the heart, although most systems will need to be connected to an arterial line catheter, therefore contraindications to arterial line placement must be determined. Current literature supports the use of noninvasive monitoring to those who are 100 % supported ventilation with tidal volumes ≥8 mL/kg. There is no support for use in those with spontaneous breaths or arrhythmia. Patients with significant variation in respiratory pattern may have results that are unreliable.

5.3 Definitions and Values (Table 5.1)

Table 5.1 Cardiac output monitoring definitions

Term	Reference range	Definition
Cardiac output	4–8 L/min	Volume of blood being pumped by the heart in 1 min
Cardiac index	2.5–4 L/min/m^2	Compares the amount of fluid being pumped by the heart with an individual's body surface area
Stroke volume	40–80 mL/beat	Volume of blood being pumped by the ventricle per beat
Stroke volume variation	10–15 %	Represents percentage of change between minimum and maximum stroke volumes and predictor of fluid responsiveness
Stroke volume index	33–47 mL/m^2/beat	Quantity of blood ejected from the heart per beat
Mixed venous saturation (SvO$_2$)	60–80 %	Percentage of oxygen bound to hemoglobin in blood returning to the right side of the heart; represents oxygen delivery and consumption at the tissue level. Usually obtained from pulmonary artery catheter
Central venous oxygen saturation (ScvO$_2$)	70 %	Surrogate marker for SvO$_2$, usually obtained from internal jugular or subclavian catheters

5.4 Materials

- FloTrac (Edwards Lifesciences, Irvine, CA, USA)
- Vigileo monitor
- Pressure bag

5.5 Procedure

1. Connect FloTrac to arterial line and distal FloTrac port to pressure bag. Inflate bag to 300 mmHg.
2. Connect green FloTrac cord to green Vigileo cord and red FloTrac cord to arterial line port on heart monitor.
3. Turn the port on the FloTrac off to the patient and zero on your heart monitor as you would with an arterial line. At the same time, you should have pressed the "Enter" knob on the Vigileo system, scrolled it to "zero arterial pressure," and pressed the knob again for that function.
4. You should now have an arterial waveform on your heart monitor and your cardiac values on the Vigileo system.

5.6 Pearls and Pitfalls

- Some components of the values obtained are not reliable on spontaneously breathing patients and it is critical to check the ventilator waveforms for these breaths.
- Patients without adequate tidal volumes (at least 8 mL/kg) will also have unreliable values.
- Stroke volume variation usually is more reliable when greater than 13 % and indicates fluid responsiveness rather than when it is less than 13 % (similar to a low central venous pressure (CVP) being more informative than normal or high CVP levels).

Selected Reading

1. Alarcon LH, Fink MP. Chapter 13. Physiologic monitoring of the surgical patient. In: Brunicardi FC, Andersen DK, Billiar TR, et al., editors. Schwartz's principles of surgery. 9th ed. New York: McGraw-Hill; 2010.
2. Holcroft JW, Anderson JT, Sena MJ. Chapter 12. Shock and acute pulmonary failure in surgical patients. In: Doherty GM, editor. CURRENT diagnosis & treatment: surgery. 13th ed. New York: McGraw-Hill; 2009. Available at http://www.accesssurgery.com/content.aspx?aID=5212482. Accessed 22 Aug 2012.
3. Edwards Critical Care Education. Available at http://www.edwards.com/education/Pages/cceducationmap.aspx. Accessed 11 Dec 2012.

Peripheral Venous Cutdown

6

Jeffrey Kile, Katrina John, and Amish Aghera

6.1 Indications

- Distorted anatomy of peripheral venous access sites
- Unavailability of cannulable veins (e.g., in hypovolemia, burn victim, traumatic anatomy, sclerosed veins, etc.)
- Emergency venous access for infusion/transfusion
- Unavailability of central venous access or less invasive means peripherally

6.2 Contraindications

- Absolute
  - Availability of less invasive or less time-consuming means of vascular access
  - Overlying infection, traumatic tissue, burn, etc., at cutdown site
  - Traumatic injury proximal to cutdown site
- Relative
  - Coagulation disorders

6.3 Materials and Medications (Fig. 6.1)

- Sterile gloves
- Antimicrobial solution and swabs
- 4″×4″ gauze sponges
- Local anesthetic (1 % lidocaine 5 mL)
- 5-mL syringe
- Blunt needle
- 25- or 27-gauge needle
- Scalpel
- Vein dilator/lifter
- Peripheral intravenous catheter
- Curved hemostat
- 0-0 silk sutures or 4.0 nylon sutures
- Iris scissors
- Intravenous infusion tubing
- Adhesive tape

J. Kile, MBBS, PhD, MPH (✉) • K. John, MBBS
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: jeffrey.kile@gmail.com; trenjohn@me.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_6

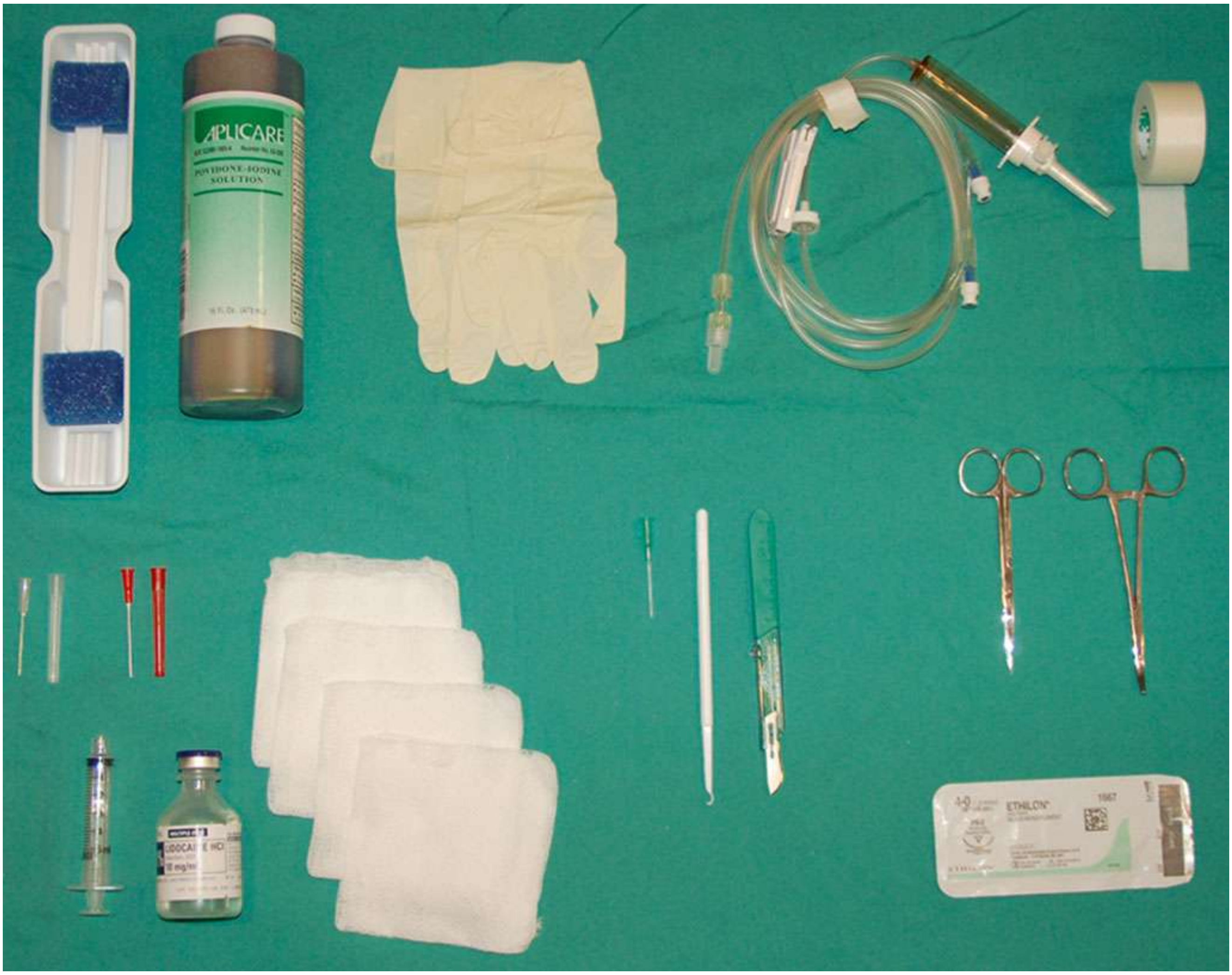

Fig. 6.1 Materials and medications

6.4 Choice of Vessel for Cutdown

- Greater saphenous vein: this vessel is the longest vein in the body, is predominantly subcutaneous, and is exposed with minimal blunt dissection just anterior to the medial malleolus at the ankle.
- Basilic vein: this vessel is reliably located 1–2 cm lateral to the medial epicondyle on the anterior aspect of the humeral region, typically catheterized just superior to the antecubital fossa approximately, and its diameter permits its localization relatively easily even in the hypotensive patient.
- Cephalic vein: this vessel runs anteromedially from the radial aspect of the wrist to the antecubital fossa, is superficial and large in diameter, and is most easily cannulated at the distal flexor crease in the antecubital fossa.

6.5 Procedure

6.5.1 Standard Venous Cutdown Technique

1. Apply antimicrobial solution liberally to the skin surrounding the incisional area.
2. Establish a sterile field by placing drapes around the incisional area.
3. Apply a tourniquet proximal to the planned cutdown site to maximize visualization of vein to be cannulated.
4. Inject local anesthetic to raise a small (0.5 cm) wheal using 25- or 27-gauge needle and then insert the tip of the needle through wheal to infiltrate skin superficial to the artery with approximately 4 mL of anesthetic.
   - Injection of local anesthetic into the vessel may precipitate arrhythmia, so draw back on the plunger prior to infiltration to ensure the tip of the needle is not inside the vessel.
5. Incise the skin with scalpel perpendicular to the course of the vein through all cutaneous layers until subcutaneous fat is visualized (Fig. 6.2).
   - Some practitioners prefer using a longitudinal incision to reduce the risk of transecting neurovascular structures, but this may not produce sufficient exposure of vein.
6. Using a curved hemostat or gloved finger, bluntly dissect the subcutaneous tissue to isolate and mobilize approximately 2–3 cm of the vein (Fig. 6.3).
   - A small self-retaining retractor or tissue spreader can be used in this step to improve visualization of vein if desired
7. Pass suture under the vein distal to the planned venous puncture site using hemostat to stabilize the vein and tie the suture over the vein (Fig. 6.4).
8. Pass a second suture under the vein proximal to the planned venous puncture site using hemostat (Fig. 6.5).
   - This step enables increased visualization, vessel control, and hemostasis during incision.
   - Leave the ends of both sutures long to facilitate maneuvering the vein.
9. Incise one-half to one-third of the diameter of the vein using a scalpel or iris scissors held at a 45° angle to the vessel (Fig. 6.6).
10. Grasping the proximal edge of the incision with a hemostat to apply counter traction (in a distal direction), insert the tip of the catheter into the venous incision (Fig. 6.7).
    - Do not force the catheter if it does not easily advance.
    - Catheter can be introduced directly through the skin incision or via skin puncture adjacent to the skin incision.
    - If the catheter lacks a tapered tip, cut the distal end of the cannula at a 45° angle to fashion a beveled tip.
11. Thread catheter into vein (Fig. 6.8).
12. Aspirate any air which may have entered the cannula during insertion.
13. Connect hub of catheter to intravenous tubing.
14. Tie the proximal suture around the vein just proximal to the venous incision, encircling both the vein and the intraluminal cannula with the suture.
15. Remove tourniquet.
16. Secure the catheter hub to the skin using nylon (4.0) sutures as follows. Take a 0.5 cm bite of skin under the catheter hub with the suture needle, tie several knots in the suture without pinching the skin, then tie a second set of knots around the hub of the catheter firmly.
17. Close the incision using nylon (4.0) sutures.
18. Dress the wound with appropriate self-adhesive sterile dressing or sterile gauze pads and adhesive tape.

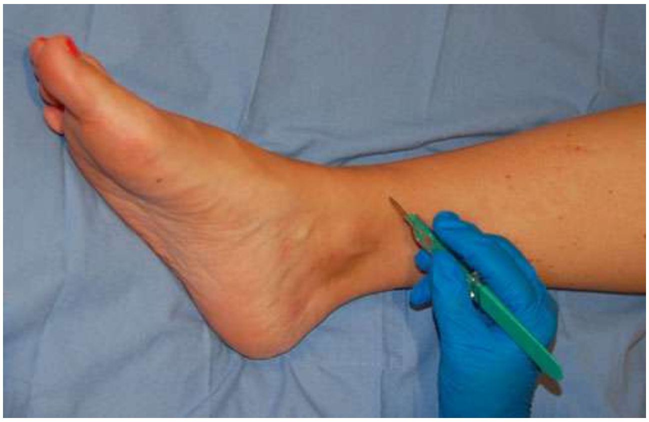

Fig. 6.2 Incision of skin

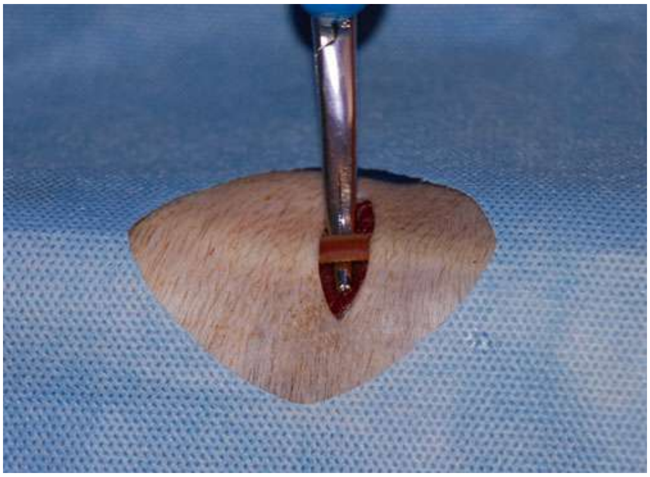

Fig. 6.3 Mobilization of vein

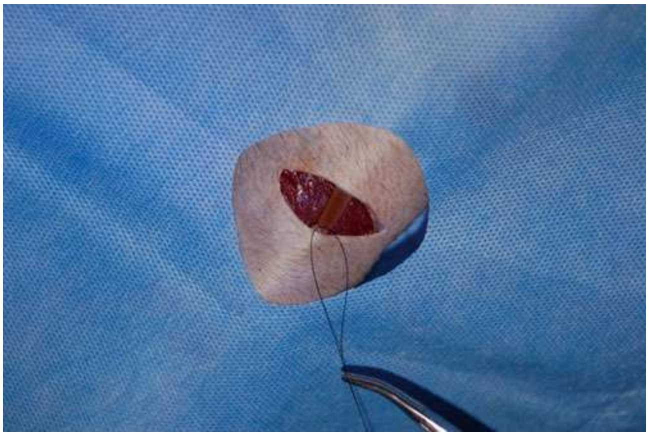

Fig. 6.4 Distal ligature tied around vein

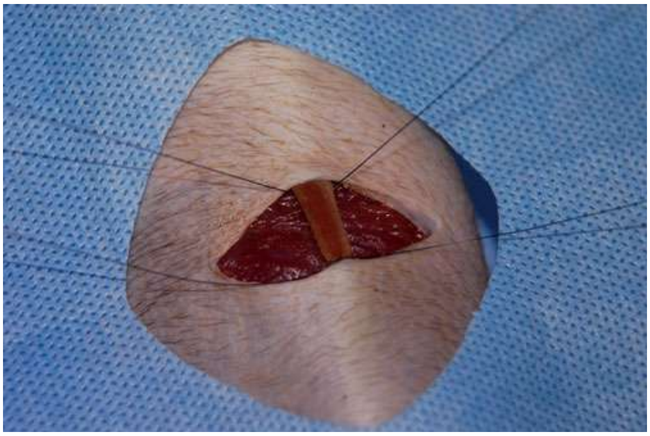

Fig. 6.5 Distal and proximal ligatures in place

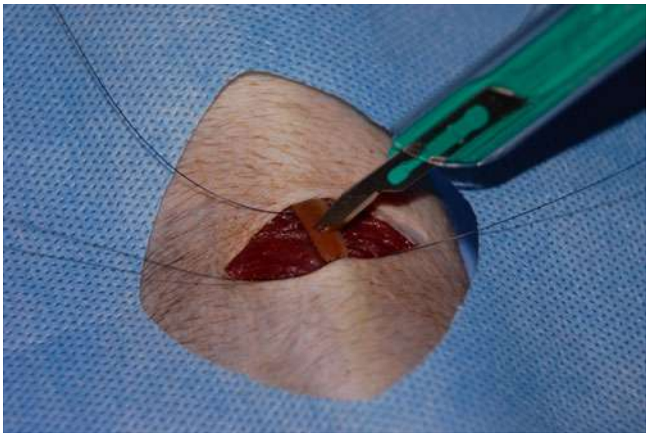

Fig. 6.6 Incision of vein

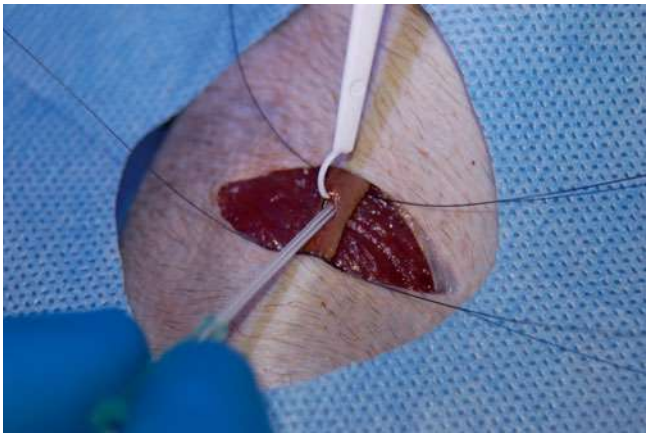

Fig. 6.7 Catheterization of vein

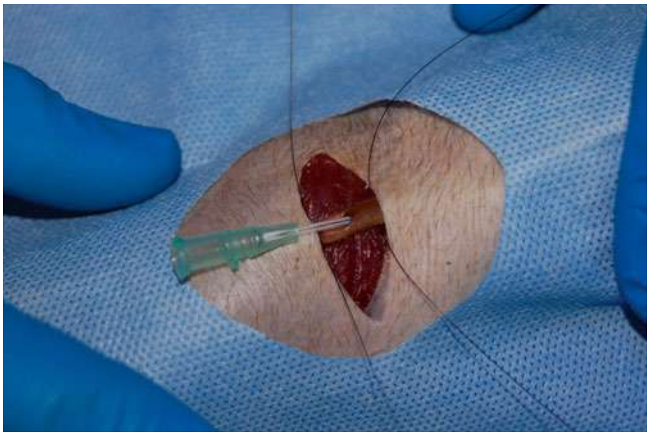

Fig. 6.8 Catheter threaded into vein

6.5.2 "Mini-Cutdown" Technique

(Perform steps 1–6 from the "Standard Venous Cutdown Technique," above, and then proceed with the steps below.)

1. Puncture the vein using a standard over-the-needle venous catheter.
   - Catheter can be introduced directly through the skin incision or via skin puncture adjacent to the skin incision.
2. Thread catheter into the vein over the needle.
3. Remove and discard the needle.
4. Aspirate any air which may have entered the catheter during insertion.
5. Connect catheter to intravenous tubing.

(Continue with steps 15–18 from the "Standard Venous Cutdown Technique," above.)

6.5.3 Modified/Guide Wire Technique

(Perform steps 1–6 from the "Standard Venous Cutdown Technique," above, and then proceed with the steps below.)

1. Insert the blunt end of the guide wire into the incised vein.
2. While stabilizing the guide wire at its insertion site, thread the dilator and sheath assembly over the free end of the guide wire until it is approximately one inch from the skin.
3. Grasp the free end of the guide wire protruding from the tail end of the assembly.
   - If it does not protrude from the tail end of the assembly, the guide wire must be removed sufficiently from the artery to be securely grasped. It must protrude visibly from the tail end of the dilator throughout the subsequent process of threading the dilator into the vein.
   - Never let go of the guide wire during this step, as insertion of the dilator and sheath assembly can otherwise push the guide wire completely into the vein.
4. Holding it firmly near its tip, thread the assembly over the wire into vessel with a gentle back-and-forth twisting motion.
5. Holding the sheath securely in the vein, remove and discard the dilator and guide wire.
6. Aspirate any air which may have entered the sheath during insertion.
7. Connect sheath to intravenous tubing.
8. Remove tourniquet.
9. Secure the sheath to the skin.
10. Close the incision using nylon (4.0) sutures.
11. Dress the wound with appropriate self-adhesive sterile dressing or sterile gauze pads and adhesive tape.

6.6 Complications

- Hematoma
- Infection
- Sepsis
- Phlebitis
- Embolization
- Wound dehiscence

6.7 Pearls and Pitfalls

- Fluids are infused most quickly via short, large-bore catheters.
- If the line is inserted for slow infusion of intravenous drugs, catheter lumen size is relatively insignificant.
- In larger children and adults, intravenous plastic tubing, small-bore pediatric feeding tubes, and Silastic catheters may be used as infusion catheters.
- Threading a 10-gauge intravenous catheter or intravenous tubing directly into the incised vein achieves excellent flow rates.
- If difficulty is encountered while threading the catheter into the incised vein, ensure an appropriately sized catheter has been used and that the vessel lumen has been correctly identified and that no false passage has been created in the adventitia.
- As compared to the standard venous cutdown technique, the mini-cutdown technique is easier and also preserves the vein, permitting repeated catheterization if necessary.
- As compared to the standard venous cutdown technique, the modified/guide wire technique reduces procedure time and increases the likelihood of vein salvage in the event of vessel transection.

Selected Reading

1. Chappell S, Vilke GM, Chan TC, et al. Peripheral venous cutdown. J Emerg Med. 2006;31:411–6.
2. Klofas E. A quicker saphenous vein cutdown and a better way to teach it. J Trauma. 1997;43:985–7.
3. McIntosh BB, Dulchavsky SA. Peripheral vascular cutdown. Crit Care Clin. 1992;8:807–18.
4. Nocito A, Wildi S, Rufibach K, et al. Randomized clinical trial comparing venous cutdown with the Seldinger technique for placement of implantable venous access ports. Br J Surg. 2009;96:1129–34.
5. Shockley LW, Butzier DJ. A modified wire-guided technique for venous cutdown access. Ann Emerg Med. 1990;19:393–5.

Part II

Airway and Vascular Procedures

Bag-Valve-Mask Ventilation

7

Braden Hexom and Tatiana Havryliuk

7.1 Indications

- Hypoxia
- Hypoventilation/apnea
- Rescue maneuver if failed intubation

7.2 Contraindications

- Absolute
  - Inability to ventilate due to lack of seal (thick beard, deforming facial trauma)
  - Inability to ventilate secondary to complete upper airway obstruction
  - Active, adequate spontaneous ventilation
- Relative
  - Full stomach (aspiration risk)
  - After induction and paralysis during rapid sequence intubation (aspiration risk)

7.3 Materials (Fig. 7.1)

- Bag valve mask (BVM) with reservoir
- Oxygen connector tubing
- Nasal pharyngeal airway/oral pharyngeal airway
- Lubricant jelly

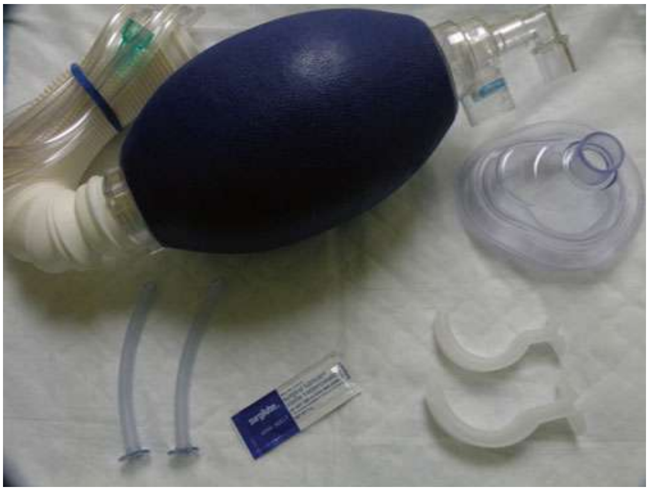

Fig. 7.1 BVM supplies: bag, mask, oral airways, nasopharyngeal airways, lubricant

B. Hexom, MD (✉)
Department of Emergency Medicine, Mount Sinai Hospital,
New York, NY, USA
e-mail: braden.hexom@mssm.edu

T. Havryliuk, MD
Department of Emergency Medicine, University of Colorado
Denver, Denver, CO, USA
e-mail: tatiana.havryliuk@ucdenver.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_7

7.4 Procedure

1. Position patient in "sniffing" position.
2. Open the airway with chin-lift/head-tilt or jaw thrust maneuvers.
3. Place airway adjuncts to maintain airway patency. Use oral airway (Fig. 7.2) in unconscious patients. Use nasal airway (Fig. 7.3) in semiresponsive patients.
4. Attach oxygen tubing to high-flow oxygen (15 L/min).
5. Place appropriately sized mask on patient's face covering the nose and mouth.
    - For one-handed technique (Fig. 7.4), use nondominant hand to make a "C" with index finger and thumb on top of the mask and form an "E" with the rest of the fingers using them to pull up on the mandible (E–C technique). Use the dominant hand to provide bag ventilations.
    - For two-handed (Fig. 7.5), two-person technique (preferred), make two semicircles with index fingers and thumbs of both hands on top of the mask and use the rest of the fingers to pull up on the mandible.
6. Consider the Sellick maneuver (cricoid pressure) to compress the esophagus against the cervical vertebrae, preventing gastric insufflation.
7. Ventilate patient providing reduced tidal volume breaths (500 mL) at a rate of 10–12 breaths per minute.
8. Give each breath gently over 1–1.5 s to avoid high peak pressures, avoiding gastric insufflation.
9. Prepare for definitive airway as dictated by the clinical scenario.

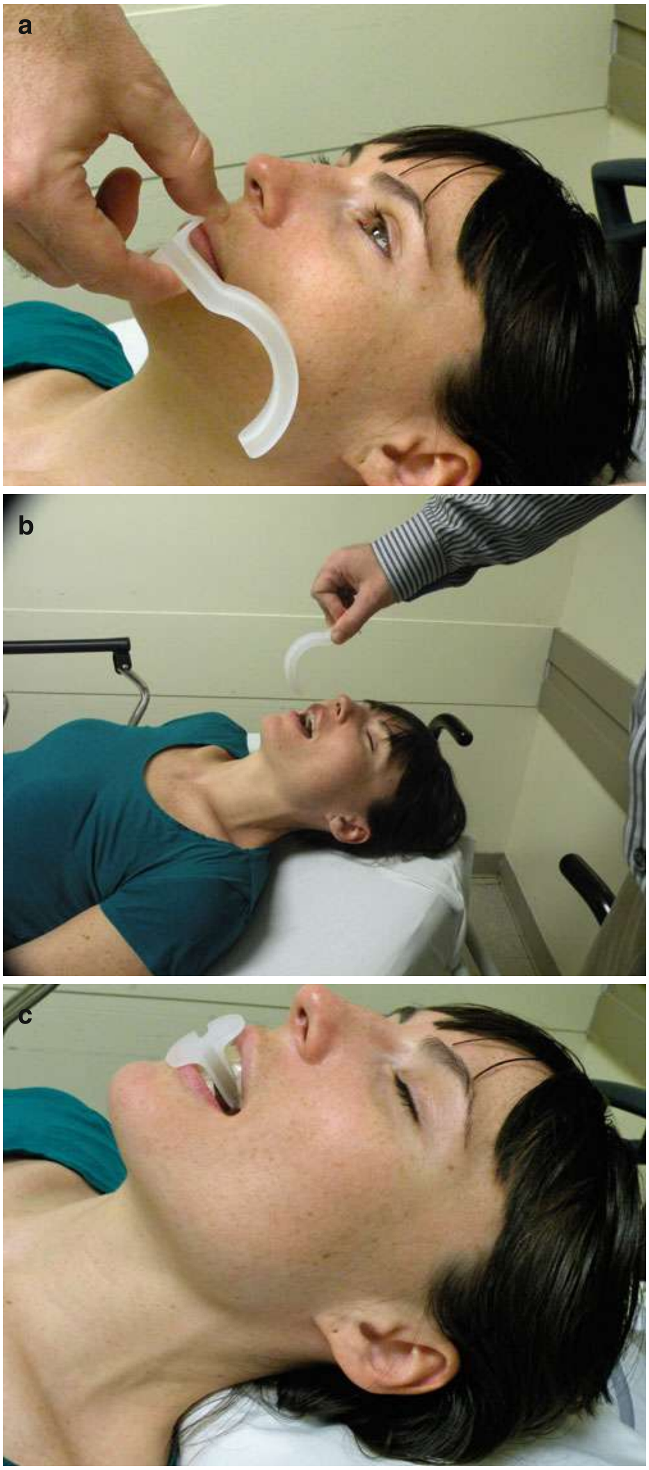

Fig. 7.2 (**a**–**c**) Oral airway insertion

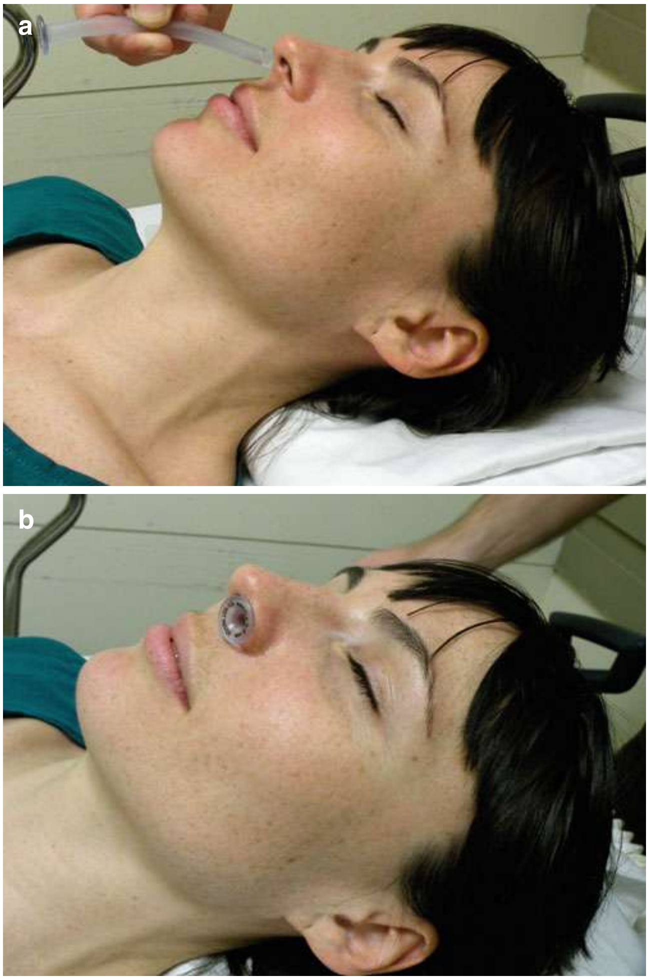

Fig. 7.3 (**a**, **b**) Nasal airway insertion

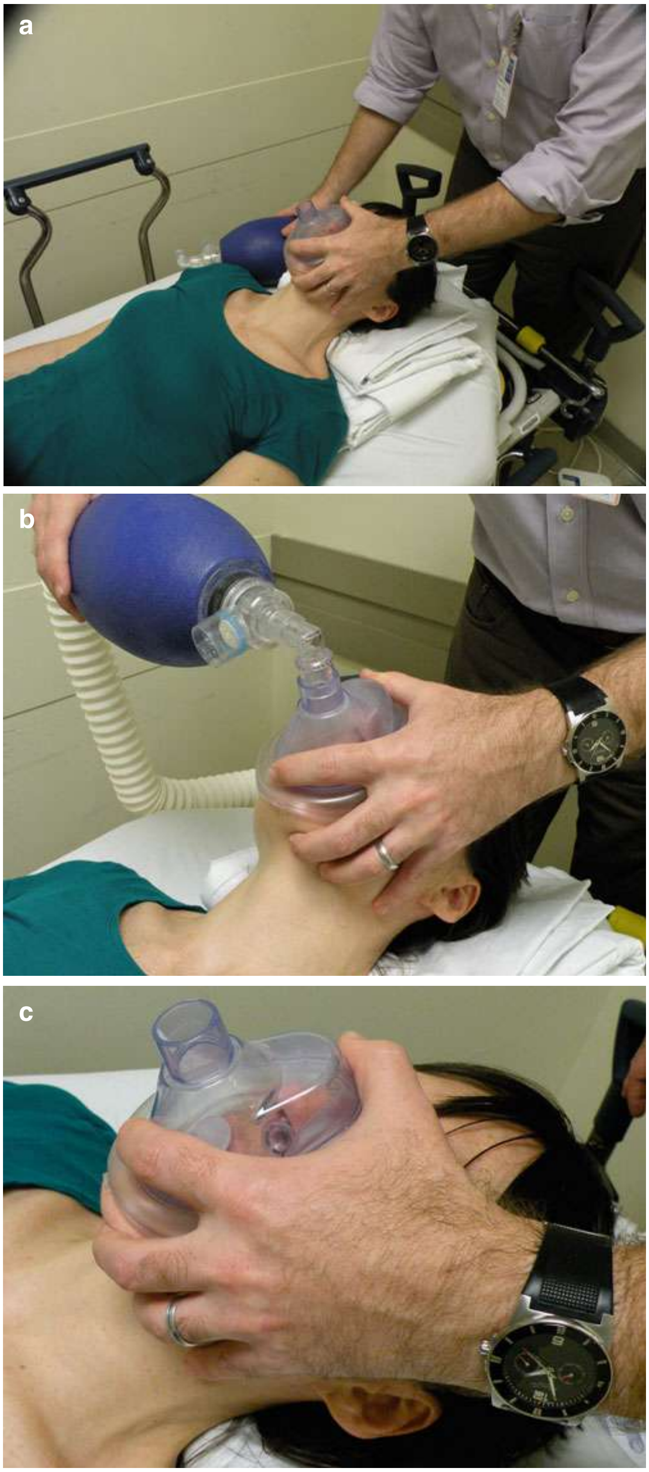

Fig. 7.4 (**a**–**c**) One-handed seal technique

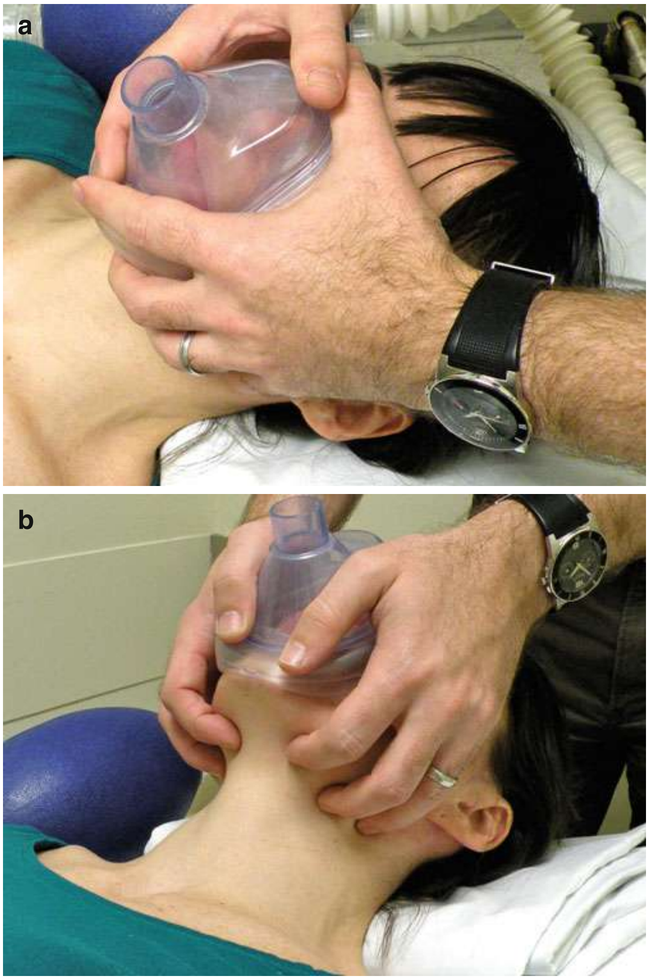

Fig. 7.5 Two-handed seal technique: (**a**) two semicircles, (**b**) alternative thumbs method

7.5 Complications

- Stomach inflation may lead to vomiting and aspiration.
- Increased positive thoracic pressure may cause decreased preload, worsening cardiac output, and/or hypotension.
- Hypoventilation (inadequate O_2 tidal volume, airway patency, or mask seal).

7.6 Pearls and Pitfalls

- Pearls
  - Use jaw thrust to open mouth for patients with possible cervical spine injury.
  - Use airway adjuncts whenever available, especially if prolonged BVM ventilation is anticipated.
  - Use lubricant jelly to insert nasal airway; do not insert in patients with severe facial trauma.
  - Mask should be placed on patient's face prior to attaching to bag.
  - Apply pressure to the bony part of the mandible only to avoid soft-tissue injury.
  - Provide just enough tidal volume to see a chest rise and deliver each breath gently over 1–1.5 s to prevent gastric insufflation.
  - Ensure good seal:
    - Select appropriate mask size.
    - Choose two-handed technique over one-handed, if possible.
    - Keep well-fitting dentures in place if present (and remove before intubation).
    - Lift the mandible toward the mask (as opposed to pushing the mask onto the face).
    - Rock the mask on face until no leak is present.
    - Apply K–Y jelly to beard to improve the seal.
- Pitfalls
  - Overcompression during the Sellick maneuver can compress the cricoid ring, preventing ventilation.
  - Cricoid pressure (Sellick maneuver) is not recommended during cardiac arrest resuscitation.
  - Difficult BVM ventilation: MOANS
    - *M*ask seal
    - *O*besity/obstruction
    - *A*ge
    - *N*o teeth
    - *S*tiff

Selected Reading

ECC Committee, Subcommittees and Task Forces of the American Heart Association. American heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation. 2010;122:S685–705.

Joffe AM, Hetzel S, Liew EC. A two-handed jaw-thrust technique is superior to the one-handed "EC-clamp" technique for mask ventilation in the apneic unconscious person. Anesthesiology. 2010;113:873–9.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. Philadelphia: Saunders Elsevier; 2010.

Walls RM, Murphy MF. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Awake Orotracheal Intubation

8

Benjamin M. Mahon, Justin Bennett, and Lars K. Beattie

8.1 Indications

- Urgent but not emergent endotracheal intubation is required in a patient who:
  - Is awake
  - Is currently protecting his airway
  - Is not a candidate for a supraglottic airway (LMA)
  - The patient
    - Is predicted to have a difficult airway
    - Has structural abnormalities of the airway
    - Will not tolerate a period of apnea
    - May lose his airway (anaphylaxis, angioedema, traumatic airway)
- Patients requiring urgent but not emergent intubation in whom paralytics are contraindicated (i.e., allergies, myasthenic crisis)

8.2 Contraindications

- Absolute
  - Surgical airway indicated
  - Emergent crash airway needed
  - Obtunded patient
  - Allergies to medications needed (lidocaine, glycopyrrolate)
- Relative
  - Inability to maintain airway or tolerate secretions

8.3 Materials and Medications

- Suctioning equipment
- Intravenous access equipment, cardiac monitor, pulse oximetry, blood pressure cuff
- 4 % lidocaine solution
- 2 % viscous lidocaine jelly
- Nebulizer
- Mucosal atomization device
- 4×4 gauze
- Tongue depressor
- Glycopyrrolate/atropine
- Sedation: ketamine, propofol, Versed, and/or fentanyl
- Intubation equipment
- Backup emergency airway adjuncts
- Bag valve mask
- Laryngoscope, fiber optics, oral airway, etc.

8.4 Procedure

1. Preparation
   (a) Establish IV access.
   (b) Place the patient on a cardiac monitor with continuous pulse oximetry.
   (c) Keep backup RSI emergency airway medication and equipment at the bedside.
2. Administer 0.2–0.4 mg of intravenous glycopyrrolate (or a small dose (0.5–1 mg) of atropine to decrease secretions) 15 min prior to procedure.
3. Nebulize 2 mL of 4 % lidocaine with oxygen at 5 l O2 per minute to anesthetize the pharynx (Fig. 8.1).

B.M. Mahon, MD
Ponciana Medical Center,
Kissimmee, FL, USA
e-mail: drfaust2k9@aol.com

J. Bennett, MD
Department of Emergency Medicine, Wake Forest University
School of Medicine, Winston-Salem, NC, USA
e-mail: jubennet@wakehealth.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_8

4. Use Yankauer suction (with the patient's assistance) to dry out the mouth as much as possible. Dabbing the tongue with gauze can assist in this step.
5. Continue preoxygenation.
6. Immediately after the nebulized solution is applied, give the patient a "lidocaine lollipop" (Fig. 8.2).
    (a) A 2 ml dollop of 2 % viscous lidocaine is to be placed on the end of a tongue depressor and is given to the patient to place in his mouth (like a lollipop).
    (b) Have the patient copiously gargle, then swallow the viscous lidocaine.
7. Using a mucosal atomizer, spray 2 ml of 4 % lidocaine in the posterior oropharynx and as far down toward the glottis as possible (Figs. 8.3, 8.4, and 8.5).
8. Sedation
    (a) It is feasible to proceed with the awake intubation in an un-sedated, wide awake but cooperative patient.
    (b) Sedation can be initiated using institutional preferences, but some options include:
        (i) Midazolam 2 mg IV
        (ii) Ketamine 1 mg/kg IV
        (iii) Propofol 1 mg/kg IV
        (iv) Ketofol (ketamine and propofol both at concentrations of 10 mg/ml, 5 ml of each mixed in a 10 cc syringe) titrated at 1–3 ml aliquots
    (c) More atomized lidocaine can be provided prior to endotracheal tube (ETT) passage, but one must be aware of the upper lidocaine dose for your patient.
    (d) Adequate anesthesia is confirmed by the absence of a gag reflex upon direct palpation (Fig. 8.6).
9. Intubation, induction, and gentle direct laryngoscopy can be performed at this point to place the ETT (Fig. 8.7).
    (a) Induction (if no prior sedation) and paralytic agents should be available to immediately administer after ETT placement.
    (b) Thorough discussions on intubation techniques can be found in other chapters in the atlas.

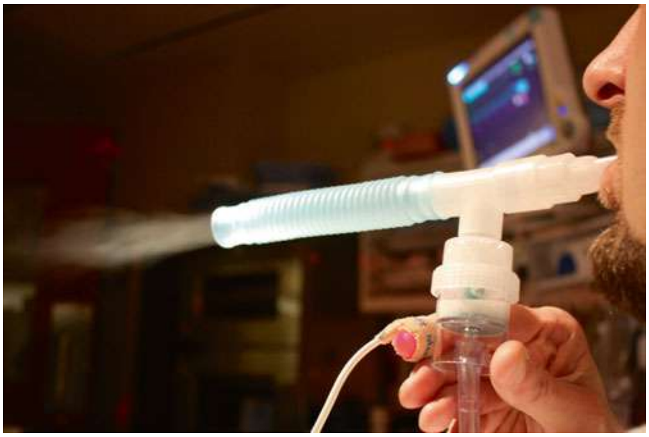

Fig. 8.1 Nebulization of 4 % lidocaine

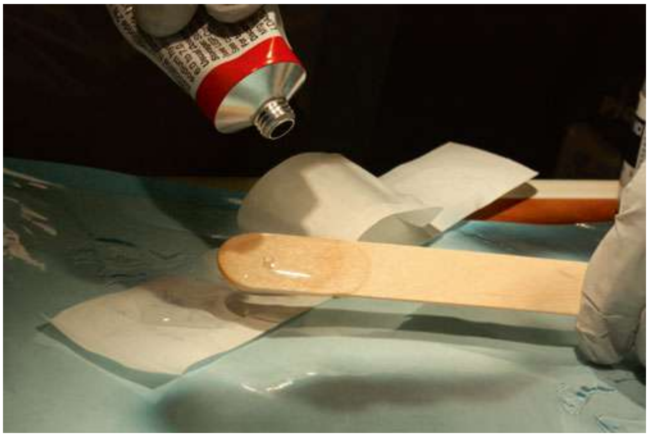

Fig. 8.2 Lidocaine lollipop

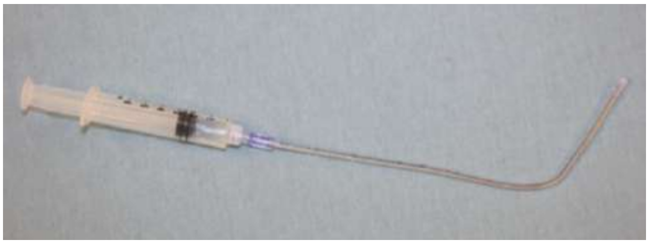

Fig. 8.3 Mucosal atomizer attached to syringe

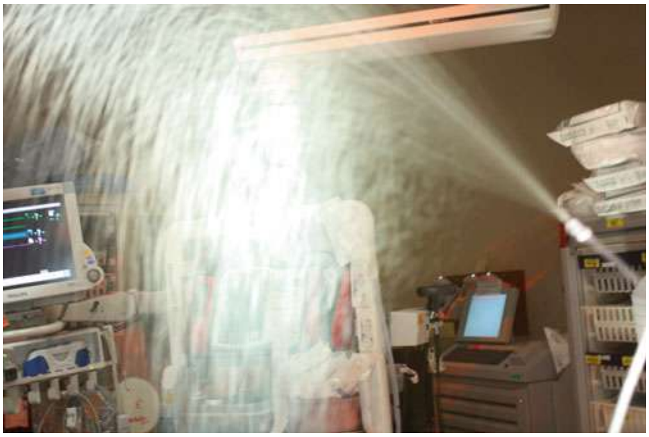

Fig. 8.4 Pushing syringe plunger atomizes lidocaine

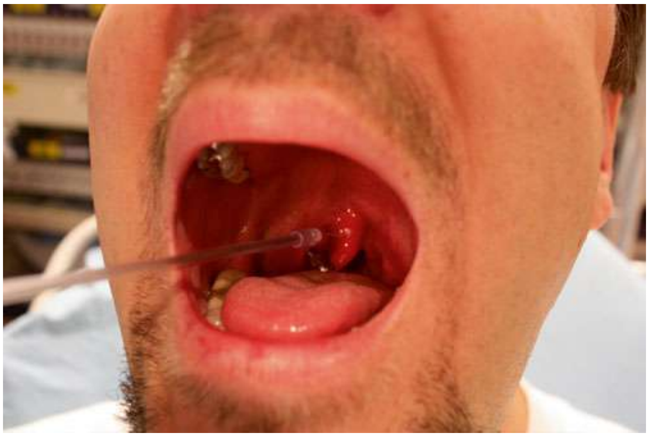

Fig. 8.5 Atomized lidocaine being administered to the posterior pharynx

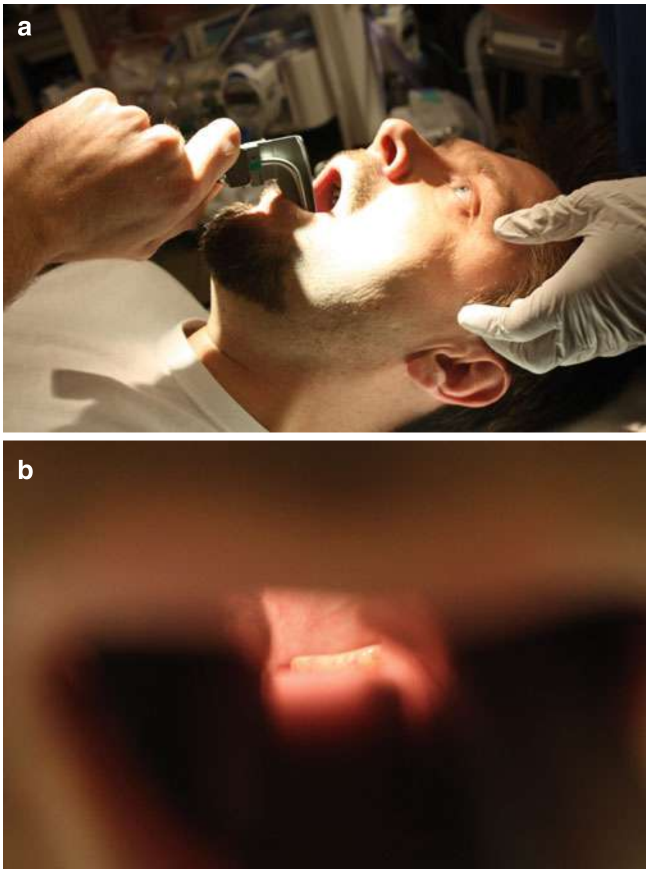

Fig. 8.6 (**a**, **b**) Adequately anesthetized awake patient with laryngoscopic view of epiglottis

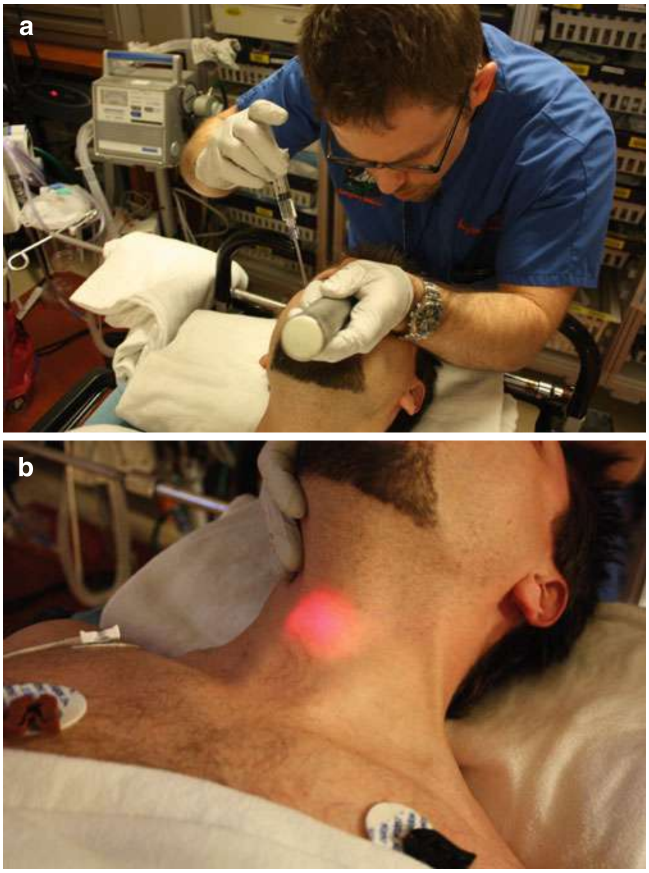

Fig. 8.7 (**a**) Final lidocaine atomization of deep structures and trachea, (**b**) intubation using a lighted stylet

8.5 Pearls and Pitfalls

- Pearls
  - Steps 3–7 should be done successively and as quickly as possible, to capitalize on the short half-life of lidocaine.
  - Simultaneous preoxygenation and anesthetization can be achieved by nebulizing the 4 % lidocaine through a face mask.
  - If the patient has been sedated, soft restraints may help prevent the patient from inadvertently grabbing the tube or your intubating equipment.
  - A nasotracheal intubation can be performed by simply anesthetizing the nares through which the ETT will be placed using lidocaine jelly and/or atomized lidocaine.
- Pitfalls
  - Failure to prepare all equipment beforehand may extend the procedure beyond the lidocaine half-life.
  - The toxic dose of lidocaine is 300 mg or 3–5 mg/kg. The dosages listed are intended for a 70 kg patient and yields a total dose of 280 mg (4 mg/kg). This dose may need to be decreased in smaller individuals.

Selected Reading

Agro F, Hung OR, Cataldo R, Carassiti M, Gherardi S. Lightwand intubation using the Trachlight: a brief review of current knowledge. Can J Anaesth. 2001;48(6):592–9.

Rhee KY, Lee JR, Kim J, Park S, Kwon WK, Han S. A comparison of lighted stylet (Surch-Lite) and direct laryngoscopic intubation in patients with high Mallampati scores. Anesth Analg. 2009;108(4): 1215–92.

Walls RM, Murphy MF. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams and Wilkins, a Wolters Kluwer Business; 2008. Chap 11.

Rapid-Sequence Intubation

9

Ram A. Parekh

9.1 Indications

- Failure to oxygenate
- Failure to ventilate
- Unable to protect airway patency or reflexes
- Projected clinical course deterioration

9.2 Contraindications

- Absolute
  - Complete upper airway obstruction
  - Significant facial and airway trauma with loss of landmarks for orotracheal intubation
- Relative
  - Anticipated difficult intubation
    - Not an absolute contraindication.
    - Patient scenario requires a careful preintubation assessment and plan.
    - Consider an "awake" intubation.
    - Consider alternative airway adjuncts (e.g., extraglottic devices, video laryngoscopy, laryngeal mask airway [LMA]).
  - Induction or paralytic agent-specific contraindications given clinical circumstances
    - Caution: induction agents that lower blood pressure in hypotensive patients
    - Caution: succinylcholine in potentially hyperkalemic patients
  - Crash airway
    - Apneic, arrest, and periarrest situation

9.3 Materials and Medications

- Laryngoscope with appropriate blade (choice based on proceduralist's preference and patient anatomy) (Fig. 9.1)
- Intubating stylet
- Endotracheal tubes (ETTs)
- Syringe, 10 mL (to inflate ETT cuff)
- Surgilube
- Suction catheter
- Oral and nasal airways (Fig. 9.2)
- Ambu bag and mask attached to oxygen source
- Induction, pretreatment, and paralytic agents
- ETT confirmation device—EZ capnometry, quantitative end-tidal carbon dioxide concentration ($EtCO_2$) detection

---

R.A. Parekh, MD
Department of Emergency Medicine,
Icahn School of Medicine at Mount Sinai,
Elmhurst Hospital Center, Elmhurst, NY, USA
e-mail: ram.parekh@mssm.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_9

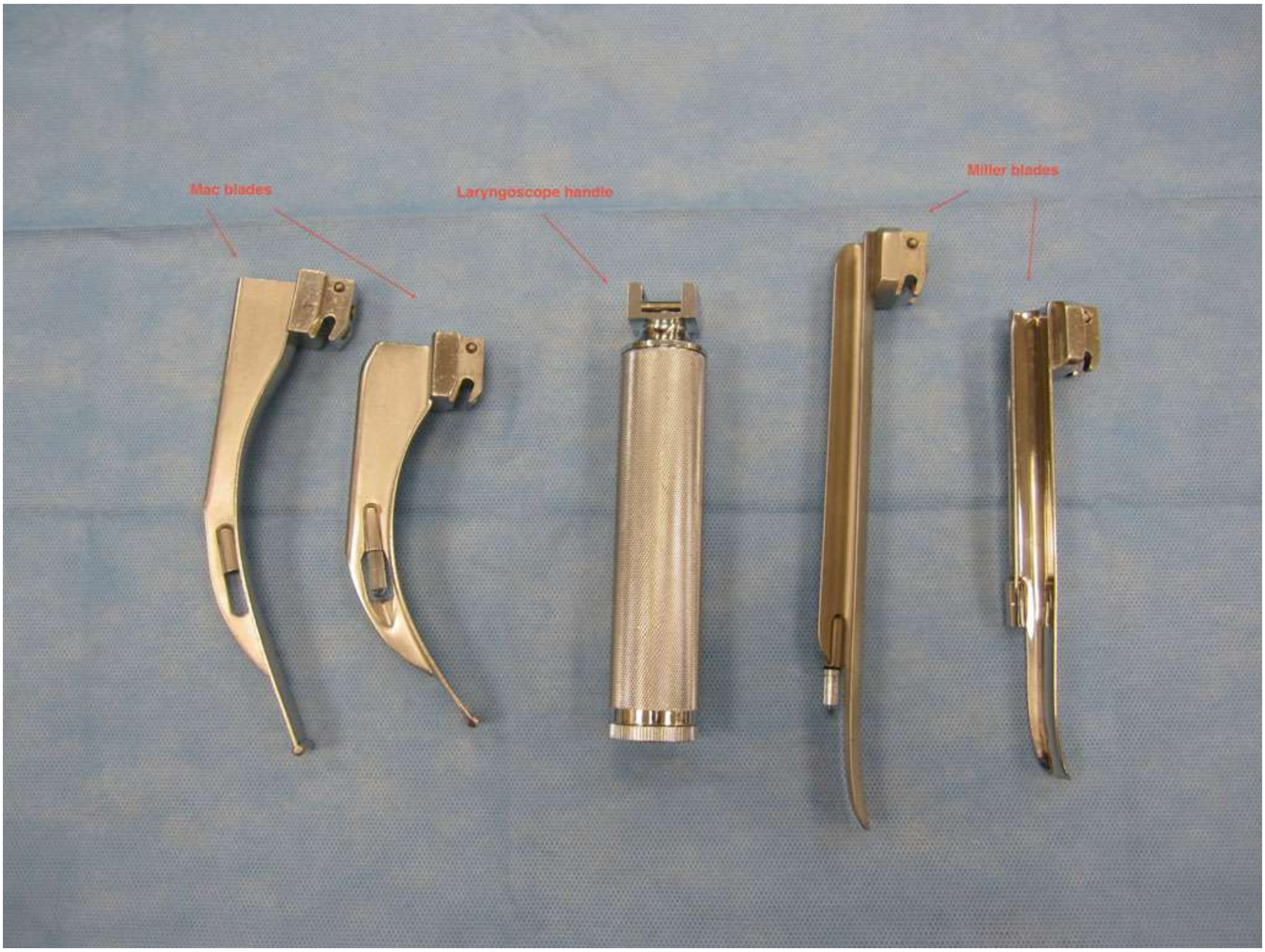
Mac blades
Laryngoscope handle
Miller blades

Fig. 9.1 Laryngoscope and blades

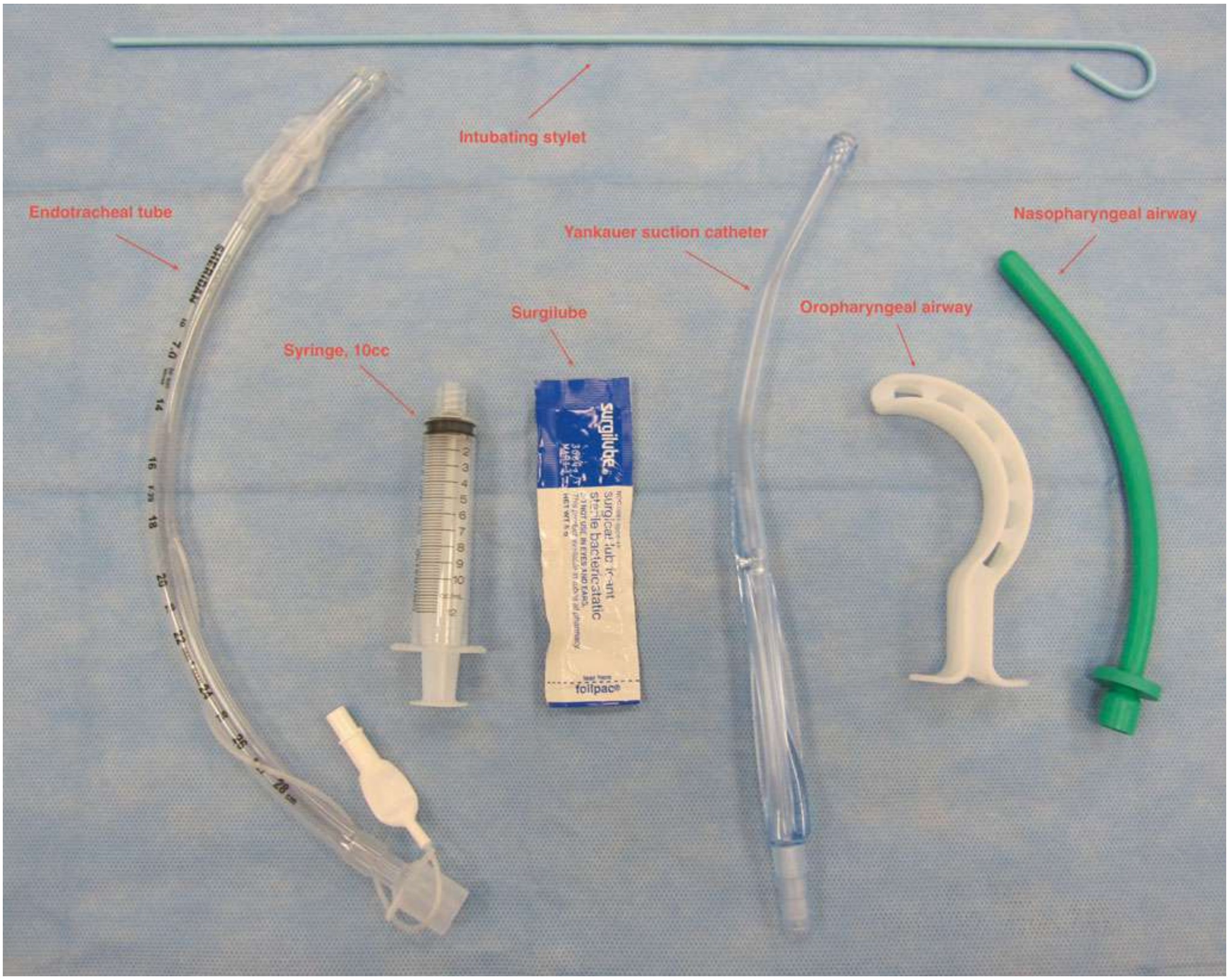
Intubating stylet
Endotracheal tube
Yankauer suction catheter
Nasopharyngeal airway
Surgilube
Oropharyngeal airway
Syringe, 10cc

Fig. 9.2 Oral and nasal airways

9.4 Procedure

1. Preparation
   - Thoroughly assess patient for difficulty of intubation (Fig. 9.3).
   - Develop fallback plans for failed intubation attempt.
   - Establish at least one, but preferably two, secure intravenous (IV) lines.
   - Place on cardiac monitor with pulse oximetry, blood pressure monitoring, and continuous capnography.
   - Yankauer suction device attached to suction, suction on.
   - Pharmacological agents, drawn and labeled.
   - Laryngoscope and blades tested to ensure proper functioning of light source.
   - Desired ETT size, prepared for intubation:
     - Intubating stylet in position (tip at eye of ETT)
     - 10-mL syringe attached to ETT
     - ETT configured per proceduralist's preference (e.g., hockey stick, curved)
     - Cuff tested for air leak
2. Preoxygenation
   - Administer high-flow oxygen for 3–5 min (Fig. 9.4).
     - Nitrogen is exchanged for O_2 in the functional residual capacity of the lungs.
     - Establishes oxygen reservoir within lungs (primarily), blood, and body tissue.
     - Also known as nitrogen "washout."
   - This can be done using:
     - Non-rebreather masks—delivers 65–70 %
       - Difficult intubation not anticipated
     - Well-fitting bag-valve-masks (without positive-pressure ventilation)—delivers greater than 90 % oxygen (Fig. 9.5)
     - Noninvasive positive-pressure ventilation (NIPPV)—delivers 100 % oxygen
       - Consider NIPPV in high-risk patients with moderate to severe shunt physiology.
3. Pretreatment
   - Administer pharmacological agents to mitigate adverse physiological effects of intubation, induction, and paralysis, which may be undesirable in certain clinical circumstances (Table 9.1).
4. Induction and paralysis
   - Administer a rapidly acting induction agent to produce rapid loss of consciousness via IV push (Table 9.2).
   - Immediately follow induction agent with a neuromuscular-blocking agent via IV push (Table 9.3).
5. Positioning
   - If no cervical spine injury suspected, place the patient in the "sniff" position (Fig. 9.6):
     - Flex neck
     - Extend head
   - Ideally, the patient's pinna will be at the level of the sternum.
6. Direct laryngoscopy (see Chap. 10)
7. Proof of placement
   - Visualize ETT passing vocal cords
   - Confirm tube placement via $EtCO_2$:
     - Qualitative detection device—EZ Cap
     - Quantitative continuous $EtCO_2$ waveform on monitor (preferred) (Fig. 9.7)
   - Auscultation of breath sounds:
     - Lung fields bilaterally
     - Epigastric region (ensuring no breath sounds in the stomach)
8. Postintubation management
   - Secure ETT (Fig. 9.8)
   - Initiate mechanical ventilation.
   - Postintubation sedation and analgesia.
   - Postintubation chest x-ray.

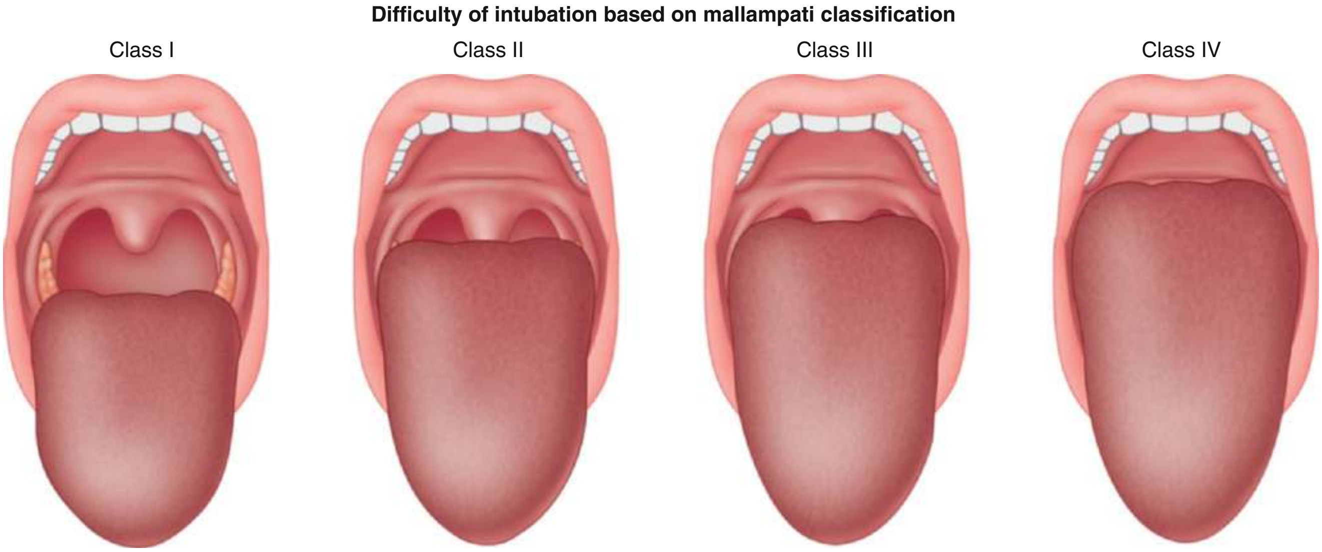
Difficulty of intubation based on mallampati classification
Class I
Class II
Class III
Class IV

Fig. 9.3 Assess patient for difficulty of intubation

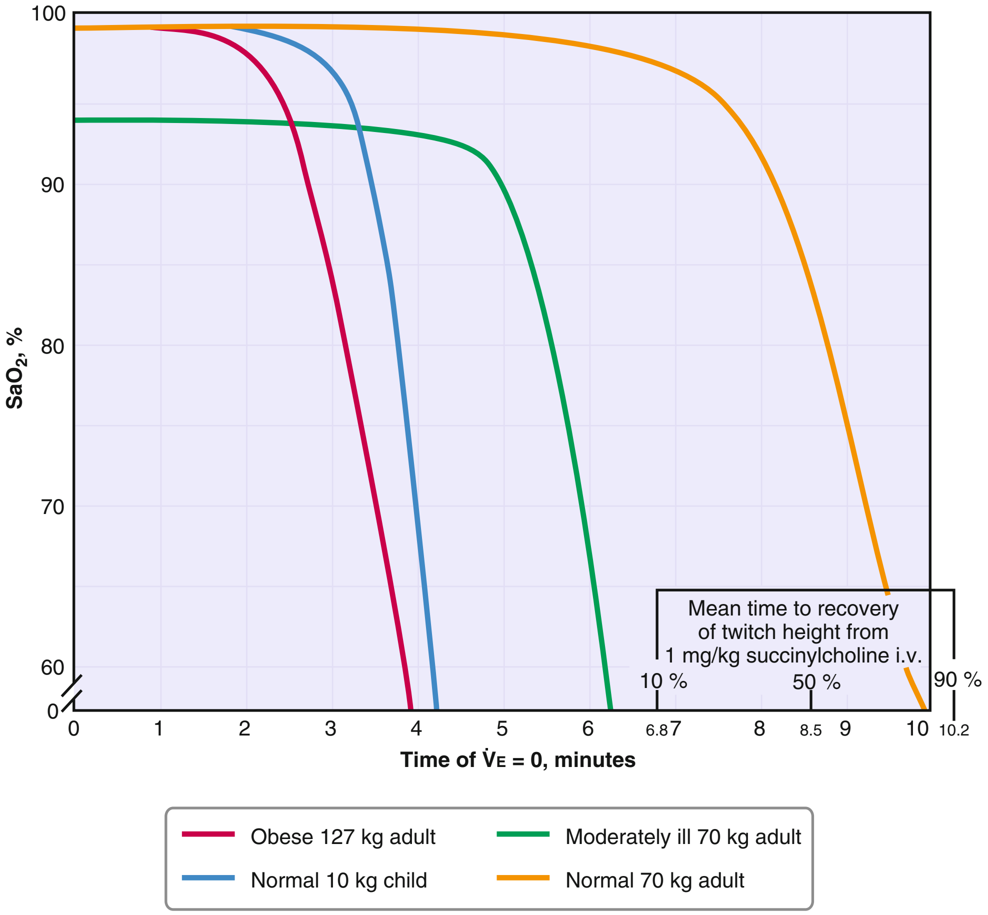

Fig. 9.4 *FaO2* fractional concentration of alveolar oxygen, *SaO2* arterial oxygen saturation, *Ve* expired volume per minute

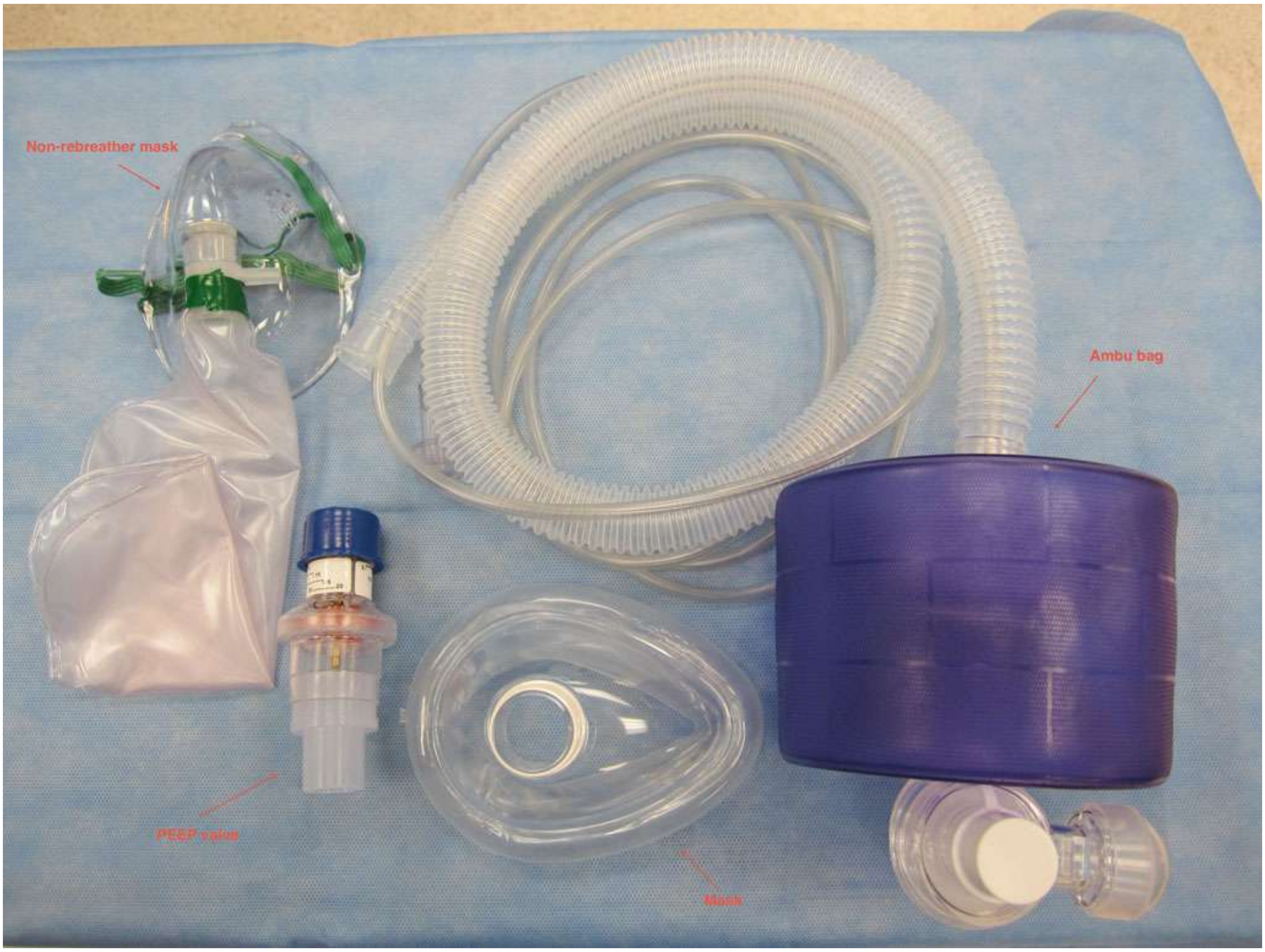

Fig. 9.5 Preoxygenation materials

Table 9.1 Pretreatment: pharmacological agents used to mitigate adverse physiological effects of intubation, induction, and paralysis

Agent	Dose (IV)	Indication
Lidocaine	1.5 mg/kg Rapid push	Use in tight brain to attenuate ICP increase from laryngoscopy/intubation; use in tight lungs to blunt bronchospastic response
Fentanyl	1–3 mcg/kg Slow push	Use in tight brain, tight heart, and tight vessels to blunt reflex sympathetic response to laryngoscopy

ICP intracranial pressure

Table 9.2 Induction: rapidly acting induction agents used to produce rapid loss of consciousness

Agent	Dose (IV) (mg/kg)	Onset (sec)	Duration (min)
Midazolam	0.2–0.3	60–90	15–30
Etomidate	0.3	15–45	3–12
Thiopental	3	<30	5–10
Ketamine	1.5–2.0	45–60	10–20
Propofol	1.5	15–45	5–10

Table 9.3 Paralysis: neuromuscular-blocking agents administered immediately after induction agent

Agent	Dose (IV)	Onset (sec)	Duration (min)
Succinylcholine	1.5 mg/kg	45	6–10
Rocuronium	1.0 mg/kg	60–75	40–60
Vecuronium	0.01 to prime, then 0.15 mg/kg	75–90	60–75

Fig. 9.6 Patient in the "sniff" position. *OA* oral axis, *LA* laryngeal axis, *PA* pharyngeal axis

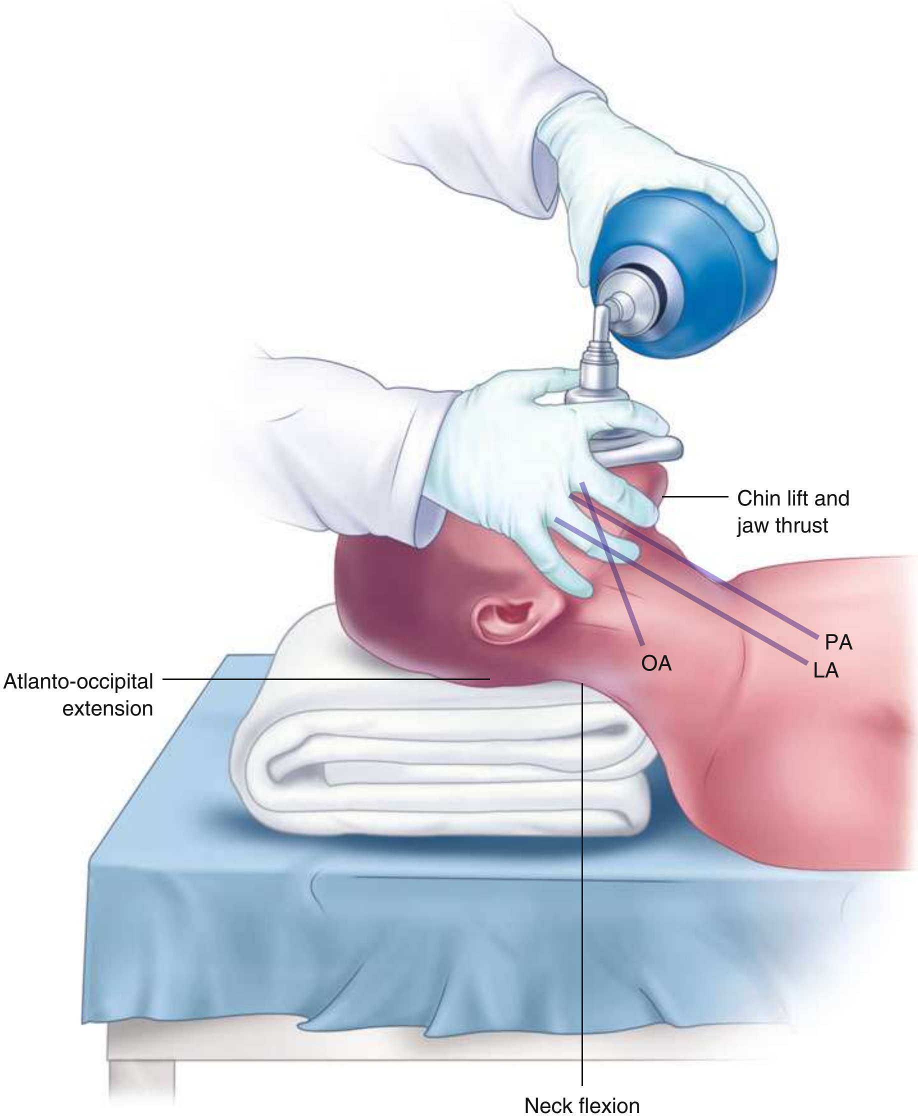
Chin lift and jaw thrust
PA
OA
LA
Atlanto-occipital extension
Neck flexion

Fig. 9.7 $EtCO_2$, end-tidal carbon dioxide concentration: *a* Qualitative detection device—EZ Cap, *b* quantitative continuous $EtCO_2$ waveform on monitor

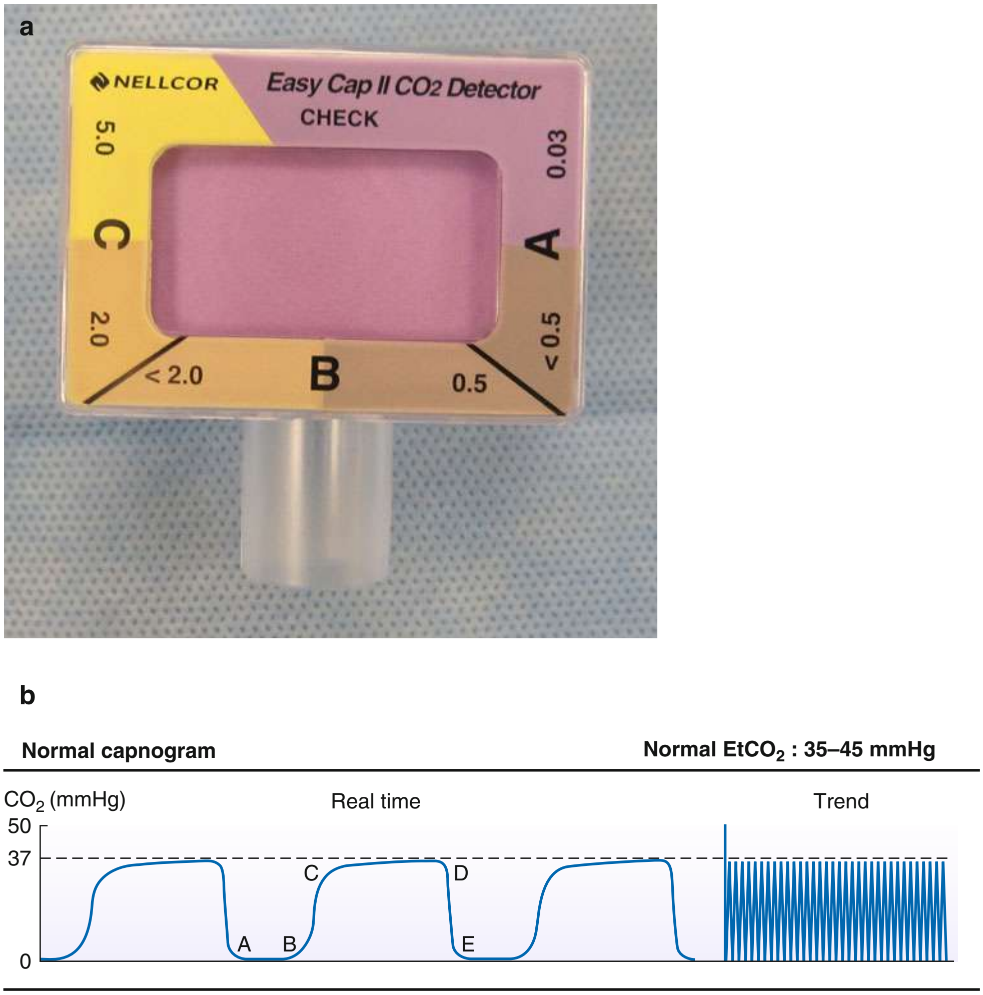

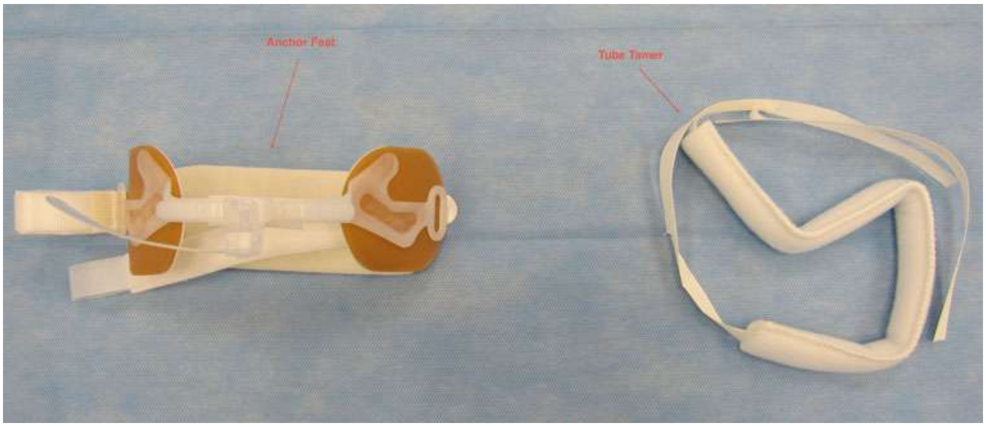
Anchor Fast
Tube Tamer

Fig. 9.8 Equipment used to secure ETT

9.5 Complications

- Esophageal intubation
- Right mainstem intubation
- Pneumothorax from laryngeal trauma
- Aspiration
- Dental trauma
- Vocal cord injury
- Hypotension
  - Induction agent
  - Decreased venous return from positive pressure
  - Pneumothorax
- Hyperkalemia (succinylcholine used in mildly hyperkalemic patient)
- Iatrogenically obstructed airway
- Failure to intubate

9.6 Pearls and Pitfalls

- Utilization of oral and nasal airways will greatly increase the ease of preoxygenation and reoxygenation.
- Inadequate preoxygenation will cause premature desaturation, decreasing time for intubation.
- Suboptimal patient positioning can prevent vocal cord visualization during direct laryngoscopy.
- Consider alternative airway devices before intubation to have a plan in place if a difficult airway is encountered.
- Underdosing induction or paralytic agent will prevent adequate time to perform procedure or create patient discomfort.
- Inappropriately sized McIntosh laryngoscopic blades.
  - Too small—more difficulty in moving tongue and epiglottis out of way for vocal cord visualization
  - Too large—easier to overshoot and go past vocal cords into the esophagus
- Excessive cricoid pressure (Sellick maneuver) may lead to disrupted laryngoscopic view and difficulty passing the ETT.
  - Disrupted view: readjust larynx using dominant hand to allow cord visualization.
  - Difficulty passing ETT: ask for release of some cricoid pressure to allow for ETT passage.
- Inadequate postintubation sedation and analgesia, especially when long-acting paralytics are used.
- Acidic gastric contents can cause CO_2 qualitative capnometry to change to yellow, falsely indicating tracheal placement of the ETT.

Selected Reading

Walls R, Murphy M. Manual of emergency airway management. Philadelphia: Lippincott Williams & Wilkins; 2008.

Weingart SD. Preoxygenation, reoxygenation, and delayed sequence intubation. J Emerg Med. 2011;40:661–7.

Direct Laryngoscopy

10

Bharath Chakravarthy and Weston Seipp

10.1 Indications

- Orotracheal intubation
  - Maintenance of oxygenation/ventilation
  - Airway protection
- Visualization of laryngeal anatomy
- Foreign body retrieval

10.2 Contraindications

- Absolute
  - None
- Relative
  - Presumed difficult airway
    - Anatomical limitations
      - Small oral opening (less than three of the patient's fingers)
      - Small mandible (hyomental distance less than three fingers)
      - Hyoid-thyroid distance (less than two fingers)
    - Clinical limitations
      - Patient with unstable cervical spine
      - Patient with multiple facial or neck trauma
      - Patient with history of tracheal stenosis, irradiation, or history of tracheal mass or surgery

---

B. Chakravarthy, MD, MPH (✉)
Department of Emergency Medicine,
University of California Irvine,
Orange, CA, USA
e-mail: bchakrav@uci.edu

W. Seipp, MD
Department of Emergency Medicine,
University of California Irvine Medical Center,
Orange, CA, USA

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_10

10.3 Materials and Medications (Fig. 10.1)

- Laryngoscope handle
- Laryngoscope blade with light
  - Macintosh blade ("Mac" or "curved blade")
  - Miller blade ("straight blade")
- Bag valve mask attached to 100 % O_2 source
- Endotracheal tube (ETT)
- 10-mL syringe
- Yankauer suction
- End-tidal CO_2 ($EtCO_2$) monitor (colorimetric or quantitative)
- McGill forceps (for foreign body retrieval)
- Postintubation chest radiograph

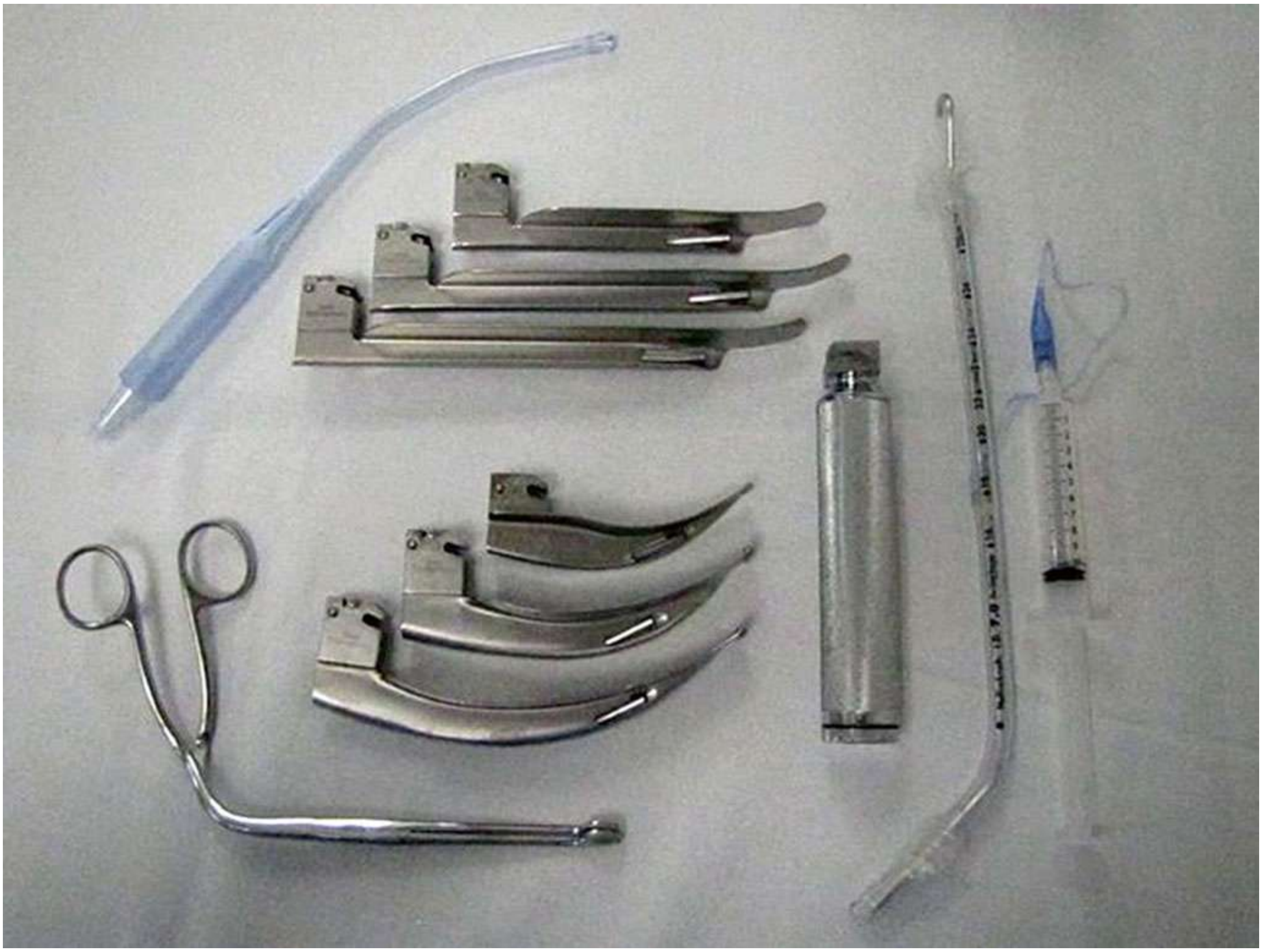

Fig. 10.1 Basic laryngoscopy supplies. *Clockwise from top left*: Yankauer suction, Miller Blades, endotracheal tube, 10-mL syringe, laryngoscope handle, Macintosh blades, and McGill forceps

10.4 Procedure

1. Check the laryngoscope handle and blades to ensure that the light is functioning.
2. Choose the appropriate blade based on patient size.
    (a) The Macintosh or Miller 3 size is appropriate for the majority of adults, and a 4 can be used for a larger body habitus.
    (b) The Macintosh blade is generally preferred in adults owing to increased space for ETT passage [1].
    (c) The Miller blade may be preferable in cases in which the patient has limited mouth opening (owing to its smaller vertical height), or in which the airway is particularly floppy (such as in infants and children) [1].
3. Position the patient (Fig. 10.2).
    (a) Raise the bed so that the patient's oral opening is at the level of the xiphoid process of the laryngoscopist.
    (b) The optimal laryngeal view is obtained in the neck flexion/head extension or "sniffing" position. To achieve this, place towels under the patient's occiput to raise it approximately 6–9 cm [2].
    (c) If patient is in cervical spine precautions, then an assistant must hold the cervical spine in midline immobilization throughout the laryngoscopy; elevation of the occiput is therefore contraindicated.
4. Provide 100 % O_2 via face mask to preoxygenate the patient before laryngoscopy.
5. After ensuring adequate anesthesia and neuromuscular blockade (if performing laryngoscopy for intubation), perform the scissor technique to open the patient's mouth and to lift the tongue base from the glottic opening.
6. Macintosh blade insertion (Fig. 10.3).
    (a) Insert the laryngoscope into the patient's mouth, starting from the right side, and slowly advance into the oropharynx, using the blade's vertical flange to "sweep" the tongue to the left and away from the glottic opening.
    (b) Advance the blade along the tongue toward the tongue base, until the epiglottis and posterior arytenoids are viewed.
    (c) In order to expose the cords, insert the Macintosh blade into the vallecula, which is the potential space anterior to the epiglottis and posterior to the tongue base. This will act as a fulcrum and raise the epiglottis, exposing the vocal cords.
    (d) To further expose the cords and/or expose the vallecula, exert force outward at a 45° angle to the patient. Do not "rock" the laryngoscope because this may cause injury to the teeth.
7. Miller blade insertion (Fig. 10.4).
    (a) Insert the blade into the right side of the mouth and slowly advance along tongue toward the tongue base. The Miller blade does not have a flange for isolating the tongue, and thus the Macintosh may be preferable in patients with large tongues.
    (b) Advance the blade along the right side of the tongue until the epiglottis and posterior arytenoids are visible.
    (c) In contrast to the Macintosh blade, the Miller blade is used to directly isolate the epiglottis and expose the vocal cords. Using the tip of the Miller, move the epiglottis anteriorly to expose the vocal cords.
    (d) As with the Macintosh, exert force outward at a 45° angle to the patient to increase the view of vocal cords. Do not "rock" the laryngoscope because this may cause injury to the teeth.
8. Assessing the glottic view (Fig. 10.5).
    (a) With the epiglottis either directly or indirectly lifted from the glottic opening, assess the Cormack-Lehane laryngeal view grade.
        (i) Grade I—view of entire laryngeal opening, including cords
        (ii) Grade II—view of posterior laryngeal cartilages
        (iii) Grade III—visualization of epiglottis only
        (iv) Grade IV—no structures visualized
    (b) A lower grade (higher quality, better) view is predictive of intubation success [3].
    (c) In the event of a higher grade view, the operator may request airway adjuncts, such as a bougie.
9. Improving the glottic view [4].
    (a) With the laryngoscope in the desired position, it is possible to improve the glottic view by exerting backward pressure on the thyroid cartilage either with the operator's right hand (bimanual laryngoscopy) or with an assistant applying BURP (backward-upward-rightward pressure) (Fig. 10.5a).
    (b) Backward pressure increases the vertical distance between the epiglottis and the posterior cartilages, thereby increasing the likelihood of vocal cord visualization.
10. Passing the ETT (Fig. 10.6).
    (a) With the optimal view of the cords obtained, pass the ETT from the right corner of the mouth through the vocal cords, to a depth of 21 cm at the incisors in females and 23 cm in males.
    (b) Inflate the ETT cuff with approximately 5 cc of air until the pilot balloon is firm to touch.
11. After completing intubation or after completion of laryngoscopy, slowly remove the blade from the mouth, taking care to avoid dental or lip trauma.
12. Attach the capnography device to the ETT tube to ensure $EtCO_2$ return.
    (a) Colorimetric devices will turn from purple to yellow in the presence of $EtCO_2$.
    (b) Quantitative devices will return a CO_2 waveform.
13. Attach the ETT to ventilator or bag-valve-mask connected to an oxygen source.

14. Auscultate breath sounds in both lung fields and ensure absence of breath sounds over the epigastrium (which could signify esophageal intubation).
15. Obtain a postintubation chest x-ray to ensure no right mainstem intubation or pneumothorax.

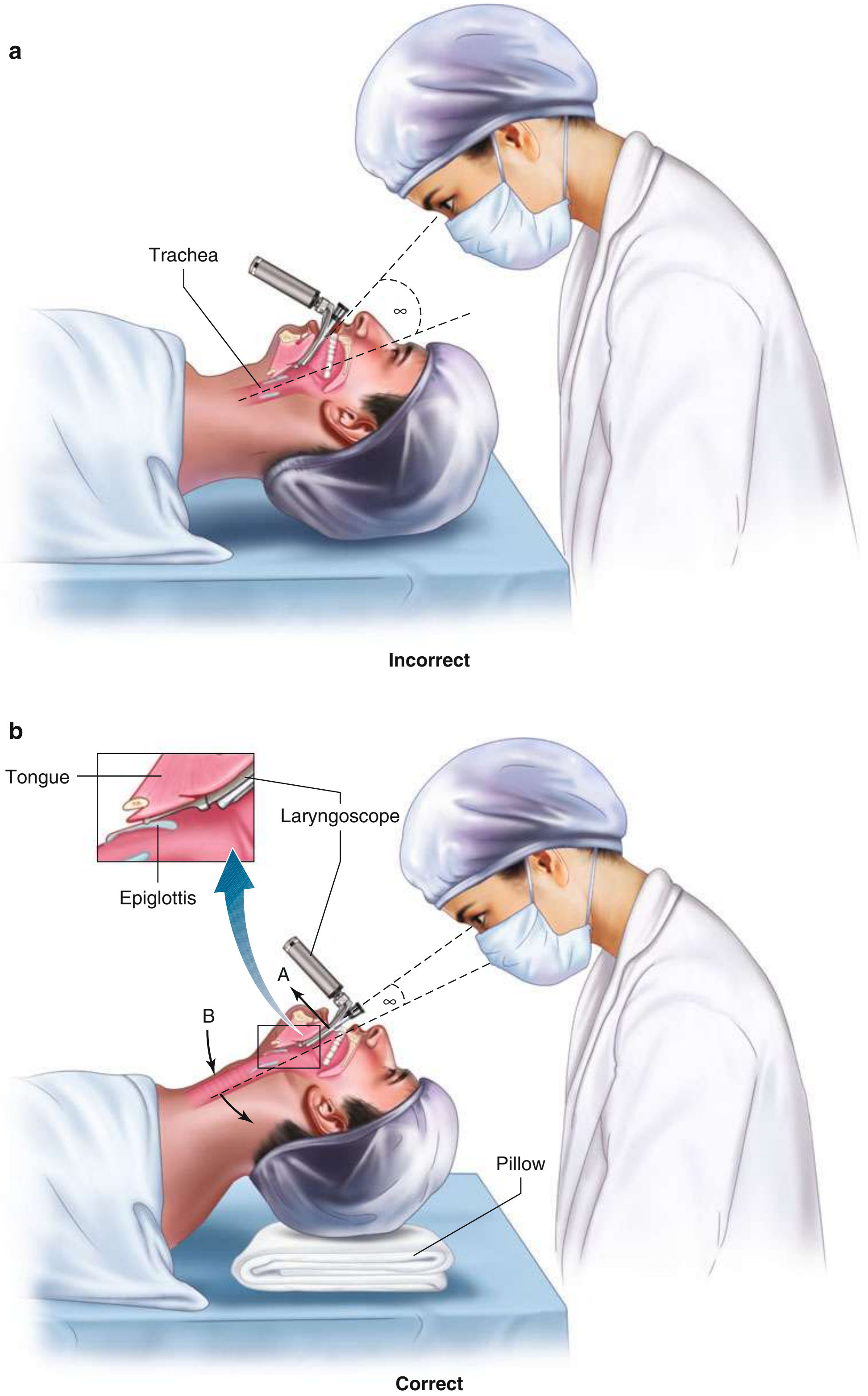
a
Trachea
∞
Incorrect
b
Tongue
Laryngoscope
Epiglottis
A
B
∞
Pillow
Correct

Fig. 10.2 Visualization axis and sniffing position: (**a**) The patient's occiput is not elevated and the neck is not in flexion, thereby creating a steep visual axis, (**b**) the occiput is correctly elevated 6–9 cm, placing the patient in sniffing position and allowing the visual axis to align with the airway axis

Fig. 10.3 Macintosh blade insertion. The blade is inserted into the vallecula, which raises the epiglottis and exposes the glottic opening

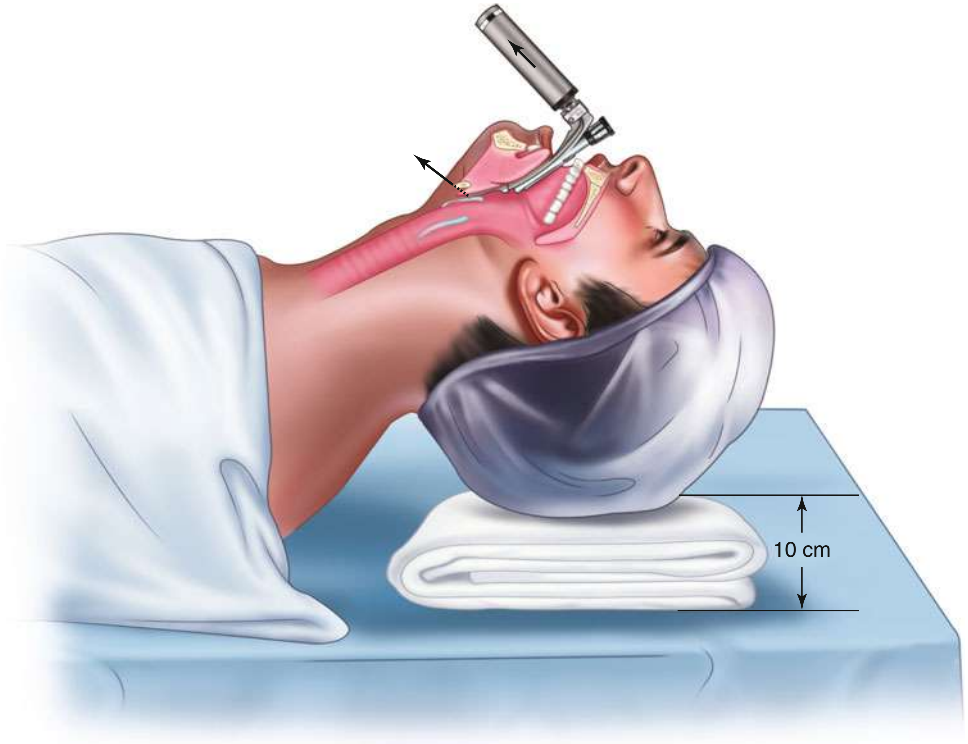

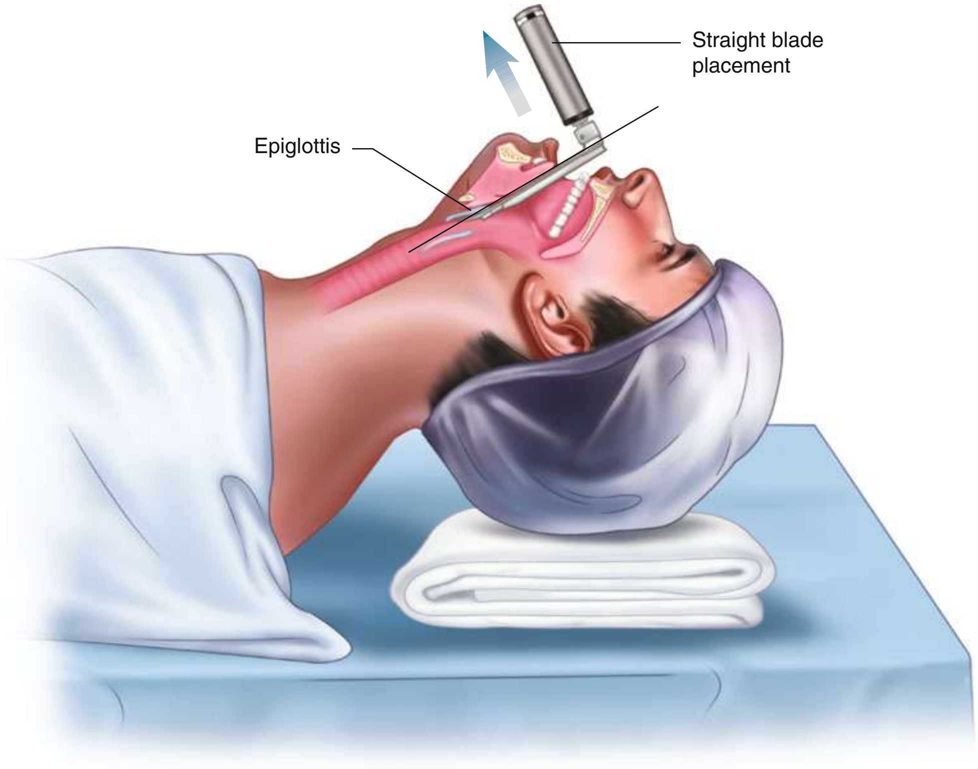
Straight blade placement
Epiglottis

Fig. 10.4 Miller blade insertion. The blade is used to elevate the epiglottis directly, exposing the glottic opening

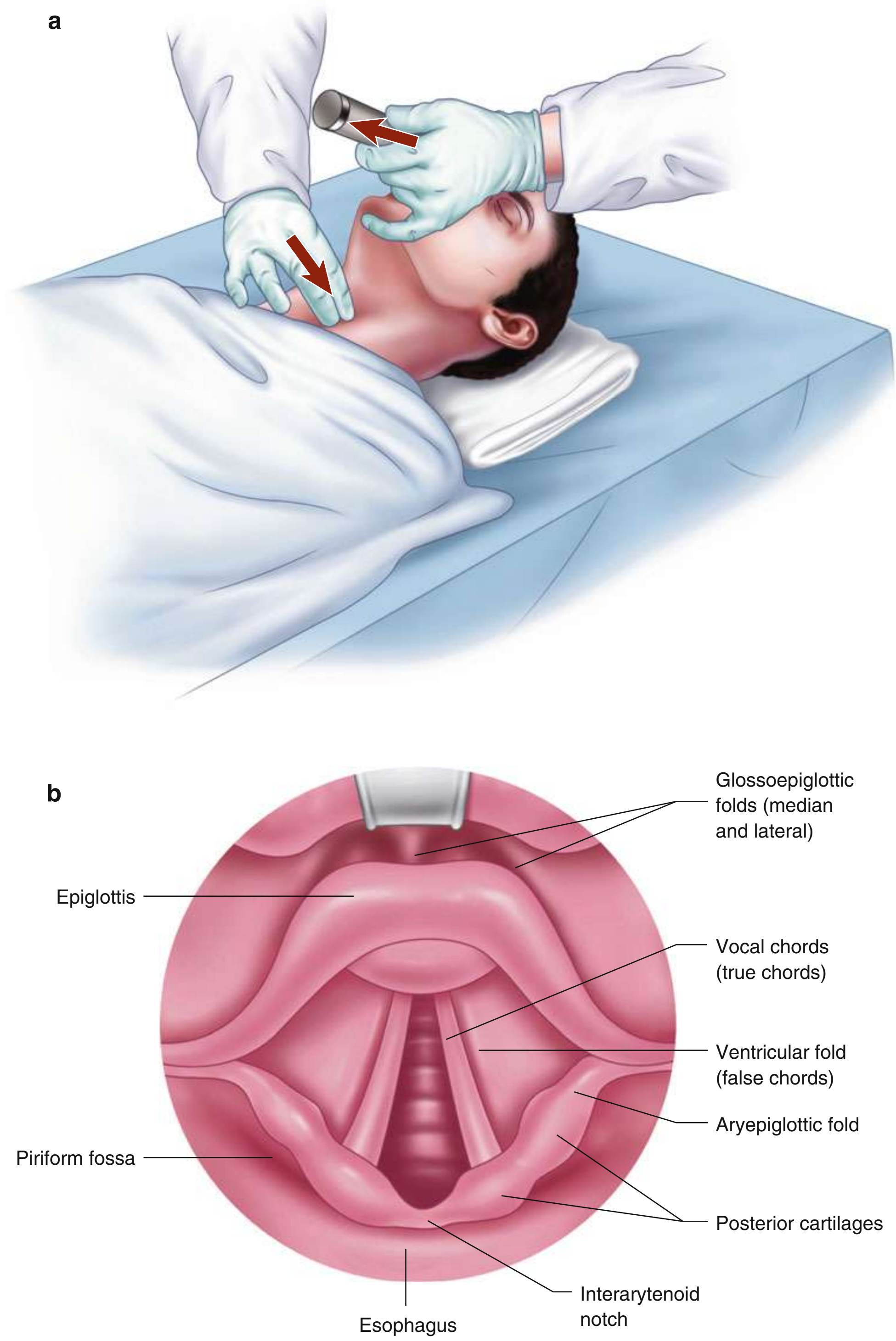
a
b
Glossoepiglottic folds (median and lateral)
Epiglottis
Vocal chords (true chords)
Ventricular fold (false chords)
Aryepiglottic fold
Piriform fossa
Posterior cartilages
Interarytenoid notch
Esophagus

Fig. 10.5 (**a**) Bimanual laryngoscopy. The force on the neck is opposite the direction of lift by the laryngoscope, (**b**) laryngoscopy view, (**c**) Cormack and Lehane

Fig. 10.5 (continued)

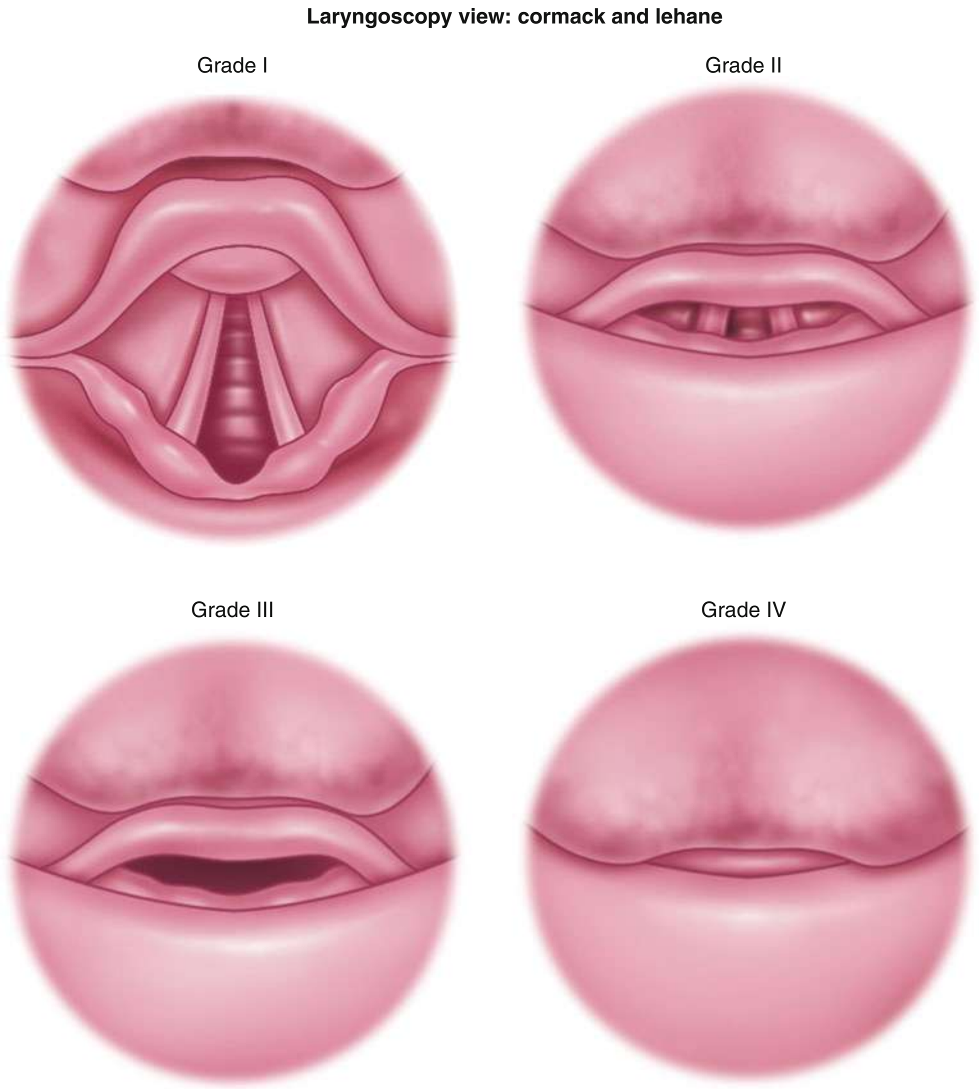
Laryngoscopy view: cormack and lehane
Grade I
Grade II
Grade III
Grade IV

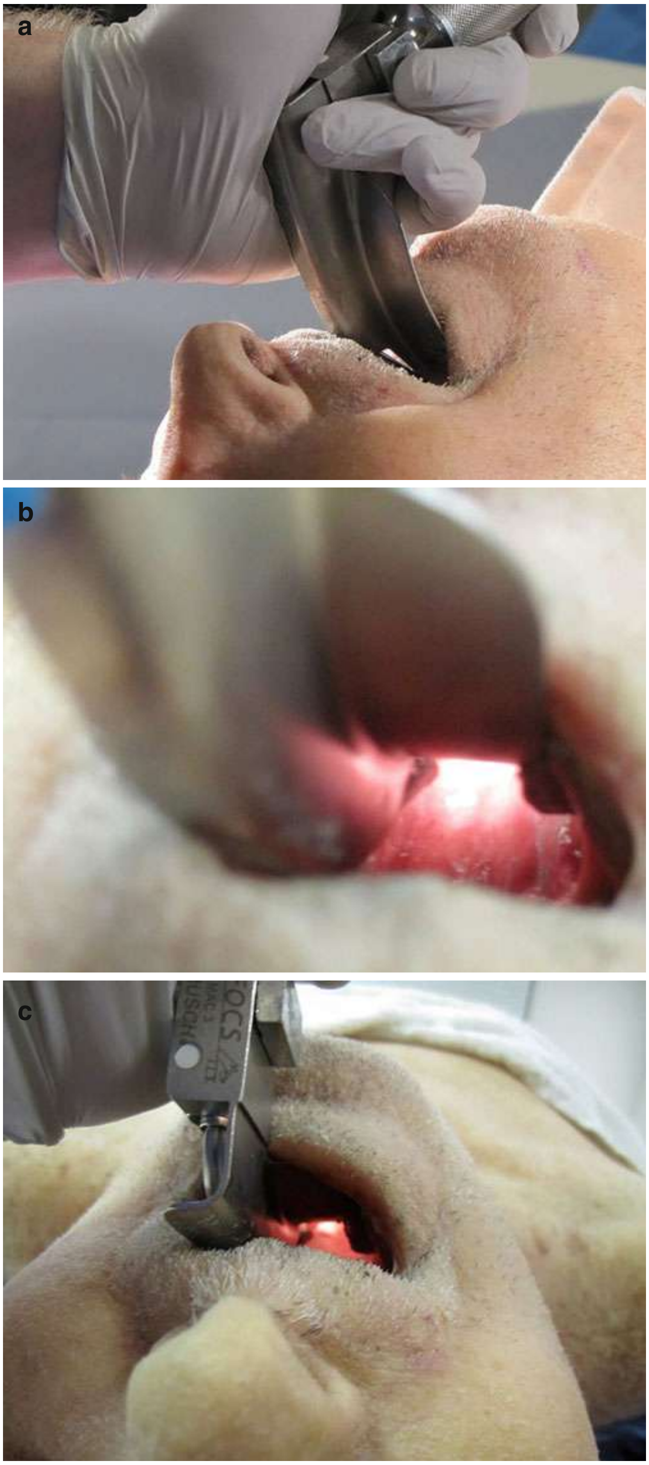

Fig. 10.6 (**a**) Insert the laryngoscope into the mouth, (**b**) sweep the tongue to the left, (**c**) advance the laryngoscope until the epiglottis is visible, and insert the blade into vallecula to expose the cords. When using the Miller blade, insert the blade until the epiglottis is seen. Slide the blade under the epiglottis and lift to expose the cords

10.5 Complications [3, 5]

- Common (1–4 %)
  - Esophageal intubation—*can be fatal if unrecognized*
  - Mainstem bronchus intubation
  - Tachycardia
  - Aspiration
  - Hypotension
- Uncommon (<1 %)
  - Dental/oral trauma
  - Oropharyngeal edema or bleeding
  - Laryngospasm
  - Dysrhythmia
  - Pneumothorax
  - Cardiac arrest

10.6 Pearls and Pitfalls

- Pearls
  - Positioning is of key importance—all patients with a stable cervical spine should be placed in the "sniffing" position to maximize view.
  - Consider the "ramping" position in obese patients with stable cervical spines—elevation of the head and shoulders allows redundant tissue to fall and gives an improved glottic view.
  - Always have suction readily available to remove blood, vomitus, or edema.
  - If structures are not readily visible, withdraw the blade gradually because it is common to insert the blade too deep.
- Pitfalls
  - "Rocking" the laryngoscope instead of lifting outward
  - Failure to recognize esophageal intubation
  - Failure to evaluate postintubation chest x-ray

References

1. Hagberg CA. Benumof's airway management. Maryland Heights, MO: Mosby; 2007. p. 363–5.
2. Park SH, Park HP, Jeon YT, Hwang JW, Kim JH, Bahk JH. A comparison of direct laryngoscopic views depending on pillow height. J Anesth. 2010;24:526–30.
3. Martin LD, Mhyre JM, Shanks AM, Tremper KK, Kheterpal S. 3,423 emergency tracheal intubations at a university hospital: airway outcomes and complications. Anesthesiology. 2011;114:42–8.
4. Levitan RM, Kinkle WC, Levin WJ, Everett WW. Laryngeal view during laryngoscopy: a randomized trial comparing cricoid pressure, backward-upward-rightward pressure, and bimanual laryngoscopy. Ann Emerg Med. 2006;47:548–55.
5. Walls RM, Brown CA 3rd, Bair AE, Pallin DJ, NEAR II Investigators. Emergency airway management: a multi-center report of 8937 emergency department intubations. J Emerg Med. 2011;41:347–54.

Laryngeal Mask Airway

11

Sohan Parekh

11.1 Indications

- Rescue device in a failed intubation
- Initial device in a predictably difficult airway
- Temporizing airway prior to definitive endotracheal intubation or surgical airway

11.2 Contraindications

- Absolute
  - Inadequate mouth opening
- Relative
  - Neck trauma/injury/radiation
  - High risk of aspiration

S. Parekh, MD
Department of Emergency Medicine,
University of Texas at Austin Dell Medical School,
Austin, TX, USA

Department of Emergency Medicine,
University Medical Center Brackenridge,
Austin, TX, USA
e-mail: sohan.parekh@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_11

11.3 Types (Fig. 11.1)

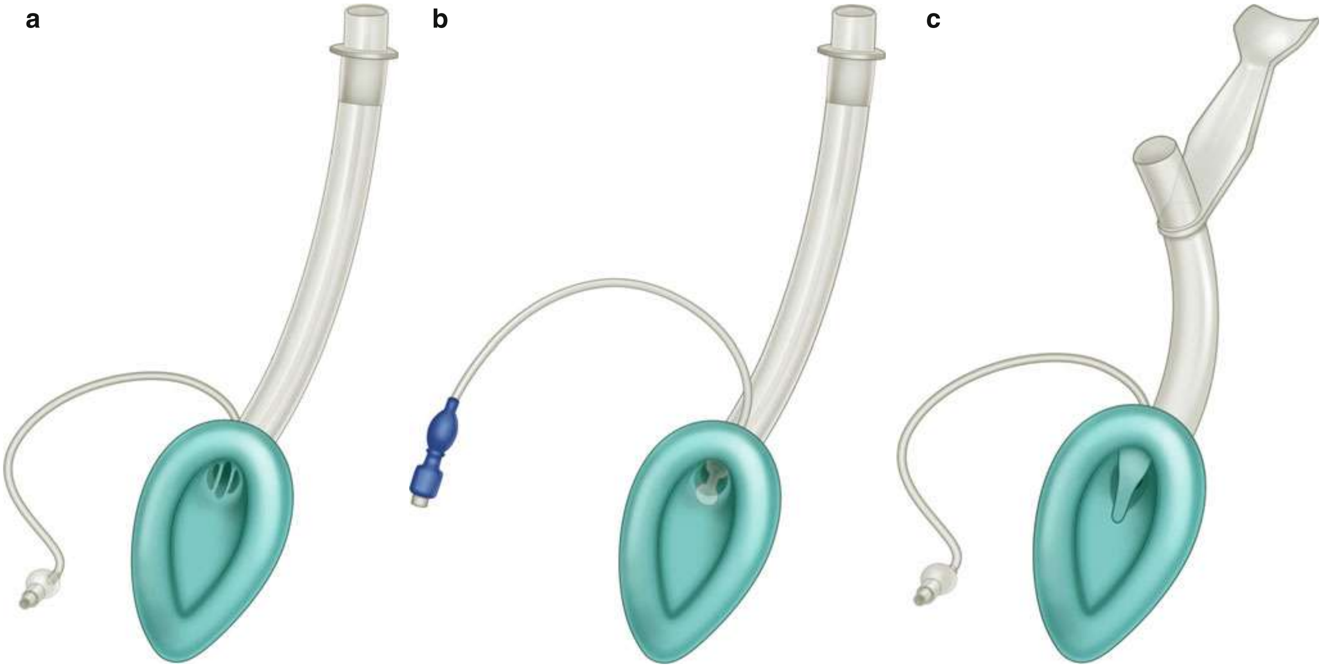

Fig. 11.1 Types of laryngeal mask airway: (**a**) LMA Unique, (**b**) LMA Classic Excel, (**c**) LMA Fastrach

11.4 Materials and Medications

- Appropriately sized laryngeal mask airway (LMA; LMA Unique/LMA Classic Excel/LMA Fastrach) and corresponding syringe (Table 11.1)
- Surgilube
- Bag valve mask
- Continuous end-tidal CO_2 ($EtCO_2$) or colorimetric $EtCO_2$ detector
- 8-mm or smaller endotracheal tube (ETT) (for Fastrach intubating LMA only)

Table 11.1 Laryngeal mask airway sizing

Size	Patient weight (kg)	Maximum cuff inflation volume (mL)	LMA product availability
1	<5	4	Unique
1½	5–10	7	Unique
2	10–20	10	Unique
2½	20–30	14	Unique
3	30–50	20	Unique, Classic Excel, Fastrach
4	50–70	30	Unique, Classic Excel, Fastrach
5	70–100	40	Unique, Classic Excel, Fastrach
6	>100	50	Unique

LMA laryngeal mask airway

11.5 Procedure

11.5.1 LMA Unique or Classic Excel

1. If using a reusable LMA Classic Excel, ensure that it is sterile and inspect it for any damage or wear.
2. Tightly deflate the cuff using a syringe such that it forms a spoon shape (Fig. 11.2).
3. Lubricate the posterior surface of the LMA with sterile lubricating jelly.
4. Stand behind the patient at the head of the bed as in direct laryngoscopy.
5. Place the patient's head in the sniffing position and ensure proper induction and paralysis.
6. Hold the LMA with the index finger of the dominant hand positioned at the juncture of the tube and cuff (Fig. 11.3).
7. Widely open the mouth with the nondominant hand and insert the LMA with the flattened tip flush with the palate.
    - Ensure that the tip of the device does not fold over during insertion.
8. Using the index finger, push the LMA along the curvature of the hard and soft palate (Fig. 11.4).
9. Continue to insert the LMA into the hypopharynx until resistance is felt. (At this point the tip of the LMA is in the esophagus.)
10. Stabilize the tube with the nondominant hand and remove index finger of the dominant hand from the LMA.
11. Inflate the cuff of the LMA to at least half of the maximum value using a syringe.
    - The LMA might move slightly outward during cuff inflation as the LMA positions itself in the hypopharynx.
12. Confirm placement and adequate gas exchange with continuous $EtCO_2$ capnography or colorimetry.

Fig. 11.2 Deflate the laryngeal mask airway (LMA) cuff using a syringe

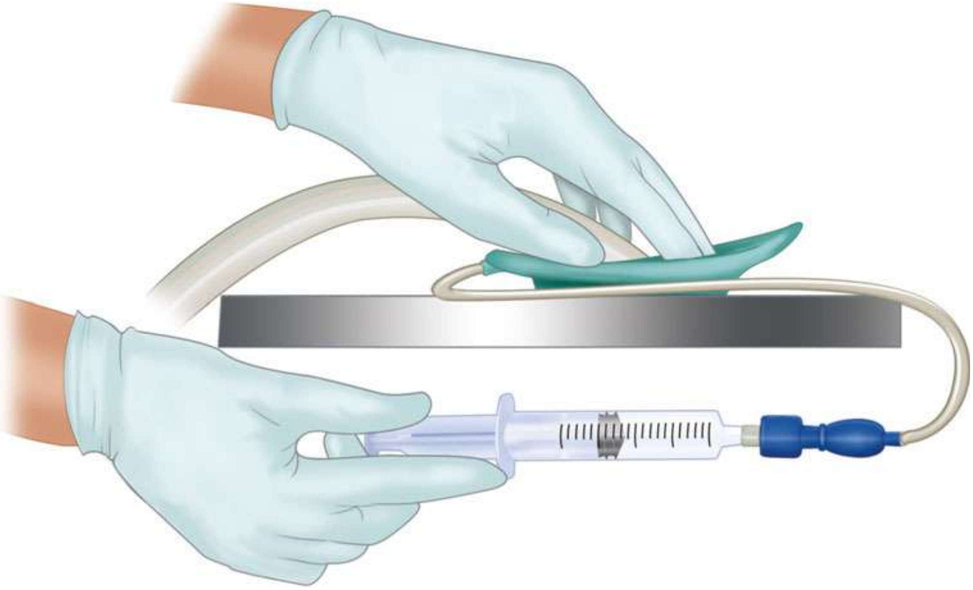

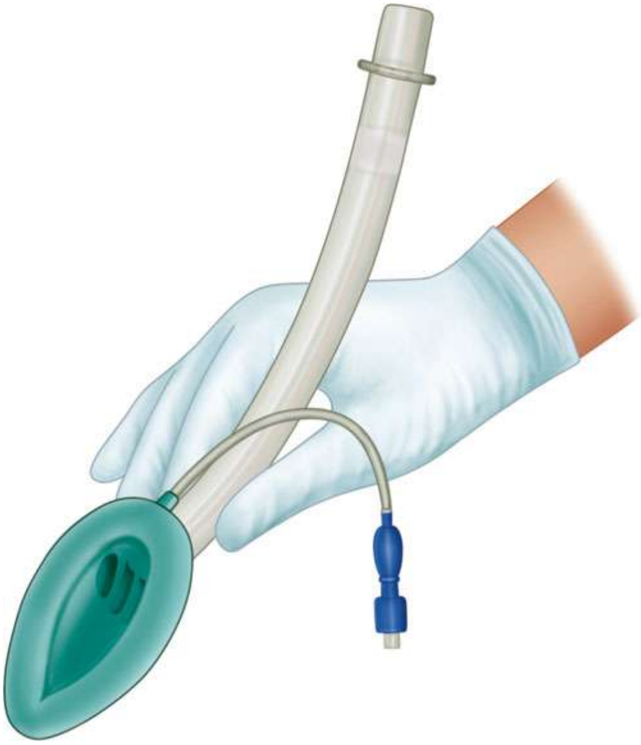

Fig. 11.3 Hold the LMA with the index finger positioned at the juncture of the tube and the cuff

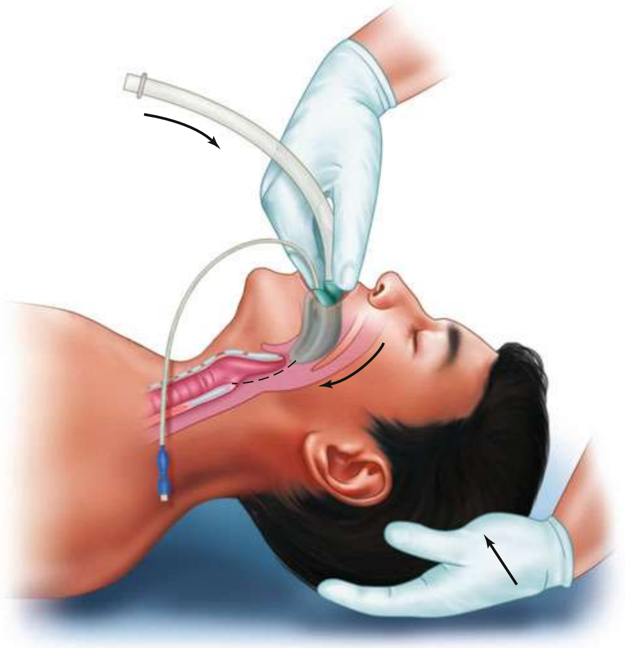

Fig. 11.4 Use the index finger to guide the LMA along the hard and soft palate

11.5.2 LMA Fastrach

1. If using a reusable LMA Fastrach, ensure that it is sterile and inspect it for any damage or wear.
2. Tightly deflate the cuff using a syringe such that it forms a spoon shape.
3. Lubricate the posterior surface of the LMA with sterile lubricating jelly.
4. The LMA Fastrach may be inserted from any position with respect to the patient's head.
5. Position the patient's head in the neutral position. Do not extend the head.
6. Widely open the mouth with the nondominant hand.
7. Holding the handle of the LMA Fastrach, insert the device into the mouth, placing the deflated cuff flush with the superior palate.
    - Distribute the lubricant over the superior palate using a side-to-side motion to allow for easier insertion.
    - Ensure that the tip of the device does not fold over during insertion.
8. Using the handle, gently advance the LMA Fastrach directly into the oropharynx until the curved portion of tube comes into the contact with the patient's chin (Fig. 11.5).
9. At this point use the handle to rotationally advance the device further into the oropharynx following the natural curvature of the palate and posterior pharynx (Fig. 11.6).
    - Do not initiate any rotation until the tube is in contact with the patient's chin.
10. Once resistance is felt, inflate the cuff of the device to at least half of the maximum value using a syringe.
    - Note that the tube is directed slightly caudally when properly inserted.
    - Confirm placement and adequate gas exchange with $EtCO_2$ capnography or colorimetry.

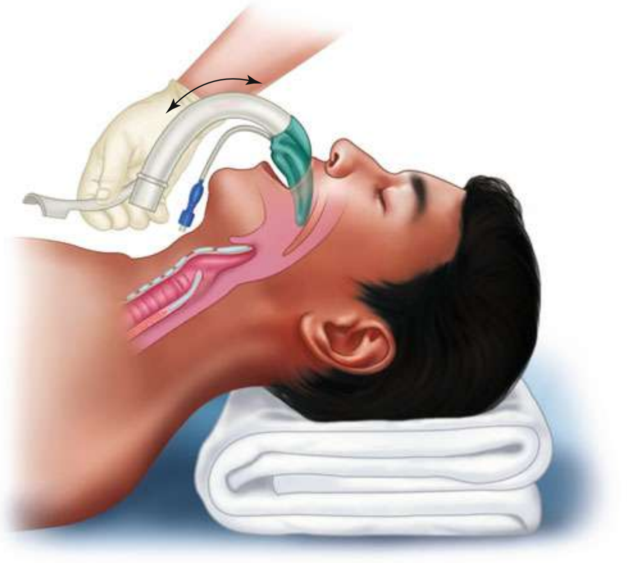

Fig. 11.5 Using the handle, insert the LMA Fastrach such that the posterior surface is in contact with the superior palate

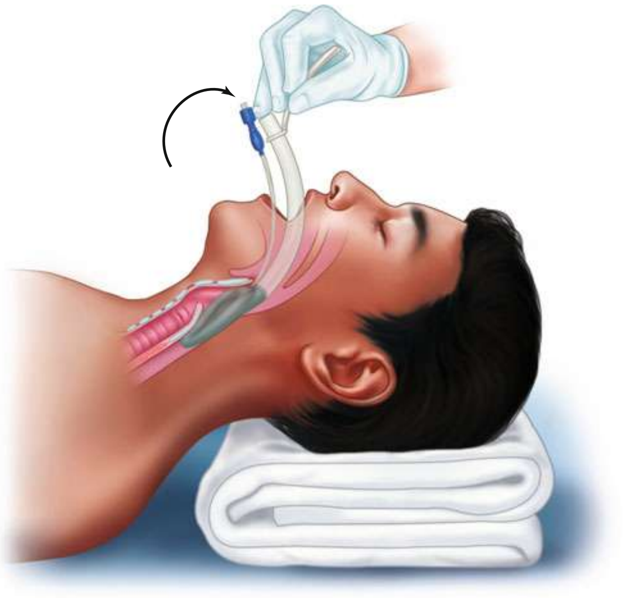

Fig. 11.6 Once the tube is in contact with the chin, use the handle to rotationally advance the device further into the oropharynx

11.5.3 Endotracheal Intubation through the LMA Fastrach

1. Ensure that the ETT will pass freely in the LMA.
2. Lubricate the cuff of the ETT.
3. Firmly hold the handle of the LMA Fastrach with the nondominant hand and insert the ETT to a depth of 15 cm (Fig. 11.7), which places the ETT tip at the point of emergency from LMA Fastrach.
    - Ensure that the tube does not pass beyond 15 cm at this point.
4. Using the handle of the LMA Fastrach, draw the device outward in order to displace the larynx slightly to accommodate insertion of the ETT (Fig. 11.8).
    - Use a lifting rather than a levering motion.
5. Carefully advance the ETT slightly further. If no resistance is felt, continue with insertion of the ETT (Fig. 11.9).
6. Confirm placement and adequate gas exchange with $EtCO_2$ capnography or colorimetry.
7. Once successful confirmation of intubation is established, deflate the cuff pressure on the LMA Fastrach.

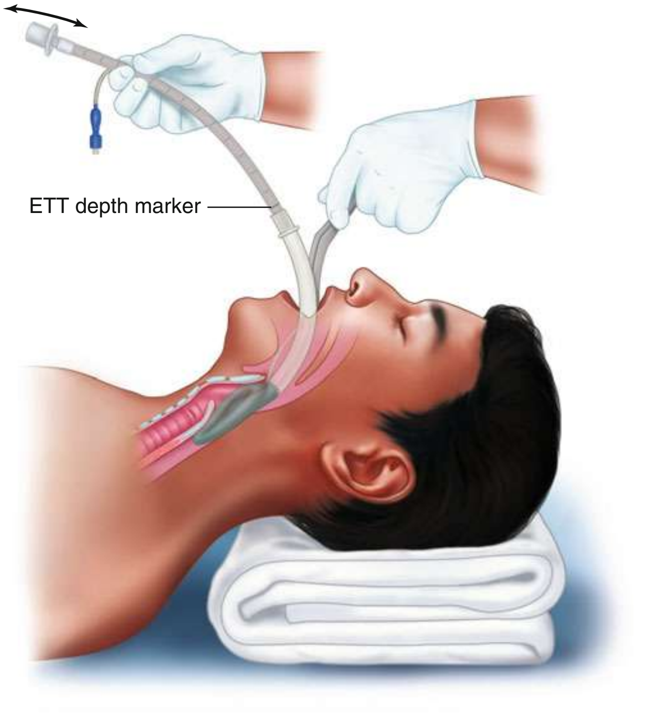

Fig. 11.7 While holding the handle of the LMA, insert the endotracheal tube (*ETT*) to the 15-cm mark

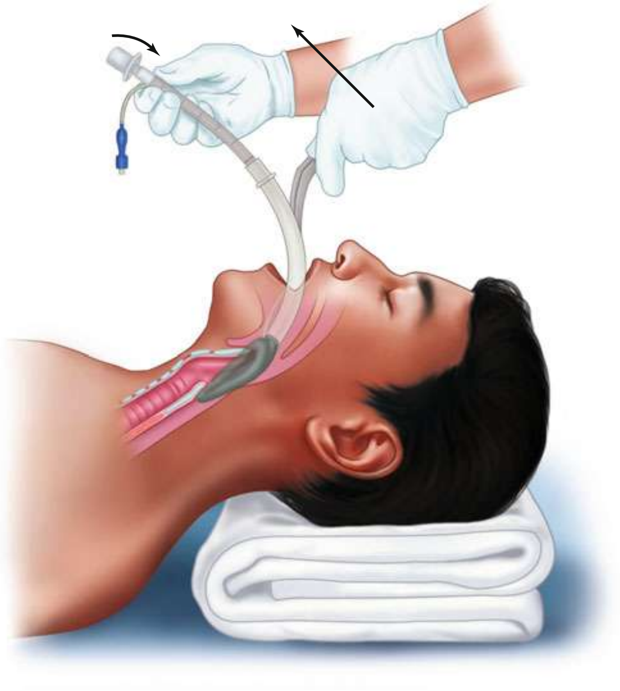

Fig. 11.8 Lift the handle outward to open the glottis for the ETT

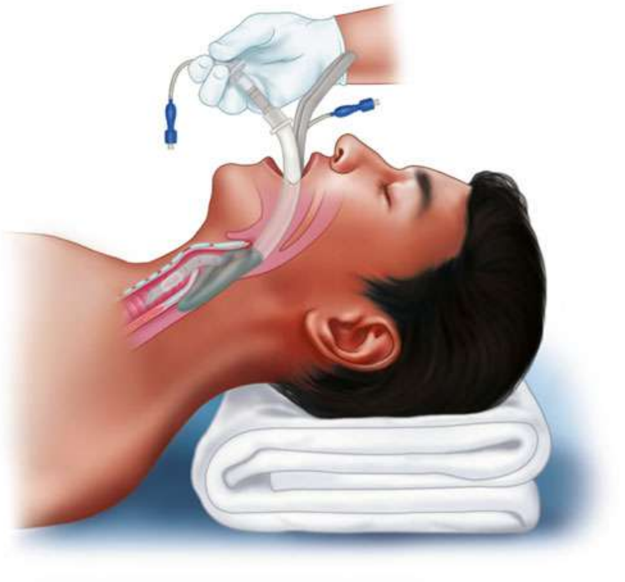

Fig. 11.9 If no resistance is felt during further insertion of the ETT, the ETT may be fully advanced

11.5.4 Removing the LMA Fastrach after Successful Intubation

1. The LMA Fastrach need not be removed immediately, but if this is desired, first adequately oxygenate the patient and then disconnect the patient from the circuit.
2. Remove the airway connector from the proximal end of the ETT.
3. Ensure that the cuff of the LMA Fastrach is entirely deflated.
4. Stabilize the ETT with the nondominant hand, and using the dominant hand, gently ease the LMA Fastrach out by rotating the handle caudally (Fig. 11.10).
5. Once the tube of the LMA Fastrach reaches the proximal end of the ETT, use the stabilizer rod to maintain the position of the ETT while continuing to remove the LMA Fastrach using the handle (Fig. 11.11).
6. After the cuff of the LMA Fastrach has been fully removed from the oral cavity, release the stabilizer rod and ensure stability of the ETT by grasping it distally at the mouth with the nondominant hand (Fig. 11.12).
7. Continue to ease the LMA Fastrach out from around the ETT, ensuring that the pilot balloon and inflation line of the ETT cuff pass through the device (Fig. 11.13).
   - Take care not to rupture the pilot balloon or tear the inflation line of the ETT.
8. Replace the airway connector on the proximal end of the ETT and reconnect the patient to the circuit.

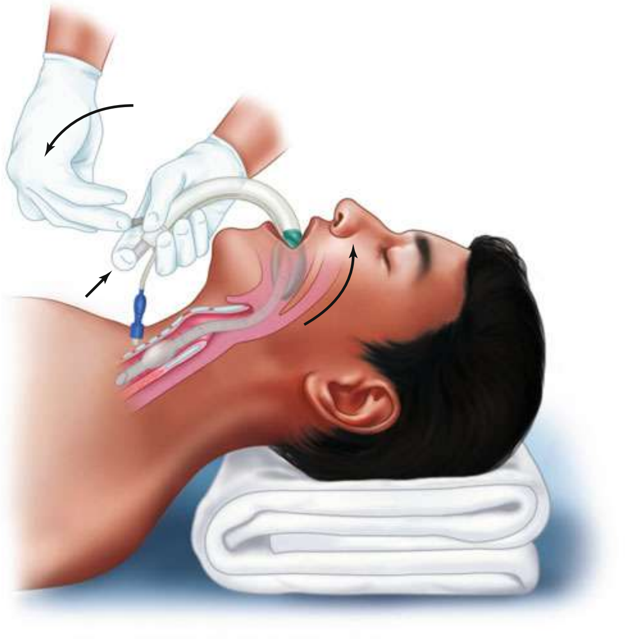

Fig. 11.10 Rotate the handle caudally to gently ease the LMA Fastrach out of the pharynx

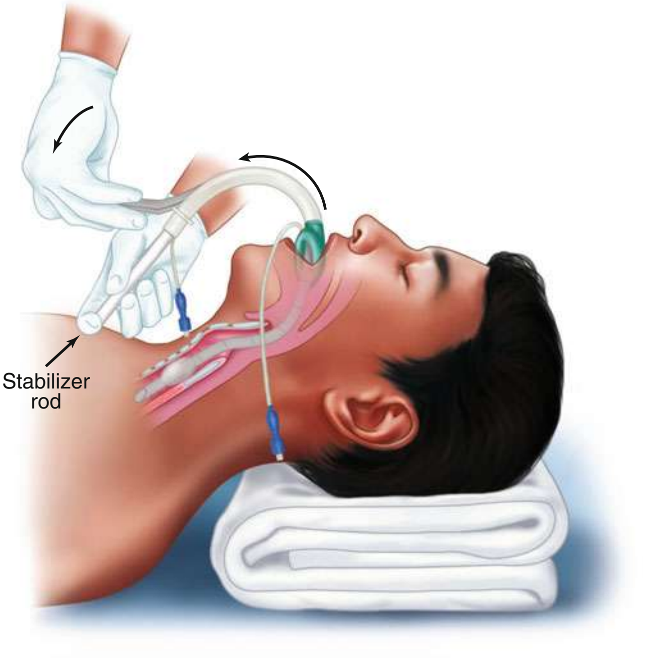

Fig. 11.11 Use the stabilizer rod to allow for further removal of the LMA Fastrach

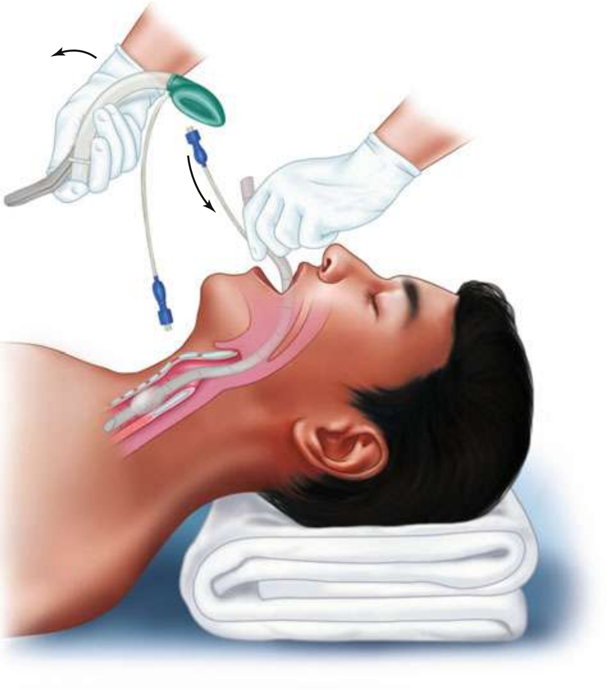

Fig. 11.13 Carefully pass the pilot balloon and inflation line of the ETT cuff through the tube of LMA Fastrach as it is removed

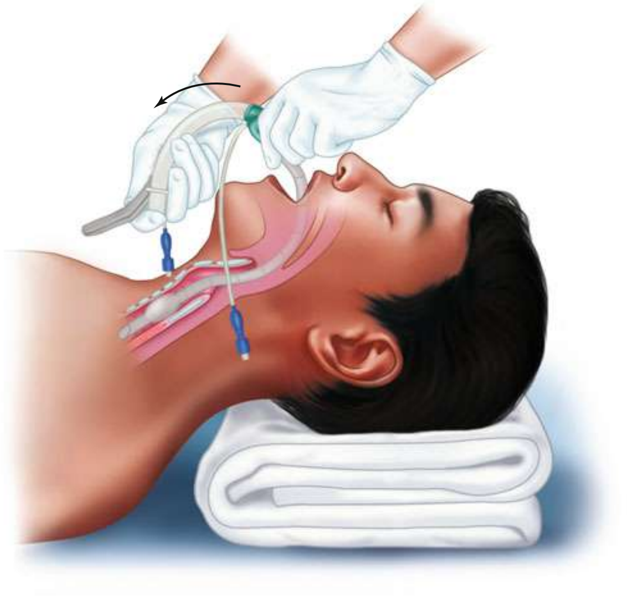

Fig. 11.12 Once the cuff of the LMA Fastrach is out of the mouth, grasp the ETT distally and remove the LMA entirely

11.6 Complications

- Aspiration with resulting pneumonitis
- Ineffective seal resulting in insufficient ventilation
- Coughing, bucking, or breath holding
  - Ensure that the patient is adequately sedated.

11.7 Pearls and Pitfalls

- Cricoid pressure can push the tip of the LMA out of the esophagus and prevent optimal placement.

Selected Reading

Barata I. The laryngeal mask airway: prehospital and emergency department use. Emerg Med Clin North Am. 2008;24:1069–83.
LMA classic LMA flexible LMA classic single use LMA flexible single use instruction manual. Victoria: The Laryngeal Mask Company Limited; 2010.
LMA fastrach LMA fastrach single use instruction manual. Victoria: The Laryngeal Mask Company Limited; 2010.
Pollack CV. The laryngeal mask airway: a comprehensive review for the emergency physician. J Emerg Med. 2001;20:53–66.
Walls RW, Murphy MF, editors. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Combitube

12

Clint Masterson

12.1 Indications

- Need for ventilation and oxygenation in an unconscious, unresponsive, or paralyzed patient
- Rescue airway needed after failed intubation

12.2 Contraindications

- Absolute
  - Awake, responsive patient
  - Intact gag reflex
  - Known esophageal disease
  - Ingestion of caustic substances
  - Child (no Combitubes are made for children)
- Relative
  - D50 or naloxone about to be given
  - Facial trauma

12.3 Materials

- Combitube sized based upon height (Fig. 12.1)
  - >5 ft—size 41 French (cuff inflation 15 and 100 mL)
  - >4 ft to < 5.5 ft—size 37 French (cuff inflation 12 and 85 mL)

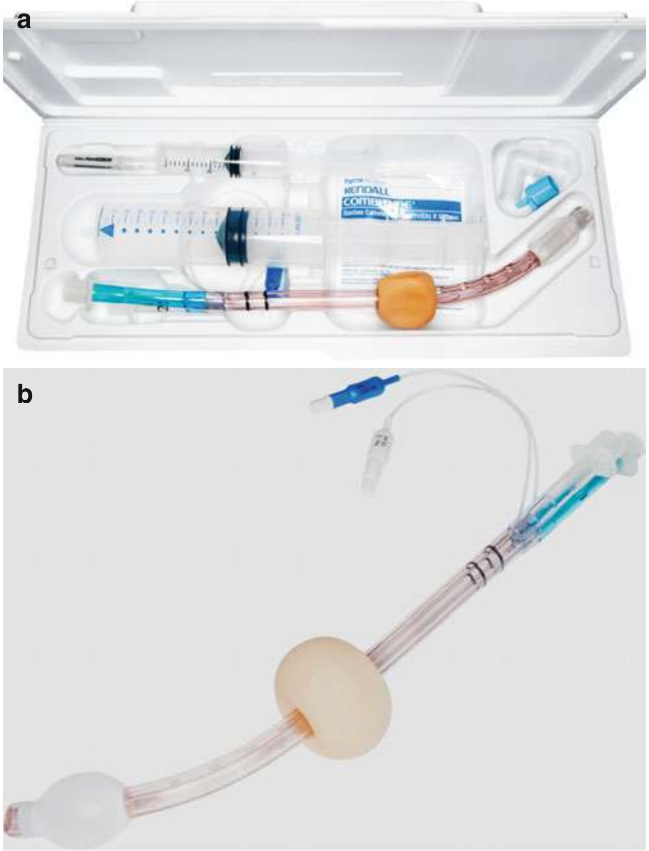

Fig. 12.1 (**a**, **b**) Combitube equipment

C. Masterson, MD
Department of Emergency Medicine,
Mayo Clinic Health System in Fairmont,
Fairmont, MN, USA
e-mail: clint@clintm.us

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_12

12.4 Procedure

1. Test both balloons and cuffs for leaks as one would an endotracheal tube (ETT).
2. Open up the airway.
    (a) Use a laryngoscope to move the tongue and open the oropharynx.
        OR
    (b) Use the left hand to elevate the chin, elevating the tongue and pharyngeal tissue.
3. Insert Combitube blindly into the oropharynx until the teeth lie between the two black bands on the proximal Combitube (Fig. 12.2).
4. Inflate the proximal blue cuff until air pressure is produced or the manufacturer-recommended pressure is reached.
    (a) 85 mL for 37 French Combitube
    (b) 100 mL for 41 French Combitube
5. Identify placement and attach to oxygen.
    (a) Ventilate through tube #1 (blue).
    (b) Auscultate the stomach and lungs.
        (i) If breath sounds are heard, the Combitube is in its more common esophageal location.
        (ii) Attach tube #1 to bag valve mask and O_2.
    (c) ONLY IF gurgling is present over the stomach when tube #1 is ventilated:
        (i) Ventilate through tube #2.
        (ii) If breath sounds are heard, the Combitube is in the less common tracheal location.
        (iii) Attach tube #2 to bag valve mask and O_2.
6. If no breath sounds are heard in either location:
    (a) Consider obstruction—Combitube may be obstructing the glottis or collapsing the trachea owing to deep proximal cuff inflation.
        (i) Deflate the cuffs.
        (ii) Withdraw 3 cm.
        (iii) Reinflate and start from step 4.
    (b) Consider equipment failure.
        (i) Check that balloons are maintaining pressure and intact.
    (c) Consider reinsertion.
7. Confirm placement with capnogram and pulse oximetry.
8. Secure the Combitube in position (Fig. 12.3).

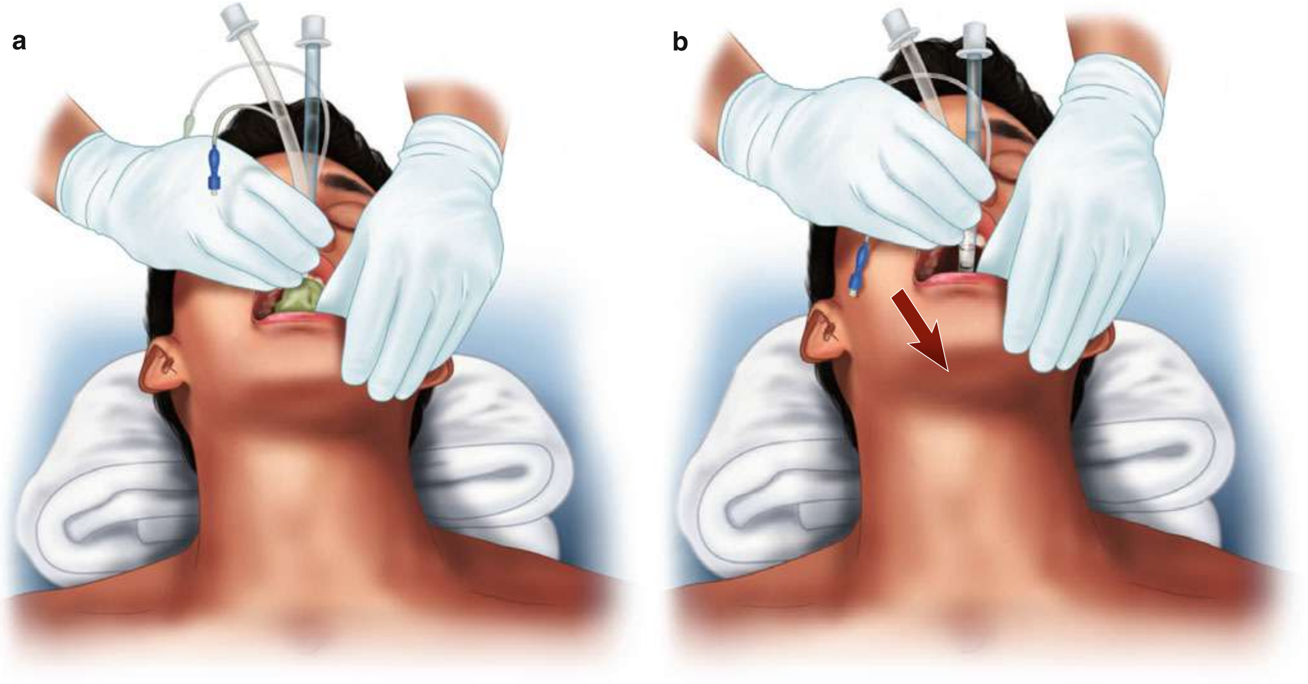

Fig. 12.2 (**a**) Insertion of Combitube. (**b**) Teeth should lie between the two black bands on the proximal Combitube

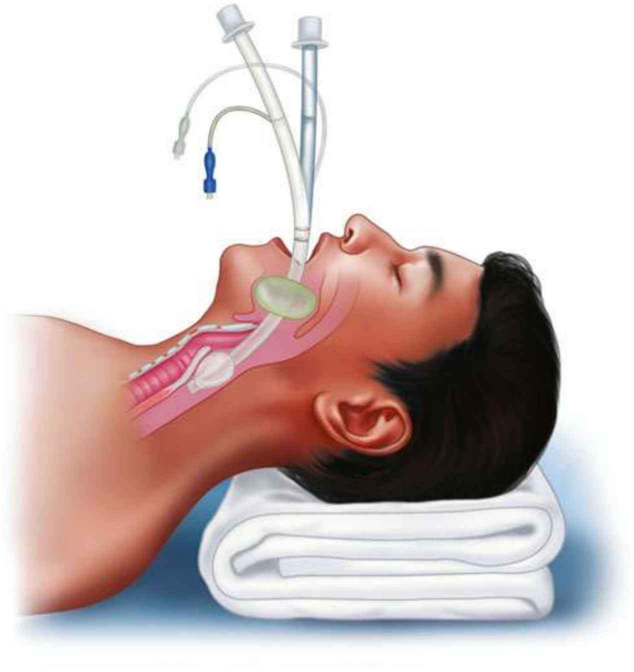

Fig. 12.3 Secure the Combitube in position

12.5 Pearls and Pitfalls

- Pearls
  - In an esophageal intubation situation, a suction tube may be threaded down using tube #2 to decompress the stomach.
  - The Combitube comes with an L-shaped piece that may also be attached to the end of tube #2 to deflect gastric contents away from practitioners.
- Pitfalls
  - After Combitube placement, a definitive airway should be placed when possible.
    - Gastric contents may aspirate despite placement of the Combitube.
    - Combitube should be considered a bridging airway device.
  - Combitubes are associated with a more pronounced hemodynamic stress response than ETTs or laryngeal mask airways (LMAs).
  - Balloon overinflation can lead to esophageal rupture (albeit rare).
  - Combitubes are associated with an increased incidence of sore throat, dysphagia, and upper airway hematomas than ETTs and LMAs.
  - Piriform sinus perforation.

Selected Reading

Agro F, Frass M, Beunmof JL, Krafft P. Current status of the Combitube: a review of the literature. J Clin Anesth. 2002;14:307–14.

Laurin E, Bair A. Devices for difficult airway management in adults. www.uptodate.com. Accessed 14 Mar 2014.

Liem EB. Combitube insertion. University of Florida Department of Anesthesiology, Center for Simulation, Advanced Learning and Technology, Virtual Anesthesia Machine Web site; 2006. http://vam.anest.ufl.edu/airwaydevice/combitube/index.html. Accessed 14 Mar 2014.

Walls R, Murphy M. Manual of emergency airway management. Philadelphia: Lippincott Williams & Wilkins; 2008.

Assessment of the Difficult Airway

13

Melinda W. Fernandez and Lars K. Beattie

13.1 Indications

- Should be performed on all patients that require airway management, conditions permitting
- Respiratory distress
- Airway protection

13.2 Materials and Medications

- None required

13.3 Procedure

1. Anticipating a difficult airway in emergency department patients is the first step to avoiding an unexpected surgical airway.
2. Two mnemonics can be applied quickly and easily to aid in an airway assessment: MOANS and LEMON.

13.3.1 Predicting Bag-Valve-Mask Difficulty

1. Use the mnemonic *MOANS* to assess for possible bag-valve-mask (BVM) difficulty.
   M—mask seal. Will you be able to get a good seal on the face? Predictors of difficulty include facial hair such as a beard, elderly, or emaciated owing to loss of muscle tone in the face.
   O—obesity. Body mass index (BMI) >30.
   A—age (>55 years). Loss of facial muscle tone can make bagging difficult.
   N—no teeth. Although being edentulous makes for an easier intubation, it makes bagging more difficult.
   S—stiff lungs. Acute or chronic lung disease can make a person difficult to bag. In the setting of Trauma, pulmonary contusion(s) and/or other direct lung injuries may increase BVM difficulty.

13.3.2 Predicting Difficult Laryngoscopy

1. Attempts should be made, if at all possible, to assess for a potentially difficult airway. This does not mean you cannot perform direct laryngoscopy if you are anticipating a difficult airway. It does, however, force you to consider all options and to have a solid backup plan in place with backup equipment readily available in the room.
2. Use the mnemonic *LEMON* to predict difficult direct laryngoscopy.
   L—look. A quick look at the patient will tell you a lot. Are there facial injuries; facial anomalies; obesity; short, thick neck; and small mouth or mandible?
   E—evaluate. Use the 3-3-2 rule to quickly assess for the strongest predictors of difficult laryngoscopy.
   - *3:* Open the patient's mouth and three vertically aligned fingers should fit between the incisors.
   - *3:* Three finger widths should fit along the length of the mandible from the mentum to the hyoid bone. Shorter or longer distances may make for a difficult intubation.
   - *2:* Thyromental distance should ideally be two fingers. Measure this from the hyoid to the thyroid.

M.W. Fernandez, MD
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: mindyfernandez@ufl.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_13

M—Mallampati classification (Fig. 13.1). If patient's condition and situation allow, have the patient open the mouth wide, stick out the tongue, and say "Ahh." Evaluate for visible structures.

- *Class I:* Tonsillar pillars and the entire uvula are visible.
- *Class II:* More than the base of the uvula is visible but no pillars are visible.
- *Class III:* Only the base of the uvula is visible.
- *Class IV:* No uvula or soft palate is visible. Only the hard palate is visible.
- These classifications correlate with the Cormack-Lehane grading system for laryngoscopic views. A Mallampati class I will correlate with a grade 1 view about 99 % of the time, whereas a Mallampati IV will be a grade 3 or 4 view all of the time and a rescue plan with backup equipment immediately available should always be in place [1, 2].

O—obstruction. Observe for anything that can get in the way (e.g., the tongue, dentures, blood, vomit, foreign body, edema, redundant tissue).

N—neck mobility. If patient's condition and situation allow, have the patient flex and extend the neck to evaluate mobility. Many patients in the emergency department have limited neck mobility. Examples include the trauma patient who arrives in cervical collar immobilization or a patient with degenerative or rheumatoid arthritis.

3. The *"6-D"* method is another assessment tool that can be used to predict difficult laryngoscopy and intubation. This method can be remembered by the fact that the word "difficult" begins with the letter "D":

Disproportion

- Increased tongue size in relation to pharyngeal size
- Airway swelling or trauma

Distortion

- Neck mass, hematoma, abscess, previous surgical airway, arthritic neck changes

Decreased thyromental distance

- Anterior larynx and decreased mandibular space.
- Look for a receding chin or greater than three fingerbreadths from the mentum to the hyoid bone.

Decreased inter-incisor gap

- Reduced mouth opening.
- Look for less than two to three fingerbreadths placed vertically in the patient's open mouth.

Decreased range of motion in any joints of the airway

- Limited head extension
- Previous neck radiation and/or surgery
- Neck contractures

Dental overbite

- Oversized, angled teeth disrupt the alignment of airway axes.
- Can decrease the interincisor gap.

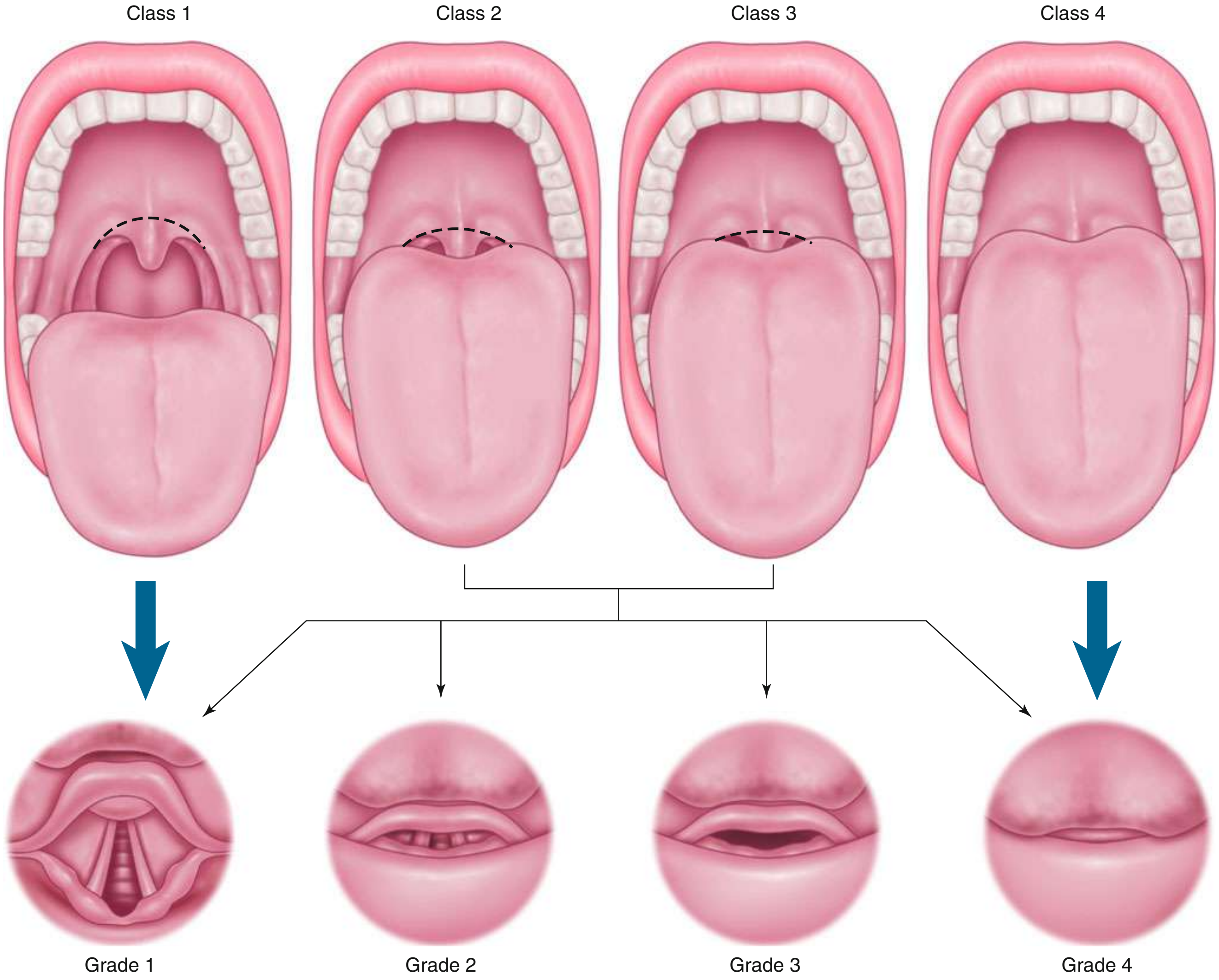
Class 1
Class 2
Class 3
Class 4
Grade 1
Grade 2
Grade 3
Grade 4

Fig. 13.1 Mallampati classification

13.3.3 Optimizing Laryngoscopy in the Obese Patient

- It is essential that emergency medicine physicians are able to successfully intubate the obese patient.
- Proper assessment and positioning will increase the success rate:
  - The goal is to ensure alignment of the oropharyngeal-pharyngeal-laryngeal (OA-PA-LA) airways by placing the patient in the head-elevated laryngoscopy position (Fig. 13.2a).
  - Align the external auditory meatus with the sternal notch along a horizontal line by positioning the patient on a "ramp."
  - The ramp can be created by stacking blankets/towels under the lower back ramping up to the neck and head (Fig. 13.2b).
- While the patient is in position on the ramp, the support is adjusted to minimize head flexion and allow for positioning in the sniffing position.
  - Because of the increased height, a step stool may be required to adequately visualize the airway from the head of the bed.

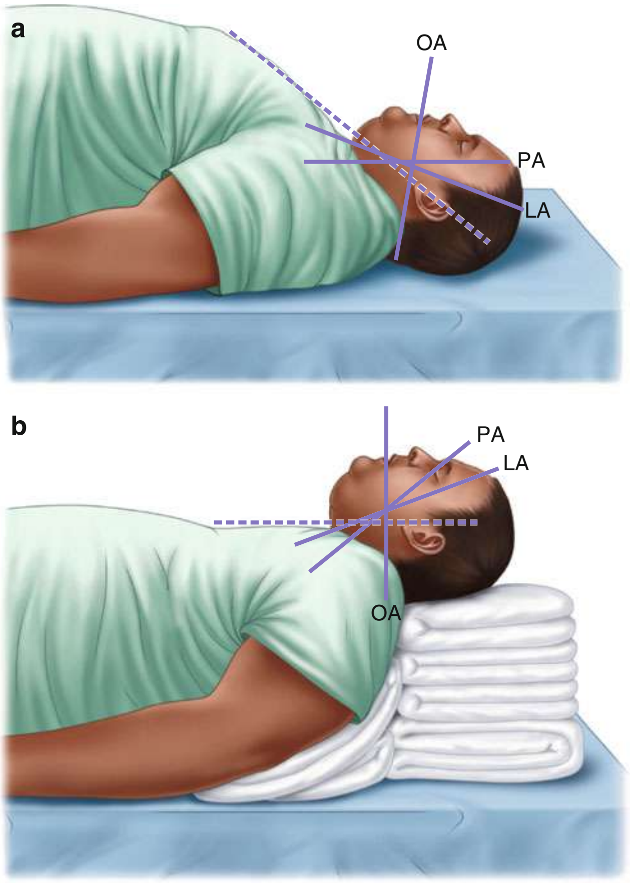

Fig. 13.2 (**a**, **b**) Ramping the obese patient will align the axes of the airway and allow easier direct laryngoscopy when viewed from the head of bed. (*LA* laryngeal airway, *OA* oropharyngeal airway, *PA* pharyngeal airway)

13.4 Pearls and Pitfalls

- Owing to time-sensitive patient care situations, emergency physicians are often not able to perform a thorough airway evaluation on every patient.
- With every airway that you manage and before pushing any drugs, always ask yourself:
  - Will I be able to ventilate this patient?
  - Will I be able to intubate this patient?
  - What is my difficult airway plan if I encounter trouble?
  - Will I be able to perform a surgical airway, if necessary?
- Be sure you have a solid backup plan A, B, and C before pushing any drugs.

References

1. Lee A, Fan LT, Gin T, Karmakar MK, Ngan Kee WD. A systematic review (meta-analysis) of the accuracy of the Mallampati tests to predict the difficult airway. Anesth Analg. 2006;102:1867–78.
2. Boschert S. Think L-E-M-O-N when assessing a difficult airway. ACEP News. Nov 2007.

Selected Reading

Murphy M. Bringing the larynx into view: a piece of the puzzle. Ann Emerg Med. 2003;41:338–41.

Rick J. Recognition and management of the difficult airway with special emphasis on the intubating LMA-Fastrach/whistle technique: a brief review with case reports. BUMC. 2005;18:220–7.

Roberts J, Hedges J. Clinical procedures in emergency medicine. 5th ed. Philadelphia: WB Saunders; 2009. p. 60–2.

Wilson W. Difficult intubation. In: Atlee J, editor. Complications in anesthesia. Philadelphia: WB Saunders; 1999. p. 138–47.

GlideScope

14

Sohan Parekh

14.1 Indications

- Initial device in a predictably difficult airway
- Rescue device in a failed intubation

14.2 Contraindications

- Absolute
  - Inadequate mouth opening
- Relative
  - Blood, vomit, or other secretions that can coat and obstruct the camera lens

14.3 Materials and Medications

- GlideScope video monitor with video cable (GlideScope Video Laryngoscope [GVL] system) (Fig. 14.1) or appropriate-size video baton (Cobalt System) (Fig. 14.2)
- Appropriate-size reusable video laryngoscope (GVL) or single-use laryngoscope blade (GVL Stat) (Table 14.1).
- Endotracheal tube (ETT)
- Malleable stylet or GlideRite rigid stylet
- 10 mL syringe
- End-tidal CO_2 ($EtCO_2$) capnography or colorimetry

S. Parekh, MD
Department of Emergency Medicine,
University of Texas at Austin Dell Medical School,
Austin, TX, USA

Department of Emergency Medicine,
University Medical Center Brackenridge,
Austin, TX, USA
e-mail: sohan.parekh@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_14

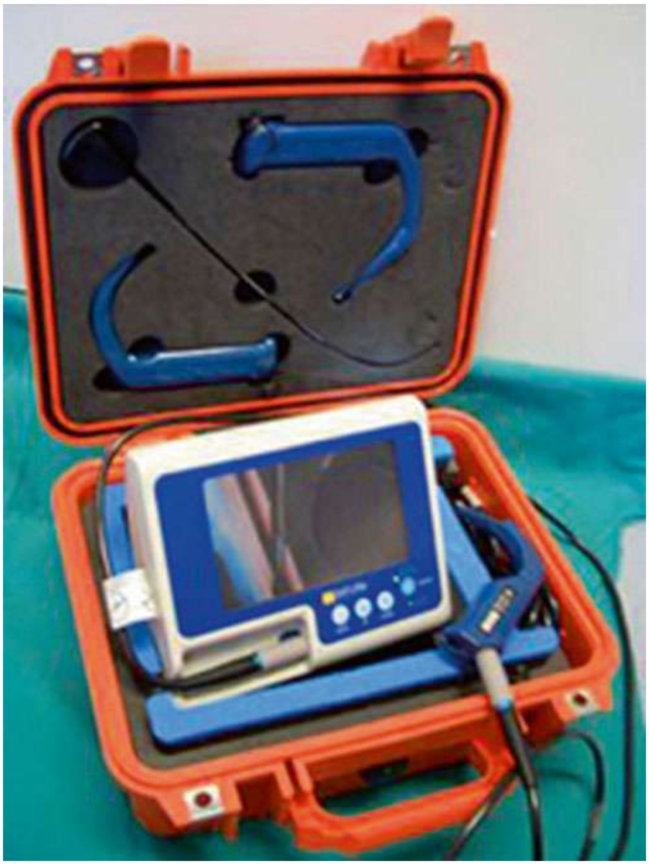

Fig. 14.1 GlideScope GVL system (With kind permission from Springer Science+Business Media: Noppens RR, Werner C, Piepho T. Indirekte Laryngoskopie. *Der Anaesthesist*. 2010;59(2):149–61)

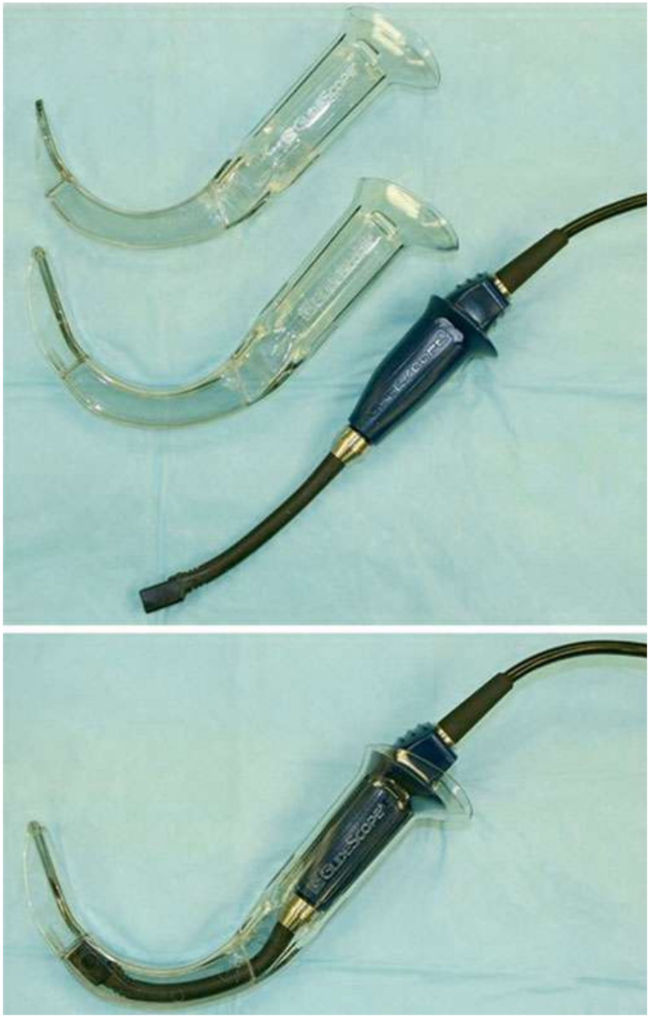

Fig. 14.2 GlideScope Cobalt system (With kind permission from Springer Science+Business Media: Jones PM, Turkstra TP, Armstrong KP, et al. Comparison of a single-use GlideScope® Cobalt videolaryngoscope with a conventional GlideScope® for orotracheal intubation. *Can J Anesthe/Journal canadien d'anesthésie*. 2010;57(1))

Table 14.1 GlideScope sizing

				Video Baton 1–2			Video Baton 3–4	
GVL 1	GVL 2	GVL 3	GVL 4	Stat 0	Stat 1	Stat 2	Stat 3	Stat 4
1.8–10 kg	10 kg—adult	40 kg—morbidly obese	40 kg—morbidly obese	<1.5 kg	1.5–3.6 kg	1.8–10 kg	10 kg—adult	40 kg—morbidly obese

14.4 Procedure

1. Insert the video cable (GVL system) or the video baton (Cobalt system) into the GlideScope video monitor (Fig. 14.3).
2. If using the GVL system, insert the distal end of the video cable into the port on the handle of the reusable video laryngoscope (GVL) (Fig. 14.4).
3. If using the Cobalt system, insert the video baton into the GVL Stat (Fig. 14.5).
   (a) Align the logo on the side of the video baton with the logo on the side of the single-use laryngoscope blade (GVL stat).
   (b) The video baton should slide smoothly and click into the place.
4. Turn on the GlideScope at for at least 30–120 s before use to fully activate the antifog mechanism.
5. Insert a stylet into the ETT. If using a malleable stylet, shape the curvature of the distal end of the tube to conform to the 60° curvature of the laryngoscope blade.
6. Firmly hold the laryngoscope handle in the left hand and ensure that an image can be clearly seen on the video monitor.
7. After ensuring adequate sedation and paralysis, open the mouth wide and insert the laryngoscope blade in the midline underneath the tongue (Fig. 14.6).

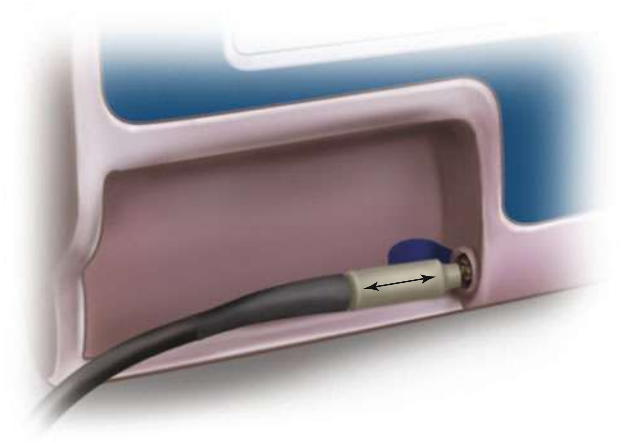

Fig. 14.3 Cable insertion into the video monitor

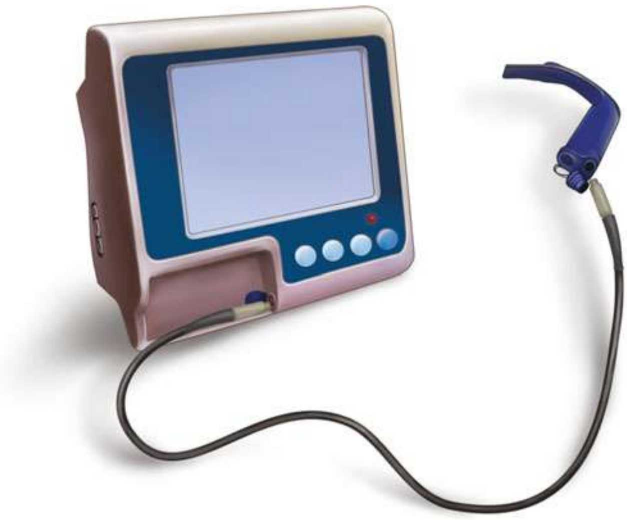

Fig. 14.4 Connect the distal end of the video cable the port on the handle of the GVL (GVL System)

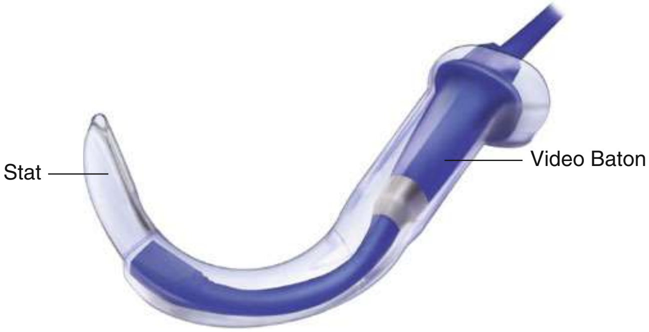

Fig. 14.5 Slide the video baton into the GVL Stat, Cobalt System

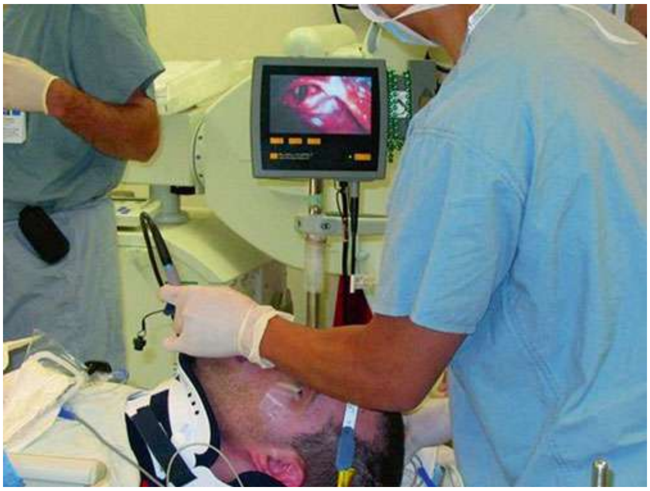

Fig. 14.6 Insert the laryngoscope blade in the midline beneath the tongue (with kind permission from Springer Science+Business Media: Osborn IP, Kleinberger AJ, Gurudutt VV. Chapter 8: Airway emergencies and the difficult airway. In: Levine AI, Govindaraj S, DeMaria S, editors. *Anesthesiology and otolaryngology*. 2013)

8. Looking at the video monitor, advance the laryngoscope blade further into the oropharynx in order to obtain a view of the epiglottis.
   (a) Do not look directly into the oropharynx.
   (b) Movements and adjustments should be guided by the image on the video monitor.
9. Place the laryngoscope blade in the vallecula (analogous to a Macintosh blade) and apply a gentle backward tilt to expose the glottis.
10. In the event that a satisfactory glottic view cannot be obtained, the laryngoscope blade may be advanced and used like a Miller blade to lift the epiglottis out of the way.
11. Directing attention back toward the patient, insert the ETT into the mouth adjacent to the laryngoscope blade.
12. Guide the ETT toward the tip of the laryngoscope such that the end of the ETT emerges on the video monitor.
13. Looking at the video monitor, advance the ETT toward the glottis, and maneuver the tip of the tube between the vocal cords by rotating and altering the angle of the ETT.
    (a) If the ETT tip is posterior to the arytenoids:
        (i) Pull the ETT superiorly, rotate it over the left arytenoid, and gently twist the tube over the epiglottic aperture.
        (ii) Apply external laryngeal manipulation.
        (iii) Withdraw the blade to reduce tilting of the laryngeal axis and lessen the angle of introduction.
    (b) If the ETT abuts the false vocal cords, turn the ETT in the clockwise direction while withdrawing the stylet (Fig. 14.7).
14. Using the thumb, partially withdraw the stylet a few centimeters from the ETT.
    (a) The distal end of the tube should be free of the stylet.
    (b) An assistant can perform this task to allow for greater control and stability of the ETT.
15. Insert the ETT to the desired depth.
16. Fully remove the stylet, inflate the cuff of the ETT using a syringe, and confirm placement with $EtCO_2$, capnography, or colorimetry.

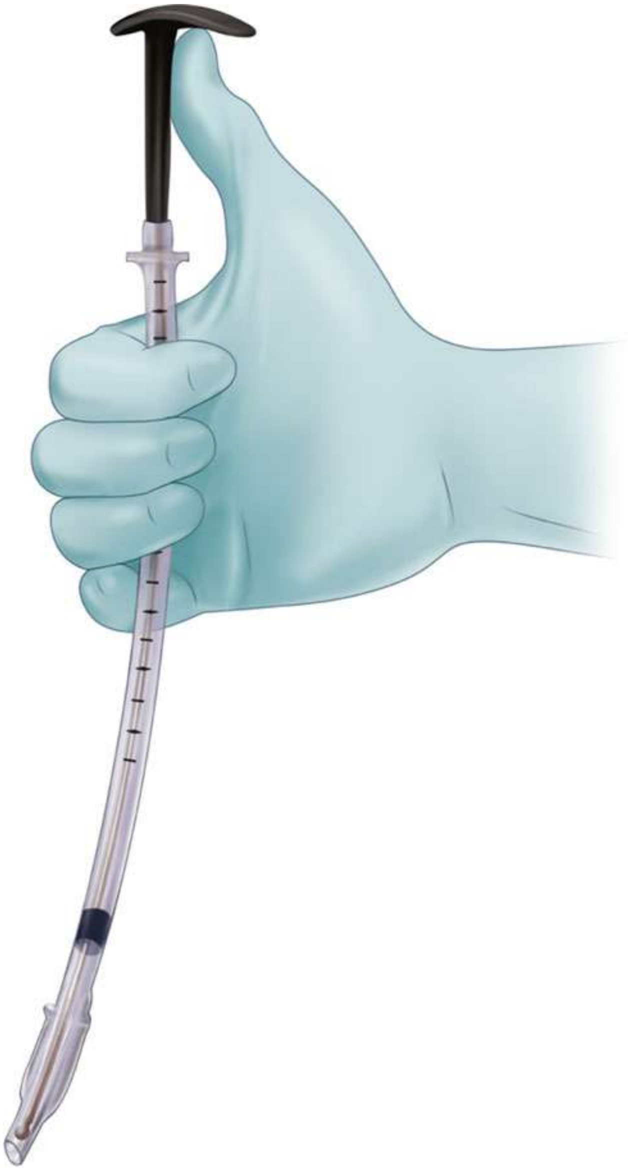

Fig. 14.7 Partially withdraw the stylet from the ETT to allow for passage through the vocal cords

14.5 Complications

- Dental injury
- Airway trauma

14.6 Pearls and Pitfalls

- Unlike conventional laryngoscopy, there is no need to displace the tongue.
- The greatest challenge when intubating with the GlideScope is maneuvering the ETT into the glottis aperture. Alternatives to the conventional technique are:
  – Make a 90° bend in the ETT just proximal to the cuff and insert it in the horizontal direction with the tip toward the right cheek. Once the tube is advanced past the flange of the laryngoscope, rotate it counterclockwise, at which point it should be pointed at the glottis. The tube can then be gently rotated into the glottis.
  – Consider inserting the laryngoscope slightly to the left of the midline upon initial insertion to allow greater space for advancement and maneuvering of the ETT.
- Do not overly lever the laryngoscope or use excessive lifting force after insertion into the vallecula. An adequate view of the glottis is generally easily obtained with minimal effort.
- Display settings can be adjusted using the menu button on the video monitor.

Selected Reading

Cho JE, Kil HK. A maneuver to facilitate endotracheal intubation using the GlideScope. Can J Anaesth. 2008;55:56–7.

GlideScope GVL and Cobalt user's manual & quick reference guide. Bothell: Verathon Inc; 2009–2011.

Kramer DC, Osborn IP. More maneuvers to facilitate tracheal intubation with the GlideScope. Can J Anaesth. 2006;53:737.

Lim HC. Utilization of a GlideScope videolaryngoscope for orotracheal intubations in different emergency airway management settings. Eur J Emerg Med. 2009;16:68–73.

Walls RW, Murphy MF, editors. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Endotracheal Tube Introducer (Bougie)

15

Joseph Rabinovich

15.1 Indications

- During orotracheal intubation, when only epiglottic visualization or partial glottic view is obtained during laryngoscopy.
- Particularly useful when neck mobility is limited, leading to inadequate visualization of the glottis (as in the case with cervical spine immobilization).
- When the glottic opening is narrowed either from pathological causes (burns, trauma, tumor, or other anatomical variation).
- When the direct view of the airway is very narrow, as with limited mouth opening or large tongue. In these scenarios, the endotracheal tube (ETT) can obstruct one's view of the cords during placement.

15.2 Contraindications

- When a failed airway occurs (three unsuccessful attempts at endotracheal intubation and inability to adequately oxygenate)
- When surgical airway is indicated (i.e., upper airway obstruction that prevents passage of the ETT via the orotracheal route)

15.3 Materials and Medications

- ETT introducer (bougie) (Fig. 15.1)
- Water-based lubricant
- Lubricated ETT 6 mm or larger *without* stylet (pediatric bougies are available that accommodate smaller ETTs)
- Standard orotracheal direct laryngoscopy (Miller or Macintosh blade) or video laryngoscopy setup
- Assistant

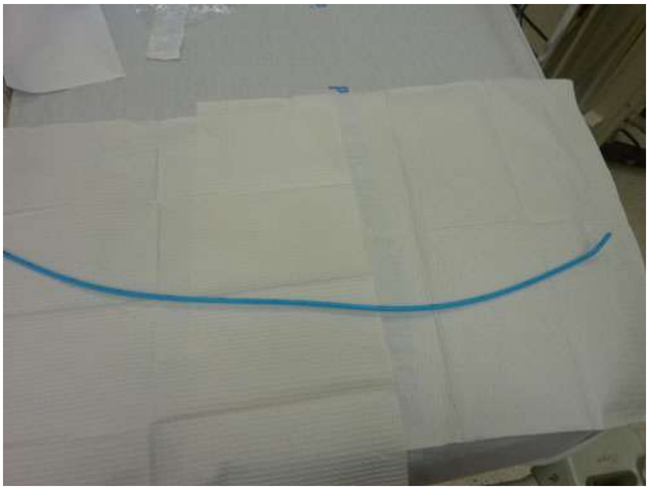

Fig. 15.1 Endotracheal tube introducer (bougie)

J. Rabinovich, MD
Department of Emergency Medicine,
Mount Sinai School of Medicine,
Elmhurst Hospital Center, Elmhurst, NY, USA
e-mail: joerab@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_15

15.4 Procedure

1. The operator first optimizes airway visualization (Fig. 15.2). (Bougie use should not be a substitute for poor technique.)
2. Without losing sight of the airway, the operator asks the assistant to hand him or her the bougie with the coudé tip facing up.
3. The operator directs the bougie tip underneath the epiglottis (Fig. 15.3).
4. Confirmation of placement can be done visually or by tactile sensation:
    (a) A ratchet-like sensation may be felt as the bougie tip is advanced into the airway and slides over the tracheal rings.
    (b) As the bougie is further advanced, the operator may feel the bougie rotate as it enters the bronchus and/or will get a "hold up," the most reliable sign that the bougie is in the trachea [1]. (The "hold-up" sign occurs when the Bougie encounters a terminal bronchus [typically at around 35 cm] and stops advancing.) [2]
5. Once placement is confirmed, the bougie needs to be partially withdrawn to about 25 cm at the lip line.
    (a) Some brands will have a thick black indicator line.
    (b) A sufficient amount of the bougie needs to extend out beyond the proximal end of the ETT.
6. While the operator holds the bougie in place, the assistant threads the ETT over the bougie (Fig. 15.4).
7. The operator now grasps the ETT in her or his right hand and advances it over the bougie.
8. Simultaneously, the assistant holds and stabilizes the proximal end of the bougie.
9. The ETT should be advanced to approximately 23 cm in males and 21 cm in females. The assistant removes the bougie as the operator holds the ETT in place (Fig. 15.5).
10. As the operator holds the ETT firmly in position, the assistant inflates the ETT balloon and withdraws the bougie.
11. Confirmation of proper ETT placement is achieved through traditional means (end-tidal CO_2 detection, auscultation of breath sounds).

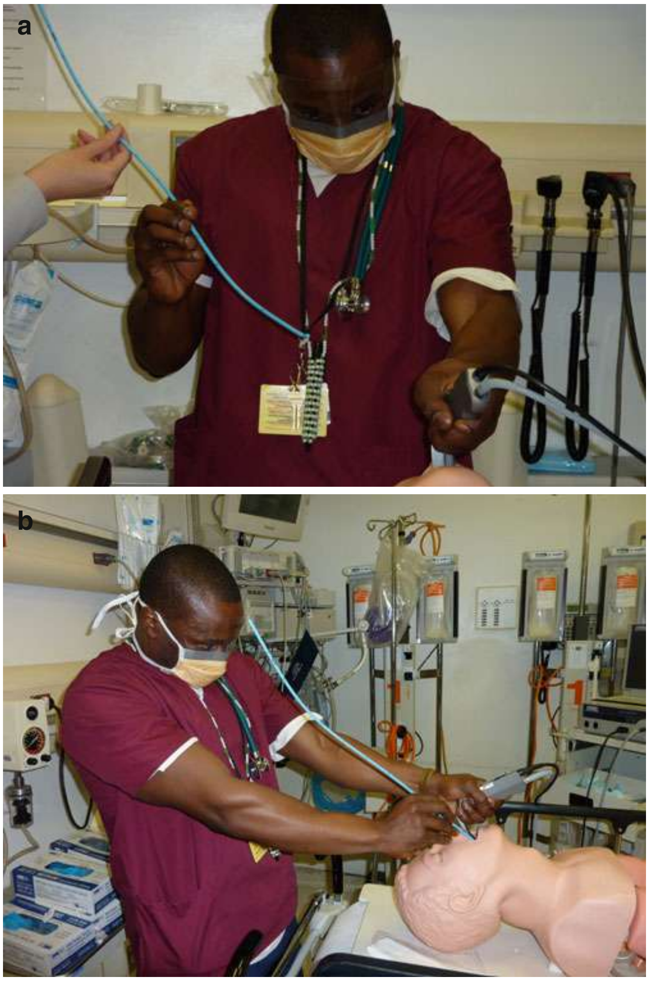

Fig. 15.2 (**a**) Assistant hands operator bougie with coudé tip directed upward, while operator maintains his focus on the target. (**b**) Bougie being placed, parallel with line of sight, underneath epiglottis

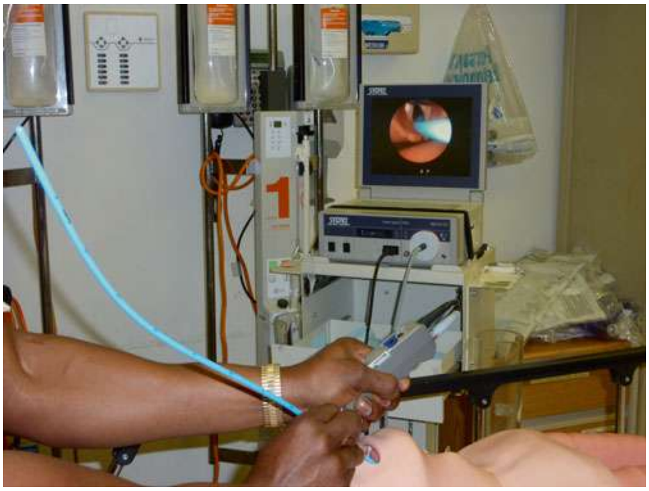

Fig. 15.3 Bougie can be placed into the glottis using direct laryngoscopy or video assistance

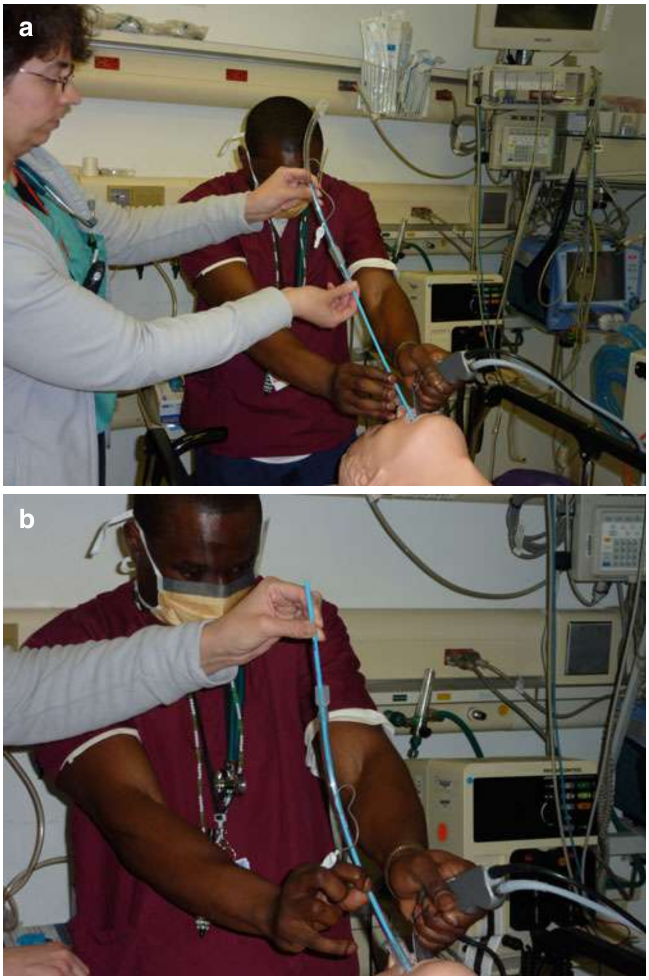

Fig. 15.4 (**a**) Assistant places ETT over bougie and operator withdraws bougie until it protrudes out the top of the ETT, (**b**) while assistant stabilizes the protruding portion of the bougie, the operator railroads the ETT into the airway. The operator continues to support the soft tissues with the laryngoscope blade to facilitate placement

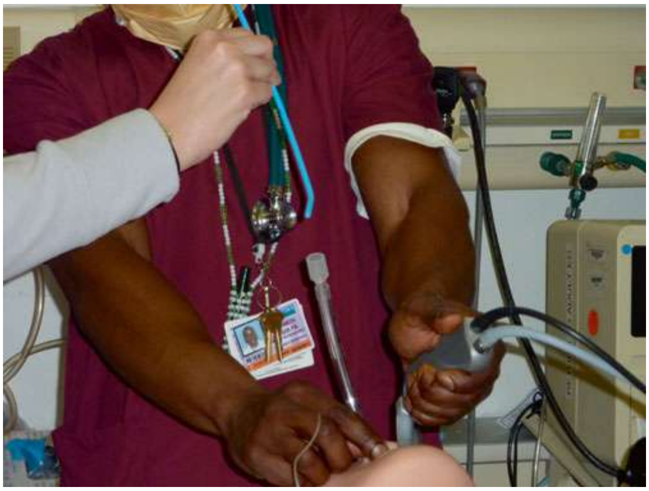

Fig. 15.5 The assistant removes the bougie while the operator stabilizes the endotracheal tube

15.5 Complications

- Trauma to the esophagus, larynx, trachea, or bronchus, including perforation [3, 4].
- In general, complications are rare.

15.6 Pearls and Pitfalls

- The line of sight should be as parallel as possible to the axis of the bougie as it is being passed, allowing better eye-hand coordination. This augments more accurate placement of the bougie tip.
- Maintain the view of the airway as the assistant hands the operator the bougie.
- Keep the laryngoscope in place to support the soft tissues, as the endotracheal tube is slid over the bougie, to facilitate placement.
- If resistance is met during passage, withdraw the ETT slightly (~2 cm) and rotate the ETT counterclockwise one-quarter turn (90°) and reattempt passage.
  - This changes the position of the leading edge of the ETT, which may catch on the posterior laryngeal inlet [2].
  - By rotating the ETT, the leading edge now is anterior facing and is less likely to catch on the arytenoid cartilage along with other laryngeal inlet structures.
- If encountering resistance to ETT placement, consider releasing cricoid pressure (if used).
- Measurement markings on the bougie are aligned with the coudé tip. If not sure of orientation and the loose site of the tip, use the markings to properly orient the tip.

References

1. Kidd JF, Dyson A, Latto IP. Successful difficult intubation. Use of the gum elastic bougie. Anaesthesia. 1988;43:437–8.
2. Murphy MF, Hung OR, Law JA. Tracheal intubation: tricks of the trade. Emerg Med Clin North Am. 2008;26:1001–14.
3. Kadry M, Popat M. Pharyngeal wall perforation – an unusual complication of blind intubation with a gum-elastic bougie. Anaesthesia. 1999;54:404–5.
4. Smith BL. Haemopneumothorax following bougie-assisted tracheal intubation. Anaesthesia. 1994;49:91.

Lighted Stylet Intubation

16

Benjamin M. Mahon and Lars K. Beattie

When lighted stylet intubation is done correctly, the procedure can be very safe, with very little difference in outcome from that of primary laryngoscopy. (*Note*: *several lighted stylet devices*, *such as the TrachlightTM and Light WandTM*, *are no longer being manufactured*, *but these devices are still in use*.)

16.1 Indications

- Difficult/impossible direct laryngoscopy [1, 2]
  - Congenital abnormalities of airway
  - High Mallampati grade [3]
  - Dental appliances
- Failed direct laryngoscopy

16.2 Contraindications

- Absolute
  - Morbid obesity
  - Airway foreign body
  - Expanding neck mass
- Relative
  - Abnormal airway anatomy
  - Airway lesions (e.g., abscess, mass, epiglottitis) that change oropharyngeal anatomy
  - Acute care where concomitant resuscitation requires a well-lit room
  - Lack of familiarity or experience with procedure
  - “Can’t oxygenate, can’t ventilate” situation

16.3 Materials and Medications

- Intravenous (IV) access, O_2, and monitor
- Ambu bag with supplemental oxygen
- Suction (Yankauer and tubing)
- Lighted stylet (LS)
- Endotracheal tube (ETT) 2.5-mm larger than LS with 10-cc syringe
- Surgilube
- Intubation medications (this procedure may be performed as an awake or a rapid-sequence intubation)

B.M. Mahon, MD
Ponciana Medical Center,
Kissimmee, FL, USA
e-mail: drfaust2k9@aol.com

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_16

16.4 Procedure

1. Preoxygenate.
2. Positioning.
    - (a) Sniffing position, pinna at the level of the sternal notch (Fig. 16.1).
    - (b) Skip sniffing position if cervical spine injury is suspected.
3. LS-ETT unit preparation.
    - (a) Insert the wire stylet into the device.
    - (b) Check the LS light.
    - (c) Lubricate the LS with K-Y Jelly.
    - (d) Position the LS just distal to the Murphy eye.
    - (e) Curve the LS to user preference at the line labeled "Bend Here."
4. Administer intubation medications.
5. Have an assistant to apply cricoid pressure.
6. Grasp and elevate the patient's jaw near the corner of the mouth with the operator's thumb, index, and middle fingers, elevating the tongue and epiglottis along with it.
7. Using the free hand, insert the LS-ETT unit into the oropharynx and advance (Fig. 16.2).
8. Use the midline glow in the neck to guide insertion of the LS-ETT (Fig. 16.3).
9. Bright light *below* the thyroid prominence indicates correct placement of the ETT tip.
10. Dim or blurred light or light at the thyroid prominence suggests incorrect positioning (Fig. 16.4).
11. If the transilluminated light is dim, off center, or not seen, esophageal positioning must be considered.
    - (a) Withdraw the LS-ETT unit approximately 2–5 cm.
    - (b) Reposition the patient's head and neck.
    - (c) Reattempt according to steps 5–8.
12. Placement of the ETT (Fig. 16.5).
    - (a) Hold the LS-ETT unit steady with one hand.
    - (b) Check the depth of the ETT and adjust accordingly.
    - (c) Release the LS latch that holds the ETT to the LS.
    - (d) While holding the ETT in position, gently slide the LS out from the ETT.
    - (e) Inflate the ETT balloon.
13. Confirm ETT placement (continuous end-tidal CO_2 [$EtCO_2$], colorimetric capnometry).
14. Secure the ETT.

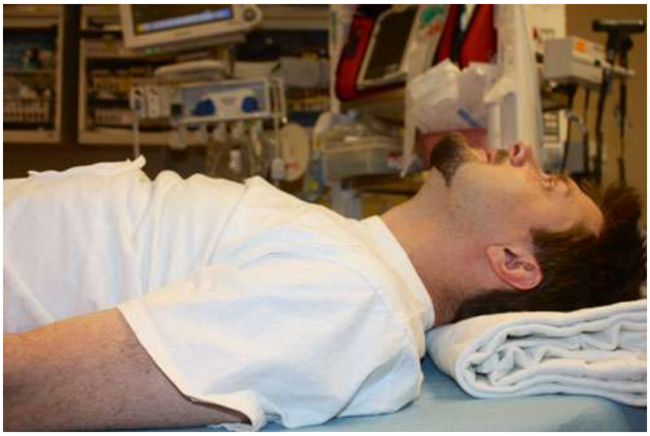

Fig. 16.1 Sniffing position, pinna at the level of the sternal notch

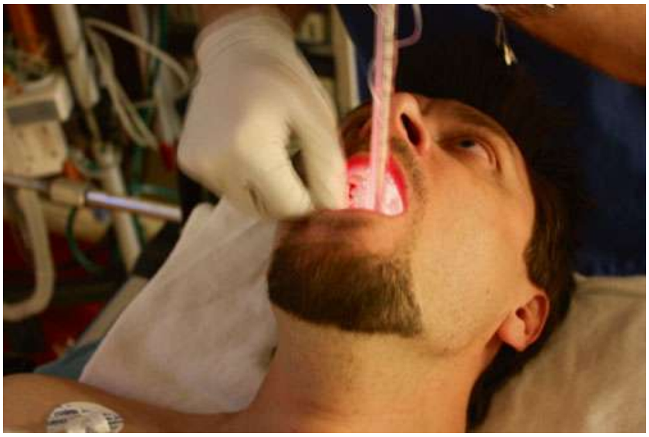

Fig. 16.2 Grasp and elevate the patient's jaw near the corner of the mouth with the operator's thumb, index, and middle fingers, elevating the tongue and epiglottis along with it. Using the free hand, insert the LS-ETT unit into the oropharynx and advance

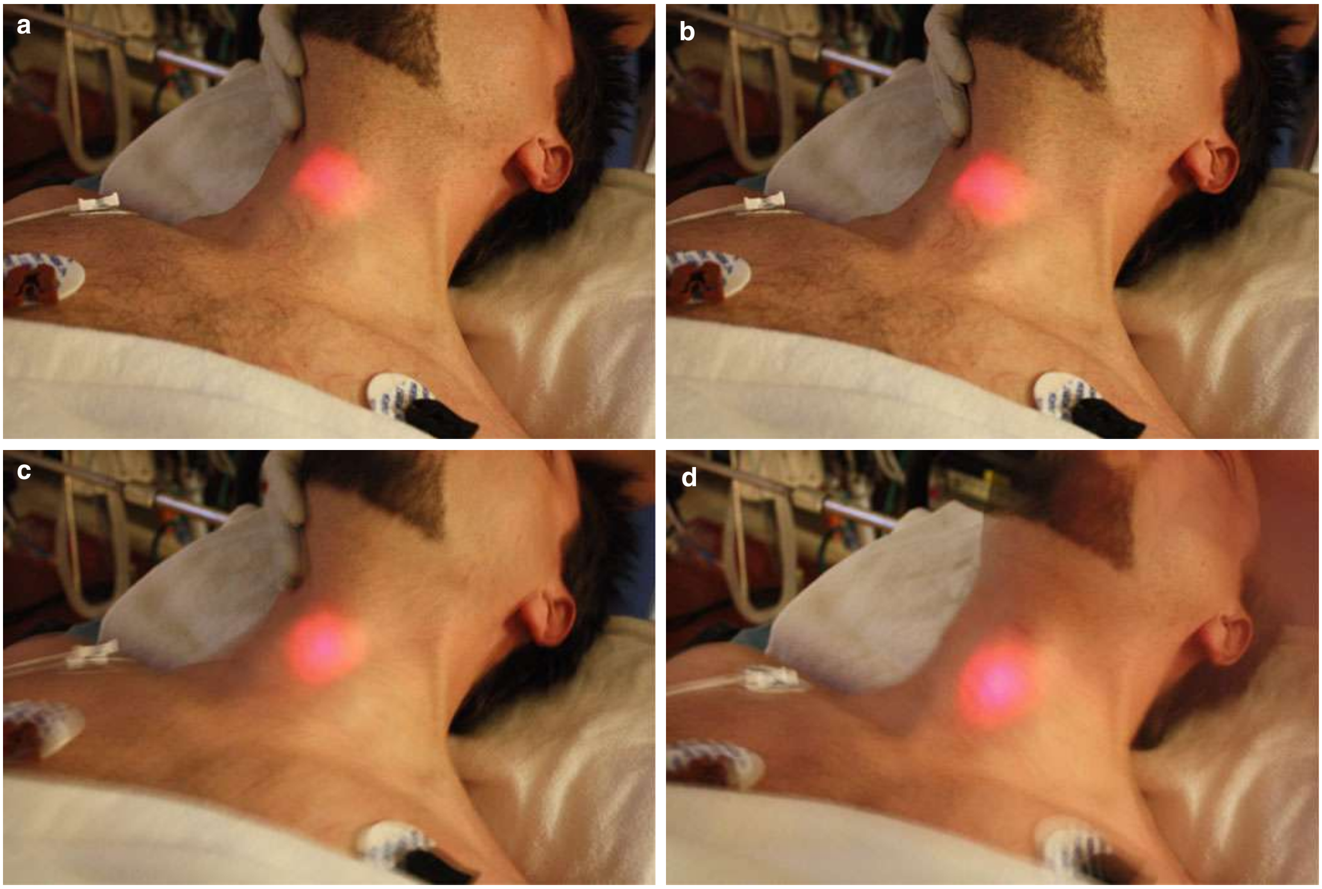

Fig. 16.3 (**a**–**c**) Use the midline glow in the neck to guide insertion of the LS-ETT. (**d**) Bright light *below* the thyroid prominence indicates correct placement of the ETT tip

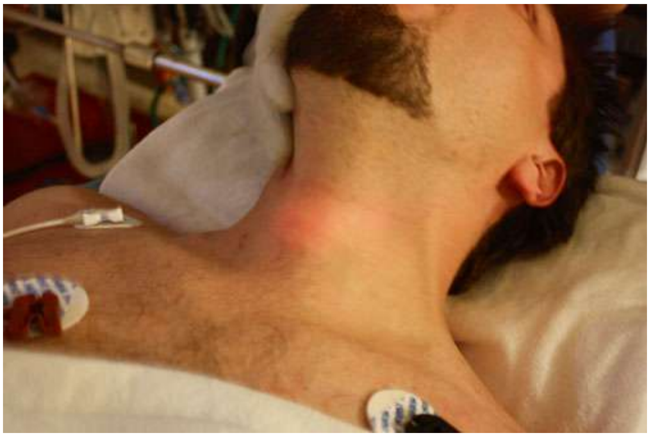

Fig. 16.4 Dim or blurred light or light at the thyroid prominence suggests incorrect positioning

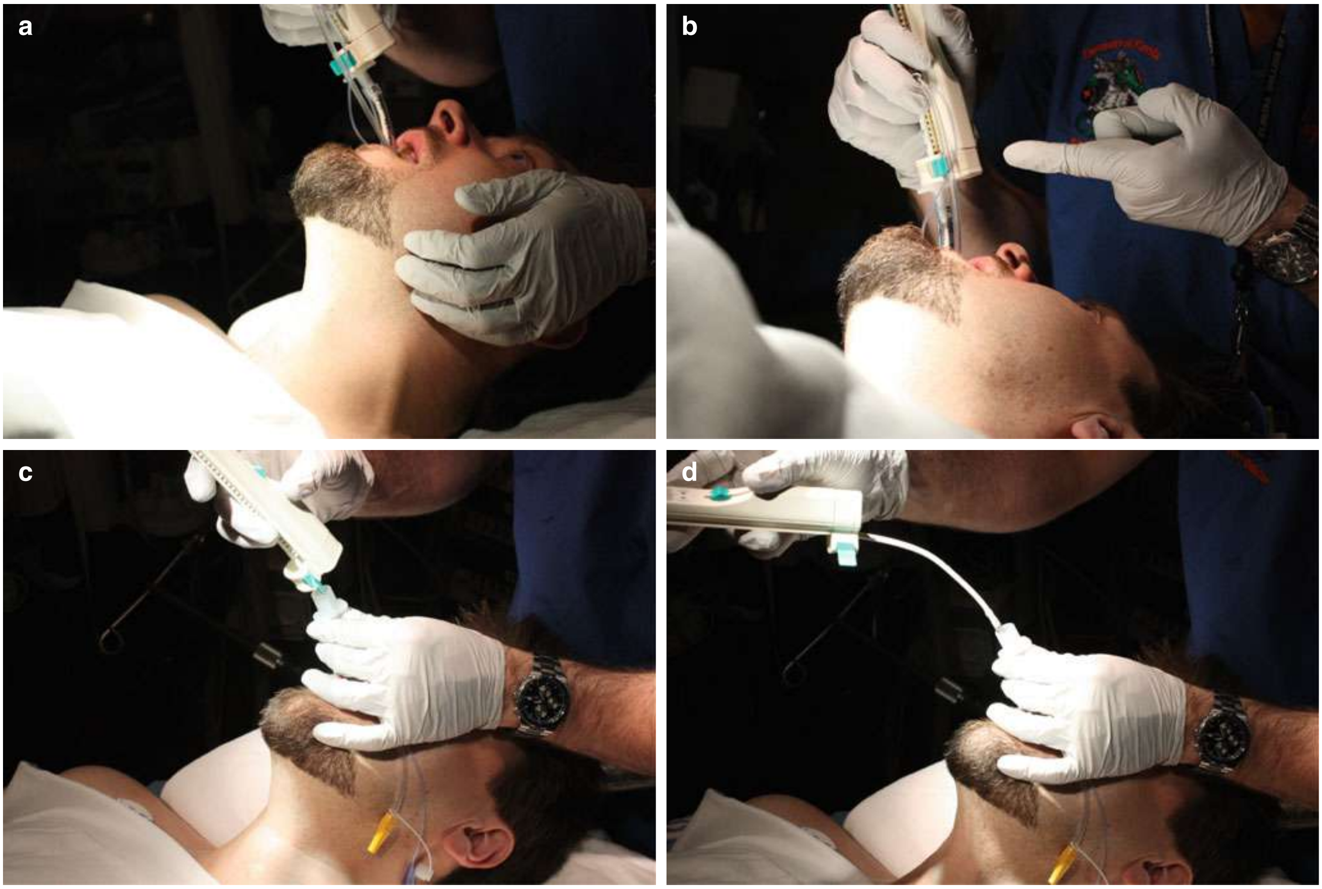

Fig. 16.5 (**a**) Hold the LS-ETT unit steady with one hand. (**b**) Check the depth of the ETT and adjust accordingly. (**c**) Release the LS latch that holds the ETT to the LS. (**d**) While holding the ETT in position, gently slide the LS out from the ETT

16.5 Pearls and Pitfalls

- Pearls
  - LS-ETT complex—Typically the classic “hockey-stick” shape with the 90° curve just proximal to the cuff is recommended [2].
  - Dimming the room lights will enhance transillumination.
  - Pulling the wire stylet out from the LS-ETT unit will make it more pliable and may facilitate its placement in the trachea and removal of the LS.
  - Some LS devices start to blink after 30 s to prevent bulb overheating.
  - The LS may be used with nasotracheal intubation, intubation through a laryngeal mask airway (LMA), or conventional laryngoscopy to enhance success.
- Pitfalls
  - LS intubation should not be used as an emergency airway alternative by a proceduralist unfamiliar with the technique:
    - It is technically complicated and more challenging than many other airway adjuncts in the standard difficult airway algorithm.
    - One study compared the use of four rescue airway devices in the difficult airway algorithm. A success rate of only 20 % was achieved with the Trachlight™ on the first attempt when in the hands of the novice physician when used as a rescue device in their difficult airway algorithm [4].
  - In very thin patients, transillumination may be visualized quite well even when the LS-ETT unit is in the esophagus:
    - When the unit is in the esophagus, typically the light it will be more diffuse.
    - When the unit is in the trachea, the transilluminated area will be well circumscribed.
  - In obese patients or patients with significant neck tissue, the transilluminated light from the LS-ETT unit may be dim despite correct positioning in the trachea.

References

1. Agro F, Hung OR, Cataldo R, Carassiti M, Gherardi S. Lightwand intubation using the Trachlight: a brief review of current knowledge. Can J Anaesth. 2001;48:592–9.
2. Davis L, Cook-Sather SD, Schreiner MS. Lighted stylet tracheal intubation: a review. Anesth Analg. 2000;90:745–56.
3. Rhee KY, Lee JR, Kim J, Park S, Kwon WK, Han S. A comparison of lighted stylet (Surch-Lite) and direct laryngoscopic intubation in patients with high Mallampati scores. Anesth Analg. 2009;108:1215–9.
4. Aikins NL, Ganesh R, Springmann KE, Lunn JJ, Solis-Keus J. Difficult airway management and the novice physician. J Emerg Trauma Shock. 2010;3:9–12.

Selected Reading

Langeron O, Birenbaum A, Amour J. Airway management in trauma. Minerva Anestesiol. 2009;75:307–11.

Walls RM, Murphy MF. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. Chap. 11.

Fiber-Optic Stylet Intubation (Rigid and Semirigid)

17

Joseph Rabinovich

17.1 Indications

- For use in routine and predicted difficult oral intubations.
- Similar to a flexible fiber-optic scope with the specific advantages of:
  - Less setup time
  - Less time to perform the procedure
  - Appropriate for routine intubations (and easier to accumulate experience)
  - Rigid enough to lift up the epiglottis
  - Easier to navigate through tissue
  - Less susceptible to being obscured by blood and secretions
  - More durable, more portable, easier to clean, and less expensive
- Particularly useful when neck mobility or mouth opening is restricted.
- Advantageous in awake intubations because it can minimize tissue contact, resulting in less stimulation to the patient's airway and better tolerance.
- Certain stylets can be used to intubate through supraglottic airways such as laryngeal mask airways (LMAs).

17.2 Contraindications

- Complete upper airway obstruction where surgical airway is indicated
- Oral pharyngeal swelling requiring a nasotracheal or surgical approach
- Failed airway and unable to adequately maintain oxygenation

17.3 Relative Contraindications

- Large amounts of blood and secretions may obscure visualization of the airway and cords.
- Very distorted airways. Compared with flexible endoscopy, this device is less maneuverable.

17.4 Materials and Medications

- Endotracheal tube (ETT) 5.5 mm or greater (Fig. 17.1) (Pediatric stylets are also available.)
- Water-soluble lubricant
- Defogging agent
- Optional: Swivel adaptor and meconium aspirator

J. Rabinovich, MD
Department of Emergency Medicine,
Mount Sinai School of Medicine, Elmhurst Hospital Center,
Elmhurst, NY, USA
e-mail: joerab@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_17

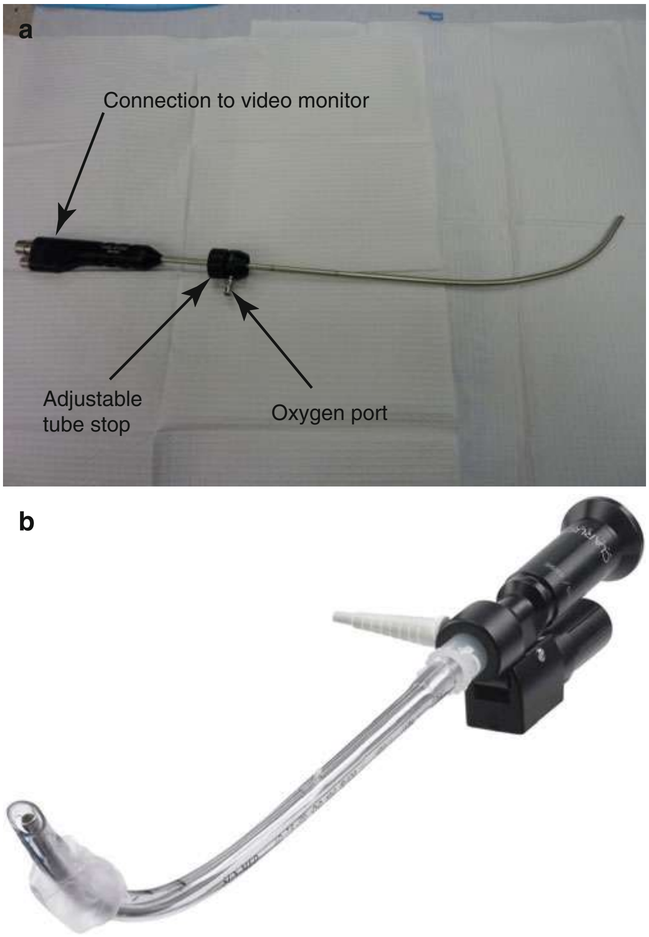

Fig. 17.1 (**a**) Bonfils rigid fiber-optic stylet, Karl Storz endoscopy, (**b**) Levitan FPS malleable fiber-optic stylet

17.5 Procedure

1. Use standard preparation for rapid sequence intubation (RSI) or for awake intubation.
2. Place the ETT over the stylet.
    (a) The ETT should extend slightly beyond the stylet tip.
    (b) If using the malleable stylet:
        (i) Without a laryngoscope: Bend tip to about 70° [1].
        (ii) With a laryngoscope: Bend tip to 35° [1].
    (c) Lubricate the tip of the ETT.
3. Depending on specific scope capability, connect oxygen tubing to the port on the scope.
    (a) This keeps secretions away from the tip while providing an oxygen source.
    (b) Keep flow less than 6 L/min [2].
4. Lens fogging prevention.
    (a) Warm the tip of the scope with the hand or immerse the tip in warm saline.
    (b) Apply the defogging agent.
    (c) Alternatively, chlorhexidine is an effective defogger.
5. Scope insertion.
    (a) Use the nondominant hand to pull the jaw forward while holding the tongue.
        (i) In an awake patient, have the patient protrude the tongue and the operator grasps it with 4×4 gauze. (Alternatively, a Macintosh laryngoscope blade can be used.)
        (ii) The goal is to move the base of the tongue off the posterior pharyngeal wall.
    (b) Initially position the scope horizontally and to the right of the patient's mouth.
    (c) Once the tip is in the oropharynx, reposition the scope vertically (Fig. 17.2a).
    (d) The scope tip should be in the midline or in the retromolar position (per scope design).
    (e) Position the scope and the tip of the ETT in front of the uvula.
    (f) Refer to the eyepiece or video screen to see if there is a clear image of the uvula (Fig. 17.2b).
    (g) Advance the scope very slowly to maintain a view of landmarks, avoiding tissue contact (Fig. 17.3).
6. Once the epiglottis is visualized:
    (a) Continue to advance slowly.
    (b) To get underneath the epiglottis, the tip of the scope may need to be moved posteriorly (by tilting the operator's hand slightly forward) (Fig. 17.4).

7. Once underneath the epiglottis, tilt the scope back, to advance into the more anterior directed airway.
   (a) Make sure the glottic opening is well centered on the screen to facilitate placement (Fig. 17.5).
   (b) If resistance is felt, operator may need to rotate the scope clockwise, or tip scope slightly forward, to prevent the ETT from abutting the anterior aspect of the trachea.
   (c) The ETT may need to be advanced off the rigid stylet to allow further advancement.
8. To remove scope:
   (a) Twist the proximal end of the ETT clockwise.
   (b) Stabilize the tube with the nondominant hand.
   (c) Use the dominant hand to pull the scope forward, following the curvature of the stylet (Fig. 17.6).
   (d) An assistant may be of use during this step.

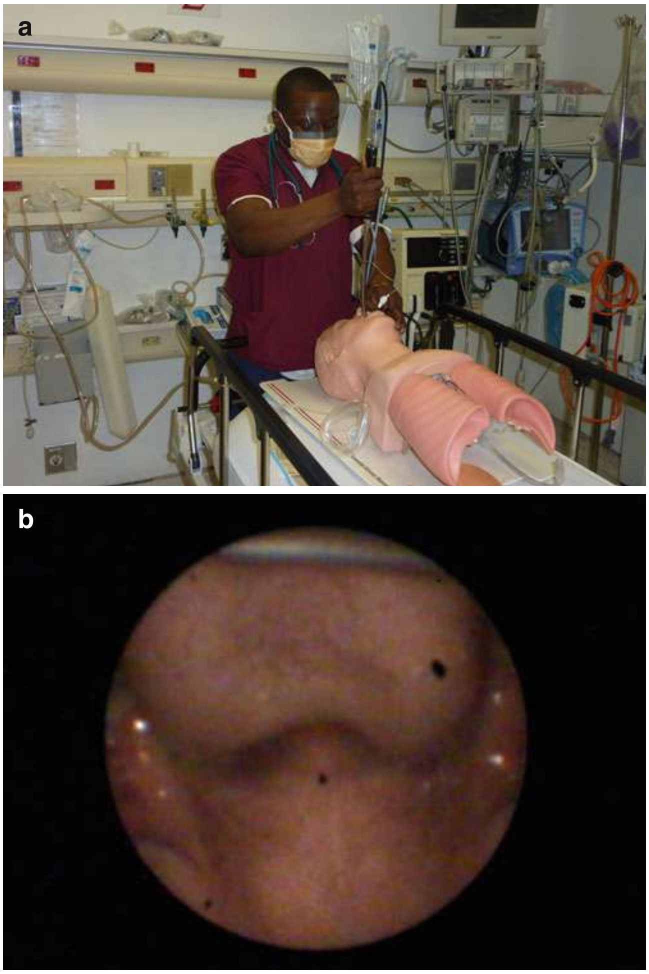

Fig. 17.2 (**a**) Initial placement of scope, under direct visualization; (**b**) tip is positioned in from the uvula

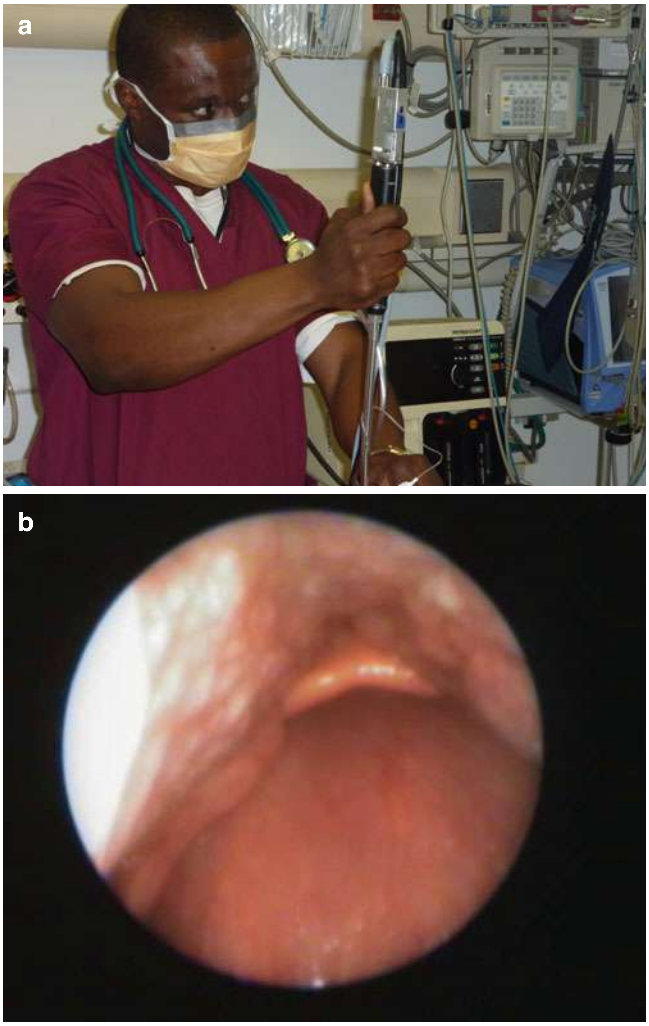

Fig. 17.3 (**a**) Using the video monitor, or through an eyepiece, the operator advances to the next landmark, (**b**) the epiglottis

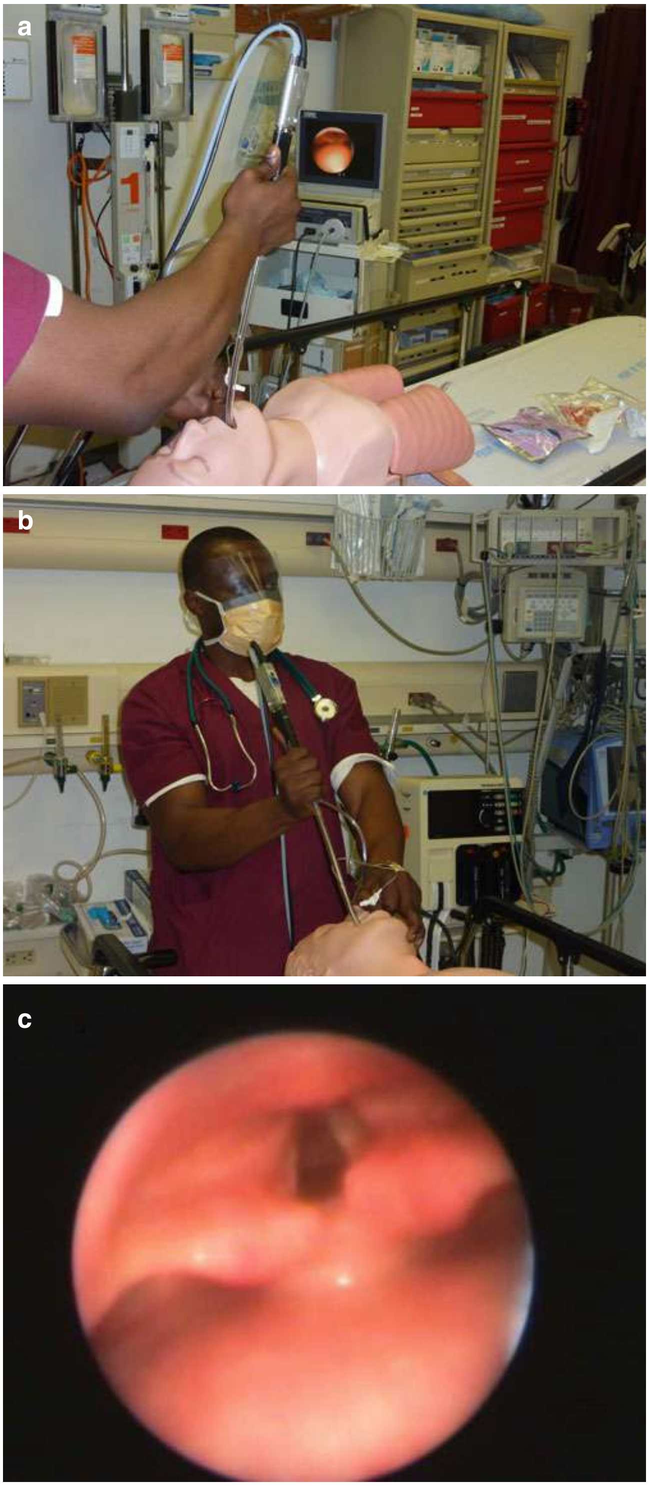

Fig. 17.4 (**a**) Operator tips scope forward to get underneath the epiglottis while advancing; (**b**) once under the epiglottis, the scope may need to be tipped back to advance to the glottic opening; (**c**) operator should try to keep the glottic opening in the center of the screen

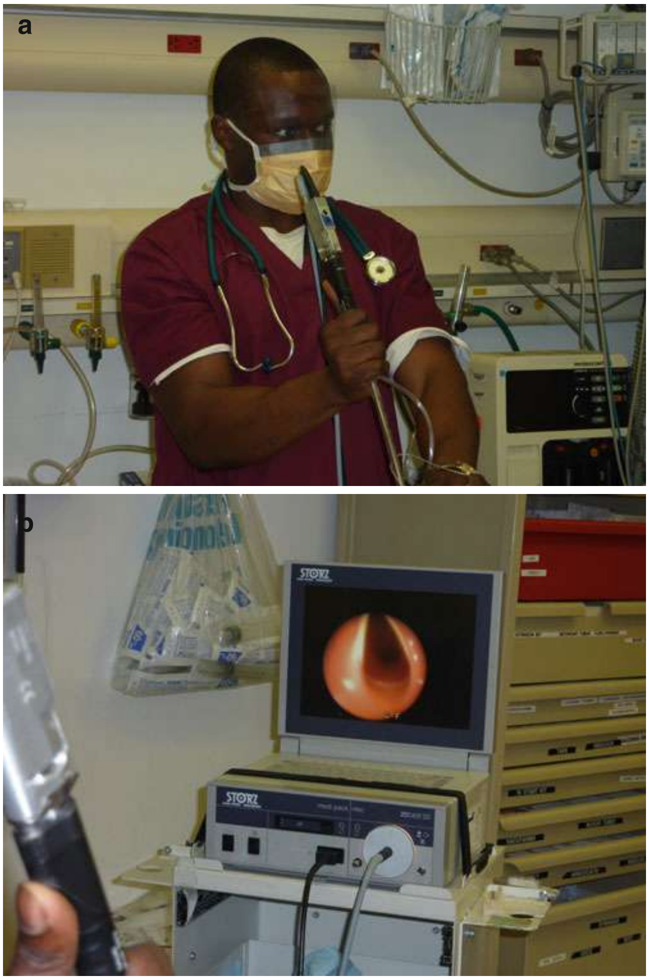

Fig. 17.5 (**a**) Operator is advancing the scope, (**b**) through the glottic opening keeping the image centered

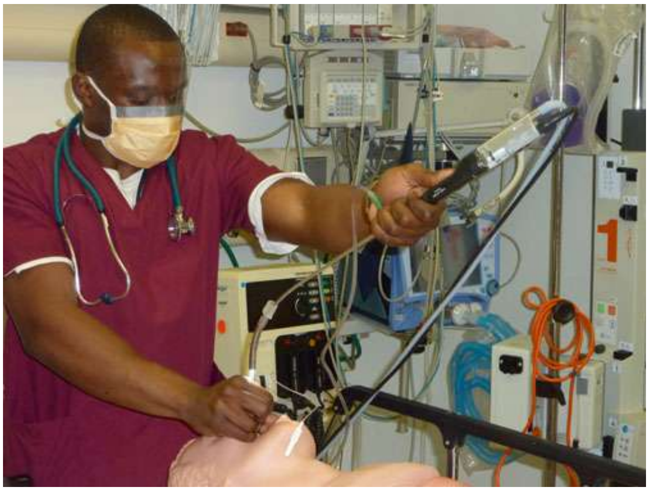

Fig. 17.6 To remove scope, operator must pull it forward while stabilizing the endotracheal tube

17.6 Pearls and Pitfalls

- Pearls
  - When the operator loses perspective or a clear view, withdraw the scope back to the point where identifiable structures are visualized and proceed.
  - The operator can suction through the scope by attaching a swivel adaptor and a meconium aspirator (Fig. 17.7) [3].
- Pitfalls
  - If the scope is advanced too quickly, orientation can be lost.
  - Structures that are too close to the scope will become blurred and unidentifiable.
  - If the scope tip abuts pharyngeal tissue, visualization can become blurred.
  - Flow greater than 6 L/min connected to the oxygen port may result in subcutaneous emphysema (single case report) [2].

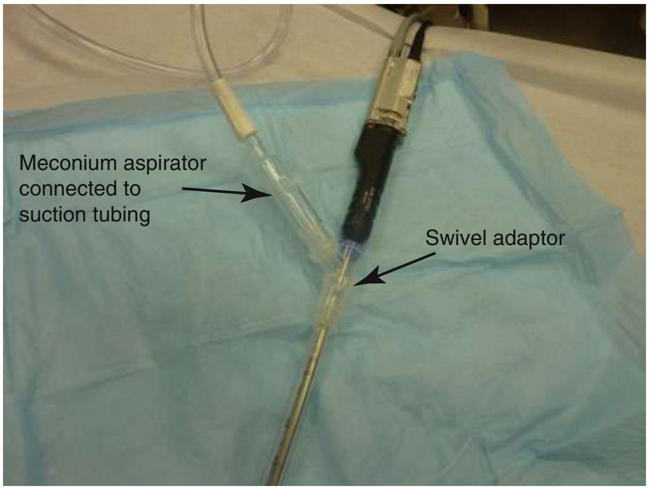

Fig. 17.7 By placing the stylet through the rubber valve of a swivel adaptor, which is then connected to suction via a neonatal meconium aspirator, the operator can now suction through the endotracheal tube

References

1. Levitan RM. Design rationale and intended use of a short optical stylet for routine fiberoptic augmentation of emergency laryngoscopy. Am J Emerg Med. 2006;24:490–5.
2. Hemmerling TM, Bracco D. Subcutaneous cervical and facial emphysema with the use of the Bonfils fiberscope and high-flow oxygen insufflation. Anesth Analg. 2008;106:260–2.
3. Weingart SD, Bhagwan SD. A novel set-up to allow suctioning during direct endotracheal and fiberscope intubation. J Clin Anesth. 2011;23:518–9.

18

Storz Video Laryngoscope

Joseph Rabinovich

18.1 Indications

- Orotracheal intubation for both routine and predicted difficult airways.
- Teaching traditional direct laryngoscopy to novice intubators.
- Ideal for unanticipated difficult airway with the option of intubating indirectly if an adequate direct view is unobtainable.
- An excellent tool when cervical spine precautions need to be taken: Because the video view of the airway is generated by a camera at the tip of laryngoscope blade, less manipulation is required for optimal glottic views.

18.2 Contraindications

- Absolute
  - When orotracheal intubation is contraindicated, e.g., for massive facial trauma, complete upper airway obstruction precluding orotracheal access to the airway
  - In a failed airway (three unsuccessful attempts with inability to maintain adequate oxygenation)
- Relative
  - Blood or copious secretions may prevent indirect viewing of the airway but does not always preclude the use of this device.

J. Rabinovich, MD
Department of Emergency Medicine,
Mount Sinai School of Medicine,
Elmhurst Hospital Center, Elmhurst, NY, USA
e-mail: joerab@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_18

18.3 Materials and Medications (Fig. 18.1)

- Standard materials and medications for endotracheal intubation. Operator should have a backup laryngoscope in case of equipment failure.
- Endotracheal tube (ETT) with or without stylet.
- Water-based lubricant.
- Antifogging agent (not required for C-Mac).

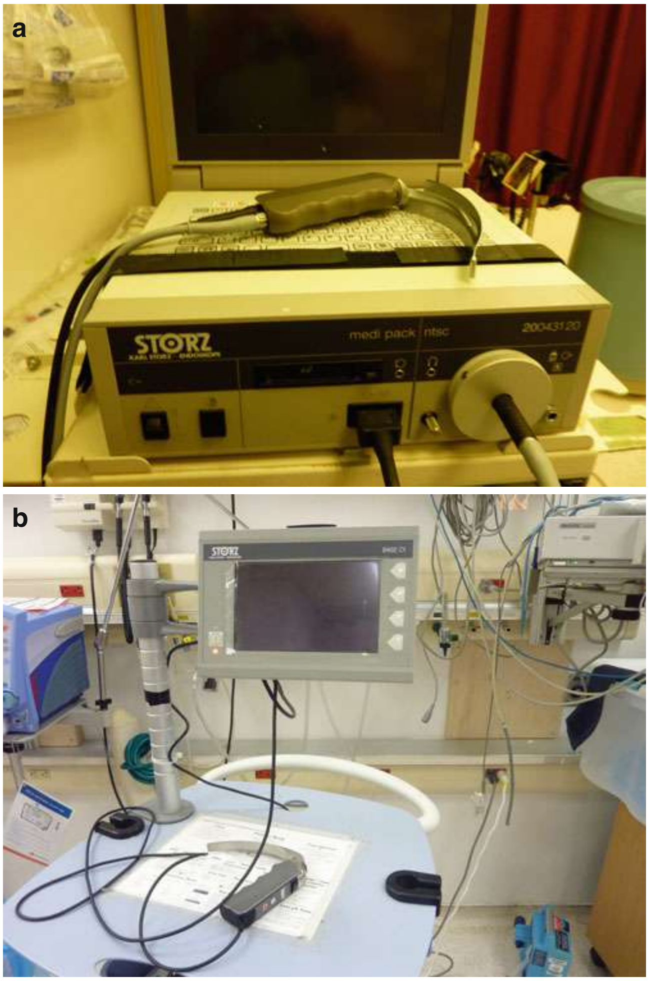

Fig. 18.1 (**a**) Storz video laryngoscope (older version), (**b**) Storz C-Mac (newer version)

18.4 Procedure

1. Standard preparation for orotracheal intubation. If there are no cervical spine precautions, then align the external auditory meatus with the sternal notch [1].
2. Apply antifogging drops to lens at tip of the blade, and/or hold the hand over the blade tip to warm it up to body temperature (older V-Mac model) (Fig. 18.2a).
3. Because blade geometry is the same as in standard laryngoscopes, the insertion technique is identical to that of standard laryngoscopy with a Macintosh blade (Fig. 18.2b).
4. Obtain the best direct view possible.
5. Consider the addition of the backward-upward-rightward pressure (BURP) maneuver [2].
6. Airway maneuvers may be performed by the operator or by the assistant using the video screen as a guide along with operator feedback (Fig. 18.3).
7. The operator has the option of intubating directly with adequate view or indirectly if visualization is improved.
8. Consider using an ETT introducer (bougie—see Chap. 15) if the view is inadequate.
9. Place the ETT, with or without a stylet, into the airway under direct or indirect visualization. If using a stylet, bend the distal end of the tube to approximately 35° as for a standard intubation (Fig. 18.4) [3].
10. Remove the stylet, inflate the cuff of the ETT using a syringe, and confirm placement with end-tidal CO_2, capnography, or colorimetry.

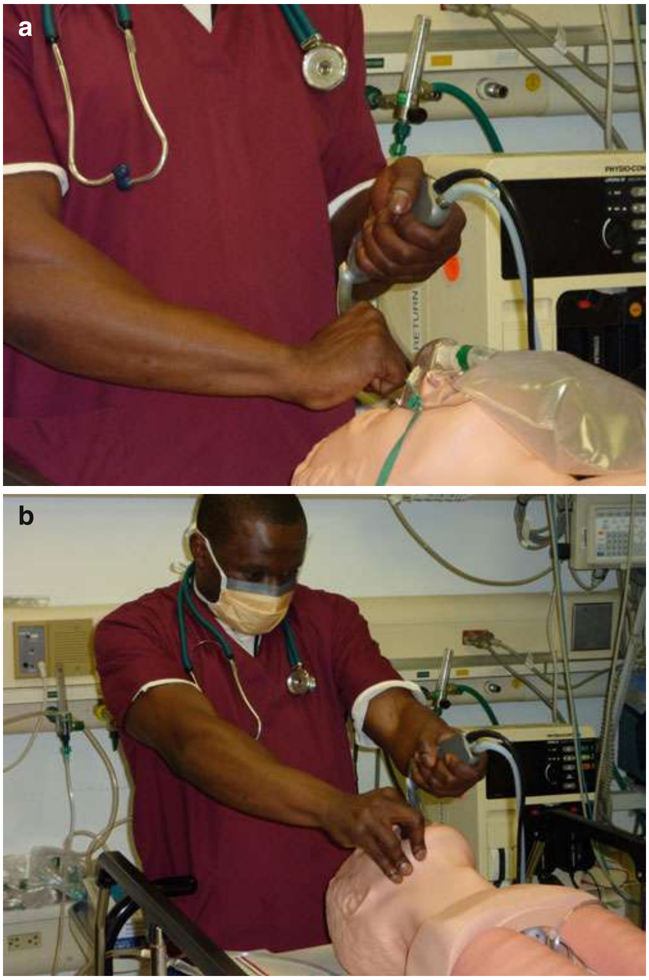

Fig. 18.2 (**a**) Operator warming the blade to prevent fogging, (**b**) laryngoscope blade insertion using the standard direct technique. Operator visually places the blade and optimizes the glottic view

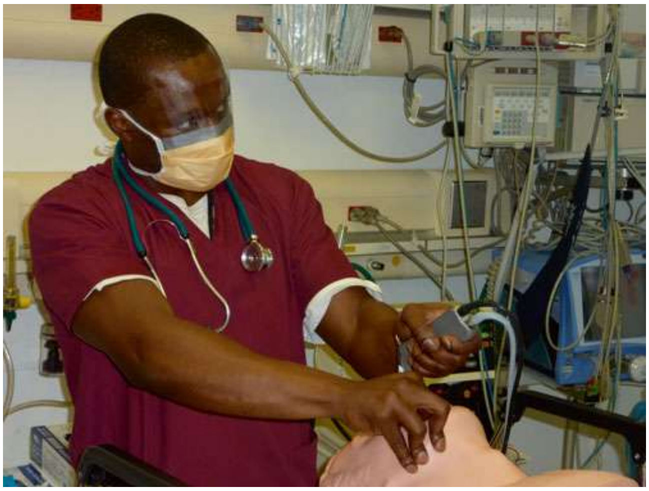

Fig. 18.3 Operator has the choice of using direct visualization (preferred when a novice intubator is learning laryngoscopy) or the indirect video view of the glottis

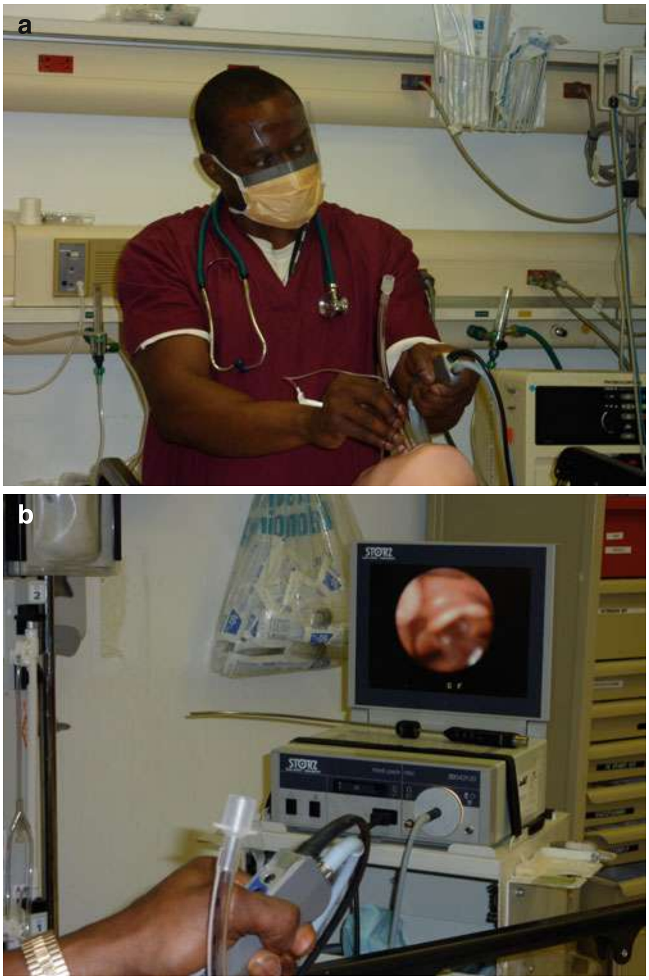

Fig. 18.4 (**a**) Operator initially places the endotracheal tube (ETT) into the oropharynx using direct visualization to avoid injury; (**b**) the tip of the ETT passing through the cords can be confirmed by direct visualization or by watching the video image

18.5 Complications (As with any Direct Laryngoscopy)

- Dental trauma
- Oropharyngeal trauma
- Vocal cord injury

18.6 Pearls and Pitfalls

- The initial placement of the laryngoscope blade and ETT should be done under direct visualization to avoid trauma to the oropharyngeal structures.
- As with direct laryngoscopy, the ETT should not be placed blindly, and the ETT must be seen to pass through the vocal cords to avoid placement in the esophagus.
- View can be obscured by secretions or fogging. If this occurs, the operator may need to remove the blade, wipe it down, and reinsert. The newer model, the C-Mac, is less likely to fog owing to design improvements.
- Observing the ETT pass through the vocal cords can sometimes be easier using the video image because the camera is angled to provide the most optimal view.
- Placement of the blade into the oropharynx can be awkward because the handle of laryngoscope is bulkier than a standard laryngoscope. Its handle is larger and has cables attached to its base. In patients with large anteroposterior diameter chests, the operator may need to rotate the laryngoscope handle toward the right corner of the mouth in order to introduce the blade into the oral cavity and then rotate it back to the proper position.
- These devices are ideal for teaching laryngoscopy. With same blade geometry, the technique is the same as with standard laryngoscopy. As the operator attempts intubation directly, the instructor can observe on the video screen and guide the student. The instructor will also be able to visually confirm that the ETT is entering the trachea.

References

1. Greenland KB, Edwards MJ, Hutton NJ, Challis VJ, Irwin MG, Sleigh JW. Changes in airway configuration with different head and neck positions using magnetic resonance imaging of normal airways: a new concept with possible clinical applications. Br J Anaesth. 2010;105:683–90.
2. Knill RL. Difficult laryngoscopy made easy with a "BURP.". Can J Anaesth. 1993;40:798–9.
3. Levitan RM, Heitz JW, Sweeney M, Cooper RM. The complexities of tracheal intubation with direct laryngoscopy and alternative intubation devices. Ann Emerg Med. 2011;57:240–7.

Selected Reading

Brown 3rd CA, Bair AE, Pallin DJ, et al. Improved glottic exposure with the video Macintosh laryngoscope in adult emergency department tracheal intubations. Ann Emerg Med. 2010;56:83–8.

Niforopoulou P, Pantazopoulos I, Demestiha T, Koudouna E, Xanthos T. Video-laryngoscopes in the adult airway management: a topical review of the literature. Acta Anaesthesiol Scand. 2010;54:1050–61.

Cricothyroidotomy

19

Henry Young II, Shannon Toohey, Bharath Chakravarthy,
and Lars K. Beattie

Up to seven intubation attempts in 1000 end up in a "can't intubate/can't ventilate" situation in the emergency department. These are considered failed airways that may require a surgical airway to maintain ventilation and oxygenation.

19.1 Indications

- Endotracheal tube (ETT) placement attempts unsuccessful
- Failed bag valve mask, laryngeal mask airway, or Combitube ventilation
- Severe facial trauma affecting the upper airway
- Severe oropharyngeal hemorrhage or profound emesis
- Obstruction (foreign body, mass, mass effect)

19.2 Contraindications

- Airway protection achievable using a less invasive strategy
- Tracheal transaction
- Pediatric patients younger than 8 years

19.3 Techniques

- Scalpel-bougie – minimalist
- Scalpel-Trousseau – standard

19.3.1 Scalpel-Bougie

19.3.1.1 Materials and Medications (Fig. 19.1)

- Betadine or chlorhexidine
- Scalpel #11 blade
- ETT (≥6 cm)
- Bougie
- Surgilube
- Bag valve mask

H. Young II, MD
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: hyoungii@ufl.edu

S. Toohey, MD
Department of Emergency Medicine, University of California
Irvine Medical Center, Orange, CA, USA
e-mail: stoohey@uci.edu

B. Chakravarthy, MD, MPH
Department of Emergency Medicine,
University of California Irvine, Orange, CA, USA
e-mail: bchakrav@uci.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_19

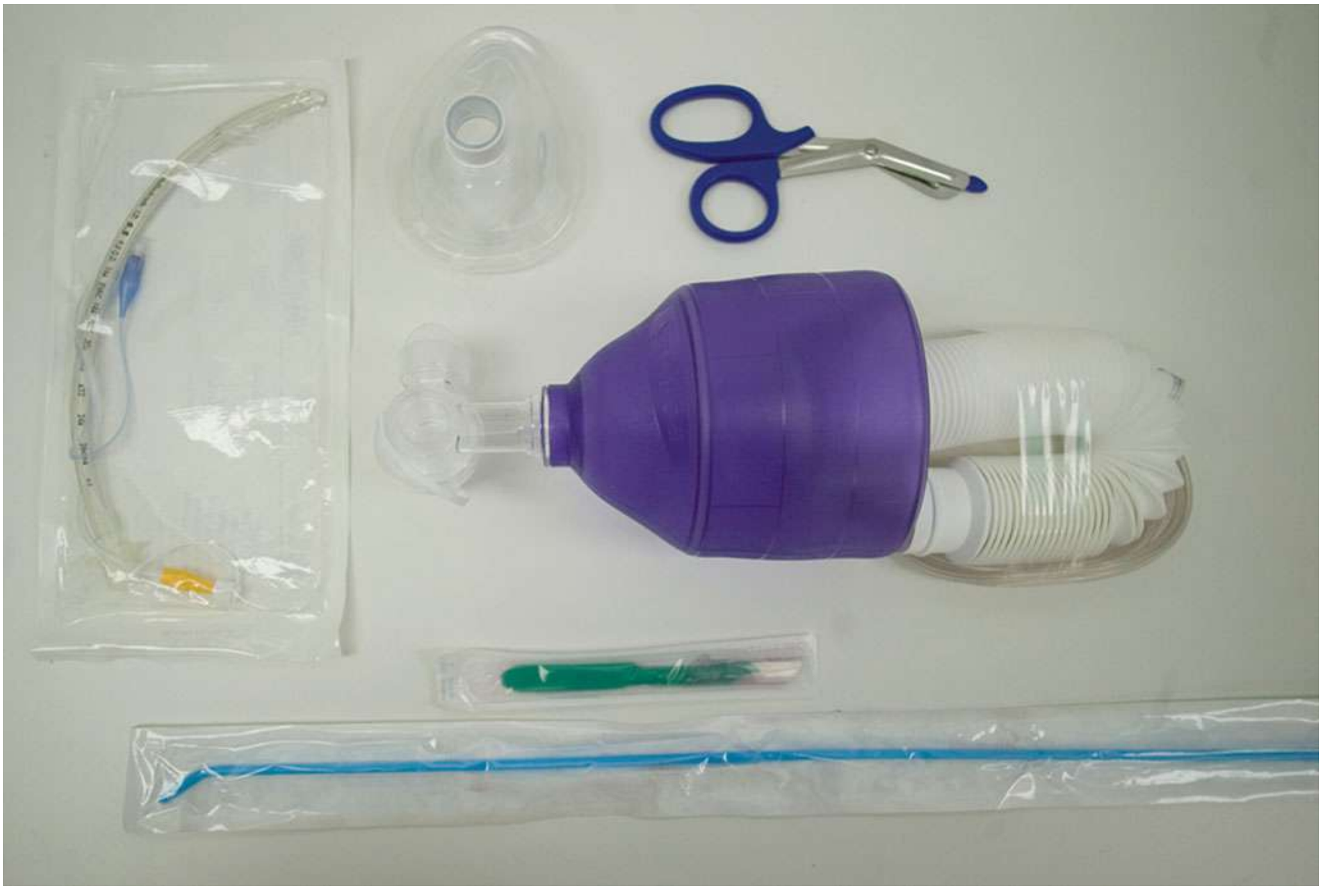

Fig. 19.1 *Right to left*, *top to bottom*: ETT (≥6 cm); bag valve mask; trauma shears; scalpel #11 blade; bougie

19.3.1.2 Procedure

1. Apply topical antiseptic.
2. Remove the 15-mm ETT ventilator connector from the ETT end.
3. Place copious Surgilube on the bougie and railroad over the end of the bougie.
4. Palpate the thyroid notch, cricothyroid membrane, and hyoid bone for orientation (Fig. 19.2a).
5. Stabilize the thyroid cartilage between the thumb and the middle finger of the nondominant hand.
6. Make a vertical skin incision (2–3 cm) over the cricothyroid membrane (Fig. 19.2b).
7. Use the index finger to palpate the cricothyroid membrane.
8. Turn the scalpel 90° and make a 1.5-cm horizontal incision through the lower half of the cricothyroid membrane (Fig. 19.3a).
9. With the scalpel still in the incision, turn it 90°, and insert the bougie into the incision, using the blade as a guide (Fig. 19.3b).
10. Advance the bougie caudally 5–6 cm. Stop if resistance is encountered.
11. Slide the ETT over the bougie into the incision (Fig. 19.4a).
12. Inflate the ETT cuff and ventilate the patient (Fig. 19.4b).
13. Verify the position of the ETT via auscultation, end-tidal CO_2 ($EtCO_2$), and chest radiograph.
14. Secure the ETT.

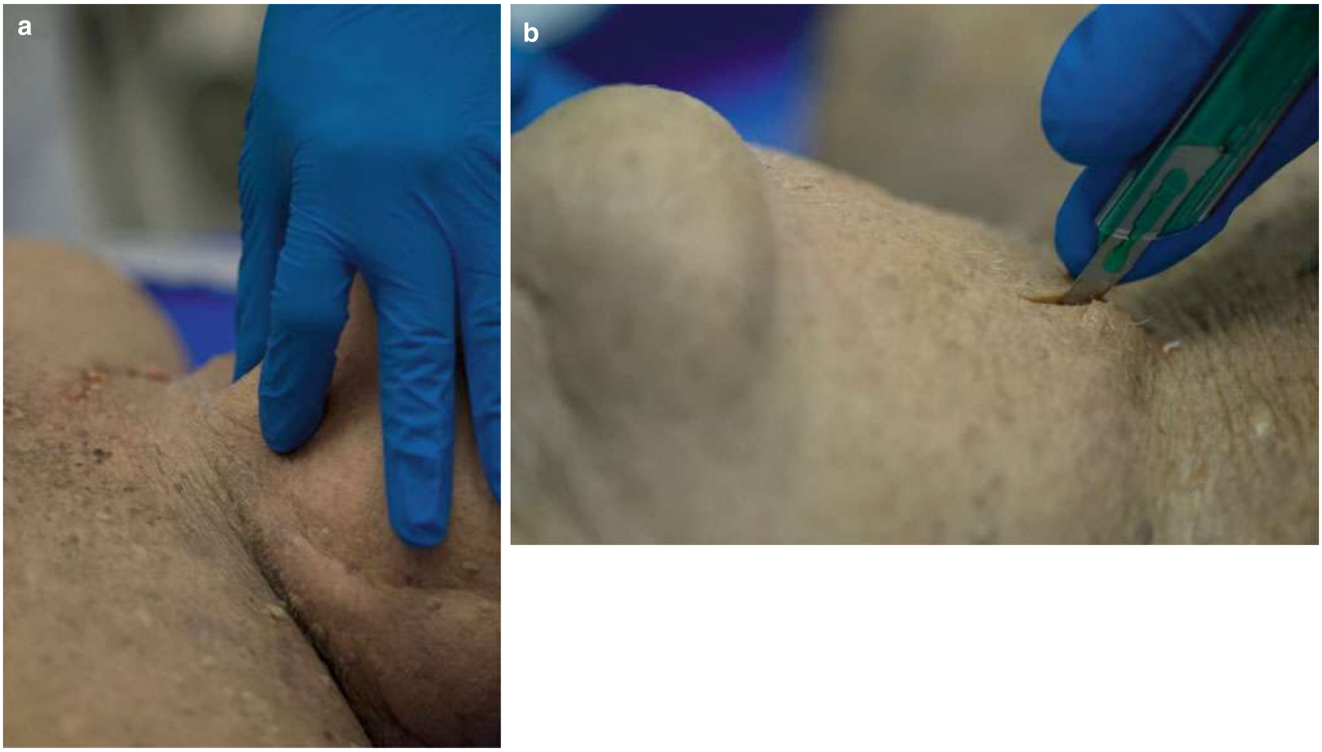

Fig. 19.2 (**a**) Palpate and stabilize the thyroid cartilage, (**b**) make a vertical 2–3 cm incision over the cricothyroid membrane

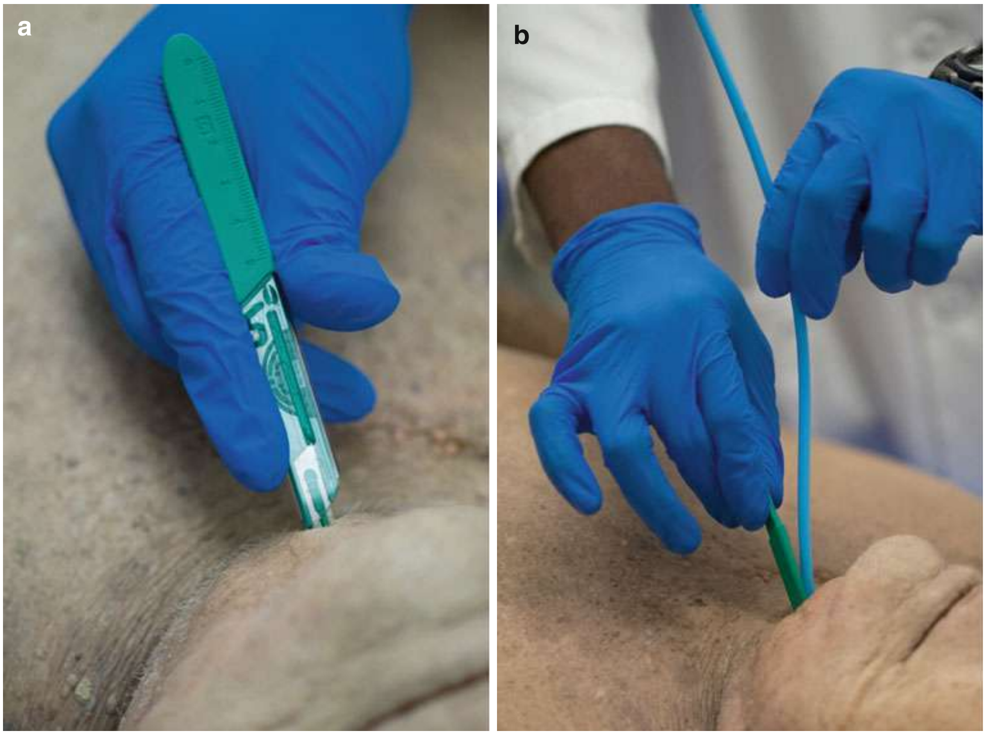

Fig. 19.3 (**a**) Make a horizontal incision into the cricothyroid membrane, (**b**) insert the bougie into the incision made in the cricothyroid membrane

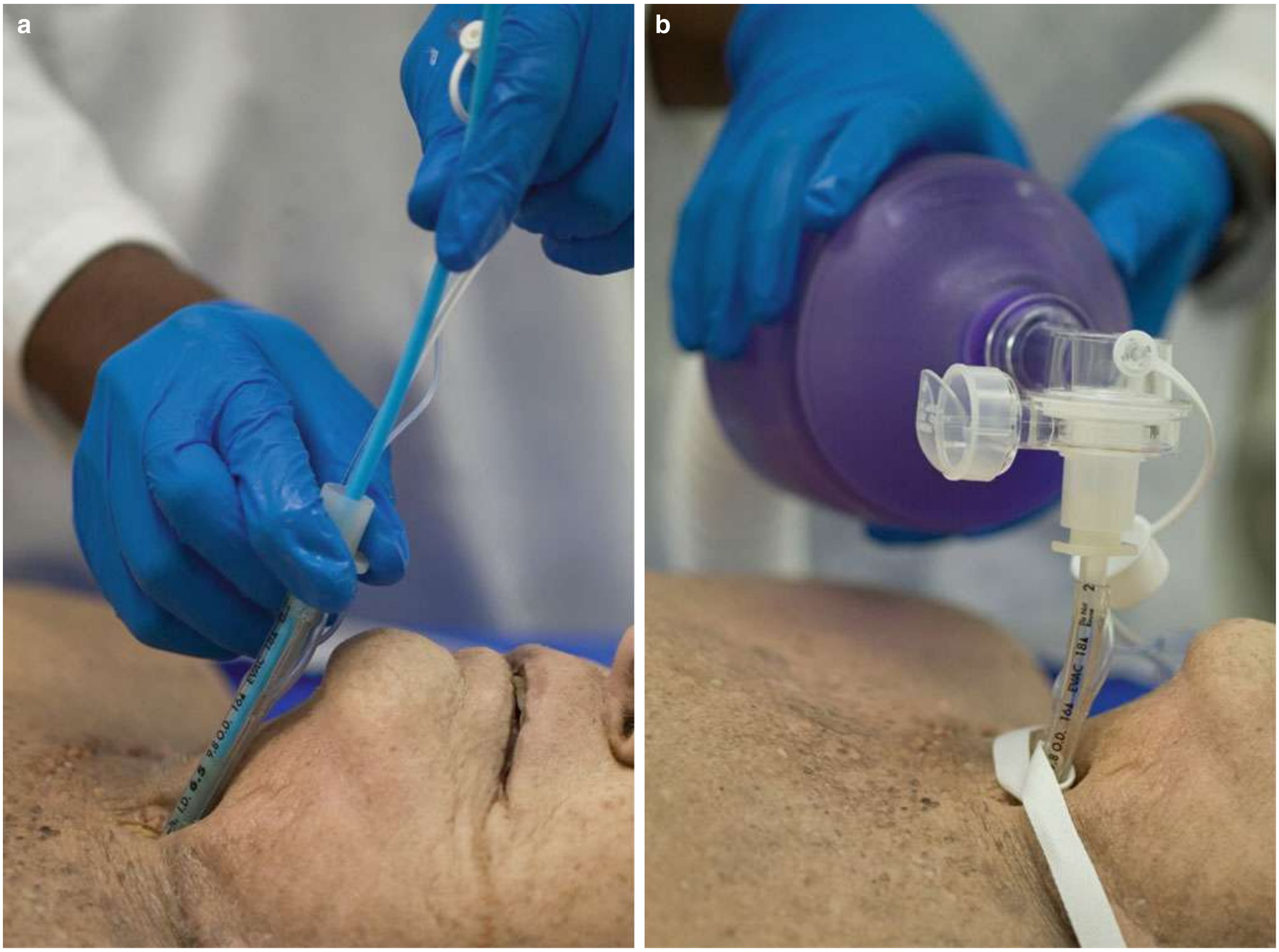

Fig. 19.4 (**a**) Slide the endotracheal tube (ETT) over the bougie into the trachea, (**b**) ventilate the patient

19.3.2 Scalpel-Trousseau

19.3.2.1 Materials and Medications (Fig. 19.5)

- Scalpel with #11 blade
- Tracheal hook
- Trousseau dilator
- Cuffed tracheostomy tube (TT) (6.5 or 7.0) or ETT (5.0, 5.5, or 6.0)
- Antiseptic preparation

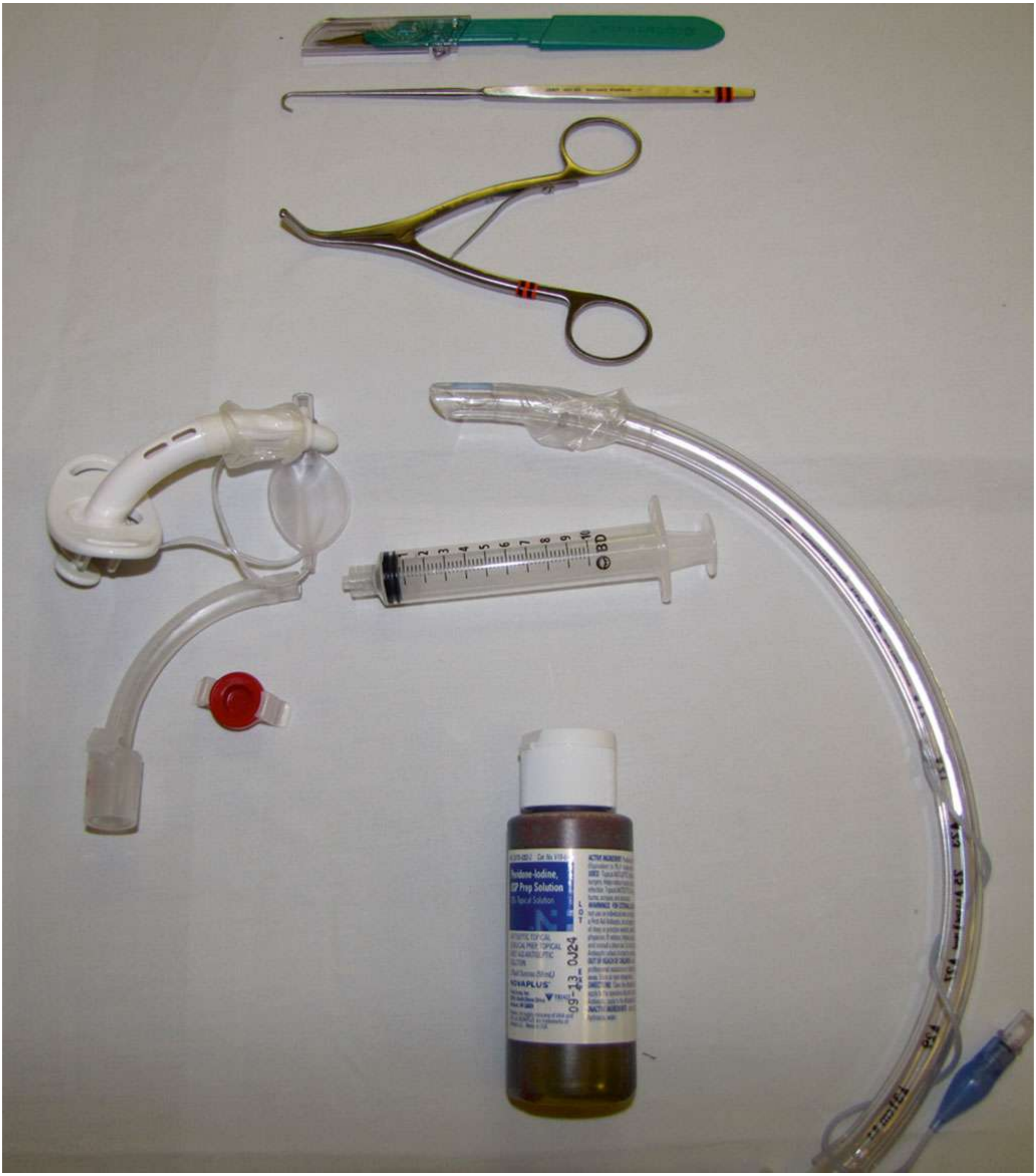
Povidone-Iodine, USP Prep Solution
BD
NOVAPLUS

Fig. 19.5 *Top to bottom*, *right to left*: Scalpel with #11 blade; tracheal hook; Trousseau dilator; cuffed tracheostomy tube (TT) (6.5 or 7.0) or ETT (5.0, 5.5, or 6.0); antiseptic preparation

19.3.2.2 Procedure

1. Apply topical antiseptic.
2. Palpate the thyroid notch, cricothyroid membrane, and hyoid bone for orientation.
3. Stabilize the thyroid between the thumb and the middle finger of the nondominant hand (Fig. 19.6a).
4. Make a vertical skin incision (2–3 cm) over the cricothyroid membrane (Fig. 19.6b).
5. Palpate with the index finger to verify the cricothyroid membrane location.
6. Use stabilization of the thyroid and palpation to maintain orientation of the anatomy.
7. Make a 1.5-cm horizontal incision through the lower half of the membrane (Fig. 19.7a).
8. Insert a tracheal hook into the incision, then rotate such that hook faces superiorly (Fig. 19.7b).
9. Withdraw at a 45° angle in a cephalad direction, applying gentle traction to the thyroid cartilage.
10. Place the Trousseau dilator into the incision transversely and open the membrane incision vertically (Fig. 19.8a).
11. Insert a cuffed ETT (5.0–6.0) or TT (6.5–7.0) into the incision between the prongs of the dilator in the horizontal access (Fig. 19.8b).
12. Rotate both the dilator and the ETT toward the head of the patient and then direct the tube downward into the trachea while removing the dilator.
13. Inflate the ETT cuff and ventilate the patient.
14. Verify the position of the ETT via auscultation, $EtCO_2$, and chest x-ray.
15. Once placement of the tube has been verified, the tracheal hook can be removed.
16. Secure the ETT.

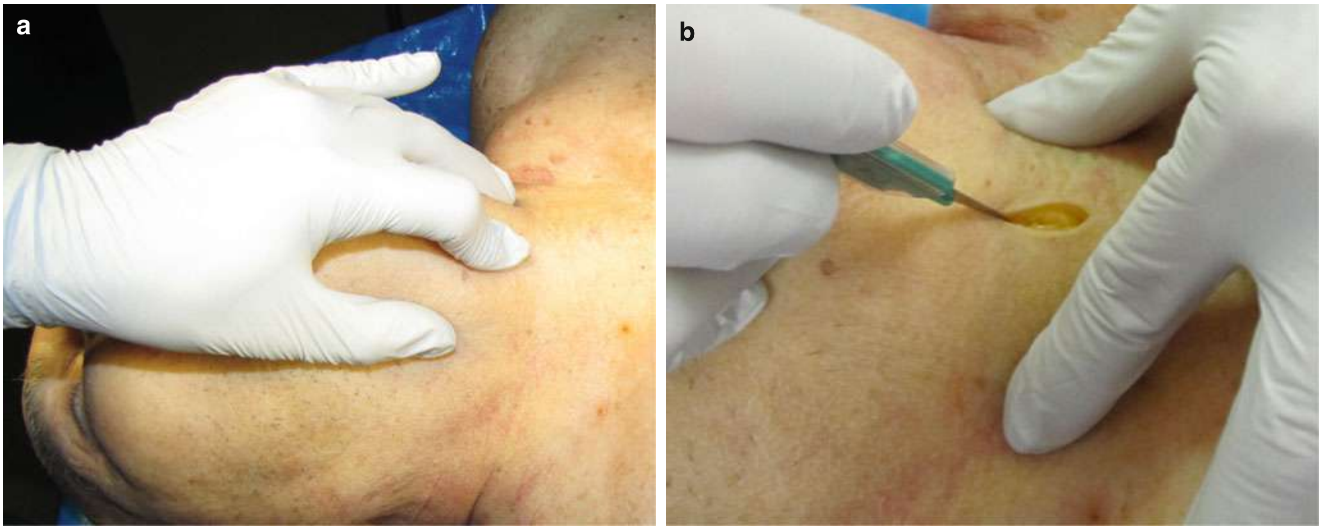

Fig. 19.6 (**a**) Stabilize the thyroid cartilage between the thumb and the middle finger of the nondominant hand; (**b**) make a vertical incision over the cricothyroid membrane

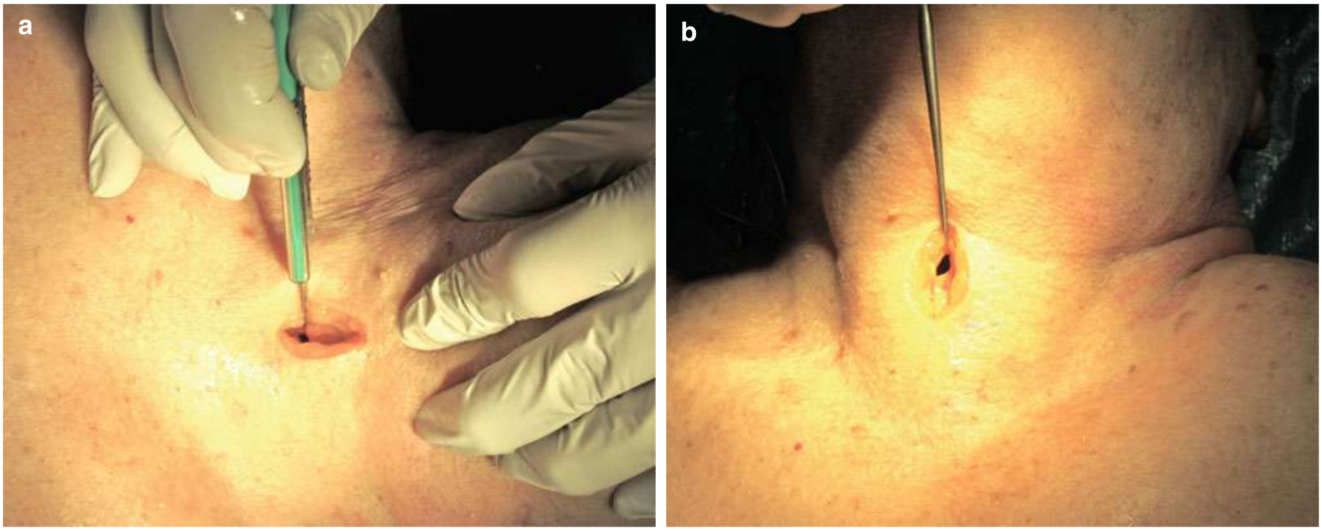

Fig. 19.7 (**a**) Make a horizontal incision in the lower half of the cricothyroid membrane. (**b**) Insert a tracheal hook and apply gentle traction superiorly at a 45° angle

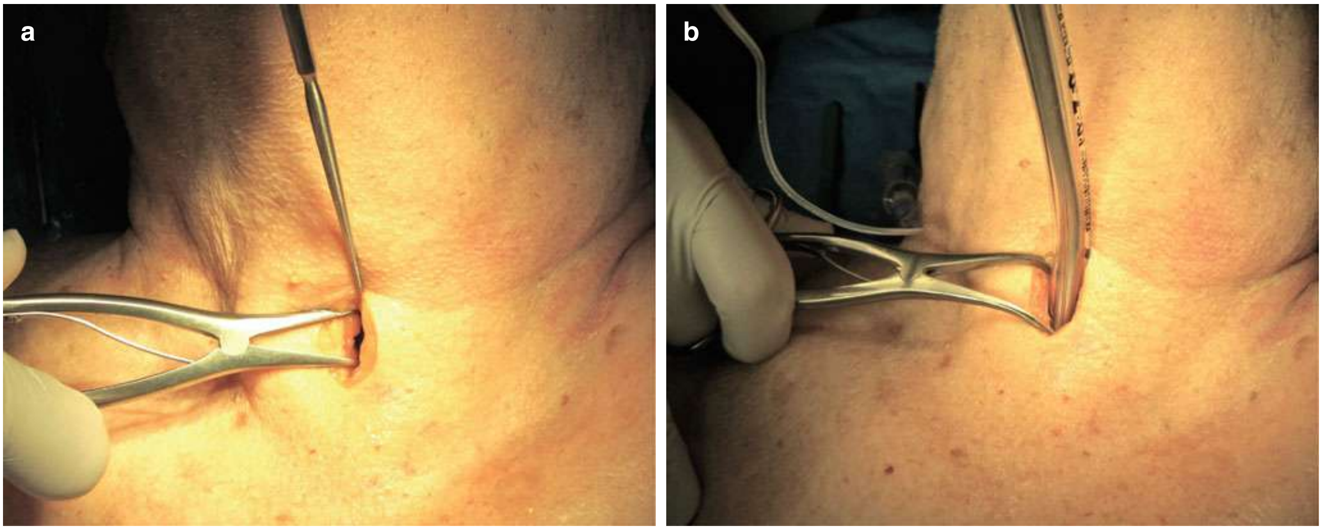

Fig. 19.8 (**a**) Insert the Trousseau dilator into the incision, and open the path for placement of the tracheostomy tube (TT) or the ETT. (**b**) Insert the TT or the ETT into the expanded incision between the Trousseau dilator prongs

19.4 Complications

- Bleeding
- ETT misplacement (false passage, through the thyrohyoid membrane, unintentional tracheostomy)
- Hoarseness, dysphonia, or vocal cord paralysis
- Subglottic or laryngeal stenosis
- Damage to thyroid cartilage, cricoid cartilage, or tracheal rings
- Perforated esophagus
- Infection
- Aspiration

19.5 Pearls

- Predictors of difficult cricothyrotomy: "SHORT" [1]
  - Prior *S*urgery or *S*car tissue
  - *H*ematoma
  - *O*bese
  - Prior *R*adiation
  - *T*umor/abscess
- The incision should cut through the skin and subcutaneous tissue down to the cricothyroid membrane and cartilages.
- Vertical incisions allow for extension in either direction if the cricothyroid membrane is above or below initial incision.
- Blood in the field may hinder visualization of the membrane, but the airway should be established before attempts to control any bleeding.
- Cricothyroid arteries are located cephalad to the cricothyroid membrane.
- Either a TT or an ETT can be used for the procedure.
  - ETTs are more ubiquitous.
  - TTs are easier to secure.
- If an ETT is used, a stylet can help direct placement.
- Cricothyroidotomy is preferred in the emergency setting over tracheostomy owing to the increased risks of bleeding, the mobility of the trachea, and the risk of lacerating the underlying thyroid gland [2].
- Ultrasound can be used to visualize landmarks in patients in whom landmarks are difficult to identify.

References

1. Walls RM, Murphy MF, editors. Manual of emergency airway management. 4th ed. Philadelphia: Wolters Kluwer; 2012.
2. Boon JM, Abrahams PH, Meiring JH, et al. Cricothyroidotomy: a clinical anatomy review. Clin Anat. 2004;17:478–86.

Selected Reading

DiGiacomo C, Neshat KK, Angus LD, et al. Emergency cricothyrotomy. Mil Med. 2003;168:541–4.

Hamilton PH, Kang JJ. Emergency airway management. Mt Sinai J Med. 1997;64:292–301.

Helm M, Gries A, Mutzbauer T. Surgical approach in difficult airway management. Best Pract Res Clin Anaesthesiol. 2005;19:623–40.

Sagarin MJ, Barton ED, Chng YM, Walls RM. Airway management by US and Canadian emergency medicine residents: a multicenter analysis of more than 6,000 endotracheal intubation attempts. Ann Emerg Med. 2005;46:328–36.

Walls RM. Cricothyroidotomy. Emerg Med Clin North Am. 1988;6:725–36.

Tracheostomy Tube Malfunction

20

Deena Bengiamin and Bharath Chakravarthy

Tracheostomy tube (TT) malfunction is the source of airway compromise in patients requiring these airway devices. TT malfunction may create an airway emergency, and the timely replacement of TTs is a challenging procedure in the most experienced hands.

20.1 Indications

- Cuff rupture
  - Can lead to dislodgment
- Dislodgment
  - Most common emergency department TT complication
  - Can lead to air passage obstruction
- Obstruction
  - Caused by blood or thick, dry secretions (formed in the absence of nasopharyngeal air humidification).
  - Dried secretions or blood can act as a one-way valve, allowing air in but restricting outward flow.

20.2 Materials and Medications

- Airway suction catheter
- Oxygen humidifier
- Bougie/nasogastric (NG) tube (12 French)
- Saline
- N-Acetyl-cysteine (NAC)
- Appropriately sized endotracheal tube (ETT) or TT

20.3 Procedure

1. Supply high-flow humidified oxygen through a bag valve mask (BVM), or a non-rebreather face mask.
2. Assess clinical indicators to determine TT problem.
   (a) Indicators of cuff rupture: Air leak with BVM and loose TT.
   (b) Indicators of dislodged TT (if TT is still in the stomal opening): Subcutaneous emphysema, crepitus, and diminished or absent breath sounds.
   (c) Indicators of an obstructed TT: ±stridor and diminished or absent breath sounds.
   (d) If time allows, chest radiograph, continuous capnography, and oxygen saturation can be helpful.
3. Obstruction
   (a) Remove the inner cannula and inspect for obstruction—clean if necessary.
   (b) If this fails, 5–10 mL of saline or NAC can be squirted directly down the TT to loosen secretions.
   (c) Suction thoroughly with a suction catheter (Fig. 20.1).
   (d) In refractory cases, the TT will need replacement.
4. Replacement
   (a) Ideally, the replacement tube should be of the same type and size as the original TT.
   (b) A smaller size TT or ETT can be helpful in settings of airway compromise.
   (c) A 6–7.5-cm ETT tube may be used if a TT is unavailable.
   (d) Remove the existing TT.
   (e) Hyperextend the patient's head and neck to maximize visualization of the stoma.

D. Bengiamin, MD
Department of Emergency Medicine, University of San Francisco Fresno, Fresno, CA, USA
e-mail: dbengiamin@fresno.ucsf.edu

B. Chakravarthy, MD, MPH (✉)
Department of Emergency Medicine, University of California Irvine, Orange, CA, USA
e-mail: bchakrav@uci.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_20

- (f) Note: Careful inspection of the area is paramount because the thyroid isthmus may obscure visualization of the tracheal stoma.
- (g) Techniques
  - (i) Direct insertion
    1. As soon as possible, insert the new TT or ETT into the stoma to prevent stomal narrowing (Fig. 20.2).
    2. Inflate the new TT/ETT cuff.
  - (ii) Bougie or NG tube
    1. Lubricate the TT or ETT tube.
    2. Lubricate a bougie or 12-French NG tube.
    3. Insert the lubricated bougie or NG tube into the TT or ETT.
    4. Insert the bougie or NG tube into stoma and advance into the trachea (Fig. 20.3a).
    5. Direct the NG tube or bougie caudad toward the lower tracheobronchial tree.
    6. Do not advance more than 7 cm.
    7. If resistance is noted:
       - (a) Either the operator has reached a terminal bronchiole or is in a false passage.
       - (b) Do not force bougie/NG tube farther at this point.
       - (c) Use clinical judgment (bougie/ETT depth, palpation) to determine likely placement.
    8. Advance TT or NG tube over bougie or NG tube is in trachea (Fig. 20.3b).
    9. After the TT is in place, remove the bougie/NG tube.
  - (iii) The fingertip technique
    1. Insert a gloved forefinger into tracheal stoma (Fig. 20.4a).
    2. Formulate a mental plan as to the direction and path of the stoma.
    3. Then place a TT or an ETT into the stoma as the finger is withdrawn (Fig. 20.4b).

5. If placement of a TT or ETT is not possible through the stoma, consider endotracheal intubation.

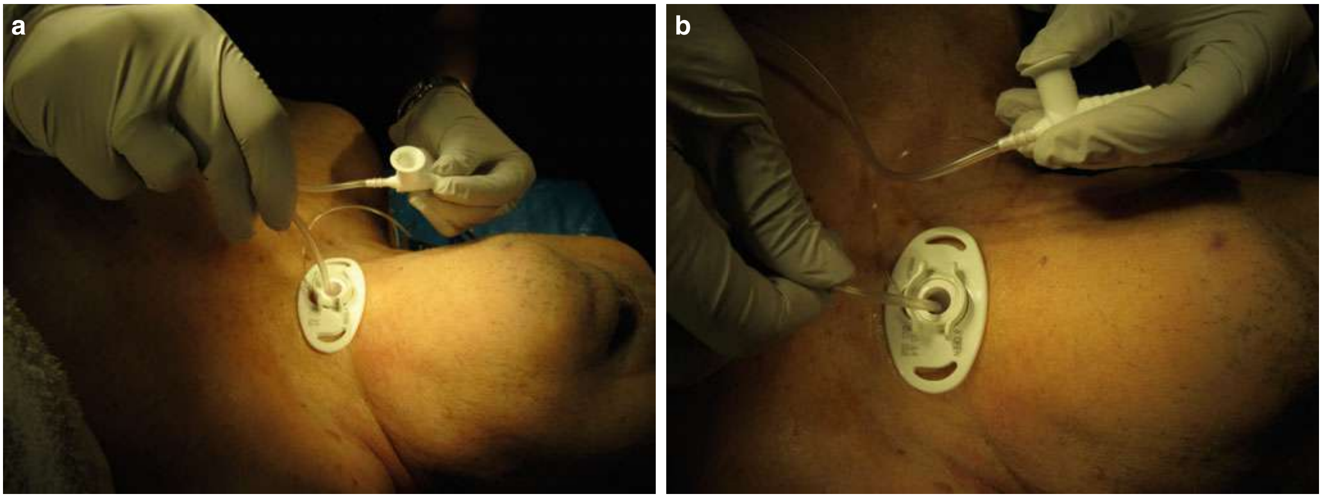

Fig. 20.1 (**a**) Insert the suction catheter to the appropriate depth, keeping the suction port open. (**b**) Slowly withdraw the catheter in a circular motion, keeping the suction port closed

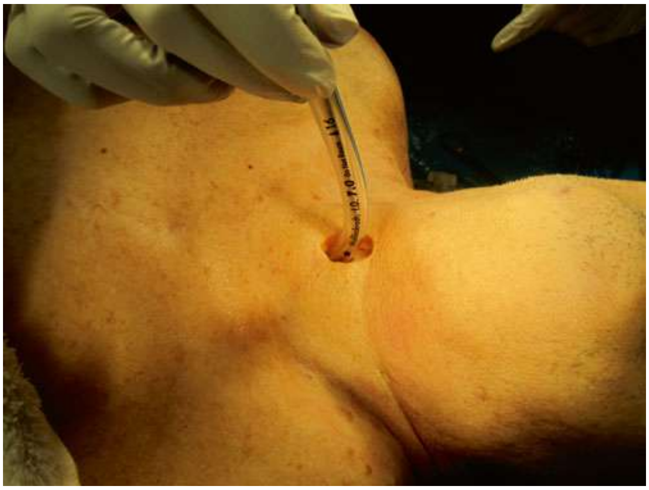

Fig. 20.2 Insertion of the endotracheal tube (ETT) directly into the stoma

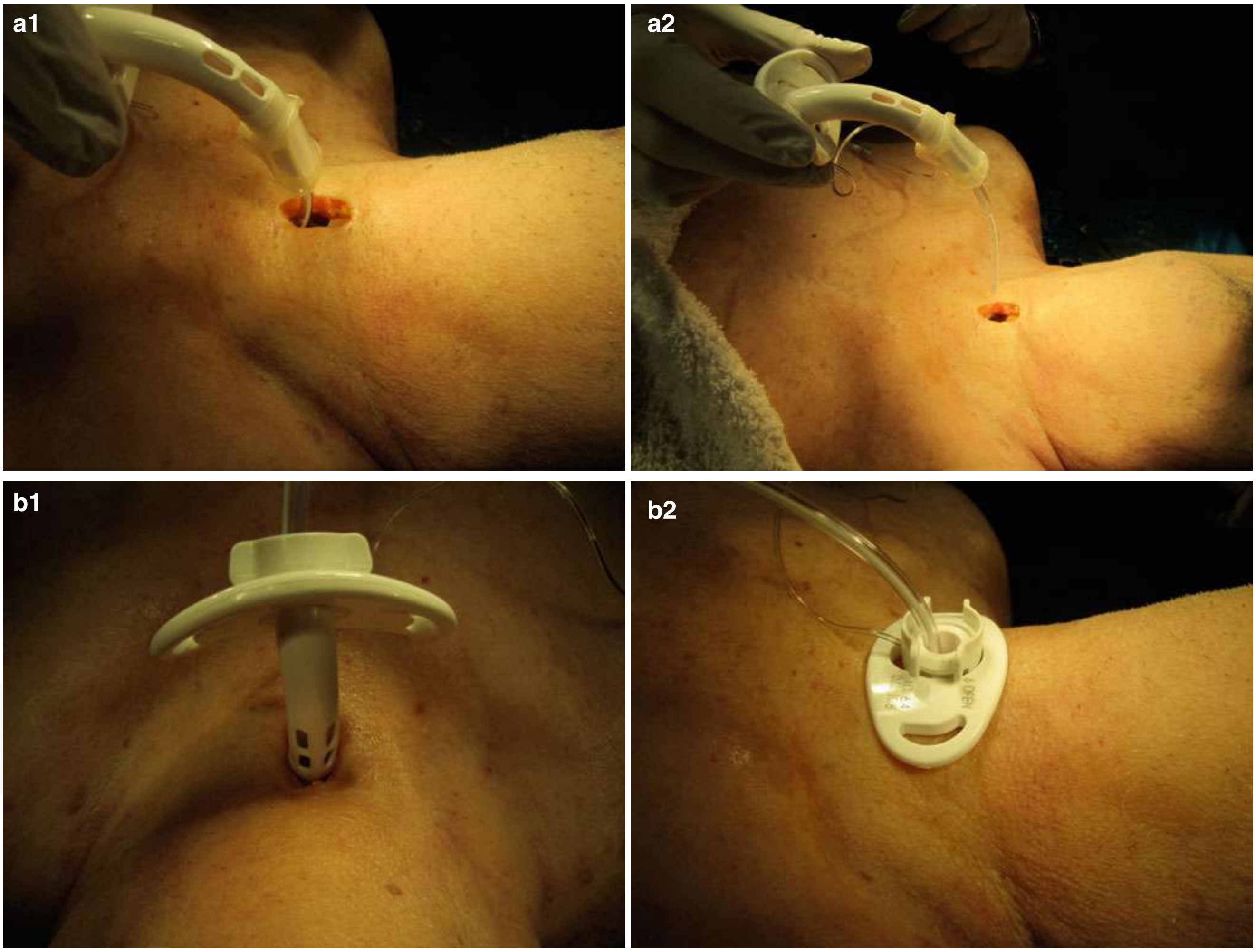

Fig. 20.3 (**a1, a2**) Insert the tracheostomy tube (TT) over the nasogastric (NG) tube or bougie, which is first inserted into the trachea. (**b1, b2**) Advance the TT over the NG tube or bougie, which serves as a guidewire

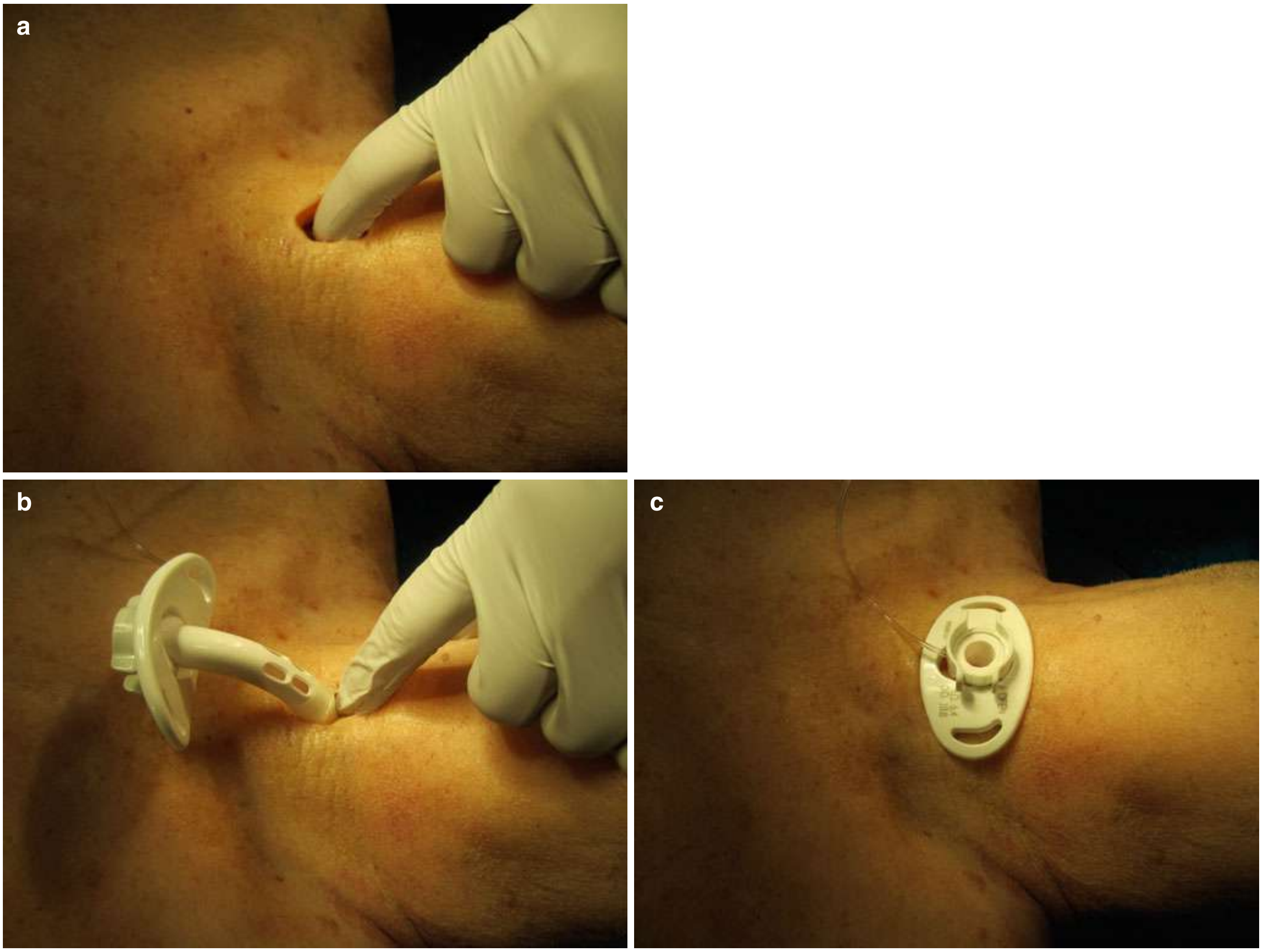

Fig. 20.4 (**a**) Insert a gloved finger into the stoma. (**b**) Insert the TT along the path while withdrawing the gloved finger. (**c**) TT in place

20.4 Pearls and Pitfalls

- Stomal closing
  - – Stomal constriction begins as soon as the TT is removed or displaced.
  - – Forceful attempts to replace a large TT may result in false passages and trauma.
- Tracheal stenosis
  - – Constant TT cuff pressure may cause necrosis, ulceration, and granulation tissue formation, leading to tracheal narrowing.
  - – Complicates TT replacement in the setting of dislodgment or obstruction.
  - – Bougie use can be helpful to place a small(er) ETT until surgical dilation and/or resection can be achieved.
- Creation of a false lumen during recannulation
  - – Subcutaneous emphysema will be an early indicator of this.
  - – Early confirmation of correct placement of TT by confirming:
    - Continuous end-tidal CO_2 monitoring
    - Equal chest rise
    - Bilateral breath sounds
- Unrecognized trachea–innominate artery fistula (Fig. 20.5)
  - – Usually occurs within 3–4 weeks of placement
  - – Presentation: Bleeding around the tracheostomy tube (>10 mL) or massive hemoptysis
  - – Requires
    - ETT cuff overinflation to compress the fistula.
    - Digital pressure on stoma may be helpful to tamponade the bleeding.
    - Place stomal ETT deep to bleeding fistula to protect airway.
    - Definitive surgical intervention in operating room.
  - – Associated with high mortality
- Unrecognized tracheoesophageal fistula
  - – Usually iatrogenic injury from TT placement or NG tube erosion
  - – Presentation: Dyspnea, copious TT secretions, recurrent food aspiration, and gastric distention
  - – Requires
    - Bronchoscopy or swallowing studies to confirm diagnosis
    - Surgical repair or stenting

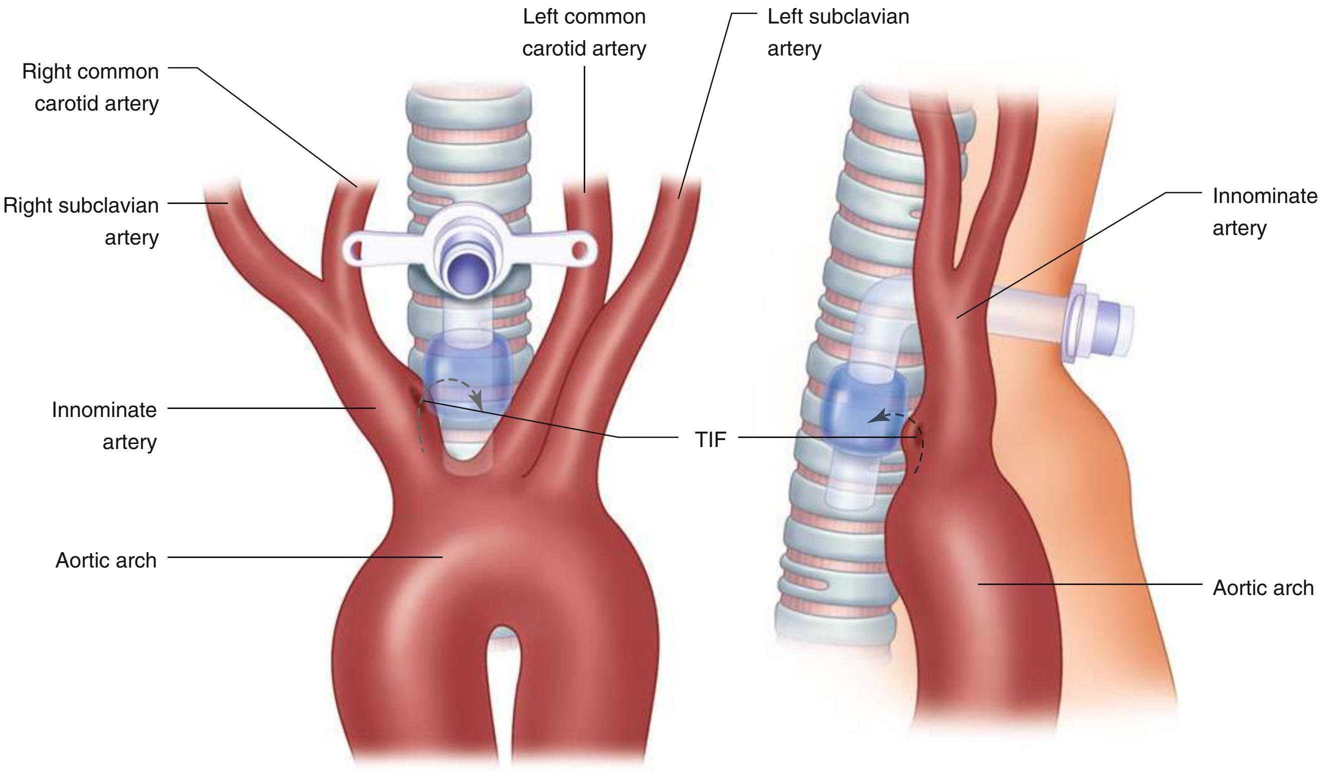

Fig. 20.5 Anatomical relationship between the trachea and the innominate artery. *TIF* trachea–innominate artery fistula

Selected Reading

De Leyn P, Bedert L, Delcroix M, et al. Tracheotomy: clinical review and guidelines. Eur J Cardiothorac Surg. 2007;32:412–21.

Dobiesx VA, Miller SA, Pitzele MJ. Complications of tracheostomies. In: Wolfson AB, Hendey GW, Ling LJ, Rosen CL, Scheider JJ, Sharieff GQ, editors. Harwood-Nuss' clinical practice of emergency medicine. 5th ed. Philadelphia: Lippincott Williams & Wilkins; 2009.

Epstein SK. Late complications of tracheostomy. Respir Care. 2005;40:542–9.

Friedman M, Ibrahim H. The dislodged tracheostomy tube: "fingertip" technique. Oper Technol Otolaryngol. 2002;13:217–8.

Young JS, Brady WJ, Kesser B, Mullins D. A novel method for replacement of the dislodged tracheostomy tube: the nasogastric tube "guidewire" technique. J Emerg Med. 1996;14:205–8.

Percutaneous Transtracheal Jet Ventilation

21

Clint Masterson

21.1 Indications

- Failure to control the airway by other means
- As a temporary measure while preparing for definitive airway control
- Securing the airway in crash airways in infants and small children

21.2 Contraindications

- Absolute
  - Transection of the trachea below the cricothyroid membrane
- Relative
  - Inability to identify the cricothyroid landmarks
  - Anatomical distortion to the cricothyroid membrane
  - Supraglottic obstruction (preventing gas exhalation)

21.3 Materials and Medications (Fig. 21.1)

- Betadine, chlorhexidine, or similar skin sterilization solution
- 12- to 16-gauge angiocatheter or transtracheal jet ventilation (TTJV) purpose-specific catheter
- 10-mL syringe filled with 4 mL of normal saline, 2 % lidocaine, or viscous lidocaine
- Hand-operated regulator valve
- Attach oxygen supply.
  - Connect kit tubing to wall oxygen OR.
  - Connect 7–0 endotracheal connector to bag valve mask (BVM) attached to oxygen.

C. Masterson, MD
Department of Emergency Medicine, Mayo Clinic Health System
in Fairmont, Fairmont, MN, USA
e-mail: clint@clintm.us

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_21

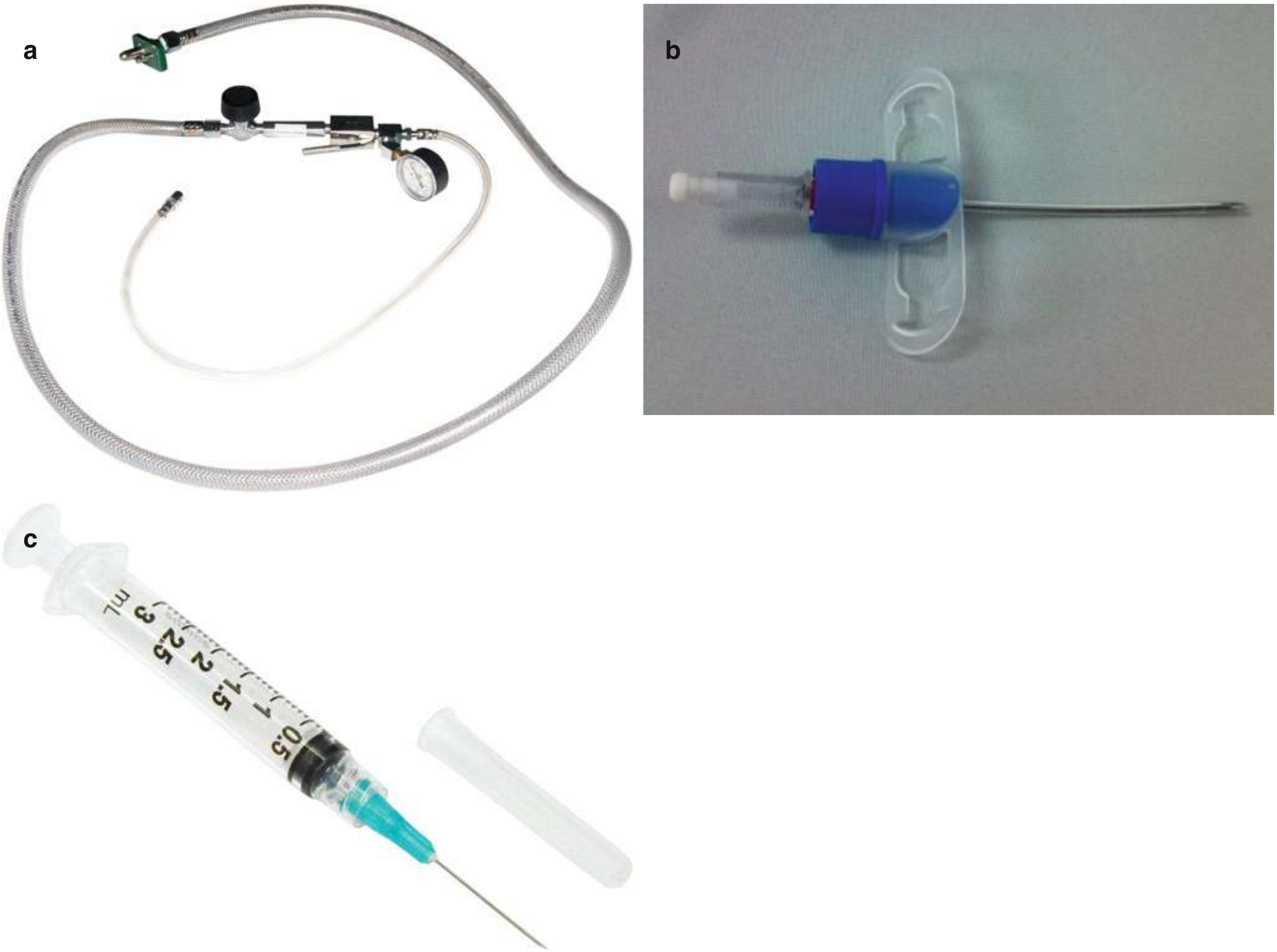
a
b
c
mL
3
2.5
2
1.5
1
0.5

Fig. 21.1 (**a**) Tubing and regulator valve, (**b**) commercially available catheter, (**c**) 3-mL or 10-mL syringe

21.4 Procedure

1. Attach the tubing and the hand-operated regulator valve to wall oxygen (Fig. 21.2a), and place the distal end of the tubing near the patient in preparation for ventilation.
2. Adjust regulator to maximum pressure, 50 psi if possible (Fig. 21.2b).
3. Palpate the cricothyroid membrane just distal to the thyroid prominence (Fig. 21.3).
    (a) Sterilize the area with a suitable cleansing agent.
    (b) Use the thumb and index finger of the nondominant hand to stabilize the trachea for the procedure.
4. Attach the TTJV catheter (or angiocatheter) to the syringe (Fig. 21.4).
5. Advance the catheter through the cricothyroid membrane at a 30–45° caudal direction while aspirating with the syringe (Fig. 21.5).
6. Return of air confirms entry into the trachea.
7. If lidocaine is utilized, it can then be injected to prevent spasm during the procedure.
8. Fully advance the angiocatheter and secure it while the needle and syringe are withdrawn.
9. Remove the needle, secure it to the skin, and connect it to the regulator hose.
10. Secure the distal end of the oxygen tubing (distal to the hand-operated valve) to the catheter (Fig. 21.6).
11. If a BVM is used as the oxygen source:
    (a) Attach a 3-mL syringe to the angiocatheter.
    (b) Attach the BVM with the 7–0 endotracheal tube (ETT) connector to the end of the plungerless 3-mL syringe (Fig. 21.7).
12. Operate the valve 12–20 times a minute with long periods to allow gas exhalation and exchange (Fig. 21.8).
13. Preparations should be made for a definitive airway as soon as possible—preferably within 15 min.

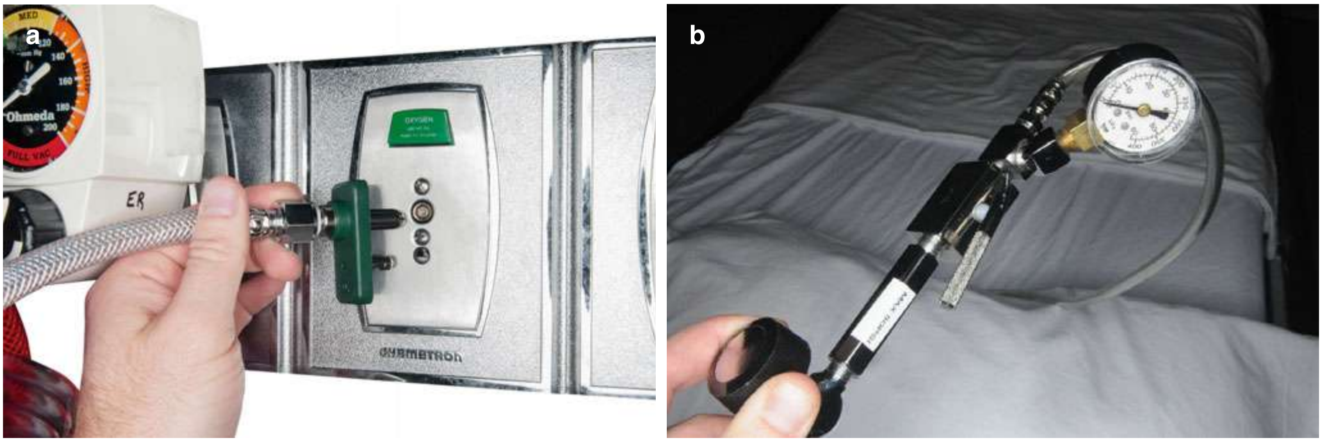

Fig. 21.2 (**a**) Attach tubing and the hand-operated regulator valve to wall oxygen, and (**b**) adjust regulator to maximum pressure (50 psi if possible)

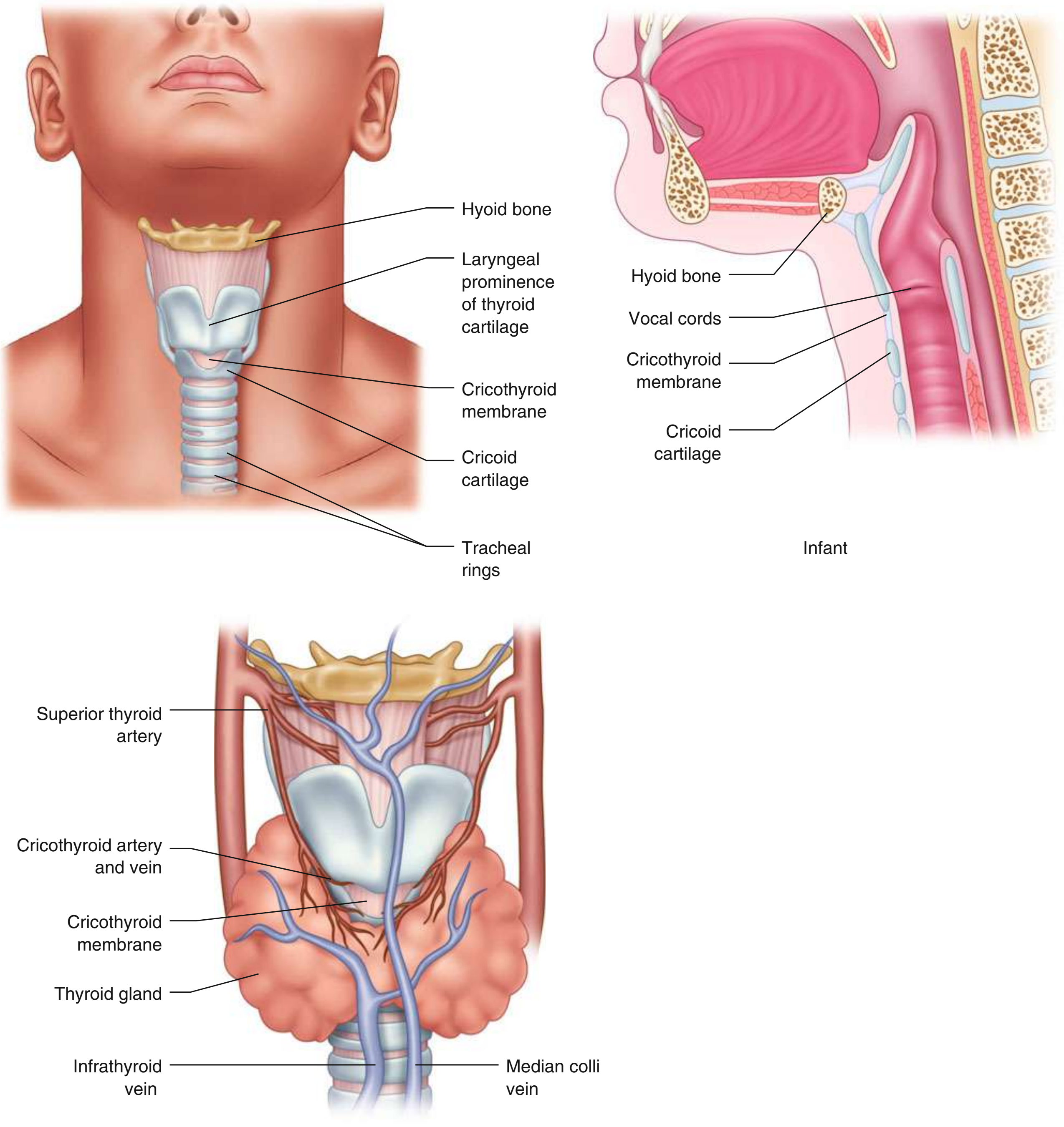
Hyoid bone
Laryngeal prominence of thyroid cartilage
Cricothyroid membrane
Cricoid cartilage
Tracheal rings
Hyoid bone
Vocal cords
Cricothyroid membrane
Cricoid cartilage
Infant
Superior thyroid artery
Cricothyroid artery and vein
Cricothyroid membrane
Thyroid gland
Infrathyroid vein
Median colli vein

Fig. 21.3 Airway anatomy

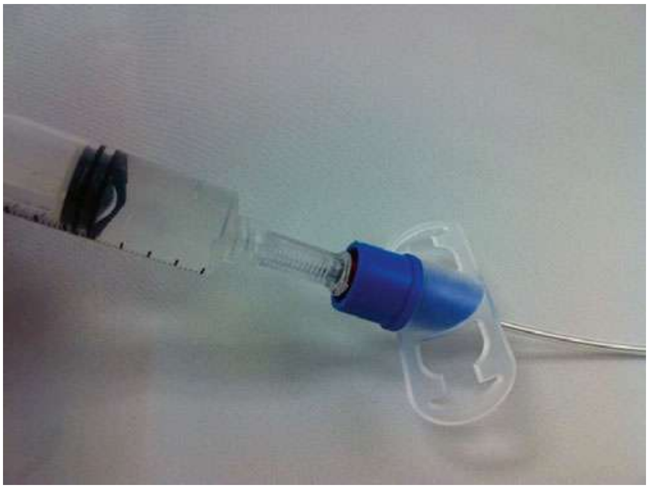

Fig. 21.4 Attach the TTJV catheter to the syringe

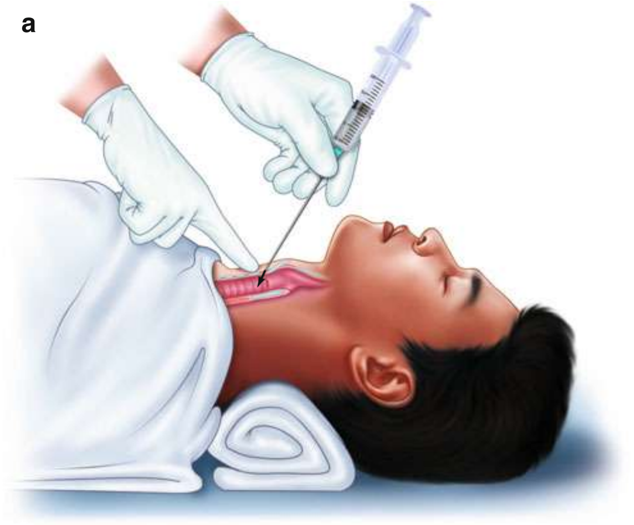

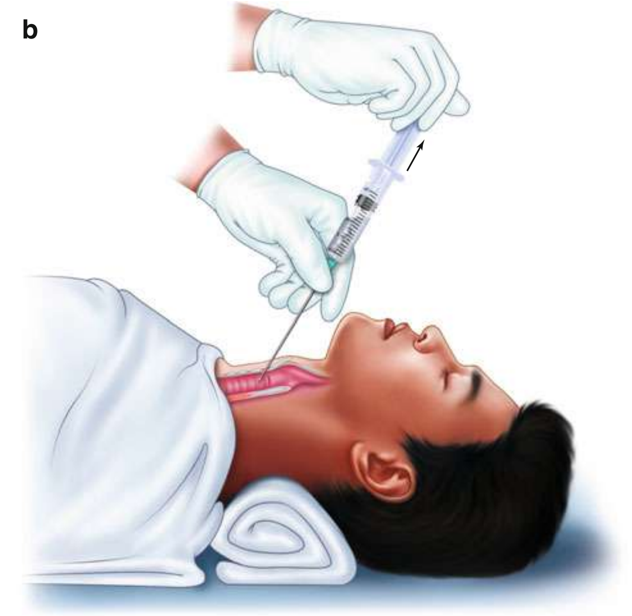

Fig. 21.5 (**a**) Advance the catheter through the cricothyroid membrane at a 30–45° caudal direction (**b**) while aspirating with the syringe

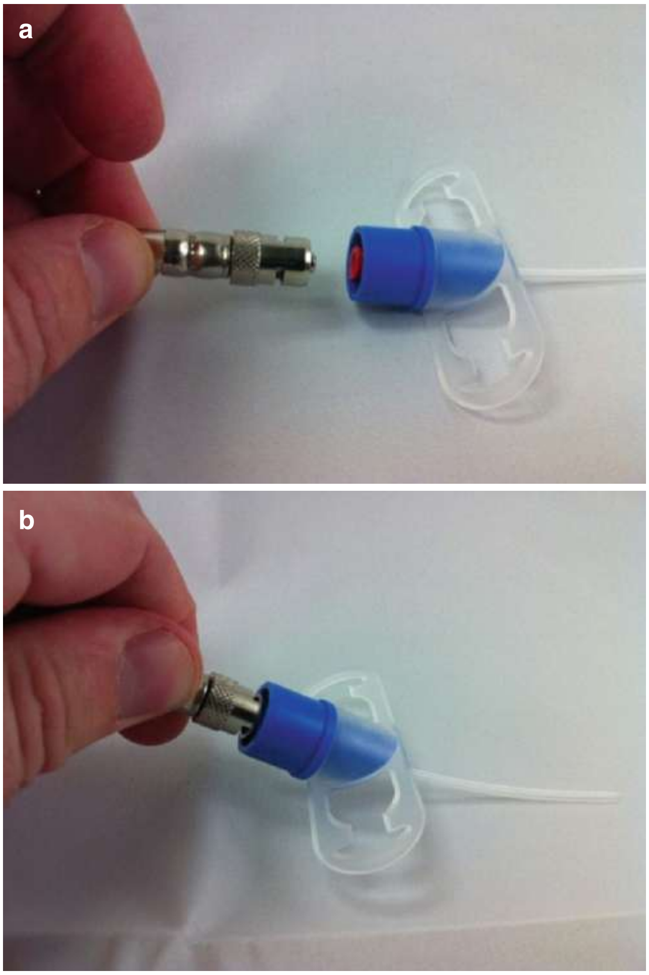

Fig. 21.6 Secure the distal end of the oxygen tubing to the catheter

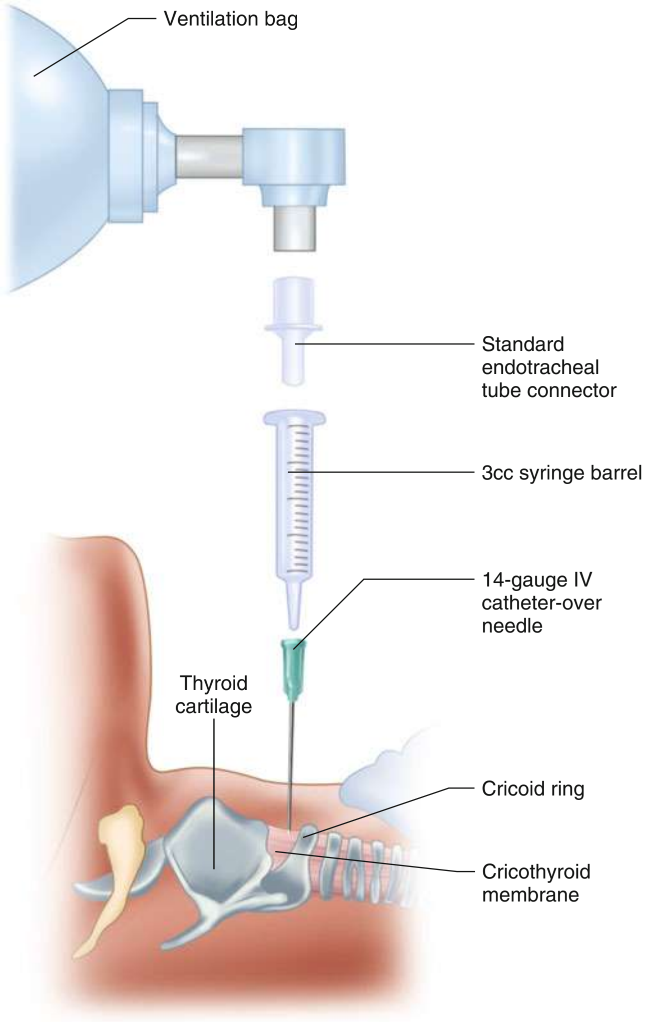
Ventilation bag
Standard endotracheal tube connector
3cc syringe barrel
14-gauge IV catheter-over needle
Thyroid cartilage
Cricoid ring
Cricothyroid membrane

Fig. 21.7 Attach the BVM with the 7–0 endotracheal tube (ETT) connector to the end of the plungerless 3-mL syringe

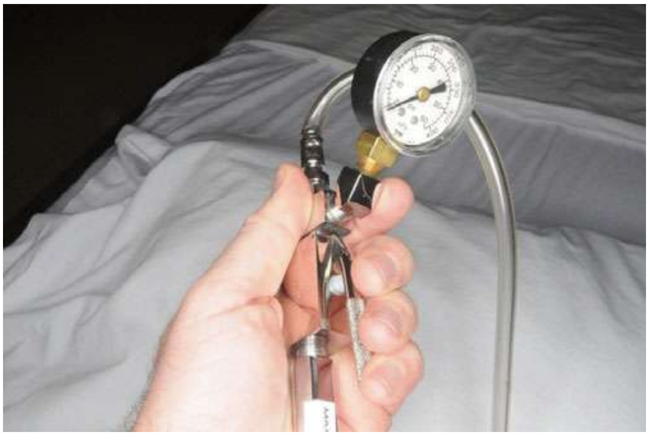

Fig. 21.8 Operate the valve 12–20 times a minute with long periods to allow gas exhalation and exchange

21.5 Complications

- Pneumothorax
- Pneumomediastinum
- Subcutaneous emphysema
- Catheter kink or misplacement
- Hypercarbia and respiratory acidosis
  - Use of TTJV for prolonged periods of time without adequate ventilation will elevate CO_2.
- Barotrauma
- Coughing in conscious patients
- Aspiration
- Persistent stoma

21.6 Pearls and Pitfalls

- If the wall connector does not have a pressure regulator, it can still be used although the risk of barotrauma is greater. Use the endpoint of chest rise to determine the end of each ventilation burst in this case.
- Higher pressures and lack of supraglottic air exchange are risk factors for pneumothorax. If the supraglottic area is obstructed, a Y catheter can be attached to allow gas to escape before the next insufflation.
- TTJV may or may not allow sufficient gas exchange to prevent hypercarbia. Preparations should be made to obtain a definitive airway as soon as possible.
- Endotracheal intubation may be facilitated by the high pressures insufflated in the trachea, and a repeat attempt may be performed after the transtracheal ventilation is achieved.

Selected Reading

Patel R. Percutaneous transtracheal jet ventilation. A safe, quick and temporary way to provide oxygenation and ventilation when conventional methods are unsuccessful. Chest. 1999;116:1689–94.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. Philadelphia: Saunders Elsevier; 2010.

Tintinalli J. Tintinalli's emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw Hill; 2010.

Walls R. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Part III

Thoracic Procedures

Needle Thoracostomy

22

Lucas McArthur and Christian Fromm

22.1 Indications

Needle decompression thoracostomy is a procedure used in the emergent treatment of a tension pneumothorax. Tension pneumothorax is a clinical diagnosis. Decompression treatment should not be delayed in order to obtain radiographic confirmation. The following scenarios illustrate some of the clinical signs that *may* be present in such patients:

- *Awake patient* with suspected or confirmed tension pneumothorax
  - Chest pain
  - Respiratory distress
  - Decreased breath sounds with hyperresonance and/or subcutaneous emphysema
  - Trachea deviated away from the side of the pneumothorax
  - Tachycardia
  - Falling pulse oximetry (SpO_2)
  - Shock
- *Ventilated patient* with suspected or confirmed pneumothorax (often insidious)
  - Increased resistance to ventilation
  - Hypotension
  - Elevated central venous pressure
  - Tachycardia
  - Decreased breath sounds with hyperresonance and/or subcutaneous emphysema
  - Trachea deviated away from the side of the pneumothorax
  - Falling SpO_2
  - Shock
- *Injured patient* (especially with penetrating chest trauma) with suspected or confirmed tension pneumothorax
  - In arrest
  - Unexplained hypotension
  - Apnea
  - Decreased breath sounds with hyperresonance and/or subcutaneous emphysema

22.1.1 Absolute Indications

- Patient in acute respiratory distress with rapid decompensation secondary to suspected or confirmed tension pneumothorax
- Injured patient in extremis with apnea, unexplained hypotension, or arrest

22.2 Contraindications

- No absolute contraindications.

22.3 Materials

- Large-bore needle/angiocatheter (minimum of 16 gauge)
- 10-mL syringe (optional)
- One-way valve (optional)
- Betadine (povidone-iodine) swab/chlorhexidine scrub
- Tape

L. McArthur, MD (✉)
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: LMcarthur@jfkhealth.org

C. Fromm, MD
Department of Emergency Medicine, Maimonides Medical Center,
SUNY Downstate College of Medicine, New York, NY, USA
e-mail: cfromm@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_22

22.4 Procedure

1. Expose the anterior chest at the level of the second intercostal space on the affected side (Fig. 22.1). Alternatively, expose the chest wall at the level of the anterior axillary line in the fourth or fifth intercostal space on the affected side.
2. Cleanse the area with a Betadine swab or chlorhexidine scrub (Fig. 22.2).
3. Using a gloved hand, locate the second intercostal space at the midclavicular line.
    (a) The first rib is normally not felt.
    (b) The second rib is felt just below the clavicle.
    (c) The second intercostal space is the area between the second and the third ribs.

    Note: Alternatively, this procedure may also be performed on the midaxillary line in the fourth intercostal space of the affected side. The same general steps listed later are employed in this approach and care is taken to avoid the neurovascular bundles inferior to the fourth rib.
4. Insert the needle/angiocatheter perpendicular to the chest wall into the second intercostal space just above the superior edge of the third rib to avoid the intercostal neurovascular bundle (Fig. 22.3).
    (a) This step may be done with or without a syringe attached.
    (b) Local anesthesia is usually unnecessary but may be used if the patient is not in extremis.
5. Carefully walk the needle over the third rib and advance until the pleural space is entered.
    (a) Entry into the pleural space is accompanied by a "popping" sound or a sensation of "giving way."
6. If you are able to withdraw air with the syringe or hear a "hiss" of air escaping through the angiocatheter during expiration and inspiration, then placement is considered successful.
7. After removing the needle, secure the angiocatheter in place with tape (Fig. 22.4).

    Caution: Do not reinsert needle into the angiocatheter owing to the danger of sheering the angiocatheter.
8. Assess the patient and evaluate the effectiveness of the procedure.
    (a) The patient should exhibit immediate and obvious improvement in respiratory status including improved lung sounds and vital signs.
    (b) The procedure may be repeated if the patient is not improving.
    (c) Excess pleural air may be aspirated through the angiocatheter with a syringe.
9. Obtain a chest radiograph to confirm success.
    (a) Repeat in 6 h.
10. Because needle decompression is only a temporizing measure, tube thoracostomy (see Chap. 23) must be performed for definitive management of the pneumothorax.

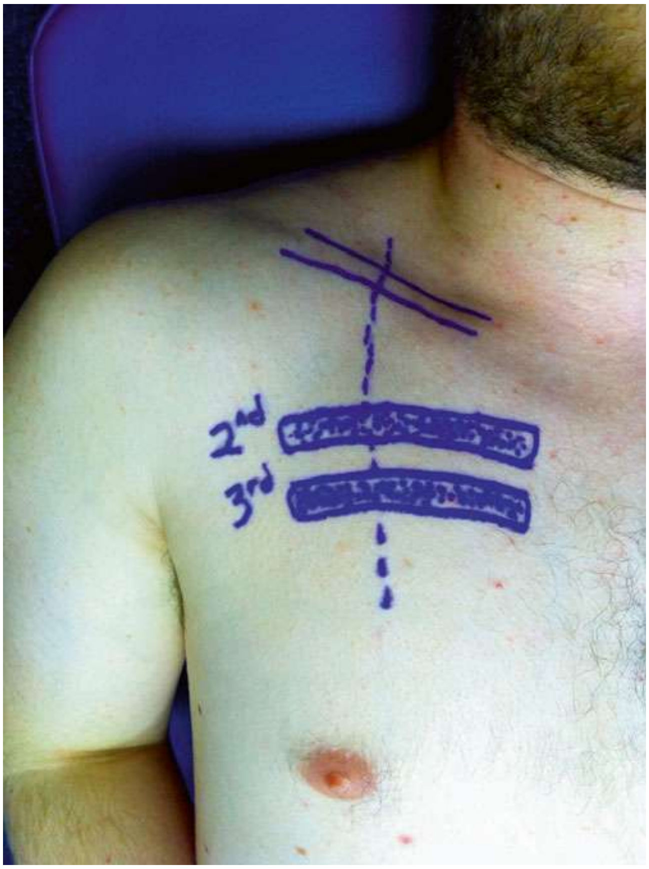

Fig. 22.1 The preferred site for needle thoracostomy is the second intercostal space in the midclavicular line

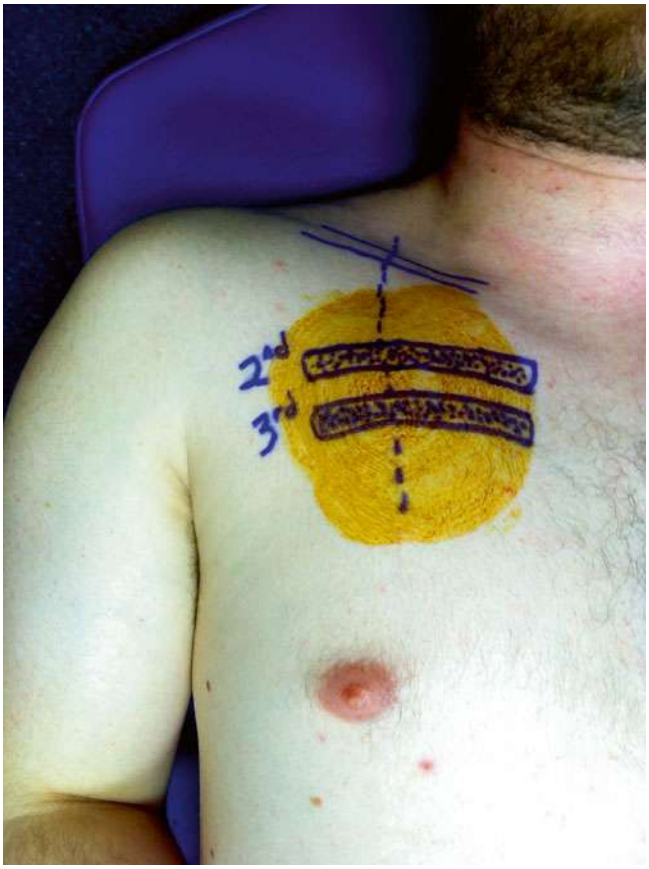

Fig. 22.2 Prepare the skin with povidone-iodine or chlorhexidine

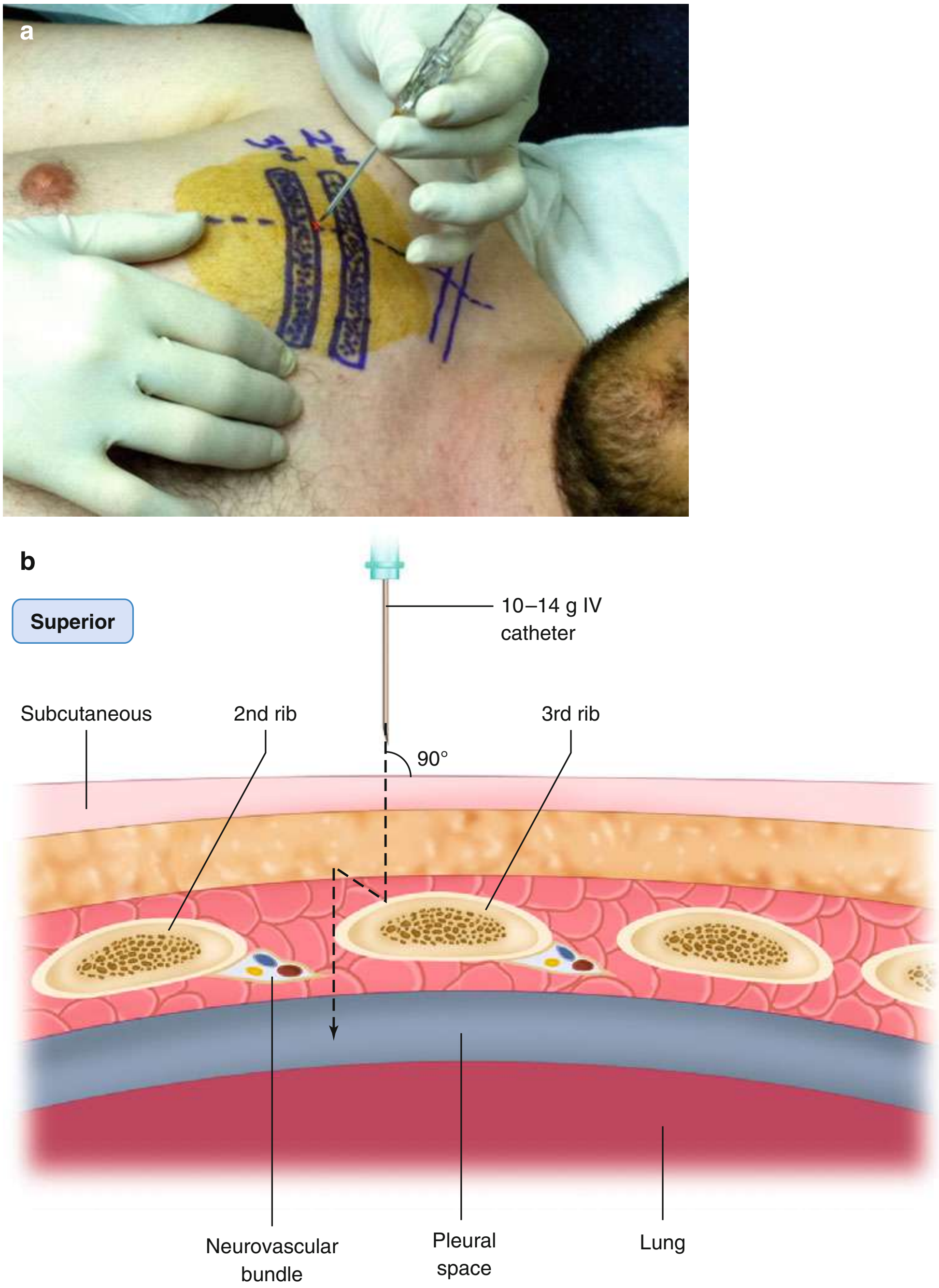
a
3rd
2nd
b
Superior
10–14 g IV catheter
Subcutaneous
2nd rib
3rd rib
90°
Neurovascular bundle
Pleural space
Lung

Fig. 22.3 (**a**) Insert the needle into the second intercostal space just above the superior edge of the third rib and (**b**) avoid the neurovascular bundle by approaching the skin with the needle perpendicular to the chest wall just above the superior edge of the third rib

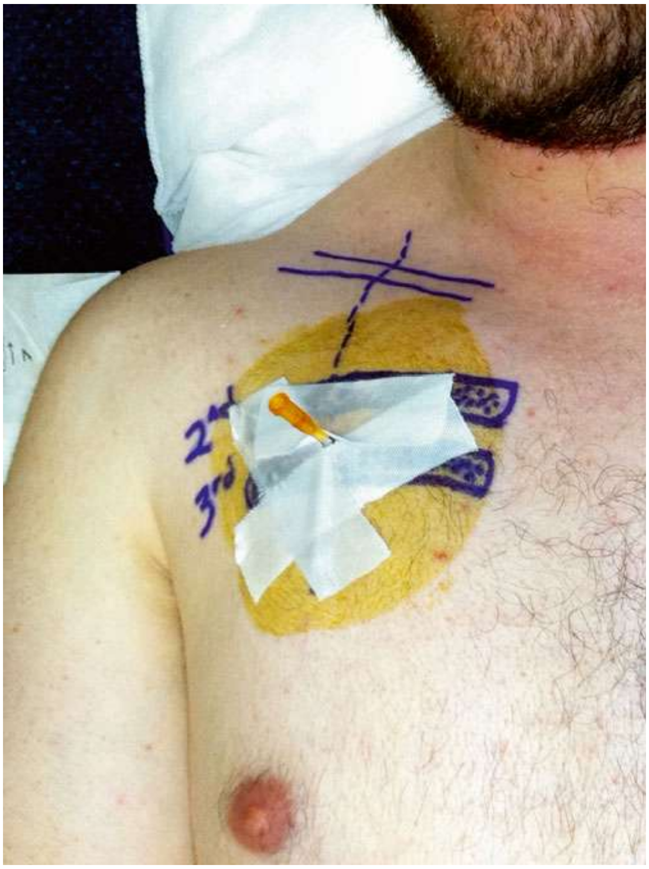

Fig. 22.4 After removing the needle, secure the angiocatheter in place with tape

22.5 Complications

- Failure to resolve the tension pneumothorax.
  - Obese or muscular patients may require a longer needle and catheter to reach the pleural space or, alternatively, may require proceeding immediately to tube thoracostomy.
- Iatrogenic pneumothorax.
- Laceration of intercostal artery or nerve.
- Rapid re-expansion may result in the development of pulmonary edema.
- Infection.

22.6 Pearls and Pitfalls

- Use the sternum as a landmark to more easily locate the second and third ribs.
- The same procedure may also be done on the midaxillary line in the fourth intercostal space, which is usually landmarked with the nipple.
- Primary pneumothorax is unusual in those older than 40 years. Consider the presence of underlying disease in this population.

Acknowledgments We would like to thank Antonios Likourezos, MA, MPH, and Abraham Lederman for assisting with the photographs.

Selected Reading

Britten S, Palmer SH. Chest wall thickness may limit adequate drainage of tension pneumothorax by needle thoracocentesis. Emerg Med J. 1996;13:426–7.

Custalow CB. Color atlas of emergency department procedures. Philadelphia: Saunders; 2005.

Leigh-Smith S, Harris T. Tension pneumothorax—time for a re-think? Emerg Med J. 2005;22:8–16.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. 3rd ed. Philadelphia: Saunders; 1998.

Chest Tube Thoracostomy

23

Brandon R. Allen and Latha Ganti

23.1 Indications

- Spontaneous pneumothorax (large and/or symptomatic)
- Tension pneumothorax (or suspected)
- Iatrogenic pneumothorax
- Penetrating chest injuries
- Hemopneumothorax in acute trauma
- Patient in extremis with evidence of thoracic trauma
- Complicated parapneumonic effusions (empyema)
- Chylothorax/hemothorax
- Post-thoracic surgery
- Bronchopleural fistula

23.2 Contraindications

- Absolute
  - Emergent thoracotomy
- Relative
  - Coagulopathy
  - Pulmonary bullae
  - Pulmonary, pleural, or thoracic adhesions
  - Loculated pleural effusion or empyema
  - Skin infection over the chest tube insertion site

23.3 Materials and Medications

- Tube thoracostomy tray
  - #10 scalpel; 18-, 22-, and 25-gauge needles; 10-mL syringes; forceps; clamps; scissors; drape; abdominal pads; 0 or 1–0 silk suture; needle driver; curved clamp (Fig. 23.1a)
- Betadine (povidone-iodine) or other skin antiseptic preparation solution
- Lidocaine (1 % or 2 % with epinephrine)
- Appropriate chest tube size (approximate)
  - Adult male: 28–36 French
  - Adult female: 28 French
  - Child: 12–24 French
  - Infant: 12–16 French
  - Neonate: 10–12 French

Vaseline gauze
Chest drainage system (Fig. 23.1b)

B.R. Allen, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: brandonrallen@ufl.edu

L. Ganti, MD, MS, MBA (✉)
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_23

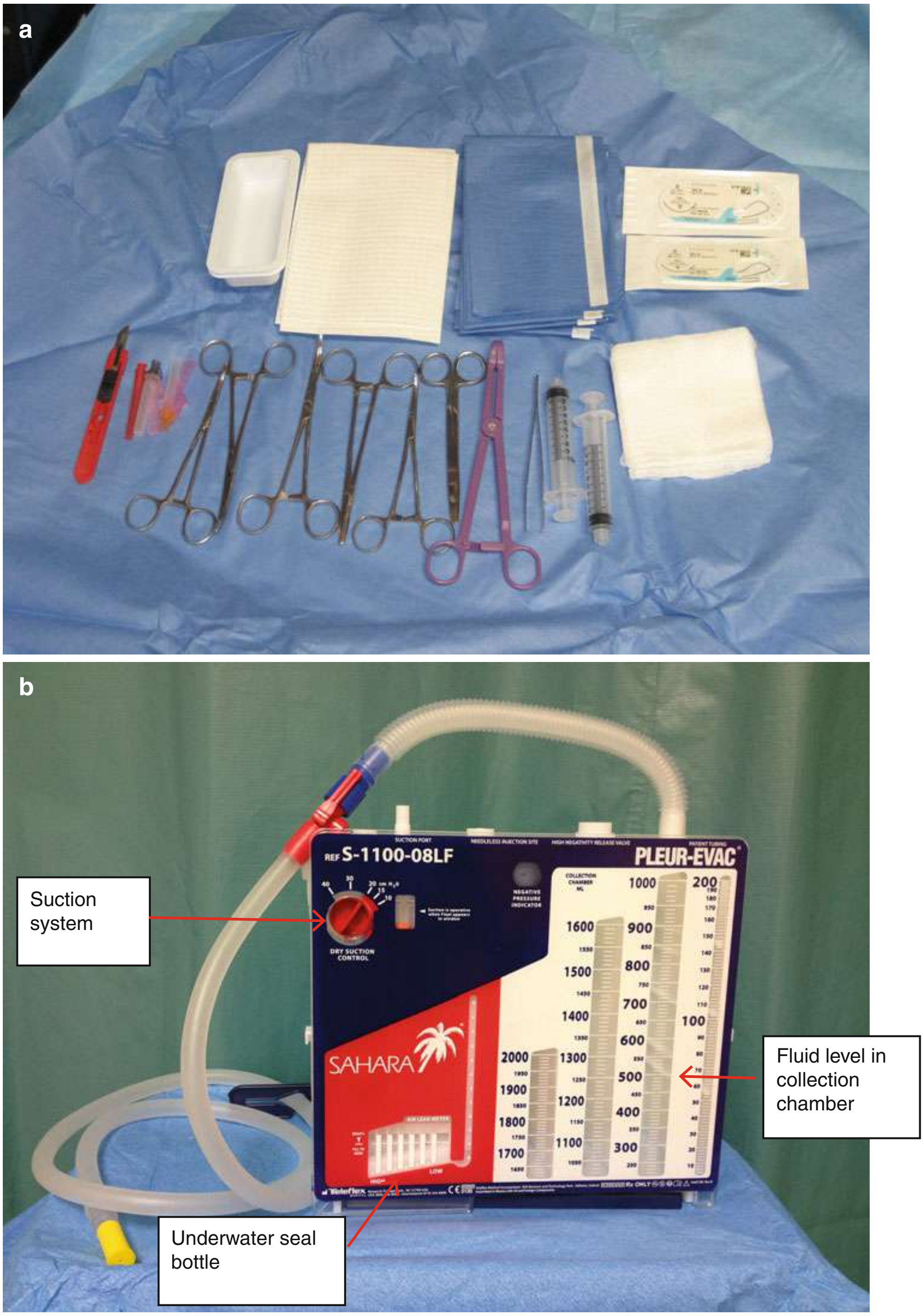
a
b
Suction system
Fluid level in collection chamber
Underwater seal bottle
REF S-1100-08LF
PLEUR-EVAC
SAHARA

Fig. 23.1 (**a**) Tube thoracostomy tray, (**b**) chest drainage system

23.4 **Procedure** (Fig. 23.2)

1. Sterile skin preparation with sterile drape.
2. Anesthetize the appropriate area subcutaneously up to and including the rib periosteum with 5 mL of 1 % lidocaine with epinephrine (Fig. 23.2a).
3. Using a #10 or #11 blade, make an approximately 4-cm skin incision over the desired intercostal level of entry (most often the fourth or fifth intercostal space in the midaxillary line) (Fig. 23.2b, c).

   If the incision is placed *below* the fifth intercostal space, the risk of subdiaphragmatic placement into the abdominal space is increased.
4. Bluntly dissect with a hemostat or Kelly clamp through the subcutaneous tissue to the level of the intercostal muscles with intermittent opening of the dissection instrument during advancement (Fig. 23.2d, e).
5. Digitally palpate the selected intercostal space and the superior margin of the inferior rib (pay careful attention to avoid the neurovascular bundle lying inferiorly) (Fig. 23.2f).
   - If time permits, additional analgesia is recommended at this point of the procedure.
6. Guiding the closed Kelly clamp over the upper margin of the rib, enter the chest wall into the pleural cavity. (This will require some controlled force and a twisting motion.) Once the pleural space is entered, a rush of air or fluid should occur (Fig. 23.2g).

   Uncontrolled force and a lunging motion can result in penetration to the lung, heart, liver, or spleen.
7. Open the Kelly clamp while still inside the pleural space and then withdraw while the clamp is still open to enlarge the dissected tract of entry and allow easier passage of the thoracostomy tube (TT).
8. Explore the dissected tract with a sterile finger to appreciate lung tissue and possible adhesions.
9. To estimate the length the TT is to be inserted, measure the distance between the skin incision and the apex of the lung. If preferred, place a clamp over the tube at the estimated length (Fig. 23.2h).
10. Grasp the proximal end of the TT with the large Kelly clamp and pass the tube through the thoracic cavity along the previously dissected tract.
11. Release the Kelly clamp and continue to advance the tube posteriorly and superiorly.

    Make sure all of the fenestrated holes of the TT are within the thoracic cavity to prevent unnecessary manipulation and/or replacement of the TT.
12. Once the TT is in the desired position, connect the tube to the drainage device (Fig. 42.1b). Once connected, release the cross clamp on the distal end of the TT.
13. Secure the TT to the skin with 0 or 1–0 silk or nylon suture. A simple, interrupted suture above and below the TT with each stitch wrapped tightly around the TT is recommended.

    Incomplete security of the TT leads to dislodging of the tube with routine patient movements.
14. Apply petrolatum gauze over the skin closure surrounding the TT and then apply a support dressing with 4×4 gauze and adhesive tape (4 in.).
15. Obtain a chest radiograph to confirm placement of the TT.

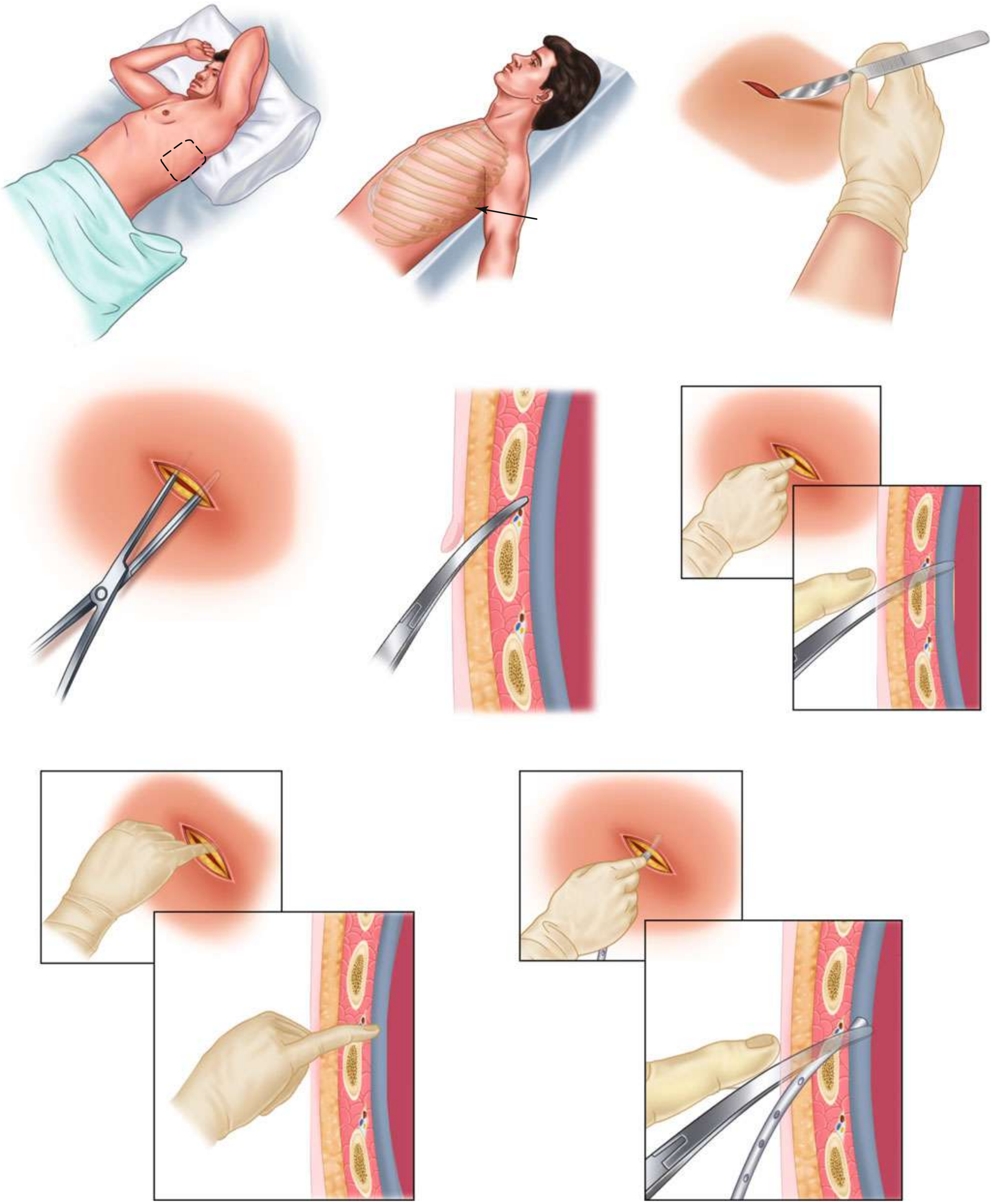

Fig. 23.2 Chest tube thoracostomy procedure

23.5 Complications

- Improper placement for pneumothorax
  - Reposition if:
    - Horizontal (over diaphragm)
    - Subcutaneous
    - Placed too far into the chest (against apical pleura)
  - Remove if:
    - Placed into the abdominal space
- Bleeding (local vs. hemothorax)
- Hemoperitoneum (liver or spleen injury)
- Tube dislodgment
- Empyema (TT introduces bacteria into the pleural space)
- Retained pneumothorax (may require second TT)
- Re-expansion pulmonary edema
- Subcutaneous emphysema

23.6 Pearls and Pitfalls

- Water seal acts as a one-way valve; if the system bubbles, there is an air leak.
- In the Pleur-evac® systems, there is an orange floater which, when static, means the desired suction pressure (usually 20 cmH_2O) has been reached.
- The negative pressure in the chest cavity equals the amount of water in water seal plus amount of suction.
- A chest tube can be removed when there is no air loss or blood for 24 h.
- When removing the tube, have the patient exhale and remove as quickly as possible.
- Leave petrolatum gauze in place for 48 h before changing it (allows wound to heal better).

Selected Reading

Ball CG, Lord J, Laupland KB, et al. Chest tube complications: how well are we training our residents? Can J Surg. 2007;50:450.

Collop NA, Kim S, Sahn SA. Analysis of tube thoracostomy performed by pulmonologists at a teaching hospital. Chest. 1997;112:709.

Dalbec DL, Krome RL. Thoracostomy. Emerg Med Clin North Am. 1986;4:441.

Miller KS, Sahn SA. Chest tubes. Indications, technique, management and complications. Chest. 1987;91:258–64.

Millikan JS, Moore EE, Steiner E, et al. Complications of tube thoracostomy for acute trauma. Am J Surg. 1980;140:738.

Thoracentesis

24

Lee Richard Donner and Michael Anana

24.1 Indications

- *Therapeutic thoracentesis* is performed to relieve dyspnea, hypoxia, or otherwise compromised respiratory function due to a large pleural effusion.
- *Diagnostic thoracentesis* is performed to aid in the diagnosis and workup of:
  - Pleural effusions of unknown cause
  - Unilateral pleural effusions
  - Pleural effusions originally determined to be due to heart failure but persisting after 3 days of diuresis

24.2 Contraindications

- Absolute
  - None
- Relative:
  - Coagulopathy, thrombocytopenia.
  - Small or loculated pleural effusion. These will increase the risk of missing the effusion and causing lung injury.
  - Positive-pressure ventilation.
  - Skin infection over the needle insertion site.

24.3 Materials and Medications

- Thoracentesis tray (commercially available kits generally include the items in the following list) (Fig. 24.1)
  - (1) Fenestrated drape
  - (1) 25-gauge × 1-in. needle
  - (1) 21-gauge × 1.5-in. needle
  - (1) 8-French catheter over 18-gauge needle
  - (1) Small plastic syringe, 5 mL
  - (1) Small plastic syringe, 10 mL
  - (1) Large plastic syringe, 50–60 mL
  - (1) Three-way stopcock
  - Specimen vials and caps
  - (1) Collection bag, 1500 mL, or vacuum container
  - (1) Tubing set
  - (1) Hemostat
  - Betadine (povidone-iodine) or other skin antiseptic preparing solution
  - 10-mL lidocaine 1 % without epinephrine

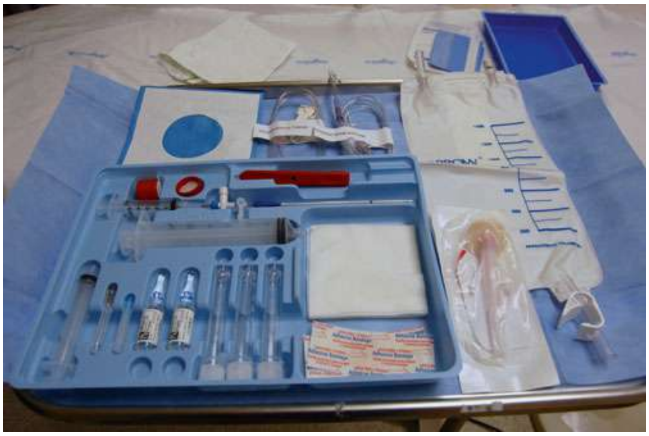

Fig. 24.1 Typical commercial thoracentesis tray

L.R. Donner, MD (✉)
Emergency Medicine Department, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: LeeDonnerMD@gmail.com

M. Anana, MD
Emergency Department, University Hospital, Rutgers New Jersey Medical School, Newark, NJ, USA
e-mail: ananami@njms.rutgers.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_24

24.4 Procedure

1. Place patient in seated position with arms at rest on a bedside table (Fig. 24.2).
2. The location and height of the pleural effusion are confirmed by physical examination. Auscultation of decreased or absent breath sounds, dullness to percussion, and decreased tactile fremitus are physical findings to confirm the location and height of the effusion (Fig. 24.3).
3. Determine and mark the site of needle insertion. This will be at the midscapular line one or two intercostal spaces below the maximum height of the effusion as determined by a combination of imaging and physical examination (Fig. 24.4).
    - *Do not* attempt thoracentesis inferior to the eighth intercostal space because respiratory cycle and anatomical variation place the diaphragm and intra-abdominal organs at risk.
4. Prepare the skin with the sterile skin preparation and sterile drape.
5. Anesthetize the appropriate area subcutaneously using a 25-gauge needle on a 10-mL syringe. Create a wheal in the skin and then infiltrate up to and including the periosteum of the rib inferior to the landmarked space with 5 mL of 1 % lidocaine.
    - Remember that the neurovascular bundle is found at the inferior border of each rib. Keep this in mind as you approach the rib and throughout the rest of the procedure to avoid injury to these structures (Fig. 24.5).
6. Using a 22-gauge needle, slowly walk the needle up and over the superior border of the rib. Continue to advance the needle along the superior border of the rib with the syringe withdrawn and infiltrating lidocaine intermittently along the way.
7. Pleural fluid will be aspirated once the pleural space is reached. Do not advance the needle further. Deposit 1–2 mL more lidocaine. A hemostat can be placed on the needle at the level of the skin to mark the depth of the pleural space and the needle can be removed.
8. Some commercial kits may come with an adjustable depth guard to be positioned at the determined depth. On an 18-gauge catheter-over-needle device, position the depth guard to the appropriate depth determined from the prior step. If a depth guard is not available, use the index finger and thumb on the catheter at the appropriate depth. With the 5-mL syringe attached, advance the device over the superior border of the rib while aspirating, expecting pleural fluid to return again at the determined depth (Fig. 24.6).
9. When the pleural space is reached, do not advance the needle further. Advance the catheter over the needle until the hub reaches the skin. Remove the needle during expiration and immediately cover the open hub with a gloved finger to prevent development of pneumothorax. Some kits provide catheters with one-way valves to prevent air entry.
10. Attach the 50–60-mL syringe to the catheter via the three-way stopcock. Pleural fluid can be drained and transferred to appropriate collection vials for diagnostic thoracentesis. A collection bag may be attached with tubing to the third port of a three-way stopcock for larger volume evacuation in the case of a therapeutic thoracentesis. Employ a syringe pump method to drain 50–60 mL of fluid at a time to the collection bag. Fill the syringe by withdrawing the plunger while the stopcock is closed to the bag. Then, close the stopcock to the patient and pump the contents of the syringe to the bag. Next, close the syringe closed to the bag, and repeat the cycle until the desired volume is drained. A vacuum container is an alternative that simply attaches via tubing to the stopcock (Fig. 24.7).
    - If using a three-way stopcock and a device that does not have a one-way valve on the catheter, be sure to always keep the stopcock closed to the patient unless withdrawing fluid in order to decrease the risk of pneumothorax.
11. When the desired amount of pleural fluid is obtained, remove the catheter during expiration and apply an occlusive dressing.

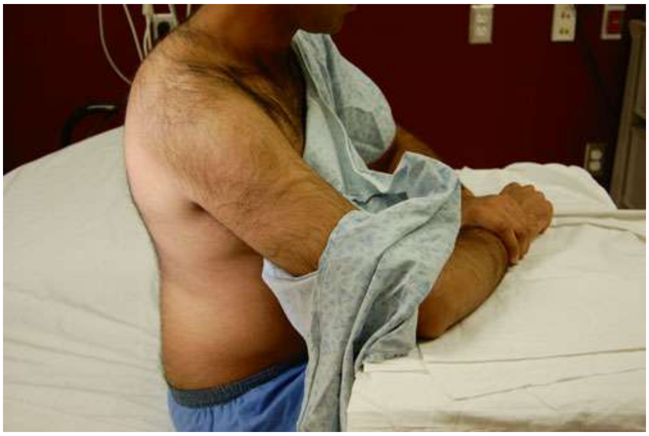

Fig. 24.2 Patient in upright, seated position

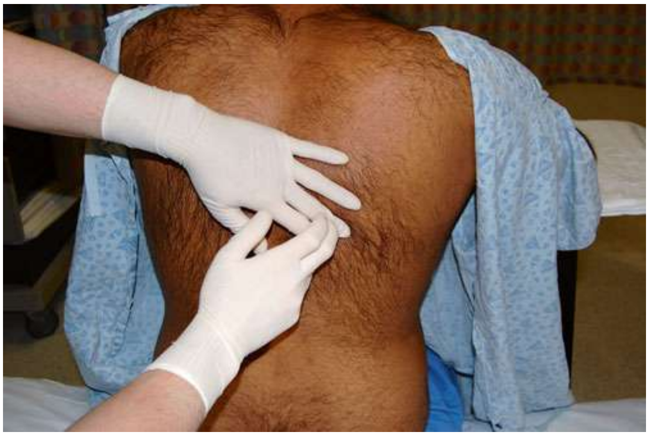

Fig. 24.3 Determining the location and height of the pleural effusion

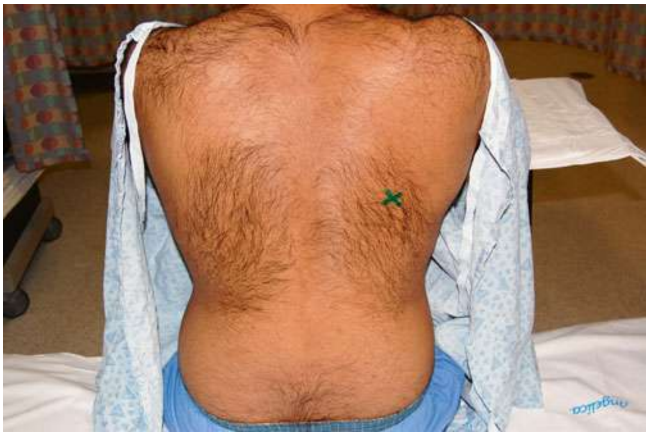

Fig. 24.4 Marked site for needle insertion

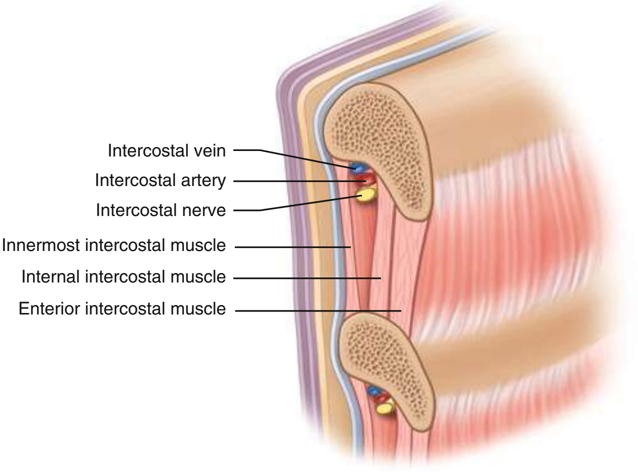
Intercostal vein
Intercostal artery
Intercostal nerve
Innermost intercostal muscle
Internal intercostal muscle
Enterior intercostal muscle

Fig. 24.5 The intercostal neurovascular bundle

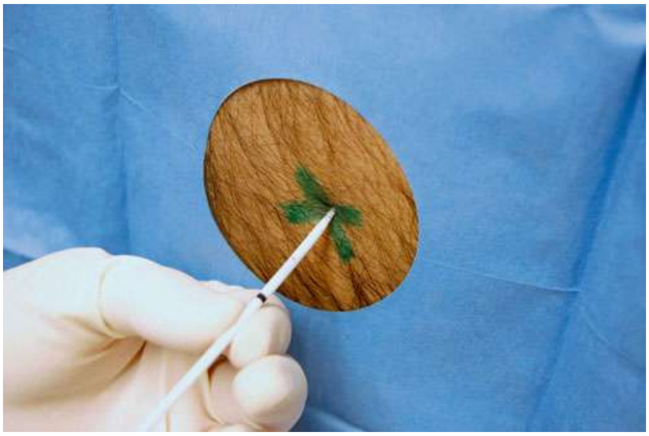

Fig. 24.6 Needle and catheter insertion, thumb and index finger at desired depth

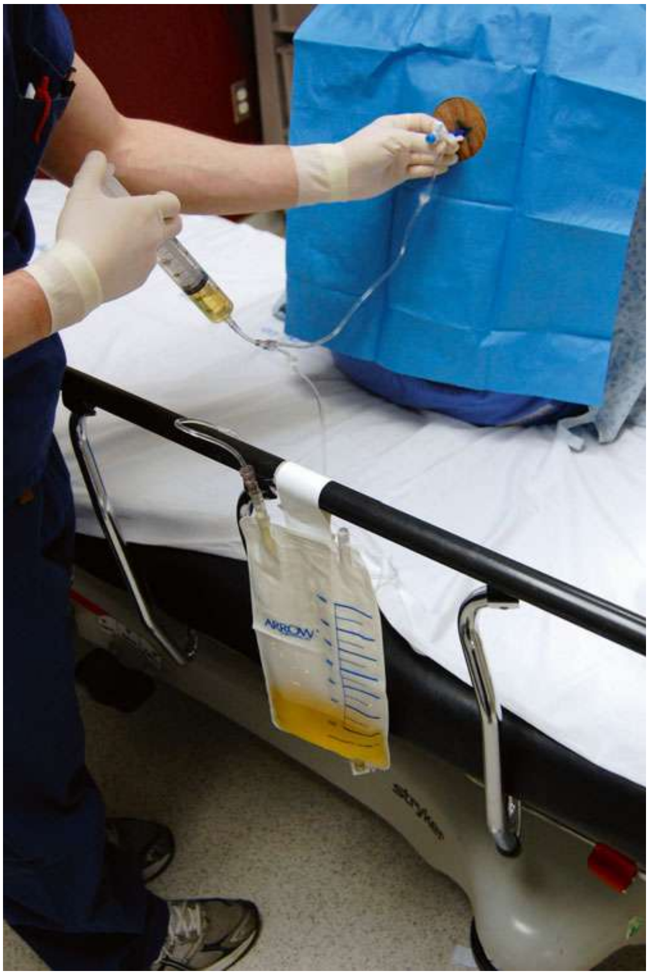

Fig. 24.7 Pleural fluid collection via syringe pump method

24.5 Complications

- Pneumothorax
- Re-expansion pulmonary edema
- Hemothorax, hematoma
- Intra-abdominal organ injury
- Air embolism
- Empyema

24.6 Considerations

- If available, the use of bedside ultrasound is highly recommended because ultrasound guidance has been shown to substantially reduce the risk of pneumothorax. Before the procedure, the height, width, and depth of the effusion can be appreciated by scanning the chest and viewing the effusion through the intercostal spaces. The use of ultrasound aids in selecting the needle insertion site by:
  - Visualizing the distance the needle must pass to reach the parietal pleura
  - Confirming the thickness of the effusion in the site selected is at least a minimum of 1.5 cm
  - Providing the clinician with a view of the effusion and surrounding structures through the complete respiratory cycle

  With these items in mind, the needle insertion site can be selected with confidence and marked before beginning the procedure.
- In addition, the use of bedside ultrasound in real time will allow the clinician to visualize the needle as it passes toward and enters the pleural space. This use requires sterile probe covers.
- Re-expansion pulmonary edema is a rare but feared complication of thoracentesis. The cause is not fully understood. Historically it was thought that re-expansion pulmonary edema was caused by removing too large a volume of fluid from the pleural space (>1–1.5 L). Another theory is that re-expansion pulmonary edema is caused when great negative intrapleural pressures (<20 cmH_2O) are generated during the procedure. The low incidence of this complication has yielded inconclusive evidence.
  - In light of this, it is prudent to continue to limit the volume of pleural fluid removed to no more than 1–1.5 L. Pleural manometry is not widely available for use in the emergency department, but should also be considered if available to maintain intrapleural pressures from reaching more negative values.

24.7 Pearls and Pitfalls

- Bedside ultrasound reduces the risk of complications.
- Never attempt thoracentesis below the eighth intercostal space.
- Thoracentesis should not be performed on pleural effusions demonstrated to be less than 1–1.5 cm thick by ultrasound or on a lateral decubitus film.
- Positive-pressure ventilation mandates extreme care while performing thoracentesis because the lungs can be punctured during inflation. In addition, ventilated patients will not be able to sit upright and will need to have the procedure performed in lateral decubitus position with the effusion side down and the posterior axillary line as the needle insertion site. Alternatively, the patient can be placed supine with the head of the bed elevated to 45° and the midaxillary line as the needle insertion site.
  - Routine post-procedure radiographs are not necessary to exclude pneumothorax. Indications for post-procedure imaging include onset of chest pain during the procedure, persisting cough or chest discomfort after the procedure, air aspiration along any step of the procedure, or positive-pressure ventilation.
  - Pneumothorax rates are higher for inexperienced providers. Although an effusion compromising respiratory function is considered a clear indication for therapeutic thoracentesis, the performance of nonurgent diagnostic thoracentesis might best be delayed for those more practiced in the procedure.

Selected Reading

Dewitz A, Jones R, Goldstein J. Additional ultrasound-guided procedures. In: Ma OJ, Mateer JR, Blavias M, editors. Emergency ultrasound. 2nd ed. New York: McGraw-Hill; 2008. p. 546–50.

Feller-Kopman D, Berkowitz D, Boiselle P, Ernst A. Large-volume thoracentesis and the risk of reexpansion pulmonary edema. Ann Thorac Surg. 2007;84:1656–61.

Gordon CE, Feller-Kopman D, Balk EM, Smetana GW. Pneumothorax following thoracentesis: a systematic review and meta-analysis. Arch Intern Med. 2010;170:332–9.

Light RW. Pleural effusion. N Engl J Med. 2002;346:1971–7.

Thomsen TW, DeLaPena J, Setnik G. Videos in clinical medicine. Thoracentesis. N Engl J Med. 2006;355, e16.

Open Chest Wounds and Flail Chest

25

Jacob J. Glaser and Carlos J. Rodriguez

25.1 Background

Thoracic injuries are commonly associated with penetrating and blunt abdominal trauma and are implicated in 50–70 % of trauma deaths [1]. Cardiac tamponade, tension pneumothorax, massive hemothorax, airway obstruction, flail chest, and open pneumothorax represent the six immediately life-threatening injuries attributed to chest trauma [2]. Accordingly, they must be accurately identified and dealt with urgently.

Open pneumothorax ("sucking" chest wound) is seen in penetrating chest injuries. If the associated chest wound is greater than 2/3 the diameter of the trachea (generally anything greater than 1.5–2 cm), air can preferentially enter the intrapleural space, via the trachea, with each inspiration [3] (Fig. 25.1). This allows equilibration of pressure between the pleural space and the atmosphere, causing the lung to collapse and leading to profound hypoventilation and hypoxia.

Flail chest results from high-energy blunt, crushing chest trauma causing two or more fractures in two or more contiguous ribs. Classically, the fractures are lateral or sternal. Posterior rib fractures rarely cause flail physiology (Fig. 25.2). Flail chest has been reported to have mortality as high as 16 % [4]. This injury pattern is associated with a high incidence of underlying pneumothorax, hemothorax, pulmonary contusion, and chest wall instability. Mortality from flail chest is thought to be correlated with the degree of underlying pulmonary contusion and attendant hypoxia [2].

Both open pneumothorax and flail chest are immediately life threatening and require early appropriate management.

J.J. Glaser, MD (✉)
Combat Casualty Care Directorate, Naval Medical Research Unit,
San Antonio, TX, USA
e-mail: Jacob.glaser1@gmail.com

C.J. Rodriguez, DO
Division of Trauma Surgery, Surgical Critical Care, Walter Reed
National Military Medical Center, Bethesda, MD, USA
e-mail: Carlos.J.Rodriguez.mil@health.mil

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_25

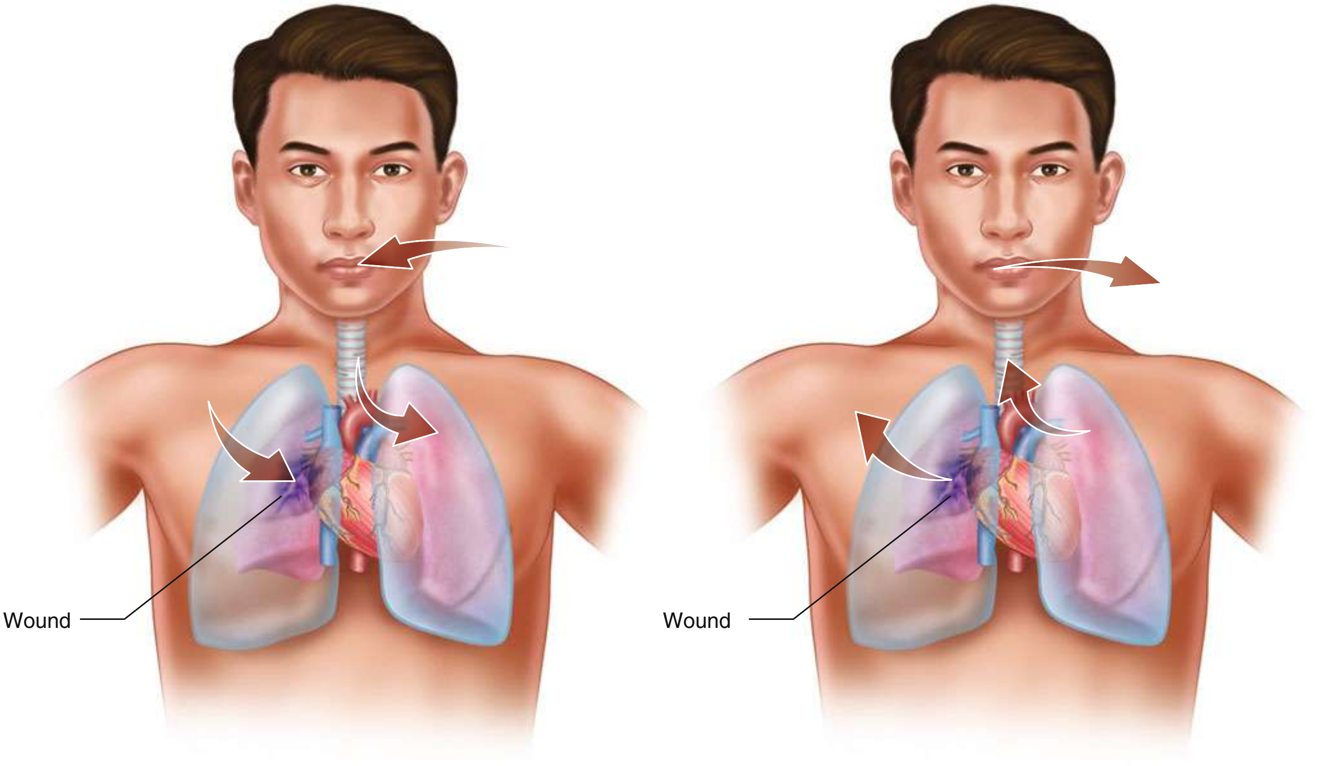
Wound
Wound

Fig. 25.1 Air preferentially will enter the chest via the wound, collapsing the lung on the affected side

Fig. 25.2 Flail chest: two or more rib fractures in two or more segments. Lateral flail (most common) and sternal flail segments represented below

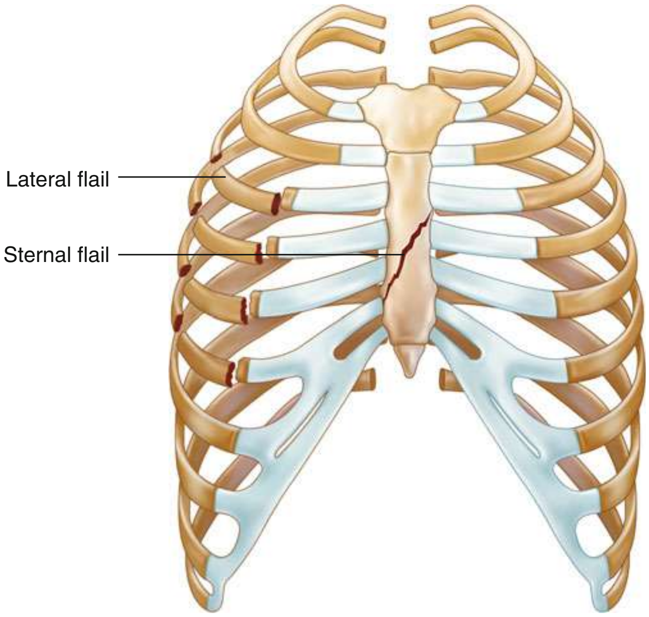
Lateral flail
Sternal flail

25.2 Initial Evaluation of Thoracic Trauma

- To the best extent possible, obtain good patient handover from the prehospital caregivers.
- Initial physical evaluation.
  - Appropriate attention should be given to the ABCs (airway, breathing, circulation) of ATLS (Advanced Trauma Life Support) management.
  - Evaluation and resuscitation are to be *concurrent* with diagnostic procedures and immediate interventions.
  - Maintain a high degree of suspicion for open chest wound in impalement injury and destructive penetrating trauma (blast injury or shotgun).
  - Maintain a high degree of suspicion for flail chest in high-energy direct impact trauma (motor vehicle crash, fall, crush injury).
- Administer high-flow O_2 with a non-rebreather mask.
  - If patient is in respiratory distress, is unstable, or has an obvious chest wall defect, consider early intubation to secure the airway.
  - Inspect the chest wall for occlusive dressings.
- Decompress the chest if tension physiology is present or suspected.
  - Immediate decompression of a suspected tension pneumothorax can be readily accomplished by removing any existing occlusive dressings *or*
  - Place a large-bore cannula over the rib, second intercostal space, or midclavicular line.
- Specific immediate management appropriate to open chest wound and flail chest (see later).
- Monitor continuous pulse oximetry and electrocardiogram.
- Initiate crystalloid resuscitation via large-bore intravenous (IV) access.
  - Early mobilization of blood products if ongoing hemorrhage or expectation of excessive blood loss.
  - Placement of resuscitative lines concurrent with management of respiratory parameters.
- Early surgical consultation for management of intrathoracic injuries and management of chest wall defect.

25.3 Open Pneumothorax ("Sucking Chest Wound")

- Immediate management requires attention to airway and respirations.
  - If in respiratory distress, *intubate*.
- Close the chest wall defect with an occlusive three-sided dressing.
  - This includes a valve mechanism that allows trapped air to escape, preventing tension (Fig. 25.3).
  - An IV bag cut to fit the wound and then taped on three sides can be useful in an emergent situation.
  - Commercial products are available and appropriate for smaller wounds, including the Asherman Chest Seal and HyFin Vent.
- A completely occlusive dressing may quickly convert an open chest wound into a tension pneumothorax [2, 3, 5] and therefore should *never* be done.
  - The patient and dressing must be serially checked to ensure that trapped air is allowed to escape.
  - If there is any doubt, immediately remove the dressing and replace it with an appropriate dressing.
- These maneuvers are a bridge to definitive care.
- When the timing is appropriate (i.e., time and resources are available), perform a formal tube thoracostomy and convert to a completely occlusive dressing over the wound.
  - Avoid placing the tube through the open wound.
- Once tube thoracostomy, placement of occlusive dressing, and the airway are secured, the pathophysiology of the open pneumothorax becomes inconsequential.
- Immediate consultation with surgery for definitive care of associated intrathoracic injuries is required.
  - The patient may need urgent thoracotomy to treat associated injuries.
  - Irrigation and debridement should take place in the operating room.
  - Depending on injury severity, the patient may need chest wall reconstruction.

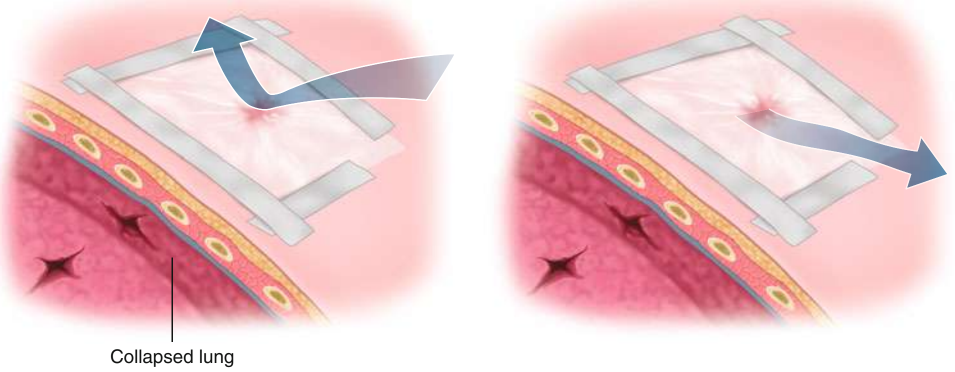

Fig. 25.3 Three-sided dressing, allowing a valve to decompress tension in the chest

25.4 Flail Chest

- Diagnosis is made from mechanism and examination, not radiographically.
  - With inspiration, the affected chest will move inward with negative pressure.
  - With expiration, the affected chest will move outward.
  - Patients who are intubated on positive pressure (and not spontaneously breathing) will often not show this paradoxical chest movement.
- Attention *must* be paid to presumed underlying blunt lung injury and contusion.
  - The degree of underlying contusion (not the flail segment itself) is directly related to the degree of hypoxia and associated morbidity and mortality [2, 3].
  - There should be a low threshold for intubation to manage respiratory distress, hypoxemia, or hemodynamic instability.
  - The patient is at high risk for hemothorax and pneumothorax requiring tube thoracostomy.
- Abdominal injuries may be present in up to 15 % of patients [2].
- After initial stabilization, treatment is supportive.
  - Intensive care unit admission for management of underlying pulmonary contusion.
  - Pain control in the form of epidural or regional block for excellent pain control.
  - Close attention to pulmonary toilet and lung re-expansion.
  - Patients require observation and treatment in a monitored setting until ensured that respiratory parameters and oxygenation are improving.
- Surgical stabilization of the chest wall is rarely performed.
  - Early surgical consultation for chest wall fixation in questionable cases is warranted.

References

1. LoCicero J, Mattox KL. Epidemiology of chest trauma. Surg Clin North Am. 1989;69:15–9.
2. Pietzman AB, Schwab CW, Yealy DM, editors. The trauma manual. 2nd ed. Philadelphia: Wolters Kluwer Health; 2000.
3. Weinberg JA, Croce MA. Chapter 33: Chest wall injury. In: Flint L, Meredith JW, Schwab CW, editors. Trauma: contemporary principles and therapy. 1st ed. Philadelphia: Lippincott Williams & Williams; 2007.
4. Clark GC, Schecter WP, Trunkey DD. Variables affecting outcome in blunt chest trauma: flail chest vs. pulmonary contusion. J Trauma. 1988;28:298–304.
5. Borden Institute Walter Reed Army Medical Center. Emergency war surgery. 3rd ed. Washington: Office of the Surgeon General U.S. Army, Borden Institute; 2004.

Emergent Resuscitative Thoracotomy, Open Cardiac Massage, and Aortic Occlusion

26

Kevin M. Jones and Jay Menaker

26.1 Indications

- Penetrating chest trauma with recent or immanent loss of vital signs
- Consider in blunt trauma with pericardial tamponade or exsanguination where aortic occlusion may provide proximal control

26.2 Contraindications

Absolute
  - Prolonged cessation of vital signs
  - Injury profile obviously incompatible with life
  - Absence of surgical services to whom care can be transferred

Relative
  - None

26.3 Materials and Medications (Fig. 26.1)

- Betadine (povidone-iodine) for rapid skin preparation
- #10 Scalpel
- Mayo or long Metzenbaum scissors
- Finochietto retractor (rib spreader)
- Long DeBakey or other tissue forceps (2)
- Satinsky vascular clamp and/or straight vascular clamp
- Long needle holders (2)
- Lebsche knife or sternal osteotome with hammer
- Lap sponges or gauze pads

K.M. Jones, MD (✉)
Trauma and Surgical Critical Care, R Adams Cowley Shock Trauma Center, University of Maryland Medical Center, Baltimore, MD, USA
e-mail: KJones12@umm.edu

J. Menaker, MD
Departments of Surgery and Emergency Medicine, R Adams Cowley Shock Trauma Center, University of Maryland Medical Center, Baltimore, MD, USA
e-mail: jmenaker@umm.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_26

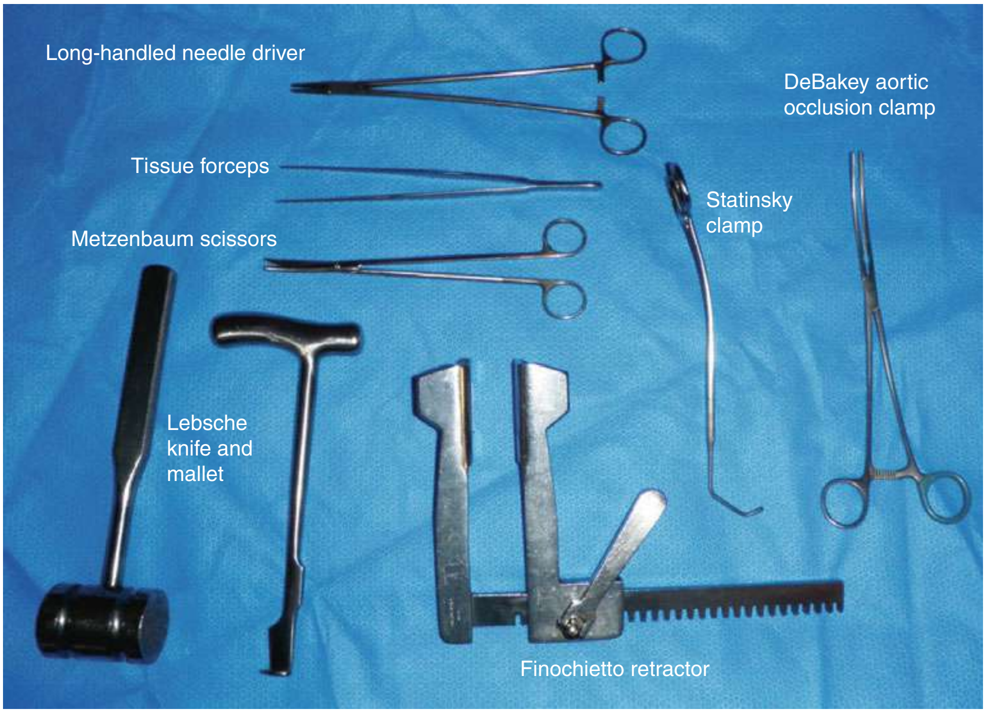
Long-handled needle driver
DeBakey aortic occlusion clamp
Tissue forceps
Statinsky clamp
Metzenbaum scissors
Lebsche knife and mallet
Finochietto retractor

Fig. 26.1 Basic components of an emergency thoracotomy tray: Lebsche knife and mallet for crossing sternum (bone-cutting forceps or sternal osteotome would also suffice), Finochietto retractor, atraumatic vascular clamps (a Satinsky clamp and a DeBakey aortic occlusion clamp), long-handled needle driver, tissue forceps, and Metzenbaum scissors. *Not illustrated* Scalpel with #10 or #20 blade, Mayo scissors

26.4 Procedure

26.4.1 Resuscitative Thoracotomy and Open Cardiac Massage

1. Rapidly prepare the entire anterior and bilateral chest with Betadine.
2. Using a scalpel, incise the skin and subcutaneous tissue from just right of the sternum to the anterior margin of the left latissimus dorsi, following the curvature of the inframammary crease or the fourth or fifth intercostal space (Fig. 26.2).
   - This incision is often made too low on the patient's chest. It should extend across the sternum, not at the level of the xiphoid. Upward retraction of the breast may help provide access to the fourth or fifth intercostal space, where the incision should be located.
3. Bluntly enter the right pleural space through the fourth or fifth intercostal space.
4. Using scissors, cut the intercostal muscles, dividing between the fourth and the fifth ribs from the sternum to the posterior axillary line.
5. If better exposure to the heart is desired, some practitioners advocate extending the incision across the sternum using a Lebsche knife, sternal osteotome, or bone-cutting forceps at this time (Fig. 26.3).
   - If the thoracotomy incision extends to or through the sternum, tie off the internal mammary arteries before closing the chest should the patient be successfully resuscitated, because these will have been divided with the incision.
6. Insert a Finochietto retractor and retract the ribs in order to gain access to the left chest and expose the pericardium (Figs. 26.4 and 26.5).
   - Insert a Finochietto retractor with the rack and pinion bar down and lateral, as in Fig. 26.4, so as not to interfere with extension of the thoracotomy across the sternum into a clamshell maneuver if needed.
7. If massive left pleural hemorrhage is encountered, investigate and control the source at this time.
8. Using tissue forceps, raise a portion of the pericardium anterior to the phrenic nerve, and enter the pericardium using scissors (Fig. 26.6).
9. Widely open the pericardium with scissors, cutting in a cranial-caudal direction anterior to the phrenic nerve and deliver the heart (Fig. 26.7).
10. If hemopericardium is encountered, investigate for and initiate appropriate repair of identified cardiac injuries.
11. Initiate open cardiac massage by cupping the heart between the flattened palmar aspect of the fingers of both hands and rhythmically compressing the heart from apex to base, relaxing completely between compressions to allow filling (Fig. 26.8).

26.4.2 Aortic Occlusion

1. Expose the posterior aspect of the left mediastinum by having an assistant retract the left lung superomedially, dividing the inferior pulmonary ligament if necessary (Fig. 26.9).
2. Bluntly dissect the pleura separating the pleural and mediastinal space just anterior to the vertebral bodies, exposing the aorta (Fig. 26.9).
3. Completely encircle the aorta with the finger of the nondominant hand (Fig. 26.10).
   - Differentiating the aorta from the esophagus when the patient is in a state of profound shock is very difficult. Having an assistant pass an orogastric tube may help distinguish the two. The aorta should be the most posterior structure, lying immediately on the anterior aspect of the vertebral bodies.
4. With the aorta completely encircled, place a vascular clamp across the aorta and verify by sight and feel that the complete vessel is occluded within the clamp (Fig. 26.11).

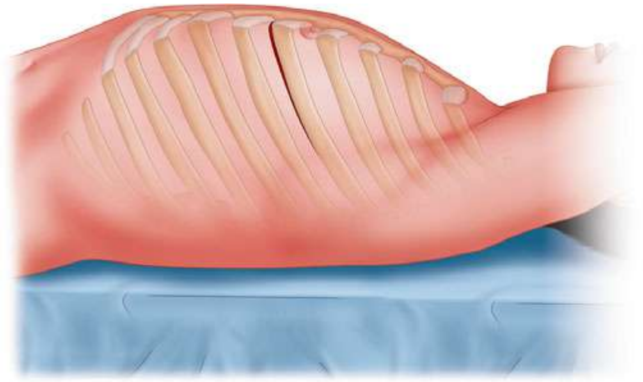

Fig. 26.2 Raise the left arm above the head. Make an incision along the left fourth intercostal space, just below the nipple in a male or at the inframamillary crease in a female

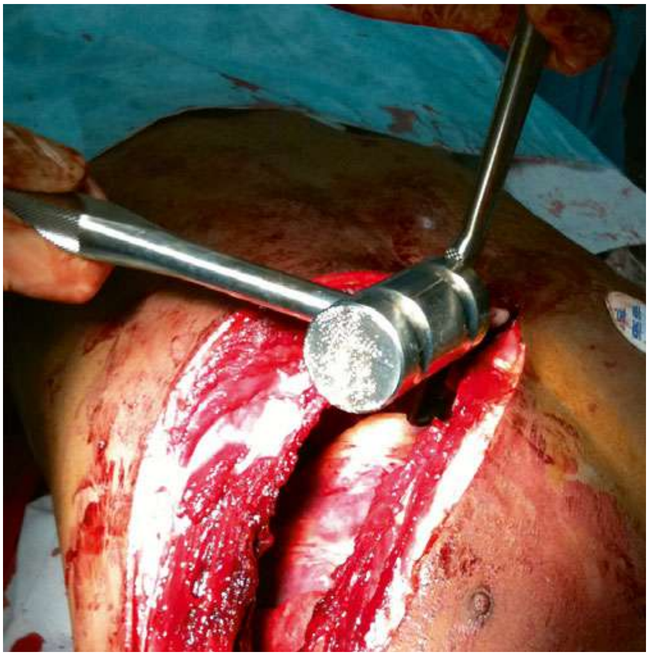

Fig. 26.3 Use a Lebsche knife to extend the incision across the sternum to improve exposure to the heart

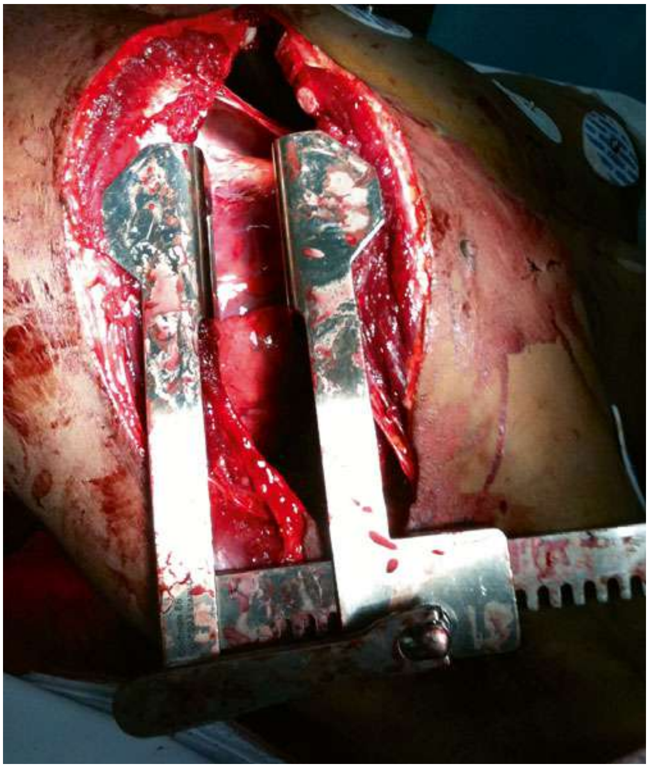

Fig. 26.4 Finochietto retractor placed through the fourth intercostal anterolateral incision. Note the rack and pinion bar placed posterolaterally, where it will not impede access to the midline

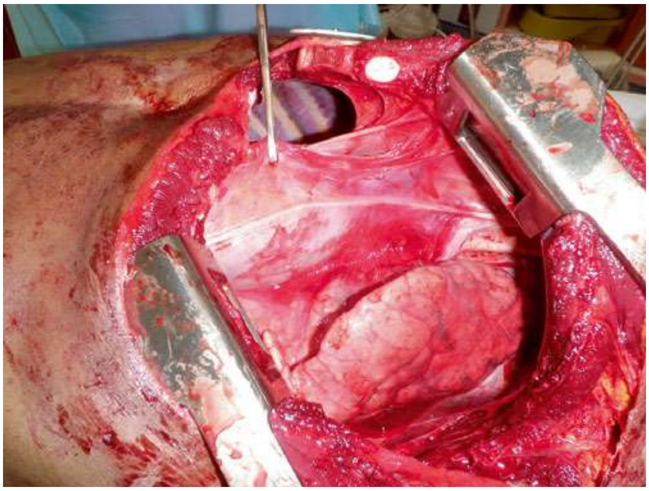

Fig. 26.5 Finochietto retractor extended, exposing the left pleural space and the pericardium. In this case the sternum has been divided as above, and the resultant window into the right pleural space is seen anterior to the pericardium. The lung is deflated in this postmortem picture but would be far more an obstacle in the actively ventilated patient. The clamp seen in the upper portion of the picture is reapproximating the pericardium, which has previously been divided, for the sake of illustration

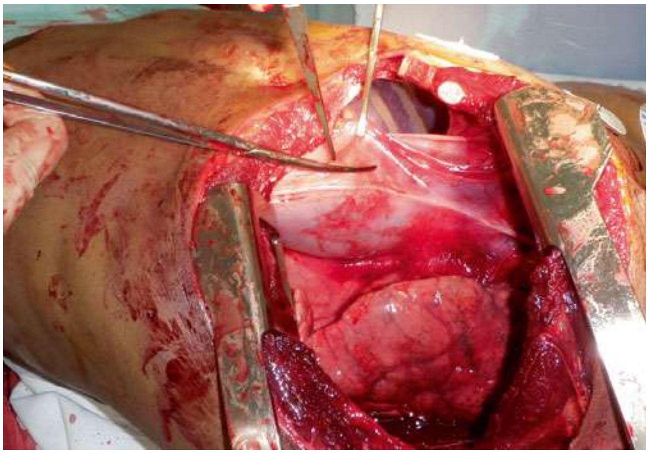

Fig. 26.6 The pericardium is lifted with tissue forceps and opened with a nick using scissors. The phrenic nerve is easily visualized running cranial-caudal just below the scissors in this picture

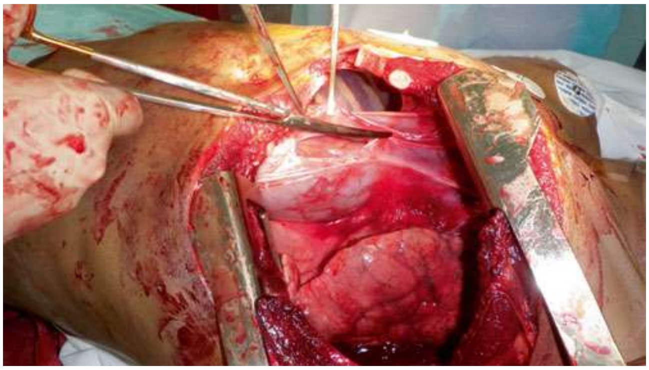

Fig. 26.7 Open the pericardium widely in the cephalad-caudad plane anterior to the phrenic nerve, taking care not to damage the phrenic nerve

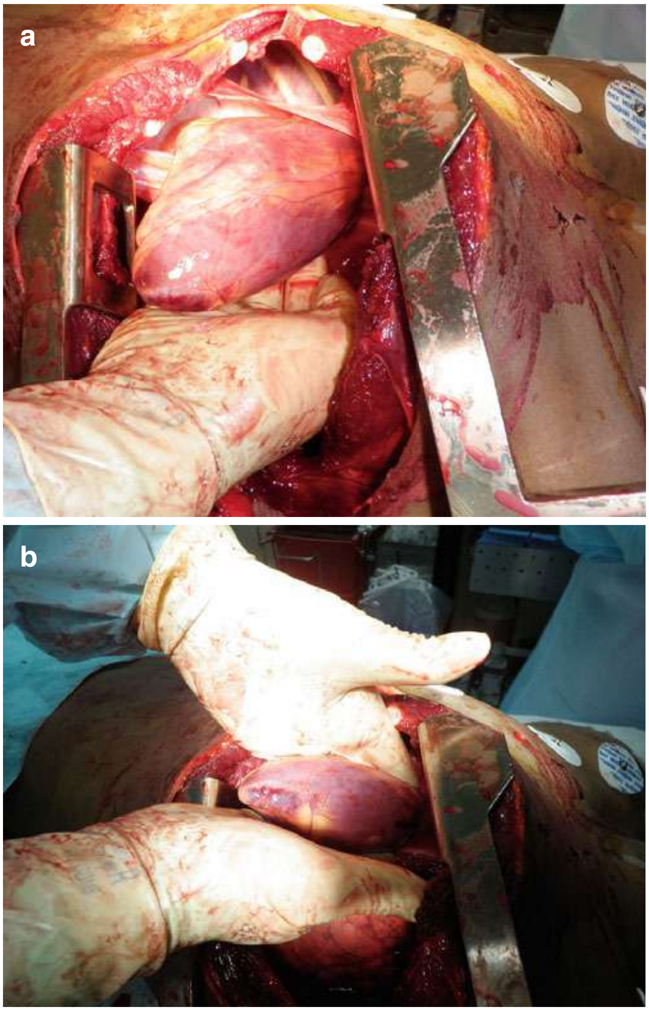

Fig. 26.8 (**a**) Deliver the heart from the pericardium, rapidly assess for cardiac injury requiring damage-control repair; (**b**) initiate open cardiac massage

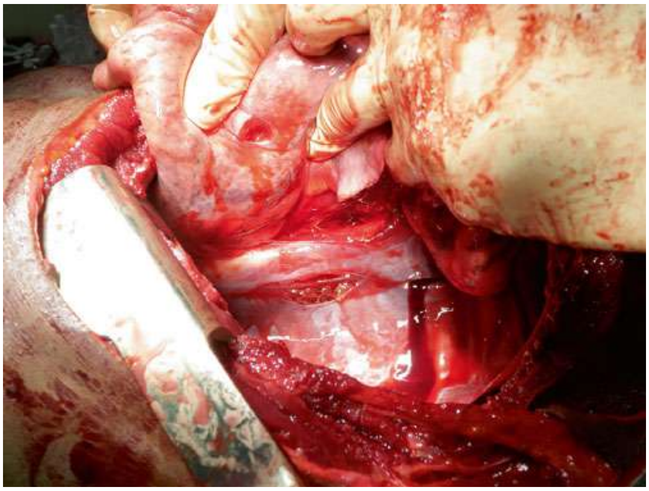

Fig. 26.9 The heart and lung are retracted superomedially, allowing visualization of the left posterior mediastinal pleura; the aorta lies just anterior to the vertebral bodies and has been previously isolated through the pleural interruptions seen here. The heart and lung are assertively retracted here for the benefit of illustration of the posterior mediastinum. Such retraction would completely occlude venous return

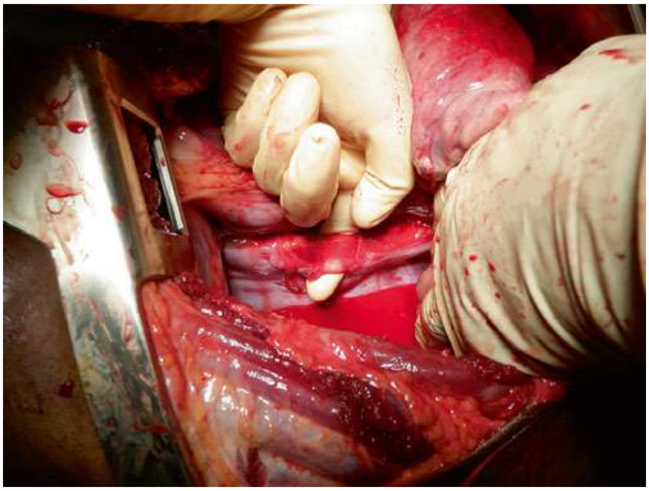

Fig. 26.10 After bluntly dissecting the mediastinal pleura, the aorta is looped using a finger of the nondominant hand

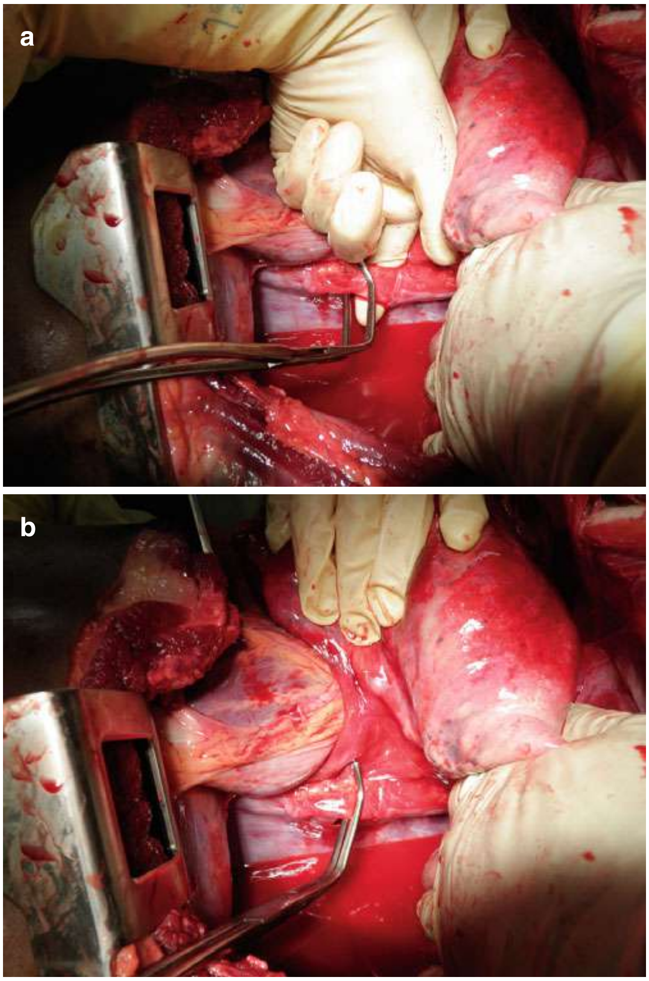

Fig. 26.11 (**a**) A vascular clamp (a Satinsky clamp is used here, although any large atraumatic vascular clamp can be used) is applied across the descending aorta, (**b**) followed by visual and tactile confirmation that the aorta is completely occluded

26.5 Complications

- Injury to care providers, by means of scalpel, needlestick, or sharp foreign body, is the principal concern.
- Post-emergency department thoracotomy infections are rare, even given the less than optimal sterile conditions.
- Damage to lung parenchyma during the initial incision is common and often leads to air leak in survivors.
- Neglect of the mammary arteries, often divided during emergent thoracotomy and not briskly bleeding in the shock state, will result in intrathoracic hemorrhage if not tied off.

Selected Reading

Bartlett RL. Resuscitative thoracotomy. In: Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 3rd ed. Philadelphia: WB Saunders; 1998.

Feliciano DV, Mattox KL. Indications, technique, and pitfalls of emergency center thoracotomy. Surg Rounds. 1981;4:32.

Siemans R, Polk Jr MC. Indications for thoracotomy following penetrating thoracic injury. J Trauma. 1977;17:493.

Wall Jr MJ, Huh J, Mattox KL. Indications and techniques of thoracotomy. In: Feliciano DV, Mattox KL, Moore EE, editors. Trauma. 6th ed. New York: McGraw Hill; 2008.

Lung Ultrasonography

27

Ali H. Dabaja, L. Connor Nickels, and Rohit Pravin Patel

27.1 Indications

- Evaluation of respiratory failure and insufficiency due to pneumothorax, pleural effusion, pulmonary edema, acute respiratory distress syndrome (ARDS), and alveolar consolidation (atelectasis, pneumonia, aspiration)
- Monitoring progress of diseases such as pulmonary edema and pneumothorax
- Procedural guidance during pleural fluid removal or pneumothorax treatment
- Procedural guidance for chest tube placement for complex pleural effusions, hemothorax, pneumothorax, and other pleural diseases

27.2 Contraindications

- None

27.3 Materials

- Ultrasound equipment/machine, ultrasonography gel
- Microconvex or phased-array cardiac probe for sufficient evaluation of lung artifact (use abdominal preset). Linear-array probes, or vascular probes, can be used for more detailed evaluation of the pleura, although this is insufficient for penetration to deeper structures.
- Sterile materials and equipment where appropriate

27.4 Preparations

- Clean ultrasound equipment and use sterile operating procedures and ultrasound conductive gel whenever adequate.
- Ultrasound positioning. Place ultrasound device where it is easily accessible and in adequate view.
- Proper environment. Minimize light interference. Close shades and lights where appropriate.
- Position the patient. This helps optimize and expedite the examination. Supine position should be adequate for critically ill patients. Upright positioning is best for all others, with arms abducted above the head or spaced away from the chest to allow space for the probe. Patients should be moved to the edge of the bed. Uncover appropriate areas, including anterior and lateral chest walls.
- Equipment use. Turn on equipment, enter patient data, and select proper probe (cardiac probe or abdominal probe with abdomen presets). Depth and gain should be adjusted accordingly, where depths in the range of 4–10 cm give proper evaluation of more superficial and deeper structures, respectively.

A.H. Dabaja, DO
Division of Critical Care, Department of Anesthesia, University of Florida Health, Gainesville, FL, USA
e-mail: ali.dabaja@gmail.com

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_27

27.5 Procedure

1. Probe position (Fig. 27.1). Start examination at the midclavicular level at the space between the second and the third ribs. Probe positioning should be perpendicular to the ribs (longitudinal positioning) with the ultrasound marker pointed cephalad. This should place the most superficial structures at the top, with deeper structures at the bottom of the monitor. Upon completion, probe positioning can be mapped to evaluate three or four additional areas, typically between the anterior and the posterior axillary lines. Lateral views with the probe should be most posterior, typically along the posterior axillary line, tracking caudad toward the diaphragm. PLAPS (posterolateral alveolar and/or pleural syndrome) pointed posteriorly should also be evaluated, specifically in supine patients. PLAPS is lateral to the scapula and typically requires that the patient be lifted off the bed from one side.
2. Identify the "bat sign" (Fig. 27.2). The initial view observed should be a window of lung flanked by two rib shadows. This view, termed the "bat sign," should now allow evaluation of the parietal and visceral pleura, seen most superficial as echogenic line (approximately 0.5 cm below start of rib shadows), subsequent lung sliding, and other findings such as "A" and "B" lines, as well as abnormal lung tissue.
3. Identify "A lines" (Fig. 27.3). A lines indicate air. These are multiple echogenic lines appearing horizontally in sequence deep to the pleural line. This artifact represents reverberations of the pleura and can be found in aerated lungs, which can be normal or abnormal (e.g., pulmonary embolism, chronic obstructive pulmonary disease [COPD]). The first true A line, denoted "A1," is found equidistant from the chest wall to the pleural line. Many other A lines might be seen and are denoted "A'" lines. Subsequent equidistant A lines are "A2," "A3," and so on.
4. Identify "B lines" (Fig. 27.4). These artifacts appear in well-aerated lung and are vertical echogenic lines (ray, flashlight, lung rockets) transmitted from the pleura to the deeper parts of the lung on the ultrasound monitor field. They are due to thickened interlobular septa and represent alveolar fluid surrounded by air. True B lines arise from the pleural line and shoot all the way down to the far lung fields, whereas "comet tails" are seen only close to the pleural and are sometimes referred to as "shimmering" or "glimmering" during movement of the pleural line. When multiple B lines are seen in a patient, it is sometimes referred to as "lung rockets" or "flashlights" because many rays are shooting from the pleura. Even though most of the time B lines represent pulmonary edema, they can be seen in other conditions such as aspiration, pulmonary fibrosis, acute respiratory distress syndrome (ARDS), and pneumonia.
5. Identify lung sliding (Fig. 27.5). Lung sliding identifies movement of a normal parietal-visceral interface. Patient breathing causes a rhythmic movement of parietal against the visceral pleura appearing as movement of the hyperechoic line. M-mode can be used to show a timed clip of this through a still image and should only be used as a method of reporting or saving for documentation purposes. A "lung point" is an area adjacent to lung sliding (parietal pleura) that is devoid of movement. This is highly specific for pneumothorax. If one suspects a pneumothorax, quantification of the size can be made by evaluation at the anterior lung points and movement toward the lateral wall of the chest. The more rib spaces found to have absent lung sliding, the larger the pneumothorax. Lung sliding (Fig. 27.5) can be evaluated with M-mode, which can help identify a normal parietal-visceral interface at that level. Obtain an adequate two-dimensional view ("bat sign") and press the "M-mode" option on the equipment. A normal interface appears as multiple hyperechoic lines, the pleura (termed "seashore"; Fig. 27.5a), followed by a sandlike pattern, the lung tissue. This pattern together is termed a "seashore sign." Air that disrupts the parietal-visceral interface, as found within a pneumothorax, is identified as horizontal repeating echogenic lines, similar to a barcode (Fig. 27.5b), and is termed "stratosphere sign." M-mode can be discontinued.
   - Lung sliding can be absent in conditions other than pneumothorax: apnea, right or left bronchial intubation, lung collapse (blebs), pneumonia, and pulmonary fibrosis.
6. Identify lung pulse. This appears as a shimmering of the pleural line due to cardiac activity. This is most apparent on the left side of the chest, closest to the heart. This helps to exclude pneumothorax as well.
7. Move posteriorly. Move the ultrasound probe laterally and posteriorly to the PLAPS point. The transducer can be directed toward the center of the patient's body in supine patients. Pleural effusions and consolidations are found in the dependent areas of the lung.
8. Move caudally. With the marker still pointing cephalad, move along the posterior axillary line in two or three additional rib spaces. Identification of pleural disease and other pathology requires multiple views and will aid in evaluating the extent of the disease. This will also allow for identification of boundaries of the lung, such as the diaphragm. Identification of the diaphragm is most critical to determine location of fluid.

9. Identify the diaphragm and liver and/or spleen (Fig. 27.6). Along the posterior axillary line, or the posterior chest wall, move the probe caudad to identify the diaphragm. This appears as an echogenic curvilinear structure, with the liver or spleen being subdiaphragmatic and typically of different echogenicity than the lung. Many times the diaphragm is very high in the supine critically ill patient. Massive edema and obesity may also degrade image quality in this location.
   - Always identify the diaphragm. Hypoechoic fluid surrounding the liver or spleen can appear as a pleural effusion and must not be mistaken as such. In addition, lung tissue may mimic hepatic tissue in certain diseases such as dense consolidations termed "hepatization" of the lung. Proper probe positioning, clear identification of the diaphragm, subdiaphragmatic structures, and lung are crucial. This is a common error in novice operators owing to the confusion of the hepatorenal or splenorenal recess for the diaphragm.
   - Identifying the diaphragm can be technically difficult depending on patient position, size, and clinical condition. It maybe useful to start below the diaphragm, first identifying the hepatorenal recess (liver and kidney view interface) and then moving cephalad until the lung and diaphragm are visualized. In addition as ribs change their orientation anatomically, the probe may need to be adjusted while still in the longitudinal axis. Moving the probe clockwise and counterclockwise may be of benefit to bring into view the lung, the diaphragm, and the subdiaphragmatic structures.
10. Identify pleural effusions (Fig. 27.6). Confirming the presence of pleural effusions requires identifying anechoic material between the pleura and the lung. This can be seen as lung movement in an undulating pattern, which typically is facilitated by cardiac activity and respirations. This is termed "jellyfish sign," where the lung flaps as it freely floats in the effusion. Floating debris can also confirm effusion, termed "plankton sign." It is also important to identify the depth of the chest wall to the pleural fluid in order to determine the best location/depth of needle insertion when attempting thoracentesis or chest tube insertion. The challenge is to find a safe path for needle insertion. The key when using ultrasound as guidance is that the angle of the needle/syringe assembly must duplicate the angle of the probe. The time between scan and needle insertion must also be minimized. Real-time guidance is not required. In patients who are obese or edematous, skin indentation during probe placement can result in underestimation of depth of needle insertion required and must be taken into account. Safety margin of thoracentesis is thought to be 10 mm of visceral-parietal distance. Dry taps may be due to loculations the blocking needle, needle plugs, patient movement from scan to tap, and poor angle selection.
    - Exudates, empyemas, and hemothoraces may appear more echogenic, unlike, for example, a transudative effusion that could be anechoic. Complex effusions can also appear as heterogeneous and echogenic. The consistency of the effusion can make identification technically difficult because this can limit lung motion. Sometimes the operator may think there is no effusion when there is an echo-dense effusion.
11. Identify consolidations (Fig. 27.7). Compressed lung appears with alveolar consolidation pattern (tissue-like sign). Alveolar consolidations are devoid typically of air and appear as tissue density; these can be atelectasis, pneumonia, aspiration, or other diseased lung. "Hepatization" of the lung is typical, where the images mimic liver tissue. Images may also have hyperechoic foci representing air bronchograms, which would indicate pneumonia. Probe location should be correlated with an anatomical lobular or segmental area.
    - Lung may slide into the effusion during the respiratory cycle and can be problematic during needle insertion, causing pneumothorax or abnormal wire placement during the performance of pigtail chest tube catheters. This is called a "curtain sign."
12. Sinusoid sign (Fig. 27.8). M-mode is placed in the center of the visible lung when a large amount of pleural fluid is seen. A sinusoid sign strengthens the operator's determination that pleural fluid is present and that the pleural fluid is not necessarily compromising lung dynamics. If the sinusoid sign is absent, it may indicate a "trapped" lung dynamic.
13. Assessment and clinical decision making. Upon completion of ultrasonography of bilateral lung fields, clinical decision-making tools may be of benefit, especially in undifferentiated respiratory failure. A protocol has been developed to organize the exam of a respiratory failure patient (on noninvasive or invasive ventilation only). The BLUE protocol assesses patients based on findings (e.g., A lines, B lines, lung sliding) of both lungs and incorporates them into an algorithm. With acceptable sensitivities and specificities, practitioners can diagnose pulmonary edema, pneumonia, pneumothorax, and COPD/asthma with the BLUE protocol.

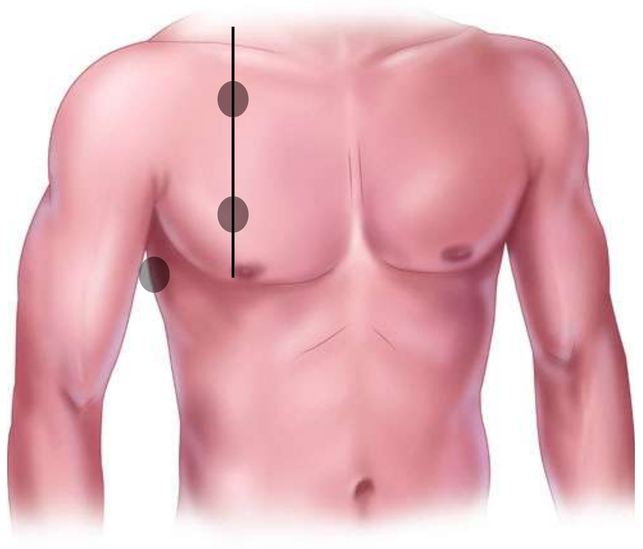

Fig. 27.1 *Black circles* indicate possible ultrasound view points. These are just guides, and one may view a few interspaces left/right and cephalad/caudad from each point to obtain better views

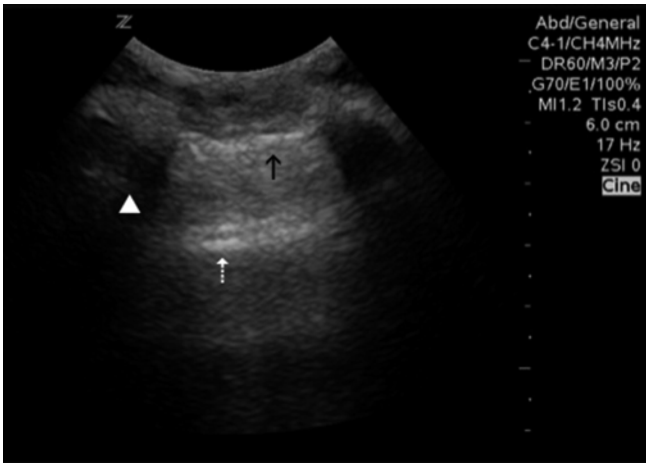
Abd/General
C4-1/CH4MHz
DR60/M3/P2
G70/E1/100%
MI1.2 TIs0.4
6.0 cm
17 Hz
ZSI 0
Cine

Fig. 27.2 Typical bat sign with rib shadows (*arrowhead*) that surround the pleura (*black arrow*) and lung tissue. Incidentally, A lines (*white arrow*) are seen

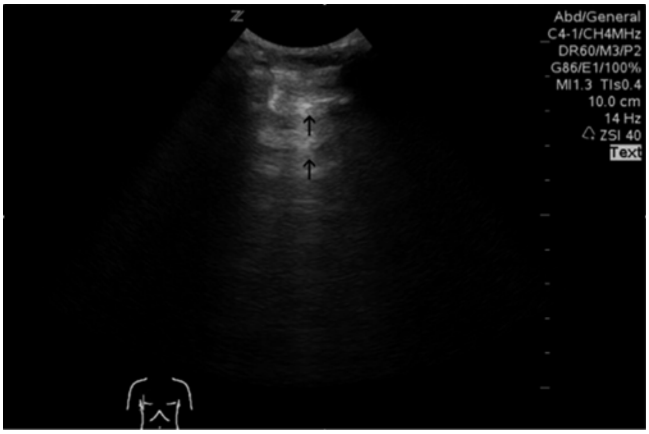
Abd/General
C4-1/CH4MHz
DR60/M3/P2
G86/E1/100%
MI1.3 TIs0.4
10.0 cm
14 Hz
ZSI 40
Text

Fig. 27.3 Rib shadows can be seen, with multiple A lines (*arrows*)

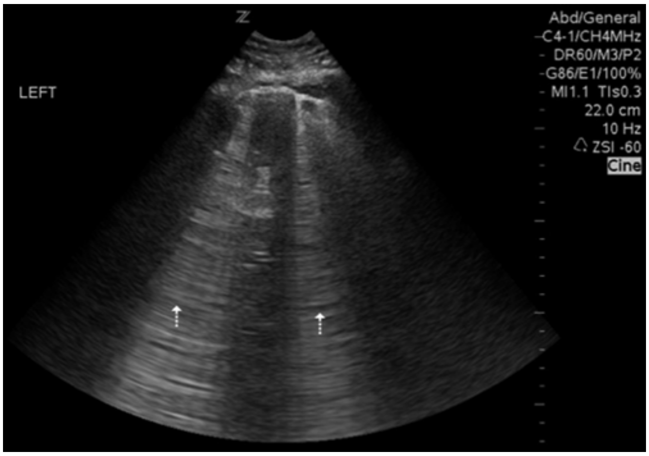
LEFT
Abd/General
C4-1/CH4MHz
DR60/M3/P2
G86/E1/100%
MI1.1 TIs0.3
22.0 cm
10 Hz
ZSI -60
Cine

Fig. 27.4 Two B lines can be appreciated (*arrows*)

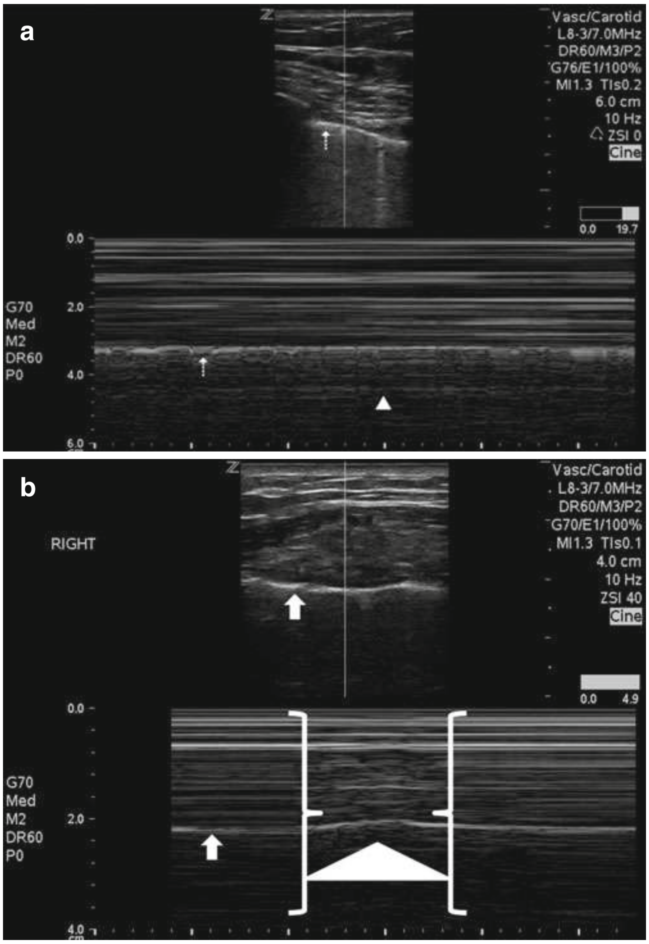
a
Vasc/Carotid
L8-3/7.0MHz
DR60/M3/P2
G76/E1/100%
MI1.3 TIs0.2
6.0 cm
10 Hz
ZSI 0
Cine
0.0
19.7
0.0
2.0
4.0
6.0
G70
Med
M2
DR60
P0
b
RIGHT
Vasc/Carotid
L8-3/7.0MHz
DR60/M3/P2
G70/E1/100%
MI1.3 TIs0.1
4.0 cm
10 Hz
ZSI 40
Cine
0.0
4.9
0.0
2.0
4.0
G70
Med
M2
DR60
P0

Fig. 27.5 (**a**) M-mode identifies the seashore sign, with the pleura (*arrows*) and the lung parenchyma (*arrowhead*) together producing an image similar to a seashore. (**b**) M-mode shows the stratosphere or barcode sign. *White arrows* indicate the pleura, and the *large triangle* shows an area of artifact from movement. Artifact from movement can be distinguished by a similar pattern change above and below the pleural line. On either side of the bracket is a continued horizontal line pattern typical of absent lung sliding

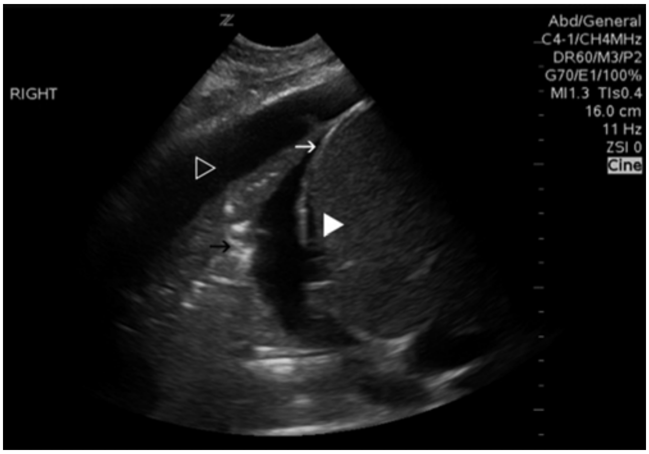
RIGHT
Abd/General
C4-1/CH4MHz
DR60/M3/P2
G70/E1/100%
MI1.3 TIs0.4
16.0 cm
11 Hz
ZSI 0
Cine

Fig. 27.6 Diaphragm (*white arrow*), liver (*white arrowhead*), lung (*black arrow*), and anechoic fluid, likely an effusion (*open arrowhead*). Incidentally, hyperechoic foci in the lung tissue, likely representing air bronchograms, can be seen

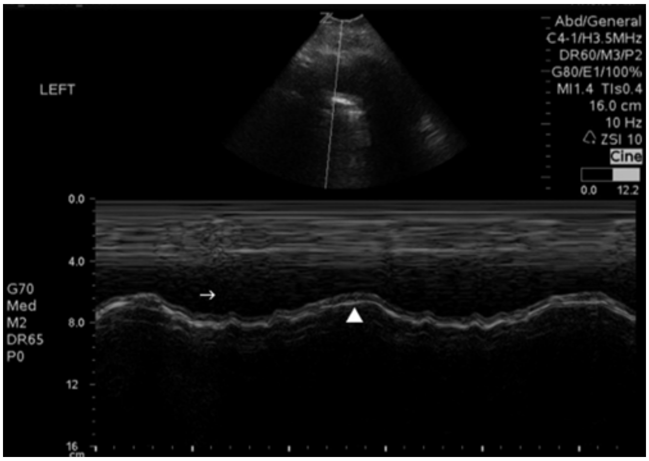
LEFT
Abd/General
C4-1/H3.5MHz
DR60/M3/P2
G80/E1/100%
MI1.4 TIs0.4
16.0 cm
10 Hz
ZSI 10
Cine
0.0
12.2
0.0
4.0
8.0
12
16
G70
Med
M2
DR65
P0

Fig. 27.8 In M-mode, identification of pleural effusion becomes evident with the sinusoidal sign (*arrowhead*) and identification of pleural fluid (*arrow*)

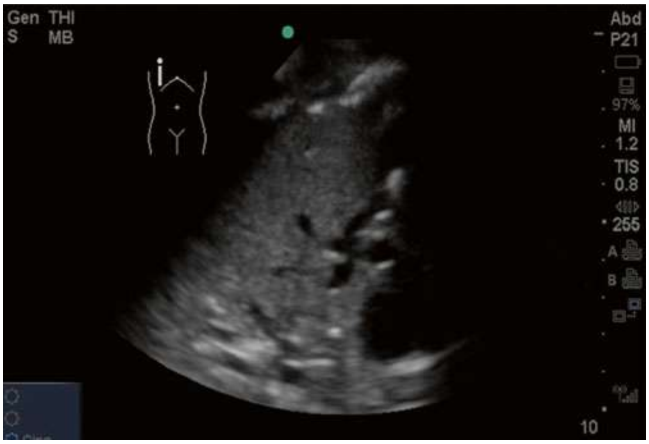
Gen THI
S MB
Abd
P21
97%
MI
1.2
TIS
0.8
255
A
B
10

Fig. 27.7 Lung here appears as tissue density, likely representing the alveolar consolidation pattern seen in pneumonia, and is termed "hepatization" and "tissue-like sign"

Selected Reading

Levitov A, Mayo PH, Slonim AD. Ultrasound evaluation of the lung. In: Critical care ultrasonography. New York: McGraw-Hill; 2009.

Lichtenstein DA, Menu Y. A bedside ultrasound sign ruling out pneumothorax in the critically ill. Lung sliding. Chest. 1995;108:1345–8.

Lichtenstein DA, Meziere GA. A lung ultrasound sign allowing bedside distinction between pulmonary edema and COPD: the comet-tail artifact. Intensive Care Med. 1998;24:1331–4.

Lichtenstein DA, Meziere GA. Relevance of the lung ultrasound in the diagnosis of acute respiratory failure: the BLUE protocol. Chest. 2008;134:117–25.

Mayo PH, Doelken P. Pleural ultrasonography. Clin Chest Med. 2006;27:215–27.

Mayo PH, Goltz HR, Tafreshi M, Doelken P. Safety of ultrasound-guided thoracentesis in patients receiving mechanical ventilation. Chest. 2004;125:1059–62.

Part IV

Cardiac Procedures

Repair of Cardiac Injuries

28

Ronald Tesoriero

Whether performed by an emergency physician or a surgeon, the majority of cardiac repairs performed at emergency department thoracotomy will be temporary in nature and require further revision in the operating room. This chapter's main focus is the temporary control of cardiac injuries.

28.1 Indications

- Wounds to the heart in patients presenting with pulseless electrical activity (PEA) or asystole with evidence of cardiac tamponade
  - Penetrating wounds: <15 min of prehospital cardiopulmonary resuscitation (CPR)
  - Blunt wounds: <10 min of prehospital cardiopulmonary resuscitation (CPR)

28.2 Contraindications

- Absolute
  - Presenting rhythm of asystole and no evidence of pericardial tamponade on Focused Assessment with Sonography for Trauma (FAST)
- Relative
  - None

28.3 Materials and Medications

- Diagnostic ultrasound
- Emergency department thoracotomy tray (Fig. 28.1)
  - Sterile drapes, #10 scalpel, curved Mayo scissors, Finochietto retractor, Lebsche sternal knife and mallet, forceps, curved Metzenbaum scissors, surgical skin stapler, Foley catheter, clamps, needle driver, 2–0 and 3–0 polypropylene suture on MH or SH (noncutting) needle, Satinsky vascular clamps
- Betadine (povidone-iodine) or other skin antiseptic preparing solution
- Defibrillator and internal cardiac panels
- Epinephrine

R. Tesoriero, MD
Department of Surgical Critical Care,
University of Maryland School of Medicine,
R Adams Cowley Shock Trauma Center,
Baltimore, MD, USA
e-mail: rtesoriero@umm.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_28

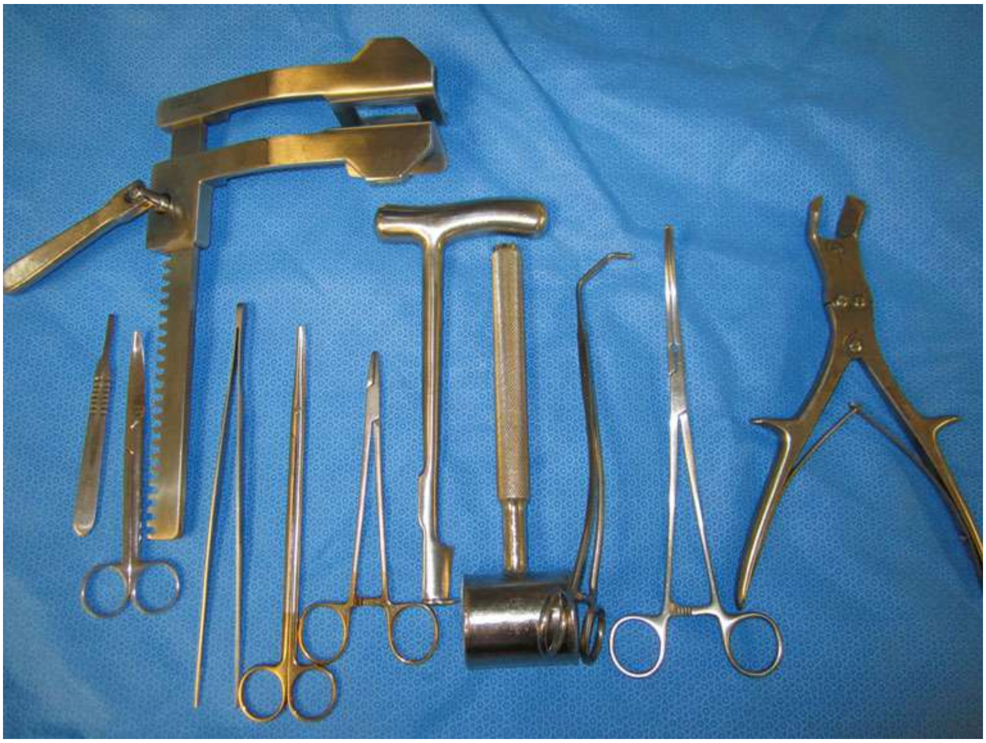

Fig. 28.1 Emergency department thoracotomy tray. *Left to right*: Scalpel, curved Mayo scissors, Finochietto rib spreader, DeBakey forceps, Metzenbaum scissors, needle driver, Lebsche knife and mallet, Satinsky vascular clamp, aortic clamp, bone cutter

28.4 Procedure

1. Identify the pericardium anterior to the phrenic nerve, pinch it between the fingers, and enter it by making a nick with the Metzenbaum scissors. Then open the pericardium longitudinally anterior to the phrenic nerve.
2. If the heart is not contracting effectively (PEA, asystole, fibrillation), immediately begin internal cardiac massage.
   - Unless the injury is so large that it cannot be controlled, taking time out at this point to repair the injury may lead to significant acidosis and inability to reestablish a perfusing rhythm.
3. While continuing internal cardiac massage, identify the area of injury and attempt to control it with manual pressure; if a larger wound, use a Foley catheter to prevent continued bleeding.
4. If necessary owing to the location of injury, extend the incision to a bilateral anterolateral thoracotomy and transect the sternum with a Lebsche knife and mallet or heavy shears.
5. If, after several minutes of internal massage and appropriate red blood cell and plasma transfusion, the patient remains in PEA or asystole, irrigate the heart with warmed saline and administer intravenous (IV) or intracardiac epinephrine.
6. If the heart enters ventricular fibrillation or ventricular tachycardia, cardiovert with internal paddles applied directly to the heart with an energy between 10 and 30 J.
7. Once a perfusing rhythm is reestablished, or if on initial evaluation the wound is so large that it cannot be controlled and will require immediate repair before internal massage can be effective, proceed to cardiac repair.
8. Choose the simplest method that will allow control of the injury until definitive repair can be performed in the operating room.
   (a) Injuries to the atrium: The pliable nature of the atria will often allow placement of a Satinsky clamp for control followed by repair with a running 3–0 polypropylene suture (Fig. 28.2).
   (b) Small injuries to the ventricles: Control the injury with direct manual pressure and close with either interrupted 3–0 polypropylene suture or a surgical skin stapler (Figs. 28.3 and 28.4).
   (c) Medium to large injuries to the ventricle: Attempt to control the wound by placing a Foley catheter through it, then blow up the balloon and apply gentle traction. Either staple or suture the wound closed before deflating the balloon and removing the catheter (Fig. 28.5).
      - Place an occluding clamp on the open end of the catheter or blood loss will continue through the catheter
      - Avoid excessive traction on the Foley because it will pull through the ventricle and make the hole larger
   (d) Extensive or inaccessible injury: Perform temporary inflow occlusion to the heart by manually compressing the right atrium against the heart so that it cannot fill. The heart will likely immediately enter PEA, fibrillation, or asystole giving the physician a couple of minutes to gain control of the injury before the patient becomes unrecoverable.
   (e) Injuries in proximity to coronary vessels: To avoid compression of the vessel, perform a horizontal mattress suture that passes beneath the artery. Teflon pledgets may assist in the repair (Fig. 28.6).

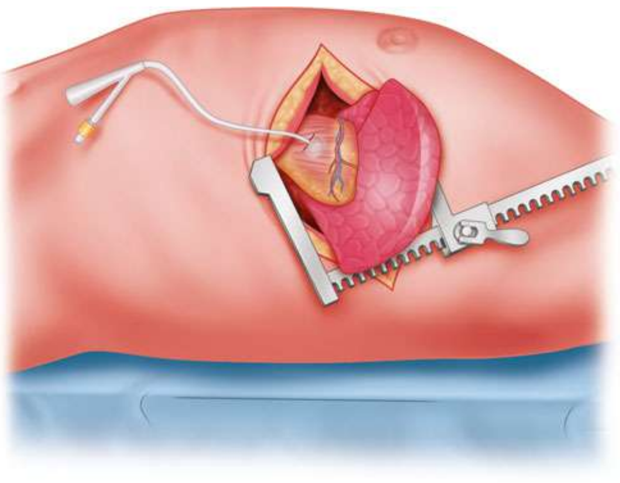

Fig. 28.2 Position the Finochietto retractor with the closed end toward the axilla. A Foley catheter may be used for initial control of large cardiac lacerations

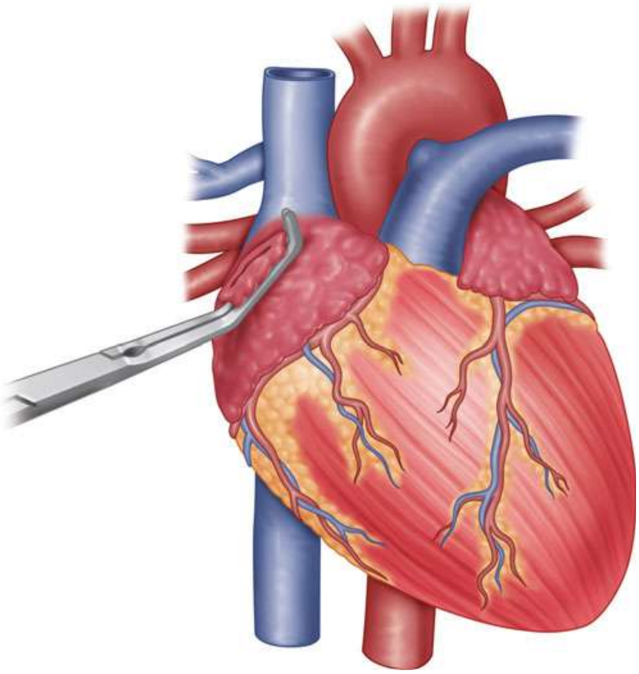

Fig. 28.3 Atrial injuries may be quickly controlled with a Satinsky clamp followed by repair with a 3–0 polypropylene suture

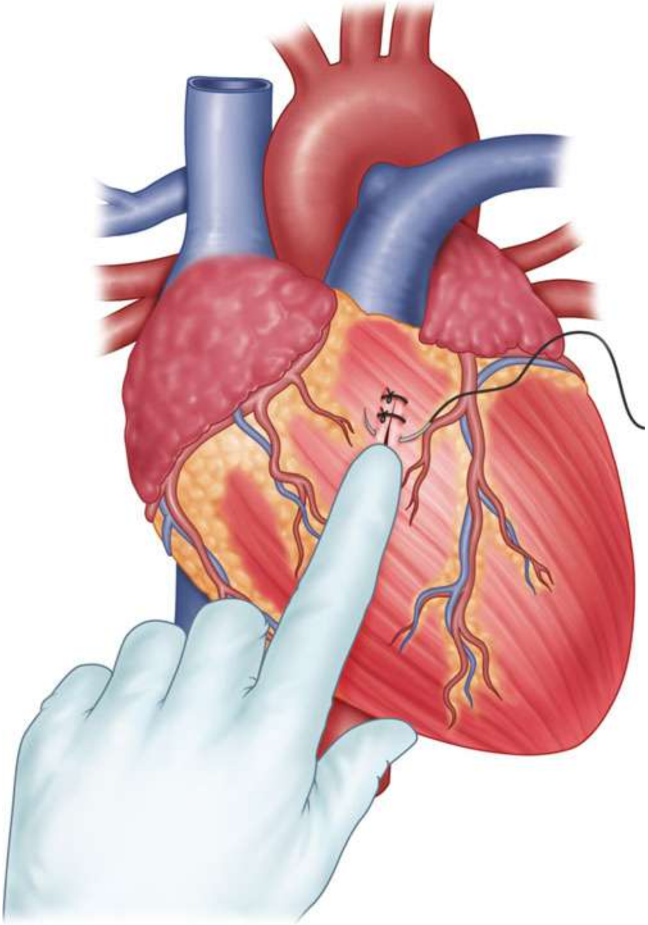

Fig. 28.4 Injuries may be controlled with direct manual pressure and closed with interrupted 3–0 polypropylene suture

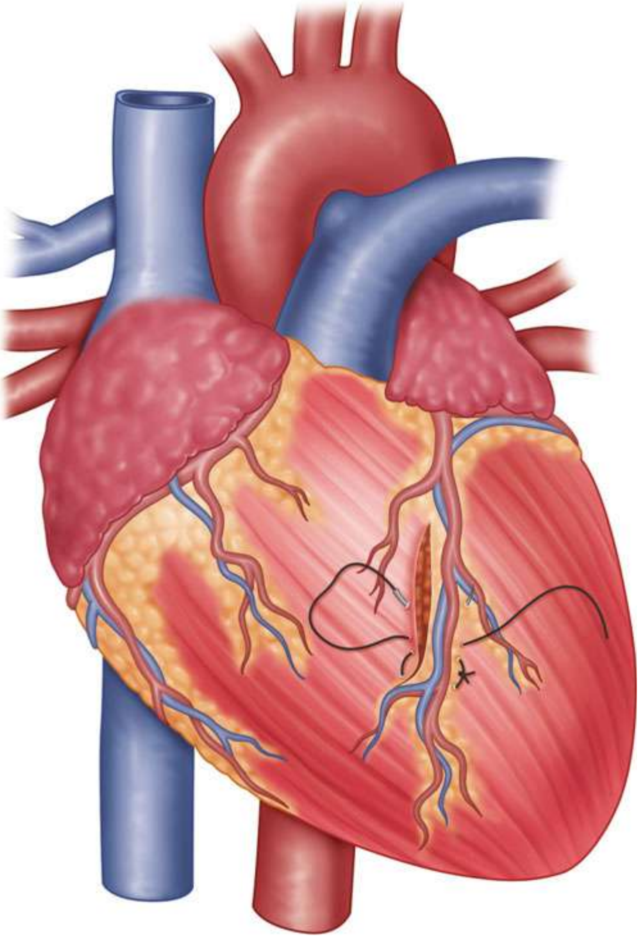

Fig. 28.6 Lacerations in close proximity to coronary vessels can be controlled with horizontal mattress sutures passed beneath the artery to avoid coronary arterial compression

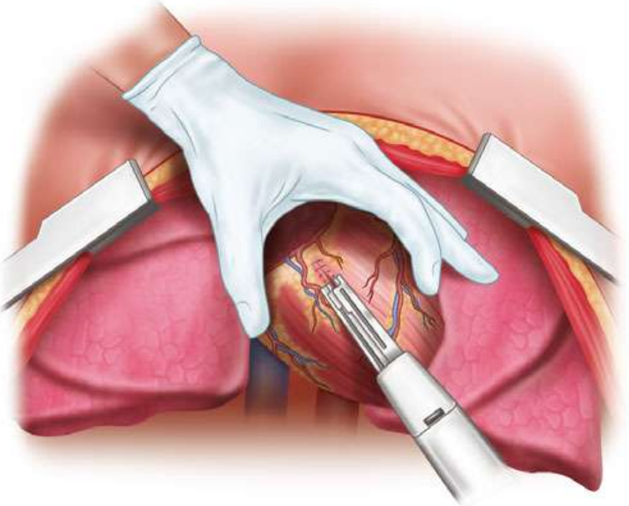

Fig. 28.5 A surgical skin stapler is a quick way to gain initial control of simple cardiac lacerations with minimal risk to the provider

28.5 Complications

- Cardiac and/or pulmonary laceration during entry into the chest
- Bleeding
- Delayed hemorrhage (failure to control internal mammary artery, disruption of sutures or staples)
- Infections: empyema and sternal infection
- Missed intracardiac valvular or septal injury (echocardiography should be performed after repair is complete)
- Air embolus

28.6 Pearls and Pitfalls

- Pearls
  - There may be more than one wound to the heart (especially with gunshot wounds). Look for them. However, if the wound is posterior and not bleeding with the heart in its natural position, it will be more prudent to leave the injury alone until the patient can be transported to the operating room. Elevating the heart can cause both inflow and outflow obstruction, leading to dysrhythmia that may be difficult to recover from.
  - To avoid sutures pulling through in patients with thin, edematous, or friable myocardium, consider horizontal mattress rather than simple sutures. These may be buttressed with Teflon pledgets for added security.
- Pitfalls
  - The myocardium tears easily. When tying sutures, take care to not tighten them too forcefully.

Selected Reading

Asensio JA, Trunkey DD. Current therapy of trauma and surgical critical care. Philadelphia: Mosby/Elsevier; 2008.

Feliciano DV, Mattox KL, Moore EE. Trauma. 6th ed. New York: McGraw-Hill Medical; 2008.

Hirshberg A, Mattox KL. Top knife: the art & craft of trauma surgery. Castle Hill Barns: TFM; 2005.

Moore EE, Knudson MM, Clay CB, et al. Defining limits of resuscitative emergency department thoracotomy: a contemporary Western Trauma Association perspective. J Trauma. 2011;70:334–9.

Wall MJ, Mattox KL, Chen CD, et al. Acute management of complex cardiac injuries. J Trauma. 1997;42:905–12.

Synchronized Electrical Cardioversion

29

Jason Jones, Ann Tsung, and Marie-Carmelle Elie

Tachycardia is defined as >100 beats per minute

29.1 Indications

- Non-sinus-rhythm tachycardias with a pulse including:
  - Atrial fibrillation
  - Atrial flutter
  - Monomorphic ventricular tachycardia (VT)
  - Refractory or unstable supraventricular tachycardia (SVT)
- Unstable signs and symptoms including acute coronary syndrome, decreased level of consciousness, chest pain, dyspnea, pulmonary edema, and hypotension

29.2 Contraindications

- Absolute
  - Ventricular fibrillation and pulseless or polymorphic (irregular) VT require unsynchronized electrical cardioversion (defibrillation), not synchronized cardioversion.
  - Known atrial thrombus.
  - Sinus tachycardia.
- Relative
  - Digitalis toxicity-related tachycardia
  - Atrial fibrillation of greater than 48 h duration without anticoagulation
  - Multifocal atrial tachycardia
  - Electrolyte abnormalities
  - Left atrial diameter greater than 4.5 cm
  - Patients with low probability of maintaining sinus rhythm and readily return to atrial fibrillation
  - Patients with sick sinus syndrome or sinoatrial blockage who will require a pacemaker for maintenance of stable rhythm

29.3 Materials and Medications

- Airway management equipment (laryngoscopes, endotracheal tubes)
- Cardiac monitoring, pulse oximetry, end-tidal CO_2 monitoring
- Cardioverter/defibrillator
- Sedation and analgesic medications

29.4 Procedure

1. Obtain a 12-lead electrocardiogram (ECG) and intravenous (IV) access.
2. If possible, correct underlying electrolyte abnormalities that may cause or contribute to the patient's arrhythmia.
3. Discuss risks, benefits, and alternatives (including pharmacological cardioversion) with the patient and obtain consent.
4. Prepare airway equipment and Advanced Cardiac Life Support (ACLS) code drugs.
5. Consider IV sedation (e.g., propofol, midazolam).
6. Provide IV analgesia (e.g., fentanyl, morphine).
7. Place defibrillator adhesive pads (8- to 12-cm diameter in adults) or paddles on the patient. Pediatric-sized pads/paddles should be used if the patient is less than 10 kg.

J. Jones, MD • M.-C. Elie, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: jasonjones@ufl.edu; elie@ufl.edu

A. Tsung, MD
Department of Emergency Medicine, University of Florida Health, Gainesville, FL, USA
e-mail: TSUNGANN@UFL.EDU

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_29

8. The first paddle/pad is placed to the right of the sternum at the second/third intercostal space. The second paddle/pad can be placed in one of two equally efficacious positions:
   (a) Anterolateral position—left fourth/fifth intercostal space in the midaxillary line (Fig. 29.1)
   (b) Anteroposterior position—between the spine and the edge of the left scapula (Fig. 29.2)
9. Turn the defibrillator/cardioverter into synchronized mode—marker above R-waves will be present (Fig. 29.3).
10. Select the energy level to be delivered based on the underlying rhythm
    (a) Regular VT (with pulses)—Adults: 100 J (monophasic or biphasic), 200 J for subsequent shocks
    (b) Atrial fibrillation—120–200 J (biphasic), 200 J (monophasic), 360 J for subsequent shocks
    (c) Atrial flutter and paroxysmal SVT—50–100 J (biphasic), 100 J for subsequent shocks
    (d) Pediatric dosage (regular and pulsed VT or SVT)—0.5–1 J/kg, up to 2 J/kg for subsequent shocks (Fig. 29.4)
11. Announce that you are going to deliver the shock on the count of three, and ensure that everyone is clear of the patient.
12. Deliver the shock by pressing button marked “SHOCK.”
    - If using paddles, apply firm pressure and keep paddles in place until shock is delivered.
13. Reassess the patient’s pulse and cardiac rhythm.
14. Repeat with escalating energy in a stepwise fashion if cardioversion is unsuccessful.

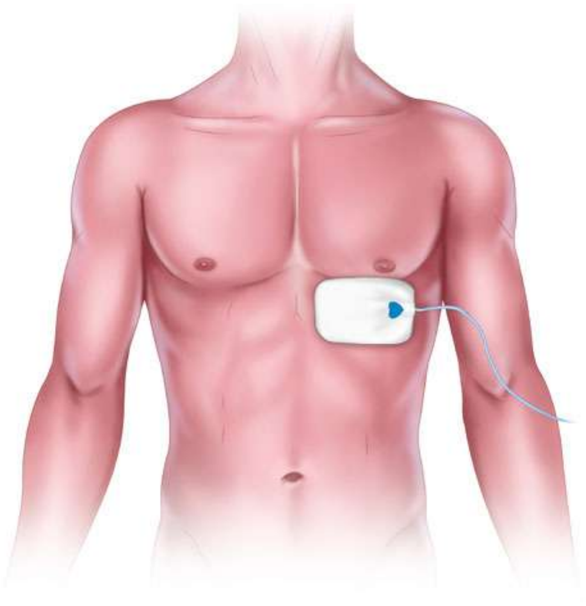

Fig. 29.1 Anterolateral pad placement

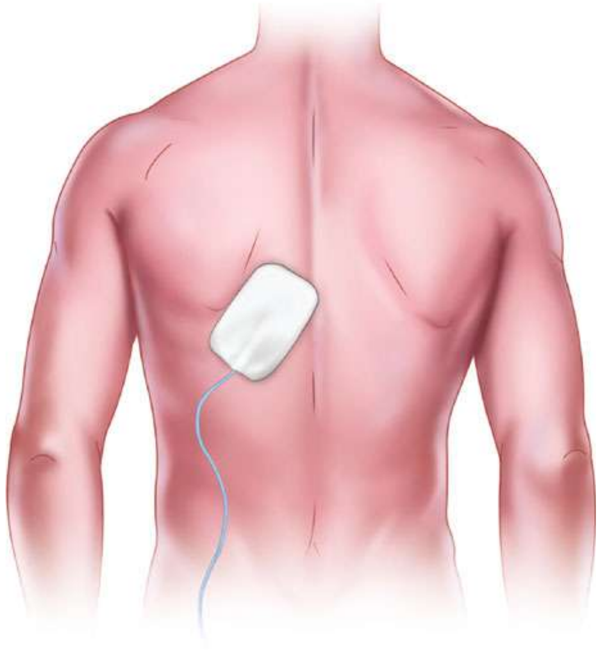

Fig. 29.2 Anteroposterior pad placement

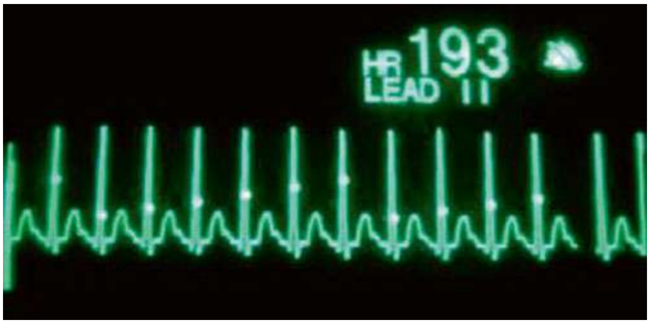

Fig. 29.3 Synchronized cardioversion—mark on R wave

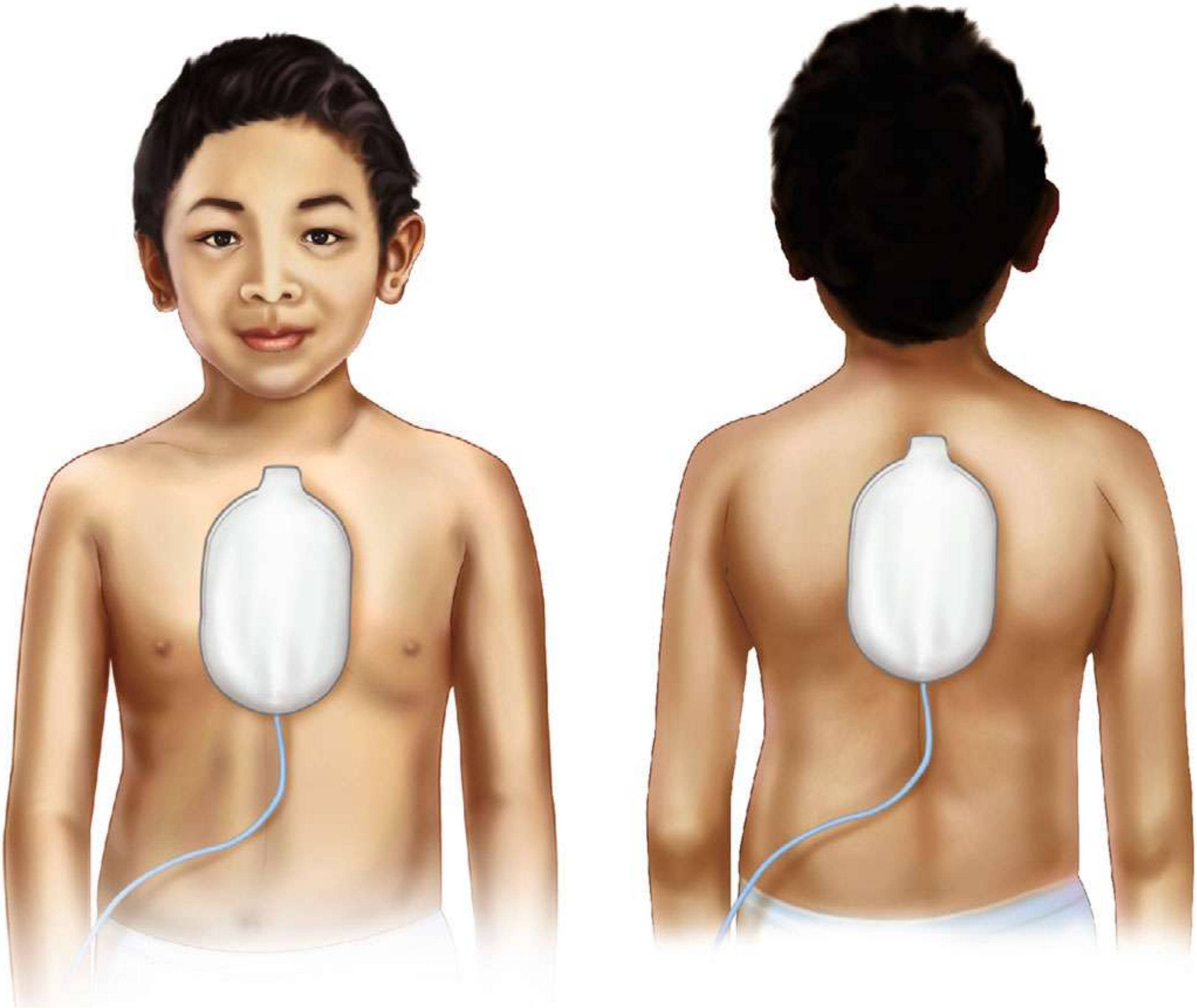

Fig. 29.4 Pediatric pad placement

29.5 Complications

- Superficial burns if there is inadequate gel.
- Induced arrhythmias (bradycardia in patients with previous inferior myocardial infarction, atrioventricular block, VT, ventricular fibrillation, asystole).
- Improperly synched cardioversion may rarely induce ventricular fibrillation.
  - Ectopy of the atria or ventricle in first 30 min after cardioversion
  - Atrial clot embolization in patients without adequate anticoagulation
- Apnea, hypoxia, hypercarbia, or hypotension may occur from sedation/analgesia.
- Medical professionals who incidentally touch the patient during shock delivery may be shocked or burned.
- Rarely, fire has occurred as a consequence of poor pad placement and a hyperoxygenated environment.

29.6 Pearls and Pitfalls

- Hirsute males should be shaved at the pad/paddle placement sites.
- Placing the cardioverter in "synchronized" mode avoids delivering a shock during the relative refractory segment, which could induce ventricular fibrillation.
- Cardioversion of pregnant patients is performed as in the general adult population.
- Keep pacemakers and implantable cardioverter-defibrillators at least 10 cm away from contact with paddles.
- If patient has implanted pacemaker, position pads so that they are not directly over the device.

Selected Reading

Gowda SA, Shah A, Steinberg JS. Cardioversion of atrial fibrillation. Progr Cardiovasc Dis. 2005;48:88–107.

Kleinman ME, Chameides L, Schexnayder SM, et al. Part 14: pediatric advanced life support. 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010;122 Suppl 3:S876–908.

Link MS, Atkins DL, Passman RS, et al. Part 6: electrical therapies: automated external defibrillators, defibrillation, cardioversion, and pacing. 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010;122 Suppl 3:S706–19.

Mayeaux EJ. The essentials guide to primary care procedures. Philadelphia: Lippincott Williams & Wilkins; 2012. p. 88–92.

Sirna SJ, Ferguson DW, Charbonnier F, et al. Factors affecting transthoracic impedance during electrical cardioversion. Am J Cardiol. 1988;62:1048–52.

Trohman RG, Parrillo JE. Direct current cardioversion: indications, techniques, and recent advances. Crit Care Med. 2000;28:170–3.

Unsynchronized Cardioversion (Defibrillation)

30

Matthew R. Tice, Zachary B. Kramer,
and Marie-Carmelle Elie

Unsynchronized cardioversion or defibrillation is the delivery of a *high-energy* shock as soon as the button is pushed on defibrillator. This means it can be delivered anywhere in the cardiac cycle. By contrast, synchronized cardioversion (see Chap. 29) delivers a low-energy shock at the peak of the R wave in the cardiac (QRS) cycle.

30.1 Indications

- Ventricular fibrillation (VF)
- Pulseless ventricular tachycardia (VT)
- Cardiac arrest due to or resulting from VF

30.2 Contraindications

- Absolute
  - Conscious patient
  - Presence of a pulse
  - Pulseless electrical activity (PEA)
  - Asystole
  - Multifocal atrial tachycardia
  - Defibrillation without knowing the rhythm
  - A second defibrillation before 2 min (or five cycles) of CPR
  - Advanced Directive, Physician Order for Life-Sustaining Treatment (POLST) indicating no cardiopulmonary resuscitation (CPR) or do not resuscitate (DNR)
- Relative
  - Potential electrical catastrophe (explosive environment [i.e., operating rooms])
  - Dysrhythmias due to enhanced automaticity such as in digitalis toxicity and catecholamine-induced arrhythmia (because mechanism of tachycardia remains after the shock)
- Factors that are not contraindications
  - Pregnancy.
  - Chest trauma.
  - Automatic implantable cardiac defibrillators (AICDs).
  - The patient is on a wet or moist surface.
  - Piercings on the chest.

30.3 Materials and Medications

- Electrocardiogram (ECG) monitor/defibrillator.
- Self-adhesive defibrillation pads or defibrillation paddles (paddles may be more successful than self-adhesive pads, but they have more complications and pose more danger to operators).
- Conductive gel for defibrillation paddles (not ultra sound gel).
- ECG electrodes.
- Supplemental oxygen.
- Intubation equipment as needed.

M.R. Tice, MD • Z.B. Kramer, MD • M.-C. Elie, MD, RDMS (✉)
Department of Emergency Medicine,
University of Florida, Gainesville, FL, USA
e-mail: mtice@ufl.edu; kramer@ufl.edu; elie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_30

30.4 Procedure

Defibrillation is an emergent maneuver and, when necessary, should be promptly performed in conjunction with or before administration of induction or sedative agents to facilitate intubation.

1. Assess the ABCs (airway, breathing, circulation).
   - Open the airway with a head tilt/chin lift (or jaw thrust in a suspected traumatic patient). If the patient is apneic, provide breaths with a bag-valve-mask (BVM) and observe chest rise.
   - Check for pulses. If absent, start CPR.
   - CPR should be initiated before any shock while getting all equipment ready for at least 2 min to provide adequate circulation to the brain and heart.
2. Wipe off the patient's chest if moist or wet.
3. Remove transdermal patches, jewelry, and piercings if possible.
4. Attach ECG electrodes to the patient.
   - The self-adhesive defibrillation pads or defibrillation paddles can be used as ECG electrodes to access the rhythm.
5. Paddles: With conductive gel applied to the metal surface, place one paddle on the patient's right chest, just below the clavicle, near the sternal border. The other should be on the left chest, midaxillary line above the fifth or sixth intercostal space (Fig. 30.1).
   - The long axis of the paddles should be perpendicular to the ribs to allow for better transduction of current through the chest.
   - Pads: Same placement as paddles except that pads can be placed in any orientation as long as they are in full contact with the chest.
   - If a lot of breast tissue is present, push the tissue to one side or lift it away and place the paddles or pads underneath.
   - An error in pad or paddle placement can distort the rhythm into looking like a rhythm that does not require defibrillation.
6. Place the ECG monitor/defibrillator into a mode to acquire a rhythm from the pads or paddles.
   - Steps 1–3 can be done simultaneously.
7. Stop CPR and assess the rhythm and pulse for no longer than 10 s.
8. If VF or pulseless VT is observed, then switch the defibrillator to charge mode.
9. Charge to 200 J.
   - Continue CPR while the defibrillator is being charged.
10. When the defibrillator indicates that it is charged, clearly order that everyone stop touching the patient and observe that there is no physical contact before defibrillating the patient.
    - If using the paddles, apply extra force to the chest through the paddles to deflate the lungs to allow for better defibrillation.
    - The operator should observe a muscle twitch during defibrillation.
11. Restart CPR for 2 min or five cycles.
    - Another operator may charge (but not fire) the defibrillator while CPR is being performed to expedite the time between pulse/rhythm check and the initiation of a shock.
12. After 2 min or five cycles of CPR, assess the rhythm and pulse and repeat steps 6–8 and give appropriate Advanced Cardiac Life Support (ACLS) medications.
13. If successful return of spontaneous circulation (ROSC) occurs, initiate the hypothermia protocol per hospital guidelines.

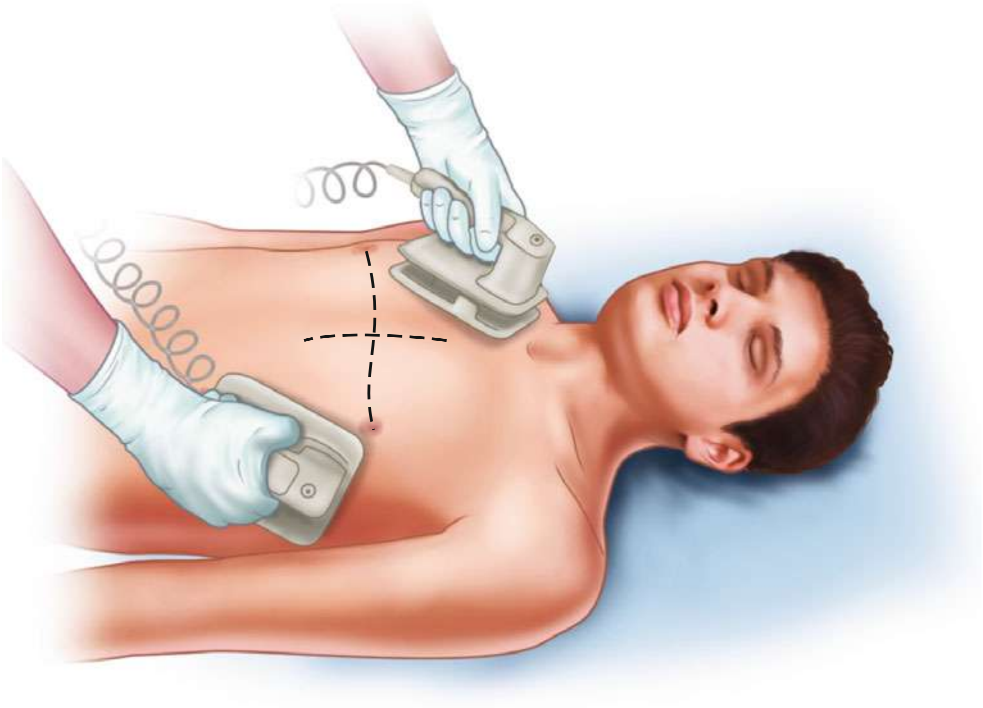

Fig. 30.1 Proper placement of defibrillation paddles

30.5 Complications

- Skin burns (most common and likely due to improper technique).
- Injury to cardiac tissue (myocardial necrosis secondary to burn).
  - ST segment elevation that lasts longer than 2 min usually indicates myocardial injury unrelated to the shock.
- Abnormal heart rhythms (usually benign like atrial, ventricular, and junctional premature beats).

30.6 Pearls and Pitfalls

- Pearls
  - Post-defibrillation cardiac dysrhythmias are more common following prolonged VF and higher-energy-level countershocks. Early defibrillation at the recommended energy levels minimizes this complication. Follow ACLS protocols to manage the resulting dysrhythmias.
  - The presence of liquid (body fluids, medications, or intravenous fluids) may cause electrical arcing thermal burns to the skin and soft tissue and produce ineffective defibrillation by allowing the current to pass across the trunk rather than transthoracically. To minimize this potential complication, ensure that any body fluids or liquids are wiped away from the skin before defibrillation attempts.
- Pitfalls
  - Myocardial and epicardial injury (not direct thermal injury) may result from the electrical current applied in defibrillation. Use the minimal recommended energy levels to minimize injury.
  - Electrical injuries to health-care providers can result if participants remain in contact with the patient during delivery of a countershock because they can serve as a ground for the current discharged.
    - This can be minimized by wearing gloves and using biphasic machines and electrode pads.
  - Fires that can result from sparks in the presence of nitroglycerin patches or ointment, flammable gases, or an oxygen-rich environment can also be a source of injury to the patient or health-care providers. Ensure “all clear” from the patient before delivery of shock to avoid these complications.

Selected Reading

Advance cardiovascular life support: provider manual. Dallas: American Heart Association; 2011.

Deakin CD, Sado DM, Petley GW, Clewlow F. Is the orientation of the apical defibrillation paddle of importance during manual external defibrillation? Resuscitation. 2003;56:15–8.

Pagan-Carlo LA, Spencer KT, Robertson CE, Dengler A, Birkitt C, Kerber RE. Transthoracic defibrillation: importance of avoiding electrode placement directly on the female breast. J Am Coll Cardiol. 1996;27:449–52.

Walker C, Callies F, Langenfeld H. Adverse effects of direct current cardioversion on cardiac pacemakers and electrodes: is external cardioversion contraindicated in patients with permanent pacing systems? Eur Soc Cardiol. 2004;6:165–8.

Transcutaneous Pacing

31

Nour Rifai and Christian Coletti

31.1 Indications

- Hemodynamically unstable (i.e., hypotension, pulmonary edema, chest pain, shortness of breath, or evidence of decreased cerebral perfusion) bradyarrhythmias refractory to medical therapies
- As a bridge to a transvenous or permanent pacemaker
- As an overdrive pacer in tachyarrhythmias
- Controversially, within the first 10 min of a witnessed asystolic cardiac arrest
- In children only with bradycardia associated with a known congenital cardiac defect or after cardiac surgery

31.2 Contraindications

- Absolute
  - None
- Relative
  - Bradyarrhythmia associated with hypothermia (ventricles are more prone to defibrillation-resistant fibrillation)
  - Prolonged cardiac arrest (>20 min)
  - Bradyarrhythmia in children (usually secondary to hypoxia or a respiratory issue)
  - Patient is unable to tolerate the procedure despite sedation and analgesia

31.3 Materials and Medications

- Pacemaker device (modern units offer combined pacer and defibrillator functions) (Fig. 31.1)
- One set of standard electrocardiogram (ECG) electrodes
- One set of pacer pads
- Code cart and airway equipment (prophylactically)
- Sedation and analgesia (typically a short-acting benzodiazepine and an opioid)
  - Midazolam: 0.2–0.10 mg/kg intravenous (IV) push and may repeat with 25 % of initial dose after 3–5 min. Do not exceed 2.5 mg/dose or a cumulative dose of 5 mg.
  - Fentanyl: 1–2 mcg/kg IV slow push over 1–2 min, may repeat dose in 30 min. (Fentanyl is the opioid of choice because it is less likely to exacerbate any hypotension.)

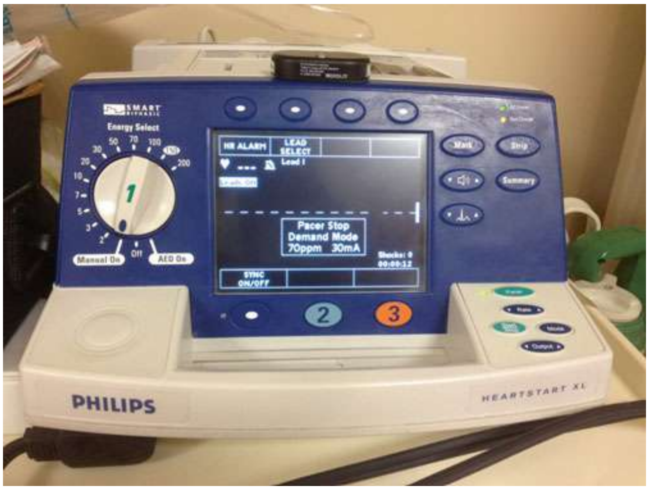

Fig. 31.1 Pacemaker device

N. Rifai, MBCHB • C. Coletti, MD (✉)
Department of Emergency Medicine,
Christiana Care Health System,
Newark, DE, USA
e-mail: nourifai@gmail.com; chris.coletti@gmail.com

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_31

31.4 Procedure

1. Time permitting, clean and dry the skin, and shave any excess hair off the chest.
2. Administer any appropriate sedation and analgesia.
3. Attach the ECG electrodes to both the input port of the pacemaker unit and the patient. On the patient, the white lead is placed just above the right clavicle, the black lead is just above the left clavicle, and the red lead is around the left midaxillary line.
4. Attach the pacer pads either in the anteroposterior or anterolateral positions as pictured (avoid placement over an implanted pacemaker or defibrillator; Fig. 31.2).
5. Turn the machine on and switch it to synchronous (or on-demand) mode.
    - Asynchronous (or fixed) mode fires impulses with no regard to the intrinsic cardiac cycle, increasing the likelihood of an R on T phenomenon, which could result in ventricular tachycardia or fibrillation.
    - Synchronous (or on-demand) mode will not fire an electrical impulse when a QRS complex is sensed; this is the preferred mode for transcutaneous pacing.
6. Set the desired heart rate: typically 60–80 beats/min to achieve adequate perfusion.
7. Select a lead on the pacemaker unit and then press Start.
8. Slowly increase the output current until electrical capture is denoted by a visible pacemaker spike, which will precede every QRS complex on the ECG monitor (Fig. 31.3).
    - Electrical capture is usually achieved between 50 and 100 mA.
    - If a patient is unconscious or truly deteriorating quickly or in cardiac arrest, it may be prudent to set the initial currents at maximum to ensure rapid capture and then decrease the current to just above that at which electrical capture was achieved.
9. After electrical capture is appreciated on the monitor, assess for mechanical capture by palpation of a pulse at a rate that corresponds to that which the machine is set at. An improved blood pressure or resolution of chest pain, shortness of breath, or altered mental status also suggests that the heart rate has improved and perfusion has been restored.
10. When pacing in overdrive for tachyarrhythmias, the pacer rate is set 20–60 beats/min faster than the detected tachycardic rate.
    - Bear in mind that rhythm acceleration or the induction of ventricular fibrillation is a possibility with pacing, hence the recommendation of having a code cart and airway equipment in the room at all times.

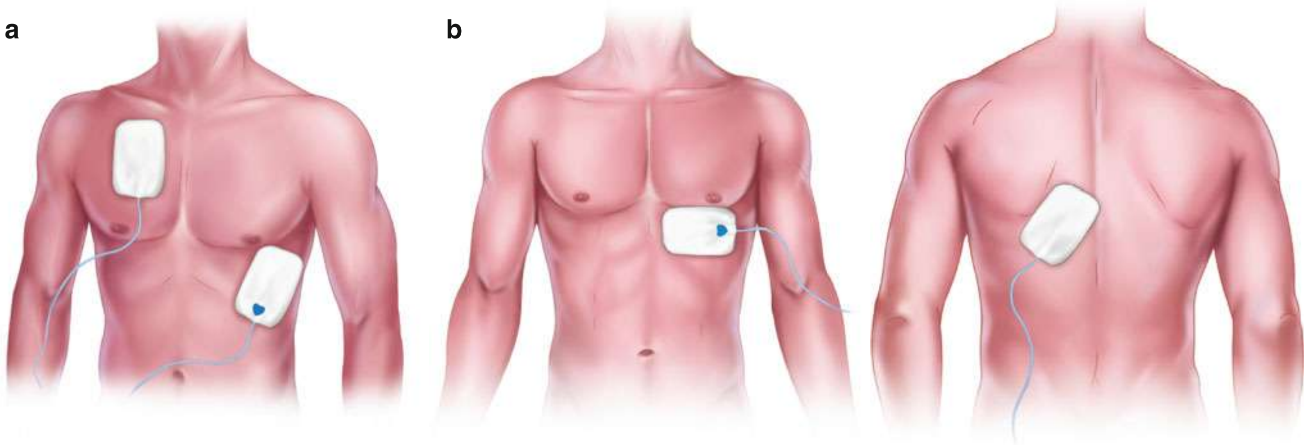

Fig. 31.2 Pacer pads attached in the anterolateral (**a**) or anteroposterior positions (**b**)

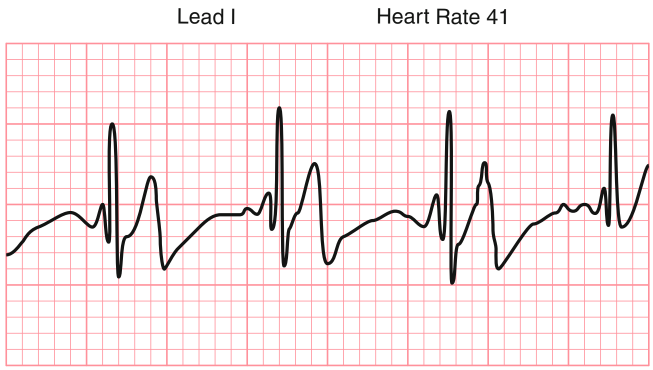
Lead I
Heart Rate 41

Bradycardia prepacing attempt

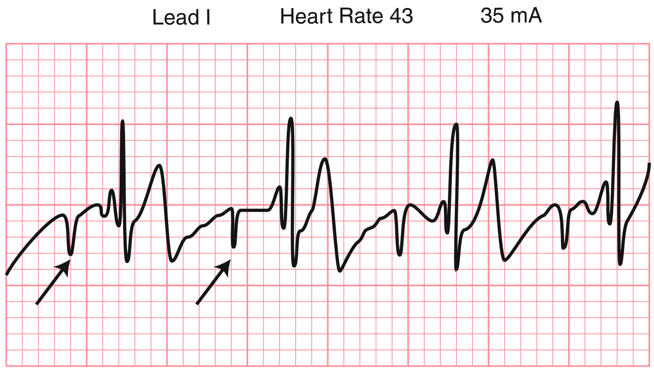
Lead I
Heart Rate 43
35 mA

Pacing attempted, note pacing stimulus indicator (arrow) which is below threshold, no capture

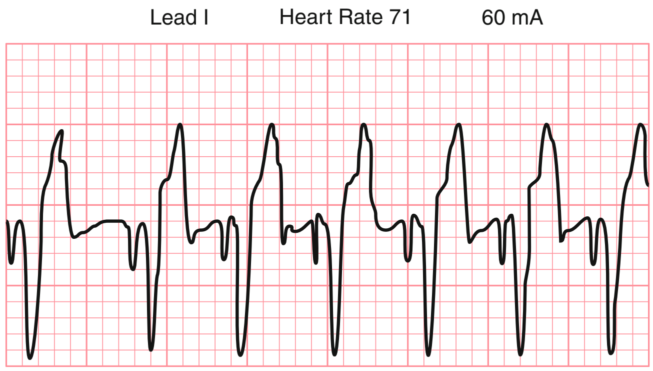
Lead I
Heart Rate 71
60 mA

Pacing above threshold (60mA), with capture (QRS complex broad and ventricular, T wave opposite QRS)

Fig. 31.3 A visible pacemaker spike will precede every QRS complex on the ECG monitor

31.5 Pearls and Pitfalls

- Pearls
  - Pacer pads can be prophylactically placed on bradycardic but stable patients in preparation for potential decompensation.
  - It may be prudent to discuss with a stable but bradycardic patient the option of trying the pacemaker not only to ensure that the machine and pads are working but also to gauge an idea of the output current that will be needed to electrically capture for quicker, more efficient, and potentially lifesaving pacing that may prove necessary.
  - When palpating for mechanical capture, the femoral pulse may be easier because carotid palpation may prove difficult with muscular contractions induced by the pacer impulses.
  - If impulses are not capturing, try slightly readjusting the pacer pads, making sure the leads are still on, the settings were not changed, the machine's battery did not die, or that the machine is plugged in.
  - If electrical capture is achieved but mechanical is not, increase the rate until adequate pulses are palpated.
  - Be careful to not miss an underlying ventricular fibrillation if the monitor is not blanked or dampened by an ECG screen.
  - Cardiopulmonary resuscitation (CPR) can be continued even while transcutaneous pacing is taking place because the pacer pads are insulated and the power delivered in each impulse is minimal, making the risk of injury to health-care workers very low.
- Pitfalls
  - Certain factors including a large body habitus, the presence of large pericardial effusions, scarring secondary to intrathoracic surgeries, or large amounts of intrathoracic air associated with obstructive pulmonary diseases may not only increase the threshold for capture but also may even cause a failure to capture, in which case transvenous pacing should immediately be considered.
  - Be aware that long-term pacing and pacing in children increases the likelihood for cutaneous and soft tissue damage.

Selected Reading

Bonow JS, Mann DL, Zipes DF, editors. Braunwald's heart disease: a textbook of cardiovascular medicine. 9th ed. Philadelphia: Saunders Elsevier; 2008.

Pfenninger JL, Fowler GC, editors. Pfenninger and Fowler's procedures for primary care. Philadelphia: Saunders Elsevier; 2010.

Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders Elsevier; 2010.

Tintinalli JE, Stapczynski JS, Cline DM, Ma OJ, Cydulka RK, Meckler GD, editors. Tintinalli's emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw-Hill Medical; 2011.

Transvenous Cardiac Pacing

32

Katrina John, Jeffrey Kile, and Amish Aghera

32.1 Indications

Hemodynamic compromise in the presence of:

- Sinus node dysfunction
- Second- and third-degree heart block
- Atrial fibrillation with slow ventricular response
- New left bundle branch block (LBBB)
- Right bundle branch block (RBBB)
- Bifascicular block
- Alternating bundle branch block (BBB)
- Implanted pacemaker malfunction

32.2 Contraindications

- Absolute
  - Prosthetic tricuspid valve
- Relative
  - Bradycardia in the presence of severe hypothermia

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_32

32.3 Materials and Medications (Fig. 32.1)

- Sterile gloves
- Sterile gown and drapes
- Face mask and surgical cap
- Two 10-mL syringes
- One 3-mL syringe
- Local anesthetic (1 or 2 % lidocaine without epinephrine)
- 22-gauge needle
- Povidone-iodine or chlorhexidine/isopropyl alcohol
- Several 4 × 4-in. gauze sponges
- Sterile dressing pack
- 4-0 nylon or silk sutures
- Scalpel (#11 blade)
- Needle holder
- Scissors
- Introducer needle
- Guidewire
- Dilator
- Introducer sheath
- Collapsed sterile extension sheath
- Balloon-tipped pacing catheter
- Pacing generator with working battery
- Insulated cable with alligator clamps or other suitable connectors
- Electrocardiogram (ECG)/cardiac monitor

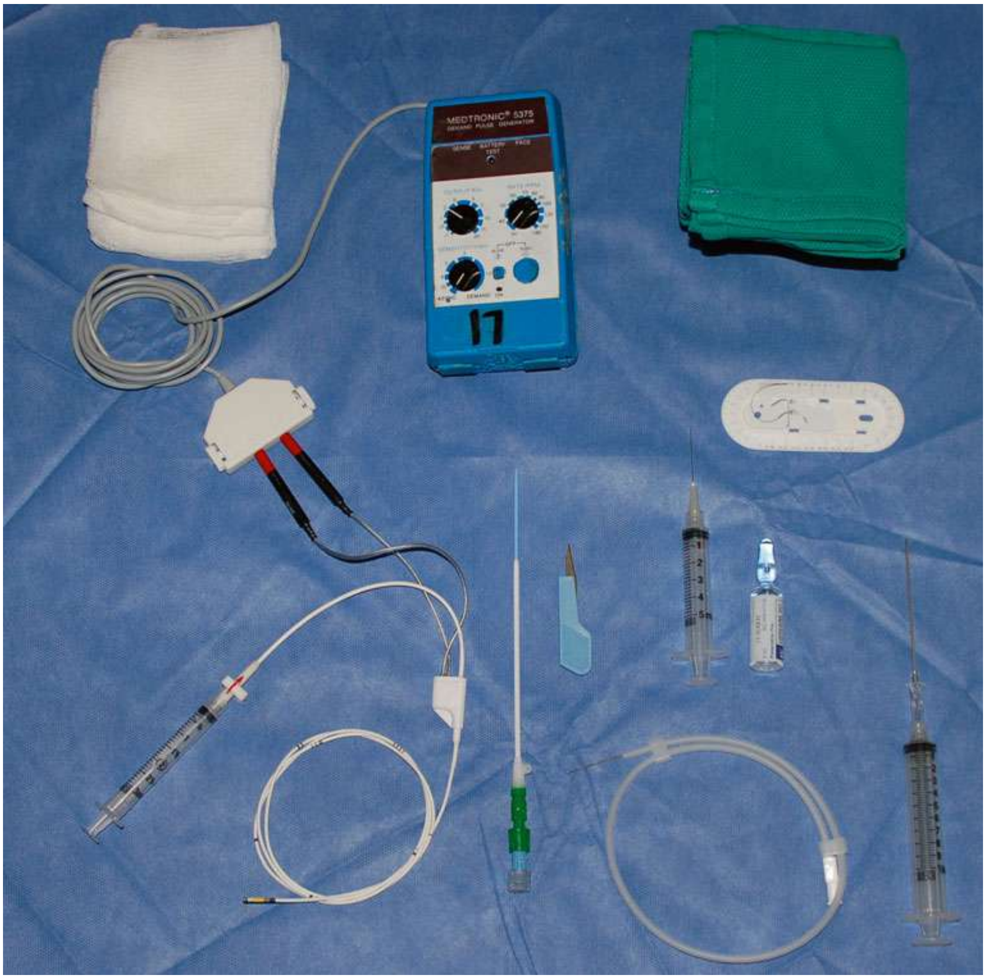

Fig. 32.1 Materials and medications

32.4 Procedure

1. Select appropriate catheterization site.
    - Four sites provide superior access to right atrium: brachial, femoral, internal jugular, and subclavian veins. The right internal jugular and left subclavian veins, respectively, are the most direct routes and, thus, the most commonly used in the emergency department setting.
2. If clinical situation permits, explain the procedure fully to the patient and obtain consent.
3. Sterilize the skin and apply sterile drapes.
4. Anesthetize the intended insertion site with lidocaine.
5. Prepare the pacing wire by inflating and deflating the balloon with 1–1.5 mL of air using a 3-mL syringe, and confirm that locking lever functions correctly to maintain balloon inflation (Fig. 32.2).
6. Using ultrasound guidance by means of a probe covered in a sterile sheath, insert the introducer needle into the vein while applying negative pressure to the attached 10-mL syringe (as with standard central line placement) (Fig. 32.3).
7. When flashback of blood is visualized in the syringe, remove the syringe, stabilizing the introducer needle firmly in place.
8. Pass the guidewire through the needle to a depth of 10–15 cm.
9. Holding the guidewire securely in place, remove the introducing needle.
10. Using a scalpel, make a small incision (approximately the width of the catheter to be used) through the dermis at the insertion site of the guidewire.
    - Avoid severing the guidewire by facing the sharp edge of the scalpel *away* from the guidewire.
11. Pass a dilator over the guidewire to make a tract in the skin and then remove the dilator.
12. Pass the introducer catheter over the guidewire until the hub is in contact with the skin and then remove the guidewire (Fig. 32.4).
13. Attach the collapsible sterile sheath to the hub of the introducer catheter.
14. Attach the positive and negative connectors of the pacing wire to their respective terminals on the cable(s) connected to the generator.
15. Set generator rate at 80 beats/min, output to 5 mA, and sensitivity to 3 mV.
16. Insert the free end of the pacing wire into introducer catheter with the balloon deflated and slowly advance the wire.
17. When the tip of the wire is within the superior vena cava, inflate the balloon, lock the valve to ensure that the balloon remains inflated, and continue to advance the pacing wire.
18. Closely watch the cardiac monitor for signs of capture (i.e., a wide QRS preceded by a pacing spike).
    - Markings on the pacing wire indicate its depth of insertion. At 20 cm, the wire should be in the right ventricle. If there is no capture observed by 25 cm, then the wire should be slowly withdrawn and advanced again.
19. Once capture is achieved, deflate the balloon.
20. Advance the wire a further 1–2 cm.
21. Coil the wire that remains outside the extended sterile sheath and suture it loosely but securely to the skin.
22. Suture the introducer hub to the skin.
23. Reduce generator output to zero and then increase it slowly to determine the minimum pacing threshold (i.e., the minimum voltage at which capture is achieved) and then set the output to twice this value.
24. Set the generator rate.
25. Stabilize the generator near the catheterization site.
    - For example, the generator can be attached to an intravenous (IV) fluid stand to the right of the patient's head if the pacing wire has been placed via the patient's right internal jugular vein.
26. Obtain a chest x-ray to confirm pacing wire position and absence of potential complications (e.g., pneumothorax) (Fig. 32.5).
27. Obtain a 12-lead ECG (Fig. 32.6).

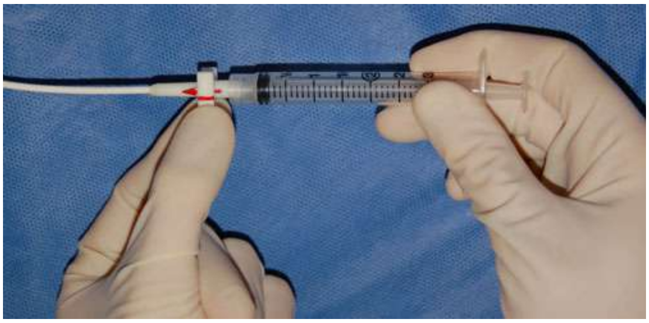

Fig. 32.2 Pacing wire balloon locking lever

Fig. 32.3 Insertion of introducer needle into internal jugular vein with ultrasound guidance

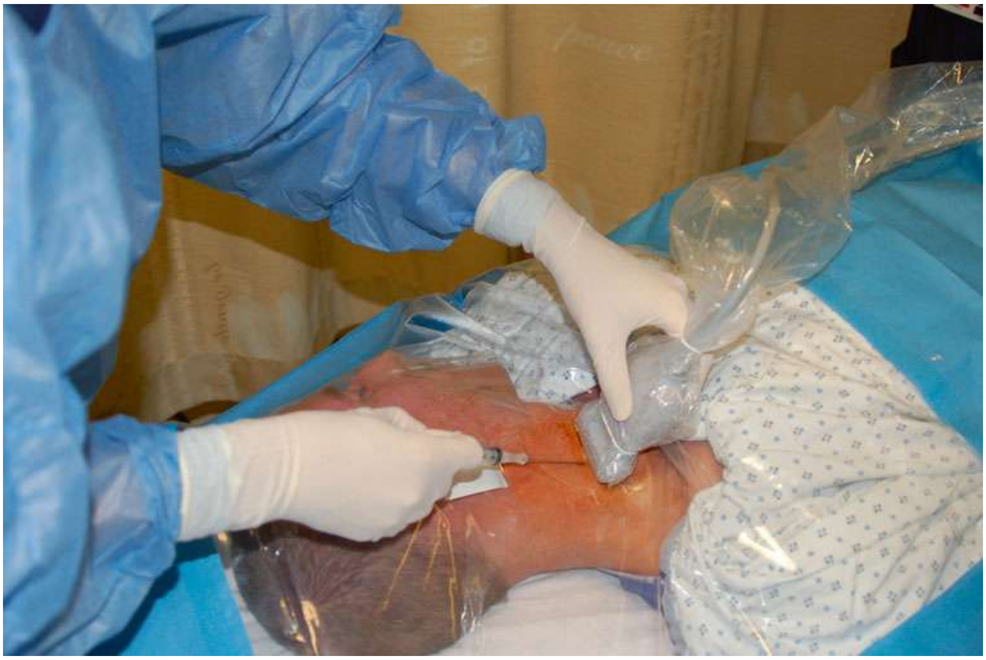

Fig. 32.4 Placement of introducer catheter over guidewire

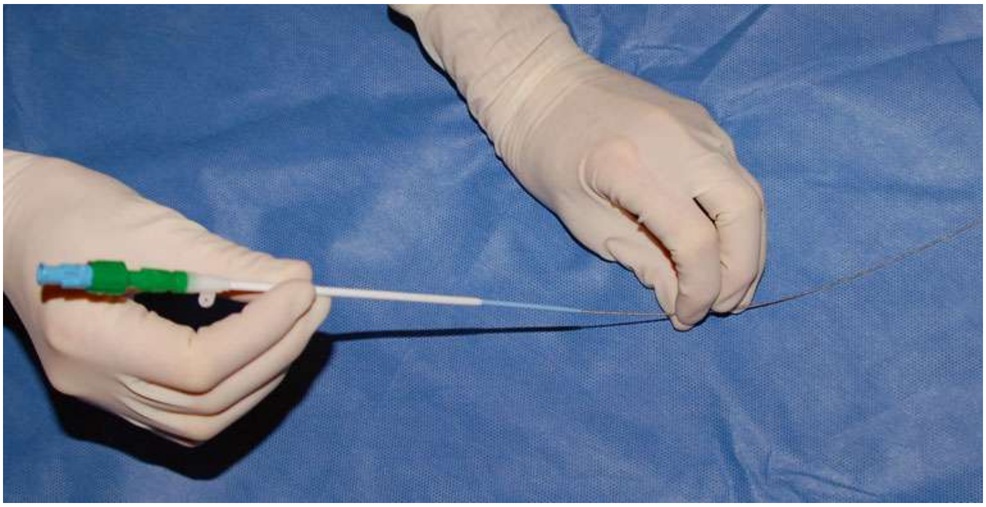

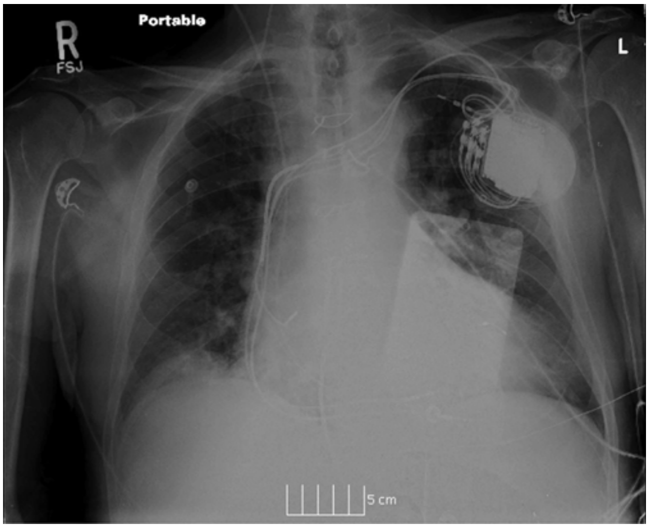
R
FSJ
Portable
L
5 cm

Fig. 32.5 X-ray displaying correct position of pacing wire in the right ventricle

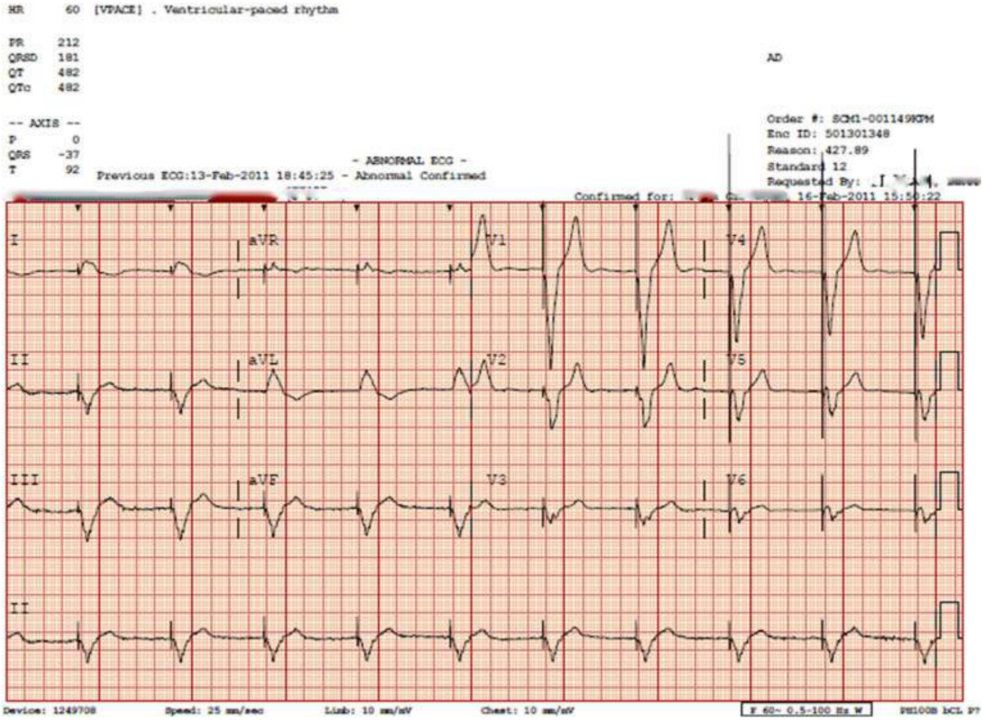
HR 60 [VPACE] . Ventricular-paced rhythm
PR 212
QRSD 181
QT 482
QTc 482
-- AXIS --
P 0
QRS -37
T 92
Previous ECG:13-Feb-2011 18:45:25 - Abnormal Confirmed
- ABNORMAL ECG -
AD
Order #: SCM1-001149KPM
Enc ID: 501301348
Reason: 427.89
Standard 12
Requested By: [redacted]
Confirmed for: [redacted] 16-Feb-2011 15:50:22
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
Device: 1249708
Speed: 25 mm/sec
Limb: 10 mm/mV
Chest: 10 mm/mV
F 60~ 0.5-100 Hz W
PH100B bCL P?

Fig. 32.6 Cardiac monitor displaying wide-complex QRS and pacing spike

32.5 Complications

- Arterial puncture
- Venous thrombosis
- Thrombophlebitis
- Pulmonary embolus
- Pneumothorax
- Fracture of guidewire with embolization
- Hemothorax
- Thoracic duct laceration
- Indwelling line infection (local or systemic)
- Dysrhythmias/premature ventricular contractions (PVCs)
- Insertion of pacing wire into coronary sinus/pulmonary artery
- Left (rather than right) ventricular pacing through atrial septal defect/ventricular septal defect (ASD/VSD)
- Septal perforation
- Ventricular perforation
- Entrapment/twisting of preexisting permanent pacing wires by the temporary pacing wire
- Balloon rupture
- Chordae tendineae rupture
- Pacing wire fractures
- Loss of capture owing to wire displacement or fracture
- Generator failure

32.6 Pearls and Pitfalls

- Confirm all necessary equipment is present and functional before beginning the procedure. If a dedicated transvenous pacing set is unavailable, individual components required (e.g., pacing catheter, connector cables, introducer catheter, guidewire) may need to be "cannibalized" from several different procedure kits in an emergency department.
- Become familiar with the locking valve for the pacing wire balloon before beginning the procedure. This small part is essential for inflating and deflating the balloon but is frequently difficult to adjust owing to stiffness (and all the more so if covered in blood).
- The ideal position for the pacing wire tip is against the diaphragmatic aspect of the right ventricle between its midpoint and its apex. This position is confirmed by development of an LBBB pattern with left axis deviation on the cardiac monitor. If difficulty is encountered achieving this position, the tip of the pacing wire can be moved to the right ventricular outflow tract, where the wire is less stable, but adequate pacing can still be achieved. In this latter position, the threshold will need to be kept relatively high and the monitor should display an LBBB pattern with an inferior axis.
- The use of balloon-tipped catheters enables the catheter tip to be directed with blood flow and its position (and successful capture) to be confirmed with a cardiac monitor. Nonballoon-tipped catheters, however, have typically required insertion with guidance of either constant ECG monitoring or fluoroscopy.

Selected Reading

Aguilera PA, Durham BA, Riley DA. Emergency transvenous cardiac pacing placement using ultrasound guidance. Ann Emerg Med. 2000;36(3):224–7.

Birkhahn RH, Gaeta TJ, Tloczkowski J, Mundy T, Sharma M, Bove J, Briggs WM. Emergency medicine-trained physicians are proficient in the insertion of transvenous pacemakers. Ann Emerg Med. 2004;43(4):469–74.

Dalsey WC, Syverud SA, Hedges JR. Emergency department use of transcutaneous pacing for cardiac arrests. Crit Care Med. 1985;13(5):399–401.

Harrigan RA, Chan TC, Moonblatt S, Vilke GM, Ufberg JW. Temporary transvenous pacemaker placement in the Emergency Department. J Emerg Med. 2007;32(1):105–11.

Silver MD, Goldschlager N. Temporary transvenous cardiac pacing in the critical care setting. Chest. 1988;93(3):607–13.

Pericardiocentesis (Optional: Ultrasound Guidance)

33

Katrina Skoog Nguyen, L. Connor Nickels, and Rohit Pravin Patel

33.1 Indications

- Treatment of hemodynamic compromise from cardiac tamponade
- To diagnose the cause or presence of a pericardial effusion

33.2 Contraindications

- Absolute
  - Aortic dissection
  - Need for immediate surgery for trauma patients
- Relative
  - Coagulopathy
  - Anticoagulant therapy
  - Thrombocytopenia

33.3 Materials and Medications

- Antiseptic (e.g., Chloraprep)
- 1 % lidocaine
- 25-gauge needle, 5/8-in. long
- 18-gauge catheter-type needle, 1½-nch long
- Syringes (10, 20, and 60 mL)
- Ultrasound (US) machine and cardiac/phased array probe
- Sterile US probe cover
- Cardiac monitor

33.4 Procedure

1. Identify the point of maximal effusion with US. Evaluate for hypoechoic or anechoic (dark) effusion around the heart, between the pericardial sac and the myocardium. A patient with hemodynamic compromise from a pericardial effusion or tamponade will have right ventricular collapse, septal bulging, and dilation of the inferior vena cava. Diastolic collapse of the right ventricular free wall can be absent in elevated right ventricular pressure and right ventricular hypertrophy or in right ventricular infarction.
2. Measure the distance from the skin surface to the effusion border to assess the expected needle depth.
3. Choose the needle trajectory based on the point of maximal effusion in the path with the fewest intervening structures. The most commonly used approaches are left parasternal, apical, and subxiphoid. For complex loculated posterior pericardial effusions, optional techniques such as transatrial and transbronchial may be performed by specialists. These types of loculated effusions can occur in autoimmune diseases, infective pericarditis, after cardiac surgery, and after radiotherapy.
4. Sterile preparation: prepare the skin of the entire lower xiphoid and epigastric area with antiseptic. Prepare the US transducer with a sterile sleeve.
5. Local anesthetic: if the patient is awake, anesthetize the skin and planned route of the needle.
6. Pericardial needle insertion: depends on approach used.

K.S. Nguyen, DO
Northwest Community Hospital, Arlington Heights, IL, USA
e-mail: katrinaskoog@hotmail.com;

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida
Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_33

33.4.1 **Subxiphoid Approach** (Fig. 33.1)

- The US transducer is placed just inferior to the xiphoid process and left costal margin.
- Insert the needle between the xiphoid process and the left costal margin at a 30–45° angle to the skin.
- Aim for the left shoulder.

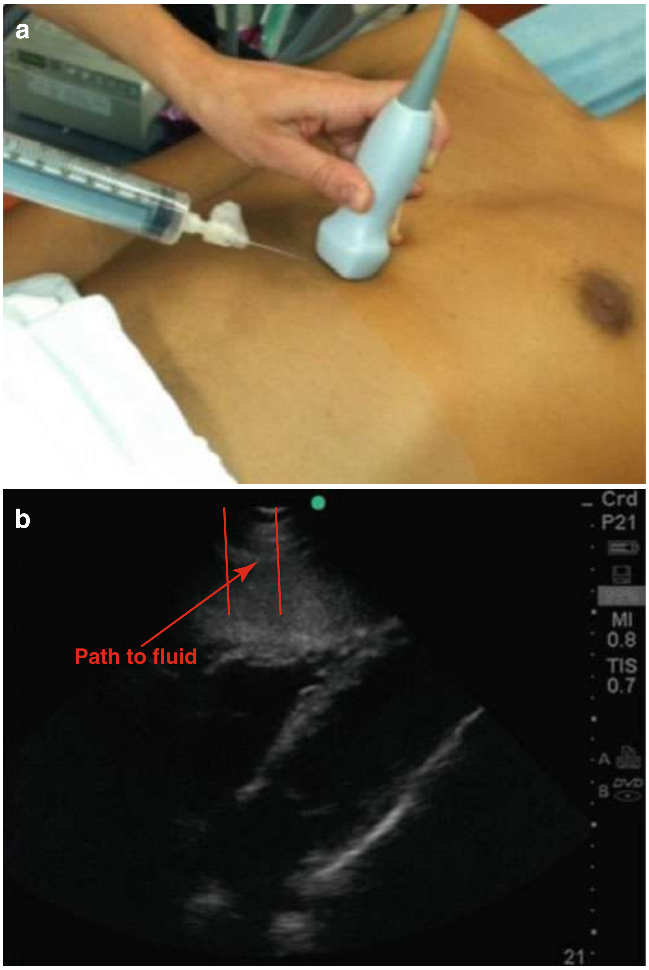
a
b
Path to fluid
Crd
P21
MI
0.8
TIS
0.7
21

Fig. 33.1 Subcostal approach (**a**) and corresponding ultrasound image (**b**). *Red line* through liver to obtain fluid (no fluid present on ultrasound image)

33.4.2 **Apical Approach** (Fig. 33.2)

- The US transducer is placed at the patient's point of maximal impulse and aimed at the patient's right shoulder for a four-chamber view of the heart.
- Insert the needle in the fifth intercostal space 1 cm lateral to and below the apical beat, within the area of cardiac dullness.
- Aim for the right shoulder.
- A Mayo Clinic review showed, in 80 % of total effusions, that the distance to the effusion was least and the size was maximal in the apical approach [1].

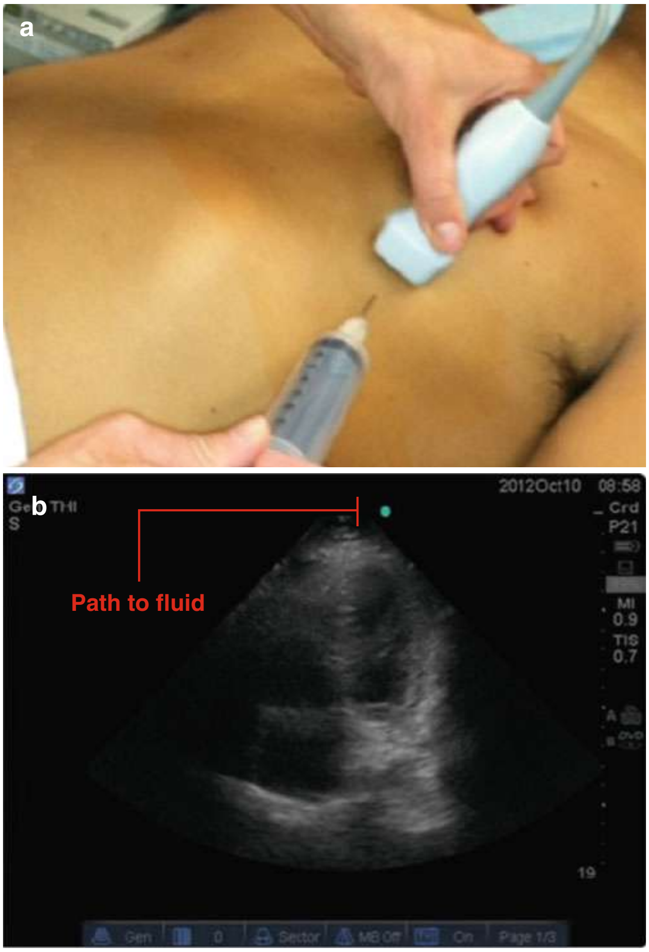
a
b
THI
Path to fluid
2012Oct10 08:58
Crd
P21
MI
0.9
TIS
0.7
19
Gen
Sector
MB Off
On
Page 1/3

Fig. 33.2 Apical approach (**a**) and corresponding ultrasound image (**b**). *Red line* with minimal distance to fluid if present (no fluid present on ultrasound image)

33.4.3 Parasternal Long-Axis Approach

- The US transducer is placed obliquely on the left sternal border between the fourth and fifth ribs with the transducer indicator aimed at the right shoulder.
- Insert the needle perpendicular to the skin in the fifth intercostal space medial to the border of cardiac dullness.
  - Visualize and feel a giving way as the needle penetrates the pericardium. Removal of fluid confirms successful entry. Remove fluid with the goal of restoring hemodynamic stability. Aspiration of fluid should result in improvement in blood pressure and cardiac output.
  - Remove the catheter and apply a dressing. Optional: Place a pigtail catheter using the Seldinger technique for continued drainage.

33.5 Complications

- Blind techniques are associated with 20 % morbidity and 6 % mortality. The complication rate with US-guided approaches is less than 5 % [2].
- Any vital structure within reach of the pericardial needle has the potential for injury: pneumothorax; hemothorax; coronary vessel laceration; hemopericardium; heart chamber lacerations; intercostal vessel injury; dysrhythmias; ventricular tachycardia; puncture of the liver, diaphragm, or gastrointestinal tract; bacteremia; purulent pericarditis; air embolisms; and pleuropericardial fistulas.

33.6 Pearls and Pitfalls

- Pearls
  - A scoring index is available that can be obtained at initial presentation in patients without hemodynamic compromise that identifies those who require pericardial effusion drainage later in the course of treatment. The scoring index uses echocardiographic findings, etiological information, and the size of the effusion at end-diastole [3].
  - Take care to avoid the left internal mammary artery, which travels in a cephalad-to-caudad direction 3–5 cm lateral to the left sternal border [4].
- Pitfalls
  - Tamponade should always be considered in the differential diagnosis of shock or cardiac arrest owing to pulseless electrical activity (PEA). Complications of acute coronary syndrome, aortic dissection, and decompensations in chronic advanced constriction may also need treatment [5].
  - When cardiac compensation mechanisms are exhausted, small increases in pericardial volume can lead to an increase in ventricular diastolic pressure, systemic and pulmonary congestion, and decrease preload and cardiac output.
  - Use US-guidance rather than a blind or an electrocardiogram (ECG)-alone-guided approach for pericardiocentesis to significantly decrease the risk of injury to vital structures [4].
  - Do not delay patient transport to the operating room to perform pericardiocentesis on a traumatically induced effusion unless the patient is hemodynamically unstable and on the verge of cardiac arrest.

References

1. Tsang TS, Enriquez-Serano M, Freeman WK, Barnes ME, Sinak LJ, Gersh BJ, et al. Consecutive 1127 therapeutic echocardiography guided pericardiocentesis: clinical profile, practice patterns, and outcomes spanning 21 years. Mayo Clin Proc. 2002;77:429–36.
2. Guo K, Ding ZP, Tan J. Trans-pleural pericardiocentesis: revisiting an old technique. Catheter Cardiovasc Interv. 2011;78:815–8.
3. Halpern DG, Argulian E, Briasoulis A, Chaudhry F, Aziz EF, Herzog E. A novel pericardial effusion scoring index to guide decision for drainage. Crit Pathw Cardiol. 2012;11:85–9.
4. Seferović PM, Ristić AD, Imazio M, et al. Management strategies in pericardial emergencies. Herz. 2006;31:891–901.
5. Sagristà-Sauleda J, Mercé AS, Soler-Soler J. Diagnosis and management of pericardial effusion. World J Cardiol. 2011;3:135–43.

Ultrasound Evaluation of Pulmonary Embolism and Heart Strain

34

Giuliano De Portu, L. Connor Nickels, and Marie-Carmelle Elie

34.1 Indications

- Undifferentiated respiratory distress
- Unexplained hypotension
- Evaluation for cardiac tamponade, pericardial effusion

34.2 Contraindications

- Relative
  - Morbidly obese patients
  - Patients with chest wall deformities
  - Patient with subcutaneous emphysema, pneumopericardium
  - Combative or altered patients

34.3 Materials and Medications

- Portable ultrasound machine with appropriate probes (phased array)
- Ultrasonic gel for the probe
- Drapes for the patient (if the condition allows)

34.4 Procedure

1. Begin with the parasternal long-axis view. If possible have the patient turn to the left decubitus side (that helps "move" the heart closer the chest wall).
2. Find the phased array probe, select the "cardiac examination" on the machine, and make sure the orientation marker is pointed toward the right shoulder of the patient.
3. Place the probe along the left side of the sternum over the fourth to sixth intercostal space. This should produce the image shown in Fig. 34.1.
4. Measure the diameter of the ventricle during the end of diastole (normal values, 21 mm ± 1 mm; any measurement >25–30 mm is abnormal).
5. Using the aorta as a landmark, evaluate the structures starting with the pericardium (bright white line around the heart), making sure there is no fluid around it. (Fluid above the aorta indicates a pericardial effusion. Fluid below the aorta indicates a pleural effusion.) If a hypoechoic or anechoic stripe appears in the anterior side of the heart, it is most likely a fat pad. Fluid seen "all around" categorizes it as an effusion.
6. While keeping the probe in the same place, rotate the probe marker 90° clockwise toward the left shoulder to obtain the parasternal short-axis view (Fig. 34.2).

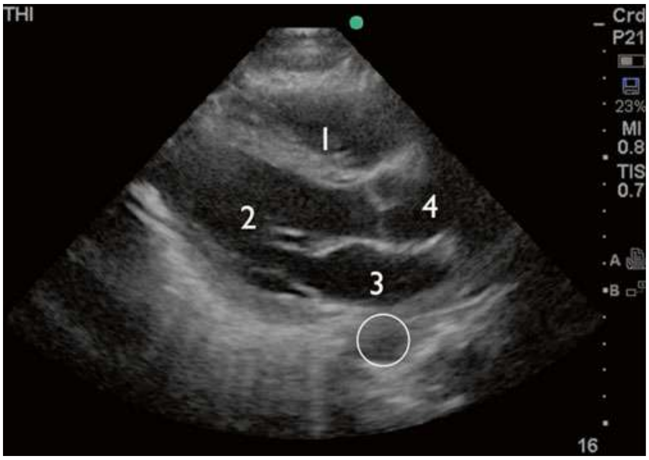

Fig. 34.1 Parasternal long-axis view (PSLA). *1* right ventricle, *2* left ventricle, *3* left atrium, *4* aortic outflow track; *circle* descending aorta (right atrium is not visualized in the PSLA)

G. De Portu, MD • L.C. Nickels, MD, RDMS
M.-C. Elie, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: gdeportu@ufl.edu; cnickels@ufl.edu; elie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_34

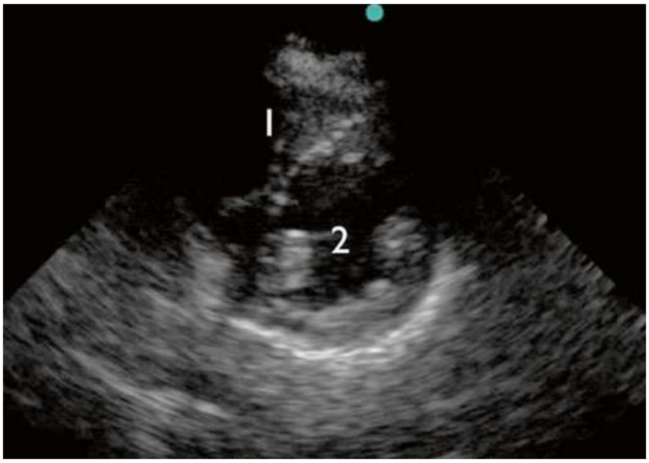

Fig. 34.2 Parasternal short-axis view (PSSA). *1* right ventricle, *2* left ventricle

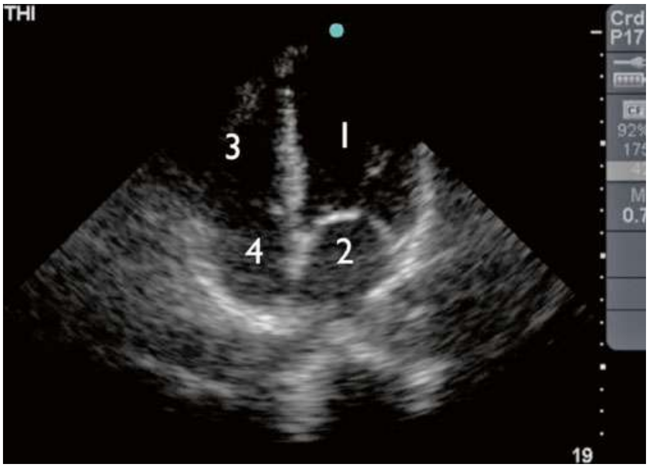

Fig. 34.3 Four-chamber view (4C). *1* left ventricle, *2* left atrium, *3* right ventricle, *4* right atrium

7. The parasternal short-axis view will give information on the contractility of the heart. The right ventricle should be anterior and to the left and the left ventricle to the right. The normal position of the septum bows slightly toward the right ventricle.
8. Obtain a four-chamber view. Place the orientation marker to the patient's left.
9. Palpate for the point of maximum impulse (PMI) and place the transducer. All four chambers should appear in one view (Fig. 34.3).
10. Compare the sizes of the ventricles and note any difference. Notice also the interventricular septum. The normal right-to-left ventricular ratio is less than 0.5. (In an unstable patient, this is probably the most useful because both ventricles can be quickly visualized and the ratio compared.)
11. A subxiphoid approach is also possible. Make sure the orientation marker is toward the right side of the patient (Fig. 34.4).

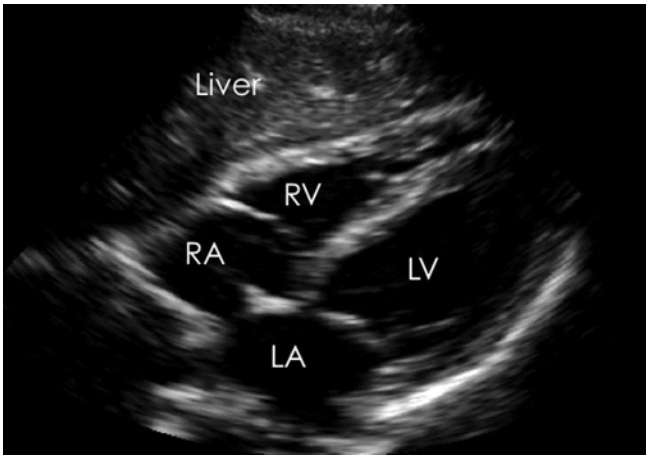

Fig. 34.4 Subcostal/subxiphoid view. *RV* right ventricle, *RA* right atrium, *LV* left ventricle, *LA* left atrium

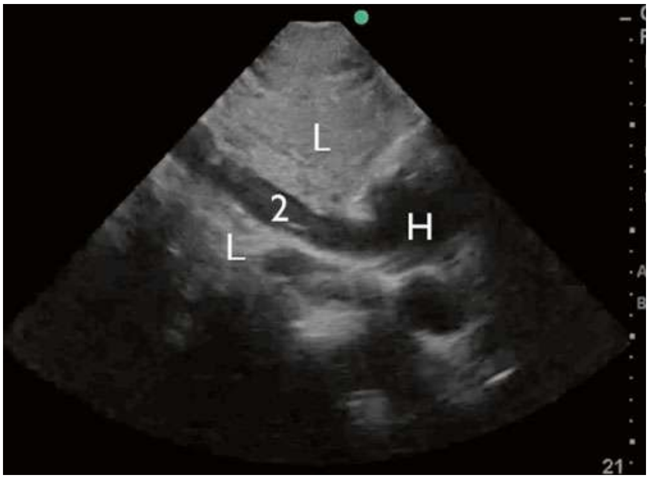

Fig. 34.5 Inferior vena cava. *L* liver, *2* inferior vena cava, *H* heart

12. Use the liver for orientation. The right side of the heart will be nearest to the liver (think that the liver is on the right of the body); again note for any differences in the size ratio.
13. Next take a look at the inferior vena cava (IVC) by placing the curved array transducer just inferior to the xiphoid in a longitudinal fashion. (Switch from the subxiphoid view to the IVC by rotating the probe counterclockwise until the IVC is seen.) Evaluate the IVC. During normal physiological inspiration, the drop in intrathoracic pressure "pulls" blood into the heart, thus decreasing the relative IVC size. If something is preventing venous return, such as a massive PE, collapse will not be as evident and the suspicion for PE increases (fluid overload and increased central venous pressure [CVP] will also account for this finding). Normal IVC diameter is 1.2–2.3 cm, and total collapse and greater than 50 % collapse are normally visualized. An increase in IVC size and less than 50 % or no change has been correlated with increased right atrial pressures (11 to >20 cm Hg) (Fig. 34.5).

34.5 Findings

A right heart that is "strained" or pumping against a higher resistance owing to a PE will show some or all of these changes:

1. Right ventricular dilation (Fig. 34.6).
2. Right ventricular hypokinesis (especially of the middle segment), McConnell's sign but normal motion of the apex.
3. Tricuspid regurgitation.
4. Abnormal septal motion: deviated toward the left ventricle (normally it relaxes during diastole toward the right ventricle); as pressure increases, the right ventricle will not empty properly and septal flattening can be seen.
5. Dilated IVC with little or complete loss of changes in diameter with respiration (variability); the IVC collapses less than 50 % during inspiration.

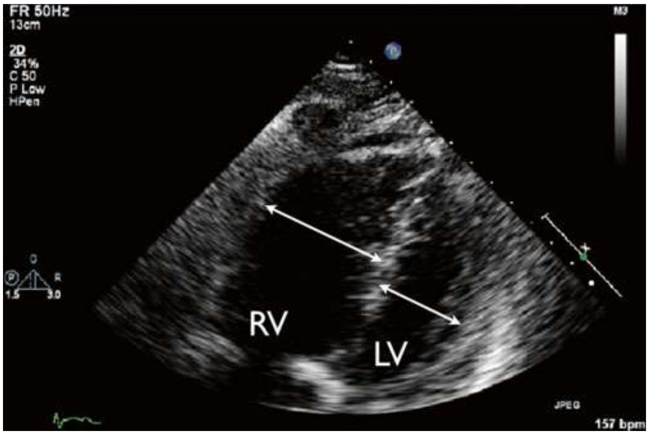

Fig. 34.6 Four-chamber view that shows right ventricular dilatation due to pulmonary embolism. Normally, the LV is greater than the RV during diastole, but in the case of increased pressure, the RV will be enlarged. *RV* right ventricle, *LV* left ventricle

34.6 Complications

- Ultrasound per se has been shown to cause no direct complications with proper use unless it is interfering with advanced airway or life-saving procedures.

34.7 Pearls and Pitfalls

- Pearls
  - Proper interpretation of the images is necessary because some normal anatomy could be confused with a positive finding. Make sure to properly identify the structures because a common mistake is to identify the aorta thinking it is the IVC.
  - Depth: increase to be sure to visualize all structures.
- Pitfalls
  - Ultrasound is user dependent, and the acquisition of images will vary with each user, creating the possibility of false-negatives if images not properly acquired.
  - Reversal of the orientation owing to transducer misplacement could "reverse" the anatomy and allow mistakenly identification of structures.
  - An enlarged right ventricular wall (>5 mm) is considered hypertrophied and would indicate a more chronic etiology of a right ventricular dilation.
  - Gain problems: adjust so the heart chambers are anechoic and the walls are echogenic.

Selected Reading

Kline JA. Thromboembolism In: Tintinalli J, Stapczynski J, Ma OJ, Cline D, Cydulka R, Meckler G, editors. Tintinalli's emergency medicine. 7th ed. New York: McGraw Hill; 2012.

McConnell MV, Solomon SD, Rayan ME, Come PC, Goldhaber SZ, Lee RT. Regional right ventricular dysfunction detected by echocardiography in acute pulmonary embolism. Am J Cardiol. 1996;78:469–73.

Reardon RF, Joing SA. Cardiac. In: Ma OJ, Mateer J, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw Hill Professional; 2007. p. 109–44.

Vanden Hoek TL, Morrison LJ, Shuster M, et al. Part 12: cardiac arrest in special situations: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010;122(18 Suppl 3):S829–61.

Pacemaker Evaluation in the Emergency Department

35

Joseph D. Romano and Christian Coletti

35.1 Pacemaker Function

There are approximately 500,000 implanted cardiac pacemakers in the USA and another 100,000 are implanted each year [1]. In the emergency department, a physician may be faced with a pacemaker that is not functioning appropriately. To understand the ways in which a pacemaker can malfunction or lead to medical complications, first it is important to understand how pacemakers work when they do so appropriately.

A common pacemaker system is composed of a pulse generator and insulated wire leads that originate in the pulse generator and end within the myocardium. The pulse generator is implanted in the pectoral region. It contains circuitry and the battery and creates the electrical impulses that depolarize the myocardium. The leads leave the pulse generator and are directed to the heart by following the venous system. Leads may terminate in the right atrium, the interventricular septum of the right ventricle, or biventricular pacing; a third lead navigates the coronary sinus to the left ventricular wall (Figs. 35.1 and 35.2).

Early pacemakers produced impulses only at a set rate. Today, however, virtually all pacemakers have sensing leads that detect intrinsic activity and react with electrical impulses when the intrinsic intervals fall outside of a set time threshold. Some pacemakers are programmed to allow for rate changes based on physical activity. To simplify the classification of different pacemaker types, the North American Society of Pacing and Electrophysiology and British Pacing and Electrophysiology Group (NASPE/BPEG) developed a five-letter code to describe each pacemaker (Table 35.1) [2].

J.D. Romano, MD (✉) • C. Coletti, MD
Departments of Emergency Medicine and Internal Medicine,
Christiana Care Health System,
Newark, DE, USA
e-mail: jromano@christianacare.org; chris.coletti@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_35

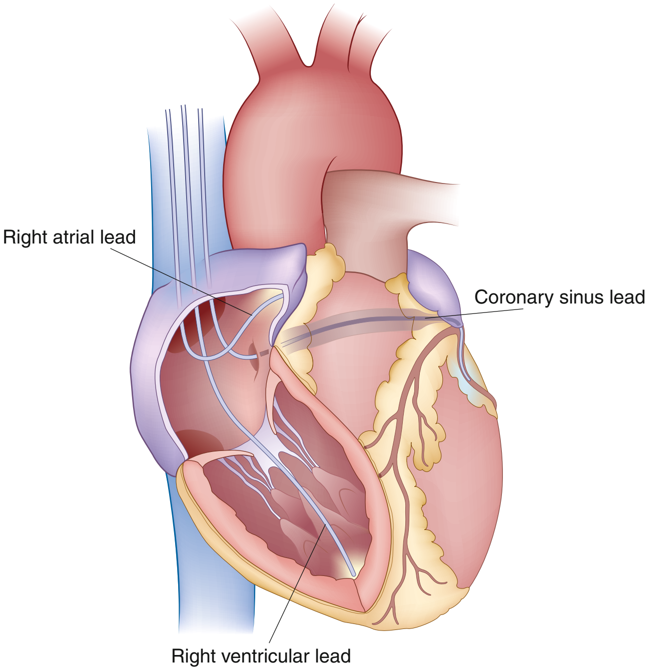
Right atrial lead
Coronary sinus lead
Right ventricular lead

Fig. 35.1 Common pacemaker lead insertion sites

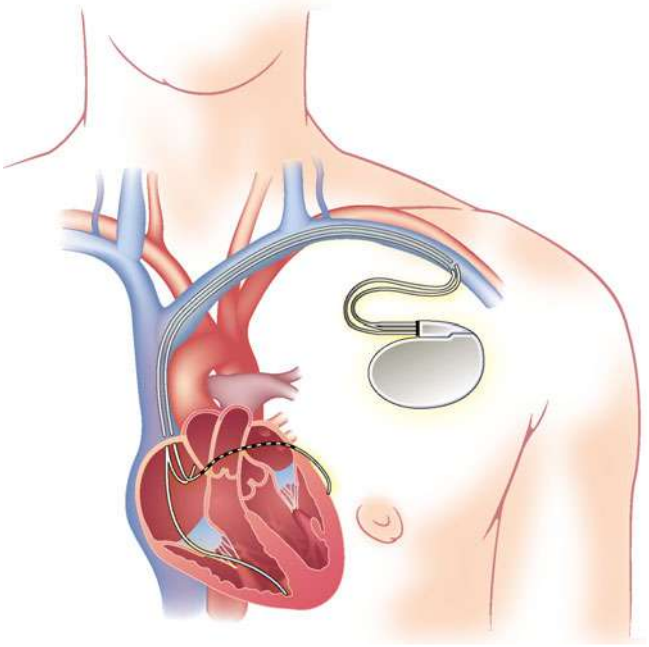

Fig. 35.2 Common generator location and path of pacemaker leads

Table 35.1 NASPE/BPEG five-letter pacemaker code

Paced chamber	Sensed chamber	Response to sensing	Rate modulation	Multisite pacing
A = atrium	A = atrium	T = triggered	R = rate modulation	A = atrium
V = ventricle	V = ventricle	I = inhibited	O = none	V = ventricle
D = dual	D = dual	D = dual		D = dual
O = none	O = none	O = none		O = none

Reproduced with permission from Bernstein et al. [2]

35.2 Common Pacemaker Codes [3]

- VVI: ventricular paced and sensed, inhibited by normal intrinsic pacing; used for patients with a need for pacing if bradycardia occurs
- VVIR: ventricular paced and sensed, inhibited by normal intrinsic pacing; adaptive to changes in intrinsic rate for physically active patients
- DDD: dual-chamber paced and sensed, inhibited by normal intrinsic atrial and ventricular electrical activity; used in third-degree atrioventricular (AV) block to allow for more physiological conduction
- DDDR: dual-chamber paced and sensed, inhibited by normal intrinsic atrial and ventricular electrical activity; adaptive to changes in intrinsic rate; used primarily in sinoatrial (SA) node dysfunction to closely mimic normal, adaptable heart conduction

35.3 Pacemaker Malfunction

Malfunctioning pacemakers can have complicated programming that is best altered after thorough "interrogation" by a trained electrophysiologist. This type of evaluation is beyond the scope of this chapter and, in many clinical settings, subspecialty support is not readily available. In the emergency department, it is vital to stabilize the patient and identify common pacemaker system issues. Pacemaker system malfunctions can be identified as a failure to sense, failure to pace, failure to capture, or pacing at an inappropriate rate [4]. Another common abnormality associated with pacing is the "pacemaker syndrome" [3].

- Failure to sense:
  - Oversensing: The sensor interprets external stimuli as a normal ventricular rate leading to inappropriate inhibition. This may be due to lead fracture, fibrosis of the lead tip, or lead dislodgment. Cross talk is present when the atrial stimulus is sensed on the ventricular lead causing inappropriate inhibition. Oversensing is rarely due to failure of the generator.
  - Undersensing: Present when there is constant pacing despite cardiac activity that has not exceeded threshold. This can be due to low-amplitude intrinsic cardiac activity, lead dislodgment, battery depletion, or metabolic abnormalities.
- Failure to pace:
  - Lack of pacing despite appropriate sensing of the intrinsic electrical activity or activity below the threshold rate.
  - Typically caused by lead fracture, battery depletion, or fibrosis of the lead tip.
  - It is rare for any specific part of the device to fail, but suspicion should be high if the patient had recent radiation therapy, electrocautery, defibrillation, electroshock therapy, magnetic resonance imaging (MRI), lithotripsy, or transcutaneous electrical nerve stimulation (TENS) treatments.
  - Rarely due to battery depletion.
- Failure to capture:
  - A lack of cardiac response despite appropriate sensing and subsequent pacer stimuli.
  - Commonly due to lead dislodgement, but can be due to myocardial perforation, lead fracture, fibrosis of the lead tip, poor lead placement, battery depletion, and antiarrhythmic medications.
  - Functional failure to capture occurs when pacer stimuli fall within the refractory period of previous depolarization.
- Pacing at an inappropriate rate:
  - Likely due to an endless-loop reentry tachycardia known as pacemaker-mediated tachycardia.
    - More common in DDD pacemakers.
    - Initiated by a premature ventricular stimulus, which is carried retrograde through the intrinsic conduction system to the atrioventricular node (AVN) and then the atria. This conduction is sensed by the atrial lead and causes triggering of pacing back in the ventricle. The ventricular depolarization is then sensed again in the atria, forming a loop. The intrinsic conduction system acts as a retrograde limb and the pacer circuit acts as the anterograde limb of a reentry tachycardia.
- The "pacemaker syndrome":
  - Constellation of symptoms found in 20 % of patients with pacemakers.
  - Symptoms include syncope, near-syncope, dizziness, fatigue, weakness, pain, shortness of breath, and cough.
  - Ventricular rates are poorly timed with atrial activity such that atrial contraction occurs against closed mitral and tricuspid valves.
  - Ventricles lose the benefit of the atrial kick, the atria enlarge, and signs and symptoms similar to congestive heart failure ensue.

35.4 Initial Evaluation of the Patient with a Pacemaker [4]

- History:
  - Symptoms of syncope, near-syncope, chest pain, palpitations, and irritation at the generator pocket
  - Brand and NASPE/BPEG code
  - Date of implantation
  - Location of generator pocket and any previous locations
  - Most recent electrophysiological interrogation
  - Medications that interfere with function, such as flecainide or lidocaine
  - Recent procedures, such as MRI, lithotripsy, or defibrillation
- Physical examination:
  - Check vital signs, listen for heart sounds (paradoxically split S2 is normal).
  - Inspect generator pocket; turning of the leads is associated with Twiddler's syndrome, in which a patient compulsively touches the skin around the device.
  - Look for jugular venous distension.
- Laboratory studies and electrocardiogram (ECG):
  - Obtain a chest x-ray to determine effusions, infiltrates, generator placement, and lead placement or fracture.
  - Electrolyte imbalances must be detected and corrected.
  - ECG: The following are examples of appropriate ECG patterns in patients based on type of pacing [3]:
    - VOO (asynchronous pacing): Regular pacer spikes lead to immediate QRS waveforms. Left bundle branch block (LBBB) is a normal finding in a right ventricular lead location.
    - VVI: Pacer spikes and an LBBB waveform should be seen if the intrinsic rate is below the threshold rate.
    - DDD: Various patterns are possible. If the intrinsic rhythm and intervals are normal, then no pacer spikes will be seen. If the atrial rate is slow and AV delay is normal, then an atrial spike will cause a P wave and a normal QRS. If AV delay is prolonged, then two spikes may be observed: a P wave and a QRS with an LBBB waveform.
- If malfunction is suspected:
  - Obtain intravenous (IV) access, place the patient on a heart monitor, and, if possible, consult cardiology for interrogation and reprogramming.
  - Obtain a ring magnet:
    - Positioning over the generator causes the pacing stimuli to revert to an asynchronous ventricular pacing mode at a set rate [5].

35.5 Interpreting the Type of Malfunction Based on ECG

The first step in evaluating the ECG of a malfunctioning pacemaker is to determine the presence of pacing spikes [6]:

- If pacing spikes are noted on the ECG, determine whether capture is present by ensuring that an appropriate waveform follows each pacer spike and that there is an associated pulse.
  - If capture is present, check the rate:
    - If the rate is appropriate, the pacer is functioning normally.
      - If the rate is slow, suspect oversensing.
      - If the rate is rapid, suspect undersensing or pacemaker-mediated tachycardia.
  - If capture is not present, consider metabolic effects or component failure.
- If pacing spikes are not noted on the ECG, determine whether the patient is in an intrinsic rhythm:
  - If the patient is in an intrinsic rhythm and the rate is appropriate, application of a ring magnet will cause pacer spikes to show up at a set rate. This is normal functioning.
    - If application of the magnet does not cause pacing, suspect mechanical failure.
  - If the patient is not in an intrinsic rhythm, place a magnet over the generator.
    - If magnet application causes pacing, consider oversensing.
    - If it does not, consider mechanical failure.
- If a patient is hemodynamically unstable and application of a magnet leads to stability, keep the magnet in place until the patient is able to have the pacemaker interrogated.

35.6 Management of Pacemaker-Mediated Tachycardia

- After ensuring adequate IV access, placing a heart monitor, and interpreting a baseline ECG, put a ring magnet over the generator [5].
  - If a normal rhythm results, keep the magnet on the chest.
  - If it does not change the rate, attempt isometric pectoral exercises by having the patient press the left hand against the right shoulder.
    - This is an attempt to overstimulate the pacemaker sensor and precipitate inhibition of pacer output.
    - If this is unsuccessful, consider transcutaneous pacing.
    - If transcutaneous pacing is unsuccessful, then the leads may require surgical adjustment or removal.

References

1. Ford-Martin PA, Spiwak AJ. Pacemakers. In: Gale encyclopedia of surgery: a guide for patients and caregivers. 2004. Encyclopedia.com: http://www.encyclopedia.com/doc/1G2-3406200337.html. Accessed 01 Apr 2011.
2. Bernstein A, Daubert J, Fletcher R, et al. The revised NASPE/BPEG generic code for antibradycardia, adaptive-rate, and multisite pacing. North American Society of Pacing and Electrophysiology/British Pacing and Electrophysiology Group. Pacing Clin Electrophysiol. 2002;25:260–4.
3. Marx JA, Hockberger RS, Walls RM, Adams J, editors. Rosen's emergency medicine. 7th ed. Philadelphia: Mosby; 2009.
4. Bonow RO, Mann DL, Zipes DP, Libby P, editors. Braunwald's heart disease: a textbook of cardiovascular medicine. 9th ed. Philadelphia: Saunders; 2011.
5. Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 4th ed. Philadelphia: Saunders; 2004.
6. Kaszala K, Huizer JF, Ellenbogen KA. Contemporary pacemakers: what the primary care physician needs to know [review]. Mayo Clin Proc. 2008;83:1170–86.

Part V

Spine Evaluation

Cervical Collar Placement

36

Justin Bennett and Lars K. Beattie

The cervical spine (C-spine) accounts for the majority of all spinal injuries. In the US prehospital setting, patients are often transported using cervical collars (C-collars) and rigid backboards in trauma and when spinal cord injuries are suspected. After airway, breathing, and circulation, every effort must be made to secure the C-spine.

There are two main subcategories of C-collars: the one-piece and the two-piece. One-piece C-collars include the Stifneck and the Ambu Perfit. Two-piece C-collars have posterior and anterior pieces, with the anterior piece usually the larger of the two. Examples of two-piece collars include the Aspen collar, the Philadelphia collar, and the Miami J collar. The basic features of a C-collar include:

- Adjustable circumference with fasteners (usually Velcro straps)
- Adjustable height with a locking device of different sizes
- Hooks for a nasal cannula
- Exposure of the anterior neck for pulse checks and advanced airway procedures
- Posterior access for cervical palpation
- Padding to protect the soft tissues of the neck

36.1 Indications

- Prehospital suspicion for spine trauma:
  - Emergency department patients or trauma patients who fail clinical rule-out criteria such as NEXUS (National Emergency X-Radiography Utilization Study) and Canadian C-spine rules (see Chap. 37)

36.2 Contraindications

- Absolute:
  - Cervical dislocation with fixed angulation
  - Impaled foreign object in the neck
  - Massive soft tissue swelling in the neck
- Relative:
  - Unsecured airway
  - Surgical airway
  - Vomiting
  - Mandible or soft tissue injuries with potential for airway compromise
  - Preexisting anatomical abnormalities

36.3 Materials and Medications

- Properly fitting C-collar
- Consider:
  - Head blocks, if needed for lateral stabilization
  - Towels or backboard pads for custom support
    - Under shoulders – pediatric patients
    - Under occiput – adults with poor C-spine mobility

J. Bennett, MD
Department of Emergency Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, USA
e-mail: jubennet@wakehealth.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida, Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_36

36.4 Procedure (Aspen Collar)

1. Address airway, breathing, and circulation while maintaining in-line immobilization, before placing a C-collar.
2. Gather personnel:
   - (a) One person is needed to apply the collar to an awake patient.
   - (b) Two or more people may be required when a patient has an altered level of consciousness:
     - (i) One to maintain in-line immobilization in the neutral position
     - (ii) One to place the C-collar
3. While the C-spine is being held in neutral position, assess the airway before placing the C-collar:
   - (a) Anticipate and prepare for airway compromise early to avoid a crash intubation.
   - (b) Place airway if necessary.
4. Palpate and inspect the C-spine, head, and shoulders for evidence of trauma *before* placing the C-collar.
5. While maintaining the neutral position of the C-spine, place the C-collar:
   - (a) Remove loose clothing, jewelry, and earrings that may cause soft tissue pressure wounds.
   - (b) Begin with the piece of C-collar that fits under the occiput (Fig. 36.1).
   - (c) Fold the Velcro straps behind the C-collar.
   - (d) Gently hold back the hair (Fig. 36.2).
   - (e) Slide the occipital section or piece of the C-collar behind the occiput (Fig. 36.3). (Use in-line C-spine stabilization in patients with an altered level of consciousness.)
   - (f) Wrap (one-piece) or place (two-piece) the anterior section of the C-collar around the circumference of the patient's neck and snugly under the chin (Fig. 36.4).
   - (g) Once the Velcro fasteners are in place, ensure that the height is properly adjusted on the C-collar to minimize C-spine mobility.
     - (i) Most C-collars have height adjustments that utilize a locking device that requires releasing a locking mechanism by pulling out.

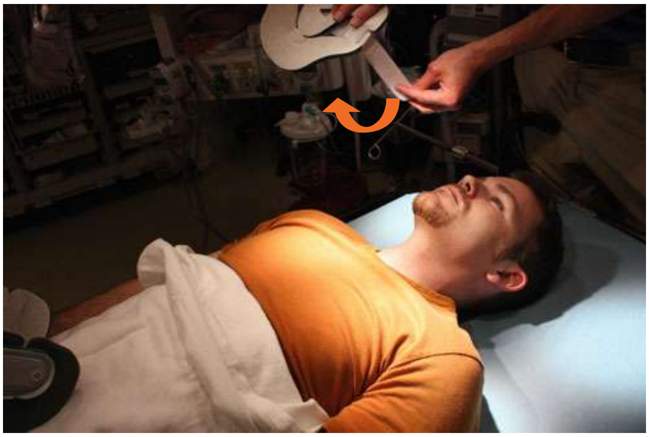

Fig. 36.1 Fold Velcro strap of C-collar posteriorly

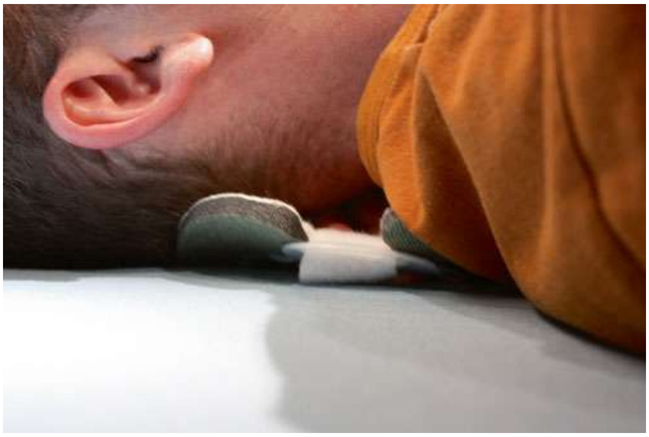

Fig. 36.2 Slide occipital section of collar behind occiput flat against stretcher

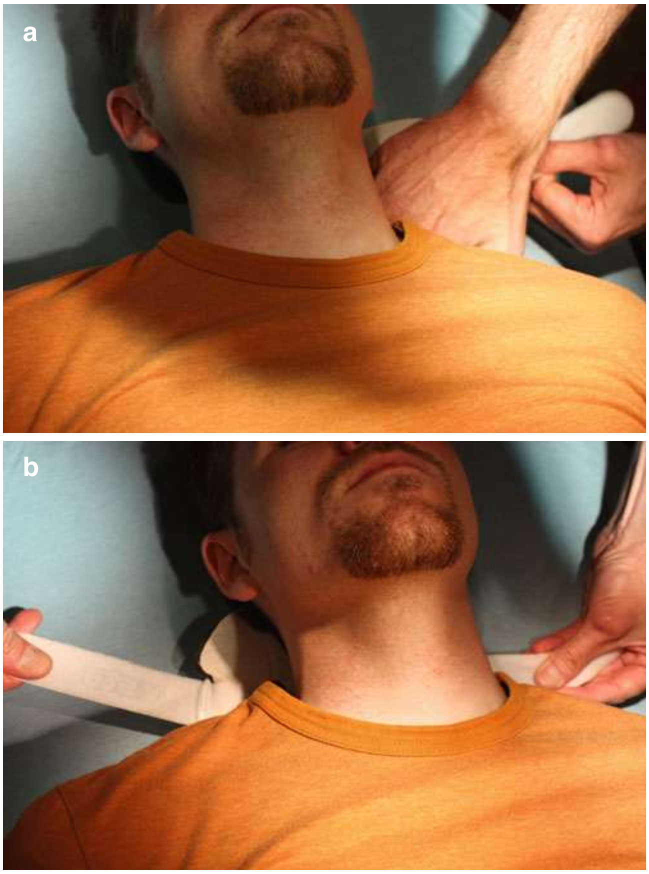

Fig. 36.3 (**a**) Use of hand to guide C-collar behind occiput; (**b**) unfold Velcro strap and position collar behind cervical spine

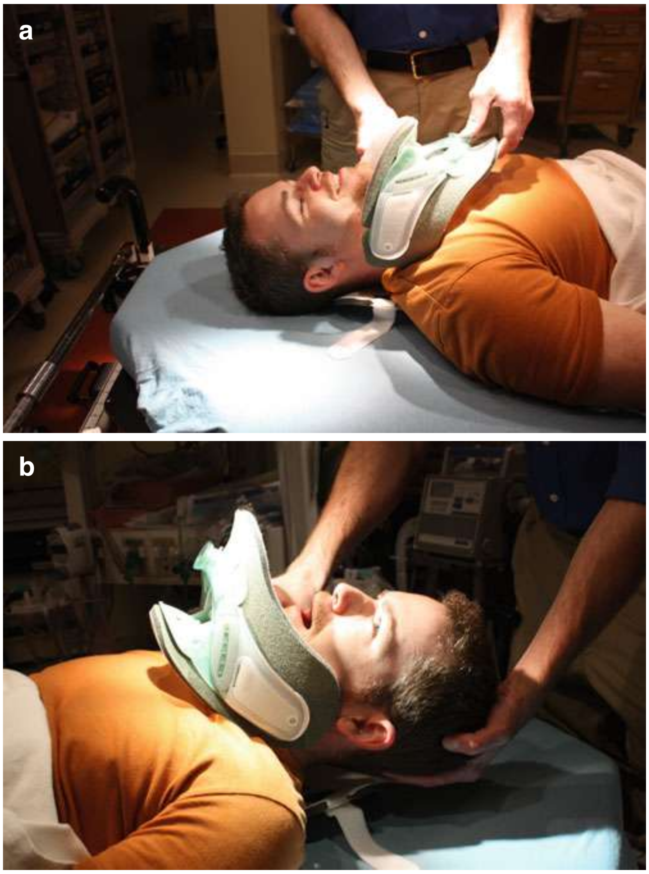

Fig. 36.4 (**a**) Correct orientation of anterior portion of C-collar; (**b**) incorrect orientation of anterior portion of C-collar may cause injury

    (ii) The Aspen collar is adjusted with a combination of pulling out while twisting a round knob at the sternal notch (Fig. 36.5).
    (iii) Pull out (away from patient) and twist for height adjustment of the Aspen two-piece collar.
    (iv) Adjust the height of the Ambu Perfit one-piece collar by simultaneously pulling (away from the patient) the two locking pins out and adjusting the height of the collar before pushing the pins back in (toward the patient) to lock the collar at the desired height (Fig. 36.6).

6. Special cases:
   - (a) Depending on the age of the patient, it may be necessary to place towels under the shoulder blades to keep the neck in a neutral position.
   - (b) Children:
     - (i) Head-to-body ratios are relatively larger than that of adults.
     - (ii) Placement on a backboard may cause significant neck flexion because the occiput rests on a flat board.
     - (iii) Towels can be placed under a child's shoulders to minimize flexion.

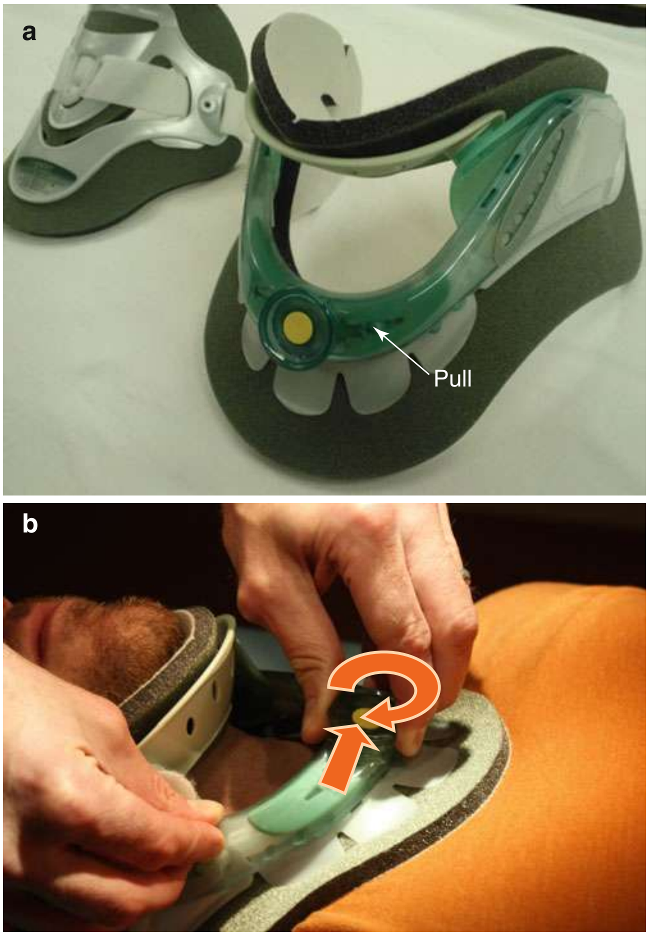

Fig. 36.5 Aspen collar is adjusted with a combination of (**a**) pulling out while (**b**) twisting a round knob at the sternal notch

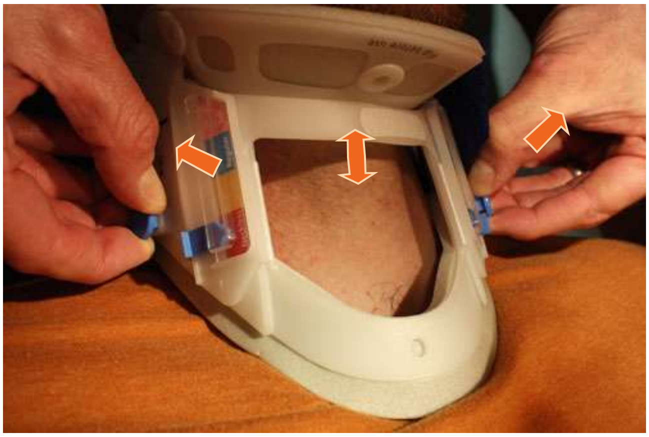

Fig. 36.6 Adjustment the height of the Ambu Perfit one-piece collar by simultaneously pulling (away from the patient) the two locking pins out and adjusting the height of the collar before pushing the pins back in (toward the patient) to lock the collar at the desired height

(c) Adults with excess soft tissue or degenerative changes that prevent C-spine straightening:
  (i) A towel or pad is placed behind the occiput to prevent hyperextension.

36.5 Pearls and Pitfalls

- Pearls
  - Airway, breathing, and circulation should be assessed before placement of the C-collar.
  - A high index of suspicion of C-spine injury is needed in intoxicated or comatose patients and the elderly.
  - Always remember to adjust the height and circumference of a C-collar for a snug fit.
- Pitfalls
  - Spinal immobilization increases the risk of aspiration in vomiting patients.
  - Failure to palpate and inspect the C-spine, head, and shoulders before C-collar placement may delay recognition of an impending airway emergency and conceal signs of critical injury needing rapid assessment and care.
  - Lateral C-spine motion is unsecured unless tape, rubber blocks, or towels are used for support.
  - Overlooking C-spine stabilization during C-collar placement leaves the C-spine at risk of further injury in patients with an altered level of consciousness.
  - C-collars that are too tight may decrease venous return from the head and may increase intracranial pressure.

Selected Reading

American College of Surgeons Committee on Trauma. Advanced trauma life support for doctors. 8th ed. Chicago: American College of Surgeons; 2004.

Hankins DG, Boggust A. Prehospital equipment and adjuncts. In: Tintinalli JE, Stapczynski JS, Cline DM, Ma OJ, Cydulka RK, Meckler GD, editors. Tintinalli's emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw Hill; 2012.

Hoffman JR, Wolfson AB, Todd K, Mower WR. Selective cervical spine radiography in blunt trauma: methodology of the National Emergency X-Radiography Utilization Study (NEXUS). Ann Emerg Med. 1998;32:461–9.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders Elsevier; 2009.

Stiell IG, Clement CM, McKnight RD, et al. The Canadian C-spine rule versus the NEXUS low-risk criteria in patients with trauma. N Engl J Med. 2003;349:2510–8.

Cervical Spine Clearance

37

Braden Hexom and Tatiana Havryliuk

37.1 Indications

- Assessment of the need for radiological imaging in trauma

37.2 Contraindications

- Known unstable cervical spine (C-spine) fracture
- Known unstable ligamentous injury
- Intoxication/altered mental status
- Presence of distracting injury

37.3 National Emergency X-Radiography Utilization Study (NEXUS) and Canadian C-Spine Rules (CCR) to Assess the Need for Imaging

- *The NEXUS clinical criteria*
  1. Tenderness at the posterior midline of the C-spine
  2. Focal neurological deficit
  3. Decreased level of alertness
  4. Evidence of intoxication
  5. Clinically apparent pain that might distract the patient from the pain of a C-spine injury
- The presence of any one of the above findings is considered to be clinical evidence that a patient is at increased risk for C-spine injury and requires radiographic evaluation.
- *CCR*

  For alert (Glasgow Coma Scale [GCS] 15) and stable trauma patients
  1. Any high-risk factor that mandates radiography? *YES → Radiography*
     - Age older than 65 years
     - Dangerous mechanism (Table 37.1)
     - Paresthesias in extremities

     ↓ *NO*
  2. Any low-risk factor that allows safe assessment of range of motion? *NO → Radiography*
     - Simple rear-end motor vehicle crash (MVC)
     - Sitting position in emergency department
     - Ambulatory at any time
     - Delayed onset of neck pain
     - Absence of midline C-spine tenderness

     ↓*YES*
  3. Able to actively rotate neck? *NO → Radiography*
     - 45° left and right

     ↓*ABLE*

     No radiography
- CCR found to have higher sensitivity (99.4 % vs. 90.7 %) and specificity (45.1 % vs. 36.8 %) and a lower rate of imaging (55.9 % vs. 66.6 %) than NEXUS criteria for C-spine clearance of low-risk alert trauma patients.

Table 37.1 Dangerous Mechanism

Fall from elevation ≥3 ft/5 stairs
Axial load to head (e.g., diving)
Motor vehicle crash at high speed (>100 km/h), rollover, ejection
Motorized recreational vehicles
Bicycle crash

B. Hexom, MD (✉)
Department of Emergency Medicine, Mount Sinai Hospital,
New York, NY, USA
e-mail: braden.hexom@mssm.edu

T. Havryliuk, MD
Department of Emergency Medicine, University of Colorado
Denver, Denver, CO, USA
e-mail: tatiana.havryliuk@ucdenver.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_37

37.4 Type of Imaging

- Bones
  - C-spine computed tomography (CT)—the new gold standard.
  - Plain X-rays—less sensitive than CT; in one study X-rays missed 45 % of injuries that were picked up by CT [1].
- Ligaments
  - C-spine magnetic resonance imaging (MRI)—consider for possible ligamentous injuries and for further evaluation of obtunded patients.
  - Flexion-extension X-rays—less sensitive than MRI for detection of ligamentous injuries in the acute phase. Delayed flexion-extension films are more sensitive for ligamentous injury than those done the day of injury, but less sensitive than MRI.

37.5 Materials

- Cervical collar (C-collar) (Fig. 37.1)

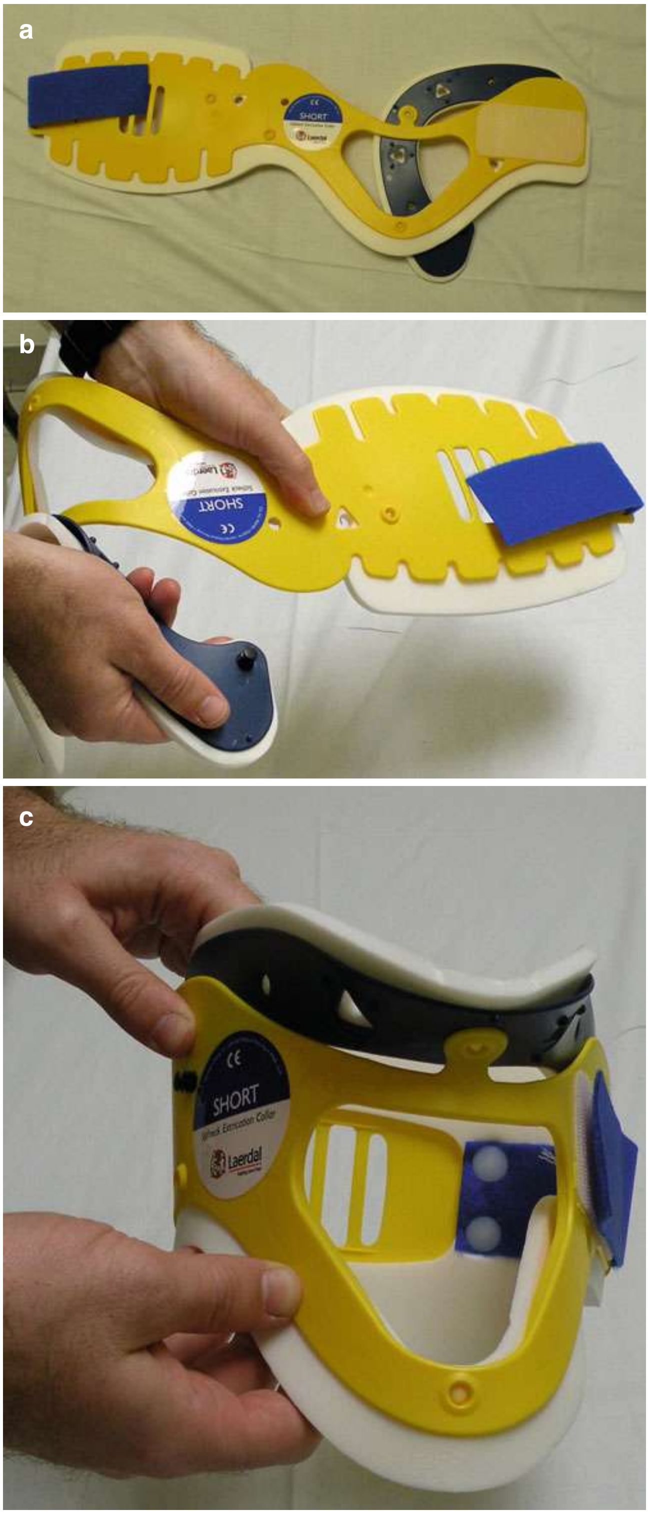

Fig. 37.1 Assembly of C-collar: Select appropriate size (**a**). Invert chinstrap and snap into place (**b**). Assembled C-collar (**c**)

37.6 Procedure (Fig. 37.2)

1. Apply C-collar to stabilize spine. Select appropriate size.
2. Perform a brief neurological examination and assess the patient's mental status; proceed only if both are normal.
3. Loosen the collar to palpate the midline while the patient holds his or her head still. If spinal tenderness exists replace the C-collar and proceed to imaging (Fig. 37.3).
4. Instruct the patient to rotate the neck 45° to each side and flex the neck. If the patient is pain free and with no neurological comprise, the C-collar may be removed and no imaging of the C-spine is required (Fig. 37.4).

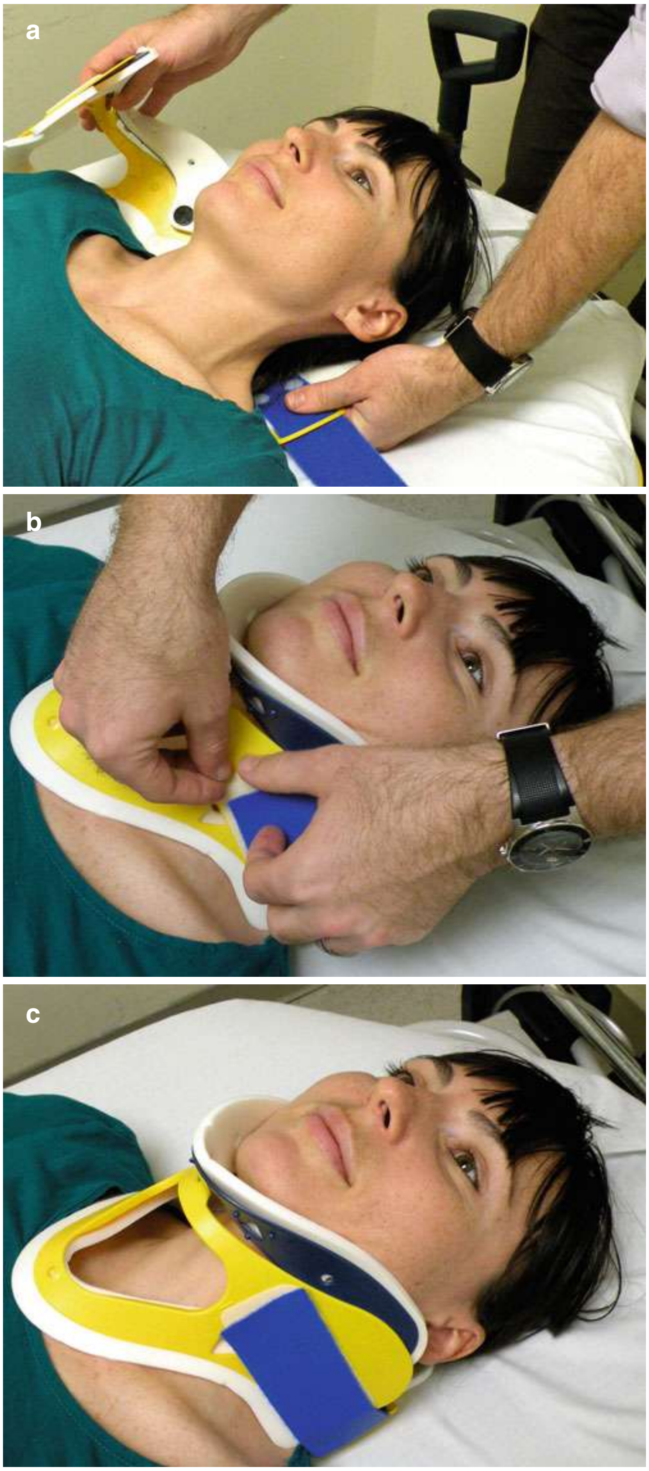

Fig. 37.2 Proper placement of C-collar: Slide C-collar under the neck while keeping neck immobilized (**a**). Secure the collar (**b**). Correct C-collar placement (**c**)

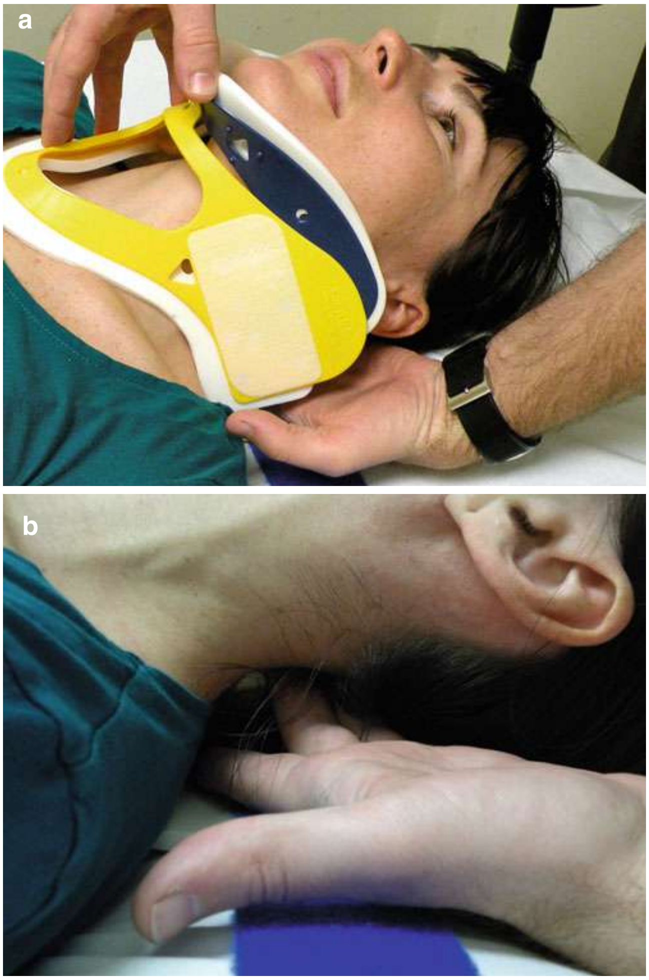

Fig. 37.3 Assessment of C spine tenderness: Loosen C-collar while keeping neck immobilized (**a**). Palpate midline of C spine (**b**)

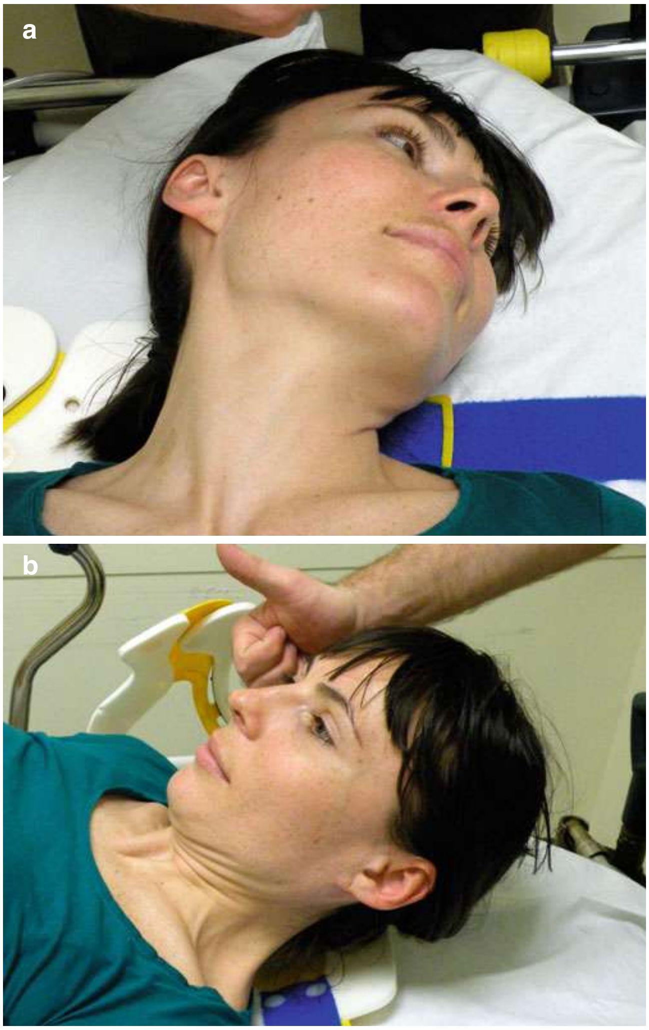

Fig. 37.4 C-spine clearance: Instruct patient to rotate neck 45° each way (**a**). Instruct patient to flex neck (**b**)

37.7 Complications

- Missing clinically important C-spine injury
- Pressure ulcer from prolonged use of C-collar

37.8 Pearls and Pitfalls

- *Pearls*
  - Select appropriate size of C-collar.
  - Ensure adequate mental status because further imaging might be required in obtunded/intoxicated patients.
- *Pitfalls*
  - Avoid prolonged use of C-collar because this can lead to a pressure ulcer.

Reference

1. Schenarts PJ, Diaz J, Kaiser C, Carrillo Y, Eddy V, Morris Jr JA. Prospective comparison of admission computed tomographic scan and plain films of the upper cervical spine in trauma patients with altered mental status. J Trauma. 2001;51:663–8; discussion 668–9.

Selected Reading

Hoffman JR, Schriger DL, Mower W, Luo JS, Zucker M. Low-risk criteria for cervical-spine radiography in blunt trauma: a prospective study. Ann Emerg Med. 1992;21:1454–60.

Platzer P, Jaindl M, Thalhammer G, et al. Clearing the cervical spine in critically injured patients: a comprehensive C-spine protocol to avoid unnecessary delays in diagnosis. Eur Spine J. 2006;15:1801–10.

Stiell IG, Clement CM, McKnight RG, et al. The Canadian C-spine rule versus the NEXUS low-risk criteria in patients with trauma. N Engl J Med. 2003;349:2510–8.

Stiell IG, Wells GA, Vandemheen K, et al. The Canadian C-spine rule for radiography in alert and stable trauma patients. JAMA. 2001; 286:1841–8.

Log Roll

38

Justin Bennett and Lars K. Beattie

38.1 Indications

- Any patient arriving on a rigid backboard in the emergency department
- Assessment of posterior traumatic injuries
- Performed as early as possible after arrival to prevent skin breakdown from pressure ulcers

38.2 Contraindications

- Improperly fitted cervical collar (C-collar) or unsecured cervical spine (C-spine) before log roll
- Unsecured endotracheal tube before log-rolling intubated patients

38.3 Materials and Medications

- Personnel: three or four people
  - One to stabilize the C-spine
  - One or two to roll the patient
  - One to palpate the length of the spine
- Properly fitting hard C-collar
- Trauma shears for removing transport straps

38.4 Procedure

1. Ensure that the airway, breathing, and circulation (ABCs) are established before attempting to remove the patient from the backboard.
2. Gather personnel, at least three, but preferably four people. Larger patients will often require additional assistance.
3. For intubated patients an additional person will be needed to secure the endotracheal tube during the log roll.
4. Ensure the height is properly adjusted on the C-collar to minimize C-spine mobility (Fig. 38.1).
5. Position the stretcher at an ergonomic lifting position for the person responsible for stabilizing the C-spine.
6. Clothing and transport straps should be removed before the log roll to improve visualization of injuries.
7. The patient is asked to cross his or her hands over the chest.
8. C-spine stabilization (at head of bed)
   - Grasp the patient's trapezoids at the midclavicular line.
   - Secure the sides of the patient's head between the operator's forearms.
   - Stabilize the patient's head in neutral anatomical position relative to the body using the operator's forearms (Fig. 38.2).

J. Bennett, MD
Department of Emergency Medicine, Wake Forest University
School of Medicine, Winston-Salem, NC, USA
e-mail: jubennet@wakehealth.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_38

9. Thoracic and lumbar spine stabilization
   - One or two (preferably two) people should stand next to the patient on the side to which the patient will be rolled.
     - One person: Place hands over the patient's shoulder and hip.
     - Two people
       - First person places hands on the patient's shoulder and hip.
       - Second person places hands on the patient's hip and knees.
   - The decision to roll to the left or right side is determined by injury sites, to minimize injury exacerbation and pain, and to minimize risk of endotracheal tube dislodgment.

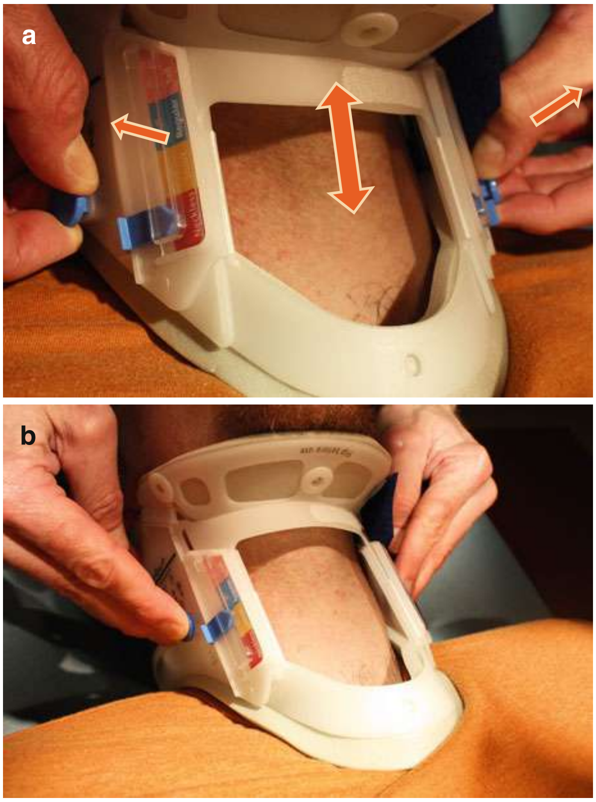

Fig. 38.1 (**a**) The height of the Ambu Perfit one-piece collar is adjusted by simultaneously pulling the two locking pins out (away from the patient) and adjusting the height of the collar, (**b**) then pushing the pins back in (toward the patient) to lock the collar at the desired height

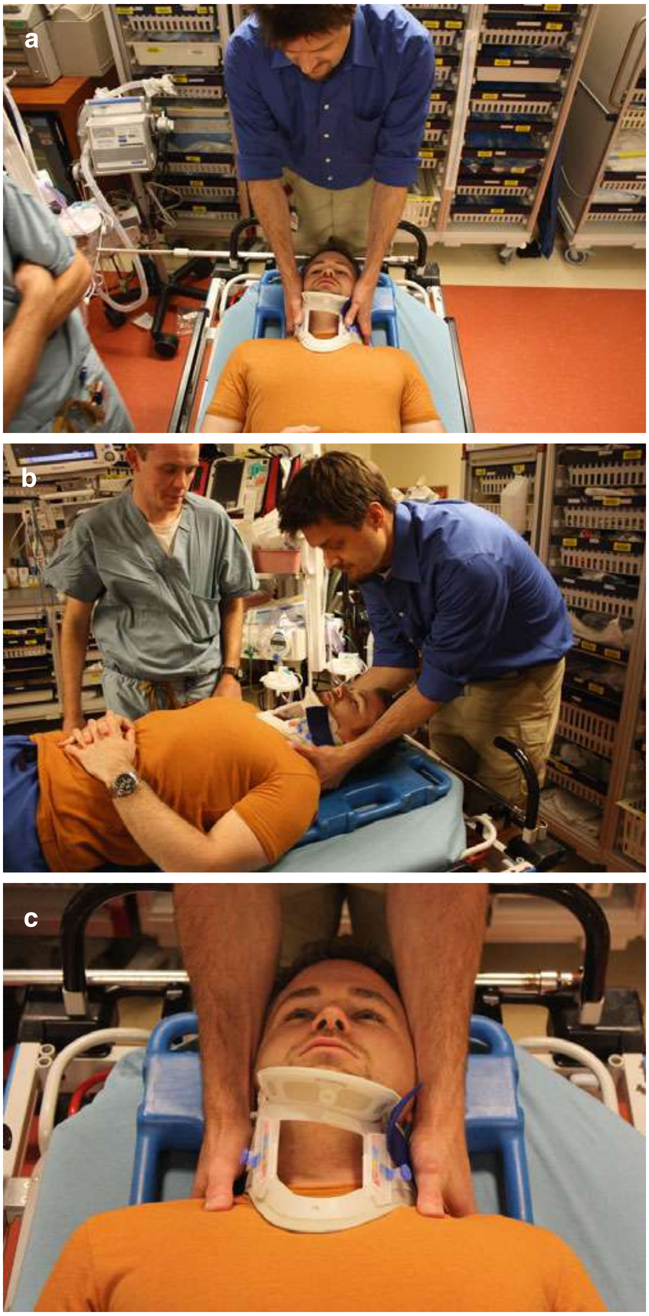

Fig. 38.2 (**a**) Incorrect: Single-point stabilization does not keep head and C-spine in line with torso (**b**, **c**). Correct: The two-point stabilization technique keeps the head and C-spine in line with the torso during the log roll

10. The person at the head of the bed signals the initiation of a synchronized log roll when everyone is in position.
11. Attention should be directed at maintaining a neutral axis along the anatomical line of symmetry during the roll (Fig. 38.3).
12. While the patient is on her or his side, the rigid backboard should be removed and secured to prevent injury of caretakers.
13. The patient's entire posterior should be exposed and examined, taking care to note lacerations and obvious deformities.
14. The entire length of the spine is examined by inspection and then by palpation to assessed for tenderness, step-offs, and deformities (Fig. 38.4).
15. Before returning the patient back onto the stretcher, ensure that any debris, glass, lumps of clothing, or blankets are removed.

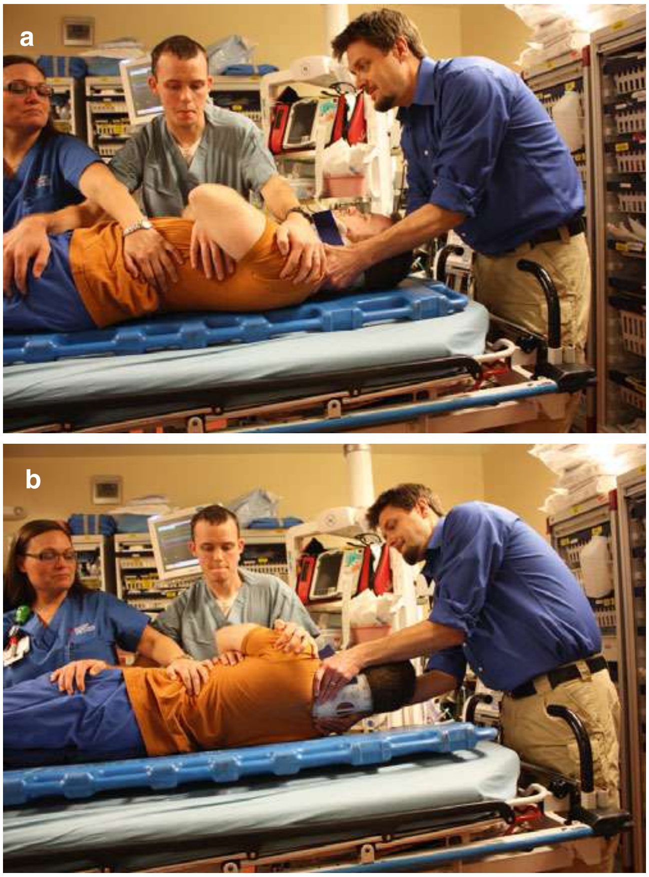

Fig. 38.3 (**a**) The interlocking hands of operators performing lateral log roll help maintain spine immobilization and minimize twisting. (**b**) Two-point cervical spine stabilization is maintained in neutral axis relative to the anatomical line of symmetry during the roll. Appropriate bed height adjustment will increase the ease of the procedure

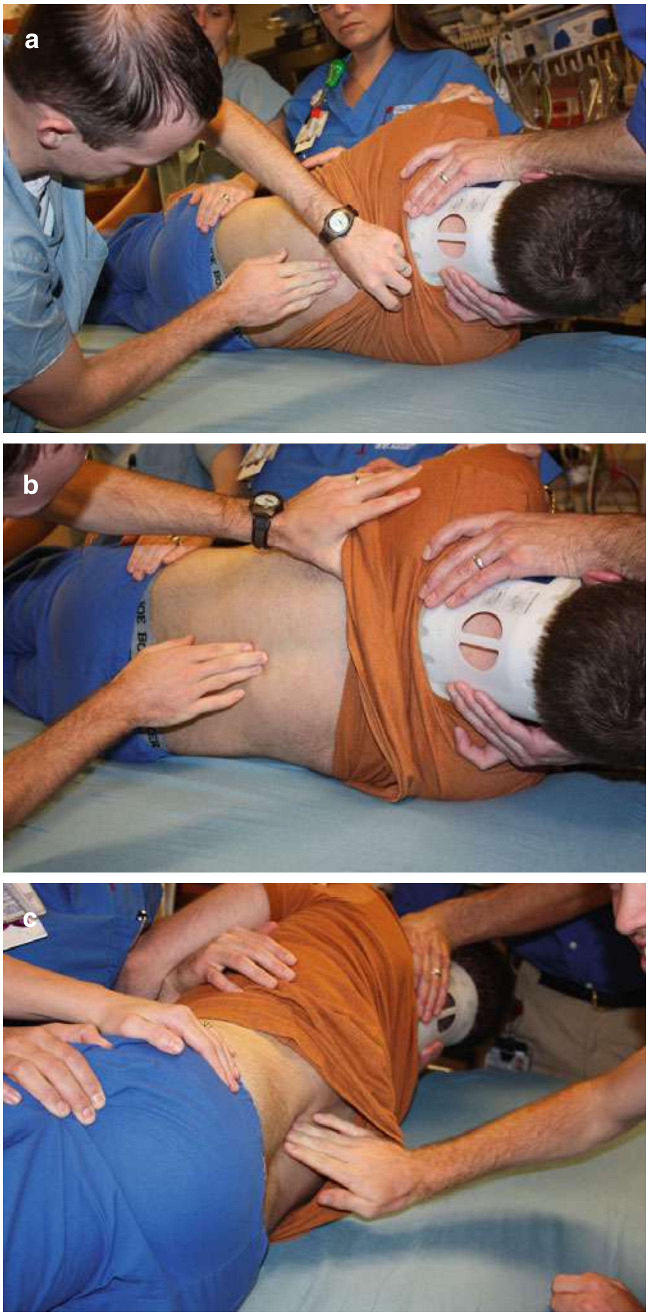

Fig. 38.4 (**a**–**c**) After backboard has been removed, the entire length of the of the spine can be appropriately palpated and assessed for injury to the thoracic and lumbar spine

38.5 Pearls and Pitfalls

- Pearls
  - The log roll should be performed in unison to avoid segmental rotation.
  - The ABCs should be established before initiating the log roll, which is part of the secondary survey in the trauma evaluation.
  - A proper log roll begins with stabilization of the C-spine because it is the most freely mobile part of the spine and, therefore, most frequently injured.
  - Take the patient off the backboard as soon as possible to prevent formation of pressure sores.
- Pitfalls
  - Failure to establish a two-point stabilization between the patient's body and the head leaves the C-spine at risk of further injury during the log roll.
  - Patients should not be left on the backboard for spinal precautions.
  - Failure to ensure that the ABCs are established before attempting a log roll.

Selected Reading

American College of Surgeons Committee on Trauma. Advanced trauma life support for doctors. 8th ed. Chicago: American College of Surgeons; 2008. ISBN 1880696312.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders Elsevier; 2009.

Part VI

Neurologic and Neurosurgical Procedures

Burr Hole Craniotomy

39

Latha Ganti

39.1 Indications

- Extradural hematoma (EDH) or subdural hematoma (SDH) with Glasgow Coma Scale (GCS) <8 and unavailability of timely neurosurgical intervention in the face of increased intracranial pressure
- Signs of increased intracranial pressure [1]
  - Deteriorating level of consciousness
  - Slowing of the pulse
  - Dilating pupils
  - Focal seizures
  - Hemiparesis
  - Extensor posturing of the limbs

39.2 Contraindications

- Absolute
  - GCS >8
  - Immediate availability of neurosurgeon
- Relative
  - Lack of imaging (in this case, decision is guided by neurologic findings and signs of increased intracranial pressure or impending herniation)

39.3 Materials and Medications

- Razor to shave area
- 2 % lidocaine with epinephrine to numb scalp
- 10 % povidone-iodine or chlorhexidine prep
- Light, suction, cautery, dressing tray
- Gelfoam
- Penrose drain
- 3-0 silk
- #10 scalpel blade and #3 handle
- Small self-retractors or rakes
- Drill and drill bits: can be manual (Fig. 39.1) or automatic stopping variety (Fig. 39.2)
- Bone wax or electric cautery apparatus
- Suction apparatus
- Saline irrigation (IV tubing connected to a saline bag with clamp set so flow is low, or a saline syringe flush)

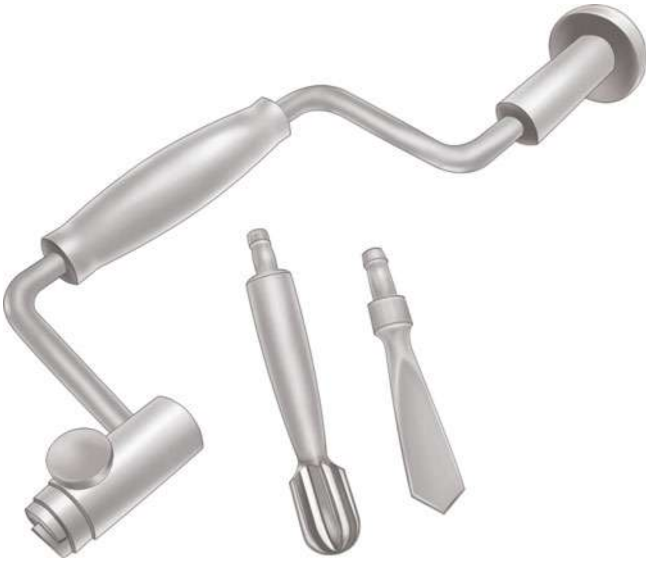

Fig. 39.1 Hudson brace, burr, and perforator

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida, Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_39

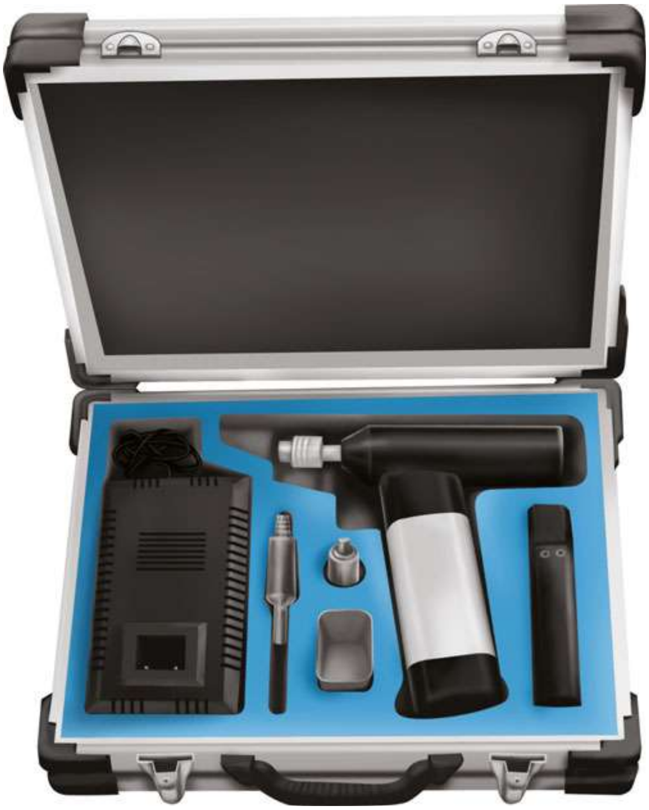

Fig. 39.2 Automatic stopping craniotomy drill set

39.4 Procedure

Preparation

1. Patient should be supine, intubated, with appropriate C-spine precautions
   - Hypertonic saline and/or mannitol can be considered for medical management of increased ICP if/as directed by neurosurgery
2. Determine location and depth for burr hole placement (Figs. 39.3 and 39.4):
   - Have CT images immediately available for viewing
   - Most common location for EDH or SDH is temporal
     - Temporal burr hole placement: 2 fingerbreadths above the ear, 2 fingerbreadths forward (Fig. 39.5)
     - Parietal burr hole placement: 2 fingerbreadths above the ear, 3 fingerbreadths backward (behind the ear)
     - Frontal burr hole placement: 3 fingerbreadths from midline, 3 fingerbreadths above hairline
   - Estimate depth of hematoma by counting the number of slices the hematoma is as seen on CT scan and multiplying by the slice thickness [2]
3. Shave hair around area of hematoma.
4. Prep skin with betadine and chlorhexidine.
5. Anesthetize scalp skin with lidocaine and epinephrine.

   Accessing hematoma
6. Make a vertical incision approximately 4 cm long down to bone.s
7. Use periosteal elevator or end of scalpel blade to scrape muscle and periosteum away from bone.
8. Place self-retaining retractors (or rakes) to keep field open (Fig. 39.6).
9. Using a drill with a clutch mechanism [2], drill through outer table (resistance), diploic space (no resistance), then inner table.
   - Burr hole needs to be placed over the *center* of the hematoma (Fig. 39.7).
10. Control bone bleeding with bone wax; control bleeding from veins and/or muscle with gelfoam sponge or cautery (diathermy).
    - Wound edges may also be cauterized or tamponaded with manual pressure +/– epinephrine gauze.
11. Once in the inner table, separate dura from bone
12. Enlarge opening by switching to a conical or cylindrical burr or use a rongeur

    Evacuating hematoma
13. EDH blood will be visible at this point and should be gently suctioned out.
14. SDH blood will be seen as a tenting of the dura and may be clotted: lift the dura with a hook or make an incision with a fresh scalpel through it to expose the clot and drain.

15. Irrigate with saline (this can be via a hand syringe or via IV tubing connected to a saline bag at low sped flow).
16. Repeat gentle suction.

Closure

17. For temporal burr holes, ligate middle meningeal artery (if visible) or cauterize.
18. Place Penrose drain (Fig. 39.8) and close the dura first with a 3-0 absorbable suture, ensuring there is no dural leak (will lead to infection if present).
    - A watertight seal of the duraplasty is essential to minimize cerebrospinal fluid leakage.
19. Loosely suture scalp using 3-0 silk.

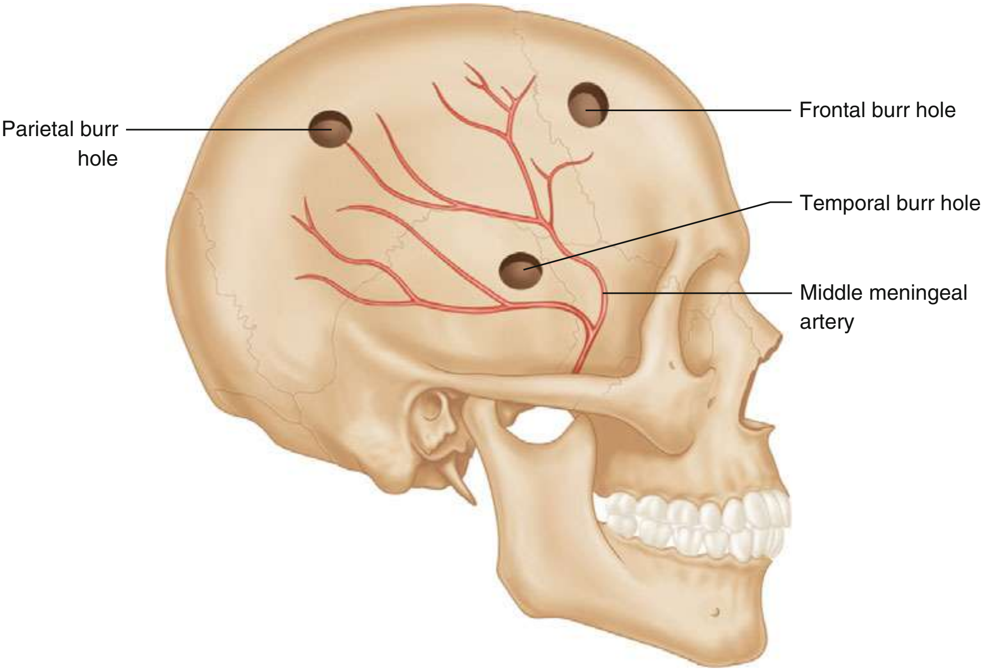
Parietal burr hole
Frontal burr hole
Temporal burr hole
Middle meningeal artery

Fig. 39.3 Positions of burr hole placement

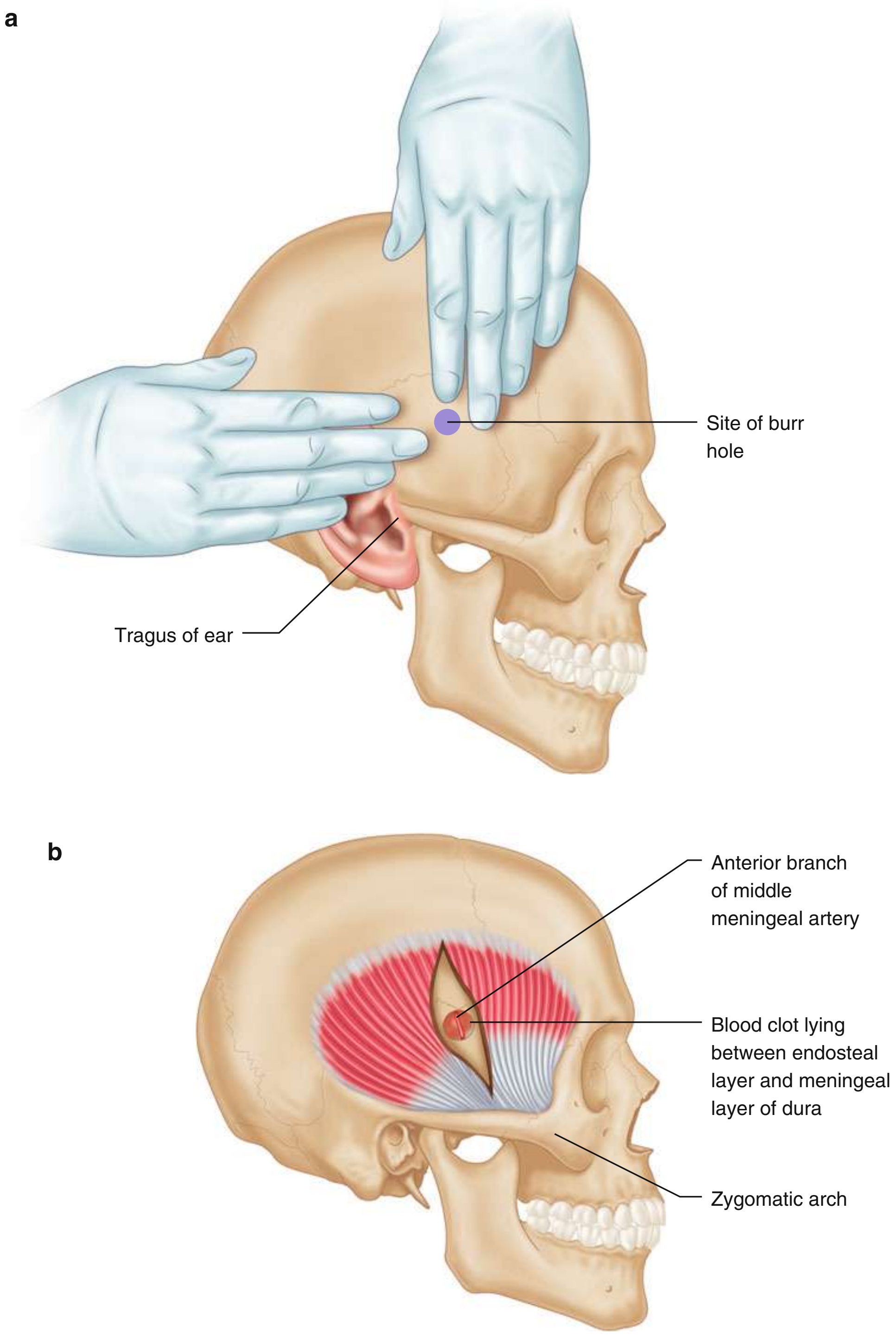
a
Site of burr hole
Tragus of ear
b
Anterior branch of middle meningeal artery
Blood clot lying between endosteal layer and meningeal layer of dura
Zygomatic arch

Fig. 39.4 (**a**, **b**) Anatomy for temporal burr hole placement (most common location)

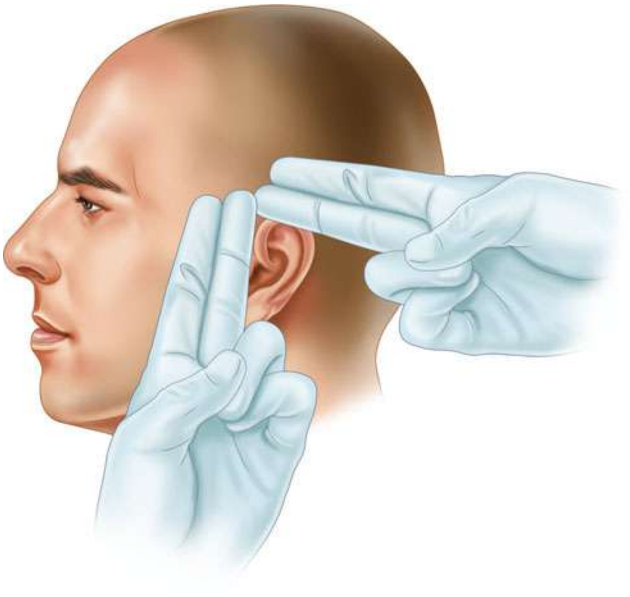

Fig. 39.5 Landmark for temporal burr hole above zygoma, make 5 cm incision over site

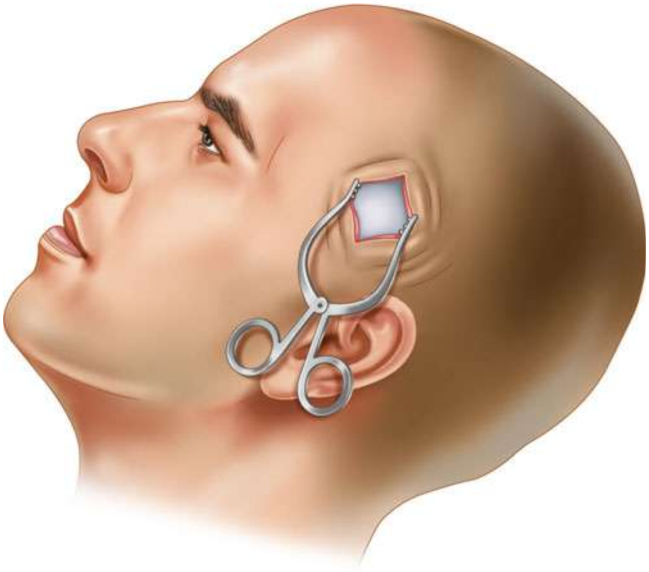

Fig. 39.6 After incision is made, use retractors (preferably self-retaining) to visualize field

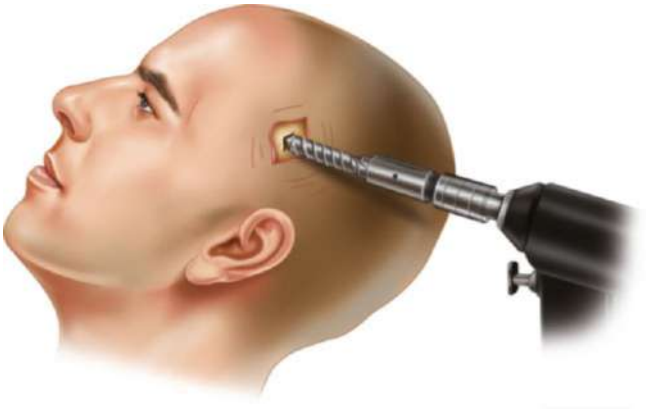

Fig. 39.7 Use penetrator drill to get through outer table. Follow up with a burr to get through the rest of the layers. Can also use an automatic stop craniotomy drill

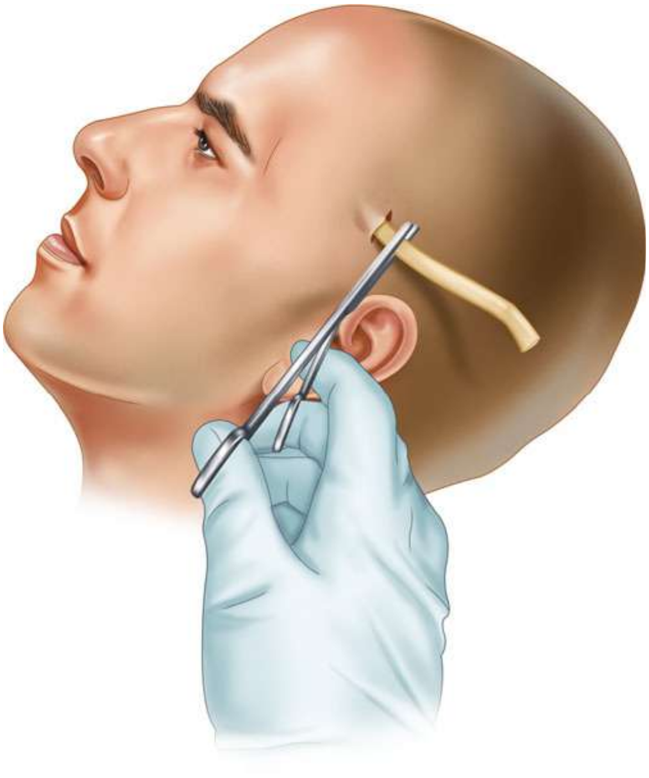

Fig. 39.8 Suture in Penrose drain and close the wound

39.5 Complications

- Brain laceration/perforation
- Temporal artery laceration
- Wrong location (minimized when viewing images during procedure)
- Facial artery laceration
- Meningitis
- Brain abscess

39.6 Pearls and Pitfalls

- Pearls
  - If brain herniates through the burr hole, there is likely a hematoma at another location.
  - Remember this is only a lifesaving procedure that is to be done when timely neurosurgical intervention is not possible. *Do not delay transfer* of the patient
- Pitfalls
  - Bone in temporal area is quite thin; if not using an automatic stop drill, be very careful not to go too far and perforate the brain.

References

1. Wilkinson DA, Skinner MF. The primary trauma care manual for trauma management in district and remote locations. World Health Organization (WHO). http://www.steinergraphics.com/surgical/006_17.6.html. Accessed 22 May 2014.
2. Wilson MH, Wise D, Davies G, Lockey D. Emergency burr holes: "how to do it.". Scand J Trauma Resusc Emerg Med. 2012;20:24. doi:10.1186/1757-7241-20-24.

Selected Reading

http://www.viewmedica.com/vm/pages/library/L_df8516c9#vm_A_ac54d3a1. Accessed 29 June 2014.

External Ventricular Drain Placement

40

Latha Ganti

An external ventricular drain (EVD; also known as a ventriculostomy tube) is placed via a burr hole in one of the ventricles of the brain (Fig. 40.1) to drain excess cerebrospinal fluid (CSF) that causes elevated intracranial pressure.

40.1 Indications

- Emergent need for intracranial pressure (ICP) monitoring and or management
- Enlarged ventricles on neuroimaging with Glasgow coma scale <12
- Subarachnoid hemorrhage with Hunt-Hess grade ≥3 (Table 40.1)
- Coma
- Obstructive hydrocephalus
- Intraventricular hemorrhage
- Signs of increased intracranial pressure

40.2 Contraindications

- Absolute
  - Immediate availability of a neurosurgeon who can do the procedure
- Relative
  - Coagulopathy
  - Scalp infection

40.3 Materials and Medications

- Intubation equipment and medications for sedation as needed
- Sterile gloves, gown, mask
- Ruler
- Surgical marking pen
- 1–2 % lidocaine with epinephrine to numb scalp
- 5 cc syringe and needles to give anesthetic
- 10 % povidone-iodine solution and swabs
- Razor to shave area
- Fenestrated clear drape
- Scalpel #11 blade (for scalp) and #15 blade (for periosteum) with #3 handle (Fig. 40.2a)
- 4×4 sterile gauze
- Adson forceps
- Mosquito forceps
- Self-retaining eyebrow retractor (Fig. 40.2b)
- 3-0 nylon suture and needle holder and/or skin stapler
- Scissors
- A hand drill with variable chuck
- One or more drill bits with depth guards in 5/32″ (3.97 mm), 13/64″ (5.31 mm), and 1/4″ (6.35 mm) sizes
- A hex wrench for depth guard adjustment
- Ventricular catheter
- Primed ventricular drainage collection system

40.4 Procedure

Patient preparation

1. Patient is intubated and placed supine in neutral position with head of the bed elevated 30–45°.
2. Administer one dose of intravenous antibiotics that covers skin flora.
3. Make precise measurement of where the hole and incision will be made. Most commonly, EVDs are placed in

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida, Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_40

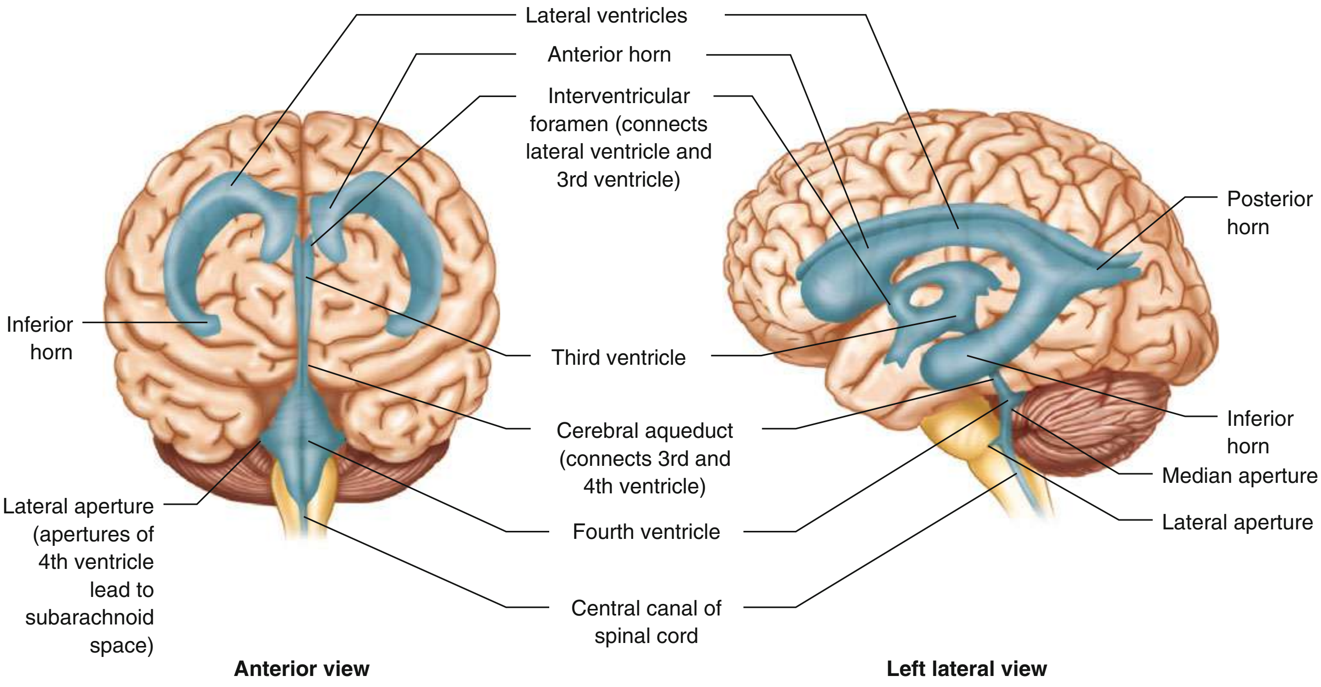
Lateral ventricles
Anterior horn
Interventricular foramen (connects lateral ventricle and 3rd ventricle)
Posterior horn
Inferior horn
Third ventricle
Cerebral aqueduct (connects 3rd and 4th ventricle)
Inferior horn
Median aperture
Lateral aperture (apertures of 4th ventricle lead to subarachnoid space)
Fourth ventricle
Lateral aperture
Central canal of spinal cord
Anterior view
Left lateral view

Fig. 40.1 Ventricles of the brain that contain CSF

Table 40.1 Hunt and Hess scale for subarachnoid hemorrhage [1]

Grade	Signs and symptoms
1	Alert and oriented, mild headache, slight or no nuchal rigidity
2	Alert and oriented, moderate to severe headache, nuchal rigidity, no neurologic deficit
3	Drowsiness, confusion, mild focal neurologic deficit
4	Stupor, moderate-severe hemiparesis
5	Coma, decerebrate posturing

the right frontal scalp, as this is the nondominant hemisphere in >95 % of the population.

4. Using a sterile skin/surgical marker, draw a line 11 cm back from nasion and then a point 3 cm to the right of that, which corresponds to the mid-pupillary line. This is called Kocher's point (Fig. 40.3).
5. Once Kocher's point is marked, shave skin and remove loose hairs so field is wide enough for subcutaneous catheter placement and tunneling.
6. Prep area with povidone-iodine.
7. Place sterile clear drape over field.
8. Prep area of incision once more.
9. Infiltrate scalp with 1–2 % lidocaine with epinephrine (the epinephrine acts as a hemostatic agent, keeping surgical field clean) (Fig. 40.4).
10. Make a 2 cm linear stab incision and extend incision to skull (Fig. 40.5).
11. Use eyebrow self-retaining retractors to hold skin edges back.
12. Drill burr hole with hand drill (Fig. 40.6).
13. Irrigate burr hole with sterile saline.
14. Put on a new pair of sterile gloves (prior to handling ventriculostomy catheter).
15. Remove ventricular tubing from sterile package. It has markings on it that are 1 cm apart.
16. Insert ventriculostomy catheter perpendicular to the skull at the point of insertion. Catheter is aimed at the ipsilateral medial canthus of the eye (anteroposterior plane) and tragus of ear (lateral plane).
17. Advance ventriculostomy tube 5–6 cm with stylet from outer skull table so it sits in the anterior horn of the lateral ventricle (Fig. 40.7).
18. Ensure ventriculostomy catheter is draining CSF.
19. Attach metal trocar to tip of ventriculostomy catheter and tunnel the trocar and catheter under the galea approximately 3–5 cm to the right of the original incision. Bring trocar out through separate stab incision in scalp.
20. Remove trocar, make sure ventriculostomy catheter is still draining CSF.
21. Place temporary cap on ventriculostomy catheter to prevent overdrainage of CSF.

22. Close the original incision with either sutures or staples.
23. Secure catheter to the scalp using staples.

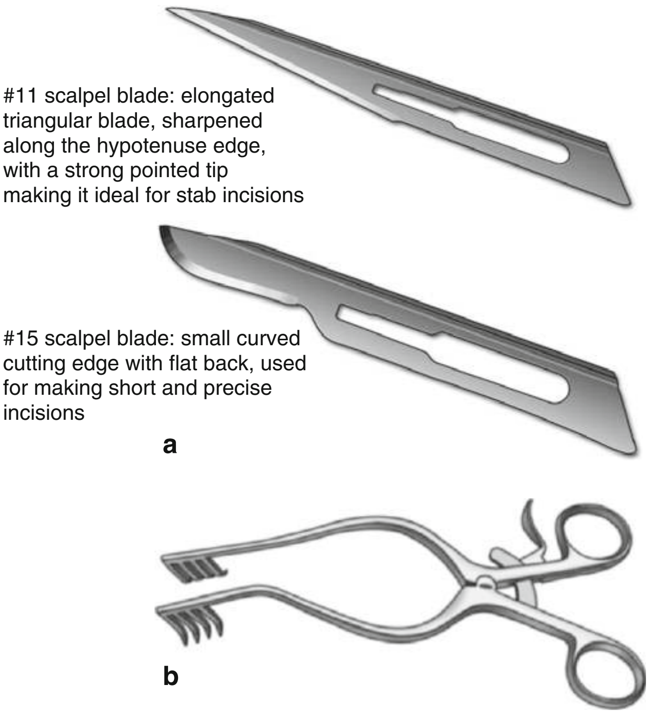
#11 scalpel blade: elongated triangular blade, sharpened along the hypotenuse edge, with a strong pointed tip making it ideal for stab incisions
#15 scalpel blade: small curved cutting edge with flat back, used for making short and precise incisions
a
b

Fig. 40.2 (**a**) Surgical scalpel blades. (**b**) Self-retaining eyebrow retractors

24. Cover incision with sterile transparent dressing.
25. Suture stopcock to ventricular catheter to ensure it does not come off (take care not to make suture too tight or it will occlude ventriculostomy catheter).
26. Remove blind end of stopcock, attach ventricular catheter to ventricular drain system once it is primed.

Priming, setup, and maintenance of ventricular drainage system

27. Prepare drainage system by priming system with sterile, preservative-free saline.
28. Place zero point of drainage system at midbrain (level of patient's ear tragus) or above midbrain at 15–20 cm of H$_2$0 (target ICP set by neurointensivist and will change depending on patient's situation) (Fig. 40.8).
29. Attach drainage tubing.
30. Patients who require an EVD should be closely monitored by nurses trained and competent in assessment and management of the drain and in recognizing signs of increased ICP in the patient.
31. Assessment of the drainage system should be done a minimum of every 4 h, which includes inspecting the EVD from the insertion site along the entire drainage system, checking for cracks in the system or fluid leaking from the insertion site [2].

Post-procedure

32. Obtain CT scan of brain to verify placement (Fig. 40.9).

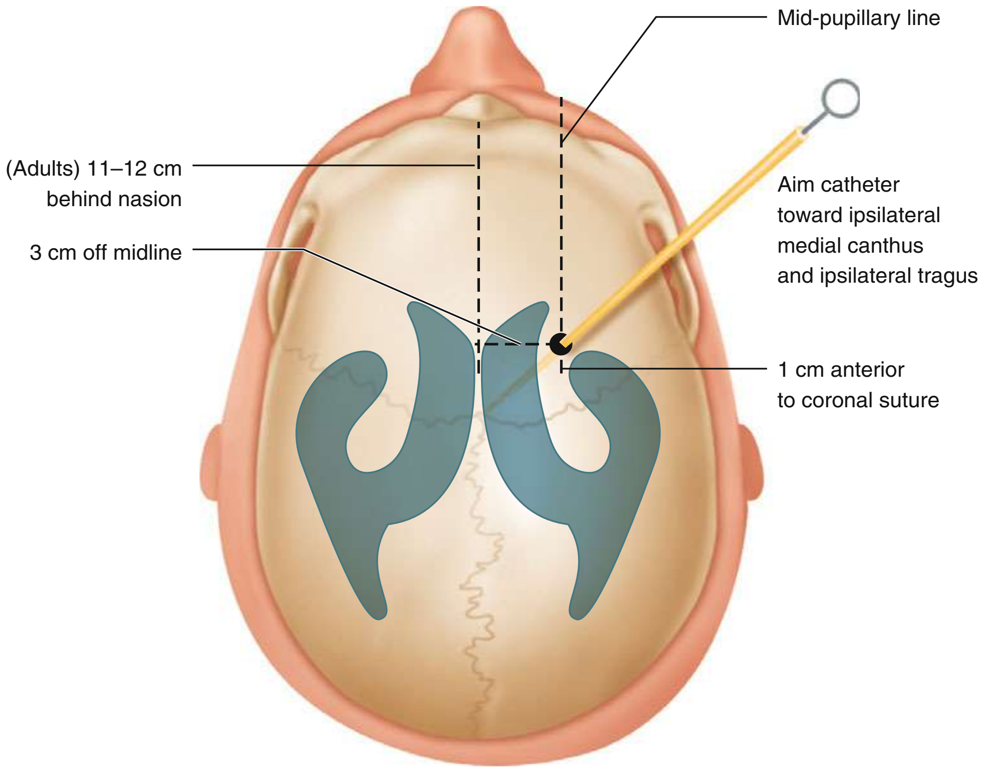
Mid-pupillary line
(Adults) 11–12 cm behind nasion
3 cm off midline
Aim catheter toward ipsilateral medial canthus and ipsilateral tragus
1 cm anterior to coronal suture

Fig. 40.3 Kocher's point for frontal EVD placement

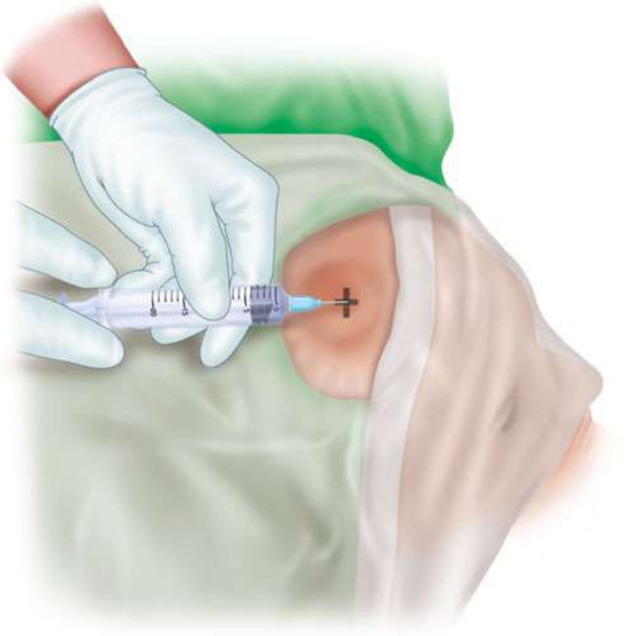

Fig. 40.4 Anesthetize skin

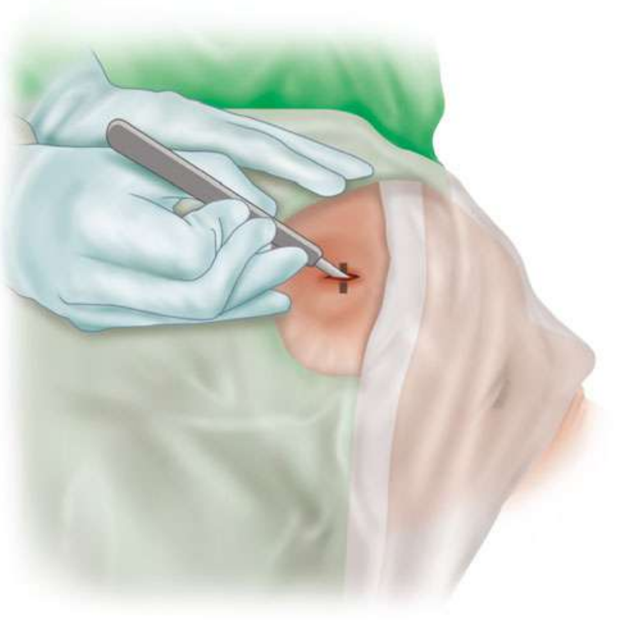

Fig. 40.5 Make stab incision

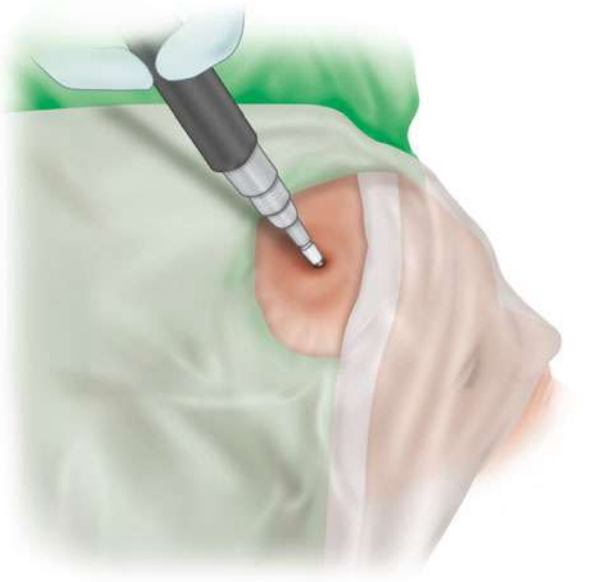

Fig. 40.6 Make burr hole

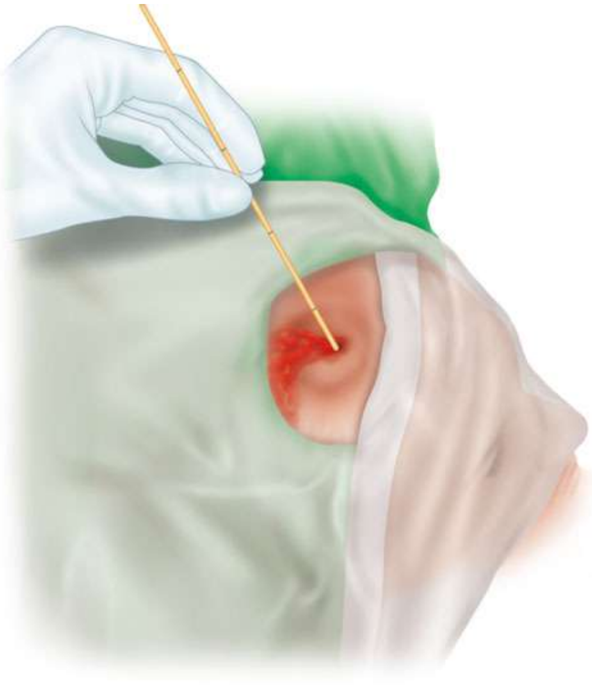

Fig. 40.7 Advance ventriculostomy catheter with stylet

40.5 Complications

- Hemorrhage
- Infection: meningitis, ventriculitis
- Tract hematoma
- Migration/dislodgement

40.6 Pearls and Pitfalls

- Pearls
  - EVD needs to be re-leveled every time the patient moves. Family and all visitors must be informed that any patient movement or change in elevation of head

Fig. 40.8 Zero point of drainage system placed at midbrain (level of patient's ear tragus)

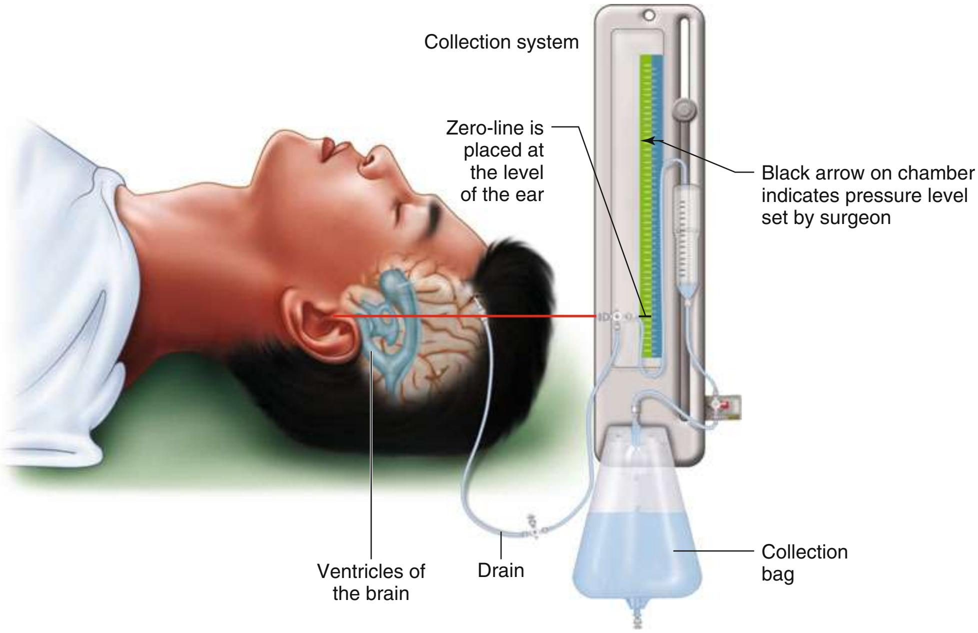
Collection system
Zero-line is placed at the level of the ear
Black arrow on chamber indicates pressure level set by surgeon
Ventricles of the brain
Drain
Collection bag

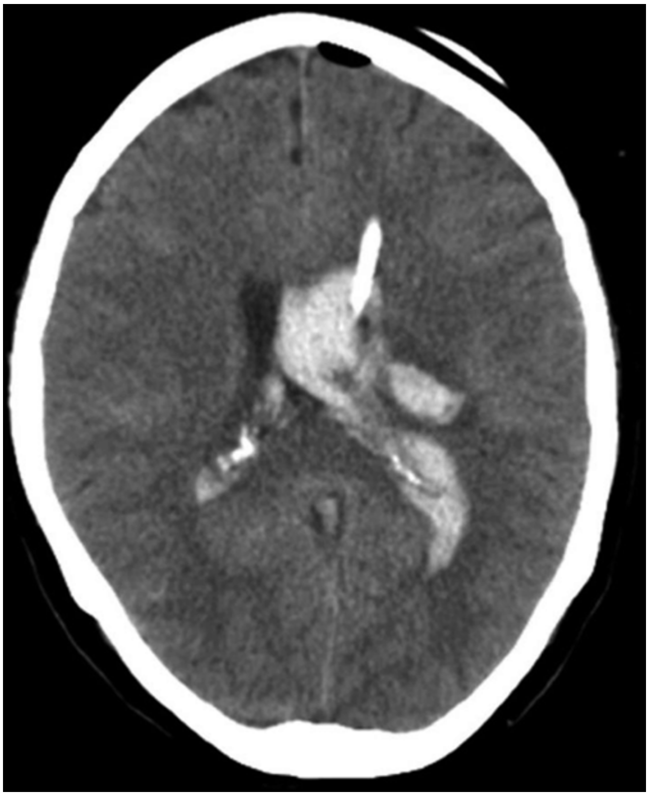

Fig. 40.9 CT scan demonstrating typical placement of ventriculostomy catheter in ipsilateral anterior horn of lateral ventricle [3] (Reproduced with permission from National Institutes of Health: Jaffe et al. [3])

of bed will require re-leveling of EVD in order to ensure it is a the appropriate level for CSF drainage.
  - CSF collection chamber should remain upright to prevent reflux/leakage.
  - EVD placement in the parieto-occipital scalp or frontotemporal or trans-sylvian locations is generally not done at the bedside due to higher rate of complications. This is performed in the operating room.
- Pitfalls
  - If the drain is leveled too low: CSF will drain too easily, which can lead to re-rupture of an unrepaired ruptured cerebral aneurysm or cause a subdural hematoma due to shearing of bridging veins.
  - If the drain is leveled too high, CSF will not drain or will not drain at desired rate, leading to hydrocephalus. The ICP will need to be higher in order for the CSF to drain.

References

1. Hunt WE, Hess RM. Surgical risk as related to time of intervention in the repair of intracranial aneurysms. J Neurosurg. 1968; 28(1):14–20.
2. Care of the patient undergoing intracranial pressure monitoring/external ventricular drainage or lumbar drainage. AANN clinical practice guideline series. 2011. http://www.aann.org/uploads/AANN11_ICPEVDnew.pdf. Accessed 27 Sept 2014.
3. Jaffe J, Melnychuk E, Muschelli J, et al. Ventricular catheter location and the clearance of intraventricular hemorrhage. Neurosurgery. 2012;70(5):1258–64.

Lumbar Puncture in Adults

41

Kevin Tench, L. Connor Nickels, and Rohit Pravin Patel

41.1 Indications

- Diagnostic
  - Evaluation for the possibility of a central nervous system (CNS) infection: viral, bacterial, and fungal meningitis and encephalitis
  - Evaluation for inflammatory processes: multiple sclerosis, Guillain-Barré syndrome
  - Evaluation for spontaneous subarachnoid hemorrhage (SAH)
  - Suspicion of CNS diseases: oncological and metabolic processes
- Therapeutic
  - Therapeutic reduction of cerebrospinal fluid (CSF) pressure
  - Procedures requiring lower body analgesia or anesthesia
  - Intrathecal antibiotic administration for some types of meningitis
  - Chemotherapy and methotrexate for some forms of leukemia and lymphomas

41.2 Contraindications

- Presence of infection in tissues at or around puncture site.
- Increased intracranial pressure (ICP) from a space-occupying lesion; patients with signs of cerebral herniation or with potential of increased ICP and focal neurological signs.
- Bleeding diathesis (thrombocytopenia, anticoagulant therapy, hemophilia); may increase risk of spinal hematoma, but level of coagulopathy that increases risk is unclear.
- Patients with cardiorespiratory compromise may worsen owing to position.

Patients with prior history of lumbar surgery, osteoarthritis, ankylosing spondylitis, kyphoscoliosis, or degenerative disk disease might have more success if lumbar puncture is performed by an interventional radiologist using imaging techniques and should be considered.

41.3 Materials and Medications (See Fig. 123.1)

- Spinal needle(s) with stylet
  - Adults: 3.5-in. 20-gauge needle; obese may require 5.0-in. 22- to 24-gauge needle
  - Children: 2.5-in. 22-gauge needle
  - Infants: 1.5-in. 22-gauge needle
- Three-way stopcock (optional: drainage catheter)
- Manometer (optional: extension tube for higher opening pressures)
- Specimen tubes (# may vary, but in general labeled 1–4, important to obtain from 1, 2, 3, 4 owing to cell count obtained from tubes 1 and 3)
- Local anesthetic (lidocaine 1 or 2 %), 5- to 10-mL syringe and needle (25-gauge) for local anesthesia
- Sterile drapes and gauze
- Mask, sterile gown, sterile gloves
- Antiseptic solution for skin preparation (Chloroprep or iodine)

K. Tench, MD
Department of Emergency Medicine, Banner Boswell Medical Center, Sun City, AZ, USA

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_41

41.4 Procedure

1. Positioning
    - Determined by practitioner preference or patient capability.
    - Options: lateral recumbent position, upright sitting position (Fig. 41.1).
    - Lateral recumbent position is preferred to obtain accurate opening pressure and to reduce the risk of post-puncture headache.
    - Both positions require the patient to arch the lower back toward the practitioner in order to open up the intervertebral spaces (obtain the “fetal position” or arch “like a cat”).
    - Shoulders and hips should remain aligned during process.
2. Landmarks
    - Determined by palpation.
    - Draw a visual line between the superior aspects of the iliac crests that intersects the midline at the L4 interspace. The L3–4 and L4–5 spaces are preferred because these points are below the termination of the spinal cord.
    - Palpate the posterosuperior iliac crests with the midpoint of a visual line that connects the two crests representing the L4 spinous process.
    - Palpate the space between the L3–4 or the L4–5 spinous processes and mark where the needle will be placed.
3. Ultrasound guidance (optional)
    - Helpful in obese patients, patients with previous surgical scarring, or anyone in whom palpation of the spinous processes is not easily done.
    - Sonographic measurement of the dura mater strongly correlates with needle depth needed to obtain CSF.
    - Identify the spinal process in the short and long axis to determine the midline and the interspinous space.
    - Identify the interspinous ligament for estimation of the depth of needle insertion.

      Commonly only the spinous processes are well visualized, and the interspinous ligament, ligamentum flavum, and subarachnoid space are less clearly seen.
    - High-frequency (5–10 MHz) linear probe to best evaluate anatomy.
    - A marking pen can be used to create a cross-hair-type figure (Fig. 41.1).
    - After placing the patient as described, locate the midline at the lumbar spine in transverse and longitudinal orientations.
    - Bright echogenic structures with shadowing posteriorly identify the spinous processes.
    - Transverse probe positioning to identify midline (Fig. 41.2), and then longitudinal probe positioning to identify interspinous space (Fig. 41.3).
4. Sterile preparation
    - After positioning and palpating the appropriate landmarks, the practitioner should then dress in the appropriate protective gear: mask, gown, and sterile gloves.
    - After dressing, the practitioner can then sterilely prepare the patient.
    - Make sure the patient’s back is completely exposed.
    - Clean the patient’s back with an antiseptic solution (Chloroprep should be scrubbed in an up/down and side/side fashion; iodine in a circular motion starting from the center of the anticipated insertion point).
    - Apply sterile drapes with the puncture site exposed.

      This is an optimal time to make sure all equipment in a standard lumbar puncture tray connects properly and that the stopcock for opening pressure measurement is assembled. Make sure the stopcock is closed away from the patient so that CSF can flow from the patient to the manometer. If the assembly is done, it will decrease the amount of CSF lost after the puncture.

- Local anesthesia
    - 1 % Lidocaine or anesthetic cream topically before preparing skin
    - For injection, form a skin wheal over the insertion site.
    - Inject into the deep tissues below the wheal in all directions while only breaking the skin once.
    - Systemic sedatives and analgesics may also be used.

5. Needle insertion
    - Needle should be inserted in the midline between the L3–4 or the L4–5 spinous process, and the stylet should be firmly in place.
    - Initially parallel to the bed, but once into the subcutaneous tissue, the needle should be angled toward the umbilicus (slightly cephalad, 15°) with the bevel facing upward (Fig. 41.4). This sagittal plane orientation spreads rather than cuts the fibers of the dural sac, which run parallel to the spinal axis.
    - If properly positioned, the needle passes through the skin; subcutaneous tissue; supraspinous ligament; interspinous ligament between the spinous processes; ligamentum flavum; epidural space including the internal vertebral venous plexus, dura, and arachnoid; into the subarachnoid space and between the nerve roots of the cauda equina.
    - In most cases, a “pop” will be felt when the needle penetrates the ligamentum flavum, entering into the subarachnoid space; then intermittent withdrawal should be done in 2-mm intervals to assess for CSF flow.
    - If bone is encountered during insertion, the needle should be withdrawn partially without exiting the skin and readjusted to a different angle more cephalad.

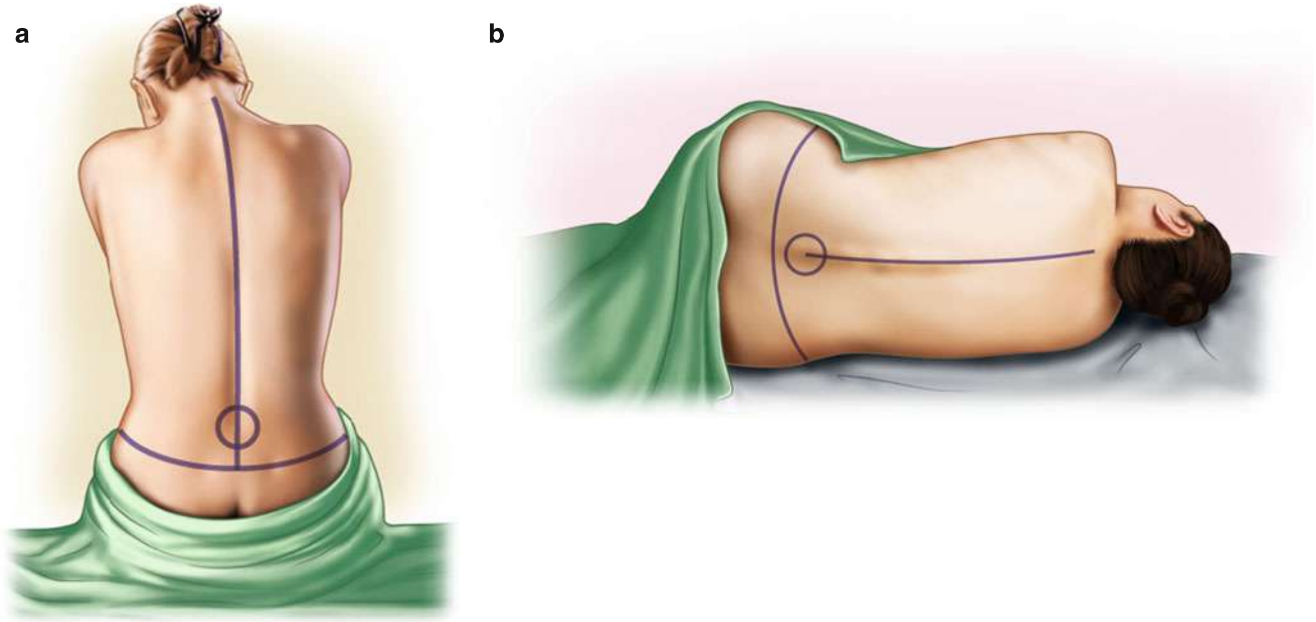

Fig. 41.1 Sitting position (**a**) left lateral decubitus position, (**b**) with general areas of insertion of needle

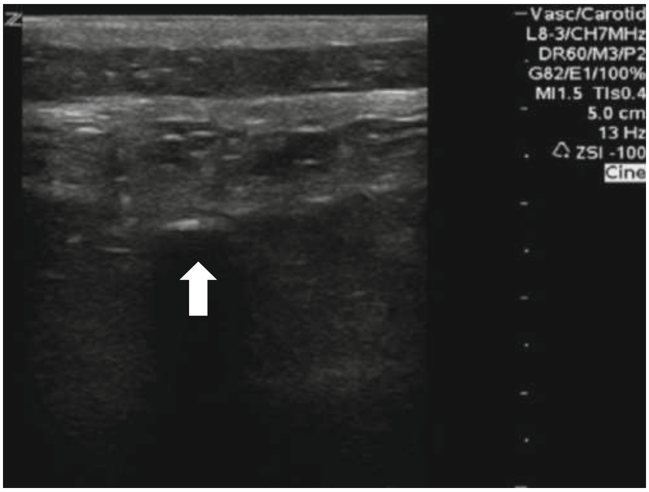

Fig. 41.2 Transverse view, *white arrow* indicates spinous process. Place in middle of ultrasound view to locate appropriate midline access point

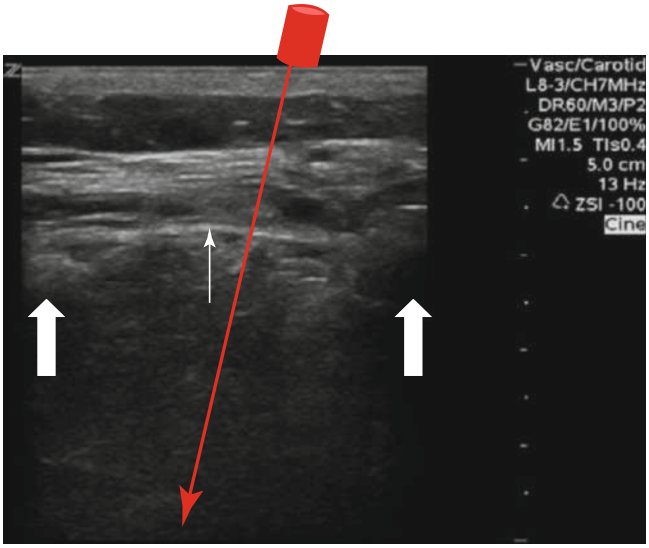

Fig. 41.3 Longitudinal view of lumbar spine. *Thick white arrows* indicate vertebral shadows; *thin white arrow* indicates supraspinous ligament. *Red arrow* indicates trajectory of needle

- If the tap is traumatic, CSF may be blood tinged but should clear as more is collected. If it does not clear, it may indicate intracranial hemorrhage or subarachnoid blood. Also in traumatic patients, clotting will be seen in the tubes; clotting does not occur in SAH owing to defibrinated blood being present in the CSF. Blood-tinged CSF can also be seen in herpes simplex virus (HSV) encephalitis.
- A dry tap is usually due to incorrect positioning and misdirection of needle, often due to a superior direction of the needle with obstruction by the lamina or spinous process of the superior or inferior vertebra. If the needle is too lateral, an inferior or superior articular process may be hit.

  If flow slows down, rotate the needle 90° because a nerve root may be obstructing the opening.

6. Opening pressure measurement
   - Must be performed in the lateral recumbent position. Although there are some conversion formulas from the sitting position, these are not standard of care.

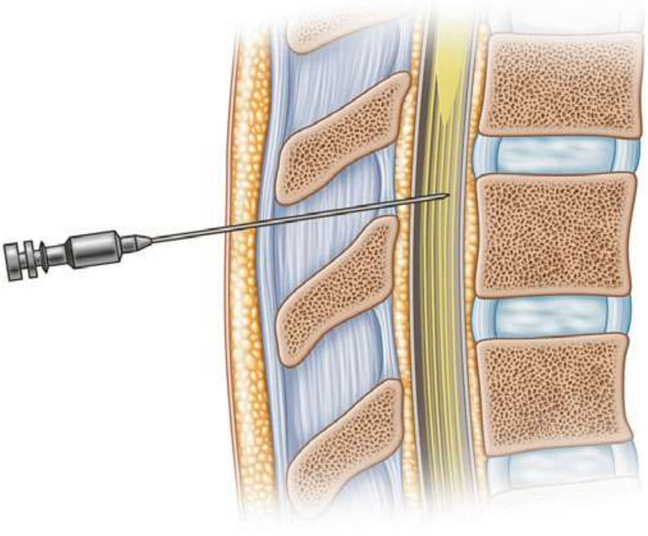

Fig. 41.4 Angle of insertion of needle, cephalad with bevel up

- Once the needle is in the subarachnoid space and CSF is flowing from the needle, the three-way stopcock should be attached to the needle and the manometer should be attached to the stopcock to take a measurement. Use the flexible tube to connect the manometer to the hub of the needle.
- Note the height of fluid in the manometer after it stops rising (normal opening pressure, <20 cm Hg); it may be possible to see pulsations from cardiac or respiratory motion.
  - Elevated CSF pressure is seen with meningeal inflammation, hydrocephalus, pseudotumor cerebri, SAH, and CHF.
  - Decreased CSF pressure is seen in leakage of CSF and severe dehydration.

7. Collecting CSF fluid
   - Collect at least 1–2 mL of CSF fluid in each tube, going from 1 to 4 and never aspirate because this can cause hemorrhage.
   - After collecting the fluid, replace the stylet and remove the needle, clean the skin, and place a bandage over the puncture site.
   - General recommendations
     - Tube 1: glucose, protein, protein electrophoresis
     - Tube 2: Gram stain, bacterial and viral cultures
     - Tube 3: cell count and differential
   - When ruling out SAH, cell count should be performed in tubes 1 and 3 or 1 and 4 to differentiate between SAH and traumatic tap.
     - Tube 4: Any special tests: myelin basic proteins, lactate, pyruvate, and smear on cell concentrates all depend on suspicion.

41.5 Complications

- Implantation of epidermoid tumors: from introducing skin plug into the subarachnoid space and can be avoided by using stylet when advancing.
- Postlumbar puncture headache: most common, occurring in 36.5 % of patients within 48 h
- CSF leak: causes headache when CSF leak through puncture site exceeds rate of production
- Bleeding: most common in patients with bleeding diathesis; may result in spinal cord compression
- Epidural hematoma
- Infection: local cellulitis, abscess (local or epidural), or meningitis
- Herniation syndromes: high risk can be identified by computed tomography but may not completely identify all patients with increased ICP
- Backache: local or referred pain
- Cardiorespiratory compromise

41.6 Pearls and Pitfalls

- Pearls
  - Positioning the patient is key to a successful procedure.
  - In adults the spinal cord may terminate higher than previously thought and it may be okay to go one interspace higher than recommended; but in infants owing to the differential in longitudinal growth of the spinal canal and cord, the spinal cord usually ends in L3. So in children the tap must go L4–5 or L5–S1.
  - Always keep the stylet in place until after the skin barrier is penetrated because this will avoid introduction of epidermoid tissue.
- Pitfalls
  - Postspinal headaches can be avoided with smaller needles and intravenous (IV) fluids.
    - Having the patient lie on the back for 1 h after the procedure has no change in incidence of headache.
    - Treatment consists of IV fluids initially, then caffeine, and, ultimately, if the headache persists, a blood patch.

Selected Reading

Boon JM, Abrahams PH, Meiring JH, Welch T. Lumbar puncture: anatomical review of a clinical skill. Clin Anat. 2004;17:544–53.

Ellenby MS, Tegtmeyer K, Lai S, Braner DA. Videos in clinical medicine. Lumbar puncture. N Engl J Med. 2006;355:e12.

Ferre RM, Sweeney TW, Strout TD. Ultrasound identification of landmarks preceding lumbar puncture: a pilot study. Emerg Med J. 2009;26:276–7.

Peterson MA, Abele J. Bedside ultrasound for difficult lumbar puncture. J Emerg Med. 2005;28:197–200.

Reflex Eye Movements (Doll's Eyes and Caloric Testing)

42

Thomas T. Nguyen, Tina Dulani, and Saadia Akhtar

42.1 Doll's Eyes (Oculocephalic Reflex Testing)

42.1.1 Indications: Doll's Eyes

- To assess brain stem function of a comatose patient
- To assess cerebral function in a comatose patient if brainstem function is intact

42.1.2 Contraindications: Doll's Eyes

- Absolute
  - Occult cervical spine injury; rule out radiographically and clinically
  - Basilar skull fracture
- Relative
  - Rheumatoid arthritis; increased risk of atlantoaxial subluxation resulting in spinal cord compression
  - Osteoporosis; increased risk of cervical spine injury
  - Cervical spine ankylosis; increased risk of cervical spine injury

42.1.3 Procedure: Doll's Eyes (Fig. 42.1)

1. Stand at the head of bed and grasp the patient's head with both hands.
2. Use the thumbs of both of hands to open the patient's eyelids.
3. Rapidly move the patient's head to one side and hold.
4. Simultaneously observe for the presence or absence of horizontal movements.
   (a) Head movement to one side should result in conjugate eye movement to the opposite side and then in spontaneous return of the eyes to the midline (normal test).
   (b) Normal oculocephalic reflex is the observation of conjugate eye movements to the opposite side of head turning. This indicates a functionally intact brainstem in a comatose patient.
   (c) Abnormal oculocephalic reflex: incomplete or absent horizontal eye movements. The eyes remain in the midline. This indicates impairment of the brainstem; caloric testing should be done if not contraindicated.
   (d) A partially abnormal oculocephalic reflex: conjugate eye movement opposite to head turning but does not return to the midline means the brainstem is intact but cerebrum function is not.
5. Repeat by rotating the head to the opposite side.
6. Vertical oculocephalic response can be tested by moving the patient's head up and down. A compensatory vertical eye movement should be observed.
   (a) This test is useful only if the horizontal oculocephalic reflex is negative. An intact vertical oculocephalic reflex with a negative horizontal oculocephalic reflex suggests a pontine lesion.
7. Document the observations.

42.1.4 Pearls and Pitfalls: Doll's Eyes

- Oculocephalic reflex may not be present in the first 10 days of life and is unreliable until 2 years of age.
- Do not attempt the doll's eye maneuver in patients with cervical spinal injuries.

T.T. Nguyen, MD (✉) • S. Akhtar, MD
Department of Emergency Medicine, Mount Sinai Beth Israel,
New York, NY, USA
e-mail: tnguyen@chpnet.org

T. Dulani, MD
Department of Emergency Medicine, New York Methodist
Hospital, Brooklyn, NY, USA
e-mail: tdulani@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_42

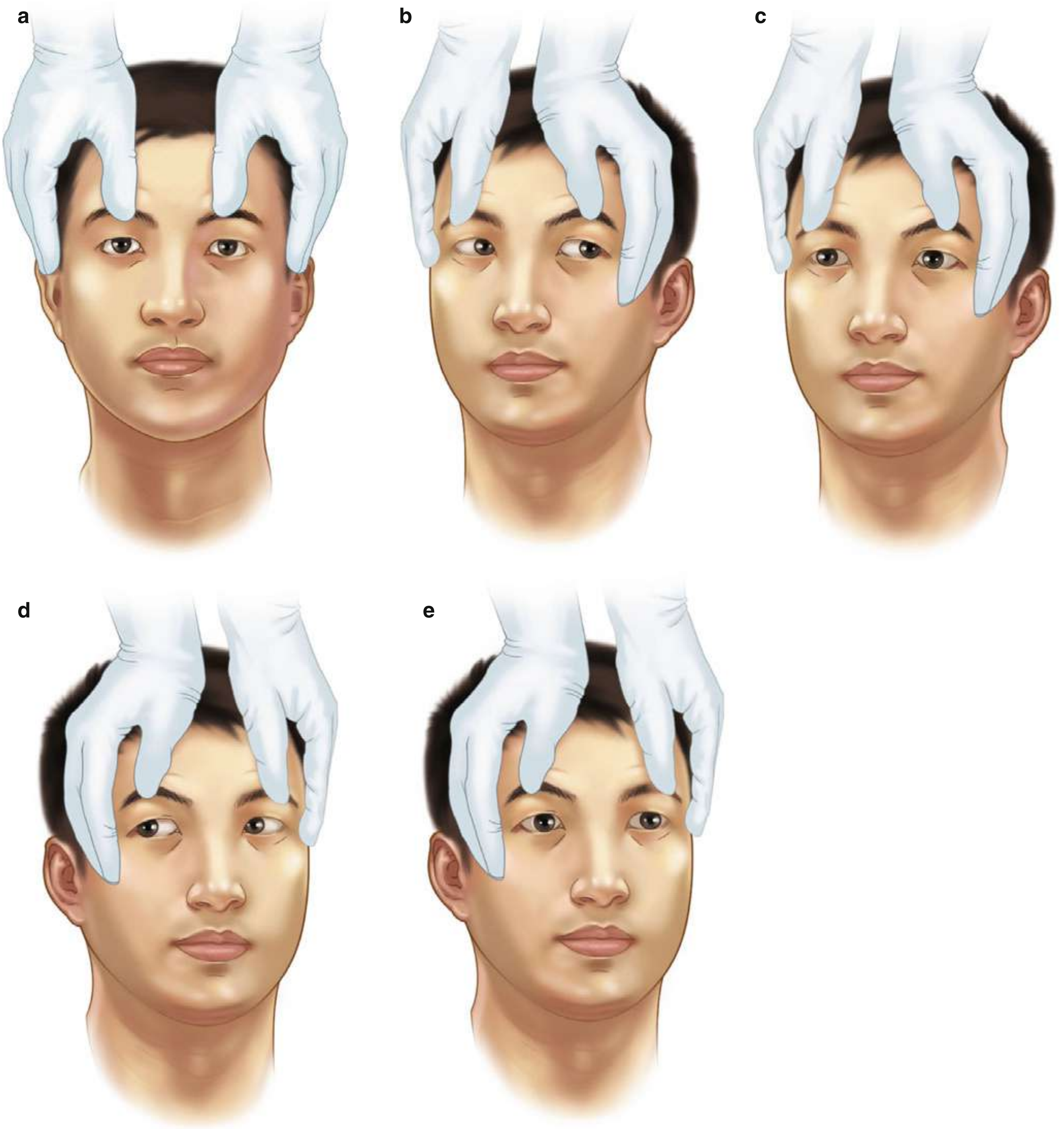

Fig. 42.1 The oculocephalic (doll's eyes) reflex in a patient with an intact brainstem. (**a**) Start with the head facing upright and grasp as depicted using both thumbs to keep the eyes open. (**b**) Rotate the head 90° to the right and the eyes deviate to the left (opposite side). (**c**) The eyes will spontaneously return to the midline. (**d**) Then rotate the head 180° to the left and the eyes should deviate to the right (opposite side). (**e**) The eyes will spontaneously return to the midline

- The doll's eye reflex can be absent or partial in patients with ocular muscle nerve palsy (e.g., cranial nerve [CN] 6).
- Make sure the patient does not have a neuromuscular blockade agent or other toxins present.
- A conscious person can suppress the doll's eye or oculocephalic reflex.

42.2 Caloric Reflex Testing (Vestibuloocular Reflex)

42.2.1 Indications: Caloric Reflex Testing

- In any comatose patient with abnormal doll's eye or if it cannot be performed
- To assess brainstem function of a comatose patient
- To assess asymmetrical function in the peripheral vestibular system

42.2.2 Contraindications: Caloric Reflex Testing

- Absolute
  - Perforated tympanic membrane
  - Presence of tympanostomy tubes
  - Basilar skull fracture, petrous bone fracture
  - Suspicion of cerebrospinal fluid (CSF) otorrhea

42.2.3 Materials and Medications: Caloric Reflex Testing

- Otoscope
- 60-mL syringe
- 16- to 18-gauge angiocatheter
- Thermometer
- Ice-cold water (30–33 °C)
- Warm water (44 °C)
- Emesis basin (to collect water)
- Towels or Chux

42.2.4 Procedure: Caloric Reflex Testing (Fig. 42.2)

1. Place the patient supine and elevate the head to 30° to bring the lateral semicircular canal into the upright position.
   (a) Careful otoscopic examination should be done before caloric testing to make sure there are no contraindications.
   (b) Remove any cerumen from the external auditory canal (EAC); the irrigation fluid should be able to reach the tympanic membrane.
2. Get a 60-mL syringe with an 18-gauge angiocatheter. Remove the plastic angiocatheter from the needle and discard the needle.
3. Place the angiocatheter on the 60-mL syringe and fill it with ice water (30 °C).
4. Place the angiocatheter in the EAC and squirt the water in over a period of 30–40 s. The water should freely enter and exit the auditory canal.
   (a) The stimulus depends upon the temperature of the water and not on water pressure.
   (b) Reflex horizontal movements may be delayed for up to 1 min after irrigation of the EAC.
5. Observe: Have an assistant hold the eyelids open with the head still and facing forward.
   (a) Normal test: Cold water induces slight deviation of the eyes to the side being irrigated followed by a latent period of approximately 20 s and then nystagmus to the opposite side (direction of the fast phase).
   (b) Abnormal test: Eyes do not deviate; this implies brainstem problems.
   (c) The quick phase may return in patients in a persistent vegetative state. Search for the slow, full eye deviation in response to caloric stimulation and not nystagmus when assessing a comatose patient.
6. Pause at least 5 min so the auditory canal can warm up.
   (a) The same EAC can be irrigated with warm water if the contralateral side cannot be tested for any reason.
7. Repeat steps 3–5 in the opposite ear.
   (a) Warm water testing can be used if there is no response to cold water testing. Use warm water (44 °C) and repeat testing, starting with one ear and then, after 5 min, with the other ear.
   (b) Warm water induces nystagmus to the same side (direction of the fast phase).
   (c) COWS: Cold opposite, warm same (referring to the compensatory fast phase of eye movement, the nystagmus).
8. Dry the patient off and reexamine the tympanic membranes to assess for testing-related injury.

42.2.5 Complications: Caloric Reflex Testing

- If contraindications are excluded, no significant complications are expected.
- Tympanic membrane injury or EAC injury can occur from the angiocatheter, forceful irrigation, or injury during cerumen removal.
- Potential complications from caloric testing include meningitis, otitis media, and vomiting.

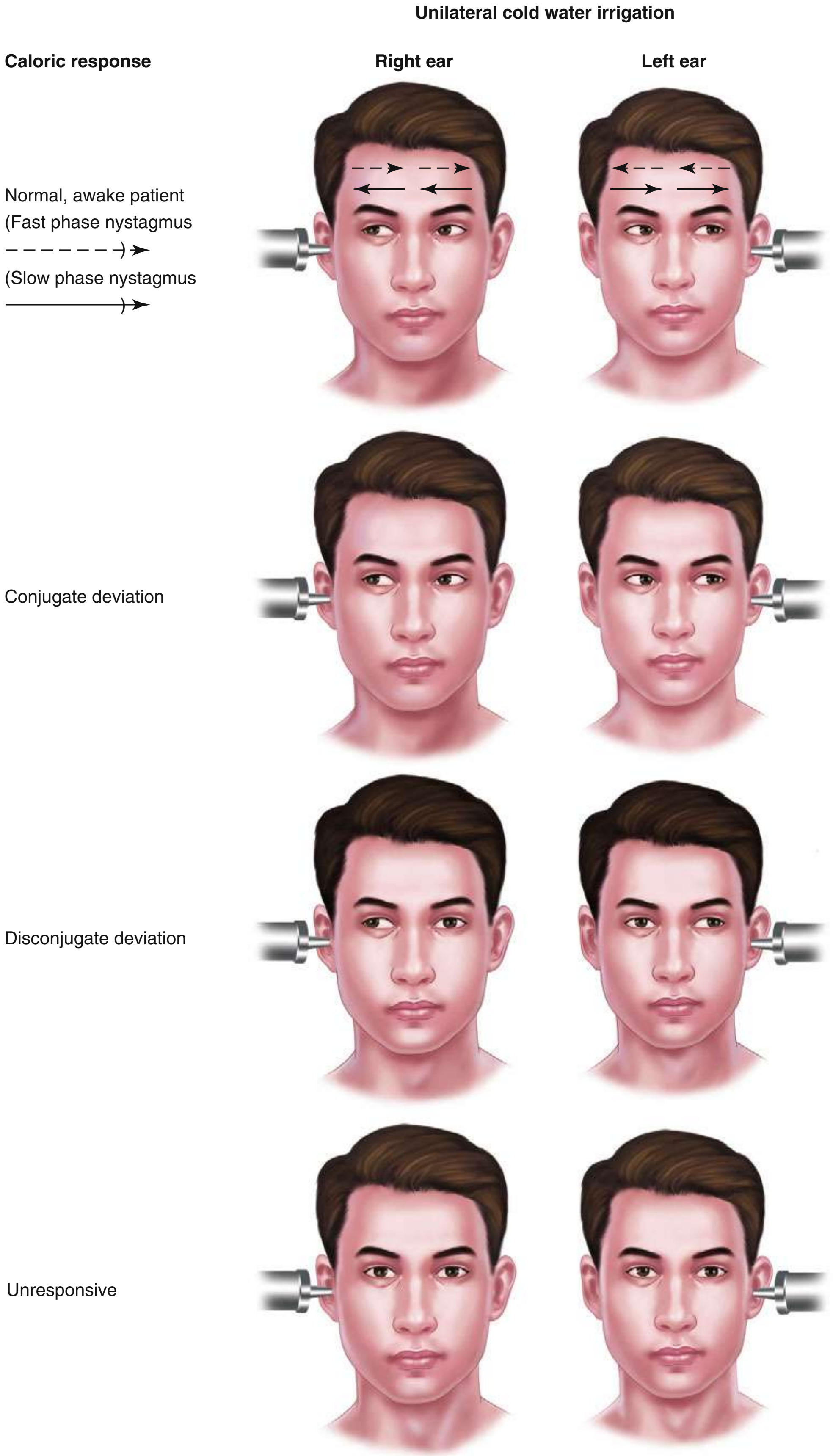
Unilateral cold water irrigation
Caloric response
Right ear
Left ear
Normal, awake patient
(Fast phase nystagmus
(Slow phase nystagmus
Conjugate deviation
Disconjugate deviation
Unresponsive

Fig. 42.2 Depiction of the vestibuloocular reflexes with unilateral cold-water irrigation

42.2.6 Pearls and Pitfalls: Caloric Reflex Testing

- A positive response to caloric testing indicates intact brainstem function.
- In an awake patient, excessive reflex vagal activity may occur (i.e., nausea, vomiting, dizziness).
- Absence of horizontal eye movement means comatose brainstem injury.
- Disconjugate or impaired reflex horizontal eye movements indicate impaired brainstem function at or below the level of the oculomotor nucleus.
- A comatose patient with intact full-reflex horizontal eye movements indicates that the lesion causing coma is in the cerebral hemisphere.

Selected Reading

Gomella LG. Central nervous system. In: Clinician's pocket reference. 11th ed. New York: McGraw-Hill; 2007.

McCann J, et al. Rapid assessment: a flowchart guide to signs and symptoms. Philadelphia: Lippincott William & Wilkins; 2004.

Reeve A, Swensen R. Auditory and vestibular function. In: Disorders of the nervous system: a primer. Hanover, NH: Dartmouth Medical School; On Line; 2009.

Reichman EF, Simon RR. Chapter 103. Reflex eye movements (caloric testing and doll's eyes). In: Emergency medicine procedures. New York: McGraw Hill Education; 2004.

Ropper AH, Samuels MA. Chapter 15. Deafness, dizziness, and disorders of equilibrium. Chapter 17. Coma and related disorders of consciousness. In: Adams and Victor's principles of neurology. 9th ed. New York: McGraw Hill; 2009.

Simon RP, Greenberg DA, Aminoff MJ. Chapter 3. Disorders of equilibrium. In: Clinical neurology. 7th ed. New York: McGraw Hill; 2009. http://www.accessmedicine.com.elibrary.einstein.yu.edu/content.aspx?aID=5146162.

Dix-Hallpike Maneuver

43

Rui Domingues and Muhammad Waseem

The Dix-Hallpike maneuver, also termed the "head-hanging positioning maneuver," is helpful in confirming the clinical suspicion of benign paroxysmal positional vertigo (BPPV). This maneuver provokes abnormal nystagmus, which is a characteristic feature of BPPV.

43.1 Indications

- BPPV is one of the most common types of vertigo.
  - The pathophysiology of BPPV, in brief, is believed to be due to free-moving densities (canaliths/otoliths) in the posterior semicircular canal; with head movement, the particles would alter the flow of the endolymph and cause the stimulation of the ampulla. The particles in the canal cause slow or even reversal of the movement of the cupula and create signals that do not correlate with the actual head movements, therefore causing the sensation of nystagmus.
  - This maneuver locates the cause of vertigo as either the inner ear or the brain; if the problem is in the ear, this maneuver helps localize which ear is affected.
- This maneuver is indicated for patients presenting with vertigo, which is evoked by a change in position and has no symptoms at rest.
- This maneuver is inexpensive, easily done, and part of the physical examination when a patient presents with the complaint of dizziness or vertigo.

43.2 Contraindications

- Severe cervical spine disease.
- Unstable spinal injury.
- High-grade carotid stenosis.
- Unstable heart disease.
- Elderly patients may not tolerate this maneuver.
- There is no need to perform this test in the presence of nystagmus at rest.

43.3 Materials

- Examination table
- Flat cushion
- Frenzel goggles: These high-powered (+20 diopters) magnifying glasses can be placed on a patient during the performance of the maneuver and have shown to increase the sensitivity of the Dix-Hallpike maneuver by preventing the patient from visually fixating on an object, thereby preventing suppression of nystagmus. They are not required to perform the maneuver; they usually are used by specialists.

43.4 Procedure (Fig. 43.1)

1. Have the patient sit at the edge of a bed. The patient is instructed to maintain eye contact with the physician throughout the maneuver.
2. With the patient seated, the examiner will extend the neck, approximately 20°, and turn the head to one side, approximately 30–45°.
3. The examiner then assists the patient by lowering the patient quickly into a supine position, so that the head hangs over the edge of the bed or table, with the neck in a hyperextended position. A flat cushion can be placed

R. Domingues, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: rbdomingues@hotmail.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_43

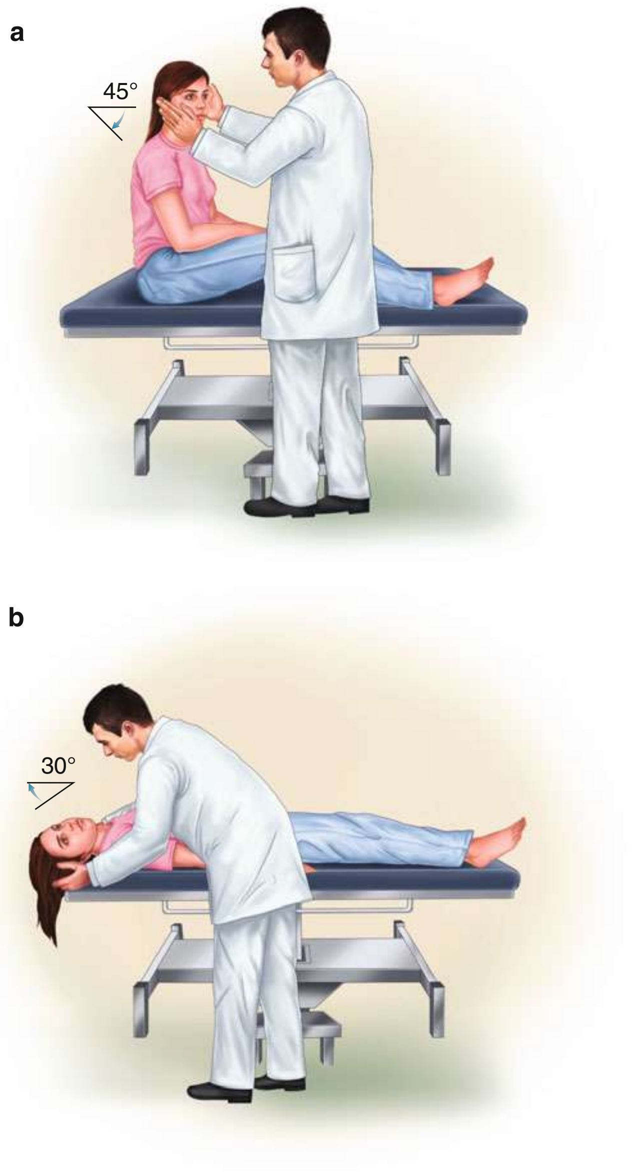

Fig. 43.1 (**a**) With the patient seated, the doctor will extend the neck, approximately 20°, and turn the head to one side, approximately 30–45°. (**b**) The physician then assists the patient by lowering the patient quickly into a supine position, so that the head hangs over the edge of the bed or table, with the neck in a hyperextended position

beneath the person's back in the shoulder blade area to assist with obtaining head extension.

4. This position is held, and the examiner observes for nystagmus for up to 60 s.
   (a) Nystagmus is a rapid, rhythmic movement of the eyes and usually appears after a brief period and lasts less than 30 s.
   (b) The direction of the nystagmus is usually up and twisted; therefore, the eyes will beat toward the ground.
   (c) If the patient becomes dizzy or the doctor sees nystagmus, the test is positive for the ear that is pointed to the floor.
5. The patient is then returned to the upright position and is again observed for nystagmus for 30 s.
6. If nystagmus is not provoked, the maneuver is repeated with the head turned to the opposite direction.
7. If nystagmus is provoked, the patient should have the maneuver repeated to the same side.
   (a) With each repetition, the intensity and duration of nystagmus should decrease.

43.5 Complications

- Vertigo
- Nausea

43.6 Pearls and Pitfalls

- The maneuver can be uncomfortable to the patient because it can cause vertigo and nausea.

Selected Reading

Brandt T, Steddin S, Daroff RB. Therapy for benign paroxysmal positioning vertigo, revisited. Neurology. 1994;44:796–800.

Buckingham RA. Anatomical and theoretical observations on otolith repositioning for benign paroxysmal positional vertigo. Laryngoscope. 1999;109:717–22.

Fife TD, Iverson DJ, Lempert T, et al. Practice parameter: therapies for benign paroxysmal positional vertigo (evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology. 2008;70:69–74.

Halker RB, Barrs DM, Wellik KE, Wingerchuck DM, Demaerschalk BM. Establishing a diagnosis of benign paroxysmal positional vertigo through the Dix-Hallpike and side-lying maneuvers: a critically appraised topic. Neurologist. 2008;14:201–4.

Herdman SJ. Treatment of benign paroxysmal vertigo. Phys Ther. 1990;70:381–8.

Lanska DJ, Remler B. Benign paroxysmal positioning vertigo: classic descriptions, origins of the provocative positioning technique and conceptual developments. Neurology. 1997;48:1167–77.

Li JC, Meyers AD. Benign paroxysmal positional vertigo. http://emedicine.medscape.com/article/884261-overview. Accessed December 20, 2015.

Von Brevern M, Radtke A, Lezius F, et al. Epidemiology of benign paroxysmal positional vertigo: a population based study. J Neurol Neurosurg Psychiatry. 2007;78:710–5.

Epley Maneuver for Vertigo (Particle Repositioning or Canalith Repositioning Procedure)

44

Rui Domingues and Muhammad Waseem

During the Dix-Hallpike test (see Chap. 43), the particles move in the canal and trigger a burst of upbeat-torsional nystagmus. The Epley maneuver causes resolution of positional nystagmus. This maneuver is effective in about 80 % of patients with benign paroxysmal positional vertigo (BPPV).

44.1 Indications

- Performed to alleviate the symptoms of posterior canal BPPV

44.2 Contraindications

- Back or spine injuries or other problems
- Presence of detached retina

44.3 Materials

- Pillow or pad
- Padded table or bed

44.4 Procedure (Fig. 44.1)

1. The patient is positioned on the bed with the head slightly extended and turned 45° to the affected side; hold this position for 60 s.
    - The patient may hold the physician's arm for support.
2. The patient is then assisted by the physician to a supine position, with the head in extension, with a pillow or pad placed at the shoulder level allowing for extension; this position is held for a 60 s. This position will likely provoke transient dizziness and vertigo.
3. The head is then turned 90° to the opposite side; this position is held for 60 s. This position will likely provoke transient dizziness and vertigo.
4. The head is then turned through a further 90° while the patient rolls onto the unaffected side; this position is held for 60 s.
5. The patient then sits up slowly, by rolling up from lying on the unaffected side; this position is held for 60 s.
6. The patient should wait for 10–15 min before discharge.
7. Discharge instructions are then provided. Home instructions:
    - Instruct the patient to sleep in a semi-recumbent position for the next two nights; the patient should sleep at a 45° angle, which is most easily done by sleeping in a recliner or with several pillows arranged on a couch.
    - Sleep on the nonaffected side.
    - During the day, try to keep the head in a vertical position; no sudden head movements to the right, left, up, or down.
    - When men shave under their chins, they should bend their bodies forward in order to keep their heads vertical.
    - Do not go to the hairdresser or dentist because these require head movements.
    - Care should be taken when putting in eye drops because it requires head extension.
    - No vacuuming or mopping the floor.
    - Try not to wear clothing that needs to be pulled over the head. Try to avoid bending down to tie shoes.

R. Domingues, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: rbdomingues@hotmail.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_44

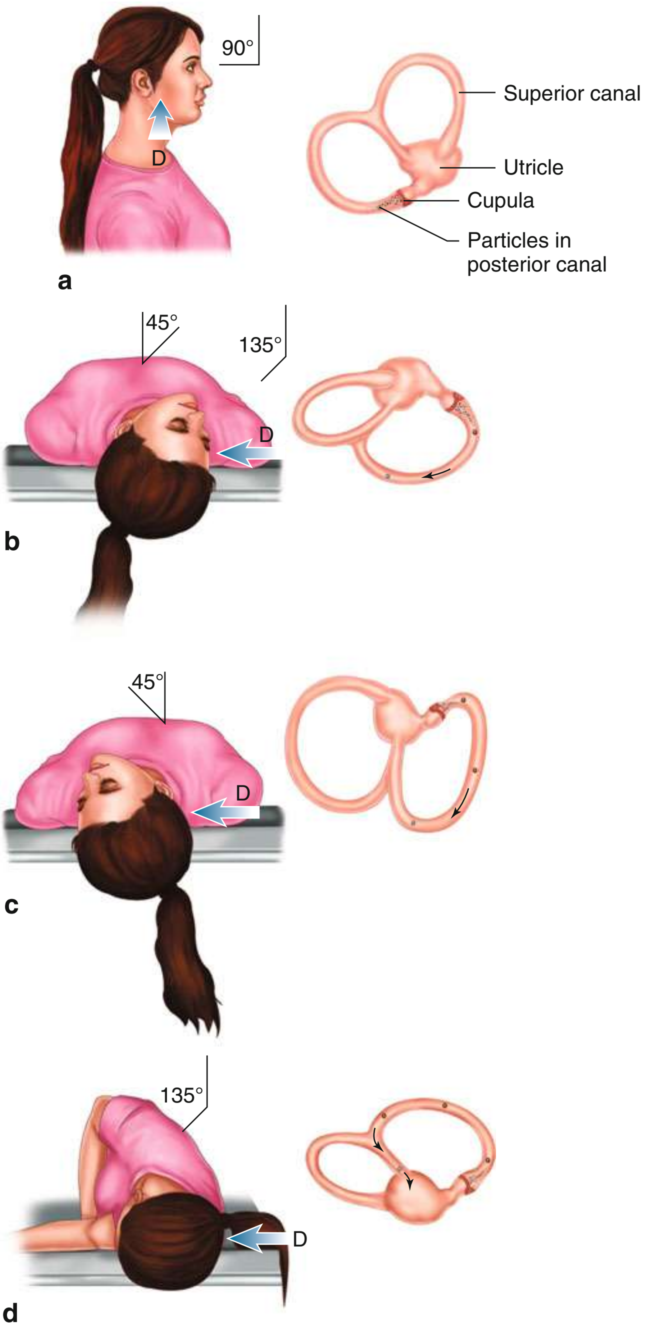

Fig. 44.1 (**a**) Patient seated on the table. (**b**) Patient in supine position, with the head turned 45° to the affected side. (**c**) Head then turned to the unaffected side (45°). (**d**) Head turned additional 90° on the unaffected side. *D* direction of view of labyrinth

44.5 Complications

When performing the previous maneuver, caution is advised should neurological symptoms occur. Occasionally such symptoms are caused by compression of the vertebral arteries; if it persists for a long period of time, a stroke can occur.

44.6 Pearls and Pitfalls

- Pearls
  - This maneuver should initially be performed by a trained therapist or medical physician.
  - It is best to perform the Epley maneuver before going to bed at night.
- Pitfalls
  - The patient should not drive herself or himself home after the procedure.
  - Avoid rapid changes in head position that might provoke BPPV.

Selected Reading

Harvey SA, Hain TC, Adamiec LC. Modified liberator maneuver: effective treatment for benign paroxysmal positional vertigo. Laryngoscope. 1994;104:1206–12.

Herdman SJ. Treatment of benign paroxysmal vertigo. Phys Ther. 1990;70:381–8.

Massoud EA, Ireland DJ. Post-treatment instructions in the nonsurgical management of benign paroxysmal positional vertigo. J Otolaryngol. 1996;25:121–5.

Parnes LS, Price-Jones RG. Particle repositioning maneuver for benign paroxysmal positional vertigo. Ann Otol Rhinol Laryngol. 1993;102:325–31.

Radtke A, Neuhausen H, von Brevern M, Lempert T. A modified Epley's procedure for self-treatment of benign paroxysmal positional vertigo. Neurology. 1999;53:1358–60.

Smouha EE. Time course of recovery after Epley maneuvers for benign paroxysmal positional vertigo. Laryngoscope. 1997;107:187–91.

Welling DB, Barnes DE. Particle repositioning maneuver for benign paroxysmal positional vertigo. Laryngoscope. 1994;104:946–9.

Clinical Brain Death Examination in Adults

45

Eric S. Papierniak, Hassan M. Alnuaimat, Tracy Timmons,
and Deborah M. Stein

45.1 Indications

- Deep coma without any identifiable reversible causes
- Suspected brain death in a mechanically ventilated patient

45.2 Contraindications

- Metabolic derangements
- Acid-base disorders
- Electrolyte abnormalities sufficient to cause the coma
- Profound hypotension
- Hypothermia (core body temperature <36°C)
- Drug overdose or toxic exposure
- Locked-in syndrome

45.3 Materials and Medications

- A bright light
- Intravenous (IV) catheter
- 30-mL syringe
- 50-mL of ice water
- Endotracheal suction catheter
- Long swab or tongue depressor
- Gauze 4×4 or saline eye drops

45.4 Procedure

1. Evaluate the motor response to pain.
    - Apply pressure to the supraorbital nerve and nail beds. No motor response is consistent with brain death (Figs. 45.1 and 45.2).
2. Test for the absence of brainstem function.
    - Check pupillary response to bright light.
    - Check the oculocephalic reflex. With the eyelids held open, quickly turn the head to the side. The normal response is for the eyes to move in the opposite direction, maintaining the focus forward. Failure of the eyes to deviate during head rotation is consistent with brain death (Fig. 45.3).
    - Check the oculovestibular reflex with cold caloric testing. Elevate the head of the bed to 30°. Attach an IV catheter to the tip of a 30-mL syringe. Inject 50 mL of cold water into the external ear canal. Observe the pupils for 1 min for deviation toward the ear with the cold stimulus. Wait 5 min before testing the other side. No deviation of the pupils is consistent with brain death (Figs. 45.4 and 45.5).
    - Check for bulbar paralysis.
        - Check for a cough response to bronchial suctioning.

E.S. Papierniak (✉)
Division of Pulmonary, Critical Care, and Sleep Medicine,
Department of Medicine, University of Florida Health,
Gainesville, FL, USA

Malcom Randall VA Medical Center, Gainesville, FL, USA
e-mail: eric.papierniak@medicine.ufl.edu

H.M. Alnuaimat, MD
Department of Medicine, University of Florida, University of
Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: Hassan.Alnuaimat@medicine.ufl.edu

T. Timmons, MD
Department of Trauma/Surgical Critical Care, R Adams Cowley
Shock Trauma Center, University of Maryland Medical Center,
Baltimore, MD, USA
e-mail: ttimmons1@umm.edu

D.M. Stein, MD, MPH
Department of Surgery, University of Maryland School of
Medicine, R Adams Cowley Shock Trauma Center, University of
Maryland Medical Center, Baltimore, MD, USA
e-mail: dstein@umm.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_45

  - Stimulate the posterior pharynx with a long swab or tongue depressor and observe for a gag.
  - Check the corneal reflex by lightly touching the cornea with a cotton swab, corner of gauze, or drops of sterile saline and observe for blinking of the eyelids (Fig. 45.6).

3. Check for respiratory effort by performing an apnea test. Before starting the apnea test, the following conditions must be met:
   (a) The patient must not be hypothermic. The core temperature must be greater than 36°C.
   (b) The patient must be hemodynamically stable with a systolic blood pressure greater than 90 mmHg.
   (c) The arterial blood gas (ABG) must demonstrate a normal arterial partial pressure of carbon dioxide ($PaCO_2$). The arterial partial pressure of oxygen (PaO_2) may be elevated with preoxygenation to minimize the risk of hypoxemia prematurely ending the examination.
   - Monitor the patient with a pulse oximeter.
   - Disconnect the ventilator and deliver oxygen at 6 L/min by cannula into the endotracheal tube.
   - Alternately the patient can remain connected to the ventilator with no applied support, which allows for an in-line negative pressure monitor to be attached to the circuit. Note: Most ventilators are too sensitive to be used as the sole method of sensing respiratory effort as they can produce false-negative results.
   - Watch the chest and abdomen for respiratory motion, or monitor for negative inspiratory force on the pressure gauge. Check an ABG every 10 min and when the test ends.
   - If the $PaCO_2$ increases by 20 mmHg or the $PaCO_2$ is greater than 60 mmHg and there was no respiratory effort, the test is consistent with brain death.
   - The test must be aborted and the patient reconnected to mechanical ventilation if the patient becomes hypotensive, desaturates, or develops cardiac arrhythmias.
   - If there is respiratory effort, the patient must be reconnected to mechanical ventilation (Figs. 45.7 and 45.8).

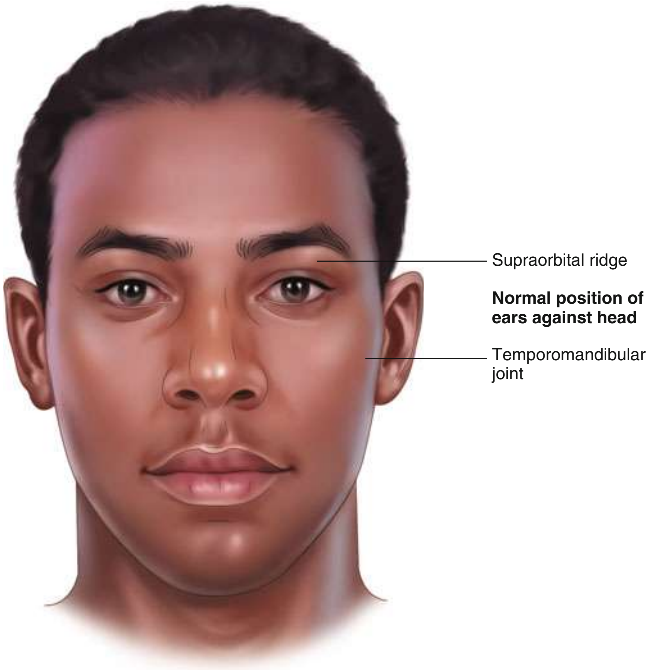

Fig. 45.1 Motor response to pain 1. Temporomandibular joint and supraorbital ridge test supraorbital nerve function

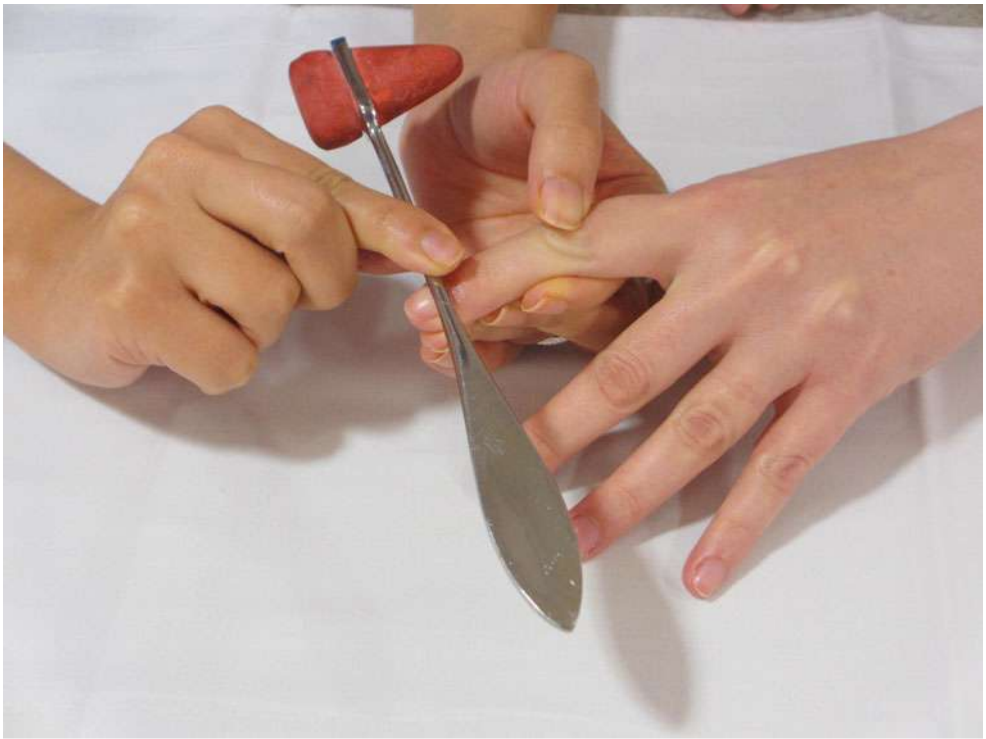

Fig. 45.2 Motor response to pain 2: Firm pressure can be applied to the nail beds with a penlight or similar instrument

a Normal reaction: Eyes move side to side when head is turned

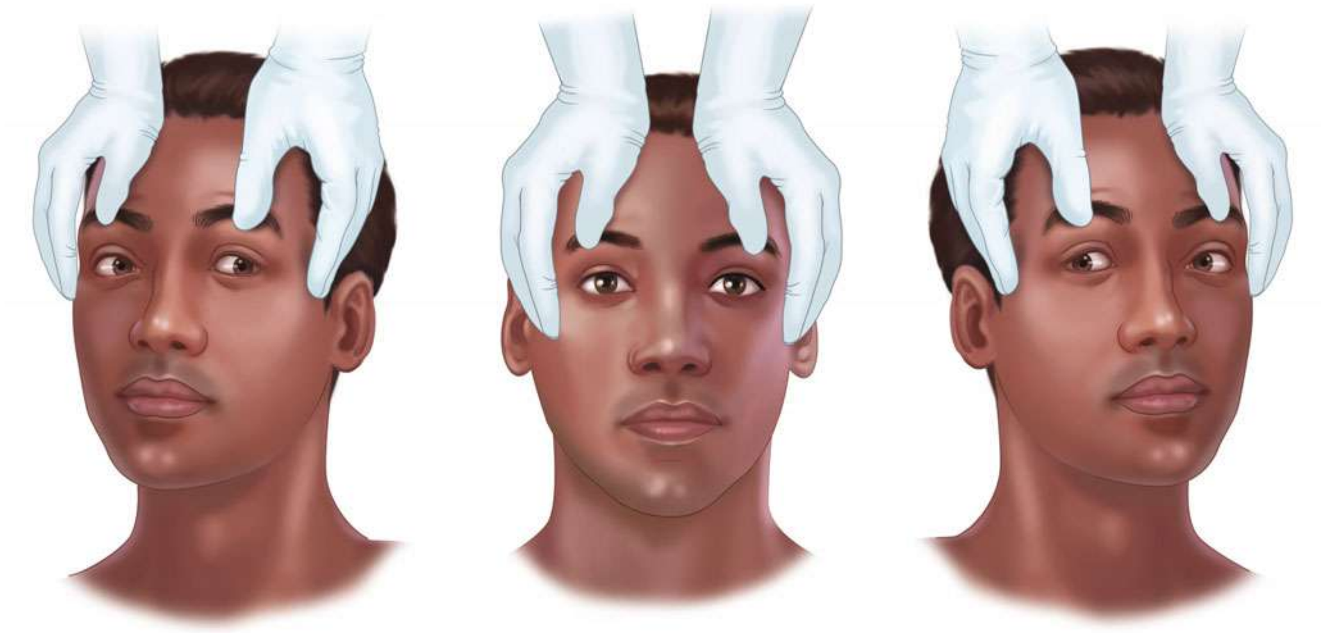

b Abnormal reaction: Eyes remain in fixed position in skull when head is turned

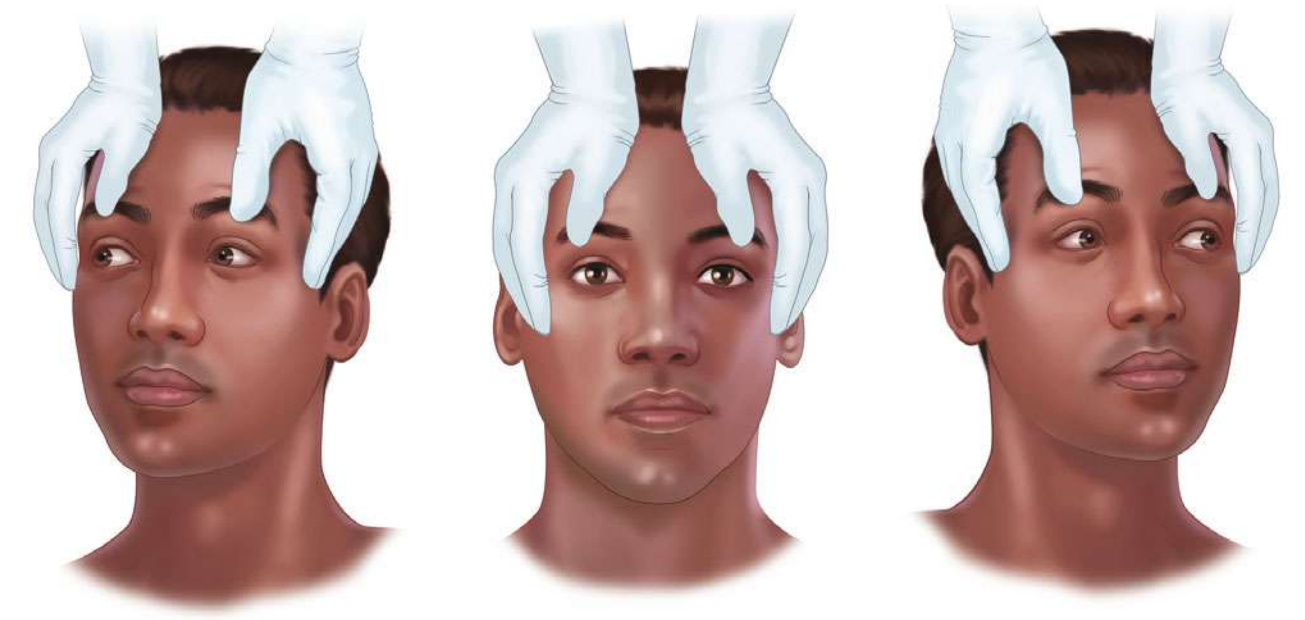

Fig. 45.3 (**a**, **b**) Oculocephalic reflex or "Doll's eyes." Movement of the eyes away from the direction of head turning (i.e., to keep the gaze forward) indicates intact functioning

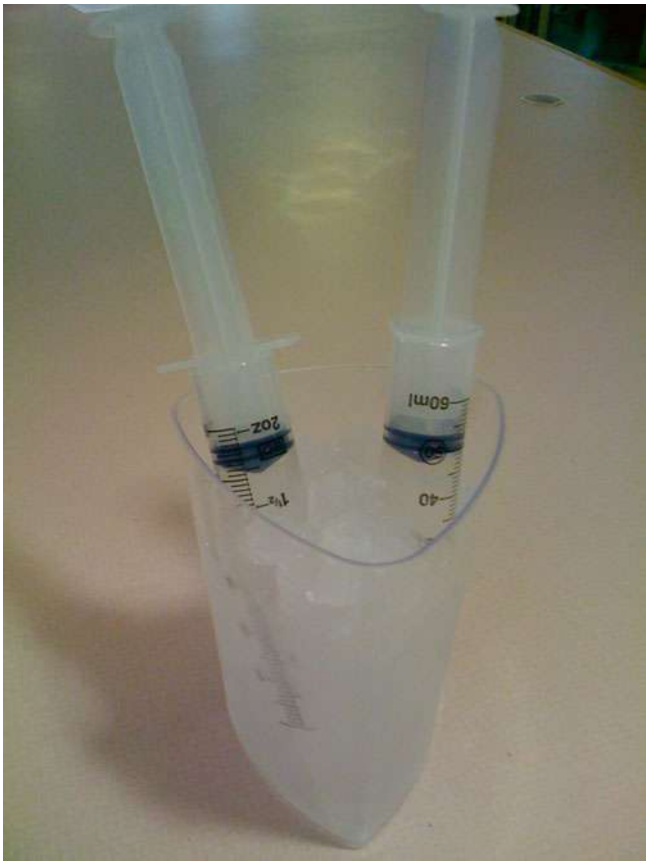

Fig. 45.4 Preparation of cold saline for testing of the oculovestibular reflex or "cold calorics"

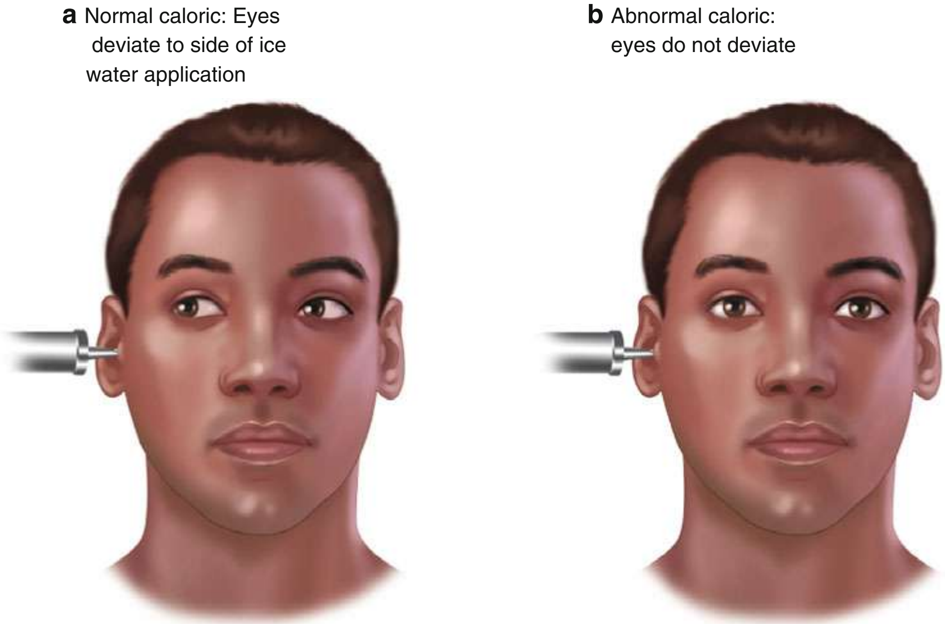

a Normal caloric: Eyes deviate to side of ice water application

b Abnormal caloric: eyes do not deviate

Fig. 45.5 (**a**, **b**) Oculovestibular reflex. Intact functioning is demonstrated by deviation of the gaze toward the side being tested

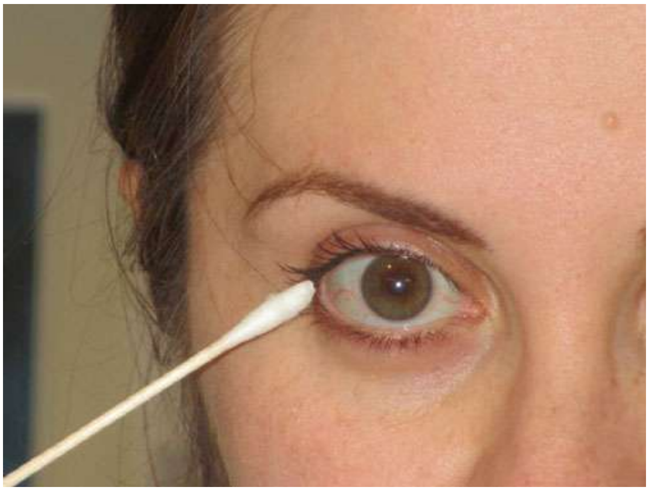

Fig. 45.6 Use of a cotton ball or unwound tip of a cotton swab for testing of the corneal reflex. A piece of sterile gauze ("4×4") is also commonly used

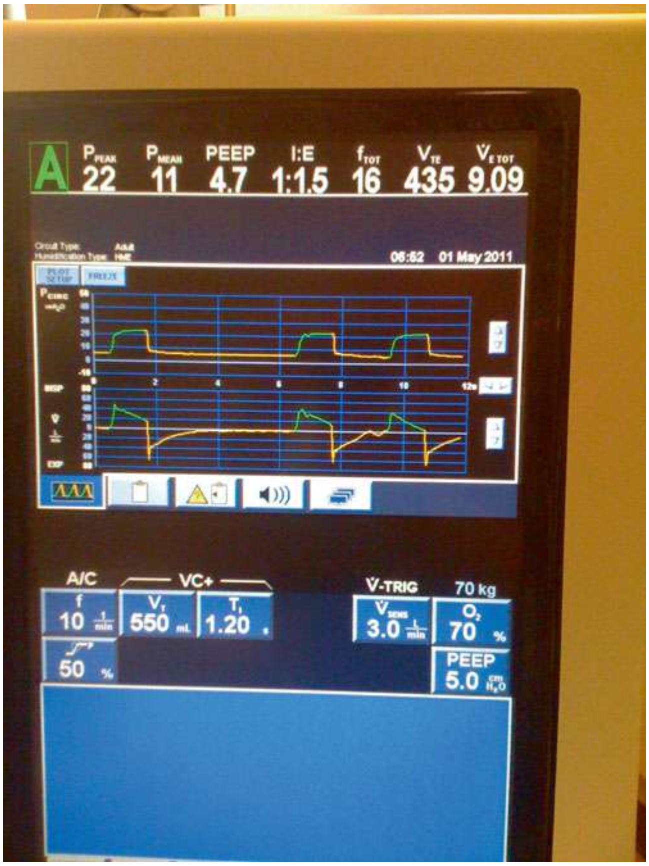

A
P_{PEAK} 22
P_{MEAN} 11
PEEP 4.7
I:E 1:1.5
f_{TOT} 16
V_{TE} 435
$\dot{V}_{E\,TOT}$ 9.09

Circuit Type Adult
Humidification Type HME
06:52 01 May 2011

A/C
VC+
f 10 $\frac{1}{min}$
V_T 550 mL
T_I 1.20 s
50 %
V-TRIG
V_{SENS} 3.0 $\frac{L}{min}$
70 kg
O_2 70 %
PEEP 5.0 cmH_2O

Fig. 45.7 Ventilator waveform before apnea test

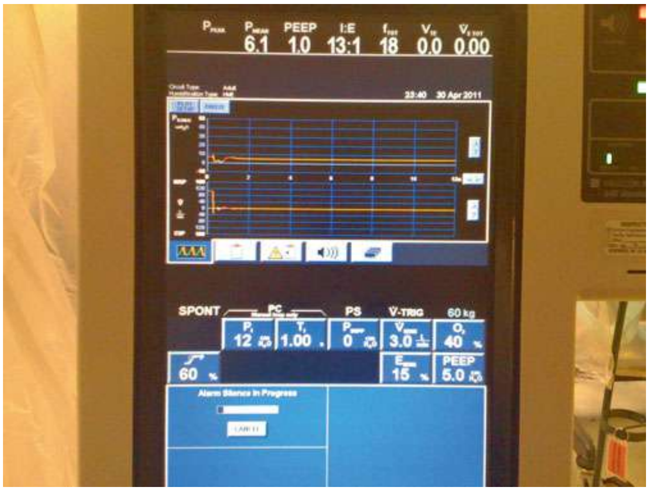

P_{PEAK}
P_{MEAN} 6.1
PEEP 1.0
I:E 13:1
f_{TOT} 18
V_{TE} 0.0
$\dot{V}_{E\,TOT}$ 0.00

23:40 30 Apr 2011

SPONT
PC
PS
V-TRIG
60 kg
P_I 12
T_I 1.00
P_{SUPP} 0
V_{SENS} 3.0
O_2 40 %
60 %
E_{SENS} 15 %
PEEP 5.0

Alarm Silence in Progress

Fig. 45.8 Ventilator waveform in a patient without spontaneous respirations during the apnea test

45.5 Pearls and Pitfalls

- Pearls
  - The clinical examination should be performed by someone with experience or training in brain death examinations. Local regulations as well as hospital policies should be reviewed in order to determine the requirements for brain death. Some locales require a repeat examination by a different practitioner.
  - A patient with a cervical spine injury causing quadriplegia may not feel nail bed pressure and may be unable to respond to pain with more than facial movements.
- Pitfalls
  - Pupillary constriction less than 3 mm is not consistent with brain death.
  - Do not test oculocephalic reflex (doll's eyes) on patients who may have a cervical spinal cord injury from trauma.
  - Do not test the oculovestibular reflex with cold caloric testing in patients with a ruptured tympanic membrane.
  - A large number of patients may demonstrate spinal reflex movements during brain death. These reflexes may be triggered by touch, noxious stimuli, or removal of the ventilator. These movements can include plantar flexion, upper extremity posturing, eye opening, and the "Lazarus sign" (the arms raise off the bed and cross). Although these movements can be disconcerting to the health care team and family, they do not preclude a diagnosis of brain death.

45.6 Controversies

- Repeat examinations
  - The American Academy of Neurology guidelines recommend repeating the brain death examination at 6 h. The guidelines acknowledge that this is an arbitrary interval. Recent evidence suggests that the second examination may be unnecessary.
- Confirmatory Testing
  - Confirmatory testing is an option if specific elements of the clinical examination cannot be performed. For example, a trauma patient with suspected cervical spinal cord injuries precluding the oculocephalic reflex test would be a candidate for confirmatory testing.
  - Conventional cerebral angiography will demonstrate no filling beyond the carotid bifurcation or within the circle of Willis. Electroencephalography demonstrates no electrical activity. Transcranial Doppler ultrasonography will demonstrate vascular resistance associated with elevated intracranial pressure. Technetium-99m hexamethylpropylene-amine oxime brain scans demonstrate the "hollow skull" sign or no uptake of isotope in the brain (Fig. 45.9).
  - Newer modalities of confirmatory testing include computed tomography (CT) angiography and magnetic resonance imaging/magnetic resonance angiography (MRI/MRA). These tests are being used in some hospitals; however, the recent review by the American Academy of Neurology finds evidence insufficient to recommend using these newer modalities to confirm brain death.

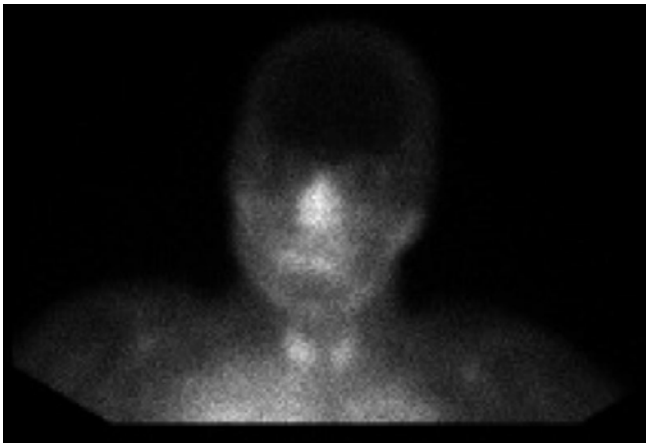

Fig. 45.9 "Empty skull sign" on nuclear medicine flow study confirms brain death

Selected Reading

A definition of irreversible coma: report of the Ad Hoc Committee of the Harvard Medical School to Examine the Definition of Brain Death. JAMA. 1968;205:337–40.

Lustbader D, O'Hara D, Wijdicks EF, et al. Second brain death examination may negatively affect organ donation. Neurology. 2011;76:119–24.

Practice parameters for determining brain death in adults (summary statement). The quality standards subcommittee of the American Academy of Neurology. Neurology. 1995;45:1012–4.

Saposnik G, Maurino J, Saizar R, Bueri JA. Spontaneous and reflex movements in 107 patients with brain death. Am J Med. 2005;118:311–4.

Wijdicks EF. Current concepts: the diagnosis of brain death. N Engl J Med. 2001;344:1215–21.

Wijdicks EF, Varelas PN, Gronseth GS, Greer DM, American Academy of Neurology. Evidence-based guideline update: determining brain death in adults: report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology. 2010;74:1911–8.

Part VII

Ophthalmic Procedures

Slit Lamp Examination

46

Bobby K. Desai

46.1 Indications

- A slit lamp magnifies structures of the eye (Fig. 46.1).
- Gives the operator a three-dimensional view of the area visualized.
- Used to delineate abnormalities that cannot be visualized by other means.
- Helpful in foreign body removal.

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_46

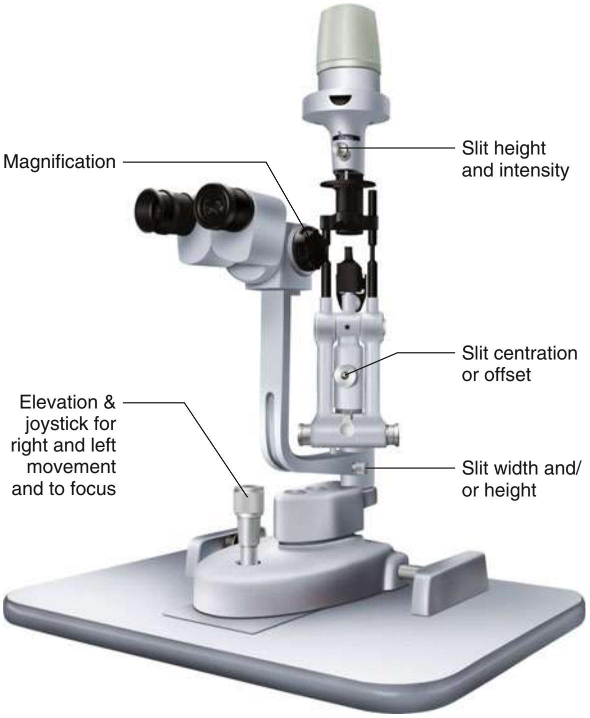
Magnification
Slit height and intensity
Slit centration or offset
Elevation & joystick for right and left movement and to focus
Slit width and/ or height

Fig. 46.1 A basic slit lamp

46.2 Contraindications

- Not to be used in patients who cannot tolerate an upright posture (e.g., trauma patients).
  - In these cases if a slit lamp examination is required, a portable slit lamp may be beneficial (Fig. 46.2).

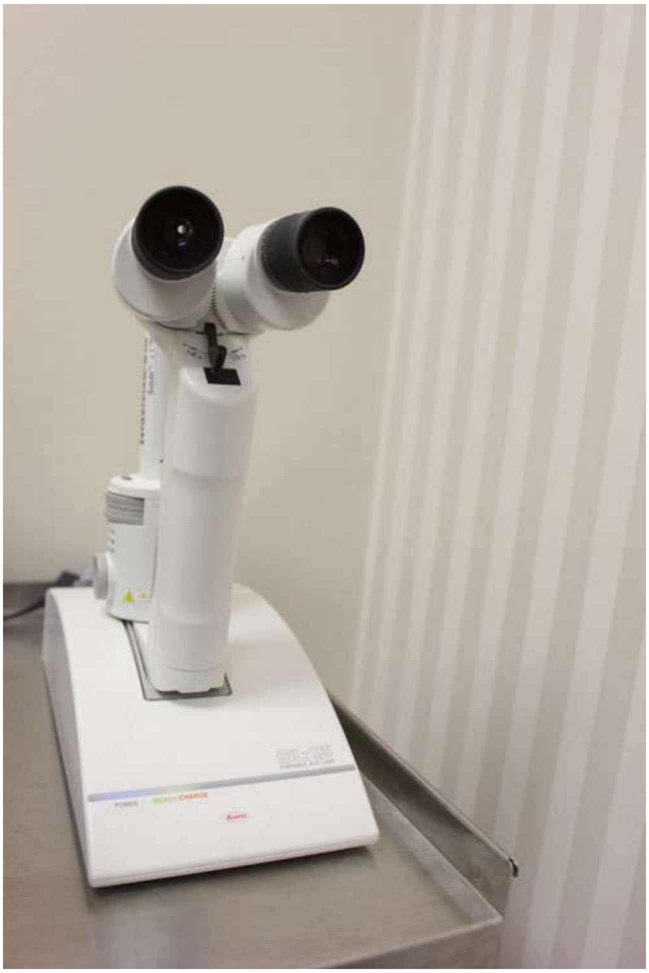

Fig. 46.2 A portable slit lamp

46.3 Materials

- The slit lamp is an eye-specific binocular stereoscope specifically designed to examine the eye and its structures and gives the practitioner a three-dimensional assessment of the eye.
- The operator has full movement of the microscope in all directions and the apparatus can be locked in place if required.
- Its light source can be manipulated to change the characteristics of the beam from its intensity to the angle at which it projects (Fig. 46.3).
  - A low-power setting is typically used for normal examination.
  - A higher-intensity beam is used when evaluating the anterior chamber with a narrow slit beam.
- It has colored filters typically built within the slit lamp.
  - Cobalt blue: Used with fluorescein dye to evaluate for corneal abrasions (Fig. 46.4) and avulsions; the dye will collect where the corneal epithelium is absent. It results in a yellow glow or hue visible through the microscope.
  - Green filter: Used to increase contrast of blood vessels. They appear black and the filter is useful for the assessment of hemorrhage.
- The operator is able to adjust the magnification of the microscopic typically through dial controls.
  - Low magnifications are most helpful for general examination.
  - Higher magnifications are used for examination of a particular area in fine detail.

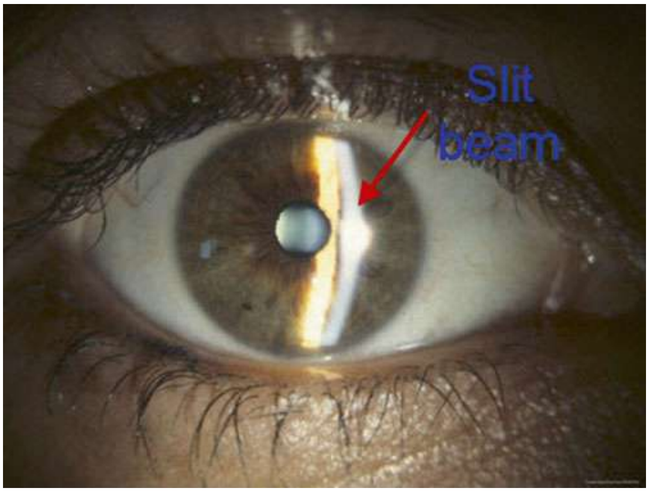

Fig. 46.3 The slit lamp beam and the reflection of the beam

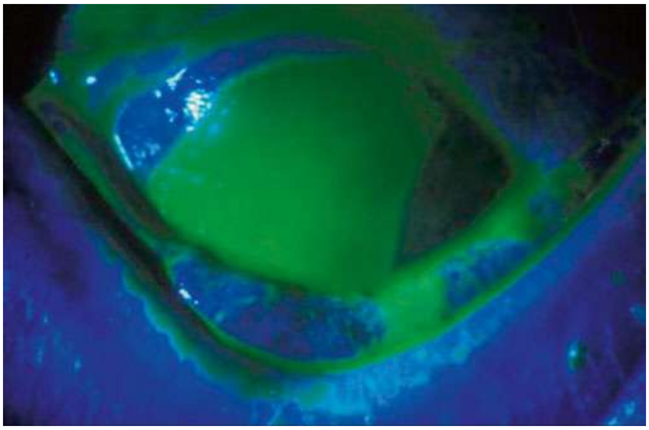

Fig. 46.4 Example of a corneal abrasion seen with a cobalt blue filter (With kind permission from Springer Science+Business Media: Das S, Chohan A, Snibson GR, Taylor HR. Capsicum spray injury of the eye. *International Ophthalmology*. 2005;26(4–5):171–3)

46.4 Procedure

1. As part of a complete eye examination, informed consent is not generally required.
2. An explanation of the procedure and reassurance are helpful.
3. Lock the slit lamp before positioning the patient because unintended movement may inadvertently damage the equipment or cause injury to the operator or patient.
4. The apparatus has a chin and a head rest. Ask the patient to place the chin and forehead in the respective areas (Fig. 46.5).
5. Adjust the height of the apparatus and the patient's chair for optimum comfort (Figs. 46.6 and 46.7). It may be necessary for children to stand during the examination.
6. The patient may place his or her hands on the sides of the table that the lamp is mounted on; this ensures stability of both the lamp and the patient.
7. Adjust the eyepieces of the microscope to fit the operator.
8. Turn on the slit lamp at its lowest setting to avoid a sudden power surge that could potentially damage the bulb.
9. Move the stage forward and narrow the beam and angle it at 45° to the patient. Aim the beam laterally so as not to cause the patient discomfort.
10. Focus the beam by manipulating the joystick to move the apparatus forward and backward so that the beam is clearly visible and its lines are sharp.
11. For each area of the eye to be examined, inspect the area thoroughly using the joystick to slowly manipulate the slit lamp across the eye in all directions, using the height adjustment of the joystick to slightly raise and lower the slit lamp as needed.
    - The operator may find that the slit lamp may move too freely; in that case she or he may find slightly tightening the locking nut of the C-arm may provide better and more precise control.

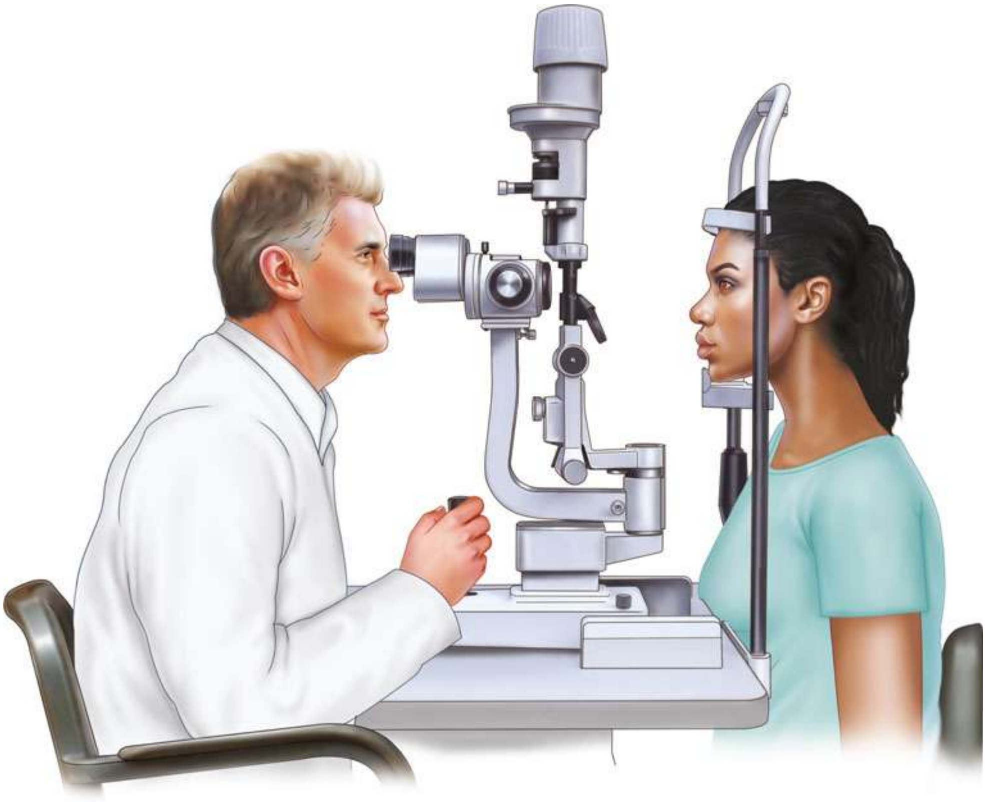

Fig. 46.5 Appropriate positioning for a slit lamp examination

Fig. 46.6 The slit lamp is too high

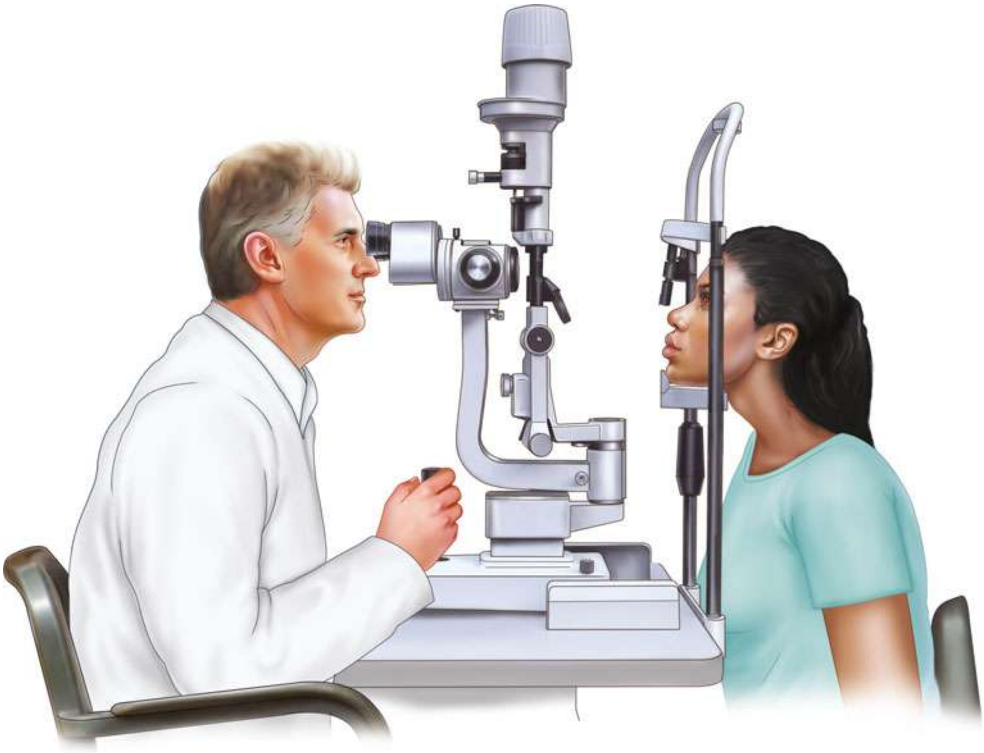

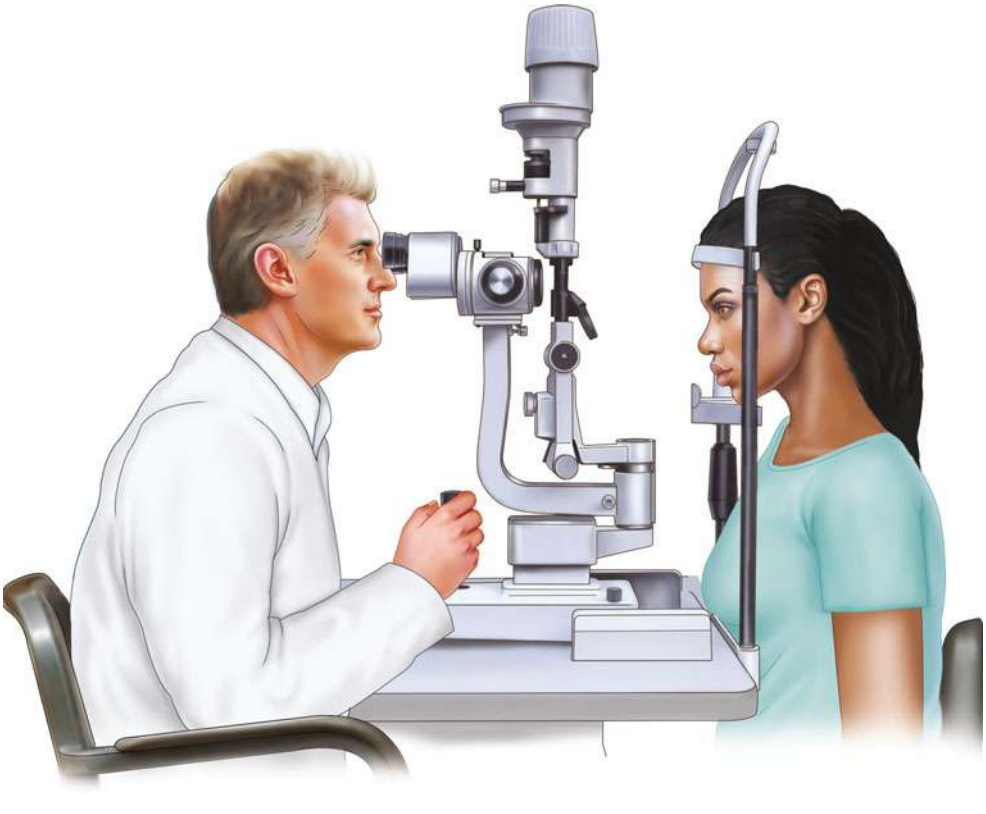

Fig. 46.7 The slit lamp is too low

46.5 Complications

There are no complications following routine slit lamp examination.

46.6 Pitfalls

- Ensure a working bulb.
- Appropriately set the microscope's oculars for the operator's pupillary distance.
- Ensure all dials and knobs are firmly set.
- Inadequate focusing
- Patient noncompliance

46.7 Fluorescein Testing

46.7.1 Indications and Uses

- Examination of the cornea under a cobalt blue filter on a slit lamp or Wood's lamp: The chemical fluoresces under blue light and appears green under these circumstances.
- Useful for detecting corneal abrasions, corneal foreign bodies, and infections of the eye.

46.7.2 Procedure

- Fluorescein is typically used after installation of a topical anesthetic, which provides patient comfort especially for those with significant pathology.
- The fluorescein strip comes packaged in single-use wrappers (Fig. 46.8).
- Hold the strip by the white end and wet the orange end *lightly* because heavily moistening the strip may cause a significant amount of dye to be present, obscuring the examination.
  - A dry strip may be used, but it may irritate the patient's eye especially if already sensitive.
- If this occurs, the clinician may use tissue paper to gently blot the excess solution away.
  - The exception to this is the performance of the *Seidel test*, which is used to assess the eye for potential perforation (Fig. 46.9).
  - The clinician will instill a large amount of dye into the eye by wetting the orange strip copiously.
  - The clinician will next examine the eye for a stream of fluid leaking from the ruptured globe.
  - This stream will fluoresce green or blue in distinction to the rest of the globe, which appears orange.
- The choice of solution to wet fluorescein strip is up to the clinician because saline, tap water, or the recently used anesthetic solution may be used safely.
- Place the now-wet orange end on the lower lid of the patient's eye.
- Ask the patient to blink several times to allow the solution to spread evenly.
- The clinician may use a Wood's lamp, penlight, or the cobalt blue filter on the slit lamp to examine the now-stained eye.
  - The slit lamp is preferable owing to the potential for missing small abrasions.

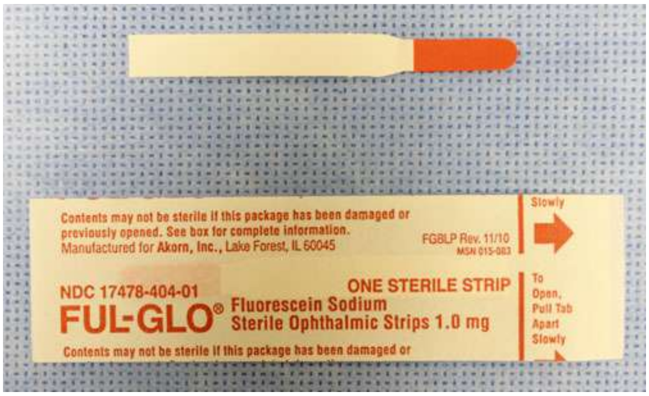
Contents may not be sterile if this package has been damaged or previously opened. See box for complete information.
Manufactured for Akorn, Inc., Lake Forest, IL 60045
FG8LP Rev. 11/10
MSN 015-083
Slowly
NDC 17478-404-01
ONE STERILE STRIP
FUL-GLO®
Fluorescein Sodium
Sterile Ophthalmic Strips 1.0 mg
To Open, Pull Tab Apart Slowly
Contents may not be sterile if this package has been damaged or

Fig. 46.8 A typical fluorescein strip package

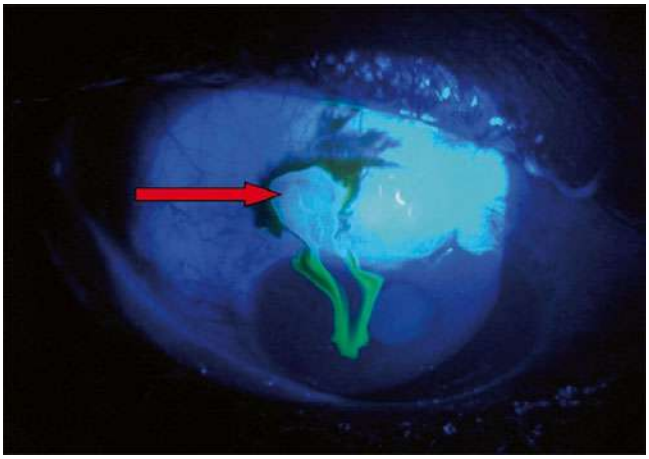

Fig. 46.9 A positive Seidel test (Reproduced with permission from: Lingam V, Panday M, George R, Shantha B. Management of complications in glaucoma surgery. *Indian J Ophthalmol*. 2011;59(Suppl1): S131–40)

46.7.3 Complications

- There is a theoretical risk of the development of superficial punctate keratitis from instillation of topical anesthetic before fluorescein testing.
- Discoloration of soft contact lenses.
- Potential for infection using premixed fluorescein solution.

46.7.4 Pitfalls

- Contact lens wearers should remove their lenses because the fluorescein will permanently stain the lens.
  - The wearer should not put the contacts back in for several hours.

Selected Reading

DuBois L. The slit lamp examination. In: DuBois L, Ledford JK, Daniels K, Campbell R, editors. Clinical skills for the ophthalmic examination: basic procedures. 2nd ed. Thorofare: Slack; 2006. p. 61–70.

Galor A, Jeng BH. Red eye for the internist: when to treat, when to refer. Cleve Clin J Med. 2008;75:137–44.

Lang GK. Ophthalmology: a short textbook. Stuttgart: Thieme; 2000.

Ledford JK, Sanders VN. The slit lamp primer. 2nd ed. Thorofare: Slack; 2006.

47

Eye Irrigation

Bobby K. Desai

47.1 Indications

- Chemical burns to the eye
- Removal of superficial foreign bodies

47.2 Contraindications

- For suspected globe perforation, extreme care must be taken to not exacerbate the injury.

47.3 Materials and Medications

- Irrigating device—the Morgan lens (Fig. 47.1)
- Saline solution or lactated Ringer's (preferably warmed)
- Topical anesthetic drops (Fig. 47.2)
- Basin to secure the extruded solution
- Intravenous (IV) tubing to attach the IV bag to the Morgan lens
- pH paper (Fig. 47.3)

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_47

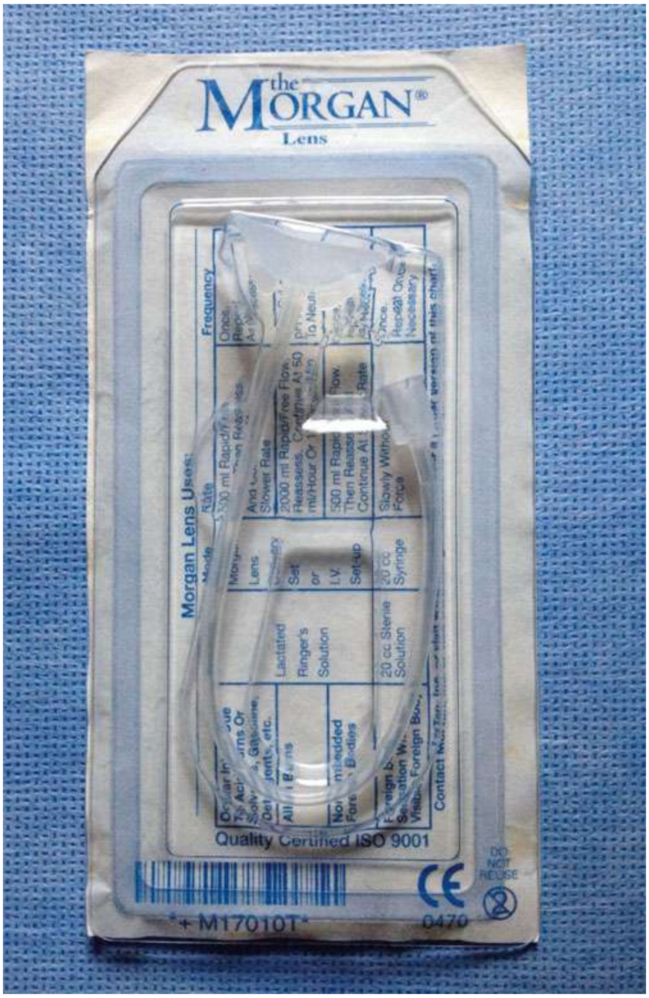
the MORGAN® Lens
Morgan Lens Uses:
Quality Certified ISO 9001
DO NOT REUSE
+ M17010T
0470

Fig. 47.1 The Morgan lens and packaging

Alcon®
Alcon Laboratories, Inc.
Ft. Worth, TX 76134 USA

Fig. 47.2 Examples of topical anesthetics

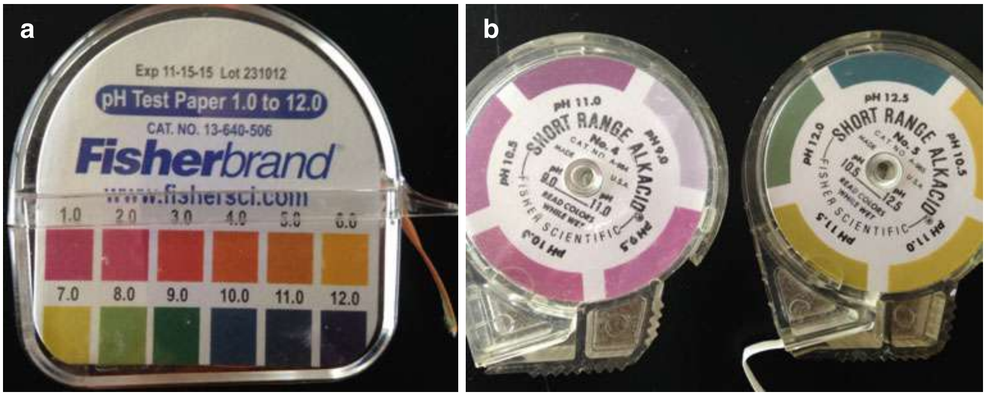
a
Exp 11-15-15 Lot 231012
pH Test Paper 1.0 to 12.0
CAT. NO. 13-640-506
Fisherbrand
www.fishersci.com
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
b
SHORT RANGE ALKACID No. 4
SHORT RANGE ALKACID No. 5
FISHER SCIENTIFIC

Fig. 47.3 (**a**, **b**) Examples of pH paper

47.4 Procedure

1. Informed consent is generally not required, although a thorough explanation to the patient is warranted.
2. Obtain pH of the eye *before* instilling anesthetic drops.
3. Anesthetize the eye(s) with topical anesthetic of choice by instilling drops within the lower lid and then asking the patient to blink several times in order to evenly distribute the solution.
4. Ensure there are no foreign bodies on the inside of the upper lids by inverting the upper lid.
    - Particulate foreign bodies may be removed with moistened cotton tip applicators.
5. After adequate anesthesia is ensured, place one end of the Morgan lens within the fornix of the upper lid (Fig. 47.4).
6. Next, gently retracting the lower lid will ensure smooth placement of the remaining portion of the Morgan lens (Fig. 47.5).
7. Using the end of the Morgan lens, screw in the prepared IV tubing (Fig. 47.6).
8. Attach the end of the IV tubing to the saline bag and place at height to allow for gravity to ensure a smooth flow of solution.
    - Continue to irrigate the eye until desired pH is obtained.
9. To remove the Morgan lens, use the opposite technique for insertion.

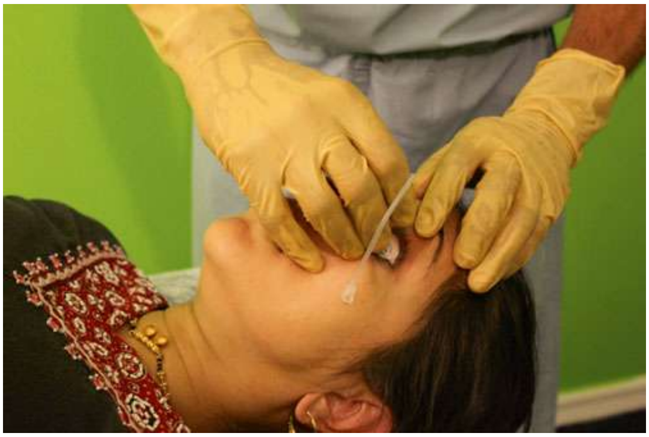

Fig. 47.4 Inserting the Morgan lens under the upper lid

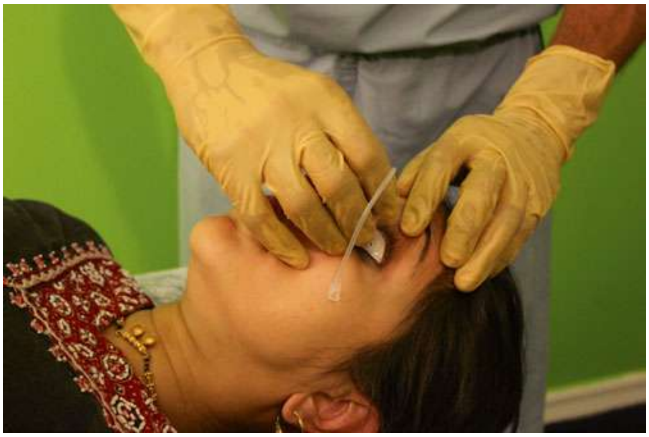

Fig. 47.5 Inserting the Morgan lens under lower lid

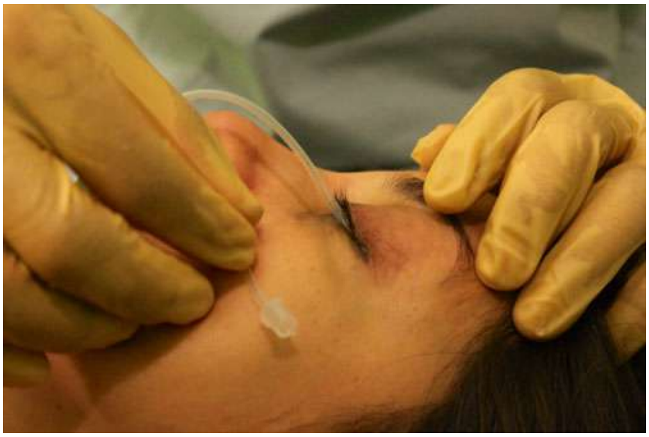

Fig. 47.6 The Morgan lens in place, ready to be attached to IV tubing for saline irrigation

47.5 Complications

- Corneal abrasions may be caused by the Morgan lens and are treated in the usual fashion.
- Deep corneal injury may occur with inadequate irrigation.

47.6 Pearls and Pitfalls

- Note that alkali burns will require significant irrigation and more topical anesthesia may be required.
- Ophthalmological consultation may be required, especially for alkaline and hydrofluoric acid burns.

Selected Reading

Lang GK. Ophthalmology: a short textbook. Stuttgart: Thieme; 2000.
Rhee DJ, Pyfer MF, Rhee DM, editors. The Wills Eye manual: office and emergency room diagnosis and treatment of eye disease. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 1999.

48

Corneal Foreign Body Removal

Bobby K. Desai

48.1 Indications

- Presence of a corneal foreign body (Fig. 48.1)

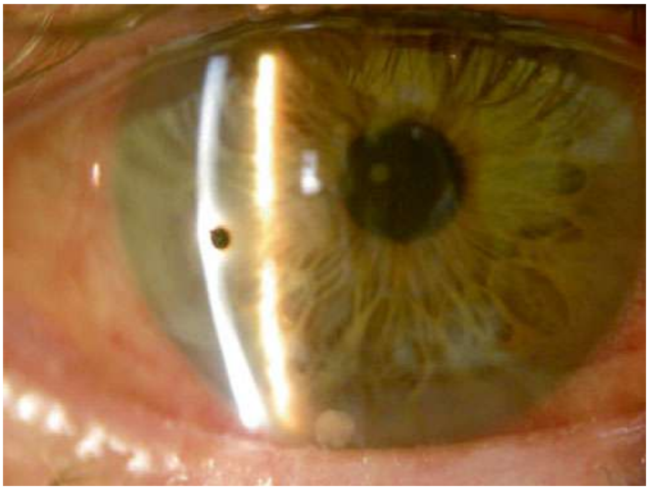

Fig. 48.1 Corneal foreign body (With kind permission from Springer Science + Business Media: Zuberbuhler B, Tuft S, Gartry D, Spokes D. Ocular Surface and Reconstructive Surgery. In: *Corneal Surgery*. 2013. 29–48)

48.2 Contraindications

- An uncooperative patient may require ophthalmological consultation as well as sedation.
- Suspected foreign bodies from high-velocity injuries must be referred to an ophthalmologist subsequent to initial evaluation.
- Any signs of globe penetration (e.g., hyphema) require emergent ophthalmological consultation.
- Overt globe rupture.
- Evidence of an inflammatory process such as iritis requires emergent ophthalmological consultation.
- Foreign bodies associated with corneal burns secondary to chemical exposure such as alkaline corneal burns will require emergent ophthalmological consultation.

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_48

48.3 Materials and Medications

- Eye chart for visual assessment
- Cotton tip applicator
- Slit lamp or other source of magnification
- 27-Gauge needle or ophthalmic foreign body needle (Fig. 48.2)
- Topical ophthalmic anesthetic
- Ophthalmic burr (Fig. 48.3)

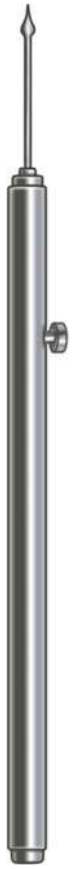

Fig. 48.2 Foreign body needle

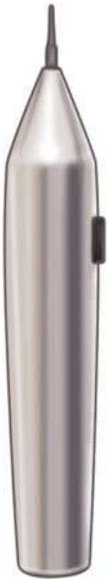

Fig. 48.3 Ophthalmic burr drill

48.4 Procedure

1. Informed consent may be required.
2. Visual acuity and a formal assessment of the eye should be done and documented.
3. Consider intraocular foreign body.
    - Computed tomography (CT) scan may be used to assess for the presence of an intraocular foreign body.
4. Eversion of the upper lid should be performed to evaluate for retained foreign body under the lid.
    - These usually can be removed by:
        - A moistened cotton tip applicator
        - Irrigation
5. Magnification is preferable, but some foreign bodies may be large enough to see with the naked eye.
6. Anesthetize the eye with topical anesthetic.
7. For multiple loose foreign bodies, removal may be attempted with irrigation after appropriate anesthesia.
8. Removal of the foreign body may be attempted with a moistened cotton swab.
    - Metallic foreign bodies recently embedded within the cornea may be difficult to remove in this fashion.
    - However, during the healing process, some metallic foreign bodies may be pushed closer to the surface during reepithelialization of the cornea. These potentially can be removed with this technique.
        - Care must be taken to avoid leaving a rust ring that could permanently stain the cornea.
9. A 27-gauge needle bent at a 90° angle may be used to remove foreign bodies by gently prying it out.
    - An advantage of using a needle is that any rust ring as well as any metallic foreign body can potentially be removed.
10. Technique: A slit lamp may be used to magnify the area.
    - See slit lamp chapter (Chap. 46) for details.
    - The patient should be informed about the procedure; formal informed consent may not be required.
    - Proper positioning is critical for success.
    - The patient's head should be fully forward and firmly placed against the head rest.
    - The patient's hands may be placed on the sides of the slit lamp stage for stabilization.
    - Apply topical anesthesia to the cornea.
11. Using the patient's cheek as a bolster potentially avoids significant movement of the needle if the patient unexpectedly moves.
    - Other modalities for stabilization include supporting the elbow by placing on a box or using towels.
12. Have the patient gaze at one point in the far distance.
    - Using the needle or burr as a scoop, gently manipulate the foreign body out.
        - Using a burr can allow for the total removal of any rust ring.
13. Once the foreign body is removed, patching is not required.
14. Consider application of topical antibiotics.
15. Tetanus prophylaxis should be given as for open other wounds.
16. Arrange follow-up with a primary care physician or preferably an ophthalmologist.

48.5 Complications

- Forceful attempts to dislodge a foreign body may result in corneal perforation.
- Incomplete removal of a foreign body.

Selected Reading

Lang GK. Ophthalmology: a short textbook. Stuttgart: Thieme; 2000.

Rhee DJ, Pyfer MF, Rhee DM, editors. The Wills eye manual: office and emergency room diagnosis and treatment of eye disease. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 1999.

Thomas SH, White BA. Foreign bodies. In: Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010. p. 715–32.

Ultrasound Evaluation of Retinal Detachment

49

Shalu S. Patel, L. Connor Nickels, and Rohit Pravin Patel

49.1 Indications

- To aid in the evaluation of vision loss or change in vision

49.2 Contraindications

- Relative: Suspicion of increased intraocular pressure
- Relative: Suspicion of globe rupture (see Chap. 50)

49.3 Materials and Medications

- Bedside ultrasound machine with high-frequency (7.5- to 10-MHz) linear transducer
- Transparent adhesive such as Tegaderm (optional)
- Sterile ultrasound gel
- Sterile gauze

49.4 Procedure

1. Position the patient supine with his or her eyes closed. The eyelid may be taped closed with a transparent adhesive if desired (Fig. 49.1).
2. Place a liberal amount of ultrasound gel over the eyelid.
3. Place the ultrasound probe over the eye in a transverse position.
4. Adjust the depth of the ultrasound so that the whole eye fills the screen.
5. Scan through the eye fully in the transverse and sagittal planes.
6. When viewing the ultrasound image, the normal eye is a circular hypoechoic structure (Fig. 49.2). The structures should be evaluated from anterior to posterior.
7. Identify the cornea. This is the first thin hypoechoic line in the anterior eye.
8. Continuing posteriorly, identify the anterior chamber, which is an anechoic area bordered by the cornea, iris, and lens. The iris is an echogenic linear structure, and the normal lens is anechoic.
9. Identify the vitreous chamber, the large anechoic region posterior to the lens.
10. Carefully evaluate the posterior globe. The normal retina cannot be distinguished from the other choroidal layers on ultrasound.
11. A detached retina will appear as a hyperechoic linear floating membrane in the posterior vitreous chamber (Fig. 49.3).
12. In the retrobulbar region, the optic nerve can be identified as a hypoechoic linear structure perpendicular to the globe (Fig. 49.3).

S.S. Patel, MD
Department of Emergency Medicine, Florida Hospital Tampa,
Florida Hospital Carrollwood, Tampa, FL, USA
e-mail: shalu314@gmail.com

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_49

Fig. 49.1 Probe positioning with Tegaderm applied to orbit: (**a**) sagittal view, (**b**) axial view

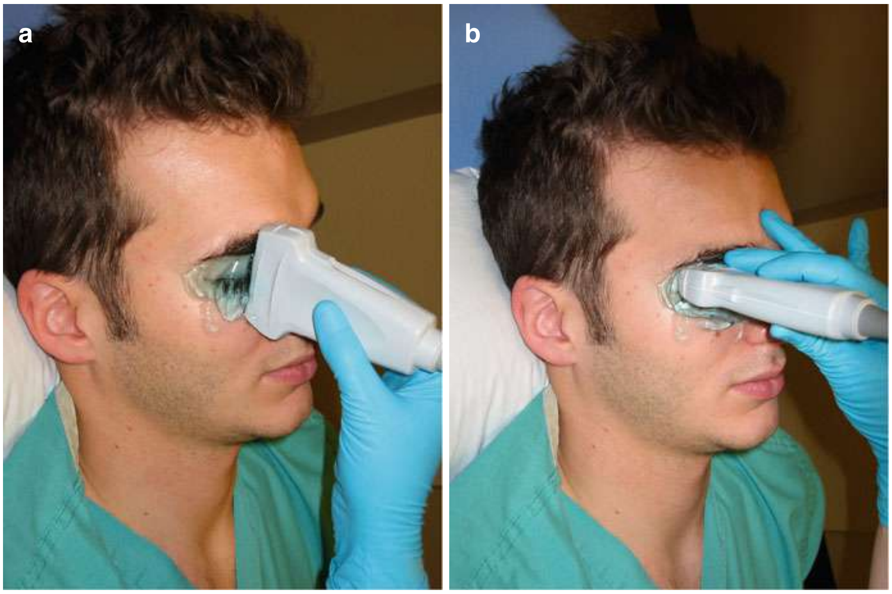

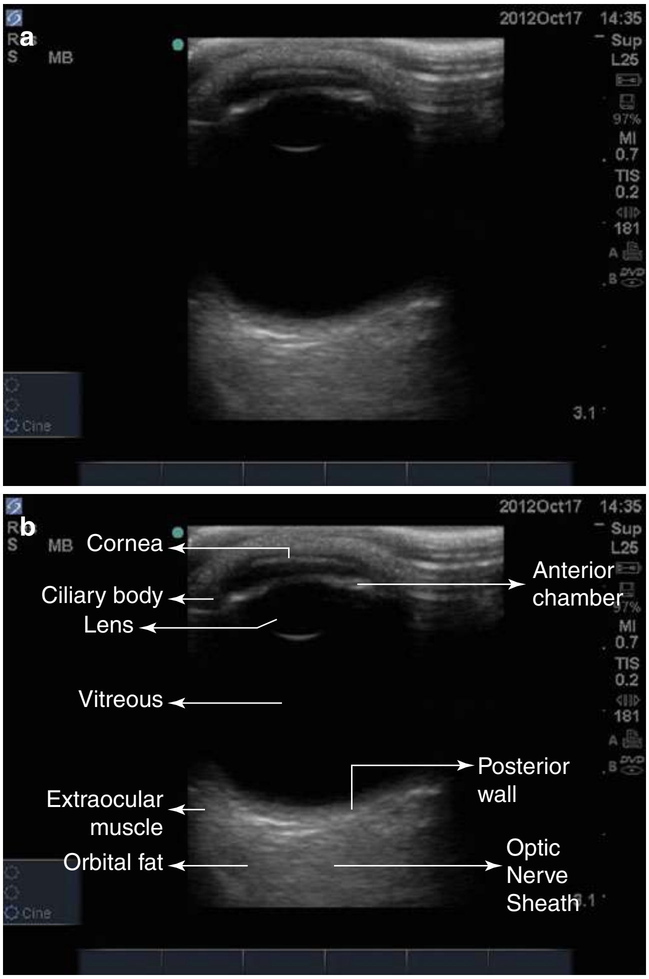
a
b
2012Oct17 14:35
Sup
L25
MB
Cornea
Ciliary body
Lens
Anterior chamber
Vitreous
Posterior wall
Extraocular muscle
Orbital fat
Optic Nerve Sheath
Cine

Fig. 49.2 Ocular ultrasound with normal anatomy findings ((**a**) without labels, (**b**) with anatomy labeled)

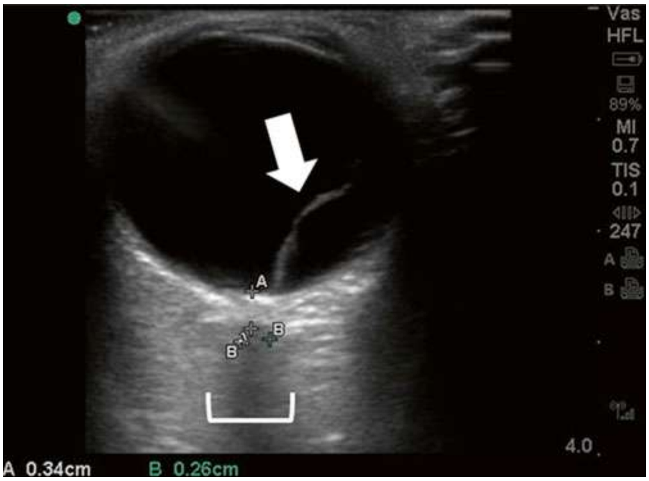
Vas
HFL
A
B
A 0.34cm
B 0.26cm
4.0

Fig. 49.3 Retinal detachment appears as a hyperechoic linear structure (*white arrow*). *White* bracket shows the optic nerve sheath shadow (see Chap. 50 for more information)

49.5 Complications

- Conjunctival infection
- Increased intraocular pressure (if too much pressure is applied)

49.6 Pearls and Pitfalls

- Pearls
  - Use a transparent adhesive such as Tegaderm to decrease the risk of conjunctival infection.
  - Use a liberal amount of gel to obtain the best images.
  - Use minimal pressure to obtain the best images.
  - False positives on ocular ultrasound may occur with disk edema or vitreous hemorrhage.
- Pitfalls
  - Subacute retinal detachments may be missed on ocular ultrasound.

Selected Reading

Blaivas M. Bedside emergency department ultrasonography in the evaluation of ocular pathology. Acad Emerg Med. 2000;7:947–50.

Blaivas M, Theodoro D, Sierzenski PR. Elevated intracranial pressure detected by bedside emergency ultrasonography of the optic nerve sheath. Acad Emerg Med. 2003;4:376–81.

Shinar Z, Chan L, Orlinksy M. Use of ocular ultrasound for the evaluation of retinal detachment. J Emerg Med. 2011;4:53–7.

Whitcomb MB. How to diagnose ocular abnormalities with ultrasound. AAEP Proc. 2002;48:272–5.

Ultrasonography in the Evaluation of Intraocular Pathology

50

Benjamin M. Mahon, Marie-Carmelle Elie,
L. Connor Nickels, and Rohit Pravin Patel

50.1 Indications

- Eye trauma
- Loss of vision or decreased vision
- Head injury
- Ocular pain
- Suspected foreign body

50.2 Contraindications

- None

50.3 Materials

- Linear probe (7.5–10 MHz)
- Tegaderm or other protective eye covering
- Copious amounts of water-soluble transmission gel

50.4 Procedure

1. Begin by asking the patient to close her or his eye, and then apply the Tegaderm or other suitable eye cover over both eyes.
2. Apply a copious amount of transmission jelly over both eyes.
3. Gently apply the ultrasound probe to the outer eyelid. If the operator has used a sufficient enough amount of jelly, the probe itself should not need to be in actual contact with the eyelid. This becomes of critical importance in suspected globe trauma or intraocular foreign body, when excess manipulation of the intraocular contents could precipitate further damage or worsening of present damage. Ensure the pressure is gentle enough to prevent this, but sufficient to obtain good imaging (see Chap. 49).
4. Image the normal eye first, to become familiar with the patient's anatomy (see Chap. 49), before proceeding to the suspected abnormal eye.
5. Start with low gain.
6. Ask the patient to move the eye slowly from the left to the right. This is called kinetic echography and will provide information elaborated upon later. Have the patient keep her or his eye straight ahead, as the operator phases up and down and side to side, between the transverse and the sagittal planes.
7. Increase gain slowly while the patient is moving the eye. Shifting from low gain to higher gain during the examination will help to prevent overlooking certain disease processes. As a fluid-filled structure, the eye provides its own acoustic window, and imaging is generally easy.

50.5 Complications

- Through excessive pressure applied to the eye, as described previously, further disruption of the intraocular contents and worsening of preexisting injury can occur.

50.6 Pearls and Pitfalls

- Pitfalls
  - Applying too much pressure to the globe and disrupting already damaged intraocular contents
  - Failing to image with full spectrum of gain
  - Using an insufficient amount of gel

B.M. Mahon, MD
Ponciana Medical Center, Kissimmee, FL, USA
e-mail: faust2k9@ufl.edu

M.-C. Elie, MD, RDMS (✉) • L.C. Nickels, MD, RDMS
R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: elie@ufl.edu; cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_50

50.7 Specific Disease Processes/Pathology [1]

50.7.1 Vitreous Hemorrhage (Fig. 50.1)

- Best seen in high gain.
- Clinical: Very common pathology. Seen as "floaters" on the visual field. Can cause blindness if large.
- Ultrasound: Appearance may vary. Early on in the evolution of a hemorrhage, these can be seen as small dots or mobile linear opacities that float freely in the back of the posterior chamber with eye movement. Has been described as a "snow-storm appearance." As hemorrhage ages, the blood organizes and membranes form, sometimes layering inferiorly owing to gravity.

50.7.2 Vitreous Detachment

- Best seen with higher gain.
- Clinical: Common in older patients. Presents as "flashers." Vitreous gel loses its attachment from the internal membrane lining.
- Ultrasound: Seen as "swaying seaweed." There are layering opacifications. Also seen as "snow-globe appearance." Concurrent vitreous hemorrhage is also common.
- Note: A small percentage of patients also have a retinal tear, so ophthalmology follow-up is necessary.

50.7.3 Retinal Detachment (Figs. 50.2 and 50.3)

- Best seen with lower gain.
- Clinical: Acute loss of vision, usually painless, often preceded by "flashers" and/or "floaters" as the neurosensory component of the retina pulls away from the retinal pigment epithelium. *An ophthalmological emergency.* Can be caused by ocular trauma or no known precipitant at all.
- Ultrasound: Seen as a taut, hyperechoic, linear opacity that moves with the eye, attached to the optic disc, in the posterior region of the globe.

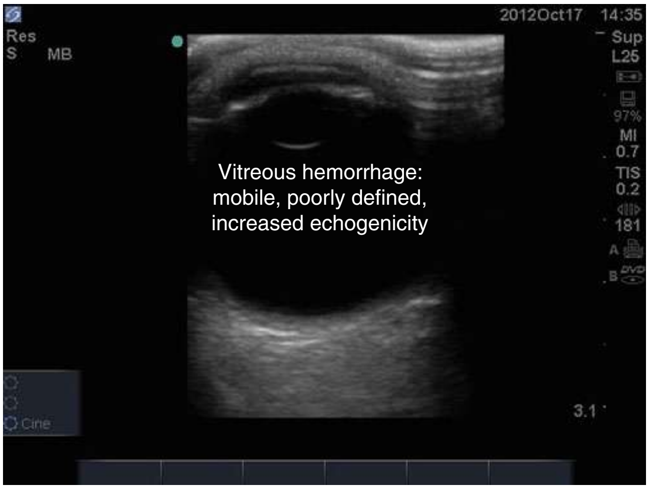

Vitreous hemorrhage: mobile, poorly defined, increased echogenicity

Fig. 50.1 Vitreous hemorrhage

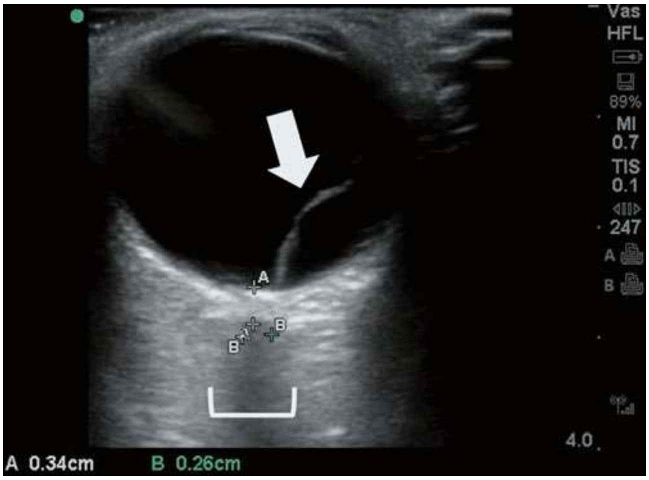

Fig. 50.2 Retinal detachment ultrasound image

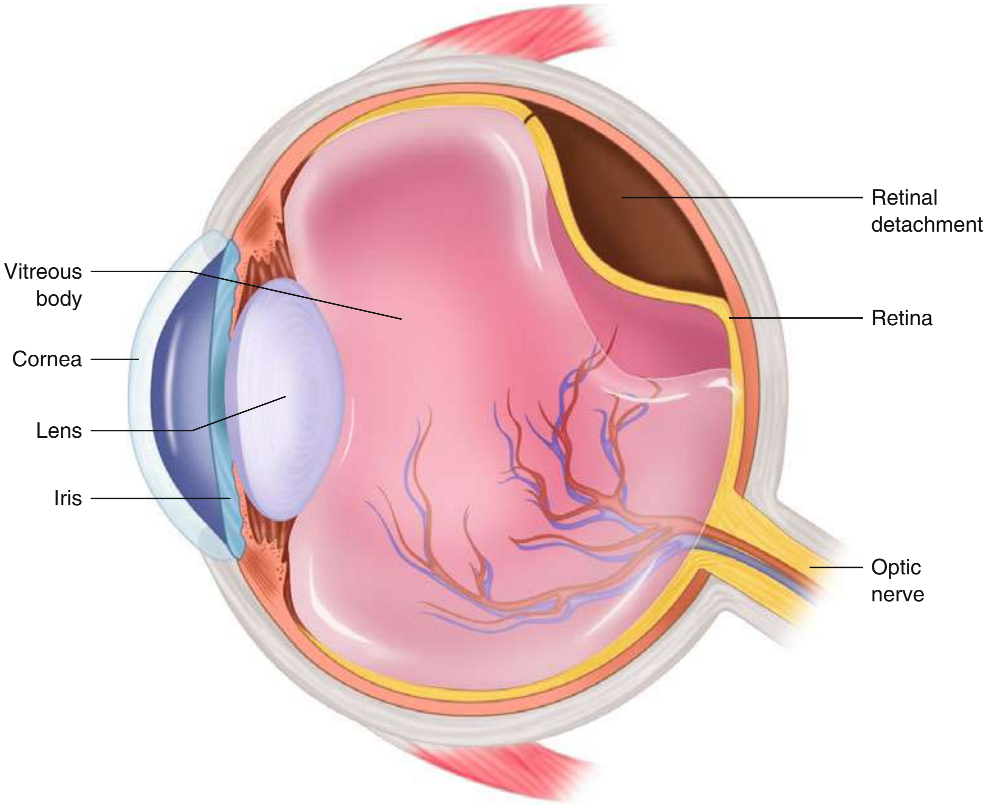
Vitreous body
Cornea
Lens
Iris
Retinal detachment
Retina
Optic nerve

Fig. 50.3 Retinal detachment graphical representation

50.7.4 Intraocular Foreign Body (Fig. 50.4)

- Clinical: Usually a preceding episode of trauma, often from metalworking or landscaping.
- Ultrasound: Foreign bodies are seen as hyperechoic substances in the eye often with distal shadowing. Certain shadow patterns distal to the foreign body, as well as comet tails, can help to differentiate the type of foreign body material.

50.7.5 Globe Rupture (Fig. 50.5)

- Clinical: Usually the result of trauma. Frequently, the mechanism causing the injury has subsequently made the eye examination itself impossible, if edema, blepharospasm, hyphema, or distorted anatomy makes visualization of the posterior chamber too difficult. In these circumstances ultrasound can be immeasurably helpful.
- Ultrasound: Findings include decreased size of the anterior chamber, posterior chamber, or both, often with buckled sclera. Globe rupture should be suspected when the injury mechanism is present and any major distortion of the ocular anatomy is appreciated.

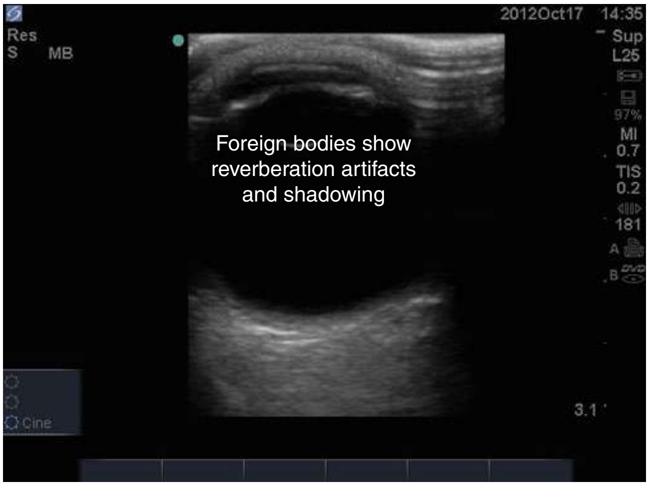
2012Oct17 14:35
Res
S MB
Sup
L25
97%
MI
0.7
TIS
0.2
181
Foreign bodies show reverberation artifacts and shadowing
Cine
3.1

Fig. 50.4 Foreign bodies

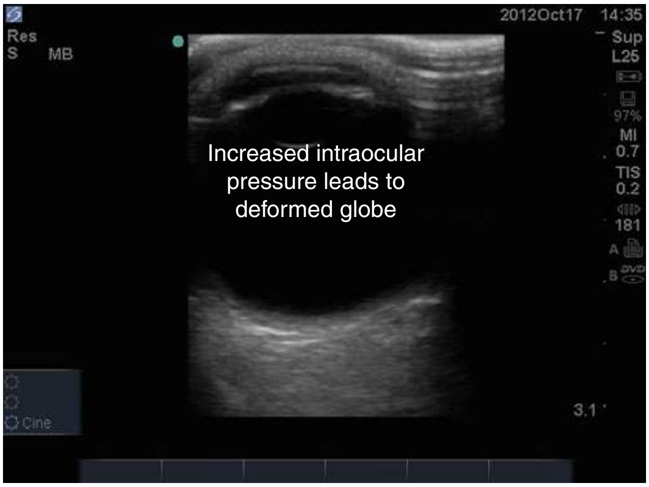
2012Oct17 14:35
Res
S MB
Sup
L25
97%
MI
0.7
TIS
0.2
181
Increased intraocular pressure leads to deformed globe
Cine
3.1

Fig. 50.5 Globe rupture

50.7.6 Increased ICP

- Clinical: Increased ICP is encountered in a number of pathological processes, such as closed head injury from trauma, altered mental status, pseudotumor cerebri, space-occupying lesion, or any other process causing intracranial injury. Physical examination may be suggestive of an intracranial process, but often key examination findings such as papilledema are not appreciated, owing to both limitations from the patient and their mental status, as well as limitations in the skill of the ED doctor, or the ability to dilate the eyes sufficiently, to appreciate the finding.
- Ultrasound: The size of the *optic nerve sheath diameter* is an indirect measurement of ICP. It is visualized as a hypoechoic linear strip radiating posteriorly from the edge of the back of the globe. A normal diameter is 5.0 mm or less. This measurement is taken at 3 mm back from the posterior edge of the globe (Fig. 50.6). Take measurements bilaterally to compare. Two or more measurements should be taken and then averaged on each side.
- A diameter greater than 5.0 mm, at 3.0 mm distal from the posterior globe border, is strongly suggestive of increased ICP and warrants further evaluation and imaging [2].

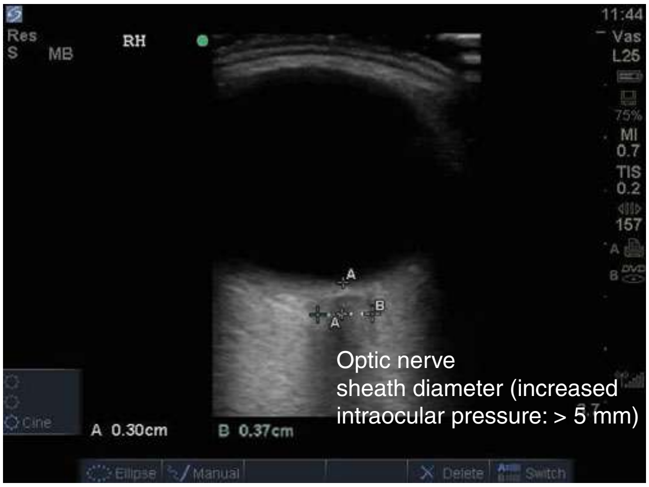

Optic nerve
sheath diameter (increased
intraocular pressure: > 5 mm)

Fig. 50.6 Optic nerve sheath measurement

References

1. Blaivas M, Theodoro D, Sierzenski P. A study of bedside ocular ultrasonography in the emergency department. Acad Emerg Med. 2002;9:791–9.
2. Blaivas M, Theodoro D, Sierzenski P. Elevated intracranial pressure detected by bedside emergency ultrasonography of the optic nerve sheath. Acad Emerg Med. 2003;10:376–81.

Selected Reading

Mustafa M, Montgomery J, Atta H. A novel educational tool for teaching ocular ultrasound. Clin Ophthalmol. 2011;5:857–60.

Qureshi MA, Laghari K. Role of B-scan ultrasonography in preoperative cataract patients. Int J Health Sci (Qassim). 2010;4:31–7.

Shinar Z, Chan IL, Orlinsky M. Use of ocular ultrasound for the evaluation of retinal detachment. J Emerg Med. 2011;40:53–7.

Vodapalli H, Murthy SI, Jalali S, Ali MJ, Rani PK. Comparison of immersion ultrasonography, ultrasound biomicroscopy and anterior segment optical coherence tomography in the evaluation of traumatic phacoceles. Indian J Ophthalmol. 2012;60:63–5.

Whitcomb MB. How to diagnose ocular abnormalities with ultrasound. AAEP Proc. 2002;48:272–5.

Yoonessi R, Hussain A, Jang TB. Bedside ocular ultrasound for the detection of retinal detachment in the emergency department. Acad Emerg Med. 2010;17:913–7.

Tonometry

51

Bobby K. Desai

51.1 Indications

- To measure intraocular pressure (IOP)

51.2 Contraindications

- A relative contraindication is the presence of an active or indolent infection about the cornea or conjunctiva, unless a one-time-use device is used (e.g., a Tono-Pen XL).
- Recent trauma to the eye.
- Uncooperative patients because improper technique may cause damage to the eye.

51.3 Overview

- Tonometry is the measurement of IOP obtained by evaluating the resistance of the eye to indentation by a force applied to it.
- Can be obtained by several methods.
- Two methods are routinely used in an emergency department setting:
  - *Impression tonometry*: measures the indentation made by a plunger carrying a known amount of weight.
  - More weight can be added to the apparatus.
  - The more weight needed to indent the cornea results in a higher IOP reading.
    - The Schiøtz tonometer is the most commonly used apparatus to utilize this method (Fig. 51.1).
- *Electronic indentation tonometry*: does not exert pressure on the eye (Fig. 51.2).

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_51

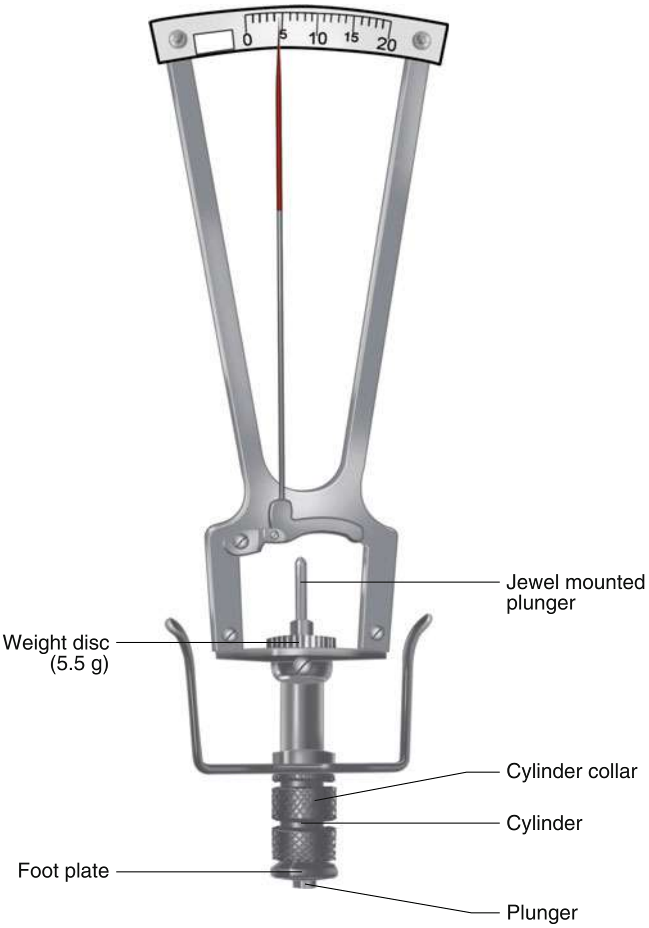
0
5
10
15
20
Jewel mounted plunger
Weight disc (5.5 g)
Cylinder collar
Cylinder
Foot plate
Plunger

Fig. 51.1 Schiøtz tonometer

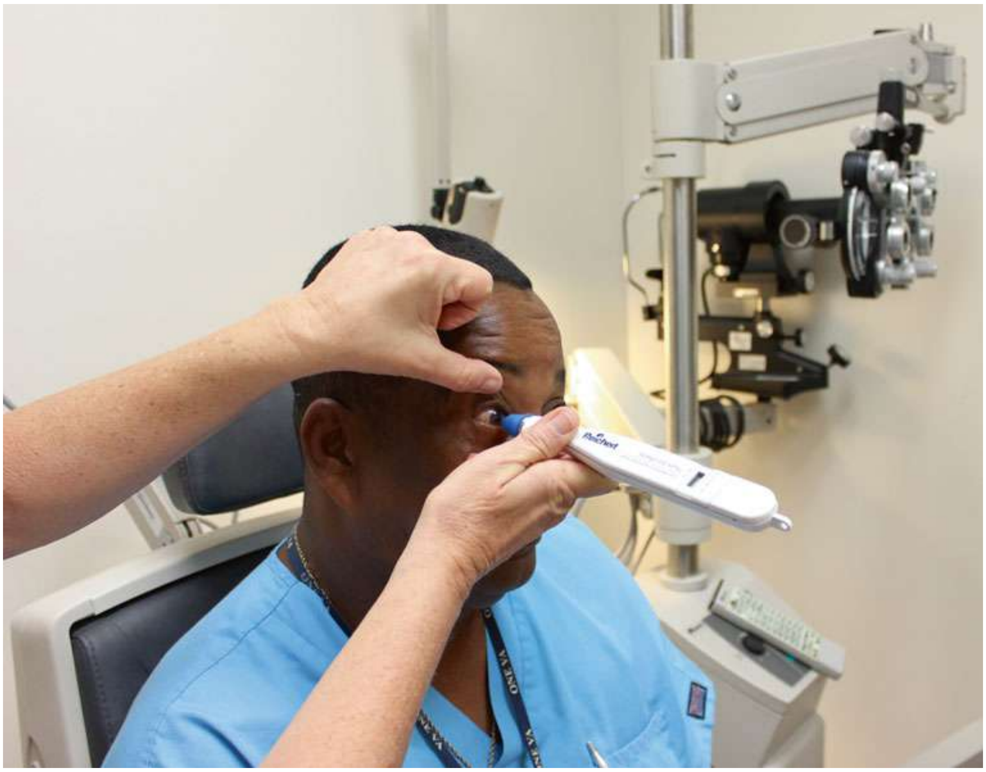

Fig. 51.2 Tono-Pen

51.4 Schiøtz Tonometry

51.4.1 Procedure: Schiøtz Tonometry (Fig. 51.3)

1. Carefully clean the apparatus between each patient.
    - Rubbing alcohol may be used.
    - Because it is metal, it can be autoclaved.
2. Calibrate the Schiøtz tonometer with the metal test block provided.
    - Test by placing the apparatus directly on the metal block.
    - The needle should be at "0" at the far end of the scale.
        - If not, loosen the screw at the base of the needle to rezero it.
        - Ensure that the needle is completely straight because any bend will produce an erroneous reading.
3. An explanation of the procedure is helpful because patient cooperation is critical for accurate results.
4. Anesthetize the eye with topical ophthalmic anesthetic of choice.
5. The patient should be in a recumbent position.
    - Have the patient focus on an area of the ceiling.
6. Hold the instrument with the aid of the curved arms at the side of the tonometer.
    - The operator can rest her or his hand on the patient's cheek or forehead to maintain stability.
7. Gently rest the tonometer on the patient's eye such that the instrument is centered on the eye and the instrument is completely vertical; no pressure should be exerted on the eye.
8. Note the scale reading.
9. Lift the Schiøtz directly off the cornea to avoid injury.
10. Using the table provided with the Schiøtz, the operator may convert the scale reading into the IOP.
    - The scale is inversely proportional to the actual IOP.
    - If the scale is low (i.e., high IOP), the additional weights provided with the instrument may be used and the patient retested.

51.4.2 Pitfalls

- Ensure the plunger is clean because it can transmit infection.
- False readings may be obtained without proper calibration.
- Placing pressure on the instrument will cause false readings.

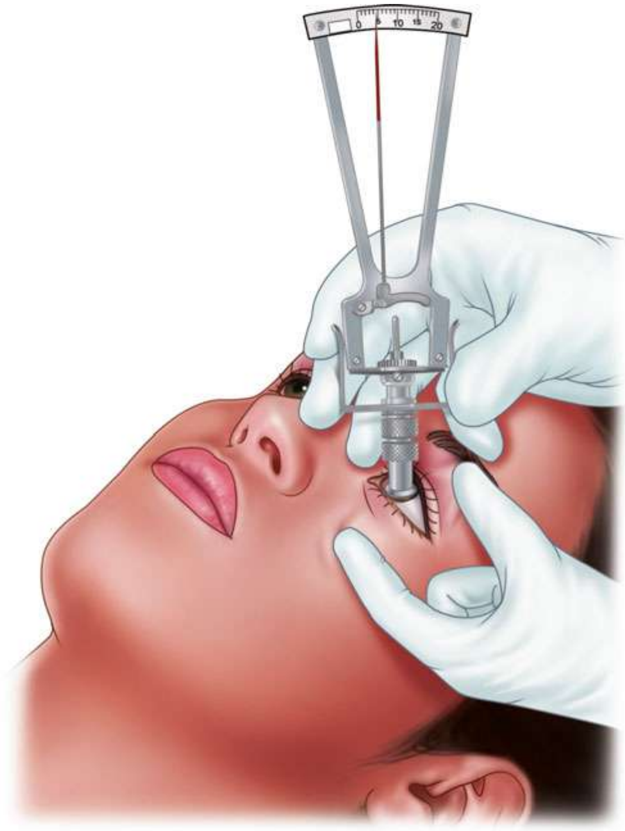

Fig. 51.3 Correct use of Schiøtz tonometer

51.5 Tono-Pen Tonometry

51.5.1 Overview

- Electronically measures IOP
  - Combines applanation and indentation tonometry
- Uses pressure-sensitive electronics to average four successive readings and displays the reading and a reliability factor digitally

51.5.2 Calibration of the Tono-Pen

- Should be performed once daily
- Hold the instrument with the tip down (Fig. 51.4).
- Press the black button twice.
  - The letters “CAL” will appear in the display.
- Press the black button again.
  - After a few seconds the word “UP” will appear.
- Rotate the instrument so that the tip points upward.
  - If “GOOD” appears, the Tono-Pen is ready for use.
  - If “BAD” appears, repeat the process until “GOOD” appears.
    - If “GOOD” does not appear, the device cannot be used.

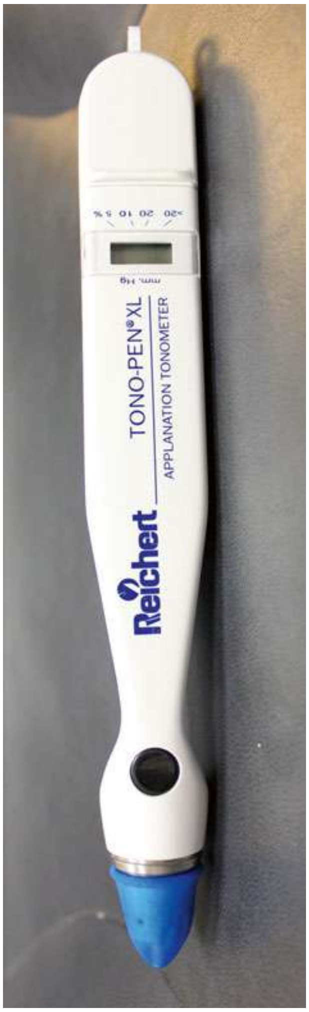

Fig. 51.4 Tono-Pen calibration position

51.5.3 Procedure: Tono-Pen Tonometry

1. An explanation of the procedure is helpful because patient cooperation is critical for accurate results.
2. Anesthetize the eye with the topical ophthalmic anesthetic of choice.
3. Place a probe cover on the unit.
4. The patient should be in a comfortable position because the unit can be used in any position.
    - Have the patient focus on a specific area.
5. The best way to hold the instrument is similar to that of a pen.
    - Ensure that the digital readout is visible.
6. The operator can rest his or her hand on the patient's cheek or forehead to maintain stability.
7. Hold the unit perpendicular to the patient's cornea.
8. Press the black button only once.
    - If "ICALI" is seen, followed immediately by a single row of dashes [– – – –], it indicates that the Tono-Pen requires calibration before it will measure.
    - If "====" is seen and a "beep" is heard, the unit is ready.
    - Proceed with measurement.
9. Gently touch the Tono-Pen to the cornea and withdraw (Fig. 51.5). Repeat several times.
    - Indentation is not required.
10. The unit will "chirp" and a digital reading will be displayed if a valid reading is obtained.
11. After four valid readings are obtained, the average of these measurements as well as a single bar that signifies statistical reliability will appear on the readout.
    - After the final beep if the liquid crystal display (LCD) readout shows "----,“ not enough valid readings were obtained.
    - In this case, the measurement must be repeated.

51.5.4 Pitfalls

- Unsuccessful calibration mandates a repeat attempt at calibration.
- Loosen the Ocu-Film tip cover and repeat for multiple failed attempts.
- Press the reset button and reattempt calibration.
- Replace the battery if necessary.
- If all else fails, use another device or Schiøtz tonometer.

Fig. 51.5 Proper use of Tono-Pen

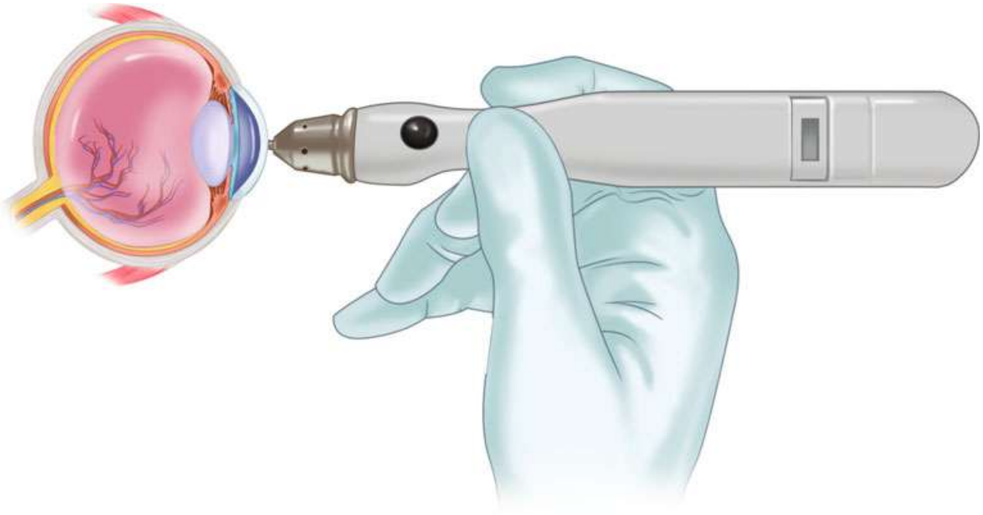

Correct

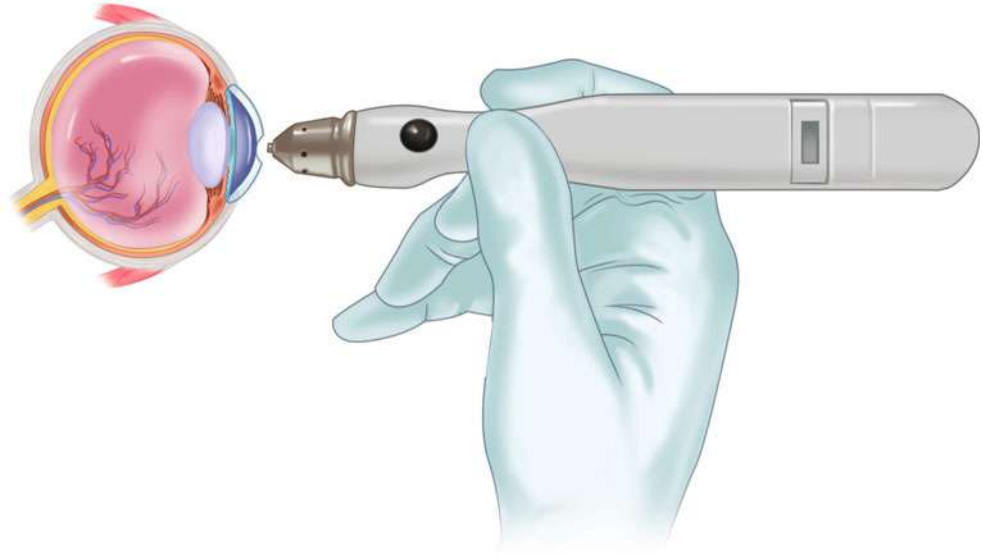

Incorrect

51.6 Complications

- Unusual if proper technique is used.
- If too much force is applied, a corneal abrasion may result.
- Infection when using an improperly sterilized Schiøtz tonometer.

Selected Reading

Lang GK. Ophthalmology: a short textbook. Stuttgart: Thieme; 2000.

Rhee DJ, Pyfer MF, Rhee DM, editors. The Wills eye manual: office and emergency room diagnosis and treatment of eye disease. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 1999.

Lateral Canthotomy

52

Benjamin M. Mahon and Bobby K. Desai

52.1 Indications

- Absolute indications: In the presence of presumed or confirmed retrobulbar hemorrhage
  - Acute visual loss
  - Intraocular pressure (IOP) greater than 40 mmHg (normal IOP is 10–20 mmHg)
  - Proptosis not amenable to retropulsion
- Relative indications: In the context of presumed or confirmed retrobulbar hemorrhage (Fig. 52.1)
  - Ophthalmoplegia
  - Cherry red macula
  - Profound eye pain
  - Afferent pupillary defect (Marcus Gunn pupil)
    - This defect is seen with the *swinging flashlight test.*
      - Shine a light into both eyes; a normal response is equal constriction of both pupils.
      - In those patients with an afferent papillary defect, when light is swung from an unaffected pupil to an affected pupil, the pupil will seem to paradoxically dilate, rather than constrict. The pupil on the nonaffected side will similarly dilate as light is shown into the affected eye. This results from injury to the afferent fibers of cranial nerve (CN) II on the affected side, while the efferent fibers, innervated by CN III, remain intact.
  - This procedure is most effective if performed as soon as possible because irreversible vision loss can occur secondary to ischemia in as little as 90 min.

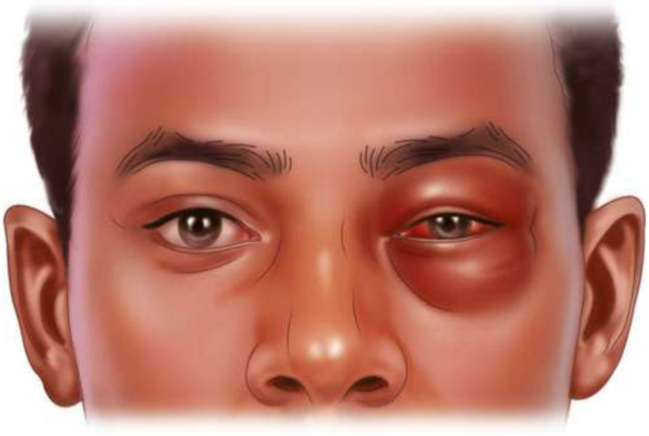

Fig. 52.1 Retrobulbar hemorrhage

B.M. Mahon, MD
Ponciana Medical Center, Kissimmee, FL, USA

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_52

52.2 Contraindications

- Absolute
  - Globe rupture
- Relative
  - Suspected globe rupture: Heralded by a number of signs including hyphema, irregularly shaped pupil, exposed uveal tissue, or profound restriction of extraocular movement

52.3 Materials and Medications

- Straight hemostat
- Suture or iris scissors
- Forceps
- 1–2 % lidocaine with epinephrine
- Sterile gloves
- Sterile towels/drapes
- 4×4 gauze
- Face shield/mask
- Large-bore needle to withdraw anesthetic, 25-gauge needle to infiltrate
- 10-cc syringe
- Normal saline

52.4 Procedure

1. Position the patient
   - While waiting for the procedure to begin, the patient should initially be upright to produce any decrease in IOP that can be provided.
   - At the start of the procedure, lay the patient supine.
   - A cooperative patient is absolutely necessary because even slight movements can cause devastating iatrogenic injury.
   - If the patient is anxious or conscious, sedation may be required.
   - In extreme cases of altered mental status or combative trauma patients, endotracheal intubation and mechanical ventilation may be required.
2. Gently irrigate the affected eye to remove any debris.
3. Inject approximately 1–2 mL of 1–2 % lidocaine with epinephrine into the lateral canthus of the affected eye (Fig. 52.2).
4. Using a curved hemostat, gently crimp the skin over the lateral corner of the patient's eye down to the orbital rim for 3 min to establish hemostasis and set the boundaries for the incision.
   - Using forceps, use the hemostats to pick up the skin just crushed.
   - Raise the skin with forceps and then use scissors to make a 1- to 2-cm incision from the lateral corner of the eye extending laterally outward (Fig. 52.3).
     - This incision can often sufficiently decrease IOP.
     - It is feasible to remeasure the IOP at this point, and if it is still greater than 40 mmHg, proceed to the next step.
   - Visualize the lateral canthal tendon by retracting the inferior orbital lid inferiorly.
   - Direct the scissors along the lateral side of the orbital rim away from the globe, and cut the lateral canthal tendon inferiorly (Fig. 52.4).
   - At this point, again measure IOP, and if greater than 40 mmHg, proceed to the next step.
   - Direct the scissors along the lateral side of the orbital rim away from the globe, this time directed superiorly, and cut the superior crux of the lateral canthal tendon.

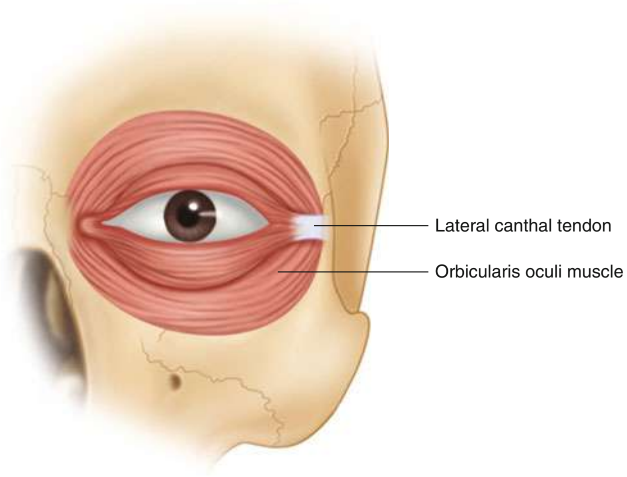
Lateral canthal tendon
Orbicularis oculi muscle

Fig. 52.2 Lateral canthal tendon

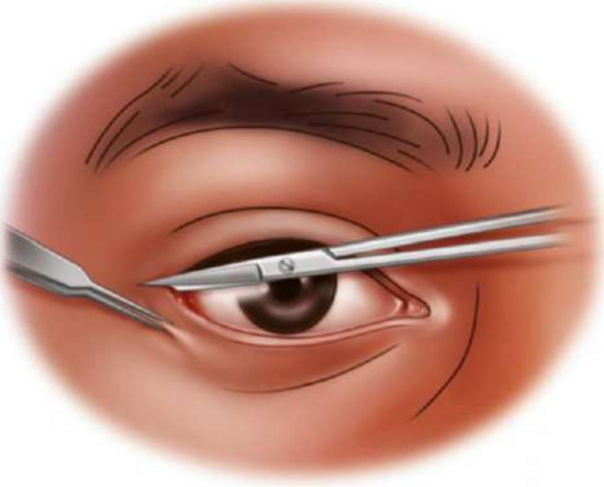

Fig. 52.4 Further cut to reduce intraocular pressure

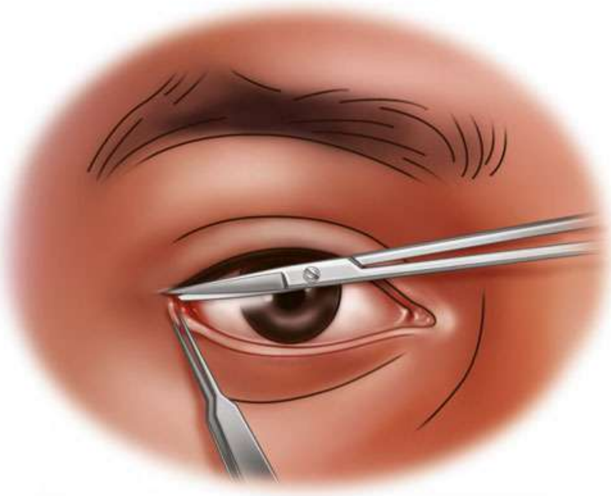

Fig. 52.3 Initial cut

52.5 Complications

- If one is making a superior incision, caution must be taken during this procedure to avoid injury to the lacrimal gland or artery.
- An additional complication of the procedure is ptosis due to iatrogenic injury to the levator aponeurosis, either partially or completely.
- The most obvious complication involves direct injury to the globe itself due to either operator error or poor patient control and immobility.
- Hemorrhage from inadequate hemostasis.
- Infection.

52.6 Pearls and Pitfalls

- Abrupt decrease in pain, decrease in IOP, and resolution of afferent pupillary defect will herald a successful procedure, assuming it was performed within an early enough time frame.
- Having an assistant present is beneficial.
  - In this case, the role of the assistant is to provide lateral retraction of the tissues to decrease likelihood of globe rupture.
- After cantholysis of the inferior component of the lateral canthal tendon, the lower lid will become lax as its attachment to the lateral wall is separated. This signifies a successful incision.
- Incisions made during this procedure generally heal without the need for suturing.
- Emergent ophthalmological consultation should ideally be sought before beginning this procedure.
- Emergent ophthalmological consultation is mandatory subsequent to the procedure.

Selected Reading

Goodall KL, Brahma A, Bates A, Leatherbarrow B. Lateral canthotomy and inferior cantholysis: an effective method of urgent orbital decompression for sight threatening acute retrobulbar haemorrhage. Injury. 1999;30:485–90.

McInnes G, Howes DW. Lateral canthotomy and cantholysis: a simple vision-saving procedure. CJEM. 2002;4:49–52.

Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: WB Saunders; 2009: Chap. 63.

Vassallo S, Hartstein M, Howard D, Stetz J. Traumatic retrobulbar hemorrhage: emergent decompression by lateral canthotomy and cantholysis. J Emerg Med. 2002;22:251–6.

Part VIII

Otorhinolaryngologic Procedures

Epistaxis Control

53

Benjamin M. Mahon and Bobby K. Desai

53.1 Etiology

- Anterior source: approximately 90 % of nosebleeds, usually Kiesselbach's plexus (Fig. 53.1)
  - Nose picking ("epistaxis digitorum")
  - Trauma
  - Infection
  - Nasal foreign body
  - Dry air
  - Atmospheric pressure alterations (e.g., increased altitude, lower arterial partial pressure of oxygen [PaO_2])
  - Allergies
  - Blood dyscrasias
  - Malignancy (e.g., leukemia, lymphoma)
- Posterior source: approximately 10 % of nosebleeds, usually sphenopalatine artery (Fig. 53.2)
  - Iatrogenic coagulopathy (e.g., warfarin, heparin, high-dose aspirin)
  - Blood dyscrasia
  - Liver failure
  - Renal failure
  - Malignancy
  - Older age

B.M. Mahon, MD
Ponciana Medical Center, Kissimmee, FL, USA

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_53

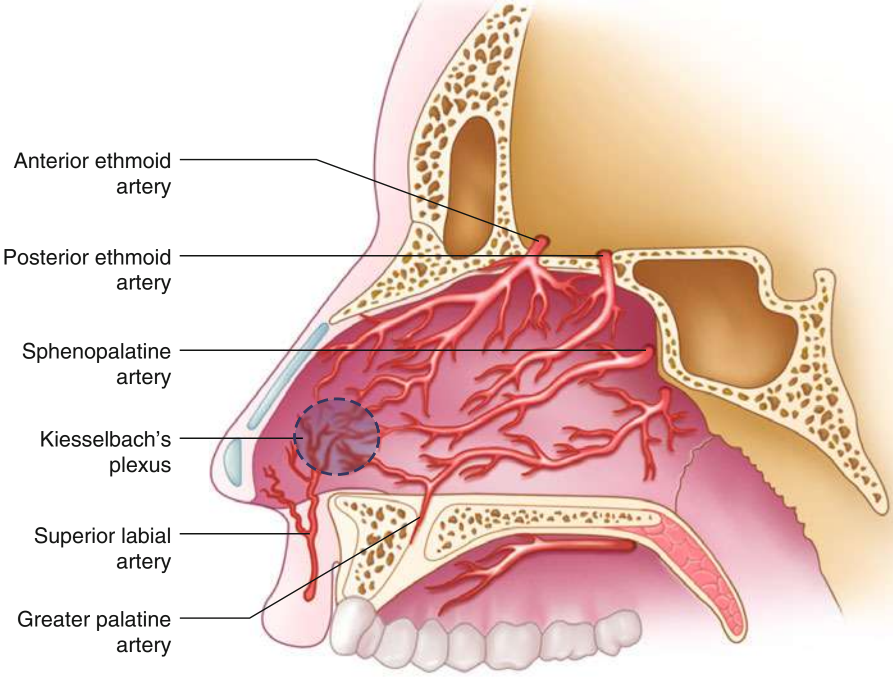
Anterior ethmoid artery
Posterior ethmoid artery
Sphenopalatine artery
Kiesselbach's plexus
Superior labial artery
Greater palatine artery

Fig. 53.1 Vascular supply to the nasal septum

Fig. 53.2 Vascular supply of the lateral wall of the nose

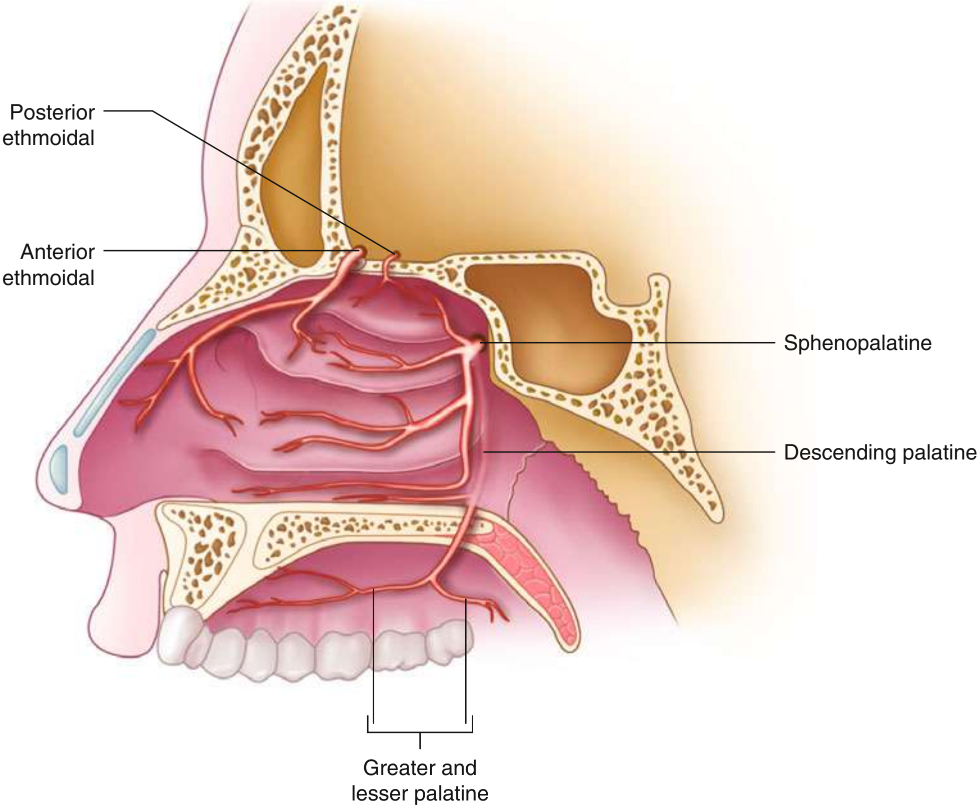
Posterior ethmoidal
Anterior ethmoidal
Sphenopalatine
Descending palatine
Greater and lesser palatine

53.2 Indications

- Acute/recurrent epistaxis

53.3 Contraindications

- Resolution of epistaxis episode before arrival
- Massive facial trauma

53.4 Materials and Medications

- Headlight with focused beam
- Gown
- Gloves
- Mask with visor
- Full-body drape or sterile towels
- Nasal packing material (any or all of the following, as needed [see later]): several cotton pledgets, 3–5 feet of ½-inch ribbon gauze (preferably enriched with petroleum jelly and bacitracin), anterior epistaxis balloon nasal pack
- Topical vasoconstrictor (e.g., 1 % pseudoephedrine or 1:1000 epinephrine)
- Topical anesthetic (e.g., 4 % lidocaine solution or 2 % tetracaine)
- Nasal speculum
- Bayonet forceps
- Anterior epistaxis balloon (e.g., Rhino Rocket)
- Any of the following commercial products, as needed (see later): Gelfoam, Surgicel, Crosseal, and FloSeal
- Foley catheter

53.5 Procedure: Anterior Epistaxis

- Initial preparation
  1. Assemble necessary equipment at the bedside.
  2. Have the patient seated upright, with head and neck in the sniffing position.
  3. Universal precautions are mandatory: gown, glove, and mask.
  4. Drape the patient with towels or a large body drape.
  5. If initial evaluation using the headlamp and nasal speculum with gentle spreading in a vertical fashion or a simple visual examination reveals the source (anterior vs posterior), proceed with the appropriate management pathway as determined later.
     – If clot obstructs visualization, it is recommended for the patient to gently blow her or his nose once to remove any easily friable clots, and then proceed with immediate inspection. If anterior nose bleeds, go to the next step.
- Management of anterior epistaxis:
  1. Initial step in the management of anterior nosebleeds involves pinching the patient's nostrils firmly for at least 10–15 min.
     – This can be done by the patient.
     – Simple pressure is very often enough to provide appropriate hemostasis in anterior nosebleeds. If it fails, proceed to the next step.
  2. A topical vasoconstrictor of choice (e.g., cocaine, oxymetazoline, 1 % pseudoephedrine) can be applied to the septum and lateral walls of the nose either topically with cotton-tipped applicators or by using soaked pledgets.
  3. Continue to apply firm pressure to the nares for 10–15 min, and recheck the nose for evidence of bleeding.
  4. If the bleeding site can be easily visualized, chemical cautery may be attempted with silver nitrate (Fig. 53.3).
     – This can be accomplished with the help of a nasal speculum and a headlight.
     – It is advised to cauterize both the site itself and 0.5 cm around the site, by holding the silver nitrate stick to the site for at least 20 s until the bleeding stops.
     – If bleeding then stops, cover the site with an absorbable gelatin foam or an oxidized cellulose such as Gelfoam or Surgicel or simply apply topical antibiotic to the site.
  5. If there is continued bleeding, the clinician may apply either two elongated cotton pledgets or a commercially available substitute (e.g., a Rhino Rocket or similar device [Fig. 53.4]) presoaked in a few milliliters or a 1:1 mixture of a topical vasoconstrictor and an anesthetic. Combinations can include the previously cited two agents, plus 2 % tetracaine or 4 % lidocaine.
  6. Universal instructions for the insertion of a commercially available anterior nasal pack include (Fig. 53.5):
     - Soak the fabric via the manufacturer's recommended solution if required.
     - Insert antibiotic ointment in the nares to facilitate insertion.
     - Insert the entire length of the balloon along the inferior surface of the nasal cavity.
     - The fluid within the nasal cavity should cause it to expand spontaneously while in the nares, after which the remaining exposed string can be taped to the cheek.
     - For those devices that require air instillation, read the manufacturer's instructions on how much air to instill within the device.

7. Another option is to provide an anterior nasal packing using a few feet of ½-inch ribbon gauze (Fig. 53.6).
8. It is preferable if this gauze is impregnated with petroleum jelly and is enriched with bacitracin ointment.
   - If it is not, bacitracin can be applied to the strip before its insertion into the nasal cavity.
9. To facilitate the insertion, bayonet forceps can be used to gently lay each successive strip on top of the one before it.
10. Begin inferiorly, laying each strip on top of the other, about three or four layers at a time, with the end points protruding from the opening of the nasal cavity.
11. Pack the strands firmly.
12. It may take several feet to pack the entire nasal cavity.
13. This may be both uncomfortable for the patient and time consuming for the provider, so ensure adequate time and copious topical anesthetic.
14. Parenteral analgesic or anxiolytics may be required for any of these techniques.
15. If none of these resources is available, another technique requires the use of two cotton pledgets gently inserted into the nostrils after vasoconstrictor/anesthetic topical solution has been sprayed into the nose either manually or with the addition of bayonet forceps, where they will remain for 20 min, supplemented by firm pressure to the nares as described previously.
    - If the bleeding has then stopped, they can be removed at this time. If it does not, it is advised to repeat the previous step a second time, this time with a longer period of pressure, before proceeding to the next step.
16. If all of these techniques fail to control bleeding, the clinician may apply a sealant spray or foam enriched with thrombin to enhance clotting.
    - Commercially available options include Crosseal or FloSeal.
17. Finally, if none of these techniques manage to halt what one is sure is only an anterior, albeit persistent, nosebleed, consult the otolaryngologist for further guidance.
    - Another consideration at this point may be that the uncontrollable anterior nosebleed is really a posterior nosebleed from an undetermined location.
    - It may be prudent to proceed to the algorithm for treatment of posterior epistaxis, even if a posterior bleed is not specifically confirmed.

53.5.1 Complications

- Using an inadequate amount of packing when inserting the ribbon gauze, the entire product may serve as a plug with potential aspiration risk, rather than as a tool for hemostasis.
- Otitis, sinusitis, and toxic shock syndrome.

53.5.2 Pearls and Pitfalls

- Pearls
  - Be sure to perform a reevaluation of both the posterior oropharynx and the posterior nasopharynx after insertion of any of the previously discussed devices, and inspect for continued bleeding/oozing to ensure hemostasis has been accomplished.
  - For those patients with nasal packs in place, consider the use of oral antibiotics to avoid subsequent sinusitis, otitis, or toxic shock syndrome.
    - Appropriate antibiotic choices include cephalexin, clindamycin, or amoxicillin–clavulanic acid.
    - The use of antibiotics for short-term packing has not been proven.
  - Consider admission for elderly patients or for those patients with potential airway complications such as those patients with chronic obstructive pulmonary disease (COPD) and for those at risk for aspiration.
  - Have the patient follow up with otolaryngology within 2–3 days for reevaluation and packing removal.
  - Provide adequate discharge instructions warning the patient against nose blowing, sneezing with the mouth closed, or any movements or actions that cause a Valsalva maneuver.
  - The patient should try to keep his or her head elevated for the next 24–48 h even while sleeping to prevent potential aspiration.
- Pitfalls
  - Discharging a patient without an adequate period of observation.
  - Failing to ensure appropriate follow-up.
  - Providing inadequate discharge instructions.
  - Mistaking an anterior nosebleed for a posterior nosebleed.
  - Leaving silver nitrite on exposed nasal mucosa can lead to iatrogenic nasal septal perforation.
  - Not informing the patient that minimal oozing of blood from the nose can be expected.

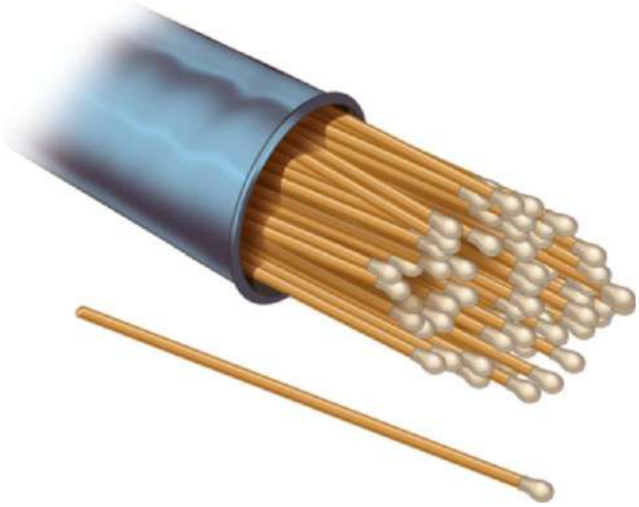

Fig. 53.3 Example of silver nitrate stick and packaging

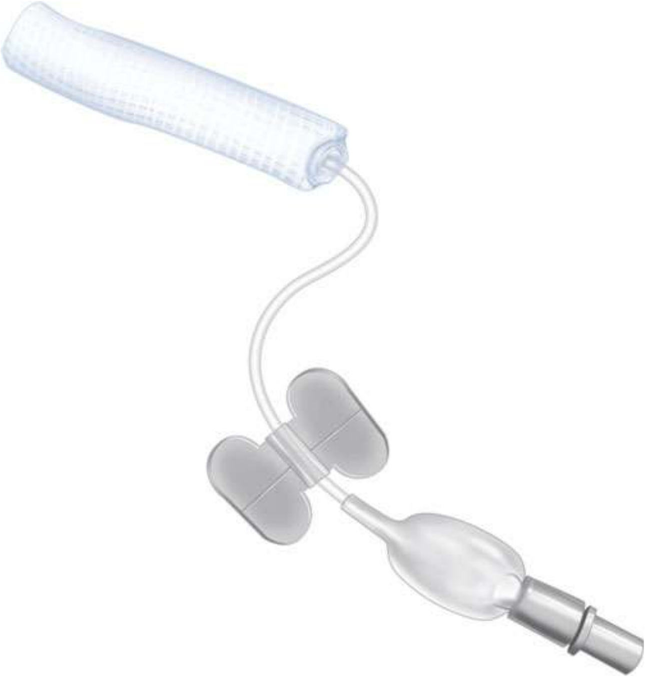

Fig. 53.4 Example of commercial nasal pack

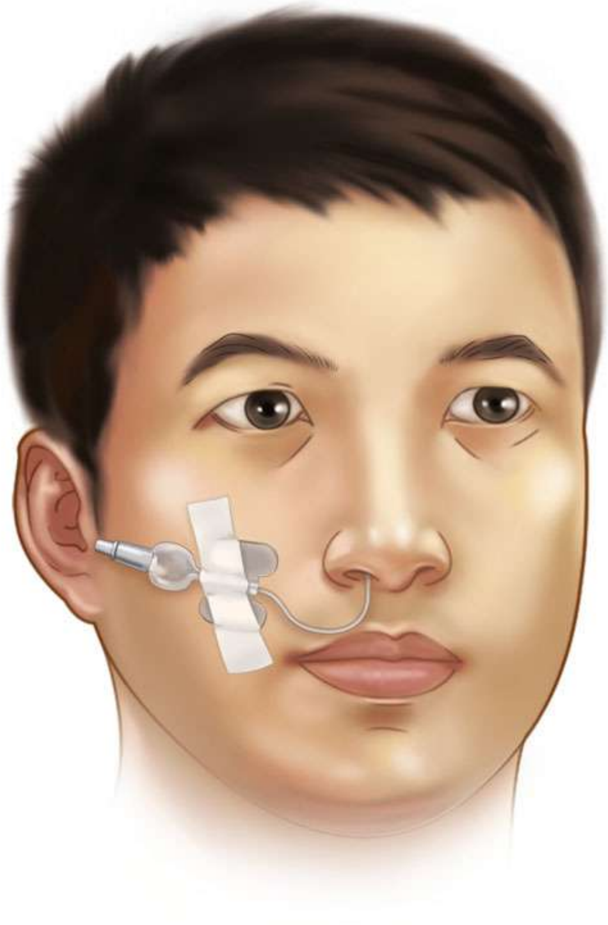

Fig. 53.5 Proper taping of a commercial anterior nasal pack

Fig. 53.6 Gauze method

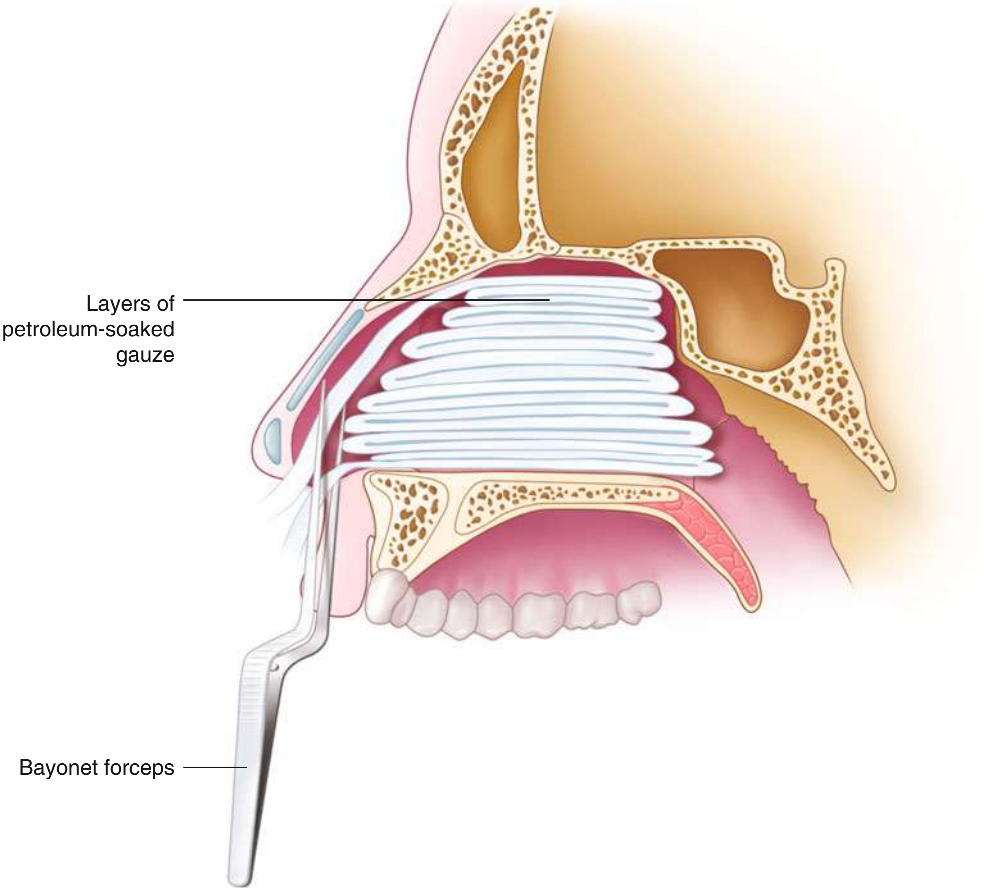
Layers of petroleum-soaked gauze
Bayonet forceps

53.6 Procedure: Posterior Epistaxis

- Preliminary Steps
  1. Owing to the possibility of extreme exsanguination from a posterior source of epistaxis or airway obstruction from a large clot dislodgment or large volume of blood aspiration, the stability of the patient's airway, breathing, and circulation must be ensured first.
     - Appropriate interventions entirely depend on the presentation of the patient and therefore are not described in this chapter.
     - Patients with posterior epistaxis can lose a large volume of blood quickly, so this needs to be an initial consideration.
  2. Consider establishing intravenous access and placing the patient on a cardiac monitor.
  3. Because patients with posterior epistaxis tend to be either older, on various blood thinners, or to potentially have some type of coagulopathy, consider obtaining a complete blood count, prothrombin time, and partial thromboplastin time as part of the preliminary workup.
  4. Consider a blood type and screen.
- Management of posterior epistaxis:
  1. Inspect the nares as described in the previous section on anterior nosebleeds.
  2. Several of the techniques discussed may be attempted but may not be successful.
  3. Assemble necessary equipment at the bedside.
  4. Have the patient seated upright, with the head and neck in the sniffing position.
  5. Universal precautions including a gown, glove, and mask should be employed.
  6. Drape the patient with towels or a large body drape.
  7. Apply a combination of a topical vasoconstrictor and anesthetic to the nose.
     - An appropriate choice (e.g., 1 % pseudoephedrine or 1:1000 epinephrine in a 1:1 ratio with 4 % lidocaine or 2 % tetracaine)
     - Be aware, however, that there are situations in which the bleeding may be too brisk, and this step may need to be avoided owing to time constraints and the volume of bleeding. So it is not unheard of to be required to intervene in posterior epistaxis without this initial step.
  8. Otolaryngology should be consulted as early as possible in the workup of this condition, both for admission and in case the following techniques fail.
  9. An advised initial step for the treatment of posterior epistaxis is to proceed directly to insertion of a unilateral or, preferably, bilateral posterior nasal pack or balloon.
     - An example of these is the elongated version of the Rhino Rocket as described previously.
     - Insertion and securing of the product proceed essentially the same way, except that it is inserted deeper into the posterior oropharynx.
  10. Consider the use of a dedicated posterior balloon, such as a Nasostat or Epistat (Fig. 53.7).
      - These devices are double-balloon systems that are lubricated preferably with bacitracin (sterile jelly is also appropriate) and inserted into the posterior nasopharynx.
      - The posterior balloon is then inflated with the product-specific quantity of air.
      - The product is then withdrawn slightly to ensure that the posterior balloon is situated firmly in its desired location of the posterior nasal cavity.
      - The anterior balloon is then inflated with the device-specific quantity of normal saline, and the device is secured.
- If the previously discussed resources are not available, a clinician can insert an adult-caliber Foley catheter into the posterior nasopharynx, with inflation of the balloon, and gentle traction anteriorly to ensure the bulb is situated similarly to the Nasostat or Epistat, firmly in the posterior nasal cavity to tamponade off the bleeding (Fig. 53.8).
- If all of these techniques still fail to control bleeding, an otolaryngologist will likely need to intervene, performing techniques outside the scope of practice of the emergency physician, including electrocautery, submucosal lidocaine/epinephrine injections, or other surgical interventions.
- The final step in the management of this patient includes admission to a monitored bed.

53.6.1 Pearls and Pitfalls

- Pearls
  - Consult otolaryngology early in the workup as soon as a posterior bleed is identified.
  - Patients receiving packing will still need to be placed on antibiotics, such as a first-generation cephalosporin, clindamycin, or amoxicillin–clavulanic acid.
  - Perform a good physical examination concomitantly with the management of the bleed specifically geared toward assessing the patient's volume status (orthostatics, capillary refill, heart rate), because the patient may have significant volume loss depending on the duration and quantity of the bleed.
  - Take a good history including prior bleed, prior admission requirements owing to epistaxis, any anticoagulant use, blood dyscrasias, and so on.

- Pitfalls
  - Misidentifying a posterior bleed as an anterior bleed. This can be catastrophic if the patient is discharged. Reference the “Etiology” section earlier.
  - Discharging the patient.
  - Consulting otolaryngology too late.
  - Underestimating the volume of blood loss. It may be helpful to think of these patients as trauma patients in their initial management and to proceed down the standard primary survey/intervention/secondary survey/intervention approach, ensuring that the volume needs are identified early.
  - Infection.
  - Pressure necrosis of the septum.
  - Hypoxia.
  - Aspiration.
  - Arrhythmias.
  - Dysphagia.
  - Dislodgment of the pack.

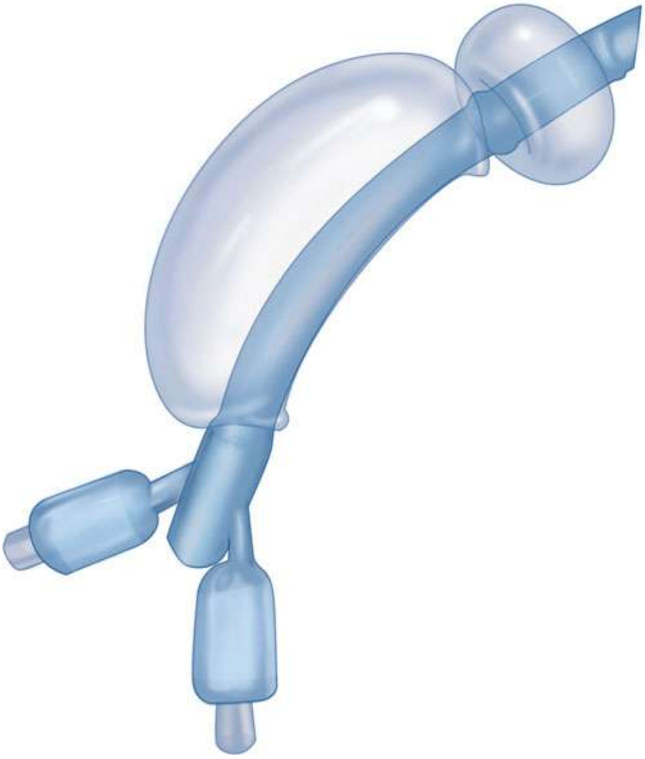

Fig. 53.7 Example of posterior balloon

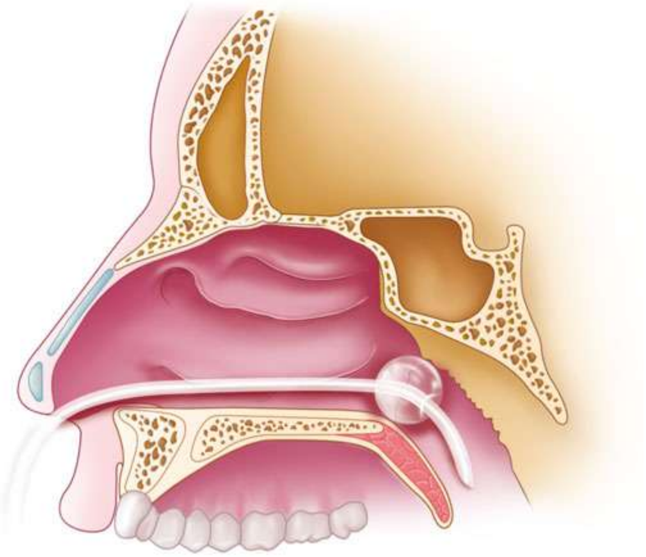

Fig. 53.8 Foley catheter in the posterior nasopharynx

Selected Reading

Buttaravoli P. Minor emergencies, splinters to fractures. 2nd ed. Philadelphia: Mosby; 2007.

Kucik CJ, Clenney T. Management of epistaxis. Am Fam Physician. 2005;71(2):305–11.

Marx JA, Hockberger RS, Walls RM, editors. Rosen's emergency medicine: concepts and clinical practice. 6th ed. Philadelphia: Mosby; 2006.

Pope LE, Hobbs CG. Epistaxis: an update on current management. Postgrad Med J. 2005;81(955):309–14.

Singer AJ, Blanda M, Cronin K, et al. Comparison of nasal tampons for the treatment of epistaxis in the emergency department: a randomized controlled trial. Ann Emerg Med. 2005;45:134–9.

Summers SM, Bey T. Chapter 239. Epistaxis, Nasal Fractures, and Rhinosinusitis. In: Tintinalli's Emergency Medicine, 7th ed. New York, NY: McGraw Hill; 2011.

Treatment of Septal Hematoma

54

Bobby K. Desai

54.1 Indications

- Septal hematomas occur after force applied to the nasal cartilage results in leakage of blood from the perichondrium (Figs. 54.1 and 54.2). They may be unilateral or bilateral.
- Untreated, this hematoma can expand and mechanically obstruct the blood vessels that supply the nasal cartilage.

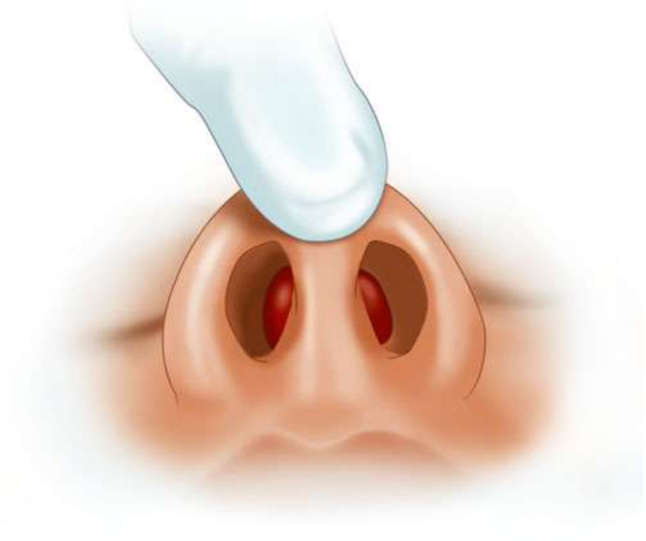

Fig. 54.1 Example of a septal hematoma

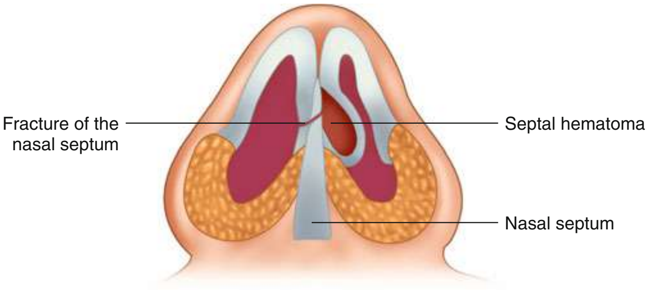

Fig. 54.2 Diagram of a septal hematoma

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_54

54.2 Contraindications

- In the setting of massive facial trauma, airway protection and consultation with otolaryngology or oral surgery may be required.

54.3 Materials and Medications

- 5-cc syringe
- 20-gauge needle (may use larger-gauge needle, but smaller ones may not allow for adequate suctioning).
- Topical antiseptic
- Sterile drapes, gown, eye protection, and gloves
- Light source—preferably a head lamp
- If using procedural sedation, specific medications of choice are needed as well as monitoring and resuscitation equipment
- Nasal speculum
- 4 % liquid cocaine or 4 % liquid lidocaine for anesthesia
- Topical vasoconstrictor of choice (oxymetazoline or phenylephrine)
- Scalpel (#15 blade)
- Suction setup
- Saline for irrigation
- Sterile gauze
- Iodoform packing or sterile rubber band
- Anterior nasal pack (Rhino Rocket)
- Intravenous antibiotics to cover *Staphylococcus aureus*, group A β-hemolytic streptococci, *Streptococcus pneumonia,* and *Haemophilus influenzae*

54.4 Procedure

- Preprocedure
  - Explain the procedure to the patient.
  - Informed consent may be required.
- Procedure—unilateral (Fig. 54.3):

1. Place the patient in a seated position if the procedure is to be done without procedural sedation.
    - If performing the procedure under procedural sedation, the patient may be placed in the supine position with appropriate monitoring and resuscitation equipment on standby.
2. Provide local anesthesia by applying the topical anesthetic of choice to cotton pledgets, wringing out the excess before insertion.
    - Remove after 10 min.
    - If using lidocaine, a topical vasoconstrictor may be required that may be mixed with the lidocaine or applied directly on the nasal mucosa before the procedure.
3. Using a nasal speculum, open the affected nares vertically as wide as possible. Horizontal use may obscure the hematoma.
4. Using a #15 blade, make a vertical incision at the hematoma extending the incision posteriorly at its base.
    - Take care not to make too deep an incision because septal perforation may occur.
5. Suction and gauze should be at the ready to remove as much blood as possible.
6. Pack the incision site with iodoform gauze or a sterile rubber band in order to keep the incision open to prevent accumulation of blood.
7. This packing can be held in place with an anterior nasal pack.
    - The pack should be placed in both nares to prevent deformity of the nasal septum.
8. Consider administration of a dose of intravenous antibiotics.
9. Send the patient home with a prescription for oral antibiotics.
10. Arrange follow-up for 24 h.

- Procedure—bilateral (Fig. 54.4):

1. If bilateral hematomas are present, otolaryngology consultation should be sought.
2. One method involves the incision of one side only through the nasal septum so that both may drain through one side.
    - Care must be taken NOT to penetrate the overlying mucosa of the other side.
3. The second option is to repeat the unilateral drainage procedure on the contralateral side.
4. Care must be taken to not have the incision exactly in the same location as the contralateral side owing to the risk of permanent septal perforation.

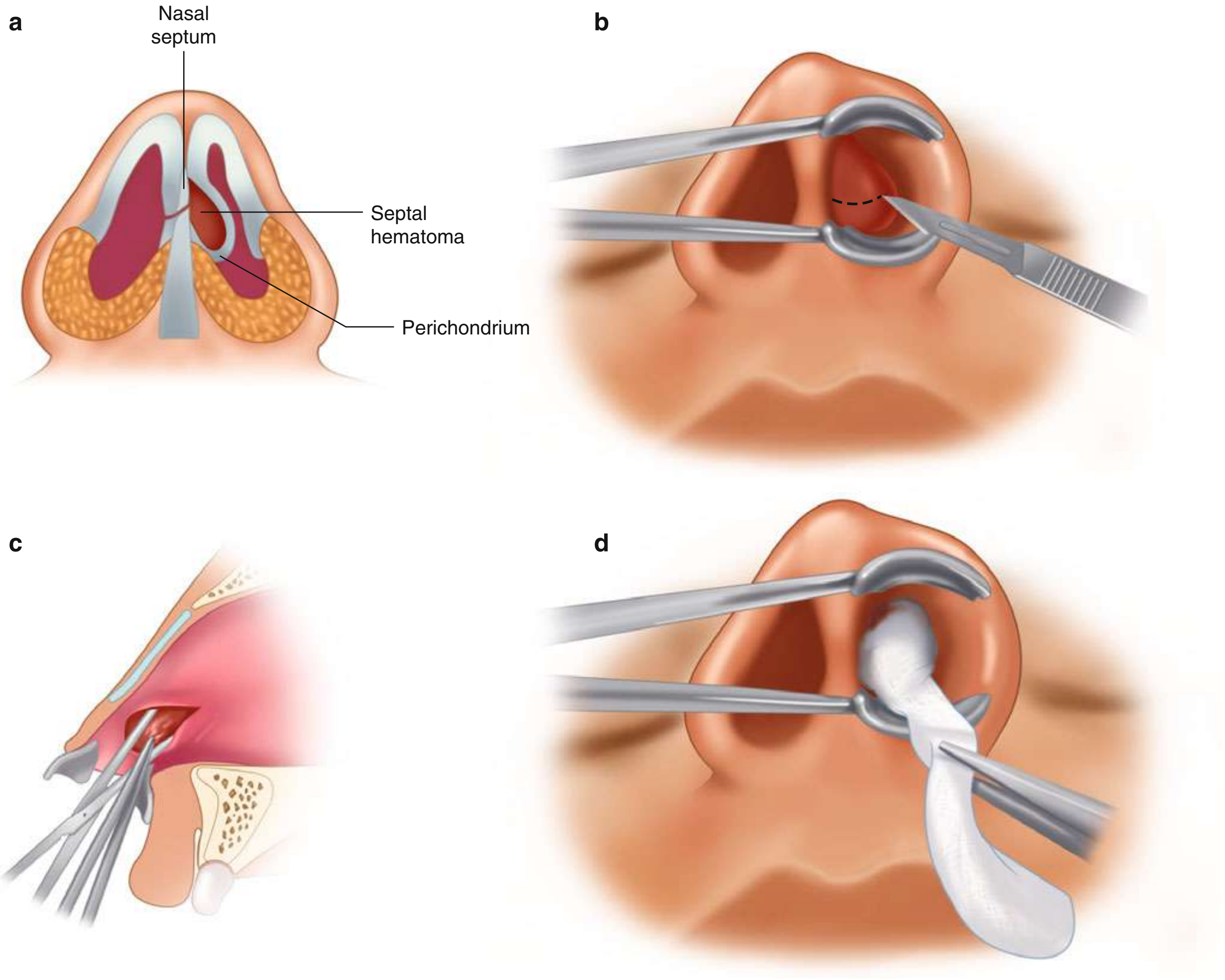
a
Nasal septum
Septal hematoma
Perichondrium
b
c
d

Fig. 54.3 (**a**–**d**) Technique for drainage of a unilateral septal hematoma

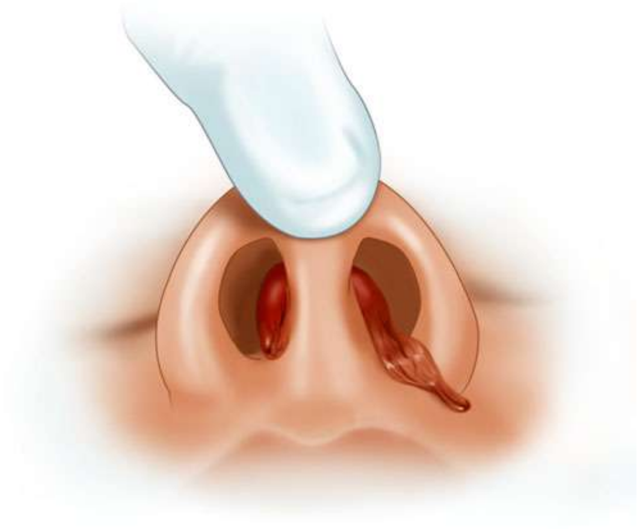

Fig. 54.4 Technique for drainage of a bilateral septal hematoma

54.5 Pitfalls and Complications

- Bleeding
- Infection: toxic shock syndrome
- Deformity of the nose
- Permanent septal perforation
- Inadequate drainage
- Reaccumulation of blood
- Development of septal abscess

Selected Reading

Ginsburg CG. Consultation with the specialist: nasal septal hematoma. Pediatr Rev. 1998;19:142–3.

Savage RR, Valvich C. Hematoma of the nasal septum. Pediatr Rev. 2006;27:478–9.

Nasal Foreign Body Removal

55

Bobby K. Desai

55.1 Indications

- Visualized foreign body (FB) on inspection
- Continued unilateral nasal discharge of unknown etiology
- Recurrent epistaxis

55.2 Contraindications: Need Urgent Referral

- Those cases in which the emergency physician is not confident of success.
- Combative patients who are not candidates for conscious sedation.
- Embedded FBs.
- Penetrating injuries with an FB.
- Button batteries that cannot be removed.
- Patients with bleeding diathesis.
- Respiratory distress.
- FBs superior and medial to the middle turbinate owing to risk of puncture to the cribriform plate.
- Chronic FBs may be difficult to visualize and these should be referred.

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_55

55.3 Materials and Medications

- 1 % lidocaine without epinephrine (solution) for anesthesia
- 2 % lidocaine jelly
- 0.5 % phenylephrine to vasoconstrict vessels
- Nasal speculum (Fig. 55.1)
- Headlamp or other direct lighting instrument
- Ambu bag (Fig. 55.2)
- Alligator forceps (Fig. 55.3)
- Curved hook (Fig. 55.4)
- Foley catheter
- Suction: schuknecht catheter (Fig. 55.5)
- Eye and biohazard protection for the clinician

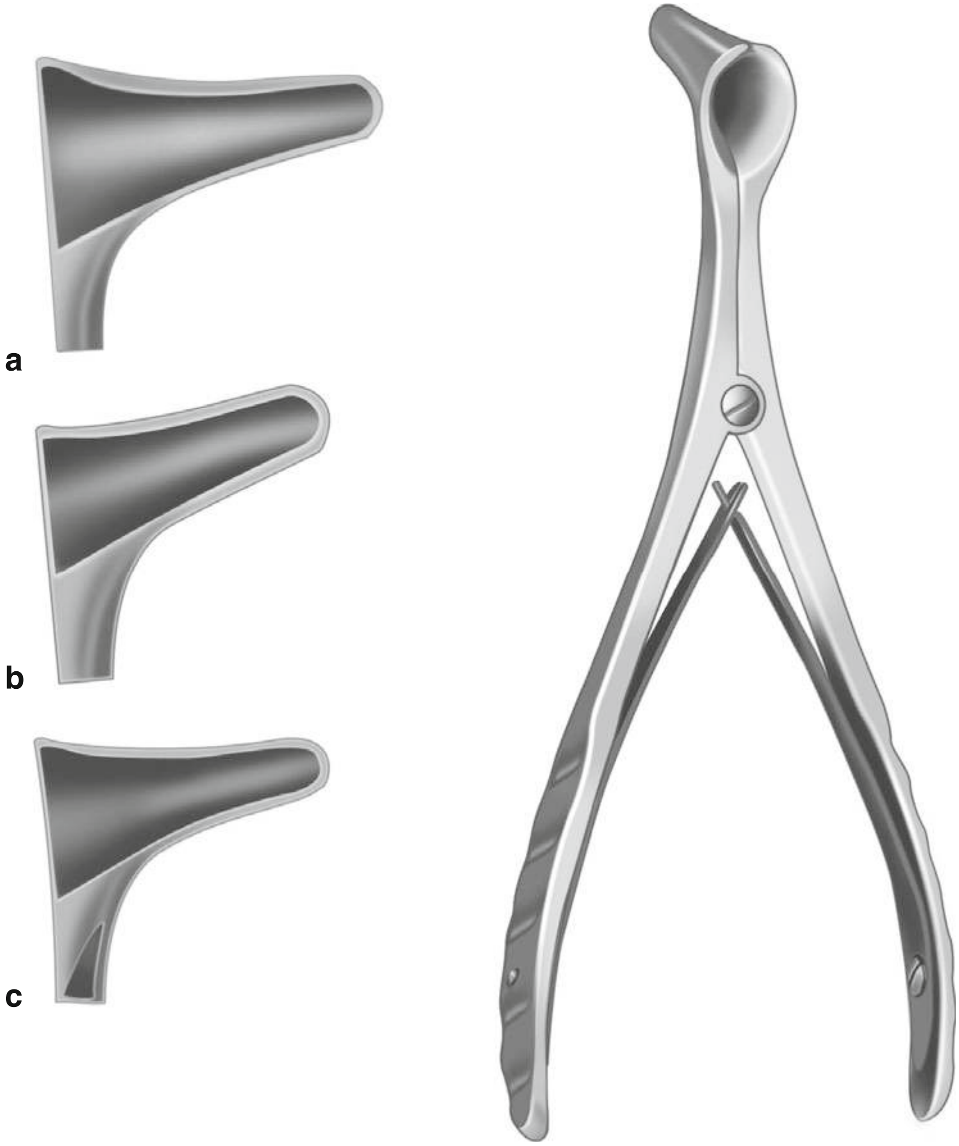

Fig. 55.1 (**a**–**c**) Example of nasal speculum

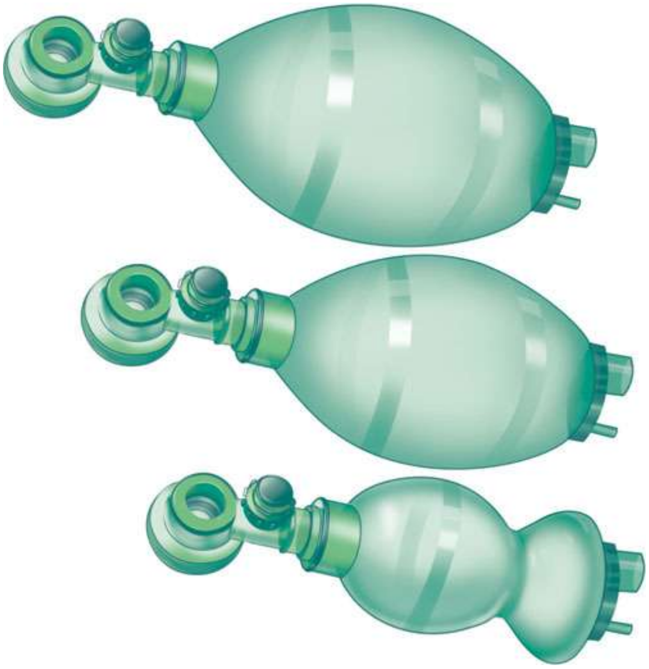

Fig. 55.2 Example of Ambu bags

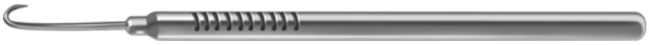

Fig. 55.4 Example of a curved hook

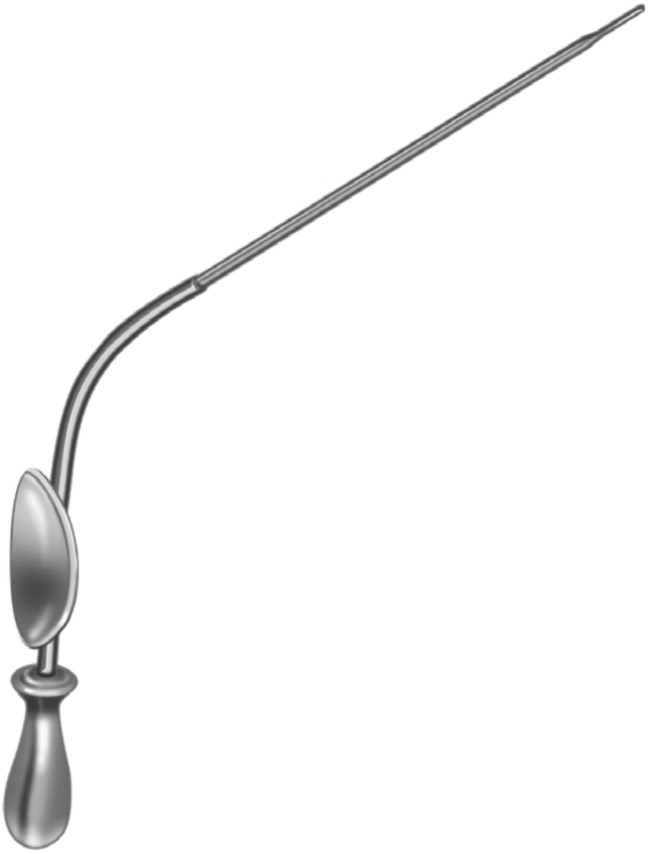

Fig. 55.5 Example of a schuknecht suction tube

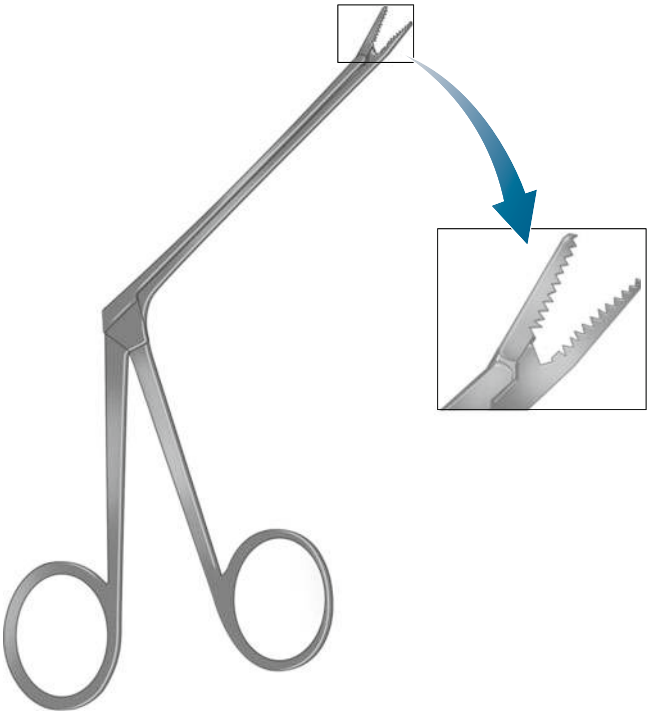

Fig. 55.3 Example of alligator forceps

55.4 Procedures (Based on the Method Chosen)

55.4.1 Alligator Forceps

- Perhaps the easiest method to use for easily visualized FBs.
- Care must be taken for those FBs that are easily broken apart (e.g., peas) to ensure that all of the matter has been removed.
- Using a nasal speculum, spread the nares as much as possible to maximize visualization.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Attempt to grasp the FB with the forceps.
- After removal, ensure the complete removal of the object and ensure that no other FBs are present.

55.4.1.1 Pitfalls and Complications

- Pushing the object deeper into the nares.
- Inability to totally remove the FB.
- Attempted removal of an embedded FB may cause significant bleeding.

55.4.2 Curved Hook

- Used for nongraspable objects especially in the anterior nares.
- Using a nasal speculum, spread the nares as much as possible to maximize visualization.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pass the hook behind the object, or if the object has holes, the hook may be placed inside the hole to facilitate removal.
- Slowly withdraw the hook, thereby removing the FB.
- After removal, ensure the complete removal of the object and ensure that no other FBs are present.

55.4.2.1 Pitfalls and Complications

- This technique cannot be used if the object cannot be directly visualized because significant trauma may arise owing to the misplacement of the hook.
- Attempted removal of an embedded FB may cause significant bleeding.

55.4.3 Foley Catheter Removal

- Use a 5 or 6 French catheter.
- Anesthetize the nares with lidocaine with or without epinephrine and vasoconstrict the nasal vessels with phenylephrine.
- Ensure that the Foley balloon has no leaks.
- Place the patient in a supine position.
  - Consider procedural sedation for children and/or a papoose board.
- Lubricate the Foley tip and balloon with lidocaine jelly.
- Advance the tip past the object.
- Blow up the balloon with 2–3 mL of air and gently withdraw the catheter.
  - The amount of air may need to be adjusted depending on the size of the object to be withdrawn.
  - Pulling too quickly may cause a soft FB to break apart.

55.4.3.1 Pitfalls and Complications

- This technique may be used if the FB is not visible.
- In order for this technique to work, the catheter must be able to slide posterior to the object in order for the balloon to pull out the object.

55.4.4 Suction

- A small metal suction catheter—the schuknecht catheter—may be used.
- The suction is placed against the FB and slowly withdrawn.
- This technique works best for round smooth objects visualized in the anterior nares.

55.4.4.1 Pitfalls and Complications

- May not work for posteriorly displaced FBs.
- May not work for objects too tightly entrenched in the nares.

55.4.5 Nasal Positive Pressure

- This technique works best for round FBs occluding the nares.
- Place the patient in a supine position.
  - Consider a papoose board to secure the patient.
- Apply pressure to the contralateral nares to occlude it.
- Connect an Ambu bag to high-flow oxygen.
  - Use a facemask that covers only the mouth.
- Place the mask on the patient's mouth and allow it to expand by covering the thumbhole.
- If this pressure is insufficient to expel the object, the Ambu bag may be compressed to add pressure to the attempt.
- This may allow the object to be seen and grasped in the anterior nares.

55.4.5.1 Pitfalls and Complications

- Too much pressure could theoretically rupture the tympanic membranes.

Selected Reading

Backlin SA. Positive-pressure technique for nasal foreign body removal in children. Ann Emerg Med. 1995;25:554–5.

Chan TC, Ufberg J, Harrigan RA, et al. Nasal foreign body removal. J Emerg Med. 2004;26:441–5.

Kadish H. Ear and nose foreign bodies: it is all about the tools. Clin Pediatr. 2005;44:665–70.

Cerumen Removal

56

Bobby K. Desai

56.1 Indications

- Hearing loss
- Profound dizziness
- Pain

56.2 Contraindications

- Uncooperative patient
- Potential for foreign bodies
- Prior surgery to ear or mastoid
- History of middle or outer ear disease

56.3 Materials and Medications

- Ear speculum (Fig. 56.1).
- Eye and biohazard protection for the clinician
- Headlamp or other direct lighting instrument
- Cerumen-softening solutions
  - Carbamide peroxide
  - Mineral oil or other oil-based agents including almond or olive oil
  - Liquid docusate sodium
  - Acetic acid
  - Sodium bicarbonate
- Alligator forceps (Fig. 56.2)
- Curved hook or cerumen spoon (Figs. 56.3 and 56.4)
- Suction—Schuknecht catheter
- Irrigation setup
  - Ear syringe (Figs. 56.5 and 56.6)
- – 20- to 60-mL syringe with an 18-gauge angiocath attached
  - Commercial ear irrigation setup
  - Water (tap water may be used)
  - Basin or other collection device
  - Towels

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_56

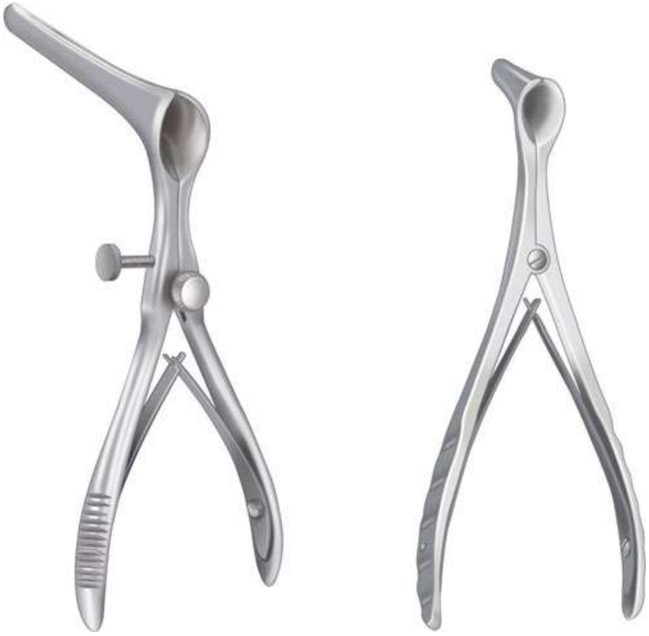

Fig. 56.1 Example of an ear speculum

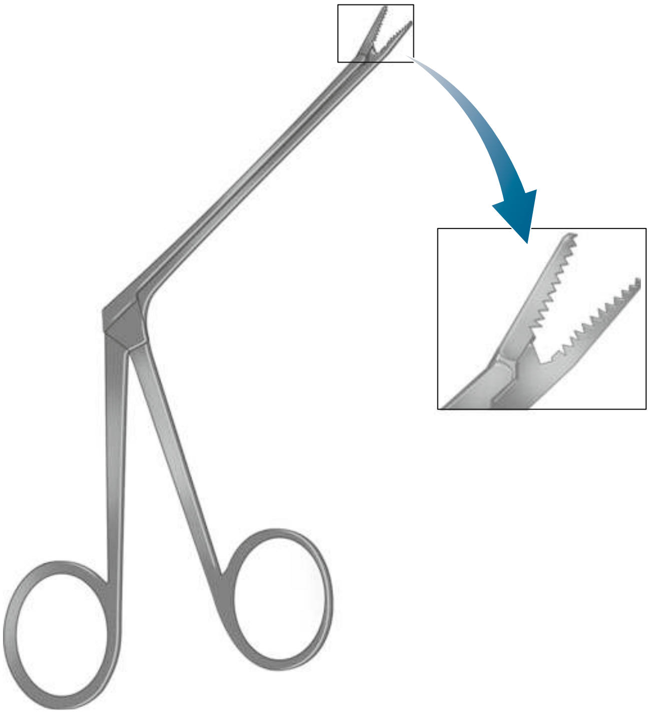

Fig. 56.2 Example of alligator forceps

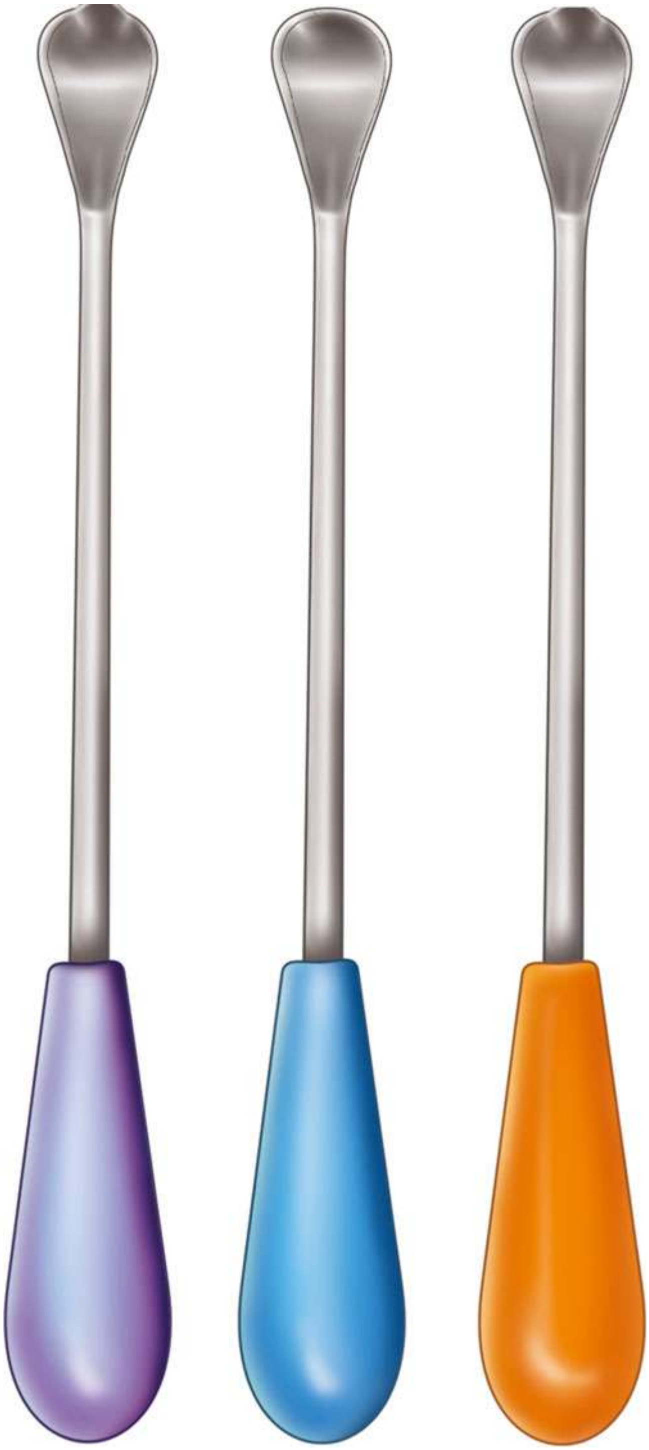

Fig. 56.3 Examples of cerumen spoons

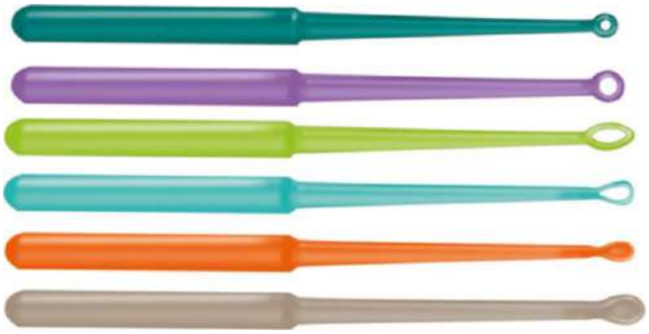

Fig. 56.4 Examples of ear curettes

Fig. 56.5 Example of a typical bulb syringe used in irrigation

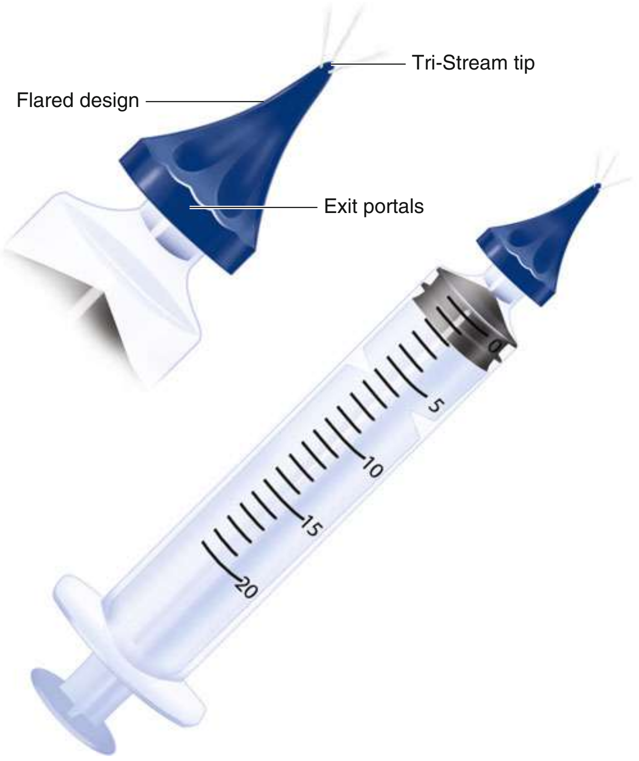
Tri-Stream tip
Flared design
Exit portals
0
5
10
15
20

Fig. 56.6 Example of a specialized ear syringe

56.4 Procedure

1. Explain the procedure to the patient to facilitate cooperation.
    - Some discomfort may be involved, including a sensation of pressure and the cold, wet feeling from the water used.
    - Nausea, vomiting, or a sensation of vertigo may also be experienced.
        - Consider pretreatment of vertigo with agent of choice.
2. The patient may be seated or supine with the head slightly turned to the affected side to facilitate the collection of fluid.
3. Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
4. Pull the ear up and out to straighten the ear canal.
5. Before irrigation, if the cerumen is hard, alligator forceps or a cerumen spoon may be used under direct visualization to remove as much cerumen as possible.
6. Use a cerumen-softening agent before irrigation by instilling and leaving in for 30 min to maximize softening.
7. Place a collection device next to the patient's ear, or alternatively, have the patient hold a basin near the ear to collect any liquid.
8. Attach an angiocath to a 20- or 60-mL syringe, place the tip of the catheter within the ear canal, and aim toward the tympanic membrane.
9. Slowly inject the solution into the ear canal.
    - Care must be taken to limit the force used owing to the potential for tympanic membrane rupture.
    - The fluid can be injected in short bursts or all at once.
10. Direct the stream as much as possible toward the rear of the cerumen to be removed.
11. Continually reassess the ear canal to ensure that all cerumen has been removed.
    - If one large piece remains, use a cerumen hook to facilitate removal.
12. If the patient complains of severe sharp pain, stop the procedure and assess the tympanic membrane for rupture.
    - Consider referral or otolaryngological consultation at this point.
13. After successful removal, dry the ear canal with cotton-tipped applicators to prevent otitis externa.

56.5 Pitfalls and Complications

- Owing to instillation of room temperature water, a caloric response may occur including vertigo, nausea, and vomiting.
- Rupture of the tympanic membrane.
- Unsuccessful irrigation secondary to inadequate time given to softening agents before attempting removal.
- Bleeding of the external auditory canal.
- Otitis externa.
- Hearing loss.

Selected Reading

Blake P, Matthews R, Hornibrook J. When not to syringe an ear. N Z Med J. 1998;111:422.

McCarter D, Courtney AU, Pollart SM. Cerumen impaction. Am Fam Physician. 2007;75:1523–8.

Roland PS, Smith TL, Schwartz SR, et al. Clinical practice guideline: cerumen impaction. Otolaryngol Head Neck Surg. 2008;139:S1–21.

Ear Foreign Body Removal

57

Bobby K. Desai

57.1 Indications

- Visualized foreign body (FB) on inspection.
- Continued unilateral otorrhea of unknown etiology mandates a search.
- Recurrent bleeding.
- History of FB placement by patient.

57.2 Contraindications: Need Urgent Referral

- Those cases in which the emergency physician is not confident of success.
- Combative patients are not candidates for conscious sedation.
- Embedded FBs.
- Penetrating injuries with an FB.
- FBs close to the tympanic membrane.
- Consider referral of spherical FBs because these may be difficult to grasp.
- Chronic FBs may be difficult to visualize; these should be referred.
- Consider referral for those patients who have attempts made at other institutions.

57.3 Materials and Medications

- Ear speculum (Fig. 56.1)
- Eye and biohazard protection for the clinician
- Headlamp or other direct lighting instrument
- Alligator forceps (Fig. 56.2)
- Curved hook (Fig. 57.1) or cerumen spoon (Fig. 56.3)
- Suction—Schuknecht catheter (Fig. 57.2)
- Irrigation setup
  - Ear syringe (Fig. 57.3)
  - 20- to 60-mL syringe with an 18-gauge angiocath attached
  - Water (tap water may be used)
  - Basin or other collection device
  - Towels
- Cyanoacrylate (Fig. 57.4)

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_57

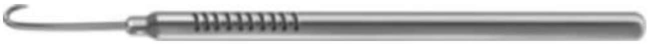

Fig. 57.1 Example of a curved hook

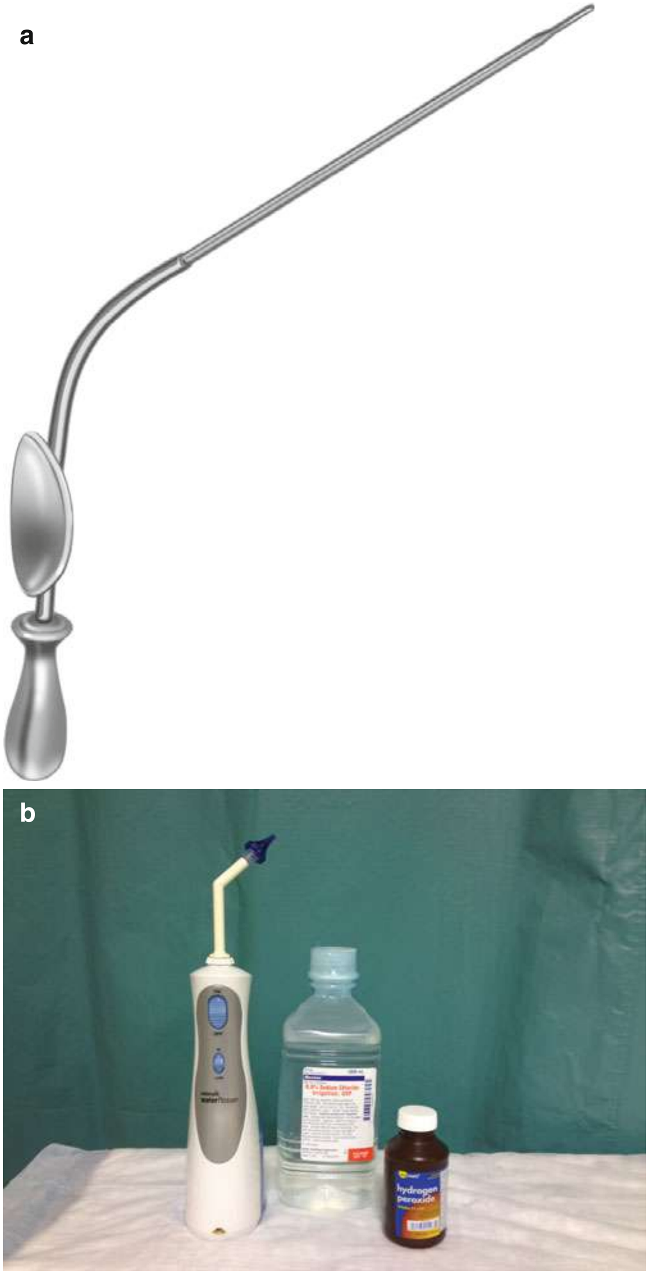
a
b
hydrogen peroxide

Fig. 57.2 (**a**) Example of a Schuknecht suction tube, (**b**) example of ear irrigation setup using a Waterpik system

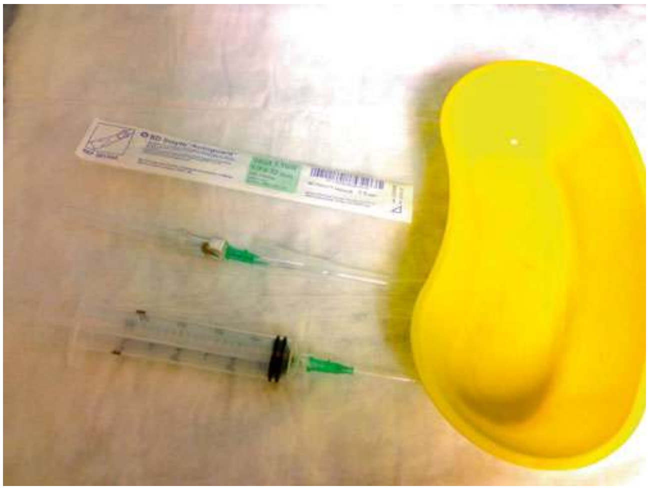

Fig. 57.3 Example of a typical syringe used in irrigation

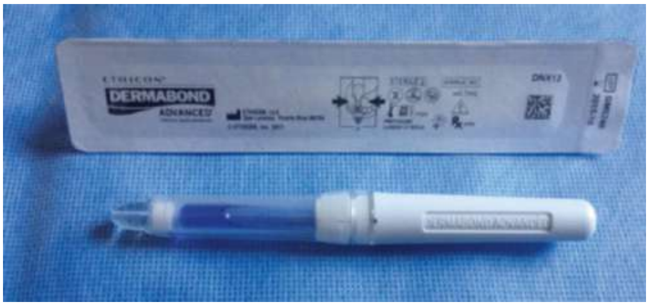
ETHICON
DERMABOND
ADVANCED

Fig. 57.4 Example of cyanoacrylate

57.4 Procedures (Based on the Method Chosen)

57.4.1 Alligator Forceps

- Ideal for easily visualized FBs, especially graspable FBs in the lateral third of the ear canal.
  - May have limited success with rounded objects.
- Care must be taken for those FBs that are easily broken apart (e.g., peas) to ensure that all of the matter has been removed.
  - May need irrigation subsequent to the attempt.
- Explain the procedure to the patient to facilitate cooperation.
- Pull the ear up and out to straighten the ear canal.
- Consider using a specialized otoscope with magnification (an operating otoscope) to maximize visualization.
- The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Attempt to grasp the FB with the forceps.
  - If using an ear speculum, attempt to bring the object as close as possible to the speculum in order to stabilize it and slowly remove both.
- After removal, ensure the complete removal of the object and ensure that no other FBs are present.

57.4.1.1 Pitfalls and Complications

- Pushing the object deeper into the auditory canal.
- Inability to totally remove the FB.
- Attempted removal of an embedded FB may cause significant bleeding.

57.4.2 Curved Hook

- Ideal for spherical objects in the lateral third of the auditory canal.
- Can be used for non-graspable objects especially in the lateral third of the auditory canal.
- Direct visualization is necessary to avoid trauma to the ear canal.
- Consider using a specialized otoscope with magnification (an operating otoscope) to maximize visualization.
- Explain the procedure to the patient to facilitate cooperation.
- The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pull the ear up and out to straighten the ear canal.
- Pass the hook behind the object, or if the object has holes, place the hook inside the hole to facilitate removal.
  - In order for this technique to succeed, there must be sufficient area near the object to facilitate the hook going behind it.
- Slowly withdraw the hook, thereby removing the FB.
- After removal, ensure the complete removal of the object and ensure that no other FBs are present.

57.4.2.1 Pitfalls and Complications

- Pushing the object deeper into the auditory canal.
- This technique cannot be used if the object cannot be directly visualized because significant trauma may arise owing to the misplacement of the hook.
- Attempted removal of an embedded FB may cause significant bleeding.
- Do not use this technique if the object is close to the tympanic membrane owing to risk of rupture.

57.4.3 Suction

- A small metal suction catheter—the Schuknecht catheter—may be used.
- Ideal for hard rounded objects in the lateral third of the ear canal.
  - Does not work as well for those objects closer to the tympanic membrane owing to lack of suction and the narrow ear canal
- Direct visualization of the object is mandatory.
  - The catheter must be placed directly on the object for this technique to succeed.
- Explain the procedure to the patient to facilitate cooperation.
- The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pull the ear up and out to straighten the ear canal.
- Place the suction against the FB and slowly withdraw.

57.4.3.1 Pitfalls and Complications

- Pushing the object deeper into the auditory canal
- May not work for objects too tightly entrenched in the ear canal

57.4.4 Irrigation

- Useful for small particulate matter unable to be grasped by forceps or too small for a hook to pull out.
  - Cannot be used for button batteries or vegetable matter.
    - Irrigation may cause an alkaline necrosis with the button battery.
    - Vegetable matter may expand, increasing the difficulty of removal.
- Explain the procedure to the patient to facilitate cooperation.
  - Some discomfort may be involved, including a sensation of pressure and a feeling cold wetness from the water used.
- The patient may be seated or supine with the head slightly turned to the affected side to facilitate collection of fluid.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pull the ear up and out to straighten the ear canal.
- Place a collection device next to the patient's ear, or alternatively, have the patient hold a basin.
- Attach an angiocath to a 20- or 60-mL syringe and place the tip of the catheter within the ear canal, or use a Waterpik system.
- Slowly inject the solution into the ear canal.
  - Care must be taken to limit the force used owing to the potential for tympanic membrane rupture.
- Direct the stream as much as possible toward the rear of the object(s) to be removed.
- Continually assess the ear canal to ensure that all foreign matter has been removed.
- If the patient complains of severe sharp pain, stop the procedure and assess the tympanic membrane for rupture.
  - Consider referral or otolaryngological consultation at this point.

57.4.4.1 Pitfalls and Complications

- Owing to instillation of room temperature water, a caloric response may occur including vertigo, nausea, and vomiting.
- Rupture of the tympanic membrane.

57.4.5 Cyanoacrylate

- Useful for objects that can be directly visualized.
- Ideal for hard rounded objects in the lateral third of the ear canal.
  - Does not work as well for those objects closer to the tympanic membrane owing to lack of visualization
- Direct visualization of the object is mandatory.
  - The stick with the cyanoacrylate must be placed directly on the object for this technique to succeed.
- Consider using a specialized otoscope with magnification (an operating otoscope) to maximize visualization.
- Explain the procedure to the patient to facilitate cooperation.
- The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pull the ear up and out to straighten the ear canal.
- Place a small amount of cyanoacrylate on the end of a cerumen spoon or wooden end of a cotton-tipped swab.
- Using direct visualization, place the stick in the ear canal and touch the FB.
- Leave the stick on the object for approximately 30–60 s to facilitate drying of the glue.
- Remove the stick and the FB.

57.4.5.1 Pitfalls and Complications

- This technique cannot be used if the object cannot be seen.
- The object must be relatively smooth to facilitate placement of the glue.

Selected Reading

Hanson RM, Stephens M. Cyanoacrylate-assisted foreign body removal from the ear and nose in children. J Paediatr Child Health. 1994;30:77–8.

Kadish H. Ear and nose foreign bodies: it is all about the tools. Clin Pediatr. 2005;44:665–70.

Treatment of Auricular Hematoma

58

Bobby K. Desai

58.1 Indications

- Presence of auricular hematoma
  - These occur after significant force is applied to the ear.
  - This type of hematoma will separate the perichondrium from the cartilage.
    - Untreated, this hematoma prevents the development of new cartilage, subsequently deforming the auricle, causing *cauliflower ear* (Figs. 58.1 and 58.2).

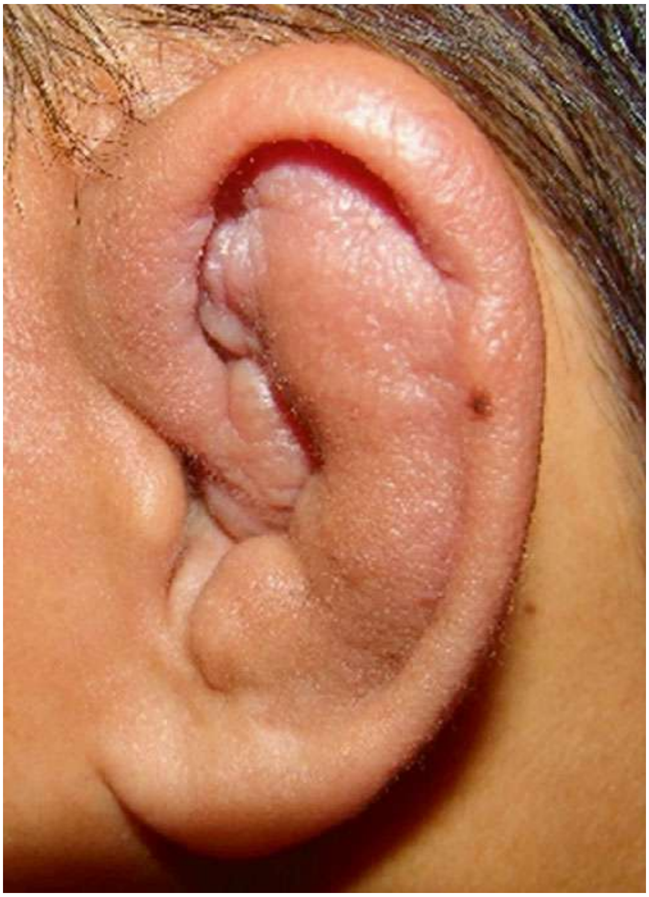

Fig. 58.1 Example of a cauliflower ear

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_58

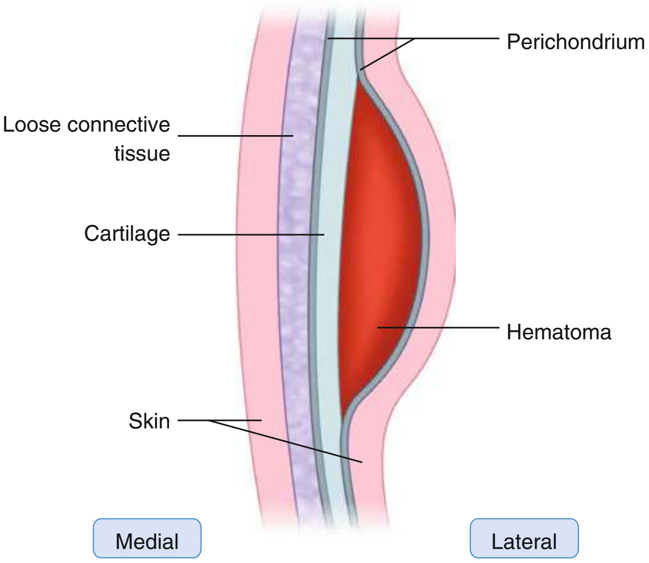

Fig. 58.2 Schematic of auricular hematoma

58.2 Contraindications

- Uncooperative patient.
- Evidence of infection mandates plastic surgery or otolaryngology consultation.
- Significant trauma to area around the ear.

58.3 Materials and Medications

- 5-cc syringe
- 20-gauge needle (may use larger-gauge needle, but smaller ones may not allow for adequate suctioning)
- Topical antiseptic
- Sterile drapes, gown, eye protection, and gloves
- 1 % lidocaine for anesthesia
- Scalpel (#11 or #15 blade)
- Suction setup
- Saline for irrigation
- Hemostat and forceps
- Penrose drain or plain gauze packing
- Xeroform gauze
- Laceration tray setup
  - Needle driver
  - 4–0 Nonabsorbable suture
- Antibiotic ointment
- Gauze bandages

58.4 Procedure (Needle Aspiration)

1. Explain the procedure to the patient.
2. Informed consent may be required.
3. The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
4. Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
   - Consider procedural sedation for children.
5. Carefully clean the area with an antiseptic solution of choice (povidone-iodine [Betadine] or chlorhexidine solution) and allow to dry.
6. Anesthetize the area where the hematoma is of greatest diameter with 1 % lidocaine.
7. Inject the hematoma itself to promote increased anesthesia.
8. The clinician should prepare for the procedure by practicing universal precautions.
9. Drape the patient's ear.
10. Attach a 20-gauge needle to a syringe.
    - A 5-cc syringe should be sufficient.
11. Attempt to aspirate the hematoma contents with the prepared syringe (Fig. 58.3).

12. "Milking" the hematoma may be required to fully remove the clot.
    - If the hematoma has been present for sufficient time for the blood to completely clot, needle aspiration will fail, and an alternative technique, described later, may be necessary to completely evacuate the hematoma.
13. Place antibiotic ointment around the injection site.
14. Loosely dress the area.
15. Refer to an otolaryngologist or a plastic surgeon for follow-up care.

If needle aspiration fails, a larger incision may be required using the following procedure:

1. Perform the previously discussed steps.
2. Consider a local field block of the greater auricular nerve.
3. Make an incision over the hematoma, following the anatomy of the helix (Fig. 58.4).
    - The incision should be 5–6 mm in length and curvilinear, following the concavity of the helix.
    - More than one incision may be required.
4. Remove the hematoma with hemostats or forceps.
    - Suctioning may be used as well.
5. Copiously irrigate the cavity with saline to make certain of complete removal of the hematoma.
6. Insert the Penrose drain or plain packing into the incision.
7. Antibiotic ointment may be applied to the incision site.
8. Place the Xeroform gauze over the incision site.
9. Apply a compression dressing to the ear (Fig. 58.5).
    - Place dry cotton into the ear canal to protect from possible drainage and subsequent infection (Fig. 58.6).
    - Mold Xeroform gauze in and around the contours of the helix (Fig. 58.7).
    - With the aid of an assistant, place gauze behind the pinna and over the Xeroform gauze already molded to the anterior portion of the ear (Fig. 58.8).
    - Place more fluffed gauze over the ear (Fig. 58.9).
    - While the assistant holds the anterior and posterior gauze in place, use a web roll (Kerlix) around the head to keep the gauze in place. Then wrap the head with an ACE wrap or stretchy bandage (Figs. 58.10 and 58.11).
    - These steps are critical to maintain equal compression of the ear, which will help prevent future deformity.
    - Alternatively, dental rolls can be secured with suture to the anterior and posterior site of the drained hematoma (Fig. 58.12).
10. Place the patient on prophylactic antibiotics.
11. Refer to an otolaryngologist or a plastic surgeon for follow-up care.

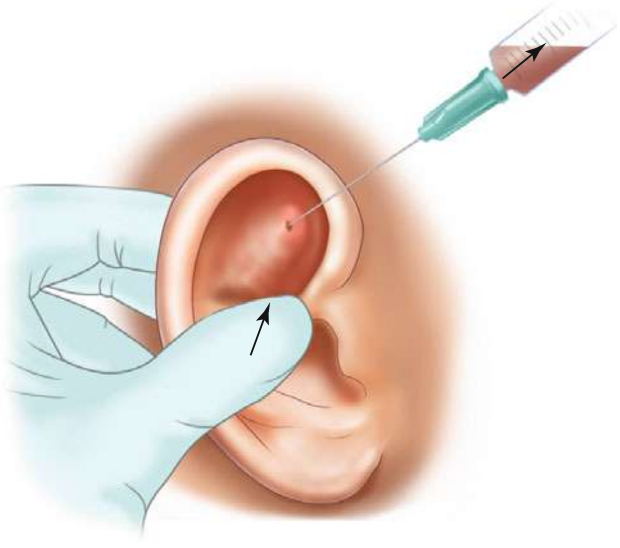

Fig. 58.3 Aspiration of a hematoma

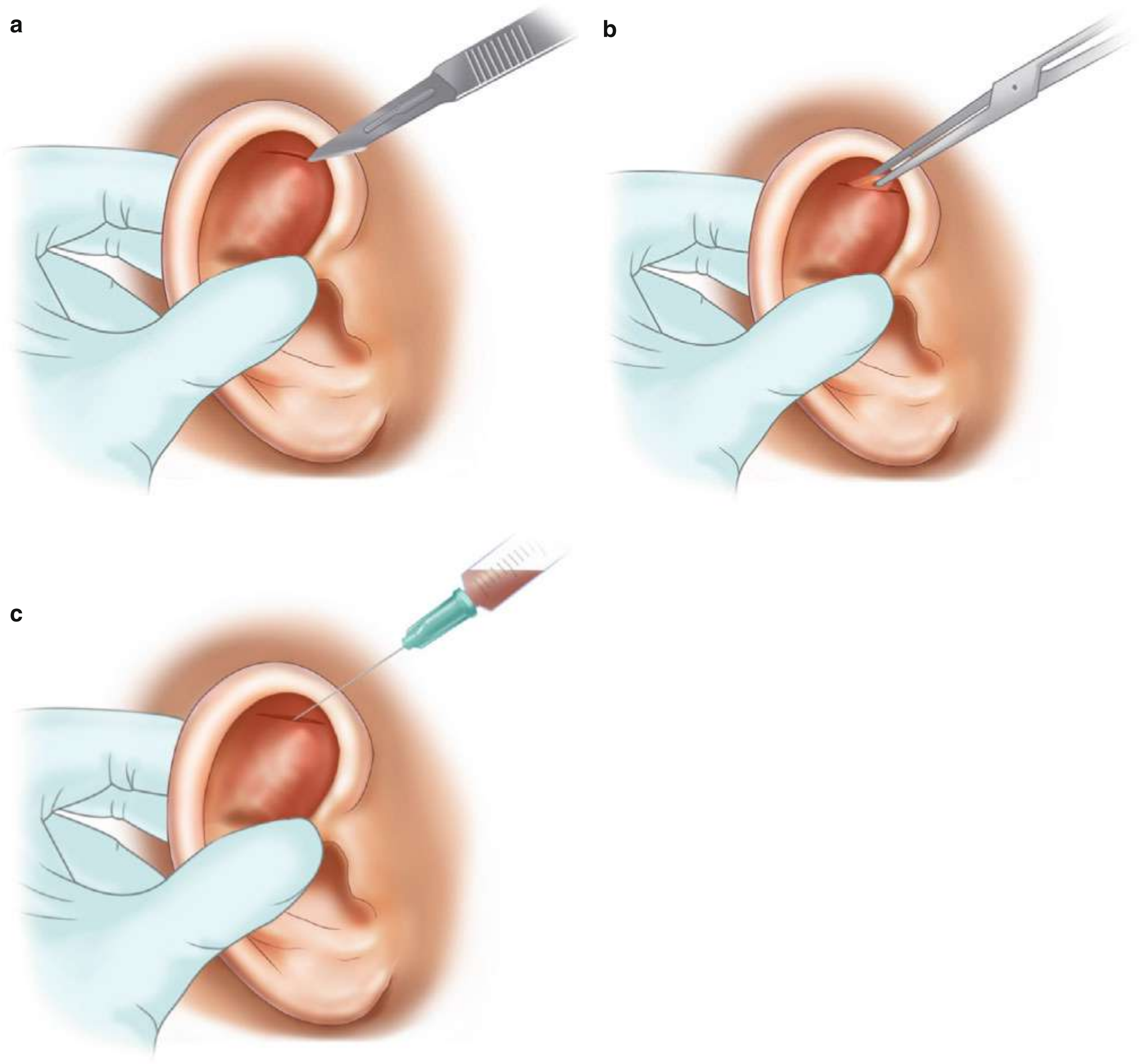

Fig. 58.4 Incision of an auricular hematoma. (**a**) Incision; (**b**) removal of clot with forceps; (**c**) aspiration of blood with syringe

Fig. 58.5 Materials needed for compression dressing

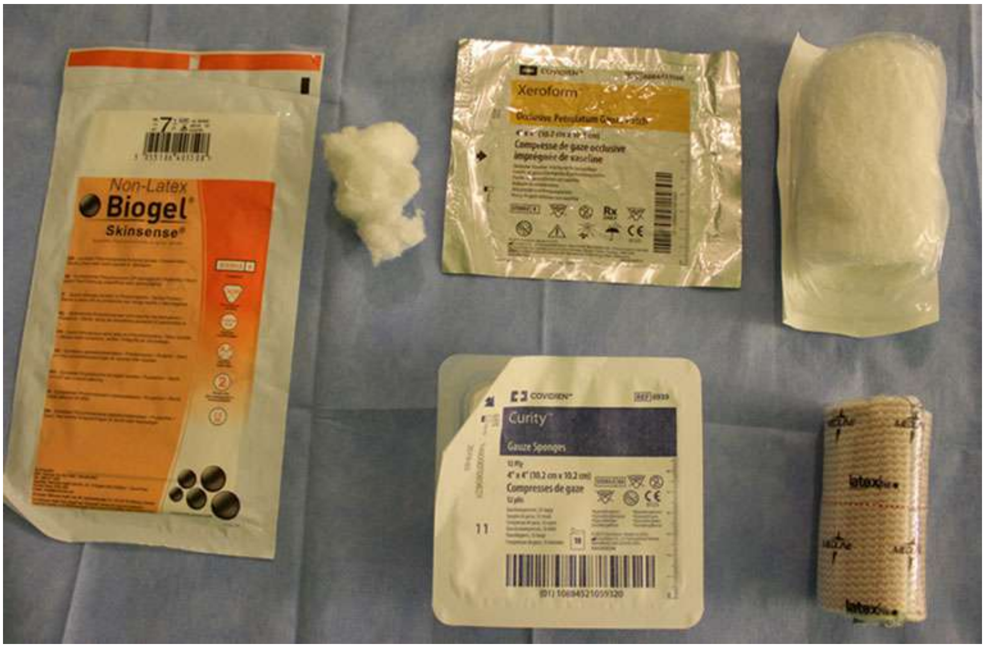
Non-Latex
Biogel
Skinsense
Xeroform
Occlusive Petrolatum Gauze
Compresse de gaze occlusive imprégnée de vaseline
COVIDIEN
Curity
Gauze Sponges
4" x 4" (10.2 cm x 10.2 cm)
Compresses de gaze
latex

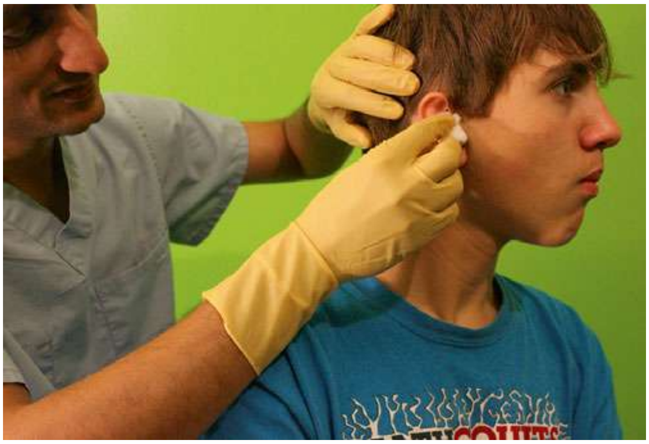

Fig. 58.6 Place cotton in the ear to prevent drainage into ear

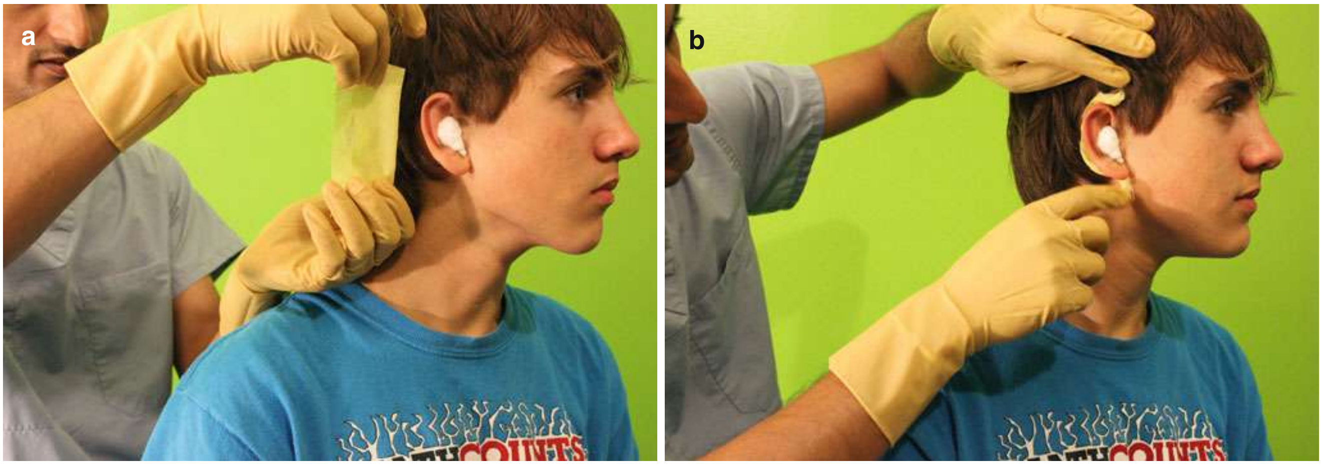

Fig. 58.7 (**a**, **b**) Mold Xeroform gauze around ear helix

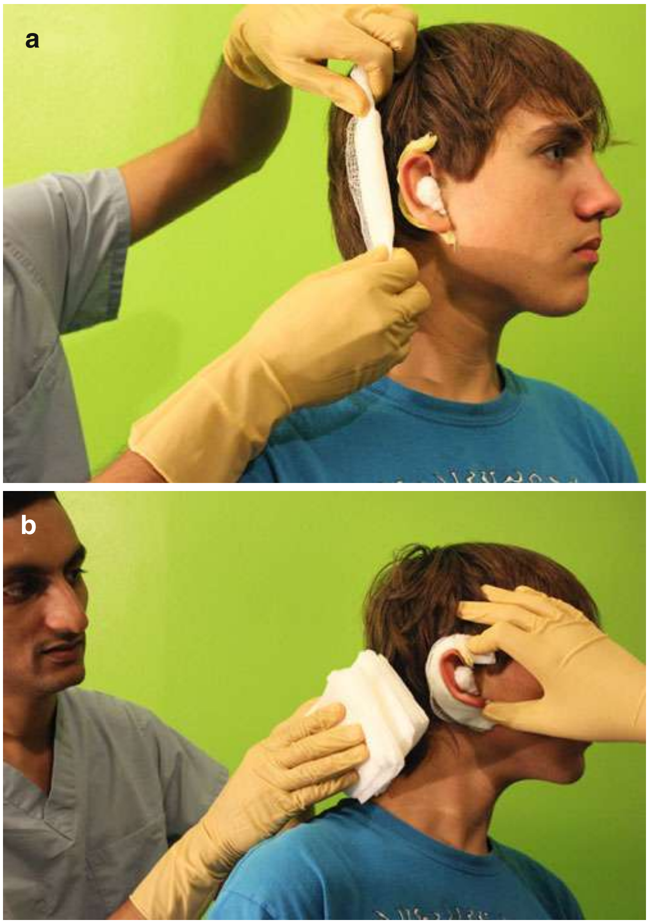

Fig. 58.8 (**a**, **b**) Mold sterile gauze on top of Xeroform gauze around ear helix

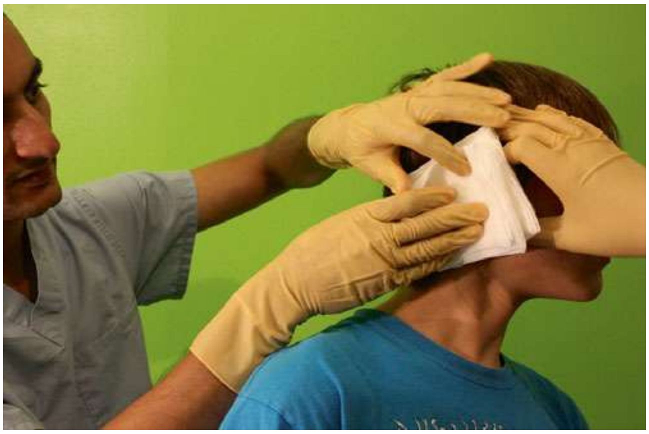

Fig. 58.9 Pad outside of ear with 4×4 gauze pads

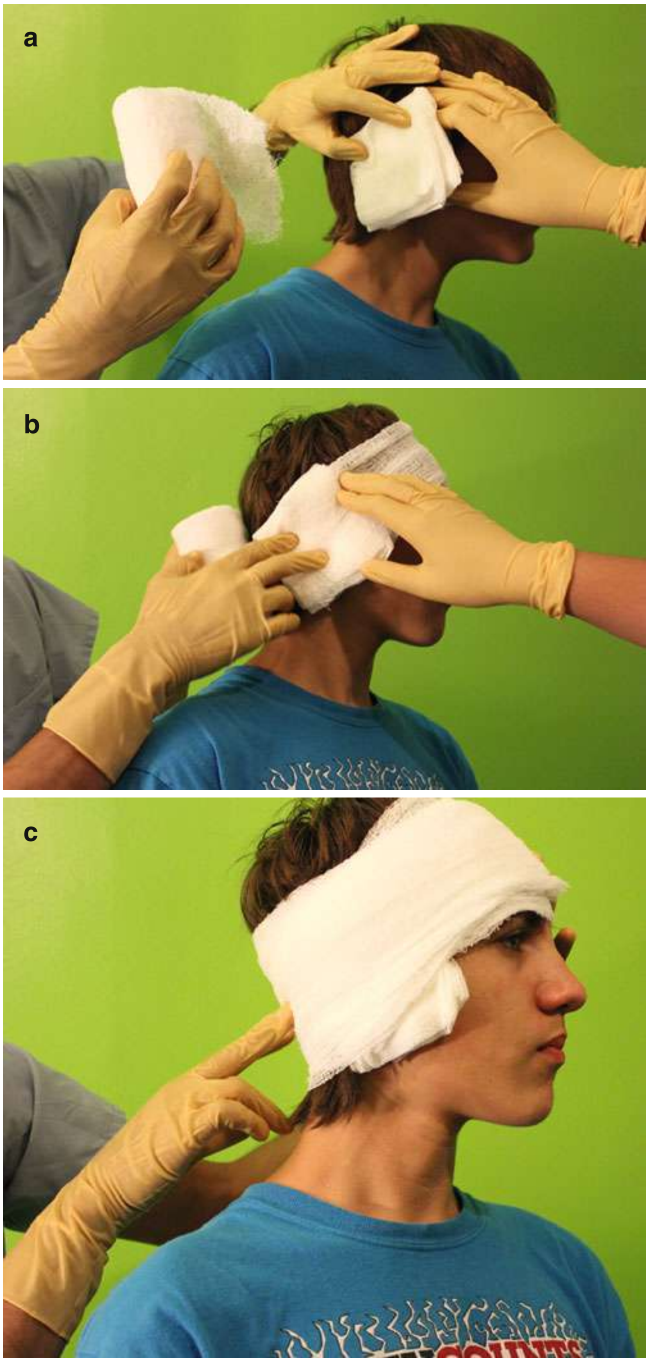

Fig. 58.10 (**a**–**c**) Wrap head with Kerlix (sterile gauze roll)

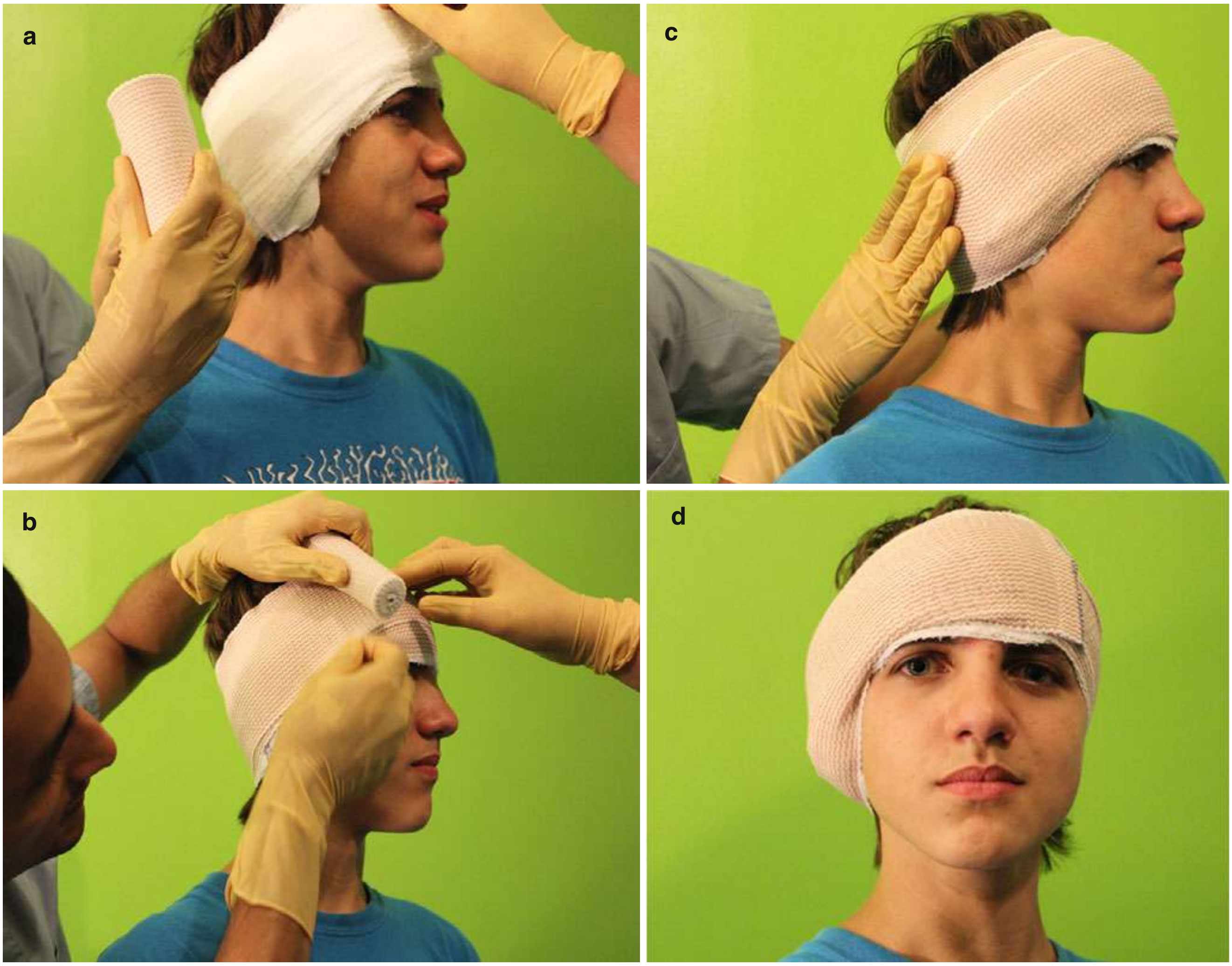

Fig. 58.11 (**a**–**d**) Wrap head with elastic bandage

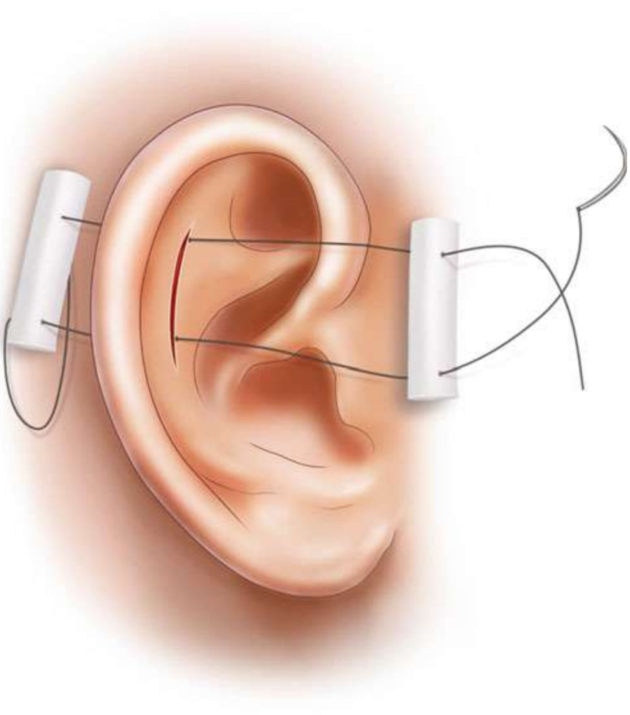

Fig. 58.12 Dental rolls secured to the ear

58.5 Pitfalls and Complications

- Bleeding
- Infection—perichondritis
- Deformity of the auricle
- Inadequate clot removal
- Hematomas present for several days warrant a plastic surgery or an otolaryngology consultation

Acknowledgement The authors would like to thank Thor Shiva Stead for serving as the subject in many of the photographs in this chapter.

Selected Reading

O'Donnell BP, Eliezri YD. The surgical treatment of traumatic hematoma of the auricle. Dermatol Surg. 1999;25:803–5.

Starck WJ, Kaltman AI. Current concepts in the surgical management of traumatic auricular hematoma. J Oral Maxillofac Surg. 1992;50:800–2.

Incision and Drainage of Peritonsillar Abscess

59

Melinda W. Fernandez and Bobby K. Desai

59.1 Indications

- Peritonsillar abscess

59.2 Contraindications

- Absolute
  - Malignancy
  - Vascular malformations
- Relative
  - Pediatric patient
  - Severe trismus
  - Uncooperative patient

59.3 Materials and Medications

- #11 or #15 scalpel
- 27-gauge 1.5-inch needle, 18- to 20-gauge 1.5-inch or longer needle
- 5-mL syringe, 10- to 20-mL syringe
- Tape
- Trauma shears
- Viscous lidocaine
- Lidocaine 1 % with epinephrine
- Laryngoscope with MAC 3 or 4 *or* tongue blade(s) and headlamp or other light source
- Suction setup with Frazier or tonsil suction tip

59.4 Procedure: Aspiration

1. Informed consent may be required.
2. Raise head of bed to at least 60° and place a pillow or other support behind the patient's head.
3. Prepare an 18- to 20-gauge 1.5-inch needle on a 10- to 20-mL syringe with a needle guard.
   - This can be accomplished by using trauma shears to cut 1–1.5 cm off the distal plastic needle cover and replacing the cover over the needle.
   - This now creates a guard to prevent deep penetration into vascular structures (Fig. 59.1).
4. Topically anesthetize area with viscous lidocaine (alternatively, can use Cetacaine [benzocaine, tetracaine hydrochloride, and butamben] spray)
5. Using lidocaine 1 % with epinephrine and a 27-gauge 1.5-inch needle on a 5-mL syringe, infiltrate 1–2 mL into the area. Blanching should be apparent.
6. Palpate the oropharynx with the gloved finger to evaluate for fluctuance.
7. Assemble a MAC 3 or 4 intubation blade on a laryngoscope handle and open into the light-on position. Insert the blade into the patient's mouth, and advance the blade as far posteriorly as possible without inducing gagging. Have the patient hold the laryngoscope handle. Alternatively, use a tongue blade and head lamp or other light source, but the laryngoscope technique allows for an unobstructed view of the area and the weight of the handle helps hold the patient's mouth open.
8. Using the previously prepared needle guard on the long 18- to 20-gauge needle on a 12-mL syringe, insert the needle into the most fluctuant area (as determined from the previous examination), aspirating as the needle advances (Fig. 59.2). The most fluctuant area will usually be located in the superior pole of the tonsil.

- It is very important to be careful not to angle the needle laterally toward the carotid artery. Also remember, this is a peritonsillar abscess so do not aspirate the tonsil itself.

M.W. Fernandez, MD
Department of Emergency Medicine, University of Florida Health, Gainesville, FL, USA
e-mail: mindyfernandez@ufl.edu

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_59

9. After aspiration, if there is a pus return, remove as much purulent drainage as possible. Typical drainage is between 2 and 6 mL. If a large amount of pus is obtained (>6 mL), incision and drainage (I&D) may be indicated. (See later for details on I&D.)
10. If the aspirate did not return pus, attempt aspiration again by moving the insertion site 1 cm inferiorly to the middle pole of the peritonsillar space. If there is still no pus aspirated, make a final attempt by moving again 1 cm inferiorly to the lower pole of the peritonsillar space.
11. Suction should be set up, readily available, and turned on. Use suction to prevent the patient from aspirating or swallowing any purulent drainage.
12. Expect a small amount of bleeding when the procedure is complete and the needle is removed.

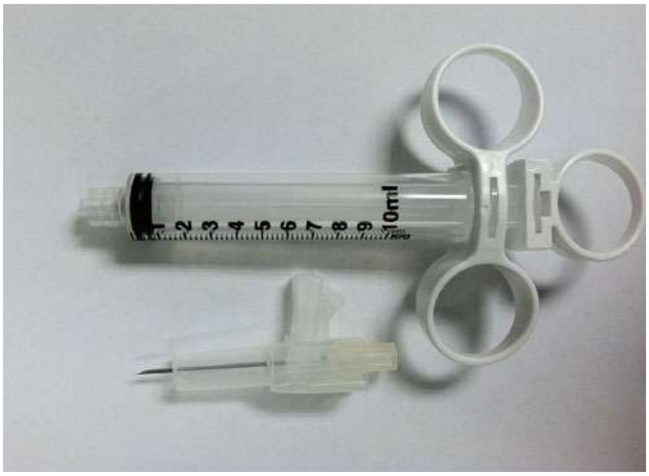

Fig. 59.1 Guard to prevent deep penetration of vascular structures

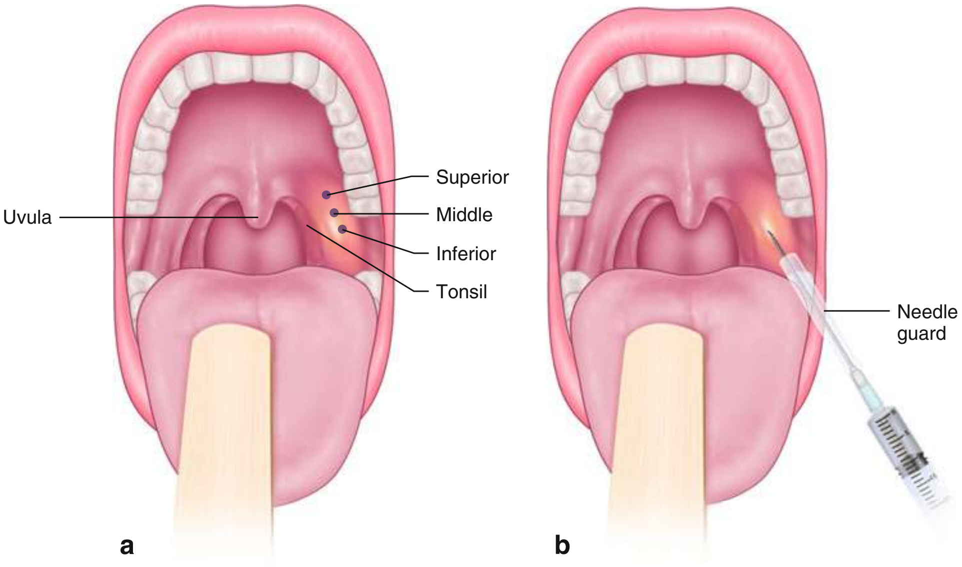

Fig. 59.2 Aspiration of a peritonsillar abscess: (**a**) Aspirate superior pole first. If no pus aspirated, move 1 cm inferior to the middle pole. If still no pus, make final attempt by moving 1 more cm inferiorly to the inferior pole. (**b**) Demonstrates use of needle guard

59.5 Procedure: Incision and Drainage

1. Raise the head of the bed and anesthetize the area as described previously.
2. Have suction set up with a Fraser or tonsil tip, available, and turned on.
3. Have the patient hold the laryngoscope handle after the blade has been inserted orally as described previously.
4. Fashion a blade guard on a #11 or #15 blade scalpel by taping over the blade with only 0.5 cm of the blade exposed at the tip. Incise the area that was previously aspirated with a small stab incision. Suction the area.
5. Insert a curved Kelly clamp into the incision and open gently to enter the abscess cavity and break up loculations (Fig. 59.3). Suction as necessary.
6. Have the patient gargle with a saline solution and expectorate.
7. Do not pack the abscess cavity.
8. Observe for 2–4 h for bleeding.

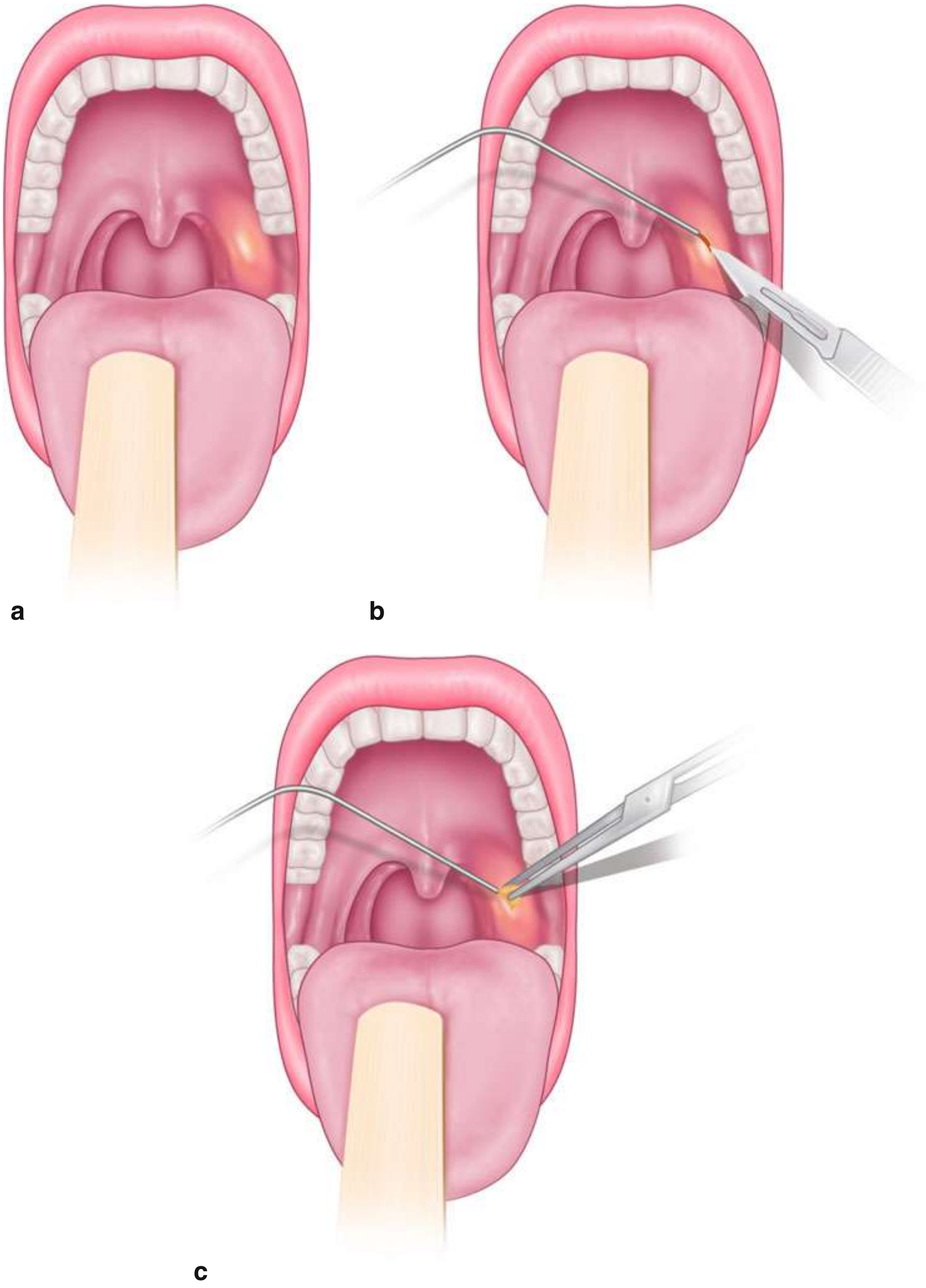

Fig. 59.3 Incision and drainage of peritonsillar abscess: (**a**) Gain access with appropriate lighting. (**b**) Fashion a blade guard on a #11 or #15 blade scalpel by taping over the blade with only 0.5 cm of the blade exposed at the tip. Incise the area that was previously aspirated with a small stab incision. (**c**) Insert a curved Kelly clamp into the incision and open gently to enter the abscess cavity and break up loculations. Do not pack!

59.6 Complications

- Aspiration or incision of the carotid artery
- Excessive bleeding
- Aspiration of purulent material
- Pain

59.7 Pearls and Pitfalls

- Pearls
  - Airway protection should always be considered. Intubation may be necessary for very large abscesses with airway compromise.
  - Bedsides ultrasound is a valuable tool in confirming the diagnosis of peritonsillar abscess. Perform using the intracavitary probe covered with a sterile glove or other appropriate probe cover and insert into the oropharynx. The abscess will appear as any other abscess: an encapsulated, hypoechoic structure. Doppler flow can also be used to locate the carotid artery relative to the abscess.
  - Empirical oral antibiotics after aspiration or I&D are considered usual care and should cover for group A streptococcus and oral anaerobes.
    - Amoxicillin/clavulanate or clindamycin is the most commonly used.
  - A single dose of high-dose steroids may be helpful in relieving symptoms.
  - Patients who cannot tolerate oral fluids, cannot take oral antibiotics, or who appear to have a toxic response should be admitted. Others may be discharged with 24-h follow-up.
- Pitfalls
  - The carotid artery is located approximately 2.5 cm posterolaterally to the tonsils. Take care that the needle is not inserted too far laterally or the risk of aspirating the carotid artery is increased.
  - Incision into the tonsil itself may cause excessive bleeding. Aspiration or incision into the tonsil will likely miss the abscess altogether and may result in misdiagnosis.
  - There is a 1–15 % failure and recurrence rate.

Selected Reading

Afarian H, Lin M. Tricks of the trade—say “ah!”—needle aspiration of peritonsillar abscess. ACEP News. 2008;27(5):38.

Braude DA, Shalit M. A novel approach to enhance visualization during drainage of peritonsillar abscess. J Emerg Med. 2008;35:297–8.

Galioto NJ. Peritonsillar abscess. Am Fam Physician. 2008;77:199–202.

Ozbek C, Aygenc E, Tuna EU, Selcuk A, Ozdem C. Use of steroids in the treatment of peritonsillar abscess. J Laryngol Otol. 2004;118:439–42.

Roberts J, Hedges J, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders; 2009. p. 1184–9.

Vieira F, Allen SM, Stocks RM, Thompson JW. Deep neck infection. Otolaryngol Clin North Am. 2008;41:459–83.

Incision and Drainage of Sublingual Abscess

60

Melinda W. Fernandez and Bobby K. Desai

60.1 Indications

- Sublingual abscess (Fig. 60.1)

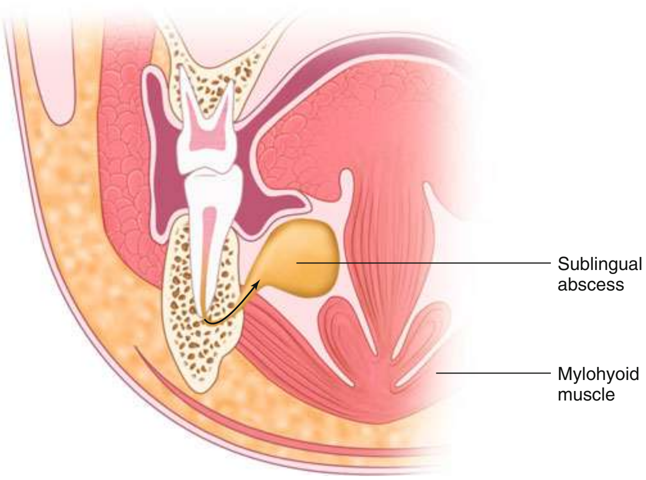

Fig. 60.1 Sublingual abscess

M.W. Fernandez, MD
Department of Emergency Medicine, University of Florida Health, Gainesville, FL, USA
e-mail: mindyfernandez@ufl.edu

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_60

60.2 Contraindications

- Relative
  - Pediatric patient
  - Severe trismus—may need sedation or drainage in the operating room
  - Uncooperative patient—may need sedation or drainage in the operating room
  - Coagulopathy, patients taking anticoagulants, or patients with a known bleeding disorder

60.3 Materials and Medications (Fig. 60.2)

- #11 scalpel
- Hemostats
- Penrose drain
- 4-0 silk suture
- 25- to 27-gauge 1.5-inch needle for injecting anesthetic
- 18- to 20-gauge needle for drawing up anesthetic
- 5- to 10-mL syringe or control syringe
- Viscous lidocaine (or other topical anesthetic)
- Lidocaine 1 % with epinephrine
- Light source
- Suction setup with tonsil suction tip
- Culture swab

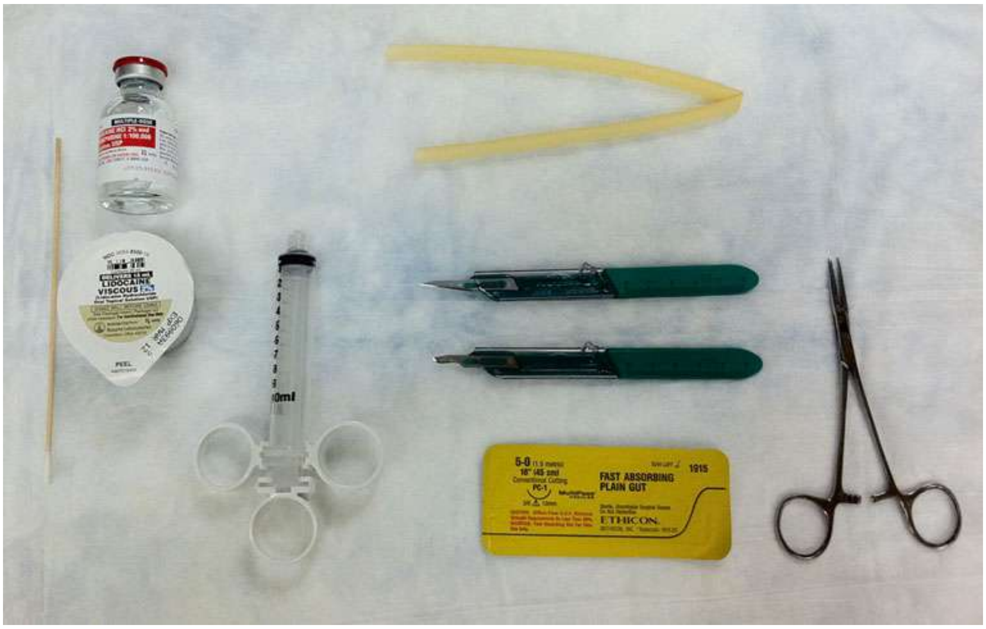

Fig. 60.2 Materials needed: #11 scalpel, hemostats, Penrose drain, 4-0 silk suture, 25–27-gauge 1.5-inch needle for injecting anesthetic, 18- to 20-gauge needle for drawing up anesthetic, 5- to 10-mL syringe or control syringe, viscous lidocaine (or other topical anesthetic), lidocaine 1 % with epinephrine, light source, suction setup with tonsil suction tip, culture swab (optional)

60.4 Procedure: Incision and Drainage

1. After explaining the procedure, risks, and benefits to the patient, put the head of the bed in the most comfortable working position for the clinician and the patient. Alternatively, the clinician may also elect to position the patient in an upright procedure chair that has multiple positions.
2. Apply viscous lidocaine topically with a cotton-tipped swab to the area to be injected with anesthetic. Leave in place for a minute or 2. Alternatively, spray the area with a topical anesthetic spray.
3. Have suction set up with a tonsil tip, available, and turned on.
4. Draw up the appropriate amount of lidocaine with epinephrine into a 5-mL syringe (control syringe if available).
5. Direct light source to area to be worked on. *Adequate lighting is essential.* Options include a headlamp or an overhead adjustable light.
6. Change needle to the 27-gauge needle and inject 1–2 mL of lidocaine with epinephrine using an inferior alveolar block. Alternatively, anesthetize the floor of the mouth around the most fluctuant area. Avoid injecting through infected tissue to avoid possible spread into deeper spaces.
7. Using the scalpel, make an intraoral stab incision superficially at the lowest point of the pus accumulation (Fig. 60.3). This will facilitate evacuation of pus under gravity. Have the suction catheter in the mouth and allow the purulent material to drain into the suction.
8. Obtain a specimen of the purulent fluid for culture and sensitivity.
9. Insert a hemostat into the incision to facilitate drainage, but do not open it up to avoid injury to neurovascular structures. Gently massage the soft tissue surrounding the abscess to assist drainage. Suction as necessary to avoid swallowing or aspirating the pus.
10. Once adequate drainage has been achieved, place a small Penrose drain (or other rubber-type drain) into the cavity, and stabilize on one side with a silk suture that goes through the drain and the mucosa (Fig. 60.4).
11. Have the patient rinse and spit with a half-strength peroxide solution followed by either water or saline.
12. Watch for signs of bleeding or upper airway symptoms.
13. Ensure that the patient can tolerate oral fluids before discharge.

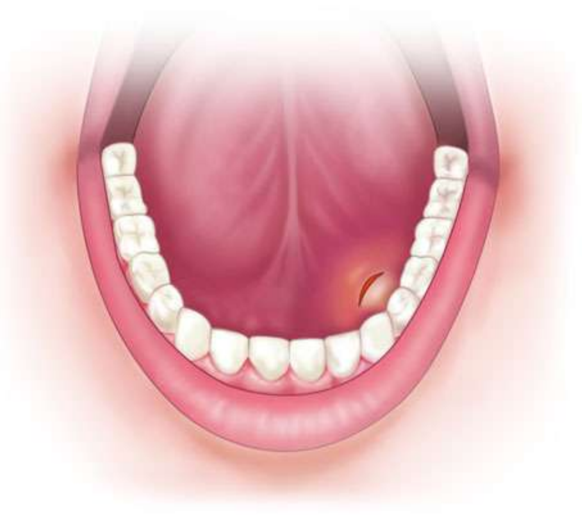

Fig. 60.3 Incise lowest portion of abscess

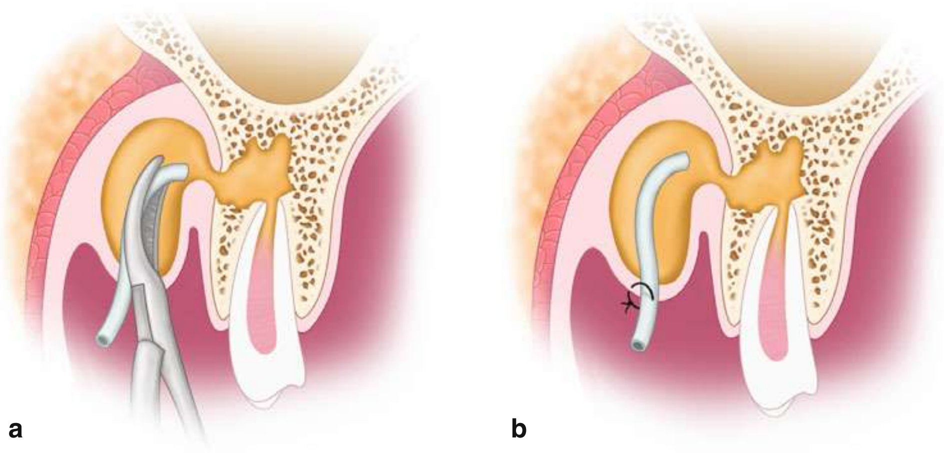

Fig. 60.4 (**a**) Using hemostats, place small Penrose drain into the abscess cavity. (**b**) Suture Penrose in place with a silk suture that goes through the drain and the mucosa

60.5 Complications (Generally Minimal)

- Excessive bleeding (apply pressure—to avoid potential nerve injury, do not cauterize or ligate unless absolutely necessary)
- Aspiration of purulent material (have the patient sit upright and use suction as the abscess is incised)
- Pain

60.6 Pearls and Pitfalls

- Pearls
  - Airway protection is paramount. Infections in this space can quickly spread to the submandibular region and can compromise the airway owing to swelling (Ludwig's angina).
    - Assess for elevation of the floor of the mouth and tongue, ability to lie supine, drooling, stridor, and restlessness.
    - If these signs are present, emergent surgical airway will be necessary before incision and drainage (I&D).
  - The sublingual space is bounded superiorly by the oral mucosa and inferiorly by the mylohyoid muscle.
  - Infected premolars and first molars drain into this space because the apices of their roots are located superior to the mylohyoid muscle.
  - The source molar should be extracted as soon as possible after I&D.
  - Empirical oral antibiotics after I&D should cover for group A streptococcus and oral anaerobes. Penicillin remains the drug of choice, but clindamycin or amoxicillin–clavulanic acid can be substituted.
- Pitfalls
  - The lingual artery, vein, and nerve are contained in the posterolateral area of the floor of the mouth, and the hypoglossal nerve is nearby. These must be avoided when performing I&D for sublingual abscesses.

Selected Reading

Flynn TR, Shanti RM, Levi MH, Adamo AK, Kraut RA, Trieger N. Severe odontogenic infections, part 1: prospective report. J Oral Maxillofac Surg. 2006;64:1093–103.

Reichman E, Simon R, editors. Emergency medicine procedures. New York: McGraw-Hill Education; 2003. p. 1342–5.

Roberts J, Hedges J, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders; 2009. p. 1184–9.

Vieira F, Allen SM, Stocks RM, Thompson JW. Deep neck infection. Otolaryngol Clin North Am. 2008;41:459–83.

Incision and Drainage of Parotid Duct Abscess

61

Melinda W. Fernandez and Bobby K. Desai

61.1 Indications

- Parotid duct abscess

61.2 Contraindications

- Absolute
  - None
- Relative
  - Pediatric patient
  - Severe trismus—may need sedation or drainage in the operating room
  - Uncooperative patient—may need sedation or drainage in the operating room
  - Coagulopathy, patients taking anticoagulants, or patients with a known bleeding disorder

61.3 Materials and Medications (Fig. 61.1)

- #11 scalpel
- 4×4 gauze
- Hemostat
- Penrose drain or ¼-inch packing gauze
- Light source (headlamp or overhead light)
- Culture swab
- 4-0 silk suture
- 25- to 27-gauge needle, 1.5–2 inches long
- 18-gauge needle to withdraw anesthetic from vial
- 5-mL syringe
- Viscous lidocaine or other topical anesthetic
- Lidocaine with epinephrine
- Suction setup with Frazier or tonsil suction tip

M.W. Fernandez, MD
Department of Emergency Medicine,
University of Florida Health, Gainesville, FL, USA
e-mail: mindyfernandez@ufl.edu

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_61

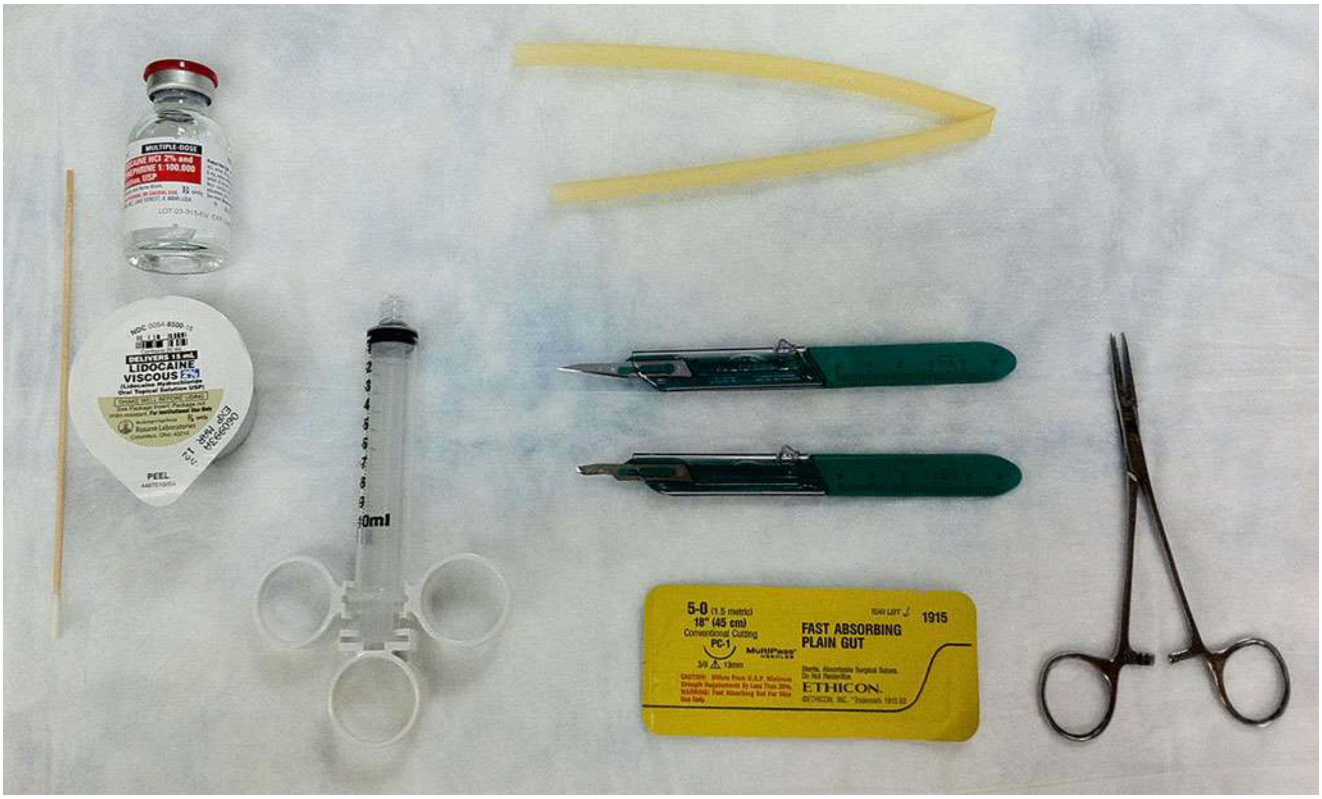

Fig. 61.1 Materials needed: #11 scalpel, hemostats, Penrose drain, 4-0 silk suture, 25- to 27-gauge 1.5-inch needle for injecting anesthetic, 18- to 20-gauge needle for drawing up anesthetic, 5- to 10-mL syringe or control syringe, viscous lidocaine (or other topical anesthetic), lidocaine 1 % with epinephrine, light source, suction setup with tonsil suction tip, culture swab (optional)

61.4 Procedure (Figs. 61.2 and 61.3)

1. After explaining the procedure, risks, and benefits to the patient, put the head of the bed in the most comfortable working position for the operator and the patient. Alternatively, position the patient in an upright procedure chair that has multiple positions.
2. Find the most fluctuant area. Apply viscous lidocaine topically with a cotton-tipped swab to the area to be injected with anesthetic. Leave in place for a minute or 2. Alternatively, spray the area with a topical anesthetic spray.
3. Have suction setup with a tonsil tip, available, and turned on.
4. Draw up the appropriate amount of lidocaine with epinephrine into a 5-mL syringe (control syringe if available).
5. Direct light source to area to be worked on. *Adequate lighting is essential.* Options include a headlamp or an overhead adjustable light.
6. Change needle to a 25- or 27-gauge needle. Inject 1–2 mL of lidocaine with epinephrine just beneath the mucosal surface. Avoid injecting through infected tissue to avoid possible spread into deeper spaces.
7. Using the scalpel, make an intraoral stab incision superficially into the area of greatest fluctuance (Fig. 61.4). Have the suction catheter in the patient's mouth and allow the purulent material to drain into the suction.
8. Obtain a specimen of the purulent fluid for culture and sensitivity.
9. Insert a hemostat into the incision to facilitate drainage, but do not open it up in order to avoid injury to neurovascular structures. Gently massage the soft tissue surrounding the abscess to assist drainage. Suction as necessary to avoid swallowing or aspirating the pus.
10. Have the patient rinse and spit with a half-strength peroxide solution.
11. Place a small Penrose drain (or other rubber-type drain) into the cavity, and stabilize on one side with a silk suture that goes through the drain and the mucosa. Alternatively, cut a strip of ¼- inch packing gauze and insert into the abscess cavity.
12. Have the patient rinse and spit with a half-strength peroxide solution followed by either water or saline.
13. Watch for signs of bleeding or upper airway symptoms.
14. Ensure that the patient can tolerate oral fluids before discharge.

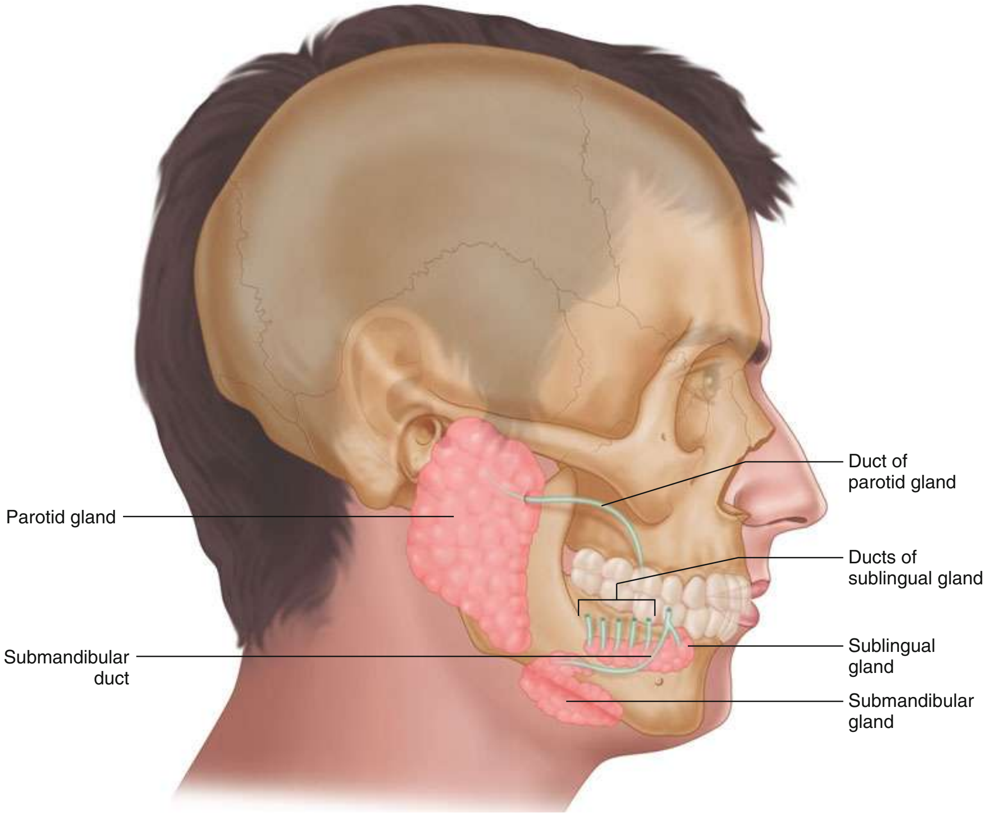

Fig. 61.2 Anatomy of the salivary glands

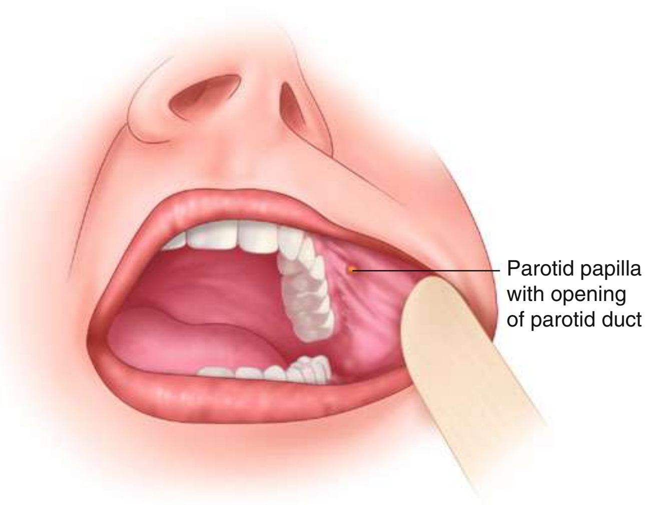

Fig. 61.3 Parotid papilla: opening of the parotid gland located adjacent to third upper molar

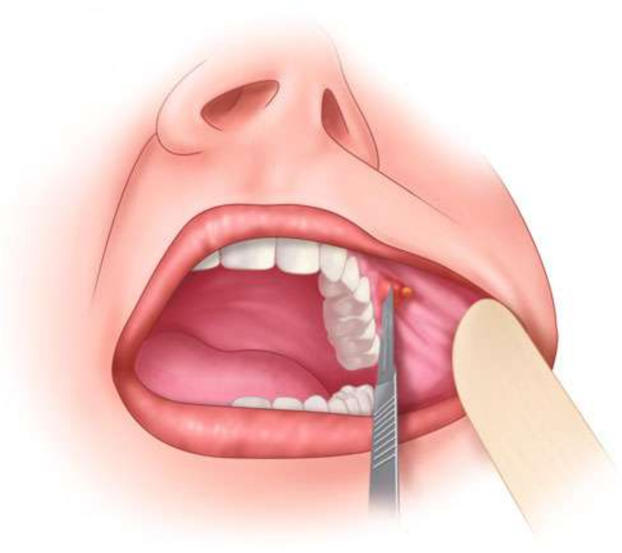

Fig. 61.4 Parotid abscess; make a superficial stab into most fluctuant area

61.5 Complications

- Complications are generally minimal but may include:
  - Excessive bleeding (apply pressure; to avoid potential nerve injury, do not cauterize or ligate unless absolutely necessary)
  - Aspiration of purulent material (have the patient sit upright and use suction as the abscess is incised)
  - Pain

61.6 Pearls and Pitfalls

- Airway protection should always be considered.
- Empirical oral antibiotics after incision and drainage are considered usual care and should cover for group A streptococcal and oral anaerobes. Penicillin remains the drug of choice, but clindamycin or amoxicillin–clavulanic acid can be substituted.
- Patients who cannot tolerate oral medications, cannot take oral antibiotics, or who appear to have a toxic response should be admitted. Pediatric patients should be admitted. Others may be discharged with 24-h follow-up.

Selected Reading

Reichman E, Simon R, editors. Emergency medicine procedures. New York: McGraw-Hill Education; 2003. p. 1346–9.

Roberts J, Hedges J, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders; 2009. p. 1184–9.

Part IX

Dental Procedures

Techniques of Mandibular Anesthesia

62

Susana Perry, Joshua Perry, and Rosalia Rey

62.1 Inferior Alveolar Nerve Block

62.1.1 Nerves Anesthetized

- Inferior alveolar, branch of the posterior division of the mandibular nerve (V3, branch of the trigeminal nerve)
- Incisive
- Mental
- Lingual (usually)

62.1.2 Areas Anesthetized (Fig. 62.1)

- Mandibular teeth to midline
- Body of the mandible
- Buccal mucoperiosteum, mucous membrane anterior to the mandibular first molar
- Anterior two thirds of the tongue and floor of the mouth (via the lingual nerve)
- Lingual soft tissues and periosteum (via the lingual nerve)

S. Perry, DMD (✉)
Department of Pediatric Dentistry, University of Florida College of Dentistry, University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: sperry@dental.ufl.edu

J. Perry, DMD
Department of Prosthodontics, University of Florida College of Dentistry, University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: jperry@dental.ufl.edu

R. Rey, DDS
Department of Restorative Dental Sciences, University of Florida College of Dentistry, University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: RREY@dental.ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_62

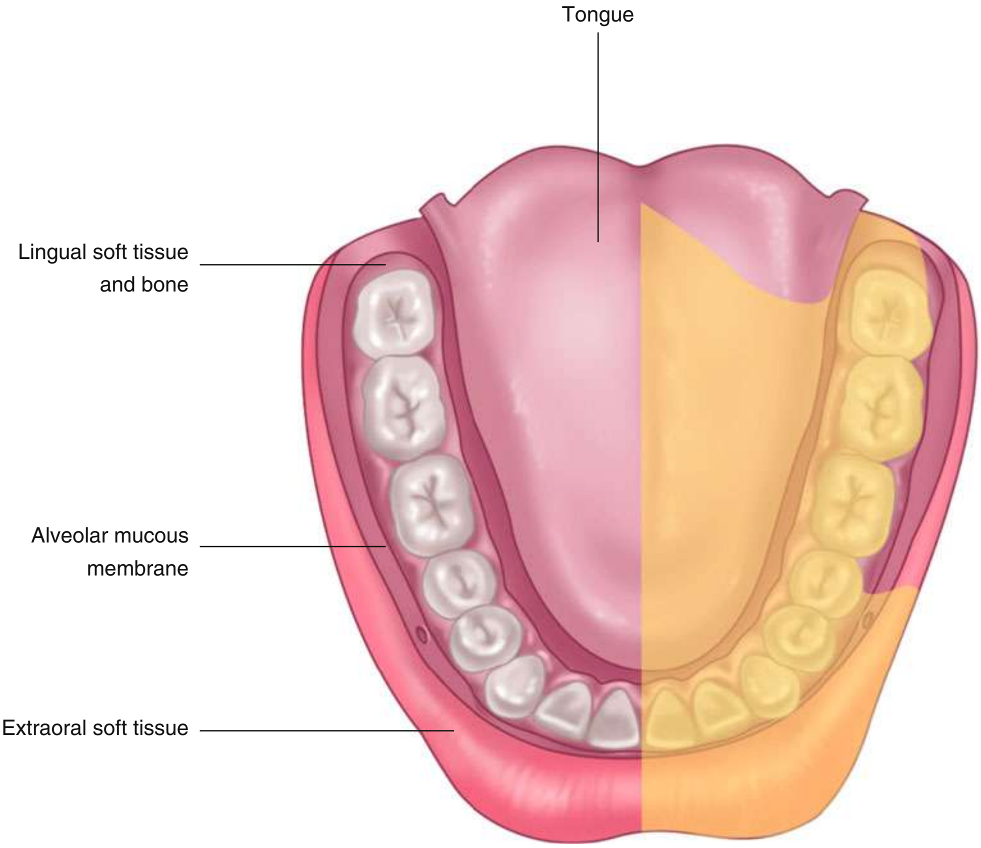

Fig. 62.1 Areas anesthetized with inferior mandibular nerve block

62.1.3 Indications

- When buccal soft tissue anesthesia is necessary
- When lingual soft tissue anesthesia is needed
- When performing procedures on multiple mandibular teeth in one quadrant

62.1.4 Contraindications

- Infection or acute inflammation in the area of injection
- Patients who might bite their lip or tongue (e.g., very young child or physically or mentally handicapped adult or child)
- In relation to local anesthetic use

Absolute
- Local anesthetic allergy
  - Avoid all local anesthetics in the same chemical class (e.g., esters).
- Bisulfite allergy
  - Avoid vasoconstrictor-containing local anesthetics.

Relative
- Atypical plasma cholinesterase
- Methemoglobinemia (idiopathic or congenital)
- Significant liver dysfunction (American Society of Anesthesiologists [ASA] III–IV)
- Significant kidney dysfunction (ASA III–IV)
- Significant cardiovascular disease (ASA III–IV)
  - Avoid high concentrations of vasoconstrictors.
  - Use local anesthetics with epinephrine concentrations of 1:200,000 or 1:100,000 or 3 % mepivacaine or 4 % prilocaine.
- Clinical hyperthyroidism (ASA III–IV)
  - Avoid high concentrations of vasoconstrictors.
  - Use local anesthetics with epinephrine concentrations of 1:200,000 or 1:100,000 or 3 % mepivacaine or 4 % prilocaine.

62.1.5 Materials and Medications

- Local anesthetic carpule (1.7–1.8 mL)
  - Mepivacaine 3 % (+epinephrine 1:100,000)
  - Articaine HCI 4 % (+epinephrine 1:100,000 or 1:200,000)
  - Lidocaine HCI 2 % (+ epinephrine 1:50,000 or 1:100,000) (Fig. 62.2)
  - Bupivacaine HCI 0.5 % + epinephrine 1:200,000
- Aspirating syringe (Fig. 62.3)
- Needle (Fig. 62.4)
  - Gauge refers to the lumen of the needle: The smaller the number, the greater the diameter of the lumen.
  - Needles are color coded by gauge: red = 25 gauge, yellow = 27 gauge, and blue = 30 gauge.
  - *Recommendations: For inferior alveolar nerve (IAN) block, it is best to use a 25-gauge long needle.*
- Mouth props
- Retractors

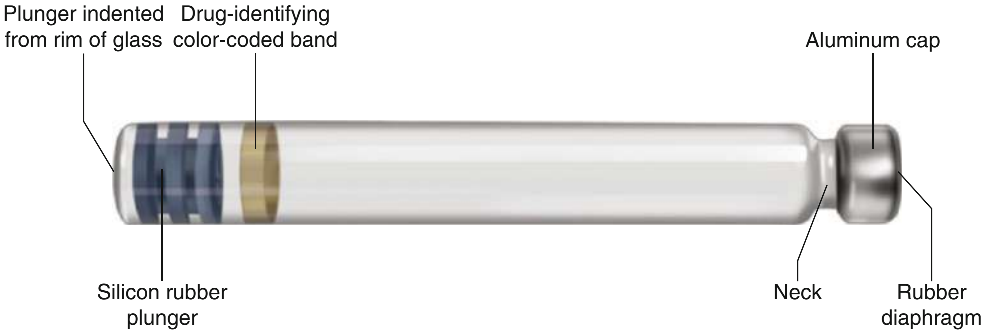
Plunger indented from rim of glass
Drug-identifying color-coded band
Aluminum cap
Silicon rubber plunger
Neck
Rubber diaphragm

Fig. 62.2 Local anesthetic carpule (1.7–1.8 mL)

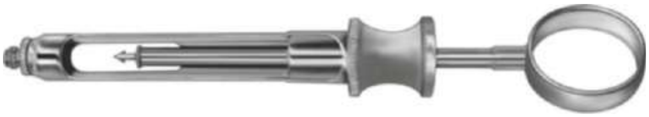

Fig. 62.3 Aspirating syringe

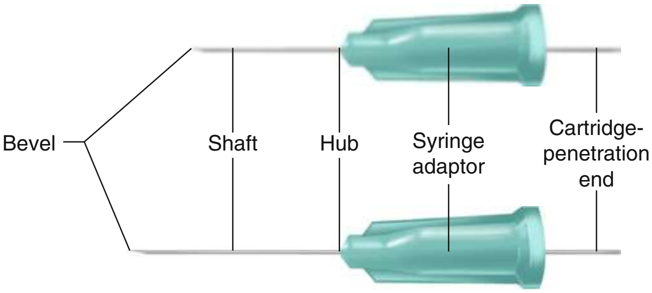
Bevel
Shaft
Hub
Syringe adaptor
Cartridge-penetration end

Fig. 62.4 Needle

62.1.6 Procedure

1. Target area: IAN as it passes downward toward the mandibular foramen.
2. Landmarks.
   (a) Coronoid notch
   (b) Pterygomandibular raphe
   (c) Occlusal plane of mandibular posterior teeth
3. Procedure.
   (a) Correct position for operator.
      (i) For a right IAN block, a right-handed administrator should sit at the 8 o'clock position facing the patient.
      (ii) For a left IAN block, a right-handed administrator should sit at the 10 o'clock position facing in the same direction as the patient.
   (b) Recommended to position the patient supine and with the mouth wide open.
   (c) Place thumb on the coronoid notch and index finger extraorally on the posterior border of the ramus in order to estimate the distance between these two points (Fig. 62.5).
      (i) The needle insertion should be three fourths of the anteroposterior distance from the coronoid notch to the deepest part of the pterygomandibular raphe.
   (d) Place the barrel of the syringe in the corner of the mouth on the contralateral side, usually corresponding to the premolars.
   (e) Slowly advance the needle until bony resistance is met.
      (i) For anxious or sensitive patients, a small volume of anesthetic may be deposited as the soft tissue is penetrated.
      (ii) Average depth of penetration to bony contact will be 20–25 mm, approximately two thirds to three fourths the length of a long needle.
      (iii) If the bone is contacted too soon (less than half the length of a long needle), the needle tip is usually located too far anteriorly (laterally) on the ramus. To correct:
         - Withdraw it slightly from the tissues and bring the syringe barrel anteriorly toward the lateral incisor or canine; reinsert to the proper depth.
      (iv) If the bone is not contacted, the needle tip is usually located too far posterior (medial). To correct:
         - Withdraw it slightly in tissue (leaving approximately one fourth its length in tissue), and reposition the syringe barrel more posteriorly (over the mandibular molars).
         - Continue the insertion until contact with the bone is made at an appropriate depth (20–25 mm).
   (f) Aspirate. If negative, slowly deposit 1.5 mL of anesthesia over 60 s.
   (g) Wait 3–5 min before commencing the dental procedure.
4. Precaution: *Do not deposit anesthesia if the bone is not contacted*. The needle tip may be resting within the parotid gland near the facial nerve (cranial nerve VII), and a transient paralysis of the facial never may occur if solution is deposited.

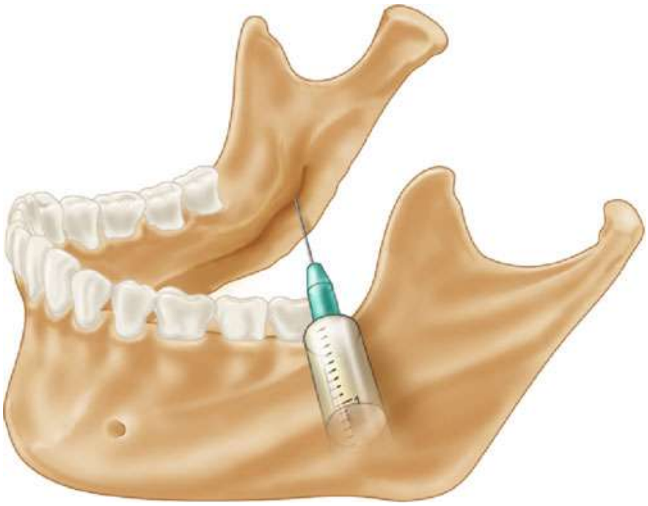

Fig. 62.5 Needle orientation for inferior alveolar nerve block

62.2 Buccal Nerve Block

62.2.1 Nerve Anesthetized

- Buccal nerve, a branch of the anterior division of the mandibular nerve

62.2.2 Area Anesthetized (Fig. 62.6)

- Soft tissues and periosteum buccal to the mandibular molars

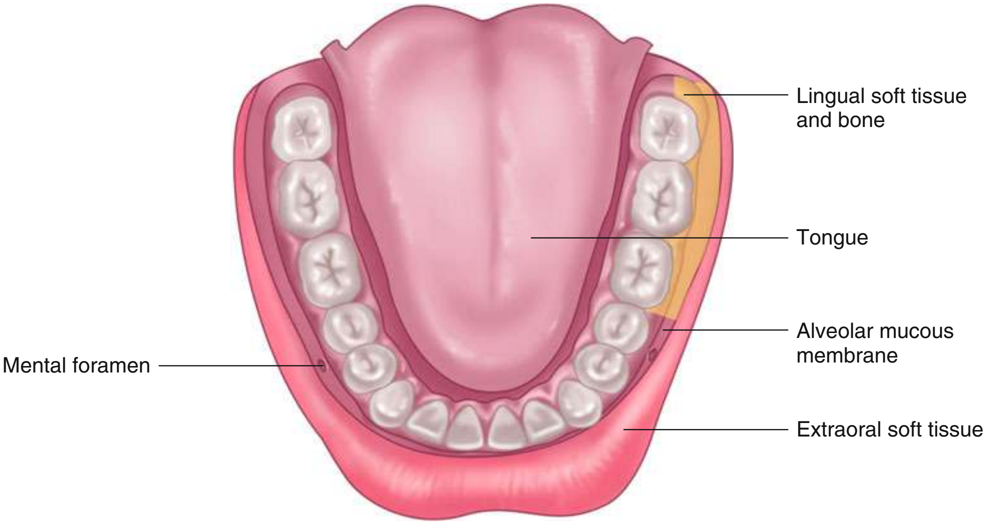
Lingual soft tissue
and bone
Tongue
Alveolar mucous
membrane
Extraoral soft tissue
Mental foramen

Fig. 62.6 Areas anesthetized with buccal nerve block

62.2.3 Procedure

1. A 25-gauge long needle is recommended.
2. Landmarks: mandibular molars and mucobuccal fold.
3. Orient the bevel of the needle *toward* the bone during injection.
4. Correct positioning.
   (a) For a right buccal nerve block, a right-handed administrator should sit at the 8 o’clock position directly facing the patient.
   (b) For a left buccal nerve block, a right-handed administrator should sit at the 10 o’clock facing in the same direction as the patient.
5. Procedure.
   (a) With the index finger, pull the buccal soft tissues in the area of injection laterally to allow for better visualization.
   (b) Align the syringe parallel to the occlusal plane of the teeth.
   (c) Penetrate the mucous membrane at the injection site, distal and buccal to the last molar (Fig. 62.7).
6. If tissue at the injection site becomes swollen, stop depositing solution.

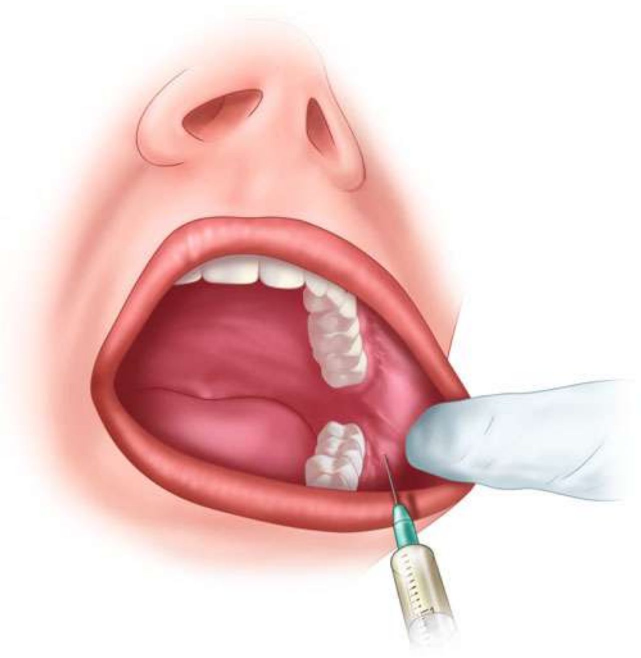

Fig. 62.7 Needle orientation for buccal nerve block

62.3 Mental Nerve Block

62.3.1 Nerve Anesthetized

- Mental nerve, a terminal branch of the inferior alveolar nerve

62.3.2 Area Anesthetized (Fig. 62.8)

- Buccal mucous membranes anterior to the foramen (around the second premolar) to the midline and skin of the lower lip

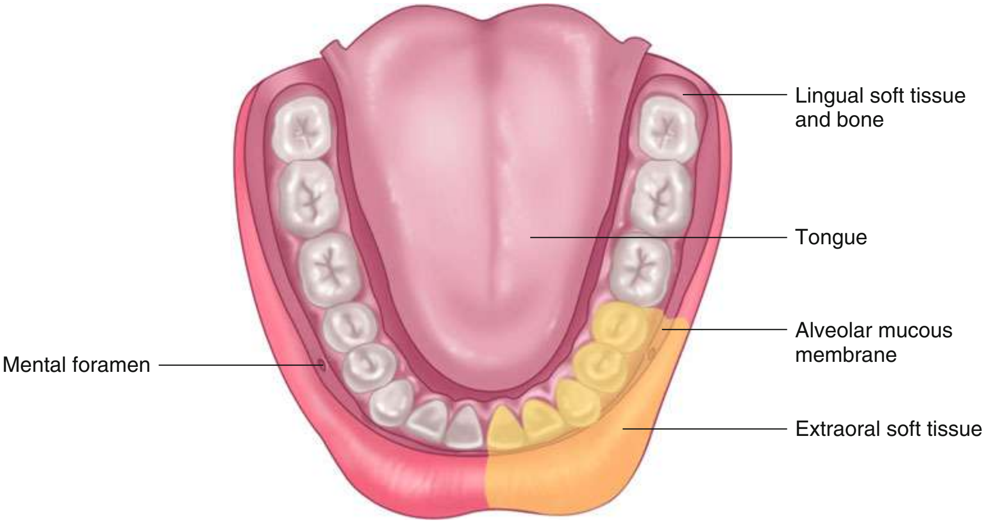
Lingual soft tissue and bone
Tongue
Alveolar mucous membrane
Extraoral soft tissue
Mental foramen

Fig. 62.8 Areas anesthetized with mental nerve block

62.3.3 Indications

- For buccal soft tissue anesthesia in procedures such as:
  - Soft tissue biopsies
  - Suturing of soft tissues

62.3.4 Procedure

1. Area of insertion: mucobuccal fold at or just anterior to the mental foramen.
2. Orientation of bevel should be *toward* the bone during injection.
3. Operator should sit in front of the patient so that the syringe is below the patient's line of sight.
4. Locate the mental foramen.
   (a) Place the index finger in the mucobuccal fold and press against the body of the mandible in the area of the first molar.
   (b) Move the finger anteriorly until the bone beneath the finger feels somewhat concave.
   (c) Mental foramen is usually found around the apex of the second premolar.
   (d) Orient the needle with the bevel directed toward the bone.
   (e) Penetrate the mucous membrane and advance needle slowly; penetration depth is usually 5–6 mm.
   (f) If aspiration is negative, deposit approximately one third of the cartridge over 20 s.
      (i) If the site balloons, stop the deposition of anesthetic and remove the syringe.

Selected Reading

Bennett CR. Monheim's local anesthesia and pain control in dental practice. 6th ed. St. Louis: Mosby; 1978.

Gow-Gates GAE. Mandibular conduction anesthesia: a new technique using extraoral landmarks. Oral Surg. 1973;36:321–8.

Jastak JT, Yagiela JA, Donaldson D. Local anesthesia of the oral cavity. Philadelphia: WB Saunders; 1995.

Malamed SF. The Gow-Gates mandibular block: evaluation after 4275 cases. Oral Surg. 1981;51:463.

Malamed SF. Handbook of local anesthesia. 5th ed. St. Louis: Mosby; 2004.

Reduction of Dislocated Temporomandibular Joint

63

Christopher J. Spencer and Geraldine Weinstein

63.1 Indications

- Open lock: associated with yawning, vomiting, or opening the mouth wide
- Open lock: associated with a dental procedure
- Open lock: associated with endoscopy
- Open lock: associated with oral intubation
- Time duration: acute to 3 weeks or less duration

63.2 Contraindications for Closed Reduction

- Absolute
  - Head trauma with fracture of the skull, maxilla, mandible, or mandibular Condyles
- Relative
  - Dislocation of 30 days or longer (will likely be unable to accomplish reduction without general anesthesia and/or open surgical approach)

63.3 Materials and Medications

- Local anesthetic syringe.
- Lidocaine 2 % 1–2 mL.
- 25- to 27-gauge needle (long or approximately 2 inches long).
- Betadine (povidone-iodine) or other skin antiseptic preparation.
- Gauze padding for thumbs.
- Consider a muscle relaxant.
- Consider conscious sedation.

C.J. Spencer, DDS • G. Weinstein, DDS (✉)
Department of Restorative Dental Sciences, University of Florida
College of Dentistry, Gainesville, FL, USA
e-mail: cspencer@dental.ufl.edu; gweinstein@dental.ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_63

63.4 Procedure

63.4.1 Manual Closed Reduction *without* Local Anesthesia (Figs. 63.1)

1. Position the patient in an upright posture with the mandible at the physician's flexed elbow height (physician's comfortable position).
2. Place the thumbs on the mandibular molars with wrapping around the thumbs to protect from possible biting force once the mandible reduces.
3. Apply bilateral firm force in an inferior direction.
4. The mandible will move rapidly in an inferior and then a posterior direction as the condyles slide back over the height of their respective articular eminences.

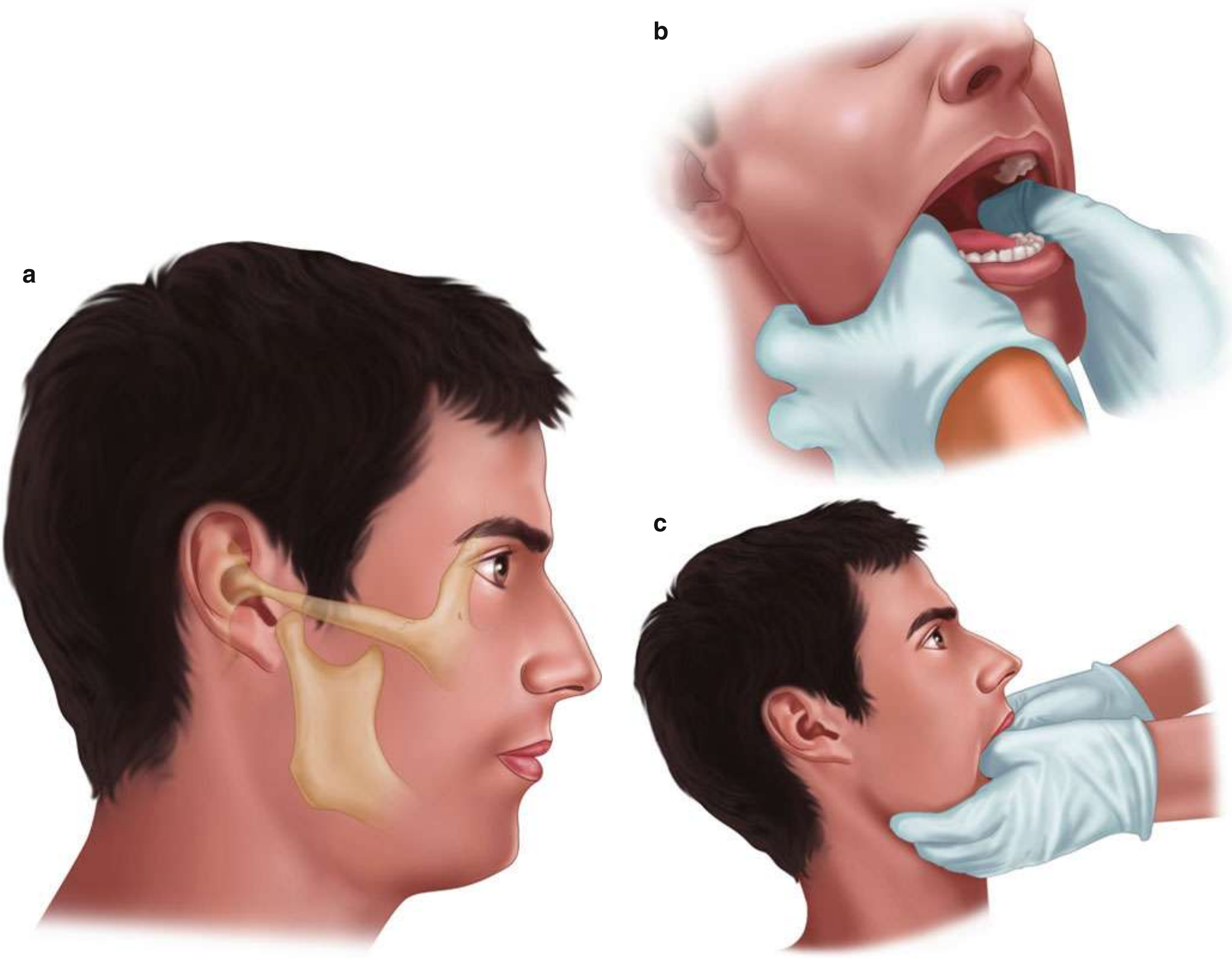

Fig. 63.1 Reduction in progress: (**a**) In a mandibular dislocation, the condyle will be anterior and superior to the articular eminence. (**b**) Position the thumbs on mandibular molars, and apply firm pressure in an inferior direction to distract the TM joint condyles so that they can reposition themselves into the glenoid fossa so that the TM joint can be reduced. (**c**) Lateral view of the distraction force with direction in an inferior direction to distract the condyle of the TM joint

63.4.2 Manual Closed Reduction *with* Local Anesthesia

- If the mandible will not respond to closed reduction with just thumb pressure, likely the masticatory muscles are contracting with sufficient force to prohibit the condyles from being sufficiently distracted owing to pain.
  1. Reduction of pain in the temporomandibular joint (TMJ) with local anesthesia.
  2. *Auriculotemporal block of V3:*
     - The auriculotemporal nerve that innervates the TMJ may be anesthetized inferior to the TMJ capsule. It can be accessed through the skin just anterior to the tragus.
     - With the patient's mouth wide open (it already is in this case), a triangular-shaped hollow will be evident inferior and posterior to the mandibular condyle. Insert the needle at a 20-degree anterior inclination, in the horizontal plane, at the level of the inferior border of the tragus of the ear (Fig. 63.2). The bevel of the needle should be anterior.
     - The needle should be inserted behind the (posterior) ramus and approximately 2 cm deep (aiming for the medial aspect of the posterior border of the ramus). If the posterior border of the ramus is contacted, the needle will need to be directed in a more posterior direction. Then deposit 1–2 mL of lidocaine 2 %.
  3. Then, as before, place the thumbs in a bilateral position on the patient's mandibular molars, and depress the mandible to distract the condyles in an inferior direction.
  4. Conscious sedation may be utilized if the reduction procedure has been arduous and stressful for the patient.

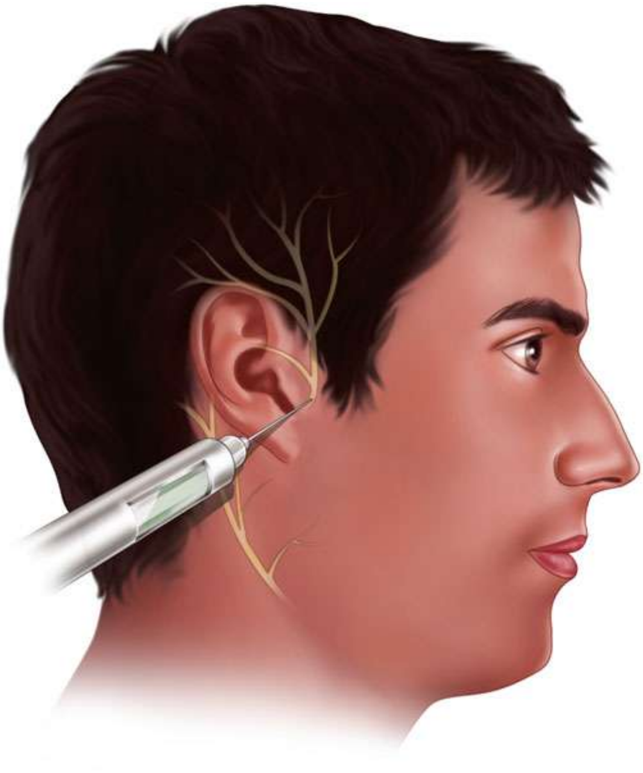

Fig. 63.2 Infiltration of cranial nerve V, the auriculotemporal branch (V3)

63.5 Complications

- Inability to reduce the condyles manually which may lead to more invasive procedures.
- If the condition is acute (≤24 h) and not associated with trauma, there are few if any significant complications or risks for this procedure.

63.6 Pearls

- The manual pressure required on the mandibular molars needs to be sustained and very firm.
- If both condyles are dislocated, it is likely beneficial to attempt one side at a time.
- The clinician needs to protect the thumbs from the impact of the patient's molars during the sudden successful reduction because the biting forces are significant in the molar region.

Selected Reading

Chan TC, Harrigan RA, Ufberg J, Vilke GM. Mandibular reduction. J Emerg Med. 2008;34:435.

Donlon WC, Truta MP, Eversole LR. A modified auriculotemporal nerve block for regional anesthesia of the temporomandibular joint. J Oral Maxillofac Surg. 1984;42:544.

Huang IY, Chen CM, Kao YH, Chen CM, Wu CW. Management of long-standing mandibular dislocation. Int J Oral Maxillofac Surg. 2011;40:810–4.

Prabhakar V, Singla S. Bilateral antersuperior dislocation of the intact mandibular condyles in the temporal fossa. Int J Oral Maxillofac Surg. 2011;40:640–3.

Thagarajah T, Mcculloch N, Thangarajah S, Stocker J. Bilateral temporomandibular joint dislocation in a 29-year-old man: a case report. J Med Case Rep. 2010;4:263.

Dry Socket (Alveolar Osteitis, Fibrinolytic Osteitis)

64

Michael A. Abraham, Amir Azari, Jennifer Westcott,
and Franci Stavropoulos

64.1 Indications (Fig. 64.1)

- Definition: severe pain occurring 2–3 days after tooth extraction
- Recent tooth extraction, especially of a mandibular tooth or an impacted third molar
- Partially or completely visible bone socket
- Intense radiating pain (often to the ear)
- Fetid odor without suppuration
- Absence of swelling, lymphadenitis, or bacteremia
- Foreign bodies present in the extraction socket

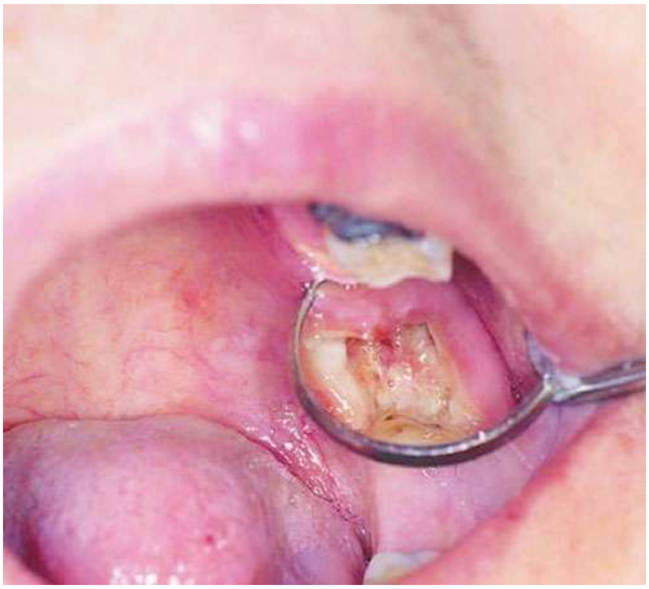

Fig. 64.1 Clinical photograph suggestive of a dry socket. Clinical correlation is necessary

---

M.A. Abraham, DMD (✉)
United States Air Force, Dental Corps, Minot, ND, USA
e-mail: mikeabrahamdmd@gmail.com

A. Azari, DMD
Department of Oral and Maxillofacial Surgery, Oregon Health and Science University, Portland, OR, USA
e-mail: Azari@OHSU.edu

J. Westcott, DMD
Private Practice, Palm Beach Gardens, FL, USA
e-mail: Jwestcott561@gmail.com

F. Stavropoulos, DDS
Department of Dental Specialties – Oral and Maxillofacial Surgery, Gundersen Health System, La Crosse, WI, USA
e-mail: Mfstavro@gundersenhealth.org

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_64

64.2 Contraindications

- Absolute
  - Osteomyelitis
  - Jaw fracture
- Relative
  - Retained roots
  - Infection

64.3 Materials and Medications

- Warm saline or 0.12 % aqueous chlorhexidine solution
- 12-mL syringe with curved tip (Monoject® syringe)
- 25-gauge needle, syringe
- Local anesthetic, 2 % lidocaine 1:100,000 epinephrine
- Suction with small tip or gauze
- Socket dressing
  - Gelfoam or small gauze strips
- Socket medicament
  - Sultan dry socket paste® (guaiacol, balsam of Peru, eugenol, 1.6 % chlorobutanol), iodoform, or eugenol® (zinc oxide eugenol dental cement)
- Curved forceps

64.4 Procedure

1. Administer local anesthesia as necessary.
2. Remove any sutures closing the extraction site.
3. Irrigate the wound gently with warm saline or 0.12 % aqueous chlorhexidine.
4. Carefully suction or gently dry any excess saline; the socket area should be isolated from saliva by using gauze or cotton rolls.
5. Gently place iodoform-soaked gauze, Gelfoam soaked in eugenol, or Sultan dry socket paste in the extraction socket with forceps/Monoject syringe.
6. Rinse with saline and replace the dressing as needed for the first 2–3 days and every 2–3 days thereafter.
7. Remove the dressing, if it does not dissolve, without replacement once the pain has resolved.
8. Pain medication (nonsteroidal anti-inflammatory drugs [NSAIDS] or narcotics) should be prescribed if necessary.
9. Follow-up with dentist.

64.5 Complications

- Delayed healing
- Wound dehiscence

64.6 Pearls and Pitfalls

- *Pearls*
  - Wound irrigation may be so painful at the first visit that administration of a local anesthetic without a vasoconstrictor should be considered.
  - The patient should experience profound pain relief within minutes of placement of the soaked medicated dressing.
  - If a medicated dressing is necessary for more than 2 weeks, reevaluate for development of osteomyelitis.
  - "Dry socket" is not a progressive disease but may persist for 10–14 days whether treated or not; therapy is palliative.
  - Instruct the patient to avoid the following, which can cause changes of pressure in the mouth:
    - Smoking
    - Using a straw
    - Spitting
    - Drinking carbonated beverages (e.g., soda, seltzer water, beer)
- *Pitfalls*
  - Avoid over manipulating the socket because this will increase the amount of exposed bone and pain.

Selected Reading

Bloomquist D, Hooley J, Whitacre R. A self-instructional guide: surgical complications. 3rd ed. Seattle: Stroma; 1983.

Matocha DL. Postsurgical complications. Emerg Med Clin North Am. 2000;18:549–64.

Roberts G, Scully C, Shotts R. Dental emergencies. West J Med. 2001;175:51.

Postextraction Hemorrhage

65

Michael A. Abraham, Amir Azari, Jennifer Westcott,
and Franci Stavropoulos

65.1 Indications

- Recent tooth extraction site, presenting with more than a slight oozing of blood
- Full evaluation indicating amount of blood loss, present physical condition, and reason for hemorrhage including coagulopathy or medication use

65.2 Contraindications

- Absolute
  - None
- Relative
  - None

65.3 Materials and Medications

- 2×2 gauze pad
- Saline
- 25-gauge needle, syringe
- Local anesthetic without vasoconstrictor—2 % lidocaine plain
- Gelfoam® (absorbable gelatin-compressed sponge) or oxidized cellulose
- Topical thrombin
- Suture kit with 3-0 chromic gut suture or 3-0 Vicryl® suture (synthetic absorbable sterile surgical suture composed of a copolymer made from 90 % glycolide and 10 % L-lactide)
- Hemostat

65.4 Procedure

1. Use suction and saline irrigation to gently rinse the affected area. If a "liver" clot is present, irrigate and remove it with suction.
2. Determine the source of hemorrhage without local anesthesia, if possible, because the use of local anesthetic with an added vasoconstrictor may obscure bleeding sites.
3. Moisten a folded 2×2 gauze pad with saline and place it directly onto the extraction site.
4. Instruct the patient to apply firm biting pressure, and observe for 1 h, changing gauze as necessary.
5. If bleeding persists, an intraoral nerve block should be performed.
   - Blocks are preferred to infiltrations; anesthetic with epinephrine infiltrated near the bleeding site will produce only temporary local hemostasis from vasoconstriction.
6. Gently curette the tooth extraction socket and remove areas of old blood clot or granulation tissue.
7. Check soft tissue for associated arterial bleeding.
   - If hemorrhage is localized to soft tissue, use pressure or tie off vessels.
8. Fold Gelfoam® into a small cylinder to fit into the extraction socket.
9. Place Gelfoam® with topical thrombin or Surgicel® (absorbable hemostat) into the socket and hold in

M.A. Abraham, DMD (✉)
United States Air Force, Dental Corps, Minot, ND, USA
e-mail: mikeabrahamdmd@gmail.com

A. Azari, DMD
Department of Oral and Maxillofacial Surgery, Oregon Health and Science University, Portland, OR, USA
e-mail: Azari@OHSU.edu

J. Westcott, DMD
Private Practice, Palm Beach Gardens, FL, USA
e-mail: Jwestcott561@gmail.com

F. Stavropoulos, DDS
Department of Dental Specialties – Oral and Maxillofacial Surgery, Gundersen Health System, La Crosse, WI, USA
e-mail: Mfstavro@gundersenhealth.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_65

position with a figure-of-eight stitch using 3-0 chromic gut suture or 3-0 Vicryl suture (Figs. 65.1 and 65.2).

10. Fold 2×2 gauze, moisten it with saline, and place it over the suture.
11. Instruct the patient to bite down with firm pressure for 30 min; repeat as necessary.
12. Follow-up with dentist.

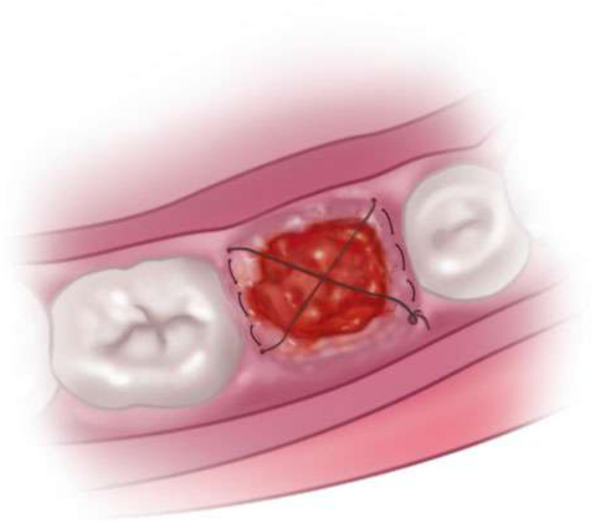

Fig. 65.1 A resorbable figure-of-eight suture placed over an extraction socket

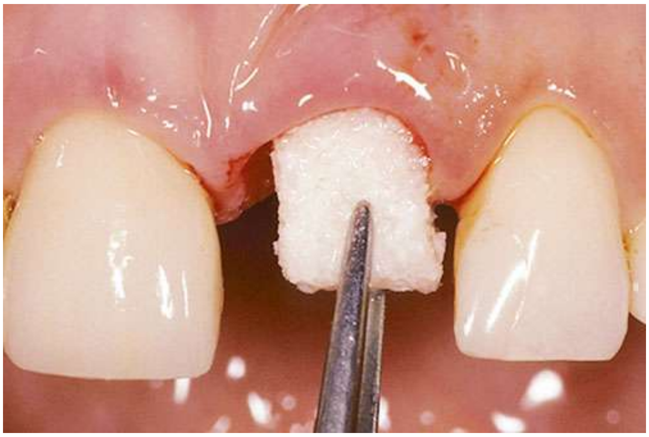

Fig. 65.2 Gelfoam® (absorbable gelatin-compressed sponge) being placed into the extraction socket (Photograph courtesy of Dr. Michael Abraham)

65.5 Complications

- Continued bleeding due to lack of patient compliance.
- Continued bleeding due to coagulopathy or medication use.
- If bleeding persists and coagulopathy is identified, the administration of intravenous blood replacement products may be necessary.

65.6 Pearls and Pitfalls

- Pearls
  - Minor bleeding concerns may be addressed at home by instructing the patient to bite on the affected area with a tea bag for 30 min (tannic acid in tea is a vasoconstrictor).
  - It is normal for an extraction socket to ooze slight amounts of blood for 12–24 h; it is normal for patients to see some blood on their pillow after waking.
  - The patient should be instructed to avoid the following, which can cause changes of pressure in the mouth:
    - Smoking
    - Using a straw
    - Spitting
    - Drinking carbonated beverages (e.g., soda, seltzer water, beer)
- Pitfalls
  - Small amounts of blood mixed with saliva may deceptively appear as large amounts of blood.

Selected Reading

Bloomquist D, Hooley J, Whitacre R. A self-instructional guide: surgical complications. Seattle: Stroma; 1983. pp. 50–5.

Hupp JR, Ellis III E, Tucker MR, editors. Contemporary oral and maxillofacial surgery. 5th ed. St. Louis: Mosby Elsevier; 2008. p. 195–7.

Fractured Tooth

66

Geraldine Weinstein

66.1 Indications (Fig. 66.1)

- Temporary repair of an acute dental fracture until follow-up by a dentist can be secured.

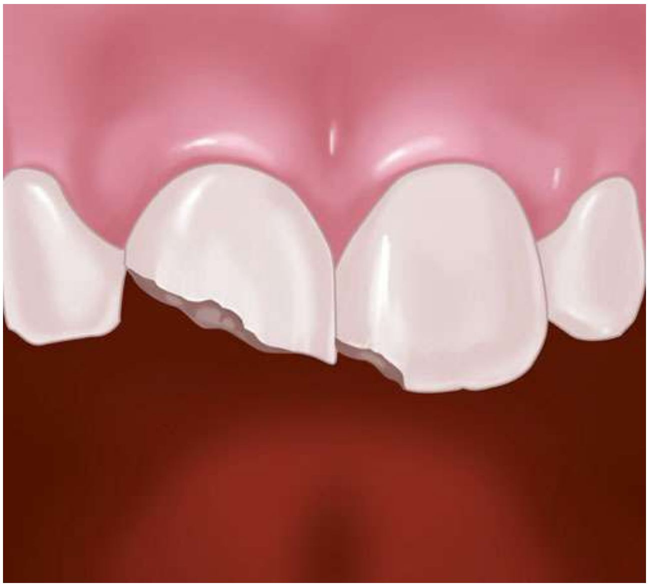

Fig. 66.1 Examples of fractured teeth

66.2 Methods of Sustaining Dental Fracture

- Traumatic injury to the head/facial area
- Falling down
- Extensive tooth decay that has undermined the integrity of the tooth structure
- Biting down on something hard

G. Weinstein, DDS
Department of Restorative Dental Sciences,
University of Florida College of Dentistry,
Gainesville, FL, USA
e-mail: gweinstein@dental.ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_66

66.3 Four Types of Fractured Tooth (Fig. 66.2)

- *Type 1*: contained to the enamel of the tooth, asymptomatic, and can be treated easily with a composite filling by a dentist.
- *Type 2*: involves a fracture through the dentin layer of the tooth. The patient may experience some sensitivity to temperature changes and chewing. Depending on the severity, treatment may include a root canal and a restoration by a dentist.
- *Type 3*: involves the pulp of the tooth and will require endodontic treatment by a dentist.
- *Type 4*: a root fracture in the tooth that makes it nonrestorable and requiring extraction. It is diagnosed by means of a periapical radiograph taken in a dental office.

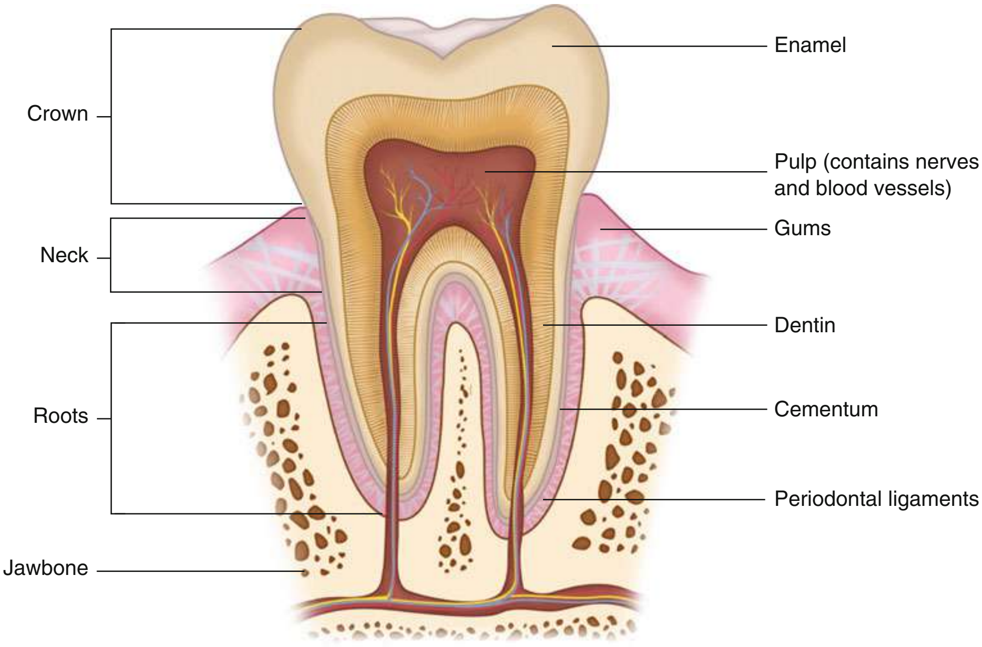
Crown
Neck
Roots
Jawbone
Enamel
Pulp (contains nerves and blood vessels)
Gums
Dentin
Cementum
Periodontal ligaments

Fig. 66.2 Anatomy of a tooth

66.4 Contraindications

- *Relative*
  - Patients at high risk of aspiration owing to intoxication and altered mental status

66.5 Materials and Medications

- Warm saline or 0.12 % aqueous chlorhexidine solution for irrigation of tissues and tooth
- Irrigating syringe
- Suction with a small tip
- Gauze to control hemorrhage
- Resorbable sutures and local anesthetic as needed for soft tissue lacerations
- Temporary tooth restoration material like intermediate restorative material (IRM) and glass ionomer (like Fuji)

66.6 Procedure

1. Have the patient rinse the mouth out with warm water to clean out any debris. Apply an ice pack to the affected cheek to reduce swelling.
2. Anesthetize the area, either locally at the tooth apex or with a nerve block (mental or inferior alveolar).
3. Irrigate the area; assess the fracture intraorally and check for soft tissue lacerations. Use the gauze with finger pressure to control bleeding in the soft tissue or the tooth.
4. Check if the tooth or bony segment is mobile. If so, a referral to a dentist or oral surgeon is necessary as soon as possible for proper assessment.
5. Type II fractures (fracture limited to dentin layer): cover the exposed surface with a temporary dental cement. In a pinch, 2-octyl cyanoacrylate (Dermabond) is an acceptable secondary alternative (Fig. 66.3).
6. Type III (pulp involved)
   - Provide immediate dental follow-up and analgesics.
   - Initiate antibiotic coverage with penicillin or clindamycin.

50% MORE
Than Other Brands
12+
REPAIRS
net wt. 2.64 g (0.09 oz)
Repair
Lost Fillings & Loose Caps
DenTek
Temparin Max
One Step
Formula
MAXIMUM
STRENGTH
Dentist
Recommended
Instant Pain Relief
Easy To Apply
NEW!
Applicator
Tip
Spade Tool
DenTek

Fig. 66.3 Example of temporary dental cement (Reproduced with permission from DenTek Oral Care, Inc.)

66.7 Complications

- Loss of a tooth
- Infection or abscess
- Aspiration of a segment or a whole tooth
- Cosmetic deformity

66.8 Pearls

- Be certain to perform a thorough intraoral examination, looking for tooth fragments or lacerations that may be hiding fragments.
- Dental blocks are very useful for pain control.
- If a tooth is not mobile and the *pulp is exposed*, immediate referral (within a few hours) to a dentist is necessary for extraction or endodontic (root canal) treatment of the tooth. Placement of a temporary-type restoration on this tooth is *not* recommended at this time because it may exacerbate symptoms. Prescribe pain medication and possibly antibiotics when the tooth's pulp is exposed and the patient is unable to see the dentist within 24 h.
- If the tooth is not mobile and *the pulp is not exposed*, a temporary restoration can be placed on the tooth and the patient referred to a dentist for treatment. The fractured part of the tooth should be saved in the event that it can be used. If temporary tooth restoration is unavailable in the emergency department, advise the patient that it is readily available at local pharmacies.
- ALL DENTAL FRACTURES, EXCEPT TYPE I, REQUIRE DENTAL FOLLOW-UP WITHIN 24 h.

Dental Avulsion Management

67

Laura Tucker and Abimbola O. Adewumi

67.1 Indications

- The tooth is completely displaced *out* of its socket, leading to severance of the neurovascular pulp supply and separation of the periodontal ligament (Fig. 67.1).
- Diagnosis
  - Clinically, the socket is found empty or filled with coagulum.
  - Imaging (occlusal, periapical, and lateral views of the affected tooth and surrounding area) (Fig. 67.2):
- Confirm vacuous socket.
- Ensure that the missing tooth is not intruded.
- Diagnose root fracture or alveolar fracture.

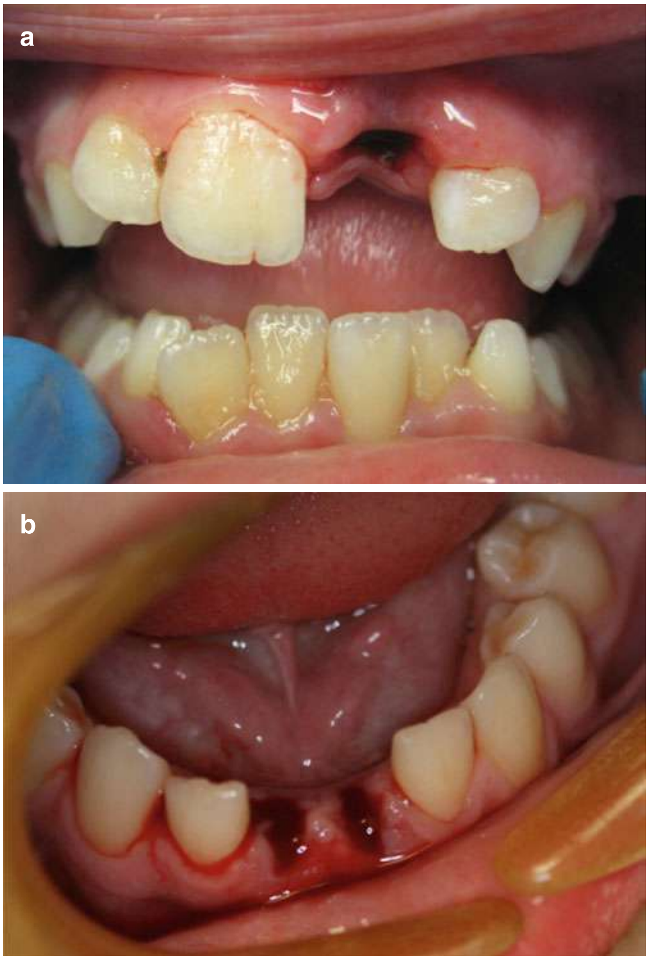

Fig. 67.1 (**a**, **b**) Empty socket following traumatic dental avulsion

---

L. Tucker, DDS • A.O. Adewumi, BDS, FDSR (✉)
Department of Pediatric Dentistry, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: lauragracesullivan@gmail.com; aadewumi@dental.ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_67

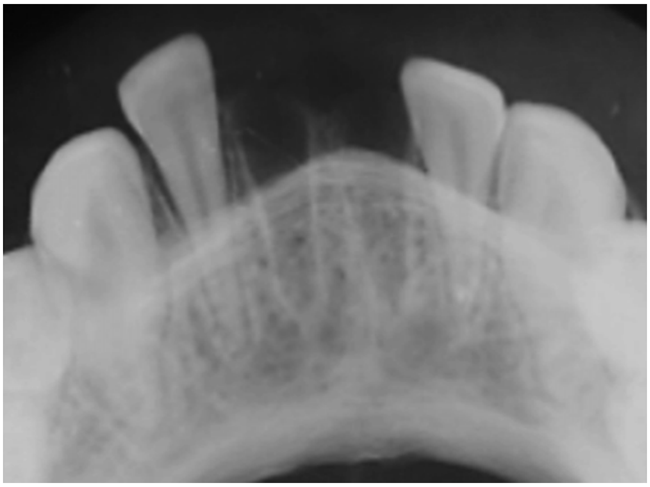

Fig. 67.2 Lower occlusal radiograph shows a complete avulsion of the mandibular right and left central incisors

67.2 Contraindications

- Absolute
  - Avulsed tooth is a primary tooth.
    - DO NOT REPLANT OR REPOSITION AVULSED PRIMARY TEETH.
    - Replantation of primary teeth increases the potential for damage to the developing permanent tooth owing to the increased frequency of pulpal necrosis.
- Relative
  - Fractured root (further intervention required before replantation)
  - Alveolar fracture (further intervention required before replantation)
  - Prolonged extraoral dry time and out of compatible solution (>1 h)
  - Immunocompromised host or congenital heart disease
  - Severe seizure disorder that may put tooth at risk for coming out while the airway is compromised
  - Patient with potential to lose airway reflexes

67.3 Materials and Medications

- Hank's Balanced Salt (Save-A-Tooth) solution or milk in which to preserve tooth until it can be replanted
  - Normal saline if neither of those is available
- Normal saline rinse
- 20- or 18-gauge cannula for gentle irrigation of the socket
- Absorbable suture for gingival lacerations, if present
- Flexible splint materials
  - Round dental wire
  - A flat pliable metal long enough to cover the affected tooth and the two teeth on other side (e.g., the metal nasal bridge from a respirator mask)
- Fixative
  - Dental adhesive
  - Dermabond or some other brand of cyanoacrylate for adhesive skin closures

67.4 Procedure

1. Be certain the tooth is a permanent one and not a primary tooth.
2. If not done by the patient, gently wash the tooth under water for approximately 10 s.
    - Be certain to hold the tooth by the crown, not the root (Fig. 67.3)
3. If the tooth cannot be replanted immediately, place the tooth in Hank's solution or milk.
4. If no such media is available, instruct the patient to hold the tooth inside his or her mouth between the cheek and the gums.
5. Gently replant the tooth, using digital pressure only into as anatomical a position as possible (Fig. 67.4).
    - Assess clinically for alignment.
    - Radiograph for confirmation.
6. Suture any gingival lacerations if present.
7. Apply a flexible splint, securing the affected tooth to the teeth on either side.
    - Consider using skin adhesive both to secure the tooth to its neighbors and, perhaps, to apply a makeshift splint until the patient can be seen by her or his dentist.
        - Towel dry the teeth as best as possible.
        - Apply skin adhesive (using the standard applicator) to the lateral edges of the avulsed tooth where it will make contact with its adjacent teeth.
        - If dental wire is available, apply adhesive to the buccal surfaces of the three teeth (the avulsed tooth central to the other two) and apply length of metal to the Fixodent (Fig. 67.5).
8. Systemic antibiotics with anaerobic coverage is empirical.
9. Ascertain tetanus status for the patient; update if uncertain.

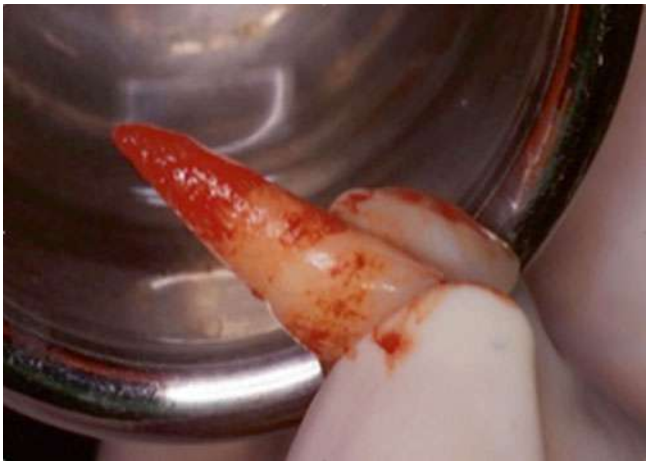

Fig. 67.3 Proper way to hold an avulsed tooth

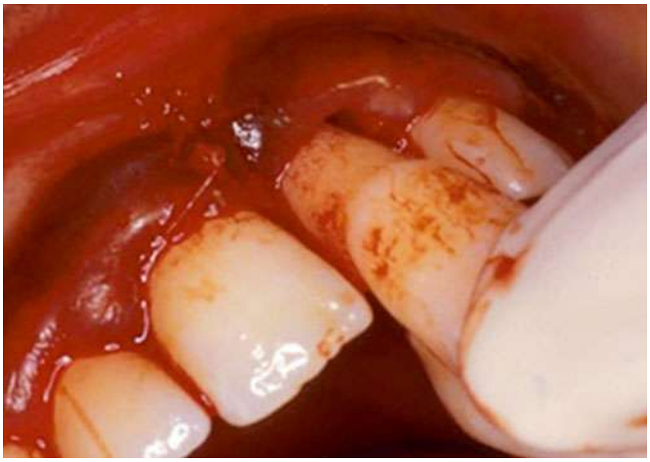

Fig. 67.4 Gentle replantation using digital pressure

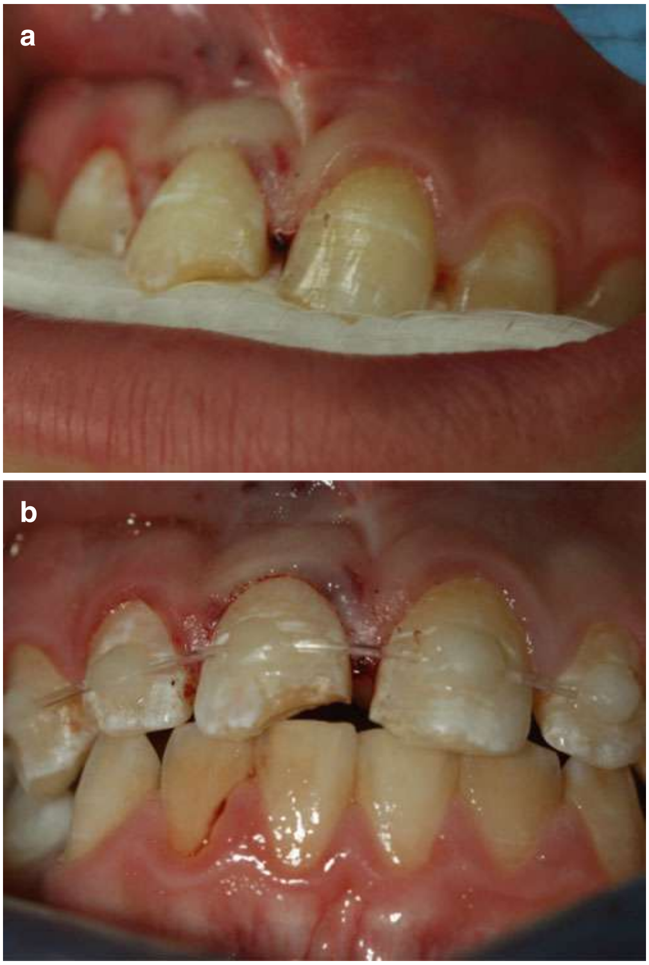

Fig. 67.5 (**a**, **b**) Splint stabilization

67.5 Complications

- Primary teeth
  - Dilaceration (bend) in the permanent tooth crown
  - Enamel defect of the lower permanent incisors as a result of avulsion of the preceding primary tooth
- Permanent teeth
  - Discoloration as a result of loss of vitality of the avulsed tooth.
  - Ankylosis of the alveolar ridge, leading to functional and aesthetic changes.
  - Replacement resorption occurs when the replanted tooth is slowly replaced with bone.
  - External inflammatory resorption is a progressive loss of tooth associated with destruction of adjacent alveolar bone.
  - Infection
  - Aspiration of an inadequately secured replanted tooth.

67.6 Pearls and Pitfalls

- Pearls
  - At the initial examination, make sure that all avulsed teeth are accounted for.
    - If not, a radiographic examination is necessary to ensure that the missing tooth is not completely intruded (pushed into the gum) or has sustained a root fracture with loss of the coronal fragment.
  - In children, always consider the likelihood of nonaccidental trauma (abuse).
  - Short-term and long-term dental follow-up cannot be emphasized enough.
- Pitfalls
  - If the avulsed tooth cannot be accounted for, aspiration is a possibility.
- Prognosis
  - Depends on extraoral dry time (length of time the tooth has been out of the mouth and not stored in an appropriate medium):
    - Ideally, tooth should be implanted within 5 min.
    - Extraoral dry time greater than 60 min has a poor prognosis for periodontal healing.
  - Depends on stage of root development of the avulsed tooth (Fig. 67.6)
- The more advanced the root development, the lower the probability of pulp healing and survival.

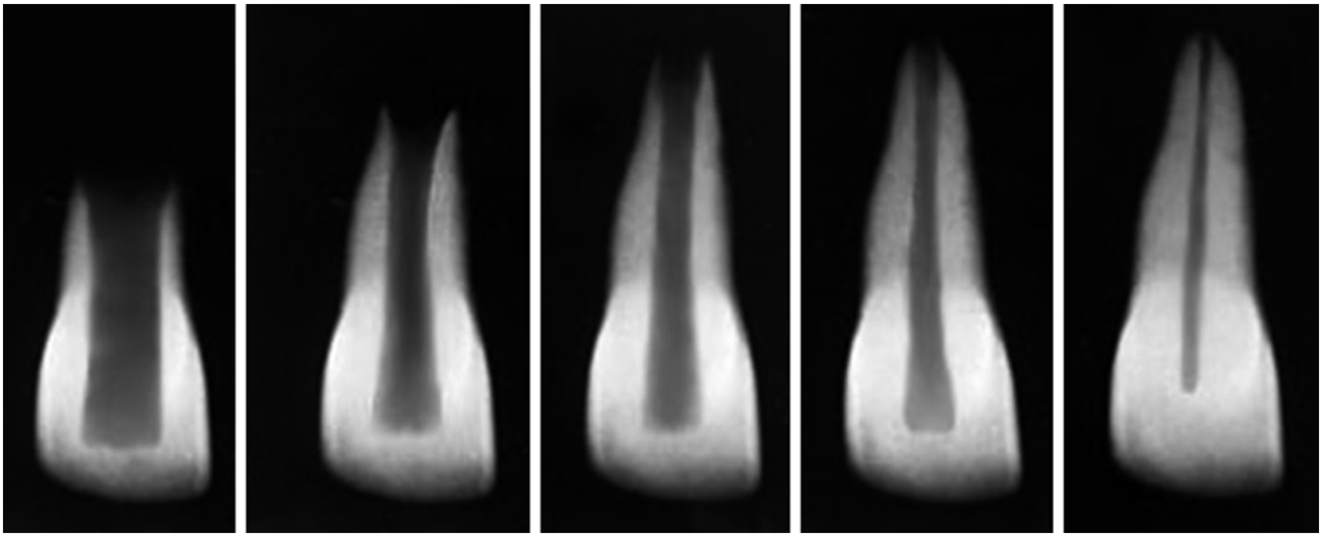

Fig. 67.6 The various stages of root development, from the less advanced (*open apex, left*) to the more advanced (*closed apex, right*)

Selected Reading

AAPD Council on Clinical Affairs. Guideline on management of acute dental trauma. AAPD reference manual. Chicago: American Academy of Pediatric Dentistry; 2010–2011. pp. 202–12.

Andreasen FM, Andreasen JO. Avulsions. In: Andreasen JO, Andreasen FM, Andersson L, editors. Textbook and color atlas of traumatic injuries to the teeth. 4th ed. Oxford: Blackwell; 2007. p. 444–88.

Andreasen JO, Jensen SS, Sae-Lim V. The role of antibiotics in preventing healing complications after traumatic dental injuries: a literature review. Endod Topic. 2006;14:80–92.

Finucane D, Kinirons MJ. External inflammatory and replacement resorption of luxated, and avulsed replanted permanent incisors: a review and case presentation. Dent Traumatol. 2003;19: 170–4.

Flores MT, Andersson L, Andreasen JO, et al. Guidelines for the management of traumatic dental injuries. II. Avulsion of permanent teeth. Dent Traumatol. 2007;23:130–6.

Hile LM, Linklater DR. Use of 2-octyl cyanoacrylate for the repair of a fractured tooth. Ann Emerg Med. 2006;47:424–6.

The dental trauma guide. Available at: www.dentaltraumaguide.org.

Part X

Gastrointestinal Procedures

Excision of Thrombosed External Hemorrhoid

68

Latha Ganti

68.1 Indications

- Acute pain within 72 h of thrombosis onset
- The thrombosis will be visible as a bluish-purplish painful mass in perianal area (Fig. 68.1).

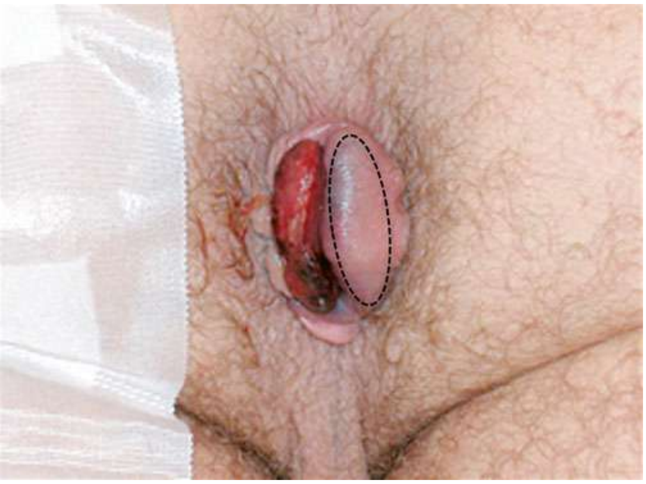

Fig. 68.1 Bluish-purplish appearance of an external thrombosed hemorrhoid. The *ellipse* denotes the area of the elliptical incision to be made (Reproduced with permission from: Fargo and Latimer [3])

68.2 Contraindications

- Absolute
  - Severe coagulopathy
  - Hemodynamic instability
  - Concurrent internal hemorrhoid with rectal prolapse
  - Painless rectal mass (external hemorrhoids are almost always painful, so a painless mass is not a thrombosed external hemorrhoid; also, the whole point of doing this procedure is to provide pain relief).
- Relative
  - Allergy to local anesthetics
  - Perianal infection
  - Inflammatory bowel disease
  - Serious systemic illness or comorbidity that could increase risk of procedure

68.3 Materials and Medications

- Sterile gloves and drape
- Alcohol swabs or pads
- 10 % povidone-iodine or chlorhexidine prep
- 2 % lidocaine with epinephrine
- 5 ml syringe with 25 or 27 gauge and 18 gauge needles
- #11 or #15 scalpel blade and handle
- Direct light source
- Forceps
- Iris scissors
- 4×4 gauze pads
- Adhesive tape
- 3-0 absorbable suture
- 1/4 inch iodoform packing
- Silver nitrate sticks
- Sterile dressing

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida, Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_68

68.4 Procedure

1. Patient should be in either prone, left lateral decubitus, or jackknife position (Fig. 68.2).
2. Place 2 vertically oriented pieces of tape down each buttock from lower back to upper thigh. Next, place a perpendicular (horizontal) strip spreading buttocks to either side, securing gurney (Fig. 68.3).
3. Place sterile drape over field and center direct lighting over field (Fig. 68.4).
4. Wipe area with alcohol.
5. Inject 1–2 cc of anesthetic into base of hemorrhoid.
6. Clean area with povidone-iodine or chlorhexidine prep.
7. Make an elliptical incision in the roof of the hemorrhoid, being careful to avoid anal sphincter muscle.
8. Remove blood clot(s); multiple clots are often present.
9. If profuse bleeding is a problem, consider cauterization with silver nitrate sticks.
10. Wound can be closed with figure-of-8 absorbable suture OR can be loosely packed with1/4 inch iodoform gauze if not suturing (do not suture wound closed with packing inside).
11. Cover wound with 4×4 gauze folded in half and taped into place (Fig. 68.5).

- Discharge medicines:
  - Antibiotics generally not necessary.
  - Prescribe ibuprofen and/or acetaminophen for analgesia. Avoid opiates, which are constipating.
  - Prescribe stool softeners, to be taken two to three times daily.
- Discharge instructions to patient:
  - Sitz baths 3–4 times daily, for 20 min, warm not hot water.
  - Packing should fall out spontaneously in 2 days.
  - Keep well hydrated.
  - Use gauze to protect underclothing from soilage/blood stains.
  - Return to ED if pain persists beyond 48 h.

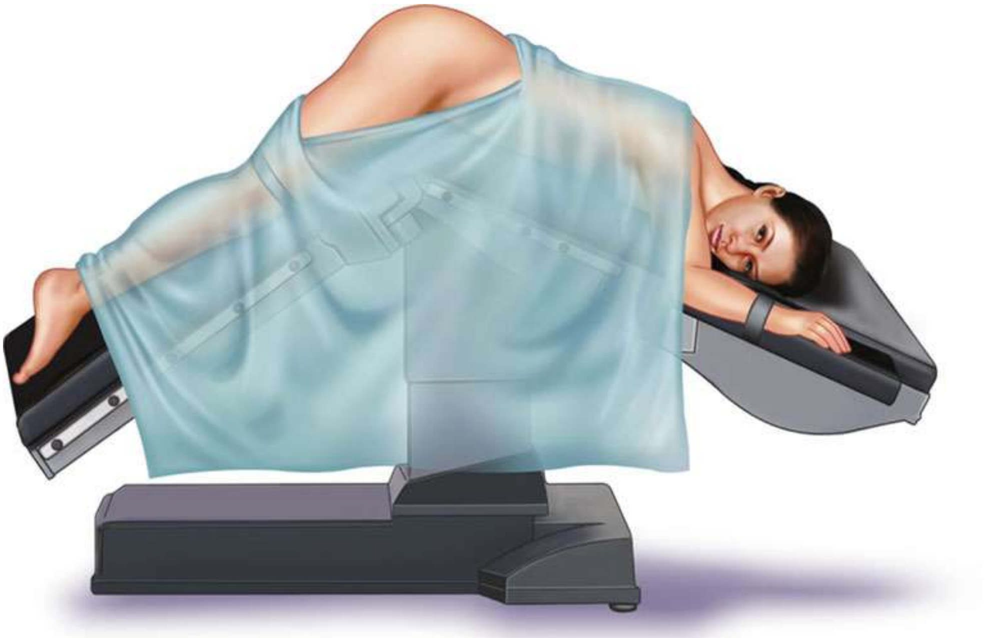

Fig. 68.2 Jackknife position

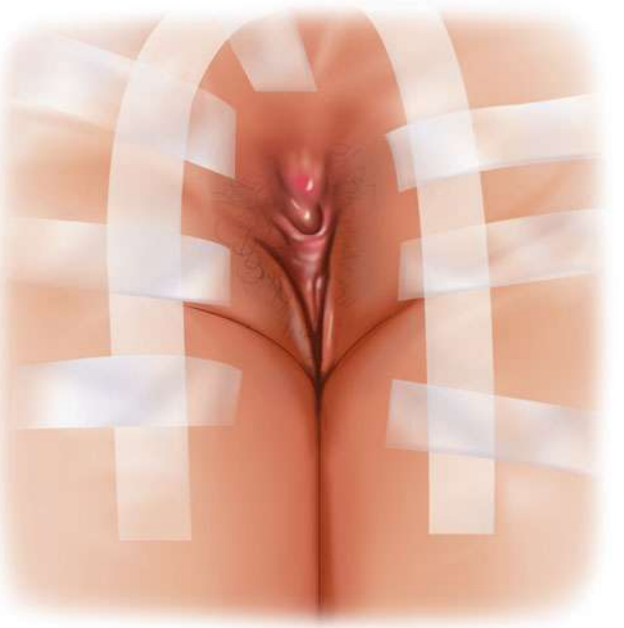

Fig. 68.3 Taping of buttocks to maximize visualization of hemorrhoid

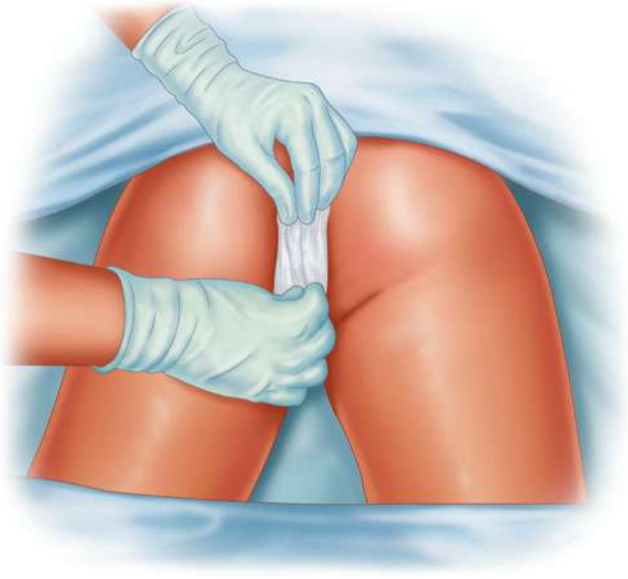

Fig. 68.5 Dress wound with sterile gauze

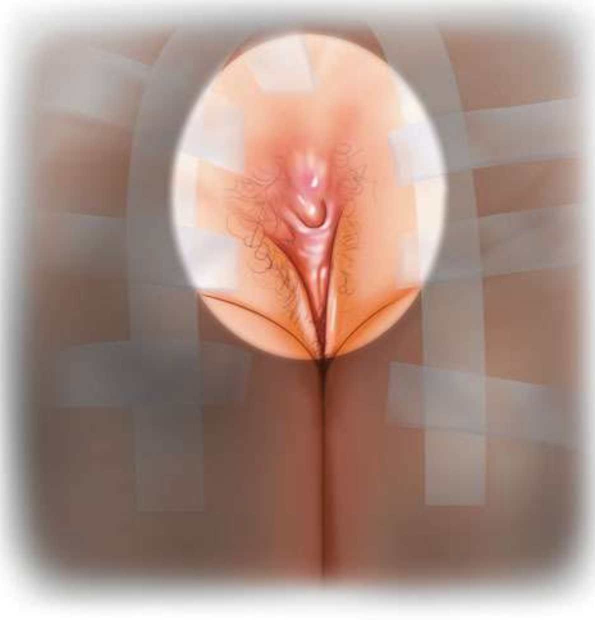

Fig. 68.4 Direct lighting over field

68.5 Complications

- Common
  - Bleeding: usually self-limited. Can apply cautery or figure-of-8 suture if not previously done
  - Pain: usually controlled with ibuprofen or acetaminophen
  - Perianal skin tag: benign
- Rare
  - Infection rate is 5 % [1].
  - Recurrence rate is 5–19 %, vs. 30 % for simple lancing [2].
  - Stricture and/or incontinence: prevented by avoiding underlying external anal sphincter muscle

68.6 Pearls and Pitfalls

- Pearls
  - Elliptical excision of the hemorrhoid results in much lower recurrence rate than simple lancing.
  - Risk factors for thrombosed external hemorrhoids include constipation, 2nd or 3rd trimester pregnancy, and traumatic vaginal delivery.
- Pitfalls
  - Excision of multiple hemorrhoids in circumferential fashion on all sides of the anal canal can cause anal stenosis.
  - Excision of a painless mass: if it is *painless*, it is *not* a thrombosed external hemorrhoid.

References

1. Lorber BW. Thrombosed external hemorrhoid excision. Medscape.com. www.emedicine.medscape.com/article/81039. Accessed 27 July 2014.
2. Rivadeneira DE. Outpatient and surgical procedures for hemorrhoids. UpToDate.com. http://www.uptodate.com/contents/outpatient-and-surgical-procedures-for-hemorrhoids. Accessed 27 July 2014.
3. Fargo MV, Latimer KM. Evaluation and management of common anorectal conditions. Am Fam Physician. 2012;85(6):624–30.

Selected Reading

Fargo MV, Latimer KM. Evaluation and management of common anorectal conditions. Am Fam Physician. 2012;85(6):624–30.

Jongen J, Bach S, Stübinger SH, Bock JU. Excision of thrombosed external hemorrhoid under local anesthesia: a retrospective evaluation of 340 patients. Dis Colon Rectum. 2003;46(9):1226–31.

Diagnostic Peritoneal Lavage

69

Latha Ganti

69.1 Indications (Table 69.1)

- Inability to perform FAST exam due to lack of equipment or operator
- Hemodynamically unstable patient in whom FAST exam is negative or equivocal

Table 69.1 Comparison parameters for DPL, FAST, and CT [1]

	DPL	FAST	CT
Speed	10–15 min	Fastest: <5 min	Variable
Repeatable	Yes, but rarely done	Yes, and frequently done	Yes, but not done often
Cost	$	$$	$$$
Invasive	Yes	No	No
Mobile	Yes	Yes	No
Advantages	Most sensitive for mesenteric and hollow viscus injuries	Highest specificity	Highly accurate but can be hampered by patient movement
Disadvantages	Misses retroperitoneal and diaphragm injuries	Hampered by subcutaneous or intra-abdominal air, obesity, and pelvic fractures Significant false-negative rate	Misses diaphragm, small bowel, and pancreatic injuries Small but significant risk of radiation-associated malignancy Cannot be done at bedside

---

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_69

69.2 Contraindications

- Absolute
  - Indication for laparotomy already exists.
- Relative
  - 2nd or 3rd trimester pregnancy
  - Previous lower abdominal surgery
  - Inexperienced operator
  - Abdominal wall infection
  - Coagulopathy
  - Cirrhosis
  - Morbid obesity

69.3 Materials and Medications

- 10 % povidone iodine prep
- 1 % lidocaine with epinephrine
- Fenestrated drape
- #10 scalpel blade and scalpel holder
- Skin retractors
- Hemostats
- Diagnostic peritoneal lavage (DPL) catheter (standard peritoneal dialysis catheter)
- 10 cc syringe
- Warmed lactated Ringer's or normal saline solution
- Skin stapler
- Simple suture tray with suture material

69.4 Procedure

69.4.1 Patient Preparation

- Place patient in supine position.
- Ensure nasogastric and urethral catheter (Foley) are in place.
- Prep and drape the area from the umbilicus to the symphysis pubis.
- Anesthetize the skin using 1 % lidocaine with epinephrine in the midline where incision will be made (Fig. 69.1).

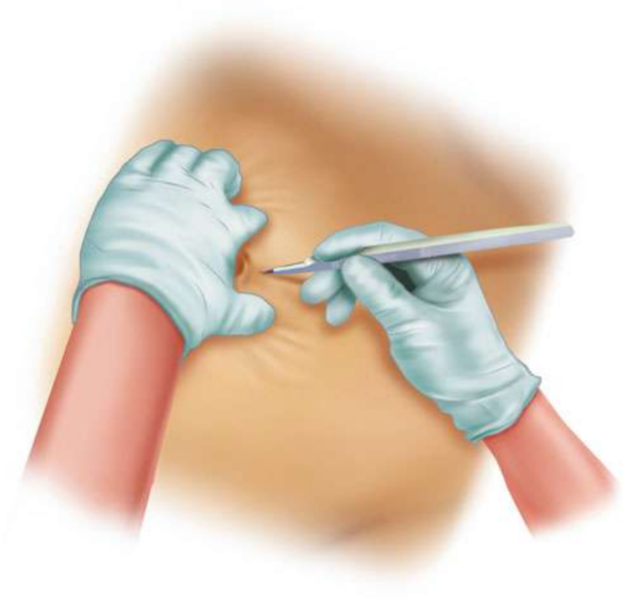

Fig. 69.1 Anesthetize skin where incision will be made (supraumbilical incision shown here, can also make infraumbilical incision)

69.4.2 Three DPL Techniques

- Semiopen (Seldinger) technique
  1. Using a #10 scalpel blade, make a 2 cm incision either superior or inferior to the umbilicus.
  2. Dissect subcutaneous fat until linea alba is exposed.
  3. Hold incision open with skin retractors (Fig. 69.2).
  4. Grasp fascia with hemostats on either side of midline.
  5. Insert 18 gauge needle at 45° angle toward pelvis (Fig. 69.3).
  6. First "pop" will be heard once fascia is penetrated.
  7. Second "pop" will be heard once peritoneum is traversed.
  8. Pass guidewire through needle into pelvis (should pass easily without resistance).
  9. Remove needle while keeping wire stable.
  10. Pass dilator over the wire through fascia, and remove (Fig. 69.4).
  11. Slip DPL catheter over guidewire aiming toward pelvis.
  12. Aspirate peritoneal contents with syringe; aspiration of blood is a positive DPL and means immediate laparotomy (can stop DPL procedure here).
  13. If no blood is immediately obvious, then connect the DPL catheter to a liter of warmed lactated Ringer's (LR) or normal saline (NS) solution for lavage (ensure setup has no one-way valves as solution and peritoneal fluid need to be able to freely mix).
  14. Place LR or NS bag on floor once it is almost empty (minimum 300–350 ml for adults or 10–15 ml/kg for children) and allow intra-abdominal fluid to return (Fig. 69.5).
  15. Send fluid for analysis (Table 69.2).
  16. Irrigate wound, and close skin only with staples or sutures.
- Open technique
  1. Make a 5 cm incision inferior to the umbilicus over linea alba and directly visualize peritoneal cavity.
  2. Both fascia (absorbable suture) and skin (nonabsorbable suture) need to be closed.
- Closed technique
  1. Access peritoneal cavity via percutaneous needle access.
  2. No surgical closure required.

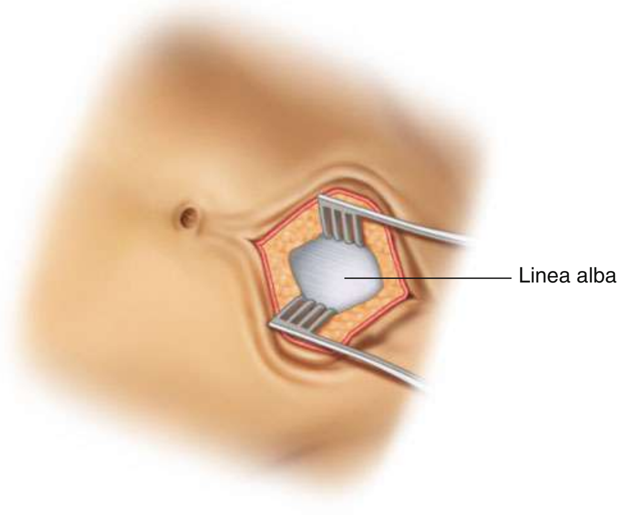

Fig. 69.2 Hold incision open with skin retractors

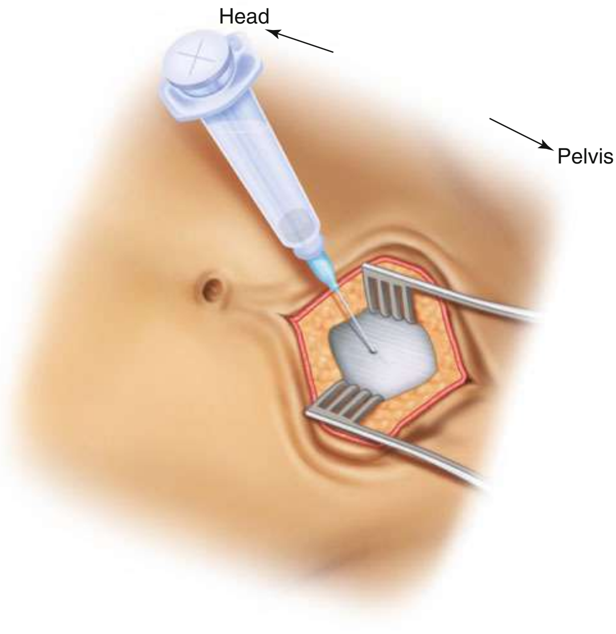

Fig. 69.3 Insert 18 gauge needle at 45° angle toward the pelvis

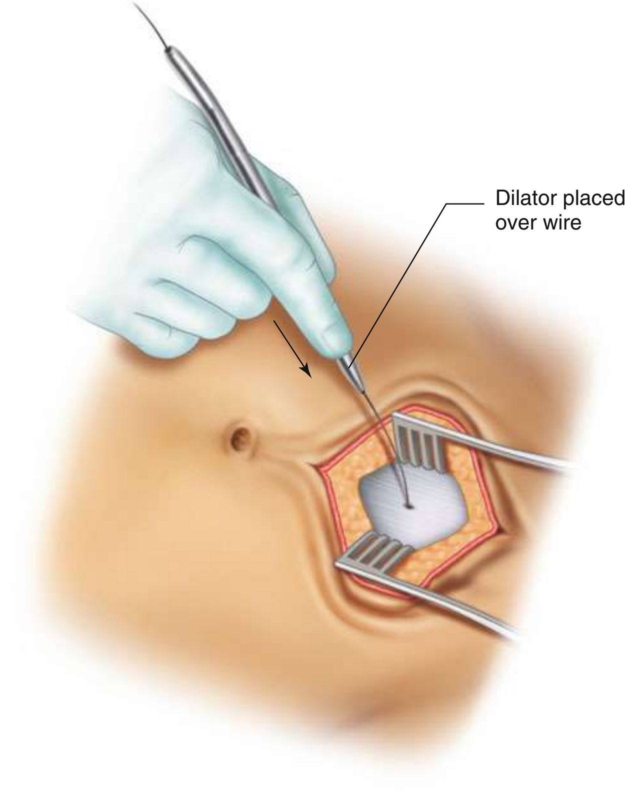

Fig. 69.4 Pass dilator over the wire through fascia and remove

Table 69.2 Diagnostic peritoneal lavage red blood cell criteria (per mm^3) [2]

	Positive	Indeterminate
Immediate gross return of blood via catheter	Any amount	
Immediate return of food particles/intestinal contents	Any amount	
Aspiration of blood	10 cc	
RBC in blunt trauma	100,000	20,000–100,000
RBC in penetrating trauma	10,000	5000–10,000
RBC in gunshot wound	5000	1000–5000
Amylase level (IU/L)	≥175	
Alkaline phosphatase level (IU/L)	≥3	
WBCs (per mm^3)	>500	250–500

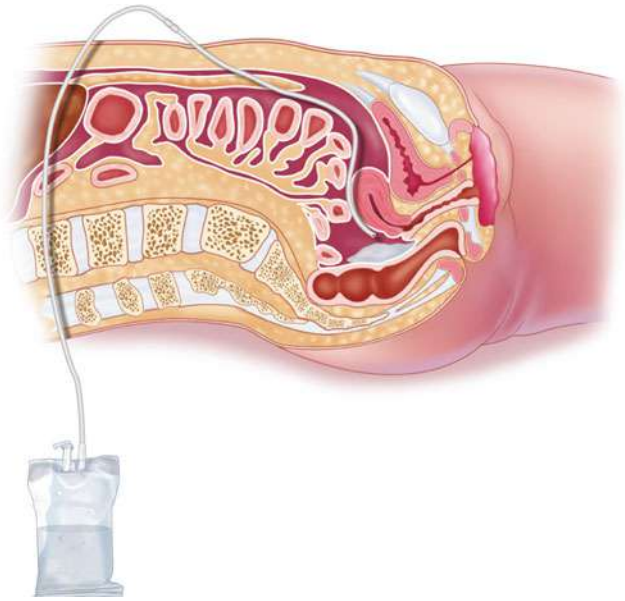

Fig. 69.5 Lavage

69.5 Complications

- Wound infection or dehiscence
- Intraperitoneal injury to organs or vessels (iatrogenic hemoperitoneum)
- Unnecessary laparotomy due to false-positive result from bleeding within rectus sheath or from site of incision
- Potential failure to recover lavage fluid due to:
  - Inadvertent placement of the catheter into the preperitoneal space
  - Compartmentalization of fluid by adhesions
  - Obstruction of fluid outflow (e.g., by omentum)
  - Fluid pooling in the intrathoracic cavity due to diaphragmatic injury
- Incisional hernia

69.6 Pearls and Pitfalls

- Pearls
  - When properly done, complication rate for DPL is low.
  - Prophylactic antibiotics are generally not indicated.
- Pitfalls
  - Inadequate decompression of stomach and urinary bladder increases the chance of injury to these organs; thus, nasogastric and Foley decompression is an important step in patient preparation.

References

1. Jagminas L. Diagnostic peritoneal lavage. Medscape.com. http://emedicine.medscape.com/article/82888-overview#a17. Accessed 28 Aug 2014.
2. Marx JA. Diagnostic peritoneal lavage. In: Ivatury RR, Cayten CG, editors. The textbook of penetrating trauma. Baltimore: Williams & Wilkins; 1996. p. 337.

Selected Reading

Whitehouse JS, Weigelt JA. Diagnostic peritoneal lavage: a review of indications, technique, and interpretation. Scand J Trauma Resusc Emerg Med. 2009;17:13.

Manual Reduction of Abdominal Hernia

70

Latha Ganti

An abdominal wall hernia is a protrusion of the intestine through an opening or area of weakness in the abdominal wall. See Table 70.1 for types of abdominal hernias and Fig. 70.1 for locations along the abdominal anatomy.

Table 70.1 Types of abdominal hernias

Type	Defect	Most commonly seen in	Notes
Inguinal	Intestine or bladder protrudes through abdominal wall or into inguinal canal in the groin	Men because of a natural weakness in this area	96 % of all groin hernias are inguinal; 4 % are femoral
Femoral	Intestine enters canal carrying femoral artery into the upper thigh	Women, especially those who are pregnant or obese	
Incisional	Intestine pushes through abdominal wall at the site of previous abdominal surgery	Elderly or overweight people who are inactive after abdominal surgery	
Umbilical	Part of the small intestine passes through abdominal wall near the navel	Newborns and obese women or those who have had many children	In children, not repaired until age five because often resolve on their own
Hiatal	Upper stomach squeezes through hiatus, an opening in the diaphragm through which the esophagus passes		

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_70

Fig. 70.1 Types of abdominal wall hernias

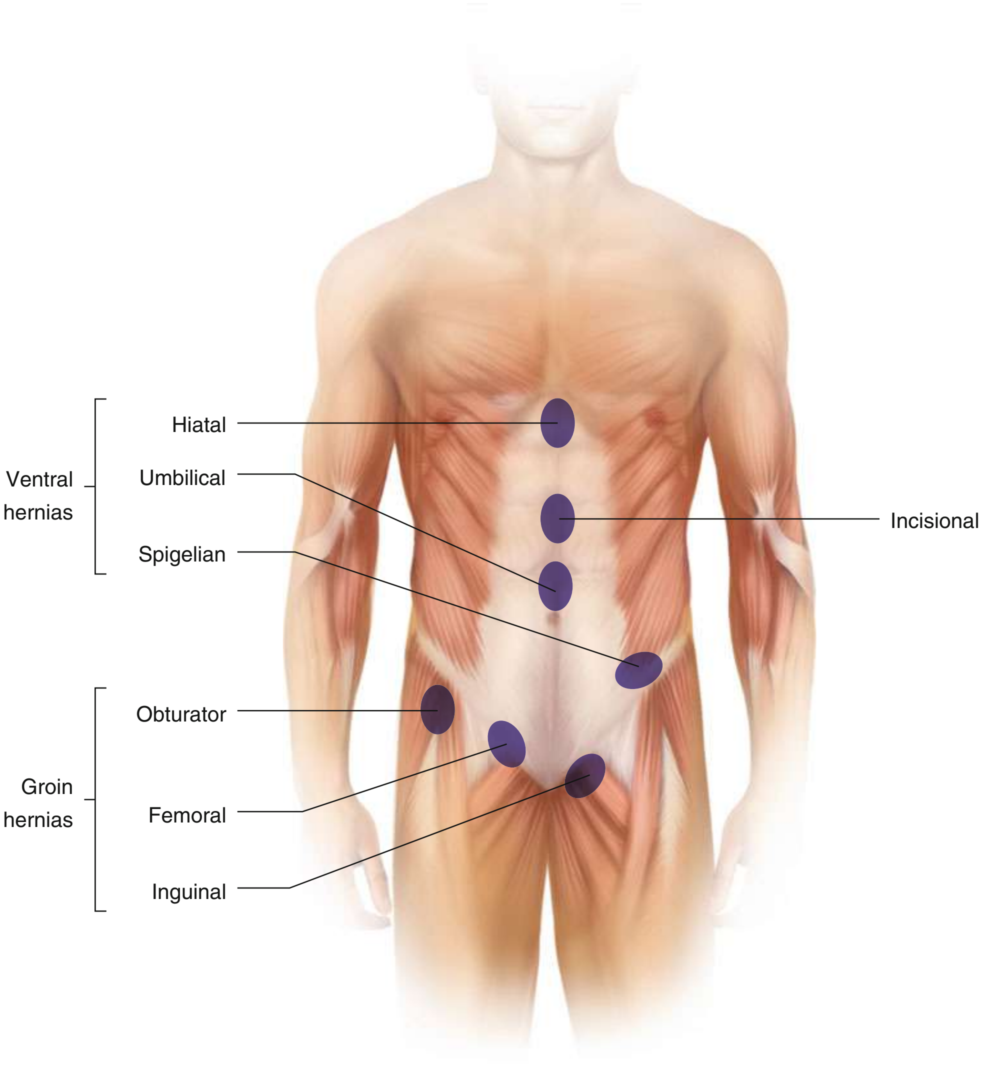
Ventral hernias
Hiatal
Umbilical
Spigelian
Incisional
Groin hernias
Obturator
Femoral
Inguinal

70.1 Indications

- Incarcerated hernia

70.2 Contraindications

- Absolute
  - Strangulated hernia (could result in placing dead bowel into abdominal cavity)
- Relative
  - Inability to get patient relaxed enough
  - Previous unsuccessful attempts

70.3 Materials and Medications

- Ice or cold compress
- Stretcher or gurney that can tilt to provide Trendelenburg position
- Moderate sedation drugs if providing moderate sedation
- Truss for post-procedure

70.4 Procedure

1. Patient positioning:
   - For abdominal hernia: place patient supine.
   - For groin hernia in adult: place in 20 ° of Trendelenburg.
   - For groin hernias in children: place in unilateral frog leg position (Fig. 70.2).
2. Apply ice or cold compress directly over hernia site to reduce swelling.
3. Administer opiate analgesia or moderate/procedural sedation.
4. Wait up to 30 min as hernia may reduce spontaneously after swelling has gone down and patient is relaxed.
5. Gently apply steady pressure distally on the tissue at the neck of the hernia with one hand and with other hand, guide hernia proximally through fascial defect. Too much pressure distally can cause hernia to balloon further, making manual reduction difficult. Takes up to 15–20 min. Do not rush procedure.
6. Once hernia is reduced, pain will improve.
7. An external support garment or truss (Fig. 70.3) can be helpful to hold reduced hernia in place and serve as a temporizing measure until surgical repair can be done.
8. Advise patient to schedule elective surgical repair.
9. If unable to reduce the hernia, obtain surgical consultation. Do not force repeated attempts.

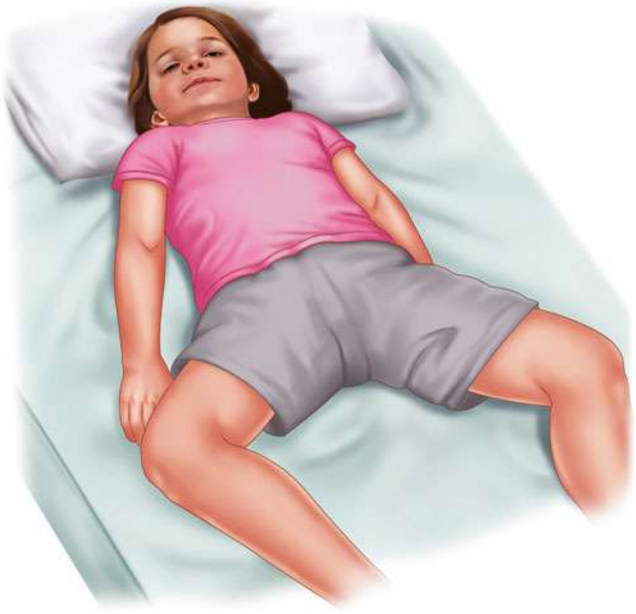

Fig. 70.2 Frog leg position in child

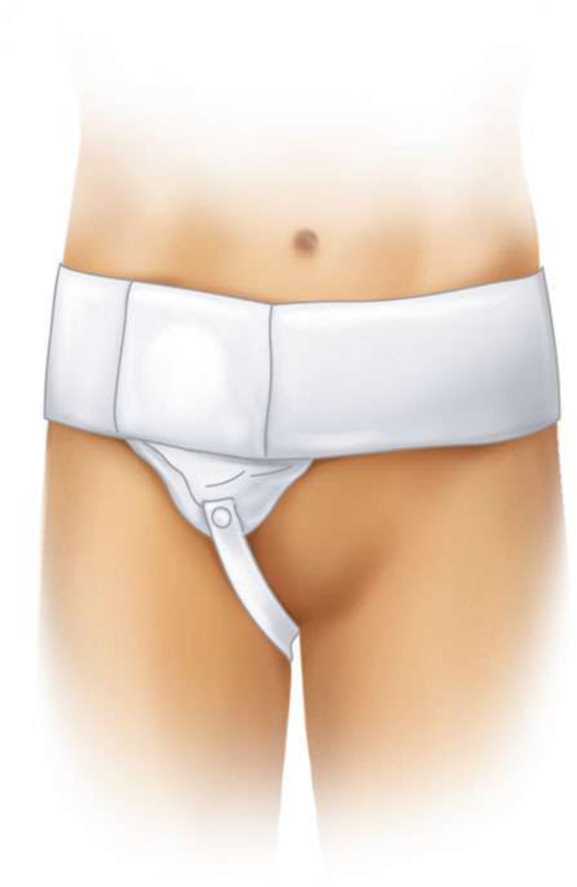

Fig. 70.3 An example of a truss, or external support, that can be useful as a temporizing measure until definitive hernia repair can be done

70.5 Complications

- Pain.
- Inability to achieve manual reduction, leading to strangulation of the hernia.
- Strangulation can result in peritonitis and sepsis.
- Recurrence.
- Hydrocele.

70.6 Pearls and Pitfalls

- Pearls
  - Definitive treatment for a hernia is surgery (herniorrhaphy). Without surgery, hernias grow larger over time; they do not disappear. Small hernias are easier to fix and result in fewer complications.
  - The only hernia that can resolve on its own is an umbilical hernia in a child.
  - Trusses, bandages, and tape may provide some comfort but do not reduce risk of incarceration or strangulation.
  - Note that if a truss is worn, it should be in place after reduction of the hernia. Also, it can be impractical in hot climates.
- Pitfalls
  - When the constricting neck and the protrusion are both reduced into the abdomen together (known as a reduction *en masse*), without actually reducing the hernia itself, strangulation ensues even though it appears one has reduced the hernia.
  - If there is still considerable pain after the reduction, it is likely the reduction was not successful or that dead bowel has been reduced into the abdominal cavity.
  - Not recognizing strangulation leads to gangrenous bowel, peritonitis, and sepsis.

Selected Reading

Campanelli G, Canziani M, Frattini F, et al. Inguinal hernia: state of the art. Int J Surg. 2008;6 Suppl 1:S26–8.

Jenkins JT, O'Dwyer PJ. Inguinal hernias. BMJ. 2008;336(7638):269–72.

Moses S. Hernia reduction. 2014. http://www.fpnotebook.com/mobile/Surgery/GI/HrnRdctn.htm. Accessed Sept 15, 2014.

Extended Focused Assessment with Sonography for Trauma

71

Coben Thorn and L. Connor Nickels

71.1 Indications

- Blunt abdominal or chest trauma
- Penetrating abdominal or chest trauma
- Undifferentiated hypotension
- The “E” in EFAST refers to the “extended” ability to detect lung pathology such as a pneumothorax or hemothorax during the otherwise standard trauma FAST exam using the same equipment with or without an additional transducer probe.
- Specific findings that can be detected on extended focused assessment with sonography for trauma (EFAST):
  - Pericardial fluid
  - Pleural fluid
  - Free intraperitoneal fluid
  - Pneumothorax
- Free fluid appears as anechoic or black.

71.2 Contraindications

- Need for immediate operative intervention

71.3 Materials and Medications

- Ultrasound machine
- Probe(s): phased array probe (5 to 1 MHz) or curved array probe (5 to 2 MHz)
  - Phased array has a smaller footprint, allowing easier access between intercostal spaces (Fig. 71.1); however, curved array provides better resolution of images

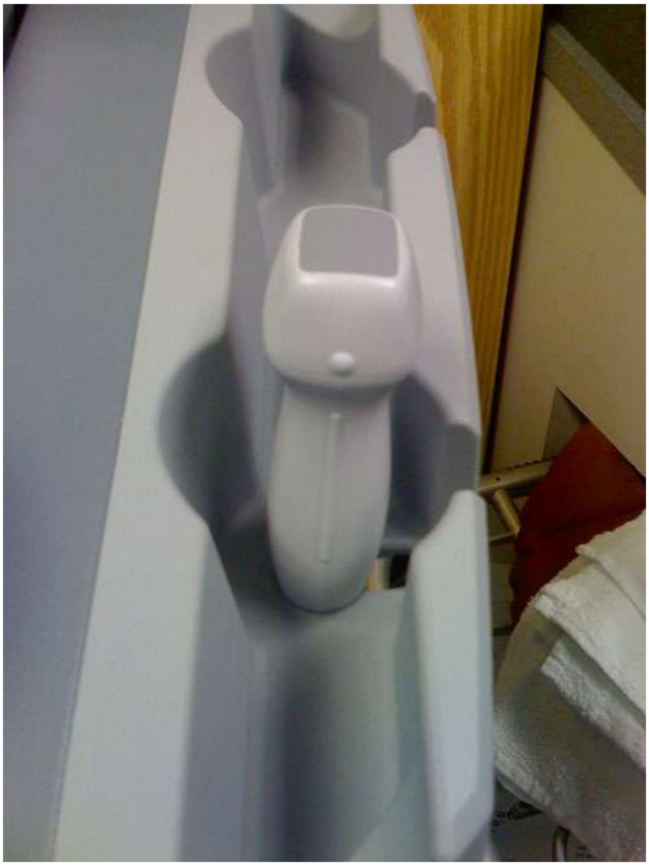

Fig. 71.1 Phased array transducer (P17) with a small footprint that is helpful to fit between the ribs and can be used for focused assessment with sonography for trauma (FAST) examination

C. Thorn, MD
Department of Emergency Medicine, Bon Secours St. Francis Health System, Greenville, SC, USA
e-mail: cobenthorn@gmail.com

L.C. Nickels, MD, RDMS (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_71

(Fig. 71.2). The linear array transducer (L38, 10–5 MHz) is good for lung images.

- Gel
- Skilled ultrasound operator
- ± Laboratory work, cardiac monitor, and two large-bore intravenous (IV) needles
  - All trauma alerts and unstable patients must have all of these.

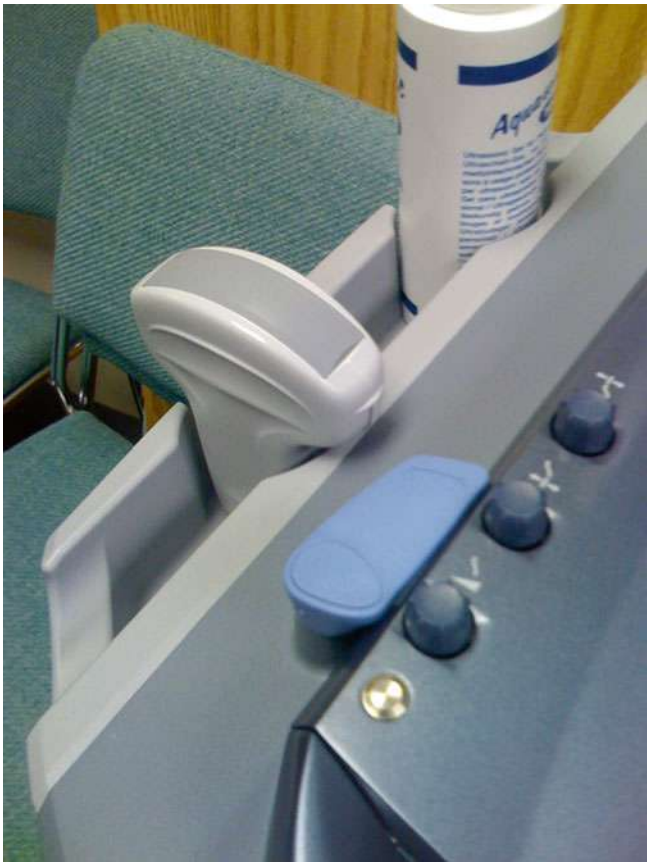

Fig. 71.2 Curved array transducer (C60) with a larger footprint and better resolution for deeper imaging that can be used for FAST examination as well as lung examination

71.4 Procedure

1. Ultrasound machine in the abdominal preset.
2. Patient in the supine position.
3. Phased array or curved array probe for focused assessment with sonography for trauma (FAST) and linear array for lung.
4. Begin scanning the patient in a systematic fashion.
   - All the views should be scanned by thoroughly sweeping through the area in question in order to maximize the information obtained.
   - All views should be obtained in the same order every time.
   - Obtain all four views, five views if pneumothorax is included.

71.4.1 Subxiphoid Four-Chamber View (Fig. 71.3)

1. Examine for free pericardial fluid.
    - Anechoic (black) stripe seen between the myocardium and the pericardium
2. Probe is placed in the subxiphoid area.
3. Indicator is to the patient's right.
4. Probe is directed toward the patient's left shoulder.
5. Use a shallow angle in the head to feet direction.
6. Should adequately visualize the following:
    - Liver edge superficially
    - Right ventricle
    - Left ventricle
    - Right atrium
    - Left atrium
7. If unable to obtain this view, proceed to *parasternal long-axis view*:
    - Probe is placed perpendicular at the left parasternal border.
    - Third to fourth intercostal space.
    - Indicator is to the patient's right shoulder.
    - Coronal section through the heart's long axis should adequately visualize the following:
        - Right ventricle most superficially
        - Left ventricle
        - Mitral valve
        - Left atrium
        - Aortic valve
        - Aortic outflow tract

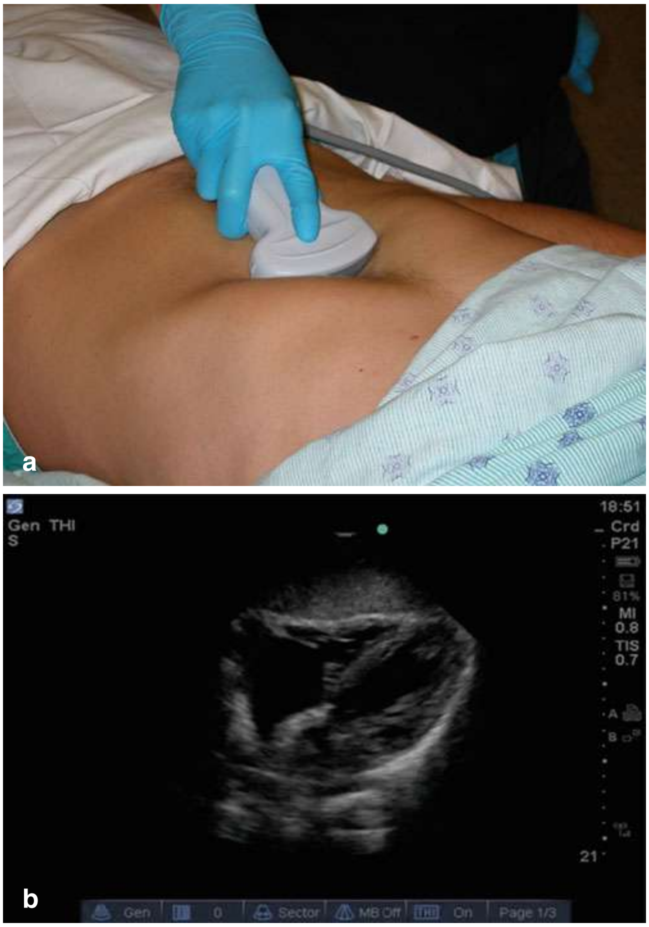

Fig. 71.3 (**a**) Image shows how to obtain the subxiphoid four-chamber view of the heart using the curved array transducer (C60) (Photograph courtesy of F. Eike Flach, MD).(**b**) Ultrasound image of four-chamber view of the heart (Used with permission from First aid for the emergency medicine clerkship 3rd Ed, McGraw Hill, 2011.) *RV* right ventricle, *LV* left ventricle, *RA* right atrium, *LA* left atrium

71.4.2 Right Upper Quadrant View (Fig. 71.4)

1. Examine for free fluid in all of the following areas:
   - Right intrathoracic space
     - Anechoic area above the diaphragm
   - Morison's pouch: hepatorenal space
     - Anechoic stripe between the liver and the kidney
   - Right paracolic gutter
     - Anechoic collection surrounding the inferior tip of the kidney
2. Probe is placed in the midaxillary line on the right.
3. Indicator is directed toward the patient's head.
4. Probe is in the coronal plane, angle can be aimed obliquely while scanning anterior to posterior.

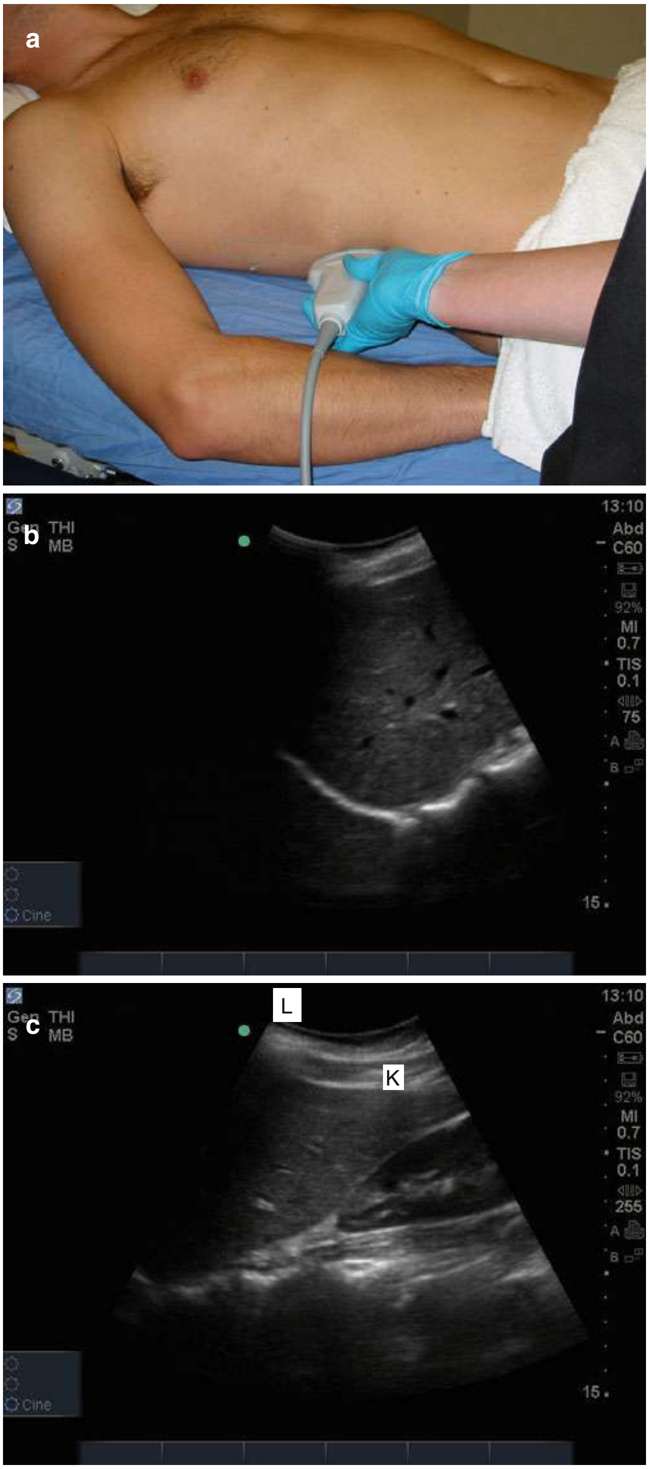

Fig. 71.4 (**a**) Image shows how to obtain the right upper quadrant view of the FAST exam using the curved array transducer (C60). The probe is aimed slightly obliquely in the coronal plane to get a better view between the ribs (Photograph courtesy of F. Eike Flach, MD). (**b**) Ultrasonographic view of the liver. (**c**) Ultrasonographic view of the liver–kidney interface (Morrison's pouch). *L* liver, *K* kidney

71.4.3 Left Upper Quadrant View (Fig. 71.5)

1. Examine for free fluid in all of the following areas:
   - Left intrathoracic space
     - – Anechoic area above the diaphragm
   - Subphrenic space
     - – Anechoic stripe below the diaphragm and above the spleen
   - Splenorenal space
     - – Anechoic stripe between the spleen and kidney
   - Left paracolic gutter
     - – Anechoic collection surrounding the inferior tip of the kidney
2. Probe is placed in the midaxillary line on the left.
3. Indicator is directed toward the patient's head.
4. Probe in the coronal plane, angle can be aimed obliquely while scanning anterior to posterior.

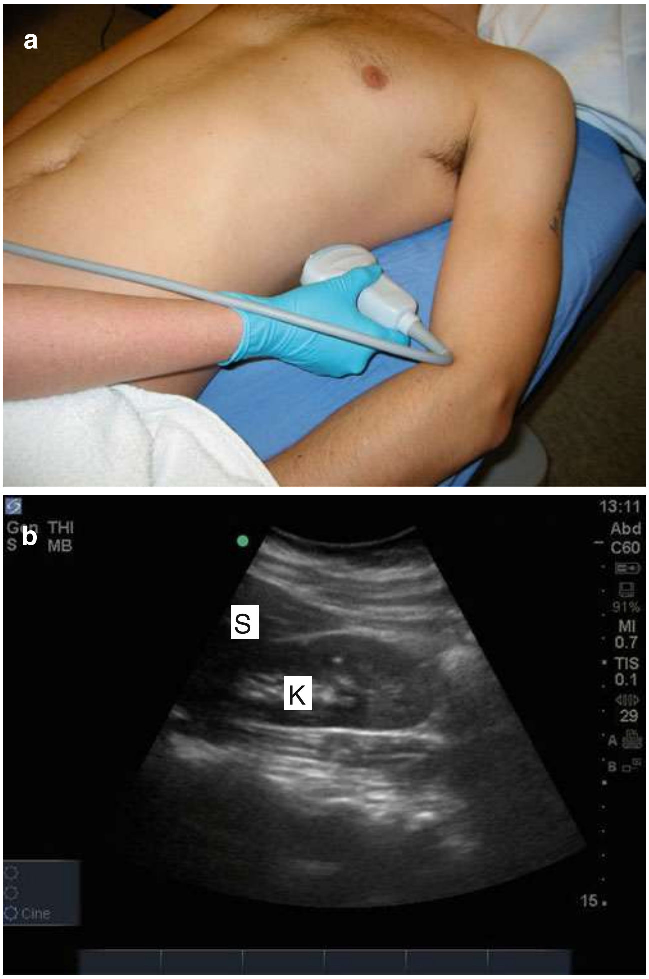

Fig. 71.5 (**a**) Image shows how to obtain the left upper quadrant view of the FAST examination using the curved array transducer (C60). Again, the probe is aimed slightly obliquely and is placed more superiorly in the midaxillary line (Photograph courtesy of F. Eike Flach, MD). (**b**) Ultrasonographic view of spleen–kidney interface. *S* spleen, *K* kidney

71.4.4 Pelvic View (Figs. 71.6 and 71.7)

1. Examine for intraperitoneal free fluid in the pelvis:
   - Anterior pelvis, above the bladder
     - – Anechoic fluid above the bladder
   - Posterior cul-de-sac (pouch of Douglas)
     - – Anechoic fluid posterior to the bladder or uterus
2. Probe is placed above the pubic symphysis over the bladder.
3. Scan through in both planes:
   - Transverse plane (Fig. 71.6)
     - – Indicator is to the patient's right.
     - – Scan through the bladder in the head to feet direction.
   - Sagittal plane (Fig. 71.7)
     - – Indicator is aimed to the patient's head.
     - – Scan through the bladder in a right to left direction.

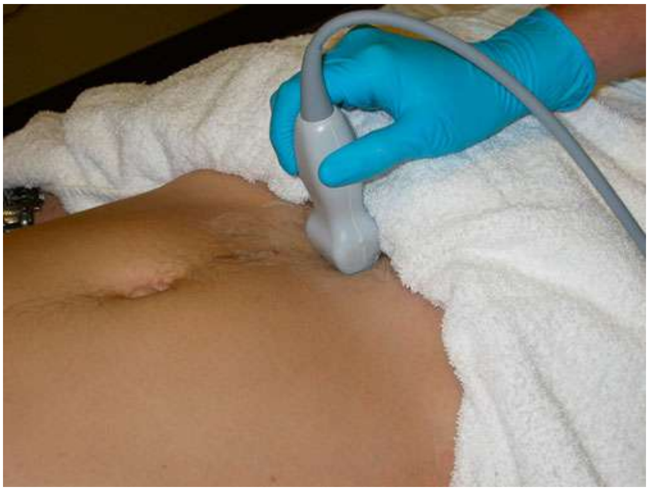

Fig. 71.6 Transverse pelvic view of the FAST examination using the phased array transducer (P17). With gentle force, the probe is pressed downward in order to look back behind the pubic symphysis and view the bladder (Photograph courtesy of F. Eike Flach, MD)

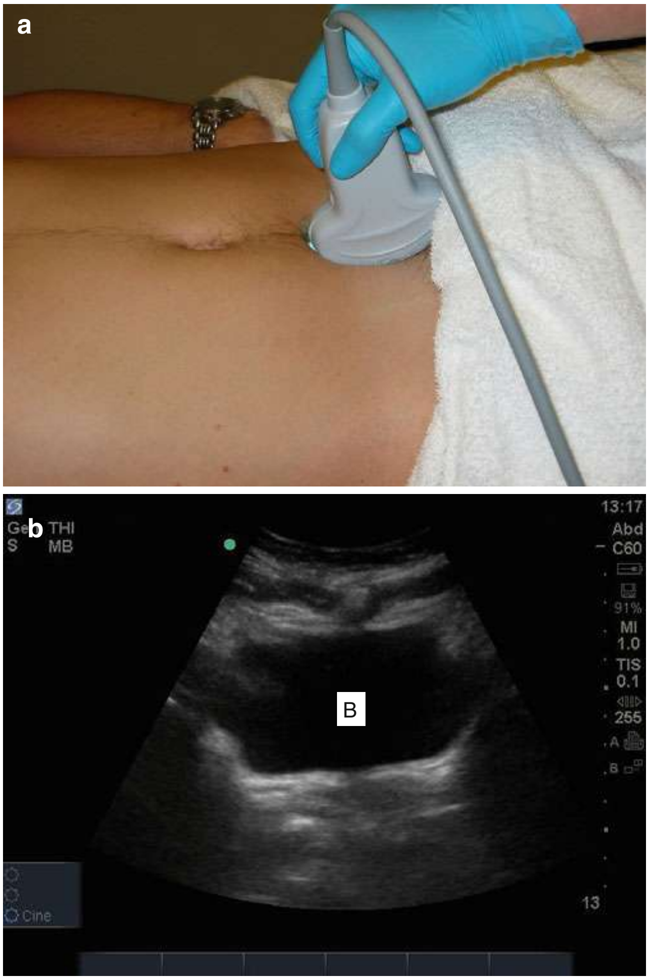

Fig. 71.7 (**a**) Sagittal pelvic view of the FAST exam using the curved array transducer (C60). With gentle force, the probe is pressed downward in order to look back behind the pubic symphysis and view the bladder (Photograph courtesy of F. Eike Flach, MD). (**b**) Ultrasonographic view of the bladder. *B* bladder

71.4.5 EFAST with Lung Views (Figs. 71.8 and 71.9)

1. Examine for pneumothorax:
   - Lung sliding
     - Absence: pneumothorax
     - Presence: normal lung
   - M mode tracing
     - Stratosphere sign: pneumothorax
     - Seashore sign: normal lung
2. Probe is placed on the anterior chest in the midaxillary line.
3. Level of the second to fourth intercostal spaces.
4. Sagittal position.
5. Center the probe over the pleural line between the ribs.
   - Find the rib and then slide the probe toward the head or feet to center the pleural line.
6. Observe for lung sliding.
7. Press M mode and move the line over the pleural line and press M mode again to get the tracing.
8. Examine multiple other areas anteriorly, moving distally, and in midaxillary line laterally, moving from superior to inferior.

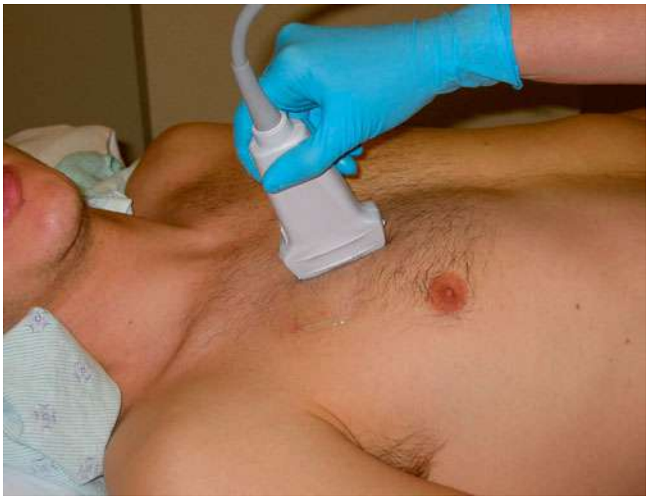

Fig. 71.8 Right lung view of the extended focused assessment with sonography for trauma (EFAST) examination using the linear array transducer (L38). The probe is placed in the sagittal plane on the anterior chest in the midaxillary line approximately at the second intercostal space and centered over the pleural line (Photograph courtesy of F. Eike Flach, MD)

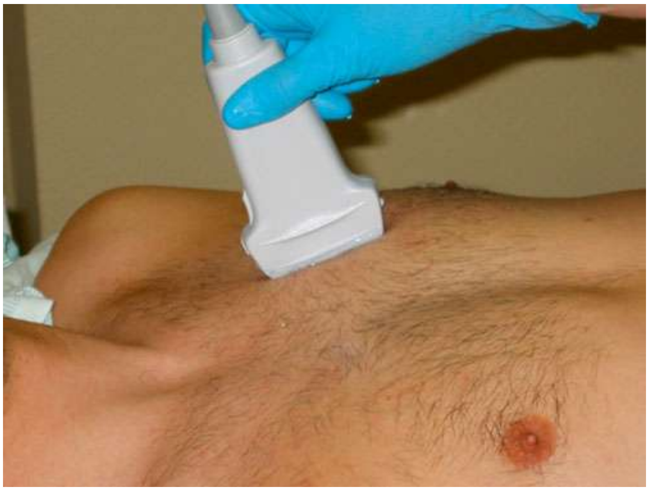

Fig. 71.9 Left lung view of the EFAST examination using the linear array transducer (L38) (Photograph courtesy of F. Eike Flach, MD)

71.5 Complications

- Overreliance on ultrasound to rule out abdominal injury:
  - FAST examinations do not detect retroperitoneal bleeding, solid organ injury, contained subcapsular hematomas, and bowel injuries.
- Not scanning through the object in question could lead to false-negative results.

71.6 Pearls and Pitfalls

- Always follow the ABCs (airway, breathing, circulation) first in any unstable patient.
- Always make sure the depth is set adequately.
  - Recommend starting deeper to make sure positive findings are not missed and then adjustments can be made from there.
- The curved array probe may be used throughout the entire EFAST for convenience if necessary.

71.6.1 Subxiphoid Four-Chamber View

- For larger body habitus, need to parallel the probe with the body in the subxiphoid area and use firm pressure to press the entire probe downward so as to look up under the xiphoid process at the heart.
- Moving the entire probe more to the patient's right in the subxiphoid area while still looking toward the left shoulder may improve visualization by using the liver as a window.
- Fat pad:
  - May be mistaken for pericardial fluid
  - Contains echoes and, therefore, is hypoechoic rather than anechoic
  - Should only be present anteriorly
    - Fluid should be gravity dependent, completely encircling the heart, and seen in multiple views.

71.6.2 Right Upper Quadrant View

- Normal artifacts of mirroring and loss of the spine are obscured when pleural fluid is present and, instead, the anechoic fluid is seen and there is loss of mirroring and continuation of the spine.

71.6.3 Left Upper Quadrant View

- Same as right upper quadrant view
- May be more difficult view to find than in right upper quadrant view for all of the following:
  - Spleen and kidney are more posterior and superior than in right upper quadrant view.
  - Spleen is smaller and less of a window for viewing.

71.6.4 Pelvic View

- Bowel can be mista0ken for free fluid or vice versa, but holding the probe still and observing can sometimes help distinguish the two.
  - Peristalsis will occur with bowel.
  - Internal echoes may be present in bowel.

71.6.5 Lung Views

- Ultrasound is more sensitive than a supine portable chest x-ray.
  - Apex anteriorly in midaxillary line.
- Rib
  - Hyperechoic horizontal line with a dense shadow posteriorly
  - Evenly spaced along the chest
- Pleural line
  - First hyperechoic line deep to the rib.
  - Actually includes the visceral and parietal pleura, but appears as one line.
  - Lung sliding is present in normal lung.
  - Comet tail artifact.
  - M mode tracing will be the same in normal lung and pneumothorax above the pleural line and different below the pleural line.
    - Seashore sign (Fig. 27.5a):
      - Appears as waves washing up on the shore.
      - Granular appearance represents movement.
    - Stratosphere sign (Fig. 27.5b):
      - Appears as straight lines
      - Bar code appearance

Selected Reading

Brunett P, Cameron P. Trauma in adults. In: Tintinalli J, Stapczynski J, Ma OJ, Cline D, Cydulka R, Meckler G, editors. Emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw Hill; 2012. p. 1671–5.

Ma JO, Mateer JR, Blaivas M. Trauma. In: Emergency ultrasound. Course Materials; New York, NY: McGraw Hill; 2008. pp. 7–109.

Saul T, Rivera M, Lewiss R. Ultrasound image quality. ACEP News. 2011;4:24–5.

Nasogastric Tube Placement

72

David P. Nguyen, L. Connor Nickels,
and Giuliano De Portu

72.1 Indications

- Evaluation of upper gastrointestinal (GI) bleeding (history of melena, bright red blood per rectum, or coffee-ground emesis)
  - Only in the cases in which frank blood is obtained, the sensitivity/specificity in detecting upper GI bleeding is poor.
  - It should not be used for diagnostic purposes. It is used to remove blood that is irritating the stomach and to determine whether bleeding is still occurring (lavage does not clear).
- Commonly used in decompression of the GI tract (partial/complete small bowel obstruction)
- Prevents aspiration and gastric dilation in intubated patients
- Used during gastric lavage and/or removal of toxins (activated charcoal) for acute overdose or poisonings

72.2 Contraindications

- Absolute
  - Facial trauma with possible cribriform plate fracture
- Relative
  - Severe coagulopathy (consider orogastric tube placement)
  - Esophageal strictures and alkali ingestions (possible esophageal perforation)
  - Esophageal varices (studies show that it is actually safe)

72.3 Materials and Medications

- For awake patients, consider pretreatment: lidocaine gel (2 % viscous)/nebulized lidocaine (4 or 10 %), vasoconstrictors (e.g., phenylephrine 0.5 %), and antiemetic (e.g., ondansetron 4 mg).
- 16- or 18-French sump tube lubricating jelly
- 50- or 60-mL syringe stethoscope

D.P. Nguyen, DO
Department of Emergency Medicine, Rush-Copley Medical
Center, Aurora, IL, USA
e-mail: davidpnguyen@yahoo.com

L.C. Nickels, MD, RDMS (✉) • G. De Portu, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; gdeportu@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_72

72.4 Procedure

- Preparation
  1. Awake patients, should receive antiemetics 15 min before procedure.
  2. Anesthetize both nares at least 5 min before placement.
     - Spray vasoconstrictor into both nares.
     - Inject about 5 mL of lidocaine gel along the floor of the nose.
     - Nebulized lidocaine via facemask also reduces both nasal and pharyngeal discomfort.
  3. Elevate the head of the bed to an upright position (when possible).
  4. Estimate tube insertion distance by measuring the tube from the xiphoid to the earlobe and then to the tip of the nose. Add 6 in. to this estimate and note the total distance. This helps with placement in the stomach and prevents esophageal placement or coiling in stomach. Mark the tube with markers or tape at the desired length.
  5. Lubricate the nasogastric (NG) tube.
- Insertion (Fig. 72.1)
  1. Always insert the tube gently into the nares along the floor of the nose under direct visualization. Always point inferiorly (do not point upward).
  2. If resistance is encountered, try to apply a small amount of pressure. STOP if unable to advance. Try the other side. It is necessary to prevent bleeding or dissecting the tissues.
  3. Have the patient flex his or her head forward when the tube is in the nasopharynx. This helps direct the tube toward the correct placement in the esophagus and not the trachea. Have the awake and cooperative patient sip water from a straw and swallow as the tube enters the oropharynx.
  4. Making the tube more rigid by placing it in cold water will help advance it because the "warmer" tube will tend to coil.
  5. Once the tube is in the esophagus, rapidly advance the tube into the stomach, taking into consideration the previously marked depth.
- Confirmation of tube placement
  1. Insufflate air into the end of the NG tube, via a 50- or 60-mL syringe, while auscultating for a rush of air (borborygmi) over the stomach.
  2. Aspiration of gastric contents (pH<4, there is >90 % gastric placement).
  3. The awake and cooperative patient should be able to talk, and if coughing or severe discomfort occurs, consider that esophageal or bronchial placement might have occurred.
  4. Radiographic evaluation:
     - "Gold standard" is to evaluate simple radiograph for position.
     - Consider in comatose patients.
- Secure the tube
  1. Tape the NG tube in place by taping both the tube and the nose. A butterfly bandage is typically used. Some companies produce a specific fixation for the tube.
  2. Secure the tube to where it does not press on the medial or lateral nostril (can lead to bleeding/necrosis).

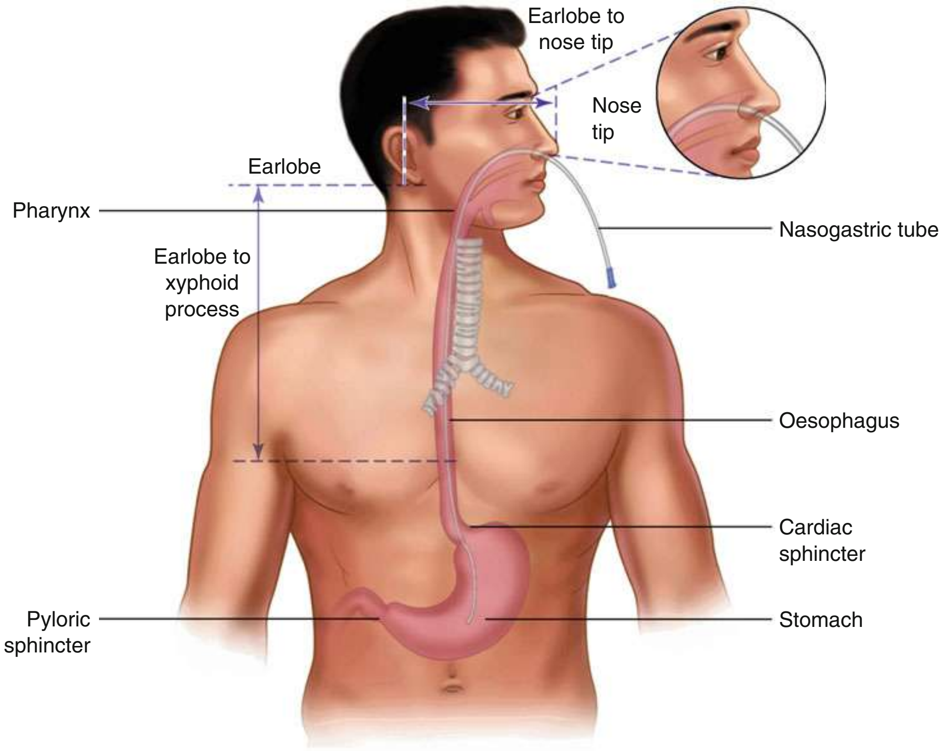

Fig. 72.1 NG tube placement

72.5 Complications

- Inability to pass the tube
- Bleeding
- Curling of the NG tube in the patient's mouth
- Pulmonary placement
- Nasal necrosis

72.6 Pearls

- Pearls
  - NG tube placement was ranked #1 as the most painful procedure in the emergency department so it is imperative to maintain patient's comfort by using anesthetics and even maybe intravenous anxiolytics.
  - Estimate the proper length of the tube before passage to avoid placing the tip of the tube in the esophagus or excessively coiling it in the stomach.
  - If leaving the tube for a prolonged period of time, make sure that the suction is set "intermittent" or "off" to prevent irritation to the gastric mucosa owing to direct pressure.

Selected Reading

Chun DH, Kim NY, Shin YS, Kim SH. A randomized, clinical trial of frozen versus standard nasogastric tube placement. World J Surg. 2009;33:1789–92.

Goff JS. Gastroesophageal varices: pathogenesis and therapy of acute bleeding. Gastroenterol Clin North Am. 1993;22:779.

Henneman PL. Gastrointestinal bleeding. In: Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Tho PC, Mordiffi S, Ang E, Chen H. Implementation of the evidence review on best practice for confirming correct placement of nasogastric tube in patients in an acute care hospital. Int J Evid Based Healthc. 2011;9:51–60.

Esophageal Foreign Body Removal

73

David P. Nguyen, L. Connor Nickels,
and Rohit Pravin Patel

73.1 Indications

- Patient presenting with any one or combination of the following:
  - Foreign body sensation in throat, neck, substernal chest, or epigastric area
  - Clear history of ingestion
  - Dysphagia
  - Airway compromise
  - Drooling
  - Inability to tolerate fluids
  - Inability to tolerate solids
  - Evidence of perforation
  - Active bleeding
- All unstable patients should have immediate airway management and urgent endoscopy.
- FBs lodged superior to the epiglottis may be retrieved by emergency physicians in an emergent situation, but generally, a consultant should be present, whether ear, nose, and throat, gastroenterology, or general surgery.

73.2 Contraindications

- Absolute
  - None
- Relative
  - Performing rapid sequence intubation (RSI) in a patient with an FB that could compromise the airway.
  - Generally, if the patient is breathing on their own, collaborate with a consultant on the best method to secure the airway (emergency department, intensive care unit, operating room).
  - Treating with glucagon repeatedly if it is inducing vomiting and/or not working.
  - Conservatively managing a patient who should otherwise undergo a procedure for removal.

73.3 Materials

- Esophagoscopy/endoscopy is the definitive diagnostic and therapeutic procedure for impacted esophageal FBs.
  - Generally, should not be performed by an emergency physician.
  - Devices used include forceps, baskets, polypectomy snares, and nets.
  - Endoscopic techniques include push into stomach, push plus fragmentation, pull with retrieval forceps, and pull with various items (basket, snare, nets).
- Foley catheter removal
  - Widely used technique for recently ingested single, smooth, blunt, and radiopaque objects
  - #12 to #16 French Foley catheter
  - Forceps (bayonet and Magill) of various sizes
  - Often done under fluoroscopic guidance
- Bougienage
  - A single, smooth object, such as a coin, lodged less than 24 h in a patient with no respiratory distress or esophageal disease can be advanced successfully into the stomach by using bougienage.
  - Dilator size is selected based on age:
    - 1–2 years: 28 French
    - 2–3 years: 32 French
    - 3–4 years: 36 French
    - 4–5 years: 38 French
    - Longer than 5 years: 40 French
- Relaxation of the lower esophageal sphincter (LES)
  - Some FBs lodged at the LES can be medically managed by relaxation of the LES.

D.P. Nguyen, DO
Department of Emergency Medicine, Rush-Copley Medical Center, Aurora, IL, USA
e-mail: davidpnguyen@yahoo.com

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_73

- Most ingested FBs and impacted food boluses eventually pass spontaneously.
    - 1–2 mg of glucagon intravenously
    - 0.4–0.8 mg of nitroglycerin sublingually
    - 5–10 mg of nifedipine sublingually
    - Carbonated beverage

73.4 Procedure

- Push technique and push with fragmentation technique (generally performed by specialists)
    1. First accepted endoscopic method.
    2. Gentle pressure is applied with the tip of the endoscope on the esophageal food bolus after air insufflation.
    3. If pressure does not disimpact the bolus, fragmentation can be attempted but is generally avoided owing to unknown pathology behind the food bolus.
- Foley catheter removal
    1. Moderate sedation and nasopharyngeal topical anesthesia may be used.
    2. Place the patient in a head-down Trendelenburg position.
    3. Check for symmetrical balloon inflation of the Foley catheter.
    4. Under fluoroscopy, visually pass the catheter distal to the FB.
    5. Fill the balloon slowly with 3–5 mL of saline or contrast agent.
    6. Using steady, gentle traction, withdraw the catheter with the balloon inflated distal to the FB.
    7. Grasp the object with fingers, forceps, or clamp once it is visualized in the oropharynx.
- Bougienage
    1. Topical anesthesia is recommended.
    2. Blind esophageal bougienage resembles placement of an orogastric tube.
    3. Place the patient in a sitting position.
    4. Pass a well-lubricated, appropriately sized bougie posteriorly along the roof of the mouth, following the natural curve of the soft palate caudally to the hypopharynx.
    5. Encourage the patient to swallow (to help pass the dilator through the cricopharyngeus muscle).
    6. Ask the patient to phonate to help exclude accidental laryngeal intubation.
    7. Once past the cricopharyngeus muscle, extend the head to aid the bougie in passing distally to the stomach.
    8. Post-procedure radiograph is used to confirm passage into the stomach.
- Relaxation of the LES
    1. Premedicate with an antiemetic, such as ondansetron.
    2. Administer 1–2 mg of glucagon intravenously (0.02–0.03 mg/kg in children, not to exceed 0.5 mg) with the patient in a sitting position over 1–2 min.
    3. Carbonated beverages given after glucagon ingestion have shown to have higher success rates.
    4. An alternative is to use either sublingual nitroglycerin (1–2 0.4 mg tabs) or 5–10 mg of nifedipine to relieve LES tone.
    5. This procedure does not work in patients with structural abnormalities.

73.5 Complications

- Esophageal FBs may cause esophageal pressure leading to edema, necrosis, infection, laceration, and/or perforation.
- Be cognizant of time (risk of complications is higher the longer the FB is left in place) and treatment side effects (i.e., do not continue to give patient water or glucagon if these induce vomiting).
- Aspiration and perforation during procedures listed previously.
- Late complications: esophageal stricture, abscess, mediastinitis, tracheoesophageal fistula, vascular injuries, pneumothorax, pericarditis, aspiration pneumonia, and vocal cord paralysis.

73.6 Pearls

- Esophageal foreign bodies can be lodged in the upper (proximal), middle, or lower (distal) one third:
    - Proximal: cervical web, Zenker's diverticulum
    - Middle: eosinophilic esophagitis, cancer, radiation structure, spastic dysmotility
    - Distal: peptic stricture, eosinophilic esophagitis, cancer, achalasia, esophageal diverticula, spastic dysmotility
- Because food bolus impactions are generally associated with pathology, follow-up evaluation for these abnormalities should be considered.
- Esophagus foreign bodies should not be allowed to remain in the esophagus beyond 24 h from presentation.
- Button/disc batteries in esophagus (emergent removal)
    - Considered an emergency, because liquefaction necrosis and perforation can occur rapidly.
    - Most common ingestions are hearing aid batteries.
    - If in stomach, and patient is a symptomatic, can wait up to 24 h.
- Sharp objects (emergent removal)

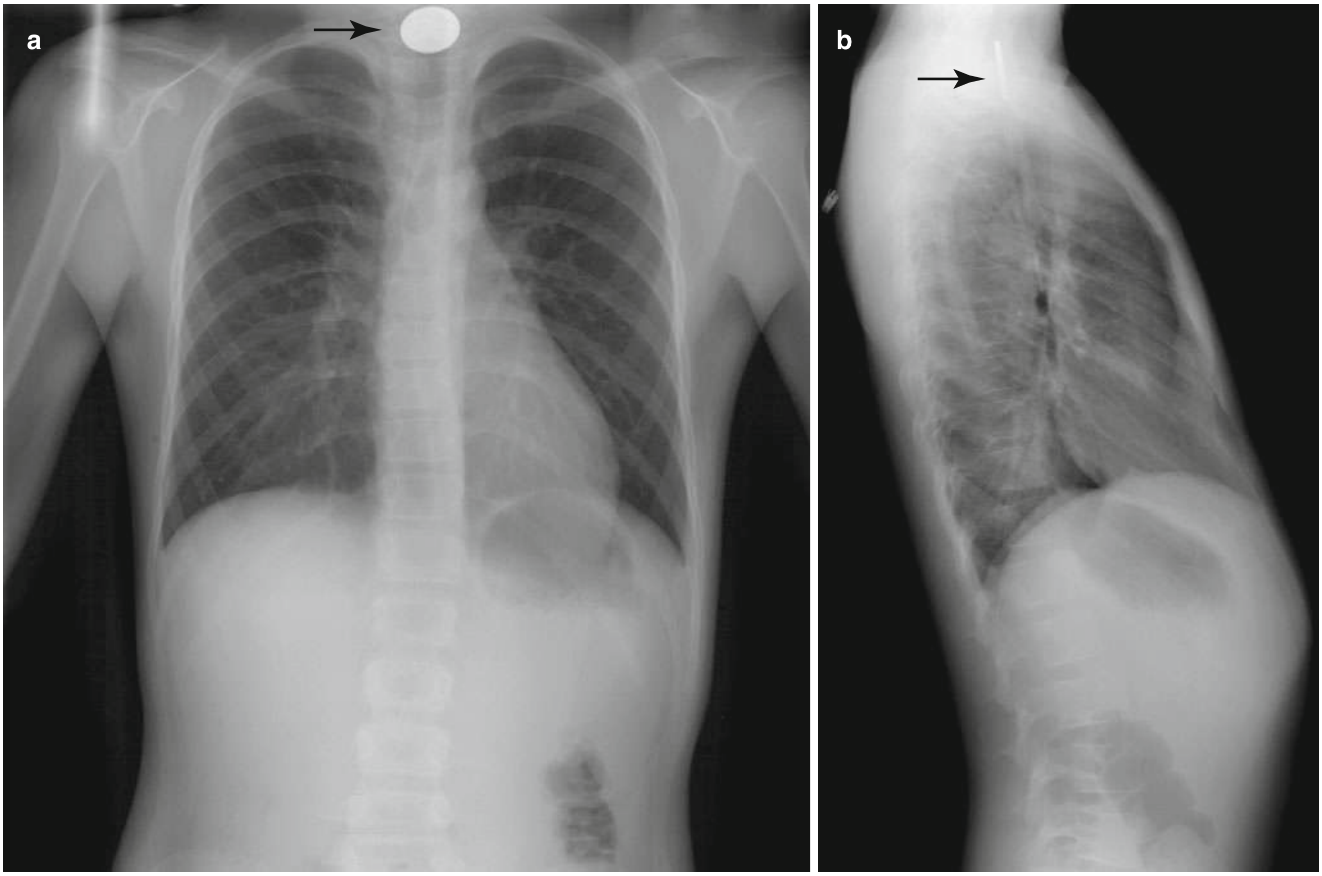

Fig. 73.1 (**a**) Anteroposterior (AP) and (**b**) lateral views demonstrating a coin in the esophagus. A coin in the trachea would present in the opposite manner—the coin would be seen on edge in the AP view and flat on the lateral view (Reproduced with permission from McGraw-Hill: Stead LG, et al. *First Aid for the Pediatrics Clerkship*. New York: McGraw-Hill, 2010)

  - Cause the majority of complications (~35 %) with esophageal FBs.
  - Direct visualization with endoscopy is the only appropriate removal technique.
- Magnets (urgent removal)
  - Can cause necrosis and fistula formation due to the way they adhere to the mucosa
- Esophageal coins (remove within 24 h) (Fig. 73.1)
  - Up to 80 % of coins at the LES will pass spontaneously within 24 h without interventions. The watchful waiting approach is used only in patients with single coins and who are asymptomatic.
  - Common complications of these procedures include mild bleeding, lip laceration, bradycardia with Foley catheter insertion, and teeth injuries.
  - Some protocols include RSI as part of the management process and should be considered if lifesaving.

Selected Reading

ASGE Standards of Practice Committee. Management of ingested foreign bodies and food impactions. Gastrointest Endosc. 2011; 73:1085.

Bhargava R, Brown L. Esophageal coin removal by emergency physicians: a continuous quality improvement project incorporating rapid sequence intubation. CJEM. 2011;13:28–33.

Conway WC, Sugawa C, Ono H, Lucas CE. Upper GI foreign body: an adult emergency hospital experience. Surg Endosc. 2007; 21:455–60.

Katsinelos P, Kountouras J, Paroutoglou G, et al. Endoscopic techniques and management of foreign body ingestion and food bolus impaction in the upper gastrointestinal tract: a retrospective analysis of 139 cases. J Clin Gastroenterol. 2006;40:784–9.

Activated Charcoal

74

Deylin I. Negron Smida and Judith K. Lucas

74.1 Indications

- Single-dose activated charcoal (AC) (Fig. 74.1)
  - Does not meet criteria for gastric emptying.
  - Gastric emptying may be too harmful.
  - Ingestion of toxic xenobiotic is known to be adsorbed by AC.
  - Ingestion occurred with a time frame amenable to adsorption by AC, or clinical factors are present that suggest that not all of the xenobiotic had already been systemically absorbed.
  - Ingestion of extended- or sustained-release formulations.
- Multiple-dose activated charcoal therapy (MDAC)
  - Life-threatening ingestion of:
    - Carbamazepine
    - Phenobarbital
    - Quinine
    - Theophylline
    - Dapsone
  - Life-threatening ingestion of another xenobiotic that undergoes enterohepatic recirculation and is adsorbed to AC
  - Ingestion of a significant amount of a slowly released xenobiotic
  - Ingestion of a xenobiotic known to form concretions or bezoars, such as aspirin

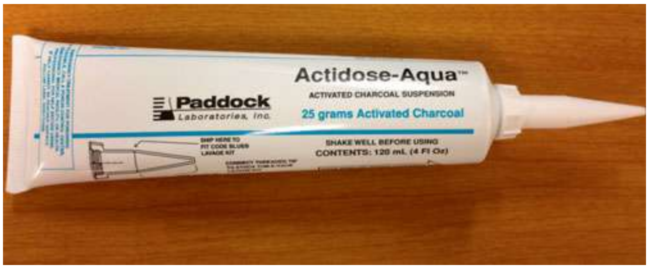
Actidose-Aqua™
Paddock
Laboratories, Inc.
ACTIVATED CHARCOAL SUSPENSION
25 grams Activated Charcoal
SHAKE WELL BEFORE USING
CONTENTS: 120 mL (4 Fl Oz)

Fig. 74.1 Activated charcoal (AC)

D.I.N. Smida, MD
Department of Emergency Medicine, University of Pittsburgh
Medical Center, Saint Margaret Hospital, Pittsburgh, PA, USA
e-mail: Deylin.negron@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_74

74.2 Contraindications

- Absolute
  - Gastric perforation
  - Gastrointestinal ileus, obstruction, or diminished peristalsis
  - Nonintubated patients with the potential of losing protective airway reflexes
  - Intestinal obstruction
  - Ingestion of:
    - Corrosives
    - Petroleum distillates
- Relative
  - Altered or decreased level of consciousness unless intubated.
  - Vomiting.
  - Xenobiotic has limited toxicity at almost any dose.
  - Dose ingested is less than the dose expected to produce significant illness.
  - Presentation many hours after ingestion.
  - Minimal signs or symptoms of poisoning.
  - Ingested xenobiotic has a highly efficient antidote.
  - Administration of charcoal may increase the risk of aspiration (i.e., hydrocarbons).

74.3 Materials and Medications

- Nasogastric (NG) tube/orogastric (OG) tube (Fig. 74.2)
- Baby bottle with split nipple (designed for drinking of slurry solutions, such as thickened formulas) or sippy cup without the valve
- Absorbent pad
- Basin
- Water-soluble lubricant
- Tubing connected to suction device
- Flavored syrup

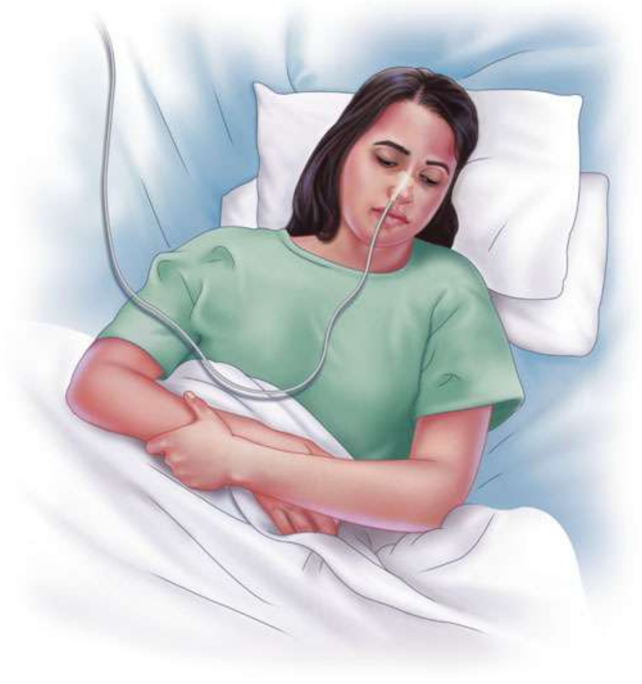

Fig. 74.2 Drinking AC (by cup, sippy cup, or bottle) is effective, but it may also be infused via nasogastric or orogastric tube

74.4 Procedure

- Single-dose administration
  - Adult
    - Can be taken via a cup and straw (drunk) if the patient is cooperative.
    - The optimal dose of AC is unknown.
    - 50–100 g/dose (1 g/kg), administered at a rate no less than 12.5 g/h or its equivalent.
    - If vomiting is anticipated, an intravenous antiemetic is recommended.
  - Children
    - 1 g/kg or 10:1 ratio of AC to drug ingested.
    - After massive ingestion, give 2 g/kg.
    - Many children will drink the suspension from a bottle or sippy cup, especially if it is mixed with juice or flavored syrup (e.g., chocolate or strawberry).
- MDAC
  - Adults: 0.5–1 g/kg every 2–4 h for 24–48 h
  - Children (<12 y old): 0.25–0.5 g/kg every 2–4 h or rate of 0.2 g/kg/h for 24–48 h
- Mixtures
  - Ready to drink
  - Powder form
    - Add eight parts water to the selected powdered form.
    - Gatorade or juices can also be used to help hide the flavor and texture.
    - In children, the AC can be mixed with cold chocolate or some other flavored syrup, which also hides the flavor.
  - Shake liquid suspension well for 1 min.
  - If the patient vomits, the dose should be repeated. Smaller, more frequent, dosing may be better tolerated, and an antiemetic may be needed.

74.5 Complications

- Aspiration pneumonitis
- Transient constipation
- Intestinal bezoars
- Bowel obstruction
- Diarrhea, dehydration, hypermagnesemia, and hypernatremia with coadministered cathartics or MDAC
- Vomiting
- Corneal abrasion if spilled in the eyes

74.6 Pearls and Pitfalls

- Pearls
  - If an OG or NG tube is used, time should be allowed for the last dose to pass through the stomach before the tube is removed. Suctioning the tube before removal may prevent subsequent AC aspiration.
  - With children, the colder and sweeter the solution and if the color is camouflaged (cup with a lid or a sippy cup), there will be increased success at oral administration (vs NG/OG).
- Pitfalls
  - No evidence-based literature supports the assertion that AC changes clinical outcome.
  - Xenobiotics and AC adsorption (Table 74.1).
  - Incorrect application (e.g., into the lungs) results in pulmonary aspiration, which can be fatal if unrecognized.
    - Incorrect placement of NG/OG tube into trachea.
    - Administration of AC to a patient with an ileus (e.g., in anticholinergic overdoses).
  - No specific contraindication for AC in pregnant women; however, diarrhea or hypernatremia in the mother may adversely affect the fetus.

Table 74.1 Absorption of xenobiotics by AC

Good absorption	Poor absorption
Acetaminophen	Alkali
Bupropion	Chlorpropamide
Caffeine	Doxepin
Carbamazepine	Ethanol or other alcohols
Chlordecone	Ethylene glycol
Dapsone	Fluoride
Digitoxin	Heavy metals
Nadolol	Imipramine
Phenobarbital	Inorganic salts
Phenylbutazone	Iron
Phenytoin	Lithium
Salicylate	Methotrexate
Theophylline	Mineral acids
	Potassium
	Tobramycin
	Valproate sodium
	Vancomycin

Selected Reading

American Academy of Clinical Toxicology; European Association of Poisons Centres and Clinical Toxicologists. Position statement and practice guidelines on the use of multi-dose activated charcoal in the treatment of acute poisoning. J Toxicol Clin Toxicol. 1999;37:731–51.

Chyka PA, Seger D. American Academy of Clinical Toxicology; European Association of Poisons Centres and Clinical Toxicologists. Position statement: single-dose activated charcoal. J Toxicol Clin Toxicol. 1997;35:721–41.

Gude A, Hoegberg LCG. Techniques to prevent gastrointestinal absorption. In: Nelson LS, Lewin NA, Howland MA, et al., editors. Goldfrank's toxicologic emergencies. 9th ed. New York: McGraw-Hill; 2011. p. 93–7,431.

Lie D. Use of activated charcoal in drug overdose. Medscape family medicine. 25 Mar 2004. www.medscape.com/viewarticle/471331

Olson KR. Emergency evaluation and treatment. In: Olson KR, Anderson IB, Benowitz NL, et al., editors. Poisoning and drug overdose. 5th ed. New York: McGraw-Hill; 2007. p. 1–56.

Gastric Lavage

75

Deylin I. Negron Smida and Judith K. Lucas

75.1 Indications

- Recent ingestion (<30–60 min).
- Life-threatening exposure where there is a high suspicion that a xenobiotic is still present in the stomach and evacuation is expected to contribute to an improved outcome (e.g., iron, tricyclic antidepressants).
- Ingested agent is not absorbed with activated charcoal (e.g., pesticides, hydrocarbons, iron, alcohols, lithium, and solvents).
- Activated charcoal is unavailable.
- Ingestion exceeds adsorptive capacity of initial activated charcoal dosing (e.g., >100 mg/kg of pills).
- Ingestion of an agent likely to form a durable mass or bezoars after overdose.

75.2 Contraindications

- Vomiting
- Unintubated patients with potential to lose airway protective reflexes
- Ingestion of a xenobiotic with aspiration potential (e.g., hydrocarbon) without intubation
- Ingestion of caustic substances (alkali or acidic)
- Ingestion of sharp metals
- Ingestion of a foreign body (e.g., drug packet)
- Risk for hemorrhagic gastrointestinal perforation
- Ingestion of xenobiotic in a form known to be too large to fit into the lumen of the orogastric tube
- Nontoxic ingestions

75.3 Materials and Medications

- Orogastric tube (Ewald tube or the Tum-E-Vac) (Fig. 75.1)
  - Adults and adolescents: 36–40 French
  - Children: 22–28 French
- Pen or tape to mark the length of the tube
- Water-soluble lubricant
- Suction
- Emesis basin
- Absorbent pad
- Catheter-tip syringe with 2 mL water/saline to check position of the tube
- Room temperature irrigation fluid
- Bite block or oral airway to prevent patients from biting down on the tube

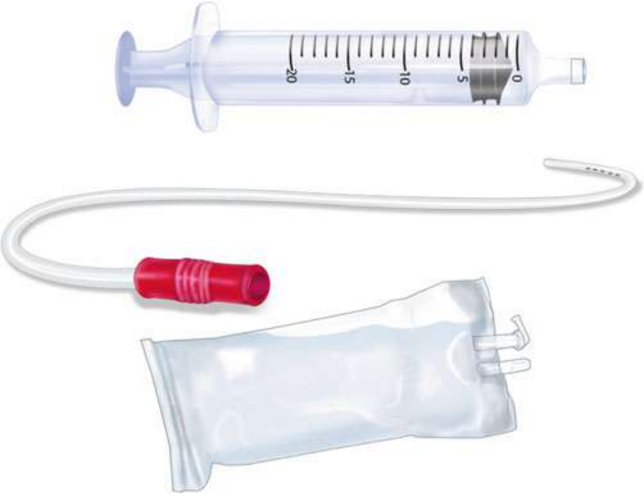

Fig. 75.1 Materials needed for gastric lavage include a large-bore nasogastric tube, a 60-cc non-Luer-Lok syringe, and a solution, typically normal saline, for lavage

---

D.I.N. Smida, MD
Department of Emergency Medicine, University of Pittsburgh
Medical Center, Saint Margaret Hospital, Pittsburgh, PA, USA
e-mail: Deylin.negron@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_75

75.4 Procedure

1. If there is potential airway compromise, endotracheal or nasotracheal intubation should precede orogastric lavage.
2. Place an oral airway or a bite block to prevent biting of the endotracheal tube if the patient recovers consciousness or has convulsions during the procedure.
3. Ensure suction apparatus is available and functioning.
4. Place the patient in an upright-seated position if awake and alert.
5. Place patient in the left lateral decubitus position if obtunded.
6. Before insertion, the proper length of tubing to be passed should be measured from the mouth, back to the ear, and down anterior to the chest and abdomen, beyond the point where any side ports on the tube would be beyond the level of the estimated lower esophageal sphincter (Figs. 75.2 and 75.3).
7. If the patient is still awake, insert the gastric tube to the level of the glottis, and encourage the patient to swallow.
8. Pass the tube to the stomach.
    (a) Coughing, airflow, or fog from the tube raises the concern for inadvertent tracheal positioning.
9. After the tube is inserted, it is essential to confirm that the distal end of the tube is in the stomach, by "popping" 5–10 mL of air into the tube while someone is listening with a stethoscope over the stomach.

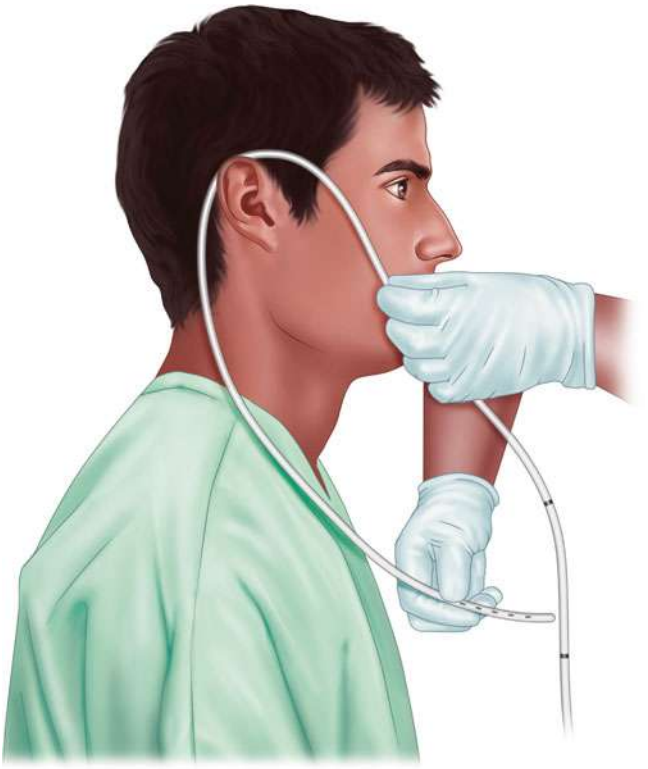

Fig. 75.3 Measuring correct placement of the tube. Place the distal tip over the stomach and wrap the tube up behind the ear (usually the right ear because the tubes generally pass easier through the right nares) and around the nares. The black line or centimeter mark at the level of the nares is the point of insertion when passage stops

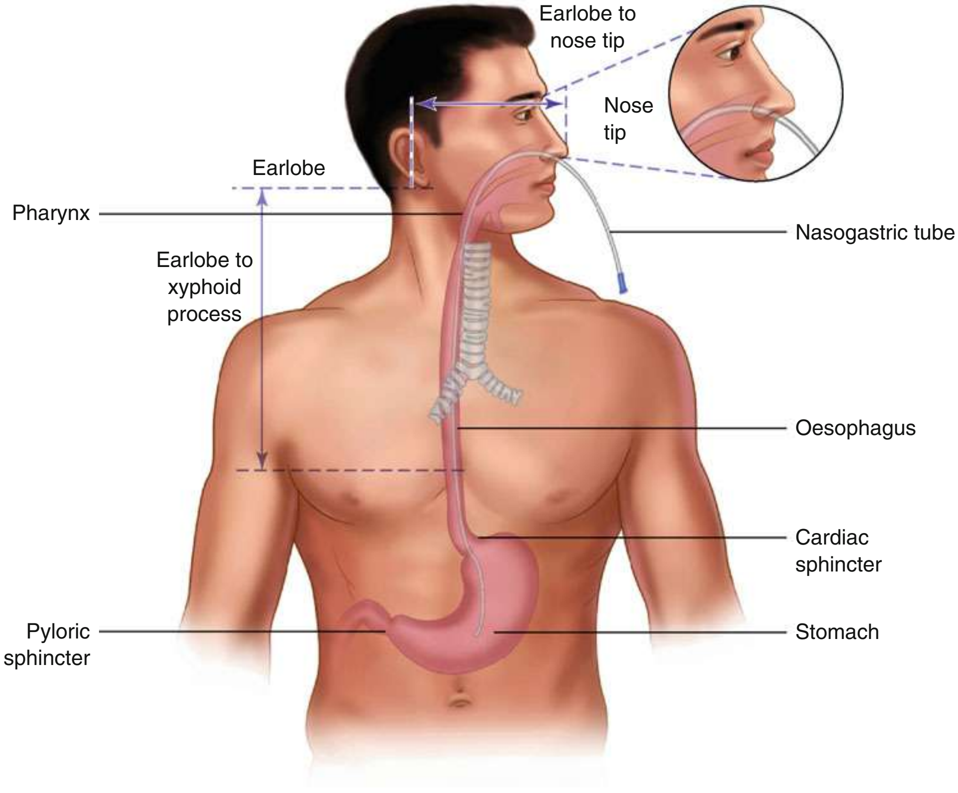

Fig. 75.2 Diagram illustrates appropriate placement of the lavage tube

10. In adults, 250-mL aliquots of a room temperature saline lavage solution are instilled via a funnel or lavage syringe. In children, aliquots should be 10–15 mL/kg to a maximum of 250 mL and suctioned back out of the tube attached to low to moderate continuous wall suction. Instillation of lavage solution and suction is repeated (Fig. 75.4).
11. Orogastric lavage should continue for at least several liters in an adult and/or at least 0.5–1 L in a child if the return is free of debris or until no particulate matter returns and the effluent lavage solution is clear.
12. Those caring for the patient must remain protected at all times, using goggles, mask, gown, and gloves. If the ingested poison is toxic via pulmonary or skin absorption, isolate the ingestant immediately in a self-contained wall suction unit.
13. Any material still in the stomach should be withdrawn, and immediate instillation of the activated charcoal should be considered for large ingestions of xenobiotics known to be adsorbed by activated charcoal.

75.5 Complications

- Vomiting
- Esophageal tears or perforation after orogastric tube insertion
- Inadvertent tracheal intubation and/or airway trauma
- Aspiration pneumonitis

75.6 Pearls and Pitfalls

- Pearls
  - You must use a large-bore orogastric tube for maximal efficacy.
  - The left lateral decubitus position is recommended because the pylorus points upward in this orientation. This position theoretically helps prevent the xenobiotic from passing through the pylorus during the procedure.
- Pitfalls
  - Large drug packets, adherent masses of pills, and plant and mushroom fragments will not pass through a 40-French lavage tube.

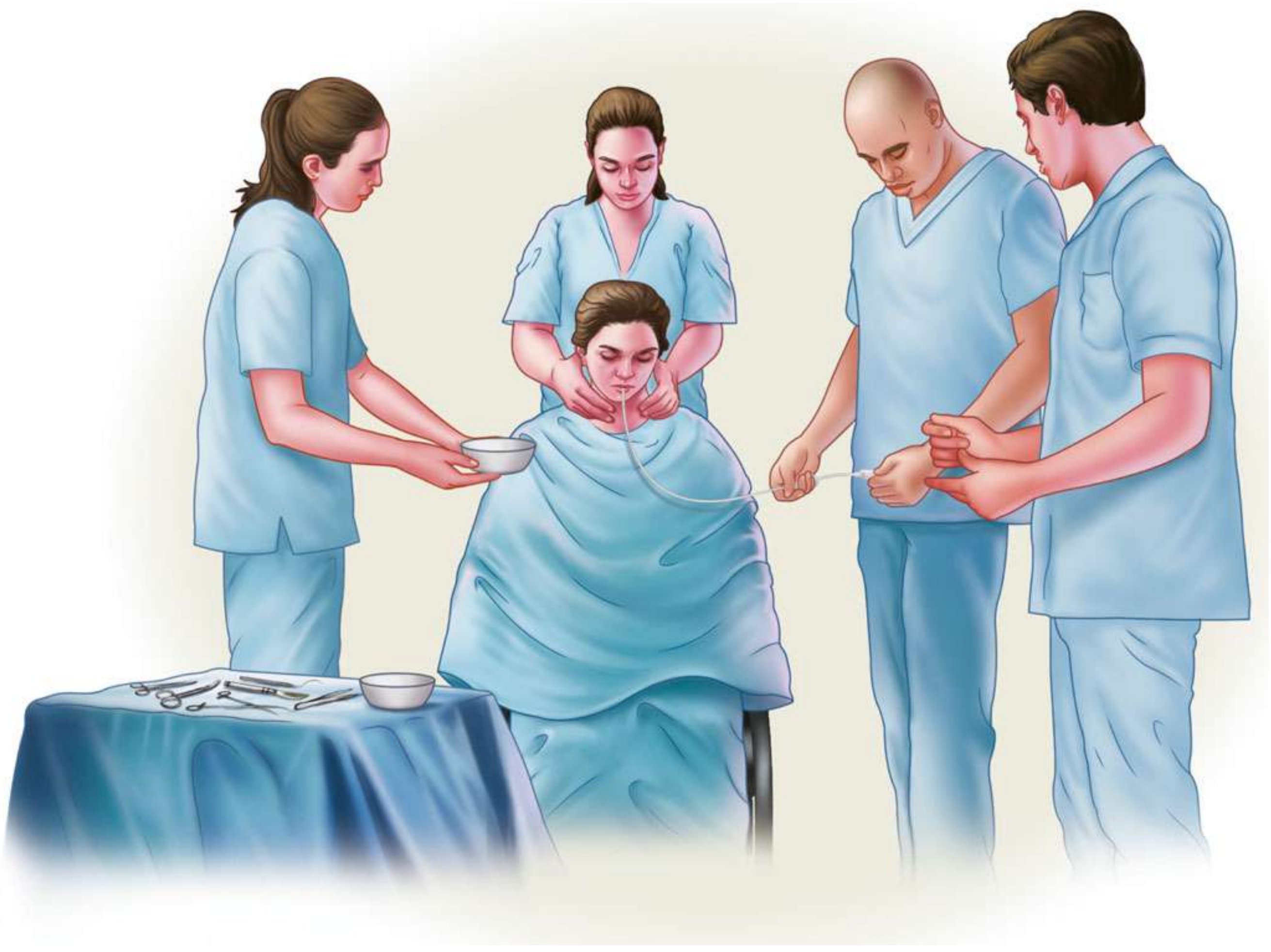

Fig. 75.4 Lavage in progress

Selected Reading

Gude A, Hoegberg LCG. Techniques to prevent gastrointestinal absorption. In: Nelson LS, Lewin NA, Howland MA, et al., editors. Goldfrank's toxicologic emergencies. 8th ed. New York: McGraw-Hill; 2006. p. 91–3.

Olson KR. Poisoning & drug overdose. In: Olson KR, Anderson IB, Benowitz NL, et al., editors. Emergency evaluation and treatment. 6th ed. New York: McGraw-Hill; 2012.

Smilktein MJ. Techniques used to prevent gastrointestinal absorption of toxic compounds. In: Nelson LS, Lewin NA, Howland MA, et al., editors. Goldfrank's toxicologic emergencies. 7th ed. New York: McGraw-Hill; 2002. p. 46–8.

Whole-Bowel Irrigation

76

Judith K. Lucas

76.1 Indications

- Whole-bowel irrigation (WBI) should not be used routinely in the management of the poisoned patient (because there is no clinical proof it will change clinical outcome).
- Ingestion of significant amount of medications.
  - Not adsorbed by activated charcoal
    - Lead, lithium, arsenic, and zinc
    - Substantial amounts of iron (high morbidity and no other effective method to gastrointestinal decontamination)
  - Sustained-release medications or enteric-coated drugs
  - Disk batteries distal to the pylorus
  - Whole transdermal patches (fentanyl, clonidine, nicotine)
  - Drug concretions
  - Ingested packets of illicit drugs

76.2 Contraindications

- Absolute
  - Bowel obstruction
  - Bowel perforation
  - Ileus
  - Hemodynamic instability
  - Compromised or unprotected airway
  - Intractable vomiting
- Relative
  - Concurrent or recent administration of activated charcoal (may decrease the effectiveness of activated charcoal)

76.3 Materials and Medications

- Topical anesthesia, although not mandatory, will reduce the pain of nasogastric (NG) tube placement.
  - 10 % lidocaine spray
  - Lidocaine gel
- Small-bore (12-French) NG tube (Fig. 76.1).
- Tape for securing the NG tube.
- Reservoir or feeding bag used for NG tube feedings (Fig. 76.2).
- Intravenous pole.
- Bedside commode or toilet (Fig. 76.3).
- Polyethylene glycol-electrolyte solution (PEG-ES) (Fig. 76.4).
- Antiemetic.
  - No absolute indication for prophylactic use
  - May be helpful if vomiting ensues during infusion
    - Metoclopramide
      - Antiemetic
      - *Increases gastric motility*

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_76

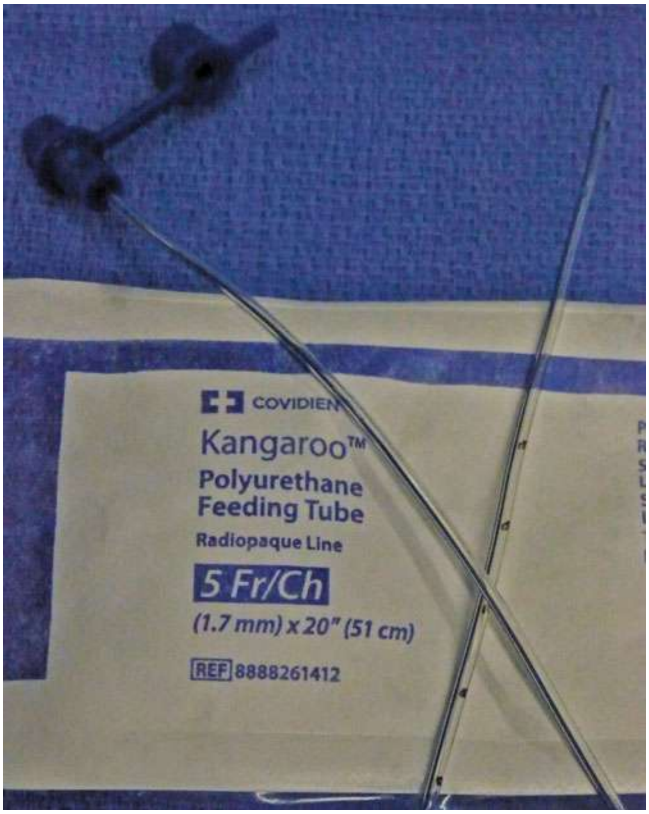

Fig. 76.1 Nasogastric (NG) tube. Typically, the infusion of the lavage solution is too rapid to be taken orally, so an NG tube can be placed. Since the irrigation solution is of low viscosity, a small-bore NG tube should be used for comfort

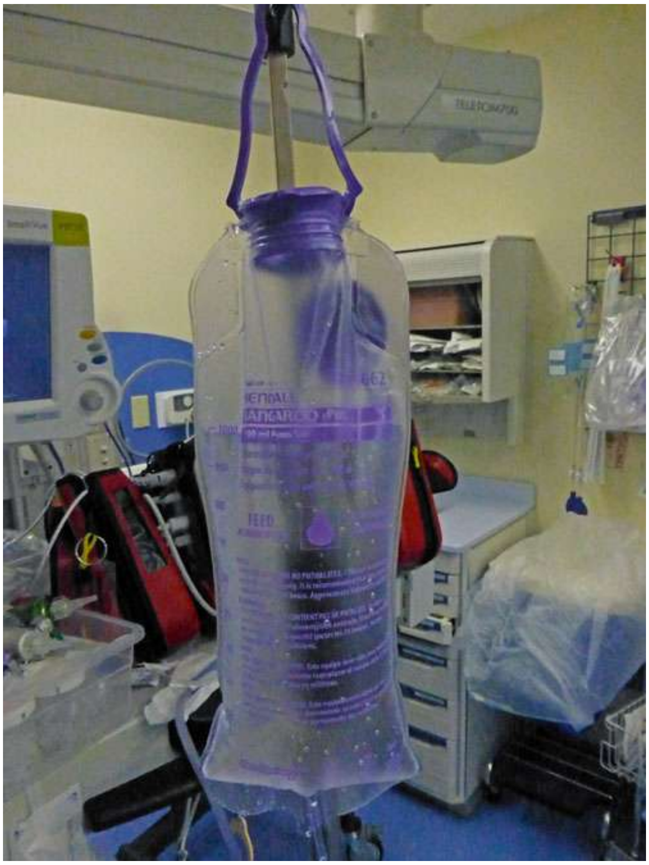

Fig. 76.2 Bag from which the lavage solution will drain; it is similar to the bags used for gastrostomy tube feeding

Fig. 76.3 Almost always, the patient will need to be seated on or very near a portable commode, as once the irrigation solution starts to move through the bowels, defecation will occur rapidly

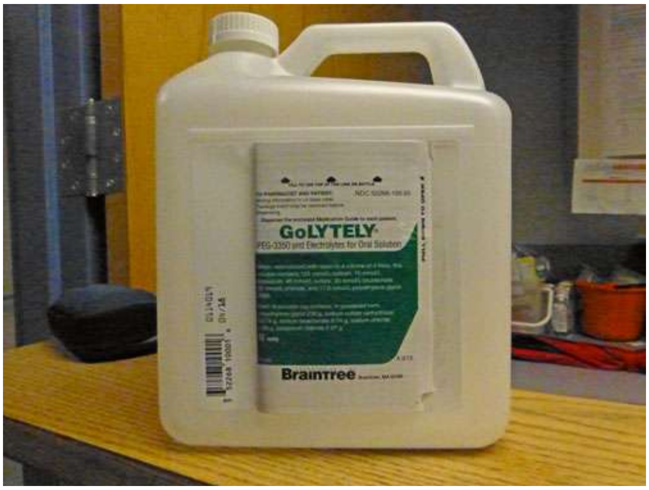

Fig. 76.4 Example brands of intestinal irrigation electrolyte solutions

76.4 Procedure

1. An NG tube is required because most patients will not drink the PEG-ES at the necessary rate.
2. Place a small-bore (12-French) NG tube to a sufficient distance that the tip lies in the central portion of the stomach.
3. Confirm NG placement with a radiograph.
4. Attach the tube to the reservoir bag of PEG-ES and hang from an elevated site (an extended intravenous pole).
5. The patient should be seated in an upright position.
    - Promotes settling of the intoxicant in the distal portion of the stomach
    - Decreases the likelihood of vomiting
6. Dosing:
    - Children 9 months to 6 years: 500 mL/h
    - Children 6–12 years: 1,000 mL/h
    - Adolescents/adults: 1,500–2,000 mL/h
7. Collect effluent.
8. Continue infusion.
    - Until the rectal effluent is the same color as the influent (i.e., clear), usually between 4 and 6 h.
    - You may continue beyond clear effluent if clinical evidence indicates ongoing effectiveness:
        - Continued pill fragments or drug packets are present in the effluent.
        - Radiographic evidence that pills, pharmacobezoars, or packets are still present.

76.5 Complications

- Nausea, vomiting, and bloating
- Misplacement of the NG tube
- Esophageal perforation owing to NG tube placement
- Aspiration pneumonitis in the unprotected airway

76.6 Pearls

- Overall, WBI is probably more effective than gastric lavage, but probably less effective than activated charcoal in preventing poison absorption (when the intoxicant can be adsorbed to charcoal).
- Vomiting.
    - Usually secondary to the ingestant (i.e., emetogenic toxins, such as iron)
    - May be due to rate of infusion
        - Slow rate by 50 % for 30–60 min.
        - Then return to original rate.
- If resistance is encountered during NG tube placement, do not force passage. Remove and redirect.

Selected Reading

Bailey B. To decontaminate or not to decontaminate? the balance between potential risks and foreseeable benefits. Clin Pediatr Emerg Med. 2008;9:17–23.

Hanhan UA. The poisoned child in the pediatric intensive care unit. Pediatr Clin North Am. 2008;55:669–86. xi.

Lheureux P, Tenenbein M. Position paper: whole bowel irrigation. American Academy of Clinical Toxicology/European Association of Poison Centres and Clinical Toxicologists. J Toxicol Clin Toxicol. 2004;42:843–54.

Othong R. Whole-bowel irrigation. MedScape Reference: drugs, diseases, and procedures. Updated: Aug 2011

Postuma R. Whole bowel irrigation in pediatric patients. J Pediatr Surg. 1982;17:350–2.

Sengstaken-Blakemore Tube

77

Thomas T. Nguyen, Etan Eitches,
and Stephanie Wetmore-Nguyen

77.1 Indications

- Life-threatening esophageal variceal bleed refractory to endoscopy and medical therapy
- Life-threatening esophageal variceal bleed refractory to medical therapy in the absence of possible endoscopy

77.2 Contraindications

- Absolute
  - Known esophageal rupture
  - Unable to intubate or maintain airway
- Relative
  - History of prior esophageal trauma or strictures
  - Recent surgery of the gastroesophageal junction
  - Resolved or resolving variceal bleeding

77.3 Materials and Medications

- Sengstaken-Blakemore (SB) tube (Fig. 77.1)
- 60-mL syringe with catheter tip (Fig. 77.2)
- Sphygmomanometer (Fig. 77.3) or cuffalator (Fig. 77.4)
- Y-Tube connector (Fig. 77.5) or 3-way stop-valve connector (Fig. 77.6)
- Vacuum suction device and tubing (Fig. 77.7)
- Tube clamps (4) (Fig. 77.8)
- Lubricant (water soluble)
- Lidocaine (Xylocaine) spray or gel
- Anchoring device such as a football helmet or catcher's mask (Fig. 77.9)
- Cup of water and straw if the patient is awake
- Scissors (Fig. 77.10)
- Intubation equipment
- Sterile water

T.T. Nguyen, MD (✉)
Department of Emergency Medicine,
Mount Sinai Beth Israel, New York, NY, USA
e-mail: tnguyen@chpnet.org

E. Eitches, MD
Department of Emergency Medicine,
Beth Israel Medical Center, New York, NY, USA
e-mail: eeitches@gmail.com

S. Wetmore-Nguyen, MD
Department of Emergency Medicine,
New York-Presbyterian Hospital/Columbia University Medical Center,
New York, NY, USA
e-mail: slw9004@nyp.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_77

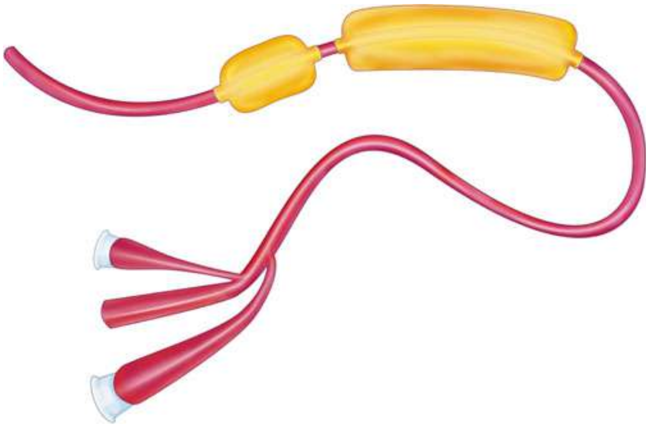

Fig. 77.1 Sengstaken-Blakemore (SB) tube

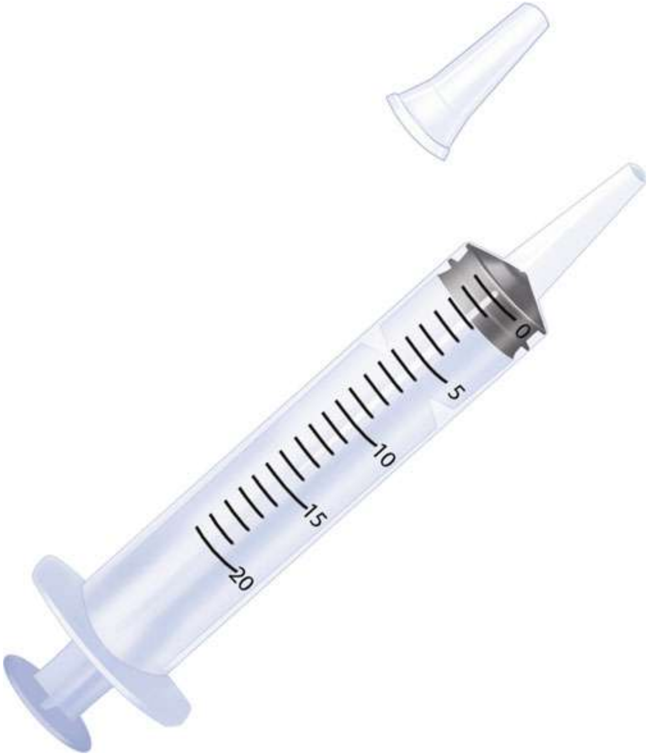

Fig. 77.2 Syringe

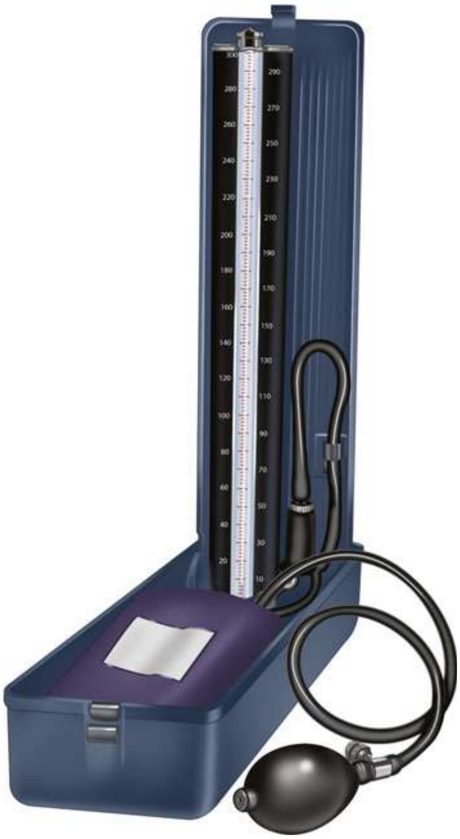

Fig. 77.3 Sphygmomanometer

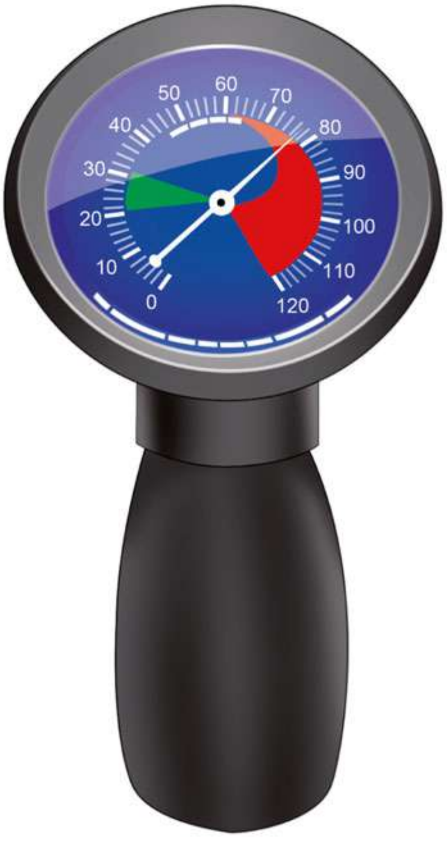
0
10
20
30
40
50
60
70
80
90
100
110
120

Fig. 77.4 Cuffalator

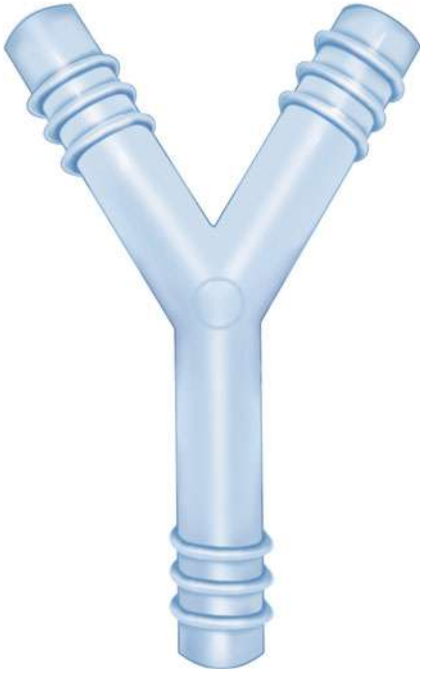

Fig. 77.5 Y-Tube connector

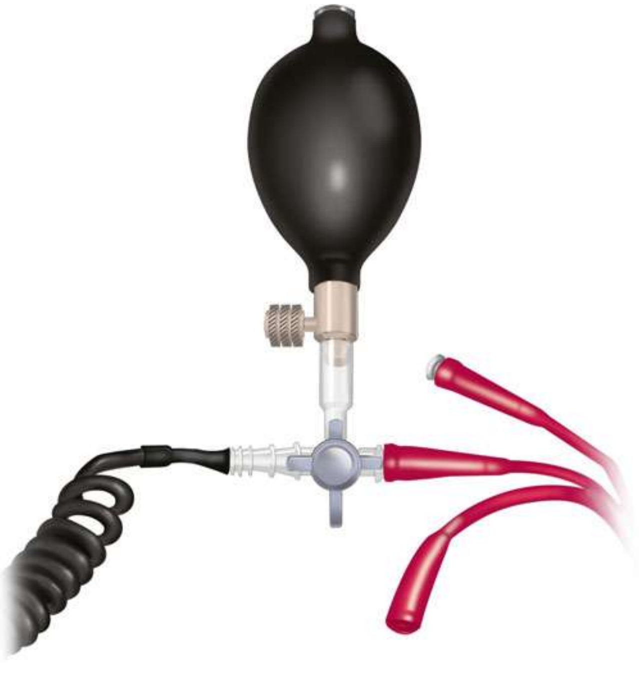

Fig. 77.6 3-way stop-valve tube connector

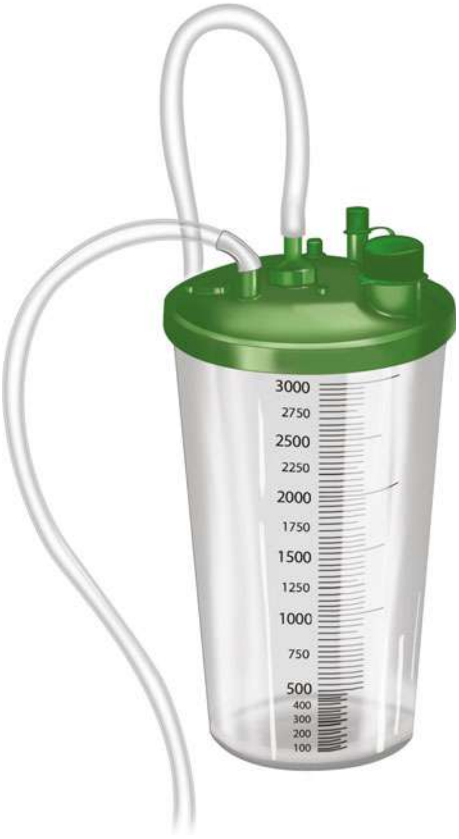
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
400
300
200
100

Fig. 77.7 Suction device and tubing

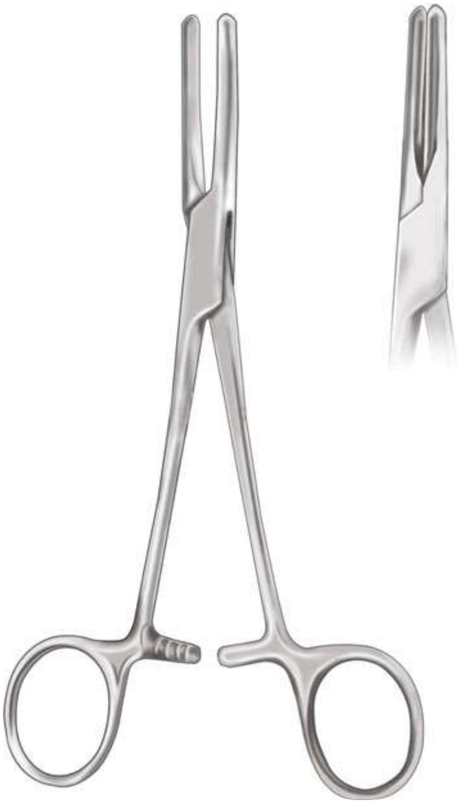

Fig. 77.8 Tube clamps

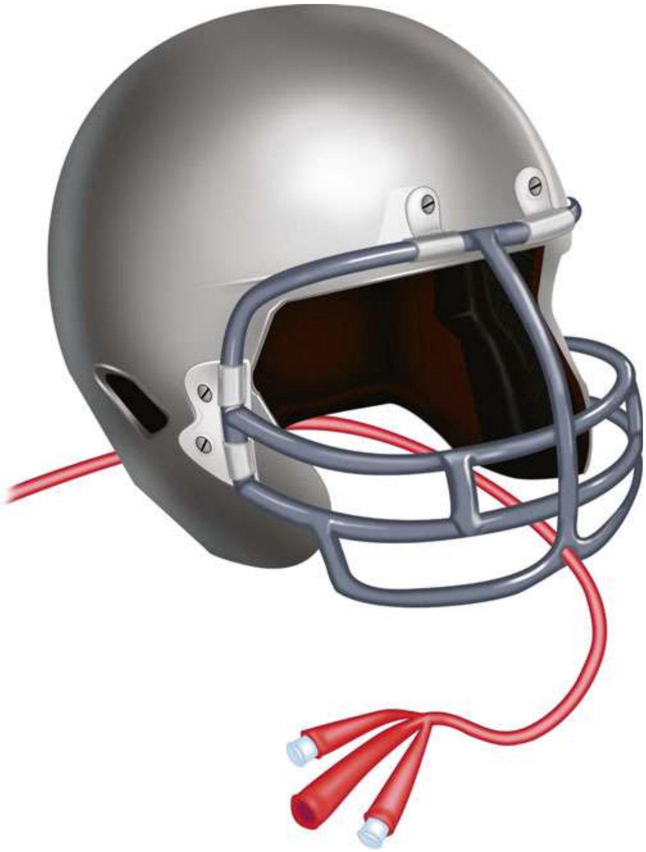

Fig. 77.9 Helmet traction setup

Fig. 77.10 Scissors

77.4 Procedure

1. Sedate and/or intubate the patient for adequate control of the patient during the procedure.
2. Ensure that the SB tube balloons are functional by inflating and deflating the balloons to ensure the absence of leaks.
3. Perform gastric lavage and irrigate the stomach with copious amount of sterile water.
4. Coat the distal and proximal portions of the SB tube with a thin layer of lubricating jelly or lidocaine gel. Spray the nasal passage with lidocaine spray.
5. Pass the SB tube via the nasogastric (NG) or the orogastric route (in intubated patients) to the 50-cm line. You may confirm placement with x-ray.
6. Inflate the gastric balloon to 200 mL of air and clamp the tube.
7. Apply gentle traction of 1–2 lb (0.4–2 kg) of force until it is felt that the gastric balloon has lodged at the gastroesophageal junction (Fig. 77.11).
8. Secure the tube to an anchor (e.g., football helmet or catcher's mask) placed on the patient's head.
9. Aspirate and lavage the gastric aspiration port. If it is clear of blood, do not inflate the esophageal balloon.
10. If it is not clear of blood, connect the esophageal tube of the SB tube to the sphygmomanometer/cuffalator using the 3-way stop-valve device (Fig. 77.12). You may use the Y-Tube connector instead.
11. Inflate the esophageal balloon to the lowest pressure determined to stop bleeding, typically 20–45 mmHg. Clamp the balloon.
12. Place an NG tube until it is felt overlying the top of the esophageal balloon of the SB tube. Check for further proximal esophageal bleed through aspiration and gentle lavage. Attach this NG tube to intermittent section to aid in the clearance of secretions.
13. Obtain a portable radiograph to confirm the position of the SB tube.
14. The esophageal tube should be at the lowest pressure that prevents bleeding and kept inflated for 24 h or until other definitive treatment is obtained.

Fig. 77.11 Proper placement of the balloons

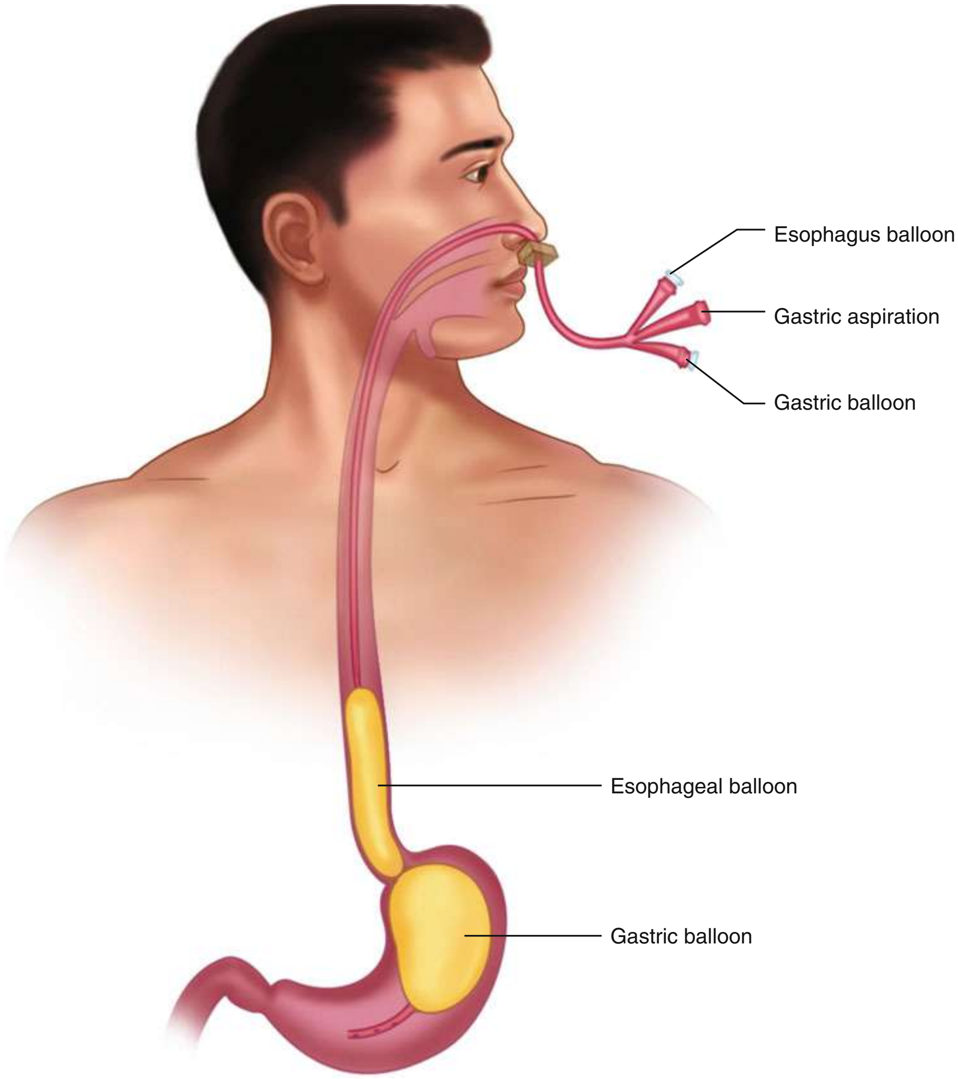
Esophagus balloon
Gastric aspiration
Gastric balloon
Esophageal balloon
Gastric balloon

Fig. 77.12 Connection of the sphygmomanometer to the esophagus balloon port

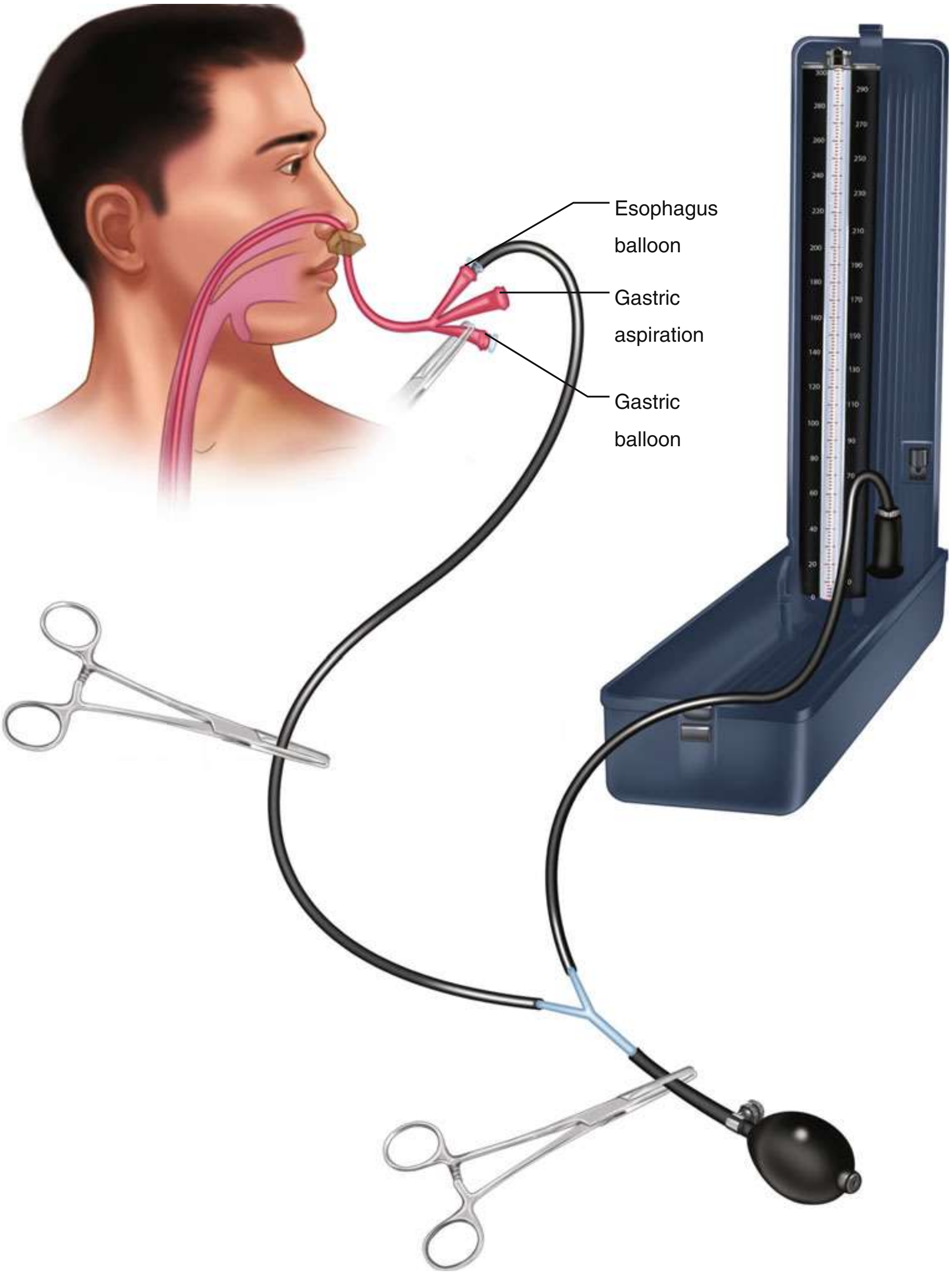
Esophagus balloon
Gastric aspiration
Gastric balloon

77.5 Complications

- Esophageal rupture occurs owing to esophageal erosion and necrosis owing to a balloon tamponade effect on tissue perfusion or overzealous balloon inflation.
- Airway obstruction owing to gastric tube deflation or failure, allowing esophageal tube to move up and occlude airway. Keep scissors near the patient to cut the SB tube lumens and remove the tube as necessary.
- Regurgitation and aspiration pneumonia from failure to adequately suction oropharyngeal secretions.

77.6 Pearls and Pitfalls

- Pearls
  - The esophageal balloon should not be inflated if the gastric balloon alone stops the bleeding.
  - Never inflate the esophageal balloon without the inflating the gastric balloon first. This will prevent it from slipping proximally into the oropharynx and obstructing the airway.
  - Nausea, vomiting, or aspiration is highly likely to occur. Use antiemetics and lavage the stomach before the procedure.
  - Intubate if there is airway compromise or risk of aspiration into the lungs.
  - Inflate the esophageal balloon only to the minimum pressure necessary to stop the variceal bleeding.
  - Using a catcher's mask may be more practical and comfortable for the recumbent patient.
- Pitfalls
  - The SB tube may induce hiccups.

Selected Reading

Bauer J, Kreel I, Kark A. The use of the Sengstaken-Blakemore tube for immediate control of bleeding esophageal varices. Ann Surg. 1974;179:273–7.

Henneman PL. Gastrointestinal bleeding. In: Rosen P, Barkin RM, editors. Emergency medicine. 6th ed. St. Louis: Mosby; 1998.

Remonda G, Morachioli N, Petruzzelli C. The use of the Sengstaken-Blakemore tube for immediate control of bleeding esophageal varices. Ann Osp Maria Vittoria Torino. 1981;24:115–20.

Sengstaken RW, Blakemore AH. Balloon tamponage for the control of hemorrhage from esophageal varices. Ann Surg. 1950;131:781–9.

Treger R, Graham T, Dea S. Sengstaken-Blakemore tube. Available at http://emedicine.medscape.com/article/81020-overview#a01. Accessed 18 May 2014.

Gastrostomy Tube Placement

78

Nathaniel Lisenbee and Latha Ganti

There are several types of gastrostomy tubes and related procedure variations:

- Procedure variations:
  - Open gastrostomy tube (G tube)
  - Percutaneous endoscopic gastrostomy (PEG) tube
  - Laparoscopic G tube
- Tube types:
  - PEG tube (Fig. 78.1)
  - Malecot tube (Fig. 78.2)
  - Balloon G tube (Fig. 78.3)
  - Low-profile G tube (nonobturated, button) (Fig. 78.4)
  - Low-profile G tube (obturated, button) (Fig. 78.5)

N. Lisenbee, MD
U. S. Air Force Medical Center Keesler, Biloxi, MI, USA
e-mail: nate4085@ufl.edu

L. Ganti, MD (✉)
Professor of Emergency Medicine, University of Central Florida, Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_78

Fig. 78.1 Percutaneous endoscopic gastrostomy (PEG) tube

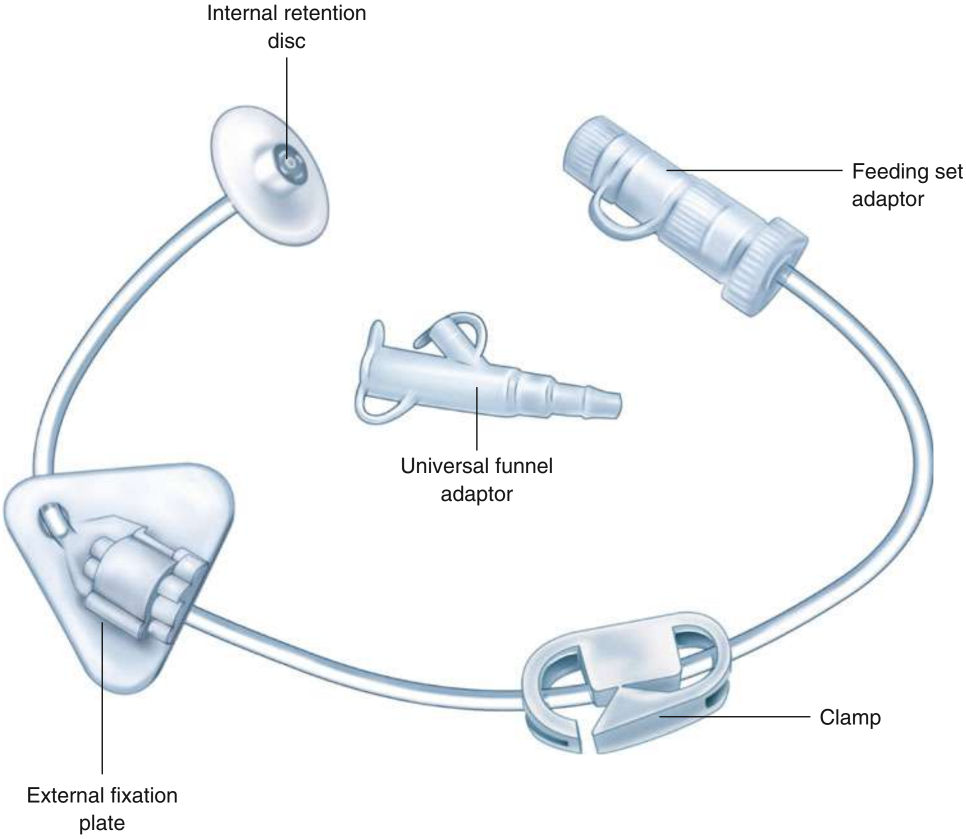
Internal retention disc
Feeding set adaptor
Universal funnel adaptor
Clamp
External fixation plate

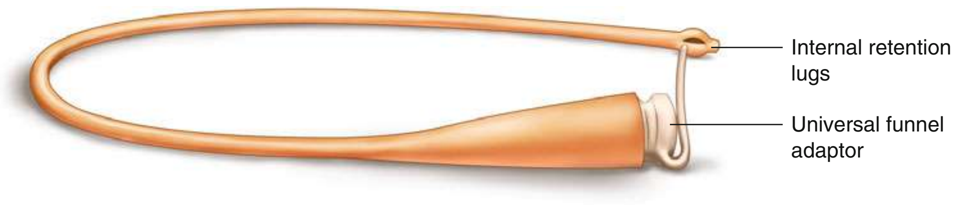
Internal retention lugs
Universal funnel adaptor

Fig. 78.2 Malecot tube

Fig. 78.3 Balloon gastrostomy tube

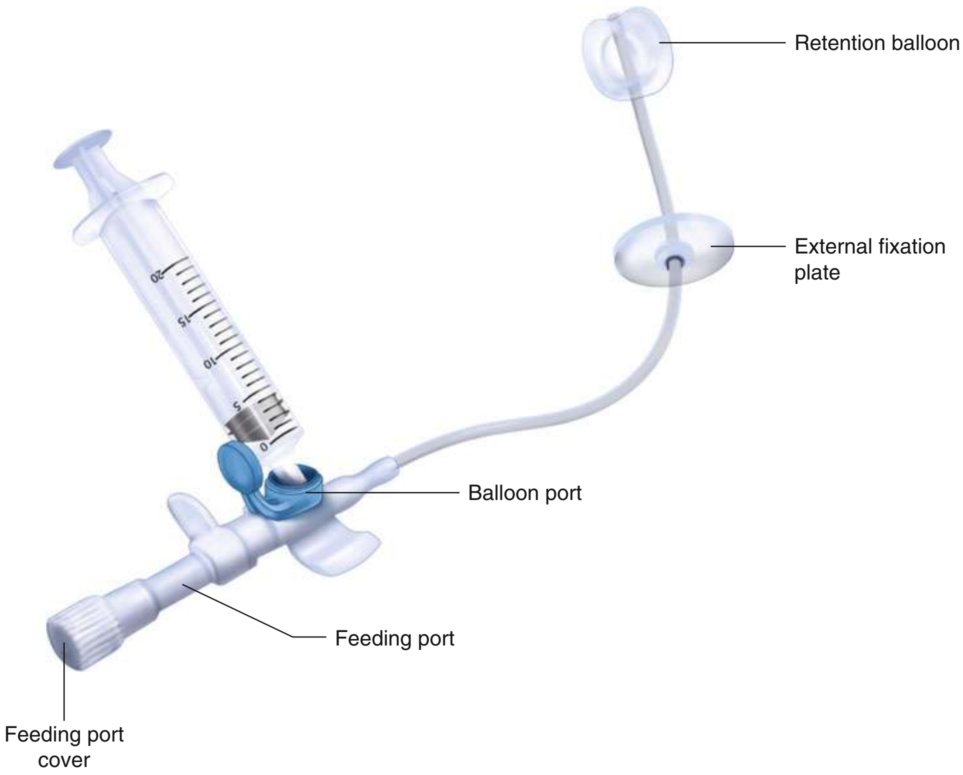
Retention balloon
External fixation plate
20
15
10
5
0
Balloon port
Feeding port
Feeding port cover

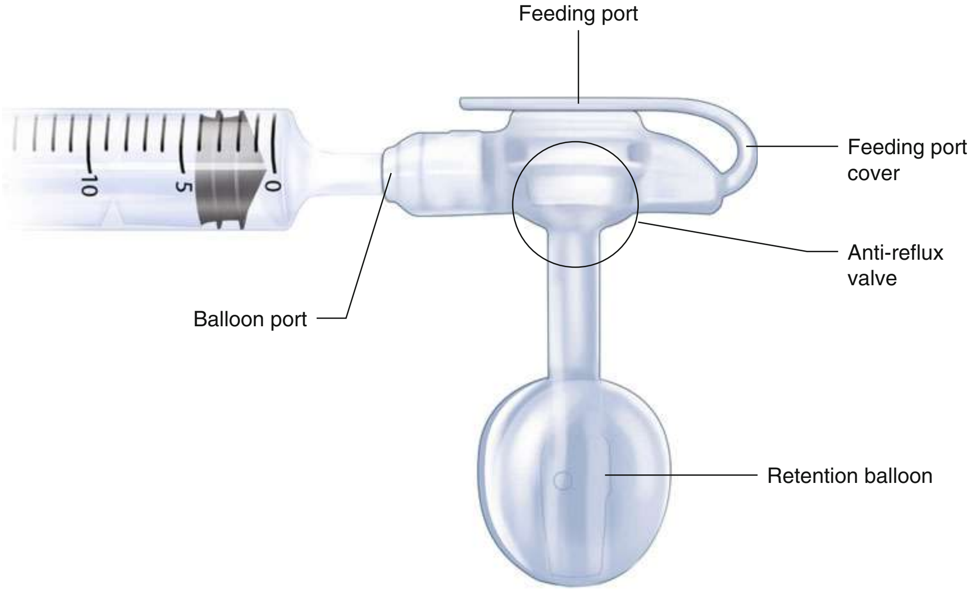
Feeding port
10
5
0
Feeding port cover
Anti-reflux valve
Balloon port
Retention balloon

Fig. 78.4 Low-profile gastrostomy tube (nonobturated, button)

Fig. 78.5 Low-profile gastrostomy tube (obturated, button)

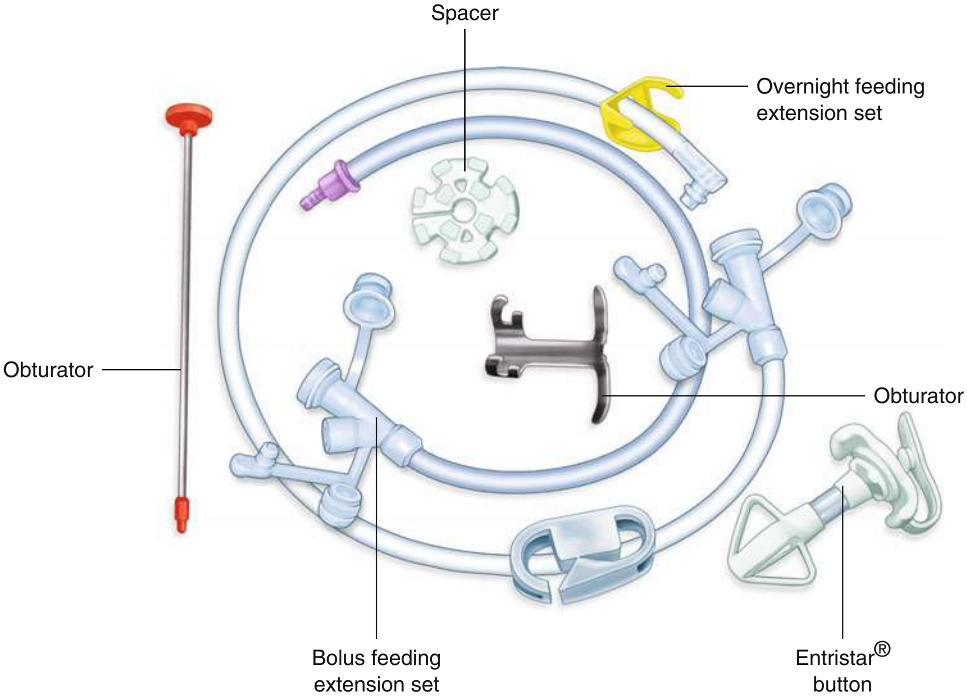
Spacer
Overnight feeding extension set
Obturator
Obturator
Bolus feeding extension set
Entristar® button

78.1 Indications

- Inability to swallow
  - Neurological deficit
  - Head trauma
  - Facial burns
  - Decreased mental status
- Need for gastric decompression
  - Gastric outflow obstruction
  - Small bowel obstruction (SBO), ileus, or volvulus
  - Intra-abdominal malignancy

78.2 Contraindications

- Absolute
  - Peritonitis
  - Ascites
- Relative
  - Hemodynamic instability
  - Coagulopathy
  - Abdominal wall infection at surgical site
  - History of gastric resection
  - Portal hypertension
  - Gastric varices

78.3 Materials and Medications

- Materials
  - Gloves
  - Lubricant
  - G tube (commercial kit) or Foley catheter
  - Sterile saline
  - External bolster
  - Multiple syringes
  - Suture material
  - Needle driver
  - Scissors
  - Stethoscope
- Medications
  - PEG tube placement
    - Moderate sedation (e.g., propofol, midazolam, fentanyl)
    - Local anesthesia (e.g., lidocaine, bupivacaine)
  - Open or laparoscopic G tube placement (not an emergency department procedure)
    - General anesthesia

78.4 Procedure (PEG Tube Replacement)

1. If the G tube is only partially removed upon patient presentation, the tube must first be removed.
2. Remove the G tube by deflating the balloon and pull gently on the tube while applying pressure to the abdominal wall at the surgical site.
3. The G tube should slide out easily with gentle traction, and the procedure should be discontinued if it does not.
4. Initially, it is important to assess the tract to determine the size and potential need for dilation.
5. If necessary for tube passage, the tract can be dilated with a cotton-tipped applicator or hemostat; however, be sure to dilate gently because it is possible to create a false tract.
6. Once the tract has been assessed, obtain the appropriate tube for replacement.
7. Initially, attempt replacement of the patient's G tube with an identical tube.
8. If an identical tube is not available, attempt placing a small Foley catheter in the tract to ensure that it stays open.
9. To initiate placement, place lubricant on the tube and carefully advance the tube into the tract.
10. Once the tube is in place, secure an air-filled syringe to the tip of the tube and insufflate a small amount of air into the stomach while auscultating to confirm passage of air into the stomach.
11. Secondary procedural confirmation can be performed by aspirating gastric contents from the tube.
12. Once placement is verified, do not forget to inflate the balloon with saline and then pull the tube backward until it abuts the inside of the stomach wall (Fig. 78.6).
13. Finally, it is very important to secure the external portion of the tube to prevent the tube from being lost into the stomach.
14. Commercial G tubes are accompanied by a bolster made specifically for the specific type of G tube in order to provide security of tube placement.

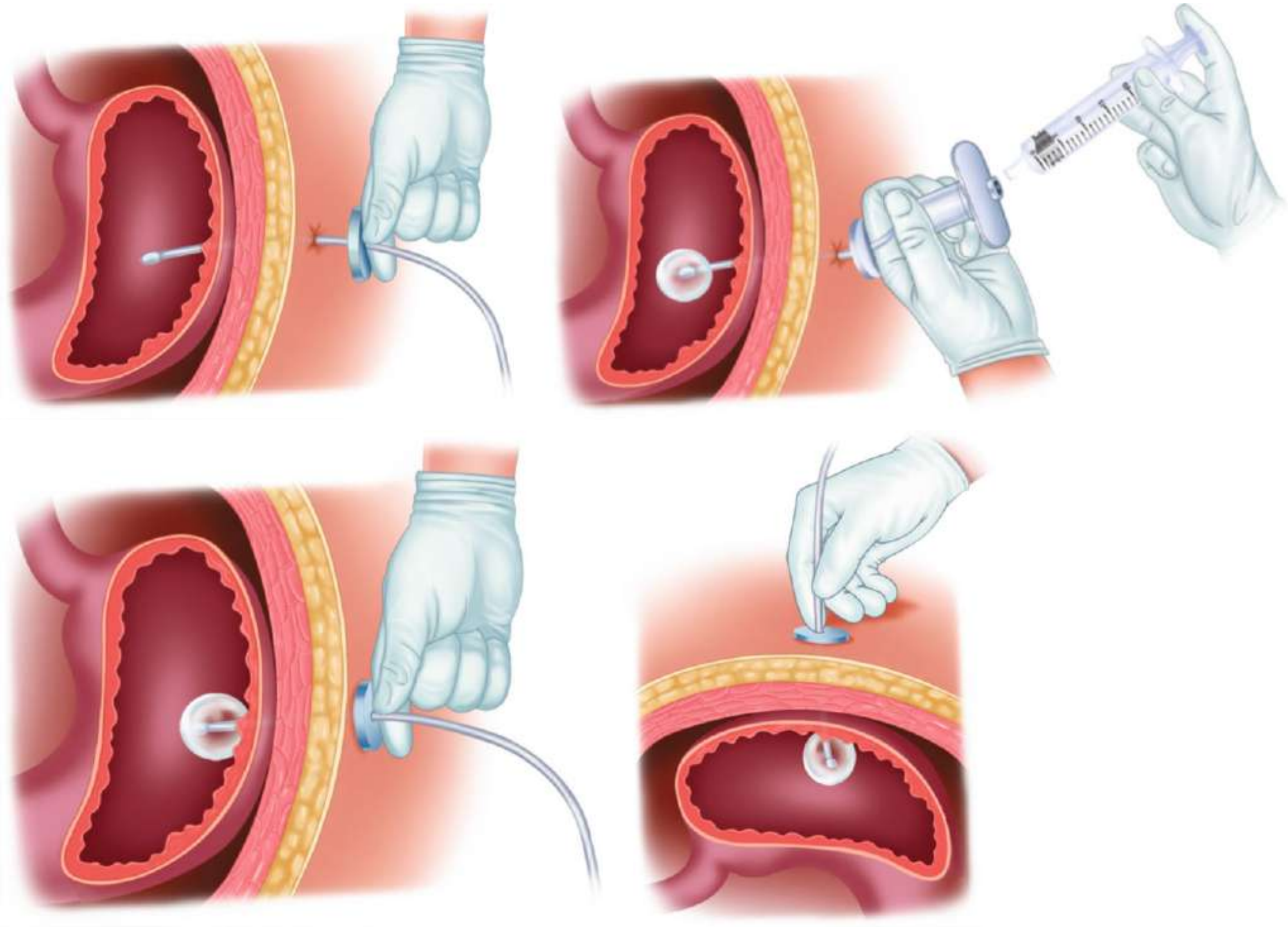

Fig. 78.6 Placement of low-profile gastrostomy tube

15. If a Foley catheter is used to maintain tract patency, an external bolster must be created using the following steps:
    - Trim a 2- to 3-in. portion from the tip of the catheter.
    - Cut two small holes just above each other on the 2- to 3-in. trimmed catheter portion.
    - Slide the external portion of the Foley catheter through both holes of the bolster.
    - Slide the bolster down the Foley catheter to the abdominal wall.
    - Secure the two ends of the bolster to the skin in order to maintain patency of the Foley catheter.
    - An interventional radiologist may also be contacted to advance the tube over a wire under fluoroscopic guidance.
16. Radiographic confirmation.
    - Typically, G tube placement is confirmed by injecting approximately 20 mL of Gastrografin® (diatrizoate meglumine, diatrizoate sodium) water-soluble contrast into the G tube, followed shortly by an abdominal x-ray.
    - Proper placement will result in an abdominal radiograph showing contrast outlining the stomach (Fig. 78.7).

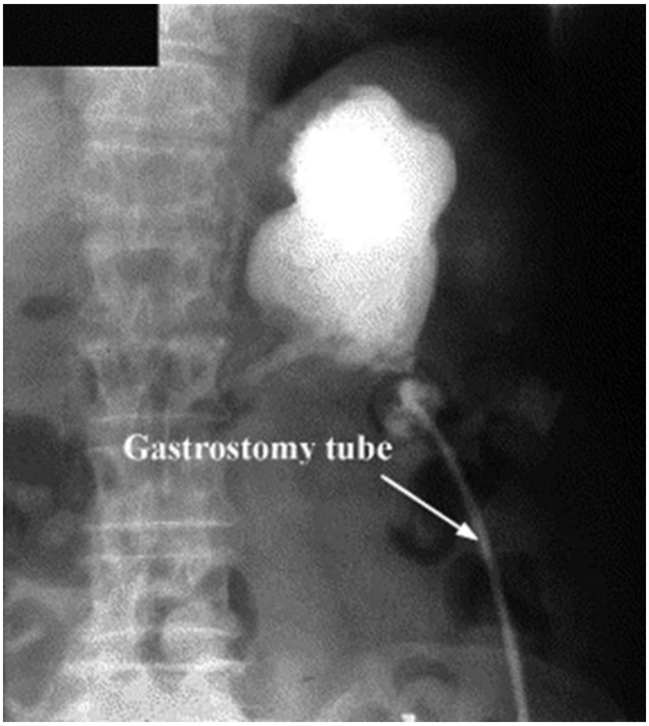

Fig. 78.7 An anteroposterior abdominal radiograph after PEG tube placement and injection of 25 mL of gastrografin. The tube can be seen projecting up the left side of the abdomen, and contrast medium appears to enter the stomach lumen. The balloon is not visualized in the stomach (Reproduced with permission from: Burke DT, El Shami A, Heinle E, and Pina BD. Comparison of gastrostomy tube replacement verification using air insufflation versus gastrografin. Archives of physical medicine and rehabilitation. 2006;87(11): 1530–3.)

78.5 Pearls and Pitfalls

- Pearls
  - Even if a patient's G tube is only partially removed upon presentation to the emergency department, the tube likely needs complete removal and replacement.
  - Most G tubes are easily removed at the bedside; however, some are not able to be safely removed in the emergency department setting. Thus, if the tube does not withdraw easily, attempt to contact the proceduralist who placed the tube in order to inquire about the best method for removal.
- Pitfalls
  - Never use barium contrast when confirming G tube placement radiographically because barium can cause significant intra-abdominal damage if accidentally injected in the intraperitoneal cavity.

78.6 Complications

- Aspiration
- Surgical site infection
- Bleeding
- Pneumoperitoneum
- Accidental perforation of the colon or small bowel
- Tube dislodgment
- Peritonitis

Selected Reading

Arora G. Medscape: percutaneous endoscopic gastrostomy tube placement. Retrieved 3 Jan 2013, from http://emedicine.medscape.com/article/149665-overview#a09.

Gauderer MW, Ponsky JL, Izant Jr RJ. Gastrostomy without laparotomy: a percutaneous endoscopic technique. J Pediatr Surg. 1980; 15:872–5.

Great Ormand street hospital for children website. Retrieved 3 Jan 2013, from www.gosh.nhs.uk/EasySiteWeb/GatewayLink.aspx?alId=102263.

Sarani B. Percutaneous endoscopic gastrostomy tube placement. In: Falter F, editor. Bedside procedures in the ICU. Philadelphia: Springer; 2012. p. 113–22.

Tawa Jr NE, Fischer JE. Metabolism in surgical patients. In: Townsend Jr CM, Beauchamp RD, Evers BM, Mattox KL, editors. Sabiston textbook of surgery. 18th ed. Philadelphia: Elsevier; 2007.

Paracentesis

79

Shalu S. Patel and Bobby K. Desai

79.1 Indications

- Diagnosis of infection in ascites
- Diagnosis of malignant ascites
- Diagnosis of hemoperitoneum in traumas
- Relief of abdominal pressure/pain or respiratory compromise secondary to ascites

79.2 Contraindications

- Severe coagulopathy
  - Prothrombin time (PT) > 21 s
  - International normalized ratio (INR) > 1.6
  - Platelets < 50,000/mm^3
- Skin infection over the needle insertion site
- Acute abdomen that requires surgery
- Pregnancy
- Distended bowel
- Intra-abdominal adhesions

79.3 Materials and Medications

- 18- to 22-gauge 1.5- to 3.5-in. needle or angiocatheter, 25-gauge needle
- Lidocaine 1 or 2 % (10 mL)
- Syringes
  - 10 mL (1), 50 mL (2)
- 1-L vacuum bottle (4) (if therapeutic tap)
- Thoracentesis kit tubing or any high-pressure connection tubing (if therapeutic tap)
- Sterile gloves
- Surgical pen (recommended)
- Povidone-iodine (Betadine) or other skin antiseptic
- Sterile drape
- Sterile gauze (4 × 4)
- Band-Aid
- Bedside ultrasound (recommended)

79.4 Procedure (Fig. 79.1)

1. Position the patient supine. If possible, adjust the head of the bed to make a 45° angle to help the fluid accumulate in the pocket. Sometimes, it may also be beneficial to have the patient lie recumbent toward the site of drainage.
2. Scan the abdomen with an ultrasound to determine whether there is a pocket of fluid that can be drained. This also allows the physician to see how far the needle needs to be inserted and how deep it can be placed without risking injury to the bowel (Fig. 79.2).
3. Mark the optimal needle insertion site with a surgical pen.
4. Prepare the skin and drape in a sterile fashion.
5. Using lidocaine, anesthetize the appropriate area subcutaneously and then continue to insert the needle, and inject anesthetic through the deeper tissues until ascitic fluid can be drawn back.
6. Withdraw the needle.
7. When ready for the paracentesis, stretch the skin caudad and insert the needle or angiocatheter (connected to a syringe) while aspirating. Then, release the skin and continue to insert the needle or angiocatheter through the peritoneal wall until fluid is retrieved. This will create a "Z-track" that will decrease leakage of peritoneal fluid through the skin.

S.S. Patel, MD
Department of Emergency Medicine, Florida Hospital Tampa,
Florida Hospital Carrollwood, Tampa, FL, USA
e-mail: shalu314@gmail.com

B.K. Desai, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_79

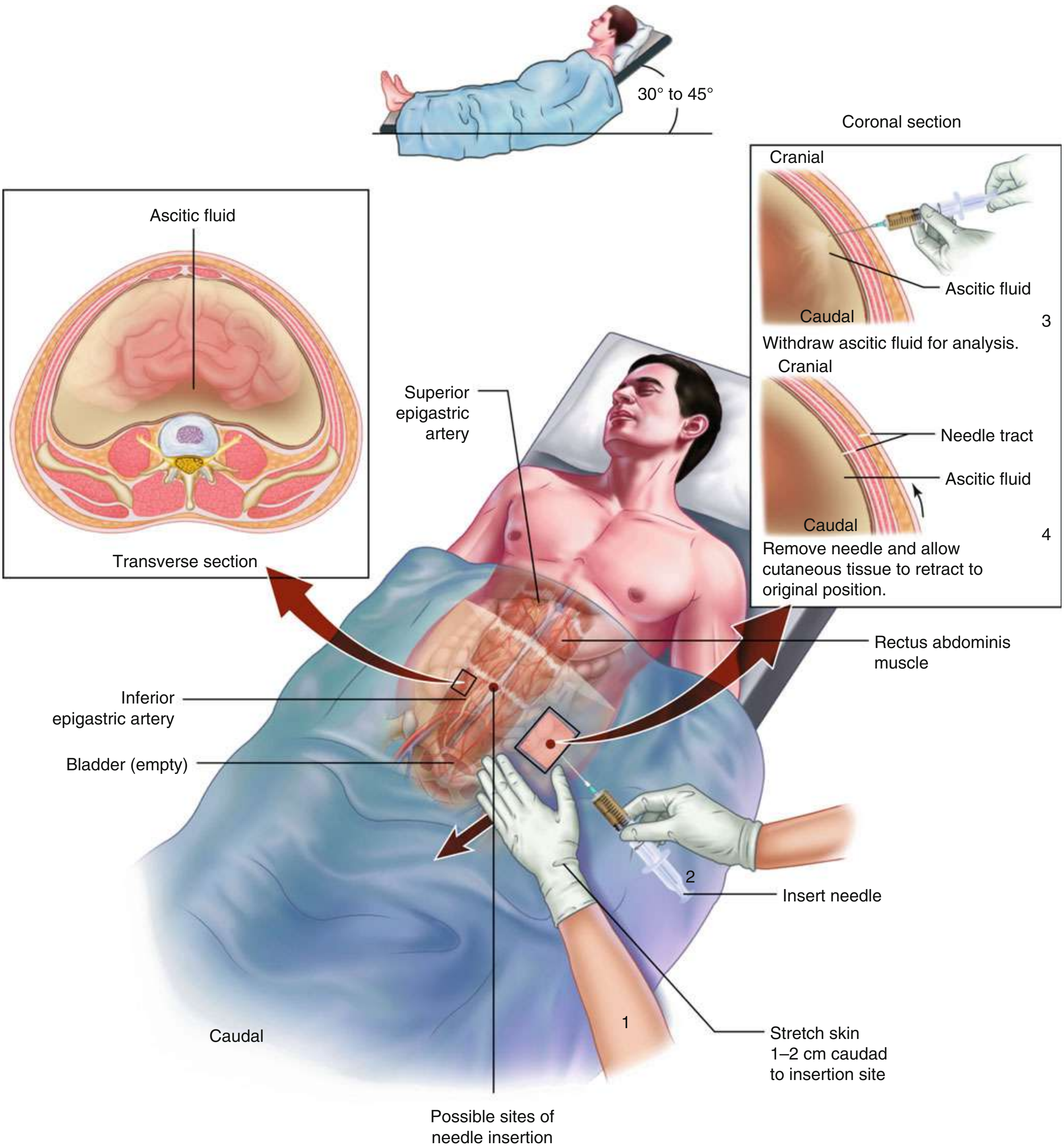
30° to 45°
Coronal section
Cranial
Ascitic fluid
Caudal
3
Withdraw ascitic fluid for analysis.
Cranial
Needle tract
Ascitic fluid
Caudal
4
Remove needle and allow cutaneous tissue to retract to original position.
Ascitic fluid
Transverse section
Superior epigastric artery
Rectus abdominis muscle
Inferior epigastric artery
Bladder (empty)
2
Insert needle
1
Caudal
Stretch skin 1–2 cm caudad to insertion site
Possible sites of needle insertion

Fig. 79.1 Paracentesis procedure

8. Once fluid is retrieved, push in the catheter and remove the needle portion (if used) or hold the needle steady.
9. Aspirate from the catheter to ensure that it is in the appropriate location.
10. If fluid easily is aspirated, unscrew the syringe and connect a 50-mL syringe to the needle or catheter and fill it with fluid. This may be done twice. Alternatively, if the procedure is done for therapeutic purposes, attach the tubing that is already connected to the vacuum bottle to the catheter and allow the vacuum to withdraw fluid into the collection bottles.
11. If fluid cannot be aspirated easily, the catheter can be repositioned further in the pocket or turned by 45° sequentially as needed.

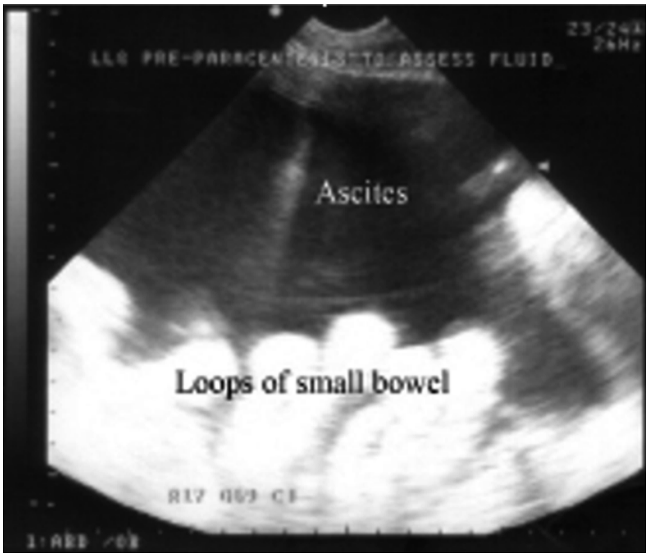

Fig. 79.2 Ultrasound to determine whether there is a pocket of fluid that can be drained

12. Once the fluid is aspirated, pull out the needle or angiocatheter and hold pressure with gauze. Bleeding should be minimal.
13. Place a Band-Aid or other dressing over the site.
14. Send the fluid to the laboratory. Generally, laboratory analyses include protein, albumin, specific gravity, glucose, bilirubin, amylase, lipase, triglyceride, lactate dehydrogenase (LDH), cell count and differential, culture and sensitivity (C&S), Gram stain, acid-fast bacillus (AFB), fungal culture, cytology, and pH.

79.5 Complications

- Persistent leakage from the needle insertion site
- Abdominal wall hematoma
- Bowel perforation
- Introduction of infection
- Hypotension (after a large-volume paracentesis)
- Dilutional hyponatremia
- Hepatorenal syndrome
- Bleeding
- Postparacentesis circulatory dysfunction

79.6 Pearls and Pitfalls

- Pearls
  - The preferred site of entry is in the midline of the abdomen, below the umbilicus.
  - The serum-ascites albumin gradient (SAAG) can be used to identify the cause of the ascites. It is calculated by subtracting the albumin concentration in the ascites from the albumin concentration in the serum. A high gradient (>1.1 g/dL) suggests portal hypertension, whereas a low gradient (<1.1 g/dL) suggests other causes.
  - Postparacentesis circulatory dysfunction (PPCD) occurs secondary to hypovolemia after large-volume paracentesis (>4 L) in cirrhotic patients. It is associated with worsening hyponatremia, renal dysfunction, shorter time to ascites recurrence, and increased mortality. Prevention of PPCD has been demonstrated with the administration of 6–8 g of albumin per liter of ascites removed.
- Pitfalls
  - Polymorphonuclear lymphocyte (PMN) count greater than $250/\text{mm}^3$ is diagnostic of spontaneous bacterial peritonitis.

Selected Reading

Ginès P, Cárdenas A, Arroyo V, Rodés J. Management of cirrhosis and ascites. N Engl J Med. 2004;350:1646–54.

Ginès P, Tito L, Arroyo V, et al. Randomized comparative study of therapeutic paracentesis with and without intravenous albumin in cirrhosis. Gastroenterology. 1988;94:1493–502.

Ruiz del Arbol L, Monescillo A, Jimenéz W, et al. Paracentesis-induced circulatory dysfunction: mechanism and effect on hepatic hemodynamics in cirrhosis. Gastroenterology. 1997;113:579–86.

Runyon BA. Paracentesis of ascitic fluid. A safe procedure. Arch Intern Med. 1986;146:2259–61.

Wong CL, Holroyd-Leduc J, Thorpe KE, Straus SE. Does this patient have bacterial peritonitis or portal hypertension? How do I perform a paracentesis and analyze the results? JAMA. 2008;299:1166–78.

Anal Fissure Management

80

David P. Nguyen, L. Connor Nickels, and
Giuliano De Portu

80.1 Indications

- An anal fissure is a small ulcer of the mucosa at the anal verge (Fig. 80.1).
- It is the most common cause of intense sudden rectal bleeding.
- Posterior midline anal fissures are the most common type (90 %).
  - Mostly found in young adults (30–50 y) but can occur at any age.
  - Usually associated with constipation (firm, large-caliber, painful bowel movements) or chronic diarrhea.
  - Most uncomplicated fissures resolve in 3–4 weeks.
- Can be extremely painful, during and after defecation.
- Classified as acute or chronic.
- Now believed to be caused by reduced anal blood flow in the posterior midline, anal sphincter hypertonia, and thus mucosal ischemia.

D.P. Nguyen, DO
Department of Emergency Medicine, Rush-Copley Medical Center, Aurora, IL, USA
e-mail: davidpnguyen@yahoo.com

L.C. Nickels, MD, RDMS (✉) • G. De Portu, MD
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; gdeportu@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_80

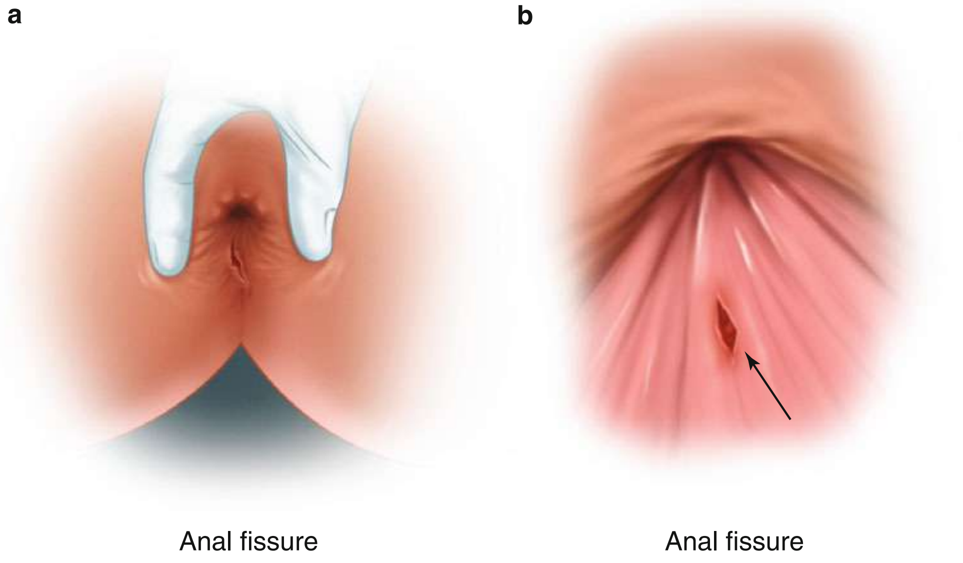
a
b
Anal fissure
Anal fissure

Fig. 80.1 (**a**, **b**) Anal fissures

80.2 Contraindications

- Digital rectal examination should be avoided unless the diagnosis is in doubt.
- Surgical procedures are generally reserved for when medical management has failed after 1–3 months of treatment.

80.3 Materials and Medications

- Standard precautions barrier protection for the provider.
- Good light source.
- Optional emergency department treatments:
  - Topical anesthetic/preparation (Anusol [pramoxine hydrochloride; zinc oxide] with cortisone).
  - Nitroglycerin (0.2 %) or nifedipine gel (2 %) is second-line therapy (relaxes muscles and promotes blood flow).

80.4 Procedure

1. In a private, calm environment, gently spread the buttocks for complete visual inspection.
   - This may cause an increase in the patient's pain and spasming.
   - If a fissure is clearly identified, stop here.
2. Apply topical anesthetic/preparation for symptomatic relief (optional, as the physician may want to just start with the treatments that follow).
3. Discharge the patient with conservative therapy management.
4. In acute anal fissures (onset of 3–6 weeks), medical management is indicated along with dietary modifications (WASH regimen [warm baths, analgesia, stool softeners, high-fiber diet]):
   - Warm sitz baths.
     - Usually 20 min soaking each time
     - Recommended after every bowel movement
     - At least twice per day if not having regular bowel movements
   - High-fiber diet with fiber supplements.
   - Increase fluid intake.
   - May add stool softeners, if needed.
   - If chronic or the previous regimen has been exhausted, one of the following may be considered:
     - 0.2–0.4 % nitroglycerin cream applied to anal area
       - May cause headache
       - Recommend wearing a glove to prevent absorption through digital skin
     - Calcium channel blockers:
       - Topical nifedipine
       - 2 % diltiazem cream
     - Botulinum toxin A injection:
       - Controversial; may have poorer success rates than surgery
5. Provide surgical referral for nonhealing wounds.
   - Lateral internal sphincterotomy is the surgical procedure of choice.

80.5 Complications

- Infection
- Abscess
- Bleeding
- Chronic fissure formation
- Postsurgical fecal incontinence

80.6 Pearls and Pitfalls

- Pearls
  - Multiple or recurrent fissures are associated with Crohn's disease, tuberculosis, syphilis, human immunodeficiency virus (HIV), and malignancy.
- Pitfalls
  - Suspect child abuse if an anal fissure is found in a child.

Selected Reading

Feldman M, Friedman LS, Brandt LJ, editors. Sleisenger and Fordtran's gastrointestinal and liver disease. 9th ed. Philadelphia: Saunders; 2010.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine. 7th ed. Philadelphia: Mosby; 2010. Chapter 94: Disorders of the anorectum

Oztürk H, Onen A, Dokucu AI, Otçu S, Yağmur Y, Yucesan S. Management of anorectal injuries in children: an eighteen-year experience. Eur J Pediatr Surg. 2003;13:249–55.

Townsend CM Jr, Beauchamp RD, Evers BM, Mattox KL, editors. Sabiston textbook of surgery. 19th ed. Philadelphia: Saunders; 2012. Chapter 53: Anus

Part XI

Genitourinary Procedures

Bladder Catheterization

81

Maritza A. Plaza-Verduin and Judith K. Lucas

81.1 Indications

- Obtaining a sterile urine specimen
- Preventing or relieving urinary retention
- Close monitoring of urine output for fluid balance with an indwelling urinary catheter
- Urgent cystourethrography
- Child with contusion or burns to the perineum and at risk for meatal swelling and obstruction to urine outflow
- Temporary measure to relieve lower urinary tract obstruction
- Neurogenic bladder
- Anesthesia-induced and/or surgery-induced urinary retention has occurred.

81.2 Contraindications

- Absolute
  - Potential urethral injury from trauma
    - Pelvic fractures
    - Known trauma to the urethra
    - Blood at the meatus
- Relative
  - Recent genitourinary surgery (consult with a urologist before placing a catheter)

81.3 Materials and Medications

- Bladder catheterization kit:
  - Sterile gloves
  - Sterile drapes
  - Povidone-iodine (Betadine) solution
  - Cotton sponges or applicators for sterilizing solution
  - Lubricant
  - Specimen collection cup
  - Catheter
    - 5-French feeding tube for neonates
    - 8-French catheters for Infants
    - 10- to 12-French catheters in older children
- Local anesthetic (if desired—2 % lidocaine hydrochloride jelly)
- Absorbent pad

M.A. Plaza-Verduin, MD
Pediatric Division, Department of Emergency Medicine, Arnold Palmer Hospital for Children, Orlando, FL, USA
e-mail: mari.plazav@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_81

81.4 Procedure

1. Inspect the urinary catheterization tray for all the appropriate materials.
2. Place the patient supine with an absorbent pad under the buttocks.
   (a) Girls should be placed in the frog-leg position (Fig. 81.1).
3. Before sterilizing the field, locate the urethral opening.
4. Remove any powder, ointments, or medicated creams the child might have on the perineum.
5. If needed, apply anesthetic to the area.
   (a) Soak a cotton ball with anesthetic (2 % lidocaine hydrochloride jelly) and hold over the urethra opening for 2 min.
   (b) 0.5–2.0 mL of anesthetic can also be injected into the urethra.
6. Sterilize the area and place drapes appropriately, exposing the genitalia.
7. Catheterization of males
   (a) If uncircumcised, gently retract foreskin, *if possible*, for cleaning and visualization of the meatus.
   (b) Hold the penis using the nondominant hand at a 90° angle from the body (Fig. 81.2).
   (c) Lubricate the catheter tip.
   (d) Insert the lubricated catheter into the meatal opening and advance it while applying gentle traction to the penis from the base of the penis.
   (e) If resistance is met, maintain gentle pressure with the catheter.
      - *Do not attempt to force the catheter that could create a false tract or traumatic fistula.*
   (f) Advance the catheter until urine is obtained, approximately inserting the catheter to just beyond the penile length.
   (g) Once completed, gently withdraw the catheter.
   (h) Clean the area, wiping away the Betadine solution.
   (i) If uncircumcised, pull the foreskin over the glans to avoid paraphimosis.
8. Catheterization of females
   (a) Sterilization of the area should occur from anterior to posterior.
   (b) Have an assistant hold the labia majora apart.
      (i) If no assistant is available, use the nondominant hand to hold the labia apart.
         - Holding the labia majora with a gentle outward, lateral, and upward traction will help visualize the meatus (Fig. 81.3).

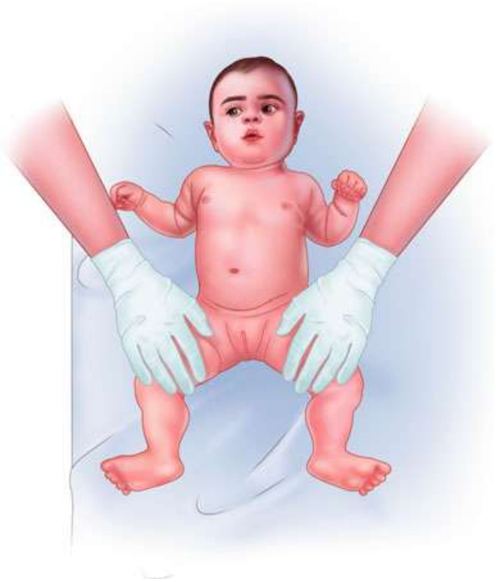

Fig. 81.1 Infant held in the frog-leg position for catheterization

Fig. 81.2 Bladder catheterization of a male; penis should be held perpendicular to the body

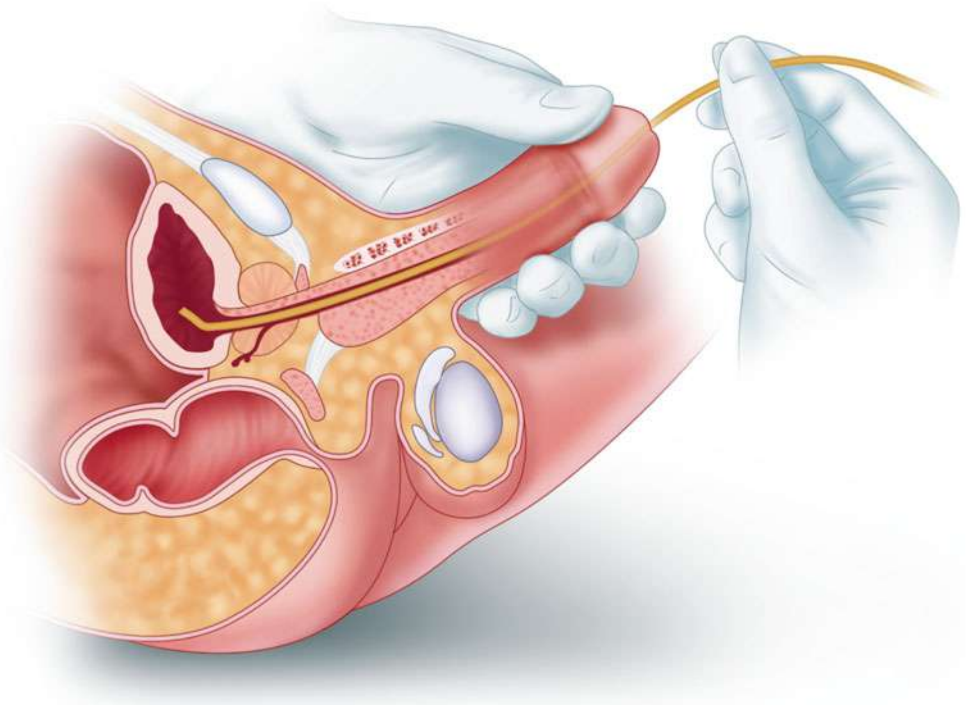

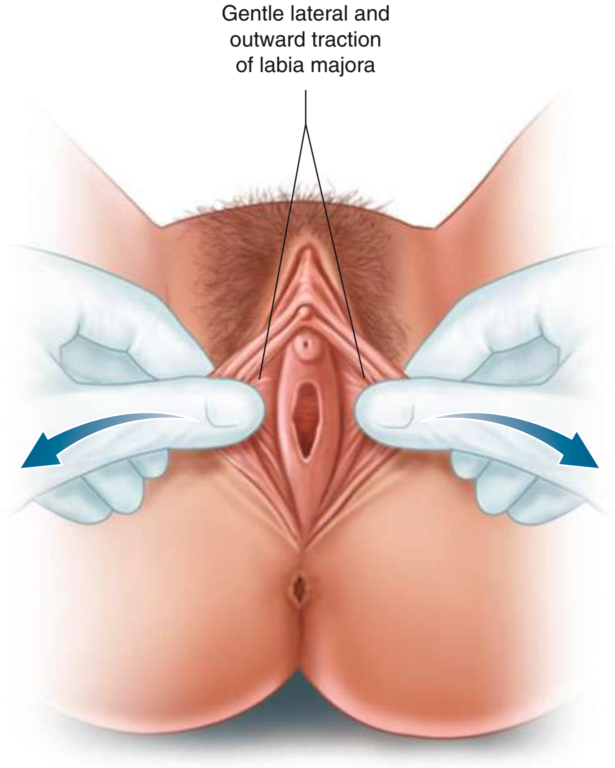
Gentle lateral and
outward traction
of labia majora

Fig. 81.3 Positioning of labia for better visualization of the meatus

  - Downward displacement of the cephalad aspect of the vaginal introital fold with a cotton-tipped applicator can help visualize the urethral meatus (Fig. 81.4).

(c) Lubricate the catheter tip.

(d) Insert the lubricated catheter into the meatal opening. Advance slowly until urine is obtained (Fig. 81.5).

(e) Once completed, gently withdraw the catheter.

(f) Clean area, wiping away the Betadine solution.

Fig. 81.4 Better visualization of the meatus is achieved with downward displacement of the introital mucosa

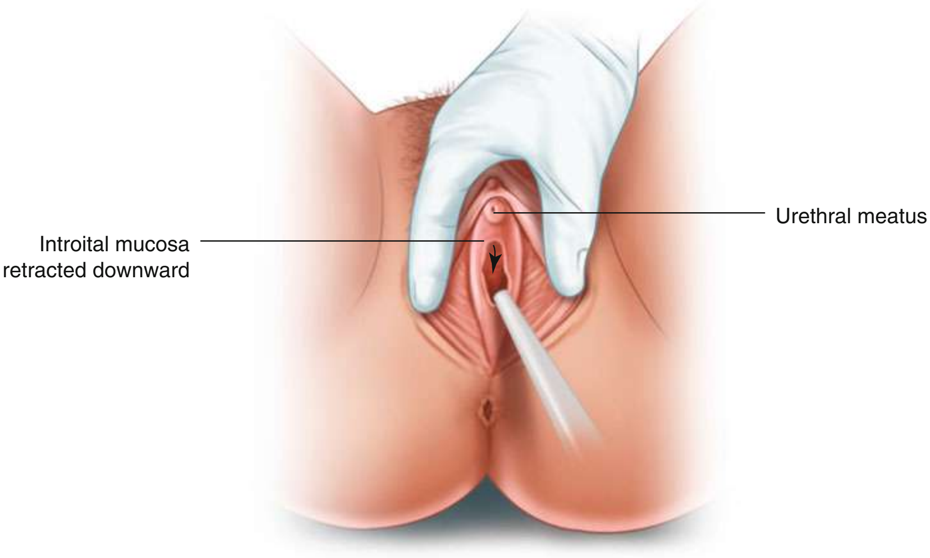

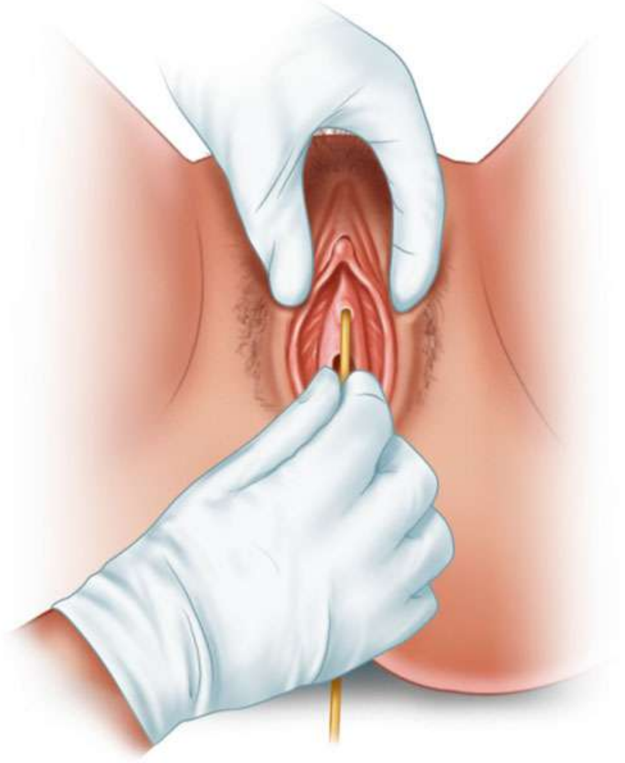

Fig. 81.5 Bladder catheterization of a female

81.5 Complications

- Urethral or bladder injury
- Infection if sterile field not maintained
- Paraphimosis owing to failure to restore a retracted foreskin to its normal position

81.6 Pearls and Pitfalls

- Pearls
  - It is not necessary to fully retract a foreskin. This only causes trauma and increases the likelihood of paraphimosis. As the infant/boy ages, the foreskin will loosen and the naturally occurring adhesions will spontaneously release.
  - The urethral meatus in an infant female is usually tucked just above the redundant hymen (as opposed to the more anteriorly located meatus in the adult woman) and often looks like a dimple or small blind pouch.
  - Using viscous lidocaine in lieu of, or blended with, lubricant anesthetizes the meatus and urethra as the catheter passes.
  - Have the specimen cup at the ready because, at times, once the (usually cold) Betadine or antiseptic solution is applied, the infant often releases the urine and it can be caught, literally, midstream.
  - In the uncircumcised male, be certain to return the foreskin over the glans to avoid paraphimoses.
- Pitfalls
  - If catheterizing a child in search of infection, send a urine culture regardless of the urinalysis results because the younger infants can have false-negative urinalysis and still have positive cultures.

Selected Reading

American Academy of Pediatrics, Subcommittee on urinary tract infection, steering committee on quality improvement and management. Urinary tract infection: clinical practice guideline for the diagnosis and management of the initial UTI in febrile infants and children 2 to 24 months. Pediatrics. 2011;128:595–609.

Beno S, Schwab S. Bladder catheterization. In: King C, Henretig FM, editors. Textbook of pediatric emergency procedures. 2nd ed. New York: Lippincott Williams & Wilkins; 2008.

Cheng YW, Wong SN. Diagnosing symptomatic urinary tract infection in infants by catheter urine culture. J Paediatr Child Health. 2005;41:437–40.

Gerard LL, Cooper CS, Duethman KS, et al. Effectiveness of lidocaine lubricant for discomfort during pediatric urethral catheterization. J Urol. 2003;170:564–7.

Kozer E, Rosenbloom E, Goldman D, et al. Pain in infants who are younger than 2 months during suprapubic aspiration and transurethral bladder catheterization: a randomized, controlled study. Pediatrics. 2006;118:e51–6.

Pelvic Examination and Wet Preparation

82

Nauman W. Rashid, Elaine B. Josephson,
and Muhammad Waseem

82.1 Indications

- Lower abdominal or pelvic pain
- Vaginal bleeding or discharge
- Cancer screening
- Pregnancy
- Exposure to sexually transmitted disease
- Sexual assault

82.2 Contraindications

- Physical or mental disability
- Recent gynecological surgery
- Third-trimester pregnancy with bleeding
- Premenstrual females (may not be indicated in adolescents, who are not sexually active, unless there is discharge, bleeding, suspicion for abuse, or a foreign body)
- If a speculum examination is necessary, examination under general anesthesia should be considered.

82.3 Materials and Medications

- Examination table with stirrups (Fig. 82.1)
- Reliable light source
- Appropriately sized speculum (Fig. 82.2)
- Endocervical brush or spatula
- Culture swab for gonorrhea and chlamydia
- Large cotton swabs for vaginal discharge or bleeding (Fig. 82.2)
- pH paper
- Saline and potassium hydroxide dropper bottles for wet preparations
- Lubricating gel
- Disposable gloves (Fig. 82.2)
- Small stool or chair for examiner

82.4 Procedure

1. Obtain permission from patient before beginning examination.
2. Chaperone should be present (medical staff member).
3. Make sure the examination table is clean and appropriately draped.
4. Have the patient in a loose-fitting gown.
5. Place the patient on the examination table in the lithotomy position with both feet in the stirrups and have the patient's pelvis as close to the edge of the table as possible.
6. Turn on the light source and adjust for optimum illumination. Put on the disposable gloves.
7. Communicate the procedure well to the patient.
8. Begin the examination with inspection and palpation of the abdomen.
9. Examine the external genitalia. Evaluate the skin, labia minora and majora, clitoris, urethral meatus, vaginal

---

N.W. Rashid, MD
Department of Emergency Medicine, WellStar Kennestone Hospital, Marietta, GA, USA
e-mail: naumanrashid@hotmail.com

E.B. Josephson, MD
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, Weill Cornell Medical College of Cornell University, Bronx, NY, New York
e-mail: elaine.Josephson@nychhc.org

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_82

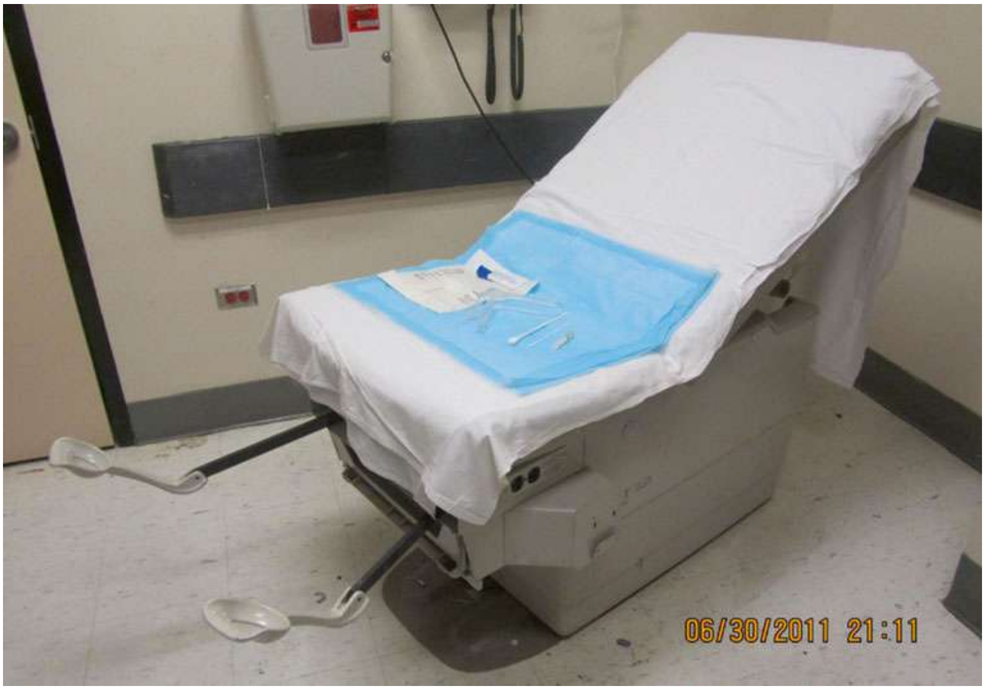

Fig. 82.1 Examination table with stirrups

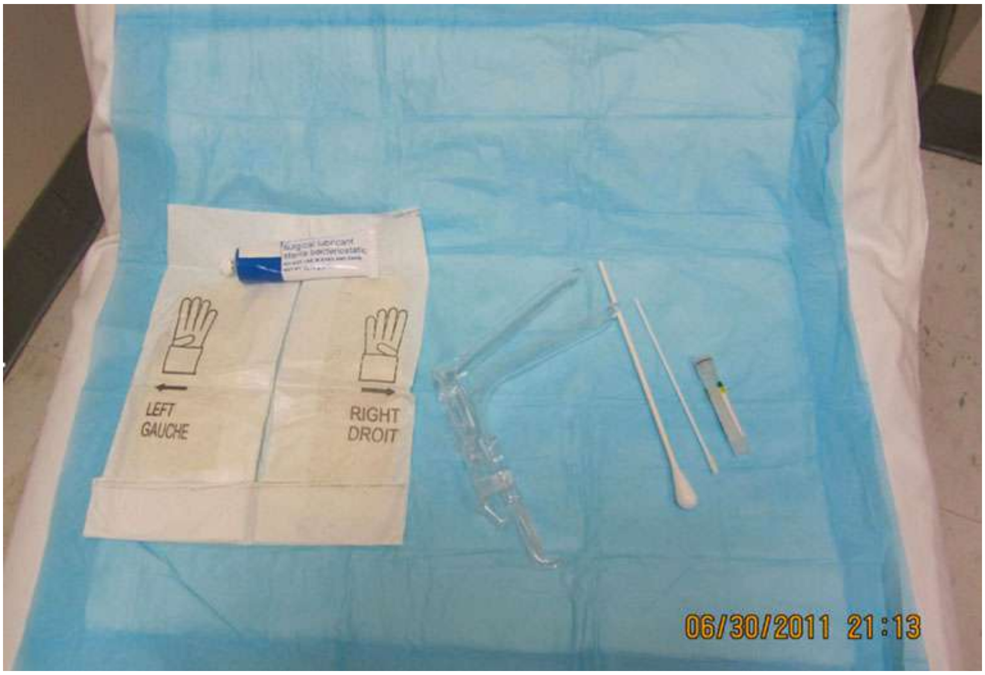

Fig. 82.2 Gloves, speculum, and swabs

canal, and Bartholin glands (Fig. 82.3). Look for skin abnormalities, lesions, masses, rashes, excoriation, abscesses, discharge, bleeding, or trauma. Palpate for tenderness.

10. Lubricate the appropriate-size speculum (mostly medium size). Insert the speculum through the vaginal opening with gentle downward pressure. The speculum should advance without any resistance until the cervical os is visualized.
11. Inspect the vaginal walls for any lesions or masses (Fig. 82.4). The cervical os is inspected to see if it is open or closed. Cervical cultures for gonorrhea and chlamydia are obtained with a cotton swab and sent for microbiology.
12. A sample of the discharge or bleeding is taken with a large cotton swab. The color, odor, and amount should be noted. The pH of the vaginal discharge can be evaluated. Normal pH is less than 4.5. An elevated pH indicates an infection (Table 82.1).

Fig. 82.3 Female external genitalia

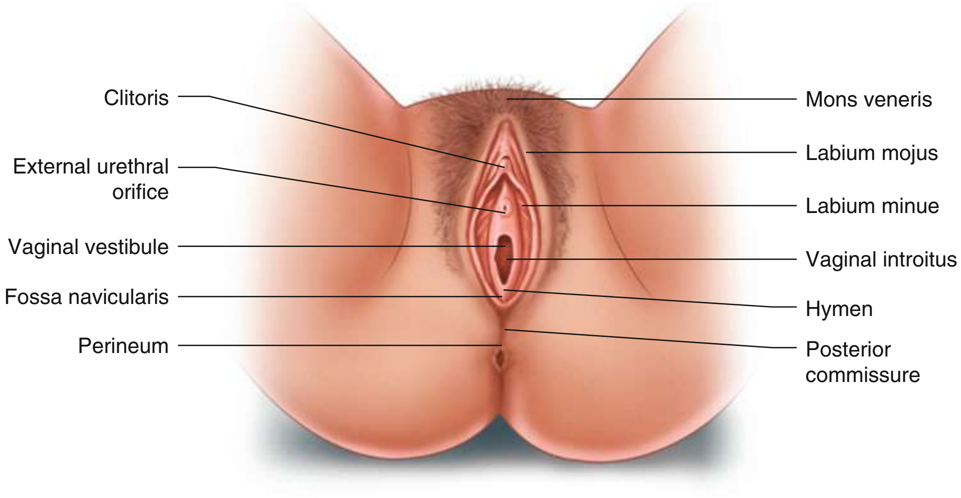
Clitoris
External urethral orifice
Vaginal vestibule
Fossa navicularis
Perineum
Mons veneris
Labium mojus
Labium minue
Vaginal introitus
Hymen
Posterior commissure

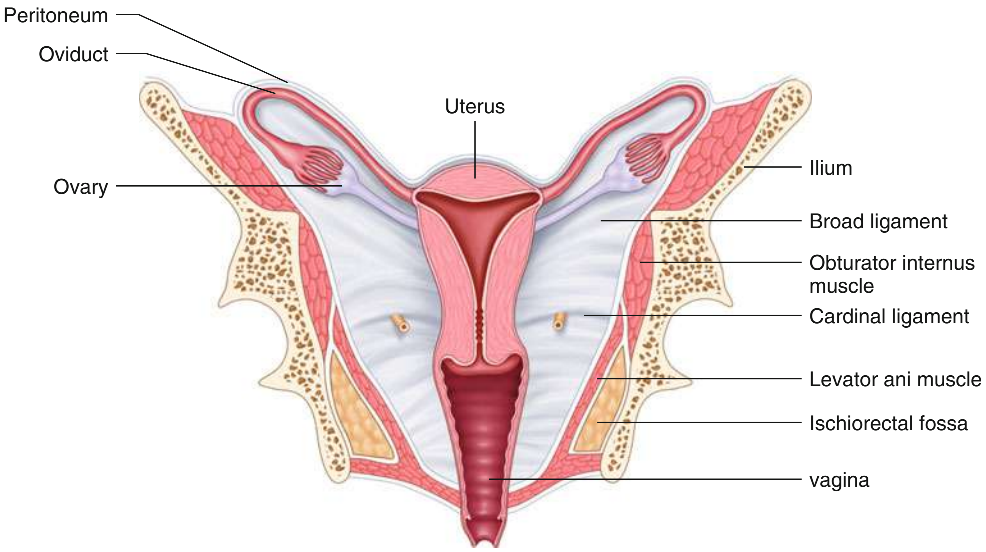
Peritoneum
Oviduct
Uterus
Ovary
Ilium
Broad ligament
Obturator internus muscle
Cardinal ligament
Levator ani muscle
Ischiorectal fossa
vagina

Fig. 82.4 Female internal genitalia

Table 82.1 Wet preparation interpretation

Organism	Preparation	pH	Microscope	Cervix	Appearance of discharge
Bacterial vaginosis	Saline	>4.5	Clue cells	Redness	Thin, milky, fishy odor
Trichomoniasis	Saline	>4.5	Motile flagella	Strawberry red	Yellow-green, foamy
Yeast	Potassium hydroxide	3.8–4.5	Budding yeast pseudohyphae	Normal	White, cottage cheese

13. Next, a bimanual examination should be performed (Fig. 82.5). Lubricating gel is applied to the nondominant gloved hand and the index and middle finger are inserted into the vagina until the cervix is felt. The other hand is placed on the abdomen to palpate the uterus and ovaries. Pressure is applied to the abdomen while the vaginal hand is elevated upward.
14. The cervix is palpated to elicit any cervical motion tenderness. The uterus is palpated and the size, position, and mobility are noted. The adnexa are examined for masses and tenderness. If a mass is palpated, the size, mobility, consistency, and tenderness are noted.
15. The final part of the pelvic examination is the rectovaginal examination. Lubricate the index and middle fingers

of the left hand. Place the index finger in the vagina and the middle finger in the rectum. Palpate for any fistulas or masses. With the finger, also palpate the uterosacral ligaments, the broad ligaments, and the pelvic side walls. The finger is then gently removed and any feces are inspected for mucous or occult blood.

16. A wet preparation is made by obtaining a sample of the vaginal discharge and placing it in a vial mixed with saline solution. A drop of the solution is placed on a microscopic slide and examined under high magnification for the presence of clue cells (Fig. 82.6) diagnostic for bacterial vaginosis and trichomonads (Fig. 82.7) diagnostic for trichomoniasis. For yeast, two drops of the solution is mixed with two drops of potassium hydroxide. Presence of hyphae is diagnostic of candida (yeast) species (Fig. 82.8)

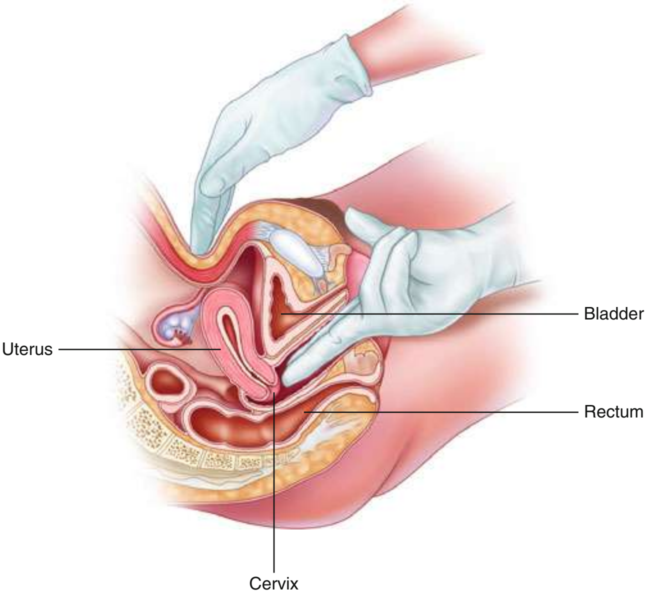

Fig. 82.5 Bimanual pelvic examination

82.5 Complications

- Urinary tract infection
- Vaginal bleeding
- Cramping

82.6 Pearls and Pitfalls

- Pearls
  - Good communication is essential to ensure the patient is comfortable and not anxious.
  - The chaperone should be a medical staff member.
  - Do not skip the pelvic examination if the patient is menstruating.
- Pitfalls
  - Do not forget to perform a complete abdominal exam along with the pelvic exam to rule out any GI etiology.
  - In older females (>50), perform the DRE for a stool occult sample as a possible source of bleeding.

Fig. 82.6 Photomicrograph of a vaginal smear specimen depicting two epithelial cells, a normal cell, and an epithelial cell with its exterior covered by bacteria giving the cell a roughened, stippled appearance known as a “clue cell” (From the CDC Public Health Image library)

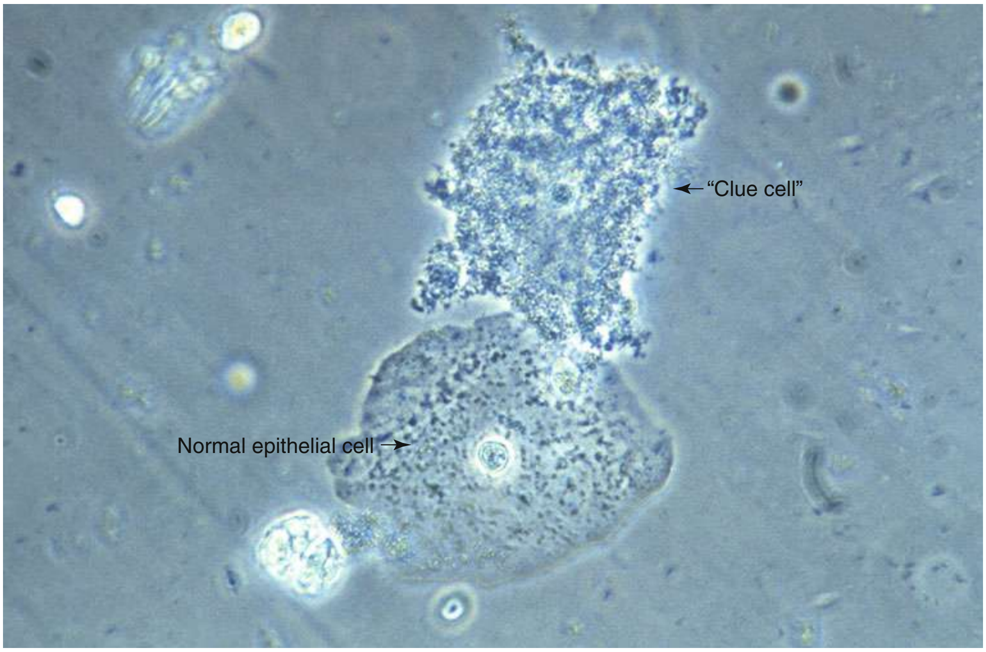

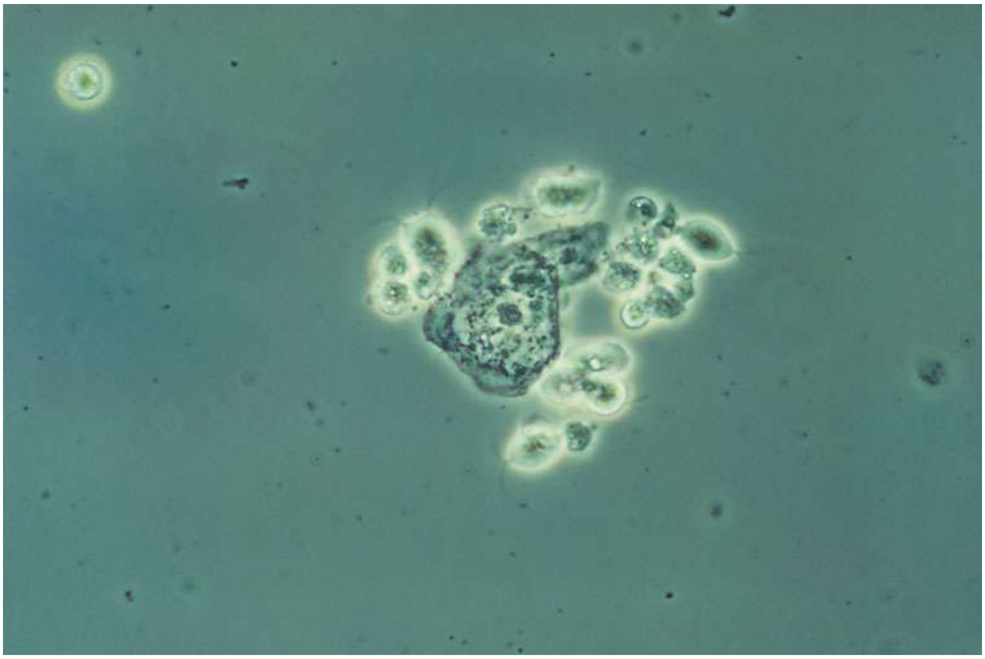

Fig. 82.7 Photomicrograph of trichomonads in wet mount prepared witvh physiological saline (From the CDC Public Health Image library)

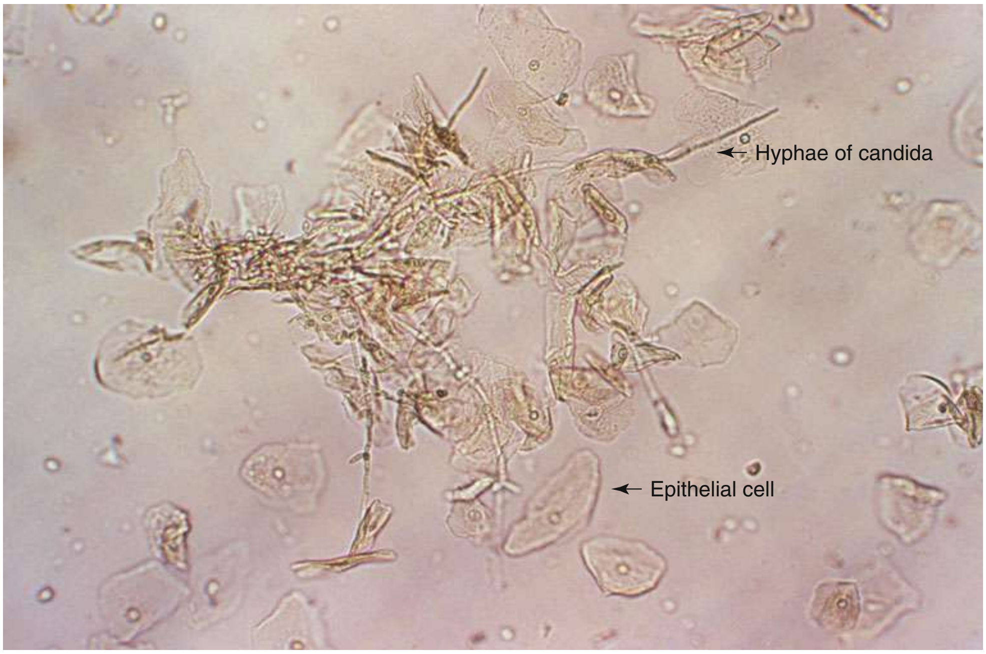

Fig. 82.8 *Candida albicans* from vaginal wet prep (From the CDC Public Health Image library)

Selected Reading

Brown J, Fleming R, Aristzabel J, Gishta R. Does pelvic exam in the emergency department add useful information? West J Emerg Med. 2011;12:208–12.

Butler J, Barton D, Shepherd J, Reynolds K, Kehoe S. Gynaecological examinations. Good not bad medicine. BMJ. 2011;342:d1760.

Carr SE, Carmody D. Outcomes of teaching medical students core skills for women's health: the pelvic examination educational program. Am J Obstet Gynecol. 2004;190:1382.

Katz VL, Lentz G, Lobo RA, Gershenson D, editors. Comprehensive gynecology. 5th ed. Philadelphia: Mosby; 2007.

Tiemstra J, Chico P, Pela E. Genitourinary infections after a routine pelvic exam. J Am Board Fam Med. 2011;24:296–303.

83

Bartholin Gland Abscess/Cysts Drainage

Holly H. Charleton, Marylin Otero, Diane F. Giorgi, and Joseph A. Tyndall

83.1 Indications

- Bartholin gland cysts larger than 1 cm or painful
- Bartholin gland abscess

83.2 Contraindications

- Absolute
  - None
- Relative
  - Recurrent/complex abscess requiring general anesthesia in operating room
  - Coagulopathy

83.3 Procedure Types

- Incision and drainage (I&D)
- Iodoform packing
- Word catheter
- Jacobi ring catheter
- Silver nitrate stick ablation (after I&D)

H.H. Charleton, MD
Emergency Department, The Brooklyn Hospital Center,
Brooklyn, NY, USA
e-mail: Hcmd08@gmail.com

M. Otero, MD
Department of Emergency Medicine, Franklin Hospital,
Valley Stream, NY, USA
e-mail: momd42@hotmail.com

D.F. Giorgi, MD
Department of Emergency Medicine, Mount Sinai Queens,
New York, NY, USA
e-mail: dgiorgi@earthlink.net

J.A. Tyndall, MD (✉)
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: tyndall@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_83

83.4 Materials

- Sterile gloves
- Sterile skin preparatory solution/swabs and drapes
- Lidocaine 1 % (local anesthesia); may give oral or intravenous sedatives or analgesics
- Needles 25 or 27 gauge with 3-mL syringe for lidocaine injection
- Scalpel #11 for incision
- Culture swab
- Hemostat or suture kit
- Word catheter (if choosing that method) (Fig. 83.1)
- Jacobi ring catheter (if choosing that method) (Fig. 83.2)
- 3-mL syringe with saline for inflation of Word catheter balloon
- Silver nitrate stick (if choosing that method)
- Gauze pads for bleeding and effluents
- Iodoform packing

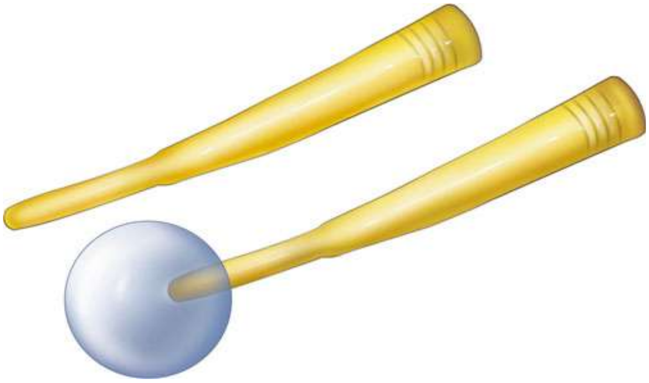

Fig. 83.1 Word catheter (inflated/deflated)

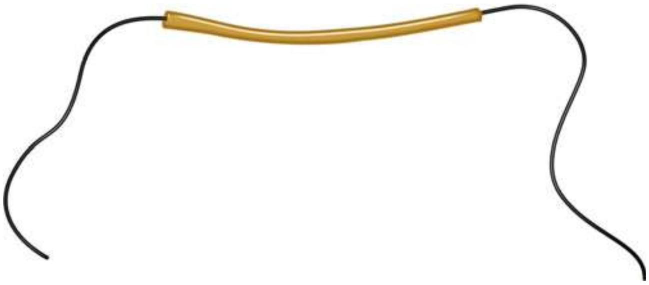

Fig. 83.2 Jacobi ring catheter

83.5 Procedures

83.5.1 Incision and Drainage

1. Obtain informed consent from the patient.
2. Place the patient in the lithotomy position.
3. Prepare the cyst/abscess and the surrounding area with sterilizing fluid/swabs and drape the area, leaving the cyst/abscess accessible for the procedure (Fig. 83.3)
4. May need to hold traction to the labia to fully expose the cyst/abscess (Fig. 83.4).
5. Inject 1–4 mL of lidocaine at the planned site of incision (Fig. 83.5).
6. Hold one side of the cyst/abscess with a forceps or hemostat to maintain traction while incising.
7. With a #11 scalpel, make an incision approximately 0.5–1 cm and 1.5 cm deep in the introitus or behind the hymnal ring to prevent vulvar scarring. The incision should be made through the fluctuant area of the abscess on the mucosal surface. The incision should be linear and large enough to fit the Word catheter (if using that method).
8. Drain the cyst/abscess completely, using the hemostat to break up the loculations.
9. Culture the abscess with a swab and send to microbiology.
10. All previous steps should be done regardless of the procedure type to follow.

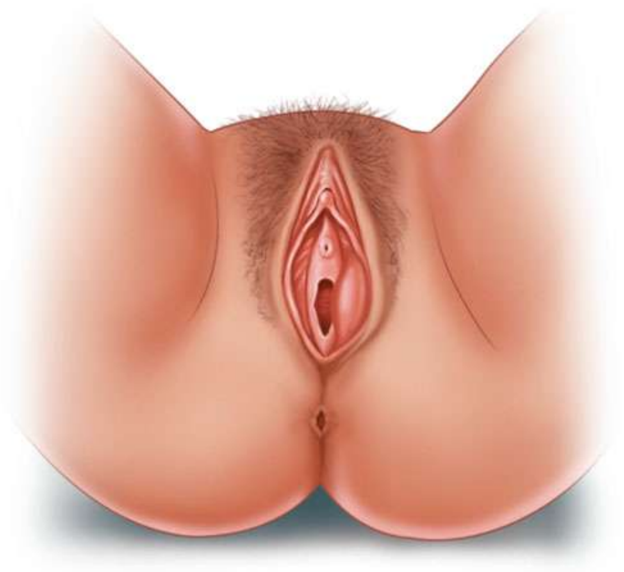

Fig. 83.3 Cyst/abscess accessible for the procedure

Fig. 83.4 May need to hold traction to the labia to fully expose the cyst/abscess

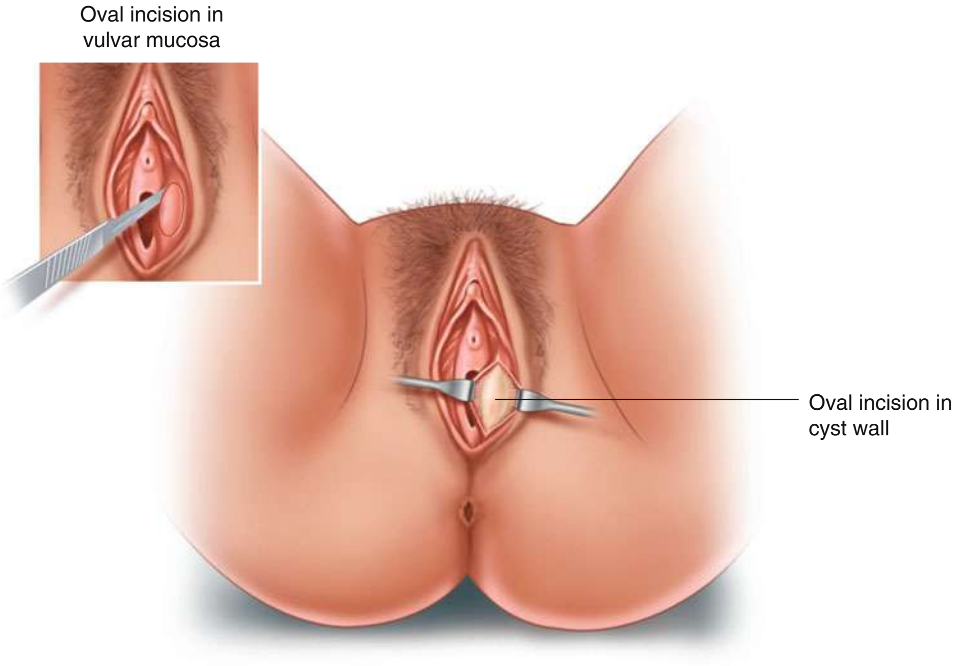
Oval incision in vulvar mucosa
Oval incision in cyst wall

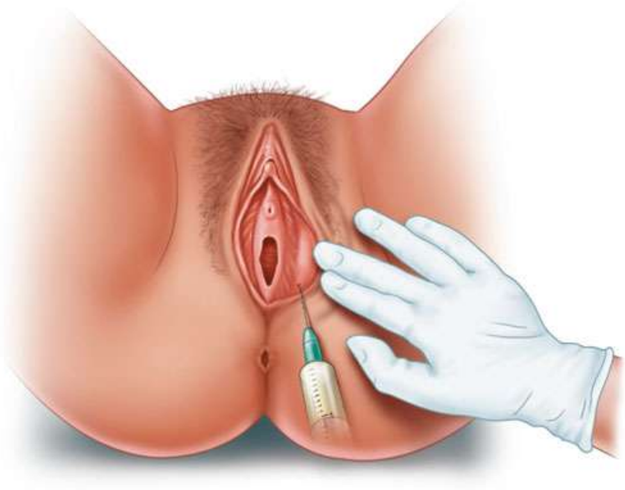

Fig. 83.5 Inject 1–4 mL of lidocaine at the planned site of incision

83.6 Iodoform Packing

1. Pack with iodoform packing, grasping the end of the packing material with a hemostat and inserting deeply within the cavity.
2. The cavity should be filled with the packing material and a small piece left exposed (for ease of retrieval during follow-up).
3. Clean the site and cover with gauze.

83.6.1 Word Catheter

- After culturing the abscess
  1. Place the Word catheter into the incision site as deep as possible (if the incision is too large, the catheter will slip out) (Fig. 83.6).
  2. Inflate the balloon of the Word catheter with 2–3 mL of saline or water injected into the hub with a needle and syringe (Fig. 83.7).
  3. Tuck the end of the Word catheter into the vagina for comfort purposes.
  4. The catheter should remain in place for 4 weeks to allow for epithelialization of the tract.

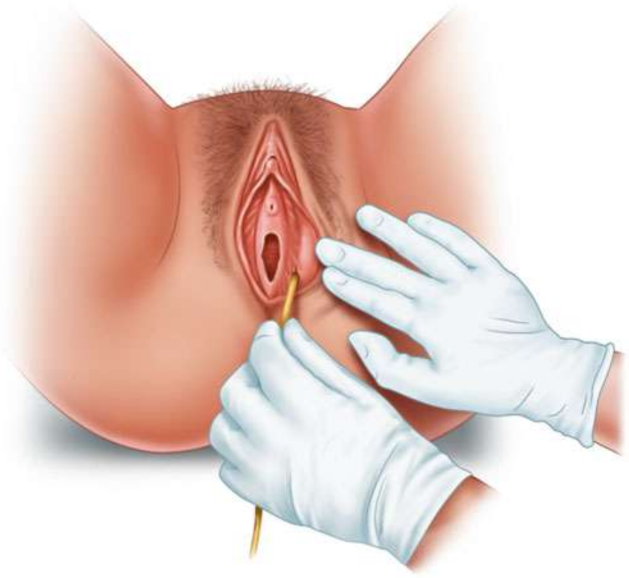

Fig. 83.6 Place the Word catheter into the incision site as deep as possible

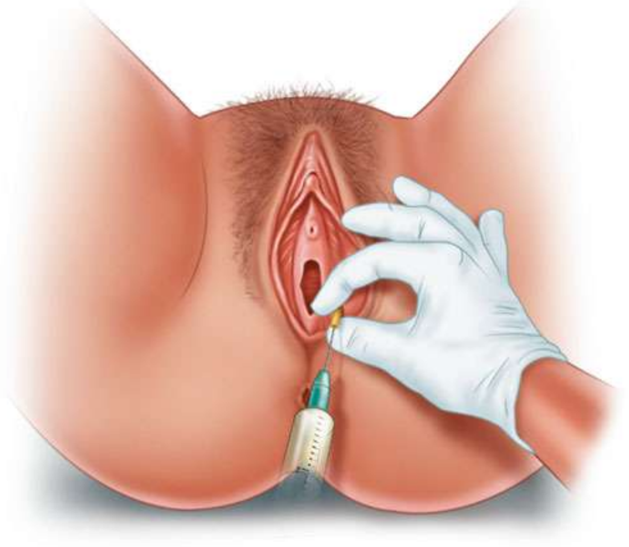

Fig. 83.7 Inflate the balloon of the Word catheter with 2–3 mL of saline or water injected into the hub with a needle and syringe

83.6.2 Jacobi Ring Catheter

1. Grasp one end of the Jacobi ring with a hemostat and pass it through the initial incision site.
2. At this time, use a hemostat to break loculations and culture the material for microbiology.
3. Pull the Jacobi ring through the abscess cavity (be careful not to pull the suture out of the catheter) and make a second incision to pull the catheter out through it.
4. The two ends of the catheter are then tied, forming a ring.

83.6.3 Nitrate Stick

- After sending the culture
  1. Take the silver nitrate stick and place it deep within the cyst/abscess cavity (Fig. 83.8).
  2. The patient is instructed to return in 48 h for removal of the remaining nitrate material and necrotic tissue and wound cleaning.
  3. The patient should be warned of side effects including pain, chemical burns of nearby tissue, edema, discharge, and scarring.

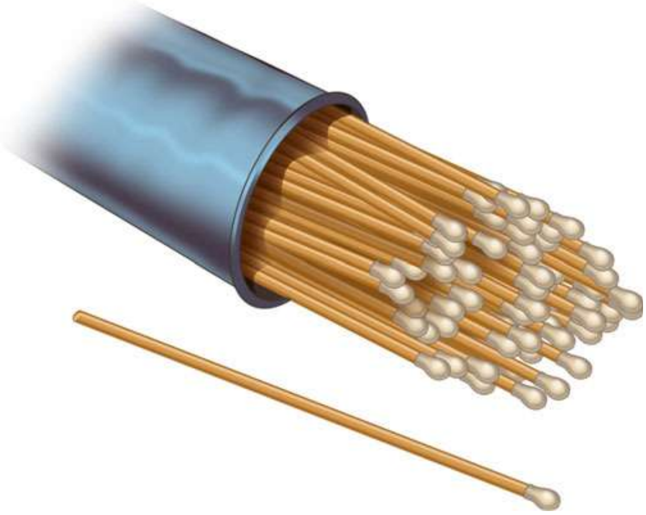

Fig. 83.8 Silver nitrate sticks

83.7 Aftercare

- Follow-up care is required after each of these procedures.
- A high-risk patient (e.g., diabetic) may need to be covered with broad-spectrum antibiotics.
- Pregnant women are also considered high risk and should be given antibiotics and followed closely.
- The patient should be instructed to remain on pelvic rest (nothing in vagina) while the catheter is in place, wear a pad owing to discharge, use sitz baths and analgesics for pain control, and follow up 1–2 days after the procedure.

Although a few other methods are available for draining Bartholin cyst/abscess including marsupialization, carbon dioxide laser vaporization, and excision, these methods are typically done by gynecologists and rarely, if ever, performed by an emergency department physician.

Selected Reading

Chen KT, Robert LB, Falk SJ. Disorders of Bartholin's gland. Available at: WWW.Uptodate.com.

Chen KT, Robert LB, Falk SJ. Word catheter placement for treatment of Bartholin's cysts and abscesses. Available at: WWW.Uptodate.com.

Gennis P, Li SF, Provataris J, et al. Jacobi ring catheter treatment of Bartholin's abscesses. Am J Emerg Med. 2005;23:414–5.

Vitaly KA, Mosquera C. Novel technique for management of Bartholin gland cysts and abscesses. J Emerg Med. 2008;36:388–90.

Wechter ME, Wu JM, Marzano D, Haefner H. Management of Bartholin duct cyst and abscesses: a systematic review. Obstet Gynecol Surv. 2009;64:395–404.

88.3 Materials and Medications (Fig. 88.1)

- 1 % Xylocaine with or without 1:200,000 epinephrine, 0.25 % bupivacaine solution. 8.4 % (1 mL/mL) sodium bicarbonate (optional)
- 18-, 25-, or 27-gauge needles, syringes up to 10 mL
- Sterile and nonsterile gloves, face shield
- Alcohol pads and povidone-iodine swabs

Fig. 88.1 Local anesthesia materials

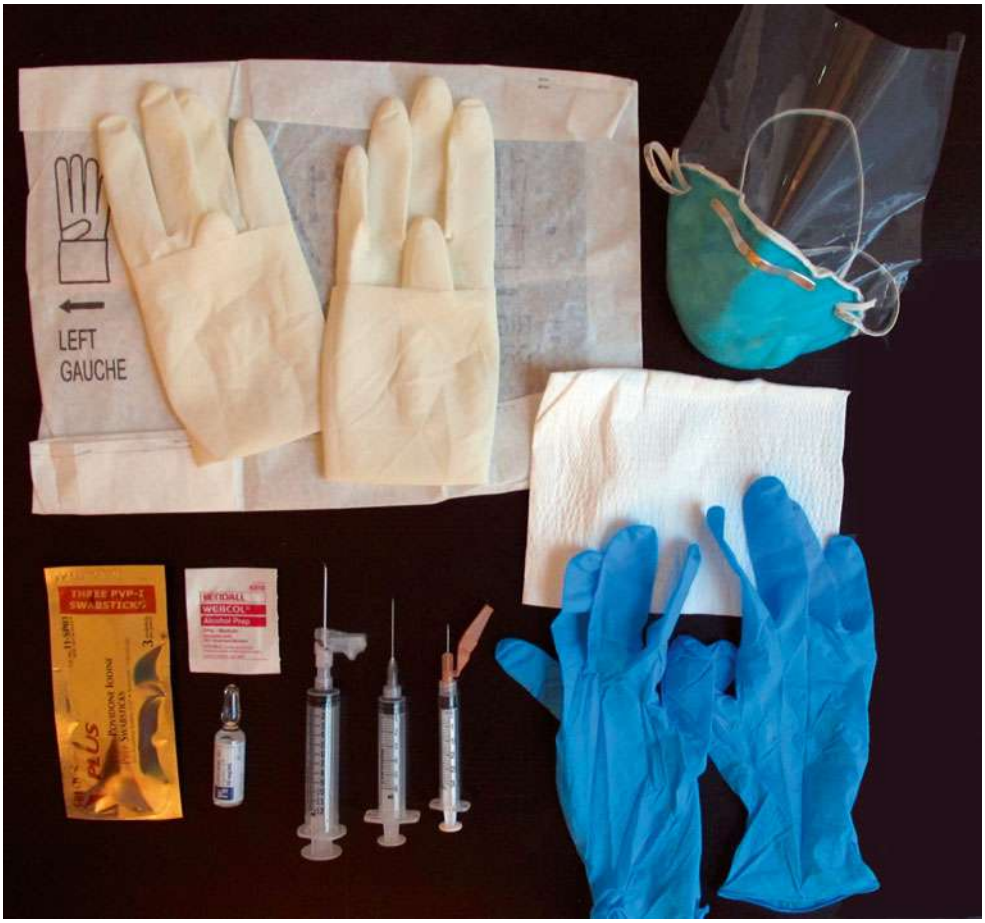
LEFT
GAUCHE
THREE PVP-I SWABSTICKS
PLUS
POVIDONE IODINE
KENDALL
WEBCOL
Alcohol Prep

Regional Anesthesia (Nerve Blocks)

89

Derek Ailes and Muhammad Waseem

89.1 Indications

- Repair of wounds where preserving anatomical landmarks or having precise anatomical alignment is important (e.g., vermillion border of lip)
- Pain control in dislocation or fracture reductions
- Incision and drainage of abscesses
- Burn and wound care
- Extensive or multiple lacerations (reduces total amount of local anesthetic needed)
- Foreign body removal

89.2 Contraindications

- Allergic to local anesthetic (see Chap. 88)
- History of coagulopathy or bleeding disorder
- Injection through infected tissue

D. Ailes, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: derekailes@yahoo.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_89

89.3 Materials and Medications (Fig. 89.1)

- Povidone-iodine, alcohol swabs
- Sterile gloves and drapes
- Local anesthetic solution (e.g., lidocaine, Marcaine with or without epinephrine)
- 18-gauge, 20- to 30-gauge needles 2 in. in length
- 22- to 24-gauge spinal needles
- Syringes up to 60 mL

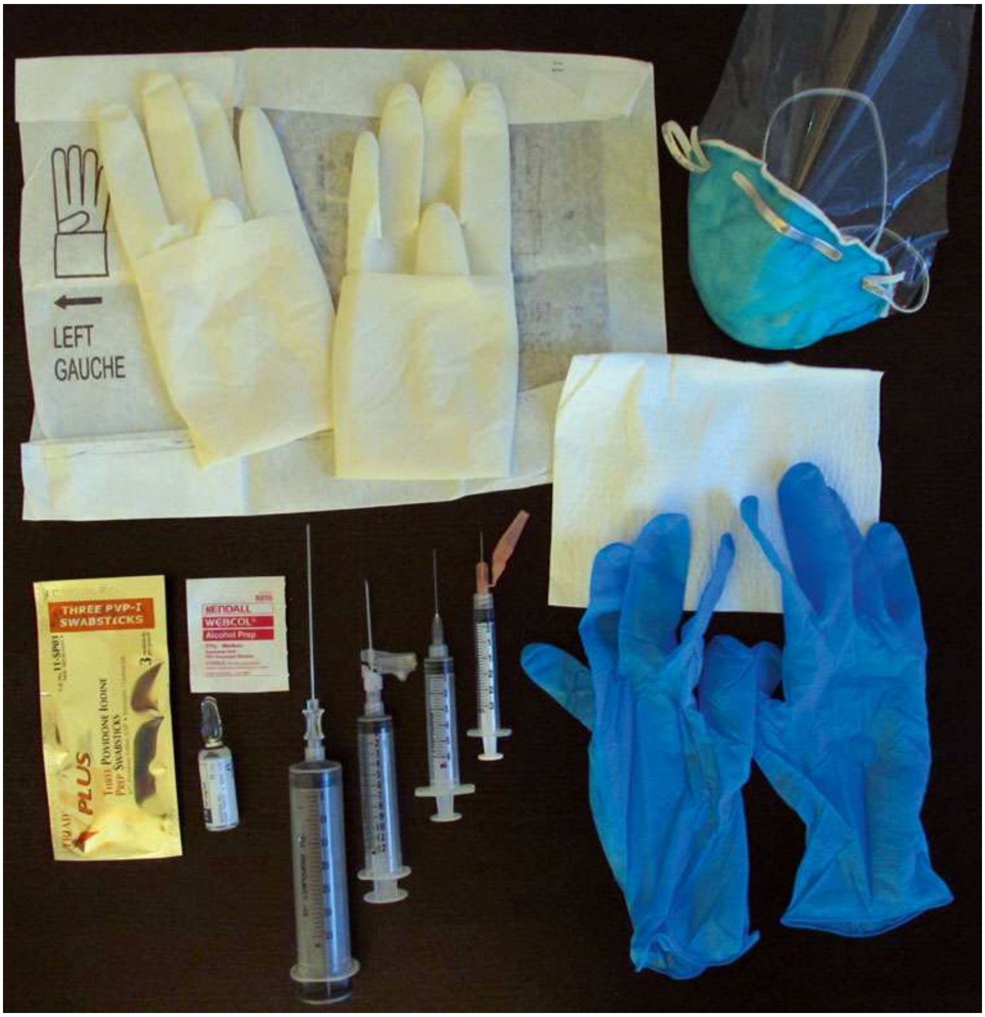
LEFT
GAUCHE
THREE PVP-I SWABSTICKS
PLUS
Three Povidone Iodine Prep Swabsticks
KENDALL
WEBCOL
Alcohol Prep

Fig. 89.1 Materials

89.4 Procedure: General Block

1. Obtain consent after explaining the risks of procedure including temporary paresthesias and expected duration of block. Perform neurological examination before procedure, documenting any preexisting deficits.
2. Position the patient comfortably, preferably supine, anticipating vasovagal response.
3. Identify landmarks for the block. Clean area, prepare with povidone-iodine, and surround with sterile drapes.
4. A small skin wheal of local anesthetic may be placed at the site of needle entry before block.
5. Insert the needle into the site while aspirating to ensure it is not in a vessel.
6. If paresthesia is elicited, withdraw the needle slightly allowing paresthesia to improve and inject.
7. Wait 5–15 min for the block to reach full effect.
8. Test for sharp sensation in the anesthetized area and document.

89.5 Complications

- Infection
- Hemorrhage
- Hematoma
- Allergic reaction
- Systemic toxicity (exceeded maximum dose or inadvertent injection into vasculature)
- Paresthesias, pain
- Intraneural injection causing ischemia
- Intra-arterial injection of epinephrine causing vasospasm and tissue ischemia

89.6 Pearls and Pitfalls

- Pearls
  - It is important to aspirate before injecting anesthesia when performing regional anesthesia because, unlike local techniques, the needle is deeper and in proximity to larger vessels.
  - Shooting pain and/or paresthesias occurs when the needle contacts the nerve. When this happens, withdraw the needle 1 mm and wait for the paresthesia to resolve before injecting.
  - Injury can occur to a limb or digit if the patient manipulates it before the anesthesia wears off. The patient should be cautioned not to use the affected area until motor and sensation return. If an extensive block was done, monitor the patient in the emergency department until return of baseline neurological function.
- Pitfalls
  - It is traditionally taught to avoid epinephrine-containing solutions in blocks in end-artery blocks (e.g., digital blocks). However, evidence for any vascular insufficiency and necrosis as a result is lacking in standard commercially available lidocaine with epinephrine preparations. They should, however, be avoided in patients with peripheral artery disease [1, 2].

89.7 Selected Specific Blocks

- Contraindications and materials are the same as general block.

89.7.1 Facial Blocks: Trigeminal Nerve (Fig. 89.2) [2, 3]

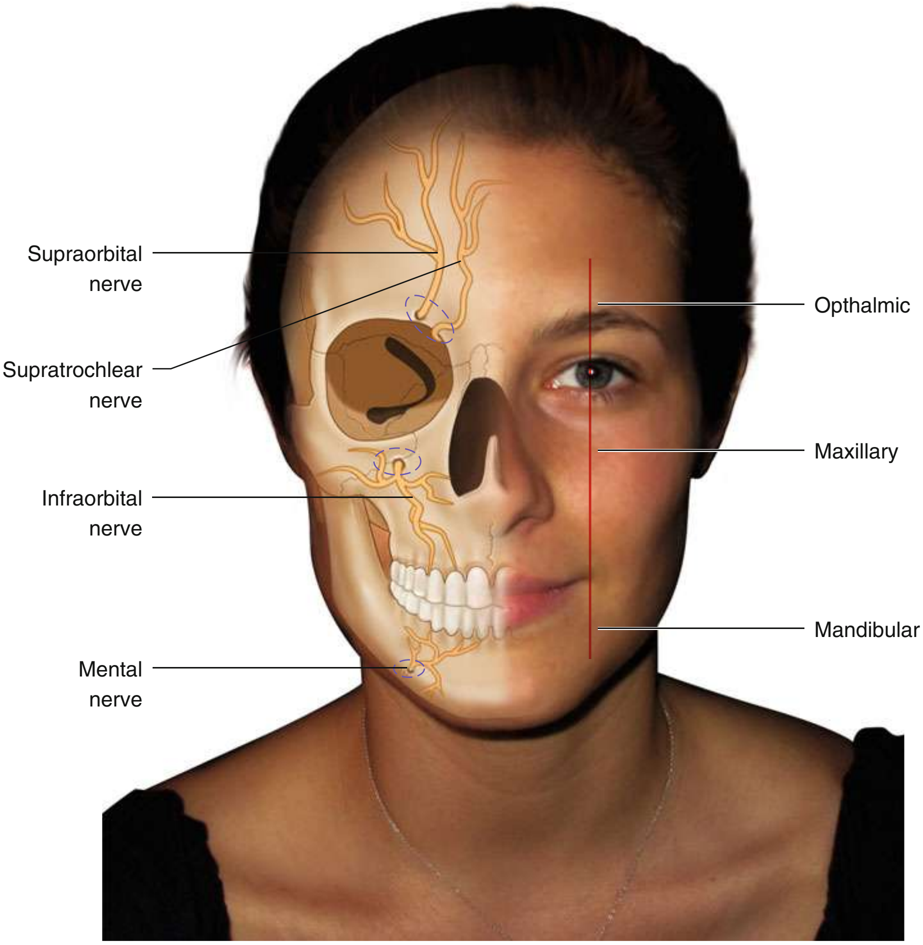
Supraorbital nerve
Supratrochlear nerve
Infraorbital nerve
Mental nerve
Opthalmic
Maxillary
Mandibular

Fig. 89.2 Vertical plane through the midposition of the pupil shows the position of the supraorbital foramen, infraorbital foremen, and mental foramen

89.7.1.1 **Supraorbital** (Fig. 89.3) and Supratrochlear Nerve Block

- Indications
  - To anesthetize the forehead from the orbital ridge to the vertex of the scalp. The supraorbital nerve emerges from the supraorbital foramen/notch and is a branch of the ophthalmic division of the trigeminal nerve. The supratrochlear nerve also is a branch of the ophthalmic division of the trigeminal nerve and exits through the superior medial aspect of the orbit.
- Procedure
  1. Inject local anesthetic solution over the midline of the forehead at eyebrow level.
  2. Inject a 25- or 27-gauge needle through skin wheal aimed laterally while injecting 3–5 mL local anesthetic subcutaneously. Stop infiltrating when the needle reaches the midline of the orbit.

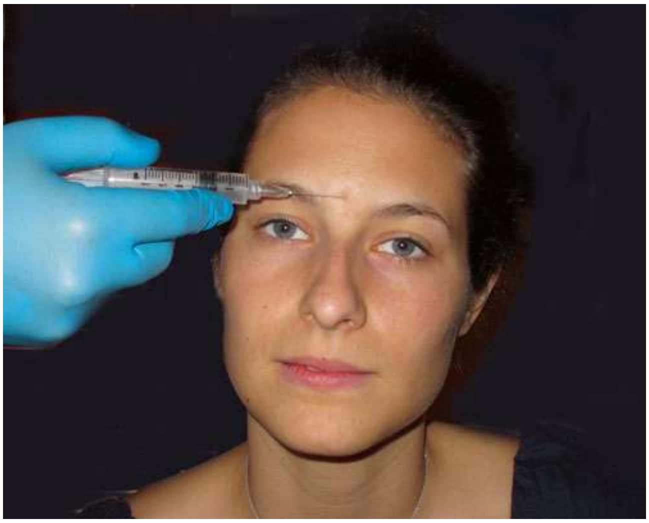

Fig. 89.3 Supraorbital nerve block

89.7.1.2 Infraorbital Nerve Block

- Indications
  - To anesthetize the medial cheek, upper lip, philtrum, skin between the lips and the nose, and nasal ala. The infraorbital nerve emerges from the infraorbital foramen and is a branch of the maxillary division of the trigeminal nerve. Anesthesia to the infraorbital nerve will also provide anesthesia to its terminus, the superior alveolar nerves.
- Procedure: Extraoral Approach (Fig. 89.4)
  1. Palpate the inferior orbital foramen in its midline position. The infraorbital nerve is often tender on palpation as it exits the foramen.
  2. Inject a 25- or 27-gauge needle just above the infraorbital foramen injecting 1–2 mL of local anesthetic.
  3. Take care not to inject into the foramen because there is an increased risk of intraneural injection.
  4. Hold a finger on the inferior orbital rim to avoid ballooning of the lower eyelid with injection.
  5. Intraoral approach is possible and preferred because it is less painful.

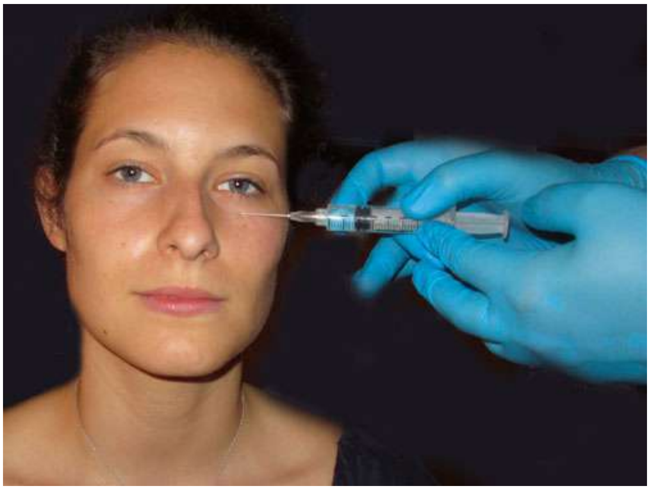

Fig. 89.4 Extraoral approach to the infraorbital nerve block

- Procedure: Intraoral Approach (Fig. 89.5) [3]
  1. Apply topical benzocaine or lidocaine gel to the point of insertion, which is the height of the mucobuccal fold over the first premolar, which is the site of insertion. Wipe off after 1–3 min.
  2. Palpate with the finger of the noninjecting hand over the inferior border of the inferior orbital rim. Retract the lip with the noninjecting hand.
  3. Using a long 25- to 27-gauge needle, with the bevel toward the bone, advance the needle at the insertion site toward the infraorbital foramen. Once the target is reached, aspirate and inject 1 mL of local anesthetic.
  4. Exert pressure on the foramen for 1 min after injection to force the anesthetic through the infraorbital foramen.
  5. If the needle is difficult to advance and the patient experiences pain on insertion, redirect the needle laterally and advance.
  6. If analgesia is attained for the lip but not the eyelid, the analgesia was placed inferior to foramen, and if analgesia is attained for the eyelid but not the lip, placement was superior to the foramen.

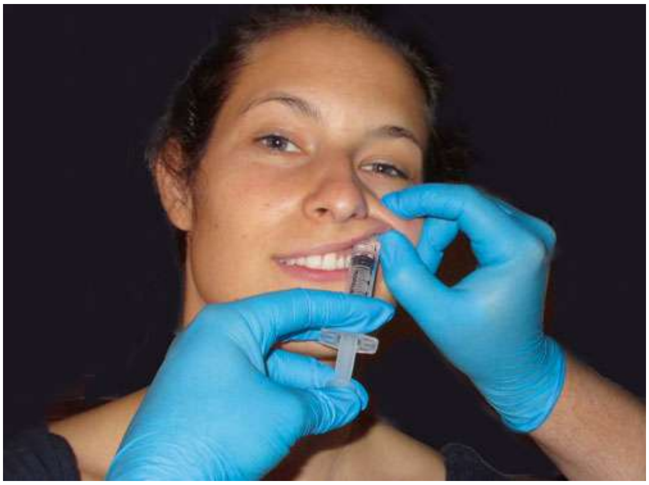

Fig. 89.5 Intraoral approach to the infraorbital nerve block

89.7.1.3 Mental Nerve Block

- Indications
  - To anesthetize the lower lip and chin and is especially useful in laceration repair at those sites. The mental nerve emerges from the mental foramen and is a branch of the mandibular division of the trigeminal nerve. Mental foramen lies in the vertical plane with the midpoint of the pupil and sits in the middle of the body of the mandible.
- Procedure: Extraoral Approach (Fig. 89.6)
  1. Inject local anesthetic solution over the identified location of the mental foramen, creating a skin wheal.
  2. Advance a 25- or 27-gauge needle through the skin wheel until the mandible is contacted, injecting 1–2 mL of local anesthetic.
  3. Intraoral approach is possible and preferred because it is less painful.
- Procedure: Intraoral Approach (Fig. 89.7) [3]
  1. Apply topical benzocaine or lidocaine gel to the point of insertion, which is the mucobuccal fold between the apices of the first and the second premolars. Wipe off after 1–3 min.
  2. Insert a 25- to 27-gauge needle, with the bevel toward the mandible, aimed toward the mental foramen.
  3. After advancing one-third the depth of the mandible and contacting the mandible, inject 1–2 mL of local anesthetic.
  4. By pressing firmly on the mental foramen for 2–3 min after the mental foramen has been blocked, an incisive nerve block is also created. This is useful if anesthesia to the lower anterior teeth is also desired.

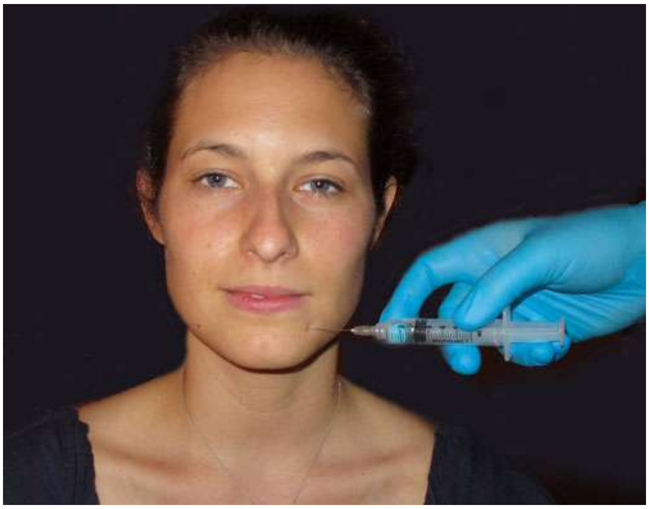

Fig. 89.6 Extraoral approach to the mental nerve block

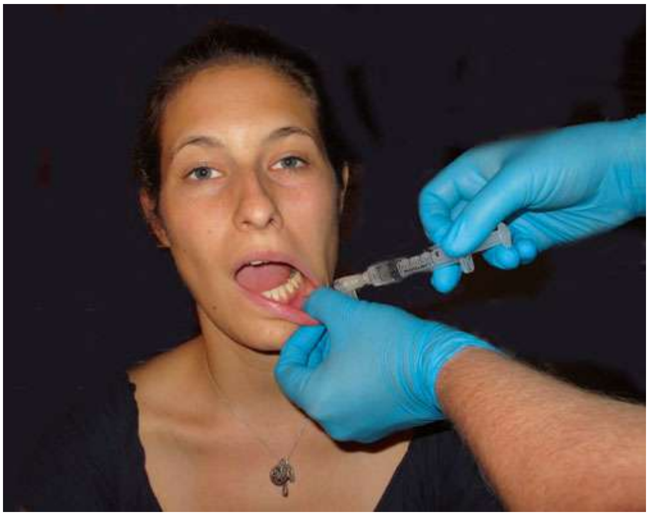

Fig. 89.7 Intraoral approach to the mental nerve block

89.7.2 External Ear Block (Fig. 89.8) [4]

- Indications
  - To anesthetize the entire external ear, excluding the external auditory canal and the concha
  - Especially useful in large lacerations of the ear and surrounding skin, hematoma evacuations, or incision and drainage of abscess.
- Procedure: Auricular Ring Block (Fig. 89.9) [4]
  1. Using a 25- to 27-gauge needle, insert the needle just inferior to the earlobe directing it toward the tragus.
  2. Aspirate and advance the needle superiorly subcutaneously injecting 3–4 mL of local anesthetic (Fig. 89.9, #1).
  3. Withdraw the needle without fully removing it and redirect it posterosuperiorly along the inferior posterior auricular sulcus, aspirating and injecting as before (Fig. 89.9, #2).
  4. Remove the needle and insert it just superior to the point of helix insertion into the scalp.
  5. Advance the needle and aspirate and inject in the direction of the tragus. Inject into the subcutaneous tissue while avoiding the ear cartilage (Fig. 89.9, #3).
  6. Withdraw and redirect the needle posteriorly and inferiorly toward the skin behind the ear, injecting as before (Fig. 89.9,#4).
  7. Beware of inadvertent cannulation of the superficial temporal artery, which crosses the zygomatic arch and crosses medial to the ear. If the artery is violated, it requires 20–30 min application of firm pressure.

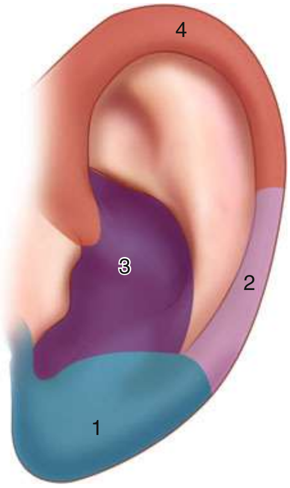

Fig. 89.8 Auricular block anesthetizes four nerves that innervate the auricle. *1* Great auricular nerve, *2* lesser occipital nerve, *3* auricular branch of vagus nerve, *4* auriculotemporal nerve

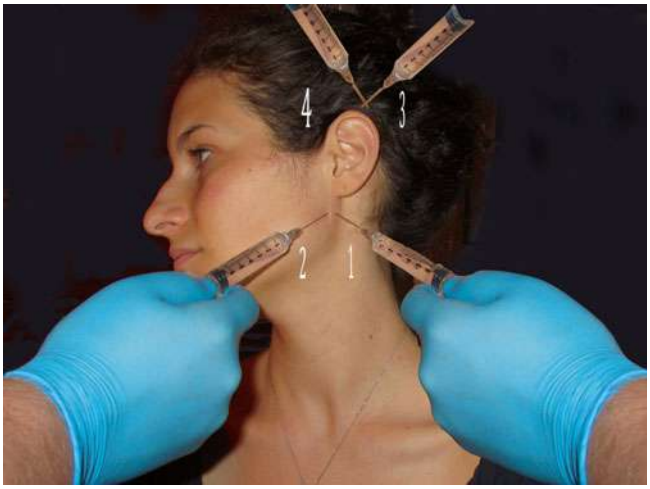

Fig. 89.9 Auricular ring block technique

89.7.3 Wrist Block (Fig. 89.10) [3, 5, 6]

- Indications
  - To anesthetize the hand in preparation for laceration repair, fracture or dislocation reduction, or pain relief

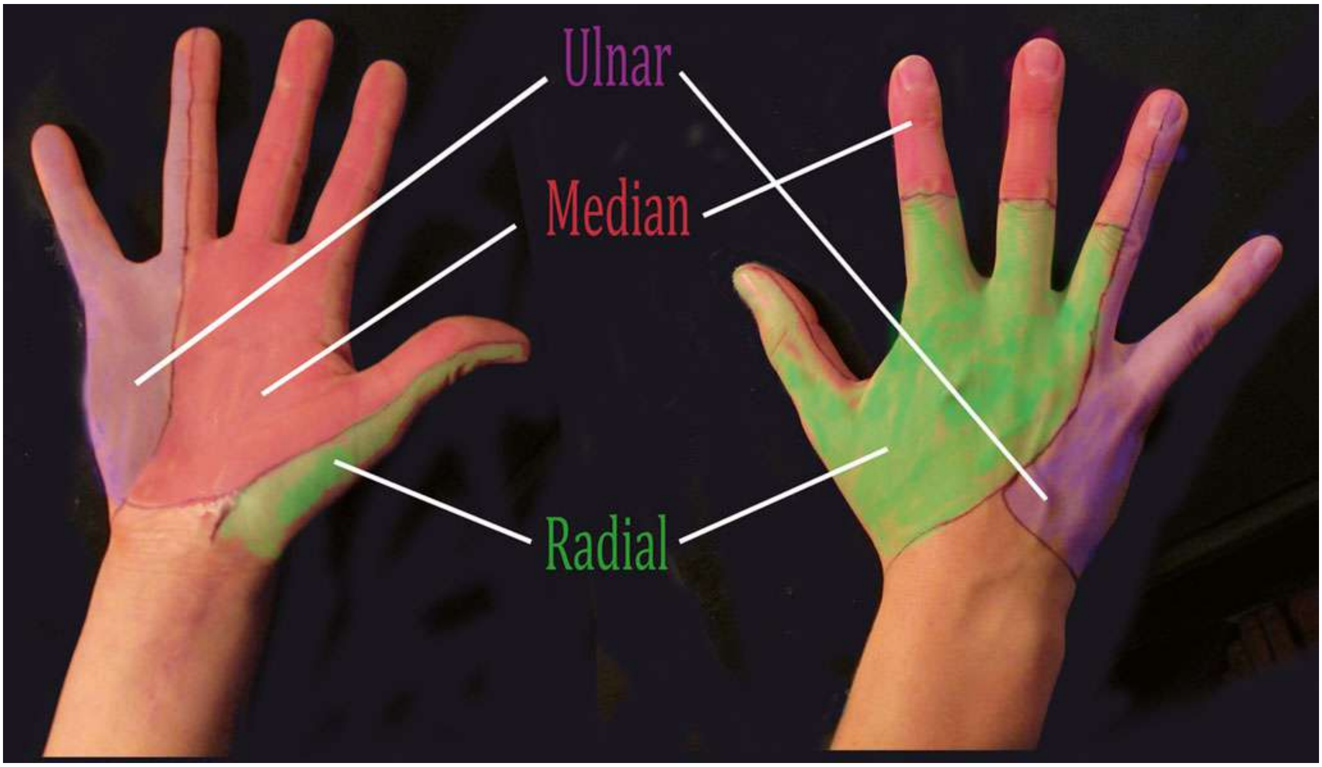

Fig. 89.10 Nerve distribution in the hand

89.7.3.1 Wrist Block: Median Nerve (Fig. 89.11)

- Procedure
  1. Position the patient supine with the palmar surface of the hand face up.
  2. Have the patient make a fist and slightly flex the wrist so that the palmaris longus and flexor carpi radialis tendons become prominent.
  3. Create a skin wheal of local anesthetic between the two tendons between the proximal skin crease and the distal skin crease at the wrist. Alternatively, anesthetic can be injected in line with the ulnar styloid process at the proximal skin crease.
  4. The palmaris longus tendon is absent normally in 10–20 % of the population. In this case, inject over the midpoint of the proximal skin crease at the level of the styloid process.

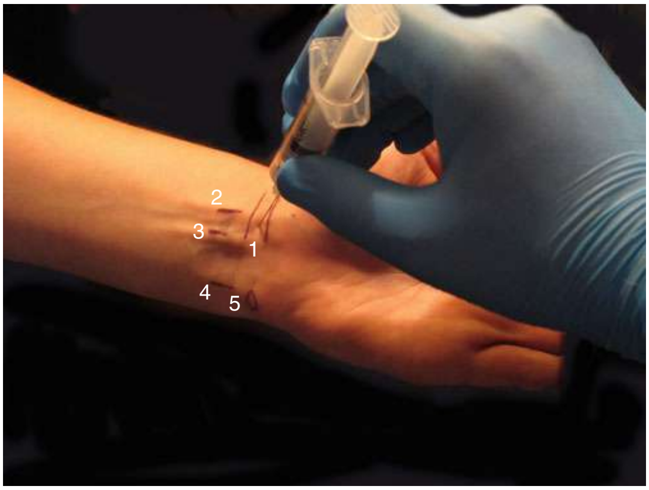

Fig. 89.11 Median nerve block at the wrist: *1* proximal and distal wrist creases radial artery, *2* flexor carpi radialis tendon, *3* palmaris longus tendon, *4* ulnar artery, *5* styloid process

89.7.3.2 Wrist Block: Radial Nerve (Figs. 89.12 and 89.13) [3, 7]

- Procedure
  1. To block the radial nerve, block multiple peripheral branches on the dorsal and the radial aspects of the lateral wrist.
  2. A field block in and around the anatomical snuffbox may be performed, requiring roughly 5–6 mL of local anesthetic. This anesthetizes the terminal branches of the radius arising from the forearm.
  3. Position the patient's palm face up and inject a skin wheal to the area 1 mm lateral to the radial pulse and in line with the proximal wrist crease. Inject 2 mL of local anesthesia with a 25- to 27-gauge needle. This anesthetizes the terminal trunk of the radial nerve.

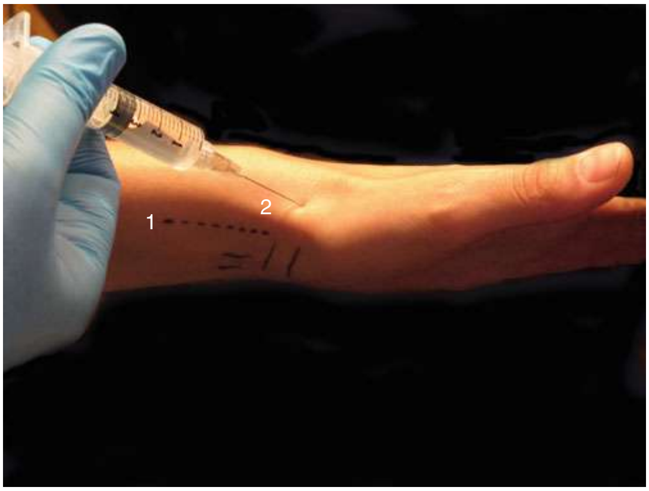

Fig. 89.12 Radial nerve block at the wrist: *1* radial artery, *2* anatomical snuffbox

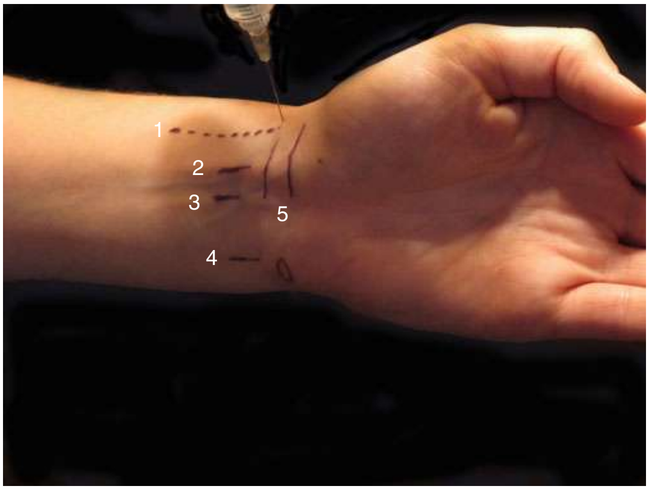

Fig. 89.13 Radial nerve block at the wrist: *1* radial artery, *2* flexor carpi radialis tendon, *3* palmaris longus tendon, *4* ulnar artery, *5* proximal and distal wrist creases

89.7.3.3 Wrist Block: Ulnar Nerve (Fig. 89.14)

- Procedure
  1. Position the patient with the hand face up.
  2. Palpate the styloid process of the ulna and pisiform and find the ulnar artery pulse.
  3. Create a skin wheal of local anesthetic between lateral to the ulnar artery and medial to the flexor carpi ulnaris tendon just proximal to the styloid process. This is at the level of the proximal wrist crease.

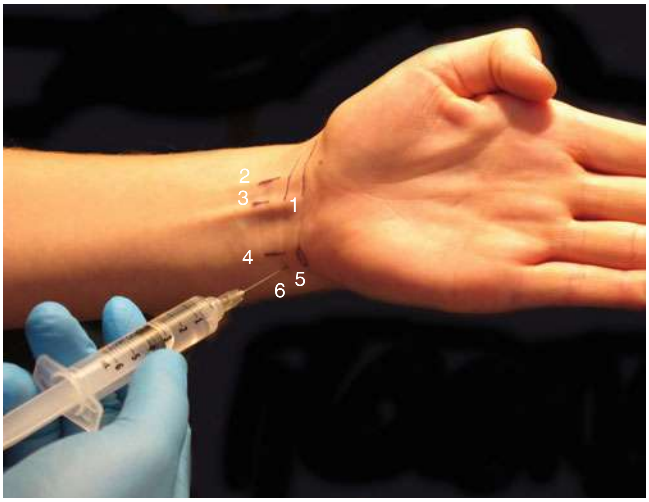

Fig. 89.14 Ulnar nerve block at the wrist: *1* proximal and distal wrist creases, *2* flexor carpi radialis tendon, *3* palmaris longus tendon, *4* ulnar artery, *5* styloid process, *6* flexor carpi ulnaris

89.7.4 Digital Nerve Blocks: Ring, Web Space, and Tendon Sheath (Fig. 89.15) [3, 8, 9]

- Indications
  - To anesthetize the digits in preparation for laceration repair, nail bed repair, joint reduction, or pain relief

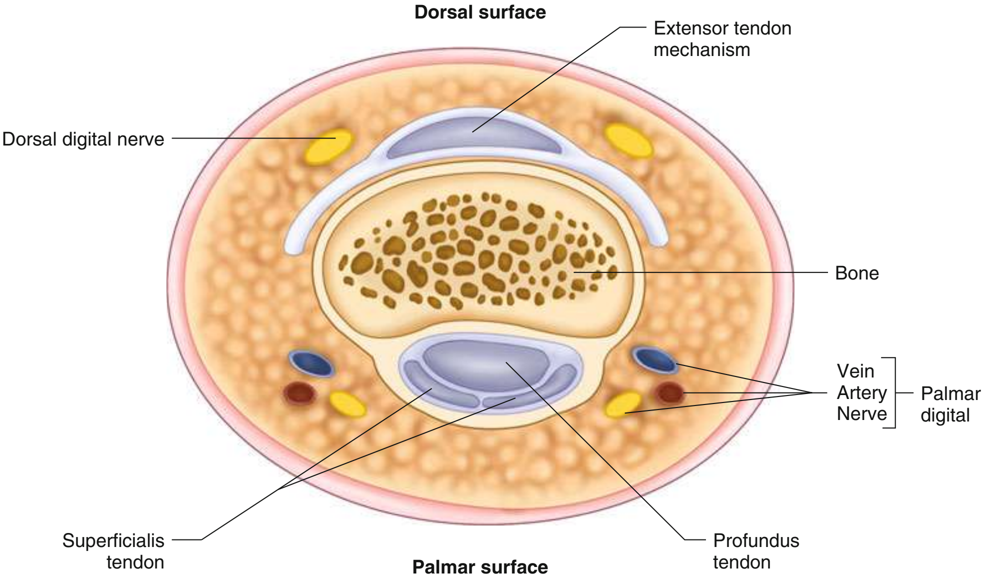
Dorsal surface
Extensor tendon mechanism
Dorsal digital nerve
Bone
Vein
Artery
Nerve
Palmar digital
Superficialis tendon
Palmar surface
Profundus tendon

Fig. 89.15 Cross section of finger

89.7.4.1 **Ring Block** (Fig. 89.16) [3]

- Procedure
  1. Insert a 25-gauge needle on the dorsal surface of the proximal phalanx of the digit to be anesthetized. Inject 1 mL along the dorsal surface and withdraw the needle.
  2. Reinsert the needle again perpendicular to the last injection and running on the lateral surface of the phalanx. Inject 1–1.5 mL of local anesthetic to just past the phalanx base.
  3. Repeat the injection in the same fashion on the medial aspect of the phalanx.
  4. Do not inject more than 5 mL into a digit.
  5. Toe blocks are similar to finger ring blocks, except that the great toe requires plantar surface injection as well, owing to its unique nerve supply.

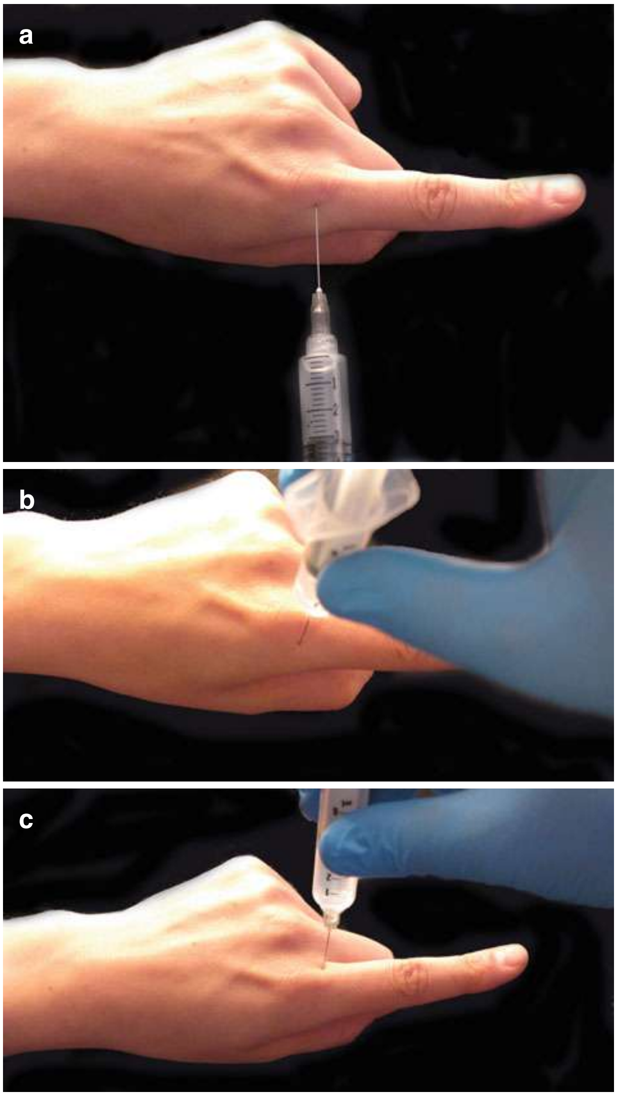

Fig. 89.16 Dorsal surface injection of digital nerve ring block. Digital nerve block is performed by injecting onto the (**a**) dorsal, (**b**) lateral, and (**c**) medial surfaces of the proximal phalanx

89.7.4.2 **Web Space Digital Block** (Fig. 89.17) [9]

- Procedure
  1. Have the patient abduct the fingers.
  2. Palpate the metacarpophalangeal joint and then insert a 25- to 27-gauge needle into the lateral web space subcutaneously, directing it dorsally. Aspirate then inject 1 mL of local anesthetic.
  3. Withdraw the needle but before the exiting skin, redirect toward the palmar aspect until the tip is next to the metacarpophalangeal joint, and inject 1 mL of local anesthetic.
  4. Repeat the procedure on the medial web space of the digit. Each digit blocked requires injection on both the lateral and the medial web spaces.

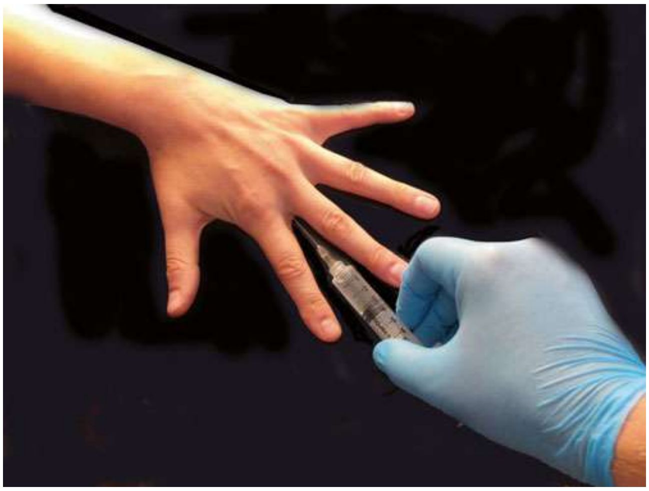

Fig. 89.17 Web space approach to the digital block, requiring injection on the medial and lateral web spaces for a blocked digit

89.7.4.3 Intrathecal Digital Block: Flexor Tendon Sheath (Fig. 89.18) [8]

- Procedure
    1. Inject anesthetic directly into the flexor tendon sheath. Palpate on the palmar surface over and proximal to the metacarpophalangeal joint. Gentle flexion of digit may better reveal the sheath. Have the patient abduct the fingers.
    2. Insert a 25-gauge needle at a 45° angle to the skin and along the long axis of the digit directly into the flexor tendon sheath at the level of the distal skin crease.
    3. Inject 2 mL of local anesthetic. The anesthetic should flow freely if it is in the sheath. If it does not, it is likely in the tendon and should be withdrawn slightly.
    4. Contraindications to intrathecal block are local infection and preexisting flexor tendon injury.
    5. Risk of tenosynovitis; sterilize the skin before introducing the needle.
    6. If laceration has involved the tendon, anesthetic may leak from the wound.

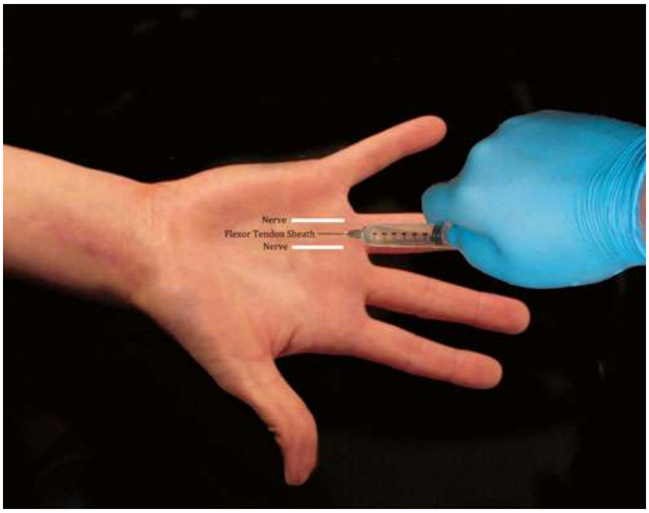

Fig. 89.18 Intrathecal/tendon sheath approach to the digital block

Acknowledgment The authors would like to thank Katy Howard for serving as the subject in many of the photographs in this chapter.

References

1. Muck AE, Bebarta VS, Borys DJ, Morgan DL. Six years of epinephrine digital injections: absence of significant local or systemic effects. Ann Emerg Med. 2010;56:270–4.
2. Waterbrook AL, Germann AC, Southall JC. Is epinephrine harmful when used with anesthetics for digital blocks? Ann Emerg Med. 2007;50:472–5.
3. Reichman EF, Simon RR, editors. Emergency medicine procedures. New York: McGraw-Hill Medical; 2004. p. 961–3.
4. Benko K. Fixing faces painlessly: facial anesthesia in emergency medicine. Available at: Emergency Medicine Practice (ebmedicine.net). 2009;11.
5. Rosh AJ. Ear anesthesia. Medscape reference. Available at: http://emedicine.medscape.com/article/82698-overview#a15.
6. Butterworth JF. Atlas of procedures in anesthesia and critical care. Philadelphia: WB Saunders; 1992. p. 160–4.
7. Brown DL. Atlas of regional anesthesia. Philadelphia: WB Saunders; 1992. p. 52.
8. Morrison WG. Transthecal digital block. Arch Emerg Med. 1993;10:35–8.
9. Mueller J, Davenport M. Digital nerve block (web space and tendon sheath). New York: McGraw-Hill's Access Emergency Medicine. Available at: http://www.accessemergencymedicine.com/videosPDF/DigitalNerveBlock.pdf.

Wound Management and Hemostasis

90

Rich Teitell and Muhammad Waseem

90.1 Indications for Primary Wound Repair

- Wounds that have been sustained less than 8 h before presentation may be closed primarily.
- If there is suspicion of broken glass or the presence of a foreign body, the patient should receive an x-ray before primary wound repair. Of note, x-ray has poor sensitivity for radiolucent foreign bodies such as wood and plastic.
- Wounds that are free of foreign bodies or show no gross evidence of infection or dead tissue may be closed primarily.

90.2 Contraindications for Primary Wound Repair

- History must address the factors that predispose a wound to greater risks of infection and not implement primary wound closure in such wounds that have had a duration greater than 8–12 h or contamination with saliva, stool, or foreign matter and a wound that has been sustained by blunt or crush mechanism.
- When wounds are highly contaminated with debris or if devitalized tissue is present, the wound should be left open for 3–4 days and then reevaluated.
- After a proper physical examination has been performed that assesses distal pulse and sensory and motor function of the surrounding region and distal extremity, if any of these are compromised, appropriate orthopedic and/or vascular consultation should be sought before wound closure, especially in the setting of a suspicion for open fracture.
- Consider delayed closure in wounds over joints.
- Wounds that have a significant loss of skin should be left to heal by secondary intention.

90.3 Procedure for Wound Closure and Hemostasis

1. Wound should be thoroughly prepared with Betadine (10 % povidone with 1 % iodine), which has a rapid onset with greatest antimicrobial effect.
2. Skin surrounding the wound should be covered with sterile drapes (Fig. 90.1) so as to reduce additional contamination within the open wound.
3. The most important determinant for preventing wound infection is adequate high-pressure irrigation.
4. Normal saline is the best solution for irrigation. Use pressures above 7 psi and an 18-gauge needle with a syringe of 30 mL or more; this gauge needle will achieve pressure greater than 7 psi.
5. It has been suggested that tap water irrigation may be just as effective. Because this approach will reduce costs, it can be considered.
6. Sharp debridement should be performed when foreign matter is found in an open wound that does not easily rinse out with saline. In this circumstance, consider wound closure by secondary intention or delayed primary closure.
7. At this point, wounds that do not meet criteria for primary closure must be left open, covered with a sterile dressing, and reevaluated in 3–4 days for consideration of delayed primary closure versus healing by secondary intention.
8. Wounds that do meet criteria as previously discussed for primary closure should be closed with either sutures or staples.
9. Staples are the preferred method of closure of scalp wounds and can be used under low wound tension.

R. Teitell, MD
Department of Emergency Medicine, Waterbury Hospital,
Waterbury, CT, USA
e-mail: rjt0013@gmail.com

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_90

10. The angle of the stapler is critical and must be positioned perpendicular to the intended area of delivery. A two-person approach can be utilized for optimal placement of the staple, using forceps to approximate the edges of the tissue.
11. Synthetic or monofilament sutures have a lower rate of infection than braided sutures (for closure techniques with sutures, see Chap. 92).
12. Dermabond should be utilized only for the most superficial linear lacerations and confers the benefit of efficiency of time and less discomfort to the patient. However, it should be mentioned that Dermabond does not offer any statistically significant difference in cosmesis when compared with suturing or other adhesives (see Chap. 93).
13. When a wound presents with deep underlying tissue involvement and direct bleeding is evident, direct pressure may be used to achieve hemostasis within a wound. Surgicel may also be applied with pressure to achieve control of small capillary and venous oozing (Fig. 90.2).
14. When a persistent bleed is visible within a wound that does not respond to direct pressure, a figure-of-eight suture (Fig. 90.3) may be used.
15. If a bleeding vessel is visible within a wound, an alternate method of hemostasis may be achieved by utilizing a hemostat to clamp the vessel and then tying an absorbable suture (Fig. 90.4).

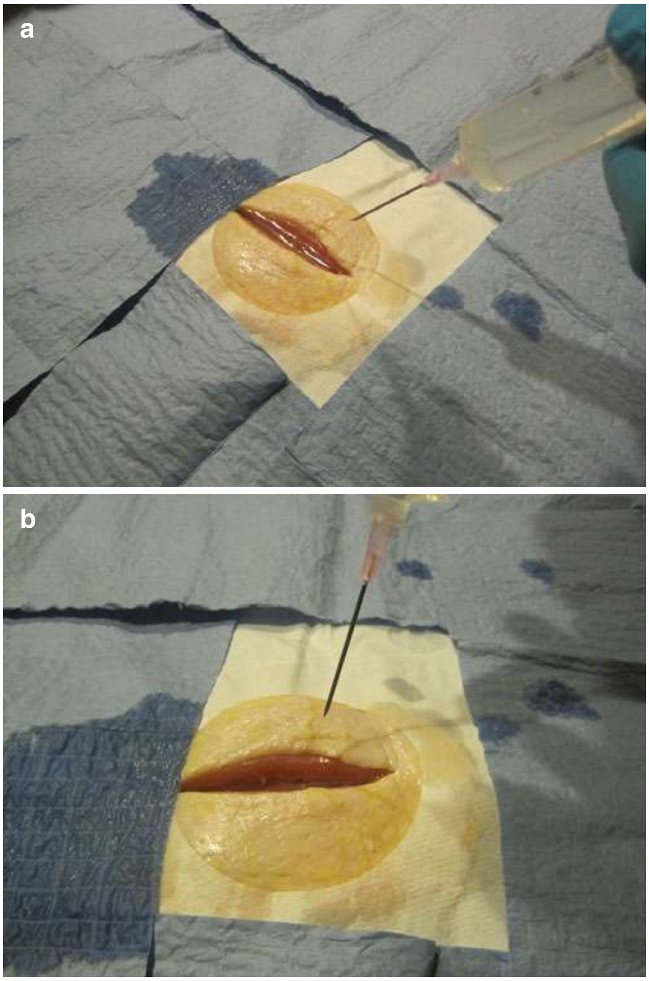

Fig. 90.1 (**a**) High-pressure irrigation using an 18-gauge needle (**b**)

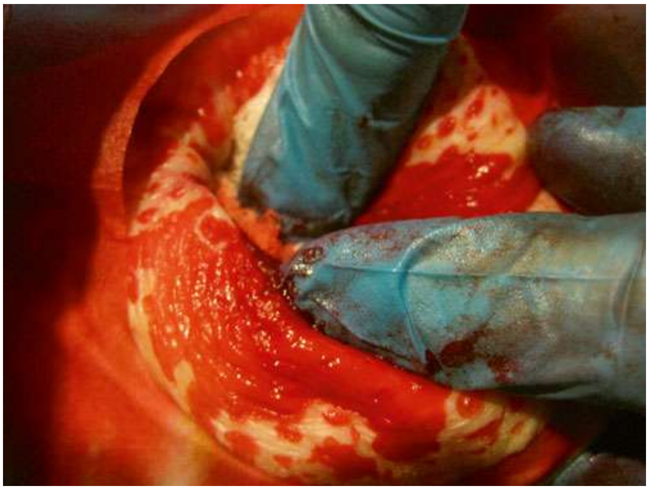

Fig. 90.2 Direct pressure with or without Surgicel may be used to achieve hemostasis within a wound and also to achieve control of small capillary and venous oozing

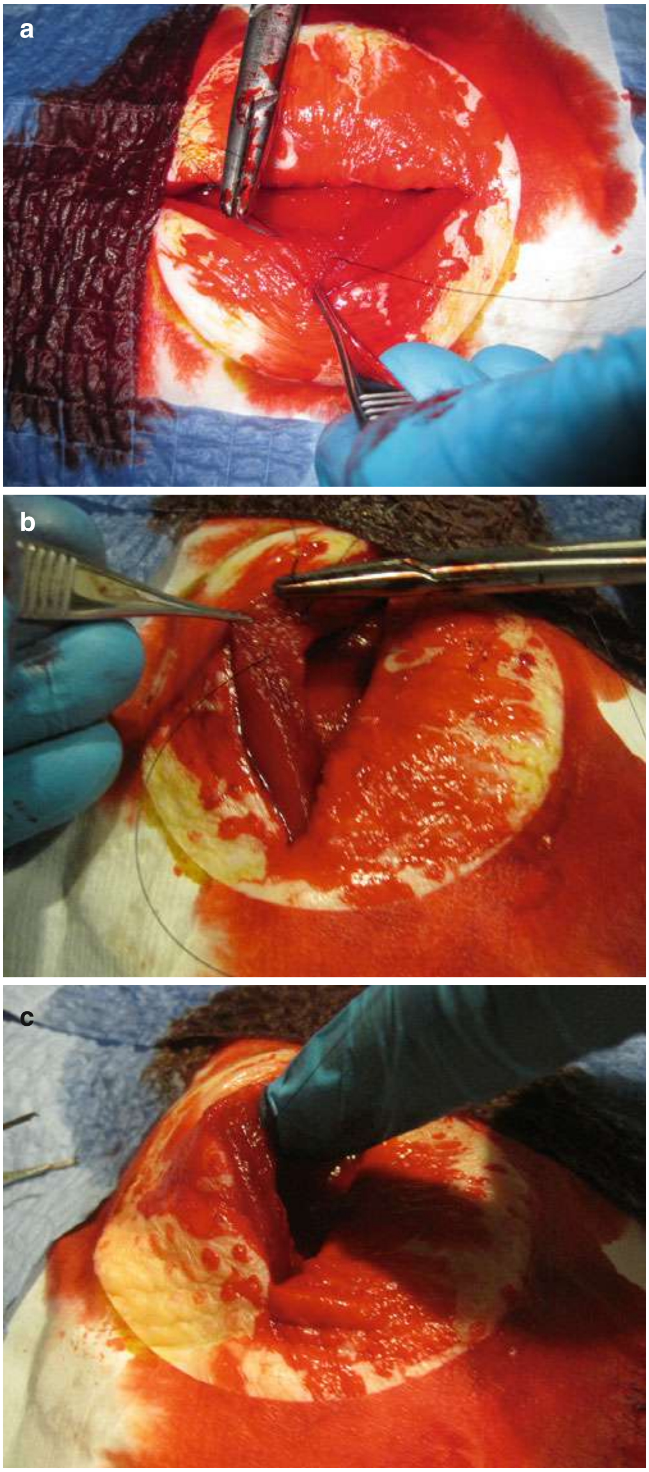

Fig. 90.3 (**a**) The location of the bleeding vessel in the subcutaneous tissue is identified and an absorbable suture is threaded through the region. (**b**) In a diagonal direction, a second suture is placed, which when tied will form the shape of an eight. (**c**) Once tied, the suture forms a figure eight in the tissue and hemostasis is achieved

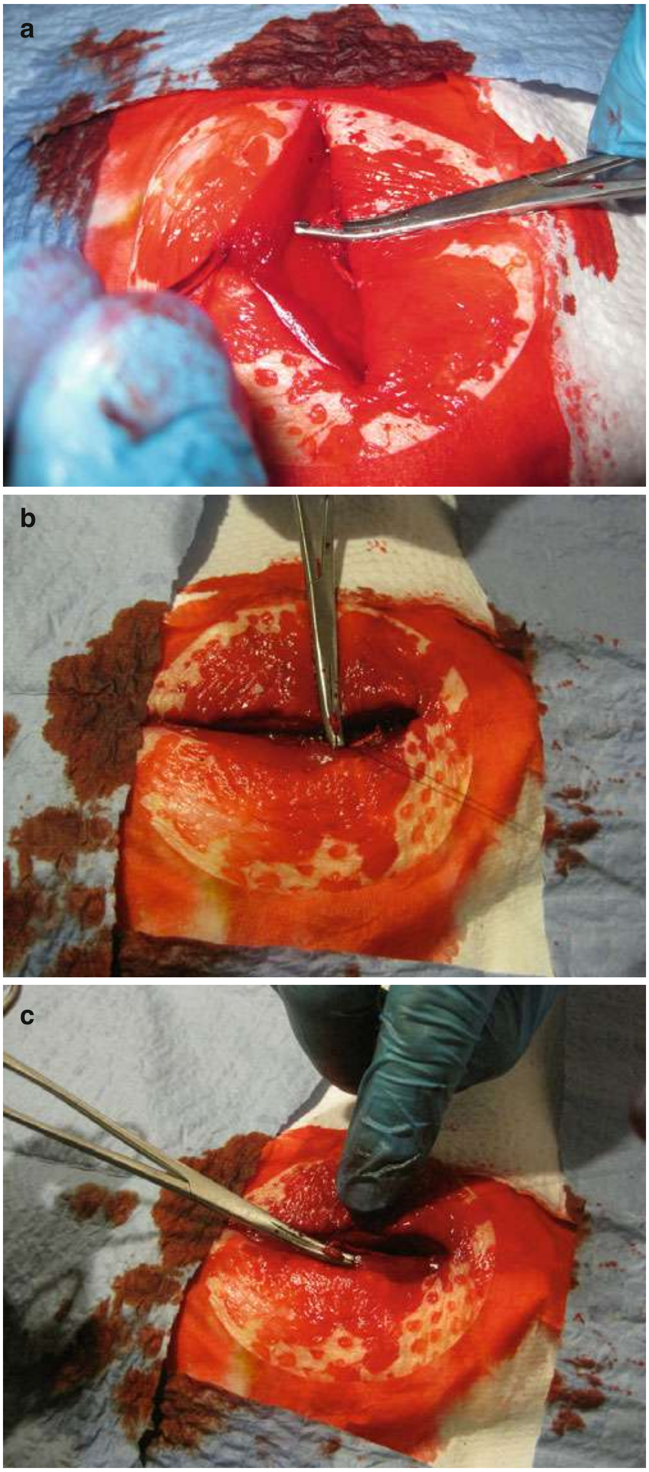

Fig. 90.4 (**a**) The culprit vessel is identified and clamped with a hemostat. (**b**) An absorbable suture is wrapped around the vessel inferior to the hemostat. (**c**) Once the first tie is placed being driven down with index finger, the hemostat may be removed and the knot can be completed for a total of three or four passes

90.4 Complications

- Hematoma formation
- Infection
- Scar formation

90.5 Pearls and Pitfalls

- Pearls
  - Antibiotic coverage should be employed judiciously depending on the wound and location.
  - Wounds to the face that are not heavily contaminated generally do not require antibiotics because the face and scalp have a significant vascular supply.
  - Patients with diabetes have higher incidences of wound infection, and the physician should consider prophylactic antibiotics in such individuals.
  - Wounds that are clean-contaminated and sutured in the emergency department should have a nonadherent dressing applied for 48 h until epithelialization has occurred. After that, the dressing may be removed, and soap and water gentle washing should commence to keep area clean. Only small amounts of bacitracin should be applied to wound in the interim period before sutures are removed.
- Pitfalls
  - Wounds located on the extremities have greater incidences of infection than those on the trunk or head/face.

Selected Reading

Blankenship RB, Baker T. Imaging modalities in wounds and superficial skin infections. Emerg Med Clin North Am. 2007;25:223–34.

Dimick AR. Delayed wound closure: indications and techniques. Ann Emerg Med. 1988;17:1303–4.

Edlich RF, Rodeheaver GT, Morgan RF, Berman DE, Thacker JG. Principles of emergency wound management. Ann Emerg Med. 1988;17(12):1284–302.

Farion K, Osmond MH, Hartling L, et al. Tissue adhesives for traumatic lacerations in children and adults. Cochrane Database Syst Rev. 2002;(3):CD003326.

Hall S. A review of the effect of tap water versus normal saline on infection rates in acute traumatic wounds. J Wound Care. 2007;16:38–41.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 6th ed. Philadelphia: Mosby-Elsevier; 2006. p. 702–3.

Ritchie AJ, Rocke LG. Staples versus sutures in the closure of scalp wounds: a prospective, double-blind, randomized trial. Injury. 1989;20:217–8.

Singer AJ, Hollander JE, Quinn JV. Evaluation and management of traumatic lacerations. N Engl J Med. 1997;337:1142–8.

91

Burn Care

Thomas Parry and Jeffrey Pepin

Burns result from exposure to heat, caustic chemicals, electricity, or radiation. Damage to the natural barrier provided by skin results in rapid fluid losses and risk of infection. Permanent scarring is a common long-term complication. They may also be complicated by sensory deficits due to loss of nerve connections and limb loss due to circulatory compromise.

91.1 Burn Description (Fig. 91.1)

- First degree or superficial: burn that remains confined to the epidermis (e.g., sunburn)
- Second degree or partial thickness: burn that extends into the dermis (e.g., blistering scald burns)
- Third degree or full thickness: burn involving the entire depth of the dermis and epidermal appendages

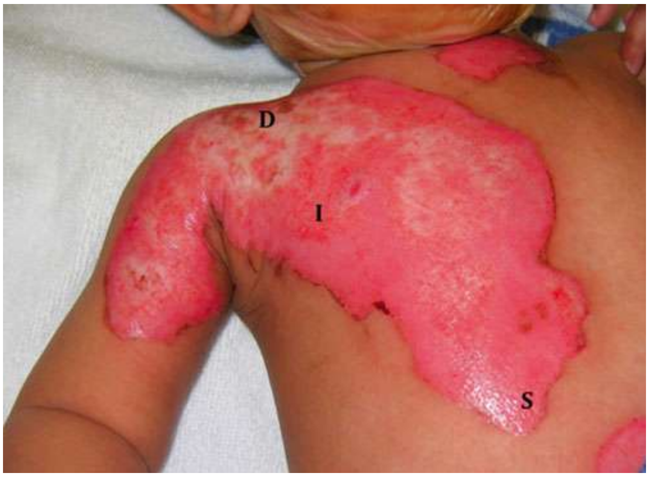

Fig. 91.1 Tea scald over the chest and shoulder of a child showing heterogeneity of burn depth. *D* deep (second or third degree), *I* intermediate (second degree), *S* superficial (first degree) (Reproduced from: Enoch S, Roshan A, Shah M. Emergency and early management of burns and scalds. *BMJ*. 2009;338:b1037, with permission from BMJ Publishing Group Ltd.)

T. Parry, MD (✉)
Department of Emergency Medicine, Lincoln Hospital and Mental Health Center, New York, NY, USA
e-mail: Thomas.parry@nychhc.org

J. Pepin, MD
Department of Emergency Medicine, University of Minnesota Medical Center Fairview, Minneapolis, MN, USA
e-mail: medicone76@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_91

91.2 Indications

- Superficial, deep second degree, and third degree.
- First-degree burns generally need only supportive care.

91.3 Materials and Medications

- Appropriate analgesia
- Antibiotic ointment
- Silver sulfadiazine (avoid on face)
- Sulfamylon for ears
- Alternative agents include bacitracin or polymyxin B ointments
- Cleansing solution such as chlorhexidine prep and water
- Basin
- Petroleum gauze
- Several 4×4 gauze pads
- Rolled gauze
- Tape
- Gloves

Airway, breathing, and circulation and cervical spine should be immediately assessed before any burn wound management. It is important to remember that any patient suspected of an inhalation injury or carbon monoxide poisoning should receive 100 % humidified oxygen until carboxyhemoglobin returns to normal. Inhalation injuries may not present themselves until after fluid resuscitation has already been started, so early intubation is recommended in these cases.

91.4 Sizing

Total body surface area (TBSA) of body parts is estimated by multiples of 9 (Rule of Nines)

- Adults
  - Head and neck: 9
  - Arms: 9 each
  - Legs: 18 each
  - Trunk: 18 front and 18 back
  - Perineum and palms: 1
- Infants/children
  - Head and neck: 18
  - Arms: 9 each
  - Legs: 14 each
  - Trunk: 18 front and 18 back

A second way to estimate TBSA in smaller burns is by using the palm surface area: using the patient's palm as a guide, each palmar surface equals 1 %.

A third way to estimate TBSA is via the use of the Lund and Browder chart (Fig. 91.2).

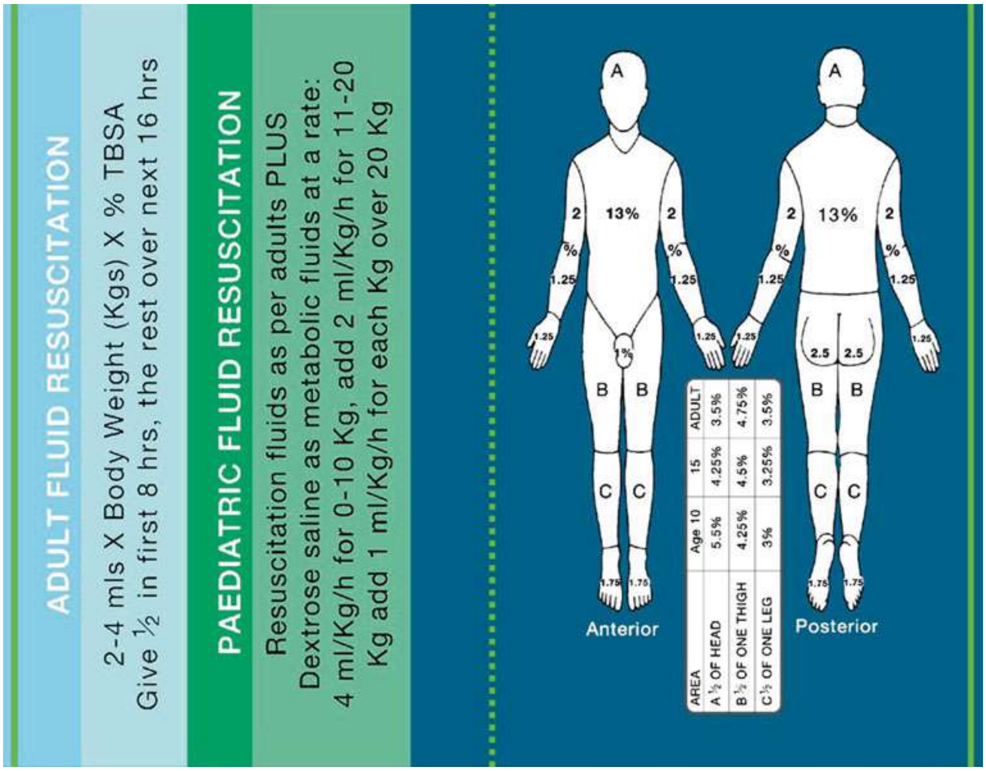

ADULT FLUID RESUSCITATION

2-4 mls X Body Weight (Kgs) X % TBSA
Give ½ in first 8 hrs, the rest over next 16 hrs

PAEDIATRIC FLUID RESUSCITATION

Resuscitation fluids as per adults PLUS
Dextrose saline as metabolic fluids at a rate:
4 ml/Kg/h for 0-10 Kg, add 2 ml/Kg/h for 11-20 Kg add 1 ml/Kg/h for each Kg over 20 Kg

AREA	Age 10	15	ADULT
A ½ OF HEAD	5.5%	4.25%	3.5%
B ½ OF ONE THIGH	4.25%	4.5%	4.75%
C ½ OF ONE LEG	3%	3.25%	3.5%

A
13%
2
2
%
%
1.25
1.25
1.25
1.25
1%
B
B
C
C
1.75
1.75
Anterior
A
13%
2
2
%
%
1.25
1.25
1.25
1.25
2.5
2.5
B
B
C
C
1.75
1.75
Posterior

Fig. 91.2 Lund Browder chart for estimating total body surface area of burns, with suggested fluid resuscitation guidelines (Reprinted from Malic CC, Karoo RO, Austin O, Phipps A. Resuscitation burn card–a useful tool for burn injury assessment. *Burns*. 2007;33(2)195–9, with permission from Elsevier)

91.5 Fluid Resuscitation

Because fluid resuscitation is absolutely essential in the early aspects of burn care, the Parkland formula has been used to estimate fluid requirements in burn patients. The patient's weight in kilograms is multiplied by the percent body surface area (BSA) involved; this number is multiplied by 4 mL of lactated Ringer's solution. Half of this amount is given during the first 8 h and the remaining amount is given over the next 16 h of resuscitation. The goal is to keep urine output approximately 0.5–1.0 mL/kg/h.

In pediatric cases, the Galveston formula may be used instead: 5000 mL/m^2 BSA burned plus 2000 mL/m^2. TBSA of 5 % dextrose in lactated Ringer's solution intravenously over the first 24 h, half in the first 8 h, and the other half over the next 16 h.

91.6 Procedure

1. Provide appropriate analgesia.
2. Clean the wound with antiseptic solution and water (if it is a dry chemical burn, make sure to brush off as much chemical before using copious amounts of water to clean the wound).
3. Debride any loose skin or foreign debris using a dry 4×4 or rolled gauze.
4. Apply antibiotic ointment or cream to the burn.
5. Apply petroleum gauze in a single layer just over the affected skin.
6. Use loose 4×4 gauzes and "fluff" them to make a thick layer of padding to place over the petroleum gauze.
7. Either wrap the entire area with a rolled gauze or tape a small layer of 4×4 gauze over the "fluffed" layer of 4×4 gauze.

91.7 Complications

- Wound infection
- Nonhealing wound requiring skin graft (deep second-degree and third-degree burns)
- Compartment syndrome (circumferential burns may require escharotomy)
- Rhabdomyolysis

91.8 Pearls and Pitfalls

- Pearls
  - Determination of the depth of burns on initial presentation is difficult (especially when covered with petroleum). A good rule of thumb is if it blanches and/or hurts, it is a partial thickness burn.
  - First-degree burns are not included in burn size estimations for fluid resuscitation calculation.
  - Burn size determines fluid requirements and transfer decision.
  - Burn of greater than 20 % TBSA should receive intravenous fluids.
  - Patients who have inhalation injury, greater than 10 % TBSA, high-voltage electrical injury, or chemical burns should be referred to burn unit.
  - Fluid resuscitation should be adjusted according to physiological response such as urine output (30–50 mL/h in adults and 1 mL/kg/h in children).
  - Assure tetanus is up to date.
- Pitfalls
  - Keep the patient warm in the first few hours. There is no need to apply ice.
  - All jewelry and rings should be removed.
  - Prophylactic antibiotics are not recommended.
  - Silver sulfadiazine should be avoided in facial burns because of the risk of staining of the skin.

Blister care is a very controversial topic. Current research suggests that it may be beneficial to keep the blister intact unless it appears to be tense or over a joint. Most blisters will rupture in 2–4 days. Ruptured blisters should be debrided with all the extra skin removed. Most burn units will scrub everything off once they receive a patient.

Wounds should be kept clean to prevent an environment that will increase the chances of infection. Wrap in saline-soaked sterile gauze prior to transfer.

Dressing changes should be done daily with all previously applied antibiotic ointment removed before a reapplication of new ointment. It is important to provide analgesic 30 min before a dressing change.

91.9 Admission Criteria

- Partial-thickness burns of noncritical areas not including the eyes, ears, face, hands, feet, or perineum that total a BSA of 10–20 % in adults
- Partial-thickness burns of noncritical areas involving 5–10 % of BSA in children younger than 10 years
- Suspicious of non-accidental trauma
- Patients unable to care for wounds in outpatient settings
- Prompt referral to a burn specialist is required in the following cases:
  - Partial-thickness and full-thickness burns greater than 10 % of the TBSA in patients younger than 10 years or older than 50 years of age.
  - Partial-thickness and full-thickness burns greater than 20 % of the TBSA in other age groups.

- Partial-thickness and full-thickness burns involving the face, eyes, ears, hands, feet, genitalia, or perineum or the skin overlying major joints.
- Full-thickness burns greater than 5 % TBSA in any age group.
- Electrical burns, including lightning injury: significant volumes of tissue beneath the surface may be injured and result in acute renal failure and other complications.
- Significant chemical burns.
- Inhalation injury.
- Burn injury in patients with preexisting illness that could complicate management, prolong recovery, or affect mortality.
- Any burn patient in whom concomitant trauma poses an increased risk of morbidity or mortality may be treated initially in a trauma center until stable before transfer to a burn center.
- Children with burns seen in hospitals without qualified personnel or equipment for their care should be transferred to a burn center with these capabilities.
- Burn injury in patients who will require special social and emotional or long-term rehabilitative support, including cases involving suspected child abuse and neglect.

Selected Reading

Bezuhly M, Fish JS. Acute burn care. Plast Reconstr Surg. 2012;130(2):349e–58.

Rex S. Burn injuries. Curr Opin Crit Care. 2012;18(6):671–6.

Wasiak J, Cleland H, Campbell F, Spinks A. Dressings for superficial and partial thickness burns. Cochrane Database Syst Rev. 2013;(3):CD002106.

92

Wound Closure

Oliver Michael Berrett, Jeffrey Joseph Harroch,
Karlene Hosford, and Muhammad Waseem

92.1 Indications

- Open wound of skin or mucosal tissues
- Purpose
  - Preserve function
  - Control bleeding
  - Promote healing
  - Cosmesis

92.2 Contraindications

- Wounds caused by animal or human bites
- Contaminated, infected, or puncture wounds
- Complex wounds (may require operating room)

92.3 Methods

- Suture placement
- Tissue adhesives
- Adhesive tapes
- Staples

92.4 Preparation

- Obtain a complete history of injury
  - Mechanism
  - Time since injury
  - Tetanus status
  - Comorbidities
- Examine the extent of the wound, remove any contaminants, debride devitalized tissue, inspect for foreign bodies. Obtain a radiograph if foreign body is suspected.
- Copious irrigation with normal saline.
- Clean around the wound with a povidone-iodine solution.
- Drape the area in sterile manner.
- Inject the anesthetic agent into the wound edges using a 25- or 27-gauge needle. The most common agent is lidocaine 2 % with or without epinephrine 1 % (maximum dose is 3 mg/kg without epinephrine and 5 mg/kg with epinephrine).
- Close the wound using the appropriate technique (see later).
- Dress the wound.
- Update tetanus and diphtheria vaccination if needed.

O.M. Berrett, MD • K. Hosford, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: omberrett@gmail.com; krlhos@aol.com; waseemm2001@hotmail.com

J.J. Harroch, MD
Department of Emergency Medicine, University of Miami Miller School of Medicine, University of Miami Hospital,
Miami, FL, USA
e-mail: jharroch@med.miami.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_92

92.5 Suture Repair

92.5.1 Materials and Medications

- Commercial kits commonly contain all the following except the suture material (Fig. 92.1)
  - Povidone-iodine solution
  - Normal saline for irrigation
  - 5- to 12-mL syringe with a 25-gauge needle
  - Anesthetic agent
  - Needle holder
  - Pickups
  - Suture scissors
  - Suturing material
  - Sterile drape or sheet, gloves, gauze

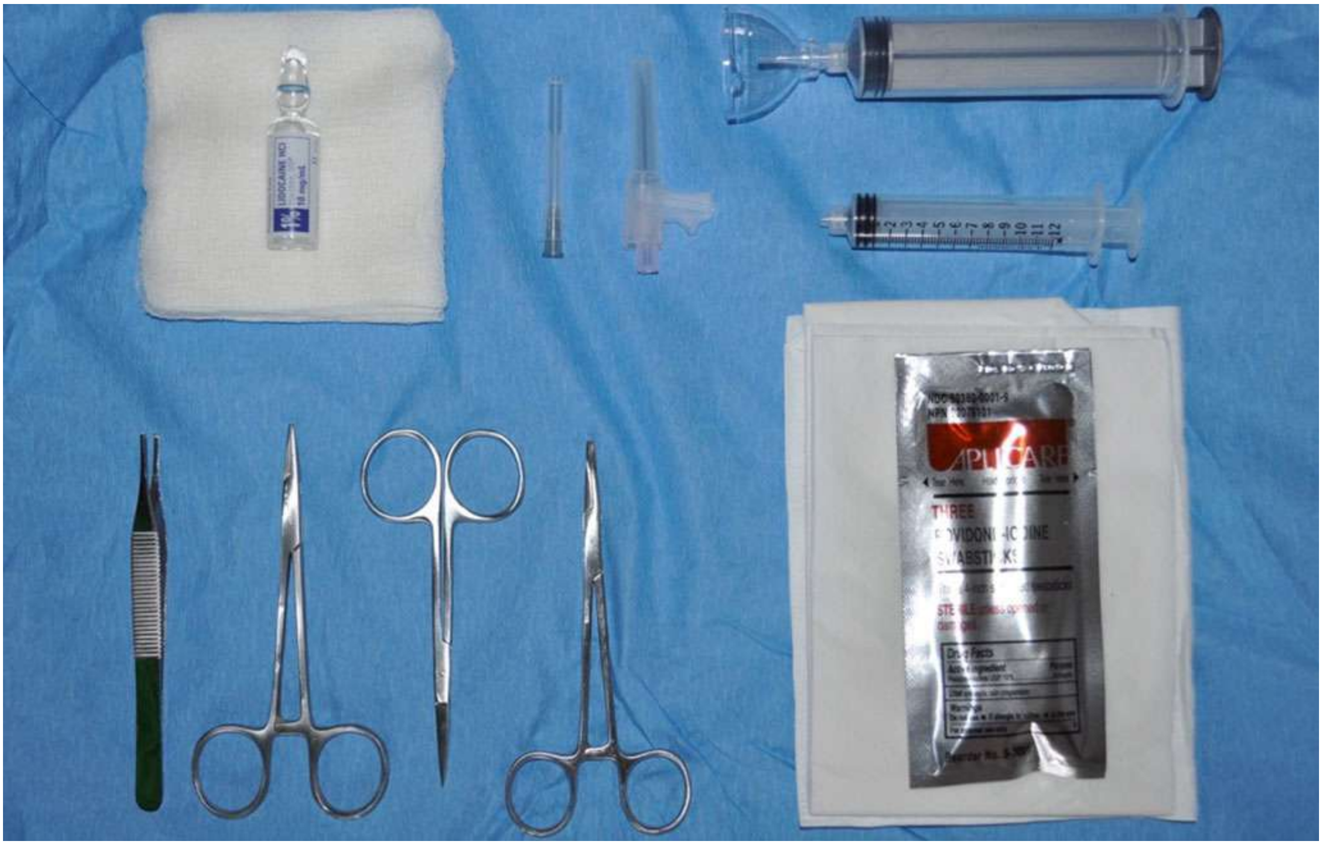

Fig. 92.1 Suture kit

92.5.2 General Guidelines

- Local anesthetic lidocaine 1 % or lidocaine 1 % with epinephrine.
- Minimize direct use of instruments on the tissues.
- Wound edges should be everted to maximize healing and cosmetic effect. This is achieved by inserting the needle at 90° to the skin.
- Sutures should be evenly spaced, placed 1–3 mm apart and 2 mm from the wound edge.
- Optimal tension is achieved by tying the sutures so the edges lightly approximate.

92.5.3 Suture Material

- Nonabsorbable
  - Silk: for specialty use, reactive and weak
  - Nylon (Ethilon), polypropylene (Prolene): good strength, good overall material for cutaneous wounds
  - Polypropylene: good strength, difficult to use
  - Require removal at a specified time
- Absorbable
  - Undergo rapid degradation in tissues, losing their tensile strength within 60 days
  - Indication: buried suture to reduce wound edge tension
  - Vicryl: subcutaneous placement, mucous membranes
  - Chromic: use for intraoral lacerations
  - Removal not required

92.5.4 Suture Size

- 0.0–2.0: thick material for large wounds, trunk
- 0.3–0.4: used on medium-sized wounds, extremities, scalp
- 0.5–0.6: fine sutures, used on facial wounds

92.5.5 Suture Techniques

- Simple interrupted sutures (Fig. 92.2)
  - Most common method
  - Position needle 2 mm from the wound edge at a 90° angle. Enter the needle into the skin and arc through the wound edge and into the opposing edge at the same level, exiting the skin on the opposing side 2 mm from wound edge, tie.
- Deep dermal suture (Fig. 92.3)
  - Also known as buried sutures, used to minimize tension in a wound.
  - Use absorbable suture material.
  - The needle is placed at the base of the wound wall and arched upward, exiting the ipsilateral wall more superficially. The needle is then directed across to the opposing wound wall at the same level and directed downward, exiting deep, tie. Note: The knot will be deep in the wound.
- Simple running suture (Fig. 92.4)
  - This method provides rapid closure of long and relatively linear lacerations.
  - Place the initial suture in the same manner as a simple suture, tie, cut the free strand, and leave the needle attached. Reintroduce the needle into the skin on the opposite side so the suture crosses the wound superficially at a 65° angle. The needle is then inserted perpendicular to the skin, emerging on the opposite side about 3 mm from the wound edge. Repeat without tying until closure is complete, maintaining appropriate tension. When the last stitch is placed, leave a loose loop of suture on one side so that both ends can be tied together.
- Vertical mattress suture (Fig. 92.5)
  - Provides the benefits of both simple and deep techniques. For use on deeper, gaping wounds and wounds over high-tension areas such as joints.
  - Position the needle 1 cm from skin edge, at a 90° angle, drive a deep arc perpendicular through the wound, exiting the skin on the opposing side the same distance from the wound edge. Next, reinsert the needle on the ipsilateral side 2 mm from the edge, emerging on the opposing side and approximate, tie.
- Horizontal mattress suture (Fig. 92.6)
  - For large wounds with tension.
  - Place the initial suture in the same manner as a simple suture, only do not tie. Reposition the needle on the ipsilateral side, horizontally 5 mm to the side of the exit at the same distance from the wound edge and drive through to the other side, tie. Note: The tie lies parallel to the wound.
- Half-buried horizontal mattress suture (Fig. 92.7)
  - Used on wounds with skin flap.
  - On one side, drive the needle percutaneously, and then pass horizontally through the dermal tissue of the tip of flap, finally passing into the dermis of the opposing edge and exiting the skin, tie.

92.5.6 Recommendations

- Face
  - 5.0–6.0 nylon.
  - Remove after 3–5 days.
- Scalp
  - 2.0–3.0 nylon or staple.
  - Remove at 8–10 days.

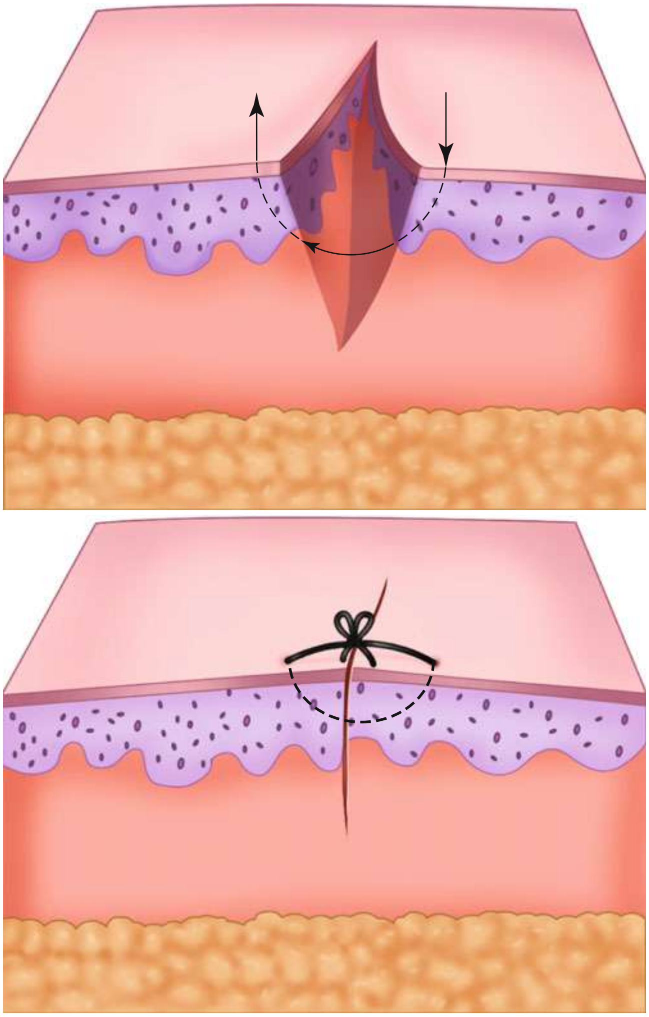

Fig. 92.2 Simple suture

- Hand
  - 4.0–5.0 nylon, consider vertical or horizontal mattress.
  - 5.0–6.0 Monocryl for nail bed.
  - Remove at 10–14 days, use for a longer time if directly over the joint.
- Extremity
  - Nonmobile skin: 3.0–4.0 nylon, remove at 8–10 days
  - Over joint: 3.0–4.0 nylon, remove at 10–14 days
- Trunk
  - Anterior trunk: 3.0–4.0 nylon, remove at 8–10 days.
  - Posterior trunk: 2.0–3.0 nylon, remove at 10–14 days.
  - Consider staples.
- Oral mucosa
  - Thin mucosa: 4.0 Vicryl will absorb; duration, 5–7 days
  - Tongue: 3.0 Vicryl will absorb; duration, 5–7 days

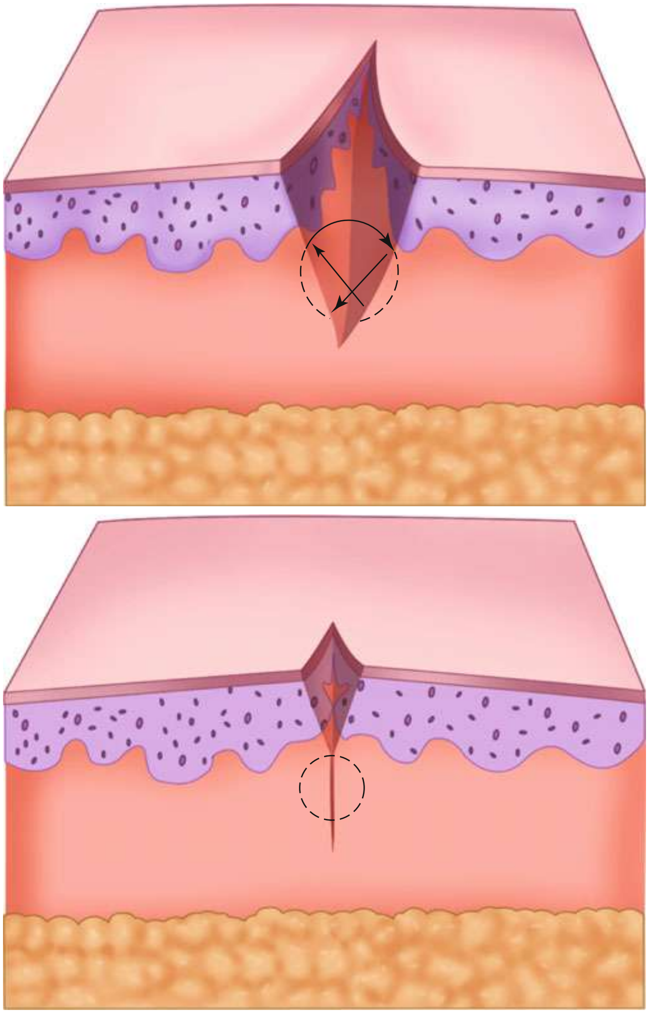

Fig. 92.3 Deep suture

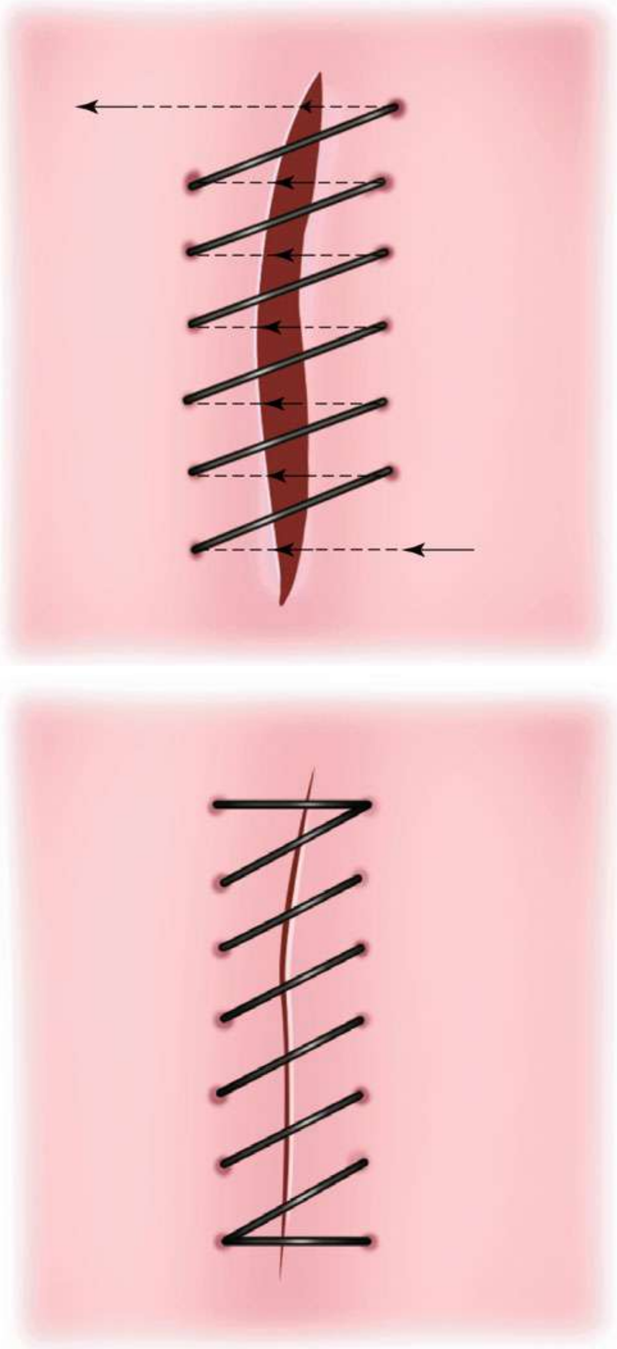

Fig. 92.4 Running suture

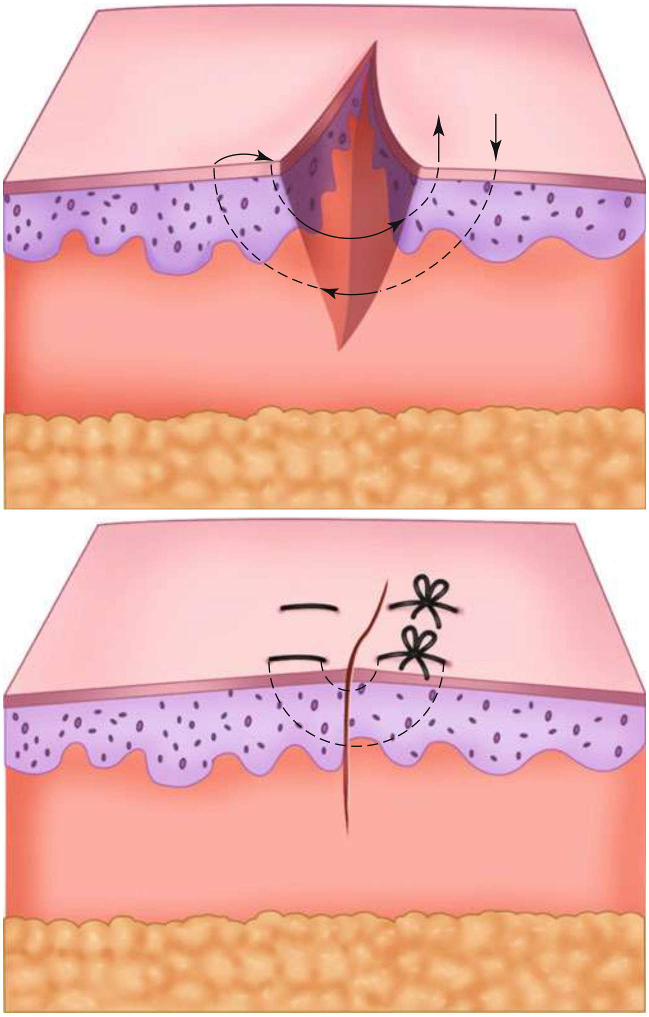

Fig. 92.5 Vertical mattress suture

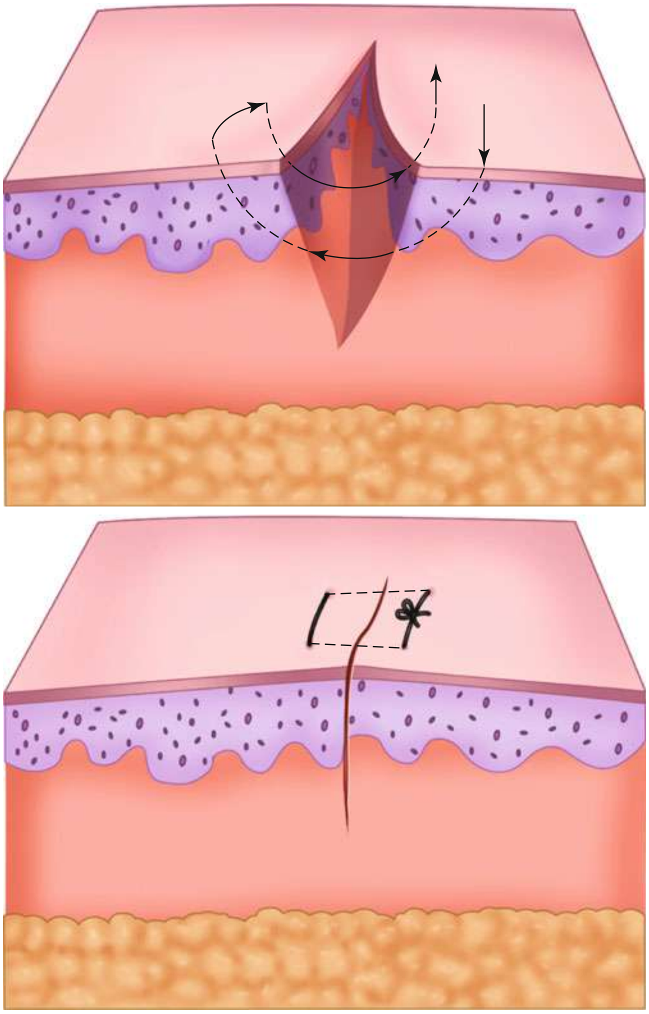

Fig. 92.6 Horizontal mattress suture

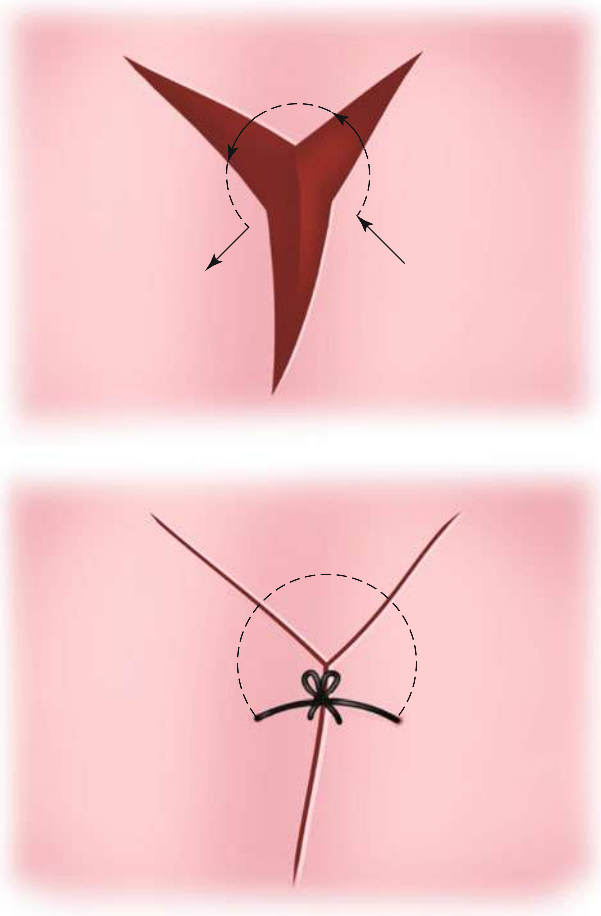

Fig. 92.7 Half-buried horizontal mattress suture

92.6 Alternative Methods of Wound Closure

92.6.1 Cyanoacrylate Tissue Adhesives

See Chap. 93.

92.6.2 Adhesive Tape

- Advantages
  - Rapid and painless application
  - Inexpensive
  - Good cosmetic result
- Disadvantages
  - Minimal strength
- Contraindications
  - Allergy to product
- Precautions
  - For use with wounds under little tension
- Procedure
  - Thoroughly clean the wound as described previously.
  - Approximate the wound edges.
  - Apply adhesive tape directly over the wound with 2–3 mm of space between strips.
  - An adjunct adhesive, such as benzoin, may be used to improve durability.

92.6.3 Staple Closure (Fig. 92.8)

- Advantages
  - Rapid
  - Inexpensive
- Disadvantages
  - Minimal strength
- Contraindications
  - Wounds on face, hands, or feet
- Procedure
  - Thoroughly clean the wound as described previously.
  - Anesthetize the wound edges.
  - Approximate the wound edges (may require an additional set of hands).
  - Apply staples.

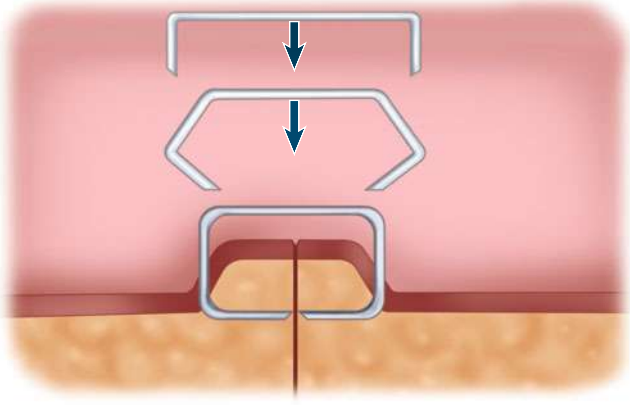

Fig. 92.8 Staple

92.7 Pearls and Pitfalls

- Pearls
  - Sutures should not remain in place longer than the recommended period. If the wound is not completely healed, sutures should be removed and an adhesive tape may be used for edge approximation.
  - Remember to verify tetanus status.
  - Antibiotics should be used judiciously.
- Pitfalls
  - Wound infection.
  - Incomplete healing may lead to wound separation.
  - Tissue reaction to suture or adhesive materials.
  - Allergy to anesthetic agent.

Selected Reading

http://apps.med.buffalo.edu/procedures/repairoflacerations.asp?p=17. Accessed 19 May 2014.

Singer AJ, Hollander JE. Methods for wound closure. In: Ma OJ, Cline DH, Tintinalli JE, Kelen GD, Stapczynski JS, editors. Emergency medicine manual. 6th ed. New York: McGraw Hill; 2003: Chap. 13, Fig. 13–14.

Singer AJ, Hollander JE, editors. Laceration and acute wounds: an evidence-based guide. Philadelphia: FA Davis; 2003. p. 122.

University of Connecticut Health Center, suturing 101. fitsweb.uchc.edu/suturing101. Accessed 19 May 2014.

Zuber TJ. The mattress sutures: vertical, horizontal and corner stitch. Am Fam Physician. 2002;66:2231–6.

Wound Closure with Tissue Adhesive

93

Pratik S. Patel and Latha Ganti

93.1 Indications

- Small superficial skin incisions or laceration repairs which require 5.0 or smaller-diameter sutures.

93.2 Contraindications

- Absolute
  - Large irregular/stellate lacerations
  - Infected/contaminated wounds
  - Animal/human bites
  - Puncture wounds
  - Crush wounds
  - Skin ulcers
  - Mucous membranes and mucocutaneous junctions
  - Axillae and perineum (owing to high moisture)
- Relative
  - Wounds on extremities (unless kept dry)
  - Joints (unless kept immobilized with a splint)

93.3 Materials and Medications

- Dermabond (2-octyl cyanoacrylate)
- Betadine (povidone-iodine) solution
- 0.9 % normal saline solution
- 20-mL sterile syringe
- Sterile gloves
- Dry 4×4 gauze

93.4 Optional Materials

- Topical anesthesia
- 1:1000 epinephrine solution
- Forceps
- Bacitracin ointment or sterile petroleum jelly ointment
- Gown and protective eyeglasses
- Splint

P.S. Patel, MD
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: drpratik@gmail.com

L. Ganti, MD, MS, MBA (✉)
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_93

93.5 Procedure

1. Have the patient rest comfortably on a chair or bed.
2. Use universal precaution measures: sterile gloves (gown and eye-screen, if necessary for wound irrigation).
3. Wash the wound with 0.9 % normal saline irrigation.
4. Use a topical anesthetic such as LET (lidocaine, epinephrine, tetracaine) or EMLA (eutectic mixture of local anesthetics) cream (lidocaine and prilocaine), or a 1:1000 epinephrine solution soaked into gauze can be used to achieve hemostasis in a bleeding wound.
5. Approximate the edges of the wound with fingers. Toothed forceps or other skin approximation devices may be used as an adjunct. Apply bacitracin ointment or Vaseline to the tips of the forceps and wipe off the excess to prevent sticking of Dermabond glue to the forceps.
6. Crush Dermabond vial between the thumb and the finger while in the inverted position until the adhesive is seen at the applicator tip (Fig. 93.1).
7. Squeeze gently until a drop of adhesive forms at the applicator tip.
8. Gently brush the adhesive at the applicator tip over the approximated wound edges (Fig. 93.2). (Do not force or press the applicator tip over the wound.)
9. Cover the entire wound with single coat of adhesive.
10. Hold the wound edges for 30 s to 1 min until it dries.
11. Apply two or three more coats of adhesive over and around the wound in a circular or oval movement to provide extra strength.
12. Wipe off extra adhesive in the surrounding skin with gauze if needed.
13. Apply a splint (optional) to provide wound stability over the joints.
14. Recommend the patient to keep the wound dry for 4–5 days. Patients may shower but should be instructed to pat dry instead of rubbing a towel over the skin.
15. No topical antibiotic ointment is required before or after application of Dermabond.

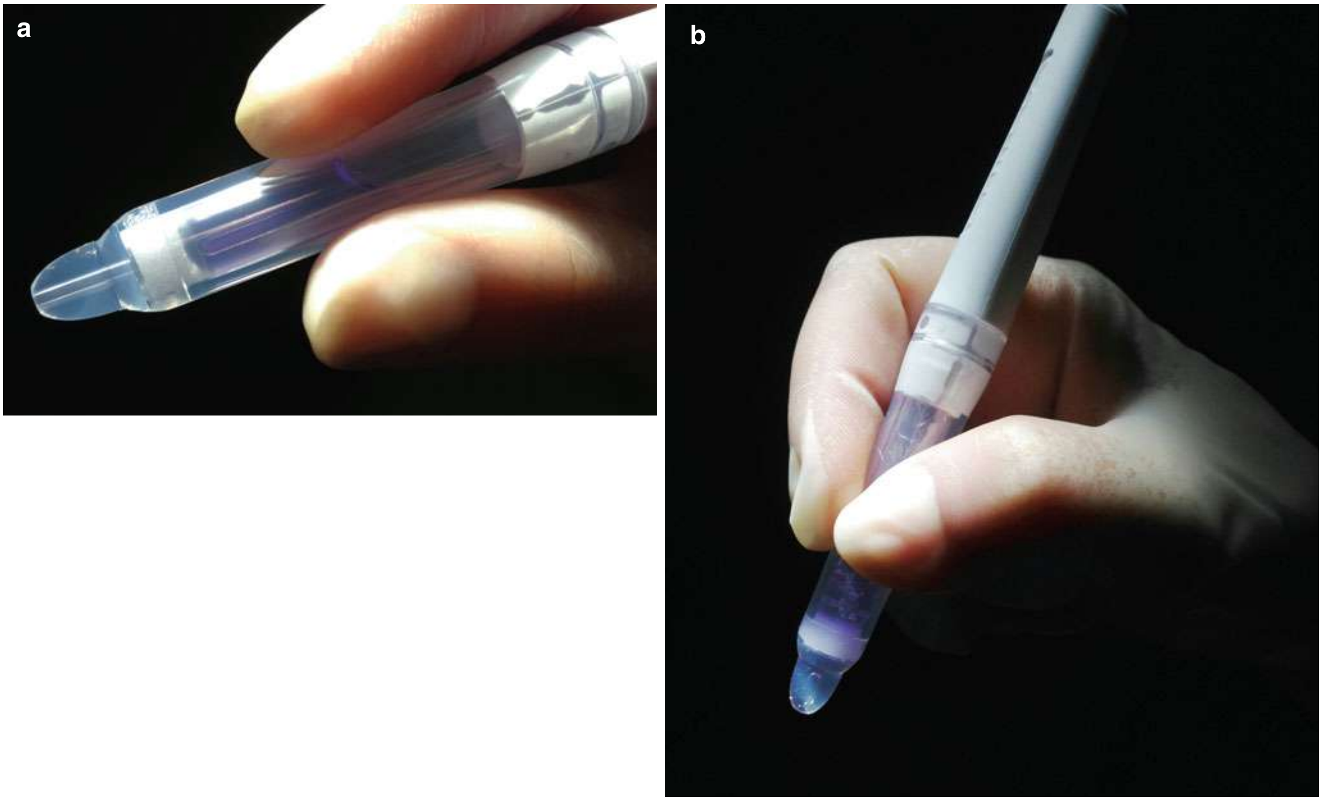

Fig. 93.1 (**a**) Before the tip has been crushed. (**b**) The purple Dermabond is in the tip

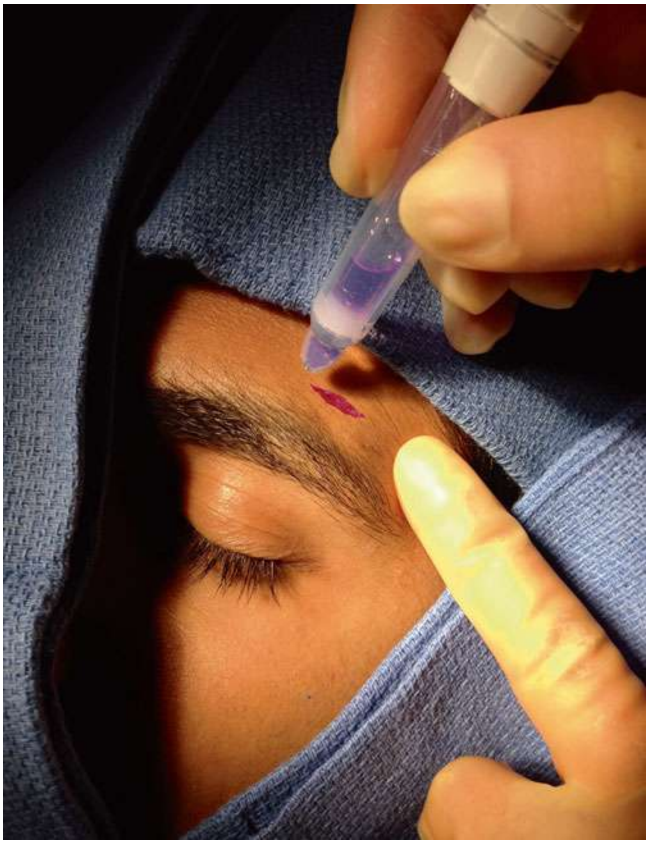

Fig. 93.2 Closing the wound: use tip of Dermabond pen to “paint” over the laceration

93.6 Complications

- Wound dehiscence
- Wound infection

93.7 Pearls and Pitfalls

- Pearls
  - Advantages of tissue adhesives for wound repair compared with sutures include faster repair time, better acceptance by patients (especially children), water-resistant covering, and no need for a second visit to remove sutures (sloughs off in 5–10 days).
- Pitfalls
  - Dermabond is a super adhesive. Take care not to have the glove, finger, drape, gauze, or instrument inadvertently stuck to the wound or the patient by having a bacitracin or petroleum jelly coating around the wound, on gloved fingers, and on forceps as needed.

Selected Reading

Bruns TB, Worthington JM. Using tissue adhesive for wound repair: a practical guide to Dermabond. Am Fam Physician. 2000;61:1383–8.

Farion K, Osmond MH, Hartling L, et al. Tissue adhesives for traumatic lacerations in children and adults. Cochrane Database Syst Rev. 2002;(3):CD003326.

Fishhook Removal

94

Judith K. Lucas

94.1 Indications

- Removal of a fishhook from nonvital structures (Fig. 94.1)

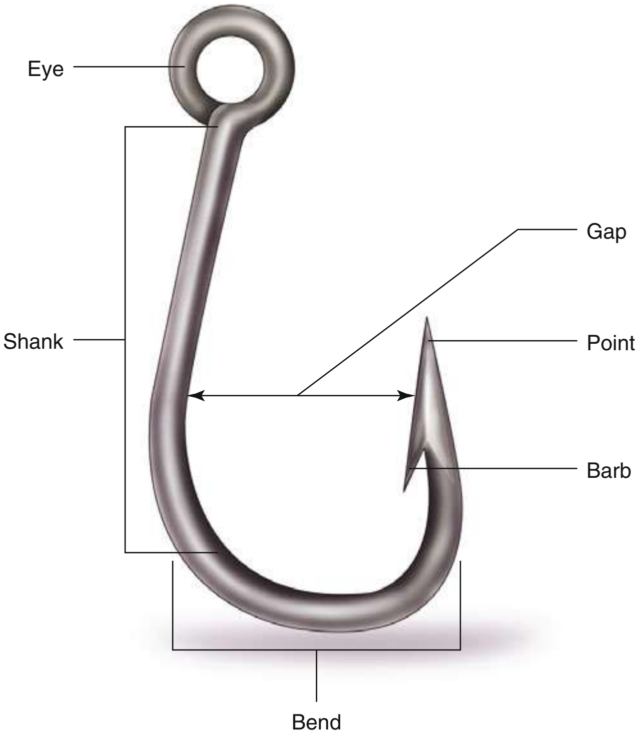

Fig. 94.1 Anatomy of a fishhook

94.2 Contraindications

- Removal of hooks located near/in eyes or eyelids, embedded near or within neurovascular structures, or embedded within vital structures such as peritoneum, testicle, or urethra.
  - Fishhooks in these areas require specialist consultation.

94.3 Materials and Medications

- Antiseptic cleansing solutions
  - Betadine (povidone-iodine)
  - Chloraprep
- Local anesthetic
  - 1 % lidocaine, with or without epinephrine
- Needle drivers or pliers
- 18- or 20-gauge needle
- 3-0 silk suture or umbilical tape
- Wire cutters
- Protective eyewear

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_94

94.4 Procedures

- Retrograde technique (Fig. 94.2): simplest, least traumatic, but least successful; good for small to medium hooks, superficially embedded hooks, and hooks with no barbs or a single barb.
  1. Detach extra hooks, line, or foreign materials (e.g., worms, fish, debris).
  2. Cleanse the puncture site and surrounding tissue with antiseptic solution.
  3. Infiltrate the entry site and surrounding area with local anesthetic.
  4. Apply downward pressure on the shank or shaft of the hook, near the eye, thus disengaging the barb from the tissue.
  5. Back the hook out of the skin along the path of entry.
- String and yank technique (Fig. 94.3): Good for small to medium hooks. Often requires no anesthesia. Good for areas of deep soft tissue penetration. Cannot be used on parts of the body that are not fixed, such as earlobes.
  1. Detach extra hooks, line, or foreign materials (e.g., worms, fish, debris).
  2. Cleanse the penetration site and surrounding area with antiseptic wash.
  3. Consider infiltrating the entry site with local anesthetic.
  4. Wrap 3-0 silk suture, fishing line, or umbilical tape several times around the bend of the hook (at the point of greatest curvature).
  5. The loose ends of the string need be held tightly. Sometimes the loose ends can be more firmly held if wrapped around a pencil or tongue depressor.
  6. The skin around the entry site should be well stabilized, while simultaneously depressing the shank, close to the eye of the hook.
  7. While stabilizing the skin and applying downward pressure on the shaft, quickly and firmly yank the string, in a parallel line to the shaft. Be certain downward pressure is applied along the shank.
  8. The hook will "fly" out quickly. Be certain to wear eye protection.
- Needle cover technique (Fig. 94.4): good method to remove large hooks with a single barb, especially if superficially embedded
  1. Detach extra hooks, line, or foreign materials (e.g., worms, fish, debris).
  2. Cleanse the penetration site and surrounding area with antiseptic wash.
  3. Infiltrate the entry site and surrounding area with local anesthetic.
  4. Advance an 18-gauge needle along the entrance wound of the hook. Pass the needle parallel to the shank. The bevel should be pointed downward, toward the barb and point.
  5. The needle is advanced until it disengages the barb, entrapping it within the needle lumen.
  6. The hook and needle are advanced just enough to disengage the barb.
  7. The hook and needle are withdrawn along the track of the wound.
- Advance and cut technique (Fig. 94.5). Almost always successful but does cause additional tissue trauma.
  1. Detach extra hooks, line, or foreign materials (e.g., worms, fish, debris).
  2. Cleanse the penetration site and surrounding area with antiseptic wash.
  3. Infiltrate the entry site and surrounding area with local anesthetic.
  4. If the hook has a single barb:
     - Grip the hook on the shank, near the bend, with either a needle driver or pliers.
     - Push the hook through along its natural trajectory, until the point and barb pass completely through the skin.
     - Clip the point proximal to the barb. Be certain to wear eye protection because the point can fly off in an unpredictable direction.
     - Withdraw the hook out back along its entry path.
  5. If the hook has multiple barbs (Fig. 94.6):
     - Grip the hook on the shank, near the hook's bend with either a needle driver or pliers.
     - Push the hook through along its natural trajectory, until the point and barb pass completely through the skin.
     - Clip the eye of the hook with wire cutters.
     - Grasp the point and withdraw the hook through the skin forward along its natural course.

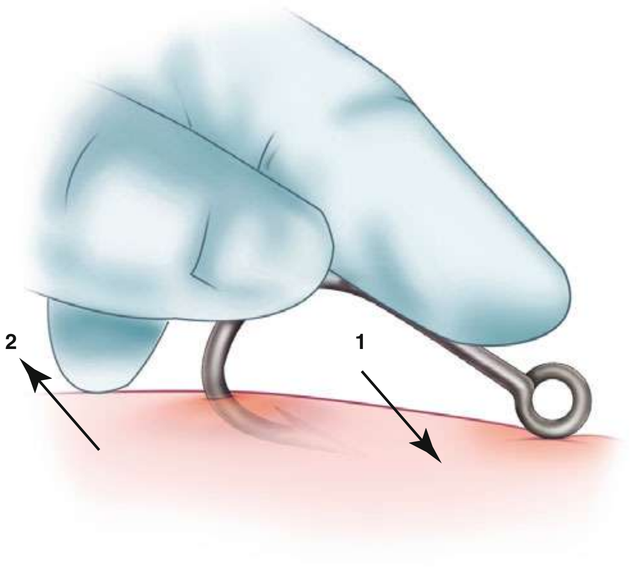

Fig. 94.2 Retrograde technique. *1* Push down along whole shank and slightly forward to release the barb. *2* Pull back and out along hook's entry path, applying steady downward pressure

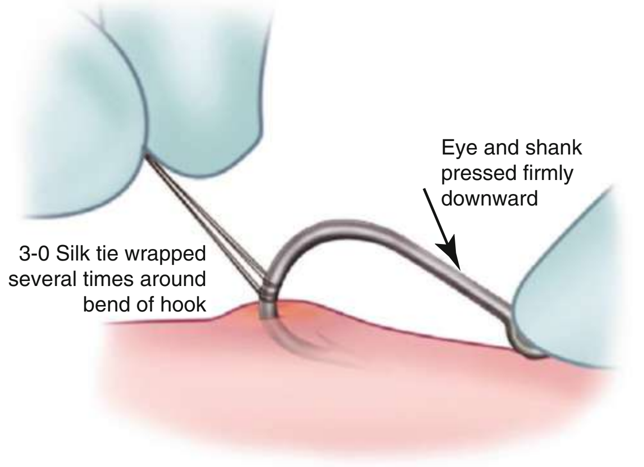
Eye and shank
pressed firmly
downward
3-0 Silk tie wrapped
several times around
bend of hook

Fig. 94.3 String and yank technique

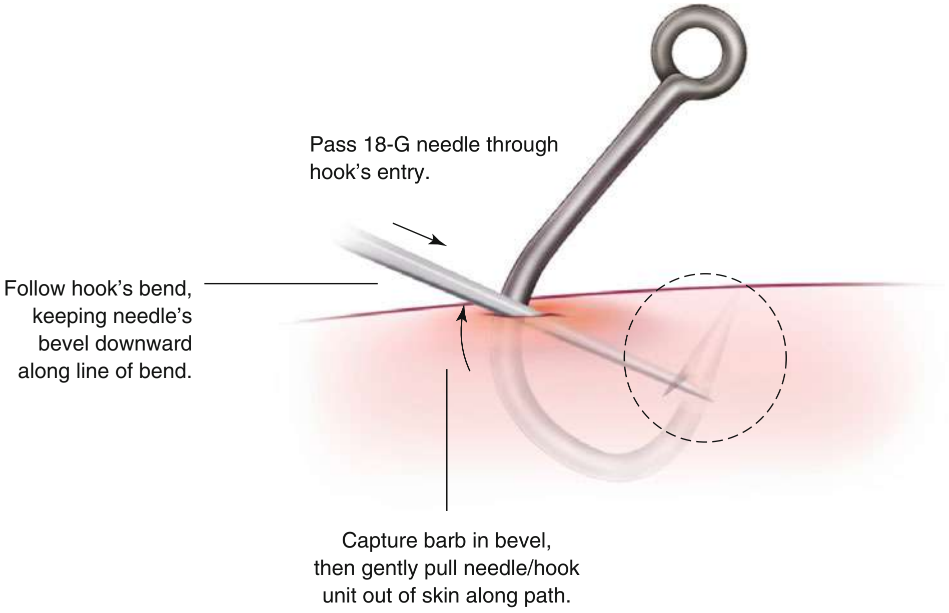
Pass 18-G needle through
hook's entry.
Follow hook's bend,
keeping needle's
bevel downward
along line of bend.
Capture barb in bevel,
then gently pull needle/hook
unit out of skin along path.

Fig. 94.4 Needle technique

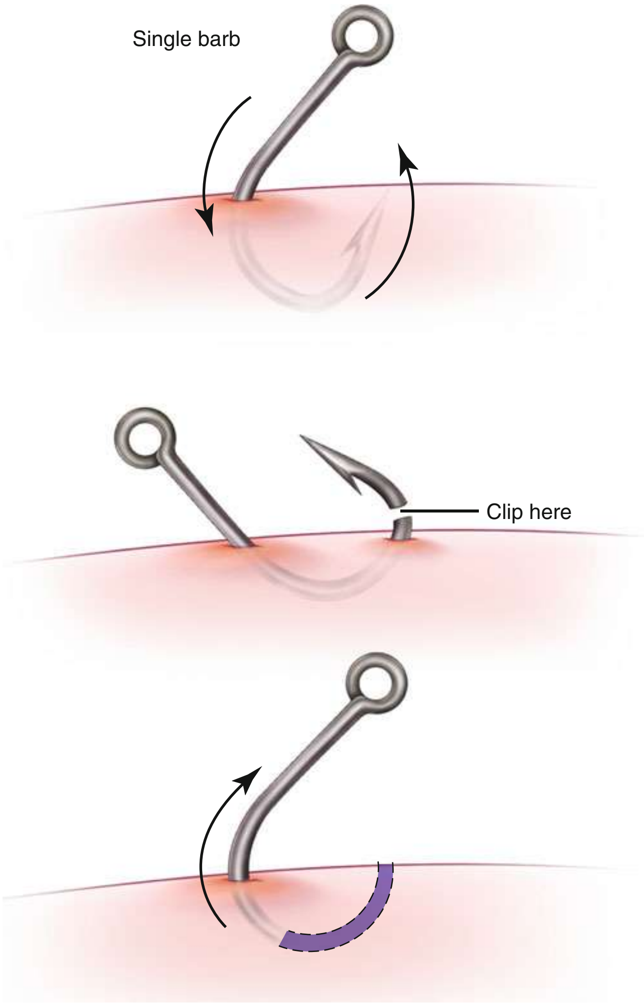
Single barb
Clip here

Fig. 94.5 Advance and cut technique

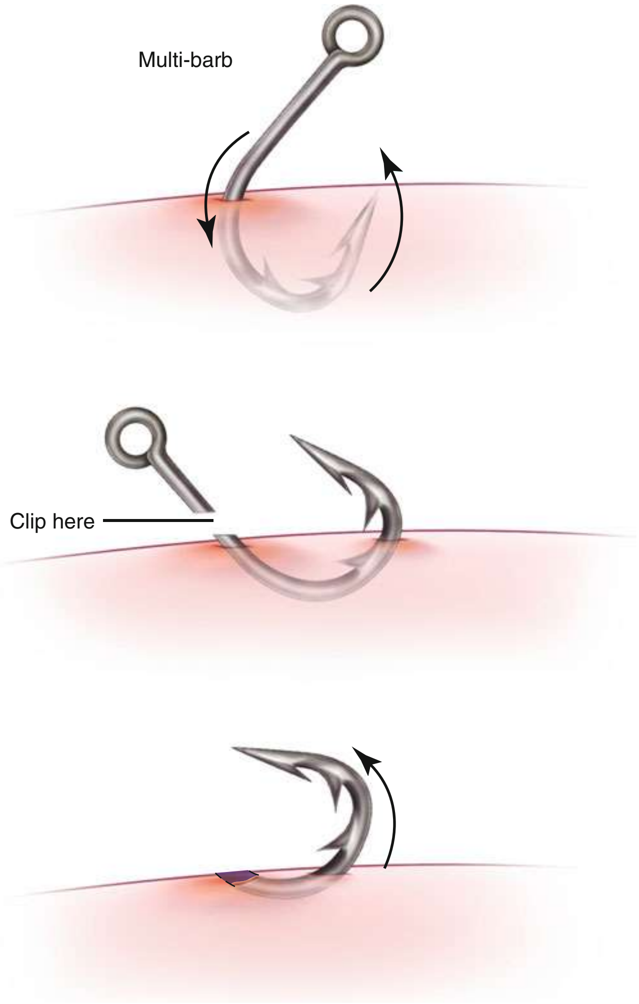
Multi-barb
Clip here

Fig. 94.6 Advance and cut a hook with multiple barbs present

94.5 Postremoval Wound Care

- Immune competent, without peripheral vascular disease:
  1. Explore the wound for possible foreign bodies.
  2. Irrigate or scrub the wound copiously with soapy solution.
  3. Apply antibiotic ointment and sterile dressing.
  4. Tdap (or Dtap) should be given to anyone in whom the last tetanus booster has been longer than 5 years.
  5. Wound check in 24–48 h with care provider.
- Immunocompromised or a patient with peripheral vascular disease:
  1. As previously.
  2. Give strong consideration to treating prophylactically with antibiotics, choosing a fluoroquinolone, third-generation cephalosporin, or aminoglycoside.

94.6 Complications

- Infection
- Retained foreign body
- Injury to neurovascular structures, if method of removal is not carefully selected
- Injury to provider if adequate (eye) protection not used

94.7 Pearls and Pitfalls

- Pearls
  – Begin with either the retrograde method or the string-yank method, because these result in less tissue damage and are the easiest to perform, although they have the lowest rate of success.
  – Eye protection is imperative, especially if utilizing the string-yank method or clipping any portion of the hook (eye or point) because the retraction of the hook, or its parts, is generally at a high velocity and travels an unpredictable path.
  – When trying the advance and cut technique, stop immediately if, when advancing the hook, impasse, or resistance is met because this may indicate bone or neurovascular structures are blocking the natural path of the hook.
  – Fishhooks that embed into or near the eye or lids should be covered with a metal patch or cup and the patient should be sent (immediately) for ophthalmologic consultation.
  – Close follow-up is imperative to watch for signs of infection.
- Pitfalls
  – When utilizing the advance and cut technique, do not cut anything until certain there is another portion of the hook on which to grasp.

Selected Reading

Bothner J. Fish-hook removal techniques. Available at: www.UpToDate.com. Literature review version 19.2, May 2011. Topic last updated: Sep 2010. Accessed 23 June 2014.

Gammons M, Jackson E. Fishhook removal. Am Fam Physician. 2001;63:2231–7.

Wakeman K. Fishhook removal. Available at: www.fishgame.com. Texas Fish and Game, LLC. July 2003. Accessed 23 June 2014.

Tick Removal

95

David N. Smith and Judith K. Lucas

95.1 Indications

- Tick attachment to the skin

95.2 Contraindications

- None

95.3 Materials and Medications

- Gloves
- Skin disinfectant (commercially available product, such as Chloraprep, isopropyl alcohol, or Betadine [povidone-iodine])
- Fine-toothed forceps

95.4 Procedure

1. Comfortably position the patient with the tick site exposed.
2. Grasp the tick as close to the skin surface as possible (e.g., grasp the mouth parts).
3. Gently pull upward with steady, nontwisting, even traction (Figs. 95.1 and 95.2).
4. After removal, clean the bite area and apply antibiotic ointment.

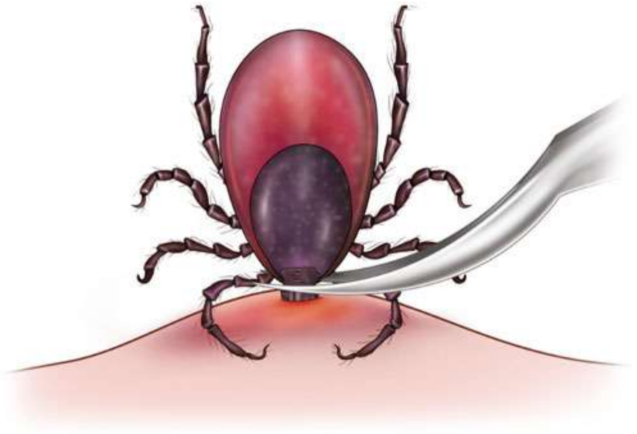

Fig. 95.1 Place the forceps as close to the mouth of the tick as possible, hold firmly, and pull straight up with steady gentle traction

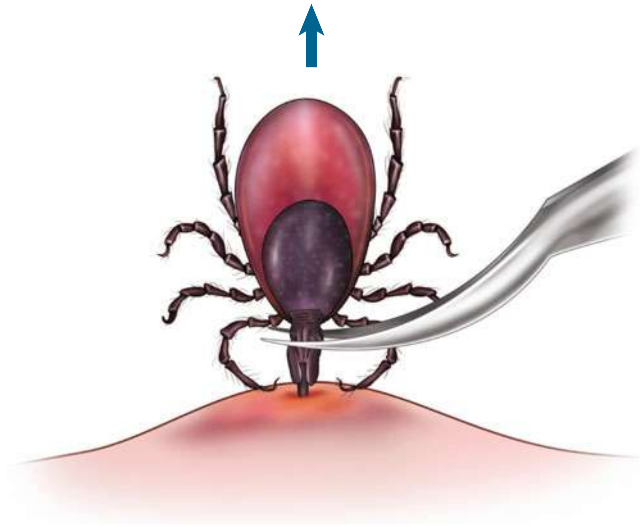

Fig. 95.2 Steady upward traction

D.N. Smith, MD
Department of Pediatrics, University of Alabama at Birmingham,
Children's of Alabama, Birmingham, AL, USA
e-mail: dsmith@peds.uab.edu

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_95

95.5 Complications

- Multiple diseases including:
  - Lyme disease
  - Human granulocytic and monocytic ehrlichiosis
  - Babesiosis
  - Relapsing fever
  - Rocky Mountain spotted fever
  - Colorado tick fever
  - Tularemia
  - Q fever
  - Tick paralysis
- Secondary infection (methicillin-resistant *Staphylococcus aureus* [MRSA] and group A streptococcus).
- Scratching can lead to lichenification.
- Rare cases of alopecia when tick located in the scalp.

95.6 Pearls and Pitfalls

- Pearls
  - Do not twist or jerk the tick out (may cause breakage of mouth parts; they may remain in the skin).
  - Do not squeeze the body of the tick.
  - Do not use a hot match, gasoline, or other noxious stimulus for removal (causes irritation of tick and release of internal contents).
  - Lyme disease transmission increases significantly after 24–48 h of attachment, so early removal is the key.
  - Patients should be monitored for up to 30 days for signs of tick-borne diseases, including erythema migrans (bull's-eye rash) indicating Lyme disease (Fig. 95.3).
- Prophylactic antibiotic treatment, a single dose of doxycycline, is used only in patients with an identified *Ixodes scapularis* tick that has been attached for longer than 36 h and if treatment can start within 72 h of tick removal.
  - Serological testing for Lyme disease is not indicated for a reported tick bite.
- Pitfalls
  - Use of lidocaine subcutaneously can irritate the tick and cause it to regurgitate its stomach contents, increasing the risk of disease transfer.

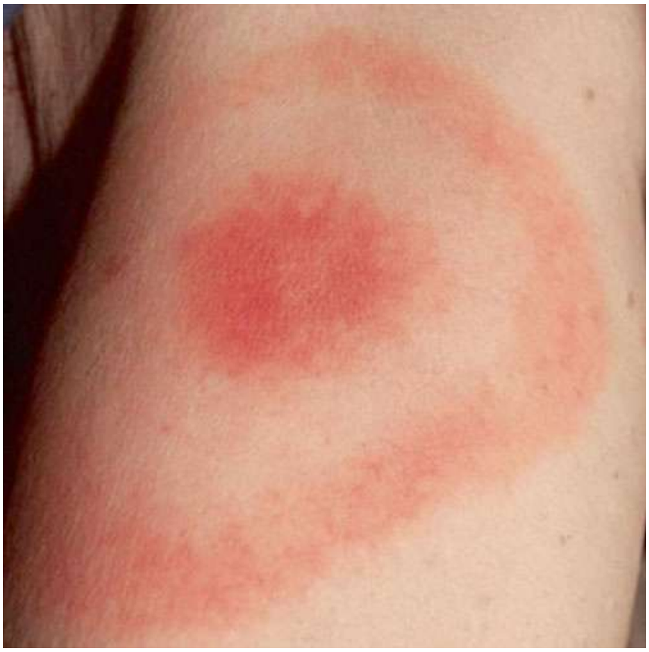

Fig. 95.3 Erythema migrans (the bull's-eye rash), the rash typically associated with Lyme disease. Often the rash is so pale as to go unnoticed

Selected Reading

Centers for Disease Control and Prevention. Tick removal. In: Ticks. 2014. http://www.cdc.gov/ticks/removing_a_tick.html. Accessed 23 Feb 2014.

Needham GR. Evaluation of five popular methods for tick removal. Pediatrics. 1985;75:997–1002.

Sexton DJ. Evaluation of a tick bite for possible Lyme disease. Up To Date; 2010. http://uptodate.com.

Sloan S. Background. In Tick removal. 2014. http://emedicine.medscape.com/article/1413603-overview. Accessed 23 Feb 2014.

Subungal Hematoma Drainage

96

Pratik S. Patel and Latha Ganti

96.1 Causes

- Crushing injury to the finger or toe
- Ill-fitted shoes/inadequate space for toes

96.2 Indications

- Pain that is not tolerable with nail edges intact

96.3 Contraindications

- Relative
  - Disrupted nail edges
  - Tolerable pain that can be managed conservatively
  - Skin infections around toe/finger
  - Bleeding disorder

96.4 Materials and Medications

- 18-gauge needle
- Betadine (povidone-iodine) solution
- Gloves
- Topical antibiotic

96.4.1 Optional materials

- Electrocautery tool
- Paper clip or sewing needle and sterilizing flame
- Finger splint
- Lidocaine (1 or 2 % with epinephrine)

P.S. Patel, MD
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: drpratik@gmail.com

L. Ganti, MD, MS, MBA (✉)
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_96

96.5 Procedure

1. Have the patient rest the affected finger comfortably on a flat surface.
2. Use universal precaution measures: gloves, gown, and eye screen.
3. Sterile skin preparation with Betadine.
4. Optional (for complicated subungal hematoma): use 1 or 2 % lidocaine with epinephrine for providing local digital block.
5. Using the thumb and index finger gently twist the 18-gauge needle with light pressure over the base of the nailbed or in the center of the hematoma until no resistance is felt. Do not apply any further pressure in order to avoid nailbed damage. Nail penetration is confirmed by return of dark blood from the hole. This procedure is known as trephination (Figs. 96.1 and 96.2).
6. Apply light pressure around the tip of the finger and hematoma to facilitate drainage. In case of continuous bleeding elevate the digit and apply firm and continuous pressure over the nail with a gauze piece.
7. Apply topical antibiotic ointment (e.g., bacitracin) over the puncture hole.
8. Apply a gauze dressing or a bandage over the wound site/fingernail.
9. Apply a finger splint (optional) to provide additional comfort.
10. Recommend the patient to keep the finger or toe dry (avoid soaking) and elevated for 1–2 days.

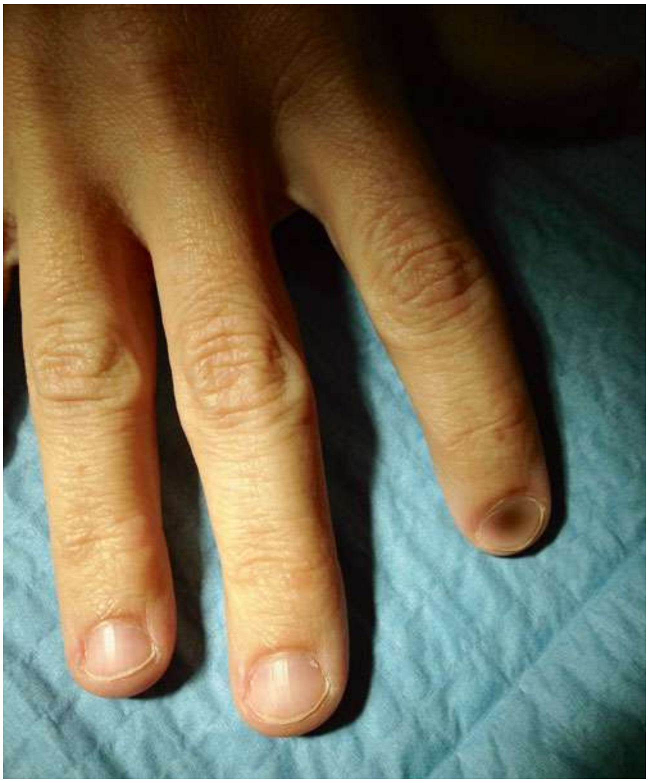

Fig. 96.1 Subungal hematoma of right index finger

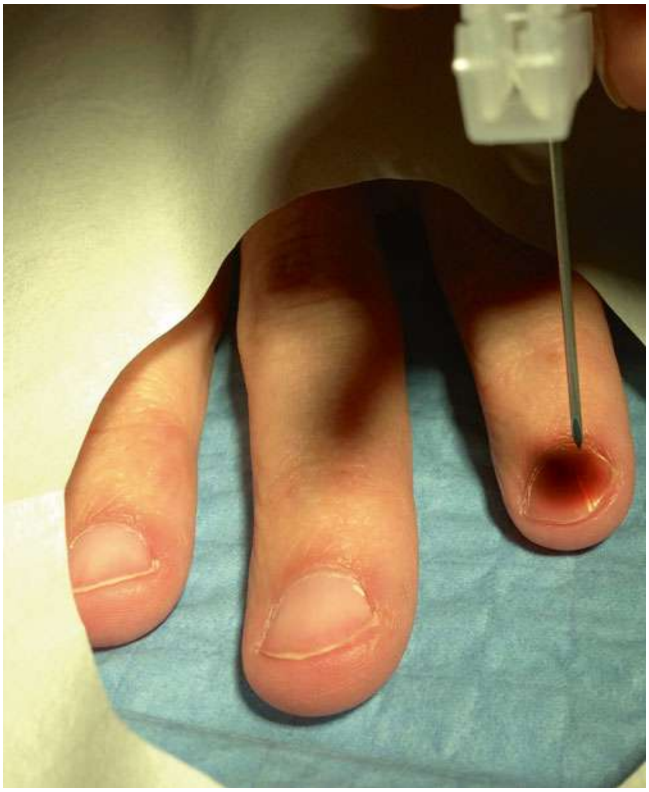

Fig. 96.2 Positioning of needle to perform subungal hematoma drainage

96.6 Complications (Rare)

- Infection
- Injury to nail bed or underlying bone if too much pressure is applied with puncture
- Onycholysis when there is incomplete drainage in rare cases

96.7 Pearls and Pitfalls

- Pearls
  - Multiple holes may be necessary for appropriate drainage of the hematoma.
  - When using electrocautery for hole puncture, execute high caution with acrylic nails because they are flammable.
  - When using a paper clip or regular needle, make sure to sterilize the tip over a flame.
  - Take a radiograph of the finger whenever necessary to rule out phalangeal fracture.
  - Check for avulsion of the extensor tendon.
  - Inform the patient that the existing nail may fall off and will regenerate in few months if the nailbed is intact.
  - Systemic or oral antibiotics are not recommended.
- Pitfalls
  - Development of a dark color change over the nailbed without any history of trauma should raise suspicion of a tumor and should be evaluated accordingly.
  - Do not remove the nail to evaluate injury to nailbed.

Selected Reading

Brown RE. Acute nail bed injuries. Hand Clin. 2002;18:561–75.

Dean B, Becker G, Little C. The management of the acute traumatic subungual haematoma: a systematic review. Hand Surg. 2012;17:151–4.

Incision and Drainage of Abscess

97

Nicholas D. Caputo, Karlene Hosford,
and Muhammad Waseem

97.1 Indications

- Abscess greater than 5 mm in diameter and in accessible areas (e.g., axilla, extremities, trunk)

97.2 Contraindications

- Absolute
  - Absence of fluctuation
  - Large, deep, and complicated (multiloculated) abscesses
  - Location
    - Perianal
    - Mastoid
- Relative
  - Location
    - Face (e.g., nose, nasolabial fold)
    - Palms
  - Coagulopathy
  - Recurrent pilonidal cysts (may mandate operative excision)
  - Area of cosmetic importance where aspiration may be preferred

97.3 Materials and Medications

- Incision and drainage tray (Fig. 97.1)
  - Drape
  - Betadine (povidone-iodine) swabs
  - 1 % lidocaine
  - 18- and 27-gauge needles
  - 12-mL syringes, gauze pads
  - #11 scalpel, mosquito clamps (hemostat)
  - Iodoform packing of appropriate size
- Ultrasound machine (Fig. 97.2)

N.D. Caputo, MD, MSc
Emergency Department Critical Care, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: Ncaputo.md@gmail.com

K. Hosford, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: krlhos@aol.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_97

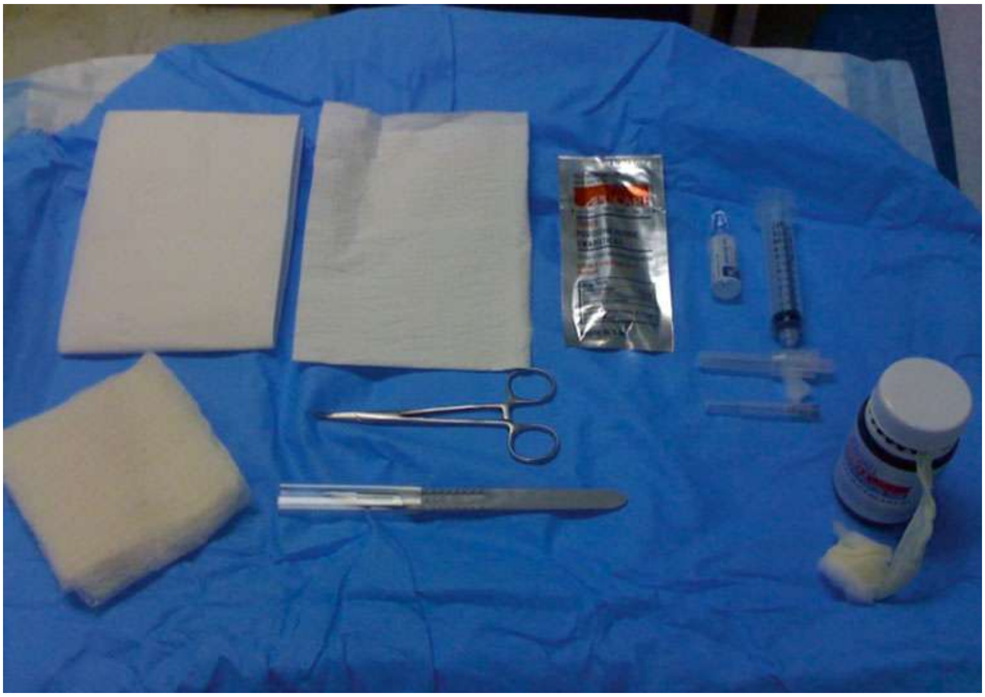

Fig. 97.1 Supplies necessary for incision and drainage

Fig. 97.2 Bedside SonoSite ultrasound

97.4 Procedure

1. Ultrasound (optional) may be helpful when abscess is suspected in the absence of fluctuation. Using the vascular probe (7 mHz), confirm the clinical suspicion of abscess and check the depth and width of the abscess (Fig. 97.3).
2. Sterile skin preparation with Betadine swab and sterile drape.
3. Anesthetize the appropriate area subcutaneously with 5 mL of 1 % lidocaine by inserting the 27-gauge needle at an acute angle into the intradermal space (Fig. 97.4a, b).
4. Using a #11 blade, make an approximately 1- to 2-cm skin incision over the desired area parallel to the Langer lines. The incision must approach into the abscess cavity (Fig. 97.4c).
   - Some physicians still advocate the technique of making a cruciate incision. This may leave a larger scar and should be discussed with patient before doing so because of cosmetic consequences.
5. Allow for spontaneous drainage. After resolution of drainage, you may express more pus with gentle downward pressure.
6. Using the hemostat, enter the incision to break any suspected loculations. This should be done with the clamps closed and curved part down. The clamps should then be opened and removed slowly (Fig. 97.4d).
7. After clearing the remaining loculations, the wound should be packed.
   - Evidence suggests that packing the abscesses does not prevent recurrence; however, this is still practiced. If packing the wound, follow the next step.
8. Take the iodoform packing with the hemostat, and place the packing into the incision site until no further packing will fit.
9. Cut the packing leaving a tail out of the incision site (Fig. 97.4e).
10. Apply dressing with 4×4 gauze and adhesive tape (2 in.).

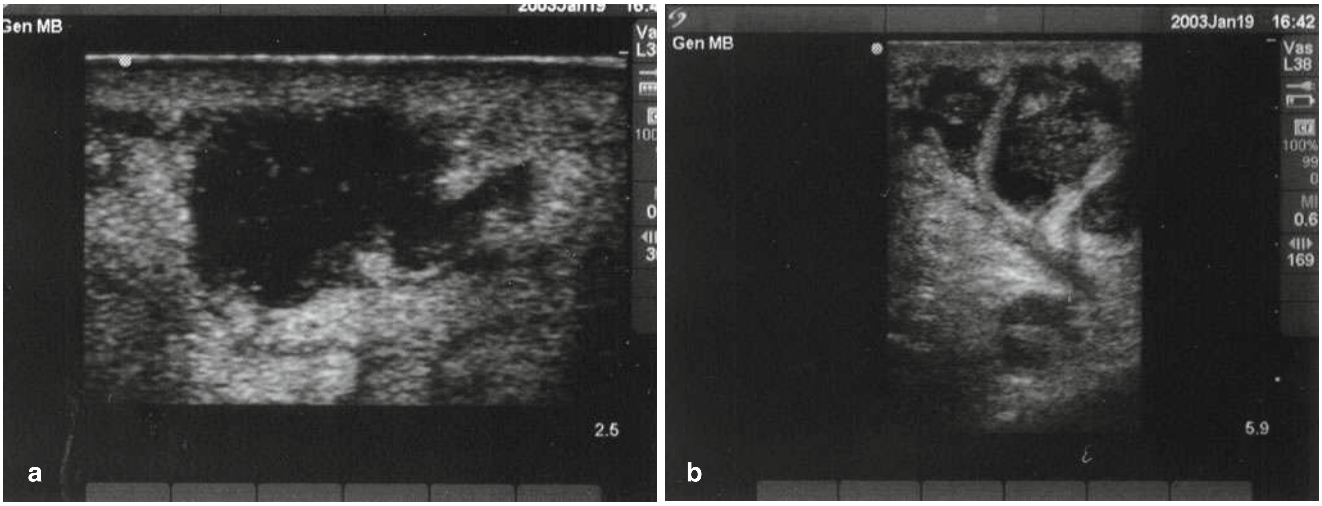

Fig. 97.3 (**a**) Example of an abscess as viewed on bedside ultrasound. (**b**) A multiloculated abscess

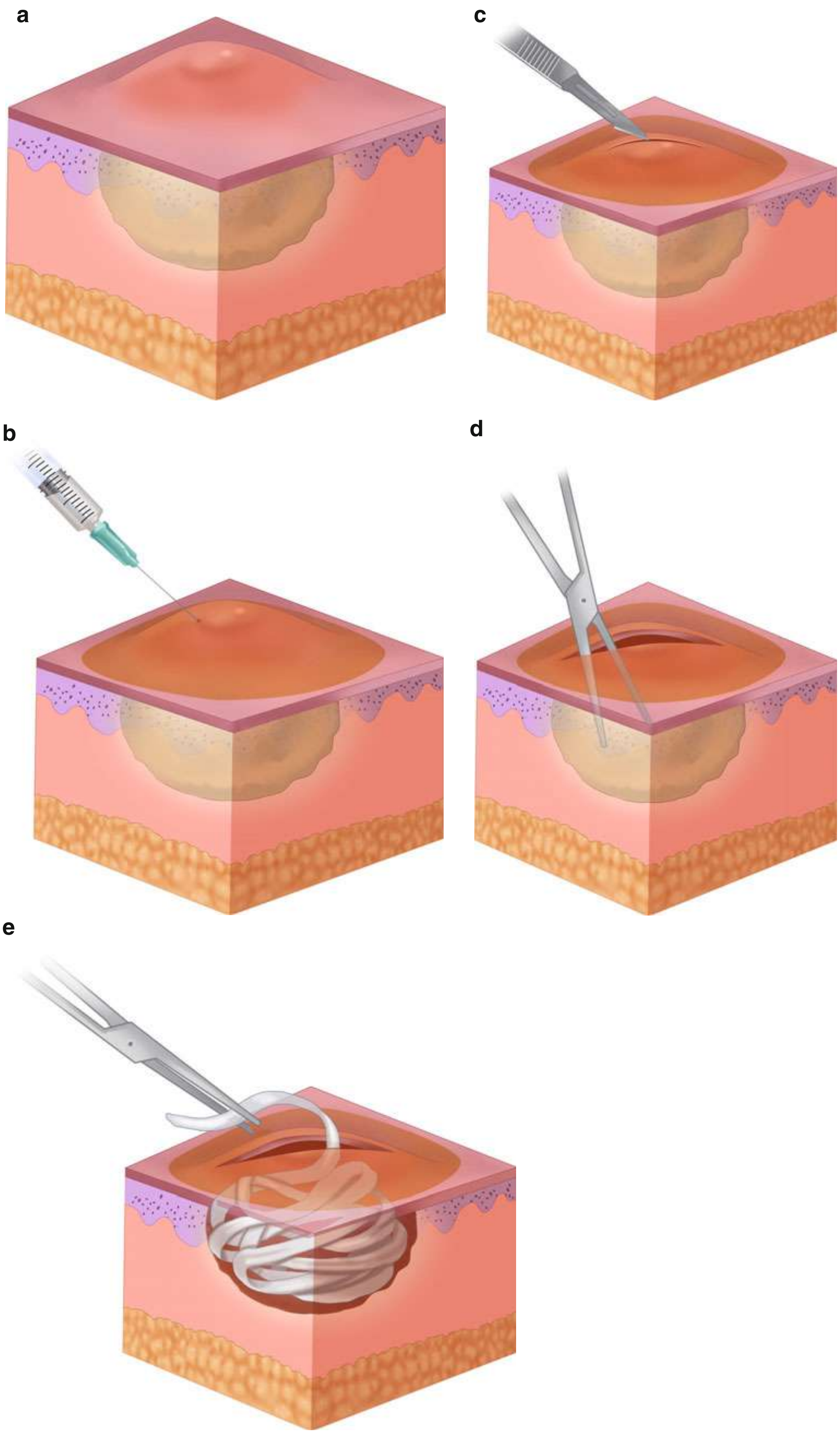

Fig. 97.4 (**a**) Abscess with overlying erythema. (**b**) Lidocaine injection in the superficial layer. (**c**) Linear incision with #11 blade. (**d**) Expression of purulent material and breaking of loculations with clamps. (**e**) Optional placement of packing

97.5 Complications

- Recurrence of abscess
- Progression of cellulitis
- Neurovascular injury to adjacent structures

97.6 Pearls

Antibiotic coverage is a controversial topic. Methicillin-resistant *Staphylococcus aureus* (MRSA) is a concern not only in the immunocompromised and diabetic patients. *S. aureus* has been detected in up to 51 % of patients with abscesses. Of these isolates, approximately 75 % were MRSA. Bactrim (trimethoprim/sulfamethoxazole) should be utilized for all prophylactic measures.

Selected Reading

Barnes SM, Milsom PL. Abscess: an open and shut case. Arch Emerg Med. 1988;5:200–5.

Burney RE. Incision and drainage procedures: soft tissue abscesses in the emergency service. Emerg Med Clin North Am. 1986;4:527–42.

Duong M, Markwell S, Peter J, Barenkamp S. Randomized, controlled trial of antibiotics in the management of community-acquired skin abscesses in the pediatric patient. Ann Emerg Med. 2010;55:401–7.

Fitch MT, Manthey DE, McGinnis HD, Nicks BA, Pariyadath M. Abscess incision and drainage. N Engl J Med. 2007;357:e20.

Frazee BW, Lynn J, Charlebois ED, Lambert L, Lowery D, Perdreau-Remington F. High prevalence of methicillin-resistant *Staphylococcus aureus* in emergency department skin and soft tissue infections. Ann Emerg Med. 2005;45:311–20.

Hankin A, Everett WW. Are antibiotics necessary after incision and drainage of a cutaneous abscess? Ann Emerg Med. 2007;50:49–51.

O'Malley GF, Dominici P, Giraldo P, et al. Routine packing of simple cutaneous abscesses is painful and probably unnecessary. Acad Emerg Med. 2009;16:470–3.

Part XIII

Orthopedic Procedures

Splinting

98

Christopher H. Stahmer and Muhammad Waseem

98.1 Indications

- Need for immobilization for fracture, dislocation, or soft tissue injury
- Suspicion for occult injury of an extremity
- Immobilization for pain management

98.2 Contraindications

- Absolute
  - Open fracture (requires operative intervention)
- Relative
  - Infection
  - Compartment syndrome

98.3 Materials and Medications (Fig. 98.1)

- Plaster of Paris
- Fast drying: 5–8 min to set
- Extra fast drying: 2–4 min to set
  - Variety of widths depending upon splint of choice:
  - Splints may take up to 2 days to dry and achieve maximum strength.
- Prefabricated splinting materials
  - Plaster OCL® (Orthopedic Casting Laboratories)
    - 10–20 sheets of plaster with padding and cover
    - Faster setup time but less customizable
  - Fiberglass splints
    - Cure rapidly
    - Less messy
    - Less moldable
    - Stronger and lighter
- Stockinette.
  - Protects the skin.
  - Variety of sizes available.
- Soft wrap (Webril™).
  - Provides padding.
  - Five to six layers depending on anticipated swelling.
    - Too much padding reduces the stability of the splint.
  - Use extra padding over bony prominences.
  - Pad between digits for splinting of digits.
  - Avoid wrinkles, which generate pressure points.
  - Do not wrap circumferentially.
    - Increased risk of ischemia.
- Ace wraps.
  - Variety of sizes depending.
    - Larger widths over legs.
    - Narrow widths around fingers and joints.
      - Avoid bunching by using narrow widths at joints.
- Water
  - Warm water and splint sets more quickly but increases the risk of burns.
    - Splint drying is an exothermic or heat-releasing reaction.
  - Hot water leaves less time to mold the splint.

C.H. Stahmer, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: stahmer@gmail.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_98

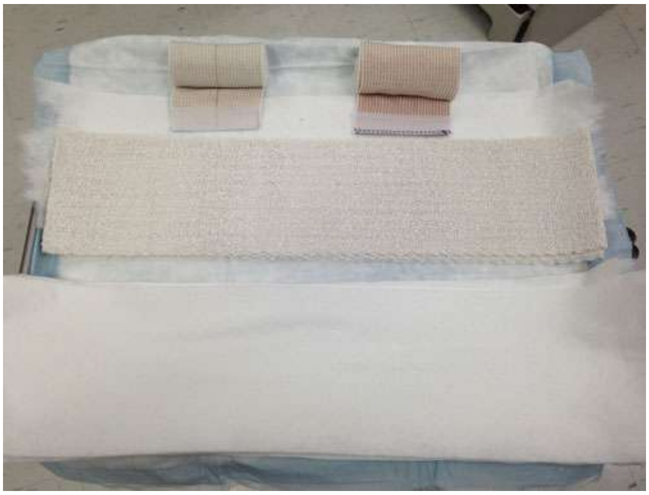

Fig. 98.1 Posterior splint materials: plaster of Paris, Ace wraps, soft roll. Note there are two layers of soft wrap: the inner layer to face the patient (eight layers) and the outer layer to pad the exterior (two layers)

98.4 Procedure

1. Completely expose and examine the afflicted body part for tissue, vascular, or neurological injury.
   - Address respective injuries before proceeding.
2. Lay out all splinting materials before initiating procedure.
   - Layer plaster of Paris.
     - Upper extremity: 8–10 layers.
     - Lower extremity: 12–15 layers.
     - Up to 20 for a large person.
     - More layers of plaster of Paris increase the risk of burn and the weight of the splint.
3. Administer appropriate anesthesia.
   - Conscious sedation
   - Hematoma block
   - Intra-articular injection
   - Intravenous pain medication
   - Oral pain medication
4. Hang fractures as indicated for improved success of reduction to relax muscles before reduction attempt.
5. Reduce afflicted extremity.
6. While maintaining reduction, apply respective splint.
7. Apply in the following order for plaster of Paris splint.
   - Stockinette (not necessary).
   - Soft wrap.
   - Select appropriate layers of plaster of Paris.
   - Prepare plaster of Paris to create splint:
     - Layer plaster with no overlap.
     - Submerge completely into water.
     - Crumple into ball without letting go of the ends of the splint.
     - Release the lower end of the splint while holding the top tightly together.
     - Run fingers in a “squeegee” manner from top to bottom to smooth the splint (Fig. 98.2).
       - This also removes excess water.
       - Repeat until the splint is smooth and free of dripping water.
     - Apply soft wrap layers to the splint.
       - Apply thicker layer to the patient’s body.
       - Apply two or three layers of soft wrap to the exterior of plaster of Paris for padding and to facilitate drying.
     - Apply Ace wrap to hold the splint and assist in contouring the splint to the patient’s extremity (Fig. 98.3).
       - Applying the Ace wrap too tightly may cause ischemia. Observe the patient after splinting for 30 min for tingling, burning, pain, or discomfort.
     - Mold the splint without making indentations with the fingertips (Fig. 98.4).
       - An indentation may cause a pressure point, which may result in an ulcer.

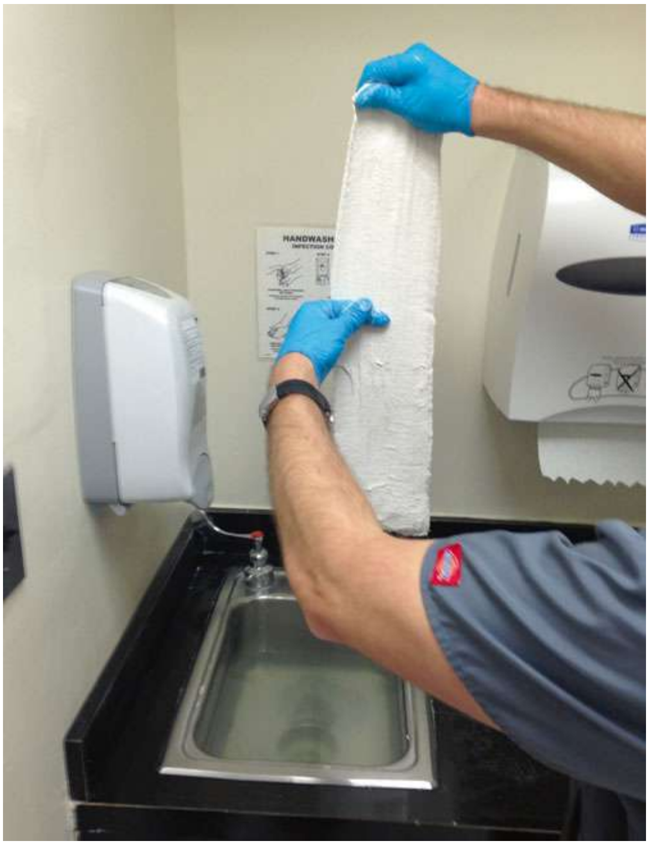

Fig. 98.2 Hold the top of the saturated plaster securely with one hand while removing excess water with the other hand

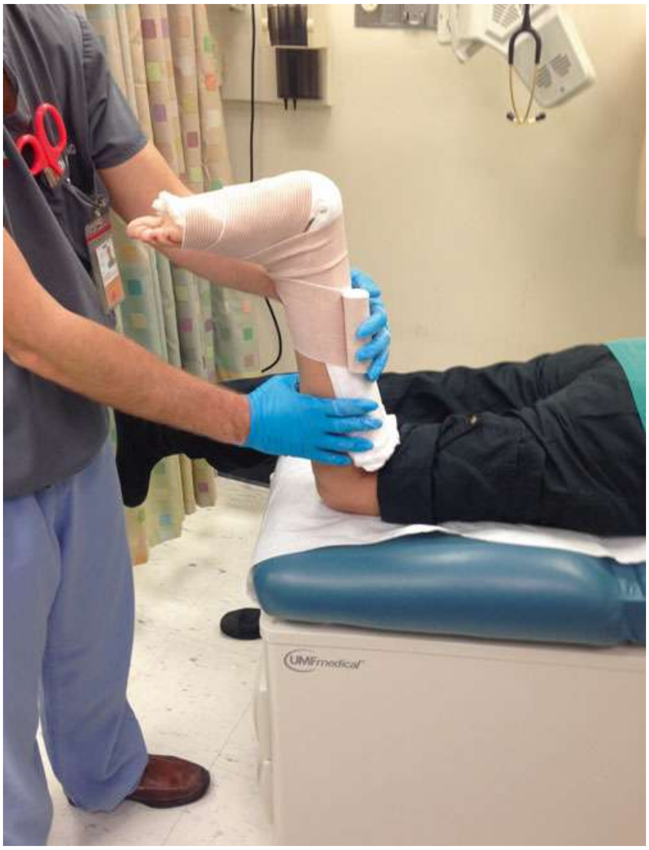

Fig. 98.3 Apply Ace wrap to hold the splint and assist in contouring the splint to the patient's extremity

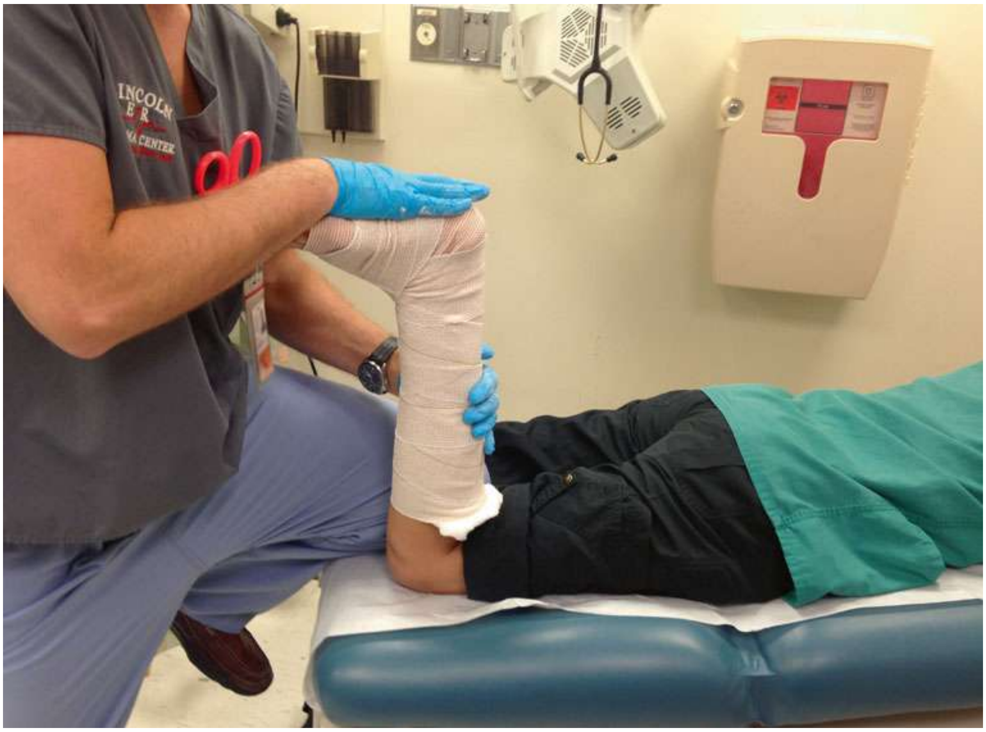

Fig. 98.4 Hold the splint in a neutral anatomical position while taking care not to make indentations with the fingertips

  - Allow the splint to cure while the practitioner maintains the appropriate position. This will take approximately 5 min depending upon water temperature and splint thickness.

98.5 Complications

- Ischemia may result in compartment syndrome.
  - Advise the patient to unwrap the splint for the following indications.
    - Increasing pain.
    - Discoloration of fingers, toes, or the splinted extremity.
    - Loss of sensation of splinted extremity.
- Burns
  - Plaster drying releases heat.
  - Increased risk with limited layers of padding.
  - If pain is troubling the patient, remove the splint and add more padding.
- Pressure sores
  - Apply ample padding.
  - Smooth all wrinkles.
  - Instruct the patient to return for increased discomfort.
- Infection
  - Clean and débride all devitalized tissue before application.
  - Requires close follow-up to reevaluate wounds.

Selected Reading

Fitch MT, Nicks BA, Pariyadath M, McGinnis HD, Manthey DE. Basic splinting techniques. N Engl J Med. 2008;359:e32.

Marx JA, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Simon R, Sherman S, Koenigsknecht S. Emergency orthopedics—the extremities. New York: McGraw-Hill; 2007.

Ulnar Gutter Splint

99

Jeffrey Kile, Katrina John, and Amish Aghera

99.1 Indications

- Fractures and soft tissue injuries of the ring or little finger
- Fractures of the neck, shaft, or base of the fourth or fifth metacarpal

99.2 Contraindications

- Relative
  - Evidence of compartment syndrome or any neurovascular compromise

99.3 Materials (Fig. 99.1)

- Splint roll (plaster of Paris, prefabricated foamcore, or fiberglass)
- Stockinette
- Cotton padding, such as simple cotton sheet wadding or a newer alternative such as Webril (Curity), Specialist (Johnson & Johnson), and so on
- Elastic bandages, such as Ace (3M) or similar
- Shears
- Adhesive tape
- Gloves

J. Kile, MBBS, PhD, MPH (✉) • K. John, MBBS
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: jeffrey.kile@gmail.com; trenjohn@me.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_99

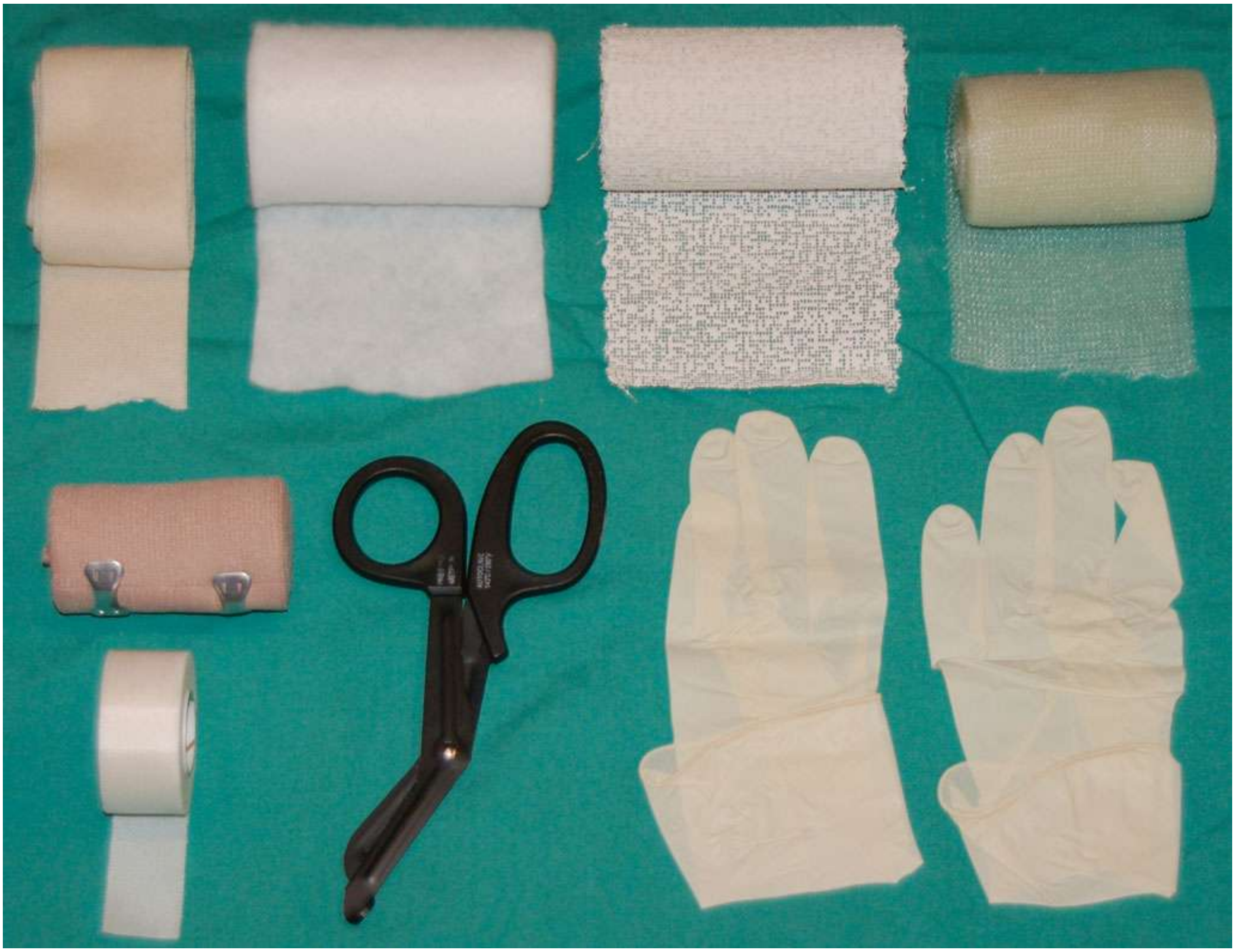

Fig. 99.1 Materials and medications

99.4 Procedure

1. Select the appropriate diameter stockinette according to the size of the forearm.
    - Generally, 3-in. diameter stockinette is used for the ulnar gutter splint.
2. Cut the stockinette to a length approximately 20 cm longer than the total length of the desired splint.
3. Apply the stockinette to the injured limb so that the stockinette extends approximately 10 cm beyond the region to be splinted proximally and distally.
4. Cut the distal aspect of the stockinette to free the thumb and also the index and middle finger (Fig. 99.2).
5. Insert cotton padding cut to the appropriate size between the ring and the little fingers to prevent skin degradation (Fig. 99.3).
6. Wrap cotton padding circumferentially around the entire region to be splinted, with each turn overlapping the previous turn by approximately 25 % of its width.
    - Generally, 3-in.-wide padding is used for the ulnar gutter splint.
    - Apply padding to an approximate thickness of 1 cm, with extra padding placed over the bony prominences of the wrist and carpometacarpal joints.
    - Extend the padding at least an inch beyond the desired splint length so that it may later be folded back over the jagged ends of the plaster or fiberglass splint roll (Fig. 99.4).
7. Unwrap and prepare an appropriate length splint.
    - The width of splinting material used should be approximately one-half the circumference of the extremity measured at the wrist, such that it is wide enough to extend approximately halfway around the distal forearm once applied.
    - Layer the splint eight sheets thick if using plaster roll and six sheets thick if using fiberglass roll.
    - It is best to begin with a generous length of splinting material because both plaster and fiberglass splint rolls shrink slightly once moistened. If the wet splint turns out too long, the ends can be either folded back or cut with shears before hardening.
8. Apply the dry splinting material to the ulnar aspect of the extremity from the midforearm to just beyond the distal interphalangeal joint of the little finger and cut the splint to approximately 20 cm longer than the desired length.
9. Submerge the dry splint in water until bubbling stops.
10. Remove the splint from water, place on a hard flat surface, and smooth out the excess water to ensure no wrinkles are present.
11. Apply to the splint to the limb over the cotton padding.
    - If the splint is too long at this stage, the ends may be folded back or cut with shears to the proper length.

12 Once positioned properly, fold the side of the splint up around the ulnar aspect of the forearm and hand to form a gutter (Fig. 99.5).

13. Fold the underlying cotton padding and stockinette back over the ends of the splint, which both protects the skin and holds the splint in place.
14. Secure the loose end(s) of the elastic bandage with adhesive tape.
    - Avoid using the metal clips often packaged with elastic bandages because these can become displaced and embed in the skin.
15. Manipulate the forearm and hand into the "neutral" position: (1) the wrist in slight extension (10–20°), (2) the metacarpophalangeal joints in 50° of flexion, and (3) the proximal and distal interphalangeal joints of the ring and little finger in slight flexion (10–15°) (Fig. 99.6).
    - When splinting a boxer's (i.e., metacarpal neck) fracture, the metacarpophalangeal joint should be flexed to 90°.
16. Mold the splint to the contour of the extremity (Fig. 99.7).
    - Use only the palms of the hands when molding because the fingertips may cause indentations, resulting in excessive skin pressure.
17. Wrap the extremity with an elastic bandage in a distal-to-proximal direction to secure the splint in place (Fig. 99.8).
18. Ensure the extremity is neurovascularly intact.
19. Instruct the patient to loosen the elastic bandage if it feels too tight.

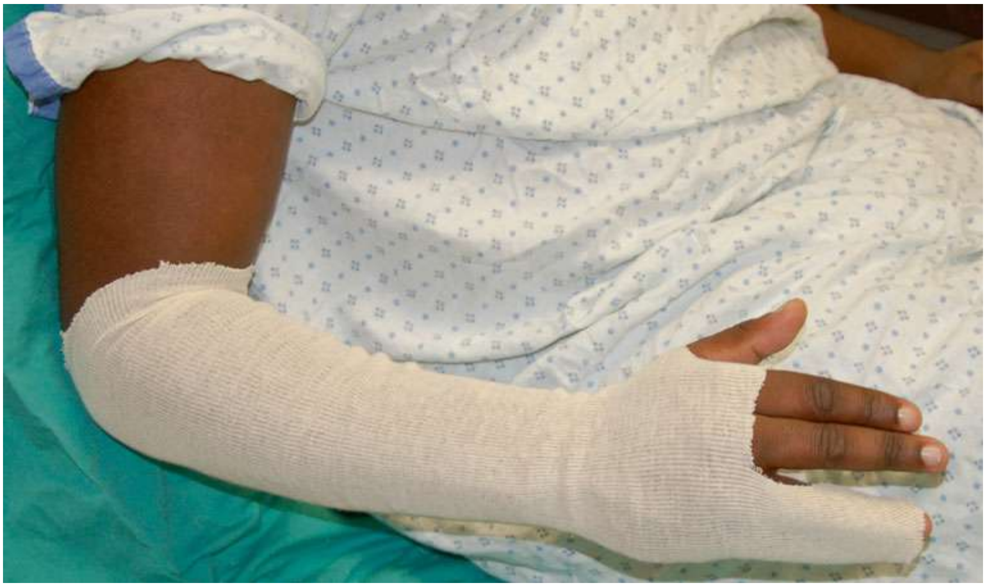

Fig. 99.2 Stockinette applied to the forearm and hand

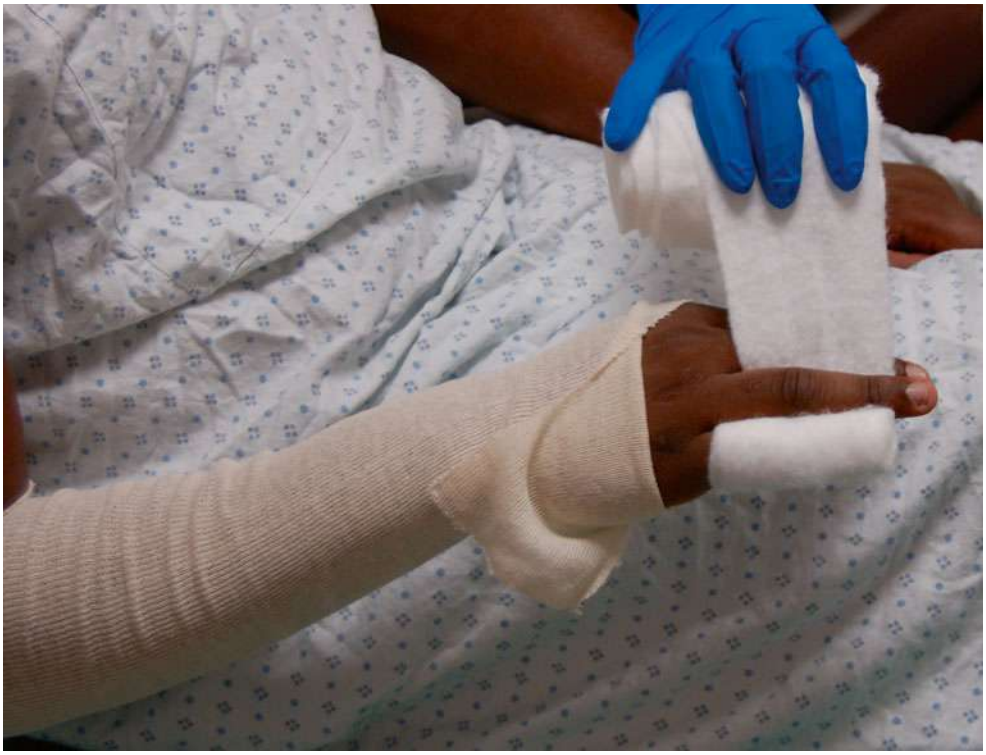

Fig. 99.3 Cotton padding applied to the ring and little fingers

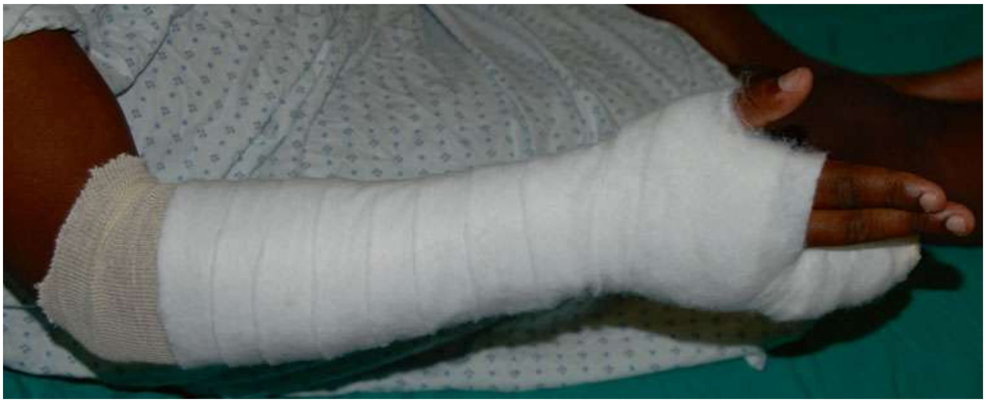

Fig. 99.4 Cotton padding applied to the forearm and hand

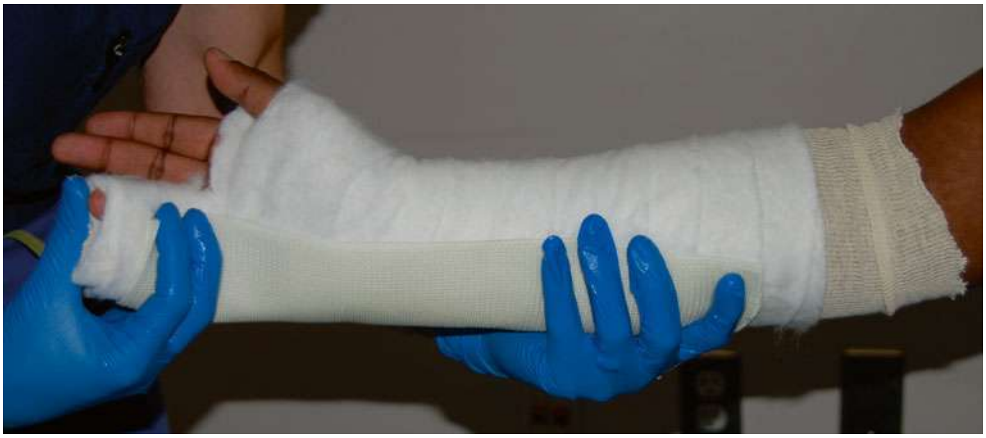

Fig. 99.5 Splint folded around the ulnar aspect of the forearm and hand

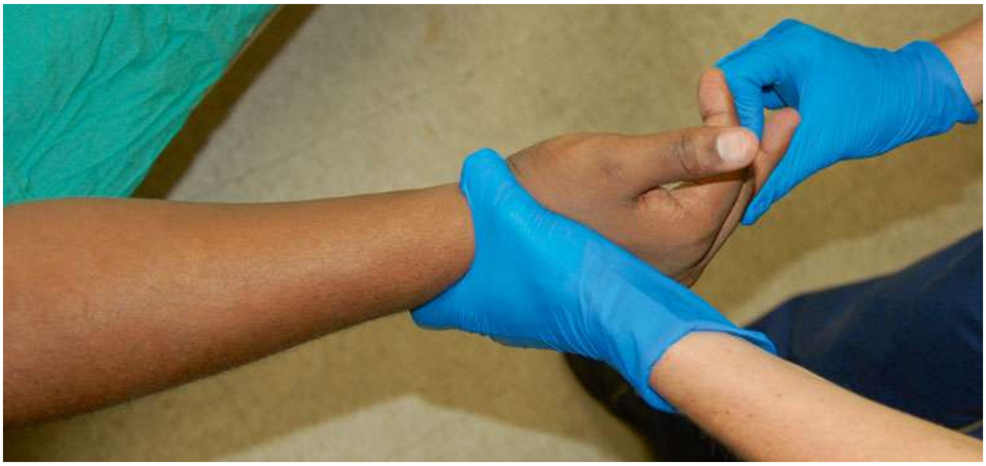

Fig. 99.6 Forearm and hand in the neutral position (without splint)

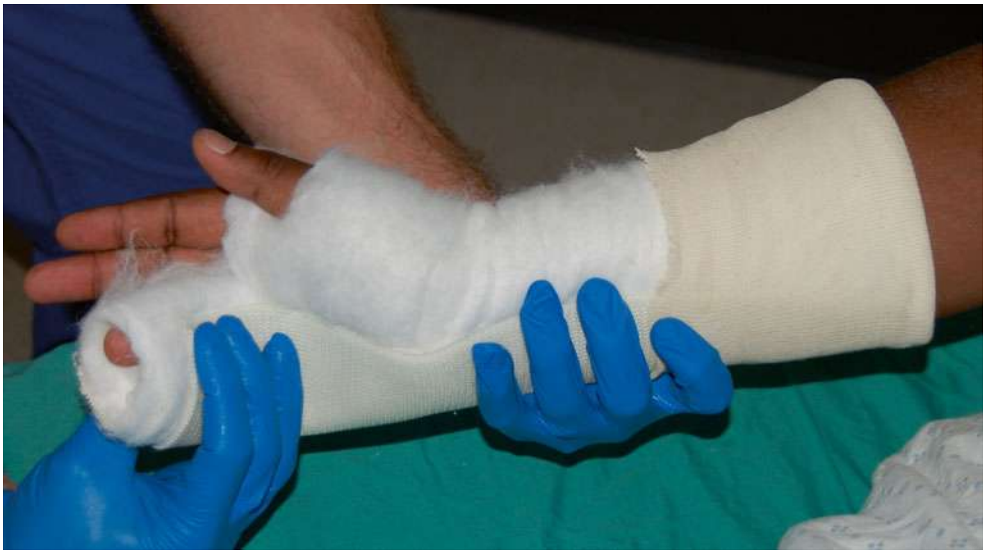

Fig. 99.7 Splint molded to maintain the forearm and hand in the neutral position

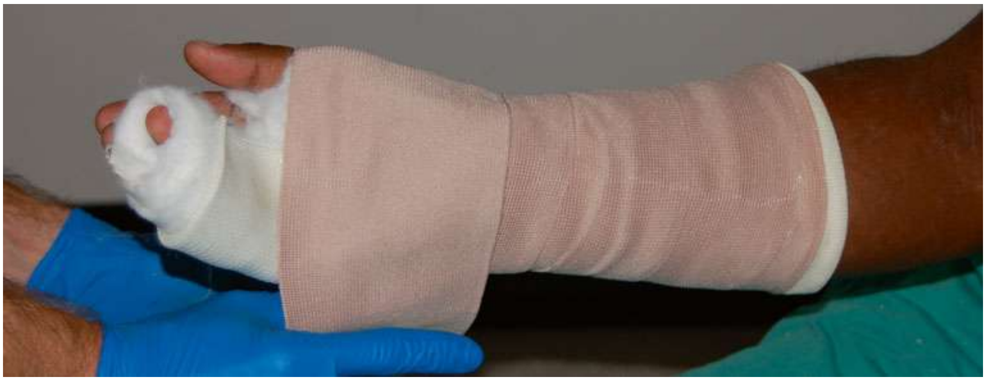

Fig. 99.8 Splint secured in place with an elastic bandage

99.5 Complications

- Neurovascular or other soft tissue compromise (if the splint/bandage is too tight)
- Soft tissue degradation (if the splint is left in place too long)

99.6 Pearls and Pitfalls

- Pearls
  - Simple plaster of Paris splints are inexpensive and allow a thoroughly customizable fit but can be damaged by water and require more time to set and more clean up than prefabricated foamcore or fiberglass splint rolls.
  - Simple fiberglass splints set quickly, are not damaged by water, are stronger and lighter than simple plaster and prefabricated splints, and offer a fully customized fit, but are not applied as quickly as prefabricated splints.
  - Prefabricated fiberglass splints are quickly applied, require virtually no clean up, and are not damaged by water, but are relatively expensive and provide a somewhat less customizable fit than simple splint rolls.
  - Stockinette protects the skin and, when folded back over the ends of the plaster or fiberglass splint roll, holds the wet splint in place before the elastic bandage is placed, and provides a padded rim with a professional appearance.
- Pitfalls
  - Avoid wrinkling the cotton padding applied between the stockinette and the splinting material because, once under the pressure of the elastic bandage, wrinkles can cause unnecessary skin pressure.
  - Avoid using more or fewer layers of splinting material than recommended. Additional layers can result in excessive heat during the setting process and a splint that is too heavy, whereas insufficient layers can result in a splint that is too weak.

Selected Reading

Harrison BP, Hilliard MW. Emergency department evaluation and treatment of hand injuries. Emerg Med Clin North Am. 1999;17:793–822.

Henry MH. Fractures of the proximal phalanx and metacarpals in the hand: preferred methods of stabilization. J Am Acad Orthop Surg. 2008;16:586–95.

Margic K. External fixation of closed metacarpal and phalangeal fractures of digits. A prospective study of one hundred consecutive patients. J Hand Surg Br. 2006;31:30–40.

McMahon PJ, Woods DA, Burge PD. Initial treatment of closed metacarpal fractures. A controlled comparison of compression glove and splintage. J Hand Surg Br. 1994;19:597–600.

Viegas SF, Tencer A, Woodard P, Williams CR. Functional bracing of fractures of the second through fifth metacarpals. J Hand Surg Am. 1987;12:139–43.

Shoulder Dislocation Reduction Techniques

100

Katrina Skoog Nguyen, L. Connor Nickels,
and Rohit Pravin Patel

100.1 Indications

- Subjective history of new-onset dislocation or recurrent dislocations combined with clinical assessment consistent with shoulder dislocation
  - Anterior dislocations (96 %)
    - Typical mechanism of injury being indirect, with combination of abduction, extension, and external rotation. Rarely, the etiology is a direct blow to the posterior shoulder.
    - Prominent acromion with a palpable drop off below the acromion and subclavicular region fullness is consistent with anterior shoulder dislocation.
  - Posterior dislocations (4 %)
    - Mechanism of injury is indirect with a combination of internal rotation, adduction, and flexion. Precipitating events include seizure, electrical shock, and falls.
    - More subtle presentation. Patient will maintain arm locked in internal rotation and adduction; he or she cannot externally rotate. Shoulder is flattened anteriorly and rounded posteriorly.
    - Ultrasound can be used to prevent missed or delayed diagnosis (Figs. 100.1 and 100.2 show probe positioning and a diagram of abnormal ultrasound anatomy).
  - Inferior dislocations (luxation erecta)
    - Arm will be held fixed in overhead position.
- Radiographs reveal shoulder dislocation.
- Ultrasound can be used to identify the nature of the dislocation (anterior or posterior) and can be determined by the position of the humeral head relative to the transducer and glenoid. Although at this point, it should not replace radiographs owing to missed fractures. Advantages may include less radiation (decreased need for postreduction x-rays) and re-sedation if reduction is not complete.

K. Skoog Nguyen, DO
Northwest Community Hospital, Arlington Heights, IL, USA
e-mail: katrinaskoog@hotmail.com

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_100

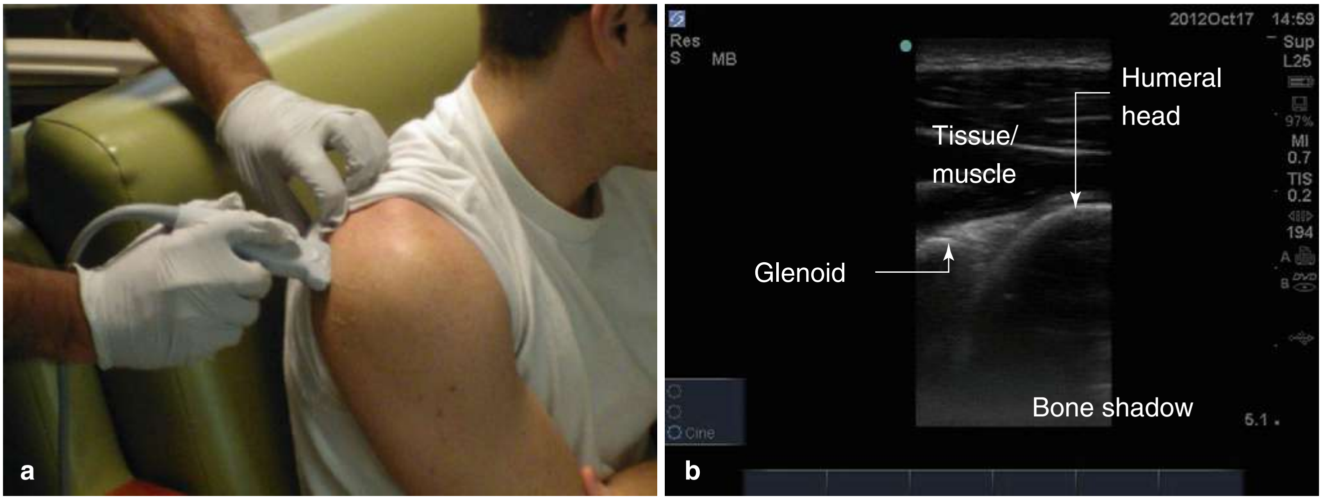
a
b
2012Oct17 14:59
Res
S MB
Sup
L25
97%
MI
0.7
TIS
0.2
194
Cine
5.1
Humeral head
Tissue/ muscle
Glenoid
Bone shadow

Fig. 100.1 (**a**, **b**) Ultrasound image of normal shoulder anatomy. "Dot fits the Dot" means when looking at the ultrasound machine from the sonographer's standpoint, the side of the probe marker corresponds to the side the marker is on the screen. This ensures when doing procedures, the direction of needle correction is the same as the orientation of the probe (Images courtesy of Dr. Rohit Patel)

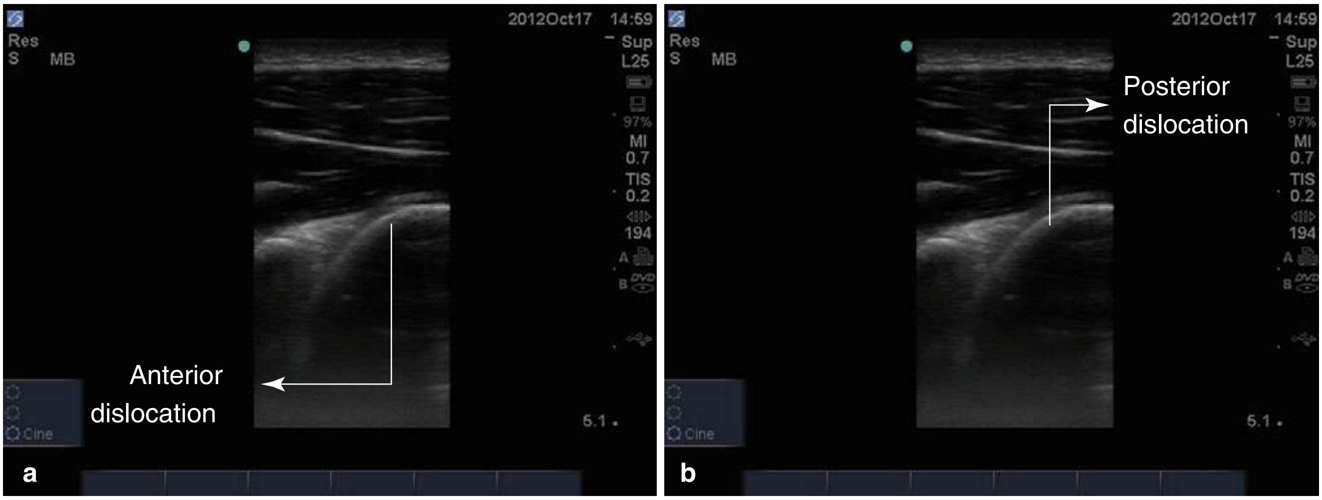
a
b
2012Oct17 14:59
Res
S MB
Sup
L25
97%
MI
0.7
TIS
0.2
194
Cine
5.1
Anterior dislocation
Posterior dislocation

Fig. 100.2 Anterior (**a**) and posterior (**b**) dislocations (Images courtesy of Dr. Rohit Patel)

100.2 Contraindications

- Associated fracture
  - This warrants orthopedic evaluation.
- Associated neurovascular deficit
  - May attempt reduction once but avoid multiple attempts.

100.3 Materials and Medications

- 1 % lidocaine, with syringe and needle and povidone-iodine prep if administering local anesthesia
- Moderate sedation medications if administering moderate sedation
- Bed sheet for traction-countertraction method
- Dangling weight for Stimson maneuver

100.4 Procedure

- Physical examination
  - Compare affected with unaffected shoulder.
  - Perform a complete neurovascular examination: test axillary, radial, ulnar, and median nerves for sensory deficit and motor function.
- Radiographs
  - Always obtain before attempting reduction for assessment of possible fracture and type and position of dislocation.
  - Obtain three views: anteroposterior, scapular Y, and axillary lateral views.
  - Anterior dislocations: humeral head appears anterior to the glenoid fossa on lateral or Y views.
  - Posterior dislocations: on anteroposterior view (vacant glenoid sign, 6-mm sign, lightbulb sign; on lateral or Y view: humeral head appears posterior to glenoid fossa).
- Pain management and sedation
  - Decide whether to use intra-articular lidocaine versus procedural sedation and analgesia.
  - For intra-articular lidocaine
    - Use 10–20 mL of 1 % lidocaine.
    - Attach a 1.5-in., 20-gauge needle.
    - Prepare the shoulder with povidone-iodine.
    - Insert the needle lateral to the acromion process and 2 cm inferiorly into the sulcus.
    - After withdrawing to ensure that the needle is not in a vessel, inject 10–20 mL lidocaine into the joint.
- Reduction techniques: it is important for the emergency department physician to be familiar with several different techniques. The following techniques are presented:

100.4.1 Stimson Maneuver (Fig. 100.3)

1. Patient is placed prone with 2.5–5 kg of weight hanging from the wrist.
2. Reduction may be facilitated by traction and external rotation of the arm.
3. A success rate of 96 % has been reported using the combined prone position, hanging weights, intravenous drug therapy, and scapular manipulation.

- Advantage: can be performed by one person only.
- Disadvantages: requires time to gather materials; the danger involved in the patient falling off the stretcher, requiring staff to monitor the patient.

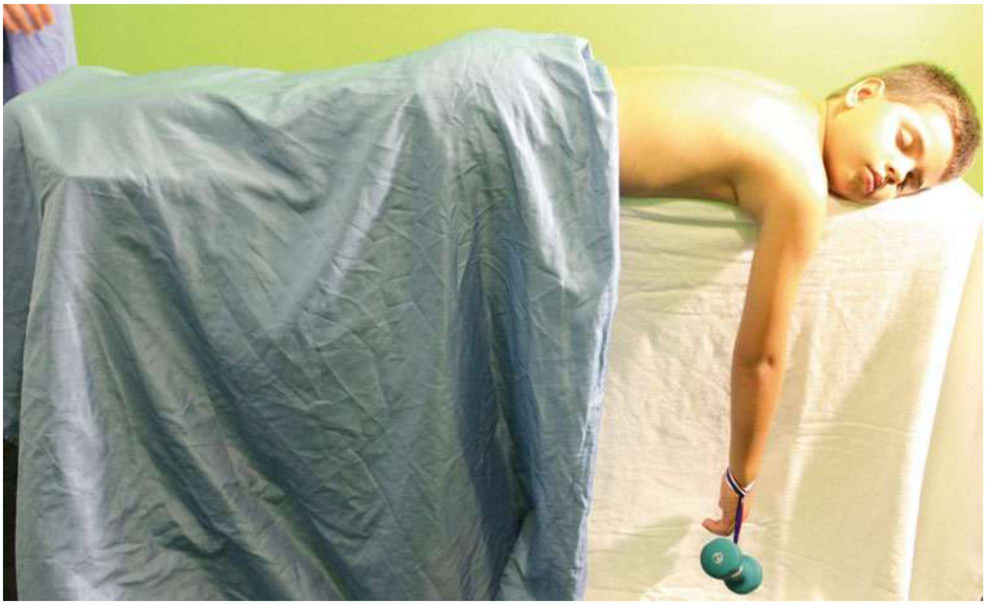

Fig. 100.3 Stimson maneuver

100.4.2 Scapular Manipulation Technique (Fig. 100.4)

1. Place the patient in the prone position with the affected arm hanging downward.
2. Apply traction down on the arm.
3. Locate the inferior tip of the scapula. Simultaneously push the inferior tip of the scapula medially toward the spine and use the other hand push the superior scapula laterally.

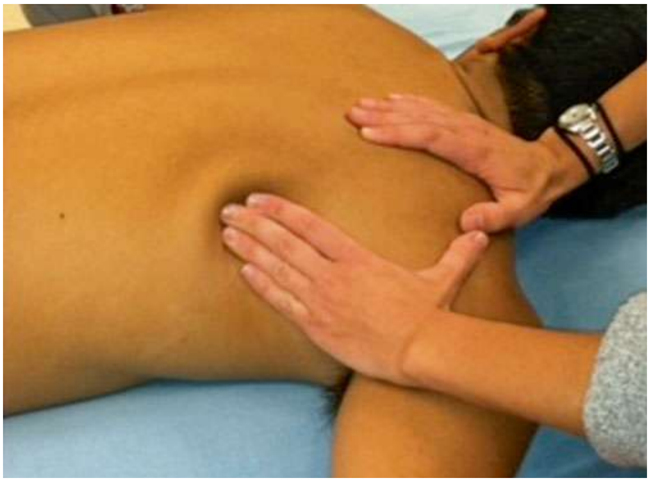

Fig. 100.4 Scapular manipulation method

- Advantages: high success rate, greater than 90 %; very safe to perform.
- Disadvantages: it requires the patient to assume the prone position; may require another person to perform traction.

100.4.3 External Rotation Method (Fig. 100.5)

1. Place the patient in the supine position with the affected arm adducted directly next to the patient's side with the elbow flexed to 90°.
2. The operator uses one hand to direct downward traction on the affected arm while maintaining it next to the patient's side.
3. The operator uses the other hand to hold the patient's wrist and guide the arm into slow external rotation.
4. Reduction usually takes place between 70° and 110° of external rotation.

- Advantages: requires no strength by operator; well tolerated by patients.
- Disadvantage: patient may have persistent dislocation during procedure, requiring operator to make adjustments.

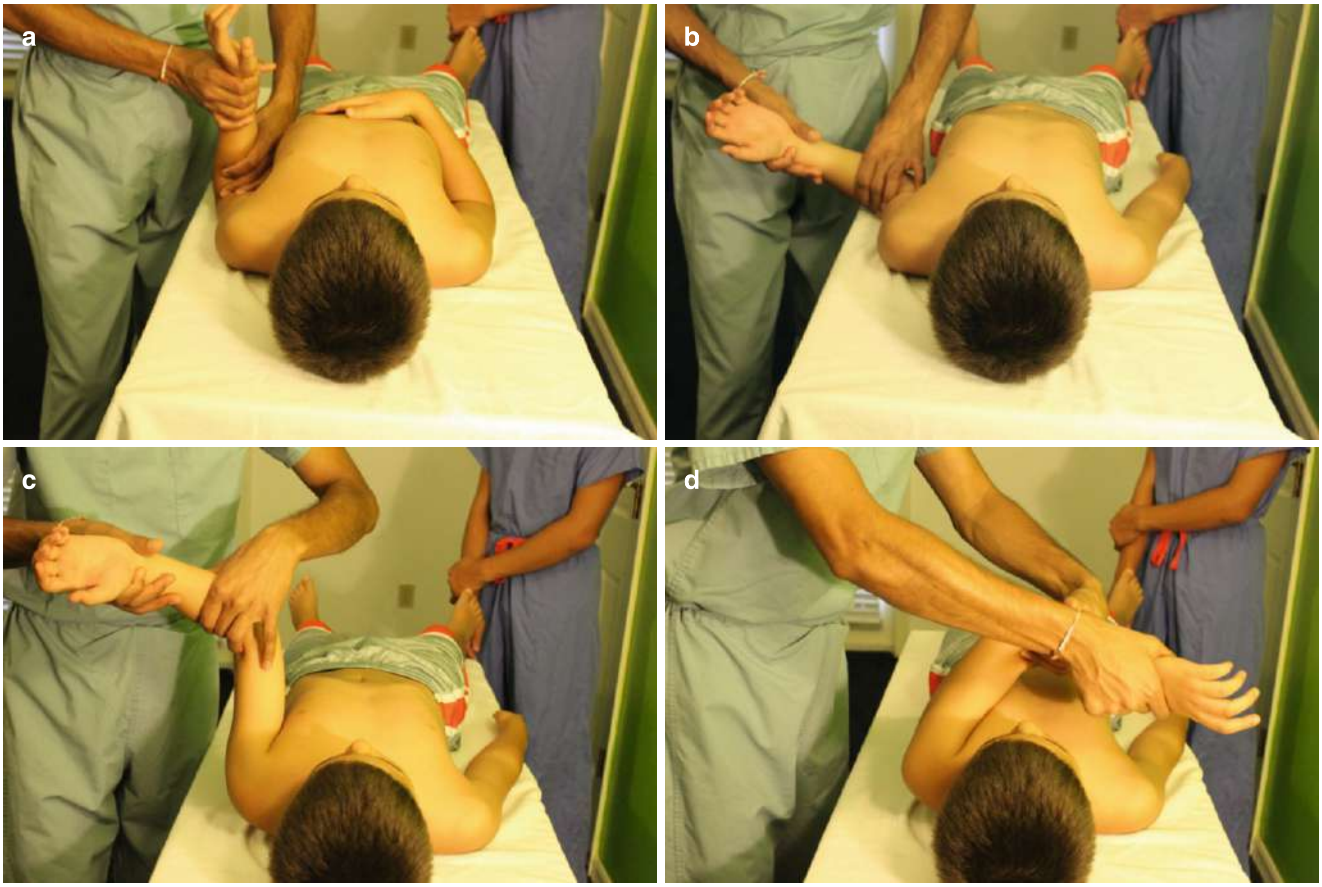

Fig. 100.5 (**a–d**) Kocher technique: external rotation method

100.4.4 Milch Technique (Fig. 100.6)

1. Technique looks as though one is reaching up to grab an apple from a tree.
2. Abduct the injured arm up to the overhead position.
3. Once in the overhead position, apply gentle vertical traction with external rotation.
4. An adjustment may need to be made if the reduction does not occur easily; push the humeral head upward into the glenoid fossa.

- Advantages: lack of complications; patient tolerance
- Disadvantage: variable success rate reported: 70–90 %

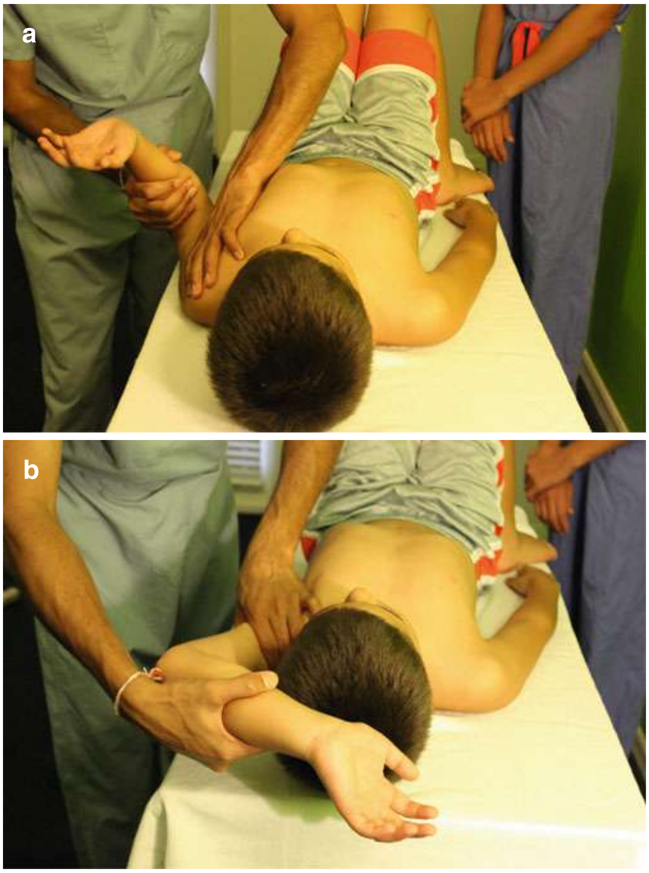

Fig. 100.6 (**a**, **b**) Milch technique

100.4.5 Spaso Technique (Fig. 100.7)

1. Place the patient in the supine position.
2. Operator grasps the affected arm at the wrist and lifts the straight arm directly upward while applying longitudinal traction.
3. Apply external rotation.

- Advantages: single operator, high level of success
- Disadvantage: may require more time to allow the shoulder muscles to relax

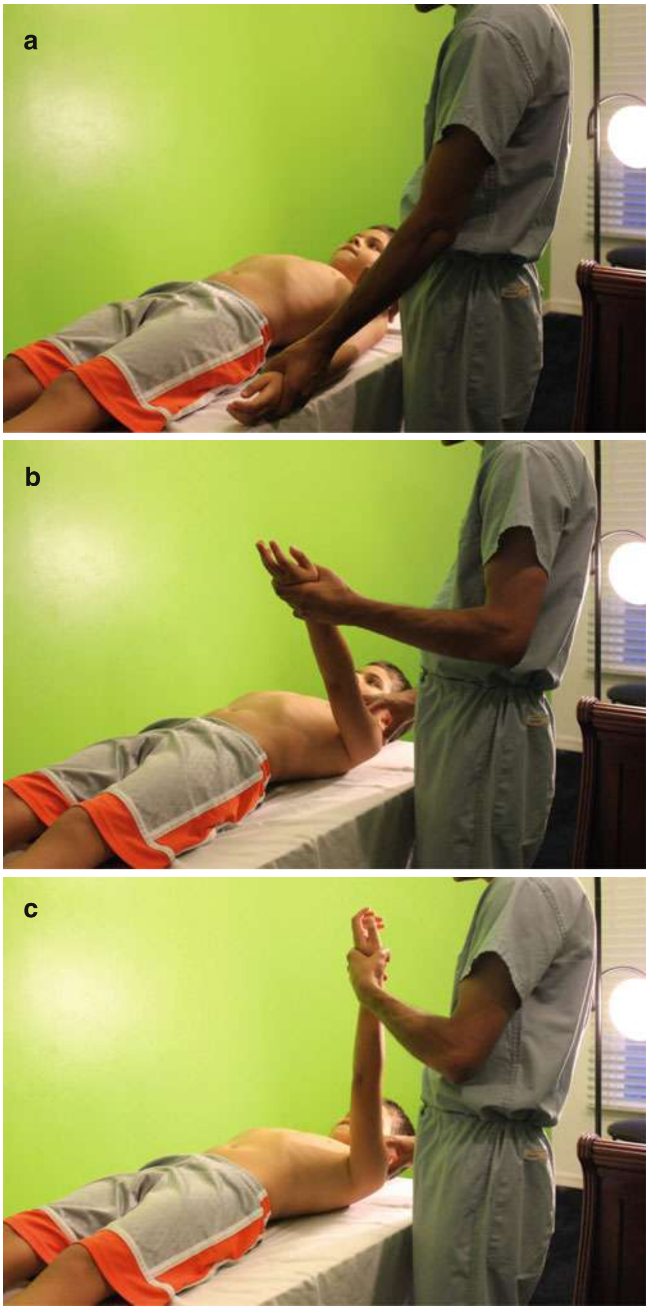

Fig. 100.7 (**a**–**c**) Spaso technique (Photographs courtesy of Dr. Pratik S. Patel)

100.4.6 Traction-Countertraction Technique (Fig. 100.8)

1. With the patient is sitting up, have an assistant wrap a sheet around the upper chest and under the axilla of the affected shoulder. Have the assistant wrap the sheet behind her or his back. Now have the patient lay supine.
2. Wrap another sheet around the flexed elbow of the affected arm and behind the operator's back.
3. Both the operator and the assistant lean back, applying gentle traction.

- Advantage: many older physicians are familiar with this method and, therefore, have a high degree of success.
- Disadvantages: requires two people; may cause skin tears on elderly patients.

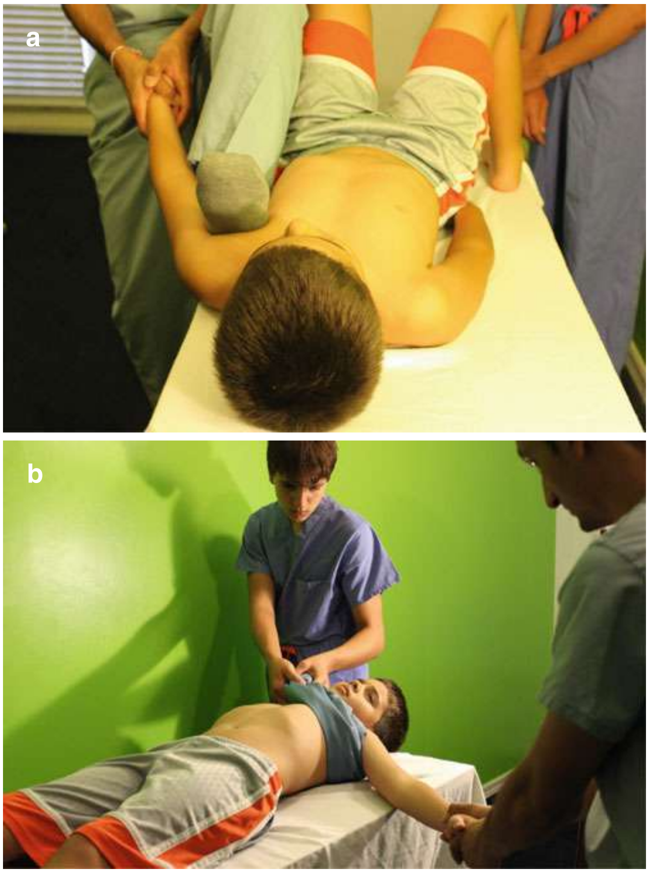

Fig. 100.8 (**a**, **b**) Hippocrates method/traction-countertraction method

100.4.7 Posterior Shoulder Dislocation Reduction

1. Give adequate premedication.
2. Place the patient supine and apply lateral traction on the proximal humerus.
3. Have an assistant apply anterior pressure to the posteriorly located humeral head.

- Advantage: logical methods for reduction
- Disadvantages: require sufficient premedication because often posterior dislocations present late; may require open reduction

100.4.8 Postreduction

- Obtain postreduction x-rays. There is some literature on using ultrasound to confirm adequate reduction, which allows repetitive assessments throughout procedure, as well as reduce radiation (see Fig. 100.2 for ultrasound of anterior and posterior dislocations).
- Do a postreduction neurovascular examination.
- Sling and swath or shoulder immobilizer for 2–3 weeks.
- Orthopedic follow-up in 1 week.

100.5 Complications

- Fractures
- Adhesive capsulitis, or frozen shoulder; especially a concern in the elderly with prolonged immobilization in sling
- Brachial plexus injury, especially of the axillary nerve
- Vascular laceration, most commonly of the axillary artery
- Rotator cuff tears

100.6 Pearls

- It is imperative to document the pre- and postreduction neurovascular status in the medical record.
- If unsure whether the reduction was successful, attempt to place the palm of the injured extremity on the contralateral shoulder. This is a good sign the reduction was successful.

Acknowledgment The authors would like to thank Karthik Stead for serving as the subject in many of the photographs in this chapter.

Selected Reading

Beck S, Chilstrom M. Point-of-care ultrasound diagnosis and treatment of posterior shoulder dislocation. Am J Emerg Med. 2013;31:449.e3–5.

Blakeley CJ, Spencer O, Newman-Saunders T, Hashemi K. A novel use of portable ultrasound in the management of shoulder dislocation. Emerg Med J. 2009;26:662–3.

Dala-Ali B, Penna M, McConnell J, Vanhegan I, Cobiella C. Management of acute anterior shoulder dislocation. Br J Sports Med. 2014;48(16):1209–15.

Simão MN, Noqueira-Barbosa MH, Muqlia VF, Barbieri CH. Anterior shoulder instability: correlation between magnetic resonance arthrography, ultrasound arthrography, and intraoperative findings. Ultrasound Med Biol. 2012;38:551–60.

Yuen CK, Chung TS, Mok KL, Kan PG, Wong YT. Dynamic ultrasonographic sign for posterior shoulder dislocation. Emerg Radiol. 2011;18:47–51.

Elbow Dislocation Reduction

101

Katrina John, Jeffrey Kile, and Amish Aghera

101.1 Indications

- Any dislocation of the elbow joint. Direction of the dislocation (i.e., anterior, posterior, lateral and divergent radius, and ulnar dislocations) is determined by the position of the ulna relative to the joint space (Fig. 101.1).

101.2 Contraindications

- Relative
  - Compound fracture dislocation

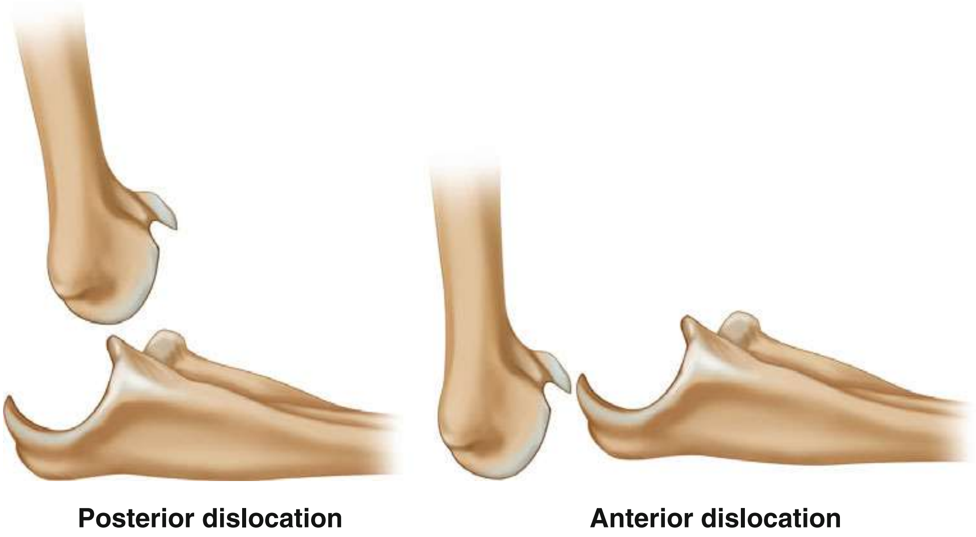

Fig. 101.1 Anatomical depiction

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_101

101.3 Materials and Medications

- Parenteral sedation and analgesia medications
- Local anesthetic for local and intra-articular anesthesia
- Splinting material
- Stockinette
- Padding
- Elastic bandage
- Tape
- Sling

101.4 Procedure for Posterior Dislocations

1. Obtain a true lateral and anteroposterior radiographs of the affected elbow.
2. Ensure adequate sedation and analgesia.
3. Consider intra-articular analgesia.
4. Check the neurovascular status of affected extremity.
5. Follow a selected method for reduction as detailed later.
6. Following successful reduction gently flex the elbow to ensure full range of motion.
7. Place a long-arm posterior splint with the elbow in at least 90° flexion and secure the arm in a regular sling.
8. Check neurovascular status.
9. Obtain a postreduction radiograph of the elbow.

101.4.1 **Method A** (Fig. 101.2)

1. Position the patient on a stretcher in the supine position.
2. Apply steady traction at the supinated distal forearm keeping the elbow slightly flexed, while an assistant applies countertraction to the midhumerus with both hands.

101.4.2 **Method B** (Fig. 101.3)

1. Position the patient on a stretcher in the supine position.
2. Extend the affected extremity over the edge of the stretcher.
3. Apply traction to the supinated forearm slightly flexed at the elbow, while an assistant holds the distal humerus with both hands and uses thumbs to apply pressure to the olecranon as if pushing it away from the humerus.

101.4.3 **Method C** (Fig. 101.4)

1. Position the patient on a stretcher in the prone position.
2. Hang the affected extremity over the side of the stretcher toward the floor.
3. Apply downward traction to the pronated distal forearm and with the other hand just above the patient's antecubital fossa lift the humerus toward you.

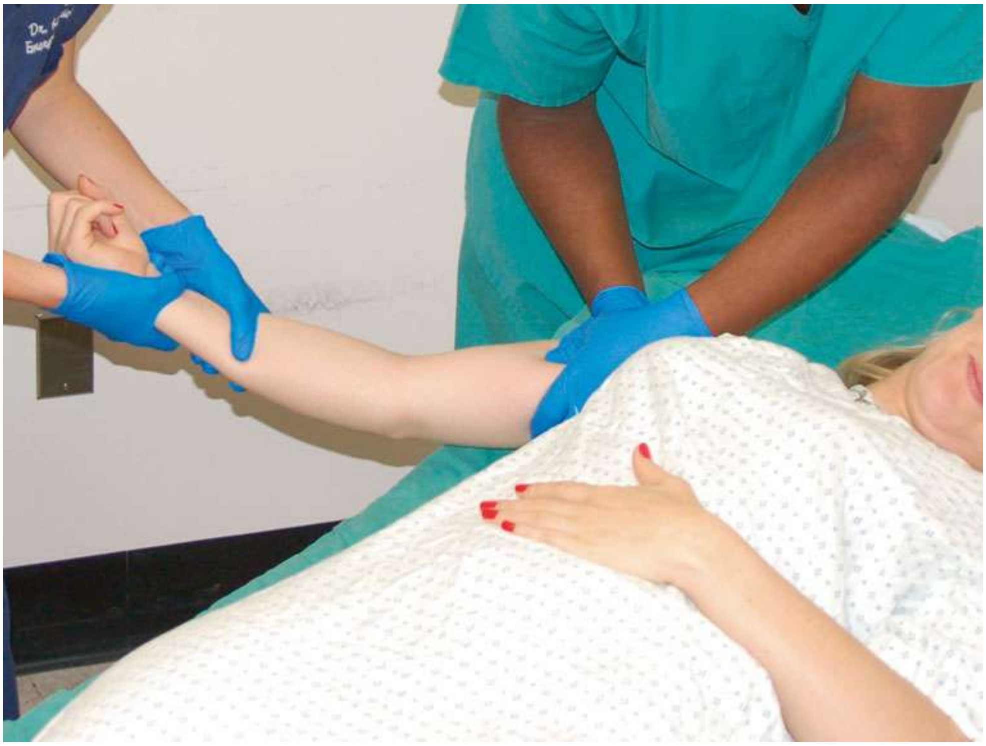

Fig. 101.2 Posterior method A

Fig. 101.3 Posterior method B

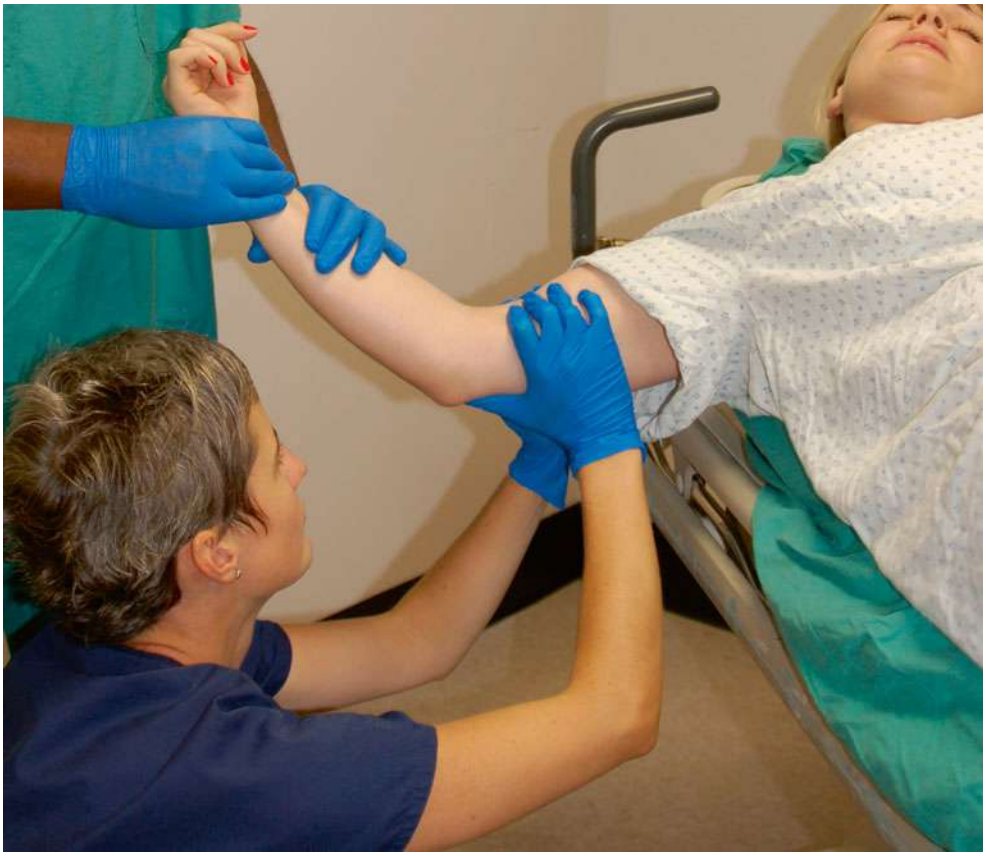

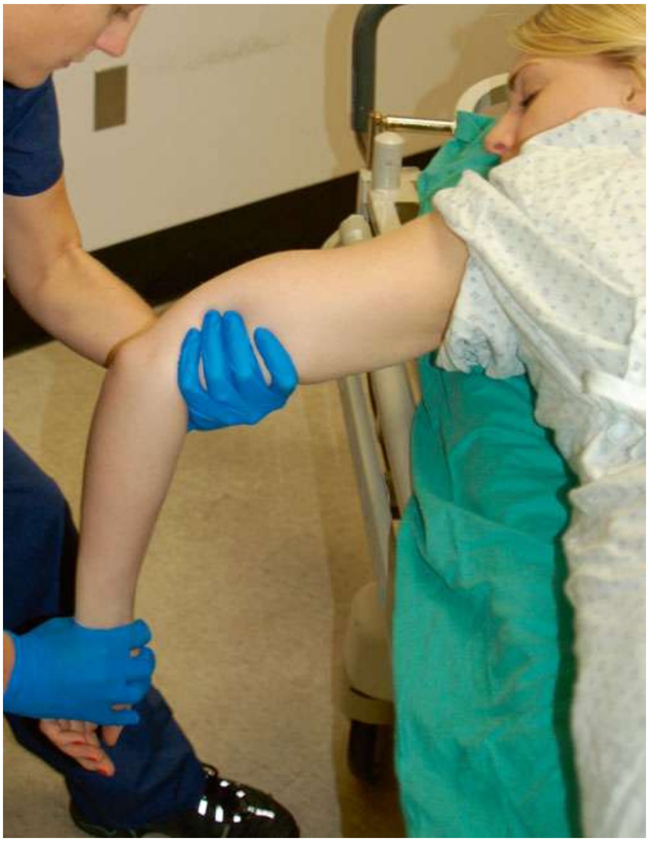

Fig. 101.4 Posterior method C

101.5 Procedure for Anterior Dislocations (Fig. 101.5)

1. Follow pre- and postprocedure steps as documented for the posterior dislocation.
2. Position the patient on a stretcher in the supine position.
3. With one hand, apply traction to the supinated distal forearm with the elbow extended, while an assistant applies countertraction with both hands around the distal humerus.
4. With the other hand apply downward and backward pressure over the proximal forearm just below the antecubital fossa.

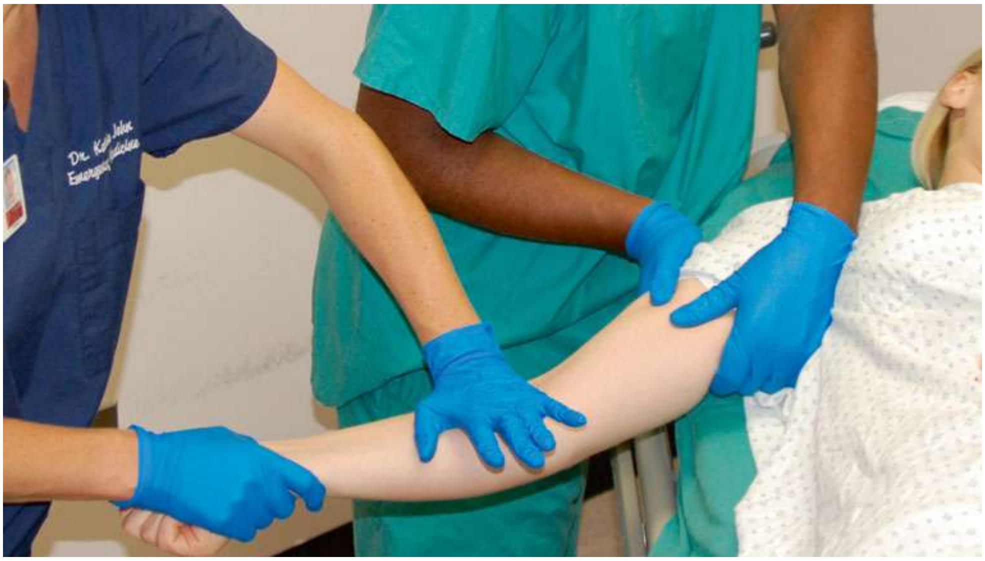

Fig. 101.5 Anterior elbow

101.6 Procedure for Radial Head Subluxations (See Also Chap. 127)

1. This procedure can normally be performed without any sedation or parenteral analgesia.
2. Position the patient, most commonly a child aged 1–3 years, facing forward on the caretaker's lap.
3. Hold the flexed elbow of the affected extremity placing your thumb firmly over the radial head.
4. With the other hand, take the child's hand and wrist, and in one continuous movement, hyperpronate and flex the forearm (Figs. 101.6 and 101.7).
5. Another method is to supinate and flex the forearm instead of hyperpronating it (Fig. 101.8).
6. Leave the room, encourage the caretaker to engage the child with distracting activities and reexamine the child in 10–20 min, at which stage, if reduction was successful, the child should be using the extremity normally again.
7. No postreduction radiograph or immobilization is required.

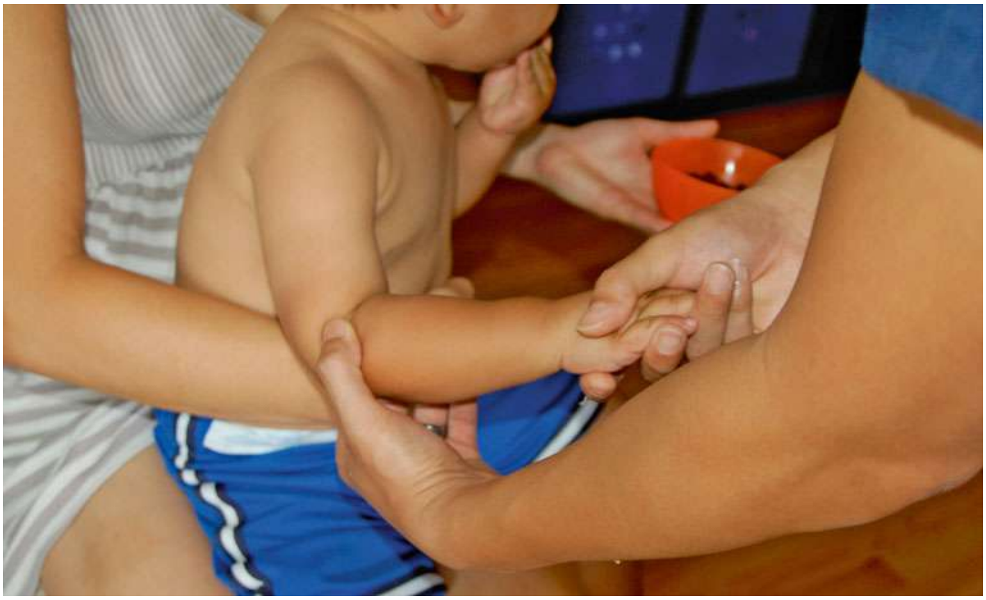

Fig. 101.6 Subluxation hyperpronated

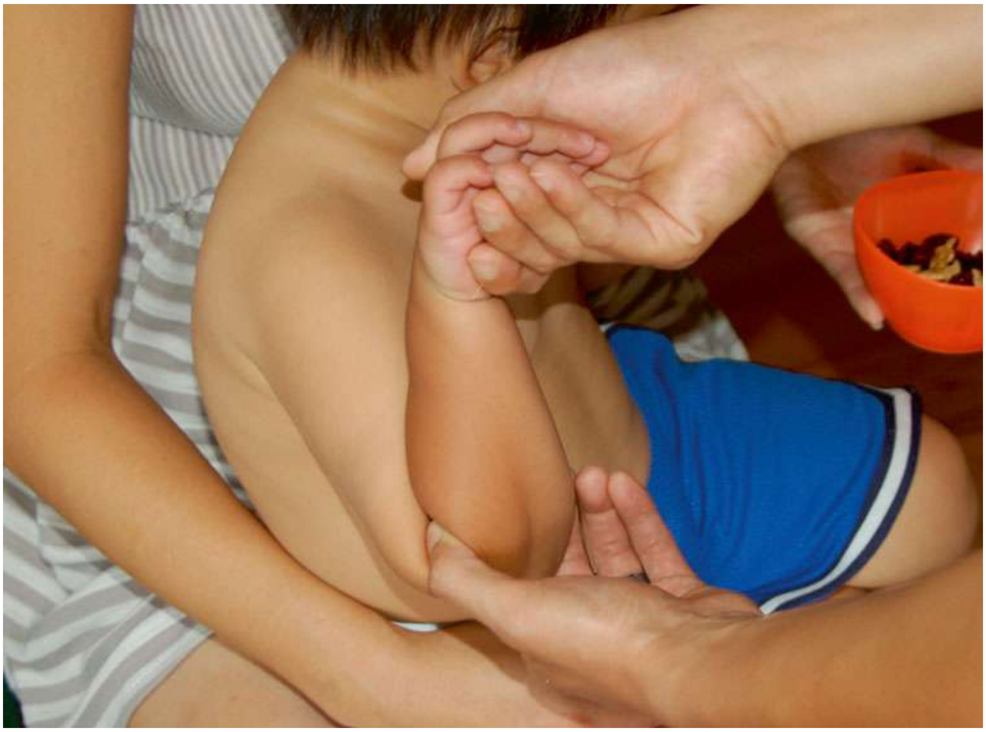

Fig. 101.7 Subluxation hyperpronated and flexed

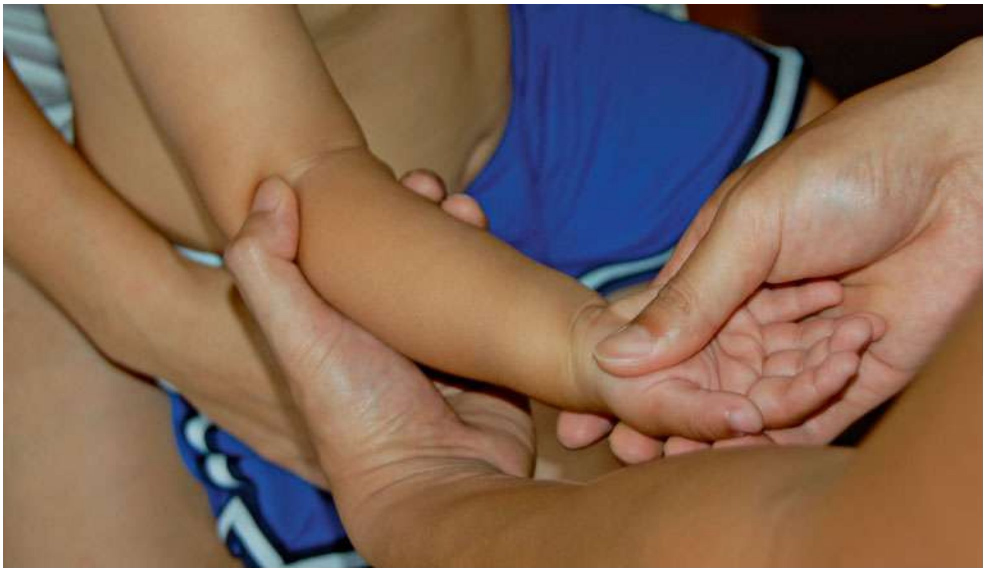

Fig. 101.8 Subluxation supinated

101.7 Complications

- Concomitant fractures
- Vascular injury, most commonly to the brachial artery
- Median nerve injury/entrapment
- Recurrent dislocation—rare

101.8 Pearls and Pitfalls

- Pearls
  - A true lateral radiograph is necessary to accurately detect and identify elbow fractures, dislocations, and soft tissue abnormalities (i.e., the fat pad sign). It is obtained with the patient's elbow in 90° flexion, in neutral rotation with the thumb pointing up and the arm and forearm resting on the radiograph cassette and the beam nearly perpendicular to the cassette. On a true lateral, the "hourglass" or "figure-of-eight" formation at the distal humerus should be clearly visible, and the rings of the capitellum and trochlea should be concentric.
  - During nursemaid elbow reduction, provide age appropriate distractions to divert the child's attention and minimize resistance.
- Pitfalls
  - On the pre- and postreduction radiographs, search for commonly associated fractures of the distal humerus, radial head, and coronoid process.
  - Inability to range the elbow after apparent reduction indicates possible trapped fracture fragments and the need for operative intervention.
  - Vascular or open injuries are common with anterior dislocations, and early orthopedic consultation is advised.
  - Ninety percent of simple elbow dislocations are posterior, and this injury is rarely associated with vascular injury. However, it does occur, and vascular evaluation after every reduction is good clinical practice.

Selected Reading

Jain K, Shashi Kumar Y, Mruthyunjaya, Ravishankar R, Nair AV. Posterior dislocation of elbow with brachial artery injury. J Emerg Trauma Shock. 2010;3:308.

Kuhn MA, Ross G. Acute elbow dislocations. Orthop Clin North Am. 2008;39:155–61.

McDonald J, Witelaw C, Goldsmith LJ. Radial head subluxation. Comparing two methods of reduction. Acad Emerg Med. 1999;6:715.

Sheps DM, Hildebrand KA, Boorman RS. Simple dislocations of the elbow: evaluation and treatment. Hand Clin. 2004;20:389–404.

Villarin Jr LA, Belk KE, Freid R. Emergency department evaluation and treatment of elbow and forearm injuries. Emerg Med Clin North Am. 1999;17:843–58.

Distal Interphalangeal Joint Reduction

102

Justin Chen and Muhammad Waseem

Distal interphalangeal (DIP) joint dislocation is rare. It occurs when an axial force is applied to the distal phalanx (Fig. 102.1).

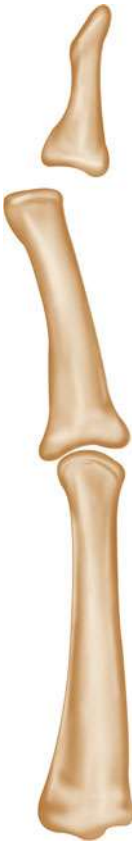

Fig. 102.1 Distal interphalangeal joint dislocation

102.1 Indications

- DIP joint reduction is performed to alleviate functional and anatomical derangements resulting from DIP joint dislocation, commonly dorsal, from axial compression.

102.2 Contraindications

- Absolute
  - Absence of radiographic confirmation (anteroposterior, true lateral, and oblique) of simple DIP joint dislocation, especially in pediatric cases
- Relative
  - Open joint dislocation, associated fracture, or entrapped volar plate
  - Digital neurovascular compromise

102.3 Materials and Medications

- Latex-free gloves
- Local anesthetic: 2 % lidocaine without epinephrine, 1.5 % mepivacaine, 0.5 % ropivacaine, or 0.5 % bupivacaine
- 25-gauge × 1.5-in. needle (can substitute with 27 or 30 gauge)
- Small plastic syringe, 10 mL
- Padded, malleable, aluminum digital splint

J. Chen, MD, MSc
Department of Emergency Medicine, North Shore University Hospital, Manhasset, NY, USA
e-mail: chen.justin@gmail.com

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_102

102.4 Procedure

1. Place the patient in the seated position with the arms at rest on a bedside table or supported by an assistant.
2. Pronate the patient's hand, remove rings if present, and rest on a flat surface.
3. Insert a 25-gauge needle at the dorsolateral aspect of the base of the finger to form a wheal to reduce patient discomfort.
4. Advance the needle and direct anteriorly toward the phalangeal base.
5. Inject 0.5–1 mL of local anesthetic as the needle is withdrawn 1–2 mm from the point of bone contact.
6. Inject an additional 1 mL of local anesthetic continuously as the needle is withdrawn.
7. The injection should never render the tissue tense nor be circumferential.
8. Hyperextend the DIP joint while applying longitudinal traction, followed by immediate joint flexion at the base of the distal phalanx.
9. Place finger(s) in an aluminum digital dorsal splint in slight flexion for 2 weeks (Fig. 102.2).
10. Postreduction radiograph is recommended for confirmation.

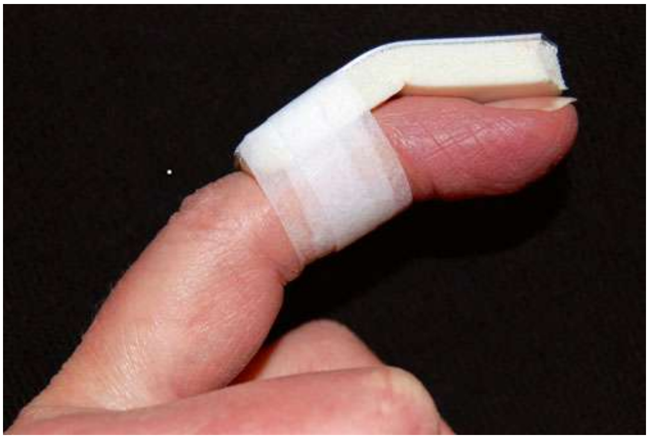

Fig. 102.2 Padded aluminum splint applied to block the DIP joint in flexion but allow further flexion, which encourages active flexion of that joint when the PIP joint flexes (Reproduced with permission from: HandLab Clinical Pearls Feb 2011, No 12. www.handlab.com)

102.5 Complications

- Irreducible dislocations
- Stiffness
- Recurrent dislocation
- Extensor lag in joints with residual subluxation
- Associated with dorsal joint prominences, swan-neck/ boutonnière deformity, and degenerative arthritis

102.6 Pearls and Pitfalls

- Pearls
  - Lidocaine without epinephrine is preferred owing to the risk of vasoconstriction of the digital vessels with epinephrine.
  - Regardless of the mechanism of trauma, all joints (DIP, proximal interphalangeal, metacarpophalangeal) should be assessed for instability.
  - Joint dislocations involving volar plate entrapment may require surgical repair (open reduction internal fixation) for successful reduction.
- Pitfalls
  - Irreducible DIP joint dislocations may be due to entrapment of an avulsion fracture, the profundus tendon, or the volar plate.

Selected Reading

Calfee RP, Sommerkamp TG. Fracture-dislocation about the finger joints [review]. J Hand Surg Am. 2009;34:1140–7.

Knoop KJ. Atlas of emergency medicine. 3rd ed. New York: McGraw-Hill Professional; 2010.

Simon RR, Sherman SC, Sharieff GQ. Emergency orthopedics. 6th ed. New York: McGraw-Hill Medical; 2011.

Stone CK, Humphries RL. Current diagnosis & treatment emergency medicine. 6th ed. New York: McGraw-Hill; 2008.

Tintinalli JE, Stapczynski JS, Ma OJ, Cline D, Cydulka R, Meckler G, editors. Tintinalli's emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw-Hill; 2012.

Hip Dislocation Reduction

103

Katrina John, Jeffrey Kile, and Amish Aghera

103.1 Indications

Displacement of the femoral head in relation to the acetabulum without concomitant femoral neck, head, or acetabulum fractures:

- Posterior hip dislocations make up 80–90 % of cases.
- Anterior hip dislocations make up 10–15 % of cases. These are classified into obturator, pubic, iliac, central, or inferior types. Central dislocations are associated with comminuted acetabulum fractures, and inferior dislocations are a rare occurrence normally occurring in children younger than 7 years of age.
- Prosthetic hip dislocations

103.2 Contraindications

- Absolute
  - Femoral neck fracture: attempted reduction may increase the displacement of the fracture and increase the probability of avascular necrosis.
- Relative
  - Fractures in other parts of the affected lower extremity: these may limit the pressure that can be applied necessary for traction during reduction.

103.3 Materials and Medications

- Parenteral sedation and analgesia medications
- Sheet or belt to fix the pelvis to the stretcher
- Knee immobilizer
- Abduction pillow

103.4 Procedure

1. Check the neurovascular status of the affected extremity.
2. Obtain anteroposterior (AP) views of the pelvis and lateral views of the hip.
3. Ensure adequate parenteral sedation and analgesia.
4. Decide upon a technique, as detailed later, and position the patient accordingly.
5. Once the hip has been successfully reduced, test the joint for stability by moving it gently thought its range of motion.
6. Place a knee immobilizer and an abduction pillow between the knees.
7. Check the neurovascular status.
8. Obtain repeat AP films of the pelvis.

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_103

103.4.1 Stimson Maneuver

1. Place the patient prone on the stretcher with the affected extremity hanging over the edge and the hip flexed to 90°.
2. Flex the knee and the foot to 90°.
3. Apply downward pressure to the area just distal to the popliteal fossa with a hand (Fig. 103.1) or knee (Fig. 103.2) while using the opposite hand to internally and externally rotate the hip at the ankle.
4. Have an assistant simultaneously manipulate the displaced femoral head into position with both hands, applying downward pressure over the affected buttock (Fig. 103.3).

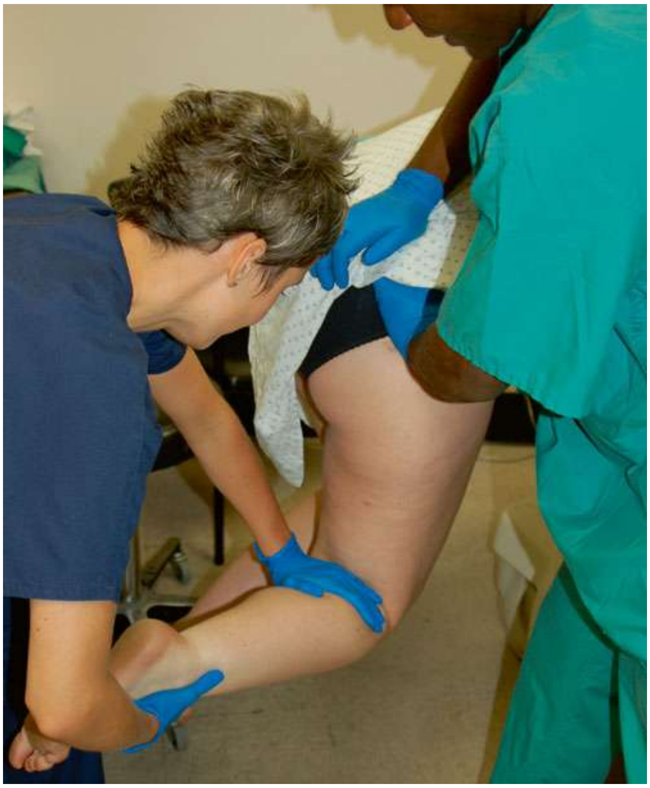

Fig. 103.1 Stimson maneuver with hand

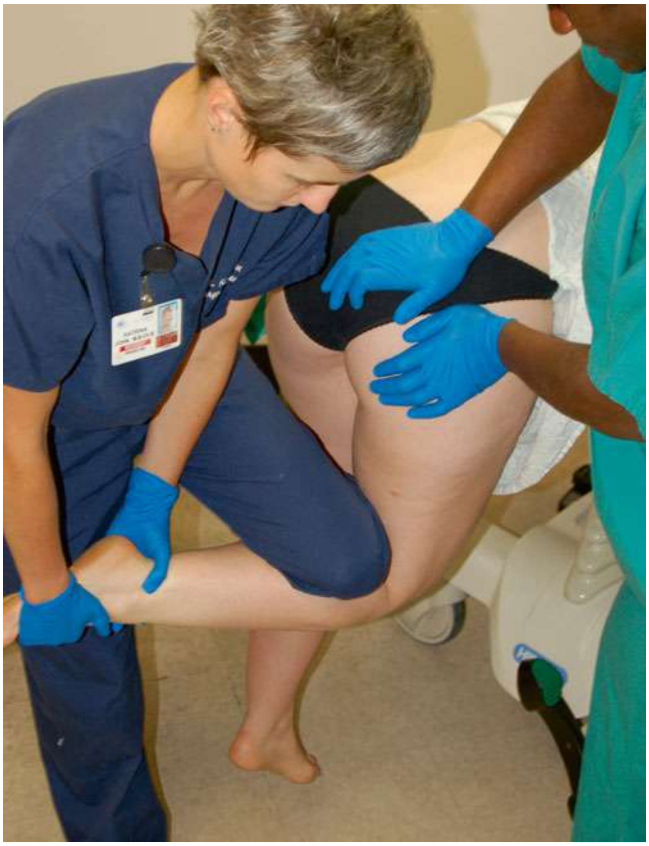

Fig. 103.2 Stimson maneuver with knee

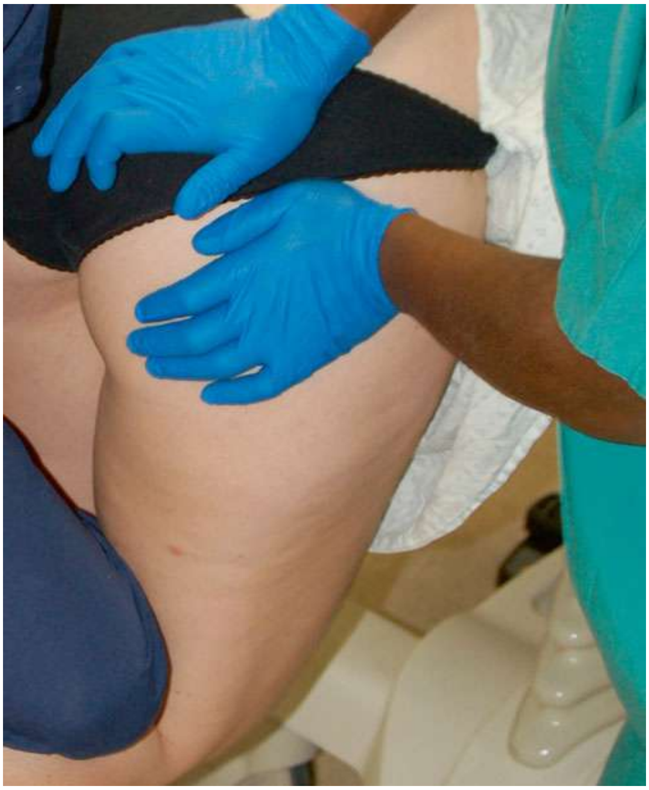

Fig. 103.3 Manipulation of the femoral head

103.4.2 Allis Maneuver

1. Position the patient supine on the stretcher.
2. The operator should stand on the stretcher to achieve maximum leverage or have the patient on a backboard on the ground.
3. Have an assistant apply downward pressure to both iliac crests.
4. Apply constant, gentle upward traction in line with the deformity while maneuvering the hip to 90° flexion and through internal and external rotation (Fig. 103.4).
5. Have a second assistant provide lateral traction to the midthigh.
6. Once the femoral head has cleared the outer lip of the acetabulum, continue traction while keeping the hip in external rotation and gently abducting and extending the hip (Fig. 103.5).

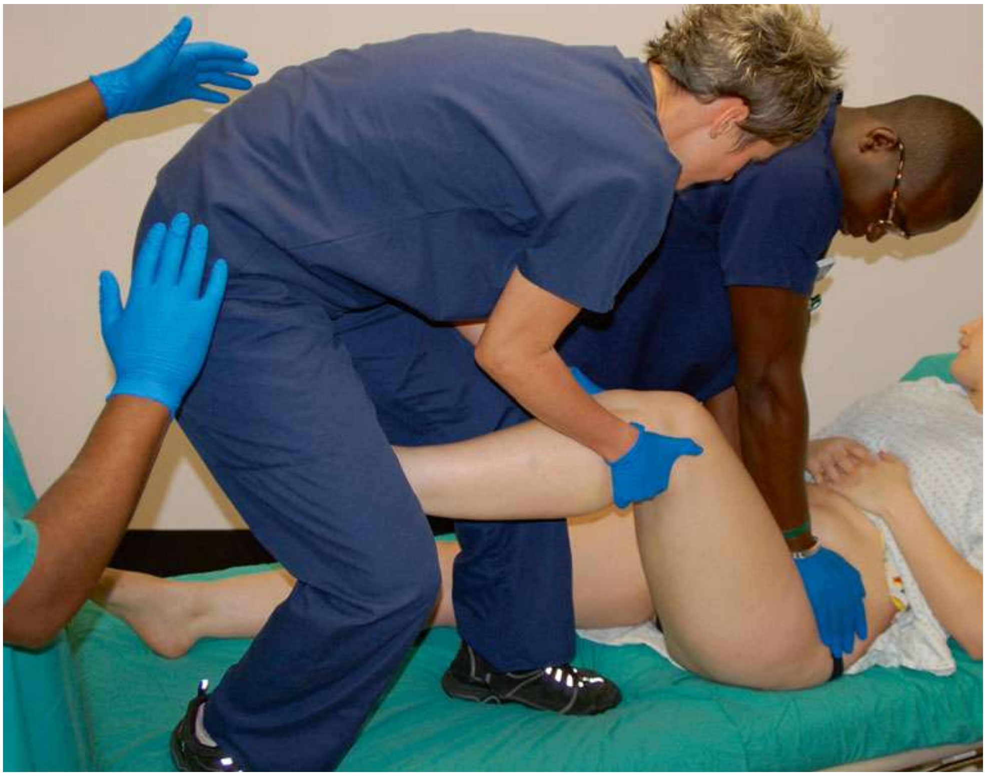

Fig. 103.4 Allis flexion

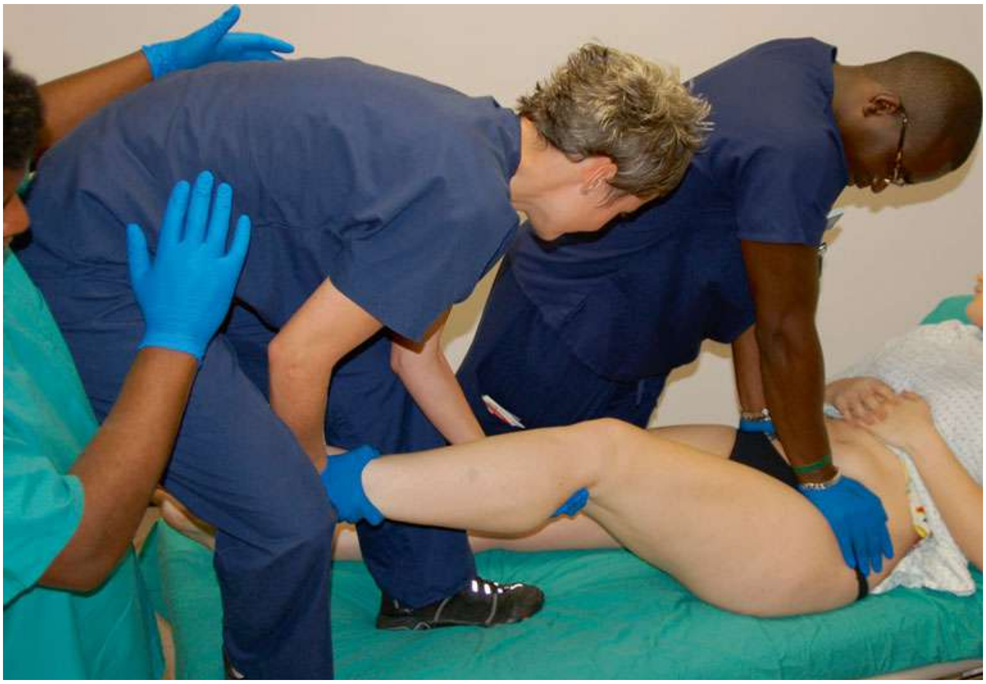

Fig. 103.5 Allis extension

103.4.3 Whistler Technique

1. Position the patient supine on the stretcher with the knee and hip flexed to 45°.
2. Have an assistant stabilize the pelvis with downward pressure on both iliac crests.
3. Stand on the side of the affected extremity and place one arm under the knee, resting the hand on the flexed knee of the unaffected extremity.
4. Secure the ankle of the affected extremity with the other hand and elevate the shoulder of the opposite arm, providing upward traction at the distal thigh and a strong fulcrum to reduce the dislocation (Fig. 103.6).
5. Internal and external rotation can be achieved with the opposite hand at the ipsilateral ankle.

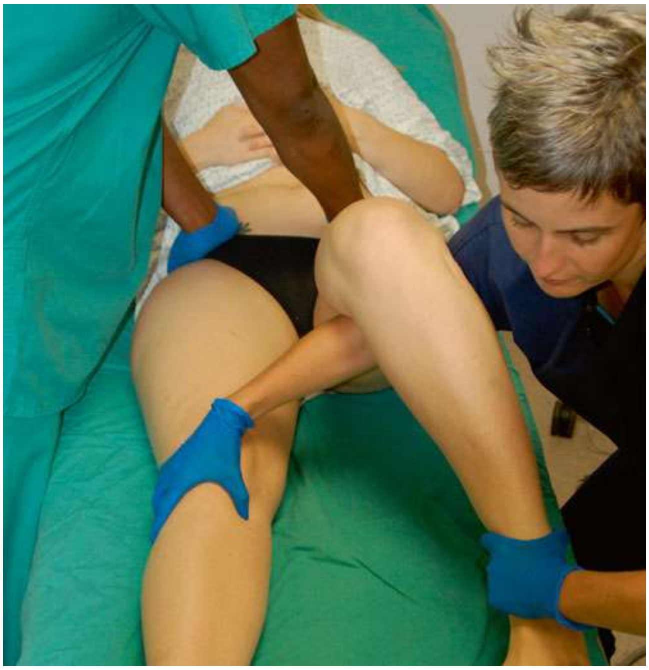

Fig. 103.6 Whistler technique

103.4.4 Captain Morgan Technique

1. Position the patient supine on the stretcher with the knee and hip flexed to 90°.
2. Stabilize and fix the pelvis with a sheet tied securely over the pelvis and under the stretcher.
3. Standing on the side of the affected extremity, the operator's foot should be resting perpendicular on the stretcher with the knee placed under the patient's knee.
4. With the opposite hand, apply downward pressure to the ankle and provide a sustained upward force to the patient's thigh by elevation of the knee through plantar flexion of the toes and upward pressure of the other hand placed behind the patient's knee.
5. Internal and external rotation can be applied simultaneously if necessary by gently twisting the ankle (Fig. 103.7).

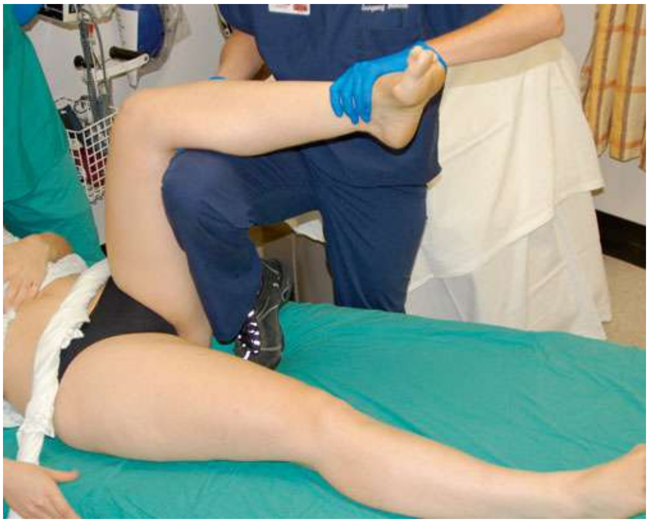

Fig. 103.7 Captain Morgan technique

103.5 Complications

- Sciatic nerve injury
- Avascular necrosis of the femoral head due to delay in adequate reduction
- Inability to perform reduction due to occult fractures and fracture fragments, incarceration of the joint capsule, or associated tendons
- Unstable or irreducible dislocations
- Traumatic arthritis and joint instability

103.6 Pearls and Pitfalls

- Pearls
  - On AP radiograph, posterior dislocations can be more easily detected by the presence of a smaller femoral head compared with the unaffected side and poor visualization of the lesser trochanter.
  - On AP radiographs, anterior dislocations can be detected by a larger femoral head and a clear lesser trochanter seen in profile alongside the femoral shaft.
  - Pay close attention to the femoral vessels and the sciatic nerve. Injury to the sciatic nerve most commonly affects the common peroneal branch, therefore causing weakness in great toe extension and foot dorsiflexion. Sensation may also be reduced over the dorsum of the foot.
  - Check the femoral head is intact and clearly in the acetabulum and for intact Shenton lines, symmetrical intra-articular spaces, and clear outlines of the lesser trochanters.
  - For any of the techniques requiring stabilization of the pelvis, an alternative is to fix it to the stretcher using a sheet or belt.
  - To overcome the powerful muscles that oppose successful reduction, it is important to provide adequate muscle relaxation and steady, prolonged traction.
  - An assistant should stand on the floor behind to support the operator if standing on a stretcher.
- Pitfalls
  - Owing to the force necessary to dislocate a native hip, this injury should serve as a red flag to the physician to consider other potentially life- or limb-threatening occult injuries.
  - Hip dislocation is a true orthopedic emergency and must be treated without delay. Delay in reduction, especially greater than 6 h, results in increased incidence of avascular necrosis of the femoral head and sciatic nerve injury.
  - Review imaging carefully because associated fractures of the femoral head, neck, and acetabulum are often present.
  - It is recommended that anterior dislocations be reduced by orthopedic surgeons under general anesthetic in the operating room. These are often more complicated and difficult to reduce, and failure at closed reduction in the operating room can be followed by an open procedure.
  - Multiple attempts at reduction should not be performed in the emergency department because these are unlikely to be successful and will only delay definitive management and lead to an increased risk of complications.

Selected Reading

Hendey GW, Avila A. The captain Morgan technique for the reduction of the dislocated hip. Ann Emerg Med. 2011;58:536–40.

Newton EJ, Love J. Emergency department management of selected orthopedic injuries. Emerg Med Clin North Am. 2007; 25:763–93.

Nordt WE. Maneuvers for reducing dislocated hips. Clin Orthop Relat Res. 1999;360:160–4.

Rupp JD, Schneider LW. Injuries to the hip joint in frontal motor-vehicle crashes: biomechanical and real-world perspectives. Orthop Clin North Am. 2004;35:493–504.

Walden PD, Hamer JR. Whistler technique used to reduce traumatic dislocation of the hip in the emergency department setting. J Emerg Med. 1999;17:441–4.

104 Knee Dislocation Reduction

Katrina John, Jeffrey Kile, and Amish Aghera

104.1 Indications

- Dislocation of the knee/fibular head/patella

104.2 Contraindications

- Absolute
  - None
- Relative
  - Immediate availability of orthopedic consultation

104.3 Materials and Medications

- Parenteral sedation and analgesia medications
- Knee immobilizer or splinting materials

104.4 Procedure

104.4.1 Knee (Femur/Tibia) Dislocation Reduction

1. Assess neurovascular function.
2. Pretreat the patient with sedation or analgesia as appropriate.
3. Position the patient supine with the affected leg fully extended.
4. Instruct an assistant to stand near the patient's hip and, facing the patient's affected knee, grasp the distal femur firmly with both hands to fix it in place.
5. Stand near the patient's foot and, facing the patient's affected knee, grasp the distal tibia and apply straight traction in a distal direction.
   - Longitudinal traction-countertraction alone, as described previously, will usually reduce the dislocation. If reduction does not occur, proceed with the following steps.
6. While applying straight traction in a distal direction to the tibia with the dominant hand, with the nondominant hand:
   (a) Anterior dislocation: push the proximal tibia in a posterior direction (Fig. 104.1)
   (b) Posterior dislocation: lift the proximal tibia in an anterior direction (Fig. 104.2)
   (c) Lateral dislocation: push the proximal tibia in a medial direction (Fig. 104.3)
   (d) Medial dislocation: push the proximal tibia in a lateral direction (Fig. 104.4)
   (e) Rotary dislocation: rotate the proximal tibia into proper linear alignment with the femoral condyles (Fig. 104.5)
      - Reduction may be facilitated by the use of two assistants rather than just one. The second assistant grasps the distal tibia and applies straight traction in a distal direction, freeing the operator to manipulate the proximal tibia as described previously using both hands.
7. After reduction, reassess neurovascular function and, if available, obtain angiography.
8. Immobilize the knee in 15° of flexion in a knee immobilizer or long-leg posterior splint.

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_104

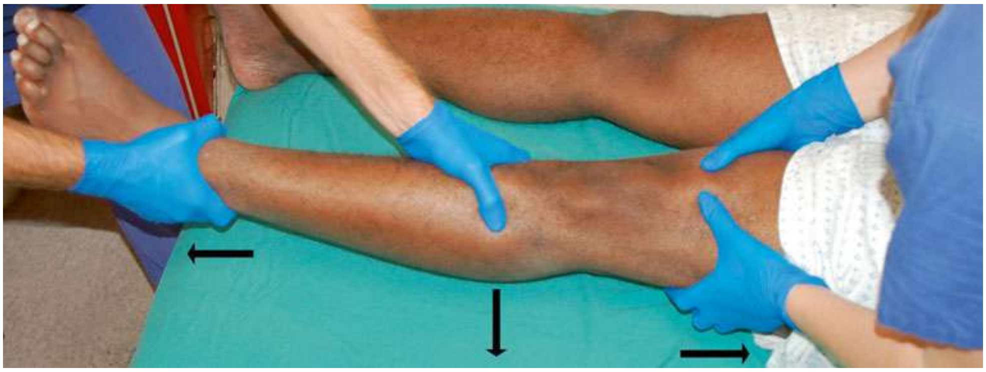

Fig. 104.1 Anterior dislocation of the knee: the proximal tibia is pushed in a posterior direction. The *arrows* indicate the direction in which force should be applied by the operator during reduction of dislocation

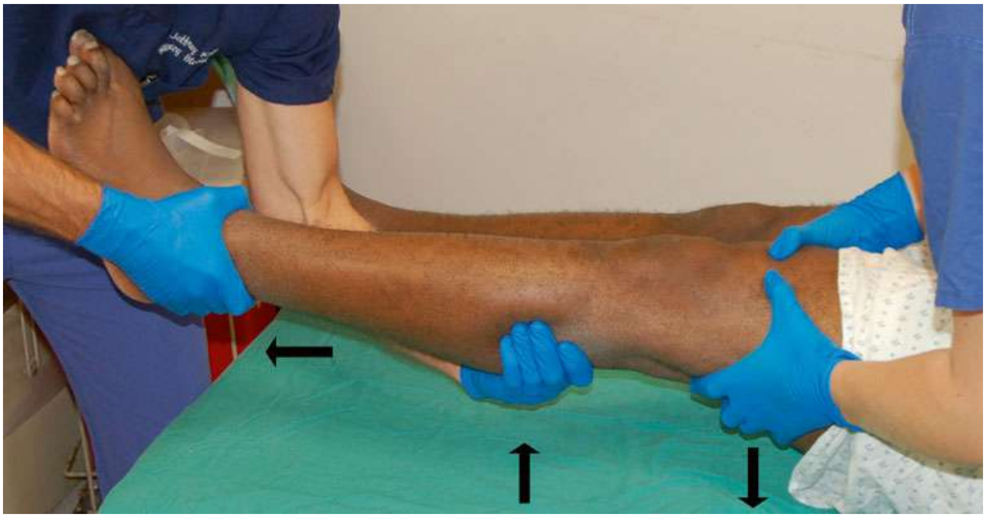

Fig. 104.2 Posterior dislocation of the knee: the proximal tibia is pushed in an anterior direction. The *arrows* indicate the direction in which force should be applied by the operator during reduction of dislocation

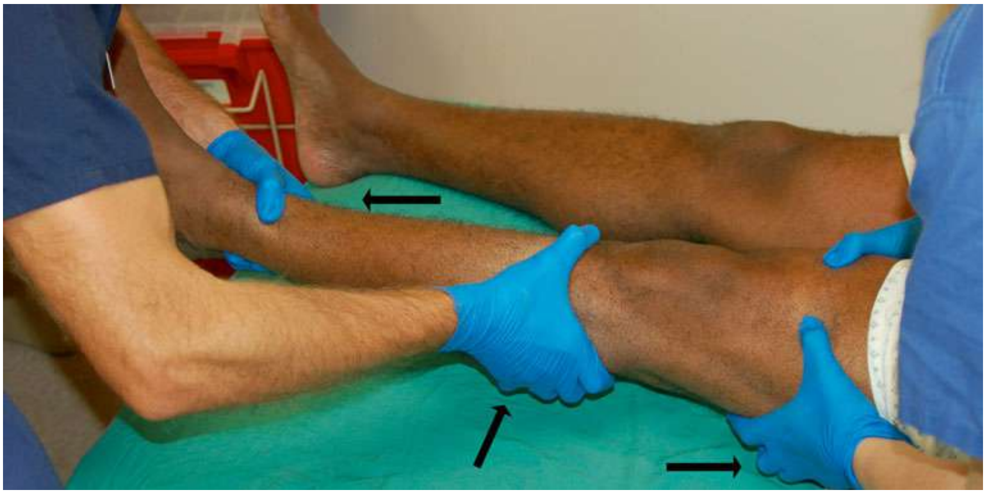

Fig. 104.3 Lateral dislocation of the knee: the proximal tibia is pushed in a medial direction. The *arrows* indicate the direction in which force should be applied by the operator during reduction of dislocation

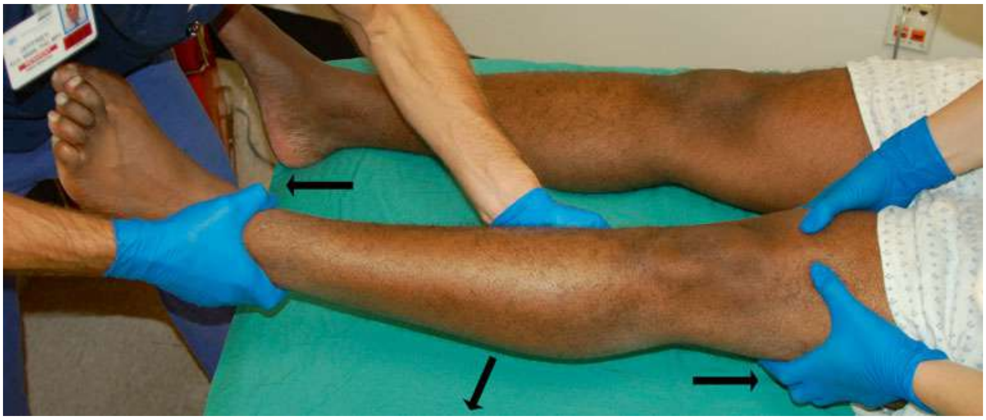

Fig. 104.4 Medial dislocation of the knee: the proximal tibia is pushed in a lateral direction. The *arrows* indicate the direction in which force should be applied by the operator during reduvction of dislocation

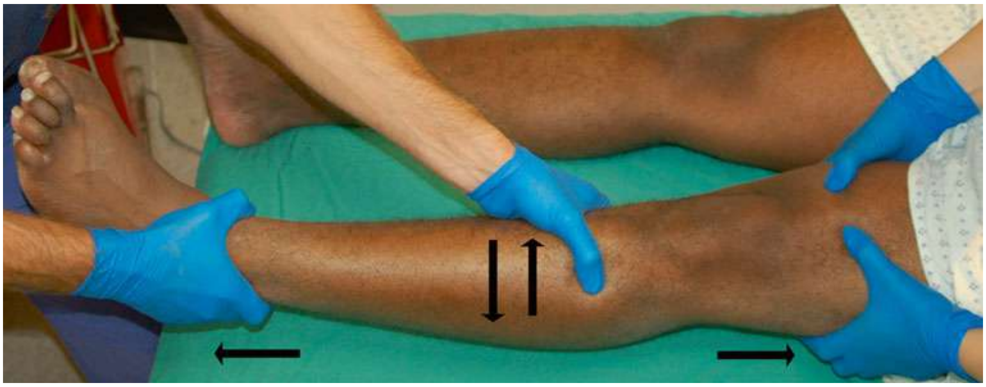

Fig. 104.5 Rotary dislocation of the knee: the proximal tibia is rotated into proper alignment with the femoral condyles. The *arrows* indicate the direction in which force should be applied by the operator during reduction of dislocation

104.4.2 Fibular Head Dislocation Reduction

1. Assess neurovascular function.
2. Pretreat the patient with sedation or analgesia as appropriate.
3. Position the patient supine.
4. Flex the knee to 90° to relax the biceps femoris tendon.
5. Instruct an assistant to stand near the patient's hip and, facing the patient's affected knee, grasp the distal femur firmly with both hands to fix it in place.
6. Stand near the patient's foot and, facing the patient's affected knee, grasp the distal tibia and apply straight traction in a distal direction with the dominant hand and with the nondominant hand.
   (a) Anterior dislocation: push the fibular head in a posterior direction (Fig. 104.6)
   (b) Posterior dislocation: push the fibular head in an anterior direction (Fig. 104.7)
      - Reduction may be facilitated by the use of two assistants rather than just one. If a second assistant is available, instruct the second assistant to stand near the patient's foot and, facing the patient's affected knee, grasp the distal tibia and apply straight traction in a distal direction. This enables the operator to grasp and move the proximal fibula as described previously using both hands.
      - Reduction is often signified by a palpable and audible click as the fibula snaps back into position.
7. After reduction, reassess neurovascular function and, if available, obtain angiography.
   - After reduction, patients should receive orthopedic referral, avoid weight-bearing for the first 2 weeks, and then gradually increase weight-bearing over the next 6 weeks.
   - Typically, immobilization is not required following reduction of an i.solated fibular head dislocation.

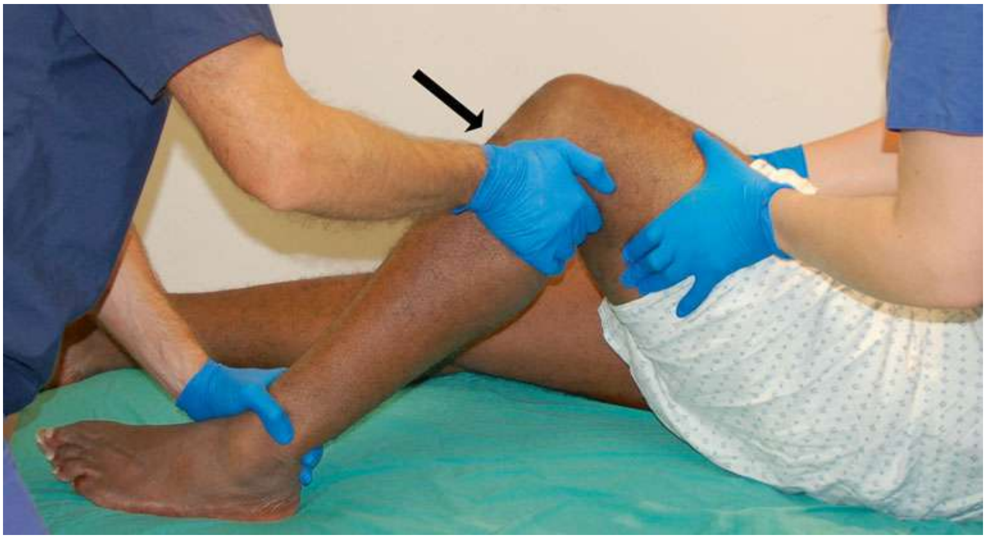

Fig. 104.6 Anterior dislocation of the fibular head: the fibular head is pushed in a posterior direction. The *arrow* indicate the direction in which force should be applied by the operator during reduction of dislocation

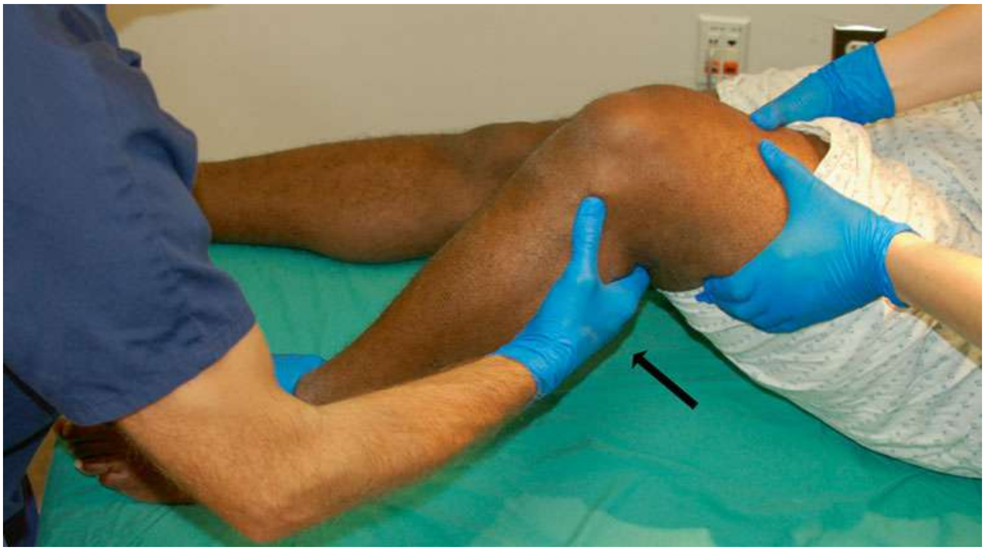

Fig. 104.7 Posterior dislocation of the fibular head: the fibular head is pushed in an anterior direction. The *arrow* indicate the direction in which force should be applied by the operator during reduction of dislocation

104.4.3 Lateral Patellar Dislocation Reduction

1. Pretreat the patient with sedation or analgesia as appropriate.
2. Stand at the side of the affected knee and, facing the knee, grasp the distal tibia and slowly extend the knee with one hand, and with the other hand simultaneously apply gentle pressure to the patella in a medial direction.

- The lateral edge of the patella may be lifted slightly to facilitate its travel over the femoral condyle during reduction (Fig. 104.8).
- After reduction, the knee should be immobilized in full extension in a knee immobilizer or long-leg posterior splint, and the patient should receive orthopedic referral, avoid weight-bearing for the first 2 weeks, and then gradually increase weight-bearing over the next 6 weeks.

Fig. 104.8 Lateral dislocation of the patella: the patella is pushed in a medial direction. The *arrow* indicate the direction in which force should be applied by the operator during reduction of dislocation

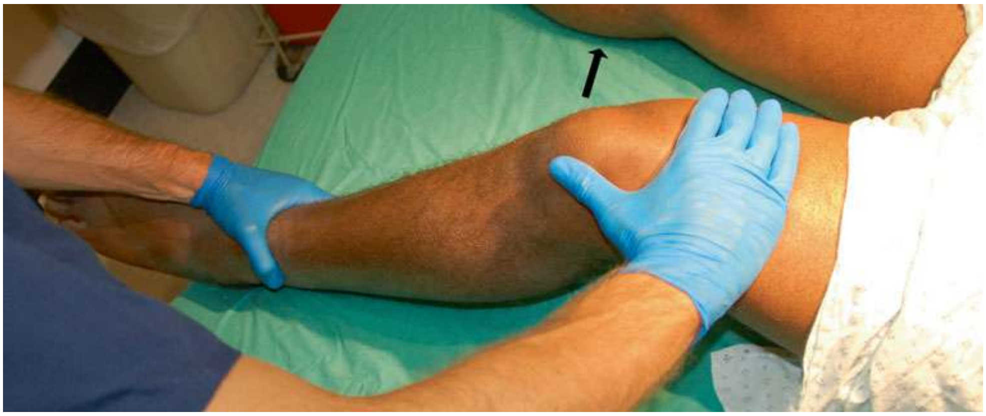

104.5 Complications

104.5.1 Knee (Femur/Tibia) Dislocations

- Distal ischemia (even requiring amputation)
- Degenerative arthritis
- Joint instability due to ligamentous injury

104.5.2 Fibular Head Dislocations

- Peroneal nerve injury
- Fibular head instability/subluxation
- Degenerative arthritis

104.5.3 Patellar Dislocations

- Failure of reduction
- Degenerative arthritis
- Recurrent dislocation/subluxation

104.6 Pearls and Pitfalls

104.6.1 Knee (Femur/Tibia) Dislocations

- Pearls
  - Dislocations of the knee are described in terms of the tibia's position in relation to the femur.
  - All knee dislocations require orthopedic evaluation at the earliest possible opportunity.
  - Owing to the frequency of associated popliteal artery and peroneal nerve injury, a neurovascular examination should be performed before and after any attempts at reduction or manipulation of the knee.
  - Dislocations of the knee should be reduced as soon as possible, particularly if distal neurovascular compromise exists.
  - Operative ligamentous repair is often required approximately 2 weeks postreduction (once acute swelling has resolved) to achieve the maximum functional recovery.
- Pitfalls
  - If the knee hyperextends more than 30° when the horizontal leg is lifted by the foot, the knee is considered severely unstable. This is likely due to a previous dislocation, and thus, the knee should be evaluated for the neurovascular complications of dislocation.
  - Because the joint capsule is commonly disrupted during knee dislocation, synovial fluid may diffuse into the surrounding tissue, such that an effusion is not always present.
  - A posterolateral dislocation may be irreducible because the medial femoral condyle traps the medial capsule within the joint.

104.6.2 Fibular Head Dislocations

- Pearls
  - Fibular head dislocations are usually anterolateral, but these do not result in neurovascular compromise.
  - A knee joint effusion is usually not seen in a fibular head dislocation because the tibiofibular ligaments are contained within a separate synovium.
  - Anterior dislocations typically result from a fall on the flexed, adducted leg, often combined with ankle inversion.
  - Flexion of the knee relaxes the fibular collateral ligament, reducing the stability of the tibiofibular joint.
  - Superior dislocation is accompanied by interosseus membrane damage and proximal displacement of the lateral malleolus.
- Pitfalls
  - Posterior fibular head dislocations usually result from direct trauma to the flexed knee and may be accompanied by peroneal nerve injury.

104.6.3 Patellar Dislocations

- Pearls
  - Patellar dislocation occurs most frequently among adolescents.
  - Patellar dislocation typically occurs in the setting of external rotation combined with a strong valgus force and quadriceps contraction.
  - Patellar dislocations are described in terms of the patellar relationship to the normal knee joint.
  - The most common patellar dislocations are lateral.
  - If a spontaneous reduction has occurred, a knee effusion and tenderness along the medial aspect of the patella are likely to be present on examination, and the patellar apprehension test will be positive.
  - To perform the patellar apprehension test, flex the knee to 30° and push the patella laterally. If the patient senses an impending redislocation, the test is considered positive.

- Isolated lateral patellar dislocations do not usually require hospitalization, but orthopedic follow-up is recommended owing to the likelihood of persistent instability.
  - Intracondylar and superior dislocations require surgical reduction.
  - Patients with an isolated patellar dislocation typically present with the knee in 20–30° of flexion and the patella displaced laterally.
- Pitfalls
  - Dislocations tend to be recurrent, particularly in patients with patellofemoral anatomical abnormalities.

Selected Reading

Martinez D, Sweatman K, Thompson EC. Popliteal artery injury associated with knee dislocations. Am Surg. 2001;67:165–7.

Peskun CJ, Levy BA, Fanelli GC, et al. Diagnosis and management of knee dislocations. Phys Sportsmed. 2010;38:101–11.

Rihn JA, Groff YJ, Harner CD, Cha PS. The acutely dislocated knee: evaluation and management. J Am Acad Orthop Surg. 2004;12:334–46.

Roberts DM, Stallard TC. Emergency department evaluation and treatment of knee and leg injuries. Emerg Med Clin North Am. 2000;18:67–84. v–vi.

Wascher DC, Dvirnak PC, DeCoster TA. Knee dislocation: initial assessment and implications for treatment. J Orthop Trauma. 1997;11:525–9.

105

Ankle Dislocation Reduction

Katrina John, Jeffrey Kile, and Amish Aghera

105.1 Indications

- Dislocation of the ankle joint. This is defined by the articulation of the talus with the mortise that is formed by the distal tibia and fibula. Dislocations can be posterior, anterior, superior, or lateral and are classified by the position of the talus in relation to the tibial mortise.

105.2 Contraindications

- Relative
  - Open dislocations where there is no evidence of acute neurovascular compromise are better managed definitively in the operating room to avoid further contamination.

105.3 Materials and Medications

- Parenteral sedation and analgesia medications
- Local anesthetic for local and intra-articular anesthesia
- Splinting material
- Stockinette
- Padding
- Elastic bandage
- Tape
- Sheet

105.4 Procedure

1. Check the neurovascular status of the affected foot and ankle.
2. If there is no evidence of critical neurovascular compromise, obtain a lateral and an anteroposterior radiograph of the affected ankle.
3. Ensure adequate parenteral sedation and analgesia to maximize success and limit pain and suffering.
4. Position the patient on a stretcher with the knee flexed at 90° over a folded pillow or rolled-up sheet or with the lower leg and knee hanging over the edge of the stretcher.

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_105

105.4.1 Posterior Dislocations

1. Hold the heel in one hand and pull with longitudinal traction.
2. With the other hand, hold the top of the foot and gently plantarflex it downward, while an assistant provides countertraction at the back of the midcalf (Fig. 105.1).
3. Continue longitudinal traction at the heel and countertraction at the calf.
4. Dorsiflex the foot while another assistant applies downward pressure to the distal anterior leg (Fig. 105.2).
5. Examine foot for restoration of normal anatomy and for any new lacerations or defects to the skin.
6. Recheck neurovascular integrity.
7. Place the leg in a sugar-tong splint with the foot at 90°.
8. Recheck neurovascular integrity.

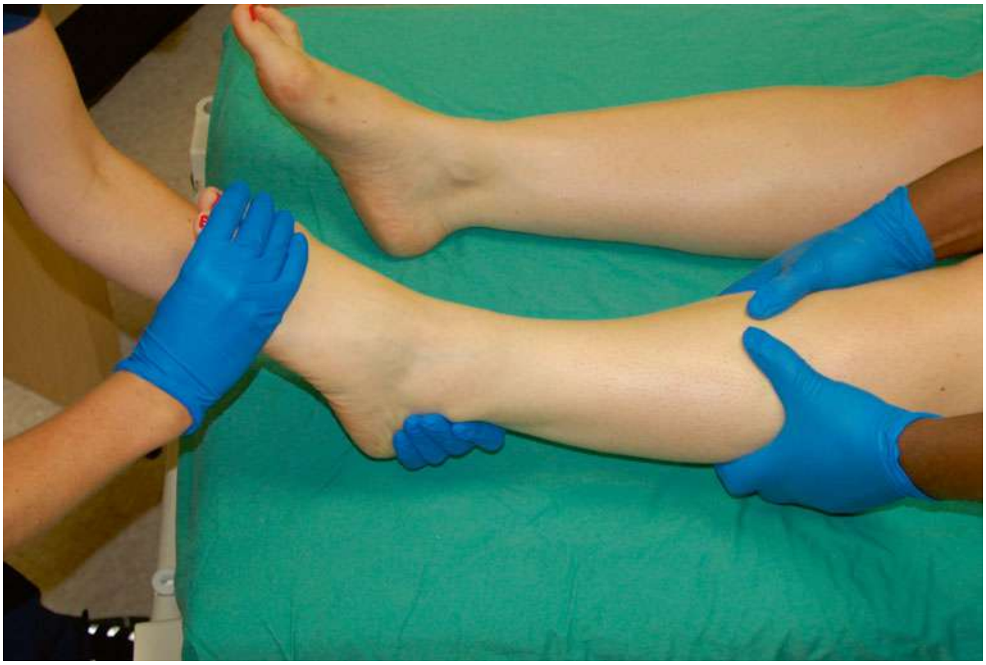

Fig. 105.1 Plantarflexion with longitudinal heel traction

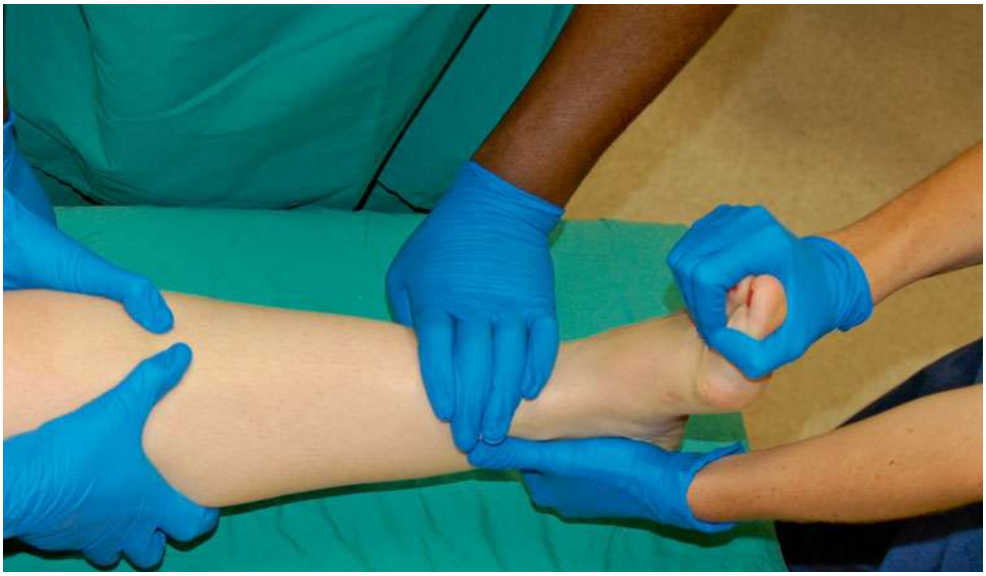

Fig. 105.2 Dorsiflexion with longitudinal heel traction

105.4.2 Anterior Dislocations

1. Hold the heel in one hand and pull with longitudinal traction.
2. With the other hand, hold the top of the foot and dorsiflex, while an assistant provides countertraction at the back of the midcalf (Fig. 105.3).
3. Continue longitudinal traction at the heel and countertraction at the calf.
4. Keeping the foot at 90° to the leg, hold the foot firmly and push the foot downward toward the floor while another assistant applies upward pressure to the distal posterior leg (Fig. 105.4).
5. Examine the foot for restoration of normal anatomy and for any new lacerations or defects to the skin.
6. Recheck neurovascular integrity.
7. Place the leg in a sugar-tong splint with the foot at 90°.
8. Recheck neurovascular integrity.

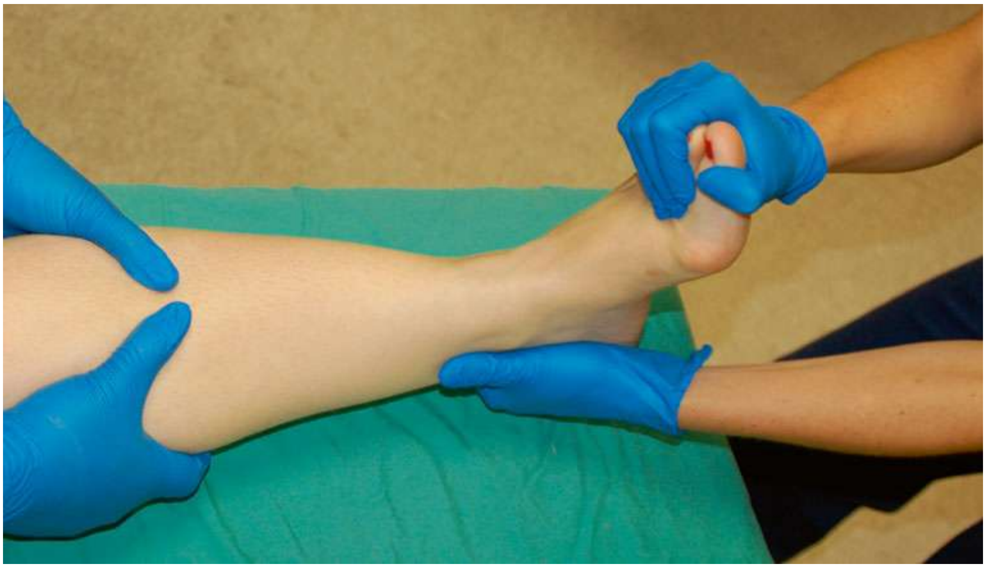

Fig. 105.3 Dorsiflexion with longitudinal heel traction

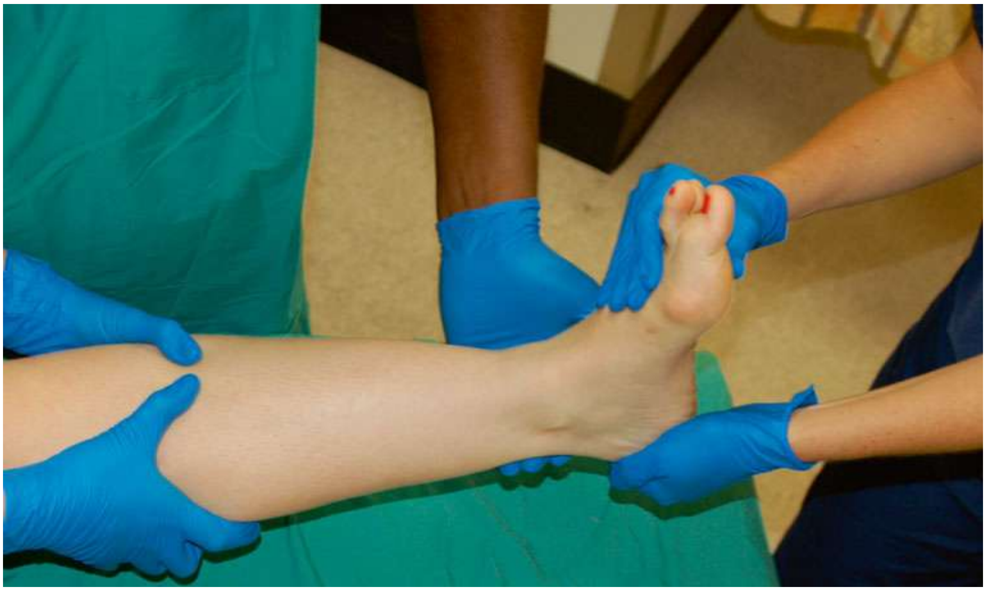

Fig. 105.4 Downward movement of foot (toward the floor) with longitudinal heel traction

105.5 Complications

- Compound fractures
- Neurovascular injury
- Skin and soft tissue damage
- Compartment syndrome

105.6 Pearls and Pitfalls

- Pearls
  - The ankle rarely dislocates without associated fractures.
- Pitfalls
  - Ankle dislocation is an orthopedic emergency, and reduction should not be delayed by imaging if there is evidence of neurovascular impairment. Complications that are exacerbated by delay in management include concomitant fractures, gross deformity of the ankle, severe stretching and tenting of the skin with resultant skin blisters, skin necrosis, and possible conversion to a compound fracture.
  - Be sure to check the radiograph carefully for commonly associated fractures notably of the malleoli.

Selected Reading

Collins DN, Temple SD. Open joint injuries: classification and treatment. Clin Orthop. 1989;243:48.

Hamilton WC. Injuries of the ankle and foot. Emerg Med Clin North Am. 1984;2:361.

Kelly PJ, Peterson FP. Compound dislocations of the ankle without fractures. Am J Surg. 1986;103:170.

Simon RR, Sherman SC, Koenigsknecht SJ, editors. Emergency orthopedics—the extremities. 5th ed. New York: McGraw-Hill; 2007. p. 264.

Wedmore IS, Charette J. Emergency department evaluation and treatment of ankle and foot injuries. Emerg Med Clin North Am. 2000;18:85.

Arthrocentesis

106

Shalu S. Patel and Bobby K. Desai

106.1 Indications

- Diagnosis of septic joint
- Diagnosis of traumatic effusion
- Diagnosis of inflammatory effusion
- Diagnosis of crystal-induced arthritis
- Therapeutic relief of pain from effusion

106.2 Contraindications

- Severe coagulopathy
- Skin infection over the needle insertion site
- Joint prosthesis
- Patients with bacteremia or sepsis (except to diagnose a septic joint)

106.3 Materials and Medications

- Betadine (povidone-iodine) or other skin antiseptic
- Sterile gloves
- Sterile towels
- Lidocaine 1 % or 2 % (5 mL) or other anesthetic of choice
- 18- to 22-gauge needle, 25-gauge needle
- Syringes (5 mL, 5–50 mL)
- Sterile gauze (4×4)
- Band-aid

106.4 Procedure

1. Informed consent may be required.
2. Position the patient appropriately. The joint should be placed in slight flexion.
3. Palpate the joint and identify anatomical landmarks.
    (a) For knee arthrocentesis, the needle should be inserted at the midpoint of either the medial or the lateral side of the patella (Fig. 106.1).
    (b) For acromioclavicular (AC) joint arthrocentesis, the needle should be inserted at the superior surface of the AC joint (Fig. 106.2).
    (c) For glenohumeral joint arthrocentesis, there are two approaches.
        (i) In the anterior approach, the needle is inserted into the groove lateral to the coracoid process (Fig. 106.3).
        (ii) In the posterior approach (preferred), the needle is inserted below the posterior border of the acromion process and lateral to the border of the scapula (Fig. 106.4).
4. Prepare the skin and drape in a sterile fashion.
5. Using lidocaine (drawn up in 5-mL syringe), anesthetize the skin with the 25-gauge needle.
6. Secure the 18- to 22-gauge needle on the 5- to 50-mL syringe (depending on the size of the joint) and insert it into the skin.
7. Advance the needle slowly into the joint space while aspirating until joint fluid can easily be withdrawn. While inserting the needle into the joint space, avoid scraping the needle against the bone.

S.S. Patel, MD
Department of Emergency Medicine, Florida Hospital Tampa,
Florida Hospital Carrollwood, Tampa, FL, USA

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health
Shands Hospita, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_106

8. If fluid cannot be aspirated easily, the catheter can be repositioned further in the joint space or turned by 45° sequentially as needed.
9. Once the joint fluid is aspirated, pull out the needle and hold pressure with gauze. Bleeding should be minimal.
10. Place a band-aid or other dressing over the site.
11. Send the synovial fluid to the laboratory. Generally, laboratory analyses may include crystals, protein, glucose, cell count and differential, culture and sensitivity, and Gram stain.

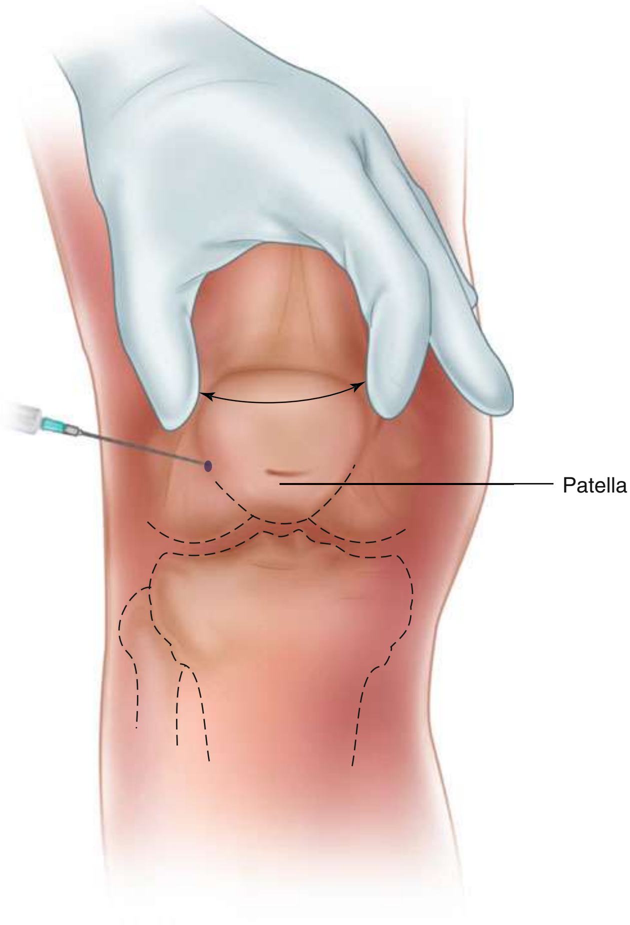

Fig. 106.1 Knee arthrocentesis

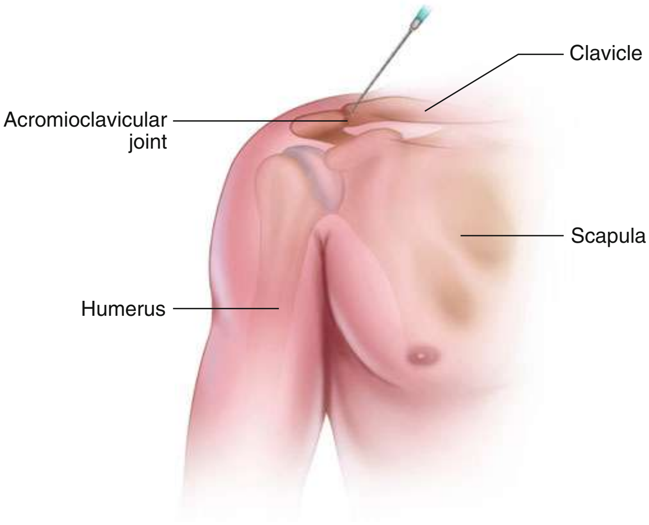

Fig. 106.2 Acromioclavicular joint arthrocentesis

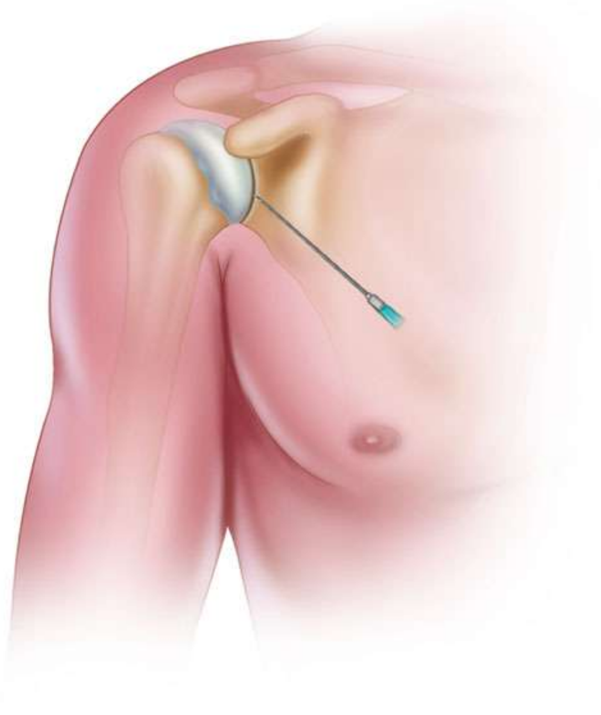

Fig. 106.3 Glenohumeral joint arthrocentesis: anterior approach

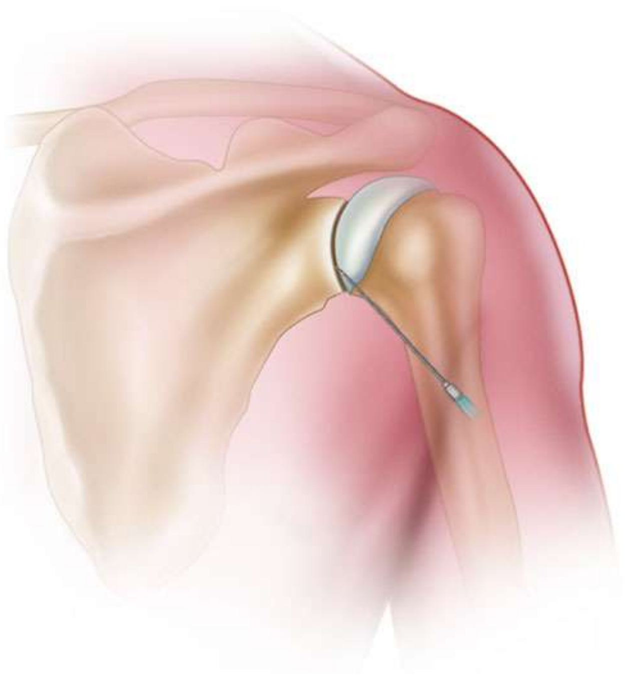

Fig. 106.4 Glenohumeral joint arthrocentesis: posterior approach

106.5 Complications

- Introduction of infection
- Bleeding

106.6 Pearls and Pitfalls

- The preferred site of entry is over the extensor surface of the joint. This will reduce the risk of damage to tendons, ligaments, and blood vessels.
- When assessing synovial fluid, the Rule of Twos may be used to differentiate among normal, inflammatory, and septic fluid. Normal synovial fluid has less than 200 white blood cells (WBCs)/mm^3. Noninflammatory synovial fluid has 200–2000 WBCs/mm^3. Inflammatory synovial fluid has greater than 2000 WBCs/mm^3 (but <50,000 WBCs/mm^3). Septic synovial fluid has greater than 75,000 WBCs/mm^3.
- Only septic synovial fluid will have a positive Gram stain and culture.

Selected Reading

Biundo JJ, Roberts N, Deodhar A. Regional musculoskeletal complaints. In: Stone JH, editor. A clinician's pearls and myths in rheumatology. New York: Springer Science; 2009. p. 433–4.

Parrillo SJ, Fisher J. Arthrocentesis. In: Roberts JR, Hedges J, editors. Clinical procedures in emergency medicine. 4th ed. Philadelphia: Saunders; 2004. p. 1042–57.

Self WH, Wang EE, Vozenilek JA, del Castillo J, Pettineo C, Benedict L. Dynamic emergency medicine. Arthrocentesis. Acad Emerg Med. 2008;15:298.

Thomsen TW, Shen S, Shaffer RW, Setnik GS. Arthrocentesis of the knee. N Engl J Med. 2006;354:e19.

Intra-articular Injection

107

Bharat Kothakota and Muhammad Waseem

107.1 Indications

- Aspiration of fluid (arthrocentesis)
  - For diagnosis: to rule out infection
  - To relieve pressure from large, painful joint effusion
- Injection of joints with inflammatory arthritis
  - Tendinitis
  - Bursitis
  - Rheumatoid arthritis (RA)
- Injection of joints with osteoarthritis (OA)
  - Injection of large weight-bearing joints
  - Injection of small joints of hands
- Intra-articular anesthetic
  - Shoulder reduction
  - Ankle impingement syndrome

107.2 Contraindications

- Cellulitis
- Bacteremia
- Fracture

107.3 Materials and Medications

- Glucocorticoid
  - Duration of effect inversely proportional to solubility
    - Less soluble → longer acting
  - Choice of steroid is the personal preference of the physician
  - Methylprednisone (Depo-Medrol) and triamcinolone acetonide (Kenalog)
    - Less likely to induce postinjection flare
  - Kenalog and triamcinolone hexacetonide (Aristospan)
    - Longest-acting agents
- Local anesthetic
  - 1 % Lidocaine
- Needle
  - 18- to 22-gauge used for knee, ankle, hip, elbow, and shoulder
  - 25-gauge or smaller used for smaller joints (interphalangeal)
- Syringe

B. Kothakota, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: bharatkothakota@yahoo.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_107

107.4 Procedure

1. Selecting an injection approach.
    - Knee
        - Lateral approach: 1 cm inferior to the patella laterally (Fig. 107.1)
        - Medial approach: 1 cm inferior to the patella medially (Fig. 107.2)
    - Ankle
        - Lateral approach: just inferior to the lateral malleolus (Fig. 107.3)
        - Medial approach
            - Plantar flex the foot.
            - Angle the needle cephalad to pass between the medial malleolus and the tibialis anterior tendon (Fig. 107.4).
    - Shoulder
        - Posterior approach
            - Insert the needle 1 cm inferior and 1 cm medial to the posterolateral corner of the acromion.
            - Direct the needle anterior and medial toward the coracoid process (Fig. 107.5).
2. Skin preparation.
    - Make three separate concentric outward spirals with iodine disinfectant
    - Scrub with cyclohexidine preparation
3. Mark the injection site by impressing the skin with a hard object.
    - Sterile end of needle sheath
    - Ball point pen with tip retracted
4. Local anesthesia.
    - 1 % Lidocaine injected into the skin and subcutaneous tissue
    - Short burst of ethyl chloride spray before iodine preparation
    - Mixing lidocaine with glucocorticoid preparation
5. Always aspirate joint fluid before injecting the corticosteroids.
    - Use a 1.5-inch 18-gauge needle for aspiration.
    - Confirm that the needle is in the joint space.
    - Reduced effusion size before injection can improve outcomes.
    - Compress the opposite side of the joint to aid in aspiration (Fig. 107.6).
6. After aspirating, change the syringe.
    - Use a sterile hemostat or hand to stabilize the needle within the joint space (Fig. 107.7).
    - Avoid injecting corticosteroids if the aspirate appears purulent.
7. Injection of medication.
    - Can use the same needle used for aspiration.
    - Insert needle 0.75–1.25 inch in depth for injection.
8. Remove the needle, wipe the iodine solution clean, and apply the bandage.
9. Postinjection care.
    - First 48 h: bedrest versus minimize walking
    - Next 2–3 weeks: crutches or cane

Fig. 107.1 Knee arthrocentesis, lateral approach

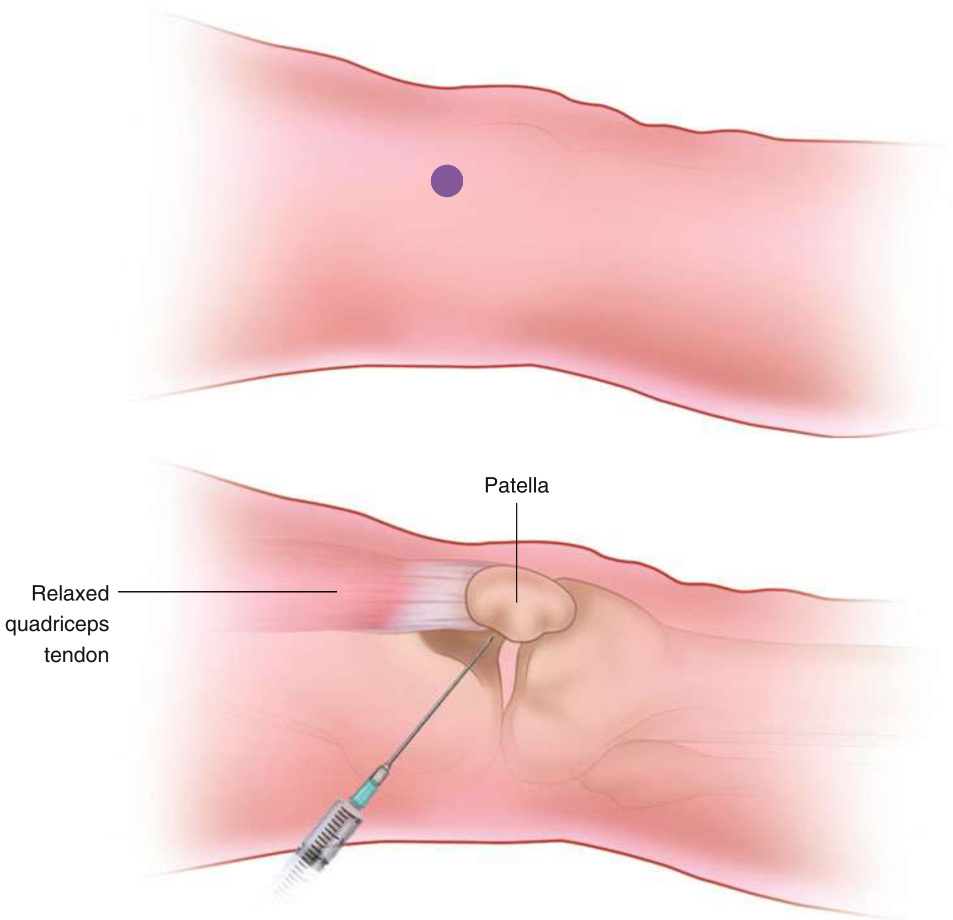
Patella
Relaxed quadriceps tendon

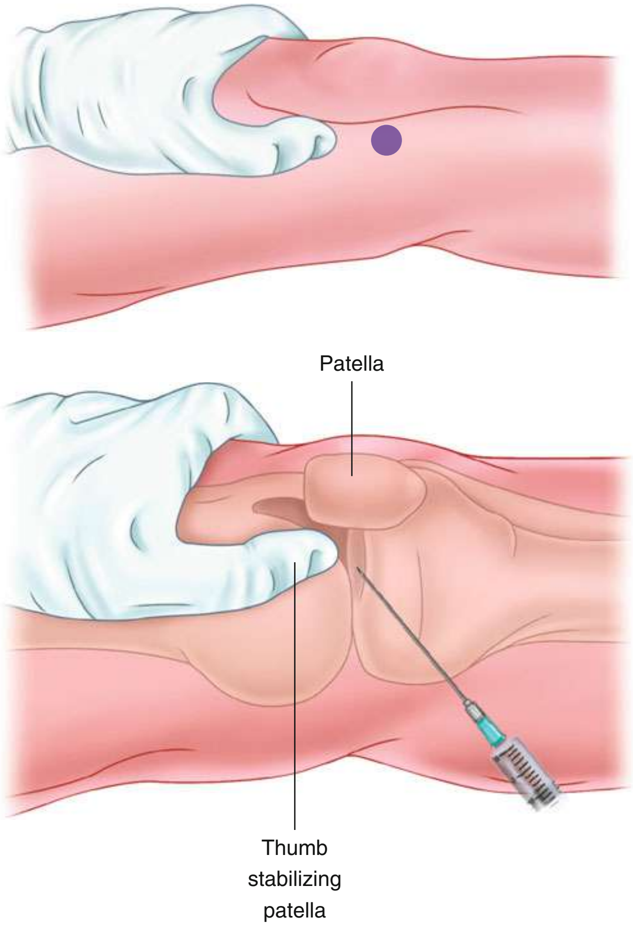
Patella
Thumb stabilizing patella

Fig. 107.2 Knee arthrocentesis, medial approach

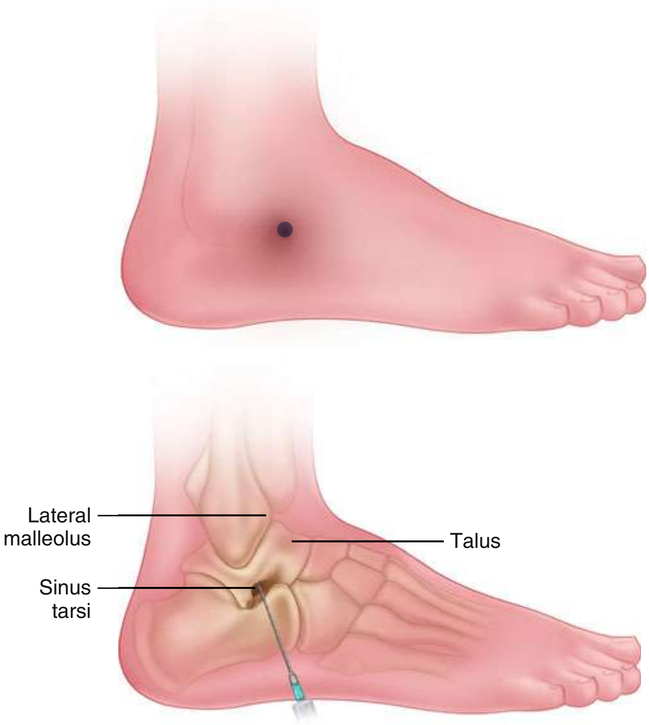
Lateral malleolus
Talus
Sinus tarsi

Fig. 107.3 Ankle arthrocentesis, lateral approach

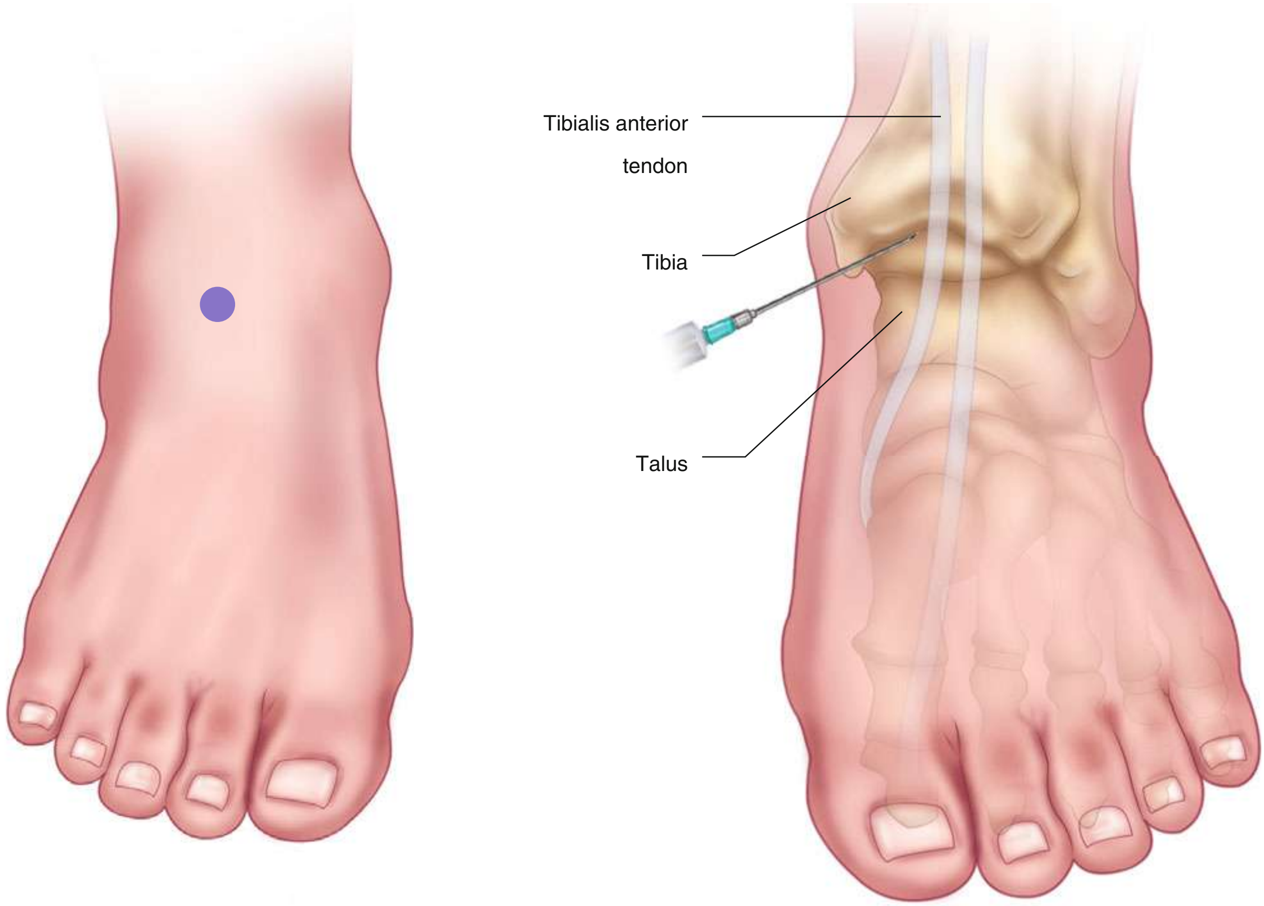
Tibialis anterior tendon
Tibia
Talus

Fig. 107.4 Ankle arthrocentesis, medial approach

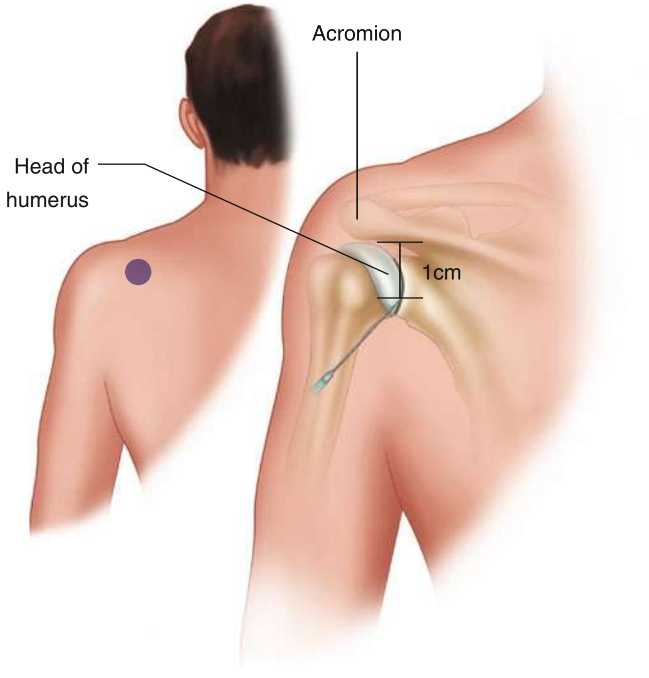
Acromion
Head of humerus
1cm

Fig. 107.5 Shoulder arthrocentesis, posterior approach

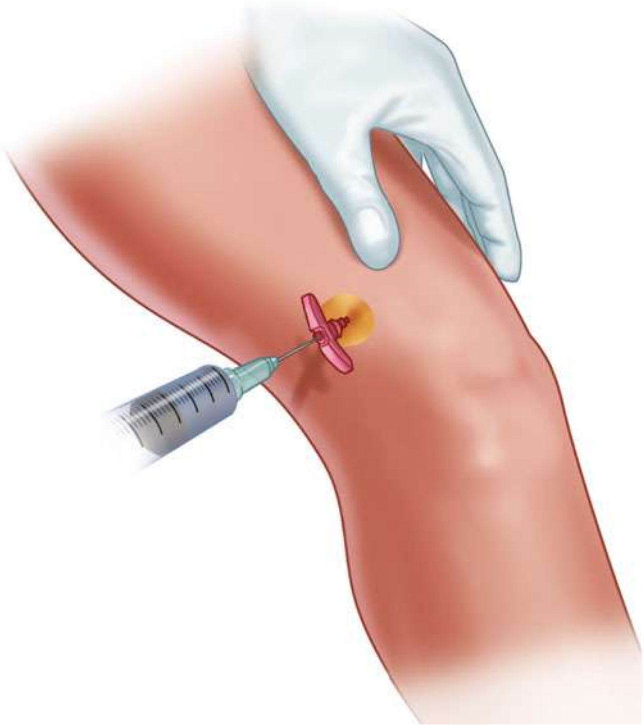

Fig. 107.6 Compress opposite side of joint to aid in aspiration

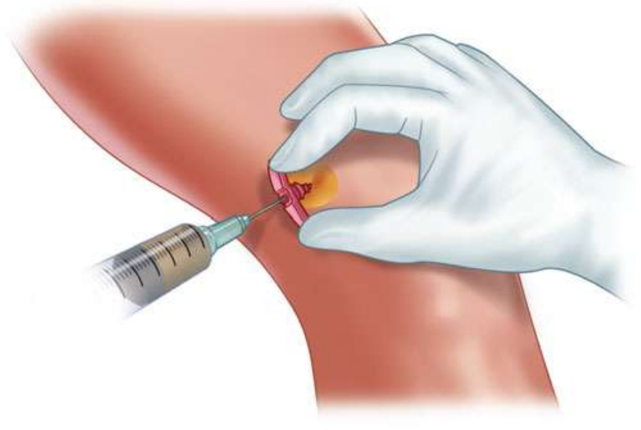

Fig. 107.7 Use hemostat or hand to stabilize needle within joint space

107.5 Complications

- Local postinjection flare
  - Irritation of the synovium by steroid microcrystals
  - Can be confused for infection
  - Occurs and resolves within 48 h after injection
  - Treat with ice and appropriate analgesics.
- Iatrogenic joint infection
  - Suspect if it begins later than, or lasts longer than, flare
  - Increasing pattern of pain
  - Fever, malaise, redness, or drainage around injection site
  - *Staphylococcus aureus* most common
- Subcutaneous atrophy and depigmentation
  - Leakage of corticosteroids into soft tissues
- Aspiration of blood
  - Indicative of trauma or bleeding disorder (hemophilia)
- Systemic absorption
  - Water-soluble preparations
  - Dose dependent
  - Injection into multiple joints
  - Transient hyperglycemia in diabetic patients
  - Avascular necrosis of the femoral head

107.6 Pearls

- Mixing lidocaine with glucocorticoids
  - Reduces pain caused by injection of steroids into joint space
  - Less likely to cause soft tissue atrophy and tendon rupture
  - Immediate relief from anesthetic indicates proper injection
- Limiting intra-articular glucocorticoid injections per joint
  - OA
    - Four injections per lifetime of the joint
    - Injections reduce the rate of accelerated degeneration in joints
  - RA
    - Limit of one injection per month
    - No evidence of glucocorticoid-induced cartilage loss

Selected Reading

Aponte EM, Schraga ED. Joint reduction, shoulder dislocation, anterior. Available at: http://emedicine.medscape.com/article/109130-overview#a08.

Cianflocco AJ. Intra-articular injections of the knee: a step-by-step guide. J Fam Pract. 2011;60(Suppl):S48–9. Available at: http://www.jfponline.com/pages.asp?id=10062.

Lavelle W, Lavelle ED, Lavelle L. Intra-articular injections. Anesthesiol Clin. 2007;25:835–62.

Molis MA, Young CC. Ankle impingement syndrome. Available at: http://emedicine.medscape.com/article/85311-overview.

Neustadt DH. Intra-articular injections for osteoarthritis of the knee. Cleve Clin J Med. 2006;73:897–911.

Roberts WN. Intraarticular and soft tissue injections: what agents(s) to inject and how frequently? Available at: www.uptodate.com.

Roberts WN. Joint aspiration or injection in adults: techniques and indications. Available at: www.uptodate.com.

108

Sugar-Tong Splint

Katrina John, Jeffrey Kile, and Amish Aghera

108.1 Indications

- Fractures to the wrist or forearm
- To prevent motion at the wrist and elbow
- To prevent supination and pronation

108.2 Contraindications

- Relative
  - Evidence of compartment syndrome or any neurovascular compromise

108.3 Materials (Fig. 108.1)

- Stockinette
- Padding
- Splint material: fiberglass/plaster of Paris or prefabricated splint rolls, 2, 3, 4 inches depending on age and body habitus
- Trauma shears/scissors
- Elastic bandage
- Tape
- Container with water
- Gloves, eyemask, sheet
- Sling

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_108

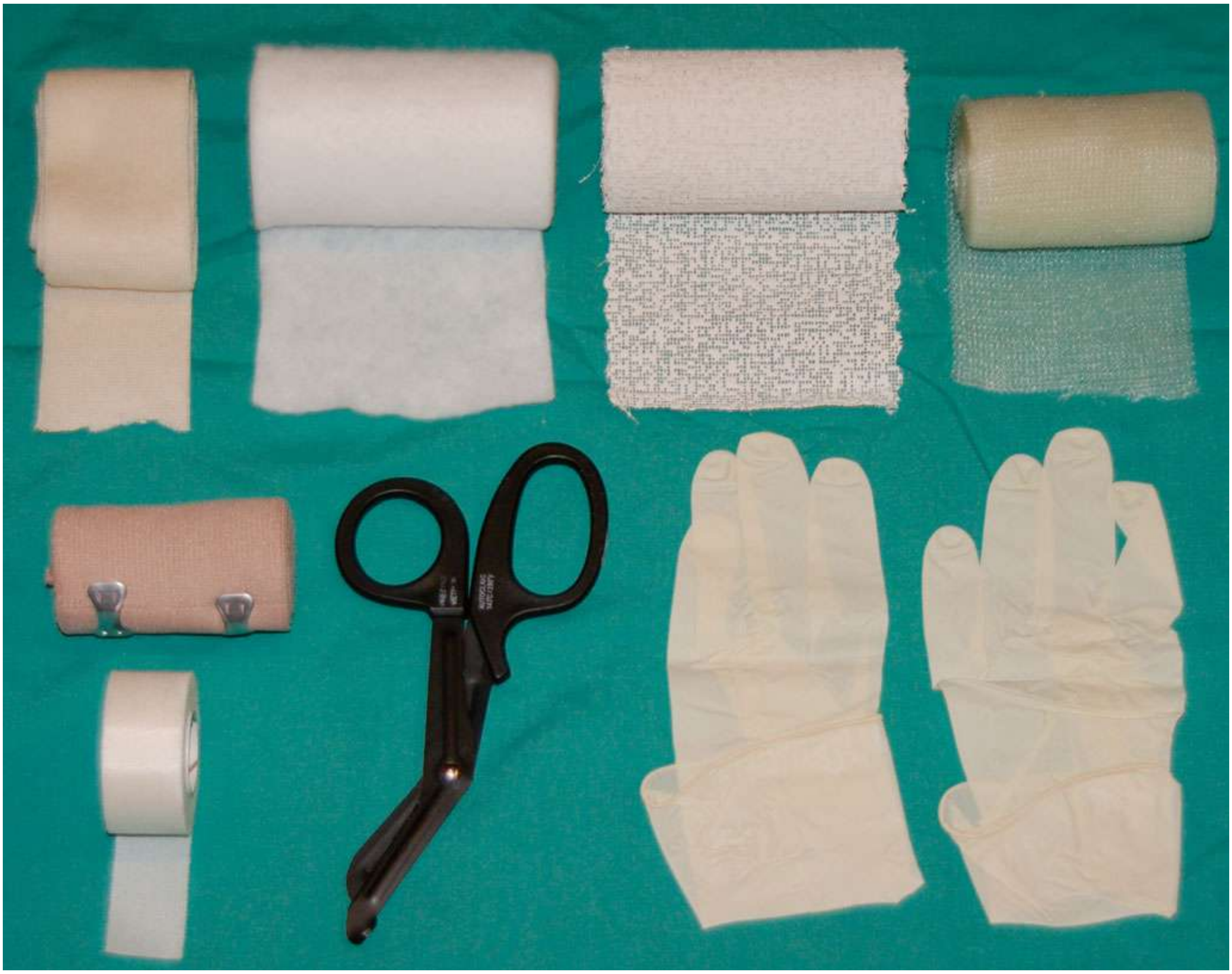

Fig. 108.1 Equipment

108.4 Procedure

1. Ensure the skin of the affected extremity is clean, dry, and intact.
2. Cover the patient with a sheet or gown to protect the patient's clothing and the surrounding area.
3. Position the patient's arm abducted at 90° at the shoulder and internally rotated with the elbow flexed at 90° (Fig. 108.2).
4. Measure the splinting material by running a single layer from the metacarpal heads of the dorsum of the hand along the extensor surface of the forearm over the elbow and humeral condyles and back down the flexor surface of the forearm to the palmar aspect of the hand to the metacarpal heads.
5. If using non-prefabricated splint rolls, lay the measured piece of splinting material out on a flat surface and multiply the layers to the same length, 6–8 layers for fiberglass and 10–12 layers for plaster.
6. Measure the stockinette from the finger tips to the midhumerus and cut a hole for the thumb.
7. Place the stockinette on the arm (Fig. 108.3).
8. Use a 3- to 4-inch padding roll to apply several layers of circumferential padding extending from the metacarpal heads to the midhumerus below the level of the stockinette (Fig. 108.4).
9. Wet the already prepared and measured splinting material and remove the excess water.
10. Ensure the forearm is in the aforementioned position, and apply the splinting material from the metacarpal heads of the dorsum of the hand along the extensor surface of the forearm over the elbow and humeral condyles and back down the flexor surface of the forearm to the palmar aspect of the hand to the metacarpal heads (Fig. 108.5).
11. Fold each end of the stockinette down over the padding and splinting material.
12. An extra layer of padding can be added at this stage.
13. Secure the entire splint with two elastic bandages/ace wrap and apply tape to ensure the bandages stay in place (Fig. 108.6).
14. Place the arm in a sling.

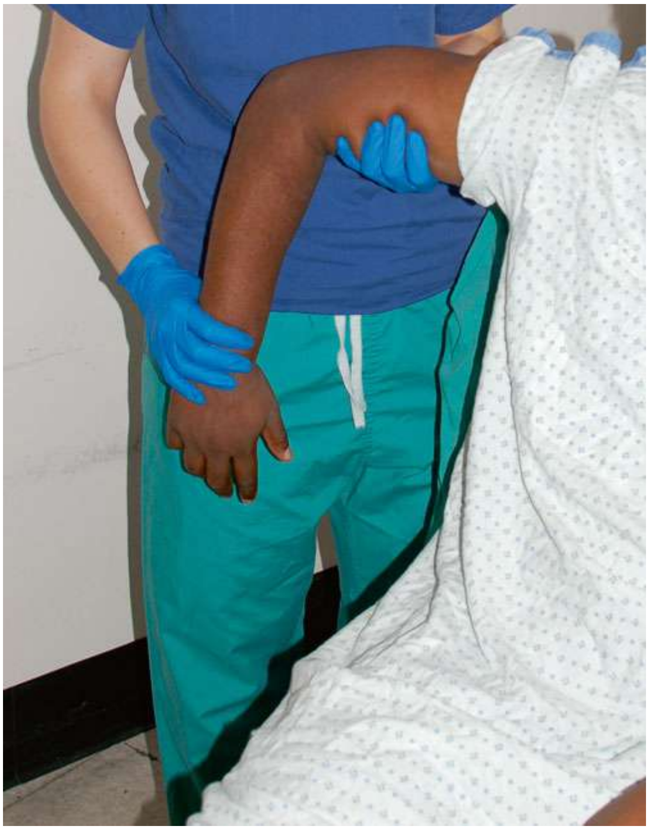

Fig. 108.2 Arm positioning

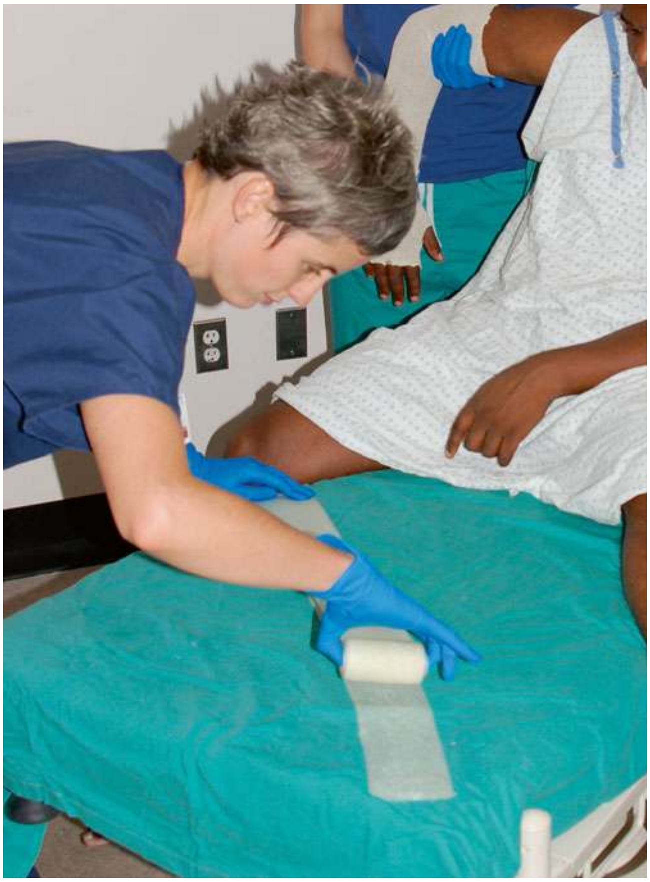

Fig. 108.3 Fiberglass being measured and stockinette on patient

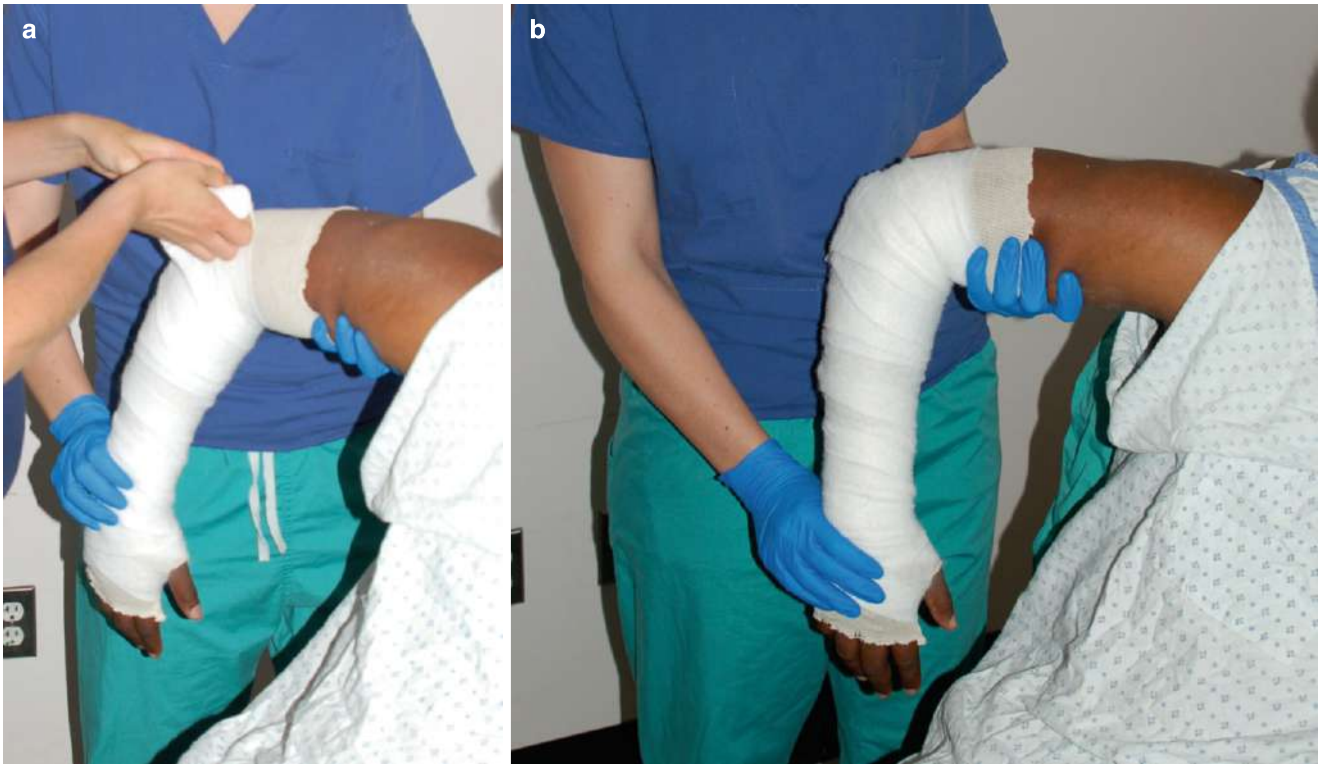

Fig. 108.4 (**a**) Application of padding over the stockinette (partial), (**b**) Application of padding over the stockinette (complete)

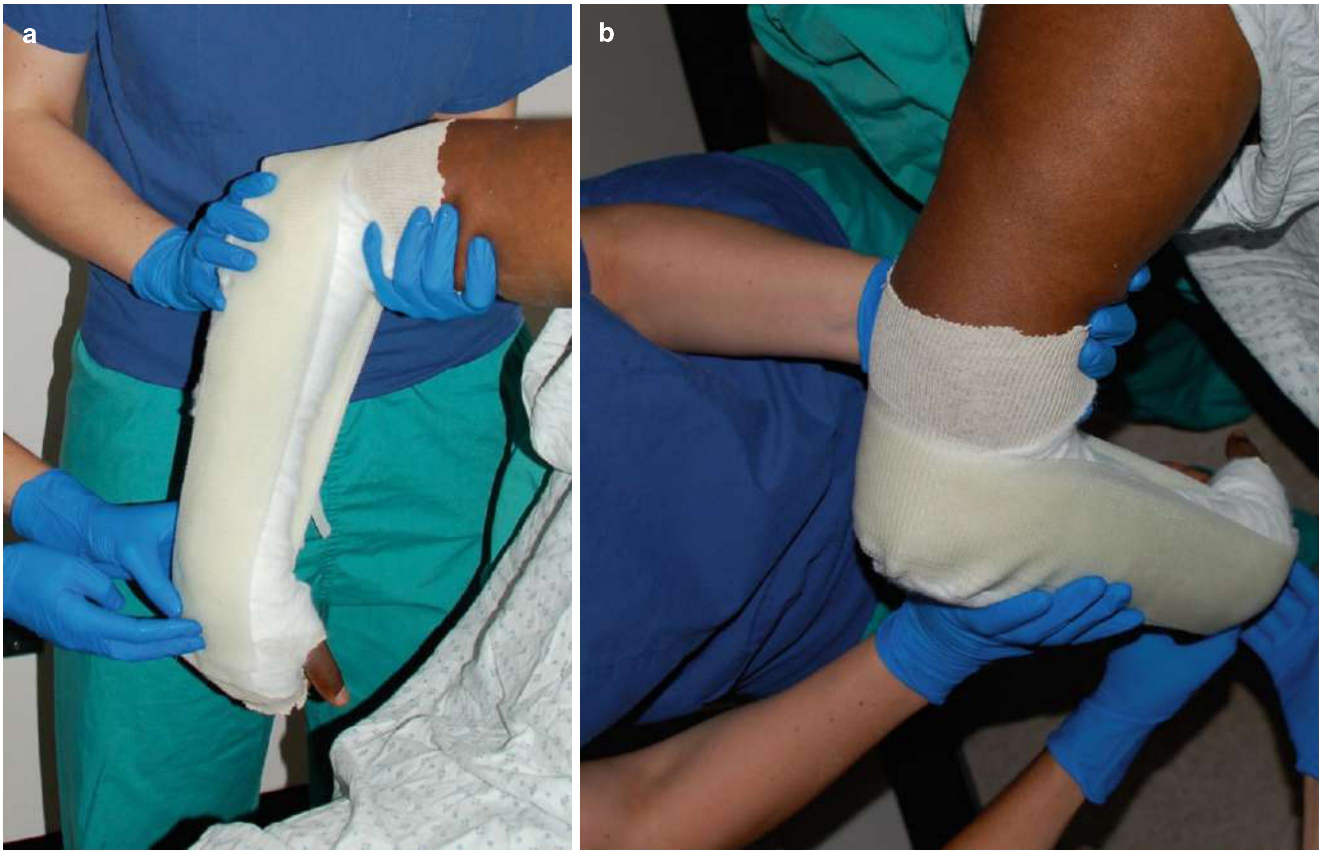

Fig. 108.5 (**a**) Application of splinting material over padding (anteromedial view), (**b**) Application of splinting material over padding (anterolateral view)

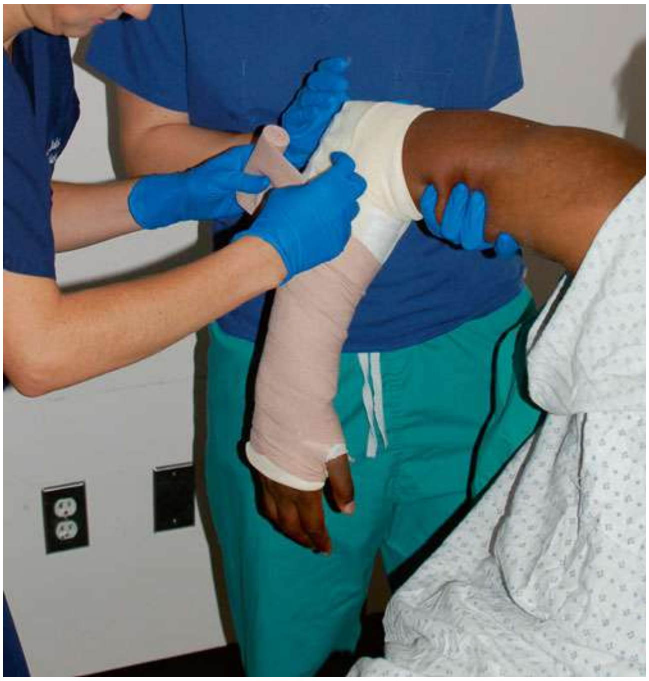

Fig. 108.6 Application of elastic bandage over splinting material

108.5 Complications

- Compartment syndrome

108.6 Pearls and Pitfalls

- Pearls
  - This procedure is best done with an assistant to hold the extremity in the desired position and to prevent the splint slipping as it is secured.
  - Having the patient in the illustrated position enables the practitioner to use gravity to hold the splint in the correct position while securing it; especially helpful if an assistant is not available.
- Pitfalls
  - If the splint is too short, it fails to immobilize the wrist.
  - If the splint is too long, it will cause reduced motion and stiffness at the metacarpophalangeal joints and swelling of the fingers due to immobility.

Selected Reading

Bong MR, Egol KA, Leibman M, Koval KJ. A comparison of immediate postreduction splinting constructs for controlling initial displacement of fractures of the distal radius: a prospective randomized study of long-arm versus short-arm splinting. J Hand Surg Am. 2006;31:766–70.

Denes AE, Goding R, Tamborlane J, Schwartz E. Maintenance of reduction of pediatric distal radius fractures with a sugar-tong splint. Am J Orthop (Belle Mead NJ). 2007;36:68–70.

Gartland JJ. The sugar tong splint. Am J Orthop. 1963;5:131.

McGeorge DD, Stilwell JH. The sugar tong splint: a reliable method of arm splintage in the child. J Hand Surg Br. 1989;14:357.

Simon RR, Koenigsknecht SJ, editors. Emergency orthopedics—the extremities. 3rd ed. New York: McGraw Hill; 1995.

Part XIV

Obstetric Procedures

Fetal Heart Rate Monitoring

109

Nathaniel Lisenbee and Joseph A. Tyndall

109.1 Indications

- Fetal heart rate (FHR) monitoring is important because it provides basic patterns that can be correlated to the acid–base status, circulatory volume, and oxygenation status of the fetus through brainstem detection and subsequent cardiac response. It has numerous indications during the antepartum and intrapartum stages [1].
- Antepartum indications include:
  - Nonstress test (consists of monitoring FHR in conjunction with fetal movements)
  - Contraction stress test (consists of monitoring FHR during contractions, which are induced pharmacologically)
  - Biophysical profile (BPP; consists of a nonstress test with an additional ultrasound)
- Intrapartum indications include monitoring FHR during:
  - Uterine contractions
  - Pain medications/anesthetic administration to the mother during labor
  - Procedures performed during labor
  - Second stage of labor
  - High-risk pregnancies, which can be defined by a number of conditions including [2, 3]:
    - Maternal diabetes, asthma, preeclampsia/eclampsia
    - Multiple gestations
    - Intrauterine growth restriction
    - Premature rupture of membranes
    - Lack of prenatal care

109.2 Contraindications

- Contraindications for internal FHR monitoring
  - Presence of placenta previa
  - Lack of ability to identify the portion of the fetal body where device application is being considered
  - Active herpes, active hepatitis, or human immunodeficiency virus (HIV) in the mother
- Contraindications for external FHR monitoring
  - None

109.3 Methods

- Two methods for FHR monitoring:
  - Auscultation monitoring
    - Defined as auscultating FHR every 15 min in the first stage of labor and auscultating every 5 min in the second stage
    - Does not provide strips with information on FHR variability or the shape of FHR accelerations and decelerations
- Electronic FHR monitoring
  - Allows for real-time continuous monitoring of FHR activity
  - Provides strips with information on FHR variability or the shape of FHR accelerations and decelerations
  - Can be performed by Doppler ultrasound or internal fetal electrocardiography (ECG)

When comparing the two methods for electronic FHR monitoring, both are equally as reliable in most settings. Thus, external monitoring is the preferred method because it is noninvasive. However, in instances in which external monitoring becomes difficult owing to poor quality or technical difficulties, invasive monitoring is indicated.

N. Lisenbee, MD
U. S. Air Force Medical Center Keesler,
Biloxi, MI, USA
e-mail: nate4085@ufl.edu

J.A. Tyndall, MD, MPH (✉)
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: tyndall@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_109

109.4 Equipment and Procedures

Multiple methods exist for electronic FHR monitoring [4]. The most commonly used are external monitoring by Doppler ultrasound and internal monitoring by fetal ECG.

109.4.1 Doppler Ultrasound is a Noninvasive Method to Monitor FHR (Fig. 109.1)

- Equipment
  - Electronic FHR monitor
  - Contraction monitor sensor with belt
  - FHR sensor with belt (consists of ultrasound transducer and ultrasound sensor)
  - Ultrasound coupling gel
- Procedure
  1. Place the patient in a supine position.
  2. Palpate the fetal anatomy through the maternal abdomen to find the approximate location of the fetal heart.
  3. Place ultrasound coupling gel on the maternal abdomen at the sight of suspected fetal cardiac activity.
  4. Place the transducer probe on gel and locate the fetal heart tones.
  5. Once the fetal heart tones are located, secure the FHR sensor to the maternal abdomen with the attached belt.
  6. Place the contraction monitor sensor near the fundus in order to monitor uterine contractions.
  7. Attach the FHR sensor and contraction monitor to the electronic FHR monitor to obtain printouts of FHR and uterine contractions.

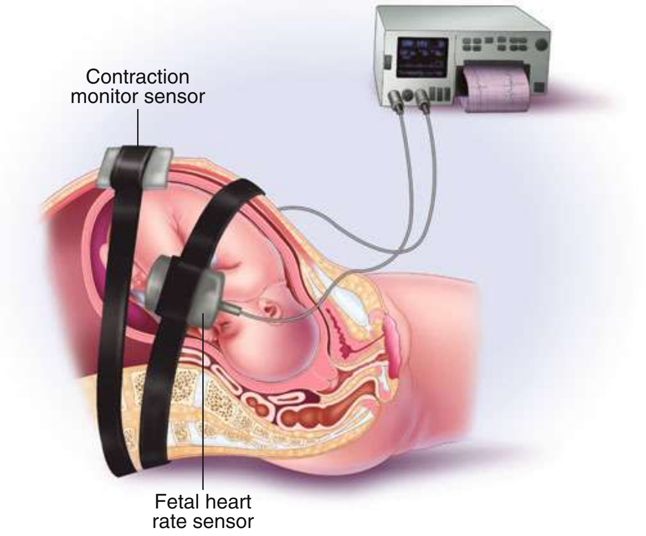

Fig. 109.1 External fetal heart rate monitoring

109.4.2 Internal Fetal ECG is an Invasive Method to Monitor FHR and is Used Only in the Intrapartum Period (Fig. 109.2)

- Equipment
  - Fetal scalp monitoring electrode
  - Leg plate electrode
  - Sterile vaginal lubricant
  - Electronic FHR monitor
- Procedure (Fig. 109.3)
  1. Place the patient in a dorsal lithotomy position.
  2. Sterilize the perineal area.
  3. Perform a bimanual vaginal examination to identify the presenting fetal head. (Note: rupture of membranes must occur before scalp electrode placement.)
  4. Place the spiral electrode guide tube on the fetal scalp and advance the electrode until it contacts the scalp.
  5. Rotate the drive tube clockwise approximately one rotation while maintaining pressure on the guide tube and drive tube.
  6. Release the electrode locking device by pressing together the arms on the drive tube grip.
  7. Carefully slide the drive and guide tubes off the electrode wires while holding the locking device open.
  8. Attach the leg plate to the inner thigh of the mother as a means to eliminate electrical interference.
  9. Attach the spiral electrode wires to the color-coded leg plate, which is then connected to the electronic fetal monitor.
  10. Do not forget to sterilize the area of electrode placement after delivery is completed and the scalp electrode is removed.

When comparing the two methods, both are equally reliable in most settings. Thus, external monitoring is the preferred method because it is noninvasive. However, in instances in which external monitoring becomes difficult owing to poor quality or technical difficulties, invasive monitoring is indicated.

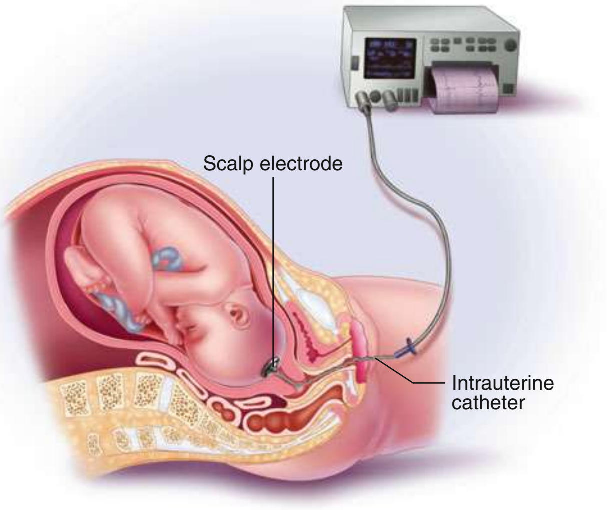

Fig. 109.2 Internal fetal heart rate monitoring

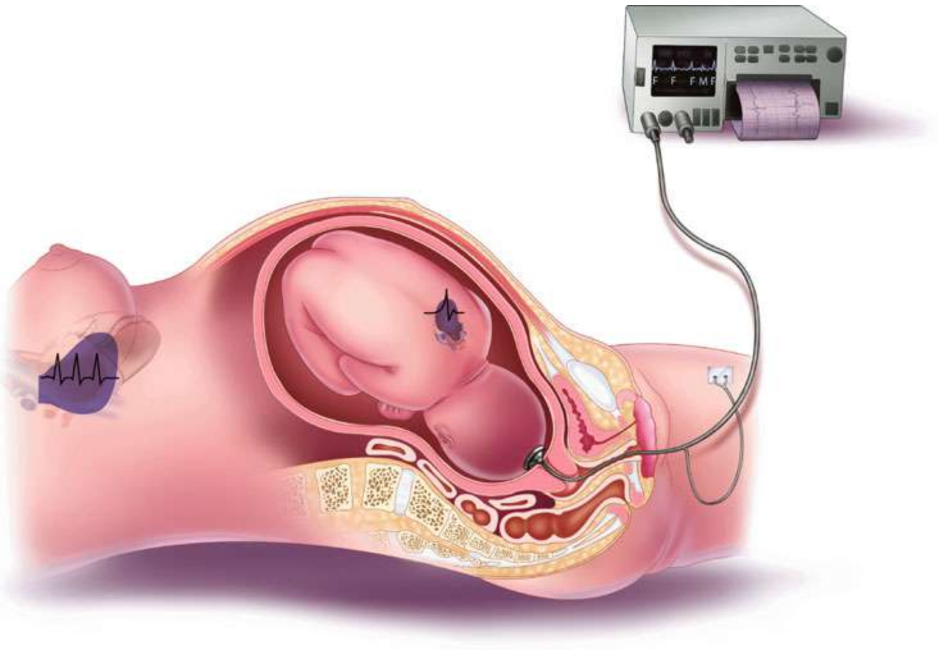

Fig. 109.3 Internal fetal heart rate monitoring

109.5 Complications

- Complications of external FHR monitoring
  - Confusing maternal aortic pulsations with FHR
  - Inability to locate FHR
- Complications of internal FHR monitoring
  - Fetal or maternal hemorrhage, fetal infection (usually scalp abscess at the site of insertion)
  - Uterine perforation
  - Subsequent fetal infection due to the invasive nature of the procedure

References

1. Hobel CJ. Intrapartum clinical assessment of fetal distress. Am J Obstet Gynecol. 1971;110:336–42.
2. Byrd JE. Intrapartum electronic fetal heart rate monitoring (EFM) and amnioinfusion. In: Advanced life support in obstetrics course syllabus. Kansas City: American Academy of Family Physicians; 1996. p. 97–106.
3. Queenan JT, Hobbins JC, Spong CY. Protocols for high-risk pregnancies. New York: Wiley; 2010. Retrieved 16 Jan 2012, from http://lib.myilibrary.com?ID=268955.
4. External and internal heart rate monitoring of the fetus. New Haven: Yale Medical Group; 2012. Retrieved from http://www.yalemedicalgroup.org/stw/.

Selected Reading

Alfirevic Z, Devane D, Gyte GM. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database Syst Rev. 2006;(3):CD006066.

American College of Obstetricians and Gynecologists. Fetal heart rate patterns: monitoring, interpretation, and management, ACOG technical bulletin, vol. 207. Washington, DC: ACOG; 1995.

Cunningham FG, Leveno KJ, Bloom SL, Hauth JC, Rouse DJ, Spong CY. Chapter 18: Intrapartum assessment. In: Williams obstetrics. 23rd ed. New York: McGraw Hill; 2010.

Freeman RK. Problems with intrapartum fetal heart rate monitoring interpretation and patient management. Obstet Gynecol. 2002;100:813.

Gomella LG, Haist SA. Chapter 13: Bedside procedures. In: Clinician's pocket reference. 11th ed. Columbus: McGraw-Hill; 2007.

Sweha A, Hacker TW, Nuovo J. Interpretation of the electronic fetal heart rate during labor. Am Fam Physician. 1999;59:2487–500.

Ultrasonography for Ectopic Pregnancy

110

L. Connor Nickels

An ectopic pregnancy is a pregnancy occurring outside of the uterine cavity (fundus) (Fig. 110.1).

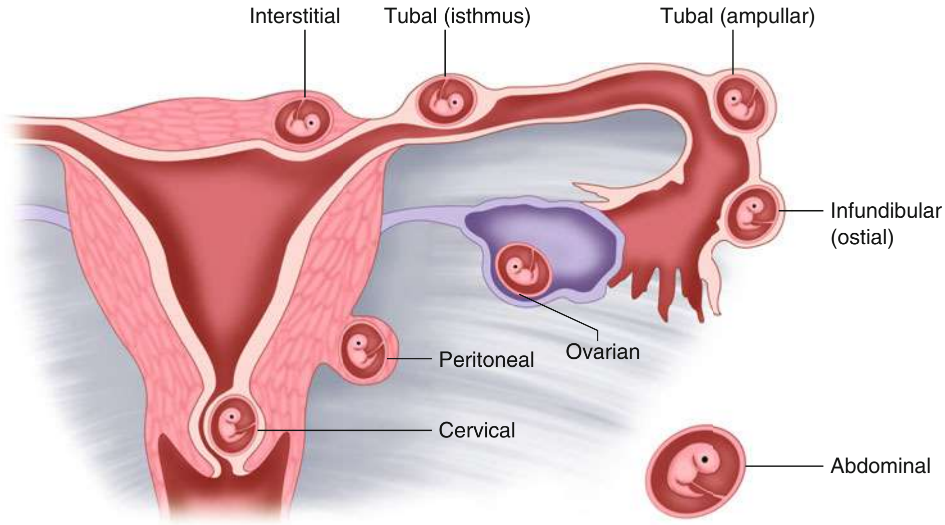
Interstitial
Tubal (isthmus)
Tubal (ampullar)
Infundibular
(ostial)
Ovarian
Peritoneal
Cervical
Abdominal

Fig. 110.1 Ectopic pregnancy diagram: pregnancy occurring outside of the uterine cavity

---

L.C. Nickels, MD, RDMS
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_110

110.1 Indications

- Patient in the first trimester of pregnancy with any combination of the following:
  - Vaginal bleeding.
  - Acute pelvic pain.
  - Hypotension or shock.
  - Dizziness or syncope.
  - Positive pregnancy test.
  - Adnexal mass.
  - Cervical tenderness.
  - Abnormal rise in the serum human chorionic gonadotropin (hCG).
  - No specific sign or symptom is absolute; therefore, the index of suspicion must be high.
- Risk factors for ectopic pregnancy
  - Pelvic inflammatory disease (PID)
  - Previous or current intrauterine device
  - Infertility treatment
  - Previous tubal surgery
  - Tubal ligation
  - Advanced maternal age
  - Previous ectopic pregnancy
- Usually a sonographic diagnosis
- Sonographic signs of an ectopic can include any of the following:
  - Gestational sac seen in one of the following:
    - Adnexa with any of the following:
      - Yolk sac (Figs. 110.2 and 110.3)
      - Fetal pole with or without cardiac activity
      - Both of these
    - Low position in the cervix (cervical ectopic)
    - Seemingly in the uterus, but off to one side and with minimal surrounding myometrium (interstitial ectopic)
    - Within the peritoneal cavity, outside of the tubes (abdominal ectopic)
  - Pseudogestational sac seen in the uterus
    - Uterine enlargement or decidual reaction (single outline only) in the endometrium without a gestational sac
  - Other unidentifiable adnexa mass
  - Free fluid in the pelvis or other gravity-dependent area (i.e., Morison's pouch in the right upper quadrant in a supine or reverse Trendelenburg patient) (Fig. 110.4)
    - Small free fluid: tracks less than one third of the posterior cul-de-sac
    - Moderate free fluid: tracks less than two third of the posterior cul-de-sac
    - Large free fluid: tracks greater than two third of the posterior cul-de-sac
    - Right upper quadrant free fluid: 100 % predictability for ectopic
  - Empty uterus with serum hCG >1000 mIU/mL
- Patient in the first trimester of pregnancy with a serum hCG at or above the discriminatory zone (i.e., 1000 mIU/mL) without a sonographically normal gestational sac visualized within the uterus has an ectopic pregnancy until proven otherwise and should have an obstetrics consult in the emergency department.
  - Discriminatory zone may differ depending on the reference, but typically serum hCG between 1000 and 2000 mIU/mL.
  - For this chapter, we have used serum hCG greater than 1000 mIU/mL [1].
- Yolk sac should be first sign of definitive intrauterine pregnancy for emergency physicians because the decidual reaction and gestational sac are not 100 % accurate.
- May be treated surgically or medically depending on findings.
  - Decision to be made by obstetrics consultant

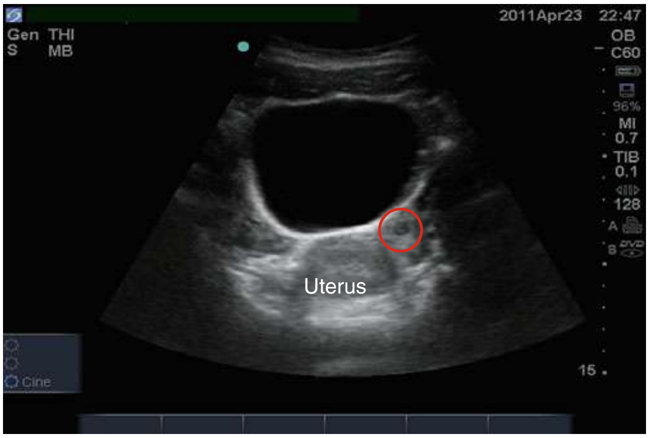
2011Apr23 22:47
Gen THI
S MB
OB
C60
96%
MI
0.7
TIB
0.1
128
Uterus
Cine
15

Fig. 110.2 Transabdominal transverse image of the uterus shows an ectopic pregnancy in the left adnexa (red circle) with the yolk sac present (Photo courtesy of L. Connor Nickels, MD, RDMS)

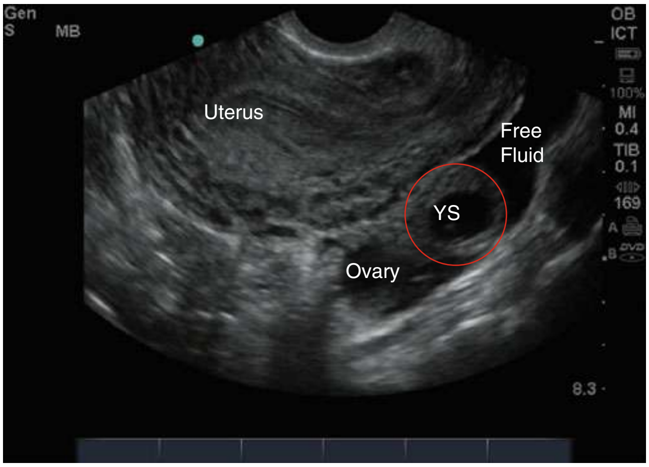
Gen
S MB
OB
ICT
100%
MI
0.4
TIB
0.1
169
Uterus
Free
Fluid
YS
Ovary
8.3

Fig. 110.3 Transvaginal sagittal image of an empty uterus with an ectopic pregnancy (in red circle) noted just posterior in the adnexa, adjacent to the ovary. The gestational sac contains a yolk sac and there is free fluid surrounding the ectopic, concerning for rupture (Photo courtesy of L. Connor Nickels, MD, RDMS). YS–yolk sac

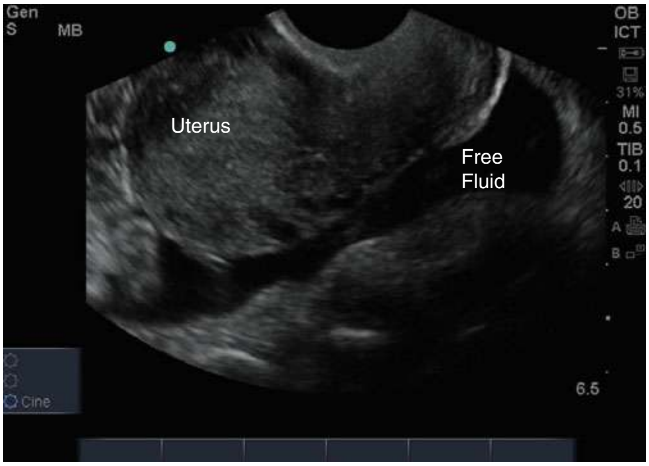
Gen
S MB
OB
ICT
31%
MI
0.5
TIB
0.1
20
Uterus
Free
Fluid
Cine
6.5

Fig. 110.4 Transvaginal sagittal image of an empty uterus and large free fluid in the posterior cul-de-sac in a patient with a presumed ectopic pregnancy (Photo courtesy of L. Connor Nickels, MD, RDMS)

110.2 Contraindications

- Treating ectopic pregnancy medically when there is a fetal pole with fetal cardiac activity present
- Failing to consult obstetrics/gynecology when the patient has a serum hCG above the discriminatory zone and no sonographic findings to diagnose intrauterine pregnancy

110.3 Materials and Medications

- Ultrasound machine
- Probes: transabdominal and transvaginal
- Gel
- Skilled ultrasound operator
- Endocavitary probe covers
- Pelvic setup (speculum and cultures)
- Cardiac monitor, two large-bore intravenous needles
- Laboratory work: serum quantitative hCG, hemoglobin, group and Rh, type and screen
  - May add additional laboratory tests depending on the stability and symptoms of the patient

110.4 Procedure

- Ultrasound machine in obstetrics preset
- Transabdominal
  1. Place the patient in the supine position.
  2. Ideally, the bladder will be full for a good acoustic window.
  3. Using a curvilinear probe, 3.5–5.0 MHz.
  4. Begin scanning the patient in a sagittal position to identify the uterus as it lies in position with the bladder. Scan through the uterus completely in this plane, looking for signs of intrauterine pregnancy.
  5. Imaging in the transverse plane should be done in the same fashion.
  6. Any signs of an intrauterine pregnancy should be clearly identified and measured.
     - Gestational sac diameter
     - Yolk sac diameter
     - Fetal pole with crown rump diameter
       - If present, fetal cardiac activity should be recorded by using M mode to obtain a tracing and measure the fetal heart rate.
  7. Although sometimes limited transabdominally, an attempt to identify the adnexa should be performed in both planes bilaterally.
  8. Any abnormalities identified should be noted.
     - Free fluid surrounding the uterus (anterior or posterior cul-de-sac) or ovaries
     - On pelvic imaging or FAST examination
     - Intrauterine contents not consistent with an intrauterine pregnancy and/or not clearly, centrally visualized within the fundus of the uterus
     - Masses or contained fluid collections outside the uterus
     - Yolk sac or fetal pole with or without fetal cardiac activity seen outside the uterus
  9. Transvaginal scanning should be performed if a definitive intrauterine pregnancy is not identified on transabdominal imaging.
- Transvaginal
  1. Place the patient in the lithotomy position.
  2. Bladder is preferably empty.
  3. Use a transvaginal probe, 5–7.5 MHz.
  4. Repeat same steps as transabdominal imaging previously.
- The procedure is the same for all pelvic ultrasounds including transabdominal and transvaginal imaging because this should be performed in a systematic fashion so as to not miss pertinent findings. Therefore, the findings may change, but the examination remains the same.
- Yolk sac should be first sign of definitive intrauterine pregnancy for emergency physicians because the decidual reaction and gestational sac are not 100 % accurate.

110.5 Complications

- Bleeding: internal and/or external
- Maternal death if ectopic ruptures
  - Nine percent of pregnancy-related deaths
  - Leading cause of maternal death in the first trimester
- Sterility if tube(s) are damaged or surgically removed

110.6 Pearls and Pitfalls

- Pearls
  - Following the algorithm despite the patient being asymptomatic can help avoid missed ectopic pregnancies.
  - Cervical ectopic pregnancy can be difficult to distinguish from a spontaneous abortion. If the patient is aborting, the ultrasound findings should change quickly and the patient should have vaginal bleeding.
  - Double decidual reaction versus pseudogestational sac can be a very subtle distinction and should not be made by emergency physicians; hence, the statement made earlier, requiring a yolk sac as the earliest definitive sign of an intrauterine pregnancy.

  - Fibroids, bicornuate uterus, and eccentrically located normal pregnancy can all appear similar to a cornual pregnancy.
  - The majority of times when the patient is pregnant, has an empty uterus on sonographic imaging, and has vaginal bleeding, the final diagnosis is still unknown because it could still be an early normal pregnancy or an ectopic pregnancy.
- Pitfalls
  - Failing to obtain a pregnancy test in all reproductive-age women who have not undergone a hysterectomy
  - Failing to identify subtle signs of ectopic pregnancies

Reference

1. Ma OJ, Mateer JR, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw-Hill Professional; 2007.

Selected Reading

Gabbe SG, Niebyl JR, Galan HL, et al., editors. Obstetrics: normal and problem pregnancies. 5th ed. Philadelphia: Churchill Livingstone; 2007.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

Stead LG, Behera SR. Ectopic pregnancy. J Emerg Med. 2007;32(2):205–6.

Ultrasonography for Hydatidiform Mole

111

L. Connor Nickels

Definition: molar pregnancy = hydatidiform mole = anomalous growth of trophoblastic tissue

- Complete: 46,XX or 46,XY
  - Completely paternal in origin
  - Contains no fetal tissue
  - Most recognizable by clinical symptoms
- Incomplete (partial mole): 69,XXX or 69,XXY
  - Maternal and paternal in origin
  - Contains fetal tissue
  - More subtle clinical presentation

111.1 Indications

- Clinical presentation of any combination of the following:
  - Vaginal bleeding
  - With or without vomiting, persistent hyperemesis gravidarum
  - High blood pressure
  - Uterine size large for dates
- Serum human chorionic gonadotropin (hCG) very elevated (more than would be consistent for dates)
  - Usually greater than 100,000 mIU/mL

111.2 Contraindications

- Conservative management

111.3 Materials and Medications

- Ultrasound machine
- Probes: curvilinear (abdominal) and endocavitary (transvaginal)
- Gel
- Skilled ultrasound operator
- Endocavitary probe covers
- Pelvic setup (speculum and cultures)
- Lab work: serum quantitative hCG, hemoglobin, group and Rh
  - May add additional laboratory tests depending on the stability and symptoms of the patient

111.4 Procedure

- Ultrasound machine in obstetrics preset
- Transabdominal imaging
  1. Place the patient in the supine position.
  2. Ideally the bladder will be full for a good acoustic window.
  3. Using a curvilinear probe, 3.5–5.0 MHz, begin scanning the patient in a sagittal position to identify the uterus as it lies in position with the bladder. Scan through the uterus completely in this plane, looking for signs of intrauterine pregnancy.
  4. Rotate the probe to a transverse position (counterclockwise), with the probe indicator toward the patient's right. Scan through the uterus completely in this plane, looking for signs of intrauterine pregnancy.
  5. Any signs of an intrauterine pregnancy should be clearly identified and measured.
  6. Sonographic findings of a molar pregnancy include:
     - Uterus filled with heterogeneous material.
       - Echogenic material interspersed with anechoic areas known as the "snowstorm" is the most common appearance of a mole (Fig. 111.1).

L.C. Nickels, MD, RDMS
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_111

    - Can be confused with missed abortion or fibroid, underscoring the importance of the serum hCG because this will be markedly elevated, which would make the alternative diagnoses less likely.
  - Theca lutein cysts are present in the adnexa.
    - Multiple cysts that occur with trophoblastic disease, multiple pregnancies (e.g., twins, triplets, quadruplets), and induced ovulation.
7. Although sometimes limited transabdominally, an attempt to identify the adnexa should be performed in both planes bilaterally.
8. Any abnormalities should be noted.
9. Transvaginal scanning should be performed if a definitive intrauterine pregnancy is not identified on transabdominal imaging.

- Transvaginal imaging
  1. Place the patient in the lithotomy position.
  2. Bladder is preferably empty.
  3. Use an endocavitary probe, 5–7.5 MHz.
  4. Repeat the same steps as for transabdominal imaging previously discussed (#2).
  5. After scanning through the uterus in sagittal and transverse planes, the adnexa should be scanned in the same format bilaterally, looking for any abnormalities, as mentioned previously.
  6. Again, any abnormalities should be noted:
     - Free fluid surrounding the uterus (anterior or posterior cul-de-sac) and/or ovaries
     - Masses or contained fluid collections outside the uterus
     - Yolk sac or fetal pole with or without fetal cardiac activity seen outside the uterus
- The procedure is the same for all pelvic ultrasounds, transabdominal and transvaginal imaging, because this should be performed in a systematic fashion so as to not miss pertinent findings. Therefore, the findings may change, but the examination remains the same.
- Yolk sac should be the first sign of definitive intrauterine pregnancy for emergency physicians because the decidual reaction and gestational sac are not 100 % accurate.

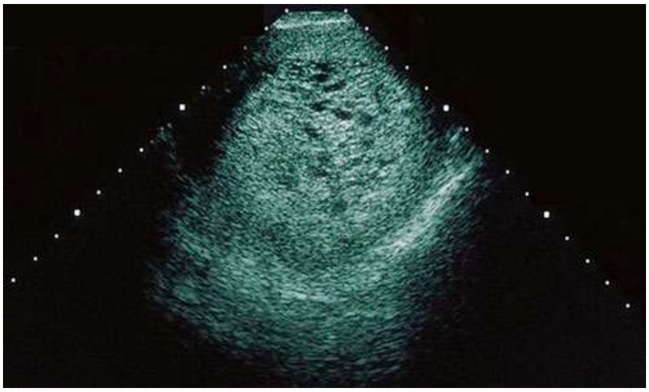

Fig. 111.1 Uterus with "cluster of grapes" represents a molar pregnancy (With kind permission from Springer Science+Business Media: Swisher E, Greer B, Montz FJ, Stenchever M. Chapter 14. In: *Atlas of Clinical Gynecology*. Vol. 4. 2002)

111.5 Complications

- Multiple complications can occur and include the following:
  - Invasive mole: when a hydatidiform mole recurs after a dilation and curettage and, subsequently, invades the muscle of the uterus.
    - Molar tissue is extremely vascular and necessitates an ultrasound with color flow to visualize any residual invasive tissue.
  - Choriocarcinoma: molar tissue develops into an aggressive malignancy, metastasizing early throughout the body.
    - In this scenario, the tumor takes on a very cystic appearance with an echogenic rim. When this is seen, one should attempt to view the liver as well for heterogeneous appearance consistent with metastasis or consider other imaging tests (more comprehensive ultrasound, computed tomography, or magnetic resonance imaging) as indicated to further assess concerns for malignancy.
    - Extremely sensitive to chemotherapy, but most favorable results are seen when diagnosed and treated early.

111.6 Pearls and Pitfalls

- Pearls
  - An obstetrics consult must be obtained in the emergency department when this diagnosis is made.
    - Patient must have a dilation and curettage to evacuate the mole.
    - Serum hCG must be followed to less than detectable levels because they have a high rate of recurrence (invasive mole) and potential for malignancy (choriocarcinoma).
  - May occur with intrauterine or ectopic pregnancies or after spontaneous abortions or full-term pregnancies
  - Qualitative and quantitative β-hCG results may be falsely negative in the setting of a molar pregnancy secondary to the “high-dose hook effect” found in sandwich immunoassays in which there is an overabundance of antigen. This error can be corrected by diluting the urine or serum sample and repeating the test.
- Pitfalls
  - Many pitfalls with potentially life-threatening consequences surround this diagnosis. Meticulous observation of the entire picture, including clinical presentation, laboratory results, and ultrasound findings, will help to avoid these outcomes.

Selected Reading

Gabbe SG, Niebyl JR, Galan HL, et al., editors. Obstetrics: normal and problem pregnancies. 5th ed. New York: Churchill Livingstone; 2007.

Hunter CL, Ladde J. Molar pregnancy with false negative â-hCG urine in the emergency department. West J Emerg Med. 2011;12:213–5.

Lentz GM, Lobo RA, Gershenson DM, et al., editors. Comprehensive gynecology. 6th ed. Philadelphia: Mosby; 2012.

Ma OJ, Mateer JR, Blaivas M. Emergency ultrasound. 2nd ed. New York: McGraw Hill Professional; 2008.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

Ultrasonography for Blighted Ovum (Anembryonic Gestation)

112

Katrina Skoog Nguyen and L. Connor Nickels

112.1 Indications

- A blighted ovum or anembryonic gestation is a pregnancy in which the embryo never develops in the gestational sac (Fig. 112.1).
- Presents as first-trimester vaginal bleeding and/or pelvic pain or cramping
- Criteria for defining an anembryonic pregnancy:
  - Based, in part, on mean gestational diameter (MGD), which is an average of three orthogonal measurements of the gestational sac
    - MGD greater than or equal to 25mm with no embryo
    - 5 MHz or less
      - MGD greater than or equal to 20 mm without a yolk sac
  - May have abnormally low sac position
- References differ in their measurements depending on the approach, transabdominal versus transvaginal, and the frequency of the probe (e.g., 5 MHz vs. 6.5 MHz or greater).

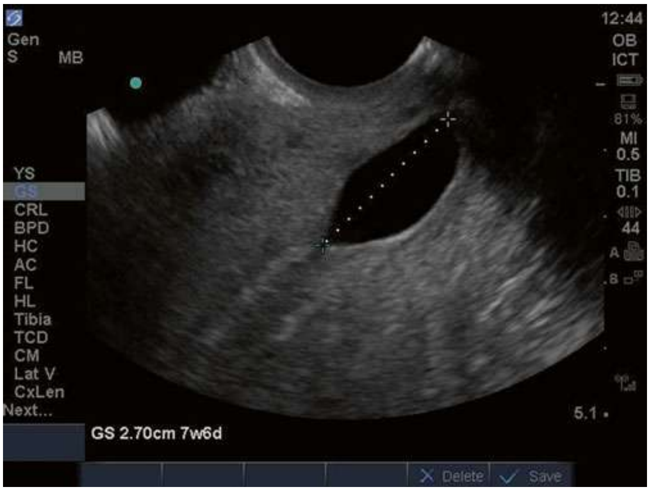

Fig. 112.1 Low-lying empty gestational sac consistent with blighted ovum (Photo courtesy of L. Connor Nickels, MD, RDMS)

K.S. Nguyen, DO • L.C. Nickels, MD, RDMS (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: katrinaskoog@hotmail.com; cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_112

112.2 Contraindications

- Misinterpreting images as blighted ovum when any of the following could potentially be present:
  - Early intrauterine pregnancy
  - Pseudogestational sac
  - Molar pregnancy
- Discussions with the patient should always involve an obstetrics consult or referral for definitive decisions and treatment.

112.3 Materials and Medications

- Ultrasound machine
- Probes: transabdominal and transvaginal
- Ultrasound gel
- Endocavitary probe covers
- Skilled ultrasound operator
- Laboratory tests: serum quantitative human chorionic gonadotropin (hCG), hemoglobin, group, and Rh
  - May add additional laboratory tests depending on the stability and symptoms of the patient

112.4 Procedure

- Ultrasound machine in obstetrics preset
- Transabdominal
  1. Place the patient in the supine position.
  2. Ideally the bladder will be full for a good acoustic window.
  3. Use a curvilinear probe, 3.5–5.0 MHz.
  4. Begin scanning the patient in a sagittal position to identify the uterus as it lies in position with the bladder. Scan through the uterus completely in this plane, looking for signs of intrauterine pregnancy.
  5. Imaging in the transverse plane should be done in the same fashion.
  6. Any signs of an intrauterine pregnancy should be clearly identified and measured.
     - In this case, gestational sac and any other intrauterine contents
  7. Although sometimes limited transabdominally, an attempt to identify the adnexa should be performed in both planes bilaterally.
  8. Any abnormalities should be noted.
     - Free fluid surrounding the uterus (anterior or posterior cul-de-sac) or ovaries
     - Masses or contained fluid collections outside the uterus
     - Yolk sac or fetal pole with or without fetal cardiac activity seen outside the uterus
  9. Transvaginal scanning should be performed if a definitive intrauterine pregnancy is not identified on transabdominal imaging.
- Transvaginal
  1. Place the patient in the lithotomy position.
  2. Bladder is preferably empty.
  3. Use a transvaginal probe, 5–7.5 MHz.
  4. Repeat the same steps as for transabdominal imaging previously discussed (#2).
- The procedure is the same for all pelvic ultrasounds including transabdominal and transvaginal imaging because this should be performed in a systematic fashion so as to not miss pertinent findings. Therefore, the findings may change, but the examination remains the same.
- Yolk sac should be first sign of definitive intrauterine pregnancy for emergency physicians because the decidual reaction and gestational sac are not 100 % accurate.

112.5 Complications

- Bleeding
- Retained products of conception

112.6 Pearls and Pitfalls

- Pearls
  - See “Contraindications.” If following an algorithm, a blighted ovum would fall into indeterminate category and obstetrics consult is recommended in the emergency department.
- Pitfalls
  - It is imperative to inform all pregnant patients that obstetrical ultrasonography performed by emergency physicians is limited and not used to detect fetal health and/or anatomy.

Selected Reading

Cosby K, Kendall J, editors. Practical guide to emergency ultrasound. Philadelphia: Lippincott Williams & Wilkins; 2006.

Doubilet PM, Benson CB, Bourne T, Blaivas M. Diagnostic criteria for nonviable pregnancy early in the first trimester. N Engl J Med. 2013;369(15):1443–51.

Ma OJ, Mateer JR, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw Hill Professional; 2007.

Morin L, Van den Hof MC. Ultrasound evaluation of first trimester pregnancy complications. J Obstet Gynaecol Can. 2005;161:581–5.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

Shah K, Mason C, editors. Essential emergency procedures. Philadelphia: Lippincott Williams & Wilkins; 2008.

Ultrasonography for Threatened, Incomplete, or Compete Abortion

113

L. Connor Nickels and Giuliano De Portu

Key Points

Threatened abortion: closed os, no passage of products of conception (POC).
Inevitable abortion: open os, no POC.
Incomplete abortion: open os, passage of POC.
Complete abortion: closed os, empty uterus (all POC passed).
Missed abortion: fetus has died and the uterus has failed to enlarge any further.

113.1 Indications

- Vaginal bleeding in the setting of early intrauterine pregnancy

113.2 Contraindications

- There are no contraindications to ultrasonography.

113.3 Materials and Medications

- Ultrasound machine
- Probes: curvilinear (transabdominal) and endocavitary (transvaginal)
- Gel
- Endocavitary probe covers
- Towels
- Skilled ultrasound operator
- Pelvic setup (speculum and cultures)
- Cardiac monitor and intravenous access if bleeding is significant
- Laboratory tests: serum human chorionic gonadotropin (hCG), hemoglobin, group and Rh, type, and screen

113.4 Procedure

The procedure is the same for all pelvic ultrasounds including transabdominal and transvaginal imaging because this should be performed in a systematic fashion so as to not miss pertinent findings. Therefore, the findings may change, but the examination remains the same.

1. Place the patient in the supine position.
2. Ultrasound machine in obstetrics preset with abdominal probe and gel.
3. Begin scanning the patient in a transabdominal sagittal plane to identify the uterus as it lies in position with the bladder. Scan through the uterus completely in this plane looking for signs of intrauterine pregnancy (gestational sac, yolk sac, fetal pole, fetal cardiac activity).
    - For our purposes as emergency physicians, a yolk sac (at the least) must be seen within the uterus to definitively report an intrauterine pregnancy (Fig. 113.1). The decidual reaction and/or presence of gestational sac is not 100 % accurate.
4. Transabdominal imaging in the transverse plane should be done in the same fashion.
5. If an intrauterine pregnancy is clearly identified on transabdominal scanning, appropriate measurements can be taken to estimate dates:
    - Gestational sac diameter.
    - Yolk sac diameter or crown rump length (CRL).
    - If present, fetal cardiac activity should be recorded by using M mode to obtain a tracing and measure the fetal heart rate (Fig. 113.2).
6. For a complete examination, the adnexa should be scanned in an attempt to identify the ovaries bilaterally.
7. Any abnormalities identified should be noted:
    - Free fluid surrounding the uterus (anterior or posterior cul-de-sac) or ovaries
        - On pelvic imaging or focused assessment with sonography in trauma (FAST) examination

L.C. Nickels, MD, RDMS (✉) • G. De Portu, MD
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; gdeportu@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_113

  - Intrauterine contents not consistent with an intrauterine pregnancy and/or not clearly, centrally visualized within the fundus of the uterus
  - Any masses or contained fluid collections noted outside the uterus
  - Yolk sac or fetal pole with or without fetal cardiac activity seen outside the uterus
8. Transvaginal scanning with the endocavitary probe should be performed if a definitive intrauterine pregnancy is not identified on transabdominal imaging.
9. Transvaginal imaging is performed in the sagittal and transverse planes to identify an intrauterine pregnancy, with bladder preferably empty.
10. Measurements should be taken of any intrauterine findings.
    - Same as #5.
    - CRL >7 mm on transvaginal ultrasound should have fetal cardiac activity.
11. Each adnexa should be scanned in the sagittal and transverse planes.
12. Any abnormalities identified should be noted.
    - Same as #7.
13. The uterus will be empty in the case of a complete abortion (Figs. 113.3 and 113.4).

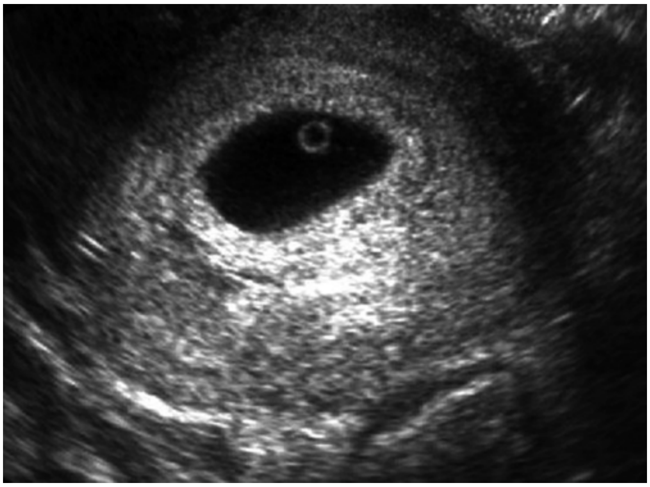

Fig. 113.1 Yolk sac present within the gestational sac with double decidual reaction (With kind permission from Springer Science+Business Media: Buja LM, Chandrasekhar C. Chapter 7: pathology of the breast and female genital tract. In: Krueger GRF, Buja LM, eds. *Atlas of Anatomic Pathology with Imaging*. 2013)

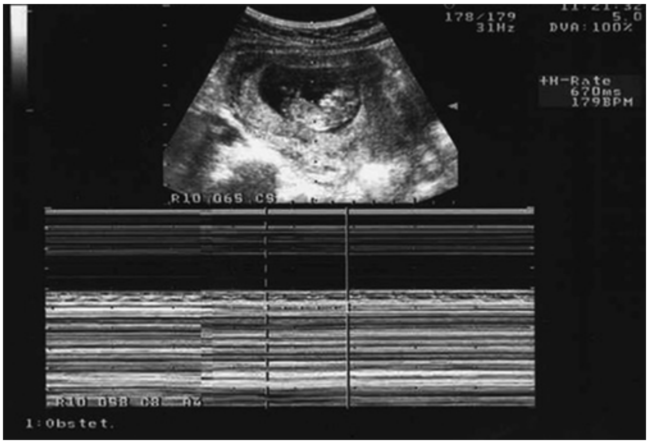

Fig. 113.2 Fetal cardiac activity recorded using M mode to obtain a tracing and measure the fetal heart rate (With kind permission from Springer Science+Business Media: Hanprasertpong T, Phupong V. First trimester embryonic/fetal heart rate in normal pregnant women. *Archives of Gynecology and Obstetrics*. 2006;274(5))

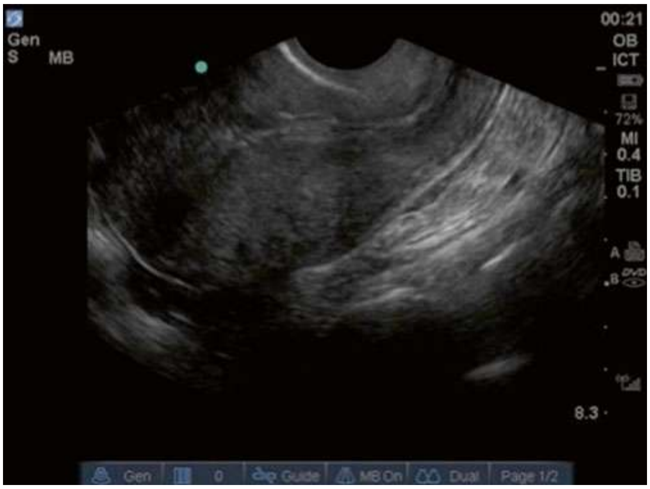

Fig. 113.3 Transvaginal sagittal image of an empty uterus (Photo courtesy of L. Connor Nickels, MD, RDMS)

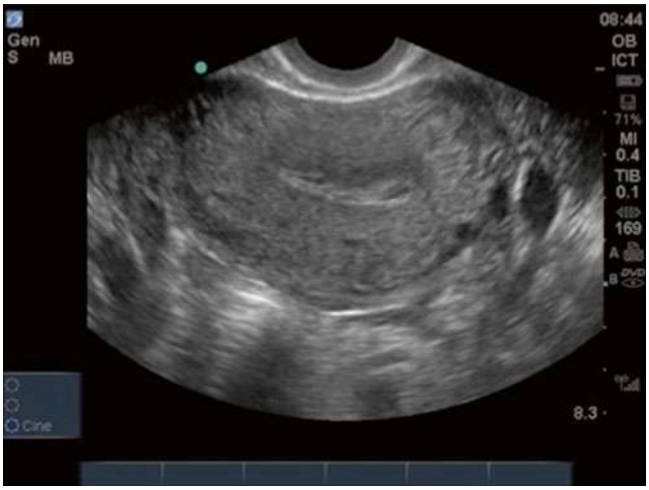

Fig. 113.4 Transvaginal transverse image of an empty uterus (Photo courtesy of L. Connor Nickels, MD, RDMS)

113.5 Complications

- Vaginal bleeding
- Miscarriage
- Rh isoimmunization

113.6 Pearls and Pitfalls

- Pearls
  - Threatened abortion:
    - Threatened abortion is not visible sonographically. It is the presumed diagnosis when a patient presents with vaginal bleeding in the first 20 weeks of pregnancy, an intrauterine pregnancy is found with dates corresponding to the patient's dates, and the cervix is closed.
    - Fetal cardiac activity suggests a better prognosis than those without.
    - Fetal heart rate less than 120 bpm suggests impending fetal death (only 6 % survival).
  - Completed abortion:
    - There are only three scenarios in which this diagnosis can confidently be made in the emergency department:
      1. Intact gestation is passed and identified in the emergency room.
      2. Ultrasound shows an empty uterus in the setting of a prior known intrauterine pregnancy.
         - Small internal echoes within the uterus may represent blood rather than retained products of conception, but this determination should be made in consultation with obstetrics.
      3. Negative pregnancy test result in the setting of a prior known intrauterine pregnancy.
    - In all other cases, the quantitative β-hCG must be followed until it is less than 2 mIU/mL.
- Pitfalls
  - Treating threatened abortion as fetal demise or abortion – many of these pregnancies actually go on to completion with a normal fetus.
  - If there is any concern for alternative diagnoses in any of the abortion types (i.e., ectopic gestation), obstetrics consult should be obtained.
  - Failing to order a quantitative serum hCG to rule out other possibilities (e.g., molar pregnancy) and to ensure that it is trending downward (if others to compare).
  - Failing to establish follow-up for an incomplete abortion to ensure appropriate management. The patient may eventually require a dilation and curettage if products do not pass naturally.

Selected Reading

Doubilet PM, Benson CB, Bourne T, Blaivas M. Diagnostic criteria for nonviable pregnancy early in the first trimester. N Engl J Med. 2013;369(15):1443–51.

Gabbe SG, Niebyl JR, Galan HL, et al., editors. Obstetrics: normal and problem pregnancies. 5th ed. Orlando: Churchill Livingstone; 2007.

Ma OJ, Mateer JR, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw-Hill Professional; 2007.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine. 7th ed. Philadelphia: Mosby; 2010.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

114

Ultrasonography for Placenta Previia

L. Connor Nickels and Giuliano De Portu

Placenta previa is a condition in which the placenta covers the internal cervical os. It is the leading cause of antepartum hemorrhage. There are four grades of the condition:

- Complete: Placenta completely covers the internal os.
- Partial: Placenta partially covers the internal os.
- Marginal : Lower margin of placenta reaches the internal os but does not cover it (within 3 cm).
- Low lying: Lower margin of placenta is located in the lower uterine segment but does not reach the internal os (Fig. 114.1).

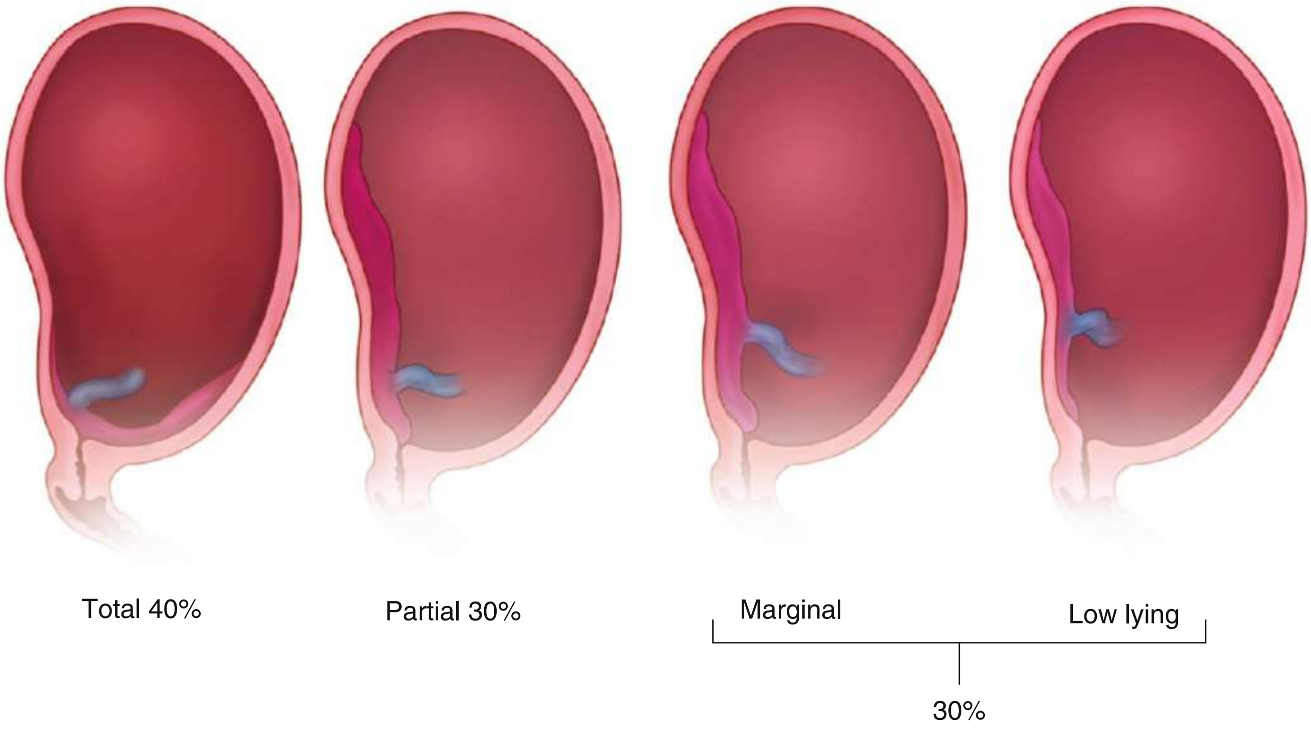

Fig. 114.1 Varying degrees of placenta previa

L.C. Nickels, MD, RDMS (✉) • G. De Portu, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; gdeportu@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_114

114.1 Indications

- A patient in the second half of pregnancy who presents to the emergency room with bright red, painless vaginal bleeding OR premature labor should be evaluated for this condition.

114.2 Contraindications

- Absolute:
  - There are no absolute contraindications to transabdominal ultrasonography.
- Relative:
  - Transvaginal ultrasound should be carried out with caution, so as to avoid hemorrhage.

114.3 Materials and Medications

- Ultrasound machine
- Probes: curvilinear or phased array, endocavitary
- Sterile gel
- Endocavitary probe covers
- Towels
- Sterile speculum (see previously)
- Cardiac monitor, two large-bore intravenous needles, laboratory tests

114.4 Procedure

1. Place the patient in the supine position.
2. Ultrasound machine with abdominal probe and gel.
3. Begin scanning the patient in a transabdominal sagittal position (Fig. 114.2) to determine the placental position and whether it is lying in the lower uterine segment.
4. The patient should then be scanned in a transverse fashion to further evaluate the exact position of the placenta.
5. If the placenta appears to lie in the lower uterine segment, the patient should be scanned in the oblique plane as well.
6. The bladder should initially be full for best visualization. However, if the placenta appears to reside low or lie over the internal os, the scanning should be repeated after the patient has voided.
   - An overdistended bladder may create the appearance of a placenta previa. The anterior wall of the uterus is compressed against the posterior wall by the distended bladder, shortening the distance between the placenta and the internal os (Fig. 114.3).
7. The following ultrasonographic findings exclude placenta previa [1]:
   - Direct apposition of the presenting part of the fetus and the cervix without space for interposed tissue
   - Presence of amniotic fluid between the presenting part of the fetus and the cervix, without the presence of placental tissue
   - Distance of greater than 2 cm between the inferior aspect of the placenta
   - Indirect visualization of the internal cervical os
8. If placenta previa cannot be ruled out with transabdominal ultrasound, the patient should then be scanned transvaginally with the endocavitary probe because this is more sensitive for diagnosing placenta previa.
9. A sterile speculum examination should be performed before transvaginal scanning to assess the cervix and ensure there are no presenting parts or bulging membranes. Transvaginal imaging is contraindicated if the patient has ruptured or bulging membranes.
10. The probes should be swapped out and a sterile cover placed on the endocavitary probe with gel inside the cover and sterile gel on the outside of the cover.
11. The endocavitary probe should be inserted into the vaginal canal, ensuring that caution is taken to stay off the cervix and distal to it, keeping the cervix in view on the screen.
12. The patient should be scanned in both the sagittal and the transverse planes to assess the inferior margin of the placenta.

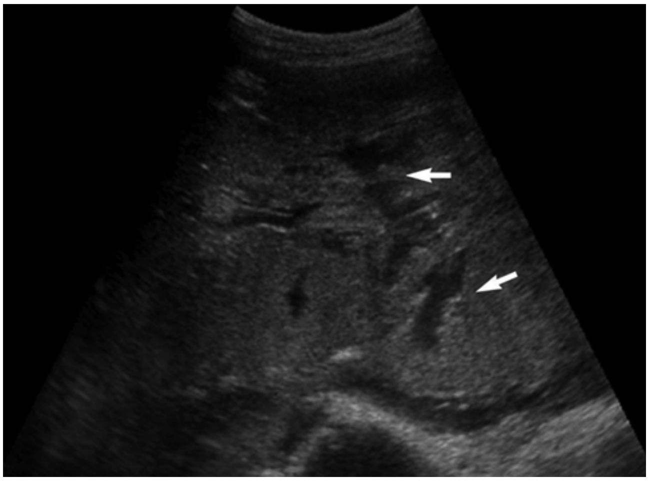

Fig. 114.2 Transabdominal ultrasound of placenta previa. Note loss of the decidual interface between the placenta and the myometrium on the lower part of the uterus and multiple intraplacental lacunae (*arrows*) (Reproduced with permission from *J Korean Med Sci*. 2010; 25(4):651–5)

13. If the inferior margin appears to be located near the internal os, then the distance should be measured.
    - Steps 8 through 13 should only be performed by the emergency physician if he or she feels confident in this ultrasound skill and there is no obstetrics available. If obstetrics is available, these steps should be performed in conjunction with them or by them to ensure the best outcome for the patient.

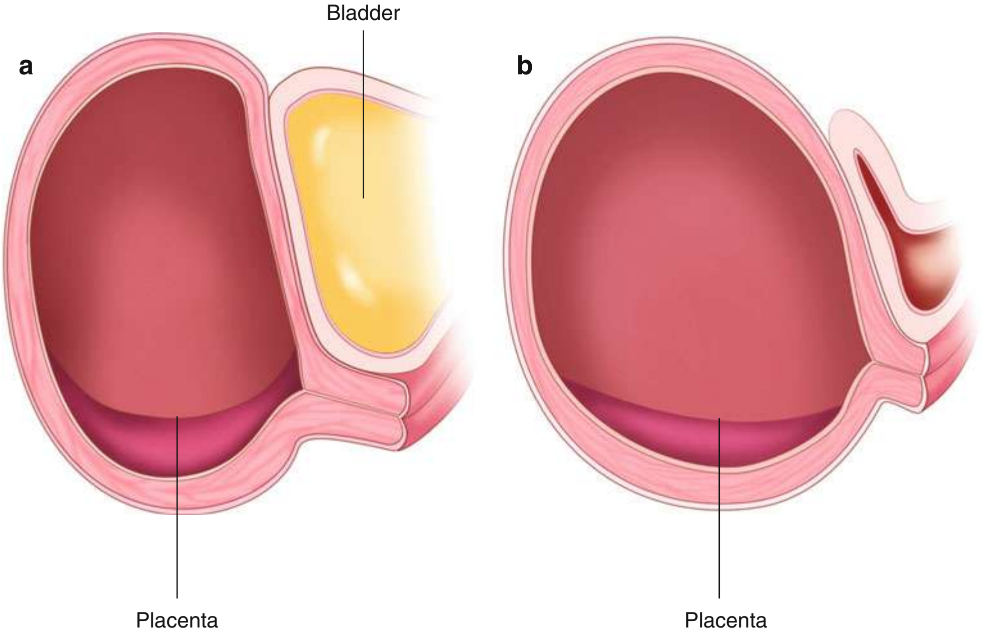

Fig. 114.3 (**a**) Overdistended bladder creating the appearance of placenta previa, (**b**) empty bladder showing the more accurate position of the placenta

114.5 Complications

- Bleeding:
  - May range from self-limited to life-threatening hemorrhage
- Maternal and/or fetal distress or death

114.6 Pearls and Pitfalls

- Pearls
  - Risk factors include:
    - Multiparity
    - Multiple pregnancy
    - Advanced maternal age
    - Previous placenta previa
    - Cigarette smoking
    - Cocaine abuse
    - Hypertension
    - Previous cesarean delivery/uterine surgeries
  - Myometrium contraction: can mimic placenta previa by temporarily displacing the placenta in the lower uterine segment.
  - If ultrasound capabilities are not available, but the patient is in the second half of pregnancy and having vaginal bleeding, do not perform a digital cervical examination.
- Pitfalls
  - Digital examination should be avoided because this may precipitate life-threatening hemorrhage and/or death.
  - Always consult obstetrics as soon as possible if this is suspected or known and the patient is symptomatic. Not consulting obstetrics could be detrimental to the mother and fetus.
  - Gentle sterile speculum examination should be done only if an obstetrician is not available. This is to ensure the bleeding is coming from the cervix. If placenta previa is suspected or known and obstetrics are available, then abdominal ultrasound evaluation alone should be sufficient for the examination of the patient.

Reference

1. American College of Radiology. Role of imaging in second and third trimester bleeding. In: ACR Appropriateness Criteri. Reston: American College of Radiology; 2001.

Selected Reading

Gabbe SG, Niebyl JR, Galan HL, et al., editors. Obstetrics: normal and problem pregnancies. 5th ed. Orlando: Churchill Livingstone; 2007.

Ma OJ, Mateer J, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw Hill Professional; 2007.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

Vaginal Delivery

115

Umarfarook Javed Mirza, Christopher Shields, and Muhammad Waseem

115.1 Indications

- Inevitable delivery of the fetus (cervix fully dilated at 10 cm and fully effaced 1 mm with crowning of the head)

115.2 Contraindications

- Absolute:
  - Indications for emergent cesarean section (C-section)
  - Prolapsed cord
  - Prior C-section with classic vertical incision
  - Placenta previa – complete/partial
  - Breech presentation – footling
- Relative:
  - Placenta previa – marginal/low lying
  - Breech presentation – complete/incomplete/frank

115.3 Materials and Medications (Fig. 115.1)

- 4 Crile clamps or Kelly clamps
- 1 Mayo scissor curved and 1 Mayo scissor straight
- 2 sponge forceps
- 2 towel clamps
- 1 Mayo-Hegar needle holder
- 1 mouse-tooth forceps
- Suction bulb
- Betadine (povidone-iodine)
- Umbilical cord clamp (Fig. 115.2)
- Incubator warmer (Fig. 115.3)
- Most importantly, help.

U.J. Mirza, DO
Department of Emergency Medicine,
Baylor University Medical Center, Dallas, TX, USA
e-mail: Umarfarook.Mirza@gmail.com

C. Shields, MD • M. Waseem, MD (✉)
Department of Emergency Medicine,
Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: cpatrickshields@hotmail.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_115

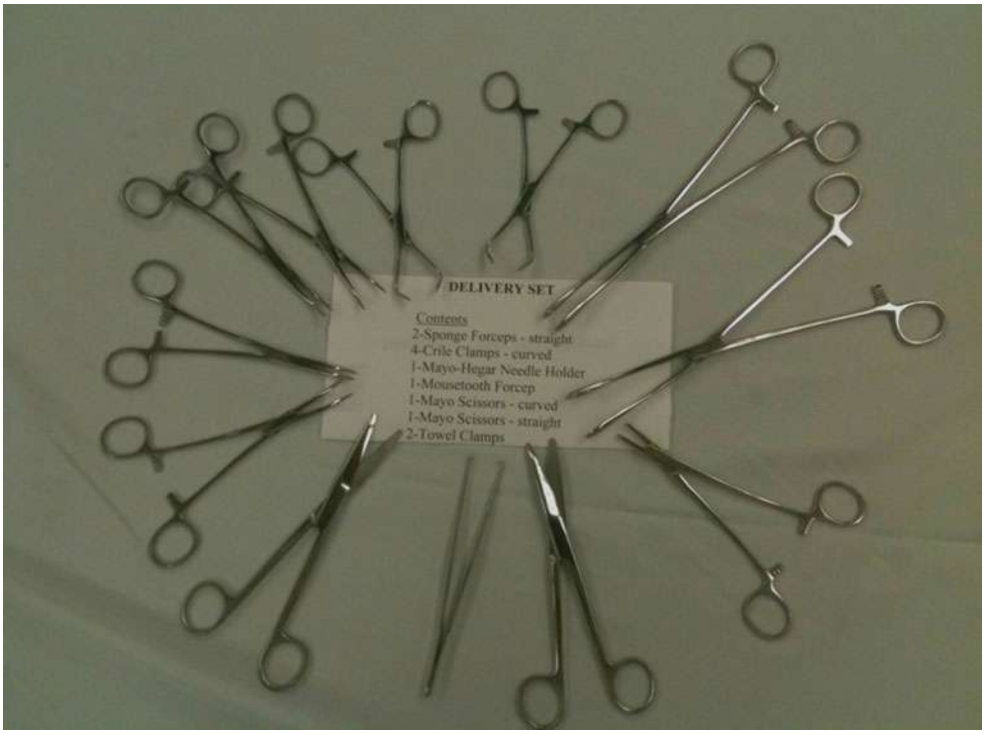
DELIVERY SET
Contents
2-Sponge Forceps - straight
4-Crile Clamps - curved
1-Mayo-Hegar Needle Holder
1-Mousetooth Forcep
1-Mayo Scissors - curved
1-Mayo Scissors - straight
2-Towel Clamps

Fig. 115.1 Equipment: delivery set

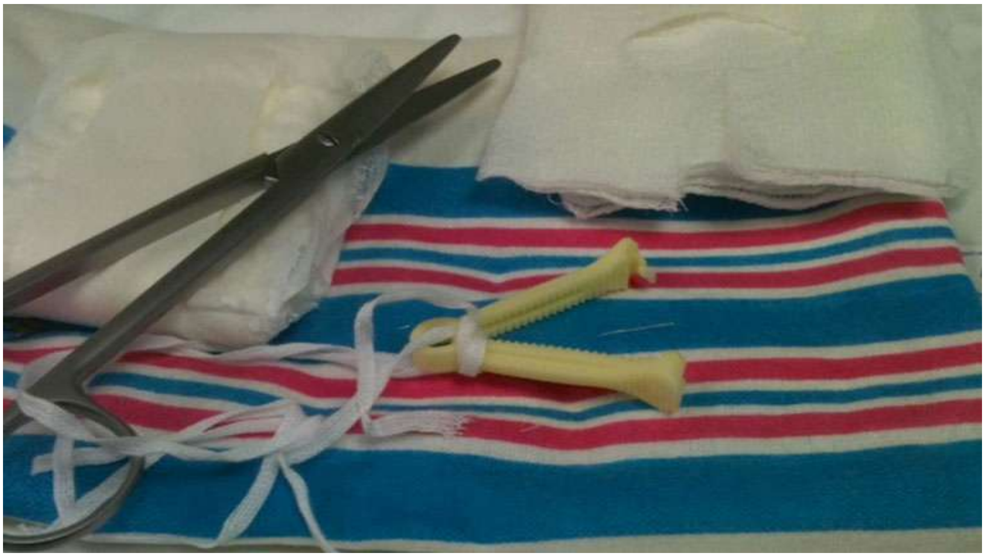

Fig. 115.2 Equipment: umbilical clamp

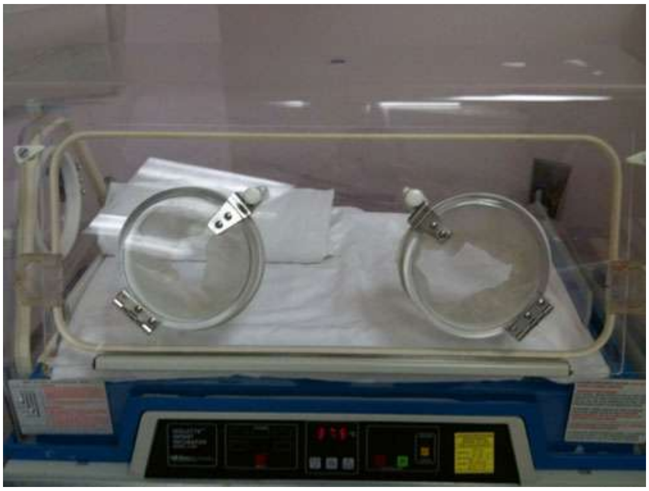

Fig. 115.3 Equipment: incubator warmer

115.4 Procedure

1. Determine the fetal presentation (breech or cephalic) with Leopold's maneuvers or bedside ultrasound.
2. If breech, follow the breech pathway.
3. Perform a vaginal examination.
4. Check for cord prolapse.
5. Check effacement.
6. Check dilation.
7. Prepare the site.
8. Position the mother in the extreme lithotomy position.
9. Deliver the head in a controlled manner.
10. Check to see if the cord is wrapped around the neck. If present, follow cord presentation pathway.
11. Deliver the anterior shoulder with downward traction of the head.
12. Check for shoulder dystocia. If present, follow the McRobert-Rubin maneuvers (see Chap. 116).
13. Deliver the posterior shoulder by gently pulling the trunk upward.
14. Suction the airway.
15. Clamp the cord 5 cm from the umbilicus in two places and cut.
16. Obtain Apgar score of the baby. Initiate resuscitation if required.
17. Keep the baby warm in the incubator.
18. Clamp the cord again closer to the vaginal opening.
19. Use this clamp to help deliver the placenta. Use gentle controlled traction on the clamp with one hand while placing the other hand suprapubically to push the uterus upward.
20. If the placenta does not deliver easily, stop and wait a few minutes to allow it to come away from the uterine wall naturally and then try again.
21. Placenta must be fully intact. If not, check for retained intrauterine placental products, and manually remove them; retained placenta may cause postpartum hemorrhage.
    - Massage the fundus externally to prevent uterine atony, but only after all products have been removed.
22. Repair any vaginal or cervical lacerations.

115.5 Complications

- Postpartum hemorrhage due to uterine atony or retained products
- Uterine inversion
- Rectal or urethral injuries
- Shoulder dystocia
- Meconium aspiration

115.6 Pearls and Pitfalls

- Pearls
  - Postpartum hemorrhage has a high incidence during the first hour postpartum.
  - Be aware of possible second delivery.
  - Oxytocin or Methergine (methylergonovine maleate) may be used after delivery of the placenta; both help increase uterine contractions and decrease postpartum hemorrhage.
- Pitfalls
  - Rushing delivery of the head/shoulders can lead to trauma to the mother/baby.
  - Rushing delivery of the placenta, when the placenta is not ready to detach, may cause placental or cord tearing or uterine inversion. Remember: it may take up to 20 min for the placenta to detach from the uterus.

Selected Reading

Desai S, Henderson SO, Mallon WK. Labor and delivery and their complications. In: Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Liao JB, Buhimschi CS, Norwitz ER. Normal labor: mechanism and duration. Obstet Gynecol Clin North Am. 2005;32:145–64.

Norwitz ER, Robinson JN, Repke JT. Labor and delivery. In: Gabbe SG, Niebyl JR, Simpson JL, editors. Obstetrics: normal and problem pregnancies. 4th ed. Philadelphia: Saunders; 2001. p. 353–94.

Shoulder Dystocia Management

116

Irina Fox Brennan and Joseph A. Tyndall

116.1 Indications

- Retracting fetal head ("turtle sign")
- Failure of anterior shoulder delivery following gentle downward traction
- Difficult face-chin delivery
- Failure of shoulders to descend
- Factors that signal the need for cesarean section:
  - Complete/partial placenta previa
  - Cord prolapse
  - Brow/face presentation
  - Non-reassuring fetal heart rate
  - Extensive uterine surgery and previous classic uterine incision

116.2 Contraindications

- Mother with diabetes and fetus larger than 4500 g
- Mother without diabetes and fetus larger than 5000 g
- History of shoulder dystocia with prior birth

116.3 Materials and Medications

- No specialized materials/equipment required.
- Team coordination is of utmost importance.

I.F. Brennan, MD • J.A. Tyndall, MD, MPH (✉)
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: ibrennan@ufl.edu; tyndall@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_116

116.4 Procedure

1. *Accurate documentation*: Designate a team member to document the progress of labor, the position and rotation of the infant's head, the time of delivery for the head and body, the presence of episiotomy, anesthesia requirements, the duration of shoulder dystocia, whether the shoulder is anterior or posterior at the time of delivery, the onset/duration/results of maneuver(s) performed, and Apgar scores.

   The HELPERR mnemonic is commonly employed:

2. *Help recruitment*: Involve anesthesiology, pediatric resuscitation, and obstetric/gynecology colleagues either as part of advance preparations or via activation of protocol during labor.
3. *Episiotomy consideration*: Not required but might be of benefit if more space is needed for rotation maneuvers.
   - Episiotomy alone will not release the impacted shoulder.
4. *Legs positioned for the McRoberts maneuver*: Position the woman's thighs onto her abdomen by abducting and flexing her hips. Have two assistants help her, one on either side. This increases the functional size of the bony pelvis by rotating the pubic symphysis toward the mother's head, hence aiding delivery (Fig. 116.1).

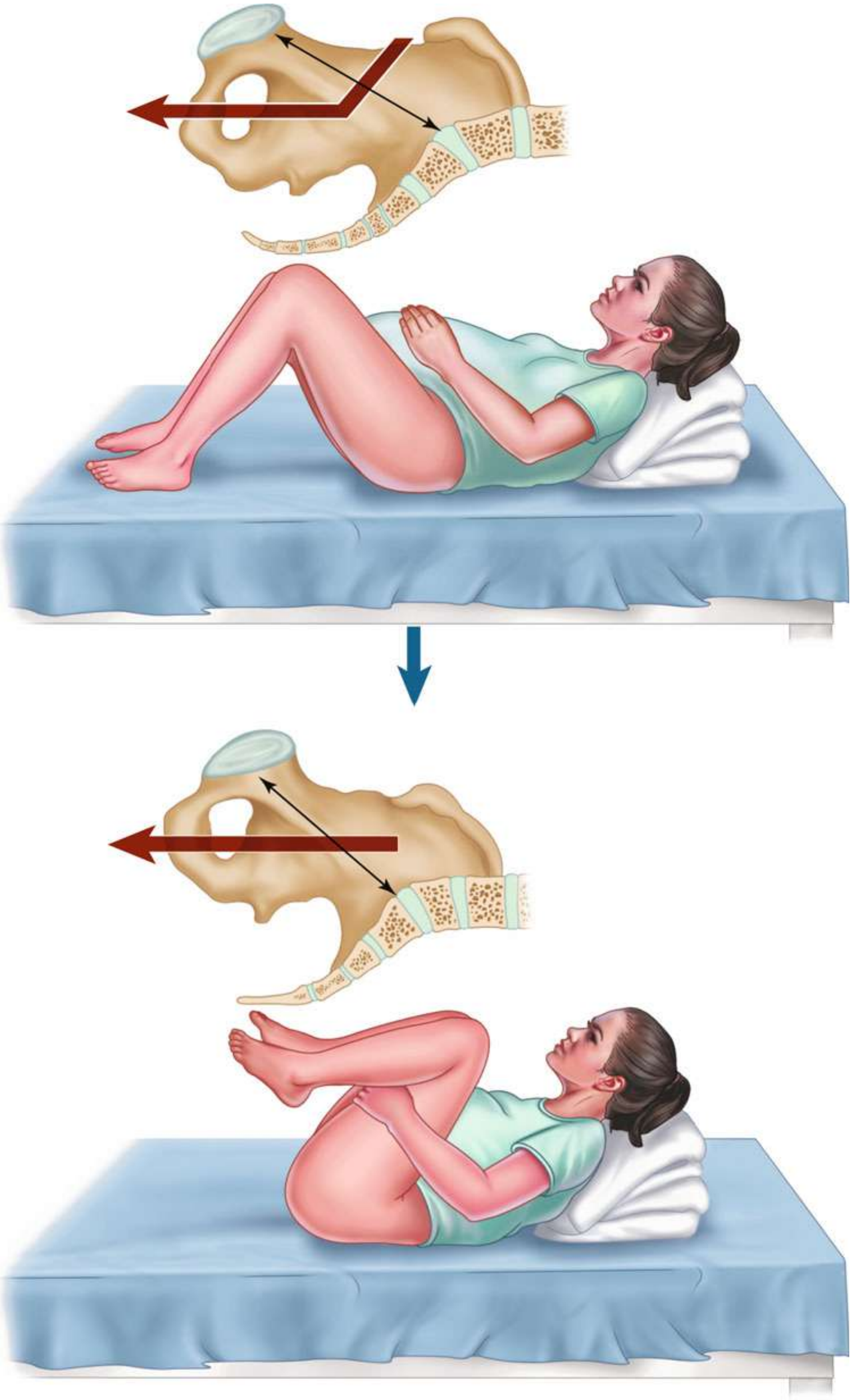

Fig. 116.1 McRoberts maneuver: hyperflexion of the maternal thighs against the abdomen

5. *Pressure suprapubically* (*Rubin I maneuver*): While continuing downward traction, have a colleague place his or her hand suprapubically on the fetal anterior shoulder and apply pressure down and laterally on its posterior aspect in 30-second increments. This allows the fetal shoulders to enter the pelvis in an oblique fashion (pelvic inlet is widest in the transverse plane) (Fig. 116.2).
6. *Enter maneuver* (*internal rotation*): Rotate the anterior shoulder into the oblique plane and under the maternal symphysis.
   - It may be necessary to push the fetus slightly up into the pelvis in order to successfully perform this maneuver.
     (a) *Rubin II maneuver*: Insert one hand vaginally behind the fetal anterior shoulder and rotate the shoulder toward the chest. This will reduce the diameter of the fetal shoulder girdle and facilitate delivery (Fig. 116.3).
     (b) *The Woods corkscrew maneuver*: Place one hand at the front of posterior shoulder and push upward gently. This can be combined with the Rubin II maneuver (Fig. 116.4).
     (c) *The reverse Woods corkscrew maneuver*: Place one hand behind the fetal posterior shoulder and rotate in the direction opposite to that of the Woods corkscrew maneuver to adduct the shoulder.
7. *Remove the posterior arm*: Flex the fetal elbow and deliver the forearm in a sweeping motion over the fetal anterior chest wall. This allows the fetus to drop into the sacral hollow, freeing the anterior shoulder impaction.
   - Avoid grasping and pulling on the fetal arm directly because it may fracture the fetal humerus.
8. *Roll the patient* (*Gaskin maneuver*): Position the patient on all fours (this acts as an upside-down McRoberts maneuver). Continue gentle traction. The turning itself, as well as gravity, will often dislodge the impacted shoulder.

   Last-resort maneuvers:
9. *Deliberate fracture of the fetal clavicle*: Apply direct pressure upward in the middle of the clavicle to reduce the shoulder-to-shoulder distance.

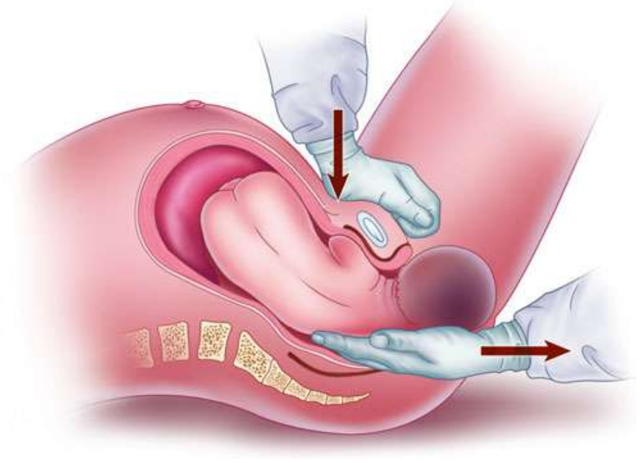

Fig. 116.2 Ruben I maneuver: suprapubic pressure is directed at the anterior shoulder

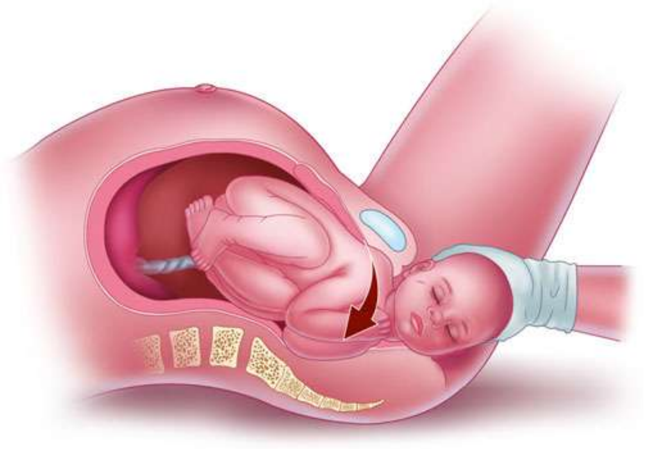

Fig. 116.3 Ruben II maneuver: pressure is applied to the most accessible part of the fetal shoulder and rotated toward the chest

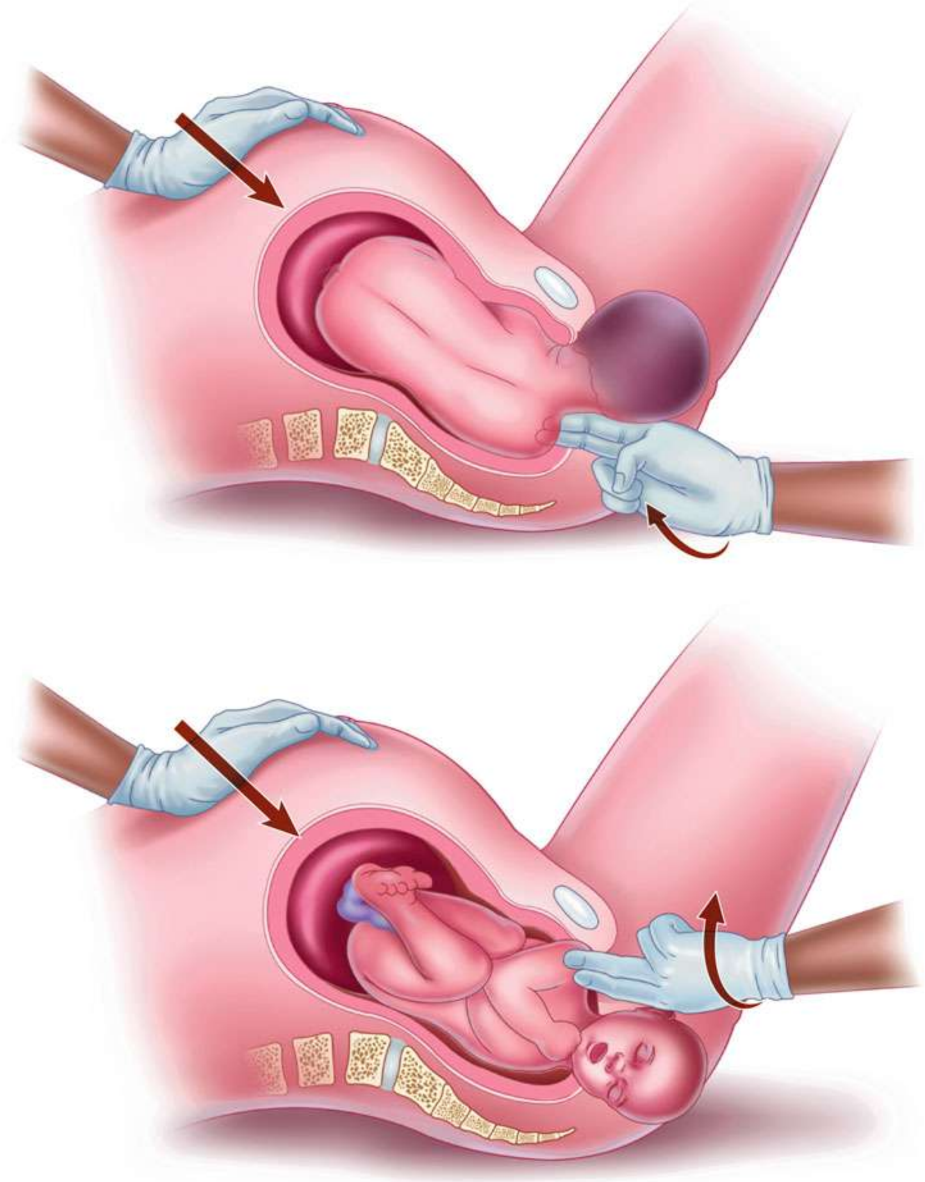

Fig. 116.4 The Woods corkscrew maneuver: pressure is applied to the clavicle of the posterior arm, enabling rotation and dislodgement of the anterior shoulder

10. *Zavanelli maneuver*: Rotate the fetal head into the direct occiput anterior position, flex, and push back into the birth canal. Continue holding upward pressure until cesarean delivery. Ensure the operating and anesthesiology teams are present (Fig. 116.5).
    - Do not perform this maneuver if the nuchal cord has been clamped or cut.
11. *Symphysiotomy*: Place the patient in the exaggerated lithotomy position, insert a transurethral catheter, displace the urethra laterally, and separate the cephalad portion of fibrous cartilage of the symphysis pubis under local anesthesia with either a scalpel blade or Kelly clamp.
    - This technique is associated with significant maternal morbidity, including bladder neck injury and infection. It is truly the last resort and should be used only when all other methods have failed and cesarean delivery is unavailable.

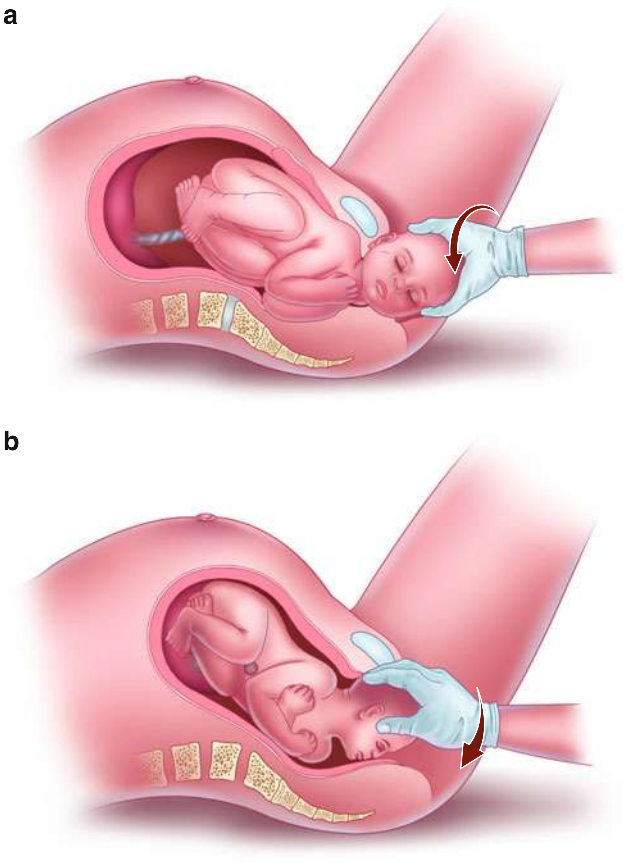

Fig. 116.5 Zavanelli maneuver: the fetal head is rotated into the direct occiput anterior position, flexed, and pushed back into the birth canal. (**a**) Rotation of the head to the occiput anterior position. (**b**) Replace head with constant firm pressure on occiput with palm of hand flexing head

116.5 Complications

- Maternal:
  - Postpartum hemorrhage
  - Fourth-degree laceration/soft tissue damage to cervix and vagina
  - Uterine rupture
  - Bladder atony
  - Symphyseal separation with or without femoral neuropathy (transient)
  - Sacroiliac dislocation
- Fetal:
  - Brachial plexus palsies (Erb, Klumpke): most resolve within 6–12 months; permanent injury occurs in less than 10 % of patients
  - Clavicular and/or humeral fractures
  - Ischemia/hypoxia/asphyxia due to umbilical cord compression with or without neurological damage
  - Significant fetal acidosis, with pH drop of 0.04 U/min between delivery of the head and trunk
  - Death

116.6 Pearls and Pitfalls

- Pearls
  - Many cases of shoulder dystocia and birth trauma are encountered in the absence of risk factors and in non-macrosomic infants. Shoulder dystocia is largely neither predictable nor preventable.
  - Presence of risk factors for shoulder dystocia should prompt advance preparation in anticipation of a difficult delivery. Risk factors include maternal diabetes, obesity, multiparity, advanced maternal age, prolonged pregnancy, macrosomia (fetal weight >4,500–5,000 g), male fetal gender, and prior birth complicated by shoulder dystocia.
- Pitfalls
  - Do not cut and clamp the nuchal cord if at all possible in order to avoid fetal hypoxia and hypotension should shoulder dystocia arise.
  - Labor induction or prophylactic cesarean delivery of macrosomic fetuses has not decreased rates of shoulder dystocia.
  - Avoid fundal pressure in shoulder dystocia because it has been shown to increase the risk of permanent neurological damage and uterine rupture.

Selected Reading

Acker DB. A shoulder dystocia intervention form. Obstet Gynecol. 1991;78:150–1.

Athukorala C, Middleton P, Crowther CA. Intrapartum interventions for preventing shoulder dystocia. Cochrane Database Syst Rev. 2006;4:CD005543.

Baxley EG, Gobbo RW. Shoulder dystocia. Am Fam Physician. 2004;69:1707–14.

Chauhan SP, Gherman R, Hendrix NW, Bingham JM, Hayes E. Shoulder dystocia: comparison of the ACOG practice bulletin with another national guideline. Am J Perinatol. 2010;27:129–36.

Gherman RB, Chauhan S, Ouzounian JG, Lerner H, Gonik B, Goodwin TM. Shoulder dystocia: the unpreventable obstetric emergency with empiric management guidelines. Am J Obstet Gynecol. 2006;195:657–72.

Gottlieb AG, Galan HL. Shoulder dystocia: an update. Obstet Gynecol Clin North Am. 2007;34:501–31.

Hoffman MK, Bailit JL, Branch DW, et al. Consortium on safe labor. A comparison of obstetric maneuvers for the acute management of shoulder dystocia. Obstet Gynecol. 2011;117:1272–8.

Patterson DA, Winslow M, Matus CD. Spontaneous vaginal delivery. Am Fam Physician. 2008;78:336–41.

Sokol RJ, Blackwell SC, American College of Obstetricians and Gynecologists. Committee on Practice Bulletins–Gynecology. ACOG practice bulletin no. 40: shoulder dystocia. November 2002 (replaces practice pattern no. 7, October 1997). Int J Gynaecol Obstet. 2003;80:87–92.

Breech Delivery in the Emergency Department

117

Kristin Stegeman, Sapnalaxmi Amin, Anton A. Wray, and Joseph A. Tyndall

Breech presentation: the buttocks enter the pelvis before the head (Fig. 117.1)

- Frank (extended):
  - Hips flexed and knees extended, buttocks presenting
  - Most common
- Incomplete/complete (flexed):
  - One or both hips and knees flexed, buttocks presenting
- Footling:
  - One or both hips and knees extended, foot presenting

The emergent delivery of a breech baby is one of the most challenging situations for an emergency physician.

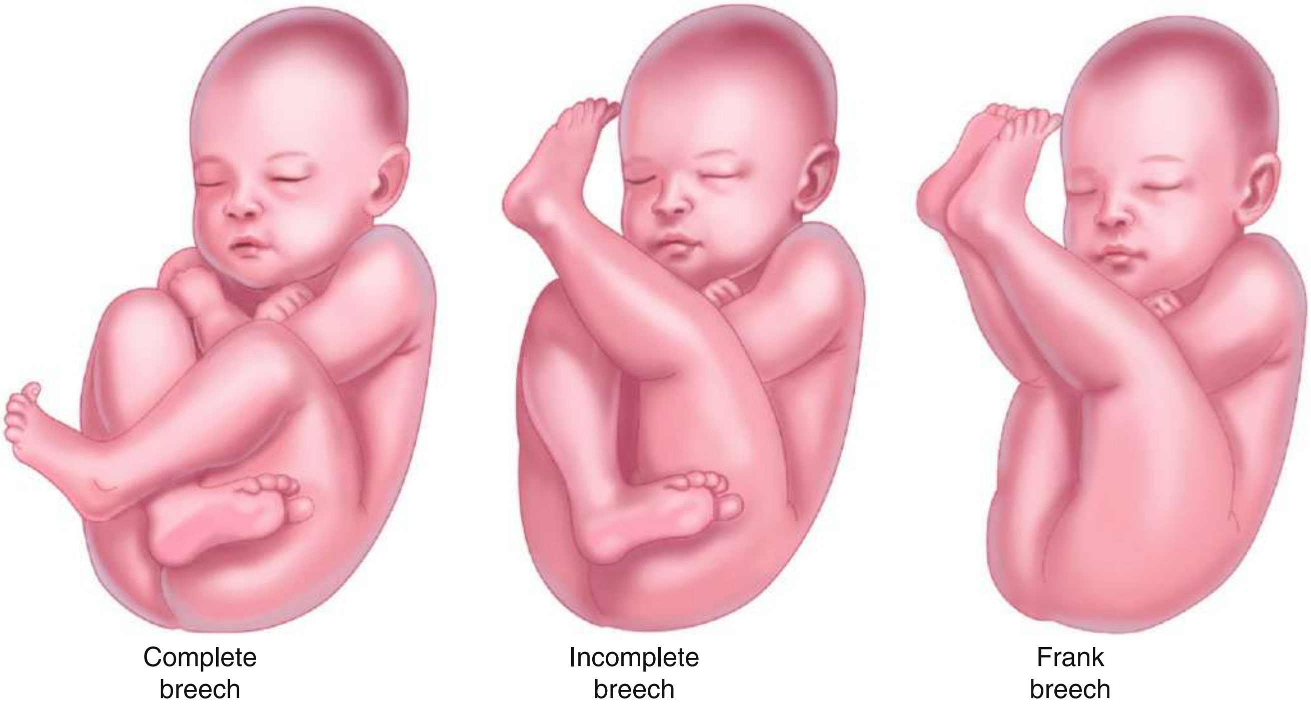

Fig. 117.1 Variations of the breech presentation

K. Stegeman, MD • A.A. Wray, MD (✉)
Department of Emergency Medicine, The Brooklyn Hospital Center, New York, NY, USA
e-mail: kristin.stegeman@gmail.com; anw9071@nyp.org

S. Amin, MD
Department of Family Medicine/Urgent Care, Bayside Urgent Care Center, Clearwater, FL, USA
e-mail: sapna_amin59@yahoo.com

J.A. Tyndall, MD, MPH
Department of Emergency Medicine, University of Florida Health, Gainesville, FL, USA
e-mail: tyndall@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_117

117.1 Indications

- Inevitable delivery of fetus with complete or frank breech presentation
- Absence of vertex presentation

117.2 Contraindications

- Footling presentation (increased risk of cord prolapse and entrapment of the after-coming head).
- There is adequate time to transfer the mother safely to labor and delivery in the knee/chest position (administer subcutaneous terbutaline before transfer).

117.3 Materials and Medications

- Supplemental oxygen for mother
- Ultrasound machine
- Piper forceps
- Sterile towels/gloves
- Betadine (povidone-iodine) or another antiseptic preparing solution
- Sterile lubricant (Surgilube)
- Instruments to cut the umbilical cord/perform episiotomy/resuscitate baby (Kelly clamps/scissors/#10 scalpel/bulb suction)
- Available emergency department staff
- Pediatric, obstetrics, and anesthesia practitioners

117.4 Procedure

1. Assess the health of the mother and the baby (vitals/physical examination/history).
2. Listen to the fetal heart with stethoscope or Doppler ultrasound.
   - Should be 120–160 beats/min.
3. Identify the type of presentation by bedside ultrasound or by digital exam.
4. Perform a sterile digital examination to confirm the position of the baby and the stage of labor.
   - Membranes should not be artificially ruptured. The amniotic sac will help to dilate the cervix, lubricate the canal, and protect the umbilical cord from compression.
5. If footling presentation: await OB/general surgery for emergency C-section.
6. If frank/complete/incomplete: instruct the patient to push when the cervix is completely dilated.
7. When the breech has descended to the perineum, consider performing an episiotomy if more space is needed. Cleanse the perineum with antiseptic and sterile lubricant beforehand.
8. Allow the baby to extrude to the umbilicus with maternal efforts alone. Do not exert traction before this time.
9. If frank: deliver the posterior leg by gently guiding the sacrum anteriorly, grasping the thigh and flexing the leg at the knee.
10. Deliver the anterior leg in a similar manner while guiding the sacrum posteriorly.
11. If incomplete: deliver the extended leg as #8 or #9 as appropriate.
12. Delivery continues as it would for a complete presentation.
13. Wrap the legs/buttocks in a clean towel to decrease trauma (create grip).
14. Grasp the upper legs with the index fingers holding the anterior iliac crests. Place the thumbs on the sacrum (Fig. 117.2).
15. Apply gentle traction as the mother pushes until the scapulae and axillae are visible.
16. If there is difficulty delivering the shoulders, deliver the posterior shoulder by rotating the trunk 90° and applying gentle downward traction to rotate the shoulder anteriorly.
17. Rotate the baby 180° so as to deliver the anterior arm in a similar manner.
18. If the arms do not spontaneously deliver, a finger can be hooked over the shoulders to bring the arm down while rotating the trunk as above (Fig. 117.3).
19. Use the McRoberts position to increase the diameter of the pelvis (Fig. 117.4).

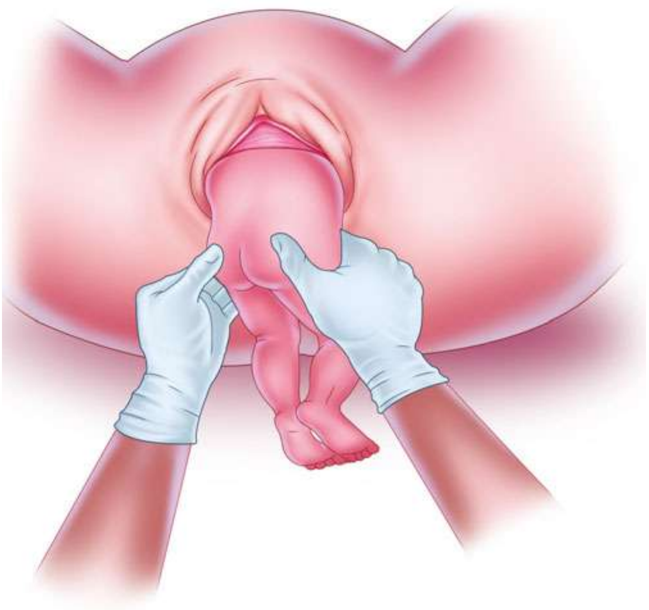

Fig. 117.2 Correct placement of hands on sacrum

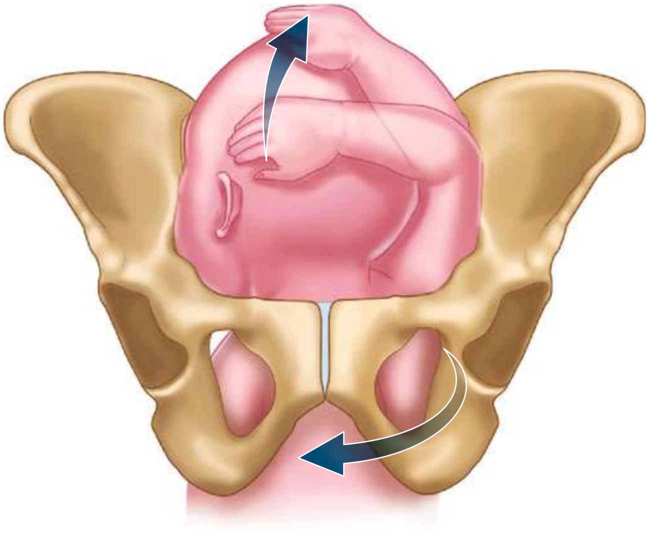

Fig. 117.3 Rotation toward maternal symphysis pubis to avoid nuchal arm

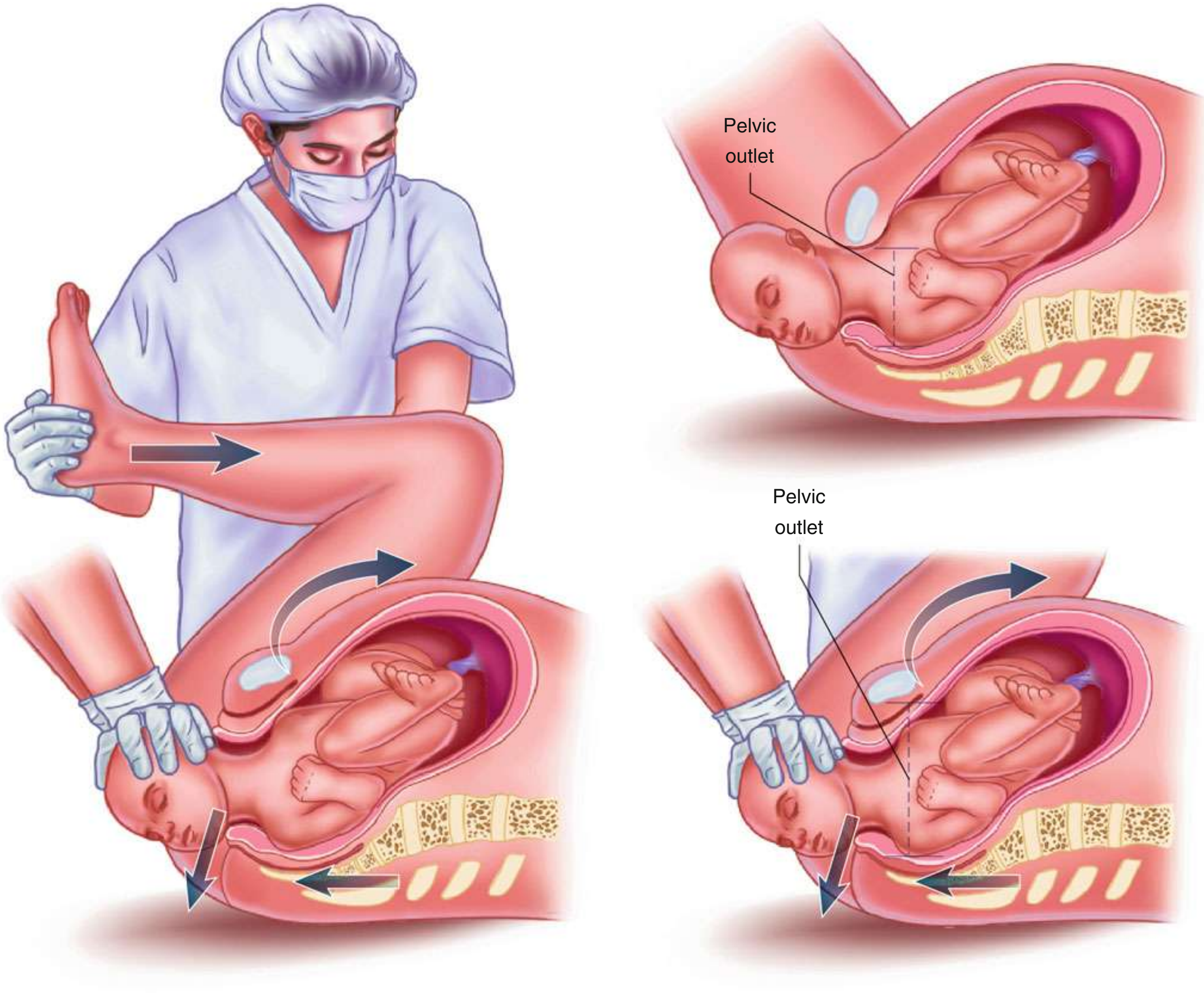
Pelvic outlet
Pelvic outlet

Fig. 117.4 The McRoberts maneuver in vertex (head-first) presentation

20. Maintain the baby in the same plane as the vagina (support the body with the forearm) and place the second and fourth fingers over the maxilla of the baby. Place the middle finger in the mouth or on the chin and the other hand on the upper back/occiput (Fig. 117.5).
    - Avoid extreme elevation of the fetus to prevent hyperextension and cervical spine injury. Piper forceps may also be used to promote flexion.
21. Deliver the head in the flexed position.
    - If the baby descends with the neck and abdomen facing anteriorly: Grasp the shoulders posteriorly with two fingers of one hand while the other hand flexes the abdomen and the baby's feet are brought upward (Fig. 117.6).
    - If the baby's neck remains extended: Leave the baby hanging (weight = traction). When the hairline appears under the symphysis, grab the baby by the feet and elevate upward (Fig. 117.7).
22. Clamp and cut the umbilical cord (collect arterial and venous samples for pH).
23. Suction the baby's mouth and nose, and resuscitate as indicated.
24. Deliver the maternal placenta.
25. Repair tears or episiotomy made during delivery.

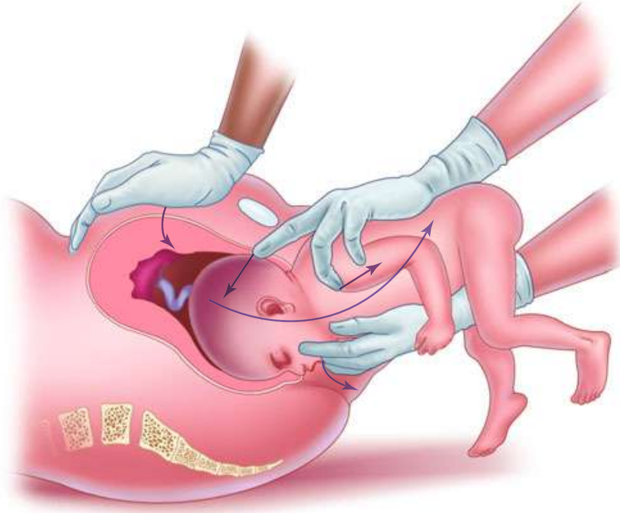
Fig. 117.5 Delivery of head in flexed position

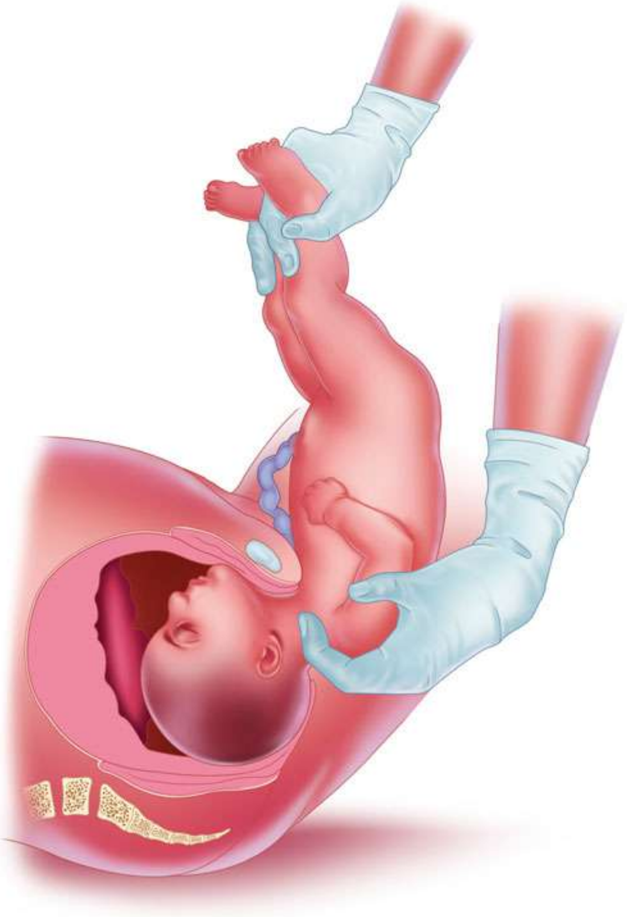
Fig. 117.6 Correct hand placement if abdomen facing anterior

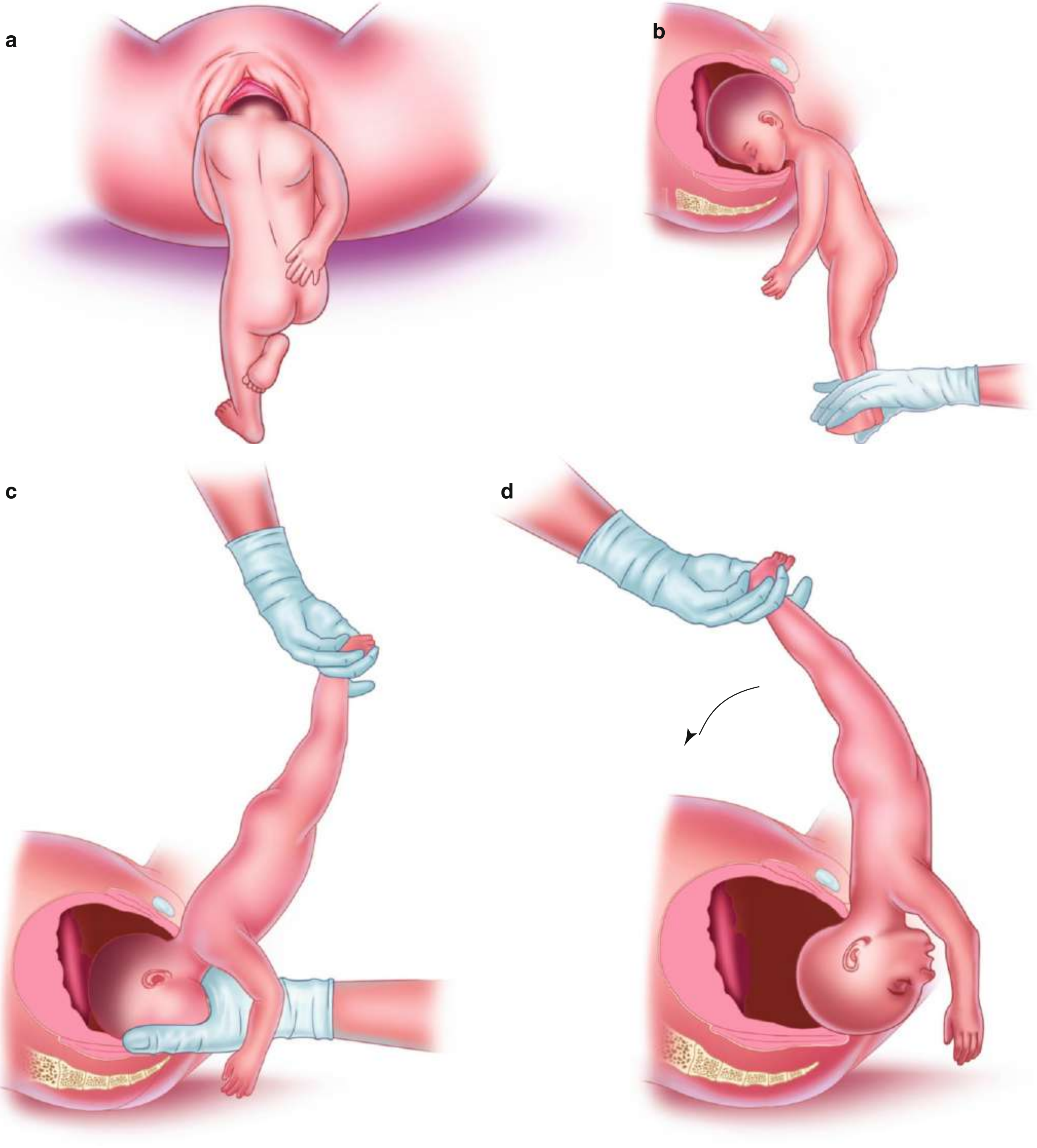

Fig. 117.7 The Burns-Marshall maneuver. (**a**) Allow baby to hang until you can see the hair at the nape of his neck; (**b**) swing the baby's head clear of the birth canal; (**c**) left hand guards and slips the perineum over fetal mouth; suction baby's air passage to clear mucus; (**d**) hold baby's feet

117.5 Complications

- Umbilical cord prolapse
- Brachial plexus injury (from the nuchal arm)
- Fetal head entrapment
- Cervical spine injury (from hyperextension of the neck)
- Birth asphyxia

117.6 Pearls and Pitfalls

- Pearls
  - Allow the uterine contractions to help deliver the baby.
- Pitfalls
  - Do not rush the delivery or use too much force. This can increase the risk of trauma to the baby and mother.
  - Beware of nuchal arm. To avoid brachial plexus injury: Rotate the face of baby toward the maternal symphysis pubis. This will reduce the tension keeping the arm around the back of the fetal head (Fig. 117.3).

Selected Reading

Auerbach PS. Gynecologic and obstetric emergencies. In: Wilderness medicine. 6th ed. Philadelphia: Elsevier; 2012.

Buckley RG, Knoop KJ. Gynecologic and obstetric conditions. In: Knoop KJ, Stack LB, Storrow AB, Thurman RJ, editors. The atlas of emergency medicine. 3rd ed. New York: McGraw Hill; 2010.

Cunningham FG, Leveno KJ, Bloom SL, Hauth JC, Rouse DJ, Spong CY. Breech presentation and delivery. In: Williams obstetrics. 23rd ed. New York: McGraw Hill; 2010.

Kish K, Collea JV. Chapter 21. Malpresentation & cord prolapse. In: DeCherney AH, Nathan L, editors. Current diagnosis & treatment: obstetrics & gynecology. 10th ed. New York: McGraw Hill; 2007.

Kotaska A, Menticoglou S, Gagnon R. Vaginal delivery of breech presentation. Int J Gynecol Obstet. 2009;107:169–76.

Probst BD. Emergency childbirth. In: Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Elsevier; 2010.

Management of Primary Postpartum Hemorrhage

118

Megan Kwasniak, Anton A. Wray, and Joseph A. Tyndall

Postpartum hemorrhage (PPH) is defined as ≥500 ml blood loss within 24 h of vaginal delivery or 1000 ml loss within 24 h of cesarean section. It is the leading cause of maternal mortality worldwide.

118.1 Indications

- Excessive vaginal bleeding with or without pain and/or hemodynamic instability within 24 h of delivery

118.2 Contraindications

- There are no absolute contraindications to the management of PPH.

118.3 Materials and Medications

- Sterile technique
- Good lighting
- Sponge forceps
- Gauze
- Towels
- IV fluids
- Type and screen/crossmatch of blood
- Absorbable suture with curved needle
- Needle holder
- Tooth forceps

118.4 Procedure

- Standard resuscitation measures: Place IV, O_2, monitor.
- Assessment and treatment should occur simultaneously.
- All techniques should be performed under strict sterile conditions.

118.4.1 Uterine Exam

1. Assess by placing the hand on the uterine fundus and checking its size and firmness.
2. High, soft, or boggy uterus implicates retained placenta or uterine atony. Start external fundal massage.
3. Non-palpable uterus implicates uterine inversion.

118.4.2 Vaginal Exam

1. Keeping one hand on the abdomen, gently examine the vaginal canal using the other sterilely gloved hand.
2. Gently scoop out any clots/retained placenta that are easily removable.
3. Look for traumatic sources of bleeding from the perineum, vaginal walls, and cervical lacerations.
4. Gauze-wrapped ring forceps can be used to assist direct visualization and clearing of clots.

M. Kwasniak, MD
Emergency Department, The Brooklyn Hospital Center,
New York, NY, USA
e-mail: megan.kwasniak@gmail.com

A.A. Wray, MD (✉)
Department of Emergency Medicine, The Brooklyn Hospital
Center, New York, NY, USA
e-mail: anw9071@nyp.org

J.A. Tyndall, MD, MPH
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: tyndall@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_118

118.4.3 If Uterine Atony Suspected

1. Continue external fundal massage with the abdominal hand. Make a fist with the vaginal hand and start bimanual massage. Raise the uterus from the pelvis and pivot it anteriorly, compressing it between the external hand and the internal fist. This maneuver will result in expression of any clots present and decrease uterine bleeding via direct compression (Fig. 118.1).
2. Administer oxytocin (20–40 units in 1 L of normal saline or lactated Ringer's solution intravenously [IV]; alternatively give 10 units intramyometrially with a spinal needle).
3. Methylergonovine may also be used at this time if oxytocin fails to reduce uterine bleeding (100 or 125 mcg IV or intramyometrially; alternatively 200 or 250 mcg may be given intramuscularly).
4. Continue massage until bleeding slows and the uterus becomes more firm. This can take 15–30 mins.

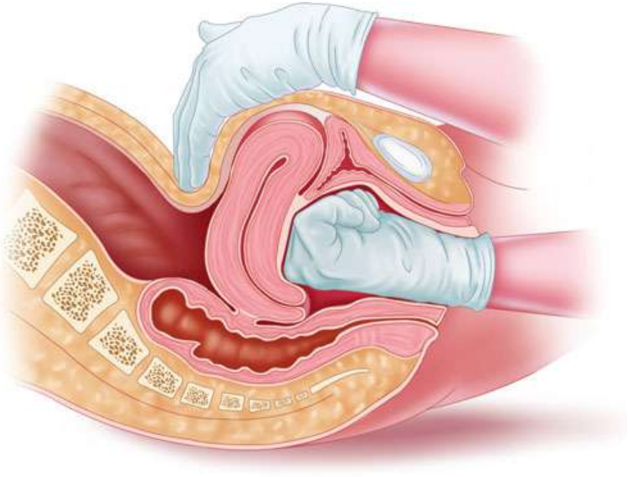

Fig. 118.1 Intrauterine massage for uterine atony

118.4.4 If Retained Placenta Suspected

1. Manual removal of the placenta or any of its retained tissue should be facilitated with sedation or additional analgesia.
2. Keeping the thumb and fingers together in a teardrop shape and using sterile technique as described previously, insert the hand through the vaginal canal and the cervix into the lower uterine segment.
3. Keep the other hand on the lower abdomen to continue gentle yet firm upward pressure and massage.
4. Find the placental edge within the uterus, grasp it gently, detach it from the uterine wall, and withdraw the hand from the patient.
5. Repeat the maneuver as necessary; remove any additional clots from the uterus and continue bimanual massage until it becomes firm and bleeding decreases.
6. If the entire placenta has been removed this way, an assistant should be available to inspect it for completeness and any torn vessels.
7. If the bleeding is particularly severe and placenta accreta is strongly suspected, pack the uterus with gauze and prepare the patient for urgent surgical intervention.

118.4.5 If Uterine Inversion Suspected

1. Using sterile technique, insert a fist through the vaginal canal and push the inverted fundus back through the cervical canal with pressure directed toward the umbilicus.
2. If the uterus has contracted, tocolytics such as IV magnesium sulfate or terbutaline can be used to aid myometrial relaxation. Alternatively, provided the patient is not overly hypotensive, nitroglycerin 50–100 mcg IV may be administered to relax the myometrium and facilitate return to normal uterine position.
3. If manual replacement is ineffective, hydrostatic reduction may be attempted. Warm fluids are run into the upper vagina under high pressure while occluding the introitus.

118.4.6 Trauma: Genital and Perineal Lacerations

1. For significant cervical lacerations: use absorbable sutures with a continuous interlocking stitch technique to close (Fig. 118.2).
2. For vaginal wound repair: place the initial and final stitch above the apices of the lacerations and grab a good amount of tissue with the needle. Small bites can lead to ongoing bleeding and hematoma formation.
3. Observe the repaired lacerations for any additional bleeding after the torn edges have been sutured.
   - Apply additional pressure to any site that continues to ooze blood; gauze-wrapped ring forceps may be used for this purpose if necessary.

If none of the above causes are apparent, consider underlying coagulopathies and treat appropriately. This may require administration of fresh-frozen plasma, platelets, or clotting factors as indicated.

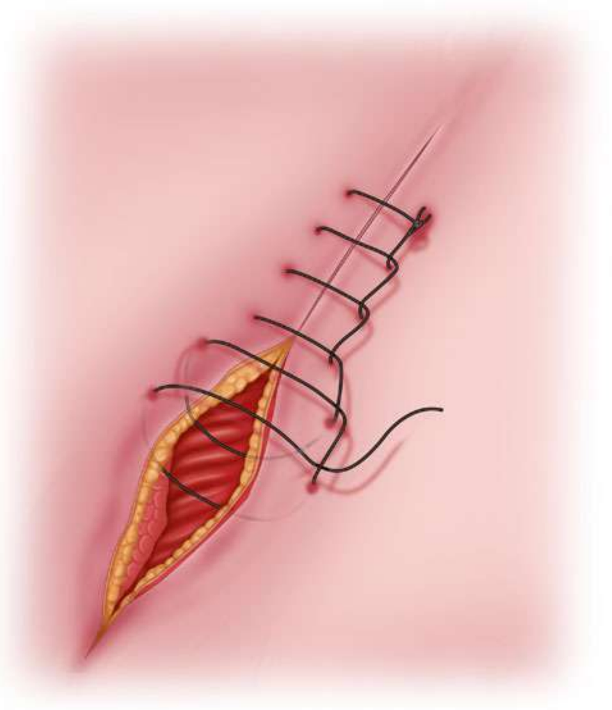

Fig. 118.2 Continuous interlocking stitch: perineal lacerations repair

118.5 Complications

- Uterine perforation and scarring
- Urinary and genital tract trauma and injury
- Genitourinary and genitointestinal fistula
- Pelvic hematoma
- Genital vascular injury
- Infection and sepsis
- Disseminated intravascular coagulation (DIC)
- Maternal death

118.6 Pearls and Pitfalls

- Pearls
  - Causes of PPH can be divided into the “5 Ts”:
    - Tone: uterine atony, occurring within the first 4 h after delivery.
    - *T*issue: retention of the placenta, especially placenta accreta and its fragments, more common at extreme preterm deliveries.
    - *T*rauma: injury to the uterus, cervix, and perineal structures after delivery of a large fetus, use of forceps and/or vacuum, frequent vaginal manipulation during delivery, and episiotomy procedures.
    - *T*hrombosis: intrinsic or acquired coagulation disorders, including idiopathic thrombocytic purpura (ITP); hemolysis, elevated liver enzymes, and low platelets (HELLP) syndrome; and disseminated intravascular coagulation (DIC), as well as preexisting conditions such as von Willebrand disease.
    - *T*raction: inversion of the uterus during placental delivery secondary to excessive traction on the umbilical cord. The uterine fundus can be within the endometrial cavity, in the cervical canal, or outside the external os and within the vaginal canal.
  - The administration of broad-spectrum antibiotics should be strongly considered following any manual removal, exploration, or instrumentation of the uterus and the genital tract.
  - Bedside ultrasonography can be very helpful for identifying uterine abnormalities, retained placental tissue, free fluid in the pelvis, and/or intrauterine hematoma.
  - Risk factors for PPH include prolonged active phase of labor, previous PPH, multiple pregnancy, and history of a bleeding disorder.
- Pitfalls
  - Failure to recognize and treat PPH early increases morbidity and mortality.
  - Underestimating the potential blood loss of PPH. The gravid uterus at term has a blood flow of 600 ml/min (non-gravid: 60 ml/h).

Selected Reading

Anderson J, Duncan E. Prevention and management of postpartum hemorrhage. Am Fam Physician. 2007;75:875–82.

Druelinger L. Postpartum emergencies. Emerg Med Clin North Am. 1994;12:219–37.

Leduc D, Senikas V, Lalonde AB, et al. Active management of the third stage of labour: prevention and treatment of postpartum hemorrhage. J Obstet Gynaecol Can. 2009;31:980–93.

Miller S, Lester F, Hensleigh P. Prevention and treatment of postpartum hemorrhage: new advances for low-resource settings. J Midwifery Womens Health. 2004;49:283–92.

Sheiner E, Sarid L, Levy A, Seidman DS, Hallak M. Obstetric risk factors and outcome of pregnancies complicated with early postpartum hemorrhage: a population-based study. J Matern Fetal Neonatal Med. 2005;18:149–54.

Tessier V, Pierre F. Risk factors of postpartum hemorrhage during labor and clinical and pharmacological prevention. J Gynecol Obstet Biol Reprod (Paris). 2004;33(8 suppl):4S29–56.

World Health Organization. World health report 2005: make every mother and child count. Available at: http://www.who.int/whr/2005/whr2005_en.pdf.

Perimortem Cesarean Section

119

Jordana J. Haber, Elaine B. Josephson,
and Muhammad Waseem

119.1 Indications

- Maternal arrest with a viable fetus (gestation >24 weeks)

119.2 Contraindications

- Stable mother
- Fetus less than 24 weeks' gestation
- Extreme fetal prematurity
- Maternal hypoxia longer than 15 min

119.3 Materials and Medications

- Cesarean section instrument tray if available
  - #10 or #11 scalpel blade, scissors, bladder retractor, 2 large retractors, gauze sponges, hemostats, suction, forceps, and straight and curved clamps
- Skin antiseptic preparing solution, such as Betadine (povidone-iodine)
- Silk suture with needle driver or skin stapler
- Sterile drapes
- Sterile gloves
- Obstetrical pack (See Chap. 115)
  - Bulb syringe and umbilical cord clamp
- Clean blanket or towels for delivery
- Neonatal resuscitation equipment

Owing to the rarity of this procedure in the emergency department, it is unlikely to have a prepared cesarean section tray available. In this case, a thorocotomy or thorocostomy tray combined with an obstetrical pack would contain all the supplies needed. At a minimum, a scalpel and an obstetrical pack are necessary.

J.J. Haber, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: Jordana.haber@gmail.com

E.B. Josephson, MD
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, Weill Cornell Medical College of Cornell
University, Bronx, New York, NY, USA
e-mail: elaine.Josephson@nychhc.org

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_119

119.4 Procedure

1. Prepare skin with antiseptic solution and a sterile drape.
2. Insert a Foley catheter to empty the bladder.
3. Continue cardiopulmonary resuscitation until delivery.
4. Obtain emergent obstetrician and neonatologist consult if available, but do not delay procedure.
5. Using a #10 or #11 blade, make a vertical midline incision beginning 4–5 cm below the xiphoid process and extend the incision to the pubic symphysis (Fig. 119.1).
6. Incise through the subcutaneous fat no further than the rectus sheath.
7. Lift the rectus sheath with a toothed forceps and make an incision with scissors to expose the uterus (Fig. 119.2).
8. With forceps and scissors, lift and incise the peritoneal membrane in the midline.
9. Identify and lift the bowel and cover it with saline-soaked gauze.
10. Retract the rectus sheath and bladder with a bladder retractor or, if not available, use saline-soaked gauze or a towel.
11. Create a 2- to 4-cm midline vertical opening in the uterus.
12. Place a finger in the opening directed caudally to protect the fetus while making a superior incision through the uterine wall. Once complete, repeat this step in the inferior direction.
13. Use a clamp to rupture the amniotic membranes. Immediately deliver the fetus and clamp the umbilical cord.
14. Expulse the head by placing a hand between the pubic symphysis landmark and the fetal occiput. Then, gently flex the fetus while simultaneously moving the head superiorly and anteriorly until delivery (Fig. 119.3).
15. Suction the mouth and nose with a bulb syringe immediately.
16. Deliver the shoulders, followed by the torso and extremities. Secure the umbilical cord with a hemostat or umbilical cord clamp 10 cm distal to the fetus and a second clamp 2 cm distal to this clamp. With scissors, incise the umbilical cord between the two clamps.
17. Immediately begin resuscitation of the infant (Fig. 119.4).
18. If the patient is still alive or regains vital signs, prepare to deliver the placenta. Begin with an oxytocin infusion at 20 U in 1 L at 10 mL/h. Apply cautious traction to the umbilical cord until the placenta separates from the uterus (Fig. 119.5).
19. Following delivery, the uterus should be closed using two layers with either 2-0 or 1-0 suture. In the case of maternal death, skin staples or a running stitch is an acceptable method of skin closure.

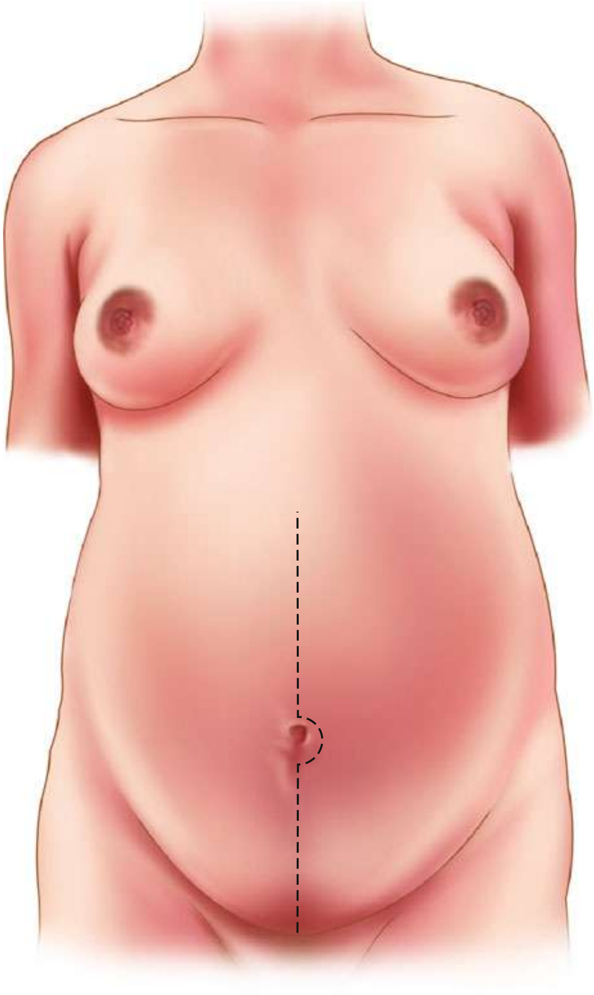

Fig. 119.1 Vertical incision

Fig. 119.2 Exposing rectus sheath

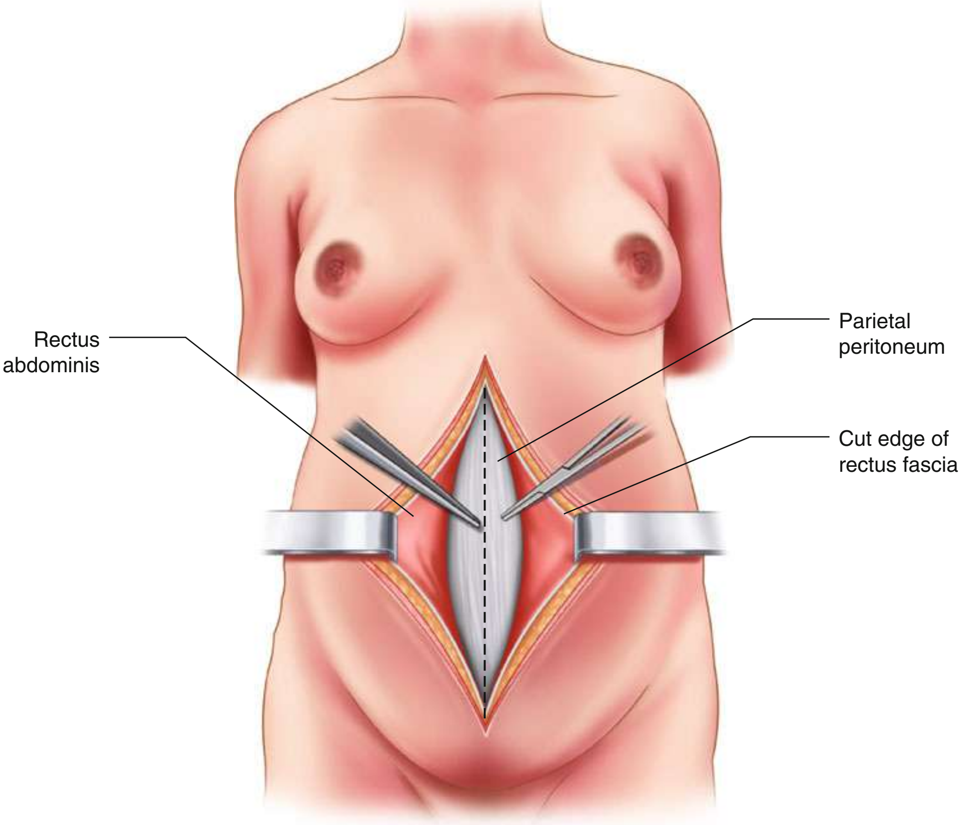
Rectus
abdominis
Parietal
peritoneum
Cut edge of
rectus fascia

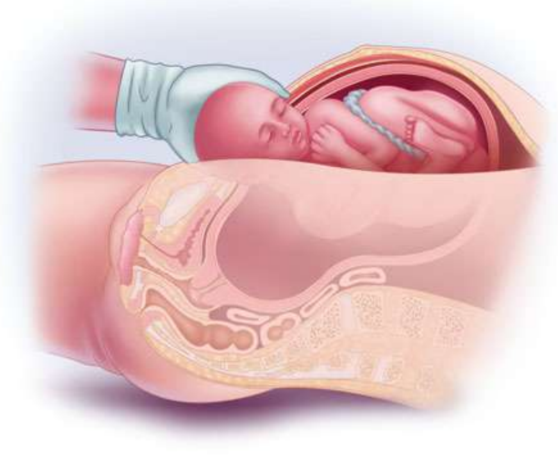

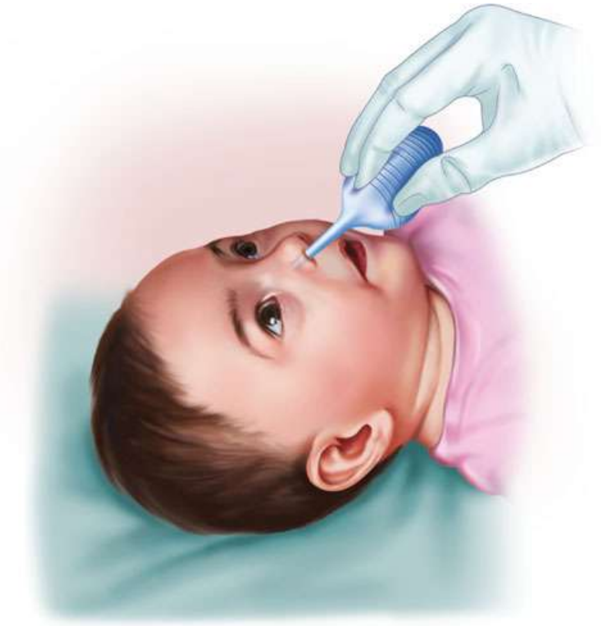

Fig. 119.3 Delivery of the fetus

Fig. 119.4 Suctioning newborn as part of resuscitation

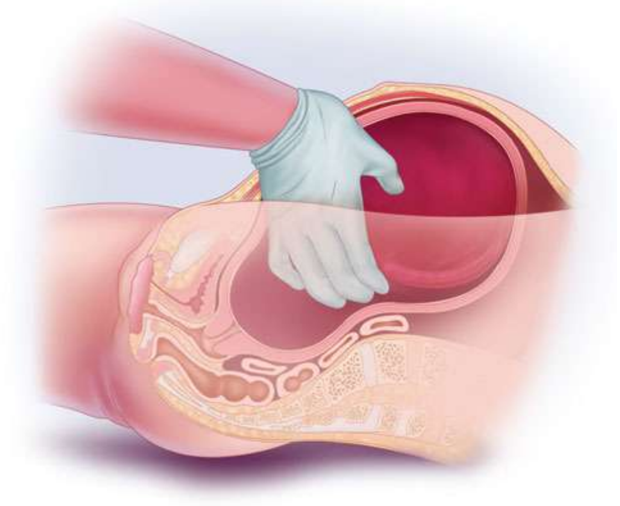

Fig. 119.5 Delivery of the placenta

119.5 Complications

- Maternal sepsis
- Maternal visceral injury
- Maternal hemorrhage
- Maternal death
- Fetal injuries and laceration
- Fetal sepsis

119.6 Pearls and Pitfalls

- Pearls
  - Perimortem cesarean section, although rarely performed, should be considered in any maternal arrest when the fetus is greater than 24 weeks' gestation.
  - In addition to saving the life of the fetus, this procedure may aid in resuscitation of the mother. Emptying of the uterus may improve thoracic compliance and, therefore, improve maternal ventilation.
- Pitfalls
  - The decision to perform an emergency cesarean section must be made early. There is a higher chance of survival if performed no more than 5 min after the onset of maternal cardiac arrest.

Selected Reading

Doan-Wiggins L. Emergency childbirth. In: Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 4th ed. Philadelphia: Saunders; 2004. p. 1117–43.

Flippin A, Hendricks S. Perimortem cesarean section. In: Reichman EF, Simon RS, editors. Emergency medicine procedures. New York: McGraw Hill Medical; 2004. p. 1070–8.

Gianopoulos JG. Emergency complications of labor and delivery. Emerg Med Clin North Am. 1994;12:201–17.

Jeejeebhoy FM, Zelop CM, Windrim R, Carvalho JC, Dorian P, Morrison LJ. Management of cardiac arrest in pregnancy: a systematic review. Resuscitation. 2011;82:801–19.

Whitten M, Irvine LM. Postmortem and perimortem caesarean section: what are the indications? J R Soc Med. 2000;93:6–9.

Part XV

Common Pediatric Procedures

120

Peripheral Venous Catheterization

David N. Smith and Judith K. Lucas

120.1 Indications (See Also Chap. 2)

- Fluid resuscitation
- Medication administration
- Blood draws

120.2 Contraindications

- Relative
  - Avoid catheterizing areas of trauma in which extravasation of fluid is possible (e.g., burns, open wounds, or severe edema in tissue).
  - Avoid catheterizing in an area of local infection for risk of inoculating the circulation with bacteria (e.g., cellulitis).
- Absolute
  - None

120.3 Materials and Medications

- Gloves
- Skin disinfectant (isopropyl alcohol, chlorhexidine, or Betadine [povidone-iodine])
- Appropriate-sized catheter (18- to 24-gauge [IV]) (Fig. 120.1)
  - Large child: 18 to 20 gauge
  - Infant or small child: 22 to 24 gauge
- Tourniquet
- Sterile 2×2 gauze
- Appropriate-sized Tegaderm transparent dressing
- Adhesive tape
- IV bag with solution set (tubing flushed and ready) or saline lock
- Sharps container

D.N. Smith, MD
Department of Pediatrics, University of Alabama at Birmingham,
Children's of Alabama, Birmingham, AL, USA
e-mail: dsmith@peds.uab.edu

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_120

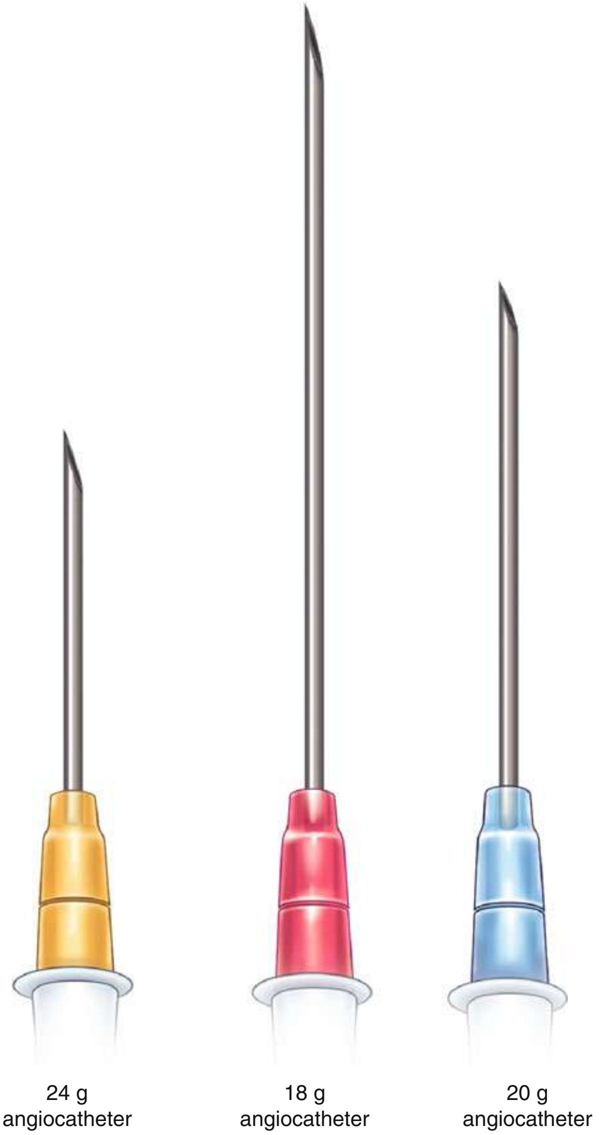

Fig. 120.1 24-Gauge angiocatheter (*yellow*), 18-gauge angiocatheter (*red*), 20-gauge angiocatheter (*blue*)

120.4 Procedure

1. Comfortably position the patient with the site exposed.
2. Assemble the equipment and don a pair of (nonlatex) examination gloves.
3. Apply the tourniquet to the extremity above the site to be catheterized (Fig. 120.2).
4. Visualize and palpate the vein.
5. Cleanse the site with a disinfectant swab using an expanding circular motion.
6. Prepare and inspect the catheter and flush the tubing; be certain that the stylet and catheter separate easily, then fit again into the notch, aligning the bevel with the hub.
7. Stabilize the vein and apply countertension to the skin, being careful not to touch the cleansed area.
8. Insert the stylet through the skin and then reduce the angle while advancing through the vein (Fig. 120.3).
9. Observe for “flashback” as blood slowly fills the flashback chamber.
10. Advance the needle approximately 1–2 mm further into the vein, depending on the gauge and age of the patient, to ensure that the catheter is within the vein.
11. Slowly advance the catheter into the vein while keeping tension on the vein and skin (Fig. 120.4).
12. While advancing the catheter, be certain to hold the stylet portion with the thumb and forefinger, so as to avoid advancing the needle portion into and through the opposite side of the vessel, thus “blowing” the vein.
13. When the catheter is advanced about halfway, slowly withdraw and remove the stylet while simultaneously continuing to advance the catheter to its hub.
14. Attach a 3-mL non-Luer-Lok syringe to the hub.
15. Remove the tourniquet.
16. Gently attempt to aspirate blood. The blood should be free flowing.
17. Secure the catheter by either placing a transparent occlusive dressing (e.g., Tegaderm) over the lower half of the catheter hub or taping over the catheter hub in a cruciate

fashion, taking care not to cover the IV tubing connection (Fig. 120.5).

18. Remove the cover from the end of the IV tubing and insert the IV tubing into the hub of the catheter (the tubing must have been flushed with IV solution before connecting with the catheter hub: the unit from the solution bag/bottle through the catheter must be air free).
19. Open up the IV roller clamp and observe for drips forming in the drip chamber.
20. Place a piece of tape over the catheter hub then make a small (kink-free) loop in the IV tubing and place a second piece of tape over the first piece to secure the loop (Fig. 120.6).
21. Place a third piece of tape over the IV tubing above the site.
22. Ensure that the IV is properly secured and infusing properly.
23. Ensure that all “sharps” are placed in the sharps container.

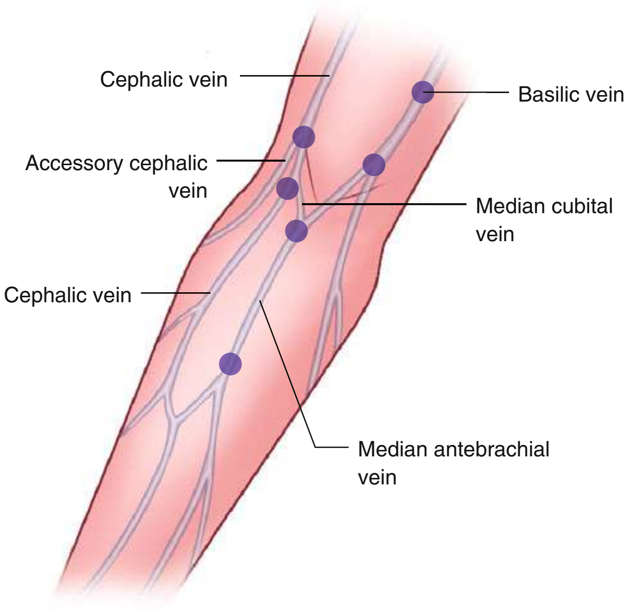
Cephalic vein
Basilic vein
Accessory cephalic vein
Median cubital vein
Cephalic vein
Median antebrachial vein

Fig. 120.2 Anatomy of the volar surface antecubital fossa and forearm

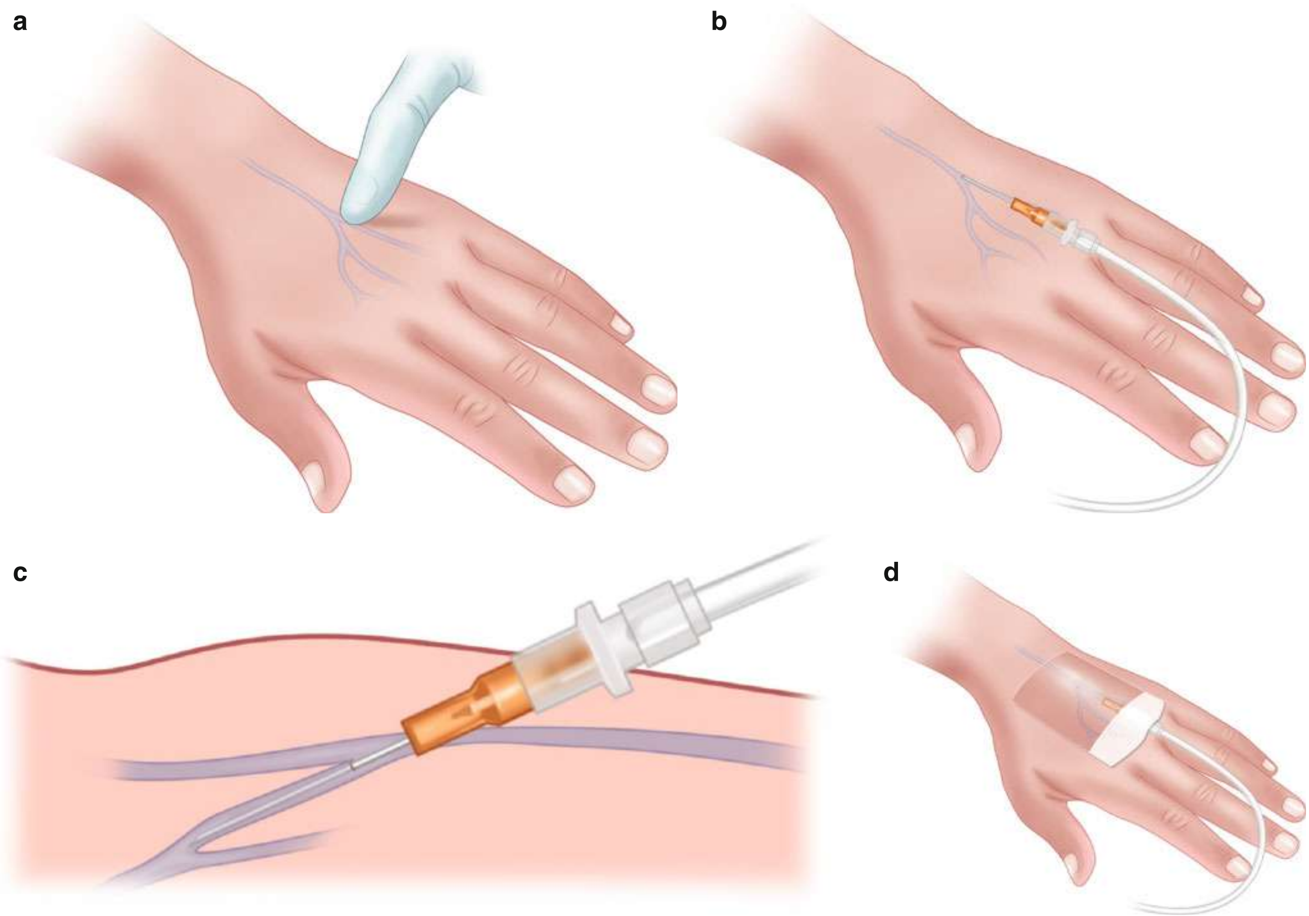

Fig. 120.3 (**a**–**d**) After applying the tourniquet, palpate a vein, as straight as possible, and ideally without many “knots” (i.e., valves)

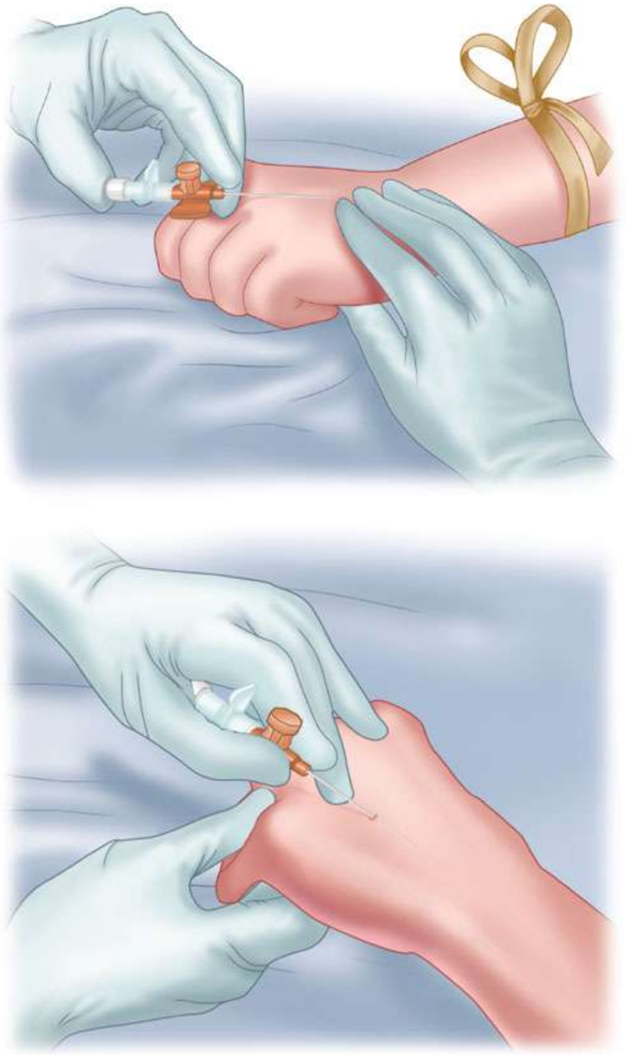

Fig. 120.4 With the tourniquet "up," and the vein distended, apply traction to the skin, pierce the skin, and pass the catheter tip (into the vein) until blood return is noted in the catheter hub

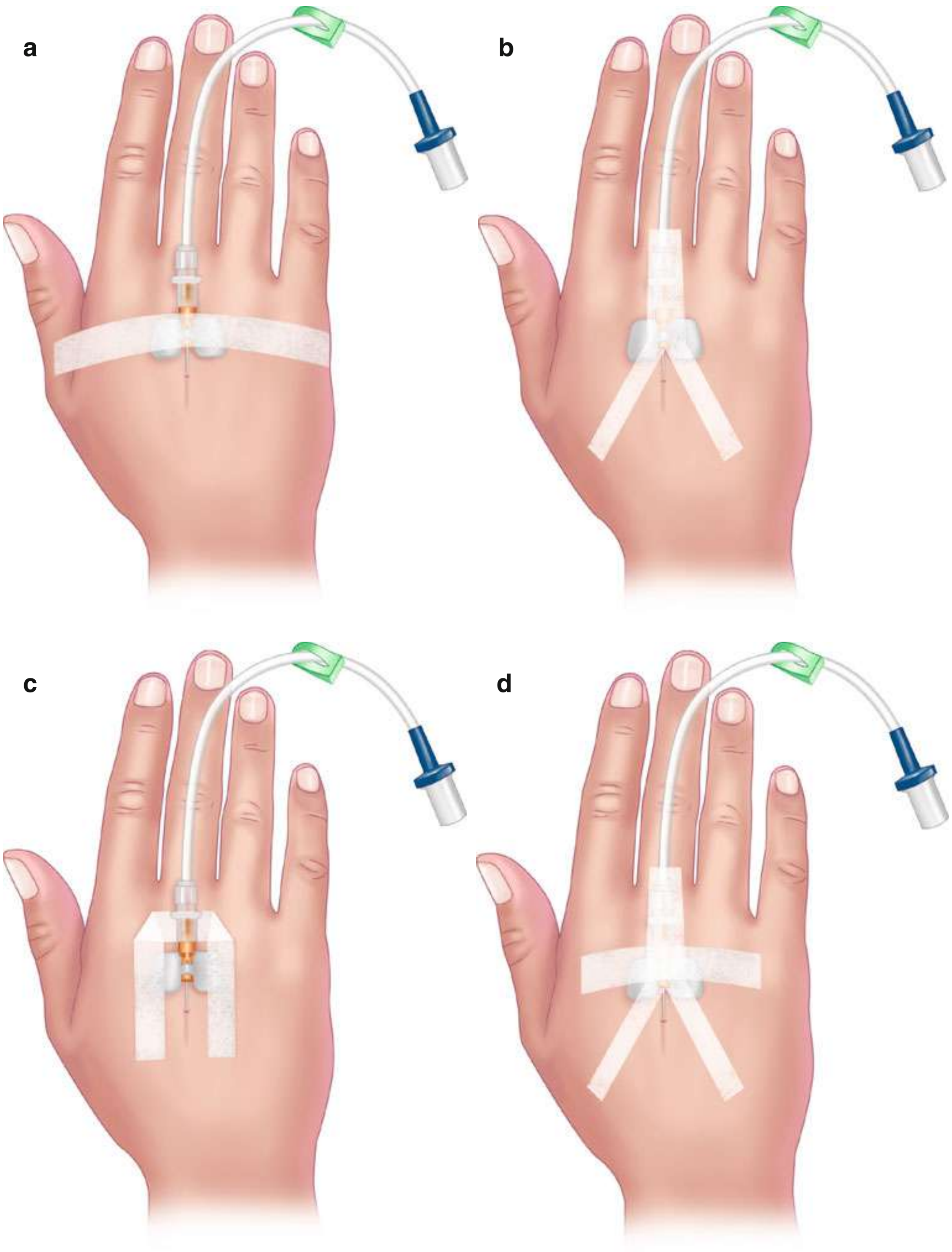

Fig. 120.5 (**a**–**d**) Securing the intravenous line utilizing the cruciate taping style

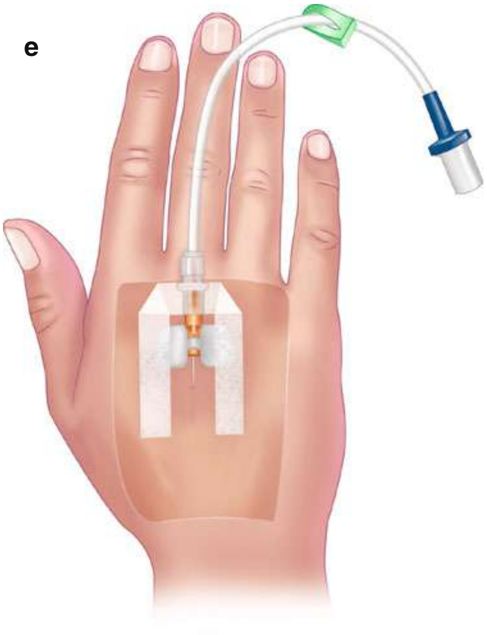

Fig. 120.5 (continued)

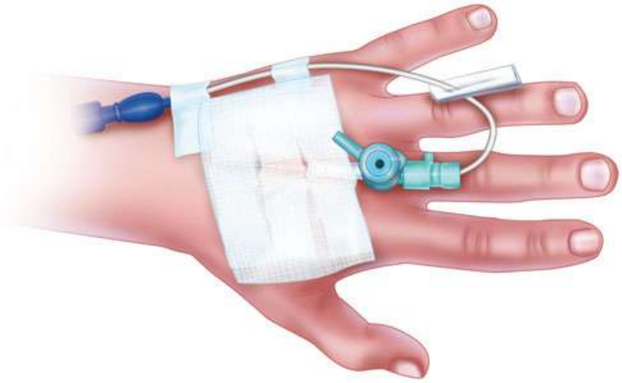

Fig. 120.6 Commercially available hub stabilizer, minimizing the need for excessive tape

120.5 Pearls and Pitfalls

- Pearls
  - Start catheter attempts distal in the extremities and move proximally with each subsequent attempt.
  - The use of ultrasound or a light source in infants can aid in location of the vessel and placement of the line.
  - In an emergent situation, in which fluids or medications are needed quickly, intraosseous access (see Chap. 122) can be obtained if venous catheterization fails.
- Pitfalls
  - The use of lidocaine subcutaneously can improve patient comfort, but it does disrupt anatomical landmarks.

Selected Reading

Bailey P. Vascular (venous) access for pediatric resuscitation and other pediatric emergencies. UpToDate. http://www.uptodate.com/contents/vascular-venous-access-for-pediatric-resuscitation-and-other-pediatric-emergencies. Accessed 23 June 2014.

Department of Emergency Medicine, University of Ottawa. Peripheral intravenous access. 2003. http://www.med.uottawa.ca/procedures/iv/. Accessed 23 June 2014.

Kost S. Ultrasound-guided vascular (venous) access. UpToDate. http://www.uptodate.com/contents/principles-of-ultrasound-guided-venous-access. Accessed 23 June 2014.

Nursing Resource Administration. Medical procedure: insertion of peripheral IV line. http://nursing-resource.com/iv-insertion/. Accessed 23 June 2014.

Umbilical Venous Catheters (Insertion and Removal)

121

Judith K. Lucas

121.1 Indications

- Temporary vascular access for infants up to roughly 10 days of life (between 7 and 14 days) with shock or cardiopulmonary failure
- Emergency vascular access in this age group, when peripheral intravenous (IV) access cannot be rapidly obtained
- Preferred vascular access in infants less than 1000 g

121.2 Contraindications

- Omphalitis
- Omphalocele
- Necrotizing enterocolitis
- Peritonitis

121.3 Materials and Medications

- Anesthetic: not necessary; procedure is painless
- Soft ties to restrain infant's extremities
- Sterile gloves/gowns
- Antiseptic solution
- Sterile towels/drapes
- 3.5- (infants <1500 g) or 5-French (>1500 g) umbilical venous catheter
- 5-French feeding tube
- Three-way stopcock
- 10-mL syringe with heparinized saline flush (1 U/mL)
- Umbilical tape or 3-0 silk on a cutting needle
- Non-toothed forceps
- Small hemostats (2)
- #11 scalpel and blade
- Scissors
- Graph depicting length of catheter insertion, if placing umbilical venous catheter (UVC) above the diaphragm in either very small infants or infants for whom measurement of central venous pressure (CVP) is indicated (Fig. 121.1)

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_121

Fig. 121.1 (**a**) Instrumentation suggested for umbilical venous catheter placement. (**b**) Umbilical venous catheters (in an emergency, 5 F feeding tube is an acceptable alternative)

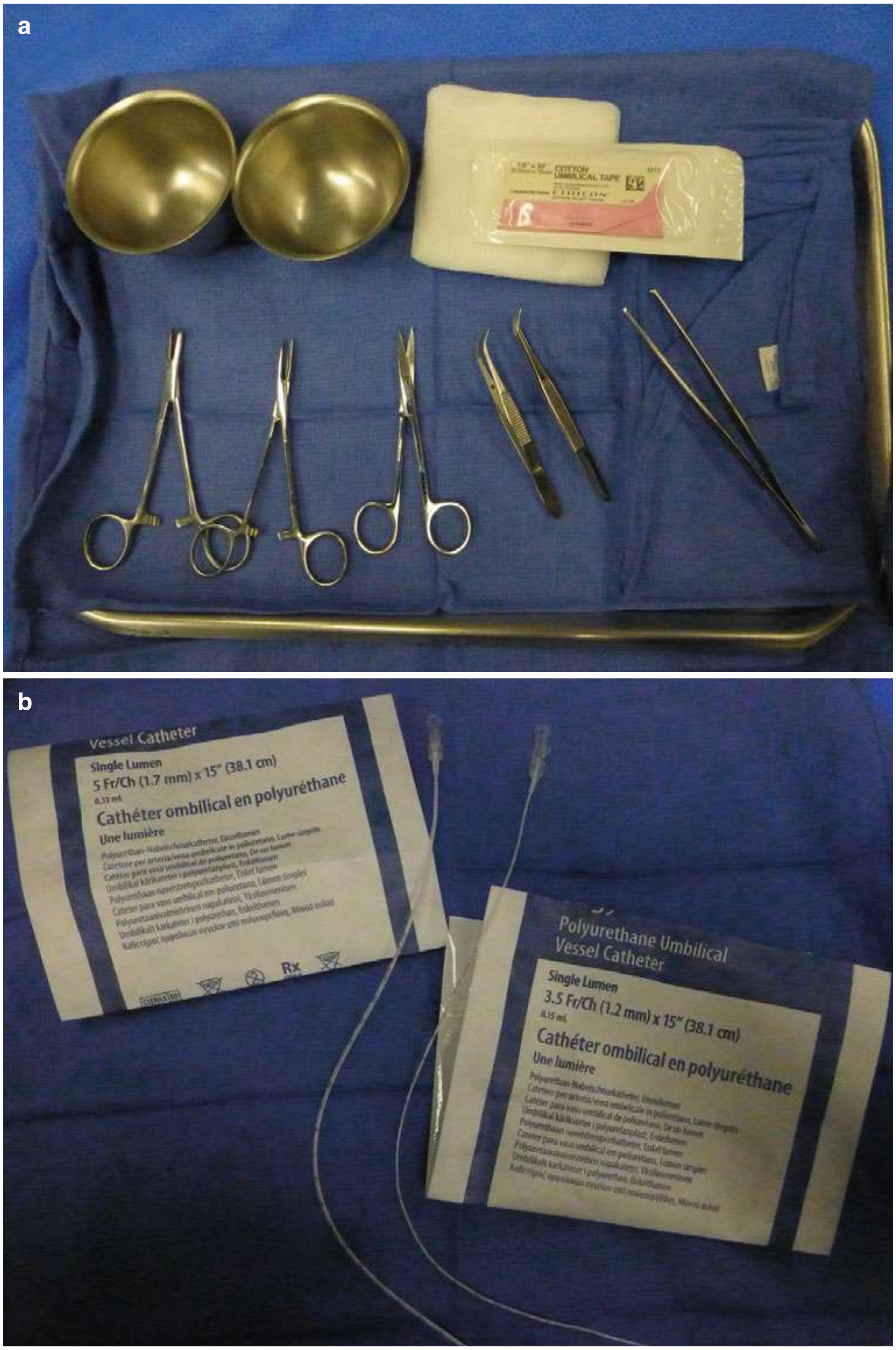
a
COTTON UMBILICAL TAPE
ETHICON
b
Vessel Catheter
Single Lumen
5 Fr/Ch (1.7 mm) x 15" (38.1 cm)
Cathéter ombilical en polyuréthane
Une lumière
Rx
Polyurethane Umbilical Vessel Catheter
Single Lumen
3.5 Fr/Ch (1.2 mm) x 15" (38.1 cm)
Cathéter ombilical en polyuréthane
Une lumière

121.4 Procedure (Insertion)

1. Place the infant under a radiant warmer.
2. Using soft ties, restrain the infant's extremities.
3. Scrub the umbilicus and surrounding abdomen with antiseptic solution.
4. Drape the umbilicus and area in sterile manner (leave the infant's head exposed).
5. Tie a loose loop with the umbilical tape around the base of the cord OR run the 3-0 silk through the skin *of the cord* in a purse-string fashion.
    - This will be used later to anchor the line after placement and to provide hemostasis should the line accidentally be pulled out and bleeding ensue.
6. Using the scalpel blade, cut the umbilical cord horizontally approximately 2 cm above the junction between the cord and the skin.
7. Identify the umbilical vessels
    - The vein is thinner walled, larger in diameter, and somewhat floppy appearing, relative to the umbilical arteries, and typically located at the 12 o'clock position.
    - The arteries are smaller, thick walled, and paired (a single umbilical artery often signifies the presence of a congenital malformation/syndrome) and located at the 4 and 8 o'clock positions (Fig. 121.2).
8. Place the stopcock on the receiving end of the umbilical catheter or the 5-French feeding tube and flush with heparinized saline solution, and then close the stopcock.
    - It is imperative that there is no air in the catheter.
9. Introduce a closed smooth-surfaced forceps into the lumen of the umbilical vein and allow the forceps to separate, allowing the vein to dilate (Fig. 121.3).
10. Insert the catheter (or feeding tube) into the lumen and gently advance, directing the catheter toward the right shoulder (Fig. 121.4).
11. Advance the catheter only until good blood flow is noted and then another 1 to 2 cm (this should be a total of only 4–5 cm in a term infant).
    - *Do not force the advancement.*
    - At this level the tip of the catheter should still be inferior to the liver (Fig. 121.5).
12. Tighten the umbilical tie or the purse-string suture.
13. Secure the catheter with a tape bridge (Figs. 121.6, 121.7, and 121.8).
14. Although if placed only up to point of blood return, the catheter tip should be below the liver; it is best to get an abdominal x-ray to confirm; many solutions are caustic to the liver and can result in complications.

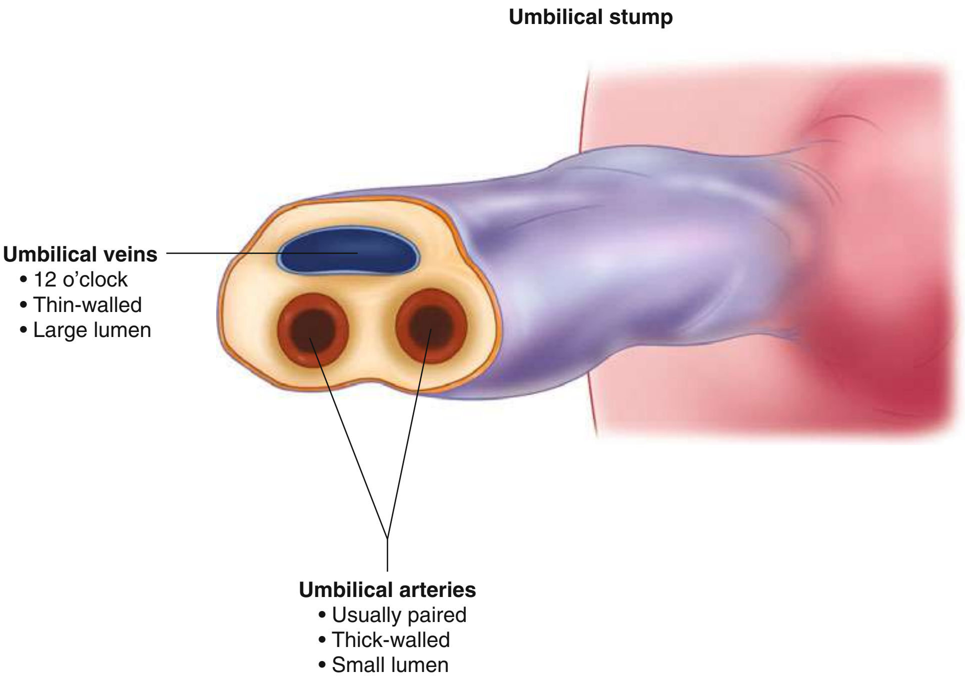

Fig. 121.2 Anatomy of the umbilical cord when cut transversely approximately 2 cm from abdominal wall

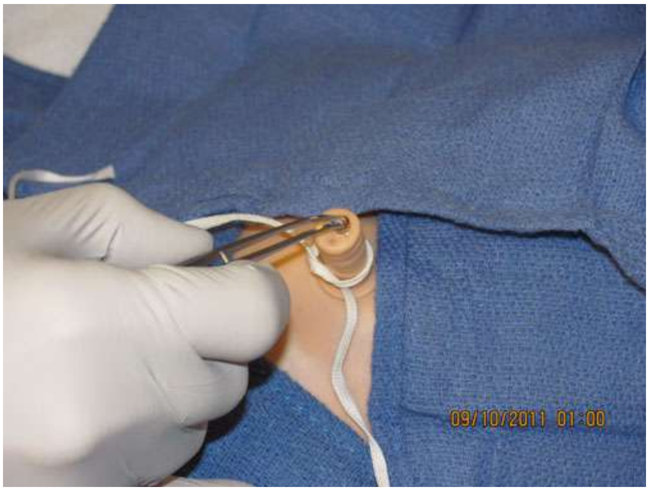

Fig. 121.3 Sterilely draped umbilicus, with umbilical tape loosely tied at base of umbilical cord. This tie will be cinched and secured once the catheter is placed, but meanwhile, can assist with homeostasis should the catheter be inadvertently dislodged during placement. Gently dilate the umbilical vein with a smooth toothed forceps. Insert the distal most couple of mm of a closed forceps into umbilical vein and relax, so the forceps tips smoothly separate

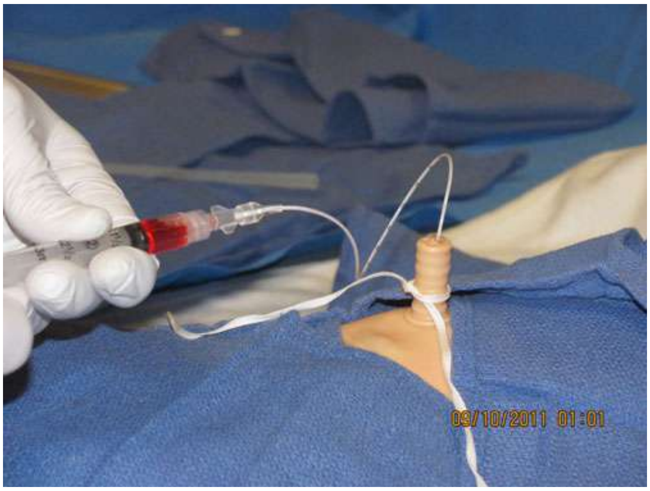

Fig. 121.5 Once there is easy blood return, pass the catheter an additional 1–2 cm. The catheter should still be inferior to the liver at this point

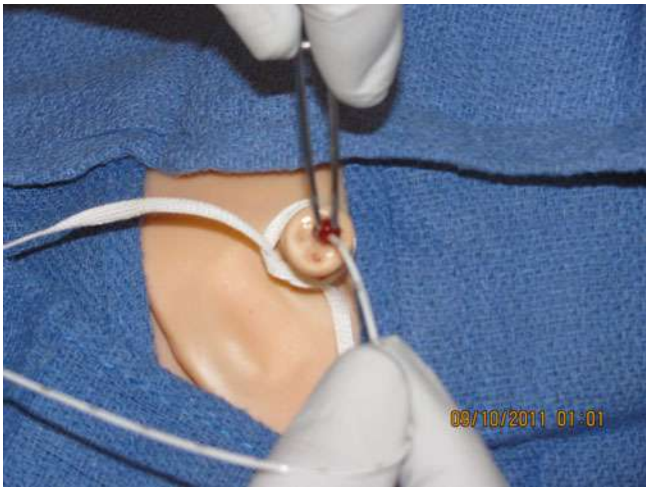

Fig. 121.4 Gently pass the umbilical venous catheter (or 5 F feeding tube) until there is easy blood return

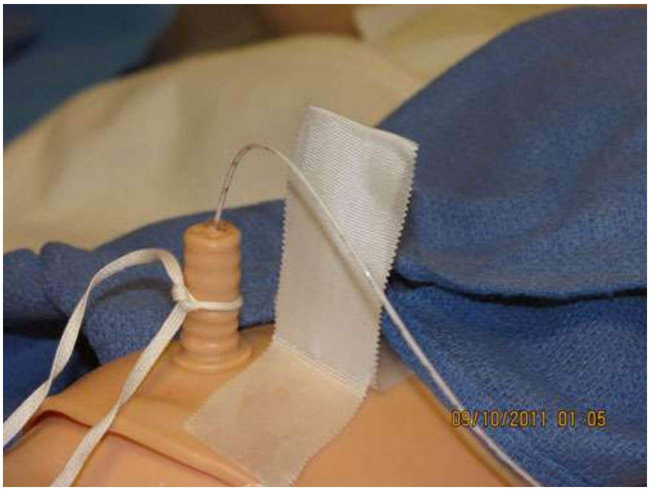

Fig. 121.6 Creating an umbilical catheter tape bridge: the uprights

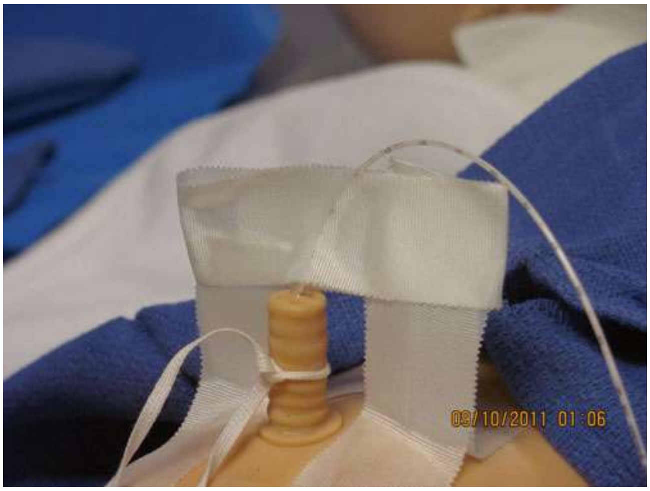

Fig. 121.7 Umbilical tape catheter bridge: the "cross bar" secures the catheter without applying tape to the umbilical stump

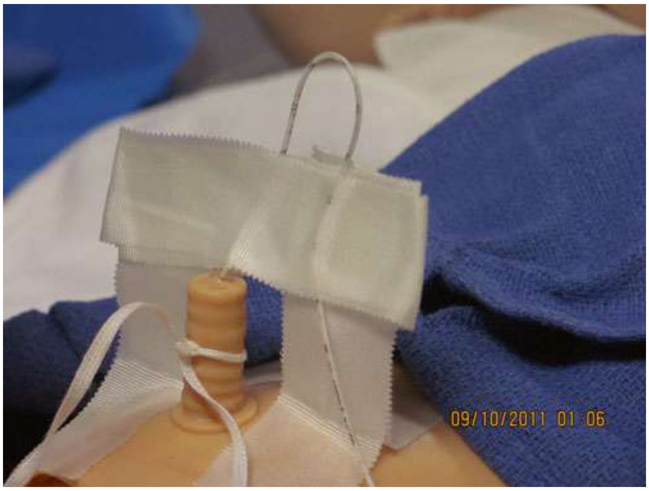

Fig. 121.8 Fold the catheter over to create a "U." Be careful not to kink the catheter, and secure with a second taped cross bar to lend additional security to the catheter and prevent displacement

121.5 Procedure (Removal)

1. The UVC should be removed as soon as adequate peripheral venous access is obtained (unless in an infant estimated to weigh <1000 g).
2. Turn infusions off.
3. Be certain the stopcock is closed to the infant.
4. It is imperative that there be no air in the catheter before withdrawal (if air is present and infant takes inspiration, the negative pressure generated can pull significant amount of air into the central vasculature).
5. Remove the securing tape from the infant.
6. Withdraw the catheter gradually as a single maneuver.

121.6 Complications

- Infection
- Bleeding due to disconnection of tubing (*always use Luer-Lok connections*) or perforation of vessels
- Arterial injury by accidental perforation
- Hepatic injury and necrosis if the catheter sits within a portal vein
- Thrombosis
- Air embolus
- Dysrhythmia or pericardial tamponade or perforation if catheter is advanced too far

121.7 Pearls and Pitfalls

- Pearls
  - The umbilical vein is 2–3 cm long before it widens into the umbilical recess, just before intersecting with the left portal vein and the ductus venosus.
  - Be certain to include the length of the umbilical stump in any calculations for placement.
  - If using the calculating graph, measure from the right shoulder to the umbilicus (Fig. 121.9).
- If choosing the above-diaphragm site for placement, corkscrewing the catheter clockwise while passing the catheter will encourage it to pass through the ductus venosus.
  - A kidney, ureter, and bladder study is mandatory in high UVC placement.
- Pitfalls
  - When preparing with Betadine (povidone-iodine), be certain to wipe away and remove any pooled Betadine along the infant's side because this will cool the infant initially; as the Betadine is warmed, it becomes highly irritating to newborn skin.

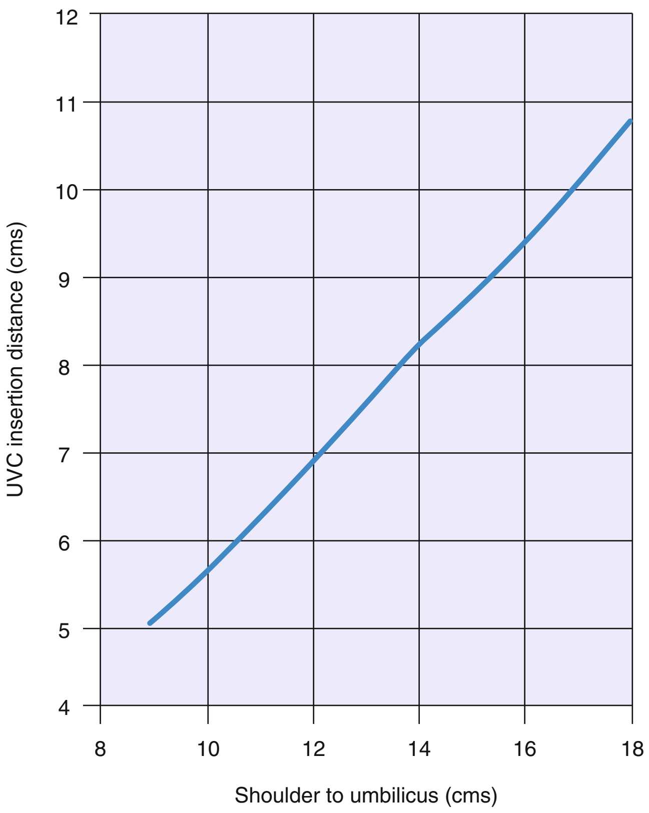

Fig. 121.9 Graph for estimation of insertion distance of the catheter; be certain to add the length (cm) of the umbilical stump to the insertion distance

Selected Reading

Sudbury, Jones, Bartlett. Emergency vascular access. In: APLS: the pediatric emergency medicine resource. 5th ed., p. 741.

Magnan JP. Umbilical Vein Catheterization. http://emedicine.medscape.com/article/80469-overview.

Schlesinger AE, Braverman RM, DiPietre MA. Neonates and umbilical venous catheters: normal appearance, anomalous positions, complications, and potential aid to diagnosis. Presented at the 2001 annual meeting of the American Roentgen Ray Society, Seattle, Apr 2001.

Intraosseous Access

122

Judith K. Lucas

122.1 Indications

- During cardiac arrest: Failure to attain vascular access after three peripheral intravenous attempts or 90 s, whichever comes first
- Inability to gain vascular access in pediatric patients presenting in shock due to hemorrhage (trauma), sepsis, profound dehydration, or cardiac failure

122.2 Contraindications

- Absolute
  - Fracture of the long bone considered for intraosseous (IO) access
- Relative
  - Previous IO access attempt in the same long bone
  - Cellulitis over the insertion site
  - Inferior vena cava injury (circulatory access proximal to the injury site is preferred).
  - Osteogenesis imperfecta

122.3 Materials and Medications

- Preparation materials, such as an antiseptic solution and sterile drapes, IF the patient's stability offers the time.
- Lidocaine without epinephrine: If it becomes necessary to place an IO in a conscious patient, use lidocaine to anesthetize the skin to the bony cortex; then once accessed, infiltrate 2–3 mL into the marrow to alleviate some of the pain of medications infusing through the marrow.
- IO needle (a few examples shown) (Figs. 122.1, 122.2, and 122.3).

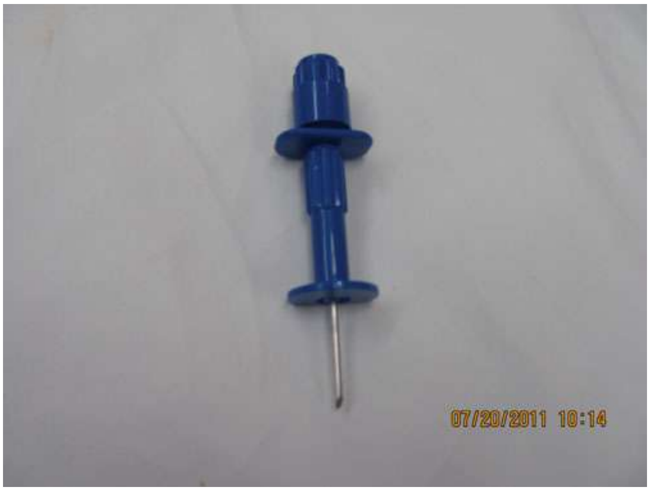

Fig. 122.1 Jamshidi disposable sternal/iliac aspiration needle (Jamshidi, Cardinal Health Dublin, OH)

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_122

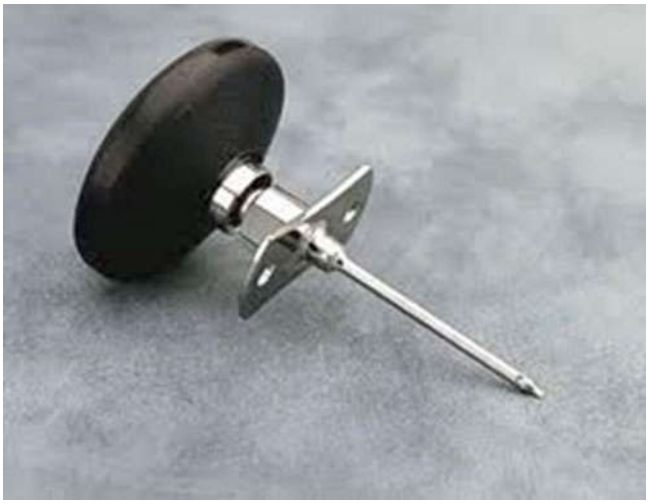

Fig. 122.2 Cook intraosseous needle (Cook Critical Care, Bloomington, IN)

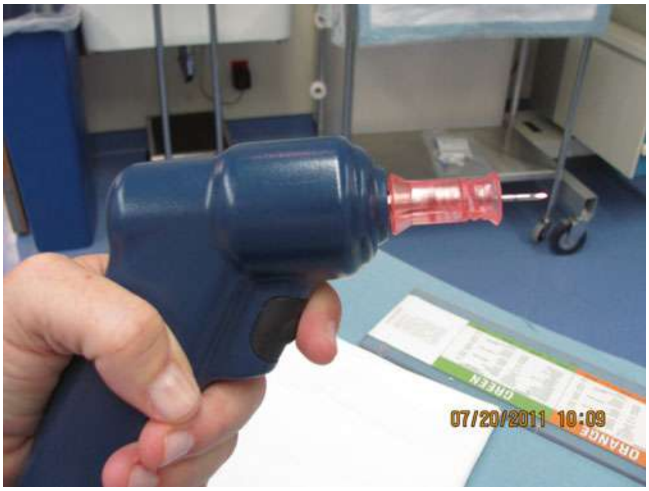

Fig. 122.3 EZ-IO (Vida-Care, San Antonio, TX)

122.4 Procedure

- Most common site: proximal tibia
  1. Patient should be supine, with the intended leg slightly externally rotated and flexed at the hip. Flex the knee about 90°.
  2. Place a towel underneath the knee and the proximal lower leg.
  3. Palpate the tibial tuberosity. Then move fingers 2 cm distal to the tuberosity (1–2 fingerbreadths) and 2 cm medial. This area is consistently flat and is distal to the growth plate (Figs. 122.4 and 122.5).
  4. If the patient is stable, create a sterile field focused on the insertion site.
  5. Insertion site should be cleaned with an antiseptic solution.
  6. Pass the needle through the skin and subcutaneous tissue to the cortex.
  7. Once bone is reached, stabilize the IO with the thumb and first fingers adjacent to where the needle penetrates the skin.
  8. Apply steady and firm pressure downward and with a twisting motion (back and forth) with the palm on the end of the device. Although inserting the needle perpendicularly to the bone is totally acceptable, ideally angling the needle caudally slightly (~15°) will avoid the growth plate (Fig. 122.6).
  9. As the cortex is fully penetrated indicating entrance into the marrow space, one *may* feel an ease in resistance. Do not advance the needle further. The needle should be able to stand up without support.
  10. Remove the inner trocar (Fig. 122.7).
  11. Attempt to aspirate marrow/blood. Not being able to do so DOES NOT mean inadequate or inaccurate placement. Instead, if marrow cannot be aspirated, affix a 10-mL syringe with normal saline and attempt to infiltrate. Resistance to flow should be minimal, and one should not appreciate either extravasation around the insertion site, coolness, or tissue expansion posterior to the site (indicating the fluid is passing through the IO).

12. Once placement is confirmed, connect an intravenous line, using a three-way stop-cock.
13. Secure the needle with tape and gauze (Fig. 122.8).

- Alternative sites
  - Distal tibia
    - The landmarks are the medial aspect of the tibia (the flat portion), 2 fingerbreadths proximal to the medial malleolus.
    - Again, externally rotate and abduct the hips, with the knee flexed about 60° (as with the proximal tibia).
    - Angle the needle toward the knee (cephalad) about 10–15° to avoid the distal tibia growth plate.
    - The remainder of the IO insertion is identical to the procedure for proximal tibia placement.
  - Distal femur
    - Slightly flex and externally rotate the hip.
    - Flex the knee enough that the quadriceps muscle group is relaxed.
    - The landmarks for the distal femur are the anterior thigh, midline, about 3 fingerbreadths proximal to the medial and lateral condyles.
    - The IO is inserted perpendicular to the bone (because there is no growth plate in the distal femur other than the condyles).
    - The remainder of the procedure is identical to the procedure for placement in the proximal tibia.
  - Proximal humerus (Fig. 122.9)
    - The patient should be supine, with the shoulder, upper arm, and elbow as close to the body as possible, yet still on the bed. The elbow should be flexed at 90°, with the forearm and palm resting on the patient's abdomen.
    - The provider slides his or her thumb up the anterior shaft of the humerus toward the shoulder until the greater humerus tubercle is palpated, which identifies the surgical neck of the humerus.
    - The insertion site is perpendicular to the humerus, approximately 1–2 fingerbreadths proximal to the tubercle.
    - The remainder of the insertion is identical to that for the proximal tibia.
  - Sternum (not usually recommended for small children).
    - Requires a special IO needle and system.
    - Method of placement is specific to the insertion system chosen.
    - Manubrium is the desired site (as opposed to the body of the sternum, which would interfere with cardiopulmonary resuscitation).
    - Risks specific to sternal placement include pneumothorax, mediastinitis, and great vessel injury.

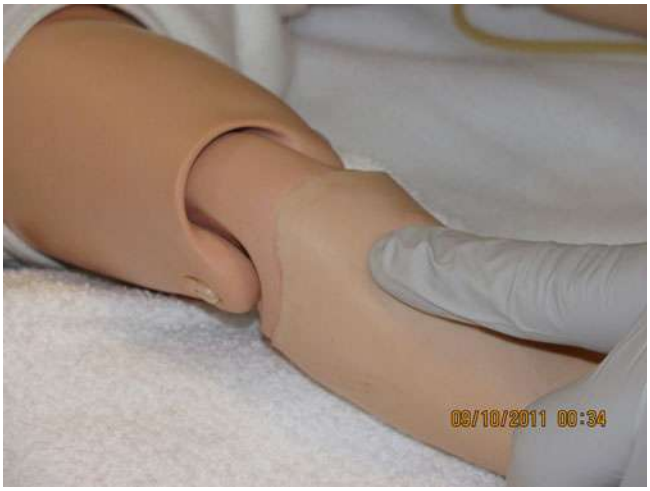

Fig. 122.4 Step 1. Identify the tibial tuberosity

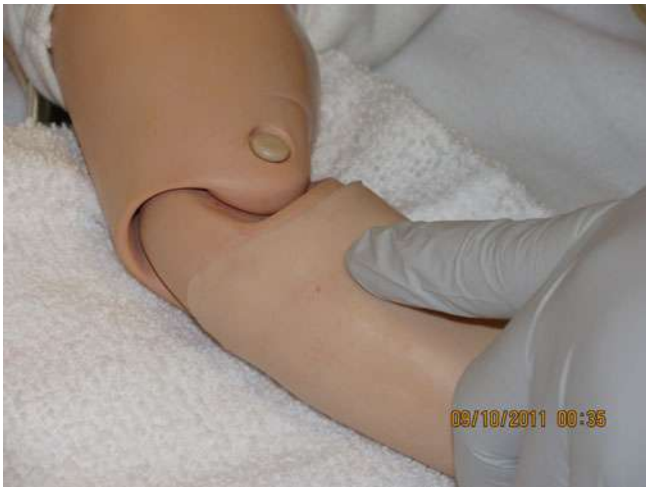

Fig. 122.5 Step 2. One to 2 fingerbreadths below the tuberosity and medially, to the flat aspect of the proximal tibia

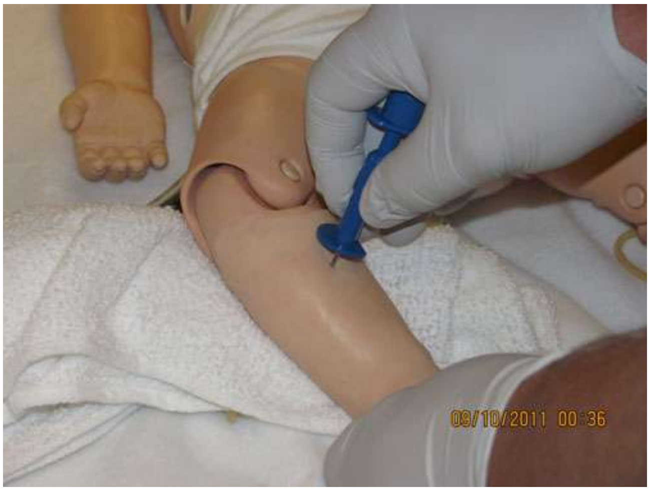
09/10/2011 00:36

Fig. 122.6 Step 3. Angle the device slightly caudal to avoid growth plate

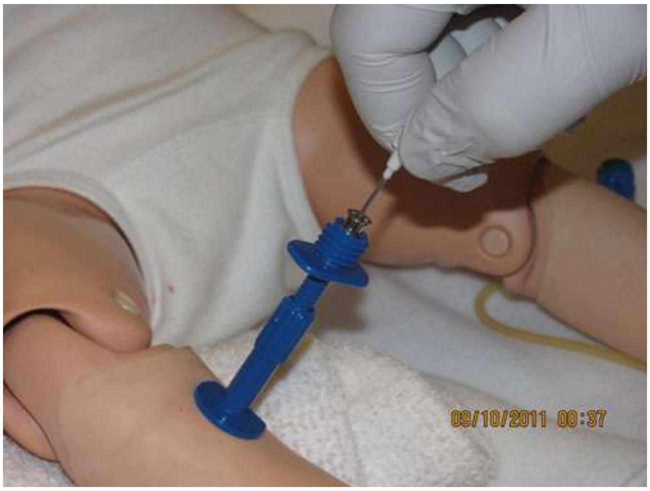
09/10/2011 00:37

Fig. 122.7 Step 4. Uncap the top of the device and remove the trocar

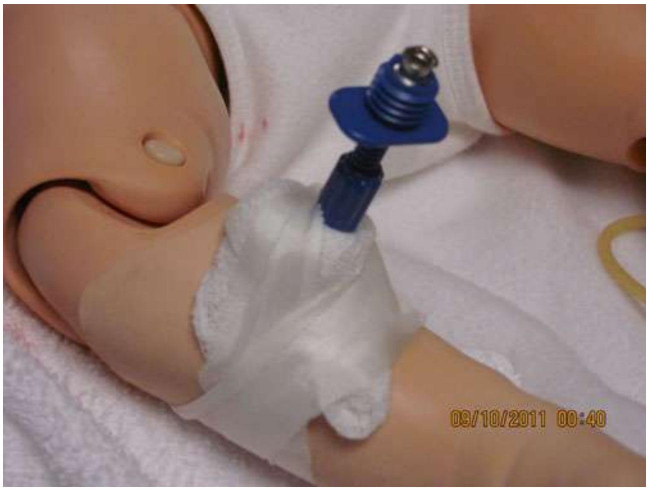
09/10/2011 00:40

Fig. 122.8 Step 5. Stabilize the intraosseous cannula

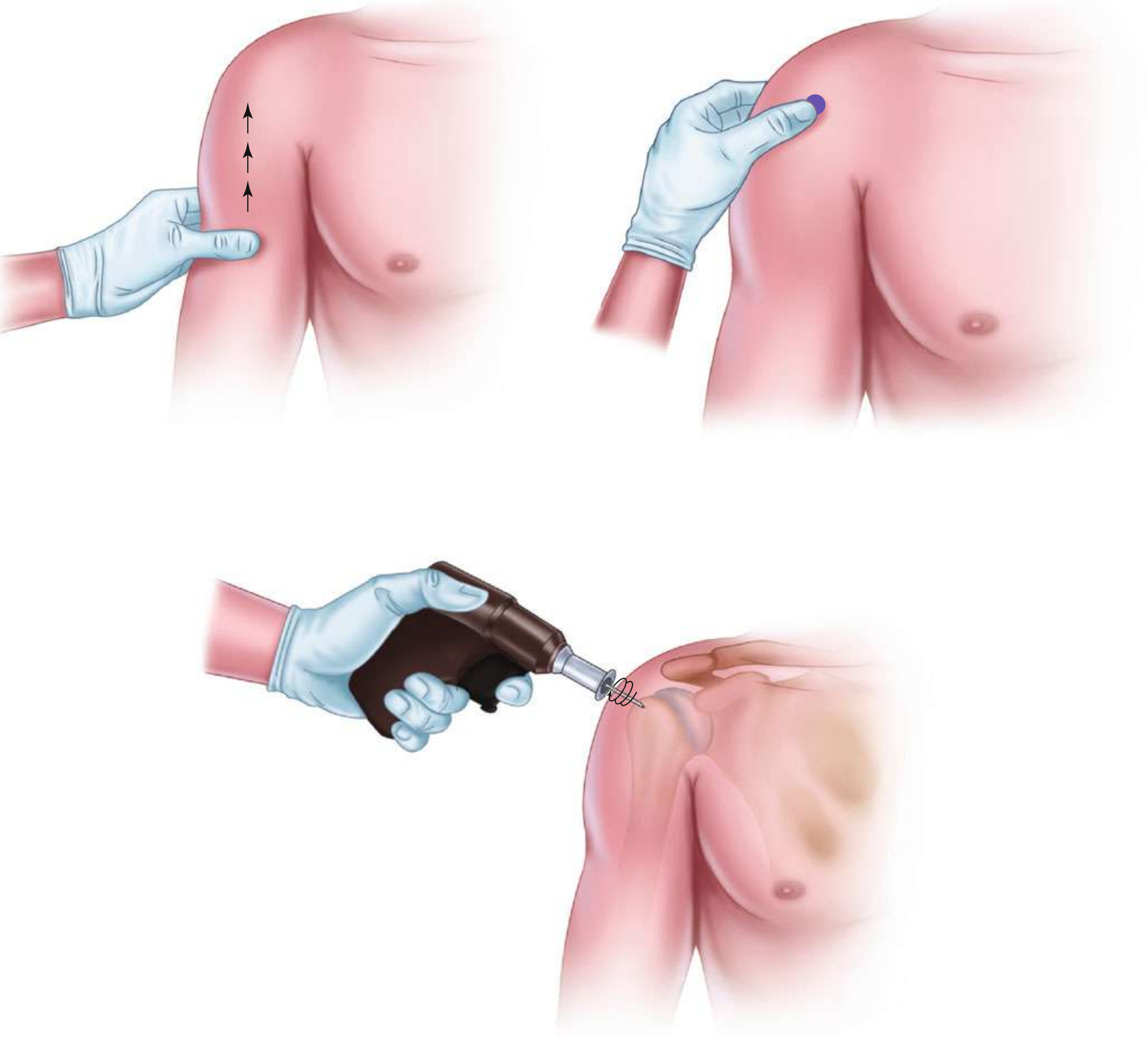

Fig. 122.9 IO insertion into proximal humerus site

122.5 Complications

- Extravasation of fluid.
  - Occurs as a result of a misplaced IO, either not completely in the marrow space anteriorly or through the cortex posteriorly.
  - Tissue necrosis.
- Compartment syndrome if extravasation is not recognized.
- Fracture and growth plate injury.
- Infection and osteomyelitis are rare complications if sterile technique is used.
- Fat embolism, considered rare and only in adult patients.

122.6 Pearls and Pitfalls

- Be certain to avoid placing a hand beneath the IO site during placement in order to prevent possible IO penetration into the provider's tissue.
- Do not rely on the sensation of a "pop" to determine appropriate penetration into the cortex, especially in infants.
- Likewise, one may not be able to aspirate bone marrow or blood, even with an appropriately placed IO. DO NOT pull out the IO if marrow cannot be aspirated. Rather, gently but firmly infiltrate 10 mL normal saline.
- Blood gases, body chemistries, and blood typing can be obtained from IO samples, but the sample CANNOT be used for hematocrit determination.

Selected Reading

Bohn D. Intraosseous vascular access: from archives to the ABC. Crit Care Med. 1999;27:1053.

DeCaen AR, Reis A, Bhutta A. Vascular access and drug therapy in pediatric resuscitation. Pediatr Clin North Am. 2008;55:909–27.

Deitch K. Intraosseous infusion. In: Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders; 2009. Chap. 25.

EMS World. Intraosseous infusion: not just for kids anymore. Posted Jan. 12, 2011. Updated from Mar 2005. EMSWorld.com. Cygnus Business Media Site.

Halm B, Yamamoto LG. Comparing ease of intraosseous needle placement: Jamshidi versus Cook. Am J Emerg Med. 1998;16:420.

Lumbar Puncture in Pediatrics

123

Maritza A. Plaza-Verduin and Judith K. Lucas

123.1 Indications

- Evaluation of cerebrospinal fluid (CSF) for infection or malignancy
- Measurement of opening pressure
- Treatment of pseudotumor cerebri
- Diagnosis of central nervous system (CNS) metastases
- Instillation of intrathecal chemotherapy
- Injection of radiopaque dye for spinal cord imaging

123.2 Contraindications

- Increased intracranial pressure
- Bleeding diathesis (platelet count <50,000)
- Overlying skin infection near the area of puncture site
- Spinal cord trauma or spinal cord compression
- Signs of progressive cerebral herniation
- Condition of the patient (e.g., unstable airway, potentially dangerous breathing problem, severe circulatory instability) that could cause an abrupt decompensation
- Known spinal cord deformity

123.3 Materials and Medications

- Lumbar puncture (LP) tray (Fig. 123.1)
  - – Sterile drapes
  - – Betadine (povidone-iodine) swabs or tray to pour Betadine
  - – Sterile sponges for preparing the puncture site
  - – Sterile 3-mL syringe with needle for lidocaine injection
  - – Sterile collecting tubes (4)
  - – Sterile spinal needle with stylet (size depending on age of patient)
    - Premature infant: 22 gauge or smaller, 1.5 inch
    - Neonate to 2 years: 22 gauge, 1.5 inch
    - 2–12 years: 22 gauge, 2.5 inch
    - Older than 12 years: 20 or 22 gauge, 3.5 inch
  - – Pressure manometer column with a three-way stopcock
- Betadine solution
- Sterile gloves
- Mask
- Lidocaine (1–2 % without epinephrine)
- 4 % Lidocaine cream (LMX-4) or lidocaine and prilocaine mixture (EMLA [eutectic mixture of local anesthetics])

M.A. Plaza-Verduin, MD
Pediatric Division, Department of Emergency Medicine, Arnold Palmer Hospital for Children, Orlando, FL, USA
e-mail: mari.plazav@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_123

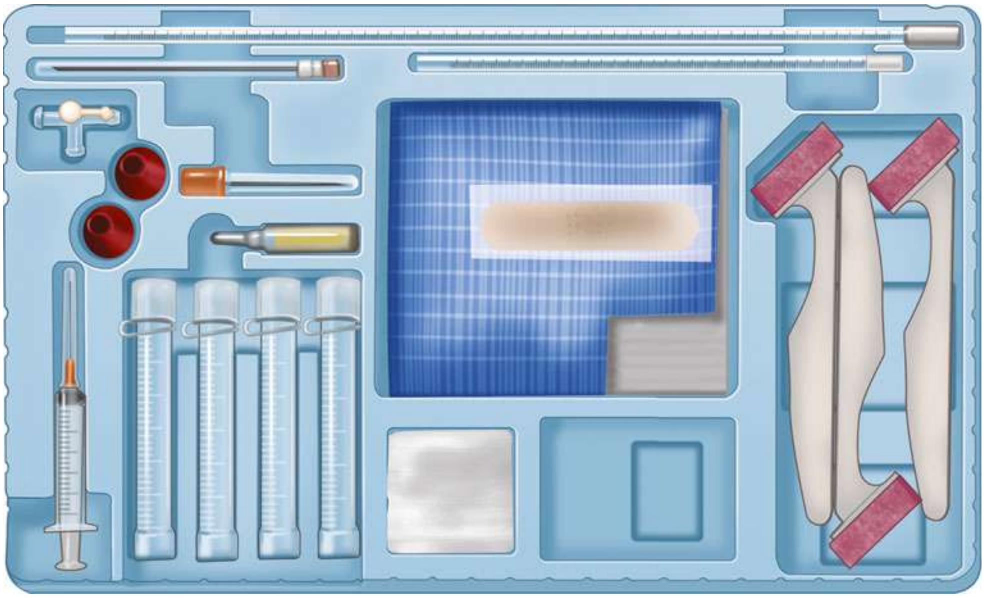

Fig. 123.1 Lumbar puncture (LP) tray

123.4 Procedure

1. Position the child (Figs. 123.2 and 123.3).
   (a) Can be positioned in either the sitting or the lateral recumbent position with the hips, knees, and neck flexed.
   (b) For small infants or any patient with any degree of cardiorespiratory compromise, keep close monitoring of heart rate, respirations, and oxygen saturation while in the flexed position.
2. Palpate the top of the iliac crest and draw an imaginary line connecting the two across the back, which should cross the midline just above the fourth lumbar spine (Fig. 123.4).
3. Palpate L3–4 or L4–5 space along this line.
4. Place EMLA or LMX-4 on the area at this time and allow some time for anesthesia to occur (can take up to 30 min to be effective).
5. Prepare the skin in sterile fashion with Betadine solution using enlarging circles that begin at the puncture site.
6. Drape the patient with sterile towels, exposing the puncture site.
   (a) If an infant, do so conservatively to be able to monitor the infant during the procedure.
7. Locate the intervertebral space (L3–4 or L4–5) once again.
   (a) Make a small mark on the chosen intervertebral space with a fingernail depression or the plastic cap of the spinal needle.
8. If desired, or if anesthetic cream was not previously used or more is needed, apply a small wheal of lidocaine at the desired puncture site using a 25-gauge needle.
9. Insert the spinal needle in the intervertebral space.
   (a) Puncture the skin in the midline just caudal to the palpated spinous process.
   (b) Bevel should be positioned so that the dura mater is pierced parallel to its fibers (which will reduce the likelihood of CSF leakage).
      (i) If in the lateral recumbent position, the bevel of the needle should be positioned horizontally.
      (ii) If in the sitting position, the bevel of the needle should be positioned vertically.
      (iii) Angle the needle slightly cephalad toward the umbilicus and parallel to the bed if the patient is in the lateral recumbent position or slightly caudal (perpendicular to the skin) if the patient is in the sitting position.
10. Advance the needle several millimeters at a time and withdraw the stylet frequently to check for CSF flow.
   (a) Advance the needle until a loss of resistance is felt or approximately 1–2 cm.
      (i) In the infants, whose dura are not so thick and a "pop" or give may be unnoticeable, after passing between the spiny processes and approximately 2 cm through the skin, remove the stylet frequently to check for CSF return. This will allow you to avoid passing through the subarachnoid space.
      (ii) After the change in resistance occurs, if resistance is met again, pull back gently on the needle to reposition in the subarachnoid space and remove the stylet to check for CSF fluid.
11. Once CSF is free flowing, attach a pressure manometer to the needle hub via the three-way stopcock (Fig. 123.5).
   (a) Make sure to hold the spinal needle with one hand in place while attaching the manometer to prevent movement of the needle.
   (b) Measure the CSF pressure once the CSF reaches the highest level in the manometer column.
   (c) CSF pressures will be best obtained in the lateral recumbent position with the neck and legs extended.
   (d) Normal CSF pressure is 5–20-cm H_2O with the neck and legs extended, 10–28-cm H_2O with the neck and legs flexed.
   (e) Have an assistant hold the top of the pressure manometer while the spinal needle is supported at the connection of the manometer and stopcock.
12. Remove the manometer and collect the CSF into sterile tubes.
   (a) Make sure to keep the needle in place when removing the manometer.
   (b) Continue to drain fluid into the collecting tubes, approximately 1 mL per tube.
   (c) The tubes should be labeled for specific studies, depending on the order of collection.
      (i) First tube for gram stain and culture.
      (ii) Second tube for quantitative glucose and protein.
      (iii) Third or fourth tube for cell count and differential.
      (iv) Leftover tube for any additional studies that may be needed.
      (v) When the LP is to assess for possible subarachnoid blood, the first and fourth tubes are sent for cell count.
13. Replace the stylet and remove the spinal needle.
14. Cleanse the puncture area to avoid staining with the Betadine solution.
15. Apply a sterile dressing to the puncture site.
16. Patients 4 years and older should remain in the supine position, with the head elevated no more than the height of a pillow (for comfort), for at least an hour to avoid spinal tap headaches.
17. Likewise, after the LP, giving a fluid bolus of normal saline (10–20 mL/kg in children; 1 L in non-volume-sensitive adults) will also assist with avoidance of spinal tap headaches.

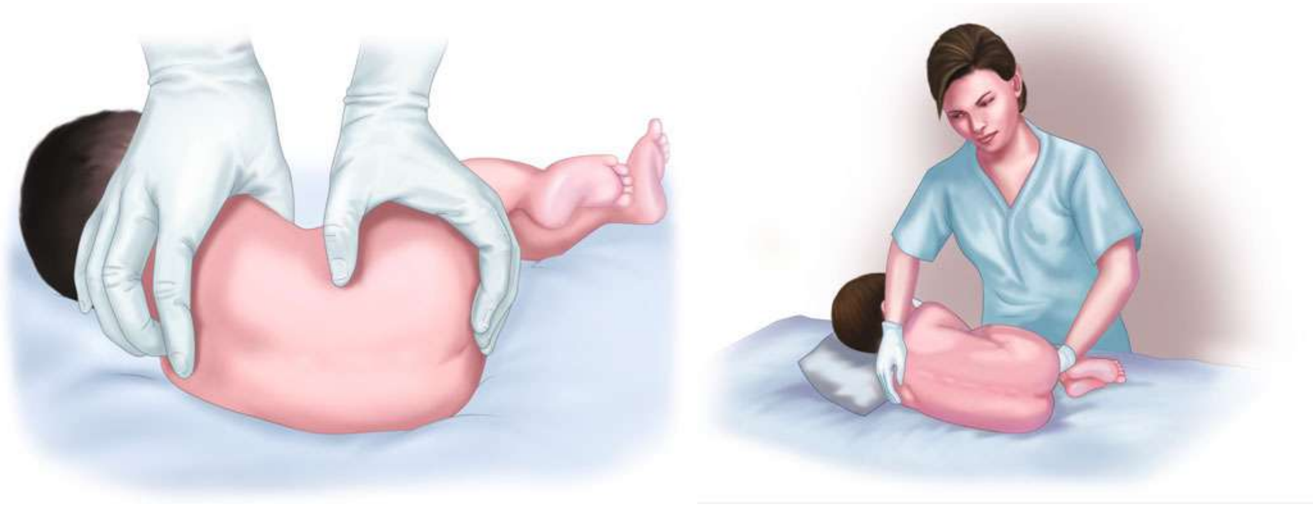

Fig. 123.2 Curling up for an LP

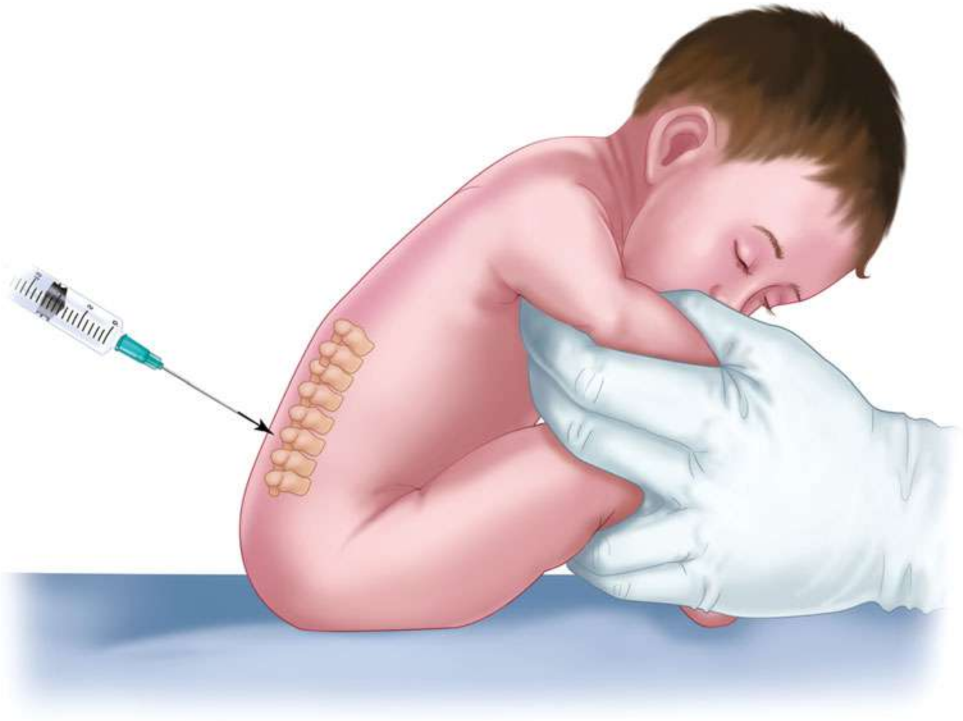

Fig. 123.3 Having the LP

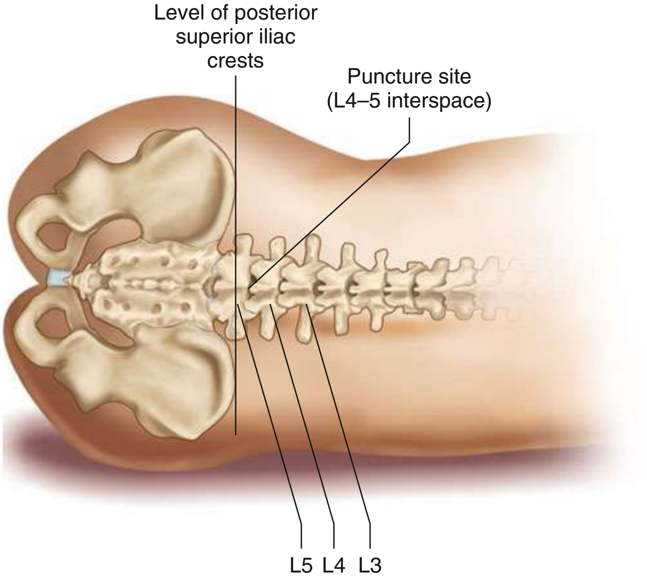
Level of posterior superior iliac crests
Puncture site (L4–5 interspace)
L5
L4
L3

Fig. 123.4 Anatomy of the lumbar spine showing the sites for dural puncture

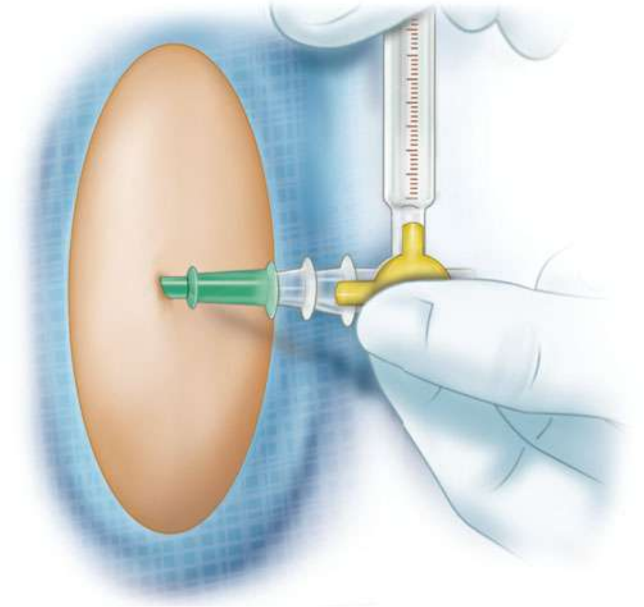

Fig. 123.5 Pressure manometer attached to the needle hub via the three-way stopcock

123.5 Complications

- Minor
  - Localized back pain
  - Transient paresthesia during procedure
  - Post-LP headache
- Major
  - Severe back pain associated with neurological signs (may be subdural or epidural spinal hematoma)
  - LP-induced meningitis
  - Cerebral herniation
  - Acquired epidermoid tumor
  - Damage to adjacent structures (disk herniation, retroperitoneal abscess, spinal cord hematoma)

123.6 Pearls and Pitfalls

- Pearls
  - Success of the LP depends on the positioning of the patient.
    - Goal of positioning is to stretch the ligamentum flavum and increase the interlaminar spaces.
    - In the recumbent position, the shoulder and hips should be perpendicular to the bed, keeping the spinal cord straight, with no rotation.
    - Sitting position is useful with older, cooperative patients or with very young infants who are unlikely to struggle and may have increased respiratory distress in the lateral recumbent position.
  - Administration of an anxiolytic (e.g., midazolam) can be used to facilitate the procedure in an older child. In some cases, procedural sedation may be required.
  - Placing the tip of the thumb on the spinous process just above the space being entered can ensure good alignment of the needle.
  - In older patients a “pop” may often be felt as a change in resistance occurs once the dura is penetrated; however, in infants and neonates, that “pop” may be exceedingly subtle or not palpable.
  - If no CSF returns, attempt the following options:
    - Ensure the needle is in the appropriate position, withdrawing the needle slowly if necessary.
    - Rotate the needle 90°.
    - If an infant, have the assistant massage the anterior fontanel to help facilitate CSF flow.
    - If the procedure continues without CSF, withdraw the needle to just under the skin and redirect it.
    - If the procedure continues without CSF, withdraw the needle and insert a new needle with stylet at an alternate site.

    - If these steps do not yield CSF, the infant may be dehydrated, not allowing for adequate CSF flow.
      - Give the patient a bolus and reattempt later.
      - Attempt putting the infant in a sitting position to increase flow.
- Pitfalls
  - Traumatic LP
    - Bloody CSF fluid, which usually clears as the CSF drains if the needle is in the correct space.
    - Occurs with improper technique (inserting the needle too far to one side into an epidural venous plexus or through the subarachnoid space into or adjacent to the vertebral body).
    - Can occur with proper technique as well.
    - If fluid does not clear and clots form in the tubes, LP should be reattempted at a different site.
  - Failed LP attempts despite proper procedure and positioning
    - Ultrasound can be used to visualize the area and determine the reason for failure or the likelihood of success with future attempts.

Selected Reading

Coley BD, Shiels WE, Hogan MJ. Diagnostic and interventional ultrasonography in neonatal and infant lumbar puncture. Pediatr Radiol. 2001;31:399–402.

Cronan KM, Wiley JF. Lumbar puncture. In: King C, Henretig FM, editors. Textbook of pediatric emergency procedures. 2nd ed. New York: Lippincott Williams & Wilkins; 2008.

Ebinger F, Kosel C, Pietz J, Rating D. Headache and backache after lumbar puncture in children and adolescents: a prospective study. Pediatrics. 2004;113:1588–92.

Friedman AG, Mulhern RK, Fairclough D, et al. Midazolam premedication for pediatric bone marrow aspiration and lumbar puncture. Med Pediatr Oncol. 1991;19:499–504.

Partin WR. Emergency procedures. In: Stone CK, Humphries RL, editors. Current diagnosis & treatment: emergency medicine. 6th ed. New York: McGraw-Hill; 2007.

Use of the Broselow Tape

126

Judith K. Lucas

126.1 Indications

- To determine the equipment size and the medication doses during a pediatric resuscitation, without having to take the time to perform calculations

126.2 Contraindications

- Premature infant or newborn whose heel, while the infant is fully extended, does not fall at least into the white area (corresponding to 3, 4, and 5 kg).
- The length of the child exceeds the distal end of the green area (36 kg).

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

126.3 Materials

- Broselow Pediatric Emergency Tape (Armstrong Medical Industries, Wilshire, IL) (Figs. 126.1 and 126.2)

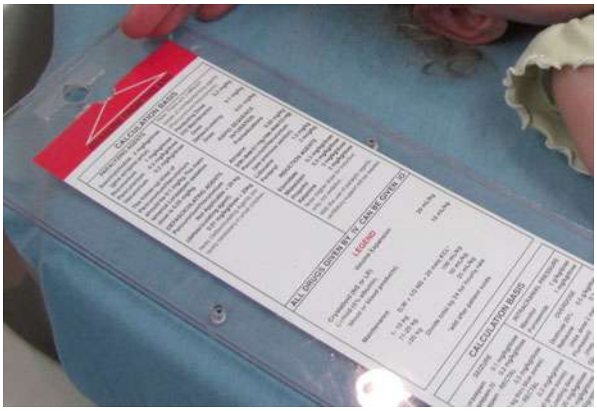
CALCULATION BASIS
ALL DRUGS GIVEN BY IV CAN BE GIVEN IO
LEGEND
CALCULATION BASIS

Fig. 126.1 Broselow tape, proximal end (at patient's head)

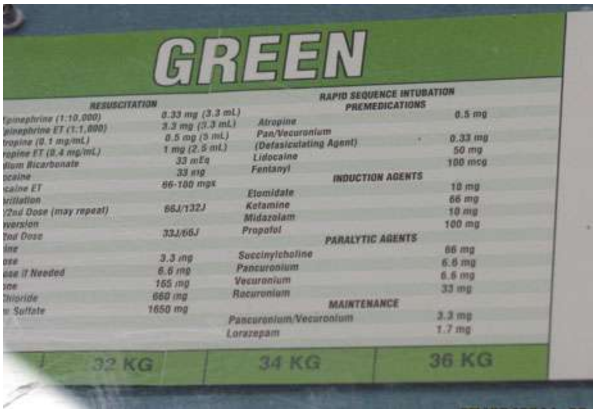
GREEN
RESUSCITATION
0.33 mg (3.3 mL)
3.3 mg (3.3 mL)
0.5 mg (5 mL)
1 mg (2.5 mL)
33 mEq
33 mg
66-100 mgs
66J/132J
33J/66J
3.3 mg
6.6 mg
165 mg
660 mg
1650 mg
RAPID SEQUENCE INTUBATION
PREMEDICATIONS
Atropine
Pan/Vecuronium
(Defasiculating Agent)
Lidocaine
Fentanyl
0.5 mg
0.33 mg
50 mg
100 mcg
INDUCTION AGENTS
Etomidate
Ketamine
Midazolam
Propofol
10 mg
66 mg
10 mg
100 mg
PARALYTIC AGENTS
Succinylcholine
Pancuronium
Vecuronium
Rocuronium
66 mg
6.6 mg
6.6 mg
33 mg
MAINTENANCE
Pancuronium/Vecuronium
Lorazepam
3.3 mg
1.7 mg
32 KG
34 KG
36 KG

Fig. 126.2 Broselow tape, distal end (the tape will accommodate a 36-kg child, at most)

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_126

126.4 Procedure (Figs. 126.3, 126.4, and 126.5)

- Place the infant or child on the bed in the supine position.
- Extend the Broselow Pediatric Emergency Tape next to the patient, placing the red line at the top of the patient's head.
- With the patient fully extended, especially through the hips and knees, align the bottom of the heel, while the ankle is flexed and toes upward, with the color on the tape adjacent to the heel.
- The color on the tape corresponds to the patient's weight.
- The side of the tape with weights noted along the bottom has medication doses for resuscitation, rapid-sequence intubation (RSI), and paralytics (Fig. 126.6).
- The top half of the other side of the tape has the appropriate dosing for managing seizures, overdoses, elevating intracranial pressure, and fluids. The bottom half of this side has the appropriate equipment sizes, such as endotracheal tubes, nasogastric tubes, chest tubes, and other equipment, that correspond to the patient's weight (Fig. 126.7).

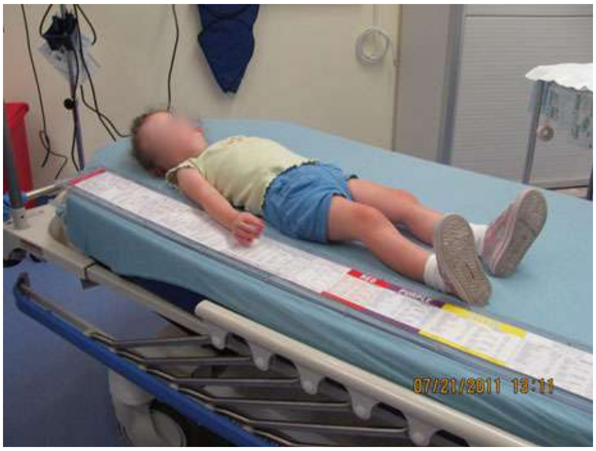

Fig. 126.3 Placement of the child on the tape

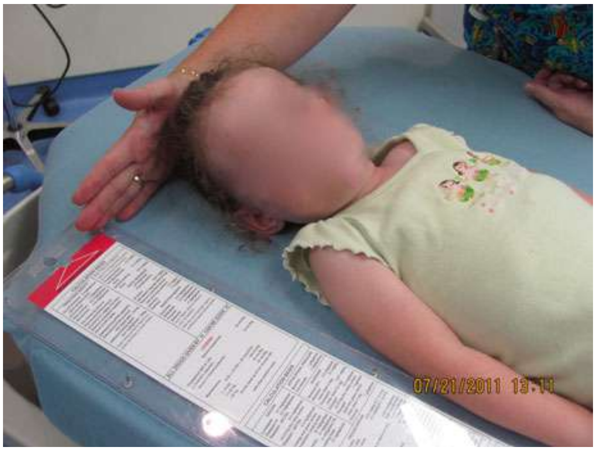

Fig. 126.4 Placement of the child on the tape

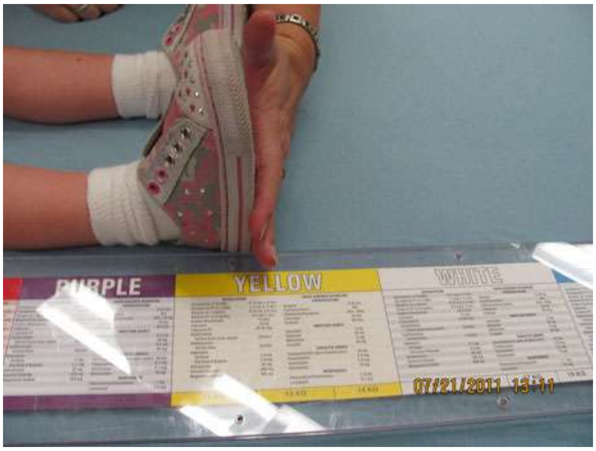

Fig. 126.5 Placement of the child on the tape

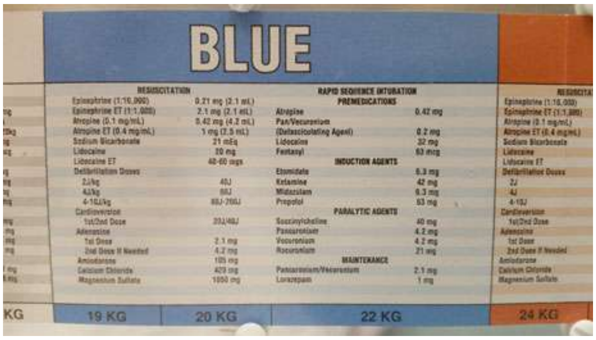

BLUE

RESUSCITATION

Epinephrine (1:10,000)	0.21 mg (2.1 mL)
Epinephrine ET (1:1,000)	2.1 mg (2.1 mL)
Atropine (0.1 mg/mL)	0.42 mg (4.2 mL)
Atropine ET (0.4 mg/mL)	1 mg (2.5 mL)
Sodium Bicarbonate	21 mEq
Lidocaine	20 mg
Lidocaine ET	40-60 mgs
Defibrillation Doses	
2J/kg	40J
4J/kg	80J
4-10J/kg	80J-200J
Cardioversion	
1st/2nd Dose	20J/40J
Adenosine	
1st Dose	2.1 mg
2nd Dose If Needed	4.2 mg
Amiodarone	105 mg
Calcium Chloride	420 mg
Magnesium Sulfate	1050 mg

RAPID SEQUENCE INTUBATION

PREMEDICATIONS	
Atropine	0.42 mg
Pan/Vecuronium	
(Defasciculating Agent)	0.2 mg
Lidocaine	32 mg
Fentanyl	63 mcg
INDUCTION AGENTS	
Etomidate	6.3 mg
Ketamine	42 mg
Midazolam	6.3 mg
Propofol	63 mg
PARALYTIC AGENTS	
Succinylcholine	40 mg
Pancuronium	4.2 mg
Vecuronium	4.2 mg
Rocuronium	21 mg
MAINTENANCE	
Pancuronium/Vecuronium	2.1 mg
Lorazepam	1 mg

19 KG 20 KG 22 KG 24 KG

Fig. 126.6 Front of tape. Weight groupings are noted at the bottom. This is the side with resuscitation and RSI medications

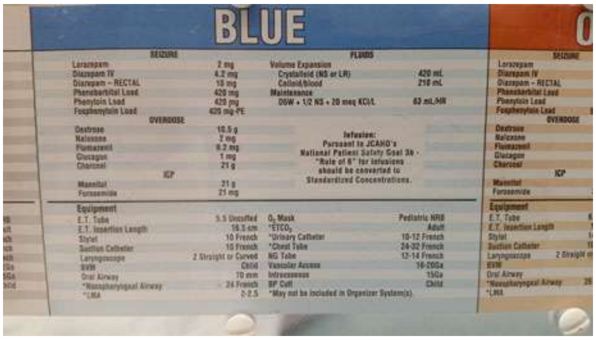

BLUE

SEIZURE

Lorazepam	2 mg
Diazepam IV	4.2 mg
Diazepam – RECTAL	10 mg
Phenobarbital Load	420 mg
Phenytoin Load	420 mg
Fosphenytoin Load	420 mg PE
OVERDOSE	
Dextrose	10.5 g
Naloxone	2 mg
Flumazenil	0.2 mg
Glucagon	1 mg
Charcoal	21 g
ICP	
Mannitol	21 g
Furosemide	21 mg

FLUIDS

Volume Expansion	
Crystalloid (NS or LR)	420 mL
Colloid/blood	210 mL
Maintenance	
D5W + 1/2 NS + 20 meq KCl/L	63 mL/HR

Infusion:
Pursuant to JCAHO's
National Patient Safety Goal 3b -
"Rule of 6" for infusions
should be converted to
Standardized Concentrations.

Equipment

E.T. Tube	5.5 Uncuffed	O_2 Mask	Pediatric NRB
E.T. Insertion Length	16.5 cm	*$ETCO_2$	Adult
Stylet	10 French	*Urinary Catheter	10-12 French
Suction Catheter	10 French	*Chest Tube	24-32 French
Laryngoscope	2 Straight or Curved	NG Tube	12-14 French
BVM	Child	Vascular Access	16-20Ga
Oral Airway	70 mm	Intraosseous	15Ga
*Nasopharyngeal Airway	24 French	BP Cuff	Child
*LMA	2-2.5	*May not be included in Organizer System(s)	

Fig. 126.7 Reverse side of tape. This side has medications for critical interventions aside from the primary survey, as well as equipment sizes

126.5 Pearls and Pitfalls

- Pearls
  - Be certain that the top of the head is at the red line and that the patient's body is fully extended. Infants tend to lay flexed at the hips and knees.
  - Determining the appropriate color code depends on noting the bottom of the heel while the ankle is flexed and the toes are pointing upward.
- Pitfalls
  - Beware of the child who is unusually heavy compared with other children of the same height.

Selected Reading

Luten R. Error and time delay in pediatric trauma resuscitation: addressing the problem with color-coded resuscitation aids. Surg Clin North Am. 2002;82:303–14, vi.

Luten R, Broselow J. Standardization of product concentration in emergency dosing: a response to Fineberg and Arendt. Ann Emerg Med. 2008;52:477–8.

Rosenberg M, Greenberger S, Rawal A, Latimer-Pierson J, Thundiyil J. Comparison of Broselow tape measurements versus estimations of pediatric weights. Am J Emerg Med. 2011;29:482–8.

Nursemaid's Elbow

127

Judith K. Lucas

127.1 Definition

- An injury commonly seen in children between 6 months and late preteens, although generally seen between 1 and 3 years
- A subluxation of the radial head, usually resulting from sudden, longitudinal traction on an extended arm with the wrist pronated
- Often occurs when a parent/caregiver is holding the child by the hand while walking and suddenly pulls the child away from a dangerous situation (Fig. 127.1)

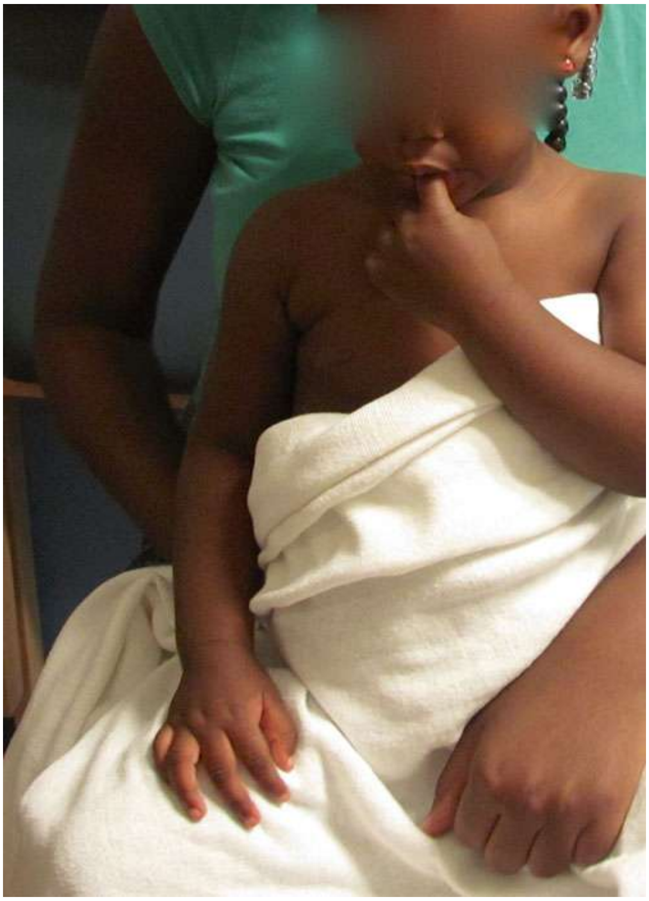

Fig. 127.1 Note that the toddler prefers to hold her right arm pronated and somewhat flexed at the elbow. When asked where it hurts, she may as often point to her wrist as to her elbow. There will be no soft tissue swelling anywhere along the upper extremity

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_127

127.2 Clinical Diagnosis

- History of pulling-type injury.
- Child presents with arm held slightly flexed at elbow and pronated.
- Patient may point to elbow or wrist as source of pain, but both areas are without any swelling and are nontender on palpation.
- Elbow can be flexed and extended, but the forearm cannot be supinated.
- Radiography is not helpful.

127.3 Indications

- Clinical presence of subluxed radial head

127.4 Contraindications

- Absolute
  - Radiographic evidence of elbow or forearm fracture
  - Swelling or pain about the elbow, forearm, or wrist
- Relative
  - Unknown history or witnessing of pulling-type injury

127.5 Materials and Medications

- None

127.6 Procedure

127.6.1 Method A: Superpronation (preferred method) (Fig. 127.2)

1. Explain to the caretaker that the reduction will cause the child *very brief* discomfort.
2. Seat the child on the lap of a parent/caregiver or an assistant, facing the operator and holding the child in such a way as to hold the child's humerus against her or his side.
3. The person performing the procedure holds the elbow at approximately 90° and grasps the elbow, with the physician's thumb over the region of the radial head (this is done in order to be able to palpate the reduction "clunk" or "click").
4. The physician then holds the patient's wrist firmly and rapidly hyperpronates the forearm. A palpable or audible "click" signifies successful reduction but may not be appreciable.
5. The child may cry for a few minutes after reduction; however, the provider should leave the bedside for 5–10 min and instruct the parents to allow the child to simply play, without focusing on the affected arm.
6. After approximately 10 min, return to the bedside and reevaluate the child, who should have full use of her or his arm by then.

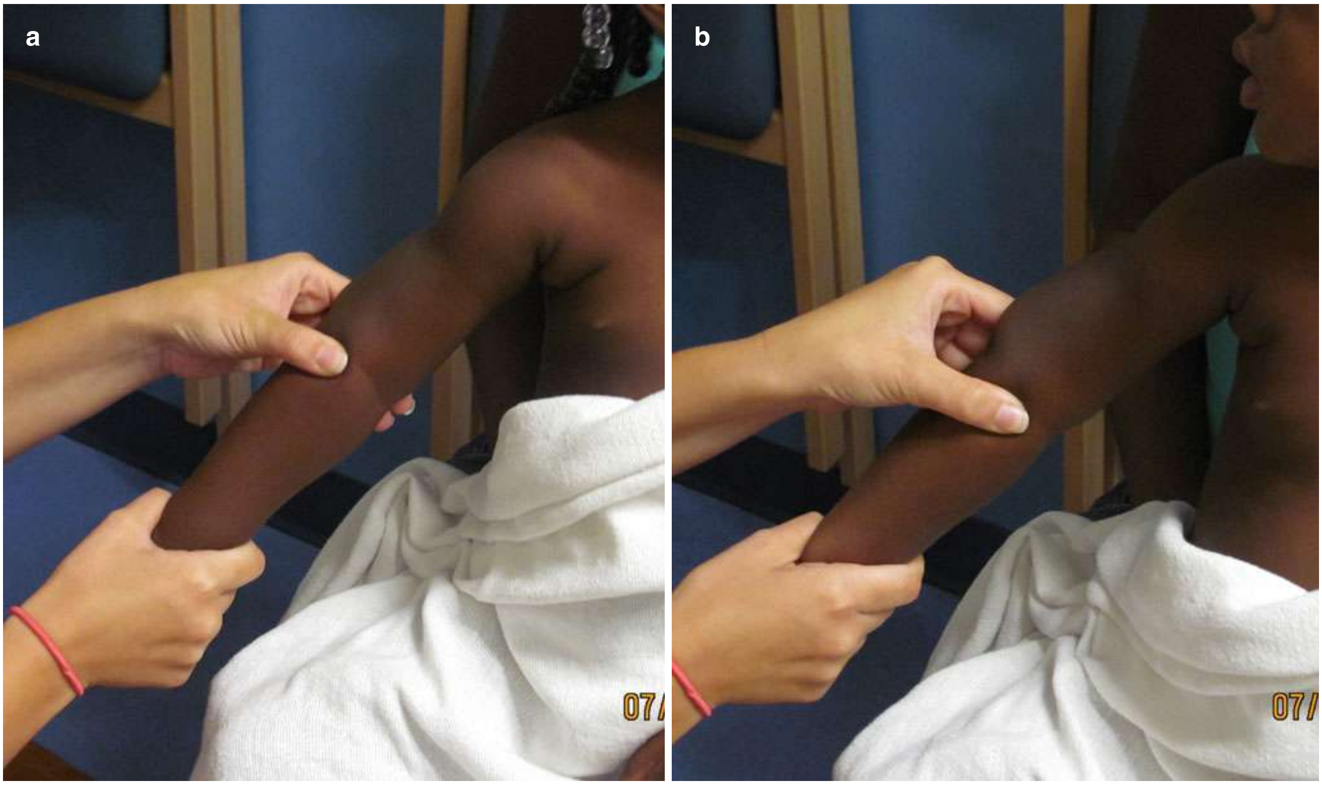

Fig. 127.2 Pronation: (**a**) The examiner's left hand holds the patient's radial head and medial condyle to better appreciate palpable reduction of the subluxation, (**b**) The examiner's right hand gently, but firmly, hyperpronates the patient's forearm from the level of the wrist. Within this maneuver, the examiner should feel or sense a "pop"

127.6.2 Method B: Supination (Fig. 127.3)

1. Follow the first three steps in Method A.
2. While holding the patient's wrist firmly, the operator steadily supinates the patient's forearm completely, followed by flexing the elbow, bringing the patient's wrist up to the shoulder.
3. Return to the fifth and sixth steps in Method A.

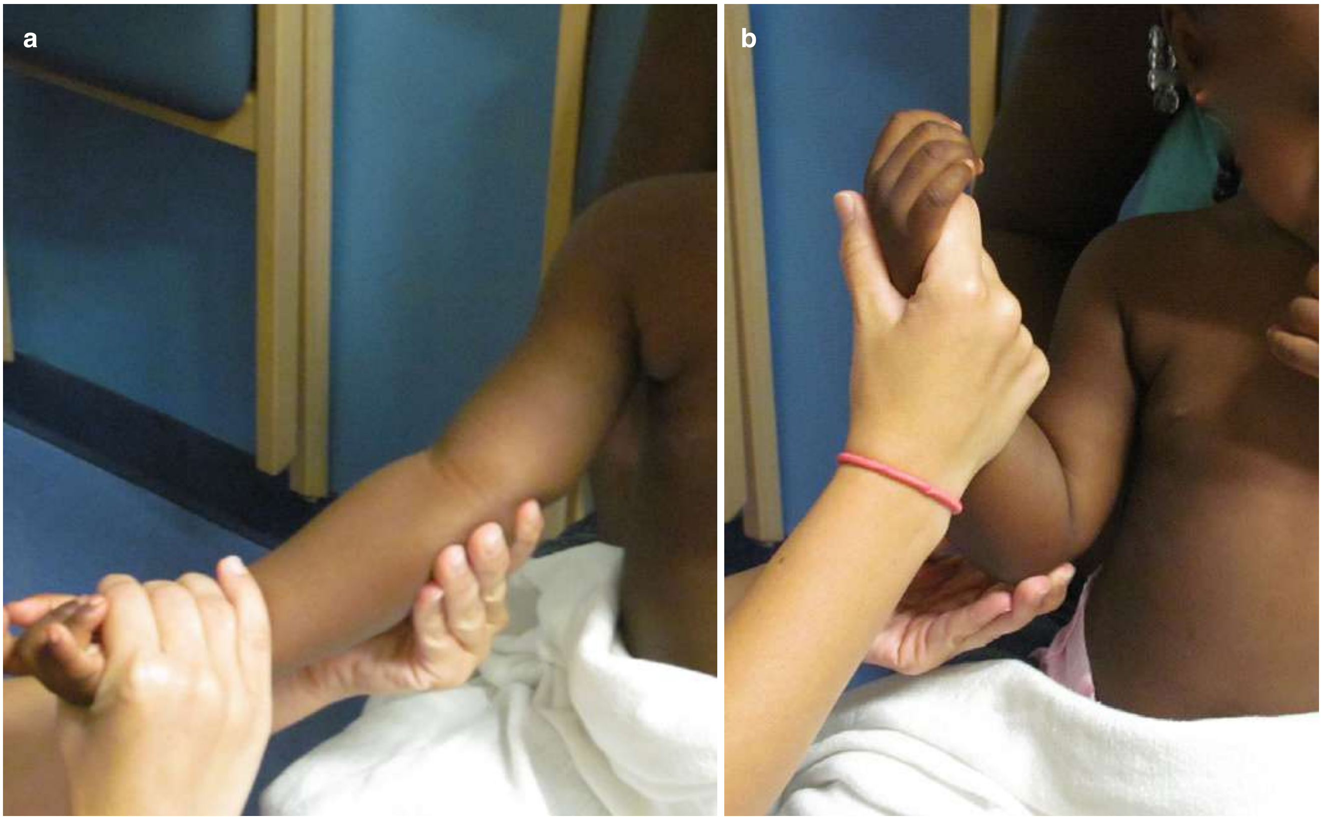

Fig. 127.3 Supination: (**a**) With the examiner's left hand in the same position as for the pronation technique, (**b**) the examiner supinates the forearm, from the level of the wrist, and fully flexes at the elbow. In either of these two maneuvers, the examiner should feel or sense a "pop"

ICRC

VOLUME 1

WAR SURGERY

WORKING WITH LIMITED RESOURCES IN ARMED CONFLICT AND OTHER SITUATIONS OF VIOLENCE

0973/002 03.2009 3,000

ICRC

WAR SURGERY

WORKING WITH LIMITED RESOURCES IN ARMED CONFLICT AND OTHER SITUATIONS OF VIOLENCE

VOLUME 1

ICRC

REFERENCE

WAR SURGERY

WORKING WITH LIMITED RESOURCES
IN ARMED CONFLICT
AND OTHER SITUATIONS OF VIOLENCE

VOLUME 1

C. Giannou M. Baldan

ICRC

International Committee of the Red Cross
19 Avenue de la Paix
1202 Geneva, Switzerland
T + 41 22 734 60 01 F + 41 22 733 20 57
E-mail: shop.gva@icrc.org
www.icrc.org
© ICRC, March 2009

Cover photos: A. Bakkour/ICRC; A. Bakkour/ICRC; V. Louis/ICRC.

PREFACE

In 1863, a small group of Swiss citizens founded the International Committee of Geneva for the Relief of Wounded Soldiers. A year later, an international diplomatic conference negotiated the first Geneva Convention for the Amelioration of the Condition of Wounded Soldiers in the Field, which to this day is one of the cornerstones of international humanitarian law, and gave the International Committee of the Red Cross its definitive name.

International humanitarian law – the law of war – has developed since, and the mandate, role and activities of the ICRC have expanded to include both protection and assistance for all the victims of armed conflict and other situations of violence. Assistance and relief programmes are now run according to a public health approach and aim to be holistic regarding human needs while respecting the dignity of each and every one.

War surgery – the care of the wounded in armed conflict and other situations of violence – remains a pillar of the ICRC's identity. Over the years, ICRC hospital teams have unfortunately been witness to a great deal of physical and mental suffering in this world. Through caring for the sick and wounded in so many different conflict zones, the ICRC and its partners in the International Red Cross and Red Crescent Movement have developed an expertise they are keen to share, and created a pool of human resources ever willing and prepared to help alleviate some of the suffering.

This new manual presents some of this expertise, gained at great human cost, in the hope that one day it will no longer be required.

Jakob Kellenberger

President
International Committee of the Red Cross

TABLE OF CONTENTS

INTRODUCTION 9

Chapter 1 SPECIAL CHARACTERISTICS OF SURGERY IN TIMES OF CONFLICT 17
1.1 Differences between surgery in times of conflict and civilian practice 19
1.2 How war surgery differs 20
1.3 “Surgeries” for victims of war 25
1.4 Differences between military and non-military war surgery: the ICRC approach 27
Annex 1. A ICRC criteria for introducing a new technology 31

Chapter 2 APPLICABLE INTERNATIONAL HUMANITARIAN LAW 33
2.1 Historical introduction 35
2.2 International humanitarian law: basic principles 36
2.3 The distinctive emblems 38
2.4 The International Red Cross and Red Crescent Movement and its Fundamental Principles 39
2.5 Rights and duties of medical personnel according to IHL 39
2.6 Responsibility of States 41
2.7 Reality check: some people do not follow the rules 43
2.8 The neutrality of a National Red Cross/Red Crescent Society 44
2.9 The role and mandate of the ICRC in situations of armed conflict 45
Annex 2. A The distinctive emblems 48
Annex 2. B The International Red Cross and Red Crescent Movement 50

Chapter 3 MECHANISMS OF INJURY DURING ARMED CONFLICT 53
3.1 The various mechanisms of injury 55
3.2 Ballistics 59
3.3 Terminal ballistics 63
3.4 Wound ballistics 70
3.5 Wound dynamics and the patient 78

Chapter 4 RED CROSS WOUND SCORE AND CLASSIFICATION SYSTEM 81
4.1 Applications of the RCWS and classification system 83
4.2 Principles of the Red Cross Wound Score 84
4.3 Grading and typing of wounds 88
4.4 Wound classification 88
4.5 Clinical examples 89
4.6 Conclusions 91

Chapter 5 THE EPIDEMIOLOGY OF THE VICTIMS OF WAR 93
5.1 Introduction: purpose and objectives 95
5.2 Public health effects of armed conflict 95
5.3 Epidemiology for the war surgeon 98
5.4 General questions of methodology 101
5.5 Aetiology of injury 104
5.6 Anatomic distribution of wounds 106
5.7 Fatal injuries 110
5.8 The lethality of context: delay to treatment 114
5.9 Hospital mortality 117
5.10 ICRC statistical analysis of hospital workload 118
5.11 Conclusions: lessons to be gained from a study of epidemiology 121
Annex 5. A ICRC surgical database 123
Annex 5. B Setting up a surgical database for the war-wounded 125

Chapter 6 THE CHAIN OF CASUALTY CARE 129
6.1 The links: what kind of care, and where? 131
6.2 Surgical hospital treating the war-wounded 133
6.3 Transport 134
6.4 Forward projection of resources 135
6.5 The reality: common war scenarios 136
6.6 Conflict preparedness and implementation 137
6.7 The pyramid of ICRC surgical programmes 138
Annex 6. A Initial assessment of a surgical hospital treating the war-wounded 140
Annex 6. B Strategic assessment of a conflict scenario 146
Annex 6. C Humanitarian intervention for the wounded and sick: typical settings 148

Chapter 7 FIRST AID IN ARMED CONFLICT 153
7.1 First aid: its crucial importance 155
7.2 First aid in the chain of casualty care 156
7.3 First aiders: an important human resource 157
7.4 Essential elements of the first-aid approach and techniques 157
7.5 Setting up a first-aid post 159
7.6 ICRC involvement in first-aid programmes 161
7.7 Debates, controversies and misunderstandings 161

Chapter 8 HOSPITAL EMERGENCY ROOM CARE 169
8.1 ABCDE priorities 171
8.2 Initial examination 171
8.3 Airway 172
8.4 Breathing and ventilation 175
8.5 Circulation 177
8.6 Blood transfusion where supplies are limited 181
8.7 Disability 184
8.8 Environment/exposure 185
8.9 Complete examination 185

Chapter 9 HOSPITAL TRIAGE OF MASS CASUALTIES 189
9.1 Introduction 191
9.2 Setting priorities: the ICRC triage system 193
9.3 How to perform triage 196
9.4 Triage documentation 198
9.5 Emergency plan for mass casualties: disaster triage plan 199
9.6 Personnel 200
9.7 Space 202
9.8 Equipment and supplies 203
9.9 Infrastructure 204
9.10 Services 204
9.11 Training 204
9.12 Communication 204
9.13 Security 205
9.14 Summary of triage theory and philosophy: sorting by priority 206
Annex 9. A Sample triage card 207
Annex 9. B Hospital emergency plan for a mass influx of wounded 208

Chapter 10 SURGICAL MANAGEMENT OF WAR WOUNDS 211
10.1 Introduction 213
10.2 Complete examination 214
10.3 Preparation of the patient 215
10.4 Examination of the wound 216
10.5 Surgical treatment 216
10.6 Retained bullets and fragments 221
10.7 Final look and haemostasis 222
10.8 Wound excision: the exceptions 223
10.9 Leaving the wound open: the exceptions 224
10.10 Dressings 225
10.11 Anti-tetanus, antibiotics, and analgesia 226
10.12 Post-operative care 226

Chapter 11 DELAYED PRIMARY CLOSURE AND SKIN GRAFTING 229
11.1 Delayed primary closure 231
11.2 Skin grafting 233
11.3 Full-thickness grafts 238
11.4 Healing by secondary intention 240

Chapter 12 NEGLECTED OR MISMANAGED WOUNDS 243
12.1 General considerations 245
12.2 Chronic sepsis: the role of biofilm 247
12.3 Surgical excision 248
12.4 Antibiotics 250
12.5 To close or not to close? 251

Chapter 13 INFECTIONS IN WAR WOUNDS 253
13.1 Contamination and infection 255
13.2 Major bacterial contaminants in war wounds 256
13.3 Major clinical infections of war wounds 257
13.4 Antibiotics 263
13.5 Neglected or mismanaged wounds 265
Annex 13. A ICRC Antibiotic protocol 266

Chapter 14 RETAINED BULLETS AND FRAGMENTS 269
14.1 The surgeon and the foreign body 271
14.2 Early indications for removal 271
14.3 Late indications 274
14.4 Technique for the removal of a projectile 275

Chapter 15 BURN INJURIES 277
15.1 Introduction 279
15.2 Pathology 279
15.3 Burn management 282
15.4 Burns presenting late 286
15.5 Nutrition 286
15.6 Care of the burn wound 286
15.7 Closure of the burn wound 289
15.8 Scar management 293
15.9 Electrical burns 294
15.10 Chemical burns 294
Annex 15. A Nutrition in major burns: calculating nutritional requirements 297

Chapter 16 LOCAL COLD INJURIES 299
16.1 Physiology of thermal regulation 301
16.2 Types of local cold injuries 301
16.3 Management 302

Chapter 17 ANAESTHESIA AND ANALGESIA IN WAR SURGERY 305
17.1 Introduction 307
17.2 Anaesthesia methods 308
17.3 Local and regional anaesthesia 309
17.4 Dissociative anaesthesia with ketamine 310
17.5 Post-operative pain management 312
Annex 17. A ICRC pain management protocols 314

Chapter 18 DAMAGE CONTROL SURGERY AND HYPOTHERMIA, ACIDOSIS, AND COAGULOPATHY 319
18.1 Resuscitative surgery and damage control surgery 321
18.2 Hypothermia, acidosis, and coagulopathy 322

ACRONYMS 329

SELECTED BIBLIOGRAPHY 333

INTRODUCTION

Our common goal is to protect and assist the victims of armed conflict and to preserve their dignity. This book is dedicated to the victims of situations which, in a better world, would not exist.

Facing the challenges

One night while on duty Dr X, an experienced surgeon working in an ICRC field hospital in the midst of a civil war, performed a craniotomy on one patient injured by a bomb, an amputation following an anti-personnel landmine injury on another, and a laparotomy after a gunshot wound on the third; not to mention the emergency Caesarean section that arrived, as always, at the most inopportune time, after midnight. She was the only surgeon available that night. This was common practice at the time, and not very much has changed in the last thirty-odd years.

Standard peacetime health services are already limited or lacking in many low-income countries, and faced with the added burden of weapon-wounded they are quickly overwhelmed. A precarious healthcare system is one of the first victims of armed conflict: the disruption of supply lines, the destruction of premises and the flight of medical personnel are all too common.

The lack of adequate resources is not limited to diagnostic and therapeutic technologies. Above all there is a dearth of human resources. Surgeons trained to practise in multidisciplinary teams find themselves alone to face the entire surgical workload and deal with subspecialties with which they have, at best, only a passing acquaintance. Reverting to the philosophy, so common 50 years ago, of the multidisciplinary single surgeon who has to "do it all" is not an easy task.

ICRC teams usually include only one or two surgeons. They are generalists, able to treat all kinds of injuries from simple soft-tissue wounds to penetrating abdominal and head injuries and complicated fractures. They must also provide emergency non-trauma surgical and obstetric care for the civilian population in the area. Ideally, they should be very general surgeons with a broad approach and wide experience.

The principles of war surgery have been known for centuries but need to be relearned by each new generation of surgeons and in every new war.

This commonly held view remains true to this day. Whether performed by military or by civilian surgeons, war surgery has its particular characteristics that are due to the special nature of the context of armed conflict, its limitations and dangers, and the particular physio-pathology of high-energy, penetrating missile and blast wounds. The care of weapon-wounded patients follows accepted surgical standards, but is performed under extreme conditions, which is why the management of a gunshot wound due to criminal violence in a civilian context cannot be easily extrapolated to surgery in a situation of armed conflict.

Where only limited resources are available the surgeon must accept that he cannot fully utilize his capacities and expertise.

Working with limited resources means that the limits of surgery that can be performed are not the expertise of the surgeon, but rather the level of anaesthesia and post-operative nursing care, and the availability of diagnostic and therapeutic equipment.

Limited resources, even in peacetime, may lead to the death of patients who would have survived had more sophisticated means been available. This is often the case in remote, and not so remote, hospitals in low-income countries; a situation exacerbated during armed conflict.

When the principles of triage are applied, saving "life and limb" for the greatest number, with the least possible expenditure of time and resources, often takes precedence.

Indeed, these characteristics mean that war surgery is very different from that practised in peacetime, when most operations are elective and most trauma is blunt, and the surgeon concentrates on doing everything he possibly can – using the full range of resources necessary – for each and every patient.

Furthermore, the work of medical staff in a situation of armed conflict is governed by a special set of rules, in addition to standard medical ethics: international humanitarian law, or the law of war. This is yet another specificity of this type of surgical care and is important for the security of patients and medical personnel alike living and working under dangerous circumstances.

International humanitarian law – the law of war – complements medical ethics in times of armed conflict or other situations of violence.

The ICRC's experience

The ICRC has been providing medical care for the war-wounded ever since its inception, for example during the Franco-Prussian War (1870). The 1970s and 80s, however, saw a tremendous increase in the already considerable humanitarian activities for the victims of war, armed conflict, and other situations of violence. These included relief efforts for refugees, internally displaced persons, and the affected resident population, and medical care for the sick and wounded. In addition, many new organizations were founded and, together with United Nations agencies, they deployed renewed efforts to respond to these humanitarian challenges.

The ICRC embarked on large-scale programmes to provide surgical care for the victims of war. Several independently-run ICRC hospitals were established and surgical staff recruited from various national Red Cross/Red Crescent Societies and Switzerland. A large number of enthusiastic and idealistic surgical staff set off on humanitarian missions. The surgeons were well-trained and experienced, but their training and experience were largely confined to sophisticated hospital facilities in industrialized countries. They faced a steep learning curve.

The ICRC also faced a steep learning curve and has acquired considerable expertise in caring for the victims of conflict in settings where the healthcare system is severely disrupted. This know-how derives from three different, but related programmes in various countries afflicted by armed conflict and other situations of violence around the world.

1. Independent ICRC-run hospitals.
2. Support to local hospitals through the short-term presence of expatriate surgical teams, with a strong focus on training and capacity building; the provision of supplies and equipment; the renovation of infrastructure and water and sanitation facilities; and financial incentives and salaries for local staff when necessary.
3. Organization of war surgery seminars, which provide opportunities for colleagues to exchange experiences and expertise.

This three-pronged approach has enabled the ICRC to develop basic clinical protocols and procedures for surgical techniques appropriate to the management of weapon-wounded patients in situations of limited resources and precarious circumstances. In addition, over the last thirty years, the ICRC has trained and maintained a pool of experienced hospital personnel who do not have to re-invent everything with each new conflict.

However, thanks to more widespread educational opportunities in recent years, there has been a sharp increase in the number of surgeons in conflict-afflicted countries. This has allowed the ICRC to switch the emphasis of its programmes from one of substitution – i.e. independent ICRC hospitals – for a poorly-functioning or non-existent hospital system to one that focuses on support and training of medical personnel in the niceties of the management of patients wounded by the weapons of war.

As part of its training programmes, the ICRC has co-organized more than 120 war surgery seminars – over a dozen a year – in the last decade. On these occasions much expertise and many ideas have been exchanged between surgeons from a number of countries – with a varying degree of experience in war trauma – and ICRC surgeons. We have all learned from these discussions, and some of the lessons are reflected in the contents of this new book.

However, in several contexts the ICRC has continued to provide direct surgical services in a neutral and impartial manner. This form of substitution contributes a fundamental element to the protection of the victims and the medical mission in situations where these humanitarian principles have been sorely tried.

While a number of manuals on war surgery have been published, they are mostly produced by and for the armies of industrialized countries. Their operational norms usually require large investments in means and personnel: helicopter evacuation of patients; well-trained field medics and stretcher-bearers; sophisticated technology; multidisciplinary teams of specialist surgeons, anaesthetists, and nurses. ICRC surgical staff refer to these manuals as references but the conditions and means described therein are seldom met in present-day theatres of armed conflict. Many of their "lessons" are inadequate – or even irrelevant – to humanitarian war surgery or the functioning of public hospitals in many countries working with limited resources.

ICRC surgical care aims to be economical, non-specialized, based on solid scientific principles, and offer good returns given the constraints. The clinical protocols and surgical techniques described in this manual are the standard procedures used among the ICRC's pool of experienced surgeons.

Putting pen to paper

To meet the challenge posed by these conditions, our predecessors in the surgical department of the Medical Division of the ICRC edited a basic reference manual for surgeons embarking on their first humanitarian mission: *Surgery for Victims of War*.

The first three editions of this book have been extensively distributed and received wide acclaim the world over from surgeons who faced the challenge of treating war-wounded patients for the first time. The general surgeon in an isolated rural hospital has perhaps drawn the greatest benefit from it.

Originally intended as a fourth edition of *Surgery for Victims of War* to address the particular needs and specific requests many of our colleagues raised during ICRC-led seminars, as well as to reflect the developments in ICRC surgical practice, it soon became clear that a new book would better serve this purpose. This book now comprises a significant amount of new material to be presented in two volumes, while maintaining the basic reasoning of the original manual.

This first volume is devoted to broad topics, with a number of entirely new chapters of a more general character whose contents are relevant not only to surgeons but also to those responsible for the organization and coordination of surgical programmes in times of armed conflict and other situations of violence. It presents the characteristics of surgical care for the victims of war, in particular the epidemiological, organizational and logistic aspects, drawing on the experience of ICRC medical staff and other colleagues. The second volume will deal with weapon-related trauma to specific body systems.

The surgical techniques presented herein share many fundamental ideas with more sophisticated medical services. However, they also derive from tried-and-tested improvisation and the use of very simple methods of treatment aiming to use technological means as appropriate as possible to the prevailing conditions of limited infrastructure, equipment, and human resources.

The explanations of techniques are geared to the level of knowledge and practice of *general surgeons in a rural hospital*. These surgeons are often the first to see patients wounded in conflict and they know that, under the circumstances, referral to more sophisticated facilities – far away in an inaccessible capital city – is impractical or impossible. This book attempts to provide surgeons who do not have specialized training with basic advice about the treatment of various weapon wounds, describing the types of operations that have proved successful in ICRC and other comparable practice.

The manual and other basic ICRC surgical publications and documents of general interest are included in the DVD attached to this volume. The disk also contains several downloadable files – forms and checklists for example – that can be used in everyday practice and adapted by the reader.

Unless stated otherwise masculine nouns and pronouns do not refer exclusively to men, for the manual is gender neutral. Any use of trade or brand names is for illustrative purposes only and does not imply any endorsement by the ICRC.

We hope that civilian and military surgeons, as well as Red Cross/Red Crescent surgeons, facing the challenge of treating the victims of armed conflict and other situations of violence for the first time under precarious and, at times dangerous, circumstances will find this book useful.

Marco Baldan

Philippa Parker
Head of Health Unit
ICRC Assistance Division

Christos Giannou
Former ICRC Head Surgeon

Marco Baldan
ICRC Head Surgeon

Acknowledgements

This manual is based on *Surgery for Victims of War*, first published by the ICRC in 1988 and edited by Daniel Dufour, Michael Owen-Smith, and G. Frank Stening. The authors included:
Bernard Betrancourt, Switzerland
Daniel Dufour, Switzerland
Ora Friberg, Finland
Soeren Kromann Jensen, Denmark
Antero Lounavaara, Finland
Michael Owen-Smith, United Kingdom
Jorma Salmela, Finland
Erkki Silvonen, Finland
G. Frank Stening, Australia
Björn Zetterström, Sweden
and was illustrated by Penelope L. Zylstra (Australia). Many of her line drawings have been included in the present book.

The second edition (1990) was revised by Robin Gray (United Kingdom) and the third (1998) by Åsa Molde (Sweden).

We owe a debt of gratitude to their pioneering work, and their clear and simple approach which has served as a model.

The present publication has benefited from the input of many colleagues with much experience within the ICRC and outside it. Critical comments and valuable advice throughout were provided by:
Ken Barrand, United Kingdom
Franco De Simone, Italy
Herman Du Plessis, South Africa
Jacques Goosen, South Africa
Åsa Molde, Sweden
Valery Sasin, Byelorussia
Harald Veen, Netherlands
Gunter Wimhoefer, Germany

Beat Kneubuehl (Switzerland) acted as the scientific adviser on ballistics and Sylvain Vité (Switzerland), an ICRC jurist, provided technical expertise on international humanitarian law and revised the appropriate sections. Massey Beveridge (Canada) served as technical adviser on burns and skin grafting and made significant contributions to the relevant chapters.

The chapter on the Red Cross Wound Score is largely based on the revised edition of the ICRC brochure by Robin Coupland (United Kingdom), who also provided essential comments and advice on ballistics and epidemiology, and has moreover played an important role through his many other pertinent publications. Holger Schmidt (Germany) and Eric Bernes (France) gave advice on first aid and emergency room trauma care. Haide Beckmann (Germany) and Thomas Walker (Switzerland) made contributions to the chapter on anaesthesia and Dieter Jacobi (Germany) provided comments for the chapter on chronic infections.

The ICRC Master Surgeons Workshop, held in Geneva in March 2002, revised the Red Cross Wound Score, ICRC triage categories, and established the ICRC antibiotic protocol. Participants included:
Marco Baldan, Italy
Massey Beveridge, Canada
Christos Giannou, Greece-Canada
François Irmay, Switzerland
Dieter Jacobi, Germany
Ben Mak, Netherlands
Valery Sasin, Belarus
Jukka Siegberg, Finland
Harald Veen, Netherlands
Gunter Wimhoefer, Germany

In addition, the ICRC Senior Anaesthetists Workshop, held in Geneva in November 2002, was the basis for the chapter on anaesthesia and established the ICRC pain management protocol. Its participants included:
Sunao Asai, Japan
Haide Beckmann, Germany
Lisa Bennett, Australia
Jeanne Frossard, United Kingdom
Christiane Gerber, Switzerland
Christos Giannou, Greece-Canada
Tuula Kangas-Saarela, Finland
Svante Linden, Sweden
Peter Mahoney, United Kingdom
Barthélémy Merzouga, Switzerland
Pascal Ollé, France
Erkki Saarela, Finland
Bernadette Sterckx, France
Vladislav Virago, Belarus
Eric Vreede, Netherlands
Jolanda Walker, Switzerland

Both workshops also helped define the ICRC criteria for the adoption of new technologies, the level of laboratory expertise required, and general strategies for ICRC surgical programmes.

Christiane de Charmant handled the editing of the final text and was responsible for the production while Pierre Gudel provided the graphic design. Their contribution is greatly appreciated.

The authors are ICRC staff members and no outside financial or material support has been received for the publication of this book.

Chapter 1

SPECIAL CHARACTERISTICS OF SURGERY IN TIMES OF CONFLICT

1. SPECIAL CHARACTERISTICS OF SURGERY IN TIMES OF CONFLICT

1.1 Differences between surgery in times of conflict and civilian practice 19

1.2 How war surgery differs 20
1.2.1 IHL: protection of non-combatants and those *"hors de combat"* and the rights and obligations of medical personnel 20
1.2.2 Specific epidemiology of war wounds 21
1.2.3 Predominance of emergency surgery 21
1.2.4 Surgery in a limited technical environment 21
1.2.5 Surgery in a hostile, violent environment 21
1.2.6 Mass casualties involving the principles of triage 22
1.2.7 Triage and surgery in successive echelons 22
1.2.8 Outcome of hospital patient care depends on the efficiency of pre-hospital echelons 23
1.2.9 Specific wound pathology: bullets, bombs, blast weapons and non-conventional weapons 24
1.2.10 Specific techniques appropriate to the context and pathology 24
1.2.11 Increased prevalence of endemic disease 25

1.3 "Surgeries" for victims of war 25

1.4 Differences between military and non-military war surgery: the ICRC approach 27
1.4.1 Military-civilian cooperation 27
1.4.2 Constraints: security 27
1.4.3 Constraints: logistics 28
1.4.4 Constraints: hospital equipment 28
1.4.5 Constraints: blood transfusion 29
1.4.6 Constraints: geography and climate 29
1.4.7 Constraints: culture shock 29
1.4.8 Constraints: the human factor 30

ANNEX 1. A ICRC criteria for introducing a new technology 31

1.1 Differences between surgery in times of conflict and civilian practice

The differences between civilian and war trauma are manifold – as are the differences between the experience of the ICRC and those of conventional military medical services.

Most surgeons today, the world over, derive their trauma training from road-traffic accidents and much that applies to the management of casualties in civilian settings will also apply to the situation of armed conflict: war surgery follows classical surgical standards. However, the generation of surgeons who mostly had to deal with accidents among agricultural or industrial workers were well aware of the dangers of gas gangrene and tetanus and the necessity of good wound excision and delayed primary closure. It was relatively easy for the surgeon to shift from this "septic" civilian surgery to war wounds. This is no longer the case for many surgeons trained today; laparoscopy, radioscopic embolization, and unreamed intramedullary nails will not get you very far when facing a landmine injury of the abdomen or a machine-gun wound to the thigh. Early specialist training of surgeons and sophisticated modern technology benefit many patients in a peacetime environment, but can be an obstacle to the practice of surgery during armed conflict.

War wounds are different. The extent of tissue destruction and contamination seen in war injuries is nothing like what is seen in everyday trauma practice. Working conditions during war are radically different from those prevailing in peacetime. Resources are limited and surgeons are often obliged to improvise or make compromises in their management decisions. Their aim should be to bring the best care possible to their patients under the circumstances, not the best care as described in the academic literature.

War surgery is a surgery of mass casualties. The logic of war triage has little to do with the routine emergency department triage of a major civilian trauma centre: war triage has a "leave to die in dignity" category unheard of in everyday civilian practice.

War surgery involves the staged surgical management of the wounded, often at different echelons of care and provided by different surgeons, especially in a military context. Even in a humanitarian context, such as ICRC surgical programmes, several surgeons deployed on short missions may participate in the treatment of a single patient. In everyday civilian practice on the other hand, the same surgeon assumes responsibility for the entire surgical management of his or her patient. While modern civilian practice often involves a "multidisciplinary approach", war surgery often demands a "multi-surgeon" approach.

> "All the circumstances of war surgery thus do violence to civilian concepts of traumatic surgery."
>
> Michael E. DeBakey[1]

These and other challenges mean that practitioners facing surgery for the victims of war for the first time will have to change their mindset, i.e. their "professional mental software".

---

[1] DeBakey M E. Military surgery in World War II – a backward glance and a forward look. *NEJM* 1947; **236**: 341 – 350. Michael E. DeBakey (1908 – 2008), an American surgeon of Lebanese origin, was a pioneer of modern cardio-vascular surgery. His treatise on the management of vascular trauma constitutes a basic reference. He invented the concept of Mobile Army Surgical Hospitals (MASH) for the US armed forces in Korea.

1.2 How war surgery differs

> War surgery is the management of an "epidemic of trauma" in a series of steps: echelons.
>
> N. I. Pirogov[2]

A number of special features characterize the practice of surgery in time of war.[3]

1. Special rules: international humanitarian law (IHL), i.e. the protection of the sick and wounded, and the rights and obligations of medical personnel.
2. Specific epidemiology of war wounds.
3. Predominance of emergency surgery.
4. Surgery in a limited technical environment.
5. Surgery in a hostile, violent environment: constraints of the tactical situation.
6. Mass casualties involving the principles of triage.
7. Triage and surgery in successive echelons of casualty care.
8. Outcome of hospital patient care is a function of the efficiency of pre-hospital echelons.
9. Specific wound pathology: bullets, bombs, blast, non-conventional weapons.
10. Specific techniques appropriate to the context and pathology.
11. Increased prevalence of endemic disease.

1.2.1 IHL: protection of non-combatants and those "*hors de combat*" and the rights and obligations of medical personnel

The Geneva Conventions of 1949 and their Additional Protocols of 1977 define categories of individuals who, by virtue of these treaties, are protected during armed conflict. These include non-combatants, combatants who no longer participate in hostilities – "*hors de combat*" – either through sickness, injury, shipwreck, or by becoming prisoners of war; and those who care for the sick and injured, namely medical and religious personnel. The latter two use the protective emblem of the red cross, red crescent or red crystal to mark the means and facilities that take care of the sick and wounded. According to the law, all these categories of protected individuals are immune from attack, as long as they do not take an active part in hostilities. International humanitarian law – the law of war – offers particular rights to medical personnel, but also ascribes obligations to them.

All health professionals are bound by medical ethics, in times of peace and of war. These are not replaced but are complemented by international humanitarian law. Compliance with these laws may create particular ethical dilemmas and security problems, and the military hierarchy does not always understand the demands of medical ethics. Civilian health staff may face particularly problematic and dangerous situations during a civil war where their own community is involved in the conflict. Chapter 2, Applicable international humanitarian law, explains the main principles and rules governing the rights and obligations of medical staff in times of armed conflict.

ICRC

Figure 1.1

Soldiers "*hors de combat*": POWs.

Japanese Red Cross Society

Figure 1.2

Soldiers "*hors de combat*": the wounded.

ICRC

Figure 1.3

Those who care for the sick and wounded.

---

[2] Nikolai Ivanovich Pirogov (1810 – 81): Russian anatomist and surgeon, professor at the Academy of Military Medicine, St Petersburg. Founded modern field surgery during the Crimean War (1854); devised the plaster cast and the mass use of anaesthesia on the battlefield and wrote a reference field surgery manual. Representative of the Russian Red Cross sent to inspect the hospitals on both sides during the Franco-Prussian War in 1870.

[3] List modified and adapted from bibliographic sources.

1.2.2 Specific epidemiology of war wounds

The nature of warfare – on land, at sea or in the air – will create a particular epidemiology of the wounded. The nature of weapons, protective body armour, and any delay in transport will affect the anatomic distribution of injuries and their severity. The understanding of these epidemiological factors will have important consequences in terms of preparation and allocation of resources: i.e. standardized supplies and specialized personnel (see Chapter 5).

1.2.3 Predominance of emergency surgery

War surgery primarily consists of emergency surgery, especially during early tactical field care. Sophisticated techniques or reconstructive procedures have no place here, except well after combat and in a distant referral hospital (see Chapters 6 and 8).

1.2.4 Surgery in a limited technical environment

The environment in times of war is bleak and harsh. The limits of surgical work are largely determined by the logistic difficulties attending the supply of remote and dangerous areas and the lack of maintenance, repair and spare parts. There is seldom enough technical support staff to ensure that infrastructure functions correctly.

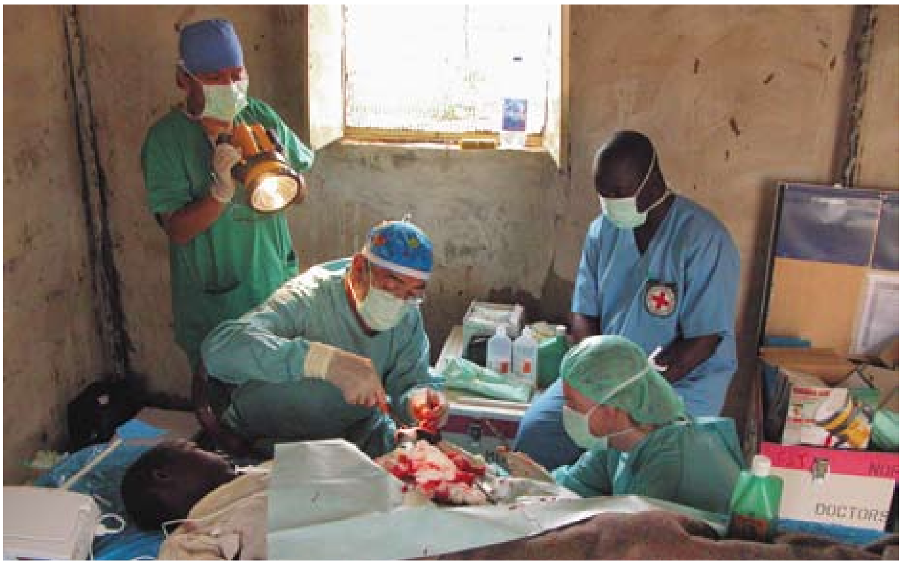

V. Louis / ICRC

Figure 1.4

Surgery in a limited technical environment.

Despite lavish outlays for field hospitals by the armies of modern industrialized countries, equipment limitations in tactical situations are well recognized. The lack of sophisticated diagnostic equipment rather than the surgeon's technical capabilities and expertise is often what determines what can be done. What is essential must be differentiated from what is "nice to have".

1.2.5 Surgery in a hostile, violent environment

The adverse conditions of a tactical situation may put the safety of patients and medical personnel at risk, and thereby create less than optimal working conditions. Dangerous evacuation routes may compromise transport and result in delays. Security must be ensured for patients and staff, by selecting suitable sites for first-aid posts and hospitals. Health facilities and ambulances should be clearly marked with the protective emblem of the red cross, red crescent or red crystal, according to the dictates of international humanitarian law.

Not all combatants are disciplined and well trained. For anyone caught up in armed conflict, there is an all too familiar syndrome seen amongst young fighting men who are under the influence of a "toxic cocktail": testosterone, adrenaline, alcohol and cannabis (and, at times, "other substances" as well).

T. A. Voeten / ICRC

Figure 1.5

Working in a hostile environment.

The setting of surgical work can change rapidly, and surgeons must be prepared to adapt to various difficult and austere environments. With few doctors and limited staff available, and wounded people pouring in, hospital facilities are easily overwhelmed. Doctors and nurses also get tired and sick, and are sometimes frightened. The physical and mental strain of working in new and strange circumstances and at times dangerous conditions means that they may not perform as well as usual.

R. Bigler / ICRC

Figure 1.6

Mass casualties: the principles of triage.

1.2.6 Mass casualties involving the principles of triage

Much has been written about the carnage of the major wars of the twentieth century, and the role of triage in the management of thousands of casualties resulting from a single battle. These lessons are still relevant in contemporary conflicts. The logic followed must be to "do the best for the most" and not "everything for everyone". This involves the most important change in the professional mindset of the surgeon.

Triage decisions are amongst the most difficult in all medical practice, possibly creating ethical dilemmas. There may also be, at times, a conflict between medical criteria and tactical ones of military necessity that require some form of compromise. Persons performing triage must be prepared to accept these necessary compromises while maintaining their personal medical integrity (see Chapter 9).

1.2.7 Triage and surgery in successive echelons

The sick and wounded are evacuated and transferred along a chain of casualty care. The principles of triage are applied at every stage in this chain. The initial surgery must not compromise the performance of later, definitive surgery. The prognosis will be much better if the wounded are evacuated rapidly to a higher echelon hospital structure. A surgeon in the field must understand the system and know what will happen to the patient at the next echelon of care and what is required of him at his echelon (see Chapter 6).

War surgery requires a logic of phased wound management. Treatment takes place in five phases often, although not necessarily, in five different places. This is the classical set-up in military planning. It involves high costs, including the means of transportation and the discipline required of such organization. Modern concepts can be integrated into this approach, such as forward surgical teams performing damage-control surgery. The five phases are as follows.

1. First aid at the point of wounding: self- or "buddy" treatment, or that provided by a field medical officer or first aider.
2. First medical treatment: vital emergency measures including the beginning of resuscitation, generally at the first-aid or dressing post, clearing station or collection point.
3. First surgical treatment: debridement/wound excision, without primary closure, at the first echelon hospital.
4. Definitive treatment: delayed primary closure of wounds and surgical treatment according to traditional principles in a referral hospital. Physiotherapy and convalescence.
5. Reconstructive surgery and rehabilitation: specialized surgery with multiple reconstructive procedures and the fitting of prostheses when required.

The rapid turnover of medical personnel treating numerous patients at different points in a chain of casualty care creates a necessity for standard protocols, which cannot be left to the discretion or whim of the individual surgeon. One cannot individualize treatment for each single patient in a tactical situation; different surgeons will operate on the same patient at the different echelons. In civilian practice, by contrast, the same surgeon manages the surgical care of the individual patient throughout that patient's course of treatment; even in a multidisciplinary approach, the same team is involved.

These five echelons for the management of the war-wounded do not always exist in a civilian or humanitarian context, where all five phases may well occur in the same medical facility – as is in fact common practice for ICRC hospital teams. However, in the humanitarian context, there is also a rapid turnover of personnel. Standardized protocols are the only means to ensure continuity of treatment and to organize surgical and nursing care on a sound footing; the protocols cannot be changed with every change of team surgeon.

Old lessons for new surgeons

Save life and limb, sacrifice limb to save life, prevent infection, and render the casualty transportable to the next echelon.

Heroic surgery will never replace good surgery.

1.2.8 Outcome of hospital patient care depends on the efficiency of pre-hospital echelons

First aid provided at the point of wounding – or at the safest place near the battlefield – and rapid evacuation are of vital importance; mortality and morbidity increase with delay. If first aid is inadequate or unavailable and the evacuation chain is long, then the outcome will be decided by nature. However, hospital mortality decreases as evacuation time increases; with very long delays, the severely injured die long before reaching hospital and surgeons spend most of their time dealing with the septic complications of the survivors (see Chapters 5 and 7).

Old lessons for new surgeons

It is more important to provide proper first aid and render the injured fit for transport, than to give early but inadequate treatment, with insufficient means, and insufficient knowledge.

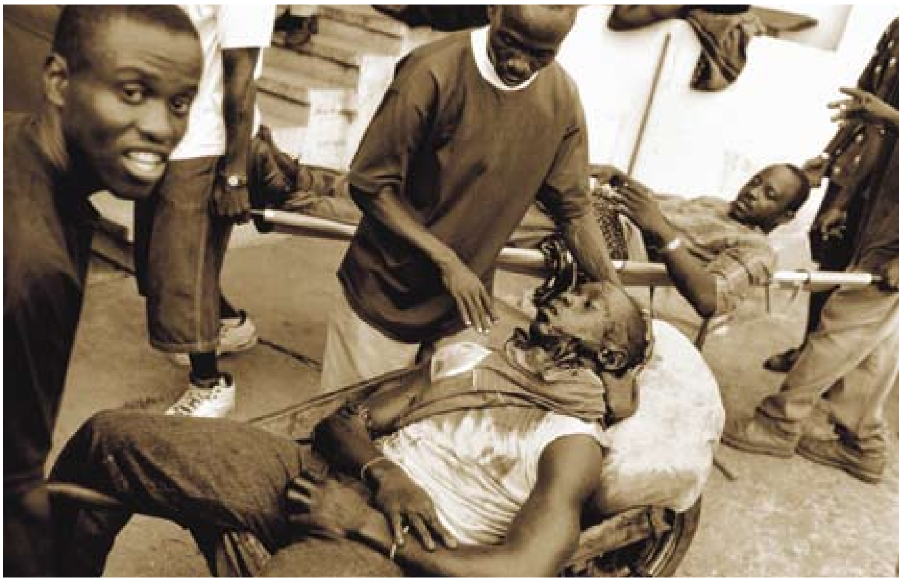

T. A. Voeten / ICRC

Figure 1.7

Inadequate pre-hospital care.

1.2.9 Specific wound pathology: bullets, bombs, blast, and non-conventional weapons

War wounds are qualitatively different from trauma seen in civilian practice: all are dirty and contaminated. Projectiles may cause massive destruction of soft tissues, bones, and important organs. Infection is the great danger and the rules of septic surgery apply (see Chapters 3 and 13).

Old lessons for new surgeons

War wounds are dirty and contaminated from the moment of injury. The rules of septic surgery apply.

Nothing in routine civilian practice is similar to the polytrauma of multiple shell fragments, traumatic amputation due to an anti-personnel landmine, or the devastating effects of a high-kinetic energy transfer from a military rifle bullet. Again, a change in the professional mindset appropriate to a modern trauma centre is required in order to adjust to the management of wounds in times of armed conflict. Surgeons practising in a low-income country, dealing with much septic pathology, will find this change much easier to handle than those whose usual work involves sophisticated technology with abundant good nursing care in an aseptic environment.

1.2.10 Specific techniques appropriate to the context and pathology

The care of numerous patients, treated in many locations by different surgeons, in austere conditions demands simplicity, security, and speed of surgical procedures. The need for speed in dealing with mass casualties with inadequate numbers of staff should not lead to confusion and disorder. Phased wound care imposes standards and a systematic approach: doing the least amount of surgical work for the greatest results, saving "life and limb", and then sending the patient on to the next stage in the chain of casualty care. The follow-up of surgical procedures by the individual surgeon in the field is difficult if not impossible, and precludes personal and individualistic techniques. Standard protocols, as mentioned, are necessary.

The majority of wounds will be to the extremities, and the objective is to treat them so that they heal as quickly as possible without becoming infected. Sepsis, potentially fatal (tetanus, gas gangrene, haemolytic septicaemia), is the greatest danger for survivors. As aforesaid: the rules of septic surgery apply.

The basic principles of management of war wounds comprise the following steps.

1. Early and thorough wound excision and irrigation.
2. No unnecessary dressing changes.
3. Delayed primary closure.
4. Antibiotics as an adjuvant.
5. Anti-tetanus vaccine and immunoglobulin if necessary.
6. No internal bone fixation.[4]
7. Early physiotherapy.

---

[4] No internal bone fixation is the rule, at least in the "acute phase". Recent experience has shown the possibility of using internal fixation once the soft tissues have healed in the absence of any infection, but only in skilled hands and with excellent conditions of hygiene and nursing care. This is not, however, standard policy in ICRC practice where no internal fixation whatsoever is used. The risk, and probability, of abuse of internal fixation precludes its availability.

Old lessons for new surgeons

The best antibiotic is good surgery.

Correct surgery gives the patient the best chance of survival with a good quality of life, and shortens the stay in hospital. High-quality physiotherapy is required to ensure early mobilization after surgery and a good functional result. Treatment is not complete, however, until the patient is rehabilitated; prosthetic workshops are needed to fit amputees and provide other suitable devices, such as orthoses, crutches or wheelchairs.

1.2.11 Increased prevalence of endemic disease

Until World War I, more soldiers died of disease than of their wounds. Battle-injury attrition was usually around 20%, and disease four times more common among soldiers. Non-battle injury attrition rates remain very high even today; infectious and communicable diseases differ according to the geography and climate, but psychological disorders and vehicle accidents are universal.

The destruction, disruption and disorganization that accompany armed conflict often cause the public health system to be among the very first to suffer. The humanitarian consequences translate into reduced availability of basic public health requirements such as water, food, shelter, etc., for the civilian population and often overwhelm the capacities of the public health system. This compounds the challenge of caring for the civilian population in a conflict zone: i.e. residents, internally displaced persons, and refugees fleeing to or from neighbouring countries (see Chapter 5).

1.3 "Surgeries" for victims of war

There is more than one type of war surgery. While the needs of the wounded are the same, the means and resources available to meet those needs vary widely from country to country and situation to situation, giving rise to different approaches to war surgery. The management of the war-wounded as performed by the military medical services of an industrialized country is not the same as that of a public rural hospital in a low-income country. Although the principles of wound management are the same in both cases, the diagnostic and therapeutic possibilities are very different. The latter should be *appropriate* to the technological, financial, and human resources at hand. Obviously many of these same constraints apply to the practice of civilian, everyday trauma and elective surgery around the world.

At least four major scenarios for the surgical management of victims of war in contemporary armed conflicts can be described.

1. Conventional army of an industrialized society, with a high level of public financing and where the military attempt to provide the same level of surgical care that is practised in civilian life. Rapid evacuation and transfer of patients to specialized facilities is routine. Safe access to adequate medical care for the sick and wounded is perceived as a right; and the duty to provide it incumbent on the armed forces.

2. Developing country with an emergent economy where, at least in the capital and other major cities, a high-level of specialized surgical care including enough trained personnel is available, even if this is not the case in rural areas. Evacuation and transfer of patients to specialized facilities is possible, if sometimes difficult. Safe access to adequate medical care is a goal to be achieved.

3. Poor country, with limited financial and human resources. A few major surgical centres may exist in the capital city but provincial and rural hospitals are largely staffed by young general surgeons or general practitioners with some surgical experience. Supplies, budgets, and human resources are chronically inadequate or even absent. Evacuation and transfer of patients are difficult or impossible. Safe access to adequate medical care is rarely guaranteed.

4. Non-State actors, guerrilla groups, populations without safe access to public structures. Field surgery is practised by a few trained doctors and nurses, because there is no alternative. Safe access to the victims by health professionals, and the victims' access to medical care, is impossible or rare, problematic, and always a challenge.

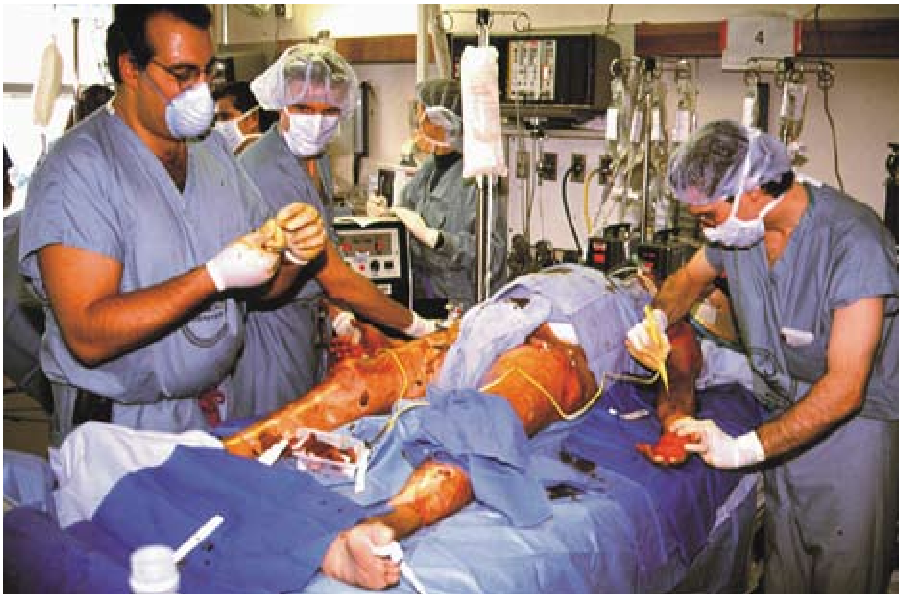

H. Du Plessis / South African Military Health Service, U. Pretoria

Figure 1.8

Surgery for victims of armed conflict in a modern hospital.

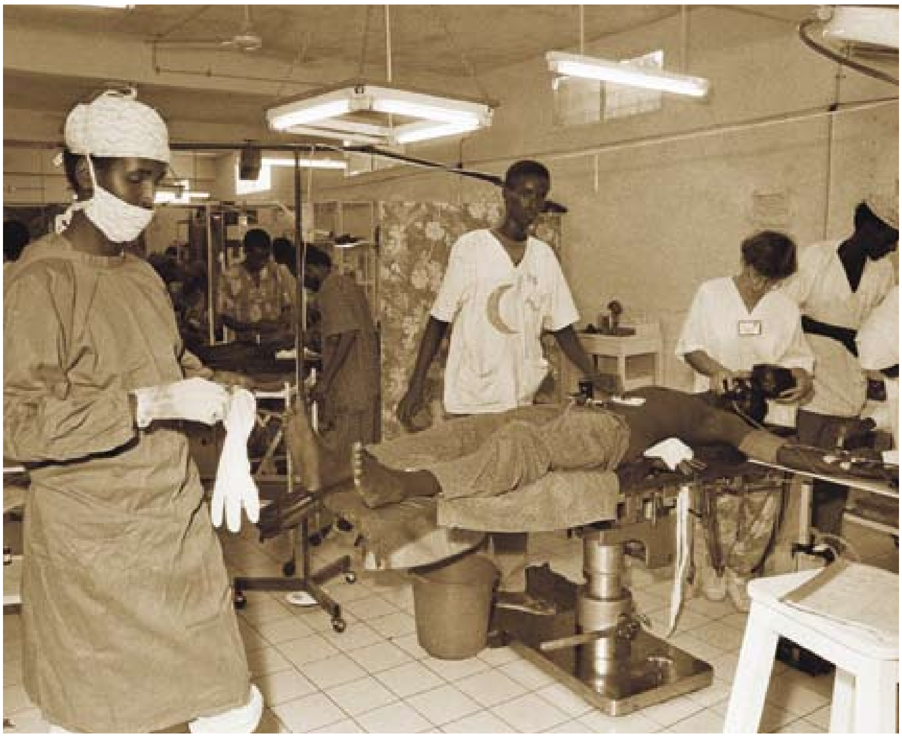

F. McDougall / ICRC

Figure 1.9

Another type of surgery for victims of armed conflict.

1.4 Differences between military and non-military war surgery: the ICRC approach

Non-military war surgery is practised by civilian medical structures (health ministry, missionary and private hospitals) and those of the ICRC or other humanitarian agencies. This section details the ICRC's experience and explains its approach.

The aims of war surgery for the ICRC are to protect the sick and wounded and help maintain their dignity by ensuring access to adequate care; save "life and limb"; minimize residual disability and assist amputees. Besides the direct victims, the ICRC also attempts to support the health system by assisting local medical colleagues to maintain the infrastructure and human resources necessary to resume functioning after the end of the conflict, thus providing the civilian population with at least the basics of healthcare. ICRC assistance to local structures may involve construction and renovation of health facilities, water and sanitation work, food supplements for patients and staff, equipment, supplies and basic salaries. Training programmes for doctors and nurses may be included. In addition, the ICRC may at times set up its own independent hospitals with expatriate staff supplemented by local personnel (see Chapter 6).

1.4.1 Military-civilian cooperation

Armed forces deployed in the field have a specific military mission. Their medical component aims primarily to give support to their soldiers in this mission, i.e. to achieve their tactical and strategic plans. The military may have many of the same "assistance and reconstruction" aims as civilian organizations, but medical criteria in a military context often must take second place to the tactical and strategic demands of military and political necessity.

The ICRC is a neutral, impartial and independent, totally humanitarian, institution. It promotes adherence to international humanitarian law and aims to protect and assist the victims of conflict – all the victims on all sides. Any cooperation with armies in the field of battle that compromises the perception of that neutrality, impartiality or independence can only compromise the humanitarian work of the ICRC, as well as that of other organizations and agencies.

The ICRC insists on maintaining its independence, and the independence of a "humanitarian space" separate from the "assistance and reconstruction" work of armies in the field. Many humanitarian organizations share the same opinion and approach.

1.4.2 Constraints: security

The ICRC often has little or no control over casualty evacuation because of security constraints. In many countries the wounded are transported by private means: taxis, donkeys, oxcarts or on foot. In some contexts, the ICRC has been able to set up first-aid posts, or assist a National Red Cross or Red Crescent Society in doing so. One extraordinary example was a 16-year medical evacuation system by fixed-wing aircraft run by the ICRC and United Nations' Operation Lifeline Sudan to assist victims of the conflict in southern Sudan: it transported over 30,000 sick and wounded patients to an ICRC hospital in northern Kenya. Even with aircraft available, the logistics and distances involved regularly resulted in a one- to three-week delay in evacuation.

While the armed forces deploy the necessary means to protect their health facilities from the "lethal chaos of the battlefield",[5] the ICRC depends on the red cross emblem and negotiations with all belligerent forces for its physical protection. The ICRC has no guns to protect itself and is just as dependent on local authorities and leaders as public hospitals are. The ICRC relies on the limits imposed by international humanitarian law and the discipline of combatants, and its diplomatic negotiating ability. Other humanitarian organizations who work in war zones face similar constraints.

---

[5] Butler F. Tactical Combat Casualty Care: combining good medicine with good tactics. *J Trauma* 2003; **54 (Suppl.)**: S 2 – 3.

1.4.3 Constraints: logistics

A hostile environment presents more than security risks. Remote areas with dangerous routes and extreme climates pose numerous logistic problems for the delivery of supplies and the maintenance of basic infrastructure, for both hospital and living quarters. The military often have lift, delivery and transport capacities that civilian institutions lack. Although the armed forces also have their logistic limits, they are of a different order of magnitude to those of the health ministry, non-governmental organizations, or the ICRC.

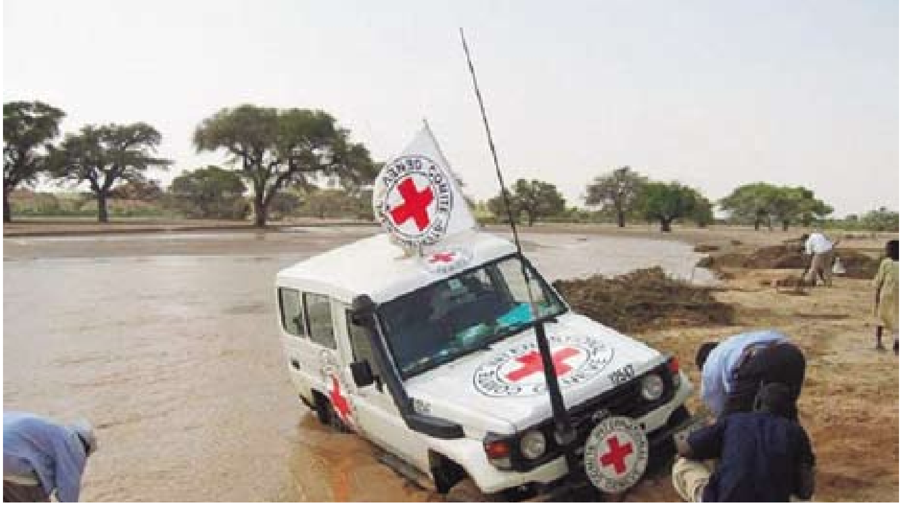
V. Mboah / ICRC

Figure 1.10
Challenging environment.

1.4.4 Constraints: hospital equipment

These limitations also affect hospital equipment. The military have lift constraints because they must also transport arms and munitions. For the ICRC, equipment limitations mean resorting to appropriate technology and mastering the tasks of maintenance, repair and availability of spare parts. This is particularly important when working in remote areas in a poor country with training programmes for local colleagues. The aim is to avoid creating a technological dependency that the permanent local staff will not be able to cope with when the conflict is over, and the ICRC leaves the hospital and the country. To this end, the ICRC has developed criteria for the introduction of any new technology into its standard list of medicines and equipment for hospital programmes (see Annexe 1. A: ICRC criteria for introducing a new technology).

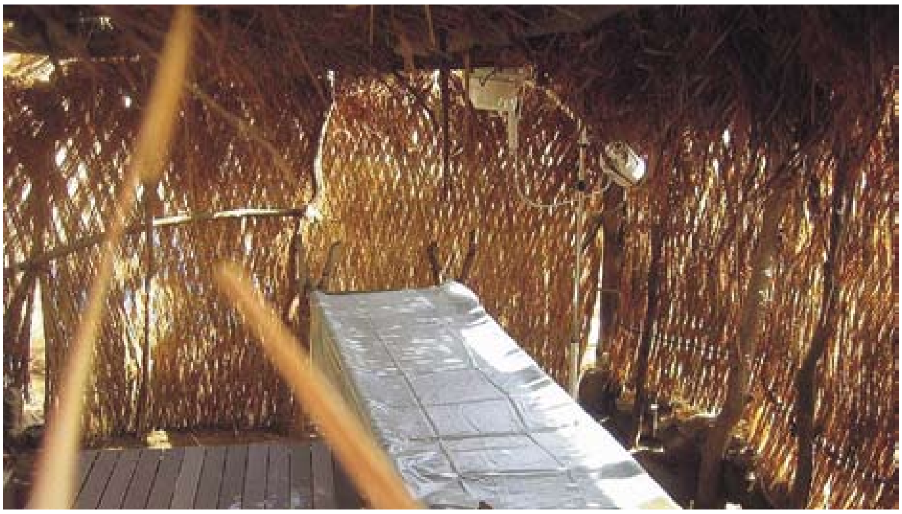
K. Barrand / ICRC

Figure 1.11
Limited surgical equipment.

A standard list is a basic, limited set of essential supplies, both medical and non-medical, which corresponds to an appropriate standard of care. All items on a standard list should always be available from a central store or reliable supplier. Standardization provides a simple framework, within which resources can be used to maximum effect, promotes continuity in patient care, helps to simplify staff training, and makes it easier to introduce new and inexperienced staff into the system. The ICRC and the International Federation of Red Cross and Red Crescent Societies have established an Emergency Items Catalogue[6] with predetermined sets and kits that run the gamut of assistance programmes.

ICRC experience shows that it is possible to perform good quality surgery with basic technology such as simple X-ray, and electronic monitoring in the theatre and post-operative ward limited to a pulse oxymeter. An ICRC laboratory set-up is also basic; there is no capacity for bacteriological culture and sensitivity and no blood components.

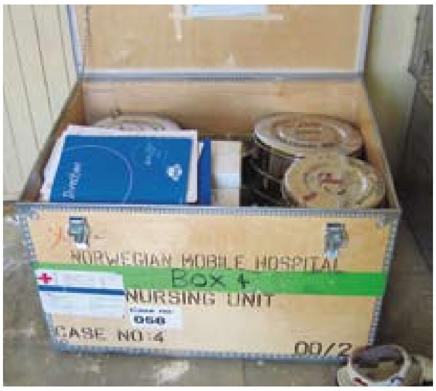

K. Barrand / ICRC

Figure 1.12

Basic supplies.

1.4.5 Constraints: blood transfusion

Blood for transfusion is often difficult to obtain in some countries because of cultural and religious restrictions and beliefs. With the steady increase in HIV infection around the world, testing must be adequate and indications for transfusion strictly limited. In some areas of the world, giving blood should probably be totally avoided. The use of blood should be restricted to vital needs and to patients with a good chance of survival, following the principles of triage. As in many, if not most, provincial and rural hospitals around the world, ICRC practice uses whole blood for transfusion, as fresh as possible, and usually donated by a relative.

1.4.6 Constraints: geography and climate

The geographic context may be important in terms of disease and the added burden it represents. The wounded may suffer from other illnesses, such as tuberculosis, malaria, typhoid and intestinal worms, as well as from malnutrition. In countries with chronic malaria infestation, there is often a peak of fever post-operatively. The surgeon must therefore try to acquire some basic knowledge of the diseases specific to the area and their treatment. Local healthcare workers are usually familiar with these conditions and more expert in their treatment than expatriate staff. These pathologies may also affect and represent a danger for expatriate staff.

1.4.7 Constraints: culture shock

Cultural constraints are another challenge that may add to the frustrations of medical work in a combat zone. In some societies, amputations and laparotomies can only be performed with the consent of the *family* of the patient. After a discussion in which the clear advantages are explained, the final decision must be left to the family. This procedure, which respects local cultural and social behaviour and norms, has to be followed and accepted even though it may be considered as a limitation and constraint by surgical and nursing staff. It is particularly difficult for committed medical personnel to see young people die because permission for necessary surgery has been denied.

In many societies, it is common for a relative to stay with a hospitalized patient, helping with nursing duties related to hygiene and feeding, and providing psychological support. This tradition must be accepted.

Adaptation to the cultural, social, and geographic context is essential.

---

[6] See Selected bibliography.

1.4.8 Constraints: the human factor

“War surgery is a surgery of complications, performed by doctors who are often ill-trained or without surgical training. It is surgery replete with adaptations and improvisations to replace that which is missing, a surgery of surprises that new means and methods of combat reveal.”[7]

This classic quotation from the Swiss army’s war surgery manual describes the situation for the new military surgeon and, as is more and more often the case in contemporary conflicts, for civilian surgeons faced with the victims of armed conflict for the first time. New ICRC surgeons coming from the National Red Cross or Red Crescent Society of a rich industrialized country face the same challenges: a constantly changing field of battle with new limits and constraints that come as a surprise. One must plan for alternatives and always keep an open mind to new options.

If military war surgery is the management of an “epidemic of trauma” in a series of echelons, this is not always the prevailing situation in non-military circumstances. Unlike a military field hospital, an ICRC hospital is responsible for all levels of medical care. It often acts as a first-aid post, field hospital, base hospital, and referral centre all in one. The “multi-surgeon” military approach to treatment in echelons gives way to a more traditional one of attending to the entire surgical history of a patient. However, since deployment of ICRC surgeons is usually for a short period of time (three months on average), a number of surgeons may be involved in caring for a single patient. Continuity of care is essential.

The modern military may “project forward” technical skills by deploying field surgical teams close to the battlefield. The aim is to perform critical surgery, often damage control surgery, as soon as possible after injury in an attempt to save lives, thereby reducing the number killed in action. The ICRC has also deployed field surgical teams – Somalia in 1992, southern Sudan in 2000, and Darfur in 2005 – but with a different aim: the protection of non-combatants and wounded fighters who no longer participate in hostilities and who would otherwise have no access to surgical care. The performance of this medical act and the protection of access to surgical care are based on the essential principles of the ICRC as a neutral and impartial humanitarian actor.

The surgeon must be able to adapt to the conditions of field surgery where “somewhat clean with soap and water” replaces a “sterile” environment and “favourite” surgical instruments are not available on the standard list. Furthermore, living conditions may resemble camping out in the bush and everyone in the team (4 members: surgeon, anaesthetist, theatre nurse and post-operative nurse) participates in the preparation of food and accommodation.

The following qualities are particularly needed by personnel working in ICRC hospitals or field surgical teams:
- professionalism;
- sound judgement and common sense;
- adaptability.

Local skills and material improvisation in some countries may bring to the attention of the surgeon efficient, cheap, and useful ways of treatment: mashed papaya for burns or autoclaved banana leaves as a non-adherent dressing, for instance. Expatriate personnel must show themselves capable of learning “new old tricks” and adapting to the circumstances. War surgery is a challenge and the work trying. Medical staff must be prepared – physically and mentally – for frustrations, fatigue and long hours, and being a witness to the results of “man’s inhumanity to man”.

War is deleterious to your health.

---

[7] War Surgery Commission of the Federal Military Department. *Chirurgie de guerre (Aide-mémoire 59.24 f) [War Surgery (a primer)]*. Bern: Swiss Army, 1970 and 1986.

ANNEX 1. A ICRC criteria for introducing a new technology

1. Needs assessment

What is the added value of this new technology? Are requested instruments "essential", "important" or "nice to have", or even "superfluous" and a luxury?

2. Maintenance requirements

What are the extra burdens for the daily maintenance of such equipment?

3. Ease of repair

Are specialized technicians necessary, and available?

4. Availability of spare parts

Is there a reliable local supplier?

5. Cost

Expense alone is not a criterion if the equipment is necessary. However, it is taken into account along with the other factors in a total cost/benefit analysis.

6. Competency required to use the technology in question

Is the expertise widely mastered and available or does it correspond to the particular practice of an individual doctor or nurse?

7. Continuity of the competency required

Can successive surgical teams use the equipment or does it depend on the expertise of a limited number of people?

8. Presence of the technology in the country

To all intents and purposes, the ICRC shall not be the first to introduce a new technology to a country; some local practice or competency must already exist.

9. Professionalism and ethical concerns

The supply of equipment and instruments must at all times meet demanding standards of professionalism in surgical care and address possible ethical concerns. (e.g. in Europe, following the outbreak of bovine spongio-encephalopathy – so-called "mad-cow disease" – catgut suture material has been banned by the European Union and Switzerland. It would not be ethical for the ICRC to continue to provide such sutures in its assistance programmes in other parts of the world, thereby creating a lower safety standard than in Europe.) Quality control of medicines and equipment is becoming a major worldwide problem as identified by the World Health Organization.

10. Sustainability

Only if the technologies can be sustained after ICRC withdrawal is it worth considering their utilization.

Chapter 2

APPLICABLE INTERNATIONAL HUMANITARIAN LAW

2. APPLICABLE INTERNATIONAL HUMANITARIAN LAW

2.1 Historical introduction 35

2.2 International humanitarian law: basic principles 36
2.2.1 Principles underlying IHL: the "law of war" 37
2.2.2 IHL and human rights law 38

2.3 The distinctive emblems 38

2.4 The International Red Cross and Red Crescent Movement and its Fundamental Principles 39

2.5 Rights and duties of medical personnel according to IHL 39

2.6 Responsibility of States 41

2.7 Reality check: some people do not follow the rules 43

2.8 The neutrality of a National Red Cross/Red Crescent Society 44

2.9 The role and mandate of the ICRC in situations of armed conflict 45
2.9.1 Health services: assistance to the war-wounded and sick 46

ANNEX 2. A The distinctive emblems 48

ANNEX 2. B The International Red Cross and Red Crescent Movement 50

2.1 Historical introduction

> "You may not be interested in war. War is interested in you."
>
> Leon Trotsky

One of the specificities of the war-wounded and those who care for them is their relationship to international humanitarian law, mainly the Geneva Conventions and their Additional Protocols.

In 19th century Europe, large-scale battles caused real carnage on the battlefields. Soldiers were regarded as cannon fodder and almost no medical services were available. Solferino, a town in northern Italy, was the site of one of these awful battles in 1859: in 16 hours 40,000 people were killed or wounded. Many wounded soldiers were left to die on the battlefield, although many of them could have been saved had relief been available. The medical services of the armies were too limited to respond to such a number of casualties – there were more veterinarians for the horses than doctors for the soldiers! Wounded soldiers lay agonizing for days without help.

Henry Dunant, a Swiss businessman travelling in the area, was struck by this drama. Out of compassion, he spontaneously organized help with the women from the neighbouring villages. Relief was given to all wounded soldiers without discrimination, regardless of their nationality. Others had performed humanitarian deeds on battlefields before; Dunant's genius lies in having taken things a lot further. Back in his hometown, Geneva, still traumatized, he wrote a book, *A Memory of Solferino*, published in 1862, in which he not only related the horrible battle and the suffering of the soldiers, but also launched an appeal around a vision and two key ideas.

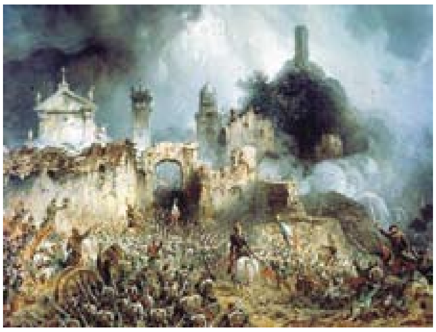
Carlo Bossoli, Museo Nazionale del Risorgimento, Turin

Figure 2.1

The battle of Solferino, 1859.

The vision was to provide neutral medical care for wounded soldiers in the field, and the ideas to accomplish this vision were the following.

- On the one hand, to create a relief society in every country which, already in peacetime, would train volunteers who could assist the armed forces' medical services, doctors and nurses ready to care for wounded soldiers in the event of war – this later gave birth to National Red Cross and Red Crescent Societies.
- On the other hand, to negotiate an international agreement that would grant protection and assistance to the sick and wounded and the medical services which cared for them, thus guaranteeing access to the wounded. This idea became the first Geneva Convention, the basis for modern international humanitarian law (IHL). Every human society throughout history has had rules about the conduct of warfare: this was the first concerted attempt to standardize and institutionalize, on an international basis, the law of war. IHL is also referred to as the "law of war" or the "law of armed conflict".

Five citizens of Geneva, including Dunant, members of a charitable association, created the "International Committee for the Relief of Wounded Soldiers" in 1863 in response to Dunant's book. This Geneva Committee persuaded the Swiss government to convene a diplomatic conference in 1864, in order to formalize the protection of medical services on the battlefield through an internationally recognized treaty: twelve governments participated. The result was the Geneva Convention of 22 August, 1864, for the Amelioration of the Condition of the Wounded in Armies in the Field. The Geneva Committee became the International Committee of the Red Cross, and the emblem of the red cross was adopted as a symbol of protection of the medical services that cared for the sick and wounded – the symbols of the red crescent, the red sun and lion, and red crystal were introduced later. Not only was this first Geneva Convention a decisive step towards formalizing the laws of armed conflict, but it also made it the duty of signatory States to create military medical services to care for their own wounded. Soldiers were no longer to be simply regarded as cannon fodder.

2.2 International humanitarian law: basic principles

Throughout history, humanity has known war. All human societies have developed customary rules that regulate how wars are fought. Over 500 cartels, codes of conduct, covenants and other texts designed to regulate hostilities had been recorded before the advent of contemporary humanitarian law. The first laws of war were proclaimed by major civilizations several millennia before our era: "I establish these laws to prevent the strong from oppressing the weak" declared Hammurabi, King of Babylon.

Just as there is no society of any sort that does not have its own set of rules, so there has never been a war that did not have some vague or precise rules covering the outbreak and end of hostilities, as well as how they were to be conducted.

The Geneva Conventions (GCs) as they stand today are the result of a long process. Over the years, the First Geneva Convention has been extended to meet the changing needs of modern warfare. Three other Conventions were adopted, one after the other, covering victims other than wounded soldiers: shipwrecked sailors, prisoners of war, and civilian populations. The four Geneva Conventions of 1949 and their two Additional Protocols of 1977, in particular, as well as other written treaties and customary law, constitute international humanitarian law (IHL): i.e. what is and what is not permitted during international and non-international conflicts. All States in the world have now ratified the Geneva Conventions, which means that they are bound by these legal instruments.

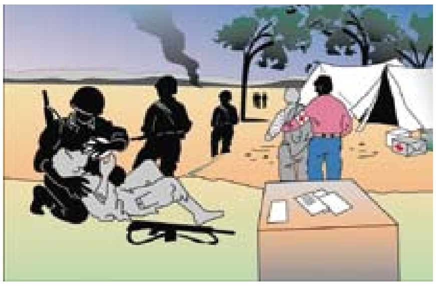
ICRC

Figure 2.2

I. Amelioration of the condition of the wounded and sick in armed forces in the field.

The four Geneva Conventions are mainly designed to regulate the behaviour of combatants and to protect persons who are not (or are no longer) taking part in hostilities in armed conflicts between States (international conflicts).

I. Convention for the amelioration of the condition of the wounded and sick in armed forces in the field (GC I: Revision of the 1864 Convention).

II. Convention for the amelioration of the condition of wounded, sick and shipwrecked members of armed forces at sea (GC II: revision of the 1899 Convention).

III. Convention relative to the treatment of prisoners of war (GC III: revision of the 1929 Convention).

IV. Convention relative to the protection of civilian persons in time of war (GC IV: new Convention adopted in 1949).

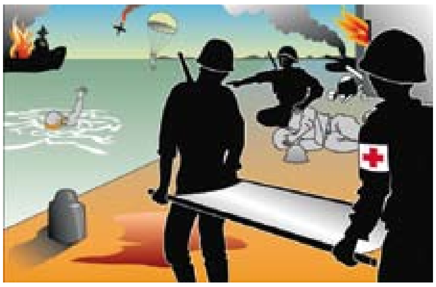
ICRC

Figure 2.3

II. Amelioration of the condition of wounded, sick and shipwrecked members of armed forces at sea.

Article 3 common to the four Geneva Conventions of 1949 also provides basic rules applicable in armed conflicts "not of an international character occurring in the territory of one of the High Contracting Parties".

In response to new needs in contemporary armed conflicts, the Conventions were developed and supplemented in 1977 by two further agreements: the Additional Protocols I and II relating to the protection, respectively, of victims of international and internal armed conflicts. A third Additional Protocol was adopted in 2005 to introduce a new protective emblem called the red crystal in addition to the red cross and red crescent emblems.

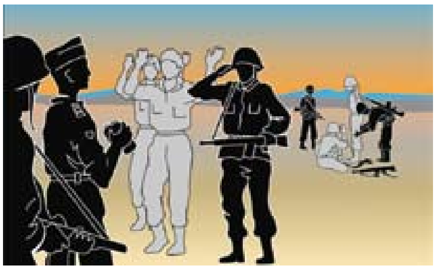
ICRC

Figure 2.4

III. Treatment of prisoners of war.

IHL regulates humanitarian issues in times of war; it aims to define the rights and obligations of the parties to a conflict in the conduct of warfare and to protect persons who are not (or are no longer) taking part in hostilities (civilians, wounded or sick soldiers, prisoners of war). These persons must be respected, protected and treated humanely by all parties. All the wounded and sick must also receive the medical care required by their condition. Those who provide such care must be respected and protected, as long as they are involved in their humanitarian tasks.

The various treaties defining IHL have become intricate, but the underlying message is simple. The human dignity of all individuals must be respected at all times without any kind of discrimination. Everything that *can* be done *should* be done to alleviate the suffering of those "*hors de combat*": persons who take no direct part in the conflict or who have been put out of action by sickness, injury, or captivity.

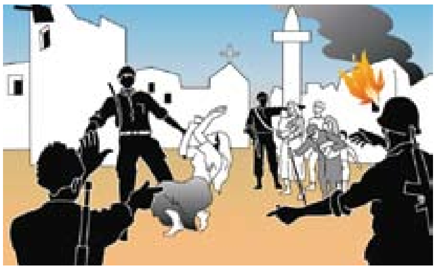
ICRC

Figure 2.5

IV. Protection of civilian persons in time of war.

International humanitarian law: main legal sources
- Customary law
- Four Geneva Conventions of 1949 (194 States party to the Conventions)
- Two Additional Protocols of 1977
- Third Additional Protocol of 2005
- The Hague Conventions of 1899 and 1907
- United Nations Convention on Conventional Weapons of 1980
- Mine Ban Convention of 1997
- Cluster Munitions Convention of 2008

2.2.1 Principles underlying IHL: the "law of war"

- The human dignity of all individuals must be respected at all times.
- Persons no longer involved in the fighting (sick, wounded and shipwrecked combatants, and prisoners of war) and those who do not take a direct part in hostilities (civilians) are entitled to respect for their lives and physical and moral integrity. They shall in all circumstances be protected and treated humanely without any adverse distinction.
- The wounded and sick shall be collected and cared for.
- Captured combatants and civilians under the authority of an adverse party are entitled to respect for their lives, dignity, personal rights and convictions. It is forbidden to kill or injure an enemy who surrenders.
- Everyone shall be entitled to benefit from fundamental judicial guarantees. No one shall be held responsible for an act he has not committed. No one shall be subjected to physical or mental torture, corporal punishment or cruel and degrading treatment. Hostage taking is prohibited.
- The choice of methods and means of warfare is not unlimited, and must be proportional to the military goals sought. It is prohibited to use weapons and methods of warfare that cause unnecessary losses, superfluous injury or unnecessary suffering.
- Attacks shall distinguish between the civilian population and combatants and between civilian objects and military objectives. Accordingly, operations shall be directed only against military objectives.
- Medical personnel have rights and obligations in times of armed conflict.

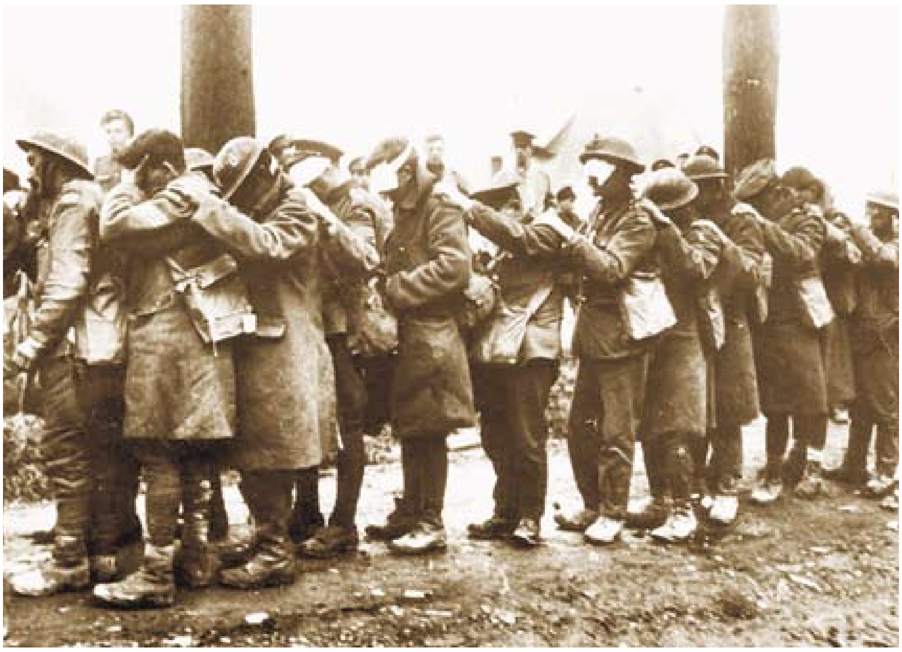

British Museum

Figure 2.6

Soldiers blinded by chemical weapons during World War I: an example of means of warfare that cause superfluous injury and unnecessary suffering.

2.2.2 IHL and human rights law

IHL applies in situations of armed conflict, whereas human rights, or at least some of them, protect the individual at all times, in war and peace alike. Some human rights treaties permit governments to derogate from certain rights in situations of public emergency and if strict conditions are fulfilled. However, certain fundamental human rights must be respected in all circumstances and may never be waived, regardless of the emergency: in particular the right to life, the prohibition of torture and inhuman punishment or treatment, slavery and servitude, and the principle of legality and non-retroactivity of the law. No derogations are permitted under IHL because it was conceived for emergency situations, namely armed conflict.

IHL mainly aims to protect people who do not or are no longer taking part in hostilities. The rules embodied in IHL impose duties on all parties to a conflict, including non-governmental groups. Human rights, being tailored primarily for peacetime, apply to everyone. Their principal goal is to protect individuals from arbitrary behaviour by their own governments. Human rights law does not deal with the conduct of hostilities.

2.3 The distinctive emblems

The distinctive emblems of the red cross, red crescent, and red crystal are meant to mark certain medical and religious personnel and equipment which must be respected and protected during armed conflict (protective use). They also serve to show that persons or objects are linked to the International Red Cross and Red Crescent Movement including in situations other than armed conflict (indicative use). Their use is strictly defined (see Annex 2. A: The distinctive emblems).

Use of the emblems is restricted to the following entities.

1. Medical staff and services of the armed forces are the very first persons authorized to use the distinctive emblem, as a sign of protection.
2. Volunteers of a National Red Cross or Red Crescent Society, when duly authorized by the competent State Authority.
3. Staff of the International Committee of the Red Cross and the International Federation of Red Cross and Red Crescent Societies.

Persons and objects displaying the emblems must not be attacked, but on the contrary must be respected, protected, and encouraged in their work.

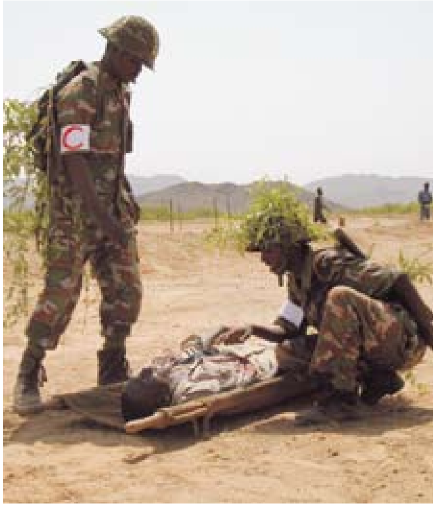

V.Louis / ICRC

Figure 2.7

Armed forces' medical services.

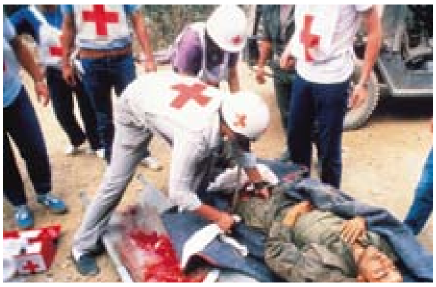

B. Fjortoft / ICRC

Figure 2.8

National Society volunteers.

B. Heger / ICRC

Figure 2.9

ICRC, International Federation and National Society staff.

2.4 The International Red Cross and Red Crescent Movement and its Fundamental Principles

The International Committee of the Red Cross and the International Federation of Red Cross and Red Crescent Societies, together with the National Red Cross and Red Crescent Societies, form the International Red Cross and Red Crescent Movement.

Fundamental Principles of the Red Cross and Red Crescent Movement[1]

Humanity

Impartiality

Neutrality

Independence

Voluntary service

Unity

Universality

In peacetime, as in time of conflict, it is the duty of the armed forces and the National Society of every country to disseminate the rights and duties enshrined in IHL. This is necessary not only for these rules to become known, understood, accepted and respected by combatants, but also so that the whole population understands and supports its Red Cross or Red Crescent Society for it to become more efficient, for the benefit of all, in times of conflict and in times of peace. Part of this dissemination includes the respect due to the red cross, red crescent and red crystal emblems as a sign of protection.

2.5 Rights and duties of medical personnel according to IHL

IHL provides medical personnel with rights in times of armed conflict, but also assigns duties to them. The duties incumbent on them are directly linked to the rights of the protected persons placed in their care. These provisions are a case-specific refinement of the basic rights and duties defined by medical ethics and the Hippocratic oath. Medical personnel are bound by medical ethics and IHL to treat patients solely on the basis of need and without regard for their nationality, race and class, religious or political beliefs. These rights and duties have been defined to enable medical personnel to perform the humanitarian task entrusted to them, born of the desire to alleviate human suffering by helping and tending the wounded and the sick, which is the very purpose of the medical mission.

The Geneva Conventions define medical personnel as:

- persons assigned by a party to the conflict, whether on a permanent or temporary basis, exclusively to medical purposes (search, collection, transportation, diagnosis and treatment of the wounded and sick, and for the prevention of disease) – this includes doctors, nurses, orderlies, first aiders, and stretcher-bearers;
- persons assigned by a party to the conflict, whether on a permanent or temporary basis, exclusively to the administration or operation of medical units or medical transport – this includes administrators, drivers, cooks, mechanics, etc.

The term "medical personnel" is thus not confined to the narrow meaning of the word. All personnel required to ensure the adequate treatment of the wounded and sick are covered by the protection granted to them, as long as they form an integral part of the medical service.

---

[1] See Annex 2. B: The Fundamental Principles of the International Red Cross and Red Crescent Movement.

Medical units, whether military or civilian, include what is set up for medical purposes:
- all buildings or installations (hospitals, clinics, first-aid posts, field hospitals, tents, etc.),
- blood transfusion and preventive medicine centres,
- medical and pharmaceutical depots and stores.

They may be fixed or mobile, permanent or temporary.

Medical equipment includes:
- stretchers,
- medical and surgical appliances and instruments,
- medical supplies, dressings, etc.

Medical transport can be organized by land, water, and air:
- ambulances, lorries or trucks,
- hospital ships, rescue craft,
- medical aircraft, etc.

Medical personnel must not be the object of attack, and must be allowed to care for the sick and wounded unhindered. They should wear the distinctive emblem of the red cross, red crescent or red crystal on a white ground and carry an identity card. Military medics may bear arms only for their own defence and to defend the wounded and sick from pillage. However, they must not use arms to prevent capture of their patients, units or themselves by the enemy; by doing so they lose their protected status.

If medical personnel fall under the control of enemy troops, they shall be allowed to continue their duties towards the wounded and sick. They shall not be compelled to perform acts contrary to medical ethics, or refrain from performing acts required by medical ethics. A patient remains a patient, and medical personnel have a responsibility to provide care according to their training and the means available. If captured medical personnel are not indispensable to the care of other prisoners, they should be repatriated. Those retained shall not be considered prisoners of war and their work shall be facilitated.

In occupied territory, civilian medical personnel should be allowed to continue providing adequate medical services for the civilian population.

The civilian population shall respect the wounded and sick, even if they belong to the enemy, and shall commit no act of violence against them. Civilians are permitted to collect and care for the wounded and sick of whatever nationality, and shall not be penalized for doing so. On the contrary, they must be aided in this work.

Medical units enjoy protected status so long as they are not used to commit acts harmful to the enemy, such as sheltering able-bodied combatants, storing arms or ammunition, or being used as military observation posts. Otherwise, their protection ceases and they become legitimate military targets. This is why strict controls must be established to safeguard the protected status of medical units and transport.

To enhance the protection of medical units and medical transport, they should be clearly marked with the red cross, red crescent or red crystal emblem, of the largest size possible. The emblem is the visible sign of protection conferred by the Geneva Conventions and their Additional Protocols.

2.6 Responsibility of States

As with any international agreement, the governments of States have a distinct responsibility whenever they become party to a treaty.

- The Geneva Conventions and their Additional Protocols are both contracts with other States and commitments towards humanity by which governments have agreed to certain rules that regulate the conduct of armed hostilities and the protection of persons who are not (or are no longer) taking part in hostilities, what is known as "law in war", or *jus in bello*. Its provisions apply to the warring parties regardless of the reasons for the conflict and whether or not the cause upheld by either party is just.
- On becoming party to the four Geneva Conventions of 1949, States agree to respect and ensure respect for their obligations in all circumstances (common Art. 1).
- Reciprocity is not supposed to be a prerequisite to a country adhering to and applying these rules, but it exists in reality and practice. It is in everyone's interest that everyone apply the law of his or her own accord.
- States are responsible for protecting those who do not participate or are no longer participating in hostilities.
- During peacetime, States are to promote IHL and train the members of their armed forces to respect IHL.
- At all times, States must take all measures necessary for the prevention and repression of any misuse or abuse of the distinctive emblems.
- On becoming party to the Geneva Conventions, States undertake to enact any legislation necessary to punish persons guilty of grave breaches of them, and are also bound to prosecute in their own courts any person suspected of having done so, or to hand over that person to another State for judgment. In other words, perpetrators of grave breaches, i.e. war criminals, must be prosecuted at all times and in all places, and States are responsible for ensuring that this is done.
- A number of independent mechanisms have been set up, each within its mandate and competency, to promote better compliance with IHL, including the ICRC, the International Fact-finding Commission, regional and United Nations bodies, and the International Criminal Court (Rome Statute of 1998). Their adequacy and sufficiency in meeting the task remains a challenge to humanity.
- Generally speaking, a State's criminal laws apply only to crimes committed on its territory or by its own nationals. International humanitarian law goes further in

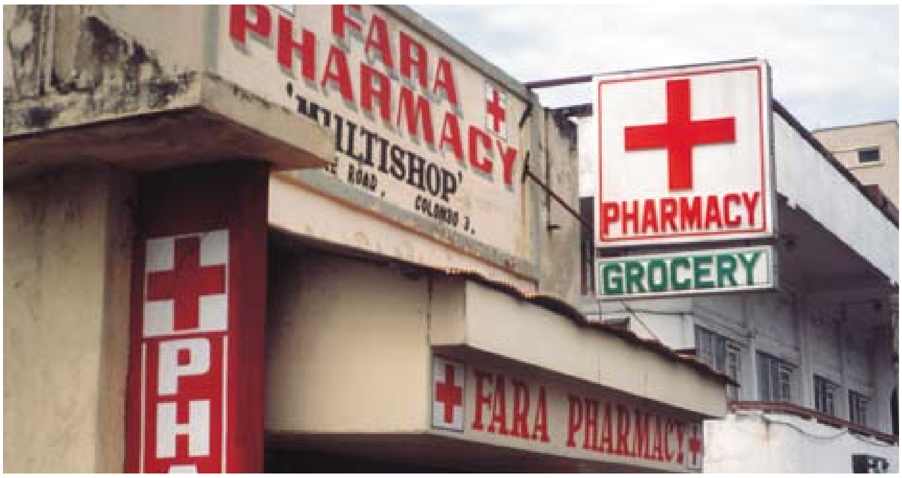

ICRC

Figure 2.10

All too often the red cross emblem is used to indicate any health service without regard to its privileged legal status, which confers protection.

that it requires States to seek out and punish any person who has committed a grave breach, irrespective of his nationality or the place where the offence was committed. This principle of universal jurisdiction is essential to guarantee that grave breaches are effectively repressed. Such prosecutions may be brought either by the national courts of the various States or by an international authority. In this connection, the International Criminal Tribunals for the former Yugoslavia and Rwanda were set up by the United Nations Security Council in 1993 and 1994, respectively, to try those accused of war crimes committed during the conflicts in those countries.

- Lastly, the international community has created a permanent International Criminal Court (Rome Statute of 1998), which is competent to try war crimes, crimes against humanity, and genocide.
- The United Nations Security Council is the main UN body responsible for the maintenance of international peace and security. To this end, it may decide on measures, including coercive ones, against any State threatening or breaching international peace (Chapters VI and VII of the United Nations Charter). The rules that outlaw war, with some exceptions, are know as "the law on the use of force" or the *jus ad bellum*. This is different from the *jus in bello*, or IHL, which is the law applicable in times of armed conflict.
- "New" or "anarchic" conflicts lead to the partial, and sometimes total, weakening or breakdown of State structures. In such situations, armed groups take advantage of the political vacuum in an attempt to grab power. It is not because a State's structures have been weakened or are non-existent that there is a legal vacuum with regard to international law. The law, and its obligations, remain.

Admittedly, humanitarian rules are harder to apply in these types of conflict. The lack of discipline among belligerents, the arming of the civilian population as weapons flood the territory, and the increasingly blurred distinction between fighters and civilians often cause confrontations to take an extremely brutal turn – in which there is little room for respect for IHL.

As a result, this is the type of situation in which particular efforts are needed to make people aware of humanitarian law. Better knowledge of the rules of law will not solve the underlying problem which led to the conflict, but it is likely to attenuate its deadlier consequences.

Insofar as a "right – or even a duty – to intervene" is tantamount to justifying armed intervention undertaken for humanitarian reasons, this is a matter not for humanitarian law but for the rules on the legality of the use of armed force in international relations, i.e. of *jus ad bellum*, according to the Charter of the United Nations.

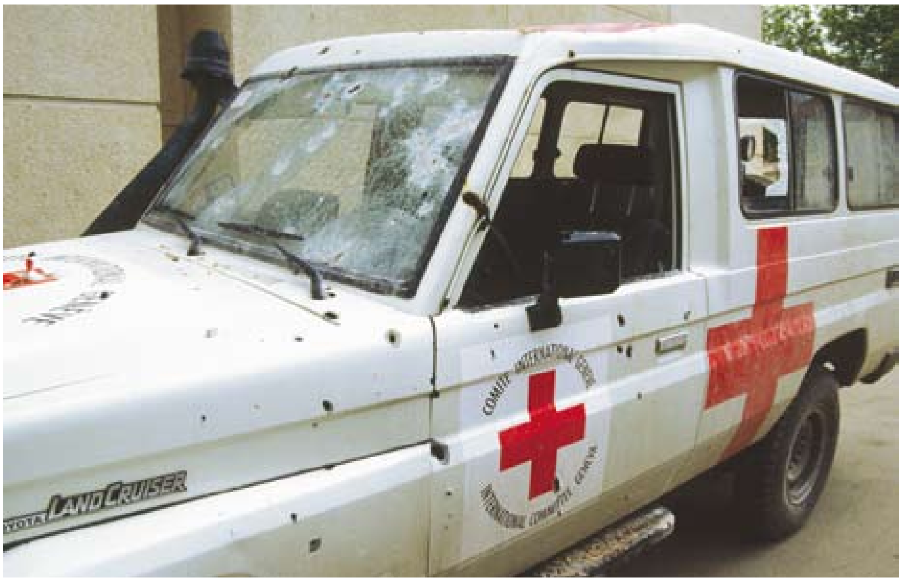

R. Sidler / ICRC

Figure 2.11

This ICRC car was targeted by "uncontrolled elements".

2.7 Reality check: some people do not follow the rules

Given what the law dictates, what is the reality on the battlefield? Whether in conflict situations or in peacetime, and whether national or international law is applicable, laws are violated and crimes committed. There are many examples of violations of IHL: surrounding a military objective with medical units so that it will not be targeted; hiding weapons in a hospital; transporting able-bodied combatants in an ambulance; using an aircraft displaying the emblem for reconnaissance missions; not respecting the non-combatant status of sick and wounded soldiers – on the contrary, all too often their survival is seen as an invitation by some to "finish the job", and atrocities are the result. All too often hospitals and medical personnel are the target of attack, or are prevented from doing their duty of caring for the sick and wounded, which is perceived by some as giving "succour and aid to the enemy". All these violations have three things in common: they seriously weaken the system of protection embodied in IHL, they divert people and objects displaying the red cross, red crescent or red crystal from their humanitarian purpose, and they put lives in danger by fostering mistrust.

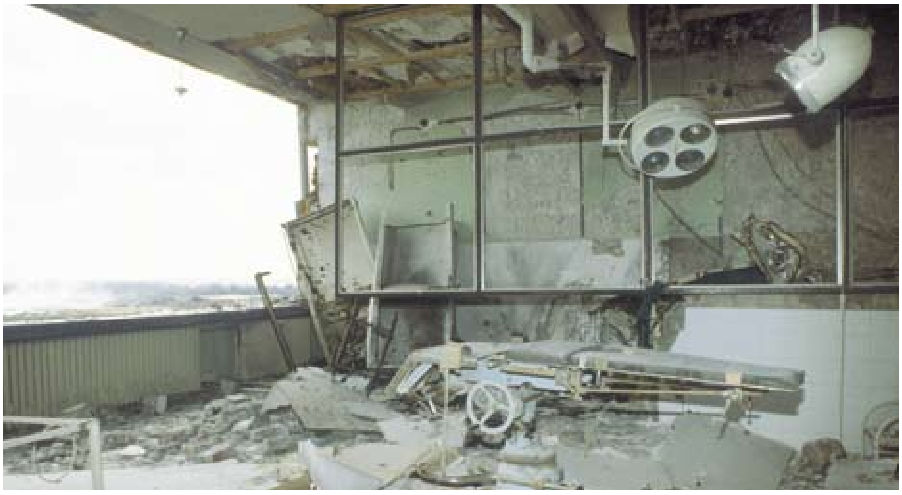
T. Pizer / ICRC

Figure 2.12

Unfortunately, hospitals are not immune to attack: this photo shows a blatant contravention of IHL.

Red Cross and Red Crescent personnel are not immune to this lack of respect for IHL. Mussolini's air force bombed a Swedish Red Cross ambulance during the invasion of Abyssinia on 30 December 1935, and 28 people were killed and 50 wounded. More recently, masked gunmen entered the residence of the ICRC hospital in the village of Novye Atagi, Chechnya (southern Russia), in December 1996 and assassinated five nurses and the construction engineer in cold blood. The authors of this text have all too often been threatened, prevented from reaching and treating the victims, or seen their hospital attacked or bombed. The world over, colleagues, both civilian and military, are time and again made to suffer "simply" because they want to fulfil their humanitarian, ethical, and legal duty to care for their patients.

Humanitarian action is based on a fundamentally "optimistic philosophy", wrote Jean Pictet.[2] But this optimism in no way detracts from the "philosophy's" realism. It is aware that humanitarian work is difficult. Its greatest enemies may well be neither weapons nor disaster, but selfishness, indifference and discouragement. One should not despair, however. Rather one should see this as an incentive to redouble efforts to educate those involved in armed conflict. Medical ethics go hand in hand with IHL and the Fundamental Principle of impartiality: the sick and wounded are entitled to treatment, whatever their origin or political persuasion.

---

[2] Jean Pictet (1914 – 2002) was Director-General and Honorary Vice-President of the ICRC. He was responsible for the preparatory works that led to the 1949 revision of the Geneva Conventions and the Additional Protocols of 1977, and his writings formed the basis of the Fundamental Charter of the International Red Cross and Red Crescent Movement, adopted in 1965.

All medical personnel should not only comply with medical ethics and IHL but should also disseminate their meaning to military and civilian authorities, ordinary soldiers and citizens.

Medical practitioners have a particularly important part to play: they have responsibilities with regard to individual patients but also towards society at large. They must convince combatants of the need to respect IHL, to refrain from attacking the medical facilities and personnel of the enemy and allow them to perform their humanitarian task. While the underlying philosophy of the Movement is *not* rooted in the principle of reciprocity, in practice medical professionals can only expect to benefit from the "protection" offered by the Geneva Conventions – in a general atmosphere of respect for the Fundamental Principles – if their colleagues practising on the "other" side also benefit from the same protection. They must inform those who bear weapons of their obligations and trust that colleagues in contact with the opposing side are doing the same. This has proved true in the past, and soldiers have often respected the principle of protection afforded to the sick and wounded and medical and religious personnel.

2.8 The neutrality of a National Red Cross/ Red Crescent Society

Another major problem faced during armed conflicts of a non-international character, that is to say internal conflicts, is the question of neutrality, especially of a National Red Cross or Red Crescent Society. The requirement of non-discrimination is of particular concern to Red Cross/Red Crescent Societies, it is in fact a condition for their recognition. They must be open to all those who wish to become members and must permit all social, political and religious groups to be represented; this representativity is the guarantee of the Societies' ability to engage in exclusively humanitarian activities and to resist all partisan considerations.

It is admittedly not always an easy task to apply the principle of neutrality, not least because everyone has personal convictions. When tension mounts and passions are aroused, every member of the Red Cross or Red Crescent must exercise great self-control and refrain from expressing his/her opinions in the discharge of his duties. Volunteers are not asked to be neutral – everyone is entitled to an opinion – but to act neutrally. That is an important distinction. The next difficulty is the fact that the parties to the conflict often take a dim view of neutral behaviour. In countries where an internal conflict is taking place, the armed forces fail to understand why the National Society does not condemn the activities of those they regard as "bandits", much less why it wants to provide assistance to any of their number no longer able to fight. As for the opposition, they are critical of the Society's connections with the authorities.

Anyone trying to work on both sides to help non-combatants is considered at best naïve, at worst a traitor. The extremely polarized nature of many struggles is such that *not taking a stand* is a hostile act in itself. This is why the Red Cross and Red Crescent's neutrality and impartiality must be explained. As one National Society first aider put it: "The best argument I have is to tell one of the parties to the conflict that if I take its side and ignore the victims on the other side, I will never again be able to bring help to its own wounded members".

It would appear, indeed, that the principles of neutrality and impartiality are ideals to be attained, inner qualities that are rarely inborn but that most often require one to overcome one's instincts. They demand from members of the Red Cross and Red Crescent Societies arduous and sustained efforts to overcome their own prejudices and preferences in order to be able to perform the purest act of impartiality, which is to give more help to the adversary who is the victim of great misfortune than to the friend whose suffering is less severe, or to care for the more severely wounded, even if guilty, before assisting the innocent whose injuries are slighter.

2.9 The role and mandate of the ICRC in situations of armed conflict

Established in 1863, the International Committee of the Red Cross (ICRC) is an impartial, neutral, and independent organization whose exclusively humanitarian mission is to protect the lives and dignity of the victims of war and internal violence and to provide them with assistance. The ICRC endeavours to prevent suffering by promoting and strengthening international humanitarian law and universal humanitarian principles.

The ICRC in a nutshell

Neutral, impartial and independent humanitarian organization.

Works in situations of armed conflict.

Entrusted with an international mandate under the Geneva Conventions.

Protects and assists victims of conflict.

Part of the International Red Cross and Red Crescent Movement.

The ICRC has been named the "guardian and promoter" of IHL by the States party to the Geneva Conventions. It works during armed conflict to protect and assist the victims: the wounded and sick, prisoners of war and other detainees, and the civilian population.

The ICRC does not take sides or determine right or wrong in a conflict. The neutrality of the ICRC, however, is not a widely appreciated principle. There are many who express indignation at its neutrality in the mistaken belief that neutrality betokens lack of commitment and courage. The ICRC, for its part, has great difficulty in getting the parties to a conflict to understand that the only thing it must grant equally to each is its willingness to serve, and that in other respects its activities are proportional to the needs, and consequently unequal when distress is greater on one side than on the other.

At the beginning of hostilities, the ICRC:

- reminds the belligerents of their obligations according to IHL;
- offers its services as a neutral intermediary for the accomplishment of humanitarian activities;
- makes an assessment of the needs;
- acts in favour of victims according to the needs.

The protection role the ICRC plays consists in acting in defence of people who are not, or no longer, taking part in combat: the wounded or sick, detainees or prisoners of war, and civilians, including the inhabitants of territories administered or occupied by a hostile power. The ICRC approaches the competent authorities to ensure that these victims receive humane treatment. The ICRC has the right to have access to prisoners of war (Third Geneva Convention), and detained persons covered by the Fourth Geneva Convention, including the right to visit detention camps.

The ICRC has an official policy of discretion. Only when it observes grave and repeated breaches of IHL and when its confidential representations have been in vain and it considers that the only means of helping the victims is to ask for the support of the international community, does it make public representations. This sometimes takes the form of an appeal to the States party to the Geneva Conventions, whose responsibility it is to respect and *ensure* respect for IHL. Such initiatives are nevertheless the exception rather than the rule.

ICRC activities

1. Protection during armed conflict:
   - protection of civilians confronted with the violence of armed conflict;
   - protection and care of the wounded, sick, and shipwrecked;
   - protection of prisoners of war and other detainees (registration and visits);
   - as a neutral intermediary, facilitation and accompaniment of the release of prisoners of war;
   - re-establishment of family links through Red Cross messages;
   - re-establishment of family links through the tracing agency for the missing and dead.
2. Assistance to victims of armed conflict – public health approach (shelter, water, sanitation, food, preventive and curative medical care):
   - economic security;
   - water and habitat;
   - health services (including assistance to the war-wounded and sick).
3. Preventive action:
   - promotion of international humanitarian law (IHL);
   - advice and assistance with national measures to implement IHL;
   - mine risk education programmes to raise awareness among people exposed to the dangers posed by mines and unexploded ordnance.

ICRC delegates must be willing to talk to all those who are responsible for violations of IHL and human rights. They cannot pass judgment on them publicly, but must speak to them on behalf of those to whom speech is denied, and who have no one else to turn to. They often do so at considerable risk to their own personal safety and their words may fall on deaf ears, but if this policy of refraining from public denunciation makes it possible to alleviate the suffering of just one man, woman or child, that is ample recompense.

As a neutral and independent institution, the ICRC is authorized by the Geneva Conventions and their Additional Protocols to administer first aid and other care in the field to victims of armed conflict. Military authorities are required to permit the ICRC to collect and care for the wounded or sick of whatever nationality, even in invaded or occupied territories. The ICRC can offer its services to the parties, in particular in the medical realm to establish neutral or hospital zones, set up hospitals for the sick and wounded, assist existing hospitals, and provide rehabilitation for amputees, especially victims of anti-personnel landmines.

The ICRC helps to organize, or provide directly, relief to victims of armed conflict. These relief supplies cover the most essential needs, such as food, clean drinking water, shelter, clothing, and medical care.

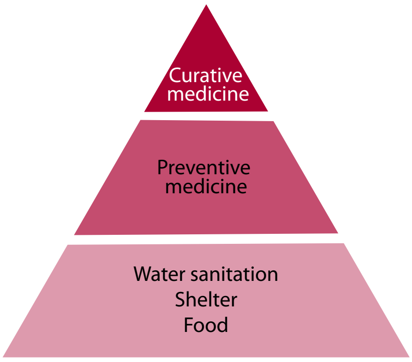

Figure 2.13

The assistance pyramid: public health analysis of population needs.

2.9.1 Health services: assistance to the war-wounded and sick

Although proper medical care of sick and wounded soldiers and civilians in times of armed conflict is taken for granted by many governments today, poverty may compromise a government's efforts to provide such care. The ICRC maintains a capacity to aid States whose authorities show a willingness to assume their responsibilities towards their own soldiers and the civilian population.

The ICRC has deployed many assistance and training programmes to this effect. One example of an offer of services by the ICRC to two countries involved in a conflict was the case of the 1998 – 2000 war between Ethiopia and Eritrea. The ICRC provided assistance to help both governments to develop the programmes described below.

ICRC EXPERIENCE

The ICRC carried out the following programmes in Eritrea and Ethiopia from 1998 to 2001.

Ethiopia

- Training in advanced first aid, triage and evacuation for the war-wounded.
- War surgery seminars.
- Specialist surgical training for the war-wounded: thoracovascular, orthopaedic and neurosurgery.
- Maxillo-facial reconstructive surgery training programme.
- Short-term ICRC surgical team presence in an emergency situation.
- Support to the Ethiopian Red Cross Society ambulance service.
- Surgical material and equipment.
- Visits to thousands of POWs and their repatriation after cessation of hostilities.
- Repatriation of thousands of interned enemy aliens.

Eritrea

- Training in advanced trauma nursing care for ambulance staff.
- Emergency Room Trauma Course for the war-wounded.
- War surgery seminar.
- Short-term ICRC surgical team presence in an emergency situation.
- Intensive care nursing training for the war-wounded.
- Development of a physiotherapy training curriculum at the Institute of Nursing Care.
- Support to the Red Cross Society of Eritrea ambulance service.
- Surgical material and equipment.
- Visits to thousands of POWs and their repatriation after cessation of hostilities.
- Repatriation of thousands of interned enemy aliens.
- Assistance to thousands of internally displaced persons (food, shelter, water, sanitation, medical care).

While the ICRC is authorized to care for the wounded and sick during armed conflict, it is also involved in post-conflict reconstruction, indeed in development assistance, and, sometimes, the ICRC achieves the right balance.

ICRC EXPERIENCE

“We want to thank the doctors and nurses of the ICRC. Thank you for coming. And, thank you for leaving.”

Rui Paolo [3]

---

[3] Rui Paolo, Director of Hospital Services, Ministry of Public Health, Dili, Timor Leste, June 2001, on the occasion of the handover of the administration of the Dili General Hospital from the ICRC to the Ministry of Public Health.

ANNEX 2. A The distinctive emblems

The distinctive emblems of the red cross, red crescent and red crystal on a white ground are the symbols of impartial humanitarian work and do not represent any particular religious belief. They provide protection for military medical services and relief workers in armed conflicts. Moreover, they are also used for identification purposes by National Societies of the Red Cross and Red Crescent Movement in each country.

In the event of armed conflict, the distinctive emblems are a visible sign of the protection conferred by international humanitarian law upon medical personnel and equipment. The persons and objects displaying them must not be attacked, but on the contrary must be respected and protected.

Their use as a *protective sign* during armed conflict is authorized exclusively for:
- medical units, transport and personnel, as well as religious personnel, of the armed forces;
- civilian medical units**,** transport, and personnel, as well as civilian religious personnel, that have received special permission by the competent authorities to use the emblem;
- medical units, transport and personnel that a Red Cross or Red Crescent National Society has put at the disposal of the medical services of the armed forces.

Persons and buildings/structures/objects displaying the emblem must not be attacked, damaged or prevented from operating, but, on the contrary, must be respected and protected, even if, for the moment, they are not caring for or housing either wounded or sick people. The perfidious use of the emblems is explicitly prohibited.

In order to ensure effective protection during wartime, the *indicative* use of the emblem must be *strictly controlled* and only used for:
- Red Cross or Red Crescent National Societies, in order to indicate that persons or goods have a connection with the Society in question (it must be small in size to avoid any confusion with the emblem used as a protective symbol);
- exceptionally, ambulances and first-aid stations exclusively assigned to the purpose of giving free treatment to the wounded and sick, with the authorization of a National Society.

The International Committee of the Red Cross and the International Federation of Red Cross and Red Crescent Societies are authorized to use the emblem for all their activities and at all times.

Misuse of the emblem is a serious problem. In peacetime, hospitals, clinics, doctors' offices, pharmacies, non-governmental organizations and commercial companies tend to use the emblem in order to benefit from its reputation, although they are not entitled to do so. This abuse clearly weakens the protective value of the emblem in wartime.

Any case of misuse of the emblem should be reported to the relevant Red Cross or Red Crescent National Society, the ICRC, or the International Federation of Red Cross and Red Crescent Societies.

Parties to the Geneva Conventions shall take the precautions necessary for the prevention or repression of any abuse of the distinctive emblems.

During peacetime staff and volunteers of the International Red Cross and Red Crescent Movement, through their behaviour, activities and awareness-raising efforts, seek to ensure that the protective value of the distinctive emblems is well known to the military and the general public.

Please note:

On 8 December 2005, a Diplomatic Conference adopted Protocol III additional to the Geneva Conventions, which recognizes an additional distinctive emblem. The "Third Protocol emblem", also known as the red crystal, is composed of a red frame in the shape of a square on edge on a white background. According to Protocol III, all distinctive emblems enjoy equal status.[4] The conditions for use of and respect for the Third Protocol emblem are identical to those for the distinctive emblems established by the Geneva Conventions and, where applicable, the 1977 Additional Protocols.

J. Perez / International Federation

---

[4] While no longer in use, the red lion and sun on a white background is still recognized by the Geneva Conventions.

Annex 2. B The International Red Cross and Red Crescent Movement

The International Committee of the Red Cross and the International Federation of Red Cross and Red Crescent Societies, together with the National Red Cross and Red Crescent Societies, form the International Red Cross and Red Crescent Movement.

Born of the compassion felt by a Swiss citizen, Henry Dunant, at the sight of the dead and wounded abandoned on the battlefield of Solferino, the International Red Cross and Red Crescent Movement believes that its first duty is to render wars that cannot be avoided less inhuman, and to alleviate the suffering that they cause. Its aim has been to bring a little bit of humanity into the horrors of war. The Movement was founded as a result of a conflict and for conflicts, with the aim of helping people in distress on the battlefield.

The International Red Cross and Red Crescent Movement endeavours, in its international and national capacity, to prevent and alleviate human suffering wherever it may be found. Its purpose is to protect life and health and to ensure respect for human beings. It promotes mutual understanding, friendship, cooperation and lasting peace amongst all peoples.

The humanitarian ideas of the Movement are reflected in the seven Fundamental Principles that guide the activities of all its components at all times: humanity, impartiality, neutrality, independence, voluntary service, unity and universality.

Components of the International Red Cross and Red Crescent Movement

The International Committee of the Red Cross (ICRC)

The International Committee of the Red Cross, created in Geneva, Switzerland in 1863, is the founding body of the Red Cross/Red Crescent Movement. It is an independent humanitarian organization. As a neutral intermediary and on the basis of the Geneva Conventions or the customary law of nations, which grant it the right of initiative, it endeavours to protect and assist the military and civilian victims of international and non-international armed conflict and internal disturbances and tension.

The countries of the world have entrusted to the ICRC the roles of promoter and custodian of international humanitarian law, and the task of working for their development and worldwide dissemination.

The functions of the ICRC are defined in its own statutes, in those of the Red Cross/Red Crescent Movement, and also in the international treaties known as the Geneva Conventions of 1949 and their Additional Protocols of 1977.

The International Federation of Red Cross and Red Crescent Societies

The International Federation is the permanent liaison body between National Red Cross and Red Crescent Societies. It attempts to prevent and alleviate human suffering through the promotion of activities by National Societies and so contribute to peace. The International Federation encourages the creation and assists the development of National Societies in providing services to the community.

The International Federation organizes and coordinates international relief efforts for victims of natural disasters and promotes the adoption of national disaster-preparedness plans. First aid is an important part of both everyday community service and disaster preparedness.

The National Red Cross/Red Crescent Society

There are 186 National Red Cross/Red Crescent Societies around the world, with more currently being created. From the outset, the goal of the Red Cross/Red Crescent Movement has been to create relief societies which, in their capacity of *auxiliaries to their countries' armed forces' medical services*, would be called upon to tend wounded or sick soldiers. The establishment of such Societies was consistent with the aims of the original Geneva Convention of 1864.

The National Societies' activities have developed steadily and diversified over the years. The Societies' concern, at first directed only at members of the armed forces, now extends to protection and assistance of all those, whether military or civilian, who are the victims of conflicts, in close cooperation with the ICRC, and of natural catastrophes, in close cooperation with the International Federation of Red Cross and Red Crescent Societies. National Societies have also developed many activities in peacetime as *auxiliaries to the public authorities* to alleviate suffering, improve health, and prevent disease.

National Society membership is open to everyone and services are provided on the sole criterion of need. National Societies must fulfil stringent conditions to achieve recognition by the ICRC and obtain International Federation membership. Among these conditions: respect for the Fundamental Principles, and recognition by their home government as a voluntary aid society, auxiliary to the public authorities.

Each National Red Cross/Red Crescent Society has its own particular characteristics, and its activities cover a wide range depending on the needs of the country and on the Society's operational capabilities. The one activity that all National Societies share is the provision and teaching of first aid.

The Fundamental Principles of the International Red Cross and Red Crescent Movement

Humanity

The International Red Cross and Red Crescent Movement, born of a desire to bring assistance without discrimination to the wounded on the battlefield, endeavours – in its international and national capacity – to prevent and alleviate human suffering wherever it may be found. Its purpose is to protect life and health and to ensure respect for the human being. It promotes mutual understanding, friendship, cooperation and lasting peace amongst all peoples.

Impartiality

It makes no discrimination as to nationality, race, religious beliefs, class or political opinions. It endeavours to relieve the suffering of individuals, being guided solely by their needs, and to give priority to the most urgent cases of distress.

Neutrality

In order to continue to enjoy the confidence of all, the Movement may not take sides in hostilities or engage at any time in controversies of a political, racial, religious or ideological nature.

Independence

The Movement is independent. The National Societies, while auxiliaries in the humanitarian services of their governments and subject to the laws of their respective countries, must always maintain their autonomy so that they may be able at all times to act in accordance with the principles of the Movement.

Voluntary service

It is a voluntary relief movement, not prompted in any manner by the desire for gain.

Unity

There can be only one Red Cross or one Red Crescent Society in any one country. It must be open to all. It must carry out its humanitarian work throughout its territory.

Universality

The International Red Cross and Red Crescent Movement, in which all Societies have equal status and share equal responsibilities and duties in helping each other, is worldwide.

Humanity and Impartiality express the objectives of the Movement.

Neutrality and Independence ensure access to those in need of help.

Voluntary service, Unity and Universality enable the International Red Cross and Red Crescent Movement to work effectively all over the world.

Chapter 3

MECHANISMS OF INJURY DURING ARMED CONFLICT[1]

---

[1] Acknowledgement: the ballistics portion of this chapter is largely based on the work of Doctor of Forensic Science Beat Kneubuehl, Institute of Forensic Medicine, University of Bern, in collaboration with Armasuisse, Centre for Military and Civilian Systems, Science and Technology, Ballistics Laboratory, Thun, Switzerland. His cooperation and collaboration with ICRC surgeons over the years have permitted a whole generation of war surgeons to gain an important insight into the wounding potential of weapons. We hope that this knowledge has permitted better surgical management of the victims of armed conflict and other situations of violence where the weapons of war are used.

3. MECHANISMS OF INJURY DURING ARMED CONFLICT

3.1 The various mechanisms of injury 55
3.1.1 Blunt injury in war 55
3.1.2 Penetrating wounds caused by weapons of war 55
3.1.3 Anti-personnel landmine wounds 57
3.1.4 Explosive blast injuries 58
3.1.5 Burns 59
3.1.6 Non-conventional weapons 59

3.2 Ballistics 59
3.2.1 Introduction 59
3.2.2 Internal ballistics 60
3.2.3 External ballistics 62
3.2.4 Summary 62

3.3 Terminal ballistics 63
3.3.1 Role of kinetic energy 63
3.3.2 Laboratory experiments 63
3.3.3 Non-deforming rifle bullet: full metal jacket (FMJ) military bullet 64
3.3.4 Deforming and fragmenting rifle bullets: dum-dum 66
3.3.5 Handgun bullets 68
3.3.6 Fragment projectiles 69

3.4 Wound ballistics 70
3.4.1 Laboratory studies *versus* human body 70
3.4.2 Projectile–tissue interaction 70
3.4.3 Tissue factors 71
3.4.4 Pathological description of the permanent wound cavity 72
3.4.5 Clinical applications 73
3.4.6 Sonic shock wave in tissues 76
3.4.7 Pressure wave in tissues and blood vessels 76
3.4.8 Fragment wounds 77

3.5 Wound dynamics and the patient 78

3.1 The various mechanisms of injury

The surgeon caring for the war-wounded will face a wide variety of injuries. During armed conflict, all the usual trauma of peacetime continues unabated and natural catastrophes occur as well. Armed conflict itself provokes blunt injuries and burns, and trauma that are specific to weapons and the circumstances of warfare. More specifically, combat involves penetrating and blast injuries; these will be the major focus of this book.

Modern armed conflict causes specific penetrating and blast injuries, as well as blunt and burn trauma.

3.1.1 Blunt injury in war

Blunt trauma is common in war. Severe injury due to blunt trauma may be more difficult to detect than in cases of penetrating trauma, and X-ray diagnosis and other more sophisticated technologies are more valuable for patient assessment.

Road traffic crashes

Military vehicles often drive at high speed over difficult terrain where safe roads do not exist. In addition, the environment of any accident may be hostile (presence of enemy forces, minefields, etc.).

Collapsed buildings and falls

The collapse of bombed buildings will cause blunt and crush injuries to anyone inside. People may fall from destroyed buildings or a balcony.

Explosions and anti-tank mines

A bomb or shell explosion can throw people against objects. A car, bus or lorry carrying passengers may drive over an anti-tank mine. The explosion of the mine overturns or destroys the vehicle and people are thrown out and onto the ground, thus suffering blunt injuries.

Beatings

Mistreatment of prisoners, officials, "suspected sympathizers" or other civilians is, alas, all too common.

3.1.2 Penetrating wounds caused by weapons of war

A moving projectile has kinetic energy. When the projectile enters the human body, it releases energy to the tissues thus causing a wound. There are two types of wounding projectiles: bullets and metallic fragments (or splinters, shards, and shrapnel)[2].

Fragment wounds

Exploding bombs, shells, rockets and grenades, submunitions (cluster bomblets) and some landmines, produce metal fragments from the weapon casing (primary projectiles). In the past these fragments were usually irregular in size and shape. In many modern weapons, however, the inner lining of the casing is scored, creating weak points that break off easily on explosion. This generates a controlled fragmentation of large numbers of pre-formed fragments that are regular in size and shape, and usually less than 1 g in weight. In other, usually improvised bombs and explosive devices, nails, bolts, steel pellets or other metallic debris surround the explosive material.

T. A. Voeten / ICRC

Figure 3.1

Different types of shells and explosives.

---

[2] The word "shrapnel" is derived from the name of Major-General Henry Shrapnel (1761–1842), an English officer, who designed a new type of artillery shell. The term originally referred only to the metal balls dispersed when a shrapnel shell bursts, but is often used to describe metal fragments intentionally included in explosive devices. For shells, bombs or other munitions, the technical term for these particles is fragments, splinters or shards, fragments being the preferred name in scientific documents and the term that is used in this manual.

Fragments are shot off at a very high speed, which decreases rapidly with the distance travelled because of their non-aerodynamic character. The farther the victim is from the explosion, the more superficial the wounds. At very close range, multiple high-energy fragments combined with the blast effect create mutilating injuries and are often fatal.

Explosions may cause stones or bricks to break up, shatter glass panes, or the force of the blast wind may transport other debris, also producing penetrating fragments (secondary projectiles), as described in Section 3.1.4.

Fragment wounds are usually multiple and the wound tract is always widest at the entry. There may or may not be an exit wound.

Bullet wounds

Handguns and military assault rifles shoot bullets at high speed. Bullet wounds seen in the injured are usually single with a small entry; multiple bullet injuries are more likely to kill. There may be no exit wound but, if there is, the size is variable. The amount of tissue damage depends on a number of factors.

According to international humanitarian law – the Third Hague Convention of 1899 – all bullets used during armed conflict are supposed to remain intact. They are designed to wound, not kill. They have a full metal jacket (usually copper or plated steel) and should be so manufactured as to prevent any fragmenting when they hit the human body. Because of various ballistic effects, some bullets do break up into fragments in the body.

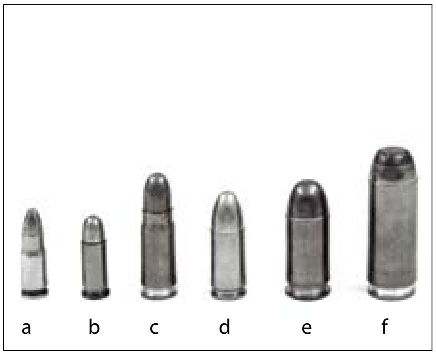

B. Kneubuehl

Figure 3.2.1

Examples of pistol ammunition:

a. 5.45 x 19 mm

b. 6.35 mm Browning

c. 7.63 mm Mauser

d. 9 mm Luger

e. 45 mm automatic

f. 50 mm AE (Action Express) semi-jacketed bullet

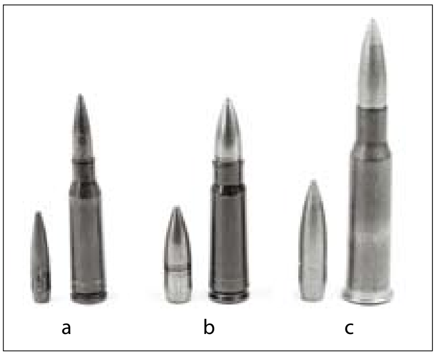

B. Kneubuehl

Figure 3.2.2

Examples of military rifle ammunition:

a. 5.45 x 39 mm Kalashnikov

b. 7.62 x 39 mm Kalashnikov

c. 7.62 x 54R Dragunov

B. Kneubuehl

Figure 3.2.3

Semi-automatic pistol 9 mm Luger (SIG-Sauer P 228).

B. Kneubuehl

Figure 3.2.4

7.62 x 39 mm AK-47 Kalashnikov military assault rifle.

Cut wounds, *"arme blanche"*

Apart from the modern soldier's bayonet, knives, machetes or *pangas*[3] may be used by combatants in "traditional" warfare in some societies. These produce incised or puncture wounds.

3.1.3 Anti-personnel landmine wounds

Anti-personnel landmines (APM) come in two basic types: blast mines with a pressure plate that explode when someone steps on them, and fragmentation mines which explode when a tripwire is touched. Unexploded ordnance is often left on the battlefield and has similar injuring effects as fragmentation mines (the lethal remnants of war consisting of cluster bomblets, bombs and shells that did not explode when fired). These weapons continue to kill and injure large numbers of civilians long after the end of hostilities and have widespread humanitarian and economic repercussions.

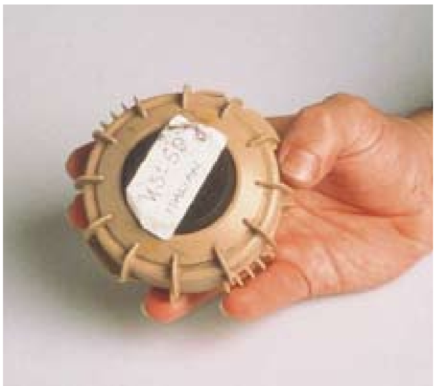

ICRC

Figure 3.3.1

Blast mine.

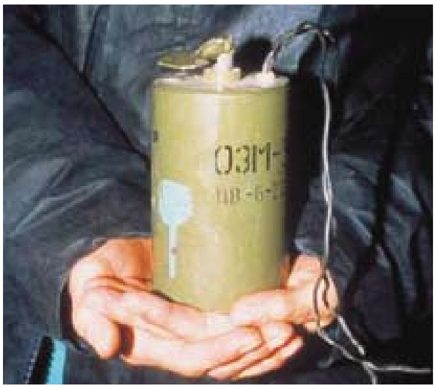

ICRC

Figure 3.3.2

Fragmentation mine.

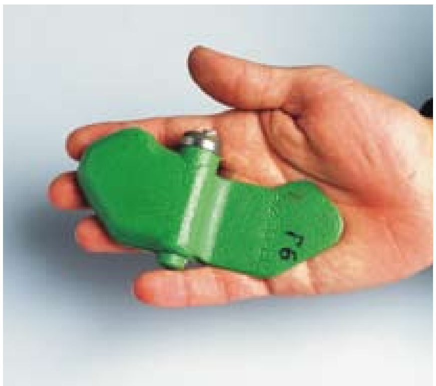

ICRC

Figure 3.3.3

PMF "Butterfly mine".

Patterns of injury

Anti-personnel mines cause three distinct patterns of injury according to the blast effect or the production of fragments.

Pattern 1

A person steps on the pressure plate of a blast mine. The explosion and local primary blast effect cause a traumatic amputation or severe injury of the contact foot and leg. There may be wounds to the other leg, genitals, abdomen or pelvis, and the contralateral arm. The severity of the wound depends on the amount of explosive in the mine compared to the body mass (Figure 3.4).

Pattern 2

A person touches the tripwire attached to a fragmentation mine, which provokes its detonation. Fragmentation mines cause the same injuries as other fragmentation devices, such as bombs or grenades, and the severity of the injury depends on the distance of the victim from the explosion.

Pattern 3

A person handles a mine: setting a mine; trying to clear a mine; a child playing with a mine. The explosion causes severe injury to the hand and arm, and frequent injuries to the face and eyes or chest.

For further information on anti-personnel landmine injuries, see Volume 2.

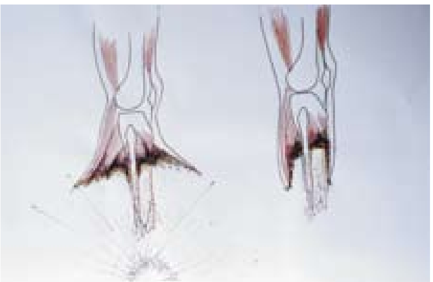

R. Coupland / ICRC

Figure 3.4

"Umbrella" effect of blast mine: superficial tissues suffer less damage than deep-tissue planes. The wounds are all severely contaminated by mud, grass, and pieces of the shoe driven deep into the tissues at the moment of explosion.

---

[3] In some countries where a "traditional" form of warfare is still practised, a large, heavy knife – the machete or *panga* – is often used as a weapon. The overhead motion strikes the victim on the head, neck or shoulder.

3.1.4 Explosive blast injuries

The detonation of high-energy explosives creates a travelling pressure blast wave in the air, (or in water for underwater explosions). The wave causes rapid and large changes in the outside atmospheric pressure: the positive pressure shock wave is followed by a negative pressure phase (Figure 3.5). Immediately behind the pressure wave is a mass movement of air: the blast wind.

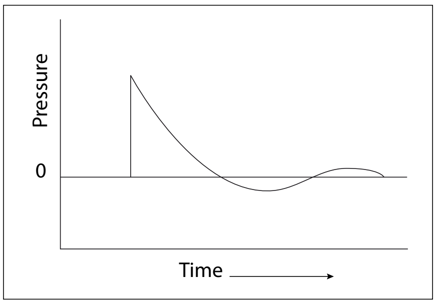

Figure 3.5

Blast wave overpressure followed by negative pressure trough.

As the blast wave passes any unprotected person it affects all parts of the body, especially those normally containing air. A victim of blast may not have any external injury. A single, large explosion may injure many people at the same time. Explosions in a closed space (building, bus, etc.) are more fatal than those out in the open.

Categorization of blast injuries

Blast injuries are commonly categorized into four types.

Primary

These are direct pressure effects. Rupture of the tympanic membrane is the most common injury. Rupture of the lung alveoli and their capillaries ("blast lung") is the most lethal injury amongst survivors.

Secondary

These injuries include missile wounds. Fragments may arise from the bomb casing or contents (primary missiles). Home-made bombs (improvised explosive devices, known as IED) may be packed with nuts and bolts, screws and ball bearings. In addition, the blast wind may mobilize various objects (secondary missiles) that then cause penetrating wounds.

Tertiary

These effects are directly due to the blast wind. It can cause total body disintegration in the immediate vicinity of the explosion; traumatic amputations and evisceration further away. The wind can make buildings collapse or throw people against objects. Trauma may be blunt, crush or penetrating.

Quaternary

Miscellaneous types of harm due to burns, asphyxia from carbon monoxide or toxic gases, or the inhalation of dust, smoke or contaminants.

The various injuries caused by major blasts cover a whole spectrum of trauma and many patients suffer several injuries from a variety of effects: i.e. multiple injuries from one single weapon system.

Other situations of blast injuries

The blast wave travels more rapidly and much further in water than in air. Blast injuries in water occur at greater distances and can be more severe. Moreover, underwater explosions tend to cause pure primary blast injury. A "fuel air explosive" – the liquid explosive material is dispersed in the air like an aerosol and then ignited – also tends to bring about pure primary blast injury and quaternary effects due to the consumption of all oxygen in the nearby air.

The blast effect of a marine mine exploding below an icy surface, or the "deck slap" of a ship hit by a torpedo, produces a shock wave that can severely fracture the bones of anyone on deck or inside the ship. Similarly, some anti-tank mines send a blast wave through the floor of the vehicle causing closed fractures of the foot and leg. The foot appears like a "bag of bones" inside intact skin, which was described in World War I as "*pied de mine*": a mine foot. Anti-personnel blast mines have a local blast effect, vaporizing the tissues of the contact foot, as described previously.

For further information on blast injuries, see Volume 2.

3.1.5 Burns

A large explosion may cause flash burns or the carbonization of tissues. Bombing may start secondary fires in buildings and an anti-tank mine may ignite the petrol tank of a vehicle. Burns are common amongst the crews of tanks, ships and aircraft hit by missiles. Some types of anti-personnel blast mines provoke burning as well as traumatic amputation of the limb.

Certain weapons cause specific burns: napalm and phosphorus bombs, magnesium flares.

3.1.6 Non-conventional weapons

International humanitarian law prohibits the use of chemical and biological weapons. Nonetheless, many countries have stocks of them. If not used during actual combat, they may nevertheless be released when storage warehouses are bombed.

Radioactive agents have come into greater use recently, such as depleted uranium in anti-tank shells. Residual radioactive shells and debris constitute a long-term hazard. A bomb surrounded by radioactive material – the so-called "dirty bomb" – is not a nuclear bomb. The explosion is caused by conventional means but radioactive material is spread out over a wide area. The bombing of nuclear medicine and other laboratory facilities, or nuclear power plants, may also release radioactive material into the atmosphere.

The reader should refer to standard military texts and those of the Organization for the Prohibition of Chemical Weapons (OPCW), and relevant documents of the World Health Organization for further details. Nuclear weapons shall not be discussed here.

3.2 Ballistics

3.2.1 Introduction

Blast and penetrating missile wounds occur in times of armed conflict and whenever the weapons of war are used in peacetime. Weapons cause specific but variable patterns of injury. While standard surgical techniques will suffice to treat simple wounds, the management of war wounds produced by high-energy weapons is based on an understanding of the mechanisms by which projectiles cause injury: wound ballistics. Only by understanding certain physical phenomena can the surgeon appreciate the different varieties of wounds seen in armed conflict and the difference between these wounds and the trauma that is seen in everyday civilian practice.

Although the study of ballistics may be "interesting" in its own right, the clinician does not always know what weapon inflicted the injury; and *never* knows the energy available at point of impact. One can only estimate the transfer of energy in the tissues from the extent of tissue damage.

The study of ballistics gives us an understanding of the basic mechanisms at work during wounding. The importance of this knowledge lies in the fact that projectile injuries should be neither under- nor over-treated. The clinical assessment of the actual wound is the most important factor determining management, and an understanding of ballistics allows the surgeon to better understand the pathology and assess the injuries that he sees, rather than explain every wound and determine specific treatment. "Treat the wound, not the weapon"[4] is not a vain concept.

Basic definitions

Ballistics is the part of the science of mechanics that studies the motion and behaviour of a projectile, and its effects on a target.

Internal ballistics

Internal ballistics deals with the processes that occur inside a gun barrel when a shot is fired: gas pressures on combustion of the propellant powder, the energy and heat released, and the course of the projectile in the barrel, are just some of its concerns.

---

[4] Lindsey D. The idolatry of velocity, or lies, damn lies, and ballistics. *J Trauma*. 1980; **20**: 1068 –1069.

External ballistics

External ballistics describes the projectile trajectory once it has left the barrel. The influences affecting the flight include gravity, air resistance and crosswind deflection, stability of the projectile (spin and yaw), as well as any contact prior to the projectile reaching the target, known as ricochet.

Terminal ballistics

Terminal ballistics describes what happens when the projectile hits the target, as well as any counter-effects produced by the target on the projectile. If the target is biological tissues, terminal ballistics is called *wound ballistics* and describes the effects on the tissues.

3.2.2 Internal ballistics

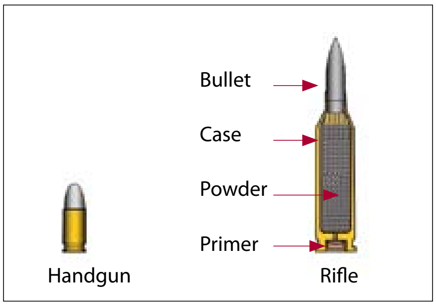

B. Kneubuehl

Figure 3.6

Main parts of a bullet cartridge.

Bullets: fundamental concepts

Figure 3.6 shows the main components of a bullet cartridge. The primer is struck by a mechanism in the gun to produce a small detonation and flame that sets off the powder in the case. This causes a very rapid burn with the production of a large volume of rapidly expanding gas, which pushes the bullet out of the barrel of the gun. The velocity with which the bullet leaves the barrel is called the muzzle velocity.

Bullet construction

Bullets are classified according to a number of parameters; one is their manufactured composition (Figure 3.7). They vary according to their calibre and mass.

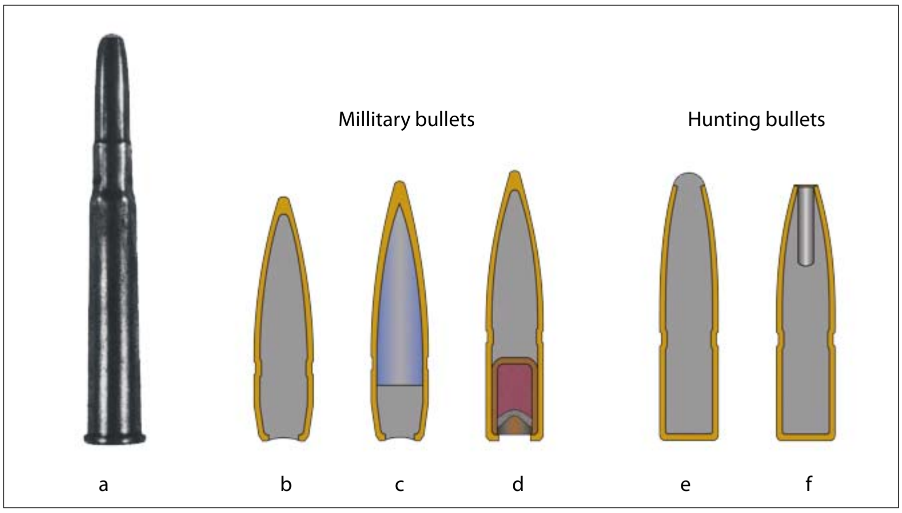

B. Kneubuehl

Figure 3.7

Bullets according to construction.

a. 303 British Mark II produced in Dum Dum, India, 1896 (see footnote number 6).

b. Full metal jacket (FMJ): military-issue bullets have a soft core (lead) with a full metal jacket of copper. This standard has been established by various conventions and treaties of IHL, especially the Third Hague Convention of 1899.

c. Hard core bullet: the lead has been replaced by steel or tungsten for better penetrating power.

d. Tracer bullet: contains a pyrotechnical substance at the tail-end that burns up in flight and illuminates the trajectory. Used to identify and pinpoint the target.

e. Semi-jacketed bullet (SJ): part of the point is not jacketed, thus exposing the lead inside. This is supposed to be used only for hunting; its use by combatants during armed conflict is illegal.

f. Semi-jacketed hollow-point (SJ-HP): also a hunting bullet with a hollow nose tip; also illegal during armed conflict.

Apart from the hunting bullets mentioned in Figure 3.7, special rifles and their ammunition exist for hunting: shotguns whose shells give off multiple lead or steel shots (pellets) available in diameters from 2 to 9 mm (Figure 3.8).

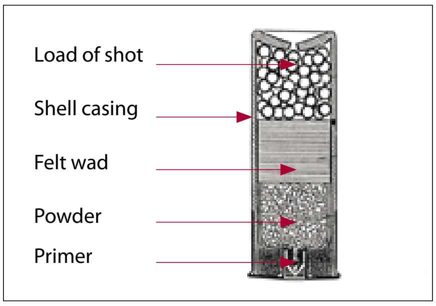

B. Kneubuehl

Figure 3.8

Shotgun shell and load of shot.

Hunting bullets are not supposed to be used by combatants in armed conflict. These bullets are constructed in such a way as to kill rather than wound. The idea is to be more "humane" when hunting wild animals and to have a "rapid and clean kill". Nonetheless, the surgeon will be faced with wounds caused by such weapons either because of accidents, criminal activity, or because combatants have illegally used them during combat.

The international legal limits placed on bullets used by the military do not apply to domestic use during peacetime, and the civilian surgeon may at times face more devastating gunshot wounds than the military surgeon on the battlefield.

Muzzle velocities

Firearms are classically divided into two categories: high-velocity (rifles) and low-velocity (handguns).The latter fire relatively heavy bullets at low muzzle velocities of 150 – 200 m/s. A typical military assault rifle fires a smaller bullet at 700 – 950 m/s. This, however, says nothing about the actual velocity when the bullet reaches the target.

9 mm Luger pistol	350 m/s
.38 Special handgun	260 m/s
.44 Magnum handgun	440 m/s
7.62 x 51 mm NATO rifle	830 m/s
5.56 x 45 mm M16 rifle	960 m/s
7.62 x 39 mm AK-47 Kalashnikov rifle	720 m/s
5.45 x 39 mm AK-74 Kalashnikov rifle	900 m/s
12 gauge shotgun	420 m/s

Table 3.1 Some examples of muzzle velocity from different firearms.

The gun barrel

A bullet is a long projectile that only flies with stability if a high spin around its longitudinal axis is attained, thus giving it a gyroscopic effect. To achieve this spin, gun barrels are constructed with interior spiral grooves (twists). This is a "rifled" barrel, used in all handguns and high-velocity rifles (Figure 3.9).

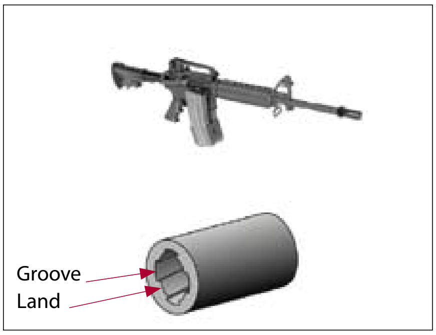

B. Kneubuehl

Figure 3.9

Cross section of a rifled gun barrel:
5.56 x 45 mm M-16 A4 military assault rifle.
The grooves and lands (the elevated sections between the grooves) are shown.

The barrels of shotguns are smooth on the inside; they are not "rifled". This limits their accuracy and distance of fire (Figure 3.10).

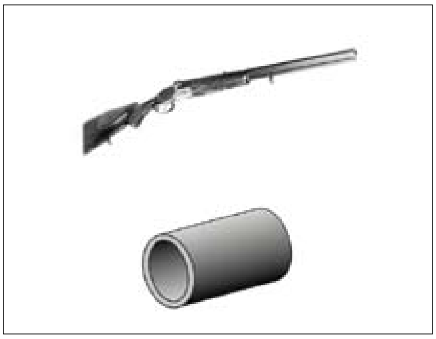

B. Kneubuehl

Figure 3.10

Typical hunting shotgun with smooth barrel.

Design of firing

Another classification of firearms is according to their design of firing. A weapon operated with a single hand is a "handgun". If the barrel and cartridge chamber are in one piece, this is a "pistol". If several chambers rotate behind the barrel, it is called a "revolver". If both hands are required to operate the weapon, it is referred to as a "rifle" or "shotgun".

Firing capacity defines how individual shots are fired. For a single-shot weapon, every shot is loaded individually. A repeating weapon includes a magazine that holds a series of cartridges which load manually one after the other. If the loading movement is repeated automatically after every shot, but every shot is fired individually, this is a semi-automatic weapon. If the automatic loading of cartridges allows several shots to be fired with one pull of the trigger, it is known as an automatic weapon.

In modern military usage, most weapons are automatic rifles and sub-machine guns or semi-automatic pistols.

3.2.3 External ballistics

Flight in air

Once shot off, several variables affecting the projectile in flight tend to influence its stability, accuracy of targeting, or velocity. The most important variables include the following.

- Spin around its own longitudinal axis, giving it gyroscopic stability.
- Distance travelled and the effect of gravity, including the shooting angle – is the bullet shot upwards or is the trajectory flat?
- Air drag: friction is responsible for 10 % of total area drag, while 90 % is due to pressure effects; faster bullets are retarded proportionately more – rifle bullets are aerodynamically shaped to decrease air drag; this is not the case with shotgun cartridges or metal fragments.
- Crosswind deflection.
- Raindrops hitting the bullet.
- Bullet hitting an obstacle before hitting the target.

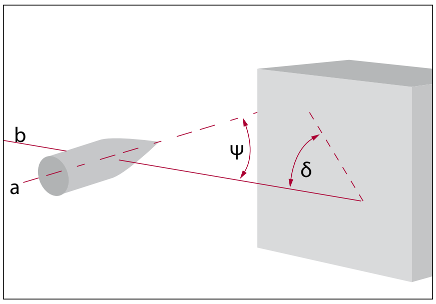

B. Kneubuehl

Figure 3.11

Angle of incidence and angle of impact:

a. longitudinal axis of bullet
b. direction of flight

Ψ. angle of incidence
δ. angle of impact

Yaw

A rifle bullet in flight does not travel in a simple straightforward motion. Because it is a spinning gyroscope, the bullet "wobbles" and undergoes some very complicated movements (nutation, precession), the most important being yaw: the nose of the bullet moves up and down, away from the line of flight, and describes an angle with the target surface at impact (Figure 3.11). This influences the wounding potential of the bullet, as an important degree of yaw will destabilize the motion of the bullet inside the target.

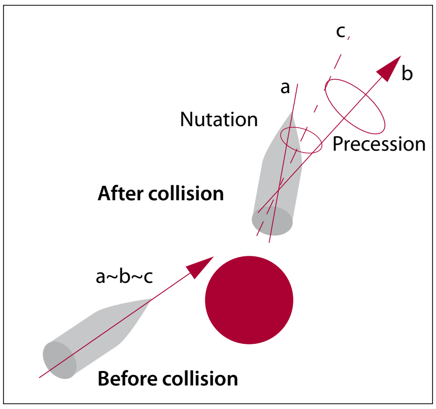

B. Kneubuehl

Figure 3.12

Ricochet: effect on a bullet after collision with an obstacle.

Ricochet

A bullet may hit an obstacle during its flight. This may be a branch of a tree, a belt buckle, a concrete wall or the ground, or a soldier's helmet or flak jacket. This causes a ricochet; the bullet is given a small "push" that destabilizes it (Figure 3.12). There is an increase in the deviation of the nose of the bullet (yaw); the bullet may even tumble, rotating around itself. Such a large yaw angle at point of impact, as mentioned above, will have important consequences for the wounding potential of the bullet in the target.

In addition, if the "push" is great enough, the bullet can be deformed or even broken up into fragments before hitting the target.

3.2.4 Summary

There are, therefore, quite a number of variables involved in determining the characteristics of a projectile *before* it hits the target and these will influence its behaviour in the target and the efficiency in creating wounds. They include:

- velocity at moment of impact and any residual velocity if there is an exit;
- mass, form and shape of the projectile, bullet construction;
- type of weapon, handgun or rifle;
- stability of the projectile in flight;
- any yaw at moment of impact.

3.3 Terminal ballistics

3.3.1 Role of kinetic energy

Any moving object has kinetic energy, whether it is a hand-wielded knife or club, a bullet fired from a gun, or a fragment given off by the explosion of a shell. This energy is described by the well-known formula:

$$E_K = ½ \, m \, v^2$$

This defines the total kinetic energy that the object possesses, but does not define the kinetic energy expended when the projectile passes into and through a target. For a bullet or fragment that does not change its mass, this is calculated from the difference in velocity between entry and exit.

$$E_{K\,EXP} = \frac{m\,(v_1^{\,2} - v_2^{\,2})}{2}$$

If there is no exit, then $v_2 = 0$, and all the kinetic energy has been transferred. If the bullet breaks up into fragments, then the mass (m) also changes, and $E_{K\,EXP}$ is modified.

Total kinetic energy is the potential for causing damage; transferred kinetic energy is the capacity to cause damage. Actual tissue damage, however, depends on the efficiency of this energy transfer, which is determined by many other variables, as will be shown.

Weapons according to E_K
Yet another classification of weapons is based on the amount of kinetic energy available for transfer.

- Low energy: knife or hand-energized missiles.
- Medium energy: handguns.
- High energy: military or hunting rifles, with a muzzle velocity above 600 m/s.

For metal fragments given off by an explosion the initial velocity is very high; this decreases rapidly with distance. The wounding potential depends on the mass of the fragment and the distance of the victim from the weapon.

3.3.2 Laboratory experiments

Many researchers have performed different experiments to describe the effect of projectiles on tissues. Targets have included human cadavers, various animals (pigs, dogs, goats) and tissue simulants.

Tissue simulants are specially-prepared blocks of gelatine or glycerine soap, which have a density and/or viscosity close to that of muscle tissue. Soap is plastic and any deformation remains unchanged, representing the maximum effects. Gelatine is elastic and deformations disappear almost entirely; they are studied with high-speed cameras. If the stress on gelatine exceeds the limit of its elasticity then it will crack and tear, showing fracture lines.

Compared to human cadavers and animals, tissue simulants have the advantage of allowing repetitions of the experiment, changing only one variable at a time. These laboratory experiments are all approximations, however, of what happens in a live human being.

Surgeons working for the ICRC have collaborated over many years with the ballistics laboratory of the Swiss Federal Department of Defence.[5] This laboratory uses gelatine and glycerine soap for its ballistics experiments. Its results have been confirmed by comparison with clinical cases treated by ICRC surgeons in various war zones throughout the world.

---

[5] See footnote 1.

Following these laboratory studies, five categories describe the behaviour of projectiles: non-deforming and deforming bullets, be they from rifles or handguns, and fragments.

3.3.3 Non-deforming rifle bullet: full metal jacket (FMJ) military bullet

When a standard military bullet hits a soft object while in stable flight, it produces a "shooting channel" that presents three distinctive phases: narrow channel; primary temporary cavity; and terminal narrow channel (Figure 3.13).

Figure 3.13

Non-deforming rifle bullet (FMJ) in soap.

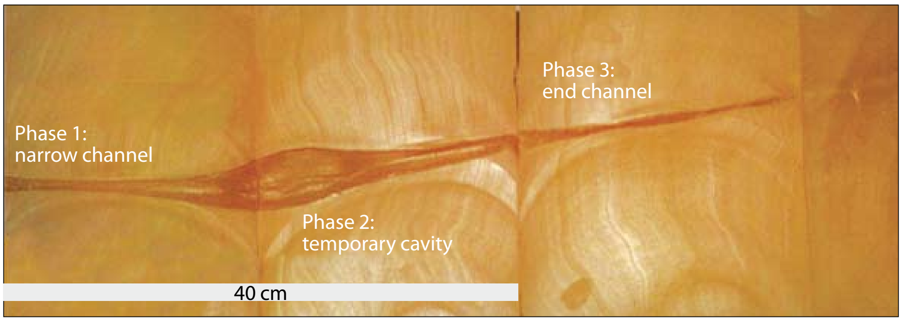

B. Kneubuehl

Phase 1
Straight narrow channel with a diameter about 1.5 times the calibre of the bullet. The greater the velocity, the wider the channel. Different types of bullets have a different length of narrow channel, typically 15 – 25 cm.

Phase 2
The channel opens into the "primary temporary cavity". The reported diameter of this cavity is anywhere between 10 and 15 times the calibre of the bullet.

Figure 3.14 shows the motion of the bullet in gelatine. It starts to yaw, and tumbles, turning a full 270°, and then advances with its tail end forward. The whole lateral surface of the bullet comes into contact with the medium, which greatly decelerates the bullet and subjects it to an important stress.

Figure 3.14

The bullet tumbles in the gelatine or soap: it rotates 270° around a transverse axis that is perpendicular to the long axis. (Graphic demonstration of bullet superimposed on soap blocks. The proportions between bullet and trajectory are exaggerated for the sake of clarity.)

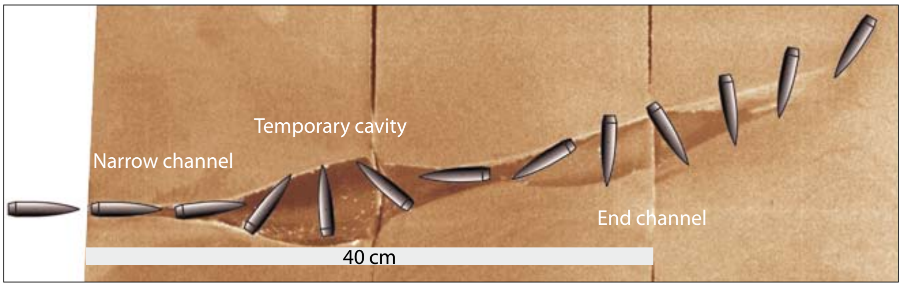

B. Kneubuehl

The slowing of the bullet involves a release of kinetic energy, which, in turn, causes the rapid and massive displacement of gelatine radially outwards, producing a cavity *behind* the projectile; there is a slight delay in the production of the cavity because of the inertia of the gelatine mass. There is an almost complete vacuum in this cavity that quickly sucks in air from the entry hole, and from the exit hole if there is one. The cavity collapses after a few milliseconds, only to reform again, with a smaller volume. The cavitation continues until all the transferred energy is used up: the cavity pulsates! In water or gelatine, there are up to 7 – 8 pulsations, in biological tissues usually 3 – 4.

The diameter of this cavity depends upon the elastic properties of the medium, as well as the amount of kinetic energy transferred. Fissures radiating from the shooting channel indicate that the shearing effect of the cavity has exceeded the elasticity of the gelatine.

Phase 3

The tumbling slows down and the bullet continues in a sideways position at a considerably reduced speed. In some cases, a narrow straight channel is observed; in others, the tumbling seems to continue but backwards, the bullet again assuming a lateral position, and a second cavity occurs. This does not reach the size of the primary temporary cavity. The bullet then creeps forward and finally stops, always with the tail end facing forwards.

In an elastic medium such as glycerine, what remains in the shooting channel at the end of the process, and all temporary effects, is called the "permanent channel".

References to these basic definitions of phases of the shooting channel will be made throughout this chapter.

Please note:

These three phases are found with all FMJ rifle bullets, yet each bullet has a specific shooting channel. The narrow channel of the 7.62 mm AK-47 is long (15 – 20 cm), while the 5.45 mm AK-74 produces a narrow channel of less than 5 cm before cavitation begins.

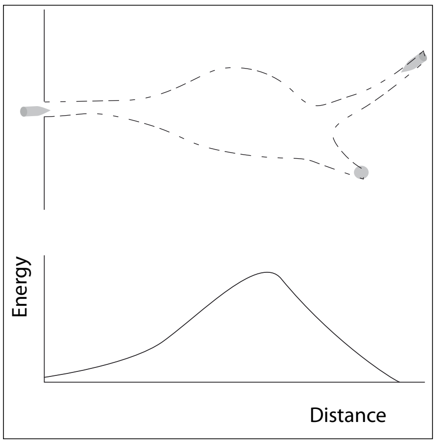

R. Coupland / ICRC

Figure 3.15

The sketch shows the position of the bullet and the extent of the shooting channel at different phases. The graph represents the transfer of kinetic energy along the bullet track: bullet fragmentation occurs at the peak of energy transfer.

Tumbling – yaw – in the target medium

Given a long enough shooting channel, all FMJ rifle bullets tumble. How early the tumbling motion begins determines the length of the narrow channel, and the onset of cavitation. This depends on the stability of the bullet – yaw – at the point of impact. The less stable the bullet in flight, the greater the yaw, which rapidly brings a larger bullet surface into contact with the medium, leading to early tumbling and a short narrow channel. How early FMJ bullets tumble is also a function of their manufacture: mass, centre of gravity and shooting distance, etc.

Fragmentation of FMJ bullets

It is during phase 2 – the temporary cavity – that certain bullets deform or even break apart because of the enormous stress exerted on them. This occurs when the bullet-medium interface is at its maximum, the cavity is widest, and the transfer of kinetic energy is highest (Figures 3.15 and 3.16). If the projectile acts on the target medium, this is a good example of the medium acting on the projectile. This fragmentation occurs only at short shooting distances, up to 30 – 100 m, depending on the bullet's construction and stability.

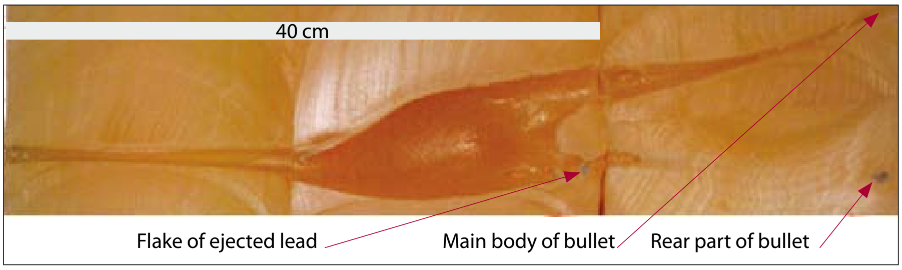

B. Kneubuehl

Figure 3.16

Fragmentation of FMJ bullet during phase 2, i.e. temporary cavitation.

The bullet flattens on its sides, bends in the middle and, finally, the jacket splits and the lead inside spills out (Figure 3.17). This gives the "shower of lead" picture often seen on X-rays (Figures 3.35, 4.5 and 10.5). If the bullet breaks, the smaller fragment typically deviates downwards. The fragmentation depends on the construction of the bullet and the velocity; with impact velocities below 600 m/s, no FMJ deforms or breaks apart.

If the bullet does fragment then the primary temporary cavity is larger than is the case with an intact bullet. This represents a much larger transfer of kinetic energy and has important clinical repercussions.

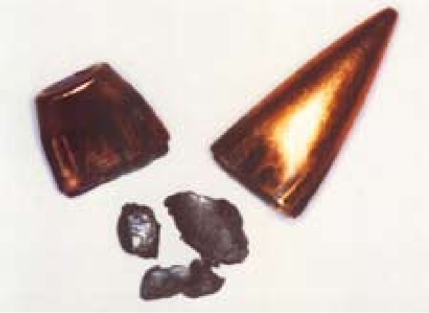

ICRC

Figure 3.17

Jacket split apart, exposing the lead within.

3.3.4 Deforming and fragmenting rifle bullets: dum-dum[6]

Some bullets (e.g. hunting bullets) are constructed to always deform or break up into fragments: hollow-nosed, semi-jacketed (SJ), soft-pointed, etc. (Figure 3.7 e and f). These are usually grouped under the term "dum-dum" bullets and are illegal for military use according to international law.

Deforming bullets are made to change their shape readily ("mushrooming"), thus increasing their cross section area, but without losing mass. The spent bullet weighs as much as the original. They are used mainly in handguns with a muzzle velocity below 450 m/s, and are available for domestic use (special police forces *and* criminals). Fragmenting bullets, on the other hand, break up and lose mass, and create a "wall" of particles, thus increasing the total effective area of cross section. They are used for hunting.

Figure 3.18 demonstrates a semi-jacketed rifle bullet mushrooming immediately on impact with a soft medium. The increased cross section causes a greater bullet-medium interaction; the bullet slows down quickly, releasing kinetic energy very early. The narrow channel almost completely disappears and the temporary cavity occurs right after impact. At first, the cavity is almost cylindrical; it then decreases conically.

Figure 3.18

Deforming semi-jacketed rifle bullet (SJ) in soap. The bullet "mushrooms" just after impact, then continues in a linear trajectory. (Graphic demonstration of bullet superimposed on soap blocks.)

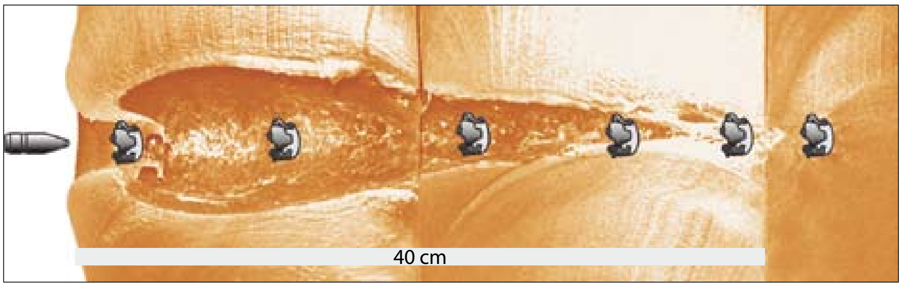

B. Kneubuehl

The principal difference between a full metal jacket and semi-jacketed bullet is at what depth of penetration the maximum energy transfer occurs in the shooting channel. The volume of the cavities is the same in both examples in Figure 3.19 indicating equal transfer of kinetic energy.

Figure 3.19

Soap blocks showing comparison between ballistic profiles of FMJ and SJ bullets – the transfer of kinetic energy occurs far sooner with the SJ bullet.

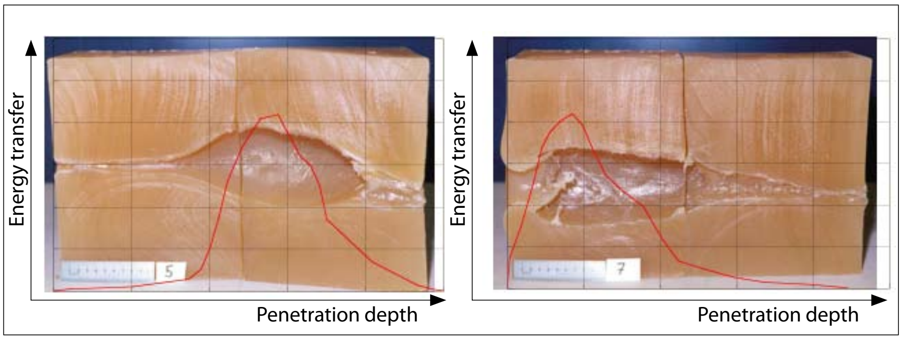

B. Kneubuehl

---

[6] The Indian Ordnance Department of the British Army developed a bullet for its colonial troops in 1897 at an ammunition factory in the town of Dum Dum, north-east of Calcutta (Kolkata) because the previous one was judged to be ineffectual and caused insignificant injuries. The bullet had a round nose with a full metal jacket of copper-nickel covering a lead core, with the exception of 1 mm at its tip where the lead was bare. The bullet was used against the Afridi Afghans in 1897 – 98, and again against the Mahdi Army at Omdurman, Sudan, in 1898 with devastating effect. The bullet was judged to be "inhumane" according to the Declaration of St Petersburg of 1868 and was banned by the Hague Convention of 1899. Projectiles causing unnecessary injuries ("*maux superflus*") were not allowed. As an example, the Conventions mentioned projectiles whose metal jacket did not fully cover the lead core. Since then, all projectiles possessing the same qualities (deformation by expanding or flattening easily, fragmentation) have received the collective name of dum-dum.

The same effect can be demonstrated when using a synthetic bone simulant encased in gelatine (Figure 3.20).

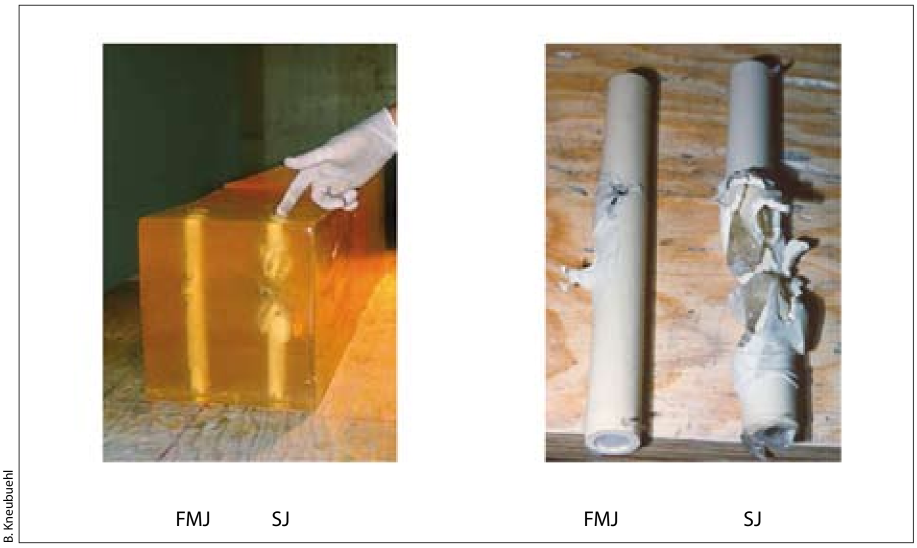

Figure 3.20

Comparison of FMJ bullet versus SJ bullet: synthetic bone at a shallow depth of gelatine. The FMJ bullet breaks the bone simulant during the narrow channel phase; the shooting channel is practically the same with or without the bone simulant. At the same depth, the SJ bullet shatters the bone simulant completely.

Ricochet effect

When a FMJ bullet strikes an obstacle before hitting the target it is destabilized. After impact, there is almost no narrow channel and the shooting channel resembles what occurs with a deforming or dum-dum bullet (Figure 3.21). This phenomenon has important clinical consequences.

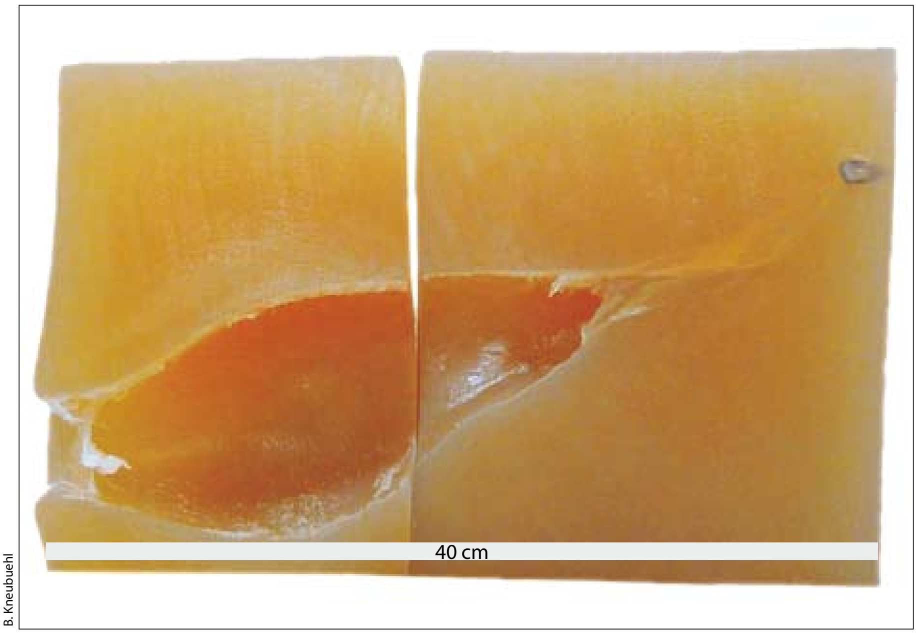

Figure 3.21

FMJ rifle bullet after ricochet effect in soap. The large impact angle after ricochet destabilizes the bullet, which tumbles easily and early in the shooting channel. Note that the cavitation occurs almost immediately on impact, similar to an SJ bullet.

Please note:

Jacket construction is not the only decisive factor in bullet behaviour be it full or semi-jacketed. It is quite possible to manufacture a bullet that will fragment at high velocities, deform at medium velocities, and retain a stable shape at low velocities.

3.3.5 Handgun bullets

Handgun bullets may be much heavier than rifle bullets.

Non-deforming bullet

The shooting channel of a non-deforming bullet shows little yaw and no tumbling; the bullet remains with the tip pointing forward with deep penetration (Figure 3.22). Temporary cavitation is long and narrow.

Figure 3.22

Standard military pistol bullet, full metal jacket in soap, no tumbling of the bullet. (Graphic demonstration of bullet superimposed on soap blocks.)

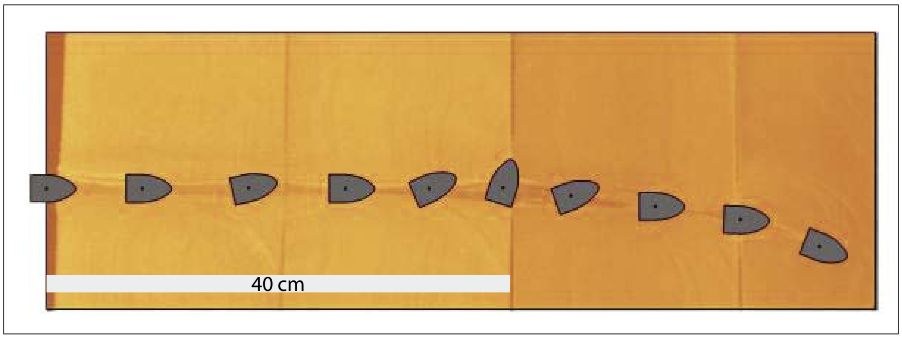

B. Kneubuehl

Deforming bullet

A soft point, deforming pistol bullet – as used by special police forces – mushrooms at entry (Figure 3.23). The large cross section area causes an abrupt reduction in velocity and transfer of kinetic energy with an immediate, and large, temporary cavity.

Figure 3.23

Deforming handgun bullet in soap: mushrooming effect. (Graphic demonstration of bullet superimposed on soap blocks.)

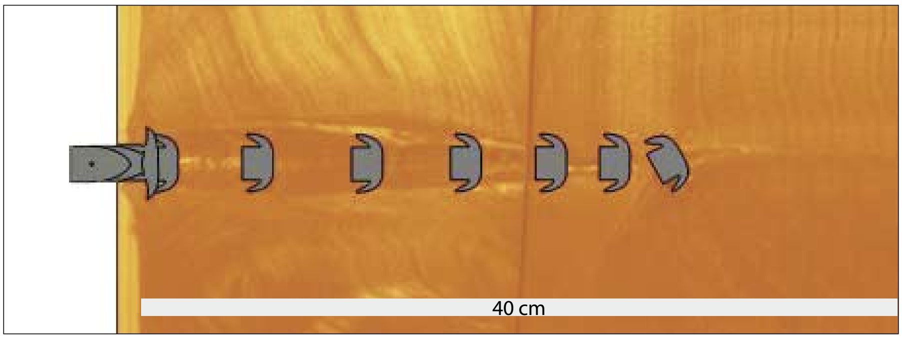

B. Kneubuehl

3.3.6 Fragment projectiles

Fragments given off from the explosion of a bomb, rocket or grenade are not aerodynamic; they have an irregular shape. Their velocity decreases rapidly with distance owing to air drag. They have an unstable flight path with an irregular rotation around an indefinite axis. On impact, the biggest cross section comes into contact with the target surface, immediately transferring a maximum of kinetic energy. No yaw or tumbling occurs in the target.

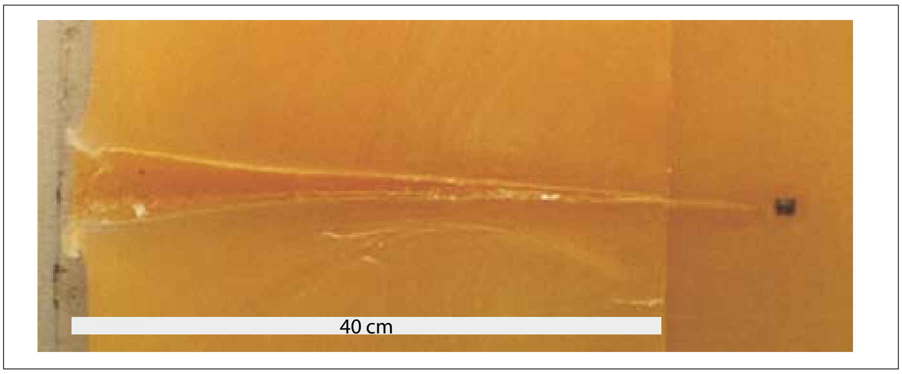

B. Kneuebuehl

Figure 3.24

Fragment profile in soap: the widest point is at entry, the cavity is cone shaped.

The shooting channel always has the largest cavity diameter at the entry, and is larger than the diameter of the fragment. The cavity then decreases constantly in a cone-like fashion (Figure 3.24).

The depth of penetration of the fragment will depend on its kinetic energy, but with a particular relationship between velocity and mass. Figure 3.25 shows two fragments with the same kinetic energy; the volumes of the cones are equal.

Thus, a light but rapid fragment tends to transfer most of its energy shortly after impact on the target; a heavy and slow fragment penetrates further and dissipates its energy along the longer shooting channel.

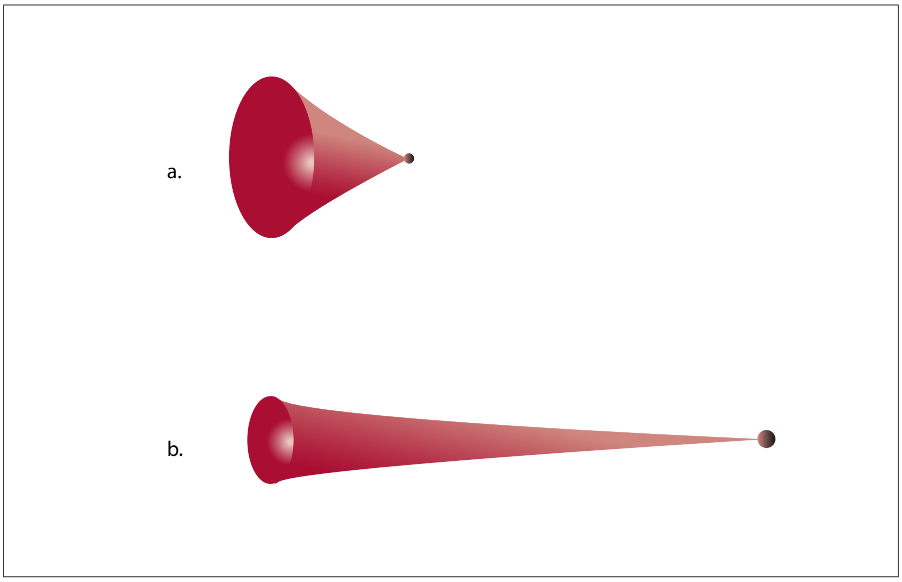

B. Kneuebuehl

Figure 3.25

Two fragments with the same kinetic energy – note the difference in energy deposition along the track, demonstrated by the difference in the cavities:

a. lightweight and fast fragment;
b. heavy and slow fragment.

3.4 Wound ballistics

Moving projectiles cause wounds by a transfer of kinetic energy to the body that destroys, disrupts, and deforms tissue. One studies wound ballistics in order to understand the mechanisms that produce this injury.

Potential for wounding is one thing; the *actual* wound facing the surgeon is another.

3.4.1 Laboratory studies *versus* human body

The experiments using tissue simulants as described above help one to understand the processes at work. Laboratory models are, nonetheless, only an approximation and show only the physical part of the process. The tissue structure of the human body is far too complex for a laboratory model to duplicate injuries. Tissue simulants have the disadvantage of being homogeneous media; living tissue is not, as described later. The laboratory results must be compared to clinical cases for confirmation and this has been the ICRC procedure in conjunction with the Armasuisse laboratory in Thun, Switzerland.

In the real world of clinical practice there are so many variables involved that one cannot be *predictive*; the surgeon cannot say with certainty that such and such a bullet will always result in a specific wound. One can however be *descriptive* and, after a good clinical examination, the surgeon can better understand the pathology that he sees, how it was created, and what sort of surgical management is required. The type of wound, its anatomic location, and the amount of actual tissue disruption are the determinant clinical factors.

3.4.2 Projectile–tissue interaction

When a bullet strikes a human body there is a projectile–tissue interaction that results in tissue damage, with a reciprocal influence of the tissues on the bullet. This interaction depends on a number of factors, which all result in the transfer of kinetic energy from the projectile to the tissues.

This transfer of kinetic energy either compresses, cuts or shears tissues, causing crush, laceration or stretch. The local energy transfer, at every point along the bullet track, is more important than the total quantity of energy transferred in producing tissue damage.

Tissue disruption is due to compression, cutting and shear: crush, laceration and stretch.

A bullet striking a human body demonstrates the *same three phases* as in the laboratory simulants – if the bullet track is long enough. In the laboratory experiments with glycerine, the "permanent channel" was defined as what remains of the shooting channel at the end of the process and all temporary effects. In biological tissues, the "permanent wound cavity" of the shooting channel is the definitive tissue damage after all temporary effects have been taken into account. This is the wound channel that the surgeon sees and is the final result of the crush and stretch of the tissues.[7]

Crush and laceration injury

A projectile causes physical compression and forcing apart of tissues along the shooting channel – crushing and laceration. This is the immediate physical effect of a penetrating foreign body; the bullet damages tissues with which it comes into direct contact, cutting through tissues as would a knife. This tissue damage is permanent and found in the final wound. This is the principal effect caused by low- and medium-energy weapons, such as handgun bullets.

---

[7] Much confusion has been created in surgical literature by authors calling the immediate crush damage of the phase 1 narrow shooting channel the "permanent cavity", as distinct from the "temporary cavity" where stretch damage occurs. In this manual, in accordance with the Armasuisse laboratory, the permanent wound cavity is the channel that remains at the end of the wounding process and constitutes the sum of crush and stretch injury.

At higher energies, when a bullet tumbles or deforms, a larger area of tissue is exposed to its effective cross section and suffers crush. The track left behind by crush is not equal along its length, but increases with bullet yaw in the tissues.

Stretch injury

Tissues have an elastic tensile strength that resists stretching. When a certain limit is reached, capillaries are ruptured and contusion of the tissues occurs. Beyond a critical limit, the tissues themselves are torn, just as laboratory gelatine shows fracture lines. The tissue damage from stretch may be permanent or only temporary.

Tissue stretch occurs during cavitation, which takes place in all projectile wounds, whatever the energy, type or motion of the projectile, and at all points along the entire shooting channel. There is even a minor cavitation effect during the phase 1 narrow channel.

The volume of the cavity is determined by the amount of energy dissipated and the tissue elasticity-resistance. This cavitation stretch performs work on tissues that have already been injured by crush, thus adding to the local immediate damage. In low- or medium-energy projectiles wounds, it is minimal.

When a bullet tumbles (or deforms or fragments), the release of kinetic energy is much greater and is superimposed on bigger tissue crush; the large phase 2 temporary cavity is the result: a momentary massive displacement of tissue in all directions away from the track of the bullet.

As in gelatine, the cavity pulsates: an elastic acceleration followed by a deceleration of the surrounding tissues: a shearing action. The near vacuum in the cavity sucks in air, contaminants and foreign bodies (textile strands from the clothing, dust, etc.), and bacteria, through the entry and any exit wound.

In high-energy wounds the *volume* of the temporary cavity can be up to 25 times that of the permanent cavity left behind, and anywhere between 10 and 15 times the *diameter* of the bullet. Although this volume is proportional to the kinetic energy dissipated, and will affect the amount of tissue damage, other factors related to the tissues can be even more important in determining the actual extent of injury.

3.4.3 Tissue factors

Resistance to crush, laceration and stretch differs widely among different types of tissues and according to anatomic structures. Tissue elasticity and heterogeneity are important factors in determining much of the bullet–tissue interaction.

Tissue elasticity

Elastic tissues tolerate stretch well, but may still suffer severe crush. The lung and skin have excellent tolerance, and leave relatively little residual damage after stretch. Skeletal muscle and the empty bowel wall have good tolerance. The brain, liver, spleen, and kidney are non-elastic, and shatter when stretched. Fluid-filled organs (heart, full bladder, full stomach and intestines) react badly owing to the incompressibility of the fluid contents; they may even “explode”.

Nerves and tendons are mobile and blood vessels elastic; they are usually pushed out of the way by the cavitation.

Cortical bone is dense and rigid and resists stretch. However, if the cavitation accelerates a large enough muscle mass with sufficient energy the bone is bent beyond its tensile strength and breaks; it may even be shattered violently, especially the diaphysis of a long bone. This phenomenon is an example of a fracture without a direct hit by a bullet. The same mechanism takes place during fracture from a blunt injury, except that the energy transfer bending the bone occurs from the outside.

Tissue heterogeneity

Apart from these considerations of tissue elasticity, there is the particular local anatomy to consider. A block of gelatine approximates skeletal muscle in its elasticity and density, but it is homogeneous. A human limb or the face are a mixture of rigid and elastic tissues comprising fascial compartments enclosing muscles, tendons and ligaments, large blood vessels and nerves, and bones. Not only does each anatomic element have a different elastic tensile strength, but the mixture of elements has its own interactive properties. The various interactions of projectile, soft-tissue organs and bony fragments can become very complicated.

Bound together and boundaries

How tightly tissues are bound together, and to nearby structures such as fascia of varying thickness, also affects the amount of permanent damage a cavity stretch can leave behind. If one side of a structure is fixed and the other side has free movement, a shearing force develops. Fascial planes may also serve as channels for the dissipation of energy – path of least resistance – to more remote tissues.

Boundaries limiting the pulsation of a temporary cavity can be very significant. They include not only fascia and aponeuroses, but also hollow, fluid-filled organs: the brain in the rigid cranium, the heart, the full urinary bladder. An empty stomach will be simply transfixed by a stable bullet. The same bullet will cause a full stomach to "explode".

A graphic laboratory example of boundary effect is found in Figures 3.26.1 and 3.26.2. The two pictures show the results with a high-energy bullet cavity. The apple literally explodes after passage of the bullet. Note that the cavitation effect *always follows* the passage of the bullet; it is not simultaneous.

Figures 3.26.1 and 3.26.2

Demonstration of boundary effect on the temporary cavitation due to a high-energy rifle bullet: the apple literally explodes after the passage of the bullet.

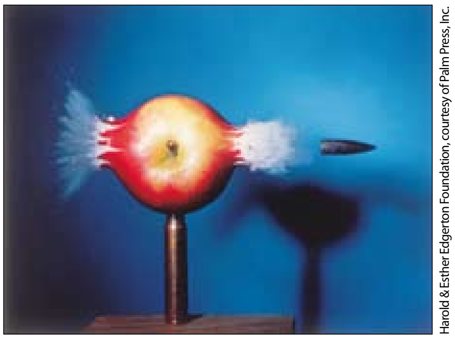

Harold & Esther Edgerton Foundation, courtesy of Palm Press, Inc.

Harold & Esther Edgerton Foundation, courtesy of Palm Press, Inc.

Clinically, these effects can create apparent paradoxes. For example, a heavy and slow bullet can cause a more severe wound in highly elastic tissue, such as lung, than a lighter, faster bullet with more kinetic energy. The heavier, slower bullet produces more crush. The faster, lighter one uses up more energy in creating the temporary cavity, which leaves little residual damage. However, the faster and lighter bullet will create more damage in less elastic tissue (liver, brain), which cannot accommodate stretch as well.

Transfer of kinetic energy does not affect all tissues to the same extent and in the same way.

3.4.4 Pathological description of the permanent wound cavity

The permanent cavity of the final wound that the surgeon sees is the result of a mixture of crushing, laceration and stretching of the tissues. Most of the damage in the majority of ballistic wounds is due to direct crush and laceration.

A number of macroscopic and microscopic changes occur. Severe vasoconstriction blanches the skin around the entry wound for 3 – 4 hours owing to the stretch. This is followed by a hyperaemic reaction that lasts up to 72 hours.[8]

---

[8] Fackler ML, Breteau DVM et al. Open wound drainage versus wound excision in treating the modern assault rifle wound. *Surgery* 1989; **105**: 576 – 584.

In the wound of skeletal muscle itself, three histological zones have been described (Figure 3.27).[9]

1. Crush zone of the wound channel filled with torn and necrotic tissues: 2 – 4 times the diameter of the bullet.
2. Contusion zone of muscle adjacent to the missile tract: the thickness of this zone is variable and averages about 0.5 cm; tissue damage is irregular and uneven.
3. Concussion zone of variable distance with congestion and extravasation of blood: the stretch is not great enough to tear tissues, but sufficient to injure capillaries. The demarcation between the contusion and concussion zones is not always well defined.

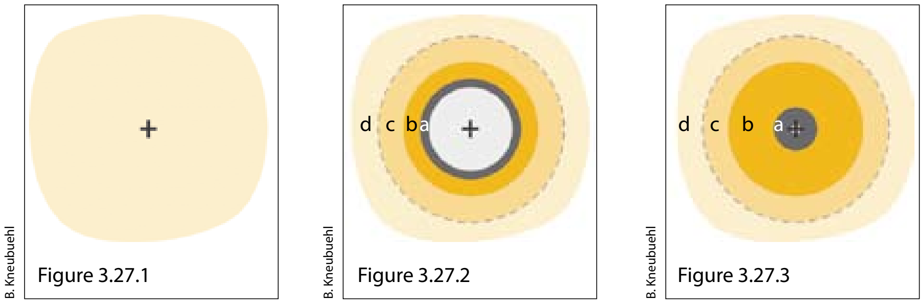

B. Kneuebuehl Figure 3.27.1

B. Kneuebuehl Figure 3.27.2

B. Kneuebuehl Figure 3.27.3

Figure 3.27

Schematic drawing of histopathological changes in the shooting channel:

3.27.1 geometric shooting channel
3.27.2 phase of maximum temporary cavitation
3.27.3 final wound channel

a: zone of crushed tissues
b: contusion zone
c: concussion zone
d: unaffected tissues

These histological changes are more severe with high kinetic-energy transfer – tumbling or deformation of the bullet – and do not change notably over 72 hours. The tissue damage is irregular along the wound track and how much of the damage is irreversible is not immediately apparent. This has obvious consequences for wound management and the extent of surgery necessary (see Chapter 10).

Bacteria are also seen within the permanent cavity. The sucking action of cavitation aspirates bacterial flora from outside, and projectiles are not sterile; the heat of firing is not high enough for long enough to sterilize the bullet.

3.4.5 Clinical applications

The external appearance of a bullet wound can be deceptive. Tiny entrance and exit holes can be associated with extensive internal damage.

The length of the shooting channel in the body, the existence of an exit wound, and the characteristics of the particular structures traversed, will all greatly affect the final wound caused by a high-energy bullet that the surgeon sees.

The position of the phase 2 temporary cavity along the wound track is clinically very relevant. Figure 3.28 demonstrates the trajectory of an FMJ rifle bullet in a laboratory soap block with a simple entry wound and three possible exit wounds.

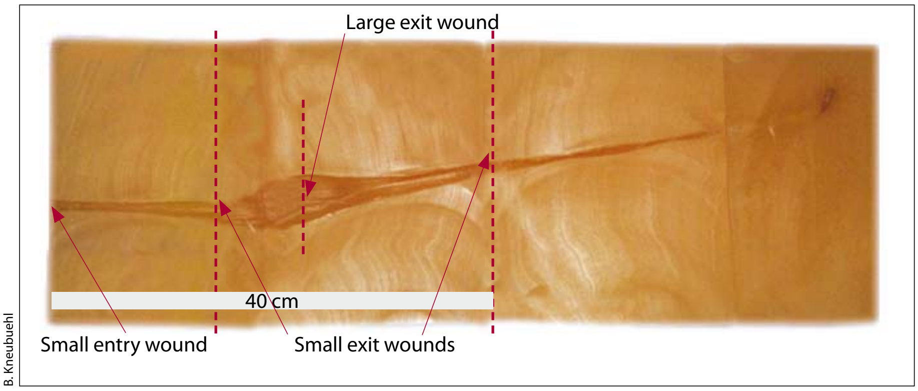

B. Kneuebuehl

Figure 3.28

Exit wounds may occur before, during or after cavitation, depending on the length of the shooting channel in the body.

---

[9] Wang Z, Feng JX, Liu YQ. Pathomorphological observation of gunshot wounds. *Acta Chir Scan* 1982; **508**: 185 – 189.

The injured body part may not be long enough for temporary cavitation to occur. Figures 3.29.1 and 3.29.2 show small entry and exit wounds along the narrow channel of phase 1, with little intervening tissue damage.

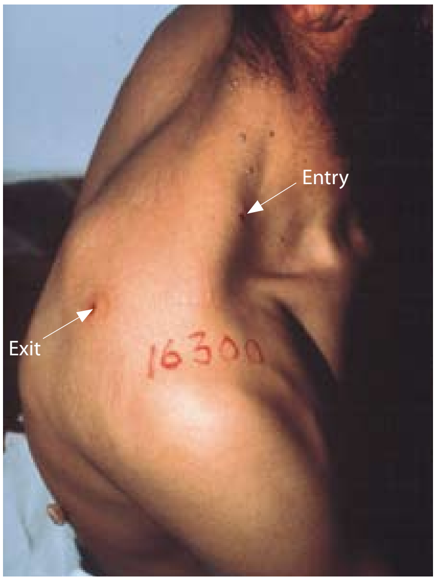

R. Coupland / ICRC

Figure 3.29.1

Small entry and exit wounds, through-and-through injury.

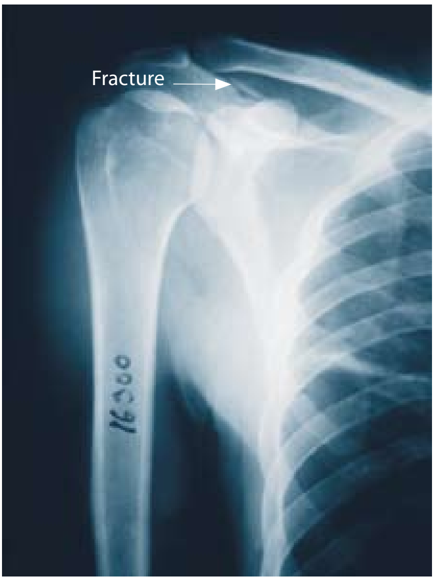

R. Coupland / ICRC

Figure 3.29.2

The radiograph shows a small drill-hole fracture of the acromion: phase 1 narrow-channel injury only.

The wound is large when the exit takes place *during* cavitation (Figures 3.30.1 – 3.30.3).

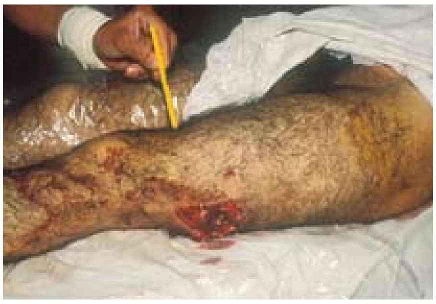

R. Coupland / ICRC

Figure 3.30.1
Gunshot wound to the thigh; the small entry is medial and the large exit on the lateral aspect.

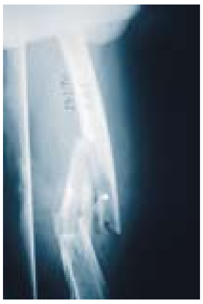

R. Coupland / ICRC

Figure 3.30.2
There is a severe comminuted fracture of the femur and fragmentation of the bullet.

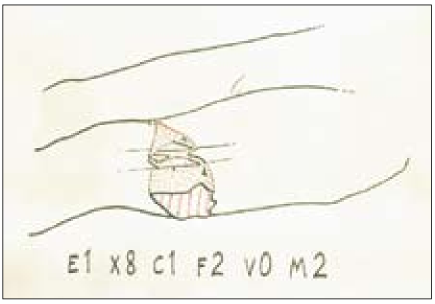

R. Coupland / ICRC

Figure 3.30.3
The exit occurred during cavitation. The fragmentation of the bullet is a tell-tale sign of severe tissue damage. Wound Score: Grade 3, Type F (see Chapter 4).

When the small exit wound occurs *after* cavitation, the *intervening* damage can be severe (Figures 3.31.1 and 3.31.2).

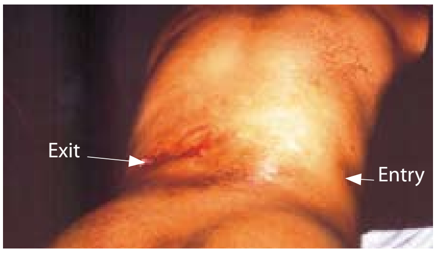

R. Coupland / ICRC

Figure 3.31.1

Small entry and exit wounds with severe damage in between.

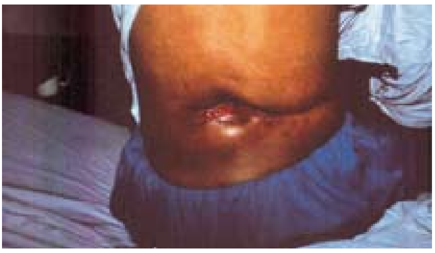

R. Coupland / ICRC

Figure 3.31.2

The same wound after excision and partial healing.

With a handgun bullet, the cavitation effects are minimal and the final wound channel is almost entirely due to crush whatever the length of the shooting channel (Figure 3.32.1 and 3.32.2).

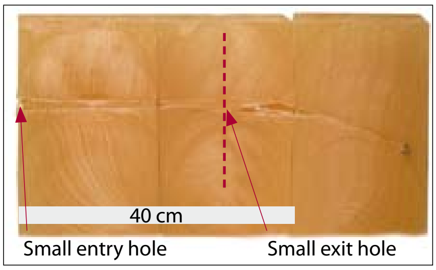

B. Kneubuehl

Figure 3.32.1

Small entry and exit wounds from an FMJ handgun bullet: demonstration in a soap block.

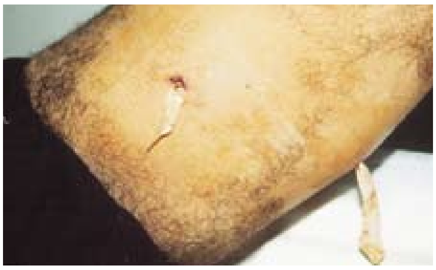

B. Kneubuehl

Figure 3.32.2

Handgun bullet wound to the thigh.

The case of bone

The direct impact of a bullet on bone is different according to where this occurs in the shooting channel. Three clinical situations arise, corresponding to the phases of the shooting channel. In the first case, the stable bullet of phase 1 causes a small "drill-hole" fracture. A small cavitation occurs, and the bone then collapses on itself. The final hole is smaller than the calibre of the bullet (Figure 3.29.2).

In the second case, the bone is broken into multiple fragments owing to a higher release of kinetic energy from the larger cross section of the tumbling bullet during phase 2 (Figure 3.33). Each bone fragment creates its own local crush damage; the muscle is cut up in between the fragments. Afterwards, the temporary cavitation acts on this shredded muscle with lowered tensile strength; the cavity is greater and the final remaining wound channel is larger. The bone fragments always remain within the temporary cavity; they do not produce a new second wound outside the cavity. Detached pieces of muscle and the bone fragments are found inside the final wound, which is usually very severe. This is of obvious clinical relevance to the management of these wounds.

The damage that occurs to bone during the phase 3 shooting channel depends on how much kinetic energy remains in the bullet. If it is a very small amount, the bone will stop the bullet without fracturing.

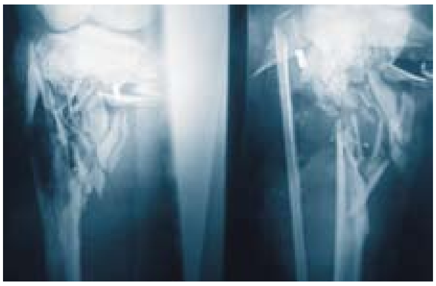

ICRC

Figure 3.33

Severely comminuted fracture of the tibia.

Ricochet

As seen in the tissue simulants, a ricochet FMJ bullet creates a wound that resembles that of a soft-nosed, dum-dum bullet: early dissipation of energy through crush and stretch (Figure 3.34). This phenomenon can be significant for a soldier wearing body armour. If the armour is pierced by a bullet, the wound may be more severe than if he had not been wearing the protective device.

Please note:

People are not injured in the anatomic position. The situation of a muscle group may alter with the patient's position so that a track is obscured, lengthened or shortened.

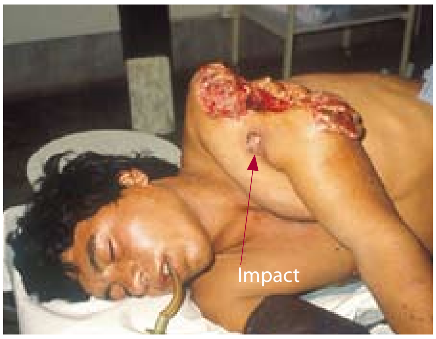

R. Coupland / ICRC

Figure 3.34

The wound of a ricochet FMJ bullet resembles that of a dum-dum bullet. The head of the humerus has literally exploded.

Disruption of the bullet and secondary fragments

At impact velocities above 700 m/s, standard FMJ bullets tend to fragment in tissues at short shooting distances (< 30 – 100 m). Some of these fragmented pieces make their own track of crushed tissue. Tissue disruption from bullet fragmentation is then synergistic: the tissues are multiply perforated by the metallic fragments, thus weakening the tissue's cohesiveness, before being subjected to cavitation stretch. The injury is severe. The same phenomenon occurs with shotgun pellets; each pellet creates its own crush track.

Clinically, an X-ray showing a "shower of lead" from a fragmented bullet should alert the surgeon to the likelihood of a large transfer of kinetic energy to the tissues and consequent severe damage (Figures 3.35, 4.5 and 10.5).

Figure 3.35

Bullet fragmentation: a "shower of lead".

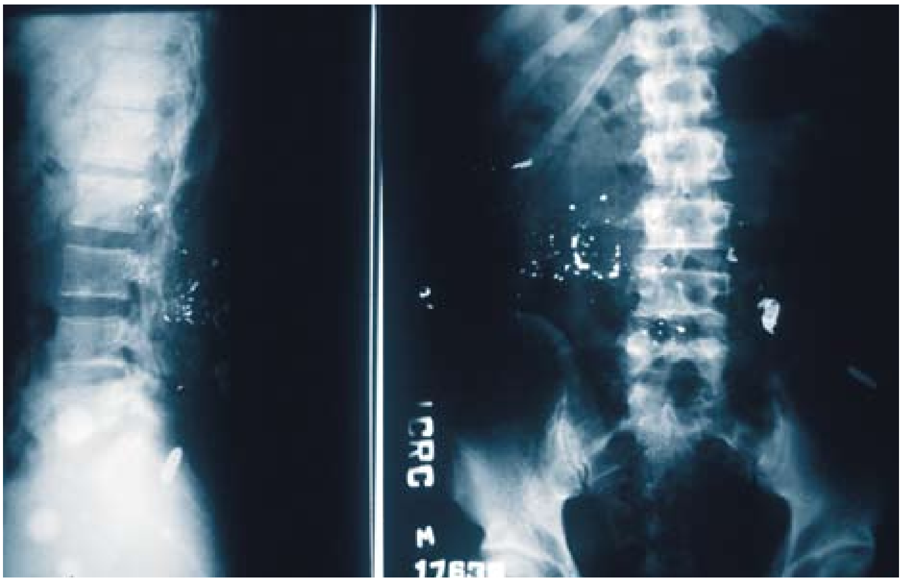

ICRC

Other secondary missiles are objects to which the bullet imparts sufficient motion through the transfer of kinetic energy. These may include foreign bodies such as a belt buckle, a small stone, the metallic contents of a pocket, or the contents of flak jackets – or autologous bodies including the teeth, dental fillings or dentures, and bone fragments (once an ICRC surgeon found a piece of a shattered mandible embedded in a patient's neck).

3.4.6 Sonic shock wave in tissues

A projectile in flight is accompanied by a series of waves propagated at the speed of sound (330 m/s in air). When the projectile strikes a person, this sonic wave is propagated throughout the body at the speed of sound in tissue (4 times the speed of sound in air).

This sonic shock wave may have high amplitude but is of very short duration, not enough to move or injure tissue. However, researchers have noted microscopic cell changes as well as stimulation of peripheral nerves, if the pressure generated reaches a certain threshold. The nerve stimulation occurs immediately, while the cell damage is apparent only after 6 hours. Occasional, short-lived neurapraxia appears to be the only clinically relevant damage.

3.4.7 Pressure wave in tissues and blood vessels

This pressure wave is part of the phenomenon of the temporary cavitation and should *not be confused* with the sonic shock wave. The outer boundary of the cavity consists of a "front" of compressed tissues, which creates a pressure wave at the limit of tissue compression. The wave develops fully only after the cavitation and decreases with distance. It is measured in milliseconds (1,000 times longer than the shock wave). The pressure wave can thus disrupt capillaries or provoke their thrombosis, rupture a full bowel or the liver, and cause retinal detachment and fractures, at a distance from the cavity.

Additionally, blood vessels contained in tissues are squeezed and emptied suddenly with tissue compression. This creates a pressure wave in the column of blood that is propagated away from the site of cavitation. The clinical result of this pressure wave along the course of a blood vessel can be thrombosis or dissection of the intima or muscularis.

3.4.8 Fragment wounds

Their non-aerodynamic shape means that fragments rapidly lose speed in air. If initial velocities are reported to be up to 2,000 m/s, the impact velocity in survivors is usually much lower. If the person is very close to the exploding device, the penetration is deep. If very far away, the casualty is only "peppered" superficially with multiple fragments.

Fragments do not tumble in the tissues, as was demonstrated in the tissue simulants. Thus, most tissue damage is due to crush. At the end of the trajectory, the sharp edges of an irregular fragment will cut tissues. By contrast, a bullet tends to push aside tissues at the end of its track. The wound profile resembles a *cone of tissue destruction*, with the largest diameter at the entry since most energy is dissipated at the surface. The diameter of the entry wound can be anywhere between 2 and 10 times the calibre of the fragment, depending on the impact velocity, mass and shape (Figures 3.36.1 – 3.36.3).

Large, slow fragments penetrate and crush tissues more; small, fast ones result in more stretch. This means that a large and slow fragment tends to cause the same type of wound whatever the tissues, while the wounds of small and fast fragments vary according to tissue elasticity. Nonetheless, in all fragment wounds the extent of tissue damage is always larger than the fragment.

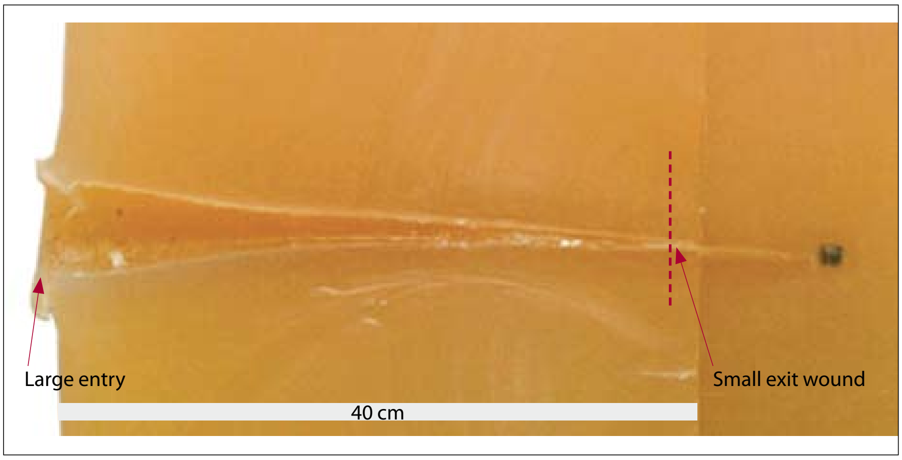

B. Kneubuehl

Figure 3.36.1

Soap block demonstration of fragment injury.

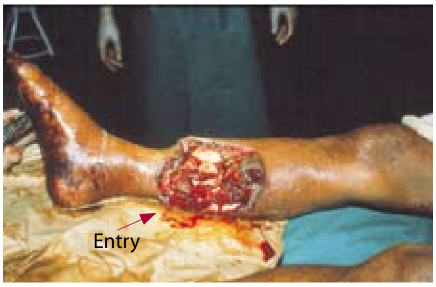

R. Coupland / ICRC

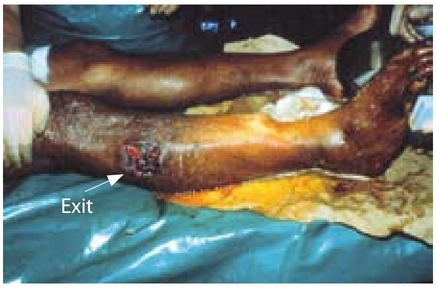

R. Coupland / ICRC

Figures 3.36.2 and 3.36.3

Fragment wound with entry larger than exit.

Wound channel: pathological description

The sharp and irregular surfaces of fragments carry pieces of skin and clothing material into the wounds. As with gunshot wounds, fragment wounds show different zones of tissue injury. The cavity contains necrotic tissues surrounded by a zone of muscle fibre fragmentation with haemorrhage both within and between fibres, further surrounded by a region of acute inflammatory changes and oedema.

The uneven distribution of tissue damage in the concussion and contusion zones has been described as "jumping" or a "mosaic".[10] This is probably because energy has been transferred both along (crush effect) and across (shear and stretch effect) muscle fibres.

---

[10] Wang ZG, Tang CG, Chen XY, Shi TZ. Early pathomorphologic characteristics of the wound track caused by fragments. *J Trauma* 1988; **28 (1Suppl.)**: S89 – S95.

3.5 Wound dynamics and the patient

The wound has a life history, from injury to complete healing. As we have seen, tissue damage depends on a number of physical factors. However, the physical effects of energy transfer do not tell us all the pathological and physiological consequences of the act of wounding.

The tissues in and around the wound undergo reversible and irreversible pathological changes, as well as inflammatory reactions. It can be extremely difficult to diagnose injured tissue that will heal from that which is nonviable and will not heal (see Chapter 10).

Total kinetic energy is the *potential* to cause damage; transferred kinetic energy is the *capacity* to cause damage. Actual tissue damage depends on the *efficiency* of this energy transfer.

The most valuable information gained from this discussion is that small missiles may cause small and minor wounds, or large and serious wounds; and a small entry wound can be associated with large internal damage. Nothing can replace a good clinical examination of the patient and the wound. The extent of the crush and stretch is well represented by the grades of the Red Cross Wound Score (see Chapter 4).

The injurious effects on the patient, however, involve even more than the local pathology. As with all trauma and many diseases, the physiological and psychological state of the victim must be taken into consideration as well. Fit young military personnel, well-trained for their role in warfare and mentally prepared for being wounded, and wounding and killing others, are not the same as civilians. Psychological status cannot be calculated by ballistics. Only these factors can explain the many anecdotes of a person being shot – even several times – and yet continuing to advance or fight in combat.

Chapter 4

RED CROSS WOUND SCORE AND CLASSIFICATION SYSTEM

4. RED CROSS WOUND SCORE AND CLASSIFICATION SYSTEM

4.1 Applications of the RCWS and classification system 83
4.1.1 Assessment of and communication about war wounds in a standardized scheme 83
4.1.2 Establishing a scientific approach to war surgery 83
4.1.3 Surgical and hospital audit 83
4.1.4 Wound information from the field 83

4.2 Principles of the Red Cross Wound Score 84
4.2.1 Examples 86
4.2.2 Notes on wound scoring 87

4.3 Grading and typing of wounds 88
4.3.1 Wound Grade according to amount of tissue damage 88
4.3.2 Typing wounds according to the injured tissues 88

4.4 Wound classification 88

4.5 Clinical examples 89

4.6 Conclusions 91

4.1 Applications of the RCWS and classification system

Surgeons from civilian practice usually have no previous experience of managing war wounds. Even experienced surgeons often do not find it possible to tell the true extent of tissue damage. As noted in Chapter 3, ballistics studies show that there is not a uniform pattern or degree of wounding. Preparation for war surgery involves an understanding of the translation of the kinetic energy of a wounding projectile into tissue damage, i.e. the actual wounds that the surgeon must deal with.

The severity of such wounds depends on the degree of tissue damage and the structure(s) that may have been injured: thus, the clinical significance of a wound depends on its size and site. The Red Cross Wound Score (RCWS) and classification system[1] is based on the features of the wound itself, and not on the weaponry or the presumed velocity or kinetic energy of the missile.

Any wound classification system will aid the surgeon if it helps to assess the severity of the injury, influences surgical management, predicts outcome, and provides for an accurate database that can be used in comparative studies. The RCWS goes a fair way to fulfilling these criteria.

4.1.1 Assessment of and communication about war wounds in a standardized scheme

The Wound Score is a useful clinical tool to communicate the severity of the wound to staff and colleagues without having to remove the dressings. It is also an element to take into consideration for the purposes of triage.

4.1.2 Establishing a scientific approach to war surgery

The RCWS permits comparison of treatments and prognoses of similar wounds. An analogy can be made with the TNM (tumour, node, metastasis) staging of different cancers, allowing comparison of treatment regimes for a similar pathology. In the case of a gunshot wound to the thigh, treatment and prognosis will differ according to the amount of tissue damage, degree of bone comminution and whether there is injury to the femoral vessels. Other wound scoring systems are primarily devised for blunt trauma. The presence of a penetrating wound in these systems often indicates a "serious" wound, with little further nuance.

4.1.3 Surgical and hospital audit

Wound Scores can be used to evaluate quality of care when combined with information on length of hospital stay, number of operations performed per patient or units of blood used, morbidity and mortality. An example relating to the adequacy of primary wound surgery is the number and cause of deaths associated with non-vital wounds, or the number of operations performed per patient for each Wound Grade.

4.1.4 Wound information from the field

ICRC surgical teams treat thousands of war-wounded every year. Analysis of a larger number of scored wounds will eventually clarify the relationship between experimental laboratory wound ballistics and the clinical management of war wounds. Worthy of note is the fact that information gained in the field served as a scientific basis for the campaign to ban anti-personnel landmines, thus promoting new standards in international humanitarian law.

---

1 This chapter is largely based on the brochure *The Red Cross Wound Classification, Revised Edition*, by Dr R. M. Coupland. The ICRC Master Surgeons Workshop held in Geneva in 2002 revised the original Wound Score (see Introduction).

4.2 Principles of the Red Cross Wound Score

Wounds are given a six-figure Score according to several parameters. The Scores can then be translated into a classification system:

- grading of the wound according to the severity of tissue injury and
- typing of the wound according to the tissue structures injured.

The RCWS is intended for quick and easy use in the field, its simplicity especially useful under conditions of stress. It is a simple clinical system that requires no extra equipment or sophisticated procedures. The time required for scoring wounds can be measured in seconds.

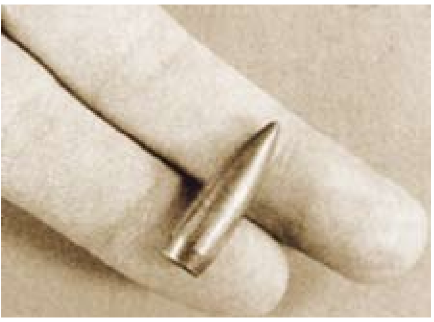
R. Coupland / ICRC

Figure 4.1

The width of two fingers is approximately the length of a military rifle bullet.

E	entry wound in centimetres	
X	exit wound in centimetres (X = 0 if no exit wound)	
C	cavity	Can the cavity of the wound take two fingers before surgical excision?
		C 0 = no
		C 1 = yes
F	fracture	Are any bones fractured?
		F 0 = No fracture
		F 1 = Simple fracture, hole or insignificant comminution
		F 2 = Clinically significant comminution
V	vital structure	Are dura, pleura, peritoneum, or major peripheral vessels injured?
		V 0 = no vital structure injured
		V N = (neurological) penetration of the dura of the brain or spinal cord
		V T = (thorax or trachea) penetration of the pleura or of the larynx/ trachea in the neck
		V A = (abdomen) penetration of the peritoneum
		V H = (haemorrhage) injury to a major peripheral blood vessel down to brachial or popliteal arteries, or carotid artery in the neck
M	metallic body	Are bullets or fragments visible on X-ray?
		M 0 = no
		M 1 = yes, one metallic body
		M 2 = yes, multiple metallic bodies

Table 4.1 Parameters of Wound Score.

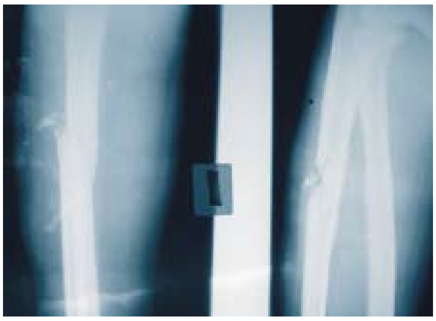
F. Herkert / ICRC

Figure 4.2.1

Gunshot wound: F1 fracture of the ulna.

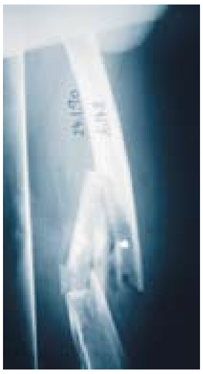
R. Coupland / ICRC

Figure 4.2.2

Gunshot wound: F2 fracture of the femur.

E (entry)	centimetres
X (exit)	centimetres
C (cavity)	C 0, C 1
F (fracture)	F 0, F 1, F 2
V (vital structure)	V 0, VN, VT, VA, VH
M (metallic bodies)	M 0, M 1, M 2

Table 4.2 Total scheme of Wound Score.

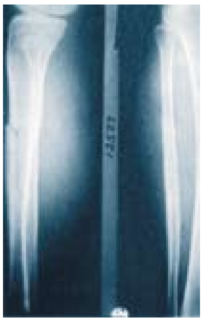
R. Coupland / ICRC

Figure 4.2.3

Gunshot wound: F 1 fracture (clinically insignificant comminution) of the fibula.

Wounds are scored on the patient's admission form after operation or after initial examination if there is no surgery.

ICRC EXPERIENCE

An ICRC admission sheet with Wound Score recorded.

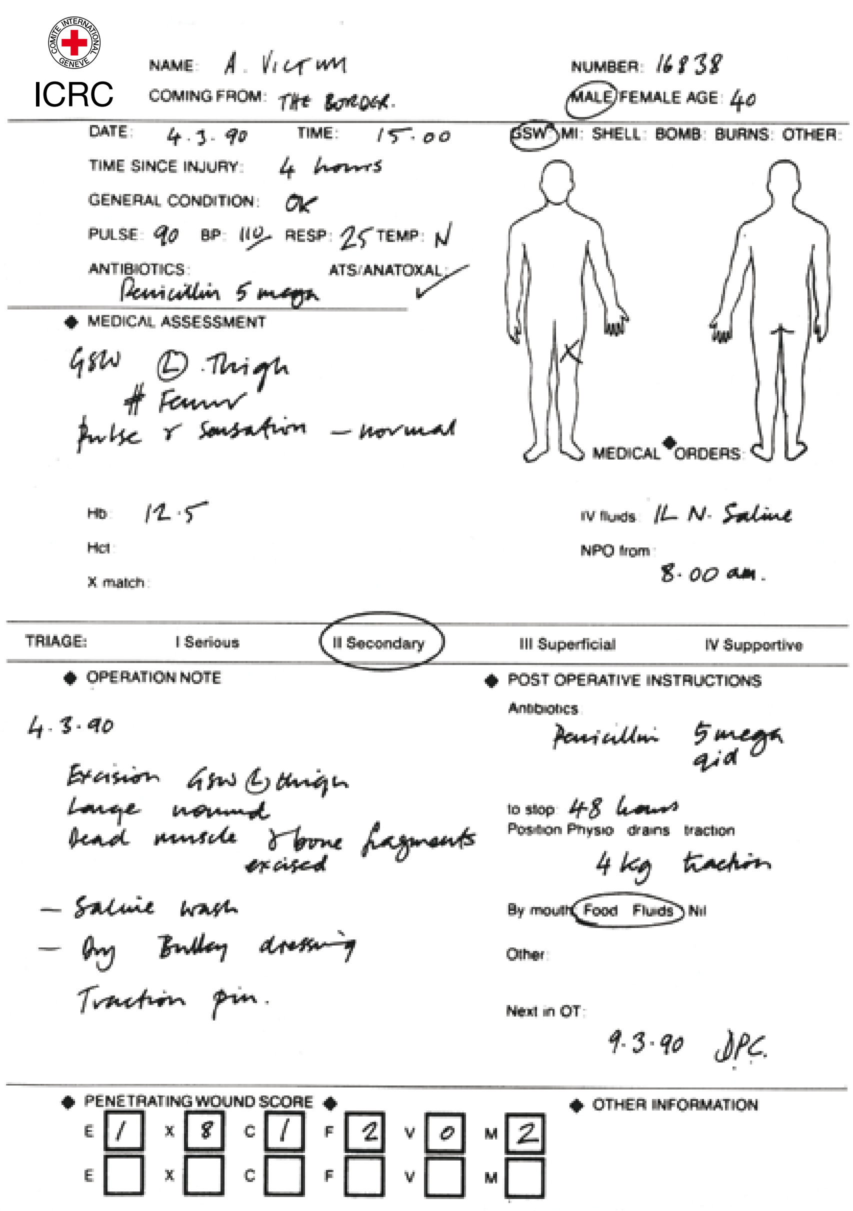

ICRC

NAME: A. Victim
COMING FROM: THE BORDER.
NUMBER: 16838
[✓ MALE] FEMALE AGE: 40

DATE: 4.3.90 TIME: 15.00
TIME SINCE INJURY: 4 hours
GENERAL CONDITION: OK
PULSE: 90 BP: 110 RESP: 25 TEMP: N
ANTIBIOTICS: Penicillin 5 mega
ATS/ANATOXAL: ✓

[✓ GSW] MI: SHELL: BOMB: BURNS: OTHER:

◆ MEDICAL ASSESSMENT

GSW (L) Thigh
Femur
Pulse & Sensation — normal

MEDICAL ◆ ORDERS:

Hb: 12.5
Hct:
X match:

IV fluids: 1L N. Saline
NPO from: 8.00 am.

TRIAGE: I Serious [✓ II Secondary] III Superficial IV Supportive

◆ OPERATION NOTE

4.3.90

Excision GSW (L) thigh
Large wound
Dead muscle & bone fragments excised

— Saline wash
— Dry Bulky dressing
Traction pin.

◆ POST OPERATIVE INSTRUCTIONS

Antibiotics: Penicillin 5 mega qid
to stop: 48 hours
Position Physio drains traction: 4 kg traction
By mouth [✓ Food Fluids] Nil
Other:
Next in OT: 9.3.90 JPC.

◆ PENETRATING WOUND SCORE ◆

E	1	X	8	C	1	F	2	V	0	M	2
E		X		C		F		V		M	

◆ OTHER INFORMATION

4.2.1 Examples

The following two sketches (Figures 4.3 and 4.4) present various wounds due to weapons, and their assessment according to the ICRC's wound scoring system.

Figure 4.3

Example of Wound Scores.

a. Simple bullet track.
b. Track produced by bullet with temporary cavitation at exit.
c. Simple track with involvement of a vital structure (artery).
d. Low-energy transfer wound with simple fracture.
e. High-energy transfer wound by fragment with comminuted fracture.

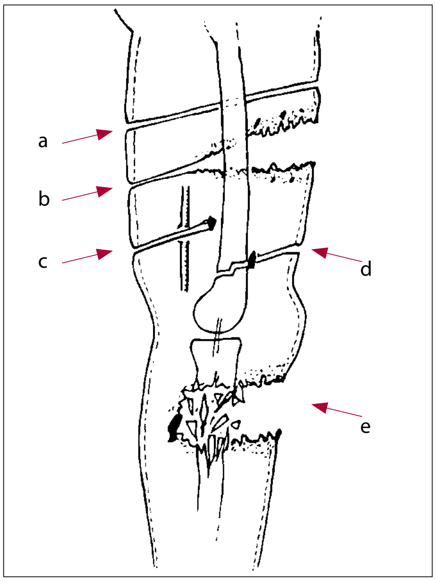

ICRC

	E	X	C	F	V	M
Wound (a)	1 ?	2	0	0	0	0
Wound (b)	1	4	1	0	0	0
Wound (c)	1	0	0	0	H	1
Wound (d)	1	0	0	1	0	1
Wound (e)	6	0	1	2	0	1

Figure 4.4

Further example of Wound Scores.

a. Through-and-through bullet wound of soft tissue with central cavitation.
b. Through-and-through bullet wound with central cavitation and comminuted fracture.
c. High-energy transfer bullet wound with cavitation and comminuted fracture.

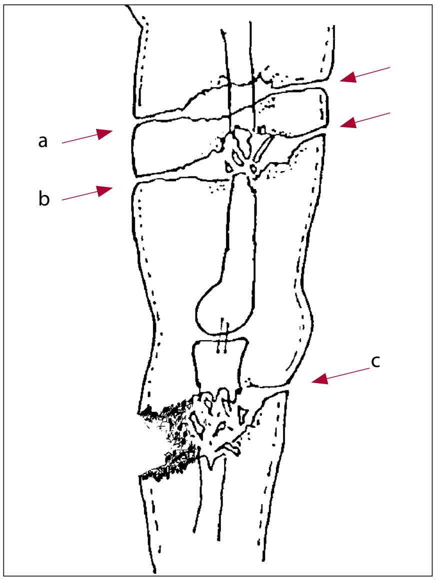

ICRC

	E	X	C	F	V	M
Wound (a)	1 ?	1	1	0	0	0
Wound (b)	1 ?	1	1	2	0	0
Wound (c)	1	6	1	2	0	2

4.2.2 Notes on wound scoring

1. When it is impossible to differentiate between the entry and the exit wounds, put a (?) between the E and X Scores.
2. With multiple wounds, only the two most serious are scored.
3. If a wound cannot be scored (unclassifiable), U/C is written on the Score. This applies to a minority of wounds.
4. When one missile causes two separate wounds (e.g. through the arm and into the chest), the 2 separate Scores are joined by a bracket.
5. Include penetrating wounds only, not superficial tangential skin wounds; e.g. skin wound 20 cm long and 1 cm wide but without penetration deep to the fascia.
6. Estimation of the cavity: the width of two fingers is approximately the length of a bullet from a military assault rifle (Figure 4.1). If the permanent cavity of a wound admits two fingers before surgical excision, then something other than crush by a bullet travelling sidelong must have taken place, i.e. stretch and laceration by cavitation. A C1 wound is likely to present significant tissue damage whatever the cause.
7. Fracture severity: it is inevitable that some wounds fall between F1 and F2, but for simplicity this is not accurately defined here. A further refinement is given in Volume 2. An example of a clinically insignificant comminution (F1) is a wound with a comminuted fibula but with an intact tibia (Figure 4.2.3).
8. A vital injury implies a more dangerous wound involving a surgical task in addition to simple wound management (e.g. craniotomy, pleural drainage or thoracotomy, laparotomy). VH includes the popliteal and brachial vessels but not those more distal. The outcome of injuries to the head, thorax or abdomen, or causing massive peripheral haemorrhage, is only partly determined by the clinical size of the wounds as determined by the RCWS (see below).
9. Metallic fragments: note the difference between an intact bullet (M1) and a fragmented bullet (M2), as shown in Figure 4.5. If the full metal jacket of a bullet has been disrupted and the lead interior has leaked out as fragments, this indicates a severe stress on the bullet and a large transfer of kinetic energy to the tissues (Figures 3.35 and 10.5).

   Note also the difference between multiple metallic fragments from a grenade or shell (Figure 4.6) – which are common and do not necessarily represent a large transfer of kinetic energy – and a disrupted and fragmented bullet, which does. If there are numerous metallic fragments, do not count them; only mark M = 2.
10. If X-rays are not available the RCWS is still valid. The F Score is judged clinically and the M parameter omitted, or included if fragments are found during surgical excision. X-rays should not be performed just to fill in the M and F Scores.
11. Traumatic amputation of limb (Figure 4.7): this injury is often seen with anti-personnel blast mines. The open end of the amputated limb equals a combined entry-exit, whose diameter represents the sum of E+X. It is equivalent to a cavity (C = 1), while the vaporized missing part of the limb qualifies as a severe fracture (F = 2). The level of traumatic amputation (above or below the knee or elbow) will determine if the Vital structure Score is 0 or H.

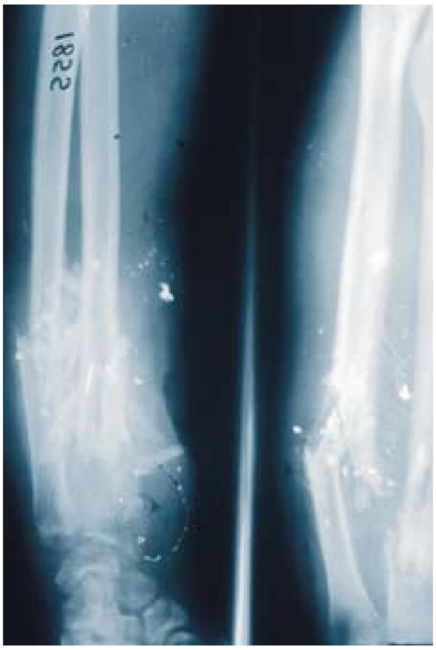

ICRC

Figure 4.5

Disrupted and fragmented bullet: M2.

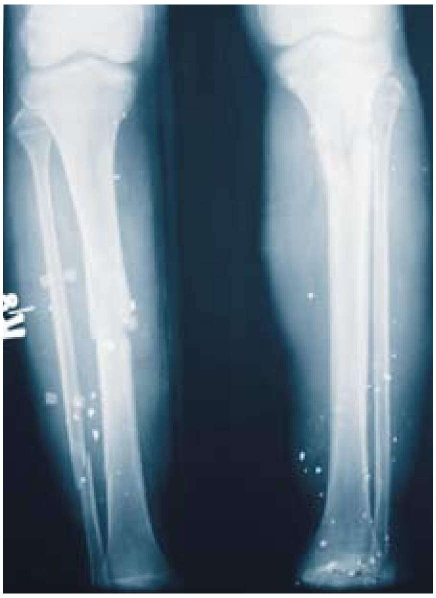

ICRC

Figure 4.6

Multiple shell fragments: M2.

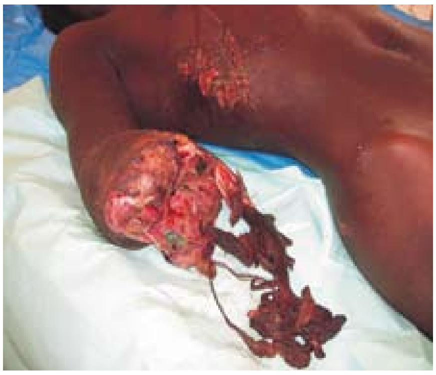

K. Barrand / ICRC

Figure 4.7

Traumatic amputation of the forearm.

E	X	C	F	V	M
20	?	1	2	0	0

4.3 Grading and typing of wounds

Once scored, the wound can be *graded* according to severity (E, X, C and F), and *typed* according to tissue type (F and V).

4.3.1 Wound Grade according to amount of tissue damage

The wound should be *graded* according to severity.

Grade 1
E + X is less than 10 cm with Scores C 0 and F 0 or F 1.
(Low energy transfer.)

Grade 2
E + X is less than 10 cm with Scores C 1 or F 2.
(High energy transfer.)

Grade 3

E + X is 10 cm or more, with C 1 or F 2.
(Massive energy transfer.)

These Grades represent the outcome of a simple clinical assessment that corresponds to the effective transfer of kinetic energy of projectiles to body tissues. Large wounds are more serious and require greater resources; this is particularly true of wounds to the limbs.

4.3.2 Typing wounds according to the injured tissues

Once *scored*, the wound can be *typed* according to tissue structures.

Type ST
Soft-tissue wounds: F 0 and V 0.

Type F
Wounds with fractures: F1 or F 2, and V 0.

Type V
Vital wounds putting the patient's life at risk: F 0 and V = N, T, A or H.

Type VF
Wounds with fractures and involving vital structures putting life or limb at risk: F 1 or F 2 and V = N, T, A or H.

4.4 Wound classification

Combining Grades and Types gives rise to a classification system divided into 12 categories.

	Grade 1	Grade 2	Grade 3
Type ST	1 ST Small, simple wound	2 ST Medium soft-tissue wound	3 ST Large soft-tissue wound
Type F	1 F Simple fracture	2 F Important fracture	3 F Massive comminution
Type V	1 V Small wound threatening life	2 V Medium wound threatening life	3 V Large wound threatening life
Type VF	1 VF Small wound threatening limb and/or life	2 VF Important wound threatening life and/or limb	3 VF Large wound threatening life and/or limb

Table 4.3 Grade and tissue Type categories.

These categories help determine surgical management for a number of wounds, though by no means all (see Chapters 10 and 12).

Further refinement of the fracture type is possible, especially for assessing the extent of bone defect. This can be useful in a specialized study of war wounds with fractures (see Volume 2).

4.5 Clinical examples

Epidemiological studies from the ICRC surgical database of over 32,000 patients have shown that the Red Cross Wound Score gives excellent prognostic results for surgical workload and morbidity due to war wounds. As demonstrated in the statistical analysis in Chapter 5, the Wound Grade is highly relevant to the number of operations per patient. This is particularly the case for wounds to the extremities. Vital Wound Scores to the extremities (V = H) also give good measures for mortality and the amputation rate.

One of the avowed weaknesses of the Classification concerns the prognostic mortality of vital central wounds. It is in the anatomic nature of the brain, heart and great vessels that even a very small wound from a projectile with low kinetic energy can be lethal if a vital centre is injured. The RCWS does classify such wounds as "potentially" lethal; that is why, by definition, they are categorized as "vital" injuries. However, a Grade 1 wound can be just as lethal as a Grade 3 wound depending on what particular structure is hit: medulla oblongata or temporal cortex; lung parenchyma or ascending aorta.

What is certain is that for most Grade 3 wounds to the head, thorax and abdomen, pre-hospital mortality is bound to be high. Without autopsies of those killed in action (KIA), analysis can only be made of the relatively few "survivors" who reach hospital; for them, wound Grade is not significant with respect to mortality (Tables 4.4 – 4.6).

Thus, the outcome of injuries to the head, thorax and abdomen are not only, nor even mainly, determined by the clinical size of the wound as defined by the RCWS. The various factors affecting the mortality from vital wounds; e.g. for the abdomen, number of organs injured, degree of faecal contamination, intraoperative blood loss, delay in operative treatment, etc., will be discussed in Volume 2.

	Number	Deaths	% mortality
Grade 1	75	14	18.7
Grade 2	70	15	21.4
Grade 3	9	3	33.3

Table 4.4 Mortality from head and neck injuries (VN) : N = 154. (ICRC Kabul, 1990-92)

	Number	Deaths	% mortality
Grade 1	82	4	4.9
Grade 2	41	2	4.9
Grade 3	3	0	0

Table 4.5 Mortality from chest injuries (VT) : N = 126. (ICRC Kabul, 1990-92)

	Number	Deaths	% mortality
Grade 1	120	9	7.5
Grade 2	70	11	15.7
Grade 3	5	0	0

Table 4.6 Mortality from abdominal injuries (VA) : N = 195. (ICRC Kabul, 1990-92)

Nonetheless, as explained in Chapter 5, a distinction should be made in patient databases between superficial and penetrating wounds to the head, thorax and abdomen to allow for a better analysis of results of treatment. Using the RCWS enables this.

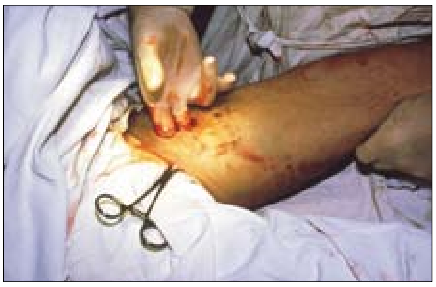

R. Coupland / ICRC

Figure 4.8.1

Two fragment wounds to the thigh. The surgeon is assessing the larger wound cavity.

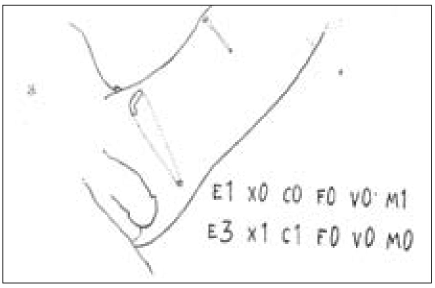

R. Coupland / ICRC

Figure 4.8.2

Both wounds are Type ST. The smaller wound is Grade 1, and the larger Grade 2.

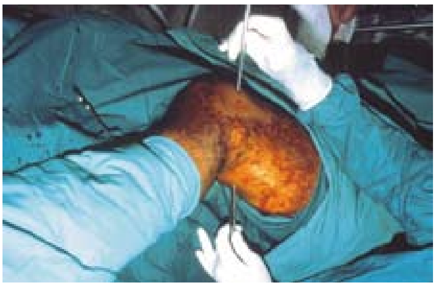

R. Coupland / ICRC

Figure 4.9.1

Through-and-through gunshot wound to the knee.

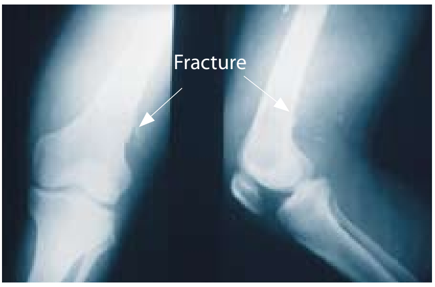

R. Coupland / ICRC

Figure 4.9.2

There is a small fracture above the lateral condyle.

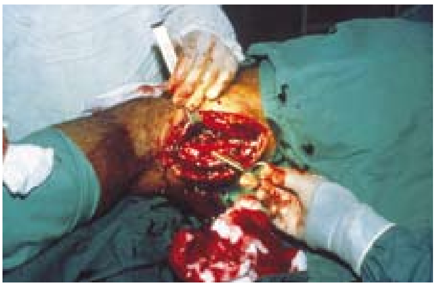

R. Coupland / ICRC

Figure 4.9.3

The popliteal artery has been injured.

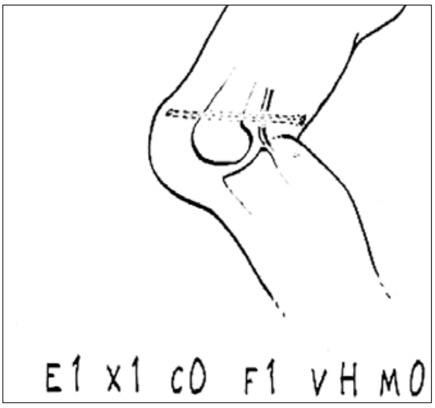

R. Coupland / ICRC

Figure 4.9.4

The wound is Type V(H)F, Grade 1.

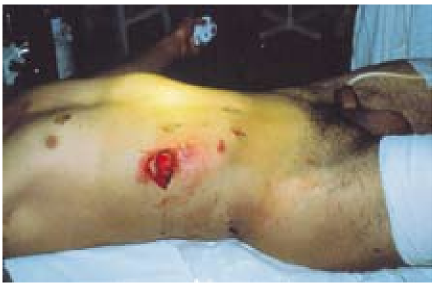

R. Coupland / ICRC

Figure 4.10.1

Gunshot wound to the abdomen.

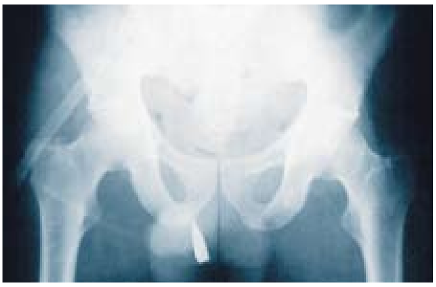

R. Coupland / ICRC

Figure 4.10.2

The bullet is lodged in the scrotum. There is a small fracture of the right pubic bone.

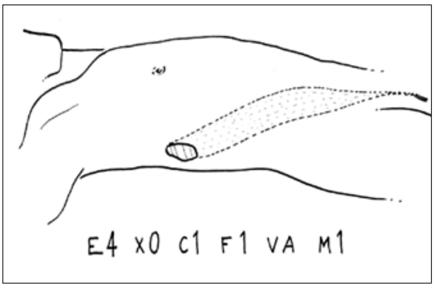

R. Coupland / ICRC

Figure 4.10.3

This is a Grade 2 wound, Type V(A)F.

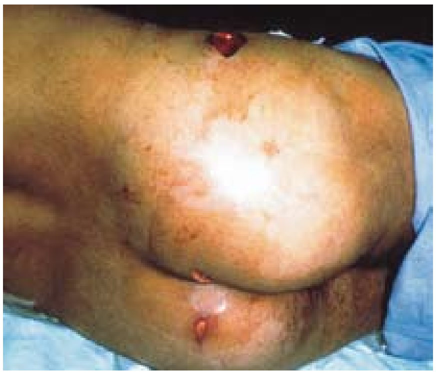

R. Coupland / ICRC

Figure 4.11.1

A bullet has grazed the left buttock and then entered the right buttock at the anal cleft. The exit is located laterally.

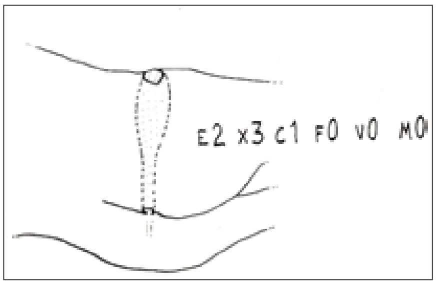

R. Coupland / ICRC

Figure 4.11.2

This is a Grade 2, Type ST wound.

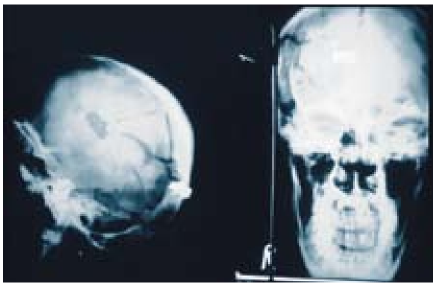

R. Coupland / ICRC

Figure 4.12.1

Fragment wound to the head with entry right parietal. Note the severe linear fractures.

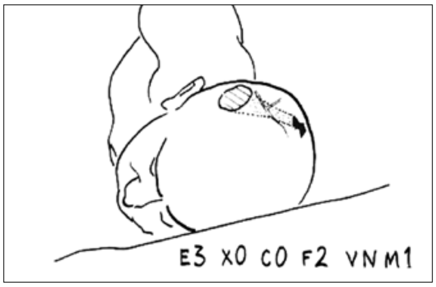

R. Coupland / ICRC

Figure 4.12.2

The cone of tissue destruction extends to the occipital lobe. Grade 2, Type V(N)F.

4.6 Conclusions

The Red Cross Wound Score enables medical staff to translate war wounds into surgical lesions rather than weaponry phenomena. It refines the heterogeneity of wounds and helps define them according to their clinical significance, and it is a good indicator of tissue damage due to the transfer of kinetic energy in penetrating wounds.

The limitations of the scoring system are recognized; complete accuracy cannot be obtained. It should be emphasized that the RCWS is for rapid use under adverse conditions, including extreme stress, and uses no additional equipment or sophisticated technology. The simplicity and advantages of the RCWS outweigh any disadvantages that may occasionally be introduced by observer error. As with any classification system, the more experience gained in using the RCWS, the more accurate the results and the fewer the observer errors.

Chapter 5

THE EPIDEMIOLOGY OF THE VICTIMS OF WAR

5. THE EPIDEMIOLOGY OF THE VICTIMS OF WAR

5.1 Introduction: purpose and objectives 95
5.1.1 The reader and this manual 95

5.2 Public health effects of armed conflict 95
5.2.1 Historical perspective 95
5.2.2 The public heath effects of social disruption 96
5.2.3 The burden of war-wounded civilians 97
5.2.4 Methodology 98

5.3 Epidemiology for the war surgeon 98
5.3.1 Preparation of the surgeon 98
5.3.2 Preparation of the institution 99
5.3.3 Surgical audit: monitoring 99
5.3.4 The surgeon and IHL 100
5.3.5 The surgeon and the specialized literature 100

5.4 General questions of methodology 101
5.4.1 Large and small numbers 101
5.4.2 Problems of data collection 101
5.4.3 Important first questions: who is counting whom? 102
5.4.4 Some definitions: the wounded and the dead 103
5.4.5 Clinical and operational consequences and IHL 104

5.5 Aetiology of injury 104
5.5.1 Warfare scenarios 104
5.5.2 Definitions of weapon-systems: methodology 105
5.5.3 ICRC statistics 106

5.6 Anatomic distribution of wounds 106
5.6.1 Body exposure 106
5.6.2 Definitions and methodology 107
5.6.3 Historical results 108
5.6.4 ICRC results 108
5.6.5 Primary tissue injury 109

5.7 Fatal injuries 110
5.7.1 Site 110
5.7.2 Trimodal distribution of trauma mortality 111
5.7.3 Ratio of dead to survivors 113
5.7.4 Lethality of weapons 113
5.7.5 Conclusions of clinical relevance 114

5.8 The lethality of context: delay to treatment 114
5.8.1 Historical developments 114
5.8.2 Forward projection of resources 115
5.8.3 Urban warfare: hospitals on front lines 115
5.8.4 The paradox of early treatment: changing mortality ratios and rates 115

5.9 Hospital mortality 117
5.9.1 Historical considerations 117
5.9.2 Hospital mortality *versus* post-operative mortality 117
5.9.3 ICRC hospital mortality 117

5.10 ICRC statistical analysis of hospital workload 118
5.10.1 Surgical audit: methodology 118
5.10.2 Number of operations per patient: all patients 118
5.10.3 Number of operations according to delay in evacuation 119
5.10.4 Number of operations according to Wound Grade 119
5.10.5 Number of operations according to Wound Grade and delay in evacuation 120
5.10.6 Number of operations according to weapon 120

5.11 Conclusions: lessons to be gained from a study of epidemiology 121
5.11.1 Setting up a surgical database for the war-wounded 122

ANNEX 5. A ICRC surgical database 123

ANNEX 5. B Setting up a surgical database for the war-wounded 125

5.1 Introduction: purpose and objectives

Epidemiological studies are standard practice in medicine. Doctors need to understand the characteristics of their patient population and the relationships between particular circumstances and the outcome of treatment. These elements help to establish the high risk factors.

As regards the victims of war, the total effects of conflict must be analysed, the impact on civilian populations as well as the military, the direct effects of war trauma and the indirect public health consequences. These help the ICRC and other organizations determine their response in terms of assistance (see Chapter 2).

5.1.1 The reader and this manual

Most readers of this manual will be civilian surgeons who are unacquainted with the practice of war surgery and its literature. Even among military surgeons many may have only a superficial acquaintance with these topics. This chapter on the epidemiology of victims of war presents a certain approach to the understanding of the many factors involved in patient care and outcome.

First the public health repercussions of armed conflict will be dealt with briefly, then the purely surgical aspects will be addressed in more detail. Questions of general methodology and definitions will begin the surgical study, and will be followed by a presentation of historical and ICRC experience and results. Relevant clinical conclusions will be emphasized throughout, as will recurring problems of methodology.

For the purpose of this discussion the authors refer directly only to the ICRC's experience in armed conflict and the importance of epidemiological studies in determining ICRC surgical activities. In addition, reference is made to the statistics in standard and well-accepted scholarly books and journal articles for historical results that have come to influence clinical practice (see Selected bibliography).

5.2 Public health effects of armed conflict

5.2.1 Historical perspective

It is commonly said that contemporary warfare causes more civilian victims than military, and that the public health effects are greater than those caused by trauma. In all major conflicts over the last one hundred years, with few exceptions (World War I), civilians have become a direct target and accounted for one-half to two-thirds of the casualties.

> "Throughout history there have been many small local wars, often involving wholesale starvation and occasionally the elimination of whole populations or tribes."
>
> G.W. Odling-Smee[1]

The heavy civilian toll in contemporary conflicts is therefore not new. Nevertheless, in certain recent conflicts – whether revolutionary wars of national liberation, civil wars, territorial and border disputes, or peasant uprisings – the destabilization of political, social, and economic infrastructures, the destruction of cultural symbols and institutions, and the psychological terrorizing of civilians have become explicit political and military goals. A number of reports have indicated that the indirect effects of social disruption have caused two to fifteen times more civilian deaths than war wounds have, and, in some cases, civilians may account for up to 90 % of the victims.

---

[1] Odling-Smee GW. Ibo civilian casualties in the Nigerian civil war. *BMJ* 1970; **2**: 592 – 596.

> "Today's armed conflicts are essentially wars on public health."
>
> Rémi Russbach, former Chief Medical Officer at the ICRC

5.2.2 The public heath effects of social disruption

Social disruption includes population displacement – either internal or refugee – pauperization, the breakdown of nutrition and sanitary conditions, the lack of safe drinking water, and the disorganization of medical systems (Table 5.1).

Direct effects	Indirect effects
Death	Economic pressures and disruption
Disability	Reduced food production and distribution Family disruption (orphans, abandoned children)
Destruction of health services	Refugees Psychological stress
Disruption of health programmes	Effects on housing, water supply, and sewage disposal
Psychological stress	Economic pressures on those caring for the war-disabled
Illness	Environmental (landmines, deforestation)

Table 5.1 Effects of political violence on health and health systems.[2]

Pathologies and the pathology of the health system

Complex emergencies during armed conflict in poor countries tend to cause death by endemic communicable disease and malnutrition. In more developed countries, increased mortality rates due to trauma and chronic disease dominate. The lack of access to care, either because of poverty, the disorganization of the medical system, or for reasons of security, aggravates the problems.

To pre-conflict poverty and fragile health services are added the problems of the collapse of medical supply and distribution, the targeting of hospitals and clinics in violation of medical neutrality, and the neglect or abandonment of preventive health programmes. In addition, health professionals are often amongst the first people to flee a conflict zone. Having to provide for war trauma also means a greater investment in curative services to the disadvantage of the preventive. The entire health system becomes organized around the conflict and there is an imbalance between the needs of victims and the available public health services (Figure 5.1).

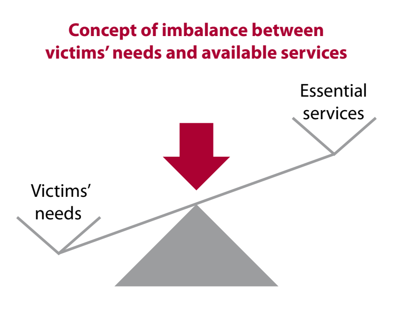
Concept of imbalance between victims' needs and available services
Essential services
Victims' needs

Increase in needs associated with:
- increase in population
- increase in morbidity due to sanitary conditions

Appearance of new needs, e.g. treatment of war wounds

Deterioration in effectiveness of services:
- insufficient means
- abandonment of facilities by personnel
- disrupted organization

Destruction of medical facilities

P. Perrin / ICRC

Figure 5.1

Victims' needs and available services during armed conflict.[3]

Moreover, all too often, increased military spending in a country at war is to the detriment of social and health investments. Another recent development is the creation of a war economy controlled by criminalized military networks. The extraction of diamonds and other precious stones, oil, timber, and various minerals, the trafficking of narcotics, and sheer banditry, take a further and severe social toll.

A public health approach shows the full range of the humanitarian consequences of the use of anti-personnel landmines, especially during the post-conflict period. In addition to death and disability, there are the long-term socio-economic effects of disrupted economic activity through the loss of farmland, livestock, water resources, and industrial capacity because of the presence of landmines.

---

[2] Zwi A, Ugalde A. Towards an epidemiology of political violence in the third world. *Soc Sci Med* 1989: **28**: 633 – 642. Cited in Lautze S, et al., 2004.

[3] From Perrin P. *War and Public Health: A Handbook*. Geneva: ICRC; 1996.

Landmines and other UXO (unexploded ordnance) are not the only weapons left over after war. Demobilized combatants with guns are not often successfully reintegrated into social and economic life: criminal violence replaces political violence; peace combined with high criminality offers little respite from war. And the social and economic costs continue to cast a shadow over everyday life.

In addition, one must count the social costs of deliberate physical, psychological, and sexual abuse as methods of warfare. Torture and rape have profound and long-term consequences.

Public health effects on the military

Even among classical armies of industrialized countries, disease caused more deaths than combat until World War II. Mortality does not tell the whole story, however. Serious disease continues to affect modern armies, still accounting for high rates of non-availability of personnel for military duty. Of course, certain natural environments (jungle, high mountains) increase the incidence of non-battle pathology.

5.2.3 The burden of war-wounded civilians

Although trauma may account for only a small number of all war-related deaths in a given civilian population, the relative proportions of mortality and morbidity due to disease and trauma vary over time and between different war zones; refugee camps are different again.

In some conflicts, the burden of war trauma is greater than the public health consequences. This was the case in the wars in the former Yugoslavia (1991 – 99) in Lebanon (1975 – 90), and Rwanda (1994), among others. One must also take into consideration the relatively small population of certain countries; a limited number of deaths may represent a proportionately high death toll.

In some conflicts, the burden of the war-wounded is greater than the consequences on public health.

The type of combat may place civilians at greater risk and war casualties may overwhelm civilian medical services, even in situations where the public health effects are greater than the direct effects of trauma (Biafra 1967 – 70, Uganda 1987, Democratic Republic of the Congo since 1997 to the time of writing).

Mortality data, however, do not reflect the extent and severity of war injuries. Morbidity and disability rates due to injuries are testimony to the long-term socio-economic burdens.

Demographics of the victims

Historical military casualty statistics concern young, fit and healthy men. In the past, women were not close to combat in most classical armies; more recently, however, various revolutionary guerrilla groups have had female combatants.

When a civilian population finds itself in the midst of war, the demographic profile of casualties is closer to its population pyramid. This has clinical consequences for the treatment of endemic diseases and chronic pathologies among the wounded.

J-P. Di Silvestro / ICRC

Figure 5.2

The number of wounded and dead by direct trauma may at times far exceed the effects on public health.

5.2.4 Methodology

Public health monitoring and data collection are notoriously difficult during the chaotic and overwhelming conditions of complex emergencies. Missing and displaced persons, constraints of time, lack of access to populations, and dangerous security situations, all create formidable problems for a limited number of qualified personnel to conduct proper studies. Furthermore, the distinction between civilian and military during a civil war is not always obvious.

The ratio of military to civilian wounded and sick may be very sensitive political and military information, easily exploited for propaganda purposes by the various antagonists. In an attempt to prevent this, should ICRC delegates determine that civilians are being expressly targeted by combatants, the organization undertakes specific *démarches* regarding the conduct of hostilities with the authorities concerned through its traditional confidential procedures.

Many authors have recently written about the public health effects of armed conflict and complex emergencies. The ICRC first published its standard *War and Public Health* in 1996 and, together with the World Health Organization and a number of universities around the world, it organizes a dozen H.E.L.P.[4] courses on the management of humanitarian aid every year.

5.3 Epidemiology for the war surgeon

What should one look for in a study of epidemiology? What information will help the individual surgeon managing war wounds for the first time? What needs to be taken into consideration to set up an efficient system for the care of the war-wounded? Is it possible to determine the "normal" levels of morbidity and mortality during armed conflict?

To respond adequately to the challenge requires preparation, on the part of the individual surgeon as well as the institution's, and a system for monitoring the results of patient management.

5.3.1 Preparation of the surgeon

The surgeon facing war-wounded casualties for the first time will ask a number of questions: "What wounds should I expect? Which injuries kill? What will the surgical workload be like?"

In many conflicts, civilian surgeons, albeit without previous military experience, become heavily involved in the management of the war-wounded. Yet, epidemiological studies clearly show that combat injuries differ from those in civilian practice: different aetiology and pathology; multiple-cause injuries; delayed access to care and austere working conditions requiring different philosophies of treatment.

Epidemiological studies show that the management of the war-wounded is different from civilian trauma. The war surgeon must develop a new clinical mindset.

Various indices have been devised to define injury patterns, which differ according to the type of conflict and the nature of the weapons used: the lethality of wounding agents; the anatomic distribution of wounds; delay to treatment; and pre-hospital and post-operative mortality are among the most important. Knowledge of these helps determine the high-risk factors affecting the outcome of surgical treatment and what set of specialist skills are needed: orthopaedic surgery, vascular, visceral, etc.

---

[4] H.E.L.P.: Health Emergencies in Large Populations. The reader is directed to the publications in the Selected bibliography for further study.

These patterns also have an impact on hospital activities and surgical workload and influence the standardization of protocols for patient management.

5.3.2 Preparation of the institution

The institution involved – military sanitary services, public health ministry, ICRC or any other humanitarian agency – must also be prepared.

The standardization of protocols and knowledge of the workload permits the establishment of ready-made kits of equipment, medicines and supplies that are easily deployed in an emergency situation as a rapid response to well-known needs. This facilitates the setting up of a surgical hospital to treat the victims of war.[5]

Preparation also implies choosing the correct profile of surgeon required, and training him or her to understand the context, the pathology, and the functioning of the institution in which they will be working, as well as the prevailing clinical protocols.

Knowing what kind of patient comes through the hospital door helps in planning a response: allocation of medical resources and personnel and the training of surgeons.

5.3.3 Surgical audit: monitoring

Monitoring of the outcome of patient management in a context of of armed conflict involves a surgical audit using an epidemiological approach: demographics, types of wounds, anatomic distribution, mechanism of injury, time since injury, post-operative mortality, number of operations and blood transfusions per patient, and morbidity, etc. There are several reasons for this.

- The adequacy of surgical expertise is important: the "fog of war" and the confusion and stress of a situation of armed conflict are not an excuse for bad surgery. Professional standards must be maintained, even when resources are limited.
- The adequacy of the clinical protocols in use must be put to the test: changes must be made in response to poor or unexpected outcomes. Many surgeons will be seeing war pathology for the first time, under conditions that do not resemble their routine work circumstances. At times, a strict and rigid application of protocols is required; on other occasions, there will be a need for improvisation and adaptation.
- The adequacy of the chain of casualty care must also be put to the test: hospital as well as pre-hospital data helps to assess the efficiency of first-aid measures in the field and that of the evacuation system. Surgeons, civilian or military, may be involved in training or helping set up pre-hospital programmes so that the injured reach them in better condition and in a timely fashion, which makes their surgery easier and ensures a better outcome for the patient.
- The role of surgeons in the development and application of international humanitarian law.

---

[5] The *Emergency Items Catalogue of the International Movement of the Red Cross and Red Crescent* has a listing of such kits for the establishment of a surgical field hospital and for the treatment of 100 war-wounded hospitalized and operated on.

5.3.4 The surgeon and IHL

Surgeons working for the ICRC believe that health professionals have a duty to gather data "to prevent and not just to treat" with regard to the law and norms of behaviour in the conduct of hostilities. They are not alone in this belief.

Health professionals were instrumental in gathering the epidemiological data from clinical studies to make the humanitarian argument for the banning of blinding laser weapons and anti-personnel landmines.[6]

In the field, ICRC delegations in areas of armed conflict monitor the compliance of belligerents with IHL. This is an essential element of the mandate of the ICRC to protect and assist the victims of armed conflict and promote respect for the law of war. Epidemiological studies, in both hospital and pre-hospital settings, can help in determining certain abuses of IHL. For example, mortality statistics can help elucidate the occurrence of executions of prisoners of war being passed off as killed in combat. The ICRC then undertakes a series of confidential *démarches* with the offending party in order to promote observance of humanitarian norms.

Other actors may do likewise, in compliance with their humanitarian responsibilities.[7] The ICRC warns, nonetheless, against the possibility of exploitation and manipulation of statistical epidemiological studies for political purposes.

5.3.5 The surgeon and the specialized literature

"Statistics can be made to say whatever you want them to say" is a well-worn cliché. It is therefore important for the surgeon reading the vast literature on war surgery to understand what is being said, what the shortcomings are, and what some of the traps and pitfalls are.

Most of the war surgery literature consists of large retrospective studies written by military surgeons of industrialized countries; some of the exceptions are included in the Selected bibliography at the end of this volume. Most surgeons are not well trained in epidemiology or statistical methods, and civilian surgeons are usually not knowledgeable about military terminology. Thus, the surgeon facing the new experience of dealing with the war-wounded who wishes to read up on the topic will often find a bewildering vocabulary and methodology.

Differences

Added to the differences between war and civilian trauma are the disparities between the experience of ICRC interventions (and other humanitarian agencies) and ministry of public health structures on the one hand, and classical military medical services on the other. The rest of this chapter deals with some of these differences and disparities.

---

[6] Respectively, Protocol on Blinding Laser Weapons 1995 (Protocol IV to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects,1980), and the Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction, 1997, known as the Ottawa Convention.

[7] Burnham G, Lafta R, Doocy S, Roberts L. Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey. *Lancet* 2006; **368**: 1421 – 1429.; Dudley HAF, Knight RJ, McNeur JC, Rosengarten DS. Civilian battle casualties in South Vietnam. *Br J Surg* 1968; **55**: 332 – 340.

5.4 General questions of methodology

There are a number of problems and difficulties relating to how epidemiological studies have been conducted. The surgeon reading the literature should be aware of them.

5.4.1 Large and small numbers

First it must be clear which phenomenon is being studied. Comparisons of data between a single battle and longer wars may not be valid; a small cohort may not be representative. Most clinical protocols take as a basis large numbers in big wars. On the other hard, the individual surgeon will face individual casualties from particular battles. Thus, the experience with small cohorts may help prepare the surgeon to face a novel situation or specific injuries.

5.4.2 Problems of data collection

Trauma registries are notoriously difficult to maintain during war. Fatigue and lack of time, a shortage of trained staff, and the personal danger that medical staff must often confront complicate the keeping of statistics. The administrative personnel and time and effort required to keep up full documentation and archives can appear to be a luxury when faced with the stress of caring for battle casualties. In an article on American casualties during the war in Viet Nam, based on data obtained from a statistical record filled out by physicians in Viet Nam at the time of discharge or transfer, the author made the following telling comment.

> " Many of the physicians who filled out these forms recognized that they were frequently completed under stress, by command, and without enthusiasm, by physicians who considered it another thankless task, and therefore the accuracy of some of the data may be questionable…"
>
> R. M. Hardaway III [8]

Besides the problems of data collection mentioned with respect to public health statistics and the displacement of populations, other discrepancies and difficulties appear when dealing with a retrospective analysis of hospital-based data.

- The quantity and quality of data from different hospitals are irregular.
- Initial admission data, operating theatre observations, and ward medical records for individual patients do not always match.
- Hospital records are often incomplete, or contain clerical errors.
- The management of casualties at different points in a military chain of evacuation compounds reporting problems, particularly with respect to patient outcome.
- Access to medical care for the wounded – especially civilians or combatants in a civil war or during irregular guerrilla warfare – is not always available or must be done "in secret".
- Patients may disappear from the hospital before completing their treatment because of personal security concerns.
- Families do not always bring their dead for registration.

As mentioned before, one must also always consider the political and military sensitivity of trauma information; ICRC health personnel have on occasion been accused of being "spies" because they were asking a hospital director for admission statistics.

---

[8] Hardaway RM III. Viet Nam Wound Analysis. *J Trauma* 1978; **18**: 635 – 643.

5.4.3 Important first questions: who is counting whom?

Foremost is the very important problem of different definitions of epidemiological categories. Some military definitions have changed over the last century or been replaced by other terms. Articles written by civilian authors have often invented their own epidemiological categories and definitions. Who exactly composes the target population under study is different from one article to the next, although all deal with the "wounded" and during the same war. Large retrospective studies are fraught with such traps.

There are important first questions the reader must ask of all war surgery studies: who is counting whom and where? When is a wounded person counted as a wounded person? Which patients are included in the nominator and denominator in any formula or equation?

Important first questions:

Who is counting whom?
When is a wound a wound?
Who are the nominator and denominator?

Who is counting?
Different hospitals – military, public, and other (mission hospitals, non-governmental organizations, ICRC) – all have their own objectives, mandates and routines. They all collect statistics; but often for very different reasons, and very different statistics (see Annex 5. A: ICRC Surgical Database).

A well-organized classical army can tell you how many of their soldiers were killed or disabled; pensions and allowances must be paid and the necessary administrative structures to do so exist. Guerrilla forces and civilian medical facilities are incapable of fully recording such statistics, with rare exceptions.

Who is being counted and where?
Are casualties calculated at the point of wounding, at first-aid posts, first echelon or referral hospitals? The total number of wounded will depend on which level of the evacuation chain is counting, and what happens to them after injury.

The wounded that arrive at hospitals – where most studies are made – are only a sample of the casualties; they do not represent the total reality of war.

When is a wounded person a wounded person?
Many studies in the military literature, even when dealing with the same war, define the wounded differently. One example: for World War II, the United States Surgeon General's Statistical Health Report indicated 724,000 wounded and 228,000 battle-related deaths among American troops; a fatality rate of 23.9 %. The Adjutant General's Report gave the figure of 593,000 battle injuries and 235,000 deaths; a rate of 28.4 %. The latter report excluded the lightly wounded.[9] Retrospective analyses do not always indicate which figure is being used.

Studies do not always make clear if all superficial and light injuries are excluded or not from the calculations. To state the obvious: a scalp laceration is not the same as an open wound to the brain. The reader should note that non-penetrating head, chest, and abdominal injuries are not always clearly differentiated from penetrating ones. The use of the Red Cross Wound Score permits this distinction (see Chapter 4).

Similarly, studies of the anatomic distribution of wounds, the wounding agent, and delay in evacuation do not always state specifically whether the authors are dealing with the survivors only, or with total combat casualties comprising the dead as well as the wounded survivors.

---

[9] Carey ME. Learning from traditional combat mortality and morbidity data used in the evaluation of combat medical care. *Mil Med* 1987; **152**: 6 –12.

5.4.4 Some definitions: the wounded and the dead

Table 5.2 includes some basic epidemiological definitions intended to standardize reporting by the armed forces of the United States of America.

Killed in action (KIA)	Those dying from battle injuries before reaching a medical facility. Comparable to civilian dead on arrival (DOA).
Wounded in action (WIA)	Those wounded who survive to reach a medical facility staffed by a physician capable of trauma life support, whether pre-hospital or hospital.
Died of wounds (DOW)	WIA but dying from wounds later on.
Returned to duty (RTD) or Carded for Record Only (CRO)	Wounded but not hospitalized and returned to active duty within 72 hours. May include patients who are registered in a pre-hospital setting: carded for record only but not evacuated.
Hospital mortality	Died post-operatively in a surgical facility.

Table 5.2 Standardized epidemiological definitions as used by USA armed forces.[10]

ICRC EXPERIENCE

In Kisangani, a city of 600,000 inhabitants in the remote east of the Democratic Republic of the Congo, major fighting broke out in June 2000. Hostilities lasted seven days. One week later, visiting ICRC delegates recorded that 4 hospitals and 62 clinics had registered 1,691 wounded; 369 were still hospitalized (over 90 % civilians, most military casualties had been evacuated). The remaining casualties were under ambulatory care. It was impossible to determine the number killed. Almost a month later, a total of 2,393 casualties had been registered and 248 were still hospitalized. The vast majority of these patients were so superficially wounded that, had they been living in a rural environment far from a health facility, they would never have presented. These are the civilian equivalent of RTD and CRO.

In nine and half months in Beirut, in 1976, an ICRC field hospital treated 1,289 war-wounded on an outpatient basis and another 696 in hospital. Similarly, three months of urban warfare in Monrovia, Liberia in 2003 saw 2,588 patients triaged by an ICRC surgical team, but only 1,015 were admitted to hospital.

These examples tell us a great deal about the sociology of medical care during armed conflict and the effects on statistical results. In urban warfare, all civilian city-dwellers who are injured, even with very minor wounds, will seek medical treatment and be registered as war-wounded. Everybody goes to the hospital! Their reasons for presenting at a hospital or clinic are many: the medical sophistication of an urban population used to seeking curative care; the need for reassurance and being in a "safe" haven; hopes of financial compensation for all those injured; and the search for simple psychological support for a traumatized civilian population.

Although the presence in hospitals and clinics of so many superficially wounded patients may not represent the true extent of the surgical workload, the burden on the emergency department (triage and first aid), nursing care (wound dressings), and non-clinical work (administration, registration, laundry and kitchen) can be overwhelming. Statistical analysis is both difficult and tedious under such circumstances (e.g. Monrovia) or almost impossible (e.g. Kisangani).

---

[10] Holcomb JB, Stansbury LG, Champion HR, Wade C, Bellamy RF. Understanding combat casualty care statistics. *J Trauma* 2006; **60**: 397 – 401.

5.4.5 Clinical and operational consequences and IHL

A number of clinical and operational elements are significant as regards IHL.

1. For armies, the total number of casualties – soldiers removed from active participation in combat or support – and the overall lethality of the battlefield as represented by the case fatality rate (CFR) are important in terms of medical planning and allocation of fighting resources.

   For civilian health structures, the total number of casualties is only one measure of the social and economic burden of war and the humanitarian consequences of the disorganization of society.

2. Killed in action (KIA) helps to determine the lethality of weapons and the type of combat. IHL concerns itself with the lethality of weapons systems and the conduct of hostilities.
3. The KIA and died of wounds (DOW) help determine the efficiency of the evacuation chain, which is dependent on access to the wounded, appropriate pre-hospital care (field triage and initial treatment), and transport logistics (evacuation time and access to a hospital for the wounded). Access of health personnel to the wounded and access of the latter to medical facilities are supposed to be guaranteed, in accordance with the protection conferred upon these persons by IHL.
4. The adequacy of medical treatment depends on the efficiency of the evacuation system, the wounding mechanism and pathology (burns, blast, penetrating wounds, anti-personnel mines, etc.), and hospital competency. DOW and hospital mortality are the main indicators. Again, IHL is interested in the medical consequences of weapons systems.
5. These various indices of mortality are important. However, they say nothing about the severity of injury in survivors – morbidity and disability – or surgical workload that may have a greater long-term effect on civilian populations and health structures. The number of operations per patient, infection rates, amputation rates, and length of hospitalization better define the severity of injury and hospital burden of war-wounded. The humanitarian consequences of such effects are of direct concern to IHL.

5.5 Aetiology of injury

5.5.1 Warfare scenarios

The type of warfare greatly influences the kinds of wounds the surgeon sees and the anatomic distribution of injuries, which obviously affects hospital workload. The widespread use of anti-personnel blast mines in guerrilla and counter-insurgency warfare results in many traumatic amputations of the legs. Close combat with assault rifles causes many gunshot wounds with severe single injuries; shelling and bombing from a distance produce multiple fragment wounds, many of which are superficial. Major contemporary wars have seen a shift from gunshot wounds to fragment injuries in the majority of casualties.

In intercommunal strife, combatants in some societies engage in “traditional” warfare using machetes and *pangas*; slash wounds to the head, neck and shoulders predominate and, secondarily, injuries to the forearms when the victim attempts to protect him or herself. Elsewhere, the ready availability of small arms means that “traditional” combat and cattle raiding are now undertaken with the use of AK-47 Kalashnikov rifles.

Weapons systems available to belligerents vary. Classical armies of industrialized countries use air bombardment as well as artillery and armour, and their infantry is often fitted out with personal protective equipment. Classical armies of low-income countries tend to rely on infantry and artillery. Rural guerrilla formations depend on ambush with personal firearms and landmines; urban guerrilla warfare is largely street fighting with rifles, rocket-propelled grenades and, sometimes, short-range mortars. The weapons systems available to combatants can also change during the course of a war.

The intensity of fighting, being on the offensive or defensive, losing or winning a given battle, the size of the unit actually engaged, and the percentage of soldiers really exposed to combat, all affect various combat casualty statistics.

Relevant clinical consequences

The numbers and relative proportion of civilian and military wounded and dead depend to a large extent on combat strategy and operational tactics, geographical site of fighting, access to medical care, and who is winning or losing.

5.5.2 Definitions of weapon-systems: methodology

Some confusion arises in surgical literature because of differing categories and definitions for wounding agents and mechanism of injury: there is no universally accepted standard.

“Landmines” include both anti-tank (ATM) and anti-personnel (APM); the two are usually not differentiated. ATM may be the agent of wounding, but the mechanism of injury can be blast, blunt, burn, or penetrating fragment. APM may be either a blast or a fragmentation device. The severity of the injury differs with the mechanism and aetiology, but this cannot always be determined from the categories in a study.

The category of “fragment wounds” often groups together injuries from shells, bombs, and grenades, together with APM. The differences are important in terms of surgical management, however. Small, superficial wounds from preformed grenade fragments often do not require surgery; this is never the case for APM fragments (see Chapter 10). Table 5.3 gives an indication of the distribution of wounds according to projectile, from a number of conflicts.

	Bullets %	Fragments %	Other %
World War I (Western Allies)	39	61	
World War II (Western Allies)	10	85	5
Korea (USA)	7	92	1
Indochina (France)	62	38	
Algeria (France)	71	23	
Borneo (UK)	90	9	1
Lebanon: Beirut 1975 – 86	49	36	14
Northern Ireland	55	22	20
Falkland Islands / Malvinas (UK)	32	56	12
Thailand 1981	38	20	42 (APM)
Lebanon 1982 (Israel)	12	53	35
Lebanon: Bourj el-Barajneh refugee camp 1986 – 87	20	60	20
Eritrea 1988 – 91	33	63	2.2 (APM)
Bosnia-Herzegovina: Sarajevo 1992 – 96 (France)	59	37	4 (APM)
Bosnia-Herzegovina 1993 (Bosnia-Herzegovina)	49	46	5
Croatia 1991 – 93 (Croatia)	25	70	6 (APM)
Yugoslavia 1991 – 92 (Yugoslavia)	41	2	52 (APM)
Somalia: Mogadishu 1992 (US helicopter)	55	31	14

Table 5.3 Aetiological distribution of the war-wounded: generally accepted historical examples. Anti-personnel landmines included in some. The country name in brackets indicates the source of information, see Selected bibliography.

5.5.3 ICRC statistics

ICRC surgical teams have worked in many different wars and combat scenarios. Table 5.4 shows the mechanism of injury in different types of warfare.

Hospital	N =	Fragments %	Bullets %	Mines %
Butare	40	8	92	–
Kabul	6,244	52	29	19
Kao-i-Dang	1,067	22	16	63
Kandahar	1,159	24	50	26
Novye Atagi	186	44	35	22
Peshawar	4,340	42	23	35
Quetta	6,570	33	39	28
Lokichokio	12,196	10	87	2
Monrovia JFK Memorial	867	38	62	–

Table 5.4 Aetiology of war-wounded in a number of ICRC hospitals.

Extremes of the distribution of wounding mechanisms are to be found in these very different conflicts, and reflected in the statistics held in the various hospitals (Figure 5.3). Southern Sudan was the scene of a rural guerrilla war in semi-desert, sahelian landscapes, and reed swamps (Lokichokio hospital in northern Kenya) where most injuries were gunshot wounds (GSW). Fighting in Monrovia, Liberia, was irregular urban warfare (JFK Memorial Hospital). Patients coming to Kao-i-Dang hospital (Thailand) had been injured in rural guerrilla warfare waged in the forests and jungles of Cambodia with widespread use of anti-personnel landmines. Needless to say, the types of pathology encountered in these three hospitals differed widely.

Figure 5.3

Different combat tactics produce very different distributions of wounding agents.

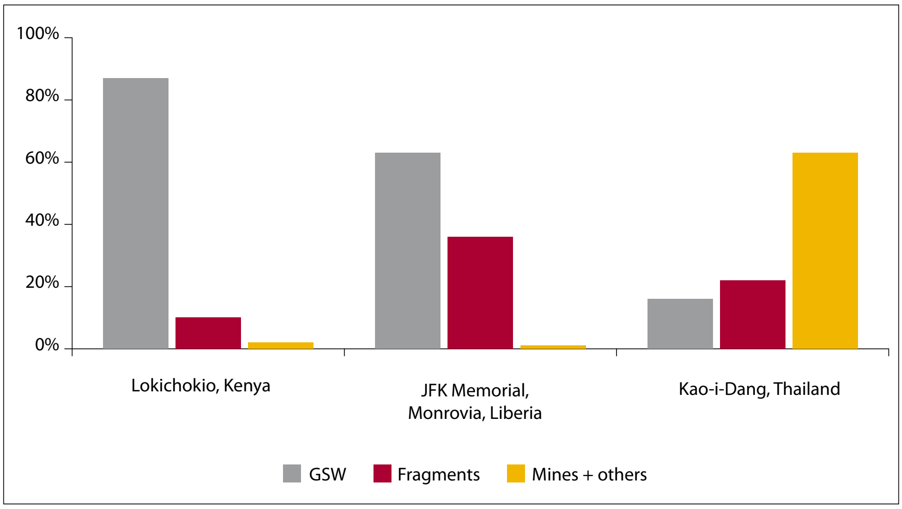

5.6 Anatomic distribution of wounds

The aetiology of wounds and their anatomic distribution change constantly depending on the development of new weapons systems, personal protective material (body armour), and the nature of combat. One constant, however, is the preponderance of wounds to the limbs amongst survivors. This distribution is important in determining surgical workload.

5.6.1 Body exposure

Body exposure to injury differs according to the type of weapon:

- anti-personnel blast mines strike the legs (Pattern 1 APM injury);
- manipulation of mines injures the upper limbs, face and chest (Pattern 3 APM injury) as described in Chapter 3;
- fragmentation devices spray the whole body with multiple hits in a random fashion; and, secondarily, according to the type of combat and military activity:

- aimed sniper fire targets the head or chest;
- trench warfare particularly exposes the head, resulting in the many mutilating facial wounds of World War I, known as "*gueule cassée*";
- purposeful observation and lookout activities of the soldier also increase the risk to the head and neck.

On the other hand, improved helmets and body armour, covering the chest and upper abdomen, effectively modify the relative exposure of anatomic regions.

Apart from aimed sniper fire, projectiles cause injuries in a random manner. Traditionally, combat exposure to injury has been estimated using as a basis the percentages of body surface area applied to assess burns, but taking into account operational necessities for soldiers (Table 5.5). It should be noted that combat exposure of the head and limbs is over-represented when compared to burns.

	Head	Neck	Thorax	Abdomen	Limbs
Body exposure in combat %	12		16	11	61
Surface area (burns) %	9	1	18	18	54

Table 5.5 Percentage body surface area versus surface area exposed in combat.[11]

5.6.2 Definitions and methodology

The anatomic distribution of wounds has been remarkably consistent over the last century. Reporting, however, has been remarkably inconsistent. Some studies cite only the primary site of injury if there is more than one; others include the category "multiple". Some count the "wounds" and not the "wounded", so that sites of injury exceed the number of patients. Many studies do not clearly define the counting method; in some reports only survivors are counted, while for others total combat casualties – KIA and WIA – are included. Again, the methodology is not always specified.

Anatomic regions are not standardized either. Different studies use different definitions for anatomic regions; there is again no universal standard. Some report only the "torso"; "pelvis and buttocks" are under a separate category from "abdomen" in some, and included in others. To be truly accurate, head, face, and neck categories should be clearly separated, which they rarely are. These latter wounds represent very different clinical problems – traumatic brain injury, asphyxia, haemorrhage – and degrees of lethality.

Anatomic distribution furthermore only indicates a potential threat; crude statistics often include superficial wounds, which is not a true indication of hospital workload and threat to life and limb. As has been mentioned above, it would be preferable to separate out penetrating from non-penetrating wounds of vital regions: head, thorax, and abdomen. The Red Cross Wound Score and Classification System attempts to address some of these problems in a simplified manner.

---

[11] Carey ME, 1987.

5.6.3 Historical results

The reader going through the literature will find very different figures for the same war, depending on the source and the methodology. This can make for frustrating reading. Nonetheless, some historical approximations are presented in Table 5.6. Wounds to the limbs predominate, from 50 % to 79 %.

Conflict	Head & Neck %	Thorax %	Abdomen %	Limbs %	Other & multiple %
World War I (Western Allies)	17	4	2	70	7
World War II (Western Allies)	4	8	4	79	9
World War II (USSR)	19	9	5	67	–
Korea (USA)	17	7	7	67	2
Viet Nam (USA)	14	7	5	74	–
Borneo (UK)	12	12	20	56	–
Northern Ireland	20	15	15	50	–
Arab-Israeli war 1973 (Israel)	13	5	7	40	31
Thailand 1981	10	12	4	66	8
Falkland Islands / Malvinas (UK)	16	15	10	59	–
Lebanon 1982 (Israel)	14	5	5	41	34
Lebanon: Bourj el-Barajneh refugee camp 1986 – 87	12	16	18	54	–
Gulf War 1991 (UK)	6	12	11	71	(32)*
Gulf War 1991 (USA)	11	8	7	56	(18)**
Afghanistan (USSR)	16	12	11	61	–
Chechnya 1995 (Russia)	24	9	4	63	–
Somalia: Mogadishu 1992 (USA)	20	8	5	65	(2)**
Croatia 1991 – 93 (Croatia)	15	11	4	69	1
Yugoslavia 1991 – 92 (Yugoslavia)	21	9	8	62	(23)**
Croatia 1991 (Yugoslavia)	12	15	8	65	–
Bosnia-Herzegovina 1992 (Yugoslavia)	14	15	9	62	–
Bosnia-Herzegovina 1993 (BH)	19	16	11	53	–
Bosnia-Herzegovina: Sarajevo 1992 – 96 (France)	11	11	14	61	3
Bosnia-Herzegovina: Sarajevo market bombing 1995	13	13	12	62	–
Eritrea 1988 – 91	20	9	6	63	2
ICRC surgical database	12.5	7.2	7.8	65.5	7
Overall average	~15	~10	~7	~65	

* Buttock and back wounds – all multiple fragment injuries – as a separate figure.

** Multiple wounds.

Table 5.6 Anatomic distribution of major wounds; some statistics include both dead and survivors, others include minor injuries. The country name in brackets indicates the source of information, see Selected bibliography.

5.6.4 ICRC results

The experience of ICRC surgical teams in various conflicts closely mirrors that of surgeons in other wars, particularly in relation to the type of conflict (Table 5.7). Many patients have multiple wounds; the distribution is given for injuries and not for patients, hence more wounds than patients. Only one wound is counted per anatomic region, however.

Pelvis and buttocks, back and soft tissue are given separately. Head, face and neck are not differentiated further.

Hospital	Number	Head & Neck %	Thorax %	Abdomen %	Pelvis & Buttock %	Back & soft tissue %	Upper limb %	Lower limb %	Limbs combined %
Kabul	8,804	15	9	10	4	3	24	35	59
Kao-i-Dang	1,660	15	8	7	4	3	24	39	63
Peshawar	6,840	18	8	6	5	3	25	35	60
Kandahar	1,396	11	9	11	3	2	24	40	64
Quetta	9,373	15	9	8	5	3	24	36	60
Butare	45	16	7	2	–	2	31	42	73
Novye Atagi	210	10	3	7	2	3	26	50	76
Lokichokio	14,203	7	8	3	7	3	29	44	73
Monrovia	904	14	13	4	4	–	21	43	64

Table 5.7 Anatomic distribution of wounds in various ICRC hospitals serving victims of differing types of combat.

5.6.5 Primary tissue injury

The anatomic distribution of wounds and the analysis of primary tissue injury permit the assessment of the surgical workload. In both, wounds to the limbs predominate; soft-tissue wounds and fractures of long bones constitute the majority of the injuries in all published series. The WDMET[12] statistics from Viet Nam are eloquent (Table 5.8); as are the figures from a French military field hospital in Sarajevo, 1992 – 1996, (Table 5.9). These analyses take into account non-penetrating wounds to vital regions.

Soft tissue (especially of limbs)	47 %
Extremities (long bone fractures)	26 %
Abdomen	8 %
Chest	4 %
Neck	2 %
Face	6 %
Head	2 %
Multiple	5 %

Table 5.8 Distribution of wounds according to primary tissue injury (USA).[13]

Soft tissue (especially of limbs)	56 %
Extremities (long bone fractures)	22 %
Abdomen	14.5 %
Chest	11 %
Face & Neck	6 %
Head	6 %
Peripheral vessels	6 %
Multiple	6 %

Table 5.9 Distribution of wounds according to primary tissue injury (French military field hospital in Sarajevo).[14]

ICRC results

In the hospitals included in the ICRC database, where evacuation was difficult for numerous patients and many with minor soft-tissue wounds did not bother to come to hospital, soft-tissue injuries still represented 36 % of wounds; fractures of the extremities 46 %; and vital central wounds 20 %.

Various other ICRC reports confirm these findings (Tables 5.10 and 5.11).

Soft tissue	33 %
Bones and joints	33 %
Vascular	11 %
Peripheral nerves	11 %
Abdomen hollow organs	17 %
Abdomen solid organs	9 %
Urogenital	5 %
Chest	9 %
Brain and spinal cord	3 %
Maxillo-facial	3 %
Eye, internal ear	2 %
Other	1 %

Table 5.10 Distribution of wounds according to tissue injury, hospitalized patients in the ICRC field hospital, Beirut 1976 (N = 696).[15]

---

[12] In the military literature, the Wound Data and Munitions Effectiveness Team (WDMET) of the United States set a gold standard for data collection. A large number of administrative staff meticulously collected information on 7,989 patients during the war in Viet Nam from 1967 to 1969 in a systematic way. The analysis was published by: Bellamy RF. Combat trauma overview. In: Sajtchuk R, Grande CM, eds. *Textbook of Military Medicine, Anesthesia and Perioperative Care of the Combat Casualty.* Falls Church, VA: Office of the Surgeon General, United States Army; 1995: 1 – 42. This report is often cited in review articles.

[13] Champion HR, Bellamy RF, Roberts P, and Leppäniemi A. A profile of combat injury. *J Trauma* 2003; **54 (Suppl.)**: S13 – S19.

[14] Versier G, Le Marec C, Rouffi J. Quatre ans de chirurgie de guerre au GMC de Sarajevo (juillet 1992 à août 1996) *Médecine et armées* 1998; **26**: 213 – 218.

[15] Translated from: Kjaergaard J. Les blessés de guerre de l'hôpital de campagne du CICR à Beyrouth en 1976. *Schweiz Z Milit Med* 1978; **55**: 1 – 23.

Parts injured	Total % (N = 1,033)	Bullet wounds % (n = 231)	Shell fragment wounds % (n = 508)	Mine injuries % (n = 294)
Soft tissue	73	67	75	70
Bone	39	52	20	63
Thorax	7	7.5	9	4
Abdomen	11	10.5	14	7.5
Brain	2.5	–	5	1
Other	4	2.5	4	4

Table 5.11 Tissue injuries according to wounding weapon, Kao-i-Dang ICRC hospital 1984 – 85. [16]

The clinically important point for the surgeon is the very large hospital workload represented by soft-tissue and orthopaedic injuries.

5.7 Fatal injuries

5.7.1 Site

It is not a simple matter to determine cause of death and site of lethal injury. Multiple injuries tend to have a synergetic effect and it can be impossible to define which of several wounds was the direct cause of death. Furthermore, many fatal war injuries involve total body disruption or severe mutilation.

A formal complete autopsy for every combat death is onerous even for the armed forces of a wealthy industrialized country, and this has rarely been common practice. Three examples of the anatomic distribution of fatal wounds are given in Table 5.12.

	US dead in World War II[17]	US dead in Viet Nam (WDMET)[18]		Israeli dead in Lebanon 1992[19]	
Head	42 %	37 %	46 %	9 %	34 %
Neck		6 %		3 %	
Face		3 %		22 %	
Chest	30 %	24 %		45 %	
Abdomen	12 %	9 %			
Multiple	–	17 %		–	
Limbs	13 %	3 %		21 %	
Soft tissue	–	1 %		–	

Table 5.12 Anatomic distribution of fatal wounds.

Central wounds predominate, as would be expected; particularly to the head, face, neck and thorax.

The great majority of war wounds are to the limbs.
The most lethal wounds are to the head and chest.

---

16 Trouwborst A, Weber BK, Dufour D. Medical statistics of battlefield casualties. *Injury* 1987; **18**: 96 – 99.

17 Garfield RM, Neugut AI. Epidemiologic analysis of warfare. *JAMA* 1991; **266**: 688 – 692.

18 Champion HR, et al., 2003.

19 Gofrit ON, Kovalski N, Leibovici D, Shemer J, O'Hana A, Shapira SC. Accurate anatomical location of war injuries: analysis of the Lebanon war fatal casualties and the proposition of new principles for the design of military personal armour system. *Injury* 1996; **27**: 577 – 581.

5.7.2 Trimodal distribution of trauma mortality

Mortality in civilian trauma was classically described by D. Trunkey in 1983[20] to fall into three categories: immediate death (50 %), early death (30 %), and late death (20 %), see Figure 5.4.

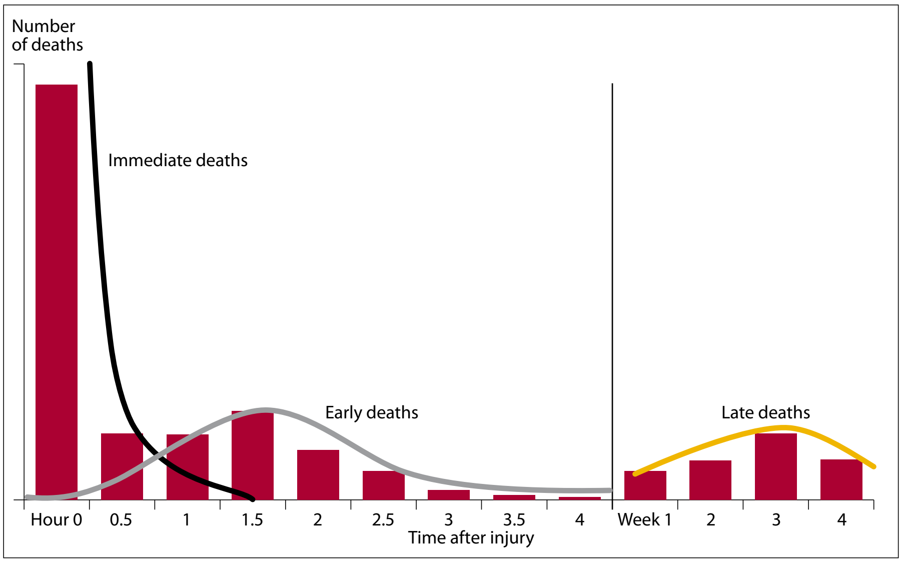

Figure 5.4
Trimodal distribution of trauma deaths.

Peak 1: immediate deaths
Most deaths occur at the moment of injury or within minutes. These are due to overwhelming trauma incompatible with life (severe brain injury, massive haemorrhage).

In war, in addition to severe injuries to the head and torso (heart, liver, and major blood vessels), some casualties suffer total body disruption or incineration through burns. It is estimated that 70 % of deaths occur within five minutes, and little to nothing can be done for these patients, which represent about 17 – 20 % of the severely wounded.

Peak 2: early deaths
These occur after the first minutes until the first few hours after injury. This peak gave rise to the idea of the "golden hour"; if certain measures were implemented rapidly, the patient could be saved.

In war trauma, there are three major causes of early death:
- continuing exsanguinating haemorrhage;
- inadequate/impaired airway due to survivable penetrating head injury;
- compromised breathing due to tension pneumothorax.

Many of these early deaths are avoidable within this "golden hour" if adequate first-aid measures in the field are implemented in time.

Peak 3: late deaths
These occur a few days to weeks after injury and are the result of subsequent complications of infection, multiple organ failure and coagulopathy, and uncontrollable increased intra-cranial pressure due to post-injury cerebral oedema (closed injury).

Good and early first-aid measures can decrease infectious and other complications in the setting of armed conflict where wounds are dirty and contaminated from the very beginning. Inadequate early care affects morbidity (sepsis, disability) as well as mortality.

The understanding of this trimodal distribution of deaths stimulated an improvement in emergency medical services and patient transport in civilian settings. Rapid evacuation times and early advanced life support in many industrialized countries, in a bid to reach the injured during the "golden hour", have reduced these three peaks to a bimodal profile: immediate deaths and late.

---

[20] Trunkey DD. Trauma. *Sci Am* 1983; **249**: 220 – 227.

Applicability to the military war-wounded

The military have found that the trimodal distribution is relevant to armed conflict. They have tended to underline three categories of patients corresponding to what can be done under various operational scenarios in the field. This has an important influence on field triage categories.

1. Unsalvageable injuries: i.e. KIA for whom nothing can be done (17 – 20 %).
2. Severe but potentially survivable injury (10 – 15 %).
3. Moderate to minor injuries (65 – 70 %).

One result of improved pre-hospital care and evacuation is that some casualties become DOW (early deaths) instead of being KIA (immediate deaths) while a few others are saved. The most important field measures undertaken include stopping haemorrhage from controllable bleeding and maintaining a patent airway and breathing by simple procedures.

ICRC EXPERIENCE

Not all penetrating head injuries are so severe as to be incompatible with life. However, precarious pre-hospital care and long and difficult evacuation routes can result in large numbers of survivors of head injuries (WIA) succumbing later (DOW) because of inadequate control of the airway, resulting in asphyxia or vomiting and aspiration.

This was the case during one recent war in Africa involving a "classical" army. Large numbers of patients with salvageable head injuries were dying during a three-day evacuation in the back of lorries travelling over dirt roads in the bush. There was no possibility of monitoring endotracheal intubation.

An ICRC surgeon advised performing a tracheotomy in a front-line field hospital before evacuation, the only way to ensure an adequate airway under the circumstances in these comatose patients. The mortality rate for these cases was halved by this simple procedure.

Applicability to civilian war-wounded

At the shelling of the Markale market place in Sarajevo, August 28, 1995, only minutes away from two referral hospitals, 104 persons were injured of whom 42 ultimately died; a fatality rate of 40.8 %.[21]

Twenty-three died instantly and another ten were dead on arrival at hospital (79 % of deaths were first peak). Five patients succumbed during surgery (second peak 12 %) and another four a week later (third peak 10 %) as shown in Figure 5.5. The trimodal distribution of deaths appears to have been preserved.

Figure 5.5

Trimodal distribution of deaths after shelling of Sarajevo's Markale market.

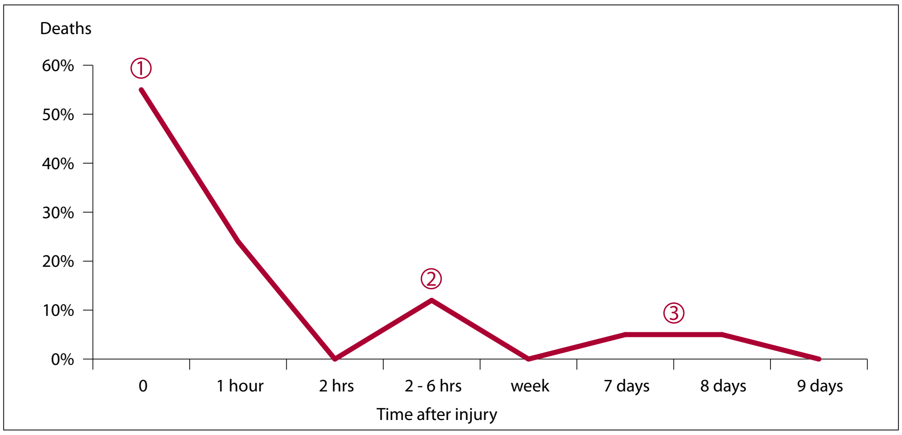

Furthermore, in remote rural areas with difficult evacuation routes, both armies and civilian populations will probably experience the trimodal distribution of deaths, as originally described. Trauma survival under such circumstances is geography specific.

---

[21] Suljevic I, Surkovic I. Medical aspects of the mass-scale civilian casualties at Sarajevo Markale Market on August 28, 1995: triage, resuscitation, and treatment. *Croat Med J* 2002; **43**: 209 – 212.

5.7.3 Ratio of dead to survivors

Many authors have noted that the ratio of dead to survivors in modern conflicts tends to be about 1:4 over the long term. This corresponds to our threshold fatality rate of about 20 – 25 %.

$$\frac{\textbf{Dead = KIA plus DOW}}{\textbf{Survivors = WIA minus DOW}} = \textbf{1:4}$$

Many factors can distort the specific results under given circumstances, including:

- the inclusion of minor wounds in the calculations – the famous methodological problem;
- the tactical situation (note the 40 % fatality rate of the Markale market place bombing mentioned above; burn injuries in tank crews or on ships; a successful surprise ambush, etc.);
- the lethality of particular weapons systems (anti-personnel landmines, napalm, etc.);
- any delay in medical evacuation; and
- the execution of wounded prisoners in contravention of the laws of war.

Improvement in personal protective equipment (body armour), better pre-hospital care, and the early availability of more aggressive surgery and enhanced intensive care facilities has changed the ratio in recent conflicts. For US armed forces in Iraq and Afghanistan, the dead to survivor ratio has become 1:8.[22] Effective body armour has decreased fatal wounds to the chest and abdomen, but increased wounds to the head and neck in survivors. No other armed forces use such improved body armour as extensively and the US experience is not easily translated to others, but it sets a standard nonetheless.

5.7.4 Lethality of weapons

If the KIA equals approximately 20 % and the DOW 5 %, this represents the total lethality of weapons during combat on land. Sea and air warfare are different.

It has long been recognized that different weapons systems have different lethality. In general, and over a large number of casualties the following percentages appear.

- Military rifle bullets: 30 – 40 % lethality; or one death in every 3 – 4 wounded.
- Randomly-formed fragments: 20 % for shells and 10 % for grenades.
- Preformed fragments: 15 % for shells and 5 % for grenades.
- Blast injury has a fatality of about 22 %.

Tactical considerations for distinct individual battles can change these figures. A well-laid ambush with small arms can easily result in more than 40 % of a small patrol being killed.

The particular case of anti-personnel landmines
All studies indicate the heavy surgical and nursing burden of injuries due to anti-personnel landmines, especially blast mines. Specific types of APM – bounding mines that jump up one metre in the air before exploding – invariably kill their victim: lethality lies close to 100 %.

Public health studies and ICRC surveys in poor countries where APM have been widely used in rural areas with little or no organized evacuation system and limited surgical facilities (Mozambique, Somalia, Cambodia, Afghanistan and Angola) indicate a case fatality rate of well over 50 % for traumatic amputations due to APM.

It is difficult to exclude the non-weapon factors in these studies and mortality statistics. Suffice it to say that the lethality of weapons includes their actual use in field conditions, and their total socio-economic and humanitarian effects. This is important in terms of IHL and was one of the significant factors in determining countries to negotiate the 1997 Ottawa Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-personnel Mines and on their Destruction.

---

[22] Holcomb JB, et al., 2006.

5.7.5 Conclusions of clinical relevance

From this analysis and further extrapolations a number of conclusions can be drawn.

1. Wounds to the head and torso are the most lethal; they represent most of the mortality. The great majority of survivors have injuries to the limbs; they represent the majority of surgical workload and morbidity.
2. Despite the fact that the head represents only 9 % of exposed body surface area, it accounts for a disproportionate number (25 %) of all battle casualties. The lethality of penetrating head wounds is approximately 75 % and they account for slightly less than 50 % of combat deaths (KIA + DOW), and about 8 % of survivors.
3. Head wounds kill either through devastating brain injury or through asphyxiation of the comatose patient who would otherwise survive the injury.
4. Death through closed head injury is relatively greater in the civilian setting than during armed conflict.
5. Uncontrolled exsanguinating haemorrhage accounts for another 50 % or so of deaths. Most of these (80 %) present central injuries to the thorax and/or abdomen, which require surgical facilities for control. The lethality of chest wounds is about 70 %.
6. The other 20 % of haemorrhage deaths are due to bleeding from peripheral, compressible blood vessels (half in the neck and half in the limbs). Thus, about 10 % of all deaths are due to haemorrhage from the limbs.
7. Death rates through haemorrhagic shock are higher in armed conflict than in the civilian trauma setting.
8. Little can be done medically for most immediate deaths and there is a threshold percentage of killed-in-action no matter how sophisticated the medical services are.

Conclusions for improvement of patient outcome

1. Some immediate and early deaths can be prevented by personal protective armour covering the torso.
2. Many early deaths can be prevented by simple measures to:
   - control limb haemorrhage;
   - relieve airway obstruction, especially in comatose patients after head injury;
   - relieve tension pneumothorax.
3. Some early deaths due to haemorrhage, particularly of the abdomen, can be prevented by rapid evacuation to a surgical facility.

5.8 The lethality of context: delay to treatment

5.8.1 Historical developments

Possibly one of the most important developments in combat casualty care over the last half-century has been improved pre-hospital care and the timely evacuation of the wounded to a surgical facility. Lasting days during much of World War I, evacuations took an average of 10.5 hours during World War II; the use of helicopters by American troops in Korea lowered this to 6.3 hours, and an average of 2.8 hours in Viet Nam. Transfer of Israeli wounded during the 1982 war in Lebanon took an average of 2.3 hours. The use of helicopters has radically changed evacuation and pre-hospital care for the armies of industrialized countries, but requires air superiority to be truly effective. It has also revolutionized equivalent civilian trauma systems. This resource is seldom available in a low-income country.

5.8.2 Forward projection of resources

Delayed evacuation may still result from operational contingencies, tactical situations, and difficult geography. Consequently, many armies have projected forward their surgical capacities close to zones of actual combat in an attempt to reach wounded soldiers earlier and avoid the mortality and morbidity related to any delay. The aim is primarily to reduce the number of "early" deaths.

Soviet deployment of "Special Surgical Teams" in forward locations in Afghanistan shortened time to reach surgery: 31 % of wounded were in surgery within 1 hour and another 39 % within 2 hours. Cumulatively, 92 % were in surgery within 6 hours with this system. This compares with an overall average during the war of 88 % operated within 12 hours, before and after the establishment, of the forward teams.

During fighting in Croatia in 1991, a Yugoslav mobile field hospital was located 5 – 10 km from the front: 61 % of wounded were evacuated within 30 minutes and another 22 % between 30 and 60 minutes of being injured.

US troops in Afghanistan and Iraq have also seconded "Forward Surgical Teams" and "Forward Resuscitative Surgery Suites" to the field; surgical care routinely begins within 1 – 4 hours following injury. Preliminary reports give a mean time of one and a half hours for evacuation of USA military wounded in Iraq to a forward surgical facility.

5.8.3 Urban warfare: hospitals on front lines

During urban warfare, fighting can take place literally in front of surgical facilities; more than once has a patient been wounded at the front door of a hospital. This was the case during much of the fighting in Beirut in the Lebanese civil war, often with evacuation times of only minutes. A French medical team in Sarajevo in 1992 – 96 (UNPROFOR – IFOR), serving military and civilian casualties, had evacuation times of 15 – 45 minutes.

Comparable times were encountered by ICRC hospital teams in Kabul in 1992 and Monrovia, Liberia, in 2003. ICRC and Somali Red Crescent surgical teams working in the Keysaney Hospital in North Mogadishu have witnessed similar short evacuation times from 1992 up until the present time of writing.

Although the vast majority of wounded patients arrive at the hospital within minutes under these circumstances, a few suffer considerable delay. Without an organized pre-hospital system and given an absolute absence of ambulances, civilians are often cut off from care facilities during street fighting and have to wait hours or even days for evacuation.

5.8.4 The paradox of early treatment: changing mortality ratios and rates

There is an apparent paradox of early evacuation and treatment: an increase in the absolute numbers who survive, but also an increase in the DOW and hospital mortality rates. Medics reach a number of former KIA in time; more severely injured casualties enter the evacuation chain earlier, but since natural automatic triage that weeds out the more severely injured does not come into play, a higher percentage of injured now die of their wounds after they have received care.

For US armed forces who died in World War II and Viet Nam, 88 % were KIA and 12 % DOW. In the conflicts in Iraq and in Afghanistan, the figures are 77 % and 23 %.[23]

---

[23] Holcomb JB, et al., 2006.

This effect has also been observed in hospital mortality rates at ICRC facilities (Tables 5.13 and 5.14).

Time until arrival in hospital	Patients (n =)	Died	Mortality
< 6 hours	3,114	172	5.5 %
6 – 24 hours	3,588	141	3.9 %
24 – 72 hours	1,668	46	2.8 %
> 72 hours	2,430	55	2.3 %

Table 5.13 Mortality according to evacuation time to ICRC hospitals: period from January 1991 to July 1993.

Longer evacuation times allow for "natural triage" to come into play. The more seriously wounded die before they reach hospital when evacuation is delayed.

Time until arrival in hospital	Patients (n =)	Died	Mortality
< 6 hours	79	5	6.3 %
6 – 24 hours	704	21	3.0 %
24 – 72 hours	210	5	2.4 %
> 72 hours	134	2	1.5 %

Table 5.14 Mortality according to evacuation time: patients transferred from ICRC first-aid posts to Peshawar ICRC hospital 1990 – 91.[24]

A new ratio: critical area to extremity wounds

A statistical method has been used to try to overcome this "bias" of logistical progress in order to better evaluate the effect of "natural triage" due to a lack of appropriate surgical care and evacuation facilities in low-income countries. In such conditions pre-hospital mortality cannot be properly determined either. The ratio of wounds in critical areas (trunk, head and neck) to wounds of the extremities is calculated among survivors: CA : Ext.

$$\frac{\text{Critical areas (head, face, neck, thorax, abdomen, pelvis)}}{\text{Non-critical areas (limbs, back, soft tissue injuries of trunk)}} = \text{CA : EXT}$$

CA = critical area, EXT = extremities and other non-critical areas.

Obviously, the wearing of body armour and the inclusion in statistics of minor or superficial, non-penetrating wounds to the head or trunk will distort the anatomic distribution of wounds, and the ratio.

In most wars, with evacuation counted in hours, the ratio is approximately 0.5. In urban areas and other situations of rapid evacuation, it tends towards 1. With increasing delay of evacuation to over 24 hours, and even days and weeks, the ratio falls.

Thus, studies of guerrilla and counter-insurgency warfare in difficult geographical terrain give much lower ratios indicating that the most seriously injured patients (critical area wounds) died before treatment (Table 5.15).

Conflict	Ratio
Thailand counter-insurgency	0.39
Eritrean war of independence	0.26
Ugandan guerrilla war	0.21
Afghan mujaheddin	0.07
ICRC Lokichokio hospital: southern Sudan guerrilla war	0.33

Table 5.15 Ratio of critical area to limb wounds in various insurgency conflicts.[25]

---

[24] Korver AJH. Outcome of war-injured patients treated at first aid posts of the International Committee of the Red Cross. *Injury* 1994; **25**: 25 – 30.

[25] Adapted from Bhatnagar MK, Smith GS. Trauma in the Afghan guerrilla war: effects of lack of access to care. *Surgery* 1989; **105**: 699 – 705.

5.9 Hospital mortality

5.9.1 Historical considerations

As mentioned, rates of killed in action for armies have remained relatively stable over the last half-century at about 20 – 25 %. Hospital mortality, however, has diminished greatly with medical progress (safe anaesthesia, blood transfusion, better understanding of the physiology of shock, antibiotics, and more aggressive surgical techniques) as shown in Table 5.16.

Conflict	Hospital mortality rate %
Crimean War 1854 – 55 (UK casualties)	16.7
American Civil War 1861 – 65 (Union casualties)	14.1
Boer War 1899 – 1901 (UK casualties)	8.6
World War I 1917 – 18 (USA casualties)	7.6
World War II (USA casualties)	4.5
American – Viet Nam War (USA casualties)	2.5

Table 5.16 Hospital mortality: examples from the past.

Hospital mortality has come to be an indicator of the efficiency of the management system, bearing in mind what has already been said about the paradox of early evacuation of the very severely injured. In calculating these figures, however, one must take into account the percentage of truly "vital" injuries and exclude or specify the superficial ones.

5.9.2 Hospital mortality *versus* post-operative mortality

A number of considerations must be taken into account if one wishes to use hospital mortality as an indication of the efficiency of patient management systems. In armed conflict, surgeons often face mass casualties. Some patients will be triaged into the "expectant" category, and given supportive treatment only, to die without pain and with dignity (see Chapter 9). These patients are recorded as DOW and are often included in hospital mortality figures.

Other patients die shortly after arrival, or "on-table", in an extreme attempt to save life. With a slightly longer evacuation time, many such patients would be KIA, dying before reaching hospital, as mentioned previously. Nonetheless, they are registered as DOW and are again included in hospital mortality figures.

Then there are the truly post-operative deaths; some from irreversible shock, others from irretrievable brain injury and, others still, from surgical complications, mostly sepsis, and additional medical pathologies.

Gross statistics of hospital mortality do not always take into consideration these very different categories of patients.

5.9.3 ICRC hospital mortality

Austere working conditions, often a limited number of professional staff and at times a perilous security situation, all make work in ICRC hospitals resemble that commonly seen in public hospitals in a poor country. Military medical services from industrialized countries may also face these constraints, but they tend to be of a different order of magnitude.

Post-operative mortality in ICRC hospitals varies from 2.2 % in Quetta and 3.1 % in Peshawar to 4.2 % in Kao-i-Dang, 4.8 % in Kabul, and 6.1 % in the Monrovia battle, where evacuation times were extremely short.

5.10 ICRC statistical analysis of hospital workload

5.10.1 Surgical audit: methodology

Without control over the pre-hospital phase and the known inadequacies of data collection, the mortality rate is not a credible statistic in ICRC surgical programmes and hospitals. The same is true of most public civilian structures. Far more substantial is the workload and morbidity as represented by the number of operations and blood transfusions per patient and the duration of hospitalization.

For statistical purposes, most patients, if well managed, undergo two operations: wound excision and delayed primary closure. There are a number of factors that explain this:

- the ICRC does not have a series of operative echelons of treatment as do the military; all operations are performed in the same hospital (see Chapters 1 and 6);
- rarely do ICRC surgeons perform planned serial debridements of war wounds, requiring more operations (see Chapter 10);
- some patients require a third operation for skin grafting to close the wound, or even more if a burn complicates a penetrating wound (see Chapter 11);
- others call for a single operation: (craniotomy, chest tube drainage, or laparotomy);
- some patients call for no operation, if they have been admitted solely for observation – expectant triage category, paraplegic, etc. (see Chapter 9);
- small superficial wounds are often treated conservatively with dressings and antibiotics, with the exception of those caused by anti-personnel landmines.

These "normal" single and multiple operations tend to balance each other out. (Pure burn injuries are not included in this discussion.)

Most patients require two operations: debridement and delayed primary closure. More than two usually denotes a complication, typically sepsis.

Performing more than two operations per patient usually denotes a complication, typically sepsis, and is a good indication of morbidity. Thus, as an appropriate approximation of workload and correct surgical management, the ICRC uses the average of two operations per patient in large studies. These premises form the basis for the following discussion.

5.10.2 Number of operations per patient: all patients

The number of operations per patient for all patients with the relevant information in the ICRC database is given in Table 5.17.

None	9 %
1 operation	16 %
2 operations	41 %
≤ 2 operations	66 %
3 operations	14 %
≥ 4 operations	20 %

Table 5.17 Number of operations per patient (N = 16,172).

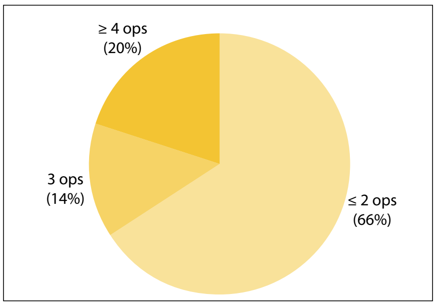

Figure 5.6

Number of operations per patient, simplified (N = 16,172).

To simplify, the number of operations are described as two or less; three; or four and more (Figure 5.6).

Fully two-thirds of all patients in the database had two operations or less, indicating a minimum of morbidity and good surgical management. There are a number of factors that influence this surgical workload: delay in evacuation, severity of the wound, type of wound, and wounding mechanism. A brief analysis of some of these factors is given in the following sections.

5.10.3 Number of operations according to delay in evacuation

Table 5.18 shows the results for the number of operations per patient according to evacuation time to the ICRC hospital in Kabul during a period of major urban combat.

Evacuation time	N = 6,140	None	1 op	2 ops	≤ 2 ops	3 ops	≥ 4 ops
< 6 hrs	3,214	7 %	30 %	47 %	84 %	7 %	8 %
6 – 24 hrs	1,606	7 %	23 %	51 %	81 %	9 %	11 %
24 – 72 hrs	605	6 %	24 %	50 %	80 %	7 %	12 %
> 72 hrs	715	9 %	26 %	42 %	77 %	9 %	14 %

Table 5.18 Number of operations per patient according to time of evacuation to ICRC Kabul hospital 1990 – 92.

These figures are confirmed if one examines the entire database (Table 5.19).

Evacuation time	N = 16,172	≤ 2 ops	3 ops	≥ 4 ops
< 6 hrs	2,409	81 %	9 %	10 %
6 – 24 hrs	3,727	70 %	13 %	17 %
24 – 72 hrs	2,785	69 %	13 %	17 %
> 72 hrs	7,251	71 %	12 %	17 %

Table 5.19 Number of operations per patient according to time of evacuation to ICRC hospitals 1990 – 99.

While hospital mortality is influenced by delay in evacuation, morbidity as determined by number of operations per patient apparently is not. The critical limit would appear to be 6 hours delay – the category "6 – 12 hours since injury" does not exist in the ICRC database, and this is a field for future study. However, before rushing to such a conclusion, another factor must be analysed: wound severity.

5.10.4 Number of operations according to Wound Grade

War wounds are categorized into one of three Grades of increasing severity according to the Red Cross Wound Score and classification system (see Chapter 4). The distribution in the ICRC surgical database is as follows.

- Grade 1: 42 %.
- Grade 2 : 37 %.
- Grade 3 : 21 %.

If one looks at the number of operations according to Wound Grade, an important difference appears (Table 5.20 and Figures 5.7.1 – 5.7.3).

Severity of wound		Number of operations per patient					
	Patients (N =16,172)	None %	1 op %	2 ops %	≤ 2 ops %	3 ops %	≥ 4 ops %
Grade 1	n = 6,729	16	23	45	84	9	7
Grade 2	n = 5,974	4	12	44	61	16	23
Grade 3	n = 3,469	3	7	30	40	19	41

Table 5.20 Number of operations per patient according to Red Cross Wound Grade.

The Grade of the RCWS shows a distinct difference in the number of operations required. Surgical workload is definitely influenced by the severity of injury, as would be expected, and the RCWS is capable of determining this.

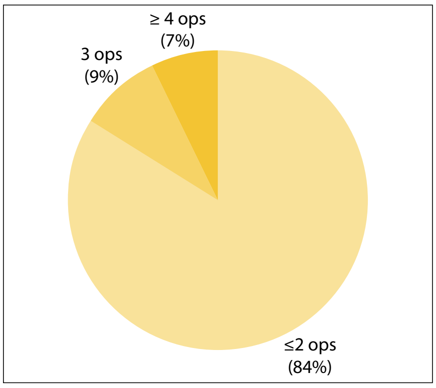

Figure 5.7.1

Number of operations per patient for Grade 1 wounds (N = 6,729).

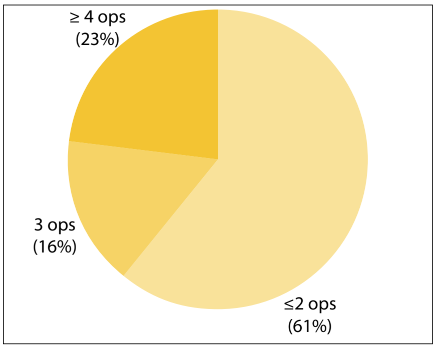

Figure 5.7.2

Number of operations per patient for Grade 2 wounds (N = 5,974).

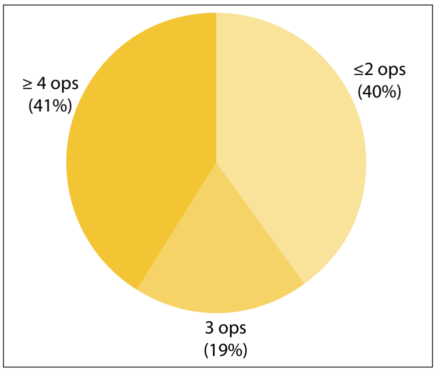

Figure 5.7.3

Number of operations per patient for Grade 3 wounds (N = 3,469).

5.10.5 Number of operations according to Wound Grade and delay in evacuation

If one now analyses the number of operations by combining the RCWS Wound Grade and the delay in evacuation the results are telling (Table 5.21).

Grade and evacuation time	Number of patients	≤ 2 operations	3 operations	≥ 4 operations
Grade 1	N = 6,729			
< 6 hrs	1,124	93 %	5 %	3 %
6 – 24 hrs	1,694	84 %	9 %	7 %
24 – 72 hrs	1,182	82 %	11 %	7 %
> 72 hrs	2,729	82 %	10 %	9 %
Grade 2	N = 5,974			
< 6 hrs	788	77 %	11 %	12 %
6 – 24 hrs	1,186	62 %	16 %	22 %
24 – 72 hrs	1,110	58 %	17 %	25 %
> 72 hrs	2,890	56 %	18 %	26 %
Grade 3	N = 3,469			
< 6 hrs	497	47 %	17 %	35 %
6 – 24 hrs	847	37 %	19 %	44 %
24 – 72 hrs	493	39 %	19 %	42 %
> 72 hrs	1,632	40 %	20 %	40 %

Table 5.21 Number of operations per patient according to Red Cross Wound Grade and time since injury.

It would thus appear that the Red Cross Wound Score better represents morbidity and surgical workload than the delay in evacuation to hospital alone, but a combination of the two is more noteworthy. Grade 3 wounds tend to be very severe and delay does not appear to make a difference. Many such patients will simply not survive to reach hospital.

5.10.6 Number of operations according to weapon

The results of an analysis of the ICRC hospitals treating a variety of weapons-induced injury (Kabul, Kandahar, Kao-i-Dang, Novye Atagi, Peshawar, and Quetta), are presented in Table 5.22.

No distinction is made in the ICRC database between anti-personnel and anti-tank mines, or unexploded ordnance. Furthermore, some wounds classified as being caused by fragments may well have been from fragmentation anti-personnel mines. Hospital staff have only the account of the patient to go by and, naturally, many patients are ignorant of weapons systems; they only know of "bombs" or "guns".

Weapon	None	1 op	2 ops	≤ 2 ops	3 ops	≥ 4 ops
Mines N = 5,236	9 %	15 %	38 %	62 %	14 %	24 %
GSW N = 5,984	9 %	22 %	44 %	75 %	12 %	13 %
Fragment N = 7,674	11 %	24 %	44 %	80 %	9 %	11 %

Table 5.22 Number of operations per patient according to weapon.

It is obvious that mines imply a much greater workload and morbidity than gunshot (GSW) or fragment wounds (Table 5.22 and Figure 5.8).

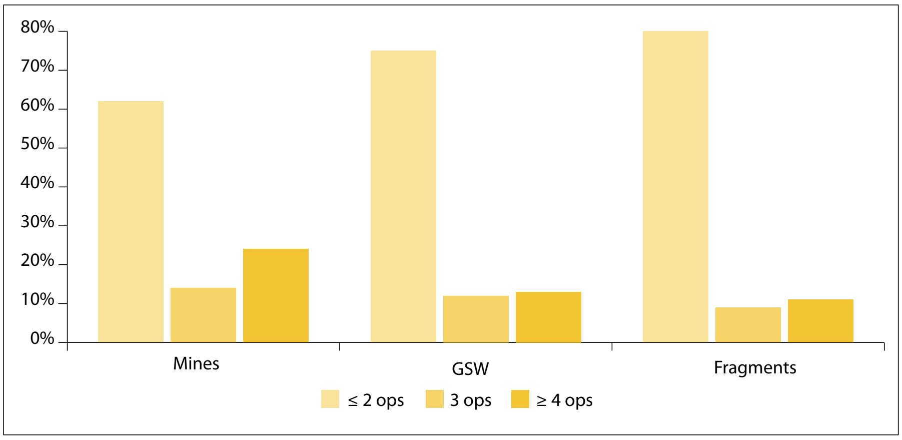

Figure 5.8

Number of operations per patient according to weapon, simplified.

Although the lethality of weapons is important, the total effects of surgical workload, morbidity and suffering, and the socio-economic consequences must also be taken into account. It is not for nothing that poison gas, bacteriological weapons, blinding laser weapons, anti-personnel mines and cluster submunitions have been proscribed and banned by international treaty.

5.11 Conclusions: lessons to be gained from a study of epidemiology

From this brief overview of the epidemiology of the victims of war, a few conclusions affecting clinical work and humanitarian action may be drawn.

1. For a civilian population in a poor country, the public health effects of war are usually greater than the effects of direct trauma. In some conflicts, war injuries carry a greater burden and post-traumatic morbidity may have more profound long-term effects than mortality. This is especially the case in a post-conflict situation with widespread contamination by anti-personnel mines whose socio-economic repercussions are long-lasting.

2. The outcome of war surgery in a civilian context is influenced by:
   - the type of injury according to wounding agent;
   - the severity of the injury;
   - the general condition of the patient (malnutrition, chronic disease, concomitant endemic disease such as malaria, etc.);
   - early and appropriate first aid;
   - the time needed for transport to the hospital;
   - the quality of hospital treatment (resuscitation, surgery, post-operative care, physiotherapy and rehabilitation);
   - possibility of evacuation to a better-equipped hospital with more experienced staff.

3. In a civilian context, and especially in a poor country, pre-hospital care is the phase most open to improvement. Much can be done to prevent death and morbidity with early and efficient first aid and life support.

   An efficient first-aid and evacuation system can prevent deaths from controllable haemorrhage and compromised airway. As pre-hospital care improves, the "killed in action" or "dead on arrival" may diminish slightly, only to see the "dead of wounds" and hospital mortality rise; more casualties are saved, but the rates become distorted. This bias should be taken into account when judging the adequacy of care.

   Long and difficult evacuation of the wounded results in an automatic "natural triage" of the most severely injured. Hospital mortality rates decrease as a result.

4. Up to 40 – 50% of the civilians wounded during urban fighting do not need hospitalization. First-aid measures – plus a simple oral antibiotic and analgesic – are all that are required. Their presence constitutes a supplementary burden on hospital management. However, discharging them from hospital can prove problematic because of security concerns, socio-economic factors and patients' fears.
5. Wounds to the head and torso are the most lethal. Injuries to the soft tissues and the limbs constitute the majority of surgical workload.
6. Gunshot wounds carry a heavier burden than injuries due to fragments; but burns and anti-personnel landmine injuries represent the heaviest hospital workload and morbidity.
7. The Red Cross Wound Score allows a good appreciation of the severity of war wounds, and the surgical workload that they represent.

5.11.1 Setting up a surgical database for the war-wounded

Various military medical services have their own categories and charts for data collection. The ICRC offers civilian health facilities an example of categories that could be placed in a simple spreadsheet (Excel ®) to allow for a collection of relevant data for further study (see Annex 5. B: Setting up a surgical database for the war-wounded).

In addition, a downloadable file form of Annex 5. B (Setting up a surgical database for the war-wounded) is available on the accompanying DVD (Annex 5. C). It can be modified by the user.

ANNEX 5. A ICRC surgical database

The ICRC set up a centralized wound database and trauma registry in 1990, originally designed to give the organization an indication of the surgical workload of its independent hospitals (i.e. hospitals that were established and run under ICRC administration). All war-wounded patients admitted to ICRC hospitals have routinely had a data form filled out on their death or discharge. Patients are not asked whether they are combatants or civilians.

Age and sex, the cause and anatomic site of injury, and the time lapsed between injury and admission are recorded for each patient. Projectile injuries are scored according to the Red Cross Wound Classification System described in Chapter 4. The emphasis is on surgical workload as described by severity of injuries, number of operations per patient, number of blood transfusions required, and length of hospitalization.

These independent ICRC hospitals have served the victims of a number of armed conflicts of very different types of combat. As at 31 December 2007, the database contained the records of 32,285 war-wounded patients. Not all are complete, however.

Hospital	Period of activity	Nature of combat
Kao-i-Dang Hospital, Thailand	1979 – 92	Jungle guerrilla warfare in Cambodia
Peshawar Hospital, Pakistan	1981 – 93	Mountain guerrilla warfare in Afghanistan
Quetta Hospital, Pakistan	1983 – 96	Mountain guerrilla warfare in Afghanistan
Karteh-Seh Hospital, Kabul, Afghanistan	1988 – 92	Irregular warfare, mostly urban
Mirweis Hospital, Kandahar, Afghanistan	1996 – 2001	Irregular warfare, mostly urban
Butare Hospital, Rwanda	1995	Irregular guerrilla warfare, mostly rural
Novye Atagi Hospital, Chechnya, Russian Federation	1996	Irregular guerrilla warfare, rural and urban
Lopiding Hospital, Lokichokio, Kenya	1987 – 2006	Rural guerrilla war in semi-desert, sahel and reed swamps of southern Sudan

The above-mentioned conflicts were different in nature. Time to treatment often included extremes – minutes to weeks. Logistic difficulties over great distances meant that the ICRC could rarely organize efficient and timely evacuation of the victims. Notable exceptions were the first-aid posts set up near the Afghan border with Pakistan and close to Kabul. The air medical evacuation programme for southern Sudan, coordinated with the United Nations Operation Lifeline Sudan, saw over 20,000 patients transferred to the ICRC hospital in Lokichokio, northern Kenya, but the distances and delays in notification of patients nevertheless seldom allowed timely evacuation.

The ICRC has also set up other hospitals, and its surgical teams have worked in local public hospitals, but their patients are not included in the database. In addition to the surgical database, other major sources of data include the Somali Red Crescent-run Keysaney Hospital in Mogadishu (1992 – present), and the JFK Memorial Hospital, Monrovia, Liberia (2001 – 2004), operated jointly by the ICRC and the hospital's Board of Governors.

When compiling ICRC statistics from the database, the limits of a retrospective analysis of prospectively collected data taken exclusively from hospital admissions must be recognized. The usual problems have been observed:

- clerical errors,
- missing patient files,
- incomplete patient files (not all categories were filled in for all patients),
- confusion as to the definition of classification categories,
- lack of continuity due to the constant turnover of hospital personnel,
- lack of properly trained administrative support staff to maintain a database.

Nonetheless, ICRC statistics are offered throughout this manual as an example of non-military experience in different combat zones as an approximation of the reality of the battlefield, especially for non-combatant populations.

ANNEX 5. B Setting up a surgical database for the war-wounded

The following categories can be recorded in an electronic database (Approach®) or listed in the *columns* of a simple spreadsheet (e.g. Excel®) and patients registered in the *rows*. Lessons have been learnt from shortcomings in the ICRC surgical database: this version has been modified accordingly.

Administrative data:

- hospital (if the database contains patients from more than one hospital)
- patient hospital number
- database number
- date of admission
- date of discharge
- number of days hospitalized
- re-admission of patient for the same injury?
- age
- sex

Time since injury:

- hours (or < 6 ; 6 – 12; 12 – 24)
- days (or 24 – 72 hrs ; >72 hrs)
- weeks

Weapon causing injury:

- rifle, gun
- bomb, shell, mortar, grenade
- anti-personnel mine (APM)
- anti-tank mine (ATM)
- unexploded ordnance (UXO), including cluster bombs
- “arme blanche”: bayonet, machete, *panga*

Mechanism of injury:

- gunshot (GSW)
- fragment
- blast
- blunt
- burn

Please note:

A bomb, artillery shell or anti-tank mine may give off penetrating fragments, cause blast or burn injury and, by destroying a vehicle or building, also cause blunt injury. One weapon is registered; several mechanisms of injury are recorded. The same can be said of anti-personnel mines.

Clinical data:

- blood pressure on admission
- number of operations
- number of anaesthesias
- number of units of blood transfused
- outcome: healed, complications, death (including cause)

Anatomic data:

Site of injury: if more than one, each injury is given a number and the appropriate number put in the column for the anatomic region. Only one injury per anatomic region is noted.

- head
- face
- neck
- thorax
- abdomen
- pelvis, buttocks
- back and soft tissue of torso
- upper limb left (to determine bilateral limb injuries)
- upper limb right
- lower limb left
- lower limb right

Red Cross Wound Score:

This should be entered for the two most important wounds, more if so desired. The number 1 wound should correspond to the number 1 anatomic region; wound 2 for number 2 anatomic region.

- Wound 1: Entry
- Wound 1: Exit
- Wound 1: Cavity
- Wound 1: Fracture
- Wound 1: Vital injury
- Wound 1: Metallic fragment
- Wound 1: Grade
- Wound 1: Type
- Wound 2: Entry
- Wound 2: Exit
- Wound 2: Cavity
- Wound 2: Fracture
- Wound 2: Vital injury
- Wound 2: Metallic fragment
- Wound 2: Grade
- Wound 2: Type

Major operation:

- craniotomy
- thoracotomy
- chest tube
- laparotomy
- peripheral vascular repair
- amputation above-elbow left
- amputation above-elbow right
- amputation below-elbow left
- amputation below-elbow right
- amputation above-knee left
- amputation above-knee right
- amputation below-knee left
- amputation below-knee right

Comments:

Please note:

This sample database is available as a downloadable spreadsheet file (Excel ®) on the accompanying DVD.

Chapter 6

THE CHAIN OF CASUALTY CARE

6. THE CHAIN OF CASUALTY CARE

6.1 The links: what kind of care, and where? 131
6.1.1 Protection: IHL 131
6.1.2 Levels and locations of medical care 132

6.2 Surgical hospital treating the war-wounded 133
6.2.1 Hospital A: rural hospital = basic trauma services 133
6.2.2 Hospital B: provincial hospital = advanced trauma services 133
6.2.3 Hospital C: major city hospital = comprehensive trauma services 133
6.2.4 Hospital assessment 133

6.3 Transport 134
6.3.1 Command, control and communications: coordination 134

6.4 Forward projection of resources 135

6.5 The reality: common war scenarios 136

6.6 Conflict preparedness and implementation 137

6.7 The pyramid of ICRC surgical programmes 138
6.7.1 Checklist 139

ANNEX 6. A Initial assessment of a surgical hospital treating the war-wounded 140

ANNEX 6. B Strategic assessment of a conflict scenario 146

ANNEX 6. C Humanitarian intervention for the wounded and sick: typical settings 148

6.1 The links: what kind of care, and where?

Modern armed conflict embraces classical war between conventional armies, urban combat between militias, and isolated and sporadic but fierce guerrilla attacks in remote rural areas. It may involve mass conflict or chronic irregular, low-intensity combat, or individual terrorist attacks. Civilians often represent the majority of victims (see Chapter 5). Field situations vary considerably, but basic medical problems for the wounded are universal. A system must be developed and adapted to deliver the best possible care in a timely manner, under all circumstances.

Modern armed conflict takes place in rural areas and cities.

Civilian populations are increasingly at risk.

Landmines and unexploded cluster munitions continue to cause victims after the conflict has ended.

Field situations vary; medical needs are universal.

The system of evacuation and treatment of the sick and wounded – the chain of casualty care – like any chain, is as strong as its weakest link. It is measured not in kilometres but in hours and days. The setting up of such a system requires planning ahead of time: an assessment of the tactical circumstances must be made; an analysis of the physical limitations and human resources carried out; and the resulting plan implemented.

Wounded patients are transferred along a chain of medical care, beginning with simple “life- and limb-saving” procedures and continuing to ever-greater levels of sophistication. The principle of echelons of care in a military system has been described in Chapter 1. In civilian practice patients also tend to follow a chain of evacuation and treatment although, in many countries, this is not very efficient.

6.1.1 Protection: IHL

Protection and prevention from further injury come first and foremost in patient care.

Article 3 common to the four Geneva Conventions and article 7 of Additional Protocol II:

The wounded and sick shall be collected and cared for.

The wounded and sick have the *right* to be cared for and to have access to appropriate medical care. The red cross, red crescent and red crystal emblems symbolize the legal protection afforded the casualty and the medical personnel, who have the right and the obligation to care for the wounded and sick. The protective quality of these emblems is a function of the training and degree of discipline of the combatant forces, and their adherence to internationally accepted norms of behaviour on the battlefield: the Geneva Conventions and their Additional Protocols. This protection begins with the first aider in the field and continues through all the levels of treatment.

Special measures must be taken to prevent further injury to any casualties and to protect them from the elements (see Chapter 7).

6.1.2 Levels and locations of medical care

The following list defines the generic places where the wounded receive different levels of medical care in the multiple-phase treatment of war wounds.

1. On the spot: first aid.
2. Collection point: first aid with/without resuscitation.
3. Intermediate stage: resuscitation with/without emergency surgery.
4. Surgical hospital: primary surgical treatment.
5. Specialized centre: definitive surgical treatment including reconstructive procedures; physiotherapy and rehabilitation, both physical and psychological.
6. Transport system for transfer from one echelon to another.

Treatment on the spot may be self- or "buddy"-treatment, or practised by a military medic or first aider. First aid starts at the point of wounding, but can be given anywhere and everywhere along the evacuation line to the place of definitive treatment. First aid is the only possible treatment on the battlefield.

The setting up of a collection point or clearing station may be a spontaneous response; alternatively, an established first-aid post (FAP), dispensary, or primary health centre might serve this purpose. Some FAP will be staffed by military stretcher-bearers and medics, others by Red Cross/Red Crescent volunteers, or other civilian personnel. The closer to the battlefield, the more prominent will be the role of military medical services. In purely civilian rural settings, village health workers, nurses or medical assistants are often the only health professionals available. Collecting the wounded at one specific location allows for the proper organization of field triage and their efficient evacuation. Besides basic first aid and trauma life support, resuscitative measures may be begun here.

A health centre or rural hospital might represent the intermediate stage, where more sophisticated resuscitation and emergency surgery are available. The military might establish a forward surgical unit.

B. Fjortoft / ICRC

Figure 6.1

National Society first aiders in the field.

J-P. Kolly / ICRC

Figure 6.2

First-aid post on the Afghan-Pakistani border.

J-J. Kurz / ICRC

Figure 6.3

Transfer of patients to the ICRC hospital in Peshawar, Pakistan.

6.2 Surgical hospital treating the war-wounded

The level of sophistication of hospital care will depend on the degree of socio-economic development within the country *before* conflict breaks out. War will then usually degrade the functioning of any pre-existing hospital. Efficient evacuation of the wounded to deficient surgical facilities is not an effective chain of casualty care.

Levels of hospital competency differ from country to country and between geographic regions. Three basic levels can be described.

M. Bleich / ICRC

Figure 6.4

ICRC field hospital, Lokichokio, Kenya.

6.2.1 Hospital A: rural hospital = basic trauma services

A district hospital (Africa) or primary health centre (South Asia) without full-time specialist doctors, especially without a fully-trained general surgeon, is a typical example. These hospitals are usually staffed by general practitioners or medical assistants with some surgical training and equipped with a minimum of proper surgical facilities. Front-line hospitals with “field surgeons” exist in some military medical services and are the equivalent. Basic resuscitation and some simple but critical operations are usually performed here.

6.2.2 Hospital B: provincial hospital = advanced trauma services

A regional hospital (Africa), district hospital (South Asia), or general hospital (Latin America) have a full-time general surgeon and proper operating theatre and sterilization facilities. Most life-threatening conditions can be dealt with adequately. The military equivalent is usually a forward field hospital specializing in damage control and resuscitative surgery.

6.2.3 Hospital C: major city hospital = comprehensive trauma services

This is a referral structure offering a broad range of specialties and sub-specialties and usually represents the highest level of care in the country or administrative region. It is often a teaching or university hospital. In military terms, this is a referral structure far from the front lines.

6.2.4 Hospital assessment

ICRC programmes aim to maintain, and upgrade if possible, the competencies of these different levels of hospital care when necessary. The ICRC has developed an analytical tool for assessing the quality of hospital management and surgical work prior to providing assistance. This tool comprises a number of factors to be considered. There are factors external to the hospital, involving an analysis of the national and provincial health systems, and internal factors: hospital infrastructure, administration and functional organization, departments, available resources, medical personnel (number and expertise), non-medical support services, finances (see Annex 6. A: Initial assessment of a surgical hospital treating the war-wounded).

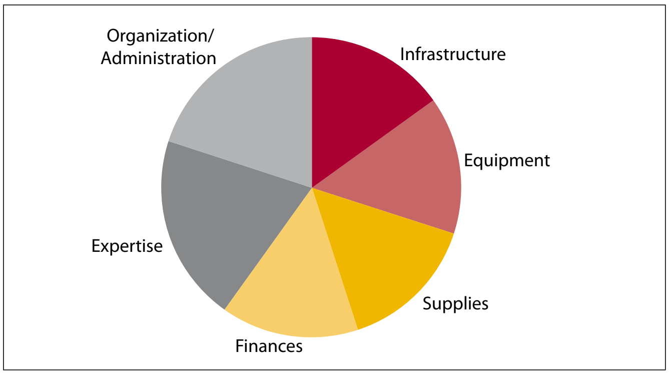

Figure 6.5

Assessment of a surgical hospital caring for the war-wounded.

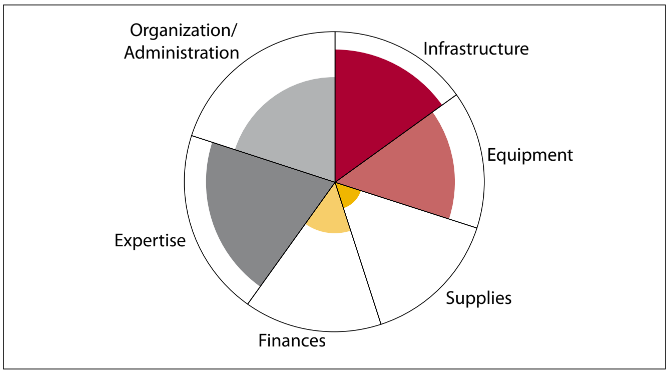

Figure 6.6

Assessment results of a typical hospital in a low-income country disorganized by war.

T. Gassmann / ICRC

Figure 6.7.1

National Society volunteers transporting patients by boat.

L. Petridis / ICRC

Figure 6.7.2

Airplane medical evacuation from southern Sudan to ICRC Lokichokio hospital.

T. Gassmann / ICRC

Figure 6.7.3

National Society "Zero Mobile" horse-drawn carriage ambulance.

The pie-charts above show the various factors that affect the functioning of a hospital dealing with the added burden of war-wounded patients compounded by the constraints of a weakened health system, and help identify dysfunctional areas. An additional factor is the emotionally stressful presence of casualties who are related to, or friends of, the medical staff.

6.3 Transport

Transport of some sort provides the connection between the different echelons of the chain of casualty care. Moving a patient has a price: transportation is a trauma in itself. It uses up additional resources and involves security risks ("mortality of the ambulance ride"), and perhaps even exposure to military activity. These extra costs must be weighed up against the likely benefits of moving the wounded. In many contexts, the availability of transport for the sick and wounded is almost a "luxury".

Moving wounded people is difficult, always takes longer than expected, adds to the trauma, and is often dangerous.

6.3.1 Command, control and communications: coordination

A chain of command is necessary to make the various echelons function correctly. A central command or dispatch centre will assume overall coordination (e.g. decisions about transfer/evacuation destinations, engagement of resources, etc.), and be responsible for contacts with related command levels of various authorities (e.g. armed forces, police, headquarters of the National Red Cross/Red Crescent Society, civil defence and national rescue service, etc.).

Circulation of information between the different levels is assured by some means of telecommunication – radio and mobile phones – if possible, or by other means (e.g. foot messengers) if not. Mobile telephone systems have a tendency to stop functioning – be turned off – during crises and armed conflict. The efficiency of the command and communication systems depends on strict observance of established procedures.

6.4 Forward projection of resources

Bringing more than basic first aid closer to the collection point is called "forward projection of resources". The availability of advanced procedures closer to the battlefield has many advantages. It allows quicker access to "life- and limb-saving" emergency measures, thus decreasing both mortality and morbidity. The need for potentially dangerous transport is reduced. The projection of resources applies particularly to treatment at the intermediate stage, but can be applied to any echelon in the chain.

> Forward projection of resources allows quicker access to life- and limb-saving measures.
>
> However, certain conditions must be met.

A number of factors limit the possibility to project resources.

1. Security (essential).
2. Human resources and expertise (essential).
3. Equipment (appropriate technology).
4. Supplies (appropriate).
5. Infrastructure (minimum requirements).
6. Possibility of onward evacuation.

When a front-line hospital is coming under bombardment, putting both patients and personnel at risk, there is little point in attempting more than first aid if evacuation to another facility is possible. On the other hand, a forward collection point may provide advanced treatment. Only a minimum of equipment is necessary for well-trained medics to perform endotracheal intubation or place a chest tube, and then quickly evacuate the patient. A secure building with adequate resources may be used as an intermediate stage where a forward surgical team may perform damage control and resuscitative surgical procedures.

All of these possibilities will depend on meeting the criteria defined above; but especially on *security and human expertise*. Infrastructure, equipment and supplies must all meet minimum requirements, and be appropriate to the task under the prevailing conditions.

> The most important factors determining the forward projection of medical care for the war-wounded are security and human expertise.

The appropriate choice of procedures to undertake for the injured outside a formal hospital setting will depend on the factors mentioned above, and will vary from country to country, and even from region to region within the same country.

The organization of any chain of casualty care for the war-wounded – military or civilian – should rely on a large dose of common sense to determine what is practical and what can realistically be accomplished to assure the best results for the greatest number, while at the same time guaranteeing the safety of the injured and the health workers. What exactly can be done for the injured outside a formal hospital setting will depend on the particular circumstances and means available. There is no dogma to be followed blindly; situations are different and improvisation and adaptation the key to success.

	On the spot	Collection point	Intermediate stage
Who ?	Family, friends, the community[1] Community health workers First aiders (Red Cross or Red Crescent, military stretcher-bearers and medics, combatants, etc.) Health professionals	Health professionals First aiders (Red Cross or Red Crescent, military stretcher-bearers and medics, combatants, etc.)	General practitioners, emergency room staff, other medical and surgical professionals Field surgical team
Where ?	In a combat zone, at the front lines	Spontaneously chosen (e.g. the shade of a tree) First-aid post Dispensary, primary health care centre	First-aid post, dispensary, primary healthcare centre, outpatient clinic Rural hospital Forward surgical hospital
What ?	Life-saving first-aid measures *The only appropriate care on the spot*	Collection of casualties Evaluation of their condition Advanced emergency care and/or stabilization Evacuation planning Routine care (fever, diarrhoea, scabies, etc.) and ambulatory care (pneumonia, non-combat trauma, etc.)	Advanced emergency care Forward life-saving surgery Occasional hospital care, although uncomplicated and requiring few days of observation Routine care (fever, diarrhoea, scabies, etc.) and ambulatory care (pneumonia, non-combat trauma, etc.)

Table 6.1 What can be done at which level?

6.5 The reality: common war scenarios

As mentioned in Chapter 1, there is more than one "surgery" for the care of the war-wounded. The exact number of levels of care and the path followed by casualties are determined on a case-by-case basis according to the sophistication of care and logistics available. In some armies or countries, the organization may be so efficient that a wounded soldier may expect to receive treatment virtually as sophisticated as the care available to him in peacetime.

In developing countries, however, the healthcare system might already be weak before the conflict and may almost cease to function because of it. Water and electricity supplies can be unreliable, trained staff often flee the area, drugs and disposable equipment cannot be replaced, budgets and salaries are not paid, and buildings are destroyed. The quality of hospital care is badly affected.

Every combination of levels of care is possible:

- the wounded from a modern industrialized army are transferred by helicopter from the scene of injury directly to a surgical hospital;
- patients in low-income countries reach health facilities on foot or by oxcart, donkey, private motor vehicle, taxi or lorry;
- during urban warfare, family and neighbours transport a casualty directly to the emergency room of a surgical hospital, which then serves as a collection point;
- in the rural areas of a low-income country, the hospital is the only health structure functioning and serves as all levels in one – forward referral of patients to a more sophisticated level of care is not available;
- rural health centres may flourish while surgical hospitals may exist only in large cities;
- a collection point or intermediate stage in a secure building is upgraded with the necessary equipment and personnel to serve as a forward surgical hospital, with or without possibility of forward referral.

In some contexts, the ICRC has resorted to deploying a Field Surgical Team that is not hospital based. This mobile team "goes to the wounded" rather than having the wounded come to a hospital; the chain of casualty care is turned upside-down. This has been necessary where casualties have not had access to medical care for reasons of personal security (Figures 1.4 and 6.8).

---

[1] In armed conflicts, under international humanitarian law, civilians are permitted to collect and care for the wounded and sick of whatever nationality, and shall not be penalized for doing so. On the contrary, they must be aided in this work. Furthermore, IHL prescribes that the civilian population shall respect the wounded and sick, even if they belong to the enemy, and shall commit no act of violence against them.

Techniques will also differ according to context. For the military, there is a balance between the needs of the wounded soldier and the necessities of combat. Some armed forces may teach the use of the self-applied tourniquet, which purportedly allows a wounded soldier to continue shooting. This logic does not apply to civilian institutions, and the ICRC's first-aid manual[2] proscribes the use of a tourniquet in the field except in very rare circumstances.

V. Louis / ICRC

Figure 6.8

The ICRC Field Surgical Team in Darfur.

6.6 Conflict preparedness and implementation

Every country should have a disaster plan. Part of emergency preparedness is the capacity to respond to a situation of armed conflict or internal troubles, as well as natural disasters. This is normal procedure for most armed forces. Ministries of public health and National Red Cross/Red Crescent Societies usually also have a disaster plan, which should be integrated into the national emergency preparedness programme.

> The objective of planning is to ensure that wounded people get the right care, at the right place, and at the right time.

Those who have to face the challenges of armed conflict should understand how to proceed in setting up a chain of casualty care. The best possible outcome for wounded people can only be achieved with proper planning and training. Plans must be realistic, flexible and reviewed regularly. Should a country be caught unawares at the outbreak of hostilities without having a pre-prepared plan, then the planning process must be accelerated and the necessary analysis take place immediately.

All plans begin with a *strategic assessment* of the possible conflict scenarios. What may happen and where? What needs are foreseen? What resources exist? (see Annex 6. B: Strategic assessment of a conflict scenario).

An *analysis of the assessment* will then answer the questions of what should be done, where, and by whom to improve the care for the wounded.

The reality check asks a few basic questions. Are the recommendations compatible with the context? Are they relevant? Practical? Reality checks are important; they influence planning, care, and training, and make certain that academic theory does not override simple pragmatism.

The organization of a chain of casualty care and in particular the assignment of resources (physical and human) to the different echelons must take into consideration a number of factors:

- the nature of the conflict, the tactical and geographic circumstances, and security concerns;
- the size of the casualty load;
- the quality of the casualty load, i.e. the types of injuries;
- the relative and proportionate number of human resources (surgical capacity depends on the number of teams working *and* the rate of arrival of casualties);
- the quality of human resources, i.e. expertise;
- logistics and supplies;
- the infrastructure.

---

[2] Giannou C, Bernes E. *First Aid in Armed Conflicts and Other Situations of Violence*. ICRC: Geneva; 2006.

Some typical situations of armed conflict and internal violence can be described. Model scenarios are then based on a consideration of the above-mentioned factors and the analytical tool of strategic assessment (see Annex 6. C: Humanitarian intervention for the wounded and sick: typical settings).

6.7 The pyramid of ICRC surgical programmes

ICRC delegates are often called upon to help organize a chain of casualty care. At times this will be in support of existing governmental or non-governmental structures. On occasion, the ICRC is required or asked to establish its own independently-run hospitals. The main reasons for doing so are either a question of protection – of the patients or the medical mission – or because of a dire shortage of local human resources.

Unlike a military field hospital, an independent ICRC hospital is often responsible for all levels of medical care. It acts as a first-aid post, field hospital, base hospital, and referral centre. In this, an ICRC hospital resembles many provincial hospitals in non-industrialized countries ravaged by war. The level of care is a combination of advanced and comprehensive trauma services. Only general surgeons are present, and no further referral of patients to a specialized centre is possible.

An independent ICRC hospital comprises all echelons in one structure.

In both instances – support to national structures or establishment of an ICRC hospital – a number of factors are taken into consideration to ensure the neutrality and independence of ICRC activities on the one hand, and the quality and professionalism of care on the other. The same logic applies to other humanitarian agencies. These factors are summarized in the following pyramid and checklist (Figure 6.9 and Section 6.7.1).

Figure 6.9

The ICRC surgical pyramid.

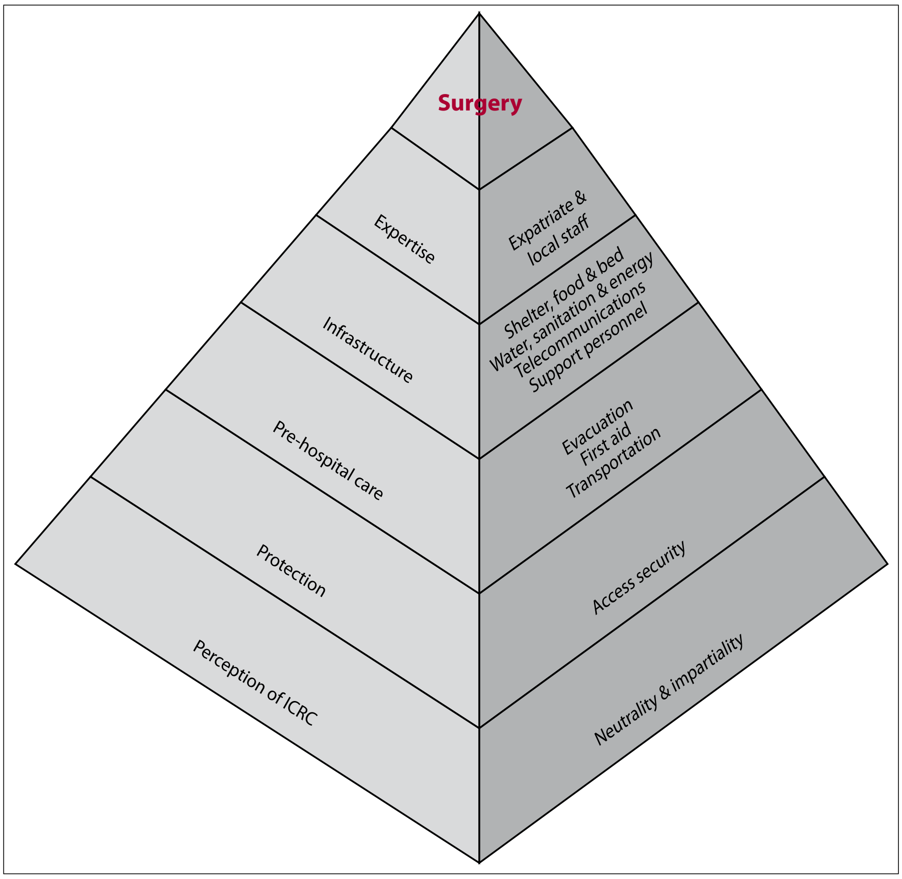
Surgery
Expertise
Expatriate & local staff
Infrastructure
Shelter, food & bed
Water, sanitation & energy
Telecommunications
Support personnel
Pre-hospital care
Evacuation
First aid
Transportation
Protection
Access security
Perception of ICRC
Neutrality & impartiality

6.7.1 Checklist

Political and operational considerations

- Perception of neutrality and impartiality, i.e. ICRC image.
- Acceptability of ICRC by political factions, the military, and the population.
- Possibility for negotiations: availability and access to interlocutors (health, defence, foreign affairs, and interior ministries, faction chiefs); contacts with owners of facilities to be rented.
- Role of the hospital in the protection of hospitalized patients.
- Role of the hospital in protecting the medical mission by providing access to neutral medical care, across enemy lines if necessary.
- Any possible competition with other health structures, private or public, i.e. "medical market".

Security

- Hospital and staff security concerns owing to the conflict or possible developments in the fighting:
  - location and environment, i.e. distance from hostilities or military targets;
  - type of building, i.e. number of storeys, ground floor, underground cellar or bomb shelter, tents.
- Incidence of banditry, hostage-taking, etc.
- Patients' safety, inside the hospital and upon discharge.

Access

- Distance and evacuation time.
- Availability of means of transportation: roads, vehicles, air evacuation.
- Safety of roads and checkpoints: military activity, political affiliation, banditry.
- Possibility of setting up a system of first-aid posts.
- Logistics: medical supply, fuel, food (local, regional, from ICRC headquarters).

Infrastructure

- Pre-existing hospital structure.
- Building capable of being transformed into a hospital (school, factory); structural integrity, possibility of expansion.
- Use of tents, prefabricated and/or temporary structures.
- Water and sanitation, electricity.
- Facilities: kitchen, laundry, residence for personnel.
- Warehouse.

Local and expatriate personnel

- Availability and number of local and expatriate Red Cross/Red Crescent medical personnel.
- Recruitment of new national staff and the question of neutrality.
- Professional competence and level of education.
- Language problems: need for translators.
- Availability of support personnel: ICRC delegates, medical and general administration, water and sanitation engineer, builder, mechanic, electrician, etc.

ANNEX 6. A Initial assessment of a surgical hospital treating the war-wounded

This form should be regarded as a guide, a kind of checklist, to help the health professional responsible for the assessment of the hospital remember certain key points.

Its purpose is to have a quick overview and understanding of the functioning of the hospital; to identify its capacity, limits and insufficiencies. It should allow for rapid and adequate decisions about the kind of *support* the hospital might need.

The form can be used *in toto* or in part according to the objectives of the assessment. It consists of questions deliberately left open to allow the health professional in charge to decide to enquire further or not according to the results expected. It is more important to try to get a correct understanding of how the hospital is working rather than to have exact figures.

The assessment is divided into six sections

General

Management and administration (including non-medical support services)

Medical support services

Clinical services *(only the surgical component is dealt with in this annex)*

Further comments

Conclusion

General

Name of the hospital: Town: Country:

Assessment done by:

Date:

Interlocutors:

1. Type (MoPH, private, military, missionary, NGO, others):
2. Catchment population:
3. Assistance from others than the authority in charge:
4. Level of reference (rural, district, regional):
5. If rural or district hospital, number of primary facilities served (first-aid posts, dispensaries, health centres):
6. Possibilities for further referral:
7. Transport system for patients (in and out):
8. Reputation of the hospital (indicate source of information):
9. Bed capacity, effective number of beds (breakdown by department):
10. Present bed occupancy:
11. Activities including specialities (surgery, medicine, paediatric, obstetric, specialized services, etc.):
12. Security (Is the area safe? Is the hospital secured? i.e. clearly marked, fenced, watchmen present, absence of arms inside the hospital compound?):
13. Endemic diseases and epidemic risk in the region:

Management & administration

I General management

1. Set-up (management team/board):
2. How are decisions taken and implemented?

II Human resources management

1. Who is in charge?
2. Does the staff receive salary/incentives?
3. Total number of personnel/breakdown by function (MD, medical assistants, nurses, students, etc.):
4. Is there a roster system in place in the hospital?

III Financial management

1. Management of finance (Is there a budget? How is the hospital financed?):
2. Is there any cost participation, "cost-recovery system"? Do the destitute have access to care?

IV Statistics

1. Management of statistics and reporting:
2. Are statistics available?
3. Is there an annual report?
4. Are there people specifically in charge of collecting data?

V Infrastructure & utilities (general condition of):

1. Wall and roof:
2. Water (running water, wells, safety of water supply, etc.):
3. Sanitation (type of toilets, etc.):
4. Electricity and/or generator (number of hours per day, fuel supply, etc.):
5. Heating/fans/air-conditioning:
6. Maintenance team (number, composition, etc.). Is there a maintenance schedule?
7. Is there a functioning workshop?

VI Waste disposal

1. Waste management systems (including toxics such as X-ray developer/fixator, etc.):
2. Incinerator (type, condition, etc.):

VII Non-medical support services

1. Kitchen (staff, nutritionist, origin of food, number of meals served per day, special diets, etc.):
2. Laundry (staff, washing by hand, machine, supplies, etc.):
3. Tailor (staff, supplies, etc.):
4. Cleaning and hygiene (system, staff, supplies, etc.):
5. Morgue (infrastructure, management, etc.):

Medical support services

I Pharmacy

1. Pharmacy staff and management:
2. Is there a standard list of medicines?
3. Are stock cards used?
4. Where do the drugs and medical equipment come from (regular supplier, local market, donations, etc.)?
5. Is there a reliable system of communication between the pharmacy and the wards (request forms, delivery forms, etc.)?
6. Did the pharmacy run out of basic drugs last month (penicillin, anti-malaria, paracetamol, ORS)?
7. What are the storage conditions (air-conditioning, refrigerator, etc.)?
8. Is medical equipment regularly maintained and serviced?

II Laboratory

1. Laboratory staff and management:
2. Tests available (haematology, chemistry, parasitology, bacteriology, serology, etc.):
3. Source of supplies:
4. Is there a reliable system of communication between the laboratory and the wards (request and results forms)?
5. Quality of the working relationship between the clinical and laboratory staff:

III Blood transfusion

1. Staff and management:
2. Policy of blood sampling and transfusion: HIV/AIDS policy?
3. Indications for blood transfusion/average number of requests:
4. How are the blood units kept? Is there a functioning refrigerator to store the blood?
5. Testing process and quality control:

IV Imaging (X-ray & ultrasound)

1. Staff and management:
2. Average number of X-rays per day:
3. Type and quality of machine:
4. Are there guidelines for the prescription of X-rays?
5. Is more sophisticated imaging equipment available?

V Other diagnostic services

1. ECG, EEG, etc.:

Clinical services

I Outpatient department (OPD)

1. Role of the OPD (consultation, follow-up of patients, admission, emergency):
2. Are there specialized OPDs?
3. Are there criteria to admit the patient to the OPD?
4. Is there a register with data about all patients seen every day?
5. Average number of cases seen every day (medicine, paediatric, surgery, obstetric, etc.):
6. Personnel in charge (MD, medical assistants, nurses):
7. Is there a clear roster?
8. Opening days and hours:
9. Main pathologies:
10. Access to laboratory and imaging:

II Admission/emergency department

1. Number of beds:
2. Is there a team on duty 24 hours a day; composition of the team?
3. Is there an on-call system in place?
4. Is there an admission book or regular procedure for admitting and registering patients?
5. Is there a regular procedure to send patients to the appropriate wards or to the OT?
6. Number and type of emergencies per day:
7. Are basic supplies and equipment available?

III Operating theatre (OT) and anaesthesia

1. Staff and roster:
2. Hygiene of the OT:
3. Is there an accurate operation book? If yes, number of surgical operations in the last month:
4. What kind of surgery is performed?
5. What kind of instruments and sets are available (laparotomy, caesarean section, debridement, skeletal traction sets, etc.)?
6. Number of operating rooms, tables:
7. Surgical linen (availability and source of provision):
8. Functioning surgical equipment (lamps, suction, diathermy, oxygen, etc.):
9. Source of materials and consumables:

IV Sterilization

1. Staff and roster:
2. Equipment (autoclaves, dry ovens):
3. Protocols in place?

V Anaesthesia

1. Staff and roster (MD and/or anaesthetist nurses):
2. Is a laparotomy performed safely with full muscle relaxation (including endotracheal intubation) by a trained anaesthetist?

3. Common anaesthesia (gas, ketamine, spinal, local):
4. Type of anaesthesia machines:
5. Availability of other equipment (pulse oxymeters, oxygen supply, etc.):

VI Nursing care

1. Is there 24-hour nursing supervision in the wards?
2. Are the patient records complete?
3. Is the nursing handover book used properly?
4. Are drugs administered on time?
5. Is a laparotomy performed safely with the patient supervised (vital signs) for 24 hours post-operatively in a room with light, and where he or she receives intravenous fluids and antibiotics?
6. What do dressings look like (clean, smelly, etc.)?
7. Are bedsores a problem?
8. Are relatives involved in patient care?

VII Frequently asked questions

1. Availability of mosquito nets for all beds:
2. Is there an admission book or regular procedure for admitting and registering patients in the ward? If yes, number of admissions to the ward in the last month?
3. Is there a person in charge of the admissions who controls a system whereby the patients are assessed and then go to the OT or receive treatment?
4. Are new admissions systematically seen by a senior surgeon/MD and within what timeframe?
5. Are there regular rounds in the wards and/or regular meetings to discuss the cases?
6. Are the diagnosis and treatment clearly formulated in the patients' files and the treatment copied onto the patients' charts?

VIII Surgical care

1. Main pathologies present in the wards (fractures, burns, chest, abdomen, etc.):
2. Management of the ward/hygiene:
3. Human resources (number, composition, roster):
4. Infrastructure and beds:
5. Is a laparotomy performed safely – patients seen a few days after operation with healing wound and eating normally?
6. Can five or more laparotomies be performed in 24 hours under good conditions including anaesthesia? If not, why?
7. What type of orthopaedic treatment is present in the surgical wards (POP, skeletal traction, external or internal fixation)?
8. What do the wounds present in the ward look like (clean, dirty, smelly, pus)?

IX Physiotherapy unit

1. Are patients walking on crutches in the wards? If not, why?
2. Management of physiotherapy department:
3. Human resources:

Further comments

1. Particularities of the context:
2. Do you have any requests for the ICRC (logic and rationality of the request)?

Conclusion

1. First general impression (cleanliness and hygiene, staff present, presence of patients):
2. Main positive findings:
3. Main negative findings:
4. Capacity to cope with mass influx of wounded:
5. Emergency/contingency plan:
6. Proposals:
7. Next step:

Please note:

This sample form may be photocopied and is also available as a downloadable document on the accompanying DVD.

ANNEX 6. B Strategic assessment of a conflict scenario

The main appraisals in a strategic assessment of a conflict situation in order to determine some of the factors influencing the chain of casualty care are the following:

1. Geography:
    a. topography of the conflict area
    b. routes of communication and transportation
    c. distribution of medical facilities available and their safety
2. Where is the fighting taking place? Safe areas, dangerous areas?
3. Where do patients come from?
4. How many wounded are there?
5. Who are the wounded?
    a. trained soldiers in a regular army
    b. guerrilla fighters, militiamen
    c. civilians
6. Who performs first aid, if anyone?
    a. National Red Cross/Red Crescent Society volunteers
    b. military services
    c. trained civilians (RC/RC community-based first aid, MoPH)
    d. untrained civilians
    e. non-governmental organizations
7. Assessment of efficiency of first-aid system
8. How are the wounded transported from point of wounding to hospital?
    a. private means
    b. public transportation
    c. ambulance service
    d. military services: air, land, etc.
9. Assessment of efficiency of evacuation system
10. Which hospitals receive the wounded ?
11. Assessment of work performed in hospitals and their capacity to receive and treat patients (see Annex 6. A: Initial assessment of a surgical hospital treating the war-wounded).

12. Which other actors are present in the area?
    a. National Red Cross/Red Crescent Society
    b. ministry of public health
    c. military medical services
    d. national non-governmental organizations
    e. missionary clinics or hospitals
    f. foreign non-governmental organizations

On the basis of the above assessment, the military or civilian health services or National Red Cross/Red Crescent Society should take action to fill in any gaps by creating the following facilities.

1. System of first-aid posts.

2. Transport system.

3. Reliable surgical units: central, regional, zonal, local; alternative sites for hospitals; convalescence houses.

The above-mentioned institutions have several means of intervention to face the challenge of establishing an efficient chain of casualty care.

1. Negotiation with the various belligerents to make sure international humanitarian law is respected in order to ensure that:
    a. medical personnel has access to the wounded and sick;
    b. the wounded and sick can reach medical care;
    c. first-aid and health personnel and structures are safe.

2. Support to existing health structures, which may take the form of infrastructure renovation, equipment, medical supplies, or re-enforcement with human resources.

3. Mobilization of local infrastructure and human resources to improve the chain of casualty care or project forward medical care for the wounded.

4. Mobilization of international agencies to supplement national efforts.

Please note:

Points 1 – 12 of this sample form are available as a downloadable document on the accompanying DVD.

ANNEX 6. C Humanitarian intervention for the wounded and sick: typical settings

Many factors affect the deployment of humanitarian medical teams: this Annex helps to analyse a number of them. The terminology used is an operational description and has no legal significance.

Possible scenarios

1. Military activity, natural disaster or major accident? (Is health infrastructure intact?)
2. Military context: classical front lines, guerrilla war, internal troubles, post-conflict (particularly the presence of landmines, cluster bombs and other unexploded ordnance)
3. Urban or rural setting?
4. Industrially-developed or low-income country: funds available?
5. Human resources: trained doctors, nurses, and first aiders in sufficient, limited numbers, or very few?

Three typical general settings for a humanitarian intervention exist.

1. **Optimal situation:**
   adequate access to medical care in spite of the violence and conflict.
2. **Austere situation:**
   poverty before the conflict has already jeopardized access to care.
3. **Dire situation:**
   very poor access to care because of the violence and conflict, sometimes compounded with pre-existing poverty.

	OPTIMAL	AUSTERE	DIRE
Location	Urban in a developed country	Poor rural area	Developing country Major destruction
Duration of trouble	Single, isolated event (e.g. act of terrorism)	Ongoing low-intensity fighting (e.g. guerrilla warfare)	Continuous heavy fighting and/or bombardment
Casualty flow	Small irregular numbers (compared to the population of the city)	Discontinuous/ intermittent flow, including massive influx	Continuous but unpredictable flow, including massive influx
Infrastructure (roads, ambulance service, health facilities)	Intact and functioning	Poor or irregular (e.g. few good roads, hospitals lacking supplies)	Severely damaged or dysfunctional (roads damaged, debris in streets, hospitals looted, etc.)
Communications	Good	Poor to moderate Irregular	None or poor
Personnel	Adequate (in number and skills)	Variable	Minimum available or complete lack
Materials and supplies	Adequate (in quantity and quality)	Irregular and inadequate	Irregular supply or non-existent
Environment	Good (daytime, good weather)	Harsh	Bleak (night, cold, heat, etc.)
Evacuation	Safe and short	Predictable but long and arduous	Uncertain or unknown
Destination of transfers	Known and available	Known but variable	Unknown or absent

Types of situations of armed conflict and violence and their effects on humanitarian medical work[3]

Example	International armed conflict	Internal armed conflict/ guerrilla warfare	Civil disturbance/revolt	Widespread banditry and other crime
Description	A straightforward war between country X, and its allies, and country Y, and its allies	Intense fighting within one country	Unpredictable; hit and run skirmishes, often with a vested interest in the continuation of anarchy	May coincide with any of the other situations
Combatants/ Fighters	Easily visible, and wearing distinctive uniforms	Not all wear uniforms Government forces opposing well-organized military groups	Armed individuals, gangs, bandits, and militia groups	Purely self-interested individuals or groups
Front lines	Well known	May not exist, or may change very quickly	Linked to constantly shifting alliances among or between forces	At the fringes of the conflict, combatants ready to take advantage of circumstances
Chain of command	Structured and with available points of contact	Tenuous points of contact on the opposing sides.	Not clear, and variable from one faction to another (often based on an individual leader surrounded by a small group and supported by part of the population)	Traditional and local leader, personalized (e.g. street gangs)
Respect for IHL	Parties aware of their obligations and try to meet them	A certain degree of respect	Very little, with a complete breakdown of law and order	IHL unknown or no regard for, or both
Humanitarian tasks	Classic	Classic	Extremely difficult	?
Risk level	Low	Growing and less predictable	Very high, perhaps bordering on the unacceptable	Very real and very dangerous threat
Obstacles to medical work	Few, if any	More restrictions, negotiations, controls and delays, etc.	Severely restricted ability to intervene – vehicles, radios, goods, etc. are extremely attractive to the warring parties	Numerous: greatest caution required

Model scenarios

In contemporary armed conflict, health staff and structures are called upon to function in various scenarios: typical ones include the following.

1. Safe urban setting

- Urban, developed environment
- Single, isolated event
- Casualty numbers relatively small compared to population of city
- Infrastructure intact: roads, emergency vehicles
- Health infrastructure intact, sophisticated hospitals
- Short evacuation time: route is secure
- Good communications
- Personnel: adequate number and quality of trained health staff
- Materials adequate
- Environment good: weather, daytime

---

[3] The terminology used is an operational description, not a legal one.

- Final destination of the wounded is known

2. Unsafe urban setting

- Low-income country: under-developed or destroyed urban setting
- Continuing danger: street fighting and bombardment in city
- Continuing and unpredictable casualty flow including massive influx of wounded
- Poor infrastructure: potholed roads, debris in streets
- Disrupted health infrastructure: hospitals damaged or looted
- Availability and length of evacuation uncertain or unknown
- No or poor communications
- Minimum health personnel available
- Material re-supply uncertain, irregular, or non-existent
- Environment poor: cold, wet, dark
- Final destination of the wounded not always obvious

3. Unsafe rural setting

- Low-income country: under-developed rural area neglected in peacetime
- Constant danger: ongoing combat, landmines
- Continuing and unpredictable casualty flow
- Poor infrastructure: badly maintained or no roads
- Poor health infrastructure: few health posts, even fewer district hospitals
- Availability and length of evacuation uncertain, long and arduous
- No or poor communications
- Minimum health personnel available
- Material re-supply uncertain, irregular, or non-existent
- Environment poor: extreme cold or heat, rainy season and dry season
- Final destination of the wounded not always obvious

4. Safe but austere setting

- Low-income country
- Continuing danger: ongoing low-intensity warfare
- Discontinuous casualty flow; includes irregular mass evacuations
- Poor infrastructure: few good roads and few vehicles
- Minimum of health infrastructure: some rural clinics or health centres, fewer district hospitals
- Evacuation predictable but long and arduous
- Poor to moderate communications
- Minimum to moderate number of health personnel available
- Minimum material re-supply
- Environment harsh
- Final destination of the wounded: distant, but known

Chapter 7

FIRST AID
IN ARMED CONFLICT

7. FIRST AID IN ARMED CONFLICT

7.1 First aid: its crucial importance 155

7.2 First aid in the chain of casualty care 156

7.3 First aiders: an important human resource 157

7.4 Essential elements of the first-aid approach and techniques 157
7.4.1 Security first and always 157
7.4.2 Basic actions 158

7.5 Setting up a first-aid post 159
7.5.1 Location 159
7.5.2 Infrastructure 159
7.5.3 Staffing, equipment and supplies 160
7.5.4 Organization 160

7.6 ICRC involvement in first-aid programmes 161

7.7 Debates, controversies and misunderstandings 161
7.7.1 "Scoop-and-run" *versus* "stay-and-treat" 161
7.7.2 Mechanism of injury and the problem of the cervical spine 162
7.7.3 ABCDE or C-ABCDE? 163
7.7.4 The tourniquet: when and how? 163
7.7.5 Resuscitation or rehydration? 164
7.7.6 Supplemental oxygen in the field? 164
7.7.7 Disability: the AVPU system 165
7.7.8 Evacuation: a risk to take 165

7.1 First aid: its crucial importance

Optimal management of the injured involves a continuum of care from the point of wounding to the surgical hospital: the chain of casualty care. The surgeon wishes to receive patients who arrive in good condition and well-stabilized, in a timely manner, and according to priority for treatment. To understand how this is achieved, or not, there are a number of things that need to be known about first aid. In addition, the surgeon working in a conflict area may well be called upon to participate in the training of first aiders to achieve a more efficient chain of casualty care.

First aid is the initial assistance given to an injured or sick person until the condition of the person is stabilized or remedied, or professional medical help is made available. How first aid is applied varies according to the security environment, the number and condition of the wounded in a particular location, the resources that can be mobilized for care, transport, and access to surgical hospitals and the latter's capacity to receive and treat patients.

Since its foundation, the International Red Cross and Red Crescent Movement has been, and continues to be, predominantly associated with the provision of first aid. It pioneered the concept of immediate response by first aiders and local communities to the consequences of war, disasters and epidemics.

The goals of a first aider in action are:
- to preserve life through the support of vital functions;
- to intervene securely and safely;
- to limit the effect of injury and to prevent further injury;
- to prevent complications and disability;
- to alleviate suffering and to provide psychological/moral support;
- to promote recovery;
- to ensure proper handover of the injured and sick to healthcare professionals when needed.

First aiders can also help mobilize their community to prepare for and respond to emergencies that happen in everyday life and during crises such as armed conflict.

Experience has shown that one of the most important factors determining the outcome of the management of the war-wounded is the pre-hospital phase. First aid saves lives and decreases morbidity. It also results in easier surgery, thus decreasing the hospital surgical burden. Furthermore, up to 40 – 60 % of the civilians and soldiers wounded during armed conflict do not require hospitalization. First aid measures – plus a simple oral antibiotic and analgesic – constitute all the treatment they need. Military terminology calls them "returned to duty" (see Chapter 5).

First aid, begun early, saves lives and can prevent many complications and much disability.

The provision of first aid is one of the fundamental responsibilities of military medical services, National Red Cross and Red Crescent Societies and, increasingly in contemporary armed conflict, of medical staff in both urban and rural public hospitals. The essential role of local communities in providing assistance on the spot should not be forgotten, as Henry Dunant, founder of the Red Cross and inspiration behind the Geneva Conventions, witnessed after the battle of Solferino on 24 June 1859.[1]

---

[1] See Henry Dunant. *A Memory of Solferino*. ICRC: Geneva; 1986.

Consequently, basic first aid, initial and refresher training, should be supported among:
- the general population;
- soldiers and members of security forces;
- community healthcare workers;
- military and other medical staff (nurses, doctors and surgeons).

Commanders do not appreciate seeing their troop levels decrease when healthy uninjured soldiers are involved in transferring their own wounded because first-aid services are inadequate in the field. The fighting capacity is then doubly reduced.

Advanced competencies can be added to the training curriculum of those specially dedicated to providing first-aid services in the field, such as military medics and the first-aid teams of the National Red Cross and Red Crescent Societies, as a forward projection of resources.

7.2 First aid in the chain of casualty care

First aid starts at the point of wounding, but can be given anywhere and everywhere along the chain of casualty care to the place of definitive treatment.

Point of wounding
On-the-spot first aid, performed in the field and often on the actual battlefield: this may be self- or "buddy-" first aid if combatants have received the proper training prior to deployment. Otherwise, it is practised by a military medic, civilian or Red Cross/Red Crescent first aider.

Collection point
It is common practice and convenient to bring all the wounded to one spot, depending on the tactical situation, to evaluate their condition, start first aid if it has not yet been given, and stabilize those for whom lifesaving measures have already been undertaken; and then to decide who needs to be evacuated for further treatment according to triage priorities. This is best served by a first-aid post (FAP).

Evacuation
The decision to transport a casualty should be assessed carefully considering the dangers and difficulties inherent in situations of armed conflict. Whatever method of transport is used along the chain of casualty care, first-aid measures should be maintained throughout.

Hospital emergency room
In the rural areas of a poor country and during urban warfare, the first site where any care is available is often the emergency reception of an established hospital. Even if there is an efficient emergency transport service, family and neighbours often do not wait for it and prefer to transport the wounded directly to the hospital, whose emergency room (ER) then serves as a first-aid post.

The degree of development and sophistication of the emergency transport and emergency medical services in a given country will determine what level of first aid and triage are performed "in the field" and what occurs only in a healthcare facility.

First aid can be performed everywhere and anywhere along the chain of casualty care.

Even in armed conflicts, ordinary life goes on. There is no let-up in road traffic crashes, accidents, or diseases among the local population, as well as combatants. The efficient work of first aiders is as necessary as usual.

7.3 First aiders: an important human resource

Usually first aiders are organized in teams with proper supervision and equipment, such as those of the Red Cross or Red Crescent or of the military. Civilian and Red Cross/ Red Crescent first aiders are an essential part of the healthcare team given that they are members of the local community, reflect its characteristics, and are well accepted by society. They take on many roles from the front lines to the healthcare facility; their availability and versatility are respected.

Therefore, it is important to respect their knowledge, and appreciate their courage and dedication. They have important rights and duties according to international humanitarian law, and should be trained accordingly. They should also be trained in the methods and practice of triage. Not only is it important to welcome them when they bring in a casualty but also to give them feedback on the measures they have taken and on the evolution of the casualties they have already brought in, to ensure a proper handover of the patient and prepare for future work. The surgeon plays a vital role in this communication with first aiders.

First aiders are often volunteers and an essential part of the healthcare team, and should be appreciated accordingly.

The presence of first aiders before, during and after emergencies helps rekindle the humanitarian spirit of individuals and communities, inspiring tolerance and ultimately building healthier and safer living environments.

7.4 Essential elements of the first-aid approach and techniques

7.4.1 Security first and always

One of the specific characteristics of armed conflict, as mentioned, is the particularly hazardous situations encountered, more dangerous than during natural catastrophes or industrial accidents. The situation is made even worse because of the will of combatants to continue fighting and inflict harm after the initial damage has been done, and the increasing unwillingness of many combatants to recognize and obey the rules of armed conflict.

First aiders are exposed and at great risk, or are at times even targeted in an ambush or a gun battle. All too often, when a bomb explodes or is dropped there is a rush of first-aid rescuers to the site; a second bomb goes off a short time later causing far more casualties than the first. Furthermore, the work space may be limited by an excited and angry crowd of bystanders and the casualties' friends, relatives and comrades, who may threaten first aiders.

The point of wounding is, by definition, a very dangerous and chaotic place.

This entails practical consequences for the first aider and imposes necessary constraints for the safety of patient and first aider alike. A wounded first aider needs help and cannot help others. The question of first-aid care under fire – for military personnel – will depend on the doctrine and training of each army.

As previously stated, the sick and wounded benefit from protection and the first aider has rights and responsibilities according to international humanitarian law. An opportunity to provide care more safely may be created by negotiations or a ceasefire. The military may choose to remove the threat or provide security by armed force. This does not mean first aiders need not take security measures prior to engagement in the field. Circumstances determine how and when care can be provided with acceptable risk.

The benefits of caring for the wounded must always be weighed against the risk of first aiders themselves being wounded.

An absence or limitation of transportation and the impossibility of deploying personnel in the field to perform first aid because of political or security constraints – all too often the case in the circumstances where the ICRC works – will cause some patients who might otherwise have survived to die, and surviving patients to have established infections and other complications which lower their subsequent chances of quality survival. The challenge is how to deploy and organize the necessary first-aid personnel, material, and equipment in the field. For the military, this is also a challenge but has a different connotation. Tactical considerations may mean that soldiers must "win" the firefight before treating or evacuating the wounded.

7.4.2 Basic actions

The basics of a routine first-aid deployment apply, while at the same time safety and security *must be* ensured.

Preventing further injury: removing casualties from danger

A wounded person is more likely to be injured again or even killed, especially if the person cannot take measures of self-protection, such as seeking cover from shooting or bombing. The prevention of further injury thus involves removing casualties from the point of wounding, and putting patients and medical personnel in as safe a place as circumstances allow. Furthermore, conscious and walking wounded must be separated from the others as a matter of scene security management, as well as triage.

Preventing aggravation of the injury: sheltering the casualties

Shelter offers some protection from further injury in a combat environment, and also provides a shield from the elements. Exposure to the sun and heat, or cold and rain, are deleterious to the wellbeing and stabilization of the injured. It is also easier and more efficient to deliver medical care when working under more comfortable physical conditions.

Performing the standard sequence of basic life support

1. ABCDE sequence leading to immediate life-saving measures.
2. Complete examination leading to stabilization measures.
3. Additional measures:
   - keeping the casualty warm;
   - ensuring proper rehydration;
   - providing psychological support.
4. Monitoring of the casualty's condition and the efficiency of the measures taken.

A = Airway

B = Breathing

C = Circulation

D = Disability (neurological status)

E = Environment and Exposure

Protection of the casualty, prevention of further injury, and lifesaving and stabilization measures are the fundamental responsibilities of the first aider in the field.

The analysis of the trimodal distribution of deaths in Chapter 5 demonstrated that first-aid measures in the first instance should specifically target those injuries that are severe but potentially survivable. Thus, the primary objectives of immediate life-saving measures are the following.

1. Establish and maintain an adequate open airway.
2. Maintain breathing.
3. Maintain circulation by controlling peripheral haemorrhage and reducing or preventing shock.

In the second instance, first-aid measures target the great majority of casualties: those who suffer fractures and soft-tissue injuries, which may cause disability.

The patient must be exposed in order to perform a correct and complete examination. Cultural and religious constraints, as well as the tactical situation, may impose limits as to how much a casualty may be undressed in the field.

All injured persons who have bled lose body heat, even in the tropics. It is important to avoid hypothermia, which can have disastrous results later in terms of coagulopathy (see Chapter 18). Casualties should be covered with a simple sheet or blanket, depending on the prevailing climate, without forgetting to put something *under* the patient since much body heat is dissipated into the ground.

National and institutional guidelines will establish the protocols for the level of treatment (i.v. fluids, antibiotics, analgesia, endotracheal intubation, etc.) that are implemented in any first-aid programme.

Simple life-saving procedures in the field – immediate life-saving measures – take precedence over more complex techniques and are based on the well-known priorities of the ABCDE system.

7.5 Setting up a first-aid post

The establishment and organization of a first-aid post should rely on a large dose of common sense to determine what is practical and realistic in a given situation, and will depend on how long it will operate (from a few minutes to a few days or weeks).

7.5.1 Location

Choosing a site for an FAP should follow certain rules. It should be placed in a secure position, far enough from the fighting not to be in danger, yet near enough to enable the rapid transfer of casualties to the post. For operational and security reasons, its location should be indicated as soon as possible to the dispatch or command centre of the chain of casualty care. Its presence should be made known to the local population and combatants, who are the potential beneficiaries of its services. A prominently displayed distinctive emblem (a red cross, red crescent or red crystal) large enough to be seen from all directions and from as far as possible helps to identify the FAP and provide protection under IHL.

Security and protection of casualties and first aiders are the primary considerations for setting up a first-aid post.

7.5.2 Infrastructure

An FAP is a functional unit: it can therefore be set up in a makeshift fashion under a tent, in a school, or any available building; or in an already existing dispensary or primary healthcare centre. Certain minimum requirements should be fulfilled: appropriate shelter against the elements; adequate size to accommodate casualties on stretchers; easy access for the "walking wounded" (e.g. avoid long stairways); and adequate in-and-out access for ambulances, and ample parking space.

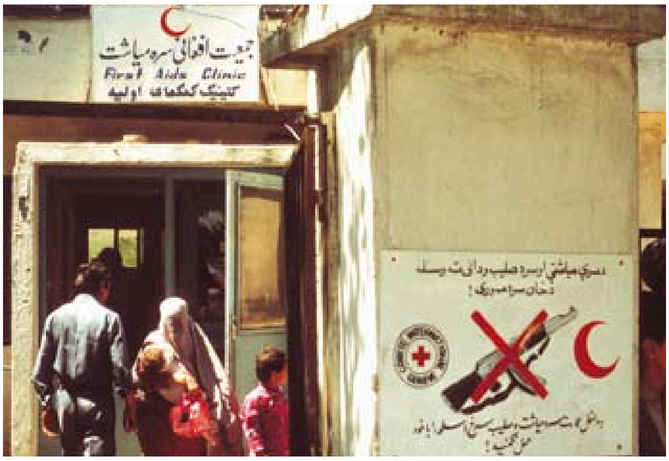

ICRC

Figure 7.1.1

First-aid post: formal.

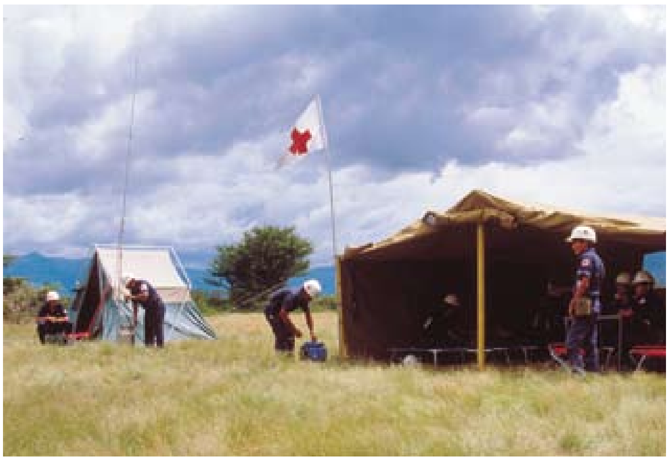

T. Gassmann / ICRC

Figure 7.1.2

First-aid post: ad hoc.

7.5.3 Staffing, equipment and supplies

The level of technical expertise of the personnel in an FAP will depend on the circumstances and standards of the country. Anyone from a first aider to a nurse, general practitioner or even a surgeon may be found working in an FAP. This allows for the "forward projection" of care for casualties (see Chapter 6). Equipment and supplies must meet minimum standards and be adequate for basic trauma care, and adapted to the personnel's skills and local standards. Telecommunication equipment, if any, should be properly installed and placed in a reserved area.

Please note:

The Emergency Items Catalogue of the International Red Cross and Red Crescent Movement contains the description of a standard first-aid and triage post, equipped for experienced nursing staff and/or a general practitioner (see Selected bibliography).

7.5.4 Organization

The premises of the FAP should be organized and the staff prepared for dealing with a mass influx of casualties. The principles and practice of triage should be well understood by all present through simulation exercises (see Chapter 9).

If the FAP stays open for a certain length of time, and the facilities allow for it, the following areas should be pre-arranged:
- admission area at the entrance to register and triage casualties;
- holding area to care for and monitor casualties awaiting evacuation;
- waiting area for relatives, friends and comrades;
- temporary mortuary;
- storage area for equipment and supplies;
- storage area for weapons removed from the wounded;
- rest area for personnel, and personal hygiene facilities.

ICRC EXPERIENCE

Occasionally, the ICRC has had to set up an ad hoc FAP where access to surgical facilities was insecure. During the long-running civil war in Sierra Leone, fighting broke out several times in the capital city, Freetown. During one such episode in 1998, the ICRC delegation was turned into a first-aid post: the parking lot became the reception and triage department and the various offices became wards and intensive care areas. Over a period of 10 days, 6 nurses and 3 Red Cross volunteers treated 244 war casualties and 228 sick patients, while a surgical team from Médecins sans Frontières (MSF) operated in the general hospital. Urban fighting often made it dangerous for patients to reach the hospital and they regularly stayed in the improvised FAP for 2 or 3 days prior to evacuation.

In addition to basic first-aid treatment, intravenous fluid resuscitation, tetanus immunoglobulin, penicillin, and analgesia were given. Many patients survived 2 - 4 days, before being transferred to the hospital, and surgery. The MSF surgical team appreciated receiving patients who had dressed wounds and were "well hydrated", ready for surgery.

Security, however, was a major concern. Looting of the ICRC delegation by armed men occurred twice and several of the 4,500 civilians, who had taken refuge in the delegation premises, effectively turning it into a displaced persons camp, were injured or killed by stray bullets. The emblem of the red cross offers only limited protection under such circumstances.

7.6 ICRC involvement in first-aid programmes

The ICRC supports first-aid programmes and activities in contexts of armed conflict and other situations of violence. These programmes include:
- deployment of medical staff to operate first-aid posts;
- training where no other organization has access (e.g. for irregular guerrilla groups);
- helping to adapt routine first-aid programmes of National Red Cross and Red Crescent Societies and other organizations to the context of conflict preparedness and response;
- development of strategic, management and planning capacities of national and local institutions and team leaders;
- provision of materials and equipment;
- assistance in the organization of an emergency care system;
- production of standards and reference documents (see Selected bibliography).

7.7 Debates, controversies and misunderstandings

Detailed first-aid techniques are described in the manuals listed in the Selected bibliography. The rest of this chapter deals only with a few topics that may give rise to controversy or misunderstanding or are specific to the scenario of armed conflict.

7.7.1 "Scoop-and-run" *versus* "stay-and-treat"

Like the rest of the chain of casualty care, the organization of first aid requires planning and training. A strategy for the efficient care and evacuation of the wounded must be established. Two well-known strategies are described by the expressions "scoop-and-run" and "stay-and-treat". These derive largely from the civilian practice of emergency medical services and may not be particularly relevant in a context of armed conflict.

In a peacetime environment, there is no impediment to patient transfer other than infrastructure and geography. The efficiency of evacuation is determined by the availability and coordination of transport, either by air (fixed wing with airports or helicopters and helipads), or by land (proper roads, necessary vehicles, etc.); and the distances involved. The context of armed conflict is not as simple.

Scoop-and-run implies basic first-aid measures, then taking the patient *immediately* to a surgical hospital. This is appropriate if the distances are short, reliable transport is available, and the nearby hospital is adequately supplied and staffed. Keeping patients on the spot until patient stabilization prior to evacuation – stay and treat – is correct if distances are great, infrastructure and transport are less than adequate, and proper surgical facilities are more difficult to access.

In a military context, the factors of ongoing combat and enemy activity must be added to those present during peacetime. Here, scoop-and-run implies air superiority if helicopters are used, safe and controlled roads for transport over land, and, in both situations, surgical facilities at a reasonable distance. These conditions are fulfilled for only a few armies in the world, and even then only under certain tactical situations. More common is the establishment of a collection point or clearing station: that is to say "stay-and-treat". How much "treatment" occurs depends on the forward projection of resources. The performance of more advanced techniques in the field has been discussed in Chapter 6.

7.7.2 Mechanism of injury and the problem of the cervical spine

The first aider, or other medical personnel, must immediately determine whether the casualty has suffered blunt or penetrating trauma. Blunt trauma above the level of the clavicles or causing unconsciousness requires immediate observation and immobilization of the cervical spine, while still giving priority to the airway. A penetrating wound to the brain causing unconsciousness requires no special care of the cervical spine.

In the case of a penetrating missile wound to the neck hitting the cervical vertebrae, any damage to the spinal cord has already been done. Unstable cervical spine injury from a projectile results in permanent cord injury. The reported mortality is over 95 %. The first aider should be careful while handling the casualty, but cannot prevent damage that is already irreversible. A combination of blunt and penetrating neck injuries should be managed in the same way as blunt ones.

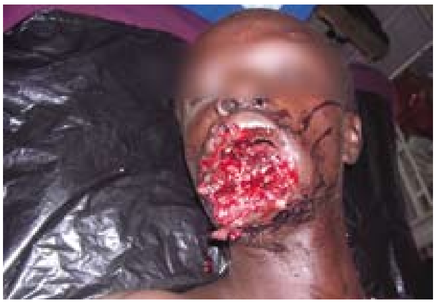

M. Dalla Torre / ICRC

Figure 7.2

Gunshot wound to the mandible: no danger to the cervical spine.

ICRC EXPERIENCE

Here are two practical examples taken from ICRC experience to demonstrate the difference between blunt and penetrating trauma:

A person wounded in a road-traffic crash with a fractured mandible and bleeding from the mouth that obstructs the airway also requires care of the cervical spine in case there is hidden injury. A bullet wound to the mandible – with bone fragments, haematoma, drooling saliva, and damaged soft tissue of the floor of the mouth – does not. If the patient is conscious, he will maintain a particular position – head thrust forward and looking down, blood and saliva flowing out – in order to maintain as patent an airway as possible. There is no use in placing a protective cervical collar on this patient and, if the first aider attempts to place the casualty in a reclining position, the wounded person will resist.

A fall may result in both loss of consciousness and injury to the cervical spine. Loss of consciousness due to a bullet wound to the head, however, does not require any specific precautions regarding the cervical spine.

7.7.3 ABCDE or C-ABCDE?

In statistical terms, severe peripheral haemorrhage is the main lethal yet salvageable injury in a context of armed conflict (see Chapter 5). Various studies show that far more soldiers die from catastrophic peripheral haemorrhage than from airway obstruction or respiratory inadequacy. Thus, the "C" which stands for Catastrophic bleeding, may be put first, simply because it is more common.

Once catastrophic external haemorrhage, which is quite obvious, has been excluded – the great majority of cases – the first aider should resort immediately to the standard routine of ABCDE.

7.7.4 The tourniquet: when and how?

The tourniquet is said to save lives, but endanger limbs. It is difficult to apply correctly and to maintain and supervise, and open to many complications if widely used when not necessary. Furthermore, the prolonged application of an improvised tourniquet is usually not effective and causes tissue damage by itself (Figures 7.3.1 and 7.3.2). There is little or no use for a tourniquet in civilian practice for the war-wounded. It may be applied *as a last resort*, when all else has failed, to temporarily control bleeding while packing the wound and applying a compressive bandage. It should then be removed.

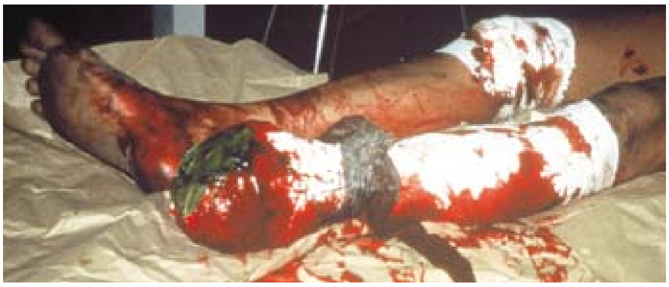

E. Winiger / ICRC

Figure 7.3.1

Improvised tourniquet: how effective?

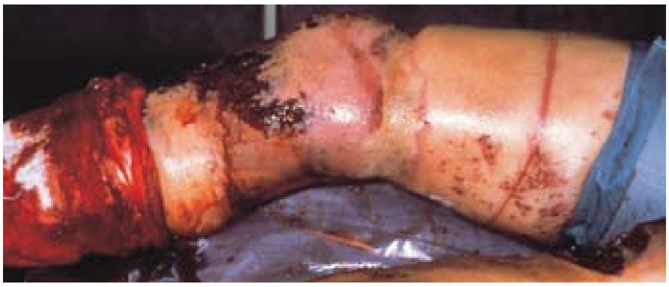

R. Coupland / ICRC

Figure 7.3.2

Improvised tourniquet: how harmful? Here a tourniquet was applied for more than six hours, resulting in a very high above the knee amputation.

By far the most common and most effective technique for a first aider to control peripheral haemorrhage is direct pressure over the wound completed by a compressive bandage; or, if insufficient, pressure over the proximal artery in the axilla or groin, tamponade of the wound with dressings, and then a compressive bandage.

The military might continue to use a tourniquet under certain tactical circumstances: to free the hands of the single medic when dealing with mass casualties; while performing first aid under direct fire, and especially at night. Some armies equip their soldiers with a pre-manufactured self-applicable tourniquet applied with one hand to allow a wounded soldier to continue participating in combat. The ICRC declines to pronounce on these tactical considerations, but warns against an all too easy abuse of the tourniquet as has been seen in many wars in the past.

The use of a tourniquet for crush injuries (people trapped under the rubble of a collapsed building whether due to an earthquake or bombardment) is justifiable, and due precautions should be taken to ensure rehydration/resuscitation and thereby avoid the complications of the otherwise inevitable myoglobinaemia.

For further information on crush injury, see Volume 2.

New locally-active haemostatic powders and sponges to promote clotting in traumatic wounds are being tested by some armies; however direct pressure must still be applied. The ICRC has no experience with these and, like most others, eagerly awaits the published results of these field trials. Cost and availability will be factors to be taken into consideration for their use worldwide.

Peripheral haemorrhage can be controlled by external pressure and a compressive bandage.

7.7.5 Resuscitation or rehydration?

The first aider can only implement simple procedures in the field. The administration of intravenous fluids requires a certain degree of medical knowledge and specific equipment that are beyond the normal scope and competencies of a first aider. He can be involved in the monitoring of a perfusion, however.

The subject of pre-hospital i.v. fluids has given rise to considerable controversy. Maintaining adequate tissue perfusion and oxygenation is balanced off against the dilution of clotting factors and the dislodging of existing clots when the blood pressure goes up ("popping the clot"). The currently recommended "hypotensive resuscitation" involves giving fluids to a just palpable radial pulse – equivalent to a systolic blood pressure of 90 mm Hg – and no more (see Chapter 8).

A great deal of this controversy concerns efficient civilian trauma systems where evacuation times are short, less than 2 hours. It is probable that only the war-wounded showing obvious clinical signs of shock require pre-hospital i.v. fluids. In addition, in many contemporary conflicts transfer to a hospital takes much longer – days or even weeks – where "popping the clot" is not really relevant. Rehydration – as part of resuscitation – may be a more appropriate term under these circumstances.

If the wounded patient is conscious and not suffering from head trauma, the first aider can rehydrate with oral fluids (e.g. oral rehydration salts), especially if evacuation is delayed and transport long. This procedure is probably not employed nearly enough. Oral rehydration can probably be given with no detriment to the patient even in cases of abdominal trauma, and especially for severe burns.

7.7.6 Supplemental oxygen in the field?

Warning!

Oxygen cylinders are the equivalent of a bomb if hit by a bullet or piece of shrapnel. In addition to the danger they represent, the cylinders are heavy and must be replaced, lasting only a short time at high flows. Furthermore, their replenishment is complex and requires special factory capacity.

Oxygen cylinders must be ruled out in the event of any deployment to a dangerous area. This is now standard ICRC policy.

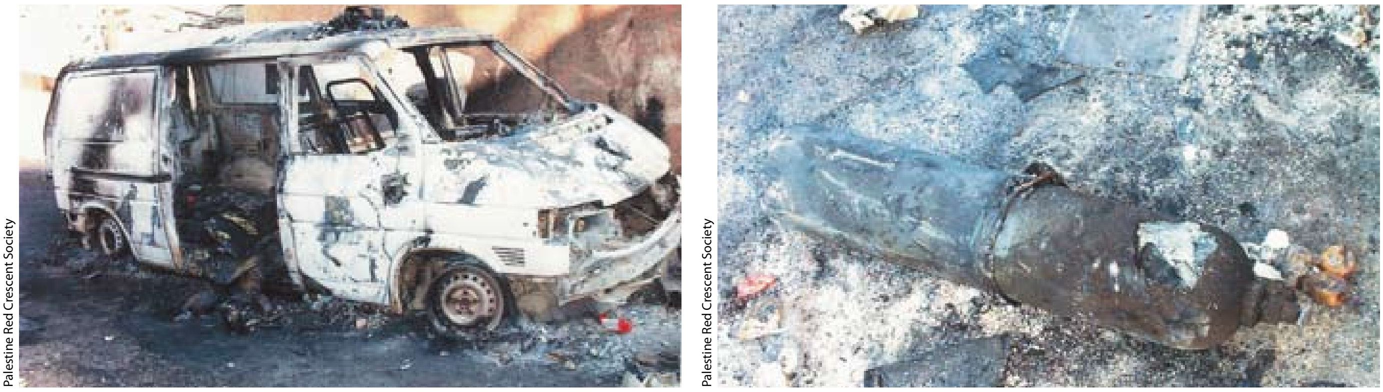

Palestine Red Crescent Society

Palestine Red Crescent Society

Figures 7.4.1 and 7.4.2

Ambulance destroyed by the detonation of an oxygen cylinder which has gone through the roof like a rocket. The photos show the ambulance and the cylinder on the ground.

Depending on security conditions, the collection point or intermediate station may have oxygen available. An oxygen concentrator (requiring an electrical supply) is preferable to compressed cylinders.

7.7.7 Disability: the AVPU system

Disability refers to the neurological status of the casualty: the state of consciousness and any paralysis due to injury to the spinal cord.

The Glasgow Coma Scale is the hospital standard. However, a simpler determination of consciousness is the AVPU system and it may be easier for first aiders to use in the field. It easily translates into the Glasgow system when the patient arrives at the hospital and gives the surgeon a good appreciation of the evolution of the patient's condition.

Alert	The casualty is awake, lucid, speaks normally and is responsive to the environment (e.g. eyes open spontaneously as you approach).
Voice responsive	The casualty is able to respond in a meaningful way when spoken to.
Pain responsive	The casualty does not respond to questions but moves or cries out in response to a painful stimulus (pinching the neck muscles, earlobe or nipple; rubbing the supra-orbital margin, or the angle of the mandible).
Unresponsive	The casualty does not respond to any stimuli.

Other conditions may influence the level of consciousness: hypoxia, shock, medication, and narcotics and alcohol (the latter all too often consumed to excess during times of armed conflict).

In the field, securing the airway is the first step to take when facing a casualty whose consciousness is compromised.

Examination of the spinal cord and vertebral column
This involves two phases: determining the movement of the limbs – comparing the two sides – and palpation of the bony prominences of the spine. Palpation of every vertebra – like "fingers playing on piano keys" – aims to identify any induration or deformity. If there is any danger to the spinal cord, then the four-person log-roll technique should be used to move the patient onto a stretcher. The cervical spine can be immobilized initially by manually controlling the head before a neck collar/Minerva splint is placed. Immobilization of the cervical spine must not increase the risk of airway obstruction, however.

The importance of the mechanism of injury, blunt or penetrating, has already been mentioned. In all cases of suspected injury to the vertebral column, proper handling of the patient and immobilization techniques are of the utmost importance; the injury may already be there, but the first aider should act in such a way that he does not worsen the condition.

7.7.8 Evacuation: a risk to take

Transportation of the wounded is always difficult, always takes longer than expected, adds to the trauma, and can be dangerous to the patient and those transporting the casualty: the famous "mortality of the ambulance ride". In situations of armed conflict, the danger of the ambulance ride is complicated by the risks of ongoing combat.

L. Berlemont / ICRC

Y. Muller / ICRC

Figures 7.5.1 and 7.5.2

Different means of patient evacuation: modern and traditional.

Speed is less important than safety of transport: ambulance crashes are notorious. All these parameters, aggravation of the patient's condition, use of resources and security, must be weighed against the likely benefits of moving the wounded.

In an urban environment, it is normal for a first aider to bypass the neighbourhood clinic if the injury is serious, and to evacuate to a clinic for a minor wound in order not to overburden the hospitals. Unfortunately, city-dwellers often spontaneously rush patients to hospital no matter how slight the injury. Evacuation and transport may be difficult and dangerous during urban fighting even if distances are very short and transfer may be delayed while waiting for a calm period or ceasefire.

There is also the confusion that reigns during street fighting and the highly-charged emotional atmosphere to contend with. Lack of discipline to perform good pre-hospital triage – and the "pressure" of bystanders – can result in chaos in the receiving hospital. The first wave of ambulances transports the dead and badly mutilated. The second wave of ambulances carries people who are shouting, hysterical, and frightened: the lucid and superficially injured. The third wave evacuates the seriously wounded who truly require emergency treatment; these patients lie still – they are usually haemorrhaging – without shouting to bring attention to themselves.

In remote rural areas, if transportation takes many hours or even days, it is logical to project forward competencies by training local providers in more advanced skills. The principles of triage apply to the choice of which patients to evacuate first (see Chapter 9). Note that there is an important *difference* in the priority to treat and the priority to evacuate, especially if evacuation is prolonged. The recognition of injuries that are not survivable helps to spare the victim and would-be rescuers the agony and frustration of unsuccessful efforts to reach a higher level of care, and to provide better care for those who can survive.

Delay in evacuation will contribute to an increase in pre-hospital mortality; the more severely injured suffer "nature's triage". As was seen in Chapter 5, longer and more difficult evacuation sorts out central injuries, with a consequent decline in hospital mortality: only those patients with a good chance of survival reach the hospital. This remains a major problem in the care of the war-wounded.

Chapter 8

HOSPITAL EMERGENCY ROOM CARE

8. HOSPITAL EMERGENCY ROOM CARE

8.1 ABCDE priorities 171

8.2 Initial examination 171

8.3 Airway 172
8.3.1 Opening the airway 173
8.3.2 Definitive patent airway 173
8.3.3 Definitive patent airway: endotracheal intubation 173
8.3.4 Definitive patent airway: surgical airway 174

8.4 Breathing and ventilation 175
8.4.1 Assisted ventilation 176

8.5 Circulation 177
8.5.1 Pericardial tamponade and pericardiocentesis 177
8.5.2 Haemorrhagic shock 178
8.5.3 Fluid replacement 179
8.5.4 Hypotensive resuscitation 181
8.5.5 Adjuvant therapy 181

8.6 Blood transfusion where supplies are limited 181
8.6.1 Clinical use of blood in ICRC practice 182
8.6.2 Two unit rule 182
8.6.3 Fresh whole blood 183
8.6.4 Autotransfusion 183
8.6.5 Total blood requirements: the ICRC's experience 183

8.7 Disability 184

8.8 Environment/exposure 185
8.8.1 "Fatal triad" of coagulopathy, acidosis, and hypothermia 185

8.9 Complete examination 185
8.9.1 Complementary diagnostic examinations and monitoring 187

8.1 ABCDE priorities

The framework for casualty management in a hospital emergency room (ER) is a continuation of the basic life support procedures of first aid. The logic is the same; the diagnostic and treatment means available are simply more advanced.

Whether or not the casualty arriving at a hospital has received first aid in the field, he/she should undergo the full ABCDE examination. The patient's condition may have changed during transport; important wounds may have been missed because of the confusion in the field; first-aid competence may have been limited or entirely absent. Efforts should again first concentrate on lifesaving treatment for asphyxia and shock, the most common remediable causes of death. The "golden hour" begins at the site of trauma, *not* on arrival at the emergency room.

Priorities: the **ABCDE** scheme of thinking

Airway before Breathing, before Circulation.

1. Assess:
Initial examination: **A**irway, **B**reathing, **C**irculation, **D**isability, **E**nvironment and **E**xposure. Triage in a mass casualty situation: see Chapter 9.

2. Act:
Emergency resuscitation: act on life-threatening conditions.

3. Assess:
Complete examination: head-to-toe palpation, front and back and sides.

4. Act:
Definitive treatment, surgical or not: stabilization.

5. Assess and Act:
Treatment or evacuation of the casualty, according to triage priority, to a higher echelon hospital for specialized care if necessary.

Please note:

In certain countries a very definite separation must be made for emergency examination and treatment of male and female patients. This can be problematic in a mass casualty situation; nonetheless, the emergency department must be organized accordingly.

8.2 Initial examination

The initial examination and emergency resuscitation are carried out simultaneously. The receiving ER doctor must automatically ask a series of questions.

1. Is the patient dead or alive?
2. Is the casualty conscious or not?
3. What is the mechanism of injury: penetrating or blunt?
4. What are the life-threatening conditions, if any, according to the ABCDE scheme?

The natural reflex when faced with a wounded person is to look at the bleeding first. It must be understood that the greatest threat to life is an inadequate intake of air. Airway takes precedence over breathing and over circulation. With practice, the airway and breathing can be secured quickly enough to then allow the doctor to deal with any visible haemorrhage. (The situation in a hospital emergency room is different from one of "catastrophic haemorrhage" in the field, as discussed in Chapter 7 on first aid.)

The ER doctor must learn to use the ABCDE scheme of thinking in a circular fashion. Determining if the patient is dead or alive already involves a rapid ABCDE examination. The dead have no air intake (A), no ventilation of the lungs (B), and no pulse (C); the pupils are dilated and do not react to light and there is no reaction to painful stimulus (D); there is no movement of the limbs and the body feels cold (E).

When mastered, the ABCDE scheme of thinking allows the doctor to answer all the above questions in one integrated process.

Most casualties are conscious, anxious or frightened, and in pain, and will tell you what happened and where "it hurts": they are alive, conscious, and the act of speaking shows that the airway is free. These signs of life may seem obvious. Nonetheless, one should make a mental checklist and practice the routine of asking the series of initial questions in order to establish an organized and coherent method of examination.

As mentioned in the previous chapter on first aid, blunt trauma above the level of the clavicles requires adequate care of the cervical spine, but not at the expense of a life-threatening airway problem. Simple measures include: manual stabilization in a neutral position with the head maintained in the axial line; semi-rigid collar; sandbags; taping; specialized backboard.

The initial examination should include the diagnosis of any life-threatening A, B, or C problems and the number, location and extent of *obvious* wounds. These two elements together are important for establishing triage priorities (see Chapter 9). A more complete examination to locate *all* wounds should follow the initial ABCDE examination.

It is important to determine the time since injury, estimate blood loss since injury, and the presence or absence of any allergies. In civilian casualties, past or pre-existing diseases and medications for chronic conditions should be taken into account. Young healthy combatants usually do not present such medical problems.

In a hospital environment the casualty should be completely undressed to allow for a correct examination – again, certain cultural or religious constraints may apply. Vital signs should be taken, i.v. lines placed and blood samples withdrawn simultaneously for blood grouping and cross-matching, and a haematocrit or haemoglobin estimation to determine a baseline for future comparison. Other tests, such as electrolytes, blood gases, glucose, etc. may be performed depending on local standards and laboratory capacity.

8.3 Airway

Airway obstruction is an emergency that requires an immediate response. Some injuries cause an immediate problem; others may result in delayed impairment of the airway.

The following list presents the common conditions that cause airway problems.

Head injury with decreased consciousness:
Apart from the danger of aspiration of vomitus, blood, broken teeth, bone or foreign bodies, the tongue and epiglottis may fall back and block the airway. Furthermore, the initially conscious casualty may slowly lose consciousness after some time.

Maxillo-facial injuries:
Even if the patient can breathe adequately in the beginning, the development of oedema of the tongue, floor of the mouth and pharynx will obstruct air intake after several hours.

Penetrating wounds to the larynx or upper trachea:
If large enough, they will create a "traumatic tracheotomy".

Blunt injury to the larynx (by a rifle butt):
This may crush the cartilage resulting in collapse of the airway.

Compressive haematoma in the neck:
This may accumulate slowly compressing the hypopharynx or larynx from the outside.

Burns to the face and neck or inhalation burns of the larynx and trachea:
Whether due to ordinary fire and smoke or specific irritating chemical agents, these require close observation for delayed obstruction or respiratory failure due to oedema.

Please note:

Inhalation of chemical agents requires special precautions for decontamination of the patient and of any equipment coming into contact with him. Not only is this an essential part of the treatment, but it is also necessary for the protection of the hospital staff and other patients.

The conditions causing delayed airway obstruction may develop slowly and there is a danger that they will not be recognized in time, especially during triage of mass casualties with inadequate supervision.

8.3.1 Opening the airway

As in first aid, the standard treatment for an actual or potential obstruction applies.

Basic steps for securing the airway

1. Open the mouth.
2. Displace the tongue forward.
3. Remove any blood or debris from the mouth and oropharynx.
4. Maintain the airway patent.

To open and clean the mouth the standard manoeuvres are jaw thrust or chin lift: displace the tongue forward, and then finger-sweep the mouth while protecting the finger, aided by mechanical suction if available. These techniques will cause some movement in the cervical spine. Manual in-line axial stabilization of the head during the manoeuvres will reduce such movement. Patency of the airway, nonetheless, takes precedence.

To maintain the airway patent a number of devices can be used:
- oropharyngeal airway (Guedel tube);
- nasopharyngeal airway;
- laryngeal mask airway;
- combitube (double lumen tube inserted blindly into the trachea and oesophagus).

All of these methods keep the airway open but cannot ensure protection against vomiting and aspiration of gastric contents.

During the reception of mass casualties, the overwhelmed ER personnel may have to revert to a simpler method as a temporary procedure. Under these conditions , the lateral recovery position (lateral security, post-tonsillectomy or semi-prone position) will be the position of choice for most patients with an airway at risk until more secure procedures can be achieved. The lateral recovery position affords relative protection against vomiting and aspiration.

8.3.2 Definitive patent airway

The following conditions require a *definitively* secured airway:
- apnoea or cardiac arrest, whatever the cause (with assisted ventilation);
- Glasgow Coma Scale score of less than or equal to 8, which is equal to being unresponsive to pain;
- continued seizures and convulsions (with assisted ventilation);
- unstable fractures of maxilla or mandible (usually bilateral fractures of the mandible, or instability of the whole mid-face at the level of the glabella and zygomata);
- large flail chest segment (with assisted ventilation);
- respiratory failure (with assisted ventilation);
- moderate to severe facial or oropharyngeal burn.

8.3.3 Definitive patent airway: endotracheal intubation

The best and simplest technique to maintain a patent airway with protection from aspiration is endotracheal intubation, through either a nasal or oral approach. Deeply unconscious patients can usually be intubated easily. Others may be restless, irritable, uncooperative and hypoxic. Intubation under these circumstances usually requires sedation. Various agents given intravenously will allow rapid intubation without struggling and without compromising the patient's hypoxic condition (diazepam, pentothal, propofol, ketamine). Cardiac arrest, whatever the cause, and severe haemorrhagic shock (Class IV, see below),require emergency endotracheal intubation. The alternative to endotracheal intubation is a surgical airway.

8.3.4 Definitive patent airway: surgical airway

The need for a surgical airway should be identified *early* and performed *quickly*. This may be the primary airway (maxillo-facial injuries, wounds to the neck involving the larynx or pharynx or haematoma accumulation, etc.) or following failure of endotracheal intubation. A surgical airway is also beneficial where there are no facilities for mechanical ventilation.

Cricothyroidotomy is preferable to tracheostomy, which can be a difficult procedure under emergency conditions and be associated with profuse haemorrhage.

Cricothyroidotomy

This is a quick, safe, and relatively bloodless procedure (Figures 8.1.1 – 8.1.4). A horizontal incision is made in the skin and extended through the cricothyroid membrane. The handle of the scalpel is inserted and turned 90° to hold the membrane open until a small tracheostomy tube can be inserted.

Figures 8.1.1 – 8.1.4

Cricothyroidotomy.

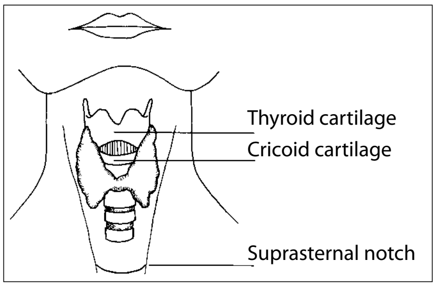

P. Zylstra / ICRC

Figure 8.1.1

Surgical landmarks: the patient's neck should be placed in extension with a pad beneath the shoulders. The thyroid and cricoid cartilages are identified by finger palpation, the cricothyroid membrane is then identified as the depression in between them.

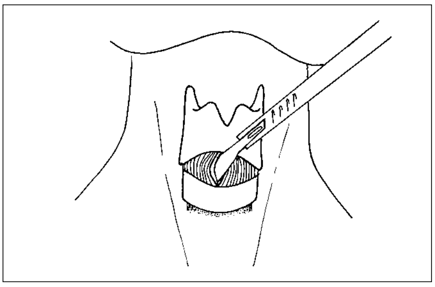

P. Zylstra / ICRC

Figure 8.1.2

A horizontal skin incision is made over the cricothyroid membrane. The wound is spread apart using the thumb and index finger.

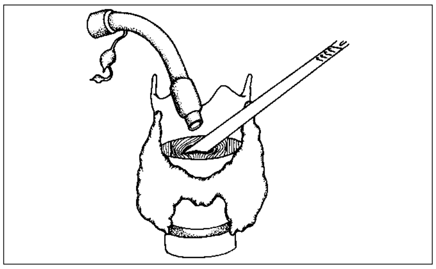

P. Zylstra / ICRC

Figure 8.1.3

The incision is carried down through the membrane and widened by insertion of the scalpel handle, which is then rotated through 90°.

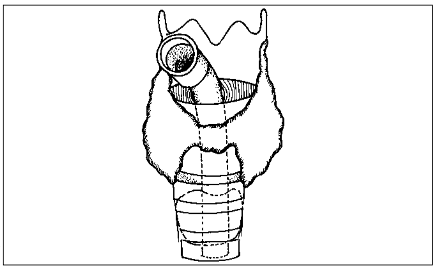

P. Zylstra / ICRC

Figure 8.1.4

A tracheostomy tube is placed through the opening and secured. The entire procedure should take no more than 30 seconds.

Tracheostomy

Tracheostomy should be an elective procedure. The only specific indication for an emergency tracheostomy in missile wounds is *direct laryngeal injury*, thus transforming a traumatic tracheostomy into a surgical one. The urgency of the problem will determine which technique is the safest and most appropriate for ensuring a secure airway.

8.4 Breathing and ventilation

The cause of respiratory distress must be found and treated. Patients with head injuries often require intubation and ventilation to support respiration, as is the case for quadriplegia, blast lung, chemical injuries and inhalation of smoke fumes. Previous disease may also impair ventilation in an injured patient.

Clinical examination may reveal a chest injury that decreases respiration including:
- flail segment of the chest;
- open pneumothorax or sucking wound to the chest;
- tension pneumothorax and haemopneumothorax.

Flail segment of the chest

This condition should be treated initially by bandaging or adhesive plaster strapping, good analgesia and positioning of the patient. More severe and complicated cases may require a chest tube and intubation with mechanical ventilation. It is usually the underlying lung trauma that causes the greatest difficulty in treatment.

For further details on flail segment of the chest, see Volume 2.

Sucking wound

A sucking wound requires a three-sided occlusive dressing in the ER. The patient then goes to theatre for debridement and closure of the wound of the chest wall and placement of an intercostal drain.

Please note:

There is a risk of transforming an open wound into a closed tension pneumothorax if all four sides of a dressing are firmly closed.

Tension pneumothorax

The diagnosis of a tension pneumothorax can be difficult on purely clinical grounds, except in the case of obvious respiratory distress; in this case no time should be wasted taking an X-ray (Figure 8.2). The condition requires an immediate wide-bore i.v. cannula to be inserted into the second or third intercostal space in the midclavicular line (needle thoracocentesis) attached to an improvised Heimlich one-way flutter valve as a temporary measure (Figure 8.3). A positive finding is indicated by the sudden rush of air as the needle enters the pleural cavity. The cannula should be long enough (8 cm) to fully penetrate the muscles of the chest wall (average thickness 4 – 6 cm). An intercostal chest tube (midclavicular or 5th intercostal space at the midaxillary line) should replace the needle thoracocentesis as soon as possible.

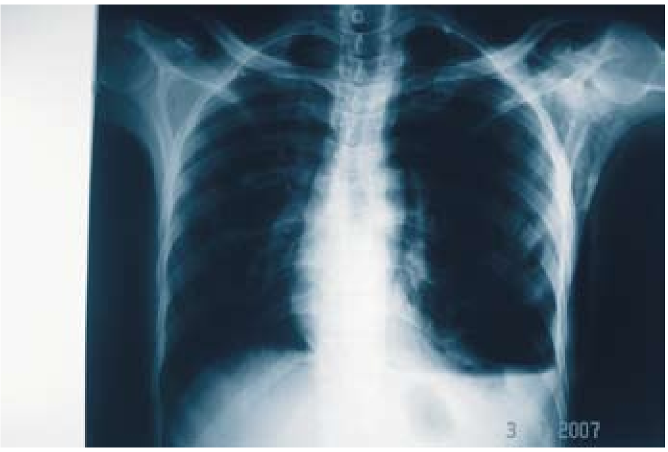

H. Nasreddine / ICRC

Figure 8.2

Tension pneumothorax with respiratory distress: the only thing wrong with the X-ray film is that it was taken. The diagnosis should have been a clinical one.

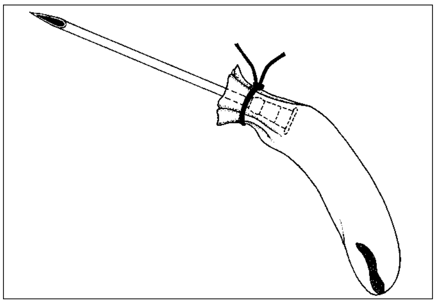

P. Zylstra / ICRC

Figure 8.3

Improvised Heimlich flutter valve: a wide bore cannula is inserted at the upper edge of a rib. A finger from a surgical glove, with a 1 cm-long incision in the end, is tied around the cannula.

Needle thoracocentesis may fail, however, even in the presence of a tension pneumothorax. A negative trial should not necessarily rule out its existence. A simple finger thoracostomy in the 5th intercostal space at the midaxillary line is a good alternative. A positive result, again, is indicated by the hiss of air rushing out; a chest tube is placed immediately.

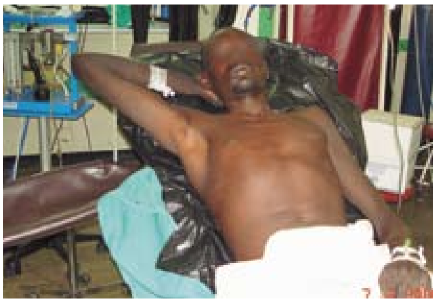

H. Nasreddine / ICRC

Figure 8.4.1

Simple, but massive, pneumothorax. Patient lying comfortably and breathing without effort. X-ray films are justified.

Other, less extreme presentations of tension pneumothorax are more common than is realized. Table 8.1 shows the important signs and symptoms for making the diagnosis. These pertain to the awake casualty and differ from those in the sedated and ventilated patient. It is acceptable to perform an X-ray in patients whose presentation may be suspicious but who are not in respiratory distress to confirm the diagnosis. The patient should be accompanied, however, by someone capable of performing an emergency thoracocentesis in case of rapid decompensation in his condition.

Reliable and early findings	Ipsilateral side	Contralateral side	Pre-terminal decompensation	Inconsistent findings
Pleuritic chest pain (universal)	Hyper-expansion of the chest	Increased mobility of the chest	Decreasing respiratory rate	Tracheal deviation to contralateral side
Air hunger (universal)	Decreased mobility of the chest		Hypotension	Distended neck veins
Respiratory distress (universal)	Hyper-resonance on percussion		Much decreased SpO_2	Surgical emphysema of chest wall
Tachypnoea	Decreased breath sounds on auscultation		Decreasing level of consciousness	
Tachycardia	Added breathing sounds – crackles, wheezes			
Decreasing SpO_2				
Agitation				

Table 8.1 Diagnosis of tension pneumothorax in awake patients.[1]

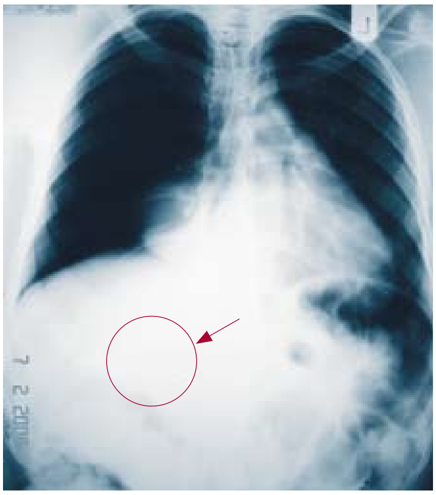

H. Nasreddine / ICRC

Figure 8.4.2

Antero-posterior radiograph showing large right-sided pneumothorax. The arrow marks the bullet.

Haemothorax

A haemothorax should be drained using a wide bore chest tube. This procedure can be lifesaving. When clinical signs of haemothorax exist, an intercostal tube should be inserted *before X-rays are taken*. Chest tubes are usually placed under local anaesthesia. If there is a wound to excise, ketamine anaesthesia would be more appropriate.

For information on insertion of chest tube – thoracostomy – see Volume 2.

8.4.1 Assisted ventilation

After intubation, assistance with respiration may be required. Common pathologies requiring such assistance include:
- head injury;
- large flail segment;
- blast injury of lungs;
- inhalation of toxic gases or smoke or flash burn to the bronchial tree;
- aspiration pneumonitis; and
- other medical causes of respiratory insufficiency.

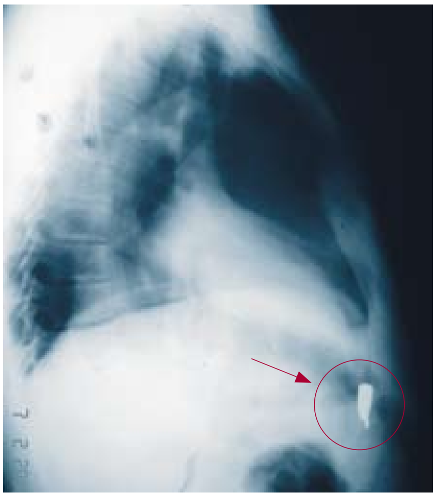

H. Nasreddine / ICRC

Figure 8.4.3

Lateral film.

Assisted ventilation may be manual:
- mouth-to-mouth or -nose for infants (use a compress barrier);
- mouth-to-mask;
- bag-valve-mask;
- bag-valve-endotracheal tube or surgical airway;

or mechanical, provided by a ventilator.

In hospital surroundings it is possible to administer supplemental oxygen, either by a central oxygen supply, compressed cylinders, or an oxygen extractor/concentrator.

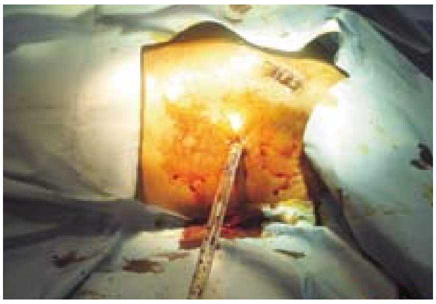

H. Nasreddine / ICRC

Figure 8.5

Intercostal tube.

The ICRC usually operates in situations of limited resources, and mechanical ventilators are not standard equipment. Few patients can be maintained for any length of time using manual ventilation by nurses and doctors. Furthermore, in a mass casualty situation the principles of triage will apply and most patients requiring assisted ventilation will be Category IV; therefore, it will not be performed (see Chapter 9).

---

[1] Adapted from: Leigh-Smith S, Harris T. Tension pneumothorax – time for a re-think? *Emerg Med J* 2005; **22**: 8 – 16.

8.5 Circulation

The main circulatory problem encountered in the war-wounded is hypovolaemic shock, and is usually due to haemorrhage or burns. In addition to any blood loss, big soft-tissue wounds sequestrate a great deal of tissue oedema with further loss of plasma and circulating volume. Dehydration complicates any previous fluid loss if evacuation is long and delayed.

Neurogenic, anaphylactic, and cardiogenic shock may also occur; septic shock is a late complication. Injury to the spinal cord causing quadri- or paraplegia will result in neurogenic shock due to a disproportion between blood volume and dilated vascular bed. Allergy to antibiotics must always be kept in mind. Traumatic cardiogenic shock is due to direct injury to the heart that is not immediately lethal (e.g. small shrapnel wound causing traumatic myocardial infarction and/or pericardial tamponade).

8.5.1 Pericardial tamponade and pericardiocentesis

In the rare instance of constrictive *haemopericardium* with cardiac tamponade due to a penetrating missile wound to the heart, pericardiocentesis may be required if there is acute decompensation, to *gain time* until emergency thoracotomy can be performed.

The procedure for pericardiocentesis is as follows.

1. A 20 ml syringe is attached to a long over-the-needle cannula (or alternatively a spinal anaesthesia needle).
2. The skin is punctured 1 – 2 cm to the left of the xiphisternal junction, at a 45° angle, and the needle advanced through fascia and muscle.
3. (The inner trocar is now removed if using a spinal needle – but not if using the needle of the cannula).
4. The needle is gently advanced while aiming at the tip of the left scapula and applying continuous suction on the syringe.
5. When the needle tip enters the pericardial sac, blood will immediately appear in the syringe. The needle is withdrawn if using a cannula, which is then advanced into the sac.
6. As much blood as possible is aspirated. If the tamponade has been relieved, the patient's condition improves immediately; if the right ventricle has been entered, there is no change.
7. At the end of aspiration, the synthetic over-the-needle cannula can be left in place with a 3-way stopcock valve. (The spinal needle is gently withdrawn, millimetre-by-millimetre, if this technique is being used).
8. Should the tamponade return quickly, the stopcock can be re-opened and aspiration repeated (or the spinal needle tap is repeated).
9. The patient should *immediately* be taken to theatre!

An emergency room thoracotomy should not be performed when the operating theatre is just down the corridor: it is as pointless as it is dangerous in *most* hospitals around the world.

8.5.2 Haemorrhagic shock

Types of haemorrhage

Haemorrhage may be arterial, venous or capillary and blood loss may be:

- peripheral and obvious (blood clot the size of a fist or an open wound the size of a hand represents 500 ml);
- peripheral and hidden:
  - closed fractures of long bones (tibia = 500 ml, femur = 1.5 litres);
  - open wound with small entry that is blocked by a piece of torn muscle;
- central (chest, abdomen, pelvis and retroperitoneum) – significant bleeding into the chest should have been diagnosed during the breathing phase of the initial examination; a massive haemothorax can represent 2 – 3 litres of blood, a severe fracture of the pelvis 3 litres.

Penetrating trauma causing central bleeding into a body cavity requires surgery of some sort, and severe internal haemorrhage may require urgent operation as part of the resuscitation process. Such patients have a high priority for surgery.

Direct pressure and elevation of the limb will stop most cases of peripheral venous and capillary haemorrhage. Fractures should be splinted.

If a bleeding blood vessel can be *seen* in the depths of a wound, *and only if it is clearly visible*, direct control may be obtained with haemostatic artery forceps.

Blind clamping must never be attempted.

Otherwise, severe peripheral bleeding can usually be stopped by packing the wound. Digital pressure is applied to the artery proximal to the wound (pressure point), while the wound is carefully packed with small quantities of gauze compress at first, followed by a more bulky dressing, and finally a firm compressive elastic bandage in order to apply even pressure to achieve haemostasis. Alternatively, and if available, a pneumatic tourniquet can be used to temporarily control severe arterial haemorrhage while bringing the patient to the operating theatre.

Once a wound has been packed and bleeding has been arrested, the dressing (or tourniquet) should not be removed until the patient has been resuscitated and is in the operating theatre, with blood ready for transfusion if available. The surgical team should be ready to arrest bleeding by rapid intervention to obtain access to and control of the major blood vessels.

In the same manner of thinking, implanted bullets and fragments, or other foreign bodies in the wound, should only be removed as part of the operative procedure.

Beware of removing packs from missile wounds, as the resultant bleeding can be torrential and difficult to control.

The body's response and classes of shock

Immediately after haemorrhage, the body initiates a number of homeostatic circulatory changes that aim to stop the bleeding as well as compensate and preserve perfusion of the vital organs. The circulating blood volume represents 7 – 8 % of body weight in the adult (5 – 5.6 litres in the 70-kg male or 70 ml/kg body weight) and 9 % in children (80 ml/kg body weight).

Haemorrhage and the shock response are traditionally graded into 4 classes according to the volume of *acute blood loss*.

Class I:

Up to 15 % blood volume lost (750 ml or less). Mild tachycardia is the only clinical sign since the body's normal homeostatic mechanisms are capable of fully compensating the loss.

Class II:
15 – 30 % blood volume lost (750 – 1,500 ml). Definite tachycardia; slight decrease in systolic blood pressure with rise of diastolic pressure (decreased pulse pressure); refilling of blanched capillary bed of fingers delayed; restlessness or anxiety.

Class III:
30 – 40 % blood volume lost (1,500 – 2,000 ml). Marked tachycardia; tachypnoea; hypotension; small urine volume; classical picture of shock. Compensatory mechanisms start to fail.

Class IV:
> 40 % blood volume lost (> 2,000 ml). Full classic symptoms of shock are present: cold, clammy and pale skin; irritability, aggressiveness and confusion leading to loss of consciousness with loss of more than 50 % circulating volume.

Class	I up to 750 ml (<15 % loss)	II 750 – 1,500 ml (15 – 30 % loss)	III 1,500 – 2,000 ml (30 – 40 % loss)	IV >2,000 ml (>40 % loss)
Pulse	<100/min Full and bounding	100 – 120/min Full	120 –140/min Weak	>140/min Thready
Systolic blood pressure	120 Normal	90 – 120	< 90 Radial not felt	< 60 Carotid not felt
Pulse pressure	Normal	Narrowed	Greatly decreased	Absent
Capillary refill	Normal	Delayed	Delayed	Absent
Respiratory rate	14 – 20/min Normal	20 – 30/min Mild tachypnoea	>30/min Marked tachypnoea	>35/min Marked tachypnoea
Urine output	>30 ml/hr	20 – 30 ml/hr	5 – 20 ml/hr	Negligible
Mental function	Lucid/ thirsty/ slightly anxious	Anxious/ frightened/ irritable	Hostile/ irritable/ confused	Confused/ lethargic/ unresponsive
Physiological status	Fully compensated	Peripheral vasoconstriction	Compensation fails, classical clinical picture	Immediately life-threatening

Table 8.2 Signs and symptoms of haemorrhagic shock according to Class.

8.5.3 Fluid replacement

The important point is to maintain adequate tissue perfusion until the haemorrhage is brought under control. Sufficient blood pressure to maintain this tissue perfusion is judged to be a systolic pressure of 90 mm Hg, equivalent to a palpable radial pulse.

The great majority of war-wounded patients are relatively healthy young adults who suffer injuries to the extremities, and in whom the amount of blood loss is not lethal. They are haemodynamically stable (Class I blood loss) and the usefulness of oral fluids for their resuscitation – in the field and in hospital – is probably underestimated. Head, thoracic and abdominal injuries, however, even with Class I blood loss, should always have an intravenous line placed.

In massive injury, with obviously gross blood loss, several wide-bore intravenous lines should be established. A venous cut-down may be required if shock is profound. Available sites include: median basilic or cephalic in the arm, greater saphenous in the groin, and the distal saphenous vein at the medial malleolus. A venous cut-down should be used for less than 24 hours, or until the patient is filled up with fluids and another i.v. line can be put in. The site of the wounds will influence where i.v. lines are placed.

Modern practice has been to give a fluid challenge and monitor the response. The challenge has usually been with 2 l of isotonic crystalloid solution[2] in 30 minutes, or 3 ml for every 1 ml of estimated blood lost as a general rule of thumb. (Except for very large estimates: 3,000 ml blood loss would require 9,000 ml Ringers lactate!). The fluid challenge for a child is calculated as 20 ml/kg body weight for the initial bolus of Ringers lactate.

---

[2] Ringers lactate is the preferred resuscitation fluid in ICRC practice.

Monitoring of clinical response

The estimation of blood loss is a very approximative measure, however. Rather than relying on this to guide resuscitation efforts, medics should observe the signs and symptoms of the *clinical response* to determine continuing fluid requirements. These include:
- pulse,
- systolic blood pressure,
- pulse pressure difference between systolic and diastolic,
- capillary refill,
- urine output, and
- mental status.

The most important simple parameter for the adequacy of fluid replacement is *urine output*: a urine flow of 0.5 – 1 ml/kg body weight/hour, and even greater in case of crush syndrome, should be aimed for.

The clinical response to a fluid challenge can be summarized as follows.

Rapid and stable response:
The pulse falls below 100, the systolic blood pressure rises above 100, and the pulse pressure widens. Urine output is good. These measurements remain stable. No further fluid resuscitation is required; the i.v. line is maintained open, however. This is the usual result with Class II and some Class III blood loss.

Transient unstable response:
An initial positive response of the pulse, blood pressure, and pulse pressure is followed by a return to subnormal values. Urine output remains low. A further challenge with Ringers lactate or a plasma expander (dextran 70, colloid, etc.) is called for. A return to normal values indicates compensated Class III blood loss, but early surgery is still required. Continuing subnormal shock values indicate continuing blood loss; the patient should be prepared for emergency surgery.

Please note:

Not more than two units of colloid or dextran should be given in any 24-hour period; besides being more expensive, colloids are less effective than Ringers lactate in diffusing into the interstitial space and can help provoke coagulation defects and interfere with cross-matching.

No response:
The patient remains in shock, indicating Class IV blood loss of more than 40 % blood volume, and requires emergency surgery – surgery as part of resuscitation – or is triaged to Category IV in a mass casualty situation (see Chapter 9);

or

other, less common pathologies that become manifest with time have been overlooked. The airway and breathing should be re-assessed to diagnose a possible cardiac tamponade, tension pneumothorax, or myocardial injury. Neurogenic shock and acute gastric dilatation should not be overlooked.

Time since injury should also be taken into account to evaluate the evolution of the shock state. If grade IV shock is present less than 1 hour after injury, emergency operation is needed for resuscitation. If grade IV shock has developed in 4 hours, resuscitation is required before operation.

Warning!

Beware of resuscitation injury: “shock lung” or acute respiratory distress syndrome in patients who receive massive crystalloid infusions that overburden the heart and lungs.

8.5.4 Hypotensive resuscitation

Giving all shocked patients a standard fluid challenge of 2 litres or more has been called into question. A rapid return to normal blood pressure *before* bleeding is controlled can be associated with hydraulic disruption of an effective blood clot ("popping the clot"), a dilution of clotting factors, and a lowering of blood viscosity that causes a decrease in resistance to flow around an incomplete clot. All of these factors may cause bleeding that has temporarily stopped to start again, especially if it is a non-compressible central haemorrhage. Thus, *moderate resuscitation* may be better than aggressive resuscitation; raising the systolic blood pressure to only 90 mm Hg rather to over 100. This is known as "hypotensive resuscitation" and should be considered in Class III and IV blood loss. It should *not* be used in head trauma where any hypotension is deleterious, or in the very young or old.

Therefore, in cases of massive haemorrhage, one must strive for a balance between death from shock after primary exsanguination or an increase in bleeding by prevention of clot formation or dislodging of formed and effective blood clot with secondary haemorrhage. In these extreme cases, surgery to stop haemorrhage is part of resuscitation: this should *never* be forgotten. Hypotensive resuscitation is widely used by ICRC surgeons today for patients who arrive quickly at a hospital.

However, if evacuation is delayed (more than 6 – 12 hours), rebleeding is less likely. In addition, delayed evacuation will result in more profound shock because of plasma loss due to post-traumatic tissue oedema and general dehydration rather than pure blood loss. Excessive sweating, vomiting or diarrhoea, rough handling of the casualty during transport, and a lack of splinting of fractured limbs will aggravate the total fluid and electrolyte loss. More aggressive resuscitation is permissible under these circumstances and the patient should be fully hydrated prior to surgery.

8.5.5 Adjuvant therapy

Never give cold fluids by rapid intravenous infusion.

No effort should be spared to prevent hypothermia in the shocked patient. Intravenous fluids should be warmed (see Section 8.8.1 and Chapter 18).

Oxygen and small doses of i.v. analgesia are equally important. The best is morphine (5 mg i.v., repeated every 10 minutes as necessary). Morphine should *not* be given if there is any question of head injury or respiratory depression. A good alternative in this case would be tramadol. Analgesia in shock should be given intravenously only.

A naso-gastric tube to prevent gastric dilatation should not be overlooked.

Steroids or a "vitamin C cocktail" should *not* be administered in cases of haemorrhagic shock.

Some clinical experiments have taken place using hypertonic saline for resuscitation. ICRC surgical teams have no experience in this and cannot comment.

8.6 Blood transfusion where supplies are limited

Where blood supplies are scarce, what should the role of blood transfusion be? This scenario is a far cry from optimal conditions in which there are relatively few limits to blood or component administration, yet it is common.

The aim of blood transfusion is to save life or to prevent significant morbidity, and *not* to restore a normal haemoglobin level. Blood is a rare and expensive commodity with serious risks attached to its administration and it should therefore be used with caution. The decision to give blood to a particular patient should be based primarily on the clinical state correlated with the laboratory findings, balanced against the risks and the shortage of supply. A better understanding of the physiology of oxygen transport, the shortage of donors and the increased risk of viral infections has led to a policy of accepting lower levels of haemoglobin than before, without overly detrimental effects to the patient. One should avoid a transfusion for reasons such as "accelerating the patient's recovery", increasing patient comfort, or providing a supplement to correct anaemia. (Certain medical conditions, such as small children with severe malaria are an exception to this last rule.)

ICRC policy is to screen donated blood for hepatitis A, B and C, HIV, syphilis, and – in non-endemic areas – malaria. In countries where malaria is endemic, it is not uncommon for trauma patients to have a malarial attack 2 to 3 days post-operatively even without a blood transfusion; this is treated when it occurs. Where donors are scare, they are not excluded because of malaria. A blood transfusion can never be 100 % safe.

The ICRC seeks blood donors in the community. Family, friends and clan members are usually the source. In certain countries, the National Red Cross/Red Crescent Society plays an important role in the collection of blood.

Culture and tradition in some societies may render blood collection extremely difficult; consequently, blood for transfusion is often in very short supply. Accordingly, and in keeping with the principles of triage, a maximum number of units of blood per intervention is set in ICRC practice. Today, this is usually about 4 units, and this figure should only be exceeded in cases of anti-personnel landmine injury with traumatic amputation and severe burn patients undergoing skin grafting. (Early tangential debridement with immediate skin grafting is not practised by ICRC teams, partly because of the risk of major blood loss.)

ICRC practice has been to transfuse whole blood, as fresh as possible. Otherwise, whole blood is stored with CPD-A anticoagulant (citrate – phosphate – dextrose – adenine). Blood components are not available in ICRC surgical programmes as is the case in many rural public hospitals. In practice, this has not been considered a handicap.

8.6.1 Clinical use of blood in ICRC practice

- Blood should not be given during resuscitation until bleeding has been controlled.
- Crystalloids and/or colloids are administered first: if the patient remains haemodynamically unstable *and* the haemoglobin is less than 6 g/dl, blood is administered. Haemoglobin less than 6 but in a stable patient is *not* an indication for transfusion. (However, there is a threshold haematocrit value of 5 – 10 %, below which continued crystalloids or colloids can provoke cardiac arrest by "washout anaemia".[3])
- In the non-resuscitation scenario (preparing the patient for a second operation), pre-operative transfusions are not administered where blood supplies are low, unless the patient is anaemic *and* symptomatic.
- The blood supply and stocks determine whether a particular patient will receive blood: the principles of triage apply.
- Consider autotransfusion where appropriate (see Section 8.6.4).

If massive transfusion of stored blood is necessary, every second unit should be supplemented with one ampoule of sodium bicarbonate (44.3 mEq) and one ampoule of calcium chloride (10 g) by a separate i.v. line. As with crystalloid fluids, blood should be warmed to body temperature to avoid increasing hypothermia. This can be achieved through the use of locally-made water baths or the body heat of staff members.

8.6.2 Two unit rule

It has been traditional practice to administer not less than 2 units of blood to any patient who requires transfusion. In the ICRC context, this rule is not always appropriate since extreme shortage of blood is so common. It is sometimes more apt to prescribe just one unit to certain symptomatic patients as this may improve their condition sufficiently, thus allowing supplies of this scarce resource to be kept for other patients in need. This is particularly true for patients in danger of "washout anaemia". World Health Organization (WHO) guidelines also accept one unit of blood in case of symptomatic anaemia.

---

[3] Takaori M, Safar P. Treatment of massive hemorrhage with colloid and crystalloid solutions. *JAMA* 1967; **199**: 297 – 302. Cited in Barkana Y, Stein M, et al. Prehospital blood transfusion in prolonged evacuation. *J Trauma* 1999; **46**: 176 – 180 and Shoemaker WC, Peitzman AB, Bellamy R, et al. Resuscitation for severe hemorrhage. *Crit Care Med* 1996; **24(2Suppl.)**: S12 – S23.

8.6.3 Fresh whole blood

This is best when screened and given within one hour of collection. Fresh whole blood is especially reserved for:
- massive haemorrhage;
- coagulopathy;
- septic shock;
- non-trauma pathologies:
  - snake bite with haemolysis
  - amniotic fluid embolism.

8.6.4 Autotransfusion

When faced with patients suffering massive haemorrhage and inadequate blood supplies, ICRC surgical teams have practised recuperation of shed blood and autotransfusion. Haemothorax and haemoperitoneum from the spleen, liver or ruptured ectopic pregnancy are the most common indications.

For further information on autotransfusion in acute haemorrhage, see Volume 2.

8.6.5 Total blood requirements: the ICRC's experience[4]

ICRC colleagues studied total blood requirements for 4,770 patients in two hospitals treating the wounded from the war in Afghanistan during a six-month period in 1990 – 91. No differentiation was made for blood transfusion given pre-operatively, peri-operatively, or immediately post-operatively: all were considered part of patient resuscitation given the difficulties and delay in obtaining blood. Guidelines at the time allocated a maximum of 6 units per patient (although this was exceeded in some cases) and a limiting haemoglobin level of 8 g/dl. Autotransfusion was not used in this series.

A comparison was made of the number of units transfused according to time since injury for all wounded patients and separately for those with central injuries (head, neck, chest, and abdomen), and according to the cause of injury.

Need for transfusion was greatest for patients arriving less than six hours after injury, steadily decreasing to those arriving after 72 hours. This was true for all patients. Surprisingly, patients with central injuries required less blood, on average, than those with peripheral wounds.

The most remarkable results were associated with the cause of injury. Blood requirement for anti-personnel landmine injuries far exceeded that for bullets or fragments, (Table 8.3) and was the main reason for the high blood use in peripheral injuries.

	Mine	Fragment	Gunshot	Burn
Patients	787	2,577	1,016	111
Percentage transfused	27.9 %	13.1 %	15.0 %	18.0 %
Average transfusion units/ patient transfused	3.7	2.6	2.7	4.1
Average units/100 patients	103.2	34.1	40.5	73.8

Table 8.3 Blood given to patients related to cause of injury: 279 patients were categorized under "other" injuries.

---

[4] Eshaya-Chauvin B, Coupland RM. Transfusion requirements for the management of war injured: the experience of the International Committee of the Red Cross. *Br J Anaesth* 1992; **68**: 221 – 223.

Consequently, ICRC recommendations for a hospital undertaking total surgical care of the wounded from conventional warfare were amended as follows:

- for every 100 war-wounded patients, the blood bank should expect to provide 45 units as a baseline;
- baseline should be increased to 60 units, if the majority of patients are admitted within 6 hours of injury;
- baseline should be increased to 100 units, if anti-personnel landmines are widely used in combat;
- burn patients require greater transfusion capacities, even in the absence of early tangential excision with immediate grafting (not practised in ICRC hospitals);
- long evacuation lines with much delay and no use of anti-personnel landmines in combat may negate the need for a blood bank, and blood for transfusion should be collected on an individual patient basis.

These recommendations may have little relevance to a modern industrialized army with very early evacuation and resuscitation but may be quite relevant in conditions of limited resources.

Please note:

ICRC guidelines have become even more stringent since: 4 unit limit and 6 g/dl haemoglobin level *with* unstable haemodynamics.

8.7 Disability

Any neurological deficit must be determined, whether central or peripheral. As mentioned previously, if the mechanism of injury is blunt trauma above the level of the clavicles then the cervical spine must be cared for in the classical manner.

According to the nursing expertise available, the AVPU system may be used or one may proceed immediately to the Glasgow Coma Scale (GCS) for the determination of the level of consciousness and any traumatic brain injury (Table 8.4). Although the GCS was originally intended for closed trauma to the head, and there are certain shortcomings when dealing with penetrating head trauma, nonetheless, it has been ICRC practice to use it.

	Response	Score
Eye opening	Spontaneous	4
	To verbal command	3
	To pain	2
	No response	1
Verbal response	Orientated and conversing	5
	Disorientated	4
	Inappropriate words	3
	Incomprehensible sounds	2
	No response	1
Motor response	Obeys verbal command	6
	Localizes pain	5
	Flexion-withdrawal	4
	Flexion-abnormal	3
	Extension	2
	No response	1

Table 8.4 Glasgow Coma Scale: the best response for each variable is scored. Maximum score is 15, minimum is 3.

A GCS < 8 indicates severe head injury and airway protection is mandatory, by intubation or cricothyroidotomy/tracheostomy.

Disability examination includes the entire vertebral column/spinal cord: presence of paraplegia, level, etc. Palpation of the vertebrae, one by one, for tenderness, induration and crepitus is more important than deformity, which may be masked by haematoma. A rectal examination to determine sphincter tonus is important for prognosis!

Proper stabilization measures should be undertaken and a urinary catheter placed. Neurogenic shock is common in paraplegic patients: fluid resuscitation and an i.v. vasopressor are often required.

For the management of paraplegic patients, see Volume 2.

8.8 Environment/exposure

Hypothermia must be avoided at all costs – rapid, yet thorough examination of the patient – and treated aggressively. With a core body temperature of 37° C, an ambient temperature of 32 – 34° C is considered neutral. Less than this and the body loses heat to the environment. After examination, the patient should be kept covered, even in a tropical climate.

8.8.1 "Fatal triad" of coagulopathy, acidosis, and hypothermia

Recently, the importance of the potentially "fatal triad" of hypothermia, acidosis, and coagulopathy in trauma patients has become recognized. Hypothermia (core temperature less than 35° C) is probably the most potent factor in causing the vicious cycle of this syndrome.

Every effort should be made to preserve heat in an injured patient, since rewarming consumes far more energy than maintaining normothermia. Warm inhaled O_2, warm i.v. fluids, and external rewarming to a maximum of 40 – 42° C should be the first step. More aggressive measures of "central reheating" such as rectal enema and gastric, bladder and peritoneal lavage (at 37° C) can be used.

Post-traumatic coagulopathy is a well-recognized danger in patients suffering severe shock and submitted to aggressive resuscitation, especially if they have received massive amounts of i.v. fluids and blood transfusions with stored blood. It results from a combination of factors. Fresh whole blood transfusion is very useful in the absence of blood components, and probably in their presence as well (see Chapter 18).

8.9 Complete examination

At this stage it is even more important than in the pre-hospital setting to undress the patient and perform a thorough examination, from head to toe, front and back and sides. In some societies, this may contravene certain cultural and religious traditions (male doctor examining a female patient). Compromises must be found.

In the more accommodating atmosphere of a hospital emergency room, a systematic approach should be used to thoroughly examine the scalp and head (mouth, nose, and ears), neck, thorax, abdomen, perineum (scrotum and urethra, rectum, and vagina), the back of the trunk and buttocks, and the extremities. The peripheral pulses, temperature and capillary refill are compared on both sides. Motor function of the main peripheral nerves is tested. The aim is to have a complete assessment of all injuries and a more accurate assessment of the organ-specific damage.

The complete examination is best described as a thorough palpation. The entry wound may be exceedingly small and not seen. This is especially the case with fragment wounds to the head or perineum where body hair matted with blood can easily hide the wound (Figure 8.6). The entry wound must be identified by close palpation. Remember also that contusion/erythema can be better felt rather than seen in dark-skinned people.

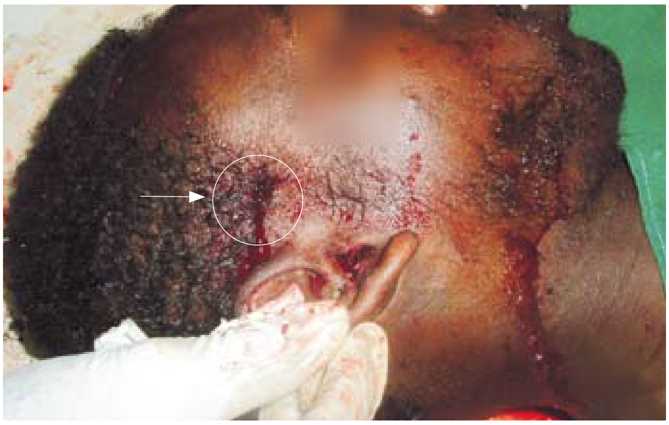

M. Baldan / ICRC

Figure 8.6

Small temporo-zygomatic entry wound hidden by hair.

One should attempt to identify the likely path through the body of the projectile. This may involve any structure between the entry and exit wounds, or the position of the projectile on X-ray if no exit exists. Remember that wounds to the chest, buttock, thigh or perineum may involve the abdominal cavity (Figures 8.7.1 – 8.7.3).

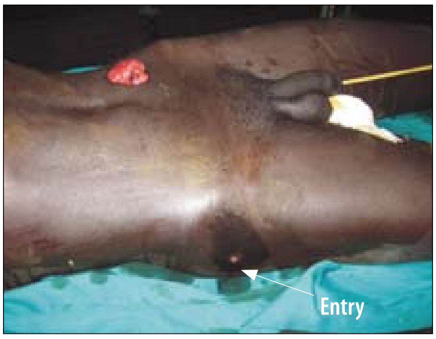

H. Nasreddine / ICRC

Figure 8.7.1

Bullet wound to the pelvis: the entry is located on the right side. A defunctioning colostomy has been performed.

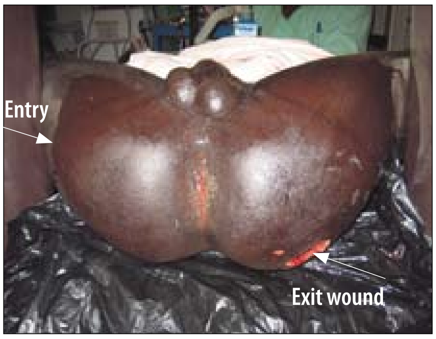

H. Nasreddine / ICRC

Figure 8.7.2

Exit wound in the left buttock.

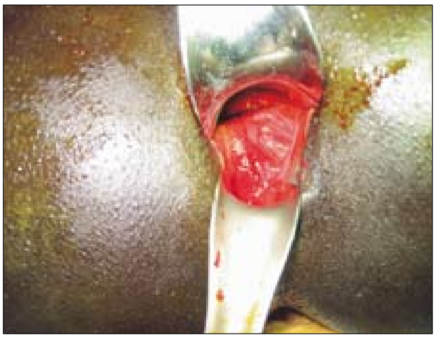

H. Nasreddine / ICRC

Figure 8.7.3

Any structure between entry and exit should be identified: here blood is seen in the rectum.

A simple outline drawing of the body on the admission chart (homunculus), front and back, is useful for recording all injuries.

Dressings on the limbs should not be removed if the casualty is haemodynamically unstable. Only once resuscitation has begun and the patient's condition is under control is it safe to examine wounds to the extremities, preferably in the operating theatre. Fractures should nonetheless be immobilized if this has not already been done in the field.

The ABCDE paradigm should be monitored for any change in condition of the patient. Resuscitation and stabilization are continued, while complementary examinations are performed. The extent of the latter will depend on the level of sophistication and competency of the particular hospital.

A basic complement is a plain X-ray, one body cavity above and below any entry or exit wound. If no exit wound exists and no projectile is evident, further radiographs should be taken to locate its position. It may be difficult to differentiate a radio-opaque bullet from a normal anatomic radio-opacity such as the heart shadow (see Chapter 10 and Figures 8.4.2 and 14.9.1).

The complete and definitive examination will involve each specialized body system. The signs, symptoms, and treatment will be described in the relevant chapters of Volume 2.

8.9.1 Complementary diagnostic examinations and monitoring

Electrocardiogram (ECG) monitoring is not routinely available in ICRC surgical hospitals, neither are computerized tomography (CT) scans, angiography, sonography, Doppler blood flow, or arterial blood gases. The use of central venous pressure lines in most circumstances where the ICRC works carries too high a risk of septicaemia. Diagnostic peritoneal lavage for abdominal injury is also not practised routinely.

The minimal ICRC standards for emergency war surgery under precarious circumstances in hospital settings with limited resources include:
- plain X-ray;
- pulse oxymeter;
- haemoglobin;
- haematocrit;
- total and differential white blood cell count;
- platelet count;
- coagulation time;
- bleeding time;
- fasting blood sugar;
- blood smear for malaria (other blood parasites where applicable);
- sickle-cell test (where applicable);
- urine analysis: dipsticks, pregnancy tests;
- blood grouping, testing and cross-matching.

If more general surgery, as well as internal medicine and paediatrics, is widely practised in the ICRC hospital, then more elaborate laboratory analyses are added. In more precarious situations, as experienced by field surgical teams, none of the above is usually available.

Chapter 9

HOSPITAL TRIAGE OF MASS CASUALTIES

9. HOSPITAL TRIAGE OF MASS CASUALTIES

9.1 Introduction 191
9.1.1 The logic of triage 191
9.1.2 Where to perform triage 192
9.1.3 A delicate balance 192

9.2 Setting priorities: the ICRC triage system 193
9.2.1 The ICRC triage categories 194
9.2.2 Notes on triage categories 195

9.3 How to perform triage 196
9.3.1 "Sift" 196
9.3.2 "Sort" 197
9.3.3 Avoid undertriage and overtriage 197
9.3.4 Avoid confusion and disagreements within the team 198

9.4 Triage documentation 198

9.5 Emergency plan for mass casualties: disaster triage plan 199
9.5.1 Planning 199
9.5.2 The team 200

9.6 Personnel 200
9.6.1 Triage team leader 200
9.6.2 Triage officer 200
9.6.3 Head nurse 201
9.6.4 The team 201
9.6.5 Triage groups/nursing teams 201
9.6.6 Surgeons and theatre personnel 202
9.6.7 Rest and relief 202

9.7 Space 202

9.8 Equipment and supplies 203

9.9 Infrastructure 204

9.10 Services 204

9.11 Training 204

9.12 Communication 204

9.13 Security 205

9.14 Summary of triage theory and philosophy: sorting by priority 206
9.14.1 Triage system: a simple emergency plan organizing the personnel, space, infrastructure, equipment, and supplies 206
9.14.2 Emergency hospital disaster triage plans differ and no two triage scenarios are the same 206
9.14.3 "Best for most" policy 206

ANNEX 9. A Sample triage card 207

ANNEX 9. B Hospital emergency plan for a mass influx of wounded 208

9.1 Introduction

9.1.1 The logic of triage

In daily, routine practice surgeons face patients one by one. They use all of the means, equipment, and supplies at hand to do the most they can for every patient; they try to do everything they possibly can for every single individual. The priority is high-intensity care for the sickest.

In a single mass-casualty incident means may be stretched to the limit, but one can still manage to do the best possible for all patients. With a massive influx of wounded, however, the hospital is overwhelmed; resources cannot meet the needs of all the casualties. This is when the logic of mass casualty triage comes into play. It is no longer possible to do everything for everyone. Medics must try to do what they can: the greatest good – but not necessarily everything – for the greatest number. This is another example of the change in professional "software" – the mindset – that is called for in surgery for the victims of war.

A change of mindset: from "everything for everyone" to "best for most".

The French term *triage*, which means to sort into groups according to quality, has been applied to the sorting of battle casualties since the time of Napoleon's surgeon of the Imperial Guard, Baron Dominique-Jean Larrey (circa 1812), who recognized the need to categorize wounded soldiers during an actual battle according to their priority for treatment.

The experience of two world wars during the twentieth century, when tens of thousands were wounded during a single battle, demonstrated the importance of sorting casualties and determining priorities for evacuation and treatment. This concept of triage has now been extended to natural disasters and single mass-casualty incidents (terrorist bombings, industrial accidents, fires in social institutions, etc.) with considerable benefit, and its use is now generally accepted worldwide.

Natural disaster triage, however, is not the same as war triage. An important natural disaster is a single mass-casualty event that overwhelms hospital services; the wounded suffer injuries at one time, but there is then a respite. The same can be said for other mass-casualty incidents. Armed conflict on the other hand can involve a continuing mass arrival of freshly wounded victims that may go on for weeks or months on end; there may be no respite until the end of hostilities. Nonetheless, many of the fundamental concepts underlying war triage apply to the disaster scenario.

Establishing priorities for treatment amongst many casualties is the most difficult decision to be taken in all of medical practice. Patients with very severe wounds whose management consumes many resources and much time, and who have little chance of survival despite treatment, may have to wait or may only receive minimal care in the interest of being able to intervene effectively for the others. One must often choose to treat first the patients for whom the smallest surgery gives the biggest results; i.e. the wounded who, though seriously injured, stand a fair chance of "good survival".

The aim in a mass casualty situation is to do the "best for most".

Some armies have practised "inverted" or "reverse" triage: the least injured were treated first so as to make them able to return more quickly to the battlefield! It should be noted that this practice contradicts the principles of international humanitarian law.

The practice of triage is not dogmatic. It is not a series of rules, but a logical approach and philosophy that must be adapted to each and every particular situation.

9.1.2 Where to perform triage

As mentioned in Chapter 1, one of the special characteristics of surgery in times of war is the staged management of patients in a chain of casualty care. The principles of triage are applied at every stage.

Triage takes place at every echelon of the chain of casualty care, including during evacuation.

The organization of a casualty collection point and effective triage permits the orderly evacuation of the wounded, the most efficient use of stretchers, ambulances or other transport, and the optimal use of medical and surgical personnel available. With the necessary training, rapid assessment of casualties can be carried out not only by doctors, but also by nurses, paramedical staff and first-aid workers. It should be possible to implement triage at *any point* along the chain of casualty care, in order to set *priorities* for first aid and evacuation of the wounded to the next echelon of treatment.

More efficient lines of evacuation and shorter duration of transport, coupled with better pre-hospital care, ensure that a greater number of the severely injured will survive to reach a surgical hospital. The number of KIA decreases; however, the DOW rate often increases (see Chapter 5).

In many low-income countries, however, civilian casualties often arrive at the hospital without having received first aid and without the help of an ambulance service. In this case the less seriously injured usually arrive first. The hospital then becomes the first triage station.

Whatever the prior treatment or sorting, however, triage must be carried out again when casualties reach the hospital; the condition of patients changes and the priorities in the hospital may be different from those in the field. All patients arriving at the hospital during an influx of casualties go through the triage process.[1]

9.1.3 A delicate balance

To set priorities for patient management a number of factors must be taken into consideration to define the needs on the one hand, and the resources available on the other. The practice of triage is a fine balance between these two.

Triage is a dynamic equilibrium between needs and resources:

- needs = number of wounded and types of wounds;
- resources = facilities at hand and number of competent personnel available.

Needs

- How many patients are arriving? 10, 50 or 100?
- Are they all suffering from penetrating injuries? Or are there many burn cases amongst them?

The number of wounded and the different pathologies should be taken into account in the total workload. Penetrating wounds will require a great deal of surgery. Burn patients require relatively little immediate surgery, but a great deal of nursing care.

---

[1] For field triage, see Giannou C, Bernes E. *First Aid in Armed Conflicts and Other Situations of Violence*. Geneva: ICRC; 2006.

Resources

- How many surgeons and anaesthetists?
- How many operating tables?
- How many instrument boxes, and what is the efficiency of the sterilization system?
- How many beds? etc.

A hospital may have three fully-equipped operating theatres, but if there is only one surgeon then only one patient can be operated at a time. If there are three surgeons and three anaesthetists, but only one theatre, then the team must improvise extra operating facilities, if there are sufficient boxes of surgical instruments.

ICRC EXPERIENCE

During fighting in Kabul in 1992 with an influx of mass casualties, four surgical teams worked in the ICRC hospital, supplemented by Afghan general practitioners with some surgical experience. There were two operating theatres in different buildings with 4 tables that could function at the same time. One team operated in one building and one in the other; the third team performed triage; and the fourth was resting!

All these factors – needs and resources – must be balanced out. No two triage situations are the same. It is this constantly changing equilibrium between needs and resources that will determine the priority for treatment among all the patients who reach the hospital. One cannot be dogmatic about which patient to operate first, since no two situations are the same. Only a full understanding of the logic and philosophy of triage will help to establish efficient priorities that will result in the best treatment possible for the greatest number of victims.

9.2 Setting priorities: the ICRC triage system

There are a number of triage category systems used throughout the world today. Some are more sophisticated than others and depend on injury severity scores and physiological parameters. There are two important factors to keep in mind when deciding upon a system to use in a hospital.

1. It should be kept as simple as possible – a mass influx of wounded always creates confusion, tension and anxiety.
2. All members of the hospital team should understand the system being used.

The ICRC uses a triage category system based upon several factors.

- Physiological appreciation of life-threatening conditions according to the ABCDE protocol – in practice, severe haemorrhage will constitute the great majority of life-threatening injuries during armed conflict (see Chapters 5 and 8).
- Anatomic factors, including the Red Cross Wound Score for penetrating injuries – in general, vital injuries (head, neck, thorax, abdomen and major peripheral vessels) will have priority, but the grade of injury must also be taken into consideration (see Chapter 4).
- Mechanism of injury – especially anti-personnel landmine injuries and burns (see Chapter 3).
- Time since injury – important for determining degree of shock and response to resuscitation as well as limb ischaemia (see Chapter 5).
- Epidemiology of the war-wounded – knowledge of the relative numbers of severe and superficial wounds (see Chapter 5).

9.2.1 The ICRC triage categories[2]

Category I: Serious wounds – resuscitation and immediate surgery

Patients who need lifesaving surgery and have a good chance of recovery. Some examples include:

- airway – injuries or burns to the face and neck requiring tracheostomy;
- breathing – tension pneumothorax, major haemothorax;
- circulation – internal haemorrhage, wounds to major peripheral blood vessels, traumatic amputation.

Category II: Second priority wounds – can wait for surgery

Patients who require surgery but not on an urgent basis. In practice this applies to many casualties, including:

- penetrating abdominal injury in haemodynamically stable patient, most probably injury to hollow organs only;
- penetrating head injuries with GCS > 8, providing definitive airway control can be maintained – if a surgical airway is necessary, then Category I for tracheostomy only;[3]
- most compound fractures: in practice, a large number of the wounded;
- major soft-tissue wounds: in practice, a large number of the wounded.

Category III: Superficial wounds – ambulatory management

Patients who do not require hospitalization and/or surgery because their wounds are so minor that they can be managed on an ambulatory basis.

They are often called the "walking-wounded". In practice, this is a very large group including those presenting superficial wounds managed under local anaesthesia in the emergency room or with simple first-aid measures.

Category IV: Severe wounds – supportive treatment

Patients with injuries so severe that they are unlikely to survive or would have a very poor quality of survival. These include the moribund, or patients with multiple major injuries whose management could be considered wasteful of scarce resources in a mass casualty situation in terms of operative time and blood. Examples include:

- penetrating head wound with GCS < 8;
- quadriplegia;
- burns > 50 % body surface area;
- major blood loss and no blood available.

---

[2] The system of triage categories employed by the ICRC was revised by the Master Surgeons Workshop held in Geneva in 2002 (see Introduction).

[3] Please note: ICRC hospitals are *not* equipped with ventilators and close monitoring of intubated patients is not always possible. A surgical airway avoids many problems, and should replace an endotracheal tube if the patient remains intubated for several days in any case.

9.2.2 Notes on triage categories

The number of casualties who require urgent treatment may exceed the surgical capacity available. A second triage within Category I is then necessary (see Section 9.3.2). On the other hand, when evacuation time to the hospital is longer than 12 hours, few patients may fall into Category I.

Many surgeons believe that all penetrating brain wounds are a Category I emergency; others believe that all penetrating brain wounds are hopeless Category IV. The use of the GCS helps to differentiate those patients who will probably survive, even with some delay in operation (Category II in the experience of ICRC surgical teams) from those whose condition is definitely a Category IV. This, evidently, provided the airway is patent.

The Category III superficially wounded may be so numerous, frightened, in pain and excited, that their uncontrolled presence in the emergency room/triage area creates great confusion. In urban contexts, they tend to be the very first evacuated to hospital, to the detriment of the more seriously wounded.

The categories are not rigid; patients waiting for surgery may change category and a single patient may fall into two categories, such as one with severe maxillo-facial injury who requires an immediate tracheostomy and basic haemostasis (Category I), while the debridement and primary reconstruction of the patient's face, which may take many hours, can wait (Category II).

Please note:

There is one exception to these categories: when an excited, and often drunk, combatant holds a gun to your head and demands that you treat his wounded comrade first. This patient immediately becomes top priority.

ICRC EXPERIENCE

Two examples of the epidemiology of triage patients recorded in recent years.

Fighting in Kisangani, Democratic Republic of the Congo, June 2000

Four hospitals and 62 clinics registered a total of 2,393 wounded. Only 25 % required hospital care, and even less a surgical operation. The vast majority were Category III and could have been treated as ambulatory cases; many were hospitalized for several days nonetheless.

Combat in Monrovia, Liberia, June – August 2003

2,567 patients were triaged in the JFK Memorial Hospital, but only 1,015 were deemed to require hospital admission (40 %), only 718 of whom (71 %) underwent operation. Of those admitted, some did not require any operation (paraplegia), while for many others there was simply not enough time or human resources available. Many patients who were originally placed in Category II – and given antibiotics, dressings and analgesia – eventually became Category III, with and without sequelae.

The equilibrium – between needs and resources – is a dynamic one; change is constant, no two triage situations are the same.

9.3 How to perform triage

Triage is a dynamic process: it requires a continuous reassessment of patients.

Patients may change triage category: their condition may deteriorate, or improve over time and with pre-operative resuscitation. As a result, a continuous reassessment of patients is absolutely necessary.

Triage is a multiple-step process: "sift and sort", then re-examine, re-examine, re-examine. "Sift" involves placing the patient in a general Category; "sort" then decides priority within that Category.

9.3.1 "Sift"

On reception of the casualties in the hospital, the triage officer must perform a rapid examination of each patient – a maximum of 30 seconds – checking the whole body, including the back. Field dressings should be changed by the nursing team as part of the examination; except for obviously large and severe wounds whose dressings are removed only in theatre. Severity of injury and probability of survival are the keys to decision-making. Clinical experience with the war-wounded – not training or reading books – is the most important preparation for this task.

To sift, the triage officer first looks for life-threatening conditions based on:
- airway, breathing, circulation;
- important physiological clues (mental status, character and rate of pulse, ease of respiration);
- anatomic site of wounds (head, chest, abdomen);
- severity of obvious wounds according to the RCWS (amputation of limb, etc.).

Based on clinical judgement alone, a first decision is made, to "sift" the casualty into a Category. The Category does *not* depend on how many patients arrive with serious wounds.

Every wounded patient must be examined *and* categorized immediately.

The idea is to quickly recognize the two extremes of injury, which, together, comprise a very large number of casualties: the very minor (Category III) and the very serious (Category IV); and to identify and separate out the dead. As little time as possible is spent on these Category III and IV patients and they must quickly be removed from the triage area to specially designated areas.

The triage officer concentrates on the critically and seriously injured (Categories I and II):
- patients who require resuscitation and immediate surgery as part of the resuscitation process;
- those who need continuing resuscitation; and
- patients who will tolerate some delay before receiving surgical attention.

The triage officer should not perform *treatment*, with one exception: if the patient is not breathing, the mouth should be opened and cleaned and the person put in the lateral recovery position, before the next patient is dealt with.

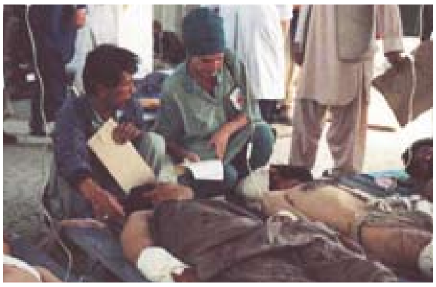
Å. Molde / ICRC

Figure 9.1
Triage officer assisted by an administrative clerk.

An "administrative team" consisting of a nurse and a clerk supports the triage officer for purposes of documentation, including all personal details of each patient, the securing of their identity papers and valuables, and the signing of any necessary consent forms. Other supporting personnel present at the entrance or in the triage area include those responsible for hospital security and crowd control.

Initial treatment is started by a dedicated team, and patients should be transferred as quickly as possible out of the triage area to theatre or the ward designated for their triage Category in order to make room for new arrivals. Resuscitation, as necessary, is continued in these assigned areas.

The use of X-rays should be *limited*; it is seldom essential under these circumstances and the triage officer should not wait for an X-ray to decide on the triage Category.

An identifying tabard or armband stating "triage officer" is useful; if, inevitably, the triage officer must move away from the triage area this can be handed over to the deputy/replacement.

9.3.2 "Sort"

After "sifting" casualties into the general Categories, a second examination is performed: "sorting". No two triage situations are equal and, therefore, according to the number of competent personnel available, the triage officer or a second physician-in-charge continues the reassessment of patients within the Category I cohort. "Sorting" decides which of the Category I patients to send to theatre first: *priority amongst the priorities*. The others, while waiting, remain under close observation and continued resuscitation.

This re-examination may reveal patients with an "emergent" condition who should receive priority over the stable: e.g. one of two Category I patients with penetrating abdominal wounds becomes haemodynamically unstable because of continuing internal haemorrhage, while the other has a stable pulse and blood pressure because bleeding has stopped.

The surgical teams should begin operating on Category I patients immediately. Since the number of casualties cannot be foreseen, it is not possible to wait for all patients to arrive and be triaged before deciding which should be taken to the operating theatre first.

When in doubt about priority amongst Category I patients, put in chest tubes and send in a laparotomy; get going!

The triage officer must keep in close contact with the operating theatre, to reassess continuously the priorities of the operating list.

The time that patients are waiting for surgery can be profitably used for completing X-rays and laboratory analyses, if necessary, but should not interfere with resuscitation efforts.

The same logic applies to the patients in the other Categories. Either the triage officer or the person in charge of the department re-examines the patients in Category II. Priorities are set for the operating list among the Category II patients.

Reassessments may result in changing a patient's Category. The condition of a patient originally triaged into Category III because he had no *obviously* serious wound may suddenly deteriorate and become Category I or II. If a second-in-charge is continuing the reassessments, a system must be available to transfer patients from one Category area to another. Referring back to the triage officer for every decision will only result in overburdening him.

Usually the hardest part of triage is having to accept that some patients may only receive analgesics and be removed to a quiet place where they can die in comfort and with dignity. After treatment of priority patients has been completed, a repeat examination of Category IV patients, if still alive, may allow the triage officer to contemplate surgery for them at this stage.

9.3.3 Avoid undertriage and overtriage

In undertriage the assessment underestimates the severity of injury and the patient is not given sufficient priority. Overtriage overestimates the injury, and a patient is assigned to a higher Category than necessary. This will divert resources from the truly seriously injured and overburden the critical care services. Repeated re-examinations will correct these errors.

It may be necessary to transfer minimal care patients to other sites for convalescence, even outside the surgical hospital. If this is done, care should be taken to identify patients for follow-up treatment and administration of medication, and to exercise some control over their whereabouts. Documentation is important: patients can easily get lost.

9.3.4 Avoid confusion and disagreements within the team

Triage decisions must be respected.

There is no time or place for disagreements during a triage of mass casualties. The decisions of the triage officer must be "dictatorial". The post-triage evaluation session is the place for "democratic" discussion and constructive criticism.

In spite of training, practising and planning, the unexpected will happen. The hospital team – led by the team leader, triage officer, head nurse – will at times have to improvise and invent new protocols and procedures to meet new circumstances. One should not be dogmatic, but rather understand the philosophy and logic of the triage process, and adapt to new situations in accordance with that logic.

9.4 Triage documentation

Good records are essential and no effort should be spared to record important aspects of the wounds, treatment, and the patient's triage category.

Each casualty should be appropriately identified, numbered, and assigned a medical chart. Large plastic bags, labelled with the patient's number, are used for clothing; smaller labelled plastic bags are used to collect patients' valuables. They are stored separately and the valuables put in a safe place.

Some system must be devised for indicating the patient's triage Category. This could be colour-coded tags tied around a hand or foot, or hung around the neck. These are easily removed and changed should the patient's triage Category change. Writing an "indelible" number on the forehead or chest only creates confusion with a change in Category.

The patient's medical chart should include basic information and be in telegraphic style: clear, concise, yet complete. It should include *at least* the following information:
- name, age, sex, time of injury, cause of injury, any first aid given;
- time of admission to hospital;
- vital signs: BP, pulse, respiration rate, neurological status;
- diagnosis: using a diagram (homunculus) is very useful;
- triage Category;
- complete pre-operative orders.

See Annex 9. A: Sample triage card.

This basic information is particularly important if patients are being transferred to another facility. A list of admitted or treated patients is necessary so that people who come looking for their relatives or friends may be informed. The local authorities may require information about the number of admissions and deaths. The media must be dealt with. These issues are addressed by the triage team leader, or replacement.

ICRC EXPERIENCE

During the post-triage evaluation, it was decided to use a plastic sleeve to hold the triage cards in order to protect them from various body fluids.

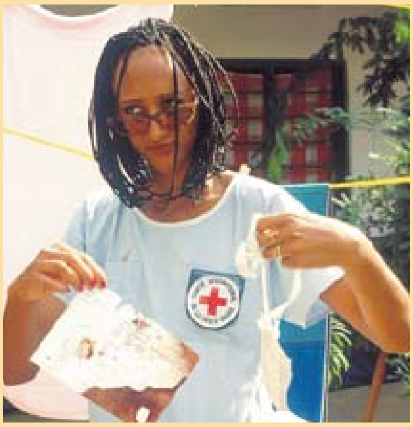

ICRC

Figure 9.2

State of the admission card after triage.

9.5 Emergency plan for mass casualties: disaster triage plan

Any hospital treating the war-wounded must be prepared to receive large numbers of casualties. Preparation means planning and training.

9.5.1 Planning

The triage tent in Figure 9.3 shows a number of key points:

- there is enough space to move around;
- the small lightweight beds/stretchers that are easily moved and inexpensive;
- small carts for emergency medical supplies;
- i.v. fluids hanging on a rope strung across the room, for flexibility of patient placement;
- a patient carried on a stretcher by dedicated bearers;
- relatively few staff members present, getting on with their allotted tasks in what appears to be a calm atmosphere.

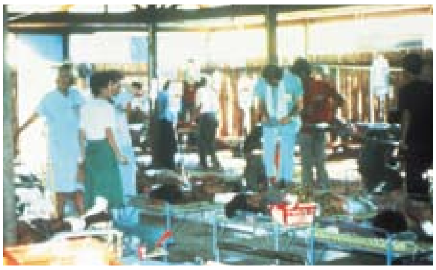

ICRC

Figure 9.3

Inside a triage tent.

This scene illustrates proper organization, which requires planning of the space, infrastructure, equipment, supplies and personnel.

Figure 9.4 shows the premises surrounding a triage tent:

- the site is temporary;
- space is available to receive ambulances or to expand the facilities;
- the structure is lightweight and easy to set up or take down;
- the structure is inexpensive;
- there is a water tower nearby;
- there is a rubbish bin outside.

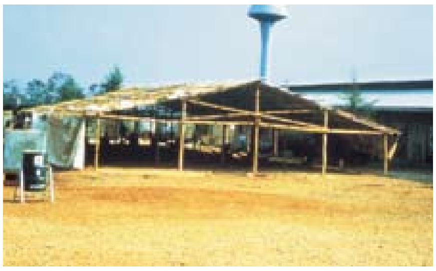

ICRC

Figure 9.4

Outside a triage tent.

This scene illustrates proper organization, which involves planning of the space and infrastructure.

A triage system cannot be organized ad hoc:
it requires planning and training.

The hospital team must be prepared for any kind of crisis: every hospital should have a disaster/triage plan (see Annex 9. B: Hospital emergency plan for mass influx of wounded). The aim of the plan is to organize:

- personnel;
- space;
- equipment;
- supplies (medical and non-medical);
- infrastructure (water, fuel for electricity generators, etc.);

- services (laundry, kitchen and cafeteria, etc.);
- training of the hospital team;
- communications;
- security.

9.5.2 The team

The hospital team should hold a series of meetings, to discuss the organization of the disaster/triage plan. Everyone working in the hospital should be aware of the plan and their respective role during a crisis. The plan should be posted so that everyone is well acquainted with it.

The plan should be put into operation as soon as notice is given of the expected arrival of mass casualties. It should include the mechanism for deciding who declares the emergency and under what conditions to implement the plan. The ordinary operating list and other routine activities should be suspended until the situation is resolved. This kind of organization does not require money or special technology; only time, effort, discipline and motivation. Any disaster plan should be an extension of the normal hospital routines, and the roles allocated to individual staff remain as close as possible to their familiar daily work. The plan should not be a complete change in the system. Introducing many new procedures will only add to the confusion.

A simple emergency plan: personnel, space, infrastructure, equipment, supplies = system.

9.6 Personnel

There are three key leadership functions in a disaster triage scenario: triage team leader, triage officer, head nurse.

9.6.1 Triage team leader

The triage team leader is the coordinator. He is usually designated to announce the onset of the hospital triage plan; he then coordinates the work of the different units and services, and makes sure that all departments are informed. The triage team leader maintains an overview of the situation, including a constant reassessment to determine the need for additional staff, supplies, and ward areas. In addition, he must be aware of events outside the hospital, maintaining contact with relevant authorities in order to anticipate any new arrival of casualties due to continuing combat.

9.6.2 Triage officer

The triage officer performs the actual clinical triage, assigning a category to each and every patient on entry. There has been much discussion about who should perform triage: surgeon or anaesthetist? Again, there is no strict rule. Each hospital team must decide according to its circumstances.

The logic of triage demands that the most experienced and respected person willing and able to take on the responsibility should do so. This person must know how to organize the emergency room/triage area and have a good understanding of the functioning and capacity of the hospital.

Even more importantly, the hospital team must be able to live with the decisions taken by the triage officer. Staff members, relatives, and military commanders might try to influence the triage decisions; nevertheless, these should be made on purely medical grounds. This can be especially difficult in a public hospital where friends and family members of the hospital staff may be among the victims. Deciding priority of treatment solely on the basis of medical need and resources available when the injured include close relatives of the personnel may be a heart-wrenching decision.

The hospital staff must be able to continue their work and live with the triage decisions for the rest of their lives. As a consequence, the experience of the triage officer and the respect with which the personnel hold the triage officer must be beyond reproach.

Triage Officer

No task in medicine requires greater understanding, skill, and judgement than the categorization of casualties and the establishment of priorities for treatment.

9.6.3 Head nurse

The head nurse organizes the nursing and paramedical personnel (laboratory, pharmacy, etc.) and non-medical support staff (trolley orderlies/stretcher-bearers, kitchen, laundry, cleaning, etc.). This is largely a coordination function. Any clinical role or supervision by the head nurse will depend on the particular circumstances of the hospital involved.

9.6.4 The team

Whether one person fills more than one leadership function will depend on the availability of competent personnel. The triage team leader may be the same as the triage officer in a small hospital; in a larger facility, this coordination function is probably best assigned to someone else, for instance an administrative officer or the head nurse. In a very large hospital, three different people should assume the three functions.

ICRC EXPERIENCE

What is the reality in the field?

In many small rural hospitals, the only surgeon is also the director. This one person often becomes the triage team leader and triage officer. After performing triage, the surgeon and the only anaesthesia nurse go to the operating theatre. A new influx of casualties arrives. Who then performs triage? Who continues the necessary coordination tasks? Either a general practitioner or the head nurse. This should be decided upon beforehand and be part of the hospital's disaster plan.

The ICRC supports the Somali Red Crescent Society in running the Keysaney Hospital in north Mogadishu, where triage has been routinely practised from 1992 until the time of writing. The triage officer in the emergency room is an "experienced female nurse, because she is the only one who can organize things".

9.6.5 Triage groups/nursing teams

Such teams should be formed for the triage area, with responsibility for the following tasks.

1. Setting up intravenous lines and taking blood for grouping and cross-matching.
2. Administration of tetanus prophylaxis, antibiotics, analgesics and other medication as prescribed.
3. Dressing wounds and splinting fractures.
4. Bladder catheterization, if indicated.
5. Arranging for an orderly flow of categorized casualties to the operating theatre or to designated areas for continuing resuscitation, delayed, or minimal treatment.

The disaster plan may be nominative, designating by name the triage officer and triage groups: who is in charge of setting up i.v. lines, dressing wounds, giving analgesics and antibiotics, etc.; *or* it may designate the function (ER doctor 1, ER nurse 3) regardless of who is on the roster that day (Figure 9.5). This depends on the staffing system, and availability of personnel, of the particular hospital.

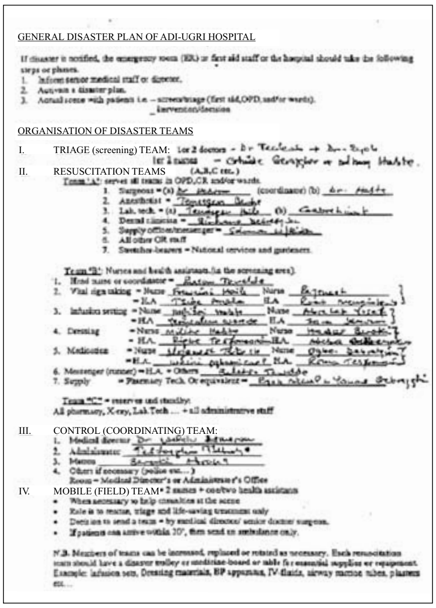

GENERAL DISASTER PLAN OF ADI-UGRI HOSPITAL

If disaster is notified, the emergency room (ER) or first aid staff or the hospital should take the following steps or phases.
1. Inform senior medical staff or director.
2. Activate a disaster plan.
3. Actual scene with patients i.e. – screen/triage (first aid, OPD, and/or wards).
_Intervention/decision

ORGANISATION OF DISASTER TEAMS

I. TRIAGE (screening) TEAM: 1 or 2 doctors –
1 or 2 nurses –

II. RESUSCITATION TEAMS (A,B,C etc.)
Team "A": serves all teams in OPD, OR and/or wards.
1. Surgeons = (a) (coordinator) (b)
2. Anesthetist =
3. Lab. tech. = (a) (b)
4. Dental clinician =
5. Supply officer/messenger =
6. All other OR staff
7. Stretcher-bearers = National services and gardeners.

Team "B": Nurses and health assistants (in the screening area).
1. Head nurse or coordinator =
2. Vital sign taking = Nurse Nurse
= H.A. H.A.
3. Infusion setting = Nurse Nurse
= H.A. H.A.
4. Dressing = Nurse Nurse
= H.A. H.A.
5. Medication = Nurse Nurse
= H.A. H.A.
6. Messenger (runner) = H.A. + Others
7. Supply = Pharmacy Tech. Or equivalent =

Team "C" = reserves and standby:
All pharmacy, X-ray, Lab. Tech .... + all administrative staff

III. CONTROL (COORDINATING) TEAM:
1. Medical director
2. Administrator
3. Matron
4. Others if necessary (police etc....)
Room = Medical Director's or Administrator's Office

IV. MOBILE (FIELD) TEAM = 2 nurses + one/two health assistants
- When necessary to help casualties at the scene
- Role is to rescue, triage and life-saving treatment only
- Decision to send a team = by medical director/ senior doctor/ surgeon.
- If patients can arrive within 20', then send an ambulance only.

N.B. Members of teams can be increased, replaced or rotated as necessary. Each resuscitation team should have a disaster trolley or medicine-board or table for essential supplies or equipment. Example: infusion sets, Dressing materials, BP apparatus, IV fluids, airway suction tubes, plasters etc....

M. Ghebrehiwet / MoPH Eritrea

Figure 9.5

Example of a nominative disaster/triage plan in a small rural hospital.

9.6.6 Surgeons and theatre personnel

These members of staff should have prepared the operating theatre and be ready and waiting for the casualties. The surgeon in theatre will see patients he or she has not previously examined and who may not have a name or signed consent form. Outside the mass casualty situation this would be regarded as negligent practice, and so another "mental adjustment" is required.

9.6.7 Rest and relief

Early consideration should be given to the *rest and relief* of staff. Triage is not an everyday routine activity. The shift system may need to be changed in order to deal with the increased workload: 2 twelve-hour shifts instead of 3 eight-hour ones, for example.

During a crisis the hospital team will be under severe emotional and physical stress. All staff members will need to pace themselves in order to perform efficiently and maintain a professional attitude. As mentioned before, during armed conflict there may be a mass arrival of new casualties every day for weeks on end. It is not possible to foresee how long the influx will last.

ICRC EXPERIENCE

At the beginning of the fighting in Monrovia in 2003, the surgical team worked 48 hours non-stop, without sleep: they were not able to continue operating. Subsequently, a strict 18-hour operating schedule was instituted: the team was able to maintain this schedule for three months.

A triage situation is tiring for everyone. Some, out of a misplaced sense of professional duty, find it difficult to accept that they *need* to rest; rest must be insisted upon. Working for extended periods without sleep only results in hospital staff no longer being able to adequately treat the new patients who continue to arrive.

9.7 Space

During an influx of mass casualties the various hospital departments must be re-arranged according to the pre-decided plan. Besides the original hospital site, any alternative sites (building, underground shelter, etc.) must be included in the plan, should the hospital require evacuation for reasons of security. The equivalent in cases of natural disaster is the destruction of the hospital premises and/or its access routes (earthquake, landslide, tsunami, etc.).

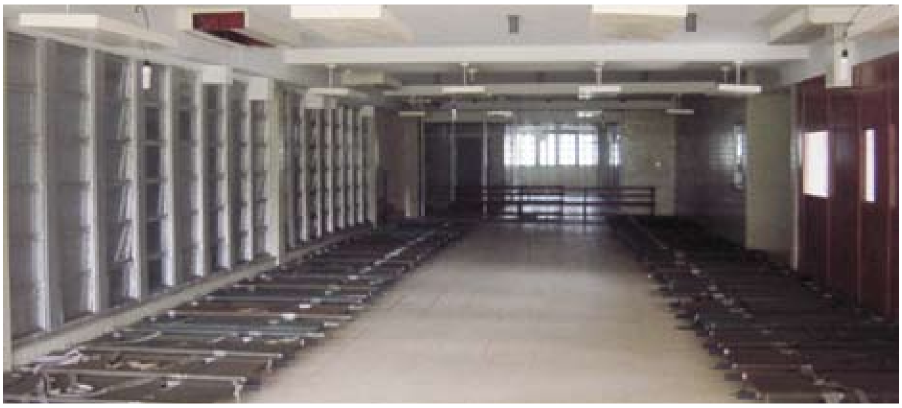
R. Aburabi / ICRC

Figure 9.6.1
Alternative triage department in an unused building: empty.

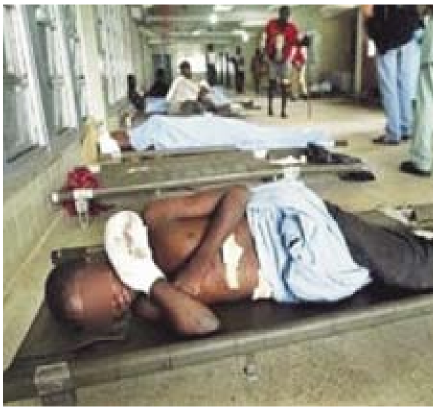
R. Aburabi / ICRC

Figure 9.6.2
The same premises: full.

The usual emergency reception/admissions room may not be large enough to accommodate a large influx of wounded. A large and extendable area capable of serving as a triage department should be pre-selected. The area should be cleared of all inpatients, and be large enough to permit easy movement of casualties and medical personnel.

In certain contexts, for religious and cultural reasons, it may be necessary to have separate male and female areas for incoming mass casualties. This must be taken into account during planning.

Once triaged, patients should be transferred out of the triage department to the operating theatre or designated wards.

Different units or wards should be identified for the different triage categories. Category I patients, requiring resuscitation and immediate surgery, might be put in a pre-operative intensive care unit near the operating theatre while waiting for an operating table to free up. Resuscitation with i.v. fluids can continue here under close supervision. The Category II cohort of patients who require surgery but not on an urgent basis can be grouped in a special ward where they can receive treatment (antibiotics, analgesics, i.v. fluids, dressings) under observation while awaiting surgery.

Category III patients suffering superficial wounds could be sent to the outpatient department or an area for convalescence outside the surgical hospital. This group is composed of many patients who are lucid, frightened (because of the shooting and shelling), panicky, and often in pain. Organization and basic staffing of this space are important in order to identify and isolate such patients, give them rapid treatment, and then discharge them.

Finally, a quiet, secluded room should be made available for Category IV patients with very severe wounds who should be left to die peacefully with dignity. An i.v. line should be put up and analgesics given if appropriate.

Some provisions should be made for the well-regulated visit of friends and relatives to all seriously-injured patients. This comes under the necessary security measures to be implemented (see Section 9.13).

9.8 Equipment and supplies

A large number of stretchers or trolleys are needed at the hospital entrance during triage, to accommodate the casualties left by incoming ambulances. Blankets and sheets are required in the triage area, as are lines on which to hang infusions. Complete sets of supplies for triage should be prepared and stored in boxes or trunks which can easily be carried to the triage area from an accessible storage space.

Triage boxes should include:
- disposable latex or plastic gloves;
- venepuncture equipment;
- i.v. fluids;
- dressings, bandages, scissors;
- catheters, naso-gastric tubes;
- drugs which are likely to be needed.

Needless to say, these triage boxes should be inspected regularly for the expiration date of their components. The drugs will probably need to be stored separately. Stocks should include the appropriate antibiotics and analgesics, as well as tetanus toxoid and anti-tetanus serum.

A reasonable stock of items that will be required for the management of extra cases in the wards should be maintained. There should be adequate numbers of bed blocks for elevation of the foot of the bed, supplies of i.v. bottle holders, and supplies of dressings, plaster of Paris, splints and traction apparatus.

Again, depending on the particular circumstances of the hospital, the pharmacy may hold designated emergency triage stocks or not. The re-supply of the hospital may be disrupted because combat has interrupted the ordinary lines of supply.

Standard documentation charts and patient folders, *each with a unique number*, should be prepared and maintained in readiness. Each folder should include a triage/admission form, a fluid balance chart, and laboratory and X-ray request forms.

9.9 Infrastructure

Plans must be made to ensure adequate supplies of water, sufficient electricity, proper sanitation, and the disposal of waste. This may include special reserves of fuel for electric generators. Spare parts are also important to keep in stock; things tend to break down in the middle of an emergency situation.

The designation of roles, responsibilities and tasks is not limited to the medical personnel. Technicians and maintenance workers to run the generators and ensure the water supply must be put on a special shift system so as to be available at all times.

9.10 Services

Hospital personnel, patients and their relatives, and volunteers must all eat. Hospital linen must be washed and theatre linen re-sterilized. The kitchen, cafeteria and laundry staff and facilities must all be included in the disaster plan. Relatives are a great nuisance in the triage area but their energies can be harnessed for the general good. They may be directed to give blood and engaged as volunteer stretcher-bearers, carriers of water, cleaners and kitchen staff, etc.

9.11 Training

The hospital team should regularly practise different triage scenarios, on its own and as part of any national disaster or conflict-preparedness plan. Volunteers from the National Red Cross or Red Crescent Society, and their first aiders, may be mobilized to work in the hospital and/or play the role of the injured.

Clinical protocols and guidelines for triage and patient management must be standardized and understood by all doctors and nurses. This helps to avoid confusion and disagreements under tense and tiring circumstances.

After every triage incident and when feasible, a general meeting of the staff should be held to discuss what went well and what went wrong. This evaluation permits the fine-tuning of the triage plan, to improve it for the next mass influx.

Some individuals or even the whole team may have found the experience very stressful. An open, frank discussion about what happened and why certain decisions were made can be very beneficial, and even therapeutic.

9.12 Communication

If the hospital is part of an integrated health system, then a means of coordination and communication with other health facilities may make it possible to transfer the wounded from the overwhelmed hospital to another which has received comparatively few patients. Alternatively, other health facilities may be able to provide help in the form of additional personnel.

The plan should include the means to contact staff who are off duty, bearing in mind that if combat is occurring in the area, hospital personnel may have difficulty in getting to work. Mobile telephone systems tend to stop functioning (or are stopped by certain authorities) during urban fighting or troubles.

In an age of immediate mass communication contact with the media is inevitable, and what is happening in the hospital can be easily exploited for political purposes. The patients, hospital and staff are all protected under IHL. Obviously, military and governmental civilian hospitals are not "politically" neutral and cannot be expected to be. However, hospital personnel must act and speak publicly in accordance with their "medical" neutrality and impartiality. An "official spokesperson" for the hospital should be duly selected.

9.13 Security

Last, but certainly not least, is the safety and security of the hospital premises, patients and staff. When armed conflict results in mass casualties every wounded person transported to the hospital is accompanied, as a general rule, by two to four friends, relatives, comrades-in-arms, or bystanders who have helped in the evacuation or transport. Curious onlookers may try to enter the hospital as well. The civilian population may be in a state of panic and consider the hospital a safe place. This well-known phenomenon is called the "convergence reaction". People's fear and excitement add to the confusion and danger, especially if many in the crowd happen to be armed. Limiting the number of people entering the hospital reduces the confusion considerably.

One major factor which creates the convergence reaction and confusion is the lack of proper pre-hospital triage, as is often seen in an urban context. As mentioned in Chapter 7, the waves of ambulances transporting casualties evacuate the dead and badly mutilated first, the superficially injured, who are shouting to call attention to themselves, second, and the seriously wounded who require emergency surgery last.

Security must be ensured by having guards posted at the gate of the hospital. Only wounded persons, possibly accompanied by a close relative according to local cultural traditions, should be allowed to enter. All weapons must be left outside the gate.

Access to the triage area should be blocked off, and a guard posted here as well to keep other patients and onlookers from mingling with new casualties.

The large numbers of relatives and friends who inevitably rush to the hospital must be prevented from crowding the premises and impeding the work of the medical staff. A system for well-regulated visits by patients' friends and relatives must be set up to reduce the possibility of "friction" occurring. Under certain circumstances, board and lodging close to the hospital compound must also be organized and provided.

Whether the guards are armed or not will depend on the circumstances of the country concerned. IHL does not prohibit armed guards if their purpose is to maintain order and protect patients and personnel.

In extreme circumstances, the hospital can organize a "security lock"[4] at the hospital or triage area entrance. This forced channel allows for better crowd control, especially of armed men.

ICRC EXPERIENCE

The ICRC hospital in Kabul, Afghanistan, used commercial, steel containers with doors cut into the ends at the hospital entrance in 1992 to create a "tunnel" that allowed the filtering of all who entered. The width of the door was calculated to allow for the passage of a carried stretcher only.

The ICRC – Somali Red Crescent Hospital in Mogadishu, Somalia, in 1992 was located in a converted prison. A first set of guarded gates led to a triage area with a large tent. A second set of gates separated the triage tent from the hospital compound itself, permitting controlled entry to the hospital.

The ICRC hospital team working in the JFK Memorial Hospital in Monrovia, Liberia, in 2003 set up a so-called "triage tent" just outside the fence and gate entrance to the hospital. No medical triage was actually performed there, it was more of a "disarmament tent". The tent provided a space for disarming any combatants, stripping the wounded of their soiled clothing, weapons and ammunition, and performing basic first aid. The wounded were then carried into the triage department – unarmed – on stretchers where the clinical triage was performed.

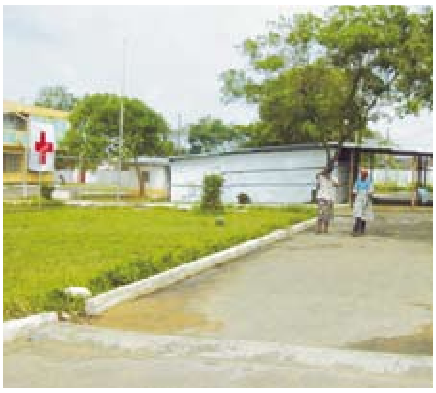
R. Aburabi / ICRC

Figure 9.7.1

Monrovia "triage tent" in front of the hospital.

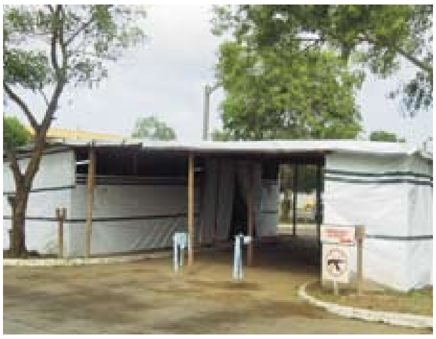
R. Aburabi / ICRC

Figure 9.7.2

"Triage tent" seen from the outside.

R. Aburabi / ICRC

Figure 9.7.3

Volunteers inside "triage tent" before the action.

R. Aburabi / ICRC

Figure 9.7.4

Volunteers after the action.

---

[4] A lock is a short section of a canal with gates at each end which can be opened or closed to change the water level, used for raising and lowering boats going through the canal. This permits control of the movement between areas at different water levels. By analogy, a security lock permits the control of people's movements between different hospital areas.

9.14 Summary of triage theory and philosophy: sorting by priority

The triage process has three components.

1. Clinical assessment to determine which patients take priority for the limited surgical time and resources available.
2. The organization and management involved in admitting large numbers of wounded to the hospital.
3. Re-assessment of the functioning of triage and its adaptation to the number of incoming wounded.

9.14.1 Triage system: a simple emergency plan organizing the personnel, space, infrastructure, equipment, and supplies

The sudden arrival of large numbers of casualties may occur at any time. Prior planning and training prevents poor performance. Unless a plan exists for the reception and triage of mass casualties, chaos will result. Hospital staff should be prepared, however, to improvise when faced with a new evolving situation.

Evaluation of hospital capacity is essential in emergency planning.

An emergency plan does not cost money; organization does not cost money. They cost time, effort, discipline and motivation.

9.14.2 Emergency hospital disaster triage plans differ and no two triage scenarios are the same

Hospital teams must regularly practise receiving a mass influx of wounded, resulting from armed conflict or natural disaster. Simulation exercises should be organized covering a variety of scenarios, and adapted to the particular circumstances of the hospital involved.

9.14.3 "Best for most" policy

Priority patients are those with a fair chance of "good survival" with the least amount of surgical work.

Triage is essential to put some order into a chaotic situation.

However good the disaster plan and extensive the training, a mass influx of casualties is *always* stressful and attended by confusion. Flexibility and adaptability of the hospital team are important. Triage is not a series of rules. It has a logic and philosophy that must be adapted to each particular situation. Triage is not a simple science; it is an art.

Please note:

Further practical information on the organization of hospital management in a triage situation is available in the ICRC manual *Hospitals for War-Wounded: A Practical Guide for Setting up and Running a Surgical Hospital in an Area of Armed Conflict* (see Selected bibliography).

In a military context, all standard war surgery manuals written by and for armed forces deal with the organization and implementation of triage under military constraints.

ANNEX 9. A Sample triage card

Triage Card No: ________________

Name: ____

Coming from: ____ Male / Female / Age ____________________________________

Date: ____ Time: ____

GSW: ☐ Mine: ☐ Fragment: ☐ Blast: ☐ Burn: ☐ Other: ________________________________

Time since injury: ____________________________________

General condition:

Pulse: __________________ BP: __________________________ Resp. rate: ____________ Consciousness: __________________

Wound assessment:

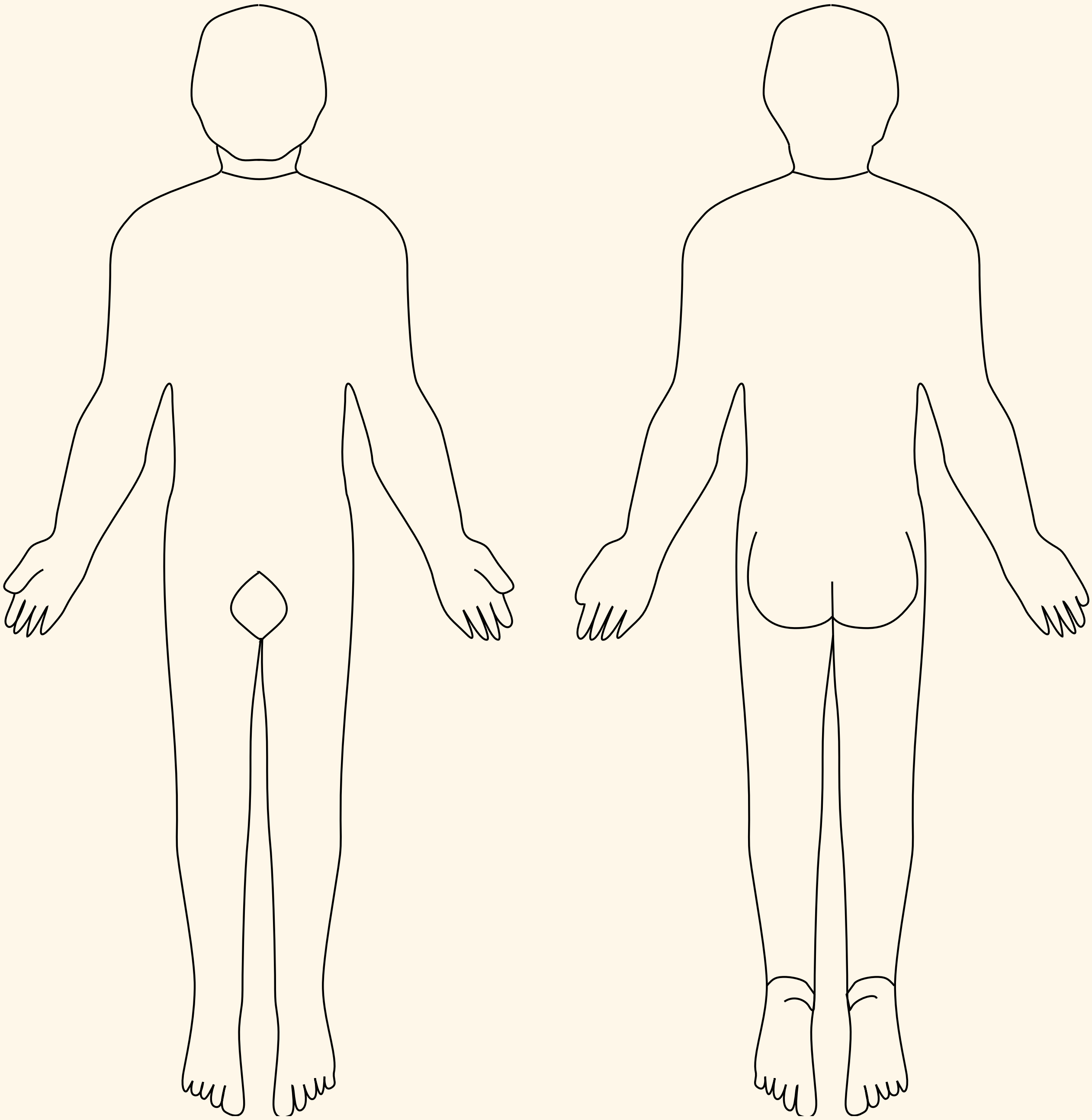

Triage Category:

I	II	III	IV

ANNEX 9. B Hospital emergency plan for a mass influx of wounded

HOSPITAL EMERGENCY PLAN FOR A MASS INFLUX OF WOUNDED										
PHASE	Influx of wounded (number of patients)	Secrity + Porters/ stretcher - bearers	TTL / TO / HN	Triage Area (ER + OPD + Physio room)	OT	Wards	Administration	Laboratory X-ray	Kitchen Laundry Tailor	Transport
1	**1 – 10**	Routine	HN informed OT informed	Routine	Stop routine Inform wards	Routine No more elective patients sent to OT	Routine	Routine	Routine	Routine
2	**11 – 20**	Extra guards at patients' entrance	TTL informed TO: INITIATE TRIAGE HN: Inform wards, OT, administration	Request triage material and staff	Keep staff on duty Check supplies	Update bed situation Inform HN/TTL	"	"	"	Standby ambulance
3	**21 – 30**	Visitors leave Visiting hours suspended	Call in on-call nurses Ensure that all staff have breaks	Call in extra staff Physio staff on standby to help in triage or in wards	Extra surgical team and staff standby	"	Keep pharmacy staff on duty	"	Prepare food and drinks for OT staff if needed	"
4	**31 – 40**	"	HN: Reassess bed situation TTL/:Monitor OT situation	"	"	Assess need for additional staff Inform HN	Keep purchaser and storekeeper on duty	Call in extra staff if needed	Laundry: priority is OT linen	Standby bus for staff transport
5	**41 – 50**	Call in extra porters/ stretcher-bearers	TTL/HN: REASSESS Define needs for staff, beds, supplies	"	Check supplies Open reserve linen stocks	"	Monitor needs for staff	"	Open emergency food stocks	"
6	**> 50**	"	REVIEW AGAIN	"	"	"	Provide resting area and sleep accommodation for staff on duty	"	Provide food and drinks for all staff on request if needed	"

ER: emergency reception room

HN: Head Nurse

OPD: outpatient department

OT: operating theatre + sterilization

TO: Triage officer

TTL: Triage Team Leader

Chapter 10

SURGICAL MANAGEMENT OF WAR WOUNDS

10. SURGICAL MANAGEMENT OF WAR WOUNDS

10.1 Introduction 213

10.2 Complete examination 214

10.3 Preparation of the patient 215

10.4 Examination of the wound 216

10.5 Surgical treatment 216
10.5.1 Technique of wound debridement 217
10.5.2 Skin 218
10.5.3 Subcutaneous fat 218
10.5.4 Fascia and aponeurosis 218
10.5.5 Muscle 219
10.5.6 Haematoma 220
10.5.7 Bone and periosteum 220
10.5.8 Arteries, nerves and tendons 220

10.6 Retained bullets and fragments 221

10.7 Final look and haemostasis 222

10.8 Wound excision: the exceptions 223
10.8.1 Management of minor Grade 1 wounds 223
10.8.2 Serial debridement 223

10.9 Leaving the wound open: the exceptions 224
10.9.1 Head, neck, scalp and genitals 224
10.9.2 Soft tissues of the chest (sucking chest wound) 224
10.9.3 Soft tissues of the abdominal wall 224
10.9.4 Hand 224
10.9.5 Joints 224
10.9.6 Blood vessels 224

10.10 Dressings 225
10.10.1 The exceptions 226

10.11 Anti-tetanus, antibiotics, and analgesia 226

10.12 Post-operative care 226

10.1 Introduction

There are many factors that determine the final outcome of the management of war-wounded patients:

- the actual injury – the clinical significance of the wound is a function of the severity of tissue damage and the anatomic structures involved i.e. size and site;
- the general condition of the patient – nutritional status, dehydration, concomitant diseases, host resistance, etc. ;
- pre-hospital care: protection, shelter, first aid, triage, evacuation time;
- resuscitation, especially in the presence of haemorrhagic shock;
- hospital triage;
- surgery;
- post-operative nursing care;
- physiotherapy and rehabilitation.

However, the single most important act that the surgeon performs for the great majority of casualties is wound excision or debridement.[1]

The basic principles of trauma wound management have been known for a considerable period of time. Ibn Sinna commented on this (Figure 10.1), as did Alexander Fleming one thousand years later.

Å. Molde / ICRC

Figure 10.1

“Damaged tissues must be removed in time.” Ibn Sinna, *Qanun fi al-Tib*.[2]

> “The severity of these [war] wound infections is merely the result of the very extensive destruction of the tissues by the projectile, thus furnishing an admirable culture medium for the bacteria out of reach of the natural protective forces of the body, and if it were possible for the surgeon to remove completely this dead tissue I am quite sure the infections would sink into insignificance.”
>
> Alexander Fleming[3]

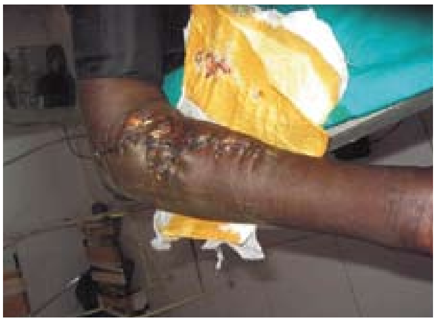

K. Barrand / ICRC

Figure 10.2.1

Wound sutured primarily to “close the hole”, without debridement.

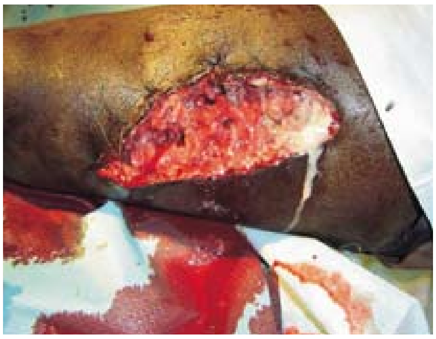

K. Barrand / ICRC

Figure 10.2.2

Sutures removed: pus pours out.

Often, ICRC health personnel have witnessed inexperienced medical staff in a low-income country dealing with severely war-wounded patients for the first time. Their initial reaction is to try to stop the bleeding by “closing the holes”. If suture material is available, they stitch the wounds shut without excising dead tissue; if not, they plug the wounds with gauze compresses. Soon, infection develops. This is treated by a frequent change of dressing and antibiotic anarchy, wasting already insufficient hospital supplies. The end of the story may be a slow recovery with a major disability, or, more frequently, major complications often leading to the death of the patient accompanied by a deep sense of frustration among the medical staff.

The basic management of war wounds is an art based on sound scientific principles. A good knowledge of wound ballistics helps the surgeon assess more accurately the extent of tissue damage and of the necessary surgery; it will not explain every wound or prescribe specific treatment for every patient. In only a few cases will knowledge of the specific causative weapon be of any clinical use.

Wound ballistics teaches us that the cavitation effect of projectiles conveys pathogenic organisms, pieces of dirty skin and clothing and dust into the depths of the wound. Anti-personnel blast mines drive pieces of the shoe or bones of the foot,

---

[1] *“Débridement”* is a French term originally used to mean the removal of a constriction, or to unbridle, and thus to promote drainage and relieve tension by incision. The English word “excision” means to cut away or cut out. A confusion of terminology occurred at the Inter-Allied Surgical Conference for the Study of War Wounds held in Paris in 1917, and the historic French expression came to mean a wound excision – the removal of dead tissue – for English-speaking surgeons.
A further misunderstanding may be caused by the fact that excision, when used with reference to cancer surgery, tends to mean *“excision en bloc”* with a wide margin of healthy, surrounding tissue. In this book, the expressions “excision” and “debridement” are used interchangeably and are understood to mean the cutting away of dead or devitalized tissue.

[2] Ibn Sinna – Avicenna (980 – 1036 CE), Persian physician and philosopher, author of *Qanun fi al-Tib* (The Laws of Medicine).

[3] Alexander Fleming (1881 – 1955), British physician, discoverer of penicillin.
Fleming A. On the bacteriology of septic wounds. *Lancet* 1915; **186**: 638 – 643.

gravel, soil, leaves and grass, and pieces of the weapon into the proximal tissues. Although the wounds are contaminated, infection does not set in for the first 6 – 8 hours. Ideally, therefore, war wounds should be debrided within six hours; this is not often the case.

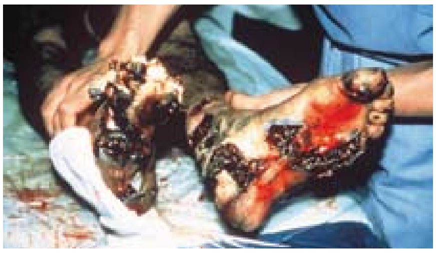
ICRC

Figure 10.3

Anti-personnel landmine injury: war wounds are dirty and contaminated.

Old lessons for new surgeons

War wounds are dirty and contaminated, from the moment of injury.

Whatever the weapon, a thorough examination of the patient and the wound(s) is essential.

Examine the patient:
- initial examination and resuscitation;
- complete examination to identify all open wounds and any closed injuries.

Examine the wounds and damaged organs:
- to operate or not to operate;
- to establish priorities if more than one surgical procedure is required;
- to plan the operation.

Please note:

The treatment of mismanaged and neglected wounds is dealt with in Chapter 12.

10.2 Complete examination

All wounds involve soft tissues, and many of them will be complicated by damage to other structures. *War wounds are often multiple* and the *pathology often multiple*: a bomb explosion can simultaneously cause primary blast injury, penetrating metallic fragments, blunt trauma and burns. Close and complete examination of the patient is important to determine the site and size of *all* wounds present, in order to determine which wounds require surgery and to best plan the sequence of operations.

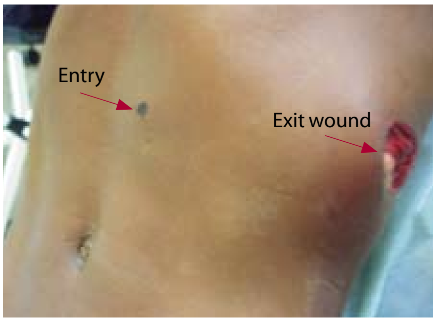

H. Nasreddine / ICRC

Figure 10.4

Entry and exit thoraco-abdominal wounds: think anatomy!

As mentioned in Chapter 8 the surgeon should try to identify the path of any projectile. The wound may perforate the body part through-and-through, or the projectile may be retained without an exit. Entry and exit wounds may be at the same or different levels. The surgeon should visualize the anatomic structures that lie along the probable projectile trajectory, which may include a vital cavity: brain, thorax, or abdomen. The surgeon must think "anatomy".

The examination may reveal a tense, tender swelling in a through-and-through wound to the calf, which indicates haematoma and severe muscle damage; by contrast, the same exit and entry wounds with soft and relaxed tissues signify minimum injury. The surgeon must think "pathology".

Fractures and peripheral vascular and nerve damage should be identified. If there has been delay in evacuation, a *bruit* or vascular murmur may be heard and a "thrill" felt, signifying a pseudoaneurysm or arteriovenous fistula. Peripheral motor function and sensation are tested. Neurapraxia is more common than a severed nerve.

Most wounds involving vital organs should already have been identified during the initial examination for airway, breathing and circulation. Nonetheless, the full-body palpation of a complete examination may reveal the small entry wound of a tiny fragment that has penetrated the meninges, pleura or peritoneum without immediately affecting the vital functions.

A full assessment may require X-rays as well – with the exception of through-and-through soft-tissue wounds, which do not require them. One body region above and below any entry or exit wound should be filmed. Deformation or fragmentation of a bullet, showing the characteristic "shower of lead particles", is a good indication of severe tissue damage. (Figures 10.5, 3.35 and 4.5). Many foreign materials, it should be noted, are *not* radio-opaque: pieces of shoe or clothing, mud, leaves and grass, plastic fragments from some landmines, for instance. On the other hand, it may be difficult to differenciate a projectile from the normal radio-opacity of some anatomic features (Figures 8.4.2 and 14.9.1). Major bone comminution also implies severe soft-tissue damage. Having radiography helps, but is not an absolute necessity for diagnosing fractures.[4]

Note that the presence of air in tissues on X-ray does not necessarily mean the gas of clostridial myositis. The temporary cavity associated with high velocity missiles often leaves behind palpable and radiographic intrafascial and intramuscular air in healthy tissue some distance from the wound. This usually indicates severe tissue damage. The diagnosis of gas gangrene is a clinical one and the radiograph is not pathognomonic: *vide* Figures 10.6 and 13.2.

More sophisticated diagnostic means can be used, of course, depending on the equipment and professional expertise of the hospital in question.

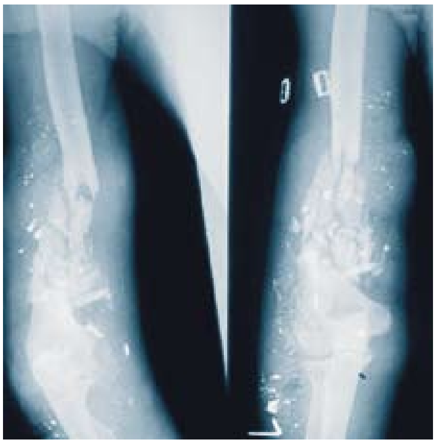

R. Coupland / ICRC

Figure 10.5

Massive comminution of humerus showing characteristic "shower of lead particles" after bullet fragmentation.

10.3 Preparation of the patient

Not only are war wounds dirty and contaminated but battlefields are dirty places as well. The wounded do not have access to basic sanitary facilities and every precaution should be taken to meet fundamental hygiene standards. The majority of casualties have wounds to the limbs and are haemodynamically stable. All stable patients should be put through a warm shower upon admission; dressings are changed as necessary for examination and triage. Only critical cases go straight to the operating theatre.

In the operating theatre, the surgeon should prepare the patient's position – in coordination with the anaesthetist and nursing staff – for the sequence of the operation before starting. A pneumatic tourniquet is very useful for limb wounds, especially before removal of a field dressing on an obviously serious injury.

In the case of multiple soft-tissue wounds, those on the posterior aspect of the trunk and limbs should be dealt with *before* those on the anterior aspect, unless there is clinical evidence of a vascular injury. Many wounds to the back have been *"forgotten"* after much work on anteriorly situated wounds, and it is easier for the anaesthetist to finish off the operation with the patient supine.

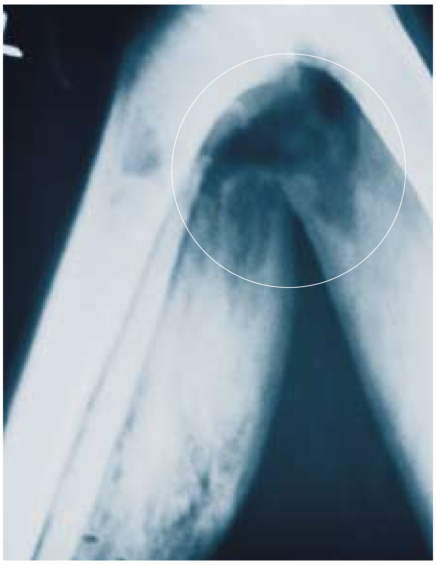

ICRC

Figure 10.6

X-ray showing air in tissues in a patient not suffering from gas gangrene.

Under anaesthesia, dressings and splints are carefully removed. The skin over a large surrounding area, including the whole circumference of the limb or torso, is cleansed with soap and water and a brush, shaved, dried, and then painted with povidone iodine. The wound is irrigated copiously.

Sterile sheets are put in place. Drapes with holes should only be used for the smallest and most superficial wounds. Most wounds will have to be extended and the operative field enlarged as a result.

The wound is then thoroughly irrigated again to remove surface dirt and debris. Under optimal conditions, sterile normal saline should be used. Under conditions of limited resources, potable water from the tap can be resorted to: "If you can drink it, you can put it into a dirty wound". Groundwater may be treated with sodium hypochlorite to reach a concentration of 0.025 % (5 ml of bleach in one litre water) if necessary.

Only in extreme cases of impending asphyxia or exsanguinating haemorrhage should the basic rules of hygiene and sterile technique be broken.

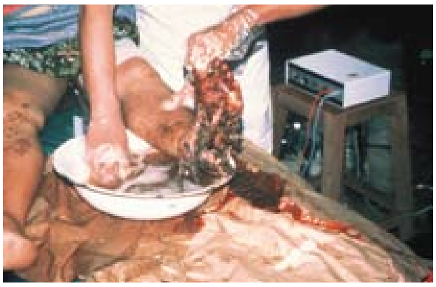

ICRC

Figure 10.7

Washing an anti-personnel landmine injury prior to surgery.

---

[4] Many ICRC surgical teams have had to operate in emergency situations with *no* X-ray equipment.

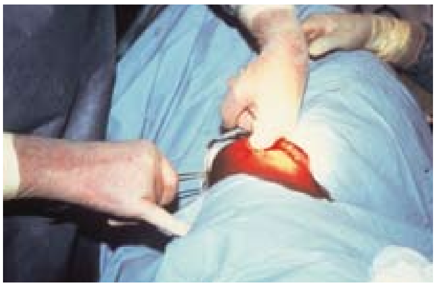
ICRC

Figure 10.8

Finger exploration of the wound.

10.4 Examination of the wound

Following the initial careful examination of the patient, full assessment of the wound(s) may require finger exploration in theatre under anaesthesia. In this era of HIV and Hepatitis B and C, the surgeon should take great care not to suffer injury from sharp bony edges.

Old lessons for new surgeons

The best instrument for wound exploration is the surgeon's (protected) finger.

The surgical work required to treat a wound depends on its site and size: i.e. the location and amount of tissue damage. Knowledge of ballistics allows only for an understanding of wounding possibilities, but knowing that certain weapons have greater wounding potential does not help the surgeon in this assessment. With the exception of anti-personnel blast mines, even knowing which weapon was used is often of no practical use. Many wounds are the result of full metal jacket bullets that ricochet, creating large injuries similar to a dum-dum bullet (see Chapter 3).

The most significant component of tissue damage is the permanent wound cavity, yet energy loss by the projectile and tissue disruption along the track is not uniform, resulting in a "mosaic of tissue damage". Injury can also occur well beyond this, owing to stretch and shear during temporary cavitation: late ischaemic necrosis in the wall of hollow organs; or intimal detachment and vascular thrombosis of vessels that still look intact from the outside. The time between injury and treatment and the possible beginning of sepsis, must also be taken into account.

War wounds are never identical, even if caused by the same weapon.

There is an art to the assessment of tissue damage that is learnt largely through experience. It is important neither to under-treat – causing sepsis and even death – nor to over-treat – resulting in excess loss of normal tissue and increased disability.

This art does have a scientific basis, nonetheless. We have seen that the Red Cross Wound Score classifies penetrating weapon wounds into different categories according to the degree of tissue damage (Grade) and the structures injured (Type). This Score helps determine the extent of surgery necessary. Careful examination is important because not all wounds require surgical excision.

10.5 Surgical treatment

The surgical treatment of a major wound is performed in two stages:

- *wound debridement*, leaving the lesion wide open, without any suturing of skin or deep structures;
- *delayed primary closure* 4 – 5 days later.

The surgical treatment of most war wounds is a staged process involving two main procedures, the first being wound debridement or excision. The resulting wound is left unsutured. The open wound then undergoes delayed primary closure (DPC) after 4 – 7 days, once the exudative phase of trauma inflammation has subsided and the proliferative phase begun. In practice, DPC is usually performed after 4 – 5 days. Closure may be by simple direct suture or require sophisticated techniques of grafting and reconstruction (see Chapter 11).

10.5.1 Technique of wound debridement

Basic principles of wound debridement

1. Stop haemorrhage.
2. Make adequate skin incisions and fasciotomies.
3. Remove dead and severely contaminated tissues to prevent/control infection.
4. Leave the wound open – unsutured.
5. Re-establish physiological function.
6. Handle gently and treat the tissues with respect, as always.

Control of haemorrhage is the first priority. In a large wound with heavy bleeding, blind clamping in the depths of a blood-filled cavity is mentioned only to be condemned. Direct local pressure should be exerted while proximal and distal control of the vessels by standard surgical exposure is performed.

For management of vascular injuries, see Volume 2.

After haemorrhage, the greatest danger to the patient is sepsis. Surgical infection requires a culture medium, which in the case of weapon wounds is a mixture of dead muscle, haematoma, bone fragments, dirty skin, foreign material (pieces of cloth or shoe, mud, gravel, leaves, dust, the missile, etc.), and, at times, extrinsic bone (bone fragments which come from the wound in another part of the body or another person and whose penetration, as secondary missiles, then creates a new wound). Injury caused by extrinsic bone fragments is seen with bomb explosions and APM.

Wound debridement is the process whereby this dead and damaged tissue, grossly contaminated with bacteria and debris, is completely cut away. This leaves an area of healthy tissue with a good blood supply and capable of combating residual surface contamination, provided the wound is not closed. Excess excision of healthy tissue, however, will increase the potential for deformity and disability.

Good vascularisation and oxygenation of the wound requires relief of tension within the wound and good drainage of the inflammatory exudate. This is provided by adequate incision of the skin and fascia, and leaving the wound unsutured.

Wound debridement involves incision and excision.

A basic instrument set is all that is needed in the vast majority of cases: scalpel, Metzenbaum (tissue) and Mayo (suture) scissors, toothed dissection forceps, bone curette, six haemostats, and retractors. Diathermy is not necessary; absorbable ligature material is preferable.

For the young surgeon, or one without experience of war wounds, it is best to excise the wound layer-by-anatomic-layer, proceeding from superficial tissues to deeper ones in order to best visualize the anatomy and pathology. At all times the surgeon should anticipate the anatomic structures present. Tissue damage, haematoma, and oedema can easily *alter the visible anatomy*, and camouflage important structures.

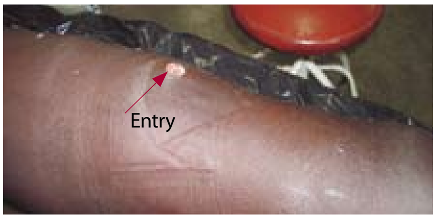

F. Jamet / ICRC

Figure 10.9.1

Patient X: entry wound on anterior aspect of thigh.

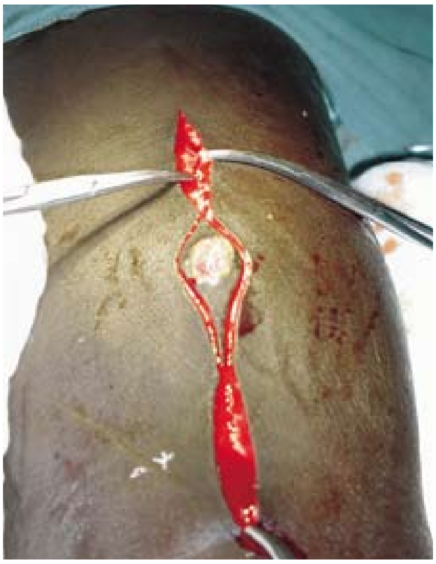

F. Jamet / ICRC

Figure 10.9.2

Patient X: longitudinal extension of the skin incision.

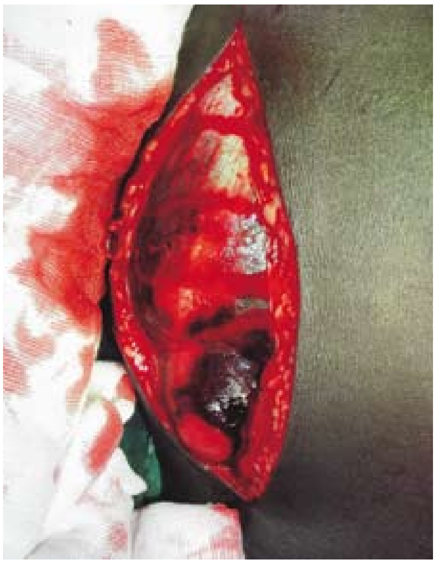

F. Jamet / ICRC

Figure 10.9.3

Patient X: opening of the fascia throughout the length of the skin incision. Note the contused and necrotic muscles.

10.5.2 Skin

Skin is elastic, with a good blood supply, very resistant to damage and remarkably viable. It should be treated conservatively. Only skin that is grossly pulped should be cut away. Usually no more than 2 – 3 mm of the skin edge need be removed at both entry and exit sites.

This is followed by as generous an incision as necessary of healthy skin in order to gain access to the depths of the wound (Figure 10.9.2). Small entry and exit wounds may hide considerable internal injury. The most common error is to attempt to make a wound excision through a small entry or exit wound. In limbs, the incision should be made in the longitudinal axis, but not over subcutaneous bone, and at flexion creases it should curve in the usual way.

The commonest mistake is to attempt a wound excision through small entry or exit holes or to make incisions too short.

This extension of the skin wound not only permits better visualization, but also allows for the proper decompression of deeper tissues and their subsequent drainage.

10.5.3 Subcutaneous fat

Subcutaneous fat has a poor blood supply and is sticky, easily holding heavy contamination. This layer should be excised generously, 2 – 3 cm all the way around the original traumatic wound.

10.5.4 Fascia and aponeurosis

Shredded fascia should also be trimmed. Large amounts of damaged muscle may lie underneath a small hole in the fascia; therefore, the muscular compartment should be opened up by a large incision of the deep fascia parallel to the muscle fibres along the entire length of the skin incision (Figure 10.9.3). This essential step allows wide and deep retraction to expose the depths of the wound. It may be necessary to add transverse cuts to the deep fascia to improve access.

Post-traumatic oedema of the wound can easily cause compartment syndrome, compromising the local circulation with resultant necrosis of the muscles. The fascial incision is left open to allow oedematous and congested muscle to swell without tension, so as to avoid interference with the blood supply and to promote drainage of the inflammatory exudate and haematoma.

After haemorrhage control, the relief of tissue tension is the most important step in wound debridement.

Deep to the fascia, the gloved finger is the best and gentlest probe to follow the track and estimate the extent of damage. Again, particular attention should be paid to the sharp edges of any fractured bone.

Fasciotomy

Compartment syndrome can occur in any fascial space, but is seen most commonly in the lower leg. Great care should be taken when dealing with any penetrating wound below the knee, with or without tibial fracture.

If there is any suspicion of compartment syndrome, decompression must be performed without delay.

For information on the technique of fasciotomy of the lower leg, see Volume 2.

10.5.5 Muscle

Dead muscle is the *ideal medium* for the development of *clostridial* infection leading to gas gangrene or tetanus as well as for the growth of many other bacteria. The track of the missile through the muscles must be opened up, layer by layer, to be properly visualized. It is vital that all grossly contaminated, obviously necrotic and detached muscle lining the track be excised.

All dead or heavily contaminated muscle must be removed.

Warning!

Muscles that are completely transected contract *away* from the wound cavity. The retracted muscle bellies must be found for inspection and excision. During wound exploration, division of uninjured muscle in a transverse direction should be avoided.

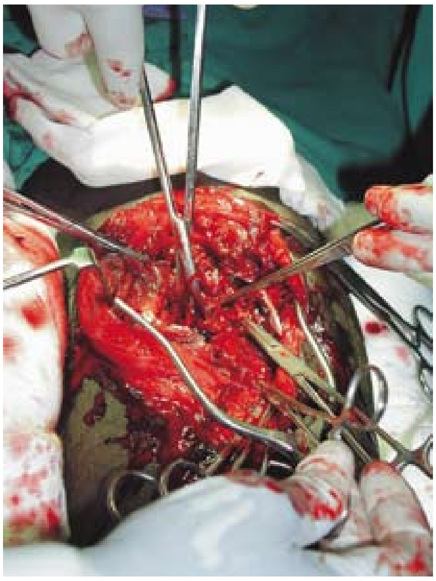
F. Jamet / ICRC

Figure 10.9.4

Patient X: wound cavity completely opened and excised.

However, not all injured muscle will undergo necrosis. How to differentiate between muscle that is injured but will heal from that which is not viable? Classically, reference is made to the four Cs of muscle viability:

- colour,
- consistency,
- contractility, and
- capillary bleeding.

All muscle that is not healthy and red, that does not contract when pinched or bleed when cut, must be excised until healthy, contractile, bleeding muscle is found.

However, confusion may arise because of certain pathological changes described below.

- As mentioned in Chapter 3, wound ballistics studies have shown that there is an intense but transient vasoconstriction lasting several hours, followed by the reactive local hyperaemia of the inflammatory response. Consequently, it is not always necessary to "cut until it bleeds". The surgeon must therefore take into account the time since wounding.
- However, minimal excision is often followed by necrosis of the muscle when the wound is looked at several days later. What is irreversible is not always immediately apparent. The wound evolves and these changes demonstrate the different stages in the "life-history" of the wound[5] and should be well understood by the surgeon.
- The surgeon should also understand that shock or the use of a tourniquet can change the colour or the capacity of muscle to bleed and that contractility is affected by hypothermia and paralysing anaesthetic agents.

The criteria of the 4 Cs are very subjective and a function of the surgeon's experience, but are nonetheless the best clinical measure available. The colour and consistency – texture – of the muscle should be noted. To test for contractility, 2 cm^3 lumps of muscle should be picked up and pinched with a forceps and, if they do not contract, excised using scissors or knife. Note if the cut edge of the muscle lump bleeds. Cutting away lumps larger than 2 cm^3 may result in the inadvertent removal of healthy tissue. Meticulous technique is the key.

Intermuscular fascia should be excised if blood-stained and contaminated.

---

[5] Pearson W. Important principles in the drainage and treatment of wounds. *Lancet* 1917; **189**: 445 – 450.

10.5.6 Haematoma

The presence of a large haematoma generally implies that a major vessel has been damaged. Dislodging the haematoma can result in sudden heavy blood loss. It is wise to be prepared for vascular control before a haematoma is evacuated. If working under a pneumatic tourniquet, the anatomical structures must be clearly identified.

10.5.7 Bone and periosteum

The Haversian vascular system of bone is a fragile one. Fragments of bone with no attachment to periosteum or muscle are already sequestrated and should be discarded, but any bone still attached should be retained. Exposed medullary bone should be curetted back to firm marrow. Any bone left *in situ* must be cleaned of dead muscle and foreign material, dirty bone ends are trimmed by a bone-nibbling forceps (*rongeur*). *Bone defect* is not important at this stage, the *wound* is of paramount importance, and every attempt should be made to avoid infection, whose development will only lead to greater bone loss. The danger of non-union from the removal of multiple bone fragments is overrated.

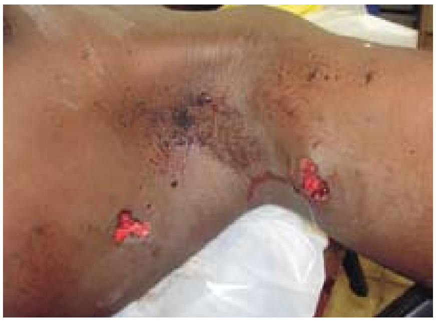
H. Nasreddine / ICRC

Figure 10.10.1

Patient Y: gunshot wound to arm and chest.

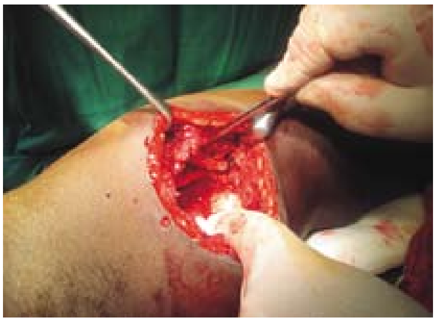
H. Nasreddine / ICRC

Figure 10.10.2

Patient Y: wound track opened up.

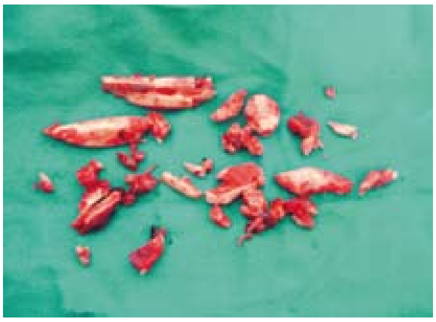
H. Nasreddine / ICRC

Figure 10.10.3

Patient Y: detached bone fragments removed.

Periosteum, on the other hand, is resilient and has a rich blood supply: it plays the predominant role in new bone formation. Its debridement should be conservative and restricted to obviously dirtied and contaminated edges.

ICRC practice is to use the simplest form of bone immobilization at the first debridement; this is usually a plaster of Paris (POP) posterior slab or some form of skeletal traction. Seldom is external fixation required at this stage and internal fixation is proscribed in ICRC practice. More definitive bone immobilization is decided during delayed primary closure of the wound.

For information on indications for different techniques of fracture immobilization and repair of bone defect, see Volume 2.

10.5.8 Arteries, nerves and tendons

As aforesaid, bleeding should be controlled if a major artery to the limb is damaged and either immediately repaired or replaced by a saphenous vein graft or temporary stent if a limb is to survive. The surgeon should pay particular attention to the possibility of a vascular injury near severely comminuted fractures with multiple fragments.

All nerves must be preserved as far as possible. Large nerves are resistant to section, although they may suffer neurapraxia. They are often the only structure left traversing a wound cavity. If they are found to be damaged, the site and degree of damage should be recorded. A non-absorbable suture may be placed in the proximal and distal ends and pulled together to facilitate their identification at a future operation. Exploration of the implicated nerve should be attempted during wound debridement *only* if it does not involve opening up healthy tissue planes.

Tendons should be trimmed and only grossly destroyed fibres removed. Severed tendons, if important and requiring later repair, should be marked with a non-absorbable suture as with nerves.

No attempt should be made at primary tendon or nerve repair, as success is unlikely in these grossly contaminated wounds. The failure of immediate repair will only make subsequent efforts that much more difficult. Also, repairs take time and energy at the "wrong time of the day"; they are best dealt with as planned elective procedures. Nerves and tendons should nonetheless be protected from extended exposure, by covering them with muscle or skin flaps or wet dressings.

For techniques on artery, nerve, and tendon repair, see Volume 2.

10.6 Retained bullets and fragments

Obviously, if the surgeon comes across a projectile during wound debridement then it should be removed, but healthy tissue should not be dissected in an attempt to find one. Otherwise, there are two conditions that require *immediate removal* of bullets and fragments, and these are related to specific proven risks and complications.

1. Projectile situated in a synovial joint – the piece of metal will cause pain, disability and progressive destruction of the joint cartilage though a mechanical action, and possible toxicity if made of lead: it should be removed as part of the wound debridement of the acute injury (Figures 10.11.1 and 10.11.2).
2. Risk that the projectile may cause erosion of an important structure (usually a major blood vessel) with the possibility of major haemorrhage or embolization (Figures 10.12.1 and 10.12.2). Should the surgeon suspect a pseudoaneurysm or an arterio-venous fistula then an operation to treat these pathologies involves the removal of the foreign body.

   Whether this removal is performed as an acute or planned elective procedure will depend on the exact anatomic location and endangered structure, the haemodynamic stability of the patient, the availability of diagnostic and operative equipment, and especially the *expertise of the surgeon*. The risks of a major procedure (removal of a bullet in the mediastinum, brain, etc.) whose morbidity may be quite high (especially in inexperienced hands) must be weighed against the benefits where the overall incidence of complications is low: for further indications and details, see Chapter 14.

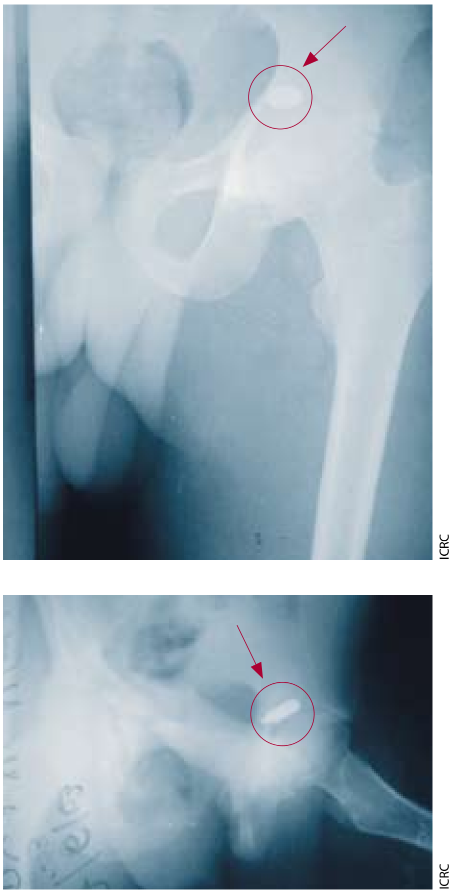

ICRC

ICRC

Figures 10.11.1 and 10.11.2

Retained bullet in the hip joint.

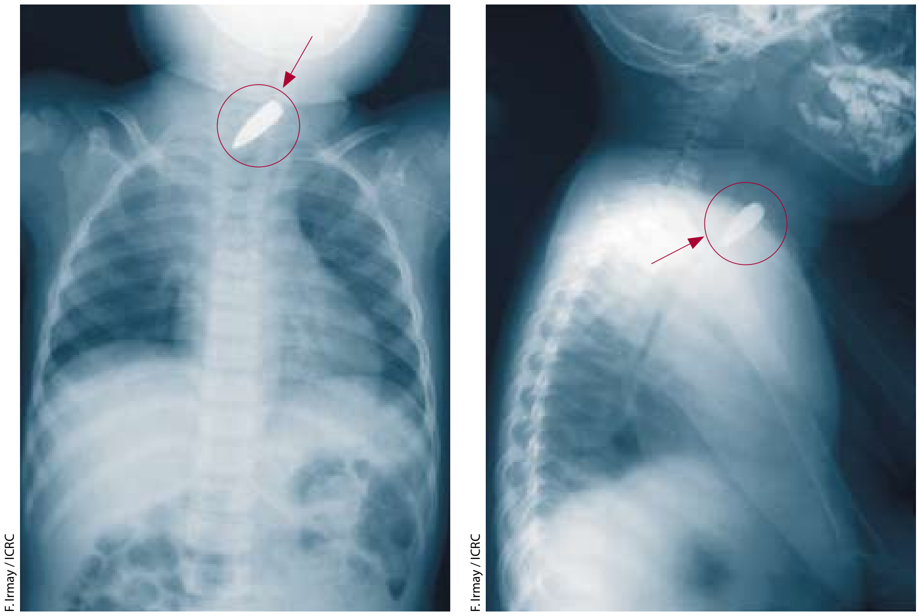

F. Irmay / ICRC

F. Irmay / ICRC

Figures 10.12.1 and 10.12.2

Retained bullet in the upper mediastinum.

10.7 Final look and haemostasis

The edges of the wound should be retracted and blood clot, dirt and missile fragments removed from the sides and depths. Gentle and copious irrigation under low pressure, preferably with normal saline at this point, will wash out any residual debris and clot and dilute any bacterial load. A plastic bottle with holes cut into the top squeezed manually with both hands provides sufficient pressure; depending on the size of the wound cavity, one to three litres of saline is used. Very large and complicated fractures may require more until the wound "looks" clean.

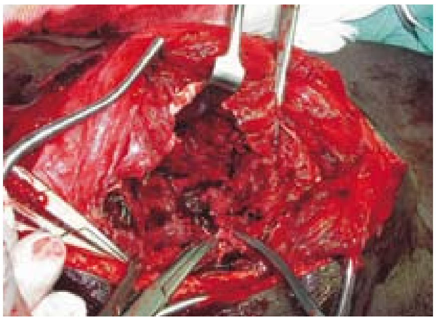
F. Jamet / ICRC

Figure 10.13
Patient X: final appearance of the wound cavity.

All the structures in the wound cavity should now be visualized and identified (Figure 10.13). The surgeon should carefully explore the wound with the finger to identify any foreign bodies or unexpected extensions of the wound.

- *Do not* open fresh planes in healthy tissue.
- *Do not* explore unnecessarily for metallic fragments.
- Remove fragments of clothing, in-driven dirt and vegetation.

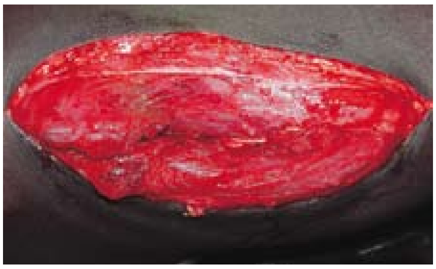
F. Jamet / ICRC

Figure 10.14
Patient X: wound is left open.

After removal of the pneumatic tourniquet, bleeding should be controlled by pressure with gauze and by fine absorbable ligatures. Electric cautery (diathermy) is best avoided as it leaves dead burnt tissue behind, which is more harmful than the foreign body of an absorbable knot.

The wound should be left wide open (Figure 10.14). There is no use in "putting in a few sutures" to partially close the wound, "just to bring the edges together a bit". This will defeat the purpose of allowing for extensive wound decompression and drainage and, after the development of reactive oedema, a loose stitch becomes tight. In addition, although the wound may look clean, it is not sterile. Bacteria and microscopic debris still exist in the wound and will only be expelled with the post-traumatic inflammatory exudate if drainage is adequate.

Should a drain be put in place? If the wound is relatively shallow and left wide open, there is no need for a drain. If there are deep pockets in the wound, which cannot be fully opened up because of anatomical constraints, then a soft penrose or corrugated rubber drain may be required. Counter-drainage through a stab wound in a dependent part might be more useful.

These procedures are *not new*. They are the rules of all septic surgery to be found in all standard surgical reference texts. "Never close infected wounds. Do not close contaminated wounds or clean wounds that are more than 6 hours old. Systematically perform wound toilet and debridement and saline irrigation until the wound is completely clean. Perform delayed primary closure as a second procedure."

Old rules for new surgeons

In war wounds, the rules of septic surgery apply.

10.8 Wound excision: the exceptions

10.8.1 Management of minor Grade 1 wounds

Many Grade 1 soft-tissue wounds, according to the Red Cross Wound Score, can be treated conservatively. Examples include:

- perforating bullet wound with small entry and exit (narrow channel wound) without swelling of intervening tissues (haematoma/oedema) or other signs of injury to important structures (Figure 3.29.1);
- multiple superficial wounds due to "peppering" with tiny fragments of obviously low velocity and low kinetic energy (e.g. hand grenade) as shown in Figure 10.15.1.

Some superficial Grade 1 soft-tissue wounds will require only simple local wound toilet. This will suffice with many low-energy small fragment wounds where the potential culture medium is so small that the body's normal defence mechanisms can deal with it. They can be cleansed with soap, water, and a disinfectant and a simple dry dressing applied; the small wound is left open to heal by secondary intention. This is especially the case if antibiotics can be administered early (see Chapter 13). This ICRC experience has been confirmed by others.[6]

Excision of entry and exit wounds under local anaesthesia may be necessary to promote free drainage in other superficial Grade 1 wounds. This may be accompanied by syringe injection of normal saline for irrigation of the wound track with/without a drain. Some such wounds, however, will require a full surgical exploration and excision, particularly in the one well-known instance where the type of weapon is of utmost importance: anti-personnel blast mines. Even small puncture wounds from landmines can be loaded with mud and grass or the mine casing, all of which must be removed (Figure 10.16).

Heterogeneity of war wounds: there is no single treatment that applies to all wounds. Wounds must be considered by Grade and Type.

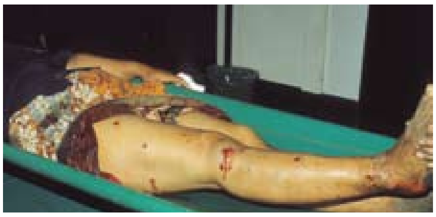
R. Coupland / ICRC

Figure 10.15.1

Superficial "peppering" by grenade fragments.

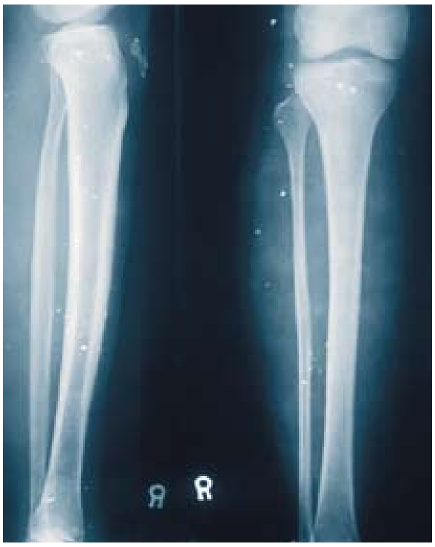

R. Coupland / ICRC

Figure 10.15.2

There is no penetration of the joint and no vascular injury – the wounds require simple wound toilet.

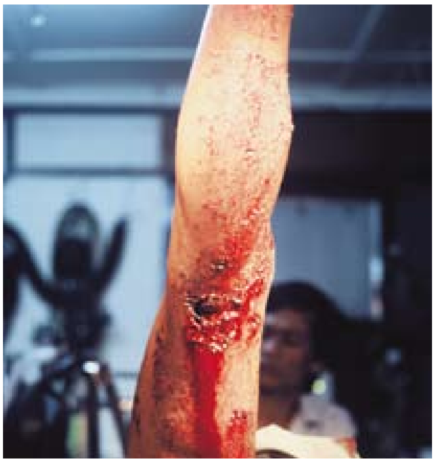
ICRC

Figure 10.16

All APM fragment wounds – large and small – require debridement.

10.8.2 Serial debridement

For some large wounds the line of demarcation between dead tissue and damaged, but viable, tissue is not clear. The life-history of a wound is such that apparently clean and living tissue may become necrotic after a few days, especially if there has been a delay between injury and debridement; and if the surgeon is not experienced in this type of surgery. The idea here is to debride obviously dead tissue; otherwise excise conservatively, and then re-examine the wound in theatre after 48 hours.

This method is called a "serial debridement", i.e. a wound excision over several sessions, and it should be *planned* as such. The surgeon should make a conscious decision that, "in this patient, because I am not certain of the viability of the remaining tissues and excessive excision of normal tissue would cause deformity or compromise function, I shall perform a second debridement later" (Figure 10.17).

In armies with very efficient means of evacuation and sufficient human resources, serial debridement may be the preferred method of treatment and takes place at different hospitals with different operating surgeons along the chain of casualty care.

Serial debridement cannot be considered a standard method of treatment in a situation of mass casualties and poor or no evacuation capacity. It requires sufficient personnel and resources to handle the number of wounded and to allow for a second look and follow-up. More often, the surgical workload is so great, or the tactical situation such, that serial debridement is not possible. The surgeon must treat each wound excision as a definitive one. In this instance, if in doubt about tissue viability, it is wiser to excise.

---

[6] Bowyer GW, Cooper GJ, Rice P. Small fragment wounds: biophysics and pathophysiology. *J Trauma* 1996; **40 (Suppl.)**: S159 – S164.

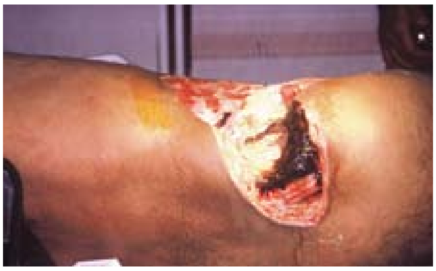

ICRC

Figure 10.17

Serial debridement of a large wound: the line of demarcation of necrotic tissue has now become apparent.

The practice of serial debridement should not be confused with an incomplete or failed wound excision. In the latter case, the patient returns to theatre for delayed primary closure after five days and the wound is found to be infected with remaining necrotic tissue. It is not ready for suture and requires re-debridement.

10.9 Leaving the wound open: the exceptions

As is usually the case in surgery, there are exceptions where wounds may, or even should, be closed.

10.9.1 Head, neck, scalp and genitals

The excellent blood supply and minimal soft tissue of these structures usually allows for immediate primary closure after wound excision. Only in the presence of severe contamination, or if in doubt, might it be wiser to leave these wounds open.

In maxillo-facial wounds, the oral mucosa is an exception in all respects and every attempt should be made to close it primarily.

Machete or *panga* wounds, especially to the face or scalp, are *not* incised wounds, but rather a mixture of crush and laceration. Dirty connecting tissue is found to underlie an overhanging edge of skin and galea aponeurotica. If less than 6 hours since wounding, immediate primary closure, after full debridement, with a subcutaneous drain is permissible. If more than 6 hours old, it is better to leave the wound open for delayed primary closure after 2 – 4 days.

In penetrating brain injury, it is best to close the dura. This can rarely be sutured directly, but the problem can easily be overcome with a patch of pericranium or aponeurosis. After debridement of the scalp wound, the skin should be closed either directly or with a rotation flap.

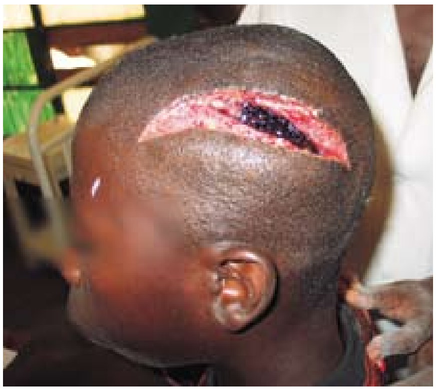

M. Baldan / ICRC

Figure 10.18

Panga wound to the head.

10.9.2 Soft tissues of the chest (sucking chest wound)

These wounds must be debrided, but healthy muscle and pleura should be closed to preserve a functional serous cavity. The skin and subcutaneous tissue should be left open and a chest tube inserted.

10.9.3 Soft tissues of the abdominal wall

As with the chest, the wound should be excised and every effort made to secure peritoneal closure. Furthermore, if the development of abdominal compartment syndrome is suspected, temporary abdominal closure is preferable (Bogotá bag, etc).

For further details on abdominal compartment syndrome see Volume 2.

10.9.4 Hand

Excision should be very conservative and all viable tissue preserved to simplify reconstruction and improve the functional result. These wounds should be left open for DPC after 2 – 4 days; however, tendons and nerves should be covered by healthy tissue, through rotation flaps if necessary. Small wounds may be closed primarily.

10.9.5 Joints

Synovial membranes should be closed; if this is not possible the capsule alone should be sutured. Little harm seems to be done if the synovium cannot be closed securely. The skin and muscle should be left open.

10.9.6 Blood vessels

Blood vessels that have been repaired primarily or by vein graft should be covered by viable muscle if possible. The skin should be left open.

For specifics of each anatomic region, see Volume 2.

10.10 Dressings

Once the wound has been adequately excised, it should be covered with a bulky absorbent dressing made of dry fluffed-up gauze re-enforced with a layer of absorbent cotton wool. This is held in place with a *loose* crepe bandage or non circumferential adhesive tape. A tight bandage wrapped around the limb and soaked with exudate that dries will have a tourniquet effect. The gauze compresses should *not* be packed tightly in the wound; this will only impede drainage. The aim is to draw inflammatory exudate out of the wound and into the dressing. Exposed tendons and joint capsules may be covered with saline-soaked compresses.

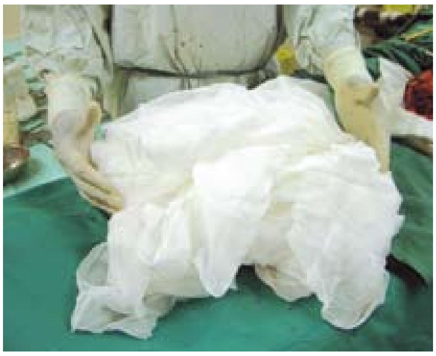

ICRC

Figures 10.19.1 – 10.19.4

Big bulky dressing of fluffed gauze covered by an elastic bandage.

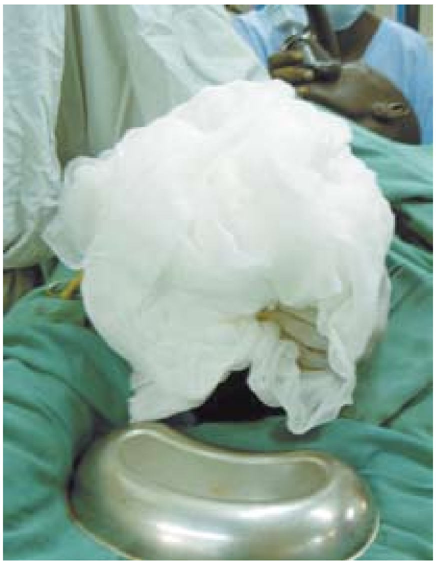

ICRC

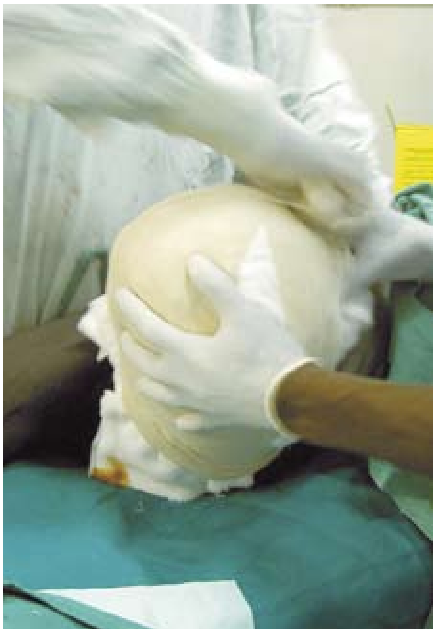

ICRC

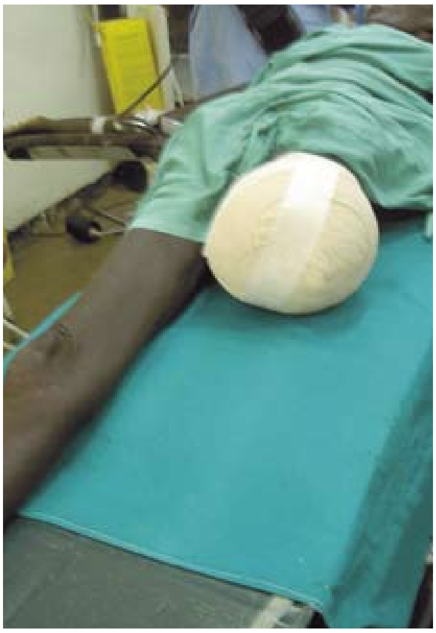

ICRC

Vaseline gauze should not be used, and the wound should not be "packed" in any way with gauze that will form a plug and prevent the free outflow of fluid.

The dressing should not be removed until taken down in the operating theatre under anaesthesia at the time of delayed primary closure. Ward dressings are an invitation to nosocomial infection. The surgeon should resist the temptation to change the wound dressing to "have a look at how it is doing". Each dressing change constitutes trauma to the healing granulation tissue and exposes it to cross-infection. A good look at the patient instead will suffice: the wound is doing quite well if the patient is smiling, eating, and sitting comfortably in bed.

If the dressing and bandage have become soaked with exudate, either over-dress with absorbent cotton, or take down the bandage and wet cotton and replace without disturbing the gauze compress that is in direct contact with the wound. The state of the dressing is not an indication of the state of the wound.

Wounds awaiting DPC should not have dressings changed until formal closure.

10.10.1 The exceptions

- Continuing *haemorrhage* requires immediate re-exploration; as do vascular changes indicating ischaemia.
- Obvious *signs and symptoms of infection:* fever, toxicity, excessive pain and tenderness, warmth, redness or shiny surface in dark-skinned people, oedema and induration, or a moist wound dressing with an offensive smell. These indicate the need for further surgical excision, which should be attended to in the operating theatre, not by changing the dressing in the ward.

All wounds awaiting DPC develop a "sour" odour after several days: the "good-bad smell" of ammoniacal products from the breakdown of serum proteins. The infected wound has a characteristic offensive odour: the "bad-bad smell".

10.11 Anti-tetanus, antibiotics, and analgesia

All patients should receive prophylaxis against tetanus. Penicillin, 5 mega-units 6-hourly i.v. should also be given from the time of admission. This should be followed by oral penicillin, 500 mg 6-hourly for a total of 5 days (see Chapter 13).

For proper rest of the injured part, and to make the patient ready for physiotherapy, good analgesia should be given (see Annex 17. A: ICRC pain management protocols).

10.12 Post-operative care

It goes without saying that proper post-operative nursing is crucial. ICRC experience has shown that the most important factor limiting the sophistication of surgical procedures performed in an ICRC hospital is not the technical expertise of the surgeon, but rather the level of post-operative nursing care. This should not be underestimated in the context of a poor, war-ravaged country.

In all cases where there is an *extensive* soft-tissue wound, even in the absence of a fracture, the entire limb should be immobilized to provide rest. This can be achieved by a plaster of Paris (POP) back slab.

The normal catabolic response to trauma must be overcome and patients offered a good nutritious diet. In a poor country, many wounded patients will arrive at hospital suffering from malnutrition. This will only compromise the body's ability to heal and resist infection.

The functional result of ultimate wound healing depends to a great extent on proper physiotherapy to retain muscle mass and joint mobility, and should be instituted early as part of the healing process.

The principles of good wound management

1. Adequate wound excision: removal of necrotic tissue, contaminated debris, organic foreign material, and blood clots.
2. Adequate wound drainage: fascial decompression, wound left open without any suturing, large bulky absorbent dressing.
3. Haemostasis.
4. Limb immobilization until healing of soft tissues.
5. Tetanus prophylaxis, antibiotics, and analgesia.
6. Nutrition.
7. Nursing and physiotherapy: mobilization of the patient.
8. No unnecessary dressing changes.
9. Delayed primary closure (4 – 5 days).

Chapter 11

DELAYED PRIMARY CLOSURE AND SKIN GRAFTING

11. DELAYED PRIMARY CLOSURE AND SKIN GRAFTING

11.1 Delayed primary closure 231
11.1.1 Methods of delayed primary closure 232
11.1.2 Dead space 232
11.1.3 Wound care 232

11.2 Skin grafting 233
11.2.1 Types of autologous skin grafts 233
11.2.2 Requisites for skin grafting 233
11.2.3 Partial thickness grafts 233
11.2.4 Reverdin pinch grafts 235
11.2.5 Application of grafts and graft meshing 235
11.2.6 Graft take 236
11.2.7 Graft care 237
11.2.8 Dressing of donor sites 237

11.3 Full-thickness grafts 238
11.3.1 Donor sites 238
11.3.2 Technique 239

11.4 Healing by secondary intention 240

11.1 Delayed primary closure

Delayed primary closure (DPC) is wound closure performed *four to seven days after injury*, which corresponds to the fibroblastic phase of wound healing. In the practice of ICRC surgical teams, the standard has been 4 – 5 days. The timing of wound closure is important; this is still defined as healing by *primary intention*.

Attempts to close wounds before they are clean must be avoided; but DPC is seldom possible later than 8 days after wound excision because of fibrosis. At this stage, healing by *secondary intention* will set in.

No wound should be closed if there is persistent contamination or infection.

For DPC, the patient is taken to theatre and the wound opened and inspected under anaesthesia. The dressing over a clean wound is dry and greenish-black in colour, and has the ammoniacal odour ("good-bad smell") of degraded serum proteins. The muscle is adherent to the gauze and, as the dressing is gently peeled away, the muscle contracts and bleeds slightly. The surface is a bright red and oozes blood. This wound is ready for DPC. Figures 11.1.1 – 11.1.5 show an example of the full sequence of wound managment with closure by skin graft.

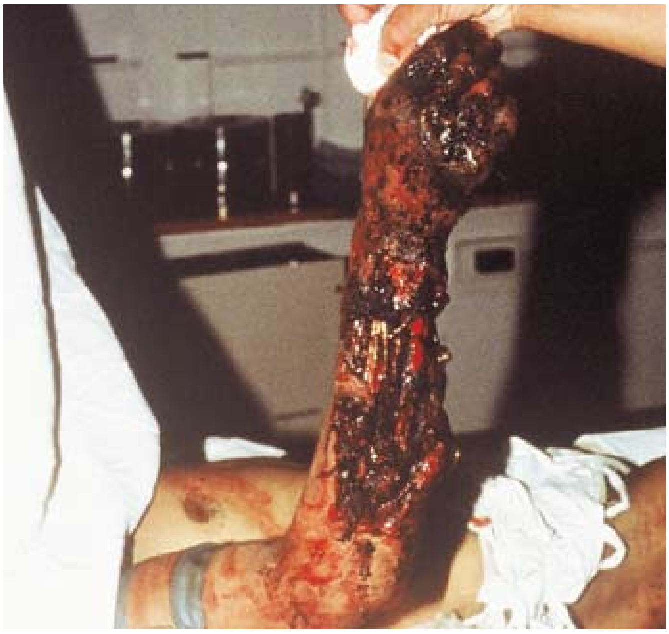
R. Coupland / ICRC

Figure 11.1.1
Very dirty wound.

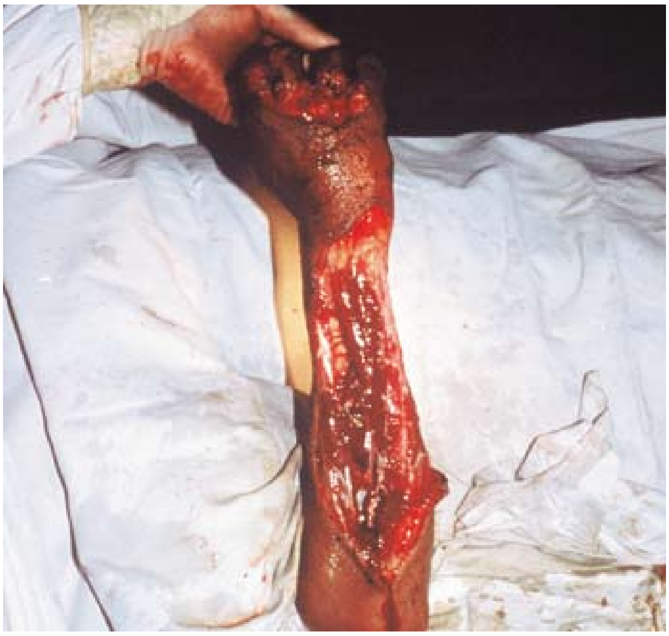
R. Coupland / ICRC

Figure 11.1.2
After debridement.

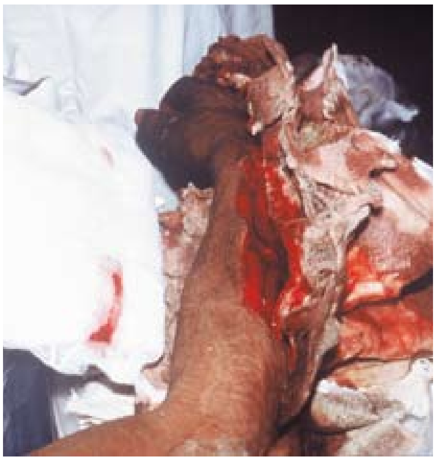
R. Coupland / ICRC

Figure 11.1.3
Five days later, removal of dressing – note the dried haemoserous discharge. As the dressing comes off, the muscle contracts and bleeds.

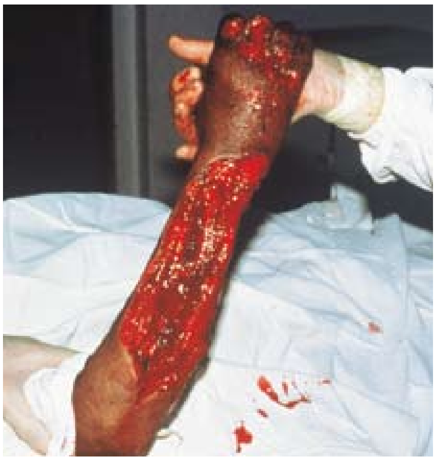
R. Coupland / ICRC

Figure 11.1.4
Clean wound ready for DPC; in this case by split-skin grafting owing to the large area of skin loss.

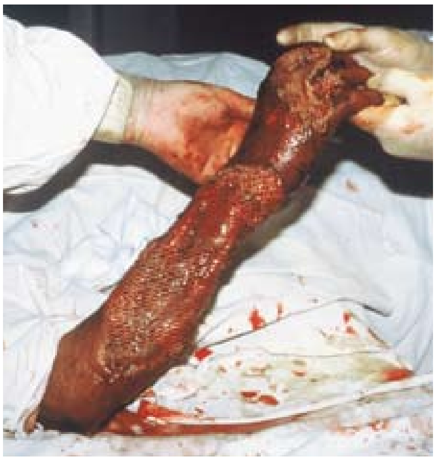
R. Coupland / ICRC

Figure 11.1.5
Wound covered by early split-skin graft.

If the wound is infected, the dressing slides off with no resistance whatsoever because there is a film of pus between it and the wound surface, which may contain areas of necrotic tissue as well as pus. The surface is a dull- or greyish-red and does not bleed. The dressing gives off the "bad-bad smell" of wound sepsis.

Such an infected wound, or one containing residual contamination or devitalized tissue, requires further surgical excision – redebridement – and is then left open again. Closure is postponed. This failed attempt at DPC is not to be confused with "serial debridement" (see Chapter 10).

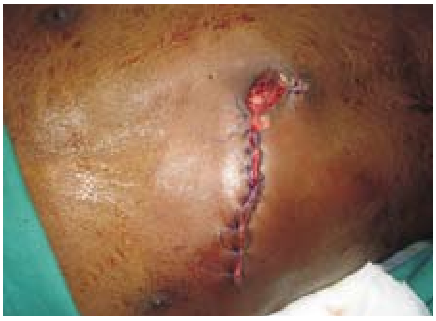
H. Nasreddine / ICRC

Figure 11.2
Direct suture as DPC.

11.1.1 Methods of delayed primary closure

This is usually accomplished by direct suture: simple approximation of the deep structures and skin with minimal mobilization of the skin edges, and *without tension* (Figure 11.2). If there is tension in the suture line, the skin edges will become necrotic and the wound will break down. Small wounds may be closed using adhesive tape to approximate the edges.

The wound should not be closed under tension.

Significant tissue loss prevents approximation of the deep structures and skin. Rotation skin flaps might be useful in some anatomic locations. If bone is exposed, a musculocutaneous flap may be necessary. Larger areas require skin grafting (Figure 11.1.5).

Adequate primary surgical excision is vital for uncomplicated wound closure by DPC.

11.1.2 Dead space

As with all wounds, the obliteration of dead space is an accepted principle of DPC management. This may be technically difficult when there has been extensive loss of deep tissue. Absorbable sutures are used to approximate deep structures, but, here again, undue tension should be avoided as it will only result in local tissue ischaemia that decreases local resistance to the development of infection and impedes healing. It is not necessary to suture fascia or subcutaneous tissues.

Drains should be *avoided* whenever possible in DPC. Drains may act as retrograde conduits for skin bacterial flora and contaminants to gain entrance into the wound, and they impair the resistance of tissue to infection; a drain constitutes an additional foreign body in the wound.

Drains should never be placed in wounds without dead space; they should only be used for those with significant dead space and where there is associated oozing of blood. Within 24 hours, the oozing has usually stopped and the drain should be removed. If used, a drain should be placed dependently and, if possible, be of the suction catheter variety. Otherwise, a simple soft penrose drain will suffice.

Avoid drains whenever possible in DPC. If used, remove within 24 hours.

11.1.3 Wound care

After DPC, the wound should be covered with some layers of dry gauze which can be left until the time of suture removal. If signs of infection develop after closure, the patient is taken back to theatre where the wound is inspected, sutures removed, and the wound re-opened for drainage and redebridement if necessary. Avoid performing this procedure in the ward, where proper inspection and drainage are difficult, as it only favours the development of cross-infection.

11.2 Skin grafting

If the wound cannot be closed by direct suture or rotational flaps because of skin loss, a skin graft can be applied, sometimes combined with partial direct suture.

11.2.1 Types of autologous skin grafts

Free skin grafts may be partial-thickness (epidermis and part of the dermis) or full-thickness (including all the dermis). Partial-thickness grafts, also called split-skin grafts (SSG), vary in thickness depending upon how much of the dermis is incorporated. The thicker the graft, the less it will contract and the more it will resemble normal skin in colour and texture; but the risk of the graft failing is greater. Conversely, thinner grafts are hardy and take more readily, but contract and distort more than thick ones and have a poorer functional and aesthetic result.

Thin partial-thickness grafts
These are used for large areas and when the development of a contracture or quality of skin cover is not important.

Thick partial-thickness grafts
These are used when skin quality is important, such as over flexion creases where contractures should be avoided. However, the recipient area must be very healthy with good vascularity.

Full-thickness grafts
These are best for a good cosmetic result in the face, or to cover a sensitive functional area on the hands or fingers.

11.2.2 Requisites for skin grafting

Skin grafts can cover any wound with enough vascularity to produce granulation tissue. This does not imply that fully mature granulation tissue must be present prior to skin grafting, and early grafting for DPC, when the wound is first opened on the fifth day post-debridement, is often used (Figure 11.1.5). Early SSG has advantages and disadvantages: the wound is closed preventing any infection from developing, which is particularly important in large wounds; but the quality of resultant skin is less and there is greater tissue fibrosis and contraction. Most full-thickness grafts are also performed early.

On the other hand, a deep wound, or one over a flexion crease, may be dressed while waiting for more granulation-tissue to grow in, and then grafted at a later date. However, excessive granulation tissue should be scraped clear prior to graft application. A limb with a wound over a flexion crease should be splinted in the extension position with a plaster back slab to avoid contracture while waiting for closure.

Tissues which cannot take skin grafts include avascular areas, such as hyaline cartilage, exposed tendon without paratenon, and exposed bone cortex without periosteum. These cases will require some form of skin or musculocutaneous flap for closure. In the case of exposed bone cortex, an alternative may be the drilling of multiple small holes in the cortex to let granulation tissue grow through from the inner cancellous bone.

11.2.3 Partial thickness grafts

Split-skin grafts (thin and thick) are taken from an area which can produce a broad area of skin. The common donor sites are the thighs, back, and arms and forearms (Figure 11.3).

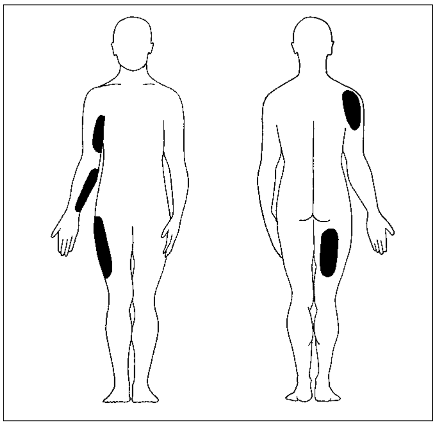
P. Zylstra / ICRC

Figure 11.3
Donor sites for partial-thickness grafts.

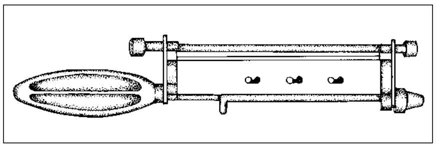

P. Zylstra / ICRC

Figure 11.4

A dermatome: various models, with disposable blades, are available – the thickness of the split skin graft harvested is controlled by the screw at the right end of the instrument; the setting is then locked using the screw at the left end.

Split-skin grafts should be taken using a dermatome, such as a Humby knife (Figure 11.4). A free hand knife, such as a De Silva knife which incorporates a razor blade, or scalpel, can be used if a dermatome is not available or if the area to be harvested is small.

The principles for the use of all hand knives are the same. The blades are disposable and are fitted into the knife. A calibration control is adjusted to set the depth of split skin required. Once this has been adjusted, a locking nut is tightened at the opposite end of the knife. It is important to check the gap between the blade and the knife before use, by holding the instrument to the light and visually checking the gap. The gap should be even throughout its length. Electric and pneumatic knives are also available but are not standard in ICRC practice.

The gap between the knife and the blade must be visually checked before use.

Harvesting the SSG

The area of skin to be removed is washed with soap and water and shaved if necessary; povidone iodine is applied. Intradermal saline or, better still, dilute adrenaline solution (1:500,000), is injected into the donor area prior to cutting, to facilitate harvesting and diminish local blood loss. The donor site, the cutting edge of the knife, and the edge of a skin board are greased with vaseline-gauze.

An assistant applies an ungreased skin board to the distal end of the donor site and pulls in such a fashion as to tense the skin tight. The assistant's other hand grasps the underside of the donor site (thigh or arm) to flatten the area from which the graft is being taken (Figure 11.5.1).

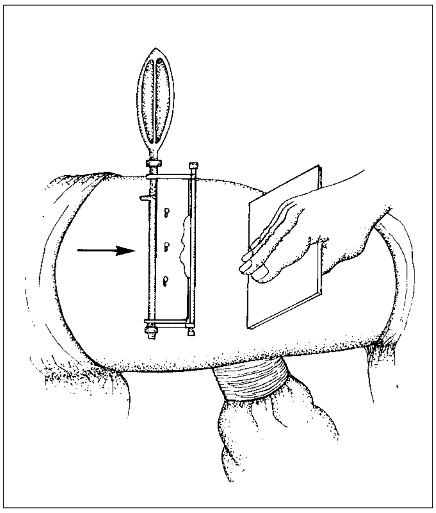

P. Zylstra / ICRC

Figure 11.5.1

Harvesting a split skin graft from the medial side of the thigh – note the assistant's left hand flattening the donor site by exerting upward pressure on the undersurface of the thigh.

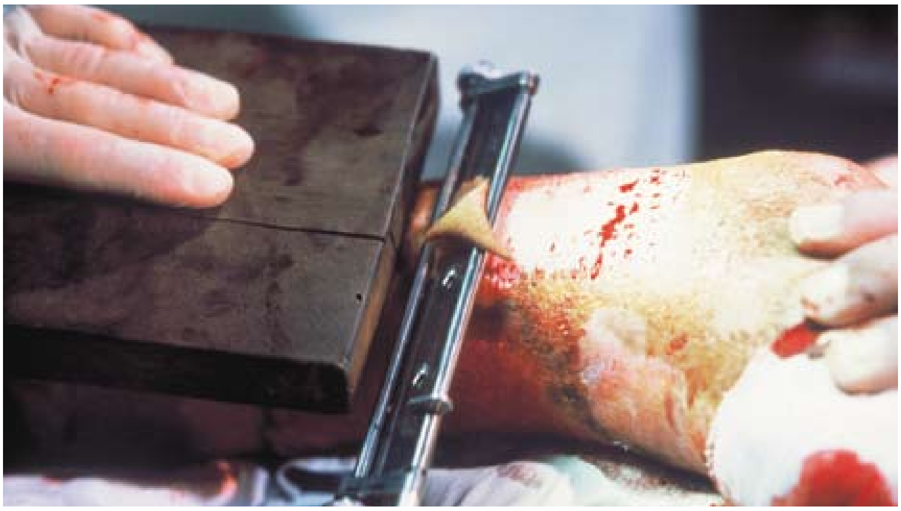

E. Winiger / ICRC

Figure 11.5.2

Surgeon's wooden board applying counter-traction.

The surgeon places the greased board about 4 – 5 cm in front of the ungreased board, and exerts counter traction to increase the skin tension (Figure 11.5.2). The knife goes into the gap between the two boards at an angle of 30° and even, gentle, side-to-side cutting movements of about 2 cm are made. Continuous movement of the knife is essential. The surgeon advances the greased board proximally while exerting *minimal effort* to advance the knife. The tendency is to cut too quickly, or exert too much pressure at too high an angle; the blade will not shave the skin off but cut deeply into the subcutaneous tissue. When the desired amount of skin has been raised, the wrist is supinated and the knife cuts through. The harvested skin is now placed on a saline-soaked gauze swab, undersurface upwards, and stored until applied to the recipient area.

11.2.4 Reverdin pinch grafts

These can be used to cover large areas of defect and where the recipient area is unfavourable, such as muscles moving in different planes. A full take is not expected and the gaps between the grafts will fill in as the skin spreads outwards and coalesces. The cosmetic result is poor.

As with SSG, an intradermal infiltration of local anaesthetic with adrenaline is injected into the donor area. This helps to avoid cutting the grafts too deeply and is often all the anaesthesia required. The grafts are cut using a scalpel blade and can vary in size up to about 2 cm^2 (Figure 11.6). The harvested skin is handled as described previously until it is applied to the recipient area.

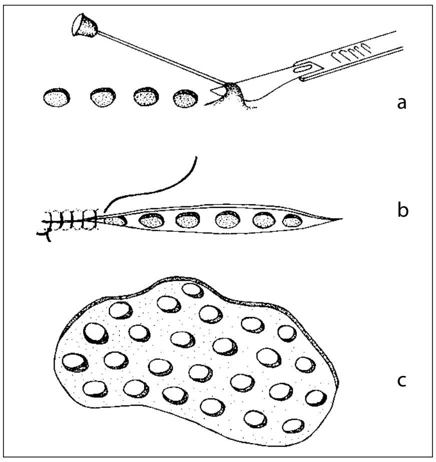

P. Zylstra / ICRC

Figure 11.6

Reverdin pinch grafts:

a. The skin is picked up with a needle and round patches of thick SSG, 1 – 2 cm in diameter, are cut with a scalpel.

b. If the patches are taken in a line, the area can be excised and the wound closed by primary suture.

c. The patches are spread over the receptor site with gaps of a few millimetres between each piece.

11.2.5 Application of grafts and graft meshing

The recipient bed is prepared by washing with saline and scraping clean any exuberant granulation tissue.

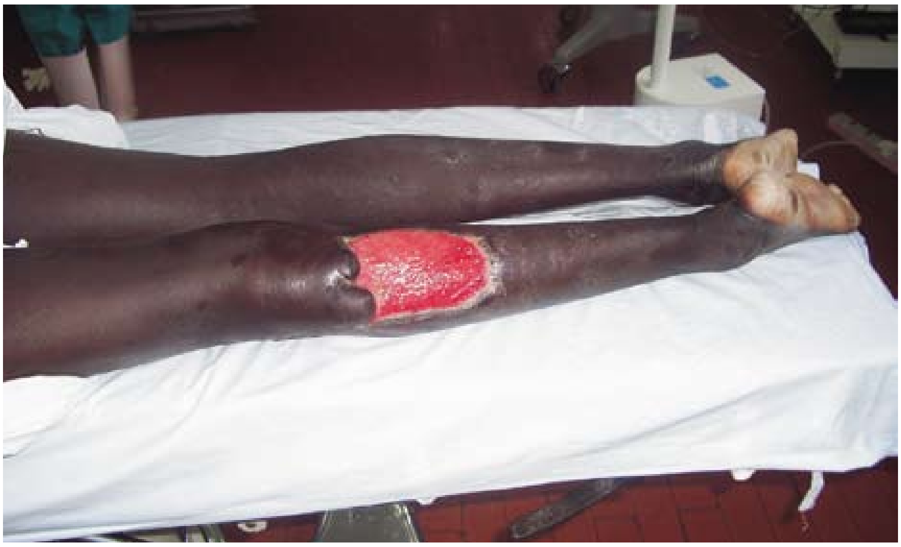

V. Sasin / ICRC

Figure 11.7

Recipient site prepared to receive SSG.

The harvested partial-thickness skin is laid across the recipient bed and cut so that it is slightly larger than the site. This allows for the cut undersurface to be in complete contact with the recipient bed and for later graft contraction. Often a graft is placed on a wound that is deeper than the thickness of the graft. In these cases the edge of the graft at the base of the wound must be carefully approximated to include the entire height of the defect.

The graft should be meshed to allow for drainage of serum and blood; this is important to avoid the graft "floating" and provide for close contact between it and the recipient bed. Meshing is accomplished by placing the harvested skin on a surface, preferably the wooden board, undersurface facing upwards, and making numerous parallel incisions through the graft with a No. 15 scalpel blade. Optimally, the ratio of gaps to skin should be 3:1 thus allowing the SSG to cover an area three times its original size (Figure 11.8). Over a period of 10 – 14 days, the skin grows across the interstices and complete healing is obtained.

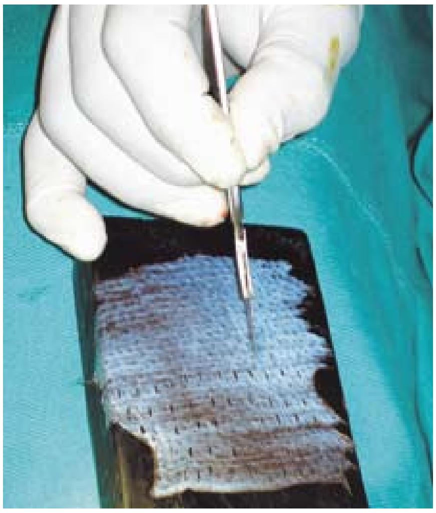

H. Nasreddine / ICRC

Figure 11.8

Meshing of a skin graft.

Graft meshing is particularly useful in cases where there is insufficient autograft to completely cover raw areas, such as large wounds or burns (see Figure 11.1 for a clinical example). Mechanical meshing machines are available – and expensive – but are not standard ICRC equipment except in hospitals with a large number of patients suffering from burns. A good improvisation is a sharp circular pizza cutter, modified with notches ground into its circumference, and used to mesh skin against the wooden board.

Grafts can often be kept in place by a suitable dressing and elastic bandaging. If graft stability cannot be ensured by the covering dressing, it should be sutured in place (Figure 11.9). One technique is a continuous 3/0 non-absorbable suture run around the edge of the graft to provide adequate contact. Another is to use several anchor sutures about the circumference, left long and used as tie-overs across the dressing; especially useful for small grafts on hollow areas (Figure 11.11). The excess skin of the graft can either be trimmed once the graft has been sutured in place, or cut 10 – 14 days later.

Figure 11.9

SSG sutured in place.

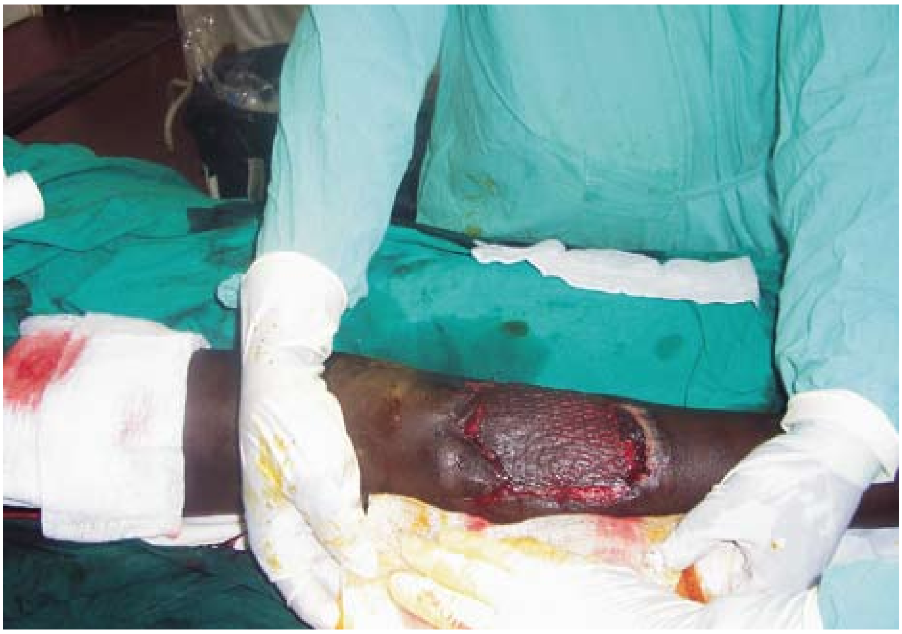

V. Sasin / ICRC

Once the graft has been sutured or placed on the recipient bed, it should be covered with a sheet of vaseline-gauze followed by a gauze compress or cotton-wool pledgets soaked with saline. These dressings should be pressed into the concavities of the defect to ensure that there is complete apposition between the graft and the bed. The dressing should then be covered with a thick layer of cotton wool held in place by a crepe bandage.

In case of consistent oozing from the recipient bed, consider delaying the application of the graft for 24 – 48 hours. The harvested skin can be stored for up to three weeks in an ordinary domestic refrigerator at 4° C without compromising the success rate. It is placed on a moist piece of gauze with the undersurface facing upwards and then put in a sterile jar containing normal saline. Jars should be stored with an airtight lid and *carefully labelled* with the patient's identification and the date of harvesting.

Delayed application of split skin grafts is often
the preferred method.

11.2.6 Graft take

Grafts adhere initially by a thin layer of fibrin and are fed by "plasmatic imbibition" from the recipient bed granulation tissue until in-growth of capillaries occurs by about the fourth post-operative day. The new capillaries link with the graft and the fibrin clot is transformed into fibrous tissue. For these reasons, there are three major factors determining the success rate of split-skin grafts.

1. A vascular recipient bed which is *free of pathogenic bacteria*. Normal bacterial flora will not necessarily interfere with graft survival. The most common pathogen causing graft failure is beta-haemolytic *Streptococcus pyogenes*, probably through its fibrinolytic enzymes. ICRC protocol suggests prophylaxis with penicillin for five days, to which this bacterium remains sensitive.

2. The *vascularity* of the graft itself. Thin split-skin grafts have a higher concentration of capillaries on their undersurface than thick grafts. For this reason, a thin split-graft has a higher chance of survival than a thick one.

3. The maintenance of *close contact* between the graft and the recipient bed. If the graft is under tension, if blood or serum collects between the surfaces, or if there is movement of the graft on its bed, satisfactory contact cannot be maintained. Thus, meshing of the graft and tight contact between the graft and recipient site are important; grafts spanning joints should be splinted to prevent graft loss from shearing forces.

Grafts should be harvested under sterile conditions and placed on a clean, well-vascularized recipient bed. It is imperative that there be tight contact between the graft and its bed.

11.2.7 Graft care

Grafts must be cared for by experienced personnel. A graft may be *saved* if a collection of serum or haematoma is evacuated early. Great care must be taken not to *lose* a good graft through clumsy or careless removal of the dressing.

Grafts require careful nursing care.

In ICRC practice, the graft is first inspected after 48 – 72 hours; the dressing is carefully removed using two pairs of tissue forceps: one holding the graft and keeping it applied to its bed, the other removing the dressing. Care must be taken not to tear away the graft. If the graft is healthy and adherent to the recipient base, another vaseline-gauze and compress are bandaged in place and no further dressing is necessary for another 10 days, until the time of stitch removal.

If a haematoma or seroma is found, the fluid should be expressed through the meshed gaps using gentle pressure applied through the forceps and a pledget of soaked cotton-wool. Another firm dressing of vaseline-gauze is applied. These grafts are then inspected on a daily basis until firmly adherent to the recipient base.

Any small pus collections should be debrided with scissors to prevent spread of infection. Intra-epidermolysis is a phenomenon where the outer layer of the graft is lost while viable epithelial cells remain at the base. Dark blisters can be seen and may be debrided but great care should be taken not to disturb underlying graft that may still be viable.

On the other hand, any graft which is dead or floating on a "sea of pus" should be removed and the wound cleaned with normal saline. (If Pseudomonas is the culprit – presence of bluish-green pus – a dilute vinegar solution is useful.) If the raw area of the wound is more than 2 x 2 square cm, it should be re-grafted when clean. If smaller, it can be left to heal by secondary intention.

At ten days, a healthy graft should be firmly adherent and can be left exposed and submitted to a simple daily inspection.

11.2.8 Dressing of donor sites

Split-thickness donor sites can bleed briskly, and in children can result in significant blood loss. As mentioned above, the donor site should be infiltrated intradermally with a 1:500,000 solution of adrenaline or a local anaesthetic with adrenaline to decrease bleeding. They can also be extremely painful, especially when dressings become firmly adherent.

The following regimen minimizes discomfort and complications.

1. After harvesting the SSG, a dry dressing should be applied immediately, kept in place if necessary with an elastic bandage, and left alone while the graft is being applied; by which time haemostasis has usually been achieved.
2. The dressing should then be removed. If there is persistent bleeding, a compress wetted with the dilute adrenalin solution should be applied and direct pressure exerted for a few minutes.
3. Next, the donor site should be covered with a firm pressure dressing of vaseline-gauze, dry compresses and elastic bandage; it should be kept from falling down the limb by adhesive strapping. The dressings should be left *in situ* for 10 –14 days, unless there are signs of underlying infection.
4. A membrane dressing, such as Opsite ®, can be used if available. If fluid accumulates beneath the membrane, it can be evacuated by pricking with a needle and pressing. The perforation site should be covered with a smaller piece of membrane. If the dressing falls off it can be reapplied. Generally, this stays in place for the desired period of about ten days.

11.3 Full-thickness grafts

Full-thickness skin grafts consist of the epidermis and entire dermis. These grafts are used principally for the *head and neck* and *hands and feet* to provide thicker, better-quality cover.

Full-thickness grafts have a number of advantages apart from better texture and colour. They allow for the transfer of hair-bearing skin, as all the dermal adnexal structures are intact, and they contract less than split-skin grafts. The main disadvantages are a lower survival rate and their limited size.

The recipient bed must be in optimal condition to take a full-thickness graft; haemostasis, in particular, must be perfect. There should be meticulous suturing and approximation between the graft and the recipient edge. Generally, a full-thickness graft is small because the plasma exudate of the recipient bed is barely sufficient to feed and oxygenate the graft, and capillaries from the recipient bed and edge and graft undersurface must grow in quickly to assure survival.

11.3.1 Donor sites

The best and most readily available donor sites are the supraclavicular fossa, in front of and behind the ear, and the flexor surface of the forearm or the groin creases. In older persons, the cheek or the neck in the line of a skin crease may also be used. The donor site is sutured primarily following harvesting of the graft.

Figure 11.10

Donor sites for full-thickness skin grafts.

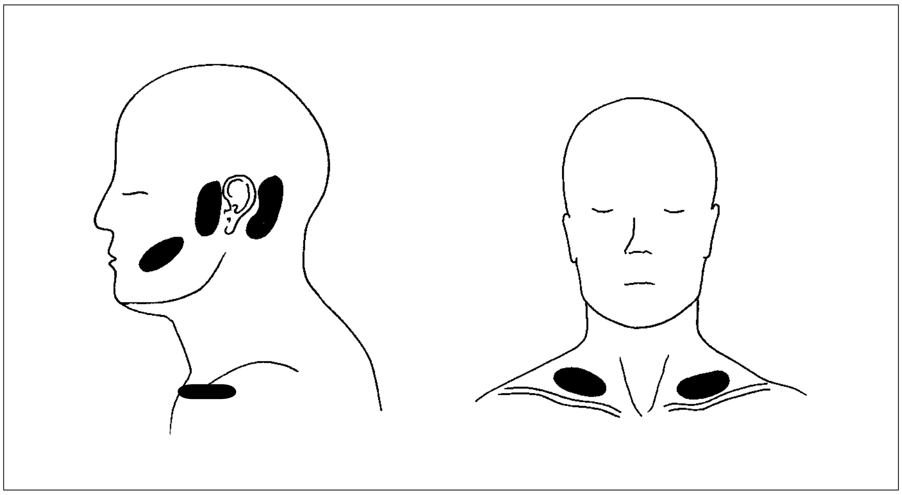

P. Zylstra / ICRC

11.3.2 Technique

The graft can be taken under local anaesthesia. The donor and recipient areas are cleansed in a sterile fashion and a template of cut gauze compress made to fit the defect exactly. The template is placed on the donor site and an outline drawn around its circumference. The donor area should be injected with local anaesthetic *with* adrenaline and the edge of the recipient area with local anaesthetic *without* adrenaline; both are left for five minutes before incision. The full-thickness skin should then be excised exactly.

The graft must be *thoroughly defatted* before application. This is best achieved by placing the moist graft, epidermis-side down, on the finger tip and scrupulously excising the fat using fine scissors. The graft is left unmeshed.

The graft is then sutured in place immediately and meticulously. Several interrupted nylon sutures are inserted close to the edge and left long for tie-overs (Figure 11.11). Small interrupted or continuous fine nylon sutures (5/0 if available) are placed around the circumference.

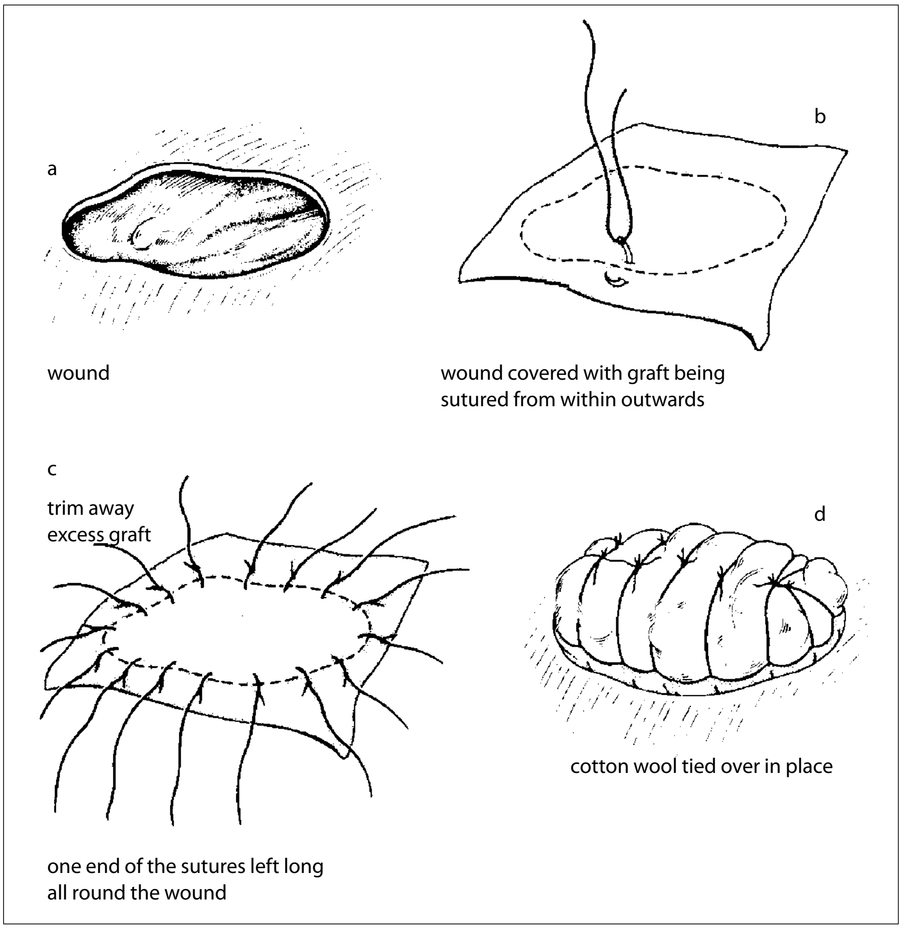
a
wound
b
wound covered with graft being sutured from within outwards
c
trim away excess graft
one end of the sutures left long all round the wound
d
cotton wool tied over in place

Figure 11.11

Tie-over method of securing skin grafts.

M. King / Primary Surgery Volume 2

Face grafts may be left open and covered with antibiotic ointment. They may be gently rolled with gauze or poked with a needle in the first 48 hours to express small serum collections.

On the hands or over a flexion crease, once the graft has been sutured in position a bolus of dressing is fashioned with vaseline-gauze and moist cotton-wool pledgets to hold the graft closely in contact with the recipient bed. The tie-over sutures are brought together over the dressing, which should be snug, but not so tight as to strangle the graft.

The dressing should be left in place for 7 – 10 days and then removed. At this point, in both light- or dark-skinned people, the graft may be any colour from pinkish or pinkish-white to bronze to black. Even a graft that has a blackish appearance on inspection can survive. No judgement on graft survival should be passed until a month from the time of application.

11.4 Healing by secondary intention

Some small wounds are difficult to suture without tension, or without very wide mobilization of skin flaps, because their surrounding edges are fibrotic. There is little benefit in another surgical operation, even skin-grafting. These wounds are best left to granulate and heal by secondary intention (Figure 12.10).

A change of dressing and gentle washing with normal saline every 4 – 5 days is usually sufficient: every dressing is a trauma to a healing wound. Daily dressings are not necessary if the wound is dry. Some traditional local dressings (sugar, honey, etc.) have an antibacterial effect and promote the formation of granulation tissue and are useful if the wound is deep. They have been used successfully by ICRC surgical teams.

Chapter 12

NEGLECTED OR MISMANAGED WOUNDS

12. NEGLECTED OR MISMANAGED WOUNDS

12.1 General considerations 245

12.2 Chronic sepsis: the role of biofilm 247

12.3 Surgical excision 248
12.3.1 Soft tissues 249
12.3.2 Bone 249
12.3.3 Irrigation 250

12.4 Antibiotics 250

12.5 To close or not to close? 251

12.1 General considerations

As described in Chapter 10, in many war zones today, characterized by irregular bush warfare, rebellions and insurrections, neglected and mismanaged wounds are amongst the commonest injuries that the surgeon sees. First aid is not available, doctors and nurses are scarce, and health services have been disrupted by poverty and conflict. Distances are long and the terrain difficult with little in the way of organized transport. Many patients reaching hospital have wounds which were inflicted more than 24 hours earlier, and some have wounds that are many days, or even weeks, old. Even if injured patients reach hospital fairly quickly, the sheer number of casualties often exceeds the surgical capacity or expertise available. The result is an inordinate delay in their treatment or frank mismanagement.

ICRC EXPERIENCE

In the ICRC hospital of Lokichokio in northern Kenya serving the victims of the conflict in southern Sudan, 12,264 war-wounded patients were registered between 1991 and 2006 in the ICRC surgical database. Evacuation was by ICRC and United Nations airplanes and few, if any, received pre-hospital treatment of any sort. Of these patients, 84 % reached the hospital more than 72 hours after injury. ICRC staff have seen similar conditions of delay in evacuation to a surgical facility in Somalia, the Democratic Republic of the Congo, Nepal and elsewhere, and have seen the same phenomenon with crush wounds after earthquakes in low-income countries.

Before reaching a proper surgical hospital, some patients receive no treatment at all; others a cursory dressing, while yet others have undergone an inadequate wound excision. Some wounds will have been sutured, generally without any form of excision.

All sutured wounds, no matter how clean they appear, must have their sutures removed and opened up for drainage. As mentioned previously, the simple rules of septic surgery apply. Never close infected or contaminated wounds primarily.

Remove all sutures from previously sutured wounds.

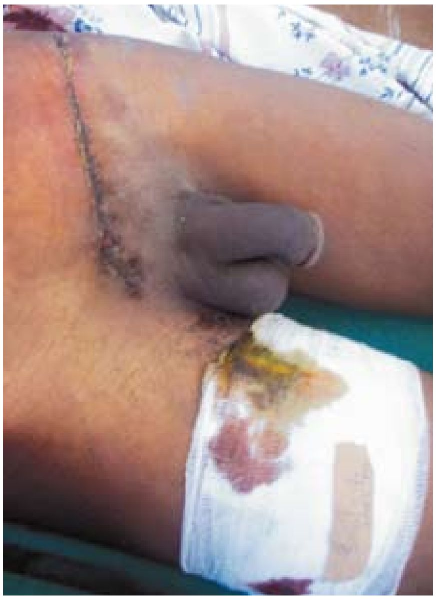

H. Nasreddine / ICRC

Figure 12.1.1

Patient A: The gunshot wound to the left iliac region was sutured 5 days previously. The wound to the right thigh underwent insufficient wound debridement.

After delay, some Grade 1 minor soft-tissue wounds are found to be healing spontaneously. Most wounds, however, will be inflamed or frankly infected with a degree of chronic sepsis, and some will be grossly putrefying. This is often seen in mismanaged wounds. Tetanus, gas gangrene, and invasive haemolytic streptococcal infection are ever present dangers (see Chapter 13). These wounds require aggressive excision.

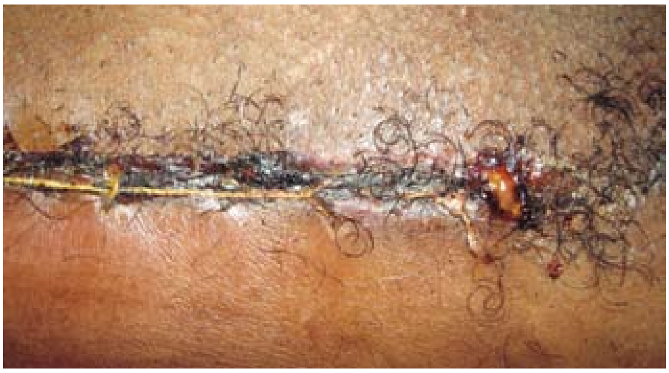

H. Nasreddine / ICRC

Figure 12.1.2

The sutured wound is infected; a pearl of pus is evident at the lateral edge.

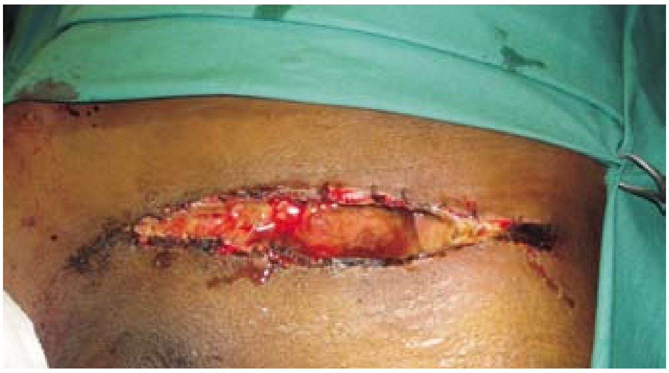

H. Nasreddine / ICRC

Figure 12.1.3

Sutures have been removed revealing abundant pus.

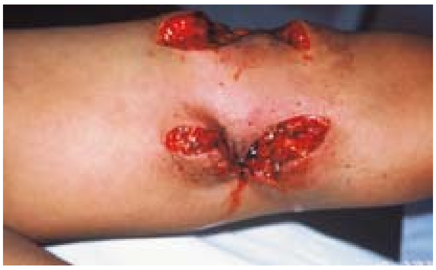
ICRC

Figure 12.2.1

Patient B: primary suture was performed – note the tension in the wound due to tissue oedema and infection. Some sutures have been removed.

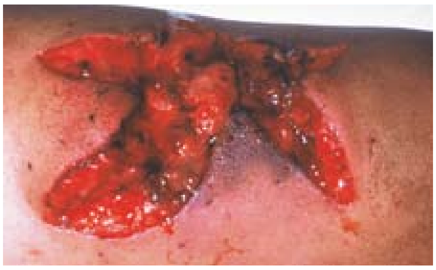
ICRC

Figure 12.2.2

All stitches have been removed – the skin edges are ischaemic and necrotic and the subcutaneous tissues oedematous.

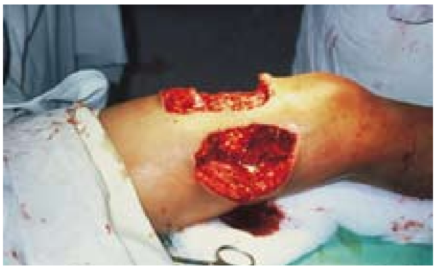
ICRC

Figure 12.2.3

After redebridement – the wound is now larger than the original injury.

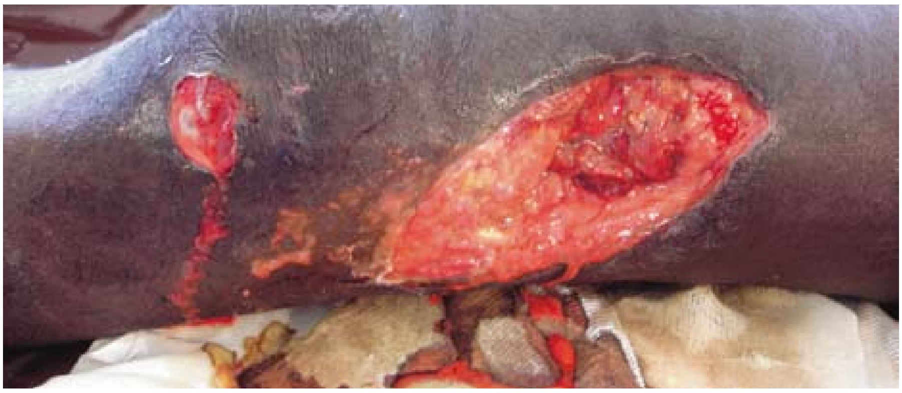
H. Nasreddine / ICRC

Figure 12.3

Neglected gunshot wound to the knee with frank sepsis.

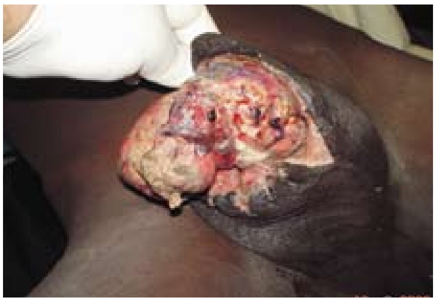
H. Nasreddine / ICRC

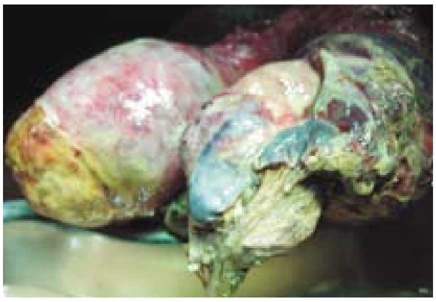
H. Nasreddine / ICRC

Figures 12.4.1 and 12.4.2

Neglected wound to the scrotum with gangrenous tissue.

Please note:

Many neglected wounds are infested with maggots. There is a long historical literature about “maggot debridement therapy”, especially for chronic wounds, and a number of surgeons around the world practise it today. Suffice it to say that, for cultural and psychological reasons, most patients do not accept such a method. ICRC experience in this domain, although often positive, can only be described as anecdotal.

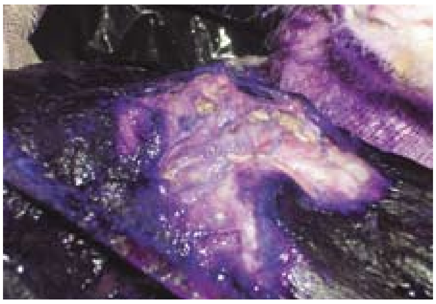
E. Dykes / ICRC

Figure 12.5

Maggot infestation of a wound. (The colouring is due to gentian violet.)

Please note:

There is much clinical and experimental evidence to back the claim that the early administration of antibiotics, penicillin in particular, can delay the advent of serious infection in war wounds. It is ICRC practice in first-aid posts to begin penicillin as soon as possible. However, adequate first-aid measures, as described in Chapter 7, are seldom available in many contemporary theatres of war and it is to this context that the present chapter is devoted.

12.2 Chronic sepsis: the role of biofilm

The chronic pyogenic sepsis of a neglected wound has its own particular pathology and bacteriology, and life history. Bacteria exist in discrete colonies only in a laboratory. The natural state of bacteria through natural selection is to become attached to surfaces, especially to inorganic or dead matter, such as sequestrated bone and cartilage. In chronic infection, bacteria secrete a glycopolysaccharide biofilm; this is the “slime” one feels on the rocks in a river. This biofilm protects the bacteria and prevents not only antibiotics, but macrophages, leukocytes and antibodies from attacking them. It must be actively broken up in order to eradicate the infection.

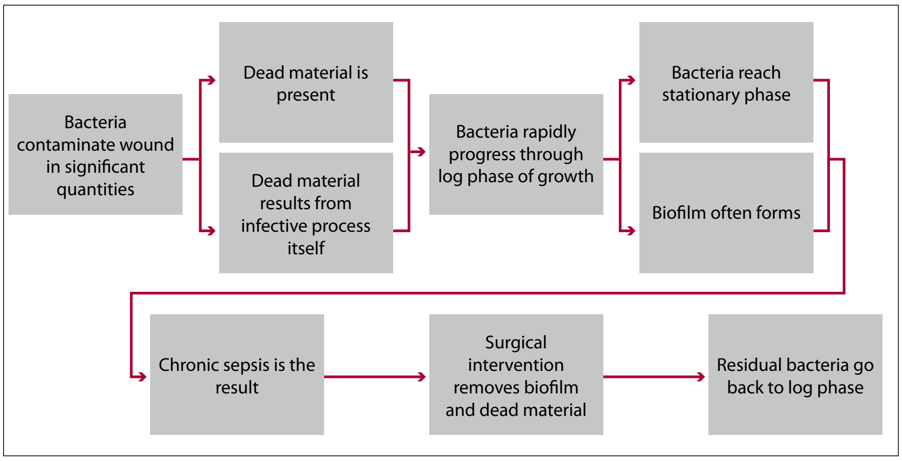
Bacteria contaminate wound in significant quantities
Dead material is present
Dead material results from infective process itself
Bacteria rapidly progress through log phase of growth
Bacteria reach stationary phase
Biofilm often forms
Chronic sepsis is the result
Surgical intervention removes biofilm and dead material
Residual bacteria go back to log phase
D. Rowley / U. of Dundee

Figure 12.6

Schema of the chronic infective process.[1]

The biofilm is secreted during the stationary phase of bacterial life. Physical disruption of this biofilm and surgical removal of the dead tissues are required to push the bacteria back into the rapid multiplication of the log phase when they are most susceptible to antibiotics and the body’s natural defence mechanisms.

---

[1] Rowley D.I., University of Dundee, Course on chronic bone infection, Seminar on the Management of War Wounds; ICRC: March 18, 2005; Geneva.

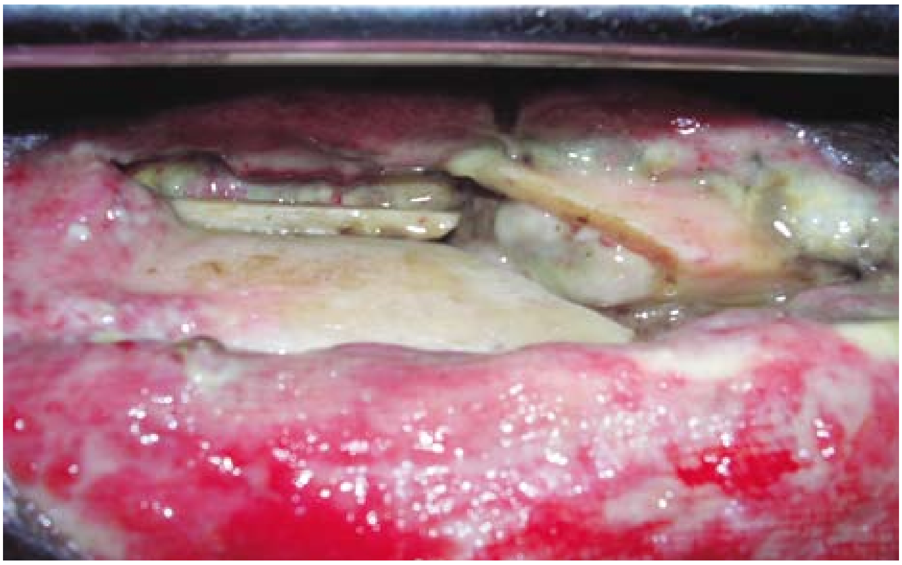

E. Dykes / ICRC

Figure 12.7.1

Neglected infected wound showing a layer of biofilm pus.

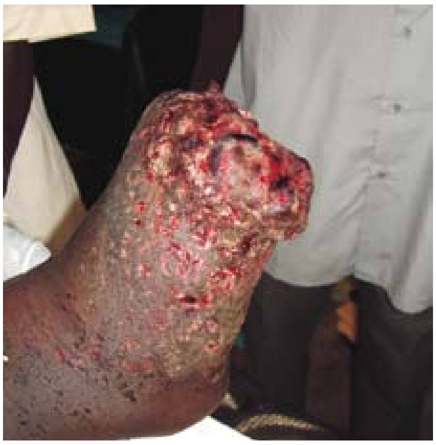

M. Dalla Torre / ICRC

Figure 12.7.2

Neglected wound of an amputation stump, without biofilm layer.

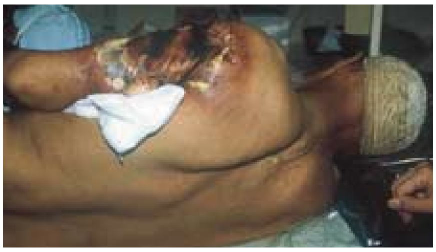

R. Coupland / ICRC

Figure 12.7.3

Neglected wound to the arm with necrotic tissue, but dry.

12.3 Surgical excision

Surgical excision is a more difficult procedure in these neglected and mismanaged wounds. The line of demarcation between viable and non-viable tissue, especially in oedematous muscle and fascia, is less obvious, and the zone of post-traumatic inflammatory hyperaemia is compounded by that of infective inflammation and the presence of the biofilm. The extent of adequate excision is more difficult to assess since there is not only a "mosaic" of ballistic tissue destruction in the permanent cavity itself but also that due to the septic process. Festering wounds that are days old have regions of infection mixed with regions of fibrous tissue healing.

The wound often resembles a cavity filled with detached soft tissue, bone fragments or the ends of fractured long bones, foreign debris, and fibrous tissue, covered with a coat of pus (Figures 12.7.1 and 12.8.2). Access may be difficult because of areas of wound contracture due to the tough fibrous tissue.

The principles of surgery remain the same, however. The skin and deep fascia must be widely re-incised and the wound cavity laid open for proper visualization and drainage (drainage following the same principle as for a simple abscess). The excision is directed towards the *removal* of all non-viable and heavily contaminated tissue and foreign debris, and the *physical disruption* of the biofilm.

Because surgical excision is more difficult in these wounds, there is a higher incidence of persistent infection. In such circumstances, several redebridements may become necessary. Depending on the experience of the surgeon, these are the patients who might most benefit from staged serial debridements.

12.3.1 Soft tissues

If present, all sutures must be removed and debridement performed as if no previous wound toilet had been done.

All macroscopic contamination should be excised. Skin and subcutaneous fat must be cut back until they ooze blood. Grossly infected fascia is usually shredded and dull grey, while healthy fascia is a glistening white. The muscle compartments are in even greater need of decompression than those of a fresh wound.

Contractility is the best indication of muscle viability. On the other hand, the surgeon must learn to distinguish between bleeding from small vessels in partially necrotic tissue, and capillary oozing from healthy but inflamed tissue. Experience brings an increase in the amount of tissue excised. Vessels and nerves should be left intact on the other hand, because they rarely turn septic.

Operative blood loss is usually considerable because of the inflammatory oedema and hyperaemia.

It is even more important *not* to open up healthy tissue planes in an attempt to remove any projectile. This will only help to spread the infection.

12.3.2 Bone

If there is an underlying fracture in these septic wounds, it usually contains *detached and necrotic* bone fragments that provide the perfect "surface" for bacterial adhesion. It is vital that these be found and removed. A plain X-ray will assist in their identification (Figure 12.9.2).

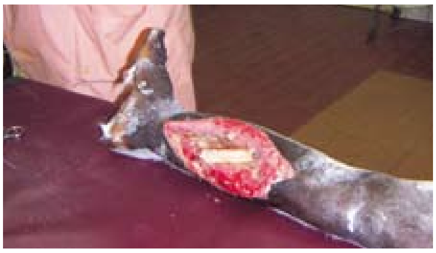
F. Jamet / ICRC

Figure 12.8.1

Gunshot wound 7 days after inadequate debridement.

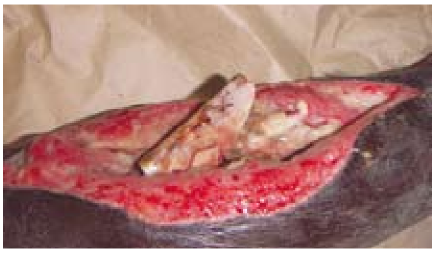
F. Jamet / ICRC

Figure 12.8.2

Obvious sequestra presenting at the wound surface – note the presence of a pus-filled biofilm covering bone and soft tissue.

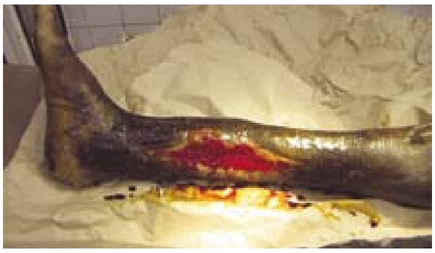
F. Jamet / ICRC

Figure 12.8.3

Wound 7 weeks after debridement.

Totally detached bone fragments are often trapped inside fibrous tissue and their localization and identification can be extremely difficult and frustrating.

- A dead bone fragment is pearly white; living bone is dull greyish-white with red capillary specks.
- If scratched with a bone curette, dead bone will give off a high and hollow resonance; living bone sounds low and dull.
- Another telltale sign can be observed when the bone fragment is grasped with forceps and a motion of pronation and supination imparted. If muscle or periosteum is holding the fragment in place, these structures will move with the pronation and supination. If the fragment is detached and held by fibrous tissue only, the movement will break the fibrous adhesion and the bone fragment will come free.

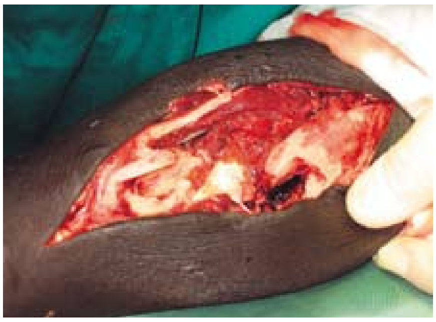

H. Nasreddine / ICRC

Figures 12.9.1 and 12.9.2

All loose bone fragments must be removed.

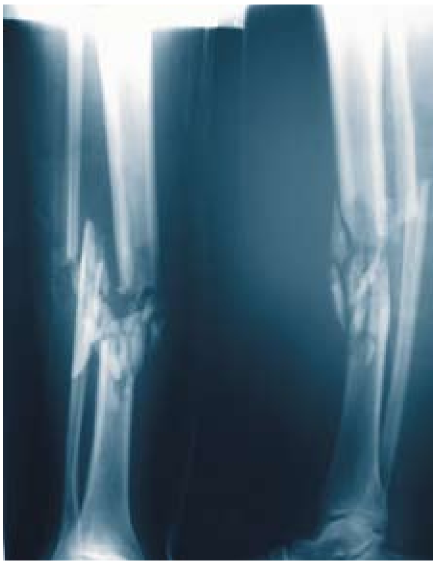

H. Nasreddine / ICRC

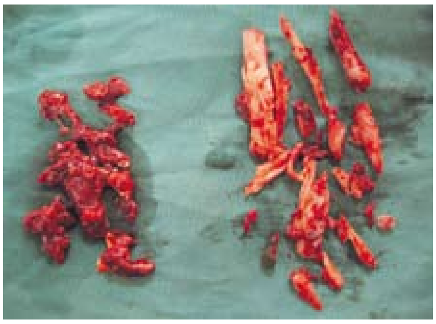

H. Nasreddine / ICRC

Figure 12.9.3

Bone fragments on the right, excised fibrous tissue on the left.

Once all loose bone fragments have been removed, the wound cavity is then curetted and a careful finger exploration performed to confirm a smooth surface: sharp bone edges may injure the exploring finger! All detached and dead bone fragments must be removed; they are already sequestra. Sometimes, several attempts may be necessary to eliminate them all, i.e. several redebridements.

All detached bone fragments are sequestra.

12.3.3 Irrigation

The wound is now copiously irrigated and dressed as for routine surgical excision. The ICRC has tested irrigation with a high-pressure pulsatile lavage and daily showering of the open cavity in these neglected septic wounds. Although both techniques appeared to clear up the biofilm and afford a better assessment of tissue viability, the results were inconclusive. Both involve a great deal of time, effort, and nursing care, and are difficult to maintain if a large number of patients are under treatment with no possibility of transfer to other less overstretched facilities.

12.4 Antibiotics

Penicillin and metronidazole are given, according to the antibiotic protocol (see Chapter 13); and gentamycin is added if there are signs of active spreading inflammation. No topical antibiotics or antiseptics are used in ICRC practice.

Bacterial culture should be done, if available. Good bacterial culture and sensitivity in a forward hospital is far more diffucult to accomplish than is usually realized. Note that clinical response does not always follow laboratory sensitivities, or lack thereof. Not only do “wild” bacteria not live in colonies, but *in vitro* sensitivity does not always reflect *in vivo* response, and the surface or discharge flora does not always represent the bacteria present *inside* the tissues.

Nonetheless, the basic rules of septic surgery still apply. Infection requires good drainage – as with an abscess – and the elimination of the bacterial culture medium, i.e. the dead tissue. Antibiotics will only be effective once the biofilm has been disrupted and the bacteria made susceptible to their action.

Persistent or recurrent wound infection requires re-excision, not ward dressings.

12.5 To close or not to close?

The primary phase of wound healing begins at the moment of wounding. It is well established by the time of presentation if the wound is a few days old. Thus, several days after surgical excision, many wounds are past the time of healing by primary intention. There is already much fibrous tissue present (Figure 12.9.3) and even more by the time the wound is clean and ready for closure. If secondary suture is attempted there is usually considerable tension on the wound edges with a high incidence of necrosis and breakdown.

The majority of these wounds are unsuitable for delayed suture and require skin grafting or rotation skin flaps for closure; or, if small, should be left to granulate and heal by secondary intention (Figure 12.10).

Most delayed or neglected wounds will be unsuitable for delayed primary closure.

It is in healing by secondary intention that several traditional local wound treatments may be of value, as mentioned in Chapter 11. ICRC surgeons and nursing staff have used dressings of honey or sugar, or hypertonic saline (add salt to normal saline until it no longer dissolves), which promote granulation tissue formation and have antibacterial properties. It must be emphasized that these topical treatments are an *adjunct to*, and *not a replacement for*, adequate surgery.

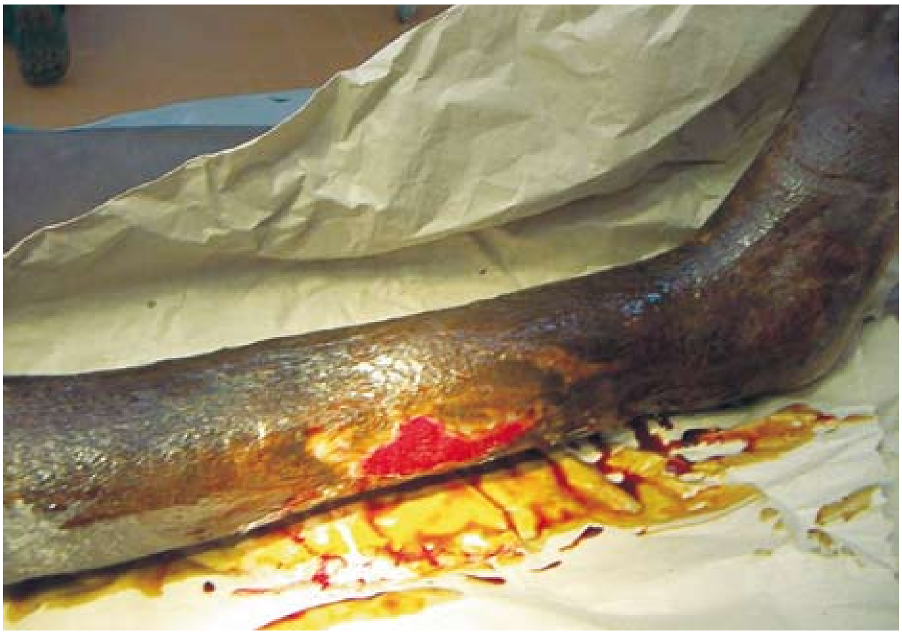
F. Jamet / ICRC

Figure 12.10
Healing by secondary intention.

Many patients with severe wound infection or frank putrefaction are malnourished, anaemic and dehydrated. As a result wound healing is poor and special measures should be taken to address these concerns.

For a discussion of post-traumatic osteomyelitis, see Volume 2.

Chapter 13

INFECTIONS IN WAR WOUNDS

13. INFECTIONS IN WAR WOUNDS

13.1 Contamination and infection 255

13.2 Major bacterial contaminants in war wounds 256

13.3 Major clinical infections of war wounds 257
13.3.1 Definitions 257
13.3.2 Gas gangrene 258
13.3.3 Tetanus 260
13.3.4 Invasive streptococcal infection 262
13.3.5 Necrotizing soft-tissue infections 263

13.4 Antibiotics 263
13.4.1 Antibiotic prophylaxis in the primary treatment of war wounds 264
13.4.2 Cephalosporins and others 265

13.5 Neglected or mismanaged wounds 265

ANNEX 13. A ICRC Antibiotic protocol 266

13.1 Contamination and infection

All war wounds are grossly contaminated by bacteria. Bullets and fragments are not sterilized on discharge and, at the entry site, bacteria are carried into the tract by the contaminated projectile itself. Furthermore, the negative pressure of the temporary cavity at both entrance and exit wounds also sucks in bacteria.

War wounds are dirty and contaminated from the moment of injury.

Other contaminants found in war wounds include clothing fragments, dust, miscellaneous organic matter (mud, grass, leaves, etc.), and wood and glass splinters from bombed buildings. Studies in Korea and Viet Nam showed that different varieties of soil are associated with different organisms – which also vary according to weather and season – with varying likelihood of producing severe infection (mud and agricultural soil contaminated with animal or human excrement are apparently the worst). In addition, the chemical composition of some soils can inactivate antibodies and impair leukocyte phagocytosis. Various home-made landmines have contained animal manure contaminant.

It is important, however, to distinguish between:
- simple contamination of a wound;
- local sepsis; and
- invasive infection.

In a contaminated wound, bacteria proliferate in dead tissues but no invasion of viable tissue takes place until the number of bacteria reaches a threshold of 10^6/gram of tissue. The presence of crushed tissues, dirt, and other foreign materials lowers the threshold of infection and invasion, by providing an environment where bacteria can grow and replicate without the patient's immune system being able to eliminate them. There are a number of *microbe-related risk factors* involved: virulence, production of enzymes and toxins, etc.

The body attempts to wall off the mass of necrotic muscle and bacteria with a fibrin barrier. Without adequate drainage, this will form an abscess; with proper drainage the necrotic mass will be expelled after 10 days.[1] In both cases, the body's natural defence mechanisms come into play to isolate the contaminants and the dead tissues.

When the local defence mechanisms can no longer cope, invasive deep tissue and systemic infection occurs. Patient immunity and resistance can also be weakened by malnutrition and chronic disease, especially HIV/AIDS. The local defence mechanisms and general resistance constitute the *host-related risk factors*.

Six hours appears to be the critical period after contamination. The efficacy of surgery and antibiotics to prevent infection decreases with time from injury, while the bacterial load increases exponentially. As with the differential grading of wounds according to the Red Cross Wound Score described in Chapter 4, a good understanding of the pathology involved is important in determining the proper and adequate procedure to follow, and the usefulness of antibiotics to *supplement* excision, drainage, and the body's natural resistance to infection. Uncomplicated soft-tissue Grade 1 wounds, as we have seen, can be treated with simple wound toilet; an antibiotic treatment helps if begun within 6 hours. Early administration of antibiotics in more severe wounds seems to inhibit bacterial growth and limit local invasion, at least temporarily. Any delay beyond six hours, however, greatly increases the risk of infection.

---

[1] Fackler ML, Breteau JPL, Courbil LJ, Taxit R, Glas J, Fievet JP. Open wound drainage versus wound excision in treating the modern assault rifle wound. *Surgery* 1989; **105**: 576 – 584.

13.2 Major bacterial contaminants in war wounds

The relative preponderance of various bacterial species differs according to geography, topology of the terrain, and climate: agricultural land, urban environment, jungle, or desert; summer or winter. The presence of community-acquired resistance in "normal" body flora also differs from region to region, and with the passing of time.

Nonetheless, certain constants remain, and the following list gives a general indication of the common pathogens involved in infections of war wounds.

Gram-positive pyogenic cocci:
Staphylococcus aureus and β-haemolytic streptococcus. These are a normal part of the skin flora. In recent years, community-acquired methicillin-resistant *Staphylococcus aureus* (MRSA) has become more of a problem in certain countries.

Gram-negative bacilli:
Escherichia, Proteus, Klebsiella, Pseudomonas, and Bacterioides are found in the gastro-intestinal tract. *Acinetobacter baumannii* is to be found in the soil and on the skin and has recently become important in causing nosocomial infections.

Gram-positive bacilli:
The Clostridia species are spore-bearing obligatory anaerobes. They are widely distributed in the environment, especially in agriculturally fertile soils, and are carried in the faeces. They are the cause of gas gangrene and tetanus.

Wounds to the upper thigh and perineum are particularly at risk from contamination by faecal organisms (Clostridia and Gram-negative bacilli), even without bowel perforation. Bowel injury demonstrably increases the risk of sepsis.

War wounds have a "bacteriological life history": contaminating bacteria change with time.

Different infections occur early or later in the wound's history.

The important point to note about this varied bacterial flora is that wound contamination is not static. That war wounds have an evolving "bacteriological life history" from the time of injury has been known since World War I and the studies made by Alexander Fleming.[2]

Many studies since have demonstrated that wound contamination – and any resulting infection – evolves with time. At the instant of wounding, there is polymicrobial contamination with particular risk from Clostridia and β-haemolytic streptococcus. Afterwards, self-contamination from skin and gastro-intestinal flora occurs; finally, hospital-acquired nosocomial bacteria may step in.

This concept is essential to understanding the roles of hygiene and environmental control measures, adequate surgery, and the use of antibiotics: i.e. the *operation-related risk factors*.[3] The widespread availability of antibiotics – and their abuse – during the last fifty years has further complicated the bacteriology of wounds because of the hospital selection of resistant strains. Whether infection follows contamination, and what type of infection develops, depends on a number of factors.

The major threats faced by the war-wounded are the development of gas gangrene, tetanus, and invasive streptococcal bacteraemia, which can all be fatal.

---

[2] Fleming A. On the bacteriology of septic wounds. *Lancet* 1915; **186**: 638 – 643.

[3] Rubin RH. Surgical wound infection: Epidemiology, pathogenesis, diagnosis and management. *BMC Infect Dis* 2006; **6**: 171. Available at: http://www.biomedcentral.com/1471 – 2334/6/171.

13.3 Major clinical infections of war wounds

The polymicrobial nature of wound contamination can result in a variety of infections. Several specific lethal infections are notorious in war wounds, especially in neglected or mismanaged wounds, and particularly if they have been stitched closed without proper excision. Common pyogenic wound infection runs through a spectrum of disease from the minor (superficial surgical site infection) to the severe (organ or space surgical site infection) and systemic (septicaemia).

13.3.1 Definitions

Simple contamination

The surgical literature classifies wounds as clean, clean-contaminated, contaminated, and infected or dirty. War wounds are considered contaminated *and* dirty.

Cellulitis/local abscess formation

Bacteria begin to spread to tissues that are contiguous to the wound, in the zone of contusion (see Chapter 3). Little systemic toxicity exists. Superficial erythema and seropurulent discharge occur; some necrotic tissue is evident, but this is related to the original injury and not to the infection. Eventually a mass of necrotic tissue and bacteria forms; if drainage is inadequate (small skin wounds), an abscess develops. In a large wound, the mass will be walled off and then expelled.

Myositis/deep tissue infection

Bacteria spread into muscles and tissues further than the permanent cavity and the zone of contusion and invade the zone of concussion and beyond. Systemic symptoms become more pronounced and, with certain infections, come to dominate the clinical picture. Overwhelming infections usually present with the classical signs of sepsis (fever, tachycardia, etc.); but they may also present without these signs, particularly in immune-compromised patients (HIV/AIDS, chronic disease, malnutrition).

In the past, deep-tissue infections were usually caused by Clostridia and invasive β-haemolytic streptococcus and accounted for the greatest amount of tissue damage, and the most severe systemic signs and symptoms. The availability of penicillin has radically changed this clinical picture.[4]

Different types of invasive tissue infection have been described:

- pyogenic deep tissue infection – this is usually a mixed infection, Gram-positive and -negative organisms, often including non-clostridial anaerobes;
- gas gangrene: this infection is still seen in neglected and mismanaged wounds;
- tetanus: poverty and the disorganizing effects of war disrupt immunization programmes in many countries – this infection is still, all too often, a major killer;
- invasive streptococcal infection – also still seen in neglected and mismanaged wounds;
- bone infection: when deep-tissue infection extends to sequestrated bone fragments chronic sepsis tends to develop with the production of a "biofilm" – this is often seen in neglected and mismanaged wounds (see Chapter 12);
- necrotizing fasciitis, synergistic gangrene and anaerobic cellulitis.

---

[4] Polhemus ME, Kester KE. Infections. In: Tsokos GC, Atkins JL, eds. *Combat Medicine: Basic and Clinical Research in Military, Trauma, and Emergency Medicine*. Totowa, New Jersey: Humana Press; 2003: 149 – 173.

13.3.2 Gas gangrene

Gas gangrene as a peculiar complication of wounds has been known since the time of Hippocrates. Its occurrence in war wounds throughout history as well as in other trauma has been widely described mainly because of its dramatic manifestations and very high associated mortality. Delay to surgery, as may occur during war or in a civilian disaster with mass casualties, greatly increases the risk of gas gangrene.

Gas gangrene is a rapidly-spreading oedematous myonecrosis occurring characteristically in severe wounds in muscles contaminated with pathogenic obligatory anaerobes, particularly *Clostridium perfringens*. Almost every case presents a mixed bacterial flora; aerobes use up available oxygen and promote the anaerobic environment required by Clostridia.

In large muscle wounds there will always be areas of ischaemia and the potential for gas gangrene. But gas gangrene can also develop even when the trauma is not very severe. If the wound is deep, contains necrotic tissue, and is *isolated from the surface*, an anaerobic environment is formed and it is possible for a clostridial infection to become established. This deep anaerobic environment is seen particularly in wounds sutured closed without debridement. The presence of foreign material in wounds, especially soil, will increase the risk of developing gas gangrene. The prolonged application of tourniquets or tight plasters and fascial compartment syndrome also present high risks. The infection is more common in wounds to the lower limb and perineum than to the upper limb.

The features of the disease result, first, from the local action of the organisms on muscle glucose, producing acid and gas, and on muscle protein, causing digestion. Second, the organisms produce soluble, very potent toxins, which diffuse into the tissues causing further tissue destruction and profound toxaemia. The muscle breakdown products are very toxic in their own right. The combination of breakdown products and specific toxins causes the relentless and profound toxaemia which, if untreated, will inevitably lead to death.

Early and adequate surgery will prevent gas gangrene in war wounds.

The *incubation period* of gas gangrene is usually short, almost always less than three days, and in the majority of cases less than 24 hours long; it may even be as short as 6 hours. Incubation periods of up to six weeks have occasionally occurred.

Typically, gas gangrene manifests itself with the sudden appearance of increasing pain in the region of the wound. The *sudden* onset of pain, sometimes so sudden as to suggest a vascular catastrophe, always suggests the possibility of gas gangrene in a wounded person. Soon afterwards the limb becomes oedematous and is associated with drainage of thin serous or serosanguinous exudate, which may become gelatinous. The pulse rate rises markedly but the patient's temperature is rarely higher than 38° C. As classically described, clinical deterioration takes place quickly and within several hours the patient becomes anxious and frightened, or may even be euphoric, and exhibits all the signs of severe sepsis.

The skin is tense, pale, often marbled with blue, and rather colder than normal. In untreated cases, local bronzing of the skin becomes more diffuse, greenish yellow areas appear in which blebs may form and become filled with dark red fluid, and patches of cutaneous gangrene may occur. The skin *may*, however, appear normal, even when lying over massive gangrene.

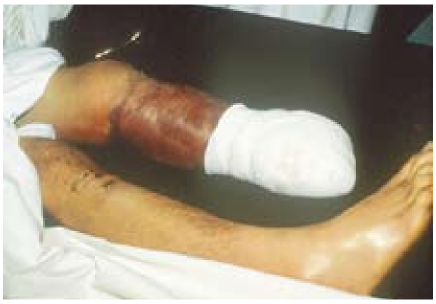

R. Coupland / ICRC

Figure 13.1.1

Gas gangrene of the leg following the application of a tourniquet.

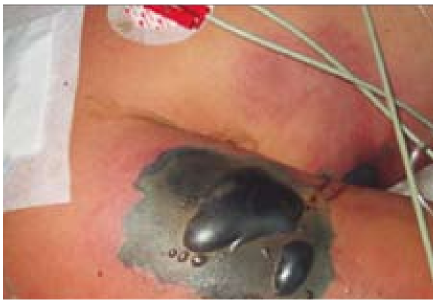

J. Goosen / South African Military Health Service, U. Witwatersrand

Figure 13.1.2

Gas gangrene of the arm.

Rapidly, swelling oedema and toxaemia increase, the serous discharge becomes more profuse and a peculiar sweetish smell may be present. The smell is variable and not pathognomonic.

Gas is usually produced at this stage and is partly responsible for the swelling of the affected part (Figure 13.2). It is produced in and between the muscle fibres, follows the fascial planes, and eventually escapes into the subcutaneous tissues under pressure through holes in the fascia, spreading rapidly beyond the confines of the infected area. Crepitus may be felt under the skin if the part is palpated. The extent of the spread of gas is not equivalent to the extent of muscle necrosis.

Infection spreads up and down the muscle from the site of the wound but has little tendency to spread to other healthy muscles; the bacteria is a *strict anaerobe*. Even in well-established gas gangrene, the bloodstream is rarely invaded by Clostridia until immediately before death.

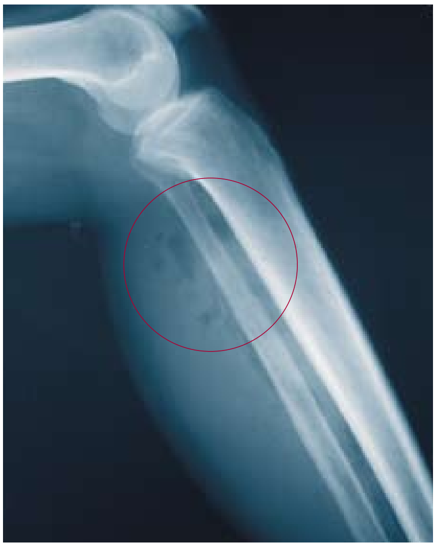

E. Dykes / ICRC

Figure 13.2

X-ray showing gas collections between muscle groups.

The muscle changes are usually only seen at operation. In the early stages, they consist of little more than swelling and pallor. Later, the colour alters to a lustreless pinkish-grey, then to brick red, and finally to a dark green-purple colour.

Management

All patients with missile wounds should receive prophylactic antibiotics that are effective against Clostridia, such as penicillin, metronidazole, or erythromycin. But the antibiotics can only reach tissues with a blood supply and good tissue perfusion. Antibiotic concentrations in the dead muscle in the depths of an anaerobic environment are ineffective. Surgery as early as possible is the goal to be attained.

The basis of management of established gas gangrene is *thorough* excision of dead tissue, which might necessitate *urgent amputation*. The limit of excision is *necrotic muscle*, not the extent of tissues containing gas. Tissues must be cut away until healthy red, bleeding, and contracting aerobic muscle is reached.

Immediate surgical intervention is essential in the management of gas gangrene.

It is often too late for surgery when a patient presents with full-blown gas gangrene because of the severe toxaemia. Nonetheless, supportive measures including oxygen, fluids, and fresh whole blood should be instituted. Appropriate antibiotics should be continued.

13.3.3 Tetanus

Tetanus is always a risk with any penetrating wound. The risk is greater for severely contaminated wounds, especially in small, deep, punctuate-type ones. Pyogenic infection in the depth of a narrow track can create the necessary anaerobic environment. The incubation period is 3 to 21 days, but can be as short as 1 day to as long as several months.

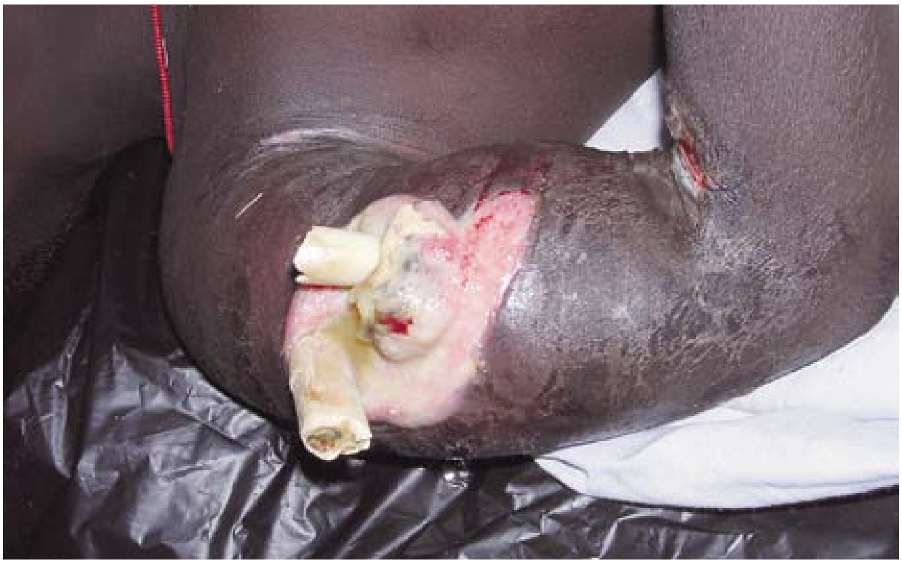
E. Dykes / ICRC

Figure 13.3

Badly infected wound giving rise to tetanus.

Virtually total protection can be obtained by active immunization with a full course of tetanus toxoid injections.

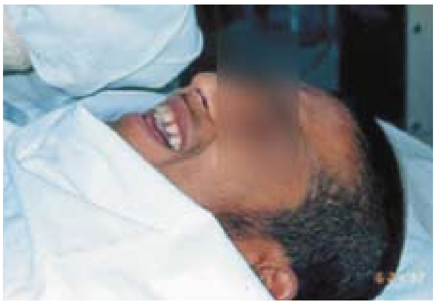
Å. Molde / ICRC

Figure 13.4

Risus sardonicus in a patient suffering from tetanus.

Clostridium tetani is a strict anaerobe. The organism produces an extremely potent toxin called tetanospasmin that spreads along peripheral nerves to the spinal cord and brainstem. The toxin affects the motor end-plate by inhibiting cholinesterase, resulting in a build-up of acetylcholine and tonic muscle spam. There is, in addition, hyperexcitability of lower motor-neurones causing muscular rigidity and dysfunction of activity in antagonistic muscles that cause unopposed reflex activity, giving rise to the typical spastic phenomena of tetanus. Early signs are neck rigidity and *trismus* (stiffness of the jaw) with difficulty in swallowing, followed by *risus sardonicus* (grimacing smile). Profound autonomic nervous system dysfunction is also present in severe cases with unstable pulse rate, blood pressure, and temperature, which are difficult to treat. Once the toxin is fixed to the nerves, antitoxin can not longer neutralize it.

The infection may affect a single muscle group or be more generalized. Three clinical degrees are described:
- mild = no generalized spasms;
- moderate = generalized spasms on provocation;
- severe = spontaneous generalized spasms with opisthotonus.

The muscular spasms are quite painful, and can last a short time or go on for several weeks. The greatest danger is asphyxiation due to laryngeal spasm or aspiration. Frequently, there is a high temperature with profuse sweating that requires careful fluid replacement. The patient remains conscious throughout.

Prophylaxis

The best precaution against tetanus is active immunization. However, many patients in developing countries where conflict has further disrupted public health programmes are not immunized and therefore the risk of tetanus in war wounds is great. In some countries, ICRC practice is to proceed with active and passive immunization for *all* patients.

Early and adequate excision of dead tissue, and leaving wounds open, is vital to prevent its development. This is particularly important in small, deep, punctuate-like wounds. Penicillin and metronidazole are the drugs of choice.

Tetanus prophylaxis for war wounds

All patients whatever their immunization status:

1. tetanus toxoid vaccine 0.5 ml i.m. (5 LF units) – this is a booster dose in a previously vaccinated patient;
2. penicillin;
3. thorough excision of the wound.

Non-immunized patients or those in whom immunization status is in doubt, addition of:

4. human tetanus immunoglobulin 500 IU i.m. (adults) or 250 IU (children under 15 years of age) – also known as human anti-tetanus serum;
5. tetanus toxoid 0.5 ml i.m. to be repeated at four weeks and again at six months.

Please note:

Vaccine and immunoglobulin should be administered through separate syringes and at separate sites.

Treatment of established tetanus

Treatment of established tetanus should follow the steps listed below.

1. Extensive debridement of the wound. Ketamine anaesthesia should be avoided if possible as patients waking from this form of anaesthesia are often restless and have hallucinations and this may stimulate spasms. Spinal anaesthesia or a regional nerve block are excellent methods.

2. Antibiotics: crystalline penicillin G (5 MIU. i.v. QID) and/or metronidazole (500 mg i.v. TID). Erythromycin, tetracycline and chloramphenicol are also active against Clostridia, in case of allergy to penicillin.

3. Anti-tetanus human immunoglobulin (3,000 – 10,000 IU): a single large dose is given intravenously and as soon as possible. The actual dose ordered is dependent on the severity of the disease and the age of the patient. It is diluted in 20ml of normal saline and given slowly over a period of 15 minutes. This can be given intramuscularly (undiluted) in the proximal part of the wounded limb if the type of immunoglobulin is not recommended for intravenous use. An intrathecal injection of 250 IU can be given e.g. with spinal anaesthesia for debridement of the lower part of the body.

   In some regions, human immunoglobulin is in short supply or not available at all. In these circumstances, one must rely on equine anti-tetanus serum. A test dose must be administered before giving the full dose (20 000 IU).

4. Control of the spasms: the patient should be nursed in an environment as devoid of stimulation as possible in a dark, quiet room. For severe cases, and in the best of circumstances, the patient should be in an intensive care unit, anaesthetized and paralysed, with mechanical ventilation. Many hospitals lack these facilities; control is based on sedation.

Intravenous diazepam, either intermittently or continuously, is used with doses up to 20 mg or more per hour depending on the duration and intensity of the spasms. If diazepam alone does not control the spasms, chlorpromazine may be added, or thiopental (pentothal) may be required. This latter is given preferably by continuous infusion, which has been found to control the spasms more effectively with lower total dosages than if given intermittently.

Warning!
Oversedation can be as fatal as the disease.

Recently, ICRC teams have used magnesium sulphate in high doses (40 mg/kg over 30 minutes, followed by i.v. infusion of 1 – 3 g/hour) with encouraging results. This regime allows lower doses of diazepam and thiopental to be used, thus avoiding oversedation: this line of treatment is still under investigation.[5]

5. Airway management is of the utmost importance to prevent aspiration. If laryngospasms persist tracheostomy may be needed, and the decision to perform one should be taken early. Care of the tracheostomy tube is instituted with particular attention to how much this stimulates spasms, and adapting the sedation accordingly. Chest physiotherapy is required to prevent respiratory complications.
6. Fluid replacement: there may be excessive fluid loss due to sweating – this needs to be carefully monitored along with urinary output. A Foley urinary catheter should be placed.
7. Nutrition: if the spasms continue for some days, a feeding gastrostomy or jejunostomy may be required as a nasogastric tube may provoke spasms and become obstructed. Again, it is important to decide to perform this procedure early. These patients require a high calorie diet.
8. Excellent nursing care is fundamental. The patient's position must be changed four-hourly. Extra sedation may be required prior to this, as movement can provoke spasms. Patient hygiene – skin, oral and bowels – is essential.

Please note:

Clinical tetanus does *not* confer normal immunity. The patient therefore needs to complete immunization after recovering from the disease.

13.3.4 Invasive streptococcal infection

Beta-haemolytic streptococcus produces powerful toxins that trigger off a rapid progression of disease from local to systemic infection. It begins with a local cellulitis with severe pain out of proportion with physical findings, and advances rapidly to fever, tachycardia, disorientation and delirium. The wound is malodorous with large quantities of thin, blood-stained seropurulent discharge. The muscles suffer from a moist oedema, and turn from pale, to bright red, then dark purple-brown. Progression can occur in a matter of hours. It is still seen in mismanaged wounds, and especially if there is a delay in treatment.

Wound debridement, drainage, and massive doses of penicillin are the basis of management, along with supportive measures. Blood transfusion is usually necessary.

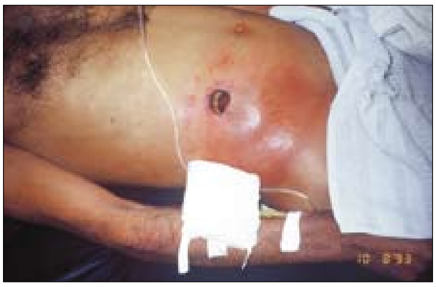
ICRC

Figure 13.5

Invasive β-haemolytic streptococcus infection of the abdominal wall.

---

[5] Thwaites CL, Yen LM, Loan HT, Thuy TTD, Thwaites GE, Stepniewska K, Soni N, White NJ, Farrar JJ. Magnesium sulphate for treatment of severe tetanus: a randomised controlled trial. *Lancet* 2006; **368**: 1436 – 1443.

13.3.5 Necrotizing soft-tissue infections

Various non-clostridial spreading gangrenous infections have been described in the surgical literature: necrotizing fasciitis, synergistic gangrene, anaerobic cellulitis, etc. There is necrosis of the skin, subcutaneous tissues and fascia, but not of the muscles. It can rapidly progress to systemic toxicity and even death. Spreading gangrene can follow trauma or thoracic and abdominal surgery (postoperative synergistic gangrene), and affect persons suffering from a general disease, such as diabetes, or malnutrition; alcoholics are particularly at risk. Infection usually affects the torso, perineum or extremities. The bacterial flora is mixed causing synergy amongst various organisms, anaerobic and aerobic, particularly microaerophilic nonhaemolytic streptococci in association with haemolytic staphylococci.

In no case are healthy muscles involved, and the clinical picture is less dramatic than in gas gangrene. Nonetheless, the systemic toxicity is life-threatening. There is severe pain disproportionate to local findings, crepitus can often be felt and soft-tissue air is present on plain X-rays. The skin is discoloured (blue, purple or black) with blistering leading to haemorrhagic bullae and induration. The necrosis can spread widely causing a large soft tissue defect. The diagnosis is primarily a clinical one, and the earlier treatment is instituted the better the result.

Management includes:

- aggressive excision of necrotic tissue and the relief of tension, which may require serial debridements – amputation may be necessary in cases of necrotizing infections of the limbs – and defunctioning colostomy in cases of infection of the perineum with faecal contamination;
- leaving the wound open for drainage;
- triple intravenous antibiotics: penicillin, gentamycin and metronidazole;
- fluid resuscitation, blood transfusion and other supportive treatment.

Reconstructive surgery to cover the defect should only be considered once the patient's condition has stabilized and the infection is fully eradicated.

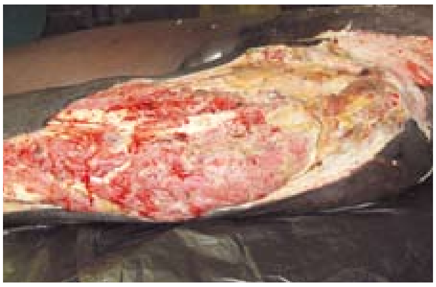

H. Nasreddine / ICRC

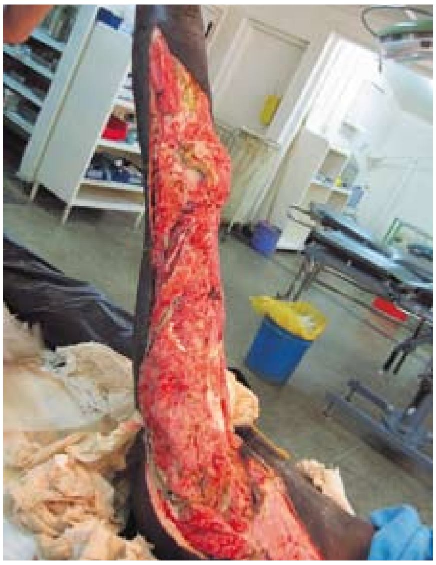

H. Nasreddine / ICRC

Figure 13.6.1 and 13.6.2

Necrotizing fasciitis with large skin loss and tissue defect.

13.4 Antibiotics

It is important to distinguish between the prophylactic use of antibiotics and their role in the therapy of established sepsis. The surgeon must also remember the life history of wounds and their changing bacterial flora.

Prophylaxis is aimed at preventing a specific infection. It is impossible to find any reasonable cocktail of antibiotics that would be effective against the entire "polymicrobial cesspool" that can contaminate a war wound. Nor is it desirable to do so. Such a practice would constitute a simple abuse of antibiotics and contribute to the development of bacterial resistance.

Old lessons for new surgeons

The best antibiotic is good surgery.

The surgeon must realize that trust in the efficiency of antibiotics will never replace good surgery. Good surgery involves good diagnosis, good clinical decision-making, and good holistic management of the patient. Sometimes "good surgery" means knowing when *not* to operate. As we have seen, uncomplicated Grade 1 wounds and many civilian gunshot wounds can be treated conservatively and expectantly.

War wounds, however, are usually understood to be different from civilian weapon trauma; not always because of the weapon, but because of the environment. The battlefield is a dirty and contaminated place and the danger of invasive infection is always present, even in minor wounds; mass casualties cannot always be followed up correctly; there is often a long delay between injury and treatment; proper hygiene and nutrition cannot always be maintained; and immunization is not always universal.

Under such circumstances, the basis of prevention of primary infection remains complete wound excision and good drainage, respect in the handling of the tissues, and leaving the wound open for delayed primary closure. Antibiotics are only an *adjunct* to good surgical practice, and cannot replace poor surgery.

A case can be made for prophylaxis against the second series of infections, those acquired from the flora of the patient (skin, respiratory and gastrointestinal tracts) if the occurrence of such infections becomes an important clinical problem in the functioning of a given hospital, and is proven by proper bacteriological studies. Infections caused by bacteria with multiple antibiotic resistance and by opportunistic organisms, such as *Pseudomonas aeruginosa*, have become more prevalent with the uncontrolled use of wide-spectrum antibiotics. Again, good surgery and proper hygiene measures and environmental control cannot be replaced by antibiotics, and their use should be regarded as an adjunct only.

The prevention of nosocomial infection is another story. Proper clinical procedures and hygiene are the correct means of prevention: frequent hand washing; no unnecessary change of dressings on the ward; isolation of infected patients; adequate sterilization; proper cleaning of the hospital premises, etc. These alone can and will stop nosocomial infection. The use of antibiotics to *supplement* such measures will depend on the *virulence* of the particular bacteria involved.

A well-functioning bacteriology laboratory is important if antibiotic use is to be anything more than a “shotgun” approach or an “educated guess”. Wound cultures are notorious for not predicting subsequent infections or infecting pathogens. In the absence of correct collection of specimens, aerobic and anaerobic, and proper bacterial culture and sensitivity techniques, an educated guess is the best that can be hoped for in most cases. Accurate bacteriological capability is a far more difficult endeavour in a forward hospital than most people realize.

13.4.1 Antibiotic prophylaxis in the primary treatment of war wounds

Antibiotics do not reach the source of infection in a missile wound – i.e. the culture medium of dead tissue, debris and foreign material – they only affect the contusion and concussion zones around the wound. However, early administration of antibiotics seems to inhibit bacterial growth and delay invasive infection. Antibiotics especially help prevent spread to the blood stream. The benefit of antibiotics to prevent invasive infections in wounds that received no other treatment for hours to days after injury has been demonstrated,[6] and ICRC clinical experience confirms this. The ICRC recommends that penicillin treatment be begun in the pre-hospital setting if at all possible.

As mentioned, historically, the biggest killers of surviving war-wounded casualties have been primary infection by β-haemolytic streptococcus and Clostridia. This is a relatively narrow spectrum of bacteria, for which penicillin remains the best antibiotic.

Prophylaxis for war wounds can only be assured for Clostridia and β-haemolytic streptococcus: penicillin is the drug of choice.

---

[6] Mellor SG, Cooper GJ, Bowyer GW. Efficacy of delayed administration of benzylpenicillin in the control of infection in penetrating soft-tissue injuries in war. *J Trauma* 1996; **40 (3 Suppl.)**: S128 – S134.

Single-dose or the administration of prophylactic antibiotics limited to 24 hours are indicated under optimal conditions of rapid evacuation, early pre-hospital first aid, and adequate infrastructural hygiene. In the ICRC context of limited resources, often less than ideal environmental control, and delayed evacuation, antibiotics are usually given for 5 days, until delayed primary closure. This period of 5 days may be regarded as an exaggeration by some colleagues, but is considered a reasonable combination of a prophylactic *and* therapeutic regime by ICRC practitioners and has proved its worth in clinical experience (see Annex 13. A: ICRC antibiotic protocol).

Topical antibiotics and the washing of wounds with antibiotic solutions are not recommended.

13.4.2 Cephalosporins and others

The ICRC does not routinely use cephalosporins or other sophisticated antibiotics in a surgical setting. In many areas where the ICRC works, typhoid fever and shigellosis, amongst other infections, are endemic diseases. Resistance to first-line antibiotics for these fatal diseases is becoming more and more of a problem in clinical practice. Often, cephalosporins are the only recourse left available. The judicious use of antibiotics is an important element to keep in mind when dealing with the surgery of the victims of war, where proper surgical, nursing, and hygiene measures should be the basis of prevention and treatment of infection.

13.5 Neglected or mismanaged wounds

The particular, although commonly encountered, condition of neglected or mismanaged wounds has been dealt with in Chapter 12. As mentioned above, these wounds are particularly at risk of developing gas gangrene, tetanus, and deeply invasive sepsis. According to the ICRC antibiotic protocol, penicillin and metronidazole should be administered; if there are systemic signs of pyogenic infection, gentamycin is added.

Antibiotics are important in the treatment of these diseases, but do not replace the essentials of septic surgery: wide excision, good drainage, copious lavage, and leaving the wound wide open.

ANNEX 13. A ICRC Antibiotic protocol

Injury	Antibiotic	Remarks
Minor soft-tissue wounds, uncomplicated Grade 1	Penicillin-V tablets 500 mg QID for five days	Anti-tetanus measures for all weapon - injured patients
Compound fractures Traumatic amputations Major soft-tissue wounds (Grades 2 and 3)	Penicillin-G 5 MIU i.v. QID for 48 hours Follow with penicillin-V tablets 500 mg QID until DPC	Continue penicillin-V for five days if closure is performed with a split skin graft If redebridement is performed instead of DPC: stop antibiotic unless there are signs of systemic infection or active local inflammation – in latter case, add (metronidazole 500 mg i.v. TID) and gentamycin 80 mg i.v. TID
Compound fractures or major soft-tissue wounds with delay of more than 72 hours Anti-personnel landmine injuries to limbs whatever the delay	Penicillin-G 5 MIU i.v. QID and metronidazole 500 mg i.v. TID for 48 hours Follow with penicillin-V tablets 500 mg QID and metronidazole tablets 500 mg TID until DPC	
Haemothorax	Ampicillin 1 gm i.v. QID for 48 hours, followed by amoxycillin tablets 500 mg QID	Continue until two days after removal of the chest tube
Penetrating cranio-cerebral wounds	Penicillin-G 5 MIU i.v. QID and chloramphenicol 1 gm i.v. TID for at least 72 hours	Continue i.v. or orally according to patient's condition for a total of 10 days
Brain abscess	Same regime as for penetrating cranio-cerebral wounds plus metronidazole 500 mg i.v. TID	Continue i.v. or orally according to patient's condition for total of 10 days
Penetrating eye injuries	Penicillin-G 5 MIU i.v. QID and chloramphenicol 1 gm i.v. TID for 48 hours	Continue i.v. or orally according to patient's condition for total of 10 days Local instillation of antibiotic eye-drops
Maxillo-facial wounds	Ampicillin 1 gm i.v. QID and metronidazole 500 mg i.v. TID for 48 hours	Continue i.v. or orally according to patient's condition for total of 5 days
Abdominal wounds:		Continue for 3 days depending on drainage
1. solid organs only: liver, spleen, kidney; or isolated bladder injury	Penicillin-G 5 MIU i.v. QID	
2. stomach, small intestines	Ampicillin 1 gm i.v. QID and metronidazole 500 mg i.v. TID	
3. colon, rectum, anus	Ampicillin 1 gm i.v. QID and metronidazole 500 mg i.v. TID and gentamycin 80 mg i.v. TID	

MIU = million international units.

QID = quater in die (4 x day).

TID = ter in die (3 x day).

Please note:

This protocol was established by the ICRC Master Surgeons Workshop held in Geneva in 2002 (see Introduction).

Chapter 14

RETAINED BULLETS AND FRAGMENTS

14. RETAINED BULLETS AND FRAGMENTS

14.1 The surgeon and the foreign body 271

14.2 Early indications for removal 271

14.3 Late indications 274

14.4 Technique for the removal of a projectile 275

14.1 The surgeon and the foreign body

There are hundreds of thousands of perfectly healthy people walking around the present and past war zones of the world who have retained projectiles in their body. These metallic foreign bodies often catch the attention of patients and relatives, who impute to them any pains and disabilities they suffer from and insist upon their removal, even after the wound has completely healed.

It can be very difficult, and frustrating, for the surgeon to point out to an insistent patient that a projectile is dangerous when moving but is no longer so when it has come to a stop. The incidence of late infection is low (2 – 3 %)[1] and vascular embolization rare and more common with fragments than bullets. The risk of embolization – "migrating bullet" – is more common in published clinical cases in the specialized surgical literature than it is in real life. One American study from Viet Nam reported a rate of 0.3 % in 7,500 cases with arterial injuries; and rarer still in the overall wound population.[2]

The frustration is directly related to the number of patients presenting with such requests, who can at times be counted literally in the hundreds. The problem tends to be psychological and cultural rather than surgical. How can one convince a young combatant that the operation to remove a bullet may cause more damage than the bullet has accomplished? Such a foreign body in the midst of a muscle mass, where there are no pain receptors, is not painful and the surgeon must find a way to overcome the patient's subjective fixation; perhaps by making the bullet's presence a matter of "pride".

The experience of the ICRC shows that many of these operations to remove a projectile are risky, useless, consume time and materials, and result in new surgical trauma with its possible complications, often without identification and removal of the foreign body. There are few surgical endeavours more frustrating for an unprepared young surgeon than to spend two hours searching ever more frantically for an "easy" fragment or bullet, and finally not finding it.

14.2 Early indications for removal

However, as stated previously, there are indications for removal of bullets and fragments, which may be divided into early and late. The most important early indications have been dealt with in Chapter 10: as an integral part of the primary operation or to address the possibility of erosion of an important structure (Figures 10.11, 10.12 and 14.1).

As mentioned in the latter case, whether the removal is an immediate or planned procedure will depend on a number of factors, especially the experience of the surgeon and the possible morbidity and lethality of the operation itself compared to the relatively low incidence of serious complications. The decision is not always a simple one and the reader may consider what should be done in the following examples (Figures 14.2 – 14.8).

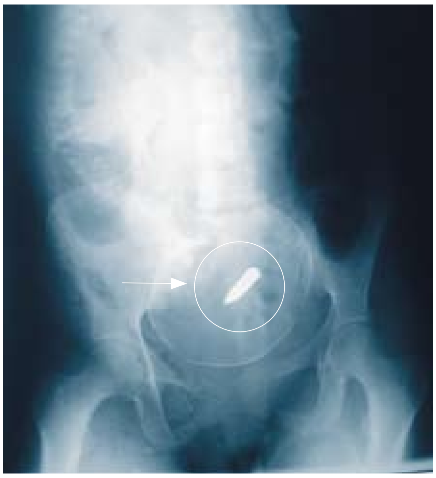
M. Baldan / ICRC

Figures 14.1.1 and 14.1.2

Bullet lodged in full-term pregnant uterus.

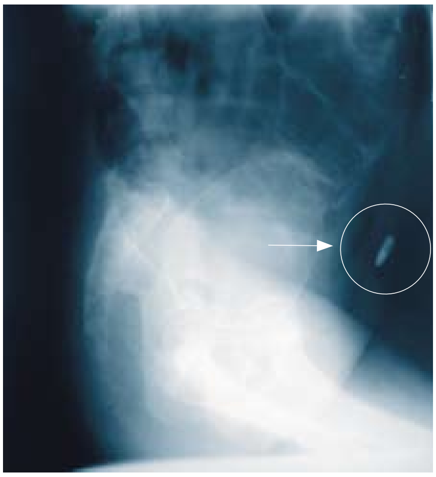
M. Baldan / ICRC

---

1 Rhee JM, Marin R. The management of retained bullets in the limbs. *Injury* 1997; **28**: 23 – 38.

2 Rich NM, Collins GJ, Andersen CA, McDonald PT, Kozloff L, Ricotta JJ. Missile emboli. *J Trauma* 1978: **18**: 236 – 239.

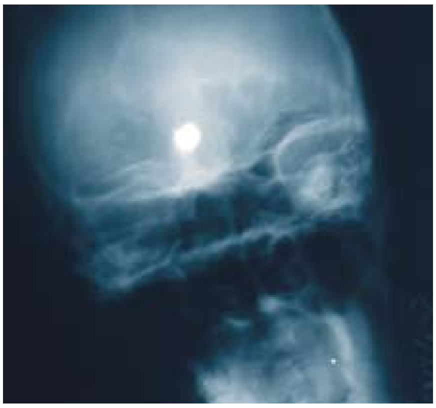
M. Baldan / ICRC

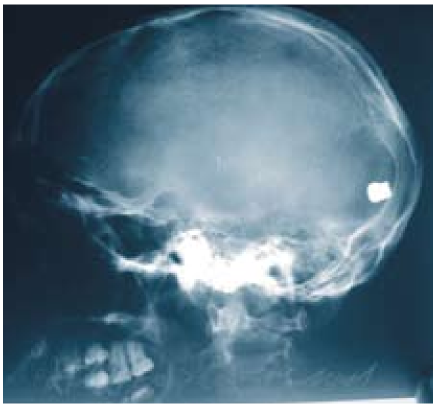
M. Baldan / ICRC

Figures 14.2.1 and 14.2.2

A fragment entered the forehead and is retained in the occipital region.

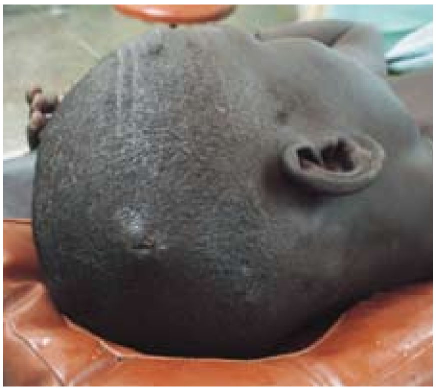
ICRC

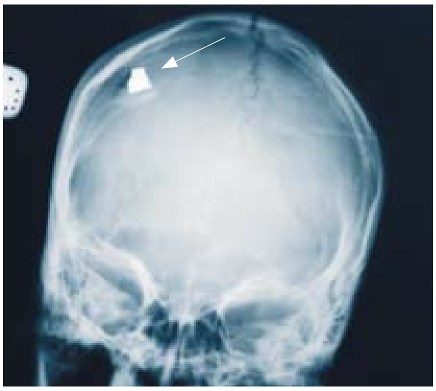
ICRC

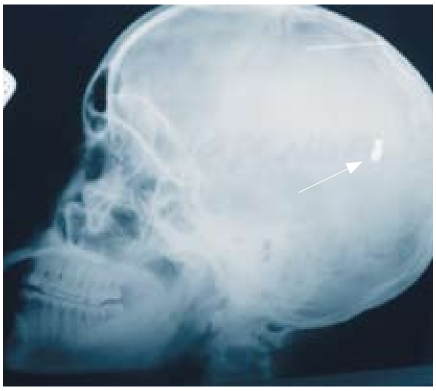
ICRC

Figures 14.3.1 – 14.3.3

A fragment has entered the parieto-occipital area.

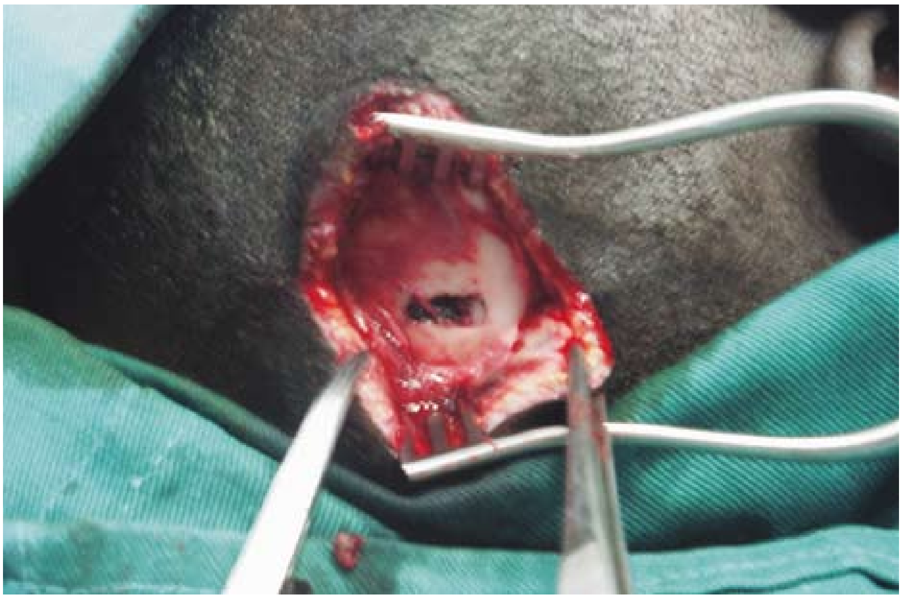
ICRC

Figures 14.3.4

The fragment is evident under the bone.

R. Coupland / ICRC

Figure 14.4.1

Nurse pointing out site of fragment entry.

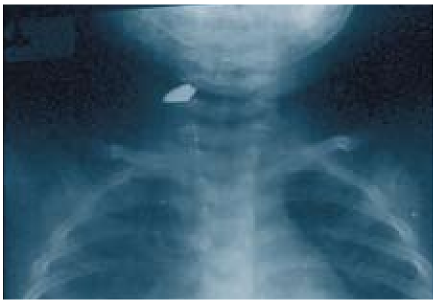

R. Coupland / ICRC

Figure 14.4.2

Radiograph showing a fragment retained in the neck.

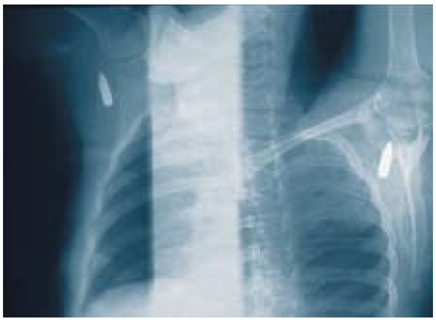

M. Baldan / ICRC

Figure 14.5

Bullet in the axilla: the vessels and nerves are intact.

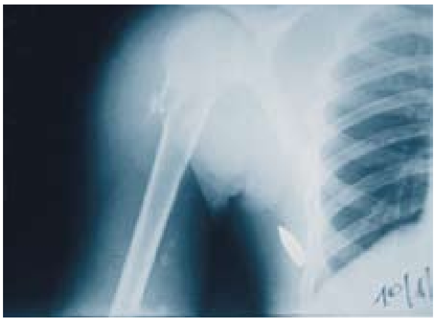

M. Baldan / ICRC

Figure 14.6

Fractured humerus with bullet retained in the subscapular muscles.

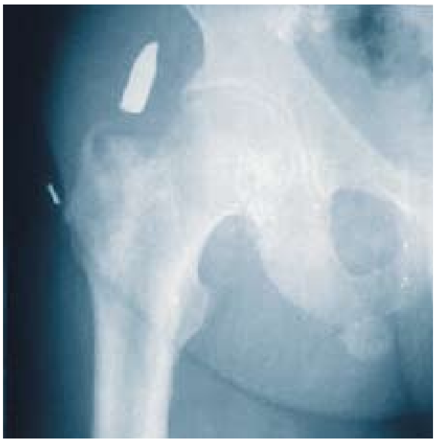

M. Baldan / ICRC

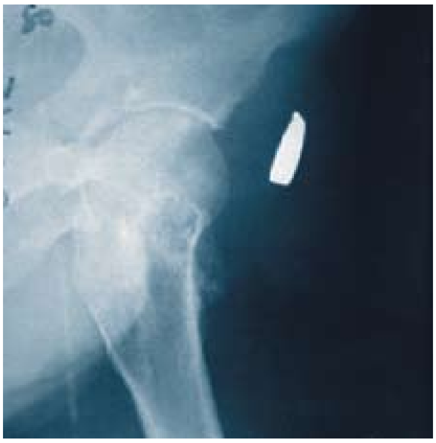

M. Baldan / ICRC

Figures 14.7.1 and 14.7.2

Extra-synovial bullet near the hip.

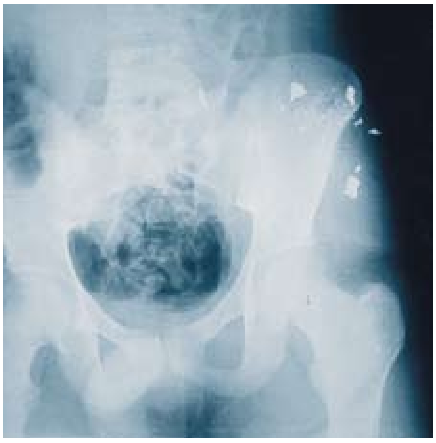

F. Herkert / ICRC

Figure 14.8

Fragmented bullet in extraperitoneal muscle mass.

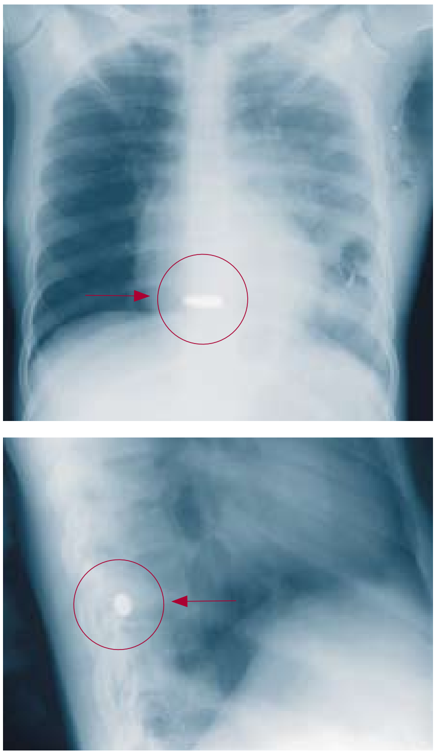
F. Irmay / ICRC

F. Irmay / ICRC

Figures 14.9.1 and 14.9.2

Bullet retained in vertebral column, the patient is paraplegic.

The following scenarios should be added to the early indications already mentioned.

- A small fragment in a heart chamber – the concomitant pericardial tamponade is an urgent condition, the removal of a retained fragment is not – or a fragment obviously within the lumen of an important vessel. The possibility, and danger, of embolization is high in these particular instances. Operation should *only* be undertaken, however, if the surgeon *and* anaesthetist are experienced and the necessary diagnostic and operative equipment are available.
- A projectile located in the spinal cord: emergency laminectomy and removal should be considered *only* if there is clear progressive neurological deficit and radiological evidence of spinal cord compression by the foreign body *and* the surgeon is experienced in this field. A definitive diagnosis of paraplegia, indicating that the spinal cord is irrecoverably sectioned, means that it is already too late.
- Small metallic fragment penetrating the eye: removal is indicated if located in the anterior chamber, again, *only* if the surgeon is experienced, *and* magnification, proper instruments and suture material are available.

14.3 Late indications

Late indications are largely the result of the complications listed below.

Sepsis

If the projectile, along with surrounding contaminants, is acting as a nidus for infection: an abscess or sinus. The removal is performed as a planned operation after the necessary diagnostic procedures (X-ray, sinugram, etc.).

Pain

If located superficially, especially if over a pressure point (palm of hands, sole of foot, elbow, etc.), the projectile will cause true pain. The removal may be performed as an elective procedure after healing from the acute trauma.

Pain

If the foreign body is impinging on a nerve and causing radicular pain or paraesthesia. As above, the removal is performed as a planned operation.

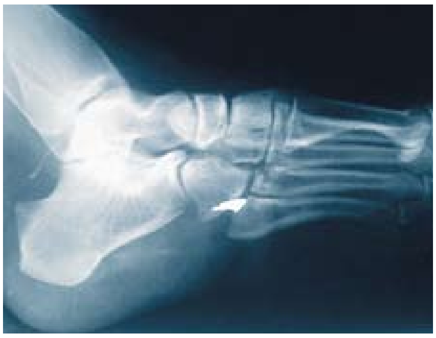
ICRC

Figure 14.10

Metallic fragment located in a pressure point: sole of the foot.

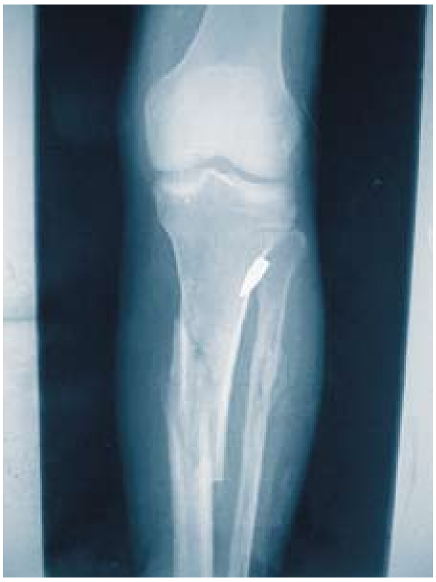
ICRC

Figure 14.11.1

Bullet in antero-lateral compartment of the leg pressing on common peroneal nerve.

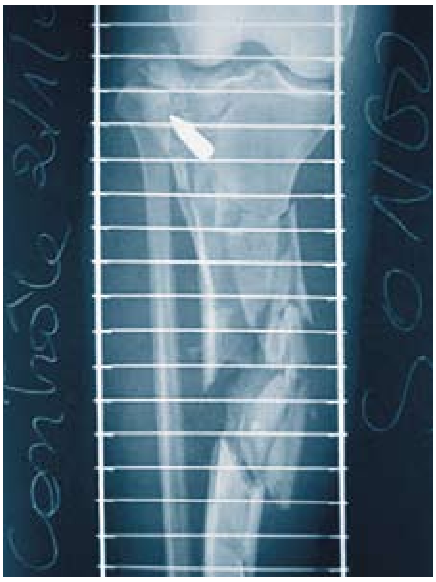
ICRC

Figure 14.11.2

Bullet not pressing on a nerve.

Lead toxicity

This is extremely rare, usually involves a synovial joint or intervertebral disc, and surgery should be contemplated only if there is documented elevation of serum lead level (above 10 micrograms/dl in children and 40 micrograms/dl in adults). The response to an EDTA (ethylene diamine tetra-acetic acid) challenge is a useful diagnostic test. Standard chelation therapy is instituted (EDTA, dimecaprol, d-penicillamine, succimer) and the surgical removal of the metal should not be performed before serum lead levels have been reduced to avoid acute lead poisoning.[3] This complication does not arise with an intact, full metal jacket bullet where no lead filling has spilled out.

We shall not deal with wounds not related to armed conflict where the bullet may be required as part of a legal and forensic investigation. National law of the country concerned determines what exactly can be done while not subjecting the patient to unnecessary harm and suffering.

14.4 Technique for the removal of a projectile

For those patients requiring removal of a deeply-seated metallic fragment or bullet, proper localization prior to surgery is of the utmost importance. Most rural or provincial hospitals do not have fluoroscopy or an image-intensifier in the operating theatre enabling the surgeon to operate under direct vision.

A simple stereotactic technique to assist localization involves taping a series of radio-opaque objects (paper clips, injection needles, Kirschner wires) to the body part, both anterior and lateral surfaces, to act as guides. Plain X-rays with antero-posterior and lateral views are taken. The radio-opaque object is then removed and its position on the skin marked with a felt pen.

By looking at the two X-ray views, the surgeon can estimate the relative distance of the bullet from the radio-opaque objects in the axes both across and in the depth of the body part – e.g. half-way between the second and third paper clips in the A-P view and one-third of the way between the first and second paper clips in the lateral view. The idea is to use two-dimensional X-rays to extrapolate into three dimensions. The films should be taken on the morning of the operation and the patient should be undressed (combatants often carry bullets in their pockets).

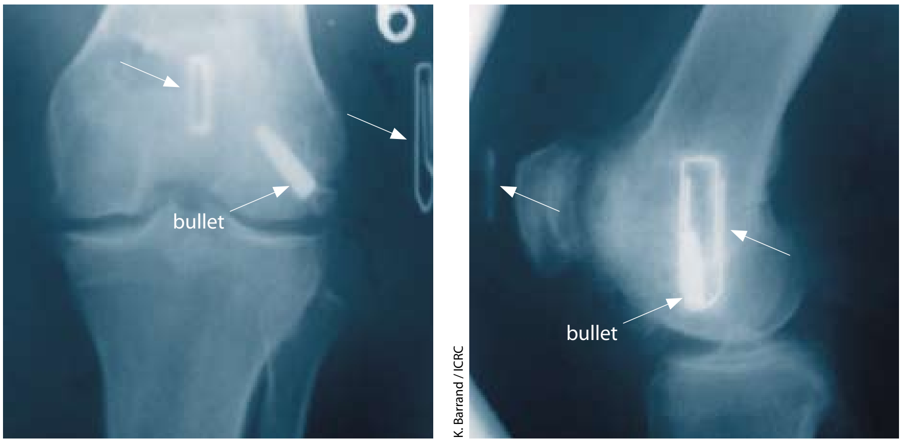

K. Barrand / ICRC

K. Barrand / ICRC

Figure 14.12.1 and 14.12.2

Stereotactic localization of metallic foreign body. Safety-pin is anterior and paper clip lateral.

In due course, the body will form tough avascular scar tissue around the metallic fragment: a foreign body granuloma. This includes other foreign matter and dirt, and should be excised together with the projectile.

---

[3] Linden MA, Manton WI, Stewart RM, Thal ER, Feit H. Lead poisoning from retained bullets: Pathogenesis, diagnosis, and management. *Ann Surgery* 1982; **195**: 305 – 313.

Chapter 15

BURN INJURIES

15. BURN INJURIES

15.1 Introduction 279

15.2 Pathology 279
15.2.1 Burn depth 279
15.2.2 Pathophysiological changes 280
15.2.3 Types of burns 281
15.2.4 Extent of burns 281

15.3 Burn management 282
15.3.1 First aid 282
15.3.2 Resuscitation 282
15.3.3 Initial fluid replacement therapy 284
15.3.4 Monitoring resuscitation 285
15.3.5 After 48 hours 285

15.4 Burns presenting late 286

15.5 Nutrition 286

15.6 Care of the burn wound 286
15.6.1 Initial wound management 287
15.6.2 Local care 288

15.7 Closure of the burn wound 289
15.7.1 Mechanical cleansing and debridement 290
15.7.2 Surgery 290

15.8 Scar management 293

15.9 Electrical burns 294

15.10 Chemical burns 294
15.10.1 Acid and alkali burns 294
15.10.2 Phosphorus burns 294
15.10.3 Napalm injuries 295
15.10.4 Magnesium 296
15.10.5 Vesicant chemical weapons 296

ANNEX 15. A Nutrition in major burns: calculating nutritional requirements 297

15.1 Introduction

Burns are a common event in a war environment. Flame weapons, explosive blast and ignition of combustible materials all create burn hazards. The causal agent may be thermal, chemical, electrical, or linked to radiation. Each cause has specific consequences that may require specific assistance.

A serious burn injury is a painful and life-threatening event requiring an inordinate amount of hospital resources and nursing care. The most immediate threat to life is an impaired airway due to inhalation of hot air and smoke with consequent oedema; the most important subsequent threats to life are from hypovolaemic shock and infection, followed by the complex pathophysiological effects which continue after injury. It is associated with many complications, prolonged morbidity, multiple operations, and great demands on equipment, materials, medical and nursing time. The long-term sequelae, physical, cosmetic and psychological, deeply affect the morale of both patients and staff. Modern burn centres have made great progress in the successful treatment of major burns, but such facilities are never available in situations of limited resources.

Nevertheless, the principles of treatment are the same, and the aim is to do the best one possibly can under the austere circumstances, where stark reality imposes its frustrating limitations. Surgery will have the greatest impact on young people with small but potentially crippling injuries such as bilateral hand burns. Patients with burns covering 40 – 50% of the total body surface will rarely survive under field conditions and are best treated by giving sufficient fluids to relieve thirst, and generous amounts of analgesia. In a triage situation of mass casualties they are considered Category IV: i.e. given supportive treatment only.

15.2 Pathology

15.2.1 Burn depth

Burns involve varying amounts of injury to the skin, partial or full thickness, and classically correspond to three degrees of burn depth of increasing severity (Figure 15.1).

First degree burns:
First degree burns, or superficial burns, are painful, red, and have no blisters. They will heal on their own.

Second degree burns:
These may be superficial or deep partial-thickness burns. They are invariably blistered and the blister floor is usually pink or mottled red in appearance and the surface is moist. The less they blanch with pressure, the deeper they are. They are painful and have some sensation to pinprick. Hairs will be difficult to pluck out if the base of the follicle is alive. Most will eventually heal by a combination of re-epithelialization and wound contraction, but often with a seriously disabling scar contracture. Some will benefit from appropriate skin grafting.

Third degree burns:
These refer to full-thickness burns. Destruction of the full thickness of the skin gives a charred, leathery or waxy appearance. Any surviving hairs pull out easily. The burns are usually dry and have no sensation. Injury may extend to the muscles and deeper tissues. They are normally caused by flames, immersion in very hot liquids, electric current or chemicals. Smaller full-thickness burns will eventually heal by contraction but this inevitably causes severe deformity and loss of function. They are best treated by skin grafting.

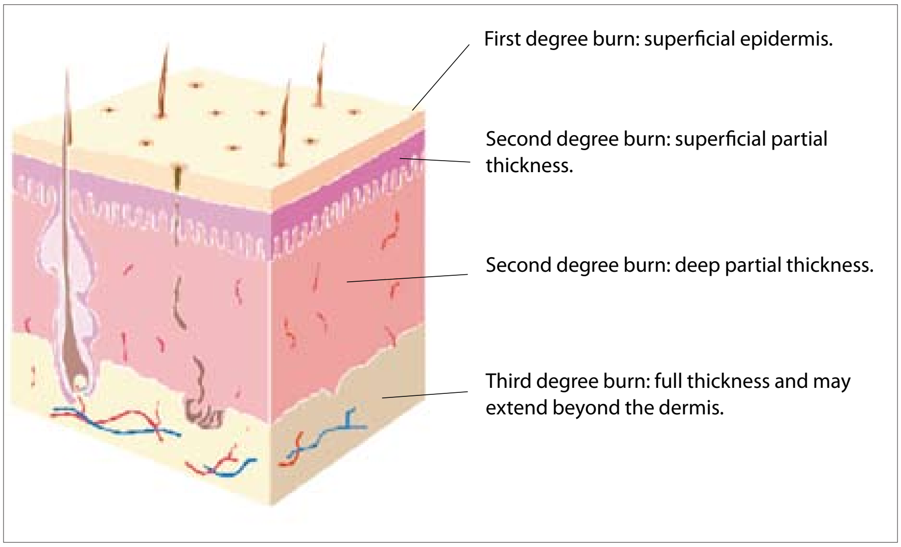
First degree burn: superficial epidermis.
Second degree burn: superficial partial thickness.
Second degree burn: deep partial thickness.
Third degree burn: full thickness and may extend beyond the dermis.
ICRC

Figure 15.1

Histology of the skin and degree of burn depth.

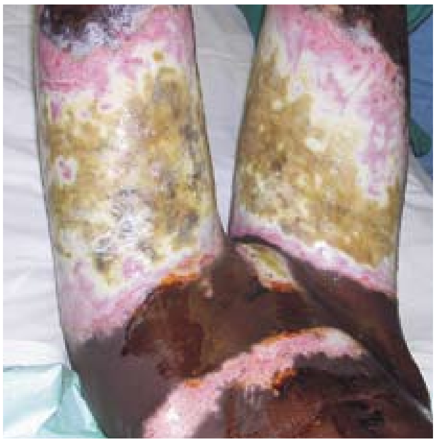

Figure 15.2

Different regions of a burn wound have different depths of damage: central eschar of full-thickness burn surrounded by areas of partial-thickness burn.

Different regions of a burn wound have different depths of damage. A burn is basically an ischaemic three dimensional wound with:

- a zone of coagulation – a central area of irreversible skin death that creates the eschar;
- a zone of stasis – a middle layer of damaged but viable tissue with a significant inflammatory reaction, but an early stasis of local blood flow;
- a zone of hyperaemia – a deep and peripheral area that looks like cellulitis, but is only hyperaemic.

Appropriate fluid resuscitation preserves cells in the zone of stasis, but subsequent infection or drying of the wound can rapidly enlarge the injury, both peripherally and in depth.

Superficial burn wounds are painful; deep burns are anaesthetic.

15.2.2 Pathophysiological changes

The most important pathophysiological change from thermal injury is increase in capillary permeability, which, if resuscitation is successful, returns to normal within 24 – 48 hours. Plasma water and proteins up to a molecular weight of 350,000 are freely exchanged between the intra- and extravascular compartments of the extracellular space. A strongly negative interstitial fluid pressure develops that creates a strong "suction" in the burnt tissue. If the affected area is very large, these changes become widespread with the loss of important amounts of fluid from the vascular compartment.

This loss is most marked in the region of the burn wound and accounts for the local oedema, which peaks at 6 – 12 hours post-burn. However, with aggressive crystalloid resuscitation in large wounds (>25 – 30% of the total body surface area or TBSA), the hypoproteinaemia causes a systemic disorder resulting in *generalized oedema* of non-injured skin and internal tissues, above all swelling of the larynx, and resultant obstruction to the airway, and abdominal compartment syndrome.

For further details on abdominal compartment syndrome, see Volume 2.

There is an associated rapid rise in the haematocrit which, together with polymerization of some plasma proteins, causes a marked increase in blood viscosity. The immediate danger linked to the sequestration of fluids in the extravascular space is hypovolaemic shock and, with haemoconcentration, acute tubular necrosis and renal failure. The loss of skin cover, and its function as a regulator of temperature, means that hypothermia, with its complication of coagulopathy, is an ever-present danger (see Chapter 18).

15.2.3 Types of burns

Flame and scald burns are by far the most common. Flame burns are usually deep and appear so at presentation, whereas scald burns may appear much less severe at first; experienced burn surgeons usually withhold prognostication until they have inspected them on the third day. Contact flame burns are typically very deep at the centre, which should be taken into consideration if they are to be operated on.

Electrical burns fall into two distinct categories. Flash burns occur when a person causes a short circuit and resulting electrical flash, but no current travels through the body. These may be treated as regular thermal injuries. High voltage (>1,000 volts) electrical conduction injuries – the current travels through the body and is characterized by the "can't let go" phenomenon – are "iceberg injuries" as they usually present with small cutaneous wounds and severe deep-tissue damage.

Chemical burns are caused by particular agents: acids, alkalis, and specific compounds (napalm, phosphorus, vesicants, etc.), with their individual characteristics.

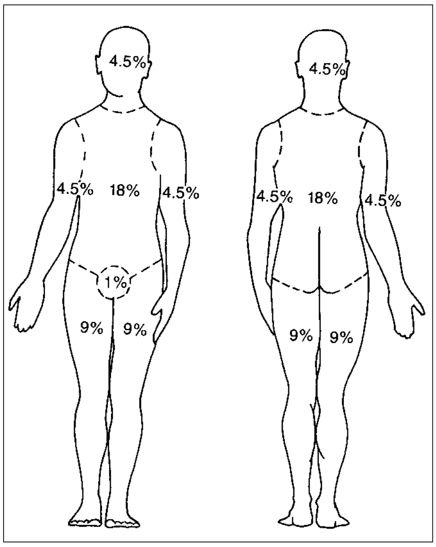

P. Zylstra / ICRC

Figure 15.3

"Homunculus" diagram showing the "Rule of Nines" to assess burnt surface area in an adult.

15.2.4 Extent of burns

The sequestration of large amounts of fluid and plasma proteins in the extravascular space is a function primarily of the extent of burnt tissues. It is thus important to estimate the total body surface area that has been burnt. The depth of the burn must also be taken into account; only second and third degree burns are included in the estimation of TBSA.

The simplest calculation is best done using the "Rule of Nines" (Figure 15.3). The size of the *patient's hand* (including palm and fingers) is about 1 % of TBSA.

The head and neck of a child under one year of age represent about 18 % and the lower extremity 14 % of TBSA (Figure 15.4). With age these figures gradually change to adult proportions.

Although it is difficult to grade the severity of burn wounds, the following provides a rule of thumb.

Minor:
- second degree less than 15 % TBSA;
- third degree less than 3 % TBSA.

Moderate:
- second degree 15 – 25 % TBSA;
- third degree less than 10 % TBSA.

Major:
- second degree more than 25 % TBSA;
- third degree more than 10 % TBSA.

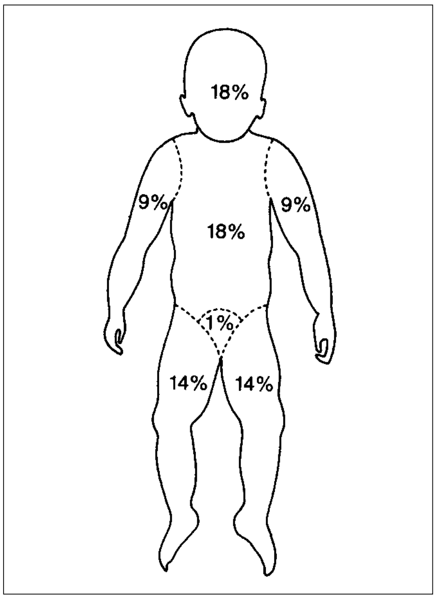

P. Zylstra / ICRC

Figure 15.4

Assessment of burnt surface area in a child "homunculus".

15.3 Burn management

The management of burns follows a series of standard measures as with any wound or trauma.

1. First aid.
2. Resuscitation:
   - airway;
   - breathing;
   - circulation / fluid replacement therapy.
3. Analgesia.
4. Prophylactic antibiotics.
5. Tetanus prophylaxis.
6. Nutrition.
7. Hypothermia.
8. Wound management.

15.3.1 First aid

Rescuers should first ensure that the scene is safe and take suitable precautions if un-ignited fuel, explosives, electricity, or chemical agents are present. The patient should be removed to safety and clean fresh air, and the vital signs assessed. If available, oxygen should be provided if there is any indication of smoke inhalation.

The burn should then be cooled with water or wet towels (for 20 minutes), and covered to reduce pain. The patient should not, however, be left wrapped in cold wet material so long as to provoke hypothermia. After cooling the *burn*, keep the *patient* warm.

If evacuation to the hospital is delayed and the airway is not at risk, the patient should be encouraged to drink plenty of small quantities of fluids often and regularly, and the colour and volume of the urine monitored.

15.3.2 Resuscitation

The following elements should be determined:

- nature of the causal agent – flame, scald, contact, electrical conduction, electrical flash, chemical;
- possible aggravating factors – additional trauma, smoke inhalation (fire in an enclosed space equals smoke inhalation);
- time since injury – fluid resuscitation is calculated from the *time of burn*, not the time the patient arrives at hospital.

As with all injured persons, examination begins with the ABCDE sequence. Deep burns to the face, neck or anterior chest cause laryngeal oedema, which increases dramatically with fluid resuscitation. Inhalation of smoke, hot gases or chemical products contributes to this swelling. However, critical laryngeal oedema can occur in *any* deep burn to this important area. The patient should be checked for burnt hair in the nostrils or soot in the nose, mouth or sputum.

The airway must be secured and kept open, preferably by tracheostomy. This should be performed *before* the airway is compromised since it will be very difficult once oedema has closed the airway and cutting through oedematous tissues can be bloody and complicated.

Please note:

Because of this oedema even the tracheostomy hole will disappear in the depths of the tissues. Rather than an ordinary tracheostomy tube, an endotracheal tube should be inserted in the tracheal opening.

True smoke inhalation induces carbon monoxide poisoning and a chemical pneumonitis related to the inhalation of hot toxic gasses. Carbon monoxide poisoning should be assumed in any person found unconscious at the scene of a fire, and is treated by administering as high a concentration of oxygen as possible for 6 hours.

Increased fluid requirements early in resuscitation suggest severe inhalation injury, which may not appear on a chest X-ray until the second or third day. Inhalation injuries increase the 24-hour fluid requirements by 1 – 2 ml per kg of body weight per percentage of TBSA (1 – 2ml/kg/%), about 50 % increase. Hypoxia and hypercapnoea despite maximal oxygen supplementation or ventilation are ominous signs. Serious cases of smoke inhalation will not usually survive without mechanical ventilation. This can be difficult to organize in a setting offering limited resources.

The sequestration of large amounts of fluid and plasma proteins in the extravascular space produces hypovolaemic shock.

Please note:

Only second and third degree burns should be included in the estimation of the percentage of TBSA. Therefore, the patient must be completely exposed and the extent and depth of the burn carefully estimated using the "Rule of Nines". The patient should be weighed and the "homunculus" diagram used to assess and mark the extent of the injury. Particular attention should be paid to circumferential burns which may require escharotomy.

The natural tendency is to *overestimate the extent* of the burn. In controlled series this has been shown to be up to 25 %. A good idea is to first calculate the burnt area and then the area which is not burnt; the total should add up to 100 %. Another natural tendency is to *underestimate the depth* of the burn; periodic re-examination will help to determine this better.

Patients with burns of moderate and major severity should get a Foley catheter placed in the bladder to monitor the hourly urine volume, which is the most important single measurement for adequacy of resuscitation. A nasogastric tube should also be passed and, if there is no acute gastric dilatation, enteral feeds can be started within the first 24 hours. Early feeding through the nasogastric tube and appropriate acid suppression (antacids, H2 blockers) prevent acute haemorrhagic gastritis, which is usually fatal. Where maintaining an i.v. line is difficult, fluid resuscitation may be administered via the nasogastric tube or even orally in smaller burns. This may be particularly useful in small children.

Adequate analgesia (intravenous narcotic) is necessary at all stages of burn management. ICRC practice is to give penicillin for the first five days to prevent invasive haemolytic streptococcal infection; any other infection is treated ad hoc. Tetanus prophylaxis must be instituted as required. Other associated injuries (penetrating wounds, fractures, etc.) must be diagnosed and treated in parallel with the burn wound.

15.3.3 Initial fluid replacement therapy

Given adequate initial resuscitation with crystalloids, the *capillary integrity* is mostly restored within 18 – 24 hours following the burn. At this point, colloid can be given and will stay within the vascular compartment, increasing the plasma volume. Cardiac output will respond to fluid replacement long before blood and plasma volumes return to normal and a first, mild diuresis begins at about 12 hours after fluid therapy. Red cell life is reduced and although replacement of red cell mass is not necessary for the first 48 hours, for major burns blood transfusion will probably be necessary after this time.

Most minor burns presenting less than 15% TBSA second degree damage do not require formal fluid resuscitation and can be treated with oral fluids, and patients can be managed on an outpatient basis. (Some surgeons prefer to hospitalize even a 3% full-thickness burn, especially to the face, hands or feet.) Moderate and major burns require hospitalization and intravenous fluids. The ICRC uses the modified Brooke/Parkland formula[1] for fluid resuscitation.

There are three phases in resuscitation therapy.

1. First 24 hours following the infliction of the burn wound (not from the start of treatment).
2. Second 24 hours.
3. After 48 hours.

The first 24 hours

Ringer's lactate is the fluid of choice. Administration should be divided into three 8-hour periods.

Ringer's lactate solution 2 – 4 ml/kg/% burn = total volume for first 24 hours:

- first eight hours: one half the volume
- second eight hours: one quarter the volume
- third eight hours: one quarter the volume

Urine output should be 0.5 ml/kg/hour.

Fluid administration should start at the low end of the formula (2 ml/kg/%; for children start at 3 ml) and the hourly urine output monitored. There is a *non-linear* relationship in burn patients between the infusion rate and the urine flow. Normal urine flow is described as 0.5 – 1.5 ml/kg/hour. The *lower* limit should be aimed for, and any increase above this should call for a lowering of the infusion rate to avoid over-resuscitation.

Apparently, there is a natural "oedema-limiting mechanism" related to the amount of fluid that is easily mobilized from plasma into the burnt areas. Giving more fluid to maintain tissue perfusion and urine flow above this level may close down this mechanism and lead to more tissue fluid sequestration (tissue is a "fluid addict"[2]); giving even more intravenous fluid does not lead to improvement.

If the urinary output is low and does not respond to increased infusion in the second eight-hour period, the third eight-hour period Ringer's lactate should be replaced with colloid, plasma or 5% albumin if available. If the vascular compartment is well-filled, however, but the patient is not making urine, the kidneys are likely to be failing and may respond to furosemide or mannitol.

---

[1] Brooke formula: 2 ml/kg/% TBSA for first 24 hours; Parkland formula: 4 ml/kg/% TBSA for first 24 hours.

[2] Sjöberg F. (Department of Anaesthesia and Intensive Care, University Hospital, Linköping, Sweden.) Monitoring of Resuscitation Adequacy. Proceedings of the International Society for Burn Injuries, 42nd World Congress of the International Society of Surgery; 2007 Aug 26 – 30; Montreal.

Other clinical vital signs must be closely watched, particularly the peripheral circulation, the patient's general condition, such as consciousness, restlessness, nausea or vomiting, and the haematocrit.

Constant monitoring is required and, after 12 hours, the patient's condition and fluid requirements should be reviewed and recalculated.

Please note:

Not only is there a natural tendency to overestimate the TBSA but, in much contemporary clinical practice, it has been discovered that over-resuscitation has become more common and more of a problem than under-resuscitation. The traditional fear of renal failure has led many clinicians to administer an excess of fluids. This has been called "fluid creep"[3] or "resuscitation morbidity" and manifests itself most commonly as pulmonary oedema and, later, abdominal compartment syndrome, delayed wound healing, increased susceptibility to infection, and multiple organ failure. It should be remembered that the fluid resuscitation formula is only a guideline, and actual amounts of i.v. fluids should be adapted accordingly.

The second 24 hours

During this phase, expansion of plasma volume can be achieved.

Plasma should be administered: 0.3 – 0.5 ml/kg/%/day.

If plasma is safe and *available*, theoretically it is best given at this stage. Albumin 5 % run at 50 ml/hour for 2 – 3 days is an expensive alternative, but the evidence to support either strategy is very slim. Otherwise, and in ICRC practice, Ringer's lactate should continue to be given at the rate of one-quarter of the first day's volume; intravenous fluid should be adjusted to urine output and enteral feeding increased as tolerated. This will provide not only essential nutrition but free water to cover evaporative losses from the burn wound.

15.3.4 Monitoring resuscitation

Clinical evaluation is particularly important in the absence of sophisticated means and laboratory measures. A clear sensorium, good tissue perfusion, good pulse, and an adequate urinary output are all signs of good progress. Less and less fluid is needed at the end of resuscitation in order to maintain urine volume. A flow sheet should be used to monitor vital signs and fluid input/output. The patient should be serially weighed if possible.

15.3.5 After 48 hours

Mobilization of burn wound oedema causes an expansion of the blood volume and results in massive diuresis, high cardiac output, tachycardia and anaemia. The better the fluid management in the first period – avoiding over-resuscitation – the less pronounced these clinical signs and the more stable the patient will be.

If available, plasma or albumin should be given to maintain serum albumin of 20 g/l and blood to maintain haemoglobin greater than 70 g/l. Deep burns cause greater anaemia. Fresh whole blood is best. Potassium, calcium, magnesium and phosphate are generally excreted in large amounts at this time, and should be supplemented where possible.

---

[3] Pruit BA Jr. Fluid and electrolyte replacement in the burned patient. *Surg Clin N Am* 1978; **48**: 1291 – 1312.

15.4 Burns presenting late

Often, patients with severe burns present late. Those presenting after a delay, but within the first 24 hours, should receive fluid resuscitation and an attempt should be made to infuse most of the calculated volume by the end of 24 hours from the time of burning.

Patients presenting after 24 hours will probably need some fluid but the amount should be guided primarily by clinical assessment of hydration and renal function. Patients who have survived the first 72 hours without renal failure have compensated the losses themselves (usually with oral fluids); they may require some rehydration but infection of the burn wound is the main concern. For those presenting weeks after the burn, infection is compounded by poor nutritional status, anaemia, and hypoproteinaemia. In these cases steps should be taken to control infection (debride grossly infected necrotic tissue) and to improve nutrition before any definitive surgery is undertaken. Early creation of a feeding gastrostomy should be considered.

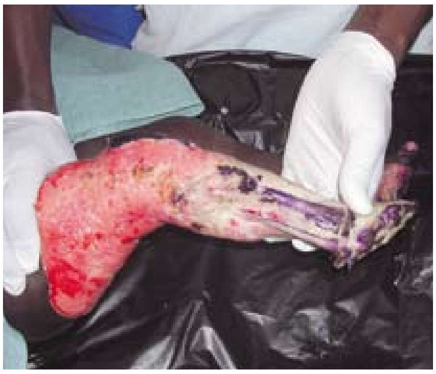

E. Dykes / ICRC

Figure 15.5

Old flame burn to the leg.

15.5 Nutrition

Catabolism is particularly elevated in burn patients, especially those losing large amounts of protein through open wounds, and healing will require an important increase in caloric and protein intake for a lengthy period. Early enteral feeding is very important in maintaining gut function (reducing gastroparesis), and preventing complications. Major burn patients may require more than twice their normal protein and calorie intake until their wounds have closed. Nutritional requirements may be easily calculated for the individual patient (see Annex 15. A: Nutrition in major burns).

Blenderized enteral feeding solutions can readily be made from commonly available local foods and can be delivered by naso-gastric tube, gastrostomy or jejeunostomy. Especially when dealing with burn wounds that are weeks to months old, the patient's nutritional status should be evaluated and improved before any attempt is made at skin grafting or both the graft and the donor site risk delayed or failed healing.

15.6 Care of the burn wound

After successful resuscitation of the patient has been achieved, the next greatest threats to life that must be addressed are the burn wound itself and the complications of sepsis.

The aim of treatment is to achieve healing of the burn by the following measures.

1. Control of bacterial colonization by removal of all dead tissue.
2. Prevention of accumulation of purulent fluid and debris.
3. Prevention of secondary bacterial contamination.
4. Maintenance of an environment which promotes wound healing.
5. Avoidance of techniques or treatment which will damage the healing burn wound.

Most of the morbidity and mortality associated with major burns are the result of infection. All the methods for cleaning the wound, excising dead tissue and treating the burn are directed towards the control of "burn wound sepsis".

Untreated, eschar desiccates and sloughs off by means of bacterial enzymatic invasion of the plane between viable and non-viable tissue. Full-thickness burns will require skin grafting to heal properly; otherwise, healing will be by fibrous tissue contraction alone with chronic open wounds and disabling scar contracture.

In partial-thickness burns, viable areas of dermis underlie the necrotic tissue and if sufficient epidermal cells remain at the base of skin appendages, such as sweat glands and hair follicles, re-epithelialization will gradually take place if given the right conditions.

Infection will transform a partial-thickness burn into a complete-thickness one. Because of the complete or partial ischaemia associated with burn wounds, systemic antibiotics may not reach the site of bacterial colonization. Local treatment, both mechanical and antimicrobial, are the basic elements of wound care.

15.6.1 Initial wound management

Initial wound toilet should be undertaken after resuscitation has begun and proceed in parallel. Once the patient's condition has stabilized, attention can be turned to more *definitive* measures.

Any constricting agents (rings, wrist-watch, jewellery, etc.) should have been removed on admission. The patient should be sedated and the burn wound gently washed with soap and water. Freely running, clean water, without excess pressure but a regular flow and at a temperature comfortable for the patient, is best. This will cool the burn, decrease pain, and remove superficial debris and adherent clothing.

Small intact blisters may be left *undisturbed* but large, bloody, or pus-filled blisters, and those which interfere with joint movement, should be *unroofed and debrided*. Large burn wounds may be more easily cleaned by placing the patient in a shower. Baths should be avoided because of the attending logistical difficulties and the danger of cross-infection under field conditions. The routine daily immersion of burn patients in filthy tubs of cold water is to be condemned.

Particular attention should be paid to circumferential deep burns. During the first 48 hours the increasing oedema of the tissues and the thick unyielding burn eschar may result in a tourniquet effect. Circumferential burns of the chest will restrict breathing and those of the limbs will cause peripheral ischaemia which can lead to amputation. This disaster is easily averted.

Escharotomy is the process of incising the eschar down to the subcutaneous fat to allow relief of the constriction.

Escharotomy should be performed with a sharp knife or electric diathermy through the burnt skin into the subcutaneous fat.

Escharotomy incisions are placed in the mid-lateral and mid-medial lines of the affected extremity and should extend up to, but never include, unburnt skin. A "T-cut" at each end of the escharotomy allows tissue to expand without causing an acute constriction at the end of the incision. On the hands, the mid-lateral incisions coming down each side of the forearm should be swung onto the dorsum, and branched to make a single escharotomy incision down the mid-lateral line of each finger on its less-used side. (i.e. ulnar side of thumb, index and ring finger; radial side of middle and little finger).

A thoracic escharotomy begins at the midclavicular lines, continues along the anterior axillary folds down to the costal margin, and across the epigastrium to the xyphoid process.

Please note:

Few patients who truly need a thoracic escharotomy will survive without mechanical ventilation.

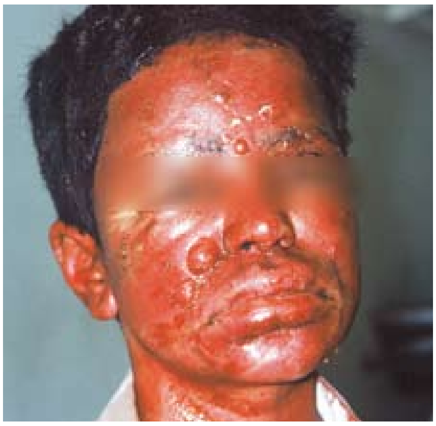
ICRC

Figure 15.6

Burn to the face with intact blisters.

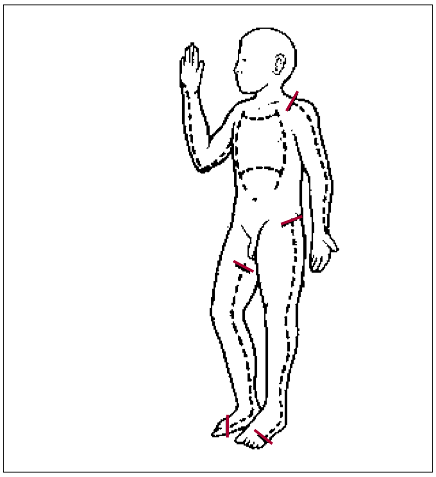
M. King / Primary Surgery Volume 2

Figure 15.7.1

Sites for escharotomy incisions.

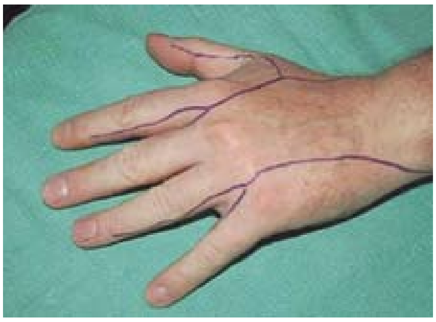
M. Beveridge / ICRC

Figure 15.7.2

Placing for escharotomy incisions on the hand.

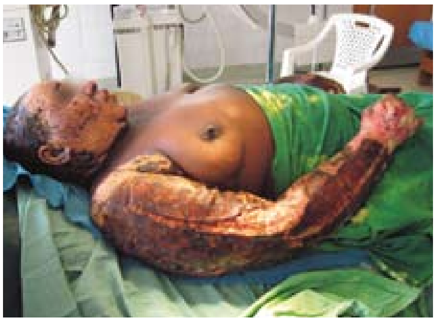
M. Baldan / ICRC

Figure 15.7.3

Escharotomy incision of the arm.

Though full-thickness burns are usually anaesthetic, the escharotomy should be done with some form of anaesthesia as the edges can be very painful and the cut goes into the subcutaneous fat. Ketamine is ideal.

Particular attention should be paid to burns with underlying fractures or burns extending deep into the fascia; these can create a compartment syndrome. Formal fasciotomy, with excision of the aponeurosis, may be necessary in addition to escharotomy.

Severe perineal burns may necessitate faecal diversion.

15.6.2 Local care

Care of burn wounds involves an extraordinary amount of nursing care. The method used depends on the depth, extent, and location of the burn. ICRC surgical teams use occlusive dressings – and its plastic bag modification – and open therapy; both with a topical antibacterial agent.

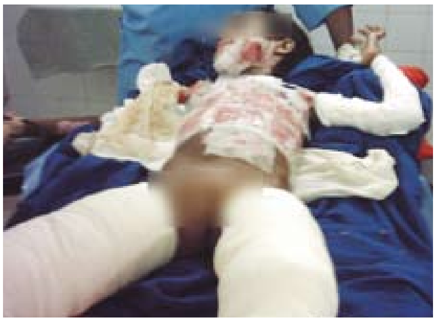
K. Yagi / U. of Khartoum

Figure 15.8

Paraffin gauze and occlusive dressings.

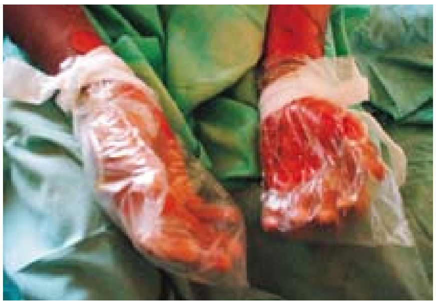
K. Yagi / U. of Khartoum

Figure 15.9

Plastic bag modification of occlusive dressing.

Many antibacterial compounds can be used. Silver sulphadiazine (Flamazine ®) and silver nitrate solution are unique in that they can penetrate the wound surface and reach the bacteria beneath it; therefore, they are best for infected full-thickness burns. Antibiotic ointment (polymyxin/bacitracin or similar) and paraffin gauze is another excellent dressing, particularly for partial-thickness burns with no eschar. Honey and ghee dressings can be used where expensive compounds such as silver sulfadiazine are scarce. Equal parts of honey and either ghee (clarified butter), or oil, are mixed together and poured over open sheets of gauze in a pan. The honey provides a hyperosmotic environment that discourages bacterial growth and the ghee or oil keeps the gauze from adhering. Silver nitrate solution (0.5 %) is effective but oxidizes easily, staining everything it contacts black. Other agents used around the world include gentian violet, which desiccates the burn, tea which tans the burn like leather, papaya and banana leaf, boiled potato skins, fermented fish sauce, and Amazonian frog skin. In some instances local methods may be preferable to imported products.

Occlusive dressings

Bulky sterile dressings relieve pain and are comfortable for the patient, and protect the burn wound from infection. They absorb serum and exudate and foster a moist healing environment by keeping the injured part immobile and warm; they contain antibiotics able to penetrate the dead eschar (i.e. silver sulfadiazine ointment).

The dressings have three components: an inner layer of a liberal application of silver sulphadiazine covered by fine-mesh gauze or paraffin gauze; a middle layer of large gauze-wrapped absorbent cotton pads, which absorb the exudate and protect the wound; and an outer layer of bandages that secures the dressing.

If the dressing becomes soaked through, the outer layers should be changed or bacterial contamination will occur by capillarity. Dressings should be changed every day or two under adequate analgesia, and the old silver sulphadiazine washed off in the shower. The wounds should be inspected and gently cleaned at dressing time, and morsels of dead eschar removed with scissors and forceps.

Occlusive dressings are best for small burn areas, especially on the limbs, or when hygienic conditions are less than optimal.

The plastic bag or surgical glove method

This method is used for burns to the hands and feet. After cleaning the burn and applying silver sulphadiazine directly to the wound with a spatula or the gloved hand, a plastic bag is used as a glove or a sock, tied around the wrist or ankle. It must not be too tight. The limb should be kept elevated to reduce swelling. The burnt area will be kept moist and movement of the joints, both passive and active, is encouraged. A surgical glove may also be used instead of a plastic bag; it allows for more mobility during physiotherapy.

Open therapy

Where resources are inadequate to allow for occlusive dressings, this is the method of choice – as well as being the standard method for burns of the face and perineum – but requires clean and isolated surroundings, and the ambient temperature must be warm; hypothermia is to be avoided at all costs.

The patient is placed on clean sheets and the burnt area left completely exposed. Silver sulphadiazine ointment is applied to the burn wound liberally with a sterile gloved hand and repeated twice daily or as required. If the room is cold, the patient may be covered with a clean sheet and blanket draped over a frame to prevent contact with the wound. The entire bed should be kept under overhanging mosquito netting.

The *advantages* of this method of treatment include simplicity of examination of the burn wound and ease of nursing care. It also allows early mobilization by physiotherapy.

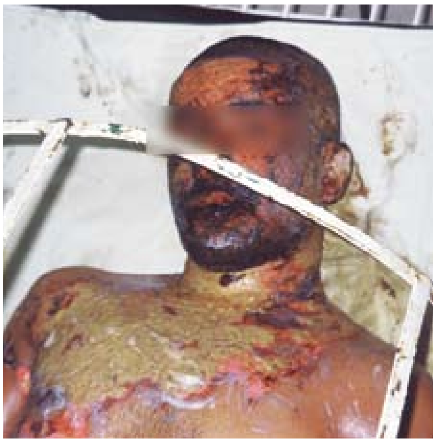
ICRC

Figure 15.10

Open burn therapy with frame.

The *disadvantages* are pain, odour, and desiccation of the wound, delayed eschar separation and hypothermia. Frequent showering is required to wash away exudate and fragments of softened eschar. The bed linen must be changed regularly; it is readily dirtied by the burn exudate. Local customs and religious practice may limit the use of this "open" method.

Facial burns are best treated open with frequent gentle cleansing and application of moist warm saline gauze soaks, interspersed with application of topical antibiotic ointment (e.g. polymyxin/bacitracin). Beards and facial hair growing through the burn should be shaved at least every two days to prevent accumulation of exudate which may harbour infection. Antibiotic eye ointment should be applied frequently to the conjunctiva if the lids are burnt and retracted, to prevent keratitis and corneal ulceration. The lids should only very rarely be sutured together (blepharoplasty) as the sutures almost inevitably pull out causing yet more damage to both lid and globe.

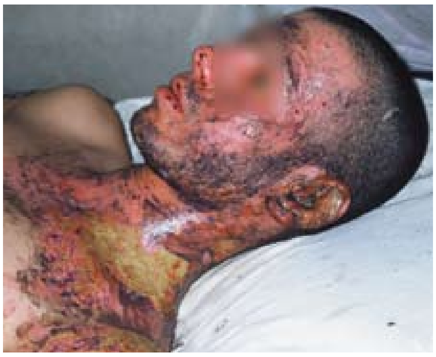
ICRC

Figure 15.11

Healing burn to the face.

15.7 Closure of the burn wound

The preparation of the burn wound and its subsequent closure are the two main steps in the surgical management of burns. The type of surgery required depends upon the skill and training of the surgeon, the specific burn injury, and the facilities available to support the treatment, especially the availability of blood for transfusion. Like all other injuries treated under difficult circumstances, great judgement is required to select the technical procedure that is appropriate.

Partial-thickness burns will generate new epithelium on their own if infection is prevented. The wounds should be inspected carefully as they progress. In superficial partial-thickness burns (and donor skin graft sites) epithelial cells grow out from around tiny skin appendages giving a typical "leopard spot" appearance in those with pigmented skin and, on close inspection, a slightly dull, silvery coat of epithelial cells may be seen growing atop the dermis (Figure 15.12.1). Tiny white pearls of epidermis herald re-growth and healing; areas of raspberry-red granulating dermis or fat do not have sufficient epidermal cells to heal (Figure 15.12.2). A "nicely granulating" burn wound is not a good thing unless there is a plan to graft the wound.

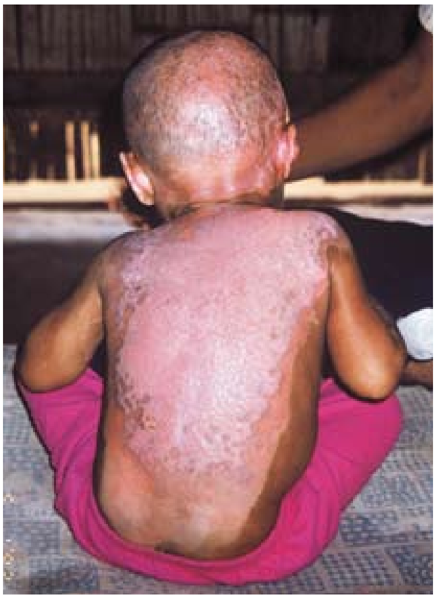

ICRC

Figure 15.12.1

Partial-thickness burn well on the way to healing.

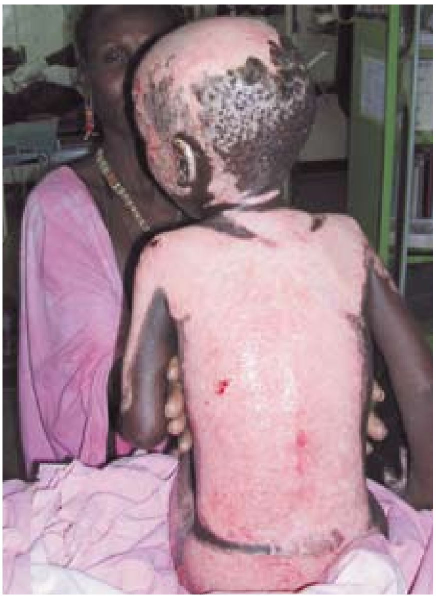

E. Dykes / ICRC

Figure 15.12.2

Granulating burn: no healing. The pale colour of the granulations indicates that the patient is anaemic.

For full-thickness burns, the eschar may be removed completely in a single procedure or in a number of stages. The aim of treatment is to prepare the wound for eventual closure and to prevent colonization by bacteria and fungi.

15.7.1 Mechanical cleansing and debridement

Thorough cleaning and removal of debris and fragments of eschar are performed at each dressing session. Gentle washing and sharp debridement of dead skin fragments should be combined with thorough irrigation with water. The surface can then be cleansed with a mild disinfectant (dilute hypochlorite solution, detergent soap) and thoroughly washed again with water. Silver sulphadiazine is reapplied. Where resources are scarce, the cycle of dressing and plucking can be continued until the wound is completely free of eschar. This gives partial thickness burns the optimum chance to re-epithelialize and minimizes the area that will require grafting. Such patients need large amounts of nursing time, dressing material, and will probably require blood transfusion even without surgery.

15.7.2 Surgery

The most difficult decisions in burn care involve the timing and extent of surgery. Judicious staging of burn surgery is highly important and, where possible, should take into account the patient's livelihood. Donor and graft areas should be matched in advance and the process broken down into manageable parts. Attention should be paid to how the body parts will be positioned for the operation: if the plan is to graft the arm, it should be seen to before the hand as the latter will be needed to hold up the arm for surgery.

The hands, feet, and joint surfaces are considered priority areas in order to re-establish function; early grafting of these parts must be balanced against the metabolic advantages of closing larger areas on the limbs and torso. At least two weeks should pass before grafting of facial burns is considered, as even quite deep burns may go on to heal on their own. Eyelids are the first priority here.

Tangential excision

The removal of the full thickness of the burn wound is performed in a single session. Early tangential excision and immediate grafting of burns reduces mortality, morbidity, suffering and hospital stay while improving functional and cosmetic results, but it requires considerable resources and is *impracticable* for areas larger than 10 % TBSA outside specialized burns centres. It is bloody surgery.

Most field surgeons should be *conservative* with regard to this technique. It is, however, recommended in ICRC practice for small burnt areas, especially on the face, hands and feet, and over joint surfaces.

In tangential excision of the eschar, the superficial layers of the burnt tissue are progressively shaved away by knife, dermatome or cutting diathermy until viable tissue is reached. Traditionally, this is determined by the appearance of copious punctuate bleeding. It is accompanied by significant blood loss and is the most important limiting factor of this technique.

Blood loss can be reduced by Esmarch bandage exsanguination of the limb and a tourniquet, and subcutaneous infiltration of a dilute adrenaline solution (1:500,000); in the face, lidocaine with adrenaline can be used. The subcutaneous infiltration of some type of fluid (saline, dilute adrenaline solution, or local anaesthetic), causes a local tumescence that facilitates the excision. With adrenaline solution, as dead burn eschar is sliced away, the surgeon should look for a level with visible pearly white dermis, or glistening yellow fat, and no capillary thrombosis. After excision, larger bleeders should be cauterized and the wound wrapped for ten minutes with gauze soaked in adrenaline solution. The wrapping should be removed, and the process repeated until there is no active bleeding, before the skin graft is applied.

When tangential excision is successful, only dead tissue is removed. It can be difficult to judge how much to take away in order to leave a viable layer that will take an immediate skin graft.

Face

The skin of the face, particularly the beard areas on males, is very thick and well populated with deep epidermal cells that will provide for re-epithelialization if given time. If there is any doubt about the depth of burn on the face it is best to wait two weeks before tangential excision.

As mentioned, face burns are treated by the open technique with warm wet gauze soaks, gentle cleaning and shaving every second day, and the application of a local antibiotic ointment. Severe face burns will require scraping and cleaning under general anaesthesia to properly evaluate which areas are healing and which will eventually need grafting. Gauze soaked in an adrenaline-saline solution (1:33,000) should be used and pressure applied to control bleeding. When clean, a thin coat of antibiotic ointment should be applied and the previous dressing routine resumed while waiting to decide whether to excise and graft.

Excision of a small deep burn to the face can be performed under local lidocaine with adrenaline anaesthesia; larger areas require general anaesthesia, but simultaneous subcutaneous infiltration with a dilute adrenaline solution will render the excision easier and less bloody.

Hands, feet, and joint surfaces

Tangential excision of the hands, feet and joint surfaces can be performed at three days onwards, once the patient is well resuscitated.

Many severe hand burns will benefit from early escharotomy since deep partial-thickness or full-thickness burns heal with extremely disabling contractures; operating these early should be considered and good, thick skin for the grafts saved. Typically people clench their fists when they sustain a burn so the palmar skin extending up to the mid-lateral lines of the fingers is usually preserved, or burnt much less deeply than the dorsum, and rarely needs grafting. Therefore most burnt hands and fingers will only require grafting of the dorsum. If the escharotomy was performed accurately along the edges of the full-thickness burns along the mid-lateral lines of the fingers, this will mark the extent of excision necessary.

Proper preparation is the key to good excision. The areas to be excised should be carefully marked with ink or gentian violet. The hand and forearm should be exsanguinated by five-minute elevation and application of a rubber Esmarch bandage, beginning with the hand and progressing proximally; a pneumatic tourniquet is then applied. (Correct use of a tourniquet is painful; the operation should be performed under general anaesthesia.) Normal saline or a weak adrenaline solution should be infiltrated subcutaneously over the dorsum of the hand. The edges of the areas to be excised should be scored with a #15 blade.

Tangential excision should be performed using a small dermatome or scalpel, preserving viable dermis where possible and being very careful not to damage tendon sheaths. The hand should be wrapped in adrenaline-soaked gauze and the tourniquet briefly released. The tourniquet should then be re-inflated for ten minutes to allow natural haemostasis, then removed. The hand should be unwrapped and diathermy used to cauterize remaining bleeders. Wrapping in adrenaline gauze and cauterizing of bleeders may need to be repeated several times to ensure perfect haemostasis prior to application of the skin grafts. Sheets of relatively thick split-thickness skin should be used. Grafts should be carefully tailored over the dorsum of the hand and fingers, and sutured in place. Each finger should be covered in paraffin gauze dressing, then wrapped independently in gauze, taking care to leave the tips of the fingers exposed in order to assess perfusion.

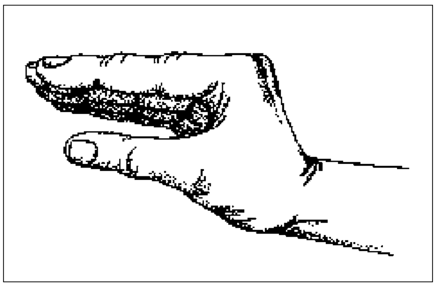
M. Beveridge

Figure 15.13

The "safe position" for splinting a hand.

Finally, the hand should be splinted in the "safe position" (Figure 15.13) with the wrist dorsiflexed 30°, the metacarpal-phalangeal joint flexed as close to 90° as possible, the fingers splayed apart and the inter-phalangeal joints straight. The initial dressing should be left for 5 – 7 days and then unwrapped carefully. The burn should then be re-dressed daily with paraffin gauze and re-splinted. Physiotherapy and mobilization should begin as soon as the grafts are solid. A hand should never be immobilized longer than ten days. K-wires placed through the fingers may help in bad cases. The same general principles apply to the feet and over joint surfaces.

Early tangential excision gives the speediest recovery and best functional and cosmetic results when it is practicable, but should *never* be undertaken for large areas unless the resulting blood loss can be properly managed.

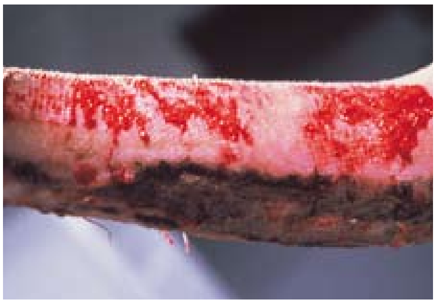
E. Winiger / ICRC

Figure 15.14.1

Meshed skin graft: recipient site.

Skin grafting

The hands, feet, and joint surfaces are priority areas for skin grafting, with or without tangential excision. The anterior chest and neck take priority over the belly and buttocks. The back has very thick skin, and so burns to it may be observed for some time while waiting to see if they will heal on their own.

Allowing burns to granulate and demarcate under dressings (2 – 6 weeks) is prudent practice where resources are scarce. This means accepting inevitable protein loss from open wounds, possible infection, delayed healing, and chronic anaemia; supplementary measures must therefore be taken to reduce these effects.

To prepare for grafting, the jelly-like granulation tissue must be scraped away with the back of a scalpel handle before the skin graft is carefully secured and dressed. The advantage of delayed grafting is that often a much smaller area need eventually be grafted.

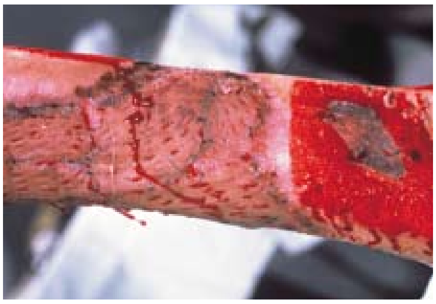
E. Winiger / ICRC

Figure 15.14.2

SSG after grafting.

Grafting on burns is time-consuming and adequate time should be allocated for these operations. The staging of the surgery should be carefully planned and just one limb or body area operated at a time. In general, the thinner the graft, the better its chances of "take", and the thicker the graft, the better the functional and cosmetic result (see Chapter 11 for details on skin grafting). A limb or digit that has to be amputated should be regarded as a prime source of donor skin. Harvesting of grafts in children whose skin may be extremely thin should be performed with *great care*. If there is no solid layer of dermis left behind, the donor site will not heal. Except on the face, hands and feet, skin grafts should be meshed to allow serum to drain through rather than lift the graft off the bed of the recipient site.

Old wounds or sites where previous grafts have failed are notorious for difficulty in achieving an adequate graft take. Good nutrition and meticulous preparation of the surface are the keys to success. Early excision of grossly dead and infected tissue, topical and systemic antimicrobial therapy, and aggressive nutritional supplementation should precede any attempt at skin grafting. The best course may be to graft critical areas while leaving some of the larger, less functional areas to granulate.

The surface of the debrided burn wound often has a superficial slime of exudate and bacterial contamination. Dressings with a supersaturated saline solution (add salt to normal saline until it no longer dissolves), changed frequently for a couple of days, will produce a clean, bright red granulating surface ready for grafting.

The occlusive dressing which is applied after skin grafting plays a great part in the survival of the graft. It must be put on with great care. The dressing must hold the graft closely applied to the recipient site for the first few days for capillaries to grow in. Any blood or serum that floats the graft away from its bed will cause it to fail.

The life of a patient with a severe burn is in danger until the dead tissue has been excised and the defect closed with a healthy skin graft.

15.8 Scar management

Burns always result in scars and the burn graft itself is best thought of as a scar. One of the most disastrous consequences of burns is the severe scar contracture that may render life horrendous for the victim afterwards. Management of the burn scar begins before grafting, during local care of the burn wound. Rigorous splinting – using plaster of Paris slabs – and stretching routines should be implemented to prevent contracture of major joints: the limb should be splinted *against* the force of contracture. If a joint is equally burnt on both sides, it should be splinted in extension. Airplane-strut type splints should be used for burns to the axilla. Patients must be given adequate analgesia for the daily passive stretching exercises.

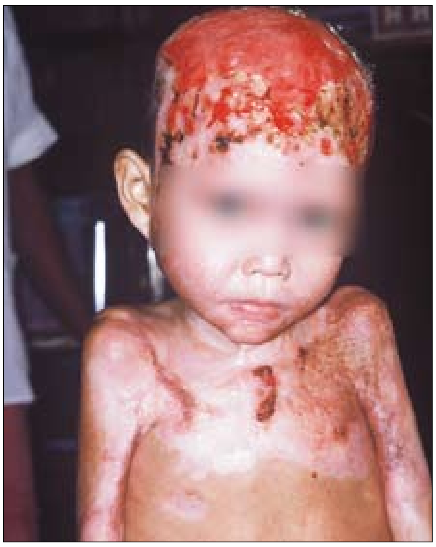
ICRC

Figures 15.15.1 and 15.15.2

Post-burn scar contractures.

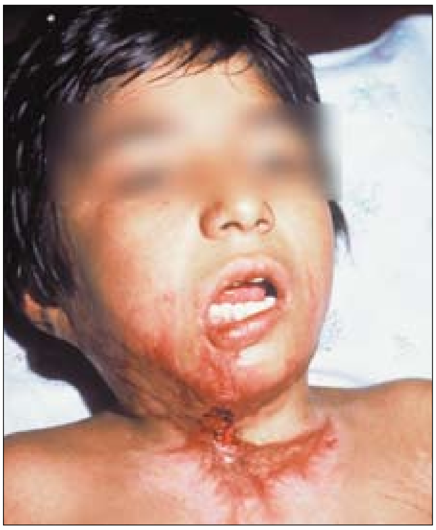
ICRC

Much, if not most of the functional benefit of burn grafting depends on assiduous splinting and stretching of the tissues afterwards to manage the process of burn scar contraction. This can be particularly active in children – a perfect operation may result in recurrence of a dreadful scar if the tissues are not managed properly over the next 6 – 12 months.

All burn grafts crossing joints should be splinted at operation with plaster slabs. Later, when the grafts have taken, a plaster slab covered in tube bandage and fitted correctly makes an excellent, reusable splint that can be worn at night and removed for therapy during the day. Dedicated staff and adequate analgesia are essential to active and passive stretching of burn scars; if this hurts too much the patient will simply not comply. Burns crossing joints should be splinted and stretched even if they have not been grafted, as this will reduce the degree of contraction as the scar heals.

Pressure garments are important to optimum scar management and their use results in much softer and more pliable scars. Where they are not available, elastic bandages and a variety of tight-fitting stretchy commercial clothing may help. Water-based skin cream will help with itching as will antihistamines. Burn scar massage is also helpful where traditional practitioners are present.

15.9 Electrical burns

As mentioned, there are two distinct categories. With flash burns, the patient typically presents with fairly deep burns to the face and one or both hands and forearms. Treatment is as with regular thermal injuries.

High voltage (>1,000 volts) electrical conduction injuries have small cutaneous entry and exit wounds, which extend deep into the muscles causing myonecrosis. The rhabdomyolysis has a *systemic* effect, with myoglobinaemia and myoglobinuria leading to acute tubular necrosis; and a *local* effect, i.e. compartment syndrome.

The patient should receive Ringers' lactate, with 50 mEq sodium bicarbonate per litre, in sufficient volume to maintain urine output at 0.5 ml/kg. If the urine is dark or bloody, or urine output ceases, the vascular space should be well filled and a bolus of 20 % mannitol given (1g/kg); furosemide may be added as well.

Any suspicious compartments should be released promptly by generous full-length fasciotomy, including carpal tunnel release in the forearm. Dead muscle should be debrided conservatively, and numerous returns to the operating theatre may be needed (serial debridement).

Numerous complications attend electrical conduction injury including arrhythmias (ECG monitoring is strongly recommended), cervical fractures associated with the jolt of electricity, intestinal perforation and a variety of unusual neurological sequelae.

15.10 Chemical burns

A number of chemical substances cause specific burns. The presence of any of them on an injured person poses a danger to first aiders, hospital staff, and other patients. Careful removal of contaminated clothing and proper decontamination measures must be undertaken – of the patient and of any equipment used – and specific protocols followed to protect the medical personnel treating the patient.

> The wounding chemical agent constitutes a danger to first aiders, hospital staff, and other patients. Proper protective measures must be taken.

15.10.1 Acid and alkali burns

In general, acids cause coagulative necrosis of the skin, and strong alkalis liquefactive necrosis and burrowing of the chemical deep into the tissues. "Acid violence" involving the throwing of concentrated sulphuric acid into the face is an increasingly common phenomenon and the chemical is readily available wherever there is a motor vehicle garage. The acid burn should be washed with very large volumes of water and the eyes thoroughly irrigated. Strong alkalis may come in dry form (NaOH crystals); any remaining material must be brushed off the patient before washing with water.

After this decontamination, the treatment of chemical burns follows the same sequence as the treatment of thermal burns. Acid attacks usually involve the face and typically cause extremely disfiguring injuries which are very difficult to reconstruct.

15.10.2 Phosphorus burns

Some modern anti-personnel weapons contain white phosphorus. This element *ignites on contact with air*, and fragments of phosphorus will be scattered throughout any wounds; it is lipid soluble and sticks to the subcutaneous fat. The burn is deep and painful and the phosphorus continues to burn as long as it is in contact with oxygen or until all the phosphorus is consumed, all the way down to the bone. Local treatment is more urgent than with conventional burns because of the aggressive nature of phosphorus. Much of the injury in an individual patient, however, results from the ignition of clothing, which causes a conventional burn.

Contaminated clothing must be removed immediately, care being taken not to contaminate the staff attending to the casualty. Visible, smoking particles can be removed with a spatula or knife, and should be placed in a basin of water to exclude them from the air. Phosphorus burn wounds must then be isolated from oxygen by being kept wet through liberal soaking with water, by covering with wet dressings, or by placing the injured part in a basin of water. On no account must they be allowed to dry out.

When surgical treatment is available, the idea is to identify and remove the remaining phosphorus particles. The wet wound can be irrigated with a neutralizing agent. A freshly prepared solution of 1% copper sulphate combines with the phosphorus to form black copper sulphide, which impedes violent oxidation and identifies the particles. The black particles can then be removed with forceps and placed in a basin of water. The solution must be very dilute, the palest blue colour, since its absorption can cause haemolysis and acute renal failure. If used, the copper sulphate solution must be washed away immediately. Or, if copper sulphate solution is not available, the operating theatre lights may be put out; any remaining particles will glow with phosphorescence in the dark and can be carefully picked out with forceps and placed in a basin of water.

Care must be taken not to allow the wound and the phosphorus to dry out and re-ignite in theatre; appropriate, non-flammable anaesthetic agents should be used. In all cases, the wound should then be excised and dressed as usual.

Phosphorus may provoke hypocalcaemia and hyperphosphataemia; intravenous calcium should be given. Absorbed phosphorus can be toxic to many organs:

- central nervous system – delirium, psychosis, convulsions, coma;
- gastro-intestinal tract – abdominal colics, melaena;
- liver – hepatomegaly, jaundice;
- kidneys – proteinuria, acute tubular necrosis;
- blood – thrombocytopenia, hypoprothrombinaemia;
- myocardium – ventricular arrhythmia, myocarditis.

15.10.3 Napalm injuries

Napalm is jellied petrol, an intensely flammable agent that clings to the clothing and skin while still burning and causes serious, deep and extensive burns. Its incomplete combustion of the oxygen in the air around the victim provokes an acute rise in carbon monoxide that can lead to a loss of consciousness and even death. The intense heat and benzene fumes easily cause inhalation burns.

Napalm burns are invariably *full thickness*, with coagulation of muscles and other deep tissues. Nephrotoxicity is a serious complication of the rhabdomyolysis, and the mortality may be high in proportion to total body surface area involved. A full-thickness burn of only 10% of the body surface area may result in renal failure. The patient should be kept well hydrated and in alkalosis; mannitol may be necessary to protect renal function.

First-aid treatment includes extinguishing the burning napalm by smothering it, i.e. excluding the oxygen. Unlike phosphorus, napalm does not then reignite in contact with the air. Surgical treatment involves removal of the napalm with a stick, spatula or knife. The wound is then excised deep to any remaining contaminant, care being taken to avoid contact ("no-touch" technique), and dressed as usual.

15.10.4 Magnesium

Flares released from aircraft to avoid heat-seeking missiles use magnesium, which gives off great heat when lighted. These may reach the ground and cause fires and injuries. The intense heat causes a full-thickness burn. The wound should be excised deep to the contaminant using the "no-touch" technique. Some reports from ICRC surgical teams (Kabul, Afghanistan) have made mention of toxic side effects due to absorption of magnesium, similar to phosphorus; this has not been confirmed.

15.10.5 Vesicant chemical weapons

Unlike other agents causing specific burns, chemical weapons have been banned by international treaties.[4] However, some States retain stockpiles, which may either be used militarily or released into the air if the storage facilities are bombed. Certain chemicals have a potential dual function: they can be used in weapons and are widely employed for civilian purposes (the disinfection of public water supplies in the case of chlorine).

Traditional chemical weapons are either neurotoxic or vesicant (blistering); the latter cause burns to the skin and inhalation injury. Vesicant agents (mustard gas, lewisite, phosgene) cause skin burns similar to flame burns.

The patient's clothing must be removed and properly destroyed. The patient is then decontaminated with abundant soap and water. Care must be taken not to contaminate hospital personnel, equipment, and other patients with the chemical agent. Correct decontamination protocols include the use of protective clothing and equipment (mask, gloves, boots, etc.) by first aiders or hospital staff.

Once decontamination has been performed, the wounds are dealt with in the traditional manner; however, a "no-touch" technique should be used during wound debridement and removed tissues disposed of with care. The patient's respiratory function must be closely monitored. Inhalation of chemical fumes will burn the respiratory mucosa with the development of acute respiratory distress syndrome (ARDS). Assisted ventilation may be required.

---

[4] 1925 Geneva Protocol for the prohibition of the use in war of asphyxiating, poisonous or other gases and of bacteriological methods of warfare and the 1993 Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction.

ANNEX 15. A Nutrition in major burns: calculating nutritional requirements

Calorie requirements = Basal energy expenditure x stress factor x activity factor

The *basal energy expenditure* is calculated as follows:

$$[66 + (14 \times \text{weight in kg}) + (5 \times \text{height in cm}) - (6.8 \times \text{age in years})]$$

The *stress factor* for minor procedures is 1.3; for skeletal trauma, 1.35; for major sepsis 1.6; and for major burns 2.1.

The *activity factor* is 1.2 for those in bed and 1.3 for those who are mobilizing. Women require about 4 % less than men for equal body size and age.

Example

For a 25-year-old male weighing 60 kg and 170 cm tall, in bed with a major burn, the calorie requirements =

$$[66 + (14 \times 60) + (5 \times 170) - (6.8 \times 25)] \times 2.1 \times 1.2 = 3{,}997 \text{ kcal/day}$$

Protein, glucose and fat requirements

Daily *protein* requirement for acute burns is 2 g/kg in adults and 3 g/kg in children. Protein provides about 4 kcal/g (120 g and 480 kcal in the example given above).

Daily *glucose* requirement is about 6 g/kg/day in burns.
Glucose provides 4 kcal/g (360 g and 1,440 kcal in the example).

The difference between the calculated energy requirement (3,997 kcal) and that provided by protein and glucose should be made up with fat.

Daily *fat* requirement = 3,997 kcal – 480 – 1,440 = 2,077 Kcal
Each gram of fat provides 9 kcal, therefore, 2,077 ÷ 9 = 231 g of fat

The larger the volume and the higher the fat concentration the more likely the patient is to develop diarrhoea. In an adult with a major burn, 3 litres of feeds per day is a reasonable target; therefore for the patient in the example, a "cocktail" containing 40 g of protein, 120 g of glucose and 80 g of fat per litre should be prepared.

Making a high-energy enteral feeding solution for burn patients

Ingredients	Glucose	Protein	Fat	kcal
Skimmed milk powder 110 g (244 ml)	44 g	40 g		385
Edible oil 80 g (80 ml)			80 g	720
Sugar 50 g (50 ml)	50 g			200
1 Banana (15 mEq potassium)	25 g			110
Add: Salt 3 g Calcium containing antacid 3 tablets Multivitamin tablet 1 daily Ferrous sulphate + folic acid tablets Codeine 30 – 60 mg per litre provides analgesia and reduces diarrhoea Eggs contain 15 g of protein each: beware of salmonellosis from raw eggs Supplement tube feeds with cooked eggs fed by mouth when possible				
Boiled and filtered water to make 1,000 ml of solution			Total 1,415 kcal per litre	

Make a paste of milk powder with a little water; add sugar, salt, crushed tablets and oil. Slowly add more water while mixing well; add mashed banana and mix thoroughly (using a blender if possible). Filter the mix through a gauze compress and refrigerate. Irrigate the feeding tube regularly with water to keep it from blocking. Use within 24 hours.

Chapter 16

LOCAL COLD INJURIES

16. LOCAL COLD INJURIES

16.1 Physiology of thermal regulation 301

16.2 Types of local cold injuries 301
16.2.1 Non-freezing injury 301
16.2.2 Freezing injury 301
16.2.3 Local signs and symptoms 301

16.3 Management 302
16.3.1 First aid and transport 302
16.3.2 Hospital treatment 302
16.3.3 After care 303

16.1 Physiology of thermal regulation

The significance of prevention and treatment of local cold injuries during armed conflict cannot be overemphasized. Although most commonly seen in arctic and subarctic climates, cold injuries can occur whenever the combination of cold, wet, and immobility exists. High altitudes, even in tropical or temperate regions, can experience cold weather. Wind is an aggravating factor in all circumstances.

Normal body temperature is maintained through a balance between heat production and heat loss, and is regulated by a hypothalamic "thermostat". At least 95 % of the heat produced by the metabolism of viscera and muscles is normally lost to the environment by conduction, convection, radiation, and evaporation, largely through the skin and lungs; the head and neck account for 20 – 30 %. The skin primarily dissipates heat by regulating its blood flow, which may vary from 50 ml/min to 7,000 ml/min.

In a cold environment, the *core temperature* (that is, the temperature of the vitally important visceral organs) is conserved by decreasing heat loss through peripheral vasoconstriction, and by increasing the production of heat by involuntary muscle contractions (shivering). If the heat loss exceeds the heat-producing resources of the body, the core temperature begins to fall and *hypothermia* develops. Wetness and wind increase the effect of cold in inducing loss of body heat.

In peripheral tissues subjected to low temperatures, wetness, wind, and contact with a cold surface (metal), local cold injury may result from vasomotor and/or cellular effects, including intracellular ice crystal formation.

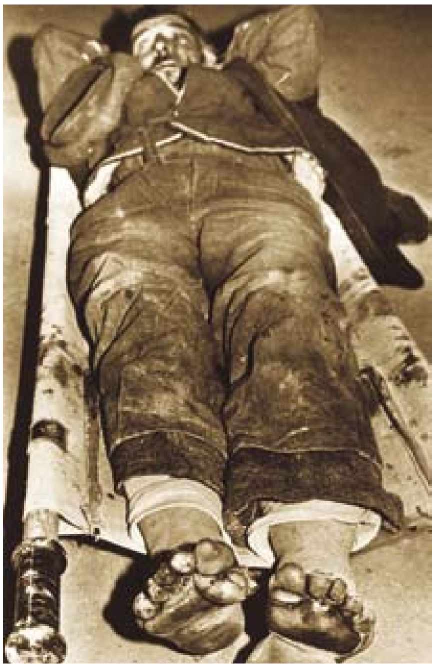

O. Litvin / ICRC

Figure 16.1

Patient with "trench foot" affecting both feet.

16.2 Types of local cold injuries

Local cold injuries can occur at temperatures above or below freezing point and are classified as non-freezing or freezing injuries.

16.2.1 Non-freezing injury

This is also known as "immersion foot" or "trench foot", and is seen in prolonged exposure to cold ambient temperatures above freezing, with high humidity and immobilization, as often witnessed during the trench warfare of World War I, hence the name. Prolonged wetness and cooling of the feet, as can occur in a jungle or rice paddy, can also provoke an "immersion foot" injury. Such injuries are diagnosed and treated like other cold injuries, except that the feet should not be immersed in warm or hot water.

16.2.2 Freezing injury

This type of injury, also known as "frostbite", may be superficial, in which only skin and subcutaneous fat are involved, or deep, reaching structures such as muscle.

Frostbite occurs in the extremities and exposed body parts: nose, ears, etc. In the early stages of injury, it is not easy to differentiate between superficial and deep wounds.

16.2.3 Local signs and symptoms

These include:
- paraesthesia;
- numbness and insensitivity to pin prick;
- pallor (wax-white or mottled blue discoloration);
- impaired movement leading to paralysis;
- firmness of the body part;
- oedematous swelling (particularly in non-freezing injuries) followed by blister formation after 24 – 36 hours.

16.3 Management

16.3.1 First aid and transport

When a body part is frozen solid for hours it runs a lesser risk of tissue loss than when it is exposed to a cycle of thawing, refreezing, and thawing. A person can continue to walk on a frozen foot while it remains frozen; but will be unable to walk after thawing because of the pain and swelling. This should be kept in mind *before* pre-hospital treatment is attempted; it might be preferable to transfer the patient to adequate shelter and facilities rather than begin treatment with inadequate means.

The patient should be moved to a sheltered place as soon as possible. Boots and socks should be removed, avoiding trauma to the skin. As with burns, constricting agents such as rings should also be removed.

As some degree of central hypothermia generally coexists with local cold injury, the general body temperature should be raised by hot drinks, blankets, or skin-to-skin contact. The frozen extremity should be *prevented* from thawing before normal core temperature is reached.

Once hypothermia has been corrected, any available form of heat transfer (skin-to-skin contact, foot-in-axilla, hand-to-nose) should be utilized to warm the body part. Rapid rewarming in warm water (40 – 42°C, or just tolerable to elbow immersion) should only be used when it is certain that refreezing can be avoided.

Analgesia is essential, as thawing of frostbite is extremely painful.

Do not:

- rub or massage injured tissue;
- apply ointments or other topical medication;
- break blisters;
- rewarm by exposure to fire, radiant heat or excessively hot water.

16.3.2 Hospital treatment

Central hypothermia should be dealt with *first* through external rewarming by blankets and warm water bath. In severe cases (core body temperature below 30° C) combined or not with frostbite, central warming has a priority over peripheral warming because of the risk of "after-drop". This represents a situation where the core temperature tends to decrease during peripheral heating of the body. Rewarming of the limbs causes a local vasodilatation with shunting of cold stagnated blood towards the central part of the body; arrhythmias and cardiac arrest may result. Central core warming measures include warm i.v. fluids, rectal enema, and bladder, gastric and peritoneal lavage at 37° C. A low-reading thermometer is necessary to satisfactorily monitor core temperature (see Chapter 18).

Hypothermia takes precedence over local cold injury.

General rewarming should precede local rewarming.

Once hypothermia has been sufficiently corrected, attention may be given to the local cold injury. Superficial injuries may be rewarmed rapidly in 40 – 42° C water. For deep injuries, if the limbs are still frozen or cold and vasoconstricted, they should be rewarmed with dry heat at 37 – 39° C. Even if only one limb is affected, both should be rewarmed together until the nail beds become pink. The patient should receive 100 % oxygen, warmed and humidified, during the warming process.

If the patient is comatose, it should be borne in mind that the patient is not dead until he is *warm* (33° C core temperature) and dead!

Aspirin still appears to be a most useful drug for pain relief and possibly for preventing tissue loss. Pethidine may be added as necessary. Amitriptyline is the preferred analgesic for immersion injury. Tetanus prophylaxis and penicillin should be given. Smoking is contraindicated.

Heparin, anticoagulants, corticosteroids, antihistamines and intravenous dextran have all been shown to be of little benefit. Disagreement exists on the effect of sympathectomy.

16.3.3 After care

Once rewarming is complete, little more can be done to alter the course of events.

The basic treatement is conservative wound care.

Good nursing care and physiotherapy is the essence of conservative treatment. The extremities should be kept on sterile sheets under cradles. Sterile cotton pledgets are placed between toes or fingers. Warm povidone iodine soaks twice daily help prevent superficial infection. As blebs appear, precautions are taken to avoid their rupture: they should not be allowed to dry up. The limb should be placed so as to avoid pressure on the injured part as much as possible. A good functional result will be helped by maintaining active exercises and elevating the affected limb.

It is difficult to predict the extent of eventual tissue loss from frostbite during the first weeks after injury. Tissue loss, however, is generally smaller than expected. It is therefore important to wait until the damaged tissue becomes necrotic and mummified, with a clear line of demarcation and spontaneous amputation of fingers or toes. As with burns, escharotomy of circumferential injuries, and even fasciotomy, might be necessary.

Wait until a clear line of demarcation appears between necrotic and viable tissue.

The *surgical treatment* of local cold injury is to avoid excision of tissues unless secondary infection intervenes. The natural process is allowed to happen: “frozen in January, amputation in July”.

Chapter 17

ANAESTHESIA AND ANALGESIA IN WAR SURGERY[1]

---

[1] Much of this chapter is based on the report of the Senior Anaesthetists Workshop held in Geneva, November, 2002 (see Introduction).

17. ANAESTHESIA AND ANALGESIA IN WAR SURGERY

17.1 Introduction 307

17.2 Anaesthesia methods 308

17.3 Local and regional anaesthesia 309

17.4 Dissociative anaesthesia with ketamine 310
17.4.1 General considerations 310
17.4.2 Ketamine i.m. and i.v. bolus 311
17.4.3 Ketamine infusion anaesthesia 311
17.4.4 Ketamine analgesia 312

17.5 Post-operative pain management 312
17.5.1 General guidelines 312
17.5.2 Pain scoring systems 313

ANNEX 17. A ICRC pain management protocols 314

17.1 Introduction

Standard anaesthetic practices, as utilized in trauma care, should be followed. However, providing safe and effective anaesthesia under circumstances of limited resources is probably the most challenging task of hospital work. Many limitations must be accepted owing to security concerns, lack of infrastructure, and difficult logistics.

A reminder to surgeons: operations are "big" or "small".
All anaesthesia is potentially fatal.

This chapter is not intended for anaesthetists, but rather presents what surgeons should know about anaesthesia when working under precarious circumstances. There are several reasons for this.

1. Surgeons perform "big" operations and "small" operations. There is no such thing as a "small" anaesthesia. Every anaesthesia is potentially fatal. The limits of operational activity in the theatre are not those of the surgeon's expertise. They are determined far more by the level of competency and sophistication of the anaesthesia department. The anaesthetist will tell the surgeon what can be done; not the other way around. The surgeon must understand, and accept, this limit. There is only one other limiting factor that is as important as anaesthesia, if not more so, in determining the level of sophistication of surgery to be performed: postoperative nursing care.
2. In wartime, a surgeon may sometimes find himself without an anaesthetist. In this case, the war surgeon must know how to administer safe and adequate anaesthesia in order to perform the most important and basic of lifesaving procedures. There is much that a surgeon alone can do in terms of anaesthesia. Good knowledge of local anaesthesia techniques – practical, cheap and safe – can be very helpful when resources are limited.
3. Moreover, in many countries practising surgeons far outnumber the anaesthetists, and this is likely to remain the case in the foreseeable future. Anaesthesia under these circumstances is usually administered by a nurse or technician under the "guidance", and medical responsibility, of the surgeon, who must understand the important indications and contraindications of the various anaesthetic techniques. Needless to say, the surgeon must know about the possible complications of the anaesthesia used and how to counter them.

It would be presumptuous to try to explain the safe administration of anaesthesia in a half-dozen pages. This account comprises the basic precepts that a surgeon should be familiar with. For a full description of appropriate anaesthetic techniques under circumstances of limited resources, practicable by a surgeon or other medical officer, the reader is referred to the excellent texts listed in the Selected bibliography.

Figures 17.1.1 – 17.1.4

Standard ICRC equipment for anaesthesia.

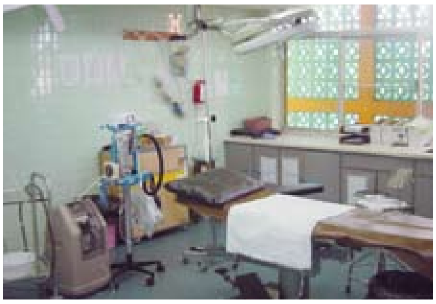
R. Aburabi / ICRC

Figure 17.1.1
Typical operating theatre. Note the inhalation anaesthesia machine and oxygen concentrator.

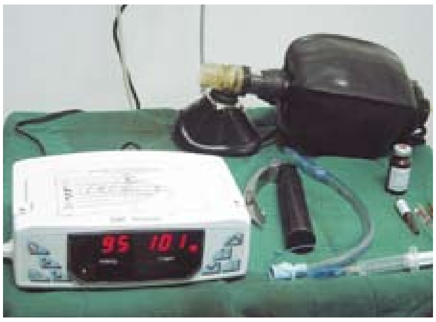
M. Baldan / ICRC

Figure 17.1.2
Pulse oxymeter for monitoring; laryngoscope and endotracheal tube, bag for manual ventilation.

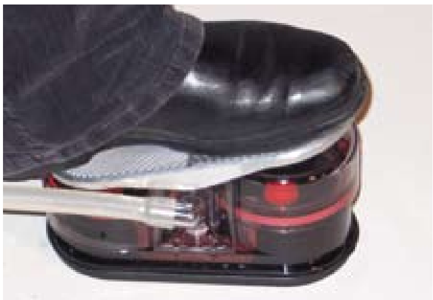
C. Gerber / ICRC

Figure 17.1.3
Foot-driven suction pump.

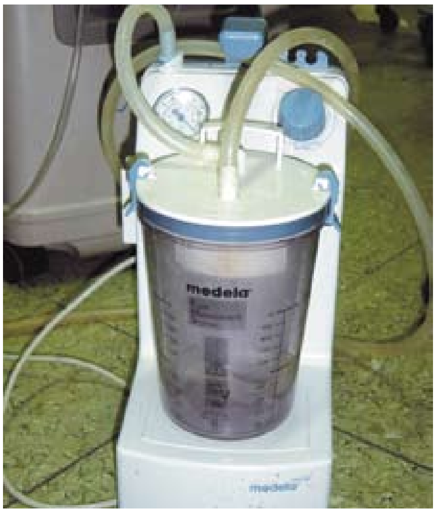

M. Baldan / ICRC

Figure 17.1.4
Electric suction pump.

ICRC EXPERIENCE

The points listed below have been found to be particularly important by ICRC anaesthetists and may be useful to other humanitarian organizations or foreign staff operating in a war environment in a new and strange context.

- The presence of a good interpreter, preferably with a medical background, is essential for communication with patients.
- Unfamiliarity with anaesthetic equipment and local conditions increases the likelihood of complications. The presence of an assistant knowledgeable about local standards is essential.
- The assistance of a second person – doctor, nurse – preferably with experience in anaesthesia, is mandatory for all anaesthetic procedures.
- Assessing the time the patient last ate is difficult, especially with breastfed babies.
- Many of the wounded, especially in hot climates, are severely dehydrated. It is important to correct any hypovolaemia before surgery.
- Routine premedication is not necessary.
- Arrangements for blood transfusion may be complicated (negotiating with family or clan members) and require local blood-collecting procedures to be activated. Likely blood requirements should be ordered early.

17.2 Anaesthesia methods

Local and regional anaesthesia techniques are excellent and can be used in many patients, although they tend to be underestimated and, therefore, underused: a below-knee amputation or Caesarean section can be performed under local anaesthesia, for example. They are safe in avoiding vomiting and aspiration, especially when the time of the last meal cannot be determined accurately.

Oxygen, nitrous oxide and other volatile anaesthetics may be in short supply. In the context of actual combat, oxygen cylinders are to be banned. Not only are the logistics for refilling cylinders and transporting them difficult and dangerous, but the presence of an oxygen cylinder is the equivalent of a bomb. Oxygen concentrators – with a pulse oxymeter for control – are standard ICRC equipment for these circumstances; however, they require an electrical supply.

Ketamine is the anaesthetic of choice in ICRC practice for major surgery. Equipment for the use of ketamine is minimal – not even an oxygen concentrator is necessary – and muscle relaxants can be added to provide full general anaesthesia. The further the hospital from the actual conflict zone, the more sophisticated the equipment and supplies that may be provided, and therefore the techniques employed.

The following anaesthetic agents and their characteristics should be well known to surgeons.

17.3 Local and regional anaesthesia

The clinically significant point in all forms of local and regional anaesthesia is to allow sufficient time for the anaesthetic to work. The most common mistake is to make a local anaesthetic infiltration, and then incise immediately.

	Technique	Drug of choice	Recommended volume	Remarks
Surface anaesthesia	Ophthalmic	0.4 % oxybuprocaine		If not available, 2 % or 4 % lidocaine (xylocaine) drops
	Mucosa	2 % lidocaine gel or 5 % lidocaine nebuliser spray		
Local infiltration anaesthesia	In general	1 % lidocaine with adrenaline	40 ml	If a volume greater than 40 ml is required, dilute with an equal volume of isotonic saline
	Fingers, toes, ears or penis	1 % lidocaine without adrenaline	20 ml	Danger of ischaemic gangrene if adrenaline is used
Nerve block	Digital block	2 % lidocaine without adrenaline	2 – 4 ml per digit	
	Axilla: brachial plexus block	2 % lidocaine with adrenaline or 0.5 % bupivacaine with adrenaline	30 – 40 ml	Where early movement is desired Provides prolonged sensory analgesia
	Intercostal block (flail chest)	0.5 % bupivacaine with adrenaline	2 ml for each nerve	Prolonged analgesia required
Regional intravenous anaesthesia		0.5 % lidocaine without adrenaline	40 ml	
Spinal anaesthesia	Subarachnoid block	0.5 % bupivacaine in 5 % dextrose		Hyperbaric, single dose ampoule
	Epidural	0.5 % bupivacaine	7 – 30 ml according to level of anaesthesia	Requires appropriate conditions of hygiene Not to be used routinely

Table 17.1 Anaesthetic agents of choice for local and regional anaesthesia.

Doses above are for adults. As a "rule of thumb", the maximum dose of lidocaine without adrenaline is 3 mg/kg (200 mg for an adult); and double with adrenaline: 6 mg/kg.

17.4 Dissociative anaesthesia with ketamine

The main components of safe and adequate general anaesthesia are:
- unconsciousness (hypnosis),
- analgesia,
- amnesia,
- immobility/muscle relaxation.

Ketamine is the anaesthetic of choice for major war surgery in conditions of limited resources.

17.4.1 General considerations

Ketamine is very safe, and can be given intramuscularly, or as an intravenous bolus or perfusion.

M. Baldan / ICRC

Figure 17.2
Drugs for general ketamine anaesthesia.

It causes amnesia, analgesia, and a state of dissociation. The patient feels detached from the environment, but his reflexes are intact, particularly the laryngo-pharyngeal reflexes that protect the airway. Opening of the eyes, shouting, and limb movements are frequent and *normal*; the patient is nevertheless anaesthetized. It may cause hallucinations and should be combined with a sedative such as a benzodiazepine (diazepam is inexpensive but burns on injection; midazolam is not irritant but is costly).

Ketamine increases the cardiac output and the arterial blood pressure and, therefore, is particularly useful in patients with haemorrhagic shock. Ketamine dilates the bronchi, but also increases bronchial secretions and saliva. Atropine should be administered first to control saliva and bronchial secretions. There are important contraindications to the use of atropine: marked tachycardia, hypertension, valvular stenosis, hyperthyroidism, or fever. Ketamine also increases the muscular tone, which is why it makes a laparotomy, and especially the closure of the abdomen, more difficult if used without myorelaxants.

There is no absolute contraindication for ketamine anaesthesia. Relative contraindications include:
- psychiatric disease,
- mitral and aortic stenosis,
- untreated hyperthyroidism,
- eclampsia,
- epilepsy,
- operation on the eye – this is due to nystagmus, which causes the globe to move and makes operating difficult.

The question of the use of ketamine in patients with a head injury or increased intracranial pressure has been controversial, although based on very few and very early reports in the 1970s. There is no longer any contraindication for its use in these patients consistent with evidence-based medicine, and ICRC standard practice includes craniotomy performed under ketamine anaesthesia (see Selected bibliography).

The *advantages* of ketamine are the following:
- easy and rapid to administer with quick onset of action;
- safe;
- provides anaesthesia, amnesia, and analgesia;
- circulation is stimulated;
- respiration remains normal (very slow intravenous injection);
- protective airway reflexes remain intact in the majority of patients – vomiting may however occur: staff should be ready to suck clean the mouth and secure the airway;
- maintains cerebral blood flow;
- particularly suitable for infants.

17.4.2 Ketamine i.m. and i.v. bolus

A simple form of ketamine administration is represented by both the intramuscular and the bolus intravenous routes. Table 17.2 compares the two.

	Ketamine i.m.	Ketamine i.v. bolus
Indications	Short duration operation (10 – 20 minutes) Anaesthesia in children (injected while the mother is holding the infant) Repeated injections for change of dressings in burn patients with poor i.v. access	Short duration operation (10 – 20 minutes) Induction of anaesthesia
Premedication	Preferable if time allows In children diazepam may be given orally or rectally	Preferable if time allows
Technique	Ketamine and atropine can be mixed in the same syringe Diazepam should be given in a separate syringe	Place an intravenous line and inject atropine Inject diluted diazepam solution very slowly over 3 minutes, until the patient becomes drowsy Inject ketamine very slowly over one minute (rapid injection can provoke respiratory arrest)
Latent period	5 – 10 minutes	Painful stimuli (skin incision) can be tolerated after one minute
Continuing anaesthesia: second dose	Second injection can be given if necessary before the effect of the first wears off	After 10 – 15 minutes, the patient begins to feel pain and reacts to stimuli with movements and speech – not to be confused with normal movements under ketamine Administer another i.v. bolus: one-third to one-half the initial dose

Table 17.2 Comparison of ketamine anaesthesia administered intramuscularly versus intravenous bolus.

In special circumstances (repeated ketamine anaesthesias with "ketamine-resistance" and/or disturbing limb movements), one or more of the following may be added according to the reaction of the individual patient:

- 50 – 100 mg tramadol or 5 –10 mg morphine,
- 10 – 25 mg promethazine,
- 50 – 100 mg thiopental.

17.4.3 Ketamine infusion anaesthesia

This is the preferred technique in ICRC practice. It is not only more economical of ketamine, but allows for a longer operation without re-injections. It can be used either after i.v. bolus induction of anaesthesia or as an induction method on its own.

A solution is made of ketamine in normal saline and placed in a different vein from that used for fluid replacement. The rate of infusion is titrated against the patient's response, both for induction and as maintenance of anaesthesia.

Ketamine infusion anaesthesia can be combined with muscle relaxants and endotracheal intubation. This is standard ICRC procedure where muscle relaxation is necessary (abdominal or thoracic surgery). As mentioned previously, the absence of mechanical ventilators means that the paralysed patient must be bagged manually.

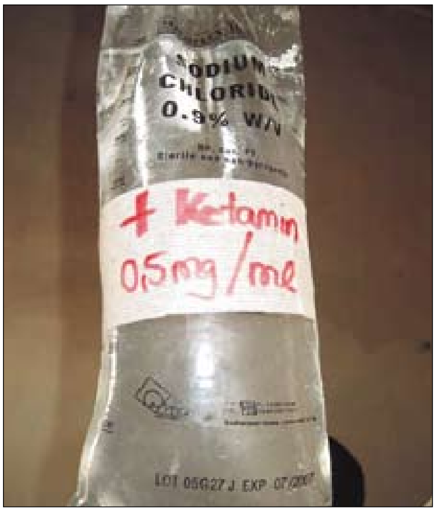

ICRC

Figure 17.3
Ketamine infusion.

ICRC EXPERIENCE

The Senior Anaesthetists Workshop recommended that, in exceptional circumstances (extreme emergency with a very small ICRC surgical team, no training of local staff intended, equipment to be withdrawn afterwards), the ICRC deploy small portable ventilators "to free the hands" of the anaesthetist. These ventilators must be able to function with an oxygen concentrator (there must be a regular supply of electricity) rather than with compressed air or oxygen. They were sent to Banda Aceh, Indonesia, after the 2006 tsunami hit this conflict-affected area because of an "epidemic" of tetanus cases encountered there.

Hallucinatory side effects do occur with ketamine, and once encountered are likely to be repeated on other occasions in the same patient. The frequency of hallucinations tends to differ with different population groups, and bizarre reactions may occur in heavy users of alcohol or other drugs. In addition, such patients do not reach the required state of general anaesthesia unless a neuroleptic and opioid are added.

The incidence and severity of hallucinations can be lessened by using diazepam 10 mg i.v. at least 5 minutes before induction and another 10 mg i.m. at the end of the procedure. After the operation, patients often sleep for hours. They should be in a calm and quiet place. On awakening, they respond to orders "show your tongue" or "raise your head". Sometimes these patients are very loud, they shout or sing, and move a great deal. This is normal and not dangerous for them.

17.4.4 Ketamine analgesia

The analgesic effect of ketamine can be profitably used for a number of conditions. One noteworthy example is for repeated change of dressings in burn patients. For analgesia, a lower dose than that given for intramuscular anaesthesia is administered to those whose venous access is compromised, as is often the case with major burns (see Annex 17. A: ICRC pain management protocols).

17.5 Post-operative pain management

Good post-traumatic or post-operative pain relief not only helps to alleviate suffering, but also allows for rapid mobilization of the patient and early physiotherapy which help attain as good a functional result as possible.

17.5.1 General guidelines

1. Analgesia should be given *before* the pain starts.
2. Combination therapy is recommended. Analgesic drugs have a better effect in combination than alone; e.g. paracetamol and a NSAID have additive actions because of their different sites of action.
3. Local anaesthetic infiltration or blocks are used whenever appropriate in conjunction with other forms of analgesia.
4. Injectable analgesics act more quickly and are more effective if administered as an i.v. bolus, titrated until the desired effect has been achieved. This is especially true in case of hypovolaemia and shock where the peripheral circulation is decreased and, therefore, the intramuscular or subcutaneous routes are unreliable.
5. The choice of drugs is made with the staff and facilities in mind; e.g. opioids may be undesirable if proper monitoring is unavailable.
6. It is advisable to institute a system of pain scoring, especially for post-operative pain relief.

17.5.2 Pain scoring systems

Pain intensity can be measured in various ways and the use of a system of measurement is recommended. The choice of scoring system depends on a number of factors, including cultural issues and the level of literacy and numeracy of patients and the nursing staff. In all of them it is the patient who assesses the intensity of pain.

Common scoring systems include the following.

Verbal score
The pain intensity is measured using words:
- none
- mild
- moderate
- severe
- extreme

Number scale
Numbers from 0 to 10 should be used:
- 0 means no pain and 10 the worst pain imaginable.

Visual analogue score
This is analogous to the number scale. The pain intensity is measured on a 10 cm line. The left end indicates "no pain" and the right-hand side means "the worst pain imaginable". The distance in cm from the left edge to the point the patient indicates is the pain score.

The availability of particular drugs for pain relief in a given country is often subject to importation and distribution restrictions. The widespread abuse of opioids in particular has created many problems in the proper management of patients. Annex 17. A includes various alternatives that may be used according to the availability of various medications. These must of course be adapted to specific situations.

ANNEX 17. A ICRC pain management protocols

Analgesic drugs

Modern analgesics can be very sophisticated, and expensive. The ICRC has the following simple analgesic drugs available for each of the three WHO levels:

Level 1:
Paracetamol tablets/syrup, aspirin tablets/injection, the non-steroidal anti-inflammatories (NSAID) diclofenac injection and ibuprofen tablets;

Level 2:
Tramadol injection/tablets;

Level 3:
Morphine injection, pethidine injection, pentazocine injection/tablets.

Opioids

1. When level 3 opioids are used in a hospital, naloxone must also be immediately available.

2. Morphine remains the gold standard for pain relief, intra-operatively as well as pre- and post-operatively.

3. The use of opioids – morphine and pethidine, but not tramadol – may be severely restricted because of the risk of respiratory depression. ICRC surgical teams use opioids only if nursing staff are sufficient in numbers and trained in patient monitoring, as well as in recognition and treatment of respiratory depression. In practice this will usually mean that post-operative opioids are only used in the recovery ward or intensive care unit (ICU).

4. Do not mix opioids (e.g. tramadol and morphine) until an appropriate time has lapsed.

5. In the emergency room, operating theatre and recovery area/ICU, the preferred route for opioid administration is intravenous. If they are used in the ward, the oral or subcutaneous routes are preferred.

6. The fear of addiction when using opioids for pain relief, although a generally held belief, is exaggerated; if a patient is in pain and needs an opioid, the development of addiction is extremely rare. The fear of addiction must not stop good pain control.

7. Sedation score: when using level 3 opioids the level of sedation should be monitored.
   0 = none: patient awake and alert
   1 = mild: occasionally drowsy, but easy to rouse
   2 = moderate: frequently drowsy, but easy to rouse
   3 = severe: drowsy and difficult to rouse
   S = sleep: normal sleep and easy to rouse

Monitoring, diagnosis, and treatment of respiratory depression
When using level 3 opioids the patient must be monitored regularly for:

- blood pressure
- pulse
- temperature
- respiratory rate
- sedation score
- pain score

All findings must be documented.

Respiratory depression

1. Diagnosis

   Respiratory depression is diagnosed by:
   - sedation score of 3: this is the earliest and most reliable sign;
   - respiratory rate 8 or less: this comes later than sedation;
   - decreasing pO_2, as shown by pulse oximetry: this is a late sign especially if the patient is also being given oxygen.

2. Treatment:
   - oxygen;
   - respiratory assistance with bag and mask if necessary;
   - naloxone i.v. in 50 mcg increments until clinical signs improve.

Remember that the duration of action of naloxone is shorter than morphine and therefore it may need to be repeated; alternatively, use a continuous infusion of 1 – 5 mcg/kg/hr.

Never must a patient be discharged to the ward with a sedation level of 3 or more, or a respiratory rate of 8 or less, or if respiratory depression has occurred.

Ketamine

A low dose of ketamine is a good alternative analgesic where opioids are not available or for high-risk patients. Give repeated doses 0.1 – 0.3 mg/kg i.v., titrated until the desired analgesia is achieved; or an i.m. bolus of 2 – 3 mg/kg. Ketamine in a low dose does not require routine atropine and diazepam as adjuvant therapy.

Pre-hospital care

For pre-hospital first aid, often through National Red Cross/Red Crescent Societies, ICRC distributes only:

- paracetamol tablets/syrup;
- tramadol injection (does not provoke respiratory depression).

Emergency room

The following are available in the emergency room:

- paracetamol
- diclofenac injection
- tramadol injection

In a hospital with adequate nursing care, the following are also provided:

- morphine injection
  - adult: 1 – 3 mg i.v. titrated
  - child: 0.05 mg/kg i.v. titrated
- low-dose ketamine (see above).

Operating theatre

All analgesic drugs are available and their intra-operative use is greatly encouraged. The choice of drug depends to a large extent on the quality of post-operative monitoring.

1. Opioids:

   note the level of post-operative monitoring.

2. NSAID:

   it is advisable to administer an NSAID before the end of surgery.

3. Ketamine:

   0.1 – 0.3 mg/kg boluses as analgesia can be used if opioids are unavailable.

4. Local and regional anaesthesia:

   The use of local anaesthetic infiltration or local and regional blocks is greatly encouraged.

Post-operative pain control

The level of pain control needed depends on the psychology of the patient, the type of surgery, and the time since surgery. The basic principles of post-operative analgesia are the following.

1. Give analgesia regularly and not on an as necessary basis (PRN).
2. Do not wait for the pain to be felt, but start immediately once the patient has regained consciousness. This also means that analgesia should be started before spinal anaesthesia has worn off.
3. Start with combination treatment and decrease doses in subsequent days.
4. Review post-operative analgesia regularly.
5. Use local anaesthetic infiltration or blocks as much as possible.

MILD PAIN

Paracetamol
+
Local infiltration or block

MODERATE PAIN

Paracetamol
+
NSAID
+
Local infiltration or block

SEVERE PAIN

Paracetamol
+
NSAID
+
Opioid
+
Local infiltration or block

Analgesic doses

ADULT			
PARACETAMOL	oral	1 g QID Max. 4 g/day	
IBUPROFEN	oral	400 mg TID/QID Max. 2.4 g/day	Caution in cases of asthma and renal impairment Max. 72 hours
DICLOFENAC	i.v./i.m.	75 mg BID Max. 150 mg/day	Caution in cases of asthma and renal impairment Max. 72 hours
TRAMADOL	oral/i.v.	50 – 100 mg 4 hourly Max. 600 mg/day	
PETHIDINE	i.m.	50 – 150 mg 3 hourly	
	i.v.	10 mg increments	Titrate to effect
MORPHINE	subcutaneous / i.m.	5 –15 mg 4 hourly	
	i.v.	2 mg increments	Titrate to effect
NALOXONE	i.v.	50 mcg increments	Repeat until clinical signs improve

PAEDIATRIC			
PARACETAMOL	INFANTS (0 – 12 months)		
	oral / suppository if available	Loading dose: 15 mg/kg Maintenance: 10 – 15 mg/kg QID Max. dose: 60 mg/kg/day	
	CHILDREN		
	oral / suppository if available	Loading dose: 20 – 30 mg/kg Maintenance: 20 mg /kg QID Max dose: 90 mg/kg/day	
IBUPROFEN	oral	20 mg/kg/day in 3 – 4 divided doses Max. single dose: 200 mg Max. daily dose: 800 mg	Not under 6 months (immature kidneys) Caution in cases of asthma and renal impairment Max. 72 hours
DICLOFENAC	i.m.	1 mg/kg TID Max. single dose: 50 mg Max. daily dose: 150 mg	Not under 6 months (immature kidneys) Caution in cases of asthma and renal impairment Max. 72 hours
TRAMADOL	Not recommended, but commonly used in some European countries for children under 1 year		
PETHIDINE	i.m.	1 mg/kg 4 hourly	
	i.v.	0.25 – 0.5 mg/kg increments	Titrate to effect
MORPHINE	subcutaneous / i.m.	0.05 – 0.1 mg/kg 4 hourly	
	i.v.	0.05 mg/kg increments	Titrate to effect
NALOXONE	i.v.	4 mcg/kg	Repeat until clinical signs improve

BID: *bis in die* (twice a day)

QID: *quater in die* (four times a day)

TID: *tri in die* (three times a day)

Chapter 18

DAMAGE CONTROL SURGERY AND HYPOTHERMIA, ACIDOSIS, AND COAGULOPATHY

18. DAMAGE CONTROL SURGERY AND HYPOTHERMIA, ACIDOSIS, AND COAGULOPATHY

18.1 Resuscitative surgery and damage control surgery 321
18.1.1 Damage control: three-stage protocol 321

18.2 Hypothermia, acidosis, and coagulopathy 322
18.2.1 Hypothermia 323
18.2.2 Acidosis 326
18.2.3 Acute coagulopathy associated with trauma 326

18.1 Resuscitative surgery and damage control surgery

Resuscitative surgery has long been practised, and re-invented by many surgeons beginning with Pringle in 1908 and Halstead in 1913. Today, this is especially the case in critical situations where blood for transfusion is in short supply. The surgeon controls massive haemorrhage with temporary measures and then stops the operation. The patient is taken to the ward to continue resuscitation and, 24 – 48 hours later when haemodynamically stable, is re-operated for the definitive procedure.

ICRC EXPERIENCE

Afghan colleagues working with ICRC surgical teams in the Jalalabad Teaching Hospital in 1993 developed a simple protocol for dealing with patients severely injured by anti-personnel landmines. Many patients had suffered a traumatic amputation of one leg and severe injuries to the other and succumbed, exsanguinated, by the time the debridement of the second leg had been performed.

Our Afghan colleagues then decided to divide the surgical procedure into two parts: the first operation was on the traumatic amputation, the other leg was simply washed and dressed, and the operation ended. Fresh whole blood was sought from family members and the patient fully resuscitated and given penicillin. Forty-eight hours later, another operation to debride the second leg was performed. This was Afghan "damage control surgery", locally invented, to face the situation of a lack of blood for transfusion.

Shortly after a war surgery seminar in Kinshasa, two young Congolese general practitioners belonging to the military medical services, with limited surgical expertise, told the ICRC surgeon of one of their experiences and asked a simple question. While working in a field hospital in the bush, they received a comrade with a bullet injury to the abdomen. Evacuation to another hospital was impossible. They operated and found a severe wound to the liver, and had no blood for transfusion.

"We couldn't stop the bleeding, and not knowing what to do, we packed the liver and stopped the operation", they recounted. They then managed to persuade some other soldiers to donate several units of blood, and re-operated the patient after 48 hours; the patient survived. "Did we do the right thing?" they asked. The ICRC surgeon replied that they had discovered on their own what was now standard practice in many parts of the world when faced with exsanguinating haemorrhage.

18.1.1 Damage control: three-stage protocol

Recently, it has become clear that fixing the tissues and organs during a long operation is not sufficient if certain physiological limits are thereby exceeded.

In response, a three-stage protocol for "damage control surgery"[1] has been defined to preclude the risk of pushing the patient beyond those limits into the "lethal triad" of hypothermia, metabolic acidosis, and coagulopathy that follows much major trauma.

First stage:
Operation, as short as possible, to control haemorrhage and contamination: do the least to overcome life-threatening conditions.

Second stage:
Resuscitation, to stabilize the patient by correction of shock, hypothermia, acidosis, and coagulopathy.

Third stage:
Operation, for definitive repair.

---

[1] "Damage control" is originally a United States Navy term denoting "the capacity of a ship to absorb damage and maintain mission integrity". In surgery, we mean the procedures necessary to maintain the physiological capacity to maintain life's vital functions.

The success of this formalized protocol depends almost entirely on the early recognition that the injury is so severe as to require a curtailment of the initial operation; and, especially, the second-stage capacity to correct certain physiological parameters in the intensive care unit (ICU).

Damage control surgery is obviously restricted to the very severely injured, a small minority. It is a much individualized form of management requiring many resources, and may not be compatible with a situation of triage of mass casualties. The patient who qualifies for damage control in a civilian context, or in a military one with a small number of casualties, will in many cases be triaged into the expectant Category IV in a war scenario.

In much of the world, sophisticated ICUs and blood components are simply not available and a full-scale damage-control approach is not compatible with limited resources. However, basic principles apply and simple measures can and must be adapted to try to prevent and correct the syndrome of the lethal triad.

In the two examples cited above, the time between operations – used to obtain fresh whole blood from family and friends and “stabilize” the patient – helped to fight hypothermia, acidosis, and coagulopathy without knowing it. Attention was paid only to the haemodynamics and the comfort of the patient (keeping him warm), but this was sufficient to overcome the triad. What started as a “resuscitative surgery” approach became a “damage control” one unknowingly.

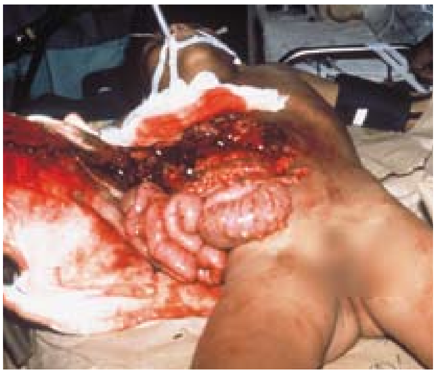

ICRC

Figure 18.1

Typical patient who would benefit from a damage control approach: eventration and a shattered liver.

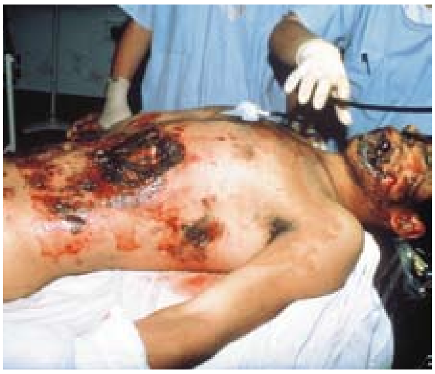

R. Gray / ICRC

Figure 18.2

Another candidate for damage control surgery: APM injury with penetrating wounds to the abdomen, thorax, left thigh and hand, and the face.

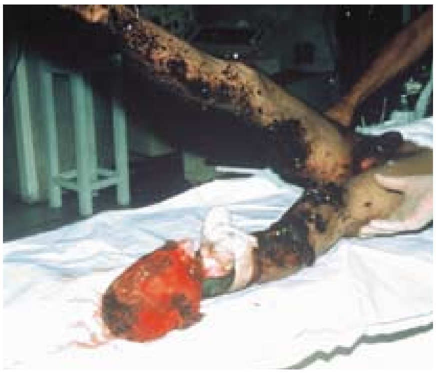

ICRC

Figure 18.3

Patient with APM blast mine injury: traumatic amputation of the left foot, penetrating injuries and burns to both legs, the perineum and genitalia, with penetration of the abdomen.

18.2 Hypothermia, acidosis, and coagulopathy

The effects of hypothermia in trauma patients have long been underestimated. It afflicts the injured even in tropical climates. While metabolic acidosis due to shock – the anaerobic metabolism resulting from reduced tissue perfusion and hypoxia – and coagulopathy are better known, the combined triad is far more prevalent than is often realized and its effects easily fatal. These three elements compound and reinforce each other, leading to a self-sustaining vicious circle. Haemorrhagic shock may not be corrected until hypothermia is reversed, and hypothermia can provoke the other two elements of the triad and is probably the most important factor initiating the vicious circle.

Early diagnosis is essential but simple preventive measures, that should be initiated during first aid and evacuation and continued in the hospital, are often not taken. Even if the context does not allow for staged damage control surgery, a great deal can be done with adequate but appropriate means to prevent or fight off the effects of the triad.

18.2.1 Hypothermia

Physiology

The body's temperature is kept in a homeostatic balance between heat production and heat loss. There are many non-traumatic causes and factors that can lead to hypothermia; cold injuries were discussed in Chapter 16. In trauma patients, haemorrhagic shock decreases tissue perfusion and metabolism and thus heat production. This is often compounded by exposure of the casualty to the elements, particularly in a context of armed conflict.

> It should be taken as a general rule that every injured person will lose body heat, even in a tropical climate.

Hypothermia is determined by central body – core – temperature, taken *per rectum*. A normal medical thermometer is useless; a special low-reading one beginning at 30° C is necessary. Classically, a core temperature below 35° C denotes hypothermia and medical classifications reach profound states of 25° C or less (immersion in a cold lake, hypothalamic disorders, drug abuse, etc.). Some patients suffering from severe hypothermia do survive. There are also reported beneficial effects of controlled therapeutic hypothermia in certain critical pathologies, but this topic shall not be considered here.

Pathology

It is very rare for patients suffering trauma *and* uncontrolled hypothermia lower than 32° C to survive. This observation is not new, although the threshold has changed.

> "It has been said that any wounded man in a state of stupor whose temperature sinks below 36° C will die."
>
> E. Delorme, 1915[2]

In modern practice, the temperature of 32° C is recognized as the crucial cut-off point. All enzyme systems in the body are temperature dependent and, therefore, all organ systems will tend to fail at this low core temperature, especially if stress is accentuated by trauma and shock.

Thus, medical classification systems of hypothermia with wide ranges of temperature are not suitable for the trauma patient. A more appropriate classification system has been proposed (Table 18.1), and takes into account the profound effects of trauma and haemorrhage with concomitant acidosis and coagulopathy. In the trauma patient, any temperature below 36° C should be considered hypothermic and it is particularly important to distinguish between Stage I and Stage II. A damage-control approach, with abbreviated first operation as mentioned above, is indicated at 34° C, now considered the critical clinical temperature.

General medical classification		Trauma classification	
Mild	35° – 32° C	I	36° – 35° C
		II	34° – 32° C
Moderate	32° – 28° C	III	32° – 28° C
Severe	28°– 20° C	IV	< 28° C
Profound	< 20° C		

Table 18.1 Classification systems of hypothermia.[3]

---

[2] Delorme E. *War Surgery*. [Translated by Méric H.] London: H.K. Lewis; 1915. [WWW Virtual Library, The Medical Front WWI website] Available at: http://www.vlib.us/medical/delorme/delorme.htm.

[3] Adapted from Kirkpatrick AW, Chun R, Brown R, Simons RK. Hypothermia and the trauma patient. *Can J Surg* 1999; **42**: 333 – 343.

The *clinical effects* of hypothermia are multiple and mimic intense sympathetic stimulation:

- shivering – the patient attempts to produce body heat by these muscular contractions but this leads to increased oxygen consumption and tissue hypoxia;
- hypoventilation – leads to postoperative hypoxaemia, with further tissue hypoxia;
- peripheral vasoconstriction – to shunt blood to the central organs, with tissue hypoxia;
- decreased oxygen liberation in the tissues (O_2 dissociation curve shifted to the left);
- decreased metabolism.

Furthermore, the tissue hypoperfusion and hypoxia of shock in the trauma patient also increase anaerobic metabolism. All of which result in a mixed metabolic and respiratory *acidosis*.

Enzyme failure results in compromised platelet function; platelet-fibrin clots are not sustained; and fibrinolytic activity is increased. There is dramatic prolongation of prothrombin time (PT) and partial thromboplastin time (PTT).

Please note:

Prolonged PT and PTT may be overlooked because the testing is performed at 37° C rather than at the patient's actual temperature.

There is in addition a rapid drop in the number of platelets and increased blood viscosity following shock. The end result is a disseminated intravascular *coagulopathy* with a marked bleeding tendency.

Other physiological effects include:

- decreased immune response;
- hyporeflexia and prolonged action of neuro-muscular blockers;
- reduced cardiac output, depressed contractility and bradycardia;
- arrhythmias, including atrial and ventricular fibrillation, which begin at 30° C;
- impaired hepatic and renal function;
- impaired cerebral blood flow with decreasing consciousness.

Clinical picture

Early signs include shivering and tremors; the patient first complains of cold then becomes confused. There is cyanosis and the axilla and groin (normally warm spots) feel cold to the touch because of intense vasoconstriction. Tachycardia and tachypnoea (with hypoventilation) occur. All are signs of sympathetic stimulation. Making the diagnosis of hypothermia is a challenge because *most of the early signs* are the normal physiological response to the stress of trauma and haemorrhage; the surgeon should be particularly attentive to the presence of shivering and tremors. It is also difficult to judge clinically when the patient passes from Stage I to II without close monitoring of the rectal temperature.

As the patient goes into Stage III, there is a general slowing down of all the vital functions:

- loss of shivering and hyporeflexia;
- respiratory and circulatory depression;
- decreased urine output;
- increasing acidosis;
- prolonged clotting time;
- stupor; and,
- with myocardial ischaemia, the onset of attacks of atrial fibrillation.

With even greater hypothermia the patient becomes unconscious; bradycardia and bradypnoea are pronounced; urine output fails; and ventricular fibrillation resistant to defibrillation occurs. The patient may appear dead: no palpable pulse or heart beat and dilated pupils. Re-warming should continue until a core temperature of at least 33° C before pronouncing death.

High risk cases

At high risk of developing post-traumatic hypothermia are patients who have:

- been severely injured and are submitted to a long delay in evacuation to hospital;
- remained trapped under rubble (earthquake, collapsed buildings due to bombing, etc.), and thus been exposed to the environment without being able to take protective measures for a considerable period of time;
- sustained major burns;
- suffered haemorrhagic shock treated with large quantities of room temperature intravenous perfusions or cold transfusions – banked blood is stored at 4° C and its administration acts as a heat sink that withdraws large amounts of body heat;
- been subjected to long, extensive laparotomy or thoracotomy in air-conditioned operating theatres – the surgeon and anaesthetist are comfortable, but the patient is dying;
- suffer from chronic illnesses involving metabolism, alcoholism, and drug abuse; and
- the very young and very old, in general.

Please note:

These high risk factors are partly intrinsic to the patient and partly iatrogenic.

Management

Treatment begins with active measures of prevention during first aid and evacuation: the casualty must be kept warm! The patient must be sheltered from cold and wind, relieved of wet clothing, and covered with a dry blanket or sheet, even in a tropical climate.

In hospital, active preventive measures must continue: the emergency room should be kept thermo-neutral (28° C for an adult) and, after the patient has been undressed, examined, and resuscitative measures begun, covered with a blanket. For the severely injured, further environmental measures should continue. In the operating theatre, air-conditioning should be turned off (the OT should be heated if necessary), and prolonged contact with wet drapes avoided. The impact of the latter should not be underestimated: the patient starts off dry and warm and ends up wet and hypothermic.

A patient shivering on the operating table is a clear warning sign!

Under muscle relaxant anaesthesia, a patient can no longer shiver and is particularly at risk. Since 20 – 30 % of body heat is lost through the head and neck, keeping these body parts warm is a priority (e.g. wrapping the head in a towel and plastic bag intra-operatively). The rest of the body, apart from the operative site, can equally be wrapped in dry towels and secured with plastic bags to preserve heat. Warm saline should be used for peritoneal or pleural lavage. Operating time should be kept as short as possible; an open abdomen or thorax inevitably entails loss of body heat which is just as critical at this stage as blood loss.

Administered oxygen should be humidified – a heat and moisture exchange filter (HME) should be used, if available – and intravenous fluids and blood transfusions warmed. Fluids and blood can be heated by being placed in a tub of hot water, comfortable to an elbow touch.

All of these simple measures, which do not require sophisticated equipment, should be standard procedure in all surgical patients, especially those suffering from trauma.

Not only will these simple measures prevent hypothermia, but their early implementation will help treat the condition.

If a damage-control approach is being used, or even if a major trauma case presents signs of hypothermia postoperatively, resuscitation should include active internal "central heating" of the patient by:

- gastric, colonic and bladder lavage with warm water (37 – 39° C); and
- continued perfusion of similarly warm i.v. fluids.

More sophisticated technologies exist – including extracorporeal circulation, pleural and peritoneal lavage in the ICU – but the vast majority of patients will be well served by these simple and inexpensive preventive and active measures.

A patient in Stage I hypothermia may have definitive surgery completed. A Stage II patient should have damage control surgery only. If Stages III or IV have been reached, damage control surgery may be slightly postponed in order to warm the patient somewhat before opening the peritoneum or pleura, but this will depend on the particular circumstance and haemodynamic stability of the individual patient. Again: rare is the haemorrhaging trauma patient with a core temperature below 32° C who survives, however good the surgery.

Type of measures	Procedures	Applicability
Standard measures	Passive external warming (warm environment, blankets and covers) + warmed intravenous fluids + warmed and humidified oxygen	Standard measures are applicable for all Stages and at all times
Active external warming	Electric blankets Environmental heater	Particularly suitable for ICU/ emergency room
Active internal warming intraoperatively	Irrigation of pleural/peritoneal cavity with warm saline during surgery	Should be routine surgical procedure
Active internal warming postoperatively	Warm gastric, colonic, and bladder lavage Warm pleural/peritoneal lavage if feasible	Particularly suitable for ICU

Table 18.2 **Summary of management of hypothermia in the trauma patient.**

18.2.2 Acidosis

As mentioned, shock will lead to tissue hypoperfusion and hypoxia with a subsequent rise in anaerobic metabolism. This acidosis is compounded by the effects of hypothermia. Full resuscitation and haemodynamic stability in a warm patient, assuring good tissue perfusion, is the best antidote. Intravenous sodium bicarbonate is risky and requires sophisticated monitoring.

18.2.3 Acute coagulopathy associated with trauma

Post-traumatic coagulopathy is also more common than realized and results from a combination of intrinsic and extrinsic factors. Amongst the intrinsic ones, the most important include:

- extensive tissue injury, which releases tissue factors resulting in consumption coagulopathy (with reduced platelet count), excessive fibrinolysis, and activation of inflammatory cascade;
- activation of the C-reactive protein inflammatory cascade independently of the release of tissue factors and entirely due to the shocked state and tissue hypoperfusion;
- haemodilution resulting from the mobilization of extravascular fluids as a homeostatic response to shock;
- decrease in the total and ionized calcium concentrations; and
- the effects of progressive hypothermia and acidosis.

The severity of the coagulopathy seems to parallel the severity of injury and shock.

Important factors resulting in post-traumatic coagulopathy:

- tissue damage
- shock: hypoperfusion and homeostatic haemodilution
- hypothermia
- acidosis
- inflammation
- therapeutic haemodilution with i.v. fluid resuscitation

There are also a number of extrinsic factors. Coagulopathy itself is often exacerbated by clinical practice in a "bloody vicious cycle".[4] Haemorrhagic shock is aggressively treated with large volumes of intravenous fluids (not warmed), leading to further haemodilution, and massive transfusion of stored blood or packed RBCs (not warmed) in the attempt to maintain blood pressure (see Chapter 8); all of which adversely affect the coagulation cascade and only produce more bleeding. Bleeding and coagulation times remain prolonged despite adequate replacement of blood, plasma and platelets if the patient remains hypothermic.

It must also be remembered that the anticoagulant preservative solution and refrigeration of stored blood or its components has profound effects on the physiological properties of transfused blood with time, the most important being greatly reduced oxygen availability to the tissues through decreased 2,3-diphosphoglycerate (2,3-DPG) and loss of red cell viability because of decreased adenosine triphosphate (ATP). These effects do not help overcome early tissue hypoxia in spite of the increase in haemoglobin. In addition, the glucose in stored blood is slowly metabolized with production of lactate and fall in the pH, exacerbating the acidosis.

Warm whole blood, as fresh as possible, is probably the best treatment; and, as often repeated in this manual, in conditions of limited resources is often all that is available. With numerous potential donors (family and friends), it may be possible to build up a small reserve.

Whole blood is screened and best given within 1 hour of collection. In ICRC practice, this freshest of blood is reserved for:

- massive haemorrhage, with incipient coagulopathy and hypothermia;
- septic shock;
- snake bites with haemolysis;
- amniotic fluid embolism.

Intravenous calcium should be administered, separately and at least one ampoule added per two units of blood transfused.

Coagulopathy also attends many patients with severe head injury, and a damage control approach is applicable to injuries to most body systems. Volume 2 will include relevant observations in the discussion of war injuries to the different anatomic regions.

---

[4] Kashuk J, Moore EE, Milikan JS, Moore JB. Major abdominal vascular trauma – a unified approach. *J Trauma* 1982; **22**: 672 – 679.

ACRONYMS

ABCDE	Airway, Breathing, Circulation, Disability, Environment/Exposure
APM	Anti-personnel mine
ARDS	Acute respiratory distress syndrome
ATM	Anti-tank mine
ATP	Adenosine triphosphate
AVPU	Alert, Voice responsive, Pain responsive, Unresponsive
BID	*Bid in die* (twice a day)
CFR	Case fatality rate
CPD-A	Citrate – phosphate – dextrose – adenine
CRO	Carded for record only
CT	Computerized tomography
DOA	Dead on arrival
DOW	Died of wounds
DPC	Delayed primary closure
2,3-DPG	2,3-diphosphoglycerate
E_K	Kinetic energy
$E_{K\,EXP}$	Kinetic energy actually expended, transferred or dissipated
ECG	Electrocardiogram
EDTA	Ethylene diamine tetra-acetic acid
EEG	Electroencephalogram
ER	Emergency (reception) room
FAP	First-aid post
FMJ	Full metal jacket
FST	Field surgical team
GC	Geneva Convention
GCS	Glasgow coma scale
GSW	Gunshot wound

H.E.L.P.	Health emergencies in large populations
HIV/AIDS	Human immunodeficiency virus / Acquired immunodeficiency syndrome
HME	Heat and moisture exchange (filter)
HN	Head nurse
ICRC	International Committee of the Red Cross
ICU	Intensive care unit
IED	Improvised explosive device
IHL	International humanitarian law
i.m.	Intramuscular
i.v.	Intravenous
IU	International unit
KIA	Killed in action
LF	Low frequency
mEq	Milliequivalents
MIU	Million international units
MoPH	Ministry of Public Health
MRSA	Methicillin-resistant *Staphylococcus aureus*
MSF	Médecins sans frontières (Doctors Without Borders)
NATO	North Atlantic Treaty Organization
NGO	Non-governmental organization
NSAID	Non-steroidal anti-inflammatory drug
OPCW	Organization for the Prohibition of Chemical Weapons
OPD	Outpatient department
ORS	Oral rehydration salts
OT	Operating theatre
POP	Plaster of Paris
POW	Prisoner of war
PRN	*Pro re nata* (whenever necessary)
PT	Prothrombin time
PTT	Partial thromboplastin time
QID	*Quater in die* (4 times per day)
RBC	Red blood cell
RC/RC	Red Cross/ Red Crescent
RCWS	Red Cross Wound Score
RTD	Returned to duty
SJ	Semi-jacketed
SSG	Split-skin grafts
TBSA	Total body surface area

TID	*Ter in die* (3 times per day)
TNM	Tumour, node, metastasis
TO	Triage officer
TTL	Triage team leader
UN	United Nations
UNPROFOR – IFOR	United Nations Protection Force – (NATO) Implementation Force
UXO	Unexploded ordnance
WDMET	Wound Data and Munitions Effectiveness Team
WHO	World Health Organization
WIA	Wounded in action

SELECTED BIBLIOGRAPHY

General texts

Please note:

The publications listed under general texts serve as basic references throughout this manual.

Military publications

Kirby NG, Blackburn G, eds. *Field Surgery Pocket Book*. London: Her Majesty's Stationery Office; 1981.

Roberts P, ed. *The British Military Surgery Pocket Book. (AC 12552)* Camberley, Surrey, UK: Department of Military Surgery, Army Medical Directorate; 2003.

Bowen TE, Bellamy RF, eds. *Emergency War Surgery NATO Handbook Second United States Revision*. Washington, DC: US Department of Defense; 1988.

Lounsbury DE, Brengman M, Belamy RF, eds. *Emergency War Surgery Third United States Revision*. Washington, DC: Borden Institute, US Department of Defense; 2004. Available at: http://www.bordeninstitute.army.mil/other_pub/ews.html.

Nessen SC, Lounsbury DE, Hetz SP, eds. *War Surgery in Afghanistan and Iraq: A Series of Case Studies, 2003 – 2007*. Washington, DC: Office of the Surgeon General, Borden Institute, US Department of Defense; 2008.

Pons J, ed. *Memento de chirurgie de guerre [War Surgery Primer]*. Paris: Ecole d'application du Service de santé pour l'armée de terre, ORA éditions; 1984.

War Surgery Commission of the Federal Military Department. *Chirurgie de guerre (Aide-mémoire 59.24 f) [War Surgery (A Primer)]*. Bern: Swiss Army; 1970 and 1996.

Civilian publications

Courbil L-J, ed. *Chirurgie d'urgence en situation précaire [Emergency Surgery under Precarious Circumstances]*. Paris: Editions Pradel; 1996.

Geelhoed GW, ed. *Surgery and Healing in the Developing World*. Georgetown, TX: Landes Bioscience; 2005.

Husum H, Ang SC, Fosse E. *War Surgery: Field Manual*. Penang, Malaysia: Third World Network; 1995.

Loefler I. Africa – Surgery in an unstable environment. Weary Dunlop Memorial Lecture. *Aust NZ J Surg* 2004; **74**: 1120 – 1122.

Mahoney PF, Ryan JM, Brooks AJ, Schwab CW, eds. *Ballistic Trauma: A Practical Guide 2nd Edition*. London: Springer-Verlag; 2005.

Trauma surgery references

Boffard KD. *Manual of Definitive Surgical Trauma Care 2nd Edition*. London: International Association for Trauma Surgery and Intensive Care, Hodder/Arnold; 2007.

Botha AB, Brooks A, Loosemore T, eds. *Definitive Surgical Trauma Skills Manual*. London: Royal College of Surgeons of England; 2002.

Hirshberg A, Mattox KL. *Top Knife: The Art and Craft of Trauma Surgery*. Shrewsbury, UK: tfm Publishing Ltd; 2005.

Integrated Management on Emergency and Essential Surgical Care. *E-Learning tool Kit* [CD-ROM]. Geneva: World Health Organization; 2005.

King M, ed. *Primary Surgery, Volume Two: Trauma*. Oxford: Oxford University Press; 1987. Available at: http://www.primary-surgery.org/ps/vol2/html/index.html. Updated May 2, 2008.

Trauma.org. (Dedicated international trauma-interest web site.)
[Web site http://www.trauma.org].

Chapter 1

Emergency Items Catalogue of the International Movement of the Red Cross and Red Crescent. Geneva: International Committee of the Red Cross and International Federation of Red Cross and Red Crescent Societies; 2004. [CD-ROM] [online] Available at: http://www.icrc.org/emergency-items.

Bowyer GW. War surgery and the International Committee of the Red Cross: a historical perspective. *Int J Orthop Trauma* 1996; **6**: 62 – 65.

Morris DS. Surgeons and the International Committee of the Red Cross. *Aust NZ J Surg* 1992; **62**: 170 – 172.

Mulli J-C. Activités chirurgicales en zone de guerre: l'expérience du Comité international de la Croix-Rouge [Surgical activities in a war zone: the experience of the International Committee of the Red Cross]. *Bulletin Medicus Mundi* 1995; **57**: 42 – 54.

Russbach R. Les unités chirurgicales du Comité international de la Croix-Rouge: le personnel, le matériel, les coûts [Surgical units of the International Committee of the Red Cross: the personnel, the equipment, the costs]. *Médecine et Hygiène* 1991; **49**: 2629 – 2632.

Vassallo DJ. The International Red Cross and Red Crescent Movement and lessons from its experience of war surgery. *J R Army Med Corps* 1994; **140**: 146 – 154.

Chapter 2

Basic references

Dunant H. *A Memory of Solferino*. Geneva: ICRC; 1950.

The Geneva Conventions of August 12, 1949. Geneva: ICRC; 1995.

Pictet J, ed. *Commentary on the Geneva Conventions I-IV of 12 August 1949*, 4 vol. Geneva: ICRC; 1952 – 1960.

Protocols Additional to the Geneva Conventions of 12 August 1949, revised Edition. Geneva: ICRC; 1996.

Sandoz Y, ed. *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949*. Geneva: ICRC; 1987.

David E. *Principes de droit des conflits armés 4e éd [Principles of the Law of Armed Conflicts 4th Edition]*. Brussels: Bruylant; 2008.

Gasser H-P. International Humanitarian Law: An Introduction. In: Haug H, ed. *Humanity for All*. Geneva: Henry Dunant Institute; 1993.

Kolb R. *Ius in bello, Le droit international des conflits armés [Ius in Bello, The International Law of Armed Conflicts]*. Brussels: Bruylant; 2003.

Palwankar U. *Symposium on Humanitarian Action and Peace-Keeping Operations, Geneva, June 22 – 24, 1994*. Geneva: ICRC; 1994.

IHL and the medical profession

Annas GJ. Military medical ethics – physician first, last, always. *N Engl J Med* 2008; **359**: 1087 – 1090.

Baccino-Astrada A. *Manual on the Rights and Duties of Medical Personnel in Armed Conflicts*. Geneva: International Committee of the Red Cross, League of Red Cross and Red Crescent Societies; 1982. (Out of print, under revision.)

British Medical Association. *Medicine Betrayed*. London: Zed Books; 1992.

The Medical Profession and the Effects of Weapons. Symposium: Montreux, Switzerland, 1996. Geneva: Media Natura, Geneva Foundation to Protect Health in War, International Committee of the Red Cross; 1996.

Coupland RM. Weapons intended to blind. [Editorial]. *Lancet* 1994; **344**: 1649 – 1650.

Coupland RM. Wounds, weapons and the doctor. *Schweiz Z Milit Med* 1995; **72**: 33 – 35.

Coupland RM. Abhorrent weapons and "superfluous injury or unnecessary suffering": from field surgery to law. *BMJ* 1997; **315**: 1450 – 1452.

Coupland RM. "Non-lethal" weapons: precipitating a new arms race. Medicine must guard against its knowledge being used for weapon development. [Editorial]. *BMJ* 1997; **315**: 72.

Coupland R, Herby P. Review of the legality of weapons: a new approach The SIrUS Project. *International Review of the Red Cross* 1999; **835**: 583 – 592.

Giannou C. The Mine Information System: The principal factors determining the severity of landmine infestation. Presented at the signing of the Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-personnel Mines and on their Destruction; 1997 Sept; Ottawa, Canada.

Chapter 3

General references

Sellier KG, Kneubuehl BP. *Wound Ballistics and the Scientific Background*. Amsterdam: Elsevier; 1994.

Kneubuehl BP, Coupland RM, Rothschild MA, Thali MJ. *Wundballistik, Grundlagen und Anwendungen [Wound Ballistics, Basics and Applications]*. Berlin, Heidelberg, New York: Springer-Verlag; 2008. (Available in German only.)

Neuenschwander J, Coupland R, Kneubuehl B, Baumberger V. *Wound Ballistics: An introduction for health, legal, forensic, military and law enforcement professionals*. [Brochure and film on DVD]. Geneva: ICRC; 2008.

Journal articles

Bowyer GW, Cooper JG, Rice P. Small fragment wounds: biophysics and pathophysiology. *J Trauma* 1996; **40 (3Suppl.)**: S159 – S164.

Brismar B, Bergenwald L. The terrorist bomb explosion of Bologna, Italy, 1980: an analysis of the effects and injuries sustained. *J Trauma* 1982; **22**: 216 – 220.

Cheng XM, Liu YQ, Guo RF, Lian WK, Wang DT. Analysis of wound ballistics in 2,414 cases of battle casualties. *Journal of Trauma (China)* 1990; **6 (Suppl.)**: S169 – S172.

Cooper GJ, Ryan JM. Interaction of penetrating missiles with tissues: some common misapprehensions and implications for wound management. *Br J Surg* 1990; **77**: 606 – 610.

Coupland RM. Clinical and legal significance of fragmentation of bullets in relation to size of wounds: retrospective analysis. *BMJ* 1999; **319**: 403 – 406.

Coupland RM, Hoikka V, Sjoeklint OG, Cuenod P, Cauderay GC, Doswald-Beck L. Assessment of bullet disruption in armed conflicts. *Lancet* 1992; **339**: 35 – 37.

DePalma RG, Burris DG, Champion HR, Hodgson MJ. Blast injuries. *N Engl J Med* 2005; **352**: 1335 – 1342.

Fackler ML, Malinowski JA. The wound profile: a visual method for quantifying gunshot wound components. *J Trauma* 1985; **25**: 522 – 529.

Fackler ML, Bellamy RF, Malinowski JA. The wound profile: illustration of the missile-tissue interaction. *J Trauma* 1988; **28 (1Suppl.)**: S21 – S29.

Fackler ML. Wound ballistics: a review of common misconceptions. *JAMA* 1988; **259**: 2730 – 2736.

Farjo LA, Miclau T. Ballistics and mechanisms of tissue wounding. *Injury* 1997; **28 (3Suppl.)**: C12 – C17.

Hayda R, Harris RM, Bass CD. Blast injury research: modelling injury effects of landmines, bullets, and bombs. *Clin Orthop Relat Res* 2004; **422**: 97 – 108.

Hollerman JJ, Fackler ML, Coldwell DM, Ben-Menachem Y. Gunshot Wounds: 1. Bullets, ballistics, and mechanisms of injury *Am J Roentgenol* 1990; **155**: 685 – 690.

Houdelette P. Notions de balistique lésionnelle concernant les armes déflagrantes légères antipersonnel [A primer on wound ballistics concerning anti-personnel explosive small arms]. *Médecine et armées* 1997; **25**: 261 – 264.

Hull JB, Cooper GJ. Pattern and mechanism of traumatic amputation by explosive blast. *J Trauma* 1996; **40 (3Suppl)**: S198 – S205.

Katz E, Ofek B, Adler J, Abramowitz HB, Krausz MM. Primary blast injury after a bomb explosion in a civilian bus. *Ann Surg* 1989; **209**: 484 – 488.

Leibovici D, Gofrit ON, Stein M, Shapira SC, Noga Y, Heruti RJ, Shemer J. Blast injuries: bus versus open-air bombings: a comparative study of injuries in survivors of open-air versus confined-space explosions. *J Trauma* 1996; **41**: 1030 – 1035.

Lindsey D. The idolatry of velocity, or lies, damn lies, and ballistics. *J Trauma* 1980; **20**: 1068 – 1069.

Peters CE, Sebourn CL, Crowder HL. Wound ballistics of unstable projectiles. Part I: projectile yaw growth and retardation. *J Trauma* 1996; **40 (3Suppl)**: S10 – S15.

Peters CE, Sebourn CL. Wound ballistics of unstable projectiles. Part II: Temporary cavity formation and tissue damage. *J Trauma* 1996; **40 (3Suppl)**: S16 – S21.

Santucci RA, Chang Y-J. Ballistics for physicians: myths about wound ballistics and gunshot injuries. *J Urol* 2004; **171**: 1408 – 1414.

Sebourn CL, Peters CE. Flight dynamics of spin-stabilized projectiles and the relationship to wound ballistics. *J Trauma* 1996; **40 (3Suppl)**: S22 – S26.

Vail S. The study of wound ballistics is based on a significant amount of science and a tremendous amount of art [Defense Review Web site]. May 14, 2006. Available at: http://www.defensereview.com/modules.php?name=News&file=article&sid=875.

Volgas DA, Stannard JP, Alonso JE. Ballistics: a primer for the surgeon. *Injury* 2005; **36**: 373 – 379.

Volgas DA, Stannard JP, Alonso JE. Current orthopaedic treatment of ballistic injuries. *Injury* 2005; **36**: 380 – 386.

Wang ZG, Feng JX, Liu YQ. Pathomorphological observations of gunshot wounds. *Acta Chir Scand* 1982; **508 (Suppl.)**: S185 – S195.

Wang ZG, Tang CG, Chen XY, Shi TZ. Early pathomorphological characteristics of the wound track caused by fragments. *J Trauma* 1988; **28 (1Suppl.)**: S89 – S95.

Chapter 4

Bowyer GW, Stewart MPM, Ryan JM. Gulf war wounds: application of the Red Cross Wound Classification. *Injury* 1993; **24**: 597 – 600.

Bowyer GW. Afghan war wounded: application of the Red Cross Wound Classification. *J Trauma* 1995; **38**: 64 – 67.

Coupland RM. The Red Cross classification of war wounds: the EXCFVM scoring system. *World J Surg* 1992; **16**: 910 – 917.

Coupland RM. *The Red Cross Wound Classification, Revised Edition.* Geneva: ICRC; 2005.

Giannou CP. Penetrating missile injuries during asymmetric warfare in the 2003 Gulf conflict. [Correspondence]. *Br J Surg* 2005; **92**: 1047 – 1048.

Hinsley DE, Rosell PAE, Rowlands TK, Clasper JC. Penetrating missile injuries during asymmetric warfare in the 2003 Gulf conflict. *Br J Surg* 2005; **92**: 637 – 642.

Rosell PAE, Clasper JC. Ballistic fractures: the limited value of existing classifications. *Injury* 2005; **36**: 369 – 372.

Rowley DI. *War Wounds with Fractures: A Guide to Surgical Management.* Geneva: ICRC; 1996.

Savic J, Cernak I, Jevtic M, Todoric M. Glucose as an adjunct triage tool to the Red Cross Wound Classification. *J Trauma* 1996; **40 (3Suppl.)**: S144 – S147.

Stewart MPM, Kinninmonth A. Shotgun wounds of the limbs *Injury* 1993; **24**: 667 – 670.

Vassalo D, McAdam G. Modification to the Red Cross Wound Classification. *Injury* 1995; **26**: 131 – 132.

Chapter 5

Public health

Aboutanos MB, Baker SP. Wartime civilian injuries: epidemiology and intervention strategies. *J Trauma* 1997; **43**: 719 – 726.

Burnham G, Lafta R, Doocy S, Roberts L. Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey. *Lancet* 2006; **368**: 1421 – 1429.

Coupland RM. The effects of weapons on health. *Lancet* 1996; **347**: 450 – 451.

Coupland RM, Meddings DR. Mortality associated with use of weapons in armed conflicts, wartime atrocities, and civilian mass shootings: literature review. *BMJ* 1999; **319**: 407 – 410.

Coupland RM, Samnegaard HO. Effect of type and transfer of conventional weapons on civilian injuries: retrospective analysis of prospective data from Red Cross hospitals. *BMJ* 1999; **319**: 410 – 412.

Giannou C. Antipersonnel landmines: facts, fictions, and priorities. *BMJ* 1997; **315**: 1453 – 1454.

Guah-Sapir D, van Panhuis WG. The importance of conflict-related mortality in civilian populations. *Lancet* 2003; **361**: 2126 – 2128.

Horton R. Croatia and Bosnia: The imprints of war – 1. Consequences. *Lancet* 1999; **353**: 2139 – 2144.

Ityavyar DA, Ogba LO. Violence, conflict and health in Africa. *Soc Sci Med* 1989; **28**: 649 – 657.

Jeffries SJ. Antipersonnel mines: who are the victims? *J Accid Emerg Med* 1996; **13**: 343 – 346.

de Jong K, Mulhem M, Ford N, van der Kam S, Kleber R. The trauma of war in Sierra Leone. *Lancet* 2000; **355**: 2067 – 2068.

Lautze S, Leaning J, Raven-Roberts A, Kent R, Mazurana D. Assistance, protection, and governance networks in complex emergencies. *Lancet* 2004; **364**: 2134 – 2141.

Meddings DR. Weapons injuries during and after periods of conflict: retrospective analysis. *BMJ* 1997; **310**: 1417 – 1420.

Meddings DR, O'Connor SM. Circumstances around weapon injury in Cambodia after departure of a peacekeeping force: prospective cohort study. *BMJ* 1999; **319**: 412 – 415.

Meddings DR. Civilians and war: a review and historical overview of the involvement of non-combatant populations in conflict situations. *Med Confl Surviv* 2001; **17**: 6 – 16.

Perrin P. *War and Public Health: A Handbook*. Geneva: ICRC; 1996.

Perrin P. *H.E.L.P. – Health Emergencies in Large Populations: Public Health Course in the Management of Humanitarian Aid*. Geneva: ICRC; 2001.

Spiegel PB, Salama P. War and mortality in Kosovo, 1998 – 99: an epidemiological testimony. *Lancet* 2000; **355**: 2204 – 2209.

Taback N, Coupland R. Towards collation and modelling of the global cost of armed violence on civilians. *Med Confl Surviv* 2005; **21**: 19 – 27.

The Sphere Project: Humanitarian Charter and Minimum Standards in Disaster Response. Geneva: The Sphere Project; 2004. Available at: http://www.sphereproject.org.

Toole MJ, Galson S, Brady W. Are war and public health compatible? *Lancet* 1993; **341**: 1193 – 1196.

Udwadia TE. Surgical care for the poor: a personal Indian perspective. *Indian J Surg* 2003; **65**: 504 – 509.

Civilian war surgery literature

Bhatnagar MK, Smith GS. Trauma in the Afghan guerrilla war: effects of lack of access to care. *Surgery* 1989; **105**: 699 – 705.

Cutting PA, Agha R. Surgery in a Palestinian refugee camp. *Injury* 1992; **23**: 405 – 409.

De Wind CM. War injuries treated under primitive circumstances: experiences in an Ugandan mission hospital. *Ann R Coll Surg Engl* 1987; **69**: 193 – 195.

Dudley HAF, Knight RJ, McNeur JC, Rosengarten DS. Civilian battle casualties in South Vietnam. *Br J Surg* 1968; **55**: 332 – 340.

Fosse E, Husum H, Giannou C. The siege of Tripoli 1983: war surgery in Lebanon. *J Trauma* 1988; **28**: 660 – 663.

Fosse E, Husum H. Surgery in Afghanistan: a light model for field surgery during war. *Injury* 1992; **23**: 401 – 404.

Nassoura Z, Hajj H, Dajani O, Jabbour N, Ismail M, Tarazi T, Khoury G, Najjar F. Trauma management in a war zone: the Lebanese war experience. *J Trauma* 1991; **31**: 1596 – 1599.

Odling-Smee GW. Ibo civilian casualties in the Nigerian civil war. *BMJ* 1970; **2**: 592 – 596.

Rukovansjki M. Spinal cord injuries caused by missile weapons in the Croatian war. *J Trauma* 1996; **40 (3Suppl.)**: S189 – S192.

Suljevic I, Surkovic I. Medical aspects of the mass-scale civilian casualties at Sarajevo Markale Market on August 28, 1995: triage, resuscitation, and treatment. *Croat Med J* 2002; **43**: 209 – 212.

Military war surgery literature

Acosta JA, Hatzigeorgiou C, Smith LS. Developing a trauma registry in a forward deployed military hospital: preliminary report. *J Trauma* 2006; **61**: 256 – 260.

Bellamy RF. Combat trauma overview. In: Sajtchuk R, Grande CM, eds. *Textbook of Military Medicine, Anesthesia and Perioperative Care of the Combat Casualty*. Falls Church, VA: Office of the Surgeon General, United States Army; 1995: 1 – 42.

Bilski TR, Baker BC, Grove JR, Hinks RP, Harrison MJ, Sabra JP, Temerlin SM, Rhee P. Battlefield casualties treated at Camp Rhino, Afghanistan: lessons learned. *J Trauma* 2003; **54**: 814 – 822.

Burkle FM Jr, Newland C, Meister SJ, Blood CG. Emergency medicine in the Persian Gulf War – Part 3: battlefield casualties. *Ann Emerg Med* 1994; **23**: 755 – 760.

Carey ME. Learning from traditional combat mortality and morbidity data used in the evaluation of combat medical care. *Mil Med* 1987; **152**: 6 – 12.

Chambers LW, Green DJ, Gillingham BL, Sample K, Rhee P, Brown C, Brethauer S, Nelson T, Narine N, Baker B, Bohman HR. The experience of the US Marine Corps' Surgical Shock Trauma Platoon with 417 operative combat casualties during a 12 month period of Operation Iraqi Freedom. *J Trauma* 2006; **60**: 1155 – 1164.

Champion HR, Bellamy RF, Roberts P, Leppäniemi A. A profile of combat injury *J Trauma* 2003; **54 (5Suppl.)**: S13 – S19.

Danon YL, Nili E, Dolev E. Primary treatment of battle casualties in the Lebanon war, 1982. *Is J Med Sci* 1984; **20**: 300 – 302.

Eastridge BJ, Jenkins D, Flaherty S, Schiller H, Holcomb JB. Trauma system development in a theater of war: experiences from Operation Iraqi Freedom and Operation Enduring Freedom. *J Trauma* 2006; **61**: 1366 – 1373.

Fekadu T. *Mass Casualty Management Under Unique War Situation: Inside Eritrea's War for Independance*. Asmara, Eritrea: self published; 2002.

Garfield RM, Neugut AI. Epidemiologic analysis of warfare. *JAMA* 1991; **266**: 688 – 692.

Gofrit ON, Kovalski N, Leibovici D, Shemer J, O'Hana A, Shapira SC. Accurate anatomical location of war injuries: analysis of the Lebanon war fatal casualties and the proposition of new principles for the design of military personal armour system. *Injury* 1996; **27**: 577 – 581.

Gofrit ON, Leibovici D, Shapira SC, Shemer J, Stein M, Michaelson M. The trimodal death distribution of trauma victims: military experience from the Lebanon war. *Mil Med* 1997; **162**: 24 – 26.

Grau LW, Jorgensen WA. Handling the wounded in a counter-guerrilla war: the Soviet/Russian experience in Afghanistan and Chechnya. *U.S. Army Medical Dept Journal* 1998; **Jan/Feb**: 2 – 10.

Hardaway RM III. Viet Nam wound analysis. *J Trauma* 1978; **18**: 635 – 643.

Holcomb JB. Current perspective on combat casualty care. The 2004 Fitts Lecture. *J Trauma* 2005; **59**: 990 – 1002.

Holcomb JB, Stansbury LG, Champion HR, Wade C, Bellamy RF. Understanding combat casualty care statistics. *J Trauma* 2006; **60**: 397 – 401.

Jackson DS, Batty CG, Ryan JM, McGregor WSP. The Falklands war: army field surgical experience. *Ann R Coll Surg Engl* 1983; **65**: 281 – 285.

Jevtic M, Petrovic M, Ignjatovic D, Ilijevski N, Misovic S, Kronja G, Stankovic N. Treatment of wounded in the combat zone. *J Trauma* 1996; **40 (3Suppl.)**: S173 – S176.

Johnson DE, Panijayanond P, Lumjiak S, Crum JW, Boonkrapu P. Epidemiology of combat casualties in Thailand. *J Trauma* 1981; **21**: 486 – 488.

Jones EL, Peters AF, Gasior RM. Early management of battle casualties in Vietnam. *Arch Surg* 1968; **97**: 1 – 15.

Mabry RL, Holcomb JB, Baker AM, Cloonan CC, Uhorchak JM, Perkins DE, Canfield AJ, Hagmann JH. United States Army Rangers in Somalia: an analysis of combat casualties on an urban battlefield. *J Trauma* 2000; **49**: 515 – 529.

Spalding TJW, Stewart MPM, Tulloch DN, Stephens KM. Penetrating missile injuries in the Gulf war 1991. *Br J Surg* 1991; **78**: 1102 – 1104.

Van Rooyen MJ, Sloan EP, Radvany AE, Peric T, Kulis B, Tabak P. The incidence and outcome of penetrating and blunt trauma in central Bosnia: the Nova Bila Hospital for war wounded. *J Trauma* 1995; **38**: 863 – 866.

Versier G, Le Marec C, Rouffi J. Quatre ans de chirurgie de guerre au GMC de Sarajevo (juillet 1992 à août 1996) [Four years of war surgery at the French surgical facility in Sarajevo (July 1992 – August 1996)]. *Médecine et armées* 1998; **26**: 213 – 218.

Vojvodic V. Management of war casualties in the Military Medical Academy (Belgrade) during combat operations in 1991/1992: an overview. *J Trauma* 1996; **40 (3Suppl.)**: S180 – S182.

ICRC war surgery articles

Bowyer GW. Management of small fragment wounds: experience from the Afghan border. *J Trauma* 1996; **40 (3Suppl.)**: S170 – S172.

Coupland RM. Hand grenade injuries among civilians. *JAMA* 1993; **270**: 624 – 626.

Coupland RM. Epidemiological approach to surgical management of the casualties of war. *BMJ* 1994; **308**: 1693 – 1696.

Kjaergaard J. Les blessés de guerre de l’hôpital de campagne du CICR à Beyrouth en 1976 [War wounded in the ICRC field hospital in Beirut 1976]. *Schweiz Z Milit Med* 1978; **55**: 1 – 23.

Korver AJH. Outcome of war-injured patients treated at first aid posts of the International Committee of the Red Cross. *Injury* 1994; **25**: 25 – 30.

Morris D, Sugrue W, McKenzie E. On the border of Afghanistan with the International Committee of the Red Cross. *NZ Med J* 1985; **98**: 750 – 752.

Scott-Findlay J, Smith FS. A Timor experience in war and civilian trauma. *Med J Aust* 1976; **2**: 90 – 92.

Trouwborst A, Weber BK, Dufour D. Medical statistics of battlefield casualties. *Injury* 1987; **18**: 96 – 99.

Chapter 6

Hayward-Karlsson J, Jeffery S, Kerr A, Schmidt H. *Hospitals for War-Wounded: A Practical Guide for Setting up and Running a Surgical Hospital in an Area of Armed Conflict*. Geneva: ICRC; 1998.

Hayward-Karlsson J. Hospital and System Assessment. In: Mahoney PF, Ryan JM, Brooks AJ, Schwab CW, eds. *Ballistic Trauma: A Practical Guide 2nd Edition*. London: Springer-Verlag; 2005: 513 – 526.

Pruit BA. Combat casualty care and surgical progress. *Ann Surg* 2006; **243**: 715 – 729.

Chapter 7

Butler FK Jr. Tactical combat casualty care: combining good medicine with good tactics. *J Trauma* 2003; **54 (Suppl.)**: S2 – S3.

Coupland RM. Epidemiological approach to surgical management of the casualties of war. *BMJ* 1994; **308**: 1693 – 1696.

Coupland RM, Molde Å, Navein J. *Care in the Field for Victims of Weapons of War: A Report from the Workshop Organized by the ICRC on Pre-Hospital Care for War and Mine-Injured*. Geneva: ICRC; 2001.

Dubick MA, Atkins JL. Small-volume fluid resuscitation for the far-forward combat environment: current concepts. *J Trauma* 2003; **54 (Suppl.)**: S43 – S45.

Giannou C, Bernes E. *First Aid in Armed Conflicts and Other Situations of Violence*. Geneva: ICRC; 2006.

Guidelines on First Aid and HIV/AIDS. Geneva: International Committee of the Red Cross /International Federation of Red Cross and Red Crescent Societies; 2001.

Husum H. Effects of early prehospital life support to war injured: the battle of Jalalabad, Afghanistan. *Prehosp Disaster Med* 1999; **14**: 75 – 80.

Husum H, Gilbert M, Twisborg T. *Save Lives, Save Limbs*. Tromsoe, Norway: Third World Network, Tromsoe Mine Victim Resource Centre; 2000.

Husum H, Gilbert M, Wisborg T, Heng YV, Murad M. Rural prehospital trauma systems improve trauma outcome in low-income countries: a prospective study from North Iraq and Cambodia. *J Trauma* 2003; **54**: 1188 – 1196.

Korver AJH. Outcome of war-injured patients treated at first aid posts of the International Committee of the Red Cross. *Injury* 1994; **25**: 25 – 30.

Mabry R, McManus JG. Prehospital advances in the management of severe penetrating trauma. *Crit Care Med* 2008; **36 (Suppl.)**: S258 – S266.

Roberts DL. *Staying Alive: Safety and Security Guidelines for Humanitarian Volunteers in Conflict Areas 2nd Edition*. Geneva: ICRC; 2006.

Royal Defence Medical College and the Royal Centre for Defence Medicine. *Battlefield Advanced Trauma Life Support (Incorporating Battlefield Resuscitation Techniques and Skills)*. London: British Armed Forces, UK Minister of Defence; 2003.

Wisborg T, Murad MK, Edvardsen O, Husum H. Prehospital trauma system in a low-income country: system maturation and adaptation during 8 years. *J Trauma* 2008; **64**: 1342 – 1348.

Cervical spine immobilization

Arishita GI, Vayer JS, Bellamy RF. Cervical spine immobilization of penetrating neck wounds in a hostile environment. *J Trauma* 1989; **29**: 332 – 337.

Barkana Y, Stein M, Scope A, Maor R, Abramovich Y, Friedman Z, Knoller N. Prehospital stabilization of the cervical spine for penetrating injuries of the neck – is it necessary? *Injury* 2000; **31**: 305 – 309.

Rhee P, Kuncir EJ, Johnson L, Brown C, Velmahos G, Martin M, Wang D, Salim A, Doucet J, Kennedy S, Demetriades D. Cervical spine injury is highly dependent on the mechanism of injury following blunt and penetrating assault. *J Trauma* 2006; **61**: 1166 – 1170.

Tourniquet use

Brodie S, Hodgetts TJ, Ollerton J, McLeod J, Lambert P, Mahoney P. Tourniquet use in combat trauma. *J R Army Med Corps* 2008; **153**: 310 – 313.

Lakstein D, Blumenfeld A, Sokolov T, Lin G, Bssorai R, Lynn M, Ben-Abraham R. Tourniquets for hemorrhage control on the battlefield: a 4-year accumulated experience. *J Trauma* 2003; **54 (Suppl.)**: S221 – S225.

Navein J, Coupland R, Dunn R. The tourniquet controversy. *J Trauma* 2003; **54 (Suppl.)**: S219 – S220.

Welling DR, Burris DG, Hutton JE, Minken SL, Rich NM. A balanced approach to tourniquet use: lessons learned and relearned. *J Am Coll Surg* 2006; **203**: 106 – 115.

Chapter 8

Trauma training

Driscoll P, Skinner D, Earlam R, eds. *ABC of Major Trauma 3rd Edition*. London: BMJ Books; 2001.

American College of Surgeons, Committee on Trauma. *Advanced Trauma Life Support for Doctors, Student Course Manual (ATLS) 7th Edition*. Chicago, IL: American College of Surgeons; 2004.

Kortbeek JB, Al Turki SA, Ali J, et al. Advanced Trauma Life Support, 8th Edition. The evidence for change. *J Trauma* 2008; **64**: 1638 – 1650.

International Association for Trauma Surgery and Intensive Care (IATSIC). *Definitive Surgical Trauma Care Course (DSTC)*. Various country and regional sessions.

Trauma training in low-income countries: some examples

Aboutanos MB, Rodas EB, Aboutanos SZ, Mora FE, Wolf LG, Duane TM, Malhotra AK, Ivatury RR. Trauma education and care in the jungle of Ecuador, where there is no advanced trauma life support. *J Trauma* 2007; **62**: 714 – 719.

Basic Emergency Skills in Trauma (B.E.S.T.). Manila: Philippine College of Surgeons.

National Trauma Management Course (NTMC). Academy of Traumatology of India and Trauma Unit, Johannesburg Hospital and Faculty of Health Sciences, University of Witwatersrand. Available at: http://www.indiatrauma.org.

Emergency Room Trauma Course (ERTC). Geneva: ICRC.

Pneumothorax

Leigh-Smith S, Harris T. Tension pneumothorax – time for a re-think? *Emerg Med J* 2005; **22**: 8 – 16.

Shock resuscitation

Alam HB, Koustova E, Rhee P. Combat casualty care research: from bench to the battlefield. *World J Surg* 2005; **29 (Suppl.)**: S7 – S11.

Bickell WH, Wall MJ Jr, Pepe PE, Martin RR, Ginger VF, Allen MK, Mattox KL. Immediate versus delayed fluid resuscitation for hypotensive patients with penetrating torso injuries. *N Engl J Med* 1994; **331**: 1105 – 1109.

Champion HR. Combat fluid resuscitation: introduction and overview of conferences. *J Trauma* 2003; **54 (Suppl.)**: S7 – S12.

Moore FA, McKinley BA, Moore EE. The next generation in shock resuscitation. *Lancet* 2004; **363**: 1988 – 1996.

Rhee P, Alam HB, Ling GSF. Hemorrhagic Shock and Resuscitation. In: Tsokos GC, Atkins JL, eds. *Combat Medicine: Basic and Clinical Research in Military, Trauma, and Emergency Medicine*. Totowa, NJ: Humana Press; 2003: 177 – 218.

Rhee P, Koustova E, Alam HB. Searching for the optimal resuscitation method: recommendations for the initial fluid resuscitation of combat casualties. *J Trauma* 2003; **54 (Suppl.)**: S52 – S62.

Rushing GD, Britt LD. Reperfusion injury after hemorrhage. *Ann Surg* 2008; **247**: 929 – 937.

Shoemaker WC, Peitzman AB, Bellamy R, Bellomo R, Bruttig SP, Capone A, Dubick M, Kramer GC, McKenzie JE, Pepe PE, Safar P, Schlichtig R, Severinghaus JW, Tisherman SA, Wiklund L. Resuscitation from severe hemorrhage. [Symposium article] *Crit Care Med* 1996; **24 (Suppl.)**: S12 – S23.

Blood transfusion

Eshaya-Chauvin B, Coupland RM. Transfusion requirements for the management of war injured: the experience of the International Committee of the Red Cross. *Br J Anaesth* 1992; **68**: 221 – 223.

Kiebooms L, Rouvillois A, Jones T. *ICRC Blood Transfusion Guidelines*. Geneva: ICRC; 2004.

Kauvar DS, Holcomb JB, Norris GC, Hess JR. Fresh whole blood transfusion: a controversial military practice. *J Trauma* 2006; **61**: 181 – 184.

Key NS, Negrier C. Transfusion medicine 3: Coagulation factor concentrates: past, present, and future. *Lancet* 2007; **370**: 439 – 448.

Klein HG, Spahn DR, Carson JL. Transfusion medicine 1: Red blood cell transfusion in clinical practice. *Lancet* 2007; **370**: 415 – 426.

Repine TB, Perkins JG, Kauvar DS, Blackborne L. The use of fresh whole blood in massive transfusion. *J Trauma* 2006; **60 (6Suppl.)**: S59 – S69.

Spinella PC, Perkins JG, Grathwohl KW, Repine T, Beekley AC, Sebesta J, Jenkins D, Azarow K, Holcomb JB. Fresh whole blood transfusions in coalition military, foreign national, and enemy combatant patients during Operation Iraqi Freedom at a U.S. combat support hospital. *World J Surg* 2008; **32**: 2 – 6.

Stroncek DF, Rebulla P. Transfusion medicine 2: Platelet transfusions. *Lancet* 2007; **370**: 427 – 438.

World Health Organization. *WHO Guidelines for Blood Transfusion* [WHO Web site]. Available at: http://www.who.int/bloodsafety/clinical_use/en/Handbook_EN.pdf and http://www.who.int/bloodsafety/clinical_use/en/WHO_BLS_98.2_EN.pdf.

Chapter 9

Further practical information for the organisation of hospital management in a triage situation is available in: Hayward-Karlsson J, Jeffrey S, Kerr A, Schmidt H. *Hospitals for War Wounded*. Geneva: ICRC; 1998.

Please note:

In a military context, all standard war surgery manuals written by and for armed forces deal with the organization and implementation of triage under military constraints.

Almogy G, Belzberg H, Mintz Y, Pikarsky AK, Zamir G, Rivkind AI. Suicide bombing attacks: update and modifications to the protocol. *Ann Surg* 2004; **239**: 295 – 303.

Barbera JA, Macintyre AG. *Janes's Mass Casualty Handbook: Hospital Emergency Preparedness and Response*. Coulsdon, Surrey, UK: Jane's Information Group; 2003.

Burkle FM, Orebaugh S, Barendse BR. Emergency medicine in the Persian Gulf war – part 1: preparations for triage and combat casualty care. *Ann Emerg Med* 1994; **23**: 742 – 747.

Coupland RM, Parker PJ, Gray RC. Triage of war wounded: the experience of the International Committee of the Red Cross. *Injury* 1992; **23**: 507 – 510.

Frykberg ER. Medical management of disasters and mass casualties from terrorist bombings: how can we cope? *J Trauma* 2002; **53**: 201 – 212.

Frykberg ER. Principles of mass casualty management following terrorist disasters. [Editorial]. *Ann Surg* 2004; **239**: 319 – 321.

Gertsch P. Assessment of hospital workload in war surgery. *Br J Surg* 1987; **74**: 831 – 833.

Gray RC. War surgery and triage. *Postgrad Doctor Mid East*. 1993; **16**: 150 – 157.

Hirshberg A, Scott BG, Granchi T, Wall MJ Jr, Mattox KL, Stein M. How does casualty load affect trauma care in urban bombing incidents? A quantitative analysis. *J Trauma* 2005; **58**: 686 – 695.

Hogan DE, Lairet JR: Triage. In: Hogan DE, Burstein JL, eds. *Disaster Medicine 2nd Edition*. Philadephia, PA: Lippincott Williams & Wilkins; 2007: 12 – 28.

Iserson KV, Moskop JC: Triage in medicine, part I: concept, history, and types. *Ann Emerg Med* 2007; **49**: 275 – 281.

Iserson KV, Moskop JC: Triage in medicine, part II: underlying values and principles. *Ann Emerg Med* 2007; **49**: 282 – 287.

Kennedy K, Aghababian RV, Gans L, Lewis CP. Triage: techniques and applications in decision making. *Ann Emerg Med* 1996; **28**: 136 – 144.

Rignault D, Wherry D. Lessons from the past worth remembering: Larrey and triage. *J Trauma* 1999; **1**: 86 – 89.

Rochat C-H. Le triage chirurgical en conditions de guerre (Kaboul 1989) [Surgical triage under war conditions (Kabul 1989)]. *Médicine Militaire* 1992; **69**: 35 – 38.

Ryan JM, Sibson J, Howell G. Assessing injury severity during general war: will the military triage system meet future needs? *J R Army Med Corps* 1990; **136**: 27 – 35.

Suljevic I, Surkovic I. Medical aspects of the mass-scale civilian casualties at Sarajevo Markale Market on August 28, 1995: triage, resuscitation, and treatment. *Croat Med J* 2002; **43**: 209 – 212.

Torkki M, Koljonen V, Sillanpää K, Tukiainen E, Pyörälä S, Kemppainen E, Kalske J, Arajärvi E, Keränen U, Hirvensalo E. Triage in a bomb disaster with 166 casualties. *Eur J Trauma* 2006; **32**: 374 – 380.

World Medical Association. Statement on Medical Ethics in the Event of Disasters. Adopted by the 46th WMA General Assembly,1994 Sept.; Stockholm: Sweden. Available at: http://www.wma.net/e/policy/d7.htm.

Chapters 10 and 11

Anglen JO, Gainor BJ, Simpson WA, Christensen G. The use of detergent irrigation for musculoskeletal wounds. *Int Orthop* 2003; **27**: 40 – 46.

Baldan M, Giannou CP. Basic surgical management of war wounds: the ICRC experience. *East Cent Afr J Surg* 2003; **8**: 35 – 38.

Bewes P. The Management of Wounds in Developing Countries. Health Development e-TALC [CD – ROM]. January, 2004. No. 5. [TALC: Teaching-aids At Low Cost Web site]. Available at: http://www.talcuk.org.

Bowyer GW, Cooper GJ, Rice P. Small fragment wounds: biophysics and pathophysiology. *J Trauma* 1996; **40 (3Suppl.)**: S159 – S164.

Bowyer GW. Management of small fragment wounds in modern warfare: a return to Hunterian principles. *Ann R Coll Surg Engl* 1997; **79**: 175 – 182.

Coupland RM. Technical aspects of war wound excision. *Br J Surg* 1989; **76**: 663 – 667.

Fackler ML, Breteau JPL, Courbil LJ, Taxit R, Glas J, Fievet JP. Open wound drainage versus wound excision in treating the modern assault rifle wound. *Surgery* 1989; **105**: 576 – 584.

Trunkey DD. Comments on the article by Fackler et al. [Editorial] *Surgery* 1989; **105**: 693 – 694.

Gray RC. Surgery of war and disaster. *Trop Doct* 1991; **21 (Suppl.)**: S56 – S60.

Gray RC. *War Wounds: Basic Surgical Management*. Geneva: ICRC; 1994.

Hamer ML, Robson MC, Krizek TJ, Southwick WO. Quantitative bacterial analysis of comparative wound irrigations. *Ann Surg* 1975; **181**: 819 – 822.

Molan PC. The evidence supporting the use of honey as a wound dressing. *Int J Low Extrem Wounds* 2006; **5**: 40 – 54.

Molde Å. Victims of war: surgical principles must not be forgotten (again)! *Acta Orthop Scand* 1998; **281 (Suppl.)**: 54 – 57.

Rautio J, Paavolainen P. Delayed treatment of complicated fractures in war wounded. *Injury* 1987; **18**: 238 – 240.

Rautio J, Paavolainen P. Afghan war wounded: experience with 200 cases. *J Trauma* 1988; **28**: 523 – 525.

Rochat C-H, Graber P, Ursprung T. Traitement des plaies par projectiles en conditions de guerre (Afghanistan). Suture primaire différée ou suture secondaire? [Treatment of projectile wounds under conditions of war (Afghanistan). Delayed primary suture or secondary suture?]. *Médecine Militaire* 1986; **1**: 20 – 22.

Rowley DI. The management of war wounds involving bone. *J Bone Joint Surg Br* 1996; **78**: 706 – 709.

Strada G, Coupland RM, Gray RC. Surgery for the victims of war: the experience of the International Committee of the Red Cross. *J Emerg Surg* 1991; **14**: 126 – 130.

Svoboda SJ, Owens BD, Gooden HA, Melvin ML, Baer DG, Wenke JC. Irrigation with potable water versus normal saline in a contaminated musculoskeletal wound model. *J Trauma* 2008; **64**: 1357 – 1359.

Verbeke JH. Initial treatment of war casualties in a field hospital. *Acta Anaesth Belg* 1987; **38**: 261 – 265.

Vermeulen H, Ubbink DT, Goossens A, de Vos R, Legemate DA. Systematic review of dressings and topical agents for surgical wounds healing by secondary intention. *Br J Surg* 2005; **92**: 665 – 672.

Chapter 12

Bhaskar SN, Cutright DE, Hunsuck EE, Gross A. Pulsating water jet devices in debridement of combat wounds. *Mil Med* 1971; **136**: 264 – 266.

Bhatnagar MK, Smith GS. Trauma in the Afghan guerrilla war: effects of lack of access to care. *Surgery* 1989; **105**: 699 – 705.

Coupland RM, Howell P. An experience of war surgery and wounds presenting after 3 days on the border of Afghanistan. *Injury* 1988; **19**: 259 – 262.

Craig G. Treating the Afghan war wounded. *J Roy Soc Med* 1993; **86**: 404 – 405.

Gross A, Cutright DE, Bhaskar SN. Effectiveness of pulsating water jet lavage in treatment of contaminated crushed wounds. *Am J Surg* 1972; **124**: 373 – 377.

Mellor SG, Cooper GJ, Bowyer GW. Efficacy of delayed administration of benzylpenicillin in the control of infection in penetrating soft tissue injuries in war. *J Trauma* 1996; **40 (3Suppl.)**: S128 – S134.

Morris DS, Sugrue WJ, McKenzie E. On the border of Afghanistan with the International Committee of the Red Cross. *NZ Med J* 1985; **98**: 750 – 752.

Rowley DI. War surgery in an African conflict. *Scott Med J* 1997; **42**: 163 – 164.

Biofilm

Evans LV, ed. *Biofilms: Recent Advances in their Study and Control*. Amsterdam: Harwood Academic Press; 2000.

Fletcher M, ed. *Bacterial Adhesion: Molecular and Ecological Diversity*. New York: John Wiley & Sons; 1996.

Monroe D. Looking for chinks in the armor of bacterial biofilms [Public Library of Science Web site]. Available at: http://www.plos.org. *PLoS Bio*. 2007; **5**: e307.

Wilson M, ed. *Bacterial Adhesion to Host Tissues: Mechanisms and Consequences*. Cambridge: Cambridge University Press; 2002.

Chapter 13

Dahlgren B, Berlin R, Brandberg A, Rybeck B, Seeman T. Bacteriological findings in the first 12 hours following experimental missile trauma. *Acta Chir Scand* 1981; **147**: 513 – 518.

Dahlgren B, Berlin R, Brandberg A, Rybeck B, Schantz B, Seeman T. Effect of benzylpenicillin on wound infection rate and on the extent of devitalized tissue twelve hours after infliction of experimental missile trauma. *Acta Chir Scand* 1982; **148**: 107 – 112.

Fleming A. On the bacteriology of septic wounds. *Lancet* 1915; **186**: 638 – 643.

Lindberg RB, Wetzler TF, Marshall JD, Newton A, Strawitz JG, Howard JM. The bacterial flora of battle wounds at the time of primary debridement. *Ann Surg* 1955; **141**: 369 – 374.

Mellor SG, Cooper GJ, Bowyer GW. Efficacy of delayed administration of benzylpenicillin in the control of infection in penetrating soft tissue injuries in war. *J Trauma* 1996; **40 (3Suppl.)**: S128 – S134.

Miclau T, Farjo LA. The antibiotic treatment of gunshot wounds. *Injury* 1997; **28 (3Suppl.)**: C1 – C5.

Munoz-Price LS, Weinstein RA. Acinetobacter infection. *N Engl J Med* 2008; **358**: 1271 – 1281.

Murray CK, Roop SA, Hospenthal DR, Dooley DP, Wenner K, Hammock J, Taufen N, Gourdine E. Bacteriology of war wounds at the time of injury. *Mil Med* 2006; **171**: 826 – 829.

Murray CK, Hospenthal DR, eds. Prevention and management of combat-related infections: clinical practice guidelines consensus conference. *J Trauma* 2008; **64 (3Suppl.)**: S207 – S286.

Petersen K, Riddle MS, Danko JR, Blazes DL, Hayden R, Tasker SA, Dunne JR. Trauma-related infections in battlefield casualties from Iraq. *Ann Surg* 2007; **245**: 803 – 811.

Polhemus ME, Kester KE. Infections. In: Tsokos GC, Atkins JL, eds. *Combat Medicine: Basic and Clinical Research in Military, Trauma, and Emergency Medicine*. Totowa, NJ: Humana Press; 2003: 149 – 173.

Rubin RH. Surgical wound infection: epidemiology, pathogenesis, diagnosis and management. *BMC Infect Dis* 2006; **6**: 171. Available at: http://www.biomedcentral.com/1471-2334/6/171.

Sebeny PJ, Riddle MS, Petersen K. *Acinetobacter baumannii* skin and soft-tissue infection associated with war trauma. *Clin Infect Dis* 2008; **47**: 444 – 449.

Tian HM, Deng G, Huang MJ, Tian F, Suang G, Liu YQ. Quantitative bacteriological study of the wound track. *J. Trauma* 1988; **28 (Suppl.)**: S215 – S216.

Tian HM, Huang MJ, Liu YQ, Wang ZG. Primary bacterial contamination of wound track. *Acta Chir Scand* 1982; **508 (Suppl.)**: S265 – S269.

Simchen E, Sacks T. Infection in war wounds: experience during the 1973 October war in Israel. *Ann Surg* 1975; **182**: 754 –761.

Tetanus

Oladiran I, Meier DE, Ojelade AA, OlaOlorun DA, Adeniran A, Tarpley JL. Tetanus: continuing problem in the developing world. *World J Surg* 2002; **26**: 1282 – 1285.

Thwaites CL, Yen LM, Loan HT, Thuy TTD, Thwaites GE, Stepniewska K, Soni N, White NJ, Farrar JJ. Magnesium sulphate for treatment of severe tetanus: a randomised controlled trial. *Lancet* 2006; **368**: 1436 – 1443.

Necrotising fasciitis

Angoules AG, Kontakis G, Drakoulakis E, Vrentzos G, Granick MS, Giannoudis PV. Necrotising fasciitis of upper and lower limb: a systematic review. *Injury* 2007; **38 (Suppl.)**: C18 – C25.

Hasham S, Matteucci P, Stanley PRW, Hart NB. Necrotising fasciitis: clinical review. *BMJ* 2005; **330**; 830 – 833.

Chapter 14

Baldan M, Giannou CP, Sasin V, Morino GF. Metallic foreign bodies after war injuries: should we remove them? The ICRC experience. *East C Afr J Surg* 2004; **9**: 31 – 34.

Linden MA, Manton WI, Stewart RM, Thal ER, Feit H. Lead poisoning from retained bullets: pathogenesis, diagnosis, and management. *Ann Surg* 1982; **195**: 305 – 313.

Rhee JM, Martin R. The management of retained bullets in the limbs. *Injury* 1997; **28 (3Suppl.)**: C23 – C28.

Rich NM, Collins GJ, Andersen CA, McDonald PT, Kozloff L, Ricotta JJ. Missile emboli. *J Trauma* 1978; **18**: 236 – 239.

Chapter 15

Arturson G: Pathophysiology of the burn wound and pharmacological treatment. The Rudi Hermans Lecture, 1995. *Burns* 1996; **22**: 255 – 274.

Cartotto R, Musgrave M, Beveridge M, Fish J, Gomez M. Minimizing blood loss in burn surgery. *J Trauma* 2000; **49**: 1034 – 1039.

Hettiaratchy S, Dziewulski P: ABC of burns: pathophysiology and types of burns. *BMJ* 2004; **328**: 1427 – 1429.

Lindahl OA, Zdolsekz J, Sjöberg F, Ängquist K-A. Human postburn oedema measured with the impression method. *Burns* 1993; **19**: 479 – 484.

Lund T, Onarheim H, Reed RK. Pathogenesis of edema formation in burn injuries. *World J Surg* 1992; **16**: 2 – 9.

Pruit BA Jr. Fluid and electrolyte replacement in the burned patient. *Surg Clin N Am* 1978; **48**: 1291 – 1312.

Sheridan RL. Burns. *Crit Care Med* 2002; **30 (Suppl.)**: S500 – S514.

Thomas SJ, Kramer GC, Herndon DN. Burns: military options and tactical solutions. *J Trauma* 2003; **54 (5Suppl.)**: S207 – S218.

Zdolsek HJ, Lindahl OA, Ängquist K-A, Sjöberg F. Non-invasive assessment of intercompartmental fluids in burn victims. *Burns* 1998; **24**: 233 – 240.

Chapter 16

Britt LD, Dascombe WH, Rodriguez A. New horizons in management of hypothermia and frostbite injury. *Surg Clin North Am* 1991; **71**: 345 – 370.

Chapter 17

Bion JF. An anaesthetist in a camp for Cambodian refugees. *Anaesthesia* 1983; **38**: 798 – 801.

Bion JF. Infusion analgesia for acute war injuries: a comparison of pentazocine and ketamine. *Br J Acc Surg* 1984; **39**: 560 – 564.

Eshaya-Chauvin B, Nyffenegger E. Anesthésie pour blessés de guerre: étude rétrospective [Anaesthesia for war-wounded: retrospective study]. *Revue Médicale de la Suisse Romande*. 1990; **110**: 429 – 432.

Husum H, Heger T, Sundet M. Postinjury malaria: a study of trauma victims in Cambodia. *J Trauma* 2002; **52**: 259 – 266.

King M, ed. *Primary Anaesthesia*. Oxford: Oxford University Press; 1986.

Korver AJH. Relation between fever and outcome in injured victims of an internal armed conflict: the experience in a war surgery hospital of the International Committee of the Red Cross. *Milit Med* 1996; **161**: 658 – 660.

Leppäniemi AK. Where there is no anaesthetist. *Br J Surg* 1991; **78**: 245 – 246.

Pesonen P. Pulse oximetry during ketamine anaesthesia in war conditions. *Can J Anaesth* 1991; **38**: 592 – 594.

Vreede E, Lasalle X, Rosseel P. *Field Anaesthesia: Basic Practice: A Guide for Anaesthetists*. Paris: Médecins sans Frontières; 2001.

Ketamine anaesthesia in head trauma

Bourgoin A, Albanese J, Wereszczynski N, Charbit M, Vialet R, Martin C. Safety of sedation with ketamine in severe head injury patients: comparison with sulfentanyl. *Crit Care Med* 2003; **31**: 711 – 717.

Gofrit ON Leibovici D, Shemer J, Henig A, Shapira SC. Ketamine in the field: the use of ketamine for induction of anesthesia before intubation of injured patients in the field. *Injury* 1997; **28**: 41 – 43.

Green SM, Clem KJ, Rothrock SG. Ketamine safety profile in the developing world: survey of practitioners. *Acad Emerg Med* 1996; **3**: 598 – 604.

Himmelseher S, Durieux ME. Revising a dogma: ketamine for patients with neurological injury? *Anesth Analg* 2005; **101**: 524 – 534.

Ketcham DW. Where there is no anaesthesiologist: the many uses of ketamine. *Trop Doct* 1990; **20**: 163 – 166.

Sehdev RS, Symmons DAD, Kindl K. Ketamine for rapid sequence induction in patients with head injury in the emergency department. *Emerg Med Austr* 2006; **18**: 37 – 44.

Tighe SQM, Rudland S. Anesthesia in northern Iraq: an audit from a field hospital. *Mil Med* 1994; **159**: 86 – 90.

Trouwborst A, Weber BK, Dufour D. Medical statistics of battlefield casualties. *Injury* 1987; **18**: 96 – 99.

Chapter 18

Damage control surgery

Burch JM, Ortiz VB, Richardson RJ, Martin RR, Mattox KL, Jordan GL Jr. Abbreviated laparotomy and planned reoperation for critically injured patients. *Ann Surg* 1992; **215**: 476 – 483.

Damage Control Surgery. *Surg Clin North Am* 1997; **77**: 753 – 952.

Hirshberg A, Mattox KL. Planned reoperation for severe trauma. *Ann Surg* 1995; **222**: 3 – 8.

Kashuk JL, Moore EE, Millikan JS, Moore JB. Major abdominal vascular trauma: a unified approach. *J Trauma* 1982; **22**: 672 – 679.

Moore EE. Staged laparotomy for the hypothermia, acidosis, coagulopathy syndrome. *Am J Surg* 1996; **172**: 405 – 410.

Parker PJ. Damage control surgery and casualty evacuation: techniques for surgeons, lessons for military medical planners. *J R Army Med Corps* 2006; **152**: 202 – 211.

Rotondo MF, Schwab CW, McGonigal MD, Phillips GR, Fruchterman TM, Kauder DR, Latenser BA, Angood PB: "Damage Control": an approach for improved survival with exsanguinating penetrating abdominal injury *J Trauma* 1993; **35**: 375 – 382.

Sharpiro MB, Jenkins DH, Schwab CW, Rotondo MF. Damage control: collective review. *J Trauma* 2000; **49**: 969 – 978.

Stone HH, Strom PR, Mullins RJ. Management of the major coagulopathy with onset during laparotomy. *Ann Surg* 1983; **197**: 532 – 535.

Hypothermia and coagulopathy

Bernabei AF, Levison MA, Bender JS. The effects of hypothermia and injury severity on blood loss during trauma laparotomy. *J Trauma* 1992; **33**: 835 – 839.

Brohi K, Singh J, Heron M, Coats T. Acute traumatic coagulopathy. *J Trauma* 2003; **54**: 1127 – 1130.

Brohi K, Cohen MJ, Ganter MT, Matthay MA, Mackersie RC, Pittet J-F. Acute traumatic coagulopathy: initiated by hypoperfusion: modulated through the protein C pathway? *Ann Surg* 2007; **245**: 812 – 818.

Brohi K, Cohen MJ, Ganter MT, Schultz MJ, Levi M, Mackersie RC, Pittet J-F. Acute coagulopathy of trauma: hypoperfusion induces systemic anticoagulation and hyperfibrinolysis. *J Trauma* 2008; **64**: 1211 – 1217.

Cosgriff N, Moore EE, Sauaia A, Kenny-Moynihan M, Burch JM, Galloway B. Predicting life-threatening coagulopathy in the massively transfused trauma patient: hypothermia and acidosis revisited. *J Trauma* 1997; **42**: 857 – 862.

Gentilello LM, Jurkovich GJ, Stark MS, Hassantash SA, O'Keefe GE. Is hypothermia in the victim of major trauma protective or harmful? A randomized, prospective study. *Ann Surg* 1997; **226**: 439 – 449.

Gregory JS, Flanebaum L, Townsend MC, Cloutier CT, Jonasson O. Incidence and timing of hypothermia in trauma patients undergoing operations. *J Trauma* 1991; **31**: 795 – 800.

Grosso SM, Keenan JO. Whole blood transfusion for exsanguinating coagulopathy in a U.S. field surgical hospital in postwar Kosovo. *J Trauma* 2000; **49**: 145 – 148.

Gubler KD, Gentilello LM, Hassantash SA, Maier RV. The impact of hypothermia on dilutional coagulopathy. *J Trauma* 1994; **36**: 847 – 851.

Hess JR, Lawson JH. The coagulopathy of trauma versus disseminated intravascular coagulation. *J Trauma* 2006; **60 (6Suppl.)**: S12 – S19.

Holcomb JB, Jenkins D, Rhee P, Johannigman J, Mahoney P, Mehta S, Cox ED, Gehrke MJ, Beilman GJ, Schreiber M, Flaherty SF, Grathwohl KW, Spinella PC, Perkins JG, Beekley AC, McMullin NR, Park MS, Gonzalez EA, Wade CE, Dubick MA, Schwab CW, Moore FA, Champion HR, Hoyt DB, Hess JR. Damage control resuscitation: directly addressing the early coagulopathy of trauma. *J Trauma* 2007; **62**: 307 – 310.

Jurkovich GJ, Greiser WB, Luterman A, Curreri PW. Hypothermia in trauma victims: an ominous predictor of survival. *J Trauma* 1987; **27**: 1019 – 1024.

Kirkman E, Watts S, Hodgetts T, Mahoney P, Rawlinson S, Midwinter M. A proactive approach to the coagulopathy of trauma: the rationale and guidelines for treatment. *J R Army Med Corps* 2008; **153**: 302 – 306.

Kirkpatrick AW, Chun R, Brown R, Simons RK. Hypothermia and the trauma patient. *Can J Surg* 1999; **42**: 333 – 343.

Luna GK, Maier RV, Pavlin EG, Anardi D, Copass MK, Oreskovich MR. Incidence and effect of hypothermia in seriously injured patients. *J Trauma* 1987; **27**: 1014 – 1018.

MacLeod JB, Lynn M, McKenney MG, Cohn SM, Murtha M. Early coagulopathy predicts mortality in trauma. *J Trauma* 2003; **55**: 39 – 44.

Niles SE, McLaughlin DF, Perkins JG, Wade CE, Li Y, Spinella PC, Holcomb JB. Increased mortality associated with the early coagulopathy of trauma in combat casualties. *J Trauma* 2008; **64**: 1459 – 1465.

Seekamp A, van Griensven M, Hildebrandt F, Wahlers T, Tscherne H. Adenosine-triphosphate in trauma-related and elective hypothermia. *J Trauma* 1999; **47**: 673 – 683.

Shafi S, Elliott AC, Gentilello L. Is hypothermia simply a marker of shock and injury severity or an independent risk factor for mortality in trauma patients? Analysis of a large national trauma registry. *J Trauma* 2005; **59**: 1081 – 1085.

Tisherman SA. Hypothermia and injury. *Curr Opin Crit Care* 2004; **10**: 512 – 519.

Watts DD, Trask A, Soeken K, Perdue P, Dols S, Kaufmann C. Hypothermic coagulopathy in trauma: effect of varying levels of hypothermia on enzyme speed, platelet function, and fibrinolytic activity. *J Trauma* 1998; **44**: 846 – 854.

MISSION

The International Committee of the Red Cross (ICRC) is an impartial, neutral and independent organization whose exclusively humanitarian mission is to protect the lives and dignity of victims of armed conflict and other situations of violence and to provide them with assistance.

The ICRC also endeavours to prevent suffering by promoting and strengthening humanitarian law and universal humanitarian principles.

Established in 1863, the ICRC is at the origin of the Geneva Conventions and the International Red Cross and Red Crescent Movement. It directs and coordinates the international activities conducted by the Movement in armed conflicts and other situations of violence.

ICRC

VOLUME 1

WAR SURGERY

WORKING WITH LIMITED RESOURCES IN ARMED CONFLICT AND OTHER SITUATIONS OF VIOLENCE

0973/002 03.2009 3,000

ICRC

WAR SURGERY

WORKING WITH LIMITED RESOURCES IN ARMED CONFLICT AND OTHER SITUATIONS OF VIOLENCE

VOLUME 1

ICRC

REFERENCE

S Devaji Rao

GENERAL SURGICAL EMERGENCIES

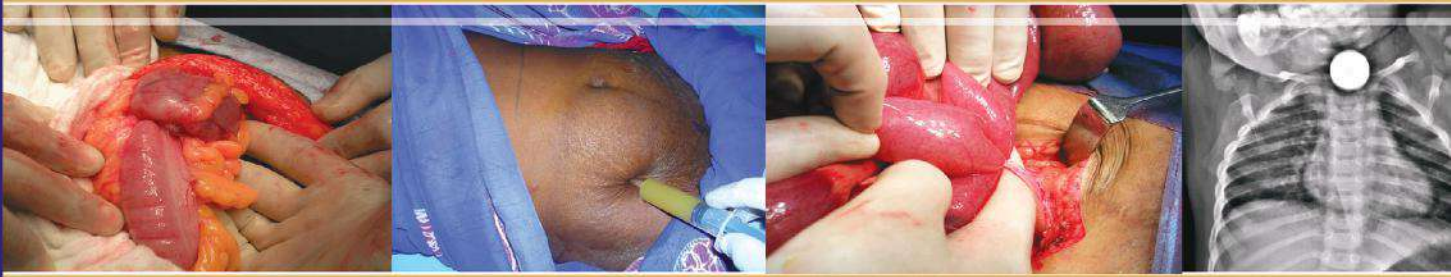

JAYPEE

General Surgical Emergencies

General Surgical Emergencies

S Devaji Rao MS MNAMS FICS
Teaching Faculty
National Board of Examinations
Senior Consultant in General Surgery
Surgical Gastroenterology and Surgical Oncology
St Isabel's Hospital, Mylapore, Chennai, Tamil Nadu, India
Chennai Meenakshi Multispecialty Hospital Ltd, Chennai, Tamil Nadu, India

JAYPEE BROTHERS MEDICAL PUBLISHERS (P) LTD.
New Delhi • Panama City • London • Dhaka • Kathmandu

Jaypee Brothers Medical Publishers (P) Ltd.

Headquarters

Jaypee Brothers Medical Publishers (P) Ltd.
4838/24, Ansari Road, Daryaganj
New Delhi 110 002, India
Phone: +91-11-43574357
Fax: +91-11-43574314
Email: jaypee@jaypeebrothers.com

Overseas Offices

J.P. Medical Ltd.
83 Victoria Street, London
SW1H 0HW (UK)
Phone: +44-2031708910
Fax: +02-03-0086180
Email: info@jpmedpub.com

Jaypee Brothers Medical Publishers (P) Ltd.
17/1-B Babar Road, Block-B, Shaymali
Mohammadpur, Dhaka-1207
Bangladesh
Mobile: +08801912003485
Email: jaypeedhaka@gmail.com

Jaypee-Highlights Medical Publishers Inc.
City of Knowledge, Bld. 237, Clayton
Panama City, Panama
Phone: +507-301-0496
Fax: +507-301-0499
Email: cservice@jphmedical.com

Jaypee Brothers Medical Publishers (P) Ltd.
Shorakhute, Kathmandu
Nepal
Phone: +00977-9841528578
Email: jaypee.nepal@gmail.com

Website: www.jaypeebrothers.com
Website: www.jaypeedigital.com

© 2012, Jaypee Brothers Medical Publishers

All rights reserved. No part of this book may be reproduced in any form or by any means without the prior permission of the publisher.

Inquiries for bulk sales may be solicited at: jaypee@jaypeebrothers.com

This book has been published in good faith that the contents provided by the author contained herein are original, and is intended for educational purposes only. While every effort is made to ensure accuracy of information, the publisher and the author specifically disclaim any damage, liability, or loss incurred, directly or indirectly, from the use or application of any of the contents of this work. If not specifically stated, all figures and tables are courtesy of the author. Where appropriate, the readers should consult with a specialist or contact the manufacturer of the drug or device.

General Surgical Emergencies

First Edition : 2012

ISBN 978-93-5025-961-0

Printed at

Dedicated to

My teachers

My parents
Uma Bai and Siva Rao

My beloved wife
Kalpana

My daughters
Bhavna and Kirthana

Preface

Many textbooks are available dealing with the details of emergency surgical conditions. Not many are available dealing exclusively with the emergency conditions, especially those related to surgery. Not that every patient coming to the casualty at odd hours with an acute symptom has a life-threatening emergency, but it becomes necessary to comfort the patient at the first instant, followed by a quick diagnosis of the clinical condition. It is also imperative that the impending emergency should be identified by the clinician, so that a catastrophe is avoided. The classical example of such confusing situation is the acute pancreatitis. This can present with a variety of symptoms which will not give a lead towards diagnosis, but the whole clinical picture establishes after some time, and takes the patient through a very tough morbid situation, and may also end in death.

This concise book has been designed in such a way that the common emergency conditions are detailed to the extent so that surgical students or the casualty doctors or young surgeons may not miss the diagnosis. It cannot replace the many well-established textbooks, but it gives sufficient information for a clinician to manage the emergency.

Color photographs are useful add-ons to these chapters, which will make the reader remember the information for a long time.

Every new doctor who is resident in the casualty and the intensive care unit will face with a variety of clinical problems, and I hope the handbook will come handy.

I also hope that I have hit the required details at the right level for the young surgeons.

S Devaji Rao

Acknowledgments

I sincerely thank Professor SM Balaji, Professor J Cornelius, Dr Gayathri, Dr M Kanagavel, Dr Kirthana Rao, Professor N Mohan, Dr R Narasimhan, Professor MG Rajamanickam, Professor R Rajaraman, Professor N Sekar, Dr K Sridhar, Professor V Srinivasan, Dr V Thulasiraman, Dr Usha Dorairajan, Professor PS Venkatesh Rao and Dr S Vijayaraghavan, who contributed their clinical photographs, from their personal collections. My special thanks to my friends Dr Mani Veeraraghavan for the endoscopy photographs and Dr V Ganesan for the ultrasound photographs. Most of the CT and MRI pictures were procured from Bharat Scans, Chennai, Tamil Nadu, India, and I sincerely thank Dr R Emmanuel for providing them.

My special thanks to my daughters Dr Bhavna for going through the script and making corrections, and Dr Kirthana for drawing the diagrams.

My very special thanks goes to my wife Kalpana for the tolerance during the preparation of this manual.

Contents

SECTION ONE: INTRODUCTION

1. Introduction **3**

SECTION TWO: ASSESSMENT

2. Assessment of Surgical Emergencies **7**

- ❖ Make the Patient Lie Down Comfortably 7
- ❖ Elicit a Quick History 8
  - Make a Thorough Clinical Examination 8
  - Come to a Quick Working Clinical Diagnosis 10
  - Ask for Essential Meaningful Investigations 10
  - Admit the Patient where Thought to be Required 11
  - Put the Patient in Intensive Care Whenever Required 11
  - Collect Investigation Reports at the Earliest 11
  - Start Emergency Treatment 11
  - Ask for Expert Opinions 12

SECTION THREE: CRITICAL CARE

3. Critically Ill Patient and Critical Care **15**

- ❖ Cardiac Support 15
- ❖ Respiratory Support 15
- ❖ Inotropic Support 16

4. Shock **17**

- ❖ Hypovolemic Shock 17
  - Treatment of Hypovolemic Shock 18
- ❖ Septic Shock 18
  - Symptoms and Signs 18
  - Pathogenesis 18
  - Treatment 19
- ❖ Anaphylactic Shock 19
  - Incidence and Etiology 19
  - Pathogenesis 19
  - Symptoms and Signs 19
  - Treatment 20
- ❖ Cardiogenic Shock 20
- ❖ Neurogenic Shock 20
  - Pathogenesis 20
  - Treatment 20

5. Acute Respiratory Distress Syndrome **21**

  - Incidence and Etiology 21
  - Pathogenesis 21
  - Clinical Presentation 21
  - Relevant Investigations 21
  - Treatment 22

SECTION FOUR: TRAUMA

6. Polytrauma **25**

- ❖ Introduction 25
- ❖ Preparation 25
  - Prehospital Communication 25

- ❖ Trauma Reception Team 26
  - Receiving the Patient 27
  - Primary Survey and Resuscitation 27
  - Airway and Cervical Spine Control 27
  - Breathing 28
  - Circulation and Hemorrhage Control 29
  - Recognition and Assessment of Hypovolemia 29
  - Fluid Resuscitation 30
  - Disability 30
  - Exposure 30
- ❖ Secondary Survey 30
- ❖ Clinical Evaluation 31
  - Eliciting History 31
  - Physical Examination 31
  - Scalp 31
  - Neurological State 31
  - Base of Skull 31
  - Neck 32
  - Eyes 32
  - Face 32
  - Thorax 32
  - Abdomen 33
  - Extremities 33
- ❖ Definitive Care 33

7. Head Injuries 34
- ❖ Brain Injury 34
- ❖ Fractures of Skull 36
  - Symptoms and Signs 36
- ❖ CSF Fistulae 39
- ❖ Scalp Injuries 39

8. Facial Injuries 41
- ❖ Fractures of Midfacial Skeleton 41
- ❖ Fractures of Mandible 42
  - Clinical Presentation 43
  - Relevant Investigations 45
  - Treatment 45

9. Spine and Spinal Cord Injuries 46
- ❖ Injuries to Bony and Ligamentous Spine 46
  - Incidence and Etiology 46
  - Symptoms 46
  - Signs 47
  - Relevant Investigations 47
  - Treatment 48
- ❖ Injuries of the Spinal Cord 48
  - Incidence and Etiology 48
  - Symptoms and Signs 48
  - Relevant Investigations 49
  - Treatment 49

10. Thoracic Injuries 50
- ❖ Rib Fractures 50
  - Incidence and Etiology 50
  - Symptoms 50
  - Signs 51
  - Relevant Investigations 51
  - Treatment 51
- ❖ Flail Chest 53
  - Incidence and Etiology 53
  - Symptom 54
  - Sign 54
  - Relevant Investigations 54
  - Treatment 54
- ❖ Sternal Fracture 54
  - Incidence and Etiology 54
  - Symptom 54
  - Sign 55
  - Relevant Investigations 55
  - Treatment 55
- ❖ Pneumothorax 55
  - Incidence and Etiology 55

  - Symptoms 56
  - Signs 56
  - Relevant Investigations 56
  - Treatment 58
- ❖ Surgical Emphysema 58
  - Incidence and Etiology 58
  - Pathogenesis 59
  - Symptoms and Signs 59
  - Relevant Investigations 59
  - Treatment 59
- ❖ Hemothorax 59
  - Incidence and Etiology 59
  - Symptoms 60
  - Signs 60
  - Relevant Investigations 60
  - Treatment 61
- ❖ Pulmonary Contusion/laceration 61
  - Incidence and Etiology 61
  - Symptoms and Signs 61
  - Relevant Investigations 61
  - Treatment 62
- ❖ Injuries of Thoracic Aorta 62
  - Incidence and Etiology 62
  - Symptom 63
  - Signs 63
  - Relevant Investigations 63
  - Treatment 64
- ❖ Injuries of Myocardium 64
  - Incidence and Etiology 64
  - Symptoms 64
  - Signs 64
  - Relevant Investigations 65
  - Treatment 65
- ❖ Esophageal Injuries 65
  - Incidence and Etiology 65
  - Symptoms 65
  - Signs 65
  - Relevant Investigations 65
  - Treatment 66
- ❖ Diaphragmatic Injuries 66
  - Incidence and Etiology 66
  - Symptoms and Signs 67
  - Relevant Investigations 67
  - Treatment 67

11. Abdominal Injuries 68
- ❖ Closed Injuries 68
- ❖ Open Injuries 69
  - Symptoms and Signs 69
- ❖ Injuries of Liver 69
  - Incidence and Etiology 69
  - Symptoms and Signs 70
  - Relevant Investigations 70
  - Treatment 70
- ❖ Injuries of Spleen 72
  - Incidence and Etiology 72
  - Symptoms 72
  - Signs 72
  - Relevant Investigations 73
  - Treatment 74
- ❖ Injuries of Mesentery 74
  - Incidence and Etiology 74
  - Symptoms 74
  - Sign 75
  - Relevant Investigations 75
  - Treatment 75
- ❖ Injuries of Duodenum 75
  - Incidence and Etiology 75
  - Symptoms and Signs 75
  - Relevant Investigations 75
  - Treatment 76
- ❖ Injuries of Small Intestine 76
  - Incidence and Etiology 76

    - Clinical Presentation 76
    - Relevant Investigations 76
    - Treatment 77
- ❖ Injuries of Large Intestine 77
    - Incidence and Etiology 77
    - Symptoms 77
    - Signs 77
    - Relevant Investigations 77
    - Treatment 77
- ❖ Abdominal Compartment Syndrome 78
    - Incidence and Etiology 78
    - Pathogenesis 78
    - Symptom 78
    - Signs 78
    - Relevant Investigations 78
    - Treatment 78

12. Urological Injuries 79

- ❖ Renal Injuries 79
    - Incidence and Etiology 79
    - Symptom 79
    - Signs 79
    - Relevant Investigations 80
    - Treatment 80
- ❖ Ureteric Injuries 80
    - Incidence and Etiology 80
    - Symptom and Sign 81
    - Relevant Investigation 82
    - Treatment 82
- ❖ Urinary Bladder Injuries 82
    - Incidence and Etiology 82
    - Symptoms 83
    - Signs 83
    - Relevant Investigations 83
    - Treatment 84
- ❖ Urethral Injuries 84
    - Incidence and Etiology 84
    - Symptom 84
    - Sign 84
    - Relevant Investigation 84
    - Treatment 85

13. Male Genital Injuries 86

- ❖ Injuries of Penile Skin 86
    - Incidence and Etiology 86
    - Symptom 86
    - Sign 86
    - Relevant Investigation 86
    - Treatment 86
- ❖ Foreign Bodies in Penile Urethra 87
    - Incidence and Etiology 87
    - Symptoms and Signs 87
    - Relevant Investigation 88
    - Treatment 88
- ❖ Scrotal Laceration 88
    - Incidence and Etiology 88
    - Symptom 89
    - Signs 89
    - Relevant Investigation 89
    - Treatment 89

14. Female Genital Injuries 90

- ❖ Coital Injuries 90
    - Incidence and Etiology 90
    - Symptom 90
    - Signs 90
    - Relevant Investigations 90
    - Treatment 90
- ❖ Injuries during Childbirth 91
    - Incidence and Etiology 91
    - Clinical Presentation 91
    - Relevant Investigations 91
    - Treatment 91

15. Hand Injuries 92

    - Incidence and Etiology 92

    - Clinical Presentation 92
    - Relevant Investigations 96
    - Treatment 96

SECTION FIVE: ORGANS AND SYSTEMS

16. Oral Cavity 99

- Acute Odontogenic Infections 99
    - Incidence and Etiology 99
    - Pathogenesis 99
    - Symptoms and Signs 100
    - Relevant Investigations 100
    - Treatment 100

17. Head and Neck 101

- Acute Suppurative Sialadenitis 101
    - Incidence and Etiology 101
    - Clinical Presentation 101
    - Relevant Investigation 103
    - Treatment 103
- Ludwig's Angina 103
    - Incidence and Etiology 103
    - Pathogenesis 103
    - Symptoms 103
    - Signs 103
    - Relevant Investigation 103
    - Treatment 104

18. Thorax 105

- Acute Mediastinitis 105
    - Incidence and Etiology 105
    - Clinical Presentation 105
    - Relevant Investigations 105
    - Treatment 106
- Acute Pleuritis 106
    - Incidence and Etiology 106
    - Symptom 106
    - Sign 106
    - Relevant Investigations 107
    - Treatment 107
- Pleural Effusion 107
    - Incidence and Etiology 107
    - Symptom 107
    - Signs 107
    - Relevant Investigations 108
    - Treatment 109
- Acute Empyema Thoracis 109
    - Incidence and Etiology 109
    - Symptom 110
    - Signs 110
    - Relevant Investigations 110
    - Treatment 110
- Spontaneous Pneumothorax 110
    - Incidence and Etiology 110
    - Symptom 110
    - Sign 111
    - Relevant Investigations 111
    - Treatment 111
- Foreign Bodies in the Respiratory Tract 112
    - Incidence and Etiology 112
    - Pathogenesis 112
    - Symptoms 112
    - Signs 112
    - Relevant Investigations 112
    - Treatment 113
- Suppurative or Aspiration Pneumonia 113
    - Incidence and Etiology 113
    - Pathogenesis 113
    - Symptoms 114
    - Sign 114
    - Relevant Investigation 114
    - Treatment 114
- Hospital Acquired Pneumonia 115
    - Incidence and Etiology 115

    - Symptoms 115
    - Signs 115
    - Relevant Investigations 115
    - Treatment 115
- ■ Acute Lung Abscess 115
    - Incidence and Etiology 115
    - Symptoms 116
    - Sign 116
    - Relevant Investigations 116
    - Treatment 116
- ■ Pulmonary Embolism 116
    - Incidence and Etiology 116
    - Symptoms and Signs 116
    - Relevant Investigations 117
    - Treatment 117

19. Breast 118

- ■ Breast Hematoma 118
    - Incidence and Etiology 118
    - Symptoms 118
    - Sign 118
    - Relevant Investigations 118
    - Treatment 118
- ■ Acute Breast Abscess 119
    - Incidence and Etiology 119
    - Symptoms 119
    - Signs 119
    - Relevant Investigations 119
    - Treatment 119

20. Spine 120

- ■ Degenerative Diseases of Disk and Facet Joints 120
    - Incidence and Etiology 120
    - Symptoms and Signs 120
    - Relevant Investigations 121
    - Treatment 121
- ■ Spondylolisthesis 122
    - Incidence and Etiology 122
    - Symptom 122
    - Signs 122
    - Relevant Investigations 123
    - Treatment 123

21. Gastroenterology 124

- ■ Acute Abdomen 124
    - Introduction 124
    - Pain 125
    - Vomiting 128
  - ❖ Acute Upper Abdominal Pain 130
    - Abdomen 131
    - Relevant Investigations 133
    - Treatment Plan 134
  - ❖ Acute Lower Abdominal Pain 134
    - Treatment Plan 138
- ■ Acute Liver Abscess 138
    - Incidence and Etiology 138
    - Pathogenesis 138
    - Symptoms 138
    - Signs 139
    - Differential Diagnosis 139
    - Relevant Investigations 139
    - Treatment 140
- ■ Acute Cholecystitis 140
    - Incidence and Etiology 140
    - Pathogenesis 140
    - Complications 140
    - Symptoms 140
    - Signs 141
    - Differential Diagnosis 141
    - Relevant Investigations 141
    - Treatment 142
- ■ Acute Cholangitis 143
    - Incidence and Etiology 143

  - Pathogenesis 143
  - Symptoms 143
  - Signs 144
  - Relevant Investigations 144
  - Treatment 146
- Primary Sclerosing Cholangitis 146
  - Incidence and Etiology 146
  - Clinical Presentation 146
  - Relevant Investigations 146
  - Treatment 147
- Splenic Abscess 148
  - Incidence and Etiology 148
  - Pathogenesis 148
  - Symptoms 148
  - Sign 148
  - Relevant Investigation 148
  - Treatment 149
- Acute Hemorrhagic Pancreatitis 149
  - Incidence and Etiology 149
  - Pathogenesis 150
  - Symptoms 150
  - Signs 150
  - Relevant Investigations 151
  - Treatment 153
- Acute Appendicitis 154
  - Incidence and Etiology 154
  - Pathogenesis 154
  - Symptoms 154
  - Signs 155
  - Relevant Investigations 157
  - Treatment 159
- Acute Mesenteric Lymphadenitis 159
  - Incidence and Etiology 159
  - Pathogenesis 159
  - Symptoms 159
  - Signs 159
  - Relevant Investigations 159
  - Treatment 160
- Acute Colonic Diverticulitis 160
  - Incidence and Etiology 160
  - Pathogenesis 160
  - Symptoms 161
  - Signs 161
  - Relevant Investigations 161
  - Treatment 162
- Acute Meckel's Diverticulitis 162
  - Incidence and Etiology 162
  - Symptom 163
  - Sign 163
  - Differential Diagnosis 163
  - Relevant Investigations 163
  - Treatment 164
- Acute Solitary Cecal Diverticulitis 164
  - Incidence and Etiology 164
  - Symptom 164
  - Sign 164
  - Relevant Investigations 164
  - Treatment 164
- Acute Ulcerative Colitis 164
  - Incidence and Etiology 164
  - Symptoms 165
  - Signs 165
  - Relevant Investigations 165
  - Treatment 165
- Acute Intestinal Obstruction 166
  - Symptoms 166
  - Signs 167
  - Relevant Investigations 167
  - Treatment 168
- Gallstone Ileus 168
  - Incidence and Etiology 168
  - Symptoms 169

  - Signs 169
  - Relevant Investigations 169
  - Treatment 169
- Acute Intussusception 170
  - Incidence and Etiology 170
  - Symptoms 171
  - Signs 172
  - Relevant Investigations 172
  - Treatment 173
- Swallowed Foreign Bodies 173
  - Incidence and Etiology 173
  - Symptoms 173
  - Signs 173
  - Relevant Investigations 173
  - Treatment 173
- Perforated Bowel Pathologies 174
  - Differential Diagnosis 175
  - Relevant Investigations 175
  - Treatment 177
- Intestinal Strictures 177
  - Incidence and Etiology 177
  - Symptoms 177
  - Signs 177
  - Relevant Investigations 177
  - Treatment 177
- Bands and Adhesions 178
  - Incidence and Etiology 178
  - Symptoms and Signs 178
  - Relevant Investigations 178
  - Treatment 178
- Enteroliths/Food Bolus 179
  - Incidence and Etiology 179
  - Symptoms and Signs 179
  - Relevant Investigations 179
  - Treatment 179
- Volvulus 179
  - ❖ Sigmoid Volvulus 179
    - Incidence and Etiology 179
    - Pathogenesis 179
    - Symptoms 179
    - Signs 180
    - Relevant Investigations 180
    - Treatment 180
  - ❖ Cecal Volvulus 180
    - Incidence and Etiology 180
    - Clinical Presentation 181
    - Relevant Investigations 181
    - Treatment 181
  - ❖ Midgut Volvulus 181
    - Incidence and Etiology 181
    - Symptom 181
    - Signs 181
    - Relevant Investigation 181
    - Treatment 181
  - ❖ Gastric Volvulus 182
    - Incidence and Etiology 182
    - Symptoms 182
    - Sign 182
    - Relevant Investigation 182
    - Treatment 182
- Intestinal Obstruction due to Herniae (Internal and External) 182
  - Incidence and Etiology 182
  - Symptoms 183
  - Sign 183
  - Relevant Investigation 183
  - Treatment 183
- Paralytic Ileus 183
  - Incidence and Etiology 183
  - Symptoms 184

- Sign 184
- Relevant Investigations 184
- Treatment 185

■ Torsion of Mesenteric Cyst 185
- Incidence and Etiology 185
- Pathogenesis 186
- Symptom 186
- Signs 186
- Relevant Investigation 186
- Treatment 186

■ Torsion of Omentum 186
- Incidence and Etiology 186
- Pathogenesis 186
- Symptoms 187
- Signs 187
- Relevant Investigation 187
- Treatment 187

■ Colics 187
- Incidence and Etiology 187
- Clinical Features 188
- Symptoms 188
- Signs 189
- Differential Diagnosis by Clinical History and Examination 190
- Relevant Investigations 190
- Radiology 190
- Treatment Plan 190

■ Gastrointestinal Hemorrhage 190
- Definitions 190
- Types of Gastrointestinal Hemorrhages 191
- Hematemesis and Melena 192
- Melena 192
- Hematochezia 192
- Eliciting History 193
- Hematochezia 193
- Past History 194
- Family History 194
- Clinical Examination 194
- Abdomen 195
- Differential Diagnosis by Clinical History and Examination 195
- Relevant Investigations 196
- Treatment 198
- Surgical Treatment 198

22. Anorectum 199

■ Acute Anal Fissure 199
- Incidence and Etiology 199
- Pathogenesis 199
- Symptoms 199
- Signs 200
- Relevant Investigation 201
- Treatment 201

■ Anorectal Abscess 201
- Incidence and Etiology 201
- Pathogenesis 202
- Symptoms 202
- Signs 202
- Relevant Investigation 203
- Treatment 203

■ Hemorrhoids 203
- Incidence and Etiology 203
- Symptoms 203
- Signs 204
- Relevant Investigation 205
- Treatment 205

■ Perianal Hematoma 206
- Incidence and Etiology 206
- Pathogenesis 206
- Symptoms 206
- Sign 206

  - Relevant Investigation 207
  - Treatment 207
- ▪ Prolapse of Rectum 207
  - Incidence and Etiology 207
  - Pathogenesis 207
  - Symptoms 207
  - Signs 207
  - Relevant Investigation 208
  - Treatment 208

23. Vascular System **209**

- ▪ Acute Limb Ischemia 209
  - Incidence and Etiology 209
  - Symptoms 210
  - Signs 210
  - Relevant Investigations 211
- ▪ Acute Intestinal Ischemia 211
  - Incidence and Etiology 211
  - Clinical Presentation 211
  - Relevant Investigations 212
  - Treatment 212
- ▪ Leaking or Dissecting Aortic Aneurysm 212
  - Incidence and Etiology 212
  - Clinical Presentation 212
  - Relevant Investigation 213
  - Treatment 213

24. Urology **214**

- ▪ Acute Retention of Urine 214
  - Definition 214
  - Incidence and Etiology 214
  - Symptoms 214
  - Signs 214
  - Relevant Investigations 215
  - Treatment 215
- ▪ Hematuria 215
  - Definition 215
  - Incidence and Etiology 215
  - Symptoms 216
  - Signs 216
  - Relevant Investigations 216
  - Treatment 217
- ▪ Renal Colic 217
  - Relevant Investigations 218
  - Treatment 218
- ▪ Ureteric Colic 218
  - Symptom 218
  - Relevant Investigations 218
  - Treatment 219
- ▪ Acute Urethritis 220
  - Incidence and Etiology 220
  - Symptoms 220
  - Sign 220
  - Relevant Investigations 220
  - Treatment 220
- ▪ Acute Prostatitis 221
  - Incidence and Etiology 221
  - Symptom 221
  - Sign 221
  - Relevant Investigation 221
  - Treatment 221
- ▪ Acute Prostatic Abscess 221
  - Incidence and Etiology 221
  - Symptom 221
  - Sign 222
  - Relevant Investigations 222
  - Treatment 222

25. Male Genitalia **223**

- ▪ Acute Scrotal Pain 223
  - Definition 223
  - Causes of Acute Scrotal Pain 223
  - Symptoms 224
  - Past History 224
  - Signs 224

- On Palpation 225
- Differential Diagnosis by Clinical History and Examination 225
- Relevant Investigations 225
- Treatment 226

■ Torsion of Testis 226
- Incidence and Etiology 226
- Pathogenesis 226
- Symptoms 226
- Signs 227
- Relevant Investigation 227

■ Torsion of Appendages of Testis 228
- Incidence and Etiology 228
- Pathogenesis 228
- Symptom 228
- Signs 228
- Relevant Investigation 228
- Treatment 228

■ Acute Epididymo-orchitis 229
- Incidence and Etiology 229
- Pathogenesis 229
- Symptom 229
- Signs 229
- Relevant Investigations 229
- Treatment 229

■ Traumatic Orchitis 230
- Incidence and Etiology 230
- Clinical Presentation 230
- Relevant Investigation 230
- Treatment 230

■ Hematocele 230
- Definition 230
- Incidence and Etiology 230
- Symptoms 230
- Signs 231
- Relevant Investigation 231
- Treatment 231

■ Pyocele 231
- Definition 231
- Incidence and Etiology 232
- Symptom 232
- Signs 232
- Relevant Investigations 232
- Treatment 232

■ Idiopathic Scrotal Edema 232
- Incidence and Etiology 232
- Pathogenesis 232
- Symptom 232
- Signs 233
- Relevant Investigations 233
- Treatment 233

■ Acute Scrotal Abscess 233
- Incidence and Etiology 233
- Symptoms 233
- Sign 233
- Relevant Investigations 233
- Treatment 234

■ Fournier's Gangrene 234
- Incidence and Etiology 234
- Symptoms 235
- Signs 235
- Relevant Investigation 235
- Treatment 235

■ Acute Filarial Scrotum 236
- Incidence and Etiology 236
- Clinical Presentation 236
- Relevant Investigation 236
- Treatment 236

■ Fracture Penis 237
- Incidence and Etiology 237
- Symptoms and Signs 237

- Relevant Investigations 238
- Treatment 238

■ Paraphimosis 238

- Incidence and Etiology 238
- Symptoms 238
- Sign 238
- Relevant Investigation 238
- Treatment 238

■ Priapism 239

- Incidence and Etiology 239
- Clinical Presentation 240
- Relevant Investigation 240
- Treatment 240

26. Hernias 241

■ Complicated Hernias 241

- Definition 241
- Anatomy of the Hernial Sac 242
- Complications 242
- Symptoms 243
- Signs 244
- Relevant Investigations 244
- Treatment 244

27. Gynecology 245

■ Acute Torsion of Ovarian Cyst 245

- Incidence and Etiology 245
- Pathogenesis 245
- Symptom 245
- Signs 245
- Relevant Investigations 245
- Treatment 246

■ Acute Salpingitis 246

- Incidence and Etiology 246
- Pathogenesis 246
- Symptoms 246
- Signs 247
- Relevant Investigations 247
- Treatment 247

■ Rupture of Lutein Cyst 247

- Incidence and Etiology 247
- Clinical Presentation 247
- Relevant Investigations 247
- Treatment 247

28. Pediatrics 248

■ Acute Intussusception 248

- Incidence and Etiology 248
- Pathogenesis 248
- Symptoms 248
- Signs 249
- Relevant Investigations 249
- Treatment 249

■ Congenital Pyloric Stenosis 250

- Incidence and Etiology 250
- Pathogenesis 250
- Symptoms 250
- Signs 250
- Relevant Investigations 250
- Treatment 250

■ Necrotizing Enterocolitis 250

- Incidence and Etiology 250
- Pathogenesis 251
- Symptoms 251
- Signs 251
- Relevant Investigations 251
- Treatment 251

■ Tracheoesophageal Fistula 251

- Incidence and Etiology 251
- Pathogenesis 251
- Symptoms 252
- Signs 252
- Relevant Investigations 252
- Treatment 252

29. Lymphatic System 253

- Acute Lymphangitis 253
  - Incidence and Etiology 253
  - Clinical Presentation 253
  - Relevant Investigations 254
  - Treatment 254
- Acute Viral and Bacterial Lymphadenitis 254
  - Incidence and Etiology 254
  - Symptoms 254
  - Signs 254
  - Relevant Investigations 255
  - Treatment 255
- Acute Filarial Lymphangitis and Lymphadenitis 255
  - Incidence and Etiology 255
  - Symptoms 255
  - Signs 255
  - Relevant Investigation 255
  - Treatment 256

30. Skin and Subcutaneous Tissues 257

- Hematoma 257
  - Incidence and Etiology 257
  - Symptom 257
  - Sign 257
  - Relevant Investigations 257
  - Treatment 257
- Erysipelas 258
  - Incidence and Etiology 258
  - Symptoms 258
  - Investigations 259
  - Treatment 259
- Furuncle 259
  - Incidence and Etiology 259
  - Symptom 259
  - Signs 260
  - Relevant Investigations 260
  - Treatment 260
- Cellulitis 260
  - Incidence and Etiology 260
  - Pathogenesis 260
  - Symptoms 261
  - Signs 261
  - Relevant Investigations 261
  - Treatment 261
- Acute Pyogenic Abscess 261
  - Incidence and Etiology 261
  - Symptoms 262
  - Sign 262
  - Relevant Investigations 262
  - Treatment 262
- Carbuncle 262
  - Incidence and Etiology 262
  - Symptom 263
  - Signs 263
  - Relevant Investigations 263
  - Treatment 263
- Burns 263
  - Definition 263
  - Pathogenesis 264
  - Clinical Features 264
  - Metabolic Effects of Burns 267
- ❖ Physical Examination 268
  - Preliminaries Before Examination 268
- ❖ Inspection 271
  - General Examination 272
  - Systemic Examination 272
  - Relevant Investigations 272
  - Treatment 273
- Necrotizing Fasciitis 274
  - Incidence and Etiology 274
  - Pathogenesis 274

  - Symptom 274
  - Sign 274
  - Relevant Investigations 274
  - Treatment 274

SECTION SIX: SPECIAL PROCEDURES AND SITUATIONS

31. Emergency Surgical Procedures **277**

- Pericardial Aspiration 277
  - Materials Required 277
  - Procedure 277
  - Complications 278
- Intubation of Trachea 279
  - Materials Required 279
  - Procedure 279
  - Complication of Endotracheal Intubation 280
- Percutaneous Tracheostomy 280
  - Advantages of Tracheostomy 280
  - Indications 280
  - Contraindications 281
  - Materials Needed for Percutaneous Tracheostomy 281
  - Procedure 281
  - Complications 282
- Cricothyroidotomy/Minitracheostomy 282
  - Indications 282
  - Contraindications 283
  - Materials Required 283
  - Procedure 284
  - Complications 284
- Insertion of Chest Drains 284
  - Indications 284
  - Type of Drain 285
  - Site of Drain 285
  - Materials Required 286
  - Procedures 286
- Passing a Nasogastric Tube 287
  - Indications 287
  - Contraindications 289
  - Materials Required 289
  - Procedure 289
- Peritoneal Tap 289
  - Materials Required 289
  - Procedure 291

32. Preparing the Patient for Emergency Surgery **292**

  - General Measures 292
  - Investigations Support 292
  - Situations which Need Special Attention 293

33. Antibiotics **294**

- Antibiotics and Emergency Surgery 294
  - Which Emergency Surgical Procedure should have Antibiotic Prophylaxis ? 294

34. Death **296**

  - Breaking the News 296

Index *299*

Section I

Introduction

1. Introduction

1

Introduction

Emergencies are defined as the clinical situations which require urgent care. Not necessarily they should have a life-threatening pathology, but even a simple innocuous pathology can create an emergency situation, such as a passing calculus in the ureter can cause such a pain which will upset the patient, the relatives and friends, and warrant emergency care and also confuse the attending doctor till he makes a final diagnosis.

Clinical diagnosis of such emergencies may be difficult in some patients forcing the clinician to run through all the investigations to make the final diagnosis. A well informed medical student or a young surgeon will take it right, will not upset himself, comfort the patient and go through the clinical examination and administer the right treatment.

In some emergencies like hemorrhagic conditions, history taking cannot be done in detail due to paucity of time, and the treatment should get initiated at the earliest, probably in the intensive care unit. Failure to recognize and manage such critical illnesses, may lead to pathophysiological changes leading to multiple organ dysfunction. In such a situation, intensive care plays a very important role. The role of intensive care includes:

- Resuscitation and stabilization
- Optimization to prevent organ failure
- Decision regarding complex surgery
- Support of failing organs
- Recognition of futility.

In well developed hospitals, the critical care is managed at two levels, namely, intensive care units (ICU) and high dependency units (HDU), only the availability of ventilators in the ICU differentiating the both. In the ICU, some patients may require one patient – one nurse combination.

Critically ill patients can be classified according to the level of medical and nursing care needed.

- Level 0: Patients who can be managed by ward based care
- Level 1: Patients who can be managed by ward based care with advise or support from the critical care team
- Level 2: Patients requiring more advanced levels of monitoring or intervention, such as management of single failing organ system
- Level 3: Patients requiring advanced respiratory support and management of multiple organ failure.

For a good outcome, the first step is to make a proper diagnosis, followed by the necessary investigations and sensible management.

Section II

Assessment

2. Assessment of Surgical Emergencies

2

Assessment of Surgical Emergencies

As soon as the patient is wheeled into the hospital or the casualty, it is required to determine whether the patient's condition warrants emergency care or not. Though there are parameters which help to determine this, many times, experience matters. For example, a patient with acute pancreatitis may not look very sick after an episode of acute abdominal pain. The following measures will help determine the seriousness of the patient.

1. Make the patient lie down comfortably.
2. Elicit a quick history.
3. Make a thorough clinical examination.
4. Come to a quick working clinical diagnosis.
5. Ask for essential meaningful investigations.
6. Admit the patient where thought to be required.
7. Put the patient in intensive care whenever required.
8. Collect investigation reports at the earliest.
9. Start emergency treatment.
10. Ask for expert opinions.

MAKE THE PATIENT LIE DOWN COMFORTABLY

It is necessary to make the patient lie down, as a walk in patient may collapse suddenly (e.g. septic shock, hemorrhagic shock).

ELICIT A QUICK HISTORY

A good history will establish the diagnosis in majority of cases.

However, when the patient is too sick, a quick questioning is only possible. They are:

- A – allergies
- M – medications
- P – previous medical history, pregnancy
- L – Last meal
- E – Events leading to admission.

If possible, and if there is enough time, full history should be elicited in the following:

- Chief complaint
- History of present illness
- Previous history of present complaint
- Past medical history
- History of drug intake and allergies
- Social and personal history
- Family history
- History of immunization.

Make a Thorough Clinical Examination

General Examination

- Skin
  - Pallor (e.g. anemia)
  - Yellow discoloration (e.g. jaundice)
  - Bluish discoloration (e.g. cyanosis)
  - Pigmentation (e.g. von Recklinghausen's disease)
  - Eruptions (e.g. macule, papule, vesicle, bulla, etc.)
  - Nodules (e.g. Heberden's nodes of osteoarthrosis)
  - Visible veins (e.g. visible neck veins of cardiac overload, caput medusae of portal hypertension).
- Eyes
  - Orbital region
    - Prominence of eyeballs (e.g. exophthalmos of thyrotoxicosis)
    - Sunken eyeballs (e.g. dehydration)

- Eyelids
    - Puffiness (e.g. trauma, renal pathology)
    - Color (e.g. black eye of head injuries)
- Conjunctivae
    Color (e.g. pale in anemics)
- Sclera
    Color (e.g. yellow in jaundiced patients)
- Cornea and iris
    Opacities and ulcers (e.g. trauma, infections)
- Pupils
    Size, shape, symmetry and reaction to light (e.g. changes in cranial nerve lesions)

- Oral cavity
    - Lips
        - Color and texture (e.g. cracked lips of exposure to extreme cold)
        - Pigmentation (e.g. telangiectasia)
    - Gums
        - Color (e.g. blue line of lead poisoning)
        - Texture (e.g. inflammation due to gingivitis)
        - Pigmentation (e.g. hemorrhages due to leukemia)
    - Teeth
        - Color and pigmentation (e.g. mottled yellow color of fluorosis)
        - Number and health status of the teeth (e.g. Hutchinson's teeth of congenital syphilis)
    - Tongue
        - Color (e.g. blue color of central cyanosis)
        - Texture (e.g. smooth tongue of vitamin B_{12} deficiency)
        - Ulcer (e.g. aphthous ulcers)
        - Altered movements (e.g. hypoglossal nerve palsy)
    - Hard palate and soft palate
        - Color (e.g. anemia)
        - Cleft (e.g. cleft palate)
        - Ulcer (e.g. malignancy).

Halitosis (Bad breath) may be due to poor oral hygiene or even oral pathologies, e.g. oral cancers or bronchopulmonary pathologies like lung abscesses.

- Hands and feet
  - Hands of the patient should be examined carefully, like
    - Shape of the hands (e.g. characteristic shape of Dupuytren's contracture, tetany)
    - State of the joints (e.g. deformed joints of rheumatoid arthritis)
    - Shape and color and deformity fingers (e.g. clubbing or nicotine staining of chronic smokers)
    - Nails (e.g. koilonychia of iron-deficiency anemia)
    - Abnormal movements of fingers (e.g. tremors of thyrotoxicosis)
  - Feet of the patient are generally tucked under the bedclothes and their examination should not be forgotten. Examination may reveal
    - Edema, unilateral (e.g. filarial leg) and bilateral (e.g. hypoproteinemia)
    - Skin changes (e.g. pigmentation or ulcers due to ischemia)
    - Ulcers (e.g. trophic ulcers)
- Look for (jaundice, anemia, cyanosis, clubbing, oedema, lymphadenopathy – JACCOL)
- Vitals (pulse, respiration, blood pressure)
- Peripheral examination (warm/cool, capillary refill)
- Jugular venous pressure
- Mental test score or GCS (Glasgow Coma Scale), if the patient appears confused or impaired consciousness).

Systemic Examination

- Site of interest
- Other systems.

Come to a Quick Working Clinical Diagnosis

- A quick working clinical diagnosis should come to the mind of the clinical examiner
- He should also keep the various differential diagnosis in mind, so that diagnosis can be revised at anytime, without bias.

Ask for Essential Meaningful Investigations

- It is essential to ask only for meaningful investigations
- If investigation packages are available in the hospital, it is worthwhile to ask for it which may include the required investigations also
- If asking for this package will delay the results, it is appropriate to ask for the relevant ones first, followed by the rest.

The blood tests in general should include:

- Complete blood count (CBC)
- Urea, creatinine and electrolytes
- Blood sugar (random)
- Serum amylase, lipase
- Liver function tests
- Coagulation profile
- Blood grouping and Rh typing
- Arterial blood gas analysis (ABG).

The radiological tests should include:

- Chest X-ray (erect)
- Abdominal X-ray (supine and erect)
- Ultrasound scan
- CT.

Admit the Patient where Thought to be Required

No time should be wasted in deciding where to admit the patient. The best bed available is the best choice.

Put the Patient in Intensive Care Whenever Required

When in doubt about the seriousness, admit the patient in the intensive care.

Collect Investigation Reports at the Earliest

In the commotion during the emergency, it is not unusual to forget to collect the reports. The clinician should stay composed, and should remember to collect all the reports, which will help in the management.

Start Emergency Treatment

- It is always wise to start an intravenous line for control of lifeline. The choice of fluid depends on the situation (e.g. normal saline in diabetics, glucose solutions in hypertensives, plasma expanders in shocked patients)
- Antibiotics may be needed if infective pathology is suspected
- Oxygen by nasal mask is a safe additive
- Urethral catheterization and hourly output determination is necessary.

- Start emergency treatment if required even in the casualty, and history taking can be followed later
- While examining a patient in emergency, site of interest should be examined first to save time so that appropriate treatment is started.

Ask for Expert Opinions

- It is always necessary to obtain expert opinions to manage the situation
- Multiple opinions may be necessary, but it should not hurt the sentiments of the experts involved
- Intravenous fluids (rate and nature) and antibiotics will vary according to the laboratory reports and revised diagnosis
- Colloids may be necessary and should be adequately given at appropriate times
- The average daily water requirements for adult are 30 to 35 ml/kg + 500 ml/day/degree of pyrexia (above 37°C)
- Normal adult requirements for electrolytes are 1 mmol/kg each of Na, K, Cl
- Normal fluid requirements for children depend on weight
  - <10 kg : 100 ml/kg/day
  - 10 to 20 kg : 1 liter + 50 ml/kg/day
  - >20 kg : 1.5 liters + 25 ml/kg/day.

Section III

Critical Care

3. Critically Ill Patient and Critical Care
4. Shock
5. Acute Respiratory Distress Syndrome

3

Critically Ill Patient and Critical Care

A critically ill patient requires high level monitoring, especially when circulatory/respiratory support is required, or if more than one organ is failing. The intensive care is started early enough to be able to reverse the condition.

CARDIAC SUPPORT

It depends on the Hb, perfusion pressure, ventilation and gas exchange. For this to be achieved, the following are required:

- Arterial lines (beat to beat BP indication, easy access for arterial blood sampling)
- Central venous pressure monitoring (measures intravascular volume, and the right heart preload)
- Pulmonary artery catheterization (measures cardiac output, ensures optimal fluid resuscitation and helps in the use of vasoactive drugs).

> CVP monitoring indicates only the right heart filling or preload, is not a reliable measure of left ventricular preload. Pulmonary artery catheterization is useful for monitoring the left ventricular filling.

RESPIRATORY SUPPORT

Respiratory support is needed when respiratory failure occurs, due to inadequate exchange of oxygen or CO_2 to meet metabolic needs, which is determined by the lack of improvement with oxygen therapy or the patient is tiring with an increasing pCO_2. The methods by which respiratory support can be given are:

- Continuous positive airways pressure (CPAP) in a spontaneously breathing patient. This is indicated when:
    - The patient is tiring with rising pCO_2
    - The patient is unable to maintain their own airway
    - Noninvasive ventilation is contraindicated
    - Need for endotracheal suction
- Noninvasive positive pressure ventilation (NIPPV) is an alternative via a face mask.

INOTROPIC SUPPORT

This is needed when the patient is not able to maintain the blood pressure and urine output, by normal crystalloid support. Different inotropic agents are used in various circumstances:

- Noradrenaline: Dose 0.01 to 0.4 mcg/kg/min IV infusion — acts on α_1-receptor causing vasoconstriction, in sepsis increases the renal blood flow and enhance urine production
- Adrenaline: Dose 0.01 to 0.30 mcg/kg/min IV infusion
    - In low doses: Acts on β-receptors (causes increase in heart rate and contraction)
    - In high doses: Acts on α-receptors (causes increased peripheral resistance)
- Dobutamine: Dose 1 to 25 mcg/kg/min via central vein—predominantly acts on β_1 (increases the heart rate and force of contraction) and also on β_2- and α_1-receptors (decreases peripheral and pulmonary vascular resistance)
- Dopamine: Acts directly on α, β and dopaminergic receptors and indirectly by releasing noradrenaline
    - Low dose: 0.5 to 2.5 mcg/kg/min (renal dose—increases renal and mesenteric blood flow. ↑ renal flow results in ↑ GFR and ↑ renal sodium excretion)
    - Moderate dose: 2.5 to 10 mcg/kg/min (cardiac dose—stimulates β_1-receptors causing ↑ myocardial contractility, stroke volume and cardiac output)
    - High dose: >10 mcg/kg/min (cardiac plus—stimulates α-receptors causing ↑ SVR, ↓ renal blood flow and ↑ potential for arrhythmias
- Dopexamine: Dose 0.25 to 0.5 mcg/kg/min—a synthetic analog of dopamine with β_1 activity with no α activity.

4

Shock

Shock is defined as inadequate organ perfusion and tissue oxygenation.

It is classified as:

- Hypovolemic shock
- Septic shock
- Anaphylactic shock
- Cardiogenic shock
- Neurogenic shock.

HYPOVOLEMIC SHOCK

This occurs due to loss of intravascular volume (blood or fluid), which results in activation of sympathetic nervous system, which causes tachycardia and vasoconstriction of skin, muscle and GI system, so that blood is directed to vital organs, e.g. brain, heart, etc.

The vasoconstriction of renal and splanchnic circulation causes renal failure, GI sloughing and hemorrhage. When shock persists, the perfusion of brain suffers causing confusion and aggression. The hyperventilation causes respiratory alkalosis, which is overtaken by metabolic acidosis due to poor tissue perfusion and anaerobic metabolism. The hypovolemic shock is classified depending on various factors (Table 4.1).

Table 4.1: Classification of hypovolemic shock

Class	*Blood loss*	*Pulse rate (per min)*	*Blood pressure*	*Urine output (ml/hr)*	*Resp rate (per min)*	*Consciousness level*	*Treatment*
I	<15% (up to 750 ml)	60–100	No change	>30	<20	No change	Crystalloid/colloid
II	15–30% (750 – 1500 ml)	>100	Pulse pressure reduced	20–30	20–30	Anxious	Crystalloid/colloid
III	30–40% (1500 – 2000 ml)	>120	BP fall	5–15	30–40	Confused	Blood
IV	>40% (>2000 ml)	>140	BP significant fall	Anuria	>40	Lethargic	Blood

Treatment of Hypovolemic Shock

- Control of hemorrhage
- Restoration of fluid volume.

- A fluid challenge of 1 to 2 liters should be given for adults and 20 ml/kg for children
- Rapid restoration of pulse and BP indicates a loss of <20 percent
- If the response is transient in spite of resuscitation, it is beter to intervene surgically.

SEPTIC SHOCK

Shock occurring due to severe sepsis called systemic inflammatory response syndrome (SIRS).

Symptoms and Signs

- Heart rate (>90/min)
- Temperature (>38 or 36)
- Repiratory rate (>20/min)
- Leukocyte count (>12000/cmm)

Note: In the absence of infection, SIRS can be said to exist when two of the above criteria.

Pathogenesis

This results due to severe infection and inflammation, mediated by acute phace cytokines, which have a generalized effect distant to the site of original insult. Leukocytes adhere to endothelial cells via adhesion molecules leading to changes in vascular permeability and edema.

Treatment

- Resuscitation
- Identification of the source of sepsis
- Treatment of focus of sepsis (antibiotics/drainage of pus)
- Critical care management
    - Fluids
    - Oxygen administration
    - Vasopressors
    - Steroids
    - Activated protein C
    - Hemofiltration.

ANAPHYLACTIC SHOCK

Incidence and Etiology

This is a hypersensitivity reaction.

The usual causes are:

- Drugs
- Blood transfusion
- Radiological contrast.

Pathogenesis

It is mediated by immunoglobulin E (IgE).

Symptoms and Signs

- Urticarial rash
- Wheeze due to bronchospasm
- Tachycardia
- Fainting
- Rhonchi
- Hypotension.

Treatment

- Stop the cause if found
- Airway management and oxygen
- IV access and fluids
- Adrenaline (1 ml in 1:1000 IM)
- Antihistamine (IV)
- Bronchodilators.

CARDIOGENIC SHOCK

Occurs due to pump failure, commonly after myocardial infarction.

NEUROGENIC SHOCK

Occurs after spinal cord injury.

Pathogenesis

Occurs due to disruption of autonomic nervous system control over vasoconstriction, which results in a decrease in peripheral vascular resistance and blood pressure, with resultant bradycardia. Temperature control may also be lost.

Treatment

- Ventilation
- Fluid management
- Inotrope support.

5

Acute Respiratory Distress Syndrome

Incidence and Etiology

- Acute respiratory distress syndrome (ARDS) indicates the acute diffuse pulmonary inflammatory response to either direct or indirect insults from extrapulmonary pathology
    - Direct—via airway or injury to chest (e.g. aspiration, toxic gases, pneumonia)
    - Indirect—blood-borne insults (e.g. sepsis, polytrauma, severe burns, drugs)
- Frequently associated with multiple organ (kidney, liver, intestines) dysfunction.

Pathogenesis

- Pathogenesis is complex
- There is transudation of fluid in the lungs, thickening of alveolar capillaries resulting in ventilation/ perfusion mismatch. The pulmonary artery wedge pressure remains normal, distinguishing it from pulmonary edema.

Clinical Presentation

They are not specific, but share many things with other pulmonary pathologies.

Relevant Investigations

- Blood gas analysis (PaO_2 / FiO_2 of less than 200 mm Hg)
- Chest X-ray shows bilateral diffuse infiltrates
- Pulmonary artery wedge pressure (less than 15 mm Hg).

Treatment

Current treatment is supportive and no specific therapy exists to modulate the sequence of events of ARDS:

- Monitoring
  - Monitoring of all vitals
- Ventilatory management
  - Mechanical ventilation to permit adequate oxygen uptake
- Nonventilatory management
  - Treatment of underlying risk factors
  - Enteral feeding
  - Maintenance of hemodynamic stability and cardiac output.

Section IV

Trauma

6. Polytrauma
7. Head Injuries
8. Facial Injuries
9. Spine and Spinal Cord Injuries
10. Thoracic Injuries
11. Abdominal Injuries
12. Urological Injuries
13. Male Genital Injuries
14. Female Genital Injuries
15. Hand Injuries

6

Polytrauma

INTRODUCTION

Management of a patient with polytrauma depends on proper and systematic clinical evaluation, which identifies immediate and potentially life-threatening conditions before the limb threatening ones, but does not omit the latter. This usually starts with a call from ambulance control.

The first couple of hours following injury are critical, as the patient is very vulnerable. This critical time period is called 'golden hour', which is usually spent in the place of accident or in the emergency department, making this period extremely crucial.

The trauma life support program consists of:

- Preparation
- Primary survey and resuscitation
- Secondary survey
- Continuous monitoring and evaluation.

PREPARATION

Prehospital Communication

A warning from the ambulance control ideally from the scene of accident provides essential information so that receiving personnel are ready to receive the patient of trauma, which provides assessment and treatment without delay. The essential prehospital information are:

- Nature of injury
- Mechanism of injury

- Number, age and sex of the casualties
- Consciousness level of the casualties
- Airway, breathing and circulatory status of casualties
- Treatment provided
- Estimated time of arrival
- Treatment options required.

TRAUMA RECEPTION TEAM

The make-up of this team varies between hospitals depending upon the resources and the time of the day. The trauma team should consist of:

- Trained medical personnel
- Supportive trained paramedical personnel
- A team leader who should assign specific tasks to each person.

The trauma reception team undertakes the following tasks:

Team Leader

- Coordinates the tasks assigned to the team members
- Questions ambulance personnel
- Assimilates clinical findings
- Determines investigations in order of priority
- Liaises with relatives and provide information to trauma team
- Liaises with specialists who are called.

Team Members

- Manage the airway
- Clear the secretions and intubate if necessary
- Manage circulation
- Establish infusion line quickly
- Takes blood for investigations
- Supportive measures
- Connect to monitors
- Urinary catheterization
- Jugular vein catheterization where necessary.

Receiving the Patient

Receiving and transferring the patient (Fig. 6.1) is very crucial and requires five people to do the job, which should be a well practiced procedure in order to protect the spinal cord if it is intact, and to prevent further injury if it is already compromised. During the transfer, the patient's head and neck are stabilized by one member of the team, three others lift from the side and the fifth member replaces the ambulance trolley with the resuscitation trolley.

Primary Survey and Resuscitation

The activities listed below are carried out simultaneously if there are enough personnel, if not should be done in alphabetical order (ABCDE).

- Airway and cervical spine control
- Breathing
- Circulation and hemorrhage control
- Disability
- Exposure.

Airway and Cervical Spine Control

A cervical spine injury should be assumed if the patient has been the victim of significant blunt trauma or if the mechanism of injury indicates that the cervical region may have been damaged. One member of the team needs manually to immobilize the cervical spine while talking to the patient which also assesses the airway.

FIG. 6.1: Patient with polytrauma

- If the patient is able to give a logical answer in a normal voice, the airway is assumed to be patent and the brain adequately perfused.
- If the patient gives an impaired or fails to reply, the airway could be obstructed, and immediate measures should be taken:
    - A simple chin lift will help in relieving the obstructing soft tissue usually the tongue
    - Saliva, blood, vomitus, tooth or other foreign bodies should be removed
    - Suction through rigid suction tube is necessary to remove secretions
    - In patients who vomit and regurgitate, head end of the bed should be dropped 20 degrees, allowing the secretions to drip down and facilitate removal by suction
    - A nasogastric tube may be inserted to aspirate the stomach contents and prevent further vomiting.

When it is established that the airway is clear and patent, 100 percent oxygen is provided via mask or by endotracheal intubation. Pulse oxymeter is connected to maintain good SaO_2, if needed with ventilatory support.

The neck should be examined for the following five signs which could indicate the presence of immediately life-threatening thoracic conditions (Table 6.1).

When below given signs in Table 6.1 are checked, the neck may be immobilized with appropriate collar if the patient is not restless, in a restless patient, semi-rigid collar is accepted.

All the multiple injured patients, particularly those who have injuries above the clavicle or a change in level of consciousness, should be treated as though they have a cervical spine injury, until it is ruled out.

Breathing

The clinical examination of chest consists of:

- Inspection
    - Marks and wounds
    - Respiratory rate
    - Inspiratory effort
    - Symmetry of chest movements

Table 6.1: Signs of life-threatening thoracic conditions

	Signs	*Conditions*
1.	Swellings and wounds	Vascular and airway injury
2.	Distended neck veins	Cardiac tamponade, tension pneumothorax
3.	Tracheal deviation	Tension pneumothorax
4.	Subcutaneous emphysema	Pneumomediastinum
5.	Laryngeal crepitus	Fracture of laryngeal cartilage

- Percussion
    - Assess ventilation at the periphery
- Auscultation
    - Assess defects in air movement.

The respiratory rate and effort are sensitive indicators of underlying lung pathology, and requires to be monitored and recorded at frequent intervals.

Circulation and Hemorrhage Control

Bleeding can result due to:

- Fractures of long bones
- Vascular injuries
- Soft tissue injuries

Application of pressure is the best way of management of hemorrhage, as tourniquets increase intraluminal pressure of the vessels, distal ischemia and tissue necrosis. Application of tourniquet is undertaken in select situations, and when used the time of application of tourniquet has to be noted so that neighboring soft tissue is not jeopardized.

Recognition and Assessment of Hypovolemia

The assessment of hypovolemia has to be done in a systematic manner:

- Skin – color, clamminess and capillary refilling
- Vital signs – heart rate, blood pressure and pulse volume
- Consciousness level

Isolated determinations of above parameters is unreliable as various organs try compensate at early stages, especially at extremes of age.

Common sites of occult bleeding are:

- Chest
- Abdomen and retroperitoneum
- Pelvis
- Long bone fractures
- External bleed into splints and dressings.

Fluid Resuscitation

Once any overt bleeding is controlled, it is necessary to maintain the circulatory volume. This is done by administering a warm crystalloid followed by blood transfusions, maintaining the radial pulse and blood pressure. This administration of fluids is done by peripheral venous cannulation, and when not possible done through central venous catheterization.

- Vital signs return to normal after less than 2 liters of fluid are administered, when the lost blood is less than 20 percent of the blood volume
- Transient responders who are actively bleeding or recommence bleeding during the resuscitation, and the improved vital signs deteriorate indicating loss of over 30 percent of the blood volume
- Little or no response indicates that the loss is more than 40 percent or no hypovolemia.

Disability

Disabilities when occur in a patient with trauma shows seriousness. It can occur in:

- Hypoxia
- Hypovolemia
- Hypoglycemia
- Increased intracranial pressure.

Exposure

The patient’s clothes have to be removed by cutting through the seams so that there is minimal patient movement. All clothes are removed only after adequate intravenous access is established, as a rapid removal of tight trousers can precipitate sudden hypotension due to the loss of the tamponade effect in a hypovolemic patient.

Once stripped, trauma victims should be kept warm with blankets when not being examined. Now the patient is rolled on and the spine examined from base of skull to the coccyx, with a rectal examination.

What to look for during rectal examination in a trauma victim:

- Is the sphincter tone present?
- Is the rectal wall breached?
- Is the prostate in a normal position?
- Is there blood on the examiner’s finger?

SECONDARY SURVEY

Once the immediately life-threatening conditions are recognized and treated, the whole of the patient is assessed. This requires head to toe, front to back assessment along with detailed medical history and appropriate investigations.

If the patient deteriorates at any stage, the airway, breathing and circulation must be immediately re-assessed as described in the primary survey.

CLINICAL EVALUATION

Eliciting History

The history is elicited quickly and in concise manner, concentrating on:

- Nature of injury
- Mechanism of injury
- Treatment provided
- Medical history.

Physical Examination

The physical examination should be done in a systematic manner.

Scalp

The scalp is examined for:

- Lacerations
- Swellings
- Depressions

The examination is carried out by inspection and palpation (by running the fingers on the scalp).

- It should be remembered to examine the occipital region, which is usually missed out
- Blind probing should be avoided as it may further damage the underlying structures
- In children, bleeding from scalp lacerations can cause hypotension and efforts to stop bleeding (application of pressure, applying self-retaining retractors) have to be taken immediately.

Neurological State

The neurological state of the patient by the Glasgow Coma Score, the papillary responses and the presence of lateralizing signs should be assessed. This examination should be done frequently so that any deterioration is detected early.

Base of Skull

Since the skull base lies along a diagonal line running from the mastoid to the eye, the signs of a fracture in this region are also found along this line.

Signs of a base of skull fracture

- Bruising over the mastoid (Battle's sign)
- Panda's eyes
- Blood and CSF rhinorrhea
- Blood and CSF otorrhea
- Hemotympanum
- Scleral hemorrhage with no posterior margin
- Subhyoid hemorrhage.

- Battle's sign and Panda's eyes appear after about 12 to 36 hours, and is not a reliable sign in a resuscitation room.
- CSF rhinorrhea may be missed as it is invariably mixed with blood.

Neck

If any deformity is found, it is necessary to splint the neck preferably with a collar.

Eyes

- Hemorrhage
- Foreign bodies (including contact lenses)
- Papillary response and corneal reflexes (in unconscious patient)
- Visual acuity (in conscious patient).

Face

The face is examined by systematic inspection and palpation, and look for:

- Soft tissue injuries
- Fractures of midface
- Fractures of mandible
- Missing or lost teeth.

- Midface fractures may be associated with fractures of base of skull
- Mandibular fractures can cause airway obstruction due to loss of stability of tongue.

Thorax

The thorax is examined for

- Soft tissue injuries
- Fractures of clavicles and ribs
- Crepitus (e.g. surgical emphysema)
- Movements of chest (e.g. flail chest).

Abdomen

The abdomen is examined in a very systematic manner, without forgetting to examine the pelvis and perineum. Percussion is an ideal way to determine the injuries of liver and spleen. In penetrating injuries, if bowel is exposed, it should be covered immediately with sterile cloth.

Urine output measurement is a good indicator in a shocked patient, and this requires catheterization for accurate measurement. Transurethral catheterization without any difficulty may indicate that is no severe urethral disruption. If the patient is not able to urinate, and catheterization is not possible, severe urethral injury should be suspected, and suprapubic catheterization should be done. The signs of urethral injury in a male are:

- Bruising around the scrotum
- Blood at the urethral meatus
- High-riding prostate.

Extremities

The limbs are examined in the traditional manner of inspection, palpation and active and passive movement.

- The fractures should be assessed and the blood loss estimated
- Distal loss of arterial pulsation may be due to profound shock or arterial injury.

DEFINITIVE CARE

Once the patient is adequately assessed and successfully resuscitated, he is moved to the next stage for definitive care, but they are vulnerable during transfers as monitoring becomes difficult during this time.

Before transfer, it is necessary to:

- Close an open wound with sterile pads
- Correct the deformity and splint the limb.

Note:

- Intra-abdominal bleed should be suspected if the patient is hemodynamically unstable, especially if the lower six ribs are fractured or there are marks on the abdominal surface
- In unexplained hypotension, pelvic bone injuries should be suspected. Compression or distraction test is useful
- Clinical examination may not be reliable in neurologically unstable patients, and investigations like ultrasound and CT or MRI may be needed. Diagnostic peritoneal lavage may be relied upon in certain situations like massive hemoperitoneum.

7

Head Injuries

Head injuries are classified into:
- Brain injury
- Fractures of skull
- CSF fistula
- Injuries of cranial nerves
- Vascular injuries (rare)
- Scalp injuries.

BRAIN INJURY

For the purpose of understanding the pathophysiology, brain injury is subdivided into:
- Primary brain injury and
- Secondary brain injury

Primary brain injury is further classified into:
- Concussion brain
- Diffuse axonal injury
- Primary brainstem injury
- Contusions and lacerations of brain

These injuries occur depending on the:
- Severity of impact
- Direction of impact force

- Movement of head:
    - *Type of injury:* closed or penetrating injury. Mild force causes ***concussion*** of brain characterized by transient loss of consciousness, Post traumatic amnesia, confused state and patient recovers completely
    - Severe force causes ***Diffuse Axonal Injury*** characterized by prolonged unconsciousness and neurologic deficits
    - More severe force causes ***Primary Brain stem Injury***. With extremely severe impact force, fatal injury occurs with death at the accident spot itself
    - To-and-fro movement of head shakes various parts of the brain within the skull causing different injuries: gray matter moves over white matter; subcortical white matter moves over basal ganglia; brain surface hits against the rough floor of cranial fossa floor and sharp edges of falx and tentorium leading to ***contusions of brain***. When the pia-arachnoid is torn it is termed ***laceration***.
- With impact force/acceleration-deceleration force, brain moves within the cranial cavity—cerebrum moves over brainstem leading to deformation of neuronal-synaptic membranes at central reticular core.

Secondary Brain Injury

- Hypoxic-Ischemic injury to brain because of impaired breathing, aspiration, airway obstruction, chest injury, cervical spine injury, hypotension, etc.
- Intracranial hematomas (viz., acute extradural hematoma, acute subdural hematoma, acute intracerebral hematoma) or expanding contusions and brain edema can cause brain shifts and increased intracranial pressure
- Metabolic abnormalities of glucose, sodium, acid-base, etc.

Acute extradural hematoma Vs Acute subdural hematoma:

- Acute extradural hematoma occurs at the site of impact and hence the scalp injury or fracture skull will give clue to the location of extradural hematoma
- Bleeding is usually from the middle meningeal branches or from fracture edges
- Acute subdural hematoma is 6 to 8 times more common than extradural hematoma and is often due to acceleration and deceleration injury
- Subdural hematoma can be contracoup, i.e. its location can be diagonally opposite to the site of scalp injury. The bleeding is from the cortical veins
- Due to associated severe primary brain injury, prognosis is in general poor compared to isolated extradural hematoma.

FRACTURES OF SKULL

- Skull fractures are classified into:
  - ***Fractures of skull vault*** which can be:
    - Linear fracture (Fissured fracture)
    - Depressed fracture
  - ***Fractures of skull base*** which can involve:
    - Anterior cranial fossa
    - Middle cranial fossa
    - Posterior cranial fossa
- They can be also be classified into:
  - Simple (or closed) fractures
  - Compound (or open) fractures
    (depending on the absence or presence of communication to atmospheric air through breach of skin or air containing cavities in the skull)
- Fractures of skull vault will indicate the location of underlying extradural hematoma when present
- Compound depressed fractures are potential source for spread of infection leading to brain abscess, meningitis, epidural /subdural abscess
- Depressed fractures of vault can lead to direct brain contusions and hematomas.

> Skull fracture indicates that the force of injury was severe but does not indicate the severity of brain injury.

Symptoms and Signs

- Fractures of anterior cranial fossa (Fig. 7.1) can cause:
  - Ecchymosis of the upper eyelids appearing after about 24 hours
  - Subconjuctival hemorrhage in the upper eyelid in anterior skull base fractures is pointed towards the cornea and posterior limit will not be seen
  - In contrast in “Black Eye” where the injury is directly over orbit, the ecchymoses occur in both upper and lower eyelids and appear within minutes of injury
  - The subconjuctival hemorrhages are actually conjuctival in plane and will move with conjunctiva
- Cranial nerve palsies occur in fractures of the base of skull across or near the cranial nerve exit foraminae:
  - Fracture of temporal bone can explain 7th and 8th nerve palsies
  - Fracture of superior orbital fissure can explain spread of injuring force close to 3rd, 4th and 6th nerves
  - Fractures near jugular foramen can cause lower cranial nerve palsies
- CSF rhinorrhea occurs in fractures involving cribriform plate or paranasal sinuses

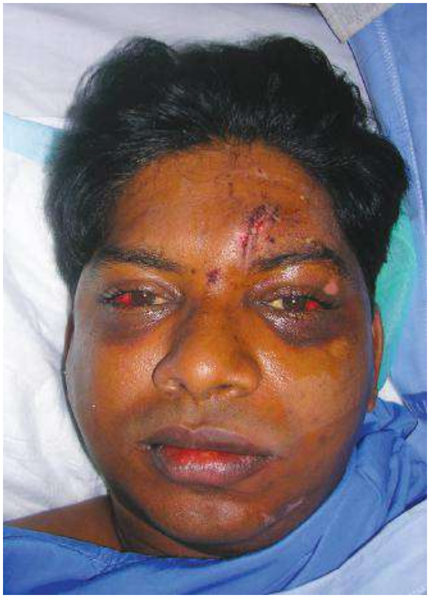

FIG. 7.1: Fracture of anterior cranial fossa

- CSF otorrhea occurs in fractures of temporal bone Eustachian tube CSF rhinorrhea. CSF leaks have the risk of meningitis
- Pulsatile proptosis in head injury can be due to carotico-cavernous fistula or more commonly comminuted fracture of orbital roof with normal pulsation of brain being transmitted to orbit. In carotico-cavernous fistula bruit may be present over orbit.

Note: Cranial nerve palsies can occur without skull fracture, e.g. olfactory nerve injury or optic nerve injury.

Relevant Investigations

X-rays (AP, lateral and oblique views) and CT of skull (Figs 7.2A to D) are diagnostic.

Treatment

- Closed depressed fractures over forehead may need elevation for cosmetic purposes
- Compound fractures will need debridement to prevent brain abscess.

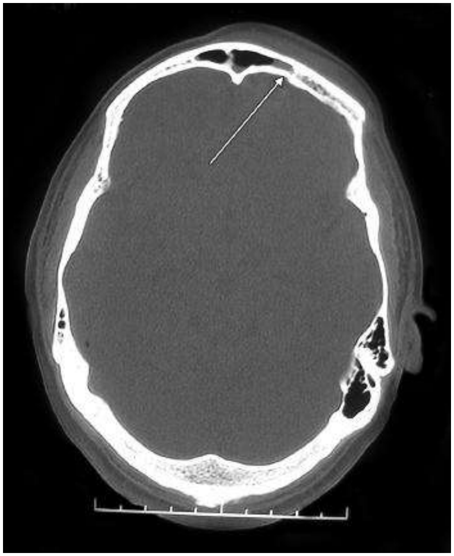

FIG. 7.2A: CT—Fracture of frontal bone

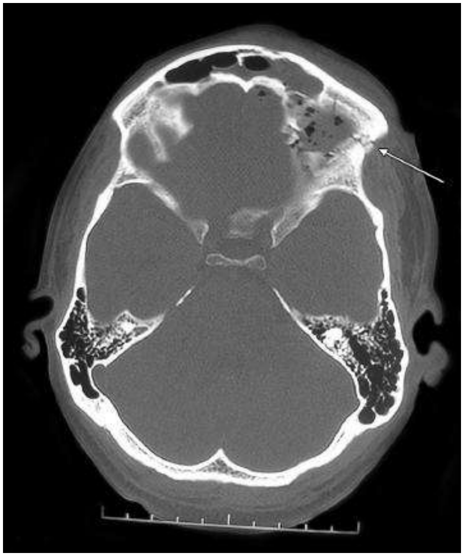

FIG. 7.2B: CT—Fracture

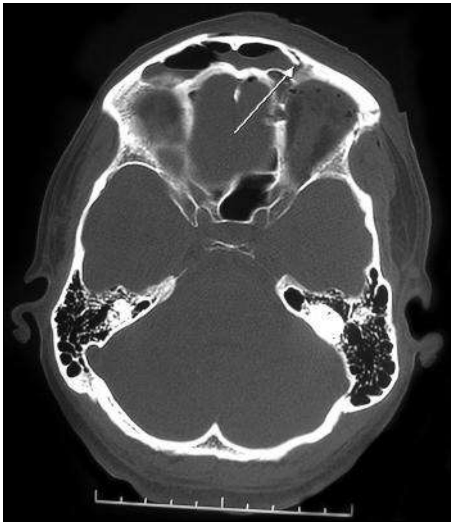

FIG. 7.2C: CT—Fracture of frontal sinus

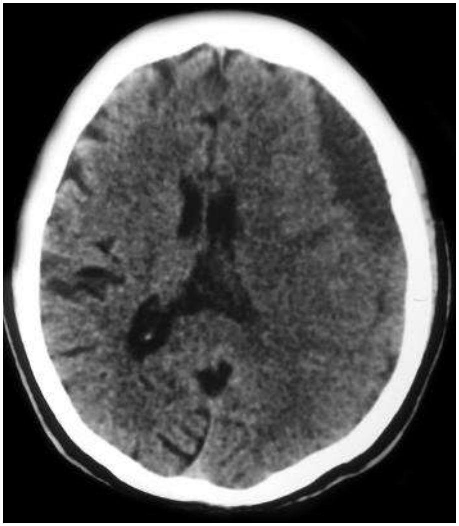

FIG. 7.2D: CT—Subdural hematoma

CSF FISTULAE

Incidence and Etiology

Leak of CSF is caused by:

- Fractures involving cribriform plate or paranasal sinuses can lead to CSF rhinorrhea
- Fractures of temporal bone can lead to CSF otorrhea and through Eustachian tube CSF rhinorrhea.

Complications: Meningitis

Symptoms and Signs

Watery discharge is pathognomonic of CSF leak.

Relevant Investigations

CT of paranasal sinuses after intrathecal contrast is useful.

Treatment

- Lumbar pucture and CSF drainage intermittently for 10 – 20 days
- Surgical repair is necessary if leak persists.

SCALP INJURIES

Scalp injuries indicate the sites of impact force.

They are of three types:

1. Contusions
2. Lacerations
3. Hematomas
    - Subpericranial—Confined to an area of one cranial bone and fixed
    - Subaponeurotic or subgaleal—More diffuse and extend beyond the margins of bones
    - Subcutaneous—Superficial and moves with the scalp.

- Fractures beneath scalp injury in unconscious patients suggest the possibility of underlying extradural hematoma
- Scalp injury over occipital region may give clue to posterior fossa hemorrhage
- The center of scalp hematomas may liquefy in the center after a few days and often give a false impression of depressed fracture to the palpating finger.

Symptoms and Signs

- Bleeding from an open wound
- Swelling in the head with or without external injury to the scalp.

Treatment

Immediate suturing of the wound, as the vessels are prevented from normal contraction by fixation of their walls to fibrous stroma of scalp.

Note: When there is scalp laceration, before wound closure, depressed fracture has to be excluded clinically under aseptic precautions.

8

Facial Injuries

FRACTURES OF MIDFACIAL SKELETON

- The fracture of facial skeleton is usually due to direct injury
- Types of fractures of midface are:
  - Floor of the orbit (Blowout fracture)
  - Zygomatic
  - Nasal
  - Maxillary [LeFort classification (Fig. 8.1)]
    - LeFort I (low level fracture): This is horizontal fracture above the level of the nasal floor (Guerin's fracture, floating fracture)

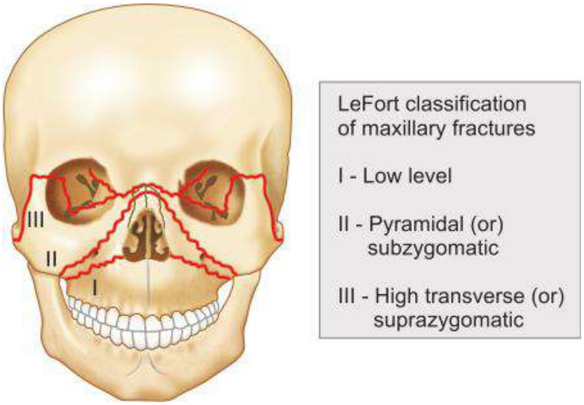
III
II
I
LeFort classification of maxillary fractures
I - Low level
II - Pyramidal (or) subzygomatic
III - High transverse (or) suprazygomatic

FIG. 8.1: Fractures of maxilla

- LeFort II (pyramidal or subzygomatic fracture): The fracture runs from the thin middle area of the nasal bones down either side crossing the maxillary processes into the medial wall of each orbit
- LeFort III (high transverse or suprazygomatic fracture): The fracture runs from near the frontonasal suture transversely backwards, parallel with the base of the skull and involves the full depth of the ethmoid, including the cribriform plate.

FRACTURES OF MANDIBLE

Fractures of the mandible (Fig. 8.2) can be divided according to the anatomical location into 8 types:

1. Dentoalveolar
2. Condylar
3. Coronoid
4. Ramus
5. Angle
6. Body
7. Parasymphysis
8. Symphysis

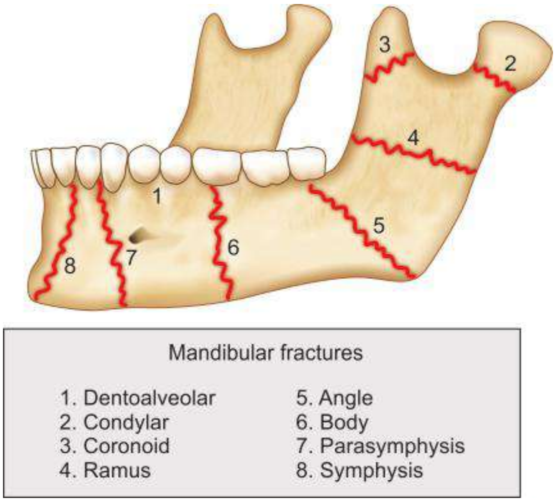
3
2
4
1
5
6
8
7
Mandibular fractures
1. Dentoalveolar
2. Condylar
3. Coronoid
4. Ramus
5. Angle
6. Body
7. Parasymphysis
8. Symphysis

FIG. 8.2: Fractures of mandible

The mandibular fractures can be:

- Unilateral
- Bilateral
- Multiple
- Comminuted.

Clinical Presentation

History

- History of injury and hearing or feeling of a bone crack
- The nature of impact (direct violence, indirect violence or excessive muscular contraction) should be determined.

Symptoms

- Pain and loss of function
- Diplopia and enophthalmos are present in LeFort III type
- But loss of function may not be a feature of impacted fracture.

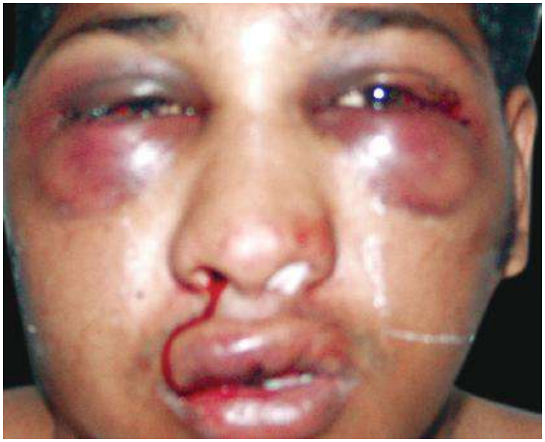

FIG. 8.3A: Ecchymosis and nasal bleed in injury of midface

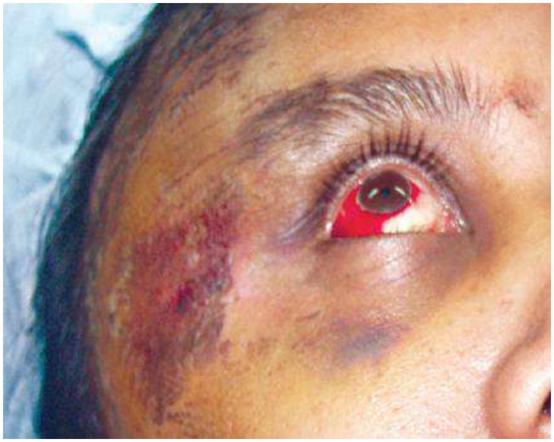

FIG. 8.3B: Subconjunctival hemorrhage

Signs

- Tenderness of the bone on applying pressure
- Swelling and ecchymosis (Fig. 8.3A)
- Nasal bleed (Fig. 8.3A) and CSF rhinorrhea are common in associated base of skull fractures
- Subconjuctival hemorrhage occurs in orbital bone fractures (Fig. 8.3B)
- Blebs on the skin overlying the fractured site
- Deformity: This is elicited by inspection, palpation and measurement. This is the most important sign of a fracture. The deformity in case of a flat bone is either depression or elevation and in case of long bone this is angular, lateral, longitudinal or rotatory. Hard palate is deformed in Le Fort I fracture (Fig. 8.4)
- Dental alignment is lost in displaced fractures (Fig. 8.5).

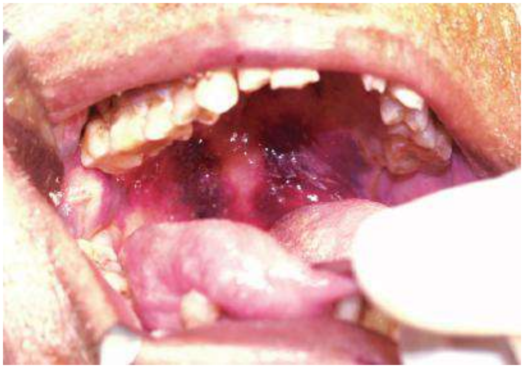

FIG. 8.4: LeFort I fracture

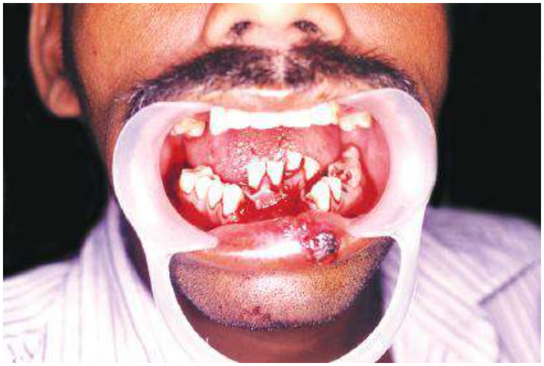

FIG. 8.5: Displaced mandibular fracture

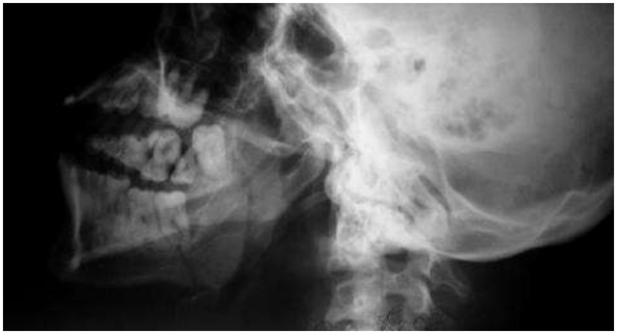

FIG. 8.6A: X-ray—Fracture of ramus of mandible

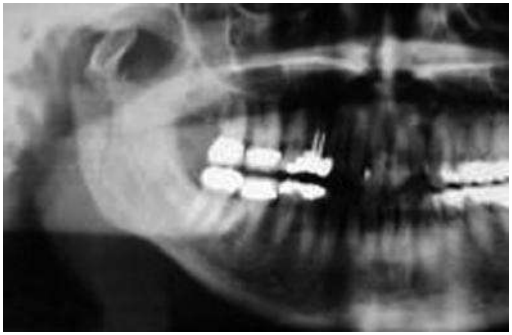

FIG. 8.6B: X-ray—Fracture of condyle of mandible

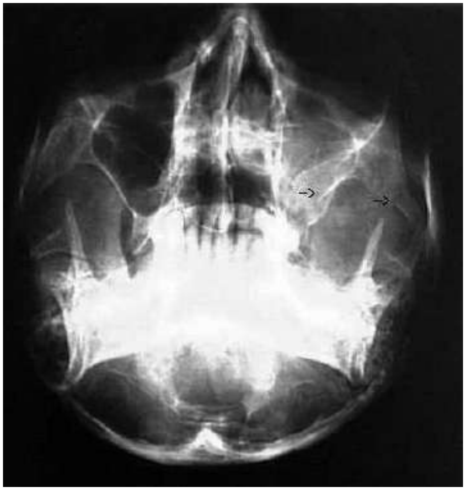

FIG. 8.6C: X-ray—Zygomatic fracture

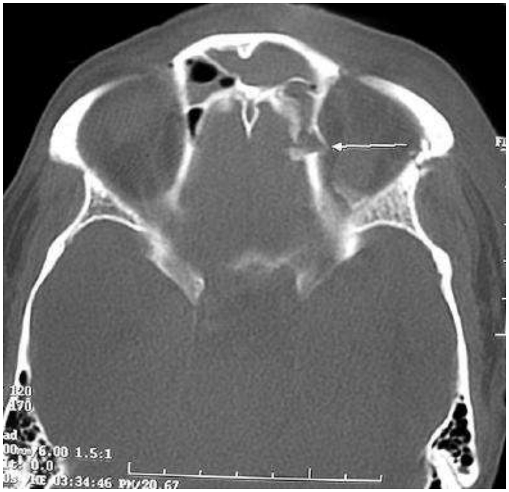

FIG. 8.7: CT—Fracture of orbital bones

- Abnormal mobility: This is present in case of impacted fractures
- Crepitus: Also present in case of impacted fractures and should be elicited gently
- Loss of function: For example, if the maxilla or the mandible is fractured, the patient may not be able to masticate properly.

Relevant Investigations

X-rays (Figs 8.6A to C) and CT (Fig. 8.7) are diagnostic.

Treatment

- Undisplaced fractures need conservative management
- Displaced fractures require surgical treatment, fixation with plates and screws (craniomaxillofascial plating) and interdental wiring, with the aim to restore precise anatomical alignment.

9

Spine and Spinal Cord Injuries

INJURIES TO BONY AND LIGAMENTOUS SPINE

Incidence and Etiology

- When an overwhelming force strikes and compresses the spine, the bone crumbles resulting in fracture of spine
- The type of the vertebral fracture depends on the direction of injuring force and the curvature of spine at the time of injury
- With increasing force of injury, two or three columns of spine get injured leading to instability and dislocation
- Generally compression forces cause fracture while distraction forces cause ligament injury
- The injuring forces can cause injury to spinal cord/root.

Symptoms

- When the spinal cord is not affected, pain on movement, stiffness and tenderness are the symptoms
- When the spinal cord is affected, neurological deficit occurs, depending on the level and completeness of damage to the cord
    - Limb paralysis like paraplegia or quadriplegia may be the presenting symptoms
    - Respiratory, circulatory and urinary bladder dysfunctions may be superadded.

Signs

Clinical examination is to determine the level and type of injury

- External injuries give clue to the mechanism of injury (e.g. chin injury may indicate hyperextension injury, occipital injury may indicate flexion-distraction injury, vertex injuries may indicate vertebral compression).

Relevant Investigations

X-rays (Figs 9.1A and B), CT (Fig. 9.2) and MRI are essential to assess the extent of injury.

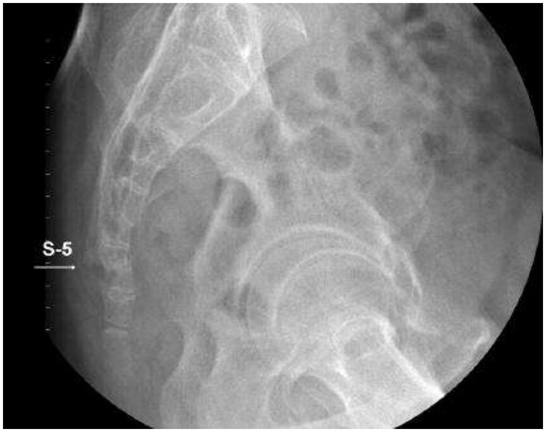

FIG. 9.1A: X-ray—Fracture of 5th sacral vertebra

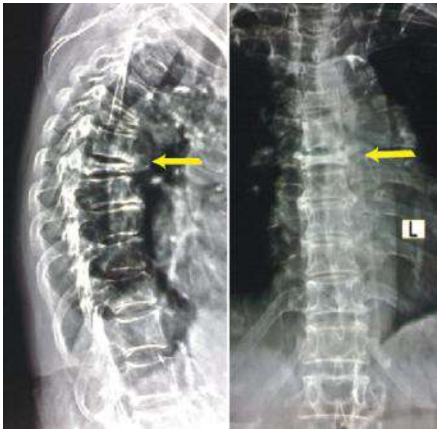

FIG. 9.1B: X-ray—Compression fracture vertebra

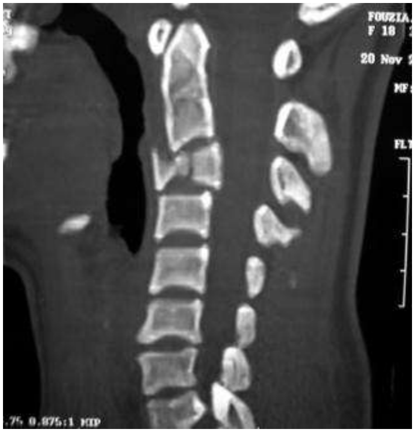

FIG. 9.2: CT cervical spine—Fracture body of C3 vertebra

Treatment

- Rest to the injured area may be sufficient with analgesic support. Many of them recover completely
- Surgical decompression and stabilization are mandatory for incomplete cord injuries. Though, the procedures are useful for bony stability in complete cord injuries, the neurologic recovery is poor.

INJURIES OF THE SPINAL CORD

Incidence and Etiology

- During the injuries of the spinal cord, the neurons suffer primary damage on impact and are prone to secondary injuries from hypoxia, hypotension, hematoma and swelling
- When the damage is severe, spinal cord function at and below the level of the lesion is abolished
- Spinal cord injuries are usually associated with injuries of vertebral column, usually fracture or fracture-dislocation
- The mechanisms of injury to the spinal cord /root are by:
    - Primary injuries: The primary injuring force itself can cause violent movement and stretching or disruption of cord with or without abnormal bony displacements
    - Secondary injuries: The fractured segments of vertebrae, extruded disk, infolded ligaments and intraspinal hematomas can cause secondary compression of cord or roots causing neurologic deficit.

Symptoms and Signs

- Paralysis occurs below the level of the lesion, e.g. thoracic cord injury leads to paraplegia and cervical cord injury leads to quadriplegia
- Flaccid paralysis with no sensation indicates complete lesion with a poor prognosis
- Signs suggestive of spinal cord injury are given in Table 9.1
- In an unconscious patient, some symptoms suggest spinal cord injury (Box 9.1).

Table 9.1: Spinal cord injuries and their signs

Complete cord injury	Total paralysis with loss of sensation below level of injury
Cord hemisection (Brown-Sequard syndrome)	Ipsilateral paralysis with contralateral loss of sensation below level of lesion
Central cord syndrome	Greater motor loss in the upper limbs than the lower limbs Variable sensory loss below level of lesion
Anterior cord syndrome	Paralysis, loss of pain/temperature sensation below level of lesion Proprioception and vibration preserved

Box 9.1: Signs suggestive of spinal cord injury in an unconscious patient

- Diaphragmatic pattern of breathing
- Unexplained hypotension
- Flaccid paralysis
- Reduced anal tone
- Urinary retention/Priapism

Relevant Investigations

X-rays and CT are useful in diagnosis.

Treatment

- The treatment should focus on maintaining stability of the vertebral column either by external or internal fixation
- Recovery is variable and rehabilitation should start early for better results
    - Establish IV access and give good volume load to support blood pressure
    - Vasopressors may be needed to maintain circulation
    - Tracheal intubation and assisted ventilation may be needed to support ventilation
    - Nasogastric intubation may be needed for gastric decompression
    - Urethral catheterization is needed.

10

Thoracic Injuries

RIB FRACTURES

Incidence and Etiology

- This constitutes the most common chest injury
- Minor fractures are those confined to one or two ribs
- Mechanism of injury
    - Upper rib injuries involve major energy transfer and are often associated with injuries to major vessels, brachial plexus and tracheobronchial tree
    - Fractures of lower ribs are frequently associated with liver and splenic injuries
    - Rib fractures in the elderly can occur after relatively low energy transfers as the bones are of low density and chest wall compliance is poor.

Fractures of brittle ribs of elderly patients cause very little underlying injuries, whereas, flexible rib injuries of younger individuals cause severe injury without obvious fractures.

Symptoms

Severe pain on deep inspiration and coughing, poor inspiratory effort, and progressive atelectasis and pneumonia due to underlying lung contusion.

Signs

- Crepitus, and bony tenderness
- Skin bruises should give the suspicion
- The hallmarks of rib fractures are intense pain, poor inspiratory effort, and progressive atelectasis and pneumonia due to underlying lung contusion.

> - Apical rib fractures are associated with injury to great vessels
> - Mid zone rib fractures are associated with pulmonary contusion
> - Basal rib fractures are associated with abdominal visceral injuries (liver, spleen).

Relevant Investigations

- Chest X-ray shows the site and number of fractures (Figs 10.1A and B), underlying pleural and lung injuries
- Chest CT (Figs 10.2A and B) gives clearer view of fractures.

Treatment

Centers around pain management:

- Oral and parenteral analgesics
- Intercostal nerve blocks
- Epidural analgesia especially in elderly or patients undergoing abdominal surgeries.

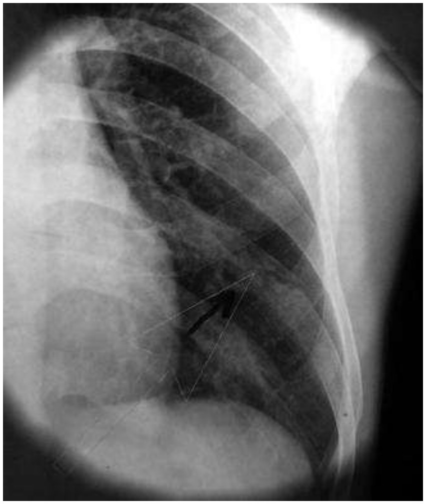

FIG. 10.1A: X-ray—Fracture rib

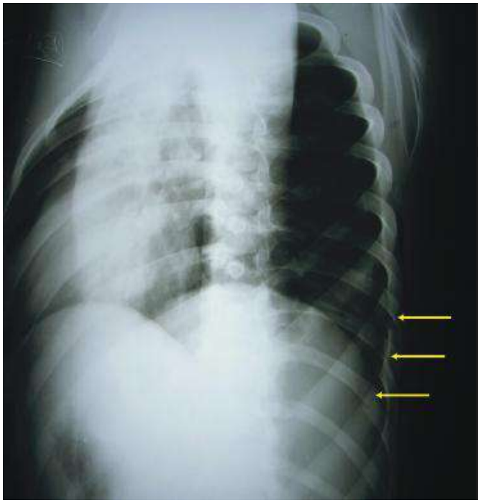

FIG. 10.1B: X-ray—Fractures of lower ribs

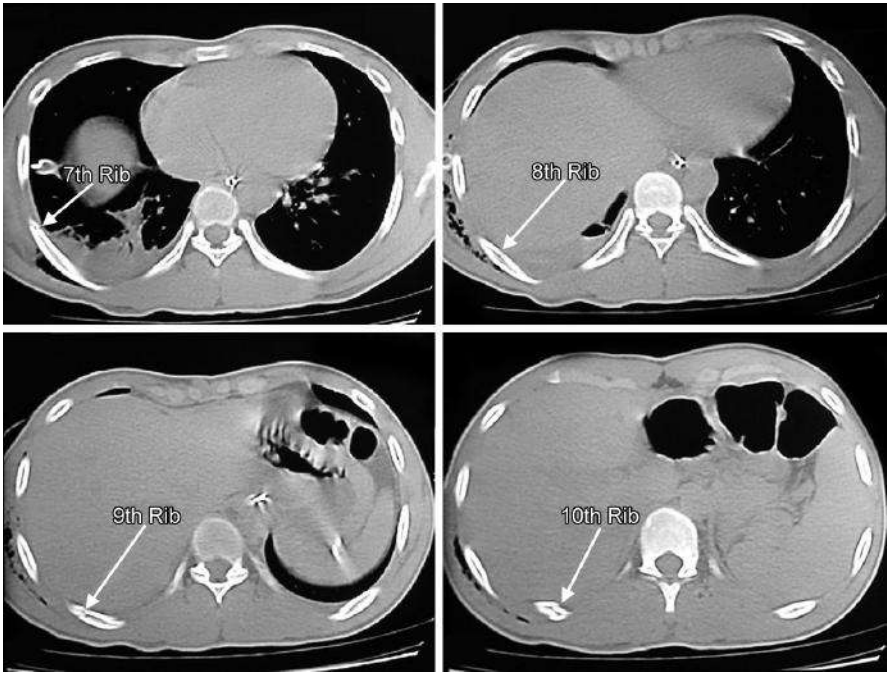
7th Rib
8th Rib
9th Rib
10th Rib

FIG. 10.2A: Chest CT—Fractures of 7, 8, 9 and 10th ribs

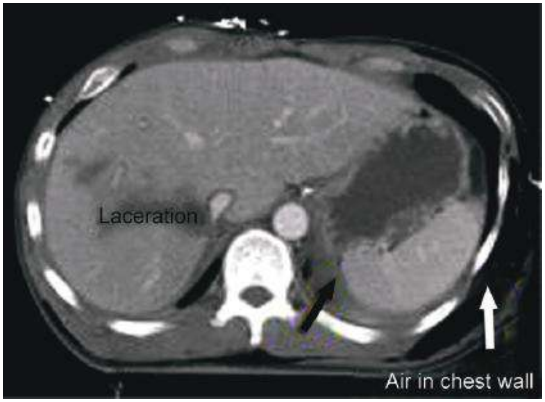
Laceration
Air in chest wall

FIG. 10.2B: CT—Laceration of liver and perisplenic collection

FLAIL CHEST

Incidence and Etiology

- When three or more ribs are fractured, each in more than one place, producing a free floating section of the chest wall with or without separation of the costochondral junction it is called a flail chest (Fig. 10.3)
- The flail segment interferes with the ventilatory function by ineffective chest wall motion (paradoxical movement) i.e. movement inward with inspiration and outward with expiration, producing pain and splinting and thereby a fall in tidal volume, hypoxia and hypercarbia.

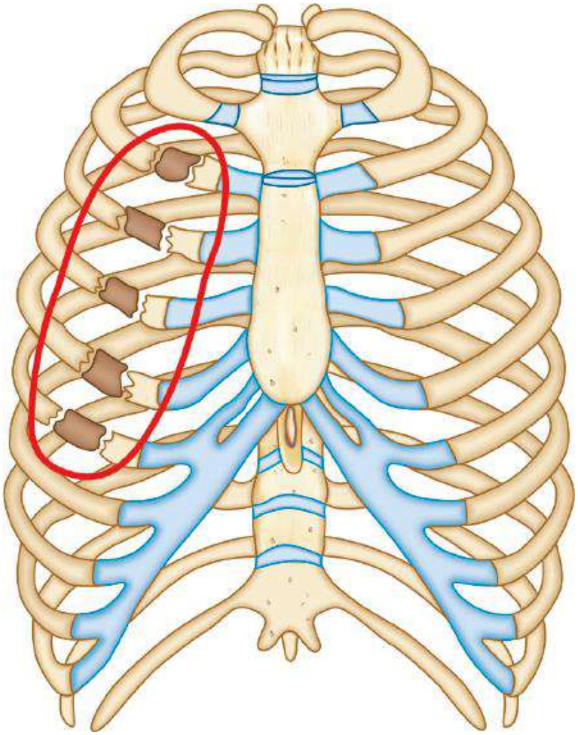

FIG. 10.3: Flail chest

- Other causes for flail chest are:
  1. Traumatic disruption of ligaments and cartilages of ribs—not seen on X-ray.
  2. Destruction of ribs from malignant disease, e.g. multiple myeloma.
  3. Metabolic disease—osteitis fibrosa cystica.
  4. Nonclosure of median sternotomy wound.

Symptom

Dyspnea.

Sign

Paradoxical respiration and hypoxia.

Relevant Investigations

- Chest X-ray—to assess fracture, lung injury, hemopneumothorax
- Arterial blood gas analysis—to aid treatment of respiratory insufficiency (ventilation perfusion mismatch).

Treatment

Treatment of flail chest is shown in Table 10.1.

Table 10.1: Treatment of flail chest

Segment of flail	*Respiratory distress*	*Respiratory function*	*Treatment*
Small	No	Good	Pain relief and observation
Moderate	Severe	Moderate	Mechanical ventilation and analgesics
Large	Severe	Bad	Chest wall reconstruction and mechanical stabilization

STERNAL FRACTURE

Incidence and Etiology

- Occurs mostly at the manubriosternal junction and is associated with very high velocity trauma
- Injury to aorta, esophagus, bronchi, myocardium and spine need to be kept in mind.

Symptom

Severe pain over the anterior chest wall.

Sign

Crepitus over the fracture site is characteristic.

Relevant Investigations

Chest X-ray lateral view and CT (Fig. 10.4) demonstrate the fracture.

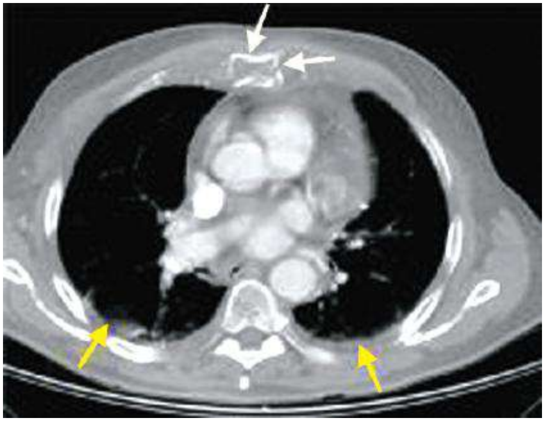

FIG. 10.4: CT—Fracture sternum

Treatment

- Sternal fractures can be managed conservatively with pain relief
- Rarely, in case of persistent chest wall instability, fixation may be necessary.

PNEUMOTHORAX

Incidence and Etiology

- Defined as air in the pleural cavity
- The types (Fig. 10.5) are:
  - *Closed pneumothorax:* Air in the pleural cavity and has no external communication (e.g. rupture of emphysematous bulla) or from outside
  - *Open pneumothorax:* Air in the pleural cavity has external communication (e.g. penetrating chest wall injury or rib fracture)

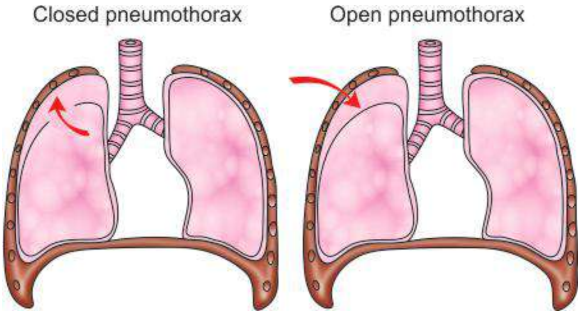

FIG. 10.5: Types of pneumothorax

- *Tension pneumothorax:* Continued entry of air into the pleural cavity, increasing the intrapleural pressure above the atmospheric pressure, which results in the shift of the mediastinum away from the side of injury.

Symptoms

Chest pain, dyspnea and tachycardia.

Signs

- On examination the neck veins are distended, the trachea and apex beat are shifted away from the side of tension, breath sounds become distant or absent on the side of tension, due to the presence of air between the chest wall and the lung substance
- The chest on the affected side is more resonant (DD—hemothorax, hydrothorax) on percussion.

Relevant Investigations

- Chest X-ray (Figs 10.6A to C) is conclusive, with shift of mediastinal structures away from the side of the pathology, with air shadow lateral to the lung parenchyma on the side of the pathology
- CT (Figs 10.7A and B) is diagnostic.

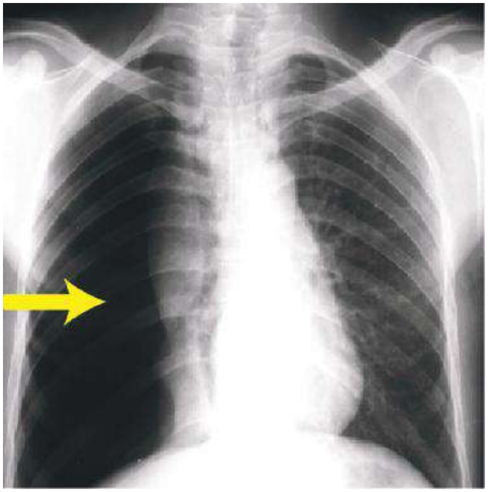

FIG. 10.6A: Chest X-ray—Right pneumothorax

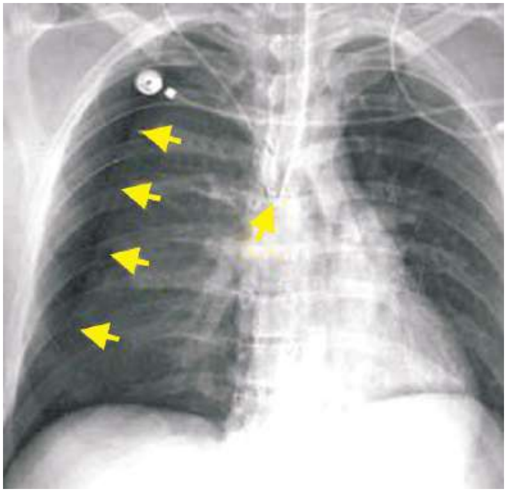

FIG. 10.6B: Chest X-ray—Right pneumothorax in patient on ventilator

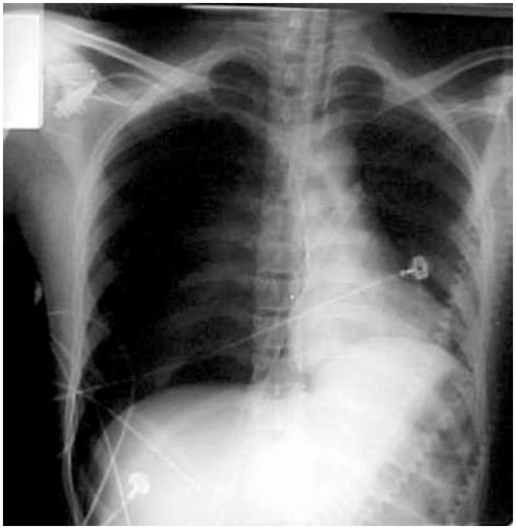

FIG. 10.6C: Chest X-ray—Tension pneumothorax

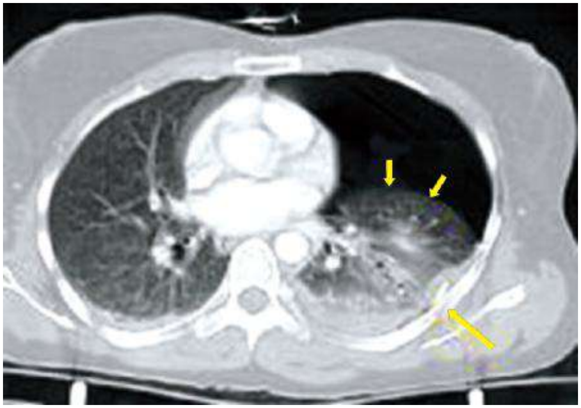

FIG. 10.7A: CT—Left pneumothorax with fractures of ribs

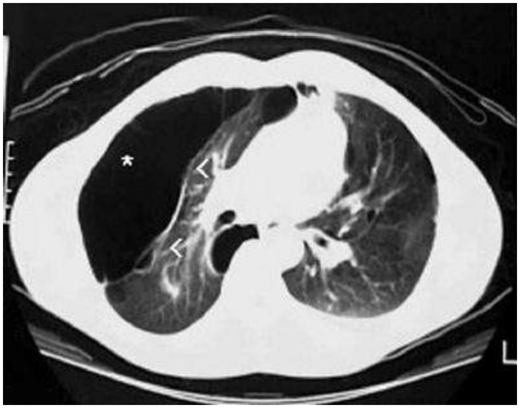

FIG. 10.7B: CT—Right tension pneumothroax

Treatment

- *Open pneumothorax:* The external wound is closed with a tape to convert it into a closed variety, supported by intercostal drainage
- *Closed and tension pneumothorax:* Simple aspiration of air from the pleural space followed by tube thoracostomy.

- Large chest wall wounds more than 75 percent the diameter of trachea allow preferential air entry through the chest wall. Any attempt to ventilate leads to movement of air in and out of the defect. No ventilation is achieved, and severe respiratory compromise occurs
- All traumatic pneumothoraces should be drained
- Massive air leaks may require bronchoscopy to exclude bronchial rupture
- Bronchial rupture should be suspected in the presence of deceleration injury, mediastinal widening, hemoptysis, first rib and clavicular fractures.

SURGICAL EMPHYSEMA

Incidence and Etiology

Defined as air in the subcutaneous tissues due to the air entry from the injured lung or external injuries like fractured rib.

Pathogenesis

The entrapped air gradually spreads along the fascial planes into the neck, mediastinum but, rarely down into the scrotum producing, a pneumoscrotum.

Symptoms and Signs

Subcutaneous tissues appear swollen and crepitus is a classical finding.

Relevant Investigations

- X-ray is diagnostic and reveals the air shadow in the subcutaneous plane, and also the underlying cause (e.g. rib fracture)
- CT (Fig. 10.8) is diagnostic.

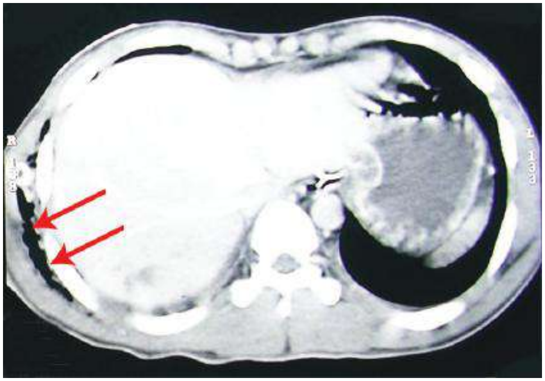

FIG. 10.8: CT—Right sided surgical emphysema with rib fracture

Treatment

- Small emphysema resolves spontaneously
- Hemodynamic instability warrants surgical intervention
- Treating the underlying cause.

HEMOTHORAX

Incidence and Etiology

Defined as blood in the pleural space, which is usually due to external (blunt or penetrating) trauma.

Symptoms

Chest pain, dyspnea and tachycardia.

Signs

- Tachycardia
- Neck veins are distended
- Trachea and apex beat are shifted away from the side of tension
- The chest on the affected side is less resonant or dull (DD—pneumothorax) on percussion
- Breath sounds become distant or absent on the side of lesion, due to the presence of blood between the chest wall and the lung substance
- Hypotension and shock may be evident depending on the amount of blood loss (bleeding from lung parenchyma is usually small but those from the intercostals and internal mammary arteries may be large).

Relevant Investigations

- Chest X-ray (Fig. 10.9A) is conclusive, with shift of mediastinal structures away from the side of the pathology, with haziness with obliteration of costophrenic angle
- CT (Fig. 10.9B) is diagnostic.

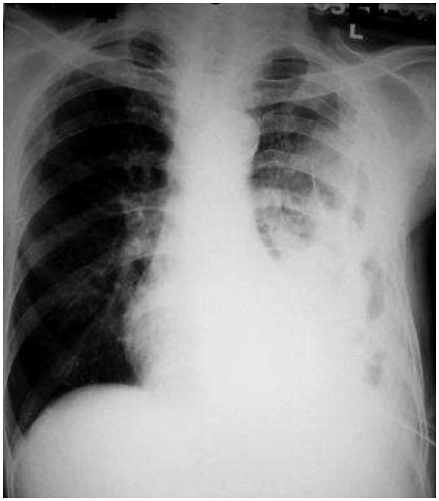

FIG. 10.9A: Chest X-ray—Left hemothorax

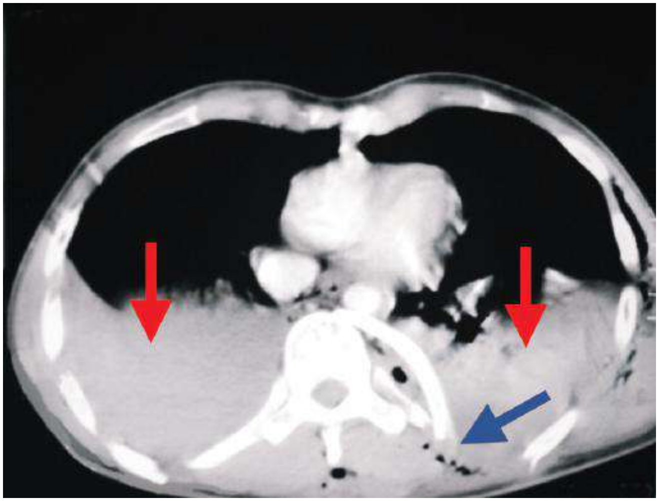

FIG. 10.9B: CT chest—Bilateral hemothorax (Red arrows) with left 10th rib fracture (Blue arrow)

Treatment

- Small collections of blood may be aspirated under aseptic conditions
- Large collections need intercostal drainage
- Thoracotomy is needed for massive bleeds (initial bleed of more than 1 – 1.5 liters), or continued bleeds of more than 200–300 ml/hr, to control the source of bleeding.

- Early drainage is the key to success in the management of hemothorax
- Once clot gets established, thoracotomy is needed
- While draining hemothorax, it is advisable to use large drains (28 – 32 F)
- Initial drainage of >600 ml or continued drainage of > 150 ml/hr will need thoracic surgical referral.

PULMONARY CONTUSION/LACERATION

Incidence and Etiology

- While contusions of the lungs are produced by blunt chest injury with hemorrhage and edema in the lung parenchyma, lung lacerations are due to penetrating injuries
- There may be associated injury to larger airways:
  - Blunt injury usually produces injuries within 2.5 cm of the carina
  - Penetrating injuries may be at any level.
- Alveolar microhemorrhages are responsible for the poor ventilatory status
- Tracheobronchial injuries may coexist.

Complications: Mediastinal emphysema in case of major airway injury and hemopneumothorax in case of peripheral bronchial injuries.

Symptoms and Signs

- Dyspnea, tachycardia and chest pain
- Small lacerations produce no symptoms.

Relevant Investigations

- Chest X-ray is diagnostic, which shows vague opacification in the injured area (usually within 1-2 hours of injury)
- CT (Fig. 10.10) and MRI are conclusive
- Bronchoscopy is needed to evaluate tracheobronchial injuries.

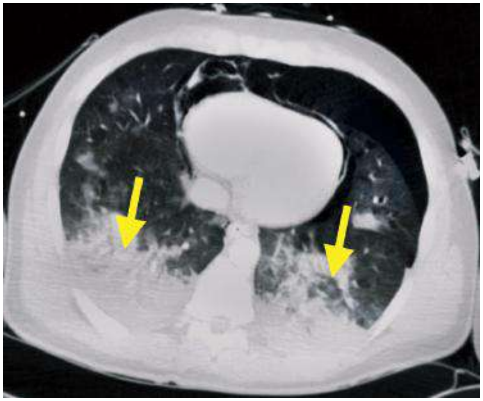

FIG. 10.10: CT chest—Bilateral pulmonary contusion

Treatment

- Most lacerations of the lung resolve spontaneously and need no treatment
- Large lacerations may require resuscitation and tube thoracostomy
- Bronchial injuries involving more than 1/3rd circumference of the bronchus require surgery
- Mechanical ventilation is warranted when there is profound V/Q mismatch.

INJURIES OF THORACIC AORTA

Incidence and Etiology

- Thoracic aorta is vulnerable for injuries at its three fixed sites:
  - Annulus
  - Ligamentum arteriosum—isthmus
  - Aortic hiatus
- The most common sites being:
  - The ascending aorta proximal to innominate artery
  - The descending aorta at the point beyond the origin of left subclavian artery (ligamentum arteriosum)
- Mechanism of injury:
  - Penetrating: More common and may involve other great vessels
  - Blunt: Rapid deceleration produces shear at the fixed portions of the aorta, or by anteroposterior compression.
- Almost 80 percent die on the way to the hospital.

Symptom

Severe respiratory distress.

Signs

- Features of massive hemothorax
- Signs of shock may be evident
- Neurological signs of cord compression may be present.

> Injuries distal to innominate artery may show pseudocoarctation syndrome (upper extremity hypertension and hypotension and low pulse in lower limbs).

Relevant Investigations

- Chest X-ray and CT (Fig. 10.11) are diagnostic with the following findings:
    - Widening of superior mediastinum
    - Depression of left main stem bronchus
    - Loss of aortic knob
    - Massive hemopnuemothorax
    - Look for associated 1st rib fracture, flail chest, sternal fracture and fracture of thoracic spine.
- Arteriography is useful for definitive diagnosis of location and extent of injury.

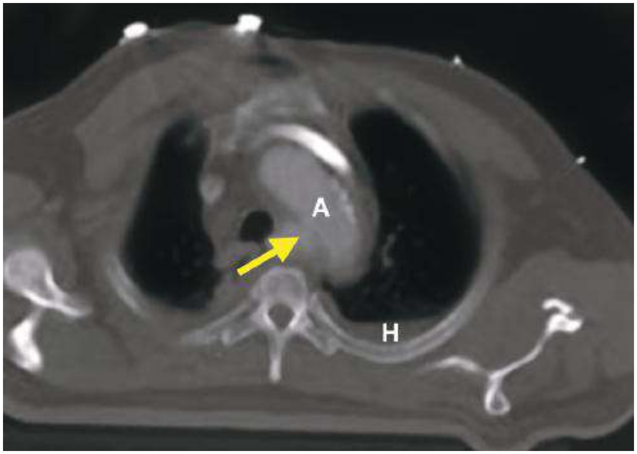

FIG. 10.11: CT chest—Aortic rupture with upper rib fractures

Treatment

- Resuscitation is the primary and emergent treatment
- Resection of damaged segment of aorta and repair/interposition grafts is curative.

INJURIES OF MYOCARDIUM

Incidence and Etiology

- Myocardial injuries are caused by:
  - Penetrating injuries (e.g. gunshot or stab injuries) where the outcome of injury depends on size of pericardial defect
  - Blunt injuries
    - Rupture into the pericardium producing pericardial tamponade
    - Myocardial contusion
    - Arrhythmias.
- Commotio cordis is the condition of sudden cardiac death or near sudden cardiac death after blunt, low-impact chest wall trauma in the absence of structural cardiac abnormality. Ventricular fibrillation is the most commonly reported induced arrhythmia in commotio cordis.
- Blunt impact injury to the chest with a baseball is the most common mechanism and does not result solely from the force of a blow as it is not seen in association with any rib or sternal fracture. It is largely the result of the exquisite timing of the blow during a narrow window within the repolarization phase of the cardiac cycle, 15 to 30 msec prior to the peak of the T wave. Survival rates for commotio cordis are low, even with prompt CPR and defibrillation.

Symptoms

Dyspnea and cyanosis.

Signs

- Examination reveals, distension of jugular veins, hypotension and narrowing pulse pressure and distant heart sounds (Beck's triad)
- The jugular venous distension raises paradoxically on deep inspiration (Kussmaul's sign), because the increased venous return cannot be accommodated within the constricted heart
- Signs of shock may be evident
- Pulsus paradoxus is a cardinal sign (drop in systolic BP >10 mm during inspiration due to CO_2 absorption).

Relevant Investigations

- Chest X-ray is contributory with cardiomegaly
- ECG may remain normal for over 36 hours but exhibit features of ischemia, dysrhythmia later
- ECHO is done to assess regional wall motion abnormalities
- CPK-MB values show elevation (check values at admission, 24–48 hours).

Treatment

- Cardiac monitoring and resuscitation are important
- Cardiac tamponade warrants pericardiocentesis/subxiphoid pericardial window
- Thoracotomy is done to create an opening of pericardial sac.

ESOPHAGEAL INJURIES

Incidence and Etiology

- Esophageal injuries are caused by:
    - *Penetrating injury* may occur at any level and are should be suspected when the injury crosses the midline (e.g. in sword swallowers as circus act), during esophagoscopy
    - *Blunt injury:* Usually following severe blow to the sternum or epigastrium. The common site of injury is at the lower 1/3rd esophagus.

Symptoms

Fever, dyspnea (due to mediastinitis or mediastinal emphysema) or tachypnea.

Signs

Features of surgical emphysema (spread of mediastinal emphysema to neck, face and chest wall) may supervene, with signs of hypoxia about 3–4 days later.

Relevant Investigations

- Chest X-ray may reveal
    - Pneumomediastinum
    - Air in the prevertebral space
    - Left pleural effusion
    - Hemo or pneumothorax in the absence of rib fracture.

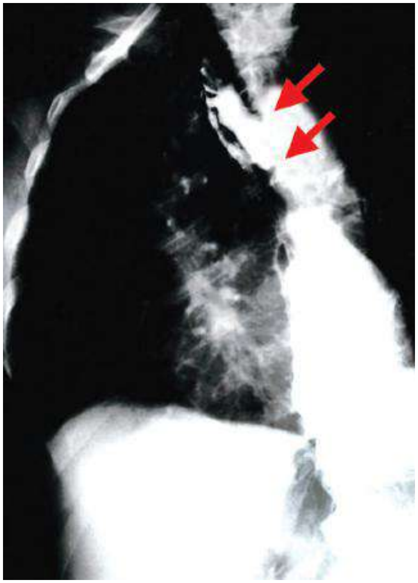

FIG. 10.12: Gastrograffin swallow—Leaking dye in esophageal perforation (penetrating injury)

- Gastrograffin swallow may show the leak (Fig. 10.12)
- Esophagoscopy may show the injury
- Intercostal drainage shows particulate food matter, and show air leak during both phases of respiration.

Treatment

- Intercostal drainage is mandatory
- Early operative repair of the esophageal tear is necessary.

DIAPHRAGMATIC INJURIES

Incidence and Etiology

- Injuries of diaphragm are caused by:
    - Blunt injuries produce large radial tears and herniation of abdominal viscera into the chest
    - Penetrating injuries are small initially and enlarge over a period of time.

Symptoms and Signs

Diagnosis is difficult unless the tears are large and allows herniation of abdominal contents into the chest cavity—dyspnea and tachycardia.

Relevant Investigations

- Chest X-ray is contributory and may show:
  - Bowel loops in the chest (in herniation of small bowel)
  - Double shadow overlying the diaphragm
  - Nasogastric tube in the chest—in patients with gastrothorax.
- Contrast studies and CT of chest are conclusive.

Treatment

Surgical repair with or without mesh is mandatory.

11

Abdominal Injuries

Abdominal injuries can be of two varieties. They are:

1. Closed injuries (caused by blunt injuries)
2. Open injuries (caused by penetrating or stab injuries).

CLOSED INJURIES

- Caused by a blunt force exerted suddenly on the abdomen, such as:
    - Fall from a height
    - Blow with a fist
    - Injuries inflicted by heavy weapons like crowbars, poles, sticks
    - Run over injuries in road traffic accidents
    - Hitting against the steering wheel during sudden deceleration or braking of a speeding vehicle (in the absence of seat belt).
- They may lead to compression of intra-abdominal organs against the vertebral column causing rupture of:
    - Solid organs
    - Hollow organs
    - Mesentery
    - Detachment of gut from the mesentery
    - Contusion of abdominal wall.

OPEN INJURIES

- Caused by any sharp instrument like knife, flying objects like bullets, missiles, pieces of wood or glass.
- The incriminating agents enter the abdominal cavity taking with them some infection resulting in peritonitis.
- The points of entry and exit of the agent will indicate the direction and the possible organs injured.
  - Hollow organs may perforate and infect the peritoneal cavity (due to contamination by the contents of the organ—feces, urine, intestinal or gastric contents), or bleed (either into the organ or outside or both).
  - Solid organs cause hemorrhage (inside the organ or outside or both), and large bleeds cause shock and sometimes death.

Symptoms and Signs

- General features:
  - *Symptoms:* Weakness, clouding of thoughts and speech, air hunger, restlessness, decreased or absent urine output
  - *Signs:* Pallor, tachycardia, hypotension, subnormal temperature.
- Local features: Depend on the organ involved and the severity of damage.

INJURIES OF LIVER

Incidence and Etiology

The liver ranks high on the list of intra-abdominal organs involved by injury:

- Blunt injuries are more common than the penetrating injuries, due to increase in motor traffic moving at high speeds, and are associated with fracture of lower ribs on the right side. The dome of the liver is involved with anterior-posterior tears, more on the right lobe (7:1).
- Spontaneous rupture of liver is seen in:
  - Primary carcinoma in adults
  - Trauma during birth in children (postmature babies) being delivered per vaginum.

Liver injuries (Fig. 11.1) are classified into:

- Transcapsular (blood and bile will seep into the peritoneal cavity)
- Subcapsular (collection of blood between the capsule and the liver parenchyma mostly on the superior surface of liver)
- Central (interruption of liver parenchyma leading to intrahepatic hematoma, abscess and hematobilia).

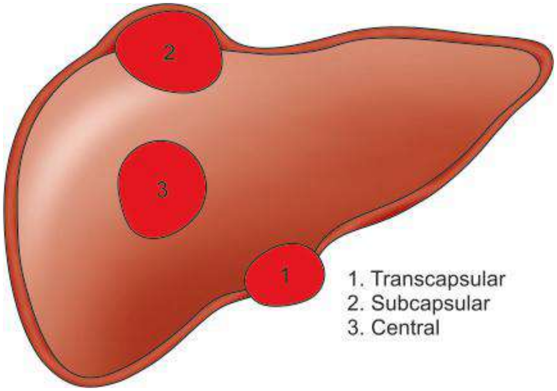

FIG. 11.1: Types of liver injuries

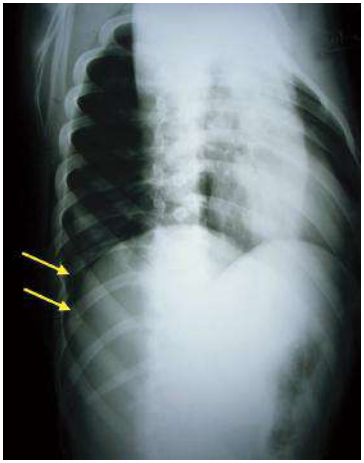

FIG. 11.2: X-ray showing lower rib fracture

Symptoms and Signs

- *Transcapsular:* Symptoms and signs of shock and peritoneal irritation (pain right upper abdomen with reference to the right shoulder, guarding and rigidity of the right hypochondrium, absent bowel sounds, shifting dullness)
- *Subcapsular:* Local tenderness and increase in area of liver dullness
- *Central:* Signs of shock may be present with hemobilia and hematemesis.

Relevant Investigations

- Plain chest X-ray (Fig. 11.2) will demonstrate fracture of lower ribs on the right side
- Plain X-ray abdomen may show haziness in the area of the liver with elevation of right dome of diaphragm
- CT (Figs 11.3A and B) and MRI are useful in localizing the damaged areas of liver and collections of blood or bile
- Peritoneal tap may be useful in identifying bile leaks
- Colloidal gold Au^{198} or Technitium-sulfur colloid Tc^{99m} scans are useful during active bleeding.

Treatment

- Correction of shock
- No surgical intervention is needed for small injuries

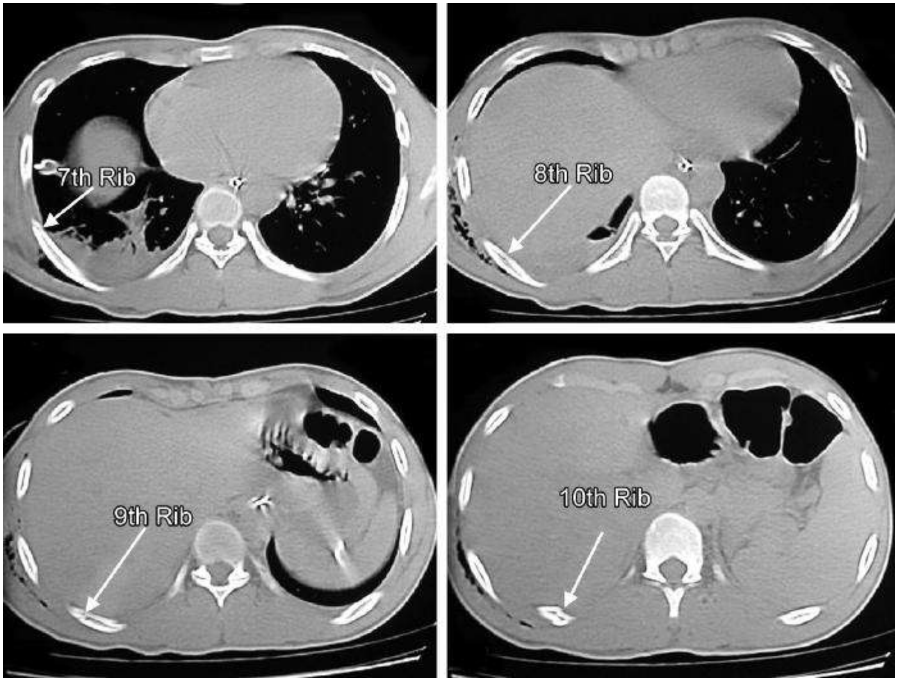
7th Rib
8th Rib
9th Rib
10th Rib

FIG. 11.3A: CT—Right lower rib fractures

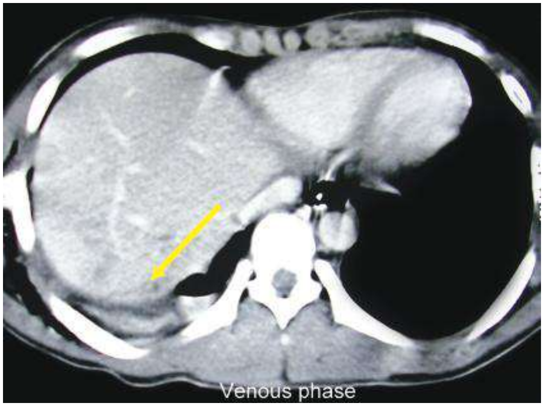
Venous phase

FIG. 11.3B: CT—Liver injury with hemoperitoneum

- Early surgical intervention, sometimes amounting to hepatectomy, is needed for large injuries associated with vascular or biliary tract injuries.

- Post-traumatic liver hemorrhage is amenable to radiological intervention (embolization).

INJURIES OF SPLEEN

Incidence and Etiology

- Spleen is the intra-abdominal organ most frequently injured by blunt trauma, usually by thoraco abdominal injuries associated with fractures of left lower ribs.
- The causes of splenic injuries are:
    - Blunt injuries of lower chest and abdomen (e.g. automobile accidents—may be associated with rib fractures, lung injuries, fracture of spine, intra-abdominal organs)
    - Transabdominal and transthoracic penetrating injuries (e.g. gunshot or knife injuries—may be associated with injuries to left kidney, colon, pancreas, vascular structures of mesentery)
    - Operative injuries (e.g. during operations on adjacent viscera—gastrectomy)
    - Spontaneous rupture or minor trauma (e.g. enlarged spleen of malaria).
- Splenic injuries may be:
    - Linear or stellate lacerations (due to excessive traction on the capsule during operations)
    - Capsular tears (due to excessive traction on the capsule during operations)
    - Subcapsular hematoma (due to excessive traction on the capsule during operations)
    - Puncture wounds (caused by penetrating injuries)
    - Intrasplenic hematomas (caused by penetrating injuries)
    - Rupture of parenchyma (due to extreme friability and vascularity of the organ).

Symptoms

Symptoms and signs of shock and peritoneal irritation.

Signs

- Pain left upper abdomen with reference to the left shoulder (Kher's sign)
- Guarding and rigidity of the left hypochondrium
- Absent bowel sounds
- Shifting dullness

- A mass of percussible area of fixed dullness in the left hypochondrium (Ballance's sign)
- Lacerations may be seen on the left lower chest.

> - The clinical presentation of splenic injury is of three types:
>   1. *Acute:* Immediate presentation of symptoms
>   2. *Delayed:* A quiescent period of 7 days to less than 2 weeks, between the injury and intraperitoneal bleeding (Latent period of Baudet), probably related to a temporary tamponade of a minor laceration or the presence of a slowly enlarging subcapsular hematoma which eventually ruptures
>   3. *Occult:* Mild symptoms caused by rupture of organized intrasplenic or parasplenic hematoma.

Relevant Investigations

- Plain chest X-ray chest will demonstrate fracture of ribs on the left side
- Plain X-ray abdomen may show:
  - Haziness in the area of the spleen
  - Elevated immobile left dome of diaphragm
  - Medial displacement of gastric shadow with indentation
  - Obliteration of psoas shadow
  - Widening of space between splenic flexure and the preperitoneal pad of fat.
- CT (Figs 11.4A to C) and MRI are useful in localizing the damaged areas of spleen and collections of blood
- Peritoneal tap may be useful in identifying blood in the peritoneal cavity.

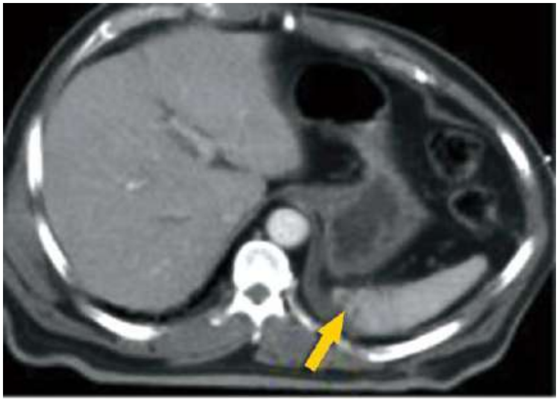

FIG. 11.4A: CT—Splenic injury with lower rib fracture

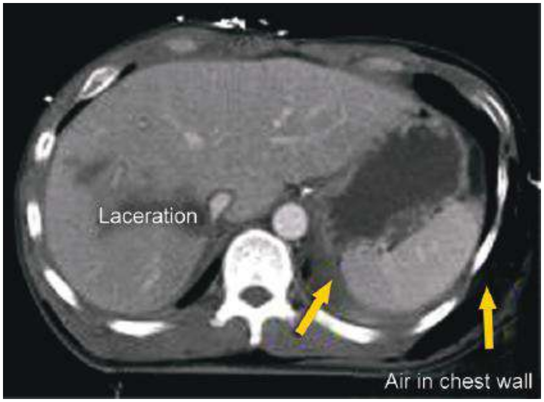

FIG. 11.4B: CT—Laceration of liver and perisplenic collection

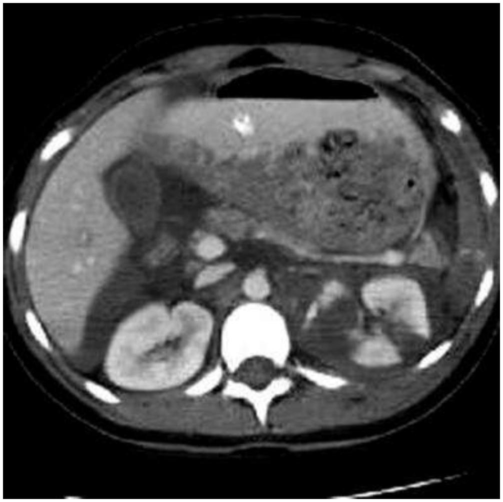

FIG. 11.4C: CT—Combined renal and splenic injuries of automobile accident

Treatment

- Once the diagnosis is made the operation should not be delayed
- Splenectomy is the recommended treatment, regardless of the type and the extent of the injury.

- Following splenectomy, to prevent the life-threatening infections (pneumococci), long-term prophylaxis with penicillin (minimum of 2 years), with immunization against *Pneumococcus*, *Meningococcus* and *Hemophilus influenzae* is required.

INJURIES OF MESENTERY

Incidence and Etiology

- Mesentery consists of its arteries and veins and they may be injured by either penetrating or non-penetrating abdominal trauma
- In most cases, associated organ injuries are found
- Isolated injury to mesenteric vessels is rare.

Symptoms

Depending on the size of the vessel lacerated, the rapidity of bleeding and associated organ injury, the patient will present with signs of shock, abdominal pain and distension.

Sign

On examination, tenderness is marked.

> When the bleeding occurs within the layers of mesentery, the clinical signs evolve slowly and the viability of the bowel is threatened.

Relevant Investigations

- Plain X-rays may show air under the diaphragm when the bowel is perforated by the injury
- Peritoneal tap may be useful in establishing the bleeding.

Treatment

- Repair of the torn mesentery is required
- Resection and anastomosis is required for nonviable intestine.

INJURIES OF DUODENUM

Incidence and Etiology

- Duodenum can be injured both by penetrating and nonpenetrating abdominal trauma
- The duodenum can rupture:
  - Intraperitoneally and cause immediate chemical irritation of the peritoneum due to the highly alkaline duodenal content
  - Retroperitoneally (more common with blunt trauma such as steering wheel injuries).

Symptoms and Signs

- Intraperitoneal rupture:
  - Abdominal pain, fever and distension with vomiting
  - Examination shows marked tenderness in the upper abdomen and later signs of generalized peritonitis.
- Retroperitoneal rupture:
  - Pain in the epigastrium and back, with pronounced vomiting
  - Testicular pain is a common feature of retroperitoneal rupture of duodenum.

Relevant Investigations

- Plain X-ray of abdomen may show air under the domes of diaphragm (intraperitoneal rupture) or large accumulation of air above the right kidney (retroperitoneal rupture). Diagnostic accuracy can be increased

by injecting air in the Levine's tube to increase the air collections. Water soluble dye injections in the tube can make the diagnosis more precise
- CT with contrast is conclusive
- Paracentesis may show bile-stained fluid, if the rupture is intraperitoneal.

Treatment

- Simple suturing may be adequate in many cases
- Supplementary gastroenterostomy is required for large tears
- Rarely, even a pancreatoduodenectomy may be necessary for extensive trauma involving the periampullary region.

INJURIES OF SMALL INTESTINE

Incidence and Etiology

- Injuries to the small intestine are more common than injuries to the duodenum and large intestine, the usual mechanism being the blunt trauma crushing the bowel against the vertebral column, more commonly the duodenojejunal flexure and the ileocecal junction, the fixed parts
- Blunt injuries cause slow necrosis of bowel and leak occurs late
- Signs and symptoms develop late—2 to 3 days later, depending on the size of the damage and leak of contents
- Penetrating injuries can also cause small bowel trauma, but less commonly, probably due to its sliding away from a knife because of its great mobility
- Associated mesenteric tears are common.

Clinical Presentation

- Abdominal pain, distension and vomiting
- Tenderness and guarding are pronounced around the damaged bowel and the patient may point it (Pointing sign).

Relevant Investigations

- Plain X-rays may show air under the domes of the diaphragm
- Paracentesis will show bile-stained fluid.

Treatment

- Simple suturing is done for simple tears
- Resections are required for large tears with nonviable bowel
- Peritoneal toileting is mandatory under cover of antibiotics.

INJURIES OF LARGE INTESTINE

Incidence and Etiology

- Large bowel injuries can be caused by penetrating and nonpenetrating injuries
- Ruptures may be:
  - Extraperitoneal for ascending and descending colon injuries, whereas
  - Intraperitoneal, when the injuries are of transverse and sigmoid colons
- Patient develops septic complications quickly as the large fluid leak is fecal and infected
- Delayed presentation is not uncommon as in small bowel injuries

Symptoms

- Abdominal pain, vomiting and distension
- High grade fever occurs due to fecal contamination.

Signs

Signs of peritonitis.

Relevant Investigations

- Plain X-rays may show air under the domes of the diaphragm (intraperitoneal ruptures)
- Paracentesis will show feculent fluid.

Treatment

- Early laparotomy is required
- Closure of tears with proximal diversion is necessary
- Peritoneal toileting is mandatory under cover of broad-spectrum antibiotics.

ABDOMINAL COMPARTMENT SYNDROME

Incidence and Etiology

Occurs when the intra-abdominal pressure rises above the venous pressure.

This can be caused by:

- Blood
- Free gas
- Free fluid
- Tissue or splanchnic edema
- Organomegaly.

Pathogenesis

Occurs due to underperfusion of intra-abdominal organs (e.g. gut, kidneys and liver).

Symptom

Falling urinary output.

Signs

Tense and quiet abdomen.

Relevant Investigations

Measurement of intra-abdominal pressure by connecting the urinary catheter to a pressure transducer (>20 mm Hg suggests abdominal compartment syndrome).

Treatment

Surgical exploration

- Patients may bleed torrentially when the abdomen is opened followed by hemodynamic instability, following visceral reperfusion
- Assisted ventilation may be required for many days.

12

Urological Injuries

RENAL INJURIES

Incidence and Etiology

- Kidneys are the most common organs to get injured in the urological system
- Injuries to the kidneys may be caused by:
    - Closed blunt abdominal trauma such as road traffic accidents, during active sports and may be associated with lower rib fractures
    - Penetrating injuries by bullets, knives are also causes of renal trauma
- Renal trauma is of three types:
    - Minor trauma (e.g. renal contusion, subcapsular hematoma)
    - Major trauma (e.g. deep cortical lacerations, extravasation of urine)
    - Renal vascular trauma (e.g. avulsion of renal pedicle).

Symptom

Hematuria is the most common symptom.

Signs

- Bruising over the ribs posteriorly or evidence of penetrating injury
- Tenderness and guarding in the loin and expanding mass
- Signs of shock may be seen in major trauma.

Relevant Investigations

- Plain X-rays may show fractures of lower ribs
- CT is the investigation of choice and accurate assessment is possible (Figs 12.1 and 12.2)
- Renal arteriography may be needed in cases where CT is not contributory, especially with renal vessel injuries.

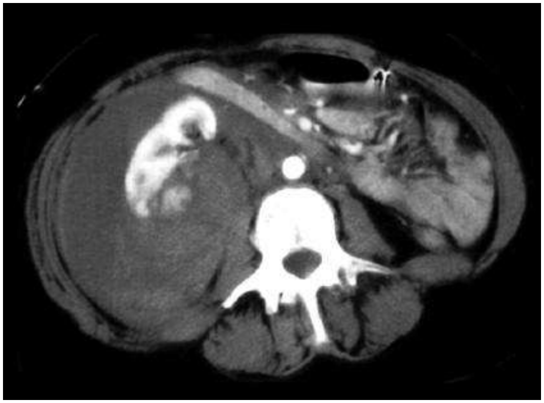

FIG. 12.1: CT—Injury of right kidney

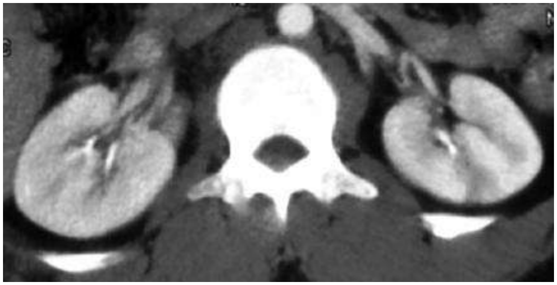

FIG. 12.2: Contrast enhanced CT—Left renal injury

Treatment

- Any patient with renal injury should rest in bed and all urine samples should be examined for blood
- Surgical exploration is warranted in closed injuries, when there is retroperitoneal bleeding, expanding loin mass, urinary extravasation, evidence of nonviable renal parenchyma and renal pedicle injuries
- Exploration of abdomen is required for penetrating injuries for assessment of other injuries and kidney should only be explored when the condition warrants.

URETERIC INJURIES

Incidence and Etiology

- Most common cause of ureteric injuries is during abdominal and pelvic operations
- Urinary leak occurs around the 5th postoperative day
- The operations associated with ureteric injuries are given in Table 12.1.

Table 12.1: Causes of ureteric injuries

Operations		
Gynecology	*General surgery*	*Urology*
Hysterectomy (abdominal or vaginal) – open or laparoscopic	Sigmoid colectomy	Excision of bladder diverticula
Ovarian cystectomy	Abdominoperineal resection	Ureterolithotomy
Anterior colporrhaphy	Surgery for aortic aneurysm	Ureteroscopy

Nature of Injuries

- Complete ligation of one or both ureters, leads to increase in intraureteric pressure, kidneys stop secreting urine, with resultant dilatation of pelvis and atrophy of kidney
- When the obstruction is incomplete (inclusion in a stitch), secretion continues to be normal, hydronephrosis and infection occur
- When the ureter is divided or suffers crushing injury, urine leaks into the retroperitoneum or peritoneal cavity which may result in a urinary fistula
- The injury may be recognized at the time of surgery
- If bilateral ligation is not recognized during surgery, it manifests as uremia.

Symptom and Sign

The clinical presentation varies according to the nature of injury (Table 12.2).

Table 12.2: Ureteric injuries and their clinical presentations

	Nature of injury	*Clinical presentation*
Symptoms	Bilateral ligation	Anuria in the immediate postoperative period
	Unilateral ligation	No symptoms/loin pain/fever
Signs	Division	Urine leak through the drain/wound/vagina
	Retroperitoneal urine leak	Abdominal distension and paralytic ileus
	Intraperitoneal urine leak	Abdominal distension and paralytic ileus with peritonitis and sepsis

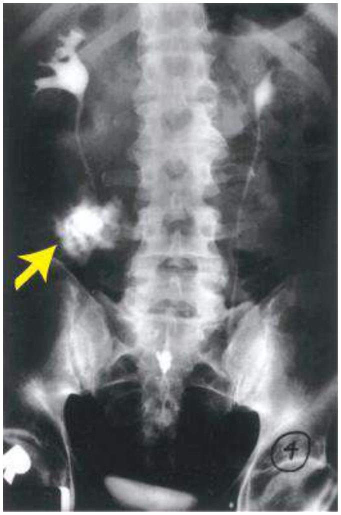

FIG. 12.3: IVU—Leakage of contrast from damaged ureter

Relevant Investigation

IVU (Fig. 12.3) or contrast enhanced CT will demonstrate the nature of obstruction and urinary leak.

Treatment

- If injury is recognized during surgery, it should be rectified immediately:
  - Spatulation and end-to-end anastomosis is done, when the length is not lost
  - Incomplete damages to the ureter may be treated by stenting and spontaneous healing may be expected.
  - Surgical corrections (ureteroureterostomy, ureteroneocystostomy) are necessary to restore the anatomy, before renal function is compromised.

URINARY BLADDER INJURIES

Incidence and Etiology

- Urinary bladder is partially extraperitoneal and is also closely covered by the parietal peritoneum, and because of this anatomy, it can rupture:
  - Extraperitoneally (80%)
  - Intraperitoneally (20%)

15

Hand Injuries

Incidence and Etiology

- Hand injuries form nearly one-third of all injuries to human body
- They can be:
  - Open
  - Closed
- Acute injuries can be classified depending on the nature of force and on contamination as tidy, untidy and indeterminable injuries:
  - *Tidy injuries:* Clean cuts and incised wounds, usually resulting from sharp force
  - *Untidy injuries:* Ragged skin wounds, may be with associated fractures, usually resulting from crush or avulsion force
  - *Indeterminable injuries:* Injuries where it is difficult to determine the extent of injury and the viability of tissues, usually caused by very severe force
- Injuries to the vessels, nerves, bones and joints may be present in any of the above injuries.

Complications: Post-traumatic deformities like contractures, neurovascular deficit.

Clinical Presentation

The patient presents to the surgeon either in the acute stage or late.

- In acute injuries (Figs 15.1A to C), the pain is excruciating, and assessment is difficult, unless pain is alleviated by analgesics. When patients arrive late, infection gets superadded, and result in post-traumatic sequelae.

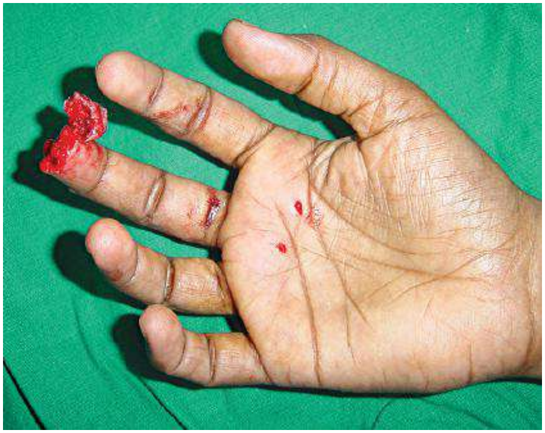

FIG. 15.1A: Traumatic amputation of tip of middle finger

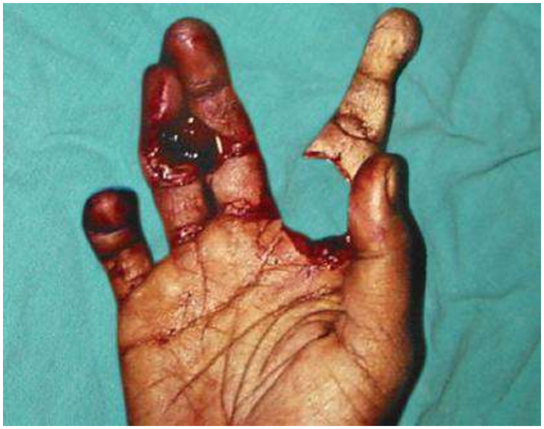

FIG. 15.1B: Traumatic amputation of entire index finger

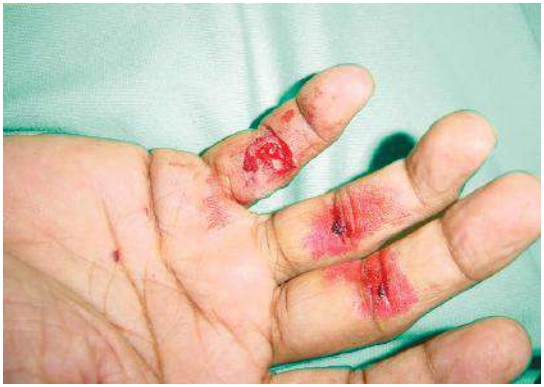

FIG. 15.1C: Penetrating injury of three fingers

- *Nail injury:* Avulsion of nail (complete or partial) (Fig. 15.2A)
- *Skin damage:* Loss of skin, especially degloving
- *Nerve damage:* Sensory or motor loss depending on the nerve affected
- *Damage to arteries:* Suspected by profuse hemorrhage (incomplete tears), gangrene (complete injuries) (Figs 15.2B to D)
- *Damage to tendon sheaths and tendons:* Lack in active movement
- *Damage to bones and joints:* Pain, deformity, tenderness and loss of movement.

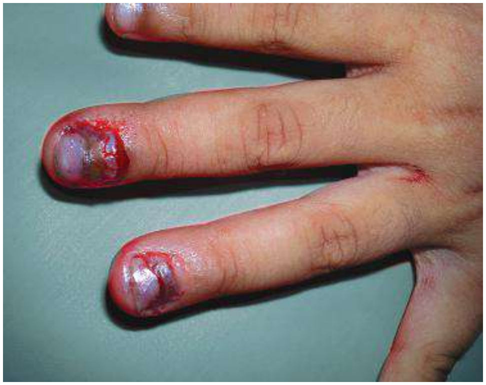

FIG. 15.2A: Avulsion of nails by crush injury

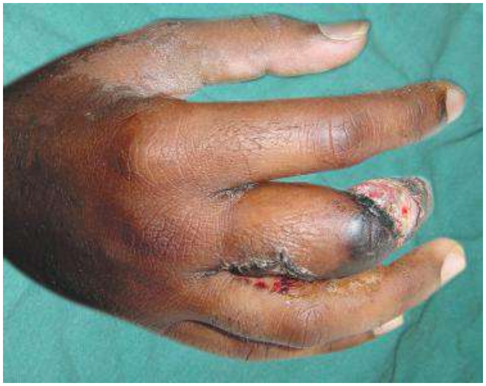

FIG. 15.2B: Traumatic gangrene of middle finger

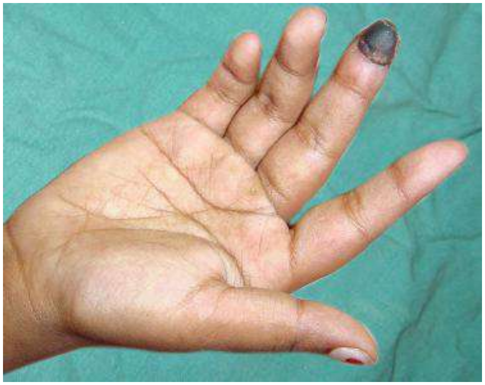

FIG. 15.2C: Gangrene of fingertip following crush injury

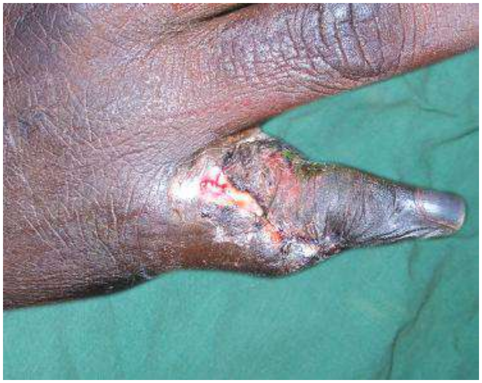

FIG. 15.2D: Traumatic gangrene of little finger

> In acute injuries, for assessment of damage, surgical exploration may be necessary when in doubt and it should be done meticulously. The structures should be considered divided until otherwise proved.

- In late post-traumatic cases detailed examination of movements and sensation can be done with ease, and damage assessed completely.
  - *Skin injuries:* Necrosis of skin, may manifest as raw area, when very late, can present with contracture due to secondary healing.
  - *Vascular injuries:* Necrosed skin and deeper tissues

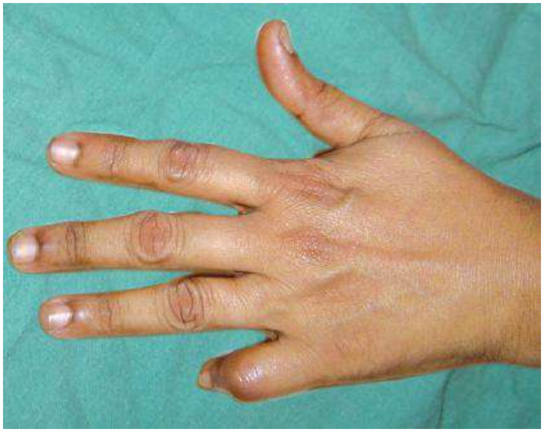

FIG. 15.3A: Boutonnière deformity of little finger

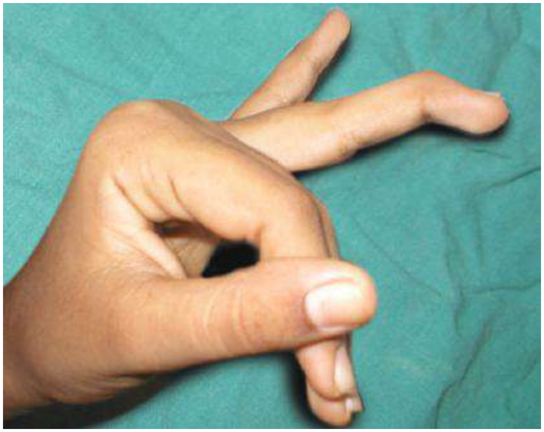

FIG. 15.3B: Mallet finger

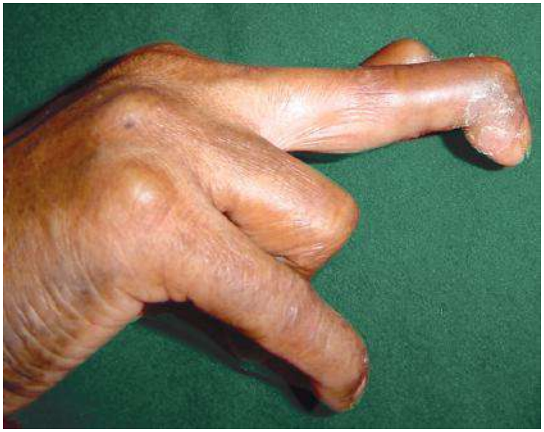

FIG. 15.3C: Swan-neck deformity

- *Extensor tendon injuries:* Dropped finger (if injury is at dorsum or proximal level), Boutonnière deformity (when injury is at proximal finger level) (Fig. 15.3A) and Mallet finger (when injury is at distal finger level) (Fig. 15.3B). Neglected or untreated Mallet fingers may progress to Swan-neck deformity (Fig. 15.3C).

  - *Flexor tendon injuries:* Lack of movement of PIP joint suggest flexor digitorum sperficialis injury and DIP joint suggests flexor digitorum profundus injury
  - *Nerve damage:* Wasting of muscle groups, lack of sweating and trophic changes
  - *Bone or joint injuries:* Fractures with or without dislocations of joints.

Relevant Investigations

- X-rays (in various views) are necessary to assess bone injuries
- Doppler studies are required to assess vascular damage.

Treatment

- Bleeding can be controlled by elevating the limb or by applying pressure directly over the site with pads
- In acute injuries:
  - *Skin injuries:* Primary repair wherever possible
  - *Vascular injuries:* Primary microvascular repair
  - *Bone and joint injuries:* Proper debridement and splinting, reduction of dislocations and immobilization
  - *Nerve injuries:* Primary repair under magnification where possible
  - *Tendon injuries:* Primary repair in clean wounds.
- In late post-traumatic cases:
  - *Skin injuries:* Wounds healing by secondary intention cause severe fibrosis and contractures, and may need release and repair and skin grafting in some cases
  - *Vascular injuries:* Grafts wherever required
  - *Bone and joint injuries:* Malunion or nonunion should be treated accordingly
  - *Nerve injuries:* Nerve grafts
  - *Tendon injuries:* Tendon transfers or grafts.

Note: Dirty wounds should be debrided well and converted into a clean wound and further treatment is carried out.

Section V

Organs and Systems

16. Oral Cavity
    - Acute Odontogenic Infections
17. Head and Neck
    - Acute Suppurative Sialadenitis
    - Ludwig's Angina
18. Thorax
    - Acute Mediastinitis
    - Acute Pleuritis
    - Pleural Effusion
    - Acute Empyema Thoracis
    - Spontaneous Pneumothorax
    - Foreign Bodies in the Respiratory Tract
    - Suppurative or Aspiration Pneumonia
    - Hospital Acquired Pneumonia
    - Acute Lung Abscess
    - Pulmonary Embolism
19. Breast
    - Breast Hematoma
    - Acute Breast Abscess
20. Spine
    - Degenerative Diseases of Disk and Facet Joints
    - Spondylolisthesis
21. Gastroenterology
    - Acute Abdomen
        - Acute upper abdominal pain
        - Acute lower abdominal pain
    - Acute Liver Abscess
    - Acute Cholecystitis
    - Acute Cholangitis
    - Primary Sclerosing Cholangitis
    - Splenic Abscess
    - Acute Hemorrhagic Pancreatitis
    - Acute Appendicitis
    - Acute Mesenteric Lymphadenitis
    - Acute Colonic Diverticulitis
    - Acute Meckel's Diverticulitis
    - Acute Solitary Cecal Diverticulitis
    - Acute Ulcerative Colitis
    - Acute Intestinal Obstruction
    - Gallstone Ileus
    - Acute Intussusception
    - Swallowed Foreign Bodies
    - Perforated Bowel Pathologies

- Intestinal Strictures
- Bands and Adhesions
- Enteroliths/Food Bolus
- Volvulus
  - Sigmoid volvulus
  - Cecal volvulus
  - Midgut volvulus
  - Gastric volvulus
- Intestinal Obstruction due to Herniae (Internal and External)
- Paralytic Ileus
- Torsion of Mesenteric Cyst
- Torsion of Omentum
- Colics
- Gastrointestinal Hemorrhage

22. Anorectum
    - Acute Anal Fissure
    - Anorectal Abscess
    - Hemorrhoids
    - Perianal Hematoma
    - Prolapse of Rectum

23. Vascular System
    - Acute Limb Ischemia
    - Acute Intestinal Ischemia
    - Leaking or Dissecting Aortic Aneurysm

24. Urology
    - Acute Retention of Urine
    - Hematuria
    - Renal Colic
    - Ureteric Colic
    - Acute Urethritis
    - Acute Prostatitis
    - Acute Prostatic Abscess

25. Male Genitalia
    - Acute Scrotal Pain
    - Torsion of Testis
    - Torsion of Appendages of Testis
    - Acute Epididymo-orchitis
    - Traumatic Orchitis
    - Hematocele
    - Pyocele
    - Idiopathic Scrotal Edema
    - Acute Scrotal Abscess
    - Fournier’s Gangrene
    - Acute Filarial Scrotum
    - Fracture Penis
    - Paraphimosis
    - Priapism

26. Hernias
    - Complicated Hernias

27. Gynecology
    - Acute Torsion of Ovarian Cyst
    - Acute Salpingitis
    - Rupture of Lutein Cyst

28. Pediatrics
    - Acute Intussusception
    - Congenital Pyloric Stenosis
    - Necrotizing Enterocolitis
    - Tracheoesophageal Fistula

29. Lymphatic System
    - Acute Lymphangitis
    - Acute Viral and Bacterial Lymphadenitis
    - Acute Filarial Lymphangitis and Lymphadenitis

30. Skin and Subcutaneous Tissues
    - Hematoma
    - Erysipelas
    - Furuncle
    - Cellulitis
    - Acute Pyogenic Abscess
    - Carbuncle
    - Burns
    - Necrotizing Fasciitis

16

Oral Cavity

ACUTE ODONTOGENIC INFECTIONS

Incidence and Etiology

Acute odontogenic infections present in three ways:

i. Abscess
ii. Cellulitis
iii. Fulminating infections.

The odontogenic abscesses present in many ways:

- Acute periapical abscess (cause – Infective necrosis of pulp)
- Acute dentoalveolar abscess
- Acute periodontal abscess
- Acute pericoronal abscess.

Causative organisms

- Staphylococci
- Streptococci
- Bacteroides.

Pathogenesis

The infection follows the path of least resistance, the deciding factor being the fascial planes:

- Mandibular 2nd and 3rd molar infections perforate the lingual cortex and spread to submandibular space, as the roots of these teeth lie below the mylohyoid line

- Mandibular premolars and first molars, involve the sublingual space, as the roots of these teeth lie above the mylohyoid line.

Spread of Infection

- *Acute periapical abscess (cause–Infective necrosis of pulp):* Organisms from infected pulp invade periapical tissue through apical foramina
- *Acute dentoalveolar abscess:* Continuation of periapical abscess
- *Acute periodontal abscess:* Arises in the periodontal membrane adjacent to a periodontal pocket
- *Acute pericoronal abscess:* Arises around the crown of impacted or erupting teeth.

Symptoms and Signs

- *Acute periapical abscess:* Severe throbbing pain in the affected tooth which may be carious
- *Acute dentoalveolar abscess:* Severe pain, submucosal and or extraoral swelling
- *Acute periodontal abscess:* Dull pain, rarely severe, pus discharge via gingival pocket
- *Acute pericoronal abscess:* Dull continuous pain, swelling of gingiva around the crown with or without pus discharge.

Relevant Investigations

Diagnosis is more clinical. X-rays are not diagnostic.

Treatment

- *Medical:* Hydration, soft diet, analgesics, mouthwashes, broad-spectrum antibiotics
- *Surgery:* Extraction of affected tooth or incision and drainage of abscess when possible
- *Conservative:* Root canal treatment for acute periapical abscess.

17

Head and Neck

ACUTE SUPPURATIVE SIALADENITIS

Incidence and Etiology

- Occurs generally in:
  - People with poor oral hygiene
  - Postoperative patients
  - Dehydration and reduced or obstructed salivary flow
  - By ascending infection from oral flora, usually *Streptococcus viridans* or pneumococci
- Parotid gland is usually involved
- Submandibular gland may also be involved.

Clinical Presentation

- Acute parotitis:
  - Painful unilateral parotid swelling with trismus, fever and tachycardia
  - On examination, the parotid is diffusely enlarged as a brawny swelling (Fig. 17.1) and extremely tender with purulent discharge through the Stenson's duct.
- Acute submandibular sialadenitis:
  - Painful swelling of the submandibular gland
  - On examination, the gland is enlarged and tender (Fig. 17.2) purulent discharge through the Wharton's duct
  - Obstructing calculus may be seen or felt in the floor of the mouth (Fig. 17.3).

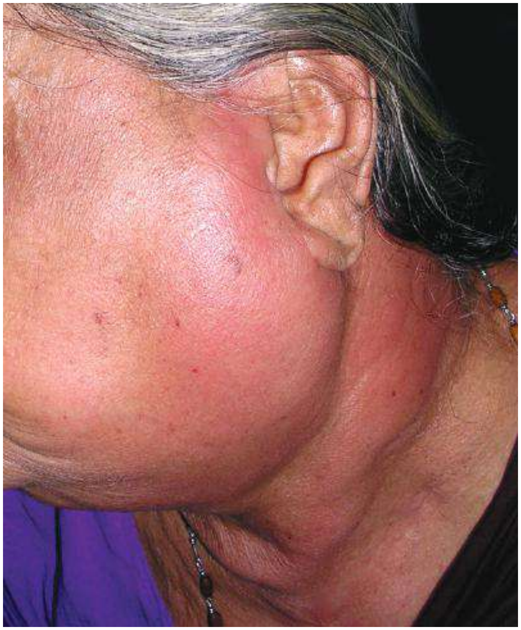

FIG. 17.1: Acute suppurative parotitis

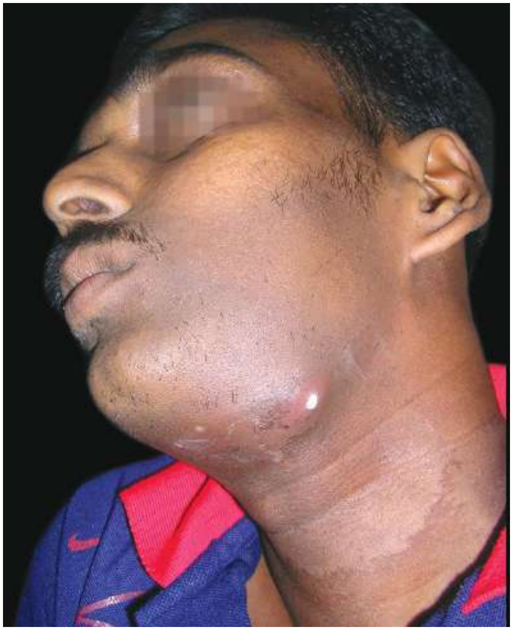

FIG. 17.2: Submandibular salivary gland abscess

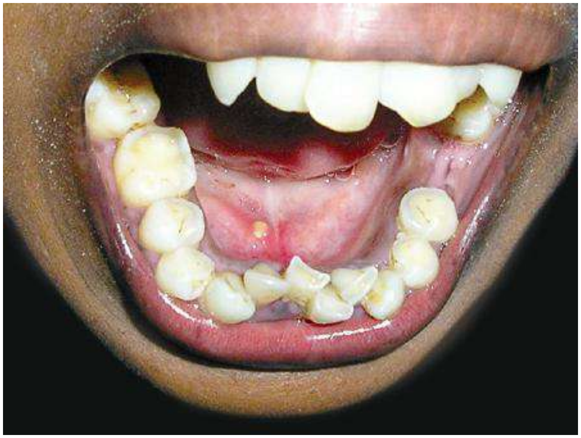

FIG. 17.3: Calculus in the Wharton's duct (*Courtesy:* Dr Jacinth Cornelius)

Relevant Investigation

No special investigation is required.

Treatment

- Broad-spectrum antibiotics are necessary
- If abscess is formed, external drainage is necessary without waiting for fluctuation to establish.

LUDWIG'S ANGINA

Incidence and Etiology

- Usually due to dental infections
- The infection is a mixture of aerobic and anaerobic organisms.

Pathogenesis

It is cellulitis involving the sublingual and submandibular spaces beneath the deep cervical fascia.

Complication: Edema of glottis.

Symptoms

- Severe pain and swelling of neck (Fig. 17.4)
- May cause respiratory embarrassment.

Signs

- Severely swollen neck
- Marked tenderness.

Relevant Investigation

Culture of pus and identifying the organism.

Oral cavity examination is mandatory.

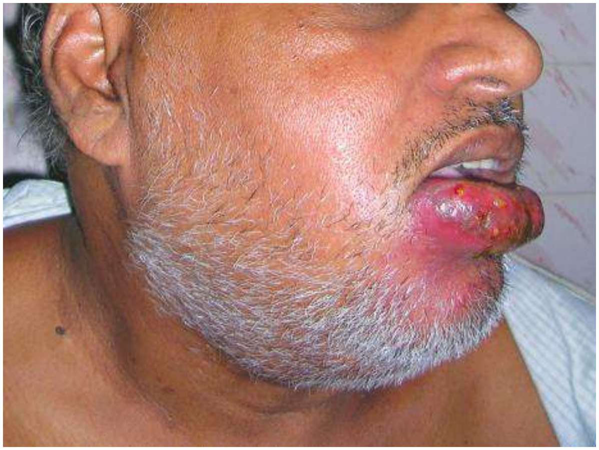

FIG. 17.4: Ludwig's angina

Treatment

- Broad-spectrum antibiotics
- Incision and drainage, if abscesses are found in the dental planes, with or without dental extractions.

18

Thorax

ACUTE MEDIASTINITIS

Incidence and Etiology

Acute suppurative mediastinitis occurs due to varied causes. They are:

- Injury to cervical or thoracic esophagus
- Extension of cervical infections
- Secondary to osteomyelitis of ribs and sternum
- Extension of retroperitoneal infections
- Suppuration of mediastinal lymph nodes.

Clinical Presentation

- Chest pain, fever and chills are commonly present
- Symptoms related to primary cause are usually present.

Relevant Investigations

- X-ray chest (Fig. 18.1) shows mediastinal widening
- CT chest is diagnostic, especially of the primary cause.

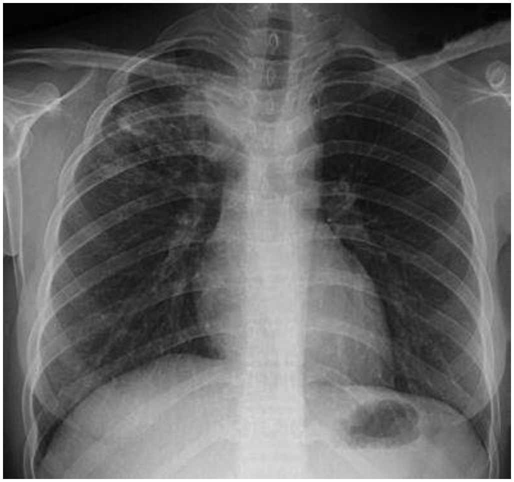

FIG. 18.1: X-ray—Superior mediastinitis

Treatment

- Broad-spectrum antibiotics are necessary
- Immediate surgery with adequate mediastinal drainage for esophageal injuries.

ACUTE PLEURITIS

Incidence and Etiology

- A simple term denoting the pleural involvement of any disease giving rise to pleuritic pain or evidence of pleural friction
- Common feature of pleural invasion by pulmonary tuberculosis or bronchogenic carcinoma.

Symptom

Pain characteristically inspirational.

Sign

Chest movement may be restricted locally with an audible pleural rub.

Relevant Investigations

X-ray chest is useful in identifying the underlying cause.

Treatment

Analgesics are required to treat pleuritic pain with treatment focused towards the primary cause.

PLEURAL EFFUSION

Incidence and Etiology

Denotes accumulation of serous fluid in the pleural space, which results due to:

- Increased hydrostatic pressure or decreased osmotic pressure (transudative effusion). The causes are:
    - Cardiac failure
    - Hepatic failure
    - Renal failure
- Increased microvascular permeability due to disease of pleura or adjacent lung (exudative effusion). The causes are:
    - Infections (e.g. bacterial, tuberculous, fungal, parasitic, viral)
    - Collagen vascular disease (e.g. rheumatoid, lupus)
    - Malignancy (e.g. mesothelioma, lung cancer, metastases)
    - Pulmonary embolism
    - Abdominal disease (e.g. pancreatitis, subphrenic abscess)
- The effusion can be:
    - Unilateral (e.g. tuberculosis, malignancy)
    - Bilateral (e.g. cardiac failure, hypoproteinemia).

Symptom

Breathlessness is the most common symptom.

Signs

- Reduced chest wall movement on the affected side, dullness on percussion and reduced or absent breath sounds and vocal resonance
- Large effusions shift the trachea to the opposite side.

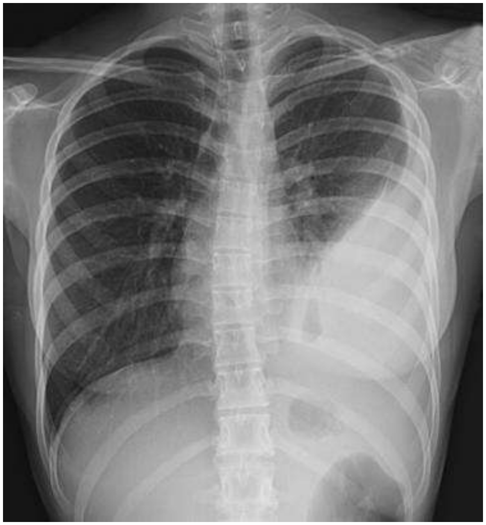

FIG. 18.2A: X-ray chest—Left pleural effusion

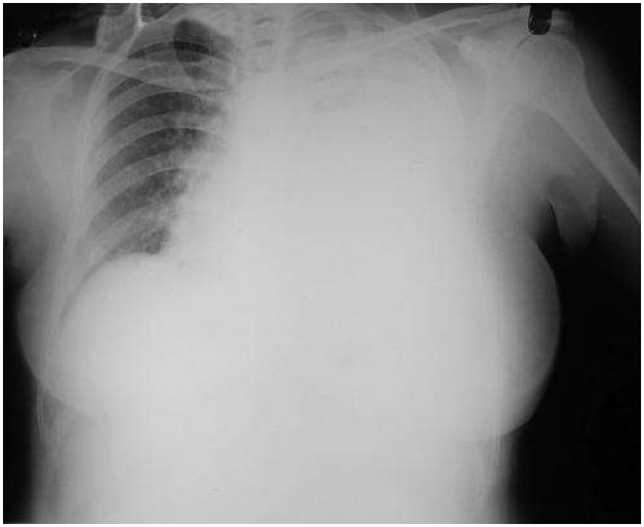

FIG. 18.2B: X-ray chest—Massive pleural effusion left side

Relevant Investigations

- Chest X-ray shows a dense uniform opacity in the lower and lateral parts of hemithorax, shading off above and medially into translucent lung (more than 400 ml is required to blunt the costodiaphragmatic angle) (Figs 18.2A and B)
- US is very useful in differentiating the effusion and tumor
- CT (Fig. 18.3) is diagnostic
- Aspiration of pleural fluid is necessary for:
  - Nature (blood – malignancy, embolus; milky – chyle; pus – empyema; straw color – tuberculosis and transudates)
  - Microbiology – culture sensitivity
  - Cytology
  - Biochemistry (Total proteins >3 g% - exudate, pleural LDH/serum LDH ratio >0.6 – exudates, high levels of triglycerides >110 mg% - chylothorax, amylase – pancreatitis, malignant effusions)
- Bronchoscopy biopsy, thoracoscopy and biopsy of enlarged lymph nodes may be needed to identify the primary cause
- Pleural biopsy yields good results.

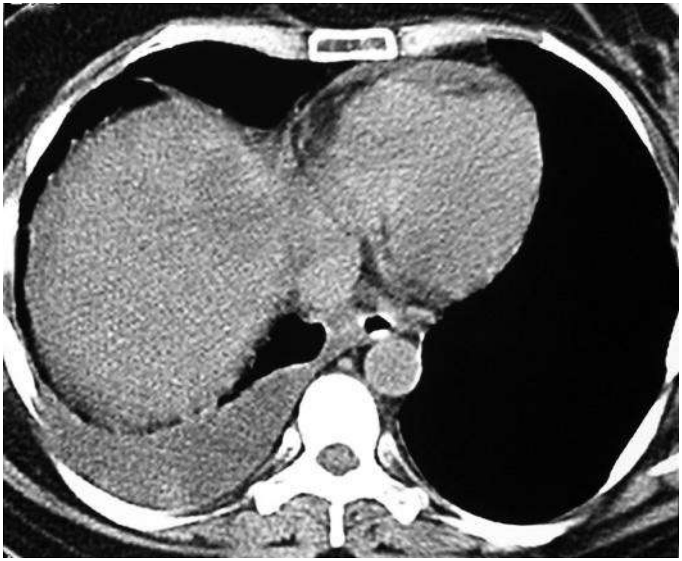

FIG. 18.3: CT—Right sided pleural effusion

Treatment

Aspiration of pleural fluid may be necessary to relieve breathlessness, but treatment of underlying cause is necessary.

ACUTE EMPYEMA THORACIS

Incidence and Etiology

- Empyema indicates pus in the pleural space, which may involve a part (loculated) or the whole of it
- Always secondary to infection in a neighboring structure, usually the lung (bacterial pneumonia and tuberculosis)
- Other causes being, infection of hemothorax or rupture of subphrenic abscess or liver abscess.

Complications: Bronchopleural fistula, empyema necessitans, (empyema burrowing through the chest wall to present as an abscess externally), pleurocutaneous fistula, pericarditis, mediastinal abscess.

Symptom

High intermittent pyrexia, with rigors, sweating, malaise and weight loss.

Signs

- Pleural pain, breathlessness and cough with purulent sputum
- Clinical signs are those of pleural effusion.

Relevant Investigations

- Chest X-ray is diagnostic
- US and CT give invaluable information
- Aspiration of pus is confirmatory
- Isolation of organism from sputum or pleural aspirate is useful in treatment.

Treatment

- General: Supportive respiratory care, physiotherapy
- Drainage of pleural space (thoracentesis or close drainage) and irrigation with normal saline is necessary to clear the pleural space of the pus, under cover of appropriate antibiotics
- Underlying cause like pneumonia and tuberculosis need active treatment.

SPONTANEOUS PNEUMOTHORAX

Incidence and Etiology

- Pneumothorax is presence of air in the pleural cavity
- This is a sudden event
- Spontaneous pneumothorax can be:
  - Primary—without any obvious evidence of pulmonary pathology (e.g. rupture of small emphysematous bulla)
  - Secondary—due to underlying pulmonary pathology (e.g. COPD, tuberculosis).

Symptom

Sudden unilateral chest pain or breathlessness.

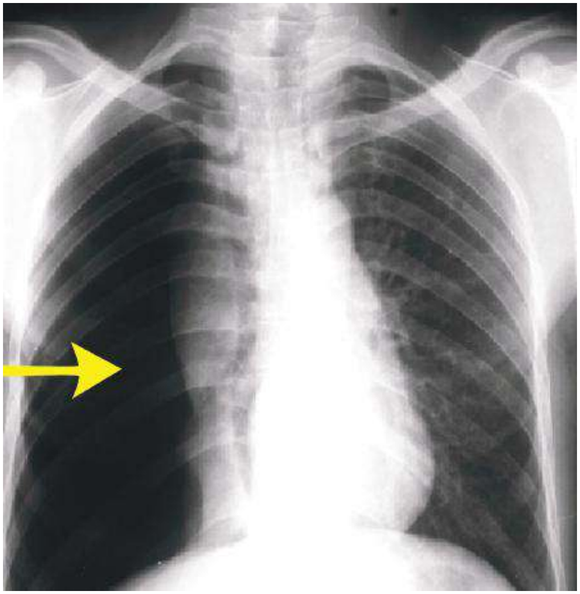

FIG. 18.4: Chest X-ray—Right pneumothorax

Sign

Small pneumothorax shows no clinical signs, whereas, large ones show decreased movement of chest wall, hyper-resonant percussion note and decreased or absent breath sounds.

Relevant Investigations

- Chest X-ray (Fig. 18.4) shows sharply defined edge of the deflated lung with complete translucency between the line and the chest wall with no lung markings
- CT is useful in defining the underlying pathology.

Treatment

- Percutaneous needle aspiration of air is necessary for full lung expansion
- Intercostal tube drainage with underwater seal may be needed [See Chapter 31 (Insertion of Chest Drains)]
- Smoking should be completely avoided and the underlying pathology treated.

FOREIGN BODIES IN THE RESPIRATORY TRACT

Incidence and Etiology

- Foreign bodies get lodged in the trachea and bronchi accidentally, and are more common in:
    - Children than in adults
    - The bronchus than the trachea
    - The right bronchus than the left, as the right bronchus is a direct continuation of trachea and is wider than the left bronchus.
- In children the objects are in the form of seeds, buttons, cells, etc. and in adults, they are dentures, loose teeth, etc.
- In adults, accidental lodging when they are under the influence of alcohol.

Pathogenesis

- Foreign bodies act like a valve and cause symptoms (Table 18.1).

Symptoms

Dyspnea, cough, stridor, cyanosis and fever.

Signs

- Rhonchi
- Reduced breath sounds.

Relevant Investigations

- X-rays reveal the radio-opaque foreign bodies, and collapsed lungs
- Bronchoscopy is diagnostic (Figs 18.5 and 18.6).

Table 18.1: Pathogenesis of foreign body obstruction in lower respiratory tract

Mechanism	*Pathophysiology*
Stop valve	Foreign body causes total obstruction and does not allow air entry (ingress) or allow air and secretions to escape (egress). The lobe of lung may collapse and consolidate
Bypass valve	When the foreign body is small or has a hole, it allows ingress and egress of air
Ball valve	During inspiration, the bronchi dilate allowing air to enter (ingress), but does not allow air to escape (egress) as the bronchi constrict during expiration. The trapped air caused pneumothorax or emphysema

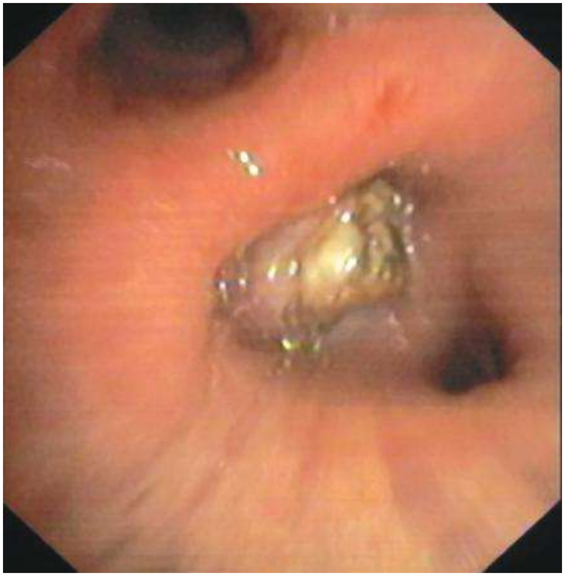

FIG. 18.5: Bronchoscopy—Tooth in the bronchus (*Courtesy*: Dr R Narasimhan and Dr Gayathri)

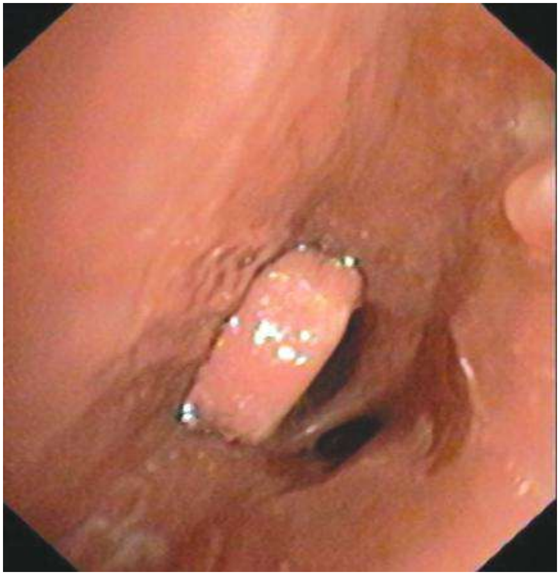

FIG. 18.6: Bronchoscopy—Tablet in the bronchus (*Courtesy*: Dr R Narasimhan and Dr Gayathri)

Treatment

Bronchoscopic removal (using a rigid bronchoscope) is curative.

SUPPURATIVE OR ASPIRATION PNEUMONIA

Incidence and Etiology

- Suppurative pneumonia or pneumonic consolidation denotes destruction of lung parenchyma by inflammatory process with microabscess formation on histology
- *Staphylococcus aureus* and *Klebsiella pneumoniae* infections cause suppuration
- Suppuration has an insidious onset whereas aspiration is acute.

Pathogenesis

- Inhalation of septic material during endotracheal anesthesia or by aspiration of gastric contents
- Aspiration may lead to severe acute respiratory distress syndrome (ARDS).

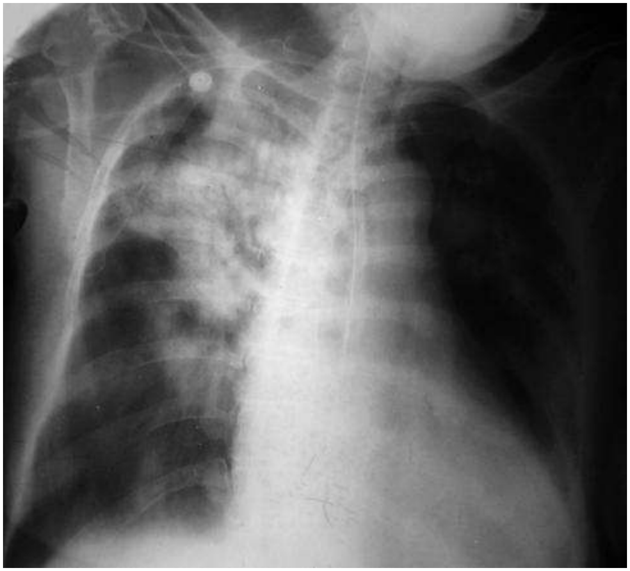

FIG. 18.7: X-ray—Aspiration pneumonia right lung

Symptoms

- Productive cough (fetid or blood stained)
- Pleural pain may be present
- High remittent pyrexia is common.

Sign

Sign of consolidation, with pleural rub.

Relevant Investigation

Chest X-ray (Fig. 18.7) shows homogeneous lobar or segmental opacity. A cavity with fluid level may indicate an abscess.

Treatment

- Broad-spectrum antibiotics form the mainstay of treatment
- Surgical intervention may be required for abscess, which does not respond to medical therapy.

HOSPITAL ACQUIRED PNEUMONIA

Incidence and Etiology

- Refers to a new episode of pneumonia 2 to 3 days after admission to hospital.
- The predisposing factors for the development of pneumonia are:
  - Immunocompromised state (e.g. corticosteroid treatment, malignancy, AIDS)
  - Reduced cough reflex (e.g. postoperative)
  - Aspiration of gastric contents (e.g. vomiting, severe reflux, nasogastric intubation)
  - Endotracheal intubation (e.g. mechanical ventilation)
  - Bacteremia (e.g. sepsis).

Symptoms

- Cough with purulent expectoration is the predominant symptom
- Breathlessness will appear soon.

Signs

- Cyanosis
- Crepitations are heard on auscultation.

Relevant Investigations

- Leukocytosis is present
- Chest X-ray will show mottled opacities in both lung fields.

Treatment

- Broad-spectrum intravenous antibiotics are necessary
- Physiotherapy is mandatory in immobile patients.

ACUTE LUNG ABSCESS

Incidence and Etiology

- A suppurative focus within the lung associated with necrosis
- The bacteria responsible for a lung abscess may reach the lung through various routes. They are:
  - Aspiration down the tracheobronchial tree (e.g. bronchiectasis)
  - By blood (e.g. septicemia)
  - Trauma (e.g. penetrating trauma)
  - Extension of adjacent suppurative focus (e.g. liver abscess).

- Other causes are:
  - Tumors of bronchus (e.g. bronchogenic carcinoma)
  - Suppuration of hematoma (e.g. penetrating trauma).

Symptoms

- Appear acutely ill
- Cough with foul smelling purulent expectoration
- Hemoptysis may occur
- Fever
- Breathlessless
- Chest pain.

Sign

Crepitations may be heard.

Relevant Investigations

X-ray and CT are diagnostic.

Treatment

- Antimicrobial therapy resolves most abscesses
- Surgical intervention (lobectomy or segmentectomy), is reserved for those which do not respond.

PULMONARY EMBOLISM

Incidence and Etiology

Majority of pulmonary emboli result from deep venous thrombosis of lower limbs, and they can be acute (minor and massive) or chronic.

Symptoms and Signs

The symptoms and signs are tabulated in Table 18.2.

Table 18.2: Symptoms and signs of pulmonary embolism

Type	*Symptoms*	*Signs*
Acute minor	Shortness of breath and hemoptysis – pleurisy	Pleural rub and signs of pleural effusion
Acute massive	Central chest pain, apprehension, low cardiac output and syncope	Sinus tachycardia, hypotension and peripheral vasoconstriction
Chronic	Exertional dyspnea, syncope and chest pain over months and years	Pulmonary hypertension, loud pulmonary component of second heart sound and a right ventricular heave.

Relevant Investigations

Chest radiographs, ECG, arterial blood gases, ventilation-perfusion lung scanning and pulmonary angiography are useful investigations.

Treatment

- *General:* Opiates to relieve pain and distress, resuscitation by external cardiac massage, with oxygen support
- Anticoagulation is necessary atleast for 5 days
- Thrombolytic therapy is used in acute massive types.

19

Breast

BREAST HEMATOMA

Incidence and Etiology

Associated with history of trauma such as seatbelt injury during a road traffic accident, or following a violent contraction of pectoralis muscles responding to a blow.

Symptoms

- Pain in the breast
- History of trauma.

Sign

Presence of bruise over the breast may be the contributory finding for diagnosis.

Relevant Investigations

Fine needle aspiration cytology (FNAC) and mammography are needed to rule out malignancy.

Treatment

Exploration and histopathological confirmation.

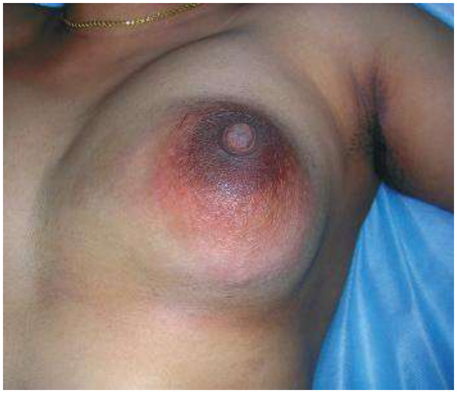

FIG. 19.1: Acute breast abscess

ACUTE BREAST ABSCESS

Incidence and Etiology

Collection of pus resulting from mastitis.

Symptoms

- Starts with dull ache, which proceeds on to throbbing pain
- Systemic manifestations like fever and malaise may develop.

Signs

- Superficial abscesses may show as fluctuant tender lumps
- Deep abscesses may show as severe cellulitis with edema of breast (Fig. 19.1) without fluctuation.

Relevant Investigations

- Polymorphonuclear leukocytosis is generally present
- Diagnostic aspiration may confirm the presence of pus.

Treatment

- Surgical drainage gives full relief
- Spontaneous or inadequate drainage may result in the formation of antibioma, chronic abscess, mammary fistula with purulent or sero sanguinous discharge.

20

Spine

DEGENERATIVE DISEASES OF DISK AND FACET JOINTS

Incidence and Etiology

Degenerative diseases of the spine are very common. They affect the lumbar, cervical and thoracic regions in decreasing order of frequency.

- Lumbar degenerative disease:
    - Prolapse of intervertebral disk: Disks between L5/S1 and L4/L5 are most often affected. Part of the gelatinous nucleus pulposus protrudes through the rent in the annulus fibrosus at its weakest part, which is posterolateral and compress the nerve roots (compressive radiculopathy) (Figs 20.1A to D). Occasionally, a large central disk prolapse at L5/S1 compresses the cauda equina (Cauda equina syndrome).
    - Facet joint disease: Degenerative changes cause root compression.
    - Lumbar canal stenosis: Increasing wear and tear, bony overgrowth and ligamentous hypertrophy can cause narrowing of the spinal canal.
- Cervical degenerative disease: Pathogenesis is the same as in lumbar disease.

Cauda equina syndrome consists of back pain, bilateral sciatica and urinary retention.

Symptoms and Signs

- Lumbar degenerative disease: Back pain is the constant feature, but varies according to the compression of the root:

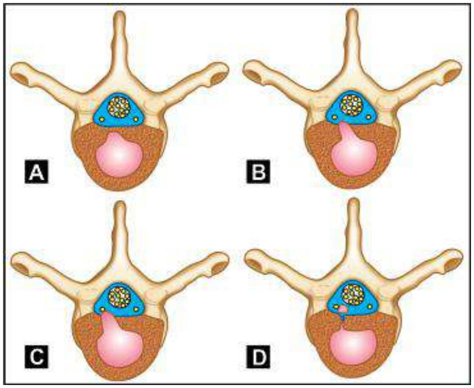

FIGS 20.1A to D: Stages of prolapse of intervertebral disk: (A) Torn annulus fibrosus with disk bulge; (B) Extrusion of nucleus pulposus protrusion not pressing the nerve; (C) Extrusion of nucleus pulposus protrusion pressing the nerve; (D) Sequestration of disk

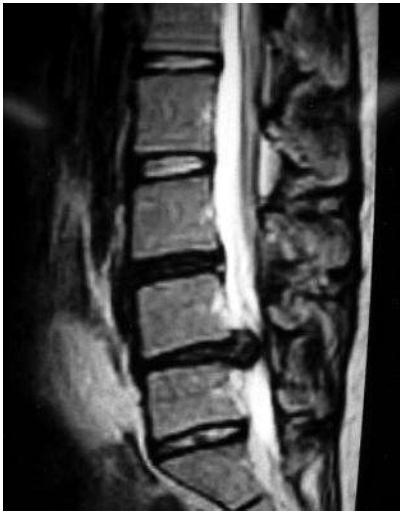

FIG. 20.2: MRI—Prolapse of L4-L5 intervertebral disk

- Prolapse of intervertebral disk: Back pain, radiating to
  - Inner thigh (L4 root compression—L3/L4 prolapse)
  - Outer aspect of the leg to the big toe (L5 root compression—L4/L5 prolapse)
  - The sole of foot (S1 root—L5/S1 prolapse), is the presenting symptom.

  Acute radiating pain and inability to move, aggravated by coughing and sneezing is the symptom of acute disk prolapse. Straight leg raising test is positive on the affected side.
- Facet joint disease presents with pain similar to the synovial joint pain (more after a period of rest like early mornings and becomes less after movement)
- Lumbar canal stenosis causes numbness and weakness of lower limbs on walking.

- Cervical degenerative disease: The presenting symptom is neck pain related to movements with radiculopathy.

Relevant Investigations

CT and MRI (Fig. 20.2) are diagnostic.

Treatment

- Acute attacks require rest and analgesics, followed by physiotherapy
- Chronic, persistent or progressive symptoms require surgery to remove the prolapsed disk or decompressive laminectomy.

SPONDYLOLISTHESIS

Incidence and Etiology

- This term signifies slipping of one vertebra over the lower, which is usually forwards, but may be backwards. In the normal spine, forward displacement of a vertebral body is prevented by engagement of the articular processes with those of the segment next below it, and this check mechanism is weak in spondylolisthesis
- The displacement is most common in the 5th lumbar vertebra, and the displacement may increase slowly over years, and reach a severe degree of spondyloptosis.
- The predisposing factors are:
    - Spondylolysis (a defect in the pars interarticularis of the neural arch)
    - Osteoarthritis of the posterior (facet) joints (degenerative)
    - Congenital malformation of the articular processes (rare).

Symptom

Chronic backache with or without sciatica.

Signs

- A step above the sacral crest
- Lumbar vertebral bodies may be felt per abdomen due to its forward displacement
- Straight leg raising (SLR) test may be positive.

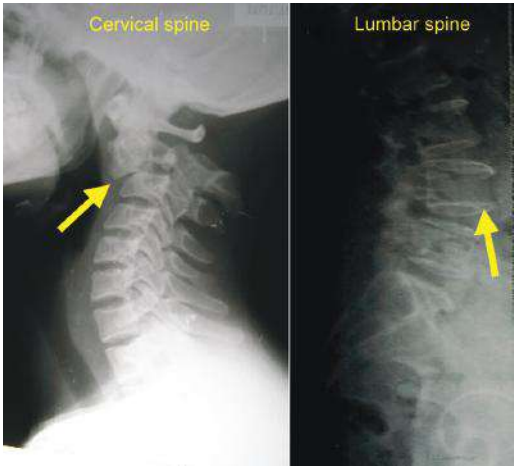

FIG. 20.3: X-rays—Spondylolisthesis

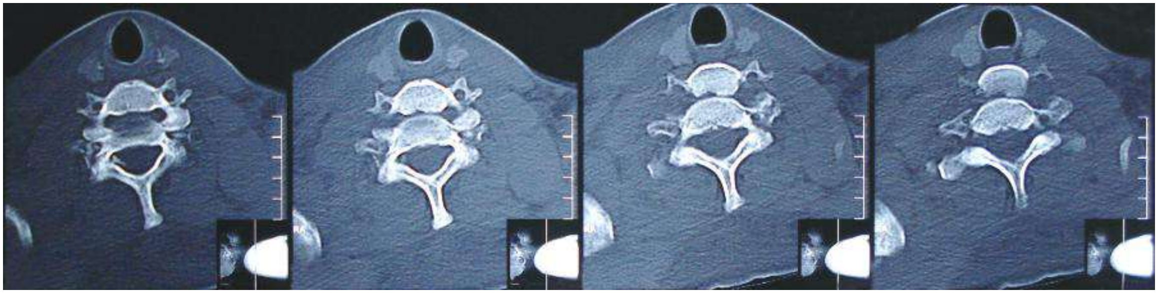

FIG. 20.4: CT—Spondylolisthesis (*Courtesy*: Dr V Thulasiraman)

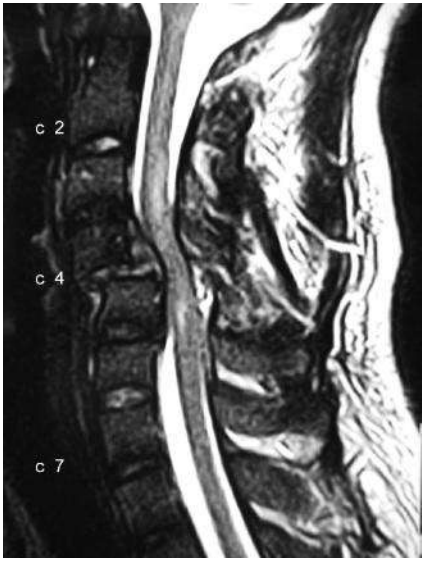

FIG. 20.5: MRI spine—C4-C5 spondylolisthesis

Relevant Investigations

- X-rays (Fig. 20.3) and CT (Fig. 20.4) are diagnostic
- MRI (Fig. 20.5) is used to assess root compression.

Treatment

- Asymptomatic patients require no treatment
- Mild cases require surgical corset
- Surgery is justified only when the disability is severe.

21

Gastroenterology

ACUTE ABDOMEN

Introduction

Acute pain often denotes the presence of a disease process or injury, which needs to be treated with elimination of cause. It is a signal of ongoing or impending tissue damage. Acute abdominal pain denotes similar intra-abdominal organ disease or pathology, which warrants emergent care.

The main visceral pain receptors in the abdomen respond to mechanical and chemical stimuli.

- *Mechanical stimuli*: Stretch, distension, contraction, compression and torsion
- *Chemical stimuli*: Bradykinin, substance P, serotonin and prostaglandins. These receptors are located on the serosal surfaces, within the mesentery and within the walls of hollow viscera
- Gut related visceral pain is usually perceived in the midline because it is a midline structure in an embryo and has bilateral symmetric innervations, except for pains originating from the gallbladder and the ascending and descending colon. Pain from other intra-abdominal organs tends to be unilateral.
    - *Pain at epigastrium*: Diseases of the foregut (abdominal esophagus, stomach and proximal half of second part of duodenum and their offshoots like liver, gallbladder, pancreas and spleen) (e.g. gastric and duodenal ulcers)
    - *Pain at the umbilical region*: Diseases of midgut (distal half of second part of duodenum, small bowel, colon up to the proximal 2/3 of transverse colon) (e.g. intestinal tuberculosis)
    - *Pain at the hypogastrium*: Diseases of hindgut (distal 1/3 of transverse colon to the anorectal junction) (e.g. colorectal and urinary bladder pathologies)

The abdomen is divided into 10 arbitrary regions for convenience of understanding and localizing. The two lateral vertical planes pass from the costal margin close to the tip of the ninth cartilage above to the femoral artery below. The horizontal plane (the subcostal plane) connects the lowest points on the costal margins and the interiliac plane connects the tubercles of the iliac crests (Fig. 21.1). These divisions will help in localizing the diseases.

Pain

Nature of Pain

- *Sudden onset pain*: Pain is sudden in otherwise healthy and asymptomatic persons. It increases in a very short time (e.g. pain due to perforations of duodenal ulcer and appendicitis)
- *Sudden onset with pain-free intervals*: The pain of quick onset can reach a peak making the patient writhe in pain and buckle up, and also quickly recede to absolutely pain-free period, only to recur again (e.g. colics)
- *Dull continuous pain increasing in severity*: A persistent pain may increase in severity over a period of time (e.g. a dull pain of subacute appendicitis in the right iliac fossa may become severe when the appendicitis becomes severe and acute due to superadded infection and inflammation)
- *Burning pain*: The pain may be of burning nature occurring suddenly (e.g. pain in acid peptic disease)
- *Constant or continuous pain*: Persistent pain without variation in intensity (e.g. peritonitis)
- *Agonizing pain*: Very severe pain which upsets the morale of the patient (e.g. pancreatitis, torsion of pedicled organ)

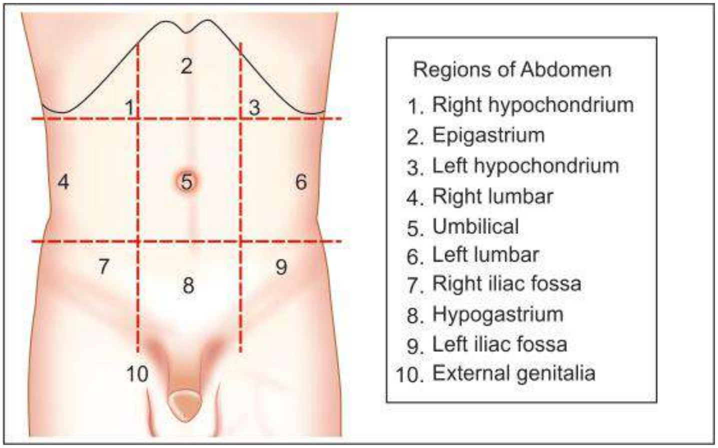
2
1
3
4
5
6
7
8
9
10
Regions of Abdomen
1. Right hypochondrium
2. Epigastrium
3. Left hypochondrium
4. Right lumbar
5. Umbilical
6. Left lumbar
7. Right iliac fossa
8. Hypogastrium
9. Left iliac fossa
10. External genitalia

FIG. 21.1: Regions of abdomen

- *Throbbing pain*: Continuous pain throbbing in nature (e.g. acute cholecystitis—due to inflammation being inside closed confines of a structure).

Shift, Migration and Referred Pain

- ***Shifting pain***: Shifting pain is defined as the pain, which originates in one region and shifts totally to another region, without any pain at the point of origin (e.g. pain in acute appendicitis originates at the umbilical region and shifts to right iliac fossa later)
- ***Migrating pain***: Migrating pain is that pain which originates at one region and shifts to another region, but does continue to stay at a lesser degree at the point of origin (e.g. pain in perforated duodenal ulcer is right hypochondrial to start with, and spreads to the right iliac fossa due to the flow of the gastric contents down the right paracolic gutter)
- ***Referred pain***: The pathology and the area of the pain are different, since the organ of pathology and the area of pain share the same nerve supply. The examples are given in Figure 21.2

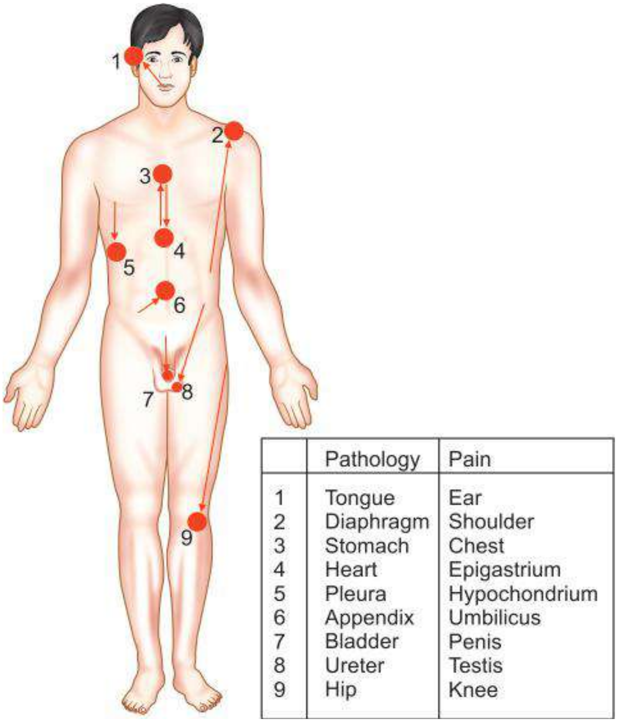

	Pathology	Pain
1	Tongue	Ear
2	Diaphragm	Shoulder
3	Stomach	Chest
4	Heart	Epigastrium
5	Pleura	Hypochondrium
6	Appendix	Umbilicus
7	Bladder	Penis
8	Ureter	Testis
9	Hip	Knee

FIG. 21.2: Referred pain

- The perception of visceral pain corresponds to the spinal segments where the visceral afferent nerve fibers enter the spinal cord. Table 21.1 shows some common spinal segments where visceral pain is perceived.

Effects of Movements, Food or Habits

The abdominal pain may be aggravated by certain factors like movements, food or habits. Some examples are given in Table 21.2.

The pain may get relieved by certain factors. Some examples are given in Table 21.3.

Table 21.1: Visceral pain and dermatomal perceptions

Organ of pathology	*Site of pain*	*Dermatome*
Stomach	Epigastrium	T5-T10
Small bowel	Umbilicus	T9-T10
Large bowel up to splenic flexure	Umbilicus	T11-L1
Large bowel from splenic flexure	Hypogastrium	L1-L2
Gallbladder	Epigastrium, scapular region	T7-T9
Pancreas	Epigastrium	T6-T10
Ureter	Loin to groin	Genitofemoral nerve (L1-L2)
Testis and ovary	Umbilicus	T10-T11

Note: Segmental nerve supply mentioned here is sympathetic supply of the viscus. Parasympathetic supply is from the vagus nerve, excepting for the hindgut and the urinary bladder, which is from the sacral segments.

Table 21.2: Aggravating factors of abdominal pain

Aggravating factor	*Pathology*
Physical movements like jolting	Appendicitis, peritonitis
Deep inspiration	Pleurisy
Intake of fatty food	Cholecystitis
Intake of spicy food, alcohol	Acid peptic disease
Lying supine	Pancreatitis
Drugs—Analgesics, NSAIDs	Acid peptic disease

Table 21.3: Relieving factors of abdominal pain

Relieving factor	*Pathology*
Vomiting	Peptic ulcer pain
Intake of bland food	Peptic ulcer pain
Local pressure	Colicky pain
Leaning forward	Pancreatitis
Drugs—antacids, H_2 blockers	Acid peptic disease

Vomiting

Vomiting is a very common feature associated with pain in acute abdominal emergencies. The patient should be asked about the following details, as each has its own significance. They are:

- *Character of vomiting*
  - *Projectile*: Involuntary projectile ejection of large quantities of vomitus (e.g. high intestinal obstruction)
  - *Regurgitative*: Effortless involuntary regurgitation of intestinal contents (e.g. peritonitis due to perforation)
- *Frequency of vomiting*
  - *Constant*: Persistent vomiting even in the absence of food intake (e.g. acute intestinal obstruction, acute pancreatitis)
  - *Periodical*: Vomiting with some periodicity or following food intake indicates bowel obstruction (e.g. acute peptic ulcer, gastric outlet obstruction)
- *Nature of vomitus*
  - Coffee ground vomitus (Brown to dark brown colour)—gastric contents with altered blood (e.g., bleeding duodenal ulcer)
  - Bloody (red in colour) (e.g. bleeding oesophageal varices)
  - Faeculent (yellowish green and foul smelling) (e.g. gastric contents followed by duodenal and intestinal contents in intestinal obstruction)
- *Quantity of vomitus*
  - Large quantities indicate distal bowel obstruction
  - Small quantities indicate gastric outlet obstruction
- *Relationship with pain*: The pain may precede, accompany or follow abdominal pain. The examples are given in Table 21.4.

Table 21.4: Relationship of vomiting with abdominal pain

Relationship of vomiting with abdominal pain	*Pathology*
Pain preceding vomiting	Acute appendicitis, acute peptic ulcer, biliary and renal colic
Pain and vomiting occurring together	High intestinal obstruction
Vomiting occurs much later than the pain	Low intestinal obstruction
Vomiting as a late feature or absent	Large intestinal obstruction

- Abdominal pain due to acute peptic ulcer may get relieved by vomiting, but gives temporary relief in colics
- Vomiting is not a constant feature in acute appendicitis as the stomach gets empty after one bout, but nausea persists
- In peritonitis, vomiting is absent but appears at a later stage.

Alteration in Bowel Habits

- ***Obstipation***: Inability to pass feces and flatus (absolute constipation) (e.g. acute intestinal obstruction)
- ***Constipation***: Inability to pass stools but able to pass flatus (due to solidified fecal mass impacted at the anal verge) (e.g. acute fissure in ano)
- ***Tenesmus***: A desire to evacuate but results in passage of flatus only (e.g. acute appendicitis)
- ***Passage of blood stained stools***: Red currant jelly stools in acute intussusception, mesenteric vascular obstruction, mesenteric thrombosis, malignancy
- ***Diarrhea***: Passage of loose stools—rarely seen in illnesses requiring emergency surgery (e.g. ulcerative colitis, Crohn's disease, acute enteritis).

Urinary Symptoms

Patients presenting with abdominal pain may have associated urinary symptoms:

- ***Frequency in micturition***: Patients presenting with renal or ureteric colic, may have associated urinary infections along with urolithiasis. Frequency is one of the common symptoms
- ***Strangury***: Frequent passing of urine with excruciating pain. (e.g. impacted stones in the urinary tract, pelvic or retrocecal appendicitis). Patient succeeds in passing a small quantity of blood stained urine
- ***Hematuria***: Passing blood in the urine (e.g. stones in the urinary tract, rarely retrocecal or pelvic appendix lying close to the ureter).

ACUTE UPPER ABDOMINAL PAIN

Causes

Right Hypochondrium

- Acute cholecystitis
- Acute cholangitis
- Acute hepatitis
- Acute hyperacidity
- Perforated duodenal ulcer

Epigastrium

- Acute hyperacidity
- Acute pancreatitis
- Perforated duodenal ulcer
- Acute hepatitis (left lobe)

Left Hypochondrium

- Acute pancreatitis
- Acute hyperacidity
- Splenic infarct

Eliciting History

1. *Nature of pain*
    - Continuous (e.g. acute pancreatitis)
    - Episodic (e.g. acute hyperacidity)
    - Colicky (e.g. biliary colic).
2. *Location of pain*
    - Epigastric pain (e.g. acute hyperacidity, acute pancreatitis, acute colitis)
    - Left hypochondrial pain (e.g. acute hyperacidity, left renal colic)
    - Right hypochondrial pain (e.g. acute cholecystitis, acute hepatitis, right renal colic).

> Renal colics can present as upper abdominal pain in their respective sides.

3. *Association of vomiting*: Presence of vomiting is not a very reliable symptom to narrow down the diagnosis, as it can be present with any severe painful pathology in the upper abdomen.
4. *Association of fever*: Fever indicates infective pathology (e.g. acute cholecystitis, acute pancreatitis, acute colitis, perforated duodenal ulcer).

5. *Association of jaundice*: Jaundice may be present with acute cholangitis, acute hepatitis or acute cholecystitis.
6. *Association of loose stools*: Association of loose stools may indicate colitis or rarely acute pancreatitis.
7. *Radiation*: Radiation to right scapula or shoulder is common with acute cholecystitis, perforated duodenal ulcer due to irritation of diaphragm.
8. *Aggravating factors*
   - Food—in acute hyperacidity
   - Lying supine—in acute pancreatitis
   - Deep breathing—in acute cholecystitis.
9. *Relieving factors*: Leaning forward while sitting—acute pancreatitis.
10. *Referred pain*: In some pathologies, the pathology and the area of the pain are different, since both of them share the same nerve supply.

Past History

- History of pain (e.g. acute on chronic cholecystitis)
- Previous surgery (e.g. cholecystectomy will rule out cholecystitis from consideration).

Family History

Gallstones

Clinical Examination

General
- Breath for fetor (e.g. alcoholic hepatitis, acute pancreatitis)
- Conjunctiva for anemia, jaundice
- Tongue for anemia
- Neck for lymphadenopathy
- Hands for signs of liver failure (e.g. clubbing, palmar erythema, liver flap, etc).

Abdomen

Inspection

Distension:
- Generalized (e.g. perforated duodenal ulcer or gallbladder with peritonitis)
- Right upper abdominal (e.g. hepatomegaly)
- Epigastric (e.g. left lobar hepatomegaly, carcinoma stomach)
- Left upper abdominal (e.g. splenomegaly)
- Scars, swellings and sinuses.

Palpation

Tenderness:

- All quadrants—generalized peritonitis
- Right upper quadrant (e.g. acute hepatitis, acute cholecystitis, acute hyperacidity)
- Epigastric (e.g. acute gastritis, acute hepatitis)
- Left upper quadrant (e.g. acute gastritis, acute pancreatitis).

Lump:

- Right upper quadrant (e.g. hepatomegaly, distended gallbladder)
- Epigastric (e.g. carcinoma stomach, left lobar hepatomegaly)
- Left upper quadrant (e.g. carcinoma stomach, splenomegaly).

Percussion

Percuss the liver for:

- Its enlargement (e.g. acute hepatitis)
- Obliteration of liver dullness (e.g. perforated duodenal ulcer).

Auscultation

- Absence of bowel sounds indicates paralytic ileus (e.g. perforated peritonitis)
- Exaggerated bowel sounds may indicate obstruction of small bowel (e.g. intestinal colic)
- Normal bowel sounds indicate that there is no gross infection of the peritoneum.

Examination of

- Groins (e.g. obstructed hernia)
- Genitalia (e.g. obstructed hernia)

Differential Diagnosis by Clinical History and Examination

- *Right hypochondrial pain*
  - Without fever and but local tenderness (e.g. acute cholecystitis, acute hepatitis, acute hyperacidity)
  - With fever and local tenderness
    - + Distension (e.g. acute cholecystitis)
    - Obliteration of liver dullness (e.g. perforated duodenal ulcer)
    - + Hepatomegaly (e.g. acute liver abscess)
  - With vomiting
    - Local tenderness (e.g. acute cholecystitis, acute hyperacidity)
    - Local tenderness with obliteration of liver dullness (e.g. perforated duodenal ulcer)
  - With diarrhea and local tenderness (e.g. acute colitis)

  - ❖ With jaundice
    - Local tenderness + hepatomegaly (e.g. acute hepatitis, acute cholecystitis, choledocholithiasis, cholangitis, acute liver abscess)
    - Local tenderness, fever, +/– abdominal lump (e.g. mucocele gallbladder, choledochal cyst)
  - ❖ Colicky in nature +/– local tenderness (e.g. biliary colic, right renal colic)
- *Epigastric pain*
  - ❖ Without fever but with local tenderness (e.g. acute hyperacidity, acute pancreatitis, acute hepatitis—left lobe)
  - ❖ With fever and local tenderness + distension (e.g. acute pancreatitis, perforated duodenal or gastric ulcer)
  - ❖ With vomiting and
    - Local tenderness (e.g. acute hyperacidity, acute pancreatitis)
    - Local tenderness and obliteration of liver dullness (e.g. perforated ulcer)
  - ❖ With diarrhea and local tenderness (e.g. acute colitis)
  - ❖ With jaundice and local tenderness
  - ❖ + tender hepatomegaly (e.g. left lobar hepatitis, left lobar liver abscess
  - ❖ Nontender hepatomegaly (e.g. metastatic liver)
- *Left hypochondrial pain*
  - ❖ Without fever but with local tenderness (e.g. acute gastritis, acute pancreatitis)
  - ❖ With fever and local tenderness + /– splenomegaly (e.g. acute pancreatitis, splenic infarct)
  - ❖ With diarrhea and local tenderness (e.g. acute colitis)
  - ❖ Colicky in nature + local tenderness (e.g. left renal colic).

- Perforated bowel can give rise to generalized abdominal distension due to generalized peritonitis
- Obstructed groin hernia is one of the common causes of intestinal obstruction causing generalized distension.

Relevant Investigations

Hematology

- Leukocytosis in infective pathologies (e.g. acute cholecystitis, perforated dudodenal ulcer, perforated cholecystitis, acute pancreatitis)
- Raised ESR in all infective pathologies.

Radiology

- *Plain X-ray abdomen*: Gas under the diaphragm (e.g. perforated hollow viscus)
- *Ultrasonography:* Radiopaque shadows in right upper abdomen (e.g. gallstones, renal stones)

Treatment Plan

- *Nonperforated pathologies:* Medical management
- *Perforated pathologies:* Early surgical management
- Diagnosis not clear and not responding to medical management—exploratory laparotomy.

ACUTE LOWER ABDOMINAL PAIN

Causes

Right Iliac Fossa

- Acute appendicitis
- Perforated appendicitis
- Acute mesenteric adenitis
- Ureteric colic
- Acute Meckel's diverticulitis
- Acute cecal diverticulitis
- Torsion of cyst of right ovary
- Pelvic inflammatory disease
- Incarcerated right inguinal hernia
- Ruptured ectopic gestation
- Right ureteric colic.

Hypogastrium

- Acute cystitis
- Acute congestive dysmenorrhea
- Uterine fibroid.

Left Iliac Fossa

- Acute diverticulitis
- Ureteric colic
- Torsion of cyst of left ovary
- Pelvic inflammatory disease
- Incarcerated left inguinal hernia
- Ruptured ectopic gestation
- Left ureteric colic.

Diabetic ketoacidosis is one of the important metabolic causes of acute lower abdominal pain.

Eliciting History

Nature of pain

- Continuous (e.g. acute appendicitis)
- Episodic (e.g. acute cystitis, ruptured ectopic gestation)
- Colicky (e.g. appendicular colic, ureteric colic, dysmenorrhea).

Association of nausea and vomiting

Presence of vomiting is not a very reliable symptom to narrow down the diagnosis, as it can be present with any severe painful pathology in the lower abdomen. Nausea is a predominant symptom of acute appendicitis, whereas vomiting is a common symptom of ureteric colic.

Association of fever

Fever indicates infective pathology (e.g. acute appendicitis, acute cystitis, acute colitis, perforated appendicitis).

Association of loose stools

Association of loose stools may indicate colitis, acute diverticulitis.

Radiation

Radiation to pain to external genitalia occurs with ureteric colic.

Past History

- History of pain (e.g. acute on chronic appendicitis, ureteric colic)
- Previous surgery (e.g. appendicectomy will rule out appendicitis from consideration).

Personal History

Menstrual irregularities

- Dysmenorrhea (e.g. congestive dysmenorrhea)
- Irregularities (e.g. pelvic inflammatory diseases)
- Amenorrhea (e.g. ruptured ectopic gestation).

Family History

- Diverticulosis
- Colonic malignancy.

Clinical Examination

General

- Conjunctiva for anemia (e.g. ruptured ectopic gestation)
- Tongue for anemia (e.g. ruptured ectopic gestation)
- Neck for lymphadenopathy (e.g. mesenteric adenitis as a part of tuberculosis).

Abdomen

- ***Inspection***

  Distension:
  - Generalized (e.g. perforated appendicitis with generalized peritonitis)
  - Right lower abdominal (e.g. ruptured appendicitis, torsion of right ovary or its cyst)
  - Hypogastric (e.g., distended urinary bladder in cystitis, enlarged uterus)
  - Left lower abdominal (e.g. torsion of left ovary or its cyst)

  Scars, swellings and sinuses
- ***Palpation***

  Tenderness:
  - All quadrants – generalized peritonitis
  - Right lower quadrant (e.g. acute appendicitis, acute mesenteric adenitis)
  - Hypogastric (e.g. acute cystitis)
  - Left lower quadrant (e.g. acute colitis, acute diverticulitis)

  Lump:
  - Right lower quadrant (e.g. appendicular abscess, mesenteric adenitis, right ovarian cyst)
  - Hypogastric (e.g. distended urinary bladder, uterine fibroids)
  - Left lower quadrant (e.g. carcinoma colon, left ovarian cyst)
- ***Percussion***

  Percuss the liver for:
  - Its enlargement (e.g. associated metastases liver)
  - Obliteration of liver dullness (e.g. perforated appendicitis and diverticulitis)
- ***Auscultation***
  - Absence of bowel sounds indicates paralytic ileus (e.g. perforation and peritonitis)
  - Exaggerated bowel sounds may indicate obstruction of small bowel (e.g. intestinal colic)
  - Normal bowel sounds indicate that there is no gross infection of the peritoneum

Differential Diagnosis by Clinical History and Examination

- *Pain in the right iliac fossa with*
  - Nausea/vomiting and fever and local tenderness (e.g. acute catarrhal/perforated appendicitis, acute typhlitis, acute Meckel's diverticulitis, acute mesenteric adenitis)
  - Vomiting and
    - +/– Local tenderness (colicy pain—e.g. right ureteric colic, appendicular colic, radiating pain—right ureteric colic)
    - Menstrual irregularities and palpable lump (e.g. right ovarian cyst torsion)
    - Amenorrhea with local tenderness and anemia (e.g. ruptured ectopic gestation)
    - Abdominal lump (e.g. ileocecal tuberculosis, Crohn's disease, cecal malignancy, right ovarian malignancy, right ovarian cyst)
  - Diarrhea or dysentery and local tenderness (e.g. acute amebic typhlitis, cecal diverticulitis)
- *Pain in the left iliac fossa with*
  - Vomiting
  - +/– Local tenderness (e.g. left ureteric colic)
    - Menstrual irregularities and palpable lump (e.g. left ovarian cyst torsion)
    - Amenorrhea with local and cervical tenderness and anemia (e.g. ruptured ectopic gestation)
    - Abdominal lump (e.g., colonic malignancy, left ovarian malignancy, left ovarian cyst)
    - Diarrhea or dysentery with or without local tenderness (e.g. acute colitis, acute diverticulitis)
- *Hypogastric pain with*
  - Urinary symptoms with
    - Local tenderness (e.g. acute cystitis)
    - Distended urinary bladder (e.g. acute distension of bladder—bladder outlet obstruction)
  - Menstrual irregularities
    - +/– With or without local tenderness (e.g. congestive dysmenorrhea, pelvic inflammation)
    - Enlarged uterus (e.g. uterine fibroids, endometrial malignancy)
- *Acute lower abdominal pain (any or all quadrants of lower abdomen) with*
  - Menstrual irregularities and local tenderness (e.g. pelvic inflammatory disease)
  - Vague symptoms and signs (e.g. metabolic causes—diabetes mellitus).

Relevant Investigations

Hematology

- Reduced hematocrit (e.g. ruptured ectopic gestation, colonic malignancy)
- Leukocytosis in infective pathologies (e.g. acute appendicitis, perforated appendicitis, and diverticulitis, dudodenal ulcer)
- Raised ESR in all infective pathologies.

Radiology

- Plain X-ray abdomen
  - Gas under the diaphragm (e.g. perforated appendicitis and diverticulitis)
  - Opaque abdomen (e.g. peritonitis)
  - Radiopaque shadows (e.g. ureteric stone)
- Ultrasonography
  - Cystic swelling (e.g. ovarian cyst torsion)
  - Collection of pus (e.g. appendicular abscess, paracolic abscess)
  - Radiopaque shadows in the lines of ureters (e.g. ureteric calculus) and bladder region (e.g. vesical calculus)
  - Dilated pelvicalyceal system or ureters (e.g. obstructed ureteric stone).

Treatment Plan

- Nonperforated pathologies—medical management (except acute appendicitis)
- Perforated pathologies—early surgical management
- Diagnosis not clear and not responding to medical management—exploratory laparotomy or diagnostic laparoscopy.

ACUTE LIVER ABSCESS

Incidence and Etiology

- *Amebic liver abscess:* Amebic infections causing hepatitis and dysentery may resolve into a liver abscess (usually in alcoholics)
- *Pyogenic liver abscess:* Septicemic patients may present with pyemic abscesses in the liver.

Pathogenesis

- *Amebic liver abscess:* Amebic infections originating in the colon as amebic colitis, travel through the portal blood to reach the liver, forming an abscess more commonly in the right lobe of the liver
- *Pyogenic liver abscess:* Septicemic patients' infected blood reach the liver through the systemic and portal circulation, forming multiple abscesses in both lobes of the liver.

Symptoms

- Constitutional symptoms with high-grade fever, tachycardia and sometime shock.
- Patients with amebic etiology may give history of preceding diarrhea or dysentery.

Signs

- Intercostal tenderness is pathognomonic of a right lobar liver abscess.

Differential Diagnosis

- Lower thoracic lesions (basal pleurisy, pneumonia, and lung abscess), which irritate the diaphragm
- Acute cholecystitis
- Acute hepatitis.

Relevant Investigations

- Plain X-ray of abdomen or chest X-ray will reveal elevation of the right dome of diaphragm (Fig. 21.3). Right pleural effusion is common
- Fluoroscopy will show reduced mobility of the right dome
- US and CT (Fig. 21.4) are diagnostic
- Aspiration is confirmatory
- Stool examination is routine
- Isolation of organism in culture is required
- Serologic test (for amebiasis) is positive in majority of cases.

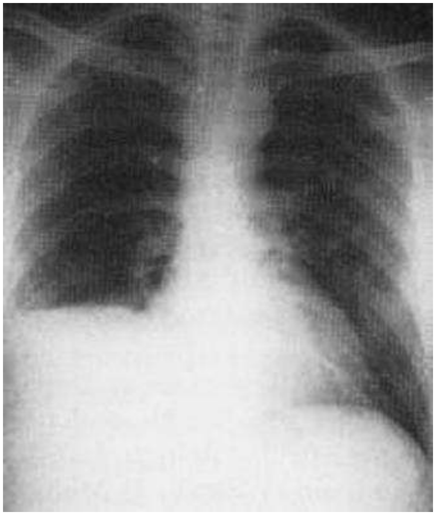

FIG. 21.3: Elevated right dome of diaphragm in amebic liver abscess

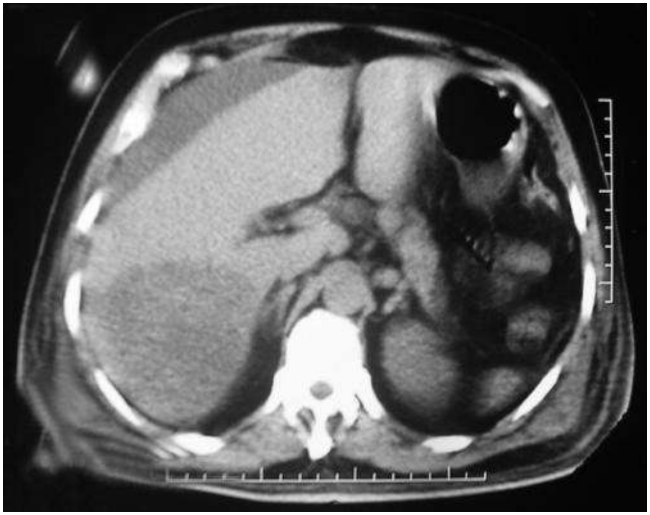

FIG. 21.4: CT—Amebic liver abscess right lobe with subphrenic collection

Treatment

- Antiamebic treatment is administered for small abscess and hepatitis of amebic etiology
- Drainage under US guidance with broad-spectrum antibiotics and antiamebic drugs is required for large abscess
- Growth of pyogenic organisms in culture of pus requires appropriate treatment.

ACUTE CHOLECYSTITIS

Incidence and Etiology

- Common in fat, flatulent, fertile, female of fifty (Five Fs)
- Ninety-five percent people with acute cholecystitis have gallstones.

Pathogenesis

- Inflammation of gallbladder occurs due to detergent action of bile (chemical cholecystitis), and infection by bacteria supervenes (commonly enteric organisms)
- Obstruction caused by a calculus in the cystic duct.

Complications

- Acute obstruction of the cystic duct causes distension and the bile can be replaced by mucus (mucocele) or pus (empyema)
- Ongoing inflammation causes gangrene and perforation of gallbladder (common in diabetics).

Symptoms

- Severe colicky pain (biliary colic) in the right hypochondrium, radiating to the inferior angle of the right scapula and the right shoulder
- Pain may be associated with vomiting.
- Fever and jaundice* are associated when there is associated cholangitis
- Symptom complex of pain, jaundice and fever with chills is called “Charcot’s triad”.

*Jaundice may occur due to the obstruction of the common bile duct caused by an impacted stone in the cystic duct (Mirizzi’s syndrome type 1)

Signs

- Tenderness at the tip of the right ninth costal cartilage (Murphy's sign)
- Hyperesthesia between the right 9th and 11th ribs posteriorly (Boas's sign)
- There may be associated guarding and rigidity in the right hypochondrium
- Palpable mass in the RUQ (inflamed and distended gallbladder).

Differential Diagnosis

- Acute appendicitis
- Acute right pyelonephritis
- Duodenal ulcer perforation
- Right basal pleurisy
- Myocardial infarction.

Relevant Investigations

- Elevated leukocyte count
- Deranged LFT (due to inflammation and edema of biliary tract)
- Plain X-ray abdomen (About 10% of gallstones are visible)
- US (Figs 21.5A and B) is diagnostic
- Rarely CT scan (Figs 21.6A to C) may be required
- Isotope scans are contributory.

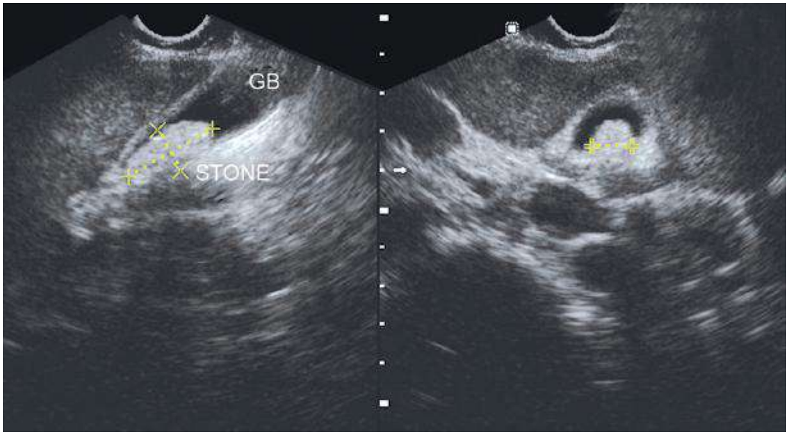

FIG. 21.5A: US—Stone in the cystic duct of gallbladder (*Courtesy:* Dr V Ganesan)

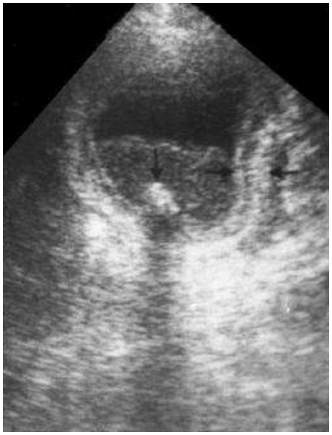

FIG. 21.5B: US—Acute edematous cholecystitis

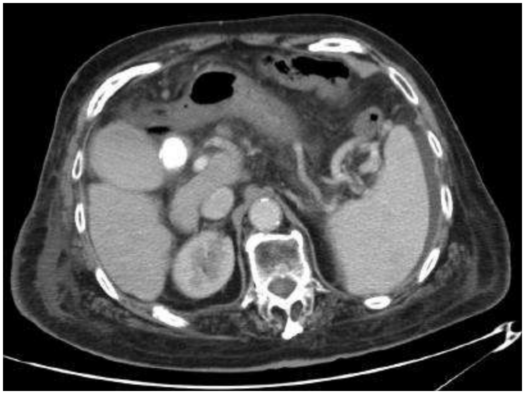

FIG. 21.6A: CT—Stone in the gallbladder

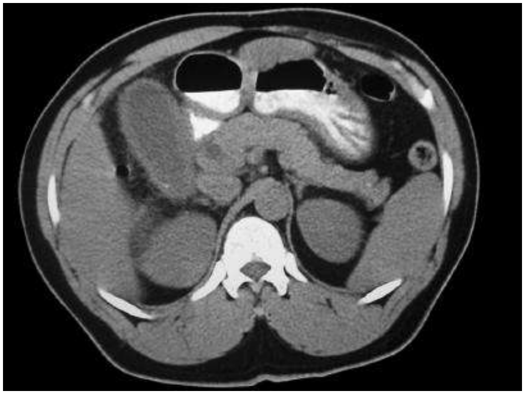

FIG. 21.6B: CT—Edematous gallbladder of acute cholecystitis

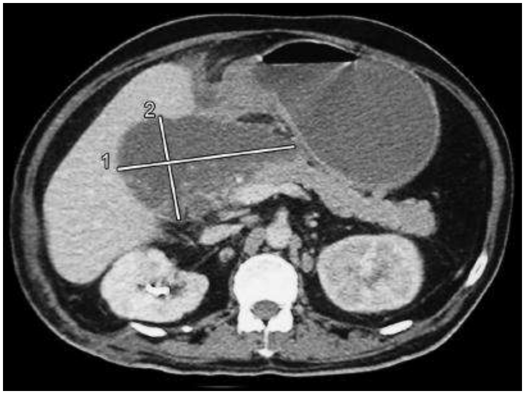

FIG. 21.6C: CT—Ruptured empyema of gallbladder

Treatment

- Acute cholecystitis is treated conservatively with antibiotics followed by elective cholecystectomy (Laparoscopic or open) **
- Common duct stones need to be removed by ERCP, sphincterotomy and basketing/stenting or by open choledochotomy, to relieve cholangitis and jaundice followed by cholecystectomy later.

** Laparoscopic cholecystectomy can be performed within 72 hours of onset of symptoms, with no higher conversion or complication rate, than a delayed operation.

ACUTE CHOLANGITIS

Incidence and Etiology

- Approximately 10 percent of patients with gallstones have associated common bile duct stones (choledocholithiasis)
- Obstructive lesions in the common bile duct can cause cholangitis (Fig. 21.7). They are:
    - Stricture
    - Tumor
    - Foreign body (e.g. stent)
    - Acute pancreatitis
    - ERCP

Pathogenesis

Obstructing lesion causes bile stasis, and this obstructed flow of bile encourages multiplication of bacteria, which have entered the biliary tract from the intestines through the sphincter of Oddi, causing ascending cholangitis.

Symptoms

RUQ pain, fever with rigors and jaundice (Charcot's triad).

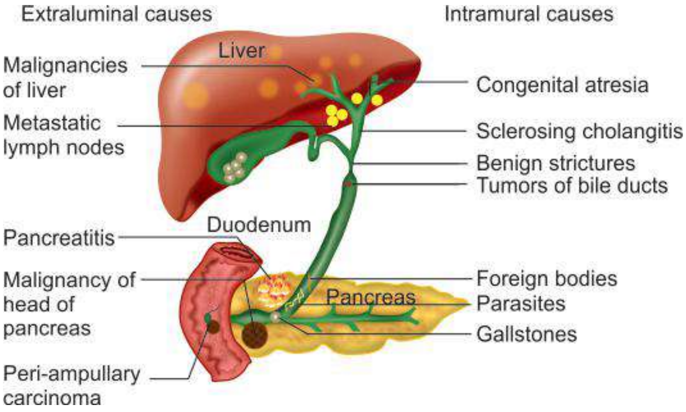
Extraluminal causes
Intramural causes
Liver
Malignancies of liver
Metastatic lymph nodes
Congenital atresia
Sclerosing cholangitis
Benign strictures
Tumors of bile ducts
Duodenum
Pancreatitis
Malignancy of head of pancreas
Pancreas
Foreign bodies
Parasites
Gallstones
Peri-ampullary carcinoma

FIG. 21.7: Causes of mechanical obstruction of biliary system

Signs

- Pyrexia, tachycardia, tachypnea
- Hypotension in shocked in individuals
- Jaundice
- Tenderness and guarding in RUQ.

Relevant Investigations

- Elevated leukocyte count
- Conjugated bilirubinemia
- Deranged LFT (due to inflammation and edema of biliary tract)
- Plain X-ray abdomen (About 10% of gallstones are visible)
- US and CT may reveal the dilated intra-hepatic radicals
- ERCP (Fig. 21.8) and MRCP (Fig. 21.9) will reveal the level and severity of stricture
- PTC (Fig. 21.10) is useful in impassable strictures
- MDCT with MPR technique (Fig. 21.11) has sensitivity of about 90 to 95 percent in localizing the causes of biliary obstruction.

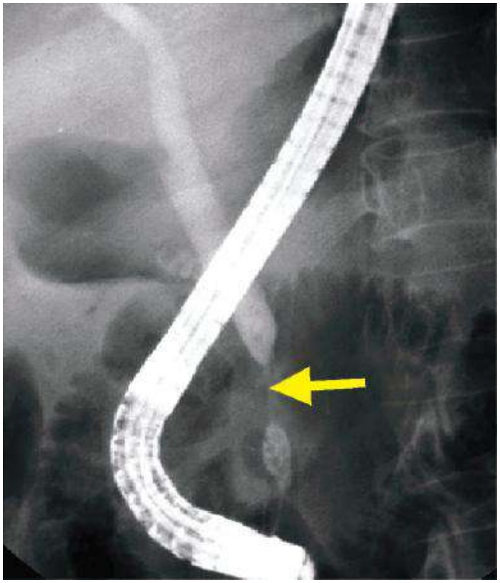

FIG. 21.8: ERCP—Benign stricture of CBD

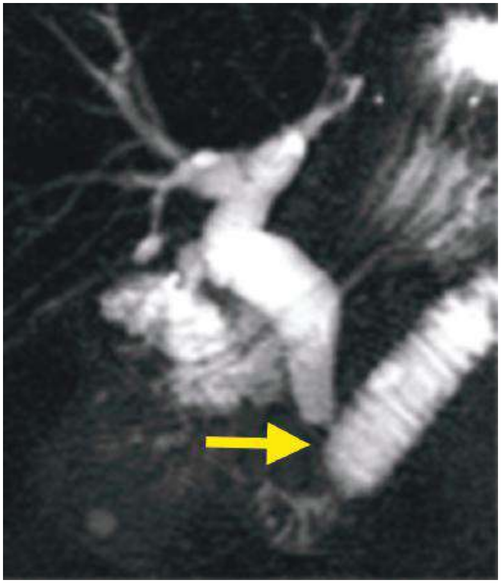

FIG. 21.9: MRCP—Benign stricture of CBD

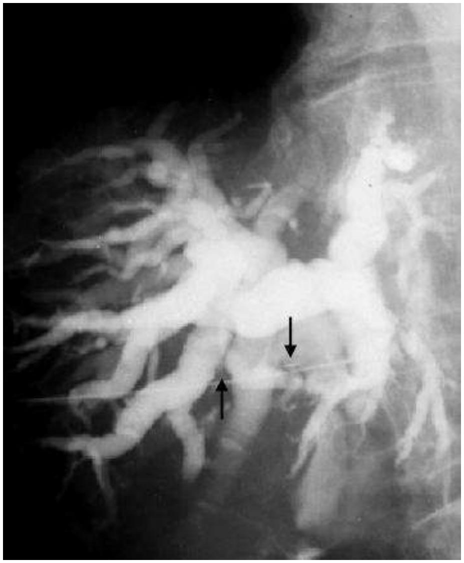

FIG. 21.10: PTC—Grossly dilated intrahepatic bile ducts due to CBD obstruction

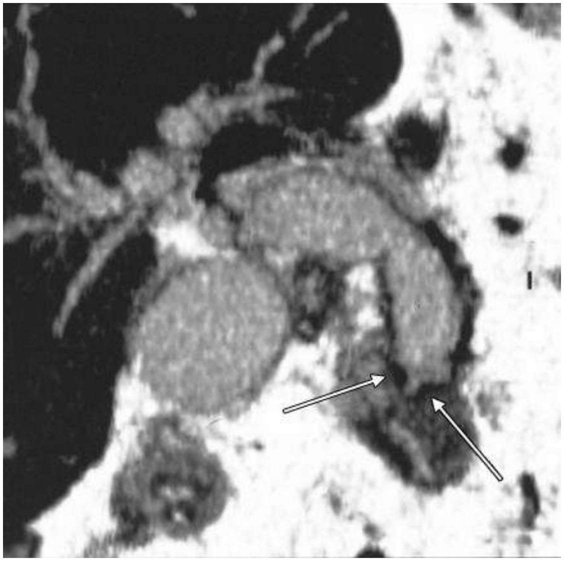

FIG. 21.11: MDCT—Cholangiocarcinoma

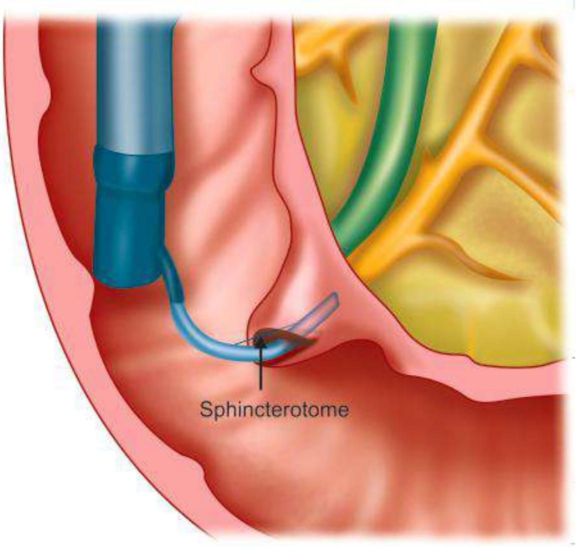

FIG. 21.12: Sphincterotomy

Treatment

- Inpatient care/Intensive care
- Analgesics
- Broad spectrum antibiotics
- Fluid resuscitation
- Emergency ERCP, sphincterotomy and basketing/stenting (Fig. 21.12) to relieve cholangitis and jaundice followed by cholecystectomy.
- Surgical treatment is required for impassable strictures
  - Resectable strictures can be resected with primary anastomosis
  - Strictures of the retropancreatic region require, choledocho-duodenostomy
  - Strictures of the common hepatic duct, require hepatico-jejunostomy.

Acute suppurative cholangitis is uncommon, but serious enough and if left untreated, has 100 percent mortality.

PRIMARY SCLEROSING CHOLANGITIS

Incidence and Etiology

- Rare condition of autoimmune origin
- Results in progressive fibrosis of the biliary system
- Causes luminal narrowing and progressive obstructive jaundice and secondary cirrhosis
- Strongly associated with inflammatory bowel disease, ulcerative colitis (75% of patients with PSC have inflammatory bowel disease and 7.5% with ulcerative colitis have PSC)
- Risk of malignant transformation (cholangiocarcinoma) is 15 percent.

Clinical Presentation

- Progressive obstructive jaundice
- Low-grade fever with chills, sweats.

Relevant Investigations

- CT (Fig. 21.13) and ERCP (Fig. 21.14) are diagnostic, classically shows diffuse stricturing and beading involving both intra and extrahepatic bile ducts, but indistinguishable from cholangiocarcinoma
- Liver biopsy may show the characteristic lesion of concentric fibrosis around small bile ducts, termed ‘onion skin’ fibrosis
- Peripheral antineutrophil cytoplasmic antibody (pANCA) is detected.

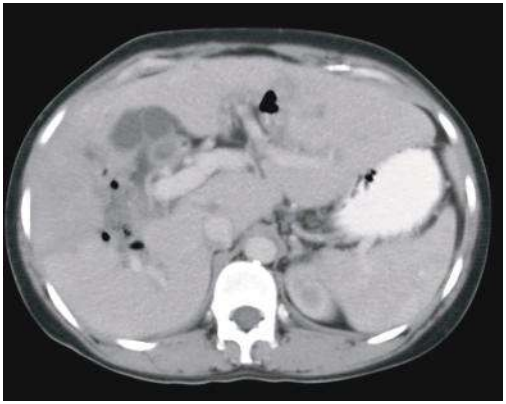

FIG. 21.13: CT—Sclerosing cholangitis

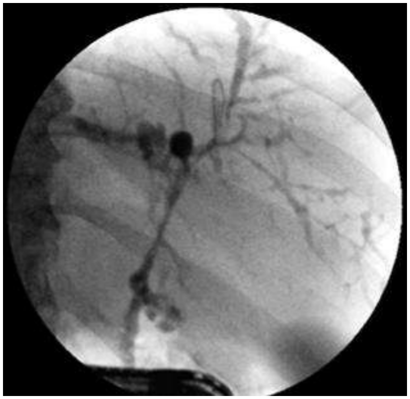

FIG. 21.14: ERCP—Sclerosing cholangitis

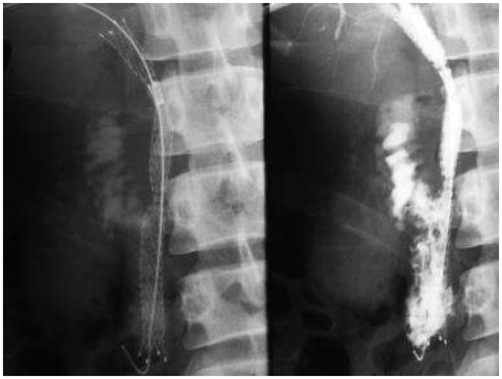

FIG. 21.15: Biliary stenting with endoprosthesis for cholangiocarcinoma

Treatment

- May settle spontaneously
- May respond to antibiotics, and UDCA
- Stenting of the biliary tree (Fig. 21.15) is very useful
- Liver transplantation is used widely for advanced disease.

SPLENIC ABSCESS

Incidence and Etiology

- Cyst of the spleen is a rare condition
- They can be:
    - Congenital
    - Parasitic (hydatid).

Pathogenesis

Severe systemic infections cause splenic abscesses which are usually multiple.

Symptoms

- Constitutional symptoms (e.g. high-grade fever)
- Previous history of severe intra-abdominal sepsis may be present.

Sign

Tender splenomegaly may be present.

Relevant Investigation

US and CT (Fig. 21.16) are diagnostic.

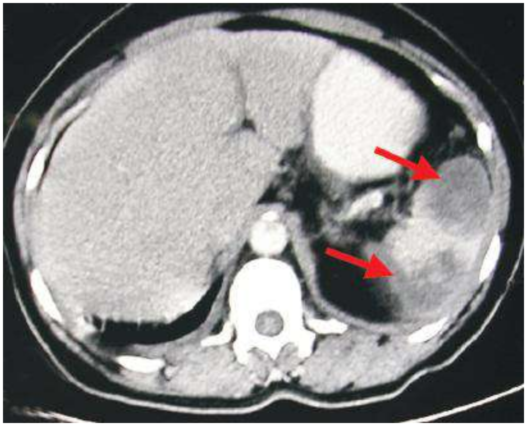

FIG. 21.16: CT abdomen—Splenic abscesses

Treatment

- Broad spectrum antibiotics
- Splenectomy is done for large spleen with abscesses.

ACUTE HEMORRHAGIC PANCREATITIS

Incidence and Etiology

- Common in the third decade
- Classification:
    - *Mild type:* Pancreatitis with minimal or no organ dysfunction and has a self-limiting course with uneventful recovery
    - *Severe type:* Pancreatitis with multiorgan failure and/or local complications like pseudocyst, necrosis or abscess
- Common causes are (nemonic: I get smashed)
    - I - Idiopathic
    - G - Gallstone (in the ampulla of Vater)
    - E - Ethanol
    - T- Trauma
    - S - Steroids
    - M - Mumps (paramyxovirus, Epstein-Barr virus, cytomegalovirus)
    - A - Autoimmune (Polyarteritis nodosa, Systemic lupus erythematosus)
    - S - Scorpion sting (Tityus Trinitatis), Snake bite
    - H - (Hypercalcemia, hyperlipidemia hypertriglyceridemia and hypothermia)
    - E - (Endoscopic Retrograde Cholangiopancreatography—ERCP)
    - D - Drugs (Steroids and Sulfonamides, Azathioprine, NSAIDs, Diuretics—SAND)
- Less common causes are:
    - Pancreas divisum
    - Long common duct
    - Carcinoma of the head of pancreas, and other cancer
    - Ascaris blocking pancreatic outflow
    - Chinese liver fluke
    - Ischemia from bypass surgery
    - Fatty necrosis
    - Pregnancy

  - Infections other than mumps, including varicella zoster
  - Repeated marathon running
  - Cystic fibrosis.
- Most common causes of pancreatitis, by demography are as follows:
  - *Western countries:* Chronic alcoholism and gallstones
  - *Eastern countries:* Gallstones.

Pathogenesis

- In acute pancreatitis, there is edema, hemorrhage and necrosis of the organ partly due to autodigestion. The inflamed pancreas may return to normal, but may recur, and likely to occur under two circumstances:
  - The initiating cause has not been removed (gallstones, alcohol consumption)
  - Major pancreatic necrosis, resulting in chronic pancreatitis or stricture of main pancreatic duct.

Complications

- Systemic complications:
  - Respiratory failure
  - Renal failure
  - Metabolic abnormalities
  - Coagulation disorders
  - Multiple organ failure.
- Local complications:
  - Pancreatic necrosis
  - Infection of pancreatic necrosis
  - Fungal infections
  - Hemorrhage
  - Pancreatic pseudocyst, pancreatic fistula, pancreatic abscess.

Symptoms

- Very severe, unbearable constant epigastric pain radiating to the back, relieved by sitting and bending forwards
- Nausea and vomiting are marked, frequent and persistent.

Signs

- Shock and cyanosis are marked
- Elevated temperature, tachycardia, tachypnea

- Epigastric tenderness but guarding and rigidity are not marked
- Retroperitoneal hemorrhage may lead to characteristic discoloration:
  - Around the umbilicus (Cullen's sign) (Fig. 21.17)
  - Of the flanks (Grey Turner's sign) (Fig. 21.18)

  (These signs are not pathognomonic of pancreatitis, as they can be seen with ruptured ectopic pregnancy as well).

Differential Diagnosis

- Perforated peptic ulcer
- Leaking aortic aneurysm
- Myocardial infarction
- Acute cholecystitis
- Acute hyperacidity

Relevant Investigations

- Elevation of serum amylase over 400 Somogyi units is indicative and more than 1000 Somogyi units is diagnostic (It usually rises 2 to 12 hours from the onset of symptoms, and normalizes within 48 to 72 hours)
- Serum lipase levels are elevated (It rises 4 to 8 hours from the onset of symptoms and normalizes within 7 to 14 days).

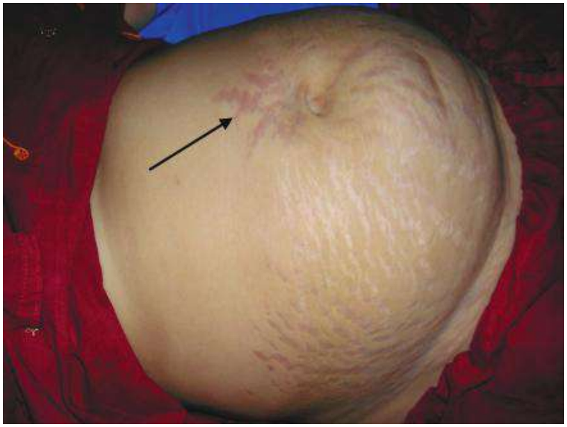

FIG. 21.17: Cullen's sign

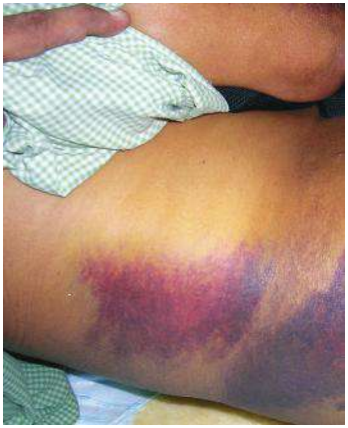

FIG. 21.18: Grey Turner's sign

Serum amylase may be normal (in 10% of cases) for cases of acute on chronic pancreatitis (depleted acinar cell mass) and hypertriglyceridemia. Reasons for false positive elevated serum amylase include salivary gland disease (elevated salivary amylase) and macroamylasemia. If the lipase level is about 2.5 to 3 times that of Amylase, it is an indication of pancreatitis due to alcohol

- Plain X-rays of abdomen show characteristic features:
  - ❖ 'Sentinel loop sign' – duodenum – which represents a focal dilated jejunal loop in the left upper quadrant
  - ❖ 'Cut off sign' – Transverse colon (Fig. 21.19) - Inflammatory exudate of acute pancreatitis extends into the phrenicocolic ligament directly spreading through the lateral attachment of the transverse mesocolon causing functional spasm and/or mechanical narrowing of the splenic flexure at the level where the colon returns to the retroperitoneum.

  (Absence of gas under the diaphragm eliminates the diagnosis of perforated duodenal ulcer)
- US may not be very useful at all times (Fig. 21.20)
- Contrast enhanced computed tomography (CECT) is very useful in assessing the size of the pancreas (Fig. 21.21), and also in determining the causes like the biliary or pancreatic duct calculi (Fig. 21.22).

Predicting the severity of an attack of acute pancreatitis is made on Glasgow or Imrie criteria (simplified Ranson criteria) Box 21.1.

Presence of three or more criteria reached before or at 48 hours of an attack predicts a severe attack and two or less predicts a mild attack.

APACHE II score (Acute Physiology and Chronic Health Evaluation) can be applied at anytime but it is cumbersome as it requires 15 different or biochemical criteria.

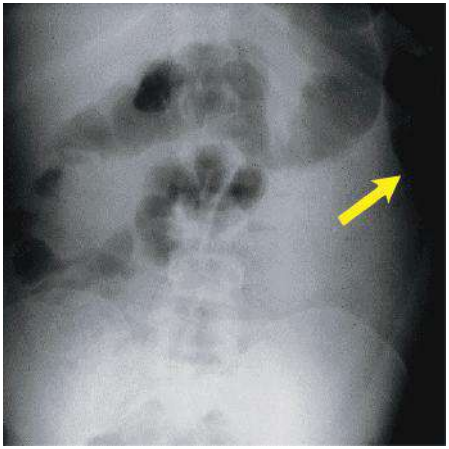

FIG. 21.19: X-ray—Colon cut-off sign

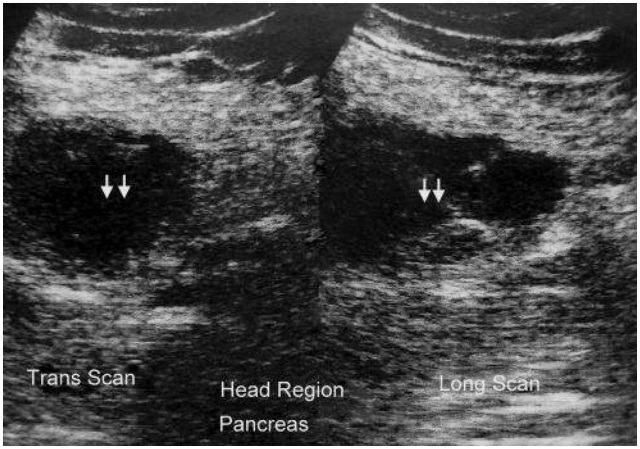

FIG. 21.20: US—Acute pancreatitis

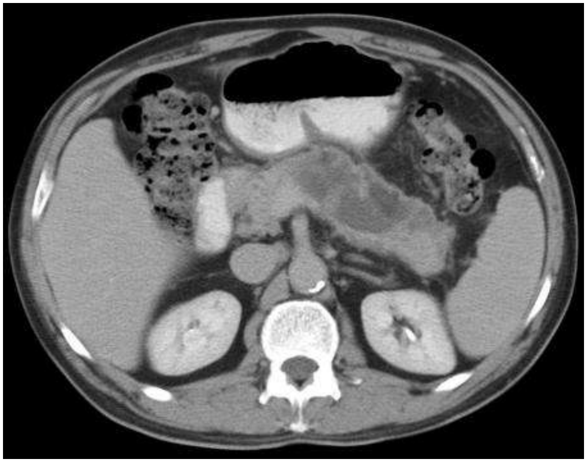

FIG. 21.21: CT—Enlarged and edematous pancreas - Acute pancreas

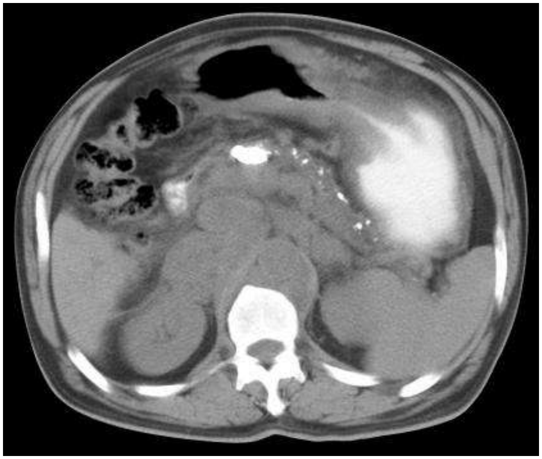

FIG. 21.22: CT—Pancreatic calculi

Box 21.1: Criteria to assess the severity of acute pancreatitis

P	Arterial PaO_2 <9 kpa
A	Albumin <32 g/L
N	Urea Nitrogen >10 mmol/L
C	Calcium <2 mmol/L
R	Raised white cell count >16 mmol/L
E	Enzyme – lactate dehydrogenase >600 mmol/L
A	Age >55 years
S	Sugar glucose >10 mmol/L

Treatment

- Initial management is conservative in intensive care unit:
    - Continuous arterial and CVP monitoring
    - Assisted ventilation if required
    - Inotropic support

  - Enteral feeding
  - Parenteral feeding if required
  - Hemodialysis of renal failure if warranted
- Endoscopic treatment:
  - ERCP sphincterotomy and extraction of stones followed by laparoscopic cholecystectomy
- Pancreatic necrosectomy is the treatment of choice with questionable outcome in a grave situation.

ACUTE APPENDICITIS

Incidence and Etiology

- Most common acute abdominal condition, occurring at any age
- Uncommon only below the age of two
- Most commonly seen between the age of 18 and 35.
- The two varieties of appendicitis are:
  1. Catarrhal appendicitis
  2. Obstructive appendicitis.

Pathogenesis

- *Catarrhal appendicitis:* Occurs due to acute inflammation of the appendix, which produces edema and even gangrene due to vascular involvement in inflammatory process
- *Obstructive appendicitis:* Caused by obstruction of its lumen by worms, fecoliths (Fig. 21.23) or hypertrophied lymphoid follicles. The appendix itself may be filled with pus (Fig. 21.24).

Complications

- When the adjacent tissues and omentum wall off the appendix or its perforation, it forms a mass called 'Appendicular mass'.
- When there is suppuration, it forms an abscess named 'Appendicular abscess', which may burst into peritoneal cavity to produce severe peritonitis and even death.

Symptoms

- *Pain:* A dull continuous ache starting at the umbilical region (visceral pain) and then localizing at the right iliac fossa (parietal pain) - catarrhal variety. Obstructive appendicitis presents with colicky pain (appendicular colic) in the right lower abdomen
- Nausea, vomiting and anorexia are usually present and are diagnostic

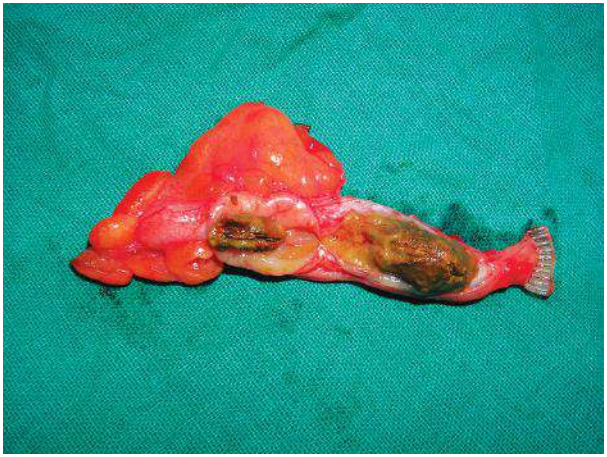

FIG. 21.23: Appendix with fecoliths

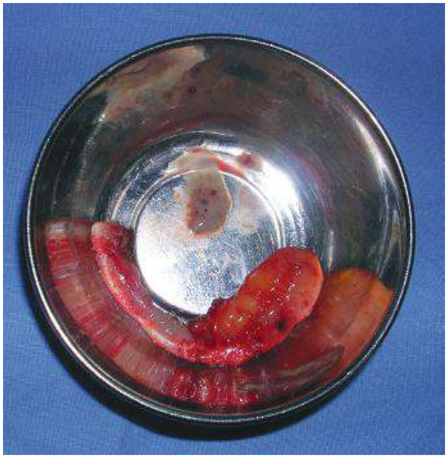

FIG. 21.24: Acute suppurative appendicitis (note the pus in the cup)

- Fever is the last to develop
  (Pain, vomiting and fever in appendicitis is called Murphy's syndrome).

Signs

- Hyperesthesia over Sherren's triangle
- Tenderness over McBurney's point
- Guarding and rigidity in the right iliac fossa
- Rovsing's sign (pain in the right iliac fossa on application of pressure in the left iliac fossa) may be elicited
- Tender mass may be felt (appendicular mass/abscess)
- Dullness on percussion (if mass already formed).

The signs and symptoms of appendicitis vary according to the position of the appendix (Fig. 21.25). They are given in Table 21.5.

Special Situations

- In children:
  - Constitutional symptoms like fever and tachycardia are more predominant
  - Use the child's hand itself for palpation, and if there is tenderness in the McBurney's point, the child will withdraw its hand
  - Appendicular mass is rare as the omentum is small in size and does not reach the appendix.

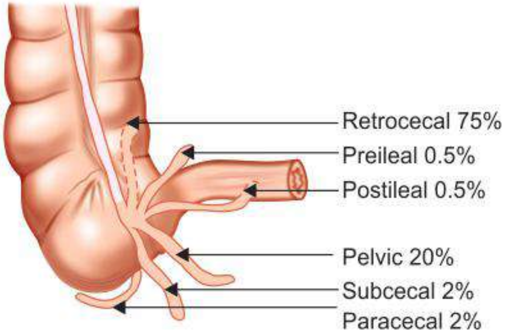

FIG. 21.25: Various positions of appendix

Table 21.5: Signs and symptoms of acute appendicitis related to the position of the appendix

Symptoms and signs	*Retrocecal and paracecal appendicitis*	*Pelvic and subcecal appendicitis*	*Pre and post-ileal appendicitis*
Pain	Right flank and back lateral to sacrospinalis muscle	Right iliac fossa	Right iliac fossa
Diarrhea	Absent	Absent	May be present
Tenderness and guarding	Not marked	Absent	Present
Positive test	Cope's psoas test Baldwing's test	Obturator test	Nil specific
Tenderness in rectal examination	Absent	Present	May be present

- In the elderly:
  - Guarding and rigidity are not pronounced as the abdominal musculature is weak
  - Incidence of gangrene is more as there is associated atherosclerosis
  - Peritonitis supervenes early.
- In pregnant women:
  - The point of tenderness is shifted up, as the appendix itself is pushed up by the enlarged gravid uterus
  - Pyelitis and cystitis of pregnancy adds to the difficulties in diagnosis of appendicitis
  - Accidental hemorrhage mimics acute appendicitis.

Cardinal signs and symptoms of acute appendicitis are:

- Periumbilical pain shifting to the right iliac fossa (very important)
- Association of nausea
- History of similar episodes in the past
- Tenderness at McBurney's point
- Guarding and rigidity in the right iliac fossa.

Differential diagnosis

- Right ureteric colic (most common)
- Right ovarian pathology
- Acute mesenteric adenitis
- Acute cholecystitis.

The differentiating features of acute appendicitis and right ureteric colic are given in Table 21.6.

Relevant Investigations

- US (Figs 21.26A and B) may be contributory. The immobile swollen appendix with free fluid in the right iliac fossa may be imaged. Mildly swollen appendix is generally not seen in an ultrasound scan. The scan helps to eliminate other lesions like the ureteric calculus, ovarian pathology, which can be imaged by US
- CT is useful in identifying inflamed appendix and (Fig. 21.27) appendicular mass. The signs of appendicitis are tabulated in Table 21.7.

Table 21.6: Differentiating features of acute appendicitis and right ureteric colic

Sign and symptom	*Pathology*	
	Acute appendicitis	*Right ureteric colic*
Pattern of pain	Periumbilical pain shifting to right iliac fossa	Loin to groin radiation on the right side
Nature of pain	Dull and continuous	Very severe and colicky with painfree intervals
Onset of pain	Slow and continuous	Sudden onset
Relationship to body movements	Aggravated	Not related to movement
Urinary symptoms	Absent	May be present
Rebound tenderness	May be present	Absent
Plain X-ray of abdomen	Nonspecific	Ureteric calculus

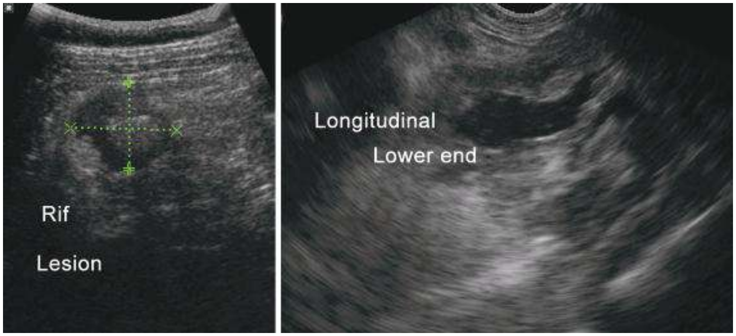

FIG. 21.26A: US—Inflamed appendix (*Courtesy*: Dr V Ganesan)

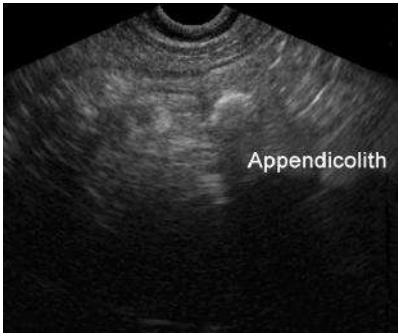

FIG. 21.26B: US—Appendicolith (*Courtesy*: Dr V Ganesan)

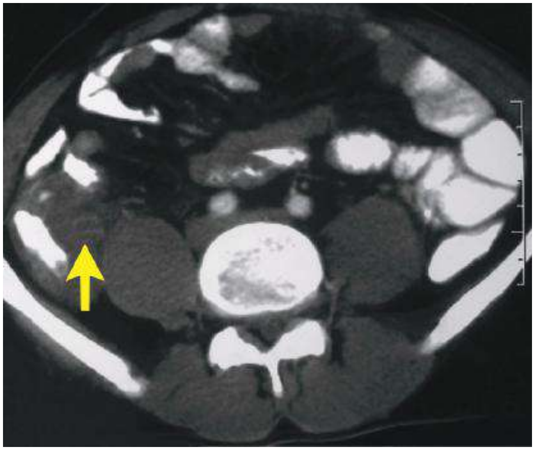

FIG. 21.27: CT—Appendicular mass

Table 21.7: CT findings associated with acute appendicitis

Appendiceal signs	*Periappendiceal signs*
Appendix >6 mm in anteroposterior diameter	Increased fat attenuation (stranding) in the right lower quadrant
Failure of appendix to fill with oral contrast or gas to its tip	Cecal wall thickening
Enhancement of appendix with IV contrast	Phlegmon in the right lower quadrant
Appendicolith	Abscess or extraluminal gas
	Fluid in the right lower quadrant or pelvis

Treatment

- Emergency appendicectomy (Laparoscopic or open) is the treatment of choice
- Conservative management is adopted for appendicular mass, subacute or chronic appendicitis (Oschner Scherren regime), followed by appendicectomy at a later date (interval appendicectomy)
- Appendicular abscess requires drainage, followed by interval appendicectomy.

ACUTE MESENTERIC LYMPHADENITIS

Incidence and Etiology

- Seen commonly in the pediatric age group
- Acute infection and inflammation of the mesenteric lymph nodes mostly of the ileocolic group (Fig. 21.28).

Pathogenesis

Commonly caused by viral infections.

Symptoms

- Periumbilical pain associated with high grade fever
- Vomiting is rare.

Signs

- The point of tenderness is usually in the right iliac fossa in the supine position, and shifts to the left side, if the patient is made to lie on the left side and vice versa ("Klein's sign")
- Guarding is not predominantly present
- In thin children, enlarged lymph nodes may be felt.

Differential Diagnosis

- Tubercular infection
- Acute appendicitis.

Relevant Investigations

- Clinical suspicion is important
- US (Fig. 21.29) and CT may show large lymph node swellings.

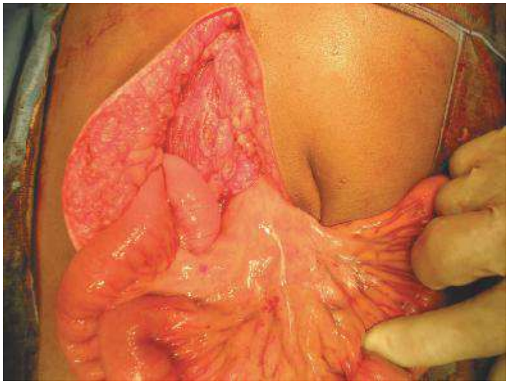

FIG. 21.28: Mesenteric lymphadenitis

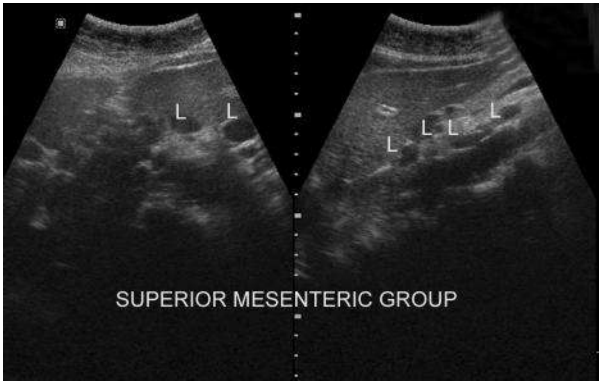

FIG. 21.29: US—Mesenteric lymphadenitis (*Courtesy*: Dr V Ganesan)

Treatment

Medical management with antibiotics and supportive measures.

ACUTE COLONIC DIVERTICULITIS

Incidence and Etiology

- Common in the West
- Disease of the elderly
- Sigmoid colon is most commonly affected.

Pathogenesis

- Acute inflammation of the diverticulum of large bowel (colonic diverticulitis)
- They are thought to arise from increased pressure in the colonic lumen, occurs at weak areas between the taeniae where vessels perforate through the submucosal layer.

Complications

- Bleeding
- Perforation
- Peritonitis
- Abscess.

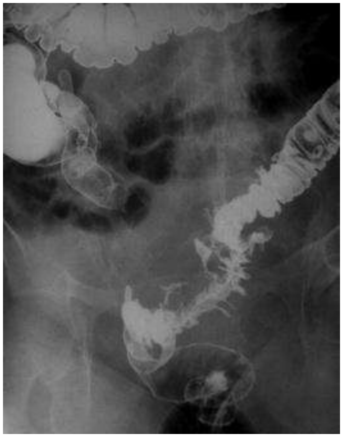

FIG. 21.30: Barium enema— Diverticulitis of colon

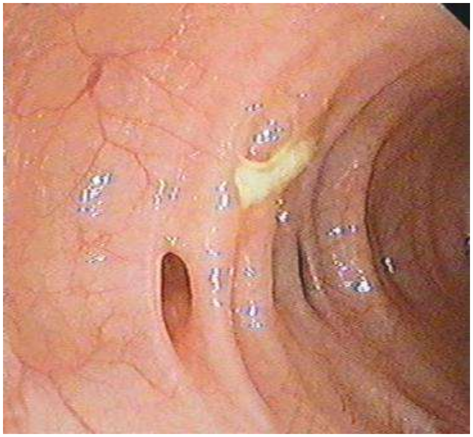

FIG. 21.31: Colonoscopy—Acute diverticulitis (*Courtesy*: Dr Mani Veeraraghavan)

The severity of diverticulitis is classified by Modified Hinchey classification

- Stage 1 – pericolic abscess
- Stage 2a – distant abscess amenable to percutaneous drainage
- Stage 2b – complex abscess associated with/without fistula
- Stage 3 – generalized purulent peritonitis
- Stage 4 – fecal peritonitis

Symptoms

- Colonic diverticulitis presents with colicky pain in the left flank
- When perforated, it may form an abscess in the paracolic region and present with high grade fever and a palpable tender lump.

Signs

- Left iliac fossa tenderness
- Tender mass (abscess) may be palpable.

Relevant Investigations

- Double contrast barium enema (Fig. 21.30) is informative
- Colonoscopy (Fig. 21.31) is diagnostic
- US and CT (Fig. 21.32) are useful in diagnosing abscesses.

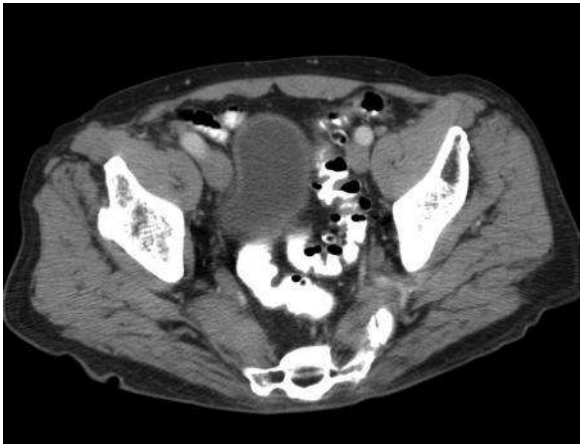

FIG. 21.32: CT—Diverticulitis of sigmoid colon

Treatment

- Uncomplicated diverticulitis needs to be treated with antibiotics
- Perforated diverticulitis with or without abscess formation needs surgical intervention.

ACUTE MECKEL'S DIVERTICULITIS

Incidence and Etiology

- Meckel's diverticulum* - An embryological remnant (Fig. 21.33) present as a free diverticulum with a wide mouth, about 25 to 30 cm from the ileocecal junction; may contain ectopic gastric mucosa (reason not known)
- Occurrence follows rough rule of 2s:
  - 2 percent of the population
  - 2 feet from the ileo-cecal junction
  - 2 inches in length
  - 2 times more common in males than in females.

Complications: Perforation and peritonitis.

---

* Meckel's diverticulum is often found incidentally during laparotomy and remains asymptomatic in majority of individuals.

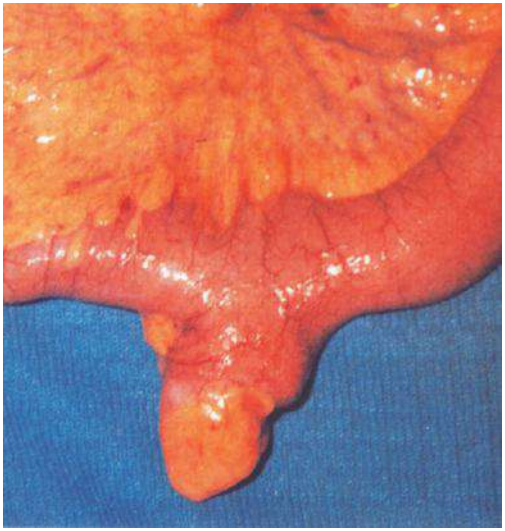

FIG. 21.33: Meckel's diverticulum

Symptom

Right iliac fossa pain.

Sign

Tenderness right iliac fossa (McBurney's point).

> Meckel's diverticulitis may present with peptic ulceration, lower GI hemorrhage, perforation, intussusception and intestinal obstruction.

Differential Diagnosis

- Acute appendicitis
- Right ureteric colic
- Right ovarian pathology in women.

Relevant Investigations

No investigation is useful.

Treatment

- Acute diverticulitis warrants diverticulectomy
- Perforation and peritonitis need appropriate management.

ACUTE SOLITARY CECAL DIVERTICULITIS

Incidence and Etiology

- Solitary diverticulum is common in the cecum
- Inflammation of the diverticulum presents like acute appendicitis and mislead the examiner.

Symptom

Pain in the right iliac fossa (similar to acute appendicitis).

Sign

A lump may be felt in the right iliac fossa.

> Clinically it mimics acute appendicitis, and many times, cecal diverticulitis is identified in second surgery, after appendicectomy.

Relevant Investigations

US and CT may be useful.

Treatment

Diverticulectomy or right hemicolectomy is the treatment of choice.

ACUTE ULCERATIVE COLITIS

Incidence and Etiology

- A chronic inflammatory bowel disease with relapses and remissions
- Acute inflammation is one of its clinical presentations.

Complications: Perforation and peritonitis.

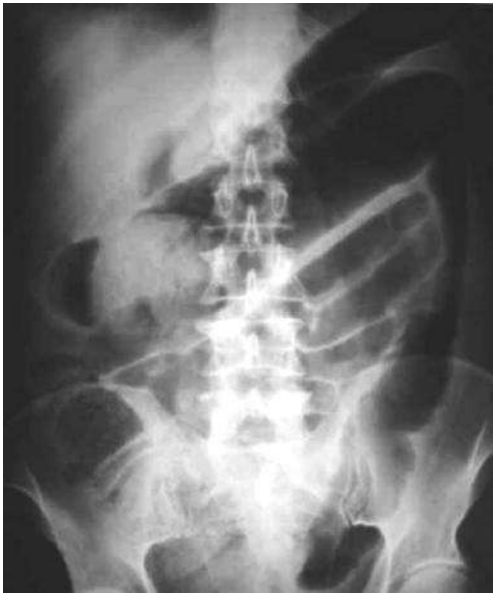

FIG. 21.34: X-ray—Toxic megacolon

Symptoms

- Incessant diarrhea, mixed with blood, mucus and pus with constitutional symptoms
- They present an emaciated appearance.

Signs

Deep ulcers involving the entire colon is called *Toxic megacolon*, presenting as grossly distended abdomen.

Relevant Investigations

- Plain X-ray abdomen (Fig. 21.34) is diagnostic
- Colonoscopy is essential for diagnosis.

Treatment

- Uncomplicated cases are treated medically
- Perforation and peritonitis need appropriate treatment.

ACUTE INTESTINAL OBSTRUCTION

- Intestinal obstructions may be dynamic or adynamic.
  - Dynamic obstruction is a physical or mechanical obstruction of the intestinal lumen due to various causes. They are:
    - Intramural causes (e.g. fecal impaction, worms)
    - Mural causes (e.g. strictures or stenosis due to tuberculosis, malignancies)
    - Extramural causes (e.g. adhesions, obstructed hernia)
  - Adynamic obstruction of the bowel is due to its neural (autonomic) paralysis, commonly seen after abdominal surgery in the immediate postoperative period, peritonitis or any other cause like fractures, tight bandages, etc.
- Intestinal obstruction may be at different levels. They are:
  - Small intestinal (high or low)
  - Large intestinal (colonic or sigmoid)
- Causes of intestinal obstruction are different for different age groups (Table 21.8).

Symptoms

1. Sudden episodic colicky abdominal pain
2. Vomiting
3. Constipation
4. Abdominal distension

   (The symptoms vary according to the level of obstruction)

- *Abdominal pain:* It is sudden and squeezing, and the patient doubles up. It is felt in the umbilical region, sometimes accompanied by the appearance of a contracting loop. There may be pain free intervals. Colonic pain presents in the hypogastrium

Table 21.8: Causes of intestinal obstruction in different age groups

Newborn	*Infants*	*Adolescents*	*Adults*	*Elderly*	*Rare causes*
Duodenal atresia	Helminths	Bands	Postoperative adhesions	Growth	Enteroliths
Pyloric stenosis	Intussusception	Intussusception	Intussusception	Intussusception	Foreign bodies
Meconium ileus		Meckel's diverticulum	Volvulus	Obstructed or strangulated hernia	Gallstones
Hirschsprung's disease		Obstructed or strangulated hernia	Obstructed or strangulated hernia		Trichobezoar
			Growth		Phytobezoar

- *Vomiting:* Vomiting is predominant in high obstructions. The vomitus consists of gastric contents, followed by the duodenal and lastly the intestinal, depending on the level of obstruction. In the late stages, the vomitus becomes feculent – ominous sign. Vomiting by itself is a late sign of chronic intestinal obstruction
- *Constipation:* The patient evacuates his bowel (contents distal to obstruction) once or twice, and constipation becomes a noticeable feature after 24 hours

Diarrhea can be a feature in certain situations like intussusception (red currant jelly stools), Richter's hernia, adynamic obstruction caused by mesenteric vascular occlusion, pelvic abscess, etc.

- *Abdominal distension:* Common feature of intestinal obstruction. Distension is
  - centrally located in small bowel obstruction (ladder pattern),
  - more on the flanks when distal colon is obstructed (asymmetrical)
  - more on the left flank in sigmoid volvulus
- *Dehydration:* When vomiting is pronounced as in high level obstructions, dehydration is a presenting feature.

Early dehydration and less abdominal distension suggests duodenal or jejunal obstruction whereas, late dehydration and more abdominal distension suggests distal ileal obstruction. Vomiting and dehydration are usually not present in isolated acute colonic obstruction.

Signs

- *General:* Pulse rate and blood pressure are maintained at normal levels in the initial stages. As dehydration becomes prominent, tachycardia and hypotension result
- *Abdomen:* Bowel sounds are not heard as obstruction worsens.

The summary of signs and symptoms related to intestinal obstruction are given in Table 21.9.

Relevant Investigations

- Plain X-rays of abdomen in the erect posture will reveal multiple air fluid levels (Fig. 21.35) and colonic obstruction may show distended colon also (Fig. 21.36).

Table 21.9: Signs and symptoms related to intestinal obstruction

Level of obstruction	*Signs and symptoms*					
	Duration of colic	*Pain free interval*	*Vomiting*	*Distension of abdomen*	*Constipation*	*Dehydration*
High	Short	Short	More	Minimal	Not constant	Severe
Low	Long	Long	Less	More	Late feature	Mild-to-moderate

(Findings of clinical examination and treatment of individual diseases are discussed at appropriate headings in the following pages)

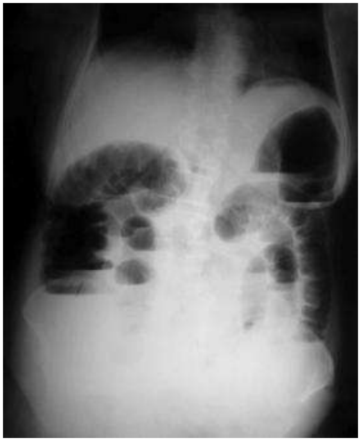

FIG. 21.35: X-ray—Multiple air fluid levels of small bowel obstruction

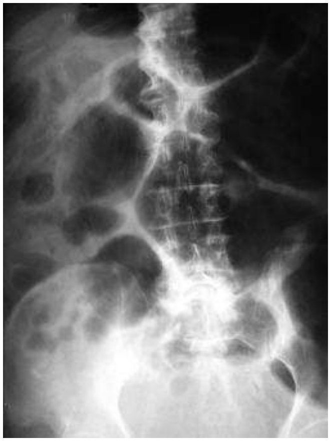

FIG. 21.36: X-ray—Distention of large bowel

Treatment

- Inpatient/intensive care
- Nil by mouth
- Intravenous fluids
- Correction of electrolytes
- Nasogastric decompression
- Urinary catheterization of better monitoring
- Identifying and treating the cause.*

---

* If adhesions are the likely cause, a trial of conservative management is justified.

GALLSTONE ILEUS

Incidence and Etiology

Pathogenesis

- Gallstone enters the bowel through a perforated gallbladder (postcholecystitis) adherent to the small bowel (cholecysto-enteric fistula)
- When the stone is big and reaches the ileocecal junction, it causes small bowel obstruction.

Symptoms

Previous history of vague attacks of right upper quadrant pain, suggesting frequent cholecystitis.

Signs

Features of intestinal obstruction.

Relevant Investigations

- Plain X-ray abdomen (Fig. 21.37) almost always shows air in the biliary tree as bowel gas passes through the cholecystoenteric fistula. A gallstone may also be seen in the right lower quadrant, if it is radiopaque
- CT (Fig. 21.38) is more informative. Rarely, a gallstone may also be seen in the small bowel, and also in the gallbladder if there are many.

Treatment

- During laparotomy:
  - Simple crushing of stone with finger from outside the bowel may be enough if it is soft
  - Simple surgical removal through an enterotomy is required if the stone is hard and big
  - Cholecystectomy must be performed with closure of fistula.

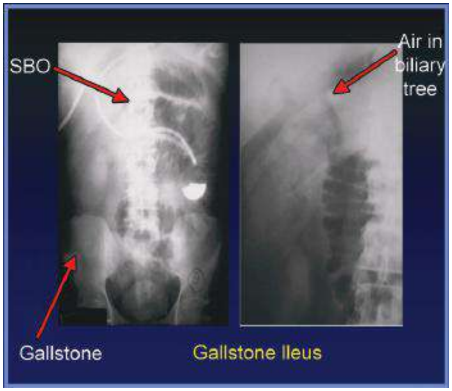

FIG. 21.37: X-ray—Gallstone ileus

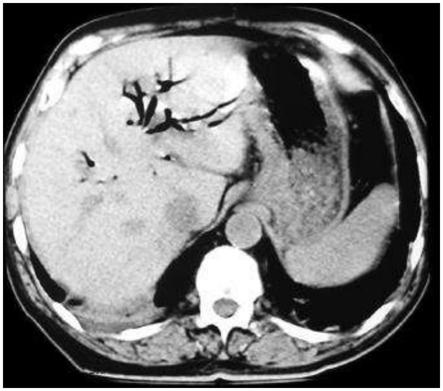

FIG. 21.38: Air in the biliary system

ACUTE INTUSSUSCEPTION

Incidence and Etiology

Intussusception is the invagination of a segment of bowel into the distal adjacent loop (proximal into the distal) (Fig. 21.39).

In Children

- Two per thousand infants are affected with male preponderance, commonly affecting the age group of 3 months to 1 year
- Commonly, it is secondary to an enlarged Peyer's patch due to viral or bacterial infections
- The other less common causes are:
    - Meckel's diverticulum
    - Duplication cyst in the bowel wall
    - Polyp.

In Adults

- Intussusception of small bowel is always secondary to a polypoid lesion, a lipoma (Figs 21.40A and B)
- In large bowel, it is due to a malignant polypoid lesion (Figs 21.41A and B).

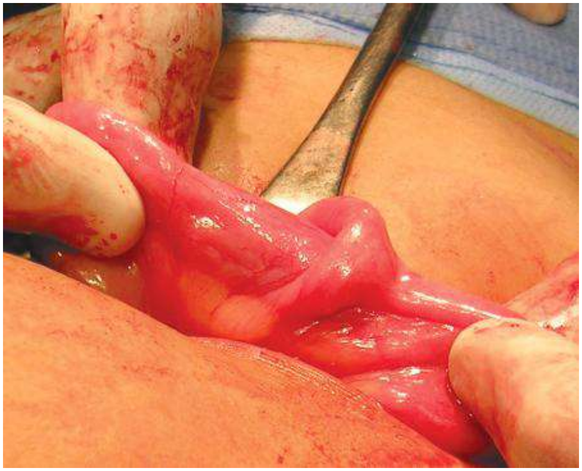

FIG. 21.39: Ileoileal intussusception

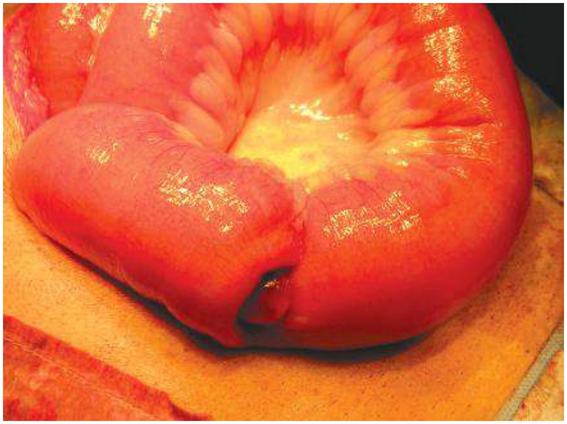

FIG. 21.40A: Ileoileal intussusception

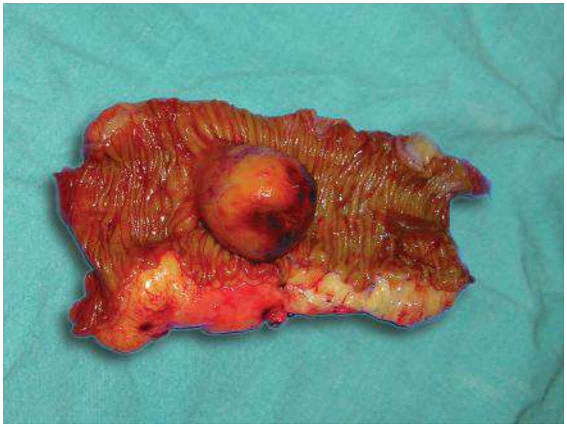

FIG. 21.40B: Lipoma the cause of ileoileal intussusception

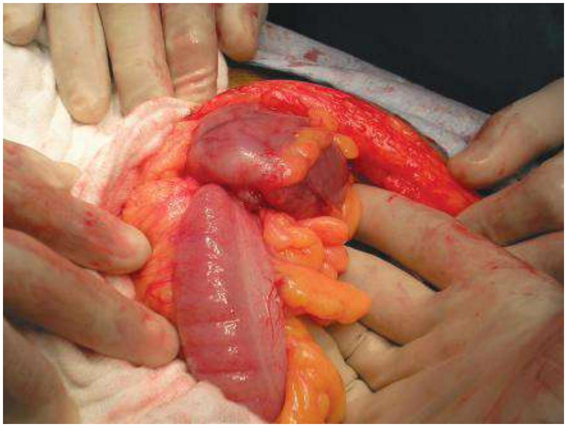

FIG. 21.41A: Colocolic intussusception

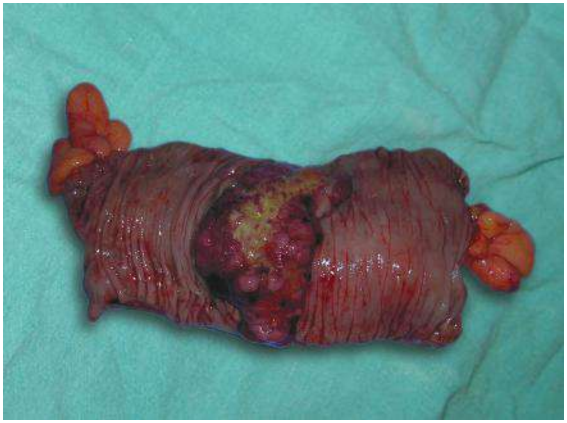

FIG. 21.41B: Malignant growth of colon cause for intussusception

Complications: When the mesentery is drawn between the loops, it may result in vascular compromise, which may lead to strangulation, gangrene and perforation.

Symptoms

- In children, there may a history of preceding gastroenteritis following a change in diet (weaning from milk to solid food)

- Severe acute colicky pain, with abdominal distension
- Passing of frequent semisolid stools with bright red blood may be predominant (red currant jelly).

Signs

- During the attacks of pain, a sausage-shaped mass may be felt, which appears during the time of colic and disappears after the colic disappears. The right iliac fossa is empty – Sign de Dance
- Rectal examination may reveal bloodstain on the examining finger (red-currant jelly)
- Colo-rectal intussusception may be felt by the examining finger on rectal examination, or it may even present through anus, resembling a rectal prolapse.

Relevant Investigations

- Plain X-ray abdomen – soft tissue shadow in the region of transverse colon with empty distal colon. Multiple air fluid levels may be seen when obstruction predominates
- Barium enema may show a filling defect called pincer-shaped filling defect (caused by the intussusceptum with the intussuscipiens)
- Colonoscopy can identify, colonic intussusceptions (Fig. 21.42A)
- US and CT (Figs 21.42B and C) will reveal the intussuscepting mass (pseudokidney appearance).

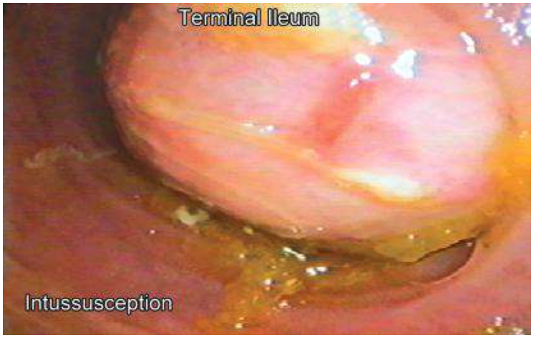

FIG. 21.42A: Intussusception of small bowel (*Courtesy*: Dr Mani Veeraraghavan)

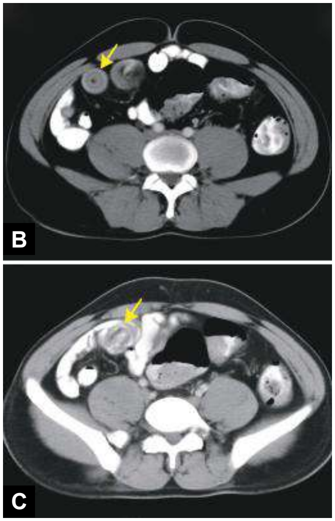

FIGS 21.42B and C: CT—Intussusception

Treatment

- Barium enema and colonoscopy, by themselves may reduce the colonic intussusception
- Laparotomy is required to reduce the small bowel intussusception, and treat the cause appropriately
- Bowel resections may be needed if the bowel segment is strangulated, and nonviable
- Perforation and peritonitis need appropriate treatment.

SWALLOWED FOREIGN BODIES

Incidence and Etiology

- Common in children and psychotic individuals
- Most foreign bodies pass through GIT without any difficulty.

Symptoms

- Abdominal pain may be present
- Vomiting may supervene.

Signs

- Signs of intestinal obstruction may appear
- Chest signs may appear.

Examine the oral cavity, oropharynx or nasopharynx for impacted foreign body, before asking for radiographs.

Relevant Investigations

- Plain X-ray of throat (Fig. 21.43), chest and abdomen may identify the foreign body
- Bronchoscopy/Gastroscopy may be required.

- By the time the patient is seen by the clinician, the foreign body might be expelled through feces, if it is small and blunt
- Since button cells have the tendency to erode through the bowel wall, it is better to monitor its position by serial radiography, and removed if stuck in a place.

Treatment

- Blunt objects reaching the stomach will pass without difficulty
- Sharp foreign bodies impacted in the GIT should be removed (endoscopically or by open surgery)
- Catastrophic bleed, obstruction and perforation peritonitis need laparotomy.

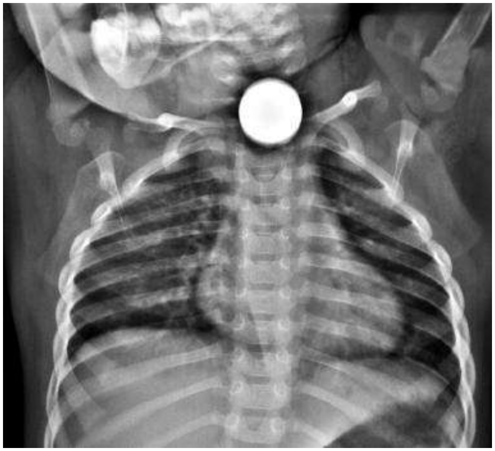

FIG. 21.43: Foreign body larynx (*Courtesy:* Bharat scans)

PERFORATED BOWEL PATHOLOGIES

There are a variety of intra-abdominal pathologies, which are serious emergencies, and also threaten life, if not recognized early and treated. Such conditions are called “acute abdominal catastrophes”. Most of these conditions are a result of infection of peritoneum called ‘peritonitis’, caused by the leak of infected intestinal contents into the peritoneal cavity. Since the peritoneum is a semi-permeable membrane, it allows the spread of infection systemically into the blood to result in septicemia, a fatal condition.

Peritonitis consists of three stages. They are:

1. *Stage of peritonism:* Occurs due to the chemical irritation of the peritoneum caused by the leaked contents of the bowel, and the patient feels sudden excruciating pain. Constitutional symptoms are not prominent, though tenderness and guarding can be elicited
2. *Stage of reaction:* The contents of the bowel are neutralized by the peritoneal exudates, and pain becomes less. Muscle guarding may be present, and shifting dullness may be elicited indicating the presence of free fluid in the abdomen
3. *Stage of spreading peritonitis:* This is an ominous stage, as there are signs of severe infection. They exhibit classic facies, called “Facies Hippocratica” – sunken eyes, hollow cheeks and anxious face. Severe tachycardia (fast thready pulse) is present. Abdomen shows board like rigidity, indicating generalized peritonitis. Death becomes imminent if not attended to, surgically.

The signs and symptoms of perforations caused by various parts of the bowel have certain indicative findings. They are given in the Table 21.10.

Table 21.10: Signs and symptoms of perforations caused by various diseases of the bowel

Signs and symptoms	*Perforation of*						
	*Duodenal or gastric ulcer (benign/ malignant)**	*Appendicitis***	*Cholecystitis*	*Typhoid/ Tubercular ulcer*	*Ulcerative colitis*	*Diverticular disease*	*Colonic malignancy*
Previous history	Frequent use of antacids	Nil or pain in the right iliac fossa in the past (treated or untreated)	Gallstones or chronic cholecystitis	Prolonged fever (high grade in typhoid and low grade in tuberculosis)	Repeated attacks of diarrhea and abdominal pain	Repeated attacks of constipation and abdominal pain	Constipation
History of drug intake	NSAIDs	Nil	Nil	Treatment for typhoid fever	Medical management	Medical management	Laxatives
Area of abdominal pain	Upper abdominal	Umbilical to start and then in the right iliac fossa	Right hypochondrial	Umbilical or lower abdominal	Flanks and hypogastric	Umbilical or flanks	Flank on the side of lesion
Gastrointestinal bleeds	Upper	Nil	Nil	Lower	Lower	Lower	Lower
Treatment	Surgical						

* Leaks of gastric contents due to perforated gastric ulcer (posteriorly placed), into the lesser sac may mask abdominal symptoms

** Clinical presentation of perforated Meckel's diverticulitis is the same as that of perforated acute appendicitis.

Differential Diagnosis

- Myocardial infarction
- Diaphragmatic irritation caused by lower lobar lung lesions
- Acute pancreatitis
- Ruptured or dissecting aortic aneurysm.

Relevant Investigations

- Plain X-rays of abdomen in the erect posture is confirmatory (gas under the diaphragm) in most cases (Figs 21.44 and 21.45) and peritonitis gives the classic 'ground glass appearance' (Fig. 21.46)
- CT (Fig. 21.47) is useful.

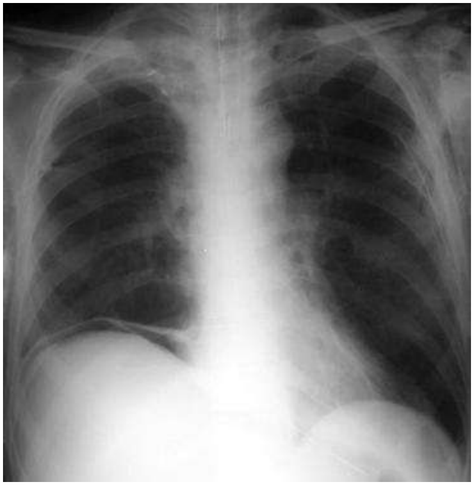

FIG. 21.44: X-ray—Air under the right dome of diaphragm-duodenal ulcer perforation

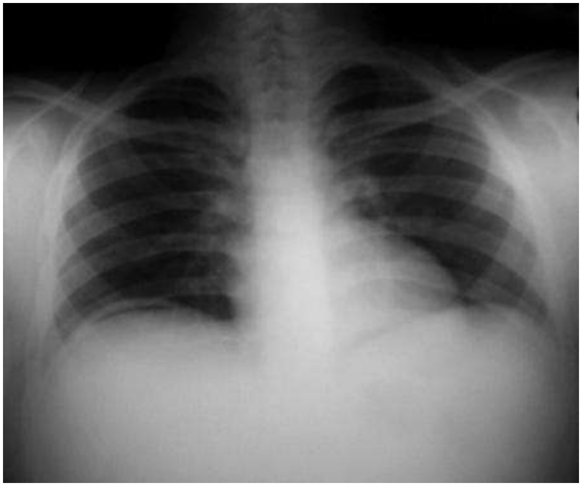

FIG. 21.45: Minimal air under right dome of diaphragm - appendicular perforation

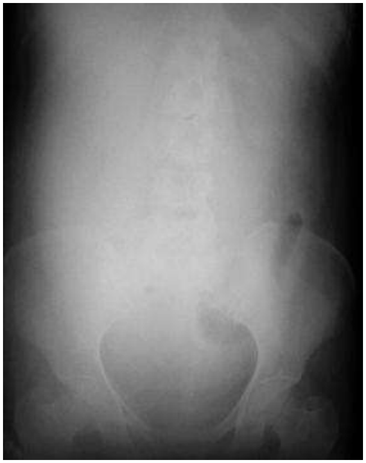

FIG. 21.46: X-ray abdomen - Ground glass appearance of peritonitis

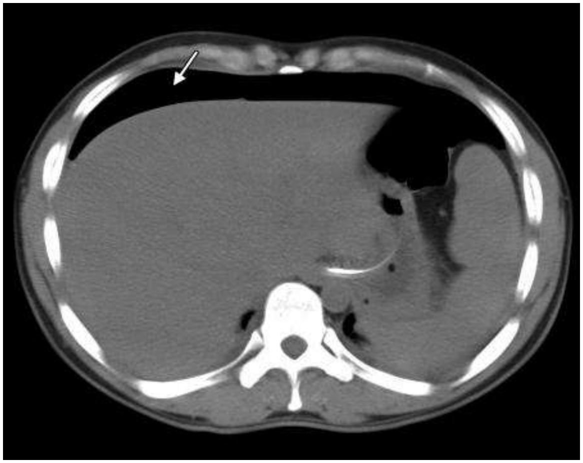

FIG. 21.47: CT—Pneumoperitoneum

Treatment

Treatment is always surgical:

- Closure of bowel perforation if it is simple and of benign cause
- Cholecystectomy (gallbladder perforations)/ appendicectomy (appendicular perforations)/appropriate bowel resections (benign or malignant bowel perforations)
- Peritoneal toileting is mandatory and should be complete.

INTESTINAL STRICTURES

Incidence and Etiology

- Obstructions can be caused by strictures due to:
  - Tuberculosis (healing lesions)
  - Malignant lesions
- Postoperative strictures are not uncommon.

Symptoms

- Symptoms of subacute or acute intestinal obstruction
- History of weight loss, low-grade pyrexia, anemia and vague abdominal pain may be present.

Signs

- Clinical examination may show a mass in the right iliac fossa (differential diagnosis—Crohn’s disease)
- Ascites may be present.

Relevant Investigations

- Plain X-ray abdomen (erect) shows air fluid levels
- X-ray chest may reveal a primary tubercular lesion.

Treatment

- Laparotomy is needed for acute obstructions
- Stricturoplasty or bypass procedures or resections are done for tubercular strictures
- Radical resections or bypass procedures are done for malignant strictures.

BANDS AND ADHESIONS

Incidence and Etiology

- Bands may be congenital, but adhesions are post surgical
- The adhesions may be like a small band or generalized, flimsy or dense, single (Figs 21.48 and 21.49) or multiple.

Symptoms and Signs

Features of subacute or acute intestinal obstruction, months or years after surgery.

Relevant Investigations

- Plain radiographs of abdomen are useful in diagnosis
- Diagnostic laparoscopy is conclusive.

Treatment

- Release of adhesions by laparoscopy or by open surgery
- Open surgery and release of adherent bowel loops for dense adhesions obstructing the bowel
- It is better to manage the acute obstructions by trial medical management with gastric suction and intravenous fluids

  Note: Surgery has the disadvantage of recurrence, especially in generalized adhesions.

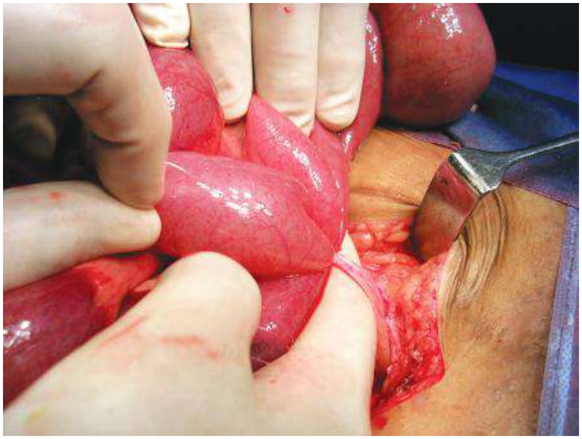

FIG. 21.48: Band in the pelvis causing internal rotation of small bowel

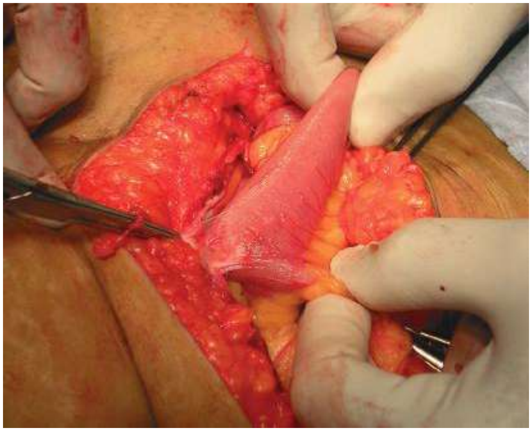

FIG. 21.49: Adhesions of small bowel to previous surgery scar

ENTEROLITHS/FOOD BOLUS

Incidence and Etiology

Obstructions may be caused by enteroliths or food bolus, resulting from poor chewing in an edentulous patient, high consumption of high fiber (e.g. orange pith), usually at a pre-existing narrowing due to tuberculosis, Crohn's disease and surgery.

Symptoms and Signs

Symptoms of subacute or acute intestinal obstruction.

Relevant Investigations

Plain X-ray abdomen (erect) shows air fluid levels. Diagnosis is more clinical and is difficult.

Treatment

Surgery is indicated for acute obstructions, and the bolus or enterolith can be milked into the large intestine and rarely it is necessary to open the bowel to remove it.

VOLVULUS

- Volvulus is defined as a twist of the bowel around its mesenteric axis
- It is more common in the large bowel (commonly the sigmoid colon) than in the small bowel
- Rotation of more than 180 degrees may result in strangulation.

SIGMOID VOLVULUS

Incidence and Etiology

Disease of the middle aged and elderly.

Pathogenesis

Rotation of the sigmoid around its axis occurs when its mesentery is unusually long.

Symptoms

- Sudden severe pain, frequently when straining to pass stool. The patient retches and develops hiccoughs
- The patient may give history of attacks of abdominal pain with constipation, relieved by passing watery stools and large volumes of flatus.

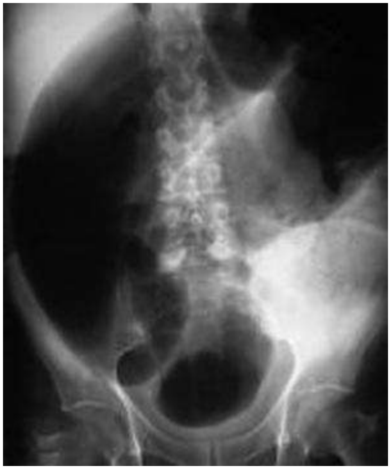

FIG. 21.50: X-ray – Sigmoid volvulus

Signs

- Abdomen rapidly distends, disproportionate to the duration of pain, and the distension is confined more to the left flank
- Rectum is empty on examination.

Relevant Investigations

Plain radiograph of the abdomen will reveal a distended sigmoid, coffee bean appearance – Freeman Dahl sign (Fig. 21.50) (Convergence of three white lines towards the base of the pedicle).

Treatment

- Untwisting of the volvulus, and fixing the colon to parietal peritoneum to prevent recurrence
- Sigmoidectomy is the treatment of choice in long redundant sigmoid colon.

CECAL VOLVULUS

Incidence and Etiology

- This occurs in those whose entire right colon has a mesentery continuous with that of the small bowel, and the cecum does not lie in the right iliac fossa
- More common in women and during pregnancy.

Clinical Presentation

- A distended, tense palpable resonant mass in the umbilical region, with an empty right iliac fossa
- Features of distal small bowel obstruction.

Relevant Investigations

Plain radiograph of the abdomen is diagnostic (Cecal bubble is seen).

Treatment

Untwisting of the volvulus with or without resection of the segment of the bowel and cecopexy is required.

MIDGUT VOLVULUS

Incidence and Etiology

- In children, this occurs due to malrotation and failure of fixation of midgut, usually before one year, rarely in neonates
- In adults, a loop of bowel rotates around a point of adhesion (to the abdominal wall or to an adjacent viscera).

Symptom

Severe abdominal pain.

Signs

Signs of intestinal obstruction.

Relevant Investigation

Contrast study shows 'Cork-screw effect'.

Treatment

Emergency laparotomy and correction of the defect with fixing of the bowel.

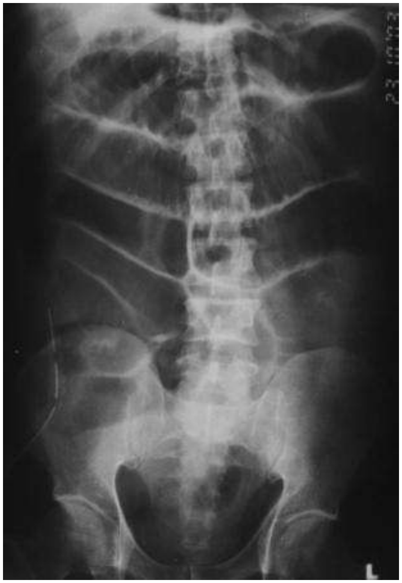

FIG. 21.53: Step-ladder pattern of paralytic ileus

Treatment

- Correction of electrolyte imbalances
- Treatment of retroperitoneal or intraperitoneal causes like hemorrhage and sepsis.

TORSION OF MESENTERIC CYST

Incidence and Etiology

- Cystic lesions of mesentery are more common than solid tumors (2:1)
- Cystic tumors of mesentery are:
  - Chylous cyst, serous cyst (developmental) – more common
  - Lymphangioma (lymphatic tissue) – more common
  - Traumatic cyst (trauma)
  - Enteric cyst, dermoid (embryonic rests)
- Majority of cysts are benign (except lymphangiosarcoma, malignant teratoma).

Pathogenesis

Torsion occurs more by its weight, and mobile nature of mesentery encourages this torsion.

Symptom

Sudden acute abdominal pain.

Signs

- Central abdominal tenderness with guarding and rigidity
- A cystic lump may be felt in the central abdomen.

Relevant Investigation

US and CT of abdomen are diagnostic.

Treatment

Emergency laparotomy is needed. The cyst should be removed.

TORSION OF OMENTUM

Incidence and Etiology

- More common in 4th to 5th decades
- Equal sex incidence
- Causes of omental torsion are:
  - *Primary:* Always unipolar, cause is unknown
  - *Secondary:* Usually bipolar, associated with adhesions (intra-abdominal inflammation—tuberculosis, postsurgical adhesions, internal and external herniae).

Pathogenesis

- The omentum twists on its long axis to an extent causing vascular obstruction
- May vary from mild vascular obstruction producing edema to complete strangulation leading to infarction and gangrene
- The situations to cause torsion are:
  - Redundant and mobile segment
  - Fixed point around which the segment can twist

- The precipitating factors for torsion are:
  - Anatomic variations (bifid omentum, accessory omentum, large and bulky omentum)
  - Venous redundancy
  - More common in right side (due to increased bulk and mobility).

Symptoms

- Acute abdominal pain usually is localized to right lower quadrant
- Movement increases the pain
- Nausea and vomiting are common.

Signs

- Tenderness, rigidity and guarding of abdomen
- A vague tender mass may be felt in the upper abdomen.

Differential Diagnosis

- Acute cholecystitis
- Acute appendicitis
- Torsion of right ovary.

Relevant Investigation

US and CT of abdomen are diagnostic.

Treatment

- Emergency laparotomy is needed.
  - Twisted omentum needs to be excised with release of adhesions
  - Associated hernia needs repair.

COLICS

Incidence and Etiology

Colic is defined as a sudden squeezing or griping pain lasting for about 3 to 5 minutes with pain free intervals.

- Nausea, vomiting and retching are common accompaniments
- The cause of a colic is partial obstruction of a tubular structure due to varied causes.

Clinical Features (Table 21.12)

Table 21.12: Clinical features of various colics

Clinical feature	*Colics*				
	Biliary colic	*Renal colic*	*Ureteric colic*	*Intestinal colic*	*Appendicular colic*
Incidence	Fat, fertile, flatulent, female of fifty	Young age	Younger age	Any age	Any age
Etiology	Gallstones	Renal calculus	Ureteric calculus, papilloma, clot, stricture	Parasitic infestations (younger age), strictures (middle age), malignancy (old age)	Fecoliths, Worms
Nature of pain	Right hypochondrial pain, referred to right scapula or shoulder	Severe pain in the lumbar region radiating towards umbilicus	Severe colicky pain radiating from loin to groin, testis or thigh	Colicky pain in the umbilical region	Colicky pain in the right iliac fossa
Associated symptom	Dyspepsia	Urinary symptoms	Urinary symptoms	Constipation or diarrhea	Repeated attacks of dull pain in right iliac fossa

Symptoms

- Severe griping pain with pain free intervals
- Location of pain is definite (Table 21.13)
- Radiation (Fig. 21.54) may be present:
  - Radiation to right scapula or shoulder is common with biliary colic
  - Radiation to external genitalia and groins (e.g. ureteric colic)
- Vomiting is usually present

Table 21.13: Location of pain related to regions of abdomen

Right hypochondrium Biliary colic Right renal colic		*Left hypochondrium* Left renal colic
Right lumbar region Right renal colic	*Umbilical region* Intestinal colic	*Left lumbar region* Left renal colic
Right iliac fossa Appendicular colic Right ureteric colic	*Hypogastrium* Uterine colic Urethral colic	*Left iliac fossa* Left ureteric colic

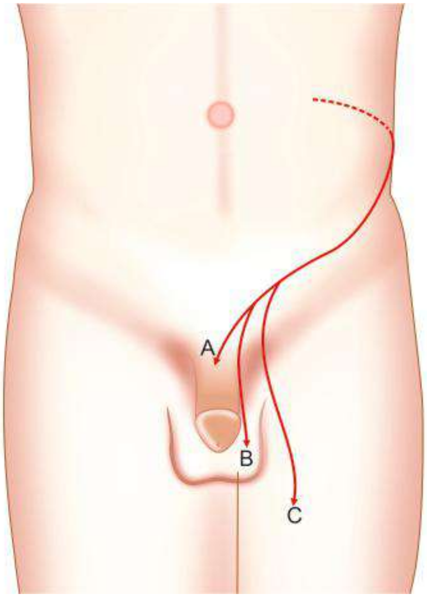

FIG. 21.54: Ureteric colic: (A) Base of penis; (B) Genitalia; (C) Thigh

- Fever may be present if infection is associated
- Jaundice may be present if cholangitis is associated
- Loose stools may be present if colitis is associated
- History of similar episodes in the past
- Family history may present regarding biliary and urinary calculi.

Signs

- Tenderness in the region of pathology
- Mass may be felt proximal to the level obstruction:
  - Right upper quadrant (e.g. distended gallbladder)
  - Right lumbar (e.g. obstructed and enlarged kidney as in PUJ obstruction)
  - Umbilical (e.g. distended bowel)
  - Hypogastric (e.g. distended uterus in dysmenorrhea).

Differential Diagnosis by Clinical History and Examination

- Biliary colic, jaundice, palpable gallbladder (e.g. cholangitis with mucocele, gallbladder choledochal cyst)
- Renal colic, urinary symptoms, palpable kidney (e.g. obstruction at PUJ)
- Ureteric colic, urinary symptoms, tenderness in iliac fossa (e.g. obstruction of ureter)
- Intestinal colic, exaggerated bowel sounds (e.g. obstruction of small bowel)
- Appendicular colic, tenderness right iliac fossa (e.g. obstructive appendicitis)
- Uterine colic, dysmenorrhea, palpable uterus (e.g. congestive dysmenorrhea).

Relevant Investigations

Hematology

- Leukocytosis in infective pathologies (e.g. acute cholecystitis)
- Raised ESR in all infective pathologies.

Radiology

- *Plain X-ray abdomen:* Gas filled loops of bowel (e.g. acute intussusceptions, acute intestinal obstruction)
- *Ultrasonography:* Radiopaque shadows in the abdomen (e.g. renal stones, ureteric stones).

Treatment Plan

- Medical management will suffice in most instances
- Obstructive pathologies may require surgery or removal to relieve the cause of obstruction
- Repeated attacks of colic will require evaluation and management.

GASTROINTESTINAL HEMORRHAGE

Definitions

- Hematemesis—vomiting blood, which may be fresh or partly altered.
- Melena—passing altered blood (meaning digested blood) per rectum.
- Hematochezia—passing blood per rectum bright in color
- Occult bleeding—bleeding not visible to naked eye, but microscopic.

Types of Gastrointestinal Hemorrhages

The blood loss due to gastrointestinal hemorrhage is divided into three types:

1. Mild—less than 500 ml.
2. Moderate—500 to 1500 ml.
3. Severe—more than 1500 ml.

- Gastrointestinal hemorrhages, when they are in large quantities, called major hemorrhages present either as vomiting of blood or passage of blood per rectum, but usually they coexist
- Many times, melena is the sole clinical presentation of upper gastrointestinal bleed.

While evaluating a GI hemorrhage, the following pathologies are to be kept in mind (Fig. 21.55)

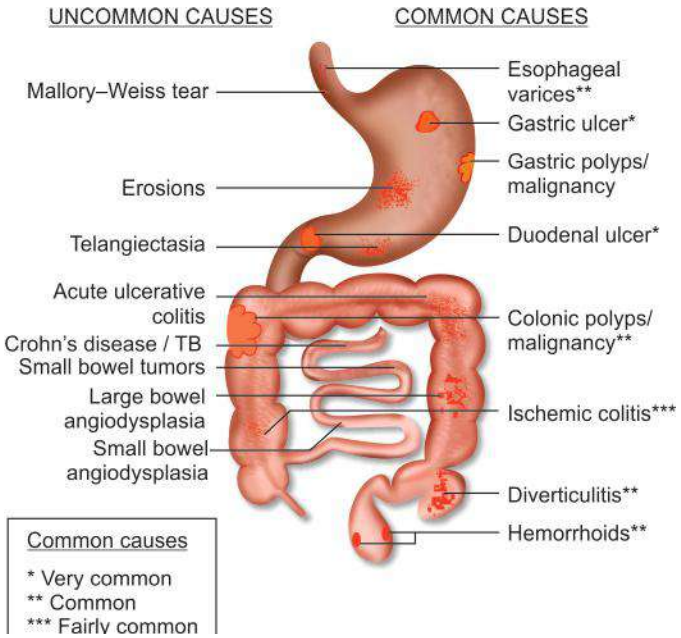
UNCOMMON CAUSES
COMMON CAUSES
Mallory–Weiss tear
Esophageal varices**
Gastric ulcer*
Gastric polyps/ malignancy
Erosions
Telangiectasia
Duodenal ulcer*
Acute ulcerative colitis
Crohn's disease / TB
Small bowel tumors
Colonic polyps/ malignancy**
Large bowel angiodysplasia
Ischemic colitis***
Small bowel angiodysplasia
Diverticulitis**
Hemorrhoids**
Common causes
* Very common
** Common
*** Fairly common

FIG. 21.55: Causes of gastrointestinal bleeding

Hematemesis and Melena

Esophagus

- Reflux esophagitis
- Esophageal varices
- Mallory–Weiss tears
- Corrosive poisoning.

Stomach and Duodenum

- Acid peptic disease
- Gastric polyps
- Gastric lymphoma
- Carcinoma stomach
- Corrosive poisoning.

Melena

Small Bowel

- Angiodysplasia
- Diverticulitis
- Radiation enteritis
- Infections and inflammations
- Ischemic disease
- Intussusception
- Richter's hernia
- Benign tumors
- Malignant tumors.

Hematochezia

Large Bowel

- Angiodysplasia
- Diverticulitis
- Radiation colitis
- Infections and inflammations

- Ischemic disease
- Inflammatory bowel disease
- Benign polyps
- Malignant tumors.

Rectum and Anus

- Polyps
- Malignant tumors
- Hemorrhoids, fissures.

Lower gastrointestinal hemorrhages do not present with hematemesis.

Eliciting History

Hematemesis

1. *Nature of bleed*
    - Frank blood (e.g. esophageal varices)
    - Altered blood (e.g. peptic ulcer, gastric malignancy).
2. *Duration of hematemesis*
    - Short duration (e.g. acute hyperacidity)
    - Recurrent attacks (e.g. chronic duodenal ulcer, gastric malignancy).
3. *Association of rectal bleed*
    - Melena (e.g. upper GI bleed)
    - Hematochezia (e.g. lower GI bleed).
4. *Association of abdominal pain*
    - Absence of abdominal pain (e.g. esophageal varices)
    - Presence of abdominal pain (e.g. chronic duodenal ulcer, gastric malignancy).
5. *Association of fever:* Fever – usually low grade (e.g. GI malignancy).
6. *Association of jaundice*: Jaundice may be present (e.g. periampullary carcinoma, hepatic failure).

Hematochezia

1. *Duration of bleed*
    - Short duration (e.g. hemorrhoids, inflammatory bowel disease, colonic malignancy)
    - Recurrent attacks and long duration (e.g. hemorrhoids, inflammatory bowel disease).

2. *Association of abdominal pain*
   - Absence of abdominal pain (e.g. hemorrhoids, malignancy)
   - Presence of abdominal pain (e.g. inflammatory bowel disease).
3. *Association of fever*: Fever (e.g. inflammatory bowel disease)
4. *Association of jaundice*

   Jaundice may be present (e.g. portal hypertension, metastatic colonic malignancy)
5. *Association of bowel disturbances*
   - Constipation (e.g. colorectal malignancies)
   - Diarrhea (e.g. inflammatory bowel disease).

Past History

- History of pain (e.g. chronic duodenal ulcer, diverticulitis)
- History of bleeds (e.g. chronic duodenal ulcer, diverticulitis, hemorrhoids)
- Previous surgery (e.g. recurrent malignancy of GIT).

Family History

- Familial polyposis
- Colonic malignancy.

Clinical Examination

General

- Vital signs (e.g. shock)
- Breath for fetor (e.g. hepatic failure)
- Conjunctiva for anemia (e.g. GI malignancy), jaundice (e.g. hepatic failure)
- Tongue for anemia (e.g. GI malignancy)
- Neck for lymphadenopathy (e.g. GI malignancy)
- Hands for clubbing, palmar erythema, liver flap, etc. (e.g. liver failure).

> - Mild hemorrhages are rarely associated with systemic signs. When the blood loss approaches about 40 percent of the blood volume, shock ensues
> - The volume of blood loss either by vomiting or through rectum, is not a very reliable measure, as large amounts stay in the bowel.

Abdomen

- **Inspection**

  Distension
  - Generalized (e.g. ascites of liver failure, malignant ascites)
  - Right upper abdominal (e.g. hepatomegaly)
  - Epigastric (e.g. left lobar hepatomegaly, carcinoma stomach)
  - Left upper abdominal (e.g. splenomegaly).

  Scars, swellings and sinuses
- **Palpation**

  Tenderness
  - Right upper quadrant (e.g. hepatic failure)
  - Epigastric (e.g. chronic duodenal ulcer)
  - Left upper quadrant (e.g. chronic duodenal ulcer).

  Lump
  - Right upper quadrant (e.g. hepatomegaly, distended gallbladder)
  - Epigastric (e.g. carcinoma stomach, left lobar hepatomegaly)
  - Left upper quadrant (e.g. carcinoma stomach, splenomegaly).
- **Percussion**

  Percuss the liver for
  - Its enlargement (e.g. chronic hepatitis, metastatic liver).
- **Auscultation**
  - Bowel sounds are usually normal
  - Exaggerated bowel sounds may indicate obstruction of small bowel (e.g. intestinal colic).

Differential Diagnosis by Clinical History and Examination

- Hematemesis and melena (e.g. pathologies of upper GI tract proximal to ligament of Treitz)
  - With upper abdominal pain, epigastric tenderness (e.g. bleeding gastric/duodenal ulcer)
  - Without abdominal pain and
    - No clinical findings (e.g. esophageal varices, vascular malformations, upper GI malignancy)
    - Abdominal lump (e.g. upper GI malignancy).
  - Without abdominal pain and
    - Increasing jaundice, abdominal lump, + / – hepatomegaly (e.g. upper GI malignancy + / – liver metastases)
    - Fluctuating jaundice, palpable GB, + / – hepatomegaly (e.g. periampullary carcinoma + / – liver metastases).

- Melena (e.g. pathologies of lower GI tract distal to ligament of Treitz)
  - With abdominal pain
    - Epigastric tenderness (e.g. bleeding duodenal/gastric ulcer)
    - Umbilical tenderness (e.g. inflammatory bowel disease)
    - Flank tenderness (e.g. ulcerative colitis)
    - Palpable lump (e.g. proximal colonic malignancy).
  - Without abdominal pain
    - No clinical findings (e.g. esophageal varices, vascular malformations)
    - Abdominal lump (e.g. upper GI malignancy).
- Hematochezia (e.g. pathologies of lower GI tract in the colorectum)
  - With pain during defecation, positive inspection findings (e.g. acute fissure in ano)
  - Without perianal pain
    - Positive proctoscopy findings (e.g. hemorrhoids, rectal pathologies)
    - Negative proctoscopy findings (e.g. colonic pathologies)
  - With abdominal pain + / – fever, local tenderness (e.g. inflammatory bowel disease, diverticulitis)
  - With colicky pain, abdominal lump (e.g. obstructing colonic malignancy)
  - With jaundice + / – abdominal pain, + /– abdominal lump, + / – hepatomegaly (e.g. colonic malignancy with liver metastases).

Whatever be the external visible loss, signs of hypovolemia should be watched for.

Relevant Investigations

Hematology

- Hemoglobin and PCV for anemia
- Total and differential leukocyte count (e.g. tuberculosis, ulcerative colitis)
- Thrombocytopenia (e.g. hypersplenism)
- ESR may be raised in infections and malignancies
- Coagulation profile (e.g. bleeding disorders).

Coagulation profile should be assessed in all cases of GI bleed, as bleed by itself the sole presentation of coagulation disorders.

Liver Function Tests

- Serum bilirubin (e.g. raised levels indicate jaundice)
- Serum transaminases (e.g. raised levels indicate liver cell disease)
- Serum alkaline phosphatase (e.g. raised levels may indicate biliary obstruction due to hepatic metastases).

- Serum proteins (e.g. reduced levels indicate undernutrition)
- Prothrombin time is prolonged in liver diseases and needs to be corrected by administration of vitamin K, in the management of gastrointestinal bleeds.

Stool Examination

Examination of stool for occult blood is required in cases of occult bleeding from GIT.

> Occult bleeding from GIT is common, but the bleed should be atleast 10 ml to identify by examination.

Radiology

- Chest X-ray (e.g. aspiration pneumonitis, mediastinal widening and hilar lymphadenopathy of esophageal malignancy)
- Contrast studies of bowel
  - Barium swallow (e.g. esophageal malignancy)
  - Barium meal (e.g. gastric malignancy, periampullary carcinoma)
  - Barium enema (e.g. colonic malignancy)
- CT/MRI scan
  - Chest (e.g. esophageal malignancy, and paraesophageal pathology like lymph nodes, pulmonary secondaries)
  - Abdomen (e.g. cirrhosis of liver, malignant deposits of liver, lymph node metastases of GI malignancies, intestinal tuberculosis)
- Ultrasonography of abdomen (e.g. malignant deposits, dilated portal vein and biliary radicals)
- Magnetic resonance cholangiopancreatography (MRCP) (e.g. periampullary carcinoma).

Endoscopy

- Upper gastrointestinal endoscopy (e.g. gastroesophageal reflux disease, esophageal tears, gastric malignancies, periampullary carcinoma)
- Lower gastrointestinal endoscopy (e.g. polyps, tumors of rectum and colon)

> - Endoscopy has the advantage of obtaining tissue for histopathology
> - Endoscopy may be used as a therapeutic tool at the same sitting (e.g. sclerotherapy in bleeding esophageal varices and endoclipping of bleeding vessel)
> - Endoscopy should be performed at the earliest opportunity and after adequate resuscitation.

Radioisotope Studies

Radioisotope scanning using the patient's own labeled red blood cells can be useful in small bowel bleeds, especially those from angiodysplasia.

Selective Arteriography

Selective arteriography (superior mesenteric arteriography) may be useful in determining the small bowel bleeds, particularly those from angiodysplasia.

Capsule Endoscopy

Swallowing a small capsule with a video camera, and recording the images of the lumen of small bowel gives tremendous information of bleeding from small bowel. Histopathology cannot be obtained by capsule endoscopy.

Histopathology

Biopsy through endoscopy (upper and lower) is confirmative.

Treatment

Medical Management

- Large bore vascular access and correction of hypovolemia and hematocrit
- Correction of coagulopathy
- Emergency endoscopy
  - For nonvariceal bleeding
    - IV PPI (80 mg 6 hrly)
    - Nasogastric suction and gastric lavage
  - For variceal bleeding
    - Insertion of Sengstaken-Blakemore tube for compression of varices
    - Administration of vasopressin (upto 20 units SC or slow IV).
    - Variceal injection or banding
    - Transhepatic intravenous portosystemic shunt (TIPSS) for uncontrolled bleeding.

Surgery is needed for persistent uncontrollable GI bleeding.

Patient requiring more than six units of blood for resuscitation for acute bleed, will require surgical management.

Surgical Treatment

Surgery is indicated in recurrent esophageal variceal bleeding, complicated inflammatory bowel disease, polyps, benign and malignant tumors of GIT and hemorrhoids.

- Rapid changes occur in the tissues distal to the occlusion and the limb survival depends on the presence of collaterals
- Skeletal muscle and nerve tissue are the most sensitive to hypoxia and suffer damage early.

Complications: Compartment syndrome, acute renal failure (acidosis), muscle contracture (Volkmann's ischemic contracture).

Symptoms

- Acute ischemia produces classical symptoms (6 Ps)
  - *Pain:* Usually excruciating, sudden in onset and continuous
  - *Pallor:* Affected limb appears pale due to lack of blood
  - *Pulselessness:* Pulses distal to the block are absent
  - *Paresthesia:* Due to nerve damage
  - *Paralysis:* Sensory motor deficit due to ischemic nerve damage secondary to severe ischemia (a late sign)
  - *Poikilothermia:* The affected limb appears cold.

Signs

- Affected muscle gets swollen, tense and is very tender, due to persistent ischemia
- On palpation, muscle gives a characteristic rubbery feel (can be mistaken for DVT)
- Later skin changes like mottling (Fig. 23.1) and blisters start and ultimately the limb becomes gangrenous (Fig. 23.2).

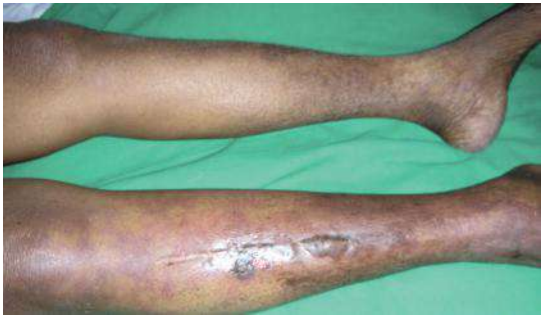

FIG. 23.1: Mottling of ischemic limb (*Courtesy*: Dr N Sekar)

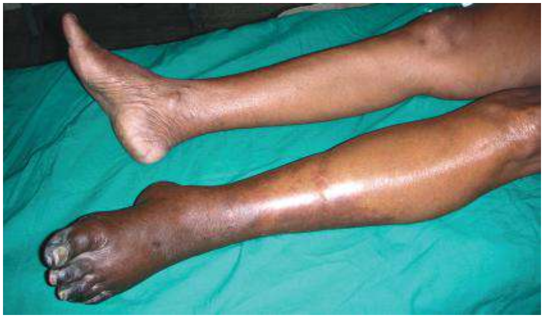

FIG. 23.2: Gangrene of acute ischemia of left lower limb (*Courtesy*: Dr N Sekar)

Relevant Investigations

All investigations should be done without delay:
- Duplex scan is done to confirm diagnosis in early cases
- Angiogram for treatment planning
- Tests for thrombophilia
- Hyperkalemia when renal shut down ensues.

Treatment

- Acute arterial occlusion is a surgical emergency. Usually irreversible damage occurs by 4 to 6 hours. Hence, revascularization should be done within 6 to 8 hours for limb salvage (Golden hour).
    - Intravenous heparin should be administered as soon as the diagnosis of acute ischemia of the limb is made. Best results are seen in those who undergo revascularization within 6 to 8 hours after the onset of ischemia
    - Embolism is best managed by embolectomy done through the brachial artery at the elbow or the femoral artery at the groin
    - Acute arterial thrombosis can be managed by surgical bypass or by catheter directed thrombolytic therapy. Frequently both modalities may be required to achieve revascularization
    - Fasciotomy will be required to avoid muscle damage in compartment syndrome caused by delayed revascularization
    - Early amputation for limbs with irreversible ischemia.

Note: Those who present late should be taken up for surgery without imaging and intraoperative angiogram should be done to reduce time delay.

ACUTE INTESTINAL ISCHEMIA

Incidence and Etiology

Intestinal ischemia occurs usually due to occlusion of mesenteric artery caused by:
- Thrombosis on an atheromatous plaque (e.g. origin of superior mesenteric artery)
- Embolus following atrial fibrillation/myocardial infarction/detached atheromatous plaque.

Clinical Presentation

- Severe acute abdominal pain with copious vomiting
- Very fast deterioration of health

- Hematemesis and melena also occur in some cases
- The clinical signs are disproportionate to the symptoms, and rarely an area of tenderness may be felt near the infarcted bowel.

Relevant Investigations

- ECG, plain X-ray and US abdomen are useful
- Selective angiography is informative.

- Systemic inflammatory response, oliguria, persistent acidosis, raised serum amylase and bloody diarrhea may all occur but none is specific.

Treatment

- Conservative management to stabilize the patient
- Laparotomy and resection of nonviable bowel will be necessary, if the patient is stable
- Balloon angioplasty or bypass grafts may be feasible in select cases.

Note: Outcome is guarded in most cases.

- Intestinal ischemia should be considered as a possible cause of deterioration, particularly in the elderly with pre-existing vascular disease or vasculitis.

LEAKING OR DISSECTING AORTIC ANEURYSM

Incidence and Etiology

- Aneurysm of aorta (> 4 cm diameter – normal 1.5 to 2.5 cm) is the disease of elderly, and rupture is the deadly complication, with a mortality of more than 80 percent
- The rupture can occur into the retroperitoneum or into the peritoneal cavity.

Clinical Presentation

- Acute severe upper abdominal pain or in the lower chest
- Radiation to the back mimics acute pancreatitis
- Aneurysm may be felt per abdomen with guarding and rigidity
- Lower limb pulses are feeble or absent
- Dissecting aneurysms present with retrosternal pain, as the pathology starts at the aortic arch.

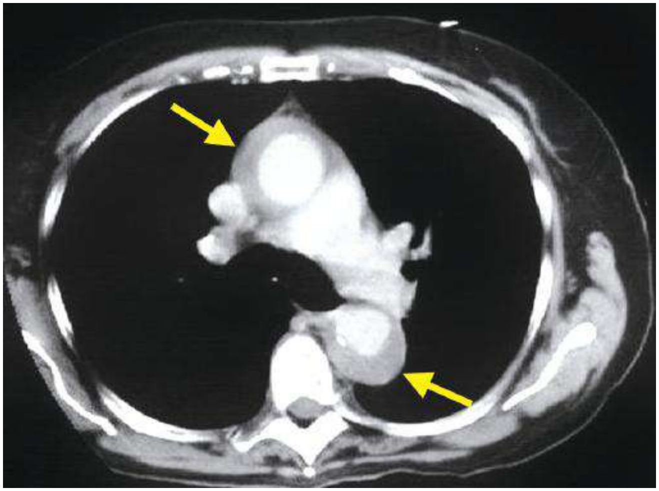

FIG. 23.3: CT chest—Dissecting aneurysm of thoracic aorta

Relevant Investigation

US and CT of the abdomen are conclusive (Fig. 23.3).

Treatment

Emergency surgical intervention is necessary, along with correction of shock.

- Only a third with rupture of aneurysm live to reach the hospital.

24

Urology

ACUTE RETENTION OF URINE

Definition

Defined as inability to micturate even when the bladder is full.

Incidence and Etiology

- This occurs when the sphincter is unable to relax or with proximal urethral obstruction like in prostatic enlargement
- Acute retention is of short duration and the bladder distension is painful
  The causes of acute retention of urine are given in Table 24.1.

Symptoms

- Severe lower abdominal pain
- Inability to void urine
- Symptoms relating to underlying cause (e.g. LUTS).

Signs

- Palpable distended tender urinary bladder
- Rectal examination: Anal sphincter spasm, enlarged prostate.

Table 24.1: Causes of acute retention of urine

Males	*Females*	*Both sexes*
Prostatic enlargement	Multiple sclerosis	Clot retention
Urethral stricture	Urethral stenosis	Calculus at bladder neck or urethra
Tight Phimosis / Meatal stenosis	Retroverted gravid uterus	Rupture of urethra
Acute urethritis	Cervical fibroid	Spinal anesthesia
Postoperative	Bladder neck hypertrophy	Neurogenic
		Fecal impaction
		Anal pain (post hemorrhoidectomy)
		Drugs (antihistamines, antihypertensives, anticholinergics, antidepressants)

Relevant Investigations

- Hemogram
- Urea/creatinine
- PSA levels may be high
- US may confirm the cause of obstruction.

Treatment

- Immediate urethral catheterization
- Urine examination for sugar and infective organisms
- Treatment of the cause.

HEMATURIA

Definition

Passing of blood in the urine is called hematuria.

- Frank hematuria: Presence of frank blood or blood clots in the urine
- Microscopic hematuria: Presence of blood discovered by microscopy or dipstick.

Incidence and Etiology

- Whatever the cause, hematuria is usually episodic:
  - Hematuria (painless) (e.g. tumors of urological system)
  - Hematuria with dysuria (e.g. urological infections)
  - The causes of hematuria are given in Table 24.2.

Table 24.2: Causes of hematuria

Etiology	*Diseases*				
	Kidneys	*Ureters*	*Bladder*	*Prostate*	*Urethra*
Congenital	Polycystic kidney				
	AV malformation				
Traumatic	Trauma		Trauma		Trauma
Inflammatory	Glomerulonephritis		Cystitis	Prostatitis	Acute urethritis
	Tuberculosis		Schistosomiasis		
Neoplastic	Malignancy	Urothelial tumors	Malignancy	Malignancy	Malignancy
Others	Calculi	Calculi	Calculi	Benign enlargement	Calculi
Nonurological	Coagulation and hematological disorders				

Note: Microscopic hematuria may represent a significant lesion in the urinary tract and should be taken seriously, though in about 5 percent of the cases, no cause can be found.

Symptoms

- Passing dark colored/blood stained urine
- Pain may or may not be present.

Signs

- Clinical signs may vary (palpable kidney in large tumors, palpable bladder in bladder obstruction)
- Anemia (in chronic hematuria or large hematuria)
- Enlarged prostate.

Relevant Investigations

- Hematocrit may be low
- Renal function may be impaired
- Coagulation profile may be altered
- Urine examination for infection (routine and C/S)
- Ultrasonography
- Intravenous urography (IVU)
- Cystoscopy
- CT/MRI may be needed while evaluating malignancies.

Treatment

- Primary treatment requires immediate urethral catheterization
- When catheterization is not possible, suprapubic cystostomy is needed
- Definitive treatment depends on the underlying cause.

RENAL COLIC

Caused by distension of renal capsule and pelvis or stretch of the peritoneum by large renal swellings:

- Nature of pain
  - Pain localized to renal angle (posterior space between the 12th rib and the sacrospinalis muscle) (Fig. 24.1) (e.g. acute pyelonephritis, renal tuberculosis)—described by putting the hand on the waist with his fingers on the renal angle and thumb pointing the umbilicus (Fig. 24.2)
  - Pain radiating to umbilicus (e.g. renal tuberculosis)
- Intensity of pain
  - Severe (e.g. acute pyelonephritis renal stones)
  - Dull and persistent (e.g. polycystic disease and malignancy)
  - Acute obstruction of the upper ureter can give a severe colicky pain in the loin called “Renal colic”.

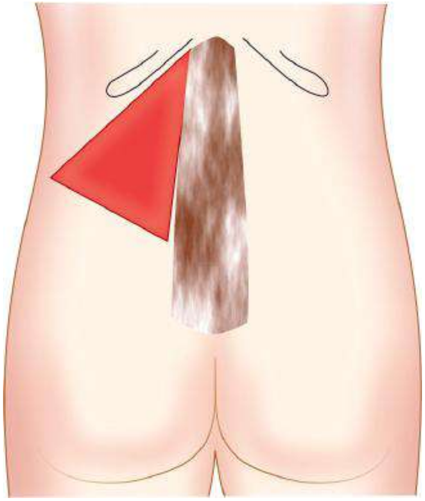

FIG. 24.1: Renal angle

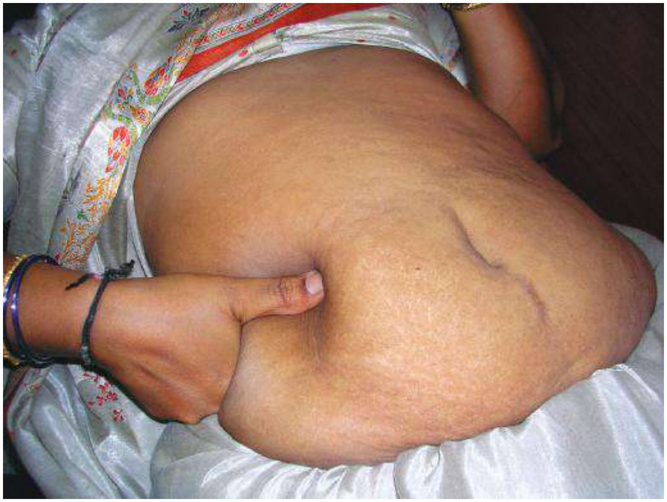

FIG. 24.2: Description of renal pain

Relevant Investigations

- Hematocrit may be low
- Renal function may be impaired
- Urine examination for infection (routine and C/S)
- Ultrasonography will show stones or proximal dilatation
- IVU will show stones, or filling defects (stones, tumors) or proximal dilatations
- Cystoscopy is routine
- CT/MRI may be needed while evaluating malignancies.

Treatment

- Emergency treatment requires antispasmodics to relieve pain, antiemetics for vomiting
- Definitive treatment depends on the underlying cause.

URETERIC COLIC

Caused by obstruction of the ureter commonly by a stone or a clot, and may correspond to the level of obstruction:

- Nature of pain
    - When the stone is lodged in the upper ureter, the pain radiates to the testicle (T11-12)
    - When the stone is lodged in the mid ureter, the pain may be in the McBurney's point on the right side and simulate diverticulitis on the left side (T12-L1)
    - When the stone is lodged in the distal ureter, the pain resembles vesicular pain or may radiate to the genitalia or inner side of thigh (L1-L2)
- Intensity of pain
    - Dull pain in the side of abdomen (e.g. chronic obstruction of ureter with calculi)
    - Severe colicky pain—(loin pain radiating to the groin, genitalia or inner thigh—genitofemoral nerve) in acute obstructions of ureter called "Ureteric colic" (Fig. 24.3).

Symptom

Severe pain as described above.

Relevant Investigations

- Hematocrit may be low
- Renal function may be impaired

27

Gynecology

ACUTE TORSION OF OVARIAN CYST

Incidence and Etiology

It can occur at any age, may follow major physical exertion, which can initiate the torsion.

Pathogenesis

Due to the large size and vulnerability to rotate over a narrow pedicle, torsion occurs.

Symptom

Severe pain in the lower abdomen, with signs of shock.

Signs

- Guarding and rigidity may be present.
- A tender lump may be palpable, whose lower border cannot be made out.

Relevant Investigations

Ultrasonography (US) and Computed tomography (CT) or Magnetic resonance imaging (MRI) of abdomen (Fig. 27.1) are diagnostic.

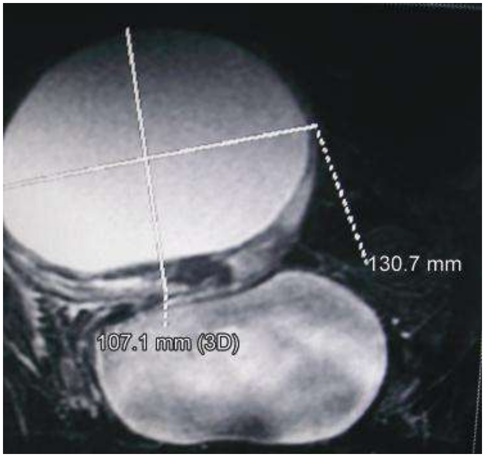

FIG. 27.1: MRI—Torsion of ovarian cyst

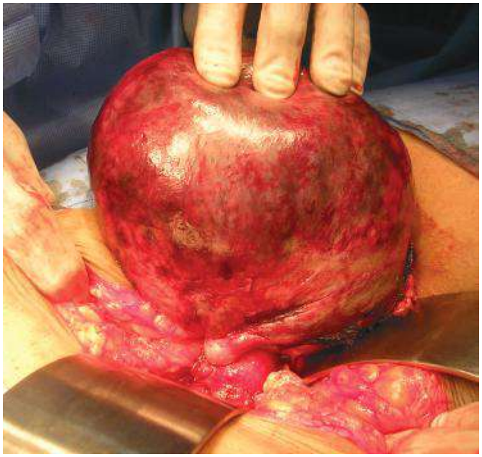

FIG. 27.2: Torsion of ovarian cyst

Treatment

Emergency laparotomy (Fig. 27.2) is needed. The cyst should be removed.

ACUTE SALPINGITIS

Incidence and Etiology

- A disease of the sexually active female, sometimes after sexually transmitted diseases, can confuse with acute appendicitis
- Usually associated with menstrual period, abortion or puerperium
- Common infective agents are *N. gonorrhoeae*, coliforms and chlamydiae.

Pathogenesis

Infective organism gains entry through the vagina to infect the salpinx.

Symptoms

Suprapubic pain, with or without white discharge per vagina (leukorrhea).

Signs

- Localized hypogastric tenderness
- Vaginal examination may reveal tender cervix and a bulky uterus.

Differential Diagnosis

Acute appendicitis is to be thought of in right sided salpingitis.

Relevant Investigations

Ultrasonography (US) is useful in diagnosing mass of the right ovary and the bulky uterus.

Treatment

Treatment with antibiotics.

RUPTURE OF LUTEIN CYST

Incidence and Etiology

A common condition in young females.

Clinical Presentation

- Acute abdominal pain associated with vomiting (right sided pathology may be confused with acute appendicitis)
- On examination:
  - A firm mass may be felt adjacent to the midline or in one of the iliac fossae
  - Guarding and rigidity may be present
- Vaginal examination may reveal a tender mass.

Relevant Investigations

Ultrasonography (US) of the abdomen is diagnostic.

Treatment

Emergency surgical intervention is mandatory.

28

Pediatrics

ACUTE INTUSSUSCEPTION

Incidence and Etiology

- Two per 1000 infants are affected with male preponderance, commonly affecting the age group of 2 months to 2 years
- Commonly, it is secondary to an enlarged Peyer's patch due to viral or bacterial infections (Fig. 21.39)
- The other less common causes are:
  - Meckel's diverticulum
  - Duplication cyst in the bowel wall
  - Polyp.

Pathogenesis

- Intussusception is the invagination of a segment of bowel into the distal adjacent loop (proximal into the distal)
- When the mesentery is drawn between the loops, it may result in vascular compromise, which may lead to strangulation, gangrene and perforation.

Symptoms

- In children, there may a history of preceding gastroenteritis following a change in diet (weaning from milk to solid food)

- Severe acute colicky pain, with abdominal distension
- Passing of frequent semisolid stools with bright red blood may be predominant (red recurrant jelly).

Signs

- During the attacks of pain, a sausage shaped mass may be felt, which appears during the time of colic and disappears after the colic disappears. The right iliac fossa is empty—Sign de Dance
- Rectal examination may reveal bloodstain on the examining finger (red-currant jelly)
- Colorectal intussusception may be felt by the examining finger on rectal examination, or it may even present through anus, resembling a rectal prolapse.

Relevant Investigations

- Plain X-ray abdomen: Soft tissue shadow in the region of transverse colon with empty distal colon. Multiple air fluid levels may be seen when obstruction predominates
- Barium enema may show a filling defect called pincer shaped filling defect (caused by the intussusceptum with the intussuscipient)
- Colonoscopy can identify, colonic intussusceptions
- US and CT (Fig. 28.1) will reveal the intussuscepting mass (pseudokidney appearance).

Treatment

- Resuscitation
- Hydrostatic or pneumatic radiological reduction

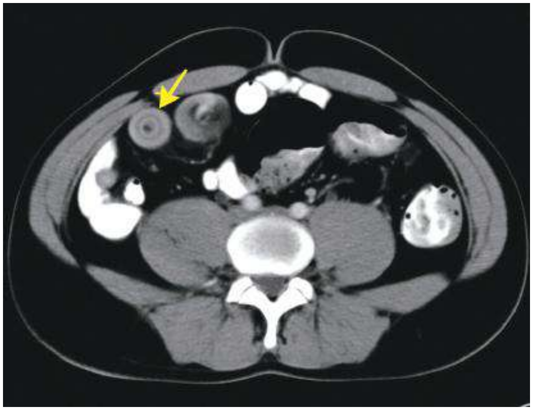

FIG. 28.1: CT—Intussusception

- Barium enema and colonoscopy, by themselves may reduce the colonic intussusception
- Laparotomy is required to reduce the small bowel intussusception, and treat the cause appropriately
- Bowel resections may be needed if the bowel segment is strangulated, and nonviable
- Perforation and peritonitis need appropriate treatment.

CONGENITAL PYLORIC STENOSIS

Incidence and Etiology

Four in 4000, with male predominance, most common in first born child.

Pathogenesis

Congenital inherited condition resulting in hypertrophy of the circular muscle fibers of pylorus.

Symptoms

- Nonbilious projectile vomiting
- Failure to thrive despite hunger.

Signs

- Visible gastric peristalsis
- Palpable mass in abdomen deep to right rectus in the transpyloric plane.

Relevant Investigations

- Serum electrolytes to be checked
- US is diagnostic.

Treatment

- Resuscitation
- Pyloromyotomy when the child is stable.

NECROTIZING ENTEROCOLITIS

Incidence and Etiology

Common in newborns.

Pathogenesis

Thought to be due to ischemia of large bowel wall with translocation of luminal bacteria resulting in systemic sepsis.

Symptoms

- Fever
- Bilious vomiting
- Bloody diarrhea
- Abdominal distension.

Signs

- Abdominal distension
- Absence of bowel sounds (may indicate perforation and peritonitis).

Relevant Investigations

- Leukocytosis
- Abdominal X-ray may show thickened dilated bowel wall containing intramural gas.

Treatment

- Fluid resuscitation
- Broad spectrum antibiotics
- Surgical resection of necrotic bowel may be needed.

TRACHEOESOPHAGEAL FISTULA

Incidence and Etiology

This fistulous communication can occur in newborn infants.

Pathogenesis

It is a developmental abnormality.

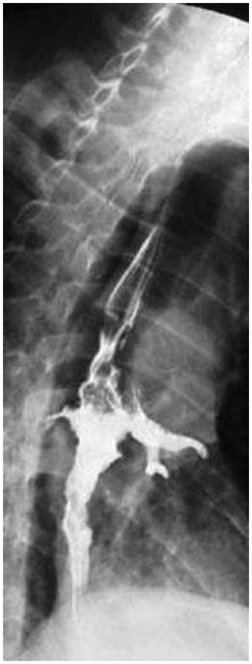

FIG. 28.2: Barium swallow—Tracheoesophageal fistula

Symptoms

Cough and fever - due to the entry of swallowed food into the respiratory tract.

Signs

Rales and rhonchi.

Relevant Investigations

Barium swallow (Fig. 28.2) allows the contrast to enter the trachea and bronchi, and is diagnostic.

Treatment

- Surgical closure of a congenital fistula is usually successful
- Malignant fistulae are difficult to handle and death becomes inevitable due to pulmonary infection.

29

Lymphatic System

ACUTE LYMPHANGITIS

Incidence and Etiology

In the tropics, the main causes are:

- Filariasis
- Tuberculosis.

Clinical Presentation

- Swelling of the limbs and genitalia associated with fever and rigor
- On examination, the limbs are swollen and erythematous with weeping eruptions (Fig. 29.1)
- Secondary infections can lead to localized patchy gangrene (Fig. 29.2).

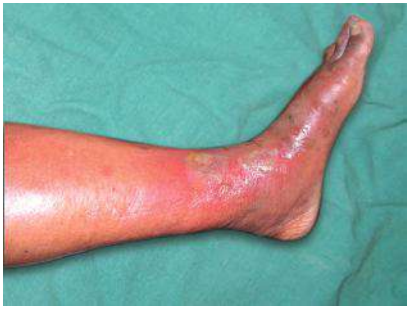

FIG. 29.1: Acute lymphangitis

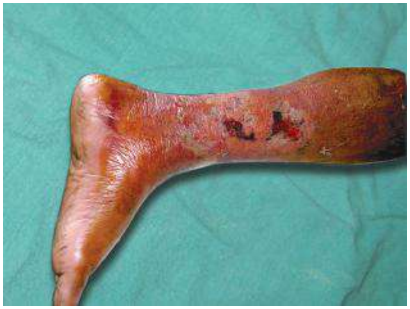

FIG. 29.2: Gangrenous patches in acute lymphangitis

Relevant Investigations

Peripheral smear for microfilaria.

Treatment

Antifilarial treatment and compression bandages.

ACUTE VIRAL AND BACTERIAL LYMPHADENITIS

Incidence and Etiology

- Caused by viruses (commonly rhinovirus), bacteria
- Bacterial infections are secondary to oral cavity infections
- Infections may result into abscesses
- Common in the neck.

Symptoms

- Painful swelling (commonly in the neck, rarely in the axilla and inguinal region)
- Fever may be present.

Signs

- Single or multiple tender soft to firm swellings with signs of inflammation on the skin (Figs 29.3 and 29.4)
- Examination of the drainage area is essential to assess the primary cause e.g. acute pharyngitis or dental infections causing secondary enlargements of neck glands.

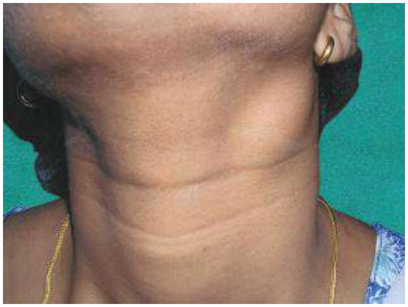

FIG. 29.3: Acute viral lymphadenitis of neck

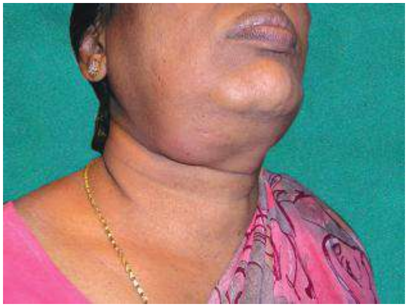

FIG. 29.4: Acute bacterial lymphadenitis of neck

Relevant Investigations

- No specific investigation is required in short lived cases
- Biopsy may be needed to differentiate persistent swellings from other causes
- Diagnostic aspiration is useful in abscesses.

Treatment

- No treatment is required for short lived cases (viral) as they resolve spontaneously
- Bacterial lymphadenitis requires broad spectrum antibiotics
- Incision and drainage for abscesses.

ACUTE FILARIAL LYMPHANGITIS AND LYMPHADENITIS

Incidence and Etiology

- Caused by thread like, parasitic filarial worms Wuchereria bancrofti or Brugia malayi
- Worms lodge in the lymphatic system, and live for 4 to 6 years producing millions of immature microfilariae that circulate in the blood
- Transmitted by mosquitoes *Culex fatigans* (develop and reach the infective stage in 7 to 21 days and migrate to biting mouth parts)
- Causes severe inflammation of the lymphatics and they settle down in the lymph nodes, commonly in the inguinal group.

Symptoms

- High-grade fever with rigor
- Swelling of the lower limbs and groin swellings.

Signs

- Inguinal lymphadenopathy (Fig. 29.5)
- Presentation with swelling of scrotum (hydrocele) (Fig. 29.6) is common.

Relevant Investigation

Demonstration of microfilaria in the peripheral blood smear (Fig. 29.7) is diagnostic.

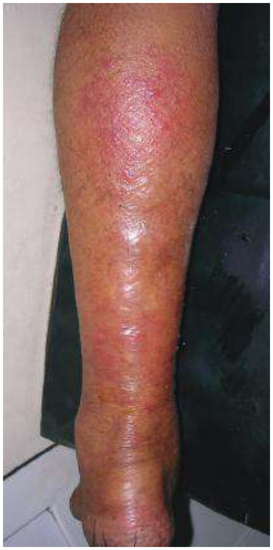

FIG. 30.2: Erysipelas

Investigations

Isolation of the organism in culture.

Treatment

Appropriate antibiotics cure the lesion.

FURUNCLE

Incidence and Etiology

- Infection of the hair follicle by *Staphylococcus aureus*
- Furuncle may be a source of systemic sepsis, in diabetics
- Cavernous sinus thrombosis is rare but a serious complication of furuncle on the face above the line drawn from the angle of the mouth to the tragus of the ear.

Symptom

A painful swelling at the hair root (Fig. 30.3).

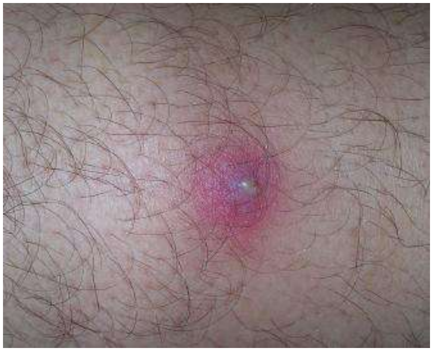

FIG. 30.3: Furuncle

Signs

- Indurated swelling containing pus in due course
- Draining lymph nodes may be involved
- Cellulitis is seen in immunocompromised individuals and diabetics.

Relevant Investigations

Culture of the pus and identifying the incriminating organism is necessary.

Treatment

- The abscess may burst spontaneously
- Some may need surgical drainage.

CELLULITIS

Incidence and Etiology

- Spreading inflammation of the subcutaneous and fascial tissues
- Commonly due to *Streptococcus pyogenes*
- Diabetics are mostly affected.

Pathogenesis

Organism gains entry through the broken skin due to trauma, and infection spreads along the subcutaneous tissue planes.

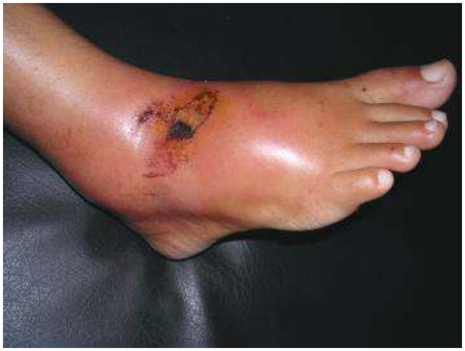

FIG. 30.4: Cellulitis foot

Symptoms

- Affected part appears grossly swollen, painful and red (Fig. 30.4)
- Constitutional symptoms like fever and toxemia are common.

Signs

- It is very tender
- The lymphatics may appear inflamed and appear as red streaks and lymph nodes may be enlarged and tender
- Abscesses may form in the subcutaneous plane and skin may undergo avascular necrosis and become gangrenous.

Relevant Investigations

Organisms should be isolated for culture from discharge if any.

Treatment

Appropriate antibiotics are necessary to control infection.

ACUTE PYOGENIC ABSCESS

Incidence and Etiology

- Abscess is defined as collection of pus
- Incomplete resolution of cellulitis and lymphadenitis are the causes
- When this occurs as a part of pyemia, they are multiple in number.

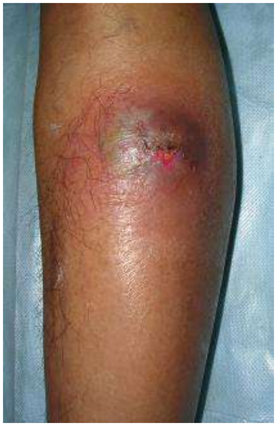

FIG. 30.5: Acute abscess leg

Symptoms

- A painful swelling (Fig. 30.5)
- Constitutional symptoms like fever, rigor and toxemia will be predominant.

Sign

Tender swelling, may show fluctuation, if the pus is liquid.

Relevant Investigations

- No special investigation is required, if it is solitary
- Isolation of organism in blood by culture if suspected to be part of pyemia
- Diagnostic aspiration may be helpful in localizing deep seated abscess.

Treatment

Incision and drainage of pus will be curative under cover of appropriate antibiotics.

CARBUNCLE

Incidence and Etiology

- An infective gangrene of subcutaneous tissue
- Caused by *Staphylococcus aureus*
- Commonly seen in diabetics.

Pathogenesis

- Thermal burns: Heat denatures cellular proteins by coagulation necrosis. The damage is directly related to the intensity of heat and the duration of contact of the incriminating agent. Usually the damage is partial thickness or full thickness
- Electrical burns: Electricity (high and low voltage) causes deep tissue destruction both at the point of entry and at the point of exit. Muscle tissue destruction is an integral part but it cannot be assessed accurately in the initial stages
- Radiation burns: Radiation causes full thickness dermal injury due to the deep penetration of ionizing radiation
- Chemical burns: Chemicals cause cell necrosis and the damage depends on the concentration of the chemical and the duration of contact with the skin. The damage ceases only when the agent is chemically expended, and majority of the times the damage is full thickness.

Clinical Features

The burn wounds are classified into four degrees based on the clinical features (Table 30.2). The clinical photographs are shown in Figures 30.7 to 30.17.

Table 30.2: Classification of burn wounds and their clinical features

Degree of burn	*Layer of skin involved*	*Clinical presentation*			*Tests*		
		Pain	*Color*	*Blisters*	*Touch*	*Pressure*	*Capillary filling*
First degree (superficial)	Epidermis	Severe	Erythema	Absent	Present	Present	Present
Second degree (partial thickness)	Superficial (superficial dermis)	Severe	Erythema/ pallor	Present	Present	Present	Present
	Deep (deep reticular dermis)	Moderate	Pallor	Usually absent	Absent	Present	Present
Third degree burn (full thickness)	Epidermis and dermis	Absent	Pallor	Absent	Absent	Absent	Absent
Fourth degree burns	Skin and subcutaneous structures	Absent	Pallor	Absent	Absent	Absent	Absent

Note: In any patient of burn injury, a combination of degrees occurs. This is important while treating a patient of burns, and repeated assessment and scrutiny of the wounds is necessary.

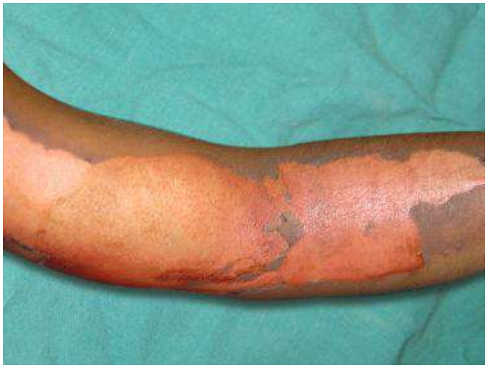

FIG. 30.7: Superficial burns

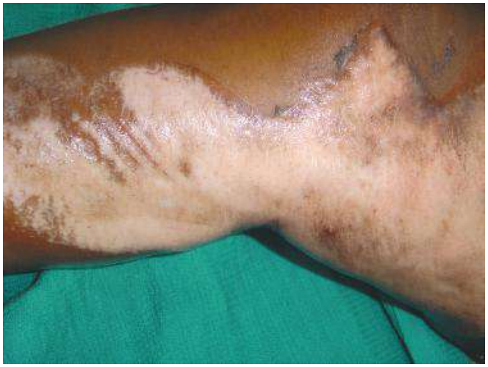

FIG. 30.8: Superficial partial thickness burns

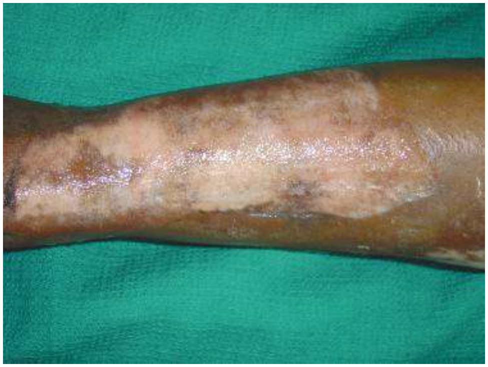

FIG. 30.9: Superficial partial thickness burns

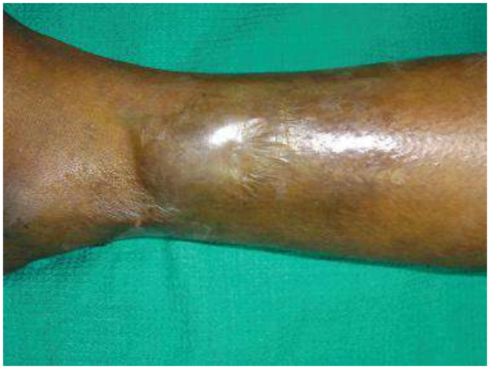

FIG. 30.10: Second degree burns with blister formation

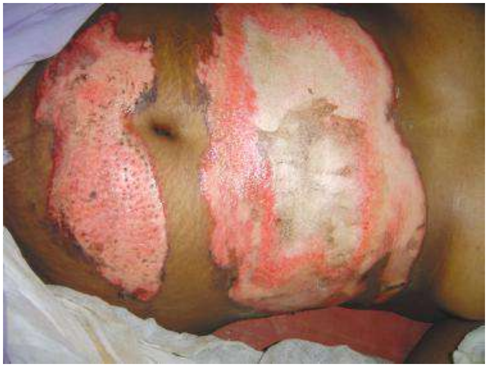

FIG. 30.11: Mixed second and third degree burns

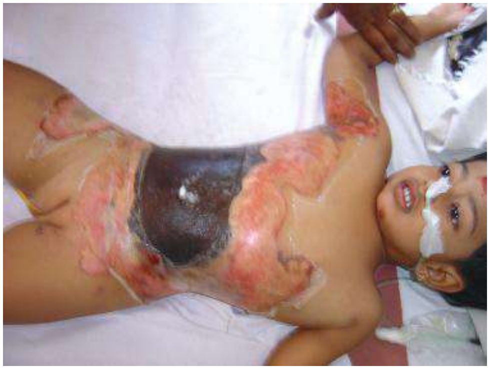

FIG. 30.12: Deep partial thickness burns (third degree)

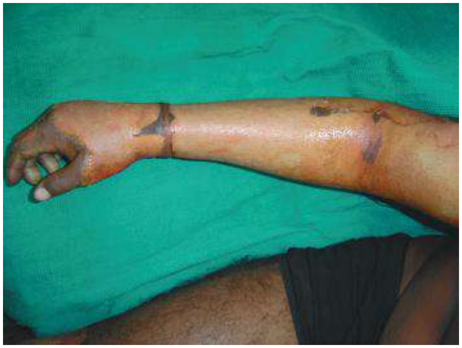

FIG. 30.13: Deep burns of the forearm

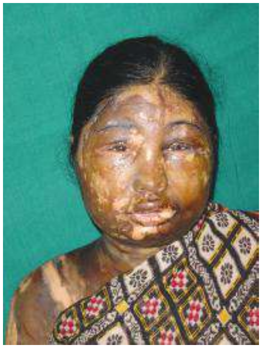

FIG. 30.14: Acid burns of the face

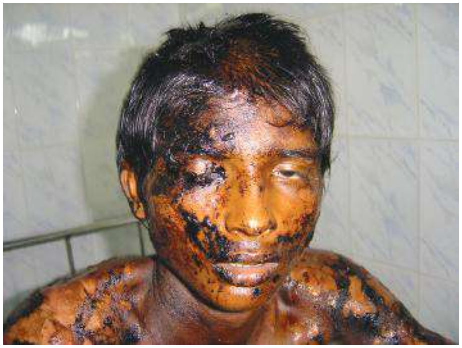

FIG. 30.15: Chemical (tar) burns—Thermal contact

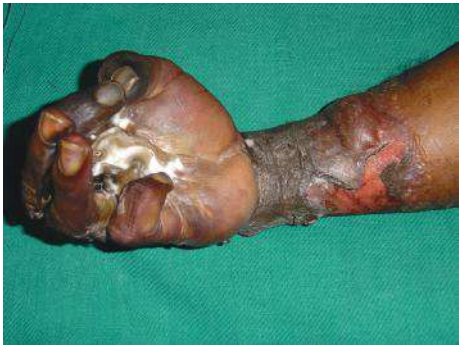

FIG. 30.16: Electrical burns

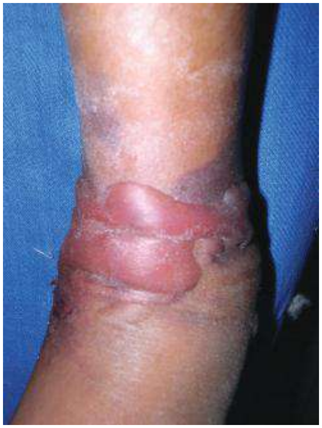

FIG. 30.17: Thermal burns caused by hyperthermic reaction of plaster of Paris

The depth of burns varies depending on the incriminating agent and its duration of contact with the skin (Table 30.3).

Table 30.3: Burns based on its etiology

Agent	*Exposure time*	*Depth of burns*	*Appearance*	*Pain*
Hot liquids	Short exposure	Superficial dermal	Wet, pink, blisters	Severe
	Long exposure	Deep dermal	Wet, red, dark	Minimal
Flame	Flash exposure	Partial thickness	Wet pink blisters	Severe
	Direct contact	Full thickness	Dry, white, waxy or brown, black leathery	Minimal
Chemicals	Direct contact	Full thickness	Light brown to light gray	Severe

Metabolic Effects of Burns

Burn injuries cause severe damage to the body structures and its metabolism. They are shown in Tables 30.4 and 30.5.

Table 30.4: Metabolic disturbances associated with burns (general effects)

Category	*Period*	*Clinical condition*	*Reason*	*Clinical presentation*
General effects	Immediate	Hypovolemic shock	Extracellular accumulation of water and excessive evaporation from burnt area	Tachycardia, tachypnea, hypotension, prerenal uremia, oliguria
		Hemoconcentration	Loss of water and plasma from burnt area	
		Dehydration	Excessive evaporation of water from burnt area	Tachycardia, tachypnea, hypotension, prerenal uremia oliguria
		Electrolyte imbalance	Retention of sodium and excessive excretion of potassium	Altered sensorium, uremia, oliguria
		Asphyxia and cyanosis	Due to burn of respiratory passages	Difficulty in breathing
	Late	Septicemia (multiorgan failure)	Systemic spread of infection	High grade fever, tachycardia, tachypnea, hypotension oliguria, death
		Acute renal failure	Fluid loss and electrolyte imbalance	Altered sensorium, uremia, oliguria or anuria
		Stress ulcers	Reason not known	Hematemesis and melena

Table 30.5: Metabolic disturbances associated with burns (late effects)

Category	*Period*	*Clinical condition*	*Reason*	*Clinical presentation*
Local effects	Immediate	Tissue damage	Direct injury of cells	Extensive wounds
		Edema	Excessive accumulation of fluid in the extracellular spaces	Swollen areas surrounding or below the burnt areas
		Water and electrolyte loss	Excessive loss from the burnt area from large wounds	Large dry wounds
		Local sepsis	Bacterial growth on the burnt area	Purulent discharge or slough formation, fever with rigors
	Late	Delayed healing	Inadequate and improper treatment, diabetes mellitus	Chronic wound with unhealthy granulation (Fig. 30.18A)
		Keloid formation	Severe scarring	Elevated thick irregular skin lesion (Fig. 30.18B)
		Marjolin's ulcer	Low grade epidermoid carcinoma arising from the epithelium covering the scar tissue after burns	A very slow growing painless ulcer, on the scar without lymphatic spread, resembling squamous cell carcinoma, but the edge is not raised and floor does not contain florid granulation tissue (Fig. 30.19)
		Contractures	Fibrosis of burnt areas around neck, joints	Contracted areas with restricted mobility (Fig. 30.20)
		Disability	Pain and restricted mobility due to contractures	Extensive fibrotic tissues resulting in debility (Figs 30.21A to D)
		Strictures of tubular organs	Annular healing by fibrosis of tubular structures	Dysphagia, if esophagus is involved and dyspnea, if trachea is involved

PHYSICAL EXAMINATION

Clinical examination of a patient of burns should be done quickly but in detail.

Preliminaries Before Examination

- The patient should be weighed using a trolley weighing machine (determination of weight is necessary to calculate the fluid requirement)
- The patient should be sedated well after ensuring the airway patency, and if required intubated
- The patient should be completely undressed.
- Two intravenous portals should be established for fluid resuscitation and medication.

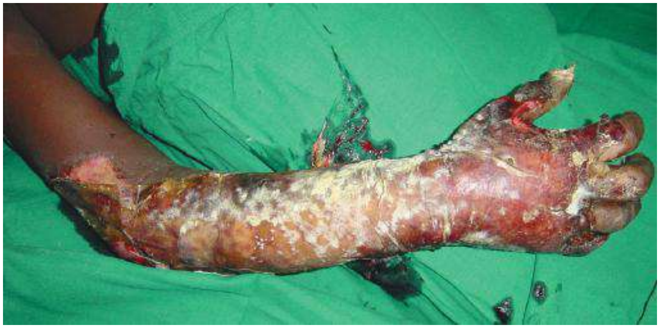

FIG. 30.18A: Infected burns wound

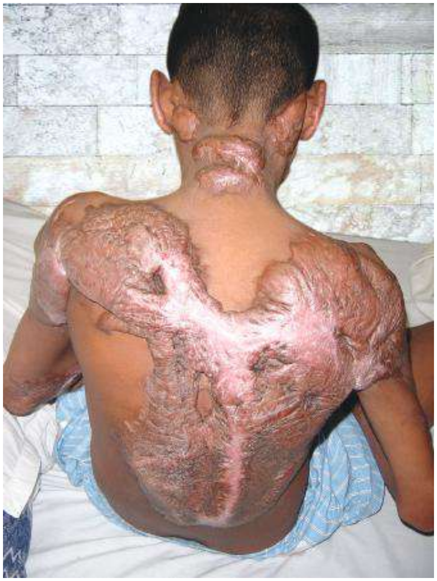

FIG. 30.18B: Keloid formation after burns (*Courtesy*: Dr K Sridhar)

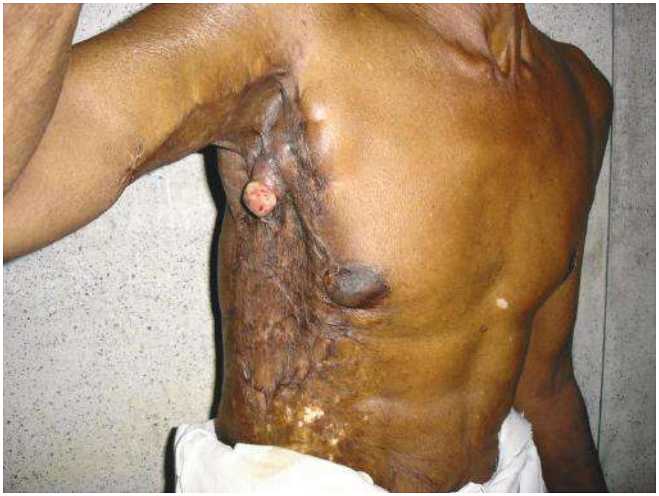

FIG. 30.19: Marjolin's ulcer chest wall (*Courtesy*: Dr R Rajaraman)

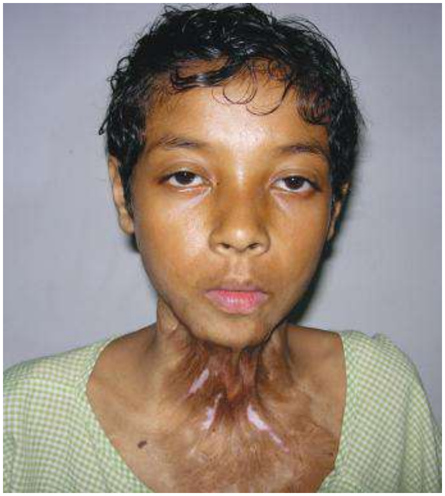

FIG. 30.20: Postburn contracture of neck (*Courtesy*: Dr K Sridhar)

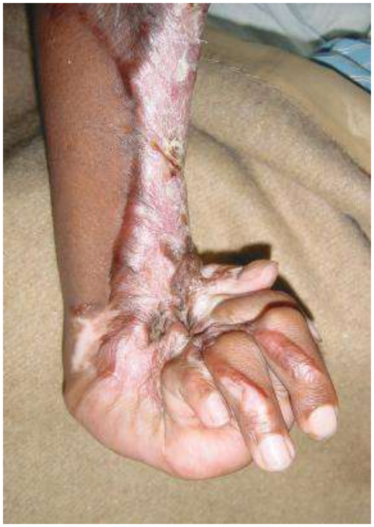

FIG. 30.21A: Postburn contractures (*Courtesy*: Dr K Sridhar)

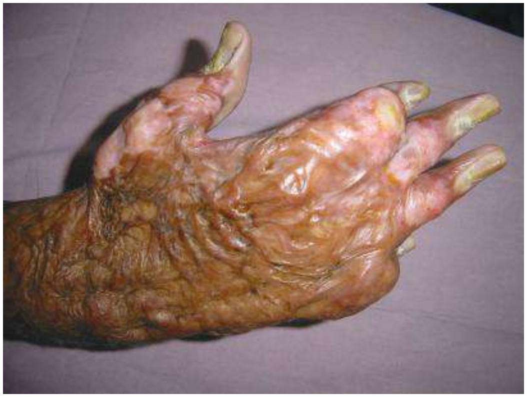

FIG. 30.21B: Postburn scarring resulting in syndactyly (*Courtesy:* Dr K Sridhar)

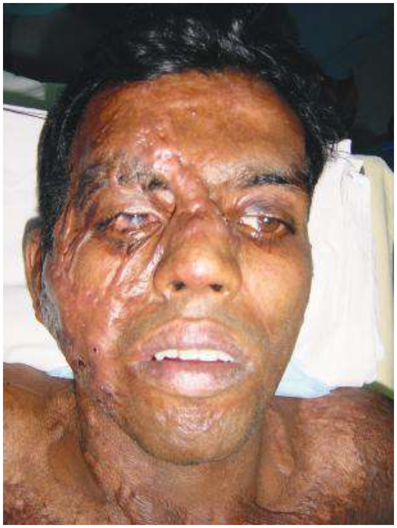

FIG. 30.21C: Postburns scarring resulting in disability (*Courtesy:* Dr K Sridhar)

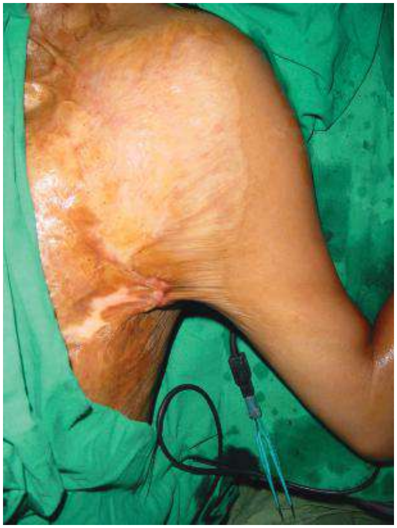

FIG. 30.21D: Postburns contracture of axillary fold (*Courtesy:* Dr K Sridhar)

Circumferential burn injuries (around the chest and limbs) especially for the third degree burns, escharotomy will have to be performed or otherwise, respiratory embarrassment and ischemic necrosis of limbs may occur.

INSPECTION

The patient should be examined head to foot, with special attention to the areas like the axillae, groins and the perineum. The extent of damage both by area and depth should be assessed quickly and precisely, as this helps in the determination of fluid replacement.

I. *Extent of area of burns:* This is calculated in two ways:
   i. For small, multiple and scattered areas: 'Rule of hand' is used; determination of surface area of burns, using the patient's own hand (with adducted fingers), which is equal to 1 percent
   ii. For large areas: 'Wallace's rule of nine is used (Fig. 30.22A). For children, the rule is modified (Fig. 30.22B), as the surface of area of the head is bigger relative to the other parts of the body.

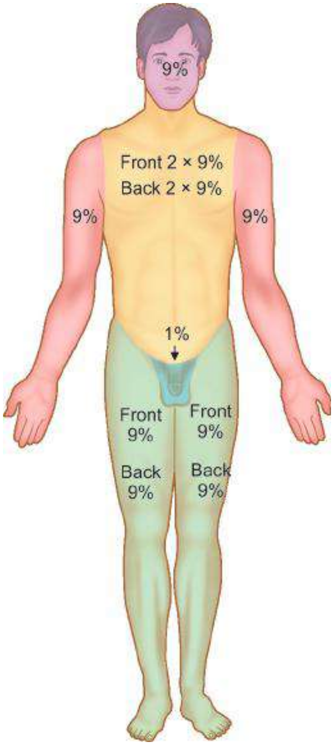
9%
Front 2 × 9%
Back 2 × 9%
9%
9%
1%
Front 9%
Front 9%
Back 9%
Back 9%

FIG. 30.22A: Wallace's rule of nine for adults

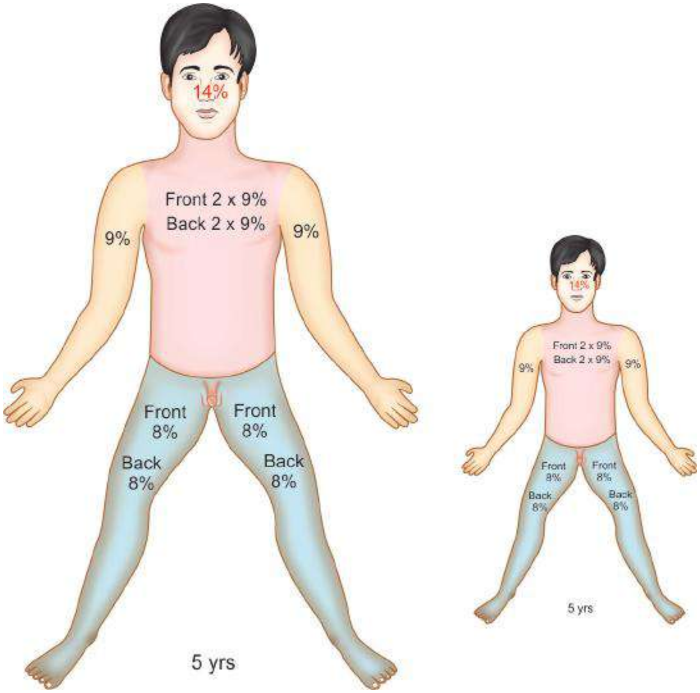
14%
Front 2 x 9%
Back 2 x 9%
9%
9%
Front 8%
Front 8%
Back 8%
Back 8%
5 yrs
14%
Front 2 x 9%
Back 2 x 9%
9%
9%
Front 8%
Front 8%
Back 8%
Back 8%
5 yrs

FIG. 30.22B: Modified rule of nine for children

II. *Depth of tissue damage:* This can be done by inspecting the burnt area, and reasonable conclusions may be reached.

It should be remembered that various degrees of tissue damage coexist in the same patient.

- Cutaneous sensation is tested by pricking the burnt area (sterile needle test) with a sterile needle. Presence of sensation is diagnostic of partial thickness burns, but it is not pathognomonic.

Plucking the hair from the burnt area can show the depth. In full thickness burns, the hair can be plucked easily, whereas in partial thickness burns it is not.

General Examination

- *Examination of sensorium:* Usually the sensorium remains normal and not willing to answer may indicate a psychiatric problem (e.g. depression or withdrawal). Inability to answer may indicate state of unconsciousness or altered sensorium (e.g. shock, hyponatremia, uremia)
- *Examination of face:*
    - *Eyes:* Sunken eyes indicate dehydration and undernutrition. Examination of conjunctivae will reveal anemia.
    - *Hairy areas:* Surging of hair in eyebrow, moustache or over forehead will indicate respiratory burns
- *Examination of skin:* The skin should be examined for dehydration and undernutrition. Generalized edema may be seen in severely burnt patients.
- *Examination of tongue:* Dry tongue indicates dehydration
- *Recording of vital signs:* Tachycardia may indicate infection or hypovolemia. Elevated temperature may indicate associated infections and toxemia (usually seen in the second week). Hyperventilation may indicate hydration, hypovolemia and is predominant in respiratory burns. Hypotension may indicate hypovolemia.

Systemic Examination

- *Examination of oral cavity:* A systematic examination of oral cavity is essential (e.g. chemical burns, burns due to inhalation of gases)
- *Examination of chest:* Auscultation—Adventitious sounds may indicate aspiration pneumonitis or respiratory infections.

Relevant Investigations

- Hematology
    - Hemoglobin for anemia
    - Total and differential leukocyte count, e.g. infections

  - ESR may be raised in infections and malignancies
    Note: Hemoglobin levels initially remain normal due to hemoconcentration and serial determinations are necessary on subsequent days.
- Biochemistry
  - *Blood sugar:* Required especially in diabetics. Repeated determination is required in patients on hyperalimentation
  - *Liver function tests:* Reduced levels of serum proteins indicate loss of proteins through burn wound and will help in the replacement with blood or plasma
  - *Renal function tests:* Uremia and creatininemia indicate deranged renal status due to dehydration and hypovolemia
  - *Serum electrolytes:* Determination of serum levels of electrolytes is important in the management of burns, as electrolyte disturbances are commonly found in burns, due to evaporation during injury and later through wound.
- Blood grouping and crossmatching: It is important as the requirement of blood may be immediate.
- Radiology
  - Chest X-ray is useful (e.g. aspiration pneumonitis, inhalation injuries of lungs)
  - Contrast studies
    - Barium swallow may be required at later stages to rule out esophageal strictures
    - Bronchogram also may be required to rule out strictures of tracheobronchial tree.
- Endoscopy
  - Esophagoscopy and bronchoscopy may be required in some cases to assess the esophageal and bronchial injuries respectively, especially when injuries occur due to inhalation of toxic fumes
  - Gastroscopy may be needed to diagnose stress ulcers of stomach.
- *Electrocardiogram (ECG):* To rule out disturbances of heart especially in electrical burns, and may have to be repeated frequently.

Treatment

Resuscitation

- Basic principles of the burns victim are the same as for any other patient.
- Problems are specifically related to thermal injuries to airways, large fluid losses and potential for infection:
  - Cover burn areas with sterile drapes or plastic film to reduce infection and fluid loss
  - Give humidified oxygen by mask
  - Endotracheal intubation is required for airway injuries
  - Monitor hematocrit and electrolytes
  - Blood may be required to maintain hematocrit

- Good IV access is needed for fluid administration
  - IV analgesia
  - IV fluids

Examples of fluid regimens for resuscitation of burn victims

Mount Vernon formula	*Parkland formula*
4.5% albumin	Ringer lactate
Volume (ml) = 0.5 × wt (kg) × % burn	Volume (ml) = 4 × wt (kg) × % burn
Over six consecutive periods of 4, 4, 4, 6, 6, and 12 hours each	Given over 24 hours

NECROTIZING FASCIITIS

Incidence and Etiology

- A rapidly spreading soft tissue infection, which can affect any part of the body
- It follows, usually a minor trauma but the exact etiology is not fully understood
- Causative organisms—Mixture of gram-negative and anaerobic organisms (bacteroides and *Clostridium* sp. anaerobic streptococci) and/or group A *streptococcus*.

Pathogenesis

The infection spreads along the fascial planes causing necrosis of skin and subcutaneous tissues. Muscle layers are usually spared. The infection may spread rapidly and can be fatal in few hours.

Symptom

Painful discolored area.

Sign

The affected area is discolored and indurated.

Relevant Investigations

- Isolation of organism in pus by culture
- Determination of blood sugar levels is mandatory.

Treatment

- Control of diabetes is necessary
- Emergency excision of slough and necrotic tissue under cover of appropriate antibiotics is curative.

Section VI

Special Procedures and Situations

31. Emergency Surgical Procedures
    - Pericardial Aspiration
    - Intubation of Trachea
    - Percutaneous Tracheostomy
    - Cricothyroidotomy/ Minitracheostomy
    - Insertion of Chest Drains
    - Passing a Nasogastric Tube
    - Peritoneal Tap
32. Preparing the Patient for Emergency Surgery
33. Antibiotics
    - Antibiotics and Emergency Surgery
34. Death

31

Emergency Surgical Procedures

PERICARDIAL ASPIRATION

Pericardial aspiration (Pericardiocentesis) is indicated in:

- Cardiac tamponade
- Large pericardial effusion
- For diagnostic pericardial fluid.

Materials Required

- Sterile gowns and gloves
- 10 ml syringe with needle
- Pericardial aspiration kit
- Sutures
- Securing tapes.

Procedure

- Position: Supine with 20 degrees head-up
- Establish IV access
- Adequate sedation
- With full aseptic technique, infiltrate the point of needle entry (just below and left of xiphisternum, between the xiphisternum and left costal margin)
- Introduce the needle directed to the left shoulder, and keep aspirating (Fig. 31.1)

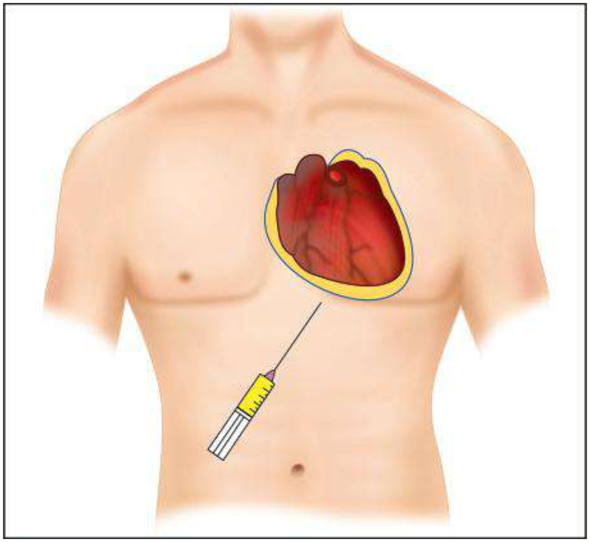

FIG. 31.1: Pericardial aspiration

- Once a straw colored fluid is drawn, keep the needle in that position, pass the guidewire and withdraw the needle
- Pass the catheter over the guidewire into the pericardial space and attach a 3-way tap
- Using a 50 ml syringe, aspirate the effusion or attach to a closed drainage system
- Suture the drain in place and strap.

Complications

- Pneumothorax
- Ventricular tachycardia
- Myocardial puncture
- Damage to coronary arteries.

- Small pericardial effusions not causing hemodynamic instability do not require pericardiocentesis
- Repeat chest X-ray and ECG

INTUBATION OF TRACHEA

Though this procedure is performed by the anesthetists it is better for the surgeons and surgical students to know, as this pertains to life-saving situation. They fall into three groups for which it is required.

- Relieving airway obstruction (e.g. tumors, head and neck trauma, surgery, airway edema)
- Protection of airway from aspiration (e.g. obtunded conscious level, impaired cough impulses)
- Facilitation of ventilation of airways (e.g. anesthesia and surgery, multiple organ failure, major trauma and brain injury).

Materials Required

- Self-inflating bags (Ambu bag)
- Face mask
- Oral/nasal airways
- Suction apparatus and suction catheters
- Laryngoscopes
- Endotracheal tubes (or various sizes)
- Sterile lubricant
- Syringe
- Anesthetic drugs/muscle relaxants
- Emergency drugs (atropine, adrenaline).

Procedure

- Oxygenate the patient well with 100 percent oxygen for about 3 to 4 minutes, which will wash out the nitrogen and fill the functional residual capacity with oxygen, thereby increasing the safety margin
- Keep the head in position (neck flexed, atlantoaxial joint extended on a firm pillow – 'sniffing the morning air' position)
- Give sedative/muscle relaxant as appropriate
- Hold the laryngoscope in the left hand and insert the blade into the right of the mouth sweeping the tongue under it. As the blade reaches the base of the tongue, the epiglottis is seen. Apply traction to gently draw the epiglottis forward exposing the V-shaped glottis behind
- Pass the endotracheal tube between the vocal cords so that cuff is distal to them
- Withdraw the laryngoscope gently
- Inflate the cuff of the endotracheal tube with air so that the tube snugly fits into the trachea
- Check the position of tube by auscultation and by observing the chest movements.

Complication of Endotracheal Intubation

Immediate

- Hypoxia (due to repeated attempts)
- Misplacement of tube
- Obstruction of airway
- Trauma to airway, teeth.

Late

- Accidental displacement of tube
- Ventilator associated pneumonia
- Injury to vocal cords
- Tracheal stenosis.

> Tracheal intubation is to done only if the student is familiar with this and has done with a senior in the past.

PERCUTANEOUS TRACHEOSTOMY

Conventional tracheostomy was performed only after patients had been intubated for about 10 – 14 days for fear of laryngeal and subglottic injury. Percutaneous tracheostomy has come into practice as the benefits are many.

Advantages of Tracheostomy

- More comfortable than the oro/nasotracheal tubes
- Need for muscle relaxants is less
- Switching from IPPV to T piece is easy without extubation
- Effective endotracheal suction can be done
- Speech is possible with cuff deflation or speaking tube.

Indications

- Actual or impending airway obstruction
- Difficult intubation
- Need for prolonged ventilator support.

Contraindications

- Distorted/abnormal anatomy
- Significant hemodynamic instability.

Materials Needed for Percutaneous Tracheostomy

- Skin disinfectant
- Local anesthetic
- Syringe
- 10 ml normal saline
- Basic surgical instruments
- Suture
- Securing tapes
- Emergency drugs
- Percutaneous tracheostomy kit (Fig. 31.2).

Procedure

- Position: supine with neck extended over a pillow
- Palpate the cricothyroid membrane and sternal notch

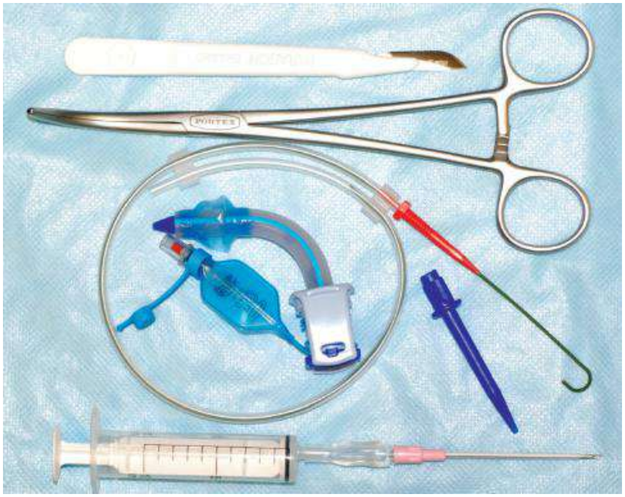

FIG. 31.2: Percutaneous tracheostomy kit

- Infiltrate with local anesthetic
- Make a 2 cm transverse incision over the membrane at the midline
- Puncture the trachea (between the first and third tracheal rings) with the introducer needle
- Confirm its position with a syringe aspiration (air/mucus)
- Pass the guidewire through the needle and remove the needle
- Dilate the trachea
- Insert the endotracheal tube
- Remove the guidewire
- Suck out all the secretions and blood clots in the trachea
- Inflate the cuff
- Place two stay sutures through the wings of tracheostomy tube, to prevent slippage, and secure with tapes
- Check the position of the tracheostomy tube (symmetrical chest expansion, X-ray).

Complications

Early

- Bleeding
- Tube misplacement
- Mucus plugging

Late

- Tracheal stenosis
- Tracheoesophageal fistula.

CRICOTHYROIDOTOMY/MINITRACHEOSTOMY

Cricothyroidotomy is an emergency procedure to access the airway, when measures like intubation have failed. It involves the insertion of a small tube through the cricothyroid membrane for attaching to the ventilator.

Minitracheostomy involved the use of a small bored and noncuffed bore, with an internal diameter of 4 mm. Kits are available for both procedures.

Indications

- Cannot intubate
- Cannot ventilate
- Severe midfacial trauma

- Cervical spine injuries
- Chemical inhalation injuries.

Contraindications

- Inability to identify landmarks (cricothyroid membrane)
- Underlying anatomical abnormality (tumor)
- Acute laryngeal disease
- Tracheal transection
- Small children.

Materials Required

- Skin disinfectant
- Sterile drape
- Syringe with local anesthetic
- Cricothyroidotomy (Fig. 31.3)/Minitracheostomy kit
- Suture
- Securing tapes.

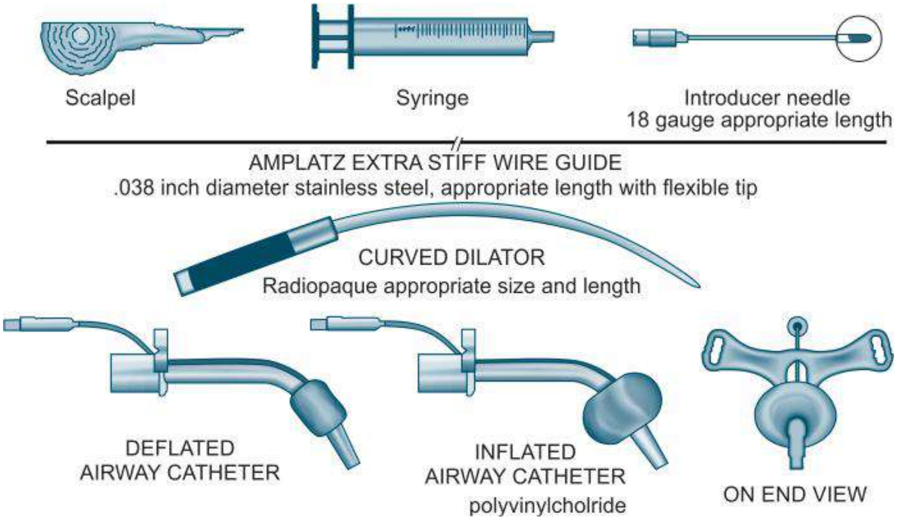
Scalpel
Syringe
Introducer needle
18 gauge appropriate length
AMPLATZ EXTRA STIFF WIRE GUIDE
.038 inch diameter stainless steel, appropriate length with flexible tip
CURVED DILATOR
Radiopaque appropriate size and length
DEFLATED
AIRWAY CATHETER
INFLATED
AIRWAY CATHETER
polyvinylcholride
ON END VIEW

FIG. 31.3: Cricothyroidotomy kit

Procedure

- Palpate the cricothyroid membrane
- Clean the skin
- Infiltrate over the cricothyroid membrane
- Make a skin incision
- Introduce needle into trachea and aspirate air
- Pass guidewire and remove needle
- Pass the introducer over the guidewire and slide the cricothyroidotomy/minitracheostomy tube off the introducer
- Remove the introducer and guidewire together, leaving the tube *in situ*
- Suck any blood or mucus
- Place two stay sutures through the wings of tracheostomy tube, to prevent slippage, and secure with tapes
- Check the position of the tracheostomy tube (symmetrical chest expansion, X-ray).

Complications

Early

- Bleeding
- Tube misplacement
- Mucus plugging

Late

- Tracheal stenosis
- Tracheoesophageal fistula.

INSERTION OF CHEST DRAINS

Insertion of a chest drain is required as an emergency treatment of life-threatening tension pneumothorax. A large bore tube preferably a 14-gauge cannula is used for this purpose.

Indications

Draining of:

- Pneumothorax
- Hemothorax
- Pleural effusion

- Empyema
- Chylothorax.

Type of Drain

Chest drains are of two types:

1. Large bore tubes
2. Seldinger drains.

Site of Drain

This is partly decided by the position of the collection clinically and radiographically.

- For draining the pneumothorax, the tube is placed in the 2nd intercostal space in the midclavicular line (Fig. 31.4).
- For draining the fluid collections, the tube is placed in the 5th intercostals space just anterior to the midaxillary line (Fig. 31.4).

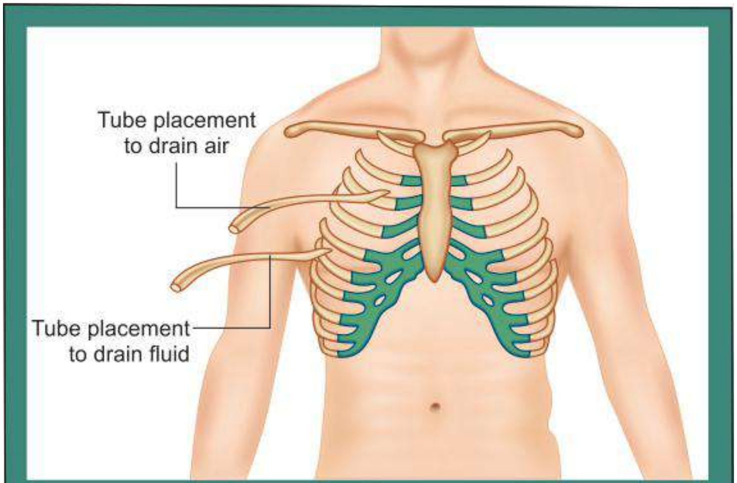

CHEST TUBE INSERTION PROCEDURE

A patient may have one or more chest tubes inserted, depending on the kind of drainage. Generally, chest tubes to drain air are placed high and anterior whereas chest tubes to drain fluid are placed low and directed posteriorly. This is because air has a tendency to rise to the apical region of the pleural space while fluid gravitates to the basal region of the pleural space.

FIG. 31.4: Placing the chest drains

Materials Required

- Skin disinfectant
- Skin drapes
- Syringe
- Local anesthetic
- Basic instruments
- Chest drain
- Silk sutures
- Adhesive tapes
- Underwater seal.

Procedures

Through Thoracostomy (Figs 31.5A to F)

- Position: supine with arm lifted, with a pillow behind the back
- Clean the field with sterile preparation
- Infiltrate the local anesthetic at the chosen site
- Make a small skin incision
- The intercostal muscles and the pleura are dissected out and plunged with artery forceps
- The chest drain is inserted into the pleural cavity
- The drain is connected to the underwater drain (Fig. 31.6)
- The drain is secured with sutures to the skin
- Adhesive strapping is done
- Order for a X-ray and the position checked.

Seldinger Technique

- Position: supine with arm lifted, with a pillow behind the back
- Clean the field with sterile preparation
- Infiltrate the local anesthetic at the chosen site
- Use the chest drain kit (Fig. 31.7)
- Advance the needle through chest until blood/fluid/air is aspirated
- Pass chest tube/pigtail catheter over guidewire
- Attach to the underwater seal
- Secure well and dress.

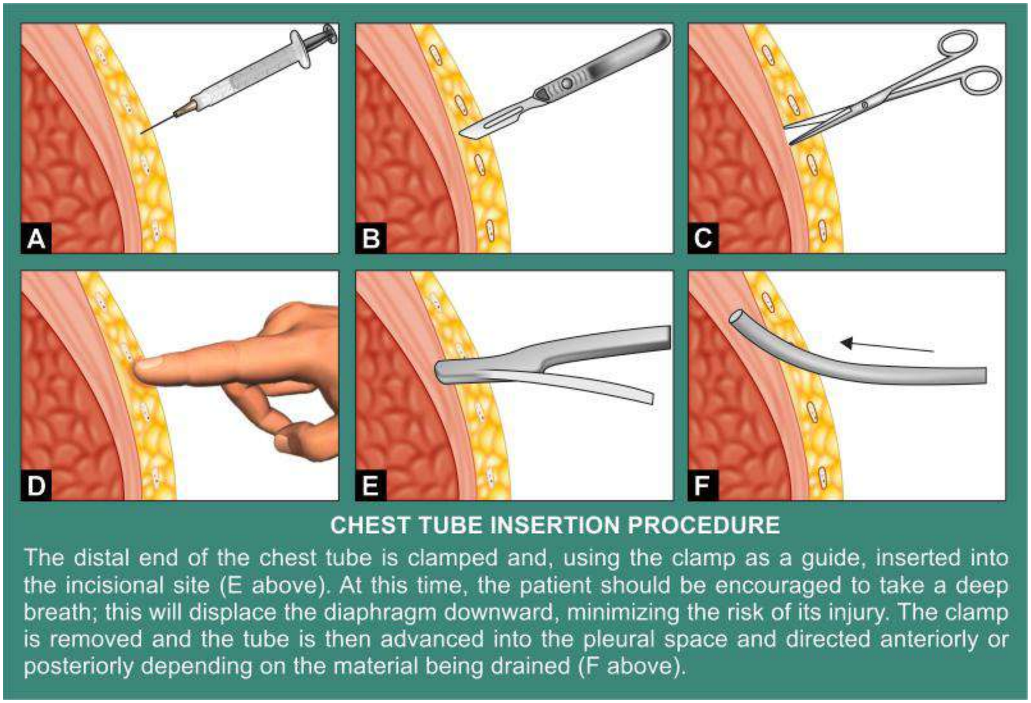

CHEST TUBE INSERTION PROCEDURE

The distal end of the chest tube is clamped and, using the clamp as a guide, inserted into the incisional site (E above). At this time, the patient should be encouraged to take a deep breath; this will displace the diaphragm downward, minimizing the risk of its injury. The clamp is removed and the tube is then advanced into the pleural space and directed anteriorly or posteriorly depending on the material being drained (F above).

FIGS 31.5A to F: Method of placing the chest drain

- Never clamp the chest drain, as clamping may produce a tension pneumothorax
- While shifting the patient, the underwater drain should be kept below the level of chest

PASSING A NASOGASTRIC TUBE

Nasogastric tube is generally used to decompress the stomach and also for feeding.

Indications

- To deflate the stomach
- To aspirate gastric contents
- To provide enteral nutrition.

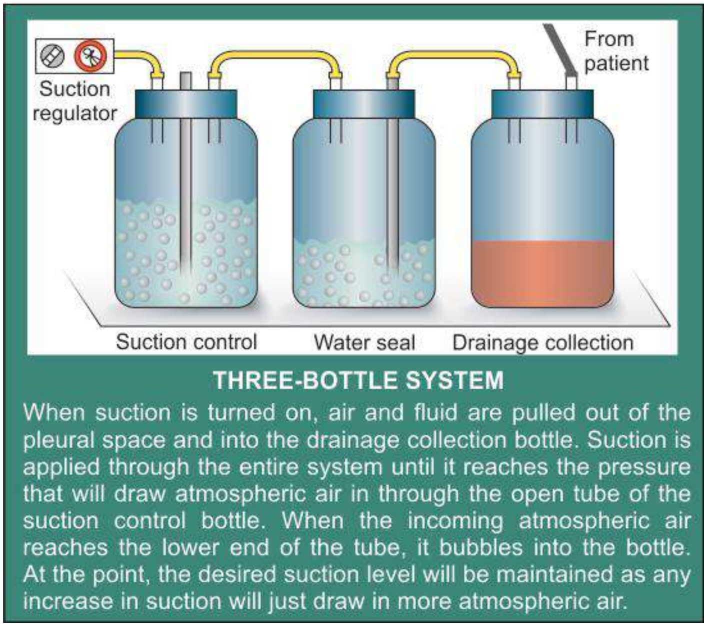
From patient
Suction regulator
Suction control
Water seal
Drainage collection

THREE-BOTTLE SYSTEM

When suction is turned on, air and fluid are pulled out of the pleural space and into the drainage collection bottle. Suction is applied through the entire system until it reaches the pressure that will draw atmospheric air in through the open tube of the suction control bottle. When the incoming atmospheric air reaches the lower end of the tube, it bubbles into the bottle. At the point, the desired suction level will be maintained as any increase in suction will just draw in more atmospheric air.

FIG. 31.6: Three chamber underwater drain

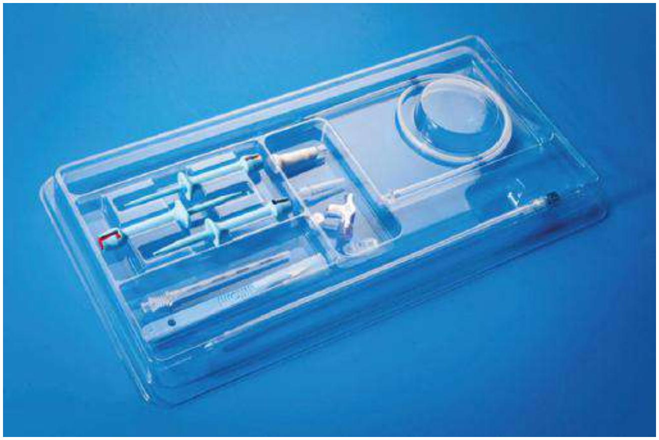

FIG. 31.7: Seldinger chest drain kit

Contraindications

Severe coagulopathy.

Materials Required

- Gloves and mask
- Nasogastric tube
- Lubricating jelly
- Laryngoscope (needed for unconscious patients).

Procedure (Fig. 31.8)

- Position: supine and head neutral or semirecumbent
- Lubricate the nasogastric (NG) tube well
- Keeping in alignment with the long axis of the patient, introduce the tube into the nostril
- If the patient is cooperative, ask him to swallow
- The tube generally flows down along with the swallowing movement
- Confirm the tip of NG tube in the stomach by auscultation or aspiration of contents
- Secure NG tube in position with adhesive tape.

PERITONEAL TAP

Peritoneal tap is required to obtain samples for diagnostic purposes and also relieve the tension of a tense ascites.

Materials Required

- Skin disinfectant
- Sterile drapes
- 10 ml syringe with local anesthetic
- Pig tail drain
- Sutures
- Adhesive strapping
- Ultrasound.

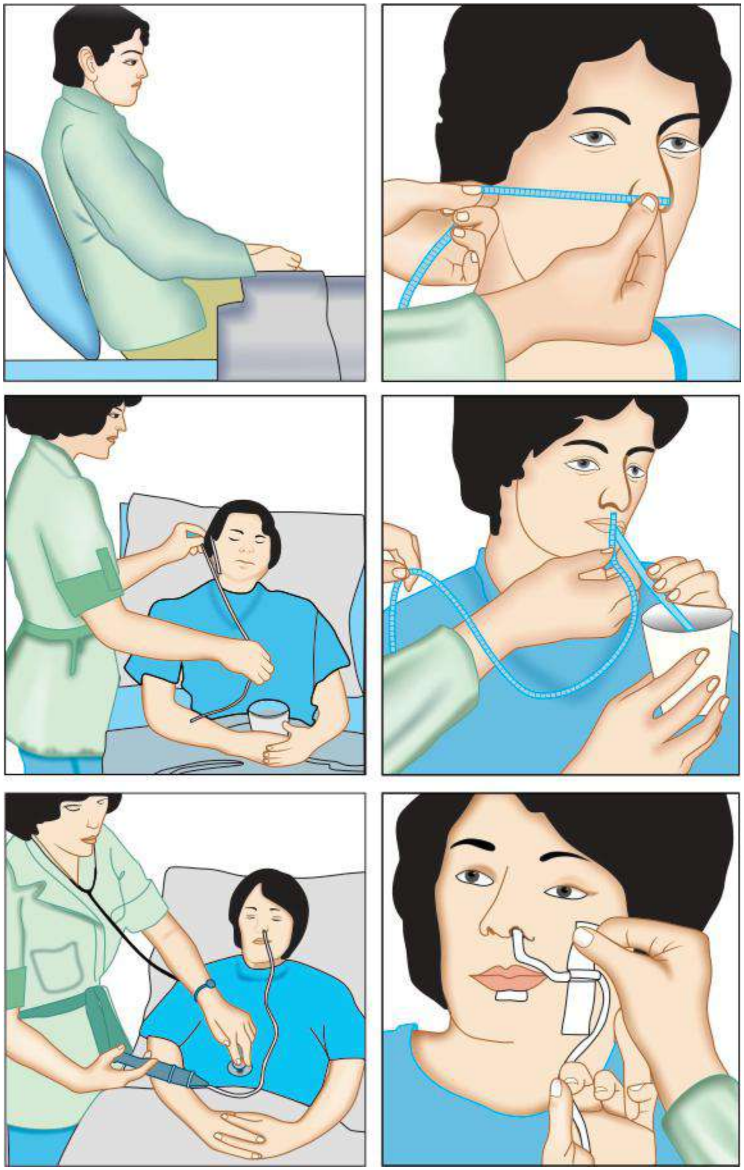

FIG. 31.8: Nasogastric tube insertion

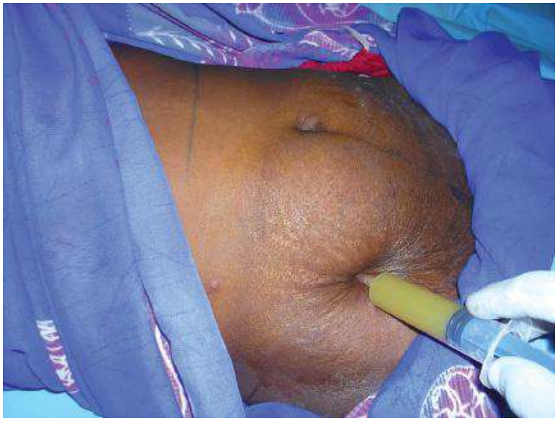

FIG. 31.9: Peritoneal tap

Procedure

- Clean the skin with disinfectant
- Confirm the presence of fluid by percussion or by ultrasound
- At the chosen site, inject the local anesthetic
- Introduce the syringe and aspirate fluid to confirm its presence (Fig. 31.9)
- Introduce pig tail drain using Seldinger technique
- Connect the drain to collecting bottle
- Strap the drain.

32

Preparing the Patient for Emergency Surgery

Any patient for surgery needs optimization, especially so the patients for emergency surgery, as the parameters are never normal. This optimization becomes necessary to reduce the perioperative complications. The optimization requires:

- Fluid and electrolyte corrections
- Correction of hematological and metabolic disturbances.

General Measures

- Nil by mouth
- Good intravenous access
- Appropriate intravenous fluid replacement
- Appropriate antibiotics
- Adequate analgesia
- Measures for DVT prophylaxis
- NG tube insertion where stomach needs to be decompressed
- Urinary catheterization
- Central venous line access
- Informed consent.

Investigations Support

- Complete blood count (CBC)
- Urea and electrolytes
- Electrocardiogram (ECG) and Chest X-ray (CXR).

Situations which Need Special Attention

- Uncontrolled glycemic status
- Ketoacidosis
- Cardiovascular complications
- Altered conscious level
- Jaundiced patient with coagulation abnormalities
- Patient on steroids
- Patient on anticoagulants.

33

Antibiotics

ANTIBIOTICS AND EMERGENCY SURGERY

Antibiotics become necessary to prevent an infection or control an infection. Prevention becomes necessary so that sepsis related complications, which carry high morbidity and mortality are averted.

Prophylactic antibiotics are administered within 1 hour before the start of the surgical procedure. Second dose becomes necessary only when contamination is encountered or the blood loss exceeds 1 liter, since it is shown that antibiotic clearance is slower in patients undergoing surgery. Postoperative administration is only required when there is heavy contamination.

Which Emergency Surgical Procedure should have Antibiotic Prophylaxis ?

Antibiotic prophylaxis is recommended for the following emergency surgical procedures

- Colorectal surgery
- Appendicectomy
- Hernia surgery (use of mesh)
- Open biliary surgery
- Lower limb revascularization surgery.

The surgeries are classified based on the contamination at the time of surgery, and the infection rates vary differently (Table 33.1).

Surgical site infections (SSIs) are defined as infective complication occurring within 30 days of surgical procedure. The infection is caused by various organisms and some antibiotics are effective against these organisms (Table 33.2).

Table 33.1: Classification of surgeries related to contamination

Classification of case	*Type of surgery*	*Wound infection rate*
Clean	Surgery in noninflamed tissue, and bowel not opened	<2%
Clean contaminated	Surgery of hollow visus other than colon, with minimal contamination	10%
Contaminated	Surgery through a hollow viscus with gross spillage	20%
Dirty	Fecal peritonitis, traumatic wound contaminated for >4 hrs, frank pus	40%

Table 33.2: Relationship of surgical procedures with potential organisms and sensitive antibiotics

Surgical procedures	*Organism*	*Antibiotics*
General surgical procedures	*Staphylococcus aureus*	Flucloxacillin, ciprofloxacin, clindamycin
	Beta – hemolytic streptococci	Ciprofloxacin, clindamycin
Upper GIT	Enteric Gram-negative bacilli	Cefuroxime, co-amoxyclav, gentamycin
Lower GIT	Enteric Gram-negative bacilli	Cefuroxime, co-amoxyclav, gentamycin
	Anaerobes	Metronidazole
Procedures using prosthesis	Coagulase + staphylococci	Vancomycin + removal of prosthesis

Table 33.3: Antibiotics related to Gram staining

Gram stain	*Antibiotic*
Gram-negative rods	Cefuroxime + / – gentamycin +/– metronidazole
Gram-positive cocci (Streptococci)	Amoxicillin
Gram-positive cocci (Staphylococci)	Flucloxacillin (Vancomycin / Teicoplain)
Gram-negative cocci	Ceftriaxone

Prior to full culture reports becoming available, a Gram stain may provide information to the likely organism. Empiric antibiotics can be used on Gram stain with fair accuracy (Table 33.3).

- Infection prevention and control measures must be adhered to
- Sepsis is usually bacterial in origin, though fungi and viruses should be considered
- Obtaining appropriate microbiology samples will help to use the right antibiotic
- Choice of antibiotic should be carefully considered.

34

Death

A death is certified only when the following are documented:

- Absent respiratory rate for 3 minutes
- Absent pulse and heart sounds for 1 minute
- Fixed dilated pupils.

Breaking the News

Breaking a bad news especially of the impending death or after death is very tricky. This will upset the family which inherently is stressed. Well informed family will accept but ill informed or less informed will not take it very easy. To overcome this situation, the clinician should appear:

- Competent
- Honest
- Clear.

It is better to choose a suitable environment for discussion with the close relatives. Crowding of all family members and relatives should be politely discouraged. While discussing this it is better to:

- Review the case record in full
- Prepare the anticipated questions with your colleagues
- Keep a colleague or nurse alongwith
- Understand the current level of understanding
- Explain in detail

AUDEL™
THE BOOKS THAT WORK™

CARPENTER'S AND BUILDER'S MATH, PLANS & SPECIFICATIONS

ALL NEW 7TH EDITION

Mark Richard Miller & Rex Miller

Audel™

Carpenter's and Builder's Math, Plans, and Specifications

All New 7th Edition

Mark Richard Miller
Rex Miller

WILEY
Wiley Publishing, Inc.

TLFeBOOK

Vice President and Executive Group Publisher: Richard Swadley
Vice President and Publisher: Joseph B. Wikert
Executive Editor: Carol Long
Editorial Manager: Kathryn A. Malm
Development Editor: Kevin Shafer
Production Editor: Angela Smith
Text Design & Composition: TechBooks

Copyright © 2005 by Wiley Publishing, Inc., Indianapolis, Indiana. All rights reserved.

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Legal Department, Wiley Publishing, Inc., 10475 Crosspoint Blvd., Indianapolis, IN 46256, (317) 572-3447, fax (317) 572-4355, e-mail: brandreview@wiley.com.

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Website is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Website may provide or recommendations it may make. Further, readers should be aware that Internet Websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Trademarks: Wiley, the Wiley Publishing logo, Audel, and The Books That Work are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates. All other trademarks are the property of their respective owners. Wiley Publishing, Inc., is not associated with any product or vendor mentioned in this book.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Control Number: 2004110825

eISBN: 0-7645-7981-9

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments xiii

About the Authors xv

Introduction xvii

Chapter 1 Nails, Screws, Bolts, and Other Fasteners 1
Nails 1
  The Penny System 3
  Nail Characteristics 3
  Selecting Nail Size 12
  Driving Nails 13
Screws 16
  Material 17
  Dimensions of Screws 18
  Shape of the Head 18
  How to Drive a Wood Screw 18
  Strength of Wood Screws 21
  Lag Screws 23
Bolts 23
  Manufacture of Bolts 23
  Kinds of Bolts 27
  Proportions and Strength of Bolts 27
Fasteners for Plaster or Drywall 33
  Expansion Anchors 33
  Hollow Wall Screw Anchors 34
  Toggle Bolts 36
  Hanger Bolts, Dowel Screws, and Toggle Studs 37
  Framing Fasteners 39
Summary 39
Review Questions 40

Chapter 2 Wood as a Building Material 41
Growth and Structure of Wood 41
Softwoods and Hardwoods 41

Lumber Conversion 43
Seasoning of Wood 44
    Moisture Content 47
    Density 49
    Strength 51
    Deadwood 53
    Virgin and Second Growth 54
    Time of Cutting Timber 54
    Air-Dried and Kiln-Dried Wood 54
    Sapwood Versus Heartwood 55
Defects and Deterioration 55
    Decay of Wood 57
    Nuclear Radiation 63
    Weathering 63
    Insect Damage 65
Summary 65
Review Questions 65

Chapter 3 Lumber, Plywood, and Other Wood Products 67
Standard Sizes of Lumber 67
Softwood Lumber Grades for Construction 68
    Stress-Graded Lumber 68
    Nonstress-Graded Lumber 69
    Appearance Lumber 70
Plywood 77
    Types of Plywood 79
    Grades of Plywood 79
Particle Board 79
Lumber Distribution 82
    Retail Yard Inventory 82
    Important Purchase Considerations 86
Engineered Lumber 87
Summary 88
Review Questions 88

Chapter 4 Strength of Timbers 89
Tension 89
Compression 90
Working Stresses for Columns 91
Shearing Stresses 92
Horizontal Shears 94
Transverse or Bending Stress 96
Stiffness 96
Modulus of Elasticity 100
Beams 100
  Allowable Loads on Wood Beams 100
  Breaking Loads on Wood Beams 102
  Distributed Load 102
  Cantilever Beams 102
Wind Loads on Roofs 104
Definitions 106
Summary 108
Review Questions 109

Chapter 5 Mathematics for Carpenters and Builders 111
Arithmetic 111
  Arithmetic Alphabet 111
  Notation and Numeration 111
  Roman Notation 113
  Definitions 113
  Signs of Operation 115
  Use of the Signs of Operation 116
  Fractions 117
  Decimals 122
  Compound Numbers 126
  Ratios 127
  Proportion 127
  Rule of Three 128
  Percentage 128
  Powers of Numbers (Involution) 129
  Roots of Numbers (Evolution) 129

Measures 131
- Linear Measure 132
- Square Measure 133
- Cubic Measure 133
- Dry Measure 135
- Board or Lumber Measure 136
- Liquid Measure 137
- Measures of Weight 137
- Time Measure 139
- Circular Measure 139
- The Metric System 140

Geometry 142
- Lines 142
- Angles 142
- Plane Figures 143
- Solids 146
- Geometrical Problems 146
- Mensuration 167

Trigonometry 182
- Trigonometric Functions 182

Functions of Numbers 197
Summary 197
Review Questions 197

Chapter 6 Surveying 199

The Level 199
- Construction of the Wye Level 199
- Lines of the Level 201
- Adjustments of the Wye Level 202
- Leveling Rod 205
- Methods of Leveling 206
- Directions for Using Level 209
- Trigonometric Leveling 211

The Transit 212
- Construction of the Transit 212
- Lines of a Transit 218
- Adjustments of the Transit 219

Adjustments of the Compass 221
Instructions for Using the Transit 222
Gradienter 224
Care of Instruments 224
The Stadia 225
Other Devices 226
Laser Levels 227
Summary 232
Review Questions 232

Chapter 7 The Design Process 233
Design Considerations 233
An Example of Design 236
Conference with Builder 236
Changes Agreed On 238
Summary 240
Review Questions 241

Chapter 8 Building Specifications 243
Example of Specifications 243
Introduction 243
Height of Ceilings 243
Interpretation of Drawings 244
Conditions 244
Mason's Work 244
Foundation 245
Chimneys 245
Mortar 245
Installing Drywall 245
Tiling 245
Other Floors 246
Coping 246
Timber 246
Framing 246
Spacing and Bridging 247
Partitions 247
Lumber 247

Sheathing and Sheathing Paper 247
Exterior Finish 247
Shingling 247
Flashing 248
Flooring 248
Window Frames 248
Sash 248
Screens 248
Glazing 248
Blinds 248
Door Frames 248
Doors 249
Interior Trim 249
Stairs 249
Mantels 249
Pantry Cabinets 250
Closet Shelving 250
Plumbing 250
Electric Wiring 250
Water Pipes 250
Kitchen Sink 251
Wash Trays 251
Bathroom Fixtures 251
Painting 251
Condition of Bids 252
Summary 252
Review Questions 252

Chapter 9 Architectural Drawings 255
Use of Drawings 255
Reading Drawings 255
Projected Views 255
Top View or Plan 256
Right-End View (Elevation) 256
Left-End View (Elevation) 258
Sections 258

  Directions of View for Sectional Views 259
  The Scale 260
Drawing Development 262
Graphic Symbols 267
Summary 278
Review Questions 279

Chapter 10 Building Styles Explored 281
The Two-Story New England Colonial House 281
A House of Modern Architecture 284
The Contemporary House 285
Utility Pole-Type Building 287
Summary 291
Review Questions 292

Appendix Construction Hardware 293

Index 349

Chapter 2

Wood as a Building Material

Wood is the most versatile, most useful building material, and a general knowledge of the physical characteristics of various woods used in building is important for carpenters and builders.

Growth and Structure of Wood

Wood, like all plant material, is made up of cells (or *fibers*) that, when magnified, have an appearance similar to (though less regular than) that of the common honeycomb. The walls of the honeycomb correspond to the walls of the fibers, and the cavities in the honeycomb correspond to the hollow or open spaces of the fibers.

Softwoods and Hardwoods

All lumber is divided as a matter of convenience into two great groups: *softwoods* and *hardwoods*. The softwoods in general are the coniferous or cone-bearing trees (such as the various pines, spruces, hemlocks, firs, and cedar). The hardwoods are the non–cone-bearing trees (such as the maple, oak, and poplar). These terms are used as a matter of custom, for not all so-called softwoods are soft, nor are all so-called hardwoods necessarily hard. In fact, such softwoods as long-leaf southern pine and Douglas fir are much harder than poplar, basswood, and so on, which are called hardwoods.

Other (and perhaps more accurate) terms often used for these two groups are the needle-bearing trees (softwoods) and the broad-leaved trees (hardwoods). In general, the softwoods are more commonly used for structural purposes such as for joists, studs, girders, and posts, whereas the hardwoods are more likely to be used for interior finish, flooring, and furniture. The softwoods are also used for interior finish and, in many cases, for floors, but are not often used for furniture.

A tree consists of the following:

- *Outer bark*—The bark is living and growing only at the cambium layer. In most trees, the bark continually sloughs away.
- *Inner bark*—In some trees (notably hickories and basswood) there are long tough fibers (called *bast fibers*) in the inner bark. In other trees (such as the beech), these bast fibers are absent.
- *Cambium layer*—This can be only one cell thick. Only these cells are living and growing.

- *Medullary rays or wood fibers*—These run radially from the center to the bark.
- *Annual rings*—These are layers of wood.
- *Pith*—This is at the very center.

Around the pith, the wood substance is arranged in approximately concentric rings (see Figure 2-1). The part nearest the pith is usually darker than the parts nearest the bark and is called the *heartwood*. The cells in the heartwood are dead. Nearer the bark is the *sapwood*, where the cells carry or store nutrients but are not living (see Figure 2-2).

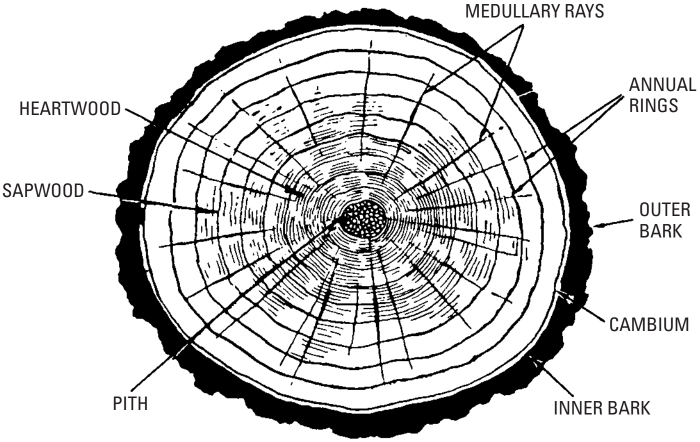

Figure 2-1 Cross-section of an oak nine years old, shows pith, concentric rings comprising the woody part, the cambium layer, and the bark. The tree grows in concentric rings (or layers) with one layer added each year. The rings are also called annual rings.

As winter approaches, all growth ceases. Each annual ring is separate and, in most cases, distinct. The leaves of deciduous trees (or trees that shed their leaves) and the leaves of some of the conifers (such as cypress and larch) fall, and the sap in the tree may freeze hard. The tree is dormant but not dead. With the warm days of the next spring, growth starts again strongly, and the cycle is repeated. The width of the annual rings varies greatly, from 30 to 40 or more per inch in some slow-growing species, to as few as 3 or 4 per inch in some of the quick-growing softwoods.

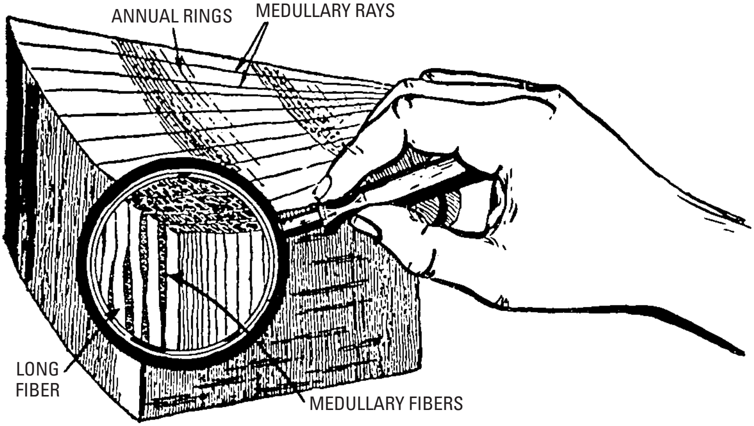

Figure 2-2 A piece of wood magnified to show its structure. The wood is made up of long, slender cells called fibers, which usually lie parallel to the pith. The length of these cells is often 100 times their diameter. Transversely, bands of other cells, elongated but much shorter, serve to carry sap and nutrients across the trunk radially. In addition, in the hardwoods, long vessels or tubes (often several feet long) carry liquids up the tree. There are no sap-carrying vessels in the softwoods, but spaces between the cells may be filled with resins.

Lumber Conversion

When logs are taken to the mill, they may be cut in a variety of ways. One way of cutting is *quartersawing* (see Figure 2-3). Here, each log is ripped into quarters, as shown in the figure. Quartersawing is rarely done this way, though, because only a few wide boards are yielded. There is too much waste. More often, wood is *rift-sawed*. The log is started as shown in Figure 2-4 and is plain-sawed until a good figure (pattern) shows, then turned over and rift-sawed. This way, there is less waste, and the boards are wide.

The result is *vertical-grain*, or *rift-grain* lumber. Vertical-grain lumber shrinks less in width. It is often used in door stiles and rails because it is less likely to warp.

The plain sawed stock is simply flat-sawn out of the log (see Figure 2-5). This results in *flat-grain* or *side-grain* lumber that is used where shrinkage and warping are less critical.

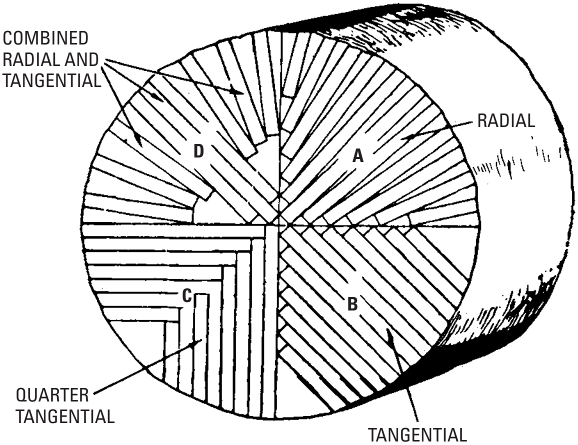
COMBINED
RADIAL AND
TANGENTIAL
RADIAL
D
A
C
B
QUARTER
TANGENTIAL
TANGENTIAL

Figure 2-3 Methods of quartersawing. These are rarely used because waste is extensive.

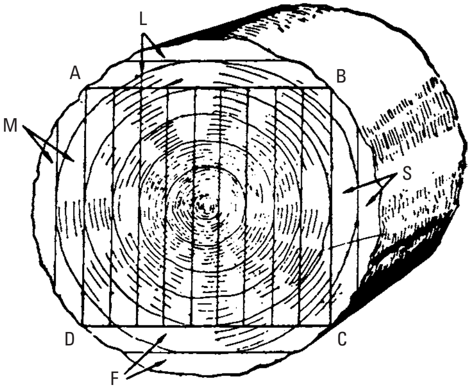

Figure 2-4 Plain or bastard sawing, sometimes called flat or slash sawing. The log is first squared by removing boards M, S, L, and F, giving the rectangular section ABCD. This is necessary to obtain a flat surface on the log.

Seasoning of Wood

Well-developed techniques have been established for removing the large amounts of moisture normally present in green wood. Seasoning is essentially a drying process, but for uses that require them, seasoning includes equalizing and conditioning treatments to improve

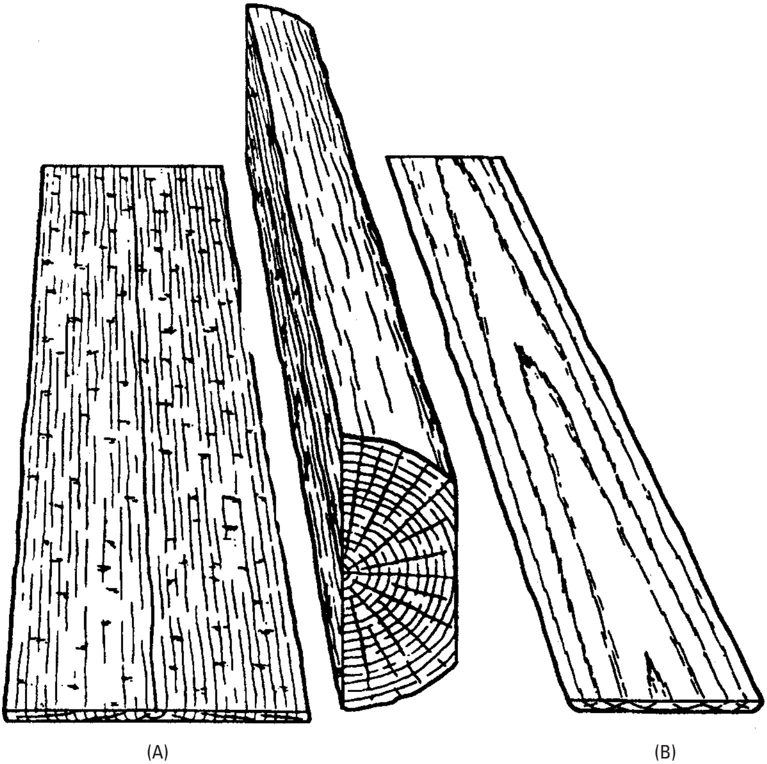

Figure 2-5 (A) Quartersawed and (B) plain-sawed boards cut from a log. *(Courtesy Forest Product Lab)*

moisture uniformity and relieve residual stresses and sets. Careful techniques are necessary, especially during the drying phase, to protect the wood from stain and decay and from excessive drying stresses that cause defects and degrade. The established seasoning methods are air-drying and kiln drying.

Drying has the following effects:

- Reduces the weight of wood, with a resulting decrease in shipping costs
- Reduces or eliminates shrinkage, checking, and warping in service
- Increases strength and nail holding power; decreases susceptibility to infection by blue stain and other fungi

- Reduces chance of attack by insects
- Improves the capacity of wood to take preservative and fire-retardant treatment and to hold paint

It is common practice at most softwood sawmills to kiln-dry all upper-grade lumber intended for finish, flooring, and cut stock. Lower-grade boards are often air-dried. Dimension lumber is air-dried or kiln-dried, although some mills ship certain species without seasoning. Timbers are generally not held long enough to be considered seasoned, but some drying may take place between sawing and shipment, or while they are held at a wholesale or distributing yard. Sawmills cutting hardwoods commonly classify the lumber for size and grade at the time of sawing. Some mills send all freshly sawed stock to the air-drying yard or an accelerated air-drying operation. Others kiln-dry directly from the green condition. Air-dried stock is kiln-dried at the sawmill, at a custom drying operation during transit, or at the remanufacturing plant before being made up into such finished products as furniture, cabinet work, interior finish, and flooring.

Air-drying is not a complete drying process, except as preparation for uses for which the recommended moisture content is not more than 5 percent below that of the air-dry stock. Even when air-drying conditions are mild, air-dry stock used without kiln-drying may have some residual stress and set that can cause distortions after nonuniform surfacing or machining. On the other hand, rapid air-drying accomplished by low relative humidities produces a large amount of set that will assist in reducing warp during final kiln drying. Rapid surface drying also greatly decreases the incidence of chemical and sticker stain, blue stain, and decay.

Air-drying is an economical seasoning method when carried out in a well-designed yard or shed, with proper piling practices (see Figure 2-6) and in favorable drying weather. In cold or humid weather, air-drying is slow and cannot readily reduce wood moisture to levels suitable for rapid kiln-drying or for use.

In kiln drying, higher temperatures and fast air circulation are used to increase the drying rate considerably. Average moisture content can be reduced to any desired value. Specific schedules are used to control the temperature and humidity in accordance with the moisture and stress situation within the wood, thus minimizing shrinkage-caused defects. For some purposes, equalizing and conditioning treatments are used to improve moisture content uniformity and relieve stresses and set at the end of drying, so the material will not warp when resawed or machined to smaller sizes or irregular

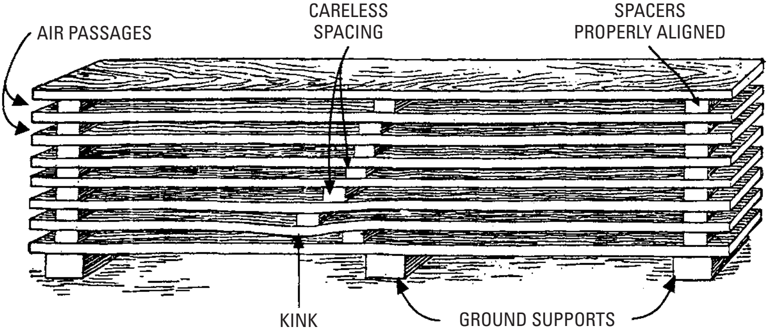

Figure 2-6 Horizontal stack of lumber for air-drying. One end of the pile should be a little higher. This is so that rainwater that falls on top or drives into the pile will drain. Each layer should be separated by three or four spacers so that the air will have free access to both sides of each board. The lowest layer should be well elevated above the ground to protect it from dampness. It takes from one to three years to thoroughly season lumber, depending upon the character of the wood, climatic conditions, and so on. When the spacers are carelessly placed so that they do not lie over each other, the weight must be supported by the board, which especially in the case of a high pile, is considerable, and will in time cause the board to sag, resulting in a permanent kink.

shapes. Further advantages of kiln drying are the setting of pitch in resinous woods, the killing of staining or decay fungi or insects in the wood, and reductions in weight greater than those achieved by air-drying. At the end of kiln-drying, moisture-monitoring equipment is sometimes used to sort out moist stock for redrying and to ensure that the material ready for shipment meets moisture content specifications.

Temperatures of ordinary kiln drying generally are between 110°F and 180°F. Elevated-temperature (180°F to 212°F) and high-temperature (above 212°F) kilns are becoming increasingly common, although some strength loss is possible with higher temperatures.

Moisture Content

While the tree is living, both the cells and cell walls are filled with water to an extent. As soon as the tree is cut, the water within the cells (or *free water* as it is called) begins to evaporate. This process continues until practically all of the free water has left the wood. When this stage is reached, the wood is said to be at the fiber-saturation point (that is, what water is contained is mainly in the cell or fiber walls).

Except in a few species, there is no change in size during this preliminary drying process, and, therefore, no shrinkage during the evaporation of the free water. Shrinkage begins only when water begins to leave the cell walls themselves. What causes shrinkage and other changes in wood is not fully understood. However, it is thought that as water leaves the cell walls, they contract, becoming harder and denser, thereby causing a general reduction in size of the piece of wood. If the specimen is placed in an oven that is maintained at 212°F (the temperature of boiling water), the water will evaporate and the specimen will continue to lose weight for a time. Finally, a point is reached at which the weight remains substantially constant. This is another way of saying that all of the water in the cells and cell walls has been driven off. The piece is then said to be *oven dry*.

If it is now taken out of the oven and allowed to remain in the open air, it will gradually take on weight, because of the absorption of moisture from the air. As when placed in the oven, a point is reached at which the weight of the wood in contact with the air remains more or less constant. Careful tests, however, show that it does not remain exactly constant—it will take on and give off water as the moisture in the atmosphere increases or decreases. Thus, a piece of wood will contain more water during the humid, moist summer months than in the colder, drier winter months. When the piece is in this condition, it is in equilibrium with the air and is said to be *air dry*.

The amount of water contained by wood in the green condition varies greatly, not only with the species but also in the same species and even in the same tree, according to the position in the tree. However, as a general average, at the fiber-saturation point, most woods contain from 23 to 30 percent water as compared with the oven-dry weight of the wood. When air dry, most woods contain from 12 to 15 percent moisture.

As the wood dries from the green state (which is that of the freshly cut tree) to the fiber-saturation point, except in a few species, there is no change other than that of weight. It has already been pointed out that as the moisture dries out of the cell walls, in addition to the decrease in weight, shrinkage results in a definite decrease in size (see Figure 2-7). It has been found, however, that there is little or no decrease along the length of the grain, and that the decrease is at right angles to the grain.

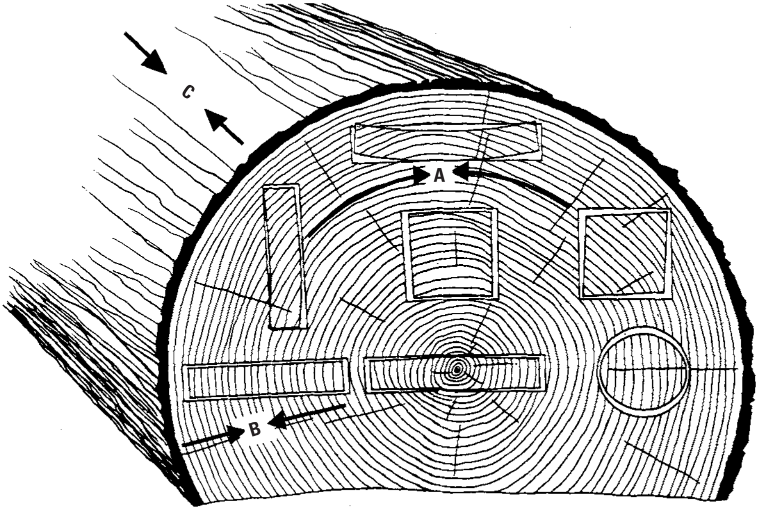

Figure 2-7 Greatest shrinkage is in the direction of the annual rings (A). Characteristic shrinkage and distortion of flat boards, squares, and rounds as affected by the direction of the annual rings. Tangential shrinkage (A) is about twice as great as radial (B). There is little or no longitudinal (C) shrinkage. *(Courtesy Forest Product Lab)*

This is an important consideration to be remembered when framing a building. For example, a stud in a wall will not shrink appreciably in length, whereas it will shrink somewhat in both the 2-inch and the 4-inch way. In like manner, a joist, if it is green when put in place, will change in depth as it seasons in the building. These principles of shrinkage also explain why an edge-grain or quarter-sawed floor is less likely to open up than a flat-grain floor.

Density

The tree undergoes a considerable impetus early every spring and grows very rapidly for a short time. Large amounts of water are carried through the cells to the rapidly growing branches and leaves at the top of the tree. This water passes upward mainly in the outer layers of the tree. The result is that the cells next to the bark (which are formed during the period of rapid growth) have thin walls and large passages. Later on, during the summer, the rate of growth slows and the demand for water is less. The cells that are formed during the summer have much thicker walls and much smaller pores. Thus, a year's growth forms two types of wood: the *springwood* (characterized by softness and openness of grain) and the *summerwood* (characterized by hardness and closeness, or density, of grain). The springwood and summer wood growth for one year is called an *annual ring*.

Density and Strength

There is one ring for each year of growth. This development of springwood and summerwood is a marked characteristic of practically all woods that grow in a temperate climate. It is evident in such trees as the yellow pines and firs and less so in the white pines, maple, and the like. Careful examination will reveal this annual ring, however, in practically all species. It follows, therefore, that a tree in which the dense summerwood predominates is stronger than one in which the soft spring wood predominates. This is a point that should be kept in mind when selecting material for important members such as girders and posts carrying heavy loads. The strength of wood of the same species varies markedly with the density. For example, Douglas fir or southern pine, carefully selected for density, is one-sixth stronger than lumber of the same species and knot limitations in which the springwood predominates. Trees having approximately one-third or more of cross-sectional area in summerwood fulfill one of the requirements for structural timbers.

Estimating Density

It must be remembered that the small cells or fibers that make up the wood structure are hollow. Wood substance itself has a specific gravity of about 1.5. That means it will sink in water. It is stated that wood substance of all species is practically of the same density. Strength of wood depends upon its density and varies with its density. The actual dry weight of lumber is a good criterion of its strength, although weight cannot always be relied upon as a basis for determining strength, because other important factors frequently must be considered in a specific piece of wood.

The hardness of wood is also another factor that assists in estimating the strength of wood. A test sometimes used is cutting across the grain. This test cannot be utilized in the commercial grading of lumber because moisture content will affect the hardness and because hardness, thus measured, cannot be adequately defined. The annual rings found in practically all species are an important consideration in estimating density, although the annual rings indicate different conditions in different species. In ring-porous hardwoods and in the conifers, where the contrast between springwood and summerwood is definite, the proportion of hard summerwood is an indication of the strength of the individual piece of wood. The amount of summerwood, however, cannot always be relied upon as an indication of strength because summerwood itself varies in density. When cut across the grain with a knife, the density of summerwood may be estimated based on hardness, color, and luster.

In conifers, annual rings of average width indicate denser material or a larger proportion of summerwood than in wood with either wide or narrow rings. In some old conifers of virgin growth in which the more recent annual rings are narrow, the wood is less dense than where there has been normal growth. On the other hand, in young trees where the growth has not been impeded by other trees, the rings are wider and in consequence the wood less dense. These facts may account for the belief that all second-growth timber and all sapwood are weak. In analyzing wood for density, the contrast between summerwood and springwood should be pronounced.

Oak, ash, hickory, and other ring-porous hardwoods in general rank high in strength when the annual rings are wide. In this respect, they contrast with conifers. These species have more summerwood than springwood as the rings become wider. For this reason, oak, hickory, ash, and elm of second growth are considered superior because of fast growth and increase in proportion of summerwood. These conditions do not always exist, however, because exceptions occur, especially in ash and oak, where, although the summerwood is about normal, it may not be dense or strong. Very narrow rings in ring-porous hardwoods are likely to indicate weak and brash material composed largely of spring wood with big pores. Maple, birch, beech, and other diffuse-porous hardwoods in general show no definite relationship between the width of rings and density, except that usually narrow rings indicate brash wood.

Strength

Wood, when used in ordinary structures, is called upon to have three types of strength: tension, compression, and shear.

Tension

Tension is the technical term for a pulling stress. For example, if two people are having a tug of war with a rope, the rope is in tension. The tensile strength of wood, especially of the structural grades, is very high.

Compression

If, however, the people at opposite ends of a 2 × 4 are trying to push each other over, the timber is in *compression*. Tension and compression represent, therefore, exactly opposite forces.

Shear

If two or three planks are placed one upon the other between two blocks, and a person were to stand in the middle, the planks would bend (see Figure 2-8). It will be noted that at the outer ends the boards tend to slip past each other.

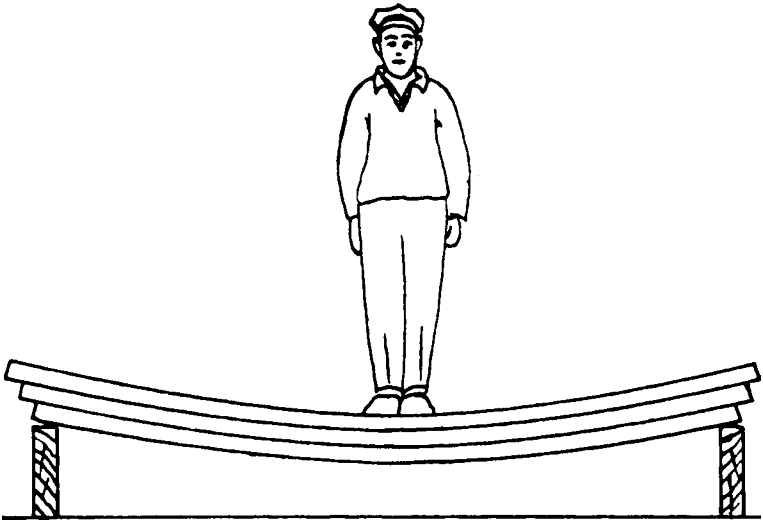

Figure 2-8 Illustrating shear in lumber.

If the planks were securely spiked through from top to bottom, the slipping would be in a great measure prevented and the boards would act more as one piece of wood. In very solid timber, there is the same tendency for the various parts of the piece to slip past each other. This tendency is called *horizontal shear*. A defect (such as a check, which runs horizontally through a piece of a timber and tends to separate the upper from the lower part) is a weakness in shear.

Suppose that the planks were spiked through at the center of span only (that is, halfway between the blocks). Such spikes would not increase the stiffness of the planks. It is clear, therefore, that there is no horizontal shear near the center of the span (see Figure 2-9), and that the shear increases as one approaches either end of the beam. This will explain why, as most carpenters have doubtlessly observed, steel stirrups are used in concrete beams (weak in shear), why there is usually none near the center, and why they are put closer and closer together near the ends of the beams.

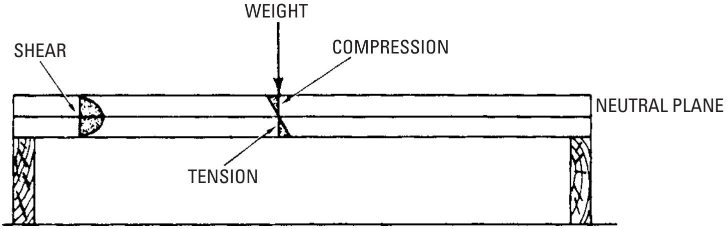

Figure 2-9 Illustrating the different proportions of tension, compression, and shear.

For all practical purposes, the compressive strength of wood may be considered to equal its tensile strength. It has been extremely difficult to make any direct measurements of the tensile strength of wood. In an experiment designed to ascertain the tensile strength of a specimen of wood, a 4-inch × 4-inch piece was selected. A portion about a foot in length near the center was carefully cut down on all four sides until it was exactly $3/4$-inch square. The test specimen was placed in a machine that gripped the 4-inch × 4-inch ends securely and a pull was exerted. The specimen did not pull apart. The $3/4$-inch-square section held and actually pulled out of the end of the 4-inch × 4-inch, leaving a $3/4$-inch square hole. This is an excellent illustration of how a piece may fail from shear rather than tension, the shear in this case being insufficient to prevent the $3/4$-inch-square piece from pulling out.

Deadwood

Because in some instances persons are prejudiced against the use of timber cut from dead trees, it is customary for individuals to specify that only timber cut from live trees will be accepted. It is true, however, that when sound trees that are dead are sawed into lumber and the weathered or charred outside is cut away, the resulting lumber cannot be distinguished from that coming from live trees except insofar as the lumber from dead trees may be somewhat seasoned at the time it is sawed. It must be remembered that the heartwood of a living tree is fully matured and that in the sapwood only a small portion of the cells are in a living condition. Consequently, most of the wood cut from trees is already dead, even when the tree itself is considered alive.

For structural purposes, it may be said that lumber cut from fire- or insect-killed trees is just as good as any other lumber, unless the wood has been subjected to further decay or insect attack.

Virgin and Second Growth

Occasionally an order calls for lumber of either virgin growth or second growth. The terms, however, are without significance, as an individual cannot tell one type from the other when it is delivered.

The *virgin growth* (which is also called *old growth* or *first growth*) refers to timber that grows in the forest along with many other trees, and, therefore, has suffered the consequence of the fight for sunlight and moisture.

The *second growth* is considered as that timber that grows up with less of the competition for sunlight and moisture that characterizes first-growth timber.

Because of environment, the virgin growth is usually thought of as wood of slow-growing type, whereas the second growth is considered as of relatively rapid growth, evidenced by wider annual rings. In such hardwoods as ash, hickory, elm, and oak, the wider annual rings are supposed to indicate stronger and tougher wood, whereas in the conifers such as pine and fir, this condition is supposed to result in a weaker and brasher wood. For this reason, where the strength and toughness are desired, the second growth is preferred among hardwoods, and virgin growth is desired in conifers. This is because of the following:

- The variety of conditions under which both virgin and second growth timbers grow
- Virgin growth may have the characteristics of second growth
- Second growth may have the characteristics of virgin timber

It is advisable in judging the strength of wood to rely upon its density and rate of growth, rather than upon its being either virgin or second growth.

Time of Cutting Timber

The time when timber is cut has very little to do with its durability or other desirable properties if, after it is cut, it is cared for properly. Timber cut in the late spring, however, or early summer is more likely to be attacked by insects and fungi. In addition, seasoning will proceed much more rapidly during the summer months and, therefore, will result in checking, unless the lumber is shaded from the intense sunlight. There is practically no difference in the moisture content in green lumber cut during the summer or winter.

Air-Dried and Kiln-Dried Wood

There is a prevailing misapprehension that air-dried lumber is stronger or better than kiln-dried lumber. Exhaustive tests have conclusively shown that good kiln-drying and good air-drying have exactly the same results on the strength of the wood. Wood increases in strength with the elimination of moisture content. This may account for the claim that kiln-dried lumber is stronger than air-dried lumber. This has little significance because, in use, wood will come to practically the same moisture content whether it has been kiln-dried or air-dried.

The same kiln-drying process cannot be applied to all species of wood. Consequently, it must be remembered that lack of certain strength properties in wood may be because of improper kiln-drying. Similar damage also may result from air seasoning under unsuitable conditions.

Sapwood Versus Heartwood

The belief is common that in some species the heartwood is stronger than the sapwood, and that the reverse is the case in such species as hickory and ash. Tests have shown conclusively that neither is the case, sapwood is not necessarily stronger than heartwood or heartwood stronger than sapwood, but that density (rather than other factors) makes the difference in strength. In trees that are mature, the sapwood is frequently weaker, whereas in young trees the sapwood may be stronger. Density, proportion of springwood and summerwood, then must be the basis of consideration of strength rather than whether the wood is sapwood or heartwood. Under unfavorable conditions, the sapwood of most species is more subject to decay than the heartwood.

The cells in the heartwood of some species are filled with various oils, tannins, and other substances (called *extratives*) that make these timbers rot-resistant. There is practically no difference in the strength of heartwood and sapwood if they weigh the same. In most species, only the sapwood can be readily impregnated with preservatives, a procedure used when the wood will be in contact with the ground.

Defects and Deterioration

The defects found in manufactured lumber have several causes. Those defects found in the natural log include the following:

- Shakes
- Knots
- Pitchpockets

Those caused by deterioration include the following:

- Rot
- Dote

Those caused by imperfect manufacture include the following:

- Imperfect machining
- Wane
- Machine burn
- Checks and splits from imperfect drying

Heart shakes (see Figure 2-10) are radial cracks that are wider at the pith of the tree than at the outer end. This defect is more commonly found in old trees rather than in young vigorous saplings. It occurs frequently in hemlock.

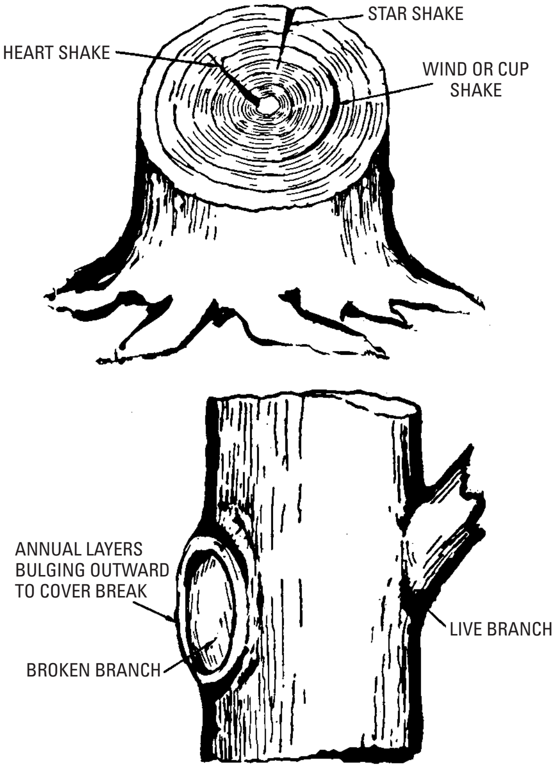
STAR SHAKE
HEART SHAKE
WIND OR CUP SHAKE
ANNUAL LAYERS BULGING OUTWARD TO COVER BREAK
LIVE BRANCH
BROKEN BRANCH

Figure 2-10 The causes of some lumber defects are found in the natural log.

A *wind shake* or *cup shake* is a crack following the line of the porous part of the annual rings and is curved by a separation of the annual rings (see Figure 2-10). A wind shake may extend for a considerable distance up the trunk. Other explanations for wind shakes are expansion of the sapwood and wrenching from high wind (hence the name). Brown ash is especially susceptible to wind shake. Wind shakes cause cup checks in lumber (see Figure 2-11A). A *star shake* resembles a wind shake but differs from it in that the crack extends across the center of the trunk without any appearance of decay at that point; it is larger at the outside of the tree. Heart and star shakes cause splits in lumber (see Figure 2-11B).

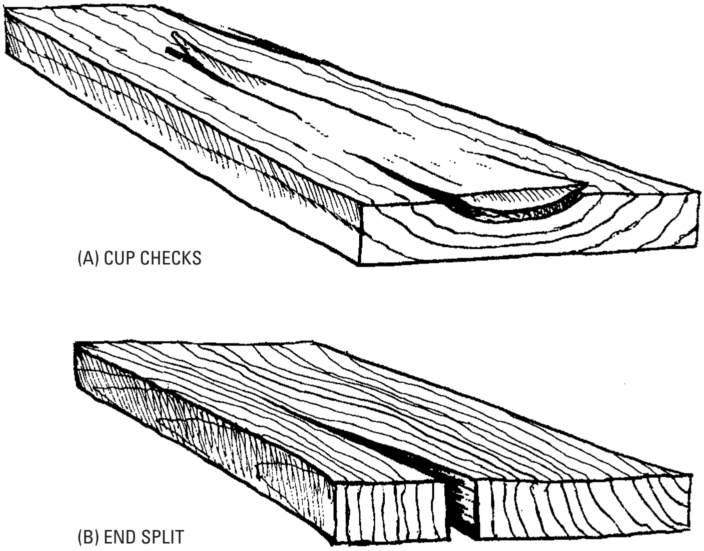

Figure 2-11 Lumber defects caused by defects in the log. (A) Cup checks are caused by wind shakes. (B) End splits are caused by star and heart shakes. *(Courtesy Practical Restoration Reports)*

Figure 2-12 shows some common causes of black or loose knots. Figures 2-13 and 2-14 show some results of improper drying practices.

Decay of Wood

Decay of lumber is the result of one cause and one cause only: the work of certain low-order plants called *fungi*. All of these organisms require water, air, and temperatures well above freezing to live, grow, and multiply. Consequently, wood that is kept dry, or that is dried quickly after wetting, will not decay.

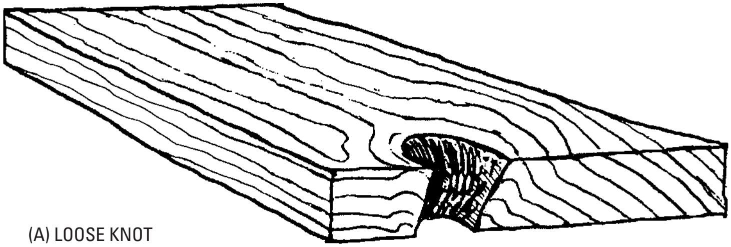

(A) LOOSE KNOT

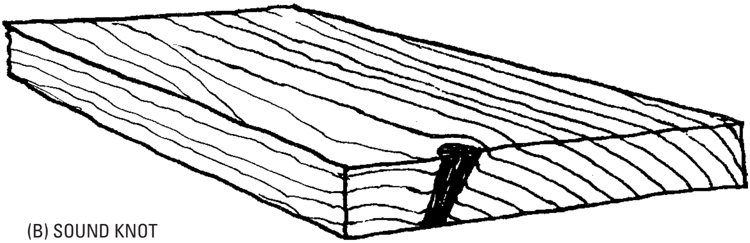

(B) SOUND KNOT

Figure 2-12 Black or loose knots (A) are caused by broken branches in the log. Bulging layers of growth cause a large swirl of cross grain at the surface of the board surrounding the knot. Tight red knots (B) are caused by a live branch in the log. The knot will not fall out and grain is relatively straight around the knot. *(Courtesy Practical Restoration Reports)*

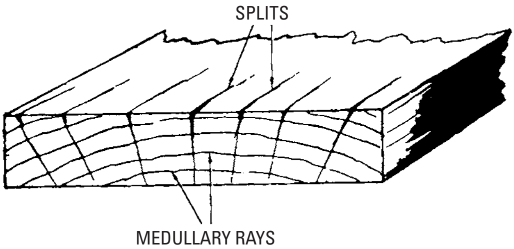

Figure 2-13 A board with splits along the medullary rays. This condition is caused by too-rapid kiln-drying.

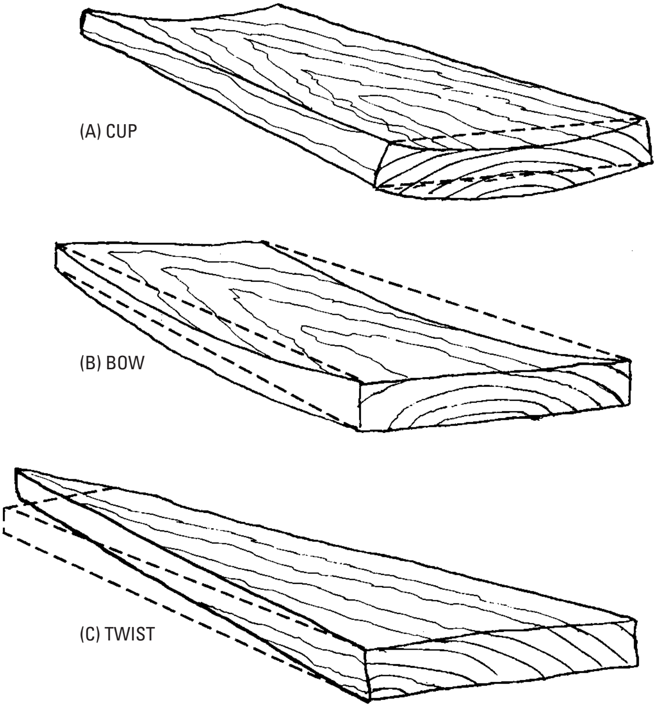

Figure 2-14 Warping is frequently caused by improper drying practices. Here the flatness of boards is distorted by irregular shrinkage. *(Courtesy Practical Restoration Reports)*

Further, if it is kept continuously submerged in water even for long periods, it is not decayed significantly by the common decay fungi regardless of the wood species or the presence of sapwood. Bacteria and certain soft-rot fungi can attack submerged wood, but the resulting deterioration is very slow. A large proportion of wood in use is kept so dry at all times that it lasts indefinitely. Moisture and temperature (which vary greatly with local conditions) are the principal factors affecting rate of decay. When exposed to conditions that favor decay, wood deteriorates more rapidly in warm, humid areas than in cool or dry areas. High altitudes, as a rule, are less favorable to decay than low altitudes because the average temperatures are lower and the growing seasons for fungi that cause decay are shorter.

The heartwoods of some common native species of wood have varying degrees of natural decay resistance. Untreated sapwood of substantially all species has low resistance to decay and usually has a short service life under decay-producing conditions. The decay resistance of heartwood is greatly affected by differences in the preservative qualities of the wood extractives, the attacking fungus, and the conditions of exposure. Considerable difference in service life may be obtained from pieces of wood cut from the same species, or even from the same tree, and used under apparently similar conditions. There are further complications because, in a few species, such as the spruces and the true firs (not Douglas fir), heartwood and sapwood are so similar in color that they cannot be easily distinguished. Marketable sizes of some species, such as southern pine and baldcypress, are becoming largely second growth. They contain a high percentage of sapwood.

Precise ratings of the decay resistance of heartwood of different species are not possible because of differences within species and the variety of service conditions to which wood is exposed. However, broad groupings of many of the native species, based on service records, laboratory tests, and general experience, are helpful in choosing heartwood for use under conditions favorable to decay (see Table 2-1). The extent of variations in decay resistance of individual trees or wood samples of a species is much greater for most of the more resistant species than for the slightly or nonresistant species.

Where decay hazards exist, heartwood of species in the resistant or very resistant category generally gives satisfactory service. However, heartwood of species in the other two categories will usually require some form of preservative treatment. For mild decay conditions, a simple preservative treatment (such as a short soak in preservative after all cutting and boring operations are complete) will be adequate for wood low in decay resistance. For more severe decay hazards, pressure treatments are often required. Even the very decay-resistant species may require preservative treatment for important structural or other uses where failure would endanger life or require expensive repairs.

Wood products sometimes are treated with preservative or fire-retarding salts (usually in water solution) to impart resistance to decay or fire. Such products generally are kiln-dried after treatment. Mechanical properties are essentially unchanged by preservative treatment.

Properties are affected to some extent by the combined effects of fire-retardant chemicals, treatment methods, and kiln-drying. A

Table 2-1 Grouping of Some Domestic Woods According to Heartwood Decay

Resistant or Very Resistant	*Moderately Resistant*	*Slightly or Nonresistant*
Baldcypress (old growth)*	Baldcypress (young growth)*	Alder
Catalpa	Douglas fir	Ashes
Cedars	Honey locust	Aspens
Cherry, black	Larch, western	Basswood
Chestnut	Oak, swamp chestnut	Beech
Cypress, Arizona	Pine, eastern white*	Birches
Junipers	Southern pine:	Buckeye
Locust, black†	Longleaf*	Butternut
Mesquite	Slash*	Cottonwood
Mulberry, red†	Tamarack	Elms
Oak:		Hackberry
Bur		Hemlocks
Chestnut		Hickories
Gambel		Magnolia
Oregon white		Maples
Post		Oak (red and black species)
White		Pines (other than long-leaf, slash, and eastern white)
Osage orange†		Poplars
Redwood		Spruces
Sassafras		Sweetgum
Walnut, black		True firs (western and eastern)
Yew, Pacific†		Willows
		Yellow poplar

*The southern and eastern pines and baldcypress are now largely second growth with a large proportion of sapwood. Consequently, substantial quantities of heartwood lumber of these species are not available.

†These woods have exceptionally high decay resistance.

variety of fire-retardant treatments have been studied. Collectively the studies indicate modulus of rupture, work to maximum load, and toughness are reduced by varying amounts depending on species and type of fire retardant. Work to maximum load and toughness are most affected, with reductions of as much as 45 percent. A reduction in modulus of rupture of as much as 20 percent has been observed. A design reduction of 10 percent is frequently used. Stiffness is not appreciably affected by fire-retardant treatments.

Wood is also sometimes impregnated with monomers (such as methyl methacrylate) that are subsequently polymerized. Many of the properties of the resulting composite are better. Generally, this is a result of filling the void spaces in the wood structure with plastic. The polymerization process and both the chemical nature and quantity of monomers are variables that influence composite properties.

Molding and Staining Fungi

Molding and *staining fungi* do not seriously affect most mechanical properties of wood because they feed upon substance within the structural cell wall rather than on the structural wall itself. Specific gravity may be reduced by from 1 to 2 percent, whereas most of the strength properties are reduced by a comparable or only slightly greater extent. Toughness or shock resistance, however, may be reduced by up to 30 percent. The duration of infection and the species of fungi involved are important factors in determining the extent of weakening.

Although molds and stains themselves often do not have a major effect on the strength of wood products, conditions that favor the development of these organisms are likewise ideal for the growth of wood-destroying (decay) fungi, which can greatly reduce mechanical properties.

Fungal Decay

Unlike the molding and staining fungi, the *wood-destroying (decay) fungi* seriously reduce strength. Even what appears to be sound wood adjacent to obviously decayed parts may contain hard-to-detect, early (incipient) decay that is decidedly weakening, especially in shock resistance.

Not all wood-destroying fungi affect wood in the same way. The fungi that cause an easily recognized pitting of the wood, for example, may be less injurious to strength than those that, in the early stages, give a slight discoloration of the wood as the only visible effect.

No method is known for estimating the amount of reduction in strength from the appearance of decayed wood. Therefore, when strength is an important consideration, the safe procedure is to discard every piece that contains even a small amount of decay. An exception may be pieces in which decay occurs in a knot but does not extend into the surrounding wood.

Blue Stain

In the sapwood of many species of both softwoods and hardwoods, there often develops a bluish-black discoloration known as *blue stain*. It does not indicate an early stage of decay, nor does it have any practicable effect on the strength of the wood. Blue stain is caused by a fungus growth in unseasoned lumber. Although objectionable where appearance is of importance, as in unpainted trim, blue stain need cause no concern for framing lumber. Precautions should be taken, however, to make sure that no decay fungus is present with the blue stain.

Nuclear Radiation

Very large doses of gamma rays or neutrons can cause substantial degradation of wood. In general, irradiation with gamma rays in doses up to about 1 megarad has little effect on the strength properties of wood. As dosage increases above 1 megarad, tensile strength parallel to grain and toughness decrease. At a dosage of 300 megarads, tensile strength is reduced about 90 percent. Gamma rays also affect compressive strength parallel to grain above 1 megarad, but strength losses with further dosage are less than for tensile strength. Only about one-third of the compressive strength is lost when the total dose is 300 megarads. Effects of gamma rays on bending and shear strength are intermediate between the effects on tensile and compressive strength.

Weathering

Without protective treatment, freshly cut wood exposed to the weather changes materially in color. Other changes due to weathering include warping, loss of some surface fibers, and surface roughening and checking. The effects of weathering on wood may be desirable or undesirable, depending on the requirements for the particular wood product. The time required to reach the fully weathered appearance depends on the severity of the exposure to sun and rain. Once weathered, wood remains nearly unaltered in appearance.

The color of wood is affected very soon on exposure to weather. With continued exposure, all woods turn gray. However, only the wood at or near the exposed surfaces is noticeably affected. This very thin gray layer is composed chiefly of partially degraded cellulose fibers and microorganisms. Further weathering causes fibers to be lost from the surface, but the process is so slow that only about $^1/_4$ inch is lost in a century.

In the weathering process, chemical degradation is influenced greatly by the wavelength of light. The most severe effects are produced by exposure to ultraviolet light. As cycles of wetting and drying take place, most woods develop physical changes, such as checks or cracks that are easily visible. Moderate to low-density woods acquire fewer checks than do high-density woods. Vertical-grain boards check less than flat-grain boards.

Because of weathering, boards tend to warp (particularly cup) and pull out their fastenings. The cupping tendency varies with the density, width, and thickness of a board. The greater the density and the greater the width in proportion to the thickness, the greater is the tendency to cup. Warping also is more pronounced in flat-grain boards than in vertical-grain boards (see Figure 2-15). For best cup resistance, the width of a board should not exceed eight times its thickness.

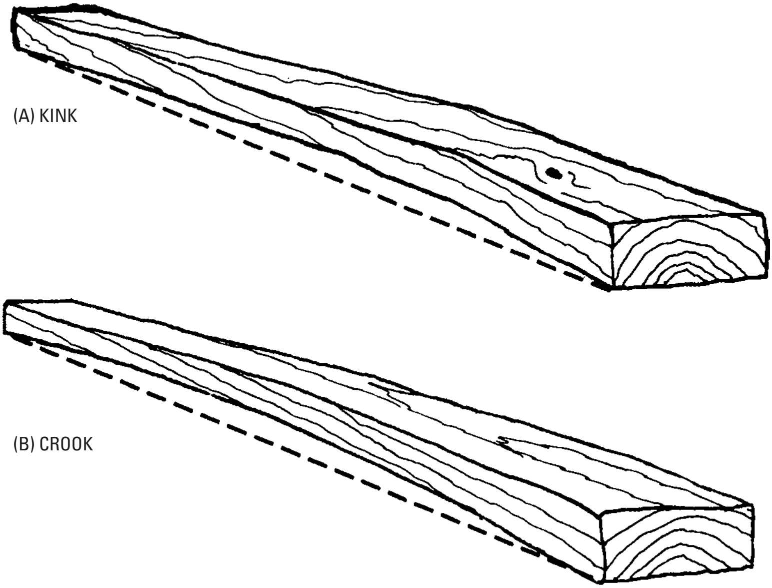

Figure 2-15 Warping can also affect straightness of a board.
(Courtesy Practical Restoration Reports)

Biological attack of a wood surface by microorganisms is recognized as a contributing factor to color changes. When weathered wood has an unsightly dark gray and blotchy appearance, it is caused by dark-colored fungal spores and mycelium on the wood surface. The formation of a clean, light gray, silvery sheen on weathered wood occurs most frequently where microorganism growth is inhibited by a hot, arid climate or where there is a salt atmosphere, such as in coastal regions.

The contact of fasteners and other metallic products with the weathering wood surface is a source of color, often undesirable if a natural color is desired.

Insect Damage

Insect damage may occur in standing trees, logs, and unseasoned or seasoned lumber. Damage in the standing tree is difficult to control, but otherwise insect damage can be largely eliminated by proper control methods.

Insect holes are generally classified as *pinholes*, *grub holes*, and *powder postholes*. The powder post larvae, by their irregular burrows, may destroy most of the interior of a piece, although the surface shows only small holes, and the strength of the piece may be reduced virtually to zero.

No method is known for estimating the amount of reduction in strength from the appearance of insect-damaged wood, and, when strength is an important consideration, the safe procedure is to eliminate pieces containing insect holes.

Summary

Wood is the most versatile building material. Softwoods and hardwoods grow as trees with a fibrous cellular structure. Logs are converted into lumber and seasoned by drying. The physical characteristics of wood affect its performance and specific uses. Defects in lumber relate to defects in the logs it was cut from and how the lumber was handled during drying and storage. After wood is installed in a building, it can deteriorate by weathering, decay, or insects unless protected.

Review Questions

1. Describe the physical structure of wood with a drawing. Label the parts.
2. What are the characteristics of softwoods and hardwoods? How is each type of wood used in building?

3. What is the difference between vertical-grain wood and flat-grain wood? How is each produced?
4. Why does wood have to be seasoned? How does the moisture content and size of wood change during seasoning?
5. How does shear affect the strength of wood?
6. What substances make some woods decay-resistant?
7. How do defects in logs relate to defects in lumber? Give some examples.
8. What causes decay in wood? How can you stop decay?
9. How does weathering affect the color of wood?
10. How long would it take for a 1-inch thick board to weather away by fiber loss?
11. Would vertical-grain or flat-grain boards hold up better on an outdoor deck? Why?

Chapter 3

Lumber, Plywood, and Other Wood Products

The basic construction material in carpentry is lumber. There are many kinds of lumber varying greatly in structural characteristics. Here, we deal with the lumber common to construction carpentry (see Figure 3-1).

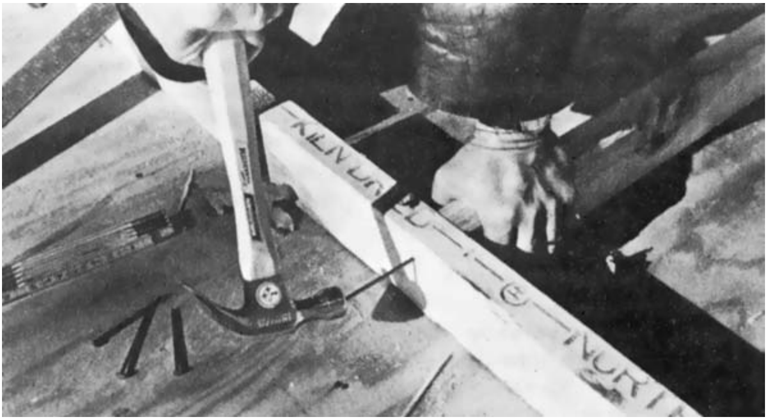

Figure 3-1 Construction lumber. Note the classification kiln-dried on one member. This stock will be used inside a structure.

(Courtesy of Vaughn & Bushnell)

Standard Sizes of Lumber

Lumber is usually sawed into standard lengths, widths, and thicknesses. This permits uniformity in planning structures and in ordering material (see Table 3-1). Standards have been established for dimension differences between *nominal size* and the *standard size*. It is important that these dimension differences be taken into consideration when planning a structure. A good example of the dimension difference may be illustrated by the common 2 × 4. As may be seen in the table, the familiar quoted size (2 × 4) refers to a rough or nominal dimension, but the actual standard size to which the lumber is dressed is $1^1/_2$ inches × $3^1/_2$ inches.

Table 3-1 Your Guide to New Sizes of Lumber

Standard (What You Order)	*Nominal (What You Get)*	
	*Dry or Seasoned**	*Green or Unseasoned†*
1 × 4	$\frac{3}{4} \times 3\frac{1}{2}$	$\frac{25}{32} \times 3\frac{9}{16}$
1 × 6	$\frac{3}{4} \times 5\frac{1}{2}$	$\frac{25}{32} \times 5\frac{5}{8}$
1 × 8	$\frac{3}{4} \times 7\frac{1}{4}$	$\frac{25}{32} \times 7\frac{1}{2}$
1 × 10	$\frac{3}{4} \times 9\frac{1}{4}$	$\frac{25}{32} \times 9\frac{1}{2}$
1 × 12	$\frac{3}{4} \times 11\frac{1}{4}$	$\frac{25}{32} \times 11\frac{1}{2}$
2 × 4	$1\frac{1}{2} \times 3\frac{1}{2}$	$1\frac{9}{16} \times 3\frac{9}{16}$
2 × 6	$1\frac{1}{2} \times 5\frac{1}{2}$	$1\frac{9}{16} \times 5\frac{5}{8}$
2 × 8	$1\frac{1}{2} \times 7\frac{1}{4}$	$1\frac{9}{16} \times 7\frac{1}{2}$
2 × 10	$1\frac{1}{2} \times 9\frac{1}{4}$	$1\frac{9}{16} \times 9\frac{1}{2}$
2 × 12	$1\frac{1}{2} \times 11\frac{1}{4}$	$1\frac{9}{16} \times 11\frac{1}{2}$
4 × 4	$3\frac{1}{2} \times 3\frac{1}{2}$	$3\frac{9}{16} \times 3\frac{9}{16}$
4 × 6	$3\frac{1}{2} \times 5\frac{1}{2}$	$3\frac{9}{16} \times 5\frac{5}{8}$
4 × 8	$3\frac{1}{2} \times 7\frac{1}{4}$	$3\frac{9}{16} \times 7\frac{1}{2}$
4 × 10	$3\frac{1}{2} \times 9\frac{1}{4}$	$3\frac{9}{16} \times 9\frac{1}{2}$
4 × 12	$3\frac{1}{2} \times 11\frac{1}{4}$	$3\frac{9}{16} \times 11\frac{1}{2}$

*19 percent moisture content or under.
†More than 19 percent moisture content.

Softwood Lumber Grades for Construction

The grading requirements of construction lumber are related specifically to the major construction uses intended, and little or no further grading occurs once the piece leaves the sawmill. Construction lumber can be placed in three general categories—stress-graded, nonstress-graded, and appearance lumber. *Stress-graded* and *nonstress-graded lumber* are employed where the structural integrity of the piece is the primary requirement. *Appearance lumber* encompasses those lumber products in which appearance is of primary importance (structural integrity, although sometimes important, is a secondary feature).

Stress-Graded Lumber

Almost all softwood lumber nominally 2 to 4 inches thick is stress-graded. Lumber of any species and size, as it is sawed from the log, is quite variable in its mechanical properties. Pieces may differ in strength by several hundred percent. For simplicity and economy in use, pieces of lumber of similar mechanical properties can be placed in a single class called a stress-grade.

Visual grading is the oldest stress-grading method. It is based on the premise that mechanical properties of lumber differ from mechanical properties of clear wood because of characteristics that can be seen and judged by eye. These visual characteristics are used to sort the lumber into stress grades (see Table 3-2). The following are major visual sorting criteria:

- Density
- Decay
- Heartwood and sapwood
- Slope of grain
- Knots
- Shake
- Checks and splits
- Wane
- Pitch pockets

Nonstress-Graded Lumber

Traditionally, much of the lumber intended for general building purposes with little or no remanufacture has not been assigned allowable properties (stress-graded). This category of lumber has been referred to as *yard lumber*. However, the assignment of allowable

Table 3-2 Visual Grades Described in the National Grading Rule*

Lumber Classification	*Grade Name*
Light framing (2 to 4 inches thick, 4 inches wide)†	Construction Standard Utility
Structural light framing (2 to 4 inches thick, 2 to 4 inches wide)	Select structural, 1, 2, 3
Studs (2 to 4 inches thick, 2 to 4 inches wide)	Stud
Structural joists and planks (2 to 4 inches thick, 6 inches and wide)	Select structural, 1, 2, 3
Appearance framing (2 to 4 inches thick, 2 to 4 inches wide)	Appearance

*Sizes shown are nominal.

†Widths narrower than 4 inches may have different strength ratio.

properties to an increasing number of former yard items has diluted the meaning of the term yard lumber.

In nonstress-graded structural lumber, the section properties (shape, size) of the pieces combine with the visual grade requirements to provide the degree of structural integrity intended. Typical nonstress-graded items include boards, lath, battens, crossarms, planks, and foundation stock.

Boards (sometimes referred to as *commons*) are one of the more important nonstress-graded products. Common grades of boards are suitable for construction and general utility purposes. They are separated into three to five different grades depending upon the species and lumber manufacturing association involved. Grades may be described by number (No. 1, No. 2) or by descriptive terms (Construction, Standard).

Since there are differences in the inherent properties of the various species and in corresponding names, the grades for different species are not always interchangeable in use. First-grade boards are usually graded primarily for serviceability, but appearance is also considered. This grade is used for such purposes as siding, cornice, shelving, and paneling. Features such as knots and knotholes are permitted to be larger and more frequent as the grade level becomes lower. Second- and third-grade boards are often used together for such purposes as subfloors, roof and wall sheathing, and rough concrete work. Fourth-grade boards are not selected for appearance but for adequate strength. They are used for roof and wall sheathing, subfloor, and rough concrete formwork.

Grading provisions for other nonstress-graded products vary by species, product, and grading association. Lath, for example, is available generally in two grades, No. 1 and No. 2. One grade of batten is listed in one grade rule and six in another.

Appearance Lumber

Appearance lumber often is nonstress-graded. However, it forms a separate category because of the distinct importance of appearance in the grading process. This category of construction lumber includes most lumber worked to a pattern. Secondary manufacture on these items is usually restricted to onsite fitting, such as cutting to length and mitering. There is an increasing trend toward prefinishing many items. The appearance category of lumber includes trim, siding, flooring, ceiling, paneling, casing, base, stepping, and finish boards. Finish boards are commonly used for shelving and built-in cabinetwork.

Most appearance lumber grades are described by letters and combinations of letters (B&BTR, C&BTR, D), as shown in Table 3-3. Appearance grades are also often known as Select grades. Descriptive terms such as prime and clear are applied to a limited number of species. The specification FG (flat grain), VG (vertical grain), or MG (mixed grain) is offered as a purchase option for some appearance lumber products. In cedar and redwood (where there is a pronounced difference in color between heartwood and sapwood, and where heartwood has high natural resistance to decay), grades of heartwood are denoted as heart. In some species and products, two (or, at most, three) grades are available. A typical example is casing and base in the grades of C&BTR and D in some species and in B&BTR, C, C&BTR, and D in other species. Although several grades may be described in grade rules, often fewer are offered on the retail market.

Table 3-3 Standard Lumber Abbreviations

Abbreviation	*Meaning*
AD	Air-dried
ALS	American Lumber Standard
AST	Antistain treated. At ship tackle (western softwoods)
AV or avg	Average
AW&L	All widths and lengths
B1S	See EB1S, CB1S, and E&CB1S
B2S	See EB2S, CB2S, and E&CB2S
B&B, B&BTR	B and Better
B&S	Beams and stringers
BD	Board
BD FT	Board feet
BDL	Bundle
BEV	Bevel or beveled
BH	Boxed heart
BM	Board measure
BSND	Bright sapwood no defect
BTR	Better
c	Allowable stress in compression in pounds per square inch
CB	Center beaded

(continued)

Table 3-3 *(continued)*

Abbreviation	*Meaning*
CB1S	Center bead on one side
CB2S	Center bead on two sides
cft or cu. ft.	Cubic foot or feet
CG2E	Center groove on two edges
CLG	Ceiling
CLR	Clear
CM	Center matched
Com	Common
CSG	Casing
CV	Center V
CV1S	Center V on one side
CV2S	Center V on two sides
DBClg	Double beaded ceiling (E&CB1S)
DB Part	Double beaded partition (E&CB2S)
DET	Double end trimmed
DF	Douglas-fir
DIM	Dimension
DKG	Decking
D/S,DS, D/Sdg	Drop siding
D1S,D2S	See S1S and S2S
D&M	Dressed and matched
D&CM	Dressed and center matched
D&SM	Dressed and standard matched
D2S&CM	Dressed two sides and center matched
D2S&SM	Dressed two sides and standard matched
E	Edge
EB1S	Edge bead one side
EB2S, SB2S	Edge bead on two sides
EE	Eased edges
EG	Edge (vertical or rift) grain
EM	End matched
EV1S, SV1S	Edge V one side
EV2S, SV2S.	Edge V two sides
E&CB1S	Edge and center bead one side
E&CB2S, DB2S, BC&2S	Edge and center bead two sides

(continued)

Table 3-3 *(continued)*

Abbreviation	***Meaning***
E&CV1S, DV1S, V&CV1S	Edge and center V one side
E&CV2S, DV2S, V&CV2S	Edge and center V two sides
f	Allowable stress in bending in pounds per square inch
FA	Facial area
FAS	Firsts and Seconds
FBM, Ft. BM	Feet board measure
FG	Flat or slash grain
FJ	Finger joint. End-jointed lumber using a finger joint configuration
FLG, Fig	Flooring
FOHC	Free of heart center
FOK	Free of knots
FT, ft	Foot or feet
FT.SM	Feet surface measure
G	Girth
GM	Grade marked
G/R	Grooved roofing
HB,H.B.	Hollow back
HEM	Hemlock
Hrt	Heart
H&M	Hit and miss
H or M	Hit or miss
IN, in.	Inch or inches
J&P	Joists and planks
JTD	Jointed
KD	Kiln-dried
LBR, Lbr	Lumber
LGR	Longer
LGTH	Length
Ltt, Lt	Lineal foot or feet
UN, Un	Lineal
LL	Longleaf
LNG, Lng	Lining
M	Thousand

(continued)

Table 3-3 ***(continued)***

Abbreviation	***Meaning***
MBM, MBF, M.BM	Thousand (feet) board measure
MC, M.C.	Moisture content
MG	Medium grain or mixed grain
MLDG, Mldg	Molding
Mft	Thousand feet
MSR	Machine stress rated
N	Nosed
NBM	Net board measure
No.	Number
N1E or N2E	Nosed one or two edges
Ord	Order
PAD	Partially air dry
PART, Part	Partition
PAT, Pat	Pattern
Pes.	Pieces
PE	Plain end
PET	Precision end trimmed
P&T	Posts and timbers
P1S,P2S	See S1S and S2S
RDM	Random
REG, Reg	Regular
Rfg.	Roofing
RGH, Rgh	Rough
R/L, RL	Random lengths
R/W,RW	Random widths
RES	Resawn
SB1S	Single bead one side
SDG, Sdg	Siding
S-DRY	Surfaced dry. Lumber 19 percent moisture content or less per American Lumber Standard for softwood
SE	Square edge
SEL, Sel	Select or select grade
SE&S	Square edge and sound
SG	Slash or flat grain

(continued)

Table 3-3 (*continued*)

Abbreviation	***Meaning***
S-GRN	Surfaced green. Lumber unseasoned, in excess of 19 percent moisture content per American Lumber Standard for softwood
SGSSND	Sapwood, gum spots and streaks, no defect
SIT. SPR	Sitka spruce
S/L,SL, S/Lap	Shiplap
STD.M	Standard matched
SM	Surface measure
Specs	Specifications
SO	Square
SORS	Squares
SR	Stress-rated
STD, Std	Standard
Std. lgths.	Standard lengths
SSND	Sap stain no defect (stained)
STK	Stock
STPG	Stepping
STR, STRUCT	Structural
SYP	Southern yellow pine
S&E	Side and edge (surfaced on)
S1E	Surfaced one edge
S2E	Surfaced two edges
S1S	Surfaced one side
S2S	Surfaced two sides
S4S	Surfaced four sides
S1S&CM	Surfaced one side and center matched
S2S&CM	Surfaced two sides and center matched
S4S&CS	Surfaced four sides and calking seam
S1S1E	Surfaced one side, one edge
S1S2E	Surfaced one side, two edges
S2S1E	Surfaced two sides, one edge
S2S&SL	Surfaced two sides and shiplapped
S2S&SM	Surfaced two sides and standard matched
t	Allowable stress in tension in pounds per square inch
TBR	Timber
T&G	Tongued and grooved

(continued)

Table 3-3 (*continued*)

Abbreviation	***Meaning***
VG	Vertical (edge) grain
V1S	See EV1S, CV1S, and E&CV1S
V2S	See EV2S, CV2S, and E&CV2S
WCH	West Coast hemlock
WDR, wdr	Wider
WHAD	Worm holes a defect
WHND	Worm holes no defect
WT	Weight
WTH	Width
WRD	Western redcedar
yp	Yellow pine

Courtesy Forest Products Laboratory

Grade B&BTR allows a few small imperfections, mainly in the form of minor skips in manufacture, small checks or stains caused by seasoning, and (depending on the species) small pitch areas, pin knots, or the like. Since appearance grades emphasize the quality of one face, the reverse side may be lower in quality. In construction, grade C&BTR is the grade combination most commonly available. It is used for high-quality interior and exterior trim, paneling, and cabinetwork, especially where these are to receive a natural finish. It is the principal grade used for flooring in homes, offices, and public buildings. In industrial uses, it meets the special requirements for large-sized, practically clear stock.

The number and size of imperfections permitted increase as the grades drop from B&BTR to D and E. Appearance grades are not uniform across species and products, however, and official grade rules must be used for detailed reference. Grade C is used for many of the same purposes as B&BTR, often where the best paint finish is desired. Grade D allows larger and more numerous surface imperfections that do not detract from the appearance of the finish when painted. Grade D is used in finish construction for many of the same uses as Grade C. It is also adaptable to industrial uses requiring short-length clear lumber.

Select Lumber

Select lumber is of good appearance and finished or dressed. See Table 3-4 for grade names, descriptions, and uses.

Table 3-4 Lumber Grades

Grade	*Description*	*Use*
Select:		
A	High quality, practically clear.	Suitable for natural finishes.
B	High quality, generally clear with a few minor defects.	Suitable for natural finishes.
C	Several minor defects.	Adapted to high-quality paint finish.
D	A few major defects.	Suitable for paint finishes.
Common:		
No. 1	Sound and tight-knotted.	Use without waste.
No. 2	Less restricted in quality than No. 1.	Framing, sheathing, structural forms where strain or stress not excessive.
No. 3	Permits some waste with defects larger than in No. 2	Footings, guardrails, rough subflooring.
No. 4	Permits waste, low quality, with decay and holes.	Sheathing, subfloors, roof boards in the cheaper types of construction.

Common Lumber

Common lumber is suitable for general construction and utility purposes and is identified by the grade names shown in Table 3-4.

Plywood

Plywood is a glued wood panel made up of relatively thin layers, or plies, with the grain of adjacent layers at an angle, usually 90° (see Figure 3-2). The usual constructions have an odd number of plies. The outside plies are called *faces* (or *face* and *back* plies), the inner plies are called *cores* (or *centers)*, and the plies immediately below the face and back are called *crossbands*. The core may be veneer, lumber, or particleboard. The plies may vary as to number, thickness, species, and grade of wood.

As compared with solid wood, the chief advantages of plywood are its having properties along the length nearly equal to properties along the width of the panel, its greater resistance to splitting, and its form, which permits many useful applications where large sheets

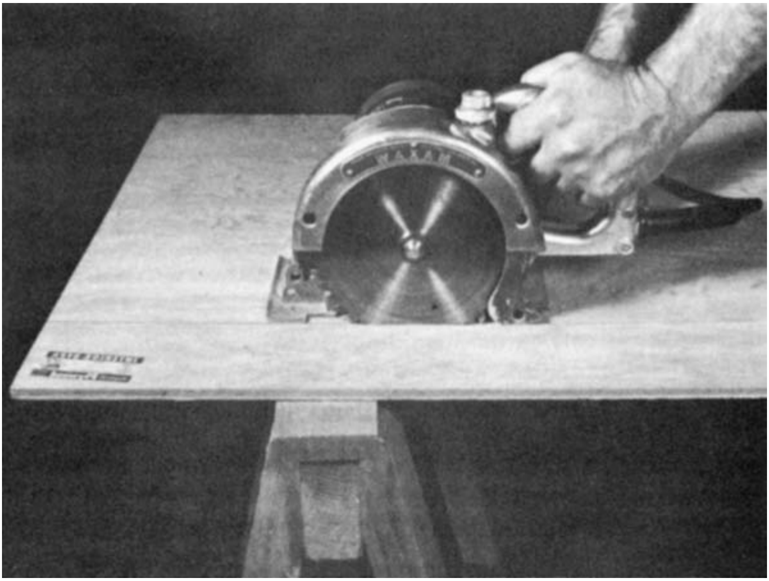

Figure 3-2 The multilayer construction of plywood gives large thin panels greater strength and stability than solid wood. *(Courtesy of the American Plywood Assn.)*

are desirable. Use of plywood may result in improved utilization of wood, because it covers large areas with a minimum amount of wood fiber. This is because it is permissible to use plywood thinner than sawn lumber in some applications.

The properties of plywood depend on the quality of the different layers of veneer, the order of layer placement in the panel, the glue used, and the control of gluing conditions in the gluing process. The grade of the panel depends on the quality of the veneers used, particularly of the face and back. The type of the panel depends on the glue joint, particularly its water resistance. Generally, face veneers with figured grain that are used in panels where appearance is important have numerous short, or otherwise deformed, wood fibers. These may significantly reduce strength and stiffness of the panels. On the other hand, face veneers and other plies may contain certain sizes and distributions of knots, splits, or growth characteristics that have no undesirable effects on strength properties for specific uses. Such uses include structural applications, such as sheathing for walls, roofs, or floors.

Types of Plywood

Broadly speaking, two classes of plywood are available: hardwood and softwood. In general, softwood plywood is intended for construction use and hardwood plywood for uses where appearance is important.

Originally, most softwood plywood was made of Douglas fir, but western hemlock, larch, white fir, ponderosa pine, redwood, southern pine, and other species are now used.

Most softwood plywood used in the United States is produced domestically, and U.S. manufacturers export some material. The bulk of softwood plywood is used where strength, stiffness, and construction convenience are more important than appearance. Some grades of softwood plywood are made with faces selected primarily for appearance and are used either with clear natural finishes or with pigmented finishes.

Hardwood plywood is made of many different species, both in the United States and overseas. More than half of all hardwood panels used in the United States are imported. Hardwood plywood is normally used where appearance is more important than strength. Most of the production is intended for interior or protected uses, although a very small proportion is made with glues suitable for exterior service. A significant portion of all hardwood plywood is available completely finished.

Plywood of thin, crossbanded veneers is very resistant to splitting and, therefore, nails and screws can be placed close together and close to the edges of panels.

Grades of Plywood

Plywood is graded according to defects on each surface and the type of glue used. Exterior or Interior plywood refers to the type of glue used to bond plies. Each face of the plywood has a letter grade—A, B, C, or D. Grade A means the face has no defects—it's perfect. Grade B means there are some defects (perhaps an area has a small patch). Grade C allows checks (splits) and small knotholes. Grade D allows large knotholes.

The commonly available grades are AC Interior, AD Interior, AC Exterior, and CDX, which is used for sheathing.

Particle Board

The group of materials generally classified as wood-base fiber and particle panel materials includes such familiar products as insulation boards, hardboards, particleboards, and laminated paperboards. In some instances, they are known by such proprietary names as

Masonite, Celotex, Insulite, and Beaver board or, in the instance of particleboards, by the kind of particle used, such as flakeboard, chipboard, or oriented strand board (OSB).

These panel materials are all reconstituted wood (or some other lignocellulose like bagasse) in that the wood is first reduced to small fractions and then put back together by special forms of manufacture into panels of relatively large size and moderate thickness. These board or panel materials in final form retain some of the properties of the original wood, but, because of the manufacturing methods, gain new and different properties from those of the wood. Because they are manufactured, they can be and are tailored to satisfy a use, need, or group of needs.

Speaking in general terms, the wood-base panel materials are manufactured by the following methods:

- Changing wood substance, essentially to fibers and then interfelting them together again into the panel material classed as building fiberboard.
- Strictly mechanical means of cutting or breaking wood into small discrete particles and then, with a synthetic resin adhesive or other suitable binder, bonding them together again in the presence of heat and pressure. These latter products are appropriately called *particleboards*.

Building fiberboards, then, are made essentially of fiberlike components of wood that are interfelted together in the reconstitution and are characterized by a bond produced by that interfelting. They are frequently classified as fibrous-felted board products. At certain densities under controlled conditions of hot pressing, rebonding of the lignin effects a further bond in the panel product produced. Binding agents and other materials may be added during manufacture to increase strength, increase resistance to fire, moisture, or decay, or to improve some other property. Among the materials added are rosin, alum, asphalt, paraffin, synthetic and natural resins, preservative and fire-resistant chemicals, and drying oils.

Particleboards are manufactured from small components of wood that are glued together with a thermosetting synthetic resin or equivalent binder. Wax sizing is added to all commercially produced particleboard to improve water resistance. Other additives may be introduced during manufacture to improve some property or provide added resistance to fire, insects such as termites, or decay. Particleboard is among the newest of the wood-base panel materials.

It has become a successful and economical panel product because of the availability and economy of thermosetting synthetic resins, which permit blends of wood particles and the synthetic resin to be consolidated and the resin set (cured) in a press that is heated.

Thermosetting resins used are primarily urea-formaldehyde and phenol-formaldehyde. Urea-formaldehyde is lowest in cost and is the binder used in greatest quantity for particleboard intended for interior or other nonsevere exposures. Where moderate water or heat resistance is required, melamine-urea-formaldehyde resin blends are being used. For severe exposures, such as exteriors or where some heat resistance is required, phenolics are generally used.

The kinds of wood particles used in the manufacture of particleboard range from specially cut flakes an inch or more in length (parallel to the grain of the wood) and only a few hundredths of an inch thick to fine particles approaching fibers or flour in size. The synthetic resin solids are usually between 5 and 10 percent by weight of the dry wood furnish. These resins are set by heat as the wood particle-resin blend is compressed in flat-platen presses.

As floor underlayment, particleboard provides (1) the leveling, (2) the thickness of construction required to bring the final floor to elevation, and (3) the indentation-resistant smooth surface necessary as the base for resilient finish floors of linoleum, rubber, vinyl tile, and sheet material. Particleboard for this use is produced in 4- by 8-foot panels commonly $^1/_4$, $^3/_8$, or $^5/_8$ inch thick. Separate use specifications cover particleboard floor underlayment. In addition, all manufacturers of particleboard floor underlayment provide individual application instructions and guarantees because of the importance of proper application and the interaction effects of joists, subfloor, underlayment, adhesives, and finish flooring. Particleboard underlayment is sold under a certified quality program where established grade marks clearly identify the use, quality, grade, and originating mill.

Other uses for particleboard have special requirements, as for phenol-formaldehyde, a more durable adhesive, in the board. Particleboard for siding, combined siding sheathing, and use as soffit linings and ceilings for carports, porches, and the like requires this more durable adhesive. For these uses, type 2 medium-density board is required. In addition, such agencies as the Federal Housing Administration have established requirements for particleboard for such use. The satisfactory performance of particleboard in exterior exposure depends not only on the manufacture and kind of adhesive used but also on the protection afforded by the finish. Manufacturers recognize the importance by providing both paint-primed panels and those completely finished with liquid paint systems or factory-applied plastic films.

Lumber Distribution

Large primary manufacturers and wholesale organizations set up distribution yards in lumber-consuming areas to more effectively distribute both hardwood and softwood products. Retail yards draw inventory from distribution yards and, in wood-producing areas, from local lumber producers.

Retail Yard Inventory

The small retail yards throughout the United States carry softwoods required for ordinary construction purposes and small stocks of one or two hardwoods in the grades suitable for finishing or cabinetwork. Special orders must be made for other hardwoods. Trim items such as moulding in either softwood or hardwood are available cut to size and standard pattern. Cabinets are usually made by millwork plants ready for installation. Many common styles and sizes are carried or cataloged by the modern retail yard. Hardwood flooring is available to the buyer only in standard patterns. Some retail yards may carry specialty stress grades of lumber such as structural light framing for truss rafter fabrication.

Some lumber grades and sizes serve a variety of construction needs. Some species or species groups are available at the retail level only in grade groups. Typical are house-framing grades, such as joist and plank, which is often sold as No. 2 and Better (2&BTR). The percentage of each grade in a grouping is part of the purchase agreement between the primary lumber manufacturer and the wholesaler. However, this ratio may be altered at the retail level by sorting. Where grade grouping is the practice, a requirement for a specific grade, such as No. 1, at the retail level will require sorting or special purchase. Grade grouping occurs for reasons of tradition and of efficiency in distribution.

Another important factor in retail yard inventory is that not all grades, sizes, and species described by the grade rules are produced, and not all those produced are distributed uniformly to all marketing areas. Regional consumer interest, building code requirements, and transportation costs influence distribution patterns. Often small retail yards will stock only a limited number of species and grades. Large yards, on the other hand, may cater to particular construction industry needs and carry drier dimension grades along with clears, finish, and decking. The effect of these variable retail practices is that the grades, sizes, and species outlined in the grade rules must be examined to determine what actually is available. A brief description of lumber products commonly carried by retail yards follows.

Stress-Graded Lumber for Construction

Dimension is the principal stress-graded lumber item available in a retail yard. It is primarily framing lumber for joists, rafters, and studs. Strength, stiffness, and uniformity of size are essential requirements. Dimension is stocked in all yards, frequently in only one or two of the general-purpose construction woods (such as pine, fir, hemlock, or spruce):

- The 2 × 6, 2 × 8, and 2 × 10 dimensions are found in grades of Select Structural, No. 1, No. 2, and No. 3. They may also be found in combinations of No. 2 & BTR or possibly No. 3 & BTR.
- The 2 × 4 grades available would normally be Construction and Standard. They are sold as Standard and Better (STD&BTR), Utility and Better (UTIL&BTR), or Stud, in lengths of 10 feet and shorter.

Dimensional lumber is often found in nominal 2-, 4-, 6-, 8-, 10-, or 12-inch widths and 8- to 18-foot lengths in multiples of 2 feet. Dimensional lumber that is formed by structural end-jointing procedures may be found. Dimensional lumber thicker than 2 inches and longer than 18 feet is not available in large quantity.

Other stress-graded products generally present are posts and timbers, with some beams and stringers also possibly in stock. Typical stress grades in these products are Select Structural and No. 1 Structural in Douglas fir and No. 1SR and No. 2SR in southern pine.

Nonstress-Graded Lumber for Construction

Boards are the most common nonstress-graded general-purpose construction lumber in the retail yard. Boards are stocked in one or more species, usually in nominal 1-inch thickness. Standard nominal widths are 2, 3, 4, 6, 8, 10, and 12 inches. Grades most generally available in retail yards are No. 1, No. 2, and No. 3 (or Construction, Standard, and Utility). These will often be combined in grade groups. Boards are sold square-ended, dressed and matched (tongued and grooved), or with a ship-lapped joint. Boards formed by end jointing of shorter sections may form an appreciable portion of the inventory.

Appearance Lumber

Completion of a construction project usually depends on a variety of lumber items available in finished or semifinished form. The following items often may be stocked in only a few species, in only a few finishes, or in limited sizes depending on the yards:

- *Finish*—Finish boards usually are available in a local yard in one or two species, principally in grade C&BTR. Redwood and cedar have different grade designations. Grades such as Clear Heart, A, or B are used in cedar; Clear All Heart, Clear, and Select are typical redwood grades. Finish boards are usually a nominal 1-inch thick, dressed two sides to $^{3}/_{4}$ inch. The widths usually stocked are nominal 2 to 12 inches in even-numbered inches.
- *Siding*—Siding, as the name implies, is intended specifically to cover exterior walls. Beveled siding is ordinarily stocked only in white pine, ponderosa pine, western red cedar, cypress, or redwood. *Drop siding* (also known as *rustic siding* or *barn siding*) is usually stocked in the same species as beveled siding. Siding may be stocked as B&BTR or C&BTR, except in cedar (where Clear, A, and B may be available) and redwood (where Clear All Heart and Clear will be found). Vertical grain (VG) is sometimes a part of the grade designation. Drop siding sometimes is stocked also in sound knotted C and D grades of southern pine, Douglas fir, and hemlock. Drop siding may be dressed, matched, or ship lapped.
- *Flooring*—Flooring is made chiefly from hardwoods such as oak and maple, and the harder softwood species, such as Douglas fir, western larch, and southern pine. Often, at least one softwood and one hardwood are stocked. Flooring is usually nominal 1-inch thick dressed to $^{25}/_{32}$ inch, and 3- and 4-inch nominal width. Thicker flooring is available for heavy-duty floors in hardwoods and softwoods. Thinner flooring is available in hardwoods, especially for recovering old floors. Vertical and flat grain (also called quartersawed and plain-sawed) flooring is manufactured from both softwoods and hardwoods. Vertical-grained flooring shrinks and swells less than flat-grained flooring, is more uniform in texture, wears more uniformly, and the joints do not open as much.

  Softwood flooring is usually available in B and Better grade, C Select, or D Select. The chief grades in maple are Clear

No. 1 and No. 2. The grades in quartersawed oak are Clear and Select, and in plain-sawed, Clear, Select, and No. 1 Common. Quartersawed hardwood flooring has the same advantages as vertical-grained softwood flooring. In addition, the silver or flaked grain of quartersawed flooring is frequently preferred to the figure of plain-sawed flooring. Beech, birch, and walnut and mahogany (for fancy parquet flooring) are also occasionally used.

- *Casing and Base*—Casing and base are standard items in the more important softwoods and are stocked by most yards in at least one species. The chief grade, B and Better, is designed to meet the requirements of interior trim for dwellings. Many casing and base patterns are dressed to $^{11}/_{16} \times 2^{1}/_{4}$. Other sizes used include $^{9}/_{16} \times 3$, $3^{1}/_{4}$, and $3^{1}/_{2}$. Hardwoods for the same purposes (such as oak and birch) may be carried in stock in the retail yard or may be obtained on special order.
- *Shingles and Shakes*—Shingles usually available are sawn from western red cedar, northern white cedar, and redwood. The shingle grades are: Western red cedar, No. 1, No. 2, No. 3; northern white cedar, Extra, Clear, 2nd Clear, Clear Wall, Utility; redwood, No. 1, No. 2 VG, and No. 2 MG.

  Shingles that are all heartwood give greater resistance to decay than do shingles that contain sapwood. *Edge-grained shingles* are less likely to warp than *flat-grained shingles*; *thick-butted shingles* less likely than *thin shingles*; and *narrow shingles* less likely than *wide shingles*. The standard thicknesses of shingles are described as $^{4}/_{2}$, $^{5}/_{2}$, $^{1}/_{4}$, and $^{5}/_{2}$ (four shingles to 2 inches of butt thickness, five shingles to $2^{1}/_{4}$ inches of butt thickness, and five shingles to 2 inches of butt thickness). Lengths may be 16, 18, or 24 inches. Random widths and specified widths (dimension shingles) are available in western red cedar, redwood, and cypress.

  Shingles are usually packed four bundles to the square. A square of shingles will cover 100 square feet of roof area when the shingles are applied at standard weather exposures.

  Shakes are handsplit, or handsplit and resawn from western red cedar. Shakes are of a single grade and must be 100 percent clear, graded from the split face in the case of handsplit and resawn material. Handsplit shakes are graded from the best face. Shakes must be 100 percent heartwood free of bark and sapwood. The standard thickness of shakes ranges from

$^3/_8$ to $1^1/_4$ inches. Lengths are 18 and 24 inches, and a 15-inch Starter-Finish Course length.

Important Purchase Considerations

The following are some of the points to consider when ordering lumber or timbers:

- *Quantity*—Feet, board measure, number of pieces of definite size and length. Consider that the board measure depends on the thickness and width nomenclature used and that the interpretation of this must be clearly delineated. In other words, nominal or actual, pattern size, and so forth must be considered.
- *Size*—Thickness in inches—nominal and actual if surfaced on faces. Width in inches—nominal and actual if surfaced on edges. Length in feet—may be nominal average length, limiting length, or a single uniform length. Often a trade designation, random length, is used to denote a nonspecified assortment of lengths. Note that such an assortment should contain critical lengths as well as a range. The limits allowed in making the assortment random can be established at the time of purchase.
- *Grade*—As indicated in grading rules of lumber manufacturing associations, some grade combinations (B&BTR) are official grades; other (STD&BTR light framing, for example) are grade combinations and subject to purchase agreement. A typical assortment is 75 percent Construction and 25 percent Standard, sold under the label STD&BTR. In softwood, each piece of such lumber typically is stamped with its grade, a name or number identifying the producing mill, the dryness at the time of surfacing, and a symbol identifying the inspection agency supervising the grading inspection. The grade designation stamped on a piece indicates the quality at the time the piece was graded. Subsequent exposure to unfavorable storage conditions, improper drying, or careless handling may cause the material to fall below its original grade.

  Note that working or rerunning a graded product to a pattern may result in changing or invalidating the original grade. The purchase specification should be clear regarding regrading or acceptance of worked lumber. In softwood lumber, grades for dry lumber generally are determined after kiln drying and surfacing. This practice is not general for hardwood factory lumber, however, where the grade is generally based on grade and size prior to kiln drying.

- *Species or groupings of wood*—Douglas fir, cypress, hemlock, fir. Some species have been grouped for marketing convenience. Others are traded under a variety of names. Be sure the species or species group is correct and clearly depicted on the purchase specification.
- *Product*—Flooring, siding, timbers, boards, and so forth. Nomenclature varies by species, region, and grading association. To be certain the nomenclature is correct for the product, refer to the grading rule by number and paragraph.
- *Condition of seasoning*—Air dry, kiln dry, and so on. Softwood lumber dried to 19 percent moisture content or less (S-DRY) is defined as dry by the American Lumber Standard. Other degrees of dryness are partially air-dried (PAD), green (S-GRN), and 15 percent maximum (KD in southern pine). There are several specified levels of moisture content for redwood. If the moisture requirement is critical, the levels and determination of moisture content must be specified.
- *Surfacing and working*—Rough (unplaned), dressed (surfaced), or patterned stock. Specify condition. If surfaced, indicate S4S, S1S1E, and so on. If patterned, list pattern number with reference to the appropriate grade rules.
- *Grading Rules*—Official grading agency name, product identification, paragraph number or page number or both, date of rules or official rule volume (rule No. 16, for example).
- *Manufacturer*—Most lumber products are sold without reference to a specific manufacturer. If proprietary names or quality features of a manufacturer are required, this must be stipulated clearly on the purchase agreement.

Engineered Lumber

Some lumber has been engineered to produce a variety of improvements for use in construction of homes and office and industrial buildings. There is glued laminated lumber that is called glulam. This engineered lumber is composed of wood laminations (lams). They are bonded together with adhesives. The grain of the lams runs parallel to the length of the glulam member. Lams are $1^3/_8$ inches thick for southern pine species and $1^1/_2$ inches for western wood species. Laminated boards up to 2 inches thick may also be used to form glulam timbers. Glulam widths are standardized at

$3^1/_8$ inches, $3^1/_2$ inches, $5^1/_8$ inches, $5^1/_2$ inches, and $6^3/_4$ inches. However, almost any width can be custom made.

Glulam dates back to the early 1900s when the first patents for glulam were obtained in Switzerland and Germany. One of the first glulam structures in the United States was built for the USDA Forest Products Laboratories in 1934. Glulam manufacture has evolved to make glulam one of the most widely used engineered lumber products in the construction industry. In the past, glulam was mainly associated with heavy timber construction, such as industrial roofing systems, bridges, and marine piers. Glulam timbers are used in residential construction as well as in commercial industrial construction.

Summary

Lumber is the basic construction material used in construction. Wood is a highly variable product so lumber is standardized by sizes and grades. The three classes of grading are stress-graded, nonstress-graded, and appearance. Plywood is made up of layers of wood veneer in wide long panels that have several advantages over solid wood lumber. Particleboard is made of small chunks of wood compressed into a large panel. Understanding lumber and wood product distribution and inventories leads to more successful and economic construction projects.

Review Questions

1. Why is lumber standardized by size and grade?
2. What grades are used for structural light framing?
3. What grade of lumber would you use for kitchen cabinet construction?
4. What are the advantages of plywood over solid lumber?
5. What grades and species of select softwood lumber are available at your local retail lumberyard?

Chapter 4

Strength of Timbers

The various mechanical properties of woods have been investigated by exhaustive testing in many laboratories, most notably the Forest Products Laboratory of the U.S. Department of Agriculture in Madison, Wisconsin. In addition, much research has been done in civil and agricultural engineering laboratories at many state universities.

Tension

A tension test is made as indicated in Figure 4-1. Although modern testing machines are by no means as simple as the apparatus shown, it serves well to show how such tests are made. The specimen is placed in the machine, gripped at each end, and the load is progressively increased until the material reaches its failure point. Usually, the elongation is not of any significance unless it is desired to determine the modulus of elasticity. The allowable fiber stresses indicated by modern stress grading (such as 1200f or 1450f) apply equally to extreme fiber stress in bending and to tensile stresses.

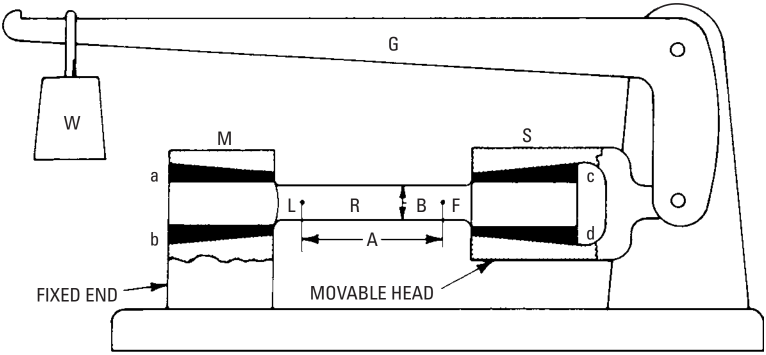

Figure 4-1 The tension test. The specimen R is placed in the wedge grips a, b, c, and d, thus applying tension between the fixed end and the movable head of the machine. The movable head is connected to the scale lever G on which the weight W slides. This arrangement is similar to an ordinary weighing scale. Two center marks (L and F) are punched on the specimen at a standard distance (A) apart. When testing, the pull on the specimen is gradually increased by moving W to the left. Dimensions A and B are then measured after each load increase.

Example A truss member is subjected to a tensile force of 50,000 pounds. What size timber of No. 1 dense yellow pine (1600f grade) will be required?

This is the calculation:

$$\frac{5000}{1600} = 31.25 \text{ square inches cross section}$$

Therefore, either a standard-dressed 4 inches × 10 inches ($3^1/_2$ inches × $9^1/_4$ inches) or a 6 inches × 6 inches ($5^1/_2$ inches × $5^1/_2$ inches) should be adequate.

Compression

A column supporting a load that tends to crush it is said to be in *compression*. Allowable compression parallel to the grain in stress-graded lumber is usually slightly less than the allowable bending and tensile stresses. For No. 1 dense timbers, 1600f grade, the compression stress is 1500 pounds per square inch, and for No. 2 dense, 1200f grade, it is 900 pounds per square inch. The builder need have no other concern than to see that the specified stresses are not exceeded. For loads applied across the grain, these stresses are much less. For the two grades mentioned, the cross-grain stress is 455 pounds per square inch for both grades.

When making a compression test (see Figure 4-2), a prepared specimen is placed between two plates, and a measured load is

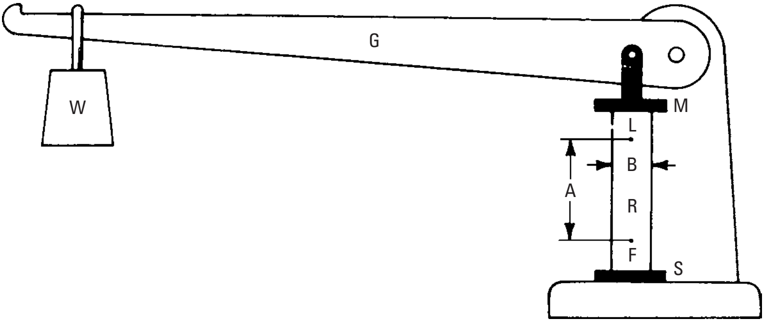

Figure 4-2 The compression test. The specimen R is placed between two plates (M and S), and a compression stress of any desired intensity is applied by moving the weight W on the lever G. As the load is gradually increased, the changes in dimensions A and B are noted, and a final result can be obtained that will indicate the amount of compression that the specimen can withstand.

applied. The load is increased progressively until the failure point is reached. Actually, modern wood-compression testing is not quite so simple. An instrument called a *compressometer* is pinned to the side of the specimen with sharp-pointed screws at points corresponding to points L and F in Figure 4-2. It is fitted with a series of levers that are connected to a dial gauge, usually reading to .0001 inch. The stem of the dial gauge is depressed until it shows a positive reading and compression in the specimen releases a part of the gauge. In this way, a sudden failure will release the gauge entirely, and jamming and ruining of the instrument is avoided. Usually, at least one of the plates through which pressure is applied is a cast iron hemisphere, thereby ensuring an evenly distributed pressure over the entire cross-section of the specimen.

Working Stresses for Columns

The amount of allowable loads on wood columns has been (and continues to be) a subject that is open to some discussion. The matter is complicated by the fact that columns of different lengths and diameters do not behave in the same manner under loadings. For the rather short column, failure will be caused by the actual crushing of the fibers of the wood, and the full compressive strength of the wood may thus be utilized. For a slightly longer column, failure may be caused by a sort of diagonal shearing action. For a long column, it will probably fail by bending sidewise and breaking. No one method can be adapted exactly for calculating allowable loads on columns of all lengths and slenderness ratios.

Column formulas are numerous. Some are based on empirical data or the results of actual testing. Only one, the *Euler* (pronounced *oiler*) *formula* seems to be based on purely mathematical calculations. It is of German origin, and, in its original form, it is so cumbersome that few designers in the U.S. care to use it. It assumes that a column will fail by *bending and breaking*. This is ensured only if the column is long and comparatively slender. However, a *modified* Euler formula is greatly favored by present-day timber designers. It is written as follows:

$$\text{Allowable load per square inch of cross section} = \frac{.3E}{(l/d)^2}$$

where the following is true:

E = modulus of elasticity of the timber used

l = unsupported length of the column (in inches)

d = least side, or diameter of a round column

This formula is applicable only when the results of its use do not indicate a higher stress than the maximum allowable unit stresses (as defined by the stress grade of the timber used). The ratio *l/d* in this equation is sometimes called the *slenderness ratio*. As an example of the use of this formula, consider the following problem.

How much of a load may safely be imposed on a 6-inch × 6-inch dressed yellow pine column that is 12 feet long?

The slenderness ratio, or *l/d*, of 6-inch × 6-inch columns 12 feet long is $^{144}/_{6} = 24$, and the modulus of elasticity of almost all good yellow pine is 1,760,000 pounds per square inch. This is the calculation:

$$\frac{0.3 \times 1{,}760{,}000}{24 \times 24} = 917 \text{ pounds per square inch}$$

The cross-sectional area of standard 6-inch × 6-inch timbers is 30.25 square inches. Therefore, the total load allowable on the column will then be:

$$917 \times 30.25 = 27{,}739 \text{ pounds}$$

An important point with respect to timber in compression is that the ends should be cut exactly square so that there will be a full bearing surface. Otherwise the timber will be subjected at the ends to more than the working stress (see Figure 4-3).

Shearing Stresses

Shearing stresses in wood are dangerous only in the direction parallel to the grain. It is almost impossible to shear the material across the grain until the specimen has been *crushed*. The crushing strength, then, and not the shearing resistance of the wood, will govern the maximum stress that can be applied to the wood.

The standard shearing test for woods with the grain was developed by the American Society for Testing Materials (ASTM), and the procedure is standardized. The specimens are standard dimensions, shaped as shown in Figure 4-4 in pairs, one with the notch at right angles to that in the other. The results of the tests on the two specimens are averaged, but usually they vary only slightly. The blocks are tested in a special shear tool that is loaded in a universal testing machine, with the load applied at a rate of .024 inch per minute.

The results of testing different types of woods vary widely. To provide for this lack of uniformity, the shear allowances in stress-rated timbers may contain a reduction factor (safety factor) of as much as

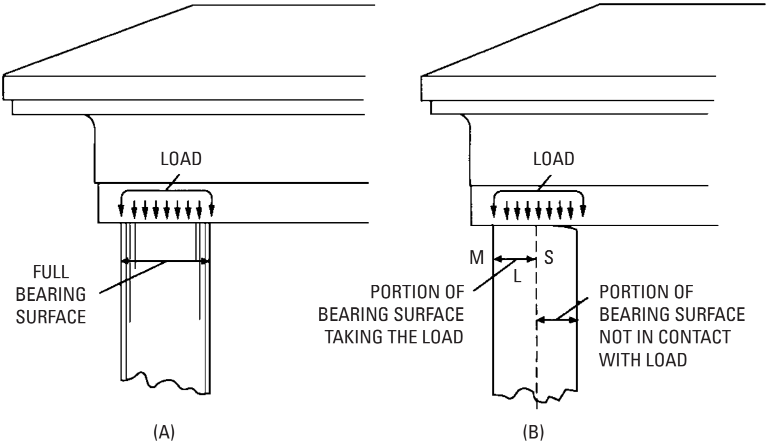
LOAD
FULL BEARING SURFACE
(A)
LOAD
M
S
L
PORTION OF BEARING SURFACE TAKING THE LOAD
PORTION OF BEARING SURFACE NOT IN CONTACT WITH LOAD
(B)

Figure 4-3 (A) Good and (B) poor column bearing surfaces illustrate the importance of squaring columns accurately when cutting them. The entire top of the column must be in contact with the load member so that the pressure per square inch of cross section on the column will correspond to the allowable working pressure for which the column was designed. If the portion MS of the bearing surface in contact is only half the entire surface, then the stress applied on the top of the column will be twice that of full contact, as shown by (A).

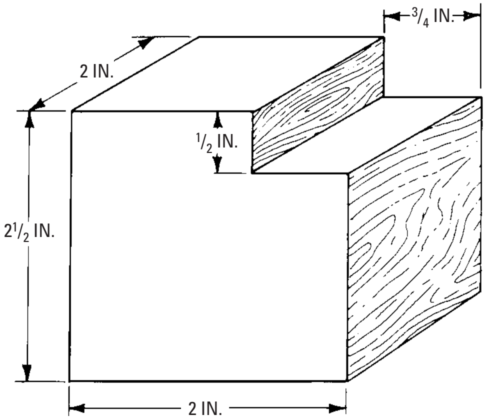
3/4 IN.
2 IN.
1/2 IN.
2 1/2 IN.
2 IN.

Figure 4-4 The ASTM standard wood-shear test specimens.

10, or it may be as little as 2 or 3. This is necessary because a piece that is below average in shearing resistance may appear anywhere.

The distinctions *across the grain* and *with the grain* should be carefully noted. Wet or green wood, in general, shears approximately one-fifth to one-half as easily as dry wood. A surface parallel to the rings (tangent) shears more easily than one parallel to the medullary rays. The lighter conifers and hardwoods offer less resistance than the heavier kinds, but the best pine shears one-third to one-half more readily than oak or hickory, thereby indicating that great shearing strength is characteristic of tough woods.

Horizontal Shears

The types of shears discussed in the preceding section are called *external* shears because they are caused by forces that originate outside the body of the material, and for the most part, they are evident and readily provided for (see Figure 4-5). In addition, in every loaded beam, a system of *internal* stresses is set up. This may be explained by observing that every shearing force results from *two* forces that are *unbalanced*. They do not meet at the same point, and a *stress couple* is set up that can be met and held motionless only by another couple that acts in opposition to it.

This is illustrated in Figure 4-6. It represents a particle of indefinite size, but possibly infinitely small, that has been extracted from the body of a beam that has been stressed by bending. In Figure 4-6A, the particle has been subjected to a vertical shearing force, or pair of forces, since there can be no action without an opposing and equal reaction. The particle is *unstable*. Since the two forces do not meet at a common point, the particle tends to revolve in a counterclockwise direction.

In Figure 4-6B, a pair of horizontal forces is supplied. These forces represent another stress couple that balances or neutralizes the vertical shearing forces. They are known as the *internal horizontal shears*. A reasonable deduction, therefore, is that at any point on the beam, there exist internal horizontal shears equal in intensity to the external vertical shears. The horizontal shears are dangerous in wood beams. That is because timbers have a low resistance to shearing with the grain.

Figure 4-6C shows how the four forces may be resolved into a single pair of concurrent forces that *do* meet at a common point. These are not shearing forces. They are, however, the result of shearing forces. When the material has approximately the same resistance to shears in all directions (such as concrete or steel), this is what occurs. Each particle tends to elongate in one diagonal

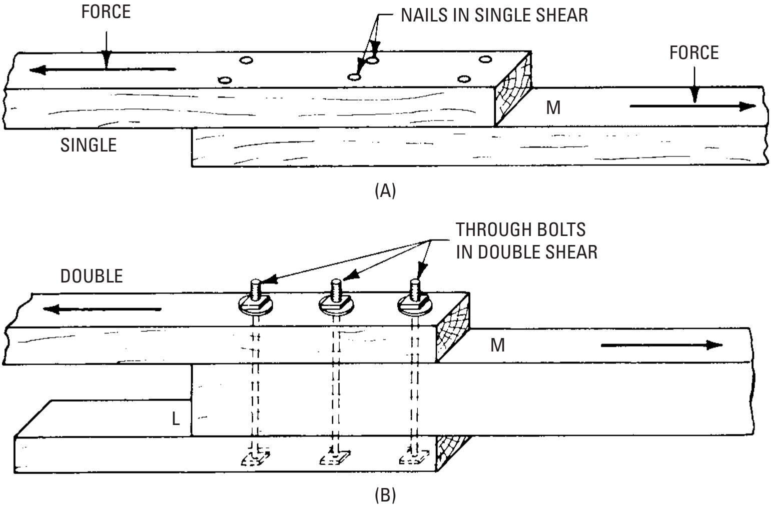

Figure 4-5 Examples of single and double shear. (A) The nails are in single shear. This is the assumption made as a rule when calculating the strength of nailed joints. However, nails may be placed in double shear, similar to the bolts in (B), if they are long enough to almost penetrate all three members. The nails in double shear will safely carry twice the load that could safely be placed on nails in single shear. However, most nailed joints are designed as if the nails were in single shear, though they may actually be in double shear. In (B), the bolts are truly in double shear, but the joint is not twice as strong as a plain lapped joint if the center member is not at least as thick as the combined thicknesses of the outer members. The outer members are usually both the same thickness.

and forms the diagonal tension that is so dangerous in reinforced concrete beams. In fact, it is the most dangerous stress in such beams, and to resist it, elaborate web reinforcing and bent-up bars are provided at points where this stress is highest (usually near the ends of the beams). In I-beams, diagonal tension results in buckling or wrinkling of the relatively thin webs. In wood beams, the forces are not resolved in this way, since wood is strong enough to resist the vertical components, but the horizontal components tend to split the beam (usually at or near the ends near the center of the height). Wood beams that are season-checked at the ends (as many are) are low in resistance to such stresses.

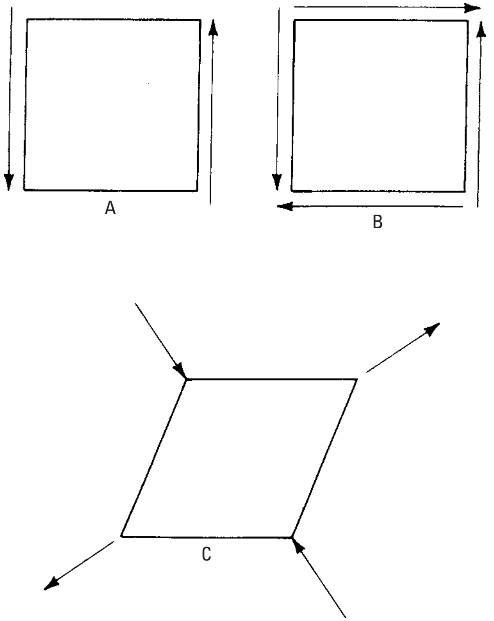

Figure 4-6 The actions of vertical and horizontal shears.

If the depth of a wood beam is greater than one-tenth to one-twelfth of its span, horizontal shears (and not bending strength) often govern its ability to carry loads. Shears are usually not dangerous in wood beams unless they are relatively deep and heavily loaded.

Transverse or Bending Stress

This is the kind of stress present on numerous building timbers (such as girders, joists, or rafters) that causes a deflection or bending between the points of support. What takes place when these or similar members are subjected to bending stress is discussed in the section titled *Beams* later in this chapter.

Stiffness

By definition, stiffness is that quality possessed by a beam or other timber to resist the action of a bending force. The action of the bending force tends to change a beam from a straight to a curved form (that is, a *deflection* takes place). When a load is applied, the beam (originally assumed straight and horizontal) sags or bends downward between the supports. The amount of downward movement measured at a point midway between the supports is the amount of deflection. The action of beams subjected to bending forces is described as follows.

If a load of 100 pounds placed in the middle of a stick that is 2 inches × 2 inches and 4 feet long, supported at both ends, bends or deflects this stick one-eighth of an inch (in the middle), then 200 pounds will bend it about $^1/_4$ inch, 300 pounds, $^3/_8$ inch, and so on, the deflection varying directly as the load. This is in accordance with Hooke's Law that states that stress is proportional to strain. Soon, however, a point is reached where an additional 100 pounds adds more than $^1/_8$ inch to the deflection—the limit of elasticity has been exceeded.

Taking another piece from the straight-grained and perfectly clear plank of the same depth and width but 8 feet long, the load of 100 pounds will deflect it by approximately 1 inch. Doubling the length reduces the stiffness eightfold. Stiffness, then, decreases as the cube of the length.

If AB in Figure 4-7 is a piece of wood, and D is the deflection produced by a weight or load, then

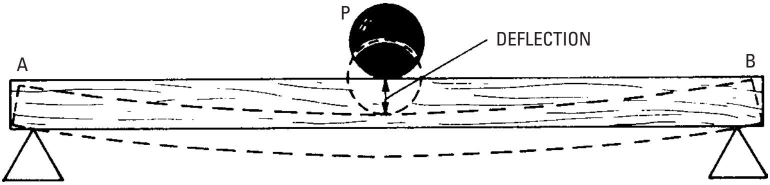

Figure 4-7 A simple beam, loaded at the middle and supported at both ends, is used to illustrate the term deflection.

$$\text{deflection } (D) = \frac{Pl^3}{48EI}$$

where the following is true:

P = the load, concentrated at the center of the span (in pounds)
l = the length of the span (in feet)
E = the modulus of elasticity of the material
I = the moment of inertia (for rectangular beams =

$$\frac{bd^3}{12}$$

where b = width and d = height).

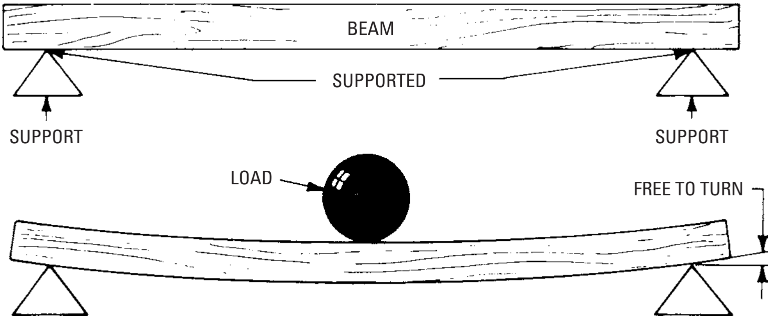

Figure 4-8 A beam supported at the ends. The ends of the beam are free to follow any deflection, thus offering no resistance and rendering the beam less stiff than when the ends are fixed.

The following rules (in conjunction with Figure 4-8, Figure 4-9, and Figure 4-10) define the stiffness and strength of practically all types of wood beams:

- For beams with a rectangular cross-section, equal depths, and equal spans, their load-carrying capacity varies *directly* as their widths.
- For beams with equal widths and equal spans, their strength varies directly as the square of their depth.

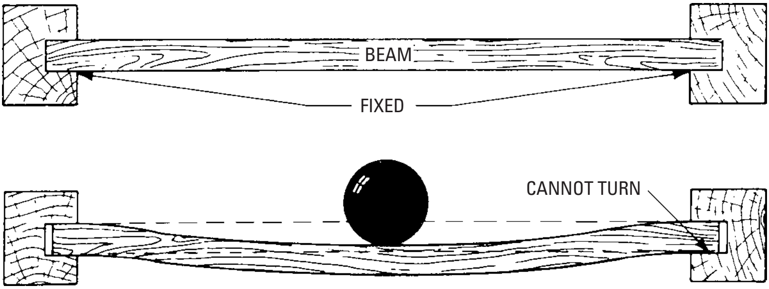

Figure 4-9 A beam supported at the ends. The ends are gripped or embedded in some unyielding substance so that they cannot turn or follow the deflection of the beam under an applied load. The beam then deflects in a compound curve, thus adding extensively to its stiffness. Therefore, a beam with fixed ends will deflect less than one with supported ends.

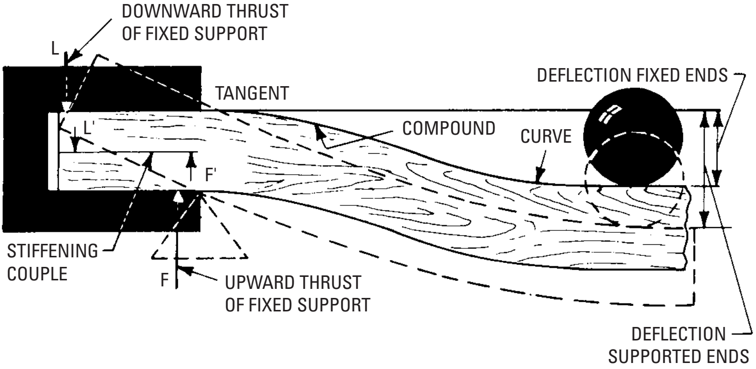
DOWNWARD THRUST
OF FIXED SUPPORT
L
L'
TANGENT
COMPOUND
CURVE
DEFLECTION FIXED ENDS
F'
STIFFENING
COUPLE
F
UPWARD THRUST
OF FIXED SUPPORT
DEFLECTION
SUPPORTED ENDS

Figure 4-10 One end of a beam illustrating the stiffening effect of fixed ends as compared with supported ends. When the ends are fixed, the deflection of the beam will be resisted by an upward thrust, indicated by F, and a downward thrust, L. These thrusts form a resisting or stiffening couple that holds the portion of the beam embedded in the bearing in a horizontal position, thereby causing the beam to deflect in a compound curve, which increases its stiffness. The dotted lines show the excess deflection for the same load if the beam were simply supported at the ends.

- If depths and widths are the same, strengths vary *inversely* as the lengths of the spans.
- Their stiffness, or resistance to deflection, will vary *inversely* as the *cubes* of their spans, other factors being equal.
- Their stiffness will vary directly as the *cubes* of their depths, other factors being equal.
- Other factors being equal, stiffness will vary *directly* as their widths.
- If a beam is split horizontally, and the two halves are laid side by side, they will carry only *one-half* as much loading as the original beam.

These relations are not strictly true for I-beams because of their irregular shapes, but they are approximately true for all types of beams. It is usually most economical with materials to use as deep a beam as can conveniently be employed. Note that double 2-inch × 4-inch trimmers over window or door heads, if set edge up, are 8 times as strong and 32 times as stiff as when placed flatwise.

Both strength and stiffness are greater in dry timber than in green or wet wood of the same species. A piece of long-leaf yellow pine is 30 percent to 50 percent stronger, and 30 percent stiffer, when in an air-dry condition than when green. In general, both strength and stiffness are proportional to densities, or dry weights, although this is not always true. Edge-grain pieces are usually stronger and stiffer because the tangent to the rings runs horizontally, but not appreciably so. There is little or no difference in the sapwood and heartwood of the same species if the densities are the same. The tool handle of red heartwood is as serviceable as the handle of white sapwood, although white sapwood handles are still called premium grade.

Modulus of Elasticity

Since it is desirable (and, for many purposes, essential) to know beforehand that a given piece with a given load will bend only by a given amount, the stiffness of wood is usually stated in a uniform manner under the term *modulus* (measure) *of elasticity*. For good grades of Douglas fir and yellow pine that are stress-rated, the modulus of elasticity is 1,760,000 pounds per square inch.

Beams

A beam is a single structural member (usually horizontal or nearly so) that carries a load or loads over a given space. At their supports, beams may be:

- *Freely supported*—This merely means the beams are resting on their bearings.
- *Restrained, or partially fixed at their bearings*—Although some designers choose to consider such restraint in their designs, the actual degree of restraint can never be accurately determined, so restrained beams are more often considered as being freely supported.
- *Fixed at their supports*—In wood beams, this condition is rarely found. In steel frames, it is not unknown. In reinforced concrete frames, it is quite common. Attempts to fix the ends of wood beams are rarely permanent. Building the ends of a beam, *any* beam, into a wall or casting it into concrete for a short distance does not fix the beams at their supports.

Allowable Loads on Wood Beams

The allowable loads on freely supported wood beams of any species are readily calculated if the timbers are stress-rated and the allowable fiber-stress is known. For a beam with a loading that is evenly distributed along its span, this is the equation to use:

$$W = \frac{f \times b \times d^2}{9 \times L}$$

where the following is true:

W = allowable evenly distributed loading, in pounds
f = allowable fiber stress in pounds per square inch
b = width of the beam, in inches
d = depth of the beam, in inches
L = length of the span, in feet

Figure 4-11 defines the dimensions of a beam—b, width; d, depth; and L, length.

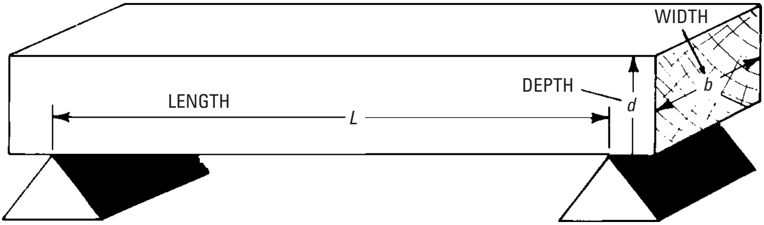

Figure 4-11 A beam resting on knife-edge supports illustrates the terms length, width, and depth.

Example What will be the maximum allowable load on a beam whose nominal size is 6 inches × 12 inches and whose actual size is $5^1/_2$ inches × $11^1/_2$ inches, with an 18-foot span? The timber is to be 1500f stress-rated. This is the calculation:

$$\frac{1500 \times 5.5 \times 11.5 \times 11.5}{9 \times 18} = 6735 \text{ pounds}$$

If the loading is to be concentrated at the center of the span, use one-half the load as calculated by the formula given. Therefore, for the same timber as calculated, with the same span, the allowable concentrated load will be

$$\frac{6735}{2} = 3367 \text{ pounds}$$

Breaking Loads on Wood Beams

Breaking loads are of no interest to the builder because the term is meaningless unless some explanation is made, since wood is extremely sensitive to the *duration* of loads. The load that would break a beam over a long period of time, for example, 10 years, will be only approximately $^{9}/_{16}$ of the load that would break it in a few minutes. The stresses specified in stress-rating lumber and timbers recognize this phenomenon. It is presumed that the full design loading will not be applied for more than 10 years during the life of the structure, and the time may be either cumulatively intermittent or continuous. It is also presumed that 90 percent of the full design loading may safely be applied for the full life of the structure. These presumptions make the use of stress-rated lumber quite conservative.

Example What is the safe working load, concentrated at the center of the span, for a full-size 6-inch × 10-inch white oak timber with a 12-foot span, stress-rated 1900f, if it is laid flatwise (see Figure 4-12A)? If it is set edge up (see Figure 4-12B)?

This is the calculation for the timber laid flatwise:

$$\frac{1900 \times 10 \times 6 \times 6}{9 \times 12} = 6333$$

$$\frac{6333}{2} = 3167 \text{ pounds}$$

For the timber set edge up:

$$\frac{1900 \times 6 \times 10 \times 10}{9 \times 12} = 10{,}555$$

$$\frac{10{,}555}{2} = 5278 \text{ pounds}$$

Distributed Load

Instead of placing the load in the middle of a beam, as in the example just given, the load may be regarded as being *distributed* (that is, the beam is uniformly loaded, as shown in Figure 4-13). Although this type of loading is actually a series of concentrated loads, it is often found in actual practice, and it is usually considered as being *uniformly* distributed.

Cantilever Beams

A *cantilever beam* (see Figure 4-14 and Figure 4-15) is firmly fixed at one end, or it is freely supported but with the end running some distance to a support above it. The loading may be either

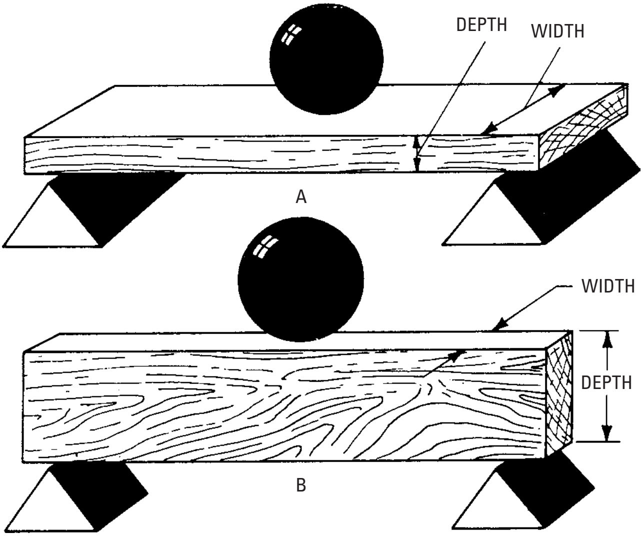

Figure 4-12 The terms width and depth of a beam depend on the position of the beam. In (A) the broad side is the width of the beam, whereas in (B), which is the same beam turned over 90°, the narrow side is the width.

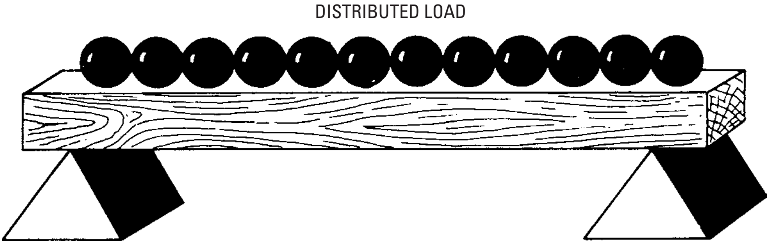

Figure 4-13 A distributed load is indicated by the iron balls equally spaced along the beam between the supports.

distributed or concentrated at any point on the span. All beams that project beyond a support and that carry a load at the free end are classed as cantilevers. Stresses set up in the beam external to the outside support are the same as when the beam is rigidly fixed at the support.

Figure 4-16 illustrates the comparison between the working loads of variously supported beams, contrasting middle and distributed

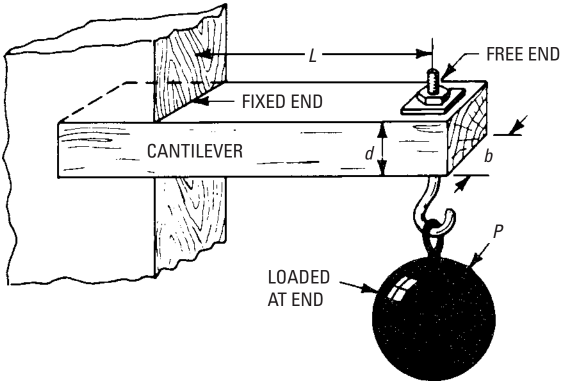
L
FREE END
FIXED END
CANTILEVER
d
b
P
LOADED
AT END

Figure 4-14 A simple cantilever beam that is fixed at one end and free at the other; this beam is supporting a concentrated load at its free end.

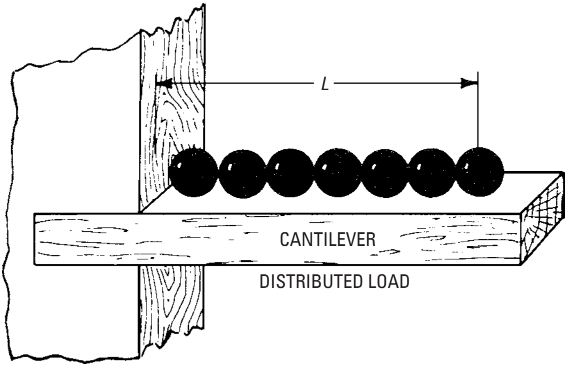
L
CANTILEVER
DISTRIBUTED LOAD

Figure 4-15 A cantilever beam with an equally distributed load.

loads of beams supported at both ends with equal loads on cantilever beams. In Figure 4-16, P= the concentrated loads, W= the evenly distributed load, f= the fiber stress, b= the width of the beam, d= the depth of the beam, and L= the length of the span.

Wind Loads on Roofs

Snow on roofs (if the roofs are of a slope that is low enough to permit the snow to lie in place) will weigh approximately 8 pounds per square foot per foot of depth when freshly fallen. When wet, its

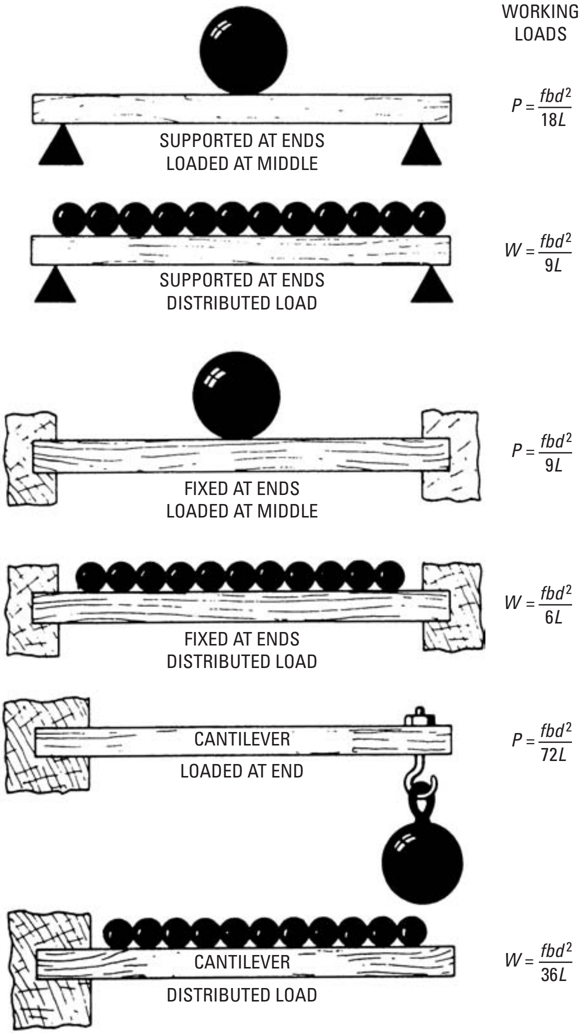
WORKING LOADS
SUPPORTED AT ENDS LOADED AT MIDDLE
$P = \frac{fbd^2}{18L}$
SUPPORTED AT ENDS DISTRIBUTED LOAD
$W = \frac{fbd^2}{9L}$
FIXED AT ENDS LOADED AT MIDDLE
$P = \frac{fbd^2}{9L}$
FIXED AT ENDS DISTRIBUTED LOAD
$W = \frac{fbd^2}{6L}$
CANTILEVER LOADED AT END
$P = \frac{fbd^2}{72L}$
CANTILEVER DISTRIBUTED LOAD
$W = \frac{fbd^2}{36L}$

Figure 4-16 The working loads for beams with different types of loads and different modes of support.

weight will vary with the water content. A commonly used figure is 10 pounds per cubic foot. In calculating snow load, use the depth of the heaviest snow on record multiplied by 10, but never less than 20 pounds per square foot anywhere. If a roof is designed for much less than 20 pounds per square foot of live load, it will not safely support workers.

In the past, there was considerable confusion regarding wind loads against sloping roofs, and some rather weird formulas were evolved to calculate their imaginary intensities (although no one had ever seen a roof blow *in*, and many persons have seen roofs blow *off*). In other words, wind loads are *negative* (that is, they create *suction*). This is now well recognized, and insurance companies are well informed on the subject.

In regions where hurricanes are not common, the insurance companies recommend that an uplift allowance of 30 pounds per square foot be provided for, acting at right angles to the slope of the roof. This will include the slight *positive* pressure commonly found inside buildings during windstorms. If eaves are wide and overhanging, it is recommended that a gross lifting force of 45 pounds per square foot be provided for. If the weight of the roof itself is appreciable, it may be deducted from the gross uplift allowance.

It should be recognized that a maximum snow load plus a maximum wind load could hardly occur at the same time. Thus, it cannot be considered compensating.

Definitions

In the shop and in the field, the fitness of any species of wood for a given purpose depends on various properties. When treating the strength, stiffness, hardness, and other properties of wood, many technical terms are used. For an understanding of these terms, the following definitions are worth noting:

- *Bending forces*—Forces that act on some members of a structural frame. They tend to deform them by flexure.
- *Brittleness*—Breaking easily and suddenly, usually with a comparatively smooth fracture. The opposite of *toughness* (sometimes incorrectly called *brashness*) that refers more to brittleness. Old and extremely dry wood is inclined to brashness. Green or wet wood is tougher, though not as strong in most cases.
- *Compression*—The effect of forces that tend to reduce or shorten the dimensions of a member.
- *Deformation*—A change of shape or dimension. Disfigurement, such as the elongation of a structural member under tension.

- *Ductile*—A term not applicable to wood. It is the property of a metal that allows it to be hammered thin or drawn into wires.
- *Elastic limit*—The greatest stress that a substance can withstand and still recover completely when the force or strain is removed.
- *Factor of safety*—The ratio between the stress at failure and allowable design stress. If the stress at failure in a bending beam is 4350 pounds per square inch (psi), and the timber is stress-graded at 1450f (allowable fiber stress in psi) the safety factor is $^{4350}/_{1450}$, or 3.
- *Force*—In common parlance, a *pull* or a *push* that would change the state of a body at rest, or would change the course of a body in motion (Newton's Law).
- *Load*—Pressure acting on a surface, usually caused by the action of gravity.
- *Member*—A part of a structure (such as a column, beam, or brace) that usually is subjected to compression, tension, shear, or bending.
- *Modulus, or coefficient, of elasticity*—Stress, tension, or compression, divided by the elongation or contraction per unit of length. Inside the elastic limit, the modulus of elasticity is approximately constant for most materials. In wood, it varies greatly with different species, with moisture content, and even in pieces sawed from the same log. In other words, it is subject to considerable natural variation.
- *Modulus of rupture*—The calculated fixed stress in a beam at the point of rupture. Since the elastic limit will have been passed at this point, it is not a true fiber stress, but it is a definite quantity, and the personal factor is not involved when obtaining it.
- *Permanent set*—When a member (either metal or wood) is stressed beyond its elastic limit, or subjected to stresses that may be far below its elastic limit for extremely long periods, it may take on permanent deformation. Permanent set in timbers does not mean or imply that the timbers have been weakened.
- *Resilience*—Synonym for *elasticity*. The property that enables a substance to spring back when a deforming force is removed.
- *Shear*—The effect of forces (external or internal) that causes bodies or parts of bodies to slide past each other.

- *Strain*—Alterations in the form of a member caused by forces acting on the member.
- *Strength*—The power to resist forces (which may be tensile, compressive, or shearing) without breaking or yielding.
- *Stress*—Distributed forces, such as pounds per square inch or tons per square foot. Within the elastic limit of materials, *stress* is approximately proportional to *strain*. This statement is called Hooke's Law.
- *Tenacity*—Synonym for *tensile strength*. The power to resist tearing apart.
- *Tension*—A force that tends to tear a body apart or elongate it.
- *Toughness*—Strong but flexible. Not brittle. Nearly the same as *tenacity*.
- *Ultimate strength*—The stress developed just before failure is evident.
- *Yield point*—This property is not evident in timbers. In steels, it occurs after the elastic limit has been passed. In materials that show no defined yield point, it may be arbitrarily assumed or defined as the stress where a permanent set occurs.

As an example of the uses for these terms, the tie rod in a truss resists being pulled apart because of its *tensile strength*. The *stress* thus applied *strains* the rod, *deforming* or *elongating* it. It is *stretched* and a *contraction* of the area of its cross-section results. If the *load* is not sufficient to *stress* the material past its *elastic limit*, the rod will return to its original length when the load is removed, depending on the duration of the load. If the *load* is heavy enough to stretch the rod past its *elastic limit*, it will not return to its original length when the *load* is removed, but it will remain *permanently set* if it is not pulled in two. The *elastic limit* is reached when *elongation* becomes proportionally greater than the *loading*. If the *load* is increased to the point where the rod breaks, or where its *tenacity* is overcome, it is *ruptured*. The rod itself is called a *member* of the truss.

Summary

In most shop work and in the field, the fitness of any species of wood for a given purpose depends on various properties. When treating the strength, stiffness, hardness, and other properties of wood, many factors must be considered. Stiffness is that quality possessed by a beam or other timber to resist the action of a bending force.

A beam is a single structural member (usually horizontal) that carries a load over a given space. The allowable loads on freely supported wood beams of any size can be calculated if the timber is stress-rated and the allowable fiber-stress is known. In many designs, the load is distributed uniformly over the length of the beam. Although this type of loading is actually a series of concentrated loads, it is often found in actual practice and is usually considered as being uniformly distributed.

Review Questions

1. What is meant by brittleness, bending forces, compression, and elastic limit?
2. What is meant by working stresses?
3. Explain shearing stresses.
4. Why is it important in some cases to distribute the load on a beam between the supports?
5. What is a cantilever beam? Explain its purpose.
6. What is a truss?
7. Where is compression experienced in construction?
8. When are shearing stresses in wood dangerous?
9. What are external shears?
10. What type of tension is so dangerous in reinforced concrete beams?

Chapter 5

Mathematics for Carpenters and Builders

An elementary knowledge of mathematics is essential to the carpenter to solve successfully the numerous problems encountered in almost any branch of carpentry. The branches of mathematics of which the carpenter should possess at least an elementary knowledge are the following:

- Arithmetic
- Geometry
- Trigonometry

Such knowledge will be found very useful, especially in making up estimates, solving steel square problems, and so forth.

Arithmetic

By definition *arithmetic* is the science of numbers and the art of reaching results by their use (see Figure 5-1).

Arithmetic Alphabet

In arithmetic figures are used to represent quantities or magnitudes, as follows:

Cipher	*one*	*two*	*three*	*four*	*five*	*six*	*seven*	*eight*	*nine*
0	1	2	3	4	5	6	7	8	9

A number is one or a collection of these figures to represent a definite quantity or magnitude as 1, 21, 517, 43,988, and so forth.

There are various kinds of numbers: simple, compound, integer, abstract, concrete, odd, even, prime, composite, and so forth.

Notation and Numeration

By definition, *notation* in arithmetic is *the writing down of figures to express a number,* and numeration is *the reading of the number or collection of figures already written.* By means of the ten figures given above any number can be expressed.

Figures have two values: simple and local. The *simple value* of a figure is its value when standing in units' place. The *local value* of a figure is the value that arises from its location.

<table>
<tr><td colspan="2"><i>Notation</i>
Basis—1—unit
Arithmetic Alphabet
0 1 2 3 4 5 6 7 8 9</td></tr>
<tr><td><i>Increased</i>
<i>By tens</i>
1, 10, 100, 1000, etc.
By varying scales
1 oz 1 lb 1 cwt
1 pt 1 qt 1 gal
1 in. 1 ft 1 yd</td><td><i>Diminished</i>
<i>By tens</i>
1, .1, .01, .001, etc.
By varying scales
$^{1}/_{4}$ $^{6}/_{7}$ $^{1}/_{3}$ oz $^{11}/_{23}$ etc.
$^{1}/_{6}$ lb $^{9}/_{8}$ oz $^{3}/_{8}$ cwt
etc.</td></tr>
<tr><td colspan="2"><i>According to the Four Ground Rules</i></td></tr>
<tr><td>Addition
Subtraction</td><td>Multiplication
Division</td></tr>
<tr><td>By involution (powers)</td><td>By evolution (roots)</td></tr>
<tr><td colspan="2"><i>Relations Expressed by</i></td></tr>
<tr><td colspan="2">Ratios 2 : 3 5 : 6 8 : 9 etc.
Proportion (equality of ratios) 2 : 3 : : 4 : 6 etc.</td></tr>
<tr><td colspan="2"><i>Practical Applications</i>
Percentage, interest, profit and loss, reduction of weights and measures, measuring, etc.</td></tr>
</table>

Figure 5-1 Scheme of arithmetic.

When one of the figures stands by itself, it is called a *unit*. If two of them stand together, the right-hand one is still called a unit, and the left-hand one is called *tens*. Thus, 79 is a collection of 9 units and 7 sets of ten units each, or of 9 units and 70 units, or of 79 units and is read as “seventy-nine.”

If three of them stand together, then the left-hand one is called *hundreds*. Thus 279 is read “two hundred seventy-nine.”

To express larger numbers other orders of units are formed. The figure in the fourth place denotes *thousands*, and in the fifth place denotes *ten thousands*. These are called *units of the fifth order*. The sixth place denotes hundred thousands, the seventh place denotes millions, and so forth.

The French method (which is the same as that used in the United States) of writing and reading large numbers is shown in Figure 5-2. This system is called Arabic notation and is the system in ordinary everyday use.

Names of periods	Billions			Millions			Thousands			Units				Thousandths		
Order of Units	Hundred-billions	Ten-billions	Billions	Hundred-millions	Ten-millions	Millions	Hundred-thousands	Ten-thousands	Thousands	Hundreds	Tens	Units	Decimal point	Tenths	Hundredths	Thousandths
	8	7	6,	5	4	3,	2	0	1,	2	8	2,	•	4	8	9

Figure 5-2 Numeration.

Roman Notation

This system is occasionally used for chapter headings, corner stones, and so forth. The method of expressing numbers is by letters, as shown in Table 5-1.

In the Roman notation, when any character is placed at the right-hand side of a larger numeral, its value is added to that of such numeral: VI is V + I; XV is X + V; MD, is M + D; and so on. I, X, and rarely C are also placed at the left-hand side of other and larger numerals, and when so situated, their value is subtracted from such numerals: IV is V – I; XC is C – X; and so on. Formerly, the smaller figure was sometimes repeated in such a position twice, with its value being in such cases subtracted from the larger: IIX is X – II; XXC is C – XX; and so on. The custom of using the Roman notation for chapter numbers, year of copyright, sections, and so forth continues!

Definitions

Following are some important definitions worth noting:

- *Arithmetic*—The art of calculating by using numbers.
- *Number*—A total, amount, or aggregate of units. By counting the units, we arrive at a certain number (such as *two* horses or *five* dozen).
- *Unit*—May mean a single article, but often it means a definite group adopted as a standard of measurement (such as *dozen, ton, foot, bushel*, or *mile)*. Most commonly used units are standardized and are defined and fixed by law.

Table 5-1 Roman Numeral Values

Roman Numeral	*Value*
I	1
II	2
III	3
IV	4
V	5
VI	6
VII	7
VIII	8
IX	9
X	10
XI	11
XII	12
XIII	13
XIV	14
XV	15
XVI	16
XVII	17
XVIII	18
XIX	19
XX	20
XXX	30
XL	40
L	50
LX	60
LXX	70
LXXX	80
XC	90
C	100
D	500
M	1000
$\overline{X}$	10,000
$\overline{M}$	1 million

- *Concrete number*—A number applied to some particular unit (such as *ten* nails, *two* dozen eggs, *six* miles).
- *Abstract number*—One that is *not* applied to any object or group (such as simply *two, four, ten)*.
- *Notation*—The art of expressing numbers by figures or letters. Our system of notation is the Arabic notation. The Roman notation uses letters (such as V for 5, X for 10).
- *Cardinal numbers*—Numbers used in simple counting or in reply to the question "How many?" Any number may be a cardinal number.
- *Ordinal numbers*—Indicate succession or order of arrangement (such as *first, second, tenth)*.
- *Integer*, or *integral number*—A whole number, not a fraction or part.
- *Even number*—Any number that can be exactly divided by 2 (such as 4, 16, 96, 102).
- *Odd number*—Any number that is *not* exactly divisible by 2 (such as 3, 15, 49, 103).
- *Factor of a number*—A whole number that may be exactly divided into the number. For example, 3 is a factor of 27, 13 is a factor of 91.
- *Prime number*—A number that has no factors other than itself and 1. Thus, 3, 5, 7, and 23 are prime numbers.
- *Composite number*—A number that has factors other than itself and 1 (such as 8, 49, and 100).
- *Multiple of a number*—A number that is exactly divisible by a given number. For example, 91 is a multiple of 7, 12 is a multiple of 3.
- *Digit*—Any number from 1 to 9, and usually 0.

Signs of Operation

Following are some important signs of operation worth noting:

- The *sign of addition* is +, and it is read "plus" or "add." Thus, 7 + 3 is read "seven plus three." The numbers may be taken in *any order* when adding (7 + 3 is the same as 3 + 7).
- The *sign of subtraction* is –, and it is read "minus." A series of subtractions *must* be taken in the order written (11 – 7 is *not* the same as 7 – 11).

- The *sign of multiplication* is ×, and it is read "times" or "multiplied by." The numbers may be taken in *any order* when multiplying (4 × 7 is the same as 7 × 4).
- The sign of division is ÷, and it is read "divided by." A series of divisions *must* be taken in the order written (100 ÷ 2) ÷ 10 = 5.
- The *sign of equality* is =, and it is read "equals" or "is equal to." The expressions on each side of an equality sign must be numerically the same. The complete expression is called an *equation*.

Use of the Signs of Operation

Following are some examples of the use of mathematical signs of operation.

Example *Use of the sign of addition.* A builder when building a house buys 1762 board feet of lumber from one yard, 2176 board feet from another, and 276 board feet from another. How many board feet did he buy?

The problem:

$$1762 + 2176 + 276 = ?$$

The solution:

$$\begin{array}{r} 1762 \\ 2176 \\ \underline{+\ 276} \\ 4214 \end{array} \text{ board feet}$$

Note how the numbers are aligned to permit addition. The units are all aligned on the right, then the tens, then the hundreds, then the thousands, each in the proper column.

Example *Use of the sign of subtraction.* A carpenter bought 300 pounds of nails for a job, and he had 28 pounds left when he finished. How many pounds of nails did he use to complete the job?

The problem:

$$300 - 28 = ?$$

The solution:

$$\begin{array}{r} 300 \\ \underline{-\ 28} \\ 272 \end{array} \text{ pounds}$$

Note that the numbers must be aligned as they were for addition.

Example *Use of the sign of division.* A carpenter's pickup truck gets an average of 17 miles per gallon of gasoline. How many gallons of gasoline would be required for him to travel 2040 miles in his truck?

The problem:

$$2040 \div 17 = ?$$

The solution:

$$
\begin{array}{r}
120\\
17\overline{)2040}\\
\underline{17}\\
34\\
\underline{34}\\
00
\end{array}
$$

Example *Use of the sign of equality.* A road contractor finds that he can lay the same amount of paving in 12 days using a 6-man crew that he can lay in 8 days using a 9-man crew. Express this statement as an equation.

$$12 \times 6 = 9 \times 8 \text{ or } 72 = 72$$

Fractions

A fraction indicates that a number or a unit has been divided into a certain number of equal parts, and shows how many of these parts are to be considered. Two forms of fractions are in common usage: the *decimal* (which is expressed in the tenths system) and the *common fraction*. The common fraction is written by using two numbers, one written over or alongside the other with a line between them. The lower (or second) number, called the *denominator*, indicates the number of parts into which the unit has been divided, and the upper (or first) number, called the *numerator*, indicates the number of parts to be considered. In the fraction $^2/_3$, the denominator shows that the unit is divided into 3 parts, whereas the numerator indicates that 2 parts are being considered (see Figure 5-3).

If the quantity indicated by the fraction is less than 1 (such as $^1/_2$, $^3/_4$, or $^5/_6$), it is called a *proper fraction*. If the quantity indicated by the fraction is equal to or greater than 1 (such as $^3/_3$, $^5/_4$, or $^7/_6$), it is called an *improper fraction*. When a whole number and a proper fraction are combined (such as $2^1/_4$, $6^1/_2$), it is called a *mixed number*.

Addition of Fractions

Fractions cannot be added without first reducing them to a common denominator.

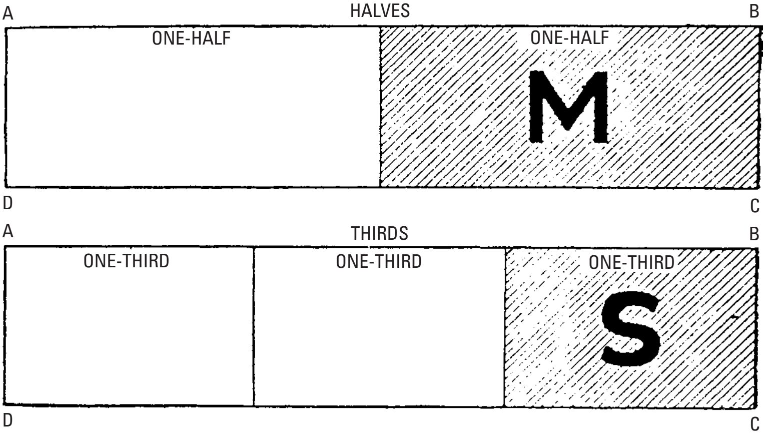

Figure 5-3 Graphic representations of fractional parts. The figures show a rectangle ABCD, representing a unit divided into two equal parts or halves and into three equal parts or thirds. The shaded section M (or one half) is larger than the shaded section S (or one third).

Example Add $^3/_4 + ^2/_9 + ^2/_3 + ^7/_{12}$

To find the common denominator, place the denominators in a row, separated by dashes. Divide them by a prime number that will divide into at least two of them without leaving a remainder, and bring down the denominators with the dividends that did not contain the divisor without a remainder. Repeat this process as often as possible until there are no two numbers remaining that can be divided by the same number. Then, multiply the divisors and the remainders together, and the result will be the smallest common denominator.

The solution:

$$
\begin{array}{ll}
2 &)\,\underline{4-9-3-12} \\
3 &)\,\underline{2-9-3-6} \\
2 &)\,\underline{2-3-1-2} \\
 & \;\;1-3-1-1
\end{array}
$$

The common denominator will then be $2 \times 3 \times 2 \times 3 = 36$. The fractions are then reduced to the common denominator of 36 by multiplying the numerator and the denominator by the same number that will produce 36 in the denominator, as follows:

$$^{3}/_{4} = {}^{3}/_{4} \times {}^{9}/_{9} = {}^{27}/_{36}$$

$$^{2}/_{9} = {}^{2}/_{9} \times {}^{4}/_{4} = {}^{8}/_{36}$$

$$^{2}/_{3} = {}^{2}/_{3} \times {}^{12}/_{12} = {}^{24}/_{36}$$

$$^{7}/_{12} = {}^{7}/_{12} \times {}^{3}/_{3} = {}^{21}/_{36}$$

The sum of the fractions is as follows:

$$\frac{27 + 8 + 24 + 21}{36} = {}^{80}/_{36} \text{ or } {}^{20}/_{9}$$

Multiplication of Common Fractions

A fraction may be multiplied by a whole number by multiplying the numerator of the fraction by that number.

Example If $^{3}/_{4}$ of a keg of nails is used for siding a garage, how many kegs of nails will be used when siding eight similar garages?

The solution:

$$8 \times {}^{3}/_{4} = {}^{24}/_{4}$$

A fraction may be simplified by dividing both the numerator and the denominator by the same number, and its value will not be affected.

Example Divide both the numerator and the denominator of the improper fraction $^{24}/_{4}$ by 4.

The result will be $^{6}/_{1}$ or 6. Therefore, six kegs of nails will be required to put the siding on the eight garages.

Fractions may be multiplied by fractions by multiplying their numerators together and their denominators together.

Example Multiply

$$^{2}/_{5} \times {}^{1}/_{4} \times {}^{5}/_{12} = \frac{2 \times 1 \times 5}{5 \times 4 \times 12} = \frac{10}{240} = \frac{1}{24}$$

Multiplication of Fractions by Cancellation

This may readily be done because any factor below the line may be divided by any factor above the line, and any factor above the line may be divided by any factor below the line, without altering the overall value of the expression.

Example The problem:

$$^{12}/_{30} \times {}^{14}/_{56} \times {}^{10}/_{24}$$

Now write it in this form:

$$\begin{array}{c} 1 \times 1 \times 1 \\ \cancel{12} \times \cancel{14} \times \cancel{10} \\ \hline \cancel{30} \times \cancel{56} \times \cancel{24} \\ 3 \times 4 \times 2 \end{array}$$

The 30 below the line may be divided by the 10 above the line with the result of 3. The 56 below the line may be divided by the 14 above the line with the result of 4. The 24 below the line may be divided by the 12 above the line with the result of 2. The result of the cancellation, then, is as follows:

$$\frac{1}{3 \times 4 \times 2} = \frac{1}{24}$$

Division of Fractions

Fractions may be divided by whole numbers by dividing the numerator by that number or by multiplying the denominator by that number.

Example

$$\frac{7}{8} \div 7 = \frac{1}{8}$$

$$\frac{7}{8} \div 7 = \frac{7}{56} = \frac{1}{8}$$

Fractions may be divided by fractions by inverting the divisor and multiplying.

Example

$$\frac{7}{8} \div \frac{2}{7} = \frac{7}{8} \times \frac{7}{2} = \frac{49}{16}$$

which is the mixed number $3^1/_{16}$.

Subtraction of Fractions

Fractions cannot be subtracted from fractions without first reducing them to a common denominator, as is done for the addition of fractions.

Example The problem:

$$^5/_6 - {}^{13}/_{16}$$

Finding the least common denominator,

$$\begin{array}{r|l} 2 & 16 - 6 \\ \hline & 8 - 3 \end{array} \qquad 2 \times 8 \times 3 = 48$$

$$\frac{13}{16} = \frac{13}{16} \times \frac{3}{3} = \frac{39}{48}$$

$$\frac{5}{6} = \frac{5}{6} \times \frac{8}{8} = \frac{40}{48}$$

$$\frac{40 - 39}{48} = \frac{1}{48}$$

To subtract a mixed number from another mixed number, it is usually most convenient to reduce both numbers to improper fractions and then proceed as shown in the last example.

To subtract a mixed number from a whole number, borrow 1 from the *minuend* (or upper number), and reduce the 1 to an improper denominator of the fraction in the *subtrahend* (or lower number), thereby reducing the whole number by 1. Then make the subtraction in the normal manner.

Example The problem:

$$14 - 6^7/_8$$

The solution:

$$14 - 6^7/_8 = 13^8/_8 - 6^7/_8 = 7^1/_8$$

Applications of Cancellation

There are countless applications where this method will save appreciable time and work, but care and thought must be given to the proper arrangement of the fractional expression if there are many factors. Also, it must be remembered that if addition or subtraction signs appear, cancellation *may not* be used.

Example A circular saw has 75 teeth with a 1-inch spacing between each tooth. To do satisfactory work, the rim of the saw should travel at approximately 9000 feet per minute. How many revolutions should this saw make per minute? (Hint: 75 inches = $^{75}/_{12}$ feet.) The solution:

$$\frac{\overset{120}{9\cancel{0}\cancel{0}\cancel{0}} \times 12}{7\cancel{5}} = 120 \times 12 = 1440 \text{ rpm}$$

Example If you go to a bank and borrow $1000 to purchase a truck, how much will the interest be, at 6 percent per annum (a theoretical example only), on the money you borrow for 1 year and 3 months? (Hint: 1 year 3 months = 15 months.)

The solution:

$$\frac{\cancel{6} \times \overset{\overset{5}{1\cancel{0}}}{1\cancel{0}\cancel{0}\cancel{0}} \times 15}{1\cancel{0}\cancel{0} \times \underset{\cancel{2}}{1\cancel{2}}} = 5 \times 15 = \$75$$

Example If eight men in fifteen 8-hour days can throw 1000 cubic yards of gravel into wheelbarrows, how many men will be required to throw 2000 cubic yards of gravel into wheelbarrows in 20 days of 6 hours each?

The solution:

$$\frac{\overset{\cancel{3}}{\cancel{15}} \times \overset{2}{\cancel{8}} \times \overset{4}{\cancel{8}} \times \overset{2}{\cancel{2000}}}{\underset{\cancel{4}}{\cancel{20}} \times \underset{\cancel{2}}{\cancel{6}} \times \cancel{1000}} = 2 \times 4 \times 2 = 16 \text{ men}$$

Example A building that is 30 feet × 30 feet with a 10-foot ceiling contains approximately 700 pounds of air. What will be the weight of the air in a room 120 feet long, 90 feet wide, and 16 feet high?

The solution:

$$\frac{\overset{4}{\cancel{120}} \times \overset{3}{\cancel{90}} \times 16 \times \overset{70}{7\cancel{0}\cancel{0}}}{\cancel{30} \times \cancel{30} \times \cancel{10}} = 4 \times 3 \times 16 \times 70 = 13{,}440 \text{ pounds}$$

Decimals

Decimal means numbering that proceeds by *tens*, and *decimal fractions* (usually simply called *decimals*) are formed when a unit is divided into ten parts. When decimals are written, the point where the numbers start is called the *decimal point*. To the left of the decimal point, the numbers read in the regular manner—units, tens, hundreds, thousands, ten thousands, hundred thousands, millions, and so forth. To the right of the decimal point, the figures are fractional, reading, from the point, tenths, hundredths, thousandths, ten thousandths, hundred thousandths, millionths, and so forth. (see Figure 5-4).

The common fraction $^{6}/_{10}$ can be expressed decimally as .6, and the fraction $^{105}/_{1000}$ can be written as .105. The mixed number $106\,^{6}/_{100}$ may be expressed decimally as 106.06. The decimal .6 is read "six-tenths," the same as the common fraction $^{6}/_{10}$, and the decimal 106.06 is read "one hundred six *and* six-hundredths," the same as the mixed number.

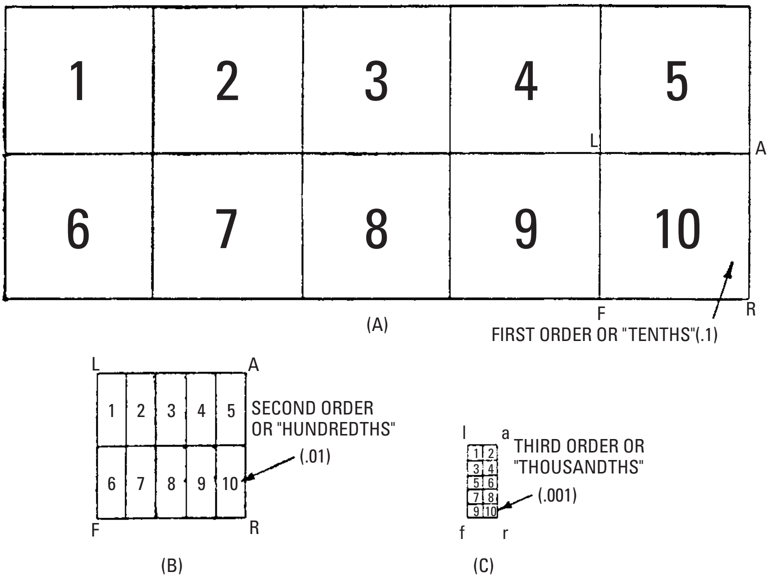
1
2
3
4
5
L
A
6
7
8
9
10
F
R
(A)
FIRST ORDER OR "TENTHS"(.1)
L
A
1
2
3
4
5
SECOND ORDER OR "HUNDREDTHS"
(.01)
6
7
8
9
10
F
R
(B)
l
a
1 2
3 4
5 6
7 8
9 10
THIRD ORDER OR "THOUSANDTHS"
(.001)
f
r
(C)

Figure 5-4 Graphic representation of decimal fractions. (A) A unit divided into ten parts—1st order or tens. (B) One of the tens as LARF, divided into ten parts—2nd order or hundredths. (C) One of the hundredths as larf, divided into ten parts—3rd order or thousandths. Similarly the process of division may be continued indefinitely.

It is not necessary to place a zero before the decimal point, as 0.06, but it is sometimes convenient when it is necessary to align a column of decimals for addition. The decimal points must *always* be aligned for addition, as shown in the following example:

$$\begin{array}{r} 0.6 \\ 6.29 \\ +10.72 \\ \hline 17.61 \end{array}$$

The position of the decimal point in the sum is established directly under the column of decimal points above the line.

Note

The number 327 is *not* read "three hundred *and* twenty-seven," but "three hundred twenty-seven." However, the decimal 300.27 is read "three hundred *and* twenty-seven hundredths." The decimal .327 is read "three hundred twenty-seven thousandths," whereas the decimal 300.027 is read "three hundred *and* twenty-seven thousandths."

Reduction of Common Fractions to Decimals

Divide the numerator by the denominator, adding zeros and carrying the division to as many decimal places as are necessary or desirable (see Figure 5-5).

Example Reduce the common fraction $^{21}/_{32}$ to a decimal.

$$
\begin{array}{rl}
 & 0.65625 \\
32 & \overline{)21.00000} \\
 & \underline{192} \\
 & 180 \\
 & \underline{160} \\
 & 200 \\
 & \underline{192} \\
 & 80 \\
 & \underline{64} \\
 & 160 \\
 & \underline{160} \\
 & 00
\end{array}
$$

Count off as many decimal places in the quotient as those in the dividend *exceed* those in the divisor. The quotient is .65625.

Subtraction of Decimals

Align the decimal points in the minuend and the subtrahend as shown for addition, and proceed as explained in subtraction of whole numbers. The decimal point in the remainder is placed in exact alignment with those decimal points above the line.

Example The problem:

$$167.02 - 27.267$$

The solution:

$$
\begin{array}{r}
167.020 \\
\underline{-\ 27.267} \\
139.753
\end{array}
$$

Note that it is necessary to add a zero to the decimal 167.02 in order to make the subtraction, but this does not change its value ($^{20}/_{1000}$ is the same value as $^{2}/_{100}$).

					1/64	.015625
				1/32		.03125
					3/64	.046875
			1/16			.0625
					5/64	.078125
				3/32		.09375
					7/64	.109375
		1/8				.125
					9/64	.140625
				5/32		.15625
					11/64	.171875
			3/16			.1875
					13/64	.203125
				7/32		.21875
					15/64	.234875
	1/4					.250
					17/64	.265625
				9/32		.28125
					19/64	.296875
			5/16			.3125
					21/64	.328125
				11/32		.34375
					23/64	.359375
		3/8				.375
					25/64	.390625
				13/32		.40625
					27/64	.421875
			7/16			.4375
					29/64	.453125
				15/32		.46875
					31/64	.484375
1/2						.500
					33/64	.515625
				17/32		.53125
					35/64	.546875
			9/16			.5625
					37/64	.578125
				19/32		.59375
					39/64	.609375
		5/8				.625
					41/64	.640625
				21/32		.65625
					43/64	.671875
			11/16			.6875
					45/64	.703125
				23/32		.71875
					47/64	.734375
	3/4					.750
					49/64	.765625
				25/32		.78125
					51/64	.796875
			13/16			.8125
					53/64	.828125
				27/32		.84375
					55/64	.859375
		7/8				.875
					57/64	.890625
				29/32		.90625
					59/64	.921875
			15/16			.9375
					61/64	.953125
				31/32		.96875
					63/64	.984375
						1.0000

Figure 5-5 Fractions of an inch and decimal equivalents.

Multiplication of Decimals

Proceed as in multiplication of whole numbers, and count off as many decimal places in the product as there are in *both* the multiplier and multiplicand.

Example The problem:

$$1.76 \times 0.06$$

The solution:

$$\begin{array}{r} 1.76 \\ \times\ 0.06 \\ \hline 0.1056 \end{array}$$

Compound Numbers

A compound number expresses units of two or more denominations of the same kind, such as 5 yards, 1 foot, and 4 inches. The process of changing the denomination in which a quantity is expressed without changing its value is called *reduction*. Thus, 1 yard and 2 inches = 38 inches, 25 inches = 2 feet and 1 inch, and so forth, are examples of reduction. Reduction problems occur and are explained with the various measures and weights.

Reduction Descending

To reduce a compound number to a lower denomination, multiply the largest units in the given number by the number of units in the next lower denomination, and add to the product the units of that denomination in the given number. Continue this process until the original number is reduced as far as desired. For an explanation of this rule, see the following example.

Example The problem: Reduce the quantity 6 yards, 2 feet, 7 inches to inches.

The solution:

$$\begin{array}{lrl} & 6 & \text{yards} \\ \text{Multiply.........} & \underline{\times\,3} & \\ & 18 & \text{feet} \\ \text{Add..............} & \underline{+\,2} & \\ & 20 & \text{feet} \\ \text{Multiply.......} & \underline{\times\,12} & \\ & 240 & \text{inches} \\ \text{Add..............} & \underline{+\,7} & \\ \text{Total....................} & 247 & \text{inches} \end{array}$$

Reduction Ascending

To reduce a number of small units to units of larger denominations, divide the number by the number of units in a unit of the next higher denomination. The quotient is in the higher denomination and the remainder, if any, is in the lower. Continue this process until the number is reduced as far as is desired.

Example The problem: Reduce 378 inches to a quantity of yards, feet, and inches.

The solution:

12) 378
3)$\underline{31}$ feet, 6 inches remainder
10 yards, 1 foot remainder

Therefore, 378 inches = 10 yards, 1 foot, 6 inches.

Ratios

By definition, a *ratio* is the relation of one number to another as obtained by dividing the first number by the second. Thus, the ratio of 2 to 4 is expressed as 2 : 4. The symbol : is read "to" in the case of a ratio and "is to" in the case of a proportion. It is equivalent to "divided by." Hence:

$$2 : 4 = {}^{1}/_{2}$$

The first term of a ratio is the *antecedent*, and the second term is the *consequent*, thus:

$$\begin{array}{ccc} \text{antecedent} & & \text{consequent} \\ 2 & : & 4 \end{array}$$

Since a ratio is essentially a fraction, it follows that if both terms are multiplied or divided by the same number, the value of the ratio is not altered. Thus:

$$2 : 4 = 2 \times 2 : 4 \times 2 = 2 \div 2 : 4 \div 2$$

Two quantities of different kinds cannot form the terms of a ratio. Thus, no ratio can exist between $5 and 1 day, but a ratio can exist between $5 and $2 or between 1 day and 10 days.

Proportion

When two ratios are equal, the four terms form a *proportion*. A proportion is, therefore, expressed by using the sign = or : : between two ratios. Thus:

(expressed) 4 : 8 :: 2 : 4

(read) 4 is to 8 as 2 is to 4

The same proportion is also expressed as follows:

$$4/8 = 2/4$$

The first and last terms of a proportion are called the *extremes*, and the middle terms are called the *means*. Thus:

$$4 : 8 :: 2 : 4$$

The product of the extremes equals the product of the means. Thus, in proportion

$$4 : 8 = 2 : 4$$

$$4 \times 4 = 8 \times 2$$

Since the equation is not altered by dividing both sides by the same number, the value of any term can be obtained as follows:

$$\frac{4 \times 4}{4} = \frac{8 \times 2}{4}$$

$$4 = 2 \times 2 = 4$$

Rule of Three

When three terms of a proportion are given, the method of finding the fourth term is called the rule of three.

Example The problem: If 5 bundles of shingles cost \$100, what will 25 bundles cost?

The solution: Let X represent the unknown term in the proportion, and, remembering that each ratio must be made up of like quantities:

$$5 \text{ bundles} : 25 \text{ bundles} = 100\ (\$) : X\ (\$)$$

Multiply the extremes by the means:

$$5 \times X = 25 \times 100$$

$$X = \frac{25 \times 100}{5} = \$500$$

Percentage

By definition, *percentage* means the rate per 100, or the proportion in 100 parts. Therefore $1/100$ of a number is called 1 percent, $2/100$ is 2 percent, and so forth. The symbol % is read as "percent." Thus 1%, 2%, and so forth. Carefully note the following explanation with respect to the symbol %. The notation 5% means $5/100$, which, when reduced to a decimal (as is necessary when making a calculation), becomes .05, but .05% means $^{.05}/_{100}$, which, when reduced to a decimal, becomes .0005 (that is, $^{5}/_{100}$ of 1%).

If the decimal has more than two places, the figures that follow the hundredths place signify parts of 1%.

Example The problem: If the list price of shingles is \$90 per 1000, what is the net cost for 1000 shingles with a 5 percent discount for cash?

The solution: Reduce % rate to a decimal.

$$5\% = {^{5}/_{100}} = .05$$

Multiply decimal by list price.

$$\$90 \times .05 = \$4.50$$

Subtract product obtained from list price.

$$\$90 - \$4.50 = \$85.50$$

Powers of Numbers (Involution)

The word *involution* means the multiplication of a quantity by itself any number of times, and a *power* is the product arising from this multiplication. Involution, then, is the process of raising a number to a given power. The *square* of a number is its second power, the *cube* is its third power, and so forth. Thus:

$$\text{square of } 2 = 2 \times 2 = 4$$

$$\text{cube of } 2 = 2 \times 2 \times 2 = 8$$

The power to which a number is raised is indicated by a small superior figure called an *exponent*. Thus, in Figure 5-6, the exponent indicates the number of times the number, or root, has been multiplied by itself.

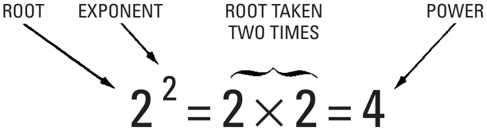

Figure 5-6 The root, exponent, and power of a number.

Roots of Numbers (Evolution)

The word *evolution* means the operation of extracting a root. The root is a factor that is repeated to produce a power. Thus, in the equation 2 × 2 × 2 = 8, 2 is the root from which the power 8 is produced. This number is indicated by the symbol √, called the *radical sign*, which, when placed over a number, means that the root of that number is to be extracted. Thus:

$\sqrt{4}$ *means that the square root of 4 is to be extracted*

The *index* of the root is a small figure that is placed over the radical sign that denotes what root is to be taken. Thus, $\sqrt[3]{9}$ indicates the cube root of 9 and $\sqrt[4]{16}$ indicates the extraction of the fourth root of 16. When there is no index given, the radical sign alone always means the *square root* is to be extracted from the number under the radical sign.

Sometimes the number under the radical sign is to be raised to a power before extracting the root, as follows:

$$\sqrt[3]{4^3} = \sqrt[3]{4 \times 4 \times 4} = \sqrt[3]{64}$$

Example The problem: Extract the square root of 186,624.

The solution:

$$\begin{array}{rl}
 & \sqrt{18'66'24} \qquad 432 \\
 & \;16 \\
83 & \sqrt{266} \\
 & \quad 249 \\
862 & \sqrt{1724} \\
 & \quad \underline{1724}
\end{array}$$

From the decimal point, count off the given number into periods of two places each. Begin with the last period counted off (18). The largest square that can be divided into 18 is 4. Put this down in the quotient, and put the square (16) under the 18. Write down the remainder (2), and bring down the next period (66). Multiply 4 (in the quotient) by 2 for the first number of the next divisor; 8 goes into 26 three times. Place 3 after 4 in the quotient and also after 8 in the divisor. Multiply the 83 by 3, placing the product 249 under 266, and subtract, obtaining the remainder 17. Bring down the last period (24), and proceed as before, obtaining 432 as the square root of 186,624.

Extracting the cube root of a number is a more complicated (though similar) process, as indicated by the following procedure:

1. Separate the number into groups of three figures each, beginning at the decimal point.

2. Find the greatest cube that can be divided into the left-hand group, and write its root for the first figure of the required root.
3. Cube this root, subtract the result from the left-hand group, and annex the next group to the remainder for a dividend.
4. For a partial divisor, take three times the square of the root already found, considered as hundreds, and divide the dividend by it. The quotient (or the *quotient diminished*) will be (or be close to) the second figure of the root.
5. To this *partial divisor*, add three times the product of the first figure of the root, considered as tens, by the second figure, and to this add the square of the second figure. This sum will be the *complete divisor*.
6. Multiply the complete divisor by the second figure of the root, subtract the product from the dividend, and annex the next group to the remainder for a new dividend.
7. Proceed in this manner until all the groups have been annexed. The result will be the cube root required, as shown in the following example.

Example The problem: Extract the cube root of the number 50,653.

The solution:

$$\begin{array}{rl} & \sqrt[3]{50'653.}\ \ (\underline{\ 37} \\ & \underline{27} \\ 2700 & 23\,653 \\ 630 & \underline{23\,653} \\ \underline{49} & \\ 3379 & \end{array}$$

Therefore, the cube root of 50,653 is 37.

Measures

To *measure* is the act or process of determining the extent, quantity, degree, capacity, dimension, volume, and so forth, of a substance by comparing it with some fixed standard, which is usually fixed by law. A measure may relate to any of these standards. There are many kinds of measures, and practically all of them are standard, but standards vary in different countries. The measures mentioned in this text are all U.S. standards unless designated otherwise. The study of measurements is sometimes called *mensuration*.

Among the many kinds of measures are the following:

- *Linear*—Measures of length
- *Square*—Used to measure areas
- *Cubic*—Used to measure volume, or volumetric contents
- *Weight*—Many systems of weights are standard
- *Time*—Almost standardized all over the world
- *Circular* or *angular*—The same all over the world

Linear Measure

Table 5-2 shows linear measurement (long) equivalents.

Table 5-2 Long Measure

Measure	*Equivalent*	*Equivalent*
12 inches	1 foot	
3 feet	1 yard	36 inches
$5^1/_2$ yards	1 rod	$16^1/_2$ feet
40 rods	1 furlong	660 feet
8 furlongs	1 mile	5280 feet
3 miles	1 league (land)	

The furlong is practically never used, except at racetracks and in some athletic events.

Table 5-3 shows land survey measurement equivalents.

Table 5-3 Land Surveyor's Measure

Measure	*Equivalent*	*Equivalent*
7.92 inches	1 link	
100 links	1 chain	66 feet
10 chains	1 furlong	660 feet
80 chains	1 mile	5280 feet

The use of the surveyor's chain, or Gunter's chain, was abandoned in the late 1800s and was superseded by the steel tape, which is much more accurate. The chain (meaning 66 feet) is still used by the U.S. General Land Office, however, when surveying very old deeds. The standard surveyor's tape is often called, from habit, a *chain*. It is

100 feet long and is graduated in feet except for the last foot, which is divided into tenths and hundredths of a foot.

Table 5-4 shows nautical measurement equivalents.

Table 5-4 Nautical Measure (U.S. Navy)

Measure	*Equivalent*
6 feet	1 fathom
120 fathoms	1 cable length
The International Nautical Mile*	6076.1033 feet
3 nautical miles	1 marine league

*Adopted in 1954.

The knot is a measure of speed, not of length, and is equivalent to 1 nautical mile per hour. A speed of 16 knots is equal to 16 nautical miles per hour.

Square Measure

Square measure is used to measure areas. In most (but not all) cases, linear units are used to measure the two dimensions, length and width, and their product is the area in square units. Expressed as an equation:

$$\text{length} \times \text{width} = \text{area}$$

The two dimensions, length and width, must be measured in the same units, but any unit of linear measurement may be used. If inches are multiplied by inches, the result will be in square inches; if feet are multiplied by feet, the result will be in square feet, and so forth (see Figure 5-7).

For the small areas commonly found in everyday life (such as tabletops or shelves), the unit most commonly used is the square inch. Plywood and lumber are commonly sold by the square foot. Carpets and other floor coverings and materials and ceilings are measured in square yards. The carpenter measures roofing by the square of 10 × 10 feet, or 100 square feet. Tracts of land are usually measured in acres or, for large areas, in square miles.

Table 5-5 shows square measure equivalents.

Cubic Measure

Cubic measure is used to determine or appraise volumes. Three dimensions are involved (length, width, and height) and their product is volume. Expressed as an equation:

$$\text{length} \times \text{width} \times \text{height} = \text{volume}$$

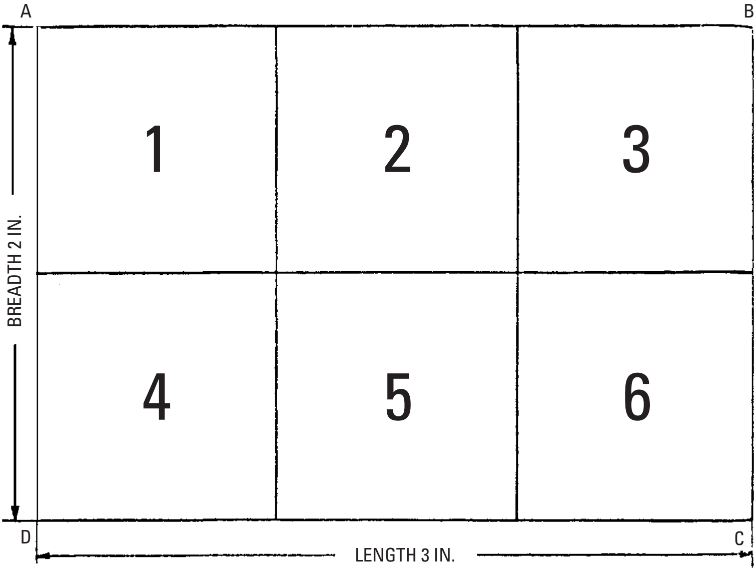

Figure 5-7 Square measure. If the rectangle ABCD measures 2 inches on one side and 3 inches on the other, and lines are drawn at each inch division, then each of the small squares will have an area of 1 square inch and the area of the rectangle will be area ABCD = breadth × length = 2 × 3 = 6 square inches.

Table 5-5 Square Measure

Measure	*Equivalent*	*Equivalent*
144 square inches	1 square foot	
9 square feet	1 square yard	
$30^1/_4$ square yards	1 square rod	272.25 square feet
160 square rods	1 acre	4840 square yards or 43,560 square feet
640 acres	1 square mile	3,097,600 square yards
36 square miles	1 township	

As with square measure, the usual linear units (inches, feet, and yards) are ordinarily used to measure these three dimensions. Most small measurements of capacity (such as small shipping cases or small cabinets) are measured in cubic inches. The contents of buildings, their cubage, are ordinarily expressed in cubic feet. Earthwork (either excavated and loose, or in place) is expressed in cubic yards (see Figure 5-8).

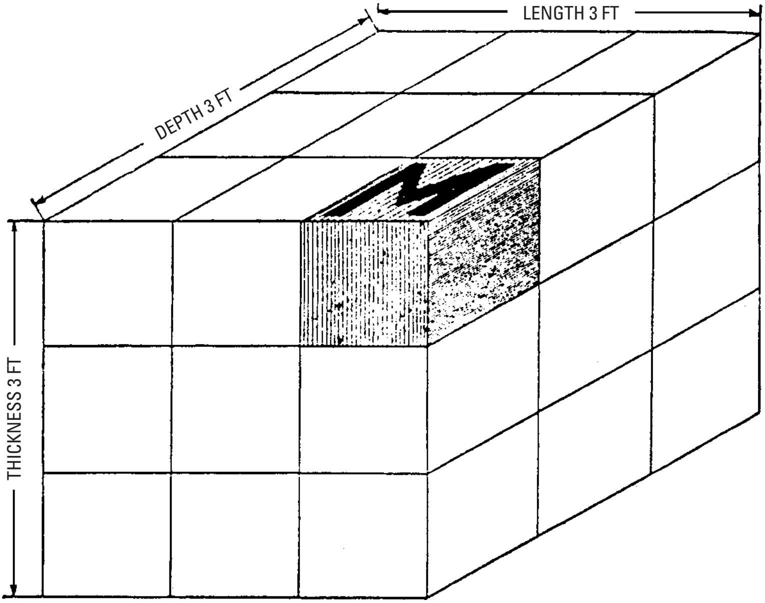

Figure 5-8 Cubic measure. If each side of the cube measures 3 feet and it is cut as indicated by the lines, each little cube as M, will have each of its sides 1 foot long and will contain 1 × 1 × 1 = 1 cubic foot. Accordingly, the large cube will contain 3 × 3 × 3 = 27 cubic feet or 1 cubic yard.

Table 5-6 shows cubic measure of volume equivalents.

Table 5-6 Cubic Measures of Volume

Measure	*Equivalent*
1728 cubic inches	1 cubic foot
27 cubic feet	1 cubic yard

Dry Measure

Quantities of loose, granular materials (such as grains, some fruits, and certain vegetables) are measured in arbitrary units that, in turn, are defined by means of cubic measures of volume, usually in cubic inches. Their value is sometimes fixed by law.

Table 5-7 shows units of dry measure equivalents for the United States and Table 5-8 shows units of dry measure equivalents for Great Britain and Canada.

Table 5-7 Dry Measure (United States)

Measure	*Equivalent*	*Equivalent*
2 pints	1 quart	67.2 cubic inches
8 quarts	1 peck	537.61 cubic inches
4 pecks	1 bushel	2150.42 cubic inches

Table 5-8 Dry Measure (British and Canadian)

Measure	*Equivalent*	*Equivalent*
1 gallon	.5 peck	277.42 cubic inches
4 pecks	1 bushel	2219.23 cubic inches

The British dry quart is not often used. It is equal to 69.35 cubic inches, or 1.032 U.S. dry quarts.

The weight, rather than the volume, of grains is the standard fixed by the U.S. government (Table 5-9).

Table 5-9 Grain Measure (U.S.)

Measure	*Equivalent*
1 bushel of wheat	60 pounds
1 bushel of barley	48 pounds
1 bushel of oats	32 pounds
1 bushel of rye	56 pounds
1 bushel of corn (shelled)	56 pounds

Board or Lumber Measure

Timbers and logs are measured in *board* or *lumber measure*. The board foot is 1 foot wide, 1 foot long, and 1 inch thick, thereby containing 144 cubic inches. In the retail market, all lumber that is less than 1 inch thick is called one inch. At the sawmills, the full sizes govern the thickness of the saw kerfs; usually about $^1/_4$ inch is allowed for and accounted as sawing loss. Actual finished (dressed) sizes of common lumber and the dimension and timbers for pine are as follows:

- The standard dressed thickness of 1-inch boards is $^3/_4$ inch.
- The standard thickness of 2-inch dimension boards is $1^1/_2$ inches.
- The standard dressed widths of lumber 2 inches thick and less are $^1/_2$ inch less for widths under 8 inches and $^3/_4$ inch less for 8-inch widths and wider.
- The standard dressed widths and thicknesses for lumber and timbers are $^1/_2$ inch less both ways under 8 inches wide and$^3/_4$ inch for 8-inch widths and over. So, a 2-inch × 8-inch board would be $1^1/_2$ inches × $7^1/_4$ inches. A 2-inch × 10-inch board would be $1^1/_2$ inches × $9^1/_4$ inches.

Liquid Measure

Liquid measure is used to measure various liquids such as oils, liquors, molasses, and water. Table 5-10 shows liquid measure equivalents.

Table 5-10 Liquid Measure

Measure	*Equivalent*	*Equivalent*
4 gills	1 pint	28.875 cubic inches
2 pints	1 quart	57.75 cubic inches
4 quarts	1 gallon	231 cubic inches

There is no legal standard barrel in the United States. By custom, a barrel of water is understood to be $31^1/_2$ gallons. The British barrel is generally 36 Imperial gallons. Crude oil is often disposed of at the wells in barrels of 50 gallons, whereas refined oils are marketed in barrels of 48 gallons. Owing to this lack of uniformity, it is safest to specify “barrels of 50 gallons,” or something of that nature, to avoid misunderstanding. The barrel is sometimes used as a dry measure unit of varying value. For Portland cement, 4 bags = 1 barrel = 4 cubic feet = 376 pounds.

Measures of Weight

The simplest definition of *weight* is the force with which a body is attracted toward the earth. It is a quantity of heaviness. The three systems (or standards) of weights used in the United States are:

- *Avoirdupois*—Used for almost all ordinary purposes (see Table 5-11)

Table 5-11 Avoirdupois Weights

	Measure	*Equivalent*
U.S.	16 drams	1 ounce
	16 ounces	1 pound
	100 pounds	1 hundredweight
	20 hundredweights	1 ton
	Measure	Equivalent
England	14 pounds	1 stone
	112 pounds	1 hundredweight
	20 hundredweight	1 ton
	2240 pounds	

Note: The 2240-pound ton is sometimes used in the United States for weighing coal at the mines and at Customs houses for evaluating shipments from England.

- *Troy*—Used in weighing precious metals and jewels (see Table 5-12)
- *Apothecaries*—Used by pharmacists when compounding drugs (see Table 5-13)

Table 5-12 Troy Weights

Measure	*Equivalent*
3.086 grains	1 carat
24 grains	1 pennyweight
20 pennyweights	1 ounce
12 ounces	1 pound

Table 5-13 Apothecaries Weights

Measure	*Equivalent*
20 grains	1 scruple
3 scruples	1 dram
8 drams	1 ounce
12 ounces	1 pound

This standard of weights is fast becoming obsolete, although pharmacists must be familiar with it. Manufacturing pharmacists and chemists are rapidly changing to the metric weights, using the metric *gram* as a basis instead of the apothecaries' scruple (1 scruple = 1.296 grams).

Time Measure

Time is defined as measurable duration (see Table 5-14). It is the period during which an action or process continues. The basis (or standard) used in our ordinary determination of time is the *mean solar day*, beginning and ending at mean midnight. The word mean as used here simply means average. The direct ray of the sun does not move in an exact and uniform path around the equator.

Table 5-14 Time Measure

Measure	*Equivalent*
60 seconds	1 minute
60 minutes	1 hour
24 hours	1 day
7 days	1 week
30 days (commonly)	1 month
365 days	1 year
10 years	1 decade
100 years	1 century
1000 years	1 millennium

The length of an *astronomical* year is 365 days, 5 hours, 48 minutes, and 45.51 seconds, or approximately 365 $^1/_4$ days. This makes it necessary to add 1 day every 4 years, thus making the leap year 366 days.

Circular Measure

This measure is used in astronomy, land surveying, navigation, and in measuring angles of all kinds. Circles of all sizes are divisible into degrees, minutes, and seconds (see Table 5-15). Note that a degree is *not* a measurement of length. It is $^1/_{360}$ of the circumference of a circle with any radius. With widespread use of calculators and the need for accuracy in missiles, the degree has now been divided

Table 5-15 Circular Measure

Measure	*Equivalent*
60 seconds	1 minute
60 minutes	1 degree
360 degrees	1 circle

in decimal form rather than minutes and seconds. Most calculators can carry the degree out to 6 places after the decimal point

The Metric System

The base, or fundamental, unit in the metric system is the *meter*. The meter is defined as the distance between two scribed marks on a standard bar made of platinum-iridium kept in the vaults of the International Bureau of Weights and Measures, near Paris, France. Of course, many other standard meter bars have been made from the measurement on this bar. It is permissible and official to use this measurement in the United States, and, in fact, the yard, the basis for the English system of measurement, has been defined as exactly $^{3600}/_{3937}$ meter, or 1 meter = 39.37 inches.

The advantage (and immeasurably greater convenience) of the metric system over the English system of units lies in the fact that it is expressed in tenths, thereby readily allowing the use of decimals. However, the American public is accustomed to the English units, and as recent experience indicates, the system should continue for a long time. The metric system is, of course, in common use all over the world with the exception of some English-speaking countries. The meter is used like the yard to measure cloth and short distances.

Units of other denominations are named by prefixing to the word meter the Latin numerals for the lower denominations and the Greek numerals for the higher denominations, as shown in Table 5-16.

Table 5-16 Denominations

Lower Denomination		***Higher Denomination***	
Greek	***Equivalent***	***Greek***	***Equivalent***
Deci	$^1/_{10}$	Deka	10
Centi	$^1/_{100}$	Hecto	100
Milli	$^1/_{1000}$	Kilo	1000
Micro	$^1/_{1,000,000}$	Myria	10,000
		Mega	1,000,000

Therefore, 1 decimeter = $^1/_{10}$ of a meter, 1 millimeter = $^1/_{1000}$ of a meter, 1 kilometer = 1000 meters, and so forth. From this explanation of the metric prefixes, the linear equivalents shown in Table 5-17 can easily be understood.

Table 5-17 Metric Table of Linear Measure

Metric Denomination		*Meter*	*U.S. Value*
	1 millimeter	.001	.0394 inches
10 millimeters	1 centimeter	.01	.3937 inches
10 centimeters	1 decimeter	.1	3.937 inches
10 decimeters	1 meter	1.	39.3707 inches 3.28 feet
10 meters	1 dekameter	10.	32.809 feet
10 dekameters	1 hectometer	100.	328.09 feet
10 hectometers	1 kilometer	1000.	.62138 miles
10 kilometers	1 myriameter	10,000.	6.2138 miles

The kilometer is commonly used for measuring long distances. The square meter (see Table 5-18) is the unit used for measuring ordinary surfaces, such as flooring or ceilings.

Table 5-18 Metric Table of Square Measure

Measure	*Equivalent*	*Equivalent*
100 square millimeters (mm^2)	1 square centimeter	0.15 + square inch
100 square centimeters (cm^2)	1 square decimeter	15.5 + square inches
100 square decimeters (dm^2)	1 square meter (m^2)	1.196 + square yards

The acre is the unit of land measure and is defined as a square whose side is 10 meters, equal to a square dekameter, or 119.6 square yards (see Table 5-19).

Table 5-19 Metric Table of Land Measure

Measure	*Equivalent*	*Equivalent*
1 centiare (ca)	1 square meter	1.196 square yards
100 centiares (ca)	1 acre	119.6 square yards
100 ares (A)	1 hectare	2.471 acres
100 hectares (ha)	1 square kilometer	0.3861 square miles

The cubic meter is the unit used for measuring ordinary solids, such as excavations or embankments (see Table 5-20).

Table 5-20 Metric Table of Cubic Measure

Measure	*Equivalent*	*Equivalent*
1000 cubic millimeters (mm^3)	1 cubic centimeter	0.061 + cubic inches
1000 cubic centimeters (cm^3)	1 cubic decimeter	61.026 + cubic inches
1000 cubic decimeters (dm^3)	1 cubic meter	35.316 + cubic feet

The liter is the unit of capacity, both of liquid and of dry measures, and is equivalent to a vessel whose volume is equal to a cube whose edge is $^1/_{10}$ of a meter, equal to 1.0567 quarts liquid measure, and 0.9081 quart dry measure (see Table 5-21).

The hectoliter is the unit used for measuring liquids, grain, fruit, and roots in large quantities. The gram is the unit of weight equal to the weight of a cube of distilled water, the edge of which is $^1/_{100}$ of a meter, and is equal to 15.432 troy grains (see Table 5-22).

Geometry

By definition, *geometry* is that branch of mathematics that deals with space and figures in space. In other words, it is the science of the mutual relations of points, lines, angles, surfaces, and solids that are considered as having no properties except those arising from extension and difference of situation.

Lines

The two kinds of lines are straight and curved. A *straight line* is the shortest distance between two points. A *curved line* is one that changes its direction at every point. Two lines are said to be parallel when they have the same direction. A horizontal line is one parallel to the horizon or surface of the Earth. A line is perpendicular with another line when they are at right angles to each other. These definitions are illustrated in Figure 5-9.

Angles

An *angle* is the difference in direction between two lines proceeding from the same point (called the *vertex*). Angles are said to be *right* (90 degrees) when formed by two perpendicular lines (see

Table 5-21 Metric Table of Capacity

10 milliliters (ml.)	= 1 centiliter	= .0338 fluid ounce
10 centiliters (cl.)	= 1 deciliter	= .1025 cubic inch
10 deciliters (dl.)	= 1 liter	= 1.0567 liquid quart
10 liters (l.)	= 1 dekaliter	= 2.64 gallons
10 dekaliters (dl.)	= 1 hectoliter	= 26.418 gallons
10 hectoliters (hl.)	= 1 kiloliter	= 264.18 gallons
10 kiloliters (kl.)	= 1 myrialiter (ml.)	
1 myrialiter	= 10 cubic meters = 283.72 + bushels	= 2641.7 + gallons
1 kiloliter	= 1 cubic meter = 28.372 + bushels	= 264.17 gallons
1 hectoliter	= $^1/_{10}$ cubic meter = 2.8372 + bushels	= 26.417 gallons
1 decaliter	= 10 cubic decimeters = 9.08 quarts	= 2.6417 gallons
1 liter	= 1 cubic decimeter = .908 quart	= 1.0567 quart liquid
1 deciliter	= $^1/_{10}$ cubic decimeter = 6.1022 cubic inches	= .845 gallons
1 milliliter	= 10 cubic centimeters = .6102 cubic inches	= .338 fluid ounces
1 centiliter	= 1 cubic centimeter = .061 cubic inches	= .27 fluid dram

Figure 5-10A), *acute* (less than 90 degrees) when less than a right angle (see Figure 5-10B), and *obtuse* (more than 90 degrees) when greater than a right angle (see Figure 5-10C). All angles except right (or 90-degree) angles are called *oblique angles*.

Angles are usually measured in degrees (circular measure) (see Figure 5-10D). The *complement* of an angle is the difference between 90 degrees and the angle. The *supplement* of the angle is the difference between the angle and 180 degrees.

Plane Figures

The term *plane figures* means a plane surface bounded by straight or curved lines, and a *plane* (or *plane surface*) is one in which any straight line joining any two points lies wholly in the surface. Figure 5-11 defines a plane surface. There is a great variety of plane

Table 5-22 Metric Table of Weight Measure

Measure	*Equivalent*	*Equivalent*
10 milligrams (mg)	1 centigram	0.15432 + grains troy
10 centigrams (cg)	1 decigram	1.54324 + grains troy
10 decigrams (dg)	1 gram	15.43248 + grains troy
10 grams (g)	1 dekagram	0.35273 + ounce avoirdupois
10 dekagrams (Dg)	1 hectogram	3.52739 + ounces avoirdupois
10 hectograms (hg)	1 kilogram	2.20462 + pounds avoirdupois
10 kilograms (kg)	1 myriagram	22.04621 + pounds avoirdupois
10 myriagrams (Mg)	1 quintal	220.46212 + pounds avoirdupois
10 quintals	1 ton	2204.62125 + pounds avoirdupois

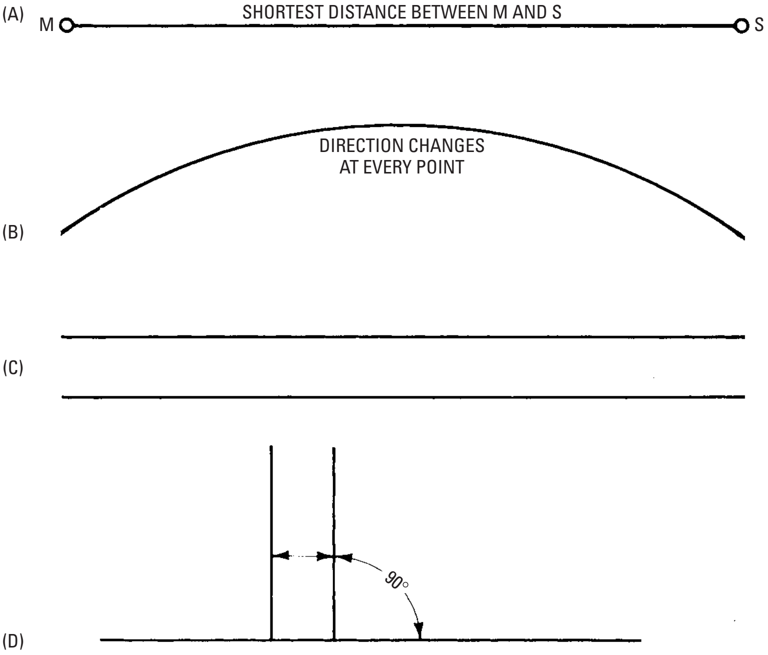
(A)
M
SHORTEST DISTANCE BETWEEN M AND S
S
DIRECTION CHANGES AT EVERY POINT
(B)
(C)
90°
(D)

Figure 5-9 Various lines: (A) straight, (B) curved, (C) parallel, and (D) perpendicular.

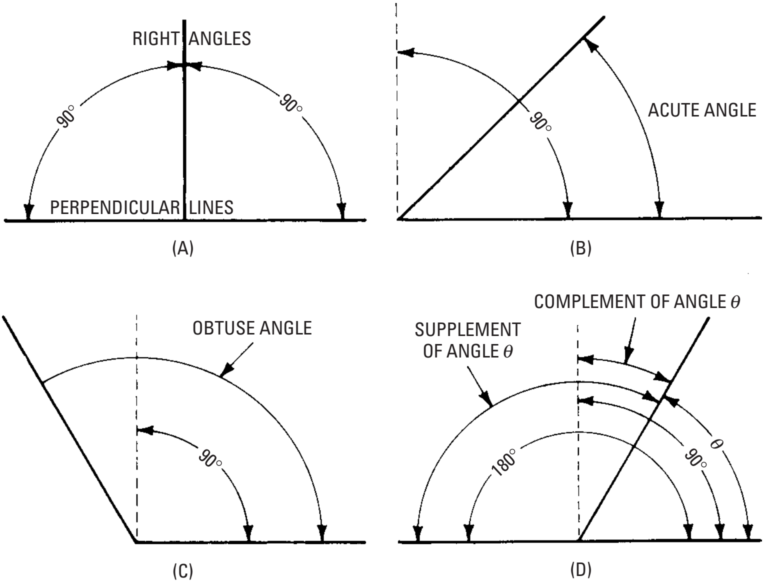
RIGHT ANGLES
90°
90°
PERPENDICULAR LINES
(A)
90°
ACUTE ANGLE
(B)
OBTUSE ANGLE
90°
(C)
COMPLEMENT OF ANGLE θ
SUPPLEMENT OF ANGLE θ
180°
90°
θ
(D)

Figure 5-10 Various angles: (A) right, (B) acute, (C) obtuse, and (D) complement and supplement of an angle.

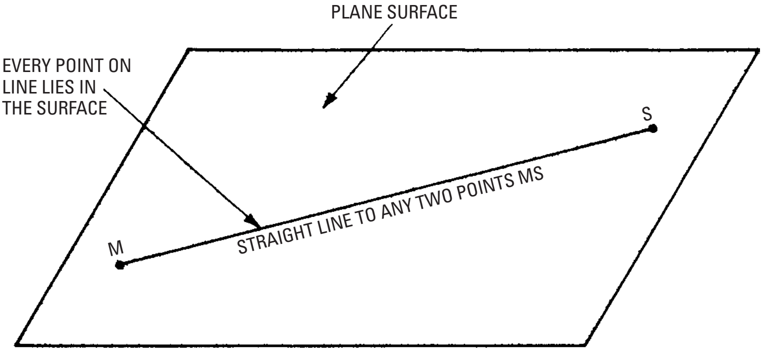
PLANE SURFACE
EVERY POINT ON LINE LIES IN THE SURFACE
S
M
STRAIGHT LINE TO ANY TWO POINTS MS

Figure 5-11 A plane surface means that every point on a straight line joining any two points in the surface lies in the surface.

figures, which are known as *polygons* when their sides are straight lines. The sum of the sides is called the *perimeter*. A regular polygon has all its sides and angles equal. Plane figures of three sides are known as triangles (see Figure 5-12), and plane figures of four sides are quadrilaterals. Figure 5-13 shows examples of these. Various plane figures are formed by curved sides and are known as circles, ellipses, and so forth, as shown in Figure 5-14. Figure 5-15 details the structure of the quadrilateral.

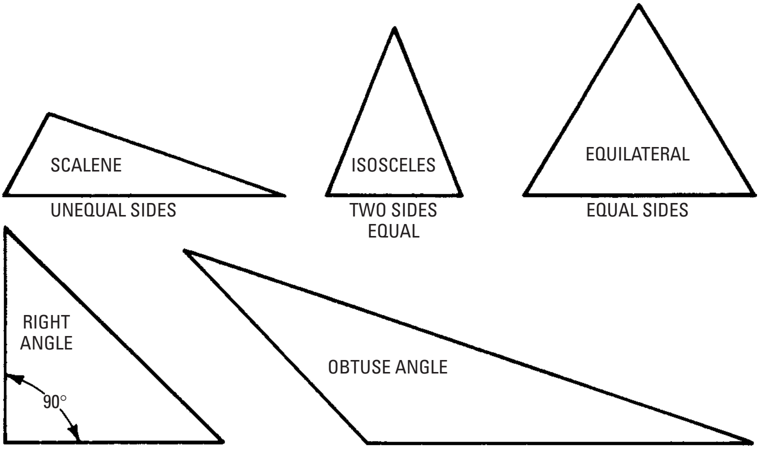

Figure 5-12 Various triangles. A triangle is a polygon having three sides and three angles.

Solids

Solids have three dimensions—length, width, and thickness. The bounding planes are called the *faces*, and the intersections are called the *edges*. A prism (see Figure 5-16) is a solid whose ends consist of equal and parallel polygons, and whose sides are *parallelograms*. The *altitude* of a prism is the perpendicular distance of its opposite sides or bases. A *parallelopipedon* is a prism that is bounded by six parallelograms; the opposite parallelograms are parallel and equal. A *cube* is a parallelopipedon whose faces are equal. One important solid is the *cylinder*, which is a body bounded by a uniformly curved surface and having its ends equal and forming parallel circles (see Figure 5-17). There are numerous other solids having curved surfaces (such as *cones* and *spheres*).

Geometrical Problems

The following problems illustrate the method in which various geometrical figures are constructed, and they should be solved by the

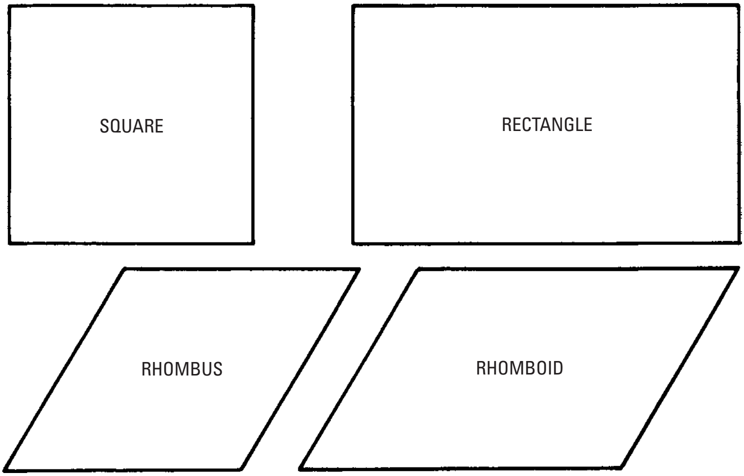
SQUARE
RECTANGLE
RHOMBUS
RHOMBOID

Figure 5-13 **Various quadrilaterals. All opposite sides of a quadrilateral are equal.**

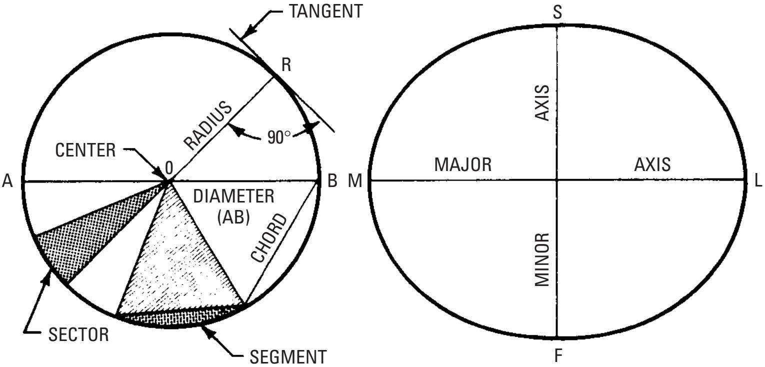
TANGENT
R
RADIUS
90°
CENTER
0
A
B
DIAMETER (AB)
CHORD
SECTOR
SEGMENT
S
AXIS
M
MAJOR
AXIS
L
MINOR
F

Figure 5-14 **Curved figures. A circle is a plane figure bounded by a uniformly curved line, every point of which is equidistant from the center point O. OR is a radius, and AB is a diameter. An ellipse is a curved figure enclosed by a curved line that is such that the sum of the distances between any point on the circumference and the two foci is invariable. ML is the major axis, and SF is the minor axis.**

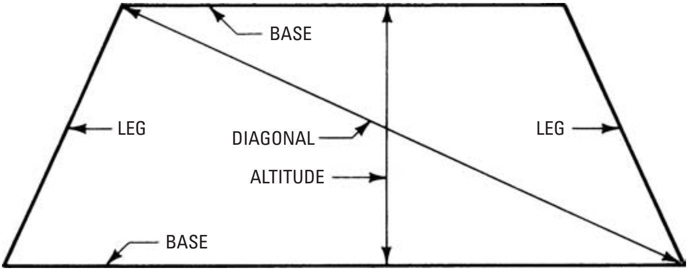
BASE
LEG
DIAGONAL
LEG
ALTITUDE
BASE

Figure 5-15 The parallel sides of a quadrilateral (four-sided polygon) are the bases. The distance between the bases is the altitude, and a line joining two opposite vertices is a diagonal.

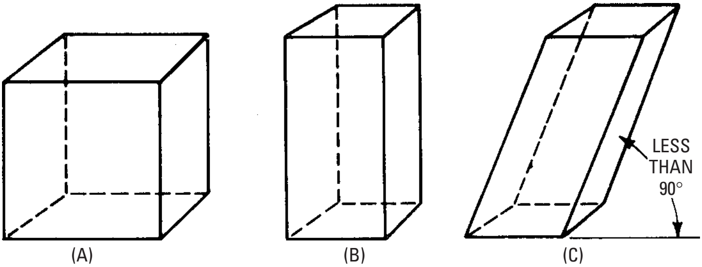
LESS THAN 90°
(A)
(B)
(C)

Figure 5-16 Various prisms: (A) cube, or equilateral parallelopipedon; (B) parallelopipedon; and (C) oblique parallelopipedon.

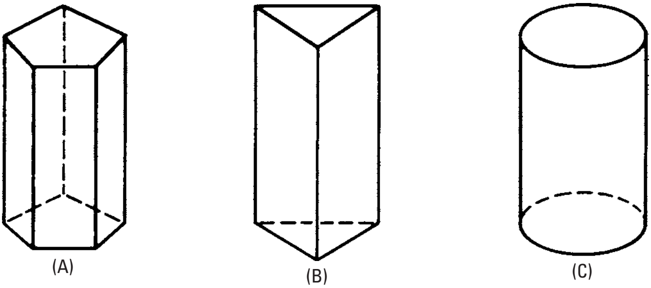

Figure 5-17 Various solids: (A) pentagonal prism, (B) triangular prism, and (C) cylinder.

use of pencil, dividers, compass, and scale. Many of these problems are commonly encountered in carpentry with layout work. Therefore, experience in working them out will be of value to carpenters and woodworkers.

Problem 1

To bisect (or divide into two equal parts) a straight line or arc of a circle.

In Figure 5-18, from the ends *A* and *B*, as centers, describe arcs cutting each other at *C* and *D*, and draw line *CD*, which cuts the line at *E*, or the arc at *F*.

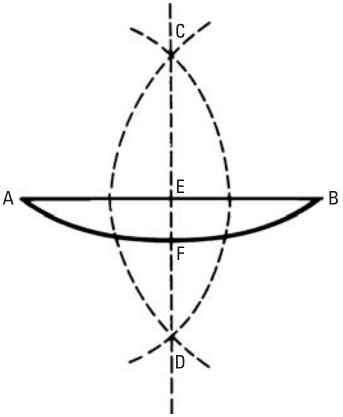

Figure 5-18 To bisect a straight line or arc of a circle.

Problem 2

To draw a perpendicular to a straight line, or a radial line to an arc.

The line *CD* is perpendicular to *AB*; also, the line *CD* is radial to the arc *AB* (see Figure 5-18).

Problem 3

To erect a perpendicular to a straight line from a given point in that line.

In Figure 5-19, with any radius from any given point *A*, in the line *BC* describe arcs cutting the line at *B* and *C*. Next, with a longer radius describe arcs with *B* and *C* as centers, intersecting at *D*, and draw the perpendicular *DA*.

Second Method

In Figure 5-20, from any point *F* above *BC*, describe a circle passing through the given point *A* and cutting the given line at *D*. Draw *DF*, and extend it to cut the circle at *E*. Draw the perpendicular *AE*.

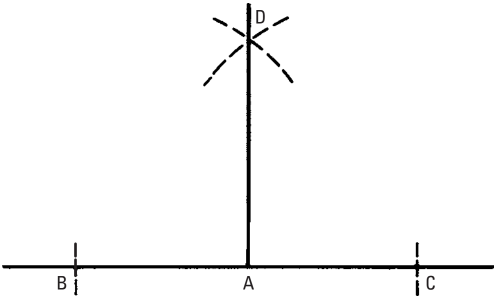

Figure 5-19 To erect a perpendicular to a straight line from a given point on that line.

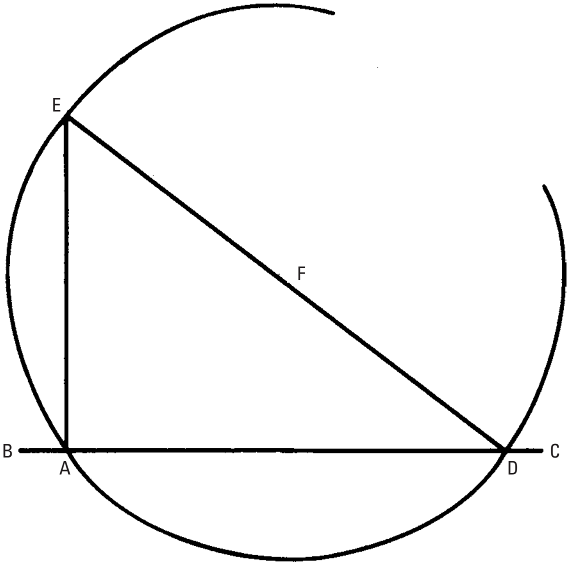

Figure 5-20 To erect a perpendicular to a straight line from a given point on that line, second method.

Third Method (Boat Builders' Layout Method)

In Figure 5-21, let *MS* be the given line and *A* be the given point. From *A*, measure off a distance *AB* (4 feet). With centers *A* and *B* and radii of 3 and 5 feet, respectively, describe arcs *L* and *F* intersecting at *C*. Draw a line through *A* and *C*, which will be the

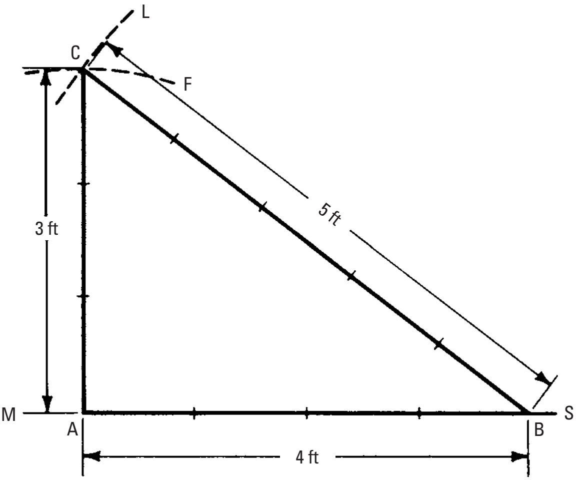

Figure 5-21 To erect a perpendicular to a straight line from a given point on that line, third method.

perpendicular required. This method is used extensively by carpenters when squaring the corners of buildings, but they ordinarily use multiples of 3, 4, and 5 (such as 6, 8, and 10, or 12, 16, and 20).

Fourth Method

In Figure 5-22, from *A*, describe an arc *EC*, and from *E* with the same radius describe the arc *AC*, cutting the other at *C*. Through *C*,

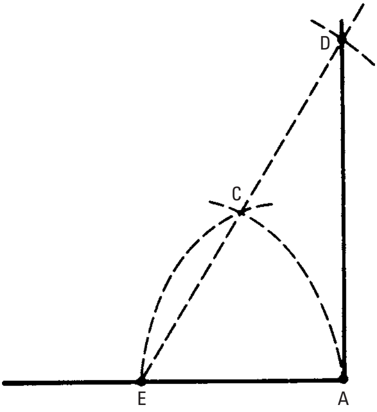

Figure 5-22 To erect a perpendicular to a straight line from a given point on that line, fourth method.

draw a line *ECD*. Lay off *CD* equal to *CE*, and through *D*, draw the perpendicular *AD*. The triangle produced is exactly 60 degrees at *E*, 30 degrees at *D*, and 90 degrees at *A*. The hypotenuse *ED* is exactly twice the length of the base *EA*.

Problem 4

To erect a perpendicular to a straight line from any point outside the line.

In Figure 5-23, from the point *A*, with a sufficient radius cut the given line at *F* and *G*, and from these points describe arcs cutting at *E*. Draw the perpendicular *AE*.

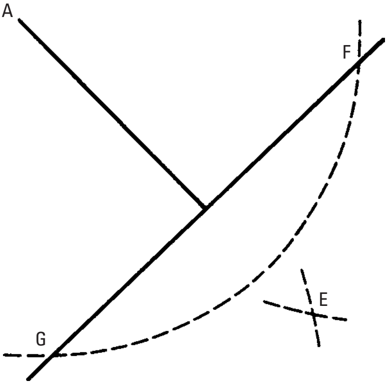

Figure 5-23 To erect a perpendicular to a straight line from any point outside the line.

Second Method

In Figure 5-24, from any two points *B* and *C* at some distance apart in the given line and with the radii *BA* and *CA*, respectively, describe arcs cutting at *A* and *D*. Draw the perpendicular *AD*.

Problem 5

To draw a line parallel to a given line through a given point.

In Figure 5-25, with *C* as the center, describe an arc tangent to the given line *AB*. The radius will then equal the distance from the given point to the given line. Take a point *B* on the given line remote from *C*, and describe an arc. Draw a line through *C*, tangent to this arc at *D*, and it will be parallel to the given line *AB*.

Second Method

In Figure 5-26, from *A*, the given point, describe the arc *FD*, cutting the given line at *F*; from *F*, with the same radius, describe the arc

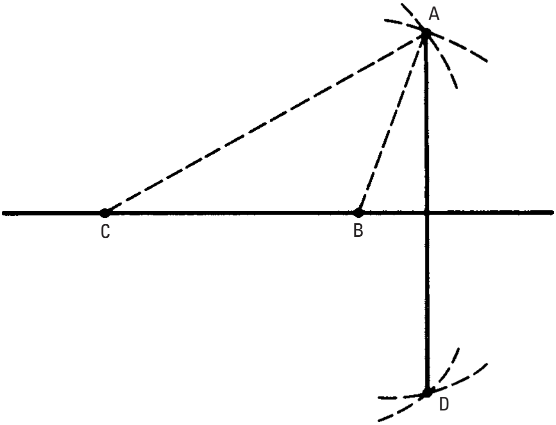

Figure 5-24 To erect a perpendicular to a straight line from any point outside the line, second method.

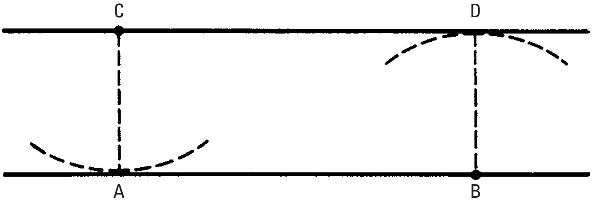

Figure 5-25 To draw a line parallel to a given line through a given point.

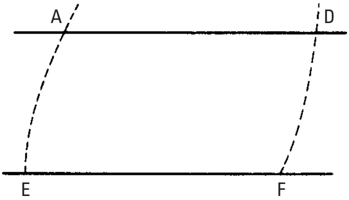

Figure 5-26 To draw a line parallel to a given line through a given point, second method.

EA, and lay off *FD* equal to *EA*. Draw the parallel line through the points *AD*.

Problem 6
To divide a line into a number of equal parts.

In Figure 5-27, assuming line *AB* is to be divided into five equal parts, draw a diagonal line *AC* of five units in length. Join *BC* at 5 and through the points 1, 2, 3, and 4. Draw lines 1*L*, 2*a*, and so forth, parallel to *BC*. *AC* will then be divided into five equal parts, *AL, La, ar, rf*, and *fB*.

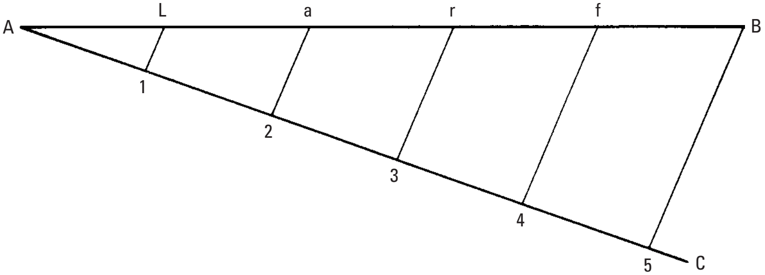

Figure 5-27 To divide a line into a number of equal parts.

Problem 7

To draw an angle equal to a given angle on a straight line.

In Figure 5-28, let *A* be the given angle, and *FG* the line. With any radius from the points *A* and *F*, describe arcs *DE* and *IH* cutting the sides of angle *A* and line *FG*. Lay off arc *IH* equal to arc *DE*, and draw line *FH*. Angle *F* is then equal to *A*, as required

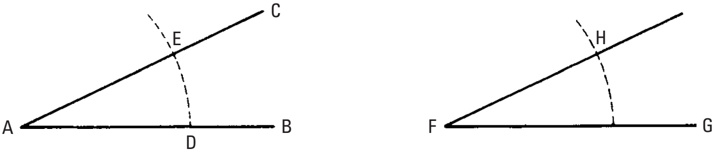

Figure 5-28 To draw an angle equal to a given angle on a straight line.

Problem 8

To bisect an angle

In Figure 5-29, let *ACB* be the angle. With the center of the angle at *C*, describe an arc cutting the sides at *A* and *B*. Using *A* and *B* as centers, describe arcs that intersect at *D*. A line through *C* and *D* will divide the angle into two equal parts.

Problem 9

To find the center of a circle.

In Figure 5-30, draw any chord *MS*. With *M* and *S* as centers, and with any radius, describe arcs *LF* and *L′F′*, and draw a line through their intersection, giving a diameter *AB*. Applying the same construction with centers *A* and *B*, describe arcs *ef* and *e′f′*. A line drawn through the intersections of these arcs will cut line *AB* at *O*, the center of the circle.

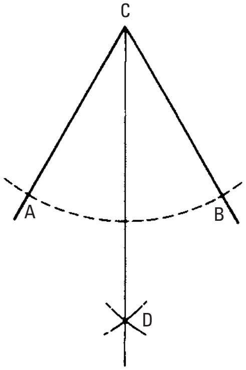

Figure 5-29 To bisect an angle.

Problem 10

To describe an arc of a circle with a given radius through two given points.

In Figure 5-31, take the given points *A* and *B* as centers, and, with the given radius, describe arcs that intersect at *C*. From *C*, with the same radius, describe an arc *AB*, as required.

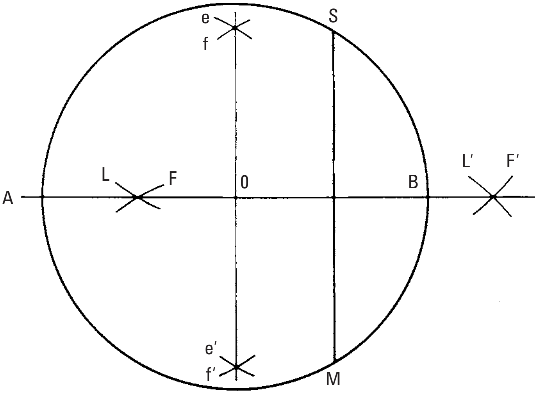

Figure 5-30 Find the center of a circle.

Second Method

In Figure 5-32, for a circle or an arc, select three points *ABC* in the circumference that are well apart. With the same given radius, describe arcs from these three points that intersect each other, and draw two lines, *DE* and *FG*, through their intersections. The point where these lines intersect is the center of the circle or arc.

Problem 11

To describe a circle passing through three given points.

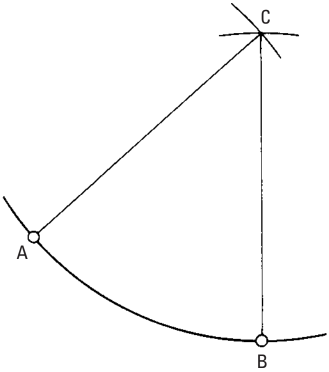

Figure 5-31 To describe an arc of a circle with a given radius through two given points.

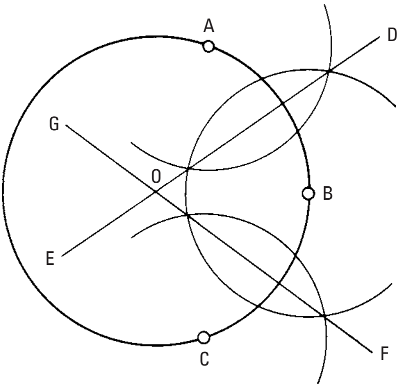

Figure 5-32 To describe an arc of a circle with a given radius through two given points, second method.

In Figure 5-32, let *A*, *B*, and *C* be the given points, and proceed as in Problem 10 to find the center *O* from which the circle may be described. This problem is useful in such work as laying out an object of large diameter (such as an arch) when the span and rise are given.

Problem 12

To draw a tangent to a circle from a given point in the circumference.

In Figure 5-33, from *A*, lay off equal segments *AB* and *AD*. Join line *BD*, and draw line *AE* parallel to *BD* for the tangent.

Problem 13

To draw tangents to a circle from points outside the circle.

In Figure 5-34, from *A*, and with the radius *AC*, describe an arc *BCD*. From *C*, with a radius equal to the diameter of the circle,

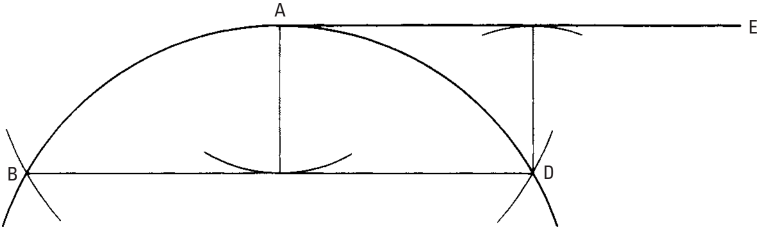

Figure 5-33 To draw a tangent to a circle from a given point in the circumference.

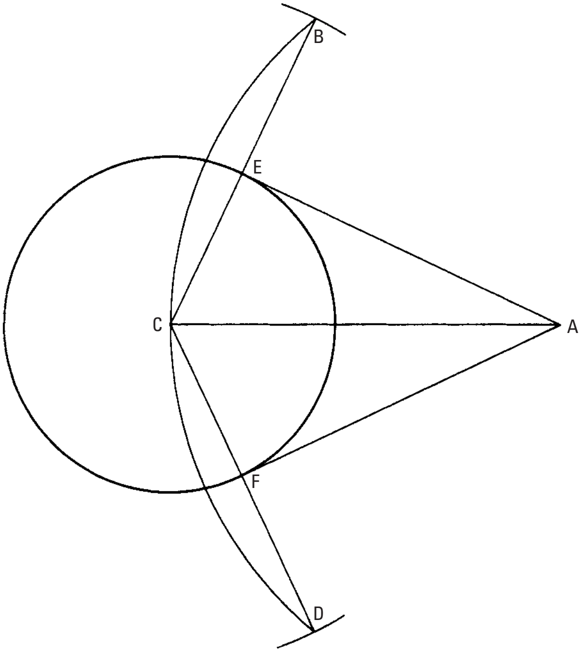

Figure 5-34 To draw tangents to a circle from points outside the circle.

intersect the arc at *BD*. Join *BC* and *CD*, which intersect the circle at *E* and *F*, and draw the tangents *AE* and *AF*.

Problem 14

To describe a series of circles tangent to two inclined lines and tangent to each other.

In Figure 5-35, bisect the inclination of the given lines *AB* and *CD* by the line *NO*. From a point *P* in this line, draw the perpendicular *PB* to the line *AB*, and on *P*, describe the circle *BD*, touching the lines and the centerline at *E*. From *E*, draw *EF* perpendicular to the center line intersecting *AB* at *F*, and from *F*, describe an arc *EG* intersecting *AB* at *G*. Draw *GH* parallel to *BP*, thus producing *H*, the center of the next circle, to be described with the radius *HE*, and so on for the next circle *IN*.

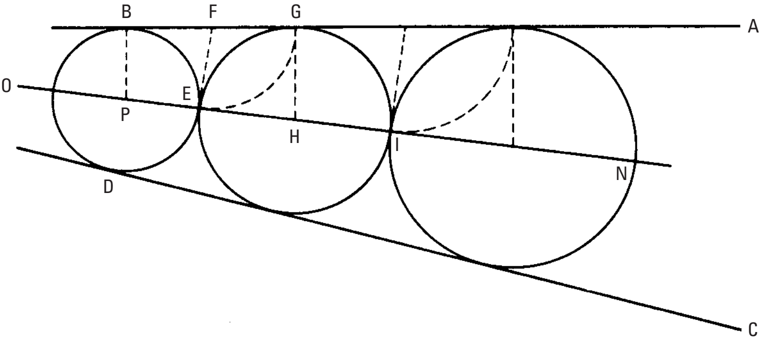

Figure 5-35 To describe a series of circles tangent to two inclined lines and tangent to each other.

Problem 15

To construct an equilateral triangle on a given base.

In Figure 5-36, with *A* and *B* as centers and a radius equal to *AB*, describe arcs *l* and *f*. At their intersection *C*, draw lines *CA* and *CB*, which are the sides of the required triangle. If the sides are to be unequal, the process is the same, taking as the radii the lengths of the two sides to be drawn.

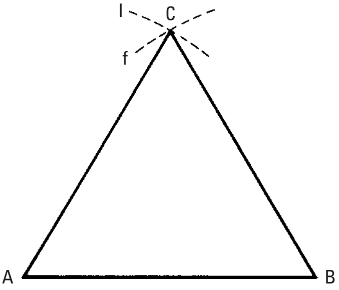

Figure 5-36 To construct an equilateral triangle on a given base.

Problem 16

To construct a square on a given base.

In Figure 5-37, with end points *A* and *B* of the base as centers and a radius equal to *AB*, describe arcs that intersect at *C*; on *C*, describe arcs that intersect the others at *D* and *E*, and on *D* and *E*, intersect these arcs *F* and *G*. Draw *AE* and *BG*, and join the intersections *HI* to form the square *AHIB*.

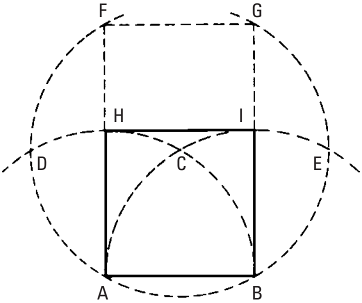

Figure 5-37 To construct a square or a rectangle on a given base.

Problem 17

To construct a rectangle on a given base.

In Figure 5-37, let *AB* be the given base. Erect a perpendicular at *A* and at *B* that is equal to the altitude of the rectangle, and join their ends *F* and *G* by line *FG*. *AFGB* is the rectangle required.

Problem 18

To construct a parallelogram given the sides and an angle.

In Figure 5-38, draw side *DE* equal to the given length *A*, and lay off the other side *DF*, equal to the other length *B*, thus forming the given angle *C*. From *E*, with *DF* as the radius, describe an arc, and from *F*, with the radius *DE*, intersect the arc at *G*. Draw *FG* and *EG*. The remaining sides may also be drawn as parallels to *DE* and *DF*.

Problem 19

To draw a circle around a triangle.

In Figure 5-39, bisect two sides *AB* and *AC* of the triangle at *E* and *F*, and from these points draw perpendiculars intersecting at *K*. From *K*, with radius *KA* or *KC*, describe the circle *ABC*.

Problem 20

To circumscribe and inscribe a circle about a square.

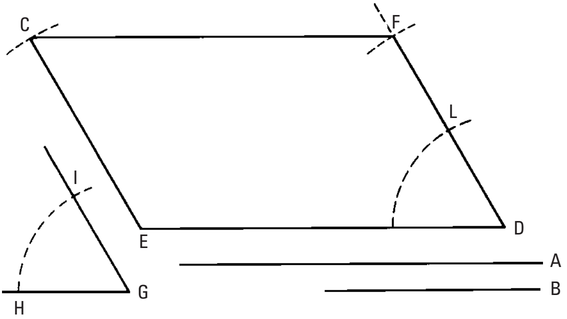
C
F
L
I
E
D
A
G
B
H

Figure 5-38 To construct a parallelogram given the sides and an angle.

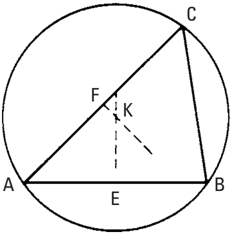

Figure 5-39 To draw a circle around a triangle.

In Figure 5-40, draw the diagonals *AB* and *CD* intersecting at *E*. With a radius *EA*, circumscribe the circle. To inscribe a circle, draw a perpendicular from the center (as just found) to one side of the square, as line *OM*. With radius *OM*, inscribe the circle.

Problem 21
To circumscribe a square around a circle.

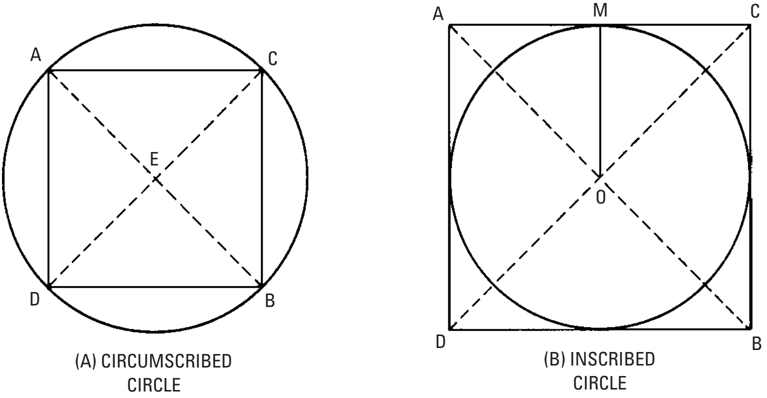
A
C
E
D
B
(A) CIRCUMSCRIBED CIRCLE
A
M
C
O
D
B
(B) INSCRIBED CIRCLE

Figure 5-40 (A) To circumscribe a circle around a square. (B) To inscribe a circle inside a square.

In Figure 5-41, draw diameters *MS* and *LF* at right angles to each other. At points *M*, *L*, *S*, and *F*, where these diameters intersect the circle, draw tangents (that is, lines perpendicular to the diameters), obtaining the sides of the circumscribed square *ABCD*.

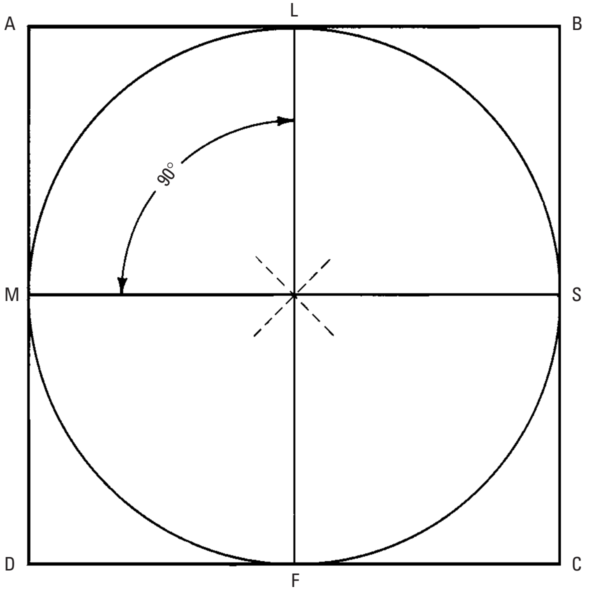

Figure 5-41 To circumscribe a square about a circle.

Problem 22

To inscribe a circle in a triangle.

In Figure 5-42, bisect two angles *A* and *C* of the triangle with lines that intersect at *D*. From *D*, draw a perpendicular *DE* to any side. With *DE* as the radius, describe a circle.

Problem 23

To inscribe a pentagon in a circle.

In Figure 5-43, draw two diameters *AC* and *BD* at right angles intersecting at *O*. Bisect *AO* at *E*, and from *E*, with radius *EB*, *AC* at *F*; from *B*, with radius *BF*. Intersect the circumference at *G* and *H*, and with the same radius, step round the circle to *I* and *K*; join the points thus found to form the pentagon *BGIKH*.

Problem 24

To inscribe a five-pointed star in a circle.

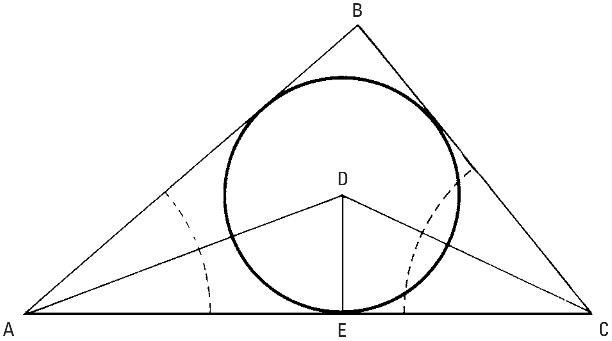

Figure 5-42 To inscribe a circle in a triangle.

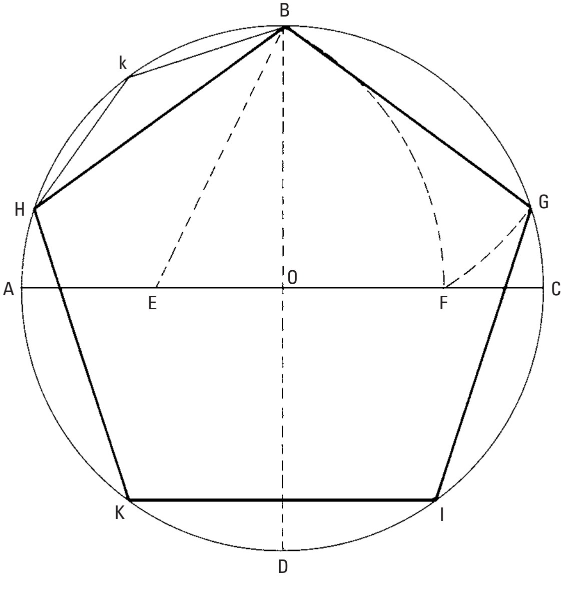

Figure 5-43 To inscribe a pentagon in a circle.

In Figure 5-44, proceed as explained for the inscribed pentagon in Problem 23. Then, connect point *B* with points *K* and *I*, point *H* with points *G* and *I*, and so forth. The star is mathematically correct.

Problem 25
To construct a hexagon from a given straight line

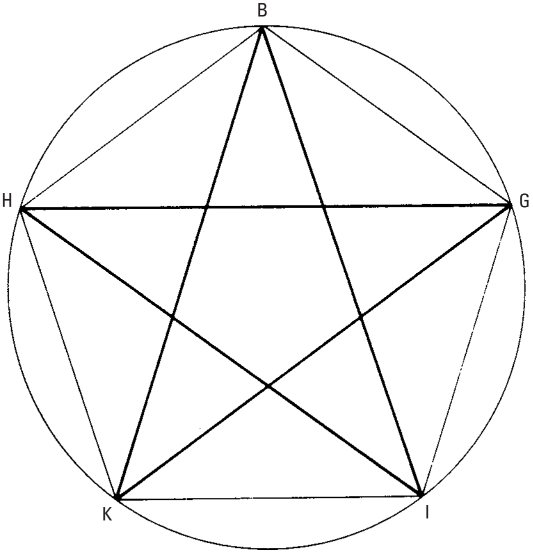

Figure 5-44 **To inscribe a five-pointed star in a circle.**

In Figure 5-45, from *A* and *B*, the ends of the given line, describe arcs intersecting at *g*. From *g*, with the radius *gA*, describe a circle. With the same radius, lay off arcs *AG*, *GF*, *BD*, and *DE*. Join the points thus found to form the hexagon.

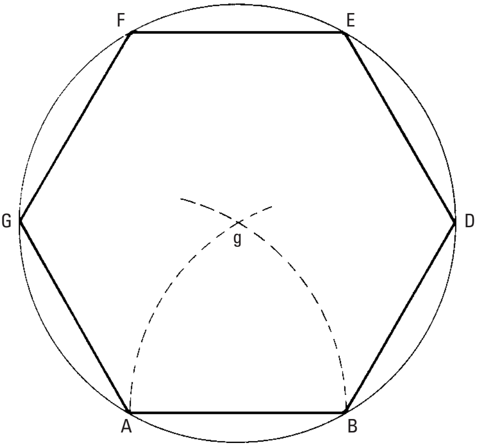

Figure 5-45 **To construct a hexagon from a given straight line.**

Problem 26

To inscribe a hexagon in a circle.

In Figure 5-46, draw a diameter *ACB*. From *A* and *B*, as centers with the radius of the circle *AC*, intersect the circumference at *D*, *E*, *F*, and *G*, and draw lines *AD*, *DE*, and so forth, to form the hexagon. The points *D*, *E*, and so forth, may also be found by stepping off the radius (with the dividers) six times around the circle.

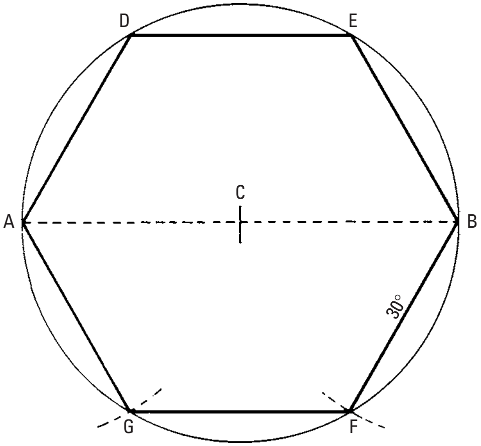

Figure 5-46 To inscribe a hexagon in a circle.

Problem 27

To describe an octagon on a given straight line.

In Figure 5-47, extend the given line *AB* both ways. Now, draw perpendiculars *AE* and *BF*. Bisect the external angles *A* and *B* by using lines *AH* and *BC*. These are made equal to line *AB*. Draw *CD* and *HG* parallel to *AE* and equal to line *AB*. Draw *CD* and *HG* parallel to *AE* and equal to line *AB*. With *G* and *D* as centers, and with the radius equal to *AB*, intersect the perpendiculars at *E* and *F*, and draw line *EF* to complete the hexagon.

Problem 28

To inscribe an octagon in a square.

In Figure 5-48, draw the diagonals of the square intersecting at *e*. From the corners *A*, *B*, *C*, and *D*, with *Ae* as the radius, describe arcs intersecting the sides of the square at *g*, *h*, and so forth, and join the points found to complete the octagon.

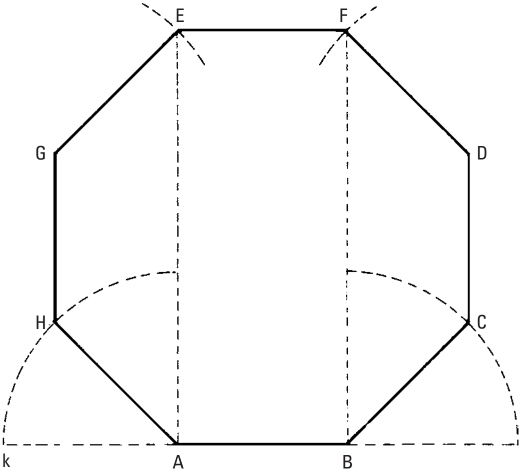

Figure 5-47 To describe an octagon on a given straight line.

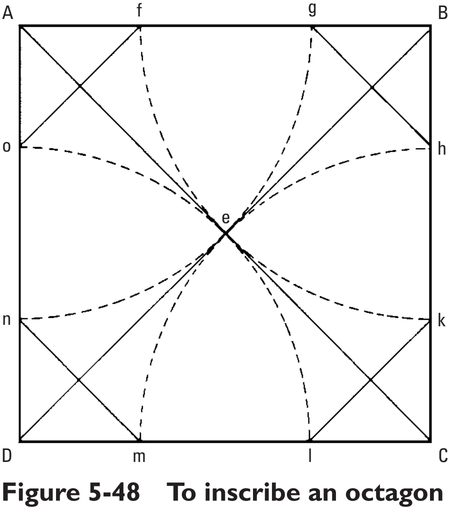

Figure 5-48 To inscribe an octagon in a square.

Problem 29

To inscribe an octagon in a circle.

In Figure 5-49, draw two diameters *AC* and *BD* at right angles. Bisect the arcs *AB*, *BC*, and so forth, at *e*, *f*, and so forth, to form the octagon.

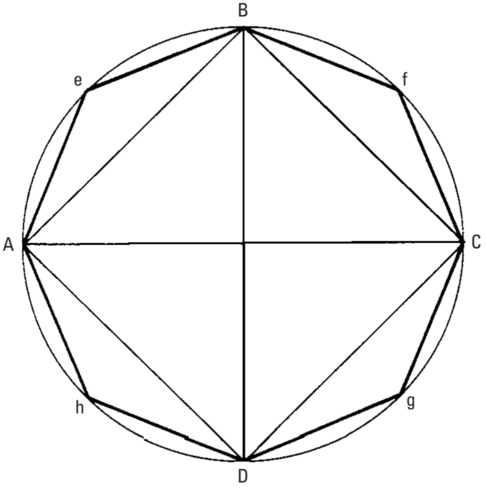

Figure 5-49 To inscribe an octagon in a circle.

Problem 30

To circumscribe an octagon about a circle.

In Figure 5-50, describe a square about the given circle *AB*. Draw perpendiculars *h*, *k*, and so forth, to the diagonals, touching the circle, to form the octagon. The points *h*, *k*, and so forth, may be found by cutting the sides from the corners.

Problem 31

To describe an ellipse when the two axes are given.

In Figure 5-51, draw the major and minor axes *AB* and *CD*, respectively, at right angles intersecting at *E*. On *C*, with *AE* as the radius, intersect the axis *AB* at *F* and *G*, the *foci*. Insert pins through the axis at *F* and *G*, and loop a thread or cord on them equal in length to the axis *AB*, so that when stretched, it reaches extremity *C* of the *conjugate axis*, as shown in dotted lines. Place a pencil inside the cord, as at *H*, and, by guiding the pencil in this manner, describe the ellipse.

Second Method

Along the edge of a piece of paper, mark off a distance *ac* equal to *AC*, one-half the major axis, and from the same point a distance *ab* equal to *CD*, one-half the minor axis, as shown in Figure 5-52. Place the paper to bring point *b* on the line *AB*, or major axis, and point *c* on the line *DE*, or minor axis. Lay off the position of point *a*.

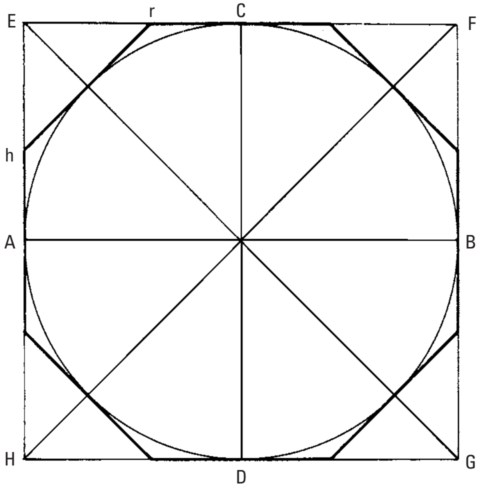

Figure 5-50 To circumscribe an octagon about a circle.

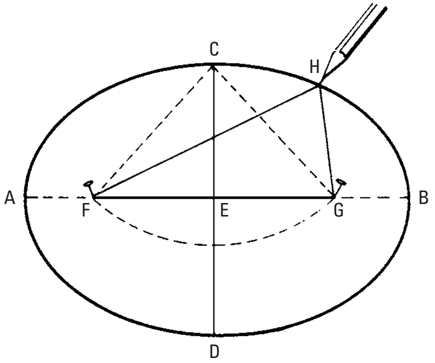

Figure 5-51 To describe an ellipse when the two axes are given.

By shifting the paper so that point *b* travels on the major axis and point *c* travels on the minor axis, any number of points in the curve may be found through which the curve may be traced.

Mensuration

As mentioned earlier, mensuration is the act, art, or process of measuring. It is the branch of mathematics that deals with finding the length of lines, the area of surfaces, and the volume of solids. Therefore, the problems that follow will be divided into three groups as follows:

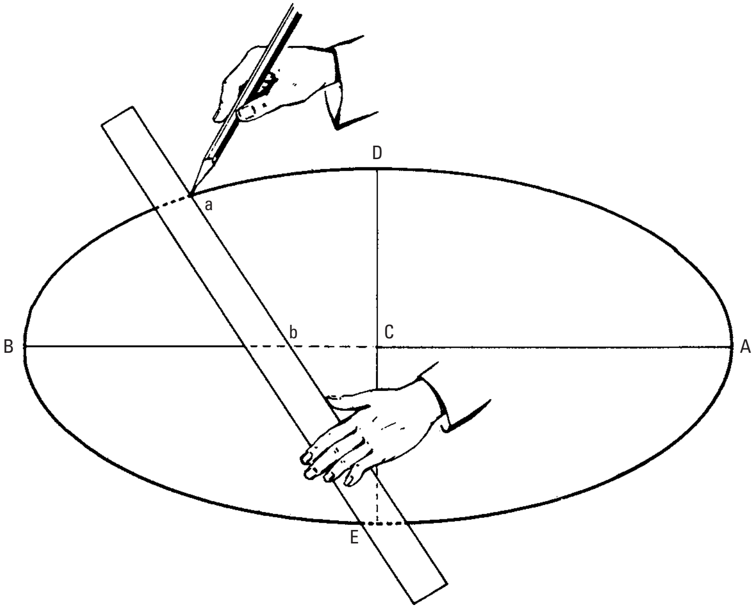

Figure 5-52 To describe an ellipse given the two axes, second method.

- Measurement of *lines*, one dimension (length)
- Measurement of *surfaces* (areas), two dimensions (length and width)
- Measurement of *solids* (volumes), three dimensions (length, width, and thickness)

Measurement of Lines—Length

Problem 1

To find the length of any side of a right triangle given the other two sides.

Rule: The length of the hypotenuse equals the square root of the sum of the squares of the two legs. The length of either leg equals the square root of the difference of the square of the hypotenuse and the square of the other leg.

Example The two legs of a right triangle measure 3 feet and 4 feet. Find the length of the hypotenuse. If the lengths of the hypotenuse and one leg are 5 feet and 4 feet, respectively, what is the length of the other leg?

In Figure 5-53A,

$$AB = \sqrt{3^2 + 4^2} = \sqrt{25} = 5 \text{ feet}$$

In Figure 5-53B,

$$BC = \sqrt{5^2 - 3^2} = \sqrt{25 - 9} = \sqrt{16} = 4 \text{ feet}$$

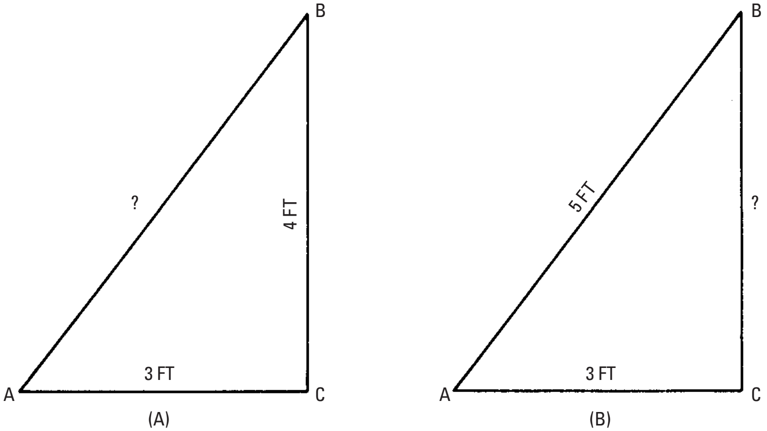

Figure 5-53 To find the length of any side of a right triangle given the other two sides.

Problem 2

To find the length of the circumference of a circle.

Rule: Multiply the diameter by 3.1416.

Example What length of molding strip is required for a circular window that is 5 feet in diameter?

$$5 \times 3.1416 = 15.7 \text{ feet}$$

Since the carpenter does not ordinarily measure feet in tenths, .7 should be reduced to inches. It corresponds to $8^1/_2$ inches from Table 5-23. That is, the length of molding required is 15 feet $8^1/_2$ inches.

Problem 3

To find the length of the arc of a circle.

Rule: Arc = .017453 × radius × central angle.

Example If the radius of a circle is 2 feet, what is the length of a 60° arc?

Table 5-23 Decimals of a Foot and Inches

Inch	*0 inch*	*1 inch*	*2 inches*	*3 inches*	*4 inches*	*5 inches*	*6 inches*	*7 inches*	*8 inches*	*9 inches*	*10 inches*	*11 inches*
0	0.0000	0.0833	0.1677	0.2500	0.3333	0.4167	0.5000	0.5833	0.6667	0.7500	0.8333	0.9167
1–16	0.0052	0.0885	0.1719	0.2552	0.3385	0.4219	0.5052	0.5885	0.6719	0.7552	0.8385	0.9219
1–8	0.0104	0.0937	0.1771	0.2604	0.3437	0.4271	0.5104	0.5937	0.6771	0.7604	0.8437	0.9271
3–16	0.0156	0.0990	0.1823	0.2656	0.3490	0.4323	0.5156	0.5990	0.6823	0.7656	0.8490	0.9323
1–4	0.0208	0.1042	0.1875	0.2708	0.3542	0.4375	0.5208	0.6042	0.6875	0.7708	0.8542	0.9375
5–16	0.0260	0.1094	0.1927	0.2760	0.3594	0.4427	0.5260	0.6094	0.6927	0.7760	0.8594	0.9427
3–8	0.0312	0.1146	0.1979	0.2812	0.3646	0.4479	0.5312	0.6146	0.6979	0.7812	0.8646	0.9479
7–16	0.0365	0.1198	0.2031	0.2865	0.3698	0.4531	0.5365	0.6198	0.7031	0.7865	0.8698	0.9531
1–2	0.0417	0.1250	0.2083	0.2917	0.3750	0.4583	0.5417	0.6250	0.7083	0.7917	0.8750	0.9583
9–16	0.0469	0.1302	0.2135	0.2969	0.3802	0.4635	0.5469	0.6302	0.7135	0.7969	0.8802	0.9635
5–8	0.0521	0.1354	0.2188	0.3021	0.3854	0.4688	0.5521	0.6354	0.7188	0.8021	0.8854	0.9688
11–16	0.0573	0.1406	0.2240	0.3073	0.3906	0.4740	0.5573	0.6406	0.7240	0.8073	0.8906	0.9740
3–4	0.0625	0.1458	0.2292	0.3125	0.3958	0.4792	0.5625	0.6458	0.7292	0.8125	0.8958	0.9792
13–16	0.0677	0.1510	0.2344	0.3177	0.4010	0.4844	0.5677	0.6510	0.7344	0.8177	0.9010	0.9844
7–8	0.0729	0.1562	0.2396	0.3229	0.4062	0.4896	0.5729	0.6562	0.7396	0.8229	0.9062	0.9896
15–16	0.0781	0.1615	0.2448	0.3281	0.4115	0.4948	0.5781	0.6615	0.7448	0.8281	0.9115	0.9948

Solution:

$$2 \times .017453 \times 60 = 2.094,$$

or approximately 2 feet $1^1/_8$ inches

Problem 4

To find the rise of an arc.

Rule: Rise of an arc =

$$\sqrt{(4 \times \text{radius}^2) - \text{length}}$$

Example If the radius of a circle is 2 feet, what is the rise at the center of a 2-foot chord?

Solution:

$$\begin{aligned} {}^1/_2\sqrt{(4 \times 2^2) - 2} &= {}^1/_2\sqrt{14} = 1.87 \text{ feet} \\ &= 1 \text{ feet } 10^1/_2 \text{ inches} \end{aligned}$$

Measurement of Surfaces—Area

Problem 5

To find the area of a square.

Rule: Multiply the base by the height.

Example What is the area of a square whose side is 5 feet (see Figure 5-54)?

$$5 \times 5 = 25 \text{ square feet}$$

Problem 6

To find the area of a rectangle.

Rule: Multiply the base by the height (that is, width by length).

Example What is the floor area of a porch 5 feet wide and 12 feet long (see Figure 5-55)?

$$5 \times 12 = 60 \text{ square feet}$$

Problem 7

To find the area of a parallelogram.

Rule: Multiply the base by the perpendicular height.

Example What is the area of a 5-foot × 12-foot parallelogram (see Figure 5-56)?

$$5 \times 12 = 60 \text{ square feet}$$

Problem 8

To find the area of a triangle (see Figure 5-57)

Rule: Multiply the base by one-half the altitude.

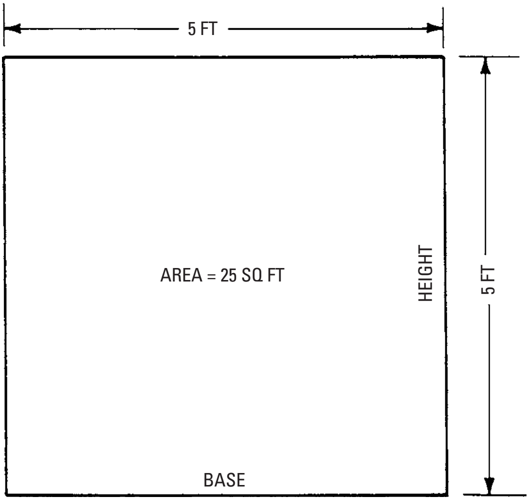
5 FT
AREA = 25 SQ FT
HEIGHT
5 FT
BASE

Figure 5-54 To find the area of a square.

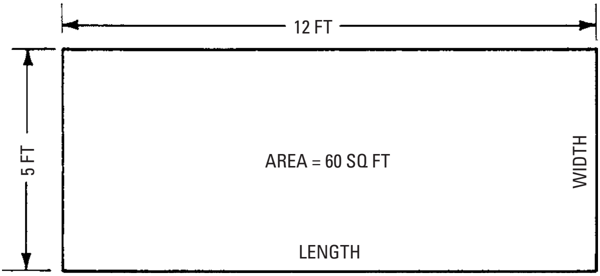
12 FT
5 FT
AREA = 60 SQ FT
WIDTH
LENGTH

Figure 5-55 To find the area of a rectangle.

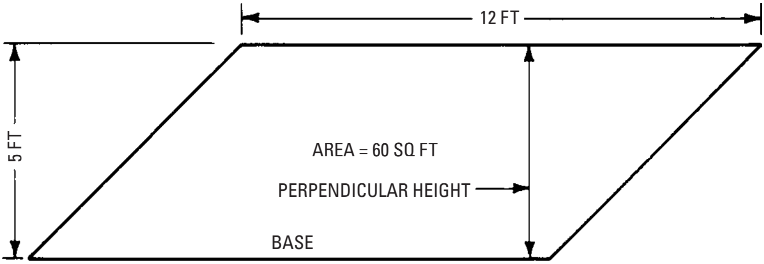
12 FT
5 FT
AREA = 60 SQ FT
PERPENDICULAR HEIGHT
BASE

Figure 5-56 To find the area of a parallelogram.

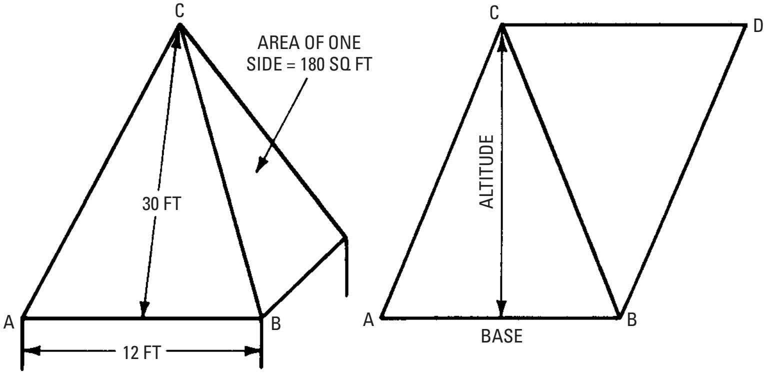

(A) (B)

Figure 5-57 To find the area of a triangle (equal to ½ area of parallelogram ABDC).

Example How many square feet of sheathing are required to cover a church steeple having four triangular sides?

Problem 9

To find the area of a trapezoid.

Rule: Multiply one-half the sum of the two parallel sides by the perpendicular distance between them.

Example What is the area of the trapezoid shown in Figure 5-58?

LA and FR are the parallel sides, and MS is the perpendicular distance between them. Therefore,

$$\text{area} = {}^1\!/_2\,(LA + FR) \times MS$$

$$\text{area} = {}^1\!/_2\,(8 + 12) \times 6 = 60 \text{ square feet}$$

Problem 10

To find the area of a trapezium.

Rule: Draw a diagonal, dividing the figure into triangles. Measure the diagonal and the altitudes, and find the area of the triangles. The sum of these areas is then the area of the trapezium.

Example What is the area of the trapezium shown in Figure 5-59? (Draw diagonal LR and altitudes AM and FS.)

$$\text{area of triangle } ALR = {}^1\!/_2\,(12 \times 9) = 54 \text{ square feet}$$

$$\text{area of triangle } LRF = {}^1\!/_2\,(12 \times 6) = 36 \text{ square feet}$$

$$\begin{aligned}\text{area of trapezium } LARF &= ALR + LRF = 36 + 54\\ &= 90 \text{ square feet}\end{aligned}$$

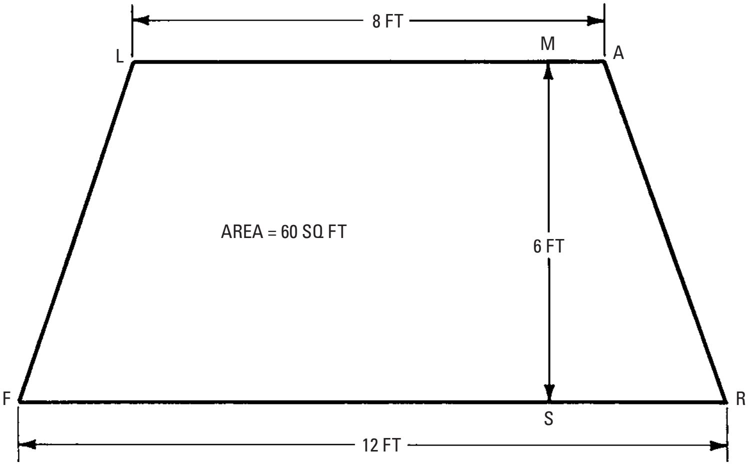
8 FT
L
M
A
AREA = 60 SQ FT
6 FT
F
S
R
12 FT

Figure 5-58 To find the area of a trapezoid.

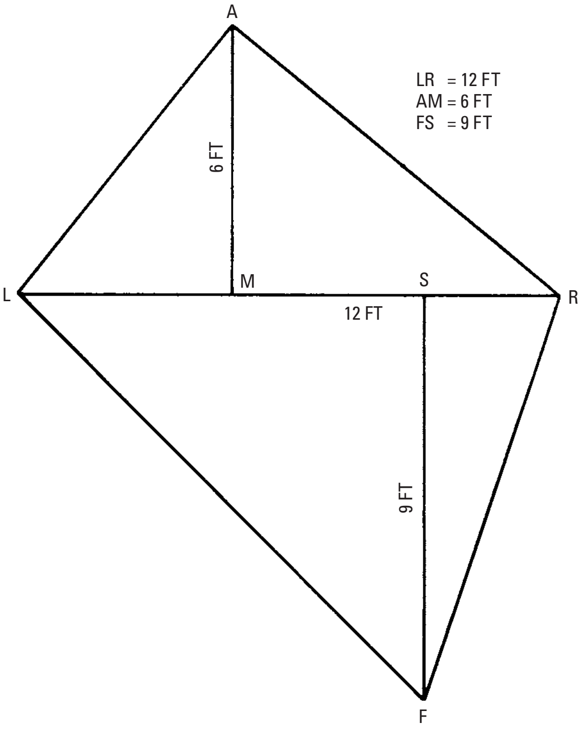
A
LR = 12 FT
AM = 6 FT
FS = 9 FT
6 FT
L
M
S
R
12 FT
9 FT
F

Figure 5-59 To find the area of a trapezium.

Problem 11

To find the area of any irregular polygon.

Rule: Draw diagonals, dividing the figure into triangles, and find the sum of the areas of these triangles.

Problem 12

To find the area of any regular polygon, such as shown in Figure 5-60, when the length of only one side is given

Rule: Multiply the square of the sides by the figure for "area when side = 1" opposite the particular polygon in Table 5-24.

Example What is the area of an octagon (8-sided polygon) whose sides are 4 feet in length?

In Table 5-24 under 8 find 4.828. Multiply this by the square of one side.

$$4.828 \times 4^2 = 77.25 \text{ square feet}$$

Problem 13

To find the area of a circle (see Figure 5-61).

Rule: Multiply the square of the diameter by 0.7854.

Example How many square feet of floor surface are there in a 10-foot circular floor?

$$10^2 \times 0.7854 = 78.54 \text{ square feet}$$

Problem 14

To find the area of a sector of a circle.

Rule: Multiply the arc of the sector by one-half the radius.

Example How much tin is required to cover a 60° section of a 10-foot circular deck?

$$\text{length of } 60^\circ \text{ arc} = {}^{60}/_{30} \text{ of } 3.1416 \times 10 = 5.24 \text{ feet}$$

$$\text{tin required for } 60^\circ \text{ sector} = 5.24 \times {}^{1}/_{2} \times 5$$
$$= 13.1 \text{ square feet}$$

Problem 15

To find the area of a segment of a circle.

Rule: Find the area of the sector that has the same arc, and also find the area of the triangle formed by the radii and chord. Take the sum of these areas if the segment is greater than 180°. Take the difference if the segment is less than 180°.

Problem 16

To find the area of a ring.

Rule: Take the difference between the areas of the two circles.

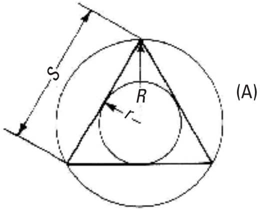

Steel square cut to miter sides at 30°

$R = 2r$, or $.577\ S$

$$\text{AREA} = .433\ S^2 = 1.299\ R^2 = 5.196\ r^2 = \frac{(R+r)}{2}\ S$$

(A) Equilateral triangle (3 sides).

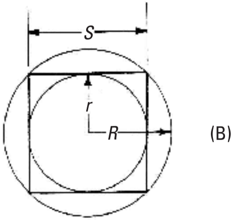

Steel square cut to miter sides at 45°

$R = .707\ S$, or $1.414\ r$

$$\text{AREA} = S^2 = 2R^2 = 4r^2$$

(B) Square (4 sides).

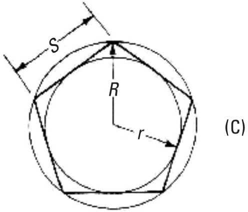

Steel square cut to miter sides at 36°

$R = .851\ S$, or $1.236\ r$

$$\text{AREA} = 1.72\ S^2 = 2.378\ R^2 = 3.633\ r^2$$

(C) Pentagon (5 sides).

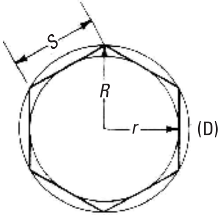

Steel square cut to miter sides at 30°

$R = S$, or $1.155\ r$

$$\text{AREA} = 2.598\ S^2 = 2.598\ R^2 = 3.464\ r^2$$

(D) Hexagon (6 sides).

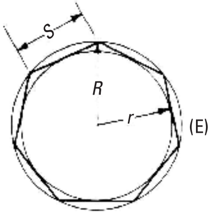

Steel square cut to miter corners at 25° 43'

$R = 1.152\ S$, or $1.11\ r$

$$\text{AREA} = 3.634\ S^2 = 2.736\ R^2 = 3.371\ r^2$$

(E) Heptagon (7 sides).

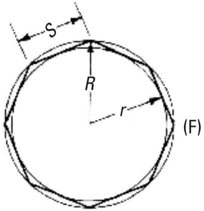

Steel square cut to miter corners at 22° 30'

$R = 1.307\ S$, or $1.082\ r$

$$\text{AREA} = 4.828\ S^2 = 2.828\ R^2 = 3.314\ r^2$$

(F) Octagon (8 sides).

Figure 5-60 Regular polygons.

Table 5-24 Regular Polygons

Number of sides	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
Area when side = 1	0.433	1.0	1.721	2.598	3.634	4.828	6.181	7.694	9.366	11.196

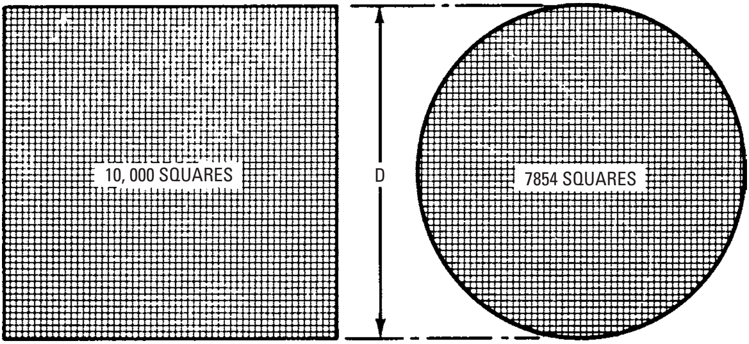

Figure 5-61 The decimal 0.7854 is used to find the area of a circle. If a square is divided into 10,000 equal parts (small squares), then a circle with a diameter D equal to one side of the large square will contain 7854 small squares. Therefore, if the area of the large square is 1 square inch, then the area of the circle will be $^{7854}/_{10,000}$, or 0.7854 square inch.

Problem 17

To find the area of an ellipse.

Rule: Multiply the product of the two diameters by 0.7854.

Example What is the area of an ellipse whose two diameters are 10 inches and 6 inches?

$$10 \times 6 \times 0.7854 = 47.12 \text{ square inches}$$

Problem 18

To find the circular area of a cylinder.

Rule: Multiple 3.1416 by the diameter and by the height.

Example How many square feet of lumber are required for the sides of a cylindrical tank (see Figure 5-62) that is 8 feet in diameter and 12 feet high? How many 4-inch × 12-feet pieces will be required?

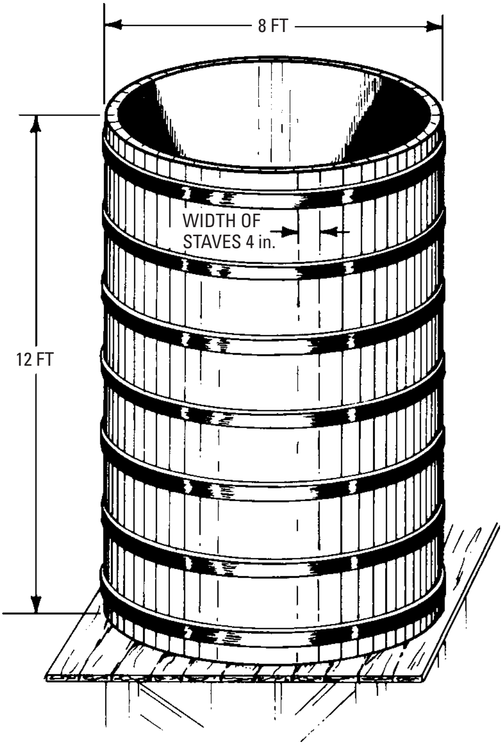

Figure 5-62 To find the area of a cylinder.

$$\text{cylindrical surface} = 3.1416 \times 8 \times 12 = 302 \text{ square feet}$$

$$\text{circumference of tank} = 3.1416 \times 8 = 25.1 \text{ feet}$$

$$\text{number of } 4'' \times 12' \text{ pieces} = \frac{25.1 \times 12}{4} = 25.1 \times 3 = 75.3$$

Problem 19
To find the area of a cone (see Figure 5-63).

Rule: Multiply 3.1416 by the diameter of the base and by one-half the slant height.

Example A conical spire with a base 10 feet in diameter and an altitude of 20 feet is to be covered. Find the area of the surface to be covered.

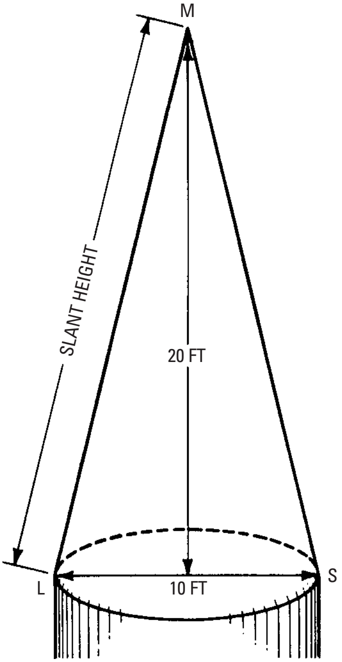

Figure 5-63 To find the surface area of a cone.

$$\text{slant height} = \sqrt{5^2 + 20^2} = \sqrt{425} = 20.62 \text{ feet}$$

$$\text{circumference of base} = 3.1416 \times 10 = 31.416 \text{ feet}$$

$$\begin{aligned}\text{area of conical surface} &= 31.416 \times {}^1/_2 \times 20.62\\ &= 324 \text{ square feet}\end{aligned}$$

Problem 20

To find the area of the frustum of a cone (see Figure 5-64)

Rule: Multiply one-half the slant height by the sum of the circumference.

Example A tank is 12 feet in diameter at the base, 10 feet at the top, and 8 feet high. What is the area of the slant surface?

$$\begin{aligned}\text{circumference of 10-foot diameter} &= 3.1416 \times 10\\ &= 31.416 \text{ feet}\end{aligned}$$

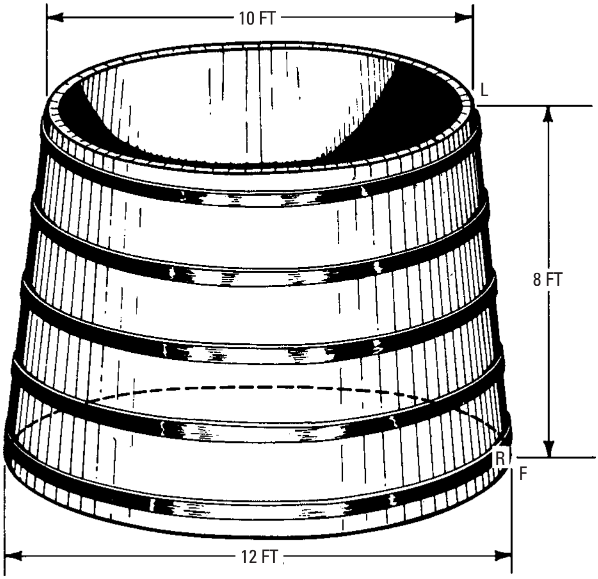

Figure 5-64 To find the area of the frustum of a cone.

$$\text{circumference of 12-foot diameter} = 3.1416 \times 12$$
$$= 37.7 \text{ feet}$$

$$\text{sum of circumferences} = 69.1 \text{ feet}$$

$$\text{slant height} = \sqrt{1^2 + 8^2} = \sqrt{65} = 8.12$$

$$\text{slant surface} = \text{sum of circumferences} \times {}^1/_2 \text{ slant height}$$

$$\text{slant surface} = 69.1 \times {}^1/_2 \times 8.12 = 280 \text{ square feet}$$

Measurement of Solids—Volume

Problem 21

To find the volume of a rectangular solid.

Rule: Multiply the length, width, and thickness together.

Example What is the volume of a 4-inch × 8-inch × 12-foot timber? (Before applying the rule, reduce all dimensions to feet.)

$$4 \text{ inches} = {}^1/_3 \text{ foot}$$

$$8 \text{ inches} = {}^2/_3 \text{ foot}$$

$$\text{volume of timber} = {}^1/_3 \times {}^2/_3 \times 12 = 2.67 \text{ cubic feet}$$

If the timber were a piece of oak weighing 48 pounds per cubic foot, the total weight would be calculated as follows:

$$48 \times 2.67 = 128 \text{ pounds}$$

Problem 22

To find the volume of a rectangular wedge.

Rule: Find the area of one of the triangular ends, and multiply the area by the distance between the ends.

Example An attic has the shape of a rectangular wedge. What volume storage capacity would there be for the proportions shown in Figure 5-65? In the illustration, the boundary of the attic is *LARFMS.*

$$\text{Area of triangular end } MLA = 20 \times {}^{10}/_{2} = 100 \text{ square feet}$$

$$\text{Volume of attic} = 100 \times 40 = 4000 \text{ cubic feet.}$$

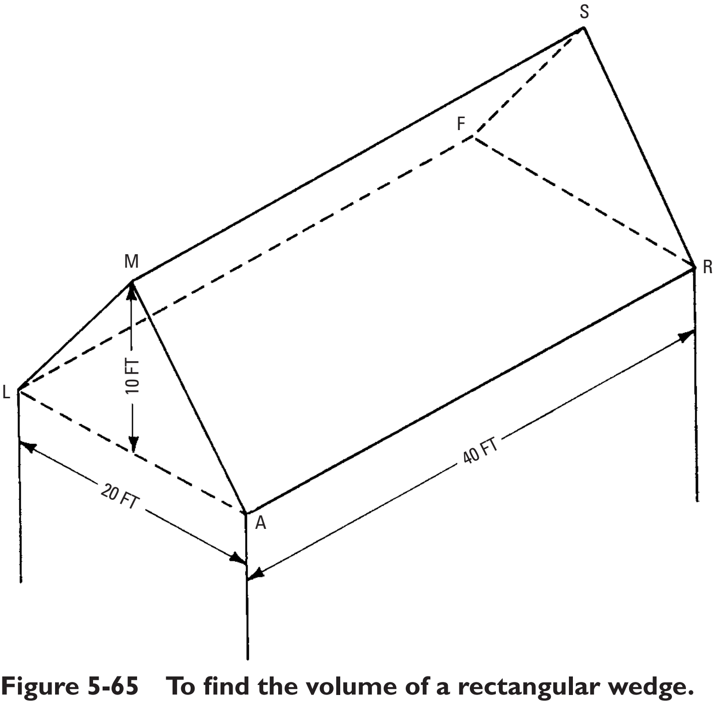

Figure 5-65 To find the volume of a rectangular wedge.

Trigonometry

Trigonometry is that branch of mathematics that deals with the relations that exist between the sides and angles of triangles, and more especially with those methods of calculating the required parts of triangles from given parts. The only branch of trigonometry useful to the carpenter and builder is *plane trigonometry*, where the lines in the triangles are straight and where they all lie in the same plane.

The six elements (or parts) in every triangle are three sides and three angles. The sum of the three angles, no matter what the lengths of the sides, will always be equal to 180 degrees.

When any three of the six parts are given, provided one or more of them are sides, the other three are calculable. The angles are measured in circular measure: in degrees (°), minutes (′), and seconds (″). The term *degree* has no numerical value. In trigonometry it simply means $^1/_{360}$ of a circle, nothing more.

To the student of trigonometry, any two radii that divide a circle into anything more than 0° or less than 360° form an angle.

The first 90° division is called the *first quadrant.* Angles in this quadrant are the *acute angles* (see Figure 5-66A) mentioned earlier in this chapter. Angles from 90° to 180° are in the *second quadrant.* These are the *obtuse angles* (see Figure 5-66B) mentioned. Angles from 180° to 270° lie in the *third quadrant*, and angles from 270° to 360° lie in the *fourth quadrant.* Figure 5-67 shows these four quadrants. Only angles in the first and second quadrants, from 0° to 180°, will be discussed in this section. Note that a straight line may be considered as an angle of 180°. Trigonometry is actually based on geometry, but it makes use of many algebraic operations that can be used by carpenters and builders.

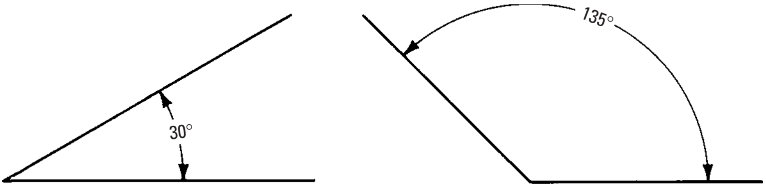

(A) An angle in the first quadrant is an acute angle.

(B) An angle in the second quadrant is an obtuse angle.

Figure 5-66 Acute and obtuse angles.

Trigonometric Functions

In mathematics, a *function* means a quantity that necessarily changes because of a change in another number with which it is connected

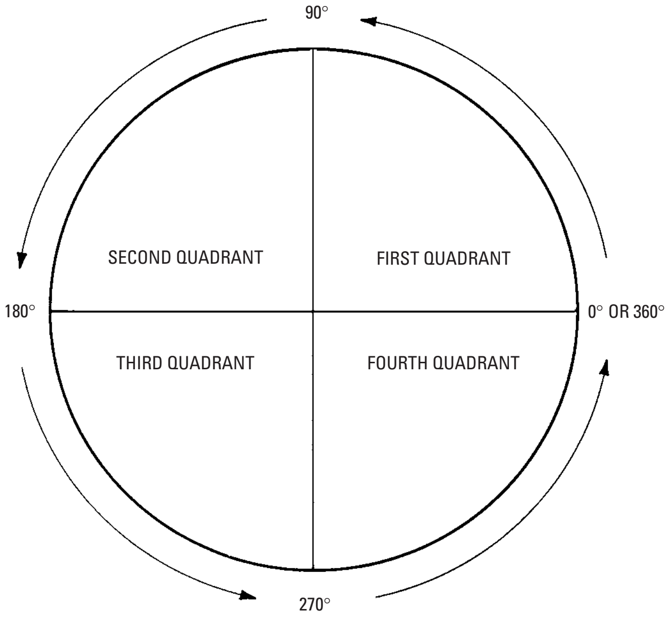

Figure 5-67 The four quadrants of a circle.

in some way. In trigonometry, it is probably less confusing to call the trigonometric functions simply *ratios*, which they truly are.

Refer to Figure 5-68 for an explanation of the following. There are six trigonometric ratios commonly used:

$$\text{Sine of angle } A = \frac{\text{opposite side}}{\text{hypotenuse}} \text{ or } \frac{BC}{AB}$$

$$\text{Cosine of angle } A = \frac{\text{adjacent side}}{\text{hypotenuse}} \text{ or } \frac{AC}{AB}$$

$$\text{Tangent of angle } A = \frac{\text{opposite side}}{\text{adjacent side}} \text{ or } \frac{BC}{AC}$$

$$\text{Cotangent of angle } A = \frac{\text{adjacent side}}{\text{opposite side}} \text{ or } \frac{AC}{BC}$$

$$\text{Secant of angle } A = \frac{\text{hypotenuse}}{\text{adjacent side}} \text{ or } \frac{AB}{AC}$$

$$\text{Cosecant of angle } A = \frac{\text{hypotenuse}}{\text{opposite side}} \text{ or } \frac{AB}{BC}$$

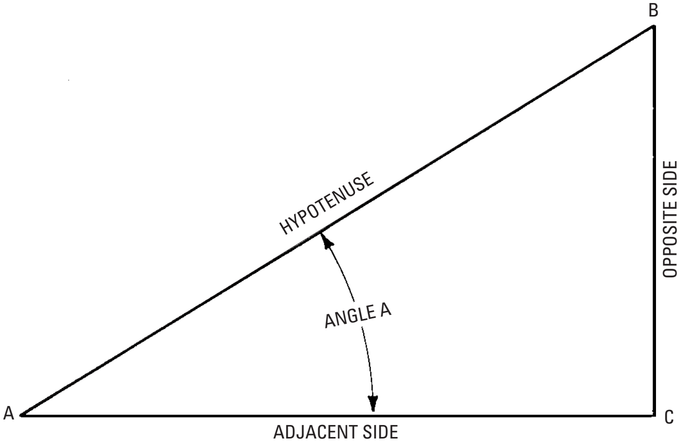

Figure 5-68 A right triangle illustrates the application of trigonometric ratios that are commonly used.

Note that these last three functions are only *reciprocals* of the sine, cosine, and tangent, respectively, or

$$\text{cosecant} = \frac{1}{\text{sine}}$$

$$\text{secant} = \frac{1}{\text{cosine}}$$

$$\text{cotangent} = \frac{1}{\text{tangent}}$$

If a proposition calls for multiplication by the sine of an angle, the same result will be obtained by dividing by the cosecant. It is convenient to do this in many calculations.

It is impossible in a discussion of this type to give a comprehensive table of the trigonometric ratios, although an adequate (but limited) number of trigonometric functions is presented in Table 5-25. Those who would like to follow up the information given here are advised to obtain a book of five- or six-place tables.

As an example of how trigonometric ratios are used to solve one of the carpenter's most common problems (determining the length of rafters given the rise and run), refer to Figure 5-69. The slope of the roof, in degrees, may be determined by dividing the opposite side, 12 feet, by the adjacent side, 18 feet. This is the tangent of the

Table 5-25 Natural Trigonometric Functions

Degree	*Sine*	*Cosine*	*Tangent*	*Secant*
0	0.00000	1.0000	0.00000	1.0999
1	0.01745	0.9998	0.01745	1.0001
2	0.03490	0.9994	0.03492	1.0006
3	0.05234	0.9986	0.05241	1.0014
4	0.06976	0.9976	0.06993	1.0024
5	0.08716	0.9962	0.08749	1.0038
6	0.10453	0.9945	0.10510	1.0055
7	0.12187	0.9925	0.12278	1.0075
8	0.1392	0.9903	0.1405	1.0098
9	0.1564	0.9877	0.1584	1.0125
10	0.1736	0.9848	0.1763	1.0154
11	0.1908	0.9816	0.1944	1.0187
12	0.2079	0.9781	0.2126	1.0223
13	0.2250	0.9744	0.2309	1.0263
14	0.2419	0.9703	0.2493	1.0306
15	0.2588	0.9659	0.2679	1.0353
16	0.2756	0.9613	0.2867	1.0403
17	0.2924	0.9563	0.3057	1.0457
18	0.3090	0.9511	0.3249	1.0515
19	0.3256	0.9455	0.3443	1.0576
20	0.3420	0.9397	0.3640	1.0642
21	0.3584	0.9336	0.3839	1.0711
22	0.3746	0.9272	0.4040	1.0785
23	0.3907	0.9205	0.4245	1.0864
24	0.4067	0.9135	0.4452	1.0946
25	0.4226	0.9063	0.4663	1.1034
26	0.4384	0.8988	0.4877	1.1126
27	0.4540	0.8910	0.5095	1.1223
28	0.4695	0.8829	0.5317	1.1326
29	0.4848	0.8746	0.5543	1.1433
30	0.5000	0.8660	0.5774	1.1547
31	0.5150	0.8572	0.6009	1.1663
32	0.5299	0.8480	0.6249	1.1792
33	0.5446	0.8387	0.6494	1.1924
34	0.5592	0.8290	0.6745	1.2062
35	0.5736	0.8192	0.7002	1.2208
36	0.5878	0.8090	0.7265	1.2361

(continued)

Table 5-25 (*continued*)

Degree	***Sine***	***Cosine***	***Tangent***	***Secant***
37	0.6018	0.7986	0.7536	1.2521
38	0.6157	0.7880	0.7813	1.2690
39	0.6293	0.7771	0.8098	1.2867
40	0.6428	0.7660	0.8391	1.3054
41	0.6561	0.7547	0.8693	1.3250
42	0.6691	0.7431	0.9004	1.3456
43	0.6820	0.7314	0.9325	1.3673
44	0.6947	0.7193	0.9657	1.3902
45	0.7071	0.7071	1.0000	1.4142
46	0.7193	0.6947	1.0355	1.4395
47	0.7314	0.6820	1.0724	1.4663
48	0.7431	0.6691	1.1106	1.4945
49	0.7547	0.6561	1.1504	1.5242
50	0.7660	0.6428	1.1918	1.5557
51	0.7771	0.6293	1.2349	1.5890
52	0.7880	0.6157	1.2799	1.6243
53	0.7986	0.6018	1.3270	1.6616
54	0.8090	0.5878	1.3764	1.7013
55	0.8192	0.5736	1.4281	1.7434
56	0.8290	0.5592	1.4826	1.7883
57	0.8387	0.5446	1.5399	1.8361
58	0.8480	0.5299	1.6003	1.8871
59	0.8572	0.5150	1.6643	1.9416
60	0.8660	0.5000	1.7321	2.0000
61	0.8746	0.4848	1.8040	2.0627
62	0.8829	0.4695	1.8807	2.1300
63	0.8910	0.4540	1.9626	2.2027
64	0.8988	0.4384	2.0503	2.2812
65	0.9063	0.4226	2.1445	2.3662
66	0.9135	0.4067	2.2460	2.4586
67	0.9205	0.3907	2.3559	2.5598
68	0.9272	0.3746	2.4751	2.6695
69	0.9336	0.3584	2.6051	2.7904
70	0.9397	0.3420	2.7475	2.9238
71	0.9455	0.3256	2.9042	3.0715
72	0.9511	0.3090	3.0777	3.2361

(continued)

Table 5-25 (*continued*)

Degree	***Sine***	***Cosine***	***Tangent***	***Secant***
73	0.9563	0.2924	3.2709	3.4203
74	0.9613	0.2756	3.4874	3.6279
75	0.9659	0.2588	3.7321	3.8637
76	0.9703	0.2419	4.0108	4.1336
77	0.9744	0.2250	4.3315	4.4454
78	0.9781	0.2079	4.7046	4.8097
79	0.9816	0.1908	5.1446	5.2408
80	0.9848	0.1736	5.6713	5.7588
81	0.9877	0.1564	6.3138	6.3924
82	0.9903	0.1392	7.1154	7.1853
83	0.9925	0.12187	8.1443	8.2055
84	0.9945	0.10453	9.5144	9.5668
85	0.9962	0.08716	11.4301	11.474
86	0.9976	0.06976	14.3007	14.335
87	0.9986	0.05234	19.0811	19.107
88	0.9994	0.03490	28.6363	28.654
89	0.9998	0.01745	57.2900	57.299
90	1.0000	Inf.	Inf.	Inf.
	–	–	–	–

angle *A* and is equal to $^{12}/_{18}$ or .6667. From Table 5-25, angle *A* is determined to be 33° 42′. The length of the rafter may be determined by the ratio:

$$\text{secant} = \frac{\text{hypotenuse}}{\text{adjacent side}}$$

or the hypotenuse (the length of the rafter) is equal to the following:

$$\text{Secant } 33°42' \times \text{adjacent side}$$

The secant of 33°42′ is equal to 1.2020. Therefore, the calculation for the length of the rafter is the following:

$$1.2020 \times 18 = 21.64 \text{ feet} = 21 \text{ feet } 7^{3}/_{16} \text{ inches}$$

Since the opposite side is known to be 12 feet, the calculation could just as easily be made by using the cosecant function.

Table 5-26 shows the slopes (in degrees) for all regular roof pitches. These pitches range from 12 × 1 to 12 × 12, and the three

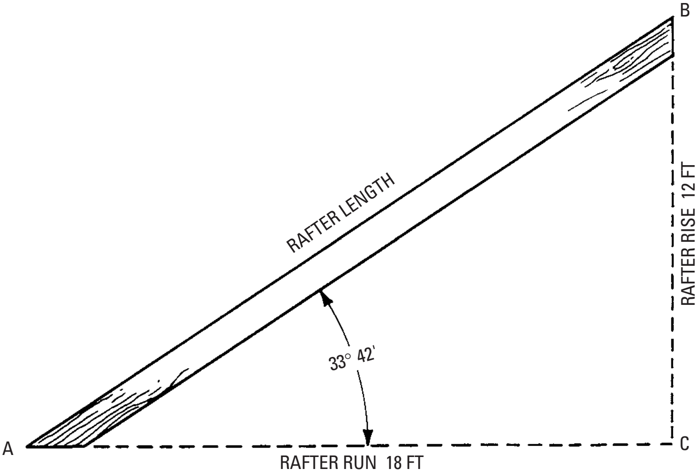

Figure 5-69 Trigonometric ratios may be used to determine the lengths of rafters for a roof.

Table 5-26 Roof Pitches in Degrees and Minutes (Measured from the Horizontal)

Pitch	*Sine*	*Cosine*	*Tangent*
12 × 1 = 4° 46′	0.083098	0.996541	0.083386
12 × 2 = 9° 28′	0.164474	0.986381	0.166745
12 × 3 = 14° 2′	0.242486	0.970155	0.249946
12 × 4 = 18° 26′	0.316201	0.948692	0.333302
12 × 5 = 22° 37′	0.384564	0.923098	0.416601
12 × 6 = 26° 34′	0.444635	0.895712	0.496404
12 × 7 = 30° 15′	0.503774	0.863836	0.583183
12 × 8 = 33° 41′	0.554602	0.832115	0.666497
12 × 9 = 36° 53′	0.600188	0.799859	0.750366
12 × 10 = 39° 46′	0.639663	0.768656	0.832183
12 × 11 = 42° 31′	0.675805	0.737081	0.916866
12 × 12 = 45° 00′	0.707107	0.707107	1.000000

main trigonometric ratios (sine, cosine, and tangent) are provided for each pitch.

Other typical examples of how trigonometric ratios can aid carpenters are shown in the following problems.

Problem 1

A grillwork consisting of radial and vertical members is to be built in a semicircular opening with a radius of 6 feet, as shown in Figure 5-70. Find the lengths of the vertical pieces MS and LF.

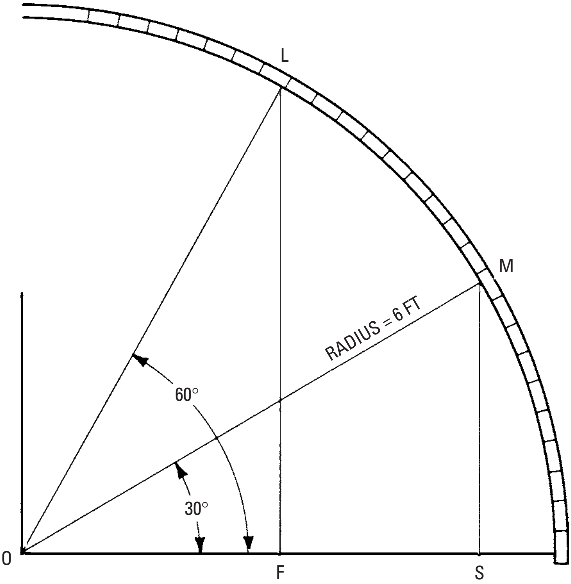

Figure 5-70 The method of finding the length of vertical pieces in grillwork with the aid of trigonometric relations.

For triangle OMS, the hypotenuse is known to be 6 feet, and angle O is 30°. Line MS is the opposite side of the triangle, and

$$\text{opposite side} = 30^\circ, \text{ or}$$

$$\frac{\text{opposite side}}{\text{hypotenuse}} = \text{sine } 30^\circ \times \text{hypotenuse}$$

$$\text{sine } 30^\circ = 0.500$$

This is the calculation:

$$\text{opposite side} = 0.500 \times 6 = 3 \text{ feet}$$

For triangle *OLF*, the hypotenuse is 6 feet, and angle *O* is 60°. Line *LF* is the opposite side, and

$$\frac{\text{opposite side}}{\text{hypotenuse}} = 60°,$$

$$\text{or opposite side} = 60° \times \text{hypotenuse.}$$

$$\text{sine } 60° = 0.866$$

This is the calculation:

$$\text{opposite side} = 0.866 \times 6 = 5.196 \text{ feet} = 5 \text{ feet } 2^3/_8 \text{ inches}$$

Problem 2

When laying out the grillwork in Figure 5-70, how far must the members LF and MS be spaced from the center O to be vertical?

The hypotenuse is known to be 6 feet and the length of adjacent side *OF* is to be found.

$$\frac{\text{adjacent side}}{\text{hypotenuse}} = \cos 60°, \text{ or}$$

$$\text{adjacent side} = \cos 60° \times \text{hypotenuse.}$$

$$\cos 60° = 0.500$$

This is the calculation:

$$\text{adjacent side } OF = .500 \times 6 = 3 \text{ feet}$$

For the length of the adjacent side OS

$$\frac{\text{adjacent side}}{\text{hypotenuse}} = \cos 30°, \text{ or}$$

$$\text{adjacent side} = \cos 30° \times \text{hypotenuse.}$$

$$\cos 30° = 0.866$$

This is the calculation:

$$\text{Adjacent side } OS = .866 \times 6 = 5.196 \text{ feet}$$
$$= 5 \text{ feet } 2^3/_8 \text{ inches}$$

Problem 3

A bridge is to be constructed from the top of a building to an opening in the roof of an adjacent building, as in Figure 5-71. If the rise OF to the point of entry L is 15 feet and the pitch of the roof is $^1/_2$, what length beams FL are required?

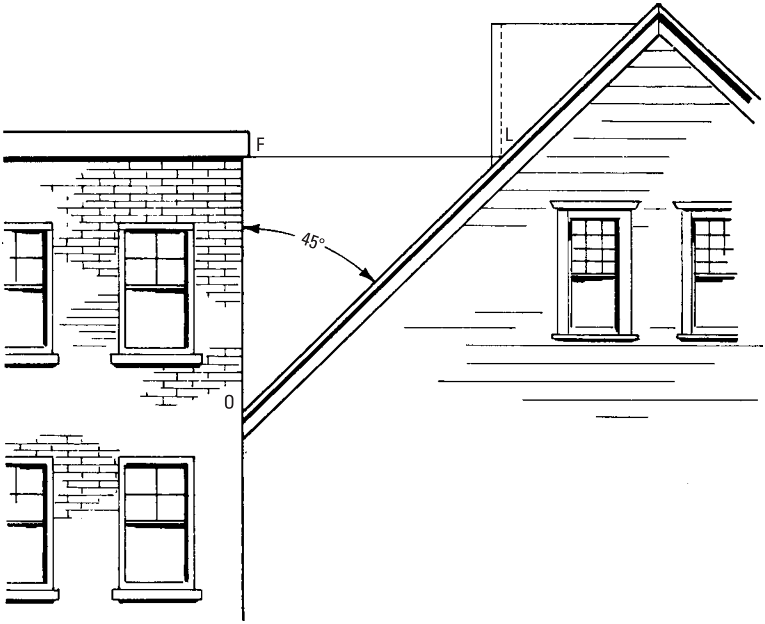

Figure 5-71 The method of finding the distance from one side of a building to a given point on an adjacent roof by employing trigonometric relations.

From Table 5-26, $^1/_2$ pitch, or 12 inches × 12 inches, is 45°. The adjacent side OF is known to be 15 feet. The required length of the opposite side = adjacent side × tan 45°. This is the calculation:

$$\text{opposite side } FL = 15 \times 1.00 = 15 \text{ feet}$$

Problem 4

When estimating the amount of roofing material necessary to cover the side of the roof from O to L in Figure 5-71, what is the distance from O to L?

It is required to find the hypotenuse with the adjacent side and included angle being given.

$$\frac{\text{hypotenuse}}{\text{adjacent side}} = \text{secant angle 0, or}$$

$$\text{hypotenuse} = 1.4142 \times 15 = 21.213 \text{ feet}$$

$$= 21 \text{ feet } 2^9/_{16} \text{ inches.}$$

Table 5-27 Functions of Numbers

							No. = Diameter	
No.	***Square***	***Cube***	***Square Root***	***Cubic Root***	***Logarithm***	***1000 × Reciprocal***	***Circum.***	***Area***
1	1	1	1.0000	1.0000	0.00000	1000.000	3.142	0.7854
2	4	8	1.4142	1.2599	0.30103	500.000	6.283	3.1416
3	9	27	1.7321	1.4422	0.47712	333.333	9.425	7.0686
4	16	64	2.0000	1.5874	0.60206	250.000	12.566	12.5664
5	25	125	2.2361	1.7100	0.69897	200.000	15.708	19.6350
6	36	216	2.4495	1.8171	0.77815	166.667	18.850	28.2743
7	49	343	2.6458	1.9129	0.84510	142.857	21.991	38.4845
8	64	512	2.8284	2.0000	0.90309	125.000	25.133	50.2655
9	81	729	3.0000	2.0801	0.95424	111.111	28.274	63.6173
10	100	1000	3.1623	2.1544	1.00000	100.000	31.416	78.5398
11	121	1331	3.3166	2.2240	1.04139	90.9091	34.558	95.0332
12	144	1728	3.4641	2.2894	1.07918	83.3333	37.699	113.097
13	169	2197	3.6056	2.3513	1.11394	76.9231	40.841	132.732
14	196	2744	3.7417	2.4101	1.14613	71.4286	43.982	153.938
15	225	3375	3.8730	2.4662	1.17609	66.6667	47.124	176.715
16	256	4096	4.0000	2.5198	1.20412	62.5000	50.265	201.062
17	289	4913	4.1231	2.5713	1.23045	58.8235	53.407	226.980
18	324	5832	4.2426	2.6207	1.25527	55.5556	56.549	254.469
19	361	6859	4.3589	2.6684	1.27875	52.6316	59.690	283.529
20	400	8000	4.4721	2.7144	1.30103	50.0000	62.832	314.159

21	441	9261	4.5826	2.7589	1.32222	47.6190	65.973	346.361
22	484	10648	4.6904	2.8020	1.34242	45.4545	69.115	380.133
23	529	12167	4.7958	2.8439	1.36173	43.4783	72.257	415.476
24	576	13824	4.8990	2.8845	1.38021	41.6667	75.398	452.389
25	625	15625	5.0000	2.9240	1.39794	40.0000	78.540	490.874
26	676	17576	5.0990	2.9625	1.41497	38.4615	81.681	530.929
27	729	19683	5.1962	3.0000	1.43136	37.0370	84.823	572.555
28	784	21952	5.2915	3.0366	1.44716	35.7143	87.965	615.752
29	841	24389	5.3852	3.0723	1.46240	34.4828	91.106	660.520
30	900	27000	5.4772	3.1072	1.47712	33.3333	94.248	706.858
31	961	29791	5.5678	3.1414	1.49136	32.2581	97.389	754.768
32	1024	32768	5.6569	3.1748	1.50515	31.2500	100.531	804.248
33	1089	35937	5.7446	3.2075	1.51851	30.3030	103.673	855.299
34	1156	39304	5.8310	3.2396	1.53148	29.4118	106.814	907.920
35	1225	42875	5.9161	3.2711	1.54407	28.5714	109.956	962.113
36	1296	46656	6.0000	3.3019	1.55630	27.7778	113.097	1017.88
37	1369	50653	6.0828	3.3322	1.56820	27.0270	116.239	1075.21
38	1444	54872	6.1644	3.3620	1.57978	26.3158	119.381	1134.11
39	1521	59319	6.2450	3.3912	1.59106	25.6410	122.522	1194.59
40	1600	64000	6.3246	3.4200	1.60206	25.0000	125.66	1256.64
41	1681	68921	6.4031	3.4482	1.61278	24.3902	128.81	1320.25
42	1764	74088	6.4807	3.4760	1.62325	23.8095	131.95	1385.44
43	1849	79507	6.5574	3.5034	1.63347	23.2558	135.09	1452.20
44	1936	85184	6.6332	3.5303	1.64345	22.7273	138.23	1520.53

(continued)

Table 5-27 (continued)

							No. = Diameter	
No.	***Square***	***Cube***	***Square Root***	***Cubic Root***	***Logarithm***	***1000 × Reciprocal***	***Circum.***	***Area***
45	2025	91125	6.7082	3.5569	1.65321	22.2222	141.37	1590.43
46	2116	97336	6.7823	3.5830	1.66276	21.7391	144.51	1661.90
47	2209	103823	6.8557	3.6088	1.67210	21.2766	147.65	1734.94
48	2304	110592	6.9282	3.6342	1.68124	20.8333	150.80	1809.56
49	2401	117649	7.0000	3.6593	1.69020	20.4082	153.94	1885.74
50	2500	125000	7.0711	3.6840	1.69897	20.0000	157.08	1963.50
51	2601	132651	7.1414	3.7084	1.70757	19.6078	160.22	2042.82
52	2704	140608	7.2111	3.7325	1.71600	19.2308	163.36	2123.72
53	2809	148877	7.2801	3.7563	1.72428	18.8679	166.50	2206.18
54	2916	157464	7.3485	3.7798	1.73239	18.5185	169.65	2290.22
55	3025	166375	7.4162	3.8030	1.74036	18.1818	172.79	2375.83
56	3136	175616	7.4833	3.8259	1.74819	17.8571	175.93	2463.01
57	3249	185193	7.5498	3.8485	1.75587	17.5439	179.07	2551.76
58	3364	195112	7.6158	3.8709	1.76343	17.2414	182.21	2642.08
59	3481	205379	7.6811	3.8930	1.77085	16.9492	185.35	2733.97
60	3600	216000	7.7460	3.9149	1.77815	16.6667	188.50	2827.43
61	3721	226981	7.8102	3.9365	1.78533	16.3934	191.64	2922.47
62	3844	238328	7.8740	3.9579	1.79239	16.1290	194.78	3019.07
63	3969	250047	7.9373	3.9791	1.79934	15.8730	197.92	3117.25
64	4096	262144	8.0000	4.0000	1.80618	15.6250	201.06	3216.99

65	4225	274625	8.0623	4.0207	1.81291	15.3846	204.20	3318.31
66	4356	287496	8.1240	4.0412	1.81954	15.1515	207.35	3421.19
67	4489	300763	8.1854	4.0615	1.82607	14.9254	210.49	3525.65
68	4624	314432	8.2462	4.0817	1.83251	14.7059	213.63	3631.68
69	4761	328509	8.3066	4.1016	1.83885	14.4928	216.77	3739.28
70	4900	343000	8.3666	4.1213	1.84510	14.2857	219.91	3848.45
71	5041	357911	8.4261	4.1408	1.85126	14.0845	223.05	3959.19
72	5184	373248	8.4853	4.1602	1.85733	13.8889	226.19	4071.50
73	5329	389017	8.5440	4.1793	1.86332	13.6986	229.34	4185.39
74	5476	405224	8.6023	4.1983	1.86923	13.5135	232.48	4300.84
75	5625	421875	8.6603	4.2172	1.87506	13.3333	235.62	4417.86
76	5776	438976	8.7178	4.2358	1.88081	13.1579	238.76	4536.46
77	5929	456533	8.7750	4.2543	1.88649	12.9870	241.90	4656.63
78	6084	474552	8.8318	4.2727	1.89209	12.8205	245.04	4778.36
79	6241	493039	8.8882	4.2908	1.89763	12.6582	248.19	4901.67
80	6400	512000	8.9443	4.3089	1.90309	12.5000	251.33	5026.55
81	6561	531441	9.0000	4.3267	1.90849	12.3457	254.47	5153.00
82	6724	551368	9.0554	4.3445	1.91381	12.1951	257.61	5281.02
83	6889	571787	9.1104	4.3621	1.91908	12.0482	260.75	5410.61
84	7056	592704	9.1652	4.3795	1.92428	11.9048	263.89	5541.77
85	7225	614125	9.2195	4.3968	1.92942	11.7647	267.04	5674.50
86	7396	636056	9.2736	4.4140	1.93450	11.6279	270.18	5808.80
87	7569	658503	9.3274	4.4310	1.93952	11.4943	273.32	5944.68

(continued)

Table 5-27 (*continued*)

							No. = Diameter	
No.	***Square***	***Cube***	***Square Root***	***Cubic Root***	***Logarithm***	***1000 × Reciprocal***	***Circum.***	***Area***
88	7744	681472	9.3808	4.4480	1.94448	11.3636	276.46	6082.12
89	7921	704969	9.4340	4.4647	1.94939	11.2360	279.60	6221.14
90	8100	729000	9.4868	4.4814	1.95424	11.1111	282.74	6361.73
91	8281	753571	9.5394	4.4979	1.95904	10.9890	285.88	6503.88
92	8464	778688	9.5917	4.5144	1.96379	10.8696	289.03	6647.61
93	8649	804357	9.6437	4.5307	1.96848	10.7527	292.17	6792.91
94	8836	830584	9.6954	4.5468	1.97313	10.6383	295.31	6939.78
95	9025	857375	9.7468	4.5629	1.97772	10.5263	298.45	7088.22
96	9216	884736	9.7980	4.5789	1.98227	10.4167	301.59	7238.23
97	9409	912673	9.8489	4.5947	1.98677	10.3093	304.73	7389.81
98	9604	941192	9.8995	4.6104	1.99123	10.2041	307.88	7542.96
99	9801	970299	9.9499	4.6261	1.99564	10.1010	311.02	7697.69

Functions of Numbers

Table 5-27 shows the functions of numbers.

Summary

All the sciences are based on arithmetic and the ability to use it. Arithmetic is the art of calculating by using numbers. A number is a total amount (or aggregate) of units. By computing the units, we arrive at a certain number or total. Similarly, a unit means a single article, often a definite group adopted as a standard of measurement (such as dozen, ton, foot, bushel, or mile).

Fractions indicate that a number or unit has been divided into a certain number of equal parts, and shows how many of these parts are to be considered. Two forms of fractions are in common usage: the decimal and the common fraction. The common fraction is written by using two numbers, one written over or alongside the other with a line between them, the lower (or second) number being called the denominator, and the upper (or first) number being called the numerator.

Geometry is a branch of mathematics that deals with space and figures in space. It is the science of the mutual relations of points, lines, angles, surfaces, and solids, which are considered as having no properties except those arising from extension and difference of situation. There are two kinds of lines—straight and curved. A straight line is the shortest distance between two points. A curved line is one that changes its direction at every point.

Trigonometry is the branch of mathematics that deals with the relations that exist between the sides and angles of triangles, especially the methods of calculating the required parts of triangles from given parts. There are six elements, or parts, in every triangle: three sides and three angles. The sum of the three angles, no matter what the lengths of the sides, will always be equal to 180 degrees.

Review Questions

1. What is the definition of arithmetic?
2. What are even numbers?
3. What are odd numbers?
4. What are fractions? How are they used?
5. What is trigonometry and how is it used in carpentry?

Chapter 6

Surveying

By definition, *surveying* means the art or science of determining the area and configuration of portions of the surface of the Earth. The two general divisions of surveying may be classified with respect to the nature of the measurements taken:

- Leveling
- Measurement of angles (transit work)

In surveying, *leveling* is the operation of determining the comparative levels of different points of land for the purpose of laying out a grade or building site, and so on, by sighting through a leveling instrument at one point to a leveling staff at another point (see Figure 6-1).

The Level

This instrument (see Figure 6-2) is employed to determine the difference in elevation between points. A common form is known as the *wye level*, so-called because its shape resembles the letter Y. It consists of a telescope mounted on two supports that from their shape are called Ys. The crossbar supporting the telescope is attached to a vertical spindle that allows it to be turned in a horizontal plane. Directly beneath the telescope and attached parallel to it is a spirit level by means of which the line of collimation of the telescope may be rendered horizontal. The *line of collimation* is the line that would connect the intersection of the crosshairs with the optical center of the objective.

Construction of the Wye Level

In construction, a circular plate is screwed to a tripod, and to this is attached a similar plate parallel to the first and connected with it by a ball-and socket joint. Four screws (sometimes only three), called *foot* or *plate screws*, hold these plates apart by resting on the lower one and passing through the other. A vertical spindle in the center of the plates supports a rod, bar, or beam, and is used to revolve the instrument. The beam is horizontal and carries at its ends two vertical standards or supports of equal size terminated by two forks of the general form of the letter Y. The inside of the Ys is Y-shaped, with an open bottom to prevent an accumulation of dirt. The top of the Ys may be closed by semicircular straps or bridles (called *clips*) that are hinged on one side and pinned on the other. The pins are tapered to permit fastening of the telescope. It is *never* clamped tightly.

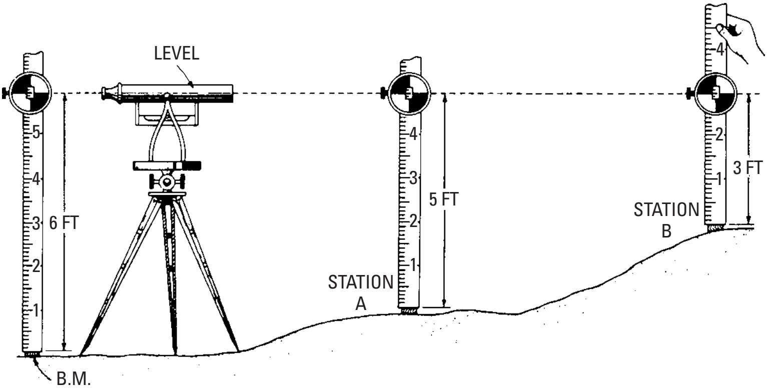

Figure 6-1 The terms backsight and foresight do not necessarily mean backward and forward. Readings taken on a point of known elevation (such as a benchmark or a turning point) are called backsights, whereas readings taken on a point of unknown elevation are termed foresights. In the illustration, the backsight (to the benchmark) is + 6 feet 0 inches, and the foresight to station A is −5 feet 0 inches. The difference in the two elevations is 6 feet −5 feet = 1 foot. If the reading at station A had been greater than the 6-foot HI (height of instrument), the calculation would have shown a negative result, thereby indicating that station A was lower than the benchmark. The elevation at station B is calculated in the same manner: 6 − 3 = 3 feet. Therefore, the elevations from any number of points can be obtained in the same way, if they can be seen from the same position of the instrument. If they cannot, a new HI must be used.

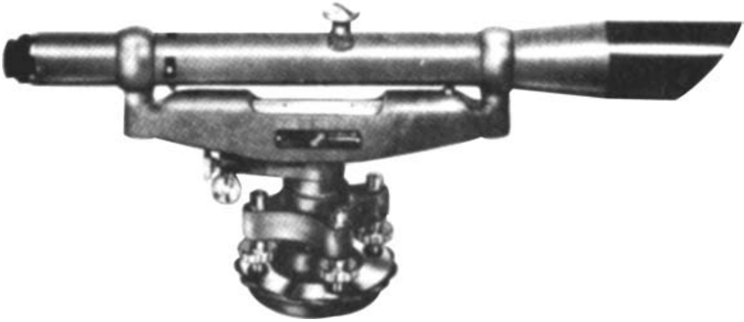

Figure 6-2 A typical builder's level. *(Courtesy David White Instruments, Division of Realist, Inc.)*

The tops of the Ys and the corresponding clips are called the *rings* or *collars*, and should be of equal diameter. A *telescope* is placed on the rings that support a spirit level. A *clamping screw* just above the upper plate serves to secure the instrument in any position desired. A *tangent screw* (also above the upper plate) provides slow motion (or *vernier*) adjustment to the instrument.

The Telescope

The *objective*, or *object glass* (so-called because it faces the object looked at) is a compound lens that is made to correct spherical and chromatic aberrations of single lenses. It gathers light and forms an image at a point in the tube where crosshairs are placed. The *ocular* (or eye) *piece* is also a compound lens through which the operator looks to see a magnified view of the image. In the best precision instruments (often foreign-made), the image is often *inverted*. A good instrument worker quickly becomes accustomed to the inverted image, but most American-made instruments have an *erecting* image that shows the object right side up. Tangent screws may be used to give motion to the tubes carrying the objective and ocular.

The Crosshairs

These are made of platinum-drawn wires or spider's threads attached to a ring within the telescope at the spot where the image is formed. The ring is secured by four capstan-headed screws that pass through the telescope tube. There are commonly two hairs, one horizontal and the other vertical, with their intersection in the axis of the telescope.

Bubble Level

The spirit level attached to the telescope can be raised vertically by means of *altitude screws* at the rear end, and it may be moved laterally to a limited extent by means of *azimuth screws* at the forward end.

The Supports

These form the Ys and are supported by the bar to that they are fastened by two nuts, one above and one below. These nuts may be moved to provide an adjustment (to move the scope in a horizontal direction).

Lines of the Level

The three principal lines of a level are as follows:

- *Vertical axis*—This passes through the center of the spindle.
- *Bubble line*—The metallic supports of the spirit level are equal, and the tangent at their top or bottom is horizontal

when the bubble is centered. This tangent is the bubble line.

- *Line of collimation*—The line that would connect the intersection of the crosshairs with the optical center of the objective is the line of collimation.

The following relations must be obtained:

- The bubble line and the line of collimation must be parallel.
- The plane described by the bubble line should be horizontal, that is, perpendicular to the vertical axis.

These conditions are generally satisfied in a new level, but exposure and use may alter these relations. Therefore, occasional adjustment of the instrument may be necessary.

Adjustments of the Wye Level

Levels and transits are expensive, precision instruments and should be treated as such. Although a passable job of leveling may be done by a relatively inexperienced worker, it is questionable if a major job of adjusting should be attempted by a novice. A perfect job of adjustment is difficult, even for an experienced adjuster, and there are few instruments in perfect adjustment. For precision work, the adjustment should be checked constantly. The first relation given earlier cannot be established directly. However, it does require several adjustments.

First Adjustment

Collimating the instrument means making the line of collimation parallel to the bottom element of the collars

1. Clamp the instrument, and unclip the collars.
2. Sight at some distant point, a point that is distinct.
3. Bring the horizontal crosshair on that point.
4. Carefully turn the telescope in the collars by one-half a revolution around its axis, and sight again. If the horizontal crosshair is still on the sighted point, the telescope is collimated with regard to that crosshair. If it is off the point, bring it halfway back by means of the capstan-headed screws and the rest of the way by the plate screws.
5. Repeat the operation over another point.
6. Collimate it with regard to the other crosshair.
7. Leave the screws at a snug bearing.

Second Adjustment

This is where you set the bubble line in a plane with the bottom element of the collars.

1. Unclip the telescope, and clamp the instrument over a pair of plate screws.
2. Center the bubble by means of the plate screws.
3. Carefully (and very slowly) turn the telescope in the collars in a small arc to the right, then to the left. If the bubble moves from center, bring it back by means of the azimuth or side screws.

Third Adjustment

Setting the bubble line parallel to the bottom element of the collars.

1. Unclip the telescope, and clamp the instrument over a pair of plate screws.
2. Center the bubble by means of the plate screws.
3. Carefully take the telescope up, replacing it carefully in the Ys in the opposite direction (that is, the objective sighting in the direction where the eyepiece originally was). If the bubble has moved, bring it back halfway by means of the altitude or foot screws of the spirit level and the rest of the way by the plate screws.
4. Repeat in another direction until the adjustment is satisfactory.

The second relation is established by making the bubble line stay in the center of the graduation during a complete revolution of the instrument around its spindle.

Fourth Adjustment

Now, make the axis of the instrument (not of the telescope) vertical.

1. Pin the clips.
2. Clamp and center the bubble over a pair of plate screws.
3. Reverse the telescope over the same pair of plate screws.
4. Bring the bubble halfway back (if it has moved) by means of the plate screws.

Fifth Adjustment

Again, make the bubble remain centered during a full revolution of the instrument.

1. Center the bubble.

2. Revolve the instrument horizontally by a one-half revolution.
3. If the bubble moves, correct it halfway by means of the support screws (at the foot of the Ys). If the rings become worn and unequal, use the two-peg method of the dumpy level.

Dumpy Level Adjustment

Figure 6-3 shows the *dumpy level* (so-called because of its compactness). It is used mostly in England, although it is used to some extent in the United States because of the better stability of its adjustments over the wye level. The dumpy level differs from the wye level mainly in that the telescope of the dumpy level is permanently attached to the supports or uprights, but these uprights are adjustable. The two-peg adjustment method is as follows:

1. Drive two stakes (pegs) several hundred feet apart.
2. Set the instrument approximately halfway between them.
3. Level up and sight the rod that is held in succession on each stake. The difference in the readings is the true difference of the

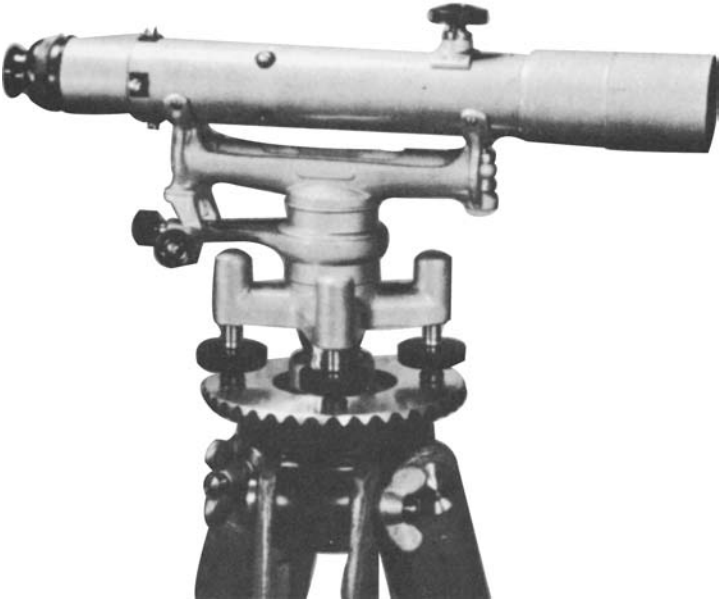

Figure 6-3 A dumpy level.

elevation of the stakes, even if the instrument is not in proper adjustment.

4. Test the instrument.
5. Set it near one of the stakes (the highest one, for instance). Level up and sight the rod held on the other stake.
6. Subtract the height of the instrument from the reading. The difference should be equal to the difference of elevation of the stakes as previously found. If these differences are not equal, set the target halfway between these readings, sight on it, and center the bubble by means of the altitude screws.
7. Repeat the operation until satisfaction is obtained.

Centering the Objective and Ocular

These adjustments are made permanently by the manufacturer. Usually, four screws hold the tubes carrying the glasses. Their heads pass through the outside tube where, after permanent adjustment, they are covered by a metallic ring.

Parallax

This is the apparent motion of the crosshairs on the object sighted when the eye is moved slightly. It shows the imperfect focusing of the ocular over the crosshairs. To correct this condition, hold a white surface (such as that of a piece of paper) slightly in front of the objective, and move the ocular tube in and out until the crosshairs are perfectly defined.

Leveling Rod

This instrument, used in leveling, is usually 6$^1/_2$ feet high, graduated to hundredths of a foot and provided with a sliding target. The rod is made in two parts, arranged so that its length can be extended to 12 feet. Precision rods are of one-piece construction and have no target. Builders' rods may be graduated in feet, inches, and eighths of an inch, with a vernier reading in 64ths of an inch. A sliding disc (called a *target*) is provided with a vernier for extremely accurate work, reading to thousandths of a foot.

In use, the rod is held in a vertical position with its lower end resting on the desired point of elevation. The target is then moved up and down until its center coincides with the crosshairs in the telescope of the level. The reading of the elevation is made from the rod on a line corresponding with the centerline of the target. There are various kinds of rods. Some are designed to be read by the rodman, while others can be read through the telescope of the level.

Methods of Leveling

The simplest type of leveling is to find the difference in level between two points that are visible from a third point, the difference in level being less than the length of the leveling rod (see Figure 6-4).

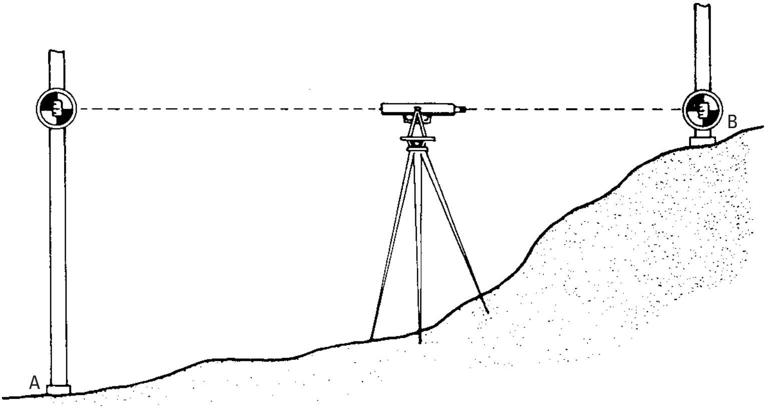

Figure 6-4 Leveling between two points whose difference in level is less than the length of the rod.

1. Set up and level the instrument at some point approximately halfway between the two points.
2. Have a worker hold the rod vertically on one of the points, and move the target up and down until its center coincides with the crosshairs of the level.
3. Take a reading. This is the HI, or height of the instrument above the benchmark *A*.
4. Turn the telescope on its spindle, have the rod held on the other point, and take a similar reading at *B*. The difference in level is equal to the difference in the readings.
5. If the difference in level is greater than the length of the rod, use the method shown in Figure 6-5. Divide the distance between the two points into sections of such length that the difference in level between the dividing points *A*, *B*, and *C* (called *stations*) are less than the length of the rod. Set up and level between points *A* and *B*.
6. Measure the distance *Aa* that is called *backsight*. Then, reverse the telescope, and take reading *Bb* that is called *foresight*.

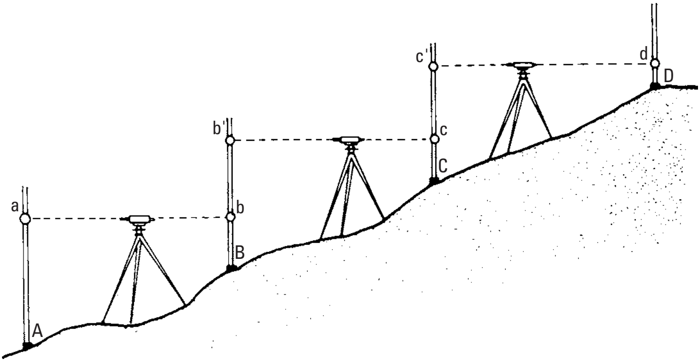

Figure 6-5 Leveling between points whose difference in level is greater than the length of the rod.

7. Next, set up and level between *B* and *C*, and take readings *Bb′* and *Cc*.
8. Repeat the operation between *C* and *D*, taking readings *Cc′* and *Dd*. The difference in level between stations *A* and *D* is equal to the sum of the differences between the intermediate stations. That is, this difference equals $(Aa - Bb) + (Bb' - Cc) - (Cc' - Dd)$, or $(Aa + Bb' + Cc') - (Bb + Cc + Dd)$.

Usually, you'll have to find the relative elevations of several points, as in grading work, in which case it is necessary to keep more elaborate notes and to measure distances between the stations. The method employed for this type of leveling is shown in Figure 6-6, and the field notes are recorded as shown in Table 6-1.

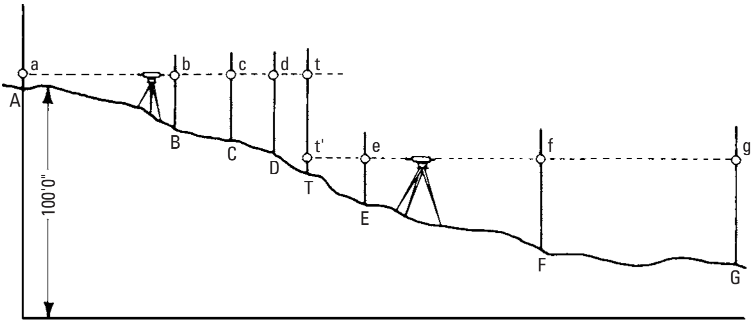

Figure 6-6 Finding the relative elevations of several points in rough terrain.

Table 6-1 Field Notes (Corresponding to the Operations Illustrated in Figure 6-6)

Station	*Distance*	*Backsight*	*Height of Instrument*	*Foresight*	*Elevation*	*Remarks*
A	0	4.2	104.2	.....	100.0	Bench mark, top of hydrant.Turning point
B	100	.....	.....	10.1	94.1	
C	60	.....	.....	7.3	96.9	
D	50	.....	.....	5.8	98.4	
T	.....	4.1	99.1	9.2	95.1	
E	70	6.8	.....	.....	92.3	
F	110	.....	.....	9.5	89.6	
G	80	.....	.....	11.5	87.6	

1. Assume a datum or reference line below the elevation of the lowest station, and refer all elevations to this line. Start at some permanently fixed point (such as a mark on a building or the top of a hydrant). This is called a *benchmark*.
2. Let *A* in Figure 6-6 be the benchmark, and assume a datum line 100 feet below the level of *A*.
3. Start with the instrument between *A* and *B*, and take a backsight on *A*. The distance *Aa* is found to be 4.2 feet, which, when added to 100 feet, gives the height of the instrument.
4. Next, take foresights on *B*, *C*, and *D*, and record these readings in the proper column. Readings *Bb, Cc*, and so on, subtracted from the height of the instrument, will give the elevations at *B*, *C*, and so on. This is done, and the results are recorded in the proper column of field notes.
5. The ground falls away so rapidly beyond *D* that it is necessary to set up the level farther along and, therefore, establish a new height of instrument. This is done by holding the rod at some convenient point (such as at *T*) called the *turning point*, and taking a foresight, which measures the distance *Tt* (9.2 feet).
6. The level is then set up in its second position between *E* and *F*, and a backsight is taken on the rod in the same position, which gives the distance Tt' (4.1 feet).
7. The distance $t't$ then equals $9.2 - 4.1 = 5.1$ feet, and this is subtracted from the previous height of instrument, thus giving the new *HI*, that is $104.2 - 5.1 = 99.1$ feet.

8. A backsight is now taken on E, and foresights are taken on F and G. These are recorded in the proper columns, and the elevations are found by subtracting these distances from the new HI.
9. The horizontal distances between the stations are measured with a tape and recorded in the second column.
10. When plotting a cross-section from notes kept in this manner, the datum line is drawn first, and perpendiculars are erected at points corresponding to the different stations. The proper elevations are then indicated on these vertical lines, and a contour line is drawn through the points so marked.

Directions for Using Level

Note carefully the following mode of procedure in leveling:

1. Center the bubble over one pair of plate screws, then over the other pair. Plate screws should have a snug bearing. When looking at the bubble or at the crosshairs, the eyes should look naturally (that is, without strain). Try to observe with both eyes open.
2. Adjust the eyepiece to the crosshairs for parallax.
3. Turn the instrument toward the target. It is better to level up facing the target.
4. Look again at the bubble.
5. Sight the target, and have it set right by motions according to a prearranged code with the worker holding the rod.
6. Look again at the bubble.
7. Read the rod or direct the target from the intersection of the crosshairs only.
8. Approve the target when sure.
9. Have the height of the target called out by the worker holding the rod.
10. Enter this height in the field book.
11. Quickly, make needed calculations.
12. Motion the worker holding the rod to a new station or to stay for a turning point and backsight, and move yourself to another position.

The following additional hints will also be useful:

- *Guarding against the sun*—Draw the telescope shade, or use an umbrella or a hat.

- *Length of sights*—Avoid sights too short and too long (250 to 350 feet should be the limit of the sights).
- *Equal sights*—The length of the backsight should practically be the same as the length of the foresight. This may be approximated by pacing, or by sighting with the stadia crosshairs in the telescope.
- *Long sights*—When sights longer than the maximum allowable in one direction only are unavoidable, correction should be made for curvature.
- *Leveling up or down a steep slope*—The leveler, after some practice, will place the instrument to take a reading near the top or the bottom of the rod (as the case may be), thus gaining vertical distance, but this produces unequal sights. The leveler may also follow a zig-zag course.
- *Leveling across a large body of water*
  - *A running stream*—Drive a stake to the water surface on each side of the stream and in a direction normal to the flow, although the line may not run so. Take a foresight reading on the first, a backsight reading on the second, and continue to and along the line. The elevations of the two stakes may be assumed equal.
  - *Across a pond*—If a pond or lake is too wide to ensure a good sighting across, use essentially the same method as for a stream. Drive stakes on each side and to the water surface. Take a foresight reading on the first and a backsight reading on the second.
- *Across a wall*—Take a foresight reading on the rod set on a stake, driven to the natural surface on the first side of the wall. Measure the height of the wall above the stake, and enter it as a backsight reading. Drive a stake to the natural surface on the second side of the wall. Measure the height of the wall on that side above the stake. Enter it as a foresight reading. Set the rod on the stake. Take a sight on it (which will be a backsight reading). Continue using this method until the leveling has been completed.
- *In underbrush*—If it cannot be cut down on the line of sight, find a high place or provide one by piling logs, rocks, and so on, to set the instrument on.
- *Through swamp*—Push the legs of the tripod down as far as possible. The leveler lies on his or her side. Two workers may

be necessary at the level. If the ground is still unsafe, drive stakes or piles to support the instrument.

- *Elevations taken at road crossings*—Take elevations both ways for some distance.
- *Elevations taken at river crossings*—Take elevations of high-water marks and flood marks, with the dates of it. Question the local residents for these dates. Also, ask for dates and data of extreme low water.
- *Proper length of sights*—This will depend on the distance at which the rod appears distinct and on the precision required. Under ordinary conditions, sights should not exceed 300 feet where elevations are required to the nearest .01 foot, and even at a much shorter distance, the boiling of the air may prevent a precision reading of this degree.
- *Correction for refraction and earth curvature*—A level line is a curved line at which every point is perpendicular to the direction of gravity, and the line of sight of a leveling instrument is tangent to this curve. This makes it necessary to consider this curve in some leveling operations. If reasonable care is used to make the lengths of backsights and foresights approximately equal, this aberration is self-correcting. However, in extremely long lines, it is approximately 2 inches in one-half mile, or about $^{2}/_{3}d^{2}$, in feet, where d is equal to the distance, in miles. This correction is usually combined with that for refraction. The combined correction is $547d^{2}$, and it is *negative*.

Trigonometric Leveling

Finding the difference in elevation of two points by means of the horizontal distance between them and the vertical angle is called *trigonometric leveling*. It is used chiefly in determining the elevation of triangulation stations and in obtaining the elevation of a plane-table station from any visible triangulation point of known elevation.

In triangulation work, the vertical angles are usually measured at the same time the horizontal angles are measured to obtain the elevations of triangulation points as well as their horizontal positions. The vertical angle is measured to some definite point on the signal whose height above the center mark of the station was determined when the signal was erected. The height of the instrument above its station should be measured and recorded.

In the most exact work, the angles are measured with a special vertical circle instrument. In less precise work, an ordinary Theodolite, whose vertical arc reads by verniers to 30 seconds or to 20 seconds, may be used, but with such instruments, only single readings can be made. The best results with such an instrument are obtained by taking the average of several independent readings, one-half of which are taken with the telescope direct and the other half with the telescope inverted. In every case, the *index correction* (or reading of the vertical arc when the telescope is level) must be recorded.

The Transit

This instrument is designed and used for measuring both horizontal and vertical angles. It consists of a telescope mounted in standards that are attached to a horizontal plate (called the *limb*). Inside the limb, and concentric with it, is another plate (called the *vernier plate*). The lower plate or limb turns on a vertical spindle or axis that fits into a socket in the tripod head. By means of a clamp and a tangent screw, it may be fastened in any position and made to move slowly through a small arc.

The circumference of this plate is usually graduated in divisions of either one-half or one-third of one degree, and in the common form of transit, these divisions are numbered from one point on the limb in both directions around to the opposite point, which is 180 degrees. The graduation is generally concealed beneath the plate above it, except at the verniers. This upper plate is the vernier plate that turns on a spindle fitted into a socket in the lower plate. It is also provided with a clamp by means of which it can be held in any position and with a tangent screw by which it can be turned through a small arc. A *vernier* is a device for reading smaller divisions on the scales than could otherwise be read.

The transit is generally provided with a compass so that the bearing of any given line with the magnetic meridian may be determined if desired. It also has a spirit level attached to the telescope, so that it may be brought to a horizontal position and made to serve as a level. Figure 6-7 shows a typical transit.

Construction of the Transit

Figure 6-8 shows the general features of the transit construction. The following sections describe these features.

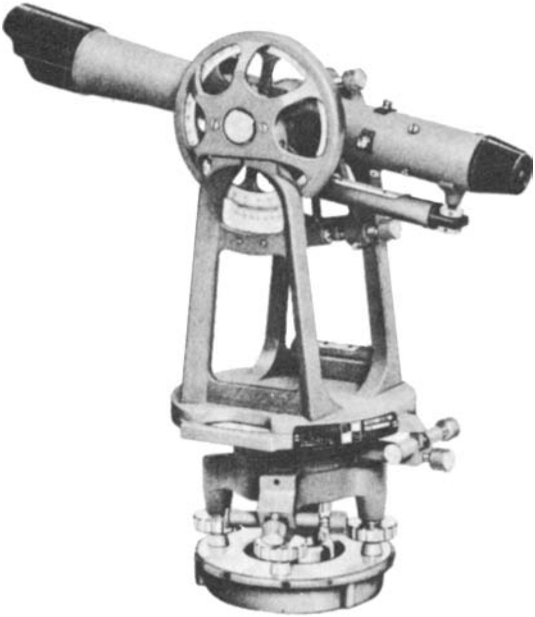

Figure 6-7 Typical transit.

Parallel Plates

There are two *plates*, one upper and one lower. The lower plate *A* is generally formed with two parts. The outside part is a flat ring and is screwed to the tripod head. The inside part is another flat ring of a diameter larger than the opening in the outside part and has a central dome, *C*, that is perforated on the top. The inside part is movable and rests on the under side of the outside part. The upper plate *B* is generally made in the form of a central nut, with four arms at right angles (or three at 120°). The upper plate carries an inverted conical shell, the lower portion of which passes through the perforation in the dome of the inside part of the lower plate, where it expands into a spherical shape and thus forms a ball joint with the lower plate. This spherical member is perforated in the center to allow the passage of a plumb-bob string.

Foot Screws

The two plates are connected by four-foot (sometimes only three) screws *F* in order to clamp the lower and upper plates (making them fast with each other and with the inverted shell) and to serve

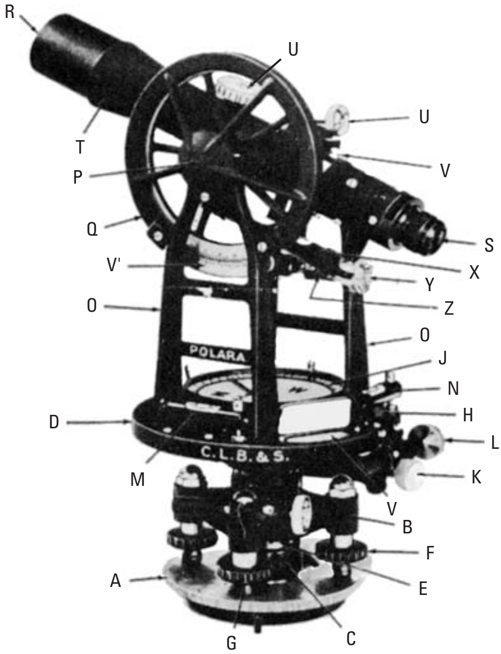

Figure 6-8 The transit. In the illustration, A represents the lower plate; B, the upper plate; C, the central dome; D, the divided limb; E, the spindle; F, the foot screws; G, foot-screw cups; H, the vernier plate; J, the compass circle; K, the clamp-screw vernier plate to divided limb; L, the tangent screw; M and N, spirit levels; O, standards or supports; P, the horizontal shaft; Q, the vertical arc; R, the objective; S, the ocular; T, the telescope; U, racks and pinions; V, the adjustable cross-hair ring; v, the divided-limb vernier; v′, the vertical-arc vernier; X, the spirit level; Y, the gradienter; and Z, the scaled index.

in leveling the instrument. The screws pass through the ends of the arms of the upper plate. They are surmounted by dust caps. There the feet fit into small cups, *G*, that rest on the top surface of the lower plate to avoid wear.

Shifting Center

Since these cups, as well as the central part of the lower plate, may be moved (after slightly loosening the foot screws), a slight motion may be given to the instrument to better set it over a given point of the ground. This arrangement is called a *shifting center*.

Outer Spindle

A second conical shell fits and may revolve in the conical shell attached to the upper plate. It is the outer spindle, and it carries projections to form attachments with the other parts of the transit.

Divided Limb

The upper portion of the outer spindle terminates in a horizontal disc of plate *D*, the limb of which is divided into 360°, subdivided into one-half, one-third, or one-quarter of one degree. Every ten degrees are numbered, either from 0° to 360° or from 0° to 180°, either way. The degree marks are a little longer than the subdivisions, and every fifth degree has a mark slightly longer yet.

Lower Motion

The outer spindle and the divided limb are also called the *lower motion*.

Inner Spindle

A solid inverted cone fits into the outer spindle and may revolve in it. It is the *inner spindle*, and, like the outer one, it is provided with some projections for similar purposes.

Vernier Plate

The upper portion of the inner spindle projects farther than the divided limb and also carries a horizontal disc, H, that moves in a plane parallel to the divided limb (that it covers), except for two rectangular openings in opposite directions through which the divisions of the limb may be seen. These openings each carry a vernier *v* by means of which the subdivisions of degrees are again divided. Some verniers read to 1 minute, others to 0.5 minute, and some to 10 seconds. To facilitate the reading of the vernier, the openings are sometimes fitted with a reflector and a magnifying glass.

Upper Motion

The inner spindle and vernier plate *H* are also called the *upper motion*. The vernier plate carries a compass circle, shown at *J*.

Compass Circle

This consists of a circular box, the bottom of which carries at its center a sharp pivot of hard metal (hard steel or iridium) on which a magnetic needle approximately 5 inches long is balanced by an agate cup fixed in the middle of its length (see Figure 6-9). The needle

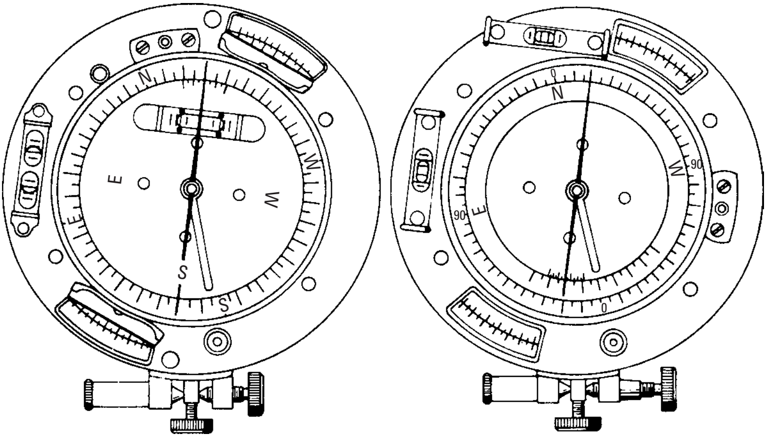

Figure 6-9 Two typical compass boxes. The compass circles are graduated to one-half degrees and numbered in quadrants. The variation plate is provided so that the magnetic declination may be set accurately. The cardinal points shift with the graduated circle. *(Courtesy C. L. Berger & Sons, Inc.)*

is strongly magnetized. Its north end is distinguished by color or ornamentation, and its balance is regulated by a small coil of fine wire wound around one arm that can be shifted. The limb that is formed by the edge of the sides of the box is divided into 360° with half degrees shown. They are numbered from two zeros marked at the ends of a diameter to 90° right and left. The bottom of the box is marked with two rectangular diameters corresponding to the graduations 0° and 90° of the vernier and two other diameters at 45° to the first.

The forward end of the diameter marked 0° is designated by the letter N, and the rear end is designated by the letter S, corresponding to north and south. The ends of the transverse diameter marked 90° are designated by the letters E on the left and W on the right, corresponding to east and west. Note that this designation is the reverse of the standard mariner's compass. Since the telescope is fixed to sight from south to north, the compass indicates the direction of the sighting. When set to an ordinary surveyor's compass, the forward end of the frame carries a vernier and a tangent screw to read fractions smaller than one-half degree.

Controlling Clamps

A screw K permits clamping of the vernier plate H to the divided limb D. Another screw attached to the upper plate permits clamping of the divided limb to the upper plate.

Tangent Screws

One *tangent* (or slow motion) *screw* L accompanies each clamp screw. It is used to complete the clamping at the exact spot where the clamp is to be made.

Spirit Levels

The *spirit levels* are attached to the vernier plate—one level M in front (north point of the box), the other N on the side—thus forming an angle of 90°.

Standards

The vernier plate carries two vertical *standards* or supports O, which are shaped like an inverted V and placed one on each side. The center of their legs is just opposite the 90° graduation of the compass box. They are made equal.

Horizontal Axis

The standards carry between and on the top of them a movable *horizontal axis* P.

Vertical Code

To the horizontal axis is attached, by means of a clamp screw, a vertical circle or arc, Q, that is divided like the horizontal circle in its vertical motion just touches a circular vernier v' carried by the left standard together with a slow-motion screw.

Telescope

In the middle of the horizontal axis and perpendicular to it is attached a *telescope* T of a description similar to that of the engineer's level, with an objective R and an ocular S, racks and pinions U for their motions, and an adjustable crosshairs ring V, with ordinary and stadia hairs.

Telescope Level

An adjustable spirit level X is also attached to the under part of the telescope, as in the engineer's level. This permits the transit to also be used as a leveling instrument, if necessary.

Motions of the Telescope

The telescope can function over the full range of the horizon and can measure any horizontal angle. In addition, since the telescope is on a horizontal axis endowed with free motion, it may move in a vertical plane carrying with it the vertical arc, and it can, therefore, measure vertical angles. In the horizontal motion, the vertical crosshair of the telescope is brought exactly on the point sighted by means of the slow-motion screw *L*, carried by the vernier plate *H*. In the vertical motion, the horizontal crosshair of the telescope is brought exactly on the point sighted by means of the slow-motion screw carried on the inside of the left-hand support and by moving the vertical circle.

Lines of a Transit

The following are the principal lines of a transit:

- Vertical axis
- Horizontal axis
- Plate level line
- Attached level line
- Line of collimation

Vertical Axis

This vertical line passes through the center of the spindle *E* (see Figure 6-8).

Horizontal Axis

The axis *P* (see Figure 6-8) of the shaft by which the telescope rests on the supports; it must be made horizontal.

Plate Level Line

The top or bottom lines of the plate level case *N* (see Figure 6-8). These are level when the bubble is centered.

Attached Level Line

The level line of the bubble level *X* (see Figure 6-8) attached to the telescope. It is employed only when the instrument is used as an engineer's level.

Line of Collimation

The line determined by the optical center of the objective and the intersection of the crosshairs.

Relations Between the Lines of a Transit

The following relations must be obtained:

- The plate levels must be perpendicular to the vertical axis.
- The line of collimation must be perpendicular to the horizontal axis.

- The horizontal axis must be perpendicular to the vertical axis.
- The attached level line and the line of collimation must be parallel.
- The zero of the vertical circle must correspond to the zero of the vernier when the telescope is horizontal.

Adjustments of the Transit

The following sections describe the necessary adjustments of the transit.

First Adjustment

Make the axis of the spindle vertical and the planes of the plates perpendicular to it.

1. Set one level over a pair of plate screws. The other level will thus be set over the other pair.
2. Level up both levels by means of the plate screws.
3. Turn the vernier plate around by a one-half revolution. If the bubbles remain centered during the motion, the vernier plate is in adjustment. If they have moved, bring them halfway back by means of the adjusting screws and the rest of the way by means of the foot screws.
4. Repeat the operation, and determine if the bubbles remain centered when revolving the divided circle. If they do not, the plates are not parallel, and the transit must be sent to the manufacturer for repairs.

Second Adjustment

Collimate the telescope.

1. Set up the transit in the center of open and practically level ground. Carefully level the instrument.
2. Drive a stake or pin approximately 200 or 300 feet away.
3. Measure the distance.
4. Take a sight on that point, and clamp the plates.
5. Revolve the telescope vertically (in altitude) by one-half a revolution, thus reversing the line of sight.
6. Measure in the new direction the same distance as first measured, and drive a pin.
7. Unclamp and revolve the vernier plate by one-half a horizontal revolution.

8. Sight again at the first point and clamp.
9. Again, revolve the telescope vertically by one-half a revolution. If the line of sight falls on the pin, the telescope is collimated. If not, drive a new pin on the last sight at the same distance as before, and drive another pin at one-fourth the distance between the first pin and the second.
10. Move the vertical crosshair by means of the capstan-headed screw and an adjusting pin, until the intersection of the crosshairs covers the last pin set.
11. Repeat the operation to be certain of collimation.

Third Adjustment

Adjust the horizontal axis so that the line of collimation will move in a vertical plane.

1. Level up carefully and sight on a high, well-defined point (such as a corner of a chimney) and clamp.
2. Slowly move the telescope down until it sights the ground, and drive a pin there.
3. Unclamp. Revolve the vernier plate one-half of a revolution, and revolve the telescope vertically one-half of a revolution, thereby reversing the line of sight.
4. Look again at the high point and clamp.
5. Slowly move the telescope down until it sights the ground. If the intersection of the crosshairs covers the pin, the horizontal axis is in adjustment. If not, correct halfway by means of a support-adjusting screw and the rest of the way by means of the plate screws.
6. Repeat the operation, and verify the adjustment.

Fourth Adjustment

Make the line of collimation horizontal when the bubble of the attached level is centered.

1. Drive two stakes 300 to 400 feet apart, and set up the instrument approximately halfway between these stakes.
2. Level up and take readings on the rod held successively on the two stakes. The difference between the readings is the difference of elevation of the stakes.
3. Next, set the transit over one of the stakes, level up, and take a reading of the rod held on the other stake.

4. Measure the height of the instrument. The difference between this and the last rod reading should equal the difference of elevation as previously determined. If it does not, correct the error halfway by means of the attached level-adjusting screw.
5. Repeat the operation, and verify the adjustment.

Fifth Adjustment

Make the vernier of the vertical circle read zero when the bubble of the attached level is centered.

1. Level up the instrument.
2. Sight on a well-defined point, and take note of the reading on the vertical circle.
3. Turn the vernier plate one-half of a revolution, and also turn the telescope vertically one-half of a revolution.
4. Again, sight on the same point.
5. Read and record the reading on the vertical circle.

One-half the difference of the two readings is the index error. The error may be corrected by moving either the vernier or the vertical circle, or the error may be noted and applied as a correction to all measurements of vertical angles.

Adjustments of the Compass

It will sometimes be necessary to adjust the compass. When an adjustment is required, it may be accomplished by using the following procedure.

First Adjustment

Straighten the needle.

1. Examine to see if the ends of the needle are set on opposite divisions. If not, fix the pivot so that they are.
2. Revolve the box by one-half of a revolution. If the needle does not set on opposite divisions, bend both ends by one-half the difference.

Second Adjustment

Place the pivot in the center of the plate. If the needle is straight, move the pivot until the needle sets on opposite divisions at points such as 0°, 45°, and 90°.

Instructions for Using the Transit

The transit requires various adjustments, as explained in the preceding section. To center the transit over a stake, rest one leg of the tripod on the ground, then grasp the other legs and place the instrument as nearly over the stake as possible. Then attach the plumb bob, and center it accurately by means of the shifting head. Avoid having the plates too much out of level because this will result in unnecessary straining of the leveling screws and plates.

Once the instrument has been centered over the stake, level it up by the spirit levels on the horizontal plate. To do this, turn the instrument on its vertical axis until the bubble tubes are parallel to a pair of diagonally opposite plate screws. Then, stand facing the instrument and grasp the screws between the thumb and forefinger; turn the thumbscrew in the direction the bubble must move. When adjusting the screws, turn both thumbscrews in or out, never in the same direction. Adjusting one level will disturb the other, but each must be adjusted alternately until both bubbles remain constant.

Figure 6-10 shows the method of measuring a horizontal angle. The process of laying-off a given angle is similar to that of measuring

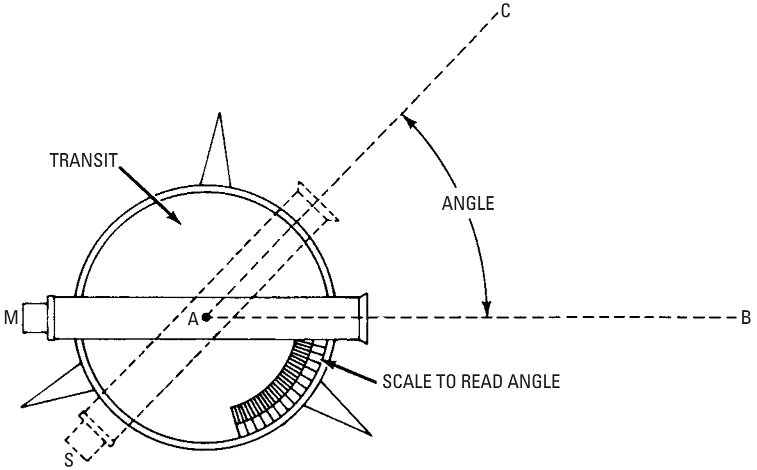

Figure 6-10 The general principle of transit work. The transit is placed over the apex A of the angle CAB that is to be measured. The telescope is sighted to stake B (position M), and a reading is taken. It is then turned horizontally and sighted to stake C (position S), and another reading is taken. The difference between these readings gives angle CAB.

the angle. The transit is set up at the vertex of the angle, the vernier is clamped at zero, and the telescope is pointed at the target, thereby marking the direction of the fixed line. The limb is now clamped, the vernier is unclamped, and the vernier plate is turned through the desired angle and clamped. A stake should now be driven in line with the vertical crosshair in the telescope, thus establishing the two sides of the angle.

When laying out the foundations of buildings, a corner stake is first located by measurement. Then the direction of one of the walls is laid out by driving a second stake. This direction may be determined by local conditions (such as the shape of the lot or the relation to other buildings). If the building is to be an extension to (or in line with) another building, the direction can be obtained by sighting along the building wall and driving two stakes in line with it. If it is to make a given angle with another building, this angle can be laid off as shown in Figure 6-11.

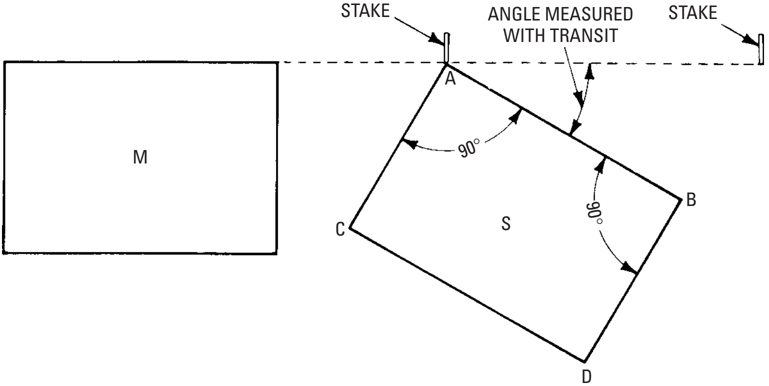

Figure 6-11 The method of laying out a new building (S) at a given angle with an old building (M). After the corner and the direction of one wall are determined, a right angle may be laid off (if the building is rectangular), thus locating two of the sides (AB and AC). The length of side AB is then measured, thereby locating corner B. The transit is set up at B, and line BD is laid off at right angles to AB. AC and BD are then laid off by the proper length, and the four corners of the building are thus located. If the building had not been rectangular, the proper angles could have been laid off instead of right angles.

Gradienter

Some transits carry a device called a *gradienter* (*Y* in Figure 6-8) that is attached to the horizontal axis by means of a clamp screw and is inside of the right-hand support. It is designed and employed for the determination of grades and distances and consists of an arm in the shape of an inverted Y with curved branches. To the extremities of this arm are attached an encased spiral spring and a nut through which moves a micrometer screw with a graduated head that revolves in front of a scaled index (*Z* in Figure 6-8) that is also carried by the arm. The ends of the screw and the spring are on opposite sides of a shoulder that is carried by the right-hand support. The head is divided into tenths and hundredths, and every revolution moves it in front of the scale by one division so that the scale gives the number of turns of the screw, and the graduated head gives the fraction of a turn.

In grading, if one revolution of the screw moves the crosshair a space of 1 foot on a rod held 100 feet away, the scope indicated by the telescope is 1 percent. To establish a grade, level up the telescope, clamp the arm of the gradienter, and turn the micrometer screw by as many divisions as are required in the grade. For example, to set the gradienter at 2.35, move the head two complete turns plus 35 subdivisions. Measure the height of the telescope from the ground. Set the rod at that height. Then hold the rod at any point on the line, raising it until the target is bisected by the crosshairs. The foot of the rod will then be on the grade.

Care of Instruments

With proper care, the usefulness of an instrument can be preserved for many years. Therefore, the following suggestions on the care of instruments should be noted:

- The lenses of the telescope, particularly the object glass, should not be removed, since this will disturb the adjustment. If it is necessary to clean them, great care should be taken, and only soft, clean linen should be used.
- To retain the sensitivity of the compass needle, the delicate point on which it swings must be carefully guarded, and the instrument should not be carried without the needle being locked. When the needle is lowered, it should be brought gently on the center pin.
- The object slide seldom needs to be removed. When removal is necessary, the slide should be carefully protected from dust.

Do not grease or oil the slide too freely; only a thin lubricant film is necessary. Any surplus of oil should be removed with a clean wiper.

- The centers, subject to considerable wear, require frequent lubrication. After a thorough cleaning, they should be carefully oiled with fine watch oil. All of the adjusting screws should be brought to a fine bearing, but they should never be tightened to such a degree that a strain is applied to the different parts. If this is done, the adjustment will be unreliable.
- When the instrument is carried on the tripod, all clamps should be tightened to prevent unnecessary wear on the centers.

The Stadia

This is a device that is used for measuring distances, and it consists essentially of two extra parallel hairs in addition to the ordinary crosshairs of the transit or a level telescope (see Figure 6-12). The stadia hairs may be adjustable, or they may be fixed permanently on the diaphragm.

When using the stadia, distances are measured by observing through the telescope of a transit the space (on a graduated rod) included between two horizontal hairs (called *stadia hairs*). If the rod is held at different distances from the instrument, different intervals on the rod are included between the stadia hairs. The spaces

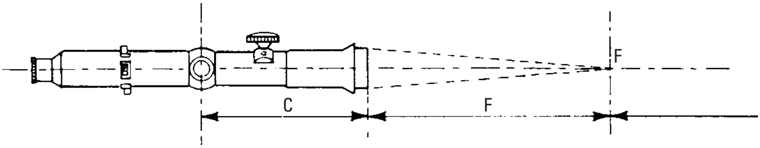

Figure 6-12 The principle of stadia operation. The fixed stadia hairs are set so that they will intercept 1 foot on a rod at a distance of 100 feet. Since the image of the crosshairs is projected to a point beyond the telescope objective equal to its focal length, the rays of light converge at that point, and measurements must begin from there. Therefore, a constant must be added to all stadia readings equal to the focal length of the object lens, plus the distance from the face of the objective to the center of the instrument. This constant is the factor F + C. For transit telescopes, it is equal to approximately 1 foot.

on the rod are proportional to the distances from the instrument to the rod so that the intercepted space is a measure of the distance to the rod. This method of measurement furnishes a rapid means of measuring distances when filling in details of topographic and hydrographic surveys.

Most transits, all plane-table alidades, and some precision leveling instruments are fitted with stadia hairs. Stadia surveying has the advantage in that the intervening country does not have to be taped, and it provides a means of measuring inaccessible distances (such as across water and up steep hills and bluffs). It is well adapted to preliminary surveys for highways and railroads because the errors tend to be compensating rather than cumulative, but it should not be used for short distances (such as farms and city lots). In sights of 200 to 400 feet, it is possible to read a rod to the nearest hundredth of a foot that represents 1 foot in distance. At 600 to 1200 feet, it is possible to read to the nearest hundredth of a yard that represents 3 feet in distance. This is the precision to be expected in stadia measurements.

The rod used is preferably a one-piece stadia rod (see Figure 6-13F), but any standard leveling rod (except builders rods graduated in inches and sixteenths) may be used. When leveling with an instrument that is equipped with stadia hairs, care should be taken not to confuse the center leveling crosshair with either of the two outside stadia hairs. It has been done. Although it is by no means obsolete, stadia surveying has been superseded by aerial photography. Neither is a substitute for careful taping.

Other Devices

For the vast majority of surveying purposes, the transit and levels described are more than sufficient. Three other devices can be even more useful, however.

The *Theodolite* (see Figure 6-14) is a transit that is more expensive but more accurate than a standard transit, and the Theodolite has more capacity. If you are doing readings over 500 feet, you may want to look into this.

The *automatic level* (see Figure 6-15) is also a tool used where great precision is required. Another device, the electronic distance measurer (EDM), is good when great distances must be read. It is a tool used on the sea and in other places (such as measuring from a point to a mountain) where there is no real point of reference. The electronic distance measurer and the Theodolite are often used together (see Figure 6-16). (By the way, no one seems to know the origin of the term Theodolite.)

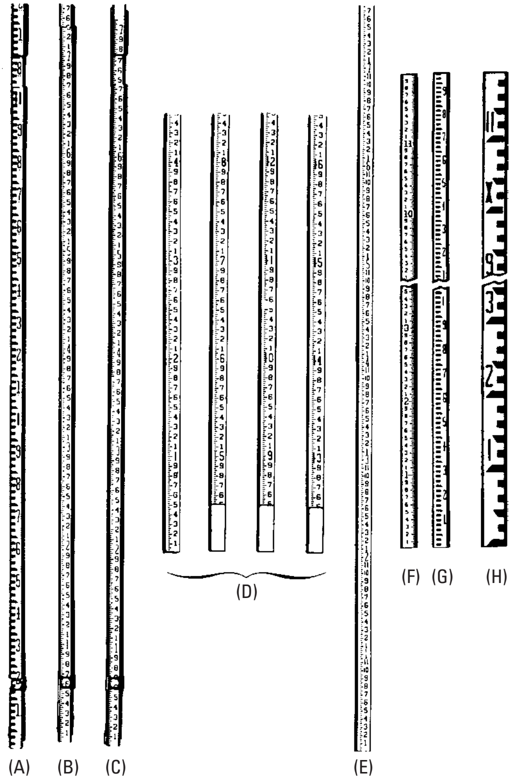

Figure 6-13 Various popular leveling rods: (A) is the Philadelphia rod, in English graduations; (B) the California rod; (C) the Philadelphia rod, in metric graduations; (D) the Chicago rod; (E) the architect's rod; (F) the stadia rod, in English graduations; (G) the stadia rod, in metric graduations; and (H) the broad stadia rod. *(Courtesy Eugene Dietzgen Co.)*

Laser Levels

The need for plumb walls and leveling moldings (as well as various other points straight and level) is paramount in house building. It is difficult in some locations to establish a reference point to check for level windows, doors, and roofs, as well as ceilings and steps.

The laser level (see Figure 6-17) has eliminated much of this trouble in house building. This simple, easy-to-use tool is accurate to within $^1/_8$ of an inch in 150 feet, and it has become less expensive recently so that even the do-it-yourselfers can rent or buy one.

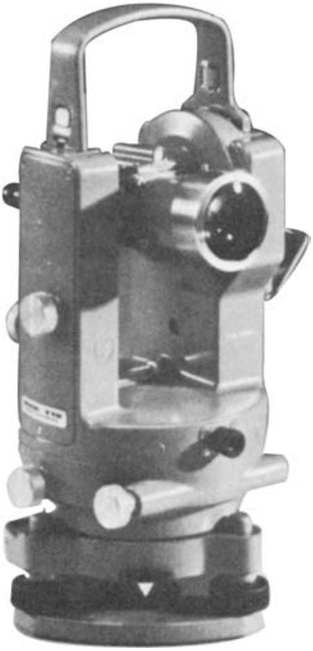

Figure 6-14 Wild T-16 Theodolite. *(Courtesy of Wild-Heerbrugg)*

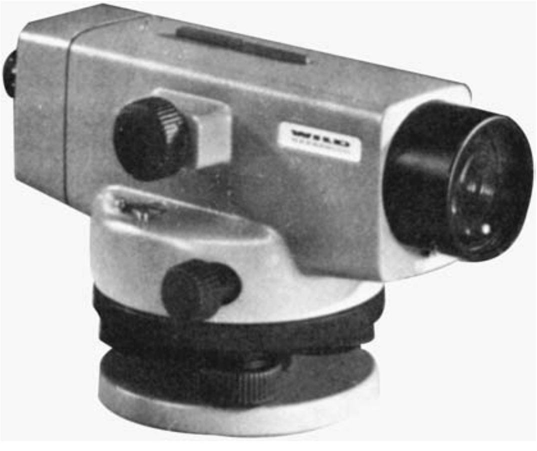

Figure 6-15 NA2 automatic level. *(Courtesy of Wild-Heerbrugg)*

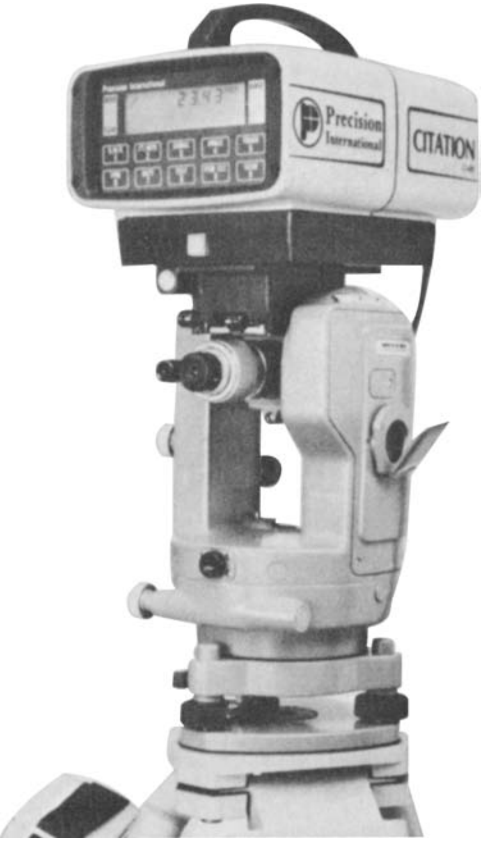

Figure 6-16 Typical EDM/Theodolite setup. *(Courtesy of Wild Heerbrugg)*

The laser level can generate a vertical reference plane for positioning a wall partition or for setting up forms (see Figure 6-18). It can produce accurate height gaging and alignment of ceilings, moldings, horizontal planes, and can accurately locate doorways, windows, and thresholds for precision framing and finishing (see Figure 6-19). The laser level can aid in leveling floors, both indoors and out. It can be used to check stairs, slopes, and drains. The laser beam is easy to use and accurate in locating markings for roof pitches, and it works well in hard-to-reach situations (see Figure 6-20). The laser beam is generated by two AAA alkaline batteries that will operate for up to 16 hours.

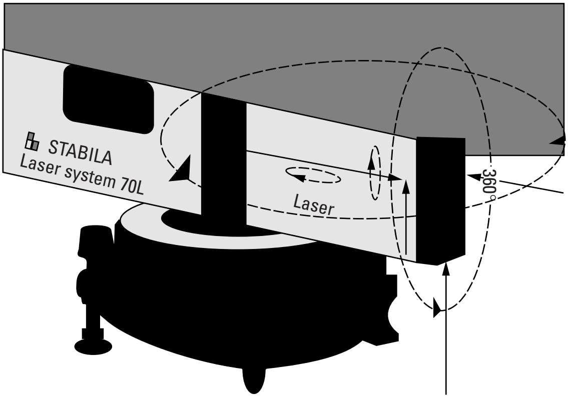

Figure 6-17 Laser spirit level moves 360° horizontally and 360° vertically with the optional lens attachment. It sets up quickly and simply with only two knobs to adjust. *(Courtesy of Stabila)*

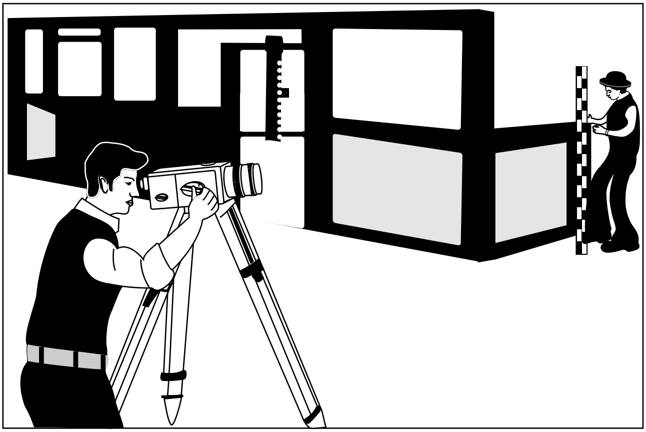

Figure 6-18 The laser level can be used to align ceilings, moldings, and horizontal planes. It produces accurate locations for doorways, windows, and thresholds for precision framing and finishing. *(Courtesy of Stabila)*

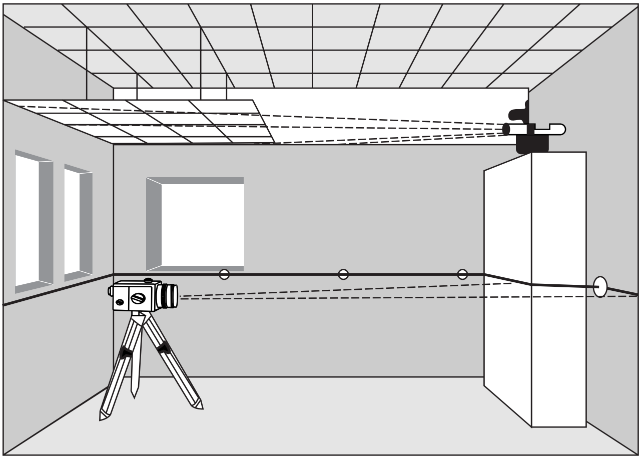

Figure 6-19 The laser level can be used for indoor or outdoor leveling of floors, stairs, slopes, drains, and ceilings, and moldings around the room. *(Courtesy of Stabila)*

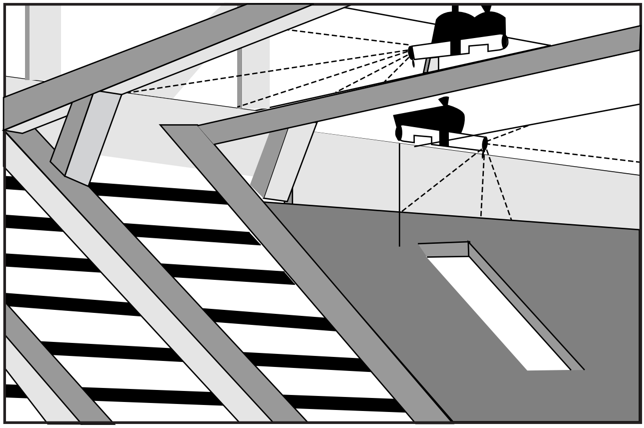

Figure 6-20 The laser beam is used to provide easy and accurate location markings on pitches and in hard-to-reach situations. *(Courtesy of Stabila)*

The combination laser and spirit level quickly and accurately lays out squares and measures plumb. No protective eyewear is needed. The laser operates on a wavelength of 635 nm and can have an extended range up to 250 feet.

Summary

Leveling, in surveying, is the operation of determining the comparative levels of different points of land for the purpose of laying out a grade or building site by sighting through a leveling instrument at one point to a leveling staff at another point.

The transit is designed and used for measuring both horizontal and vertical angles. It is a telescope mounted in standards attached to a horizontal plate, called the limb. The transit is generally provided with a compass so that the bearing of any given line with the magnetic meridian may be determined. It also has a spirit level attached to the telescope so that it may be brought to a horizontal position and made to serve as a level.

A stadia is a device used for measuring distance. It consists essentially of two extra parallel hairs in addition to the ordinary crosshairs that are used in transit or level telescopes. When using the stadia, distances are measured by observing through the telescope of a transit.

The laser level is rather inexpensive and easily used. It can be utilized in the squaring of steps, locating doors, windows, and many other squaring and leveling operations.

Review Questions

1. What are the three lines of the level?
2. What is a transit?
3. What are the fine lines of a transit?
4. Name the various leveling rods used for measuring distance.
5. Explain the terms *backsight* and *foresight* when used in leveling.
6. What is a stadia?
7. Where is the laser level most useful?
8. What is leveling in terms of surveying?
9. Why is it important to have square corners and level floors in a building?
10. How can a laser actually establish a useable level reference?

Chapter 7

The Design Process

The first consultation the designer has with the owner results in the owner setting forth what he or she wants or requires. This may include such things as the number of rooms, their sizes, the style of house preferred, the kind of materials he or she wants to use, and probably that all-important factor—how much money he or she will need, or how much is available. The first consultations may result in the designer making a dozen or more sketches.

It is the designer's function to guide the client in such matters. Usually, the owner has no skill or aptitude for building a house. Later consultations may settle such things as lighting, heating, perhaps the quality and style of plumbing fixtures, and the make or quality of such appurtenances as air-conditioning, laundry, and dishwashing equipment. The designer usually makes sketches (probably freehand) for the owner's approval.

In some instances, the purchaser comes into the office and consults with the designer or designee. The owner will decide on small changes in the chosen house plan. The designer just makes the suggested or agreed-upon changes and then draws up the blueprints. The owner picks out wall colors, kitchen cabinets, appliances, toilet-bowl designs, and many other small details. This may take a number of visits to a builder's office, where all the samples of tile, carpet, brick, and other design features are on display. If there is a model house available, it is much easier for most people to pick out the one they want and see what changes they desire.

Design Considerations

Following are some of the particulars that should be settled before final design work:

- Owners may greatly prefer a house with a basement. However, perhaps for reasons of economy, they may be willing to accept a house on a concrete slab. On the other hand, if the house is in Texas or some other location with soil problems, it will have to be a slab house. The designer should be able to give advice as to the advantages and disadvantages of the two systems along with the comparative cost of construction.
- The owner's mind may be set on a heating plant in the basement. However, in the event that a house with no basement is agreed on, the owner should understand that the heating plant must be in a utility room or in the attic.

- The owner may insist that there be a well-equipped laundry in the basement. The person who does the laundry may have some ideas on this subject.
- The owner will probably insist that there be adequate closets for each bedroom. The owner may have some ideas as to exactly what "adequate" consists of.
- The owner will want the kitchen to be convenient, with or without a garbage-disposal unit, probably with a dishwasher in a convenient location, possibly with room for a home freezer. Although saving steps is important, sufficient room in the kitchen may be more important. Few cooks appreciate a small kitchen—it greatly inhibits style.
- The owner may want a dining room, and not just dining space in one end of the living room, and may think that a counter or bar and no partition between the kitchen and dining room is not sufficiently odor-resisting when cooking fish or corned beef and cabbage.
- The owner (or perhaps the designer) may have some ideas concerning privacy in the home. There may be some objections to so-called window walls that are supposed to "bring the outdoors indoors," and a preference may be expressed for baseball-proof walls instead. Perhaps the idea of keeping large areas of glass clean is not appealing, and the owner would be quite content to leave the "outdoors" outdoors, if there is plenty of living space indoors.
- Then, there is the constant problem of sound resistance. Modern homes are often noisy, with air-conditioning, forced air heating, laundry, and dishwashing equipment, attic fans, kitchen fans, bathroom fans, radio and television, and many other noise-generating sources. It is the designer's duty and obligation to see that such noises are isolated insofar as is possible. Partitions should be noise-resistant.

The designer must understand these and many other problems. They are with us and will be with us for many years to come. The owner probably does not (possibly cannot) understand how to handle such problems. The designer can, and should. It is the designer's function to guide the owner's ideas or simple notions so that the home environment will be satisfactory as far as the owner's means will permit. The designer will probably be blamed for any serious discrepancy, no matter if the owner *did* insist on it.

It is not only desirable, but also necessary that the designer be familiar with the dimensions of the equipment, furniture, and other appurtenances found in the home. While it is rarely necessary that they be accurately detailed, space must be allotted for each one of them, and space is expensive.

All modern building is governed by codes of some sort. In all government-financed homes, the government's minimum standards are strictly enforced, and city codes are often much more restrictive. Electrical codes occasionally seem to be unreasonable, but the building designer must be governed by them. Plumbing is often seriously skimped when no one is watching. Many states have plumbing codes. However, they do not have the force of law unless augmented and enforced by local authority. There is a *National Plumbing Code*, and, although it is advisory, it is in line with good practice. It should be used where there are no local codes. State and local boards of health may make it quite difficult for the designer of an inadequate plumbing system if the occasion arises.

For economy, kitchens and bathrooms should be placed back-to-back. A 3-inch copper soil pipe will fit into a partition of 2 × 4 studs, whereas a 4-inch cast-iron soil pipe won't.

In perhaps most cases, the designer's duty is done when he or she prepares and delivers the drawings for a job. In some cases, he or she contracts to inspect the work at stated intervals, to ensure that the work is satisfactorily done.

The designer should allow for the following thicknesses of walls in drawings:

- *Standard wood outside wall*—3/4-inch plywood or insulating board sheathing, 3 1/2-inch studs, 1/2-inch sheetrock inside, 4 1/2 inches under the siding
- *Inside partitions*—3 1/2-inch studs, sheetrock 1/2 inch (both sides), 4 1/2 inches
- *Sound-resistant staggered-stud partitions*—3 1/2-inch studs staggered 2 inches, gypsum lath and plaster (both sides) 7 1/4 inches
- *Single-width brick veneer*—3/4-inch sheathing, 3 1/2-inch studs, 1/2-inch sheetrock, 9 1/2 inches
- *Concrete blocks (plastered against the masonry)*—8-inch, 8 1/2 inches
- *Concrete blocks (with furring)*—8-inch, 3/4-inch furring, 1/2-inch sheetrock, 9 1/4 inches

- *Cavity masonry wall*—3¾-inch brick, 2½-inch air space, 4-inch concrete blocks, plaster on masonry, 10⅞ inches
- *Chimneys*—minimum wall thickness, 3¾-inch brick liners, outside dimensions, 8½ inches × 8½ inches, 8½ inches × 13 inches, 8½ inches × 18 inches, 13 inches × 13 inches
- *Ceiling heights*—first floor, clear minimum, 7 feet 6 inches; basement, 6 feet 9 inches clear
- *Stair wells*—3 feet 2 inches × 9 feet, or clear headroom 7 feet above nosing of treads, vertically

An Example of Design

As an example of design, the series of figures presented in this section show the development of architectural drawings beginning with stock plans or plans that appear in newspapers and magazines from time to time. They give a prospective owner a good starting point, saving much time and study.

Certain things, however, should be kept in mind. The magazine drawings were doubtlessly prepared by a registered architect. While the publishing of a design may imply that the architect has given approval that the design can be copied, it is best to get the written consent of the original designer, or at least to secure this consent from the publisher. A registered architect's plans are protected from being copied, and court decisions have ruled that minor changes (regardless of how many of them there are) do not release the copier from liability. Purchasing a set of stock plans includes permission to build from the plans at least once. You might have to buy a second set to build the same design again. Determine where you stand legally before copying designs of any sort. It may save you from a costly and embarrassing situation.

Let us assume that we start with magazine plans (Figures 7-1 and 7-2). These two plans were prepared by a capable and experienced architect who has given us a practical and logical arrangement of a house plan, with stairs, doors, windows, closets, and so on. The architect has also indicated a proper and reasonable size for the various rooms. Any proposed changes or additions that the owner may desire may be taken up with the architect or builder, and may be easily whipped into form on the drawing board.

Conference with Builder

Let us now assume that the owner prefers to deal only with a carpenter or builder of his or her acquaintance, and asks the carpenter

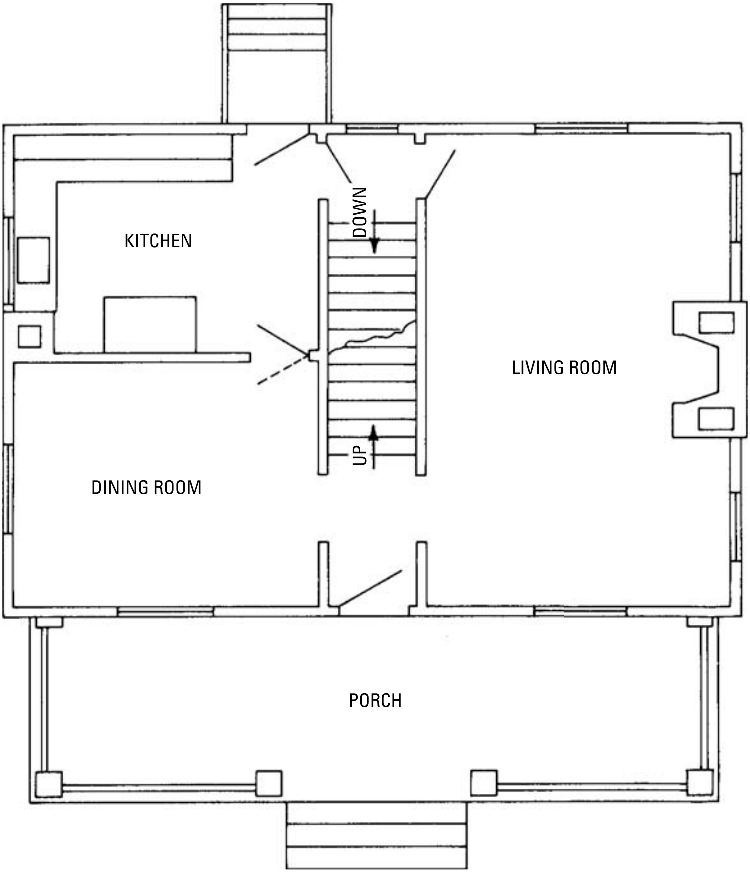

Figure 7-1 First-floor plan from a magazine.

or builder to prepare the drawings and give an estimate on construction. We will also assume that while the builder understands his or her own business from start to finish, he or she does not pretend to be a designer. A designer is, therefore, called in for advice on the preparation of the drawings. Having decided on the main features of the plans and elevations, we will start with the first-floor plan of the house, since that is always the controlling form factor. Then, following the easy and usual method, we may trace the other plans (second floor and basement) over it.

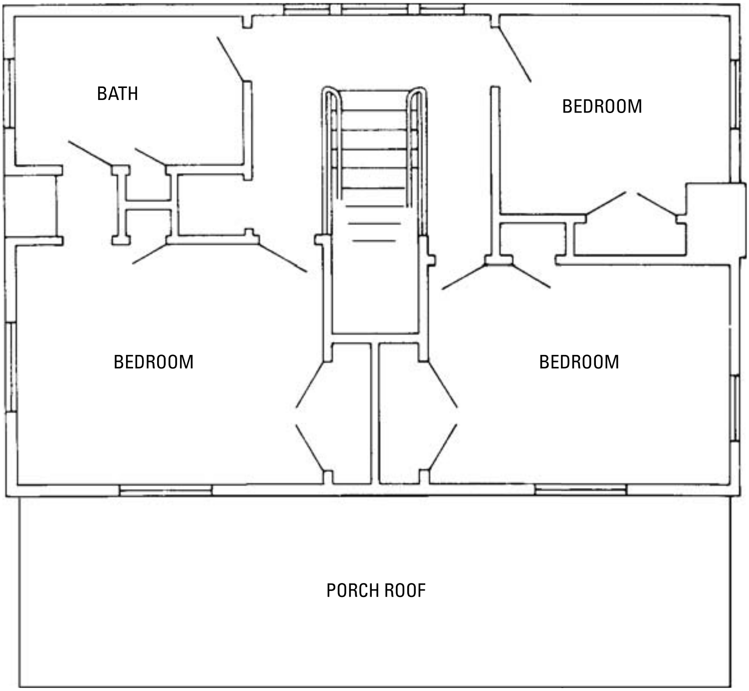

Figure 7-2 Second-floor plan from a magazine.

Changes Agreed On

After a careful study of the first-floor magazine plan, it is decided first to place the large chimney and fireplace inside the house against the stairs to conserve as much heat as possible, since considerable heat is necessarily lost from an outside wall. Next, omit the kitchen chimney, since a gas range is to be used. Third, move the large porch to the end of the house, since it provides more privacy, and change the window to a door. Fourth, provide the paved entrance platform with a settee at the front entrance. Fifth, provide a coat closet at the front entrance. Sixth, enlarge the dining room, taking the space from the kitchen. Seventh, install a first-floor lavatory in the small new wing at the rear. These, along with other minor changes, such as turning the back steps, adding two small closets near the fireplace, and substituting round for square porch columns, create a much more convenient and valuable first-floor plan. The principal second-floor changes are made to save cost and involve a lowering of the front eaves line and decreasing the size of the front rooms.

Figure 7-3 shows the modified first-floor plan after the changes have been agreed on, and Figure 7-4 illustrates the completed drawing of the first floor, with all the specifications and dimensions given in full detail.

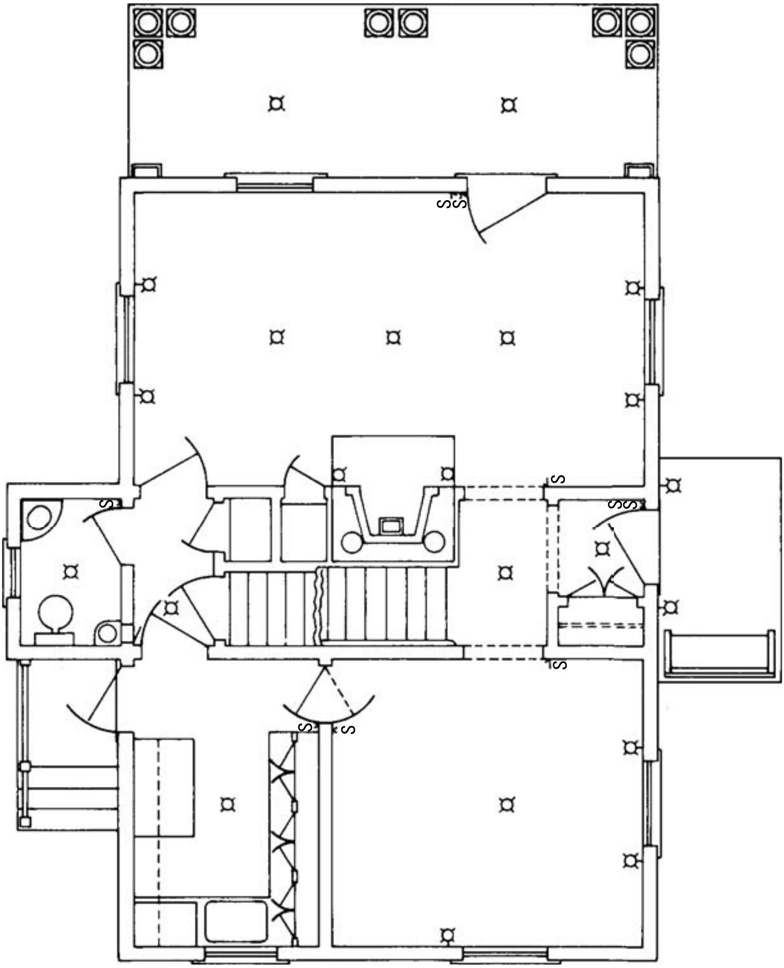

Figure 7-3 The modified first-floor plan.

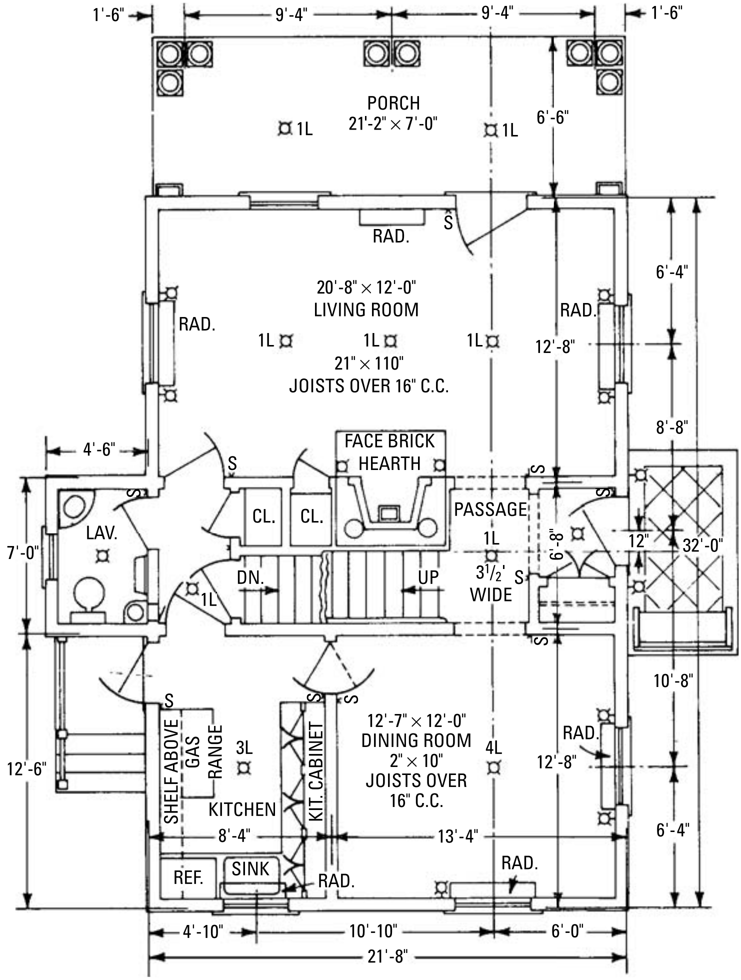
1'-6"
9'-4"
9'-4"
1'-6"
PORCH
21'-2" × 7'-0"
1L
1L
6'-6"
RAD.
S
6'-4"
20'-8" × 12'-0"
LIVING ROOM
RAD.
RAD.
1L
1L
1L
12'-8"
21" × 110"
JOISTS OVER 16" C.C.
8'-8"
4'-6"
FACE BRICK
HEARTH
S
S
S
S
CL.
CL.
PASSAGE
LAV.
7'-0"
1L
6'-8"
12"
32'-0"
DN.
UP
3½' S
WIDE
1L
S
10'-8"
S
S
S
12'-7" × 12'-0"
DINING ROOM
2" × 10"
JOISTS OVER
16" C.C.
RAD.
4L
12'-8"
SHELF ABOVE
GAS
RANGE
3L
KIT. CABINET
12'-6"
KITCHEN
8'-4"
13'-4"
6'-4"
RAD.
REF.
SINK
RAD.
4'-10"
10'-10"
6'-0"
21'-8"

Figure 7-4 The completed drawing of the first floor with all dimensions and specifications in detail.

Summary

In architectural design, the designer usually makes several sketches (probably freehand) for the owner's approval. Among the many particulars that should be settled before the designer actually begins work are: will there be a basement (which determines the location of the heating plant); how many closets in each bedroom; will there be kitchen conveniences (such as a garbage-disposer unit, dishwasher, and room for a home freezer).

These are just a few of the problems for the designer. The owner probably does not know how to understand and handle such problems. The designer should know how to handle these problems and guide the owner so that the home environment will be more than satisfactory when the work is completed.

It is, of course, necessary that a designer know the dimensions of the equipment, furniture, and other appurtenances found in a home. Electrical codes occasionally seem to be unreasonable, but the building designer must follow them. Plumbing is often seriously skimped, unless local authorities have some kind of a ruling enforcing certain regulations.

Many times the future owner prefers to deal only with a carpenter or builder, and asks him or her to prepare the drawings and give an estimate on construction. The builder then calls the designer in for advice on the preparation of the drawing.

Review Questions

1. What must a designer know from the client before drawing up plans for a house?
2. Is it generally a good idea to change floor plans after construction has started?
3. Is the kind of material used in the construction important to the designer?
4. Are wall thickness, plumbing, and electrical wiring considered when drawing up plans?
5. Why are some localities limited to slab houses?
6. What is the difference in price of a slab foundation and a basement?
7. Where is the air-conditioning and heating placed if the house is built on a slab?
8. What are some of the concerns of someone who is trying to have a house built?
9. What is the purpose of a plumbing code?
10. What is the purpose of an electrical code?
11. Why can't you use a copy of a blueprint found in a magazine to build a house?

Chapter 8

Building Specifications

By definition, a *specification* is a specific and complete statement detailing the nature and construction of the item to which it relates. As applied to the building trades, specifications describe briefly (yet exactly) each item in a list of features and materials required to complete a contract for building an entire project.

Great care should be used when reading specifications in order to avoid misunderstandings and disputes. Each item entering into the construction is defined and described with precision so that there can be no chance of misunderstanding or double interpretation.

Example of Specifications

Specifications refer to the contract form of which they are a part. This saves repetition of statements with regard to liability of contractor, owner, and so on. The following is an example of specifications.

Introduction

Specifications for a frame dwelling to be built for Mr./Ms. ________ of ____________________ in the county of ________________________ and state of ________________ on lot number ______________ on the ________ side of ________________________ street in the city or borough of ________________________, county of, ______, state of ________________________. These specifications relate to a set of attached drawings consisting of:

1. Plan of foundation and basement.
2. Plan of first floor.
3. Plan of second floor.
4. Plan of roof.
5. Four elevations (eight sheets in all).

Detailed working drawings furnished as required. All drawings are a part herewith, and are to be considered as such with any contract that may be made.

Height of Ceilings

The following dimensions for these heights are:

- Basement—7 feet 2 inches clear of finish.
- First-story—8 feet 6 inches between timbers.
- Second-story—7 feet 6 inches between timbers.

Interpretation of Drawings

For arrangement of floors, general finish, and measurements, reference must be made to the drawings. However, should any difference appear between the scale measurements and the figures, or between the wording of the specifications and the lettering on the drawings, the specification shall in all cases take precedence. If any errors that are not explained either by reference to the drawings or specifications become apparent, the contractor shall refer them to the architect for correction before proceeding with the work.

Conditions

The contractor must see that all the work on the said building is performed in a thorough, professional, and substantial manner by competent workers and must also furnish all materials (the best of their respective kinds), labor, implements, transportation, and so on, if not otherwise specified.

All painted parts of the exterior must have a prime coat of paint as fast as it is ready. The succeeding coats must not be applied within 3 days of the former, and then not in wet or freezing weather or other conditions specified by the paint manufacturer.

The contractor must protect all work while the building is in his or her hands, remove all superfluous materials or rubbish, and not obstruct the grounds around the foundation for grading and filling in as soon as the building is up. Figures are to take precedence over scale measurements.

Mason's Work

Excavate to the length, breadth, and depth required for the foundations, as shown on the architect's drawings. The topsoil is to be removed and placed in a separate pile from the other excavated materials—25 feet away from the excavation where directed. Also, excavate for a septic tank and overflow 75 feet from the foundation, as will be directed, containing 28 cubic yards to be built of concrete, with baffles, drains, and so on, all of which is set out in a special plan for same. The septic tank may be included in the plumbing subcontract. The main tank must be waterproof, although the overflow need not be. The tank shall be connected to the house at a point below the lowest fixture and below the frost line with a uniform declination of not less than 12 to 20 inches, and will have no running U trap. The drain is to be made of 6-inch socket-jointed Transite tile placed in tight cement joints from a point 4 feet outside the foundation.

Foundation

Foundations and footings, as shown on the plans, are to be made of poured concrete with 8-inch walls and 16-inch footings.

Chimneys

Build two chimneys, as shown, of the same size and shape. Use an approved hard-red-pressed brick for all exposed parts of the outside chimney and for topping out. The fireplaces in the parlor, dining room, and bedroom are to be faced with the same brick (smooth inside the fireplaces), with 8 inches on the sides and 24 inches at the top. Buff the joints, and straighten the arches on $3^1/_2$-inch × $3^1/_2$-inch angle irons, unexposed. Use firebrick for the backs. Lay the bricks in an approved refractory mortar (no fire clay). Spring trimmer arches for the hearths are to be laid with the same selected brick. All flues are to have tile linings, approved chimney pots, and clean outs.

Mortar

All mortar for brickwork is to be grade specified by building code regulations.

Installing Drywall

All walls, partitions, and ceilings, and all studded and furred places in all stories, are to be covered with $^1/_2$-inch-thick Sheetrock, which will have joints covered with tape and joint compound. Two coats of compound will be applied where wallpaper is used; three coats where paint is to be used. Panels shall be 4 feet × 8 feet and will be installed horizontally on walls, across the framing members on ceilings. The panels are to be secured with $1^3/_8$-inch blued ring-shank nails, and the pieces staggered. Inside corners of the paneling shall be covered with inside-corner molding, with nails placed 7 inches apart. Outside corners shall also be covered with metal molding. All joints will be smooth to the touch.

Tiling

The floors of bathrooms will be tiled with 3-inch × 3-inch octagonal and 1-inch-square vitrified tiling, colors to be selected. The side walls will be tiled 4 feet high of plain white glazed $2^1/_2$-inch × 4-inch molded base and nosing, with a narrow tinted stripe at the top of the sanitary base and under the nosing. The floors will be properly prepared by the carpenter by setting the rough floor $^1/_2$ inch below the top of the floor beams. All tiling will be set in adhesive recommended by the manufacturer, and the floors will be finished flush with the wood-finish floors.

Other Floors

There will be a concrete floor in the furnace room, in the shop, and in the area from the west end turning east to the cross wall, as shown in the plans. All floors will be $3^5/_8$ inches minimum, with 6-inch × 6-inch No. 11 reinforcing mesh. The kitchen hearth will be built in the same manner.

Coping

There will be 4-inch caps of blue stone on all piers showing them, edged on four sides, 3 inches larger than the piers. Cope area walls, which are to be 8 inches, with 2-inch × 10-inch blue stone where circular, fitted to radius. *No* patching of stone will be permitted.

Timber

All timber will be thoroughly seasoned, No. 1 common pine, square, straight, and free from any imperfection that will impair its durability or strength. *No* individual piece is to have moisture content of more than 19 percent. The architect will check this.

Framing

The framing will be as indicated on the drawings. Headers over openings will be the sizes indicated in detail. *No* header with checked moisture content of more than 15 percent will be acceptable. Frame so that sheathing will be flush with the foundation wall. All moldings are to be miter-spliced and mitered at angles. *No* butt ends will be showing in the finish.

Timber sizes will be as follows:

Design Element	***Timber Size***
Sills	2-inches × 8-inches
Girders	10-inch standard I-beams (25.4 pounds)
Corners	4 inches × 6 inches backed with 2 inches × 4 inches or built up
Main plate	4 inches × 4 inches (2 inches × 4 inches doubled)
Rafter plate	4 inches × 4 inches (2 inches × 4 inches doubled)
Studding (general)	2 inches × 4 inches
Closet studding	2 inches × 3 inches
Main rafters	2 inches × 6 inches
Dormer rafters	2 inches × 4 inches
Ridge boards	$1^1/_4$ inches × 8 inches
First floor joists	2 inches × 10 inches
Second floor joists	2 inches × 8 inches
Second story ceiling beams	2 inches × 6 inches

Spacing and Bridging

All studding, floor, and ceiling joists placed on 16-inch centers. In every span of flooring exceeding 10 feet, there will be a row of 1-inch × 2-inch bridging or 2-inch × 3-inch double nailed at each end. Rafters will be placed on 24-inch centers.

Partitions

All partitions are to be set plumb, well braced, and nailed. Studs at all angles and openings are to be doubled, and extra block is to be set at door openings for base nailing. All partitions that are not supported below are to be firmly trussed and braced. Ceilings to all closets will be furred down to within 12 inches of the door head except in closets more than 2 feet deep. There will be trued $^7/_8$-inch grounds at top of base and around all openings.

Lumber

All outside finish-lumber will be clear white pine unless otherwise specified. All exterior finish lumber is to be free from large or loose knots and will also be clear and thoroughly dry.

Sheathing and Sheathing Paper

Cover all the exterior walls with $^3/_4$-inch plywood sheathing nailed to each stud with 8d nails. With joints cut on studs or backed for end nailing, cover with Tyvek, which is to be well lapped, extending under all trim and around all corners to make a complete and tight job.

Exterior Finish

Windows, door casings, cornices, corner boards, water table, brackets, band courses, and so on are to be made to the detail furnished in the drawings. The stock moldings that are to be used are numbered on the drawings. The first story is to be covered with the best-grade cedar lap bevel siding, laid at $4^1/_2$ inches to the weather. The second story and gables are to be covered with 18-inch hand-split and resawn shakes, laid at $8^1/_2$ inches to the weather. Use hot-dipped galvanized nails, whose length will be approved by the architect. Window casings will be laid $2^1/_2$ inches to the weather, and the front door frames and casement windows will be according to detail shown in the plan.

Shingling

Cover all roofs with 8-inch × 16-inch Pennsylvania blue slate, laid 7 inches to the weather. All hips and other parts that require it are to be made secure against leaks by the proper use of slaters' cement and proper flashings. An ornamental galvanized-iron ridge crest will be placed on the main ridge. See details on drawings for this crest.

Flashing

Flash around chimneys, over all doors and windows, heads exposed to the weather, and where roofs join walls with 16-ounce sheet copper. Do the same in all valleys and wherever required to secure a tight job. Each side of a valley is to have a water check turned up 1 inch in the metal.

Flooring

First and second stories are to have double floors, with a subfloor of $^5/_8$-inch CDX plywood. The first- and second-story finish flooring is to be oak tongue-and-groove strip flooring, which is to be thoroughly seasoned and blind-nailed over building paper. There will be no joints in the main hall and only one joint in the run of boards in other rooms of the first floor. The second-story floors are to be cleaned and sandpapered to a smooth finish for the painter. Oak thresholds are to be set to all outside doorways, and hard rubber-tip doorstops are to be located behind all doors that open against a wall.

Window Frames

These are to be made of seasoned white pine.

Sash

All sash and frames are to be made by the Johnson Corporation and are to be of kiln-dried, vinyl-sheathed white pine. The numbers are given on the drawings.

Screens

All windows that open are to be fitted with bronze- or copper-wire window screens.

Glazing

All sash and outside doors, where indicated, are to be glazed with Johnson insulated windows or their equivalent. All hall doors are to be glazed with French plate. The plate in the Dutch door will be beveled. The basement sash is to be glazed with a single-strength glass.

Blinds

All windows, where indicated, are to be provided with an approved type of blind that will be $1^1/_8$ inches thick and made of the best grade of seasoned white pine. All blinds will move freely after painting. The blinds are to be hung on approved cast-iron blind hangers.

Door Frames

All inside door frames in finished parts of the house, first and second stories, are to be made of white pine $^{25}/_{32}$-inch thick, set plumb and true, and blocked in four places on each side. Outside doorframes are to be rabbeted for doors. All frames are to be flush with the plaster finish.

Doors

Unless otherwise specified, all inside doors are to be made of slab-type birch veneer with hollow cores and will be $1^{3}/_{8}$ inches thick. Outside doors are to be $1^{3}/_{4}$-inches thick and will be made of solid-core, slab-type birch veneer. The front doors are to be of the design shown in the drawings. Hang all doors throughout with loose-joint ball-tip butts of sufficient size to throw them clear of the architraves. Doors are to have three $3^{1}/_{2}$-inch × $3^{1}/_{2}$-inch butts on $1^{3}/_{8}$-inch doors and three 4-inch × 4-inch butts on $1^{3}/_{4}$-inch doors. A hardware schedule will be furnished. Hang both double-swing butlery doors on double-acting brass spring hinges. Furnish all nails, except those used for inside work, galvanized and all other hardware that will be necessary for the completion of the work in the proper manner.

Interior Trim

For the basement, the interior trim is to be selected cypress or redwood. For the first and second floors, the trim is to be unselected birch. There will be a 4-foot 6-inch paneled wainscoting in the dining room, first floor hall, and up stairway. This panel will be made of $^{1}/_{4}$-inch birch plywood, with trim as shown in the details. There will be a 5-inch cabinet plate shelf in the dining room, the bottom member of which will be a picture molding. This shelf will match the door head trim.

Stairs

The main staircase is to be made of unselected birch. The stringers and treads are to be $1^{1}/_{8}$ inches thick, as shown in details. The risers are to be $^{3}/_{4}$ inch thick. The risers and treads are to be housed into the wall stringer and return-nosed over the outside string. The rails are to be 3-inch × 3-inch molded, with ramps as shown in the details. Balusters will be $1^{5}/_{8}$ inches, taper turned, three to a thread, and proportionately more for increased widths. Newels and column newels are to be as shown in the details. The run on the first flight is to be $10^{1}/_{4}$ inches from face to face of the rise with 12-inch treads. The second flight and basement stairs are to be $9^{3}/_{4}$ inches of run with $11^{1}/_{2}$-inch treads. For the basement stairs, cut 2-inch × 12-inch boards for the stringers and 2-inch × 10-inch yellow pine for the treads.

Mantels

There will be two mantels where indicated on the drawings. The contractor will figure them to cost $2000 each complete, including linings and face and hearth tile. This amount will be allowed the owner to use at his or her option in the selection of it. The entire cost is to be figured in the contract price, including the setting of the mantels by the contractor.

Pantry Cabinets

There will be a cabinet, where indicated, with three glass doors, above the draining board. This cabinet will be 10 inches deep inside and will contain three shelves. The wall cabinets are to be constructed as shown in the details.

Closet Shelving

The trim on the inside of the cabinets is to be plain. There is to be an average of 10 feet of 12-inch shelving to a closet, with 6-inch clothes strips and 1 dozen clothes hooks, japanned. The kitchen closet and the closet under the kitchen stairs are to have suitable shelving and sufficient pothooks and other fixtures. There will be 25 feet of shelving in the shop closet. There will also be 1 dozen clothes hooks under the basement front stairs.

Plumbing

All necessary materials for completing the plumbing installation, as hereafter set forth, in a correct and sanitary manner are to be included in the general contract. The state plumbing code shall be strictly followed.

Electric Wiring

No. 12 Romex sheathed cable is to be installed under and subject to the requirements and regulations of the *National Electrical Code* and all state, county, and municipal codes. The locations for all electrical outlets will be shown in the drawings.

Water Pipes

Water is to be brought from the street main into the house through $^3/_4$-inch copper tubing, or plastic if preferred. Copper water tubing is to be used on all straight-line work. Place a hose-bib cock on the main at a point against the house for hose purposes, both front and rear, with a stop and waste cock in the basement.

Complete all necessary digging for the laying of sewer and water pipes to the house. No trenches are to be less than 36 inches below the grade at any point. The pipes are to enter the house in the basement at a suitable point for intersection with the inside piping system. The dirt is to be well rammed over the pipe in the trenches as it is refilled. The house sewer is to be installed by the plumber, and it will consist of a perfect 4-inch glazed socket-jointed tile pipe to a point exactly 4 feet outside the foundation wall. The soil pipe is to be 4-inch cast iron. All water pipes must have a gradual fall from the fixtures that they supply, and they must open at their lowest point for drainage purposes. All the cast-iron waste pipe will be furnished with the necessary fittings.

Furnish and install one 60-gallon gas-fired water heater of an approved type and manufacture. This water heater is to be supplied with water through a $^3/_4$-inch copper tube. Place a shutoff cock on the supply pipe. Take hot water from the water heater to and over the kitchen sink and to all other fixtures, except toilets, with $^1/_2$-inch copper tubing. The supply to the toilets will be through $^3/_8$-inch copper tubing. Cold water to all fixtures will be through separate pipelines. There must be no depressions in any pipe, and hot water must be kept rising from the boiler head.

Kitchen Sink

Furnish and install a cast-iron enameled kitchen sink, with a garbage disposer of (*specify make and model*). Furnish and install an automatic dishwasher (*specify make and model*) where indicated in drawings.

Wash Trays

Provide and install where indicated on drawings, one two-part stone tub on galvanized iron legs. Supply the tub with hot and cold water through $^1/_2$-inch copper tube and brass faucets, one for cold water, threaded $1^1/_2$ inches waste, with traps, plugs, and chains complete. The waste drain is to be connected with the soil pipe through a 2-inch copper pipe.

Bathroom Fixtures

All bathroom fixtures are to be (*specify name of manufacturer*) make, as listed in their catalog. The basement-bathroom water closet will be (*name and model*); lavoratory, (*name and model*); bathtub, 5-foot 6-inch (*name and model*). The other toilets are to be (*name and model*). The two second-story baths are to be (*name and model*), 18 inches × 27 inches. Before any wall finishing is done, all supply pipes must be proven perfectly tight by a satisfactory test, and they must be left perfect at the completion of the test.

Painting

The entire exterior woodwork, except shingles, is to be painted with three coats of (*manufacturer's name*) best grade of ready-mixed paints, thinned as necessary and as specified by the manufacturer.

The color of paint will be plain white, except for the shutters and shakes, which will be green and will be given two coats of (*manufacturer's name*) shake and shingle finish. The second-story floors will be sanded smooth and given three coats of (*manufacturer's name*) polyurethane varnish. The first-story floors are to be left bare for wall-to-wall carpeting or vinyl tiling. All interior birch trim is to be finished natural, with one coat of white shellac and two coats of (*manufacturer's name*) of flat varnish. The type and color of the interior wall paint for all interior walls will be specified by the owner.

Condition of Bids

The owner reserves the right to accept or reject any or all bids.

Summary

A specification is a statement containing a detailed description or enumeration of particulars, as of the terms of a contract, and details of construction usually not shown in an architectural drawing. Great care should be used when reading specifications to avoid misunderstandings and disputes. Each item entering into the construction is defined and described with such precision that there can be no chance of misunderstanding or double interpretation.

Framing will be indicated on the drawings. Specifications as to type of timber used (such as No. 1 common yellow pine, square, straight, and free from any imperfections that will impair its longevity) will be indicated. Headers over window and door openings will be indicated as for size and installation procedure.

Interior trim, including stair casing, will be shown. Moldings used throughout the house will indicate miter angles and type of lumber used. Ceiling height (as well as plastered or drywall construction) will be specified. Most plastered walls will be specified as to type of finish coat (which, in most cases, is at the option of the owner) and will also specify that exposed corners be protected with metal corner beads.

All necessary materials for completing the plumbing installation, as set forth in the specifications, shall meet all state and local regulations. Electrical wiring will be listed and meet the requirements and regulations of the *National Electrical Code*, or state and local codes.

Review Questions

1. Why should great care be used when reading specifications?
2. What information is included in the specifications on framing a house?

3. What code should be followed when plumbing a house?
4. Can a specification on a dwelling be changed or altered? If so, how?
5. What code should be followed when electric wiring is installed?
6. Why should the purchaser of a house not only read, but also study all specifications prepared by the builder?
7. What should the clearance of the ceiling be in the basement?
8. What is an elevation? Where is the word found?
9. What is Tyvek? Why would you want it on a house?
10. What is the height of the second-story ceiling (between timbers)? Is this higher or lower than that of the first story?
11. What does the architect place in the specifications sheet?
12. Who determines the number of bathrooms in a house?

Chapter 9

Architectural Drawings

Drawings provide designers with a practical method of communicating their ideas to the carpenters and builders. Drawings are a type of shorthand that not only reflects, but also fixes on paper, the ideas of the designers. They help avoid possible misinterpretation.

Use of Drawings

Contractors and estimators should retain prints on jobs that they complete so that they might be used to compare future jobs. They should also keep at least two sets of prints on a job in progress. One set is the working-drawings for the tradespeople. On the second set, the lead carpenter adds the changes made as the structure is built, to be turned over to the owner to update the original drawings issued. The final drawings will identify any changes that were made during construction. The owner may change the original drawings to reflect the changes, so that the drawings will be fully up-to-date to assist with maintenance, troubleshooting, and design of additions to the original building.

Reading Drawings

In the drawings (sometimes called *prints* or *plans*), various elevations are shown (such as the front, side, and sectional). A plan is a horizontal view of an object. An *elevation* is a vertical view of an object.

Projected Views

The various views are projected on imaginary projection planes, similar to the projection of a picture on glass. To illustrate the first (or front) view, place a clear pane of glass in front of the object with the glass parallel to the surface of the object being projected. Figure 9-1 shows a simple building with a shed roof. In front of the building is the pane of glass marked *V*, representing a vertical plane.

When an observer looks through the glass directly at the front of the object from a considerable distance, the observer will see only one side, in this case the side marked *ABCD*. The rays of light falling upon the object are reflected into the eyes of the observer, and in this manner, the observer *sees* the object. The pane of glass (vertical plane) is placed so that the rays of light from the object will pass through the glass in *straight parallel lines* to the eyes of the observer.

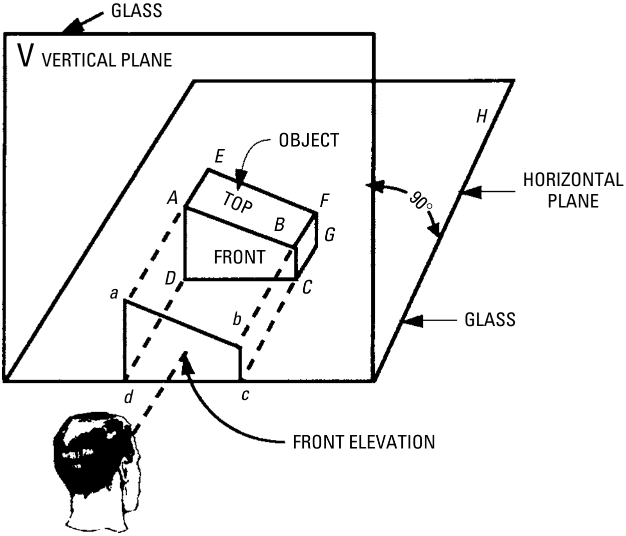

Figure 9-1 Rectangular object resting on a horizontal plane and facing a vertical plane.

The rays of light from points *ABCD* of the building pass through the glass at points *a*, *b*, *c*, and *d*. If these points (*a*, *b*, *c*, and *d*) are connected by lines, a view of the object as seen from the front is obtained, which is called *front elevation*.

The front elevation is identical in shape and size with the front side *ABCD* of the object (that is, *ab* = *AB*; *bc* = *BC*, and so on; angle *dab* = angle *DAB*; angle *abc* = angle *ABC*, and so on).

Top View or Plan

For this view, place a pane of glass in a horizontal position above the building that is resting on the horizontal plane (see Figure 9-2). Now, look at the object directly from above. Note that the rays of light from corners *AEFB* of the top pass through the glass at points *aefb*. If these points *aefb* are connected by lines, a view of the object as seen from the top is obtained, which is called the *top view*, or preferably *plan*.

Right-End View (Elevation)

A pane of glass is placed to the right of the building in a vertical position and parallel to the right side *BFGC* of the building (see Figure 9-3). Here, the pane of glass is marked *P*, which means *profile plane*, or plane from a side projection. Looking at the building directly from

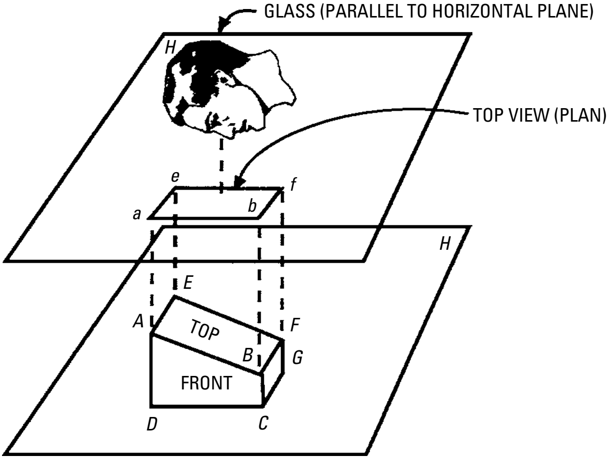
GLASS (PARALLEL TO HORIZONTAL PLANE)
H
TOP VIEW (PLAN)
e
f
a
b
H
E
A
TOP
F
B
G
FRONT
D
C

Figure 9-2 Projection of the top of an object to obtain top view.

the right side (as in position *S*), the rays of light from corners *BFGC* of the upper left-hand side (from points *AE*) pass through the glass at points *bfgc* and *ae*. If these points are connected by lines, a view of the object as seen from the right side is obtained, which is called the *right-side view*, or preferably *right-end elevation*.

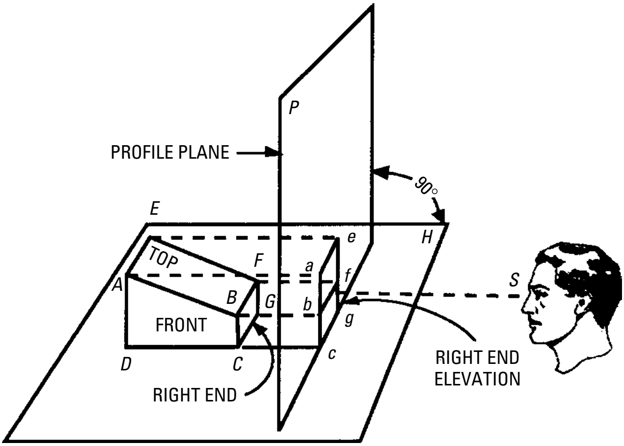
P
PROFILE PLANE
90°
E
e
H
TOP
F
a
A
f
S
B
G
b
FRONT
g
D
C
c
RIGHT END
ELEVATION
RIGHT END

Figure 9-3 Projection of the right end of an object illustrating right profile plane.

The shape of the object is such that the entire visible surface does not lie in a plane parallel to the projection plane. The points *A* and *E*, though located at the other end of the object, are visible and accordingly form part of the right-end view. Figure *aefb* does not show the top in its true size because it is projected obliquely instead of at 90°. An *oblique projection* makes an object appear smaller than its real size.

Left-End View (Elevation)

With a pane of glass shifted to the left side of the object (see Figure 9-4), and the building viewed directly from the *left* side (as position *S*), the rays of light from corners *ADHE* of the left side pass through the glass at points *adhe*. If lines connecting these points are drawn on the glass, a left side view of the object is obtained. However, the edge *FB* at the other end is invisible. It is shown by a dotted line connecting *f* and *b* projected from *F* and *B*. The completed drawing is then called a *left-side view*, or preferably a *left-end elevation*.

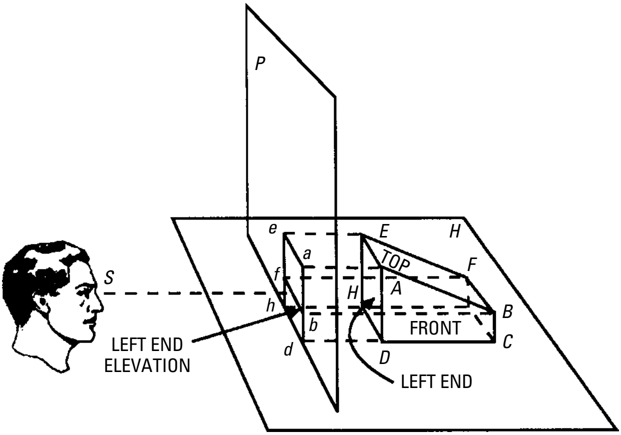

Figure 9-4 Projection of the left end of an object illustrating left profile plane.

Sections

Most buildings are so complex they cannot be clearly represented by a plan and elevation alone. In such a case, the parts that do not appear properly in these drawings are better represented by a *section*, or *sectional views*. A cross-section is a drawing of a building showing that part cut by a plane (see Figure 9-5).

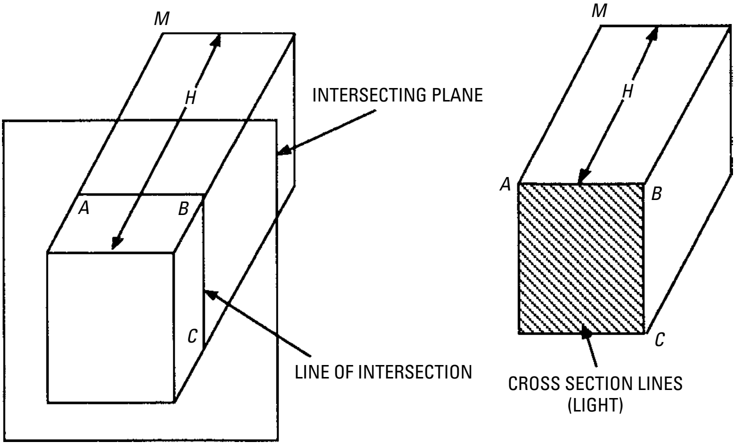
M
H
INTERSECTING PLANE
A
B
C
LINE OF INTERSECTION
M
H
A
B
C
CROSS SECTION LINES
(LIGHT)

Figure 9-5 Pictorial view of a building illustrating cross-section.

Directions of View for Sectional Views

For an unsymmetrical object, it is important to know the direction in which the sectional view is viewed. This is indicated by arrows at the end of the line representing the cutting plane (see Figure 9-6).

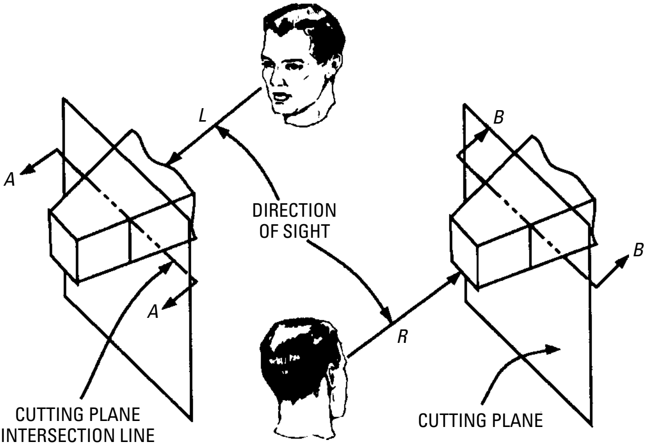
L
A
DIRECTION
OF SIGHT
B
B
A
R
CUTTING PLANE
INTERSECTION LINE
CUTTING PLANE

Figure 9-6 Direction in that a sectional view is viewed is indicated by the direction of arrows *AA*, at the ends of the cutting plane intersection line.

The arrows *AA* indicate that the object is viewed in the direction of point *L* (toward the smaller end of the object), and the arrows *BB* (in the direction of *R*) toward the larger end.

The Scale

Scale is the ratio between the actual size of the object and the size that it will be drawn. The scale is usually expressed on a drawing as full size, half size, quarter size, and so on, or it might be expressed as 1 inch = 1 foot, 1 inch = 100 feet, 1 inch = 1000 feet, or any other proportion that might be necessary to use. The scale is printed on the drawing.

On a full-size drawing, the object and drawing are of the same size. When the drawing is marked half size, the object is twice the size of the drawing. Thus, the drawing of an object is shown full size, half size, and quarter size (see Figure 9-7). If the building's height is represented by H and its diameter by D, then these dimensions will be the same for the full-size drawing. That is, $H = h$; $D = d$. For the half-size drawing, $h' = 1/2\ H$; $d = 1/2\ D$. Similarly, for the quarter-size drawing, $h'' = 1/4\ H$; $d'' = 1/4\ D$.

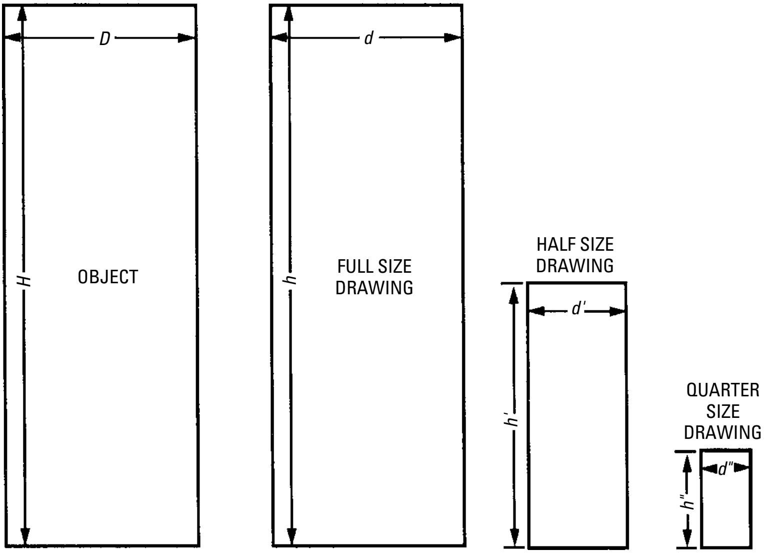

Figure 9-7 A drawing of an object to different scales: full size, half size, and quarter size.

From this it is seen that when the length of any edge on the drawing is made the same as the length of the corresponding edge on the object, the drawing is marked *full size* (sometimes *actual size*). If the length of any line on the drawing is half the length of the corresponding line on the object, the drawing is *half size*.

The scale is smaller than the building. In using prints, the size is important, since prints are used in the field, and if they are too large, they may not be easily handled.

In the case of a building, it would be impossible to have a print as large as the building. Thus, it is necessary to cut the print down in size. This necessitates the use of a scale. In the drawing of a building the building, designers usually express scale as 1 inch = 1 foot, 1/2 inch = 1 foot, and so on. This would indicate that 1 inch on the drawing would be equal to 1 foot on the actual structure, or 1/2 inch on the drawing would equal 1 foot on the actual structure, respectively. The architect uses the architect's scale. This is laid out in inches (1/8, 1/4, 1/2, and so on).

To lay off a distance of 2 feet 6 inches, place the 3/4 scale with division 2 at the given point *A*; then the zero division on the scale will be at a distance of 2 feet (see Figure 9-8). Because the end space is divided into twelfths, each division represents one inch on the 3/4 scale. Therefore, measuring off six divisions indicates that *AB* = 2 feet 6 inches. Notice the difference in actual length of this measurement on the 1 inch = 1 foot scale (see Figure 9-9).

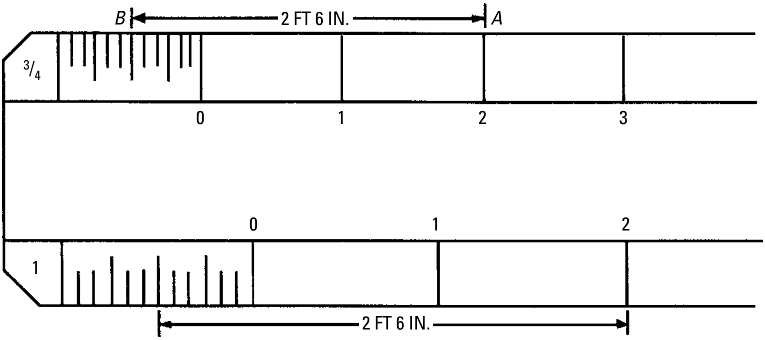

Figure 9-8 A typical architect's scale with detail showing 3/4 and 1 inch to the foot.

The scale on the original would not apply to a reproduction made by photocopying. The scale on the original (see Figure 9-10) should

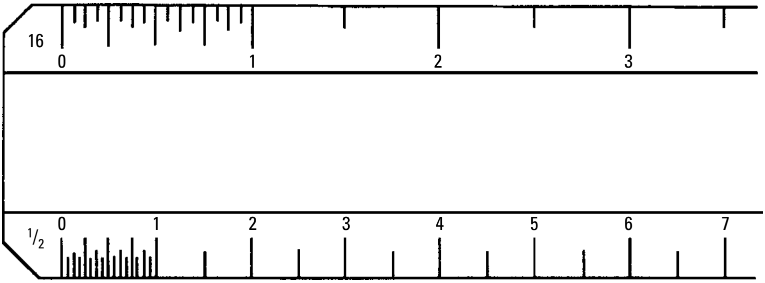

Figure 9-9 A typical architect's scale with detail with inch divisions instead of inches to foot.

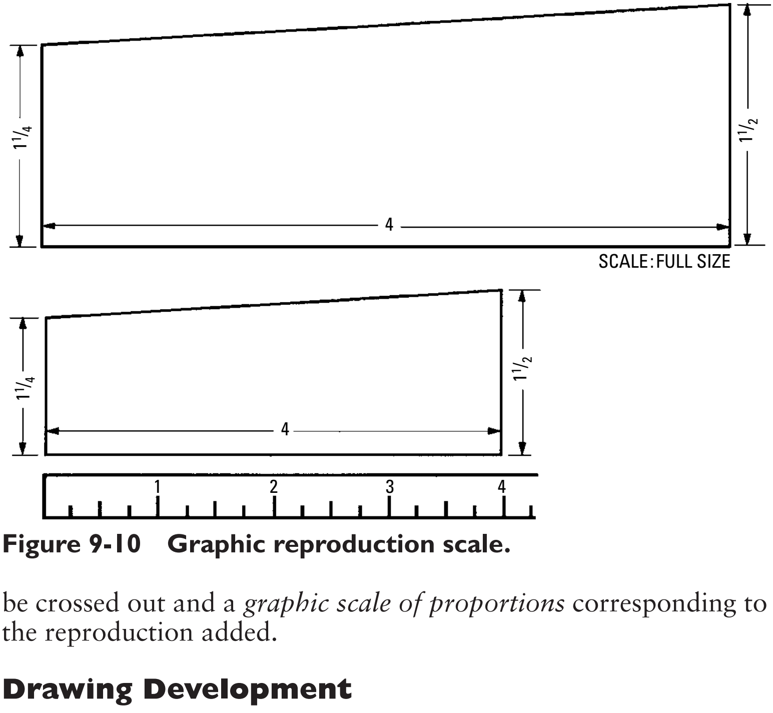

Figure 9-10 Graphic reproduction scale.

be crossed out and a *graphic scale of proportions* corresponding to the reproduction added.

Drawing Development

Knowledge of how architectural drawings are developed will help the carpenter and builder read drawings with more complete understanding. Architectural drawings are a means of transferring the thoughts of the designer to the builders and carpenters whose responsibility it is to construct the building. Graphic symbols are used to locate specific features and where they are to be placed.

The different parts or drawings that are necessary to show the structure (such as the mechanical and electrical installations) are all shown graphically. Therefore, you must become familiar with these symbols, not only those of one particular trade, but those of all the trades. This is necessary so that complete coordination may be reached between the various trades. In the construction of a building, time and money may be saved by representatives of the different trades going over the plans and laying out the pattern to be followed. No trade can work independently of the others. If independence is attempted, confusion is created and some work must be done over to make all parts of the scheme fit together. During construction, the general contractor, the plumber, the steel workers, and the electrical and mechanical contractors must lay out the work together and determine from the drawings who installs what, where, and when.

On the job, the architect may have a representative present to assist in coordinating the work and in making decisions that may be required. In designing the building, the owner or builder will often draw a rough sketch, after which he or she sits down with the architect to discuss what the owner will need and want in a building design. The requirements are noted as to space, machinery, electrical loads, numbers of persons that will occupy the building and what the future requirements might be. The owner will sometimes have a rough sketch of ideas (see Figures 9-11 and 9-12). These need not be drawn to scale or with any degree of accuracy. They are merely ideas of what the owner might want. In this discussion, no elaborate plans are given. Simple plans are used, as they show the intent and fall into line with more detailed plans.

Notice that there are no details shown, merely a sketch of the spaces to be enclosed. After the sketch is drawn, the architect and owner can sit down and discuss details, at which time, no doubt, another freehand sketch will be drawn with more details. When a tentative solution is reached, the architect will make a preliminary drawing (see Figure 9-13).

When the owner has signed construction papers, the architect will start drawing up the final plans and all details. There are many preliminary things to do, such as surveying the land to see how much excavating will be required. The location of the property lines and the general drainage plan for the immediate vicinity must be considered, and the water, sewer, gas, telephone, and power lines that exist must be considered. Local regulations regarding types of construction permitted, setbacks, and so on, all must be taken into consideration.

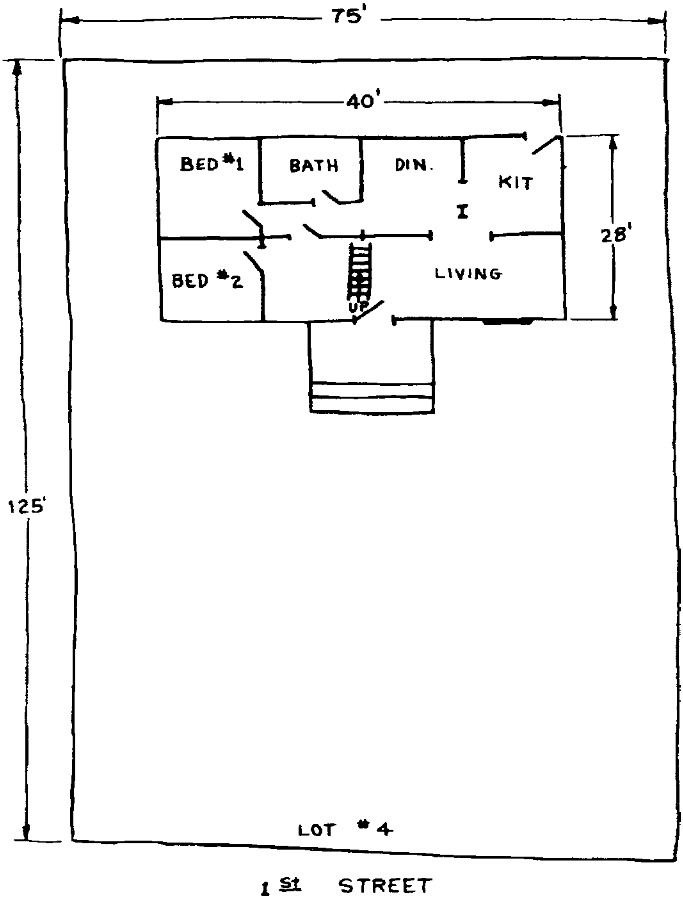

Figure 9-11 Owner's sketch of the first floor of a residence.

The final drawings are sent to the plan checkers of the inspection departments having jurisdiction. Here, they are checked to see that they conform to local codes. Corrections are noted, or the plans are approved. Most specifications that accompany plans put the burden of following applicable local codes on the builders and contractors. When questions arise (such as an electrical or mechanical problem), the contractor involved takes these problems up with the architect or an assistant who, in turn, takes them to the engineer who has performed the design work.

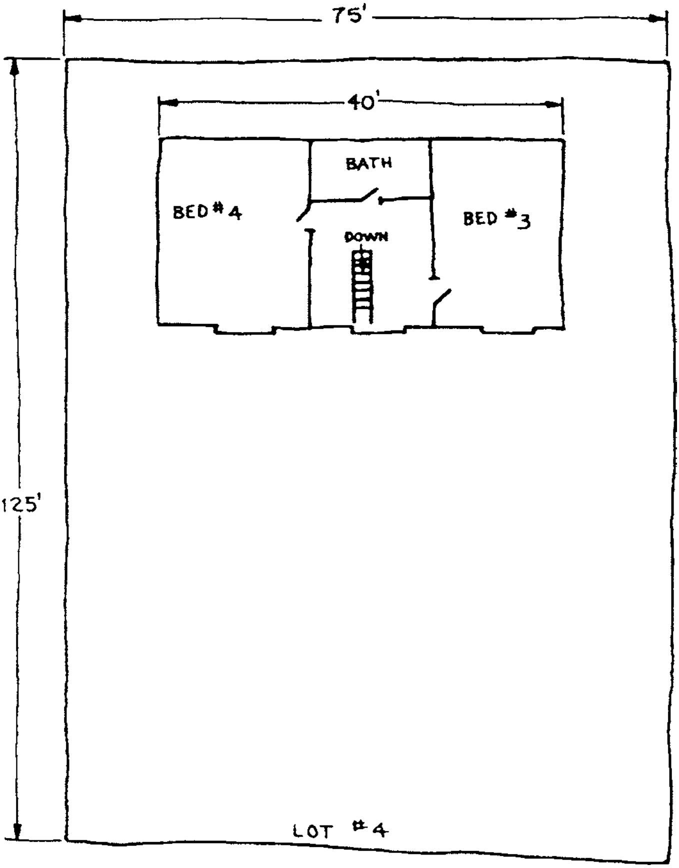

Figure 9-12 Owner's sketch of the second floor of a residence.

There are various methods of bidding on plans. Sometimes, the general contractor gives the entire bid, and then calls for bids from subcontractors. This method has some advantages in that the general contractor is responsible for the entire job. At other times, the general contractor and each subcontractor bid their parts separately. This method has an advantage. The general contractor's percentage is removed from the subcontractors' bidding, and the owner has more control over who gets the bids. The details of the architect's sketches for the preliminary part are not important to the mechanic. Therefore, they will not be covered in this book. What the mechanic is interested in is how to perform part of the work.

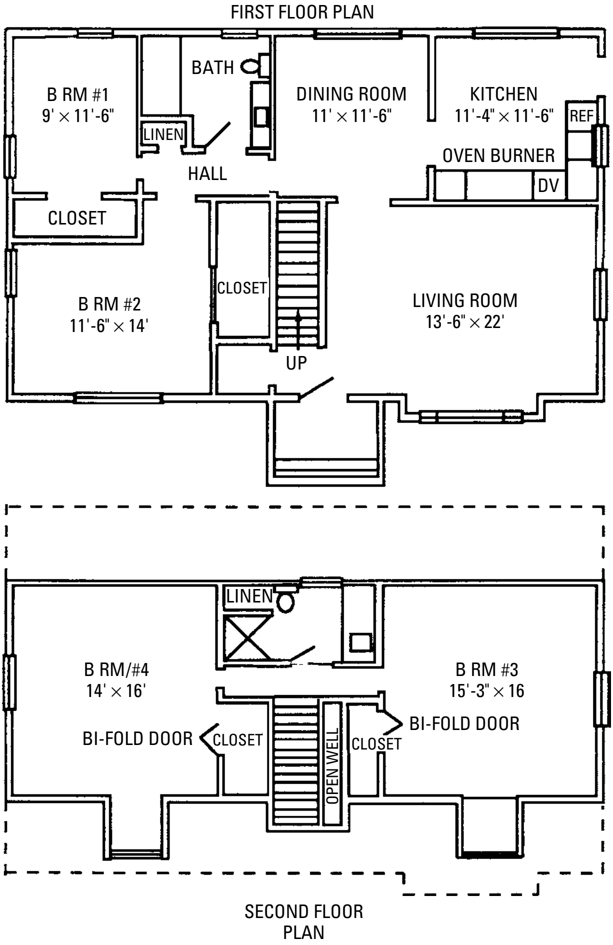
FIRST FLOOR PLAN
B RM #1
9' × 11'-6"
BATH
LINEN
HALL
DINING ROOM
11' × 11'-6"
KITCHEN
11'-4" × 11'-6"
REF
OVEN BURNER
DV
CLOSET
CLOSET
B RM #2
11'-6" × 14'
UP
LIVING ROOM
13'-6" × 22'
LINEN
B RM/#4
14' × 16'
BI-FOLD DOOR
CLOSET
OPEN WELL
CLOSET
B RM #3
15'-3" × 16
BI-FOLD DOOR
SECOND FLOOR
PLAN

Figure 9-13 Architect's drawing (preliminary), without dimensions, of the owner's idea.

Typical floor plans of the first and second floors are shown in Figures 9-14 and 9-15. Figures 9-16 and 9-17 show typical floor plans of a first and second floor, with dimensions added. Figure 9-18 is a typical plan of a basement, with dimensions added. Figure 9-19 illustrates a typical detail of the basement wall, footings, and floor. Figure 9-20 shows a typical detail of the floor joists, brick veneering, and so on as they attach to the foundation.

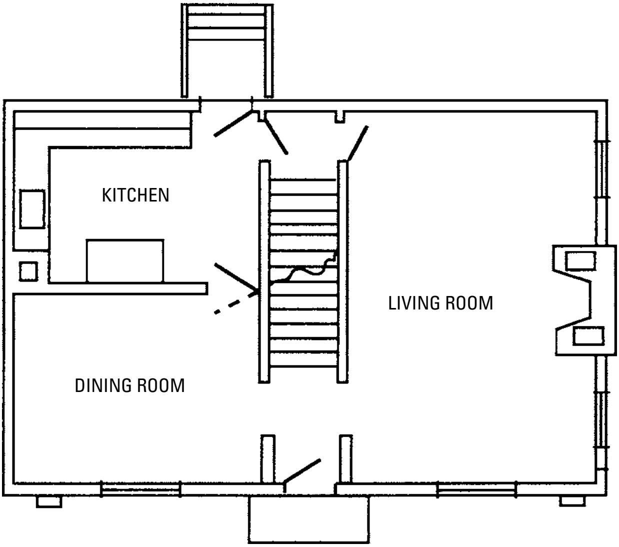
KITCHEN
LIVING ROOM
DINING ROOM

Figure 9-14 **A typical first-floor plan.**

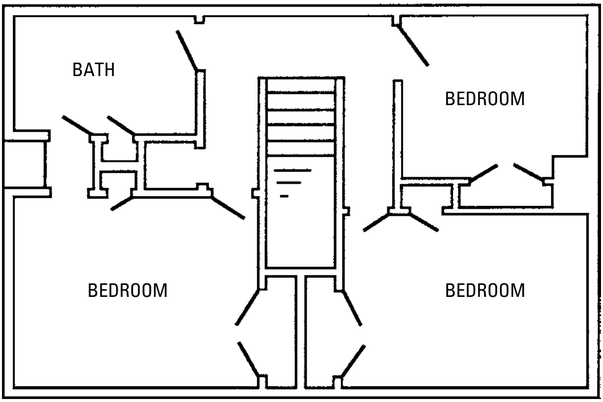
BATH
BEDROOM
BEDROOM
BEDROOM

Figure 9-15 **A typical second-floor plan.**

Graphic Symbols

In architectural drawings, a form of shorthand is used to illustrate what is to be installed and at what point or location the installation will be in the building. These are commonly known as *graphic symbols*.

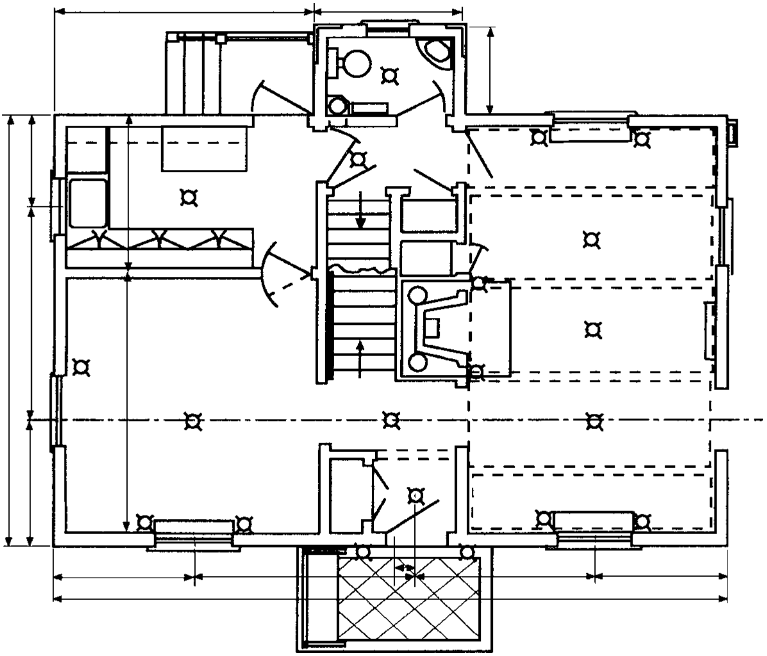

Figure 9-16 A typical first-floor plan with dimension lines.

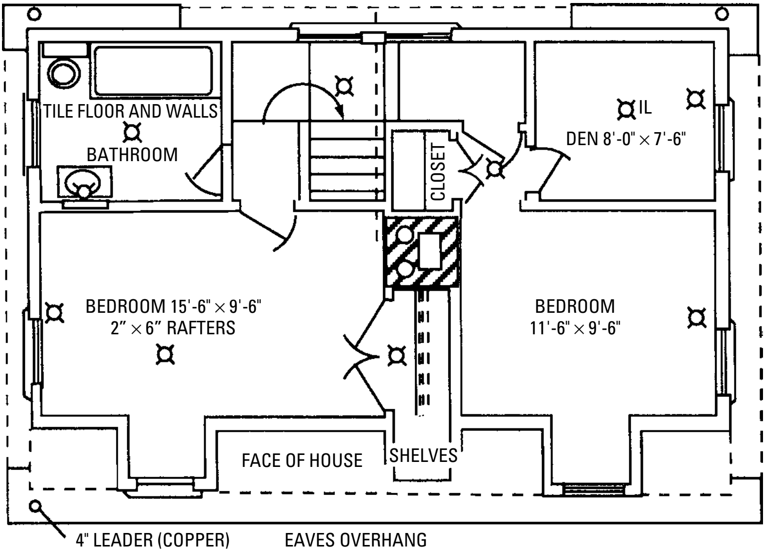
TILE FLOOR AND WALLS
BATHROOM
CLOSET
IL
DEN 8'-0" × 7'-6"
BEDROOM 15'-6" × 9'-6"
2" × 6" RAFTERS
BEDROOM
11'-6" × 9'-6"
FACE OF HOUSE
SHELVES
4" LEADER (COPPER)
EAVES OVERHANG

Figure 9-17 A typical second-floor plan with dimensions.

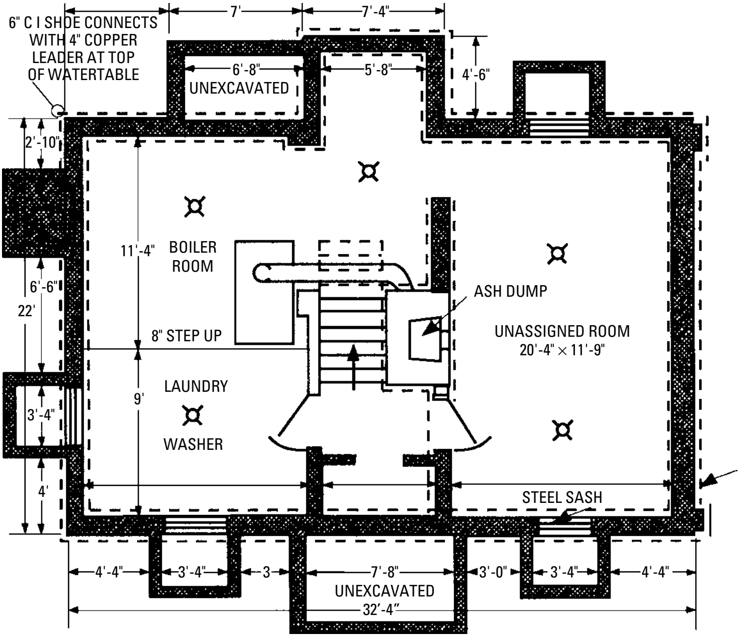
6" C I SHOE CONNECTS WITH 4" COPPER LEADER AT TOP OF WATERTABLE
7'
7'-4"
6'-8"
UNEXCAVATED
5'-8"
4'-6"
2'-10"
6'-6"
22'
3'-4"
4'
11'-4"
BOILER ROOM
8" STEP UP
9'
LAUNDRY
WASHER
ASH DUMP
UNASSIGNED ROOM
20'-4" × 11'-9"
STEEL SASH
12" × 18" CONCRETE FOOTING UNDER 10" CONCRETE WALL
4'-4"
3'-4"
3
7'-8"
UNEXCAVATED
32'-4"
3'-0"
3'-4"
4'-4"

Figure 9-18 A typical basement plan with dimensions.

Symbols are used in drawings to represent various parts and systems. You must also become familiar with these. Each trade has its own symbols, and the workers of each trade should learn to recognize the symbols of all the other trades. For example, the electrician should understand the plumber's symbols. The plumber should understand the carpenter's symbols, and so forth. In this way, each craftsperson will know what obstacles may be encountered in the work, and will be better prepared to cope with them.

In this section, the symbols shown are standards for the construction industry. However, you will find that some designers or individual institutions will deviate from these standards. Where this is done, a legend showing what the symbols mean should be added to the drawings.

A drawing consists of many different kinds of lines, each having its own purpose. Certain characteristic lines are used to convey different ideas, and the drafting practice has been rather well standardized as to the use of lines to avoid confusion in reading drawings. A good working drawing is as simple as possible, using only

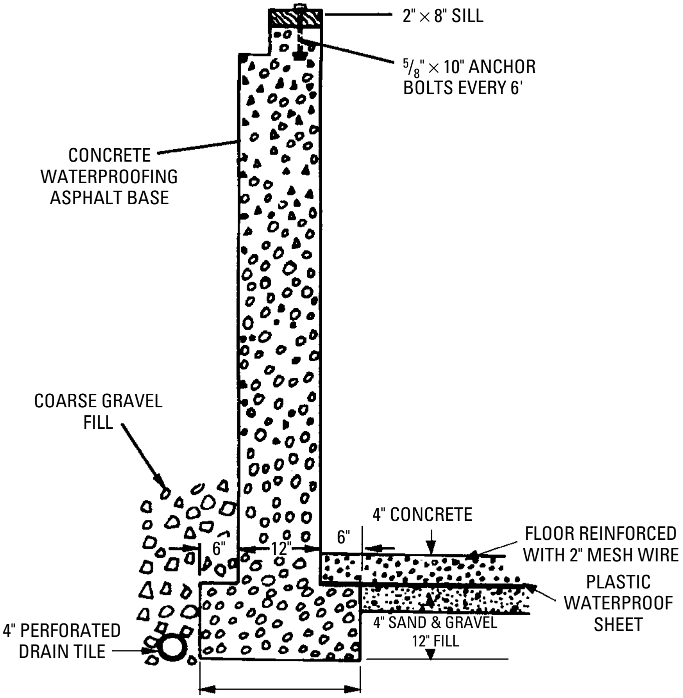

2" × 8" SILL
5/8" × 10" ANCHOR BOLTS EVERY 6'
CONCRETE WATERPROOFING ASPHALT BASE
COARSE GRAVEL FILL
4" CONCRETE
6"
FLOOR REINFORCED WITH 2" MESH WIRE
6"
12"
PLASTIC WATERPROOF SHEET
4" PERFORATED DRAIN TILE
4" SAND & GRAVEL
12" FILL

Figure 9-19 Detail of basement walls, footings, and floor.

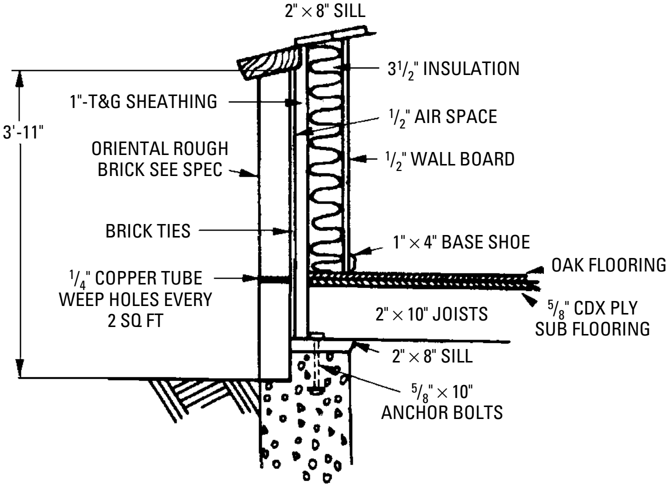

2" × 8" SILL
3'-11"
1"-T&G SHEATHING
3 1/2" INSULATION
1/2" AIR SPACE
ORIENTAL ROUGH BRICK SEE SPEC
1/2" WALL BOARD
BRICK TIES
1" × 4" BASE SHOE
OAK FLOORING
1/4" COPPER TUBE WEEP HOLES EVERY 2 SQ FT
2" × 10" JOISTS
5/8" CDX PLY SUB FLOORING
2" × 8" SILL
5/8" × 10" ANCHOR BOLTS

Figure 9-20 Detail of floor, brick veneering, and so on.

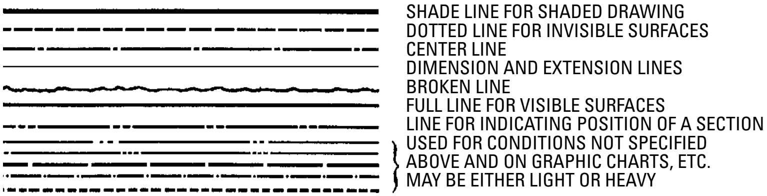
SHADE LINE FOR SHADED DRAWING
DOTTED LINE FOR INVISIBLE SURFACES
CENTER LINE
DIMENSION AND EXTENSION LINES
BROKEN LINE
FULL LINE FOR VISIBLE SURFACES
LINE FOR INDICATING POSITION OF A SECTION
USED FOR CONDITIONS NOT SPECIFIED
ABOVE AND ON GRAPHIC CHARTS, ETC.
MAY BE EITHER LIGHT OR HEAVY

Figure 9-21 Various lines used in drawing.

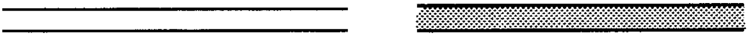

(A) WALL

(B) MASONRY WALL

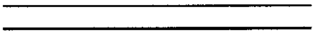

(C) WALL

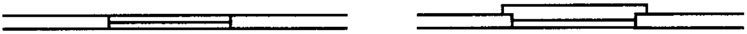

(D) SINGLE-SASH WINDOW

(E) SINGLE-SASH WINDOW
(ALTERNATE)

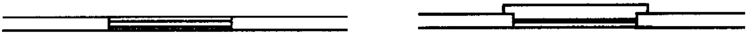

(F) DOUBLE-SASH WINDOW

(G) DOUBLE-SASH WINDOW
(ALTERNATE)

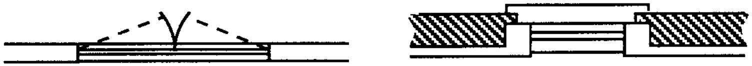

(H) DOUBLE-CASEMENT WINDOW

(I) TWO-SASH MASONRY
WALL WINDOW

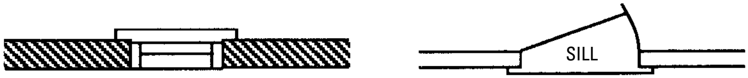

(J) BASEMENT SINGLE-SASH WINDOW

(K) OUTSIDE DOOR

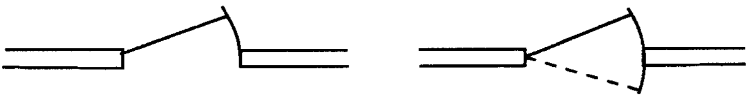

(L) INSIDE DOOR

(M) DOUBLE-SWING DOOR

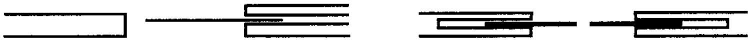

(N) SINGLE SLIDING DOOR

(O) DOUBLE SLIDING DOOR

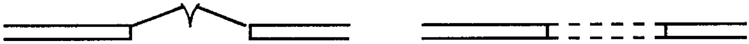

(P) DOUBLE SWINGING DOORS

(Q) ARCH OR CASED OPENING

Figure 9-22 Symbols for walls, windows, and openings.

such lines as are necessary to give all of the required information. Moreover, the reader will not have to puzzle over a mass of lines that complicate the drawing. The same thing holds true for dimensions and other data. A good drawing is accurate and complete, though simple, and is therefore easily understood (read) by the carpenter.

Figure 9-21 shows the lines generally used on drawings.

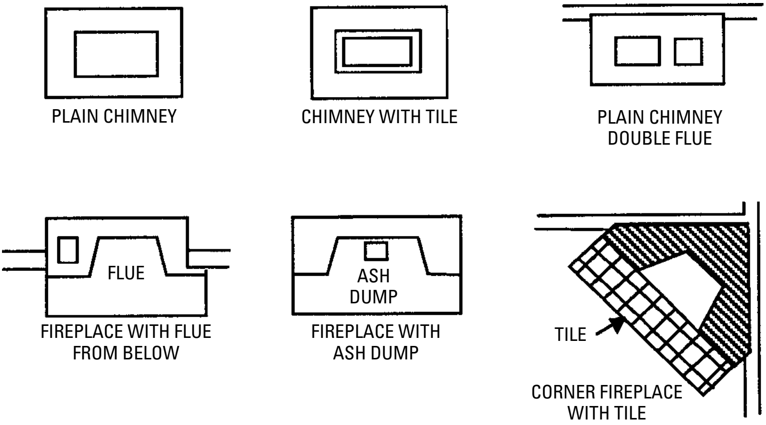
PLAIN CHIMNEY
CHIMNEY WITH TILE
PLAIN CHIMNEY DOUBLE FLUE
FLUE
FIREPLACE WITH FLUE FROM BELOW
ASH DUMP
FIREPLACE WITH ASH DUMP
TILE
CORNER FIREPLACE WITH TILE

Figure 9-23 Symbols for chimneys and fireplaces.

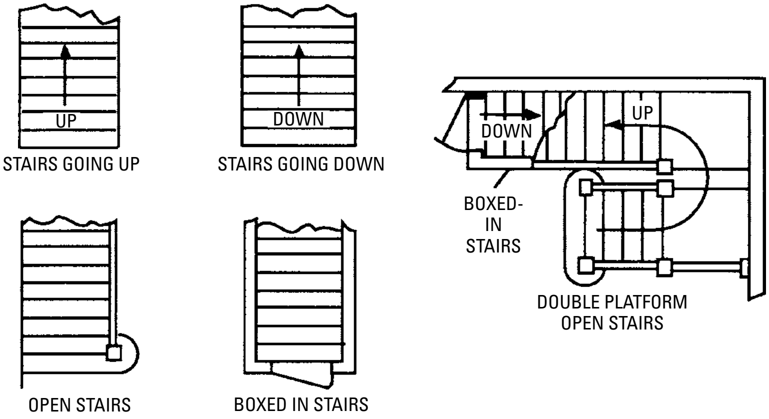
UP
STAIRS GOING UP
DOWN
STAIRS GOING DOWN
DOWN
UP
BOXED-IN STAIRS
DOUBLE PLATFORM OPEN STAIRS
OPEN STAIRS
BOXED IN STAIRS

Figure 9-24 Symbols for stairs.

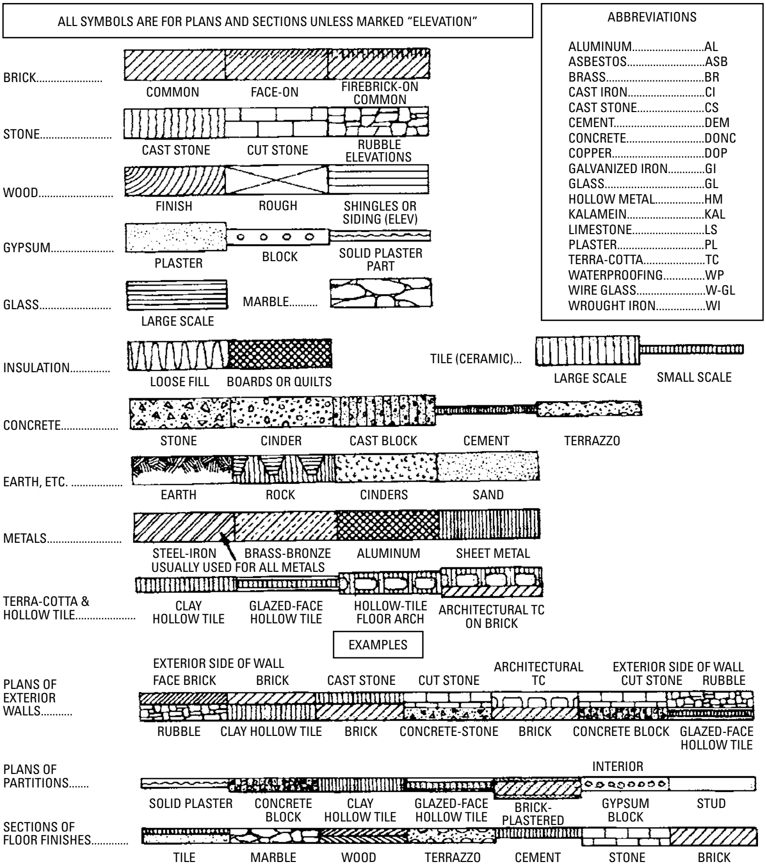
ALL SYMBOLS ARE FOR PLANS AND SECTIONS UNLESS MARKED "ELEVATION"
BRICK.......................
COMMON
FACE-ON
FIREBRICK-ON COMMON
STONE.........................
CAST STONE
CUT STONE
RUBBLE ELEVATIONS
WOOD..........................
FINISH
ROUGH
SHINGLES OR SIDING (ELEV)
GYPSUM......................
PLASTER
BLOCK
SOLID PLASTER PART
GLASS..........................
LARGE SCALE
MARBLE..........
INSULATION..............
LOOSE FILL
BOARDS OR QUILTS
TILE (CERAMIC)...
LARGE SCALE
SMALL SCALE
CONCRETE....................
STONE
CINDER
CAST BLOCK
CEMENT
TERRAZZO
EARTH, ETC. ....
EARTH
ROCK
CINDERS
SAND
METALS..........................
STEEL-IRON
BRASS-BRONZE
ALUMINUM
SHEET METAL
USUALLY USED FOR ALL METALS
TERRA-COTTA & HOLLOW TILE.....................
CLAY HOLLOW TILE
GLAZED-FACE HOLLOW TILE
HOLLOW-TILE FLOOR ARCH
ARCHITECTURAL TC ON BRICK
EXAMPLES
PLANS OF EXTERIOR WALLS...........
EXTERIOR SIDE OF WALL
FACE BRICK
BRICK
CAST STONE
CUT STONE
ARCHITECTURAL TC
EXTERIOR SIDE OF WALL
CUT STONE
RUBBLE
RUBBLE
CLAY HOLLOW TILE
BRICK
CONCRETE-STONE
BRICK
CONCRETE BLOCK
GLAZED-FACE HOLLOW TILE
PLANS OF PARTITIONS.......
INTERIOR
SOLID PLASTER
CONCRETE BLOCK
CLAY HOLLOW TILE
GLAZED-FACE HOLLOW TILE
BRICK-PLASTERED
GYPSUM BLOCK
STUD
SECTIONS OF FLOOR FINISHES.............
TILE
MARBLE
WOOD
TERRAZZO
CEMENT
STONE
BRICK

ABBREVIATIONS

ALUMINUM	AL
ASBESTOS	ASB
BRASS	BR
CAST IRON	CI
CAST STONE	CS
CEMENT	DEM
CONCRETE	DONC
COPPER	DOP
GALVANIZED IRON	GI
GLASS	GL
HOLLOW METAL	HM
KALAMEIN	KAL
LIMESTONE	LS
PLASTER	PL
TERRA-COTTA	TC
WATERPROOFING	WP
WIRE GLASS	W-GL
WROUGHT IRON	WI

Figure 9-25 Symbols for various materials.

Walls of frame buildings are represented on floor plans by two parallel lines spaced at a distance apart equal to the wall thickness (see Figure 9-22A). Masonry walls are shown on a floor plan by cross sectional lines, as shown in Figure 9-22B. Walls of all types of construction may also be shown as in Figure 9-22C, by heavy dark lines that save time in drawing and give a better print.

There are many and varied types of window construction, the symbols for some being shown in Figure 9-22D–J. The specifications should show the materials and types of construction, thickness of glass, and type of glass to be used. These details may be listed as a supplement to the specifications, or on the drawings if there is room.

There are many types of doors that will be used, the symbols for some of these being shown in Figure 9-22K–Q. There may also be special doors used, and where this is called for, a drawing showing the detail should accompany the main drawing. Detailed sketches or drawing inserts should show all details of sills, especially where masonry construction is to be used. In drawings, the dashed line should be avoided where it is intended to indicate some part that is in view. The dashed line is ordinarily intended to represent some hidden feature or part.

Figure 9-23 shows a few conventions or symbols for chimneys and fireplaces. There may be special features that should be shown in additional drawings. In each case where details are required, a notation should be added referring to the detail drawings.

Stairs must be identified as to their direction, and whether they are boxed or open. Figure 9-24 shows some methods of identification. Arrows show the direction of the stairs.

Figure 9-25 shows symbols for the identification of materials.

If confused as to an abbreviation while reading a drawing, considering the nature of the work will be helpful in interpreting the abbreviations. It should be understood that these abbreviations relate only to one part of the subject. For every field (such as carpentry, electrical work, and so on), there are many conventions relating to each individual field. Table 9-1 shows common abbreviations used in construction drawings.

Table 9-1 Common Abbreviations Used in Construction Prints

Construction Feature	*Abbreviation*
Access door	AD
Access panel	AP
Acoustic	ACST
Aggregate	AGGR
Aluminum	AL
Anchor bolt	AB
Angle	ANG
Apartment	APT
Area	A
Area drain	AD
Asbestos	ASB

(continued)

Table 9-1 (continued)

Construction Feature	***Abbreviation***
Asbestos board	AB
Asphalt	ASPH
Asphalt tile	A Tile
Automatic washing machine	AWM
Basement	BSMT
Bathroom	B
Bath tub	BT
Beam	BM
Bearing plat	BRG PL
Bedroom	BR
Blocking	BLKG
Blueprint	BP
Boiler	BLR
Bolts	BT
Book shelves	BK SH
Boundary	BDY
Brass	BR
Broom closet	BC
Building line	BL
Cabinet	CAB
Caulking	CLKG
Casing	CSG
Catch basin	CB
Cellar	CEL
Cement floor	CEM FL
Center	CTR
Center-to-center	C to C
Center line	C/L
Ceramic	CER
Channel	CHAN
Cleanout	CO
Clear glass	CL GL
Closet	CLO
Cold air	CA
Cold water	CW
Conduit	CND

(continued)

Table 9-1 (continued)

Construction Feature	*Abbreviation*
Counter	CTR
Cubic feet	CU FT or FT^3
Detail	DET
Diagram	DIAG
Dining alcove	DA
Dining room	DR
Double-acting door	DAD
Double-strength glass	DSG
Drain	D or DR
Electric panel	EP
End to end	E to E
Excavate	EXC
Expansion joint	EXP JT
Finished floor	FIN FL
Firebrick	FRBK
Fireplace	FP
Fireproof	FPRF
Flooring	FLG
Flush	FL
Footing	FTG
Foundation	FND
Frame	FR
Garage	GAR
Gas	G
Gage	GA
Gypsum	GYP
Hall	H
Hardware	HWD
Hose bibb	HB
Hot air	HA
Hot-water tank	HWT
I-beam	I
Inside diameter	ID
Insulation	INS
Iron	I
Kitchen	K

(continued)

Table 9-1 (continued)

Construction Feature	*Abbreviation*
Knocked down	KD
Landing	LDG
Lath	LTH
Living room	LR
Main	MN
Matched and dressed	M & D
Maximum	MAX
Medicine cabinet	MC
Minimum	MIN
Miscellaneous	MISC
Mixture	MIX
Mortar	MOR
On-center	OC
Pantry	PAN
Partition	PARTN
Plaster	PLAS
Plate	PL
Porch	P
Precast	PRCST
Prefabricated	PREFAB
Pull switch	PS
Radiator	RAD
Recessed	REC
Refrigerator	REF
Register	REG
Revision	REV
Riser	R
Rivet	RIV
Room	R or RM
Rubber tile	R Tile
Screen	SCR
Section	SECT
Sewer	SEW
Shelving	SHELV
Shower	SH
Single-strength glass	SSG

(continued)

Table 9-1 (continued)

Construction Feature	*Abbreviation*
Sink	S or SK
Soil pipe	SP
Square feet	SQ FT or FT^2
Stairs	ST
Standard	STD
Switch	SW or S
Storage	STG
Telephone	TEL
Thermostat	T or THERMO
Tongue-and-groove	T&G
Unexcavated	UNEXC
Vent	V
Vinyl tile	V Tile
Washroom	WR
Water	W
Water closet	WC
Water heater	WH
Weatherstripping	WS

Summary

Drawings are a means of communication between the designers and the builders. They are used on the construction site for reference by the trade workers and as a record of how the building was actually built. The building is represented with various views (such as plans and elevations). Complex views of the building are shown with sections. Sections are like a slice of the building that opens up the interior to a view. A floor plan is a horizontal slice.

Drawings are made to scale. There is a consistent relationship between the length of lines on the drawing and the size of the building and its parts.

Drawings are developed in phases that allow the designer to set the owner's requirements down on paper and show how the design can be changed to meet various functional and legal requirements.

A drawing is made up of symbols that represent the various parts of the building and where those parts will be installed during construction. The symbols are usually standardized so everyone knows what they mean. Simple symbols like lines are used to define the basic shape of the structure. Complex symbols indicate doors, windows, stairways, and other common parts of the building.

Review Questions

1. What is the basic purpose of using drawings in construction of a building? How is this accomplished?
2. Why are the drawings updated during construction?
3. What part of a building does an elevation show?
4. What part of a building does a section show?
5. Why are drawings reduced in size by scaling?
6. Who, besides the carpenters and builders, uses the drawings of a construction project?
7. What do graphic symbols do in a drawing?
8. Draw the symbol for a door and a window.

Chapter 10

Building Styles Explored

Today's homes may resemble those from the New England days of the 1600s, or they may look more like a manufactured-in-the-factory house. There is a large range of styles and looks available for those looking for a new home. Contractors have plans and even model homes to aid in making a selection. There are even certain suggested styles and limits to housing designed for the retired couple with no adult children living with them. Everywhere you look, you see evidence of a variety of tastes expressed in living quarters. Here, we will be exploring some of those styles and learn how to make an educated choice for your next home.

The Two-Story New England Colonial House

Many houses of the Colonial style of architecture, built in the late 1600s, are still standing in the New England states.

Many of the original houses were built to accommodate larger families, but certain architectural details were common to all of them (see Figures 10-1 and 10-2). One feature found in most of the original houses is moderate to steep roof slopes, often the one-third pitch. This was necessary to allow the use of wood shingles or shakes, about the only roof covering then generally available. Narrow eaves (with little or no projection at the gables) had the lap siding cut against wide corner boards. Plank frames were used for the windows, and there were no casings (or very narrow casings) outside. Entrances, however, were usually elaborate, sometimes with finely scrolled and carved pediments. Good, authentic replicas of many of these entrances are obtained today. In the later and more pretentious houses built in this era, sidelights were often used at the entrances, and sometimes the doors were double. The relatively wide pilasters at the sides were usually fluted or molded. A type of door that originated in England is paneled three high and two wide, with the top pair of panels smaller (nearly square). The rails in the upper part of the door form the Christian cross.

The front of the house was symmetrical about the central entrance. Although there were some exceptions to this design, symmetry was the rule. Invariably, second-story windows were placed directly over lower story openings. Small covered entrances were uncommon, porches and verandas virtually unknown. The central entrance hall was universal, with the stairway to the upper floor. In the earlier and smaller Colonial houses, the stairway

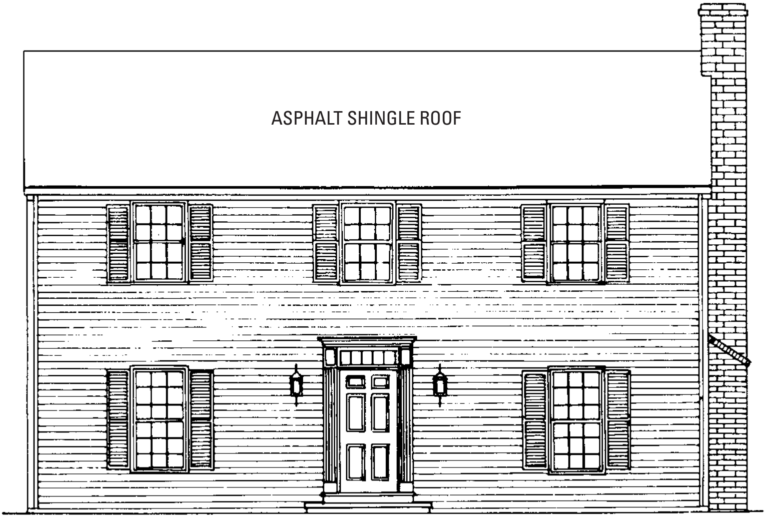
ASPHALT SHINGLE ROOF

Figure 10-1 Front elevation of a Colonial-type home.

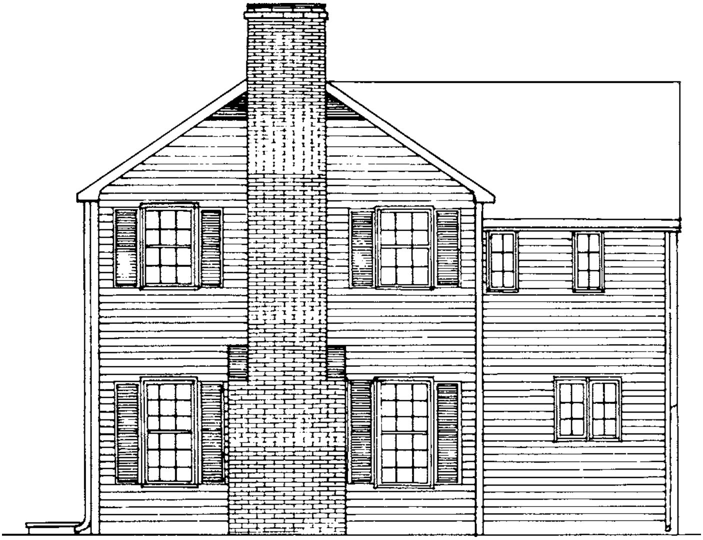

Figure 10-2 Right side elevation of a Colonial-type home.

was often steep and tortuous, often with tricky winders. However, in the later and better homes, the stairs were often beautiful, tastefully designed and elaborately carved, with superb craftsmanship. Such stairs cannot be duplicated today.

While the basic design was rectangular, many of the Colonial homes had attached ells or sheds. With a roof continuous down over a one-story shed at the rear, this house becomes the well-known *saltbox*, with a claim to fame all its own. All types of Colonial homes use windows of small glass or lights, often rectangular, sometimes diamond-shaped. In the early days, the muntins were often made of lead. Later, wood muntins were used. Large sheets of glass were almost unknown, and very expensive.

A predominant feature in the Colonial house is the *privacy* afforded, the privacy that is lacking, and so often deplored, in many modern designs. There are no unnecessarily large areas of glass to give one the eerie feeling of being spied upon at night.

Aside from the fact that this house is excellent architecturally, it need not be an expensive house (relative to other houses) to build. The downstairs bathroom is convenient, and the two complete bathrooms upstairs help prevent congestion (see Figures 10-3 and 10-4). A complete basement is suggested for convenience to heating and

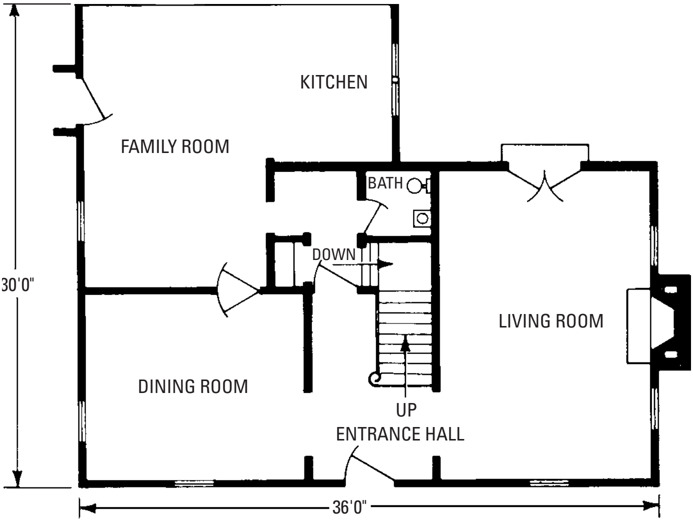

Figure 10-3 First-floor plan of a Colonial-type home.

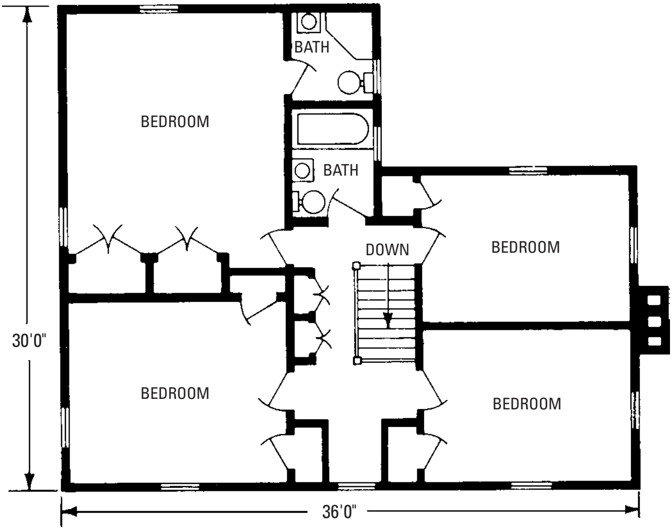
BATH
BEDROOM
BATH
DOWN
BEDROOM
30'0"
BEDROOM
BEDROOM
36'0"

Figure 10-4 **Second-floor plan of a Colonial-type home**

plumbing, plus storage space. Included in the original plans is an adjacent garage. It has a covered breezeway connecting with the outside door of the family room.

A House of Modern Architecture

Architects who have pride in their originality often prefer an unusual custom design (see Figure 10-5). This does not mean that a house of unusual appearance such as this one, no matter how

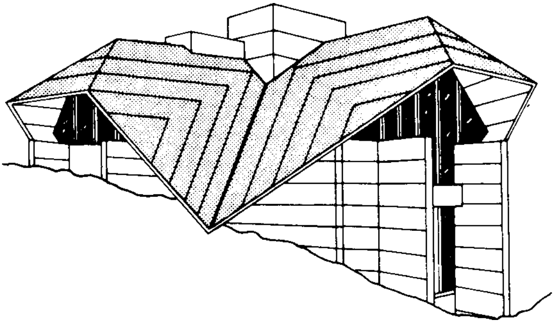

Figure 10-5 **A house of modern architecture.**

good the appearance, can be placed on the odd lot in a street of Colonial houses and have a pleasing effect. Houses such as this one need a proper setting, preferably a rugged and individual setting.

The design is bold, and reflects the lively imagination of its designer. Usually, modern practice makes use of large areas of glass, but these areas are effectively shaded by the wide overhangs, which make them much more acceptable than the large unprotected areas of glass in many modern designs.

The outstanding feature of this house is the roof. Although decidedly unusual, the pitches are regular (all 45°), and the roof is composed of simple intersecting planes. The roof framing is actually rather simple, for there are no *warped* planes. The architect has used diagonal lines on the roof instead of the commonly used vertical and horizontal accents. The effect is striking, especially on a sloping site, but working drawings are available that adapt it to level sites as well.

Although imaginative and unusual, the plan is in right-angled shapes for ease and economy of construction. The roof is formed of intersecting planes, with steep pitches. The roof treatment is important, because it dominates the entire view.

In the plans, designations of rooms, and dimensions, are intentionally omitted (see Figure 10-6 and Figure 10-7). Naturally, the uses of the rooms will depend upon the contour of the ground. The large room with the fireplace will undoubtedly be used for a living room, depending upon whether the entrance is at the second-story level (as it may be if the house is built on a sloping site), or at the first floor level (as it may be if the ground is level). The plan is flexible enough to allow considerable leeway in deciding upon the exact room arrangements and their subsequent uses.

In general, the walls are 12-inch lightweight concrete blocks, their cavities filled with insulation, with head joints cut smooth, and horizontal joints struck. The roof is sheathed with plywood (see Figure 10-8). With a roof so steep, many types of roof coverings may be used, but probably one of the modern types of mopped-down roof covers would be most acceptable, with granule-coated roll roofing as a cap sheet. The horizontal and diagonal rooflines are obtained by the use of battens.

The Contemporary House

Figure 10-9 shows a house called *contemporary architecture*. It embodies most of the features desired in homes of the present era, including the following:

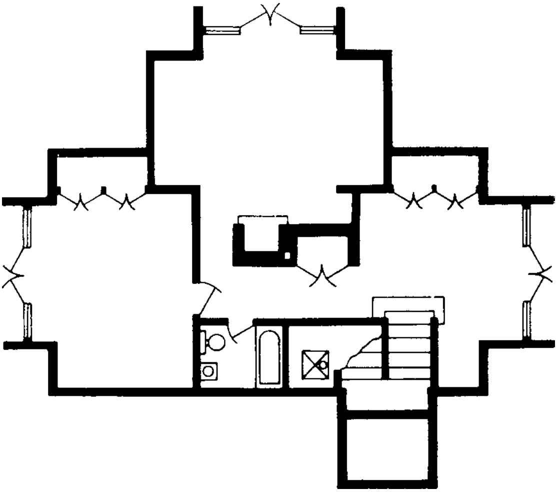

Figure 10-6 The lower-floor level of the modern house. If the entrance is at this level, the room with the fireplace will certainly be used as the living room.

- House is a split-level (not a new idea, but always popular).
- House allows adequate windows in the lower-level rooms, and the stairway down is short.
- Living room is moderately large, and, with the connected dining area, it is ample for the needs of most moderately large families.
- Bedrooms are all of practical, usable size. The spare room in the basement is a useful stand-by.
- Access door to the two-car garage is convenient.
- Family room and the party room in the basement remove some of the inevitable activities of a large family from the living room.
- Concrete patio at the rear is convenient for the usual outdoor living activities of the modern family.

Glass has been used with discretion. Although the glazed areas are moderately large, they are not unreasonably so, at the expense of privacy and to the detriment of heating and air conditioning.

Architecturally, this house is susceptible to many variations of materials. In the house illustrated in Figure 10-9, uncoursed native

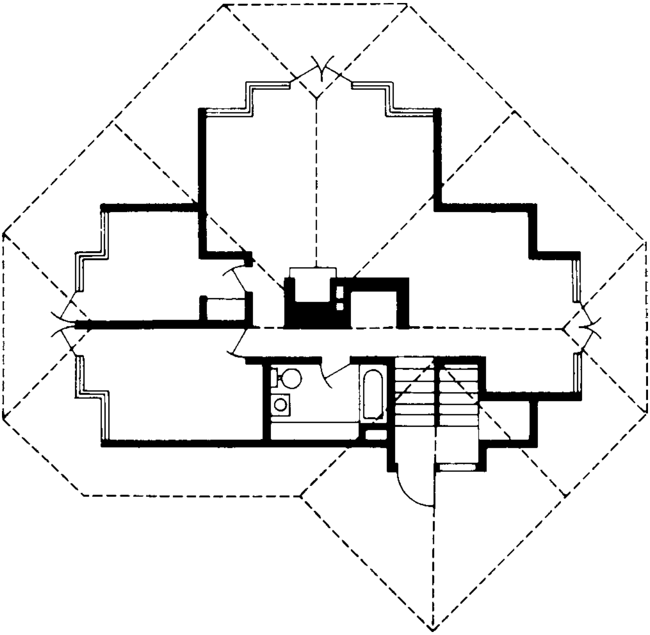

Figure 10-7 The upper-floor level. If the entrance is at this level, the room with the fireplace will certainly be used as the living room.

stone has been tastefully used, with vertical siding on the overhanging upper story.

The roof has sufficient pitch to fill its primary function that is to shed water, but it also gives a lower *spread-out* look to the house.

It is possible with the proper use of a steel beam to enlarge the party or family room. This could be done by removing the wall between party room and spare room, to make one big room in an L shape (see Figure 10-10).

Utility Pole-Type Building

The pole-type building requires no foundation (see Figure 10-11). It is mounted on round pole vertical members that are set in the ground and adequately anchored to resist uplift. For most buildings of appreciable size, the poles are set in round holes about 5 feet deep. Projecting nails or lag screws are placed in the pole 6 to 8 inches from the bottom, or holes are bored through which tight-fitting pieces of rod are driven to project a couple of inches on

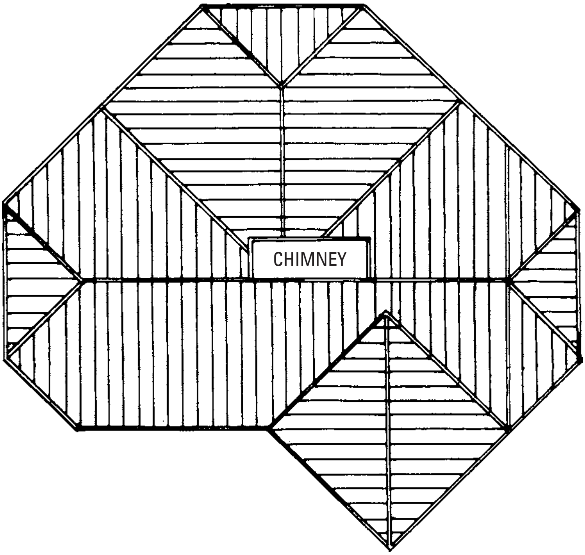

Figure 10-8 **The roof framing plan of the modern house.**

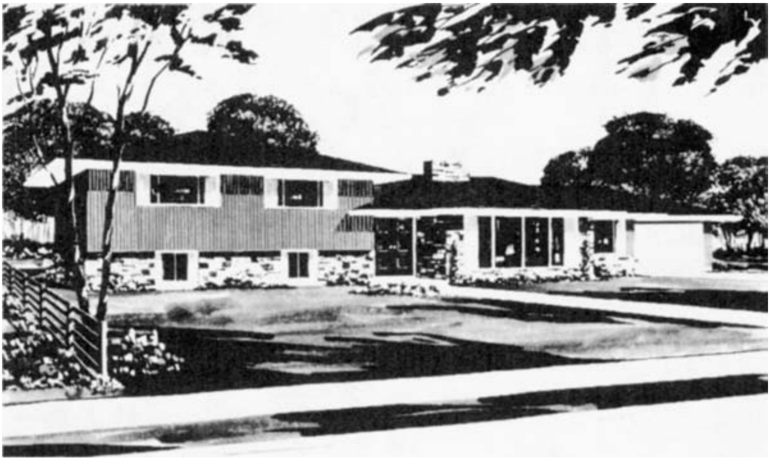

Figure 10-9 **A split-level house.**

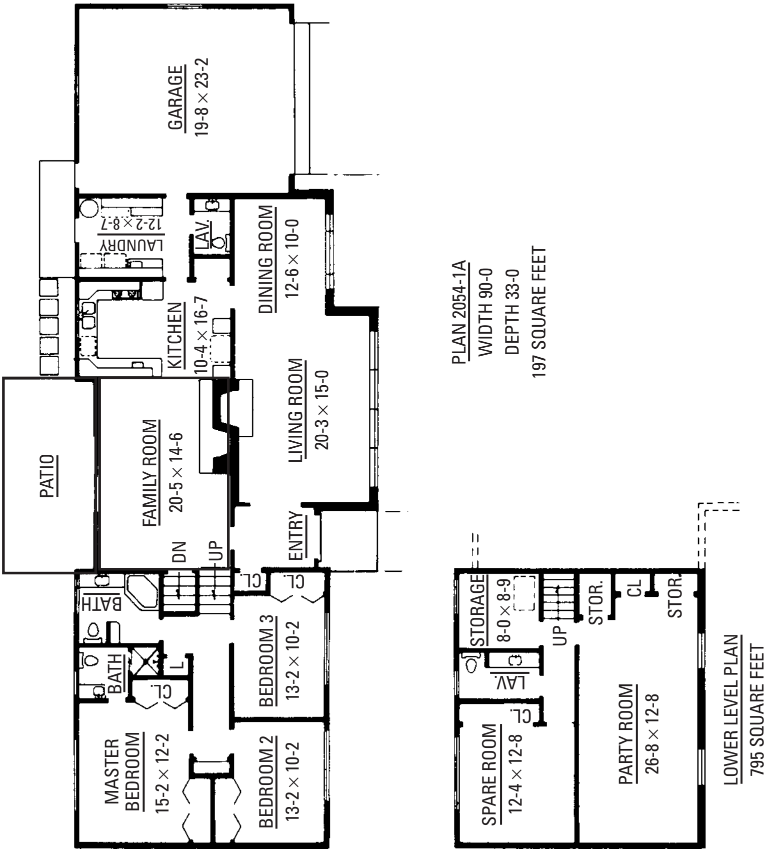
GARAGE
19-8 × 23-2
LAUNDRY
12-2 × 8-7
LAV.
DINING ROOM
12-6 × 10-0
KITCHEN
10-4 × 16-7
LIVING ROOM
20-3 × 15-0
PATIO
FAMILY ROOM
20-5 × 14-6
DN
UP
ENTRY
CL.
CL.
BATH
BATH
L
CL.
BEDROOM 3
13-2 × 10-2
MASTER
BEDROOM
15-2 × 12-2
BEDROOM 2
13-2 × 10-2
PLAN 2054-1A
WIDTH 90-0
DEPTH 33-0
197 SQUARE FEET
STORAGE
8-0 × 8-9
UP
STOR.
CL
STOR.
LAV.
CL.
SPARE ROOM
12-4 × 12-8
PARTY ROOM
26-8 × 12-8
LOWER LEVEL PLAN
795 SQUARE FEET

Figure 10-10 Floor plans for the contemporary house.

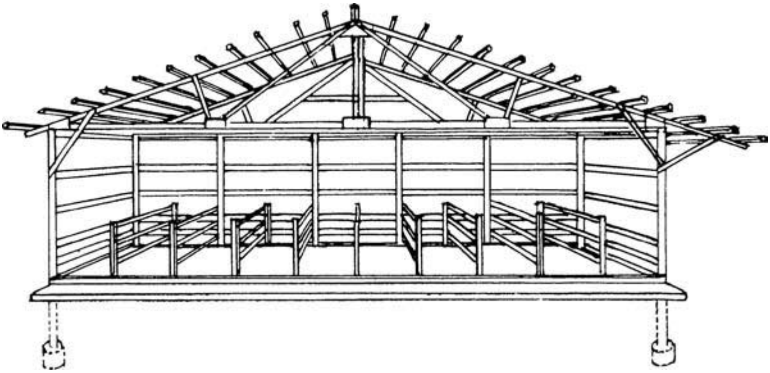

Figure 10-11 A pole-type building.

each side. After the poles are placed and positioned in the holes provided, about 12 inches of concrete is poured around the bottom ends and the holes filled with earth (see Figure 10-12).

Satisfactory joining of rafters and other framing members to the round poles may be made to properly use any standard connecting device (including plain bolts, split-ring connectors, toothed-ring connectors and plain-shank common nails).

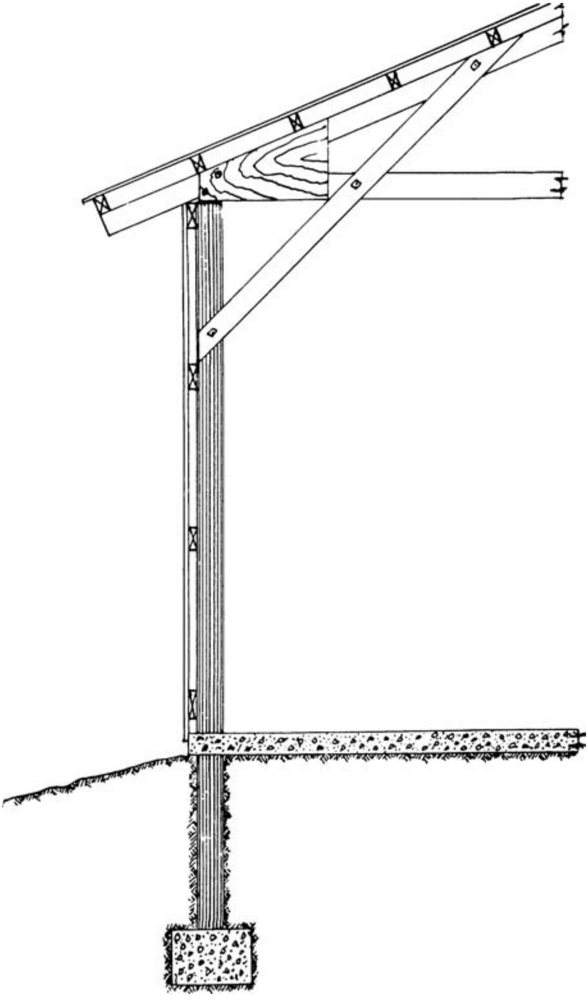

Figure 10-12 Enlarged detail of a pole-and-truss connection.

The idea of the pole-type building is not new. This type of construction was well-known at least 75 years ago, but it was generally used only on inexpensive and temporary buildings. The modern pole building can hardly be placed in this category. The poles used are now pressure-treated with a preservative and have a long life even under most adverse conditions. Modern timber fastenings are far superior to anything that our forefathers had available.

Where rafters or trusses connect to the poles, it is preferable that the poles be flattened at the contacting surfaces (see Figure 10-13). Cast-iron spike-grids, with one side curved to conform to the round side of the pole, have been satisfactorily used. However, simple bolts are also often used.

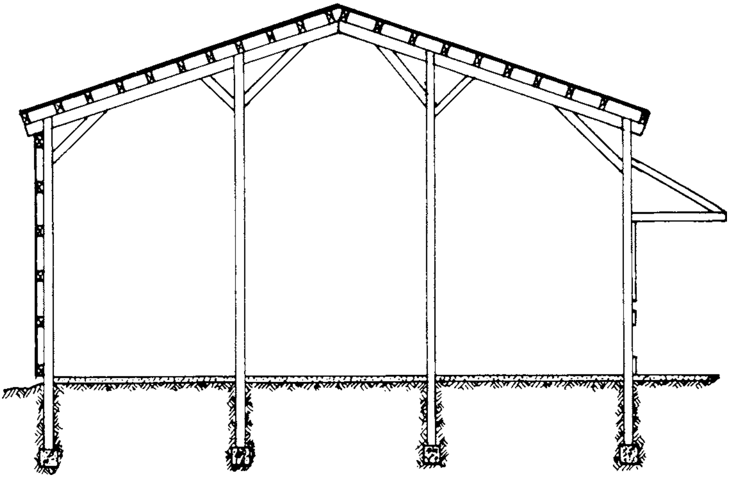

Figure 10-13 A bent of a pole-type hay storage barn, with feeding fence and canopy at the right. Roofs of these buildings may be trussed, eliminating the interior poles. Bents may be set 12 to 16 feet on centers.

Summary

Many of the two-story New England houses built in the 1600s are still standing today. Features found in most original houses were steep roof slopes, wood shingles, narrow eaves, and wide corner boards. The siding was always narrow when used in conjunction with the wide corner boards. Entrances, however, were usually elaborate, sometimes with finely scrolled and cornered pediments.

The contemporary house could be considered a modern home of the present era. Among the many styles in contemporary homes is the split-level, popularly called the tri-level. There are advantages to this type of house. The same space built in a split-level house will cost less than when it is built in a one-story.

A pole-type building (the most popular being the barn style) requires no foundation. The buildings are framed on round poles set in the ground and adequately anchored to resist uplift. The idea for pole-type buildings is not new, as this type of construction was well known at least 75 years ago.

Review Questions

1. What type of house was constructed in the 1600s?
2. What is a split-level house?
3. What are some advantages in the split-level house?
4. Give a few advantages and disadvantages of a basement.
5. What is a pole-type building?
6. What type of house is referred to as a ranch house?
7. What is the primary purpose of a roof?
8. Why would houses located in heavy snow areas have high-pitched roofs?
9. How old is the pole-type building for houses?
10. The ______ house could be considered a modern house of the present era.

Appendix

Construction Hardware

Information provided in this appendix is reproduced with permission from USP Lumber Connectors (United Steel Products Company, Inc.), a company that has been manufacturing and marketing construction hardware for more than 40 years. This information has been culled from the *Engineered Lumber Construction Hardware Product Guide and Design Manual* (Montgomery, Minnesota: United Steel Products Company, 1996).

You may contact the company's corporate offices at the following address:

Box 80
703 Rogers Drive
Montgomery, Minnesota 56069
1-800-328-5934

NDS Standards

The load resistance values presented in this appendix reflect the calculation criteria set forth in the *National Design Specification for Wood Construction* (NDS) published by the American Forest and Paper Association.

Because of this update, load values presented in previous publications may not match those presented in this catalog. The values shown in this appendix supercede those previously printed. An ongoing process is now underway to incorporate these new values in USP code reports. This will be done when each report is scheduled for reevaluation. Any questions about current code listings should be directed to the Technical Service Department at 1-800-328-5934.

NOTE

1) USP Lumber Connectors reserves the right to change specifications, designs and models without notice and liability for such changes. This appendix may not be reproduced in whole or in part without the prior written approval of USP Lumber Connectors.

2) This appendix reflects design changes and design load adjustments to some USP products. The information presented in this appendix supersedes all conflicting information published in previous documents

3) This appendix was designed as a general reference for the USP Product Line. Various specialized publications have been developed for design professionals, truss manufacturers, contractors, retail dealers, and building material wholesalers. Product load values may vary from one publication to another because of recent product testing or changes in regulations.

4) To achieve the allowable loads presented in this appendix, all specified fasteners must be used and proper installation procedures observed. Verify that the dimensions of supporting members are sufficient to receive specified fasteners. No product modifications are allowed without the written permission of USP Lumber Connectors.

5) Some connector models are listed more than once to indicate additional nailing options.

6) Nails specified as 8d, 10d, 16d, and 20d are common wire nails.

7) Bolts specified conform to ASTM A 307 standards or better

8) Products are sized for standard dimensional lumber or EWP sizes. For unusual supporting conditions, excessive shrinkage, or hostile environments, contact USP Lumber Connectors.

Code Evaluations

Most USP structural products listed in this catalog have been evaluated or are in the submittal stage for evaluation from one or more of the following code authorities:

- *CABO*—Council of American Building Officials (NER Report), National Evaluation Service.
- *ICBO*—International Conference of Building Officials.
- *BOCA*—Building Officials and Code Administrators International Inc.
- *SBCCI*—Southern Building Code Congress International, Inc.
- *METRO*—Dade County, FL.
- *DSA*—Division of State Architect, California.
- *LA CITY*—City of Los Angeles, California.
- *TDI*—Texas Department of Insurance.

Product Design Loads

The design loads listed are the lowest result obtained by one of the following methods:

- The ultimate tested load divided by three.
- Load producing 1/8-inch deflection.
- Nail I Bolt values based on 1991 NDS.

Durations of load adjustments for mechanical fastenings are as follows:

- Floor/Design Loads—No increase for duration of load.
- Roof Snow—115% of Design Load for two-month duration of load.
- Roof Nonsnow—125% of Design Load for seven-day duration of load.
- Uplift—133% or 160% of Design Load for wind I seismic loading.

All calculated load values shown in this publication reflect criteria established by the *1991 National Design Specification* (NDS).

Header Materials

- *I-Joist*—I-Joist with similar width to the supported I-joist
- *LVL*—Laminated Veneer Lumber
- *PSL*—Parallel Strand Lumber
- *LSL*—Laminated Strand Lumber
- *DF/SP*—Douglas Fir Larch/Southern Yellow Pine Grade 2 or better
- *SPF*—Spruce/Pine/Fir
- *2× Nailer*—DF/SYP Nailer

NOTE

Charts without specific header callouts ensure DF/SYP.

Testing

On all structurally rated products, USP performs full-scale testing to meet ASTM 0 1761, which is the testing standard recognized by all model agencies. All final testing is witnessed by an independent third-party testing laboratory.

Fasteners

A. *Nails*—Shall be common wire nails. Where special nails are required for the hanger, the manufacturer will furnish all 20d nails and larger, which are shorter than common wire nails of same pennyweight. Special nails shall be furnished by manufacturer.

B. *Bolts*—Where required, shall be of nominal size shown on drawings and shall be furnished by the contractor.

C. *Expanding Anchors*—Use where shown on the drawing; furnished by contractor.

D. *Lag Bolts*—Shall be nominal size shown in drawings; furnished by contractor.

Execution

Examination

A. Verify that the wood that is to support the hanger is flat, is free of large cracks or splits, and is of the species noted on the plans. If the lumber or wood is not Douglas Fir-Larch or Southern Yellow Pine, reduction of the manufacturer's published load values, in accordance with the Building Code, is required. (A design professional shall reduce the published values in accordance with UBC Chapter 23 or the National Design specification of the American Forest and Paper Association.)

Installation

A. Use only nails of size shown in manufacturer's literature.

B. Fill all nail holes with nails as shown in manufacturer's literature.

C. Except where noted in manufacturer's literature, all nails are common wire nails and should be driven at right angles to the hanger body and seated fully.

D. *Do not* use roofing nails, which are not equivalent to common wire nails. DO NOT use nails that are shorter than $1^1/_2$ inches.

E. *Do not* use pneumatic or electric driven staples.

F. If power nail drivers are used, the nails shall be driven through the hole furnished in the hanger and shall not be driven through the hanger where no hole exists.

G. If wood I-joists are used, the maximum nail that can be driven into the top or bottom flange is a 9 gauge × $1^1/_2$ inches long (10d × $1^1/_2$ inches). The maximum nail that can be driven into the web is a 16d nail.

H. When installing hangers to laminated veneer lumber (LVL), the maximum size nail that can be driven into the narrow face is 16d and the maximum size nail that can be driven into the wide face is 20d.

I. Hangers shall be installed tight against the supporting member.

J. The carried member shall be of sufficient length to maintain full bearing on the seat of the hanger. The carried member shall not fit tight (squeeze fit) against the carrying member. A gap of up to $^1/_4$ inch between the carried member and the carrying member is acceptable.

K. Before nailing the carried member into the hanger, take care to fully seat the member so that no space exists between the hanger and the bottom of the carried member. Where safe and practical, do not fasten the carried members to the hanger until floor sheathing has been installed. On some hangers, two diamond holes are located in the bottom of the seat that allow for wood screws to be installed. The screws pull the wood into the seat for positive bearing. Use 2 each #6 × $1^1/_4$-inch wood screws or #6 × $1^1/_4$-inch drywall screws.

L. Where bolting, use a wood template to accurately locate the holes. Drill hole $^1/_{32}$ inch minimum to $^1/_{16}$ inch maximum larger than the bolt diameter it is to receive. The bolt is to be tapped through the hole and connecting device with a hammer or mallet. Always use a washer under the head or nut if it is not in contact with a steel plate or hanger.

M. Lag bolts are not to be used in lieu of bolts unless specified.

N. Screws are not to be used in lieu of nails unless specified.

O. Use only stainless steel nails with stainless steel hangers.

P. Where wood connecting devices are attached to concrete, take care to install the device plumb, square, and true. If practical, use a plywood template to accurately locate the device.

Q. Anchor bolts, furnished and installed under Section 03300 *Cast-in-Place Concrete* shall be accurately located, plumb, and true before wood connecting devices are attached to the anchor bolts.

R. Where prefabricated structural wood is installed into the hanger, follow the written instructions of the prefabricated wood manufacturer in nailing, bearing support, bracing, and installation.

Wood I-Joist Installation to Wood

Sloped I-Joists

Use hangers with sloped seats whenever the slope exceeds the following: $^1/_2$:12 for seat bearing lengths of $2^1/_2$ inches or less; $^3/_8$:12 for bearing lengths between $2^1/_2$ inches and $3^1/_2$ inches; and $^1/_4$:12 for bearing lengths in excess of $3^1/_2$ inches.

Multiple I-Joists

Fasten together multiple plies of wood I-joists, in accordance with the manufacturer's installation guidelines, such that the joists act as a single unit.

Rotation on I-Joists

It may be necessary to install straps, blocking or sheathing to restrain torsional rotation of a supporting wood I-joist when using top mount I-joist hangers.

Fasteners

Install only the specified nails. The flanges of wood I-joists may split if larger diameter nails or longer nails are installed. Do not install nails larger than 16 penny common (0.162 inch diameter) into the web stiffeners in the wood I-joist.

Backer Blocks

Pattern the nails used to install backer blocks or web stiffeners in wood I-joists to avoid splitting the block. The nail pattern should be sufficiently spaced and avoid the same grain line, particularly with solid sawn backer blocks. Backer blocks must be installed on wood I-joists acting as the header, or supporting member. Install in accordance with the I-joist manufacturer's installation guidelines, with the backer blocks seated against the bottom flange of a supporting header when using the face mount hangers, and against the top flange if using top mount hangers. The nails installed to hangers mounted to an I-joist header must penetrate through the web and into the backer block on the opposite side.

Top Flange Hangers

The thickness of the metal and nail heads on top mount hangers must be evaluated for the effect on subsequent sheathing. Ensure the top mount hanger is installed such that the flanges of the hanger are not "over-spread" which tends to elevate the supported I-joist, causing uneven floor surfaces and squeaking. Similarly, ensure the hanger is installed plumb such that the face flanges of the hanger are mounted firmly against the wide-face surface of the header.

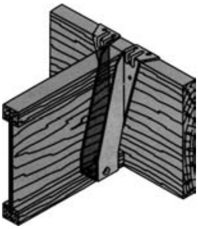
FLUSH FRAMING

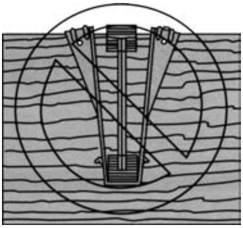
HANGER OVER-SPREAD

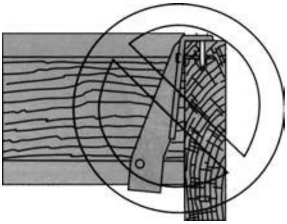
HANGER NOT PLUMB

Support Height and Lateral Stability

Hangers without web stiffeners must support I-joist top flange and provide minimum 250-pound lateral resistance with no more than 1/8″ horizontal deflection.

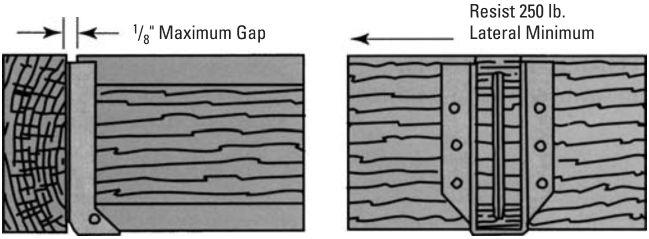
$^{1}/_{8}$" Maximum Gap
Resist 250 lb.
Lateral Minimum

Without Web Stiffeners

Hangers for joists with web stiffeners must support a minimum of 60% of joist depth.

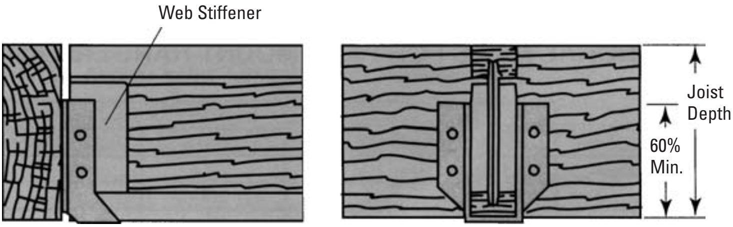
Web Stiffener
Joist
Depth
60%
Min.

With Web Stiffeners or Solid Members

Wood I-Joist Installation to Steel I-Beam

Correct Hanger Attachment to Nailer

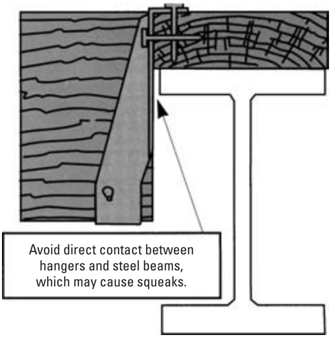
Avoid direct contact between hangers and steel beams, which may cause squeaks.

A nailer or sill plate is considered to be any wood member attached to a steel beam, concrete block wall, stem wall, or other type of support which is unsuitable for nailing, and is used as a nailing surface for top mount hangers to hold beams or joists.

Nailer Sized Correctly

Top flange of hanger is fully supported and recommended nails have full penetration into nailer, resulting in a carried member hanging safely at the proper height.

The nailer must be sized to fit the support width as shown and be of sufficient thickness to satisfy recommended top flange nailing requirements.

NOTE

A design professional must specify nailer attachment to steel beams.

Wrong Nailer Size Can Cause Component Failure

Too Narrow

Top flange not fully supported can cause nail breakout. Or, by fully supporting top flange, hanger is tilted back causing lifting of carrying member which results in uneven surfaces and squeaky floors.

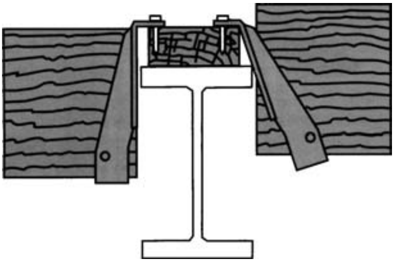

Too Wide

Loading can cause cross grain breaking of nailer. The recommended nailer overhang is 1/4″ maximum per side.

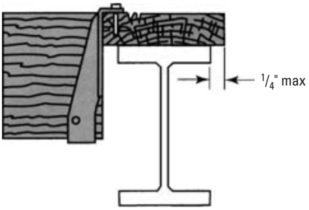

Too Thin

Top flange nailing cannot fully penetrate nailer, causing reduced allowable loads. Never use hangers which require multiple face nails since the allowable loads are dependent on all nail holes being used.

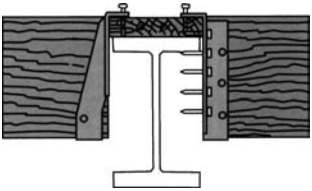

Welded Installation

Where shown in charts, hangers may be welded.
The minimum weld to the top flange is listed in the chart below.

Steel	*Weld*
10 gauge or lighter	1/8″×1 1/2″ each end
7 gauge	3/16″×1 1/2″ each end
3 gauge	1/4″×1 1/2″ each end

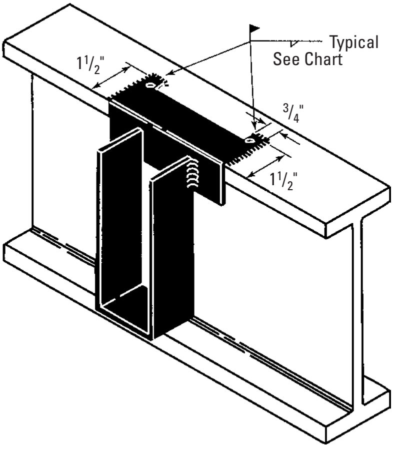

When welding galvanized steel, remove zinc coating by grinding before welding. Chip and touch up weld with primer.

Caution

Use proper respiratory protection when welding galvanized steel.

Top-Mount Hangers THO Series

This top mount hanger is engineered for I-joist to header applications. The THO offers full lateral support of the I-joist top chord which eliminates the need for web stiffeners in most applications. Raised dimple nailing guides help assure correct 45° nailing into the I-joist bottom chord. The new Seat Cleat^{TM} prongs allow builders to“pop” I-joists into hangers for quick, positive seating, Seat Cleats^{TM} are not yet available on all THO models.

Finish	G60 galvanizing
Code Listing	NER 478
U.S. Patent	#5,217,317
Seat Cleat Patent	applied for

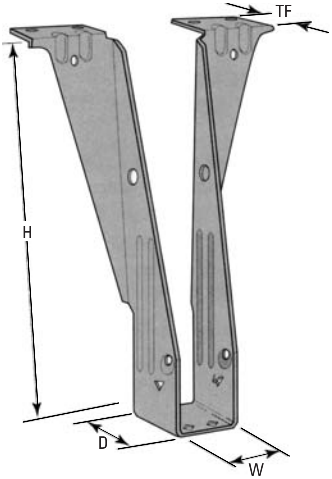

THO

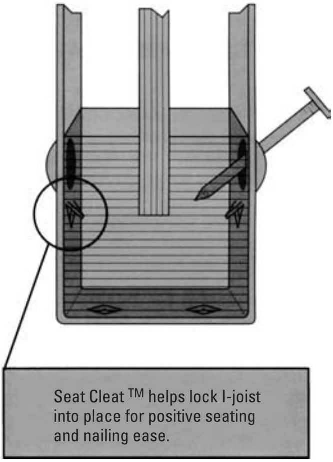

Seat Cleat^{TM} helps lock I-joist into place for positive seating and nailing ease.

Beam and Purlin Hangers BPH and HBPH Series

These BPH hangers are used to support LVL, LSL and PSL beams and headers in light-to-medium load conditions. The slantback design incorporates strength and economy. An open back design allows the builder to install the BPH after the beam is in place.

Finish	G60 galvanizing
Code Listing	NER pending

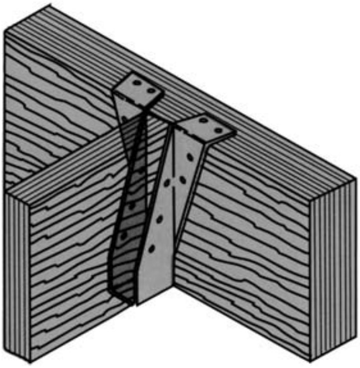

Typical **BPH** Installation

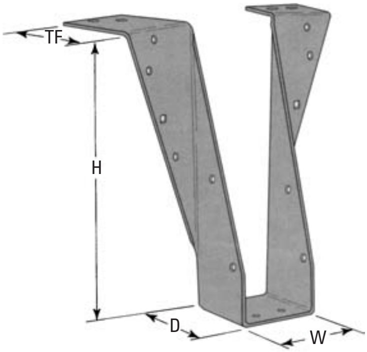

BPH / HBPH Similar

Beam Hangers LBH and HLBH Series

These heavy-duty beam hangers are specially designed for use in concentrated load conditions. The continuous top angle design offers high load values with minimum nailing. Many specialty options are available.

Finish .... paint
Code Listing .... NER pending
Specialty Options .... see page 22

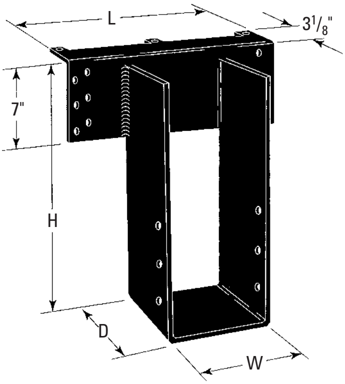

HLBH

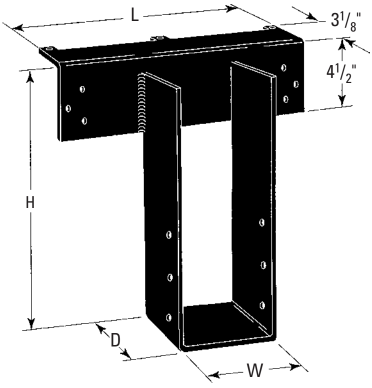

LBH / KGLT Similar

I-Joist Hangers PHI Series

The PHI series is specifically designed to support single-ply I-joists on I-joist headers. The PHI features a continuous top flange for increased strength. These hangers have also been engineered so that the top flange "depth" will match hanger seat width. This ideally places the top flange nails behind the I-joist's center web when the support joist is the same width. Joist nailing and top lateral support allows for installation without web stiffeners.

Finish	paint
Code Listing	NER pending
Specialty Options	see page 22

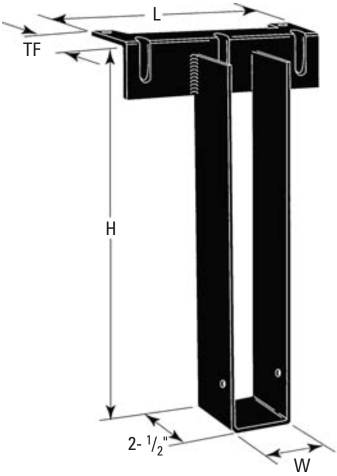

PHI

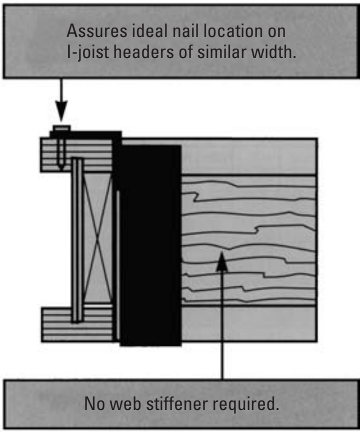
Assures ideal nail location on I-joist headers of similar width.
No web stiffener required.

Beam and Purlin Hangers PH, PHM, PHX Series

Used to hang LVL, LSL and PSL beams and headers in medium load conditions using standard nails. The continuous top flange offers the best loading capacity with minimum nailing. A wide range of specialty options is available for diverse applications.

Finish	paint
Code Listing	NER pending
Specialty Options	see page 22

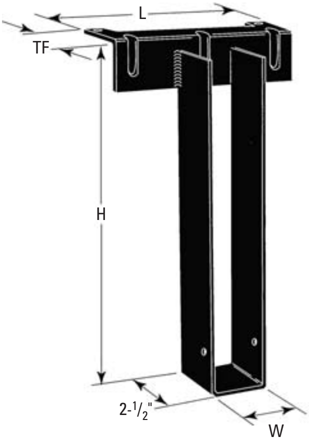

PH / PHM similar / PHX similar

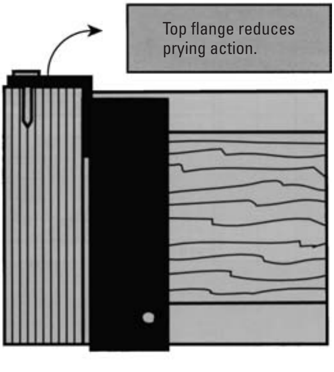

Top flange reduces prying action.

Masonry Hangers MPH Series

Specially designed to work with standard 6-inch concrete block or larger wall construction. The MPH eliminates the need for masons to fabricate special seats to support I-joists or composite wood beams. They can be easily slid into position while a mason is laying block and fastened by setting two 16d duplex nails into the concrete bed. Continuous top flange style gives strength with minimal fastening.

Finish	paint
Code Listings	NER 505
Specialty Options	see page 22

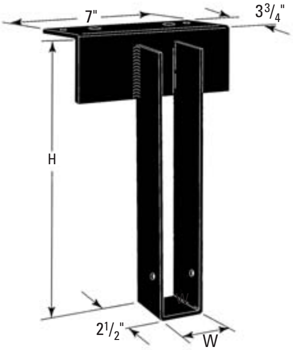

MPH

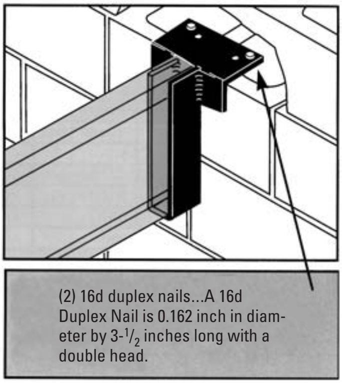

(2) 16d duplex nails...A 16d Duplex Nail is 0.162 inch in diameter by 3-$^1/_2$ inches long with a double head.

Top Mount Hangers

Joist (Inches)	USP Stock No.	Reference Number	Web STFNR Reqd	Steel Gauge	Dimensions					Fastener Schedule		100% Allowable Loads (LBS.)								
					W	H	D	L	TF	Header	Joist	I-Joist	LVL	PSL	LSL	DF/SYP	SPF	†2 × Nailer	††Welded	†††Masonry
7 × 30	HBPH7130	HB1430-2	√	7	$7^{1}/_{8}$	30	3	—	$2^{1}/_{2}$	(10)NA25	(6)NA25	—	—	—	—	5345	3980	—	—	—
	PHX7130	HW1430-2	√	$^{3}/_{10}$	$7^{1}/_{8}$	30	$2^{1}/_{2}$	10	3	(4)16d	(2)10d	—	6015	5680	5755	5045	4230	—	—	—
	LBH7130	GLTV430-2	√	10	$7^{1}/_{8}$	30	4	12	$3^{1}/_{8}$	(9) NA16DRS	(6)16d	—	6500	6560	6010	6400	6400	—	—	—
	HLBH7130	HGLTV430-3	√	7	$7^{1}/_{8}$	30	6	12	$3^{1}/_{8}$	(15)NA16DRS	(6)16d	—	10620	10370	9600	9600	8915	—	—	—
	KGLT3530-2	—	√	$^{3}/_{7}$	$7^{1}/_{8}$	30	5	10	$2^{1}/_{2}$	(10)NA25	(6)NA25	—	—	—	—	8205	—	—	—	—
7 × 32	PHX7132	HWI432-2	√	$^{3}/_{10}$	$7^{1}/_{8}$	32	$2^{1}/_{2}$	10	3	(4)16d	(2)10d	—	6015	5680	5755	5045	4230	—	—	—
	LBH7132	GLTV432-2	√	10	$7^{1}/_{8}$	32	4	12	$3^{1}/_{8}$	(9)NA16DRS	(6)16d	—	6500	6560	6010	6400	6400	—	—	—
	HLBH7132	HGLTV432-2	√	7	$7^{1}/_{8}$	32	6	12	$3^{1}/_{8}$	(15)NA16DRS	(6)16d	—	10620	10370	9600	9600	8915	—	—	—

NOTE: See HEADER MATERIALS on page 3 for specific header information.
†For 2× Nailer installation, substitute 10d × $1^{1}/_{2}''$ nails for header nailing and fill top and upper face nails, If any. Refer to nailer installation on page 9.
††Refer to welded installation detail on page 8.
†††Refer to masonry installation detail on page 10.

MPH, PH, PHI, PHM, and PHX Series Options

Option	*Skewed*	*Sloped*	*Sloped/Skewed*	*Top Flange Offset*	*Saddle*	*Sloped Top Flange*
RANGE	1° to 60°	1° to 45°	skew 1° to 45° sloped 1° to 45°	—	—	0° to 35°
ALLOWABLE LOADS	100% of table load.	100% of table load.	100% OF TABLE LOAD UP TO MAX LOAD OF: PHX - 3500 lbs. PHM - 3000 lbs. PH - 2500 lbs.	HANGER WIDTH: / %OF TABLE LOAD: $3\frac{1}{2}''$ or less / 60% $3\frac{9}{16}''$ to $5\frac{1}{2}''$ / 75% $5\frac{9}{16}''$ to $7\frac{1}{2}''$ / 85%	100% of Table Load per side.	Reduce-allowable table loads using Straight-Line Interpolation.
ORDERING	Add SK, angle required and right (R) or left (L), to product number. Ex. PH1795-SK45R	Add SL, slope required, and up (U) or down (D), to product number. Ex. PH1795-SL30D	See sloped and skewed. Ex. PH1795-SK45RSL30D	Add OS offset required, and right (R) or left (L), to product number. Ex. PH1795-OSL	Add SA, and saddle width required to product number. Ex. PH1795-SA = $5\frac{1}{2}''$	Add SF, angle required, and right (R) or left (L), to product number. Ex. PH1795-SF30L

Sloped/skewed hanger with skews greater than 15° may have all joist nailing on outside flange. Sloped or sloped/skewed hangers with slopes greater than 15° may have additional joist nails added.

All sloped, skewed or combinations require bevel cut on joist in all applications and bearing stiffeners with I-joists.

LBH and HLBH Series Options

Option	***Skewed***	***Sloped***	***Sloped/Skewed***	***Top Flange Offset***	***Saddle***
RANGE	1° to 50°	1° to 45°	skew 1° to 45° sloped 1° to 45°	—	—
ALLOWABLE LOADS	LBH - 4725 lbs. max. HLBH - 8070 lbs. max. LBH / HLBH - 50% of uplift load on skew greater than 15°.	LBH - 4110 lbs. max. HLBH - 7000 lbs. max.	LBH - 3900 lbs. max. HLBH - 6650 lbs. max. LBH/HLBH- 50% of uplift load on skew greater than 15°.	60% of table load for LBH. 45% of table load for HLBH.	100% of table load per side.
ORDERING	Add SK, angle required, and right (R) or left (L), to product number. Ex. LBH3595-SK 45R	Add SL, slope required, and up (U) or down (D), to product number. Ex. LBH3595-SL30D	See sloped and skewed. Ex. LBH3595-SK45RSL30D	Add OS, offset required, and right (R) or left (L), to product number. Ex. LBH3595-OSL	Add SA, and saddle width required to product number. Ex. LBH3595-SA=5$\frac{1}{2}$″

Sloped / skewed hanger with skews greater than 15° may have all joist nailing on outside flange. Sloped or sloped / skewed hangers with slopes greater than 15° may have additional joist nails added. All sloped, skewed or combinations require bevel cut on joist in all applications and bearing stiffeners with I-joists.

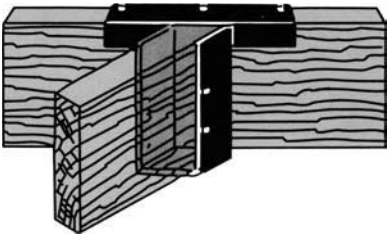

Left **Skewed** Seat

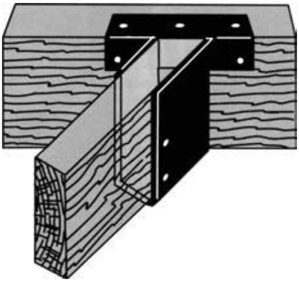

Down **Sloped**
& Left **Skewed** Seat

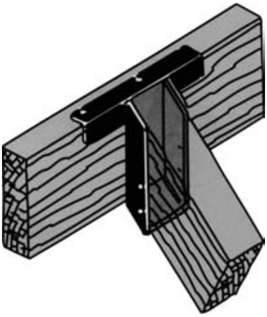

Down **Sloped** Seat

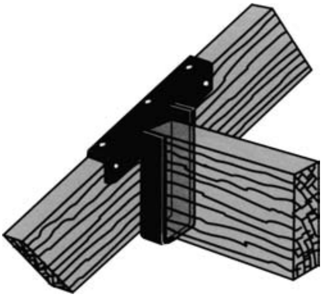

Sloped Top Flange
Down Left, Low Point

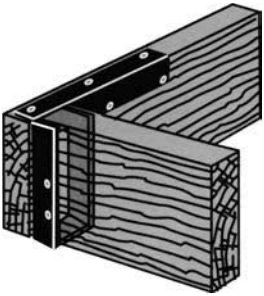

Right **Top Flange Offset**

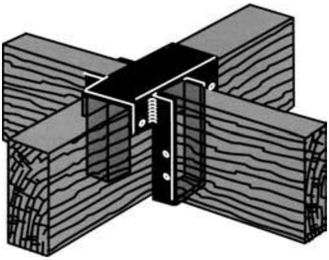

Saddle

Face Mount Hangers HD and HHD Series

These HD/HHD models are specifically designed to support LVL, LSL and PSL beams and headers in medium load conditions. The HD's slant-back design is strong but economical. Both the HD and HHD have open backs, allowing builders to install the hanger after the header or joists are in place. An extra-wide nailing flange also allows for easy field installation.

Finish	G60 galvanizing
Code Listing	NER 478
Specialty Options	see page 27

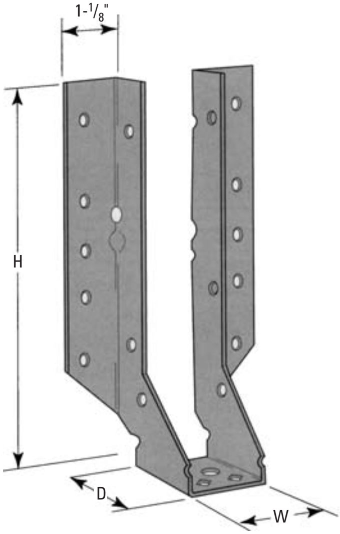

HD

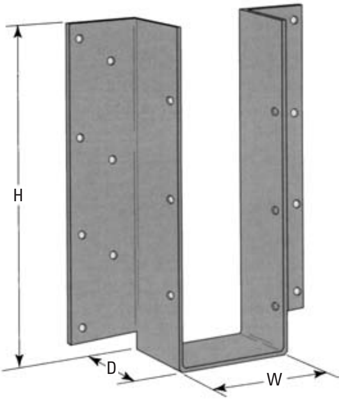

HHD

NOTE

NA20D nails included with all HHD series hangers.

Face Mount Hanger THF Series

The THF is specially designed to support the top chords of I-joists in depths up to 16 inches. This eliminates the need for web stiffeners in most applications. Reinforcement ribs give the THF added strength. Raised dimple nailing guides help assure correct 45° nailing into the I-joist bottom chord. The new Seat Cleat^{TM} prongs (patent applied for) allow builders to "pop" I-joists into hangers for quick, positive seating. Seat Cleats^{TM} are not yet available for all THF models.

Finish	G60 galvanizing
Code Listings	NER 478
U.S. Patent	#5,217,317
Seat Cleat Patent	applied for
Specialty Options	see page 27

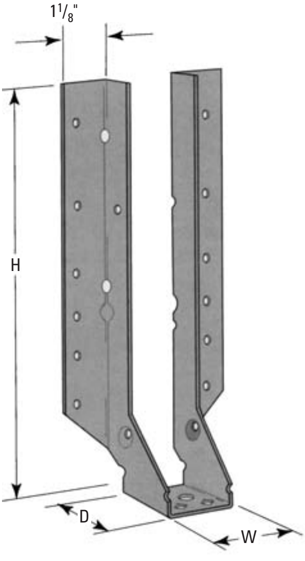

THF Single

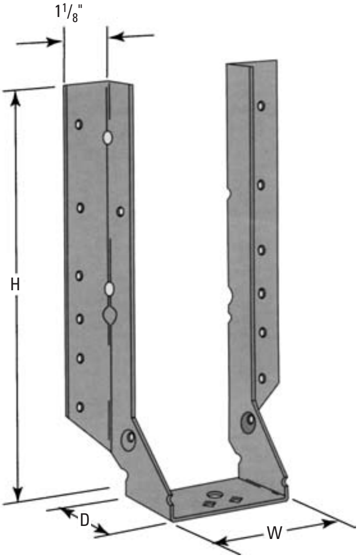

THF Double

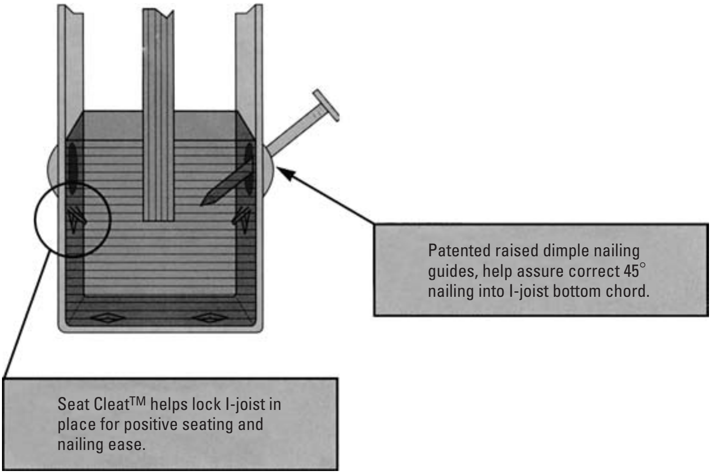

Patented raised dimple nailing guides, help assure correct 45° nailing into I-joist bottom chord.

Seat Cleat™ helps lock I-joist in place for positive seating and nailing ease.

Heavy-Duty Hangers THD Series

Deep seats allow the THD to handle higher loads. The THD offers economical costs from an automated manufacturing process. This model features wide header flanges for easier fastening using standard nails.

Finish	G60 galvanizing
Code Listings	NER 478

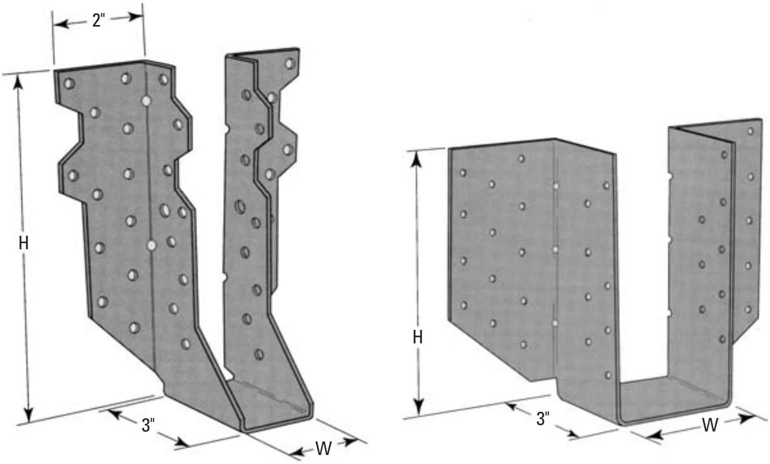

THD Configurations

Heavy-Duty Hangers THDH Series

Special slant nailing and deep seats allow the THDH to handle higher loads. This model features wide header flanges for easier fastening using standard nails.

Finish	G60 galvanizing
Code Listing	NER pending
U.S. Patent	#5,217,317

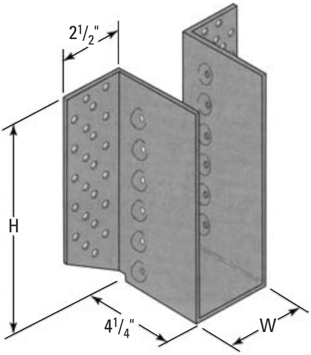

THDH

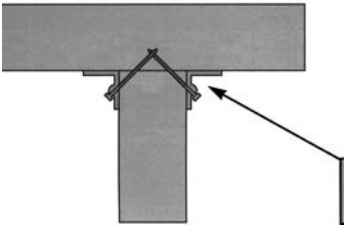

Drive joist nails into header at 45° to achieve listed loads.

HD, HHD and THF Series Options

Option	*Skewed*	*Sloped*	*Sloped/Skewed*	*Inverted Flange*
RANGE	1° to 67$\frac{1}{2}$° when width is 1$\frac{3}{4}$″ or less. 1° to 50° on all others.	1° to 45°	See sloped and skewed.	Not available in widths less than 2$\frac{1}{4}$″.
ALLOWABLE LOADS	100% of table load. 75% of uplift load on skews greater than 15°.	100% of table load.	80% of table load. 75% of uplift load on skews greater than 15°.	100% of table load. 65% of table load when nailing into the support members end grain.
ORDERING	Add SK, angle required, and right (R) or left (L), to product number. Ex. HD17925-SK45R	Add SL, slope required, and up (U) or down (D), to product number. Ex. HD17925-SL30D	See sloped and skewed. Ex. HD17925-SK45RSL30D	Add IF, to product number. Ex. HD410-IF

Sloped/skewed hanger with skews greater than 15° may have all joist nailing on outside flange. Sloped or sloped/skewed hangers with slopes greater than 15° may have additional joist nails added. All sloped, skewed or combinations require bevel cut on joist in all applications and bearing stiffeners with I-joists.

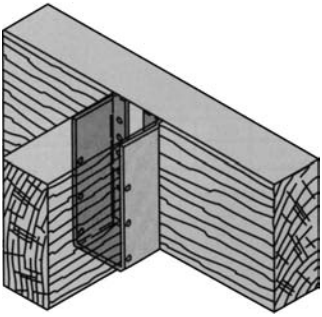

Inverted Flange

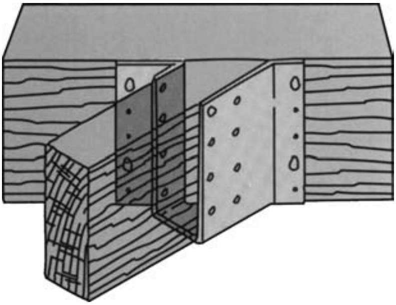

Left **Skewed** Seat

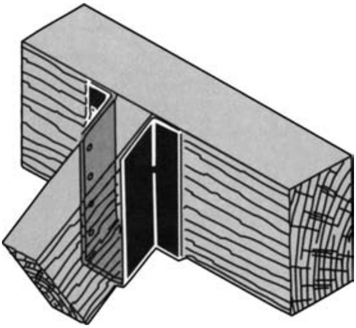

Down **Sloped** Seat

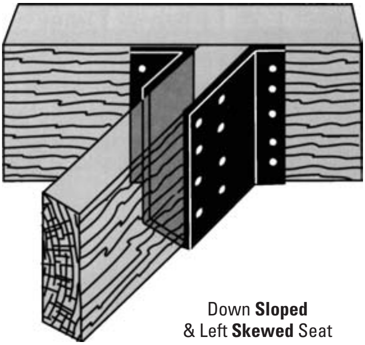

Down **Sloped**
& Left **Skewed** Seat

Knee Braces—KVB Series

USP		Dimensions			Fastener Schedule		Allowable Tension Loads (LBS)				
Stock No.	*Ref. No.*	*D*	*W*	*L*	*Beam*	*Joist*	*100%*	*115%*	*125%*	*133%*	*160%*
KVB5	VB5	10–15 Beam Depth	$3\frac{1}{4}$	$28\frac{5}{16}$	(4) NA25	(12) NA25	1050	1210	1315	1400	1680
KVB7	VB7	15–22 Beam Depth	$5\frac{1}{4}$	$39\frac{5}{16}$	(6) NA25	(12) NA25	1050	1210	1315	1400	1680
KVB8	VB8	$22\frac{1}{2}$–$28\frac{1}{2}$ Beam Depth	$5\frac{1}{4}$	$45\frac{5}{16}$	(6) NA25	(12) NA25	1050	1210	1315	1400	1680
KVB10	VB10	$28\frac{1}{2}$–36 Beam Depth	$6\frac{7}{8}$	$56\frac{5}{16}$	(6) NA25	(12) NA25	1050	1210	1315	1400	1680
KVB12	VB12	36–42 Beam Depth	$6\frac{7}{8}$	$68\frac{5}{16}$	(6) NA25	(12) NA25	1050	1210	1315	1400	1680
KVBI5	—	10–15 Beam Depth	$3\frac{1}{2}$	$28\frac{9}{16}$	(4) 10d	(12) 10d	475	545	595	635	760
KVBI7	—	15–22 Beam Depth	$5\frac{1}{4}$	$39\frac{5}{16}$	(6) 10d	(12) 10d	715	820	895	950	1140
KVBI8	—	$22\frac{1}{2}$–$28\frac{1}{2}$ Beam Depth	$5\frac{1}{4}$	$45\frac{5}{16}$	(6) 10d	(12) 10d	715	820	895	950	1140
KVBI10	—	$28\frac{1}{2}$–36 Beam Depth	7	$56\frac{7}{16}$	(6) 10d	(12) 10d	715	820	895	950	1140
KVBI12	—	36–42 Beam Depth	7	$68\frac{7}{16}$	(6) 10d	(12) 10d	715	820	895	950	1140

NA25 nails are included with the braces.

Knee Braces—KVB and KVBI Series

Knee Braces restrict lateral movement of beam under seismic stress. The KVB can be retrofitted into existing framing. NA25 nails are shipped with KVB Knee Brace and arrive attached to the connector in convenient poly bags.

The new KVBI Knee Brace is designed to be used with I-joist purlins and installs with 10d nails (not supplied).

Finish	G60 galvanizing
Material	12 gauge steel
Code Listings	ICBO 2725
	L.A. City RR25104

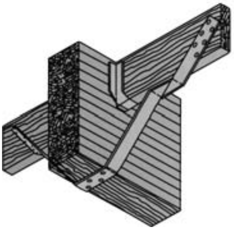
Typical **KVB** Installation

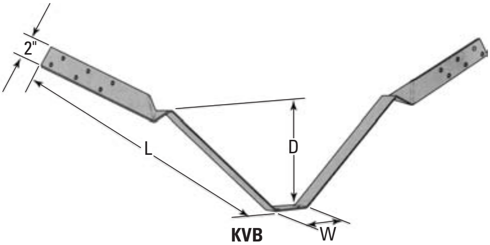

KVB

Hinge Connectors—KHC Series

			Dimensions				*Bolt Schedule*		*Allowable Loads (LBS)*	
USP Stock No.	***Ref. No.***	***Beam Width***	***W***	***Min H***	***PD***	***T***	***Carrying Member***	***Carried Member***	***560 PSI***	***660 PSI***
KHC55	HCA5-5	5 1/8	5 1/4	12	5	3/4	(2)3/4	(2)3/4	14415	16655
KHC56	HCA5-6	5 1/8	5 1/4	17	6	3/4	(2)3/4	(2)3/4	17300	19990
KHC57	HCA5-7	5 1/8	5 1/4	23 1/2	7	3/4	(2)3/4	(2)3/4	20180	23320
KHC59	HCA5-9	5 1/8	5 1/4	38 1/2	9	3/4	(2)3/4	(2)3/4	25945	29980
KHC525	—	5 1/4	5 3/8	12	5	3/4	(2)3/4	(2)3/4	14415	16655
KHC526	—	5 1/4	5 3/8	17	6	3/4	(2)3/4	(2)3/4	17300	19990
KHC527	—	5 1/4	5 3/8	23 1/2	7	3/4	(2)3/4	(2)3/4	20180	23320
KHC529	—	5 1/4	5 3/8	38 1/2	9	3/4	(2)3/4	(2)3/4	25945	29980
KHC75	HCA7-5	6 3/4	6 7/8	13	5	1	(2)3/4	(2)3/4	18985	21940
KHC76	HCA7-6	6 3/4	6 7/8	19	6	1	(2)3/4	(2)3/4	22780	26325
KHC77	HCA7-7	6 3/4	6 7/8	25 1/2	7	1	(2)3/4	(2)3/4	26580	30715
KHC79	HCA7-9	6 3/4	6 7/8	43 1/2	9	1	(2)3/4	(2)3/4	34170	39490
KHC725	—	7	7 1/8	13	5	1	(2)3/4	(2)3/4	18985	21940
KHC726	—	7	7 1/8	19	6	1	(2)3/4	(2)3/4	22780	26325
KHC727	—	7	7 1/8	25 1/2	7	1	(2)3/4	(2)3/4	26580	30715
KHC729	—	7	7 1/8	42 1/2	9	1	(2)3/4	(2)3/4	34170	39490
KHC95	HCA9-5	8 3/4	8 7/8	13	5	1 1/4	(2)3/4	(2)3/4	24610	28440
KHC96	HCA9-6	8 3/4	8 7/8	19	6	1 1/4	(2)3/4	(2)3/4	29530	34125
KHC97	HCA9-7	8 3/4	8 7/8	25 1/2	7	1 1/4	(2)3/4	(2)3/4	34450	39815
KHC99	HCA9-9	8 3/4	8 7/8	42 1/2	9	1 1/4	(2)3/4	(2)3/4	44300	51190

The minimum height is for loads shown. For heights less than the minimum shown reduce the allowable loads in direct proportion.
All bolts are 3/4 inch, and shall meet or exceed the specifications of ASTM A307.
KHC Allowable Loads are normal loads; no increase is permitted.

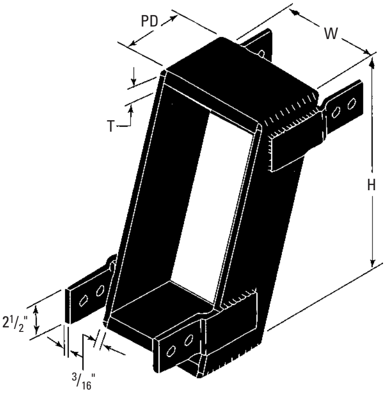

KHC

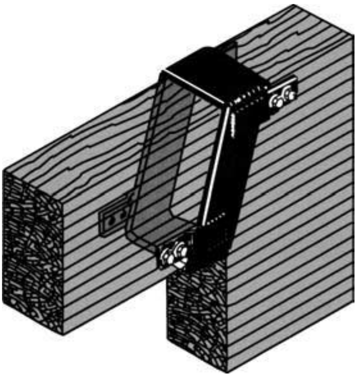

Typical **KHC** Installation

Seismic Straps—KHCST Series

USP Stock No.	Ref. No.	Dimensions L	Dimensions W	Bolt Schedule	Allowable Loads (LBS.) 133%	Allowable Loads (LBS.) 160%
KHCST2	HCST2	22	3 1/2	(4) 3/4	6065	6065
KHCST3	HCST3	28	3 1/2	(6) 3/4	8815	8815
KHCST4	HCST4	34	3 1/2	(8) 3/4	11225	11225

Allowable loads are for straps used in pairs, and are increased 33% or 60% for wind or seismic loads; no further increase is permitted.
Assume a minimum wood thickness of 3 1/8 inches.
Seismic straps shall be used with the KHC hinge connectors.
To order KHCST straps with round holes, add an "R" to the part number, as in "KHCSTR2".

Hinge Connectors KHC Series
Seismic Straps KHCST Series

KHC connectors are used to connect ends of glulam beams having equal widths. These connectors allow long continuous spans, eliminating unnecessary columns. Special dimension variations and models with seismic tabs are available upon request. KHCST straps can be installed during construction or added as a retrofit item. The strap's shape is specially designed to accommodate installation over a USP hinge connector. Order one strap per side.

Material, KHC series ……………………..
……………7 gauge 3/4″, 1″ & 1 1/4″ steel
Material, KHCST series ……. 3 gauge steel
Finish…………………………….paint
Code Listings………………… ICBO 2725
…..…………………L.A. City RR25104

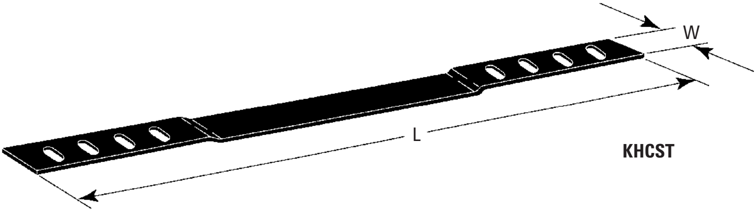

KHCST

Skewed 45° Hangers SKH Series

The SKH series provides stock face-mount hangers at preset 45-degree skews, right or left. The hanger design allows for some installation flexibility—it will accommodate a 40- to 50-degree skew. All models provide uplift capacity and some eliminate the need for miter cuts. A convenient slant nail design provides time savings and ensures proper installation.

Finish . . . . G60 galvanizing
Code Listing . . . . NER 478

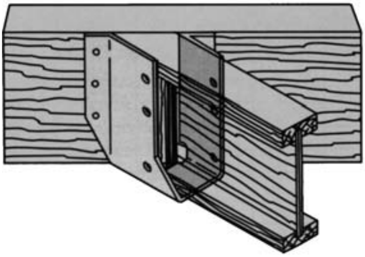
Typical **SKH** Installation

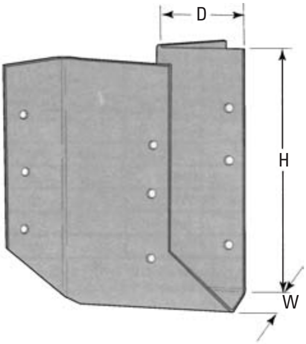

SKH

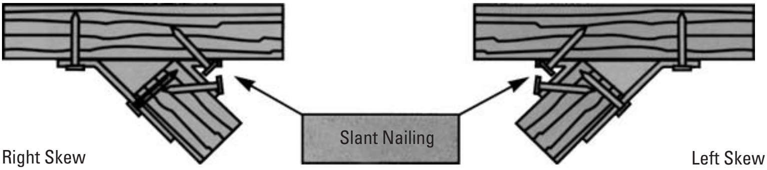

Right Skew

Left Skew

Adjustable Strap Hangers—MSH Series

The MSH is field adjustable; builders can use the flanges in overhang, face mount or combination style. The wide header strap allows for a double nailing row and higher load values. An open-back design allows installation after a member is placed in position.

Finish . . . . G60 galvanizing
Code Listings . . . . SBCCI 9494
. . . . NER pending

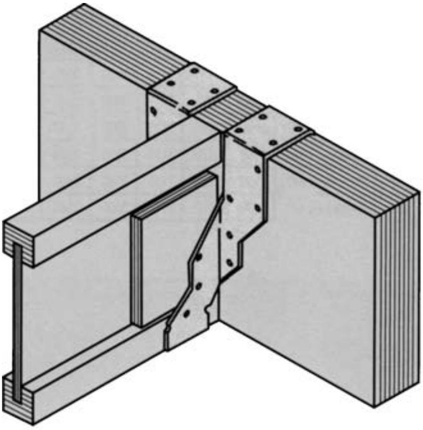

MSH Top Mount Max

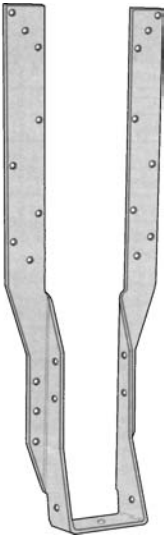

MSH422

Adjustable Rafter-to-Plate Connectors—TMP Series

								Allowable Loads (LBS.)				
								Floor	***Roof***		***Uplift***	
				Dimensions		***Fastener Schedule***			***Snow***	***Non-Snow***		
Refter Width (Inches)	***USP Stock No.***	***Ref. No.***	***Steel Gauge***	***W***	***L***	***Plate***	***Rafter***	***100%***	***115%***	***125%***	***133%***	***160%***
$1\frac{1}{2}$	TMP2	VPA2	18	$1\frac{9}{16}$	$5\frac{9}{16}$	(6)10d	(4)10d × $1\frac{1}{2}$	990	990	990	220	220
$1\frac{3}{4}$	TMP175	VPA25	18	$1\frac{13}{16}$	$5\frac{9}{16}$	(6)10d	(4)10d × $1\frac{1}{2}$	1150	1150	1150	220	220
$2\frac{1}{3}$	TMP23	VPA35	18	$2\frac{3}{8}$	$6\frac{3}{8}$	(6)10d	(4)10d × $1\frac{1}{2}$	1785	1970	1970	220	220
$2\frac{1}{2}$ or $2\frac{5}{8}$	TMP25	VPA2.68	18	$2\frac{11}{16}$	$6\frac{3}{8}$	(6)10d	(4)10d × $1\frac{1}{2}$	1825	1970	1970	220	220
3	TMP31	—	18	$3\frac{1}{8}$	$7\frac{5}{16}$	(6)10d	(4)10d × $1\frac{1}{2}$	1955	1970	1970	220	220
$3\frac{1}{2}$	TMP4	VPA4	18	$3\frac{9}{16}$	$7\frac{5}{16}$	(6)10d	(4)10d × $1\frac{1}{2}$	1970	1970	1970	220	220

No increase for duration of load for uplift is permitted; loads governed by test results.
10d × $1\frac{1}{2}$ nails are 9 gauge (0.148 inch diameter) by $1\frac{1}{2}$ inches long.

Adjustable Rafter-to-Plate Connectors TMP Series

With the TMP, builders can make rafter-to-plate connections at pitches from $^{1}/_{12}$ to $^{6}/_{12}$. This model automatically adjusts to any pitch within its range. There's no need to preset the pitches for every connector, and there's no need for time-consuming bird's-mouth notching or bevel plate installation. This model is available in the popular I-joist sizes.

Finish	G60 galvanizing
Code Listing	NER 478
U.S. Patent	# 5,217,317

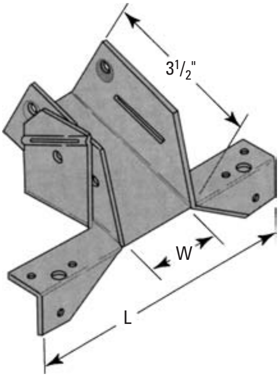

TMP

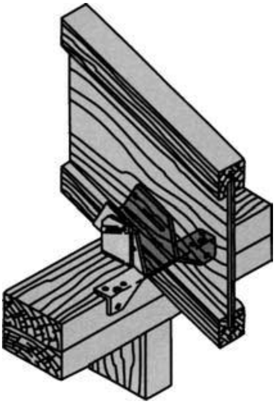

Typical **TMP** Installation

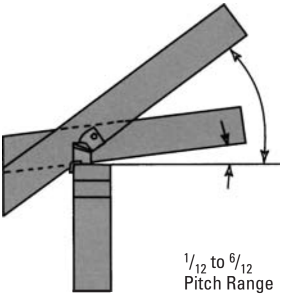

$^{1}/_{12}$ to $^{6}/_{12}$
Pitch Range

Adjustable Rafter-to-Plate Connectors—TMPH Series

Rafter Width	USP		Dimensions			Fastener Schedule		Allowable Loads (LBS.) According to Pitch									Uplift	
(Inches)	Stock No.	Ref. No.	W	L	H	Plate	Rafter	$^{6}/_{12}$	$^{7}/_{12}$	$^{8}/_{12}$	$^{9}/_{12}$	$^{10}/_{12}$	$^{11}/_{12}$	$^{12}/_{12}$	$^{13}/_{12}$	$^{14}/_{12}$	133%	160%
$1\frac{1}{2}$	TMPH2	—	$1\frac{9}{16}$	$6\frac{9}{16}$	$2\frac{1}{2}$	(10) 10d	(8) 10d × $1\frac{1}{2}$	3190	3170	3110	3000	2840	2640	2420	2180	1945	200	200
$1\frac{3}{4}$	TMPH175	—	$1\frac{13}{16}$	$6\frac{9}{16}$	$2\frac{3}{8}$	(10) 10d	(8) 10d × $1\frac{1}{2}$	3190	3170	3110	3000	2840	2640	2420	2180	1945	200	200
$2\frac{5}{16}$	TMPH23	—	$2\frac{3}{8}$	$7\frac{3}{8}$	$2\frac{1}{2}$	(10) 10d	(8) 10d × $1\frac{1}{2}$	3190	3170	3110	3000	2840	2640	2420	2180	1945	200	200
$2\frac{1}{2}$	TMPH25	—	$2\frac{11}{16}$	$7\frac{3}{8}$	$2\frac{5}{16}$	(10) 10d	(8) 10d × $1\frac{1}{2}$	3190	3170	3110	3000	2840	2640	2420	2180	1945	200	200
3	TMPH31	—-	$3\frac{1}{8}$	$8\frac{9}{16}$	$2\frac{11}{16}$	(10) 10d	(8) 10d × $1\frac{1}{2}$	3190	3170	3110	3000	2840	2640	2420	2180	1945	200	200
$3\frac{1}{2}$	TMPH4	—-	$3\frac{9}{16}$	$8\frac{9}{16}$	$2\frac{1}{2}$	(10) 10d	(8) 10d × $1\frac{1}{2}$	3190	3170	3110	3000	2840	2640	2420	2180	1945	200	200

Web stiffeners are required for all wood I-joist installations.
Design Loads may not be increased.
10d × $1\frac{1}{2}$ inch nails are 9 gauge (0.148 inch diameter) by $1\frac{1}{2}$ inches long.

Adjustable Rafter-to-Ridge Connectors TMPH Series

With the TMPH, builders can make rafter-to-plate connections at pitches from $^{6}/_{12}$ to $^{14}/_{12}$. This model automatically adjust to any pitch within their ranges. There's no need to preset the pitches for every connector, and there's no need for time-consuming bird's-mouth notching or bevel plate installation. This model is available in the popular I-joist sizes.

Finish. . . . . G60 galvanizing
Code Listing . . . . NER pending
U.S. Patent . . . . #5,230,198

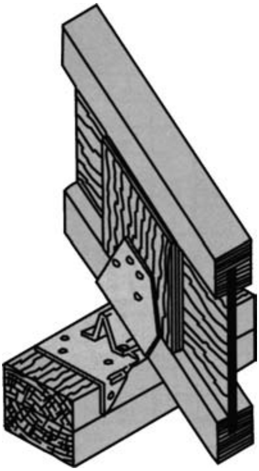
Typical **TMPH** Installation

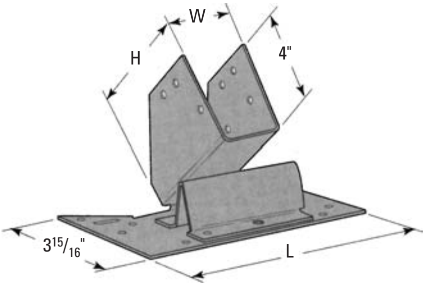

TMPH

TMP and TMPH Installation Procedure:

A) Nail clip to outside edge of wall plate.

B) Insert joist with downward pressure to seat the self-adjusting support.

C) Nail joist in place being sure to nail through the matching slots.

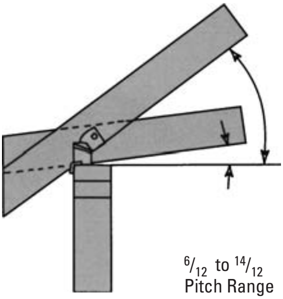

Adjustable Rafter-to-Ridge Connectors—TMU Series

									Allowable Loads (LBS.)				
									Vertical Loads				
										Roof			
Rafter Width	USP		Steel	Dimensions			Fastener Schedule		Floor	Snow	Non-Snow	Uplift	
(Inches)	Stock No.	Ref. No.	Gauge	W	H	L	Header	Joist	100%	115%	125%	133%	160%
$1\frac{1}{2}$	TMU26	LSU26	18	$1\frac{9}{16}$	$5\frac{1}{4}$	$5\frac{1}{4}$	(10) 10d	(7) 10d × $1\frac{1}{2}$	1120	1290	1305	640	640
$1\frac{1}{2}$	TMU210	LSSU210	18	$1\frac{9}{16}$	$8\frac{5}{8}$	$5\frac{1}{4}$	(12) 10d	(11) 10d × $1\frac{1}{2}$	1340	1340	1340	770	770
$1\frac{3}{4}$	TMU175	—	18	$1\frac{13}{16}$	$5\frac{1}{4}$	$5\frac{1}{4}$	(10) 10d	(7) 10d × $1\frac{1}{2}$	1120	1290	1305	640	640
$1\frac{3}{4}$	TMU179	LSSUI25	18	$1\frac{13}{16}$	$8\frac{5}{8}$	$5\frac{1}{4}$	(12) 10d	(11) 10d × $1\frac{1}{2}$	1340	1340	1340	770	770
$2\frac{5}{16}$	TMU23	LSSUI35	18	$2\frac{3}{8}$	$8\frac{7}{8}$	$6\frac{3}{4}$	(12) 10d	(11) 10d × $1\frac{1}{2}$	1345	1545	1680	1340	1605
$2\frac{1}{2}$	TMU25	LSSUI2.68	18	$2\frac{11}{16}$	$8\frac{7}{8}$	$6\frac{3}{4}$	(12) 10d	(11) 10d × $1\frac{1}{2}$	1345	1545	1680	1340	1605
3	TMU31	LSSUI210-2	18	$3\frac{1}{8}$	7	8	(12) 10d	(11) 10d × $1\frac{1}{2}$	1345	1545	1680	1340	1605
$3\frac{1}{2}$	TMU48	LSSU410	18	$3\frac{9}{16}$	7	8	(12) 10d	(11) 10d × $1\frac{1}{2}$	1345	1545	1680	1340	1605

Uplift Loads have been increased $33\frac{1}{3}$% or 60% for wind and seismic loads; no further increases shall be permitted.
10d × $1\frac{1}{2}$ nails are 9 gauge (0.148 inch diameter) × $1\frac{1}{2}$ inches long.
Web stiffeners are required for all Wood I-joist installations.
A tie strap is required for all installations when sloped down at $\frac{7}{12}$ to $\frac{12}{12}$ pitch-use LSTI22 with (4) 10d × $1\frac{1}{2}$ nails.
per end for TMU26, TMU210, TMU175 and TMU179, use KSTI223 with (5) 10d × $1\frac{1}{2}$ nails per end for all others.
10d nails require a header thickness of 1.76 inches to achieve load values shown.
If attached to 2 × lumber, use 85% of listed values.

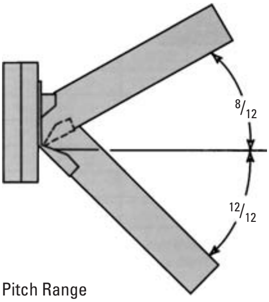

Pitch Range

Adjustable Rafter-to-Ridge Connectors TMU Series

The TMU is designed to connect rafters to ridge beams in vaulted roof structures. These connectors were originally developed to help builders avoid ordering custom sloped and skewed hangers for every situation. The TMU is easily field adjusted to meet a variety of skew and / or slope applications. Shapes to any pitch from 8/12 up through 12/12 down.

Finish	G60 galvanizing
Code Listing	NER 478
U.S. Patent	# 5,217,317

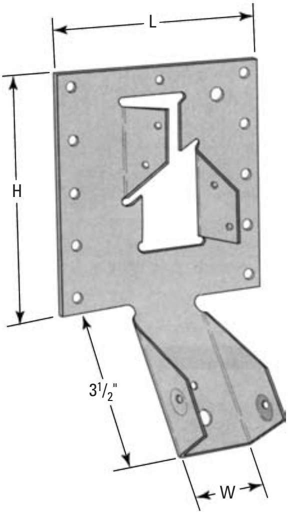

TMU

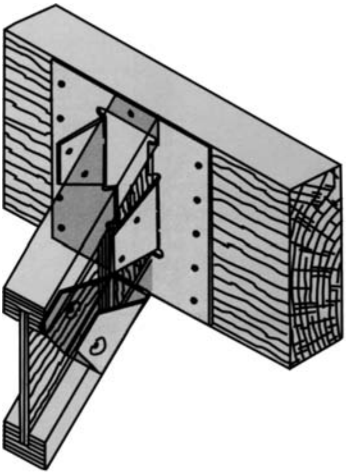

Typical **TMU** Installation

Skewed Installation

1. Install onto end of rafter using recommended nail schedule for joist. Note: When sloping TMU26/175 up, attach hanger to web before nailing to bottom flange.
2. Bend flange to desired angle, and nail in place.
3. Bend opposite flange as shown, and finish nailing.

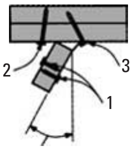

SKEW TO 45°

TMU26, 210, 175 & 179

3/4"×4"×7" Plywood Spacer with (6) 16d Nails

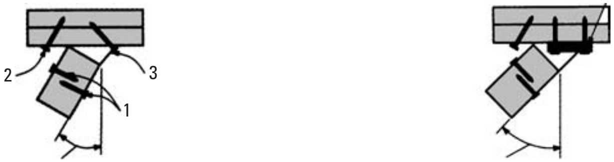

SKEW TO 30°

SKEW 31° TO 45°

TMU23, 25 ,31 & 48

TMU Installation Procedure

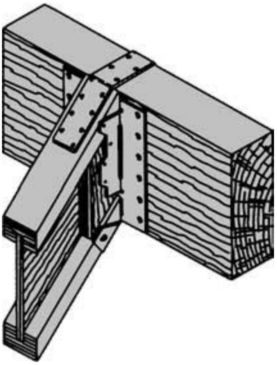

Web stiffeners required for all wood I-joist installations. Tie strap required for all wood I-joist installations when sloped 30° to 45°, or 7:12 thru 12:12. Use LSTI-22 for $1^1/_2''$ & $1^3/_4''$ width joists, use KSTI-223 for $2^5/_{16}''$ or wider joists.

Strap Ties—KSTI and LSTI22 Series

USP Stock No.	*Ref. No.*	*Steel Gauge*	*Dimensions*		*Nail Schedule*	*Allowable Loads (LBS.) Nails*	
			W	*L*		*133%*	*160%*
KSTI223	MSTI23	18	$2^{1}/_{16}$	23	(23) 10d × $1^{1}/_{2}$	1645	1975
KSTI226	MSTI26	12	$2^{1}/_{16}$	26	(26) 10d × $1^{1}/_{2}$	1735	2080
KSTI236	MSTI36	12	$2^{1}/_{16}$	36	(36) 10d × $1^{1}/_{2}$	2400	2880
KSTI248	MSTI48	12	$2^{1}/_{16}$	48	(48) 10d × $1^{1}/_{2}$	2300	3840
KSTI260	MSTI60	12	$2^{1}/_{16}$	60	(60) 10d × $1^{1}/_{2}$	4000	4800
KSTI272	MSTI72	12	$2^{1}/_{16}$	72	(72) 10d × $1^{1}/_{2}$	4510	4950
LSTI22	—	20	$1^{1}/_{4}$	24	(16) 10d × $1^{1}/_{2}$	950	950

Allowable Loads have been increased $33^{1}/_{3}$% or 60% for wind and seismic loads; no further increase shall be permitted. 10d × $1^{1}/_{2}$ nails are 9 gauge (0.148 inch diameter) by $1^{1}/_{2}$ inches long.

Strap Ties KSTI and LSTI22 Series

These strap ties are designed specifically for installation along wood I-joist flanges.

Installation Note

The LSTI22 is for use with 1-$^{1}/_{2}$″ and 1-$^{3}/_{4}$″ I-joists. The KSTI is for use with $2^{5}/_{16}$″ and larger I-joists.

Finish .... G60 galvanizing
Code Listings .... SBCCI 9494
.... NER pending

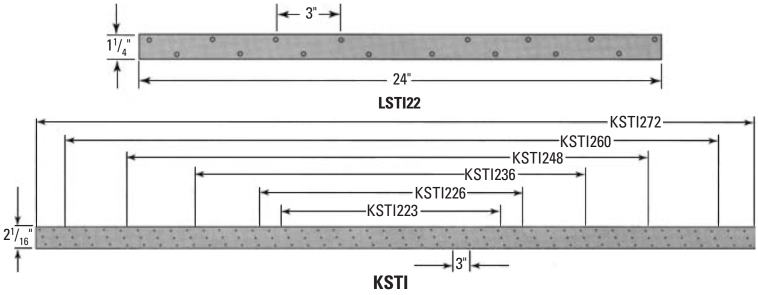

Strap Ties—LSTA and MSTA Series

USP Stock No.	Ref. No.	Steel Gauge	Dimensions		*10d Nails Per End	10d Nails Allowable Loads (LBS.)			*10d × $1\frac{1}{2}$ Nails Per End	10d × $1\frac{1}{2}$ Nails Allowable Loads (LBS.)		
			W	L		100%	133%	160%		100%	133%	160%
LSTA9	LSTA9	20	$1\frac{1}{4}$	9	4	445	590	710	4	375	500	600
LSTA12	LSTA12	20	$1\frac{1}{4}$	12	5	555	740	800	5	470	625	750
LSTA15	LSTA15	20	$1\frac{1}{4}$	15	6	670	895	945	6	570	760	910
LSTA18	LSTA18	20	$1\frac{1}{4}$	18	7	790	1055	1200	7	665	885	1065
LSTA21	LSTA21	20	$1\frac{1}{4}$	21	8	900	1200	1200	8	760	1015	1200
LSTA24	LSTA24	18	$1\frac{1}{4}$	24	9	1025	1370	1605	9	865	1150	1380
LSTA30	LSTA30	18	$1\frac{1}{4}$	30	11	1200	1605	1605	11	1055	1410	1605
LSTA36	LSTA36	18	$1\frac{1}{4}$	36	11	1200	1605	1605	13	1200	1605	1605
MSTA9	MSTA9	18	$1\frac{1}{4}$	9	4	445	590	710	4	375	500	600
MSTA12	MSTA12	18	$1\frac{1}{4}$	12	5	555	740	890	5	470	625	750
MSTA15	MSTA15	18	$1\frac{1}{4}$	15	6	665	890	1065	6	565	750	900
MSTA18	MSTA18	18	$1\frac{1}{4}$	18	7	775	1035	1070	7	660	875	1055
MSTA21	MSTA21	18	$1\frac{1}{4}$	21	8	895	1195	1260	8	760	1015	1215
MSTA24	MSTA24	16	$1\frac{1}{4}$	24	9	1035	1380	1655	9	875	1165	1395
MSTA30	MSTA30	16	$1\frac{1}{4}$	30	11	1265	1685	2000	11	1065	1425	1705
MSTA36	MSTA36	16	$1\frac{1}{4}$	36	13	1495	1995	2000	13	1260	1680	2000
MSTA48	MSTA48	16	$1\frac{1}{4}$	48	14	1500	2000	2000	16	1500	2000	2000

Loads have been increased $33\frac{1}{3}$% or 60% for wind and seismic loads; no further increase shall be permitted.
When 10d nails are used, the minimum wood thickness shall be $1\frac{3}{4}$ inches.
If used on 2× lumber, use the table for 10d× $1\frac{1}{2}$ inch nails.
*Minimum number of fasteners needed in each end to develop strap strength.

Strap Ties LSTA and MSTA Series

With a $1^1/_4$ inch width, the 20-gauge LSTA and the heavier MSTA straps are designed especially for use with $1^1/_2$ inch members. The lighter gauge allows builders to bend these straps for many framing applications.

Finish . . . . G60 galvanizing
Code Listing . . . . NER 505

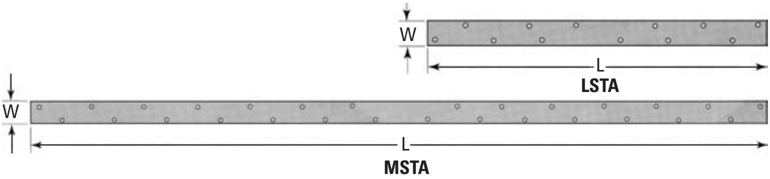

Purlin Anchors PAI Series

Purlin Anchors are engineered to meet seismic and high wind load requirements when securing I-joist purlins and beams to concrete or masonry. Loads are based on a minimum 4-inch embedment into concrete and should be reduced accordingly for masonry or other applications.

Finish . . . . G60 galvanizing
Code Listing . . . . NER 505

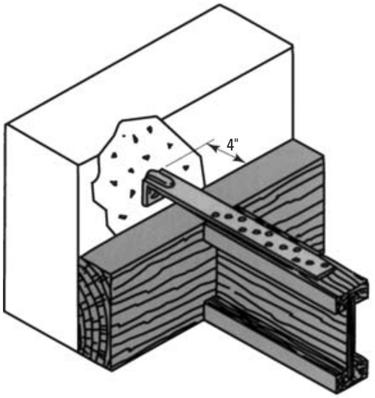

Typical **PAI** Installation

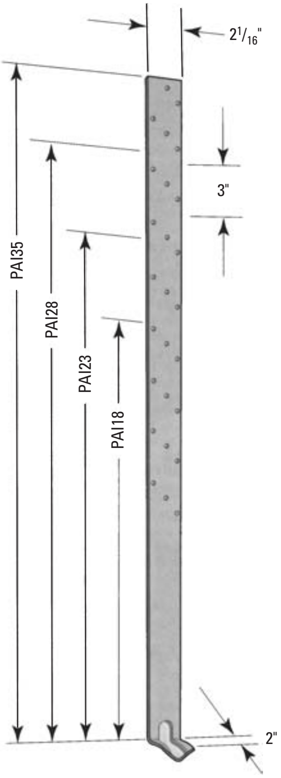

PAI

							Allowable Loads (LBS.)	
				Fastener Schedule			*Nails*	
USP Stock No.	*Ref. No.*	*Dimensions*	*Steel Gauge*	*Nails*	*Bolts*	*Minimum EMBED.*	*133%*	*160%*
PAI18	PAI18	$2 \times 18^1/_2$	12	(12) 10d × $1^1/_2$	—	4	1600	1920
PAI23	PAI23	$2 \times 23^1/_2$	12	(18) 10d × $1^1/_2$	—	4	2400	2880
PAI28	PAI28	$2 \times 28^1/_2$	12	(24) 10d × $1^1/_2$	—	4	3125	3125
PAI35	PAI35	$2 \times 35^1/_2$	12	(26) 10d × $1^1/_2$	—	4	3125	3125

Allowable Loads have been increased $33^1/_3$% or 60% for wind and seismic loads; no further increase shall be permitted.
10d × $1^1/_2$ nails are 9 gauge (0.148 inch diameter) by $1^1/_2$ inches long.

Bracing—N Series

I-Beam Height	*Joist Spacing Inches 12	16	19.2	24
	Sized for Joist to Plate Lateral Bracing			
9¼	N16	N16	N16	N27
9½	N16	N16	N16	N27
11¼	N16	N16	N27	N27
11⅞	N16	N27	N27	N27
12	N16	N27	N27	N27
14	N16	N27	N27	N30
16	N27	N27	N27	N30
18	N27	N27	N27	N30/N36
20	N27	N27	N27/N30	N36
22	N27	N30	N30	N36
24	N30	N30	N30/N36	N36
26	N30	N36	N36	N36
28	N36	N36	N36	N42
30	N36	N36	N36/N42	N42

*Requires (2) 8d common nails or NA11 nails at each end. Nails are to be driven into narrow face of each joist unless noted otherwise.

I-Beam Height	*Joist Spacing Inches 12	16	19.2	24
	Sized for Joist to Joist X Bridging			
9¼	N16	N16	N27	N27
9½	N16	N16	N27	N27
11¼	N16	N27	N27	N27/N30
11⅞	N16	N27	N27	N30
12	N16	N27	N27	N30
14	N16	N27	N27	N30
16	N27	N27	N27	N30
18	N27	N27	N27/N30	N30/N36
20	N27	N27/N30	N30	N36
22	N27/N30	N30	N30/N36	N36
24	N30	N30	N36	N36
26	N36	N36	N36	N42
28	N36	N36	N36/N42	N42
30	N36	N36/N42	N42	N42

*Requires (2) 8d common nails or NA11 nails at each end. Nails are to be driven into top/bottom face of each joist unless noted otherwise.

Bracing and Bridging N Series

N series bridging is quickly installed and economical. This style must be installed prior to the subfloor. Use (2) 8d nails at each end and leave a slight space between the units to avoid noise generating contact.

Material	20 gauge steel
Finish	G60 galvanizing
Code Listing	NER 505

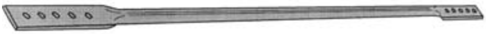

N 16

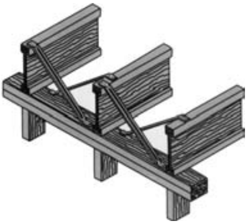

Typical **N** Bracing Installation

Wet Anchors—WE and WA Series

									Allowable Loads (LBS.)											
									L1				**L2**				**Uplift**			
USP			**Dimensions**				**Fastener Schedule**		**Nails**		**Bolts**		**Nails**		**Bolts**		**Nails**		**Bolts**	
Stock No.	**Ref. No.**	**Post Size**	**W**	**H1**	**H2**	**L**	**Nails**	**Bolts**	**133%**	**160%**	**133%**	**160%**	**133%**	**160%**	**133%**	**160%**	**133%**	**160%**	**133%**	**160%**
WE44	PB44	4 × 4	$3^{9}/_{16}$	$4^{13}/_{16}$	$3^{3}/_{8}$	$3^{1}/_{4}$	(12) 16d	$(2)^{1}/_{2}$	1135	1135	910	910	1180	1180	1020	1020	1405	1405	1430	1430
WE46	PB46	4 × 6	$5^{1}/_{2}$	$4^{5}/_{8}$	$3^{1}/_{4}$	$3^{3}/_{8}$	(12) 16d	$(2)^{1}/_{2}$	1135	1135	910	910	1180	1180	1020	1020	1405	1405	1430	1430
WA44	—	4 × 4	$3^{9}/_{16}$	4	5	2	(12) 16d	$(2)^{1}/_{2}$	1575	1575	1525	1830	675	675	1290	1290	1375	1375	1415	1695
WA45	—	4 × 6	$3^{9}/_{16}$	4	5	2	(12) 16d	$(2)^{1}/_{2}$	1575	1575	1525	1830	675	675	1290	1290	1375	1375	1415	1695
WA55	—	6 × 6	$5^{3}/_{8}$	6	5	$2^{1}/_{2}$	(12) 16d	$(2)^{1}/_{2}$	1575	1575	1525	1830	675	675	1290	1290	1375	1375	1415	1695
WA77	—	8 × 8	$7^{1}/_{8}$	$7^{1}/_{2}$	7	3	(12) 16d	$(2)^{1}/_{2}$	1575	1575	1525	1830	675	675	1290	1290	1375	1375	1415	1695

Uplift Loads have been increased $33^{1}/_{3}$% or 60% for wind and seismic loads; no further increase shall be permitted.
Not recommended for non-top-supported installations, such as fences.
H2 is minimum embedment length of anchor into concrete.
Bolt and nail values cannot be combined.

Wet Anchors WE and WA Series

Wet Post Anchors secure posts to concrete in light and medium duty applications such as patio covers, carports or decks. The built-in stand-off plate keeps the post from making contact with the ground and damaging moisture. The WE features a formed, one-piece design which offers additional economy. **Not recommended for fence post applications.**

Material	12 gauge steel
Finish	G60 galvanizing
Code Listing, WA series	NER 505
Code Listings, WE series	ICBO 2039
	NER pending
Code Listing, WE44	SBCCI 9494

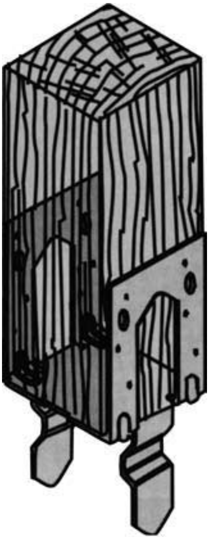

Typical **WE44** Installation

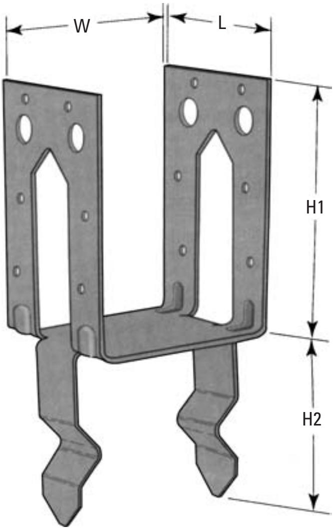

WE

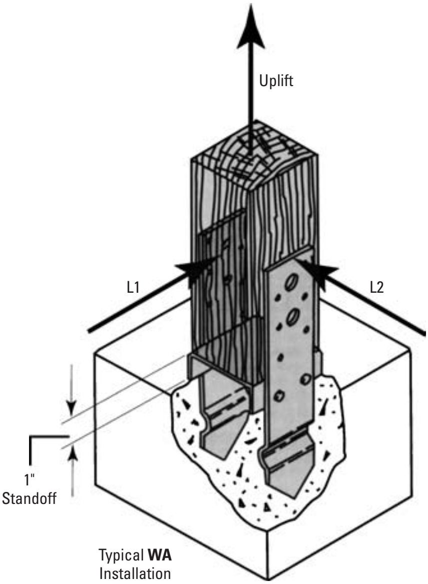
Uplift
L1
L2
1"
Standoff

Typical **WA** Installation

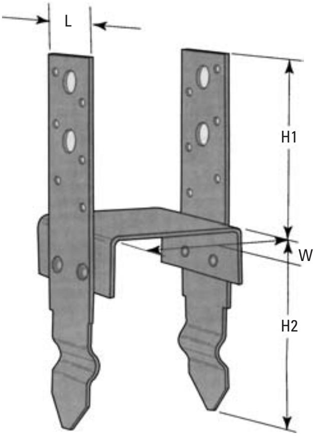

WA

Heavy Column Bases—HBM, CBE, and KCB Series

									Allowable Load (LBS.)				
									Uplift				
			Dimensions				***Fastener Schedule***		***Bolts***		***Nails***		
USP Stock No.	***Ref. No.***	***Column Size***	***W1***	***W2***	***H***	***L***	***Bolts***	***Nails***	***133%***	***160%***	***133%***	***160%***
HBM44	CBS44	$3\frac{1}{2} \times 3\frac{1}{2}$	$3\frac{9}{16}$	$3\frac{9}{16}$	$7\frac{5}{8}$	2	$(2)\frac{5}{8}$	—	5650	4300	—	—
HBM46	CBS46	$3\frac{1}{2} \times 5\frac{1}{2}$	$3\frac{9}{16}$	$5\frac{1}{2}$	$7\frac{5}{8}$	2	$(2)\frac{5}{8}$	—	5650	4300	—	—
HBM66	CBS66	$5\frac{1}{2} \times 5\frac{1}{2}$	$5\frac{1}{2}$	$5\frac{1}{2}$	$7\frac{5}{8}$	2	$(2)\frac{5}{8}$	—	5640	4285	—	—
CBE44	LCB44	$3\frac{1}{2} \times 3\frac{1}{2}$	$3\frac{9}{16}$	$3\frac{1}{2}$	$7\frac{1}{2}$	2	$(2)\frac{1}{2}$	(12)16d	3585	4300	2240	2690
CBE45	—	$3\frac{1}{2} \times 5\frac{1}{4}$	$3\frac{9}{16}$	$5\frac{3}{8}$	$7\frac{1}{2}$	2	$(2)\frac{1}{2}$	(12)16d	3585	4300	2240	2690
CBE47	—	$3\frac{1}{2} \times 7$	$3\frac{9}{16}$	$7\frac{1}{8}$	$7\frac{1}{2}$	2	$(2)\frac{1}{2}$	(12)16d	3585	4300	2240	2690
CBE55	—	$5\frac{1}{4} \times 5\frac{1}{4}$	$5\frac{1}{2}$	$5\frac{3}{8}$	$7\frac{1}{2}$	2	$(2)\frac{1}{2}$	(12)16d	3570	4285	2240	2690
KCB44	CB44	$3\frac{1}{2} \times 3\frac{1}{2}$	$3\frac{9}{16}$	$3\frac{9}{16}$	$8\frac{7}{8}$	2	$(2)\frac{5}{8}$	—	5650	6780	—	—
KCB45	—	$3\frac{1}{2} \times 5\frac{1}{4}$	$3\frac{9}{16}$	$5\frac{3}{8}$	$8\frac{7}{8}$	2	$(2)\frac{5}{8}$	—	5650	6780	—	—
KCB47	—	$3\frac{1}{2} \times 7$	$3\frac{9}{16}$	$7\frac{1}{8}$	$8\frac{7}{8}$	2	$(2)\frac{5}{8}$	—	5650	6780	—	—
KCB55	—	$5\frac{1}{4} \times 5\frac{1}{4}$	$5\frac{1}{2}$	$5\frac{3}{8}$	$8\frac{7}{8}$	3	$(2)\frac{5}{8}$	—	5640	6670	—	—
KCB57	—	$5\frac{1}{4} \times 7$	$5\frac{1}{2}$	$7\frac{1}{8}$	$8\frac{7}{8}$	3	$(2)\frac{5}{8}$	—	5640	6680	—	—
KCB77	—	7×7	$7\frac{1}{2}$	$7\frac{1}{8}$	$9\frac{3}{4}$	3	$(2)\frac{3}{4}$	—	8075	8185	—	—

Uplift Loads have been increased $33\frac{1}{3}$% or 60% for wind and seismic loads, no further increase shall be permitted.
Not recommended for fences.
Minimum edge distance is 3 inches.

Heavy Column Bases HBM, CBE, and KCB Series

These sturdy bases secure heavy columns to concrete footings or foundations. The base is set in position prior to the pour to assure proper alignment and installation. Designed for structural integrity in high wind or seismic applications. **These models are not recommended for fence post applications.**

Installation Note

For KCB installations, the base plate must be flush with the concrete surface to develop full bearing capacity.

Materials, HBM & CBE series	12 gauge steel
Materials, KCB series	$\frac{3}{16}$″ or $\frac{1}{4}$″ steel
Finish, CBE series	G60 galvanizing
Finish, HBM & KCB series	paint
Code Listings, KCB & CBE series	SBCCI 9494 NER pending
Code Listings, HBM	ICBO 2039 L.A. City RR23888
Code Listings, KCB	ICBO 2725 L.A. City RR25104

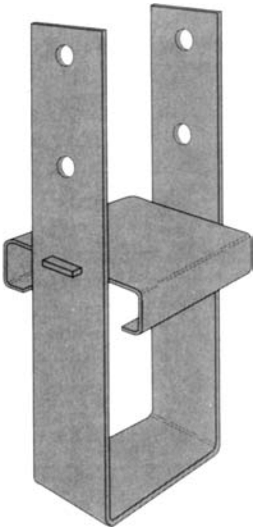

HBM

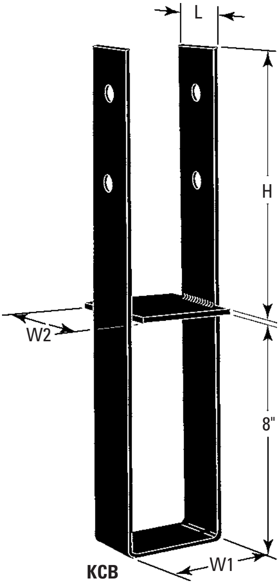

KCB

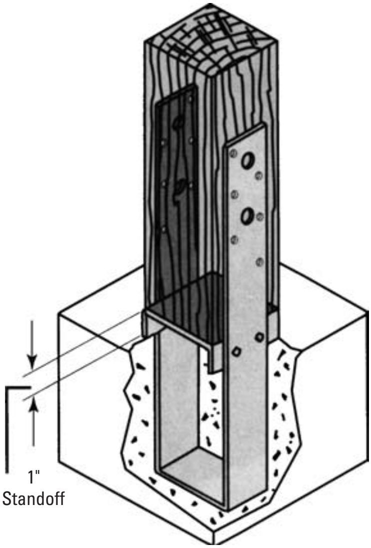

Typical **CBE** Installation

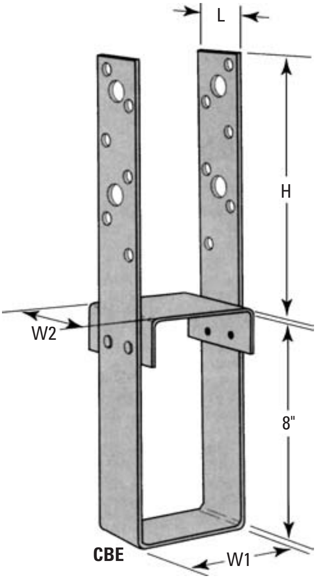

CBE

Post Caps—PCM and EPCM Series

												Allowable Loads (LBS.)					
					Dimensions				Fastener Schedule		Lateral A1		Lateral A2		Uplift	
USP Stock No.	**Ref. No.**	**Post (Inches)**	**Beam (Inches)**	**Steel Gauge**	**W1**	**W2**	**L1**	**L2**	**Post**	**Beam**	**133%**	**160%**	**133%**	**160%**	**133%**	**160%**
PCM4416	PC44-16	$3\frac{1}{2}$	$3\frac{1}{2}$	16	$3\frac{9}{16}$	$3\frac{9}{16}$	$3\frac{1}{16}$	11	(8)16d	(12)16d	1030	1030	835	835	1050	1050
PCM4516	—	$3\frac{1}{2}$	$5\frac{1}{4}$	16	$3\frac{9}{16}$	$5\frac{3}{8}$	$3\frac{1}{16}$	13	(8)16d	(12)16d	1030	1030	835	835	1050	1050
PCM4716	—	$3\frac{1}{2}$	7	16	$3\frac{9}{16}$	$7\frac{1}{8}$	$3\frac{1}{16}$	15	(8)16d	(12)16d	1030	1030	835	835	1050	1050
PCM5416	—	$5\frac{1}{4}$	$3\frac{1}{2}$	16	$5\frac{3}{8}$	$3\frac{9}{16}$	$3\frac{1}{2}$	11	(8)16d	(12)16d	1030	1030	835	835	1050	1050
PCM5516	—	$5\frac{1}{4}$	$5\frac{1}{4}$	16	$5\frac{3}{8}$	$5\frac{3}{8}$	$3\frac{1}{2}$	13	(8)16d	(12)16d	1030	1030	835	835	1050	1050
PCM44	PC44	$3\frac{1}{2}$	$3\frac{1}{2}$	12	$3\frac{9}{16}$	$3\frac{9}{16}$	$3\frac{1}{16}$	11	(8)16d	(12)16d	1930	1930	1495	1790	1050	1050
PCM45	—	$3\frac{1}{2}$	$5\frac{1}{4}$	12	$3\frac{9}{16}$	$5\frac{3}{8}$	$3\frac{1}{16}$	13	(8)16d	(12)16d	1930	1930	1495	1790	1050	1050
PCM47	—	$3\frac{1}{2}$	7	12	$3\frac{9}{16}$	$7\frac{1}{8}$	$3\frac{1}{16}$	15	(8)16d	(12)16d	1930	1930	1495	1790	1395	1395
PCM54	—	$5\frac{1}{4}$	$3\frac{1}{2}$	12	$5\frac{3}{8}$	$3\frac{9}{16}$	$3\frac{1}{2}$	11	(8)16d	(12)16d	1930	1930	1495	1790	1395	1395
PCM55	—	$5\frac{1}{4}$	$5\frac{1}{4}$	12	$5\frac{3}{8}$	$5\frac{3}{8}$	$3\frac{1}{2}$	13	(8)16d	(12)16d	1930	1930	1495	1790	1395	1395
PCM57	—	$5\frac{1}{4}$	7	12	$5\frac{3}{8}$	$7\frac{1}{8}$	$3\frac{1}{2}$	15	(8)16d	(12)16d	1930	1930	1495	1790	1395	1395
PCM75	—	7	$5\frac{1}{4}$	12	$7\frac{1}{8}$	$5\frac{3}{8}$	$5\frac{1}{2}$	13	(8)16d	(12)16d	1930	1930	1495	1790	1395	1395
PCM77	—	7	7	12	$7\frac{1}{8}$	$7\frac{1}{8}$	$5\frac{1}{2}$	15	(8)16d	(12)16d	1930	1930	1495	1790	1395	1395

(continued)

(continued)

											Allowable Loads (LBS.)					
					Dimensions				Fastener Schedule		Lateral A1		Lateral A2		Uplift	
USP Stock No.	Ref. No.	Post (Inches)	Beam (Inches)	Steel Gauge	W1	W2	L1	L2	Post	Beam	133%	160%	133%	160%	133%	160%
EPCM4416	EPC44-16	$3\frac{1}{2}$	$3\frac{1}{2}$	16	$3\frac{9}{16}$	$3\frac{9}{16}$	$3\frac{1}{16}$	$7\frac{1}{4}$	(8)16d	(8)16d	1020	1020	1055	1055	805	805
EPCM4516	—	$3\frac{1}{2}$	$5\frac{1}{4}$	16	$3\frac{9}{16}$	$5\frac{3}{8}$	$3\frac{1}{16}$	$9\frac{1}{4}$	(8)16d	(8)16d	1020	1020	1055	1055	805	805
EPCM4716	—	$3\frac{1}{2}$	7	16	$3\frac{9}{16}$	$7\frac{1}{8}$	$3\frac{1}{16}$	$11\frac{1}{4}$	(8)16d	(8)16d	1020	1020	1055	1055	805	805
EPCM5416	—	$5\frac{1}{4}$	$3\frac{1}{2}$	16	$5\frac{3}{8}$	$3\frac{9}{16}$	$3\frac{1}{2}$	$7\frac{1}{4}$	(8)16d	(8)16d	1020	1020	1055	1055	805	805
EPCM5516	—	$5\frac{1}{4}$	$5\frac{1}{4}$	16	$5\frac{3}{8}$	$5\frac{3}{8}$	$3\frac{1}{2}$	$9\frac{1}{4}$	(8)16d	(8)16d	1020	1020	1055	1055	805	805
EPCM44	EPC44	$3\frac{1}{2}$	$3\frac{1}{2}$	12	$3\frac{9}{16}$	$3\frac{9}{16}$	$3\frac{1}{16}$	$7\frac{1}{4}$	(8)16d	(8)16d	1180	1180	1495	1610	1480	1480
EPCM45	—	$3\frac{1}{2}$	$5\frac{1}{4}$	12	$3\frac{9}{16}$	$5\frac{3}{8}$	$3\frac{1}{16}$	$3\frac{1}{16}$	(8)16d	(8)16d	1180	1180	1495	1610	1480	1480
EPCM47	—	$3\frac{1}{2}$	7	12	$3\frac{9}{16}$	$7\frac{1}{8}$	$3\frac{1}{16}$	$11\frac{1}{4}$	(8)16d	(8)16d	1180	1180	1495	1610	1480	1480
EPCM54	—	$5\frac{1}{4}$	$3\frac{1}{2}$	12	$5\frac{3}{8}$	$3\frac{9}{16}$	$3\frac{1}{2}$	$7\frac{1}{4}$	(8)16d	(8)16d	1180	1180	1495	1610	1480	1480
EPCM55	—	$5\frac{1}{4}$	$5\frac{1}{4}$	12	$5\frac{3}{8}$	$5\frac{3}{8}$	$3\frac{1}{2}$	$9\frac{1}{4}$	(8)16d	(8)16d	1180	1180	1495	1610	1480	1480
EPCM57	—	$5\frac{1}{4}$	7	12	$5\frac{3}{8}$	$7\frac{1}{8}$	$3\frac{1}{2}$	$11\frac{1}{4}$	(8)16d	(8)16d	1180	1180	1495	1610	1480	1480
EPCM75	—	7	$5\frac{1}{4}$	12	$7\frac{1}{8}$	$5\frac{3}{8}$	$5\frac{1}{2}$	$11\frac{1}{4}$	(8)16d	(8)16d	1180	1180	1495	1610	1480	1480
EPCM77	—	7	7	12	$7\frac{1}{8}$	$7\frac{1}{8}$	$5\frac{1}{2}$	$13\frac{1}{4}$	(8)16d	(8)16d	1180	1180	1495	1610	1480	1480

Uplift loads and lateral loads have been increased $33\frac{1}{3}$% or 60% for wind and seismic loads. No further increase shall be permitted.
Allowable loads are for nails only.

Post Caps PCM and EPCM Series

Post caps provide a positive connection for medium duty post-to-beam applications. The extended side plate design in these products also function as tie straps where splices occur. These models fasten with standard nails.

Finish	G60 galvanizing
Code Listing	NER 505

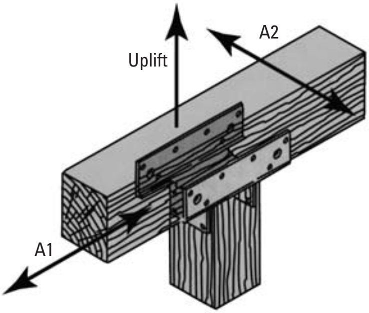

Typical **PCM** Installation

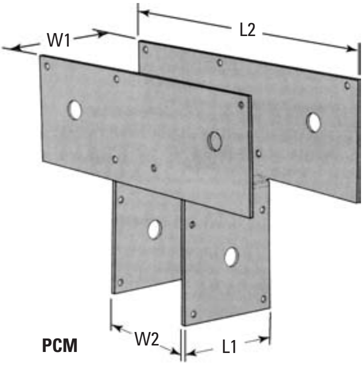

PCM

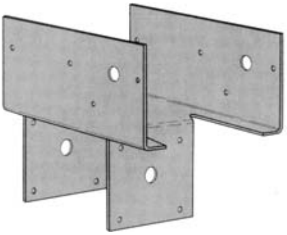

EPCM

Column Caps—KCC and KECC Series

			Dimensions					*Bolt Schedule*							
						L				*Bearing Load*		*Uplift KCC*		*Uplift KECC*	
USP Stock No.	*Ref. No.*	*Steel Gauge*	*W1*	*W2*	*H*	*KCC*	*KECC*	*Beam*	*Column or Post*	*KCC*	*KECC*	*133%*	*160%*	*133%*	*160%*
KCC44	CC44	7	$3^{5}/_{8}$	$3^{5}/_{8}$	4	7	$5^{1}/_{2}$	$(2)^{5}/_{8}$	$(2)^{5}/_{8}$	15315	12030	2985	3585	1635	1960
KCC45	—	7	$3^{5}/_{8}$	$5^{3}/_{8}$	$6^{1}/_{2}$	11	$8^{1}/_{2}$	$(4)^{5}/_{8}$	$(2)^{5}/_{8}$	24065	18595	5270	6310	3210	3850
KCC47	—	7	$3^{5}/_{8}$	$7^{1}/_{8}$	$6^{1}/_{2}$	11	$9^{1}/_{2}$	$(4)^{5}/_{8}$	$(2)^{5}/_{8}$	24065	20780	5270	6310	3210	3850
KCC54	—	7	$5^{3}/_{8}$	$3^{5}/_{8}$	$6^{1}/_{2}$	11	$7^{1}/_{2}$	$(4)^{5}/_{8}$	$(2)^{5}/_{8}$	37810	25780	5255	6310	3385	4060
KCC55	—	7	$5^{3}/_{8}$	$5^{3}/_{8}$	$6^{1}/_{2}$	11	$7^{1}/_{2}$	$(4)^{5}/_{8}$	$(2)^{5}/_{8}$	37815	25780	5255	6310	3385	4060
KCC57	CC6 - $7^{1}/_{8}$	7	$5^{3}/_{8}$	$7^{1}/_{8}$	$6^{1}/_{2}$	11	$9^{1}/_{2}$	$(4)^{5}/_{8}$	$(2)^{5}/_{8}$	37815	32655	5255	6310	3385	4060
KCC75X	CC$7^{1}/_{8}$-6	3	$7^{1}/_{8}$	$5^{3}/_{8}$	8	13	$10^{1}/_{2}$	$(4)^{3}/_{4}$	$(2)^{3}/_{4}$	54600	44100	6230	7255	4665	5595
KCC77X	CC7 1/8 - $7^{1}/_{8}$	3	$7^{1}/_{8}$	$7^{1}/_{8}$	8	13	$10^{1}/_{2}$	$(4)^{3}/_{4}$	$(2)^{3}/_{4}$	54600	44100	6230	7255	4665	5595

Bearing loads are based on 560 psi perpendicular to grain loading for glulam sizes and 8× and wider, and 625 psi for all others.
Uplift loads are increased 33% and 60% for wind or seismic loads; no further increase is permitted.
Beams shall be designed to support the required loads. Beam shear may limit loads to less than listed loads for device.
Posts are assumed adequate to support listed loads.
Bearing loads shall not be increased for duration of load.

Column Caps KCC and KECC Series

Designed for heavy duty beam-to-post connections, column caps offer strong reinforcement and support at critical framing junctions. A number of special sizes and style variations are available. Straps are normally centered on the U-bracket, but any variation may be specified. For end column caps, add an “E” after the “K” in the part number, as in “KECC325-4.” Straps may be rotated 90 degrees on special order where the W2 dimension is less than or equal to the W1 dimension. Welding is performed by certified welders and routinely inspected by a code-accepted quality control agency.

Finish . . . . paint
Code Listings . . . . ICBO 2725
. . . . L.A. City RR25104

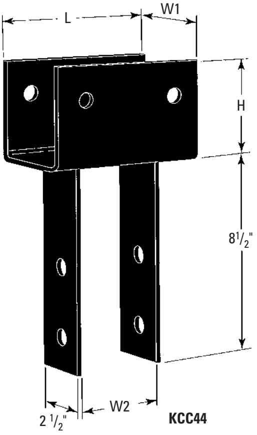
L
W1
H
8½"
2 ½"
W2

KCC44

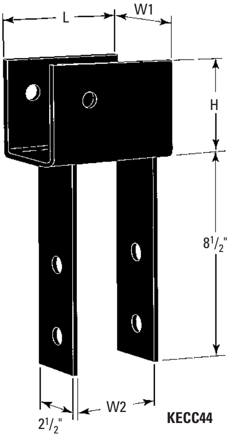
L
W1
H
8½"
2½"
W2

KECC44

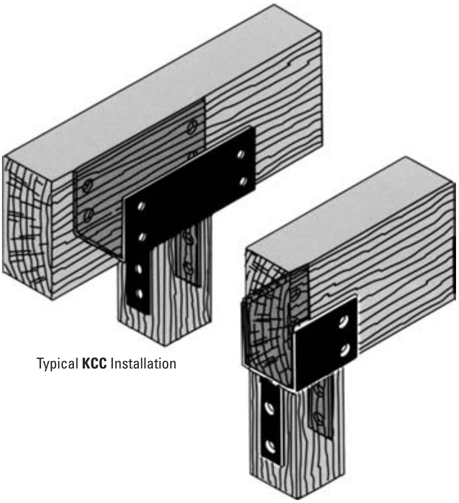

Typical **KCC** Installation

Typical **KECC** Installation

KCC and KECC Series Options

A number of size and style variations are available for KCC and KECC Column Caps on a custom order basis. Possible variations include rough or custom sizes and altered bolt pattern. A number of standard options are also available as detailed by the adjacent illustrations. These options include: KCCO, KCCT, KCCOB and 90° strap rotation.

NOTE

Consult retail price list or your technical assistance representative for details and ordering information.

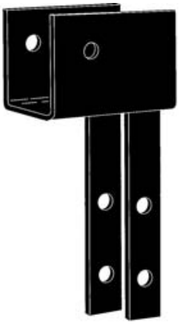

KECC Straps Rotated 90°

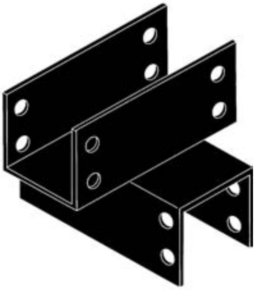

KCCOB

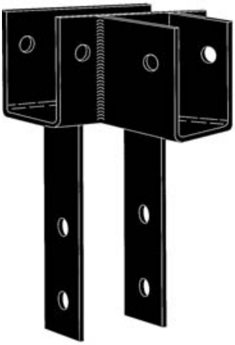

KCCT

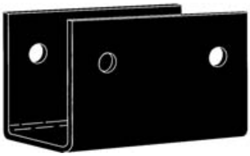

KCCO

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

This Course Consists Of:

- Compass Basics
- Orienting a Map For North Reference
- Traveling to a Target
- Backstops, Handrails and Aiming Off
- Topo Map Basics
- Pace Counting
- Finding Yourself
- UTM Coordinate System
- Using GPS with a Map
- Understanding And Adjusting For Declination

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Compass Types

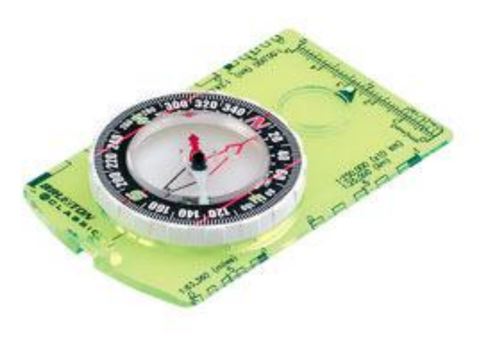

Baseplate

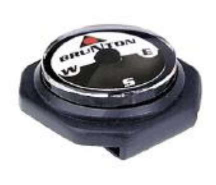

Accessory

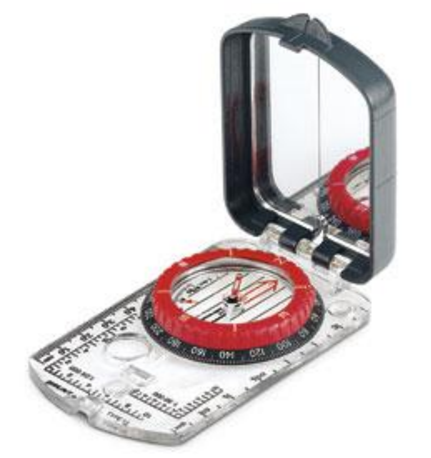

Mirrored – Sighting (Preferred)

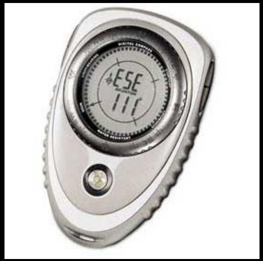

Digital

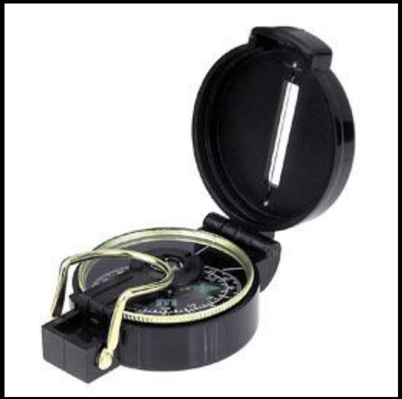

Lensatic

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

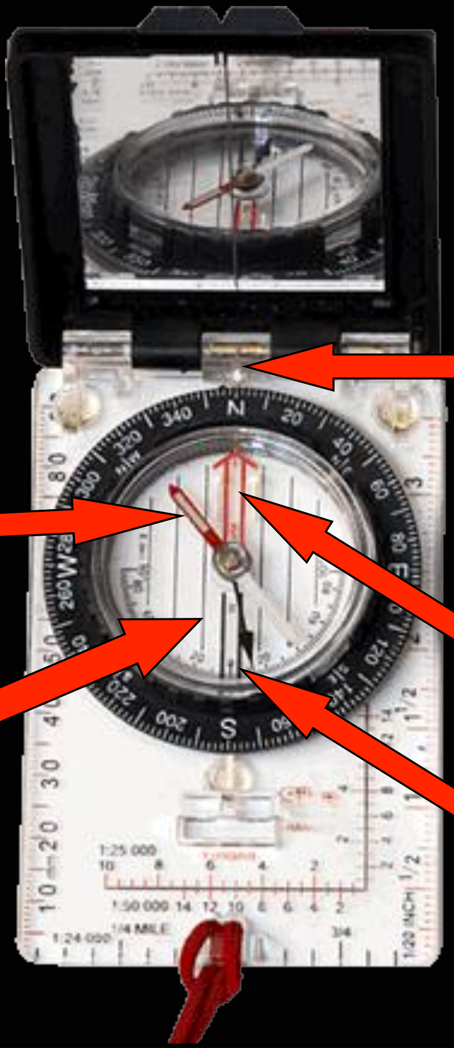

Bearing Index (Read Bearing Here)

Compass Needle – Red Points To Magnetic North

Orienting Arrow

Orienting Lines

Declination Scale

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Boxing A Compass

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

MAP ORIENTATION
With Compass Adjusted For Declination

1) Adjust dial so "N" is at index line.

2) Align edge of compass with north / south reference line.

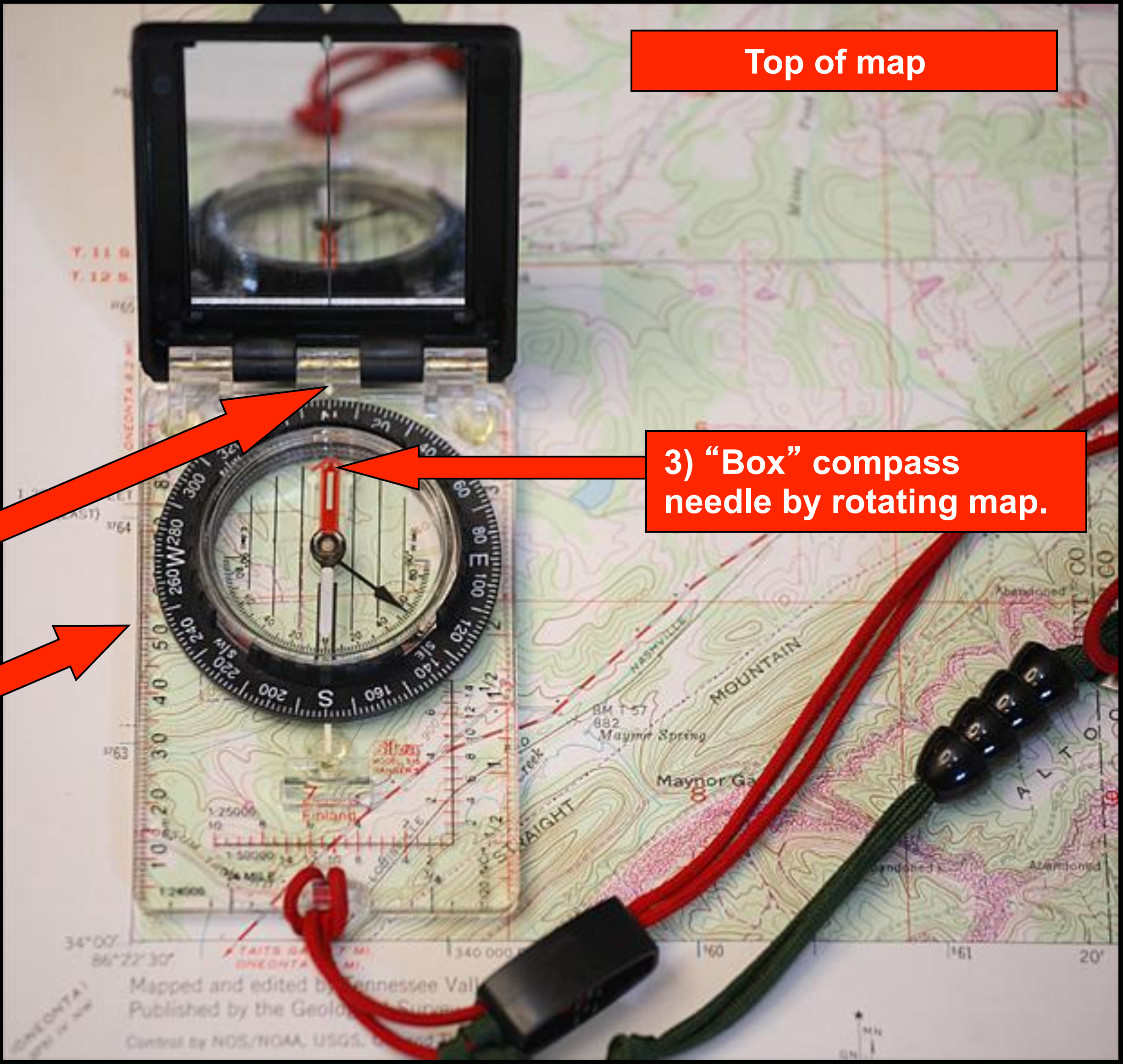

Top of map

3) "Box" compass needle by rotating map.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Finding An Azimuth With Compass

(With Map Oriented To North)

Traveling To a Target On The Map – Step 1

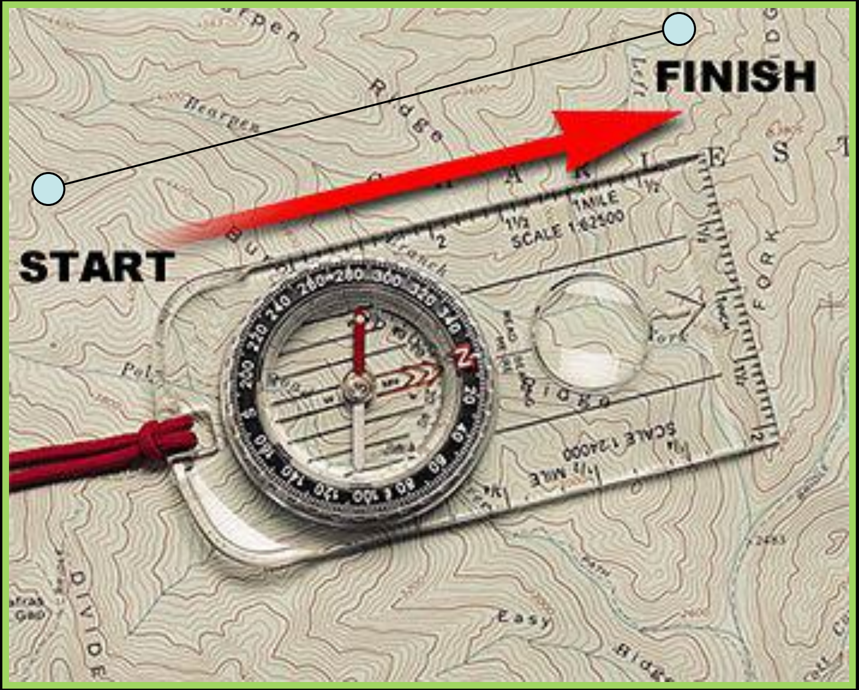

1) Draw a line on the map from your starting point to your finish point.
2) Orient the map to north.
3) Place the compass on the map with the edge of the compass on the line and the bearing arrow pointing at your destination

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Traveling To a Target On The Map – Step 2

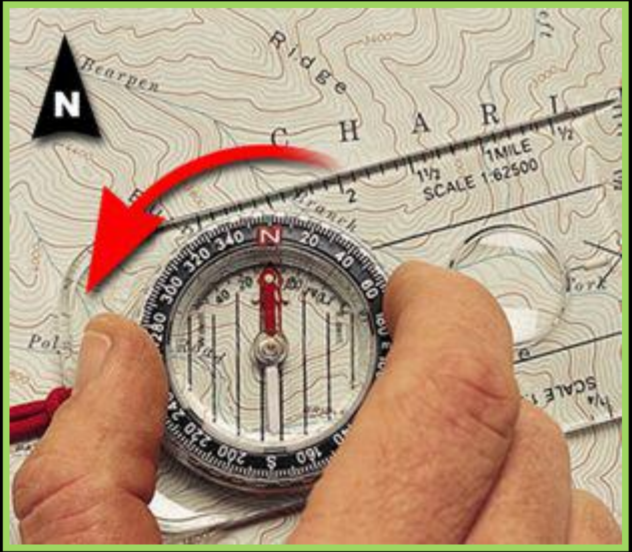

Set Compass Heading - Turn the dial on the compass until "N" outline arrow boxes the compass needle. Your direction to your target (in degrees) is read at the Index Line on the Dial.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Traveling To a Target On The Map – Step 3

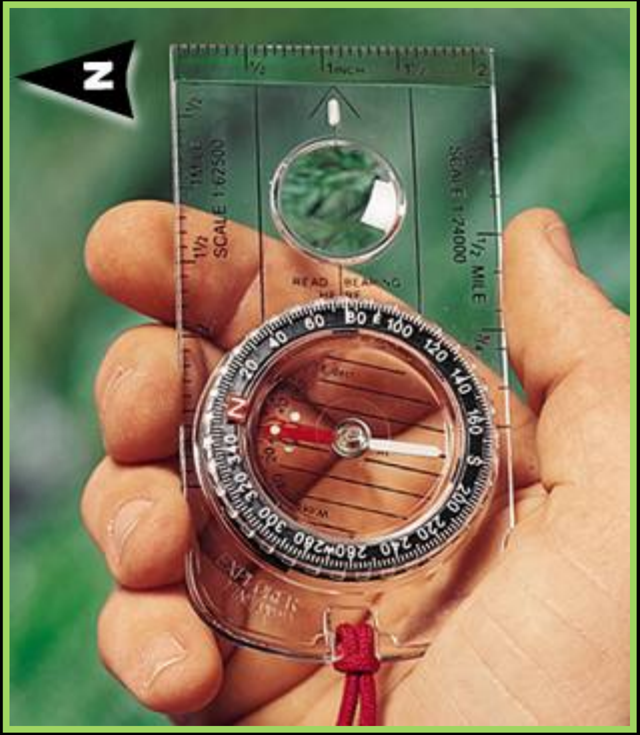

Follow Your Heading - Remove the compass from the map and hold it level, so the Magnetic Needle is free to turn. Turn your body until the red end of the Needle aligns with the Orienting Arrow (boxed) and "N" on the Dial. Using the Direction of Travel Arrow, sight a distant landmark and move to it. Repeat this process until you reach your destination.

Finding An Azimuth With Compass

(Without Map Oriented To North)

- Draw a line from start to finish.
- Lay the edge of your compass on the line in the direction you are traveling.
- Rotate the compass bezel until the orienting lines align with your North Reference Lines (Make sure that North is pointing to North on the map).
- Pick up your compass and travel to the destination.
- Note: The compass shown has already been adjusted for declination.

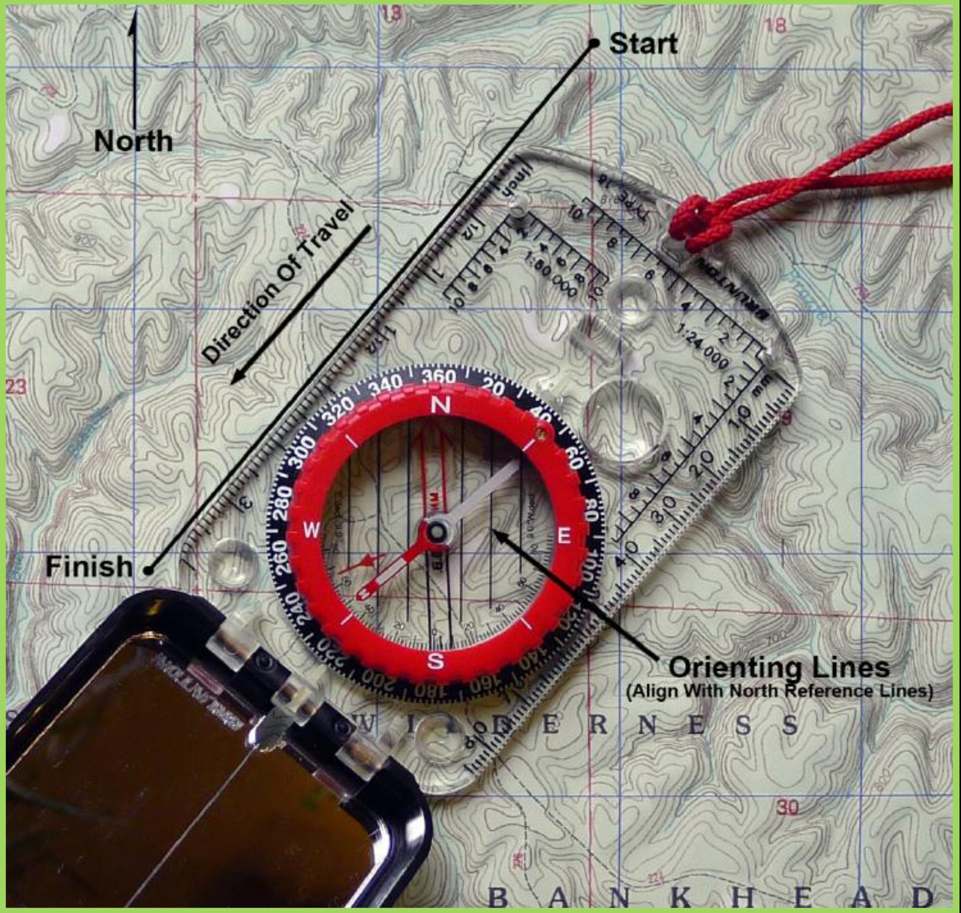
Start
North
Direction Of Travel
Finish
Orienting Lines
(Align With North Reference Lines)
WILDERNESS
BANKHEAD

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Finding An Azimuth With Compass Card

(Without Map Oriented To North)

- Draw a line from start to finish.
- Lay the compass rose on the line so the line crosses through the center of the rose.
- Align N / S line or orienting lines on the compass rose with North Reference Line.
- Read your azimuth on the “Finish” side of the compass rose.
- Note: unless you are using a compass adjusted for declination, you will have to add or subtract the declination value to the value that you acquire from the compass rose.

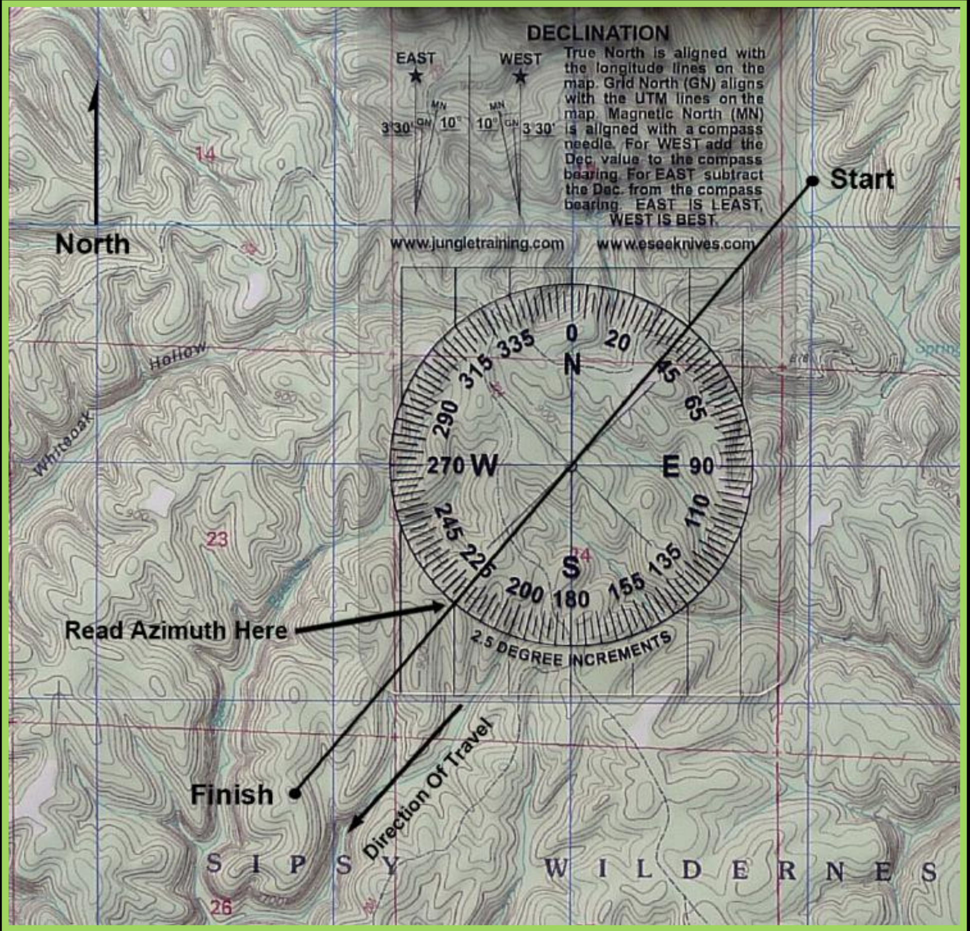
DECLINATION
EAST
WEST
MN
GN
3°30'
10°
10°
3°30'
True North is aligned with the longitude lines on the map. Grid North (GN) aligns with the UTM lines on the map. Magnetic North (MN) is aligned with a compass needle. For WEST add the Dec. value to the compass bearing. For EAST subtract the Dec. from the compass bearing. EAST IS LEAST, WEST IS BEST.
www.jungletraining.com
www.eseeknives.com
North
Start
0
20
45
65
N
335
315
290
270 W
E 90
110
245
225
S
200
180
155
135
2.5 DEGREE INCREMENTS
Read Azimuth Here
Finish
Direction Of Travel
Hollow
Whiteoak
14
23
24
26
S I P S Y W I L D E R N E S S

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Traveling To a Target With Sighting Compass

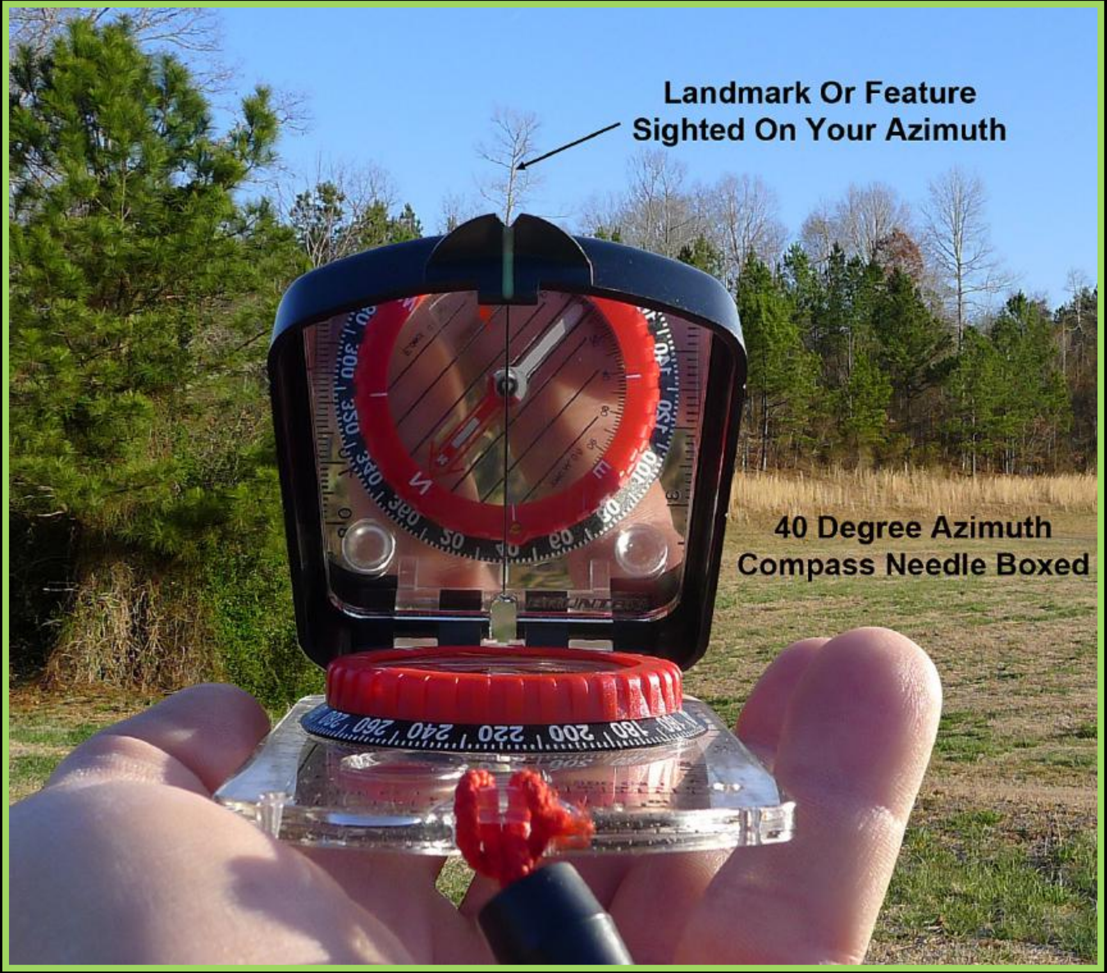

Landmark Or Feature Sighted On Your Azimuth
40 Degree Azimuth Compass Needle Boxed

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Posting Up

In addition to using natural landmarks as targets for your direction of travel, you can also post up a member of your team as a target. Another option is the Back Azimuth.

Ft. Payne, AL
3.5 Degrees West

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Back Azimuth

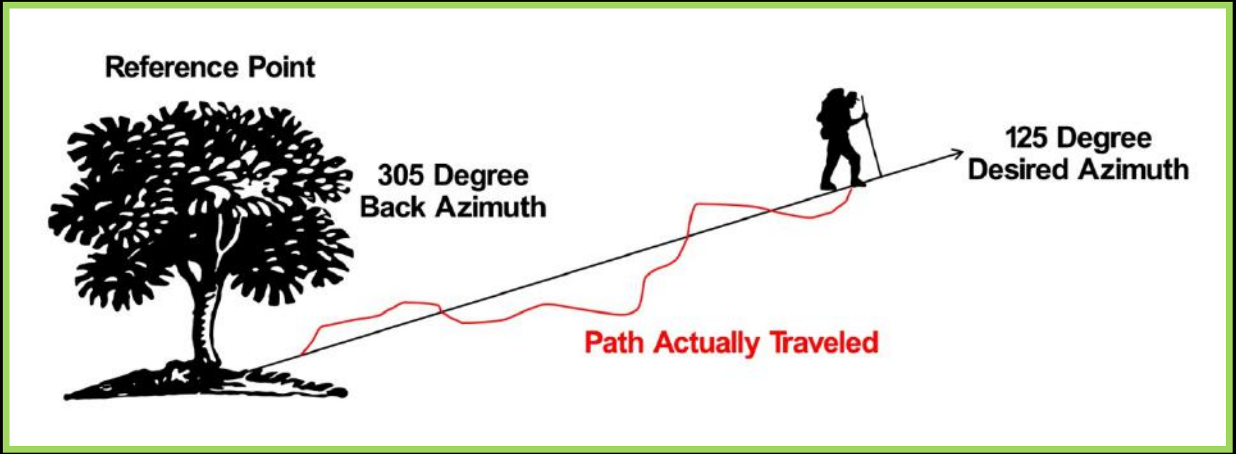
Reference Point
305 Degree Back Azimuth
125 Degree Desired Azimuth
Path Actually Traveled

Use a Back Azimuth if you need to verify you are still on your desired direction of travel. This is especially handy when crossing rivers or boxing around objects.

Back Azimuth is 180 degrees from your target azimuth. If your target azimuth is more than 180 degrees, then subtract 180. If it is less than 180, then add 180.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Boxing Around Objects

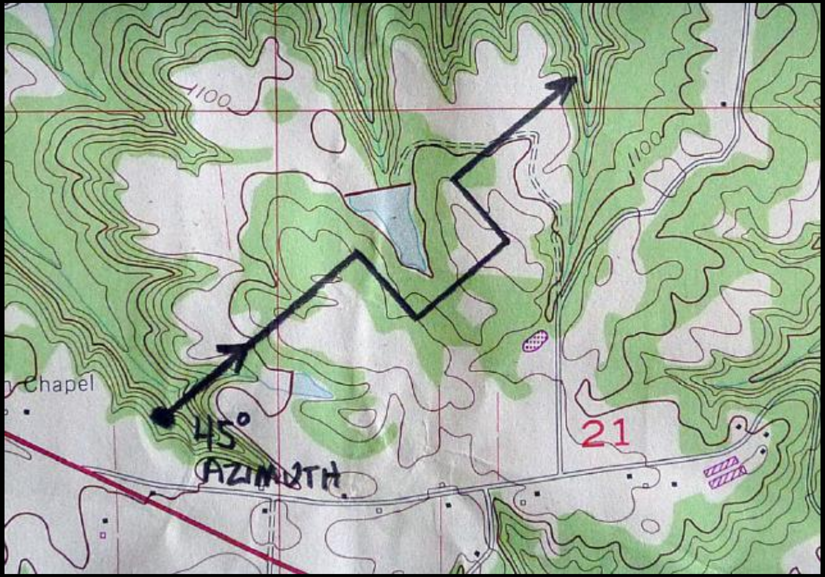

- Direction of original travel is 45 degrees
- Turn 90 degrees to a new azimuth of 135 (count your paces)
- Turn 90 degrees back to the original azimuth of 45 degrees until you are past obstacle
- Turn 90 degrees to a new azimuth of 315 and pace the same amount as the first turn
- Turn 90 degrees back to the original azimuth of 45 degrees and continue on

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

LATERAL DRIFT – MISSING YOUR TARGET

If you cannot find your target after walking your azimuth the correct distance, then mark the spot where you think it should be and began a structured grid search of the area. Make note of your azimuths, pace count and direction changes. Start your grid small then go larger if you cannot locate the target.

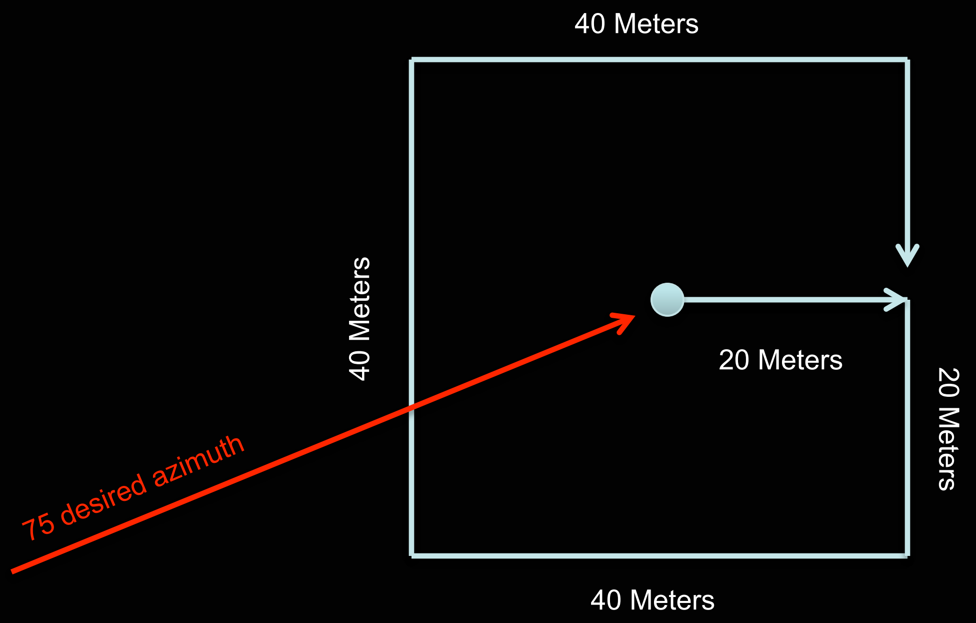
40 Meters
40 Meters
20 Meters
20 Meters
75 desired azimuth
40 Meters

Ft. Payne, AL
3.5 Degrees West

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Gross Direction Navigation

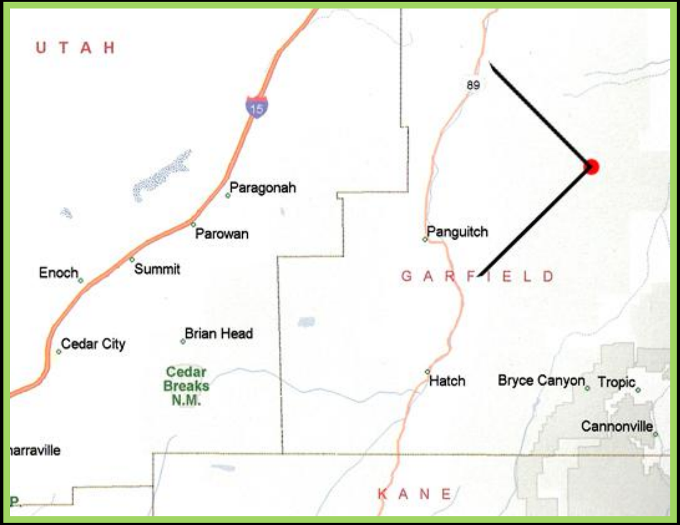
UTAH
15
89
Paragonah
Parowan
Panguitch
Enoch
Summit
GARFIELD
Brian Head
Cedar City
Cedar Breaks N.M.
Hatch
Bryce Canyon
Tropic
Cannonville
arraville
P.
KANE

As obvious from this simple road map, the lost hiker only needs to travel a westerly direction to cross a road and reach civilization. Always know which general direction crosses a main artery!

When using a map and compass always have a “panic azimuth.”

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Gross Direction Navigation Aids

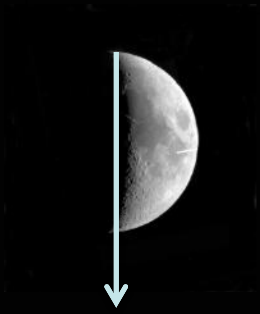

Quarter Moon

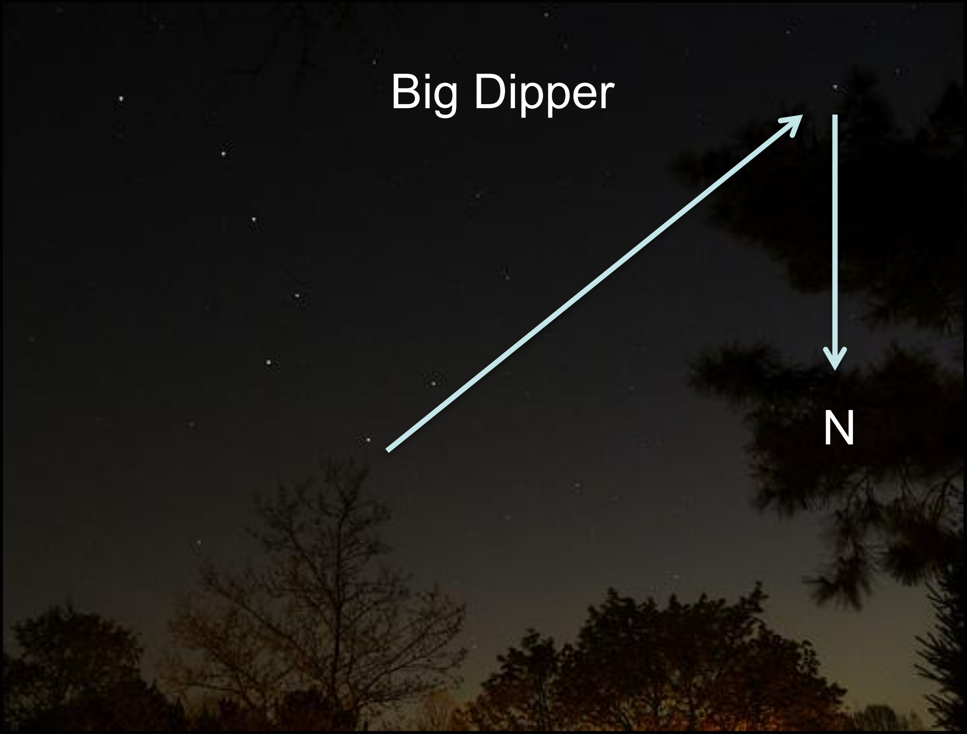

Big Dipper

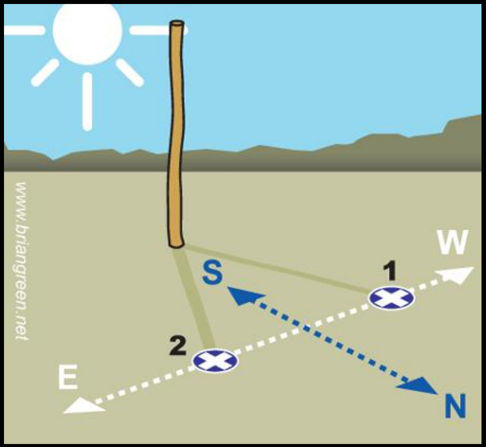
www.briangreen.net
W
S
1
2
E
N

Shadow Stick

Terrain Association Navigation

This is the most widely used method of navigation. The navigator plans his route so that he moves from terrain feature to terrain feature. An automobile driver in a city uses this technique as he moves along a street or series of streets, guiding on intersections or features such as stores, parks or houses. Like the driver, the navigator selects routes or *streets* between key points or *intersections*. These key points can be lakes, mountains, roads or any other terrain feature readily recognized on a map. Before you hike a new area, familiarize yourself with the terrain by studying maps, satellite photos or any other available information.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Backstops, Handrails & Aiming Off

BACKSTOPS are features that typically run perpendicular to your azimuth and located beyond your target point. If you run into your backstop then you know you have traveled too far and missed your target point. Backstops can be rivers, roads, mountains, railroad tracks or any other useable feature.

HANDRAILS are features that run parallel to your azimuth. They can be rivers, roads, mountains, or any other useable feature that the navigator can use to continue traveling in the desired direction.

AIMING OFF is a process of deliberately adding or subtracting to the desired azimuth so the navigator knows which way to go once they reach a feature (such as a trail or road) that leads to the target.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Backstops, Handrails & Aiming Off

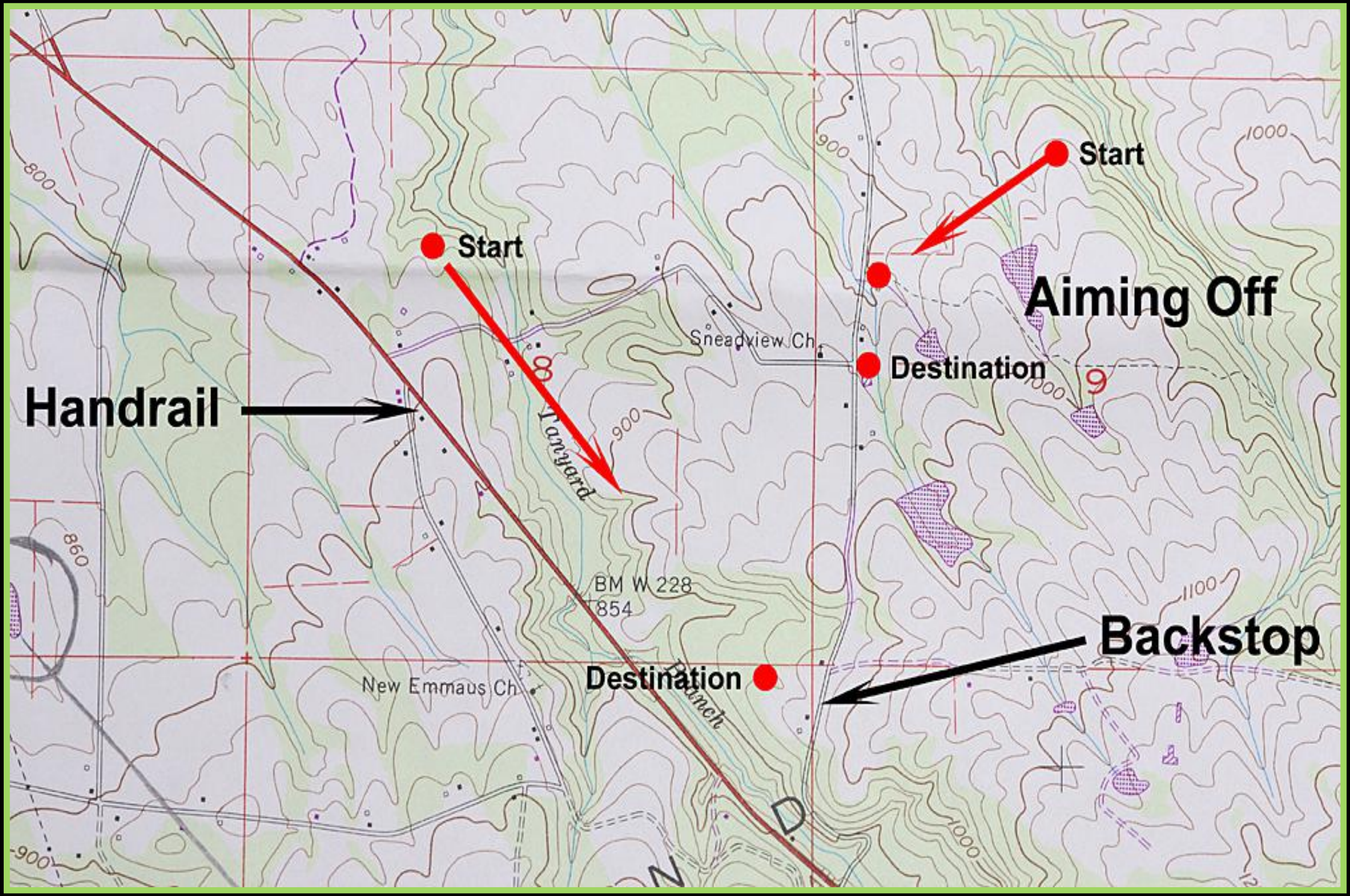
Start
Start
Aiming Off
Destination
Handrail
Destination
Backstop
Sneadview Ch
New Emmaus Ch
BM W 228
854
Tanyard
Branch
800
900
1000
1100
860

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

TOPO MAP BASICS

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

BASICS SYMBOLS

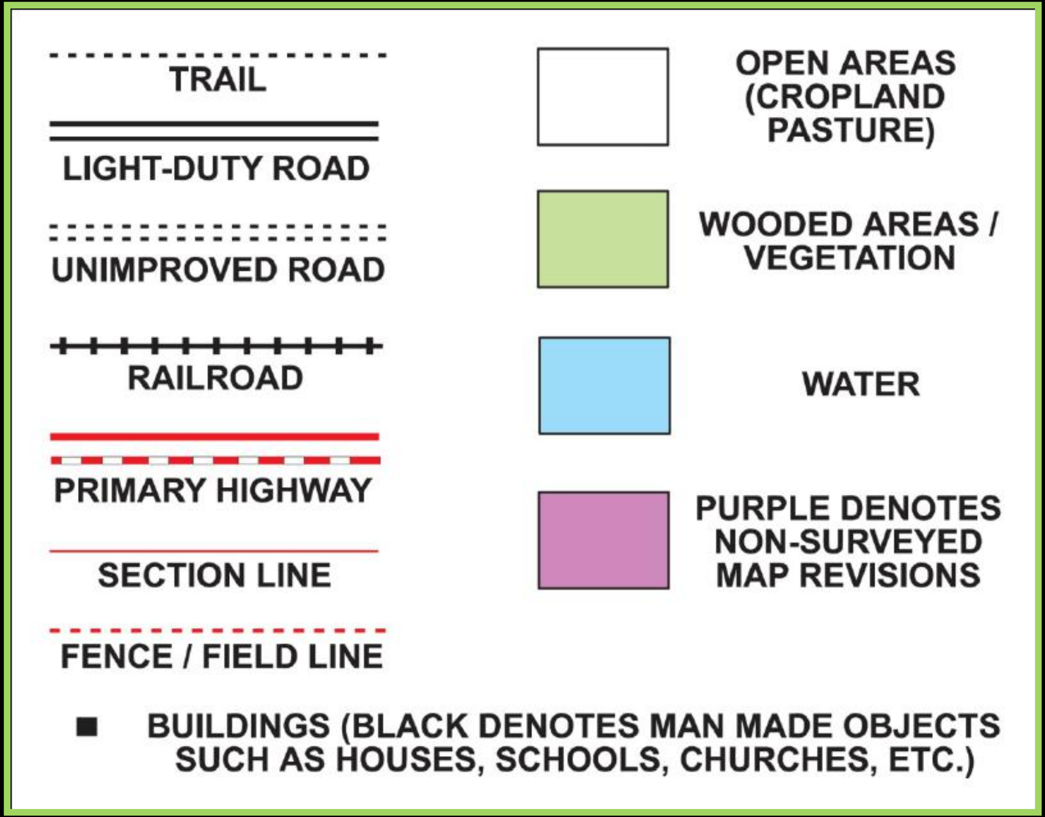

TRAIL

LIGHT-DUTY ROAD

UNIMPROVED ROAD

RAILROAD

PRIMARY HIGHWAY

SECTION LINE

FENCE / FIELD LINE

OPEN AREAS (CROPLAND PASTURE)

WOODED AREAS / VEGETATION

WATER

PURPLE DENOTES NON-SURVEYED MAP REVISIONS

■ BUILDINGS (BLACK DENOTES MAN MADE OBJECTS SUCH AS HOUSES, SCHOOLS, CHURCHES, ETC.)

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Map Scale & Datum

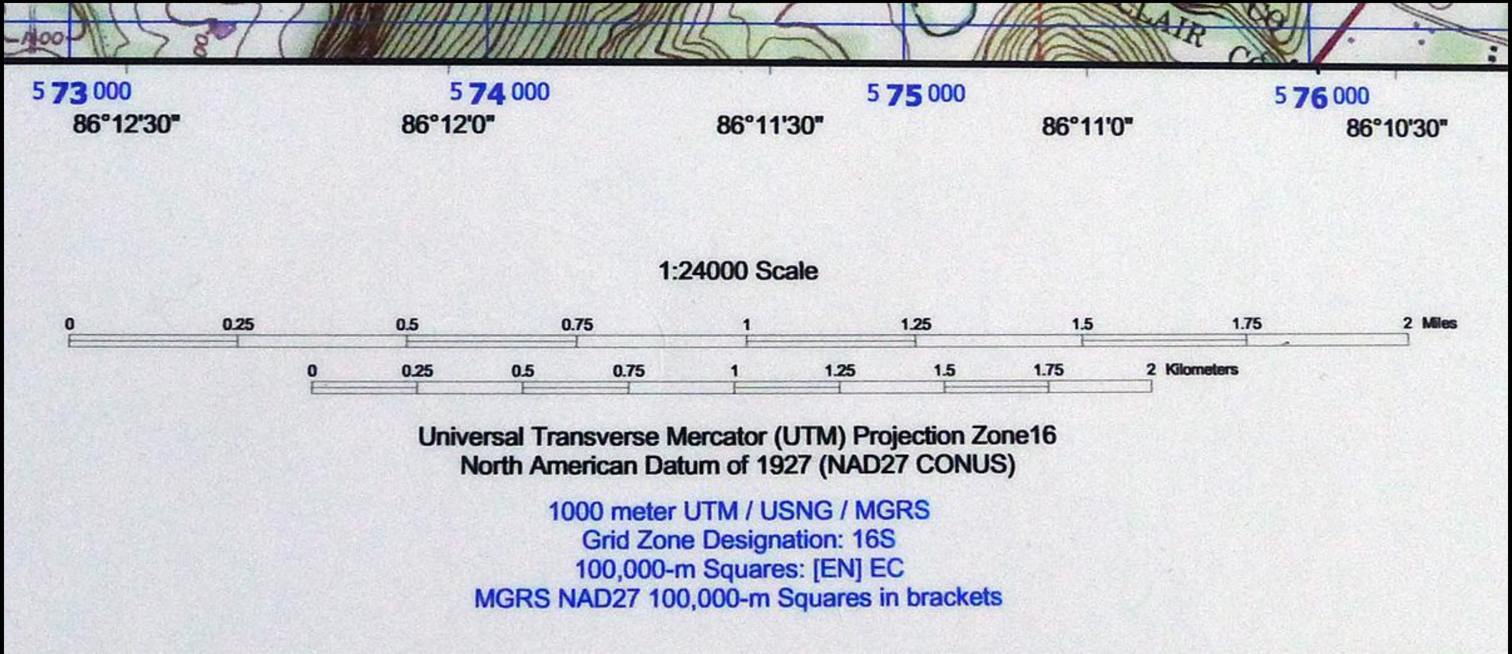

5 73 000 86°12'30"
5 74 000 86°12'0"
86°11'30"
5 75 000 86°11'0"
5 76 000 86°10'30"

1:24000 Scale

0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 Miles

0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 Kilometers

Universal Transverse Mercator (UTM) Projection Zone16
North American Datum of 1927 (NAD27 CONUS)

1000 meter UTM / USNG / MGRS
Grid Zone Designation: 16S
100,000-m Squares: [EN] EC
MGRS NAD27 100,000-m Squares in brackets

Always set your GPS to the map datum shown on your map!
Always use the proper scale card for measuring!

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

CONTOUR LINES

The closer the contour lines, the steeper the terrain. Check the map for the contour intervals. Contour lines will also have elevation markings.

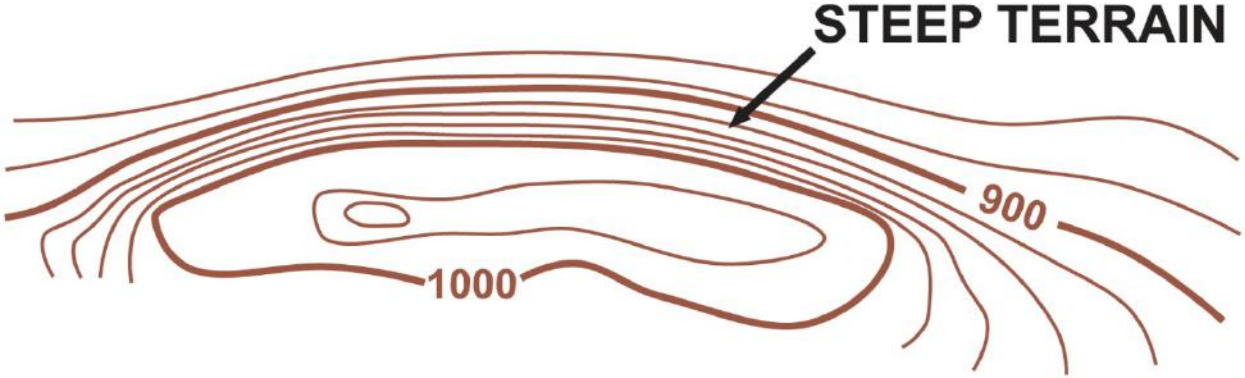

Example above shows a map with contour intervals of 20 feet.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

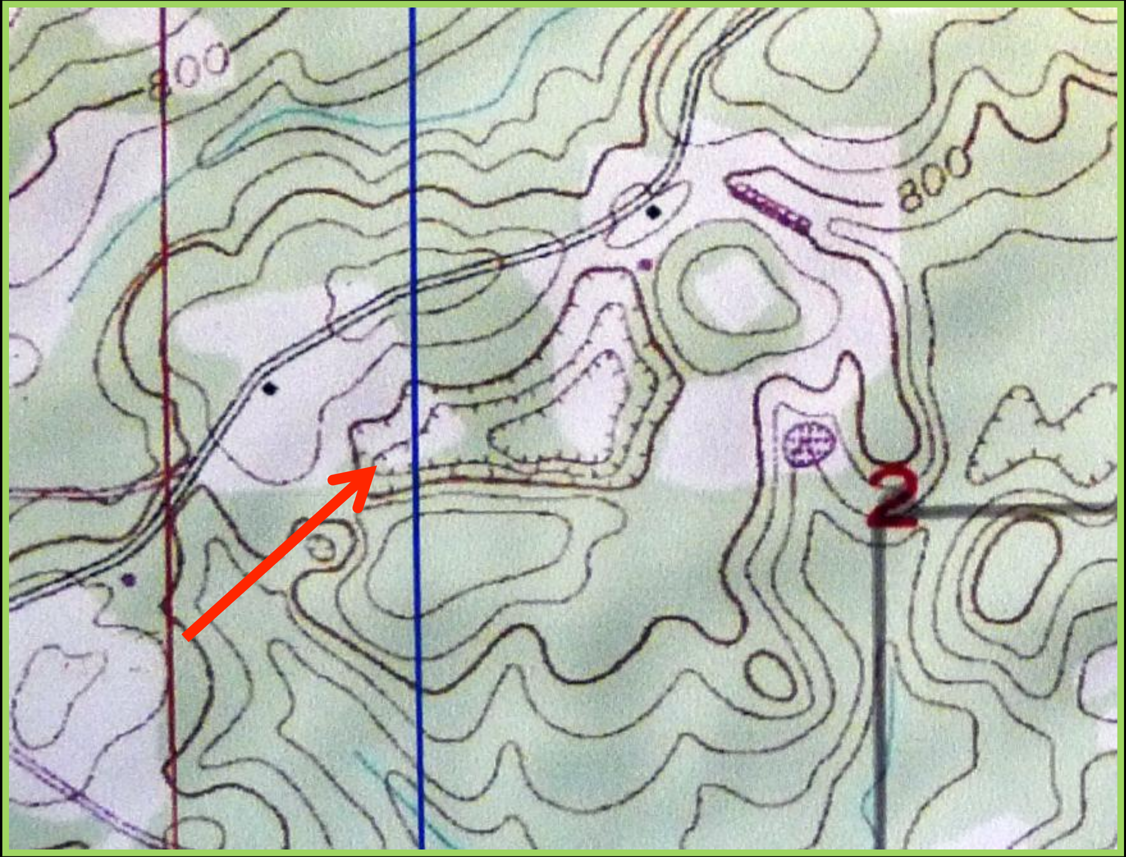

Depression

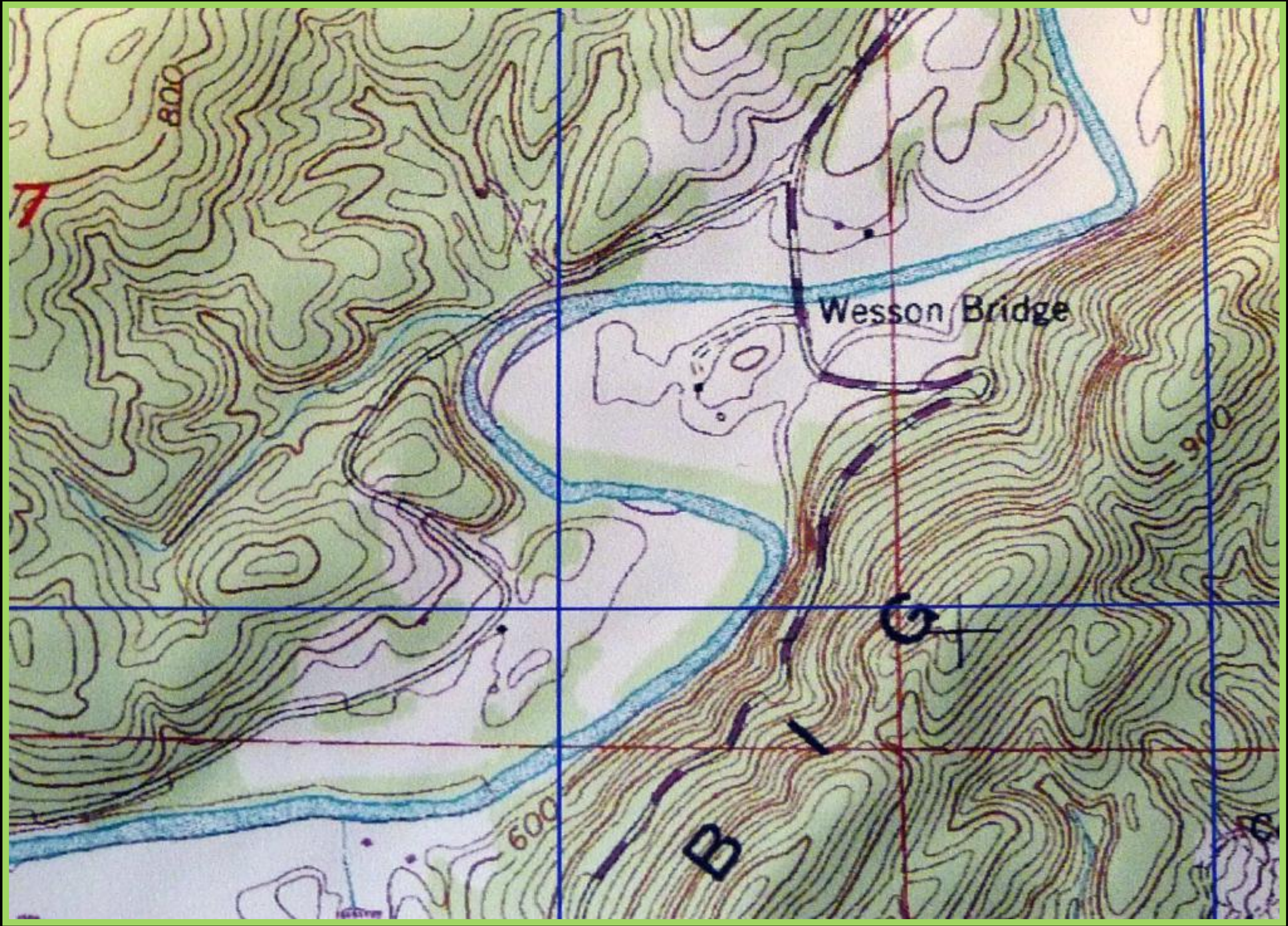
800
Wesson Bridge
900
600
B
I
G

Continuous Water Flow

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

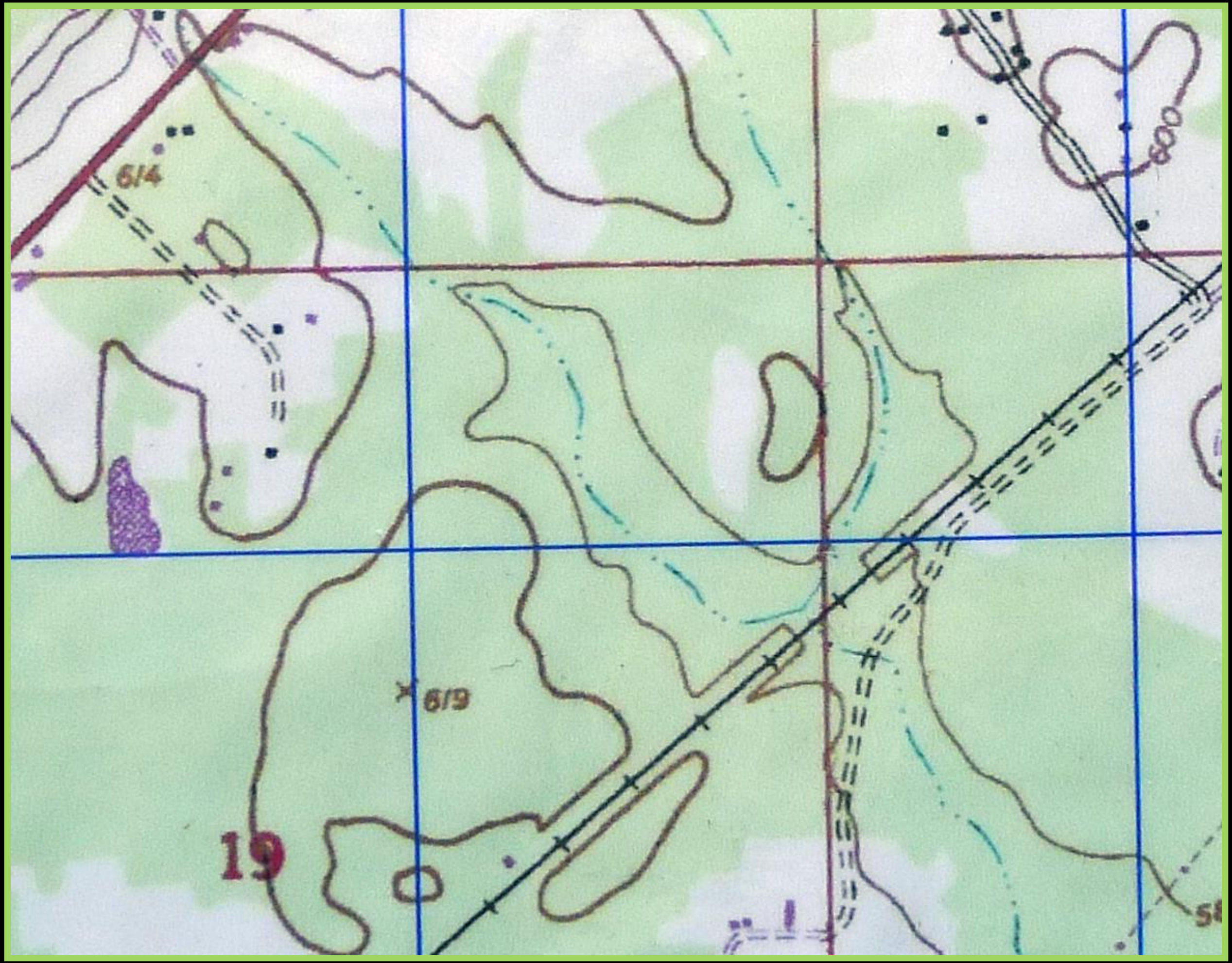

Intermittent Water Flow

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

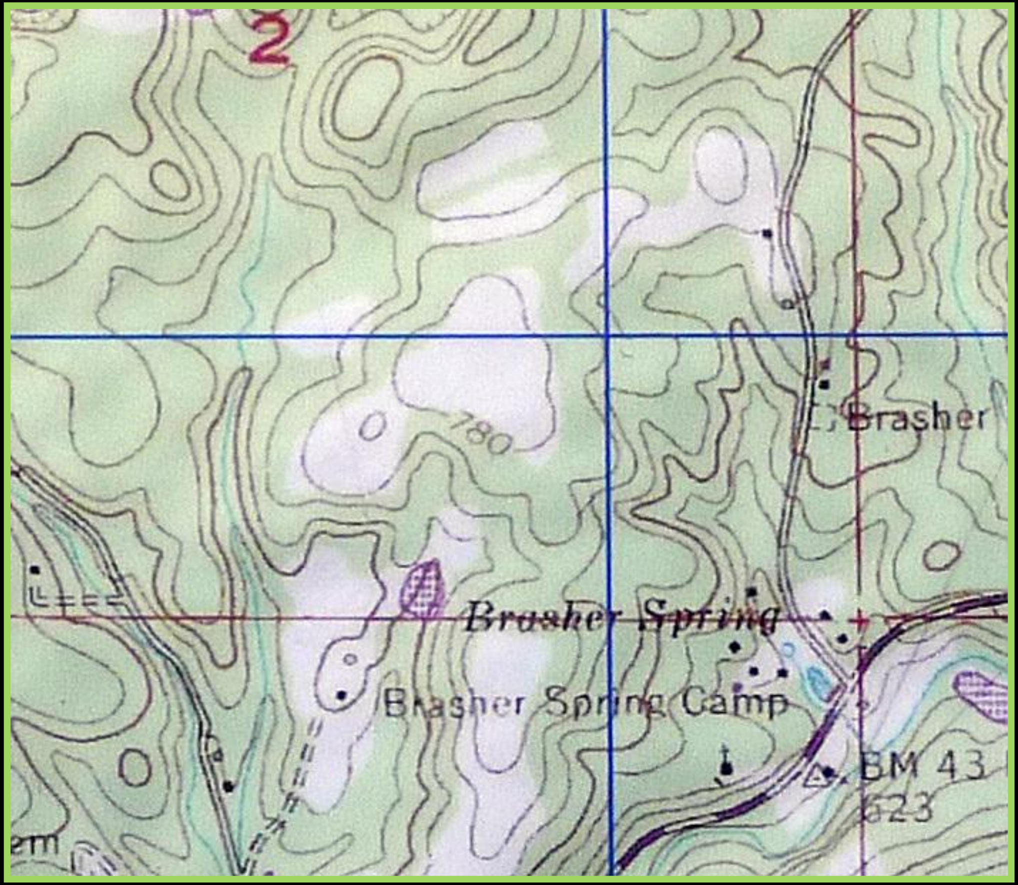
2
780
Brasher
Brasher Spring
Brasher Spring Camp
BM 43
623
em

Clear Areas / Fields

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

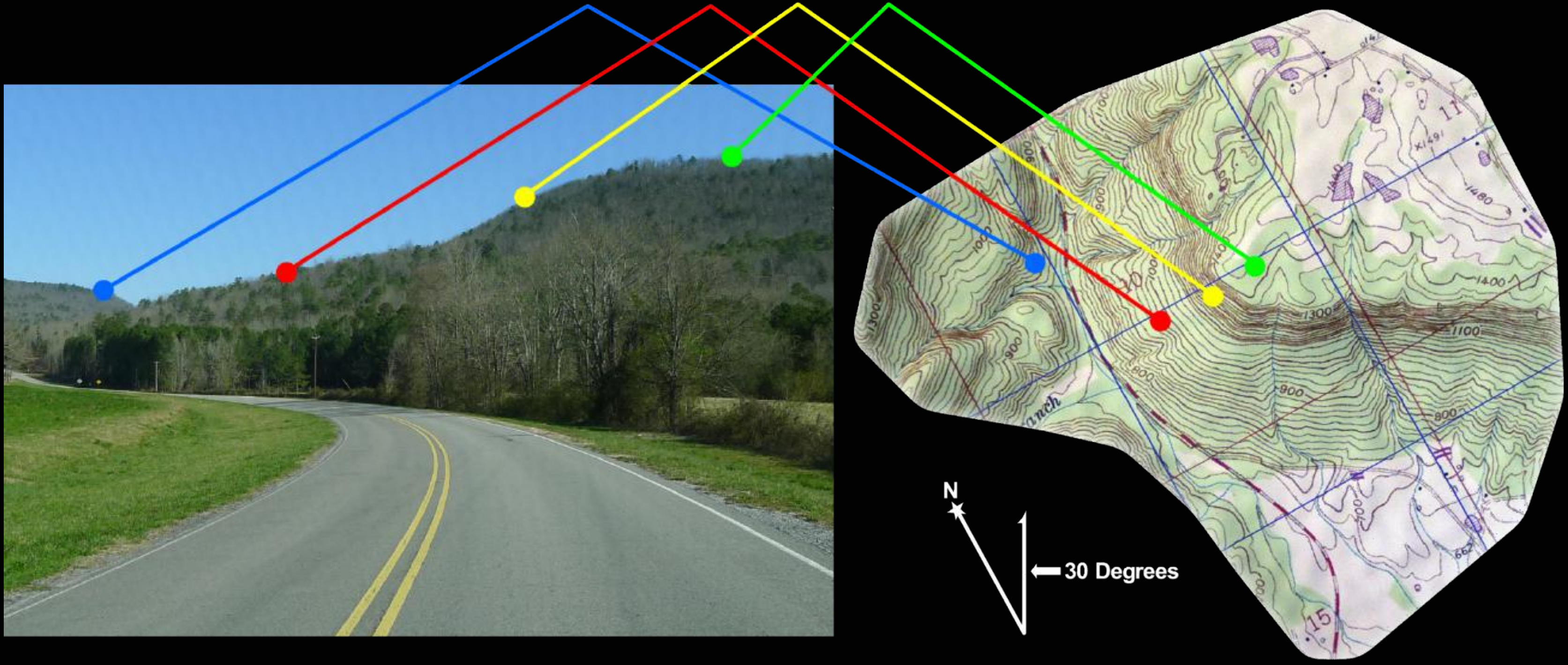

Actual View

Topo View

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

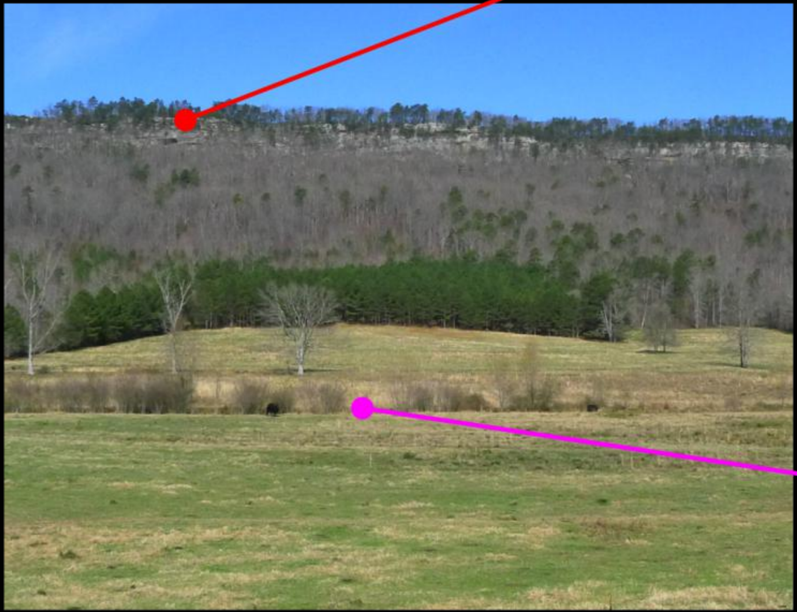

Actual View

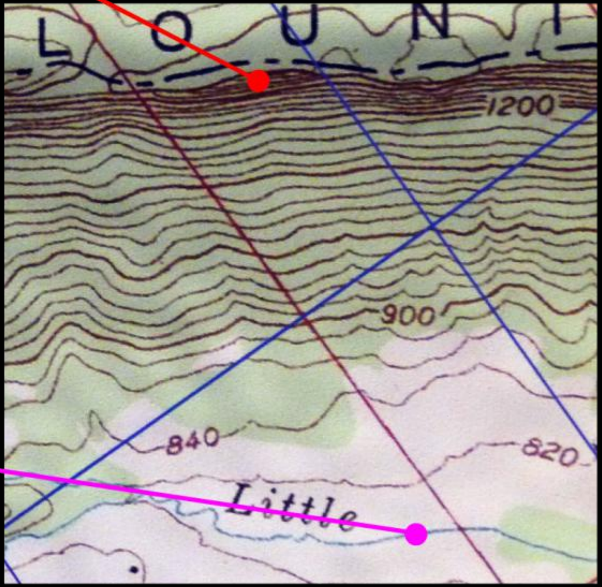
L O U N
1200
900
840
820
Little

Topo View

300 Degrees ➡

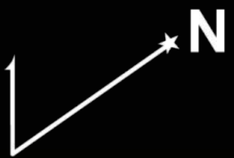

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

DETERMINING DISTANCE

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Determining Distance On Map

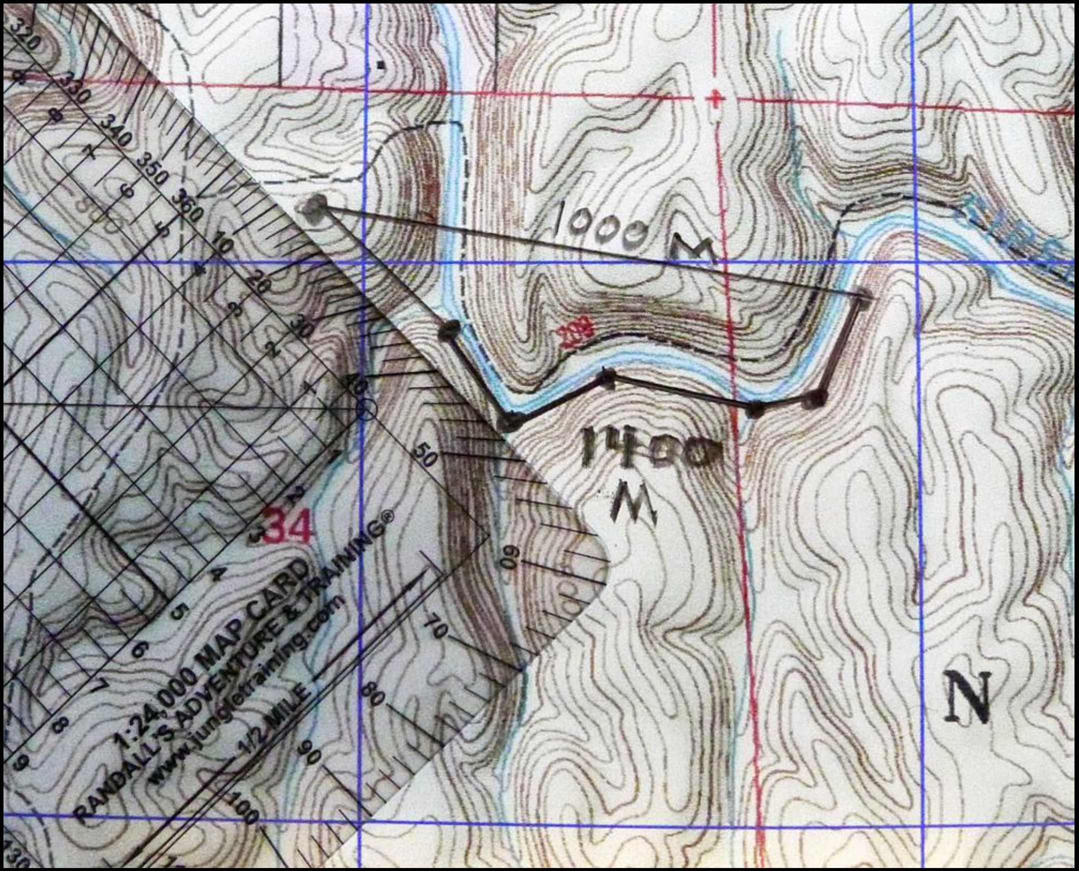
1000 M
1400 M
N
34
1:24,000 MAP CARD
RANDALL'S ADVENTURE & TRAINING®
www.jungletraining.com
1/2 MILE

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

DETERMINE DISTANCE TRAVELED BY PACE COUNT

In thick jungle, where landmarks can not always be seen to track your position, pace counting is the best way of measuring distance. Pace counting will allow the navigator to estimate where he is at any given time. To be accurate, the navigator must practice pacing over different types of terrain. First you have to do some calculations. Measure out exactly 100 meters on three types of ground. Flat easy terrain, rougher terrain with some slope and then steep hill terrain. Then on each measured course count your paces (every time your left foot touches the ground or every 2 steps = 1 pace). You will have 3 different pace counts for different types of terrain. Once finished, memorize your pace count for all 3 types.

Averages

Flat easy terrain	100 meters	65 paces
Rougher terrain with some slope	100 meters	75 paces
Steep hill terrain	100 meters	95 paces

General Rule: You have never traveled as far as you think you have.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

CONDITIONS THAT AFFECT PACE COUNT

- Slopes – Your pace lengthens on a downslope and shortens on an upgrade.
- Winds – A headwind shortens the pace and a tailwind increases it.
- Surfaces – Sand, gravel, mud, snow and similar surfaces tend to shorten your pace.
- Elements – Falling rain or snow causes the pace to be reduced in length.
- Clothing – Excess clothing or shoes with poor traction affect the pace length.
- Visibility – Poor visibility due to rain, snow, or fog will shorten your pace.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Ranger Pace Count Beads

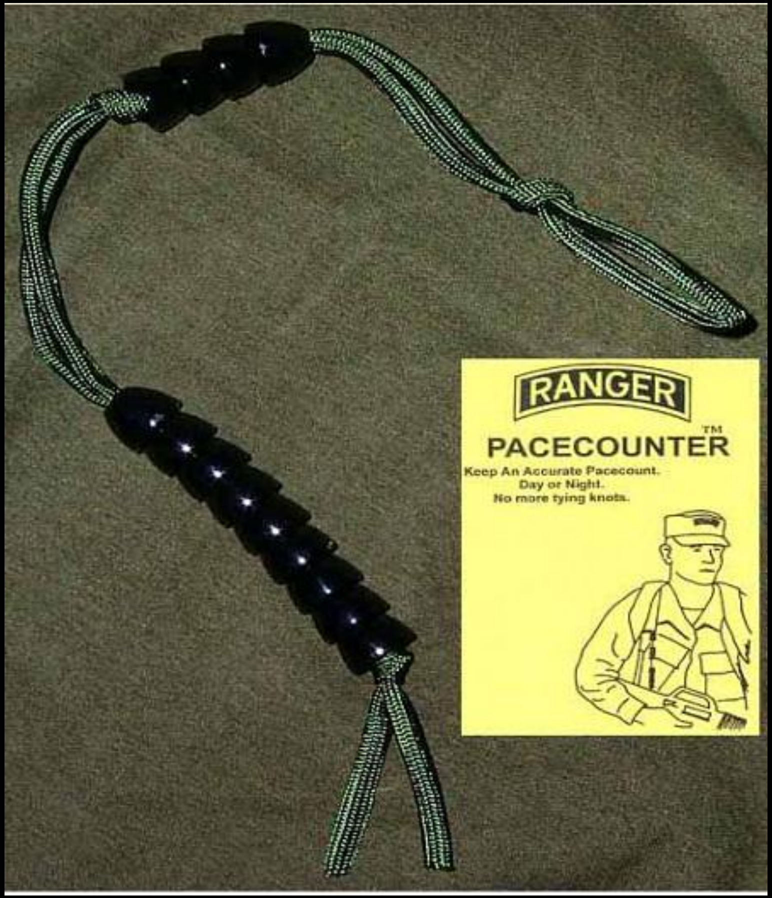
RANGER
TM
PACECOUNTER
Keep An Accurate Pacecount.
Day or Night.
No more tying knots.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

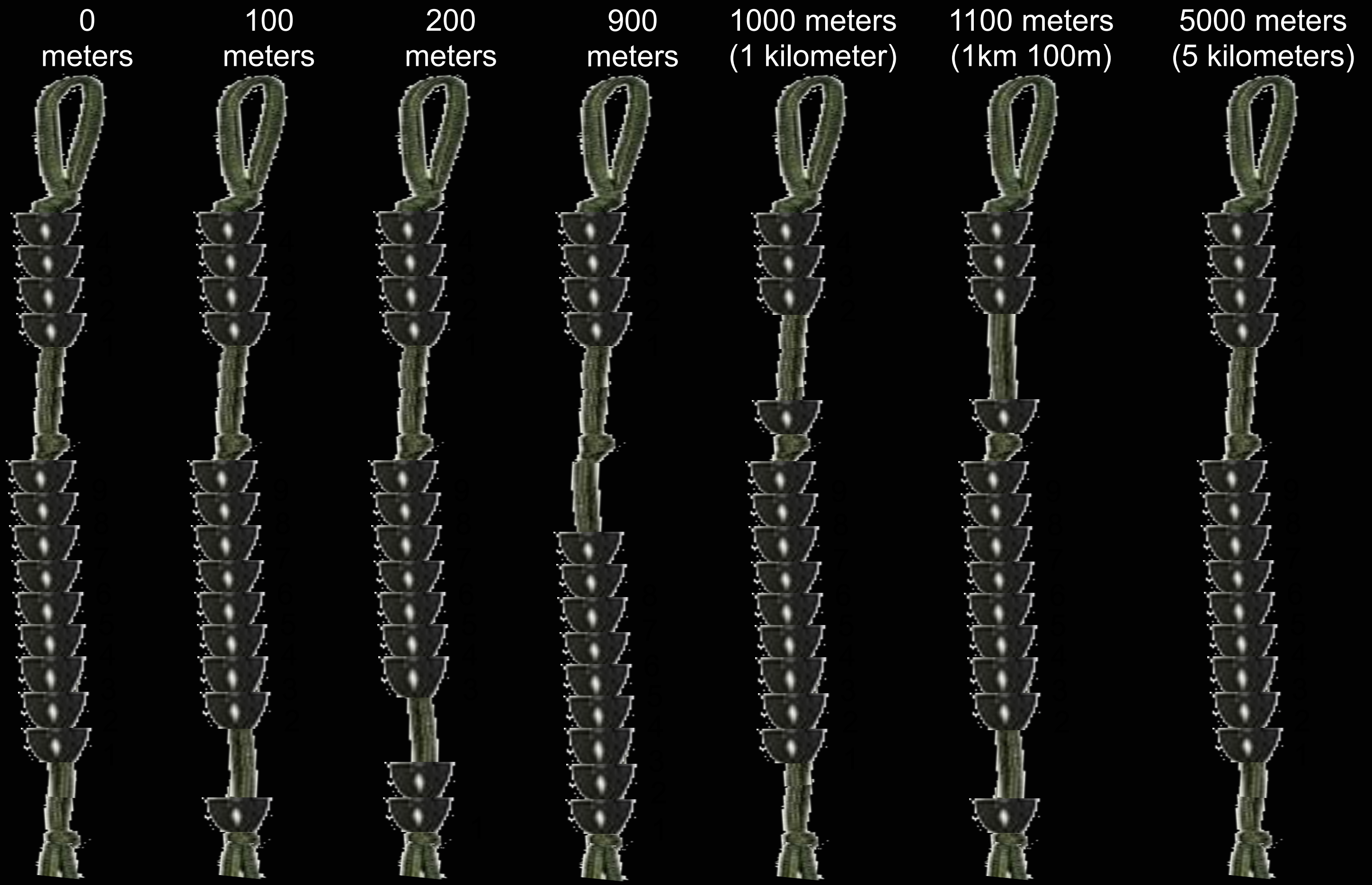
0 meters
100 meters
200 meters
900 meters
1000 meters (1 kilometer)
1100 meters (1km 100m)
5000 meters (5 kilometers)

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

FINDING YOURSELF

Baselines

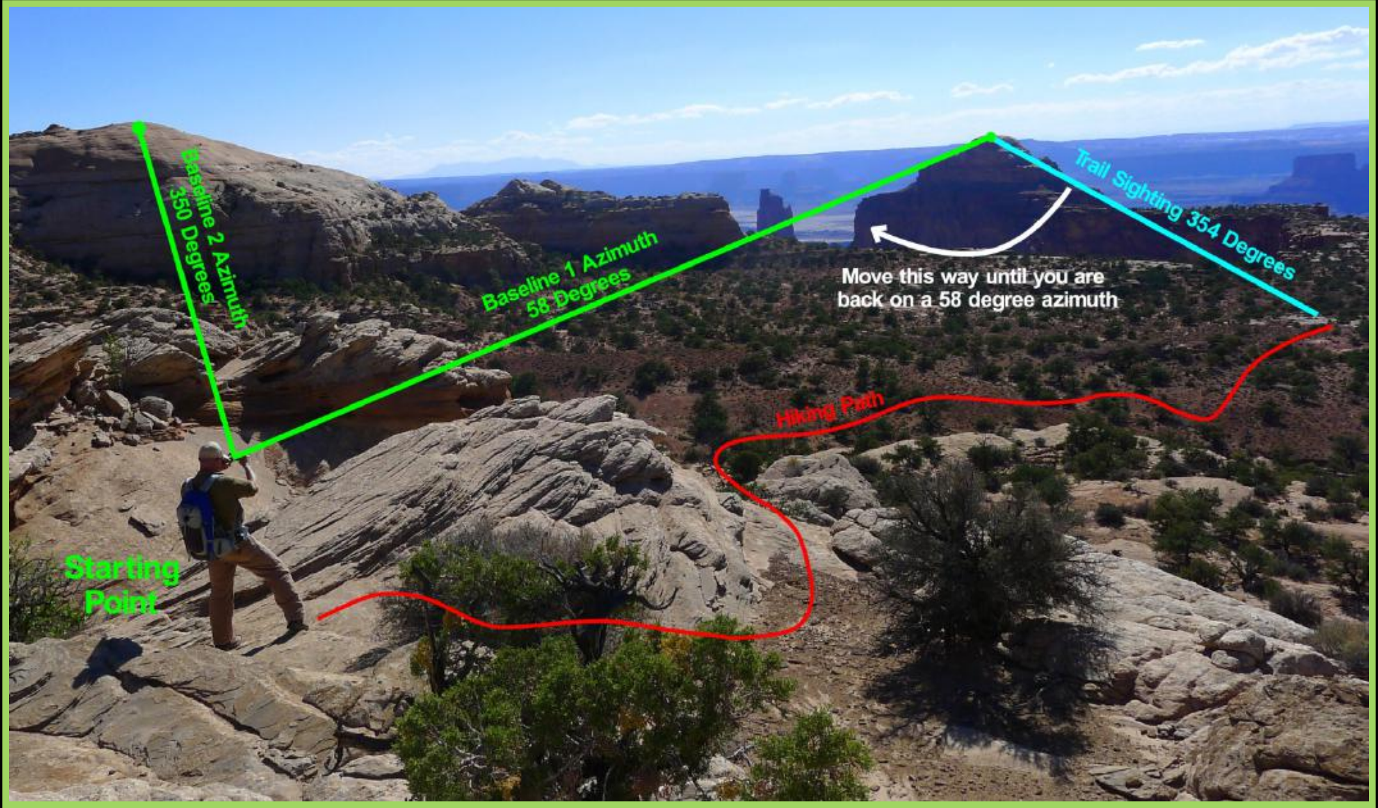
Baseline 2 Azimuth
350 Degrees
Baseline 1 Azimuth
58 Degrees
Trail Sighting 354 Degrees
Move this way until you are
back on a 58 degree azimuth
Hiking Path
Starting
Point

Find two prominent features. Shoot an azimuth to the first one and write it down. Shoot an azimuth to the second one and write it down. You can now travel freely. If you get lost, shoot an azimuth to the first feature and move until you get back on your first azimuth. Shoot an azimuth to your second feature and move until you get back on your second azimuth. You will be back at your starting point once both azimuths are the same as the ones you wrote down.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RESECTION

- Orient your map and secure its position. Locate a landmark on the map that you can actually see in the landscape. Take a compass bearing to that landmark by sighting and then rotating compass dial until the needle is boxed. Once you have a bearing, do not move the compass dial. Lay the corner of the compass on the map landmark and rotate the whole compass (not the dial) until the needle is boxed. If done properly the corner of the compass will still be over your landmark. Draw a line along the edge of the compass base.

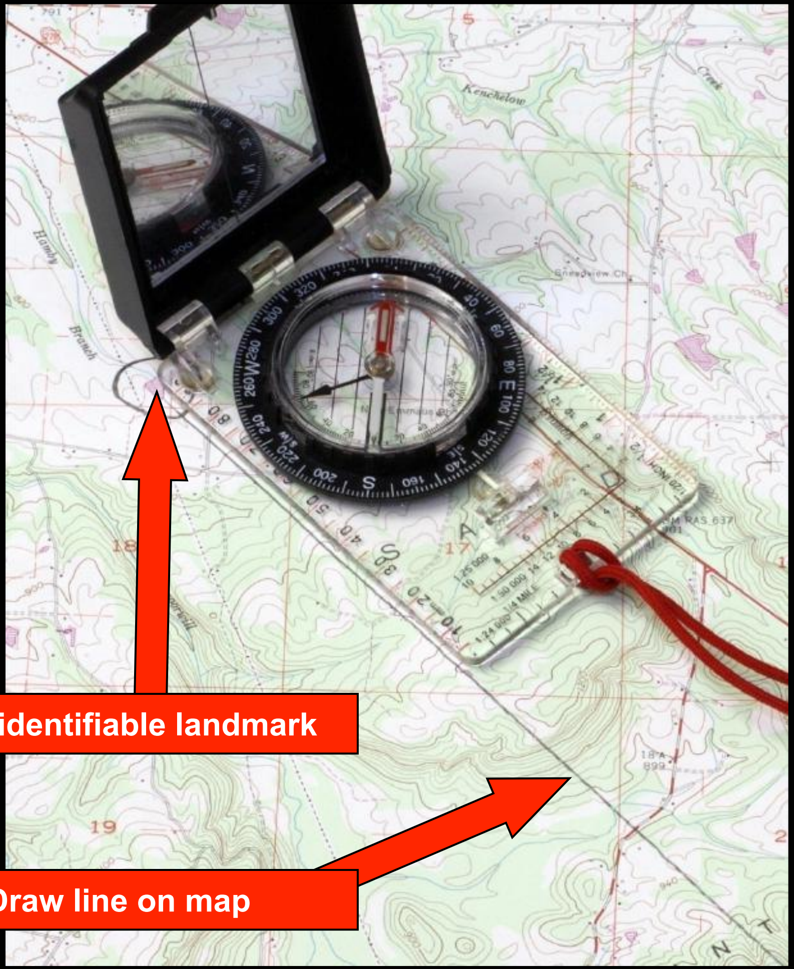

First identifiable landmark

Draw line on map

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RESECTION - Step 2

- Find a second identifiable landmark on the map and repeat the process from the previous slide. Where the two lines intersect is roughly where you are on the map.
- Finding a third point on the map and repeating the process is called "triangulation" and increases the accuracy of finding your location.

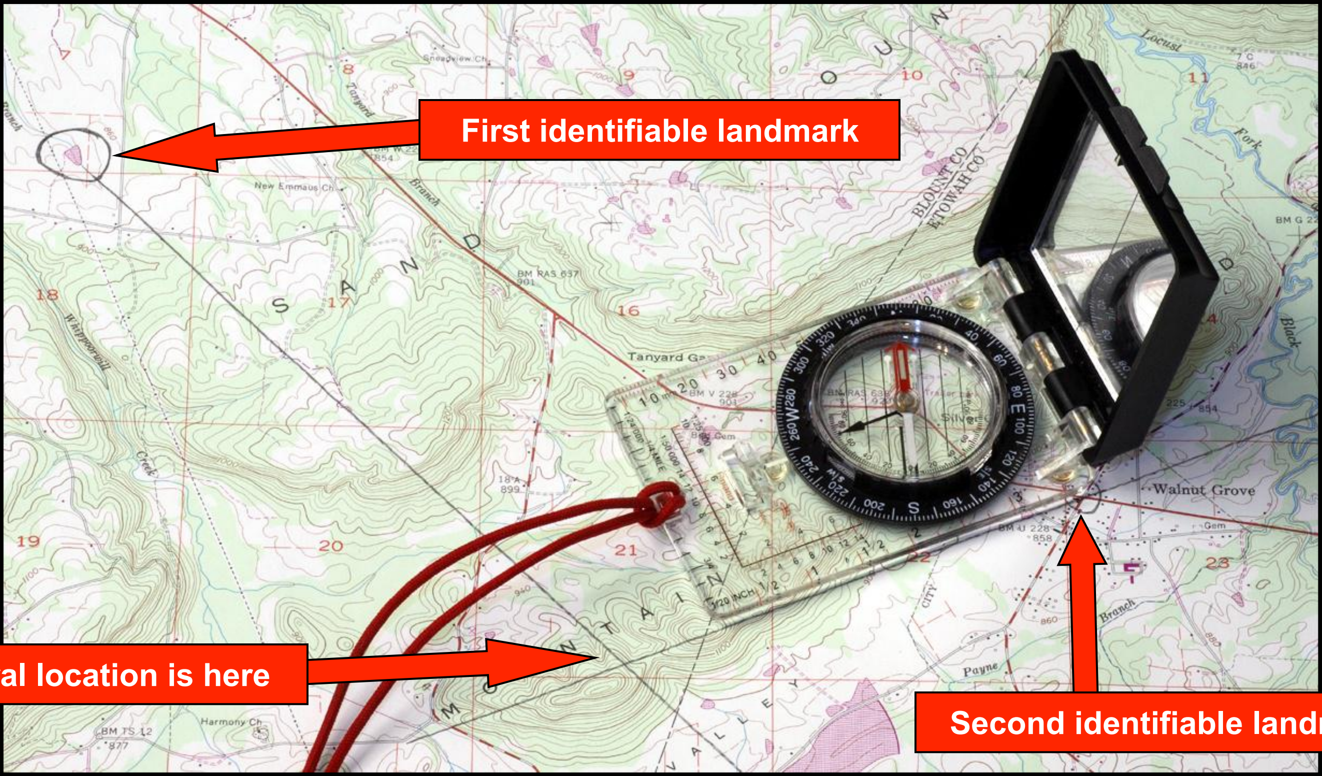

First identifiable landmark

Your general location is here

Second identifiable landmark

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

if you are already on a known location (such as a road, river or mountain ridge) but not sure exactly where you are on this feature, you may only need to shoot one azimuth to determine your exact position on the map

INTERSECTION is the reverse of RESECTION. An example of intersection is seeing smoke from a forest fire and determining where the fire is located on a map.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

PLOTTING COORDINATES

COPYRIGHT © 2013 RANDALL’S ADVENTURE TRAINING ®

Map Grids

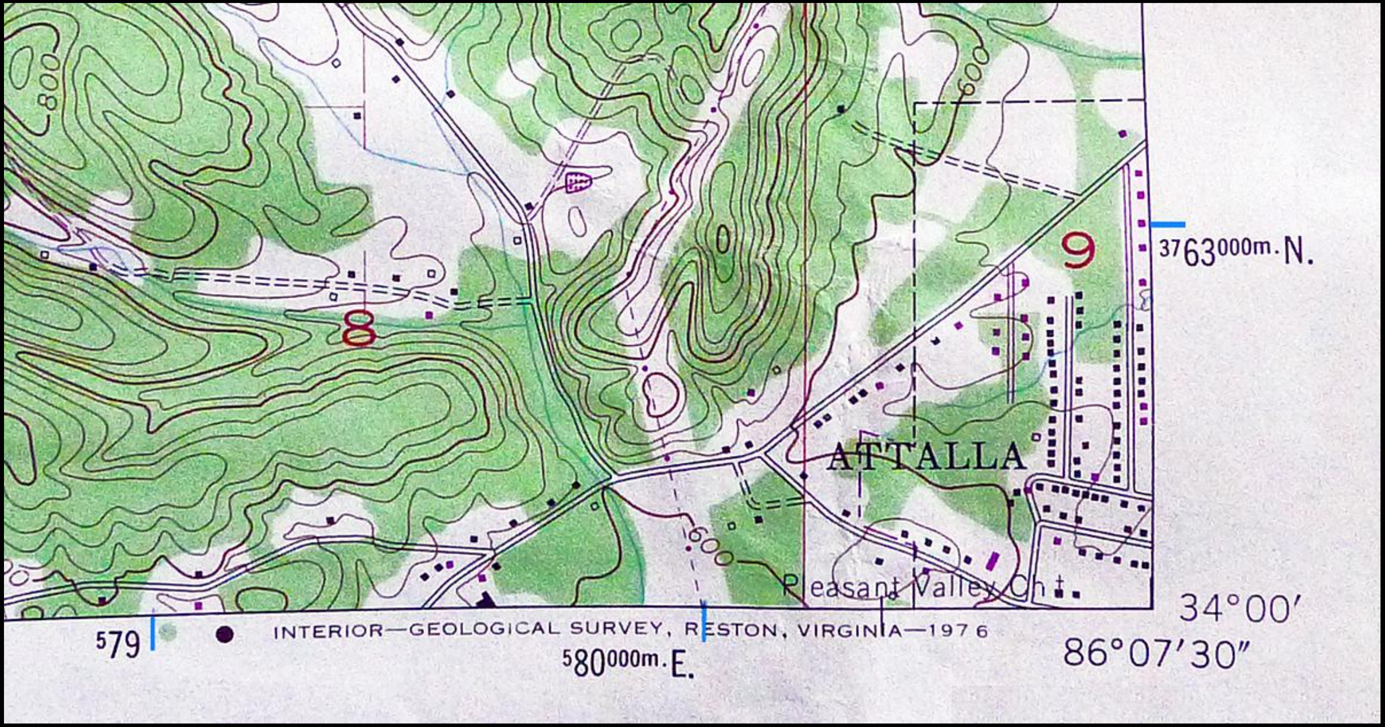
800
600
8
9
3763000m.N.
ATTALLA
600
Pleasant Valley Ch
34°00′
86°07′30″
579
INTERIOR—GEOLOGICAL SURVEY, RESTON, VIRGINIA—1976
580000m.E.

Always set your GPS to the grid system you are using on the map

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Universal Transverse Mercator (UTM)

UTM is a rectangular coordinate system based on the latitude and longitude (geographic) coordinate system. With UTM, the earth is divided into 60 zones that allows it to be projected onto maps with minimal distortion. All coordinates are expressed in meters.

The UTM grid system is the easiest method of relaying your position to others, finding yourself on a map after getting a location fix with a GPS unit, or navigating to a position on a map using a GPS unit.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

UTM: 60 Zones and 20 Latitude Bands

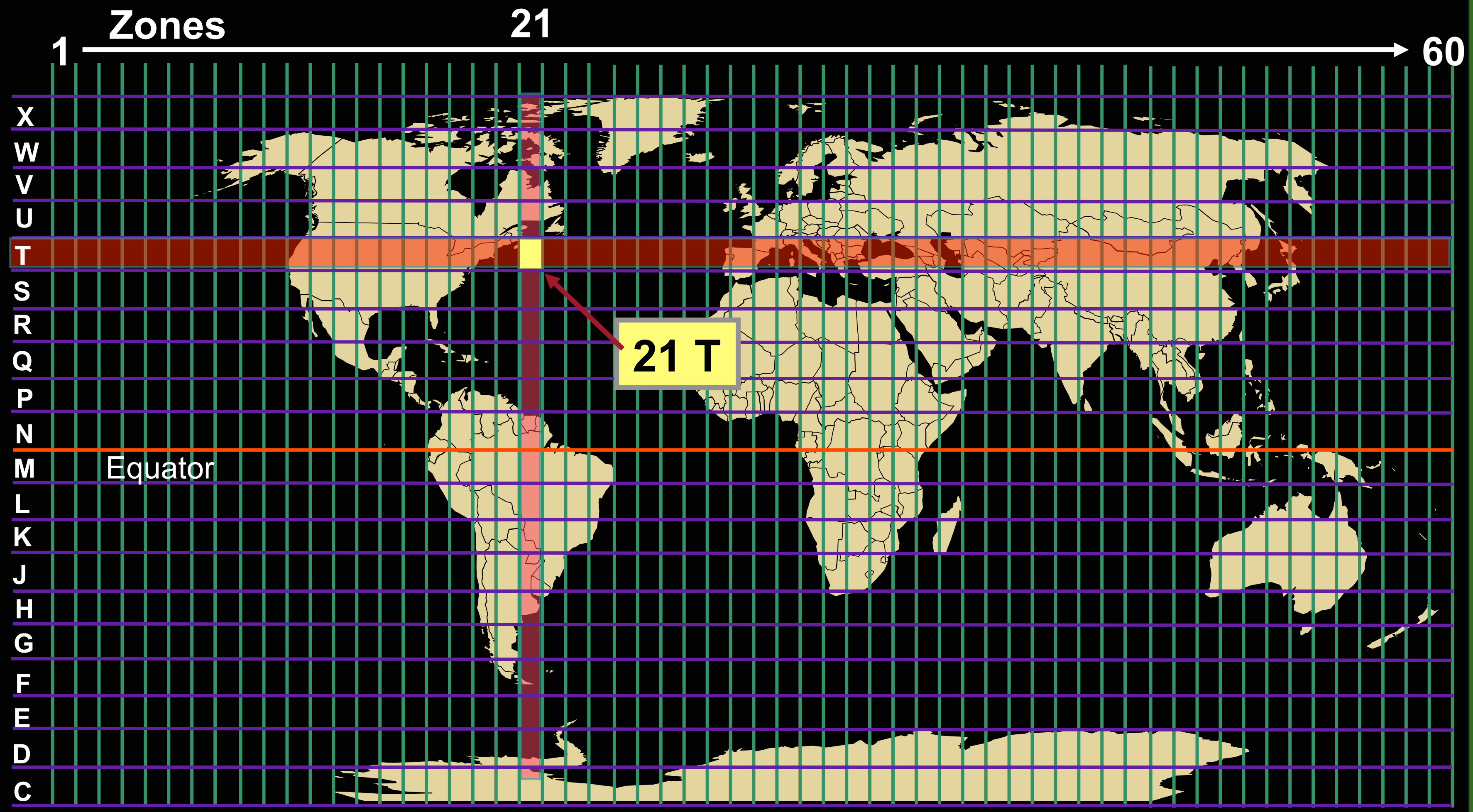
Zones
1
21
60
X
W
V
U
T
S
R
Q
P
N
M
L
K
J
H
G
F
E
D
C
Equator
21 T

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

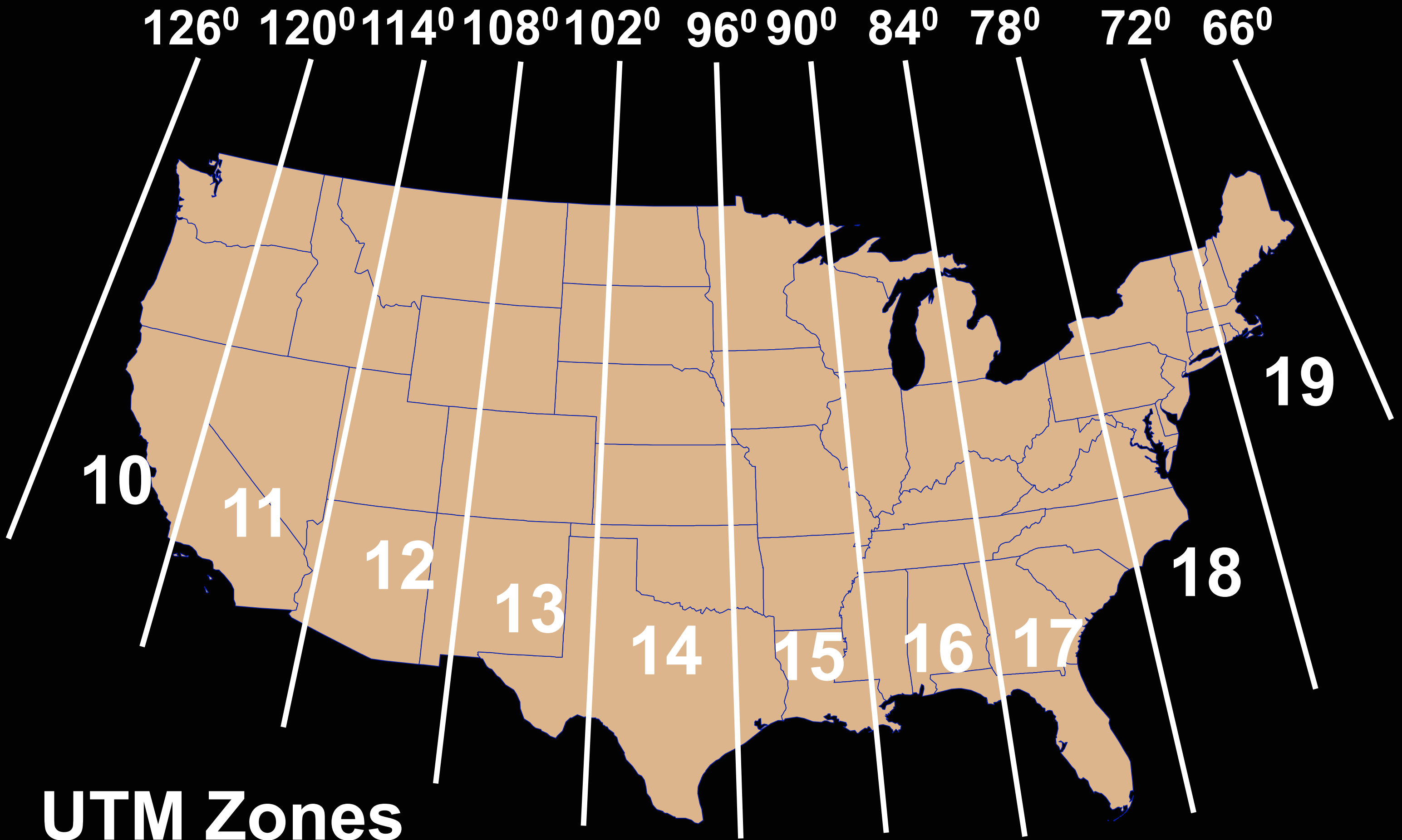
126^0
120^0
114^0
108^0
102^0
96^0
90^0
84^0
78^0
72^0
66^0
19
10
11
12
13
14
15
16
17
18

UTM Zones

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Universal Transverse Mercator (UTM)

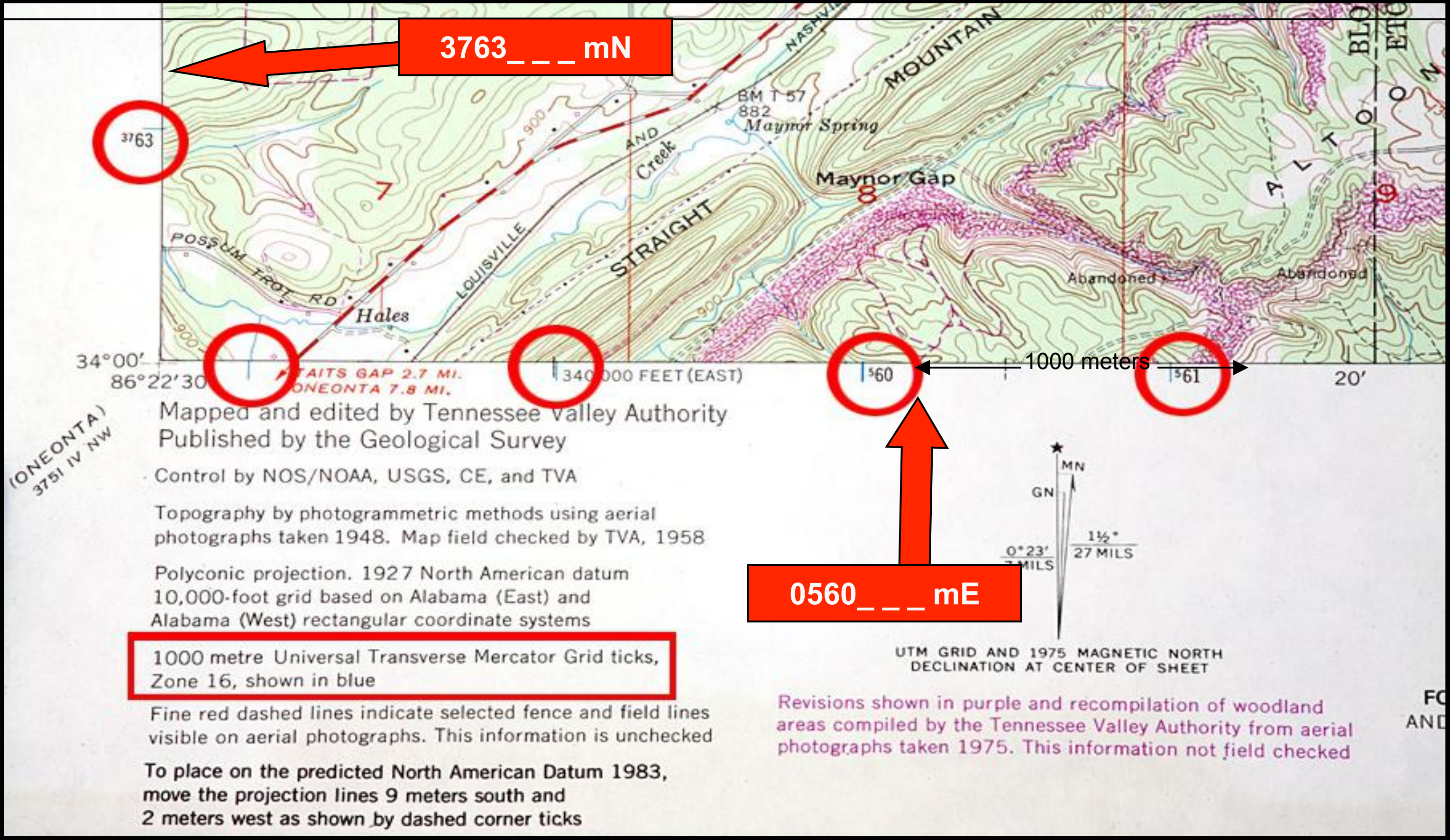
3763_ _ _ mN
³763
0560_ _ _ mE
³60
³61
1000 meters
340 000 FEET (EAST)
34°00′
86°22′30″
20′
MOUNTAIN
BM T 57
882
Maynor Spring
Maynor Gap
AND Creek
STRAIGHT
LOUISVILLE
POSSUM TROT RD
Hales
Abandoned
Abandoned
ALTOONA
STAITS GAP 2.7 MI.
ONEONTA 7.8 MI.
(ONEONTA)
3751 IV NW
MN
GN
0°23′
7 MILS
1½°
27 MILS
FO
AND

Mapped and edited by Tennessee Valley Authority
Published by the Geological Survey

Control by NOS/NOAA, USGS, CE, and TVA

Topography by photogrammetric methods using aerial photographs taken 1948. Map field checked by TVA, 1958

Polyconic projection. 1927 North American datum
10,000-foot grid based on Alabama (East) and Alabama (West) rectangular coordinate systems

1000 metre Universal Transverse Mercator Grid ticks, Zone 16, shown in blue

Fine red dashed lines indicate selected fence and field lines visible on aerial photographs. This information is unchecked

To place on the predicted North American Datum 1983, move the projection lines 9 meters south and 2 meters west as shown by dashed corner ticks

UTM GRID AND 1975 MAGNETIC NORTH DECLINATION AT CENTER OF SHEET

Revisions shown in purple and recompilation of woodland areas compiled by the Tennessee Valley Authority from aerial photographs taken 1975. This information not field checked

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Easting:

Is a 6-digit number to designate the EAST/WEST position in the grid. The grids are the widest at the equator. Since each grid has a different width, a system was created to measure the distance across a grid starting from the vertical center-line of each grid as opposed to starting from the left or right edge of a grid. The center line of each grid is called the central meridian. The central meridian is always assigned an easting value of 500,000 meters East. It is expressed as 500,000 mE. As you move west of the central meridian the easting will be a number less than 500,000. As you move to the east of the central meridian the easting will be a number greater than 500,000. An easting of zero will never occur, since a 6° wide zone is never more than 674,000 meters wide.

Northing:

The second number is called the Northing. It is a 7-digit number that designates how many meters you are north or south of the equator.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

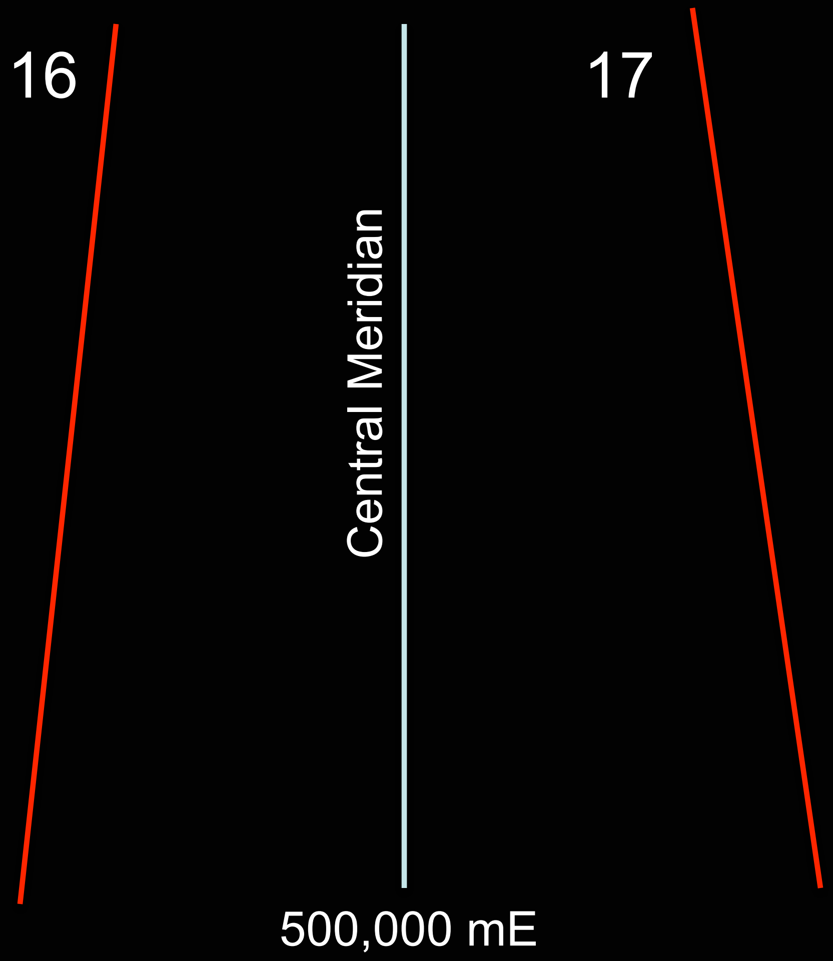
16
17
Central Meridian
500,000 mE

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

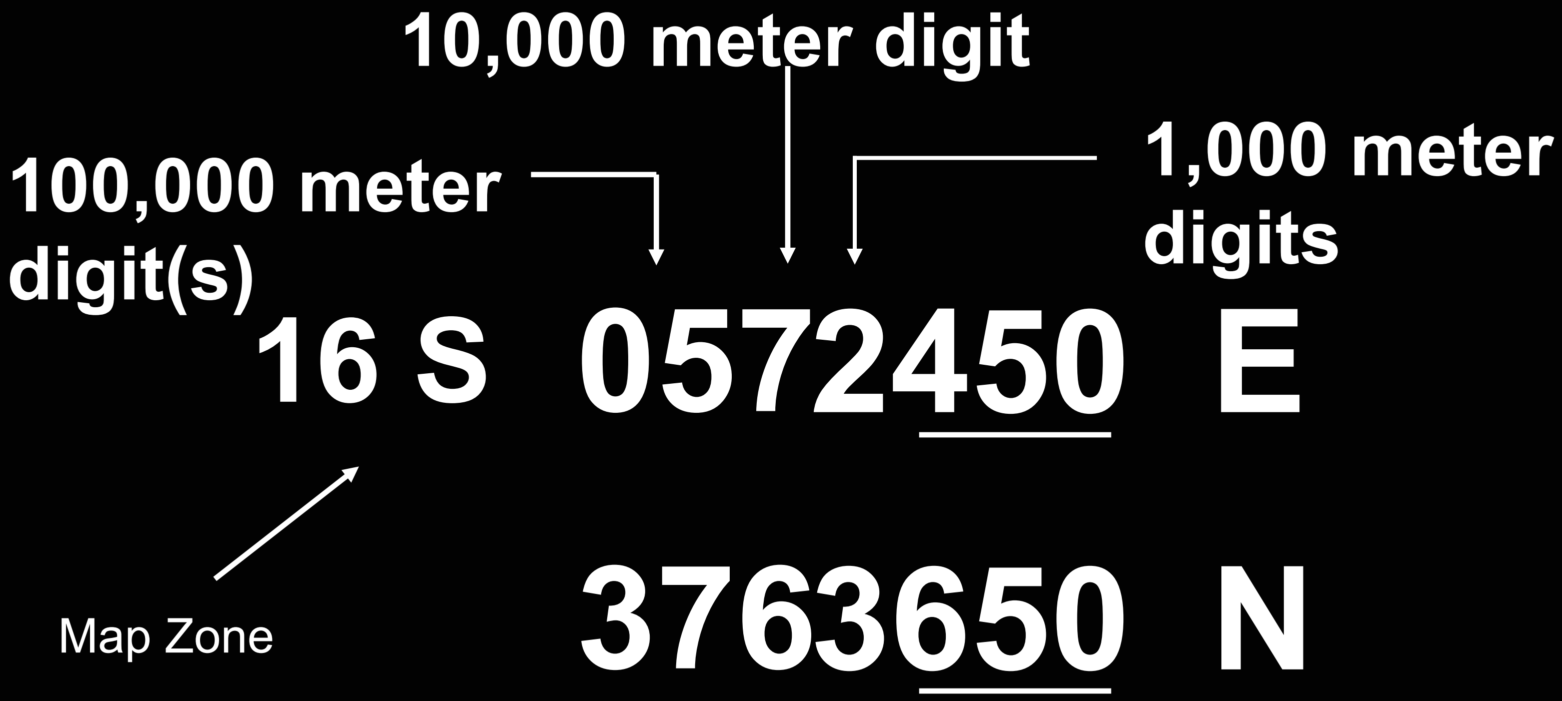

10,000 meter digit

100,000 meter digit(s)

1,000 meter digits

16 S 0572450 E

3763650 N

Map Zone

You only have to plot the last 3 numbers. The rest of the coordinate values are provided on the map

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

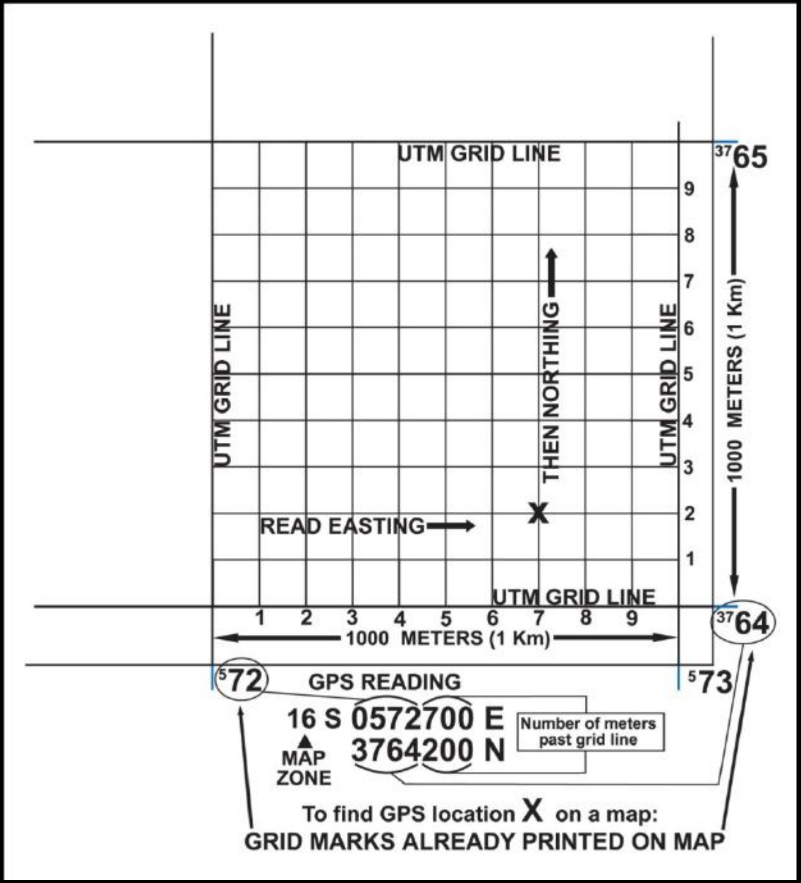
UTM GRID LINE
UTM GRID LINE
THEN NORTHING
UTM GRID LINE
1000 METERS (1 Km)
3765
9
8
7
6
5
4
3
2
1
X
READ EASTING
UTM GRID LINE
1 2 3 4 5 6 7 8 9
1000 METERS (1 Km)
3764
572
GPS READING
573
16 S 0572700 E
3764200 N
Number of meters past grid line
MAP ZONE
To find GPS location X on a map:
GRID MARKS ALREADY PRINTED ON MAP

Read EASTING First

Then Read NORTHING

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

UTM grid reader

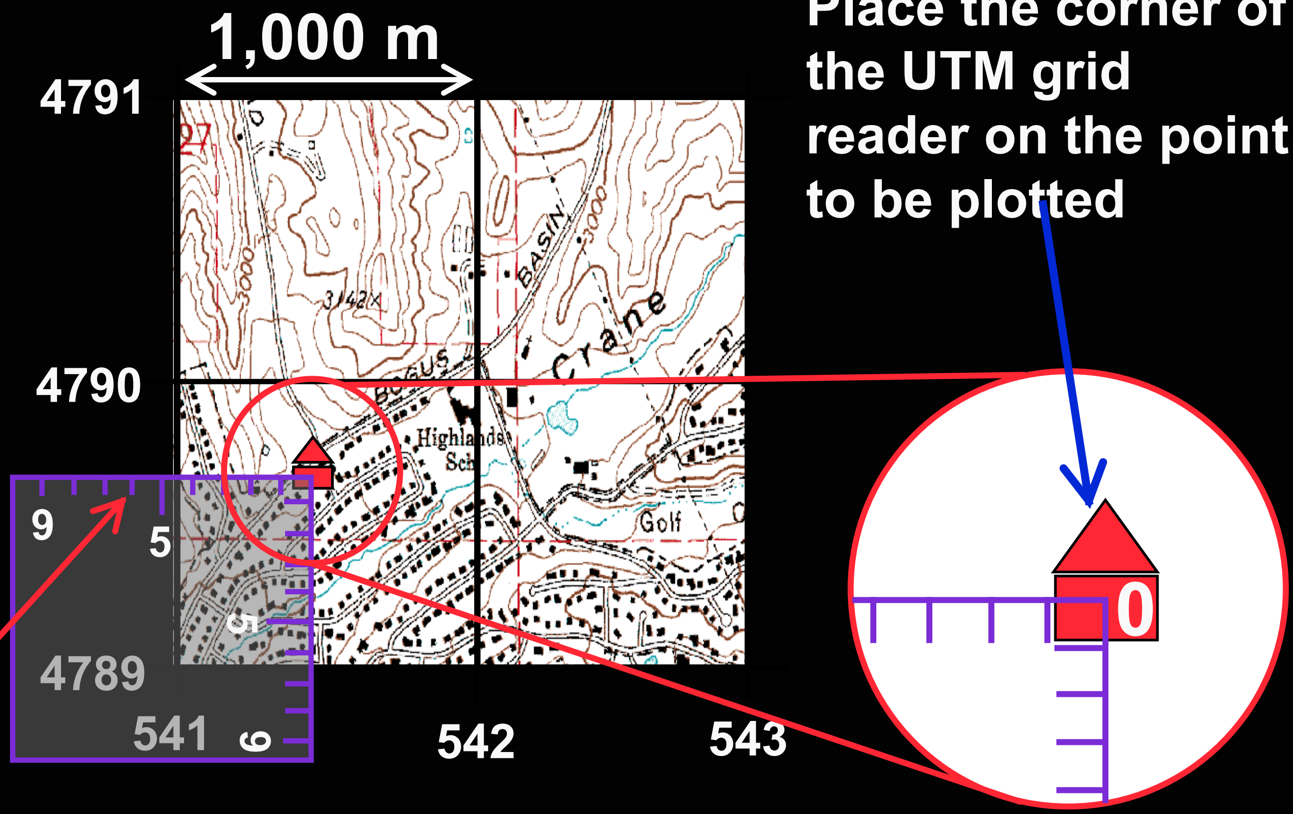
1,000 m
4791
4790
4789
541
542
543
9
5
5
9
0
27
3000
3142X
BASIN
3000
BOGUS
Crane
Highlands
Sch
Golf

Place the corner of the UTM grid reader on the point to be plotted

Each tic = 100 meters on this grid reader

House coordinates = 0541450mE 4789650mN

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Using GPS with Map

When using a GPS to acquire an azimuth to a location, setup your GPS for Magnetic North Heading if you are not adjusting declination on your compass. Use True North Heading if your are adjusting for declination on your compass.

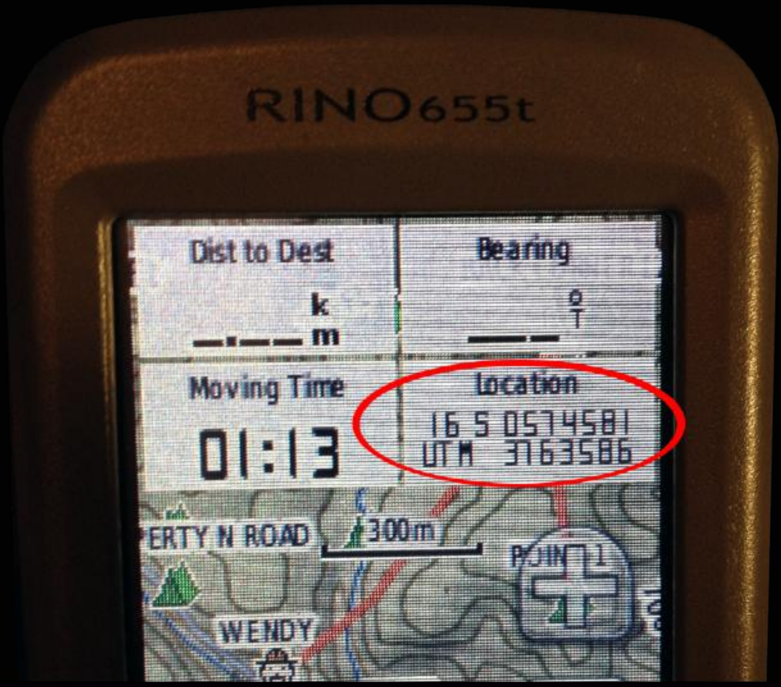
RINO655t
Dist to Dest
km
Bearing
°T
Moving Time
01:13
Location
16 S 0574581
UTM 3163586
ERTY N ROAD
300m
POINT 1
WENDY

Plotting GPS to Map

1) Find current location with GPS
2) Use scale card to plot current location on the map

Plotting Map to GPS

1) Plot location on map of where you want to go with a scale card
2) Enter plotted coordinates into the GPS as a new waypoint

MGRS System

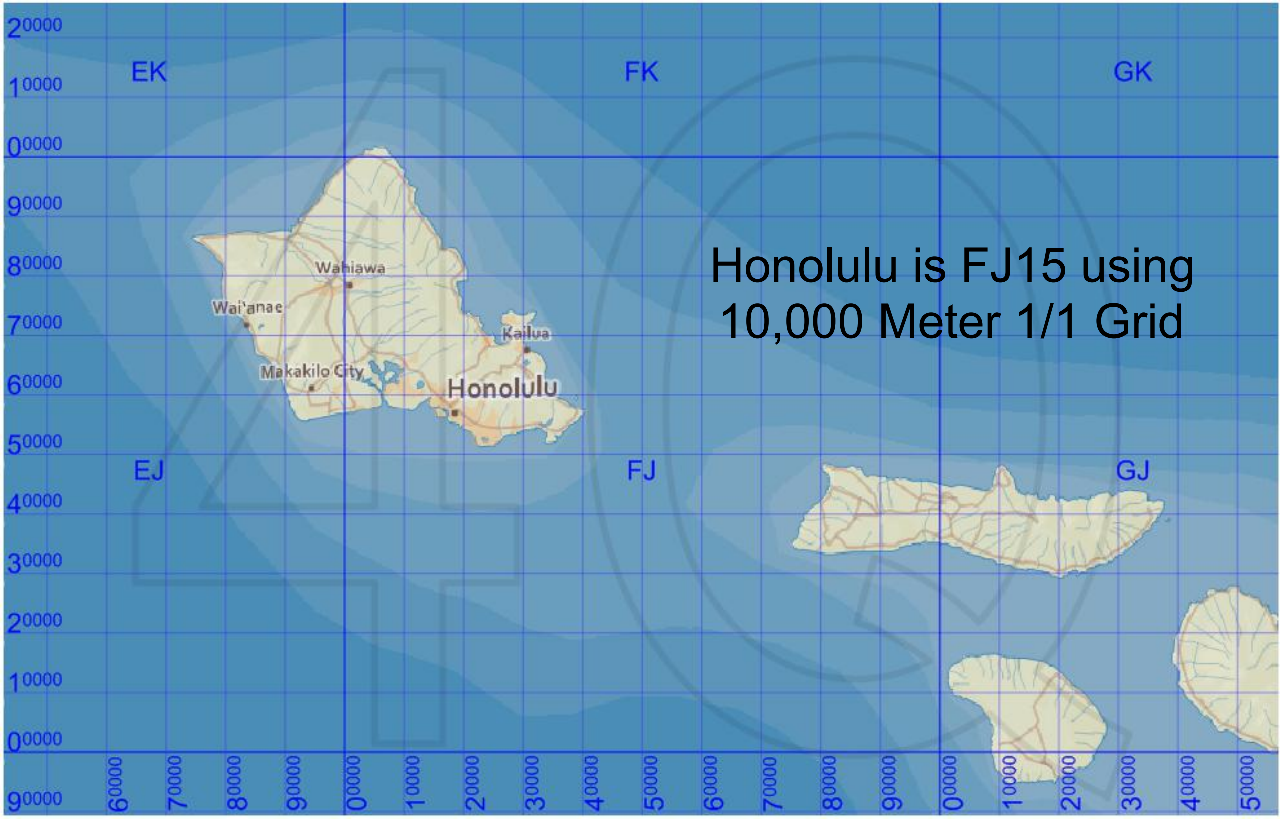
20000
10000
00000
90000
80000
70000
60000
50000
40000
30000
20000
10000
00000
90000
EK
FK
GK
EJ
FJ
GJ
Wahiawa
Wai'anae
Kailua
Makakilo City
Honolulu
60000
70000
80000
90000
00000
10000
20000
30000
40000
50000
60000
70000
80000
90000
00000
10000
20000
30000
40000
50000

Honolulu is FJ15 using 10,000 Meter 1/1 Grid

5/5 = 1 Meter, 4/4 = 10 meter, 3/3 = 100 meter,
2/2 = 1000 meter, 1/1 = 10,000 meter

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Why do we not use Latitude / Longitude for Land Nav?

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Communicating Location

General Location
Type Of Navigation Grid Used
Grid Location (Easting Then Northing)
Direction Of Travel If Moving

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Declination

Magnetic Declination is the difference between True North and Magnetic North. A compass needle ALWAYS points to Magnetic North.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

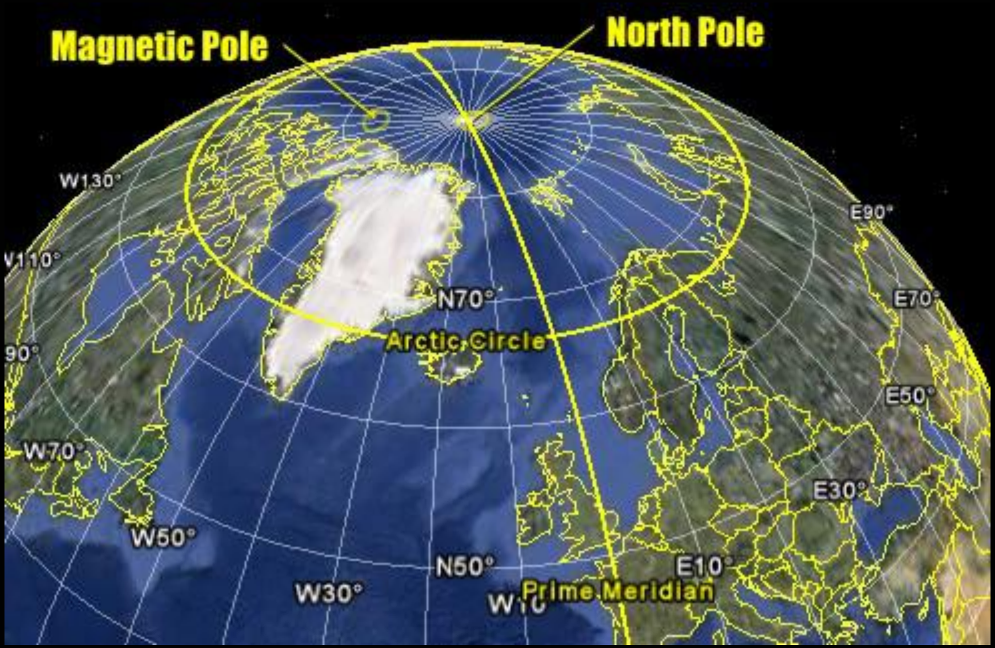
Magnetic Pole
North Pole
W130°
W110°
90°
W70°
W50°
W30°
N50°
W10°
Prime Meridian
E10°
E30°
E50°
E70°
E90°
N70°
Arctic Circle

Difference in True North
And Magnetic North

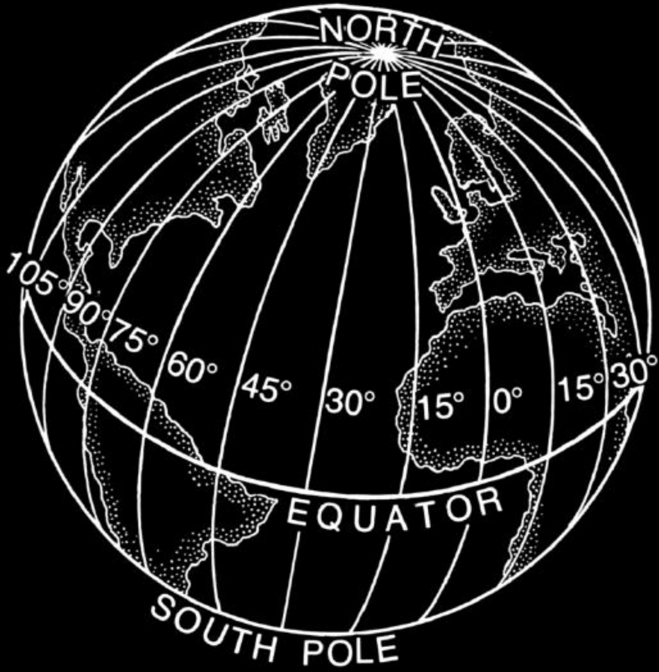
NORTH POLE
105°
90°
75°
60°
45°
30°
15°
0°
15°
30°
EQUATOR
SOUTH POLE

Longitude Lines
(True North)

Magnetic North is moving approximately 40 miles per year.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Magnetic Declination

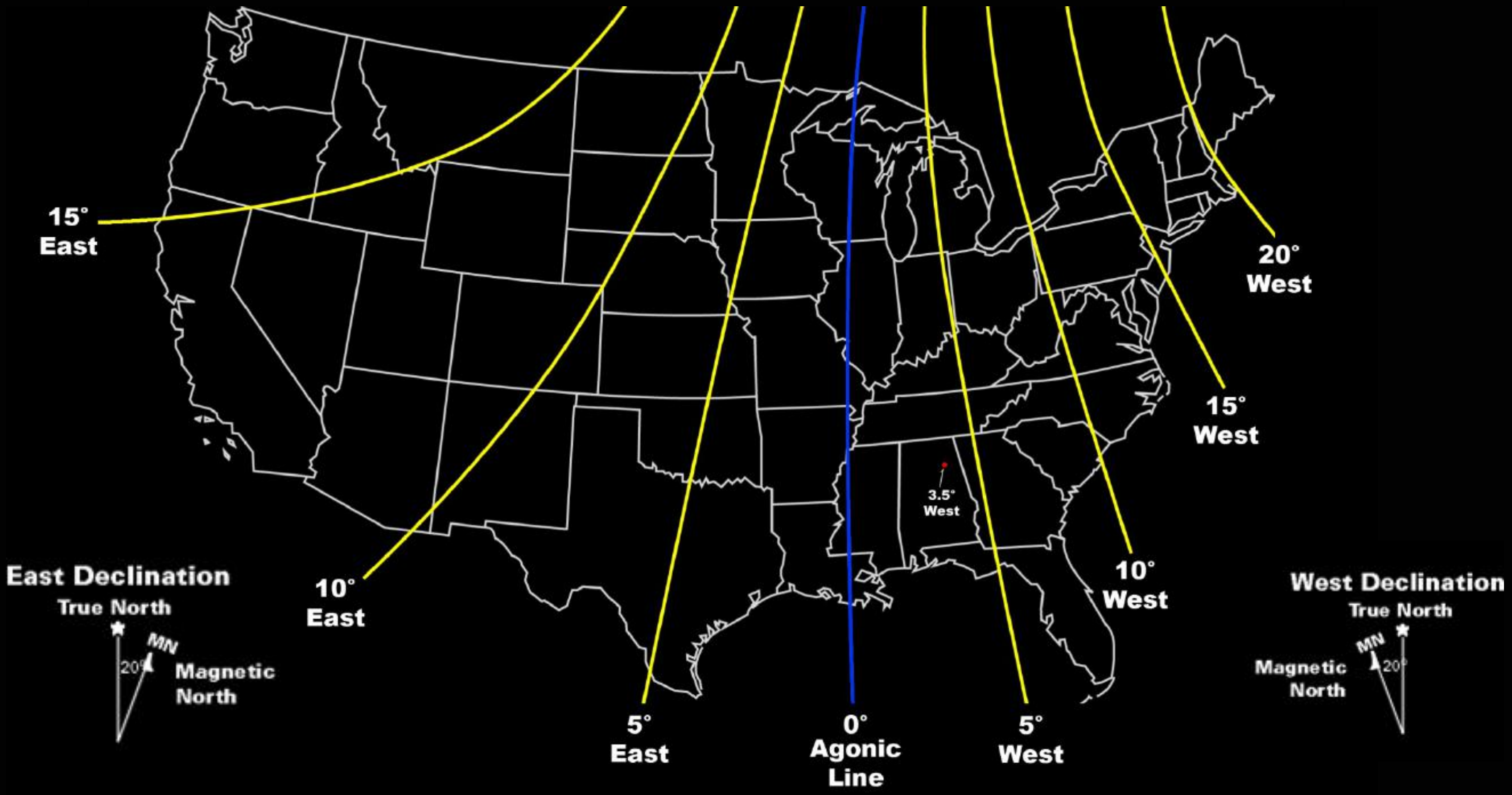
15° East
10° East
5° East
0° Agonic Line
5° West
10° West
15° West
20° West
3.5° West
East Declination
True North
MN
20°
Magnetic North
West Declination
True North
MN
20°
Magnetic North

If you do not have a compass that adjusts for declination remember to add or subtract declination to your azimuth. Add for West – Subtract for East
EAST IS LEAST - WEST IS BEST

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Depending on your map, it may printed with True North Reference lines or Grid North Reference lines

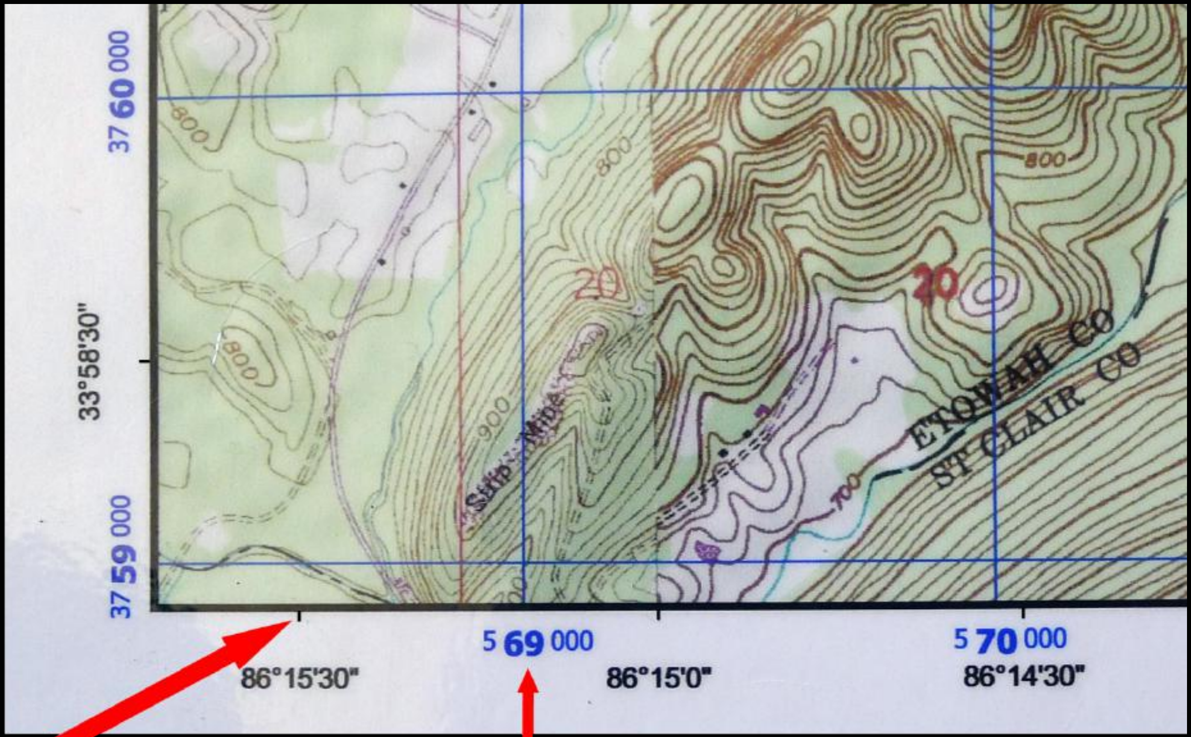
37 60 000
33°58'30"
37 59 000
5 69 000
5 70 000
86°15'30"
86°15'0"
86°14'30"
800
900
700
20
Strip Mine
ETOWAH CO
ST CLAIR CO

True North (Longitude)

Grid North (UTM Or Other Grid)

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

Adjusting For Declination For Map Use

Declination is the difference between true north and magnetic north when using the longitude lines as north reference.

Declination is the difference or spread between grid north and magnetic north when using the UTM grid lines as north reference.

At the bottom of your map, if the MN line is left of the True North Line it will always be West Declination. If the MN line is right of the True North Line it will always be East Declination

If you are using a compass that is not adjusted for declination, add the amount for west declination, subtract the amount for east declination. If your compass is adjustable for declination, set it for the north reference declination value you are using. No further addition or subtraction is required.

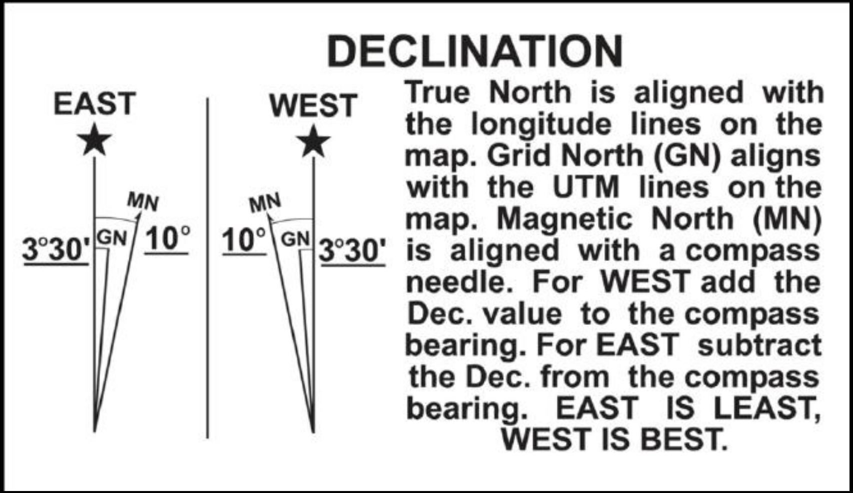

DECLINATION

True North is aligned with the longitude lines on the map. Grid North (GN) aligns with the UTM lines on the map. Magnetic North (MN) is aligned with a compass needle. For WEST add the Dec. value to the compass bearing. For EAST subtract the Dec. from the compass bearing. EAST IS LEAST, WEST IS BEST.

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

True North / Magnetic North / Grid North

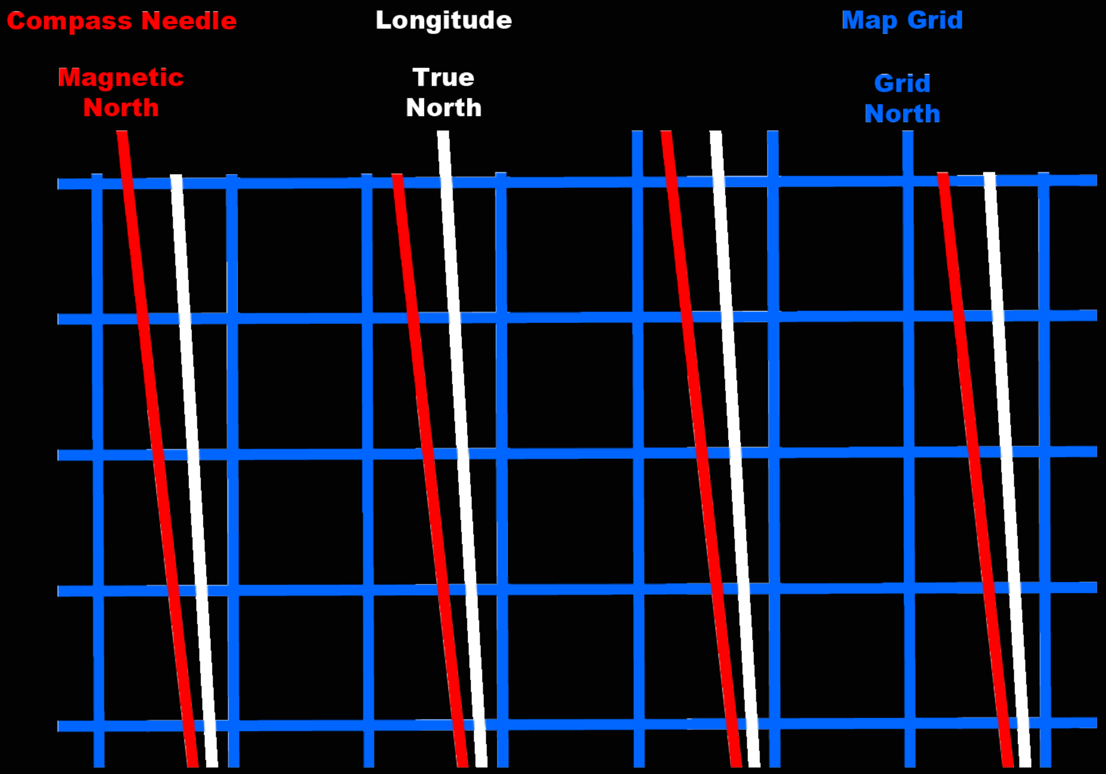
Compass Needle
Magnetic North
Longitude
True North
Map Grid
Grid North

If using Grid North as Reference then declination would be set to 6 degrees West
If using True North as Reference then declination would be set to 3 degrees West

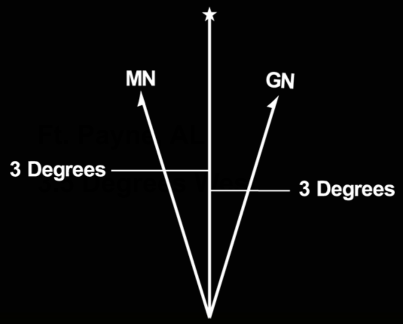
MN
GN
3 Degrees
3 Degrees

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

RANDALL'S ADVENTURE TRAINING ®
LAND NAVIGATION

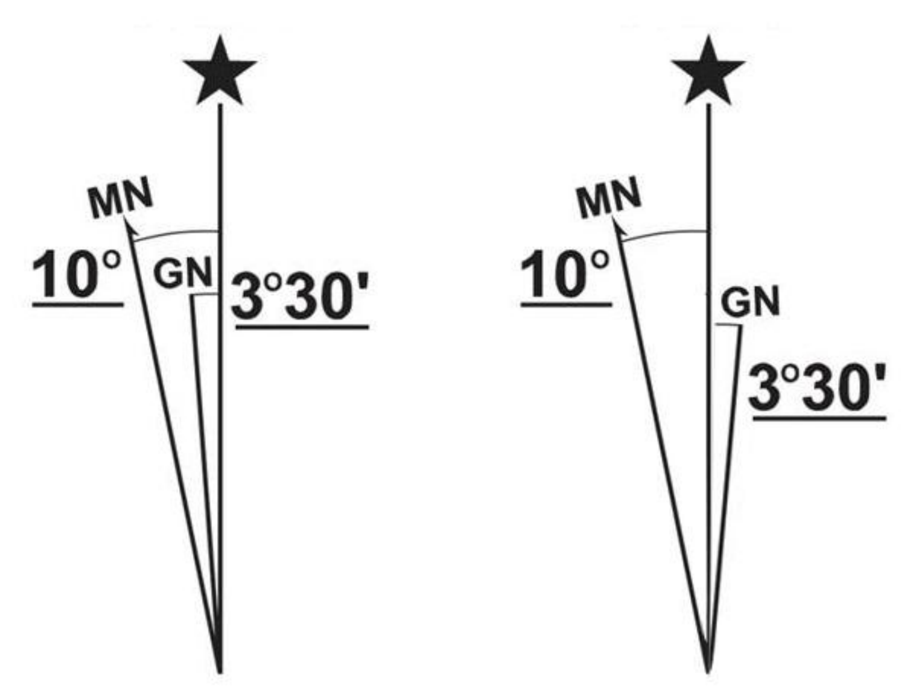
MN
10°
GN
3°30'
MN
10°
GN
3°30'

What is our declination value for each of the diagrams to use Grid North Reference?

COPYRIGHT © 2013 RANDALL'S ADVENTURE TRAINING ®

USER ADVICE
this book
could save
your life

THE SURVIVAL HANDBOOK

ESSENTIAL SKILLS FOR OUTDOOR ADVENTURE

LONDON, NEW YORK, MUNICH, MELBOURNE, DELHI

Senior Editors Nicky Munro, Richard Gilbert

Senior Art Editors Michael Duffy, Gillian Andrews

Editors Bob Bridle, Chris Hawkes,
Philip Morgan, Gill Pitts, Clare Weber

US Editor Chuck Wills

Designers Sarah-Anne Arnold, Katie Eke,
Brian Flynn, Phil Gamble, Sharon Spencer

Lead Illustrator Mike Garland
Assisting Illustrators Darren Awuah,
Phil Gamble, Peter Liddiard, Mark Walker

Production Editor Tony Phipps
Production Controller Norma Weir

Managing Editor Stephanie Farrow
Managing Art Editor Lee Griffiths

Art Director Bryn Walls
Publisher Jonathan Metcalf

First American Edition, 2009

Published in the United States by
DK Publishing, 375 Hudson Street
New York, New York 10014

09 10 11 12 10 9 8 7 6 5 4 3 2 1

SD308—April 09

Copyright © 2009 Dorling Kindersley Limited
All rights reserved

Without limiting the rights under copyright reserved above, no part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted, in any form, or by any means (electronic, mechanical, photocopying, recording, or otherwise), without the prior written permission of both the copyright owner and the above publisher of this book.

Published in Great Britain by Dorling Kindersley Limited.

A catalog record for this book is available
from the Library of Congress.

ISBN 978-07566-4279-2

DK books are available at special discounts when purchased in bulk for sales promotions, premiums, fund-raising, or educational use. For details, contact: DK Publishing Special Markets, 375 Hudson Street, New York, New York 10014 or SpecialSales@dk.com.

Printed and bound in Hong Kong by Hung Hing

Discover more at
www.dk.com

IMPORTANT NOTICE

Some of the techniques desribed in this book should be used only in dire emergencies, when the survival of individuals depends upon them. The publisher cannot be held responsible for any injuries, damage, loss or prosecutions resulting from the use or misuse of the information in this book. Do not practice these techniques on private land without the owner's permission, and obey all laws relating to the protection of land, property, plants, and animals.

CONTENTS

8 Introduction

1 BEFORE YOU GO

16 **Introduction: Prepare yourself**
18 Getting into shape
20 Mental preparation
22 Planning your journey
24 Emergency plan of action
26 **Introduction: Know your environment**
28 Temperate environments
30 Tropical environments
32 Mountain environments
34 Desert environments
36 Cold environments
38 Marine environments
40 **Introduction: Get the right gear**
42 Choosing your kit
44 Choosing a backpack
46 Dressing for the outdoors
50 Extreme survival—in the wilderness
52 Sleeping systems
56 Eating on the trail
58 Camping stoves
60 Your survival kit

2 ON THE TRAIL

64 **Introduction: Find your way**
66 Maps and map-reading
68 Get your bearings
72 Route finding
74 Navigating without a compass
76 Natural navigation
78 How weather works
80 Weather phenomena
82 Understanding local weather
84 **Introduction: Make a move**
86 Traveling on foot
88 Crossing rivers
90 Scrambling and climbing
92 Extreme survival—in the mountains
94 Moving over snow
98 Using pack animals
100 Four-wheel driving
104 Kayaking and canoeing
106 Building a raft
108 Swimming

3 CAMP CRAFT

112 **Introduction: Camp essentials**
114 Organizing your site
116 Staying clean
118 Making fire
120 The elements of fire
122 Char cloth and fire can
124 Types of fire
126 Making sparks and flames
128 Extreme survival—in the desert
130 Fire by friction
134 **Introduction: Manual skills**
136 Man-made cordage
138 Natural cordage
142 Tying knots
146 Using cutting tools

4 TAKING SHELTER

154 **Introduction: Taking shelter**
156 Hollows and caves
158 Use-anywhere shelters
162 Forest shelters
166 Tropical shelters
172 Extreme survival—in the jungle
174 Desert shelters
176 Shoreline shelters
178 Snow shelters

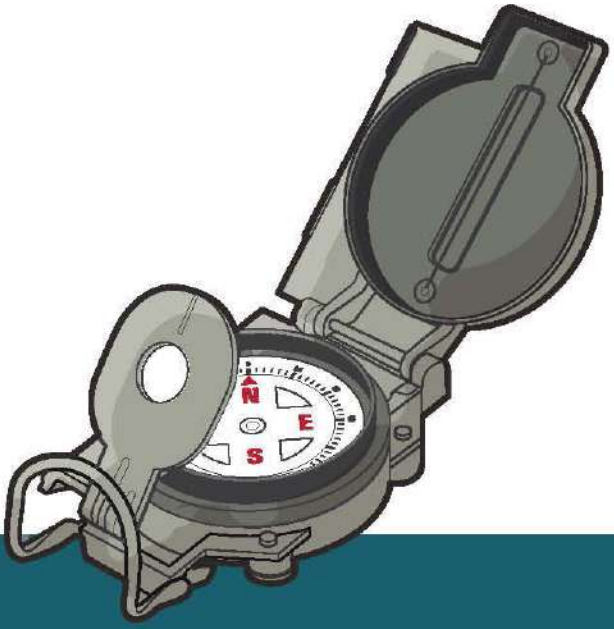

5 WATER AND FOOD

184 **Introduction: Find and treat water**
186 The importance of water
188 Finding water: Temperate climates
192 Finding water: Hot climates
194 Finding water: Cold climates
196 Finding water: At sea
198 Carrying and storing water
200 Treating water
202 **Introduction: Find and prepare food**
204 Wild cooking
206 Edible plants
208 Catching fish
212 Preparing fish
214 Extreme survival—adrift at sea
216 Trapping animals
220 Preparing small mammals
222 Preparing large mammals
224 Preparing other animals
226 Catching birds
228 Preparing birds

6 IN AN EMERGENCY

232 **Introduction: In an emergency**
234 Assessing your situation
236 Attracting rescuers
242 Wild animals
244 Extreme survival—in cold conditions
246 Environmental hazards
250 Surviving at sea

7 FIRST AID

258 **Introduction: First aid**
260 First aid essentials
262 Flesh wounds
266 Bites and stings
268 Poisonous plants and intestinal parasites
270 Bones, joints, and muscles
272 Treating exposure
274 Treating shock
275 Breathing difficulties
276 Unconsciousness
278 Moving a casualty

APPENDIX

Wild food

280 Temperate plants
282 Desert plants
284 Tropical plants
286 Plants in cold climates
288 Coastal plants
290 Mammals
292 Birds
294 Amphibians and reptiles
296 Fish
298 Invertebrates

Natural Dangers

300 Wild animals
302 Insects and arachnids
304 Poisonous plants

306 Resources
308 Glossary
312 Index
319 About the author
319 Acknowledgements

PREPARE YOURSELF

Most survival situations arise in one of two ways: either you are thrust into a situation not of your making and beyond your control, or (as is more often the case) a situation develops because of a sequence of events that could have been avoided had you recognized the danger signs and acted on them at the earliest opportunity. Unfortunately, most survival situations occur as a result of ignorance, arrogance, an overconfident belief in one's own ability, or because the forces of nature have been underestimated.

In the military forces, troops prepare for survival situations by learning the basic principles and techniques and then practicing them until they become second nature. The more you practice a particular survival skill, the more you'll understand how and why it works, and the more you'll understand your own strengths and weaknesses. Many helicopter crew members who have been involved in a survival situation, for example, report that, having ditched their aircraft in the sea, they could not remember how they subsequently

In this section YOU WILL DISCOVER...

- how keeping **in shape** can keep you **out of trouble...**
- the importance of a **positive mental attitude...**
- that where there's a **will to survive** there's a **way to survive...**
- the difference between a **coping strategy** and an **aggravating factor...**
- how **"show-stoppers"** could **scupper your trip...**
- why an **emergency plan of action** could **save your life...**

escaped from their aircraft and successfully boarded their liferaft. Thanks to their intense training regime, their actions had become instinctive and therefore subconscious.

Whether you're preparing for an overnight camping trip or a year in Africa, the more prepared you are to meet the challenges posed by a particular environment, the more likely you are to be able to cope—both physically and mentally—should you then find yourself confronted with a survival situation.

It would be foolish to think that the amount of prior preparation required for an expedition is directly related to the length of the trip or its perceived danger.

ACCIDENTS CAN HAPPEN anywhere. One of the most important things you can do to improve your chances of survival is to let people know where you're going and when you'll be back. Leave an itinerary of your trip with family or friends, and arrange a time when you will call to let them know you are safe.

In many instances, simply making sure you have a cell phone with you can prevent an accident becoming a major survival incident. In 2007, a professional athlete, out running with her dog in Moab, Utah, slipped on ice, fell approximately 60 ft (18 m), and broke her pelvis. Having left her cell phone in her vehicle, she managed to survive for two nights in the freezing desert before rescuers were finally lead to her by her dog.

In a separate incident in 2008, a farm worker in the UK caught his arm in machinery. Unable to attract attention and not having his cell phone with him, he was faced with a stark choice when the machinery caught fire: bleed to death, burn to death, or cut off his own arm with a pocket knife—he chose the latter.

> “ Never assume that the amount of **prior preparation** required is directly related to the length of a trip or its **perceived danger** ”

GETTING INTO SHAPE

IN ANY SURVIVAL SITUATION you're likely to be at your best, both physically and mentally, in the moments before the situation occurs. From that point on, through lack of sleep, food, and water, your situation will deteriorate until your rescue. Being in good physical condition will help you overcome the challenges you'll face in a survival situation.

KNOWING YOUR LIMITS

Having an idea of how fit you are and knowing how much you are capable of achieving before entering the wilderness means you can set yourself realistic goals—which will help to keep you out of danger.

THE BENEFITS OF EXERCISE

Over time, exercise induces changes such as weight loss, as well as improved posture, physique, strength, agility, mental alertness, and stamina. These are all vital attributes should you find yourself in a survival situation.

HOW FIT ARE YOU?

A fit heart pumps blood more slowly and efficiently than an unfit one. The hearts of women, children, and older people beat faster than those of young adult males. Take your pulse first thing in the morning—this is known as the "base rate." As the table below shows, the faster your pulse returns to its base rate after a period of exercise, the fitter you are. Note that < denotes "less than" and > denotes "more than".

AGE	20-29	30-39	40-49	50+
RATING	BASE RATE			
MEN				
Good	<69	<71	<73	<75
Average	70-85	72-87	74-89	76-91
Poor	>85	>87	>89	>91
WOMEN				
Good	<77	<79	<81	<83
Average	78-94	80-96	82-98	84-100
Poor	>94	>96	>98	>100

AGE	20-29	30-39	40-49	50+
RATING	BEATS AFTER EXERCISE			
MEN				
Good	<85	<87	<89	<91
Average	86-101	88-102	90-105	92-107
Poor	>101	>102	>105	>107
WOMEN				
Good	<93	<95	<97	<99
Average	94-110	96-112	98-114	100-116
Poor	>110	>112	>114	>116

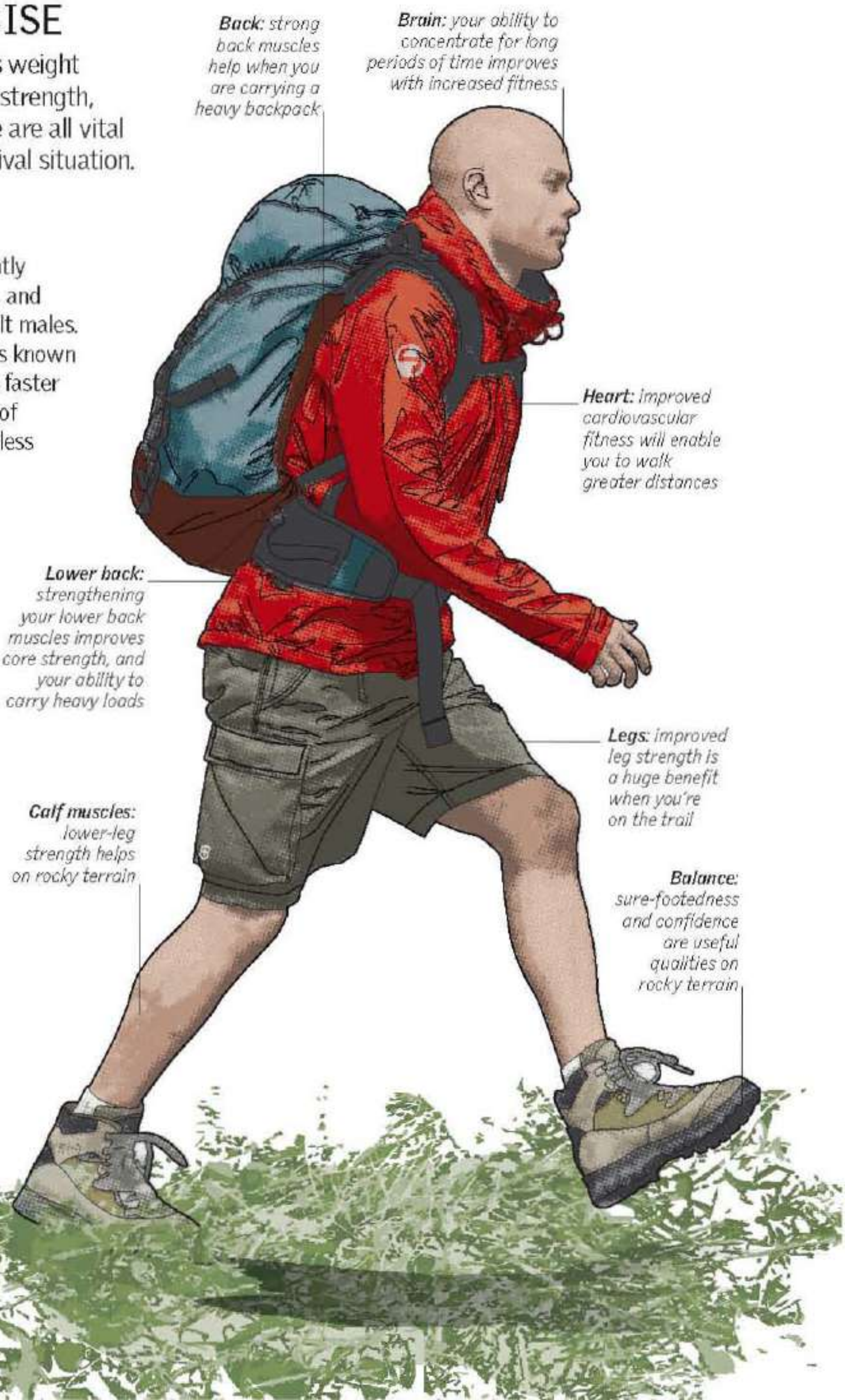
Back: strong back muscles help when you are carrying a heavy backpack
Brain: your ability to concentrate for long periods of time improves with increased fitness
Heart: improved cardiovascular fitness will enable you to walk greater distances
Lower back: strengthening your lower back muscles improves core strength, and your ability to carry heavy loads
Legs: improved leg strength is a huge benefit when you're on the trail
Calf muscles: lower-leg strength helps on rocky terrain
Balance: sure-footedness and confidence are useful qualities on rocky terrain

FIT FOR THE CHALLENGE

Survival fitness is not about trying to run the 100 meters in 10 seconds; it's about stamina and endurance, and about knowing your physical limitations and being able to work with them. It is also about understanding that you may have to push yourself beyond your limits and that having a positive mental attitude is paramount to survival—the mind will often give up long before the body has reached its limit. Bear in mind that every task in a survival situation will sap your energy.

EXERCISING REGULARLY

Any effective training schedule must include at least three 45-minute sessions per week, with exercises that raise the heartbeat to more than 120 beats per minute.

STARTING A FITNESS PROGRAM

To spend time in the wilderness with a pack on your back requires a combination of strength and aerobic fitness. Training in the gym before you set off will strengthen your heart, lungs, and leg muscles, and will increase your stamina levels when you are out on the trail.

- Seek expert instruction and guidance wherever possible.
- Start gradually and then build up your routine progressively.
- Never try to work through injuries—rest and seek medical advice.
- Devise a relevant fitness program; there are many websites and organizations that provide you with detailed fitness programs for various levels of specific activities, from training for a trek through the jungle to mountain-biking along the Appalachian Trail.
- Duplicating in training what you'd like to be able to achieve on the trail will allow you to build up reference points about how your body works, and how it copes in different situations. The more you know about how you perform, the better equipped you'll be on the trail.

WARMING UP AND COOLING DOWN

Time spent warming up and cooling down after exercise will help to improve your endurance levels and will also accelerate the recovery process. Get into the habit of starting and ending your session with a five-minute jog.

GO FOR A CHECK-UP

It's a sad fact that the majority of us don't have regular medical check-ups. Many people work on the principle that you only need to see a doctor or dentist if something is wrong. However, you should always make sure that you start any adventure or wilderness trek in top condition. Therefore, before you set out, pay a visit to your doctor and dentist and make sure that your body and teeth are in good condition.

VISIT THE DOCTOR

- Let the doctor know of any ailments that have been a concern to you during the course of the previous year.
- Tell the doctor where you are going and enquire about relevant inoculations or medication that you should be taking with you.

VISIT THE DENTIST

- Have any problems with your teeth dealt with before you go. Any minor dental problem will almost certainly turn into full-blown toothache when you least want it to.

USEFUL EXERCISES

If you already get regular exercise, you'll find yourself far more comfortable in a survival situation than your more sedentary counterparts. Any pre-expedition training regime should include plenty of stretching, aerobic exercise, and weight-training.

STRETCHING YOUR UPPER BODY AND BACK

Regularly stretching your arms, neck, chest, and shoulders will help your body to maintain a strong core. This will be of great use when it comes to activities such as rock climbing or using trekking poles. Undergoing a regular back-stretching routine will loosen the back muscles, making them more flexible and less susceptible to injury, and will increase both the back's range of motion and its endurance.

STRETCHING YOUR LEGS

Because your leg muscles bear the brunt of the work, leg stiffness is a common complaint at the end of a long day on the trail. Stretching your legs will improve your flexibility, increase blood circulation, and relax your muscles. Concentrate on your calves, quads, and hamstrings.

AEROBIC EXERCISES

Also called cardiovascular fitness, aerobic fitness refers to the ability of your heart, blood vessels, and lungs to supply oxygen and nutrients to the rest of your body during sustained physical activity. Regular aerobic exercise—such as swimming, jogging, or cycling—will reduce the risk of developing coronary heart disease or high blood pressure, will help you to manage your weight and increase your stamina levels, and will make your heart stronger and more efficient, thus improving blood flow around your body.

WEIGHT-TRAINING

Increased muscle strength will enhance your ability to perform everyday tasks such as lifting, carrying, and walking. A regular weight-training program will improve your posture, build muscle density around the joints and bones, improve your sense of balance, help your stress management abilities, aid sleep, and reduce the chances of you sustaining an injury. In addition, research has shown that a regular resistance-training program can improve your metabolic rate by up to 15 percent, which will help you shed any unwanted weight at a faster rate.

MENTAL PREPARATION

REGARDLESS OF WHETHER you're backpacking through a remote area, or on a day hike in familiar territory, a situation may arise that changes your circumstances for the worse. Very quickly you move into the unknown, which causes tremendous psychological and emotional stress, known as "psychogenic shock." Understanding this will help you to deal with it better and reduce its impact.

INDIVIDUAL REACTIONS TO DISASTER

People react to survival situations in different ways, although you can expect to find some common emotional reactions in victims who are experiencing, or have survived, a disaster situation. You may experience one or more of them during or after any survival experience or trauma.

YOUR RESPONSE TO DISASTER SITUATIONS

Your psychological response to a survival situation is crucial. Statistics show that 95 percent of people who die with psychological trauma die within the first three days. Losing the will to survive—or suffering psychological disorders that prevent you from coping with the physical conditions—is your main concern. If you break down psychologically, your chances of overcoming a situation will be compromised.

PSYCHOLOGICAL PROGRESSION

It's useful to examine how people are likely to react in a survival situation. Using this knowledge, it's possible to prepare mentally for such eventualities and, in so doing, lessen the impact if the worst should happen to you when out on the trail. Normal psychological reactions to disaster tend to occur in a set pattern of four stages: the pre-impact period, the impact period, the recoil period, and the post-trauma period (see right). Contrary to popular belief, people don't normally panic, although it can be contagious if someone does.

AGGRAVATING FACTORS	HUNGER	THIRST
Reactions to disaster can result from a direct blow to the psychological system, such as extreme shock, but they can also be brought on, or aggravated by, other factors. As with all psychological problems, knowing what these aggravating factors are, and attempting to avoid them—or at least recognizing what they might lead to—will maximize your chances of preventing or overcoming the problem. The most common aggravating factors are: hunger, thirst, fatigue, seasickness, and hypothermia.	Initially, hunger is not a problem, but a long-term lack of food will cause psychological changes to occur. Symptoms include: ▪ Apathy ▪ Irritability ▪ Depression ▪ Lack of concentration	Thirst is a serious problem, especially for survivors at sea or in the desert, and its effects are more acute than hunger. Agitation is commonplace; other symptoms include: ▪ Irrational behavior (see box, right) ▪ Delusions ▪ Visual hallucinations

COPING STRATEGIES	TRAINING	MOTIVATION
There are many things you can do to prepare yourself psychologically for a survival situation—such as learning about what you should expect if the worst were to happen—and to cope better if you find yourself in one. As with all survival skills, prior knowledge is power, and will help you to deal with a survival situation far more successfully. The main areas to think about are: training, motivation, attachment, hope, acceptance, and helping others. Developing coping strategies is an important technique for survival.	People who are properly prepared, who know their environment and how to use their equipment, and have an understanding of what to expect in a survival situation, will be far more effective if they find themselves in one. Adequate training and practice using your equipment will help you to function effectively at an automatic level. Prior knowledge is key to your survival.	Often known as "the will to survive," motivation involves a refusal to accept death, and to hang on to the belief that you were not meant to die under these conditions. It involves overcoming the emotional and physical discomforts of extreme conditions. Linked to motivation is the ability to establish goals, work out the steps to those goals, and to follow those steps through.

PANIC

Panic arises from the fear of what might happen rather than what has happened. It tends to occur when people are trapped, or if there is a time limit to their escape.

DEPRESSION

Depressed people will sit among chaos and debris vacantly gazing and not replying to questions. They're unaware of their situation and unable to help themselves, so risk further injury.

HYPERACTIVITY

Hyperactive victims are easily distracted, and are full of chatter, ideas, and often unhelpful suggestions. Sufferers can reach this stage after a state of depression.

ANGER

Aggression, anger, and hostility are common reactions to trauma. They're often irrational and may even be directed at the rescuers or medical staff trying to help them.

GUILT

Some sufferers feel guilty for surviving, and for not having done enough for others—and some irrationally blame themselves for bringing about the incident.

SUICIDE

Disaster victims have been known to commit suicide immediately after being rescued, in some cases when they're already safe in the hospital. Victims should be closely monitored.

Pre-impact period

The "pre-impact" period is divided into two stages:

- Threat: danger exists but, though obvious to those who recognize it, those who will not accept it respond with denial and under-activity.
- Warning: threat of danger is now apparent to all; response is now likely to be over-activity.

Impact Period

This is the life-threatening stage. Statistically, individuals behave in one of three ways:

- 10–20 percent of people are calm and retain full awareness.
- Up to 75 percent of people are stunned, bewildered, and unable to react rationally.
- 10–25 percent exhibit extreme behavior, such as screaming.

Recoil period

This follows on directly from the impact period; for example, victims may have escaped a sinking ship and are in the liferafts. It can last for up to three days, but generally lasts for around three hours. In most cases, it is characterized by a gradual return to normal reasoning abilities, awareness, and emotional expression.

Post-trauma period

If the recoil period is not fully successful, individuals may develop psychiatric disorders. The full impact of the incident becomes apparent and a range of emotions—guilt, depression, anxiety, aimlessness, and a feeling of bereavement—may develop. These are often referred to as Post-Traumatic Stress Disorders (PTSD).

FATIGUE

In many cases, physical exhaustion is present from the outset. At other times, it may result from sleep deprivation and the physical hardship endured over time. Most survivors agree that fatigue overwhelms them, but when they want to sleep they can't—they have an inability to relax. Fatigue causes a deterioration in mental and physical performance, followed by a psychological and physical debility.

SEASICKNESS

Seasickness often brings about an overwhelming desire to curl up and die, which in survival situations can easily become a reality. It's important not to give in to this urge. Fight seasickness with the following methods:

- Keep a fixed point such as the horizon in sight.
- Take small sips of water (not salt water) if you have sufficient supplies—but ration them if you're in a life or death situation.

HYPOTHERMIA

Hypothermia (see p. 273) produces both physical and psychological effects—the psychological consequences occur early in the condition, and cause:

- Loss of concentration
- Loss of memory
- Motor impairment
- Faulty decision-making
- Irrational behavior

IRRATIONAL BEHAVIOR

Irrational behavior can take many forms. Examples include the earthquake victims who were found collecting flowers instead of helping the injured, and famously, the band of the *Titanic*, who played while the ship sank rather than trying to save themselves.

ATTACHMENT

One of the strongest motivating forces for survival is the desire to be reunited with principal figures of attachment in your life. These may include:

- Husbands
- Wives
- Partners
- Children
- Grandchildren
- Close friends

HOPE

To hope means to entertain ideas that a distressing situation will improve and get better. In any survival situation it's important to cling onto hope, despite information or perceptions to the contrary. Thinking positively will help to ward off psychological trauma. It's often easier to be optimistic in a group situation than it is if you're on your own, as people can support each other.

ACCEPTANCE

An inability to accept one's situation or condition leads to frustration, anger, and irrational behavior, and it's important to avoid these feelings in a survival situation. The ability to accept the situation doesn't equate to giving in to it—far from it. Those who have this ability, and know when to be active and when to be passive often have a better chance of long-term survival.

HELPING OTHERS

First, monitor your own condition and check that you're really up to the task. Determine who is genuinely disturbed rather than showing "normal" reactions. Psychological first aid is only required for those who are failing to recover. Simple words of comfort and interest will make the majority who are numbed more responsive. Those who are disturbed should be monitored closely. Avoid giving sedatives.

PLANNING YOUR JOURNEY

WHETHER YOUR TRIP involves a day out with your backpack, or a full expedition over weeks, crossing continents in four-wheel-drive vehicles, you need to plan it very carefully. The plan for your day out won't be as detailed as it would be for the longer expedition, but it's equally important. It's a good idea to have a basic planning outline for your most regular trips, to which you can add supplementary information as your trips become more involved.

MINIMIZING THE "IF ONLY"

There's no way that you can plan for every eventuality—there are just too many variables—but what you can do is look at the type of trip you intend to take, and ensure that if a situation arises you're not left wishing you'd done something differently. The time to minimize the chances of an "if only" situation is during the planning stage. Look at the potential problems and risks, plan to avoid them, and equip yourself with the knowledge and/or equipment to deal with them if they arise.

WARNING!
The social customs of a country are not the same all over the world. Many cultural differences stem from ancient cultural or religious pasts, and are taken extremely seriously. While ignoring or breaking some customs may lead only to embarrassment, breaking others can result in a penalty, punishment, or even imprisonment. Cultural differences may require females in your group to dress so that they don't show bare arms or legs, or they may dictate which hand you use to greet someone politely. Always research a country's customs thoroughly when planning your trip.

THE SIX P'S

Remember the six p's: **P**rior **P**lanning and **P**reparation **P**revents **P**oor **P**erformance. Research has shown that the longer and more complicated the trip is, the less likely there is to be a major "survival" situation. This type of trip is likely to be well organized, and potential problems will have been considered. This means that they can either be avoided, or that there will be mechanisms in place to deal with them. In many ways, simply having a good understanding of how to deal with a situation, and being able to interpret the basic principles of survival, can prevent a minor problem from escalating into a disaster. In a survival situation, it may be your knowledge, combined with your ability to improvise, that determines whether you become a survivor or a statistic.

ORGANIZATIONAL PRIORITIES

When planning for a trip always start with the most important things—known as the "show-stoppers." These are generally the things that would actually stop the trip from happening in the first place if not organized in advance. If you then work backward from the show-stoppers to items that would simply make the trip more comfortable, most other things will fall into place. The chart on the right lists the things that you definitely cannot do without.

SHOW-STOPPERS	
MONEY	■ You'll need enough to cover your needs plus extra for emergencies. ■ It must be in the correct currency for the country you are visiting. ■ Ensure your ATM card is set up for use in that particular country. ■ Make sure you have a secure way of storing your money.
PASSPORT	■ Your passport must be valid. If you need to renew it, do so well in advance. ■ Some countries require your passport to be valid for a number of months after your visit. ■ Keep your passport number in several places, such as inside your survival kit. ■ Keep your passport safe, and in a waterproof container, such as a Ziploc bag.
VISA	■ Research the regulations regarding visas relevant to the countries you're traveling to. ■ Find out how to apply, how far in advance you should apply, whether or not you can apply in the country, and what other documents you need. ■ As with your passport, always keep your visa safe, dry, and in good condition.
VACCINATION	■ Check the regulations for the country—many have strict vaccination policies. ■ Ensure you have the vaccinations, and any boosters, within the correct timeframe. ■ Some inoculations last only for six months, so for extended trips you might have to have more while away. This can usually be arranged through local hospitals or clinics.
TICKETS	■ Make sure you have the correct tickets for your journey. ■ Read the tickets and make sure they have the correct names, dates, and locations. ■ Should you need to prove your movements, always keep your tickets in a safe place and never throw them away—often the return journey is included on the same ticket.
INSURANCE	■ It's advisable to get insurance against your trip being cancelled. ■ Ensure that your insurance will cover your medical expenses should the worst happen. You don't have to be climbing Everest—even a twisted or broken ankle when hill-walking could prove to be very expensive if you don't have adequate insurance.

YOUR TEAM

If you're embarking on a trip with a group, remember that team dynamics will play an important part in the success, or otherwise, of your trip. Stressful situations—and particularly survival situations—can bring out the best, or the worst, in people. When planning a long trip, it's always a good idea to plan several shorter trips beforehand, to use as practice sessions. These will not only help you decide what equipment to take, and provide you with an opportunity to practice using it, but will also allow team members to assess how they work together as a group. This can help you to organize your team more effectively.

MIXED-GENDER GROUPS

If the team has both male and female members, it's important to factor this into the planning stage. You'll need to consider sleeping and washing arrangements, who carries what, who is responsible for what, and so on. All this should be planned in detail before you set off. It should also be noted that taking it for granted that the women will cook the meals while the men make the shelters is not the best way to start an expedition.

MIXED-AGE GROUPS

When planning your trip, remember that different age groups may have different levels of fitness, which will affect the pace of your group. However, what older members may lack in pace, they may make up for in expedition experience.

PREPARATION THROUGH TRAINING

Training yourself both mentally and physically—and practicing with the equipment you'll be using—may seem like the obvious thing to do, but in many cases the obvious can often get overlooked. You'll get the most out of your trip if you're mentally and physically prepared to a level that means you can operate within your own capabilities and comfort zone. This will allow you to enjoy and appreciate the experience, as opposed to just getting yourself through it.

TRAINING YOURSELF

Duplicate in training what you intend to do on your trip. Build up your training gradually, over a period of weeks and months, and take into account the following:

- **The environment**: research the weather conditions you'll be facing, and look at the extremes as well as the average. For example, desert areas may be hot during the day, but can drop to below freezing at night.
- **Weight:** increase the amount of weight you carry until you're eventually carrying what you intend to take. This will not only condition you to the weight but also help you to decide just what's important to take and what's not.
- **Distance:** if your trip involves covering a certain distance a day, then train for that distance. This will give you an indication of whether it's achievable and sustainable.
- **Language**: if visiting a country in which your native language is not widely spoken, it will be beneficial if you can learn some useful phrases. Take a phrase book or an electronic translator.

TRAINING WITH EQUIPMENT

Use your equipment as much as you can, and find out the best way to operate it through practice under realistic conditions (see box, below). This will highlight its strengths and weaknesses, and allow you to determine both its capabilities and your own—the compass may be working perfectly, but practice may highlight the fact that you're not confident enough using it, in which case you'd need to do further training before you set off on your expedition. The wrong time to be trying to find the jack and spare wheel on your rented vehicle is at night, in the rain, with the mosquitoes looking at you as their next free meal. Before you set off, always think about the various skill demands required, and ensure that you're capable of addressing those demands.

OPERATING UNDER REALISTIC CONDITIONS

When training with new equipment, always practice using it under realistic conditions. For example, if you're going to be using your GPS in cold conditions, can you operate it using the gloves you have selected? If you're pitching a tent, have you got **all** the necessary components, and can you put it up in the dark and rain?

EMERGENCY PLAN OF ACTION

UNFORTUNATELY, EVEN THE MOST meticulously planned, thoroughly equipped, and best-executed trip can run into difficulties. Unforecasted strong winds could trap you and your kayak on an island overnight, or a sprained ankle could leave you unable to climb down the rocks you climbed up so easily. Each of these survival scenarios are difficult to predict, but could easily happen.

USING YOUR EPA

An EPA should contain up-to-date information about you as a person. This may change very little over the years, so you may only need to create a standard EPA and just update the details specific to each trip.

RUNNING INTO PROBLEMS

It's important to make sure that you have done all you can to help yourself, and help any rescue attempt that might take place if you do have any problems. Always remember that there are two sides to any survival or rescue situation: the part that you play and the part that the rescue services play. However, the effectiveness of the rescue services can be greatly increased when they have access to all relevant information—in many cases this arrives too late. It's vital to keep people informed of your intended whereabouts, so that they can raise the alarm if you deviate from your plans.

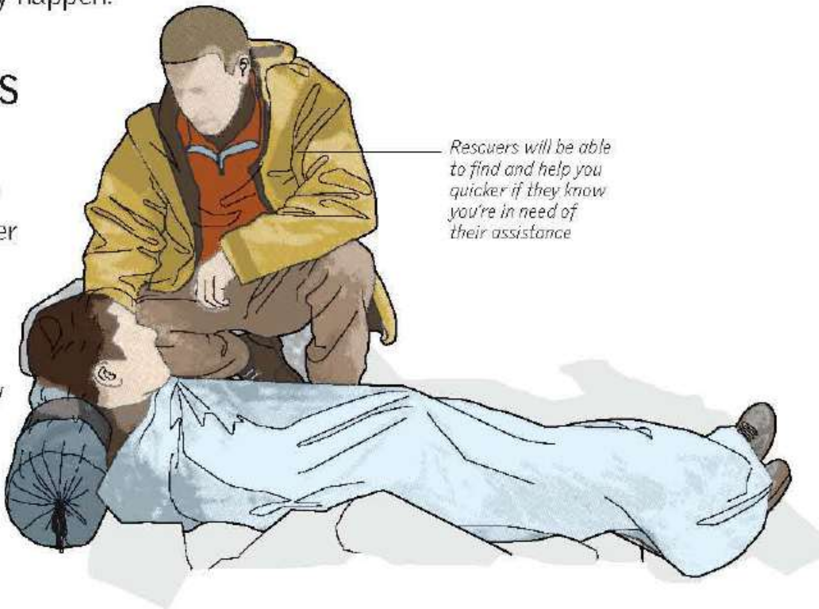

Rescuers will be able to find and help you quicker if they know you're in need of their assistance

WORST-CASE SCENARIO

In the military, every mission that's undertaken, particularly in a theater of conflict, has plans in place for the worst-case scenario. Each part of the mission is meticulously planned, and the team completes a form that states what their basic intention will be should anything happen during the various stages.

IF THE WORST HAPPENS

Having planned for the worst, should the team find themselves in difficulty, the rescue group will not have to try to second-guess what the team will do—they will have a clear indication of the team's intent and can plan effectively and decisively around this information. The team will be found quicker as a result. It's a good idea to apply this principle to your own trip.

KEEPING PEOPLE INFORMED

A good way of keeping your friends, family, and the relevant rescue services informed of your whereabouts is to write down the details of your trip, including pertinent itinerary places and dates, so that if you don't reach your destination when planned, the alarm can be raised. Similar to the "worst-case scenario" principle practised by the military (see left), an "emergency plan of action" (EPA) should be prepared by anyone venturing into the great outdoors (see right). Give a copy to your next of kin and group members, and keep a copy on your person. Where applicable, inform local services, such as park and ranger stations, of your intentions. Remember to inform these people when you arrive safely at your destination.

WRITING AN EMERGENCY PLAN OF ACTION (EPA)

The best way to determine what should be included in your plan is to look at the worst-case scenario you could find yourself in, and ask yourself what information your next of kin would need to know about you and your intentions if they felt they had to raise the alarm. In addition, if you did go missing—particularly in another country—there are many things that the relevant rescue services would need to aid their efforts, for example a recent photograph, your passport details, what equipment you have with you, what languages you can speak, and your skills. The clearer their understanding of you, your abilities, and your intentions, the easier their job will be. In addition, the more information your next of kin have, the more proactive they will feel in your rescue.

EMERGENCY PLAN OF ACTION FORM

Full name as appears on passport:
John William Smith

Date of birth:
(dd/mm/yy)
05/28/60

Height: *5′10″ (178cm)*
Weight: *168 pounds (76 kilos)*
Hair color: *Brown*

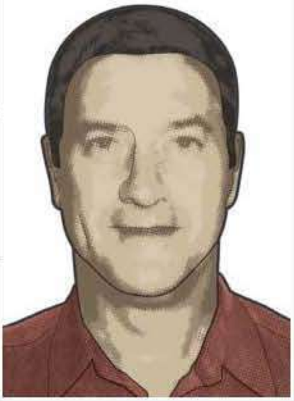

Passport number:
2008XXXXX63
Expires:
11/03/13

Driver's license number:
JHY280771smit
Expires:
12/28/22

Distinguishing marks (scars, tattoos):
Small scar—center of forehead
Large scar—right hand, middle finger
Chinese symbol tattoo—right arm

Languages spoken (fluent/basic):
English—native
French—basic
German—basic

Medication:
Medication—anti-malaria tablets

Allergies:
Amoxicillin

Swimmer: *Strong swimmer*
Outdoor skills/experience:
Attended basic military survival training.
Attended basic bushcraft course
Experienced in living outdoors

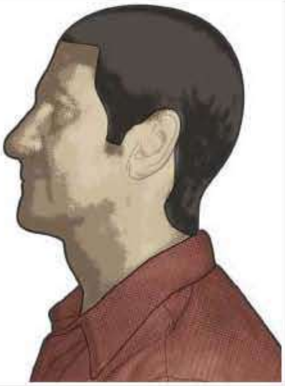

Next of kin 1: *Father*
William Smith
1018 Furlong Avenue
Brunswick, Maine,
USA 04555

Next of kin 2: *Brother*
Andrew Smith
1023 Parkglen
Ashford, Kent,
TM24 5HZ
UK

Tel: *(001) 55 555 2356*

Tel : *(0044) (0)155 555 2357*

Email: *willsmith@internet.com*

Email: *andrewsmith@internet.com*

Trip details:
Campsite 1 = Grid ST456654
Campsite 2 = Grid ST654987
Vehicles: Landrover 1 = white, reg MH55 555
Landrover 2 = blue, reg MH56 555
Group = Ben Jones, Kim Smith, and myself
Day 1: Park Landrover 2 at Campsite 2, and drive in Landrover 1 to Campsite 1
Day 2: Follow well-defined path along the Derwent Line Trail, aiming to camp overnight at Grid 4561559
Day 3: Continue along the Derwent Line Trail, aiming to be at Campsite 2 by mid-afternoon. Camp overnight at Campsite 2
Day 4: Travel in Landrover 2 to campsite 1 and retrieve Landrover 1

Foreseeable problems/intentions:
Day 1: None
Day 2: None but will use Ranger Station 18, grid 555555 (Tel. 666 6666) as an emergency rendezvous point
Day 3: None but will use Ranger Station 19, grid 666666 (Tel. 555 5555) as an emergency rendezvous point
Day 4: None

Communications plan:
Will speak to Dad on the morning of Day 1 and try to phone during the trek but am unsure of cell reception once on the trail so don't worry if you hear nothing.
Will phone Dad again when we reach campsite 2 on day 3.

My mobile: *07979 555555*
My email: *jws@internet.com*
Alt. No: *Ben 05555 555555*
Alt. No: *Kim 05555 555555*
Alt. No: *Campsite 1 555 555 55555*
Alt. No: *Campsite 2 555 555 55555*

Date: *November 23, 2009*

KNOW YOUR ENVIRONMENT

The continued survival of the human race can be attributed to our ability to adapt to our environment. While we may have lost some of our ancient ancestors' survival skills, we have, in their place, learned new skills as and when they have become necessary. The issue we face today is that the gap between the skills we once had and the skills we now have grows ever wider as we rely more heavily on modern technology. Therefore, when you head off into the wilderness, it is important to fully prepare for the environment.

Before a trip, research how the native inhabitants dress, work, and eat. How they have adapted to their way of life will help you to understand the environment and allow you to select the best gear, adopt the best techniques, and learn the correct skills. This is crucial given that most survival situations arise as a result of a sequence of events that could have been avoided—while you may have no control over the aircraft crashing, for example, you can recognize a change in the weather and choose whether to continue or turn back.

In this section YOU WILL DISCOVER...

- the difference between **tundra** and **taiga...**
- why you should **step-up** the size of your **survival footprint...**
- the **way to go** when it comes to **snowy peaks** or **swampy creeks...**
- how to avoid **getting lost** in the **permafrost...**
- the best way to **stay at the top** of the **food chain...**
- how to **feel at home** where the tropical **butterflies roam...**
- how a high-visibility **survival suit** could really **get you noticed...**

The four basic principles of survival are:

protection, location, water, and food. In most survival situations, this is also the order in which you should prioritize them.

P **PROTECTION** You must stay in a condition that allows you to be proactive in your continued survival and rescue. Physically, you should protect yourself against injury, the elements, and wildlife. Mentally, you should protect yourself against emotions that could rob you of the will to live: fear, guilt, despondency, and depression, for example. The best way to achieve this level of protection is to light and maintain a fire. Not only does it offer physical protection against the elements and wildlife, but it also provides a sense of security and familiarity that can help normalize even the most dire situation.

L **LOCATION** Your second priority is to recognize the importance of your location to your chances of survival and rescue. You will usually have two options: stay or go. Your preferred option should be to remain where you are and use anything at your disposal to mark your location to aid rescuers in their efforts to find you. If you can't stay where you are (perhaps because you are in imminent danger) you may have no option but to move to another location that provides either a better chance of survival or rescue, or both. Select a location aid that offers you the best chance of attracting attention in the environment in which you're traveling.

W **WATER** Put simply, water is the essence of life. While you may be able to survive for a few days without it, your ability to function and carry out even simple mental and physical tasks will be dramatically reduced in less than 24 hours. However, if you are injured, if the weather conditions are very hot, and if your workload is particularly heavy, for example, your survival time without water could be reduced to just a few hours. You should learn how to procure water in the particular environment in which you're traveling, and understand the ways in which a lack of water affects you.

F **FOOD** The importance of food is directly related to the length of time you are in a survival situation: the longer the situation lasts, the more important food will become in helping you stay fit and healthy. Even with a moderate workload, going without food for five to seven days will not kill you. You will, of course, feel hungry, you will grow tired, your movements will slow, and your body will lose its ability to repair itself. However, unless you're malnourished before you enter a survival situation, you are unlikely to starve to death within a week. Your body needs water to digest food, so always remember to prioritize water over food.

> “We rely so heavily on **modern technology**: water at the **turn of a tap**; heat at the **flick of a switch**; food raised, harvested and **prepared by others**”

TEMPERATE ENVIRONMENTS

THE TEMPERATE ZONES are the two regions between the Tropic of Cancer and Arctic circle, and the Tropic of Capricorn and Antarctic circle. Typified by seasonal variations in climate, conditions can include baking hot summers, freezing cold winters, and rainfall all year round. Landscape features range from forests and snow-covered peaks to grasslands and deserts. Although most temperate regions are highly populated, don't be lulled into a false sense of security—a worst-case scenario can still arise just a few miles from help.

TEMPERATE FEATURES

Although they contain a wide range of environmental features, the temperate zone's typical feature is forest, ranging from deciduous trees that shed their leaves in fall to coniferous trees that retain their leaves all year round. Grassland predominates where forests have been cleared, while highland areas contain hills and mountains. The abundance of rain means that rivers and lakes are common, and swampy wetlands form in areas with poor drainage.

WARNING!

Brown bears inhabit extensive wilderness areas in the northern temperate zone, and are the most dangerous temperate animal. The Grizzly, found in North America, is the most likely to attack humans.

TEMPERATE ESSENTIALS

Climate and terrain can vary widely in temperate areas, so preparation for a range of eventualities is essential:

- Don't underestimate the temperate environment. Although the vast majority of the human race lives in temperate zones, the diversity of terrain and weather means that survival equipment and knowledge must be broad enough to cope with a wide range of situations and conditions.
- Temperate weather can change very quickly, so check the local forecast before you set off, and carry a small AM/FM radio so you can listen to local weather reports.
- Plan a realistic and achievable route, and prepare an EPA (see pp. 24–25). Be ready to re-assess your route during the trip.
- Take clothing for the full range of conditions you may encounter.
- Carry a survival kit (see pp. 60–61), knife, emergency equipment, cell phone, and first-aid kit (see pp. 260–61)—and know how to use them.
- Carry adequate water, and equipment to collect and purify more if necessary.
- Carry some form of basic shelter, even if only going out for the day.
- Always take a map and compass, and consider using a GPS as an aid.

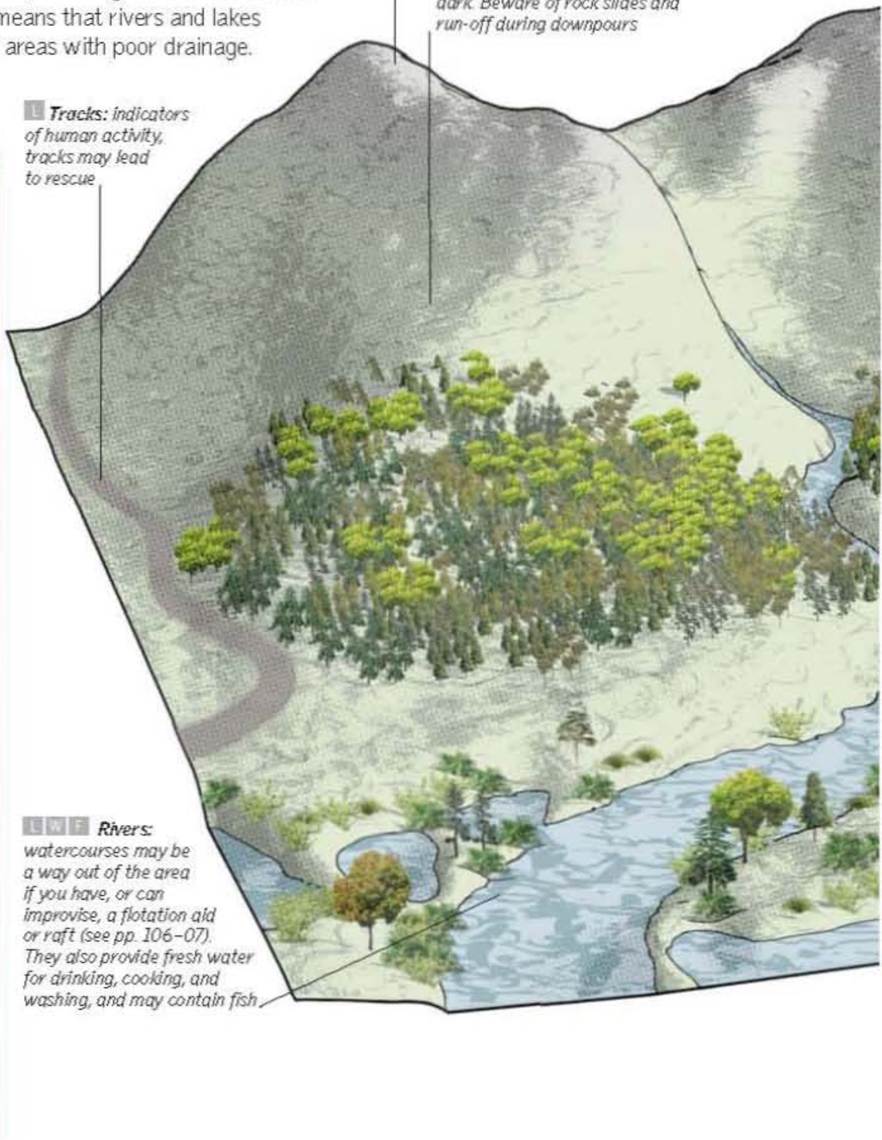

L *__High ground:__ elevated areas afford a view of your surroundings and may indicate which area offers the best prospects for your survival, or where to go to find rescue*

P L *__High ground:__ night-time temperatures are lowest at altitude, so descend to warmer levels before dark. Beware of rock slides and run-off during downpours*

L *__Tracks:__ indicators of human activity, tracks may lead to rescue*

L W F *__Rivers:__ watercourses may be a way out of the area if you have, or can improvise, a flotation aid or raft (see pp. 106–07). They also provide fresh water for drinking, cooking, and washing, and may contain fish*

WHERE TO FIND TEMPERATE AREAS

Deciduous forests are scattered throughout the temperate zones, but the largest occur in eastern North America, western Europe, and east Asia. Extensive coniferous forests are found in the higher latitudes of North America and Eurasia, while grassland is most common in continental interiors.

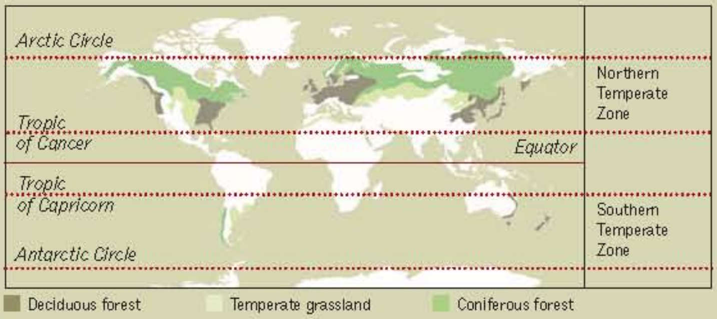

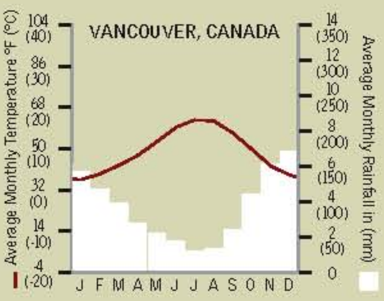

TYPICAL CLIMATE

Vancouver lies on the Pacific coast of Canada, and experiences a typical temperate climate of warm summers and cool, wet winters.

SURVIVING IN TEMPERATE AREAS

Most temperate environments have a relatively mild climate and good natural resources, making them favorable places for long-term survival. Water can be found in most areas, wood for building shelters and lighting fires is usually abundant, and different types of edible plants and animals can be found at different times of year if you know where to look. Potentially the greatest threat is hypothermia (see pp. 272–73), a particular problem in cold, wet and windy conditions, especially in winter and at night, when temperatures tend to fall even further.

WILDLIFE AND PLANT LIFE

Plant and animal life can be abundant in temperate areas and represents both a valuable food resource and a potential hazard. Shoots and stems, leaves, roots, nuts, or berries of some plants are edible, but only when positive identification—or at the least, the Universal Edibility Test (see p. 206)—is achieved. Small mammals, birds, insects, reptiles, and fish can also be found, but may be difficult to catch and kill, and must be prepared and cooked properly. Wildlife can also pose a threat—snakes, spiders, and scorpions may bite and, though rare, bears, wolves, and cougars may attack if cornered.

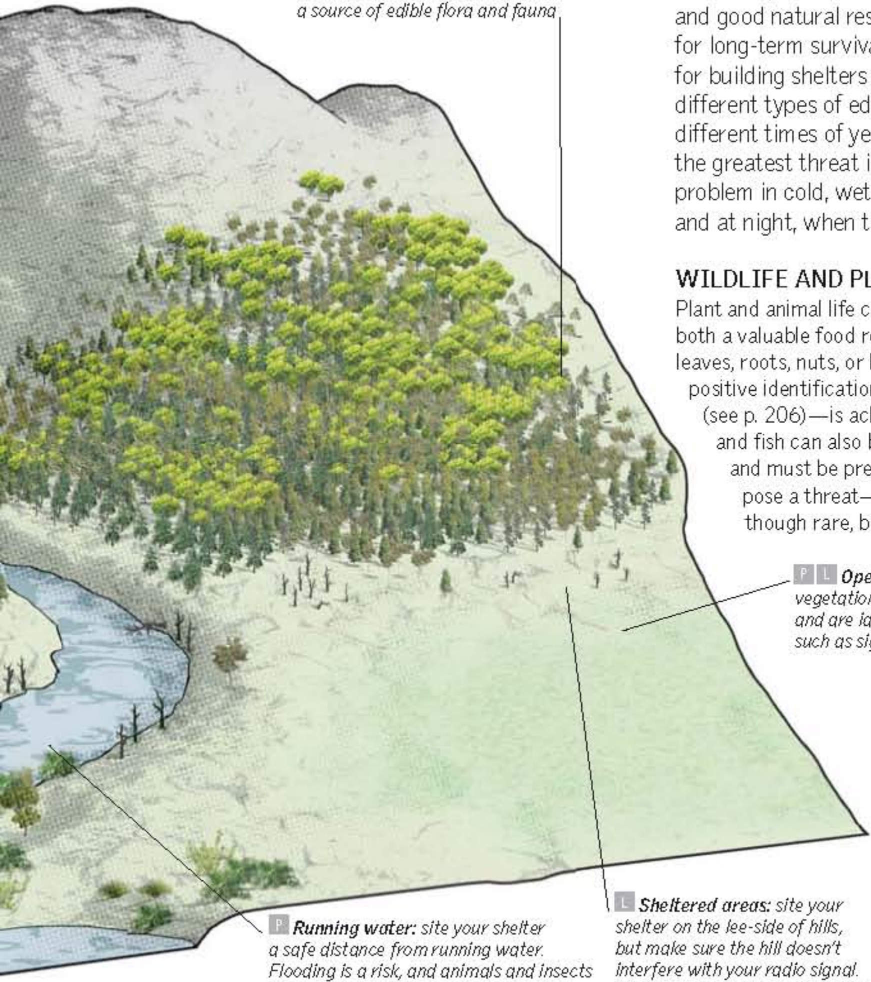

P L F ***Woodland:*** *tree cover affords protection from the elements, and wood for shelter, cooking, and signal fires. It is also a source of edible flora and fauna*

P L ***Open ground:*** *areas with no vegetation offer a clear view of dangers, and are ideal for deploying location aids, such as signal fires*

P ***Running water:*** *site your shelter a safe distance from running water. Flooding is a risk, and animals and insects are drawn to water. The noise may also hide the sound of animals or rescuers*

L ***Sheltered areas:*** *site your shelter on the lee-side of hills, but make sure the hill doesn't interfere with your radio signal. Choose a spot that faces the sun to maximize warmth and light*

FIND OUT MORE...

P **PROTECTION** **Shelters** pp. 156–65, 178–81 **Fire** pp. 118–33, 204–05 **Dangers** pp. 242–49, 300–05

L **LOCATION** **Navigation** pp. 66–77 **Movement** pp. 86–89 **Signaling** pp. 236–41

W **WATER** **Finding** pp. 188–91 **Purifying** pp. 200–01

F **FOOD** **Plants** pp. 206–07, 280–81 **Animals** pp. 208–13, 216–29, 290–99

TROPICAL ENVIRONMENTS

THE TROPICAL ZONE stretches from the Tropic of Cancer to the Tropic of Capricorn, and is centered over the Equator. A range of environmental features occurs in this area depending on the local climate, from lush, humid, biodiverse tropical rain forests to dry, more sparsely vegetated tropical scrub. With adequate preparation and caution you should be able to survive unsupported for extended periods—the jungle holds just as many resources as dangers.

WARNING!
Mosquitoes are responsible for more fatalities than any other creature in tropical areas. They carry a range of diseases including malaria and yellow fever, which kill millions of people every year.

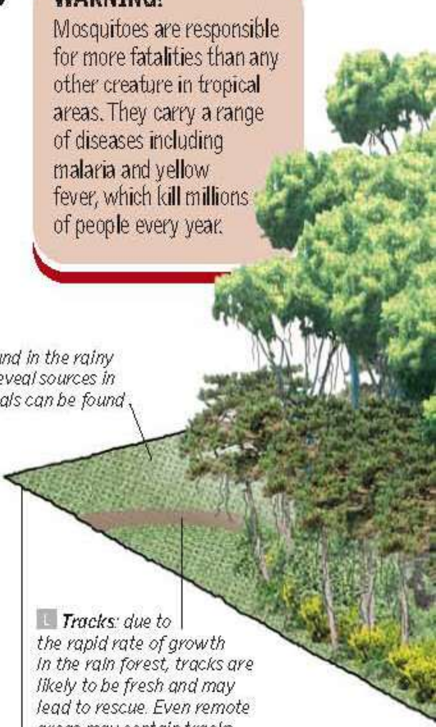

W F ***Tropical scrub****: water can be found in the rainy season, while animal movements may reveal sources in the dry season. Edible plants and animals can be found*

L ***Tracks****: due to the rapid rate of growth in the rain forest, tracks are likely to be fresh and may lead to rescue. Even remote areas may contain tracks left by loggers, prospectors, or local people*

P L ***Tropical scrub:*** *provides shade from the sun and materials for fire and shelter, but may be home to dangerous wild animals. Dense vegetation during the rainy season may impede visibility*

TROPICAL ESSENTIALS

Rain forests contain everything you need to stay alive, so remember the following tips when venturing out:

- While the jungle teems with predatory wildlife (such as big cats, crocodiles, and anacondas) that sees you as part of the food chain, it is the small creatures that can make life miserable. Most animals in the jungle want to avoid you as much as you want to avoid them—making a noise will scare most away.
- High humidity encourages infections, so keep everything covered (sleeves down, gloves on), and wash at every opportunity.
- Always sleep off the ground.
- Boil or treat all water. Some water found in plants can be safe to drink, but if in doubt, and if the water is yellow, milky, or very cloudy, treat before drinking.
- Many plants have defense mechanisms and can emit toxic liquids that may sting or burn—if you can't identify it, leave it alone.
- Dry tinder can be hard to find, so if you come across any, collect it and keep it safe and dry—lighting a fire in a moist, humid environment can be challenging.
- The smoke of a signal fire must penetrate through all the layers of the jungle, so light your fire where the canopy is sparse, such as the bend of a river or a clear-cut area.
- Navigation can be difficult, as you may only be able to see a few yards in front of you. Use dead reckoning—walking a short distance to a recognizable feature on your bearing—or pace-counting (see pp. 72–73).
- Rivers in the jungle usually run downhill to civilization and, eventually, the coast.
- Don't fight the jungle—tune in to its rhythm and work with it rather than against it.

TROPICAL FEATURES

Tropical rain forest, which occurs within 10 degrees of the Equator, is the predominant environmental feature in the tropics. But there are several other environment types just a few degrees of latitude away.

VERDANT RAIN FOREST

Tropical rain forests occur in areas with steady year-round temperatures and rainfall. Annual rainfall can reach around $6\frac{1}{2}$–10 ft (2–3 m) and daytime temperatures may reach 90°F (30°C), dropping to 70°F (20°C) during the night. Monsoon (or "seasonal") rain forests occur in areas with a wet and dry season, while montane rain forests, also known as "cloud forests," occur in mountainous areas.

SCRUB AND SWAMPS

Also known as "thorn forest," tropical scrub consists of low, woody plants with thorns, which usually grow in clusters separated by patches of bare ground (grasses are uncommon). Leaves are shed in the dry season, forming a dense, herbaceous layer in the wet season. Swamps are another common tropical feature, and may consist of fresh or saltwater. Freshwater swamps are found in low-lying, inland areas, and consist of masses of undergrowth, reeds, grasses, occasional short palms, and islands. Saltwater swamps often contain mangroves, and occur in coastal areas that are prone to tidal flooding so are best navigated by boat. Visibility in both types of swamp is poor, and movement is difficult.

SURVIVING IN THE RAIN FOREST

While natural resources are abundant in the rain forest, the heat, humidity, number of animals, and voracious rate of vegetal growth can make it an uncomfortable place. Water, materials for shelter and fire, and edible plants may all be found, although identification of plants is crucial to avoid poisonous species. Animal life is everywhere, so sleeping platforms must be built off the ground to avoid biting insects, snakes, and spiders. The greatest danger is becoming lost—navigation is difficult due to dense undergrowth at ground level, and rescuers will struggle to find you beneath the thick jungle canopy.

P L W F ***Rain forest:*** *fuel for cooking and signaling fires and material for shelter may be found, although check for deadfalls and other dangers. Water from rain and rivers, and edible plants and animals can be found*

FIND OUT MORE...

P **PROTECTION** **Shelters** pp. 158–61, 166–71 **Fire** pp. 118–33, 204–05 **Dangers** pp. 167, 242–49

L **LOCATION** **Navigation** pp. 66–77 **Movement** pp. 106–07 **Signaling** pp. 236–41

W **WATER** **Finding** pp. 192–93 **Purifying** pp. 200–01

F **FOOD** **Plants** pp. 206–07, 284–85 **Animals** pp. 208–13, 224–29, 292–99

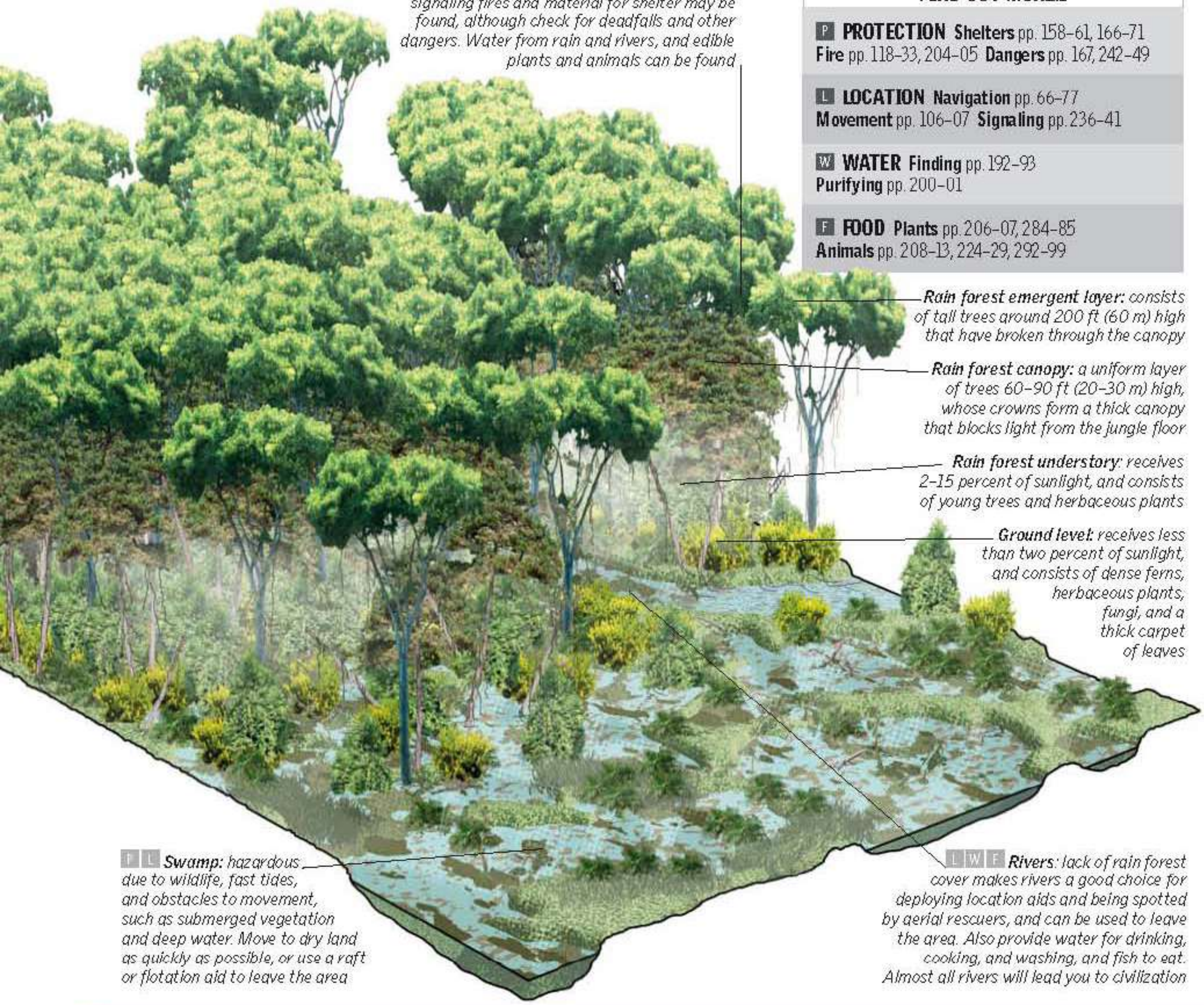

Rain forest emergent layer: *consists of tall trees around 200 ft (60 m) high that have broken through the canopy*

Rain forest canopy: *a uniform layer of trees 60–90 ft (20–30 m) high, whose crowns form a thick canopy that blocks light from the jungle floor*

Rain forest understory: *receives 2–15 percent of sunlight, and consists of young trees and herbaceous plants*

Ground level: *receives less than two percent of sunlight, and consists of dense ferns, herbaceous plants, fungi, and a thick carpet of leaves*

P L ***Swamp:*** *hazardous due to wildlife, fast tides, and obstacles to movement, such as submerged vegetation and deep water. Move to dry land as quickly as possible, or use a raft or flotation aid to leave the area*

L W F ***Rivers:*** *lack of rain forest cover makes rivers a good choice for deploying location aids and being spotted by aerial rescuers, and can be used to leave the area. Also provide water for drinking, cooking, and washing, and fish to eat. Almost all rivers will lead you to civilization*

WHERE TO FIND TROPICAL AREAS

Tropical rain forests are clustered around the Equator. The largest rain forest areas are found in Central and South America, sub-Saharan Africa, southeast Asia, northern Australia, and several Pacific islands.

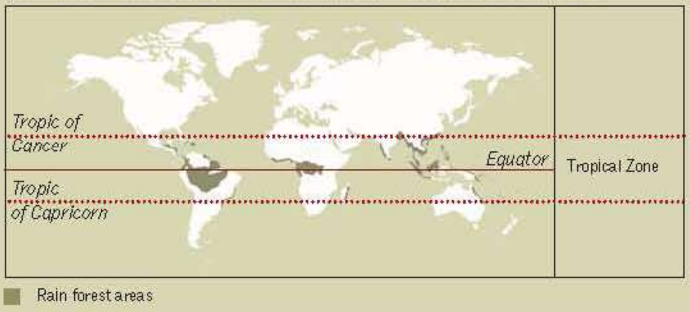

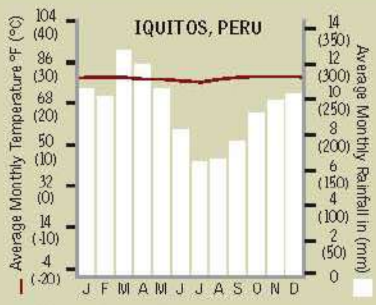

TYPICAL CLIMATE

Iquitos sits on the Amazon river in the heart of the Peruvian rain forest. It lies just south of the Equator and has a typical hot, wet tropical climate.

MOUNTAIN ENVIRONMENTS

THE MOUNTAIN ENVIRONMENT is one of the harshest places for humans to survive. Defined as landmasses with a summit above 2,000 ft (600 m), mountains can be very dangerous places due to their elevation and terrain. Lower temperatures and poor weather are more likely at higher altitudes, so there is a significant risk of hypothermia, frostbite, and altitude sickness, while snow, ice, and precipitous terrain present further hazards. Survival may depend on your ability to descend to areas with better prospects of survival and rescue.

MOUNTAIN FEATURES

Above the treeline, mountain terrain is typically barren, featuring bare rocks, gravel, boulders, and snow and ice, which are most extensive during the winter but may persist at high elevations during the summer. At lower levels, coniferous forest is the predominant feature, dissected by streams and rivers fed by snowmelt. Topographical features include scree slopes, cliffs, ravines, boulder fields, snowslopes, and glaciers.

ELEVATION CHANGE

Mountains of sufficient height may have vastly different ecosystems at different elevations, due to harsh conditions at high altitude. Plant and animal life varies at different elevations, so the natural resources available for survival will differ according to height. Mixed deciduous trees are common on low slopes and conifers grow at intermediate levels, but thin out to grass, isolated shrubs, moss, and lichen higher up. At the highest levels, high winds, frost, snow, and ice discourage any growth.

WARNING!

One of the most dangerous animals in mountain areas is the cougar, also known as the puma, panther, and mountain lion. Found in North and South America, more than 20 percent of attacks are fatal.

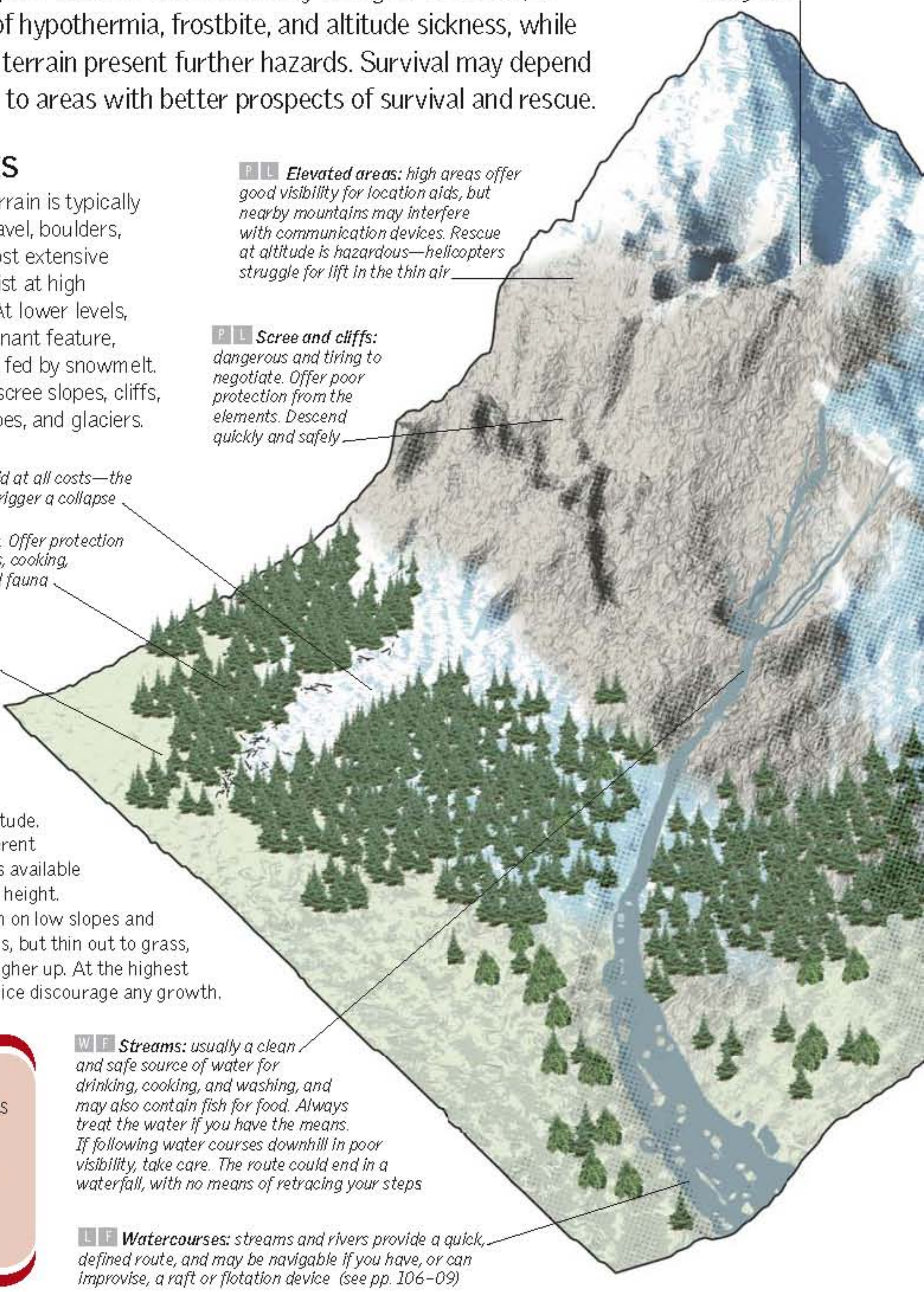

P L ***Mountains:*** *shelter is limited and dangers include unpredictable weather and rock falls. They may require special equipment and knowledge to negotiate*

P L ***Elevated areas:*** *high areas offer good visibility for location aids, but nearby mountains may interfere with communication devices. Rescue at altitude is hazardous—helicopters struggle for lift in the thin air*

P L ***Scree and cliffs:*** *dangerous and tiring to negotiate. Offer poor protection from the elements. Descend quickly and safely*

P ***Avalanche-prone slopes:*** *avoid at all costs—the smallest noise or movement can trigger a collapse*

P L F ***Forest areas:*** *rich in resources. Offer protection from the elements, material for shelters, cooking, fires, and signaling, and edible flora and fauna*

P ***Shelter:*** *always site your shelter away from areas of avalanche risk, especially if avalanche debris is present*

W F ***Streams:*** *usually a clean and safe source of water for drinking, cooking, and washing, and may also contain fish for food. Always treat the water if you have the means. If following water courses downhill in poor visibility, take care. The route could end in a waterfall, with no means of retracing your steps*

L F ***Watercourses:*** *streams and rivers provide a quick, defined route, and may be navigable if you have, or can improvise, a raft or flotation device (see pp. 106–09)*

SURVIVING IN THE MOUNTAINS

The prospects for survival are good at lower elevations, where trees provide material for shelter and fire, rivers provide water, and edible plants and animals are likely to be present. At higher elevations vegetation is scarce, so there are fewer options for shelter and food, although water from streams or snow is likely to be available. There may be a risk of avalanche on snow-covered slopes, and crevasses pose a hidden danger in glaciated areas. Cold-related injuries pose the greatest threat (see pp. 270–71).

HIGH-LEVEL HAZARDS

The primary threats to survival at high altitude are cold-related ailments, the most dangerous of which is hypothermia. This occurs when the body's core temperature falls below 95°F (35°C)—if allowed to drop as low as 86°F (30°C), it can be fatal. Frostbite may also occur in the extremities in freezing or cold and windy weather, and may lead to permanent tissue damage. Altitude sickness, which can cause pulmonary or cerebral edema—and in extreme cases, death—is also a risk above 8,000 ft (2,500 m).

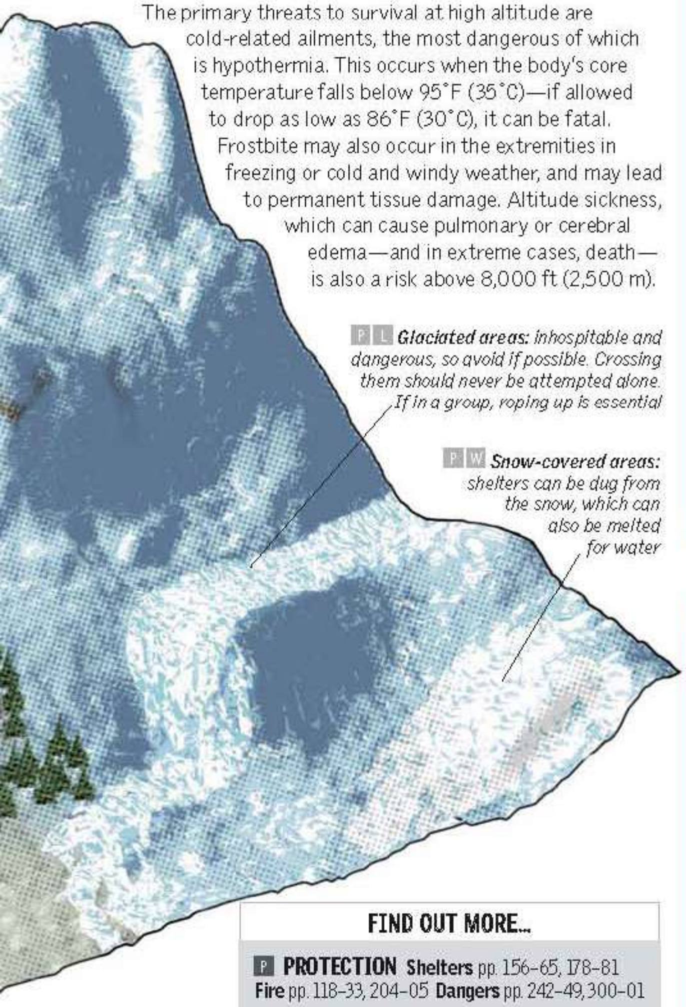

P L ***Glaciated areas:*** *inhospitable and dangerous, so avoid if possible. Crossing them should never be attempted alone. If in a group, roping up is essential*

P W ***Snow-covered areas:*** *shelters can be dug from the snow, which can also be melted for water*

FIND OUT MORE...

P PROTECTION Shelters pp. 156–65, 178–81 **Fire** pp. 118–33, 204–05 **Dangers** pp. 242–49, 300–01

L LOCATION Navigation pp. 66–77 **Movement** pp. 90–91, 94–97 **Signaling** pp. 236–41

W WATER Finding pp. 188–91, 194–95 **Purifying** pp. 200–01

F FOOD Plants pp. 206–07, 280–81, 286–87 **Animals** pp. 208–13, 216–23, 290–97

MOUNTAIN DISTRIBUTION

Mountains are found on every continent, but the major ranges are the Himalayas and Karakoram in Asia, the Andes in South America, the Rocky Mountains in North America, and the Alps and Pyrenees in Europe.

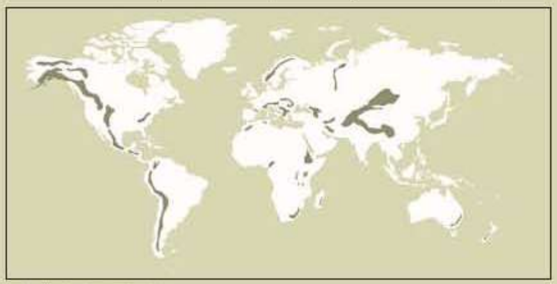

▨ Mountain ranges

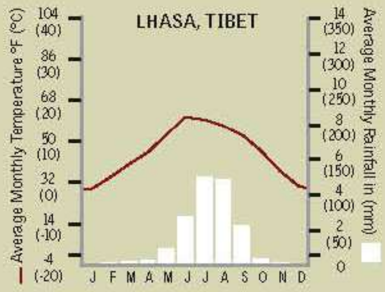

TYPICAL CLIMATE

Lhasa, the capital city of Tibet, lies at 11,972 ft (3,649 m) on the Tibetan Plateau, sometimes known as "the roof of the world." The high altitude means that winters are cool, and precipitation is relatively low because it mostly falls as snow, which has just 10 percent of the density of rain.

MOUNTAIN ESSENTIALS

High altitude and lower oxygen levels place higher-than-normal demands on even a fit body, so be prepared. Survival in the mountains is also tough on equipment.

- Respect nature and err on the side of caution—mountains are unforgiving, and rescue is unlikely to be quick or easy.
- Plan an achievable and safe route, and prepare an Emergency Plan of Action (see pp. 24–25).
- Dress in layers (see pp. 46–47). Start a walk lightly dressed (cold) and add or remove layers as necessary.
- Wear a hat (up to 70 percent of body heat is lost through the head) and gloves—cold fingers will hinder your functions.
- Tie loose items of clothing (hat, gloves, sunglasses) to your person—a lost glove could result in a lost hand, or life.
- Take a flashlight—weather changes and unforeseen problems may mean you are on the mountain in low light or darkness.
- Always carry basic equipment to give adequate protection if forced to camp overnight, such as a bivy sack and sleeping bag.
- Check avalanche warnings and carry an avalanche transceiver.
- Contact with snow makes clothing wet and less effective.
- If moving over snow, make a pair of improvised snow shoes (see pp. 94–95).

DESERT ENVIRONMENTS

THE DESERT ENVIRONMENT is a hostile place for human survival due to temperature extremes and the scarcity—or absence—of water. Most deserts are either hot and dry or cold and dry, and all typically receive less than 10 in (250 mm) of annual rainfall. They are areas of extremes, where heat exhaustion and hypothermia are threats, and flash floods may rapidly replace dry conditions. A harsh place to survive, the desert should only be entered by those who are prepared.

WARNING!

Desert areas are home to a range of venomous snakes, including the black mamba of eastern and southern Africa. It is particularly dangerous due to its aggressive nature and highly venomous bite.

DESERT FEATURES

The popular depiction of a desert is of a dry, sandy landscape with searing skies and temperatures among the highest in the world. Such areas do exist, but the desert environment has far greater diversity, from cold deserts in the polar regions to highlands, grassland, and broken, rocky terrain that contain "wadis" (seasonal stream beds).

HOT, DRY DESERTS

Temperate and tropical deserts are usually hot and dry due to high daytime temperatures, which leads to low precipitation and high evaporation. There may be months or even years between rainfall, which usually occurs in dramatic bursts but then quickly drains into seasonal stream beds (wadis), or evaporates either in the air or soon after touching the hot ground. Very little rain soaks into the ground, which results in sparse vegetation that has evolved to maximize water, usually via extensive root systems that collect moisture, waxy skin and leaves that prevent water loss, and stems or roots that hold water. At night, in winter, and at higher altitudes, freezing temperatures and frost may occur.

GRASSLANDS

Known as savanna in tropical areas, grasslands often border desert regions, and experience a similarly dry climate for most of the year. The key difference is that they also experience a wet season, which accounts for the more diverse and extensive vegetation. Temperate grassland features grasses, shrubs, and small pockets of trees, while savanna is typically more varied, with denser tree cover, bushes, and grasses that grow tall during the wet season but die back in periods of drought. Wildlife is more prevalent, including predators. Wildfires may spread during the dry season, and play an important role by destroying dead plant matter and adding fertile ash to the soil.

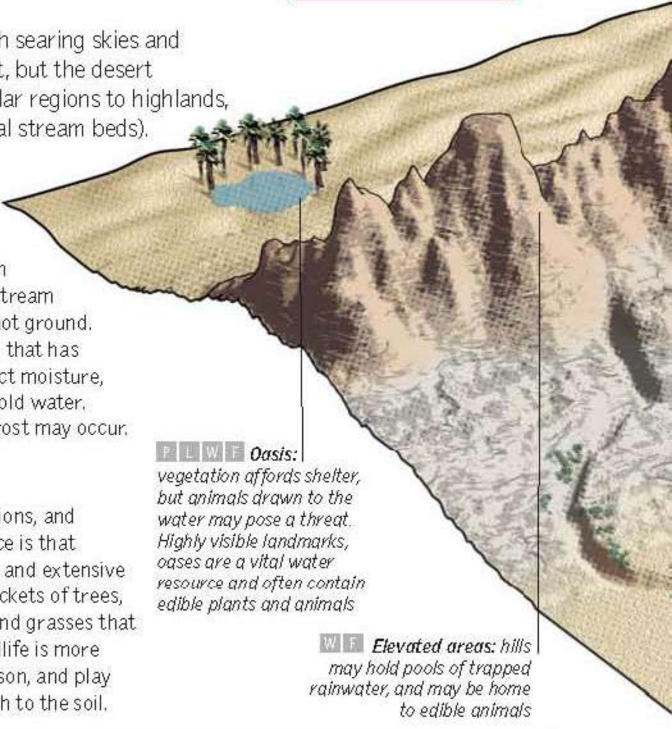

P L W F ***Oasis:*** *vegetation affords shelter, but animals drawn to the water may pose a threat. Highly visible landmarks, oases are a vital water resource and often contain edible plants and animals*

W F ***Elevated areas:*** *hills may hold pools of trapped rainwater, and may be home to edible animals*

WHERE TO FIND DESERT AREAS

Although deserts occur on every continent, the largest hot desert in the world is the Sahara in North Africa, which forms part of a desert band stretching through the Middle East and into south and central Asia.

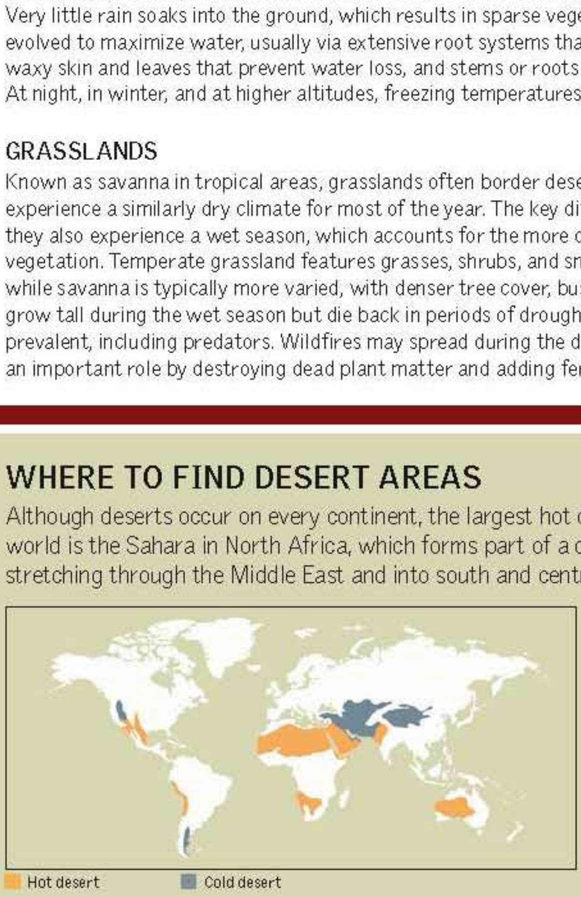

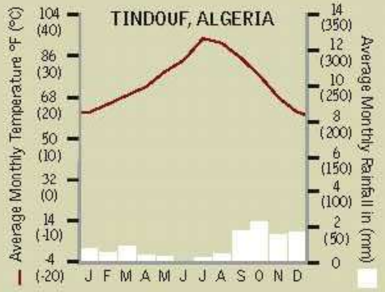

TYPICAL CLIMATE

Tindouf lies in the heart of the Sahara, and experiences a typical hot-desert climate. Precipitation is virtually zero all year round, while temperatures peak in the summer.

SURVIVING IN THE DESERT

Hostile temperatures and few natural resources limit chances of survival in the desert. Water and shelter are scarce, if not non-existent, edible vegetation is limited, and animals hide from the sun in the day. Wadis, grasslands, and higher ground all hold better prospects. The greatest dangers are dehydration and heat exhaustion (see pp. 272–73), although African savanna areas may be home to large, dangerous mammals. The main killers are big cats, hippos, rhinos, elephants, and crocodiles.

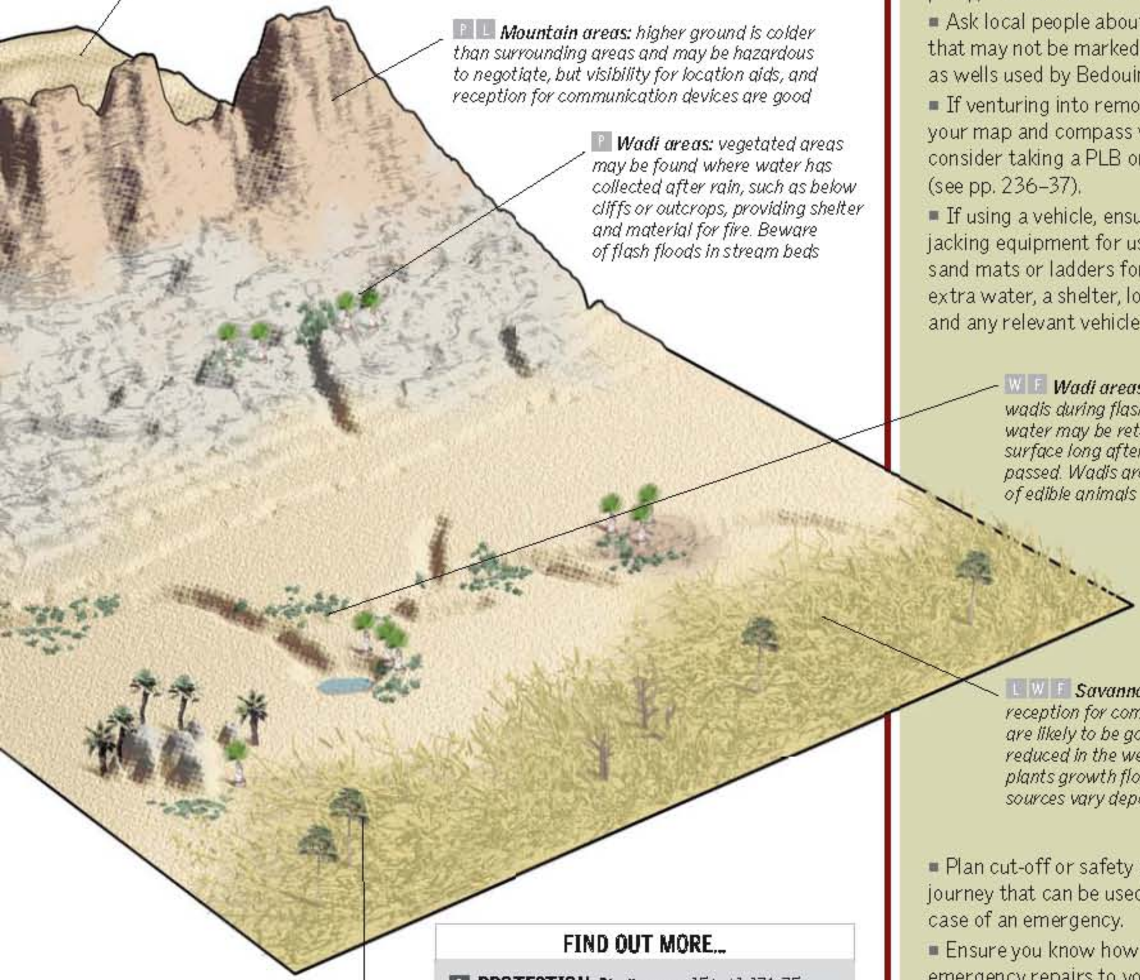

P L W ***Sand dunes:*** *shelter from the elements is poor, but visibility for location aids and for spotting rescuers is good. Observing converging animal and bird tracks or movements may indicate the direction to a water source.*

P L ***Mountain areas:*** *higher ground is colder than surrounding areas and may be hazardous to negotiate, but visibility for location aids, and reception for communication devices are good*

P ***Wadi areas:*** *vegetated areas may be found where water has collected after rain, such as below cliffs or outcrops, providing shelter and material for fire. Beware of flash floods in stream beds*

W F ***Wadi areas:*** *rivers flow in wadis during flash floods, and water may be retained below the surface long after the flood has passed. Wadis are also the habitat of edible animals*

L W F ***Savanna:*** *visibility and reception for communication devices are likely to be good but will be reduced in the wet season when plants growth flourishes. Water sources vary depending on the season*

P ***Savanna:*** *protection from the elements is limited, although material for building shelters may be found in the wet season. Sunstroke and dehydration are killers, so seek shade during the hottest parts of the day. Beware of snakes or scorpions in shaded areas, big cats and large mammals, and nocturnal animal activity*

FIND OUT MORE...

P PROTECTION Shelters pp. 156–61, 174–75 **Fire** pp. 118–33, 204–05 **Dangers** pp. 242–49, 300–05

L LOCATION Navigation pp. 66–77 **Movement** pp. 100–01 **Signaling** pp. 236–41

W WATER Finding pp. 192–93 **Purifying** pp. 200–01

F FOOD Plants pp. 206–07, 282–83 **Animals** pp. 216–21, 224–29, 292–95

DESERT ESSENTIALS

Survival in the extreme conditions of the desert is impossible without full preparation. Consider the following:

- Always prepare an EPA to notify someone of your plans before entering a desert area (see pp. 24–25).
- Water is life. Do not underestimate your needs, carry extra just in case, and carry equipment to maximize your chances of finding and procuring more (binoculars, surgical tubing, filtration/purification pump) should the worst happen.
- Ask local people about water sources that may not be marked on maps, such as wells used by Bedouin tribes.
- If venturing into remote areas, augment your map and compass with a GPS, and consider taking a PLB or satellite phone (see pp. 236–37).
- If using a vehicle, ensure you carry jacking equipment for use in soft ground, sand mats or ladders for self-recovery, extra water, a shelter, location aids, and any relevant vehicle spares.
- Plan cut-off or safety points into your journey that can be used to divert to in case of an emergency.
- Ensure you know how to accomplish emergency repairs to your vehicle.
- Ask local people about known dangers, such as impassable roads, misleading tracks, and soft-sand areas.
- When using a ladder for self-recovery, tie it to the back of the vehicle so that it's dragged behind once crossed, and can be retrieved when you reach solid ground.
- If your vehicle breaks down, leave it only if staying is no longer safe or feasible—rescuers will search for it first.

COLD ENVIRONMENTS

LOW TEMPERATURES are a potentially lethal hazard, but as long as you can conserve body heat, survival in the cold is quite possible. The ice-covered Arctic and Antarctic polar regions, sub-polar tundra and taiga, as well as parts of most temperate zones during the winter, can all be classed as cold environments. Wet, windy weather can drive moderately cold temperatures down further, increasing the risk of the most lethal cold-weather condition—hypothermia.

WARNING!
The largest predatory land mammal on Earth, the polar bear inhabits the Arctic. Attacks are most likely when the bears are hungry, and are usually fatal.

COLD ENVIRONMENT FEATURES

Outside temperate areas, the typical cold environments are the polar areas, which feature glacial ice sheets and sea ice; tundra, which contains permafrost (permanently frozen soil), small shrubs, mosses, and lichens; and taiga, with extensive conifer forests. Almost half of the northern hemisphere's landmass can be classified as a cold region, due to the influence of cold air masses (see pp. 78–83), while ocean currents and altitude can also have a cooling effect.

POLAR WASTES

The northern polar area is situated over the Arctic Ocean, so largely consists of sea ice, the extent and depth of which varies throughout the year. Because of the lack of land, melting ice is a serious danger. The southern polar area is located over the continent of Antarctica, and contains the largest glacial ice sheet on Earth. Both areas contain ice shelves—glaciers that have extended over the sea—parts of which break off to form icebergs.

TUNDRA AND TAIGA

Latitudes that neighbor the polar areas are home to tundra, an environment of permafrost and vegetation that is stunted due to low temperatures. Taiga is located further away from the poles, where temperatures are high enough to support coniferous forests.

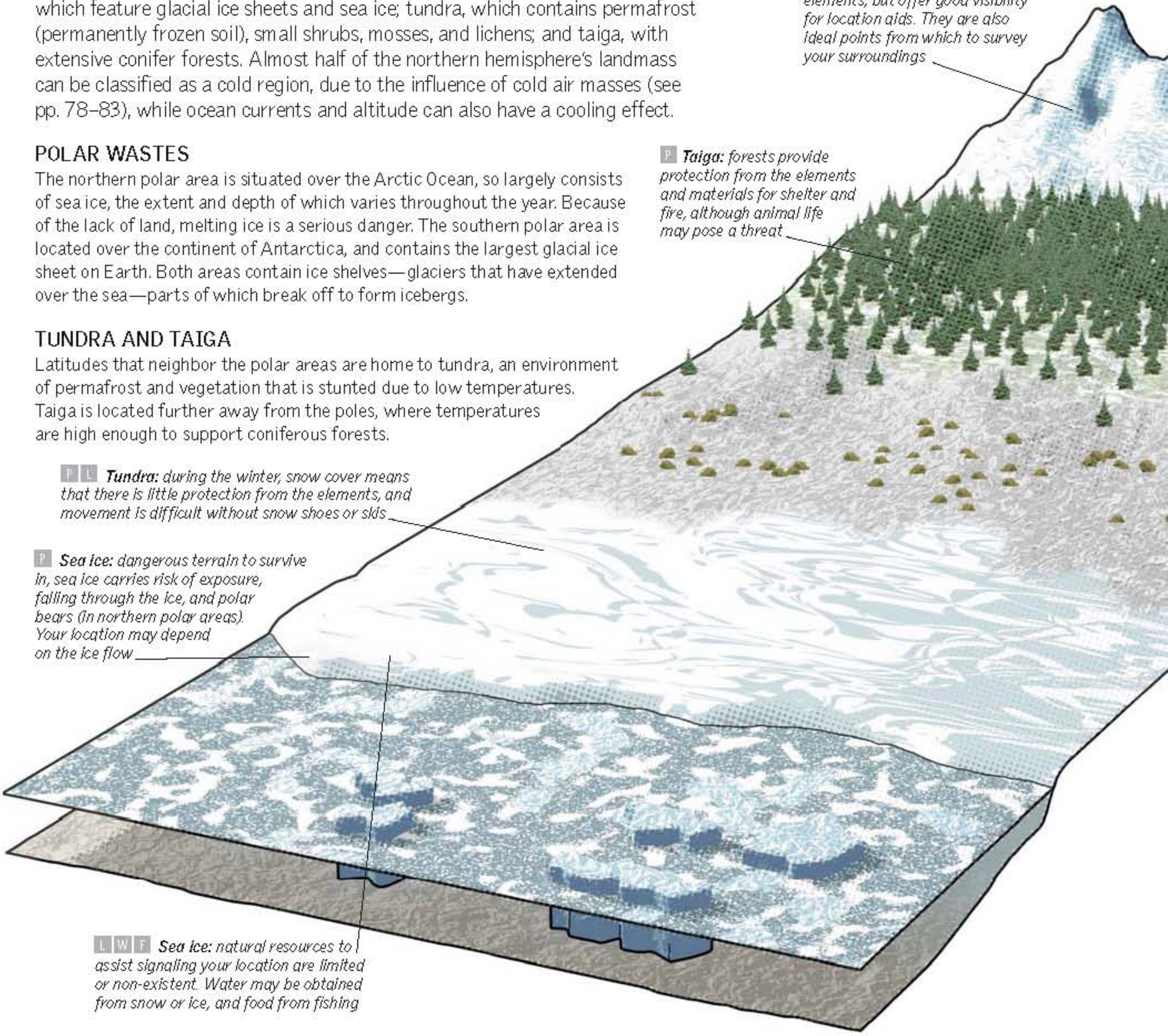

P L ***Mountains:*** *elevated areas give little protection from the elements, but offer good visibility for location aids. They are also ideal points from which to survey your surroundings*

P ***Taiga:*** *forests provide protection from the elements and materials for shelter and fire, although animal life may pose a threat*

P L ***Tundra:*** *during the winter, snow cover means that there is little protection from the elements, and movement is difficult without snow shoes or skis*

P ***Sea ice:*** *dangerous terrain to survive in, sea ice carries risk of exposure, falling through the ice, and polar bears (in northern polar areas). Your location may depend on the ice flow*

L W F ***Sea ice:*** *natural resources to assist signaling your location are limited or non-existent. Water may be obtained from snow or ice, and food from fishing*

SURVIVING IN THE COLD

Natural resources may be scarce in cold environments, so your survival is likely to depend on the equipment and supplies you have with you. Shelters can be dug from the snow to help you stay warm and avoid exposure to freezing temperatures, but material for making fire is limited in polar and tundra areas. Survival is more feasible in the taiga, where wood is available, and fresh water and edible flora and fauna can also be found. The greatest dangers are hypothermia (see pp. 272–73) and—in northern polar areas—polar bears.

LOW-TEMPERATURE DANGER

Heat loss occurs in several ways, and minimizing it is essential to survival in cold conditions. It's important to avoid over-exertion, since heat is lost through sweating, and to cover your head and wear layered clothing to trap radiated body heat. Heat is also lost through conduction—direct contact between the skin and a cooler surface—particularly when wet, since water conducts heat away from the body 25 times faster than air. For this reason, it's vital to wear insulated and waterproof clothing, and to stay dry. Respiration also involves heat loss—each inhaled breath contains cold air, which is warmed as it passes through the lungs, and laden with moisture when exhaled. Reduce heat loss by covering your nose and mouth with a ski-mask or scarf, or by breathing warmer air next to a fire or in a warm shelter.

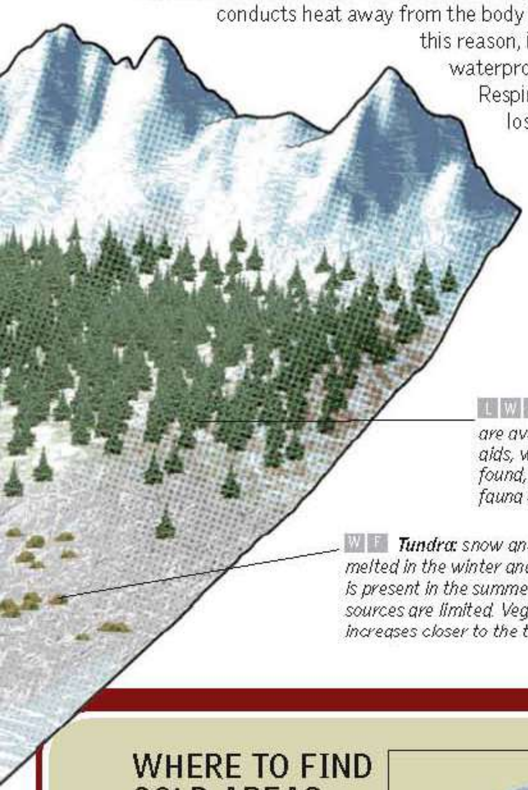

L W F *Taiga: natural resources are available for making location aids, water sources may be found, and edible flora and fauna are likely to be present*

W F *Tundra: snow and ice can be melted in the winter and ground water is present in the summer, although sources are limited. Vegetation increases closer to the treeline*

COLD ESSENTIALS

The main threats in cold environments are hypothermia and exposure, so ensure you are fully prepared:

- Dress in loose-fitting layers of clothing (see pp. 46–47), avoid overheating, and ensure your clothing stays dry and clean.
- If your hands are cold, don't warm them with your breath, which will make them wet—tuck them under your armpits.
- Regularly check your extremities (face, toes, hands, and ears) for frostnip, the first stage of frostbite (see p. 273).
- Wind-chill is dangerous, so take shelter from the wind at every opportunity, particularly if you are in a survival situation.
- Always ensure that your shelter is well ventilated—keep vent holes clear and check regularly, especially when snow falls.
- Get off the cold ground, snow, or ice—sit on your pack or make a sleeping platform using boughs to avoid losing body heat.
- If fire is your primary means of warmth, triple the amount of firewood you think you need—you will need enough to last the night.

FIND OUT MORE...

P **PROTECTION** **Shelters** pp. 156–65, 178–79 **Fire** pp. 118–33, 204–05 **Dangers** pp. 242–49

L **LOCATION** **Navigation** pp. 66–67 **Movement** pp. 94–97 **Signaling** pp. 236–41

W **WATER** **Finding** pp. 194–95 **Purifying** pp. 200–01

F **FOOD** **Plants** pp. 206–07, 286–87 **Animals** pp. 208–13, 216–23, 290–91, 296–97

WHERE TO FIND COLD AREAS

The Arctic and Antarctic polar areas are the coldest places on Earth, and are found at the furthest points from the Equator. Tundra exists next to polar regions, and is bordered by taiga, which is found in northern Eurasia and North America.

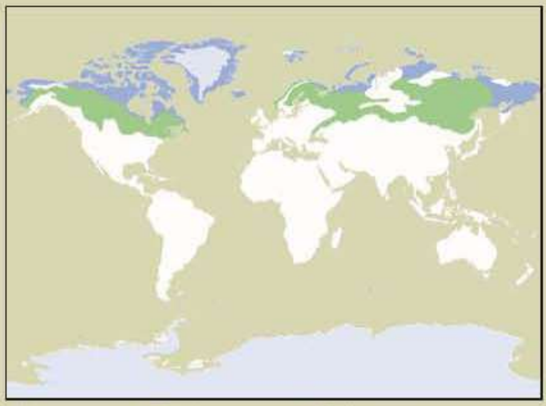

⬜ Polar areas 🟦 Tundra areas 🟩 Taiga areas

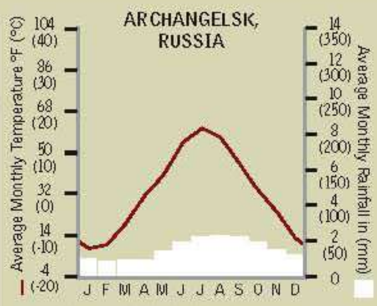

TYPICAL CLIMATE

Archangelsk is a Russian port on the Barents Sea. Located in the taiga zone, it experiences below-zero temperatures for half the year and low precipitation.

MARINE ENVIRONMENTS

POTENTIALLY THE MOST HARSH and challenging of all environments, the sea is unlike any other due to the lack of the most essential requirement for human life—fresh water. Around 70 percent of the Earth's surface is covered by saltwater, ranging from the freezing, wind-torn polar oceans to the warm tropical seas. Winds and currents govern your movement, protection is limited, and natural resources are virtually nil, so reaching land is your best hope for survival.

WARNING!

There are sharks in every ocean, but those in tropical waters are the most aggressive. Around 20 of the hundreds of species have attacked humans, and the most dangerous types are the great white, tiger, and bull sharks.

MARINE FEATURES

The marine environment ranges from the coastal margins, which are shallow enough— 650 ft (200 m) or less—to support a huge array of animal and plant life, to the open oceans, the deepest of which plunge to more than 33,000 ft (10,000 m).

SHELTERED COAST TO OPEN SEA

The coastal margins are home to the majority of ocean life, and include terrestrial areas—the shoreline and intertidal zone—and aquatic—the sea up to the edge of the continental shelf. A variety of environmental features are found, such as sand, rock, and pebble beaches, dunes, cliffs, estuaries, mud flats, mangroves, lagoons, kelp forests, and coral reefs. The open ocean is a more barren environment that supports less life, and is notable for its vast extent, comprising 92 percent of all saltwater. Survival depends on your equipment, supplies, and ingenuity.

OCEAN CONDITIONS

Your protection and location at sea depends largely on the weather—you may need to deal with anything from exposure to the hot sun, to cold, wet, windy conditions and high seas. Conditions are toughest in the South Atlantic Ocean, where winds (known as the "Roaring Forties") are strong all year round, and sea ice is extensive in the winter. Severe seasonal storms occur: hurricanes in the tropical areas of the western Atlantic, typhoons in the western Pacific. India and Southeast Asia also experience severe weather during the monsoon seasons.

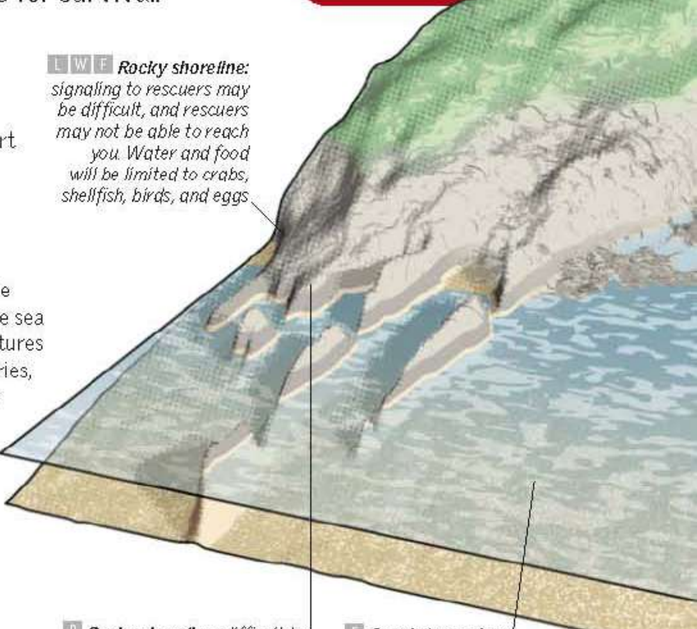

L W F ***Rocky shoreline:*** *signaling to rescuers may be difficult, and rescuers may not be able to reach you. Water and food will be limited to crabs, shellfish, birds, and eggs*

P ***Rocky shoreline:*** *difficult to make landfall and to survive on, and poses risk of injury when landing. Progress to safer areas may be difficult or impossible due to cliffs or slippery rocks. Protection from the elements and tides is limited*

F ***Coastal margins:*** *the majority of ocean life is found near the coasts, and most species live within 650 ft (200 m) of the surface*

WINDS AND CURRENTS

Winds and currents follow prevailing patterns, so if you know roughly where you are, you can figure out where they'll carry you. With a drift of just 1 knot, a liferaft can move 24 miles (38 km) from its last known position in 24 hours. It may also drift up to 35° either side of the wind, further widening the search area to more than 400 sq miles (1,000 sq km) in the first 24-hour period, and increasing exponentially over time. Conversely, favorable currents can also be used to help you drift to landfall faster.

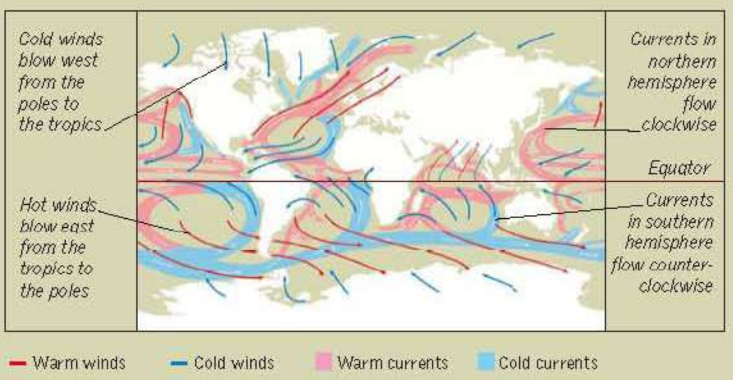

Cold winds blow west from the poles to the tropics

Hot winds blow east from the tropics to the poles

Currents in northern hemisphere flow clockwise

Equator

Currents in southern hemisphere flow counter-clockwise

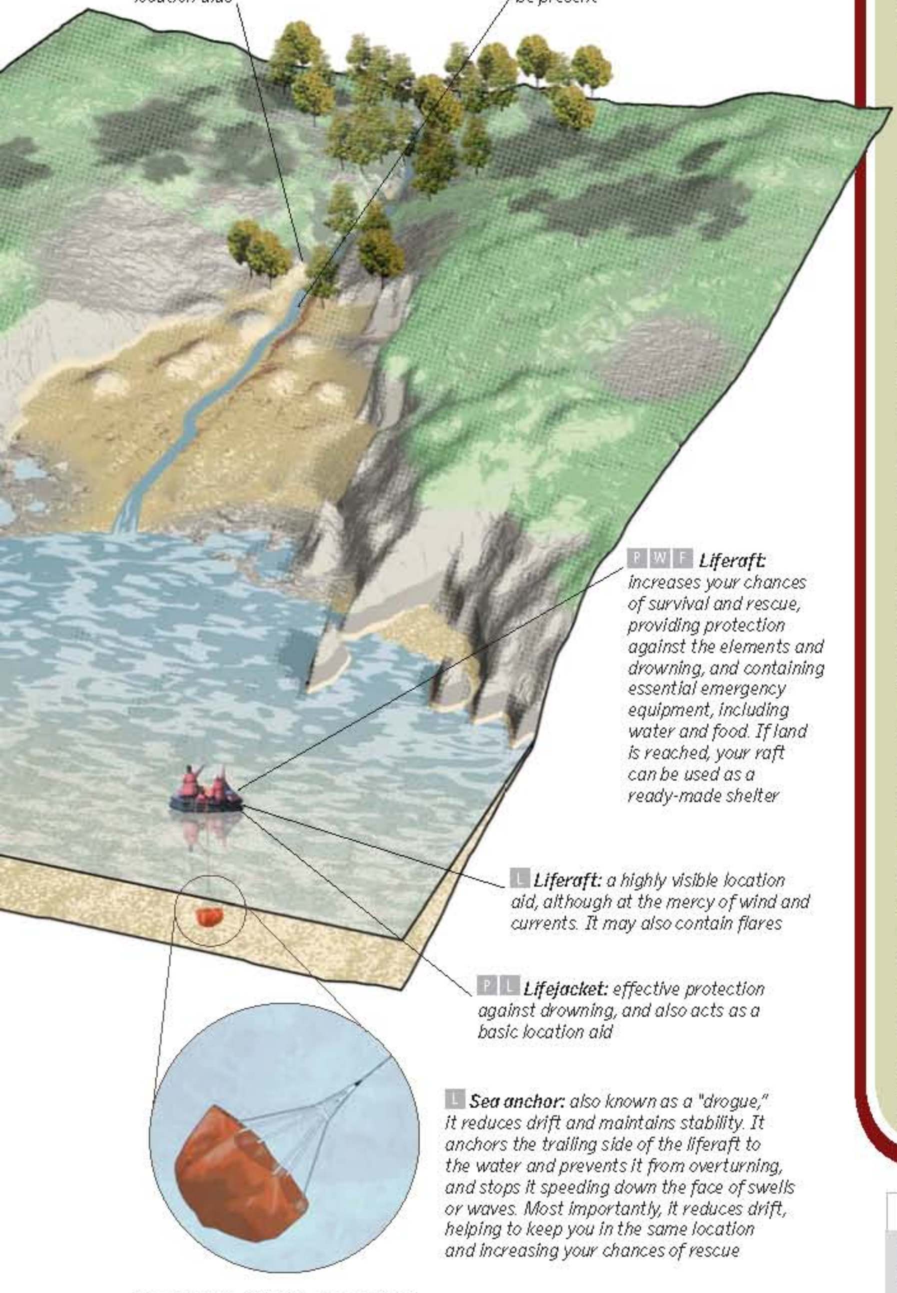

P L ***Beach:*** *above the high-tide mark, beaches may provide protection from wind and rain, and materials for building shelters, lighting fires, and improvising location aids*

W F ***Beach stream:*** *streams may drain on to beaches, providing water for drinking, cooking, and washing. Edible plants and animals may be present*

P W F ***Liferaft:*** *increases your chances of survival and rescue, providing protection against the elements and drowning, and containing essential emergency equipment, including water and food. If land is reached, your raft can be used as a ready-made shelter*

L ***Liferaft:*** *a highly visible location aid, although at the mercy of wind and currents. It may also contain flares*

P L ***Lifejacket:*** *effective protection against drowning, and also acts as a basic location aid*

L ***Sea anchor:*** *also known as a "drogue," it reduces drift and maintains stability. It anchors the trailing side of the liferaft to the water and prevents it from overturning, and stops it speeding down the face of swells or waves. Most importantly, it reduces drift, helping to keep you in the same location and increasing your chances of rescue*

SURVIVING AT SEA

Natural resources in the open ocean are virtually nil, so improvising shelter or location aids is limited to what you have with you. Seasickness, especially in a small liferaft, can seriously affect your ability to remain both hydrated and motivated. Unless you can reach land, your only drinking water will come from rain or mechanical desalination devices—consuming urine or saltwater will only increase your rate of dehydration. Reaching land or shipping lanes greatly increases your chances of rescue.

MARINE ESSENTIALS

Your chances of survival at sea will be greatly enhanced by good preparation, so consider the following:

- Pack sea-survival equipment as if your life depends on it—it might! Other environments allow you to improvise survival and location aids, but not the sea.
- Take emergency immersion-survival suits. They will protect you against the elements and aid flotation, and their color (high-visibility orange or red) is an aid to location.
- Always carry a Personal Locator Beacon (PLB), just in case.
- Take several means of obtaining water, such as a reverse-osmosis pump, which turns saltwater into drinking water, a saltwater desalination kit, a solar still, and rain catchers.
- Always keep a pre-packed emergency bag handy. It should contain; first aid kit, reverse-osmosis pump, emergency beacon, solar still, water, flares, fishing kit. (see pp. 250).
- If your vessel sinks, do not abandon it until absolutely necessary. Stay nearby unless it is unsafe to do so—anything that gives you a bigger "survival footprint" will increase your chances of being rescued, and you may be able to retrieve provisions from the vessel.
- If imminent rescue is unlikely, you will need to ration your water. Consider how much you have, how much can be procured, the chances of seasickness (which leads to dehydration), and the likelihood of being rescued later on.
- Protect yourself from the elements (sun, wind, cold, heat, and salt) as well as you can—prevention is better than cure.
- The sea, especially in rough conditions, can sap the will to live quicker than any other environment. Seasickness is literally draining, resulting in lost fluid, energy, and motivation. Take anti-seasickness tablets at the earliest opportunity—even before abandoning ship—so that you can keep them down, and keep taking them as recommended.

FIND OUT MORE...

P **PROTECTION** **Shelters** pp. 156–61, 176–77 **Fire** pp. 118–33, 204–05 **Dangers** pp. 250–55, 300–01

L **LOCATION** **Movement** pp. 106–07 **Signaling** pp. 236–41

W **WATER** **Finding** pp. 196–97 **Treating** pp. 200–01

F **FOOD** **Plants** pp. 206–07, 288–89 **Animals** pp. 208–13, 224–29, 292–99

GET THE RIGHT GEAR

Whether you're equipping yourself for a lengthy expedition or just a day trip, it's important to choose the correct equipment and clothing. The wrong time to realize that your gear isn't up to the job is when your life depends on it doing what it was designed to do.

When selecting gear, it's always a good idea to work backward from a worst-case-scenario survival situation. Think about what you would need to survive if the worst happened. This is your "first-line" equipment, which you should carry with you at all times. It should enable you to address the basic principles of survival relevant to the environment you are in, and consists of the clothing you would wear and your basic survival equipment—your survival tin and belt-order (see pp. 43, 61).

Once you have organized this basic equipment, you can then decide what pieces of essential gear would make your expedition enjoyable rather than just survivable. This will probably consist of a bivy or tent, a sleeping bag, cooking equipment and food, and something to carry it all in.

In this section YOU WILL DISCOVER...

- the **meaning** of the term "**bomb-burst gear**"...
- that it's always **good to share** when you're **in a group**...
- which **backpacks** are the **best backpacks** to pack...
- why you should **hike around the house** to **break in your boots...**
- a good reason to take a **three-season tent...**
- how to **sleep** like a **mummy...**
- what **gorp** is and **when** you should **eat it...**

It's at this stage that you should conduct a short trial to check that your chosen gear will fit in your pack and that you can comfortably carry it.

Finally, any spare capacity can be utilized for non-essential "luxury" items, such as an inflatable mattress and pillow, an MP3 player, or a book. Double-check that your gear works and that you know how to use it properly. The more you understand how and why a piece of equipment works, the better able you will be to improvise should it be damaged or lost.

With each venture into the wilderness, your confidence in your abilities and that of your equipment will increase. As it does, you may want to explore the world of ultralight camping.

ULTRALIGHT CAMPING refers to the careful selection of your three largest items of gear to dramatically reduce the load you have to carry. A standard tent, pack, and sleeping bag have a combined weight of approximately 20 lb (9kg). Ultralight camping has the potential to halve the weight of your "big three."

- Take a hooped bivy instead of a tent and choose a low season-rated sleeping bag (wear your day clothes inside the bag for extra warmth).
- You can then reduce the size of your pack because your tent and sleeping bag are smaller.

Trekking with less equipment but still staying safe allows you greater freedom to enjoy the wilderness—carrying less weight puts less strain on your body, enabling you to cover greater distances. Ultralight camping can become addictive and is safe if you use common sense—but never compromise on the quality of equipment. Try:

- Using a hexamine stove to cook food and boil water—you won't have to carry a stove and fuel.
- Taking one pot to cook food and boil water.
- Using natural materials as a mattress.

> “The **wrong time** to realize that **your equipment** isn't up to the **task in hand** is when **your life depends** on it doing what it was **designed to do**”

CHOOSING YOUR GEAR

FORWARD PLANNING IS ESSENTIAL when you're deciding on what equipment to take on your trip. You'll need to assess your own personal requirements, the likely demands of weather and terrain, and the amount of gear you'll actually be able to transport. With a little forethought, you'll be as well equipped as you can be for any situation.

PACKING FOR YOUR TRIP

You will first need to weigh up your particular kit requirements against the limitations of your chosen mode of transport (see below). It's then vital to organize and prioritize your gear (see opposite) to ensure that any items you may need in a survival scenario are always close at hand, should the worst happen.

KEY PACKING TIPS

- Pack your gear in reverse order: the things you want to get to first should be the last things you pack.
- Stash items you may need during the day in side or lid pockets, for easy access.
- Coil and tape long straps so that they aren't a snagging hazard.
- Use a dry bag inside your backpack.
- Ensure that all pockets are fully closed and zipped.
- Air out anything packed wet as soon as you stop for the day.

HOW MUCH TO PACK

The environment of the region you're traveling to will dictate the type of equipment you will need to take with you, but your proposed mode of transport is the main constraint on the quantity of kit you can take (see chart, right). The more gear you have to carry yourself, the fewer luxuries you can take: the added weight will make traveling uncomfortable, use up too much valuable energy, slow your progress, and also limit the distance you can travel each day. It's also worth considering whether you can reduce the weight and volume of the kit items you're taking (see box, below).

LIGHTENING THE LOAD

When you're going to be traveling on foot, it's important to get rid of all the excess weight and volume you can, even if it's only by a small amount. The larger and heavier your backpack, the harder you will find the journey. If you have already ruthlessly pared your gear down to the bare essentials, try to reduce the weight or volume of the kit itself. First concentrate on the big three items: tent, sleeping bag, and backpack. Consider using a hooped bivy or tarpaulin rather than a tent; look for a lighter sleeping bag—some can weigh as little as 2lb (0.9 kg); and consider buying a backpack with a lighter frame.

WEIGHT AND SPACE RESTRICTIONS	
ON FOOT	Backpacks (see pp. 44–45) are available in a wide range of volumes—from 30 liters (a day pack) to 80 liters (suitable for a trip of a week or more). Bear in mind that you'll have to carry whatever gear you take with you, and that the weight of the pack will determine what you're realistically able to achieve each day in terms of distance and activity. Pack only the bare minimum: there's no room for luxuries when traveling on foot.
MOUNTAIN/ TOURING BIKE	Pannier bags for bicycles are designed to fit on front, rear, or side pannier racks, and are available in a range of shapes, sizes, and capacities (from 5 to 50 liters). Larger bags will provide you with enough space to pack essentials, such as tent, sleeping bag, cooking equipment, and spare clothes, but you'll be restricted by the weight, and also by the need to carry cycle-maintenance gear, such as spare tires, inner tubes, cables, and tools.
MOTORCYCLE	Side panniers have a capacity of up to 80 liters and tank bags up to 25 liters, and weight is not a major issue. However, you'll need to take extra protective equipment, maintenance tools (including spare tires), and, in some cases, spare fuel, all of which will impact on available space.
PACK ANIMAL	Ideal for rough terrain, such as mountains or desert, pack animals can carry serious weight. A horse can transport up to 20 percent of its body weight; a camel can carry a weight of up to 990 lb (450 kg)—although 330 lb (150 kg) is more common—but you'll need to factor in the animal's food and water.
VEHICLE	The size of the vehicle will determine the amount and the weight of gear you can take with you, but you shouldn't have to compromise; extra equipment can be packed in peli boxes—waterproof storage containers with a hard, plastic exterior—that can be strapped to the outside of the vehicle. The main issue will be the weight-versus-fuel ratio: the heavier the vehicle, the more fuel it will use.
CANOE/ KAYAK/BOAT	Some kayaks and canoes are designed to bear loads of up to 660 lb (300 kg), including the craft's occupant(s), so you should be able to take all of your essential gear, plus a few luxuries. Given your proximity to the water, keeping your gear dry is the main priority. Weight and space restrictions don't apply for larger vessels.

PRIORITIZING YOUR KIT

However much gear you decide to take, you should group individual items into one of three categories—first-line, second-line, and third-line—according to their importance to your survival in a worst-case scenario. It's very unlikely that you'll be able to carry all of your equipment all of the time, and in an emergency situation, you may not have time to gather it all together. However, by prioritizing your gear beforehand, you'll have all your essential items on your person, or close enough to grab at a second's notice.

FIRST-LINE GEAR

Known as "bomb-burst gear" in the military, first-line gear (see right) is your basic survival equipment. If something goes wrong and you have to abandon the bulk of your gear to avoid injury, or death, what you're standing in is all you'll have to help you survive. First-line gear should therefore comprise crucial items of outdoor clothing, along with essential items for navigation and safety. Your bushcraft knife, firesteel, and belt pouch make up your belt order (see pp. 60–61). You'll need to risk-assess your particular situation and adapt your kit priorities accordingly, as conditions change, meaning that your first-line gear may alter as the day progresses. Bear in mind that if a removed layer of clothing goes into your backpack, it's no longer part of your first-line kit.

SECOND-LINE KIT

Second-line gear is everything you would need to stay safe on a daily basis under normal conditions. It can either be packed in a small day pack, or contained in pouches on a belt, the idea being that you carry it at all times. For example, a climber may decide to leave the bulk of his or her equipment (third-line gear) while attempting to reach the summit of a peak. By proceeding with first- and second-line gear, the climber is sacrificing equipment for weight and speed, but is still safe, as they have the essentials on them. Examples of second-line gear can include:

- A spare set of clothes, a bivy sack, and cordage
- Emergency rations and first-aid kit (see p. 260)
- Hexamine stove (see p. 58–59) and items to make a hot drink
- Matchless fireset (see p. 118), and a metal cup

THIRD-LINE GEAR

Also known as your "sustainment load," third-line gear is essentially the equipment you need to keep yourself alive and to function for longer periods of time. How much third-line gear you have will ultimately depend on your mode of transport and the amount of equipment you can carry. Examples of third-line gear include:

- A form of shelter—a tent or a tarpaulin
- Cooking utensils, such as a stove or cooking pot
- A backpack
- Food stores
- A sleeping bag and sleeping mat
- Any large water containers or hydration packs
- Wash kit and sanitary items

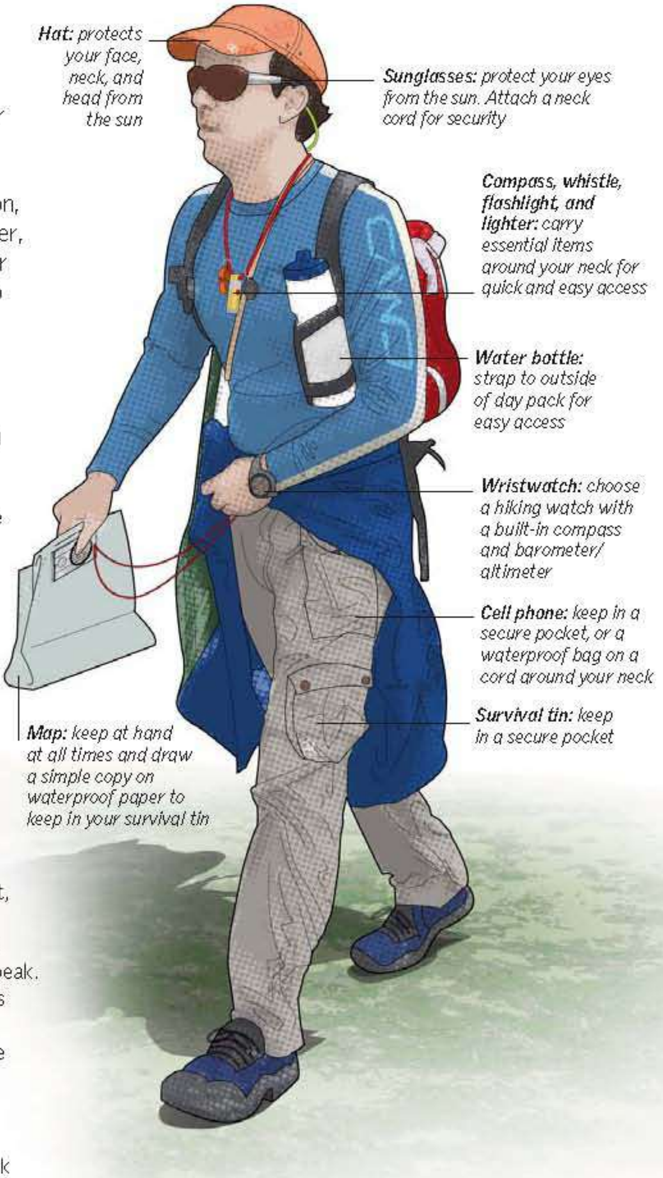
Hat: protects your face, neck, and head from the sun
Sunglasses: protect your eyes from the sun. Attach a neck cord for security
Compass, whistle, flashlight, and lighter: carry essential items around your neck for quick and easy access
Water bottle: strap to outside of day pack for easy access
Wristwatch: choose a hiking watch with a built-in compass and barometer/altimeter
Cell phone: keep in a secure pocket, or a waterproof bag on a cord around your neck
Survival tin: keep in a secure pocket
Map: keep at hand at all times and draw a simple copy on waterproof paper to keep in your survival tin

FLASHLIGHTS

A flashlight is a piece of first-line gear, even if you don't intend to be out in the dark, because situations can change unexpectedly. Modern flashlights use LEDs, and are small, lightweight, powerful, and very efficient. Opt for a "hands-free" headlamp and tape a spare set of batteries to your head-band. Also include two miniature Photon (LED) flashlights in your survival kit (see pp. 60–61).

CHOOSING A BACKPACK

BACKPACKS ARE DESIGNED to enable hikers to carry large loads comfortably. They are made from a variety of materials, most of them waterproof. However, you should always use a separate waterproof bag inside your backpack to ensure your sleeping bag and spare clothes are kept dry; they are of little use in a survival situation if they get wet.

CARRY ONLY WHAT YOU NEED

Regardless of your pack size, you are the one who has to carry it. If in a group, many communal items can be shared. Try not to double up on basics (a party of four does not need four tubes of toothpaste, for example).

WHAT DO YOU NEED?

Your choice of backpack will depend on a number of factors. How long is your trip? How much do you intend to carry? What type of gear will you be taking with you? The answers to these questions will determine the size of pack—known as the "capacity"—that is right for you. Pack sizes range from 30 liters (a day pack) to 80 liters or more (suitable for trips lasting a week or more).

BACKPACK FRAMES

If you have chosen a medium- to large-sized pack, you will have to choose between an internal- or an external-frame pack.

DAY PACK

Day hikers can sometimes find themselves in trouble because they are not carrying the appropriate equipment to protect them from the elements. A day pack provides you with the means to carry essential items such as food, water, maps, compasses, and protection from cold and wet weather on a day trip.

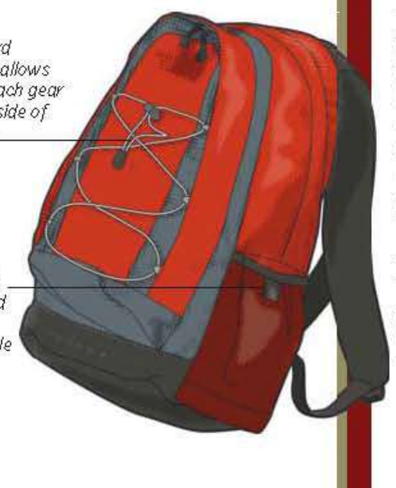

Elastic cord strapping allows you to attach gear to the outside of the pack

Side pouch can be used to carry a water bottle

INTERNAL-FRAME PACKS

Internal-frame packs are narrower in profile than their external-frame counterparts and tend to be longer, with either one or two internal compartments and very few pockets on the outside. The frame is integrated within the pack and usually consists of stays or flat bars, about 1 in (25 mm) wide and $\frac{1}{10}$ in (3 mm) thick, made from plastic. Straps on the outside allow you to compress the pack, which prevents the items inside the pack from shifting and throwing you off-balance.

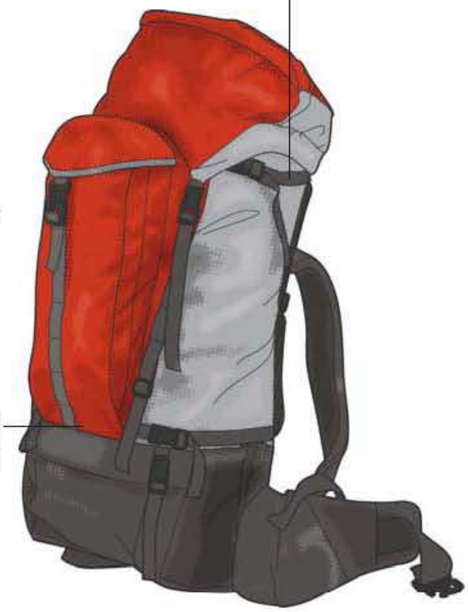

Compression straps reduce bulk and keep contents of pack in place

Additional compartment for quick access to essential items, such as wet-weather gear

EXTERNAL-FRAME PACKS

An external frame houses the pack on a lightweight tubing framework. They are great for heavy loads, as the pack sits more squarely on your hips. They are also cooler to carry than internal-frame packs, as the air can circulate between your back and the pack. External-frame packs usually have more pockets on the outside, which allows you to pack your items by category, rather than packing everything into one or two compartments. This makes locating items easier and allows you to unpack only the compartment you need, rather than the entire pack. Chest and waist straps ensure a comfortable fit.

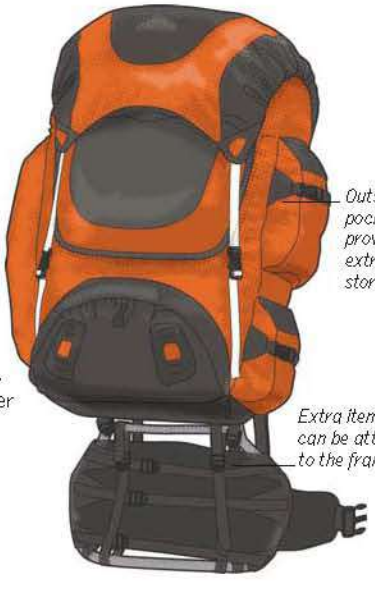

Outside pockets provide extra storage

Extra items can be attached to the frame

FITTING A BACKPACK

Once you have chosen a backpack of the right size, design, and features for your trip, you will have to make sure that it's a comfortable fit. You will need someone's help, but here are a few tips for ensuring your new pack sits comfortably on your back.

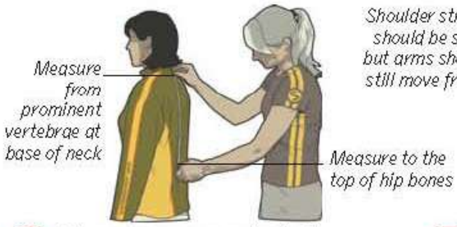
Measure from prominent vertebrae at base of neck
Measure to the top of hip bones

1 First measure your back. The pack you choose should be based on the length of your torso.

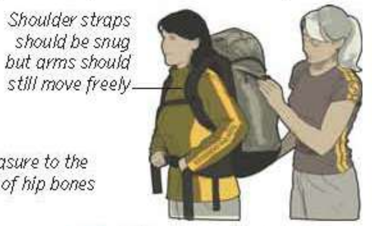
Shoulder straps should be snug but arms should still move freely

2 Make sure the pack is a comfortable fit, allowing for any extra layers of clothing.

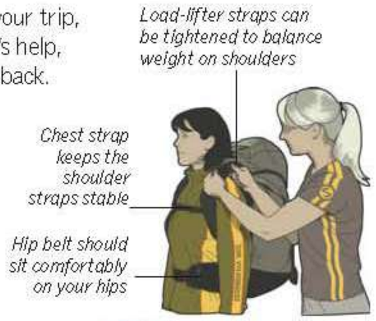
Load-lifter straps can be tightened to balance weight on shoulders
Chest strap keeps the shoulder straps stable
Hip belt should sit comfortably on your hips

3 You can now fine-tune the pack's fit by adjusting the shoulder and hip straps.

PACKING A BACKPACK

Always pack in reverse order: the items you need first should go in last. Heavy items should be placed close to your back to prevent the pack pulling away from your shoulders.

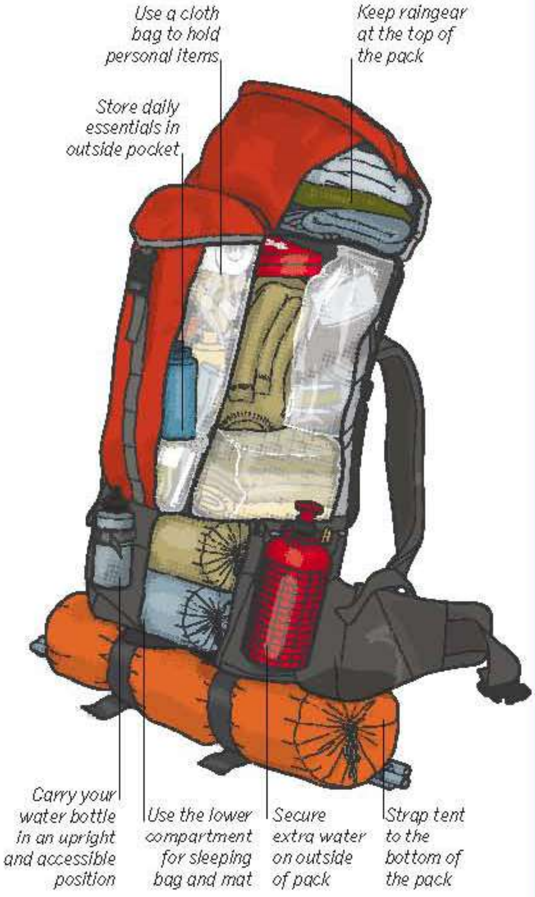
Use a cloth bag to hold personal items
Keep raingear at the top of the pack
Store daily essentials in outside pocket
Carry your water bottle in an upright and accessible position
Use the lower compartment for sleeping bag and mat
Secure extra water on outside of pack
Strap tent to the bottom of the pack

MAKING A SIMPLE PACK FRAME

In a survival situation, having something to carry your gear will be a huge advantage. All you need to make your own improvised backpack is a knife, some wood, some cord, and a waterproof sheet.

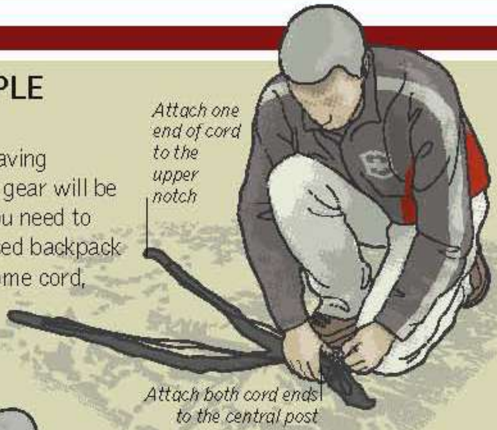
Attach one end of cord to the upper notch
Attach both cord ends to the central post

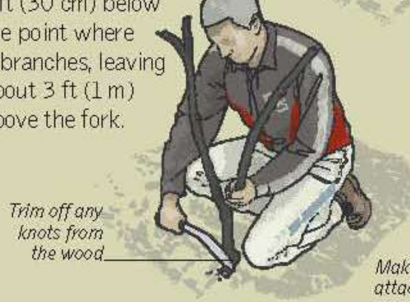
Trim off any knots from the wood

1 Cut a light bough about 1 ft (30 cm) below the point where it branches, leaving about 3 ft (1 m) above the fork.

2 Cut notches on all three ends of the bough and tie lengths of cord around the notches to serve as straps to go over your shoulders.

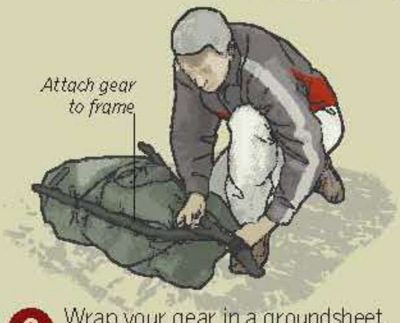
Attach gear to frame

3 Wrap your gear in a groundsheet or waterproof poncho and tie it as high as possible onto the frame on the side opposite the straps.

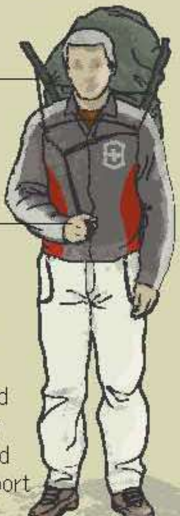
Make sure gear is attached securely to the frame
Hold straps away from shoulder to avoid chafing

4 Don't be tempted to overload your improvised backpack, even though it should be able to support a good weight.

DRESSING FOR THE OUTDOORS

MODERN OUTDOOR CLOTHING is technologically advanced and highly sophisticated. New materials and designs are extremely lightweight, durable, and versatile. To make the most of your gear, choose the fabrics and combinations of clothes most suited to the environment and conditions in which you are traveling.

REGULATING BODY HEAT

Don't be tempted to overdress. Sweat can soak you as much as rain, and a sudden change of temperature can leave you vulnerable to hypothermia. Wearing layers helps you to control your body temperature.

LAYERING CLOTHES

The rule of layering is simple: several light layers are better than one heavy layer. Wearing multiple layers gives you flexibility to fine-tune your temperature by taking off or putting on layers to regulate your body heat. Choose fabrics such as wool, fleece, microfleece, and down—all of which are good insulators.

HOW LAYERING WORKS

Layering is effective because it traps air between the layers of clothes and helps you stay warm in any environment. If you wear the correct layers in the right order, your clothing will move sweat away from your body, keep moisture out, and help to keep you insulated. Wear wicking fabrics, such as polypropylene, in hot and cold weather.

THE LAYERING SYSTEM

The top (outer) layer repels rain, while dampness is wicked away from the skin by the base layer closest to the skin. The mid-layers insulate the body and help retain warmth.

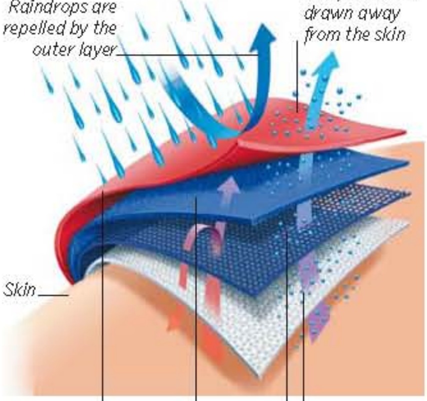
Raindrops are repelled by the outer layer
Perspiration is drawn away from the skin
Skin
Breathable waterproof outer layer
Fleece mid-layer
Synthetic mesh
Polypropylene base layer

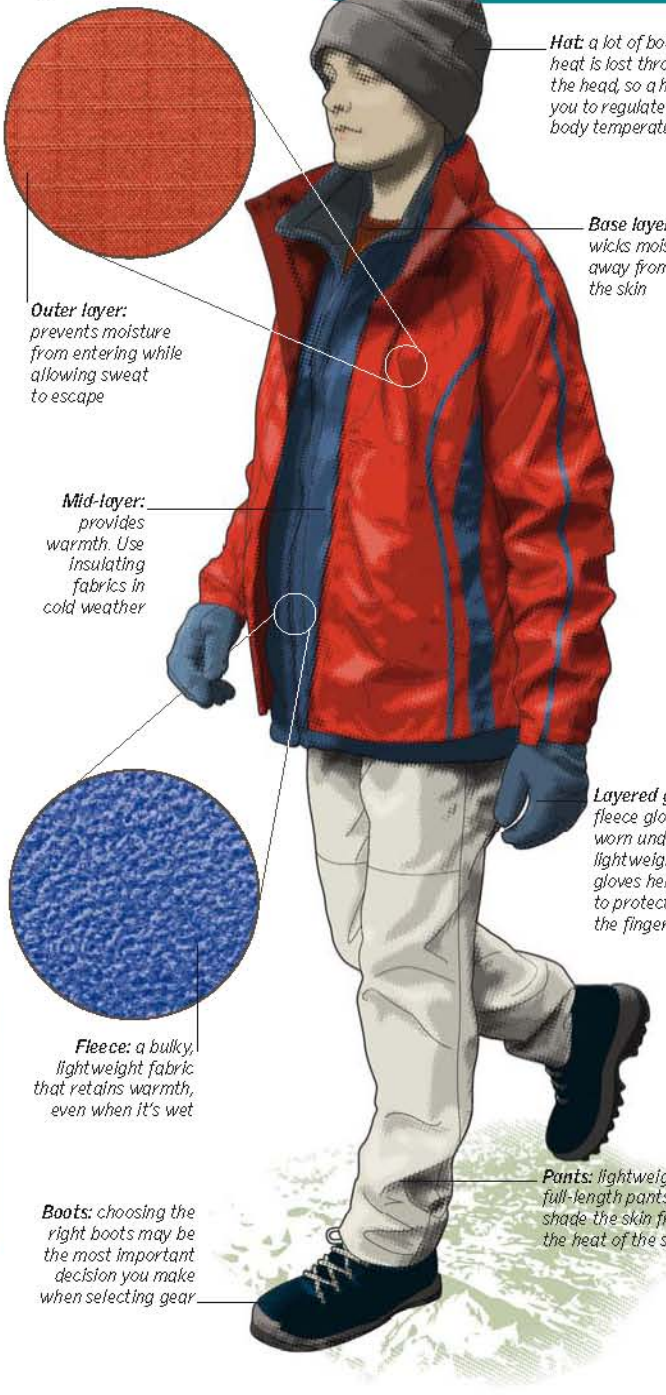
Hat: a lot of body heat is lost through the head, so a hat helps you to regulate your body temperature
Base layer: wicks moisture away from the skin
Outer layer: prevents moisture from entering while allowing sweat to escape
Mid-layer: provides warmth. Use insulating fabrics in cold weather
Layered gloves: fleece gloves worn under lightweight gloves help to protect the fingers
Fleece: a bulky, lightweight fabric that retains warmth, even when it's wet
Pants: lightweight, full-length pants shade the skin from the heat of the sun
Boots: choosing the right boots may be the most important decision you make when selecting gear

WARM-WEATHER CLOTHING

It's vital to stay as cool as possible in warm weather to avoid heat exhaustion or heat stroke (see p. 272). Too much exposure to direct sunlight can cause sunburn and dehydration, and heat rashes can be exacerbated if sweat is not able to evaporate properly. Choose breathable fabrics that will keep you cool and protect your skin from the sun. Stay hydrated and always wear a hat.

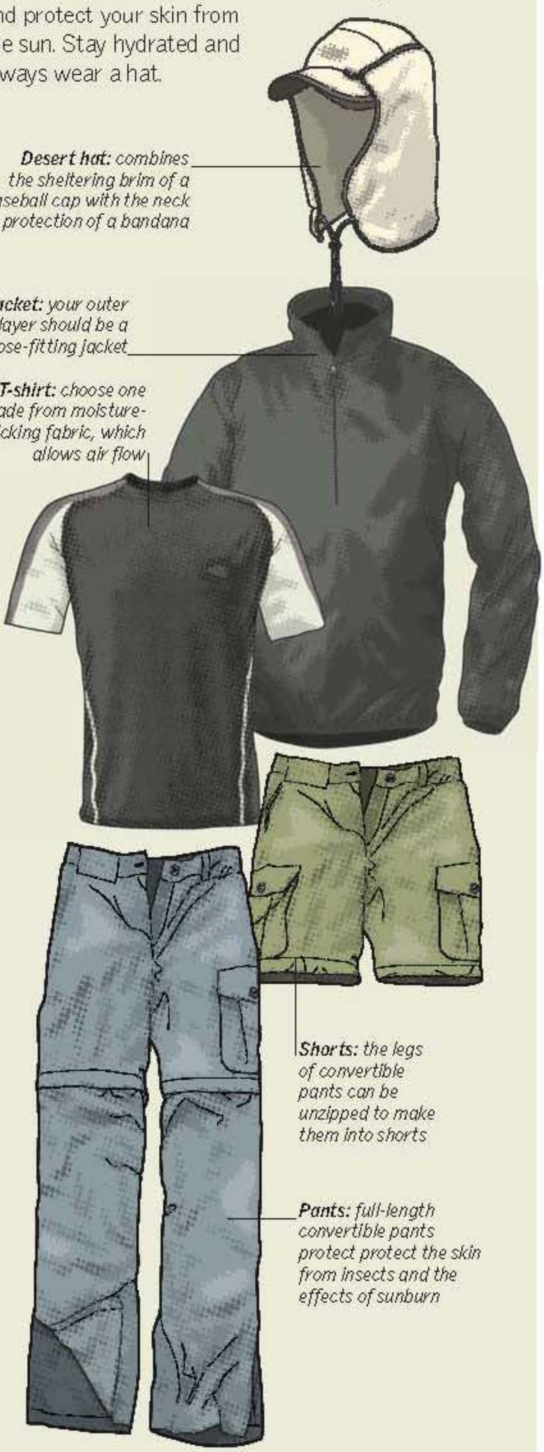

__Desert hat:__ combines the sheltering brim of a baseball cap with the neck protection of a bandana

__Jacket:__ your outer layer should be a loose-fitting jacket

__T-shirt:__ choose one made from moisture-wicking fabric, which allows air flow

__Shorts:__ the legs of convertible pants can be unzipped to make them into shorts

__Pants:__ full-length convertible pants protect protect the skin from insects and the effects of sunburn

COLD-WEATHER CLOTHING

In cold weather, you need to pay particular attention to your clothing layers. Use a lightweight base layer, which will wick moisture from your skin, several layers of warm, insulating clothing, and an outer shell that's windproof and waterproof. The secret to staying comfortable is to adjust your layers as your body temperature changes. Remove layers as your body temperature rises to avoid sweating.

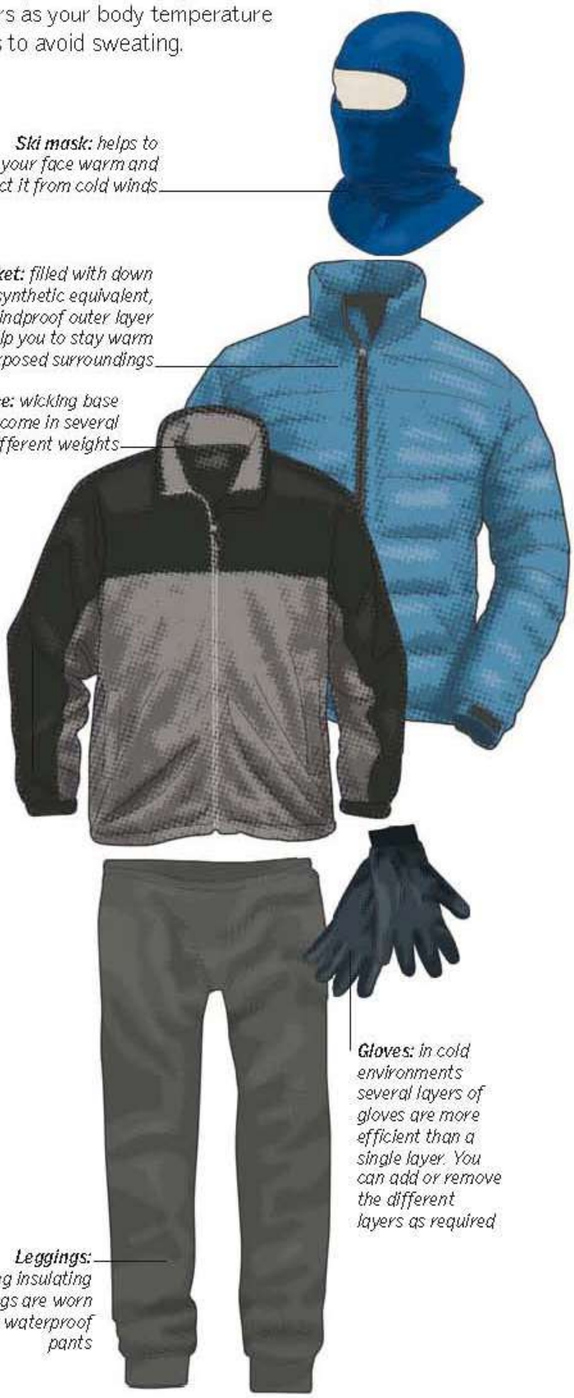

__Ski mask:__ helps to keep your face warm and protect it from cold winds

__Jacket:__ filled with down or a synthetic equivalent, a windproof outer layer will help you to stay warm in exposed surroundings

__Fleece:__ wicking base layers come in several different weights

__Gloves:__ in cold environments several layers of gloves are more efficient than a single layer. You can add or remove the different layers as required

__Leggings:__ wicking insulating leggings are worn under waterproof pants

WET-WEATHER CLOTHING

Sudden downpours can occur almost anywhere and it is easy to be caught out by them. The best fabrics to wear in such conditions are those that are breathable, waterproof, and allow sweat to escape—the best known of these is Gore-Tex®. Always carry the gear necessary to stay warm and dry, and make sure you can access it quickly if you are caught out— it can be the difference between comfort and misery.

HOW BREATHABLE WATERPROOF FABRIC WORKS

Breathable waterproof fabric is a laminate of three layers: two layers of nylon on either side of a thin layer of microporous teflon, which contains tiny pores—50 times smaller than the width of a human hair—that keep rain droplets out but allow perspiration (water vapor) to escape.

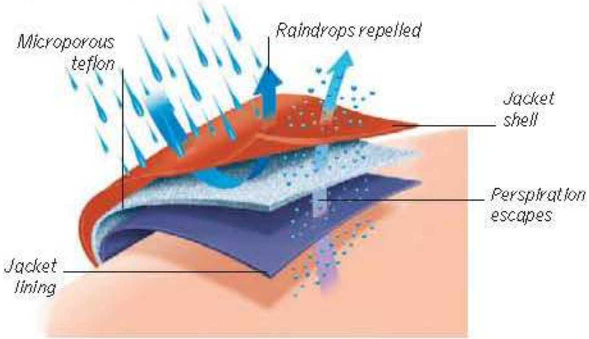
Microporous teflon
Raindrops repelled
Jacket shell
Perspiration escapes
Jacket lining

SOCKS

Socks are a vital part of your footwear system. Their function is two-fold. Firstly, they cushion the feet and prevent boots from rubbing and causing blisters. Secondly, they keep the feet warm and dry by wicking moisture away from the feet. Note that socks made from breathable waterproof fabric can be worn with regular boots but should not be combined with breathable waterproof boots.

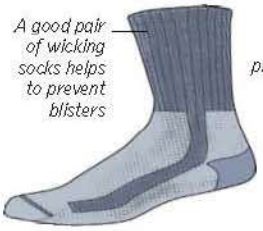
A good pair of wicking socks helps to prevent blisters

WICKING SOCK

A lightweight inner sock that helps wick moisture away from the skin and reduces friction.

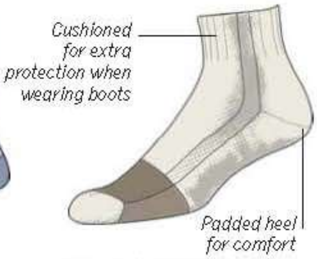
Cushioned for extra protection when wearing boots
Padded heel for comfort

WALKING SOCK

The thick fabric of the walking sock offers maximum cushioning, comfort, and protection.

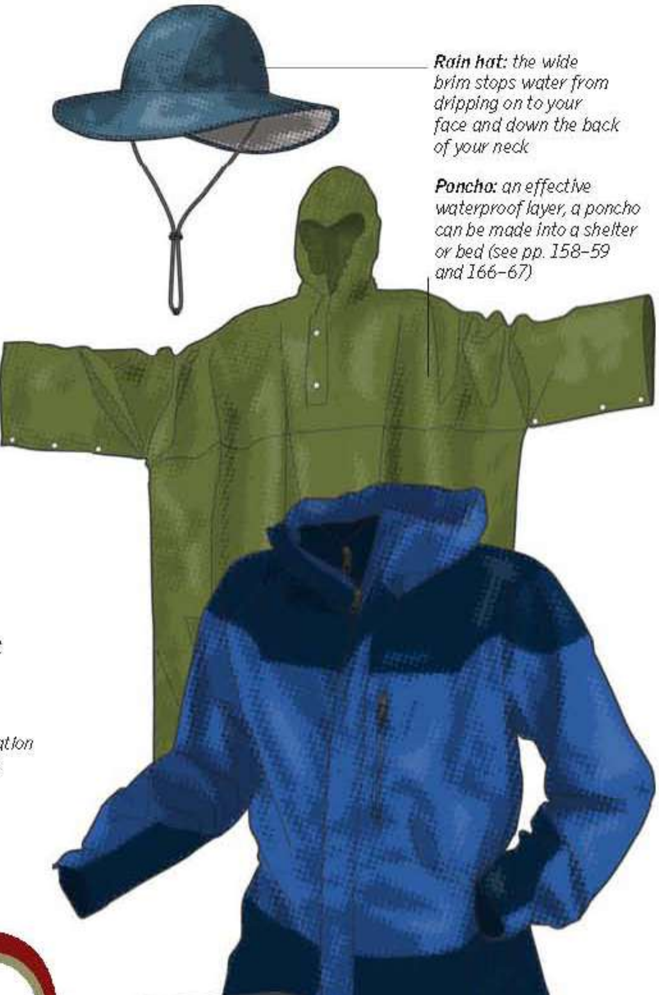

Rain hat: the wide brim stops water from dripping on to your face and down the back of your neck

Poncho: an effective waterproof layer, a poncho can be made into a shelter or bed (see pp. 158–59 and 166–67)

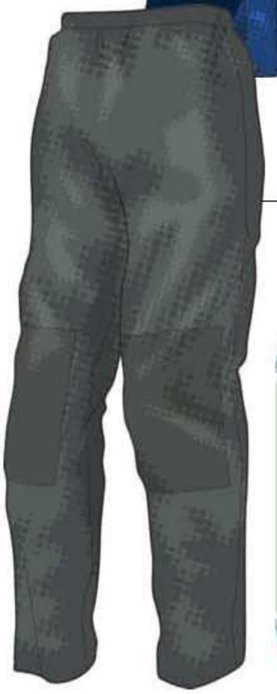

Waterproof jacket: allows sweat to escape while keeping rain out

Waterproof pants: keep your legs dry in wet conditions

GAITERS

Gaiters are useful, both in temperate areas for keeping the feet and lower pant legs dry, and in polar regions for keeping out snow.

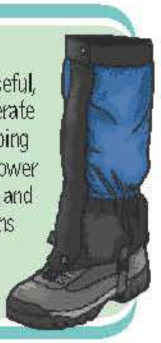

FOOTWEAR

A number of factors determine which type of footwear is right for you. The first concerns your personal requirements, which include the shape of your feet and the amount of support you need. You also need to consider how far you are traveling and the type of terrain you'll be covering. Then there is cost: footwear varies from the relatively cheap to the extremely costly. Bear in mind that if you intend to spend a lot of time in the wilderness, it pays to invest a little extra in comfortable, sturdy footwear.

BREAKING IN BOOTS

Your boots and feet should coexist in perfect harmony when you're on the trail. When you buy a pair of boots, wear them around the house and go on short hikes to make sure you have broken them in.

ANATOMY OF A SOLE

A boot sole has several layers to protect and support your feet. Look for a good tread on the outsole, and for cushioning beneath the heel and toe areas.

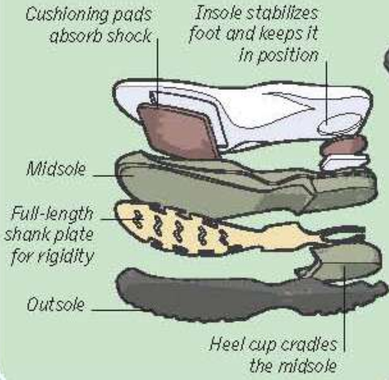
Cushioning pads absorb shock
Insole stabilizes foot and keeps it in position
Midsole
Full-length shank plate for rigidity
Outsole
Heel cup cradles the midsole

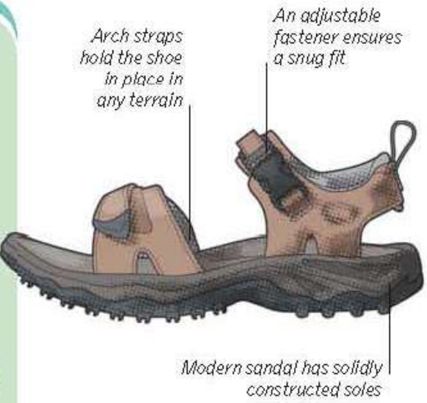
Arch straps hold the shoe in place in any terrain
An adjustable fastener ensures a snug fit
Modern sandal has solidly constructed soles

SANDALS

Modern sandal designs offer great support, and are extremely comfortable, while offering additional ventilation.

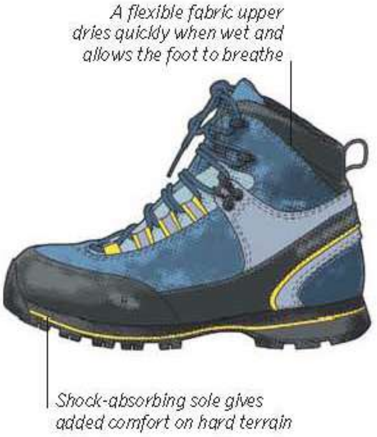
A flexible fabric upper dries quickly when wet and allows the foot to breathe
Shock-absorbing sole gives added comfort on hard terrain

LIGHTWEIGHT BOOT

Fabric and leather hybrids are increasingly popular because they combine the support and traction of a heavier shoe with the flexibility of a sports shoe, so there is less friction on the foot.

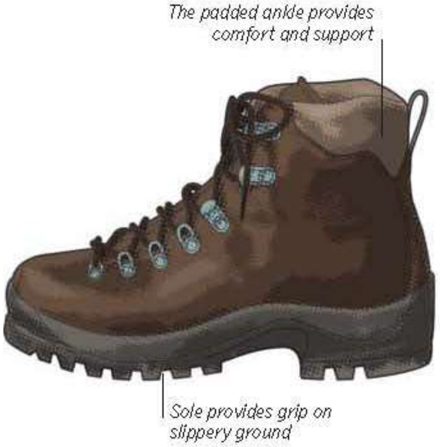
The padded ankle provides comfort and support
Sole provides grip on slippery ground

HIKING BOOTS

A compromise between weight, durability, and protection produces a good, all-round leather boot with a sole of hard-wearing rubber, such as Vibram®, water-resistant uppers, and valuable support for the ankles.

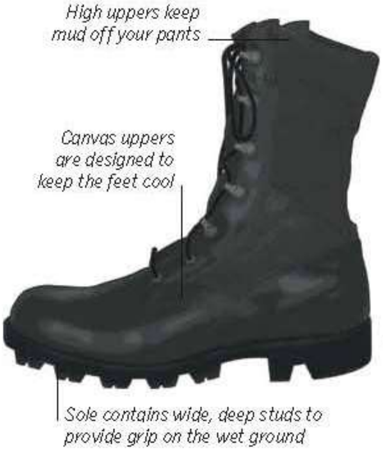
High uppers keep mud off your pants
Canvas uppers are designed to keep the feet cool
Sole contains wide, deep studs to provide grip on the wet ground

JUNGLE-TREKKING BOOT

This is a high boot made from rot-proof leather and canvas with a directly molded sole. Breather holes in the instep aid ventilation and help to drain moisture in hot or humid conditions.

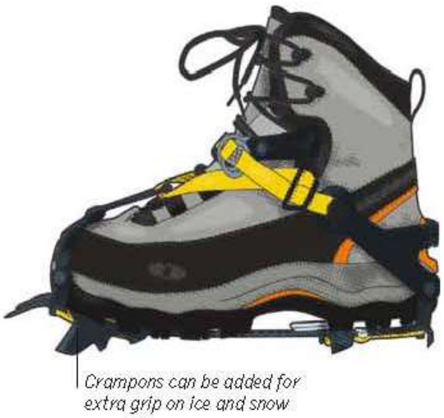
Crampons can be added for extra grip on ice and snow

CLIMBING/HIKING HYBRID BOOT

More flexible than traditional climbing boots, modern hybrid boots are designed to be worn with crampons and to keep the feet warm in very cold environments.

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—IN THE WILDERNESS

USEFUL EQUIPMENT

- Walking staff
- Pocket chainsaw
- Whistle, flashlight, signal mirror
- Animal deterrent spray
- Water purifiers
- Small AM/FM radio
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

BLAKE STANFIELD, 38, AND HIS 65-YEAR-OLD FATHER, Neil, endured being trapped beneath river ice, encounters with bears, and four days and nights in the Alaskan wilderness without food. They survived thanks to good survival skills, knowledge of the local area, and an abundance of water and firewood.

Father and son were dropped by floatplane on Friday, June 6, 2003, at a remote point on the Koyukuk River, deep inside Alaska. They planned to spend seven days rafting approximately 90 miles (145 km) to the town of Bettles. The trip had hardly begun when disaster struck—the raft crashed at speed into a massive sheet of ice, throwing the pair into the freezing water. They were both caught by the strong current and pulled under the ice. Somehow Blake was washed ashore, and he used a branch to haul his father to the riverbank.

Blake quickly realized that Neil was close to hypothermia, so used his waterproof lighter to light a fire of sticks and pine needles. Far from rescue, with their supplies lost and no one expecting their return for six days, their prospects looked bleak. With night approaching they built a shelter.

"THE RAFT CRASHED AT SPEED INTO A MASSIVE SHEET OF ICE"

The next morning, Blake decided to hike the 65 miles (105 km) to Bettles to get help. On Sunday his progress was brought to a halt by the confluence of two large rivers—the Tinayguk and Koyukuk. However, Blake realized that he had reached a landmark on the flight path of supply planes that frequently flew into the interior, so he built a signal fire and waited. Two days later, a pilot spotted him and radioed his position to a US Army base 200 miles (320 km) away in Fairbanks, then dropped him a two-way radio and supplies. The pilot then continued up the river to find Neil, dropping him a sleeping bag, tent, and supplies, before radioing his position to the Army. Both men were picked up the next day—exhausted, malnourished, and close to hypothermia, but alive.

WHAT TO DO

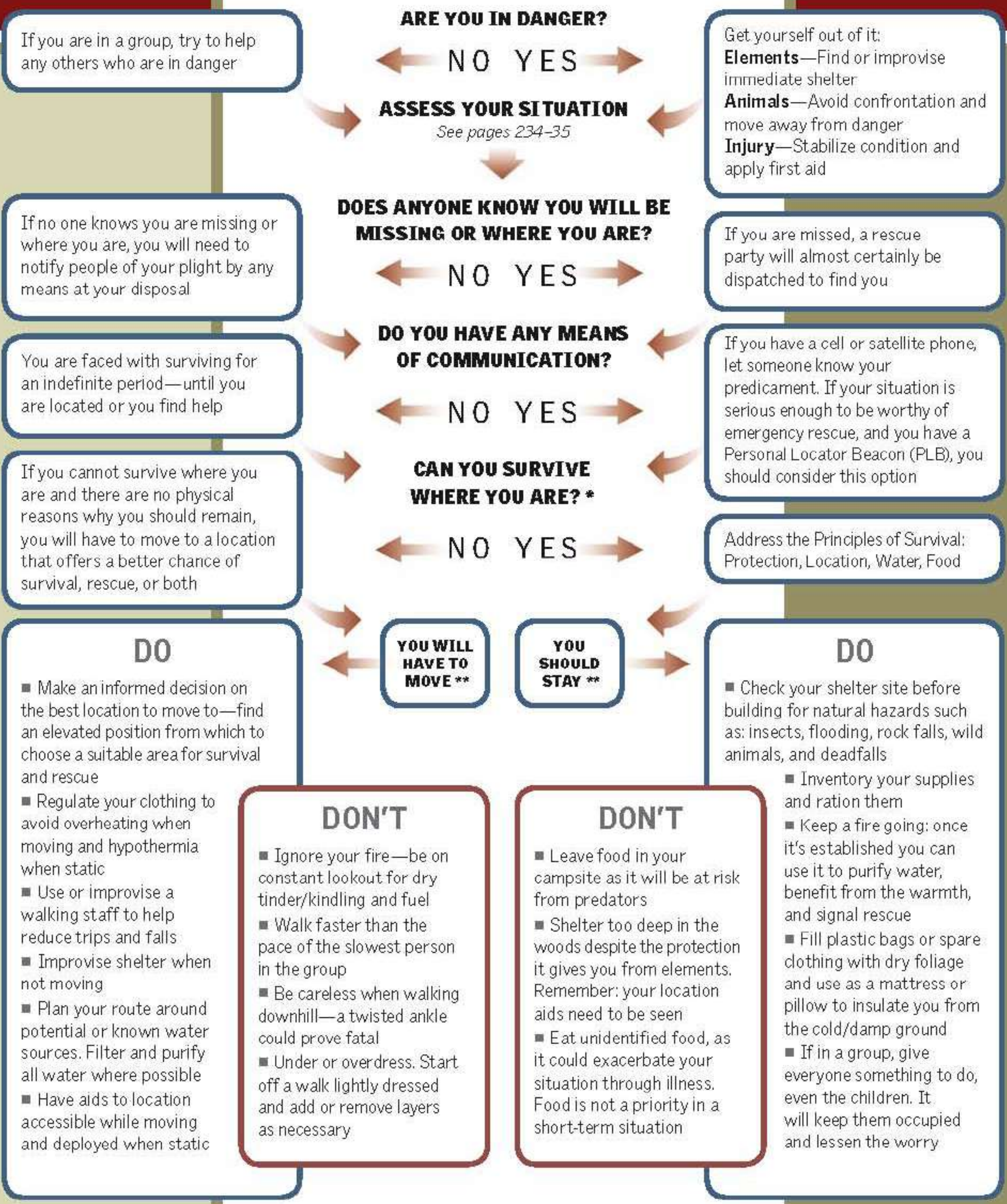

ARE YOU IN DANGER?

NO: If you are in a group, try to help any others who are in danger

YES: Get yourself out of it:
Elements—Find or improvise immediate shelter
Animals—Avoid confrontation and move away from danger
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO: If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

YES: If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO: You are faced with surviving for an indefinite period—until you are located or you find help

YES: If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO: If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers a better chance of survival, rescue, or both

YES: Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Make an informed decision on the best location to move to—find an elevated position from which to choose a suitable area for survival and rescue
- Regulate your clothing to avoid overheating when moving and hypothermia when static
- Use or improvise a walking staff to help reduce trips and falls
- Improvise shelter when not moving
- Plan your route around potential or known water sources. Filter and purify all water where possible
- Have aids to location accessible while moving and deployed when static

DON'T

- Ignore your fire—be on constant lookout for dry tinder/kindling and fuel
- Walk faster than the pace of the slowest person in the group
- Be careless when walking downhill—a twisted ankle could prove fatal
- Under or overdress. Start off a walk lightly dressed and add or remove layers as necessary

YOU SHOULD STAY **

DON'T

- Leave food in your campsite as it will be at risk from predators
- Shelter too deep in the woods despite the protection it gives you from elements. Remember: your location aids need to be seen
- Eat unidentified food, as it could exacerbate your situation through illness. Food is not a priority in a short-term situation

DO

- Check your shelter site before building for natural hazards such as: insects, flooding, rock falls, wild animals, and deadfalls
- Inventory your supplies and ration them
- Keep a fire going: once it's established you can use it to purify water, benefit from the warmth, and signal rescue
- Fill plastic bags or spare clothing with dry foliage and use as a mattress or pillow to insulate you from the cold/damp ground
- If in a group, give everyone something to do, even the children. It will keep them occupied and lessen the worry

---

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.

** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

SLEEPING SYSTEMS

REST IS AS ESSENTIAL for survival as food. A good night's sleep can offset much of the worry and stress of a difficult situation. Making sure your shelter is suitable for the environment you're in, and choosing the right sleeping bag, can be vital factors in survival.

SLEEPING BAGS

Although there are many different shapes and styles of sleeping bag, the one you choose should have enough padding to keep you warm at night even without a tent. A sleeping bag must never be allowed to get wet, so always keep it inside a waterproof cover, such as the bag's stuff sack or a bivy sack. Always try a sleeping bag before you buy it. If it's too snug, it will be less effective.

TIPS ON BUYING A SLEEPING BAG

There are many types of bag available. Here are a few things you should consider to ensure you buy the right sleeping bag for you.

- Determine the lowest temperature you're likely to encounter on your trip and choose a bag that will perform at this temperature.
- Synthetic bags are cheaper and easier to clean than down bags, and continue to provide insulation even when wet.
- Down bags are more expensive than synthetic bags, but provide a better warmth-to-weight ratio and last longer.
- A mummy-shaped bag provides better insulation than a rectangular-shaped bag, but has less room to move around in.

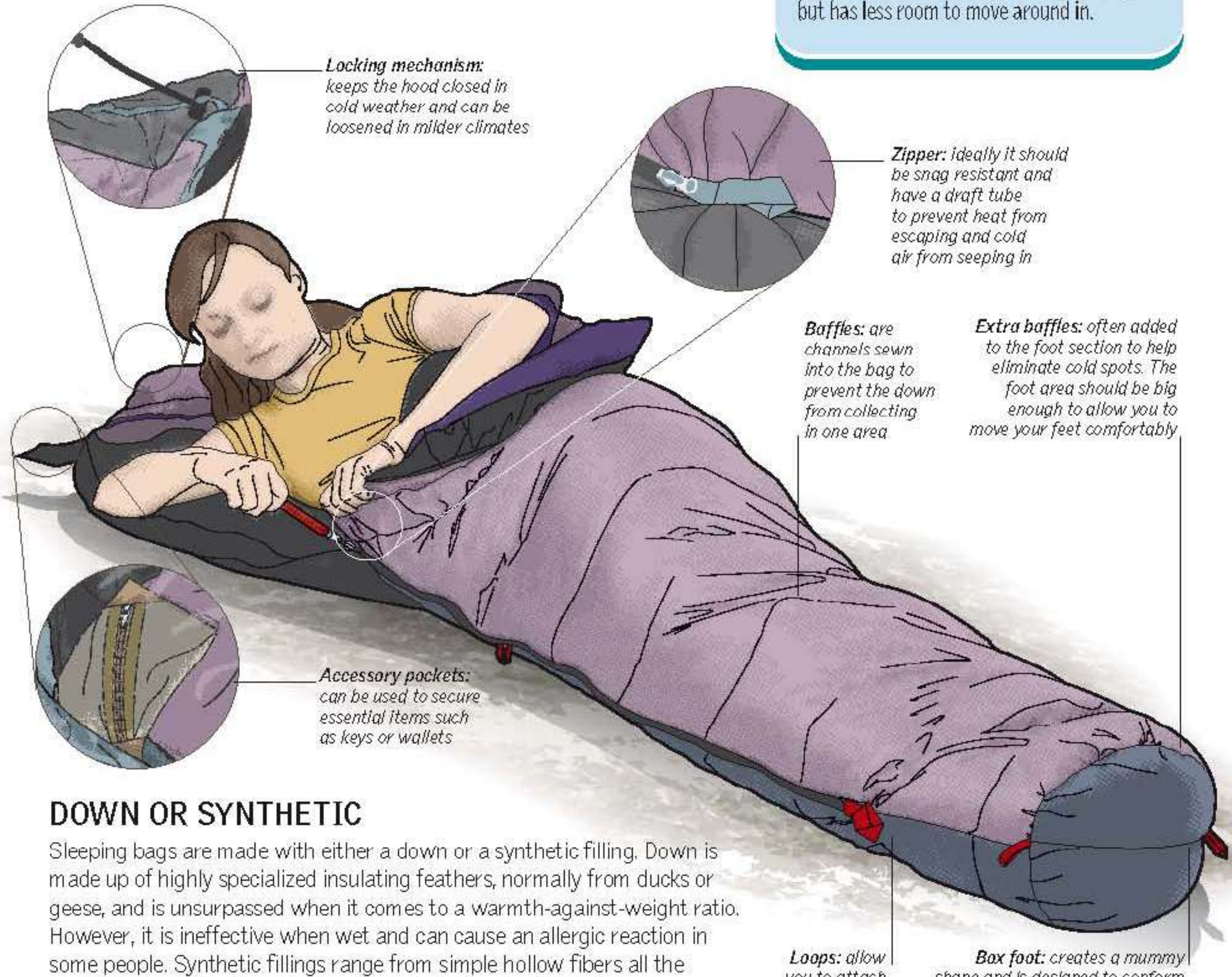

Locking mechanism: *keeps the hood closed in cold weather and can be loosened in milder climates*

Zipper: *ideally it should be snag resistant and have a draft tube to prevent heat from escaping and cold air from seeping in*

Baffles: *are channels sewn into the bag to prevent the down from collecting in one area*

Extra baffles: *often added to the foot section to help eliminate cold spots. The foot area should be big enough to allow you to move your feet comfortably*

Accessory pockets: *can be used to secure essential items such as keys or wallets*

Loops: *allow you to attach sleeping bag to the sleeping mat*

Box foot: *creates a mummy shape and is designed to conform to the position of your feet when you're sleeping on your side*

DOWN OR SYNTHETIC

Sleeping bags are made with either a down or a synthetic filling. Down is made up of highly specialized insulating feathers, normally from ducks or geese, and is unsurpassed when it comes to a warmth-against-weight ratio. However, it is ineffective when wet and can cause an allergic reaction in some people. Synthetic fillings range from simple hollow fibers all the way up to complex fibers designed to mimic the structure of down. They still retain some of their insulating capacity if they get wet.

TEMPERATURE RATINGS

Because manufacturers have no standardized way of registering how warm a sleeping bag is, temperature ratings for sleeping bags remain an inexact science. Some manufacturers give a bag both a comfort rating and a survival rating—known as a "season rating.' These give some indication as to how manufacturers expect a sleeping bag to perform.

CONSTRUCTION

The principle of all sleeping bags is the same: to trap air and prevent its circulation so that the body heats the trapped air and keeps you warm while you sleep. This can be achieved in a number of ways.

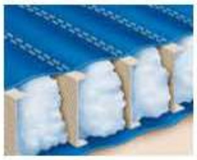

BOXWALL

Used in cold-weather down bags, the filling is contained within box-like sections to minimize bunching.

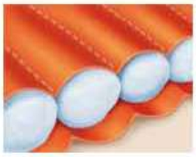

SEWN-THROUGH

Filling is in separate oval channels, although heat is still lost through the stitching.

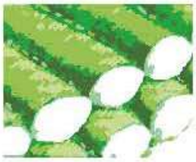

OFFSET CHANNELS

The filling is laid out in two layers to prevent bunching and heat loss through the stitching.

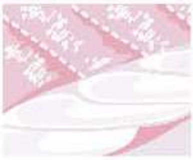

SHINGLE

Slanted layers of overlapping fibers that fill with warm air to aid insulation.

USING YOUR SLEEPING BAG

Before using your sleeping bag, always check inside for spiders and other dangers, and shake it to make sure the filling is distributed evenly. During use, if the bag is too warm, simply unzip it to cool down, or use the bag as a quilt. If it's too cold, either put on some clothing, or use a silk liner or waterproof bivi bag to increase its performance. Always air your sleeping bag well after using it.

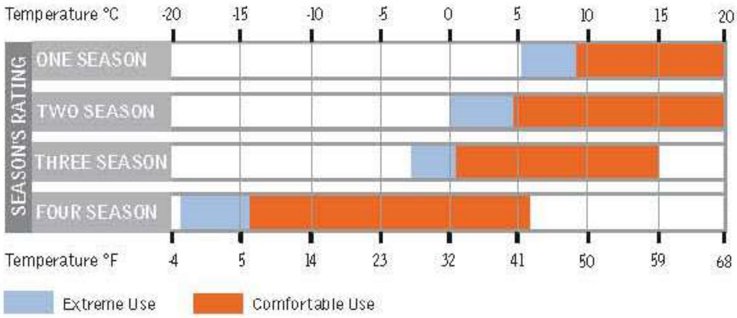

SLEEPING ACCESSORIES

You'll need to get a sleeping mat as well as a sleeping bag to ensure a comfortable, dry night's sleep. Additional gear, such as bivy sacks, sleeping bag liners, and inflatable pillows, will also increase comfort.

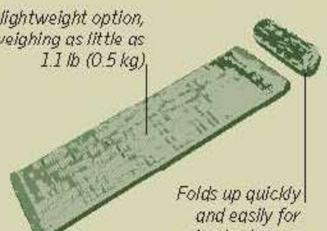

A lightweight option, weighing as little as 1.1 lb (0.5 kg)

Folds up quickly and easily for simple storage

FOAM MATS

The lightest and simplest type of mat, foam mats are best used on soft surfaces, such as sand or pine needles.

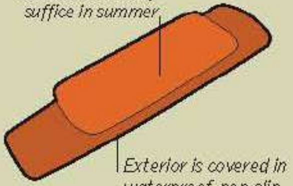

Shorter mats may suffice in summer

Exterior is covered in waterproof, non-slip polyester material

AIR MATS

They may be heavier and more expensive than foam mats, but air mats provide superior comfort and thermal insulation.

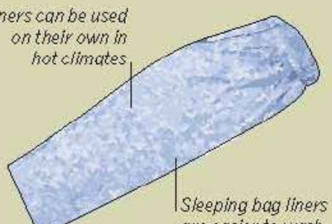

Liners can be used on their own in hot climates

Sleeping bag liners are easier to wash than sleeping bags

SLEEPING BAG LINER

Usually made of cotton or silk, sleeping bag liners trap air between your body and the sleeping bag.

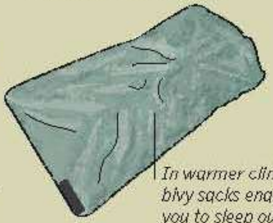

In warmer climates, bivy sacks enable you to sleep outside without a tent

BIVY SACKS

A simple waterproof bag that is used over a sleeping bag to keep it dry or used on its own as a shelter.

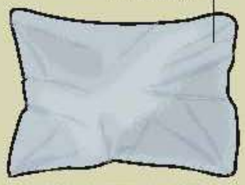

Outer layer of pillow is padded for extra comfort

INFLATABLE PILLOW

An inflatable pillow adds some luxury to expeditions. Alternatively, fill a stuff sack with spare clothing.

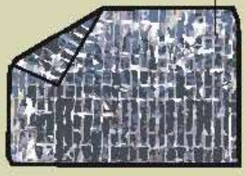

Foil prevents body heat from escaping and deflects it back to the body

SPACE BLANKET

A lightweight foil blanket for use in emergencies, space blankets can also be used as a temporary canopy.

PORTABLE SHELTERS

Shelter from the elements at night is essential for long-term survival. A variety of products are available to you, depending on the conditions you're about to enter and the amount you're prepared to carry. You can choose from lightweight options, such as hammocks and variations on bivy sacks, and the more traditional alternatives, tents, which are heavier but can accommodate up to eight people.

USING A HAMMOCK

If you have a hammock you can camp even if the ground is muddy, rocky, or on an incline. The advantage of a hammock is that it's lightweight and can be erected just about anywhere. Modern hammocks provide a shelter and bed solution in one portable package.

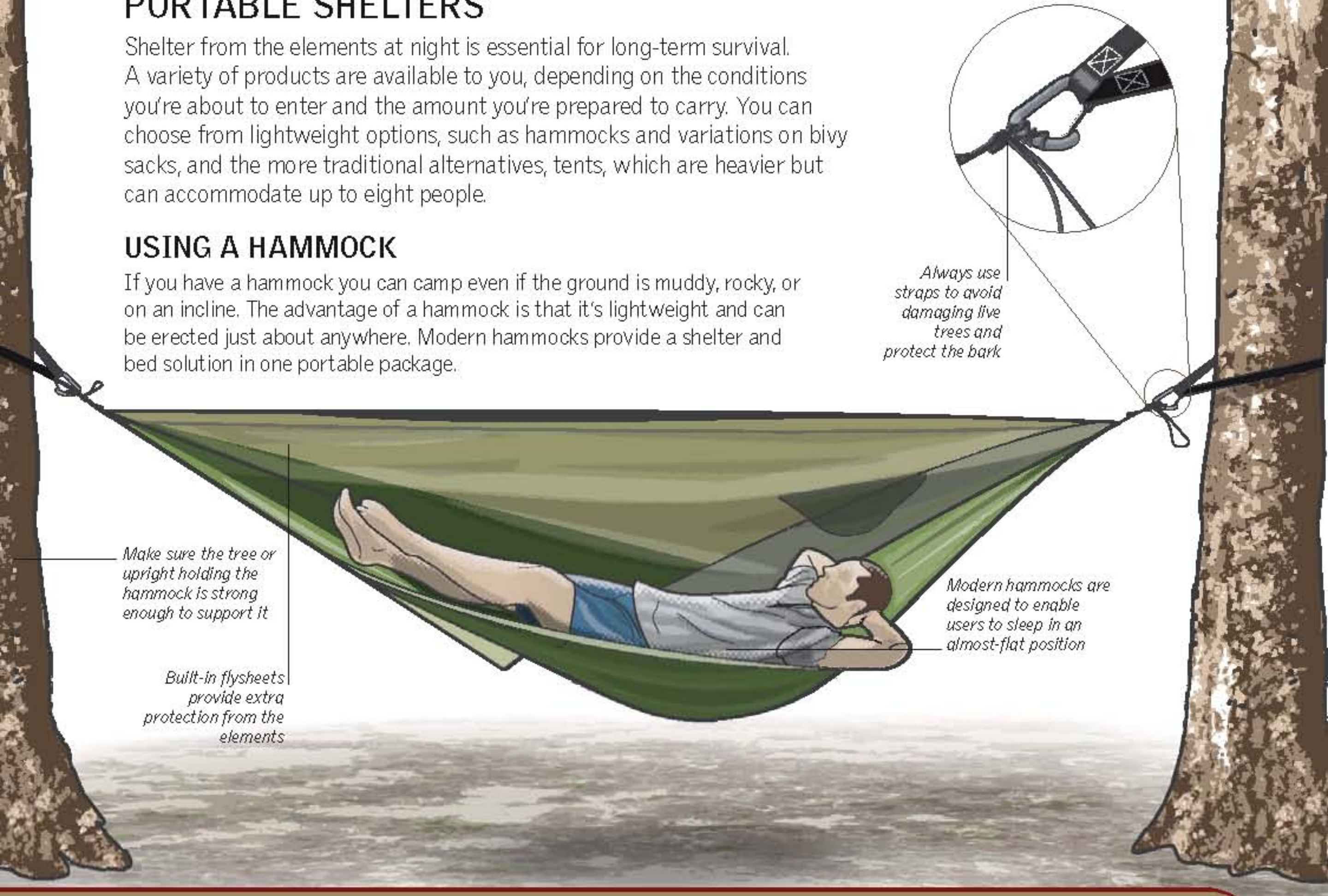

Always use straps to avoid damaging live trees and protect the bark

Make sure the tree or upright holding the hammock is strong enough to support it

Modern hammocks are designed to enable users to sleep in an almost-flat position

Built-in flysheets provide extra protection from the elements

USING A TARP TENT

A cross between a tarpaulin and a tent, a tarp tent is a highly versatile, extremely light form of shelter that's capable of withstanding even the most extreme conditions. Groundsheets should be laid out on the ground for added protection, and a suitable sleeping bag is essential. A one-person version can weigh as little as 18½ oz (0.52 kg).

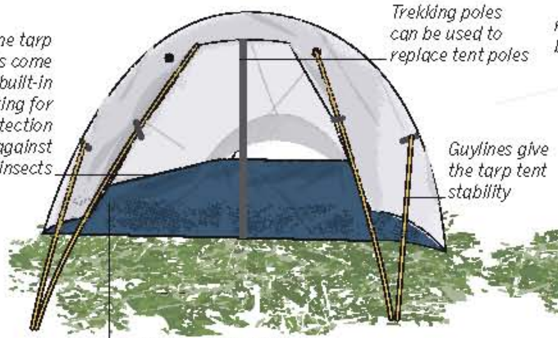

Some tarp tents come with built-in netting for protection against insects

Trekking poles can be used to replace tent poles

Guylines give the tarp tent stability

A groundsheet provides insulation from the cold, damp ground

USING A HOOPED BIVY

The hooped bivy is a waterproof and breathable bivy sack that has been designed to form a one-person shelter. It usually incorporates pegging points and a zippered and hooped entrance that form a small vestibule for your backpack. Many incorporate a heavy-duty groundsheet and can be set up and taken down in a matter of minutes.

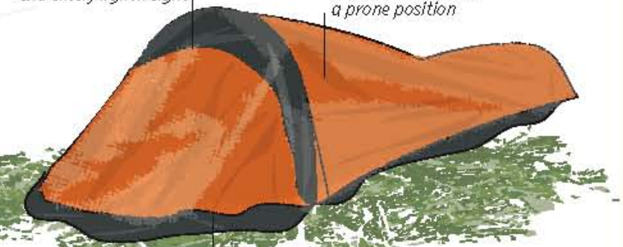

While they may have restricted space, hooped bivys pack small and are extremely lightweight

Limited height means you're restricted to a prone position

A small vestibule provides minimal storage for your gear

USING A TENT

Tents keep the rain and wind out while retaining warmth. The outer flysheet should be strong, impermeable, taut, and able to withstand high winds. The inner tent—which may be attached to a flysheet—is loose and made of permeable material to minimize condensation but retain heat. The inner and outer parts of the tent must not touch each other; if they do, heat may be lost and condensation may form on the underside of the flysheet, then seep into the tent.

WARNING!

In some climactic conditions, such as extreme high winds, severe storms, or heavy snow cover, stakes and guylines may not be enough to anchor the tent securely to the ground. This could result in the tent being blown away—and could have severe consequences for the tent's occupants. In such conditions, use extra means—such as rocks or logs—to anchor your tent securely.

THREE-SEASON TENT

Three-season tents are designed for use in a variety of climates. The inner layer is made of a lightweight material that offers both ventilation and protection against insects. The flysheet—which should extend about 4 in (10 cm) above the ground—and the groundsheet are made of waterproof fabric.

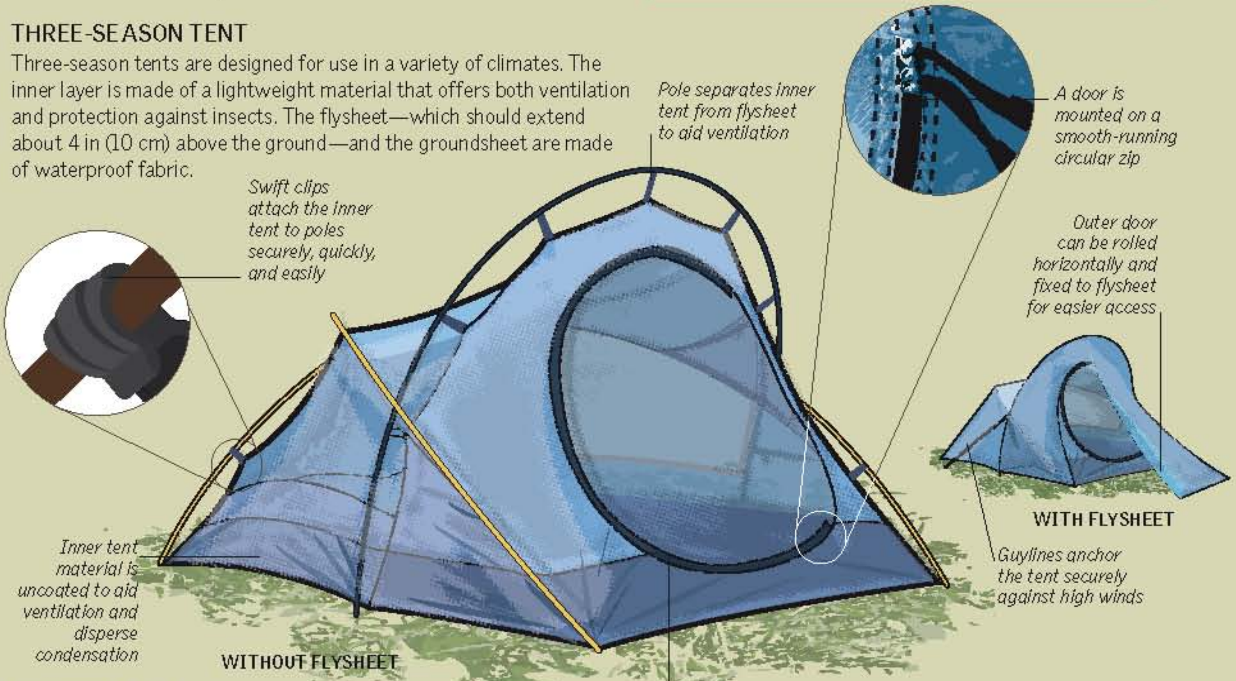

Pole separates inner tent from flysheet to aid ventilation

A door is mounted on a smooth-running circular zip

Swift clips attach the inner tent to poles securely, quickly, and easily

Outer door can be rolled horizontally and fixed to flysheet for easier access

WITH FLYSHEET

Guylines anchor the tent securely against high winds

Inner tent material is uncoated to aid ventilation and disperse condensation

WITHOUT FLYSHEET

Mesh door allows light to enter but keeps insects out

WINTER TENT

For cold-weather climates, look for a tent that offers extra-strong poles, storm windows, and ample pockets so that you can store your gear inside the tent. However, these extra features increase the tent's weight. The dome shape sheds snow well, and withstands high winds.

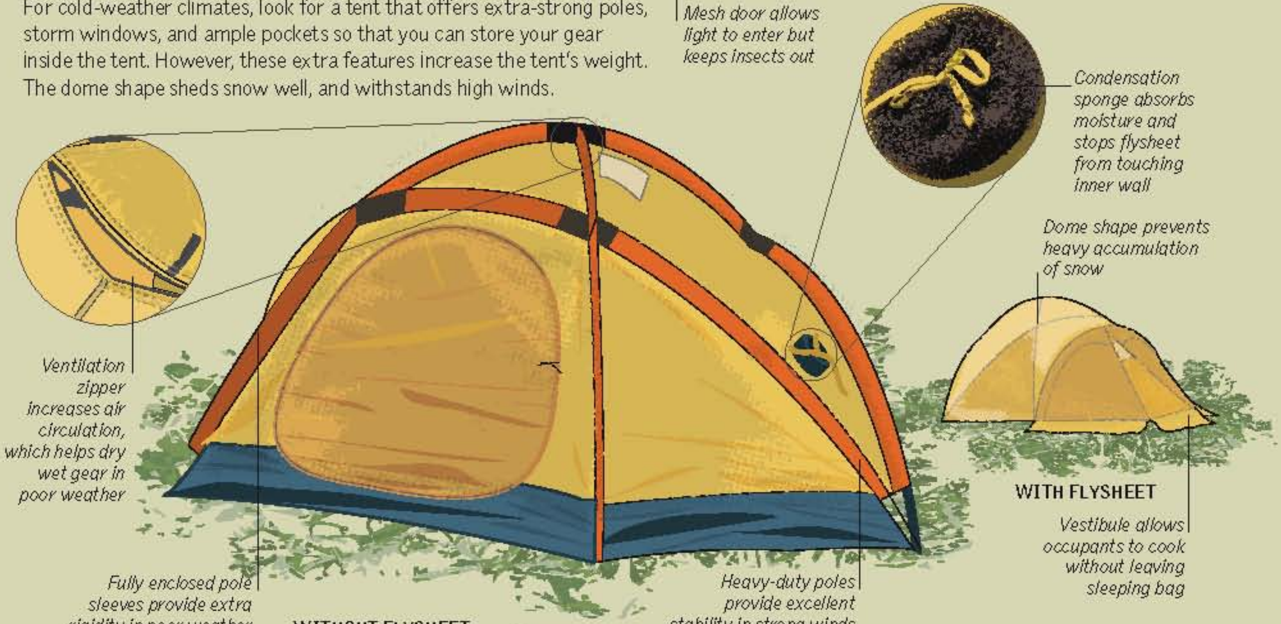

Condensation sponge absorbs moisture and stops flysheet from touching inner wall

Dome shape prevents heavy accumulation of snow

Ventilation zipper increases air circulation, which helps dry wet gear in poor weather

WITH FLYSHEET

Vestibule allows occupants to cook without leaving sleeping bag

Fully enclosed pole sleeves provide extra rigidity in poor weather

WITHOUT FLYSHEET

Heavy-duty poles provide excellent stability in strong winds

EATING ON THE TRAIL

THE FOOD YOU EAT determines how efficiently your body will function. This is never more important than when you're undertaking an expedition that will test your physical and mental endurance beyond normal levels.

FOOD FOR LIFE

Your body requires energy to fuel the chemical and biological reactions that take place within it, and to power your muscles. This, and other nutrients, comes from the food you eat (see below).

BODY FAT AND SURVIVAL

Being physically fit and active keeps you strong and less prone to injury, but being slightly overweight can actually be an advantage in a survival situation: 1 lb (½ kilo) of body fat converts to roughly 3,500 calories—the amount of energy you would usually expend during an average day.

TYPES OF NUTRITION	EXAMPLES	BENEFITS
CARBOHYDRATE	▪ Rice, pasta, bread, and cereals ▪ Fresh and dried fruit ▪ Root vegetables ▪ Chocolate and candy	▪ Major source of energy, as it converts most readily to glucose, which is your body's preferred source of fuel.
PROTEIN	▪ Meat and poultry ▪ Fish, such as tuna and salmon ▪ Eggs and dairy products ▪ Beans, legumes, and seeds	▪ Essential for growth and repair of muscle tissue. ▪ Increases your stamina and energy levels, and helps you fight fatigue. ▪ Makes you feel satisfied after eating.
FAT	▪ Nuts and seeds ▪ Dairy products, such as butter and cheese ▪ Meats, such as bacon and sausage ▪ Cooking oil	▪ Good source of energy. ▪ Makes you feel satisfied after eating.
FIBER	▪ Fresh and dried fruit, especially the skins ▪ Vegetables, beans, and seeds ▪ Whole-grain cereals ▪ Whole-meal bread, whole-meal pasta, and brown rice	▪ Provides bulk to help your digestion. ▪ Helps to make you feel full.
VITAMINS AND MINERALS	▪ Fresh and dried fruit ▪ Green vegetables ▪ Meat and fish ▪ Nuts and seeds	▪ Essential for growth and the repair of your body's tissues. ▪ Help you to maintain healthy teeth and eyes. ▪ Aids bloodflow and the production of energy.

BALANCING YOUR ENERGY LEVELS

You replace the energy that you expend on a daily basis through the food you eat. The energy obtained from food is measured in calories, and you require a certain number of calories each day to remain alive. If you consume too few calories, you will lose weight. However, eating too many calories will lead to weight gain, which can cause long-term health issues. The number of calories you require per day depends on many things, such as your age and level of activity (see right). When you don't replace the energy expended and continue to work, you start to use the energy store contained in your body's fat.

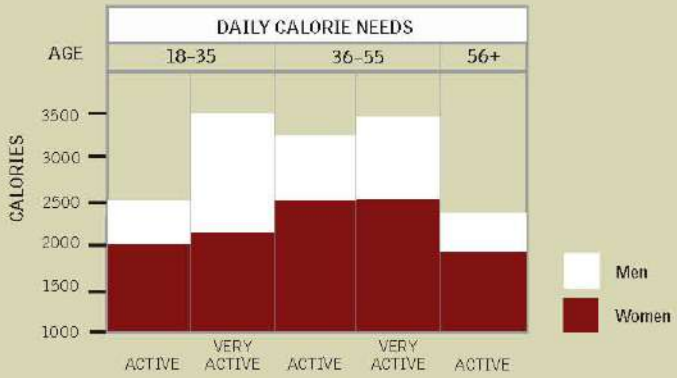

WILDERNESS DINING

Even if you're only going on a short hike, you should take food with you, such as fruit and chocolate, to replace the energy you're using. For longer expeditions, you'll need to take a number of different foodstuffs—for energy and for variety.

WARNING!

In a survival situation, don't eat for the first 24 hours unless to replace expended energy. Ration your food, as digestion uses up water.

TYPES OF FOOD	EXAMPLES	BENEFITS
BREAKFAST CEREALS AND DRIED FRUIT	▪ Trail mix—also known as "gorp" ▪ Muesli or granola ▪ Oatmeal ▪ Dried fruit	▪ Provide a vital source of carbohydrates, vitamins, and minerals to set you up at the start of the day. ▪ Contain dietary fiber to aid digestion.
TRAIL SNACKS	▪ Cookies and crackers ▪ Energy bars and cereal/granola bars ▪ Fruits, such as apples, pears, and oranges ▪ Nuts and seeds	▪ High in carbohydrates to help you maintain energy levels. ▪ Help to keep your hunger pangs at bay between main meals.
HIGH-CALORIE SNACKS	▪ Dehydrated ice cream ▪ Chocolate ▪ Hard candies ▪ Pudding mixes	▪ Provide a quick boost of energy, when required. ▪ Offer a comfort factor, although they are of limited nutritional value.
PRE-PREPARED MAIN MEALS	▪ Boil-in-the-bag meals—you can buy a wide variety of fully prepared meals, sealed in a foil bag. Cook for five to seven minutes. ▪ Freeze-dried and dehydrated meals—simply add boiling water, stir, and wait about five minutes. Extremely lightweight and take up little space. ▪ Canned meals—a great variety of them are available, but they are bulky and heavy.	▪ Provide a balanced range of food groups, vitamins, and minerals to help your body to recover and repair itself after the day's exertions. ▪ Quick and easy to prepare, and require few utensils. ▪ Many require minimal fuel to cook them, or water to rehydrate them. ▪ Can be bulked out with beans, legumes, and grains. ▪ Smaller cans may have a pull-top to open them.
MEAT, FISH, AND CHEESE	▪ Canned fish, such as sardines or tuna ▪ Cured meats, such as ham, salami, and beef jerky ▪ Hard cheese, such as cheddar and parmesan	▪ A great source of proteins and fats. ▪ Easy to store and long-lasting, although cans are heavier than other sources of food.
BEANS, LEGUMES, AND GRAINS	▪ Lentils and dried peas ▪ Dried beans and pearl barley ▪ Rice and pasta ▪ Couscous	▪ Can be used to bulk up main meals and help you to feel full. ▪ An important source of carbohydrates. ▪ Also provide fiber and protein.
MIXERS	▪ Flour ▪ Salt ▪ Suet ▪ Sugar	▪ Flour and salt can be used to make dough. ▪ Sugar and salt help to make wild food palatable. ▪ Suet is an invaluable source of fat that can be used to supplement meals.
FLAVORINGS	▪ Powdered soup ▪ Tomato puree ▪ Powdered sauces ▪ Gravy cubes	▪ Add flavor and variety to your meals. ▪ Especially useful if you're cooking with wild food.
BEVERAGES	▪ Tea ▪ Coffee ▪ Chocolate powder ▪ Dried milk	▪ Provide warmth and comfort, although they are of limited nutritional value. ▪ Chocolate powder is high in carbohydrates. ▪ Dried milk is an excellent source of calcium.

CAMPING STOVES

ALTHOUGH NOTHING CAN BEAT the satisfaction of building your own campfire, the convenience and reliability of purpose-made camping stoves make them an essential piece of cooking equipment, especially in areas where open fires are banned.

THE BENEFITS OF CAMPING STOVES

Modern camping stoves are very light and collapse down to an incredibly small, compact size. Gas-fueled stoves are by far the most common, but liquid-fueled stoves are also worth considering. Both have their own advantages and disadvantages (see below), but your choice will be dictated by space and weight constraints, the distance you're traveling, and environmental factors. A hexamine stove is a good item to have in a survival situation (see opposite).

CHOOSING A GAS- OR LIQUID-FUELED STOVE

Gas-fueled stoves are safer, cleaner, lighter, and easier to use than liquid-fueled stoves, but are less economical, and not suitable for low-temperature, high-altitude use. Liquid stoves offer greater stability, fuel versatility and economy, but you need to pump and prime (preheat) them before use, and they are less clean.

TRANSPORTING FUEL

There are many factors to consider when transporting fuel to make your expedition safer:

- When hiking, pack your fuel container upright in the middle of your backpack, and surround with clothing to protect it from knocks. Make sure you can access it easily when you reach camp. Use similar principles when traveling by other methods.
- Ensure that your fuel bottles look very different from your water bottles to prevent confusion.
- If you're going on a long expedition, check the availability of the fuel, and consider how you will dispose of cylinders safely.

		GAS-FUELED STOVES	
TYPE OF STOVE	 **BUTANE OR ISOBUTENE**	 **PROPANE**	 **BLENDED FUEL**
INFORMATION	▪ Both fuels sold in disposable, pressurized canisters.	▪ Fuel sold in disposable, pressurized canisters. ▪ Used in most portable stoves and barbecue grills.	▪ Fuel sold in disposable, pressurized canisters. ▪ Is a mix of butane, propane, and/or isobutene.
PROS	▪ Lightweight, sealed fuel container. ▪ Butane very efficient and burns at a high temperature. ▪ Isobutene more efficient than butane, and performs better in cold conditions.	▪ Lightweight, sealed fuel container. ▪ Fuel burns at a very hot, steady heat, and produces an almost sootless flame. ▪ Performs very well in cold conditions.	▪ Lightweight, sealed fuel container. ▪ Fuel safer than pure propane, and performs better than either pure butane or isobutene in cold conditions. ▪ Isobutene blends give a more efficient flame.
CONS	▪ Higher cost and lower heat output than liquid fuels. ▪ Fuel efficiency reduced at higher altitudes. ▪ Butane less efficient than isobutene in temperatures below 50°F (10°C).	▪ Higher cost and lower heat output than liquid fuels. ▪ Fuel efficiency reduced at higher altitudes. ▪ Highly combustible, so less safe than other fuels.	▪ Higher cost and lower heat output than liquid fuels. ▪ Fuel efficiency reduced at higher altitudes.

USING CAMPING STOVES SAFELY

Always take care when storing the fuel and using your stove. Use only the stove manufacturer's recommended fuel, and check all connections and fittings before each and every use. Also:

- Clear the cooking area of dry vegetation and leaves, in case you spill fuel or knock over your stove.
- Place your stove on a level and stable surface.
- Keep naked flames and sparks away from the fuel, especially when filling or refilling the stove. Before refilling, make sure the flame is out and the stove has cooled down.
- Use the cooker in a well-ventilated space. Liquid fuels, in particular, can give off poisonous fumes, such as carbon monoxide, and all cookers also burn vital oxygen.
- Don't use a stove inside your tent. In addition to the risk of poisonous fumes, fire is also a significant danger. A confined space with limited access, your tent can be reduced to ashes in minutes, along with everything inside, and you may not be able to escape in time to protect yourself. Losing your tent and your equipment is bad enough, without you receiving third-degree burns at the same time.
- Don't leave a burning stove unattended. If it topples over, it could easily start a fire. If the stove flame goes out, a spark or naked flame could also ignite the vapors and cause an explosion.

THE HEXAMINE STOVE

A very useful piece of emergency equipment, the hexamine stove (or hexi-stove) burns solid blocks of hexamethyl-enetetramine (also known as "methenamine"), and folds away compactly. The stove acts as a small windbreak and cooking platform, while the fuel blocks it uses provide a very reliable means of starting a fire in any conditions. The blocks produce a smokeless flame, don't liquefy while burning, and leave no ashes.

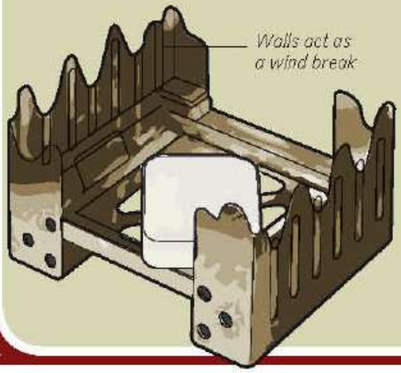
Walls act as a wind break

LIQUID-FUELED STOVES			
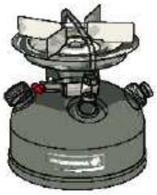 **WHITE (COLEMAN) FUEL**	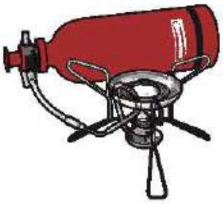 **MULTI-FUEL**	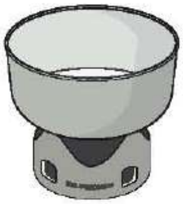 **DENATURED ALCOHOL**	**TYPE OF STOVE**
▪ Fuel sold in disposable canisters.	▪ Stove can use a variety of fuels, from unleaded gas, kerosene (paraffin), aviation fuel, and white fuel, to diesel—most of which are readily available.	▪ Fuel sold in disposable plastic containers. ▪ Stove basically consists of a vessel to burn the fuel in, and a windshield that also acts as a cooking platform.	**INFORMATION**
▪ Fuel can withstand low temperatures, high altitudes, and most weather conditions. ▪ Produces a clean flame.	▪ Stove uses a range of fuels. ▪ Fuel is non-pressurized, so can be stowed safely in almost any airtight container. ▪ Stove tends to perform reliably, efficiently, and quickly, producing a hot flame.	▪ Fuel is non-pressurized, so can be stowed safely in almost any airtight container. ▪ Economical and produces a clean flame. ▪ Stove is lightweight, stable, and safe. ▪ Usually supplied with pots and pans.	**PROS**
▪ Stove must be pumped to generate required pressure for use. ▪ Fuel needs to be primed.	▪ Potential for spillage of flammable fuel. ▪ Stove must be pumped to generate required pressure for use. ▪ Fuel needs to be primed. ▪ Some fuels produce soot and noxious fumes.	▪ Potential for spillage of flammable fuel. ▪ Stove slower to heat up than other liquid stoves, because it's not pumped. ▪ Flame vulnerable to wind and other environmental conditions.	**CONS**

YOUR SURVIVAL KIT

A BASIC SURVIVAL KIT is an essential item to take with you on any outdoor expedition. It should be compact enough to carry on your person at all times, and its contents should address the key principles of survival: protection, location, water, and food.

EXAMPLE OF A SURVIVAL KIT

Choose a tin with a waterproof seal and locking clasps. While you can buy ready-made kits, you should always adapt the contents, both to your needs and to the environment you'll be in. Ideally, items in the survival kit should be high-quality and multi-purpose—for example, the tin itself could be used as a cup, a small cooking pot, or even a signal mirror. Your kit should contain the following:

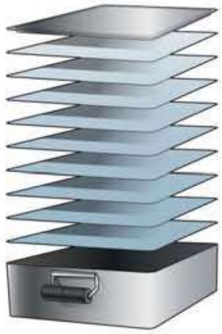

LAYERS OF A SURVIVAL TIN

LAYER 1

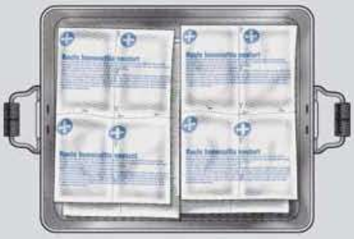

- **Blister medical pads and bandages:** useful for minor cuts and blisters. Include a range of bandages (both waterproof and fabric) in a selection of sizes. Waterproof bandages can also be used to mend holes in tents and tarpaulins.

LAYER 2

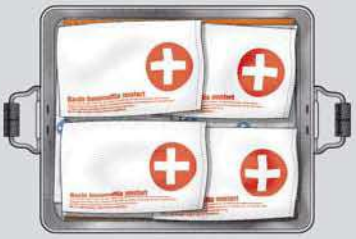

- **Antiseptic wipes:** for treating wounds and cleaning bites. Can also be iginited by a spark and used to get a fire going.

LAYER 3

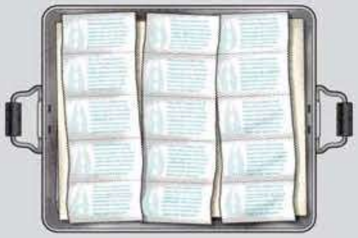

- **Water purification tablets:** choose from tablets or iodine (but don't use iodine if you're allergic to it, or to shellfish). Water will need to be filtered before being purified.

LAYER 6

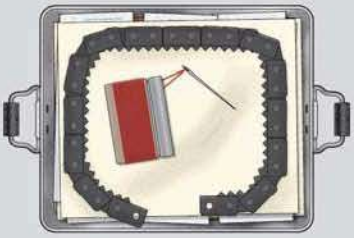

- **Survival saw, or pocket chainsaw:** can be wrapped around the inside of the tin or cut in half, if space is limited.
- **Single-edged razor:** multi-purpose—from skinning an animal to cutting cord. Store in its packaging.
- **Needle and cotton:** use strong, waxed cotton, pre-threaded through the needle.

LAYER 7

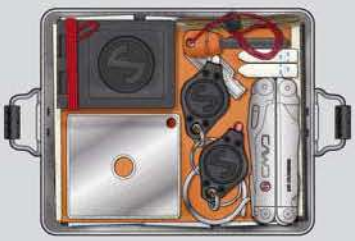

- **Flashlights:** two small Photon (LED) lights—one white and one red—taped in "off" position.
- **Mini multi-tool:** see panel, opposite.
- **Chainsaw handles:** use with pocket chainsaw.
- **High-viz card, signal mirror:** location aids.
- **Compass:** an emergency back-up.
- **Flint, fire steel, tampons:** for starting fires.
- **Snare wire:** for animal traps and lashing.

LAYER 8

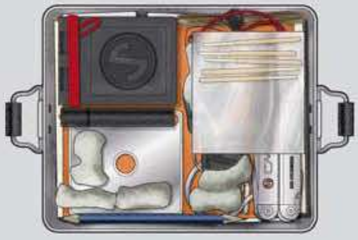

- **Waterproof matches and tinder balls:** for starting fires. Store the matches in a small, resealable plastic bag.
- **Pencil:** sharpened at both ends.
- **Potassium permanganate:** dissolve in water to sterilize water at low concentration, and to clean wounds at a high concentration. Can also be used with sugar to make fire. Store in a waterproof container.

CHOOSING A MULTI-TOOL

A useful item to carry with you on your trip, multi-tools are usually designed with a specific task or hobby in mind, and include useful features for survival, such as a small compass, pliers, saw, and various blades. Use one that has quality tools that most suit your needs, with a comfortable grip, and ensure that the blades are lockable to prevent injuries. Carry your multi-tool in a secure pocket, or in a pouch on your belt, and include a miniature version in your survival tin (see below). Bear in mind that your multi-tool should be an addition to, not a replacement for, your bushcraft knife (see p. 146).

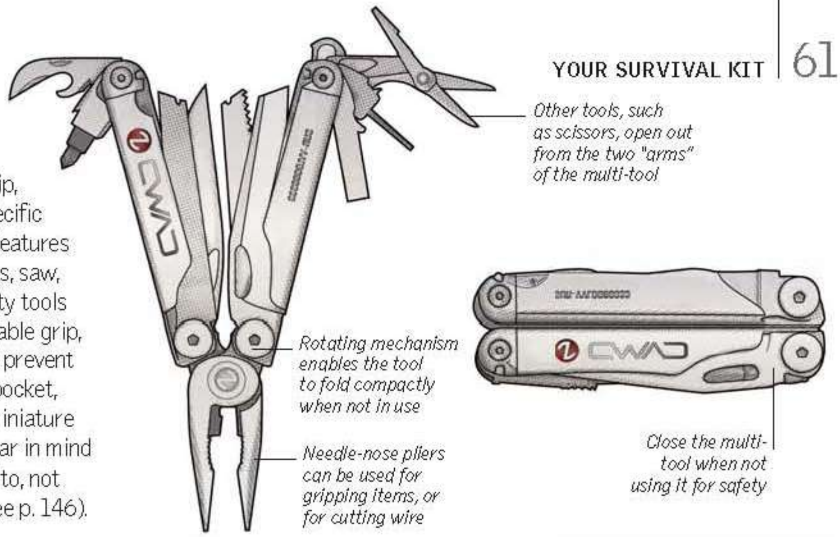

Other tools, such as scissors, open out from the two "arms" of the multi-tool

Rotating mechanism enables the tool to fold compactly when not in use

Needle-nose pliers can be used for gripping items, or for cutting wire

Close the multi-tool when not using it for safety

LAYER 4

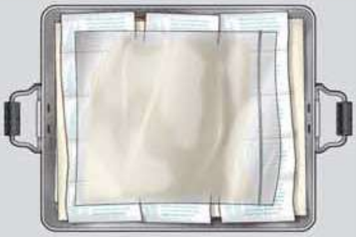

- **Petroleum jelly:** apply to chapped lips, rashes, and sores. Can also be smeared onto tampons to make them burn longer. Store in a small, resealable plastic bag.

LAYER 5

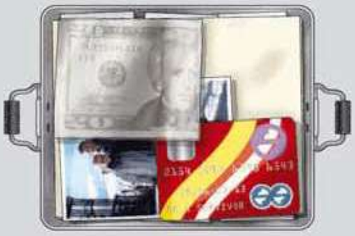

- **Waterproof notepaper:** for drawing maps or leaving messages.
- **Photograph of loved ones:** a psychological incentive in a survival scenario.
- **Credit card:** an effective means of extracting insect stings (see pp. 266–67).
- **Money:** wrapped in cellophane.

LAYER 9

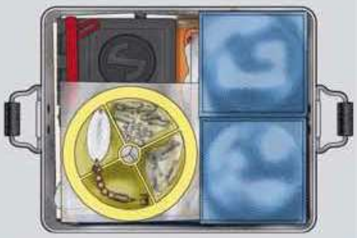

- **Non-lubricated condoms:** can be used to carry water, or as a waterproof cover for smaller items, such as your mobile phone.
- **Mini fishing kit:** if you are near water, fish can be easier to catch, prepare, and cook than mammals. Fishing line can be also used for other survival tasks. Should contain a selection of hooks, flies, swivels, and split-shots.

TIN LID SEALED WITH TAPE

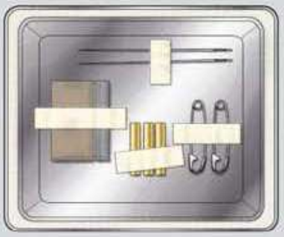

- **Sailmaker's needles:** multi-purpose—can be used as an arrow point, or for mending tents and tarpaulins. Wide eyes are best.
- **Safety pins:** for securing clothing, or mending your sleeping bag or tent.
- **Mini glowsticks (cyalumes):** useful for emergency lighting, and as a location aid.
- **Single-edged razor:** multi-purpose (see Layer 6). Store in its packaging.

ADDITIONAL USEFUL ITEMS TO INCLUDE

While your survival tin's size may be restrictive, you can always improve your kit with items that can fit on your belt, or in a belt pouch, known as a "belt order." This will form part of your first-line equipment (see p. 43).

- **Space blanket or aluminum foil:** can be used as a signaling device; for shelter; to carry, store, and heat water; or to cook in. Many are dual-sided: one silver, the other green for camouflage, or orange to stand out.
- **Plastic bags:** you can never have too many. Numerous uses—from a water carrier to a transpiration bag.
- **Medicines (such as painkillers and antibiotics):** should not replace your main first-aid kit, but ensures you have some basics if you're separated from your gear.
- **Small candle:** once lit, this will provide a reliable flame that you can build your fire around.
- **Nylon stockings:** can be used for warmth, or as an improvised water filter, mosquito net, or fishing net.
- **Small AM/FM radio:** battery- or solar-powered.
- **Surgical tubing:** enables you to reach water in otherwise inaccessible rock crevices.
- **Fire tin or matchless fire set:** self-contained methods of starting a fire, when no natural fuel is available.

ON THE TRAIL

FIND YOUR WAY

Before you set off on an expedition, you should have at least a basic understanding of how to read a map and use a compass. Your ability to correctly assess a map of the area you intend to visit will allow you to make informed decisions during your preparation. If you understand the area and terrain, your chances of getting lost will be reduced, and you will be able to continually evaluate your progress and therefore alter your plans as necessary. You will also be able to plan the safest and most appropriate route, and locate water, shelter, and areas that will allow you to use your location aids properly. If you're proficient with a map and compass, you'll have no cause to worry about getting lost or straying off-track and will be free to enjoy your outdoors experience.

In a survival situation, you will be faced with many tough decisions. You may have to decide whether to stay where you are and await rescue or move to an area that offers a better chance of survival and rescue. Your ability to navigate effectively—whether

In this section YOU WILL DISCOVER...

- how to **use your map** to tell the **lie of the land...**
- that the **path ahead** is as long as a **piece of string...**
- the difference between **eastings** and **northings** and an **ERV** and a **GPS...**
- how a **detour** could be the **most direct route...**
- how to **navigate** using your **hair...**
- why an **anvil of cumulonimbus** is no **fair-weather cloud...**

by using a map and compass or by using natural features—will play a major role in your decision-making process. Whilst a Global Positioning System (GPS) is an excellent aid, it relies on batteries and technology—both of which can fail.

An understanding of weather patterns will also allow you to evaluate conditions and make informed decisions as you travel. The ability to assess your situation and modify your plans means that you will be able to avoid many potential survival situations.

The ability to take a bearing (see pp. 68–71) and navigate using pace counting (see p. 72) could prove to be crucial skills in a survival situation.

TEST HOW ACCURATELY you can walk on a set bearing and pace out a set distance by following the exercise below. If you have stayed on your bearings and your pacing has been accurate, you should finish at your starting point. Pick a suitable area where you can walk at least 330 ft (100 m) in any direction. Don't cheat by heading for your marker! For this exercise to be useful, make sure you follow your bearings and count your paces to navigate.

① Place a marker on the ground. Dial a bearing onto your compass (110° in our example).

② Walk on this bearing, counting your paces until you think you have traveled 330ft (100m), then stop.

③ Add 120° to your original bearing. Dial this new bearing (230° in our example) onto your compass.

④ Pace out another 330 ft (100 m) on your new bearing, then stop.

⑤ Add another 120° to your latest bearing. Dial a last bearing (350° in our example) onto your compass.

⑥ Walk on your final bearing for a further 330 ft (100 m). You should be back at your starting point.

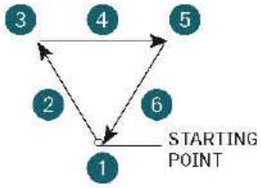

Never underestimate the **combined power** of a map, a compass, a GPS unit, and the **skills and knowledge** to use them well—your life may **depend on it**

MAPS AND MAP-READING

A MAP IS A TWO-DIMENSIONAL representation of a three-dimensional area—from a map you can determine distance and height on the ground. If you are able to read a map, and can interpret the information it contains, you can visualize what an area looks like and use these features as landmarks to make navigation easier. Your map is very important, so keep it safe.

UNDERSTANDING MAPS

Although there are many types of map available—with varying levels of detail and scale—topographic maps are best for hiking. They show important features, such as rivers, roads, railroads, paths, buildings, and forested areas, and also depict the lie of the land through the use of contour lines (see opposite) to represent height.

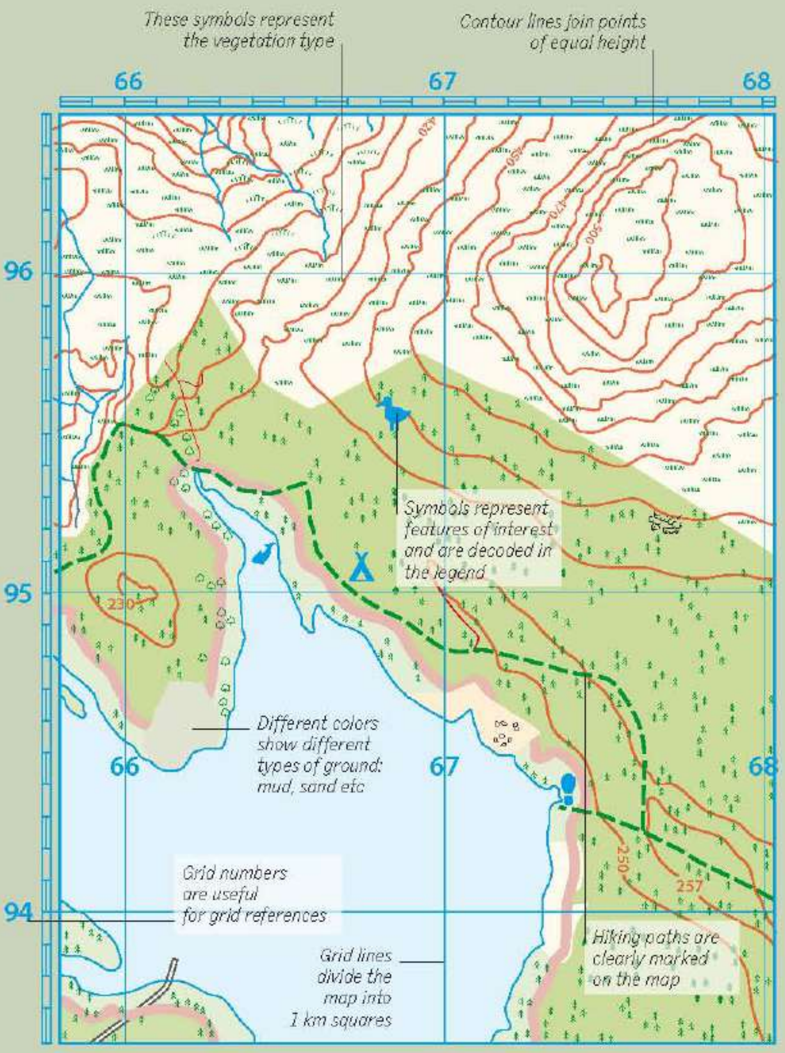

These symbols represent the vegetation type

Contour lines join points of equal height

Symbols represent features of interest and are decoded in the legend

Different colors show different types of ground: mud, sand etc

Grid numbers are useful for grid references

Grid lines divide the map into 1 km squares

Hiking paths are clearly marked on the map

THE LEGEND

Topographic maps incorporate a legend, or key, which deciphers the information shown on the map. It's important to familiarize yourself with the symbols used, as this will help you visualize what's being represented on the map more effectively. Some examples are listed below:

HEIGHTS AND NATURAL FEATURES

🟦 *Water* ⬜ *Mud*

🟧 *Sand; sand and shingle*

Vertical face/cliff

Loose rock *Boulders* *Outcrop* *Scree*

Coniferous trees

Non-coniferous trees

Managed woodland

Orchard

Scrub

Bracken, heath, or rough grassland

Marsh, reeds, or wetlands

TOURIST AND RECREATIONAL INFORMATION

Nature reserve *Campsite*

Fishing area *Walks/Trails*

THE SCALE

Hiking maps are always drawn to a scale provided in the legend. This is a ratio of how much you would have to enlarge the map to reach actual size. 1:25,000—whereby 4 cm on the map equates to 1 km on the ground (2½ in to 1 mile)—is the most useful level of detail for hikers. A smaller-scale of 1:50,000, for example, will provide a more basic overview of the terrain.

MEASURING DISTANCE

Maps are drawn to scale so that you can use them to accurately estimate distances on the ground. Being able to measure distance is important because it means you can calculate the most direct—and energy-efficient—route to your destination. In a survival situation, every last bit of energy counts, so the shorter the distance, the better.

USING THE GRID LINES

The simplest method of measuring distance on a map is to use the grid lines—on a scale 1:25,000 map each grid square represents 1 km (traveling diagonally across a square, it's approximately 1.5 km). You can also lay a piece of paper between the two points, mark the start and end of your route, and place it underneath the scale line to read the distance.

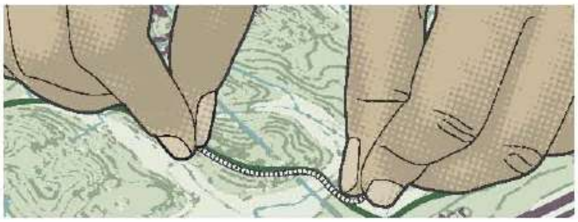

USING STRING OR SOLDER

Invariably you will deviate from a straight line and will have to navigate around obstacles or bends. A far better way of measuring distance is therefore to take a piece of string, curve it around your intended route, and transfer it to the scale line. Solder wire (use wire that is lead-free) is even more accurate because it holds its shape on the map and remains flexible.

CONTOUR LINES

Topographic maps feature lines called contour lines. A contour joins points of equal height above sea level, and allows the topography of the ground to be depicted in detail. The contour interval is specified in the legend—for maps with a 1:25,000 scale you would usually expect to see a 5-meter vertical interval between each line, although for maps of mountainous regions this interval may be 10 meters. The ability to look at the contour lines, and imagine how they translate to the ground, is a skill that takes a little while to acquire, but once mastered will allow you to read a map more proficiently.

USING CONTOURS

Knowing how steep the ground is will greatly improve your navigational skills, as well as your route-planning (see p. 73). Walking up and down hills uses a lot of energy, so it is far better to follow the contour lines on your map to go around hills instead.

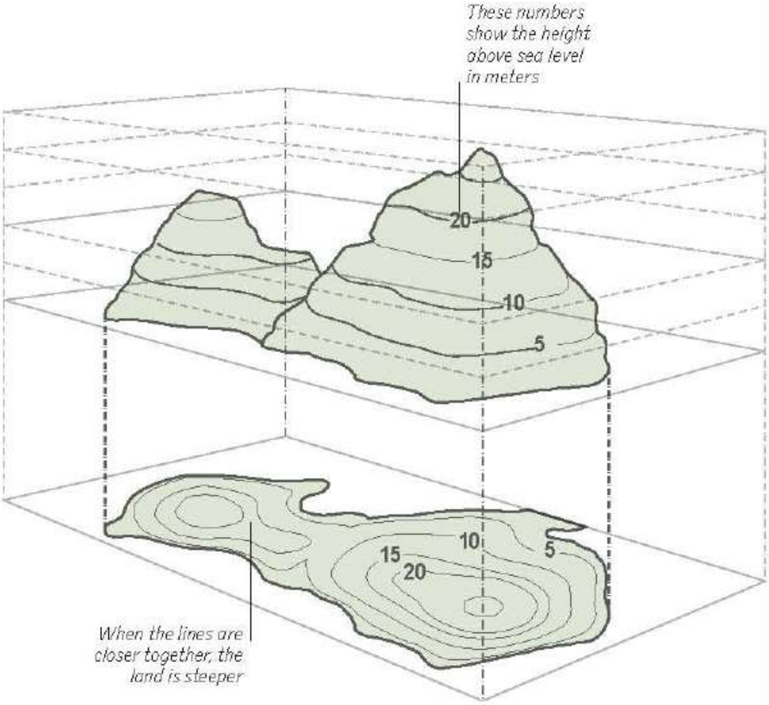

These numbers show the height above sea level in meters

When the lines are closer together, the land is steeper

GRID REFERENCES

Maps contain grid lines, which help you locate a specific point anywhere on a map, using a unique number known as a grid reference. The vertical grid lines are called "eastings," as they increase in value as they travel east on the map. The horizontal lines are called "northings," where these perpendicular lines cross creates the grid.

WORKING OUT GRID REFERENCES

Use the numbers on the grid lines and apply the easting number first. On a scale 1:25,000 map, where the grid lines are 1 km apart, the shaded area on the grid below would have a four-figure grid reference of 2046, and indicate a grid square that is 1 km by 1 km. To be more accurate, use a six-figure number—mentally divide the square into tenths. The cross would therefore have a grid reference of 185445.

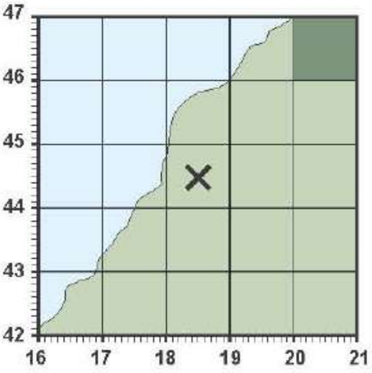

GET YOUR BEARINGS

IN ADDITION TO understanding and reading maps, it's important to know how to orientate your map to the land so you can use it to navigate. It's possible to do this by sight, although in most cases you'll need a more reliable method and for this you should use a compass. Use your compass to determine direction, orientate your map and yourself, take and plot bearings, and navigate.

CARDINAL POINTS

The four cardinal points of the compass are:

- North (N)—0°/360°
- East (E)—90°
- South (S)—180°
- West (W)—270°

HOW A COMPASS WORKS

A compass needle is a magnetized piece of metal that, when allowed to rotate freely, will orientate itself to the North and South magnetic poles. Always hold a compass level.

STANDARD COMPASS

This standard orienteering Silva compass is a good compass for hiking. It allows you to set your map and work out distances using the printed scale on the base.

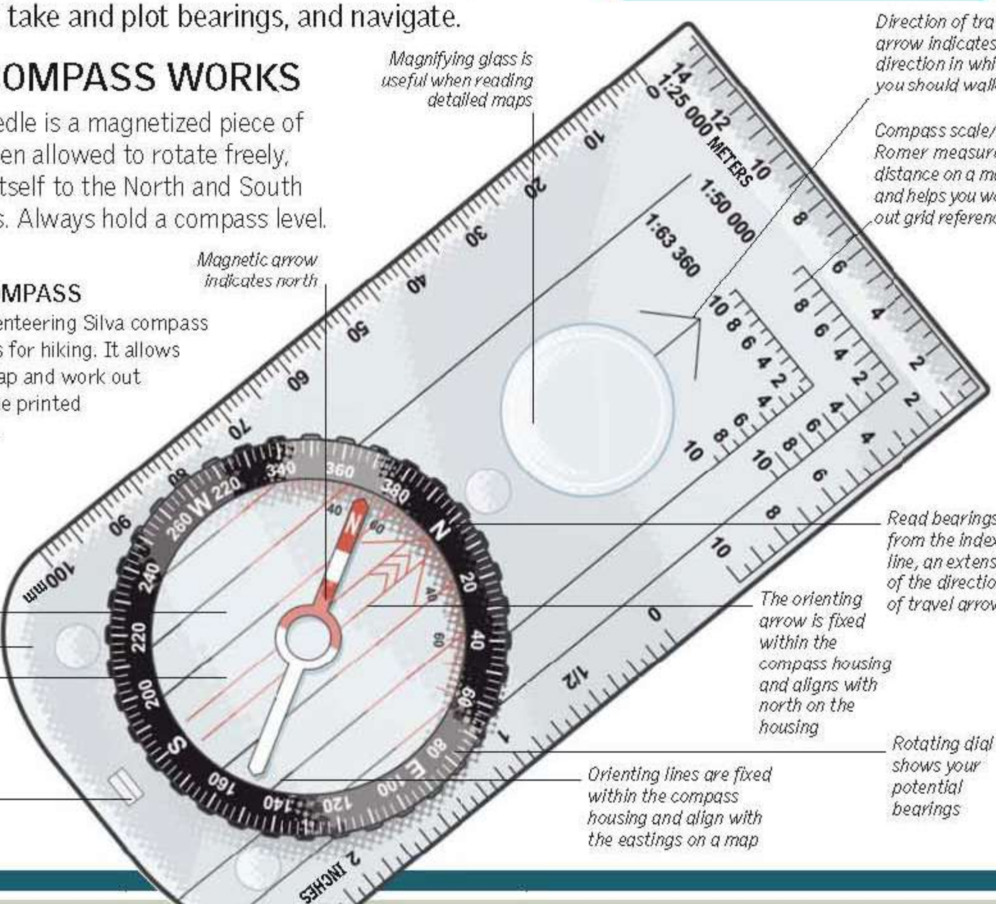

Magnifying glass is useful when reading detailed maps

Direction of travel arrow indicates the direction in which you should walk

Compass scale/ Romer measures distance on a map and helps you work out grid references

Magnetic arrow indicates north

Compass housing contains the needle and has the points of the compass marked on a circular, rotating bezel

Base plate

Liquid inside the housing dampens the needle to allow it to settle more quickly

Hole for attaching a lanyard

Read bearings from the index line, an extension of the direction of travel arrow

The orienting arrow is fixed within the compass housing and aligns with north on the housing

Rotating dial shows your potential bearings

Orienting lines are fixed within the compass housing and align with the eastings on a map

OTHER TYPES OF COMPASS

There are many different types of compass, ranging from simple button compasses to complex instruments that include features such as sighting mirrors. Use a quality compass that suits your needs, and carry a simpler backup for emergencies.

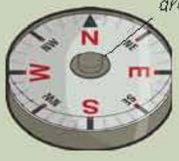

Only the points of the compass are shown

BUTTON

The simplest compass available—its small size and basic level of detail makes it ideal as a backup compass.

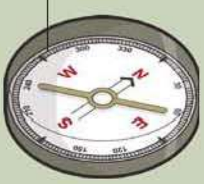

Non-rotating dial

FIXED DIAL

As well as the basic points of the compass, degrees and bearings are also shown on the immovable dial.

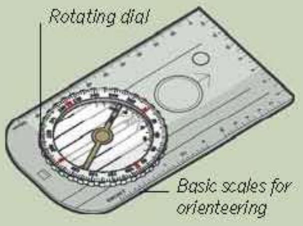

Rotating dial

Basic scales for orienteering

BASIC ORIENTEERING

With a rotating dial, this has simpler markings than the standard compass above, and is ideal for beginners.

SETTING YOUR MAP

Walking with your map set allows you to read the ground from the map as you pass over it, and to recognize and predict features as you progress, which means that you'll soon realize if you are heading off course. In some cases, when you have a good view of the terrain around you, and know the approximate area you're in, you can simply rotate your map until the features line up with the corresponding features on the ground. However, using your compass is a far more accurate method.

1 Rotate the dial so that "N" sits under the index line. Lay your map flat on the ground and ensure there is nothing nearby that could affect your compass (see box below).
- Lay the compass on the map so its edge runs parallel with a vertical grid line (easting).

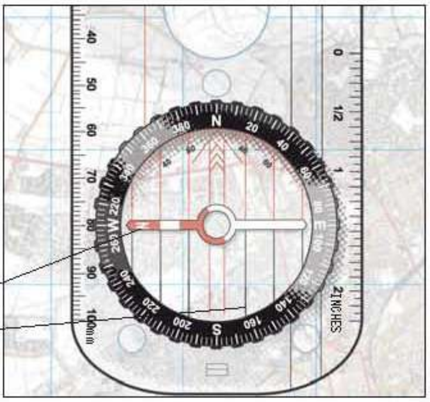

Needle will not be aligned at present

Orienting lines run parallel to the vertical grid lines on your map

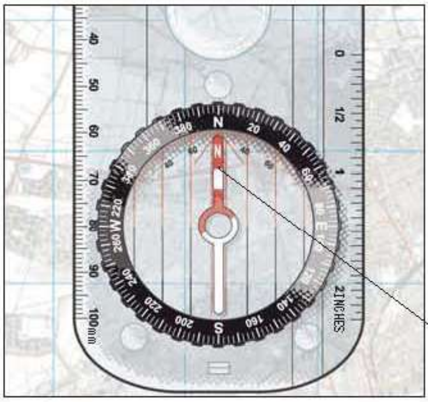

2 Keeping orienting lines aligned with the grid lines on the map, rotate the entire map until the north magnetic needle on the compass sits inside the orienting arrow. The map is now set to magnetic north and should basically line up with the features that surround you.
- If magnetic variation in your area is high (+ 5°), compensate accordingly (see right).

Needle is now aligned

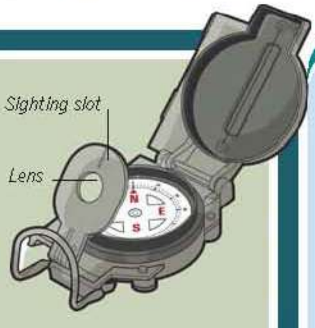

LENSATIC
The lensatic compass is an excellent instrument for taking very precise navigational measurements.

TAKING AN ACCURATE READING

Always hold the compass level to allow the needle to rotate easily. A compass is simply a magnetized piece of metal and, as such, is susceptible to interference. To avoid this, never use your compass near:
- Metal or other magnetized objects
- Electric currents, such as overhead high-tension power cables
- Buildings and vehicles—these often contain metal and electricity that might affect the accuracy of your compass readings

MAGNETIC VARIATION

Map legends refer to north in three ways: "true north," "grid north," and "magnetic north." The angle between magnetic north and grid north is known as "magnetic variation," and is provided in the map legend. True north is the direction of a meridian of longitude that converges on the North Pole. Grid north runs parallel to the vertical grid lines on a map and differs from true north because a map is flat. Magnetic north is the direction indicated by a magnetic compass.

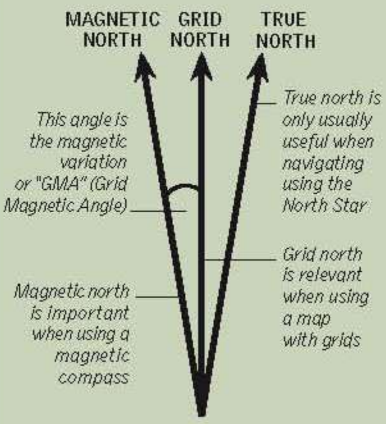

This angle is the magnetic variation or "GMA" (Grid Magnetic Angle)

Magnetic north is important when using a magnetic compass

True north is only usually useful when navigating using the North Star

Grid north is relevant when using a map with grids

COMPENSATING FOR VARIATION

When converting a magnetic bearing to a grid bearing, or vice versa, you have to adjust for magnetic variation. When the variation is west, use the phrases "Mag to grid—get rid" or "Grid to mag—add." If the variation is east, the opposite applies.

1 Check the legend of your map to establish the magnetic variation. This depends on your location, and whether it is east or west of grid north.

2 If the variation is 0°, there's no magnetic variation affecting the compass, so make no adjustments.

3 If converting a magnetic bearing to a grid bearing with a 12° west variation, take off the 12°. With an east variation, add it on.

4 If converting a grid bearing to a magnetic bearing with a 12° west variation, add on the 12°. With an east variation, take it off.

TAKING BEARINGS

Always give any compass work your full attention. Rushing it, especially when working out bearings, can lead to navigational errors that could, at best, involve more walking and, at worst, get you lost.

USING A MAP

Using your map to work out the direction in which you need to walk is simple. Use your compass as a protractor to work out your bearing, and then to keep you on track.

BACK BEARING

A back bearing enables you to find your position by taking a bearing from a feature back to you. To do this, take a bearing to a point in the normal way and either add or subtract 180 degrees. You can also read the bearing exactly opposite to the index line. This is useful when working out the bearing from a feature back to you and transferring it to your map (see below).

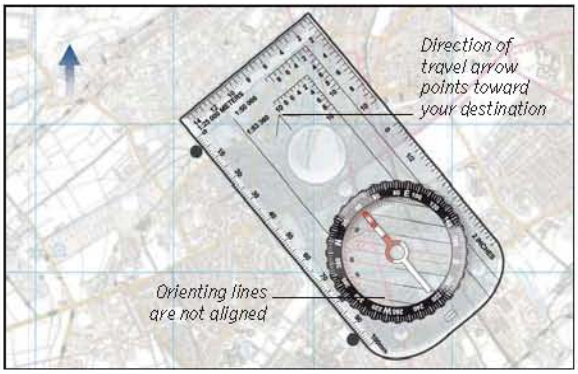
Direction of travel arrow points toward your destination
Orienting lines are not aligned

① Lay your map on a flat surface, ensuring that there's nothing nearby to interfere with the compass reading (see box, p. 69).

- Lay the edge of your compass so that it runs between the point you want to navigate from and the point you want to navigate toward.
- Ensure that the direction of travel arrow on the compass is pointing in the direction you want to travel on the map.

USING FEATURES ON THE GROUND

Sometimes you may need to take a bearing to a specific point to navigate toward it. The point may be a feature that you can see at the time but may subsequently lose sight of during your journey because of the terrain. You can also plot that bearing and others on a map in order to work out your own position (see right, and panel, far right).

TAKING A BEARING ON A FEATURE

Pointing your compass at the feature, hold the base plate steady and rotate the compass housing until the orienting arrow sits directly under the north needle. Read the bearing—this is the magnetic bearing you would walk on to navigate to the feature.

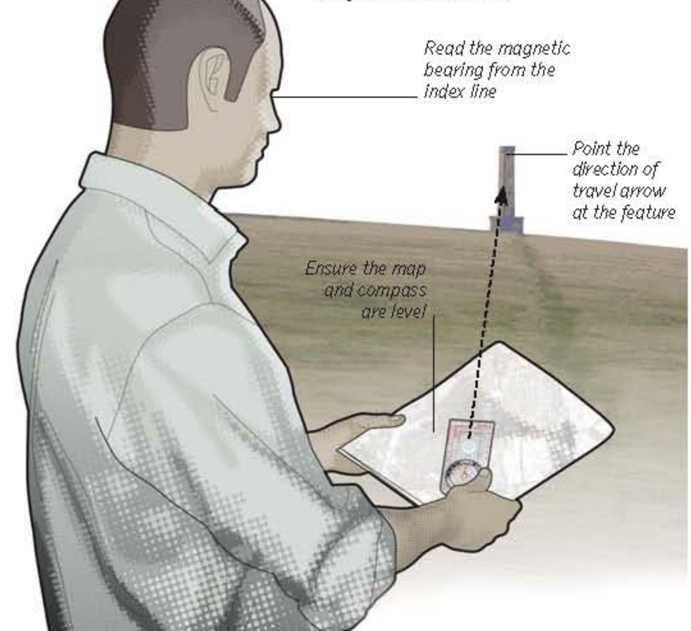
Read the magnetic bearing from the index line
Point the direction of travel arrow at the feature
Ensure the map and compass are level

TRANSFERRING A MAGNETIC BEARING TO A GRID BEARING

It is important that you know how to transfer a compass bearing from a feature (a magnetic bearing) onto a map (a grid bearing). For this example, imagine that the magnetic variation is 12° west.

- Take a bearing on your chosen feature (see left). In this example it is 45° magnetic.
- You want to plot this bearing on your map, which has a grid, so remember the phrase "mag to grid—get rid." You would therefore subtract the magnetic variation (12°) from the magnetic bearing (45°), which equals 33°.
- Dial this revised bearing into your compass.

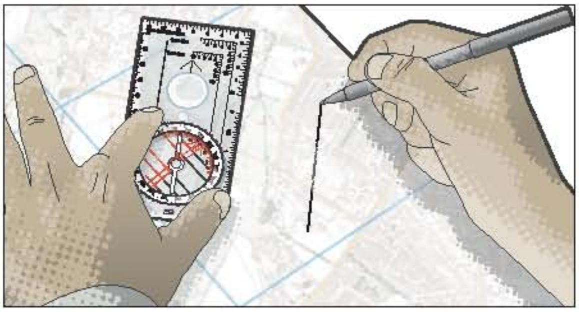

TRANSFERRING THE BEARING ONTO THE MAP

Place the top left corner of the compass base plate over the feature on your map. Keeping it there, rotate the entire compass until the orienting lines are parallel with the vertical grid lines. Draw a line from the feature down the left side of the base plate to map your bearing.

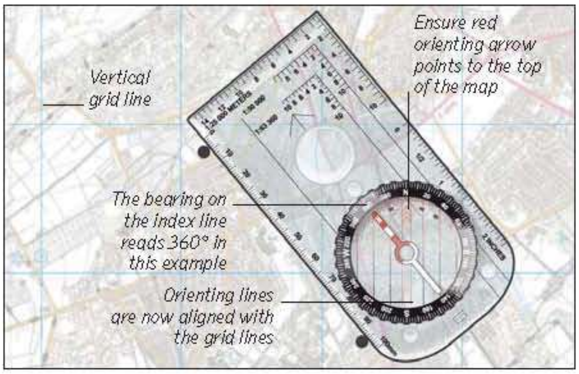
Vertical grid line
Ensure red orienting arrow points to the top of the map
The bearing on the index line reads 360° in this example
Orienting lines are now aligned with the grid lines

2 Turn the compass housing around until the orienting arrow and the orienting lines within the housing line up with the vertical grid lines on the map. Read the bearing between the two points from the index line on the compass.

- For this grid bearing to deliver you successfully to your destination, you'll need to convert it to a magnetic bearing by using the magnetic variation information on the map legend (see p. 69). Add or subtract your figure, and adjust your compass accordingly.

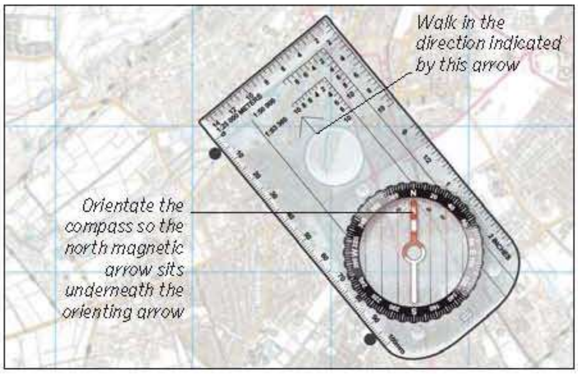
Walk in the direction indicated by this arrow
Orientate the compass so the north magnetic arrow sits underneath the orienting arrow

3 In order to walk on this bearing you must now orientate your compass.

- Hold the compass level and at a height that allows you to comfortably look squarely down on it (close to your chest is a good position).
- Turn your body until the north end of the compass needle sits inside the orienting arrow. The direction of travel arrow is now pointing exactly in the direction that you need to walk.

FINDING YOUR POSITION

If you are unsure of your position but can see features on the ground that you also recognize marked on your map, you can take bearings on these features with your compass and transfer them onto your map (see far left, and left) to accurately determine where you are. This process is called "triangulation" (also known as "resection" in military terminology). You'll need to transfer your compass bearings into grid bearings.

AT A KNOWN FEATURE

If you are at or on a known feature on a map—such as a river, road, or track—and you can see another recognizable feature, you can take a bearing on that feature and mark it on your map. Where that line crosses your known feature is your position. In the example below, you know you are somewhere along the banks of a river and can see a church that you can identify on your map.

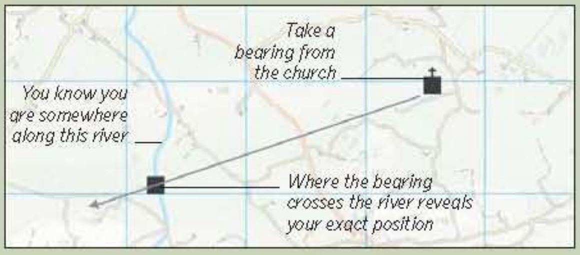
Take a bearing from the church
You know you are somewhere along this river
Where the bearing crosses the river reveals your exact position

NOT AT A KNOWN FEATURE

If you are not at a known feature, but can see other recognizable features on the ground that you can identify and locate on your map, you can take bearings on these features and transfer them to your map. In order to do this you need two features that are at least 1 km (6/10 mile) away and at least 40° apart. After you have transferred both bearings onto your map, the point at which the two lines cross will reveal your exact location. For greater accuracy, plot a third feature onto your map. The lines will cross and form a triangle—your position will be inside this triangle.

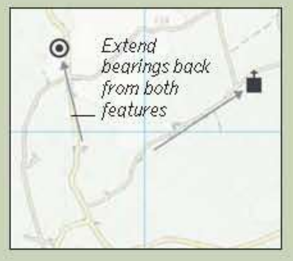
Extend bearings back from both features

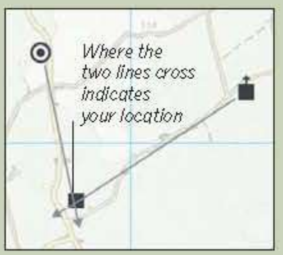
Where the two lines cross indicates your location

1 Using your compass, take bearings to the features on the ground.

- Transfer these bearings to back bearings (see box, above left), and draw them on the map from your chosen features.

2 Extend these lines further over the map until they cross each other.

- The point at which the lines cross shows your location.
- If you want to be more accurate, repeat the process with a third feature.

ROUTE FINDING

WHETHER YOU ARE PLANNING to walk a specific route, or are in a survival situation and need to move to a safer area or to one that offers better chances of rescue, being able to study your map and calculate what you can achieve in a certain time can mean the difference between reaching your destination or spending a night in the wilderness.

ROUTE PLANNING

Break your route down into small sections. This will focus your navigation and make the overall distance seem less daunting. If you're in a group, designate an emergency rendezvous (ERV) point within each section—if someone gets separated you should all head there. If you can, include water sources and a safety point, such as a campsite, on your route.

CALCULATING DISTANCE

There are several methods of calculating the distance you are walking, and a seasoned hiker will always use at least two of them at any one time.

PACE COUNTING

Pace counting involves knowing how many paces you take to cover a set distance and then counting them as you travel. It is reasonably accurate. Distances are usually calculated in meters and kilometers, and most people take approximately 60 paces (120 steps) every 100 m. Try any of these methods:

- Cut 10 notches in your walking staff. Move an elastic band down a notch every 100 m.
- Use pace counting beads (a piece of webbing with two sets of beads divided by a central knot). One side contains nine beads to count off every 100 m, and the other has four beads to count off kilometers.
- Place 10 small pebbles in your pocket. Every time you cover 100 m, transfer a pebble to the other pocket.

NAISMITH'S RULE

Naismaith's rule takes into account distance and topography, and is used for estimating the duration of hikes.

- Allow one hour for every 3 miles (5 km) you will travel
- Add 30 minutes for every 985 ft (300 m) you will climb
- Subtract 10 minutes for every 985 ft (300 m) you will descend. However, for very steep slopes you should add 10 minutes for every 985 ft (300 m) you will descend

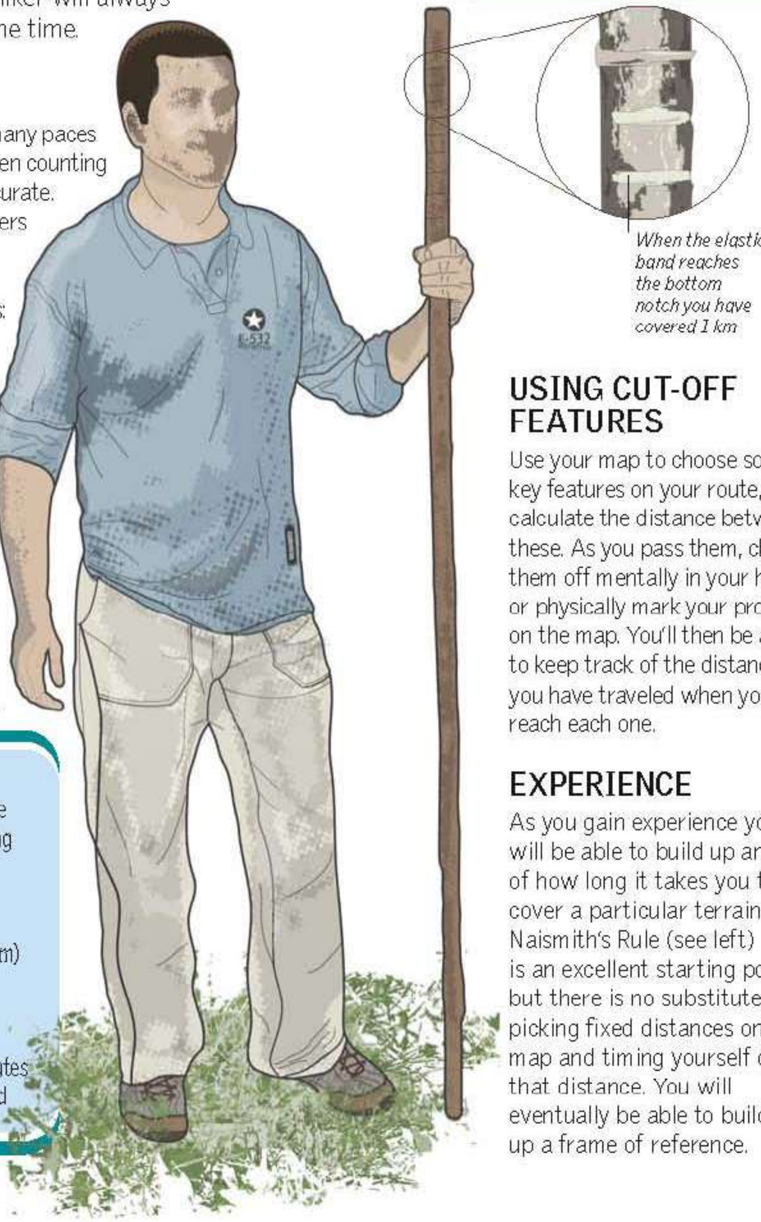

When the elastic band reaches the bottom notch you have covered 1 km

USING CUT-OFF FEATURES

Use your map to choose some key features on your route, and calculate the distance between these. As you pass them, check them off mentally in your head or physically mark your progress on the map. You'll then be able to keep track of the distance you have traveled when you reach each one.

EXPERIENCE

As you gain experience you will be able to build up an idea of how long it takes you to cover a particular terrain. Naismith's Rule (see left) is an excellent starting point, but there is no substitute for picking fixed distances on your map and timing yourself over that distance. You will eventually be able to build up a frame of reference.

NAVIGATION TECHNIQUES

When navigating across land, you're less likely to get lost if you take direct bearings from one feature to another. Unfortunately this may not always be possible; obstacles such as lakes and swamps may be directly in your path and it may be easier to walk around some features rather than walk over or through them.

PRINCIPLES OF NAVIGATION

You will never get lost if you know:

- Where you started from
- What bearing or course you have been traveling on
- How far you have traveled

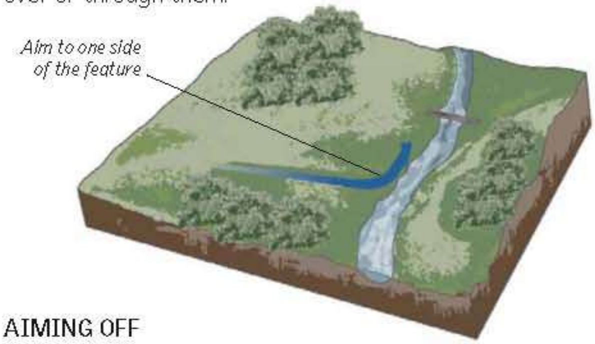
Aim to one side of the feature

AIMING OFF

It's easy to find yourself slightly off-course after a while. If you were aiming for a small footbridge over a stream and didn't arrive exactly at the bridge, you would need to guess whether to turn left or right to reach it. By deliberately aiming off to one side (also called "deliberate off-set") you can guarantee this direction.

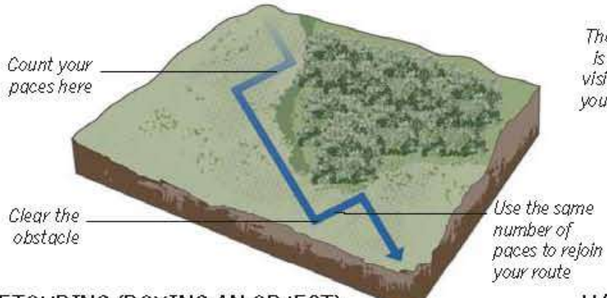
Count your paces here
Clear the obstacle
Use the same number of paces to rejoin your route

DETOURING (BOXING AN OBJECT)

If an obstacle makes a straight-line bearing impossible, use your compass to calculate four 90° turns, which you will then need to walk on to pass the obstacle. Count your paces on the first and third detours to return to your original, intended route as accurately as possible.

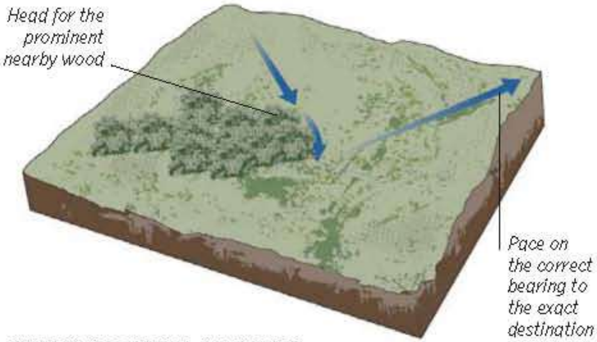
Head for the prominent nearby wood
Pace on the correct bearing to the exact destination

STAND OFF/ATTACK POINT

Useful when navigating to a specific point that may be difficult to locate, this technique involves aiming initially for a nearby prominent feature. Calculate a distance and bearing from it and use pacing to accurately locate the specific point. If you fail to find it, simply return to the prominent feature and start again.

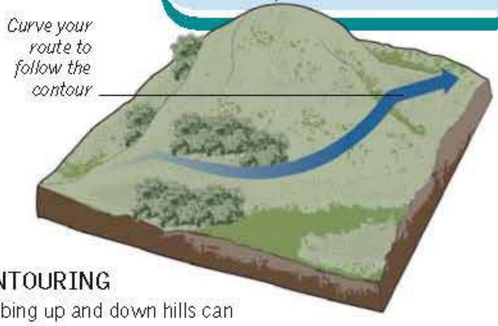
Curve your route to follow the contour

CONTOURING

Climbing up and down hills can expend a lot of energy and may not be the most effective method of navigating a particular terrain. Instead, use a technique called "contour navigation." This involves walking at the same height around a feature, which will conserve energy.

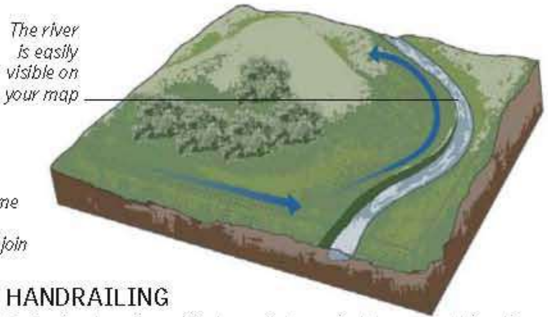
The river is easily visible on your map

HANDRAILING

Following long linear features that run in the general direction of your travel—such as rivers, roads, or paths—can be an effective way of navigating. Because you use the features rather than your compass, navigation becomes simpler, as the features are usually easy to follow. A slight detour toward one can be worthwhile.

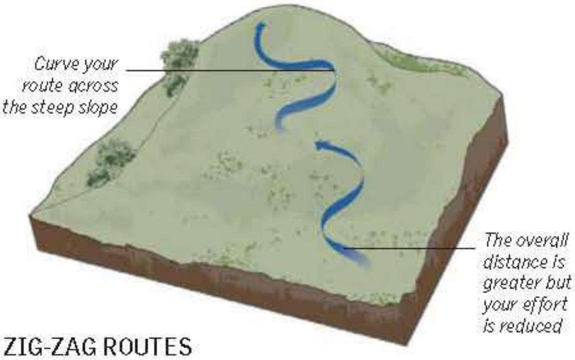
Curve your route across the steep slope
The overall distance is greater but your effort is reduced

ZIG-ZAG ROUTES

You may have no alternative but to climb a steep slope and this can be exhausting. However, if you choose a zig-zag route up the slope you can reduce the effort required to achieve the climb. This will dramatically reduce the strain on your legs, ankles, lungs, and heart. It is also effective when walking down steep slopes.

NAVIGATING WITHOUT A COMPASS

IF YOU DON'T HAVE A COMPASS you can use a few simple items from your survival kit to determine direction and enable you to navigate reasonably accurately. Advances in technology have also made satellite-navigation aids more readily available.

TOOLS AND MATERIALS

You will need an improvised needle or razor blade, plus one of the following:

- Knife
- Hair
- Silk
- Magnet
- Battery and wire; paper and tape

IMPROVISING A COMPASS

It's relatively simple to make an improvised compass using a piece of magnetized ferrous metal. How accurate your compass is depends on the materials you have available to you and your own ingenuity.

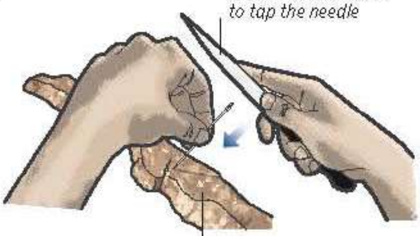

You can use the blade of your survival knife to tap the needle

Tap the needle into hard wood to heighten the effect, but make sure it doesn't become embedded

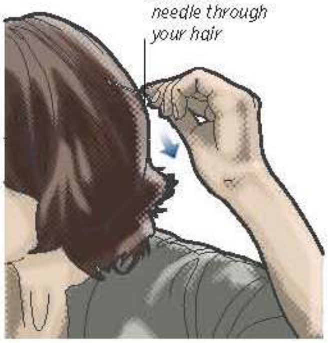

Stroke the needle through your hair

SOURCING AND MAGNETIZING THE NEEDLE

If you have a compass but it's damaged, you may still be able to use the needle from it, which will already be magnetized. If this is unusable, you'll need to find a piece of ferrous metal to magnetize. Suitable items you could use include:

- A needle from your survival tin or sewing kit
- A paperclip, opened up and straightened out
- A razor blade from your survival kit
- A small nail or straightened staple taken from a fence

Once you have your improvised needle, you will need to magnetize it. The smaller and thinner the needle is, the easier this will be. Use one of the methods shown here.

THE TAPPING METHOD

Align the needle as close to the north-south line as you can determine, hold the needle at an angle of 45 degrees, and gently tap the end of it with another piece of metal. Lightly tapping it into a hard piece of wood will increase the effect.

USING YOUR HAIR

Hold the sharp end of the needle perpendicular to your head and—taking care not to hurt yourself—stroke the needle in one direction through your hair, using careful and deliberate strokes. Repeat until the needle is magnetized.

ALLOWING THE NEEDLE TO FLOAT FREELY

Once you have made your improvised needle using one of the methods above, you need to find a way to allow it to turn freely so that it is able to indicate direction. Take care to protect your needle from elements such as the wind, which will affect its movement.

SUSPENSION METHOD

The advantage of the suspension method is that the equipment is portable and can be reused. It works best with a magnetized razor blade, which will balance well. Attach the magnetized blade to a cotton thread and suspend it inside a plastic bottle. If the bottle's neck is not wide enough to fit the blade through, remove the base of the bottle instead.

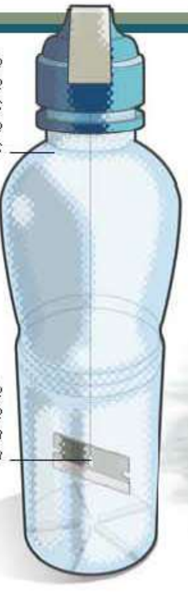

The bottle protects the compass from the elements

Use the hole in the blade to attach the cotton

FLOATING METHOD

In a sheltered place, float the needle on the surface of some water—a puddle or a small non-magnetic container filled with water, for example. Balance the needle on a small, dry leaf (or piece of paper, piece of bark, blade of grass, or inside a shortened straw).

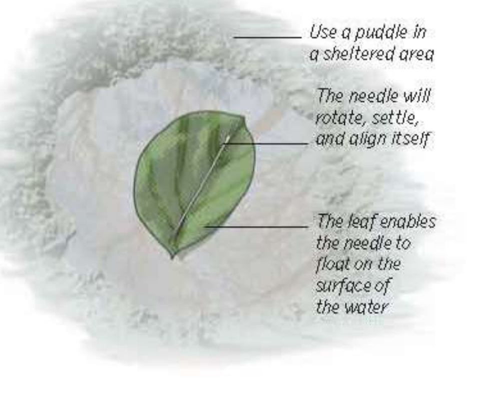

Use a puddle in a sheltered area

The needle will rotate, settle, and align itself

The leaf enables the needle to float on the surface of the water

PRINCIPLES OF MAGNETIZING

In general, the longer you work on the needle the stronger the magnetization will be and the longer it will last. To tell when it is magnetized sufficiently, hold the needle up against another metal object—if it is attracted to the metal and has enough strength to hold itself against it then it will be strong enough. Once the needle is magnetized you will need to allow it to float freely (see below), and then determine which end points north by using natural aids, such as the sun (see pp. 76–77). Mark the north end of the needle with a pen or a small scratch.

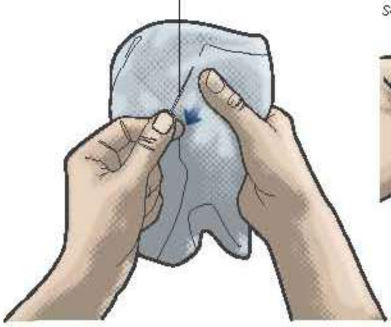

Stroke the needle 50–100 times

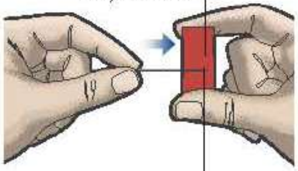

The end of the magnetized needle that attracts to the south pole of the magnet will point north

The more times you stroke, the more effective the magnetization

USING SILK

This works on a similar principle to the hair method but is more effective. If you have anything made from silk, such as a sleeping bag liner or thermal clothing, stroke the needle repeatedly in one direction against the silk.

USING A MAGNET

Stroke the magnet along the length of the needle repeatedly in one direction. It's a good idea to carry a magnet with you at all times, although you should never keep it near your compass as it will affect its accuracy.

USING ELECTRICITY

The most effective way to magnetize a needle is to pass a small electrical current around it. Use a battery and insulated wire; alternatively, use brass snare wire from your survival kit and insulate it using something non-conductive, such as paper.

1 Wrap the needle in a small piece of paper, which will insulate it from the electrical current.

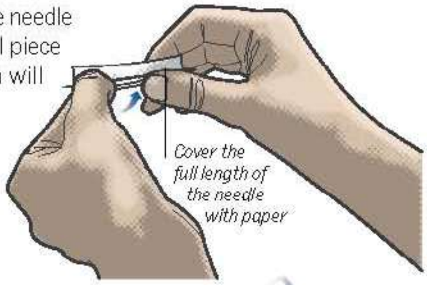

Cover the full length of the needle with paper

2 Wrap the wire tightly around the full length of the insulated needle.

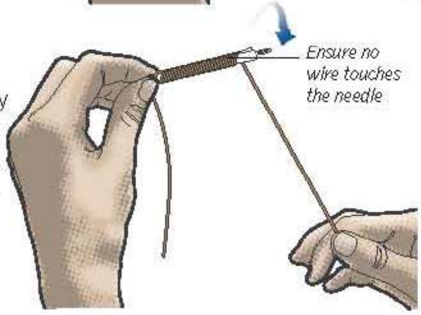

Ensure no wire touches the needle

3 Attach the wire to the battery until battery starts to get warm—this indicates the process is complete.

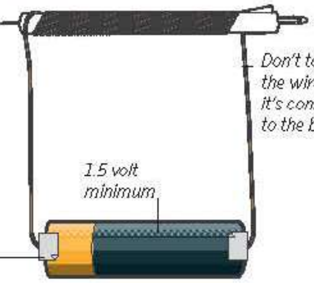

Don't touch the wire once it's connected to the battery

1.5 volt minimum

Attach the wire using some tape

USING GPS TECHNOLOGY

A Global Positioning System (GPS) is a hand-held unit that uses 24 orbiting satellites to triangulate your position on the Earth's surface to within meters. A GPS will allow you to work out straight-line distances and bearings to and from points, but unless it incorporates mapping it will not show you the best way to get there or take into account hazards on your route. Use it in conjunction with your map and compass.

TAKING EFFECTIVE GPS READINGS

A GPS needs to have a clear view of the sky. Anything that obstructs its signal, such as tall buildings or heavy tree canopy, will reduce its ability to lock onto satellites.

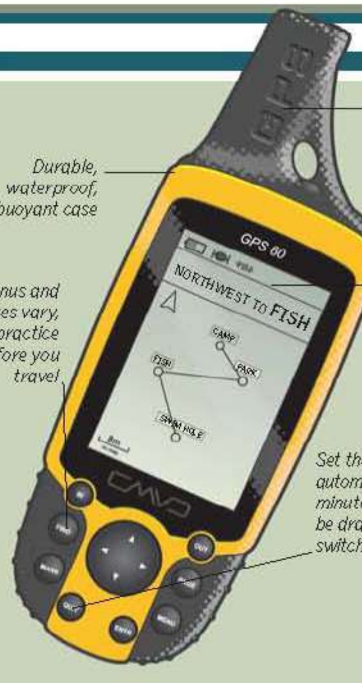

Durable, waterproof, buoyant case

Receiver locks onto at least four satellites to obtain a three-dimensional fix (latitude, longitude, and height) displayed as a grid reference

Menus and features vary, so practice before you travel

The screen lights up for use in gloomy conditions, but avoid using this feature if you can as it will drain the batteries

Set the unit to switch off automatically after two minutes so the battery won't be drained if the GPS is switched on accidentally

NATURAL NAVIGATION

IF YOU HAVE LOST or damaged your compass and don't have the materials necessary to make an improvised version, use natural indicators to determine direction. The east-west rotation of the Earth means that you can orientate yourself according to the position of the sun, moon, and stars. All you need are some very basic materials and these simple techniques.

USING THE SUN

When visible, the sun is the clearest natural signpost to the four cardinal points (north, south, east, and west). It rises approximately in the east and sets approximately in the west and, at midday, is due south in the northern hemisphere and due north in the southern hemisphere. Use the sun's course across the sky to determine direction and approximate the time.

SHADOW STICK BASICS

Use a shadow stick to determine direction and time anywhere between the Arctic (66.5°N) and Antarctic Circles (66.5°S).

- In the northern hemisphere the shadow of the stick will be on the north side of the east-west line
- In the southern hemisphere the reverse is true: the shadow of the stick will be on the south side of the east-west line.
- When the shadow is at its shortest, it is midday.

ORIENTATION

Tracking the movement of the sun across the sky using a shadow stick will provide an indication of its direction of travel. The sun moves from east to west at 15° an hour.

1 Drive a stick into a piece of level ground, and ensure it's as upright as possible.

2 Place a stone at the tip of the stick's shadow.

3 Wait three hours and place a second stone at the new position of the shadow's tip.

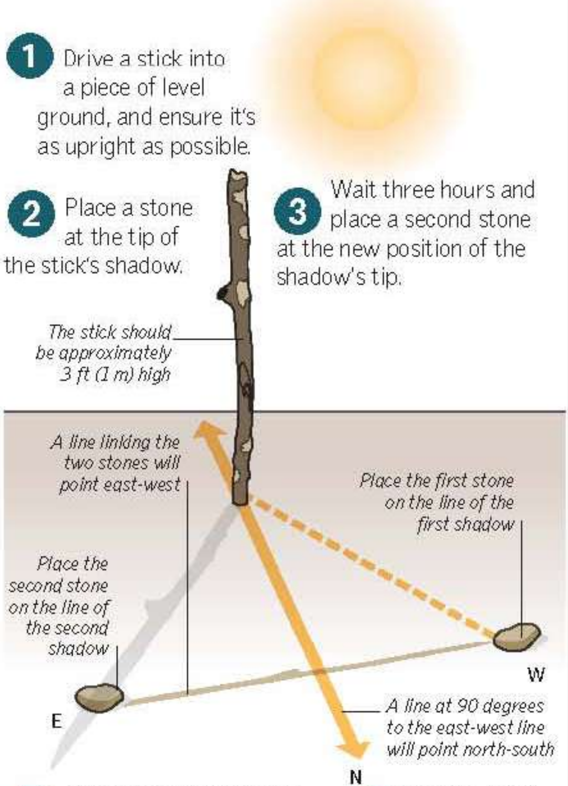

The stick should be approximately 3 ft (1 m) high

A line linking the two stones will point east-west

Place the first stone on the line of the first shadow

Place the second stone on the line of the second shadow

A line at 90 degrees to the east-west line will point north-south

4 Draw a line between the stones to find east-west; the first stone will be at due west and the second at due east.

5 To find north-south, mark a line at 90° to the east-west line.

TELLING THE TIME

Once you have established the east-west and north-south lines, you can turn the shadow-stick apparatus into a sundial, to give an approximate idea of the time.

1 Place the stick at the intersection of the north-south and east-west lines

2 Tie a piece of cord to the stick. Attach a smaller stick to the other end of the cord and use it to draw an 180-degree arc between the two marker stones.

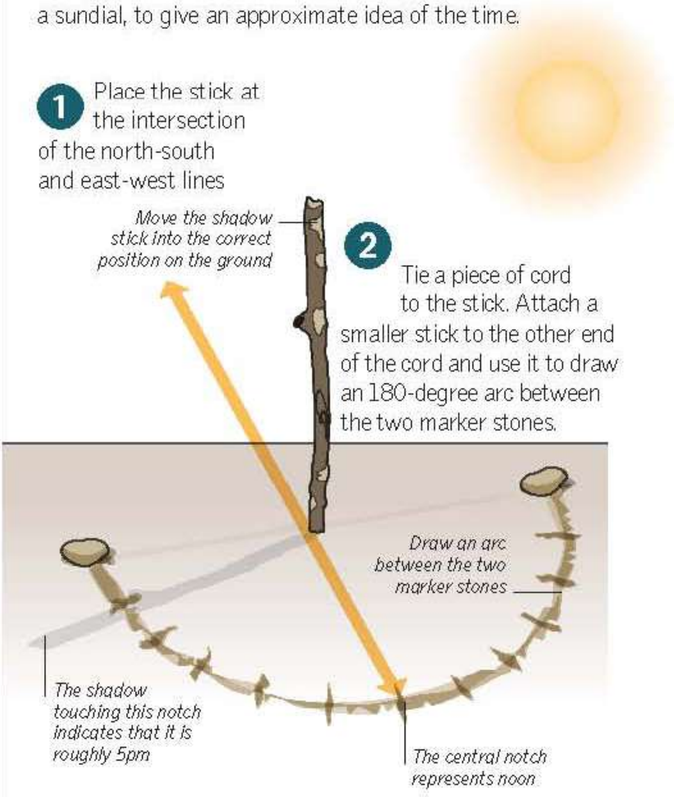

Move the shadow stick into the correct position on the ground

Draw an arc between the two marker stones

The shadow touching this notch indicates that it is roughly 5pm

The central notch represents noon

3 Divide the arc into 12 equal sections and mark each division with a notch. The notches represent one hour of time, from 6am to 6pm with noon at the middle notch.

USING THE STARS

It's possible to orientate yourself by spotting certain recognizable stars. In the northern hemisphere, Polaris is located above due north on the horizon, from which you can determine east, west, and south. Find it by locating the Plow (also known as the Big Dipper). In the southern hemisphere, you can work out the approximate position of south on the horizon by finding the Southern Cross.

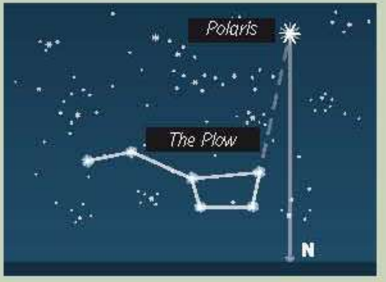

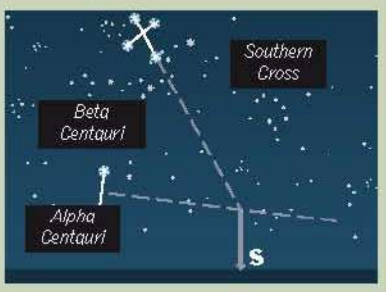

NORTHERN HEMISPHERE

Find the Plow and project a line from the top of the two stars that form its front. Follow this line until you find Polaris, which is located approximately four times the distance from the Plow as the distance between its two front stars.

SOUTHERN HEMISPHERE

Project a line from the longer axis of the Southern Cross until you find a dark area of sky. Project a second line at 90° from the mid-point between two bright stars in the Centaurus constellation. Due south is below the point where these lines meet.

USING THE MOON

Reflecting the light of the sun, the moon rises in the east and sets in the west, so can be used for orientation. A shadow stick will work on a cloudless, full-moon night.

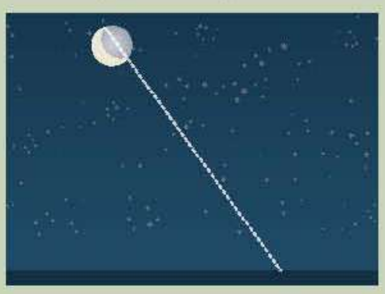

CRESCENT MOON

Although not an entirely accurate method, a line vectored between two horns of a crescent (quarter) moon will lead to a point that is approximately south on the horizon in the northern hemisphere, and roughly north in the southern hemisphere.

USING AN ANALOG WATCH

If you can see the sun, you can use an analog watch as a protractor to determine an approximate direction. Ensure it is set to the correct local time and that you have taken daylight savings (DST) into account. If you don't have a watch but know the time, simply draw a watch face on a piece of paper, marking 12 o'clock and the hour hand. This method is increasingly less effective as you near the Equator.

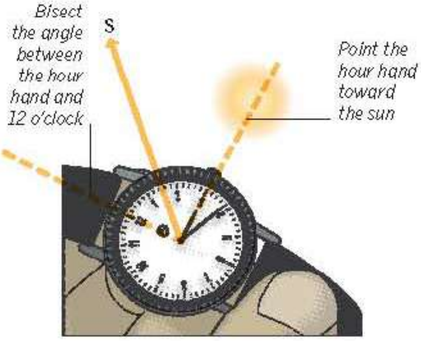
Bisect the angle between the hour hand and 12 o'clock
S
Point the hour hand toward the sun

NORTHERN HEMISPHERE

In the northern hemisphere the cardinal point nearest to the sun is south. Point the hour hand of your watch at the sun and bisect the angle between the hour hand and 12 o'clock. This will be due south.

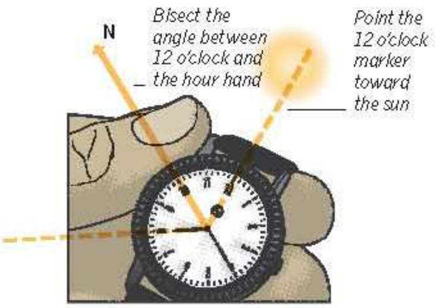
N
Bisect the angle between 12 o'clock and the hour hand
Point the 12 o'clock marker toward the sun

SOUTHERN HEMISPHERE

In the southern hemisphere, the cardinal point nearest to the sun is north. Point the 12 o'clock marker on your watch at the sun and bisect the angle between 12 o'clock and the hour hand on the watch. This will be due north.

NATURAL SIGNPOSTS

Nature responds to the elements in a range of ways, some of which can be studied for orientation tips. This is useful if you know the predominant wind direction of an area.

TREES AND PLANTS

- Windswept trees point away from the wind.
- Tree growth is most lush on the side that faces the sun (south in the northern hemisphere, north in the southern hemisphere).
- Some plants, such as the barrel cacti, twist toward the sun as they grow.
- Moss and lichen grow out of direct sunlight, on the shady side of rocks or trees.

ANIMALS AND INSECTS

- In very windy areas, small animals and birds tend to nest or burrow on the lee-side of hills.
- Spiders spin webs out of the wind, so an area with a lot of broken webs could indicate a recent change of wind direction, away from the dominant wind.

SNOW AND ICE

- In powder-snow conditions, snow "dunes" often form parallel to the prevailing wind.
- Frost erosion is most severe on slopes facing the sun (south in the northern hemisphere, north in the southern hemisphere).

HOW WEATHER WORKS

CHANGES IN THE WEATHER can have a significant effect on your expedition—or chances of survival if something goes wrong—so it's important to be as prepared as possible for all the conditions you might face. Check the weather report for the days ahead thoroughly, and pack and wear appropriate clothing. If the weather forecast is so bad that it could make traveling or navigation difficult, rethink your plans.

UNDERSTANDING THE WEATHER

Weather is created by the movement of air currents, the moisture content of the air, and the meeting of warm and cold fronts. Looking at weather maps before you leave for your trip will help you track these elements. You'll then need to use your understanding to forecast the effects of these movements on the ground.

READING WEATHER CHARTS

It can be helpful to compare weather charts with regular topographic land maps. Just as the gradient is steeper where contour lines appear closer together on land maps, the wind is stronger where isobar contours appear closer together on weather charts.

WHAT IS WEATHER?

The term "weather" relates to current conditions on the ground—such as the temperature, and whether or not it is windy or raining—while "climate" refers to a region's conditions over a longer period of time. Changes in the weather are caused by alterations in air pressure and temperature. Extreme weather, such as a hurricane, occurs when these changes are more marked than usual. Weather is monitored by meteorologists, and very accurate predictions can be made. However, weather can still cause surprises, so always prepare for all possible scenarios when packing or dressing for an expedition.

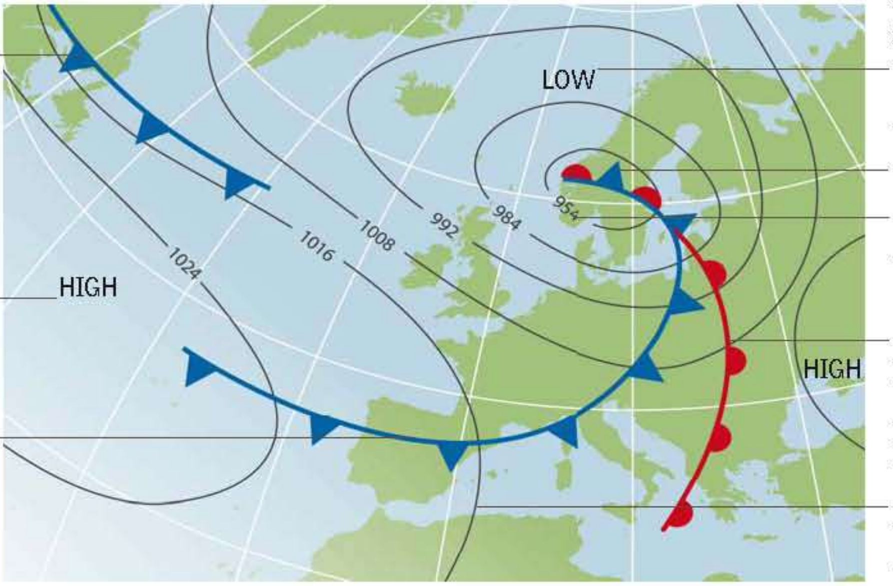

Fronts move in the direction that their symbols point toward

Areas of high pressure are generally associated with warm weather

A blue line with triangles represents a cold front, and marks the leading edge of cold air

Areas of low pressure are associated with cooler weather

Occluded fronts are fronts where two air masses merge

The lower the value of the central pressure, the more severe the wind and rain

A red line with semi-circles represents a warm front, and marks the leading edge of warm air

Isobars join areas where the air pressure is the same

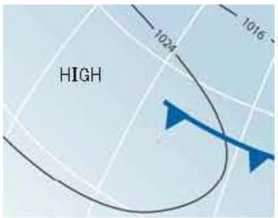

HIGH PRESSURE

Under conditions of high pressure, air descends in a spiral formation, and warms. The water vapor does not condense into clouds and we would expect to observe fairly clear skies.

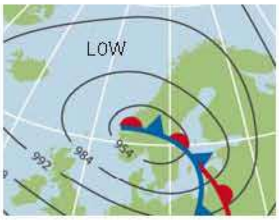

LOW PRESSURE

Under conditions of low pressure, air spirals upward—parallel to the center of the isobars—and cools. The air condenses to form clouds, and skies are usually overcast.

GLOBAL CIRCULATION

Warm air from the equator rises and moves toward the poles in each hemisphere. As it cools it descends and travels back toward the equator. Because of the Earth's rotation, three separate cells of circulating air in each hemisphere develop, which produce predictable wind and pressure patterns.

AIR MASSES

Large bodies of air with a particular temperature and humidity help weather forecasters predict the forthcoming weather and are known as air masses. They are largely defined by the area that they originated from. In general, northerly winds are colder than southerly winds, and air that has tracked over the sea ("maritime air mass") accumulates moisture and is more cloudy than air that has tracked over the land ("continental air mass"). The boundaries between tropical and polar air become warm and cold fronts.

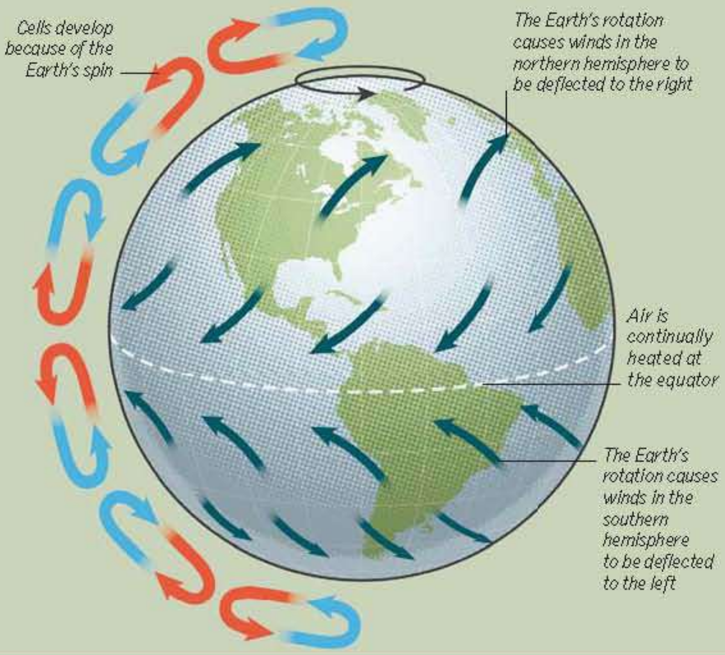
Cells develop because of the Earth's spin
The Earth's rotation causes winds in the northern hemisphere to be deflected to the right
Air is continually heated at the equator
The Earth's rotation causes winds in the southern hemisphere to be deflected to the left

ON THE GROUND

As a weather system travels over the ground, it changes in predictable ways. Knowing how the system will progress is an important part of interpreting your weather map or forecast.

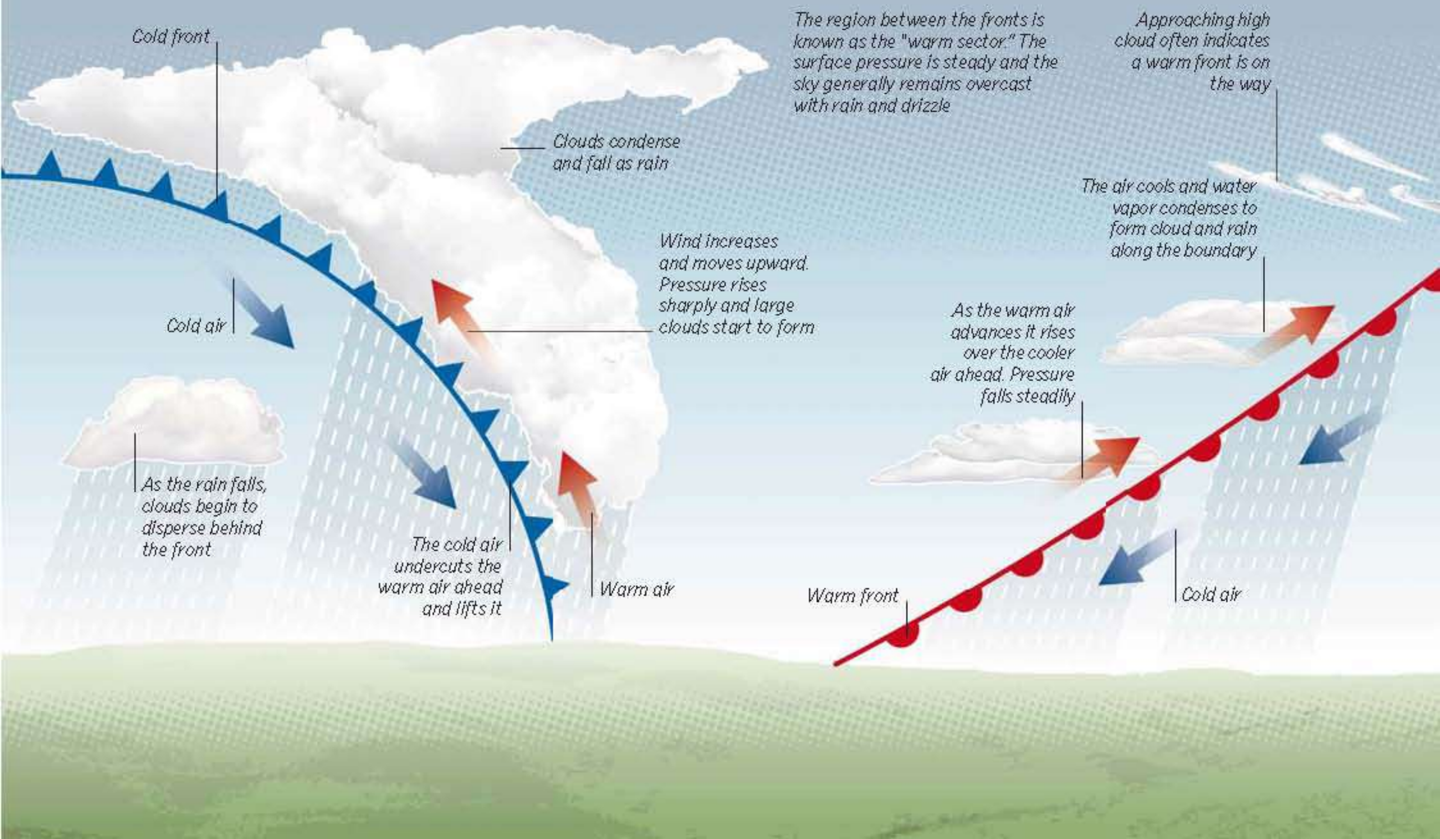
Cold front
Clouds condense and fall as rain
The region between the fronts is known as the "warm sector." The surface pressure is steady and the sky generally remains overcast with rain and drizzle
Approaching high cloud often indicates a warm front is on the way
The air cools and water vapor condenses to form cloud and rain along the boundary
Wind increases and moves upward. Pressure rises sharply and large clouds start to form
Cold air
As the warm air advances it rises over the cooler air ahead. Pressure falls steadily
As the rain falls, clouds begin to disperse behind the front
The cold air undercuts the warm air ahead and lifts it
Warm air
Warm front
Cold air

WEATHER PHENOMENA

LOOKING AT THE CLOUDS can help you to read approaching weather, which is invaluable when you're on the move with no access to weather forecasts or charts. Knowing how to recognize a storm cloud, for example, will ensure that you have enough warning to seek shelter or to change into appropriate clothing. If low cloud threatens to impair your visibility, use your compass to navigate and proceed with caution.

READING THE CLOUDS

Clouds are condensing masses of water vapor that provide precipitation and reflect solar radiation. They are categorized by height—into low, medium, and high clouds—and further defined by their shape. The shape of a cloud is determined by the way in which warm air rises, and is an indicator of the air stability.

HOW CLOUDS FORM

Clouds form by a cooling process. This process can be compared to breathing on cold glass, when condensation appears because invisible water vapor in your breath, cooled to a temperature known as the "dew point," condenses into its visible liquid state. Clouds form in a very similar way. Air temperature falls at an average rate of two degrees celcius per 1,000 ft (300 m). When air containing moisture—perhaps because it has travelled over sea—reaches the height at which it meets its dew point temperature, the water vapor condenses to form visible cloud.

HIGH CLOUDS

Fair-weather clouds are white and high in the sky. If there are no clouds, expect excellent conditions. Blankets of high cloud progressively invading the sky can signify the onset of bad weather. High clouds include:

- Anvil of cumulonimbus
- Cirrus
- Cirrustratus

MEDIUM CLOUDS

Thick layers of medium-layer cloud give heavy, persistent rain, especially if the clouds are dark and gray. The clouds you are most likely to see at this level are:

- Altocumulus
- Nimbostratus
- Altostratus

LOW CLOUDS

Low clouds have clearly defined edges and can indicate whether rain will fall in short downpours (cumulus) or persistently (stratus). Common clouds are:

- Cumulonimbus
- Stratus
- Stratocumulus
- Cumulus

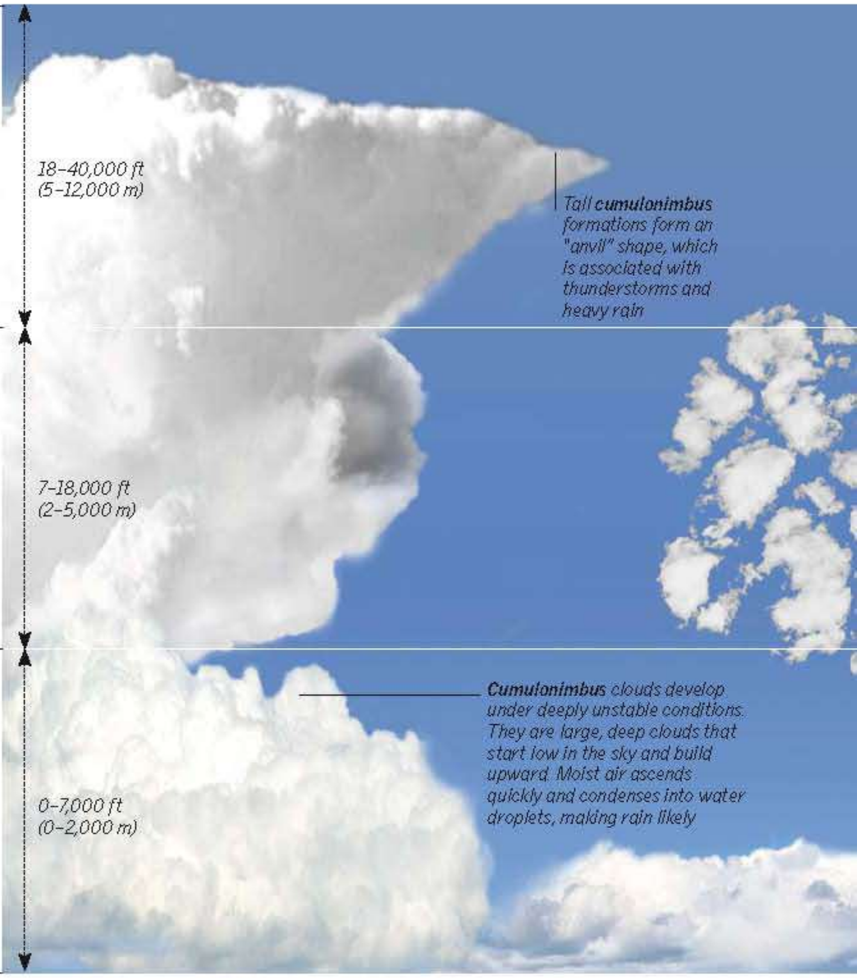

*Tall **cumulonimbus** formations form an "anvil" shape, which is associated with thunderstorms and heavy rain*

***Cumulonimbus** clouds develop under deeply unstable conditions. They are large, deep clouds that start low in the sky and build upward. Moist air ascends quickly and condenses into water droplets, making rain likely*

WEATHER RISKS

The weather can have a massive impact on your trip, bringing with it added dangers. It can affect your visibility levels, and the safety of the ground you're hiking over. In a survival situation, the temperature, and whether or not it's raining, can have a huge effect on your chances.

WARNING!

Never shelter under a solitary tree during a lightning storm—get in your vehicle if possible or find low ground.

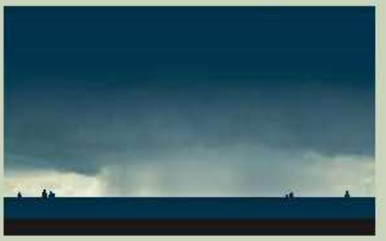

HEAVY RAIN

If heavy rain falls quickly, the ground may become saturated, slippery, or flood. Seek shelter or proceed cautiously in waterproof clothing.

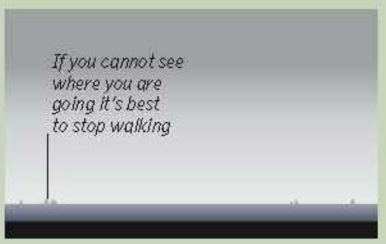

If you cannot see where you are going it's best to stop walking

FOG

Essentially a cloud in contact with the ground, fog reduces visibility, so beware of dangerous terrain, especially on mountains.

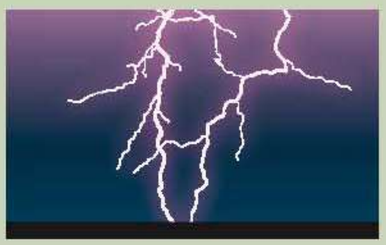

LIGHTNING

Lightning is an atmospheric discharge of electricity—it will strike the first object it encounters on its route to earth, so avoid high, exposed places.

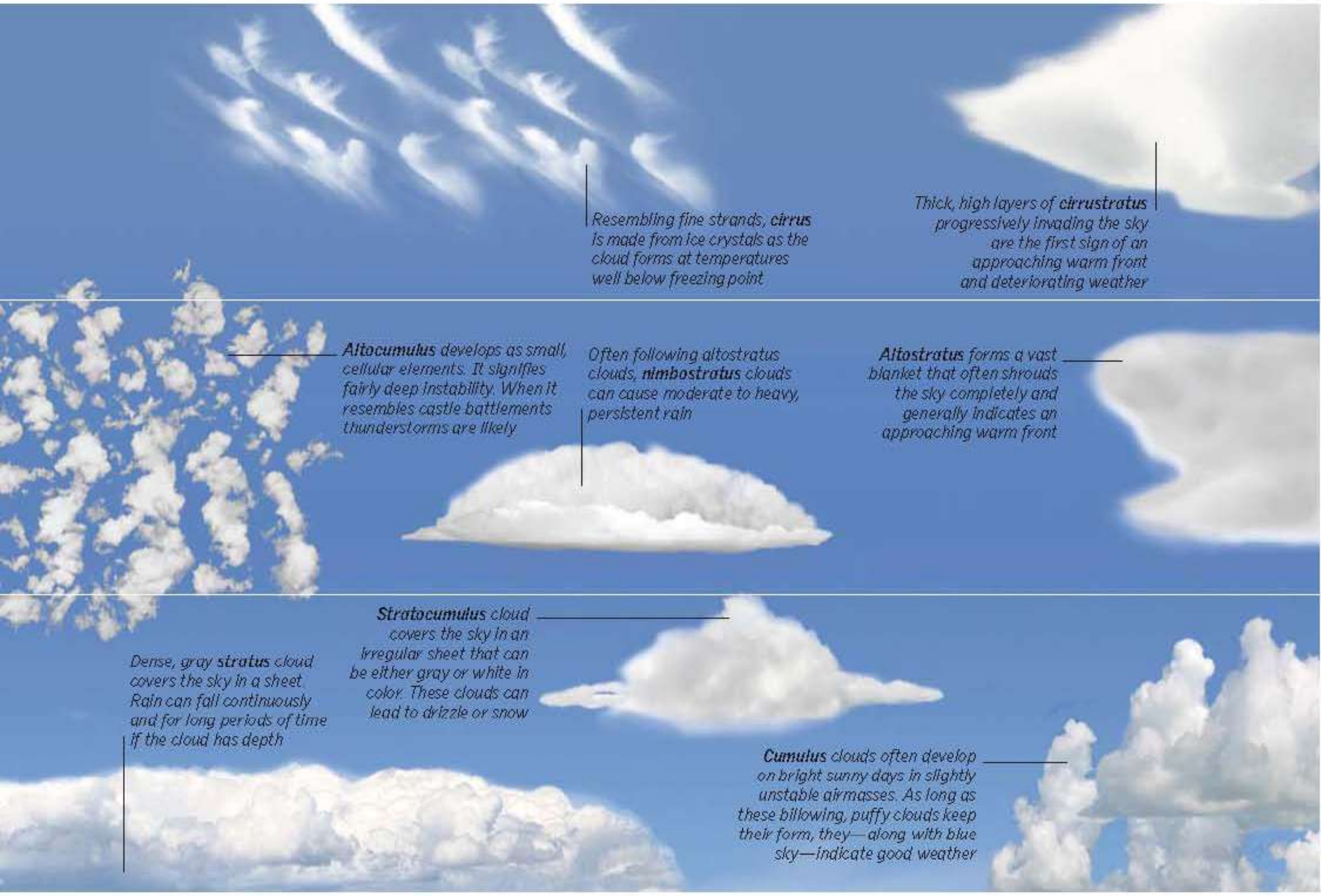

*Resembling fine strands, **cirrus** is made from ice crystals as the cloud forms at temperatures well below freezing point*

*Thick, high layers of **cirrustratus** progressively invading the sky are the first sign of an approaching warm front and deteriorating weather*

***Altocumulus** develops as small, cellular elements. It signifies fairly deep instability. When it resembles castle battlements thunderstorms are likely*

*Often following altostratus clouds, **nimbostratus** clouds can cause moderate to heavy, persistent rain*

***Altostratus** forms a vast blanket that often shrouds the sky completely and generally indicates an approaching warm front*

*Dense, gray **stratus** cloud covers the sky in a sheet. Rain can fall continuously and for long periods of time if the cloud has depth*

***Stratocumulus** cloud covers the sky in an irregular sheet that can be either gray or white in color. These clouds can lead to drizzle or snow*

***Cumulus** clouds often develop on bright sunny days in slightly unstable airmasses. As long as these billowing, puffy clouds keep their form, they—along with blue sky—indicate good weather*

UNDERSTANDING LOCAL WEATHER

WEATHER IS AFFECTED—and can be predicted—by the local geography surrounding you. High ground forces air to rise and cool, and the relative temperatures of land and sea produce predictable effects.

REGIONAL EFFECTS

It helps to understand how these predictable weather patterns occur in various geographic areas. They may influence your decisions when on the trail.

THE FOHN EFFECT

The leeward side of high ground is warmer and more sheltered than the windward side. Air rises as it travels over obstacles and, if the air contains moisture, the water vapor condenses to form cloud once cooled to a certain level. The air loses water at the summit, then descends and warms on the leeward side.

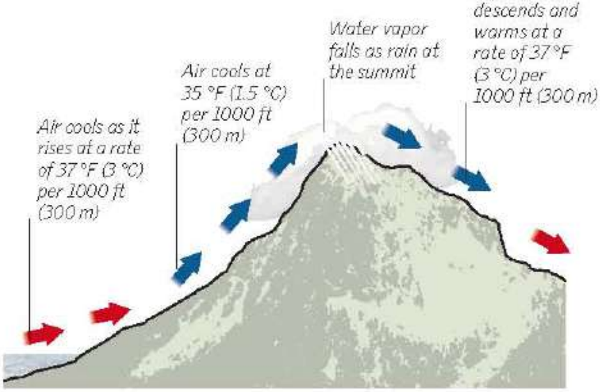
Air cools as it rises at a rate of 37 °F (3 °C) per 1000 ft (300 m)
Air cools at 35 °F (1.5 °C) per 1000 ft (300 m)
Water vapor falls as rain at the summit
Dry air descends and warms at a rate of 37 °F (3 °C) per 1000 ft (300 m)

ANABATIC AND KATABATIC WINDS

Winds that flow up and down slopes during certain atmospheric conditions are known as "anabatic" and "katabatic." They typically occur in mountainous regions.

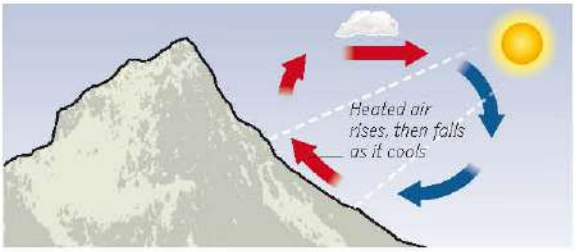
Heated air rises, then falls as it cools

ANABATIC WINDS

During the day, the surface of sloping terrain heats up. The air rises and creates a gentle upslope breeze. Anabatic winds are lighter than katabatic winds because they act against gravity.

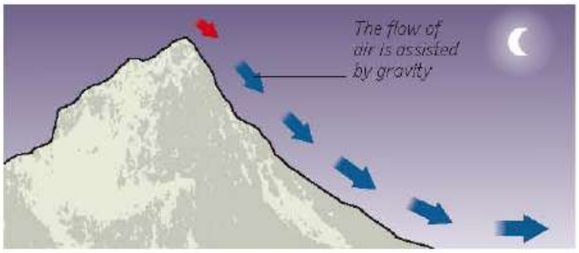
The flow of air is assisted by gravity

KATABATIC WINDS

Katabatic winds form on clear nights with a light breeze. Air in contact with the ground cools and its density increases, causing it to flow down the slopes of the hillside.

SEA AND LAND BREEZES

Sea breezes often develop along coastlines on clear sunny days, whereas land breezes tend to develop along coastlines on cloudless nights.

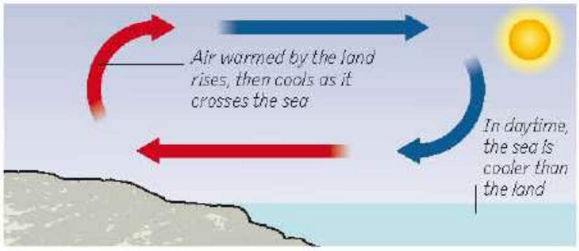
Air warmed by the land rises, then cools as it crosses the sea
In daytime, the sea is cooler than the land

SEA BREEZE

Warm air rises over land during the afternoon, drawing in cool air to replace it along the coast. The result of this coastal air circulation by day is a wind that blows from sea to land.

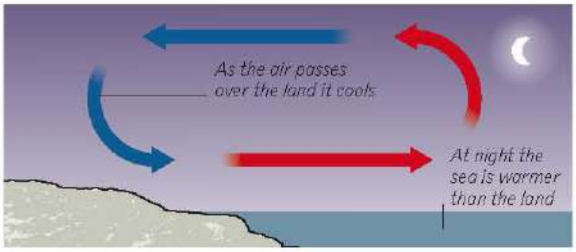
As the air passes over the land it cools
At night the sea is warmer than the land

LAND BREEZE

At night, the sea becomes comparatively warmer and air begins to rise. Air from the neighboring land is drawn toward the sea. The result is a wind that blows from land to sea.

NATURAL WEATHER FORECASTERS

These natural indicators are based on observation and can be useful if you have no alternative means of predicting the weather.

LOOKING AT THE SKY

- If the sky is red at dawn there is moisture in the air and a potential storm ahead. A red sky at night often indicates good weather to come.
- A rainbow usually indicates good weather is on the way, or a light shower.

PLANTS AND FLOWERS

- The scent of plants and flowers is often stronger before rain.
- Certain flowers, such as Scarlet Pimpernel and Morning Glory, are said to close up if bad weather is on the way.
- Pine cones are one of the best natural forecasters—their scales absorb moisture in the air and close up if wet weather is approaching, and unfurl in dry air.

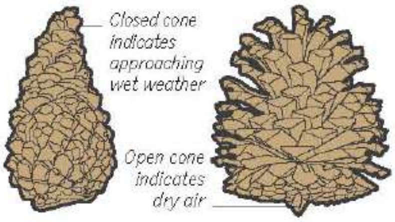
Closed cone indicates approaching wet weather
Open cone indicates dry air

WATCHING ANIMALS

- Animals migrating from high to low areas may indicate that a storm is on the way.
- Cows often lie down before it rains.
- Wool reacts to moisture in the air by swelling and straightening. It shrivels in dry weather conditions.
- Humans can also sense atmospheric change—some people suffer headaches before a thunderstorm.

CHANGING WEATHER

Watch out for alterations in the wind direction or strength as this may lead to a change in the weather. A dry, steady wind that changes direction or decreases in strength often indicates that rain is on its way.

THE BEAUFORT SCALE

Providing visual references for the effects of wind speed on land and at sea, the Beaufort Scale was designed by Sir Francis Beaufort in 1805. It provides a simple way of estimating wind speeds without the need for equipment, and is still in common usage today. The scale ranges from calm to hurricane, and is numbered from zero to 12.

BEAUFORT NUMBER	WIND DESCRIPTION	WIND SPEED MPH (KPH)	WIND EFFECT ON LAND AND AT SEA
0	Calm	0 (0)	Smoke rises vertically. Sea like a mirror.
1	Light air	1–2 (1–3)	Smoke drifts gently. Sea ripples like scales.
2	Light breeze	3–7 (4–11)	Leaves rustle. Wind felt on skin. Small wavelets.
3	Gentle breeze	8–12 (12–19)	Leaves and twigs move. Large wavelets with scattered whitecaps.
4	Moderate breeze	13–18 (20–29)	Small branches move. Small waves with frequent whitecaps.
5	Fresh wind	19–24 (30–39)	Small trees begin to sway. Moderate waves with some foam and spray.
6	Strong wind	25–31 (40–50)	Umbrella usage is difficult. Large waves of 10 ft (3 m) with some spray.
7	Near gale	32–38 (51–61)	Whole trees sway. Sea heaps up and foam begins to streak.
8	Gale	39–46 (62–74)	Walking is difficult. Moderately high waves of more than 18 ft (5 m).
9	Severe gale	47–54 (75–87)	Damage to roofs. High waves with toppling crests.
10	Storm	55–63 (88–101)	Trees uprooted or broken off. Very high waves and sea surface appears white.
11	Severe storm	64–74 (102–119)	Houses damaged. Exceptionally high waves of more than 38 ft (11 m).
12	Hurricane	More than 75 (More than 120)	Buildings destroyed. Huge waves of more than 46 ft (14 m).

MAKE A MOVE

While some survival situations are simply unavoidable, many are entered into because basic techniques relevant to a particular mode of transport have not been followed correctly. This could be due to a lack of knowledge, a loss of concentration, an individual's over-confident belief in his or her own equipment or skills, or as a result of sheer recklessness. Therefore, before you venture into an unfamiliar environment, make sure that you carefully research the type of terrain you will be encountering and investigate the best method of traveling safely and efficiently across it. Knowing the correct techniques for scrambling up a mountain or walking down a steep hillside, for example, could mean the difference between an enjoyable day out and a life-threatening survival situation arising from a fractured ankle in a remote location. Equally, being able to regain control of your vehicle after it has entered a skid could quite simply save your life. Thoroughly researching the terrain will also allow you to select the correct

In this section **YOU WILL DISCOVER...**

- how to make a **paddle** (so you're never **up a creek** without one)...
- the difference between your **finger shelf** and your **hand jam...**
- how to **pull** your own **pulk** and **scramble** over **scree...**
- when to **blaze a trail** or **ski without skis...**
- how to tell your **front-wheel skid** from your **skidoo...**
- when to let the **camel train** take the **strain...**
- how **chewing gum** can stop that **sinking feeling...**

equipment and most appropriate clothing, and enable you to familiarize yourself with the survival techniques relevant to a particular environment.

Whether you're planning to travel on foot, by boat, on horseback, in a 4WD vehicle, or by boat, you must consider not only your own capabilities and those of other members of your team, but also the capabilities of your equipment. Remember, pushing anything beyond its limits—whether that be a person, an animal, or a vehicle—will invariably lead to it failing.

A walking staff is one of the simplest yet most important survival aids you will ever need. It's the first piece of equipment to improvise should you find yourself in a survival situation.

THE "SURVIVOR'S THIRD LEG"—as a walking staff is also known—increases your ability to support yourself by allowing you to have two points of contact with the ground at any one time. This will reduce the chance of you slipping—a crucial factor given that your ability to walk is your main means of rescue; reduce your mobility and you seriously reduce your ability to survive.

A versatile tool, your survivor's third leg can be employed in many different survival situations. Among its many uses, it can be employed to:

- Support you as you walk
- Protect your face when you're walking through thickets or gorse
- Check for adequate support when crossing marshy ground
- Test the ground ahead for obstructions
- Check the depth of water when you're crossing streams and rivers
- Protect you against wild animals
- Form a ridgepole for your shelter
- Help you with your pace counting
- Spear fish or catch game
- Dig up roots or plants

> “Your ability to walk may be your **primary means of rescue**; reduce your mobility and you dramatically reduce your **chances of survival**”

TRAVELING ON FOOT

HIKING IS A GREAT WAY to explore wilderness terrain. It's important to have a decent level of fitness before you set out, and to wear and carry appropriate clothing and equipment. Hiking requires basic skills—using the correct techniques will help you move more efficiently, and ensure that your trip is safe.

BASIC WALKING SKILLS

Aim for a slow, even pace that can be maintained for the duration of the hike by all members of the group. A good way to maintain this pace is to develop a hiking rhythm. Take regular rest breaks and, if you're walking in a group, ensure everyone knows the route.

UPHILL TECHNIQUES

When traveling uphill, lean forward slightly, maintain your momentum but shorten your stride. When pushing upward, keep your feet flat on the ground.

- Carry a walking staff for support as you ascend.
- Move your legs forward from the hips.

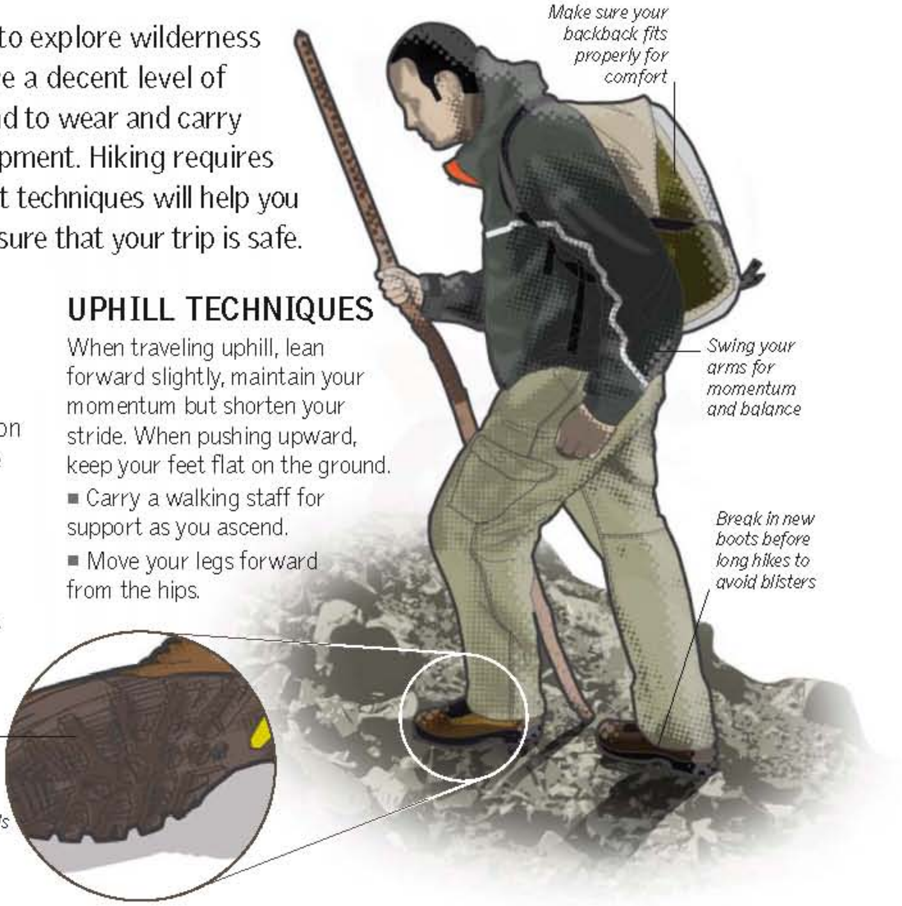
Make sure your backback fits properly for comfort
Swing your arms for momentum and balance
Break in new boots before long hikes to avoid blisters
Deep tread grips the ground and reduces slipping

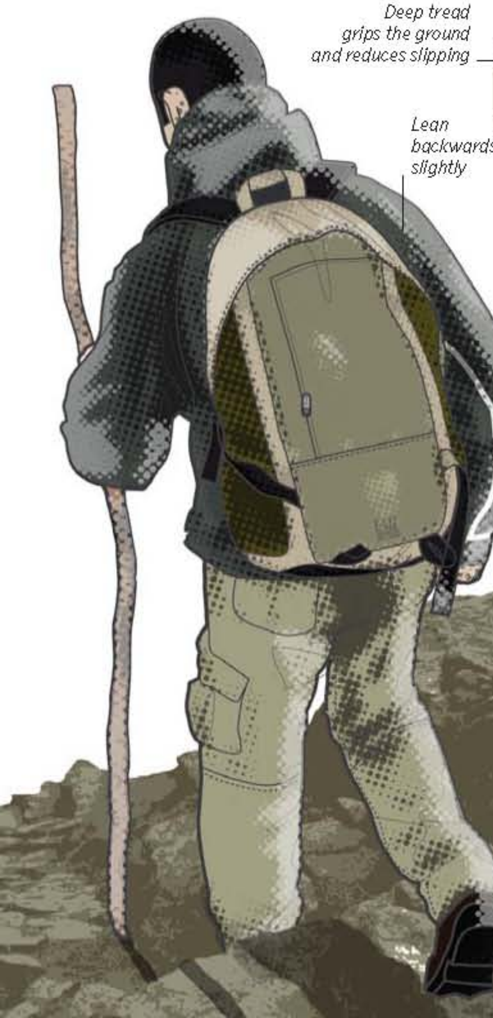
Lean backwards slightly

DOWNHILL TECHNIQUES

Walking downhill can place a lot of strain on your thighs, knees, and ankles, especially when you're carrying a heavy backpack. Be careful not to lose control or gather too much speed.

- Use your arms for balance.
- Maintain a steady rhythm.

WALKING AT NIGHT

Unless you're in the desert and it's cooler to move in darkness, avoid walking at night because of the added risks posed by navigational difficulties, and predators that hunt at night. If there's no alternative, try the following:

- Use your flashlight or improvise a torch by setting alight some birch bark or similar.
- If this isn't possible and you've time, close your eyes for 20 minutes to allow them to adjust to night vision.
- Use your walking stick or pole to feel in front of you for obstacles, tripping hazards, or sudden drops in the ground.
- Keep your pace slow and deliberate, and check your compass regularly.

NEGOTIATING DIFFICULT GROUND

One of the most challenging terrains to hike over is "scree," a mass of small rocks that slides underfoot making uphill and downhill travel difficult. The slippery nature of the surface can make progress slow, and care should be taken to avoid falling and injuring yourself. Hiking on scree is hard work, but using the correct techniques will help you to advance confidently, efficiently, and safely.

TRAVERSING SCREE

Choose a zig-zag route and look for a path that contains similar-sized rocks. Walk sideways across the slope, taking small steps and testing rocks for stability before placing your full weight on them.

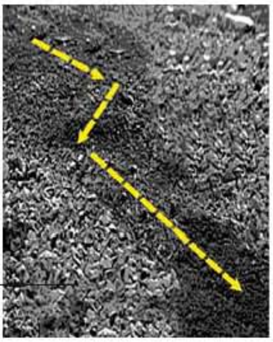
Rocks at the edge of the slope are likely to be larger and more stable

ASCENDING SCREE

Tread carefully, kicking your toes into the slope and testing the step before transferring your weight to that leg. Alternatively, walk with your feet splayed, placing your weight on the inside of each foot.

DESCENDING SCREE

"Screeing" downhill combines sliding with slow-motion jogging. When you have the technique it can be great fun, but avoid larger rocks to prevent ankle injuries.

TRAIL MARKERS

Your map and compass should be your primary method of navigation, but keep an eye out for trail markers, or "blazes," during your hike.

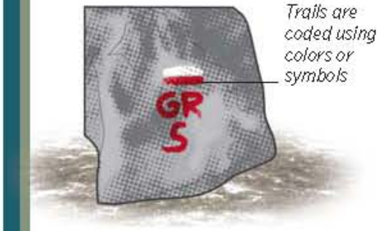
GR
5
Trails are coded using colors or symbols

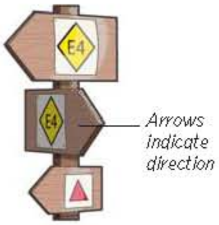
E4
E4
Arrows indicate direction

ROCKS

Trail blazes painted onto rocks and boulders are especially common in rocky terrain. They may be low on the ground so look carefully.

SIGNPOSTS

Made from wood, metal, or plastic, these signposts are particularly useful in areas where there are few rocks or trees on which to blaze.

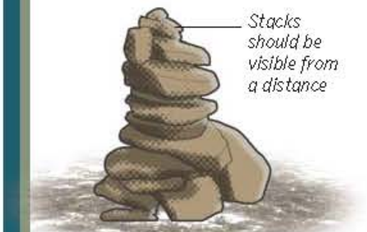
Stacks should be visible from a distance

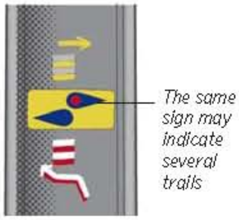
The same sign may indicate several trails

CAIRNS

Piles of rocks, known as "cairns," are designed to be visible in thick fog. Their size can vary from a few rocks to a large stack of boulders.

DIRECTIONAL SIGNS

When reading markers, look out for painted arrows or variations in the blazing, such as a bend. This indicates a change of direction in the trail.

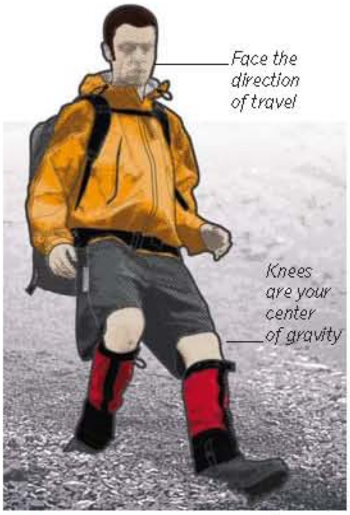
Face the direction of travel
Knees are your center of gravity

① Ensuring the scree is fine and deep, use gentle hops to launch yourself down the slope. Keep balanced and let gravity do the hard work.

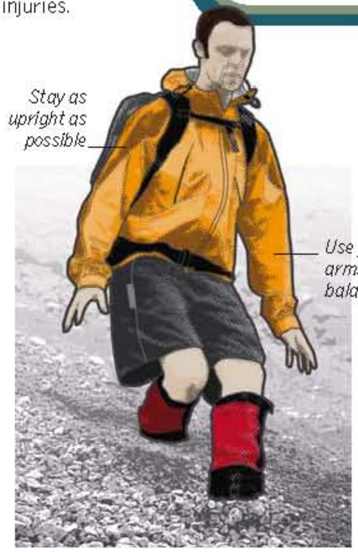
Stay as upright as possible
Use your arms for balance

② As you build up momentum and rhythm, dig your heels into the slope and slide a short distance with each step, avoiding leaning forward.

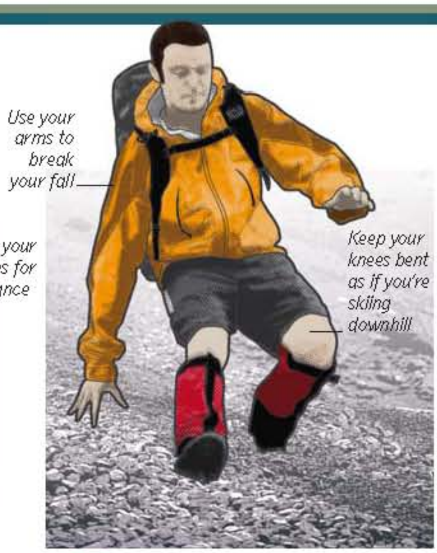
Use your arms to break your fall
Keep your knees bent as if you're skiing downhill

③ If you lose your balance temporarily, use your arms to steady yourself, relax your knees, and continue your descent.

CROSSING RIVERS

RIVER CROSSINGS ARE DANGEROUS and should be avoided unless absolutely necessary. In a survival situation, once you are cold and wet—which can lead to hypothermia—it is hard to get warm and dry. Always check your map for routes around the river, and choose the safest available crossing point.

WADING BASICS

Wear footwear when wading to protect your feet from rocks and other dangers. Remove pants to keep them dry and decrease resistance in the water. Use a walking staff for added support.

CROSSING SAFELY

Before you get into the water, make sure you have a change of clothes or means of getting dry on the other side. In cold conditions, collect everything you need to get a fire going (tinder, kindling, and dry wood), and keep this dry as you cross.

CHOOSING YOUR CROSSING POINT

Walking upstream generally leads to shallower water, but be aware that even shallow water can have strong currents, as can water that looks calm on the surface. Always check for bridges further up- or downstream.

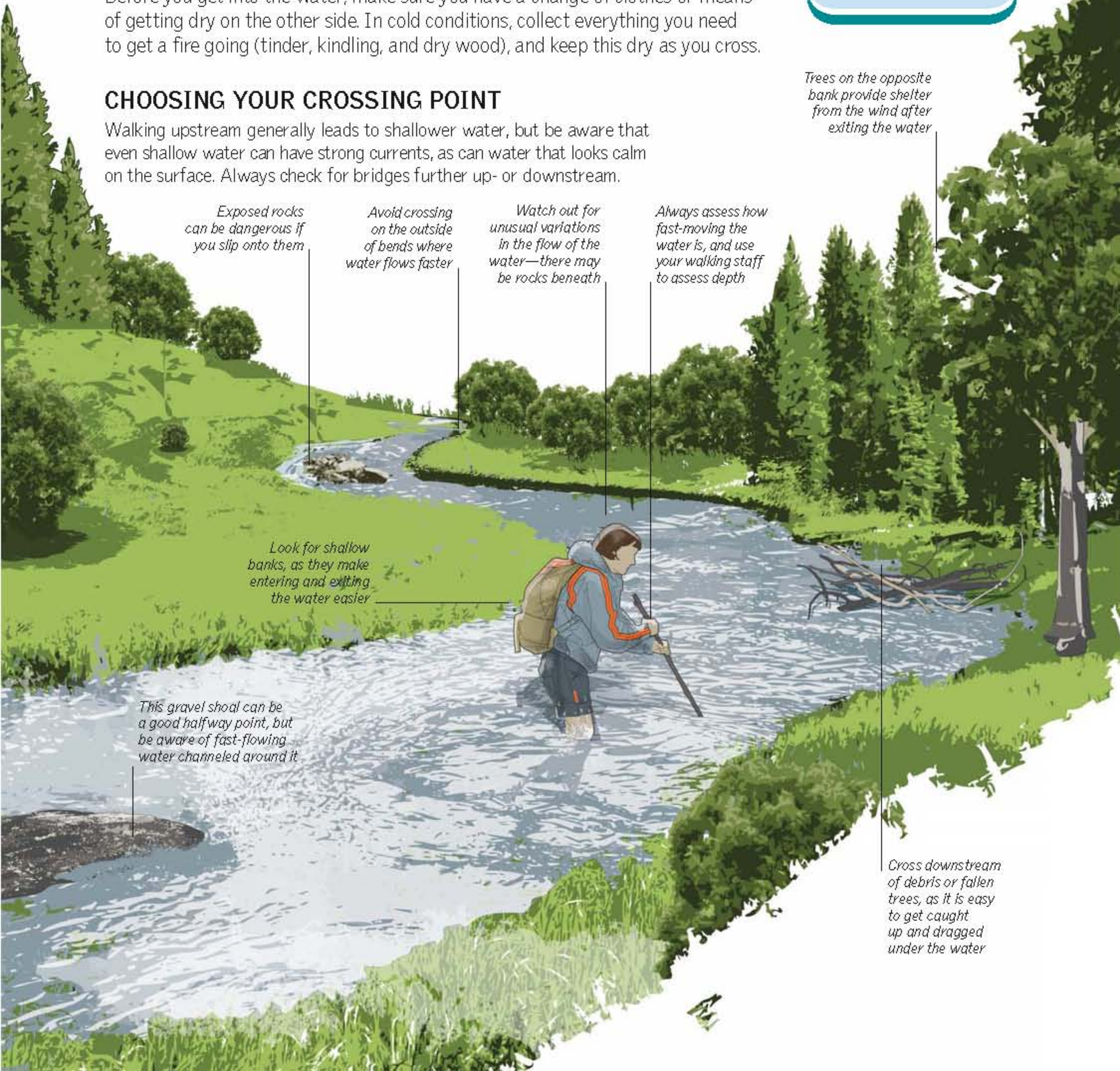
Trees on the opposite bank provide shelter from the wind after exiting the water
Exposed rocks can be dangerous if you slip onto them
Avoid crossing on the outside of bends where water flows faster
Watch out for unusual variations in the flow of the water—there may be rocks beneath
Always assess how fast-moving the water is, and use your walking staff to assess depth
Look for shallow banks, as they make entering and exiting the water easier
This gravel shoal can be a good halfway point, but be aware of fast-flowing water channeled around it
Cross downstream of debris or fallen trees, as it is easy to get caught up and dragged under the water

WADING WITH OTHERS

Crossing in a group is safer than crossing alone. Linking your arms together creates a stronger, more stable structure against the current and provides backup for anyone who falls. Loosen the straps of your backpack, and place only one arm through them, so you can release the pack quickly if you fall.

WARNING!
Never cross white water, or rivers that have flooded. In these situations it is safer not to cross at all than to risk injury.

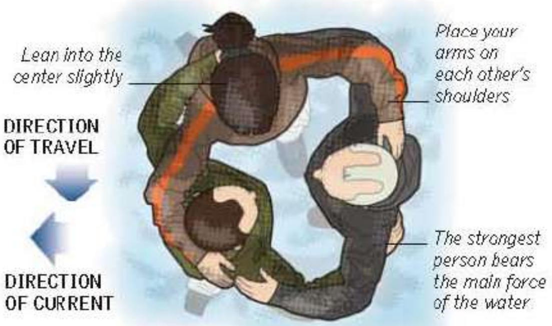
Lean into the center slightly
Place your arms on each other's shoulders
DIRECTION OF TRAVEL
DIRECTION OF CURRENT
The strongest person bears the main force of the water

CROSSING IN A HUDDLE

Positioning the strongest person upstream, with the others providing stability and support, link your arms tightly and take short, deliberate steps across the river.

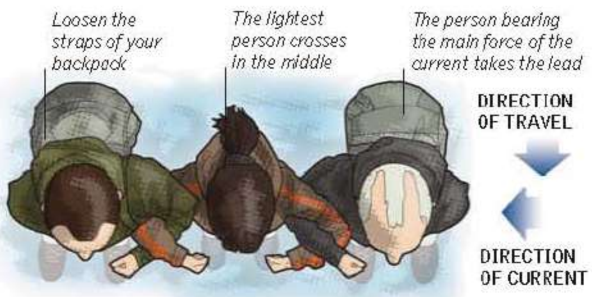
Loosen the straps of your backpack
The lightest person crosses in the middle
The person bearing the main force of the current takes the lead
DIRECTION OF TRAVEL
DIRECTION OF CURRENT

CROSSING IN A LINE

Keeping well balanced, cross the river perpendicular to the current. Move slowly and position each step carefully to avoid being swept away by the force of the water.

WADING ALONE

Wading across a river alone is not ideal, but if you have no option then your walking staff or "survivor's third leg" (see p. 72)—or simply any long stout stick—will provide additional support and balance. Use it as a probe to alert you to any sudden changes of depth in the river bed.

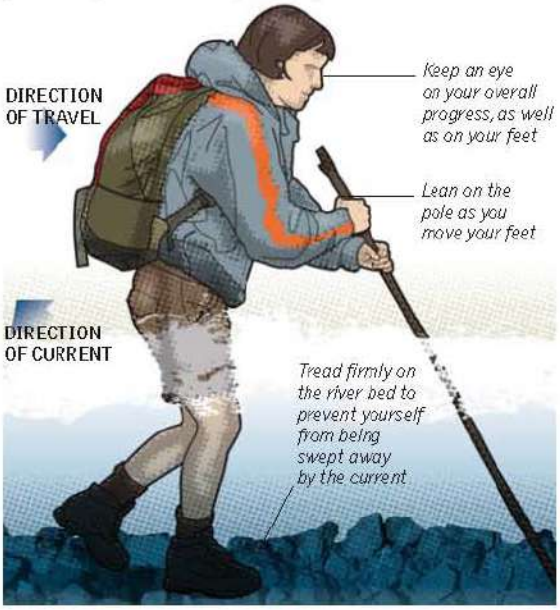
DIRECTION OF TRAVEL
Keep an eye on your overall progress, as well as on your feet
Lean on the pole as you move your feet
DIRECTION OF CURRENT
Tread firmly on the river bed to prevent yourself from being swept away by the current

USING A WALKING STAFF

Facing the current, walk diagonally across the river, placing each foot down firmly and deliberately, and leaning on your walking staff for support. Your feet and your staff are your three points of contact—keep at least two of them in contact with the river bed at all times.

CROSSING WITH ROPES

Using ropes is a good option for dangerous crossings, but they can get tangled, and drag you under the water. Always aim for the easiest and safest option—using ropes is complicated and should be your last resort.

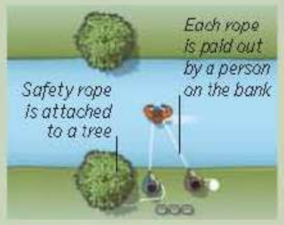
Each rope is paid out by a person on the bank
Safety rope is attached to a tree

① A "safety" rope is tied to a tree and then carried across by the strongest person, who also takes a "crossing" rope with a carabina tied centrally.

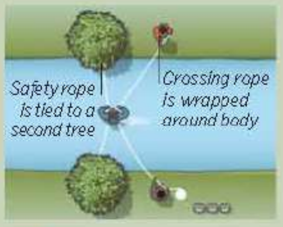
Safety rope is tied to a second tree
Crossing rope is wrapped around body

② The safety rope is tied to a tree and the carabina clipped to the safety rope. The carabina is pulled back and fixed to the next person to cross.

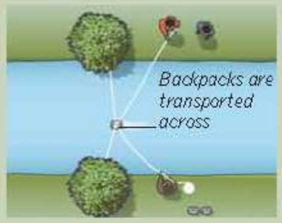
Backpacks are transported across

③ He or she crosses upstream of the safety rope, holding it for support. Backpacks can be clipped to the carabina and transported, too.

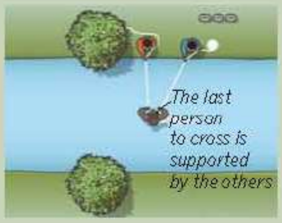
The last person to cross is supported by the others

④ The final person unties the safety rope from the tree, wraps it around his or her body, and wades across, supported by the others.

SCRAMBLING AND CLIMBING

CLIMBING WITHOUT ROPES is known as "scrambling." As with roped techniques, the idea is to maintain three points of contact with the rock at all times—either two feet and one hand or both hands and one foot. Your legs should power the climb, with your arms used predominantly for balance. Proceed carefully, making sure you are always balanced and confident of your next move.

WARNING!
Climbing is a dangerous activity that should only be considered as a last resort. Ideally, plan ahead using your map to find a way around the obstacle.

HAND AND FOOT HOLDS

When planning your route, choose your hand and foot holds carefully. Don't reach too far, and test each hold for stability before using it to support your weight.

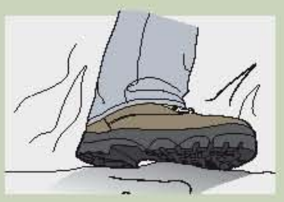

EDGES

For very small footholds, place the inside edge of your foot in the hold in the rock to take the pressure off your toes.

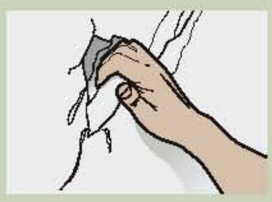

FINGER SHELVES

Curl your fingers over the rock for a secure fingerhold. The larger the fingerhold, the more secure it will be.

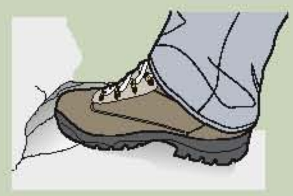

LARGE POCKETS

Placing your foot securely within the pocket, balance yourself so both your arms and legs take the strain.

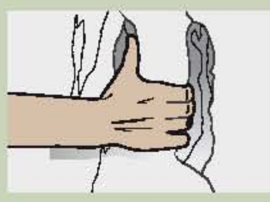

SIDE PULLS

Side pulls can be used to maintain balance or to pull yourself across the rock. Grip the rock tightly.

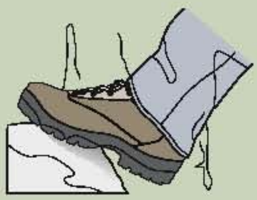

PROTRUSIONS

If possible, place the entire sole of your foot on the protrusion. If only the ball of your foot will fit, keep your heel low.

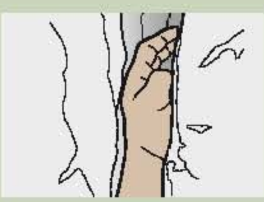

HAND JAMS

Insert your hand into the crack, push your thumb into your palm, arch your hand, and wedge it tightly inside.

BASIC CLIMBING TECHNIQUES

Always climb within your capabilities. It's important not to take any risks, as it's far safer to descend and start the climb again than to chance a fall. Before you start to climb, plan the easiest and safest route.

SCRAMBLING

When moving across the rock you will need to combine a variety of techniques to negotiate different obstacles. Always plan your moves several steps in advance, and keep three points of contact with the rock at all times.

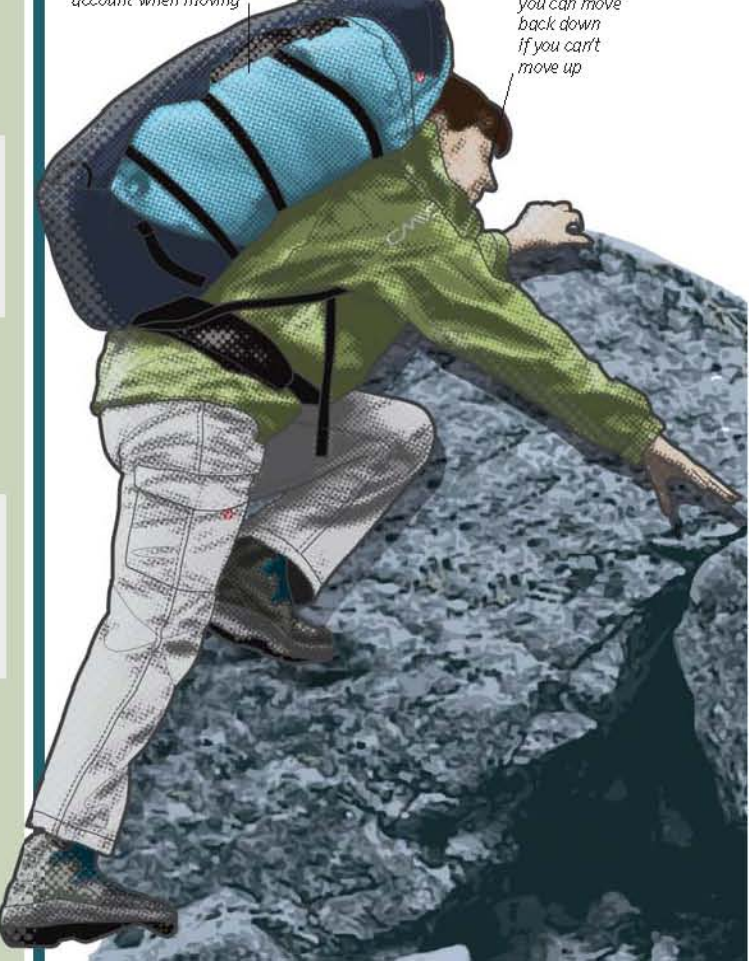

Take the weight of your backpack into account when moving

Always ensure you can move back down if you can't move up

MANTELLING

The mantelling technique is used to climb overhangs in the rock. Use your lead ankle and then your knee to lever your body up over the obstacle. It's a physically challenging technique, but useful.

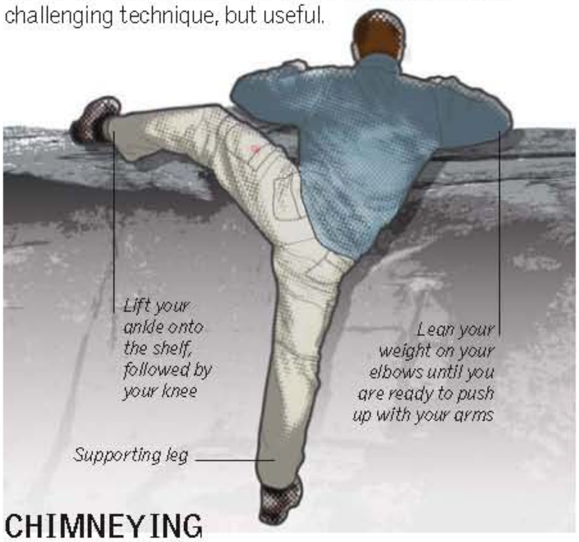
Lift your ankle onto the shelf, followed by your knee
Lean your weight on your elbows until you are ready to push up with your arms
Supporting leg

CHIMNEYING

To climb up the inside of large rock clefts, or "chimneys," use the chimneying technique. To ascend, move your back and hands up the wall, and push and follow with your legs. It's easy to get stuck at the top so plan your exit route carefully.

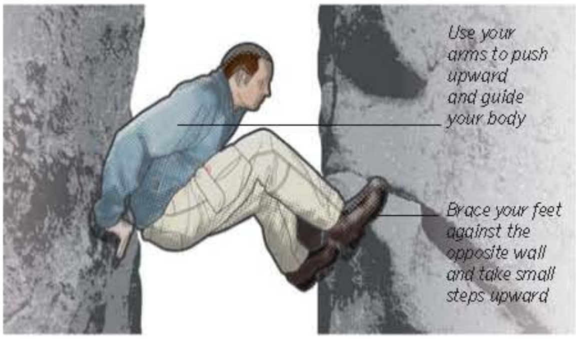
Use your arms to push upward and guide your body
Brace your feet against the opposite wall and take small steps upward

STRADDLING

If the chimney is relatively wide, you may need to alter your body position to straddle the gap. With a leg and arm on each wall, inch yourself up the rock using your legs to push your body upward.

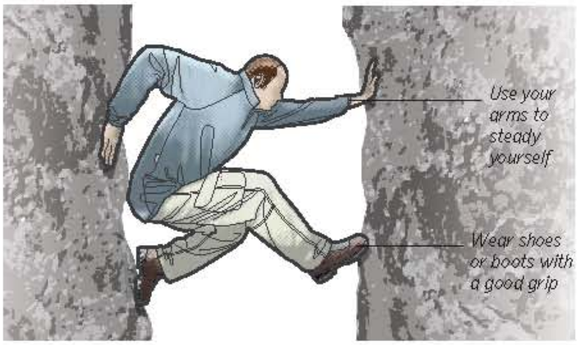
Use your arms to steady yourself
Wear shoes or boots with a good grip

CLIMBING WITH EQUIPMENT

The advantages of climbing with ropes cannot be underestimated as far as safety is concerned, although a lot of specialized equipment is required. As well as ropes, helmets, and harnesses, climbers can also use screws to secure themselves to the rock as they progress each stage.

BELAYING

Belaying is a technique used by climbing partners to safeguard each other during a climb. The belayer secures himself to the rock and stands on a suitable ledge while the lead climber ascends. Once at the top, the lead climber secures himself and acts as the belayer.

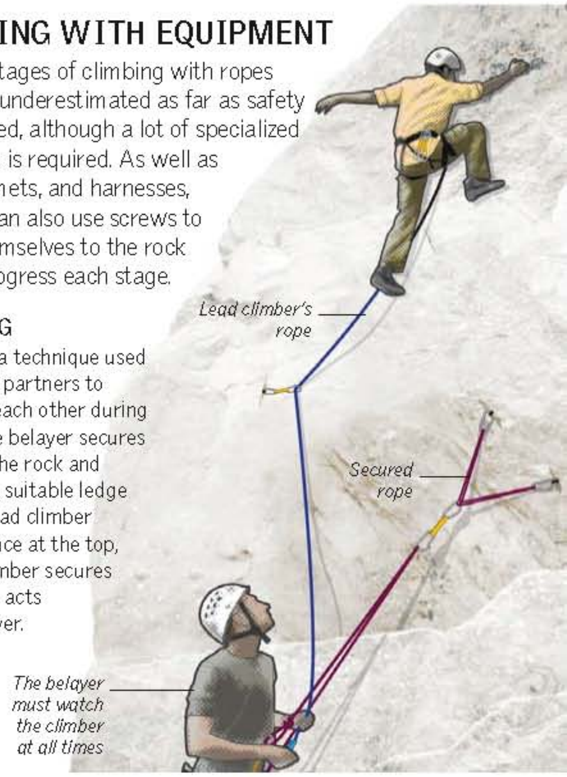
Lead climber's rope
Secured rope
The belayer must watch the climber at all times

ICE CLIMBING

Ice climbers utilize similar techniques to rock climbers, although they carry ice axes and wear crampons on their feet, to help them grip the ice.

MOVING UPWARD

As with rock climbing and scrambling, ice climbers ascend using power from their legs and arms. Crampons and ice axes are used to create holds in the rock.

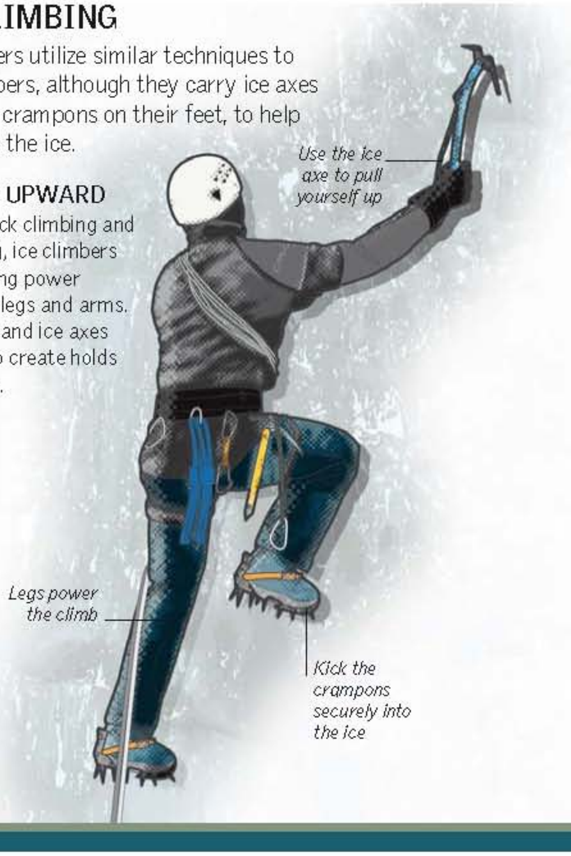
Use the ice axe to pull yourself up
Legs power the climb
Kick the crampons securely into the ice

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—IN THE MOUNTAINS

USEFUL EQUIPMENT

- Two-way radio
- Avalanche beacon and probe
- Waterproof clothing
- Spare clothing layers
- Collapsible shovel
- Climbing equipment, if needed
- Signal flares
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

STEVEN GREEN, 32, AN EXPERIENCED HILL WALKER from Dumfries, Scotland, spent four nights lying on a remote mountainside after a life-threatening fall. Suffering from a broken jaw and fractured skull, he survived via a combination of good preparation, quick thinking, and improvisation.

On Thursday, October 7, 1999, Steven set out alone for a trek in the hills of Wester Ross in the northwest Highlands of Scotland. He was well equipped with, among other things, a cell phone, bivy sack, pita bread, and water—and had taken the wise precaution (or so he thought) of informing his girlfriend of his intended route.

Following an all-day trek and an overnight camp, he was descending the mountain on Friday when he slipped on wet grass and fell down a waterfall, coming to rest in the shallow, rocky riverbed. He knew he would not survive for long in the cold water so, despite being in intense pain, he crawled to the relative safety of the bank, then wisely got into his bivy sack for warmth and waited. He sustained himself by eating pita bread, which he softened in water because he couldn't chew. His phone was damaged in the fall so he was unable to call for help. When he failed to return home on time, the alarm was raised by Steven's girlfriend, but she had forgotten to write down his intended route so no one knew his exact location.

"HE SLIPPED ON WET GRASS AND FELL DOWN A WATERFALL"

Steven experienced a further four days and nights of exposure while local and RAF rescue teams and search-and-rescue dogs combed the mountains for him. As the days passed they feared the worst, but when they located Steven on Tuesday morning—after finding his car and, inside it, his route map—they found him in good spirits. At hospital Steven was diagnosed with a fractured skull, broken jaw, missing front teeth, cuts and bruises, and a leg injury—but he was alive.

WHAT TO DO

ARE YOU IN DANGER?

NO: If you are in a group, try to help any others who are in danger

YES: Get yourself out of it:
Elements—Find or improvise immediate shelter
Animals—Avoid confrontation and move away from danger
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO: If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

YES: If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO: You are faced with surviving for an indefinite period—until you are located or you find help

YES: If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO: If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers a better chance of survival, rescue, or both

YES: Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Keep all clothing dry and clean
- Collect drinking water from fast-moving streams; filter and disinfect if you have the means
- Be on constant lookout for signs of cold-related injuries such as frostnip, frostbite, and hypothermia
- Regulate your clothing to avoid overheating when moving and hypothermia when static
- Use a walking staff to aid safe movement
- Watch the weather closely and be prepared to change your plans at short notice—mountain climates are highly unpredictable
- Improvise shelter when not on the move

DON'T

- Descend hills in a careless manner; zig-zagging across hills is less tiring and puts less strain on leg muscles
- Travel at too fast a pace—high altitude equals less air and this will place greater demands on even a fit person
- Sweat too much, as the moisture will chill you further
- Ignore any opportunities to collect dry tinder/kindling

YOU SHOULD STAY **

DO

- Select a shelter site that offers protection from the elements. Build it no bigger than it needs to be
- Fill plastic bags or spare clothing with dry foliage and use as a mattress or pillow to insulate you from the cold, damp ground
- Deploy all your aids to location and prepare for immediate use
- Check upstream for the quality of your water source
- Light a fire and (if in a group) take turns tending it to keep it going all night
- Continually re-assess your situation and adapt your actions as necessary.
- Be constantly alert for signs of rescue

DON'T

- Allow your extremities to get too cold as those areas are most susceptible to frostbite
- Overlook the dangers of carbon monoxide poisoning in cramped shelter. Don't let candles, stoves, or fires burn all night
- Breathe on your hands to warm them up: you are exhaling warm air that you will have to replace with cold

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.

** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

MOVING OVER SNOW

PREPARATION IS ESSENTIAL when hiking over frozen terrain. Not only must you be physically fit—the conditions make for slow, exhausting progress—but you must also have the right equipment and know how to use it. Wearing snowshoes or skis, and breathable and layered clothing to regulate your temperature, is essential. With the correct techniques you can hike safely and enjoy the surroundings.

WARNING!

Walking over deep snow without snowshoes is called "post-holing" and should be avoided unless absolutely necessary. Sinking into the snow will leave you exhausted and very wet, which in cold conditions can quickly lead to hypothermia. The exertion can also cause sweating and a dangerous reduction in your body temperature when you stop.

SNOW AND ICE

Knowing how to deal with different types of snow and ice is invaluable when hiking over frozen terrain.

DEEP SNOW

- If in a group, walk in single file and take turns at the front: the most strenuous position, as you're creating the path.
- Avoid rocks—in spring, they absorb heat and the snow above becomes unstable.

FROZEN CRUST

- Use a walking staff to test the snow ahead. An ice crust above deep snow may take your weight, but progress carefully.
- Later in the day you may come across melted depressions, called "sun cups." Cross on the rims to avoid sinking into the snow.

SLOPES

- Kick firmly into the slope and test your weight before ascending. When descending, you can use a technique called "boot skiing" (skiing without your skis).
- Choose your route based on the conditions: travel in straight lines if you are able to, or in zig-zags if the terrain is steep.

ICE

- Always progress carefully, using your walking staff to test the ice, especially over rivers and lakes. If in a group, rope yourselves together for safety.
- Wear crampons for extra grip. On very steep slopes, use your ice ax to cut steps.
- Use your ice ax to halt a fall by turning to face the slope and digging it into the ice.

GLACIERS

- Never attempt to cross glaciers without a guide—glaciers require specialized skills.

WAYS OF TRAVELING

When traveling over snow, your main aim is to get to your destination as safely as possible, without expending too much energy or losing too much body heat. Methods range from improvised snowshoes to motor vehicles.

USING SNOWSHOES AND SKIS

Snowshoes and skis are an effective way of traveling over snow. They work by spreading your body weight over a larger surface area, which enables you to walk on the surface of the snow rather than sink into it. Always set off cold, as you'll warm up quickly, and add or remove layers as required.

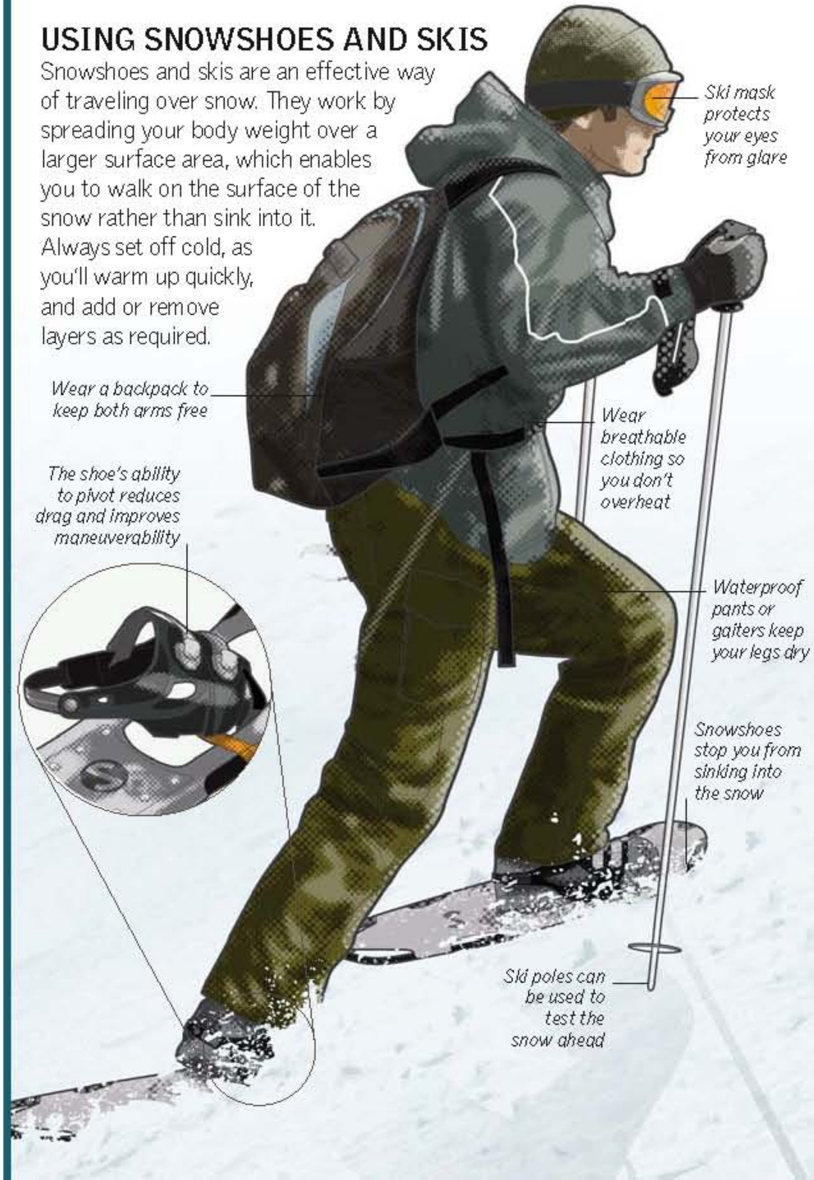

Ski mask protects your eyes from glare

Wear a backpack to keep both arms free

Wear breathable clothing so you don't overheat

The shoe's ability to pivot reduces drag and improves maneuverability

Waterproof pants or gaiters keep your legs dry

Snowshoes stop you from sinking into the snow

Ski poles can be used to test the snow ahead

MAKING IMPROVISED SNOWSHOES

If you don't have any pre-manufactured snowshoes—for example, in a survival situation—you can make some very simple shoes to help you negotiate the snow more effectively, using your knife, some green wood, and cordage. They may take time to build, but will save time and effort in the long run.

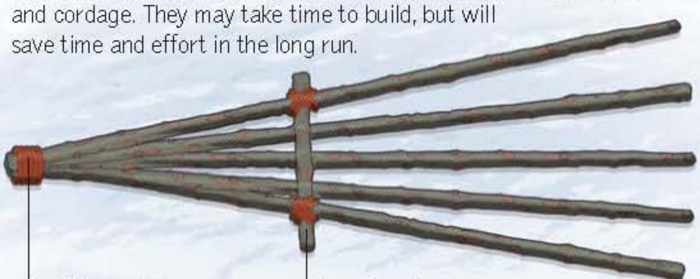
Lash the ends together securely
Cross-piece for ball of foot

1 Cut five lengths of green wood. They should be as thick as your thumb and the same length as the distance from your foot to your armpit. Cut three shorter lengths for the cross-pieces.

- Lash the ends of the five longer pieces together securely using cordage.
- Calculate where the ball of your foot will be positioned on the shoe and lash a cross-piece across the five lengths. Ensure the shoe will balance.

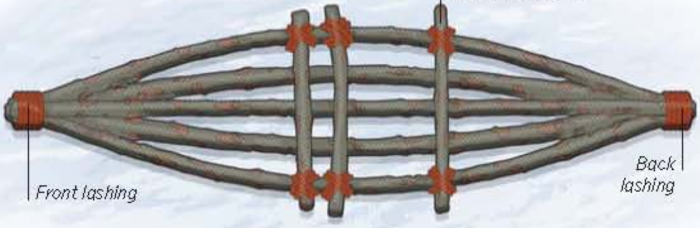
Additional cross-piece for your heel
Front lashing
Back lashing

2 Lash the five loose lengths of wood together at the back of the shoe. It's important to make this lashing as secure as possible.

- Fix the second cross-piece roughly 2 in (5 cm) behind the first.
- Lash the third cross-piece where your heel will rest.
- Repeat steps 1 and 2 to make a second shoe before progressing to step 3.

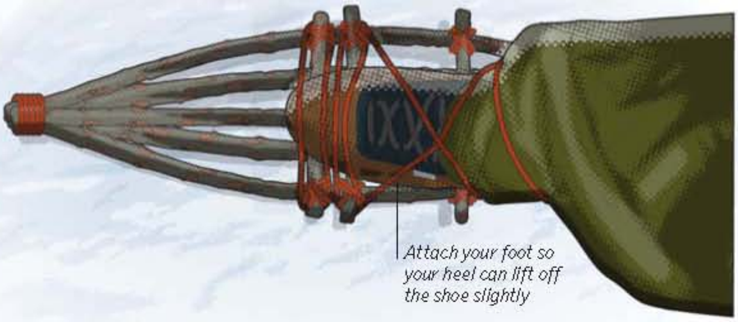
Attach your foot so your heel can lift off the shoe slightly

3 Place your foot on top of the snow shoe, ensuring that the ball of your foot sits directly over the front cross-piece and that your heel is positioned on the back cross-piece.

- Tie your boot to the snow shoe using whatever cordage you have, but ensure the heel is allowed to remain free to pivot. Repeat for the other foot.

ALTERNATIVE METHODS

These alternative methods will also spread your weight on the snow. If you have no gaiters, tie plastic bags around the bottom of your legs to keep them dry.

USING BRANCHES

Ideal for short distances, a simple way to get you out of deep snow and to a road or track is to attach branches to your feet using cordage. Select a tree, such as pine, that has strong, close branches.

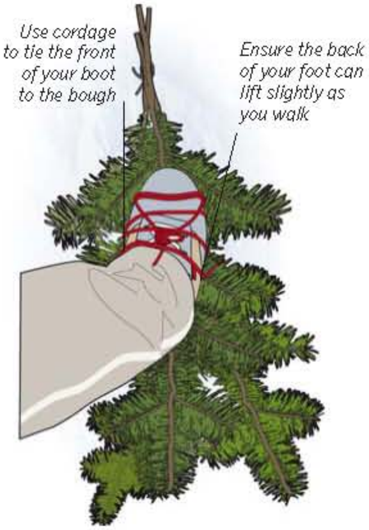
Use cordage to tie the front of your boot to the bough
Ensure the back of your foot can lift slightly as you walk

USING SAPLINGS

You will need branches that have some flexibility. Gently bend the longest branch into a teardrop shape and lash the two ends together. Heating the saplings over a fire will make them easier to bend.

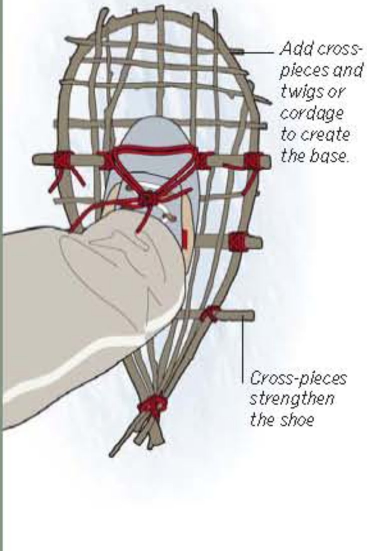
Add cross-pieces and twigs or cordage to create the base.
Cross-pieces strengthen the shoe

MAKING A SLED

If you find yourself in a survival situation with heavy equipment to carry over snow, making a simple sled or "pulk" (see panel, right) will help you transport it more efficiently. You could also build one to transport an injured person or small child. You can make your sled as large or as small as you need, providing you follow these basic principles.

TOOLS AND MATERIALS

- Survival saw or pocket chainsaw
- Knife
- Cordage
- Forked branch (for bulk of sled)
- Two branches (for bracing pieces)
- Sticks (for cross-pieces)

1 Using a short saw or penknife, cut a forked branch to form the base of the sled.

- Tie the two ends of the fork to the main branch with cord. This creates tension and forces the branches to curve and act as runners.

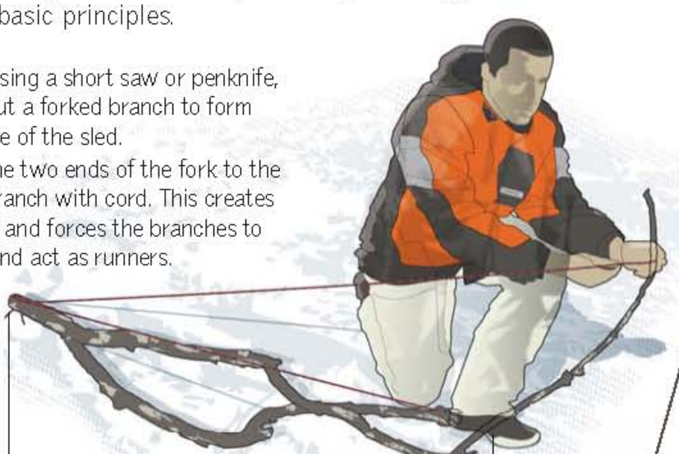

Attach the cord to the main branch using a slip knot (see p. 142)

Attach the cord to the end of the fork using a taut line hitch (see p. 145)

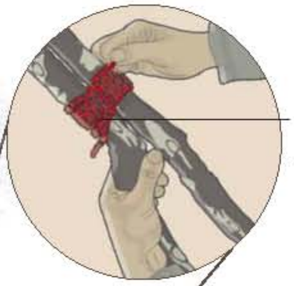

Lash the bracing piece to the main branch

2 Lash an additional branch to one of the runners to create a bracing piece.

- Repeat on the other side.
- The bracing pieces strengthen the branches and prevent any cord from touching the ground—if it does, it will wear out quickly and disintegrate once the sled is on the move.

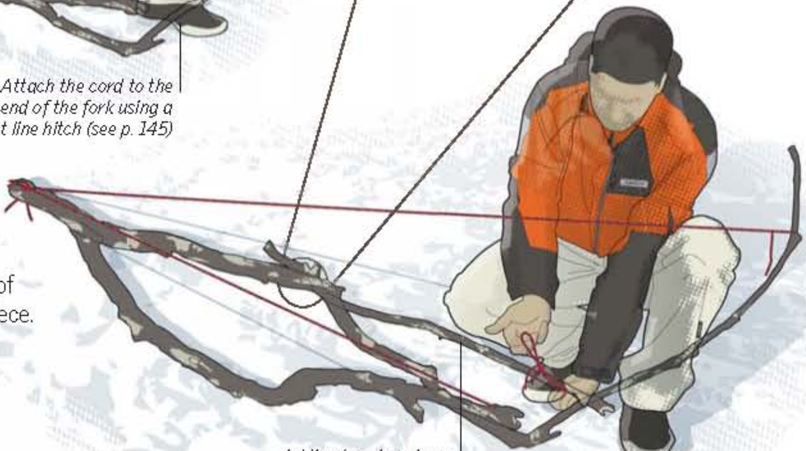

Adding bracing pieces creates a support for the carrying platform

ANIMAL POWER

Huskies, or sled dogs, are ideal for transporting people and equipment across winter terrain. Because of their thick coat they can endure very low temperatures, and their large, furred feet allow them to move quickly over snow. Huskies work well as a team and can pull heavy loads easily and efficiently.

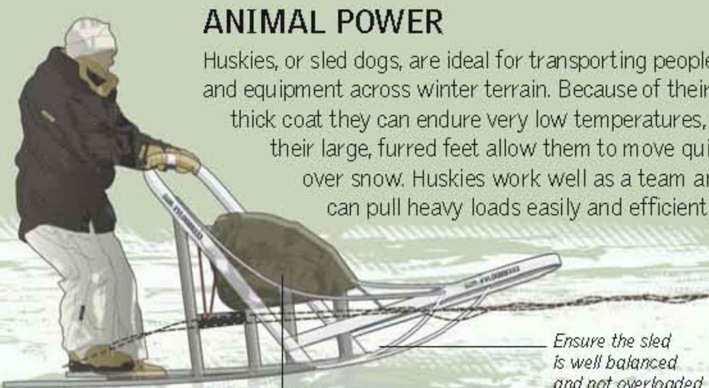

Ensure the sled is well balanced and not overloaded

Tie luggage securely to the frame

TRAVELING WITH DOGS

Although huskies are easy to care for and relatively simple to handle, you should not undertake an expedition with dogs unless you are traveling with an expert or have first undergone extensive training. The same goes for all expeditions in snowy terrain.

HUMAN POWER

Pulling your equipment behind you using a pulk is an extremely efficient way of moving over snow. Pulks are small, low-slung toboggans, typically made from lightweight plastic, and come in many different sizes.

TRAVELING WITH A PULK

Although pulks are the most efficient method of carrying heavy loads over snow using human power, they can be hard work, especially in softer snow. Wear breathable clothing to allow excess body heat to escape. Know how to release your harness quickly in an emergency. If you are in a group, one person can be harnessed to the rear of the pulk to act as a brakeman when going downhill.

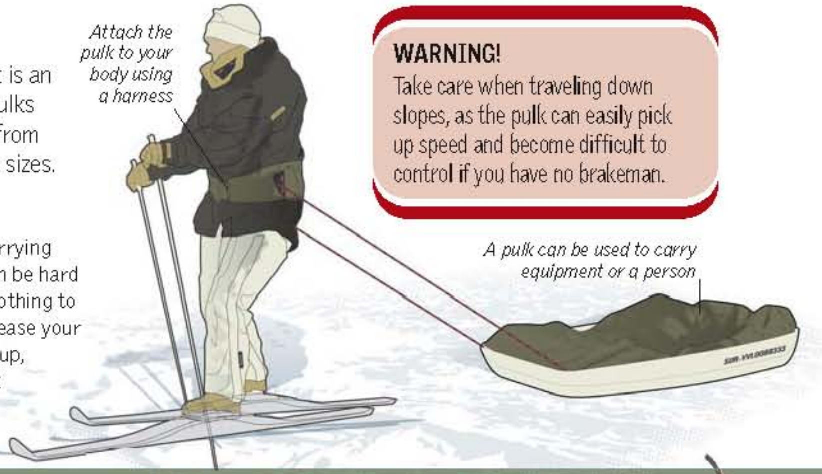

Attach the pulk to your body using a harness

A pulk can be used to carry equipment or a person

WARNING!
Take care when traveling down slopes, as the pulk can easily pick up speed and become difficult to control if you have no brakeman.

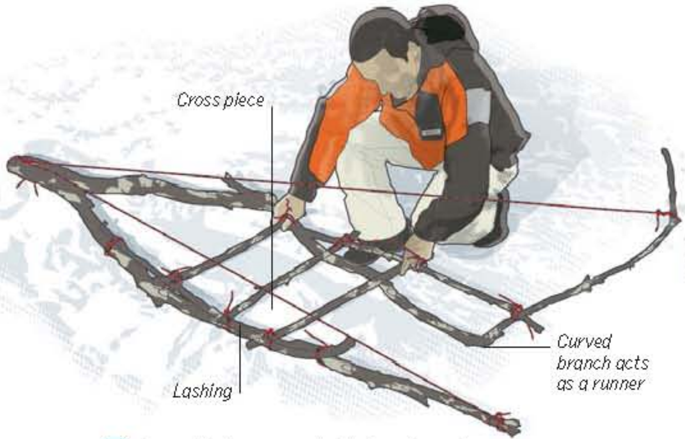

Cross piece

Lashing

Curved branch acts as a runner

3 Lay sticks across both bracing pieces and lash them together. These cross-pieces strengthen the structure.

- You can use as many sticks as you can find, but three or four should suffice.
- These cross-pieces will also form the main carrying platform of your sled.

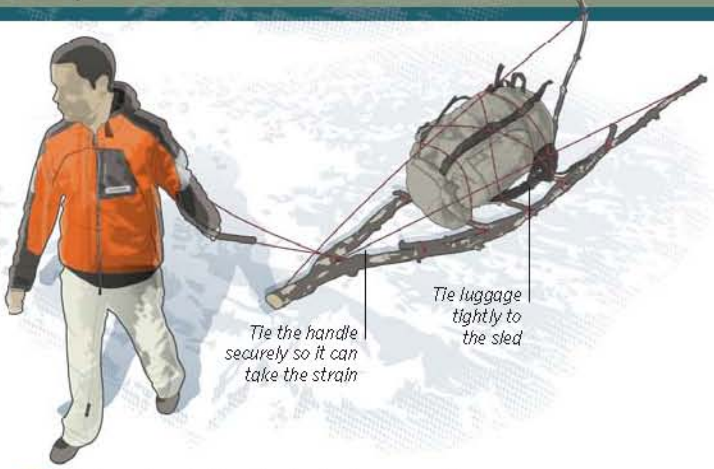

Tie the handle securely so it can take the strain

Tie luggage tightly to the sled

4 Make a handle for the sled using a small branch. The branch should be as smooth as possible to make pulling the sled more comfortable.

- Attach the handle to the front of the sled using a long piece of cord.

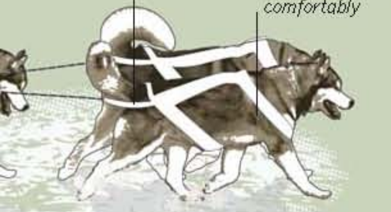

Linking the dogs together helps them to power forward in unison

Harnesses allow the dogs to pull comfortably

MOTOR POWER

Snowmobiles are practical and quick, and have revolutionized arctic travel. They are easier to handle than huskies, and can reach relatively high speeds. However, although you can cover great distances in short periods of time, if anything goes wrong you may be far from help. Always carry survival equipment when you travel.

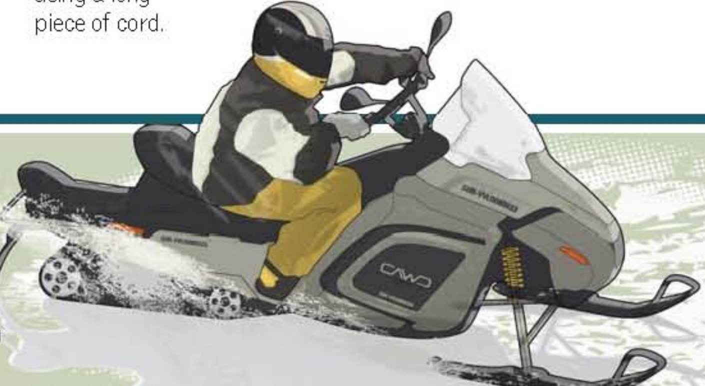

BEFORE YOU GO

Prior to a snowmobile trip, regardless of distance, it is vital to check your vehicle maintenance and ensure that you have sufficient fuel for the trip. Plan your route carefully and inform someone of your plans so that they can raise the alarm if you don't reach your destination when planned. Dress warmly, wear goggles or a ski mask to reduce glare, and drive safely.

USING PACK ANIMALS

PACK ANIMALS CAN BE USED to transport heavy loads when hiking over wilderness terrain. Ideally suited to carrying substantial loads, they can cover great distances in areas where vehicles would struggle. When you are organized, and know how to care for the animals, traveling with pack animals can be very rewarding.

PACKING TIPS

- Know the capabilities of your animal, and what weight they can safely carry over the necessary distance.
- Before loading the animal, groom it well and check for sores or ticks.
- Place heavier items at the bottom of your panniers and lighter items at the top.

LOADING THE ANIMAL

Stand on level ground, making sure that the animal has all four legs placed firmly on the ground. Before placing the loaded panniers onto the animal, lift them to ensure they are well balanced. Rearrange if one is lighter. You may need to hobble or blindfold the animal to keep it still while loading.

ANIMAL WELFARE

Remember to transport food, water, and supplies for the animal as well as for yourself. Their welfare is paramount, so treat them well.

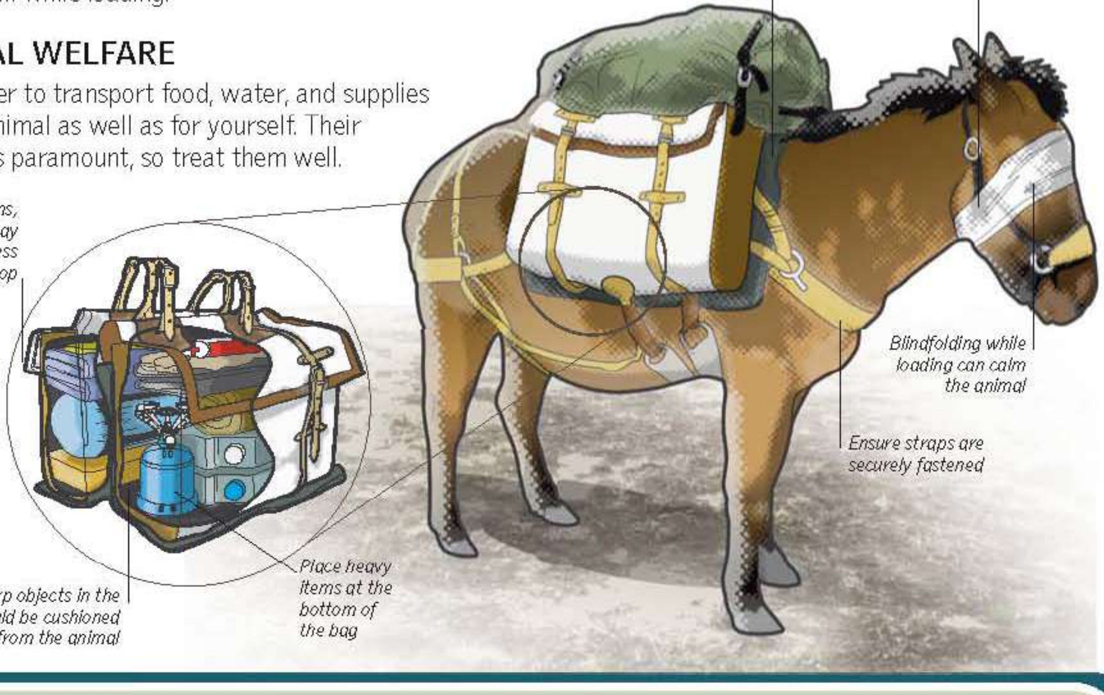

Fit a blanket for the animal's comfort

A headcollar gives you control over the animal

Place lighter items, and items you may need to access quickly, at the top

Blindfolding while loading can calm the animal

Ensure straps are securely fastened

Place heavy items at the bottom of the bag

Any sharp objects in the bag should be cushioned from the animal

PACK TRAINS

Roping animals together on the trail is standard practice when traveling with more than one animal. When tying them together, ensure the ropes are long enough to allow the animals to walk comfortably, and short enough to prevent tripping.

THE PEOPLE

Appoint a driver to lead the group and take control of navigation. You will need sufficient people to control and care for the animals.

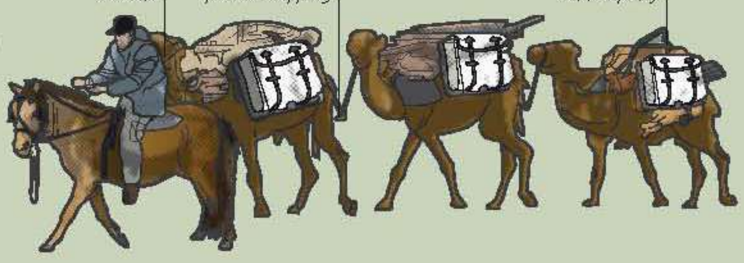

Driver leads the trail

Ropes need to be short enough to prevent tripping

Divide the loads equally

USING PACK ANIMALS TO SURVIVE

In an emergency situation, your priorities change significantly when using pack animals. Unpack the panniers and remove any heavy items not essential for survival. If necessary, you can climb on and ride to safety or, in a worst-case scenario, your animal can be used for food.

CARRYING INJURED PEOPLE

If a member of your group is injured, build a makeshift stretcher (see pp. 278–79) and use the animal to drag him or her to safety.

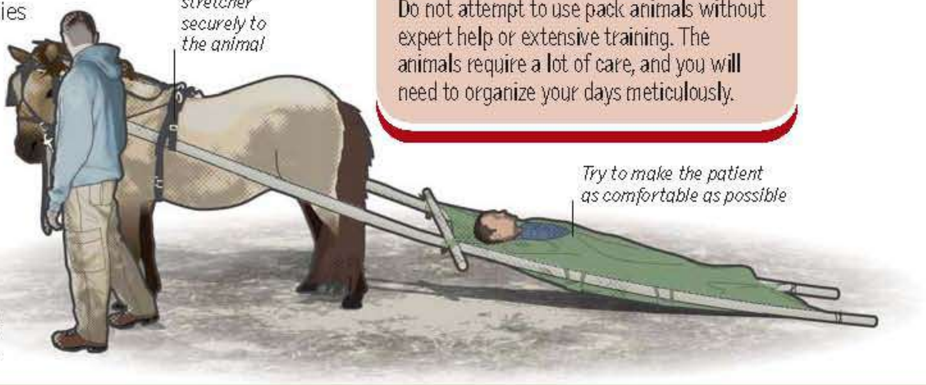

Strap the stretcher securely to the animal

Try to make the patient as comfortable as possible

WARNING!
Do not attempt to use pack animals without expert help or extensive training. The animals require a lot of care, and you will need to organize your days meticulously.

CHOOSING YOUR ANIMAL

The availability of pack animals depends on the region you are traveling in. Within any region, your choice of animal should take into account the load you are expecting them to carry, the distance you require them to travel, and the terrain you plan to cross. However, remember that individual animals' abilities can vary depending on their age and size.

ANIMAL	REGION	AVERAGE LOAD	ADVANTAGES	DISADVANTAGES
Horse	Worldwide	▪ 175–240 lb (80–110 kg) ▪ 20 percent of body weight	▪ Easy to train, with a good temperament ▪ Strong ▪ Can tolerate heat ▪ Can manage steep terrain	▪ Require grooming ▪ Can stray from camp if not securely tethered
Mule	Eurasia, Americas	▪ 120–180 lb (55–80 kg) ▪ 20 percent of body weight	▪ Will stick by a mare closely—this can be used to keep them moving and to keep them within the camp at night ▪ Very hardy and can manage steep terrain	▪ Young mules can be easily startled, so good training is essential ▪ Stubborn
Husky	Arctic	▪ 85lb (40kg) ▪ 7 dogs can pull 600lb (270 kg)	▪ Hardy—can cope well with snow and cold conditions ▪ Can travel fast	▪ Need a lot of fresh meat, which must be carried ▪ Prone to fight among themselves
Camel	Central Asia, North Africa, Australia	▪ 198–308 lb (90–140 kg) ▪ 30 percent of body weight	▪ Versatile on different terrains ▪ Can drink up to a quarter of its body weight and then go without water for several days	▪ Strong-willed and difficult to control ▪ Violent—can spit and bite
Llama	Andes	▪ 77–123 lb (35–55 kg) ▪ 25–30 percent of body weight	▪ Minimal environmental impact ▪ Travels well at high altitudes on difficult and steep terrain	▪ Makes horses and mules nervous ▪ Can be difficult to control if poorly trained
Elephant (Indian)	South Asia	▪ 1,650–2,750 lb (750–1,250 kg) ▪ 25 percent of body weight	▪ Can carry very heavy loads ▪ Can manage steep terrain	▪ Slow-moving ▪ Requires large amounts of food and water ▪ Takes time to train
Ox	Eurasia, Americas	▪ 300–450 lb (135–205 kg) ▪ 30 percent of body weight	▪ Hardworking ▪ Very strong ▪ Surefooted and can manage steep terrain	▪ Slow-moving ▪ Stubborn

FOUR-WHEEL DRIVING

A FOUR-WHEEL-DRIVE (4WD) vehicle can cover large areas of difficult terrain that would be inaccessible in a 2WD vehicle. With an experienced driver, most 4WD vehicles can tackle deep mud, water, snow, ice, and sand.

CHOOSING YOUR VEHICLE

When choosing your vehicle it's important to consider what you'll be using it for. Large vehicles have more internal space than smaller ones, but may find it harder to negotiate very rough ground without getting stuck. Similarly, very powerful vehicles can cross almost any terrain but use a lot of fuel, so may not be a viable option for long-term expeditions.

EXPEDITION CHECKLIST

- Check your map and inform others of your route and intended timeframe.
- Ensure the vehicle is fit for the journey, with all necessary spares. Check the fuel, oil, water, brakes, and hydraulic fluids, and check that there are no leaks.
- Inspect the tire treads, wheel nuts, lights, and steering arms.
- Always carry spare water, wheel(s), and fuel, as well as survival equipment, sand ladders, a 12v heavy-duty tire inflator, winching strap, and first aid kit.

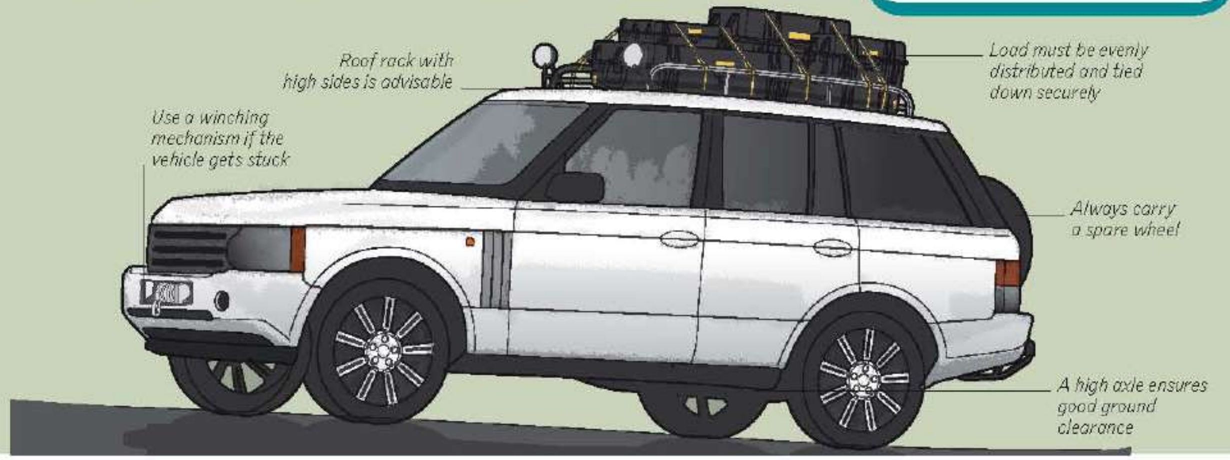

Roof rack with high sides is advisable

Use a winching mechanism if the vehicle gets stuck

Load must be evenly distributed and tied down securely

Always carry a spare wheel

A high axle ensures good ground clearance

GENERAL DRIVING TECHNIQUES

Four-wheel-drive vehicles enable you to manually switch between two-wheel-drive, for driving on roads, and four-wheel-drive, for low-traction conditions, such as on soft ground. The difference is the number of wheels powered by the engine at any one time.

CHOOSING A ROUTE

When unsure about the conditions immediately ahead, walk the route first, checking for potential problems and obstacles. Where necessary, mark the route you have walked and follow these markers when you drive. Ask yourself whether you really need to go that way and, if so, which route is best to avoid getting stuck. And, if you do get stuck, what are your options for self-recovery or escape?

WHEN TO ENGAGE 4WD

Engaging 4WD uses up a lot of fuel and should not be used on hard roads because of the risk of damaging tires and gears. Always choose the four-wheel-drive option on difficult terrain. It allows you to travel in a very low gear and gives the vehicle a far superior traction. Just before starting to cross the rough ground, stop the vehicle and engage the four-wheel-drive.

WARNING!

Don't hook your thumbs around the steering wheel when driving over rough terrain. Hitting a rut can jerk the wheel and break them.

DRIVING AS A TEAM

Driving off-road over rough terrain is both mentally and physically tiring. Take regular breaks and share the driving if in a group. You should always have at least two people in the vehicle:

- A driver—who takes responsibility for powering and steering the vehicle.
- A spotter—who takes responsibility for navigation and helps the driver pick a good route across the terrain. This may involve getting out of the vehicle and guiding the driver across a difficult section of ground.

DRIVING OVER DIFFICULT TERRAIN

A 4WD vehicle handles difficult terrain well because all four wheels can be powered by the engine at any one time (gas engines are generally more powerful, but diesel engines last longer and work well at low speeds). Using a few simple techniques, you will find that you are able to drive over terrain—and to places—that would otherwise be inaccessible. Always remember that the principle behind 4WD is to reduce the chances of you getting stuck—not to allow you to go further until you do get stuck.

MAXIMUM TRACTION

Your vehicle's traction on soft surfaces can be improved by lowering your tire pressure slightly. As a guide, place a brick ½in (1 cm) away from the side of a rear tire on flat ground, deflate the tire until it touches the brick, measure this pressure, and apply to all four tires. Make sure you can reinflate them.

DRIVING ON SAND

In soft sand the tires tend to move the sand from the front of the tire to the rear of the tire. If forward movement is halted for even a few seconds the wheel can dig itself into a hole. To prevent this, continually steer from side to side, so that the tire steers out of its own ruts. Avoid rapid changes in speed.

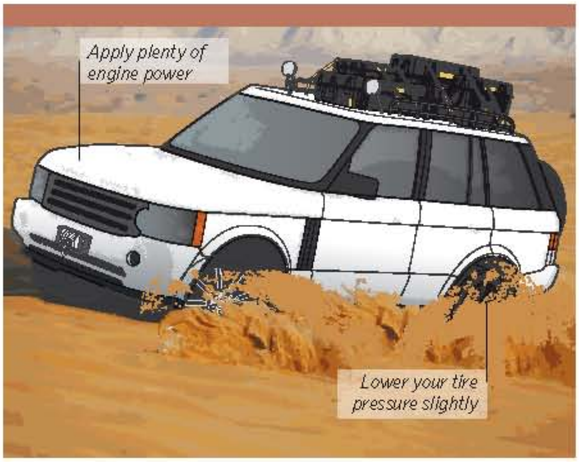
Apply plenty of engine power
Lower your tire pressure slightly

MUDDY TRACKS

Driving in mud requires concentration and the ability to adapt your driving to suit the conditions. In deep mud, use wide tires or lower the tire pressure slightly; however, if there is a hard surface below the mud this will make things worse. Steering outside existing track ruts is most likely to ensure the best traction.

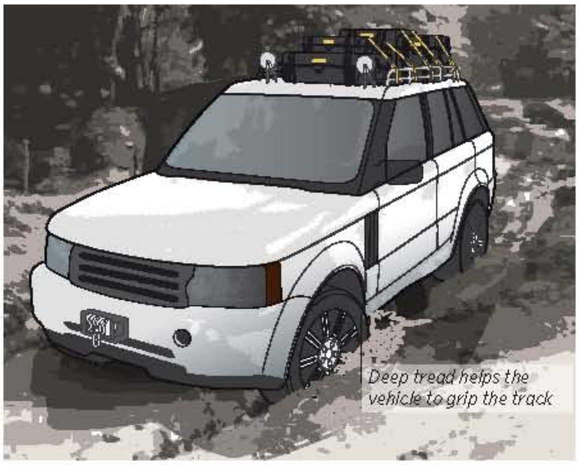
Deep tread helps the vehicle to grip the track

DRIVING ON SNOW AND ICE

Snow and ice require very smooth driving methods. Apply gradual pressure to the accelerator and brakes to avoid wheel spin, use low gears, especially when traveling downhill, and avoid changing gears unnecessarily. Using snow chains will increase traction, and you should practice fitting them before you need them.

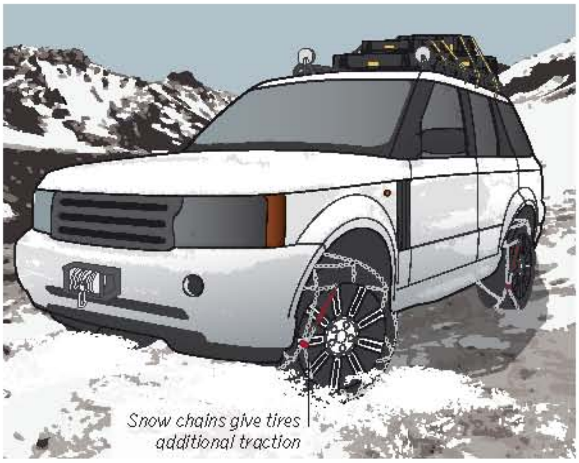
Snow chains give tires additional traction

CROSSING WATER

Always walk your route before driving across water. If the water seems too deep or the current too fast, don't attempt the crossing. It's important to drive at the correct speed—too fast will send water everywhere, but driving too slowly may flood the engine bay.

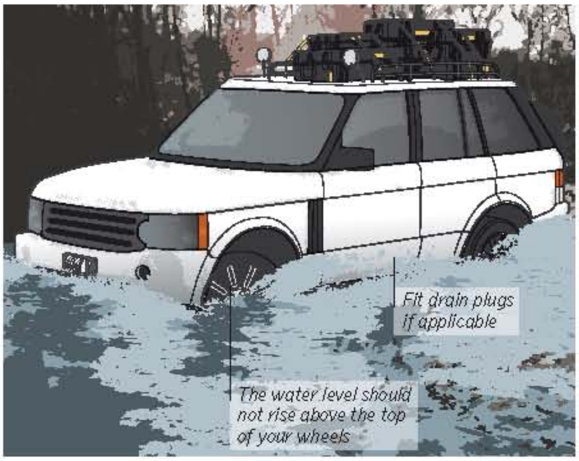
Fit drain plugs if applicable
The water level should not rise above the top of your wheels

RECOVERY FROM SOFT GROUND

Although using the correct techniques will certainly improve your chances of crossing soft ground successfully, it's important to know what to do if you do get stuck. Ideally you should never set out on a driving expedition with fewer than two vehicles. A second vehicle can be used for winching, dealing with breakdowns, or driving for help.

EQUIPMENT LIST

- Winching mechanism and strap
- Cable or tow rope
- Spare tire (inflated)
- Shovel and saw
- Sleeping bag or blanket
- Branches or stakes
- Sand ladders or mats

BASIC RECOVERY TECHNIQUES

When you get stuck, it's tempting to keep trying to aggressively free your vehicle. However, churning up what solid ground is left—and digging yourself into deeper ruts—will only make things more difficult in the long run. Stop, evaluate your options (reversing or pushing, digging, using branches, and winching), and calmly decide how best to achieve your aim. Don't act in haste—an ill-thought-out plan could leave you in more trouble than you were in originally.

REVERSING OR PUSHING

If you can't free the vehicle using four-wheel drive, try alternating between reversing and driving forward in first gear.

- If this has no effect, ask the passengers to get out and push while you drive forward.
- If you are making the situation worse, stop immediately and try an alternative solution.

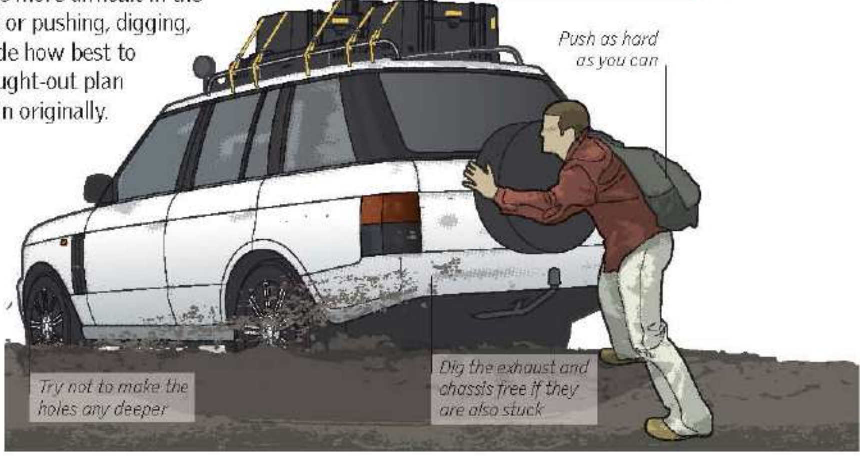

Push as hard as you can

Try not to make the holes any deeper

Dig the exhaust and chassis free if they are also stuck

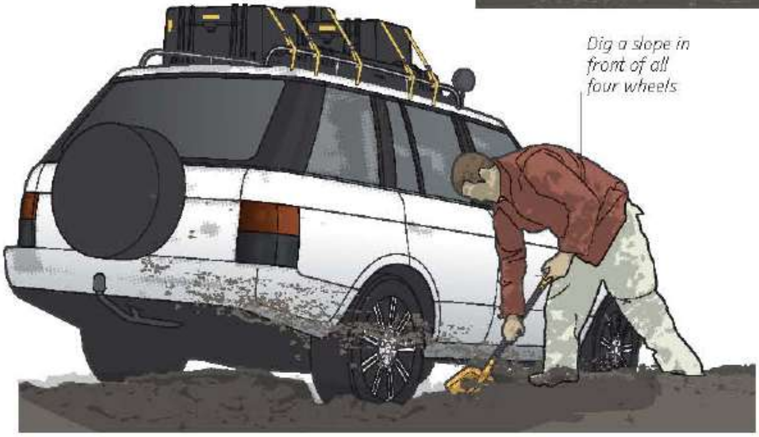

Dig a slope in front of all four wheels

DIGGING

If reversing and pushing don't work, the next step toward freeing your vehicle is to dig down in front of the wheels in an attempt to create a slope that you can then drive up.

- Dig out the sand in front of each tire to create an upward slope
- Drive very gently up the slope. Avoid revving the engine as this can cause the wheels to spin and lose their grip on the ground.

USING BRANCHES

If the vehicle still won't move, place branches, wooden planks, sand ladders, or blankets—in fact, anything that will increase your traction—in front of the wheels. The idea is to give the tires something to grip, and it should get you moving fairly quickly.

- Without revving the engine too much, gently ease the vehicle forward onto the branches or other material.
- Maintain a slow, steady speed and continue until you are back on firmer ground.
- Once clear, remember to stop and pick up your equipment, and remove any obstacles from the track.

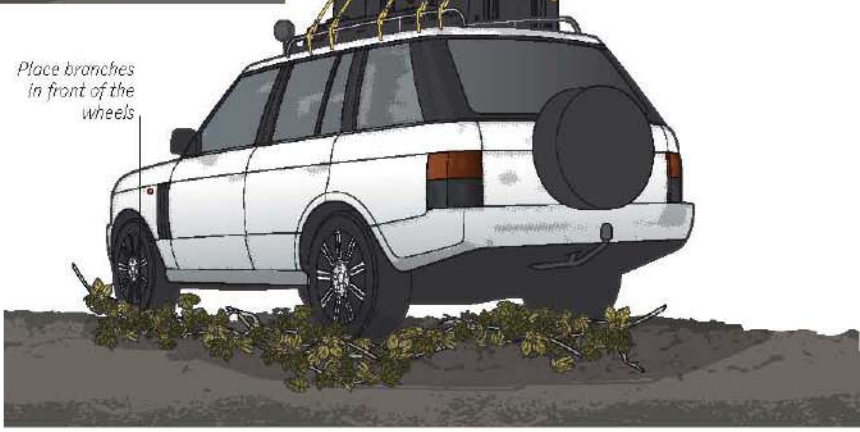

Place branches in front of the wheels

WINCHING

If none of the basic techniques work, and your vehicle is still stuck, it's time to consider winching. Use a cable attached to an electric-powered winching mechanism to pull the vehicle out of the hole via a strong anchor point. You can winch to another vehicle if you're traveling in convoy, although you do run the risk of both vehicles getting stuck. When winching, take time to assess your options—using a natural anchor point is the easiest method and should be your first approach.

WARNING!
When winching, don't stand within range of the cable in case it snaps under the strain. A breaking cable can cut a tree—or a person—in half. Place a blanket or sleeping bag over it to absorb some of the shock and reduce whiplash.

NATURAL ANCHOR POINTS

Trees, rocks, roots, or deadfalls can all be used as anchor points. When using a tree, always place the strap or cable near the ground, and use a winch strap where possible to avoid damaging the tree. If the tree looks as if it may not be able to provide the required support, tie the tree to others in the vicinity. If you intend to use rocks, ensure they are large enough and firmly embedded in the ground.

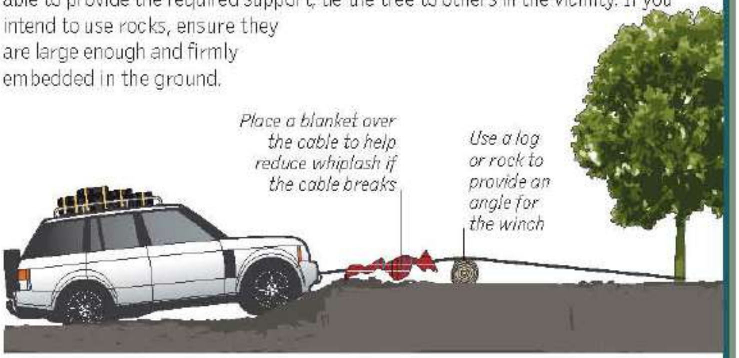
Place a blanket over the cable to help reduce whiplash if the cable breaks
Use a log or rock to provide an angle for the winch

BURIED ANCHOR POINTS

If there are no natural anchor points, you can construct an improvised anchor by burying objects such as logs or a spare wheel. Dig a hole in the ground that's at least 3 ft (1 m) deep, attach your winch cable to your improvised anchor, then bury the anchor in the hole. Refill the hole to secure it. If you're using a tire, use the tire lever behind your spare wheel as an attachment point for the cable.

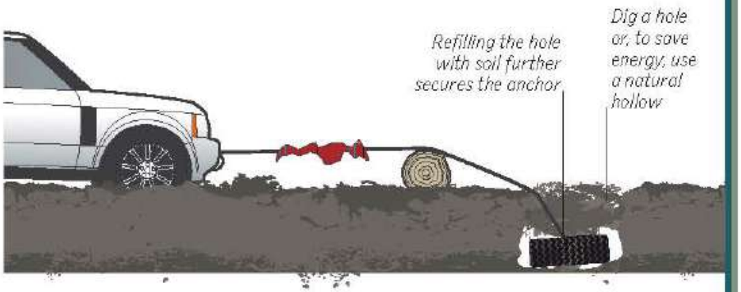
Refilling the hole with soil further secures the anchor
Dig a hole or, to save energy, use a natural hollow

WINCHING TO A STAKE

A series of long stakes in the ground can also provide an improvised anchor point. You will need a long, sturdy main stake—to which the winch cable is attached—and a series of supporting stakes, lashed together for additional strength. Push the stakes securely into the ground at a slight angle. Be careful not to stand near the stakes when winching, in case they come loose.

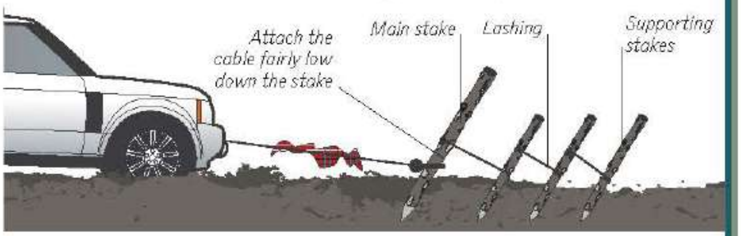
Attach the cable fairly low down the stake
Main stake
Lashing
Supporting stakes

EMERGENCY ACTION

Knowing what to do in certain emergency situations will help you stay calm if the worst happens.

FRONT SKIDS

The term "skid" covers any kind of slide in which the wheels fail to grip the road. In a front skid—also called an "understeer"—the front wheels fail to turn when you steer so the vehicle continues to travel forward.

- Steer in the direction of the skid but be careful not to oversteer. If you're on ice and skidding straight, step on the clutch or shift to neutral.
- Once you have gained control, correct your course and continue driving. Consult your driver's manual if you have anti-lock brakes.

REAR SKIDS

In a rear skid, the back wheels lose control and the vehicle rotates more than usual and can spin.

- Steer in the direction of the skid but face toward where you want to travel.
- Once facing in the right direction, bring the steering wheel back to the center and continue to steer, regardless of what the rear is doing, until you have gained total control.

BRAKE FAILURE

Before setting out on a drive, always check to see whether there is brake fluid leaking from the vehicle. If so, don't travel. If the brakes fail when on the move, do the following:

- Without turning off the engine, shift down the gears to reduce speed while maintaining control.
- Once the vehicle reaches a speed of less than 25 mph (40 kph) apply the handbrake, keeping a firm grip on the steering wheel.

ACCELERATOR MALFUNCTION

If the accelerator sticks and your engine will not slow down, you can decelerate by putting the vehicle into neutral and applying the brakes.

- If it's safe to do so, switch off the engine, but you'll lose functions like power steering and lights.
- Steer safely to a stop. If possible, avoid using the handbrake as this may cause a skid—although it may be necessary if you need to avoid a collision.

KAYAKING AND CANOEING

TRAVELING IN A KAYAK OR CANOE is a good way of negotiating river systems and open water. With your equipment safely stowed in your boat, expeditions can range from short day trips to year-long adventures.

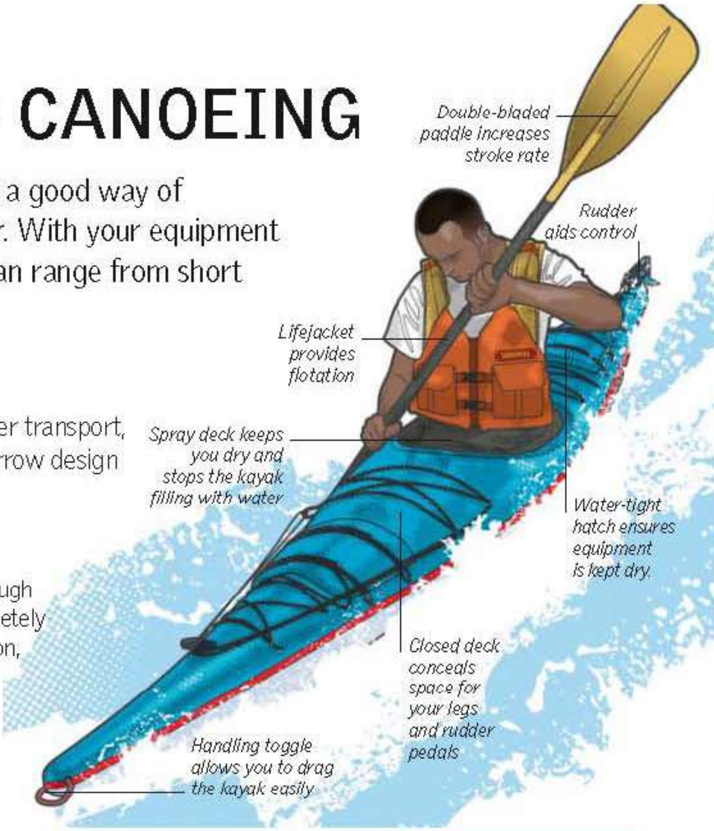
Double-bladed paddle increases stroke rate
Rudder aids control
Lifejacket provides flotation
Spray deck keeps you dry and stops the kayak filling with water
Water-tight hatch ensures equipment is kept dry.
Closed deck conceals space for your legs and rudder pedals
Handling toggle allows you to drag the kayak easily

KAYAKING

Kayaks are an extremely efficient method of water transport, especially on open sea, rivers, and lakes. Their narrow design and light frame make them easy to maneuver.

CAPSIZING IN A KAYAK

Capsizing is an almost inevitable part of kayaking in rough water. The "wet exit"—slipping out of the kayak completely while still holding on to it—is sometimes the only option, but wherever possible use an eskimo roll instead.

ESKIMO ROLL

Mastering the eskimo roll enables you to avoid getting your legs wet and having to climb back in your kayak. With practice, it should become one fluid movement.

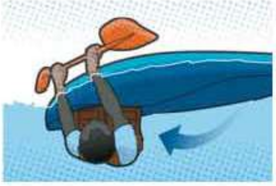

1 From the upside-down position, twist your body to the side of the kayak.

- Gripping your paddle, firmly push your arms out of the water, then rest them on the side of the kayak.

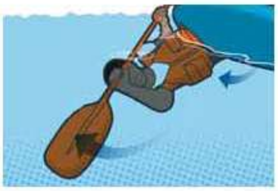

2 With your head near the surface, sweep your body and paddle away from the side of the kayak.

- At the same time, flick your hip away to start the rotation of the kayak.

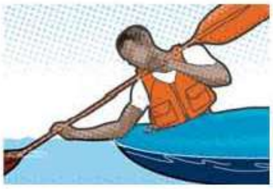

3 Continue the hip flick until the kayak returns to an upright position in the water.

- Straighten your back so you are sitting upright again, and continue to paddle your course.

WARNING!

- Always wear a lifejacket. Empty water bottles in your pockets can be used as makeshift flotation devices in an emergency.
- In a survival situation, never leave your kayak or canoe unless it is life-threatening not to do so. A kayak is highly visible, especially from the air, and is far more likely to be spotted first by rescuers.

BASIC PADDLING SKILLS

Kayak paddles have two blades that sweep alternately through the water, propelling the kayak forward. To steer, paddle on the opposite side to the direction in which you want to go. Some kayaks have a foot-operated rudder.

CATCH

Place your lead blade firmly in the water so that it enters near to your feet.

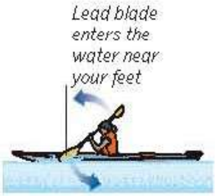
Lead blade enters the water near your feet

BRACE

Rotate your torso and lead blade in preparation for the propulsive stroke.

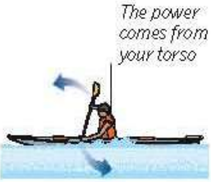
The power comes from your torso

STROKE

Sweep your lead blade firmly through the water to propel the kayak forward.

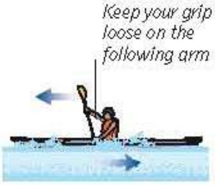
Keep your grip loose on the following arm

TRANSITION

As the blade exits the water, repeat the catch on the other side and continue paddling.

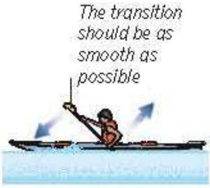
The transition should be as smooth as possible

CANOEING

As with kayaks, canoes can hold one or two people with equipment, and are powered by the body from a seated position. They are generally wider than kayaks, however, and their decks are open. Canoe paddles have one blade as opposed to two.

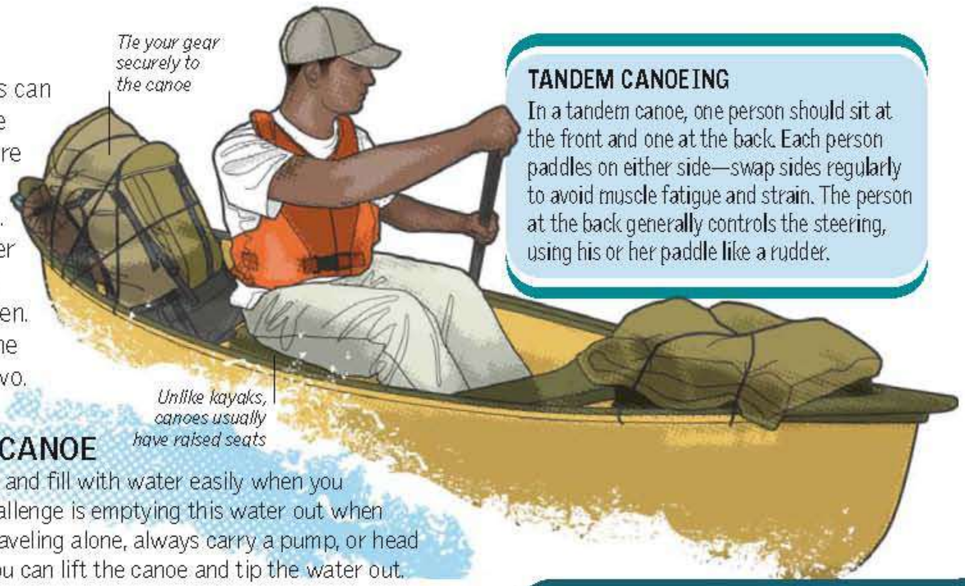

Tie your gear securely to the canoe

Unlike kayaks, canoes usually have raised seats

TANDEM CANOEING

In a tandem canoe, one person should sit at the front and one at the back. Each person paddles on either side—swap sides regularly to avoid muscle fatigue and strain. The person at the back generally controls the steering, using his or her paddle like a rudder.

CAPSIZING IN A CANOE

Because canoes are open and fill with water easily when you capsize, your greatest challenge is emptying this water out when the canoe is righted. If traveling alone, always carry a pump, or head for shallower water so you can lift the canoe and tip the water out.

CANOE-OVER-CANOE RESCUE

If you are traveling with another canoe and are far from shallow water, the best option is the canoe-over-canoe rescue. The capsized canoe is rotated upside-down and pushed onto the rescue vessel, which empties the water out. The people on the rescue craft then rotate the canoe back to an upright position and return it to the water.

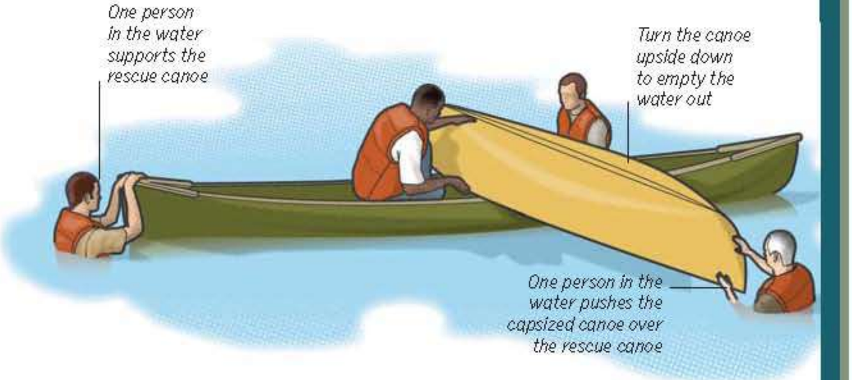

One person in the water supports the rescue canoe

Turn the canoe upside down to empty the water out

One person in the water pushes the capsized canoe over the rescue canoe

KAYAKING AND CANOEING TIPS

- Research your route and weather conditions thoroughly.
- Don't overload the boat, and ensure it's well balanced.
- Carry a high-volume pump so you can rescue yourself when alone.
- If watertight bulkheads are not fitted on your kayak, use airbags to add buoyancy and reduce the amount of water that can enter a compartment.
- Use waterproof dry bags to keep your kit dry.
- Clip valuables onto the boat or yourself. If they don't float, attach a float to them so they won't sink if they fall in the water.
- Carry waterproof material and bungee cord to repair blown hatches.
- Never enter a cave alone, or without head protection.
- Fit a leash to your paddle and attach it to the craft. If you then have to perform a rescue you can throw it in the water and keep both arms free.
- Chewing gum is great for the short-term repair of small holes in the craft, as is black masking tape and plumber's repair tape.

BASIC PADDLING SKILLS

When canoeing solo, use a "switch" style of paddling to ensure your canoe travels in a straight line—if the canoe starts to veer off course, switch sides.

CATCH

Sit in the middle of the canoe, ensuring that it's well balanced. Reach forward and thrust the blade into the water.

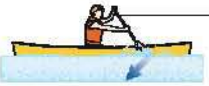

Paddle enters the water at the front of the canoe

DOWN STROKE

Drive the paddle swiftly and firmly down into the water so that it assumes a vertical position.

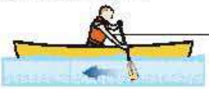

Use strong arm motions

PULL STROKE

Firmly pull the paddle blade backward through the water. This will propel the canoe forward.

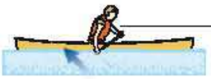

Paddle generates forward momentum

REPEAT STROKE

Lift the blade out of the water toward the rear of the canoe and return to the starting position to repeat the stroke.

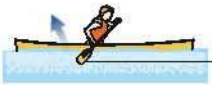

Paddle leaves the water at the rear

BUILDING A RAFT

IF VENTURING INTO the wilderness, it is useful to know how to build an improvised raft or flotation aid. You may find that a major water obstacle lies between you and rescue, or you may be in an area in which dense undergrowth makes cross-country travel difficult but river travel relatively easy. However, in a survival situation you are unlikely to have a lifejacket, so assess the risks carefully.

WARNING!
Apart from the brushwood raft (below), most improvised rafts will float half-submerged, so you'll be constantly sitting or kneeling in water. This could lead to hypothermia in certain conditions, so build an additional raised platform if necessary.

TOOLS AND MATERIALS

- Knife or small saw
- Poncho, shelter sheet, or tarpaulin
- Long sticks, brushwood, and thatch
- Cordage

MAKING A BRUSHWOOD RAFT

If you have a poncho, shelter sheet, or tarpaulin, you can construct a brushwood raft. This is a one-man raft, but if built correctly will keep you completely out of the water. Build it as close to the water as possible so you don't have to carry it far.

1 Construct your raft near a suitable launching place.

- Lay your poncho on the ground to determine the size of the finished raft. Make sure there's enough material to come up the sides and gather on top.

2 Peg out sticks to form an oval shape, the size and shape of your intended raft.

- Interlace brushwood and thatch tightly through the sticks to form the sides of the raft. The tighter you weave the brushwood, the stronger the body of the raft will be, and the better it will float. The height of the brushwood will also determine how high the raft sits in the water.

3 Form a sitting platform by pushing a selection of long poles or sticks through the sides of woven brushwood. Weave the sticks under and over each other as much as you can to create a stable structure.

- Using your knife or a small saw, trim the sticks to size so that they don't protrude from the sides.
- Remove the pegs from around the brushwood.

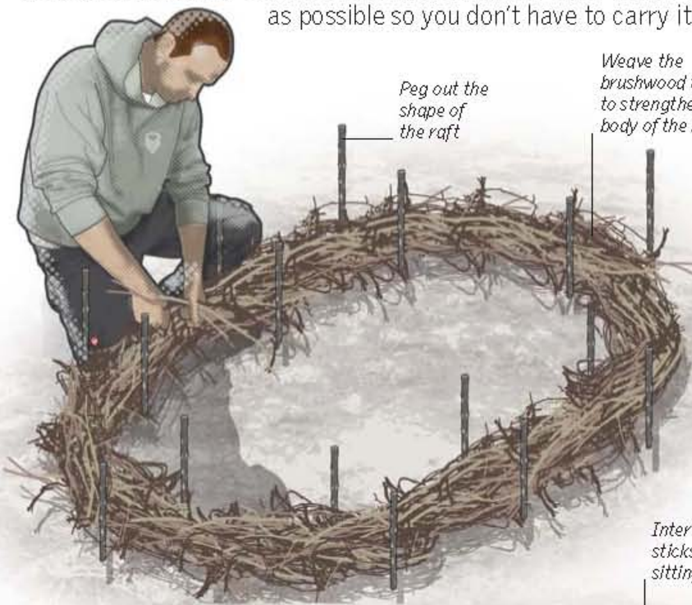

Peg out the shape of the raft

Weave the brushwood tightly to strengthen the body of the raft

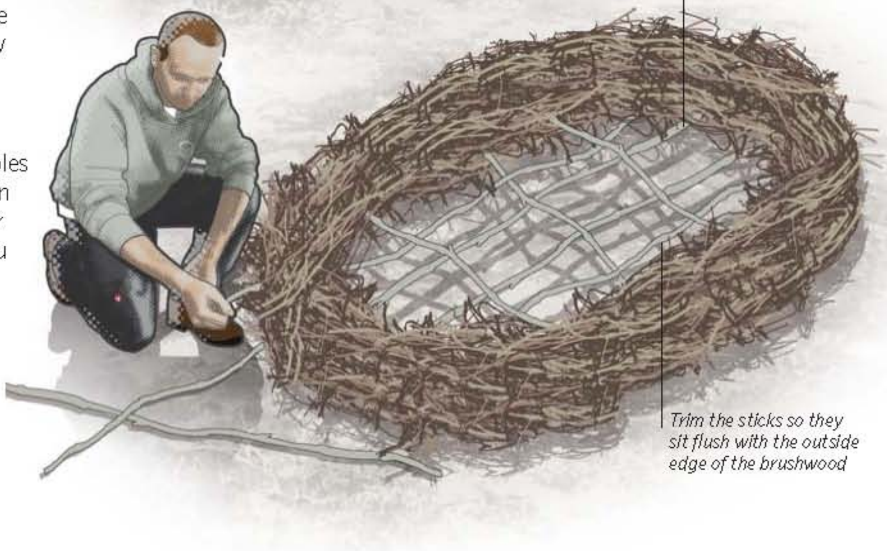

Interweave the sticks to create a sitting platform

Trim the sticks so they sit flush with the outside edge of the brushwood

MAKING A PADDLE

In most cases, you can use the current of the river for momentum rather than having to paddle, and your walking staff to maneuver the raft. However, where there is no current you will have to provide the propulsion yourself, and will need to improvise some form of paddle.

1 Find a suitable length of green wood to form the handle. This should be as wide as possible but still comfortable to grip.

- In the end of the wood, make a split long enough to form the paddle face.
- Collect smaller branches and insert one of them into the split. Lash into place.

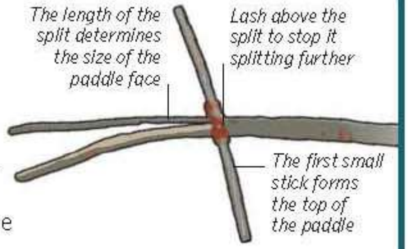

The length of the split determines the size of the paddle face

Lash above the split to stop it splitting further

The first small stick forms the top of the paddle

2 Continue to force smaller branches into the split, lashing each one as you progress until you have a sufficient paddle area.

- Securely lash the two open splits of the handle together.
- To make the paddle more rigid, lash the ends together.

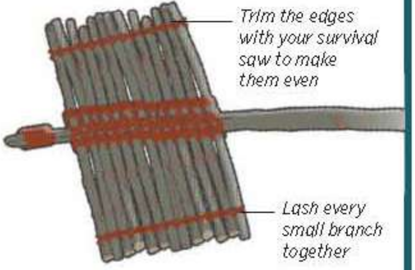

Trim the edges with your survival saw to make them even

Lash every small branch together

4 Slide your poncho underneath the raft to form a waterproof seal.

- Push the poncho hood through to the inside and tie its neck securely to ensure it's completely water tight.
- Pull the poncho up the sides of the raft and lash securely across the top.

5 Fill the space underneath the sitting platform with natural materials, such as additional brushwood, grass, or moss. You could also use any item that would aid buoyancy—such as empty plastic bottles or waterproof bags filled with air and tied off.

- Drag or push the raft into the water, checking to see whether it floats in shallower water before loading your equipment and climbing aboard.

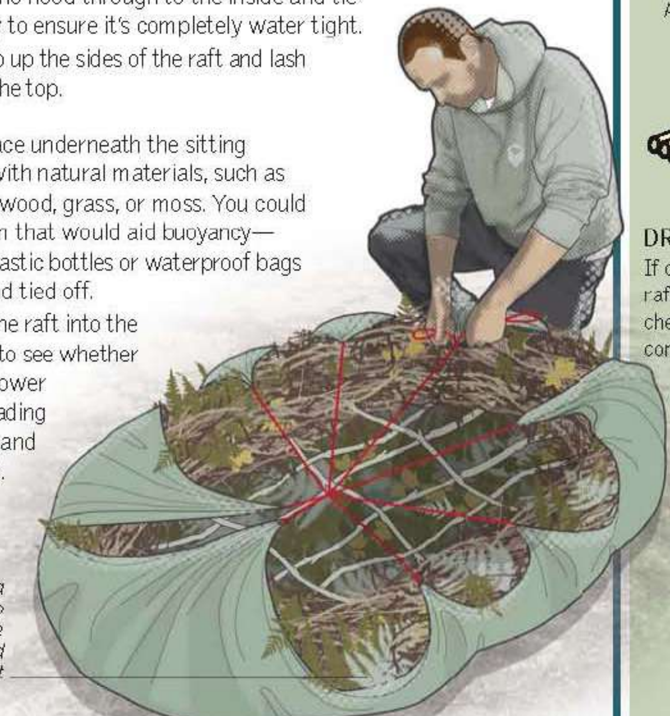

Add extra foliage to increase buoyancy and stability of raft

ALTERNATIVE RAFTS

Use whatever materials are available to you to make your raft. Look out for logs, bamboo, and discarded oil drums, as they are naturally buoyant materials that can be used to make an effective raft.

LOG RAFT

To make a log raft, use dry wood—ideally standing deadwood—as this will float higher in the water. Cut notches in the logs to allow the cross-pieces to sit snugly.

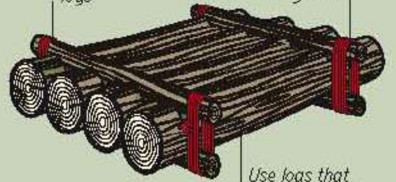

Cross-pieces tension the logs

Lash the cross-pieces together

Use logs that are roughly the same diameter

BAMBOO RAFT

Bamboo is made up of hollow compartmentalized sections, which means it is ideal for raft-making. Bamboo rafts are much lighter than log rafts.

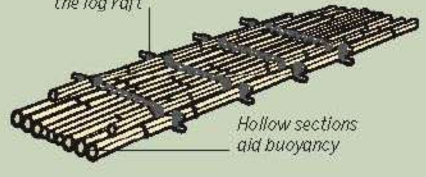

Use cross-pieces, as with the log raft

Hollow sections aid buoyancy

DRUM RAFT

If oil drums are available, they make ideal rafts. Always take care when handling chemical drums, as they may once have contained toxins.

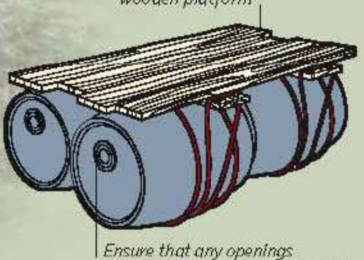

Tie the drums securely to a wooden platform

Ensure that any openings (missing filler caps, for example) sit above the water level

SWIMMING

WHEN THE WATER IS TOO DEEP to wade and you don't have the materials to build a raft, you may need to swim to negotiate a water obstacle. Before you get into the water, plan how you will get warm and dry when you get out. Making a float will help to keep clothing and equipment dry.

WARNING!
As with any survival activity involving water, try to avoid getting wet—find a way around the obstacle if possible, or find shallower water to wade across. The cold water increases your risk of hypothermia so ensure you have the necessary equipment for building a fire. Never enter water where there are likely to be dangers such as crocodiles or hippos.

SURVIVAL SWIMMING

Unless you need to cross rocky ground to enter the water, remove your boots and most of your clothing to reduce drag and keep them dry. Place them either in a waterproof survival bag or in an improvised poncho float (see opposite) and float them across.

SWIMMING ACROSS RIVERS

Choose a safe place to cross (see p. 88), and plan where you will exit the water.

- Take the current into account as you are likely to drift downstream slightly.
- Lower yourself in gently—never jump or dive into the water.
- If the water is cold, gently immerse yourself until your body has recovered from the initial shock.

THE CROSSING

Choosing a stroke you feel comfortable with (and that will expend the least energy), swim across your route with your float in one arm.

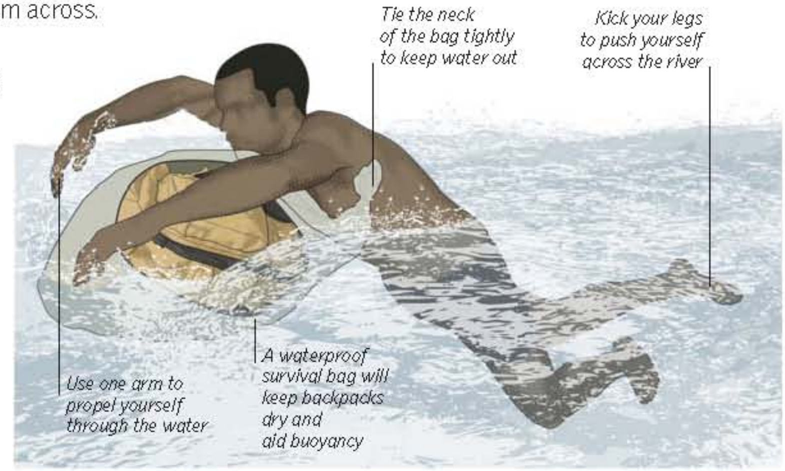
Tie the neck of the bag tightly to keep water out

Kick your legs to push yourself across the river

Use one arm to propel yourself through the water

A waterproof survival bag will keep backpacks dry and aid buoyancy

SWIMMING DOWN RIVERS

Avoid swimming up or down rivers—walking the route instead is far safer. However, if you accidentally fall in and find yourself traveling downstream in fast-moving water, there are ways of protecting yourself until you reach safety.

DEFENSIVE SWIMMING

The aim of defensive swimming is to adopt a position that keeps you as safe as possible and prevents your feet from becoming ensnared in rocks, which, in a strong current, could push you under the water. Assume the defensive swimming position until you reach shallower water, and can stand up and climb onto the bank.

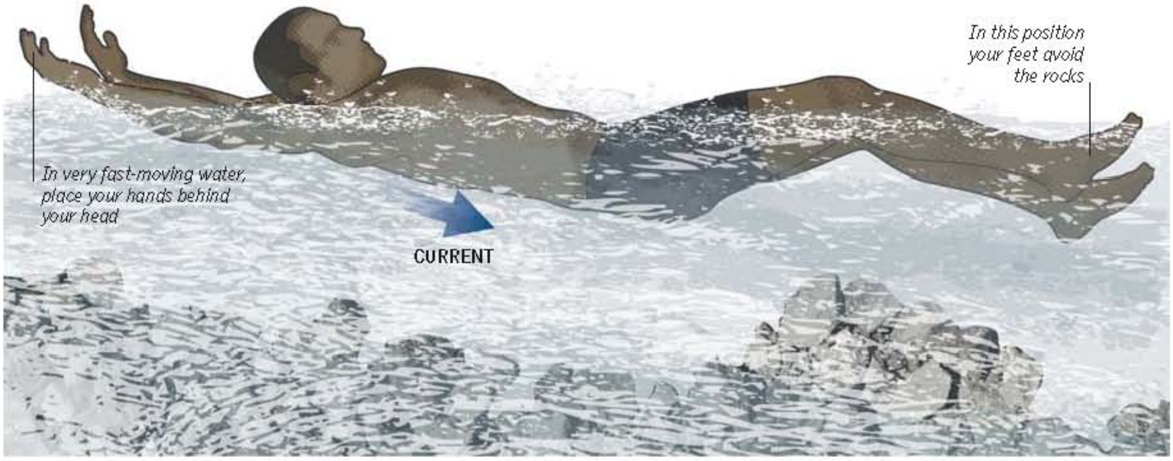
In this position your feet avoid the rocks

In very fast-moving water, place your hands behind your head

SWIMMING STROKES

It's important to know what the different swimming strokes are, although you're unlikely to stick to them strictly in a survival situation. If you're a confident swimmer, always choose the stroke that will use the least amount of energy, such as breaststroke. If you're less experienced you should avoid getting in the water unless in a life-or-death situation.

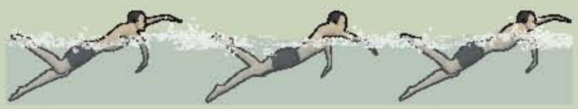

DOG PADDLE

An easy stroke suitable for less confident swimmers, this involves simply paddling forward with your arms and legs. When you paddle with your left arm, kick with your right leg and vice versa.

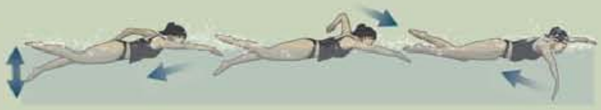

FRONT CRAWL

Also known as "freestyle," front crawl uses a lot of energy so is not the best stroke for a normal survival situation. However, you can build up a lot of speed, which may sometimes be required.

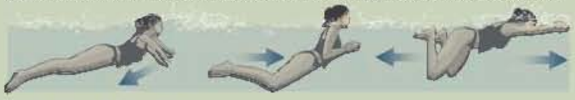

BREAST STROKE

This is perhaps the most common stroke for crossing rivers. Your arm and leg movements should be made simultaneously and smoothly. It is a good option if you're pulling a float.

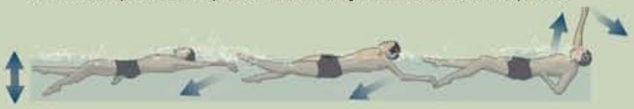

BACK STROKE

If you know the water is safe, this stroke can be a good choice, as it uses little energy. However, not being able to see where you're going can be a disadvantage, especially in rocky areas.

MAKING A PONCHO FLOAT

Constructing a poncho float will keep your belongings dry and provide limited flotation as you cross the river. If you don't have a poncho, use any large piece of waterproof material.

1 Push your poncho hood to the inner side, and tightly secure its neck with the drawstrings or cordage.

- Lay the poncho on the ground with the inner side facing upward.
- Place your gear on top.

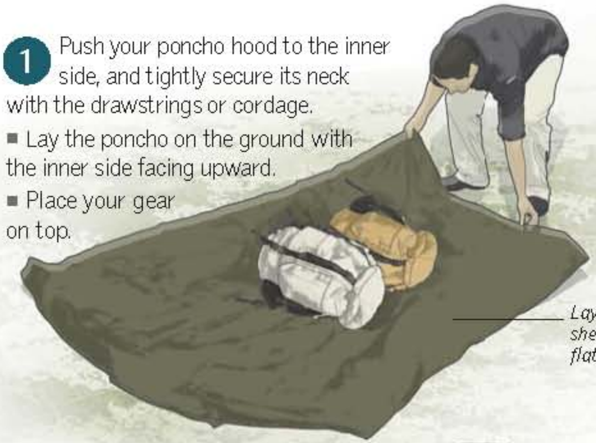

Lay the sheet on flat ground

2 Bring one side of the poncho up and over the equipment that is piled together in the middle.

- Repeat with the other three sides to create a rectangular parcel.
- Fold the corners, and ensure that they won't allow water to enter the float.

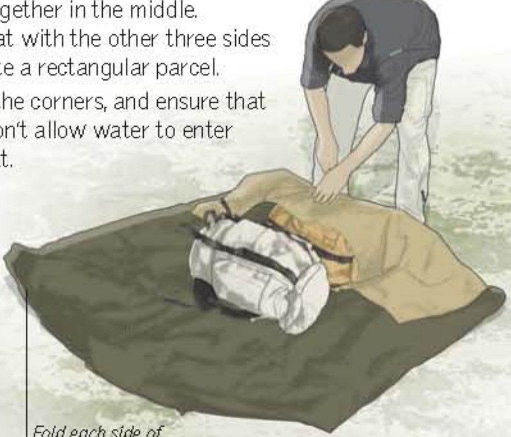

Fold each side of the poncho over your equipment

3 Wrap carefully to make it watertight.

- At this point, if you have a second poncho, repeat from step one, placing the float face down on top of the second sheet and wrapping again.
- If available, place brushwood inside the float for added buoyancy.

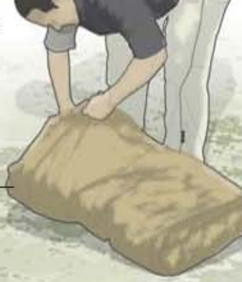

Tuck the corners in neatly

4 Tie ropes, bootlaces, vines, or improvised cordage around the float to secure it.

- As you enter the water, gently lower the float in with you. Pull it along as you cross the river.

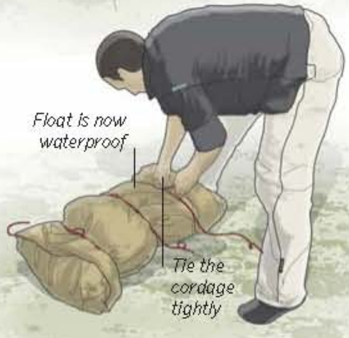

Float is now waterproof

Tie the cordage tightly

3

CAMP CRAFT

CAMP ESSENTIALS

Whether you intend to remain where you are and wait for rescue, move to a safer area and await rescue there, or walk out of the survival situation yourself, you're probably going to need to select a site on which to put up a shelter. This could be for just a single night or for a longer-term stay.

A sound understanding of what constitutes a good location will allow you to address the basic principles of survival. For example, a well-selected site will help keep you safe from danger while still allowing you to deploy your location aids effectively. It should also provide you with adequate materials for building a shelter and a fire, and offer an accessible supply of water, both for drinking and for hygiene.

A well-organized site will not only give you a sense of purpose and order, but will also provide a safe environment for yourself and your equipment. Designating an area for storing equipment and tools, for example, will help prevent vital items from being lost and will reduce the likelihood of you or members of your team being injured.

In this section YOU WILL DISCOVER...

- how to **make a shower** before you **take a shower...**
- why **keeping clean** is good for both **body and soul...**
- how to work up a **lather** with **wood ash...**
- when a **thatched screen** could **protect your modesty...**
- what turns a **hard candy** into a **firestarter...**
- the **difference** between **cattails** and **fire dogs...**

A fire is an integral part of any campsite. It can be used for warmth, for purifying water, for cooking, for signaling to potential rescuers, for protection against wild animals, and for providing light when darkness falls. It also provides a sense of security. The psychological effects of being able to start a fire in a survival situation should never be underestimated; neither should the psychological effects of *not* being able to start one. Even in a survival situation, a simple camp can give a sense of normality and "home."

You can produce a spark—and therefore light a fire—by using a device such as a firesteel or a ranger's flint and steel (see p. 127).

A FIRESTEEL COMPRISES two main parts: the material that will produce the spark (usually a rod made from ferrocerium or magnesium alloy) and a sharp striker device (usually a knife blade or short piece of hacksaw blade). When the striker is drawn over the rod, a spark is produced.

To control the spark created by these two moving parts, follow the method outlined below. It will allow you to direct the spark accurately, and reduce the chance of you knocking and scattering your tinder. This can be a problem if you hold the rod next to the tinder and strike down into it—especially if you are cold, wet, and tired and your hands are shaking.

1. Place the rod in the center of the tinder. Then position your striker onto the rod. Lock the hand holding the striker into position.

2. Pull the rod up and away from the tinder, drawing it against the striker. By drawing away from the tinder you avoid the danger of disrupting or scattering it.

3. To direct the sparks, alter the angle at which you pull the rod up and away.

> "The **psychological effects** of being able to **start a fire** should not be underestimated; neither should the effects of **not being able** to start one"

ORGANIZING YOUR SITE

WHERE YOU CHOOSE to set up your shelter depends on the environment, but always take into account the four principles of survival: protection, location, water, and food. First, make sure there are no obvious dangers, and that you're able to signal for rescue. If possible, select a location close to a water source.

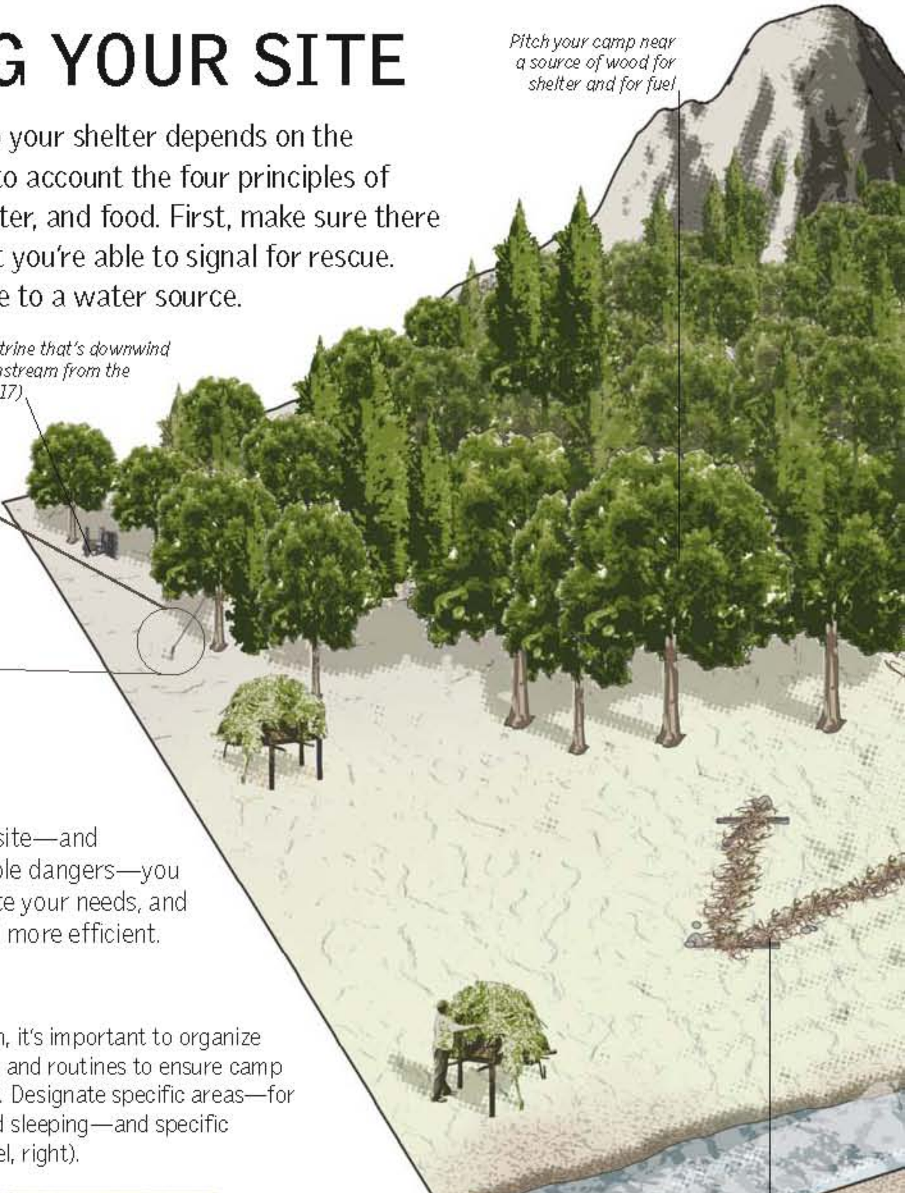

Pitch your camp near a source of wood for shelter and for fuel

Select a place for a latrine that's downwind of the camp and downstream from the water source (see p. 117)

Assign an open area for location aids, such as a large "V", and as a potential landing site for a rescue helicopter (see pp. 236–41)

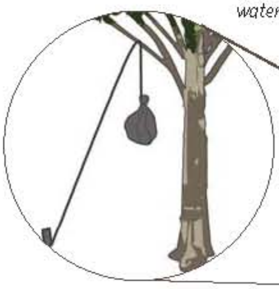

If predators are in your area, hang unused food in a bag 10 ft (3 m) off the ground and 3 ft (1 m) away from the trunk or branches of a tree

SITE FEATURES

Once you're happy with your chosen site—and confident it doesn't hold any forseeable dangers—you can organize the area to accommodate your needs, and to make your time in camp easier and more efficient.

CAMP ADMINISTRATION

Whether you're in a group or on your own, it's important to organize your site, and quickly establish disciplines and routines to ensure camp safety and to reduce the risk of accidents. Designate specific areas—for storing equipment, firewood, cooking, and sleeping—and specific routines for when you do things (see panel, right).

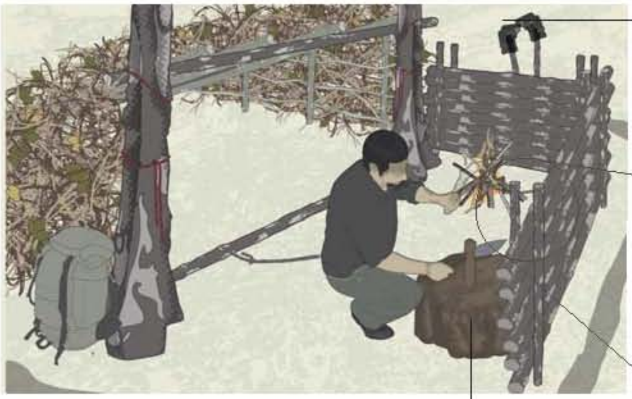

When you take off your boots, prop them upside down on a stick to ensure nothing crawls inside. Never put them too close to a fire to dry

Create a safe area for cutting wood. Use a tree stump as a platform for cutting and chopping

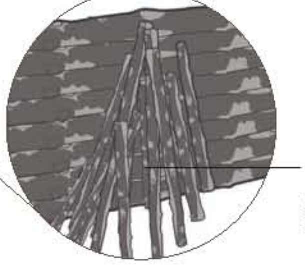

Collect and cut the fuel you need before it gets dark and keep it dry in an upright stack

ASSESSING YOUR ENVIRONMENT

You need to protect yourself from injury, threats, and risks, so have a good look around your campsite and assess it for any potential dangers, such as animals, unstable rocks or trees, and the likelihood of flooding.

ANIMALS

Look out for signs of animals, especially near water. If possible, pitch your camp against a rock face so it can only be approached from one direction. Keep a fire going all night. If you're in a group, organize a watch system. Keep things to hand that you can make a noise with to scare off prowling predators. Don't camp close to standing water where insects, such as mosquitoes, swarm.

WIND AND FLASH FLOODS

Position the entrance to your shelter at an angle to the wind. Gullies run the risk of flash floods or avalanches; inside river bends are prone to erosion and floods; and a river might burst its banks on an outside bend during a heavy downpour.

STANDING DEADWOOD

These are dead trees that haven't yet fallen. Heavy wind, or the weight of rain or snow, can make them fall. This is the best type of wood for kindling and fire fuel (see p. 121).

DEADFALLS

These are dangerous branches that have broken off a tree, but haven't yet fallen to the ground. Some trees, such as beech, ash, and yew, drop their branches without warning.

ROCKFALLS AND ICE FALLS

If you camp next to rocks, check for cracks and fissures. Fires below them can cause rockfalls. When it's cold, ice sheets can suddenly fall from the rocks.

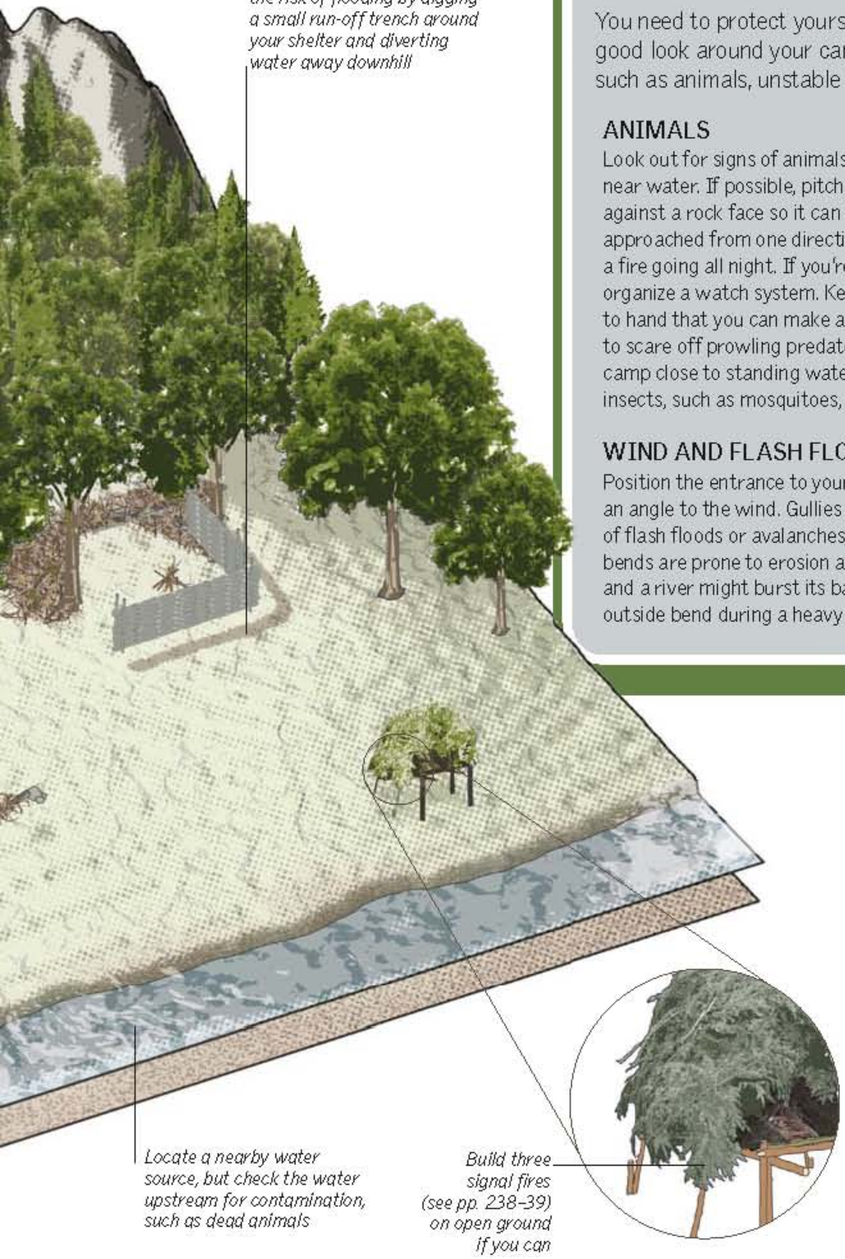

If heavy rain is likely, reduce the risk of flooding by digging a small run-off trench around your shelter and diverting water away downhill

Locate a nearby water source, but check the water upstream for contamination, such as dead animals

Build three signal fires (see pp. 238–39) on open ground if you can

WHERE NOT TO CAMP

- Don't set up your shelter on sloping, poorly drained ground.
- Don't pitch your camp too close to water because of the risk of flooding and the presence of insects and animals.
- Avoid camping near noisy water sources, such as waterfalls, because they mask any other sounds that could alert you to the presence of threats, such as wild animals, or even possible rescue, such as a helicopter or emergency whistles.

CAMP TIMETABLE

In most cases, when you arrive at an area in which you intend to remain for a while (known as "going static"), start planning your activities by working back from a cut-off time—usually when it gets dark.

THREE HOURS UNTIL DARK

You arrive in the area. Drink water, stow what is no longer needed (map, compass, and so on), and change from wet clothing, but keep one set of clothes dry at all times. Scout the area for the most suitable campsite.

THEN...

Start pitching your camp. Build your shelter, including bedding. Gather tinder, kindling, and fuel and make a safe firebase and reflector. Collect water and forage for food. Prepare all aids to location (for both night and day).

ONE HOUR UNTIL DARK

During twilight, finish organizing the site and see to your personal needs—wash, use the latrine, and check your equipment. If you're in a group, make sure everyone knows where the emergency equipment is and who has been designated each task (lighting the signal fire, shining the flashlight, blowing the whistle). Set up a watch system for the fire.

STAYING CLEAN

PERSONAL HYGIENE IN THE FIELD is an important element of protection. Keeping yourself clean and healthy helps to ensure that your body is working at its most efficient and reduces the risk of illness. How you feel physically also has a direct impact on how you feel psychologically. It's a state of mind: if you let your personal hygiene slip, it's just a matter of time before everything else starts to follow.

CLEAN CLOTHING

The condition of your clothing and equipment can affect your state of mind. If your clothing is dirty and unkempt, your attitude to survival will lack discipline. Try to keep your clothing clean and in good repair.

HYGIENE CHECKLIST

Personal hygiene is about keeping clean and healthy, so develop a daily routine that ensures you correctly use personal protection aids (tablets, insect repellents, sunblock), and safely handle food, water, and cooking and eating utensils. Keep your clothes clean, and attend to your bodily functions (see opposite).

HAIR

You don't need to shampoo your hair—let it make its own oils and minerals and establish a natural balance. Wash out incidental smells, such as fire smoke, with hot or cold water.

SCALP

Check your head for insects and bites every morning and night. Local knowledge will help—for example, look for ticks if you are in a wooded area populated by deer.

EYES

Wear sunglasses or a hat to protect your eyes from bright sunlight and snow glare. Rinse your eyes with water twice a day to protect against infections like conjunctivitis.

EARS

Carefully check your ears for foreign objects with a clean, wet finger, especially if you're sleeping rough on the ground.

TEETH AND GUMS

Use clean fingers to rub your teeth and gums, or make a toothpaste from baking soda, or a mouthwash from salt and water.

BODY

Every two days wash your armpits, crotch, hands, feet, and toes with running water to keep fungal infections at bay.

MAKING A SHOWER

Standing under a shower can work wonders for your spirit—it removes accumulated dirt and sweat and makes your body feel refreshed. An improvised shower is quick and easy to make from a metal or a plastic container.

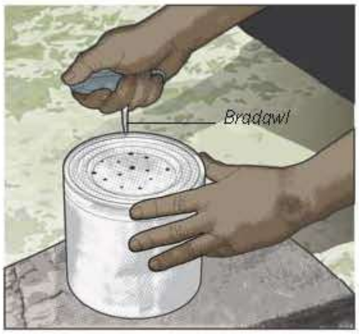

1 Turn the container upside down on a flat surface.

- Punch holes in the bottom with the bradawl on your penknife or the point of your knife.

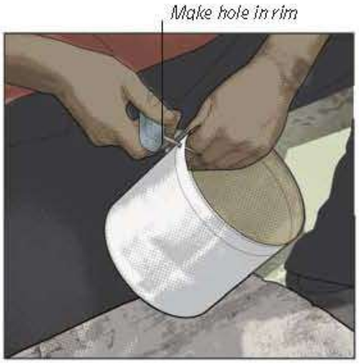

2 Make a hole about 1 in (2.5 cm) down from the rim of the container.

- Make a matching hole on the opposite side of the container.

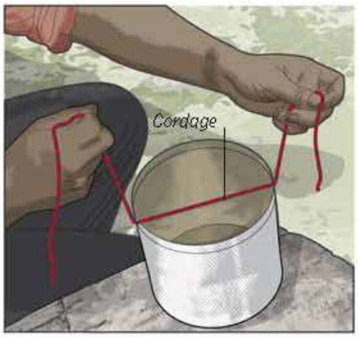

3 Smooth out any rough edges around the two holes.

- Thread a piece of cordage more than 2 ft (60 cm) long through the two holes.

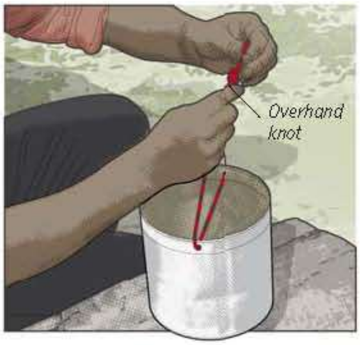

4 Pull the cord through until you have the same length—about 1 ft (30 cm)—on either side.

- Tie the two strands together with an overhand knot (see p. 143).

BODILY FUNCTIONS

In a genuine survival situation, you don't eat as much food as usual and your toilet functions—particularly with regards to solid waste—will reduce dramatically after a day or so. Nevertheless, it's important to maintain a routine, especially if you are in a group.

IN TRANSIT

If you're moving every day or so, it's not worth making a latrine, so just attend to your bodily functions as follows:

- **Urination:** choose a tree away from the water source and downwind of the camp.
- **Defecation:** dig a hole at the base of a tree, fill the hole afterward, and mark it with stones or two crossed sticks. Clean yourself (but don't use the hand you eat with) with toilet paper, leaves, or grass, then running water. Wash your hands and fingernails. Burn or bury used toilet paper.
- **Menstruation:** if you don't have tampons or sanitary napkins with you, use something cotton and washable, or even sphagnum moss. Burn or bury whatever you have used.

GOING STATIC

If you're going static by staying in one place for more than a few days, build a latrine downwind of your camp, and downhill and away from your water source. Dig a deep trench, either next to or between two trees, and make a seat from two poles. Cover the deposits with sand or soil to reduce the smell and keep flies away. When you leave, dismantle the latrine, fill the trench, and mark the area with stones or crossed sticks.

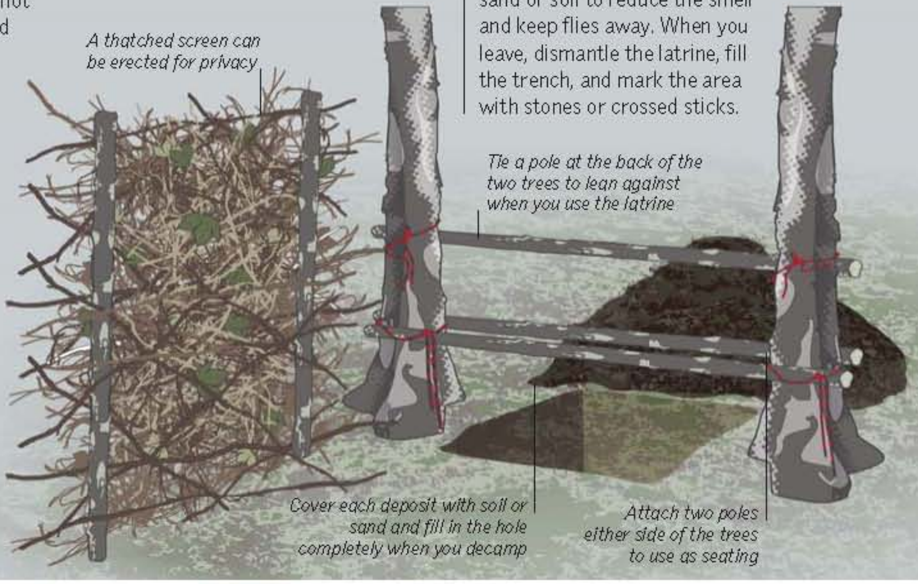

A thatched screen can be erected for privacy

Tie a pole at the back of the two trees to lean against when you use the latrine

Cover each deposit with soil or sand and fill in the hole completely when you decamp

Attach two poles either side of the trees to use as seating

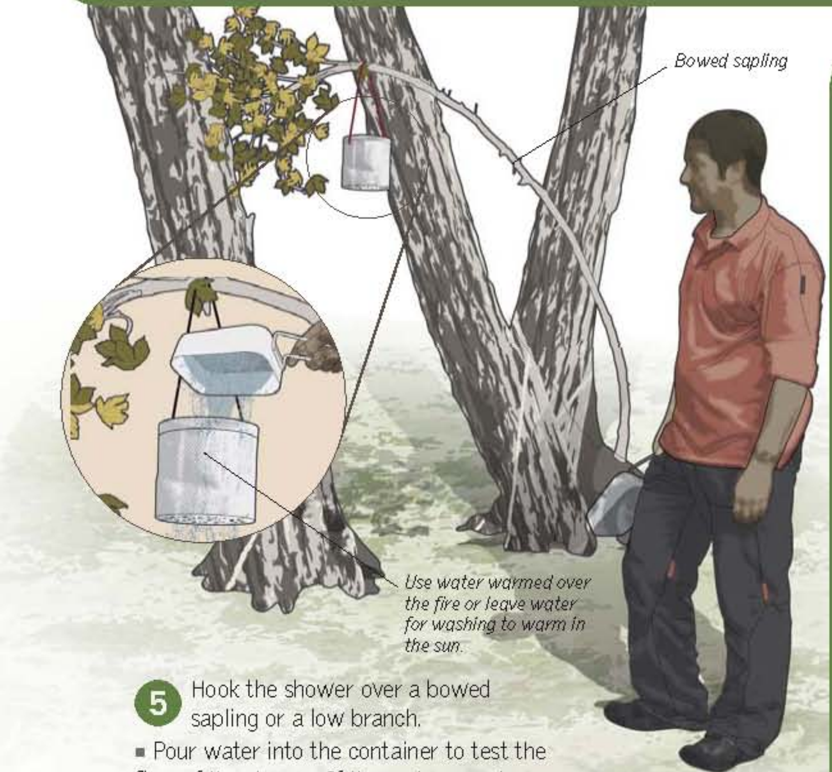

Bowed sapling

Use water warmed over the fire or leave water for washing to warm in the sun.

5 Hook the shower over a bowed sapling or a low branch.

- Pour water into the container to test the flow of the shower. If the water runs too quickly, put leaves in the bottom to slow it.

LATHERING UP

Camping soap is a concentrated antibacterial liquid soap that can be used without water. Decant a little into a small container, such as an old 35mm film container, and it will last for weeks. Alternatively, you can make soap from natural materials.

MAKING SOAP

You can make natural soap from various sources that contain a substance called saponin, which —when mixed with water—has a cleaning effect.

- **Birch leaves:** Select young leaves and place them in a container (even a plastic bag will do). Add some cold water, then add some hot or boiling water (whichever your container can handle without melting). Agitate the mixture—this allows the saponin in the leaves to dissolve into a natural soap.
- **Soapwort:** Agitate soapwort roots in water until they foam up. Let the light foam settle before using it to wash yourself and your clothes.
- **Horse chestnut:** Dip horse chestnut leaves in warm or hot water then remove. Squeeze the leaves in your hand to produce saponin.
- **Wood ash:** Mix wood ash with water. Don't use this method too often, as it can dry out your skin.

MAKING FIRE

THE ABILITY TO MAKE and maintain a fire can be a significant psychological factor between determining whether you do all you can to survive or just give up. Fire gives you a sense of "being," and, like a shelter, it can transform a clearing under a tree into a "home."

A PORTABLE KIT

It's important to gather all your materials before you start—not only your tinder, kindling, and fuel, but also your means of lighting a fire. This involves a fair amount of organization, so a portable kit that keeps it all in one place is much more convenient.

FIRE PRECAUTIONS

If you intend to have a fire in front of your shelter, bear in mind some basic rules:

- Don't build the fire too close to the shelter as it may get out of control or a spark may blow into the shelter and ignite it.
- Make sure you have fully extinguished your fire before you decamp and leave the site. Pour water (if you have enough) onto the fire and the immediate surrounding area, or use damp soil, sand, or earth.

COMPONENTS OF FIRE

Three crucial components—oxygen, heat, and fuel—need to be present in order for a fire to start and be sustained. Although you don't have any control over the resulting chemical reaction, known as combustion, you can develop the skill of lighting a fire. The key is to achieve the best balance between the components. If your fire isn't going well, get back to basics and ask yourself which one of the components is working against you.

OXYGEN

Although oxygen is vital for combustion to happen, it's all too easy to prevent it from getting to the flame by smothering the fire with too much wood. If your initial fire looks as though it's dying, try fanning it with your hand or a map to create a draft that feeds oxygen to the fire.

HEAT

Heat is essential for igniting the fuel. In most cases this heat can be generated either by a spark (such as flint and steel), by a chemical reaction and friction (such as matches or potassium permanganate), or by friction alone (such as a bow drill). (See pp. 126–27 and pp. 130–33.)

FUEL

Once the fire gets going, you need to have fuel to burn (see pp. 120–21). You should start with small, dry pieces of fuel that will catch the flame and generate enough heat to then burn increasingly large pieces of fuel.

MATCHLESS FIRESET TIN

Working on the principle of "Don't make things any harder than they need be," this matchless fireset, designed for the military, contains a one-stop solution to getting a fire going—regardless of the weather conditions.

INSIDE THE TIN

The small tin, which is waterproof when taped, contains everything you need to produce a fire: a sparking device, tinder, kindling, and fuel. Hexamine from the fuel blocks is scraped onto cotton wool, then a spark from the flint and steel ignites the cotton wool long enough for the fuel blocks to light.

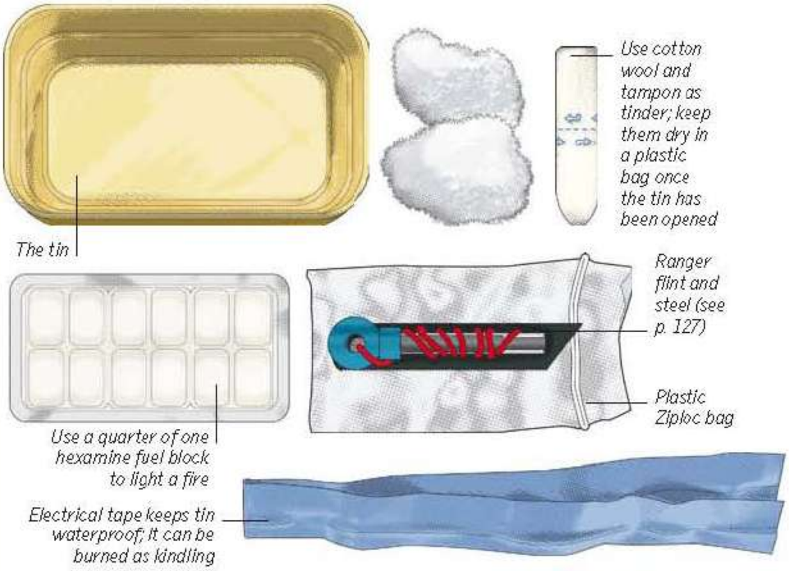
The tin

Use cotton wool and tampon as tinder; keep them dry in a plastic bag once the tin has been opened

Use a quarter of one hexamine fuel block to light a fire

Ranger flint and steel (see p. 127)

Plastic Ziploc bag

Electrical tape keeps tin waterproof; it can be burned as kindling

USING THE TIN

You can use the tin itself to boil water. Raise the tin 1 in (2.5 cm) off the ground with two sticks and light a quarter of a hexamine fuel block underneath. After a while, the water boils so you can have a potentially life-saving drink.

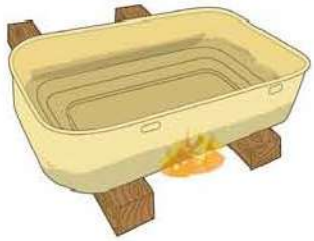

PREPARING BEFOREHAND

More than any other survival task, starting a fire needs good prior preparation. If you don't have the correct materials and in sufficient quantities, you will probably fail in your attempts to start a fire. Preparing the ground, your materials, and your equipment will usually make firelighting much easier and more likely to succeed.

CHOOSING YOUR GROUND

You need to be careful when selecting the place in which you intend to start your fire. Clear the ground before building your fire. Never light a fire directly on the ground and watch that your fire doesn't spread or burn out of control.

KEY POINTS TO REMEMBER

- Carry some form of firelighter with you at all times.
- Practice your skills before you need them—in different conditions and using different materials. You'll soon learn what works and what doesn't.
- Collect tinder as soon as you enter the forest. If it's wet, dry it in your pockets against your warm body.
- Collect everything you need to get a fire going, then multiply it by ten. If you're in a dire situation, you may have only one or two chances of getting a fire going before you either run out of matches or tinder, or you suffer from exhaustion or hypothermia.

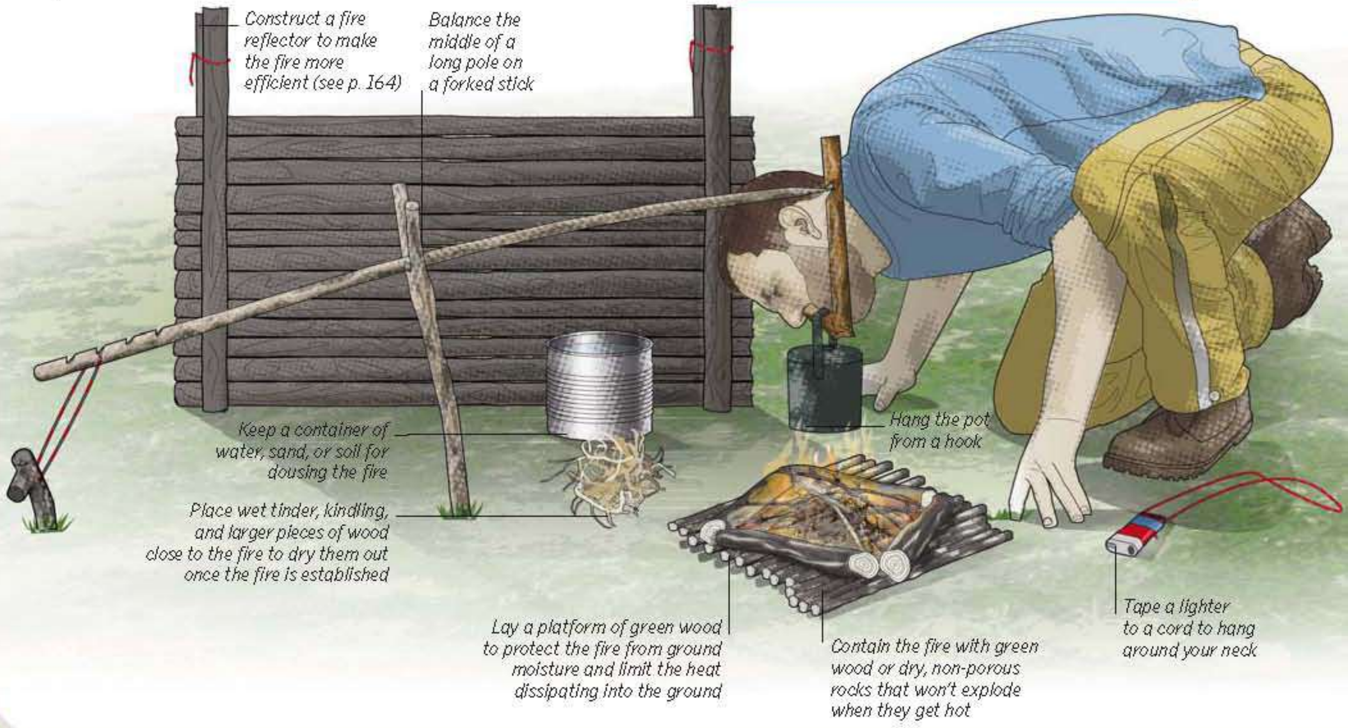
Construct a fire reflector to make the fire more efficient (see p. 164)
Balance the middle of a long pole on a forked stick
Keep a container of water, sand, or soil for dousing the fire
Place wet tinder, kindling, and larger pieces of wood close to the fire to dry them out once the fire is established
Hang the pot from a hook
Lay a platform of green wood to protect the fire from ground moisture and limit the heat dissipating into the ground
Contain the fire with green wood or dry, non-porous rocks that won't explode when they get hot
Tape a lighter to a cord to hang around your neck

CAMP FIRE DO'S

- Make sure you have enough wood close by to fuel your fire. Carrying wood to your fire uses energy, which may be at a premium in a survival situation.
- Rake the area around the spot where you intend to start the fire and brush away leaves or anything else that may ignite and start a forest fire.
- Check the ground for tree roots. Your fire could set an exposed root alight, or even one just under the surface. Once a root starts to burn underground, the heat works its way along the root and could start a forest fire.
- If you're trying to get rescued, choose a spot for the fire where passing vehicles and aircraft can see it.

CAMP FIRE DON'TS

- Don't use your hands as a rake to clear the ground, because you may get bitten by an insect or a snake. Use your feet or a branch instead.
- Don't build a fire next to old logs or fallen trees, as they may catch fire. They may not look alight but they can smolder for several days and a heavy breeze could subsequently fan the embers and start a blaze.
- Don't build your fire under an overhanging branch, or leaves, as the heat from the fire can soon dry them out and then set them alight.
- Don't position your fire so that the wind affects the way the fire burns, or blows smoke or flames into your shelter.

THE ELEMENTS OF FIRE

THE THREE MATERIAL ELEMENTS you need to build a fire are tinder, kindling, and fuel. They must be dry and collected in sufficient quantities. A well-made feather stick effectively provides tinder, kindling, and fuel on one piece of wood—and it needs only a spark to set it alight.

EMERGENCY TINDER

Carry tinder, such as a tampon or cotton wool balls, in a 35mm film canister or a Ziploc bag in your emergency kit. For each fire, use some cotton wool and a small piece of tampon. Cotton wool balls smeared with petroleum jelly make a flame last ten times longer.

MAKING A FEATHER STICK

Between four and six well-made feather sticks provide enough combustible material to get a fire going. Keep a few already made and packed away to use in an emergency. The best feather sticks are made from standing dead wood, but just about any type of dry wood will do. If you use small, dead branches that have snapped off a tree, remove the bark first.

① Choose the side of a stick with an even grain and no knots. Work on a hard, flat surface to stop the stick slipping.

- Lay the blade flat on the stick and run it all the way down without cutting into the wood—do this ten times. This gives you a feel for the wood and how your knife moves over it.

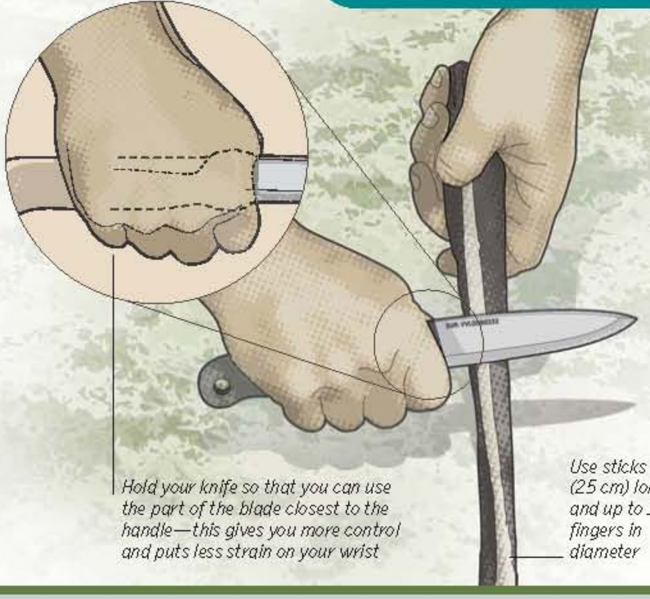

Hold your knife so that you can use the part of the blade closest to the handle—this gives you more control and puts less strain on your wrist

Use sticks 10 in (25 cm) long and up to 3 fingers in diameter

TINDER

The first element you use to make a fire is dry, combustible material called tinder. You may have tinder in your equipment (see pp. 122–23), or you may need to find natural or other man-made sources. The key to success is to experiment with what's around you before you actually need it. Make sure it's dry—leave it in the sun if it's damp. When you use some tinder, replace it at the earliest opportunity.

TYPES OF TINDER

Natural sources: Feather sticks (see above), shavings from the outside of a bamboo stem, bark shavings, fine wood dust, pine pitch, plant and animal down, powdered dung, and fire dog (a charred stick). Also, birch bark (silver birch bark can be lit even when it's wet), clematis, honeysuckle, cattail, dry grass, dead and dry moss, and some fungi. The best tinder for friction methods of firelighting (see pp. 130–33) is tree bark, dry grass, fungi, lichen, plant roots, fibers, and down.

Man-made sources: Cotton wool balls, tampons, char cloth, fire tin (see p. 123), lint, tissue paper, camera film, strips of rubber, and candle wick.

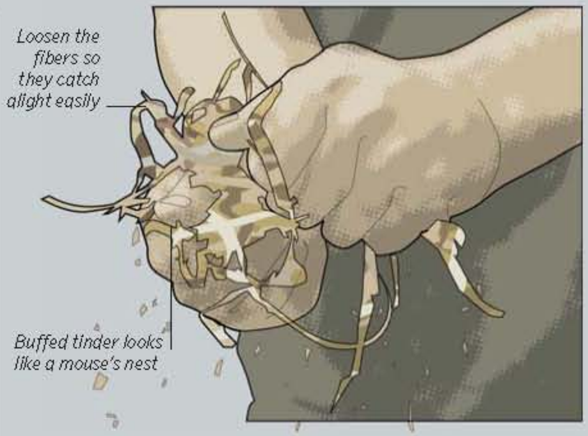

Loosen the fibers so they catch alight easily

Buffed tinder looks like a mouse's nest

TINDER BUNDLE

A good way to prepare your tinder to take a spark or a coal is to make it into a tinder bundle. Vigorously tease, rub, and pull the fibers with your fingers until it becomes a ball the size of a grapefruit. Push the finest, most combustible material into the interior of the ball. Try mixing up your tinder to make it more effective—for example, dry grass, birch bark, and cattail.

STANDING DEADWOOD

A tree that's died but hasn't fallen over has no green foliage and the bark falls off without being replaced. This is the ultimate wood for firelighting, as it can provide kindling and fuel and, when split, makes excellent feather sticks. The bottom few inches may be wet where the tree wicks up moisture from the ground.

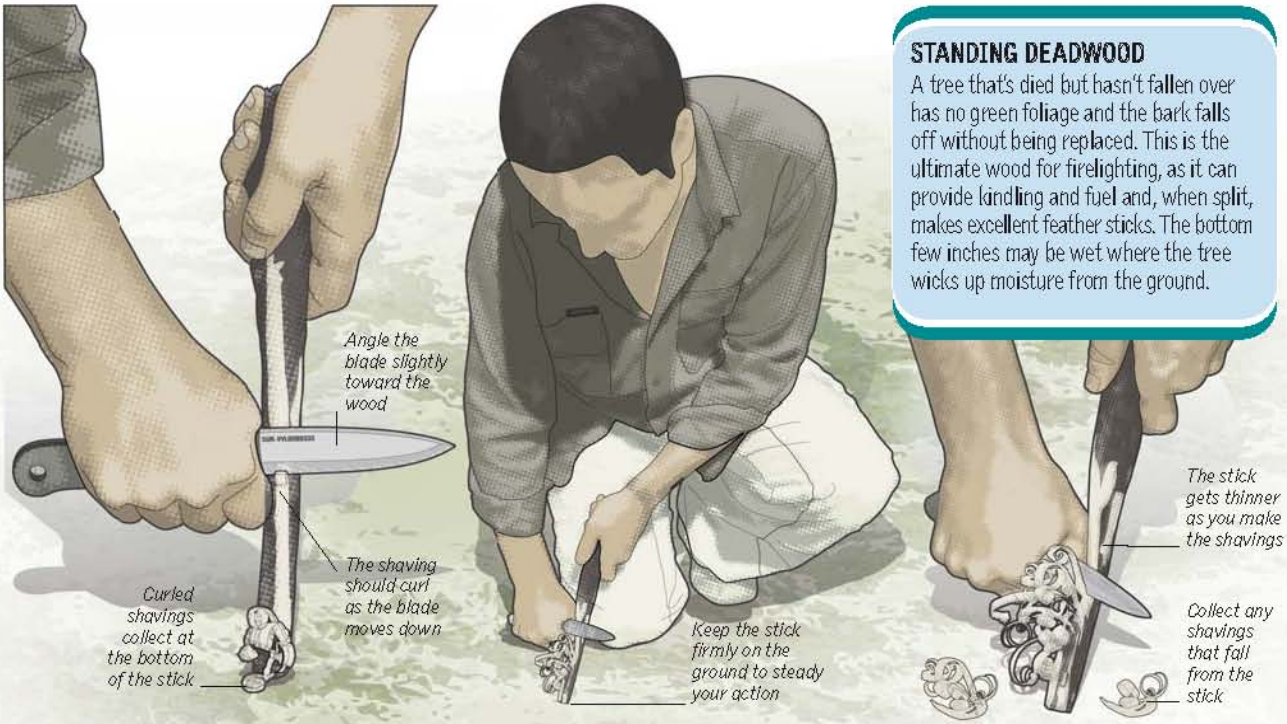
Angle the blade slightly toward the wood
The shaving should curl as the blade moves down
Curled shavings collect at the bottom of the stick
Keep the stick firmly on the ground to steady your action
The stick gets thinner as you make the shavings
Collect any shavings that fall from the stick

② Now tilt the angle of the blade slightly toward the wood.

- Keeping a part at the top of the wood to use as a handle, run the blade down the wood to cut a shaving.
- Stop just before the bottom so that the shaving stays attached to the stick.

③ Turn the stick slightly and run the knife down the edge created in step 2 to create a second shaving. Turn the stick again and repeat, working around the stick.

- Try to get a rhythm going, and put your body behind the cutting action.

④ Once you have mastered the technique, continue to feather the rest of the stick.

- When you've finished you will have a thin stick with curled shavings still attached, and ready for use as kindling.

KINDLING

Kindling is the second material element of firelighting and, when dry, is added to burning tinder. Kindling can be as thin as a match or as thick as a finger. You should be able to snap it with your hands. If it doesn't break with a crack then it's probably not completely dry. If it's damp, remove the outer bark (this retains most of the moisture) and break the kindling down into small sticks 6 in (15 cm) long.

TYPES OF KINDLING

- Soft-wood twigs are very combustible, while wood with flammable resin, such as pine, burns hot and long.
- You can use some types of tinder as kindling—for example, bark, palm leaves, pine needles, grass, ground lichens, and ferns—but you will need larger quantities for kindling than you needed for tinder.

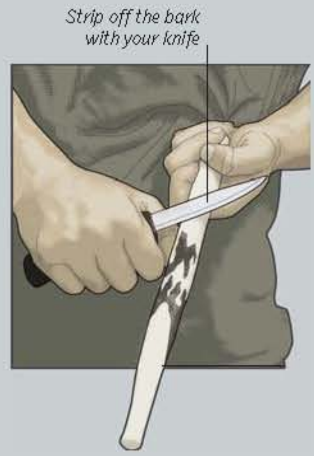
Strip off the bark with your knife

FUEL

Initially, your fire needs constant tending, but once it can sustain itself for five minutes it's established and you can add increasingly larger fuel to create a good heart—a bed of hot coals that sustains the fire with minimum effort on your part. The fuel should be about as thick as your wrist or forearm. At first, use dry fuel split into sticks that catch alight easily. Add green (live) wood and larger logs (whole or split) later, once when the fire is established. In wet conditions, build your fire under cover to keep off the rain.

TYPES OF FUEL

- Hard woods from mainly deciduous trees (such as oak, maple, ash, beech, and birch) burn hot and long, produce good coals, and are more efficient. But they are hard to get going.
- Soft woods from conifers (such as pine, fir, and spruce) are easier to light as they can contain resin, but burn faster and produce less heat than hard woods. They also make more smoke.
- Peat is found on well-drained moorland and can be cut with a knife. However, it needs a good air supply.
- Charcoal is lightweight, smokeless, and burns hot.
- Dried animal droppings provide a good smokeless fuel.

CHAR CLOTH AND FIRE CAN

BEFORE SETTING OFF ON A TRIP, add either some char cloth or a fire can to your backpack. These excellent forms of tinder are reliable and potentially life-saving. Alternatively, you can take some mayasticks. Char cloth is very easy to make, so keep some dry in your emergency fire kit or lining the bottom of your survival kit (see pp. 60–61).

TOOLS AND MATERIALS

- A can with a tight-fitting lid, such as a small shoe-polish can
- Nail
- 100 percent cotton cloth
- Knife or scissors
- Spark or flame

MAKING CHAR CLOTH

Char cloth is cotton cloth that's been combusted in the absence of oxygen (pyrolysis). It's lightweight, takes up almost no space, and produces an ember from even a weak spark extremely well. Char cloth works only when it's completely dry, so keep it in a watertight container.

① Turn the lid of a can upside-down and hammer a nail into the center.

- Experiment with the size of the hole. Usually, the smaller the hole the better.

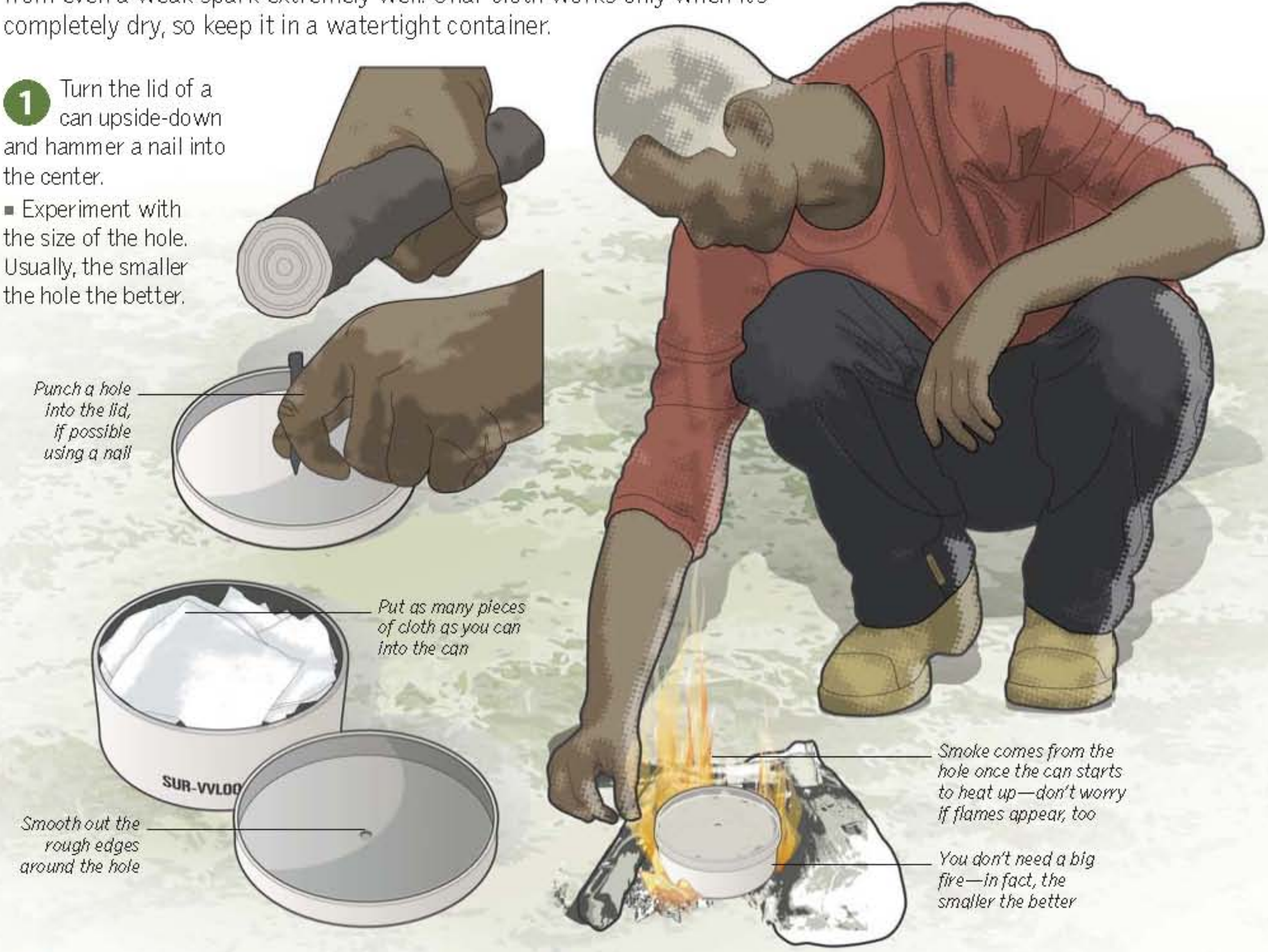
Punch a hole into the lid, if possible using a nail
Put as many pieces of cloth as you can into the can
SUR-VVLOO
Smooth out the rough edges around the hole
Smoke comes from the hole once the can starts to heat up—don't worry if flames appear, too
You don't need a big fire—in fact, the smaller the better

② Cut 100 percent cotton cloth into pieces that fit into the can without folding.

- Vary the sizes of the pieces so they don't just stack flatly on top of each other, but at the same time don't squash them in.
- Securely place the lid on the can.

③ Place the can on a fire to burn off all the oxygen inside the can, or place it on top of some good coals scraped to the side of your camp fire.

- When the smoke stops the process is complete.
- Safely remove the can from the fire.
- Don't open the can until it's cool.

MAKING A FIRE CAN

You can use a fire can to start a stubborn fire when conditions are less than perfect, or to boil water, do some basic cooking, or warm your hands on a cold day. Once lit, a fire can burns for hours with a concentrated, controllable flame that produces no smoke. When it starts to fail, you can either replenish the existing cardboard with more wax or replace the cardboard and start again.

TOOLS AND MATERIALS

- A can with a tight-fitting lid, such as a small shoe-polish tin
- Cardboard (ribbed or plain)
- Candle and a match or lighter
- Knife or scissors

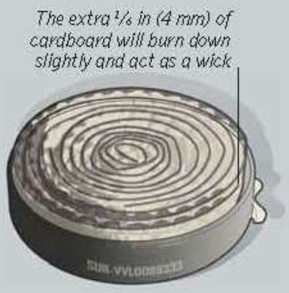

The extra ⅙ in (4 mm) of cardboard will burn down slightly and act as a wick

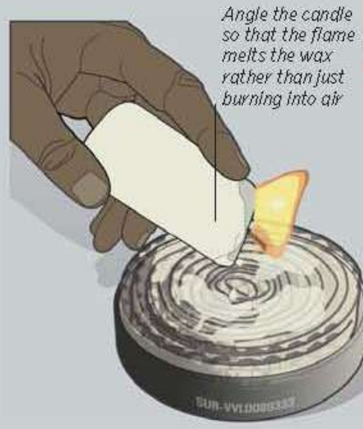

Angle the candle so that the flame melts the wax rather than just burning into air

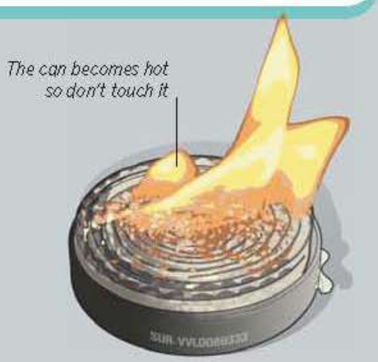

The can becomes hot so don't touch it

1 Cut out a long, thin piece of cardboard that's ⅙ in (4 mm) wider than the depth of your can.

- If it's ribbed, cut it across the grain.
- Roll it tightly along its length until the roll just fits inside the can.

2 Light your candle and let the melting wax drop into the can.

- Let the wax soak the cardboard and fill up the can—it's a slow process.
- Stop when the wax nears the top.
- Allow the wax to harden.

3 When the can has cooled down, hold it at an angle and light the top of the cardboard with the candle.

- The flame should be concentrated and spread across the top of the can.

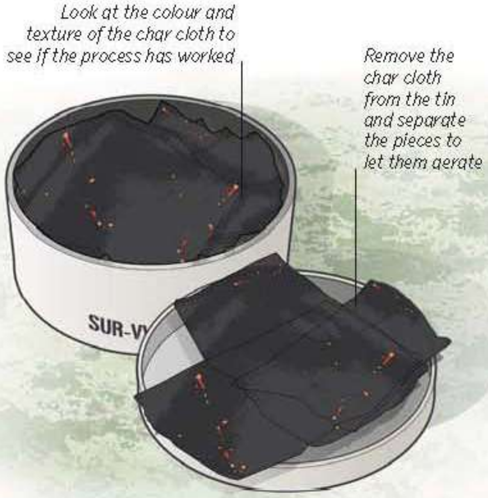

Look at the colour and texture of the char cloth to see if the process has worked

Remove the char cloth from the tin and separate the pieces to let them aerate

4 Once the tin has cooled down, remove the lid and examine the charred cloth.

- The cloth should be completely black and have a semi-rigid, but softish texture. If it's fawn or brown, put it back on the fire and leave for longer.
- Cloth that's brittle and crumbles is of no use.
- Check the cloth works by striking a spark on to a piece; the spark should create a small red ember.

FIRE WADS AND MAYASTICKS

Fire wads are rolled strips of newspaper soaked in a fuel. Once dry, they make handy waterproof tinder. Mayasticks are pieces of resinous fatwood from the highlands of Mexico and Guatemala. The pieces are easy to light, even when wet, and generate a hot flame.

MAKING FIRE WADS AND MAYASTICKS

- To make fire wads, tightly roll strips of newspaper into a tube 4 in (10 cm) long and tie them with cord. Soak them in melted wax and allow to dry. Light the center with a match.
- To light a mayastick, cut a little sliver of wood with your knife and lift it from the surface of the stick. When a little resin seeps out, set it alight with a flint and steel (see p. 127).

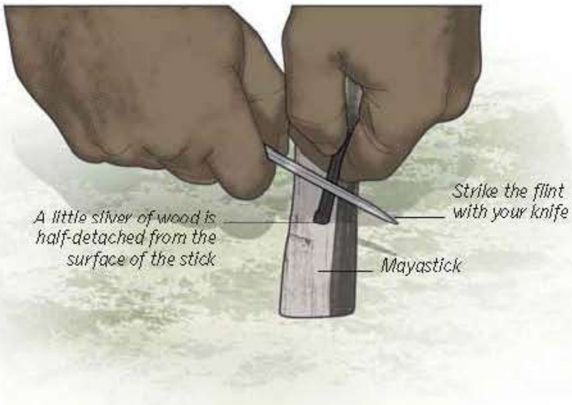

A little sliver of wood is half-detached from the surface of the stick

Strike the flint with your knife

Mayastick

TYPES OF FIRE

ONCE YOU HAVE COLLECTED your tinder, kindling, and fuel, and made sure it's dry and ready at hand, you can set about making your fire. There are many different types of fire to choose from, depending on what you need the fire for (see below). Before you start to build your fire, make sure you have prepared the ground so that everything is safe (see pp. 118–19).

CHOOSING A FIRE

If you have a choice of fuel, as well as time to invest in building a specific type of fire, look at the choices below and determine which one is likely to suit your immediate needs.

- Your primary consideration should be the function of the fire: warmth is probably your most urgent requirement, but other uses include cooking, signaling (see pp. 238–39), drying wet clothing, and disposing of waste. You may want a particular type of fire because it lasts all night.
- Consider the availability of the components you need to build the fire—for example, the right fuel and the best ground. A rule of thumb is to estimate how much you think you'll need, then double it.

FIRE ESSENTIALS

There's plenty of advice surrounding the pros and cons of making fires, but three general tips will help:

- Never make things any harder than they need to be. Choose a fire that requires the least effort for the maximum gain.
- It is more efficient to build a small fire and sit close to it, than to build a large fire and sit far away from it.
- If all your wood appears wet, remove the bark and split the wood—the center will usually be dry.
- Once the fire is established, place damp tinder, kindling, and larger pieces of wood close to the fire to dry out. Keep watch on them so that they don't dry out to the point where they catch fire!

TYPE OF FIRE	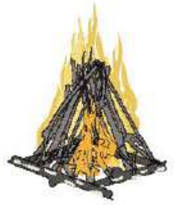 **TEEPEE**	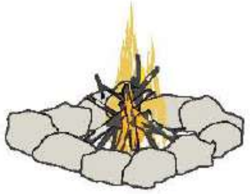 **STONE-LINED**	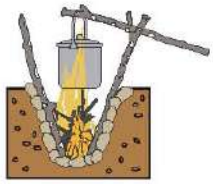 **AUTOMATIC**	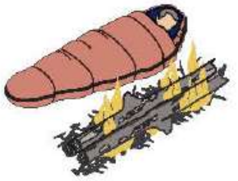 **LONG LOG**
HOW	▪ Surround the tinder ball with kindling in the shape of a teepee. Arrange small, medium, and large fuel logs in a square at the foot of the teepee.	▪ Arrange large non-porous stones in a circle, place a tinder ball in the middle, with kindling around it. Add fuel logs when the fire is established.	▪ Line a hole 3 ft (1 m) deep with non-porous stones. Put tinder and kindling inside and rest large logs against the sides so they drop down as they burn.	▪ Put tinder, kindling, and fuel logs in a depression 6 ft (2 m) long. Lay two long fuel logs on top of the burning fire.
WHERE	▪ Areas with abundant fuel, because the fire requires lots of firewood	▪ Windy locations ▪ Popular areas with existing fire rings and where low-impact camping is vital	▪ Rock-free earth or sand, because it's easy to dig	▪ Forested areas, because the fire requires long logs for fuel
WHY	▪ Quick to light ▪ Wet wood can be burned because it's dried by the heat of the inner fire	▪ Stones shelter fire from wind ▪ Using an existing fire ring reduces the fire's impact on the environment	▪ Self-feeding once lit, which means you don't need to keep adding fuel	▪ Long-lasting (the fire can stay alight all night) ▪ Emits a great deal of heat
USES	▪ Warmth ▪ Cooking ▪ Signaling (if green vegetation is burned)	▪ Warmth ▪ Cooking ▪ Signaling (if green vegetation is burned)	▪ Cooking ▪ Signaling (if green vegetation is burned)	▪ Warmth (in cold weather, build a fire either side of you). ▪ Cooking (once embers have formed)

LIGHT YOUR FIRE

There are numerous ways to build a fire and get one going, and everyone has their favorite. The following example is a tried-and-trusted method that's versatile and works well in a variety of conditions.

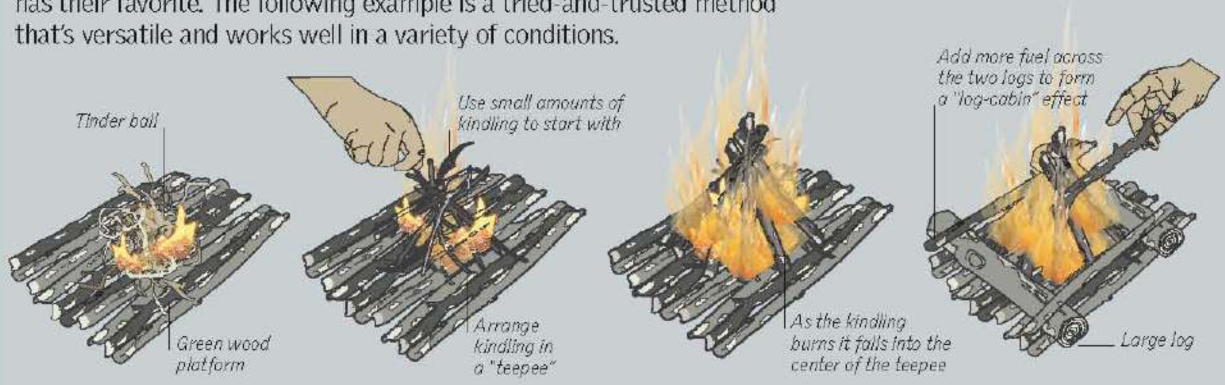
Tinder ball
Green wood platform
Use small amounts of kindling to start with
Arrange kindling in a "teepee"
As the kindling burns it falls into the center of the teepee
Add more fuel across the two logs to form a "log-cabin" effect
Large log

① Place your tinder ball (see p. 120) on a platform composed of green wood.

- Light the tinder with your chosen method (see pp. 126–27) and let it catch.

② Gently lay kindling by the tinder ball.

- Build a kindling "teepee", as this lets the fire breathe where the heat is strongest.
- Feather sticks also make ideal kindling (see p. 120).

③ As the kindling catches fire and the flames grow, add larger pieces of kindling.

- Continue to add kindling, gradually building up to split logs (see p. 148).

④ The fire is established when you can leave it alone for five minutes without it going out.

- Then lay a large log on the wind-facing side and one on the other side.

TYPE OF FIRE	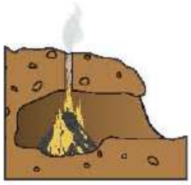 SNAKE HOLE	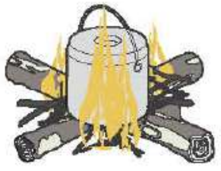 STAR-SHAPED	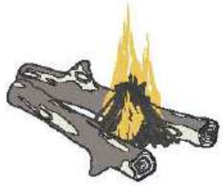 HUNTER	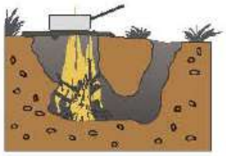 DAKOTA HOLE
HOW	▪ Create a hole in the side of a bank and a chimney up through the earth. Light a fire inside using any fuel.	▪ Build a fire from tinder, kindling, and fuel. ▪ Arrange four logs so they meet in the middle. ▪ Push logs in as they burn.	▪ Make a fire out of tinder, kindling, and any type of fuel. Place two long logs either side of the fire in a V-shape.	▪ Dig a large hole for the fire and a slightly smaller hole for the chimney, with a tunnel linking the two. Use small logs as fuel and cook at ground level.
WHERE	▪ Windy locations ▪ Rock-free earth or sand bank	▪ Forested areas, because of the logs the fire needs	▪ Cold or windy locations	▪ Anywhere you can dig a hole, because the fire requires very little fuel once established
WHY	▪ Chimney creates a draft, giving a high-temperature fire ▪ Sheltered from the elements	▪ Long-lasting ▪ Good embers for cooking	▪ Hardwood logs shelter the fire from the wind ▪ Produces a great deal of heat	▪ Concentrated heat ▪ Flames are below ground so fire is hidden
USES	▪ Warmth ▪ Cooking ▪ Waste disposal	▪ Cooking—balance a pot on the logs ▪ Heating water	▪ Warmth ▪ Cooking	▪ Cooking ▪ Keeping warm ▪ Drying wet clothing

MAKING SPARKS AND FLAMES

LIGHTING TINDER IS THE FIRST STAGE of making a fire. Matches and lighters do this instantly, but there are a number of ways to create a spark that you can then use to coax your tinder into a flame. This is relatively easy when everything is dry, but with patience and persistence it is also possible when conditions are bad.

IGNITION DEVICES

There are various ingenious methods of creating that all-important spark or flame to ignite your fire. If you don't have matches or a lighter, you'll need to use another device, such as a flint and steel. You can also improvise by using an external energy source, such as:

- Focusing the sun's heat with a magnifying glass or beverage can.
- Making sparks with a battery.
- Creating a chemical reaction using potassium permanganate.

STRIKING A MATCH

It may seem simple enough, but there's a way of striking a match—known as "commando style"—that reliably produces a flame in all kinds of conditions.

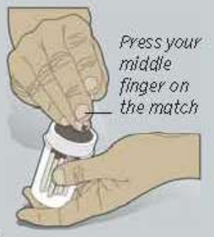
Press your middle finger on the match

1 Hold the box in one hand and a match between the thumb and first two fingers of the other.

- Strike the match firmly away from you.

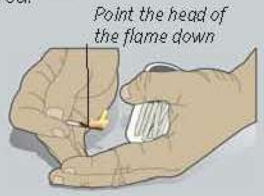
Point the head of the flame down

2 When the match lights, immediately cup your hands to protect the flame.

- Let the flame burn a little way along the stem before using it to light the tinder.

TYPE OF DEVICE	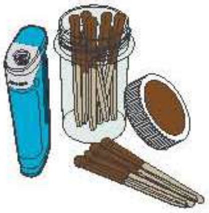 MATCHES/LIGHTERS	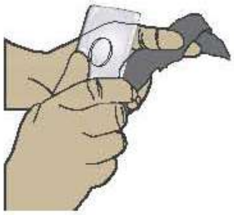 MAGNIFYING GLASS	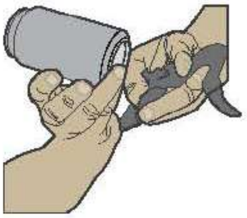 BEVERAGE CAN	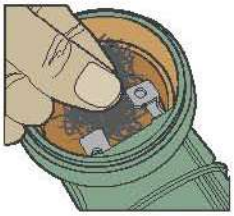 BATTERY
GEAR	▪ Waterproof box of matches. ▪ Cigarette lighter taped to a piece of cord around your neck. ▪ Dry tinder	▪ Magnifying glass ▪ Dry tinder	▪ Empty beverage can ▪ Dry tinder	▪ Flashlight and its battery ▪ Wire or wire wool ▪ Dry tinder
WHAT TO DO	▪ To light a match "commando style," strike it away from you on the box and then cup it in your hands (see above).	▪ With the magnifying glass, focus bright sunlight onto some dry tinder and create a hot spot. Hold the magnifying glass steady until the tinder catches alight.	▪ Polish the bottom of the can (see Signal mirror, p. 245). ▪ Catch bright sunlight on the shiny surface and reflect it onto tinder to create a hot spot. Hold the can steady until the tinder catches alight.	▪ Lay the wire across the battery terminals to create some sparks. ▪ Remove the bulb from the flashlight and place wire wool over the terminal. Switch on the flashlight to create sparks.
COMMENTS	▪ Waterproof matches are usually just standard matches that have been coated with wax and varnish. ▪ Always have a lighter around your neck on a piece of cord.	▪ When you choose a compass, make sure it has a magnifying glass incorporated into it (for reading the details on maps). ▪ You can also use the lens in reading spectacles.	▪ You'll need to practice this technique so that you can rely on it in a survival situation.	▪ The thinner the wire the better this will work, especially with lower voltage batteries (1.5v). ▪ Use this method for a short period only, otherwise you will drain the battery.

CREATING A CHEMICAL FLAME

Potassium permanganate is an extremely useful item to have in your survival kit, because you can use the chemical not only to sterilize water and wounds, but also to create a spark to light a fire. You will need some sugar to make it work, so use some from your survival rations or crush a hard candy.

WARNING!
Potassium permanganate is a strong oxidizer that can, when mixed with certain chemicals, create an explosive mixture. It can also stain your clothes and skin.

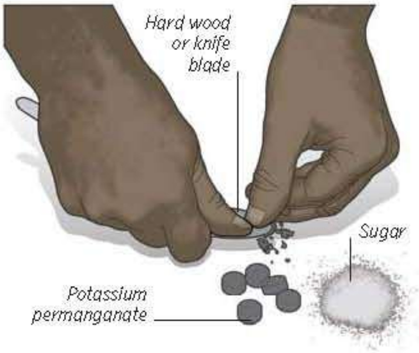

① On a hard surface, such as a flat piece of wood or rock, crush some potassium permanganate. Have an equal amount of sugar ready.

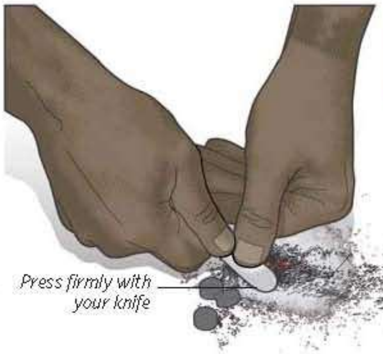

② Mix the ingredients. With the back of your knife, push down on the mixture and drag it along the hard surface to create friction.

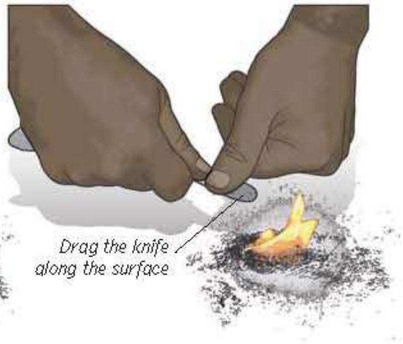

③ Continue to push down and drag your blade until you get a spark. ■ The flame dies down quickly so have your tinder at the ready.

	FIRESTEEL	FLINT AND STEEL	ONE-HANDED STRIKER	POTASSIUM PERMANGANATE
TYPE OF DEVICE	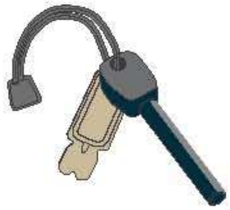	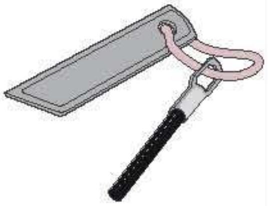	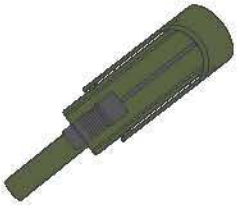	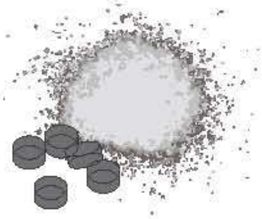
GEAR	■ Ferrocerium firesteel ■ Metal striker ■ Dry tinder	■ Magnesium-alloy rod ■ Steel striker (usually a short piece of hacksaw blade) ■ Dry tinder	■ A magnesium-alloy rod and steel device that can be operated using only one hand ■ Dry tinder	■ Potassium permanganate ■ Sugar ■ Knife ■ Dry tinder
WHAT TO DO	■ Hold striker next to the tinder. ■ Place the firesteel under the striker, directly onto the tinder. ■ Draw the firesteel back across the striker to direct the sparks onto the tinder. The firesteel moves, not the striker.	■ Follow instructions for the firesteel (left).	■ Place the flint rod in the tinder. ■ Press the thumb button onto the rod and push the handle down the length of the rod. ■ Pushing harder creates greater friction, which leads to a more intense spark.	■ Mix potassium permanganate and sugar in equal amounts on a hard surface. ■ Press down on the mixture with your knife and drag the blade along to create a spark (see above).
COMMENTS	■ Lasts for about 12,000 strikes. ■ Temperature reaches 3,000°C. ■ Works in all weathers and at any altitude.	■ There are many variations and kits; one has a magnesium block for the holder and the striker shaves slivers of magnesium into the tinder to help with ignition.	■ Works in all weathers. ■ Has a safety feature that prevents accidental use. ■ Designed for fighter pilots, who may injure an arm or hand after ejecting.	■ You can also mix a small amount of potassium permanganate with glycol or antifreeze. Quickly wrap in paper and put on the ground. Warning: stand well clear as the combustion can be sudden and dangerous.

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL— IN THE DESERT

USEFUL EQUIPMENT

- Personal Locator Beacon (PLB)
- Sunglasses and sunscreen
- Cotton scarf or hat
- Signal mirror and signal flares
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

DESPITE A LACK OF SURVIVAL KNOWLEDGE, a party of five survived searing temperatures, dehydration, hunger, and the threat of animal attack for six days when their plane crashed in the Kalahari Desert in Botswana.

On Wednesday, March 1, 2000, Carl du Plessis and three associates—Mike and Lynette Nikolic, and Nebojea Graorac—set off on a business flight from the capital of Botswana, Gaborone, bound for Maun, a popular tourist town in the north of the country. However, the aircraft developed engine trouble during the flight and was forced to crash-land. Critically, radio contact was lost, so no distress call was emitted before the plane went down. The pilot and all four passengers on board suffered burns during the crash and subsequent fire, but two sustained even more serious injuries—Lynette Nikolic damaged her spine and left arm, and Graorac suffered a punctured lung.

"CRITICALLY, RADIO CONTACT WAS LOST, SO NO DISTRESS CALL WAS EMITTED BEFORE THE PLANE WENT DOWN"

The following morning, Du Plessis and pilot Costa Marcandonatos decided to walk the estimated four days to Maun to find help. Their plan was misguided from the start—Maun was actually around some 190 miles (300 km) to the west, and other settlements to the east were much closer. They attempted to navigate by the sun but, lacking a machete, couldn't penetrate the dense bush. They were forced to follow elephant tracks to find watering holes, meeting elephants on one occasion, and heard the roar of lions at night. By Saturday they had walked 56 miles (90 km), but their zig-zag route had taken them only 20 miles (30 km) from the crash site.

Despite their mistakes, they stumbled upon a manned hunting lodge the following day. After their initial efforts to radio the authorities had failed, a helicopter finally reached the crash site on Monday, March 6, airlifting the casualties to hospital for emergency treatment six days after the crash.

WHAT TO DO

ARE YOU IN DANGER?

NO — If you are in a group, try to help any others who are in danger

YES — Get yourself out of it:
Sun/Heat—Find or improvise immediate shelter
Animals—Avoid confrontation and move away from danger
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO — If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

YES — If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO — You are faced with surviving for an indefinite period—until you are located or you find help

YES — If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO — If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers either a better chance of survival, rescue, or both

YES — Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Make an informed decision on the best location to move to
- Leave clear indications of your intent (written messages or signs) if abandoning a vehicle
- Have aids to location accessible while moving and deployed when static
- Protect yourself against glare from the sun and windburn
- Seek or improvise shelter when not moving and seize all opportunities to collect fuel for a fire—deserts get cold at night
- Be on constant lookout for signs of water or civilization—such as green vegetation, converging animal/human tracks, or circling birds

DON'T

- Sit directly on the hot ground when you stop
- Travel during the hottest part of the day
- Take risks—a twisted ankle caused by running down a sand dune could be fatal
- Force the pace—travel at the speed of the slowest person in the group
- Shelter in dry river beds, because of the potential risk of flash floods

YOU SHOULD STAY **

DO

- Seek or improvise shade and work during the coolest part of the day/night. A shelter dug even 6 in (15 cm) below ground level will provide a much cooler place to rest
- Ration sweat, not your water; you require around 1¾ pints (1 liter) per hour in temps above 100°F (38°C), half of this if it's cooler
- Continually re-assess your water situation and options for augmenting your supplies
- Leave everything out overnight that could collect dew so you can drink it
- Prepare all of your aids to location for immediate use. Be constantly alert for signs of rescue

DON'T

- Leave a broken down vehicle unless absolutely necessary—it's easier to see and is what rescuers will be looking for
- Waste energy unnecessarily—be idle unless there is something that requires doing
- Ignore your fire—use anything that burns (wood, vehicle tires) to generate heat and smoke

---

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.

** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

FIRE BY FRICTION

THIS METHOD USES PRIMITIVE EQUIPMENT, such as a bow drill, and requires knowledge, skill, practice, the correct wood, time, and effort. However, once learned it's possibly the most rewarding of all survival skills. Your first fire created from a coal produced by a fireset that you have made yourself will be a fire you will never forget. In a genuine survival situation, however, you would always start a fire with the quickest and easiest method available.

TOOLS AND MATERIALS

- Cordage such as para cord, fishing line, shoelaces, or bailing twine
- Wood
- Knife
- Bradawl (optional)

THE FOUR SKILLS

There are many ways to make fire by friction, including a hand drill, bamboo fire saw (see p. 133), and bow drill. Each has its own advantages, depending on the materials available and your practical skills. Regardless of the method, however, there are four main elements to producing fire by friction—each a skill on its own:

- Identifying and procuring the correct types of wood.
- Manufacturing the individual pieces that make the fireset.
- Using the fireset to produce a coal.
- Nurturing the coal into a flame.

MAKING A BOW DRILL

The bow drill set is one of the most efficient methods of making fire by friction. Try to find the right type of wood for each part, especially the drill and hearth board (see Testing Wood, opposite).

1 To make the bow, cut a stick to the same length as the distance between your armpit and your fingertip.

- Cut a notch (or make a hole with a bradawl) near each end.
- Tie cord with an arbor knot (see pp. 142–43) to one notch and with half hitches (see pp. 144–45) to the other. Leave only enough slack to wrap around your finger.

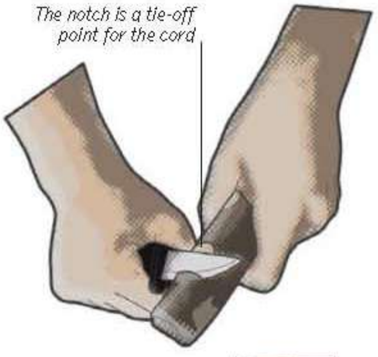
The notch is a tie-off point for the cord

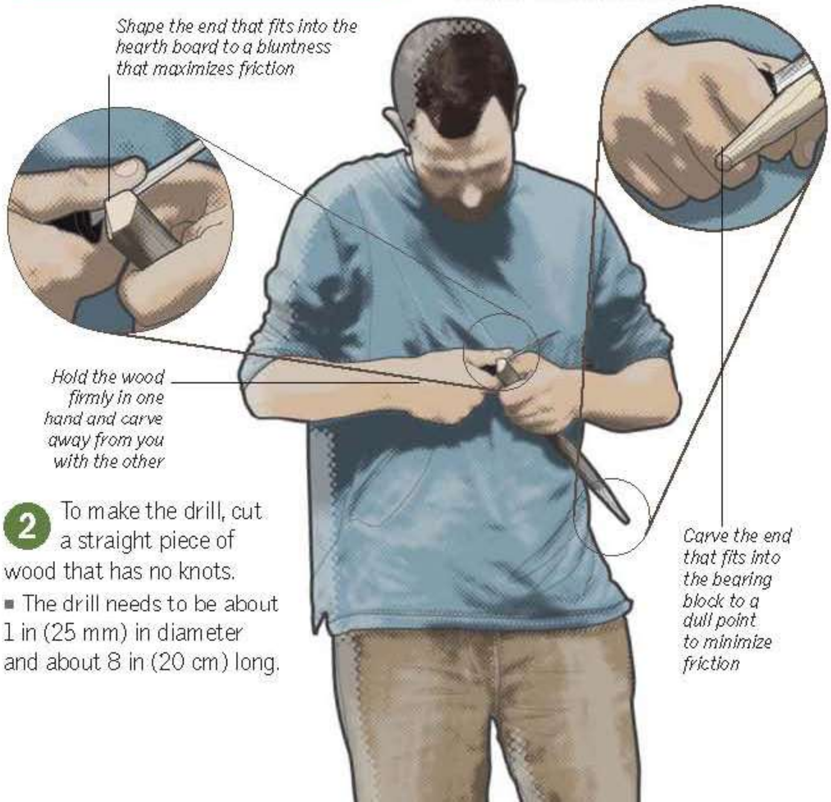
Shape the end that fits into the hearth board to a bluntness that maximizes friction
Hold the wood firmly in one hand and carve away from you with the other
Carve the end that fits into the bearing block to a dull point to minimize friction

2 To make the drill, cut a straight piece of wood that has no knots.

- The drill needs to be about 1 in (25 mm) in diameter and about 8 in (20 cm) long.

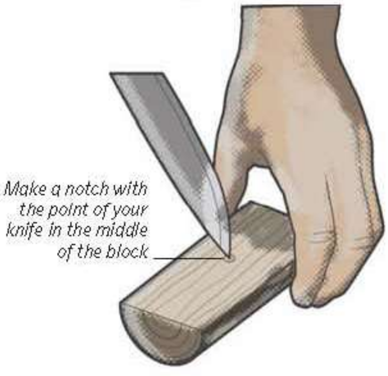
Make a notch with the point of your knife in the middle of the block

3 Make the bearing block from a piece of hard wood about 3–4 in (7.5–10 cm) across.

- Cut it to a length of 4–5 in (12–15 cm), split it in two, and make a notch in it.
- The drill's dull point fits into the notch, while you hold the block when you are bowing (see p. 132).

TESTING WOOD

Standing deadwood is ideal for the drill and hearth board as it's dry. Soft deadwood is easier to use than hard deadwood, but if you can't tell the difference, use these tests.

THUMBNAIL TEST

Cut through the bark of a branch, exposing a patch of wood beneath. If you dent this wood easily with your thumbnail, it's a soft wood. If you can only make a small impression, then it's a hard wood.

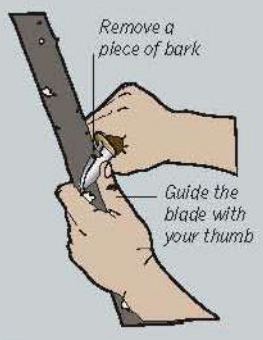
Remove a piece of bark
Guide the blade with your thumb

1 Hold your knife at a slight angle to the wood. Push it with a thumb on top of the blade.

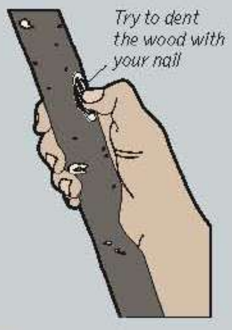
Try to dent the wood with your nail

2 Press your thumbnail at an angle of 90 degrees to the wood in the exposed patch.

CHAR TEST

Each type of wood produces a distinctive char and coal. Some, such as hazel and lime, make a coal easily but the fireset will wear out quickly. The char test helps you discover which woods make a good coal. Drill into a softwood hearth board until it smokes, continue for five seconds and then stop. Look at the char produced:

- If it's like a fine powder it will probably produce a coal.
- If it's coarse, or appears to disintegrate into nothing, it probably won't make a coal.

4 Make the hearth board from a flat piece of wood about 1 in (25 mm) thick. You may have to split some wood to get what you need.

- Hold your drill so that it sits about $^{2}/_{5}$ in (10 mm) from the edge of the board.
- Mark this point with your knife and notch out a small hole to help the drill get started.

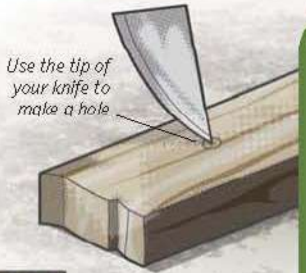
Use the tip of your knife to make a hole

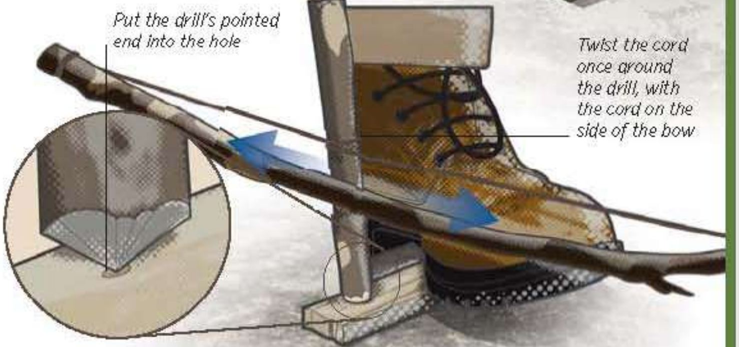
Put the drill's pointed end into the hole
Twist the cord once around the drill, with the cord on the side of the bow

5 Set up the bow, drill, and bearing block (see also p. 132) and bow the drill into the hole in the hearth board until the hole is as wide as the drill. Don't go all the way through.

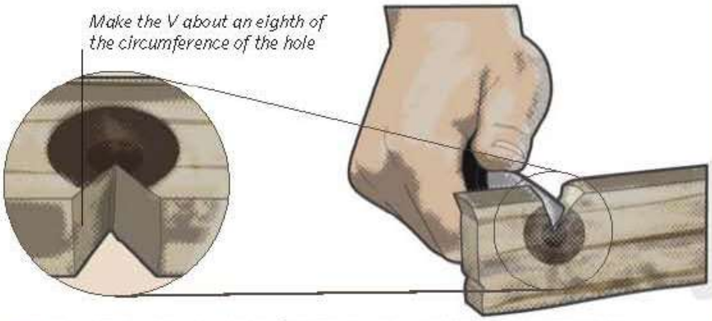
Make the V about an eighth of the circumference of the hole

6 Cut a V to the center of the hole through the hearth board. The bottom of this V is where the char collects and the ember forms in the ember pan (see p. 132).

THE FINISHED FIRESET

The bow drill set works best when the drill and hearth board are made from the same wood, because both parts then wear down evenly. This may not always be possible, so experiment with different woods.

- Good woods to start with are sycamore, hazel, ivy, lime, willow, and sotol (not found in Europe).
- Also useful are alder, birch, cedar, pine, spruce, maple, oak, and poplar.

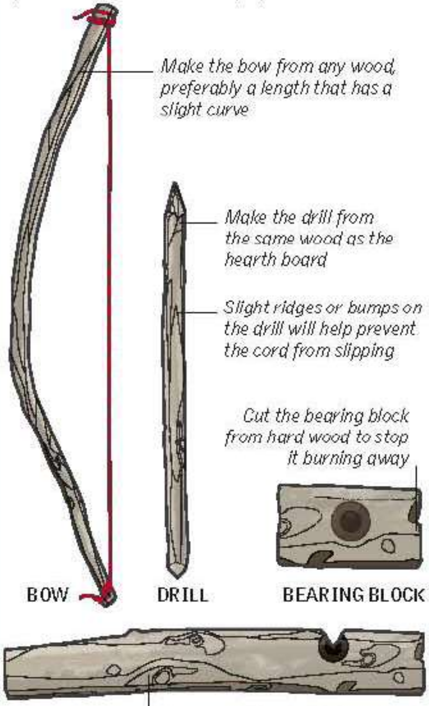
Make the bow from any wood, preferably a length that has a slight curve
Make the drill from the same wood as the hearth board
Slight ridges or bumps on the drill will help prevent the cord from slipping
Cut the bearing block from hard wood to stop it burning away
BOW
DRILL
BEARING BLOCK
HEARTH BOARD
Make the hearth board from wood that heats up and ignites easily

USING A BOW DRILL

Before you use a bow drill, have your dry tinder bundle, kindling, and fuel beside you. Until you become proficient at using a drill, you will use lots of energy, so beware of overheating and dehydration. Turning a coal into a fire is a skill in itself, so use every opportunity to practice making tinder bundles and lighting them with a small coal from your camp fire.

TOOLS AND MATERIALS

- Bow drill set (see pp. 130–31)
- Thin, dry piece of bark to use as an ember pan
- Buffed tinder, kindling, and fuel

GETTING INTO POSITION

Place the ember pan on a flat, dry surface and align it with the V on the hearth board. Rest one foot on the board. Loop the bow cord once around the drill, starting on the inside. Insert the drill into the hole on the hearth board, slot the block on top (with a green leaf in the notch for lubrication), and lock your wrist into your shin as support. Kneel on the other leg and bear down on the block.

1 Hold the bow horizontal to the ground and parallel to your body.

- Lean slightly forward to apply downward pressure to the drill and back to reduce it.
- Start bowing slowly, using the full length of the bow. Aim for an even, flowing rhythm.
- Breathing steadily as you bow, gradually increase speed and pressure until smoke appears. Vigorously bow with maximum effort ten more times—this usually produces a coal in the ember pan.

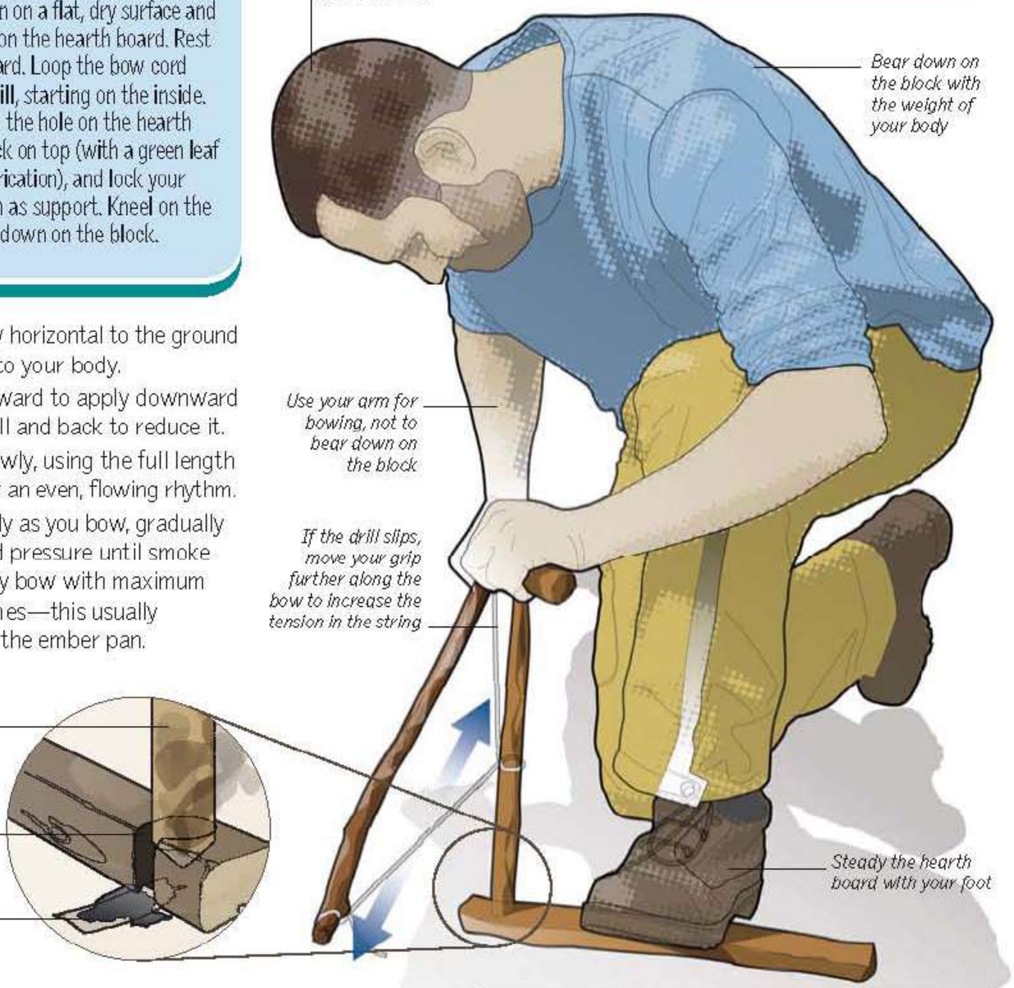
Keep your head as still as you can above the bow
Bear down on the block with the weight of your body
Use your arm for bowing, not to bear down on the block
If the drill slips, move your grip further along the bow to increase the tension in the string
Keep the drill upright as you bow
As the drill turns, friction heats the wood
A pile of black ash will collect on the ember pan
Steady the hearth board with your foot

2 Gently remove the drill and lift the hearth board up and away from the smoking coal in the ember pan.

- If the coal is sticking to the board, gently tap the board with your knife to loosen it.
- Carefully lift the ember pan off the ground and hold it in the air to see if the coal is glowing.
- If the coal in the ember is just smoking, and not glowing, then gently fan it with your hand until you see it glow.

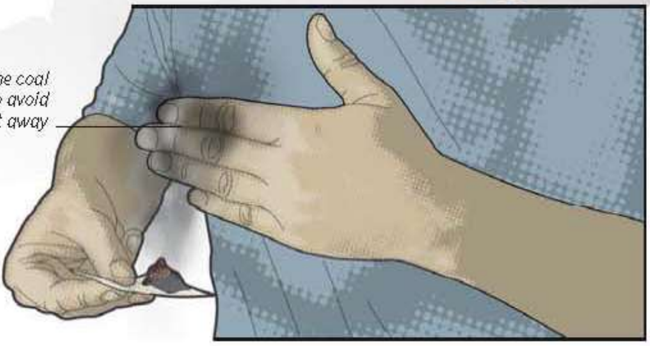
Fan the coal gently to avoid blowing it away

3 Gently transfer the glowing coal into your tinder bundle.

- The coal must touch the tinder to transfer its heat, so squeeze the tinder around the coal, but don't crush it or starve it of oxygen.

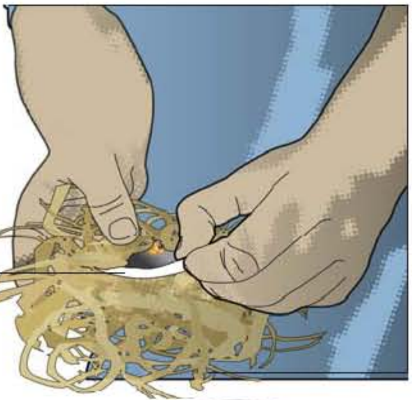
Tip the coal off the ember pan and into the center of the tinder bundle

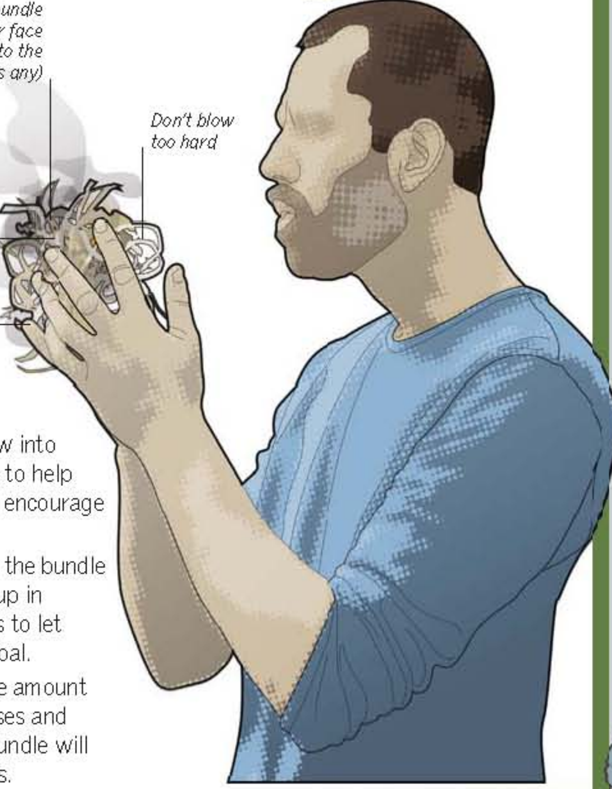
Hold the tinder bundle away from your face (turn your back to the wind if there is any)
Don't blow too hard
Air from your breath or the breeze helps with combustion
Hold the tinder bundle up, but not too close to your face

4 Gently blow into the bundle to help fan the coal and encourage it to catch.

- You can swing the bundle down and back up in between breaths to let dry air fan the coal.
- Continue as the amount of smoke increases and eventually the bundle will burst into flames.

5 Don't throw the burning bundle on the ground, just calmly place it where you intend to have your fire. Start building the fire around the bundle (see pp. 124–25).

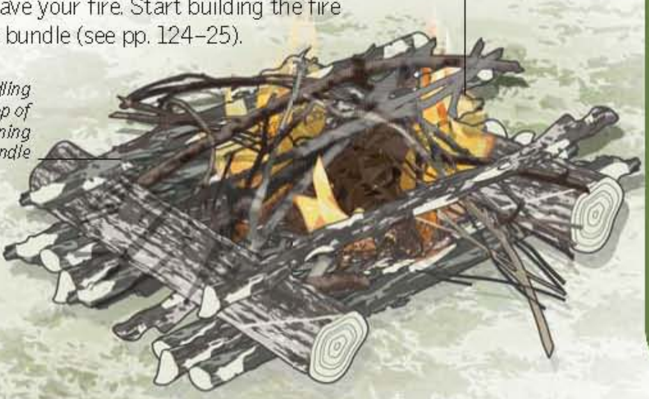
Once the kindling is set alight, you can add some fuel
Place kindling over the top of the burning tinder bundle

BAMBOO FIRE SAW

This friction method uses a saw technique to light tinder collected from scraping the outside of the bamboo with your knife.

MAKING THE FIRE SAW

You need a bamboo section 18 in (45 cm) long. Split it lengthways into two equal pieces.

1 Sharpen the edge of the first piece with your knife and then cut a V-shaped notch in the outside middle of the second piece.

2 Hold tinder over the notch in the second piece of bamboo with another bamboo strip that's snapped but not fully broken.

- Kneel down and brace the sharpened piece of bamboo vertically between the ground and your stomach (place a pad on your stomach for comfort).
- Place the notch and tinder over the sharp edge of the first piece, and rub the second piece up and down in a rhythmic movement.

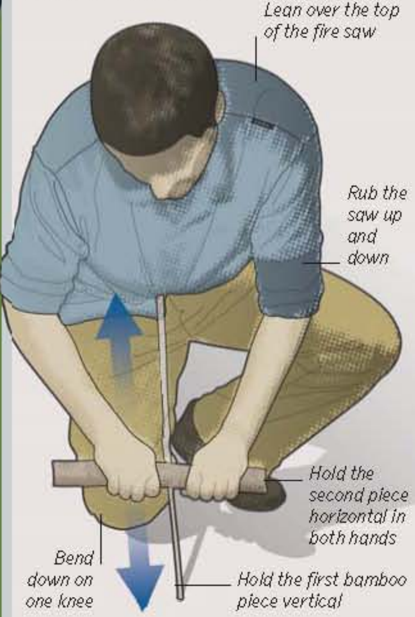
Lean over the top of the fire saw
Rub the saw up and down
Hold the second piece horizontal in both hands
Bend down on one knee
Hold the first bamboo piece vertical

3 When the friction produces lots of smoke, continue for 20 strokes more and you should have a coal.

MANUAL SKILLS

Learning skills that you can use in the wilderness is always a "work in progress." Having taught thousands of students over several years, I am yet to conduct a course in which I didn't learn a new skill. For example, this could be a more efficient way to erect a shelter, how to tie a new knot, or the best way to carry out a task more safely and efficiently.

In the military, we invariably teach students how to accomplish a particular task using only the most basic gear. By learning how to perform a task using the absolute basic equipment, they will be able to understand the important elements of the task and will have no choice but to get them right in order to succeed. This in turn helps students to appreciate how and why certain techniques do and don't work, and gives them an opportunity to use their improvisational skills to modify basic techniques in order to get the most out of them. Knowing the correct method of using a knife, for example, will not only improve safety on an expedition, but will also mean that only the minimum

In this section YOU WILL DISCOVER...

- the merits of a **hank** in **your pocket**...
- how to strip a **nettle stem** and not **get stung**...
- when **to twist** and when **to roll**...
- how to tell a **working end** from a **standing end**...
- the uses of a **sheet bend** or a **Siberian hitch**...
- the **difference** between a **parang** and a **kukri**...
- how to **fell a tree** with a **thumping stick**...

amount of valuable energy is being used to complete a task. Equally, knowing how to tie a few simple knots, which can be used in the majority of survival situations, will allow you to use whatever cordage you have in the most efficient way possible.

If you find yourself in a survival situation with more than the most basic gear, life will be much easier. However, the ability to improvise what you need when you don't have it could mean the difference between continued survival and despair.

If you give someone with little understanding of knots a length of cord and ask them to complete a task, they will invariably use the whole length of cord. If, however, you teach someone two simple knots and set them the same task, they will invariably use only the necessary amount of cord.

USING THE BEST KNOT, lashing, or hitch for a particular task means that you will use cord more efficiently and, therefore, ensure that your supply will go further. In a survival situation, where you may be moving every day, you need to be able to use the same pieces of cordage day after day.

- Use knots that don't require cutting and that can be untied quickly and easily (trying to undo a knot with cold, wet hands in the dark, and when you are tired and miserable, can be an emotional experience that usually ends in frustration and several pieces of cut cordage).
- Before you set off into the wilderness, learn a few basic knots, such as the Siberian hitch and the taut-line hitch, which are great for rigging a simple poncho shelter.
- Practice these knots until you can tie—and untie—them with your eyes closed.
- If you are trekking in a cold environment, practice tying them wearing gloves.

> “The **ability** to **improvise** what you **need** when you **don't have it** can mean the **difference** between **continued survival** and **despair**”

MAN-MADE CORDAGE

A VITAL PIECE OF SURVIVAL EQUIPMENT, man-made cordage is one of those small items that has many uses: building a shelter, repairing equipment and clothing, making traps and nets, and producing fire by friction using a bow drill. Always pack plenty.

PACKING CORDAGE

Before you set off on your travels, check how much cordage you will need. Make sure you pack it in two places—in your backpack and on your person, in case you get separated from your pack. Make sure, too, that the needle you pack in your survival kit has an eye big enough for the inner strands of the cordage.

PARACORD

Parachute cord, or "paracord," is a type of cordage originally developed for the rigging lines of parachutes and then adopted as the standard utility cordage by most of the world's military forces. It is a lightweight, nylon rope composed of 32 braided strands. Each strand contains a number of smaller threads that can also be used. In most situations, paracord is a very good option because it's readily available, strong, and packs down small.

USING CORDAGE

It is vital that you use your limited supply of cordage as efficiently as possible.

- Use only the minimum cordage that is absolutely necessary to accomplish each task.
- Use simple, strong knots that, if possible, can be released without cutting. This leaves the cordage intact for the next task.

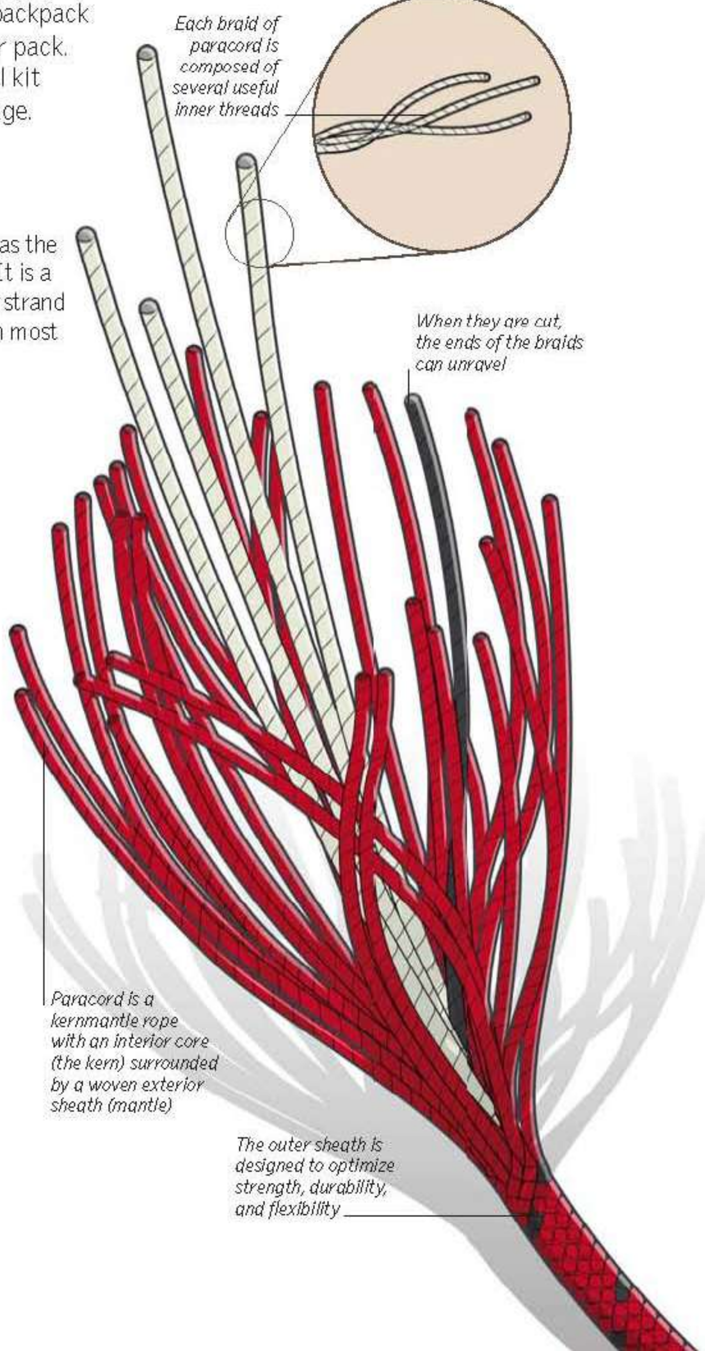
Each braid of paracord is composed of several useful inner threads
When they are cut, the ends of the braids can unravel
Paracord is a kernmantle rope with an interior core (the kern) surrounded by a woven exterior sheath (mantle)
The outer sheath is designed to optimize strength, durability, and flexibility

SURVIVAL TIP

Green is the ideal color for military paracord as it remains camouflaged. In a survival situation, red is best because it stands out—it's easier to find if you drop it and can be used as an aid to location. Don't pack just one very long piece of paracord (150ft/30m, for example). It's much better to cut the paracord into 30 ft (10 m) lengths and tie each of them into a coil called a hank (see opposite and below).

TEN HANKS

A hank can fit in the palm of your hand. It takes up very little room so you can easily stuff up to 10 of them into your pack and pockets, and leave them there until you need them.

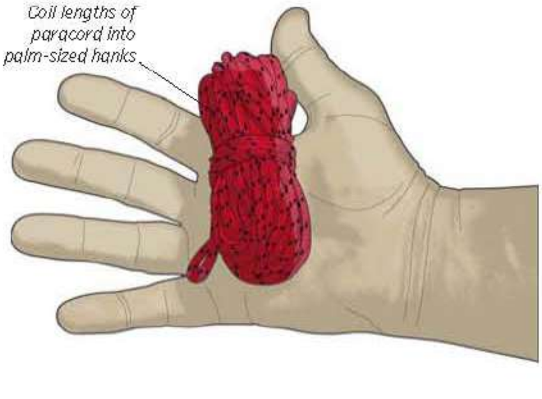
Coil lengths of paracord into palm-sized hanks

TYING A HANK

It's best to keep your cordage in a hank—you can untie it when you need a piece and retie it to store it again. Tying a hank is straightforward and soon becomes second nature, because you'll do it again and again.

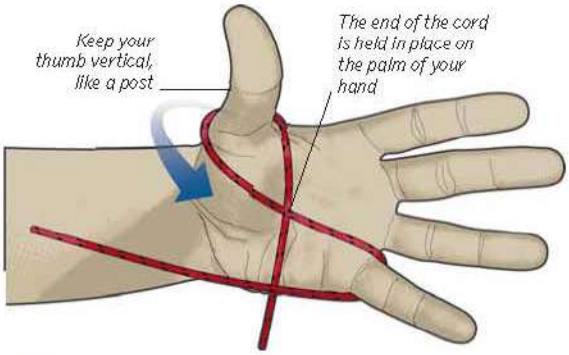
Keep your thumb vertical, like a post
The end of the cord is held in place on the palm of your hand

1 Hold out your hand (you can use either hand) with the thumb and the fingers spread.

- Lay the end of the cord on your palm, then loop the cord in a figure-eight around your thumb and little finger.

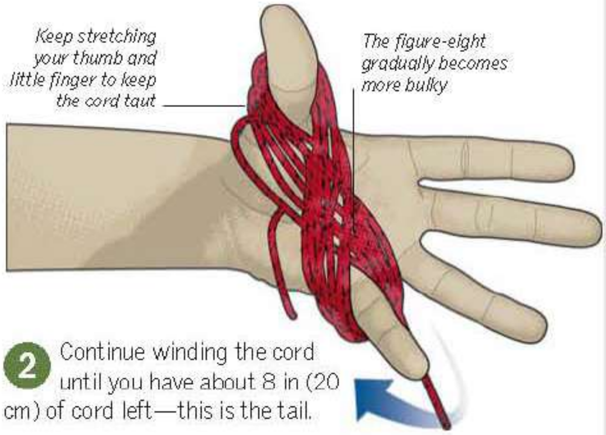
Keep stretching your thumb and little finger to keep the cord taut
The figure-eight gradually becomes more bulky

2 Continue winding the cord until you have about 8 in (20 cm) of cord left—this is the tail.

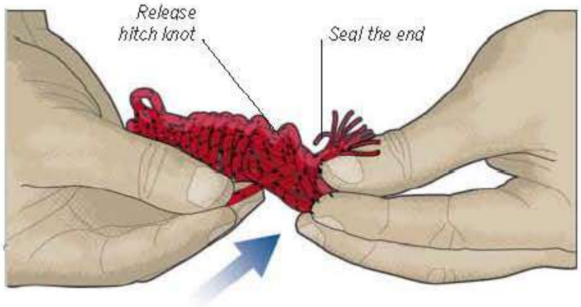

3 Take the hanked cordage off your thumb and little finger.

- Secure the hank by tying the tail around it with a release hitch knot (see p. 143).

OTHER TYPES OF CORDAGE

In a survival situation you may have to improvise and use other types of cordage, especially if you didn't bring any paracord with you or have run out. Before you start making it from natural materials (see pp. 138–41), see what else you might have that would be quicker and easier to adapt.

SHOELACES

The normal laces of shoes and boots are often strong enough to use as cord. However, the members of many military units, who are limited to the amount of equipment they can stow or carry, replace their bootlaces with paracord. They use twice as much as they need and wrap the extra length around the top of their boots before tying it. This gives them a ready supply of extra cordage that they can cut off as needed. You can easily do the same—but never sacrifice all the cordage you are using as shoelaces as you'll need to make sure you can still walk comfortably.

BELTS, CLOTHING, AND HATS

You can cut any type of belt, whatever it's made of (leather, canvas, and so on) into strips to use as cordage. You can cut clothing into strips, too. However, you don't have to destroy a shirt completely: start at the bottom hem and cut 1 in (2.5 cm) strips all the way round. You can even use your hat—a 1 in (2.5 cm) strip from the brim of a jungle hat is just over 3 ft (1 m) long. Better still, coil paracord around the crown of your hat and stitch it in place (just the front and the back).

RAIDING YOUR KIT

You may have other items in your equipment that you could use as cordage, such as guylines from tents or draw cords from clothing or sleeping bags. Dental floss, especially the waxed type, is extremely strong and can be used effectively for lashing or sewing. If you have a vehicle, you probably have a towrope; take a part of it and unravel the rope into smaller pieces of cordage.

BE A SCAVENGER

Few places are untouched by modern man. Even in remote areas you'll come across discarded items from the "civilized world"—from plastic bottles to bailing twine. In a situation where improvisation could be the difference between life and death, collect anything you could use to your advantage.

SEALING PARACORD

Paracord can be awkward to use because its ends, if left unprotected, eventually unravel. Try to seal the ends after you have cut them by holding a flame next to them until they melt. Be careful, as the melted nylon can drip and burn.

NATURAL CORDAGE

IF YOU DON'T HAVE any man-made cordage, you may be able to make a natural substitute. Use the roots and stems of plants to produce improvised rope for erecting shelters, and use sinew from the tendons of larger animals for strong whipping and sewing.

WARNING!
Take extreme care when you collect and work with plant fibers, especially removing plants such as green ferns from the ground. Bending a stem at its base and pulling it up can expose fibers that are razor-sharp. Cut the plant with a knife or wear gloves. The juice of some plants can irritate the skin, so wash your hands when you've finished.

PREPARING NETTLE FIBERS

Old stinging nettles or wood nettles with long stems have the most fibers. Wearing gloves (or with your hands covered), strip the leaves from the nettles by grasping the stem at the base and pulling your clenched fist up the full length of the stem.

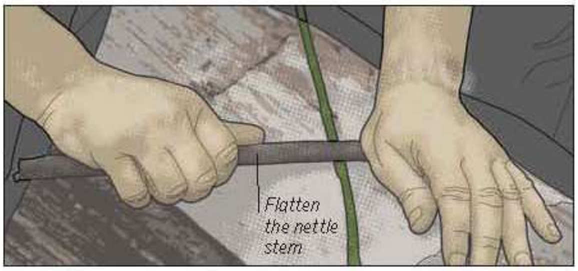

1 Sit astride a log and lay a nettle stem in front of you. Roll a smooth, rounded stick backward and forward on the stem, pressing down hard.

- Continue rolling until the whole stem is crushed.

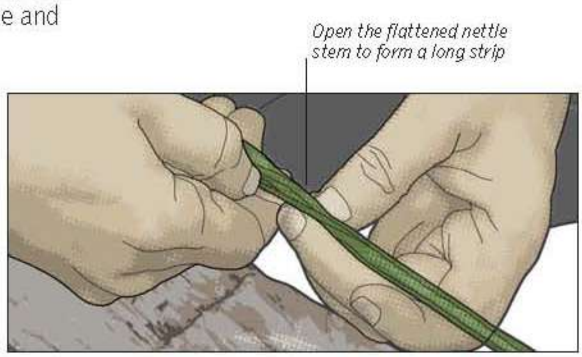
Open the flattened nettle stem to form a long strip

2 Tease open the crushed tissue along the whole length of the stem of the nettle with your thumbnail. This exposes the spongy pith within.

CHOOSING YOUR RAW MATERIALS

Many kinds of raw material can be used to make natural cordage, from bark and roots to stems and sinews. Search your immediate environment first to see what can be used—you may not have to look too far.

	BARK	ROOTS
FACTS	■ Tree trunks have layers between the bark and heartwood called the "cambium" layers, which are best for cordage.	■ Many tree roots make good cordage and lashings, but evergreen trees, such as pine, fir, spruce, and cedar, work best.
FINDING	■ The bark of trees such as willow and linden make good cordage. ■ Use other types of bark, such as clematis or honeysuckle, as they break from the tree or vine.	■ Look for new roots near the surface of the soil; these are flexible. Cut slightly thicker roots than you need, as you have to remove the bark. Remove only a few roots from each tree and repair any damage you do to the topsoil.
PREPARING	■ Strip the bark from a dead tree, then beat it to loosen the inner cambium fibers. ■ Pull off the cambium fibers in long strips. Use them as they are (clematis and honeysuckle), or braid them into stronger lengths.	■ Cut a slit along the root. Peel the bark and let the bare root dry and shrink before splitting in half or into quarters. ■ Keep the bark for tinder or kindling. You could also put it on a fire to produce smoke that keeps away flies.

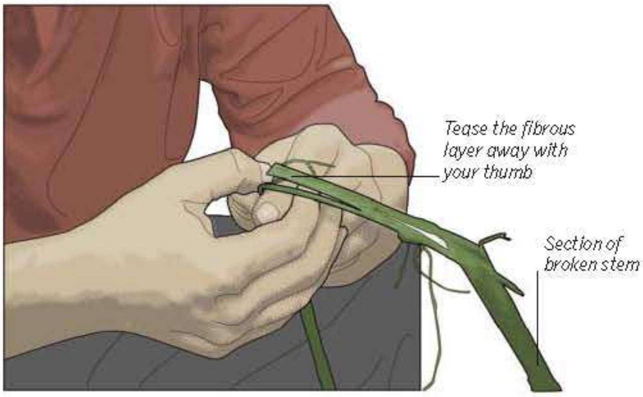

Tease the fibrous layer away with your thumb

Section of broken stem

3 Bend the stem in the middle of its length. When you bend it, the inner pith will break away from the outer skin, making it easy for you to separate them.

MAKING CORDAGE FROM PLANTS

To begin with, you must have access to enough quantities of the plant. Don't start making cordage with the only plant in the area. There should also be some basic characteristics present in order for plant material to work well as cordage.

- The fibers need to be long enough. If you have to splice them together to make workable lengths, you will weaken the line.
- If you're braiding pieces together to make a stronger line, use rougher fibers, as they will bite together better. Shiny or smooth fibers tend to unravel easily.
- The fibers need to be strong. Pull on them to see how much strain they take before breaking. They also need to be pliable enough so they don't break when you bend and tie them.

4 Carefully peel the outer skin from the inner pith.

- With the finger and thumb of one hand, gently pull the layer outward, while pushing down on it with the thumb of the other hand.
- You'll be left with long fibrous strips of outer skin, which you can make into short bindings or natural cordage (see p. 140).

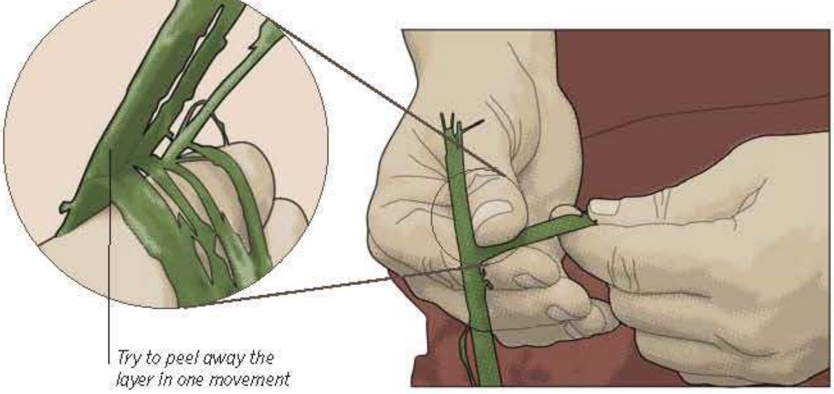

Try to peel away the layer in one movement

	WITHIES	LEAVES	SINEW
FACTS	■ The strong, flexible stems of willow, birch, ash, and hazel are called "withies" and are used for thatching and in gardening. ■ Withies make useful and robust lashing and are best used in spring or summer.	■ The leaves of many plants, such as the lily family, agave, and sisal hemp, contain useful fibers.	■ Animal sinew makes strong and versatile cordage—some sinews can support the weight of a man. ■ Sinew has been used to bind arrow heads to shafts and to sew together everyday clothing, such as leather and mulduks.
FINDING	■ Choose a young sapling or branch that's long and flexible. The fewer the stems or branches the better, as they will need to be removed.	■ To find out if the leaves of a plant contain useful fibers, simply tear a leaf apart to see if it breaks into stringy layers.	■ Sinews are the tendons connecting bone to muscle. The largest are on both sides of the spinal column, running parallel to it. ■ Shorter lengths can be taken from the calves of the hind legs.
PREPARING	■ Remove stems or branches. Grasp the wand at the base and twist until the fibers break, working your way to the tip. Bend the wand into an S-shape and crank the middle of the wand to loosen the fibers. Then cut the wand from the tree and use the withies.	■ Soak the leaves in water to make the inner layers swell and burst. Bacteria work on the cell tissue, separating it from the fibers. ■ Rinse the fibers in fresh water to stop further bacterial decay and then dry them.	■ Take a tendon from an animal's body, remove its outer sheath, then clean it and let it dry. Separate the dried fibers and use them as individual threads, or splice or braid them together to form stronger cords. You can soften dried sinew with water or saliva.

MAKING NATURAL CORDAGE

Once you have prepared a sufficient quantity of fibers, it's best to let them dry before using them—hang them up in the sun or near a fire. In a survival situation, you can use them before drying. Twisting or rolling the fibers are the two main methods of making natural cordage. As with most survival tasks, practice makes perfect.

TOOLS AND MATERIALS
Enough prepared natural fibers (see pp. 138–39) to give you the length of cordage that you require

TWISTED CORDAGE

Start your first length of cordage toward one end, not in the middle, to prevent all the subsequent splices from coinciding at the same point. This provides strength, as each splice is a potential weak point. If they are thin enough, the twisted cordage fibers can then be plaited to create a thicker cord that's even stronger.

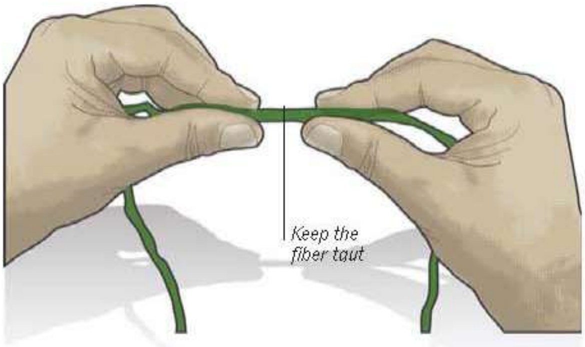

1 Grip the first length of fiber a third of the way along its length between your left thumb and forefinger.

- Grasp the fiber with the thumb and forefinger of your right hand 1 in (3 cm) from your left.

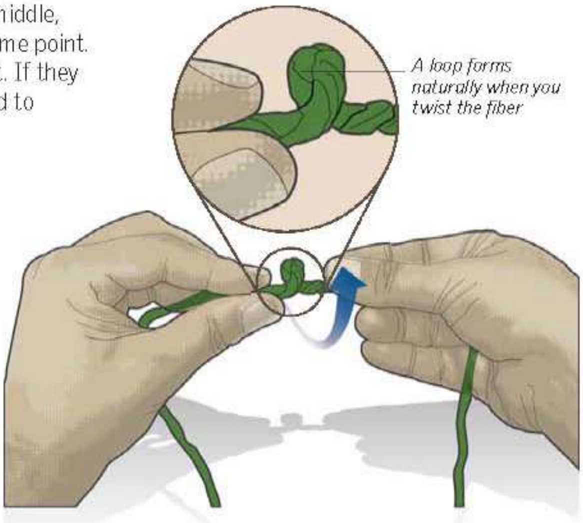
A loop forms naturally when you twist the fiber

2 Twist the fiber with your right hand until a tight loop is formed.

- Keeping the tension in the cord, inch your left thumb and forefinger forward and clamp down on the loop.

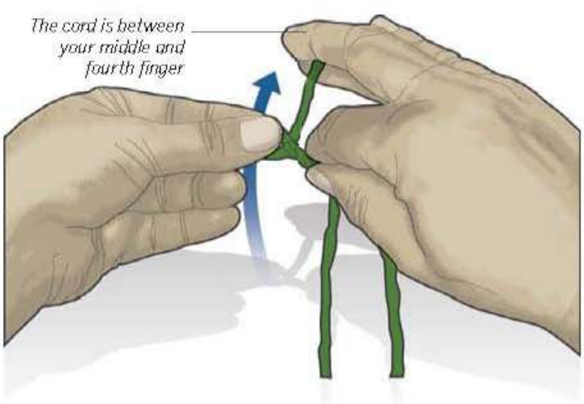
The cord is between your middle and fourth finger

4 Sweep the middle and fourth fingers of your right hand upward, pulling the left-hand cord under the captive piece.

- Inch your left thumb and forefinger forward, and clamp down on the new twist.

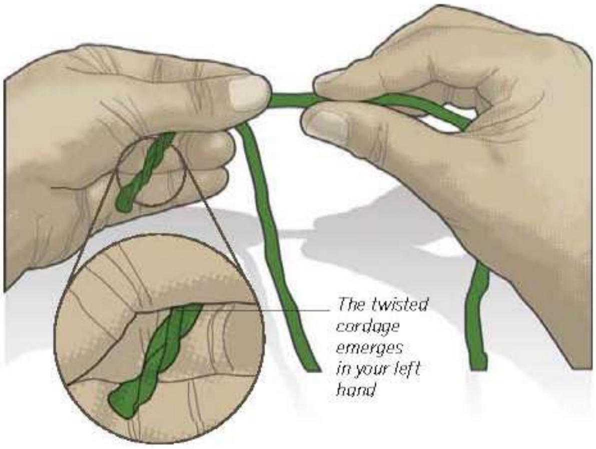
The twisted cordage emerges in your left hand

5 Now that the new twist is clamped by your left hand, you can release your right thumb and forefinger, letting the lower cord fall.

- Grip the upper cord 1 in (3 cm) from your left fingers with your right thumb and forefinger.

SIMPLE ALTERNATIVE TWIST

You need to start with two fibers of different lengths so that when you splice on additional lengths the splices are staggered. Start by tying one end of one fiber with one end of the other.

- Clamp tied ends in your left thumb and forefinger and the two tails in your right thumb and forefinger about 2 in (5 cm) down.
- Roll both fibers between your fingers away from you for one complete roll. Repeat this until you have about 3 in (7.5 cm) of rolled fibers, then clamp them in your right thumb and forefinger.
- Release your left thumb and forefinger and the rolled fibers will naturally twist together to form a strand of cordage.
- Repeat the process until you have enough for your needs.

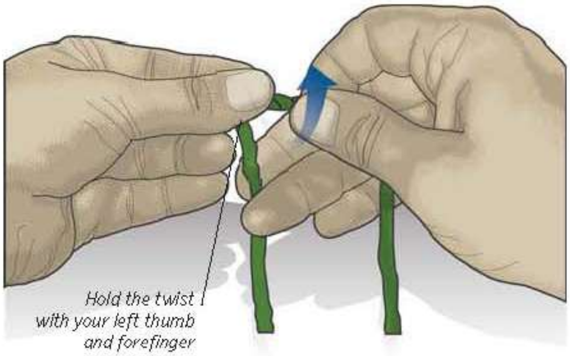
Hold the twist with your left thumb and forefinger

3 Move your right hand back so that 1 in (3 cm) of cord is visible and apply another twist.

- At the same time reach forward with the middle and fourth fingers of your right hand and grasp the cord hanging from your left fingers.

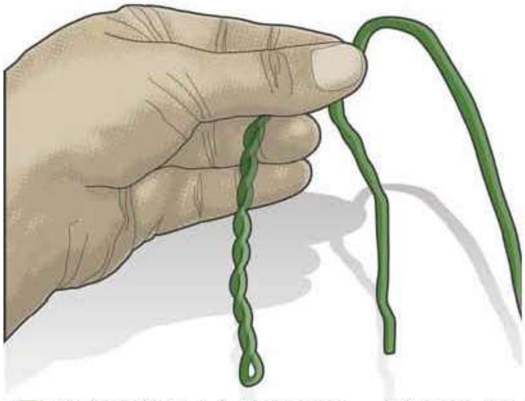

6 Repeat Steps 3, 4, and 5 as you work your way along the length of cord.

- To splice in an extra length, overlap the new piece 2 in (6 cm) along the existing fiber and twist the two together between your left thumb and forefinger. Then continue with the process as before.

ROLLED CORDAGE

This method of rolling cordage, which is also known as the "pygmy roll," is easier to do if you're sitting down. As with the simple alternative twist (see left), you need to start with two fibers of different lengths and tie one end of one fiber to one end of the other.

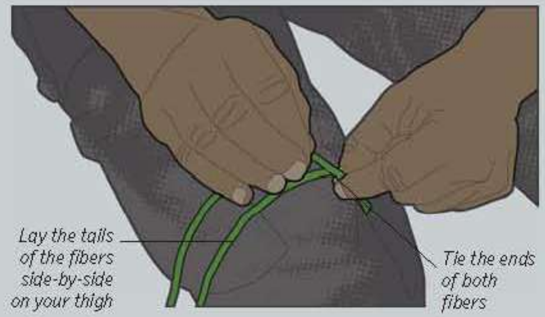
Lay the tails of the fibers side-by-side on your thigh

Tie the ends of both fibers

1 Hold the tied ends between the thumb and forefinger of your left hand.

- Lay the two tails of the fibers on your right thigh.

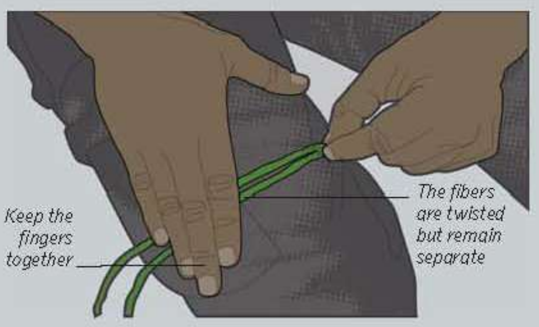
Keep the fingers together

The fibers are twisted but remain separate

2 Use the flat palm of your right hand to roll both fibers simultaneously away from you. You may find it easier to use the flat of your fingers.

- At the end of the roll, clamp the fibers down onto your thigh with the tips of your fingers.

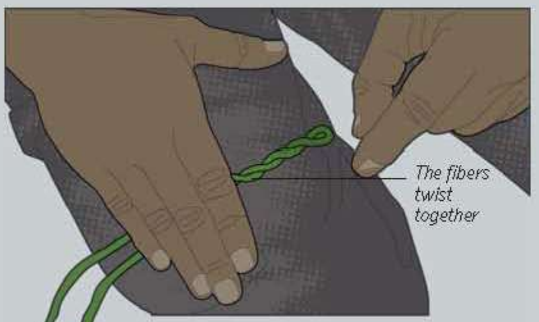
The fibers twist together

3 Release your left thumb and forefinger and the rolled fibers will naturally twist together to form a strand of cordage.

- Repeat the process, splicing in more cordage as required.

TYING KNOTS

THERE ARE HUNDREDS of different types of knot in existence, but by practicing the handful of straightforward knots shown here and overleaf, you should be able to accomplish most survival tasks. Start by tying each knot in the situation you expect to use it—for example, erecting a poncho shelter—and then practice it at least 20 times.

THE ENDS OF A LINE
The end of a rope or line that takes the most active part in knot tying is the working end. The other end is called the standing end—it's more passive, as the knot is tied around it.

SLIP KNOT

A slip knot is a good example of a simple knot that has many practical uses in survival situations.

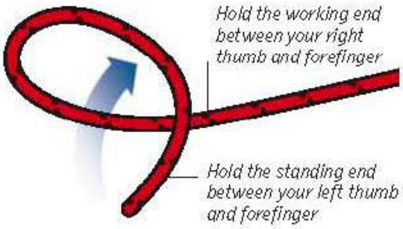
Hold the working end between your right thumb and forefinger
Hold the standing end between your left thumb and forefinger

1 Twist the standing line over the working line to form a loop.

- Hold the loop between your left thumb and forefinger.

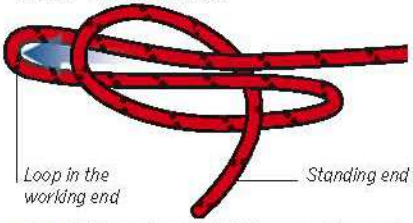
Loop in the working end
Standing end

2 Make a loop with the working end and pass it into the first loop.

- Collect the second loop with your left thumb and forefinger and pull it through.

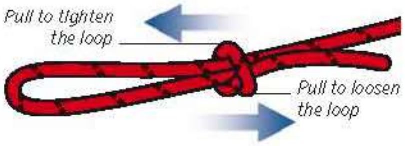
Pull to tighten the loop
Pull to loosen the loop

3 Gather both ends in your right thumb and forefinger and pull them away from the loop in your left hand.

- Put the loop over whatever you want to secure—for example, a button-tie on your bivy sack—and pull the knot tight.

ARBOR KNOT

This all-purpose knot has several survival uses. When erecting a poncho or a tarp (see Chapter Four), it can be used as a tie-off knot to secure one end of a line to a fixed point, such as a tree or peg, where no adjustment is required. It can also be used as a lashing to secure two poles together—for example, in an A-frame shelter (see p. 164)—with just one turn of cordage.

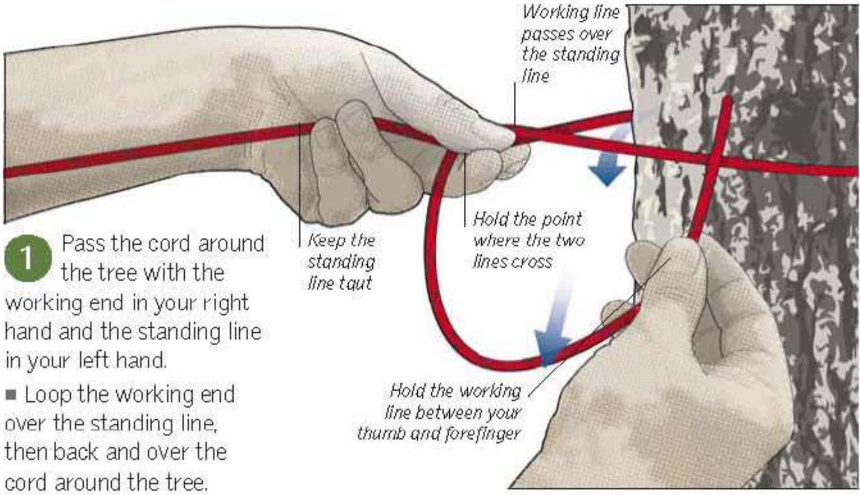
Working line passes over the standing line
Keep the standing line taut
Hold the point where the two lines cross
Hold the working line between your thumb and forefinger

1 Pass the cord around the tree with the working end in your right hand and the standing line in your left hand.

- Loop the working end over the standing line, then back and over the cord around the tree.

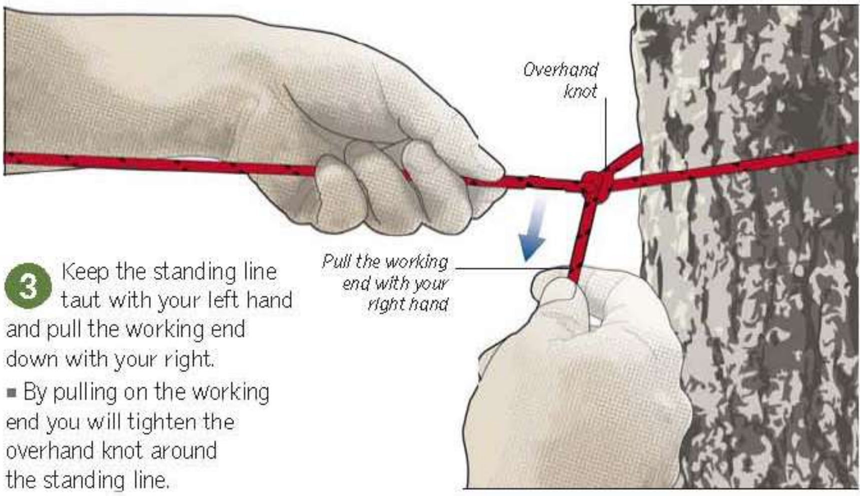
Overhand knot
Pull the working end with your right hand

3 Keep the standing line taut with your left hand and pull the working end down with your right.

- By pulling on the working end you will tighten the overhand knot around the standing line.

OVERHAND KNOTS

An overhand knot is the first part of the knot you use to tie the laces on your shoes. One overhand knot tied around another forms a stop knot or they can be tied as a double overhand knot around a fixture.

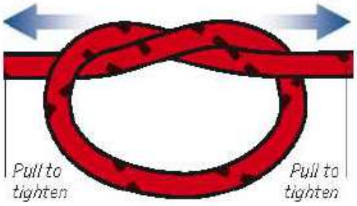
Pull to tighten
Pull to tighten

OVERHAND KNOT

Twist the left line over the right to form a loop. Hold it with your right thumb and forefinger.

- Pass the working end through the loop with your left forefinger and thumb and pull both ends to tighten.

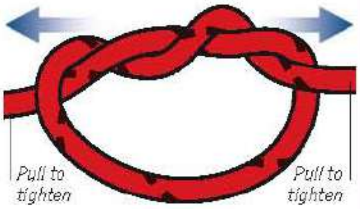
Pull to tighten
Pull to tighten

DOUBLE OVERHAND KNOT

Follow the sequence for the overhand knot (left), but pass the working end through the loop twice before pulling the two ends.

- A stop knot used at the end of a line can stop the line from slipping.

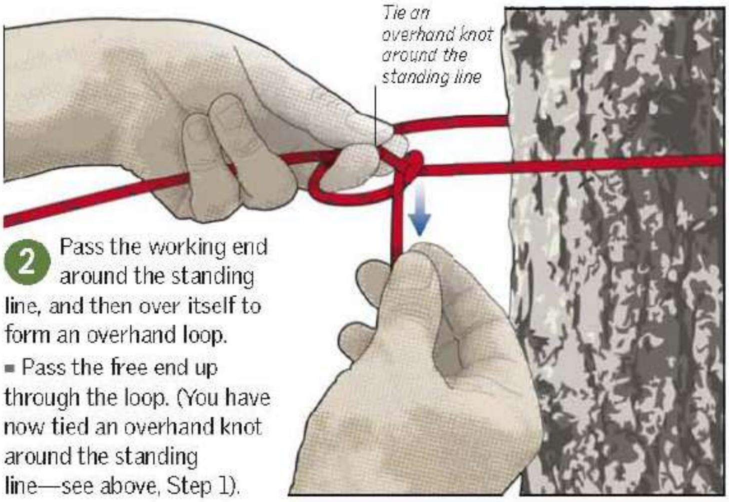
Tie an overhand knot around the standing line

2 Pass the working end around the standing line, and then over itself to form an overhand loop.

- Pass the free end up through the loop. (You have now tied an overhand knot around the standing line—see above, Step 1).

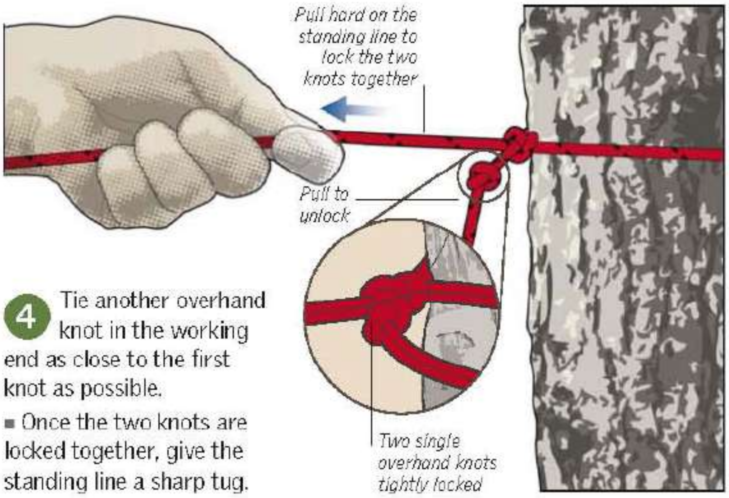
Pull hard on the standing line to lock the two knots together
Pull to unlock
Two single overhand knots tightly locked

4 Tie another overhand knot in the working end as close to the first knot as possible.

- Once the two knots are locked together, give the standing line a sharp tug.

DOUBLE CHAIN FASTENING KNOT

This knot securely attaches the free end of a line to a fixture, particularly when weight is involved and you don't want the knot to slip or come loose.

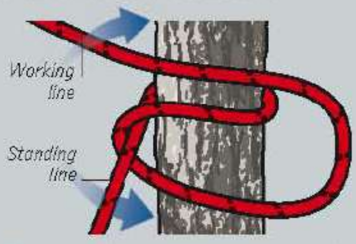
Working line
Standing line

1 Loop the working end around the fixture, such as a tree or pole.

- Pass the working end over and around the standing line.

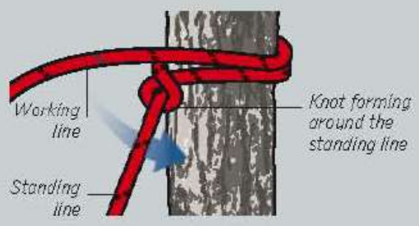
Working line
Knot forming around the standing line
Standing line

2 Pass the working end back around the fixture and then bring it back under the standing line.

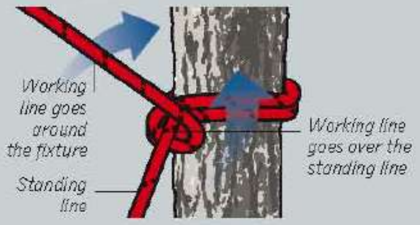
Working line goes around the fixture
Working line goes over the standing line
Standing line

3 Pass the working end over the top and around the standing line again, and once more around the fixture.

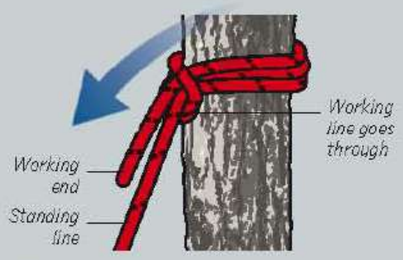
Working line goes through
Working end
Standing line

4 As you bring round the working line, thread it underneath itself to complete the knot.

- Lock the knot neatly into position.

DOUBLE SHEET BEND

The sheet bend, also known as the "weaver's hitch," is an extremely useful survival knot as it securely ties together two ropes. Double sheet bends are even more secure and are recommended for tying ropes of different thickness. In this instance you use the thinner of the two ropes to tie the knot.

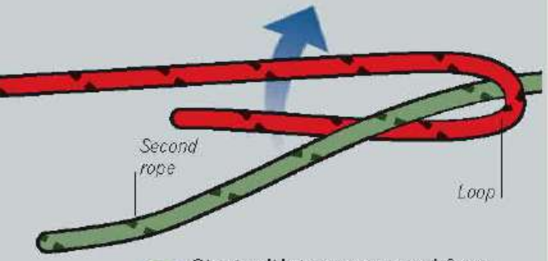

1 Start with two ropes and form a loop in the working end of one of them.

- Pass the end of the second rope through the loop.

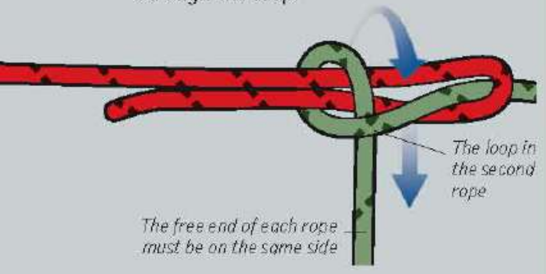
The loop in the second rope
The free end of each rope must be on the same side

2 Pass the free end of the second rope first under, then up and over the loop. Pass the end through the new loop you have created.

- To finish a single sheet bend, pull the standing end of each rope to tighten it.

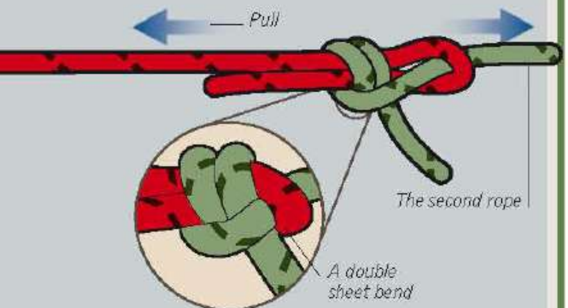

3 To tie a double sheet bend, pass the free end of the second rope through the loop (as in step 2) again.

- Tighten the knot as before by pulling the free end of each rope.

SIBERIAN HITCH

This knot, which is also known as the "Evenk knot," is good for attaching a rope to a fixture—for example, when securing a ridgeline to a tree for a poncho shelter (see p. 158).

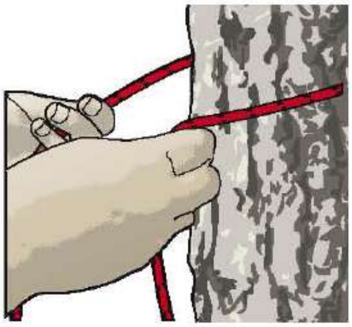

1 Pass the cord around the tree with the working line in your right hand and the standing line in your left hand.

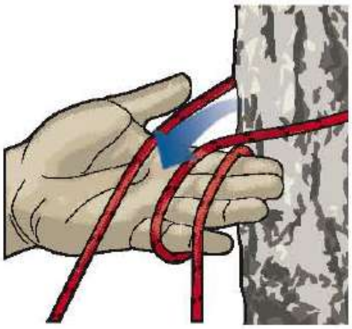

2 Lay the standing line on the palm of your left hand.

- Loop the working line one and a half times around your fingers.

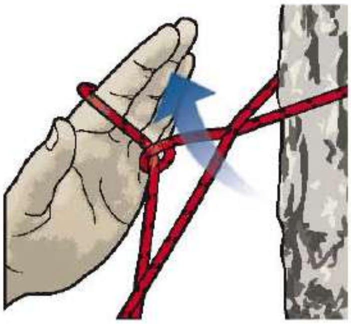

3 Bring your left hand under the standing line.

- Turn the palm of your hand so that one loop twists over another.

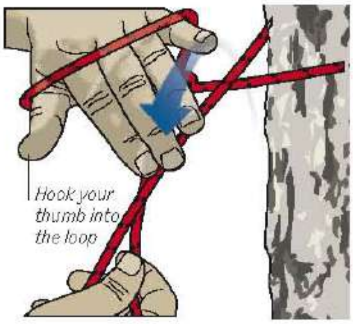
Hook your thumb into the loop

4 Hold the working end in your right thumb and forefinger.

- Bring the loop that is around your left fingers over the standing line.

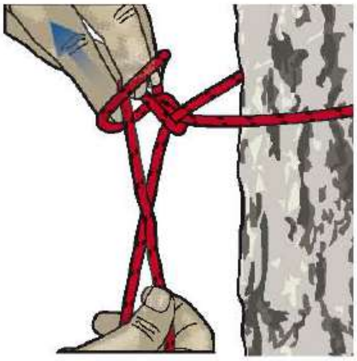

5 Pinch the working line with the fingers of your left hand.

- Pull it up through the loop to tighten the knot on itself.

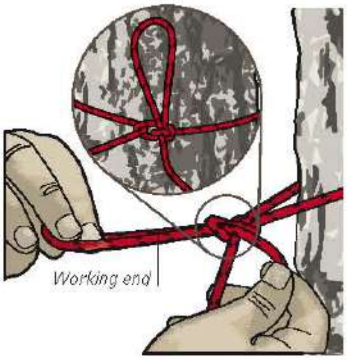

6 Pull the working end again, to tighten the knot around the tree.

- To free the knot, pull the end of the standing line.

TAUT LINE HITCH

This knot is widely used in survival and outdoor activities because it can be adjusted to increase or decrease the tension in a fixed line—for example, a guide rope on a tent or a mooring on a boat in tidal waters.

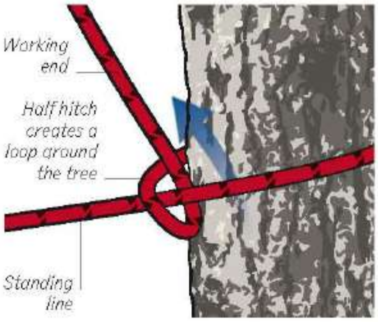
Working end
Half hitch creates a loop around the tree
Standing line

① Attach the standing line to an anchor point, such as a tent, and pass the working line around a tree.

- Tie a half hitch (see below) with the working end around the standing line.

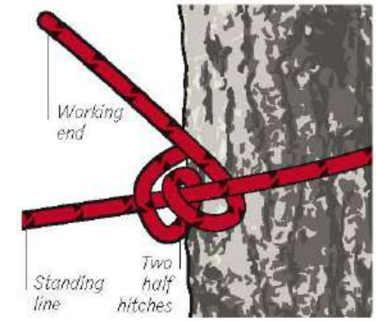
Working end
Standing line
Two half hitches

② Tie another half hitch next to the first one, so that the two half hitches sit side by side around the standing line.

- Don't tighten the half hitches yet.

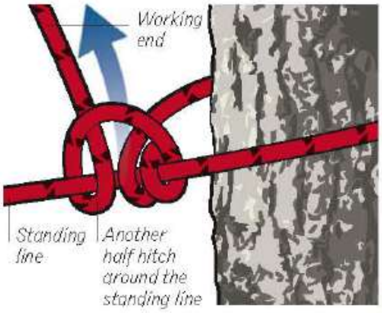
Working end
Standing line
Another half hitch around the standing line

③ Pass the working end under the standing line and back through the new loop it created (but outside of the loops in Steps 1 and 2).

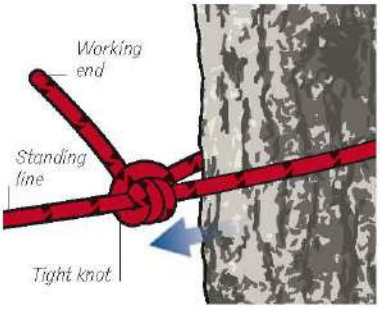
Working end
Standing line
Tight knot

④ Tighten the knot by pulling both the lines.

- Pull the knot further from the anchor point to increase the tension.

QUICK KNOT RELEASE

You can dismantle a tent or free a boat from its mooring more quickly by finishing off the taut line hitch (see left) with a quick-release knot.

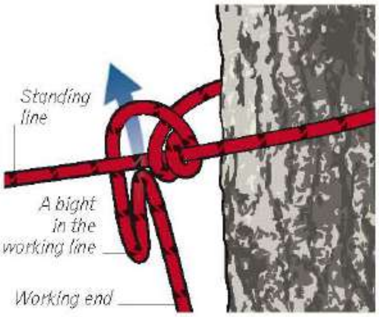
Standing line
A bight in the working line
Working end

① Repeat Step 3 of the taut line hitch (see left), but this time make a loop, known as a "bight," in the working line and pass it—rather than just the end—through the loop.

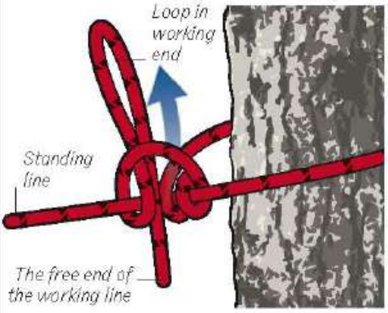
Loop in working end
Standing line
The free end of the working line

② Tighten the knot as in Step 4 (see left).

- Leave the working end free so you can pull it quickly to release the knot.

HALF HITCH

The half hitch is a simple knot that forms the basis for other knots (see taut line hitch, above). It's not a secure knot and, when used alone, it tends to slip. You can double or treble the knot to remedy this. Two half hitches under tension can be difficult to undo without cutting the cord, especially after a few days, if the knot has dried out after being wet.

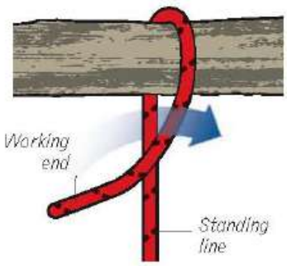
Working end
Standing line

① Pass the cord around the fixture, such as a bar or tree.

- Pass the working end around the standing line and back through the loop.

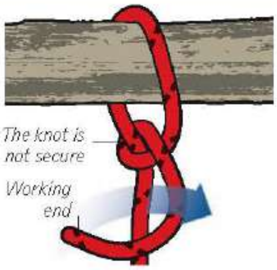
The knot is not secure
Working end

② Hold the standing line and push the half hitch knot toward the fixture to tighten it.

- Repeat the sequence and tie another half hitch.

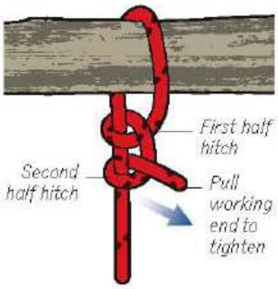
First half hitch
Second half hitch
Pull working end to tighten

③ Push the second knot tight to the first hitch knot for more security.

- If you need to, finish off the working end with an overhand knot (see p. 143).

USING CUTTING TOOLS

ANYONE VENTURING INTO THE WILDERNESS should carry some kind of knife, preferably a bushcraft knife. It's probably the most important survival tool after knowledge. Although a good knife will help you through many survival situations, look at the environment you are traveling to and choose your cutting tools according to their potential uses—for example, you may need a saw as well as a knife.

WARNING!
If you drop a bushcraft knife, never attempt to catch it because there's a very good chance you will lose your fingers!

BUSHCRAFT KNIFE

An experienced person can accomplish most survival skills, such as making feather sticks (see pp. 120–21), with a quality bushcraft knife. This type of cutting tool has a number of crucial features that will stand you in good stead in every situation.

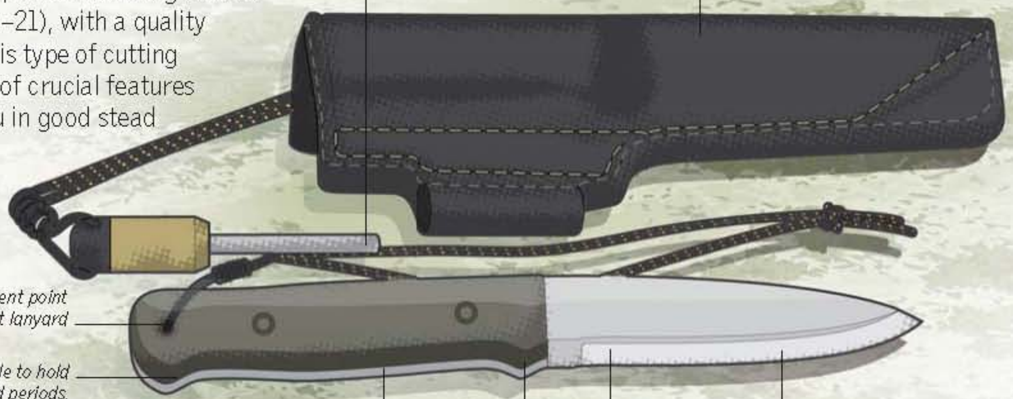

A firesteel that fits into your sheath means that, as long as you have your knife, you can also get a fire going

Fitted leather sheath holds the knife securely

Attachment point for a wrist lanyard

The handle is comfortable to hold and work for extended periods. Avoid ribbed, rubber handles

Blades that extend through to the end of the handle (full tang) have no weak points

Sharpened blade ends near the handle – working close to the handle puts less strain on your wrist

Blade is 3/16 in (4 mm) thick and made of stainless or high-carbon steel

The guard is a ridge in the handle that reduces the chance of your hand slipping on to the blade

LARGE KNIVES

All over the world, people living primitive lives have two essential items: a cooking pot and a large working blade, either a parang or machete.

- Parangs and machetes are similar types of knife. The parang is associated with the Malay people and there are many varieties available. Machetes are more common—you can still see machetes being used in Africa today that were made 40 years ago from old Landrover suspension springs! The kukri is a similar knife used by the Ghurkas of Nepal.
- A parang or machete blade is about 18–20 in (45–50 cm) long and is used primarily for cutting and slashing but, like the ax, can be used for intricate work.
- Always wear a wrist lanyard when using a machete, parang, or kukri—and pay attention to what you are doing.

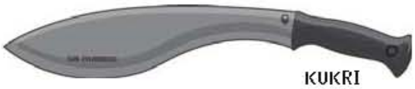

KUKRI

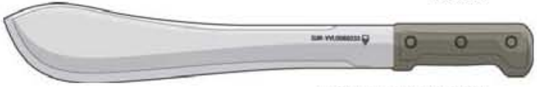

MACHETE/PARANG

KNIFE KNOW-HOW

- A bushcraft knife is used primarily for carving, cutting, and splitting small logs. If you need to chop large logs, use either a machete or a saw.
- A knife doesn't need to be longer than 8–9 in (20–22.5 cm).
- Don't use a knife with a serrated edge when working with wood, as the serrations make it hard to use the part of the blade closest to the handle (essential for making feather sticks, etc).
- Keep your knife safe when not in use. Clean and dry the blade and return it to its sheath.
- Wear your sheath on your belt or around your neck on a lanyard. Belt sheaths may be high (the handle is above the belt) or low (the handle is below the belt, which allows you to sit down without the handle poking you in the side).

SHARPENING A KNIFE

It's very important to keep the blade of your knife sharp. Blunt blades are dangerous because they are likely to bounce or slip off the wood you're working on. Also, blades that have completely lost their cutting edge are very difficult to sharpen.

SHARPENING STONES

Keep a good sharpening stone at home to ensure you always set out with a knife that is sharp.

- In the field, carry a smaller, lightweight device for quickly re-sharpening a blade that is beginning to lose its edge.
- There is a good range of small sharpening stones (medium one side, fine the other) that fit into a neat leather pouch.
- Small automatic sharpeners are available to buy that use two sets of ceramic stones to sharpen a knife pulled between them.

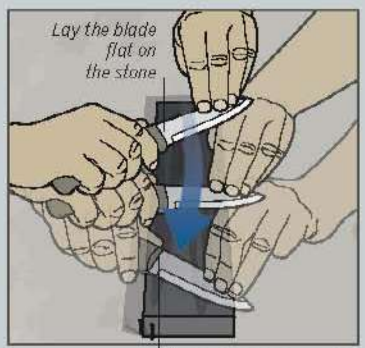

Lay the blade flat on the stone

Raise the back of the blade to lay the edge on the stone

1 Hold the knife in a forehand grip, with the blade pointing away from you. Support the tip of the blade with the thumb and fingers of your other hand.

- Push the knife so the full blade arcs over the stone in one sweep.
- Lift the blade and return to the start, repeating the action 6–8 times.

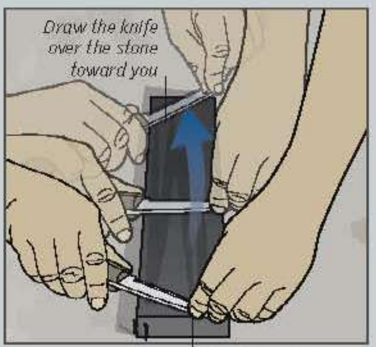

Draw the knife over the stone toward you

Start from the far end of the stone

2 Repeat the sequence for the other side of the blade, but reversing the crucial elements.

- Hold the knife with a backhand grip, with the blade pointing toward you at the far end of the stone.
- Pull the knife in an arc over the stone in one draw. Lift and return to the start, repeating the action 6–8 times.

WORKING WITH YOUR KNIFE

The way you hold your bushcraft knife can help you accomplish a variety of cutting tasks easily and safely. There are several essential grips, including the forehand, backhand, and chest lever. In most cases, you need to grip the handle firmly at the end nearest the blade so that you can cut with the part of the blade nearest the handle.

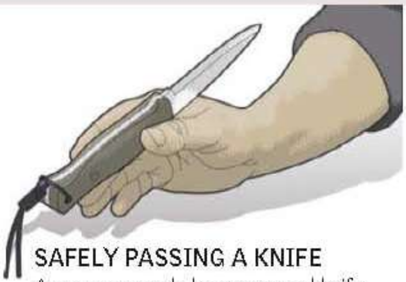

SAFELY PASSING A KNIFE

A group may only have one good knife for everybody to use, so you'll need to pass it to each other frequently. Pass the knife handle first, with the blade up. This simple and safe method is easily learned and can be adopted by everyone.

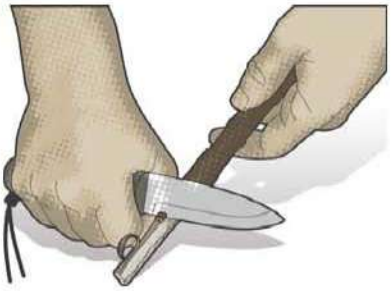

FOREHAND GRIP

This grip is the most natural and the one most often used. It allows your back and arm to provide the power and strength, while your wrist provides the control. Placing one or both thumbs on the back of the blade allows you additional control for finer, more detailed work. Make sure the blade is facing down. Otherwise you are likely to give yourself a serious cut.

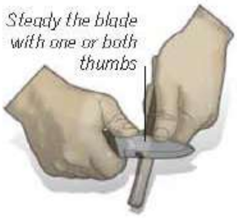

Steady the blade with one or both thumbs

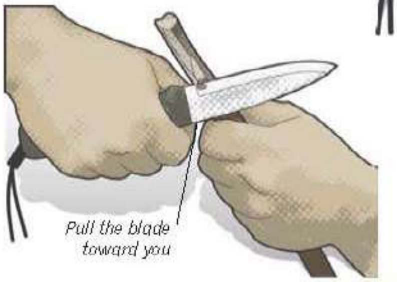

Pull the blade toward you

BACKHAND GRIP

This is the same as the forehand grip, except the blade faces up toward you. You can use it when you want to see and control exactly where the blade is going, or when you're cutting something and a follow-through would injure you—for example, when sharpening a stick.

CHEST LEVER

This is an extremely effective technique in which you hold the knife in a backhand grip and use the muscles of your arm and chest to provide power and strength. You can cut close to your body while your wrists control the cutting action.

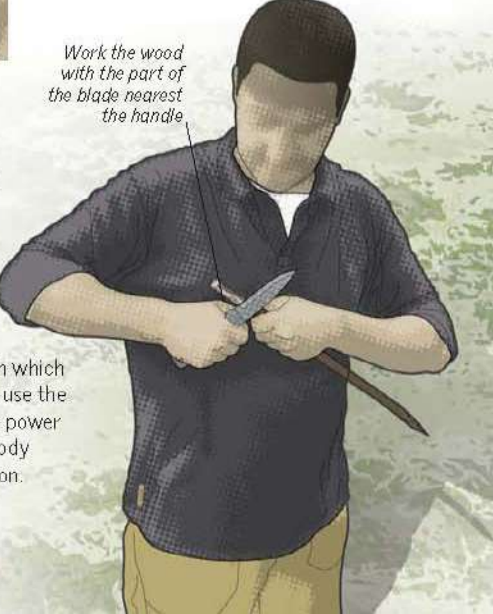

Work the wood with the part of the blade nearest the handle

TREE-FELLING WITH A BLADE

For shelters and fuel, the largest tree you are likely to cut down will be about 4 in (10 cm) in diameter. For anything larger you'd need an ax, but for smaller ones use a saw (see pp. 150–51). If you don't have a saw, use the blade of your bushcraft knife or a machete or parang. Whichever blade you use, keep its lanyard around your wrist.

1. Figure out the lean of the tree and start your first cut on the opposite side. This stops your blade jamming in the cut as the tree leans further.

- Hit the top of the blade with a thumping stick (club).
- Move the blade about 5/8 in (15 mm) to one side of the first cut, and then angle the blade at 45 degrees toward the first cut.
- Continue to alternate this action from side to side, cutting a V into the wood.

WARNING!
Ensure your blade is sharp. Keep what you're cutting between you and the blade, with the blade pointing away.

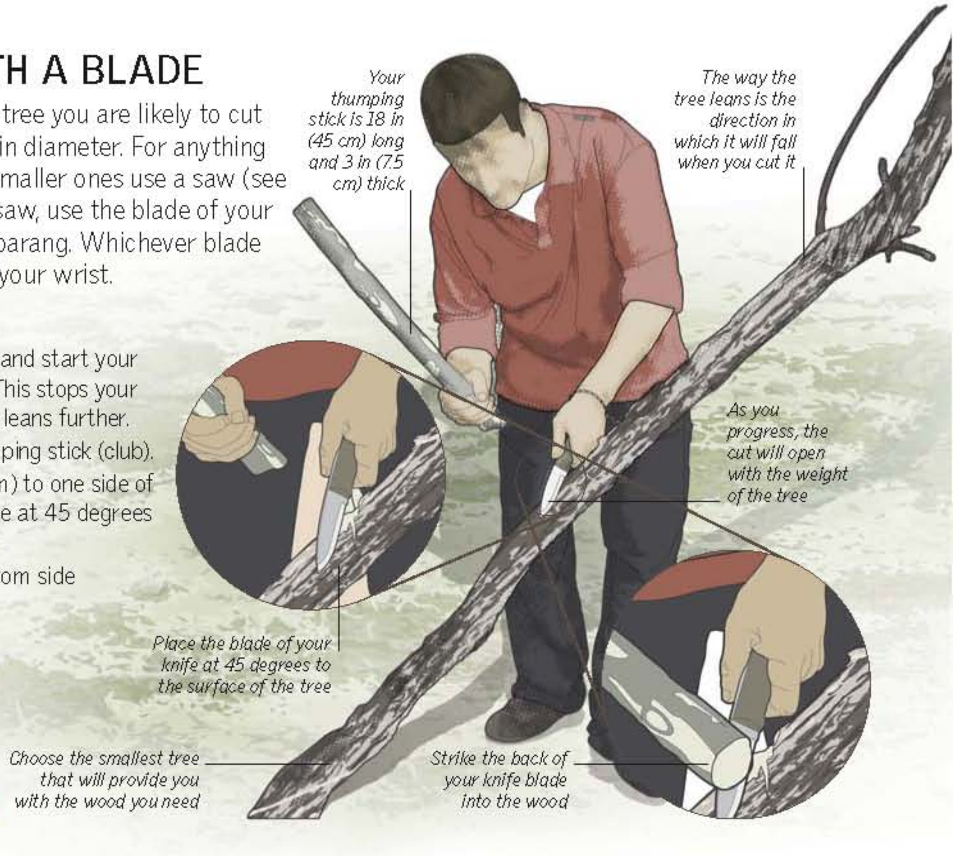
Your thumping stick is 18 in (45 cm) long and 3 in (7.5 cm) thick
The way the tree leans is the direction in which it will fall when you cut it
As you progress, the cut will open with the weight of the tree
Place the blade of your knife at 45 degrees to the surface of the tree
Choose the smallest tree that will provide you with the wood you need
Strike the back of your knife blade into the wood

SPLITTING FIREWOOD

When you split wood into fuel for a fire, use the largest blade you have, work on a solid surface, and avoid wood that has large knots as these prevent the blade from making a clean cut.

SAFE WORKING AREA
Chop wood in a safe area that other people won't wander into and where there are no snagging hazards or objects to impede your swing and knock your knife off-target.

LARGE LOGS

If you only have a bushcraft knife to work with, and you need to split large logs into smaller pieces, the following method works well.

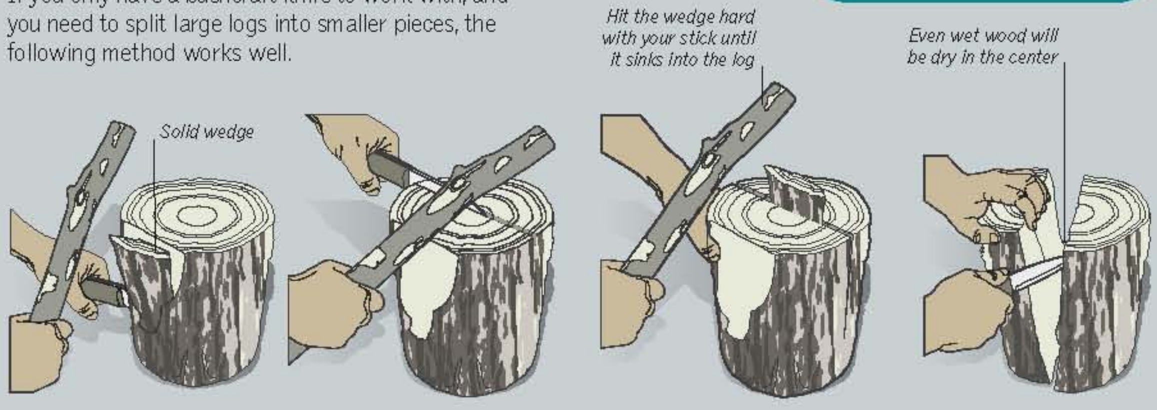
Solid wedge
Hit the wedge hard with your stick until it sinks into the log
Even wet wood will be dry in the center

1. Hit the back of your knife with a stick and cut a solid wedge from the side of the log.

2. Use your knife and a thumping stick to start the initial cut across the top of the log.

3. Hit the wedge into the initial cut. Keep doing this until the log splits. Use more wedges if you need to.

4. If the log does not split completely, force it open with your hands or pry it with your blade.

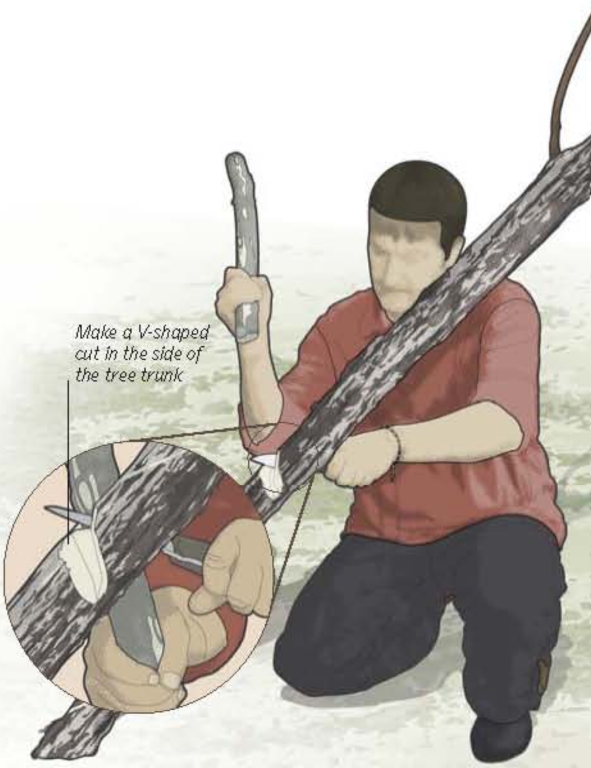

Make a V-shaped cut in the side of the tree trunk

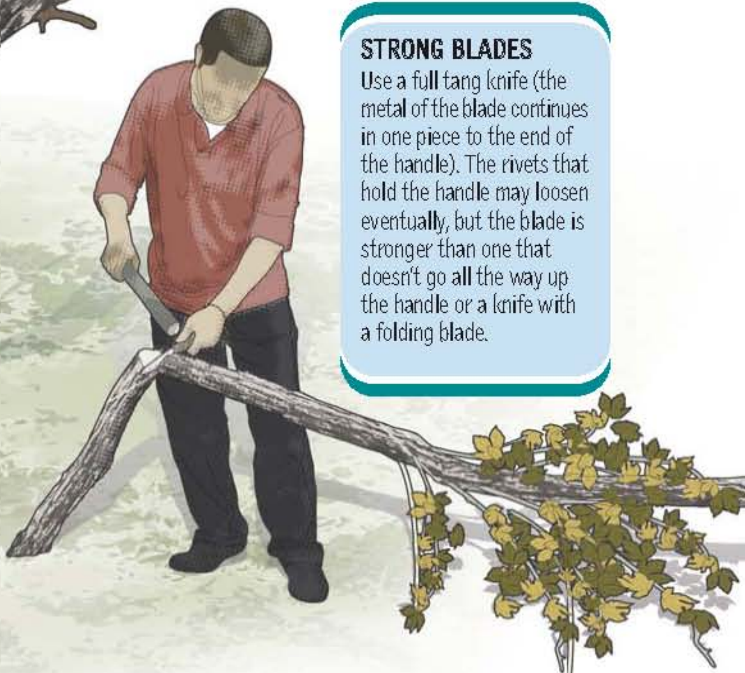

STRONG BLADES

Use a full tang knife (the metal of the blade continues in one piece to the end of the handle). The rivets that hold the handle may loosen eventually, but the blade is stronger than one that doesn't go **all** the way up the handle or a knife with a folding blade.

2 Work the cut around the tree until you've almost reached the middle from every direction except from below.

- If you try to cut completely through the wood from one side it will get increasingly difficult as the blade goes deeper.

3 Once the cuts are deep enough to weaken the wood, you will be able to snap the tree in two.

- Take care that the tree doesn't fall on your feet.

SMALL LOGS

Splitting a small log can be more tricky than splitting a large log, because it's difficult to balance it safely while you hit the blade with your thumping stick.

1 Place the log on a solid surface and put your knife in the middle of the end of the log you're about to chop.

- Hit the back of the blade firmly with your stick.

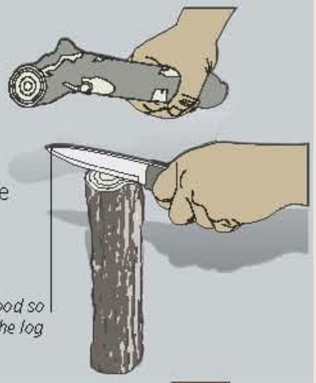

Put the blade in the wood so you can steady the log

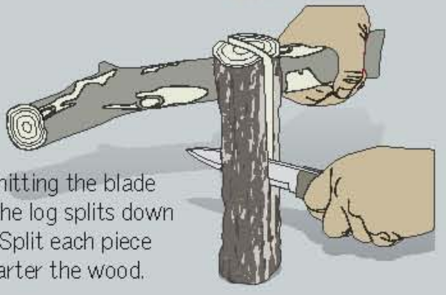

2 Keep hitting the blade until the log splits down the middle. Split each piece again to quarter the wood.

4 To trim the branches from the fallen tree, start at the bottom and work your way to the top. This allows you a clear, short swing for your thumping stick.

- Keep the tree between you and the blade, and trim each branch in the direction of its growth.
- If you're using a machete, you don't need a thumping stick, as the heavy blade has enough momentum of its own.

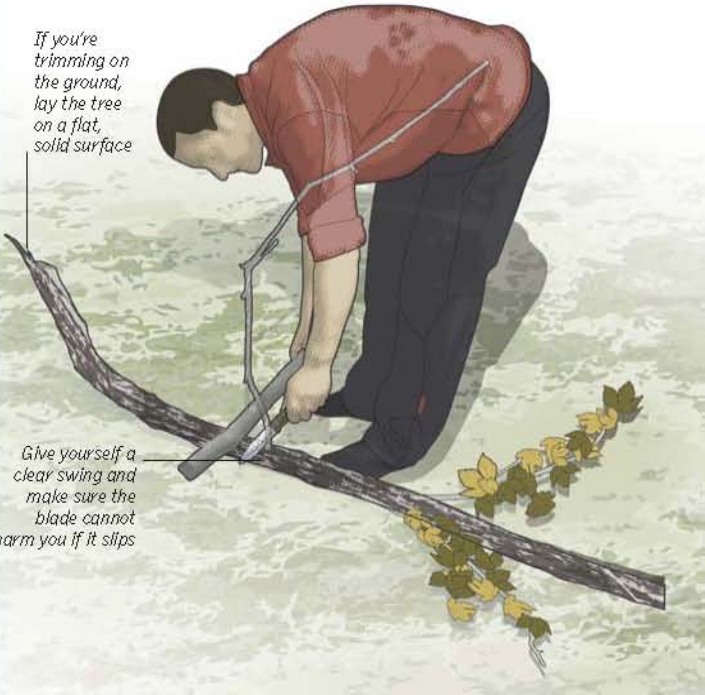

If you're trimming on the ground, lay the tree on a flat, solid surface

Give yourself a clear swing and make sure the blade cannot harm you if it slips

USING A SAW

A saw is much better than a bushcraft knife for larger jobs, such as preparing logs for a fire or cutting down the poles needed to build a shelter. There are several types of pocket saw available that are lightweight and take up very little room in your backpack or survival tin.

USING A POCKET CHAINSAW

Using a pocket chainsaw to cut down a small tree is safe and not especially hard work. You can use the same saw to trim the branches and then cut the wood into manageable sections.

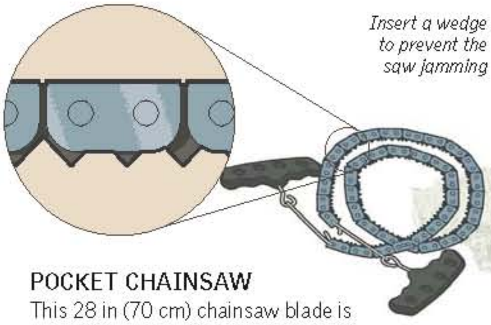

POCKET CHAINSAW
This 28 in (70 cm) chainsaw blade is compact and has attachment rings and handles. It can cut through a 3 in (7.5 cm) diameter tree in under 10 seconds.

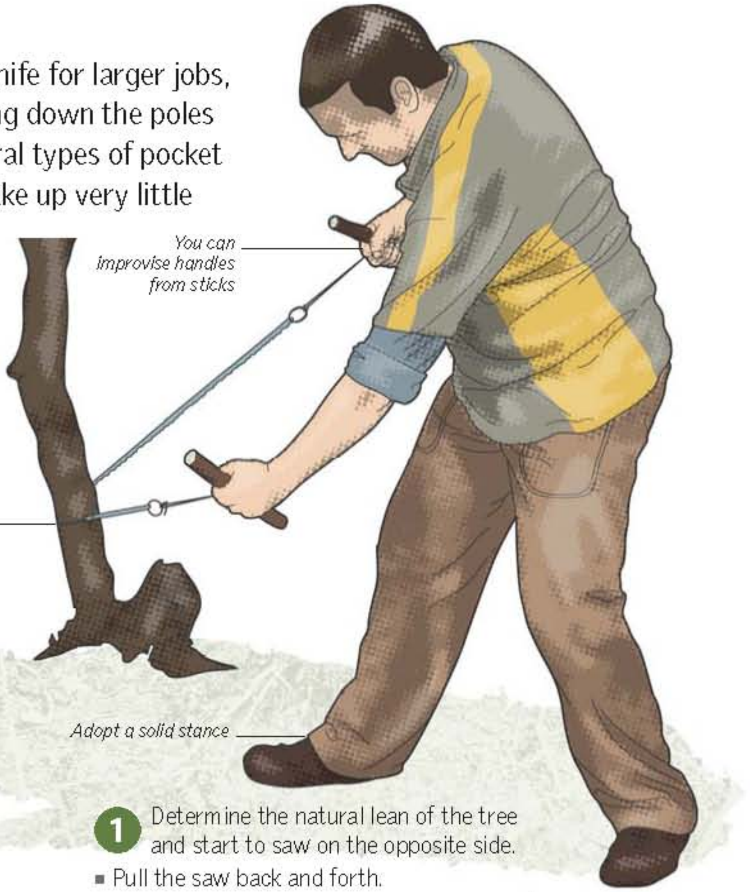
You can improvise handles from sticks
Insert a wedge to prevent the saw jamming
Adopt a solid stance

1 Determine the natural lean of the tree and start to saw on the opposite side.

- Pull the saw back and forth.
- Knock a wedge in behind the saw when halfway through to stop the tree from jamming the blade.

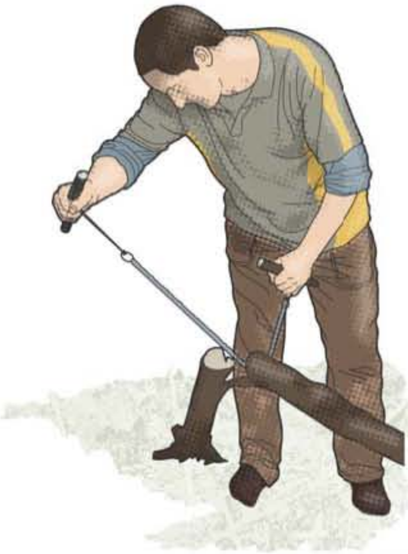

2 Slow down when the saw has nearly cut through the tree.

- Allow the weight of the tree to snap the cut and break it in two.

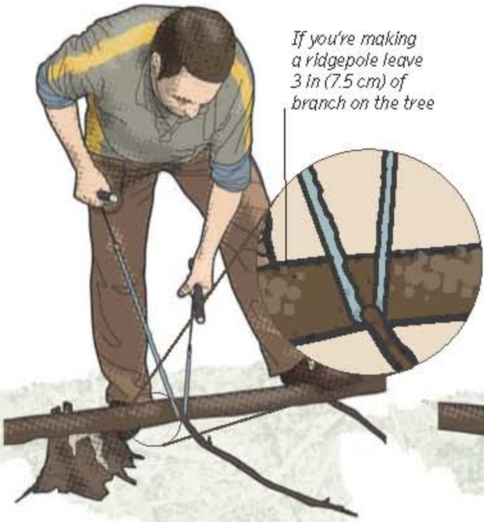
If you're making a ridgepole leave 3 in (7.5 cm) of branch on the tree

3 Use the same back-and-forth rhythm to trim the branches.

- Clear the branches away as you trim them from the tree.

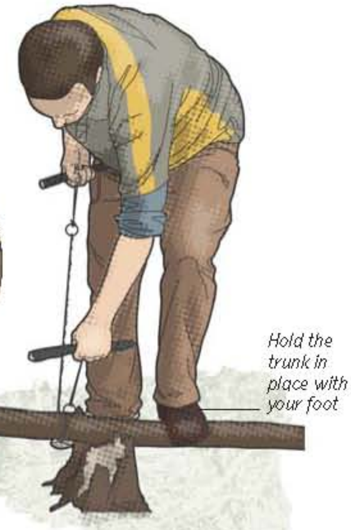
Hold the trunk in place with your foot

4 Saw the trimmed trunk into a measured section.

- When you're halfway through, prop up the trunk to make sawing easier.

FOLDING SAWS

Folding saws are strong, light, and convenient. Some, such as the Sandvik Laplander, have a serrated blade that folds into the handle (see right). Others, such as the backpacker's (see below), can be dismantled for stowing.

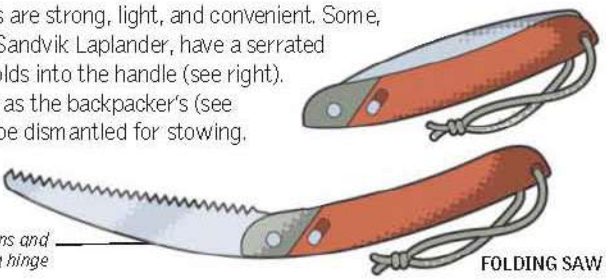

The blade opens and closes at a hinge

FOLDING SAW

BACKPACKER'S FOLDING SAW

These saws are generally lightweight and strong, but are not as effective as the pocket chainsaw. If space is tight, then you can pack only the blade and improvise a bow saw, using natural materials found in the woods.

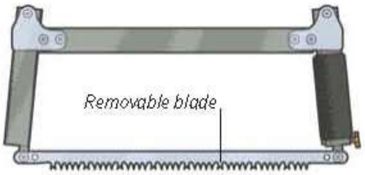

Removable blade

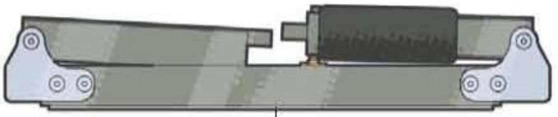

The tension frame and saw blade can usually be taken apart and stowed safely

BACKPACKER'S FOLDING SAW

COMMANDO SAW

This saw takes up very little space and consists of a serrated wire blade with a swivel ring at each end. It's a useful piece of survival equipment, although it tends to snag and snap easily unless you keep it completely straight. You can easily convert it into an improvised bow saw, and, with a little more effort, make it into a very effective tension saw.

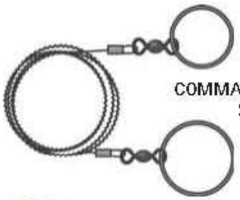

COMMANDO SAW

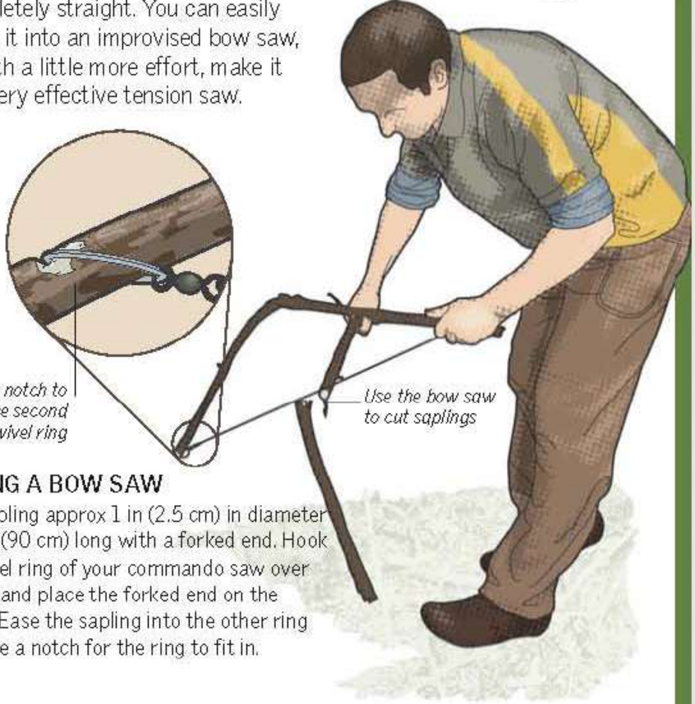

Make a notch to secure the second swivel ring

Use the bow saw to cut saplings

MAKING A BOW SAW

Cut a sapling approx 1 in (2.5 cm) in diameter and 3 ft (90 cm) long with a forked end. Hook the swivel ring of your commando saw over the fork and place the forked end on the ground. Ease the sapling into the other ring and make a notch for the ring to fit in.

IF YOU HAVE NO TOOLS

Even without any cutting tools, it's still possible to safely break wood down to size. The answer is snapping—or even burning—the wood. Don't be tempted to stand branches and sticks against a log or rock and then stamp on them in the middle. In a genuine survival situation your feet are your only means of transportation so don't do anything that risks injuring them.

SNAPPING THE WOOD

You need to limit the size of wood you're trying to break—don't try to snap branches that are too broad. Use wood that's dead, because green living wood just bends and refuses to break completely.

- Find two trees that are 1–2 ft (30–60 cm) apart—or a tree with two trunks close together—to use as a fulcrum for the wood. Place the wood in the fulcrum.
- Adopt a firm and steady stance and, with one foot in front of the other, pull the wood toward you until it breaks. Pulling toward you keeps your center of gravity over your spread feet, while pushing away from you causes you to stumble forward when the wood breaks.

Hold firmly on to the piece of wood

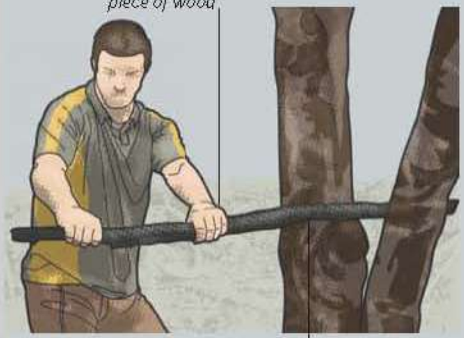

Place the wood between two trees—one is a stop or buffer, the other is a leverage point

BURNING THE WOOD

If you have a fire going, place the wood on the fire where you want it to break. Leave it on the fire until it burns all the way through. Alternatively, when it has burned halfway through, use the snapping method described above—but don't work on the wood while it's still burning or hot.

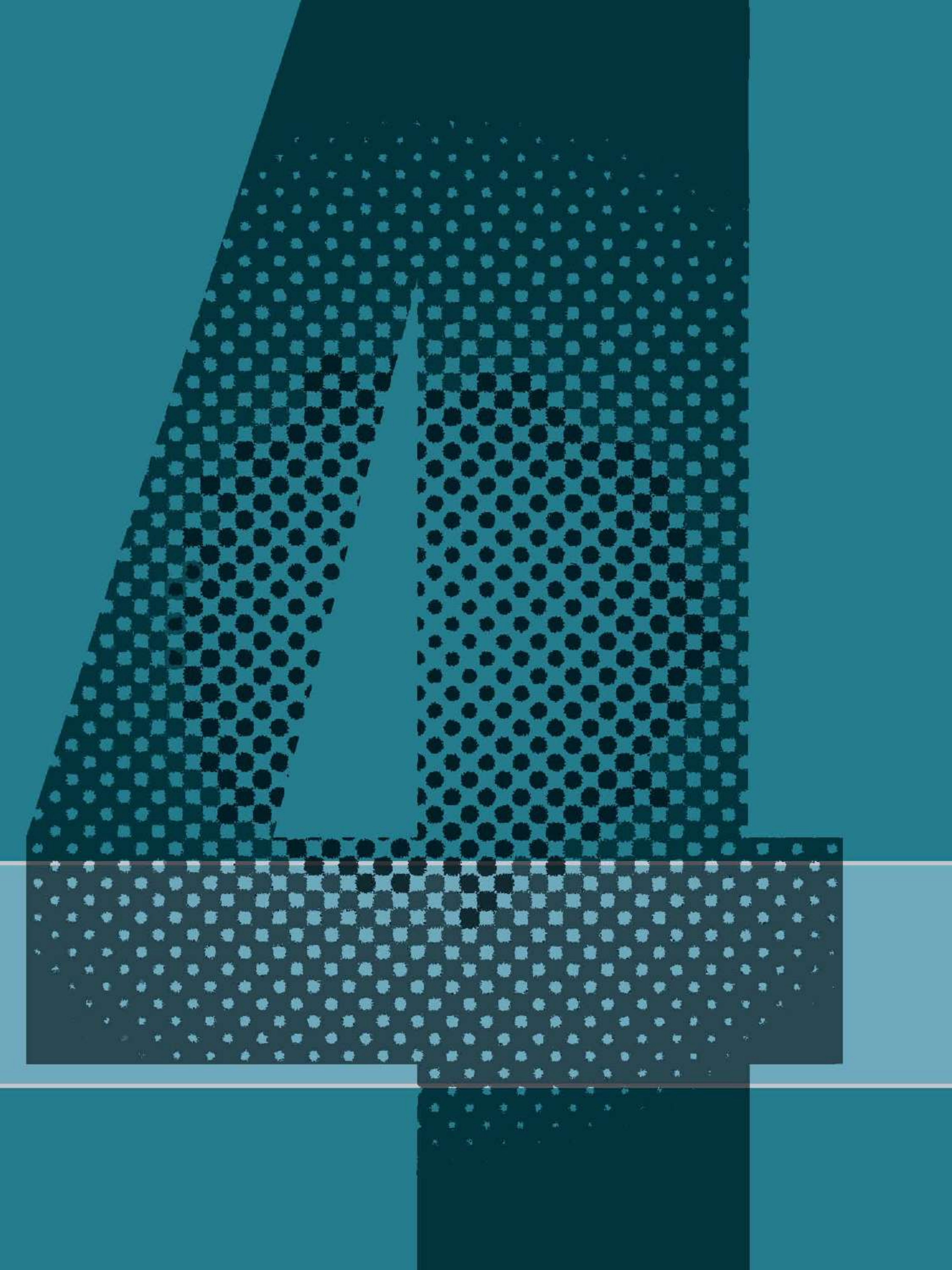

TAKING SHELTER

TAKING SHELTER

A shelter is your primary means of protection in a survival situation. It can be somewhere to simply keep dry during a sudden downpour or a place to spend several nights while out in the wilds. It's important, therefore, that you understand how to correctly construct your shelter. After all, its effectiveness could mean the difference between relative comfort and abject misery!

In a survival situation, it's important to play a proactive role in your own continued survival and rescue. You can only achieve this effectively if you're in a good condition—both physically and mentally. A day or night spent exposed to the elements will determine how well or how badly you perform the next day. A restless and uncomfortable night, for example, will lead to sleep deprivation, which can make you moody, irritable, and easily frustrated. This in turn can lead to a lack of concentration and irrational thinking—none of which are helpful in a survival situation.

Cold nights spent in the wilderness seem to last longer than they should, so time spent ensuring that you can

In this section YOU WILL DISCOVER...

- when to **lie low** in a hollow or **camp in a cave...**
- how to **modify** your **poncho** to put up a **shelter...**
- the difference between a **bothy** and a **bivy...**
- how to **bed down** with **duck down** or **wake up** in a **wickiup...**
- the **importance** of using your **noggins** to **secure** your shelter...
- how to **configure** your **fig leaves** to keep out the **rain...**

always enjoy as warm, dry, safe, and comfortable a night as possible is rarely wasted.

With shelter-building—as with other areas of survival—you should always apply the principle of expending the least amount of energy for the maximum amount of gain. This could mean simply exploiting the options nature has already provided, such as caves or hollows. Finally, plan for the worst conditions—it may be sunny when you build your shelter but pouring with rain at three in the morning!

On any expedition, you should aim to carry equipment—such as a basic bivy or a tarpaulin—that could be used to form a shelter should you need to protect yourself from the elements.

A SPACE BLANKET, however, is quite simply an essential piece of gear—if you pack nothing else, make sure you carry one and keep it with you at all times. Space blankets pack down very small and are available with one side silver (to reflect heat) and one side bright orange (to aid location). No matter what environment you're exploring, a space blanket can offer you immediate protection from the elements.

With a little ingenuity, a space blanket can also be converted into a basic shelter. You will find instructions on how to build a variety of shelters throughout this chapter, but always remember the following general tips, too:

- First look for natural sheltering places
- Look out for hazards such as signs of flooding
- Construct your shelter well before nightfall
- Clear the ground before you start to build
- Position the entrance at right angles to the wind
- Build for the worst possible weather conditions
- Make sure your shelter is sturdy and secure
- Don't over-exert yourself as you build
- Raise your sleeping area off the ground
- If on a slope, sleep with your head higher than your feet

> “The **effectiveness of a shelter** can mean the difference between relative comfort and **abject misery**”

HOLLOWS AND CAVES

IF YOU DON'T HAVE a shelter of your own, look for shelter opportunities that nature has already provided for you, such as a cave or a natural hollow in the ground. However, avoid natural hollows in low-lying areas that may flood and, if you're on a slope, keep well away from hollows exposed to run-off water.

TOOLS AND MATERIALS

- Sturdy digging stick
- Strong branches (or poles) and logs
- Saw, such as pocket chainsaw
- Tree bark, turf, leaves, forest mulch, moss, bracken, and pine boughs for the roof

BEDDING

Cover the floor of a shelter with some form of bedding to prevent your body from losing heat to the cold ground through conduction. After all the energy you've used to build the shelter, why waste it on an uncomfortable night? Use the driest materials you can find (see list below).

MAKE YOURSELF COMFORTABLE

Here are some ideas to help you enjoy a good night's sleep in your shelter:

- Collect twice as much bedding as you think you need: 6 in (15 cm) of holly, for example, compresses to about 2 in (5 cm) after you've been lying on it for a while.
- Place logs around the bedded area to stop the bedding from shifting and spreading.
- If the ground slopes, sleep with your head higher than your feet.

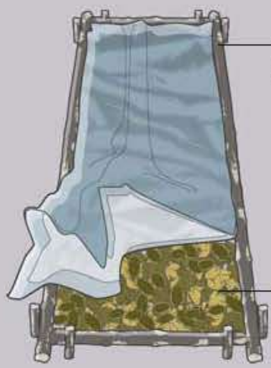

Use wooden pegs to keep the side logs in place

Bedding keeps you warm

BEDDING MATERIALS

- Feathers, such as duck and goose down, are the best bedding as they retain warmth but not moisture.
- You're unlikely to find enough feathers, so also try pine and spruce boughs, dry leaves, moss, bracken, grasses, and holly (cover it to prevent it from pricking you).

NATURAL HOLLOWS

A hollow that may be too shallow to sit in can still provide shelter from the wind if you lie in it—and, with a basic roof, keep you dry too.

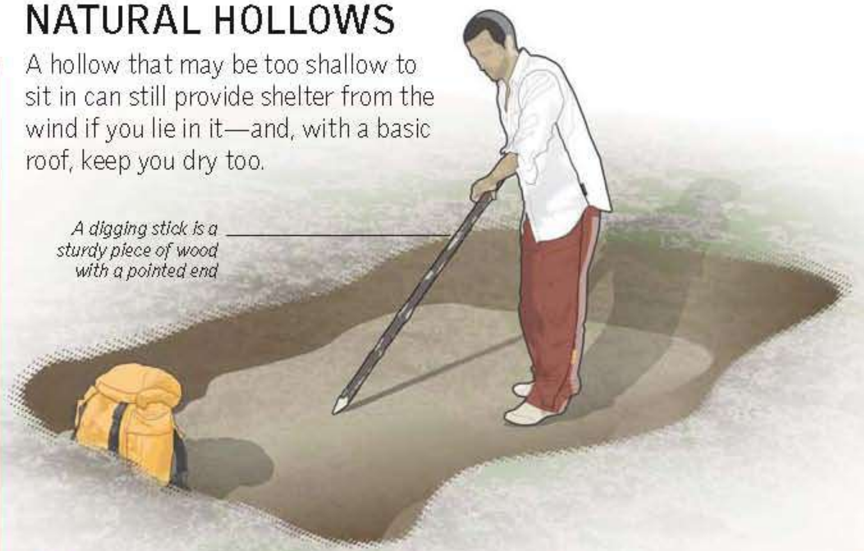

A digging stick is a sturdy piece of wood with a pointed end

1 If the hollow isn't deep enough to lie in, remove some soil with a digging stick.

- Beware wet soil, as it could be prone to flooding.
- If the ground is too hard to dig, build up the sides with logs for extra depth.
- Lay bedding material on the ground before continuing.

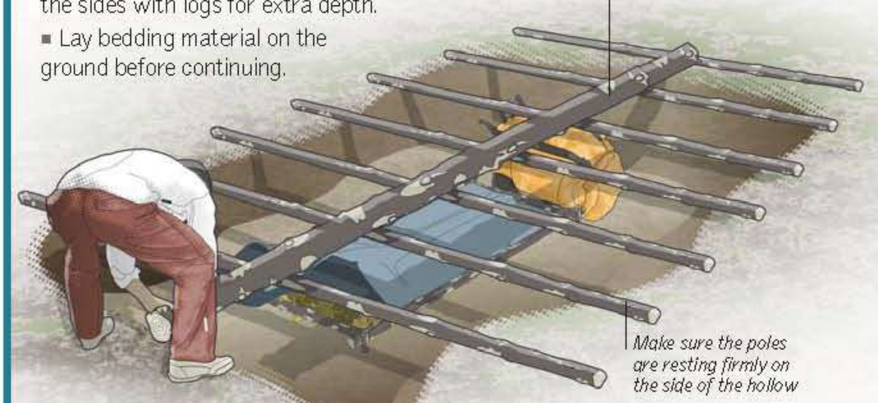

Lay a log on top of the cross poles and along the length of the hollow

Make sure the poles are resting firmly on the side of the hollow

2 With your saw, cut a number of sturdy poles or branches (see pp. 150–51) that are long enough to place across the hollow to create a supporting roof.

- Cut a thicker, longer log and place it across the poles. This will create height and slope for a pitched roof. Make sure the log isn't too heavy.

TAKING SHELTER IN CAVES

Caves are ready-made shelters that are usually dry and secure. There are some dangers, including animals, poor air quality, and water. Many caves are outlets for streams or are connected to pools. Don't go any further in than you can see, as there may be hidden drops, slippery surfaces, and crevasses. Avoid old and disused mines, as they may be prone to collapse.

ANIMAL DANGERS

Many animals, such as bears, bats, insects, spiders, and snakes, shelter in caves, so check the ground around the entrance and inside for signs of tracks and nests. Vampire bats can be dangerous as they are known to carry rabies. Bat droppings (guano) form a thick layer on the ground and, in large quantities, can be highly combustible.

POOR AIR QUALITY

If a cave makes you feel light-headed or nauseous, leave at once because the air may be foul due to excess carbon dioxide. Other warning signs are an increased pulse and breathing rate. If a flame starts to dim, or turn blue, leave the cave immediately—this may not be a sign of excess carbon dioxide but a lack of oxygen, which can be worse.

LIGHTING FIRES

If you build a fire at the front of a cave the smoke can blow back inside and the fire can block your exit if it gets too big. Build a fire toward the rear of the cave, but make sure that there is a sufficient flow of air for the smoke to escape. You can erect a small fire reflector or make a pile of rocks at the entrance to reflect heat back into the cave.

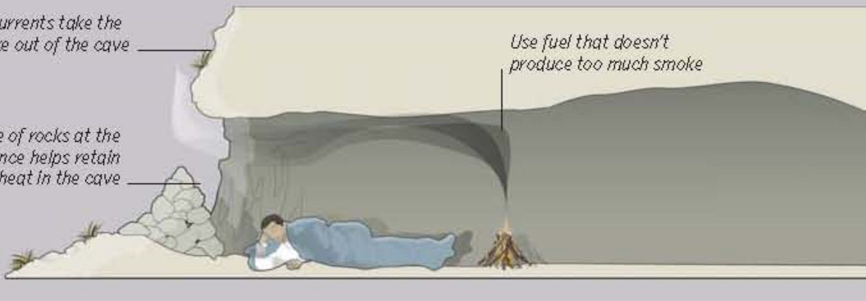

Air currents take the smoke out of the cave

A pile of rocks at the entrance helps retain heat in the cave

Use fuel that doesn't produce too much smoke

COASTAL CAVES

Coastal caves share the same dangers as any other cave, but they are also subject to the tides, which can rise very quickly—in some areas they can advance hundreds of yards in a matter of minutes.

THE DANGERS OF COASTAL CAVES

When you're assessing a coastal cave, look out for the following warning signs of flooding during high tide:

- A line of seaweed, driftwood, or flotsam and jetsam—either in the cave or further back on the beach either side of the entrance.
- A damp or wet smell. The cave may even look wet if the tide has recently gone out.
- Rock pools in the cave or in front of it.

3 Place shorter branches, sticks, or poles over the top of the log and cross branches to create the pitched roof.

- Pack these materials tightly together and make sure they're firmly in position.

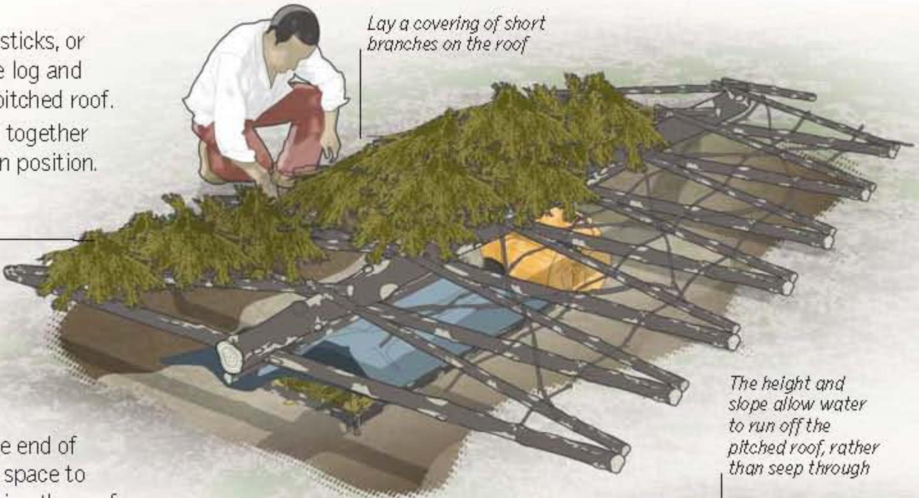

Lay a covering of short branches on the roof

The space between the support and pitched roof will help retain the heat generated inside the shelter

4 Leave an entrance at one end of the shelter with enough space to get in and out without disturbing the roof.

- Ensure the entrance doesn't face the wind—an angle of 90 degrees to the wind is best.
- Insulate the pitched roof with as much foliage as possible.
- Start the final insulation layer at the ground, working up to the top of the roof to create an overlap for rainwater to run off.
- In wet conditions secure your poncho or space blanket over the top.

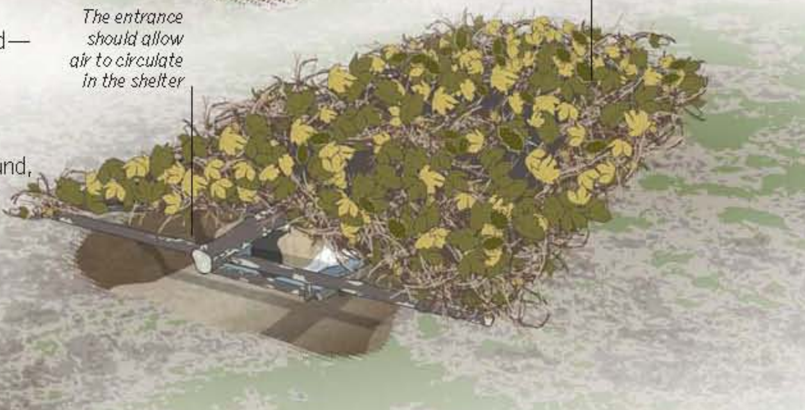

The entrance should allow air to circulate in the shelter

The height and slope allow water to run off the pitched roof, rather than seep through

USE-ANYWHERE SHELTERS

WITH A PONCHO OR BIVY SACK (see pp. 160–61), you can make a shelter for use in any environment and in any conditions. You can also adapt any type of material—a survival space blanket, ground sheet, or tarpaulin—in the same way.

TOOLS AND MATERIALS

- Poncho, bivy sack, groundsheet, space blanket, or tarpaulin
- Cordage or elastic bungees
- Pegs; stick or pole
- Rock or log for hammering
- Survival knife or penknife

ONE-POLE PONCHO

This shelter is very quick and easy to erect and take down—useful when energy levels are low or when you just want to get out of the rain. If you need to be rescued, put the most visible color on the outside to increase the chance of being spotted.

1 Cut three 2 ft (60 cm) lengths of cordage and one 6 ft (2 m) length.

- Tie the shorter lengths into loops using an overhand knot (see p. 143).
- Tie the long length to one corner with a Siberian hitch knot (see p. 144).
- Fasten the cordage loops to the corner grommets.

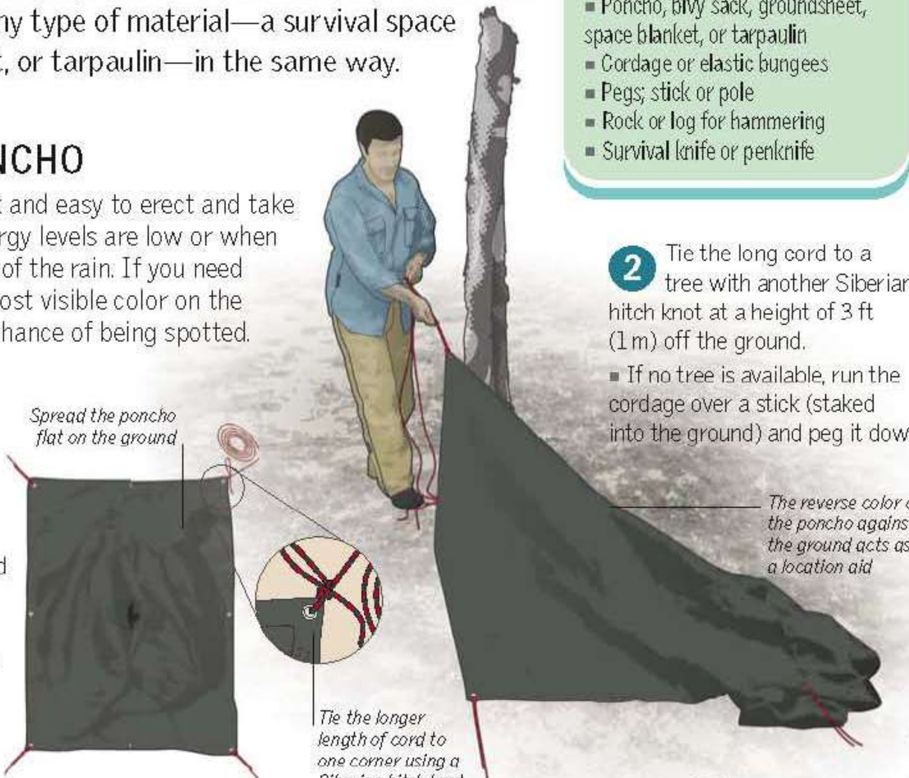

Spread the poncho flat on the ground

Tie the longer length of cord to one corner using a Siberian hitch knot

The reverse color of the poncho against the ground acts as a location aid

2 Tie the long cord to a tree with another Siberian hitch knot at a height of 3 ft (1 m) off the ground.

- If no tree is available, run the cordage over a stick (staked into the ground) and peg it down.

Make sure the top of the stick in the poncho is secure

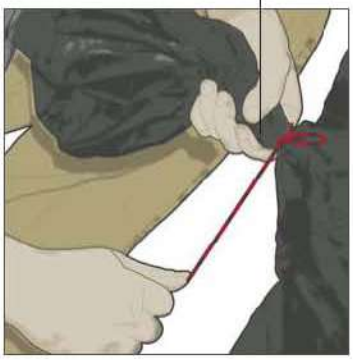

3 Place one end of a 3-ft (1-m) long stick in the hood of the poncho.

- Seal the hood by tying a simple knot.
- Wind several turns of cordage around the knot and secure it with an arbor knot (see p. 142).

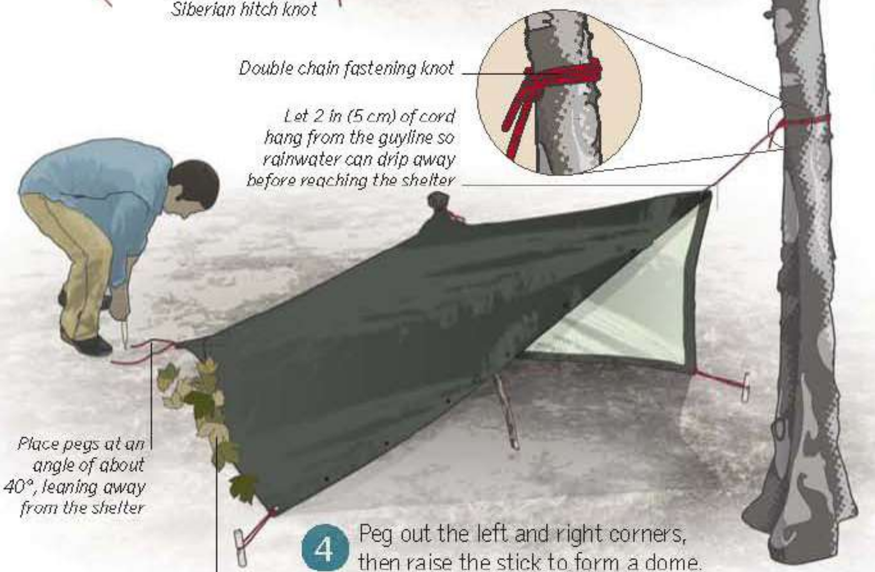

Double chain fastening knot

Let 2 in (5 cm) of cord hang from the guyline so rainwater can drip away before reaching the shelter

Place pegs at an angle of about 40°, leaning away from the shelter

Pile leaves around the edges of the shelter to keep out the wind

4 Peg out the left and right corners, then raise the stick to form a dome.

- Pull the rear corner taut and peg it out to add tension, making the structure rigid.
- To enter the shelter, curl your body around the center pole.

CORDED A-FRAME

This poncho A-frame creates a tentlike shelter between two trees. If you have enough cordage, a single ridgeline can be tied between the trees; if not, use two cordage loops, as shown below. Alternatively, you can use a ridgepole at the same height.

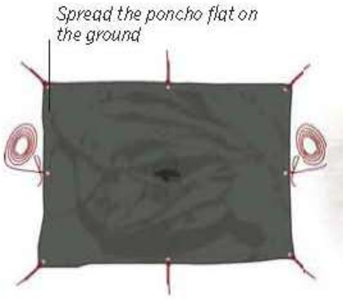
Spread the poncho flat on the ground

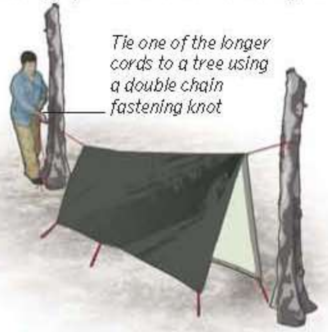
Tie one of the longer cords to a tree using a double chain fastening knot

1. Find two trees 2 ft (60 cm) further apart than the length of your poncho.

- Attach cordage loops 1 ft (30 cm) long to each grommet on the longer sides of your poncho and 3-ft (1-m) loops to the end ones.
- Tie a knot around the hood to seal it.

2. Tie the end of one of the longer cords to a tree with a double chain fastening knot (see p. 143) at a height of 3 ft (1 m).

- Tie the other long cord to the second tree with a taut line hitch knot (see p. 145) and adjust the hitch to pull it tight.

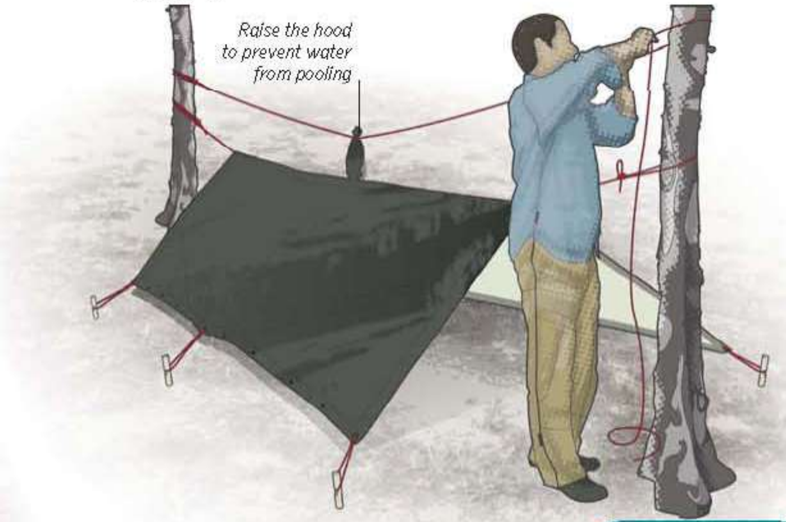
Raise the hood to prevent water from pooling

3. Peg the middle loop of each side to pull the poncho taut. Peg out the four remaining loops, pulling tight as you go.

- If you have enough cordage, raise the hood by tying it either to a second horizontal line or to a vertical line passed over a branch above.
- Place long boughs along each side of the shelter, with shorter boughs at each end to hold in bedding material.

NEED TO KNOW

- Block one end of the shelter with branches, boughs, and mulch to help retain heat.
- If your material has no grommets, use button ties instead (see p. 161).

TENT PEGS

You can make your own tent pegs and then either reuse them or replace them with new ones when necessary. Discard old pegs that naturally decompose over time.

MAKING TENT PEGS

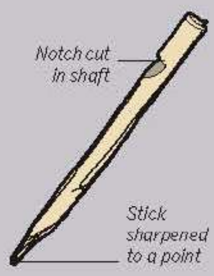
Notch cut in shaft
Stick sharpened to a point

- Pegs should be about 9in (22.5 cm) long and 1 in (2.5 cm) in diameter. Anything less will either snap under pressure or pull out in high winds.
- Cut wood for the pegs from green trees.
- Don't use wood from the ground as it will be in some stage of decay.
- Dry out new pegs slowly over the embers of a fire to harden them. Remove them when they're a light brown color. Wet wood sounds deep and dull when you tap it; as it dries the sound gets higher and crisper.
- Never drive a peg in with your foot. Use a stone or a heavy stick instead. If you misjudge it, the peg may stab you through the ankle or pierce your boot and even your foot.

ALTERNATIVE DESIGN

If you're in a group and you each have a poncho, you can clip two ponchos together and use one as a built-in floor.

USING TWO PONCHOS

First clip the poppers of both ponchos on their two long sides to form a tube. The top one is the A-frame, the bottom one the groundsheet.

- Pack all your bedding material under the bottom poncho and hold it in place by laying branches at either end.
- Assemble the shelter with the top poncho in the same way as the single poncho version.

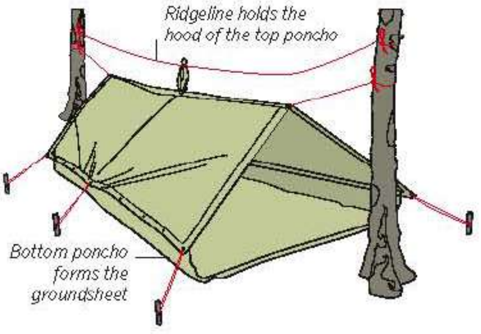
Ridgeline holds the hood of the top poncho
Bottom poncho forms the groundsheet

BIVY PUP TENT

A bivy sack is a lightweight, waterproof cover for a sleeping bag and is made out of a breathable material that cuts down condensation. It's an essential component of ultralight camping (see p. 52) and can very easily be turned into a single-person pup tent. The advantage of making a tent from the bivy sack is that, in wet conditions, the rain runs off the side and doesn't seep to the inside as quickly as it does when the bag is flat on the ground.

TOOLS AND MATERIALS

- Bivy sack
- Saplings
- Button ties
- Wooden pegs
- Cordage and knife
- Rock or log for hammering

1 Lay the bivy sack flat on the ground where you intend to set it up.

- If you have the flap of the hood on top it will form a door that hangs down. However, if the flap is on the bottom you can tie it up as a door.
- Attach a button tie (see panel, opposite) with a loop to each of the bottom corners of the bivy sack.
- Peg the button-tie loops into the ground.

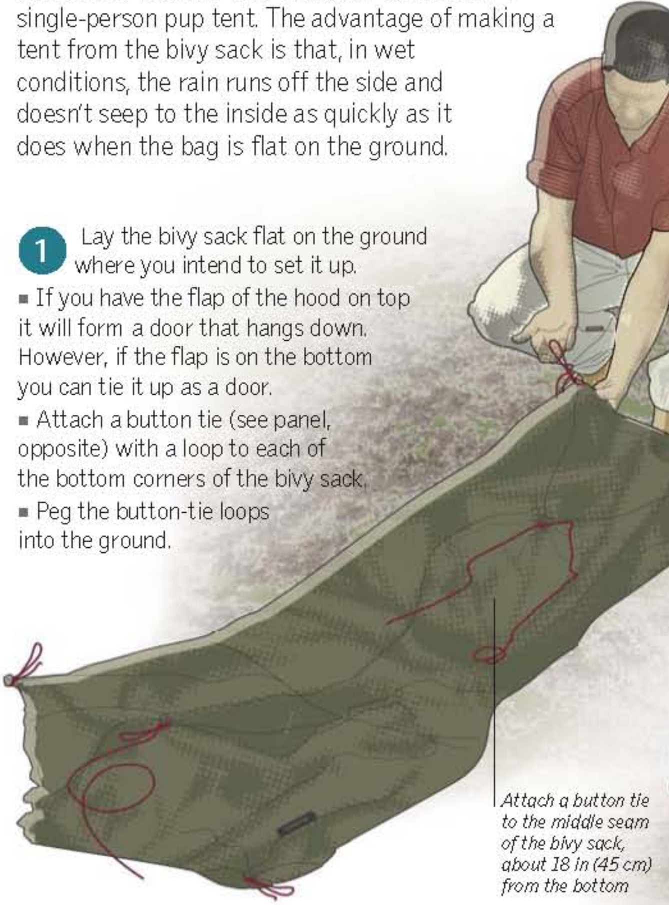

Attach a button tie to the middle seam of the bivy sack, about 18 in (45 cm) from the bottom

2 Attach a button tie with a long length of cordage to the center of the bivy sack's opening and hold it up.

- Pull down the right and left sides of the opening and place a button tie and loop where they meet the ground.
- Peg out the button-tie loops into the ground.

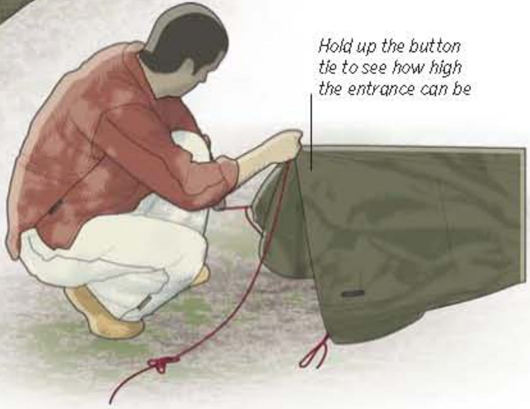

Hold up the button tie to see how high the entrance can be

4 Cut a pole the same height as the entrance (you could use a hooped sapling or crossed sticks instead).

- Stand the pole at the entrance, run the middle button-tie cord over the top, and peg out the cord tightly.
- Readjust all the pegs to ensure tight and even tension.

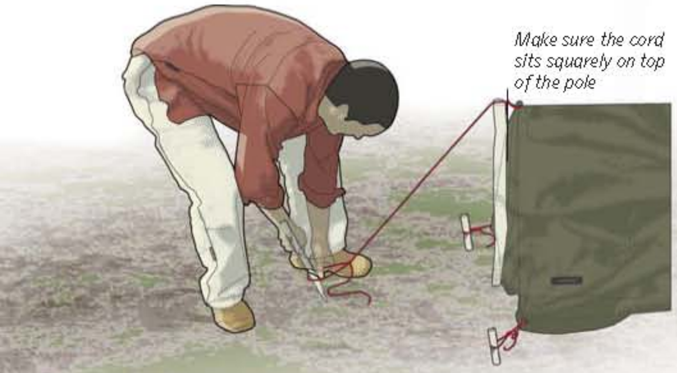

Make sure the cord sits squarely on top of the pole

5 When you want to get into the bivy pup tent, simply remove the opening pole and climb in feet first.

- Make sure you avoid the long cord as you wriggle in.

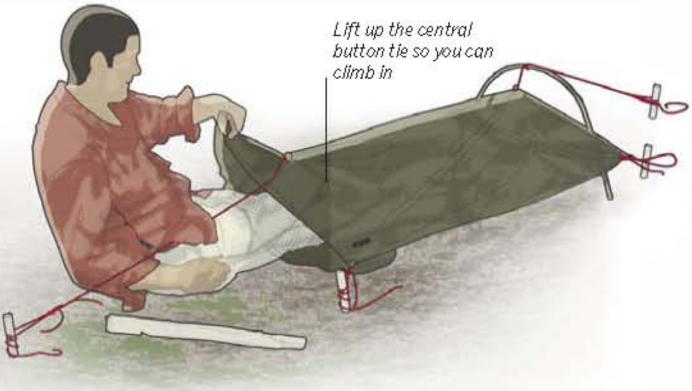

Lift up the central button tie so you can climb in

OTHER SHELTERS

There are other temporary shelters you can erect by adapting the techniques, tools, and materials outlined for both the use-anywhere shelter and the bivy pup tent. Which shelter you choose depends on whether you have a poncho, tarpaulin, or bothy bag.

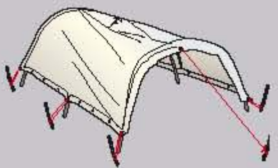

HOOPED SHELTER

If there are saplings or soft wood trees, you can form three hoops and peg out your poncho over them.

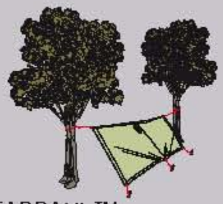

TARPAULIN

Attach one side of the tarp to a line between two trees and peg the other end down to the ground at an angle.

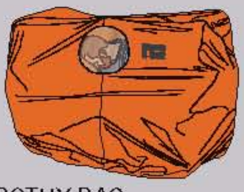

BOTHY BAG

A bothy bag, or storm shelter, is a large, light, nylon bag that protects one or two people from the elements.

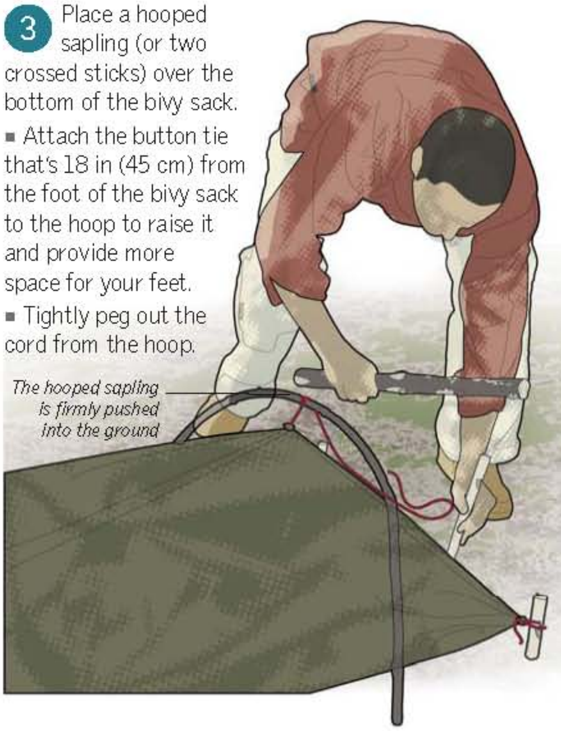

3 Place a hooped sapling (or two crossed sticks) over the bottom of the bivy sack.

- Attach the button tie that's 18 in (45 cm) from the foot of the bivy sack to the hoop to raise it and provide more space for your feet.
- Tightly peg out the cord from the hoop.

The hooped sapling is firmly pushed into the ground

6 After you have climbed in, replace the pole under the long cord.

- Use the hood of the bivy sack as a door and open and close as necessary.

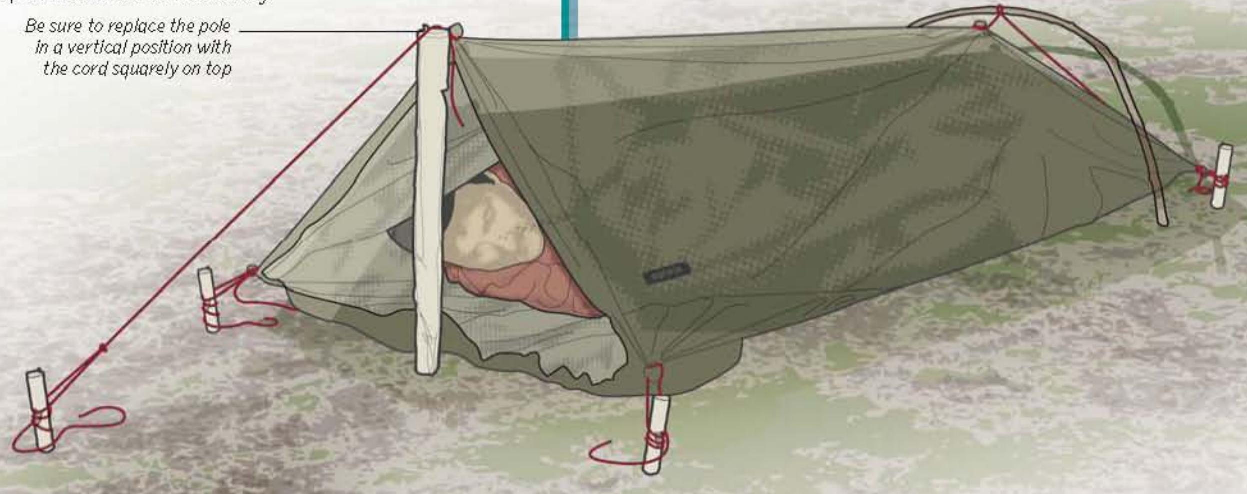

Be sure to replace the pole in a vertical position with the cord squarely on top

MAKING A BUTTON TIE

Button ties provide a simple, secure fastening for staking out ponchos and tarpaulins that have no loops or grommets to which you could attach a line. As you don't have to cut holes in the material, it remains waterproof and less prone to ripping.

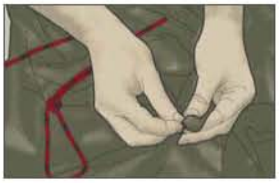

1 Encase a small, round, smooth stone (or similar) in material to form a "button". Prepare cord, with the open loop of a slip knot (see p. 142) at one end.

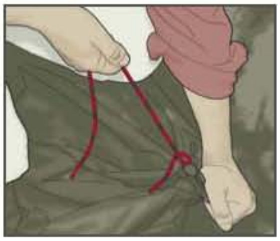

2 Place the open loop of the slip knot over the neck of the button and pull it tight.

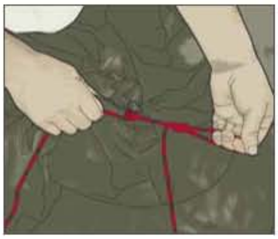

3 At the other end of the cord make a simple loop to place over your shelter pegs.

FOREST SHELTERS

IF YOU FIND YOURSELF in a survival situation in a temperate forest without a poncho, bivy sack, or tarp, you can make a shelter from natural materials. For example, you can construct a lean-to or an A-frame, or simply adapt a fallen tree. A forest lean-to is easier for a group to build because you can share the workload and don't have to use up too much precious energy.

TOOLS AND MATERIALS
- Poles and stakes
- Saw and knife
- Cordage
- Saplings
- Foliage for the roof covering

BUILDING A FOREST LEAN-TO

A lean-to has a sloping roof that leans against a horizontal ridgepole. It works best on a flat area between two trees or vertical supports secured firmly into the ground. If you're building a lean-to for a group, ask everyone to lie down next to each other and add an additional 6 in (15 cm) per person—this will establish how wide the shelter needs to be.

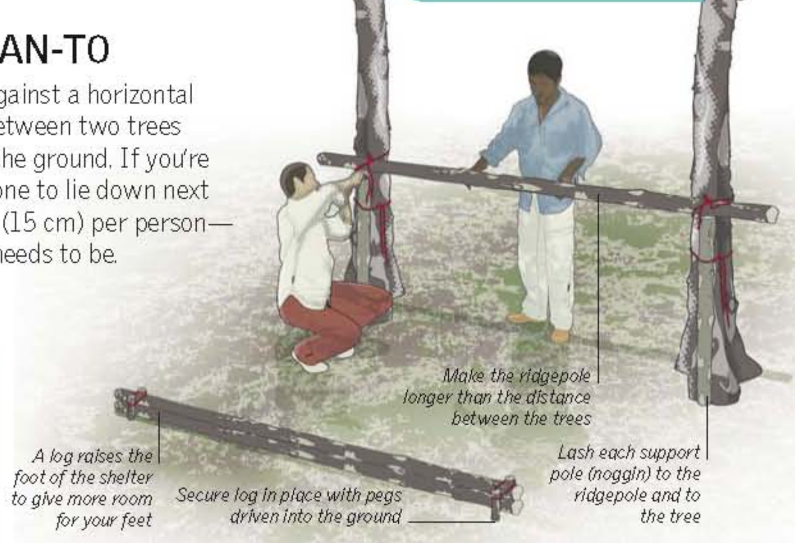

Make the ridgepole longer than the distance between the trees

A log raises the foot of the shelter to give more room for your feet

Secure log in place with pegs driven into the ground

Lash each support pole (noggin) to the ridgepole and to the tree

1 Place the ridgepole against the trees at the height you require and lash it with an arbor knot (see p. 142) to both trees.

- Place a support pole (noggin) under the ridgepole and lash to the trees.
- Firmly peg a log where you want the foot of the shelter to be.

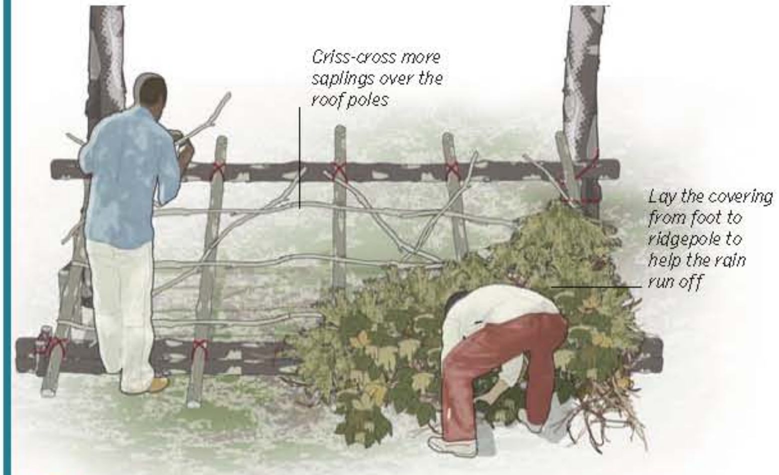

Criss-cross more saplings over the roof poles

Lay the covering from foot to ridgepole to help the rain run off

4 Apply the covering to the woven skeleton of the roof and walls. How much you apply depends on the weather you're experiencing.

- Lay large materials—for example, pine or fir boughs, large leaves, clumps of moss—as a base to stop smaller materials from falling through.

BUILDING IN A FOREST

Temperate forests, whether they consist of deciduous or coniferous trees or both, provide many opportunities for finding or building a shelter. Expend the least amount of energy for the most amount of protection—first see what nature can provide before building a shelter yourself.

POINTS TO REMEMBER

If you do need to make a shelter in a forest, here are some tips:

- Select a place close to water but away from water hazards, such as flooding, animals, and insects.
- Before you start, check for hazards such as deadfalls (sudden branch falls), rockfalls, and flash floods.
- Select a place that gives you maximum protection from the elements and with all the materials you need nearby.
- Collect everything you need before you start building and before it gets dark.
- Make sure your aids to location can be seen or activated quickly (signal fires, heliograph, flares, radio signals).
- Think safety: if possible, wear gloves to clear leaves and debris from the ground to protect you from spiders and snakes.
- Layer down your clothing as you work to prevent overheating.
- Make your shelter as waterproof as you can and ensure your bedding raises you at least 4 in (10 cm) off the ground.

TREE SHELTERS

The branches, roots, or trunk of a fallen tree can make a basic shelter. Make sure the tree is safe before you use it, as a further fall could injure you. Dead trees are probably dry, so don't light a fire under them. Check the base of the tree for snakes, spiders, and nests of insects.

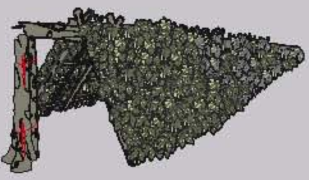

BROKEN BRANCHES

A branch that has snapped from a tree trunk but hasn't yet fallen makes a good ridgepole. Prop smaller branches and roofing material up against it.

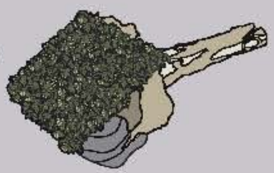

UPTURNED ROOTS

The base of an uprooted tree can give good protection from the elements. Make sure it's safe and won't flood if it rains, then add boughs and foliage.

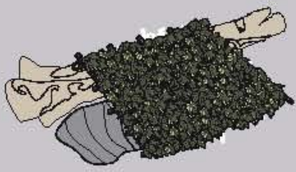

FALLEN TRUNK

Lay sticks or branches against a fallen trunk to form a pitched roof. Weave saplings across them and cover with bark, turf, moss, leaves, and mulch.

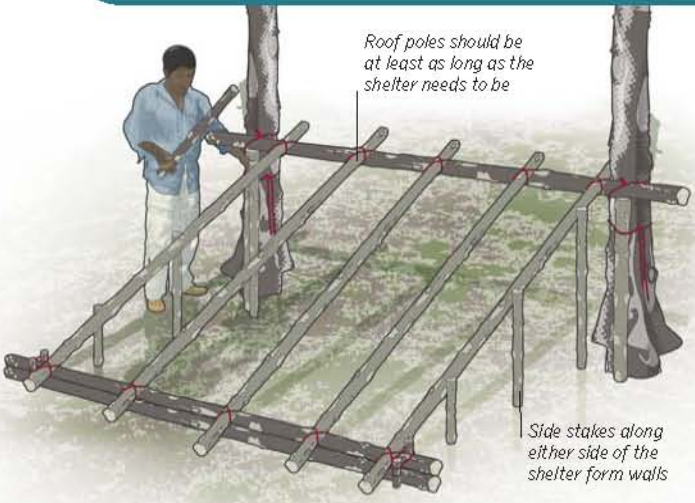

Roof poles should be at least as long as the shelter needs to be

Side stakes along either side of the shelter form walls

② Place five roof poles evenly along the ridgepole so that they slope back to the foot of the lean-to.

- Keep the outside poles between the trees to stop them moving outward.
- Lash each roof pole to the ridgepole.
- Hammer in side stakes alongside the outside poles.

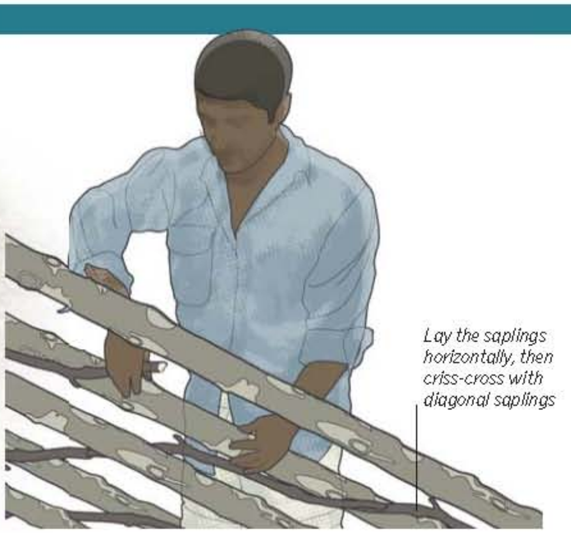

Lay the saplings horizontally, then criss-cross with diagonal saplings

③ Weave saplings in and out through the roof poles, working across and down from the ridgepole.

- Alternate the weave of each row of saplings, first horizontally and then diagonally.
- Weave smaller saplings through side stakes to form walls.

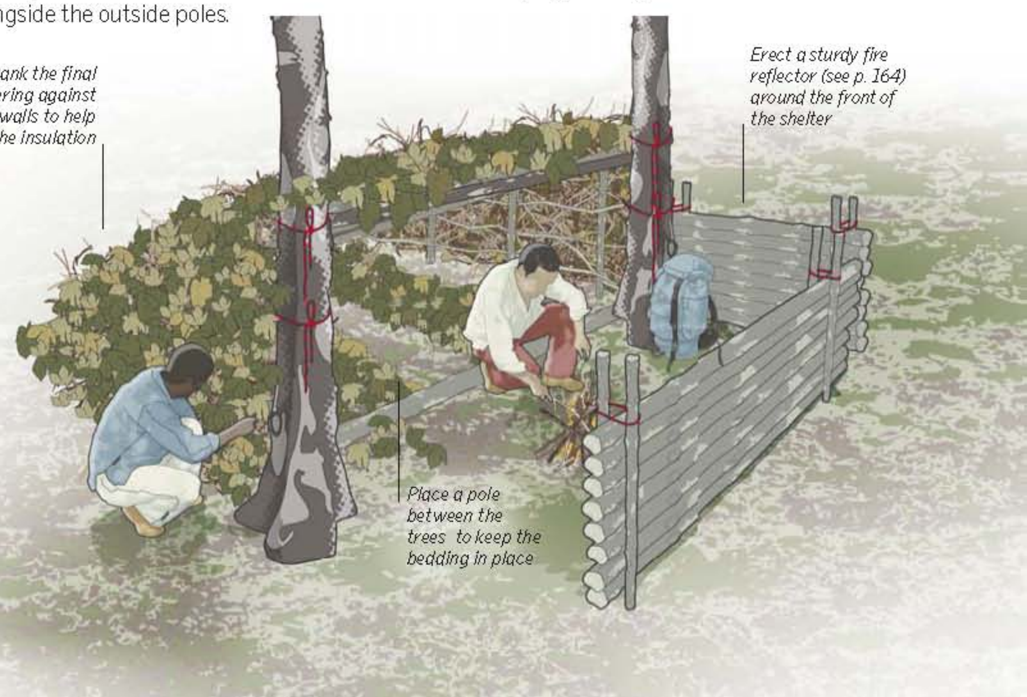

Bank the final covering against the walls to help the insulation

Erect a sturdy fire reflector (see p. 164) around the front of the shelter

Place a pole between the trees to keep the bedding in place

⑤ With everyone in the same lean-to shelter you need to light and tend only one fire.

- Layer the inside floor with dry material to form bedding.
- Establish a watch system during the night to ensure that the fire is maintained and controlled.

FOREST A-FRAME

A-frames can take several hours to build, but if you're staying in one place for a few days it may be worth expending the energy. One-person A-frames are relatively simple to construct and are easy to keep warm. They can be adapted into multi-person shelters to accommodate groups of people.

TOOLS AND MATERIALS
- Poles of various lengths
- Saplings
- Knife
- Saw (or ax)
- Cordage
- Covering for walls—such as pine boughs, leaves, and moss

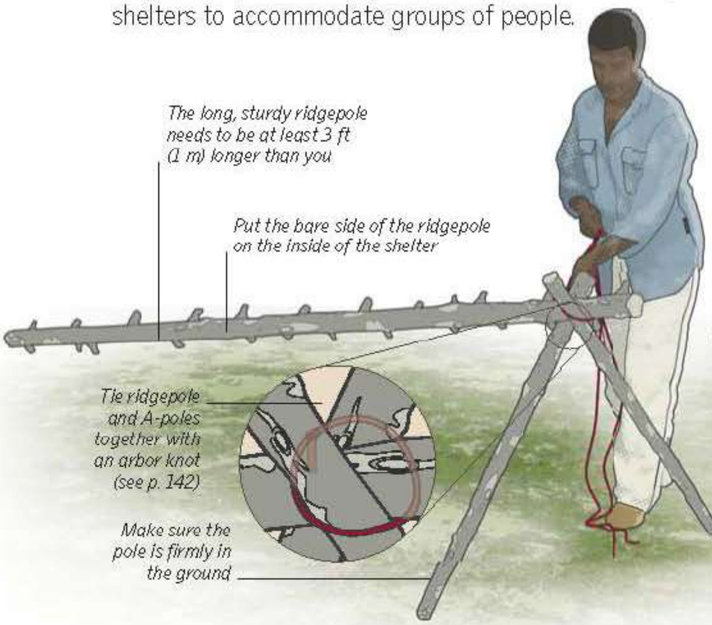
The long, sturdy ridgepole needs to be at least 3 ft (1 m) longer than you
Put the bare side of the ridgepole on the inside of the shelter
Tie ridgepole and A-poles together with an arbor knot (see p. 142)
Make sure the pole is firmly in the ground

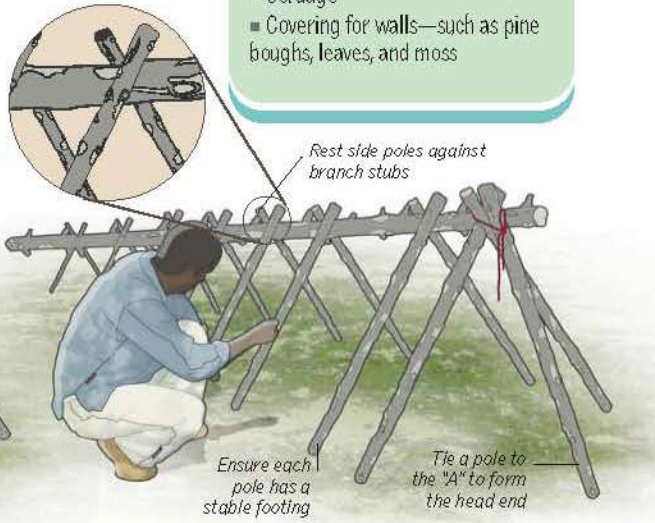
Rest side poles against branch stubs
Ensure each pole has a stable footing
Tie a pole to the "A" to form the head end

1 Smooth one side of the ridgepole with your knife, leaving one side bare and one with branch stubs.

- Hammer two poles into the ground to form an "A".
- Rest one end of the ridgepole on top of the "A" and the other end on the ground.

2 Lay a series of poles against the branch stubs on the ridgepole. These will be of decreasing height as you work your way from front (the "A") to back (the ground).

- At each junction, rest the side poles against the stubs or tie them to the ridgepole with an arbor knot (see p. 142).

FIRE REFLECTOR

A fire radiates heat in all directions—up, down, and 360 degrees around. A fire reflector makes a fire more efficient as it directs heat into your shelter.

Construct your reflector so that you can build a fire about 3 ft (1 m) from the entrance to your shelter. You can sit in the entrance and tend the fire at arms' length—any closer and you would be too hot and the shelter might catch fire; any further away and you lose too much heat. An L-shaped end to the reflector will retain more heat.

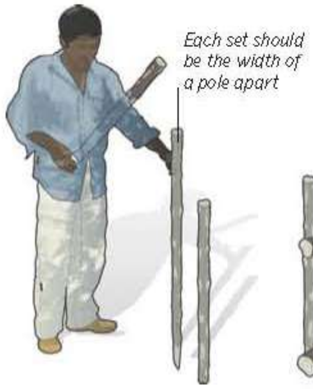
Each set should be the width of a pole apart

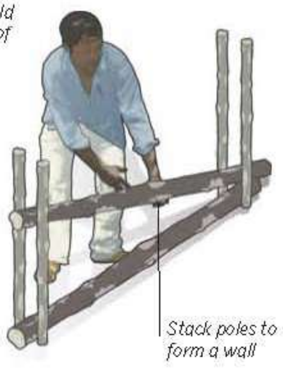
Stack poles to form a wall

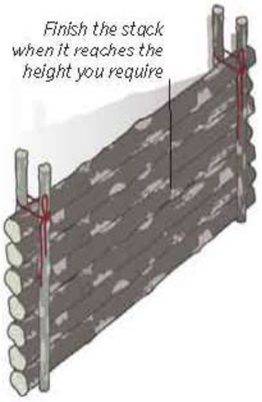
Finish the stack when it reaches the height you require

1 Check the air flow will be across the proposed reflector (to disperse smoke).

- Hammer two poles into the ground. Put a second set in at poles' length apart.

2 If the fire is going to be close to the reflector, use green wood to prevent the heat setting it on fire.

- Put poles between the sets to form a wall.

3 Tie the top of the upright poles together.

- Make sure the wall is as long as the shelter's entrance to retain heat and keep out wind and rain.

ALTERNATIVE SHELTERS

Two other options are a front-opening A-frame and a one-person lean-to. For the former, secure a ridgepole or line to a tree at the height you want the entrance to be, then adapt the forest A-frame (see opposite). For the latter, tie the ridgepole to two forked sticks hammered into the ground and apply a cover similar to the forest lean-to (see pp. 162–63).

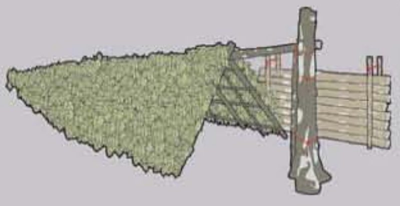

FRONT OPENING A-FRAME

Wedge a noggin under the ridgepole and lash it to the tree with an arbor knot to take some of the weight. Build a fire reflector to the side with the prevailing wind.

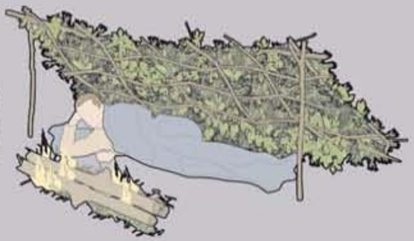

ONE-PERSON LEAN-TO

A small one-person lean-to, together with a long log fire built in front, provides you with a warm, dry, and comfortable shelter for an overnight stay.

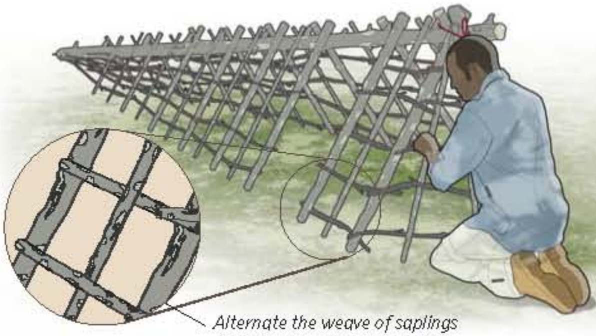

Alternate the weave of saplings between one row and the next

3 Weave four or five rows of saplings horizontally through the side poles to make the walls into a strong framework for the covering.

- Leave an entrance in the side near the front. Make sure it's large enough for you to sit inside comfortably.

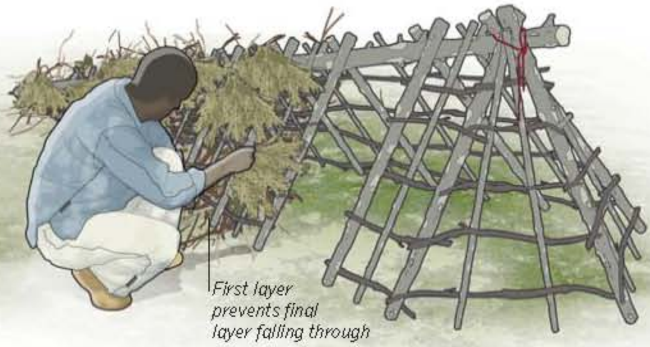

First layer prevents final layer falling through

4 Cover the framework with a thick layer of natural material, such as pine boughs, small branches, and sturdy twigs. Weave the materials into a "thatch".

- On each side, work your way gradually from the back of the shelter to the front, and from the bottom to the top.

NEED TO KNOW

Build your shelter so that the wind blows across the front of it at a slight angle, rather than into it. This maintains its effectiveness and prevents smoke entering from the fire. If the shelter backs on to the wind, there's a risk the wind will blow around it and curl into the entrance.

5 Cover the first layer with leaves, moss, and mulch, starting from the ground and working up to the ridgepole.

- Build a fire reflector (see opposite) in front of the entrance to help direct heat towards you.
- Stow your equipment and pack in the front of the shelter.

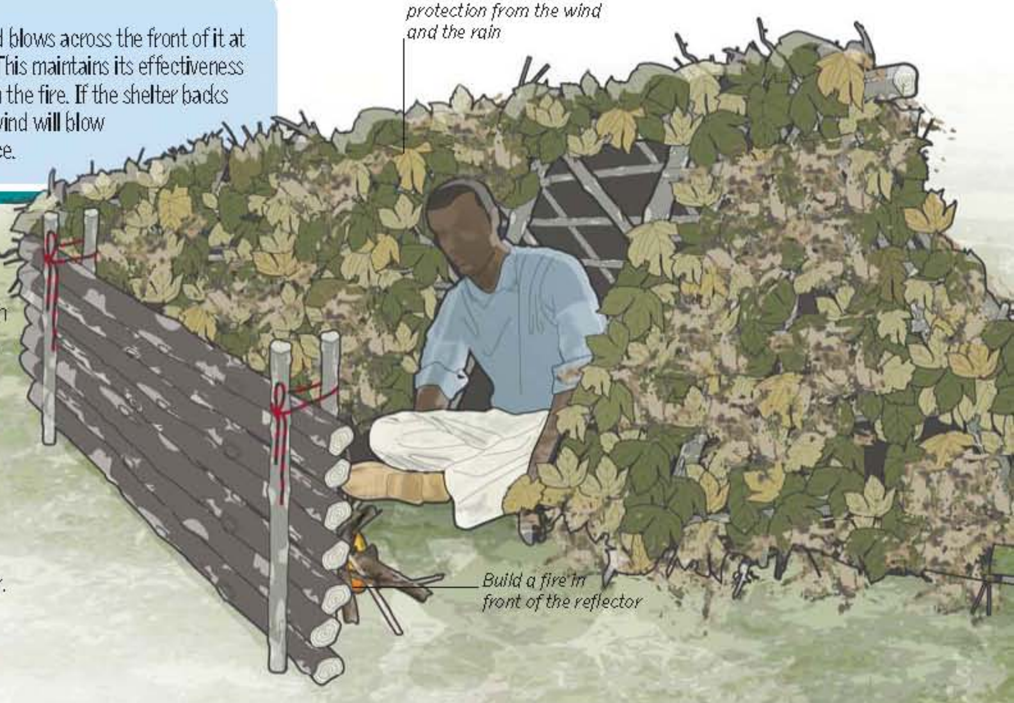

The covered walls provide protection from the wind and the rain

Build a fire in front of the reflector

TROPICAL SHELTERS

TROPICAL RAIN FORESTS CAN BE the easiest environments for building shelters because of the wealth of available materials. Once you've found a suitable location just decide what type of shelter you need. The most important thing is to raise the shelter off the ground. Make it as comfortable as you can, because quality sleep is important.

TOOLS AND MATERIALS

- Poncho fitted with press-studs
- Cordage
- Long, heavy poles or wide bamboo stalks
- Spacer bars
- Knife, saw, machete, or parang
- Needle
- Strong, waxed, cotton thread

MODIFIED PONCHO BED

Most ponchos have press-studs to allow you to clip up the edges or to clip two ponchos together. They also have grommet holes that you can use as lashing points. Here, a poncho is modified into a tropical bed.

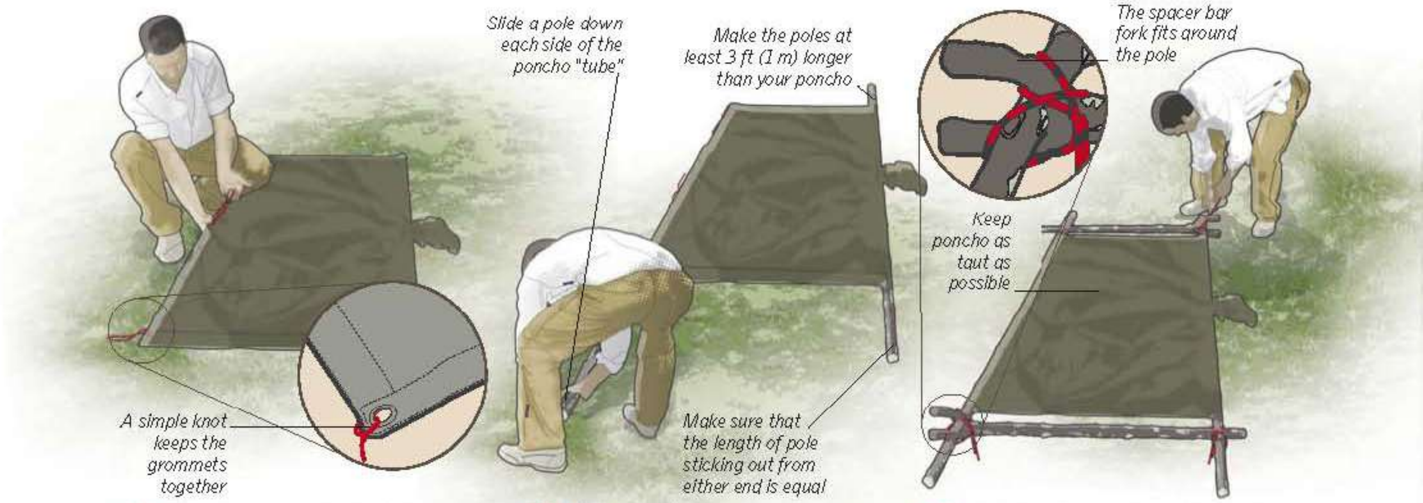
A simple knot keeps the grommets together
Slide a pole down each side of the poncho "tube"
Make the poles at least 3 ft (1 m) longer than your poncho
Make sure that the length of pole sticking out from either end is equal
The spacer bar fork fits around the pole
Keep poncho as taut as possible

1 Clip the press-studs of the two sides of your poncho together to form a kind of "tube"."

- Tie the grommets on the clipped side with pieces of cordage so that the press-studs don't come undone.

2 Cut two strong support poles that can take your weight.

- Insert the two support poles inside the "tube" of your poncho, one down each side.

3 Cut two "spacer bars" that will fit perpendicularly across the support poles or, if you can find them, cut poles with a natural fork in them.

- Tie the spacer bars into position using an arbor knot (see p. 142).

HAMMOCKS

Many types of hammock are designed for the kind of hot weather you get in the tropics, and are often fitted with mosquito nets. In fact, hammocks make the best overnight shelter, as long as you can find two trees the right distance apart. The Hennessy hammock is a good choice because it packs down small.

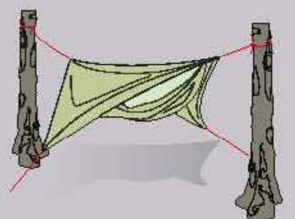

HENNESSY HAMMOCK
This ultralight, quick-drying hammock is both easy to set up and comfortable to sleep in (see p. 54). It's equipped with a self-closing door.

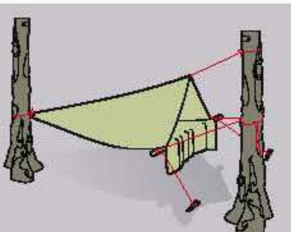

PARACHUTE SHELTER
Fold the panels of a parachute into a triangle. Lash a spacer bar to the opening and peg it out. Tie off the apex of the triangle further up the tree.

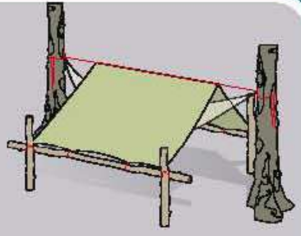

HAMMOCK AND TARP
Tie a ridgeline above the hammock. Drape the tarp over the line and anchor it on either side to a horizontal pole tied to two uprights.

ALTERNATIVE METHODS

You can easily modify other forms of sheeting, such as a tarpaulin, shelter sheet, and groundsheet. You will need a decent needle and plenty of strong, waxed, cotton thread (see Survival kits, pp. 60–61). There are two methods, which you can do either on the trail or as part of your prior preparation. Adding spacer bars keeps the poles apart and the sleeping platform tight. Both of these methods produce an excellent emergency stretcher.

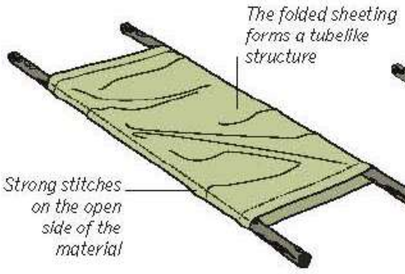
The folded sheeting forms a tubelike structure
Strong stitches on the open side of the material

METHOD ONE

Lay the material on the ground and fold it once to create a wide "tube." Stitch the length of the open sides and then insert a pole along each side of the tube.

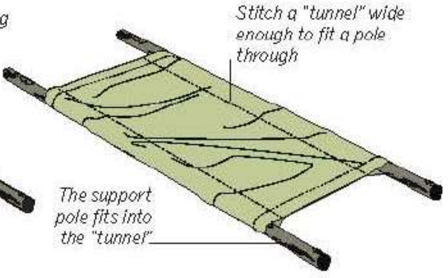
Stitch a "tunnel" wide enough to fit a pole through
The support pole fits into the "tunnel"

METHOD TWO

Fold the sheet and stitch a "tunnel" along both sides just wide enough to take a pole that will support your weight. Push the support poles through the tunnels.

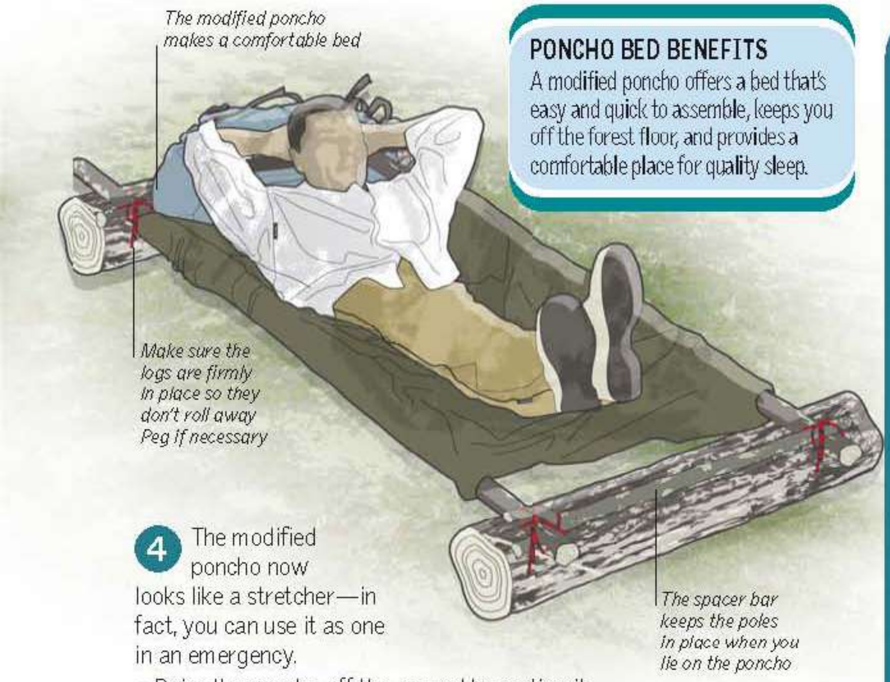
The modified poncho makes a comfortable bed
Make sure the logs are firmly in place so they don't roll away Peg if necessary
The spacer bar keeps the poles in place when you lie on the poncho

PONCHO BED BENEFITS

A modified poncho offers a bed that's easy and quick to assemble, keeps you off the forest floor, and provides a comfortable place for quality sleep.

4 The modified poncho now looks like a stretcher—in fact, you can use it as one in an emergency.

- Raise the poncho off the ground by resting it either on thick logs or sturdy poles.

ALTERNATIVE DESIGN

If you have a second poncho, shelter sheet, or tarp you can create a shelter to keep the rain off your modified poncho (see also Corded A-frame, p. 159).

- Tie a ridgeline between two trees and hang the poncho, sheet, or tarp over the top of it.
- If you're using a second poncho, tie its hood to another ridgeline.
- Peg out the four corners of the poncho, sheet, or tarpaulin.

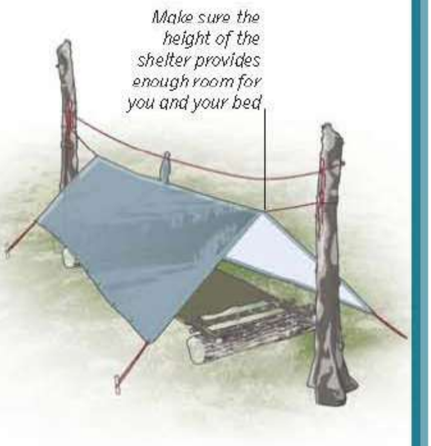
Make sure the height of the shelter provides enough room for you and your bed

BUILDING IN THE JUNGLE

Shelters in the rain forest need to be quick to erect, just big enough for your needs, and safe from animals.

CHECKLIST

If you're thinking about building a shelter in a tropical forest, the following tips may help you:

- Check for hazards such as deadfalls (sudden branch falls)—a major cause of injuries during military jungle training.
- You'll need a sharp cutting tool—ideally a parang or machete, but a decent bushcraft knife or small ax will suffice.
- Clear the ground around your shelter to deter animals. Use a makeshift brush, never your bare hands, to avoid being bitten by snakes or spiders.
- Build your shelter far enough off the ground to avoid being bothered by insects, snakes, or any other animals—particularly those that move around at night.
- Start building your shelter well before darkness sets in, which, in tropical zones, usually happens very quickly. Using a large knife in reduced light or by the light of a flashlight can be very dangerous.
- Work at a rate that you can manage. Humid conditions can very quickly lead to dehydration and heat-related injuries such as heat stress and heat stroke. Your body tries to keep itself cool by sweating, so don't work too fast, drink water frequently, and take regular breaks.
- Make your shelter secure—you don't want to make repairs at night.
- Inside your shelter use a full mosquito net and a mosquito head-net.
- A fire will deter insects and animals.

JUNGLE A-FRAME

An A-frame is relatively easy to make. If you have a poncho, tarpaulin, shelter sheet, groundsheet, or another type of sheet, you'll need to modify it first to make it into a bed (see pp. 166–67). If you do not have a poncho, lay branches across the poles to form a sleeping platform.

1 Cut seven long poles that will take your weight. Tie two poles together with an arbor knot (see p. 142) to form an "A" and tie the joint to a tree or branch of a tree (as shown).

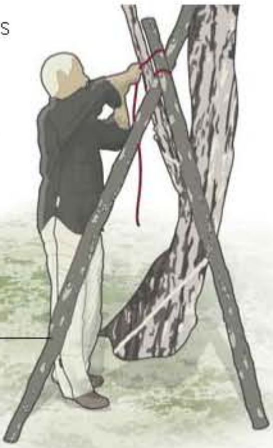

The angle formed by the "A" determines how far down the legs the platform will sit

2 Tie two more poles together to form an "A" for the other end.

- Position these in line with the first "A".
- The distance between the two "A"s needs to be at least 2 ft (60 cm) longer than your height.

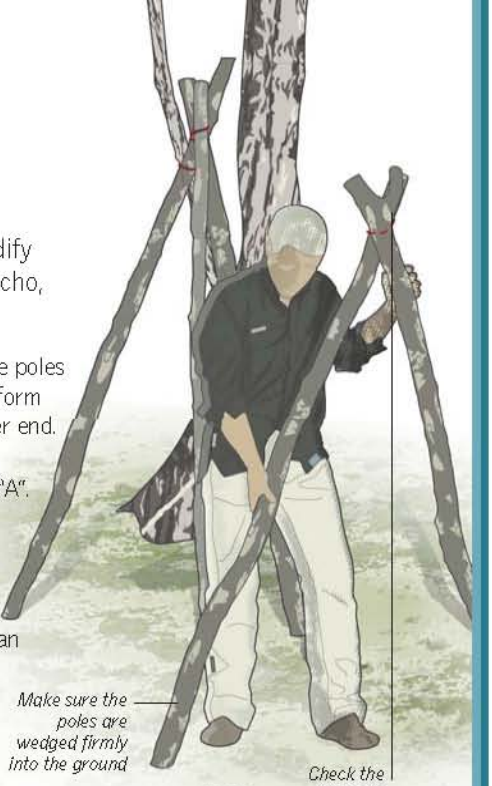

Make sure the poles are wedged firmly into the ground

Check the junctions are securely tied

3 Put a ridgepole on top of the two "A"s.

- Tie this to both junctions for extra stability.

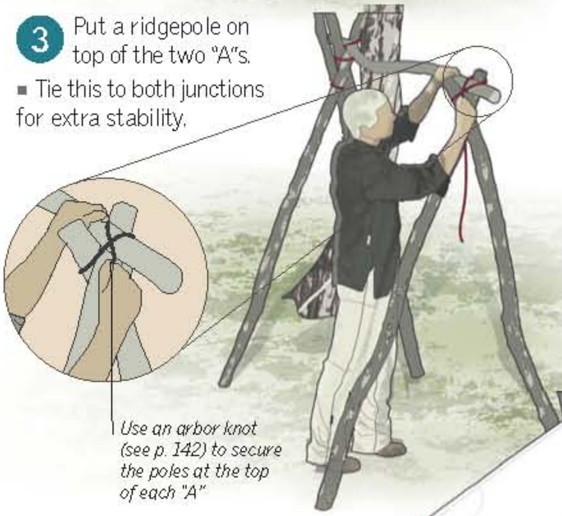

Use an arbor knot (see p. 142) to secure the poles at the top of each "A"

4 Position both poles of your poncho bed (see p. 166) on the outside of the A-frame.

- Move the poles down the outside of the frame legs until the sheet is taut.

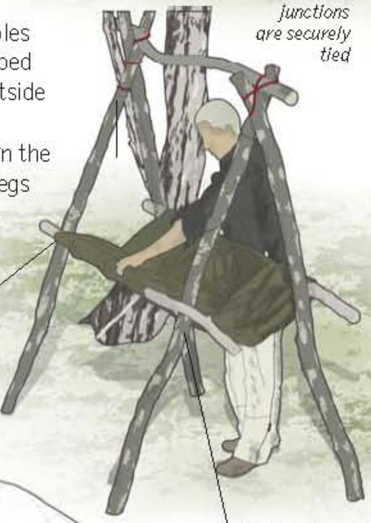

The poles of the modified poncho settle into position

Tie the poles of the modified poncho to the legs of the A-frame for extra stability

5 Place a tarpaulin or shelter sheet over the ridgepole.

- Securely tie a guyline to each corner with a Siberian hitch (see p. 144).
- Run each guyline to a peg or tree and secure it with a taut line hitch (see p. 145).

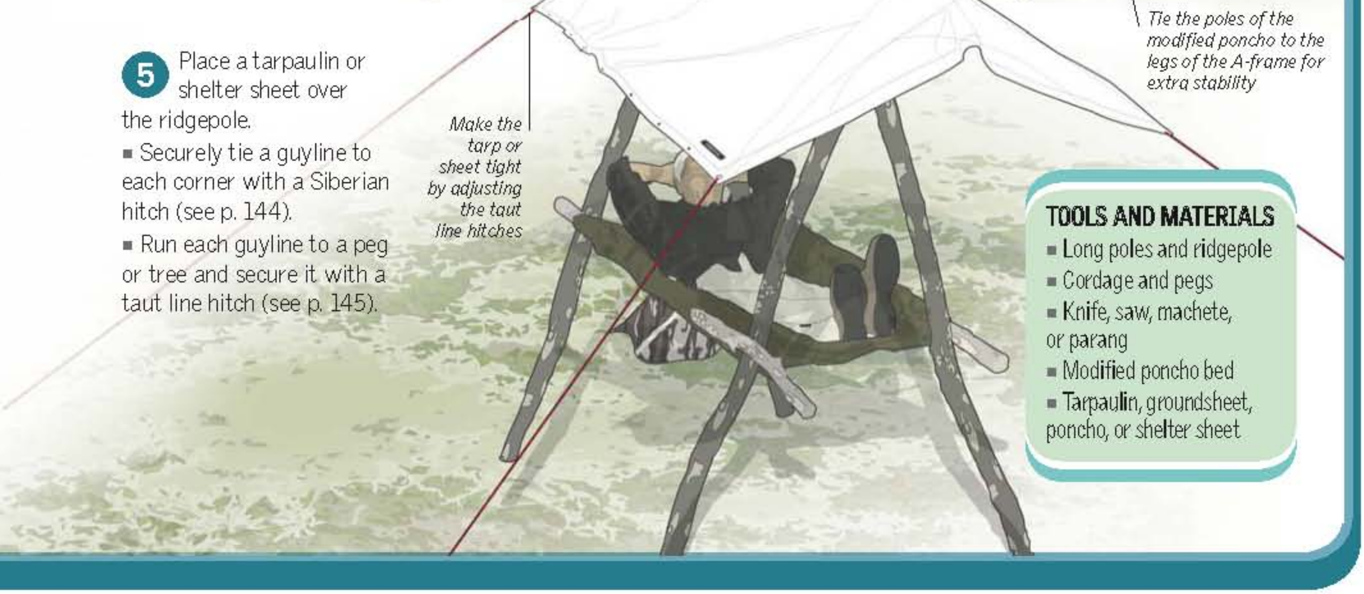

Make the tarp or sheet tight by adjusting the taut line hitches

TOOLS AND MATERIALS

- Long poles and ridgepole
- Cordage and pegs
- Knife, saw, machete, or parang
- Modified poncho bed
- Tarpaulin, groundsheet, poncho, or shelter sheet

BAMBOO LEAN-TO

If you can find enough bamboo you can make a type of lean-to shelter. Find a spot close to the bamboo supply to reduce the amount you have to carry. If you need a raised sleeping platform, adapt the method used for the jungle hut (see p. 174).

TOOLS AND MATERIALS

- Thick wooden poles and bamboo poles
- Knife, saw, machete, or parang
- Cordage and pegs
- Poncho and tarpaulin (if you have them)

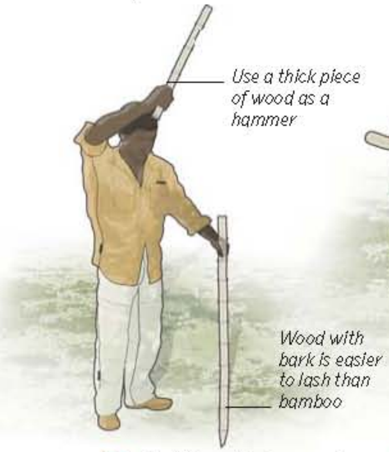
Use a thick piece of wood as a hammer
Wood with bark is easier to lash than bamboo

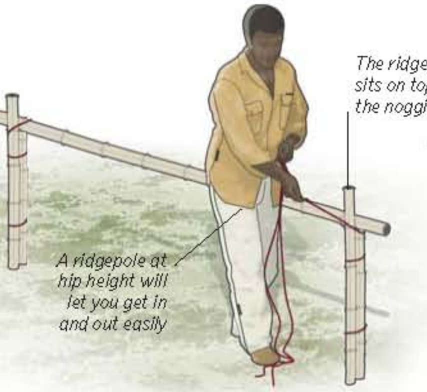
The ridgepole sits on top of the noggin
A ridgepole at hip height will let you get in and out easily

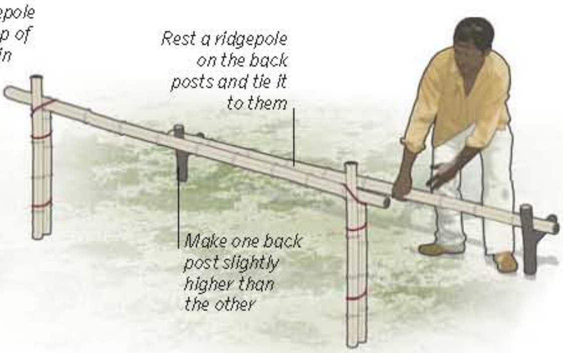
Rest a ridgepole on the back posts and tie it to them
Make one back post slightly higher than the other

1 Cut four thick, wooden support posts.

- Hammer two posts into the ground, wide enough to let you lie lengthways between them.

2 Tie a noggin to each post for support. Make sure they are level and at the height you need for an entrance.

- Place a wooden ridgepole on top of the noggins and tie it to the posts.

3 Make two short, forked back posts to give the shelter a sloped roof.

- Align them with the front posts and far enough back so the roof is wide enough to shelter you.
- Hammer them into the ground and tie a second ridgepole to them.

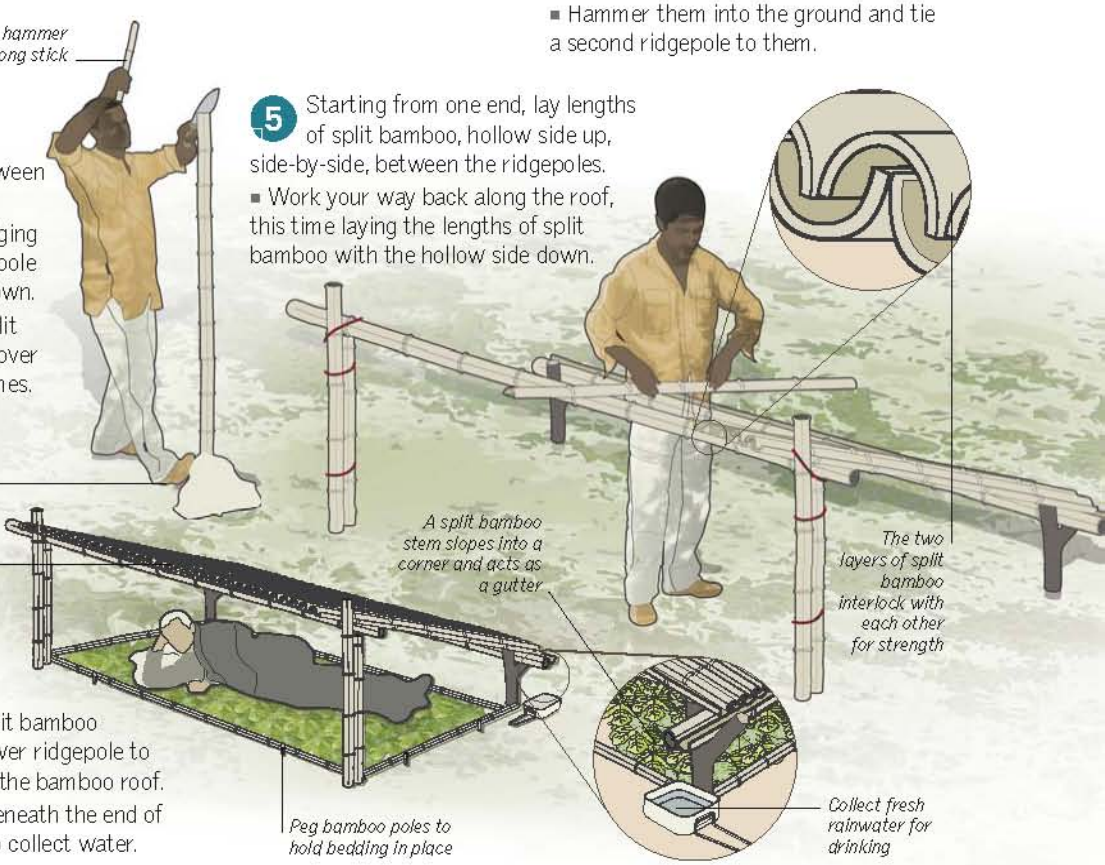
Wedge in knife then hammer it down with a strong stick
Steady the bamboo stem against a rock with your foot
The two layers of split bamboo interlock with each other for strength
A split bamboo stem slopes into a corner and acts as a gutter
The split bamboo layers of the roof direct water into the gutter
Peg bamboo poles to hold bedding in place
Collect fresh rainwater for drinking

4 Cut lengths of bamboo long enough to reach between the two ridgepoles.

- Split these by wedging your knife into each pole and hammering it down.
- Prepare enough split bamboo lengths to cover the roof area two times.

5 Starting from one end, lay lengths of split bamboo, hollow side up, side-by-side, between the ridgepoles.

- Work your way back along the roof, this time laying the lengths of split bamboo with the hollow side down.

6 Lash a long split bamboo stem to the lower ridgepole to form a gutter under the bamboo roof.

- Put a pan or pot beneath the end of this sloping gutter to collect water.

JUNGLE HUT

With a little extra effort you can make a more permanent and substantial shelter than a jungle A-frame. The raised apex roof of a jungle hut allows you to sit on the platform in relative comfort. This method uses a framework of four posts, but using two trees instead of two of the poles would make it more rigid.

TOOLS AND MATERIALS

- Strong posts and poles
- Branches for platform
- Knife, saw, machete, or parang
- Digging tool
- Broad leaves for thatching the roof

1 Dig four strong posts into the ground to form a rectangle at least 1 ft (30 cm) longer and wider than yourself.

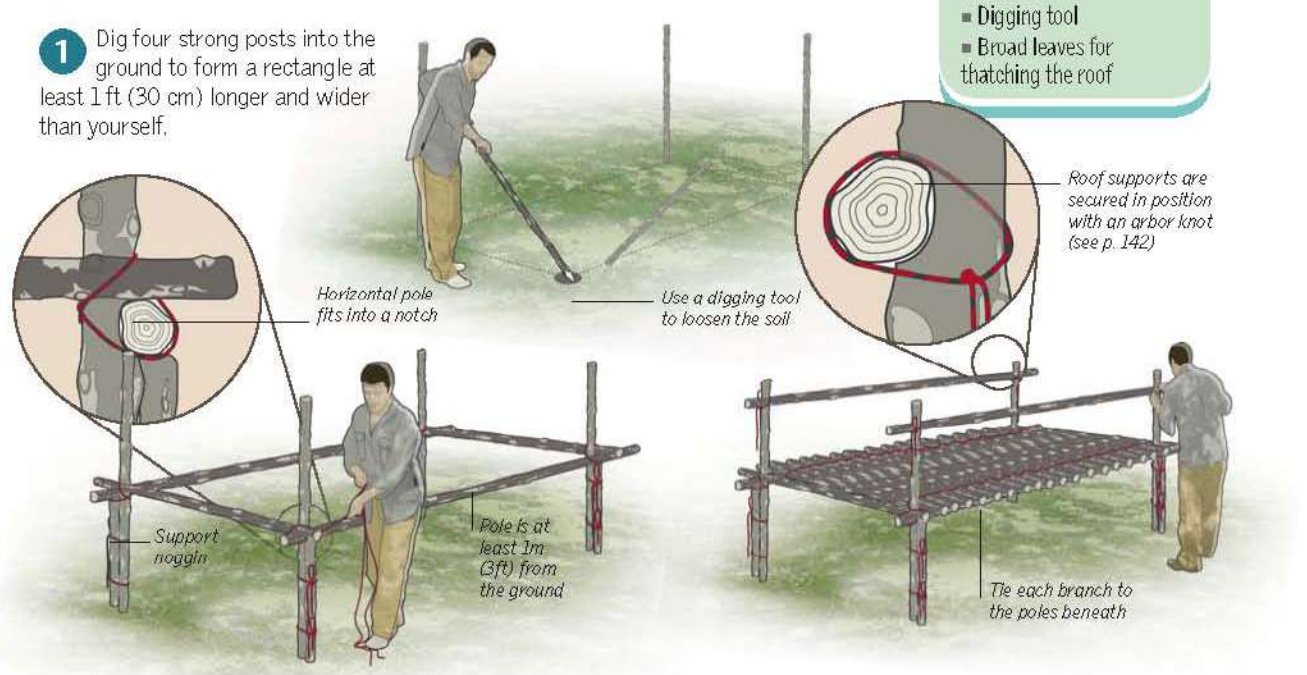

Horizontal pole fits into a notch

Use a digging tool to loosen the soil

Roof supports are secured in position with an arbor knot (see p. 142)

Support noggin

Pole is at least 1m (3ft) from the ground

Tie each branch to the poles beneath

2 With your knife or saw, cut a notch into each post deep enough to fit the curve of the horizontal pole.

- Wedge the four horizontal poles into their notches and secure the junction using an arbor knot (see p. 142).
- Lash a noggin to each post to support the poles.

3 Lay a row of stripped branches side-by-side across the horizontal poles to form a raised floor.

- To make the roof supports, cut a notch on the inside of each post about 3 ft (1 m) higher than the platform.
- Tie a pole into the notches on the posts.

THATCHING LEAVES

By taking advantage of the shapes of some tropical leaves, you can make a substantial, long-lasting roof or walls. The bigger and broader leaves involve less work. If you use rattan palm leaves you can create an alternating interweave that gives a tighter thatch than the overlapping method described above.

RATTAN LEAVES

The rattan palm has leaves that are composed of rows of smaller leaflets. You can split the leaves in two and hang the halves in layers on the framework of your roof. Alternatively, you can use them whole, by folding the leaflets from one side and individually weaving them with leaflets from the other side.

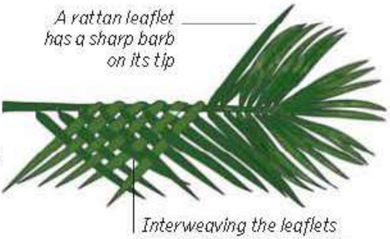

A rattan leaflet has a sharp barb on its tip

Interweaving the leaflets from two sides of the same leaf creates a tight mesh

LONG, BROAD LEAVES

Many tropical plants have long, broad leaves that are good for thatching. These include some banyan figs, rubber plants, and types of banana. Arrange these leaves side-by-side on the poles of your roof frame, tuck them over the pole, and sew them into position with strips of vine. The tips of the leaves in one row point down over the leaves of the row underneath.

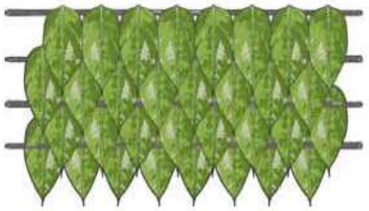

ALTERNATIVE SHELTERS

Wickiups are found in various forms around the world and are adapted as local needs and materials dictate. For example, Native Americans made them as temporary shelters in the plains, where they were also known as teepees. The pygmies of the rain forest used saplings to form a hemisphere that, once covered with natural material, created a warm, dry shelter.

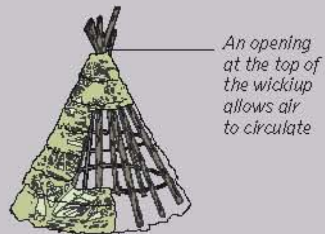
An opening at the top of the wickiup allows air to circulate

WICKIUP

The wickiup is a cluster of straight poles that are lashed together at the top, with an interwoven framework that's covered with animal hides or grass.

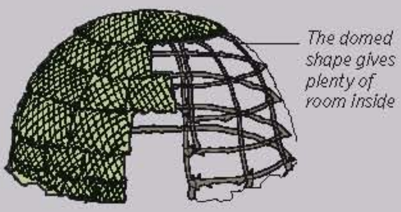
The domed shape gives plenty of room inside

PYGMY HUT

The domed pygmy hut is made from a circle of bent saplings or limber poles secured firmly in the ground, then lashed and thatched with natural materials.

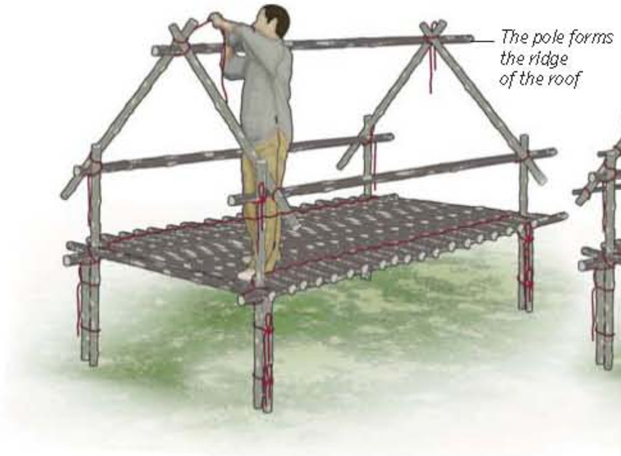
The pole forms the ridge of the roof

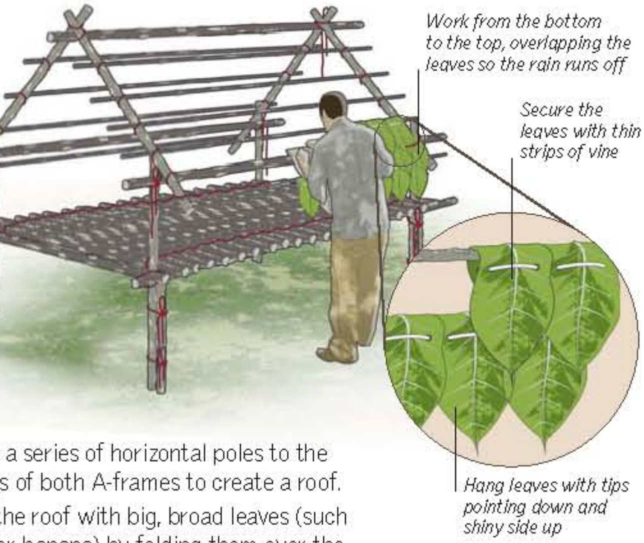
Work from the bottom to the top, overlapping the leaves so the rain runs off
Secure the leaves with thin strips of vine
Hang leaves with tips pointing down and shiny side up

4 Tie poles together to form two A-frames. Tie these on top of the roof support poles at each end of the framework.

- Place a ridgepole across the apex of the two A-frames and tie it in place.

5 Tie a series of horizontal poles to the legs of both A-frames to create a roof.

- Cover the roof with big, broad leaves (such as palm or banana) by folding them over the horizontal roof poles.

THREE-LOBED LEAVES

Leaves with three lobes, such as the familiar fig leaf, can simply be hooked over the poles of your roof with the stem pointing upward. On each row of leaves you need to hang the left and right lobes either behind the pole or in front of it in an alternating pattern. The middle lobe then hangs down over two leaves on the row underneath.

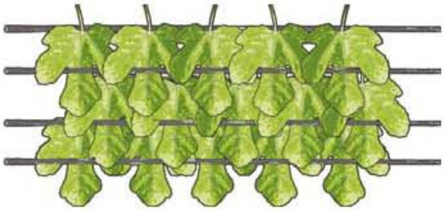

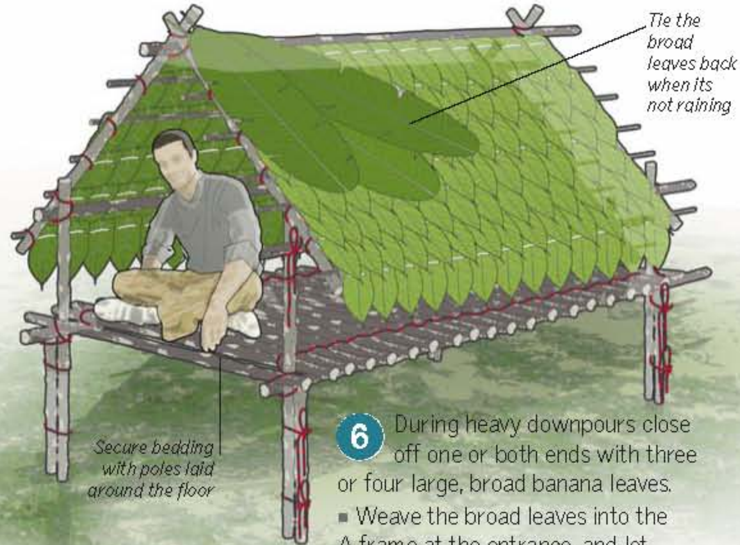
Tie the broad leaves back when its not raining
Secure bedding with poles laid around the floor

6 During heavy downpours close off one or both ends with three or four large, broad banana leaves.

- Weave the broad leaves into the A-frame at the entrance, and let them hang down.

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL— IN THE JUNGLE

USEFUL EQUIPMENT

- Machete/parang
- Tarpaulin/shelter sheet
- Hammock and mosquito net
- Insect repellant
- Signal flares and whistle
- Flashlight and batteries
- First-aid kit
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

FRENCHMEN GUILHEM NAYRAL AND LOÏC PILLOIS lost their way on a 78 mile (125 km) trek through virgin rain forest in French Guiana. They survived for 51 days in the jungle by drinking rainwater and eating palm seeds, snakes, and insects—although Nayral narrowly escaped death after eating a poisonous spider.

The pair set off from the Grand Kanori rapids on the Approuague River on Wednesday, February 14, 2007, bound for Saül, a former mining village at the center of the country. They planned to spend 12 days hiking, and carried sufficient food supplies, a compass and map, machete, a tarpaulin, and two hammocks. They soon found that the going was tough—some days it would take several hours of hacking through vines to hike just one mile. On the morning of February 26, when Nayral and Pillois had been due to reach Saül, the pair found themselves far from civilization and in unexpected terrain.

Knowing that a search operation would be launched, they set up camp and waited to be rescued. They built a shelter and divided the tasks—Nayral in charge of food and Pillois the fire, which was kept alight constantly to attract rescuers. They had abundant water from rainfall, but were reduced to scavenging for palm seeds, beetles, snakes, frogs, and spiders.

"THE PAIR FOUND THEMSELVES FAR FROM CIVILIZATION AND IN UNEXPECTED TERRAIN"

Occasionally, helicopters passed overhead, but failed to spot them through the dense canopy of foliage. After waiting for three weeks, the pair decided to abandon camp and trek west toward Saül. After another week of trekking for just three hours a day, unknowingly just 3 miles (5 km) from Saül, Nayral was incapacitated after eating a still-poisonous half-cooked spider. With no choice but to leave his companion behind, Pillois pushed on and reached Saül, returning with a helicopter to rescue Nayral—dehydrated, with intestinal poisoning, and plagued by parasites—on April 5, 51 days after beginning his trek.

WHAT TO DO

ARE YOU IN DANGER?

NO YES

If you are in a group, try to help any others who are in danger

Get yourself out of it:
Sun/Heat/Humidity—Slow down to the pace of the jungle. Find or improvise immediate shelter
Animals—Only six percent of snakes are poisonous but everything in the jungle will try to protect itself
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION
See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO YES

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO YES

You are faced with surviving for an indefinite period—until you are located or you find help

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO YES

If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers a better chance of survival, rescue, or both

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

YOU SHOULD STAY **

DO

- Make an informed decision on the best location to move to
- Use line of sight to navigate on your bearing, as visibility may be less than 33 ft (10 m)
- Improvise shelter when not moving, and sleep off the ground, clear of the damp floor and animals
- Be on constant lookout for dry tinder and fuel
- Follow water courses downstream. Transport in the jungle relies on rivers, so settlements are most likely to be found alongside rivers
- Step onto logs so that you can see what is on the other side, rather than stepping straight over onto an unseen snake

DON'T

- Use your hands to clear undergrowth—a machete or walking staff is better suited
- Drink untreated water—boil or treat all of your water before drinking
- Leave it too late in the day to stop and make camp—three hours before sunset is recommended
- Keep too quiet. Make a noise as you progress to warn animals for your approach

DON'T

- Let the oppressive nature of the jungle overwhelm you. Slow down to its pace, don't fight against it
- Let your firewood get damp—store dry tinder and split or quarter wood to get to the dry inner core
- Eat what you cannot identify as edible—this could result in you becoming so ill that you cannot function

DO

- Select a shelter site where you can sleep off the ground and where your location aids will be most effective
- Use a mosquito net if you have one; if not, put damp foliage on your fire to repel insects, or cover exposed skin with mud
- Deploy all your aids to location and prepare them for immediate use. Be constantly alert for signs of rescue
- Keep your flesh covered despite the heat—the high humidity encourages infections. Also wash at every opportunity
- Keep a fire going—it aids location and wards off insects

---

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.

** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

DESERT SHELTERS

IF YOU'RE PLANNING A TRIP into the desert, take something you can use to improvise an immediate shelter from the sun—for example, a shelter sheet, tarpaulin, poncho, space blanket, or even a couple of pieces of parachute material that will pack down into almost nothing. You can either build a shelter called a "scrape" in a natural hollow or erect a quick shelter above the ground using your poncho.

TOOLS AND MATERIALS

- Cordage and pegs
- Digging tool
- Knife
- Poncho, space blanket, tarp, or parachute material
- Rocks or ballast material
- Strong posts

BUILDING IN THE DESERT

Erecting a natural shelter in the desert is a challenge because of the heat and potential lack of materials, so try to find a site that's shaded by a tree or shrub.

POINTS TO REMEMBER

The following tips will help you deal with the extremes of temperature:

- Never try to build your shelter during the hottest part of the day.
- Ration your sweat and not your water. Avoid exerting yourself. If you start to sweat then stop what you're doing and take a break in the shade (30 minutes under an unfolded map or a space blanket and have a drink). If in a group, share the work.
- Don't leave it until the last minute to find a suitable location for your shelter. Make the decision early and plan accordingly.
- Never build a shelter in a low-lying area, dry river bed, or wadi as they are potentially at risk of flash floods.
- Avoid the top of large, isolated hills or mountains because of the danger from lightning and extreme winds.
- Try to set up your shelter on a small rise, where the temperature can be as much as 10 degrees warmer at night (cold air sinks).
- Ensure the opening faces north in the Northern Hemisphere and south in the Southern Hemisphere so the sun doesn't shine directly in during the day.
- Try to dig down to create a depression, as the ground is cooler below the surface.
- Build your shelter for the worst possible conditions and not the conditions at the time. Desert weather can change very quickly and dramatically, and the ferocious winds will rip apart any shelter that is not secure.
- If your shelter sheet has a shiny side, make sure it faces up to reflect heat and act as an aid to location from above.

DESERT SCRAPE

If you have cord, you can dig a scrape and use the cord to peg out the sheet above the scrape. If you don't have cord the sheet will have to be held in place by other means, such as soil, sand, or rocks. With all layered desert shelters try to maintain tautness and separation between the layers.

CORDLESS SCRAPE

If you don't have enough cord to peg out a sheet, you can make a scrape by either digging down or building up the sides with rocks or sand mounds placed at regular intervals. Anchor your sheet with rocks.

1 If the hollow isn't deep enough to lie in, make a digging tool and scrape out the soil from the bottom to build up the sides of the hollow.

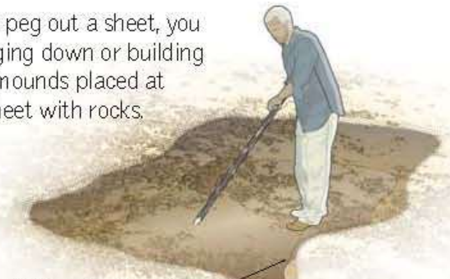

Build up the sides with rocks covered with soil

2 Spread your poncho, shelter sheet or tarp across the top of the hollow.

- Make sure the edges of the material overlap the sides of the hollow.

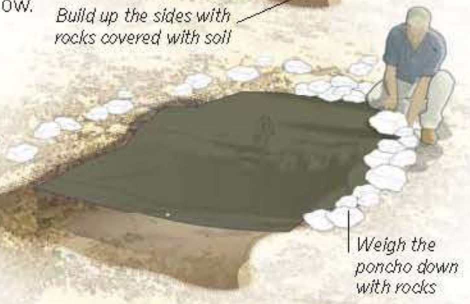

Weigh the poncho down with rocks

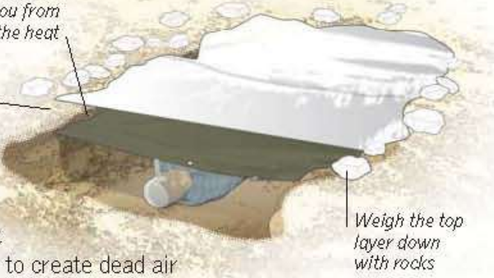

Air space helps to insulate you from the heat

Make a gap of at least 6 in (5 cm) between the layers

Weigh the top layer down with rocks

3 Use an additional sheet to make a second layer to create dead air space that helps to reduce the temperature below.

- If you only have one piece of material to use as cover, try to double it over to create the two layers.

QUICK SHELTER

If you can't find a hollow, erect a poncho shelter (see pp. 158–59) in a place that keeps you cool during the day—for example, under existing shade, such as trees or bushes, or at the top of a slight rise in order to benefit from any cool breeze.

1 Find a site beside a tree or secure a post next to where you want the shelter's opening to be.

- Rig a ridgeline to the tree or post, and peg your poncho over it to form a shelter.
- Repeat with a second poncho, shelter sheet, or space blanket to create a separate layer.

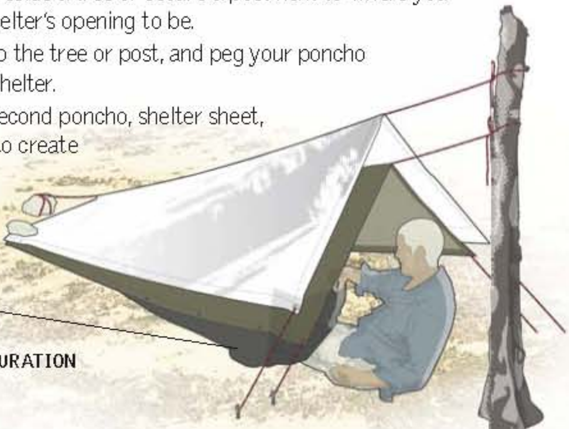
Leave sides and bottom open to air the shelter

DAY-TIME CONFIGURATION

2 At night, de-rig the outer shelter and use it as a sleeping cover for warmth.

- Prevent heat escaping by securing the sides and bottom with rocks, stones, shingle, or sand.
- Sleep on some bedding if possible.

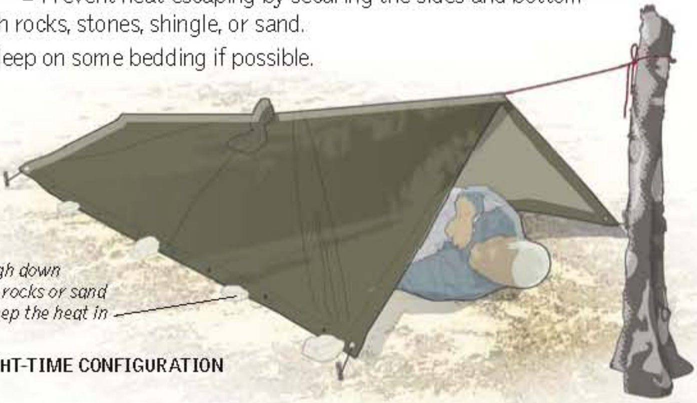
Weigh down with rocks or sand to keep the heat in

NIGHT-TIME CONFIGURATION

PARA SCRAPE

If you have cordage you can tie your sheet to four posts. If you haven't got enough cordage for four, use your backpack or a pile of stones instead.

- Prepare the depression or hollow as for a cordless scrape (see left).
- Use trees if available, or improvise posts from wood, piles of stones, or your backpack.
- Tie the sheet to the trees or the posts above the depression, leaving a gap for air to flow over.
- Create a second layer at least 6 in (15 cm) above the first.
- If foliage is available, place it between the layers to maintain separation.

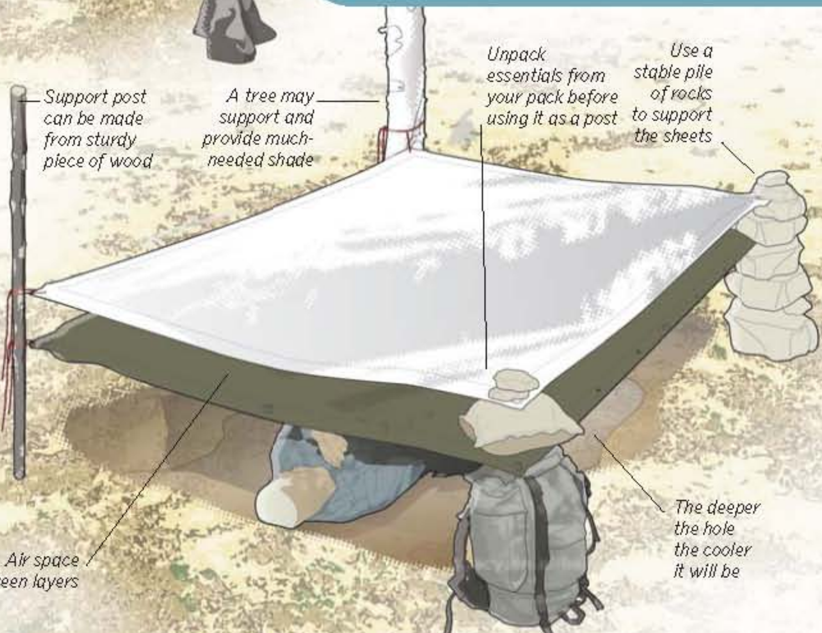
Support post can be made from sturdy piece of wood
A tree may support and provide much-needed shade
Unpack essentials from your pack before using it as a post
Use a stable pile of rocks to support the sheets
The deeper the hole the cooler it will be
Air space between layers

SUN PROTECTION

You should always carry immediate protection against the sun. You may be able to create shade using something you've brought, such as a trekking umbrella, tent, or space blanket.

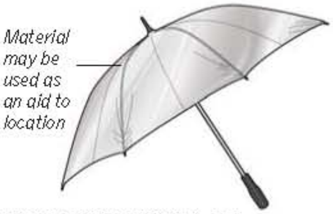
Material may be used as an aid to location

TREKKING UMBRELLA

A small, lightweight umbrella protects against sun, wind, and rain, creates cool shade, and doubles as a walking stick. Some have reflective material, or a flashlight in the handle.

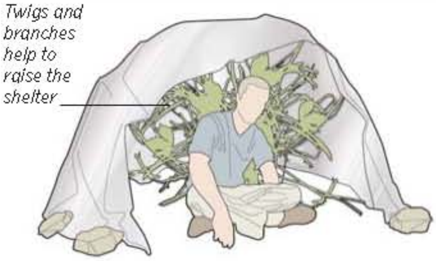
Twigs and branches help to raise the shelter

SPACE BLANKET

This should be part of your basic survival equipment and will afford immediate relief from the sun. Remember that it's almost impossible to pack a space blanket to its original size once it's unpacked.

SHORELINE SHELTERS

IF YOU FIND YOURSELF in a survival situation, and can make your way to a shoreline or beach, it may offer you particular benefits. Rescue is more likely here than further inland as your aids to location work better in open areas, people are more likely to live on the coasts, and useful flotsam washes up twice daily on even the remotest of beaches.

TOOLS AND MATERIALS

- Digging tool
- Driftwood and/or rocks
- Broad-leaved plants or grass, or poncho or tarp
- Knife
- Cordage and pegs

DRIFTWOOD SHELTER

If there's enough driftwood around on the shoreline, you can use it to make a variety of shelters covered in this chapter. A hole-in-the-ground driftwood shelter offers simple protection, but may need continued attention if the sand is very fine or gets wet.

BESIDE THE WATER

Shorelines vary with the environment—from tropical beaches, where sleeping out under the stars is not a problem, to rugged coasts in higher latitudes, where spending a night unprotected from the elements would be suicidal.

POINTS TO REMEMBER

If you intend to spend a night on a shoreline, exposed to the elements, the following tips may help you:

- Build the shelter above the highest high-tide mark on a seashore (see opposite) and above the highest watermark on a river or lake beach. If in doubt, move slightly further inland, where the protection may be better and materials more easily available.
- Try to establish what the weather's going to do by looking at the sky, watching the water, or detecting changes in the wind (see pp. 82–83).
- Plan for the worst, or at least have a back-up location should things not work out.
- Check the area for insects, such as midges, mosquitoes, gnats, and horseflies.
- Look for signs of wildlife, such as crabs and even turtles as sources of food.
- Finish your shelter and collect water and firewood before it gets dark.
- Beware the effects of sun and windburn, even on overcast days.
- Utilize driftwood and other building materials. Remember, any wood that's been in the sea is likely to be very heavy.

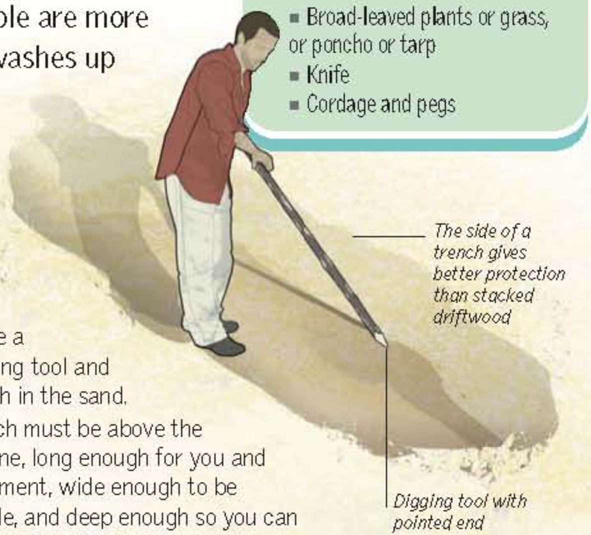
The side of a trench gives better protection than stacked driftwood
Digging tool with pointed end

① Make a digging tool and dig a trench in the sand.

- The trench must be above the high-tide line, long enough for you and your equipment, wide enough to be comfortable, and deep enough so you can roll over without disturbing the roof.

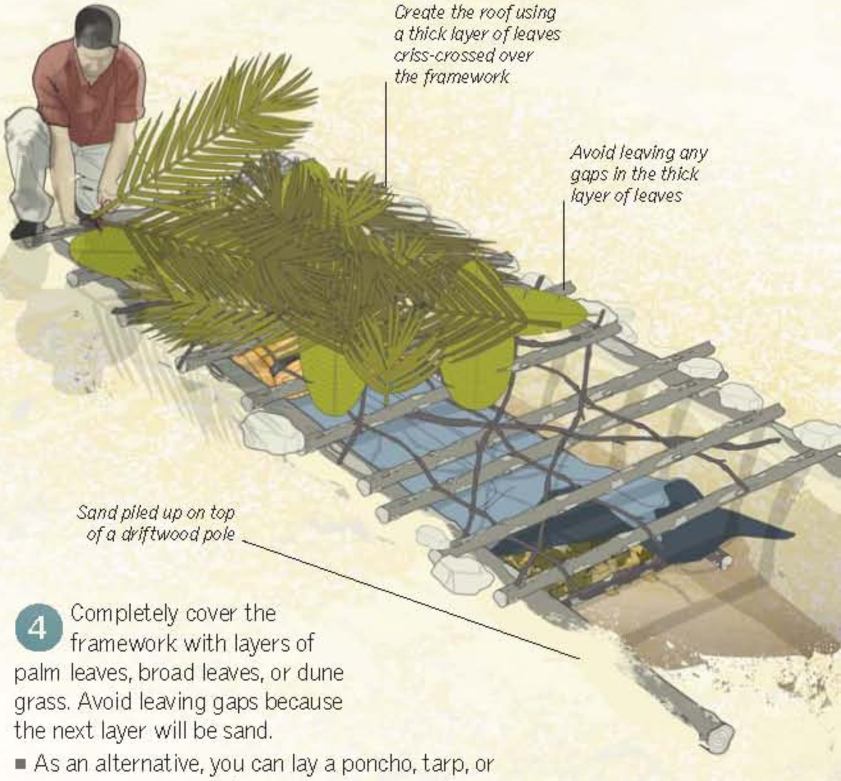
Create the roof using a thick layer of leaves criss-crossed over the framework
Avoid leaving any gaps in the thick layer of leaves
Sand piled up on top of a driftwood pole

④ Completely cover the framework with layers of palm leaves, broad leaves, or dune grass. Avoid leaving gaps because the next layer will be sand.

- As an alternative, you can lay a poncho, tarp, or shelter sheet over the framework. Weigh it down with sand or stones, or peg it out if you have cordage.

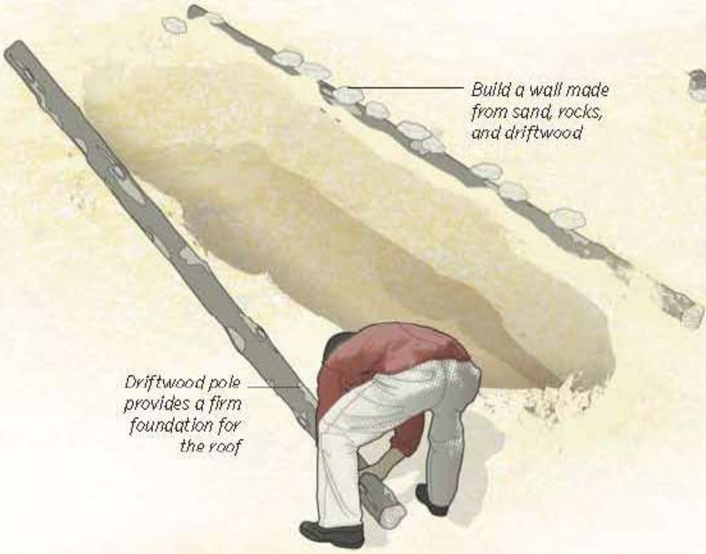
Build a wall made from sand, rocks, and driftwood
Driftwood pole provides a firm foundation for the roof

② Find two pieces of driftwood that are as long as your trench. Several shorter pieces in a line will work equally well if you can't find long pieces.

- Place the driftwood beside the trench and pile on sand and rocks to build up the height of the walls.

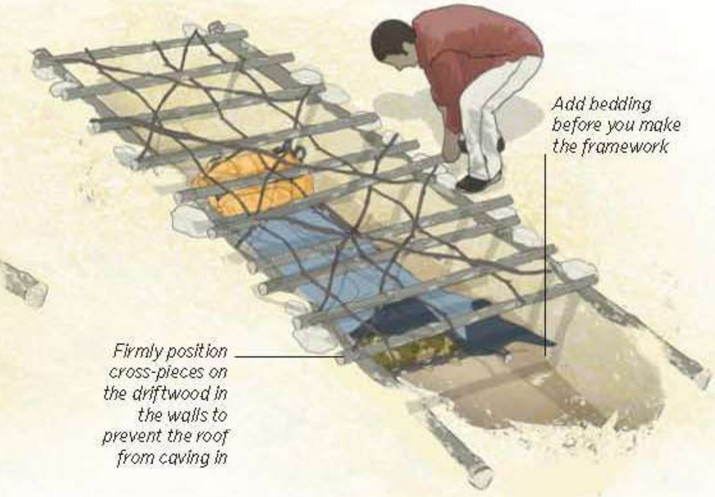
Add bedding before you make the framework
Firmly position cross-pieces on the driftwood in the walls to prevent the roof from caving in

③ Place pieces of driftwood across the trench to form the framework of a basic roof.

- Weave smaller pieces of driftwood into the framework, making sure they sit firmly on the driftwood in the walls.

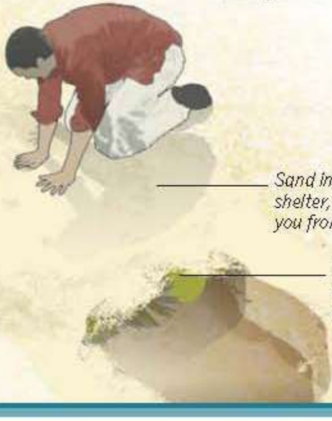
Sand insulates the shelter, and protects you from the elements
Bedding material is essential for keeping warm at night

⑤ Once you have covered the framework sufficiently, finish the shelter with a thick layer of sand or soil.

- The more cover you put on the roof, the more protected from the elements you will be. However, take care not to put so much weight on it that it collapses.

USING A LIFERAFT AS SHELTER

Remember, if you have abandoned ship in a liferaft, it can be used as a ready-made immediate shelter—even on land.

ON THE BEACH

The best place for a shelter is the landward side of the backshore: you can watch the sea, your daily fire can be seen from the sea, and you can see your shelter as you scavenge along the beach. In bad weather, however, the landward side of the dune crest is the best place.

PARTS OF A BEACH

A beach is shaped by the tides and the berm (a natural ridge) consists of deposits of the materials that make up the changing shoreline (sand, shingle, shale, and so on).

- At the top of the berm is a crest. A slope, or face, leads down from the crest to the water. At the very bottom of the face there may be a trough.
- The storm beach extends inland and this is where the wind and storms blow small particles of sand.
- Dunes form where the wind creates larger deposits of sand behind the beach.
- Longshore bars (sandbanks) may lie further out to sea and are formed where the waves first start to break.

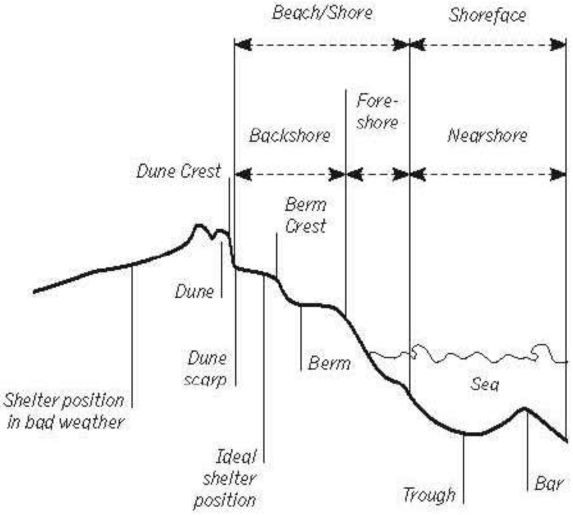
Beach/Shore
Shoreface
Fore-shore
Backshore
Nearshore
Dune Crest
Berm Crest
Dune
Dune scarp
Berm
Sea
Shelter position in bad weather
Ideal shelter position
Trough
Bar

SNOW SHELTERS

THE TYPE OF SNOW SHELTER you can build depends on the kind of snow, the equipment you have, and the opportunities offered by the environment—for example, wooded areas are usually better than open areas because they provide protection and natural materials. You could build a snow trench, snow cave, tree pit, snow ledge, or a quinzhee.

TOOLS AND MATERIALS

- Boughs and branches
- Bedding
- Long pole
- Spade or pan for shovelling snow
- Long knife or saw
- Knife (or other long blade)
- Tarpaulin or shelter sheet

COMPACTED SNOW TRENCH

If the snow is compact enough, and if you have a long knife or snow saw, you can cut out blocks of snow to form a trench and then use the blocks to form an apex roof. This design requires a lot of effort and a bit of practice, but gives you a solid shelter with some additional height.

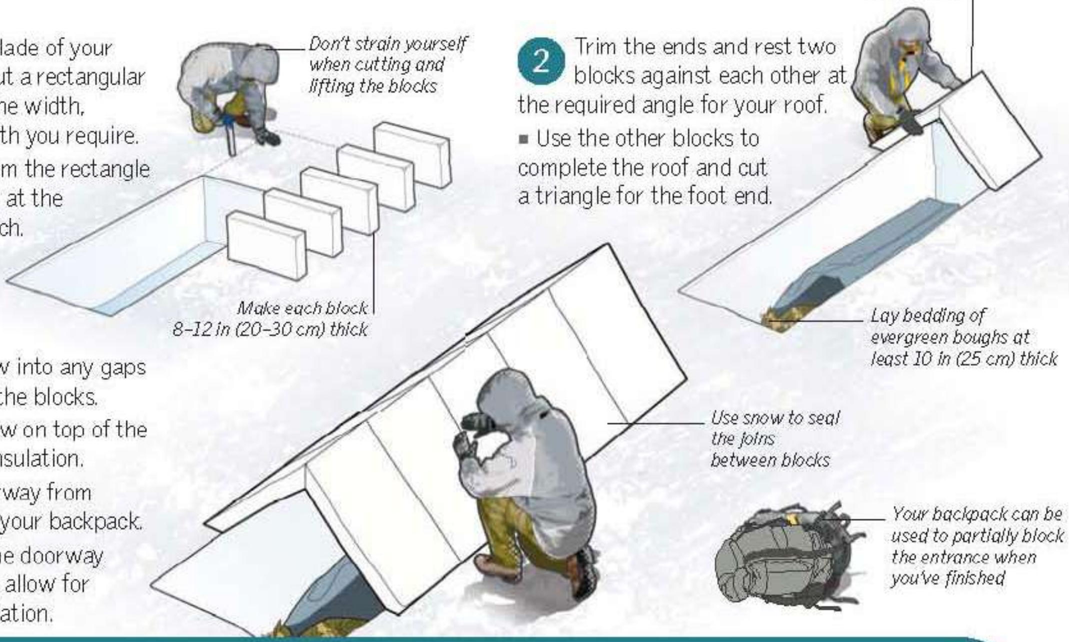

1 Use the blade of your knife to cut a rectangular trench that is the width, length, and depth you require.

- Cut blocks from the rectangle and place them at the side of the trench.

Don't strain yourself when cutting and lifting the blocks

Make each block 8–12 in (20–30 cm) thick

2 Trim the ends and rest two blocks against each other at the required angle for your roof.

- Use the other blocks to complete the roof and cut a triangle for the foot end.

Carefully balance and fit the trimmed blocks, starting at the foot end

Lay bedding of evergreen boughs at least 10 in (25 cm) thick

3 Pack snow into any gaps between the blocks.

- Pile extra snow on top of the blocks to add insulation.
- Create a doorway from snow blocks or your backpack.
- Don't block the doorway completely, but allow for adequate ventilation.

Use snow to seal the joins between blocks

Your backpack can be used to partially block the entrance when you've finished

BUILDING SNOW SHELTERS

Study the snow around you to determine whether it's compacted or uncompacted before you start.

POINTS TO REMEMBER

The following tips may help if you intend to build and stay in a snow shelter:

- Make a shelter that's just big enough for you and your equipment. Don't spend hours building a shelter for only one night: do the least amount of work for the maximum amount of protection.
- If you need to peg out guylines in the snow, tie a short stick 6–12 in (15–30 cm) long to a line or cordage and bury it in the snow. Compact the snow down on top and, when the snow sinters (hardens), your peg will be held secure.
- Snow is an excellent insulator. Fresh, uncompacted snow is typically 90–95 percent trapped air. Since the air barely moves, the snow can keep you warm and dry if used correctly.
- Check your site for hazards, such as snowdrifts, freezing winds, avalanche, cornice collapse, and big animals.
- Create ventilation holes—make one near the ground to let fresh air in and one at the top to let air escape. Check the holes are clear every 1–2 hours.
- Keep tools inside the shelter in case you have to dig your way out.
- If you leave the shelter, mark the entrance. Take basic survival equipment with you: it's better to have and not need than to need and not have.
- Brush snow off your equipment and clothing before entering the shelter.
- Tie all vital equipment to yourself so that you can't accidentally drop and lose it in deep snow.

FOUND SHELTERS

If you're lucky you may find a tree that has space under it so you don't have to dig a trench. For example, under a low bough or in a tree pit—an area where a heavy and deep snowfall has built up around a tree, leaving little or no snow under the lower branches. Use your walking stick to check the depth of the snow, and also to check for pockets of air.

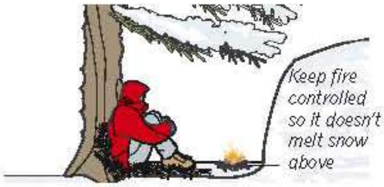
Keep fire controlled so it doesn't melt snow above

LOW BOUGH

Find an evergreen tree with a large branch lying on top of the snow. Dig down from the leeward side, excavate a pit, insulate the floor with boughs, and make sure you have adequate ventilation.

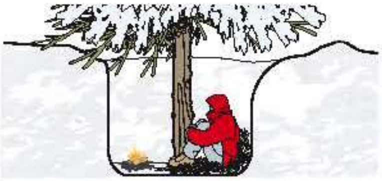

TREE PIT

Test the snow around a tree that has large, low, snowbound branches. Little or no resistance indicates a pit. Dig from the leeward side, place boughs on the floor, and ventilate the pit.

FIGHTER TRENCH

If the snow is soft, a fighter trench is quick and easy to build. In an emergency, you can even make a trench by kicking out the soft snow. First, find a location that protects you from the elements as much as possible, then test the depth of the snow with your walking stick or a pole.

WARMING AND COOLING TIPS

Whether you're building a shelter or resting inside it, the following tips may help you stay at the right temperature:

- Remove layers of clothing as you work, perhaps to your base layers with a waterproof layer on top. Keep clothing dry so that when you stop working you can put it on again.
- Keep shelters at a constant temperature. If snow melts and then re-freezes, it stops insulating.
- Avoid heating your shelter too much. Even a candle flame can raise the temperature inside by four degrees.

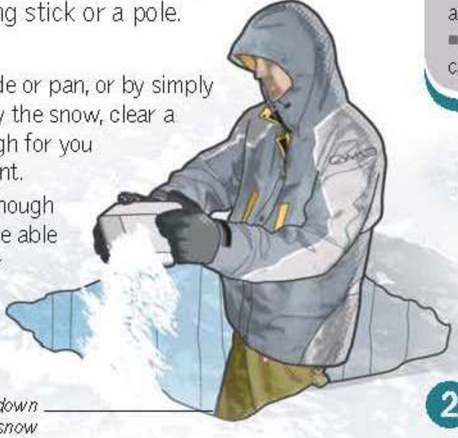
If you can, dig down through the soft snow to the ground

1. Using a spade or pan, or by simply kicking away the snow, clear a trench large enough for you and your equipment.

- Allow at least enough depth for you to be able to roll over in your sleep without disturbing the roof.

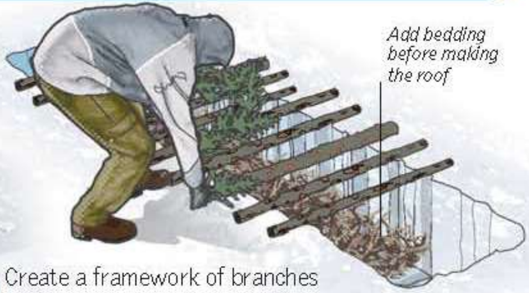
Add bedding before making the roof

2. Create a framework of branches and boughs across the trench.

- Make sure you have enough roofing material on top to stop the snow from penetrating through.

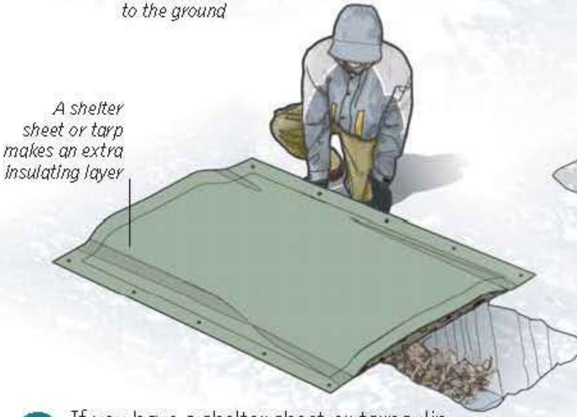
A shelter sheet or tarp makes an extra insulating layer

3. If you have a shelter sheet or tarpaulin, place it on top of the framework to form an extra insulating layer.

- Alternatively, you can use it to cover your bedding.

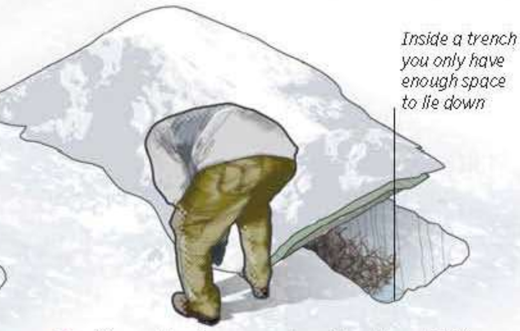
Inside a trench you only have enough space to lie down

4. Cover the framework with at least 12 in (30 cm) of snow to act as insulation.

- Dig a small pit at the entrance to allow easy access.
- At the entrance, you can make a small, controlled fire on a green log base.

QUINZHEE

A quinzhee is a dome-shaped shelter made by hollowing out a pile of settled snow. It's an overnight shelter that is easier to construct than the more permanent igloo ("igloo" means "home"), which is made from cut blocks of snow and requires skill and knowledge. You can't stand up in a quinzhee, but you can sit upright or just curl up.

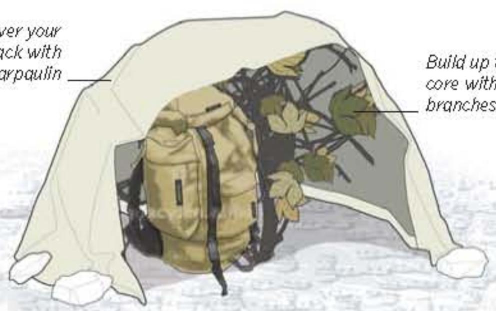

Cover your backpack with a tarpaulin

Build up the core with small branches and twigs

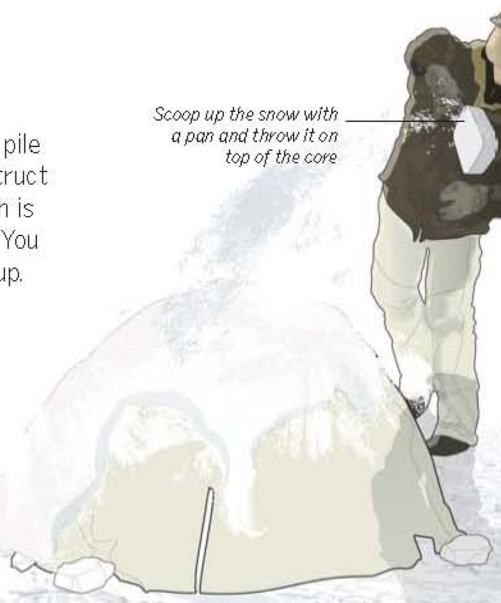

Scoop up the snow with a pan and throw it on top of the core

1 Find a relatively flat area that's covered with an abundance of snow.

- Mark a circle for your quinzhee, including walls that will be about 10 in (25 cm) thick, and stamp down the snow.
- Use your backpack and boughs or leaves covered with a tarpaulin to form the core of the shelter. Site the doorway at 90 degrees to the prevailing wind.

2 Use a container, pan, snowshoe, or other suitable item to gather as much soft snow as you can.

- Pile snow on the shelter's core to form a dome of the required height. Build up layers of snow until you have a covering of at least 10 in (25 cm), smoothing out each layer as you go.

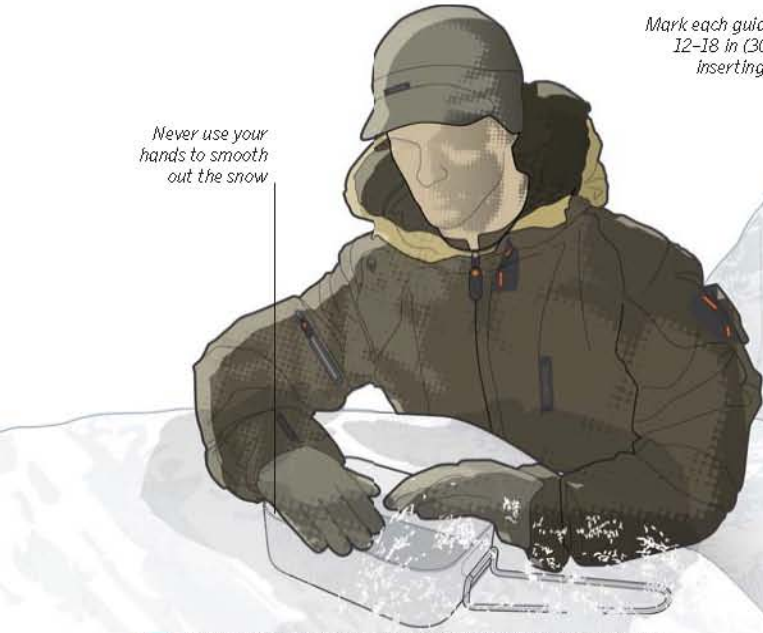

Never use your hands to smooth out the snow

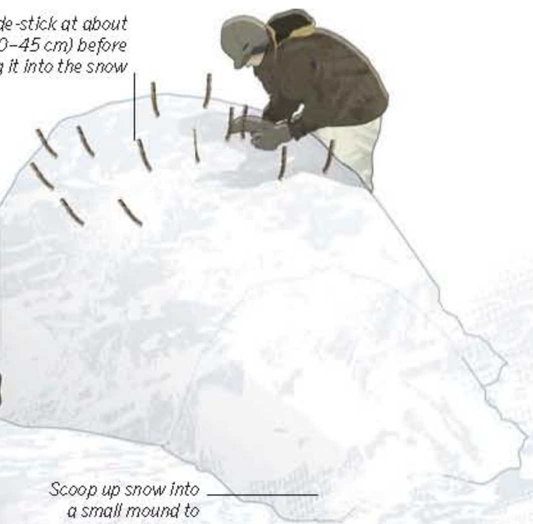

Mark each guide-stick at about 12–18 in (30–45 cm) before inserting it into the snow

Scoop up snow into a small mound to form the entrance

3 Smooth out the snow on the dome and leave it to sinter (harden) for 1–3 hours, depending on the type of snow and the ambient temperature.

- Keep active during this waiting period, particularly if it's windy: collect wood, build a fire (away from the quinzhee), and prepare yourself for the night.

4 To get an even thickness in the roof and walls, push guide-sticks of equal length through the snow toward the center of the dome.

- Build a small compact mound in front of the dome.

SNOW CAVE

A snow cave provides good protection but it takes a lot of work. Suitable locations might be compact snow on the lee side of a hill or an established snowdrift with a firm crust. Avoid snow that's newly fallen, powdery, loose, or shallow and uneven. The cave should be high enough to sit up in and deep enough for you and your equipment. Locate the entrance 45 degrees downwind to keep it free of drift.

BUILDING THE CAVE

- Dig a tunnel 3 ft (1 m) long into the snow bank, then excavate a cave on two levels. The lowest is the tunnel, while the other is a ledge at least 2 ft (70 cm) higher, and wide enough to sleep on.
- Create an arched ceiling to give the cave more support. Make it at least 18 in (45 cm) thick so it can take the weight of heavy snow.
- Mark the area with boughs to warn people where the shelter is.
- Insulate the sleeping ledge with plenty of natural material.
- Make at least one ventilation hole in the roof, but not facing into the wind.

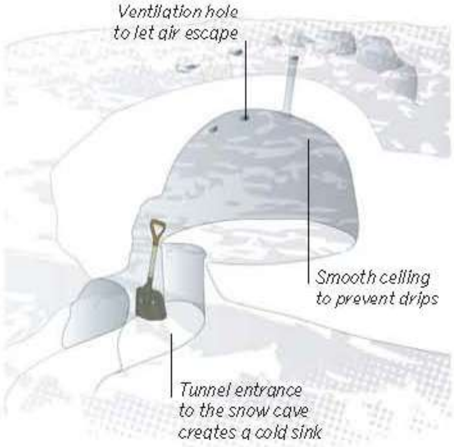
Ventilation hole to let air escape
Smooth ceiling to prevent drips
Tunnel entrance to the snow cave creates a cold sink

SNOW LEDGE

You can build a snow ledge from a well-packed drift of snow out of the wind, or at least perpendicular to it.

MAKING THE SNOW LEDGE

- Dig out blocks to form a doorway about 2 ft (70 cm) into the snow. Dig out blocks to form a horizontal rectangle above.
- Dig upward to create a sleeping platform and then form an arched ceiling.
- Place the cut blocks across the horizontal rectangle and seal any gaps with snow.
- Make a ventilation hole in the roof and one lower down to let fresh air circulate.

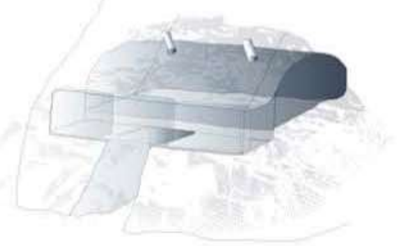

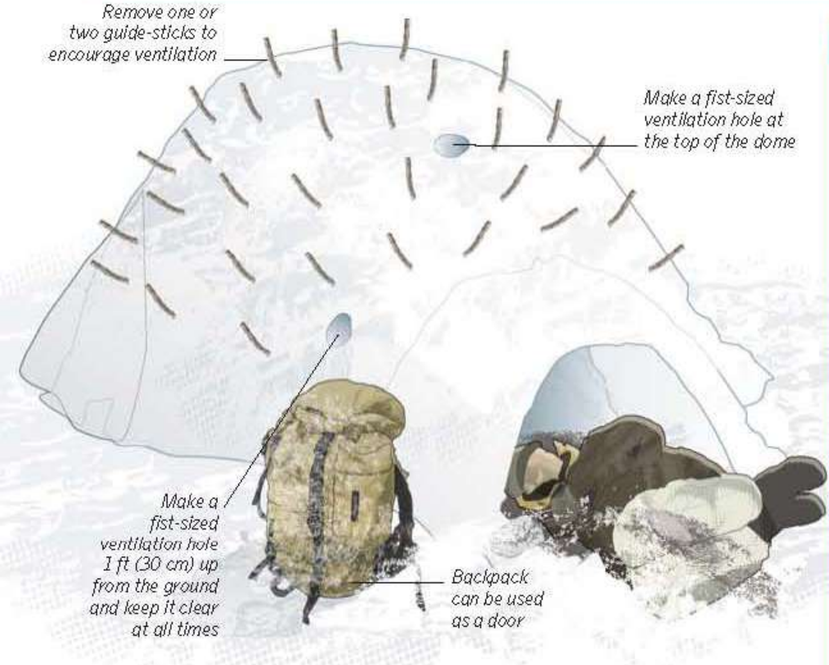
Remove one or two guide-sticks to encourage ventilation
Make a fist-sized ventilation hole at the top of the dome
Make a fist-sized ventilation hole 1 ft (30 cm) up from the ground and keep it clear at all times
Backpack can be used as a door

5 When the snow has hardened, burrow into the mound, remove your backpack and tarpaulin, and excavate snow from the core.

- Use the guide-sticks to keep the walls at least 10 in (25 cm) thick.
- Don't use your hands to dig as they will get cold and wet.
- Smooth out the snow on the inside to prevent drips from forming.
- Build a raised sleeping platform. This creates a cold sink (an area into which the coldest air falls), taking cold air away from your sleeping area.

CARBON MONOXIDE

Carbon monoxide is an odorless gas that's produced when there's not enough oxygen to create carbon dioxide from burning fuel. Carbon monoxide poisoning can be fatal in a well-insulated, non-ventilated environment.

PREVENTION IS BETTER THAN CURE

Create one or more holes, 3–4 in (7.5–10 cm) in diameter, at the base and top of your shelter. Make sure the air flows out unobstructed.

SIGNS OF POISONING

Carbon monoxide poisoning is cumulative and can build up over a few days. Mild effects are fatigue, faintness, and flu-like symptoms. As it progresses, the effects are severe headaches, nausea, and decreased mental coordination.

TAKE ACTION

Get into the fresh air at once. You need to breathe fresh air for at least four hours to reduce the amount of carbon monoxide in your system by half.

DETECTORS

Carbon monoxide detectors are widely available, but tend to be made for the home or RV. Some models are battery-operated, but they're too bulky for camping. You can wear a patch that changes color when carbon monoxide is in the air, giving an early visual warning.

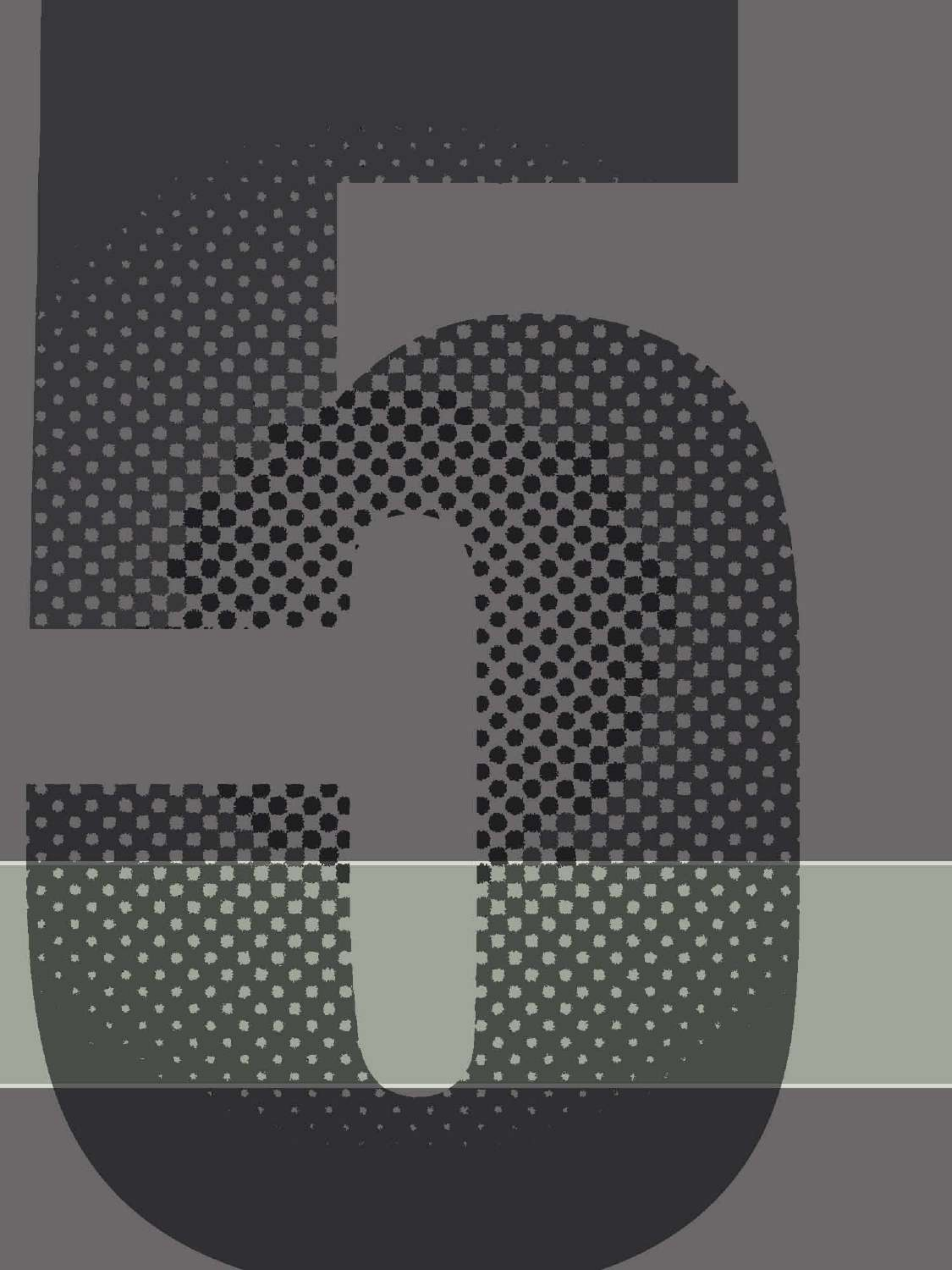
5

WATER AND FOOD

FIND AND TREAT WATER

The importance of water, even in a short-term survival situation, should never be underestimated. Water is essential to life, and a regular intake of 2–3 liters ($4\frac{1}{5}$–$6\frac{1}{3}$ pints) a day is needed just to maintain your water balance and prevent dehydration. The amount required can increase dramatically depending on factors such as the temperature of the environment, your age and physical condition, your workload, and whether you have been injured. It's not unusual for UK Royal Marines operating in desert or jungle environments to require 14 liters ($3\frac{3}{4}$ gallons) per day. We tend to take water for granted, not appreciating just how important it is until we don't have any—at which point it becomes the most important thing in the world.

You should always plan your treks around your need for water and your ability to replace it as required. There are many hydration systems available and many small and efficient methods of filtering and purifying water on the trail. In a survival situation, you should always strive to filter and purify any

In this section **YOU WILL DISCOVER...**

- that **birds** can **show you** the **way** to **water**...
- when to **suck** on a **small pebble**...
- how to **make** a **gypsy well** and a **solar still**...
- why a **Finnish marshmallow** could **save your life**...
- how to **absorb water** without **drinking it**...
- how to **improvise** a **basic bladder**...
- the **importance** of **surgical tubing**...

water before drinking it. Bringing water to the boil will kill all water-borne diseases. In the short-term, stomach bugs from contaminated water may not kill you but they can seriously affect your ability to carry out other survival tasks. However, if you have no choice, it's better to drink contaminated water than not to drink at all. That way, a doctor will at least be able to treat you, whereas dehydration will kill you—and death can't be treated! Never drink urine or salt water, as these will only dehydrate you more.

If a source of water is undrinkable, such as salt or stagnant water, or even urine, you can still produce drinking water if you have the means to start and maintain a fire.

THERE ARE MANY WAYS to purify water, but if you have a fire you will always be able to distill it—and therefore make it drinkable. Build your fire close to the water source (if the water can be poured into a container or a hole dug in the ground where it will not seep away quickly, this will make the process more efficient). When the fire is established, place rocks in it (don't use slate or other layered rocks as they might shatter). Once the rocks are heated, use a forked stick or similar to transfer them into the salt or foul water. Then, suspend a piece of absorbent material, such as a T-shirt or moss, over the hot rocks to collect the steam. The steam will condense on the material and the resulting water, which you can wring out, will be fit to drink. If you have a cooking pot or survival tin, you can simply fill it with water, boil it over a fire, and collect the steam in the same way.

If you have absolutely no means of treating or boiling water, you should try to:

- Find the clearest flowing water and collect it from the surface.
- Filter debris out, even if only through a sock.
- Remember, it's better to drink foul water than not to drink at all.

> “We tend to take **water for granted** until we don't have any—at which point it becomes the most **important thing in the world**”

THE IMPORTANCE OF WATER

YOU REQUIRE A STEADY supply of water to sustain yourself in a survival situation and without it you will dehydrate. Left unchecked, dehydration will end in death. To survive, a balance between water intake and loss must exist.

WHY YOU NEED WATER

Water is essential to life. It is needed, directly or indirectly, for every physical and chemical process that takes place in your body. Here are just a few of the functions that water performs:

- **Delivery service:** water carries oxygen, nutrients, and other essentials around the body.
- **Waste remover:** the kidneys use water to flush out toxins via urine.
- **Coolant:** water regulates the body's temperature.
- **Breathing aid:** the lungs use water to moisten inhaled air so that it doesn't irritate the sensitive pulmonary linings.
- **Sensory aid:** water helps conduct nervous impulses around the body.
- **Shock absorber:** water protects the vital organs and provides lubrication around the joints.

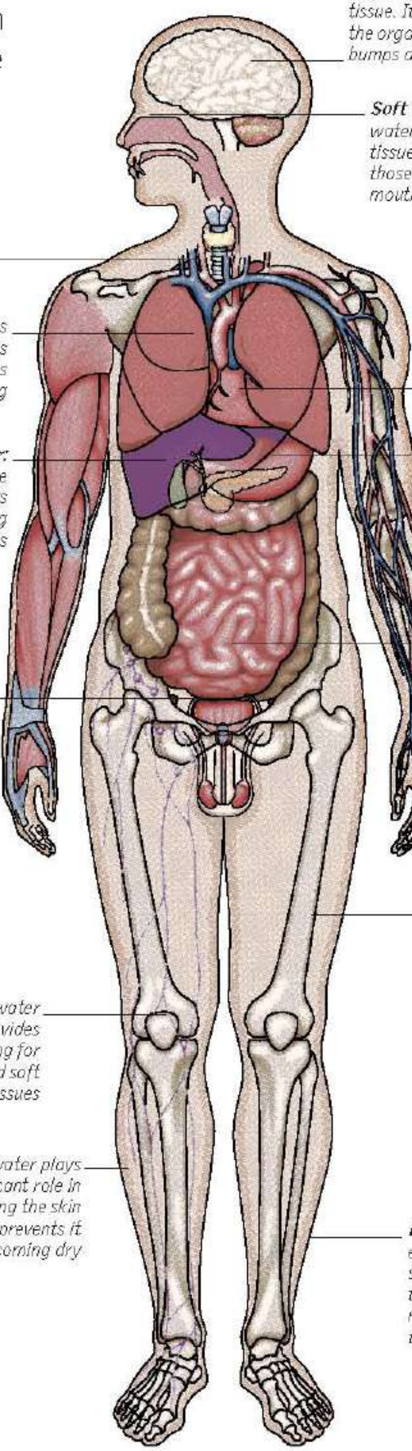

Brain: *water makes up approximately 80 percent of brain tissue. It protects the organ from bumps and jarring*

Soft tissue: *water moistens tissues such as those in the nose, mouth, and eyes*

Blood: *water plays a significant role in regulating blood pressure*

Heart: *higher water consumption helps remove fatty acids from the body and reduces the risk of a heart attack*

Lungs: *water helps moisten the lungs and assists breathing*

Stomach: *water helps the body digest food in the stomach and turn it into energy*

Kidneys and liver: *water lessens the burden on kidneys and liver by flushing out waste products*

Mineral absorption: *water helps dissolve minerals and other nutrients to make them accessible to the body*

Bladder: *water enables the body to flush out waste and toxins through the bladder*

Bones: *water makes up approximately 22 percent of bones*

Joints: *water provides cushioning for joints and soft tissues*

Skin: *water plays a significant role in hydrating the skin and prevents it from becoming dry*

Body temperature: *evaporation of sweat (water) from the skin helps regulate body temperature*

HOW MUCH WATER DO YOU NEED?

How much water you need to survive a particular situation is dictated by a number of factors, such as your physical state, the environment you are in, and your exertion levels. Even when resting in the shade, the average person will lose more than 1 liter (1¾ pints) of water each day just through breathing and urination, a figure that increases dramatically once loss of water through sweat is taken into account. A minimum of 3 liters (5¼ pints) per day is required to remain healthy in a survival situation, with this amount increasing for higher temperatures and heavier workloads.

WHAT IS DEHYDRATION?

Dehydration occurs when you fail to replace the water your body loses. It's vital to recognize the symptoms early. Factors that can lead to dehydration include high and low temperatures, humidity, work-rate, clothing, body size, fitness levels, and injury.

THE EFFECTS OF WATER LOSS		
1-5% LOST	**6-10% LOST**	**11-12% LOST**
▪ Thirst ▪ Discomfort ▪ Urine becomes darker ▪ Loss of appetite ▪ Impatience ▪ Drowsiness ▪ Lethargy ▪ Nausea ▪ Headache	▪ Dizziness ▪ Dry mouth ▪ Blueness of extremities ▪ Slurred speech ▪ Swollen tongue ▪ Blurred vision ▪ Tingling in limbs ▪ Inability to walk ▪ Difficulty in breathing	▪ Stiffness of joints ▪ Deafness ▪ Defective vision ▪ Shriveled skin ▪ Lack of feeling in skin ▪ Inability to swallow ▪ Delirium ▪ Unconsciousness ▪ Death

WATER-BORNE DISEASES

Water-borne diseases are caused by ingesting water contaminated by the faeces or urine of humans or animals that contains protozoa, viruses, bacteria, or intestinal parasites. Globally, they cause 10 million deaths a year.

	DISEASE	SYMPTOMS
PROTOZOA	Cryptosporidium	Loss of appetite, nausea, and abdominal pain, usually followed by profuse, foul-smelling, watery diarrhea, and vomiting.
	Giardiasis	Loss of appetite, lethargy, fever, vomiting, diarrhea, blood in the urine, and abdominal cramps.
VIRUS	Infectious hepatitis (Hepatitis A)	Nausea, loss of appetite, mild fever, aching muscles, dark-colored urine, jaundice, and abdominal pain.
BACTERIUM	Amoebic dysentery	Feeling of fatigue and listlessness. Feces may be solid, but will smell foul and contain blood and mucus.
	Bacillary dysentery (Shigellosis)	Fever, abdominal pain, muscle cramps, high temperature, and blood, pus, and mucus in stools.
	Cholera	Vomiting, poor circulation, cold and clammy skin, muscle cramps, rapid dehydration, and increased heart-rate.
	E. coliform	Diarrhea and vomiting. Can cause death in vulnerable groups such as the very young or the elderly.
	Leptospirosis	Jaundiced appearance, lethargy, high temperature, aching muscles, and vomiting. Can be fatal if not diagnosed early.
	Salmonella	Nausea, diarrhea, headaches, stomach cramps, fever, possible blood in the feces, and vomiting.
PARASITES	Bilharzia	Irritation to the urinary tract and blood in urine, rash or itchy skin, abdominal pain, cough, diarrhea, fever, and fatigue.
	Hookworms	Anaemia and lethargy. Larvae travel to lungs and are coughed up and swallowed into the stomach, where they grow into worms.

TOO MUCH WATER?

Hyponatraemia is a condition caused when excess water accumulates in the body at a higher rate than it can be excreted. It results in a diminished sodium concentration in the body's plasma and the swelling of the body's cells. It can lead to a swollen brain and other neurological problems and, in extreme cases, coma and death. The way to prevent hyponatraemia is to control the amount of water that you drink, and to regulate your body's salt intake. If you don't have salt or sodium tablets in your survival tin, you can filter saltwater through fabric to sift out the salt content.

RATIONING YOUR WATER

If your water supplies are limited, you'll have to use what rations you have efficiently until you're rescued. If your water rations will not last that long, you'll have to procure water yourself. There is much debate about the advantages and disadvantages of drinking no water for the first 24 hours of a survival situation, but at this early stage it's best to make sure you're adequately hydrated. Your particular circumstances will dictate what's best to do, but always consider the following:

- The incident that put you in the survival situation may have been both dramatic and stressful; this will make you thirsty.
- In the first 24 hours of a survival situation, you'll be addressing the principles of survival: protection (shelter) and location (see p. 27). This is hard, thirsty work.
- Physical factors—such as seasickness, injury, or the environment you're in, such as a desert—may dictate that water-rationing is not a viable option.
- If you only have a limited water supply, but drink nothing for the first 24 hours, you may end up being so dehydrated that what little water you have in your possession will have no positive effect on your dehydrated state.

FINDING WATER: TEMPERATE CLIMATES

THE ABILITY TO LOCATE drinking water can challenge even the most experienced survivor and the challenges can vary dramatically according to the environment and the local conditions. It is, therefore, extremely important that you're aware of all the potential sources of water for the environment you're in.

WARNING!
In a survival situation, even if you think your chances of rescue are high and that help is imminent, you should start to look for water sources as soon as you have addressed the immediate problems of protection and shelter (see pp. 154–81). Remember that you can survive for up to three weeks without food; without water, you won't be able to survive for more than a few days.

COLLECTING WATER

You can find water in a range of sources, which vary in terms of quality and accessibility. Your priority is to locate the best and most accessible source of water in your immediate environment.

CATCHING RAINWATER

Collecting rain as it falls is the safest way to procure drinking water. It will need no treatment prior to drinking, as long as the catchment device itself has not been contaminated. Any number of non-porous materials—such as tarpaulins, ponchos, flysheets, survival blankets, or even large leaves—can be used as a rain-catchment device. Bear in mind that the larger the surface area of the material, the more rainwater you'll be able to catch.

1. Select a place as close to your camp as possible where your catchment device will be exposed to the most possible rain.

2. Firmly secure the tarpaulin to four stakes (sticks of equal length could be used) using string. Make sure that one end is higher than the other to provide a natural run-off for the water.

3. Place a heavy stone in the middle of the tarpaulin, roughly two-thirds of the way toward the lower end, to create a channel into which the rain will run down from the tarpaulin's sides.

4. Place a container, such as a pan or can, beneath the end of the channel to collect the water as it starts to flow off the tarpaulin.

MATERIALS AND TOOLS
- Tarpaulin
- Four sticks
- String
- Heavy stone
- Container

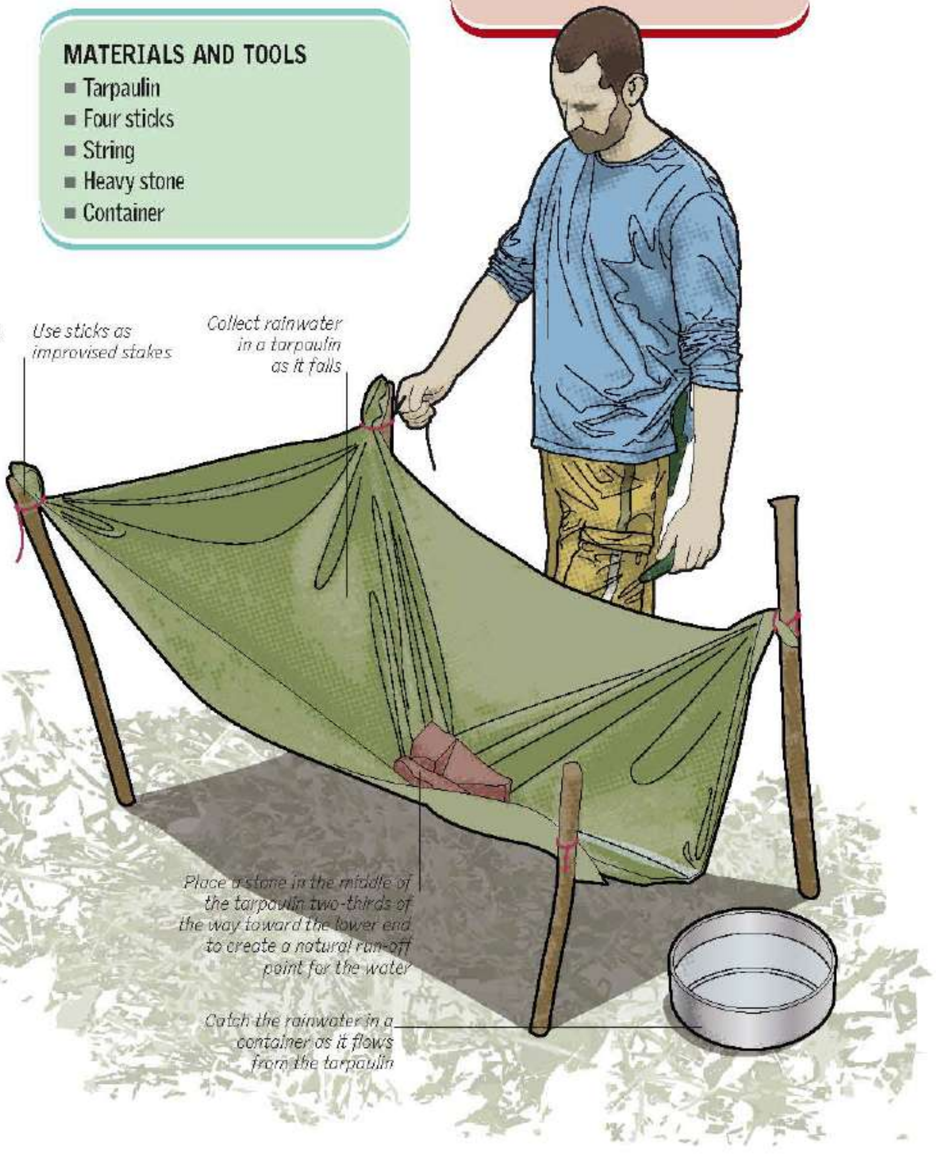

Use sticks as improvised stakes

Collect rainwater in a tarpaulin as it falls

Place a stone in the middle of the tarpaulin two-thirds of the way toward the lower end to create a natural run-off point for the water

Catch the rainwater in a container as it flows from the tarpaulin

LOCATING OTHER NATURAL WATER SOURCES

If rainwater is not available, there may be other natural sources of water, from easily visible streams and rivers, to bores and holes, which can be hidden. Whatever your source, you should always treat the water before drinking.

SOURCE	CHARACTERISTICS
SPRINGS	These occur when the water is forced to the surface as a result of subterranean pressures or from gravitational flow from higher sources. Found in low-lying areas, springs normally provide a permanent water source. Contrasting green vegetation is an indicator of their presence.
STREAMS AND RIVERS	Rivers and streams are an invaluable source of water, but there are still a few things you need to bear in mind when collecting water from them. The closer to the mountaintop the river or stream is, the clearer the fast-running water will be. The further downstream the water travels, the more likely it is to pick up minerals, debris, and pollutants that might be harmful to you. If possible, either check upstream for dead animals that may have contaminated the water, or follow the water downstream. Always try to collect the fast-flowing water near to the surface. In arid areas, rivers and streams tend only to flow during floods and will contain more pollutants and debris.
ROCK HOLES	Usually found in high ground, rock holes are natural collectors of rainwater. If the water appears to be trapped deep down, you can use your surgical tubing to retrieve it. Make sure you filter and purify the water before you drink it.
WELLS AND BORES	In some areas, you may find wells or old bore holes. Wells, which may be featured on local maps, can be deep and covered, making water procurement more difficult. In remote areas, wells are covered and marked in certain ways by the locals—find out what the markers are for your area.
LAKES AND PONDS	Rivers, streams, and water run-offs all flow into lakes or ponds. If you're collecting water from these sources, always try to collect it as it runs into the body of water, as lakes and ponds are more static than flowing water and can, therefore, become increasingly stagnant. Try to find the cleanest-looking area and avoid areas where debris has collected or where algae has formed. Note that the presence of fish indicates that the water is still oxygenated.
SEEPAGE	Usually located at the base of cliffs, high ground, or rocky outcrops, seepage is caused by slow-running channels that drain off these features.
SOAKS	Soaks are found close to rivers and creeks in low-lying areas, and are normally lower than the existing water table. The presence of soaks is often indicated by vegetation and they may be subject to pollution as a result of their use by animals. Make sure you filter and purify the water before you drink it.

AVOIDING POTENTIAL DANGERS

Most water sources are likely to be used by animals for drinking, bathing, urinating, and defecating, so always filter and purify any water collected before use (see pp. 200–01); the only exception to this rule is if you have collected rainwater. When collecting water, you should also be aware of the following dangers.

- It is possible that you'll encounter dangerous animals either using the water source, or on the way to or from the water source.
- Almost all major water sources will have a ranking system to determine which animals can use it. If all of the small gazelles suddenly disappear, ask yourself why.
- If you're collecting water from rivers, be aware of the potential dangers of river wildlife, such as crocodiles and snakes.
- If using dry riverbeds during the rainy season, be aware that flash floods can move quicker than you can run.

LOCATING HIDDEN WATER SOURCES

The presence of water is usually indicated by signs of life, such as green vegetation, animal tracks, or human habitation. Even if the terrain you're in appears lifeless, there may be plenty of indicators of a possible water source.

USE THE TERRAIN

- Observe the landscape for patches of green. Be aware that vegetation may not need obvious surface water for survival and may get its water from deep roots that tap into moisture below the surface.
- Water is subject to gravity and is more likely to be found downhill or in low-lying areas, such as valleys, dry riverbeds, narrow canyons, gullies, and at the base of cliffs or rock formations. Green vegetation forms next to a river and decreases as the ground rises away from the water source.
- Water will often seep inland on coastlines, leaving behind wetlands that contain water with tolerable levels of salt or that can be distilled in a solar still (see p. 189) to procure fresh water.

WATCH THE ANIMALS

- The Bedouin listen to the twittering of birds at dawn and dusk and follow their flight path to discover where they drink.
- Flocks of birds circling over one spot are usually flying over a water source. This does not apply to meat-eating birds, such as vultures, eagles, or hawks, who get their fluid requirements from the meat they eat.
- All finches and grain-eaters need a regular supply of water. Observe their flight patterns to locate a water source.
- Bees need water, so beehives are never far from a water source.
- Animal tracks, especially those of herd animals, will often lead to a water source. Look for converging sets of tracks.
- Flies stay close to water and the presence of mosquitoes almost certainly means that water is nearby.
- Look out for herd animals, such as buffalo, hippos, elephants, impala, and wildebeest, as they depend on water.

PROCURING WATER

Even if you find yourself in an environment without any obvious water sources, it doesn't necessarily mean that water will not be available to you. There are a number of water-procurement techniques that may mean the difference between you making it through a survival situation or not.

SOURCING WATER FROM DEW

Dew is water in the form of droplets that form on exposed surfaces in the early morning or late evening and can provide an invaluable supply of fresh water. It occurs when the temperature of a surface is low enough to allow the moisture in the warmer air above it to condense. Dew can easily be collected from any non-porous surface—such as a car roof or a tarpaulin—with a piece of cloth that can then be wrung out into a container. You can also harvest or trap dew.

HARVESTING DEW

You can harvest dew by walking through a field of long grass before sunrise or late in the evening with a piece of absorbent material—such as rags or a T-shirt—tied around your ankles.

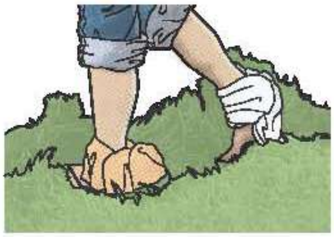

① Tie the material tightly around each ankle and walk through the dew-covered grass. It will absorb the dew as you move.

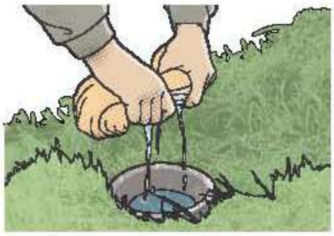

② Wring the rags to extract the water. Repeat the process until you have an ample supply of water or the dew has evaporated.

DELAYING DEHYDRATION

When water is scarce, the following points will help you delay the onset of dehydration:

- Conserve what water you have and use it as efficiently as possible.
- Work only in the coolest part of the day and avoid sweating.
- If the sun is shining, seek shade and keep covered.
- Suck a small, smooth button or pebble to help stimulate saliva and remove the sensation of thirst.
- Avoid eating protein-rich food as it requires more water to digest than those in other food groups.

MAKING A DEW TRAP

Dig a hole about 18 in (45 cm) deep, line with a plastic sheet, and fill with smooth, clean stones. Water will condense on the stones overnight. Harvest as early as possible the next morning to ensure it does not evaporate.

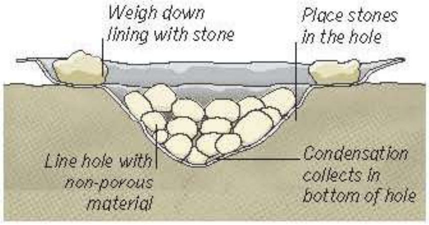
Weigh down lining with stone
Place stones in the hole
Line hole with non-porous material
Condensation collects in bottom of hole

SOURCING WATER FROM PLANTS

Transpiration is the evaporation of water from a plant, primarily from its leaves. You can collect this vapor to boost your fresh water supply. All you need is a clear plastic bag.

MAKING A TRANSPIRATION BAG

Place a smooth rock in the lower corner of the plastic bag, and place the bag over the leaves of a tree branch, tying the end. As water evaporates from the leaves, it will condense on the inside of the plastic bag and collect at its lowest point.

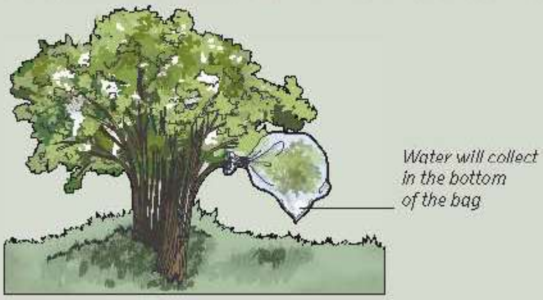
Water will collect in the bottom of the bag

MAKING A VEGETATION BAG

Cut green vegetation and place it in a plastic bag. Place a smooth rock in the lower corner of the bag and tie off the open end. Secure the bag in direct sunlight. The sun causes water in the leaves to evaporate, condensation forms on the plastic bag and drips to the lower corner of the bag.

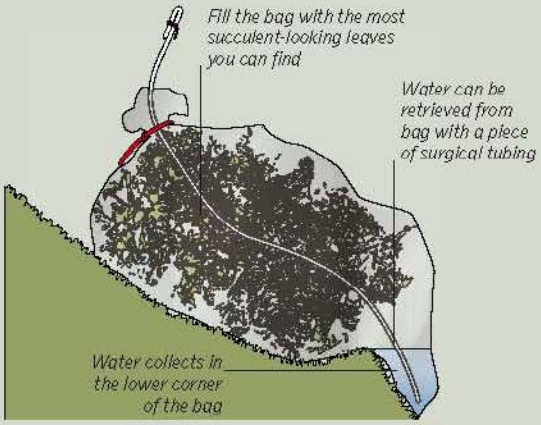
Fill the bag with the most succulent-looking leaves you can find
Water can be retrieved from bag with a piece of surgical tubing
Water collects in the lower corner of the bag

MAKING A SOLAR STILL

A solar still works using the same principle as a vegetation bag (see opposite). It collects potable water from the vapor that is produced by vegetation, water that is unfit to drink, or moisture from the ground.

1 Ideally find, or dig, a hole in the ground at least 2 ft (0.6 m) wide and 2 ft (0.6 m) deep. Place an empty container in the center of the hole. Fill the hole with vegetation, a receptacle containing undrinkable water, such as saltwater or urine, or fabric soaked in undrinkable water.

2 Cover the hole with a plastic sheet, and secure it in place with stones. Place a stone in the center of the sheet to create a run-off point for the water. The sun's heat will evaporate water from the vegetation, or distill the undrinkable water, producing pure water vapor. The water vapor, which is now free of contaminants, will condense on the underside of the plastic sheet, and drip into the container where it can be collected.

TOOLS AND MATERIALS

- Spade or shovel
- Two containers
- Vegetation and/or undrinkable water
- Piece of fabric or clothing
- Plastic sheet, tarpaulin, or space blanket
- Stones
- Surgical tubing

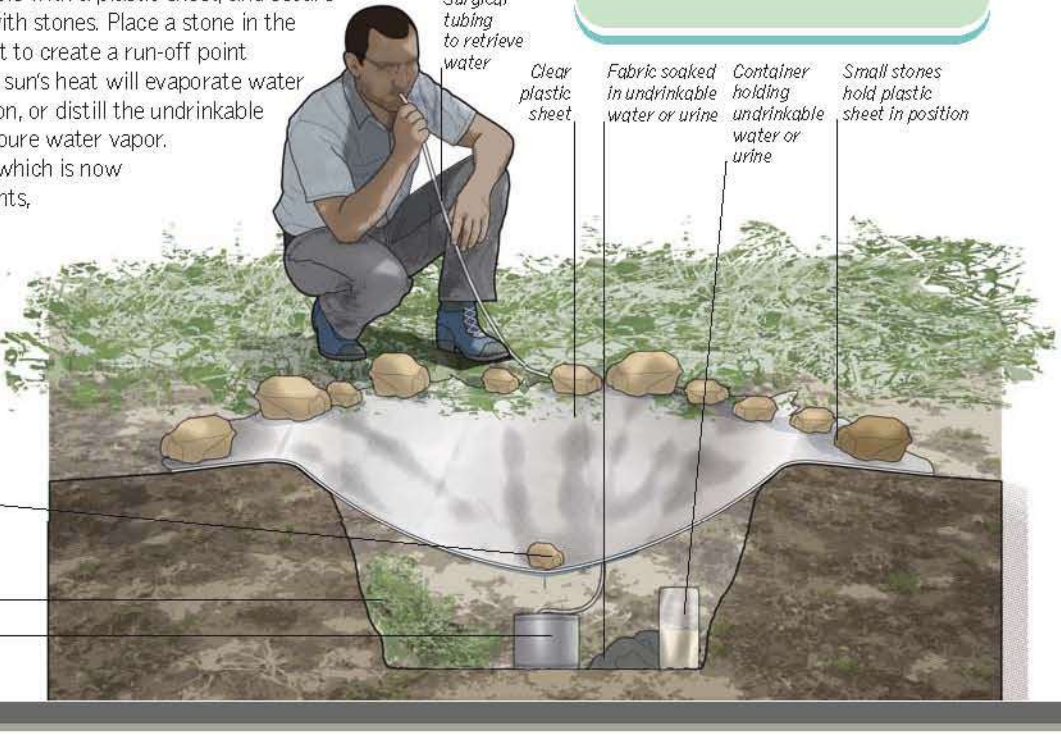
Surgical tubing to retrieve water
Clear plastic sheet
Fabric soaked in undrinkable water or urine
Container holding undrinkable water or urine
Small stones hold plastic sheet in position
Small stone forms a drip point on underside of plastic
Fresh, green vegetation
Container catches condensed water droplets

MAKING A GYPSY WELL

A good method of using the ground to create cleaner water from stagnant water sources, a gypsy well can also be used to collect water from saturated ground. The water collected through this method will still need to be treated before it is safe to drink (see p. 201).

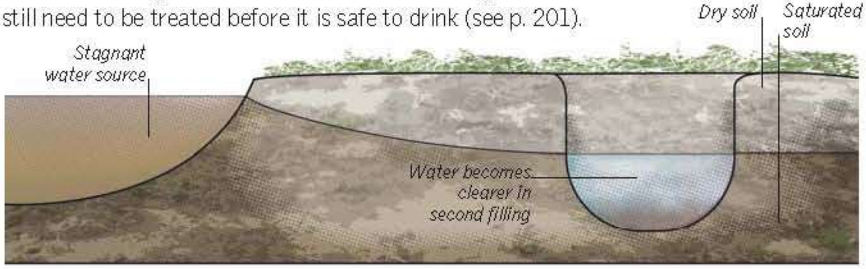
Dry soil
Saturated soil
Stagnant water source
Water becomes clearer in second filling

1 Dig a hole a few feet away from a stagnant water source. It should be wide enough for you to be able to dip into it with a container and at least 12 in (30 cm) below the first layer of saturated soil. The hole will quickly fill with water.

2 Bail out the muddy water and allow the hole to refill—you may have to repeat this process several times. Once the water is clear, it can be collected, treated, and used. Cover the well when not in use to prevent debris or small animals falling in.

SURGICAL TUBING

Always pack a length of surgical tubing in your survival kit. It doesn't take up much space and is an extremely versatile piece of equipment:

- It can be used as an improvised straw to reach water trapped in rocks, hollows, and trees.
- It allows you to access water procured from a solar still without taking the still apart.
- It can be built into transpiration and vegetation bags so that you can remove water without having to open the bag.

FINDING WATER: HOT CLIMATES

YOUR NEED FOR WATER increases in hot climates as your body starts to use more of its water supply to regulate its temperature through perspiration. If you fail to drink more than you perspire, you will start to dehydrate and even approaching the first stages of dehydration can affect your chances of survival. Hot climates can be divided into two groups: hot-humid and hot-dry.

HOT-HUMID CLIMATES

The hot-humid conditions found in jungles and rain forests mean that procuring water is rarely an issue. However, the body's need for water in these conditions should not be underestimated; it is not unusual to have to drink up to 3 ¾ gallons (14 liters) a day to avoid dehydration.

SOURCING WATER FROM VEGETATION

Many plants, such as pitcher plants, have hollow parts that collect rainfall or dew. Some trees store and catch rainwater in natural receptacles, such as cracks or hollows. In an emergency, life-saving liquid can be garnered from a tree's roots or sap. You can find water trapped in the sections of green bamboo by carefully cutting into the bamboo with your machete or knife; or collect small, unripe coconuts and quench your thirst with the fluid they contain; or make a spigot and tap into the water contained inside the water tree. Life-saving fluid is everywhere in the jungle, and you don't have to look far to find it.

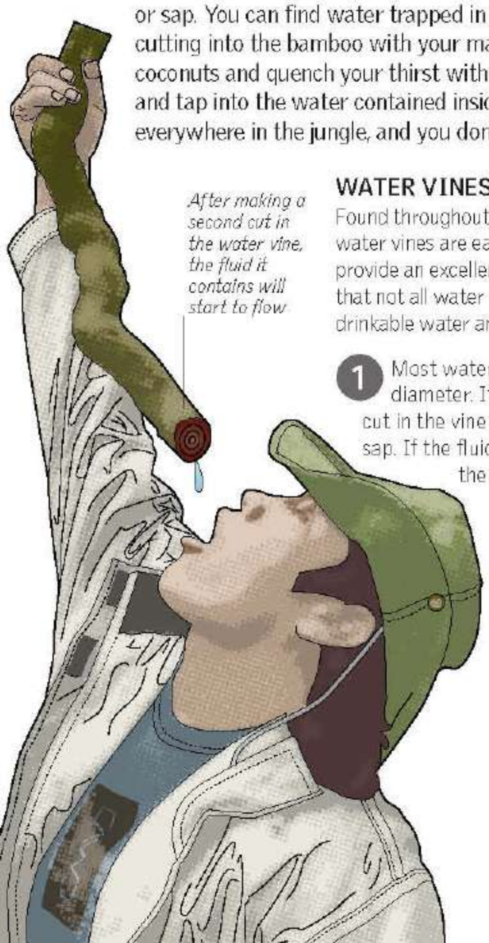

After making a second cut in the water vine, the fluid it contains will start to flow

WATER VINES

Found throughout the jungles and rain forests of tropical regions, water vines are easily identifiable by their size and shape and can provide an excellent sort of fresh water. However, bear in mind that not all water vines are water-bearing; not all contain drinkable water and some even contain poisonous sap.

1. Most water-bearing vines are about 2 in (5 cm) in diameter. If you think you have found one, make a small cut in the vine with a machete and check the color of the sap. If the fluid is milky, don't drink it; if it's clear, then the water in the vine will be safe to drink, so cut through the vine as high as you can with a knife or machete.

2. Cut off the vine at a point lower than the first cut. The liquid, which has a neutral, fruity taste, will start to flow. Don't let the vine touch your mouth as the bark may contain irritants.

3. The pores in the upper end of the vine may re-close, stopping the flow of water. To rectify this, simply cut the top of the vine again with your machete.

WARNING!
Water sources are plentiful in jungle and rain forest environments and, under normal circumstances, you should have no problem getting hold of enough water to satisfy your needs. However, rivers and streams may not be available during certain seasons, or if you have climbed too high, so knowledge of other water procurement methods could be crucial to your survival.

CATCHING RAINWATER

Catching rainwater is the best way to procure water: it is passive and requires no energy to collect once you place your containers. There are many forms of catchment device, but make sure you filter and purify the water (see pp. 200–01) before drinking it.

BAMBOO ROOF

Construct a sloped bamboo roof with a bamboo gutter. This could be the roof of your shelter (see p. 169), but if water procurement is a problem, you will have to construct additional bamboo roofs.

WIDE-LEAF ROOF

If you're in an area where there are wide-leafed plants, it's easy to construct a roof from them. Overlay the leaves as you would with roof tiles, working from the bottom to the top (see p. 170). Placed in this way, they will allow the water to run to the bottom. A length of bamboo cut in half lengthways can be placed as a collection gutter at the bottom.

BAMBOO DRAINPIPE

Observe rainwater's route down a tree trunk and tie a length of bamboo that has been cut in half lengthwise in its path. Place the other end of the bamboo into a suitable container.

DRIP RAGS

Wrapping any absorbent material around a leaning tree, such as a rag or a T-shirt, will result in the water running down the tree and soaking the material. Shape the rag to form a low point from which water will drip and place a suitable container underneath to catch the water.

HOT-DRY CLIMATES

Anyone venturing into this environment should have sufficient water for their needs, plus an emergency supply just in case, otherwise they have no right to enter the desert. Green vegetation usually signifies water or moisture in some form, and many techniques of procuring water in temperate climates (see pp. 188–91) may work in some desert conditions.

SOURCES OF EMERGENCY FLUID

If there is no surface water to be found, and you have no other means of procuring water, a water-yielding plant may be your only option. In some plants, the clear sap, fruit, or trapped rainwater, may quench your immediate thirst, but do not rely on these sources to keep you alive for long.

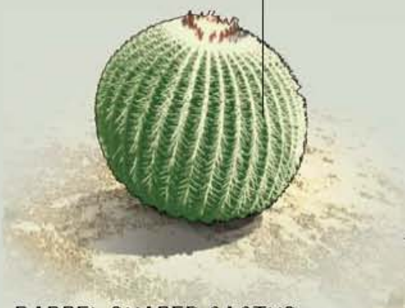
The barrel-shaped cactus is characterized by numerous ribs and hundreds of sharp spines

BARREL-SHAPED CACTUS

Barrel-shaped cacti contain a milky fluid that's safe to drink. Carefully remove the top of the cactus with a machete, mash up the flesh inside with a stick to make a pulp, and extract the juice from the pulp by sucking through a hollow grass stem. Alternatively, use a cloth to soak up as much of the fluid as possible and then wring the material to extract the fluid. The rewards from both techniques are minimal.

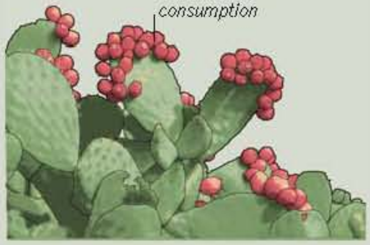
The prickly outer layer of the fruit has to be removed before consumption

PRICKLY PEARS

Found in low-spreading clumps measuring 3 ft (0.9 m) in height, and native to dry, sandy soils throughout the world, the prickly pear cactus has an edible fruit that can provide a life-saving amount of fluid.

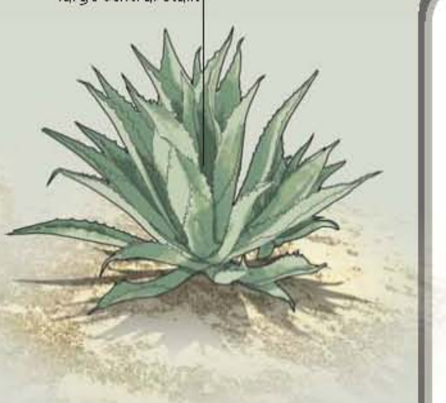
Agaves have large clusters of thick, fleshy leaves surrounding a large central stalk

AGAVES

Native to Mexico and the southern and western United States, agaves have a rosette of thick, fleshy leaves containing fluid that is safe to drink. Cut the huge flower stalk with a knife or machete and collect the juice.

FINDING UNDERGROUND WATER

Water is life in the desert; wherever there is life there'll be water. Unusual clusters of green plants may indicate a minor presence of water; an abundance of greenery may indicate a more substantial water source. Water is rarely found above a depth of 6 ft (1.8 m) below the surface, so you will have to dig for it. If you do, make sure you only do so during the coolest part of the day. Bear in mind that, despite what you may have seen in movies, or read in other books, your chances of procuring water from the inside bend of a dry wadi in a real survival situation are slim to non-existent. In fact, you're more likely to waste what water you do have looking for the dry wadi and die in the process!

THINK LATERALLY

The chances are that if you're having to procure water in the desert you'll already be in a desperate situation. Bear in mind that, in addition to the methods mentioned above, if you have knowledge of the techniques outlined in the temperate environment section (see pp.190–91)—such as solar stills, dew traps, and vegetation and transpiration bags—it could be enough to keep you alive in the desert.

CHECKLIST FOR HOT-DRY CLIMATES

Even the best-laid plans can hit unforeseen problems, but many problems encountered in a desert environment can be avoided with some prior preparation.

- Always start hydrated.
- Carry enough water for your needs plus emergency water; your emergency supply should be enough to get you out of danger.
- Monitor your progress against the water you use. If you're using more than you thought, re-evaluate what you want to accomplish. It's better to turn back and learn from the mistake than to push on and create a survival situation that need not exist.
- Cache water ahead if necessary.
- Check your map for probable water sources. Confirm the reliability of these sources with locals and ask if there are any sources, such as wells, that are not shown on the map.
- Mark your map, or waypoint your GPS, with any water sources you sight as you progress. It's better to go back to a known source than to move on with nothing more than hope.
- To keep the water cool, always keep the bottles in the shade or in a windy location.

FINDING WATER: COLD CLIMATES

ADOPT THE SAME METHODS of trying to find drinking water in cold climates as you would in temperate ones (see pp. 186–91), although the ability to procure water becomes a major problem in freezing temperatures. The cold-weather survivor faces a dilemma: he or she is surrounded by water, but could die of thirst because most of that water is frozen.

WARNING!
Never try to melt ice or snow in your mouth, as it can cause freezing injuries to your mouth and lips. What's more, your body will expend heat as it melts the ice, and this could cause you to cross that very fine survival line between being cold or being hypothermic.

CHECKLIST FOR COLD CLIMATES

You should prioritize water usage in cold-weather conditions in the same manner as you would in desert conditions (see p. 193).

- One of your first thoughts should be about procuring water. You will have to find a water source that's close to everything you need to build and maintain a fire.
- Always look for an alternative water source before trying to melt snow or ice. It's easier, less time-consuming, and more fuel-efficient to fill your water containers with natural meltwater than it is to melt snow or ice by the heat of a fire.
- Your ability to procure water in freezing conditions will be directly related to your ability to start and maintain a fire.
- Be aware that it takes time and patience to melt snow and ice. You'll also need enough fuel to maintain a fire for a considerable period of time.
- Regulate your body heat to minimize overheating and sweating.
- Keep drinking water close to your body to prevent it from freezing, but avoid having water containers directly next to your skin. Instead, keep them between layers of clothing and use the warmer air trapped between the layers to help raise the water's temperature.
- Do not use recently frozen sea water, as it contains high levels of salt.

MELTING ICE AND SNOW

If you have the choice between melting ice and snow, favor ice, as it melts more quickly than snow and is up to 17 times denser. However, if you can't find ice, use dense, compact snow. Always look for the whitest, purest-looking ice or snow.

MELTING ICE

If you have some water to begin with, pour some into a container and heat it over a fire. Break the ice into small pieces—rather than adding it in one lump—and keep adding pieces to the container to melt them. Keep the water hot, but not boiling, so that you avoid losing water through evaporation.

USING A HOT PLATFORM

If you don't already have water for the method above, you can melt ice slowly on a gently sloping platform fashioned from any flat piece of stone or wood positioned above a fire.

1. Build a fire. Search for a large stone with a flat surface and two logs, or smaller stones, that are strong enough to support it. Place the small logs or stones on either side of the fire as a support for the flat-surfaced stone. Make sure that the platform is at an angle. This will create a natural run-off for the ice as it melts. Place a block of ice in the center of the platform.

2. As the fire starts to heat up the platform, the ice will begin to melt. The meltwater will flow off the platform, where it can be collected in a container, such as a mess tin.

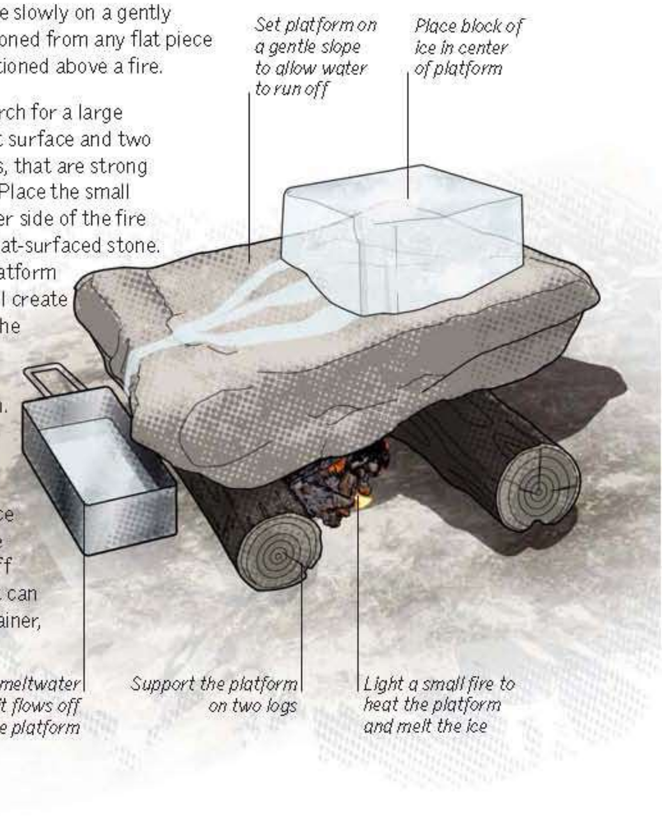

Set platform on a gentle slope to allow water to run off

Place block of ice in center of platform

Collect the meltwater in a pan as it flows off the platform

Support the platform on two logs

Light a small fire to heat the platform and melt the ice

MELTING SNOW

If you already have water, follow the techniques for melting ice (see opposite), heating a little water and adding the snow little by little. Don't pack the snow too tightly; if an air pocket forms, the heat from the fire will be absorbed by the metal container rather than by the snow and could result in the fire burning a hole through the metal container before it melts the snow.

STORING WATER

Snow is a great insulator: even if the temperature dips to -40°F (-40°C), water in a bottle will remain largely unfrozen if placed under at least 1 ft (0.3 m) of snow. Make sure you store the bottles upside down. That way, if some of the water does freeze, it will freeze at the bottom of the bottle and not at the top.

MAKING A FINNISH MARSHMALLOW

Cut a solid piece of dense snow—often referred to as a "marshmallow" or a "snowman's head"—and skewer it with a stick. Secure the stick in the ground close enough to a fire that it receives heat, and position a suitable container underneath it to capture the water as it melts.

USING A MELTING SACK

Using a similar principle to the Finnish marshmallow (see left), this technique involves using an improvised sack—made from any porous material, such as a T-shirt or a sock—suspended near to a fire. The heat from the fire will start to melt the snow, which can then be collected in a well-positioned container.

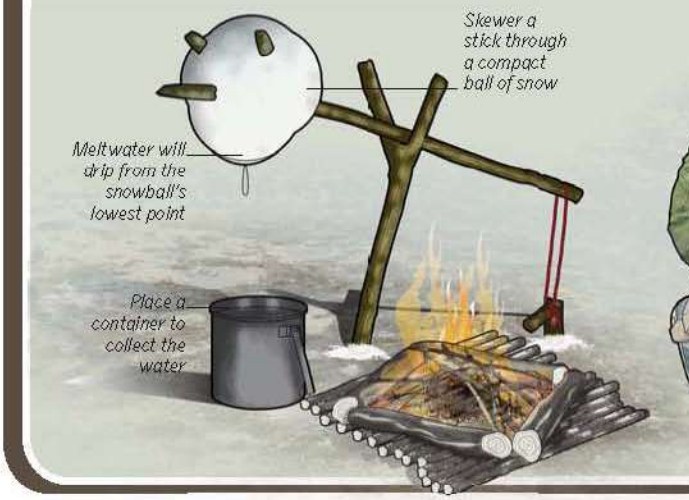
Skewer a stick through a compact ball of snow
Meltwater will drip from the snowball's lowest point
Place a container to collect the water

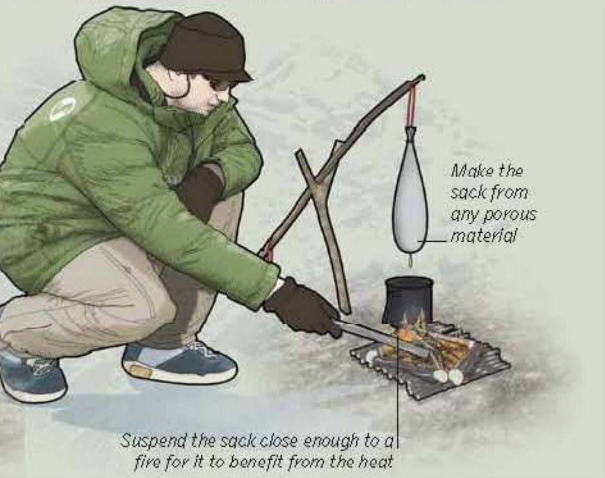
Make the sack from any porous material
Suspend the sack close enough to a fire for it to benefit from the heat

IN A WORST-CASE SCENARIO

If you're caught out on barren snow or ice, you're unlikely to have access to natural fuel to burn in order to melt snow or ice into water. As long as you have your second-line survival equipment (see p. 42), you'll have all you need to procure life-saving water.

USING YOUR FIRE TIN

Set up your fire tin on the ground and shelter it from any wind using either your body or your pack. Using your survival tin as a container, place small quantities of ice or snow into the tin. Light your fire tin and place your survival tin over the flame. Add more ice or snow as it starts to melt.

USING YOUR MATCHLESS FIRESET

Place the lid of the matchless fireset tin on the ground. Light a hexi fuel block using the cotton wool and flint/striker. Using the fireset container or your survival tin, melt small quantities of ice or snow over the flame. Once the ice or snow has melted, let the container cool so you don't burn yourself. The water is then safe to drink.

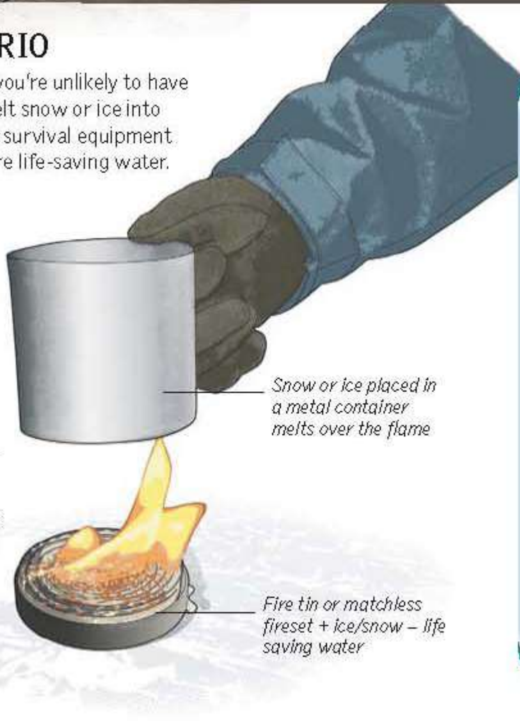
Snow or ice placed in a metal container melts over the flame
Fire tin or matchless fireset + ice/snow – life saving water

USING BODY HEAT

When Inuit hunters capture a caribou, they empty out the contents of its stomach, turn the stomach inside out, fill it with snow, and tie it shut with a length of intestine. They then put the stomach back in the cavity while they skin the caribou. By the time they have finished, the heat given off by the animal's body has melted the snow. The Inuit then open the bag carefully and suck out the water through a block of snow to filter it.

FINDING WATER: AT SEA

OF ALL ENVIRONMENTS, the sea is possibly the most difficult in a survival situation. It offers no natural resources for protection against the extremes of temperatures, wind, rain, and sea state, and provides little to aid location. Being surrounded by water that you can't drink only adds to the difficulty. Some devices are capable of making seawater safe do drink, but if you don't have one, you'll have to find a way of procuring enough fresh water to keep you alive.

WARNING!

Never drink salt water. Its salt concentration is three times higher than that of blood and ingesting it will dehydrate you. Continued use over a prolonged period will lead to kidney failure and, ultimately, death.

CONSERVING YOUR WATER RATIONS AT SEA

Rationing fresh water supplies when you're adrift at sea is a sensible precaution, as you have no idea how long it will be before you're rescued or before you reach land. Here are a few tips to help you conserve what water you have while you're at sea:

- Fix your daily water ration after taking stock of the amount of water you have, the output of solar stills and desalting kits, and the number and physical condition of your party.
- Prevent fresh water supplies from becoming contaminated by saltwater.
- Keep water supplies well shaded, both from the overhead sun and from the glare off the sea's surface.
- In hot conditions, dampening your clothes with saltwater can help to lower your body temperature—but don't overdo it. This is a trade-off between cooling yourself down and the saltwater boils and rashes that will result from continued exposure.
- Don't exert yourself. Relax and sleep whenever possible.
- Use every container you have—even a simple trash bag—to collect rainwater, and keep them well sealed and attached to the raft.
- If you don't have water, don't eat. Protein consumption will hasten the onset of dehydration.

COLLECTING FRESH WATER

If you find yourself adrift with no hope of immediate rescue, obtaining drinking water will be a major priority. If you don't have a solar still or a reverse-osmosis pump, you must find another means of procuring fresh water. Fortunately, there are several ways of doing so.

GATHERING RAINWATER

Most modern liferafts incorporate a built-in rainwater catchment system that channels rainwater and dew from the outer surface of the liferaft into collection pockets inside the liferaft. However, even if you're not in a liferaft, you can construct a similar system using a tarpaulin or any other waterproof material. Watch the clouds, be ready for the possibility of any showers (see pp. 80–81) and spread your tarpaulin in a bowl shape to catch the largest amount of rainwater possible. Always place a tarpaulin before nightfall, so that you don't miss out on any overnight rainfall.

HARVESTING DEW

At night, secure the tarpaulin like a sunshade and turn up its edges to capture dew. It's also possible to harvest any dew that may have collected on the sides of the raft using a sponge or cloth that you then wring out.

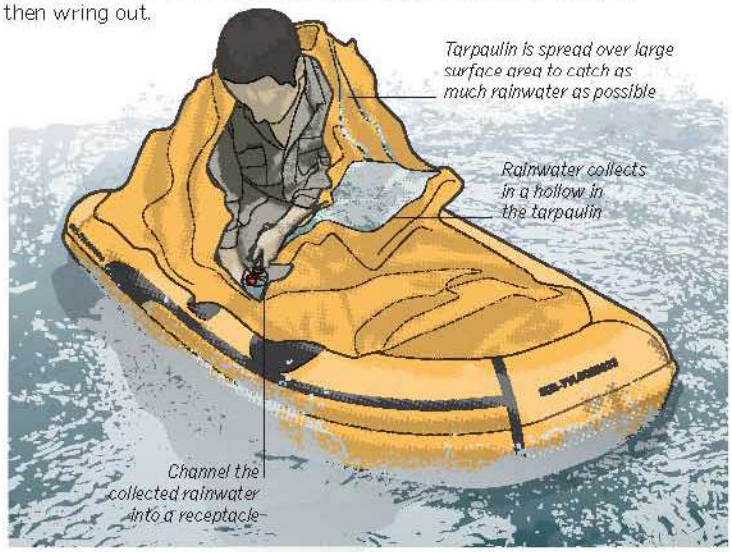

Tarpaulin is spread over large surface area to catch as much rainwater as possible

Rainwater collects in a hollow in the tarpaulin

Channel the collected rainwater into a receptacle

TREATING SALTWATER

If you have no means of collecting rainwater or dew, there are several products capable of turning undrinkable saltwater into fresh water. Although these products are standard issue on most liferafts, you should always try to have at least one of them with you if you are venturing into a marine environment.

SOLAR STILL

Solar stills are a simple way of distilling water using the power of the sun. Saltwater is placed at the bottom of the container, where it is evaporated by the sun through clear plastic. Pure water condenses on the top of the plastic and drips down to the side, where it can be collected, often via a tube. Most solar stills on modern liferafts are inflatable.

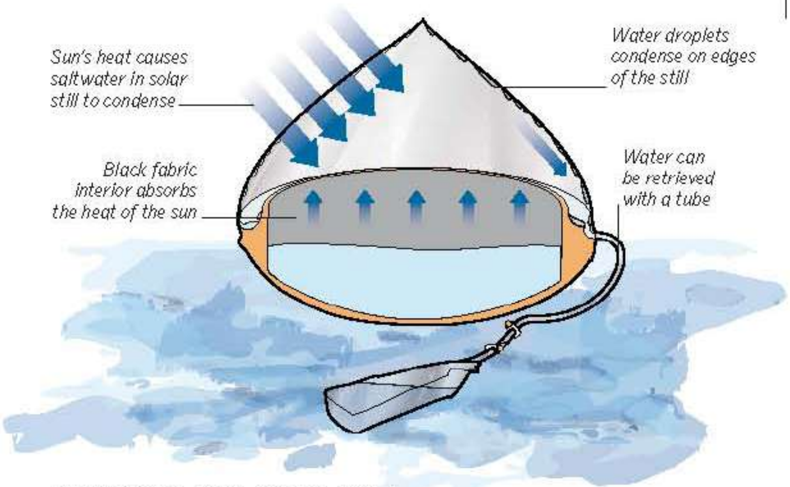
Sun's heat causes saltwater in solar still to condense
Water droplets condense on edges of the still
Black fabric interior absorbs the heat of the sun
Water can be retrieved with a tube

MAKE YOUR OWN SOLAR STILL

If you have a chance to gather together the right materials, it's easy to make a small solar still. All you need are two containers (one larger than the other), a sheet of plastic, some string, and a weight to form a natural run-off point for the water as it condenses under the heat of the sun. If you have some surgical tubing, you can use it to retrieve the fresh water without taking the still apart.

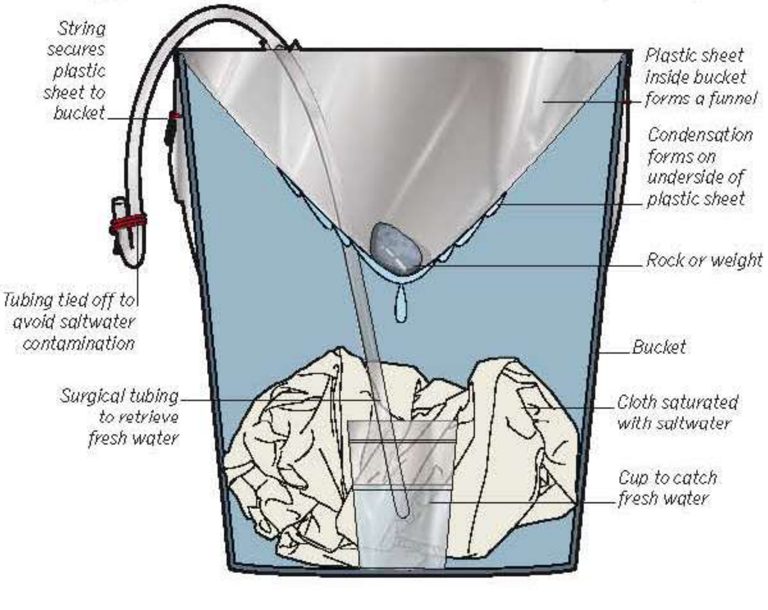
String secures plastic sheet to bucket
Plastic sheet inside bucket forms a funnel
Condensation forms on underside of plastic sheet
Rock or weight
Tubing tied off to avoid saltwater contamination
Bucket
Surgical tubing to retrieve fresh water
Cloth saturated with saltwater
Cup to catch fresh water

REVERSE-OSMOSIS PUMP

These hand-powered devices pump seawater at a very high pressure through a membrane that filters out the salt. Depending on the model, they can produce around 23 liters (49 pints) of fresh water per day.

DESALTING KIT

These kits turn seawater into freshwater through a process called "ion exchange." Because they only produce small amounts of fresh water over several hours, use desalting kits only during long periods of overcast weather when you can't use a solar still.

THE LAST RESORT

In absolute emergencies, potentially life-saving forms of liquid can be obtained from the ocean.

SEA ICE

In Arctic seas, you can obtain drinking water from old sea ice. This ice is bluish in color, has rounded corners, splinters easily and, more importantly, is nearly free of salt. New ice is gray, milky, hard, and salty. Water from icebergs is fresh, but icebergs are dangerous to approach, so you should only use them in an emergency.

FISH

Drink the aqueous fluid found along the spine and in the eyes of large fish. Carefully cut the fish in half to get to the fluid along the spine and suck the eye. If you are so short of water that you need to do this, do not drink any of the other fluids: they are rich in protein and fat and your body will use up more of its water digesting them than it obtains from them.

SEA TURTLE

Sea turtle blood has a salt concentration similar to that of humans. The blood can be collected by slitting the turtle's throat. Note that, although this may help prolong survival, sea turtles are an endangered species, so you should only kill one as a last resort.

REHYDRATION ENEMA

If you have some water that is not salty or poisonous, but is too foul to take orally, you can absorb as much as a pint a day—enough to keep you alive—through the large intestine using an improvised tubing device.

CARRYING AND STORING WATER

YOU WILL NEED TO make many decisions in any survival situation, some of which could mean the difference between life and death. One of these decisions will be whether to stay and wait for rescue or to attempt self-rescue. A major factor in this decision will be the availability of water and your ability to carry and store it.

WARNING!
Mark your containers clearly. Late at night in camp it's easy to confuse bottles that contain treated water with ones that contain untreated water or camping fuel.

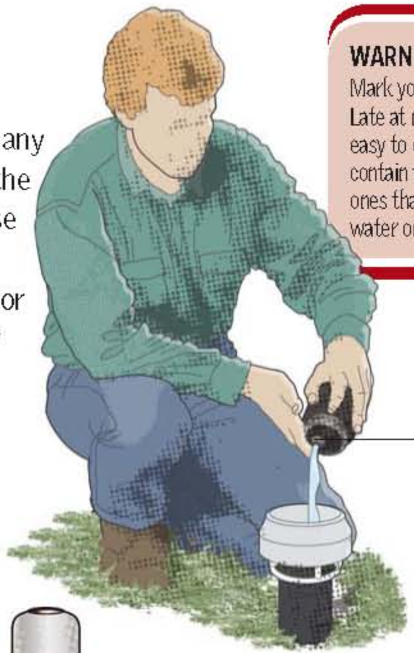
If your container has a lid, it can be used to scoop water from a stream before you pour it into a larger water bottle

WATER CONTAINERS

There are many different kinds of container for carrying both hot and cold liquids. They range from solid plastic or steel flasks or bottles, to collapsible waterproof bags that can be folded up when they're empty.

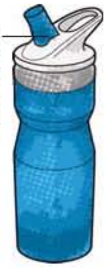
Fitted with a push-pull nozzle for ease of use while on the move

PLASTIC WATER BOTTLE

Strong and lightweight, with screw-off tops or push-pull nozzles.

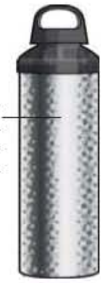
Metal container can be used to boil water

METAL WATER BOTTLE

Slightly heavier, but stronger, than plastic water bottles.

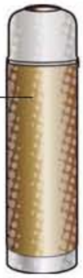
A metal exterior and a glass interior makes a thermos a heavy alternative

THERMOS FLASK

Although very heavy, thermos flasks allow you to store either hot or cold water.

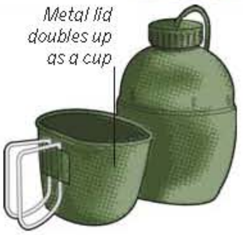
Metal lid doubles up as a cup

MILITARY WATER BOTTLE

A standard issue bottle used by most of the world's military forces.

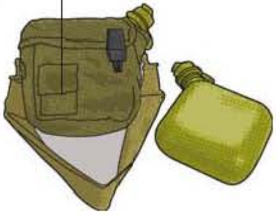
Store purification tablets in pouch

COLLAPSIBLE WATER CANTEEN

Heavy-duty plastic water canteen that can be hung round the neck.

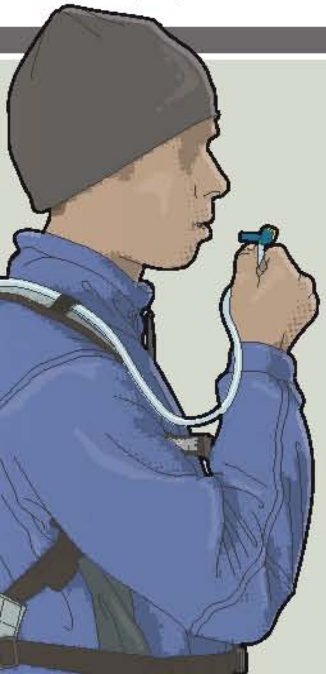

HYDRATION SYSTEMS

Hydration systems usually consist of a water-storage container (called a "reservoir"), an on-demand, one-way drinking tube, and a harness. However, a common problem with these systems is that, because you can't see the water, you can suddenly discover you have used all of it without realizing.

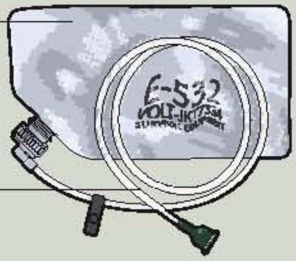
Water container is filled and then carried in a backpack

Water is accessed via a one-way drinking tube

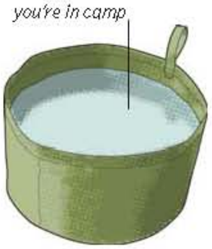
Use to store water once you're in camp

COLLAPSIBLE WATER BOWL

Easy to pack, light, and can also be used to seperate items in your backpack.

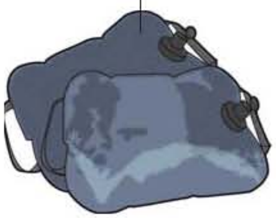
Stow water bags flat in your pack when not in use

WATER BAGS

Useful for carrying water from its source back to your camp. Can also be used to keep your equipment dry.

IMPROVISING WATER CONTAINERS

In an emergency, there's a strong chance that you won't have the luxury of having water bottles or storage systems at your disposal. However, with luck—and a little ingenuity—you should be able to find something that allows you to store and carry enough water to satisfy your needs.

EMERGENCY WATER

Sachets of emergency water can be bought and stored for emergency use. They usually come in a packet that contains five $1\frac{7}{10}$fl oz (50 ml) sachets of water, each being little more than a mouthful. These should always only be used as a last resort. Always try to procure water by other means before turning to your emergency water supply.

SEA SURVIVAL SUIT

Because they are designed to keep water out, they can also be used to keep water in. Most survival suits are packed with a layer of chalk to prevent the material from sticking together, so wash it out before use.

WATERPROOF CLOTHING

Many types of waterproof clothing can be adapted to hold water. Jacket sleeves and trouser legs can be knotted to form a basic bladder; Gore-Tex® socks will hold water; and some waterproof backpacks can be used for water storage.

DISCARDED MATERIALS

Always be on the lookout for anything that can hold water, from empty plastic bags and bottles, to large industrial-type containers. Clean and sterilize found items before use.

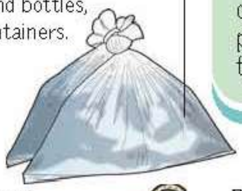

Heavy-duty plastic bags can be used to carry water

GOURDS

The shell of a hollowed-out and dried fruit—a gourd—can hold water. Gourds made from large fruits, such as squashes, pumpkins, and melons, can hold a considerable amount of water.

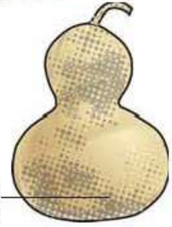

Look for a gourd with a shell thickness of ¼in (60 mm)

BAMBOO

A bamboo cane will provide a natural cup if you cut 1 in (2½cm) below one joint and then 1 in (2½cm) below the next. Take care to smooth the edges after cutting.

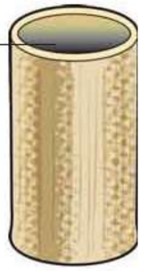

Bamboo can be cut into a ready-made cup

MAKING A MINI-RESERVOIR

Your ability to store sufficient water for your needs in a survival situation may be limited and it may be impractical —because of injury or distance—to travel back and forth for water every time you need it. If you have some basic materials, making a mini-reservoir to catch rainwater (see also p. 188) solves this problem.

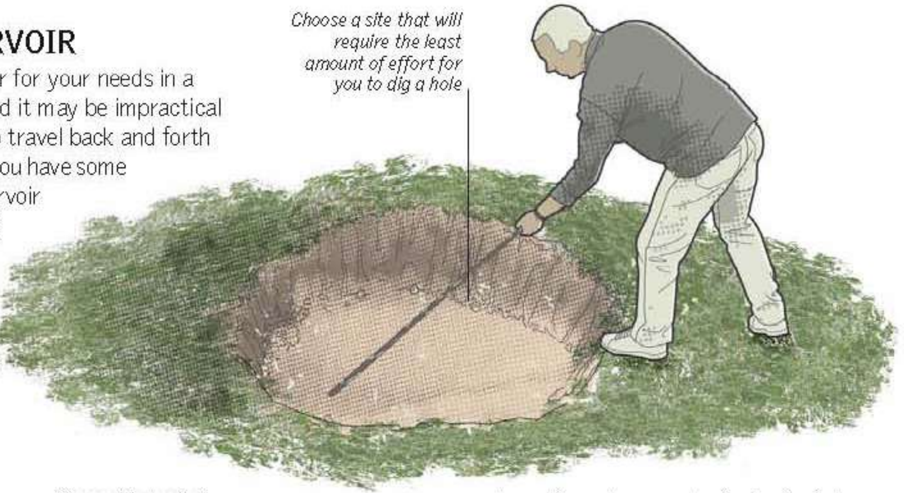

Choose a site that will require the least amount of effort for you to dig a hole

1. To reduce the amount of effort required, choose a site that offers the least resistance (such as a natural hollow). Dig a shallow pit using whatever materials you have to hand, such as a stick.

2. Smooth the edges of the pit and line it with waterproof material, such as a survival blanket. Weight the edges with rocks, soil, or logs, making sure that dirt can't run into the pond when it rains.

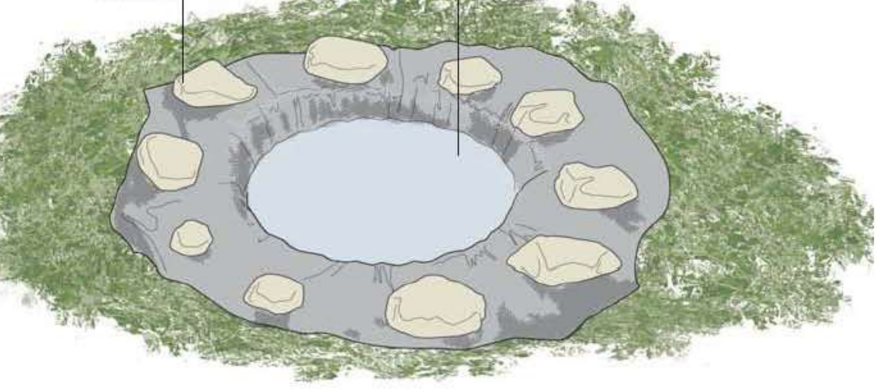

Secure the survival blanket in place using rocks

Leave the pond uncovered when it rains, but cover it at other times to reduce evaporation

WATER HYGIENE

Water stored in a mini-reservoir for any period of time will need to be filtered and purified before it's safe to drink (see pp. 200–01).

TREATING WATER

WITH THE EXCEPTION OF rainwater, all other water procured in a survival situation should be treated before it is safe to drink to remove or destroy harmful pathogens and microorganisms that could lead to gastrointestinal illness.

FILTERING WATER

If you don't have a device that filters and disinfects the water, you will have to accomplish the same task in two stages. Before purifying the water, you will have to filter it to remove any debris. You could use the popular Millbank bag (see opposite), but it may be necessary for you to construct your own improvised filter.

MILLBANK BAG
Used extensively by military forces around the world, the Millbank bag is an effective water filter that packs down small and can be used many times to produce large quantities of filtered water. Set up the bag as soon as you arrive in camp, because the process is quite slow: the Millbank bag will filter 1 liter (2.1 pints) of water in five minutes. Note that water filtered in this way will still need to be purified.

BUILDING A TRIPOD FILTER

If you're not carrying a filtration system with you such as a Millbank bag, you could make an improvised filter. All you need are three sticks to form a tripod, and some materials to create three separate layers.

1 Use three sticks or a bent sapling to form a tripod. Using any materials you have, form layers, starting at the top with the coarsest material and working your way down using finer materials—such as parachute silk or nylon—as you go.

2 Pour water into the top layer. It will become filtered as it passes through the increasingly fine materials.

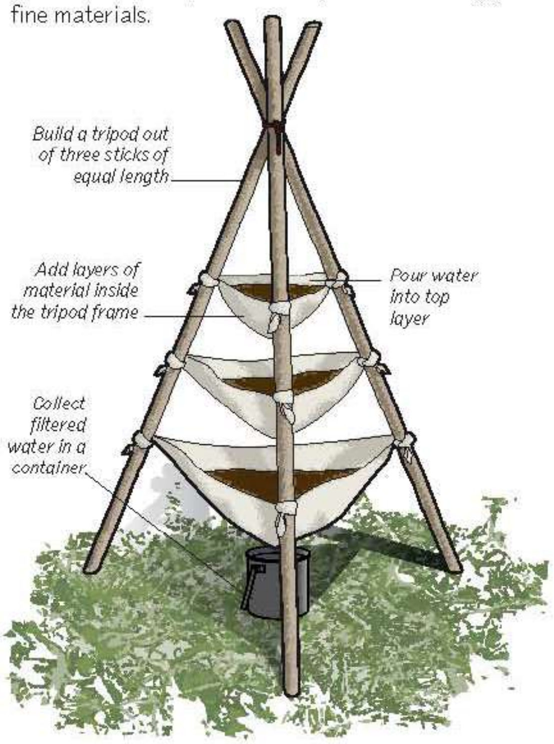
Build a tripod out of three sticks of equal length
Add layers of material inside the tripod frame
Pour water into top layer
Collect filtered water in a container

MAKING A BOTTLE FILTER

To make an improvised bottle filter, take a container, such as a plastic bottle, and cut off the bottom (or make a hole). Note that a sock used in the same manner is still effective.

1 Hang the bottle upside down from a branch. Fill the bottle with layers of different materials working from coarse to fine as you work your way down the container.

2 Pour the water into the top end of the bottle and allow it to work its way down through the layers.

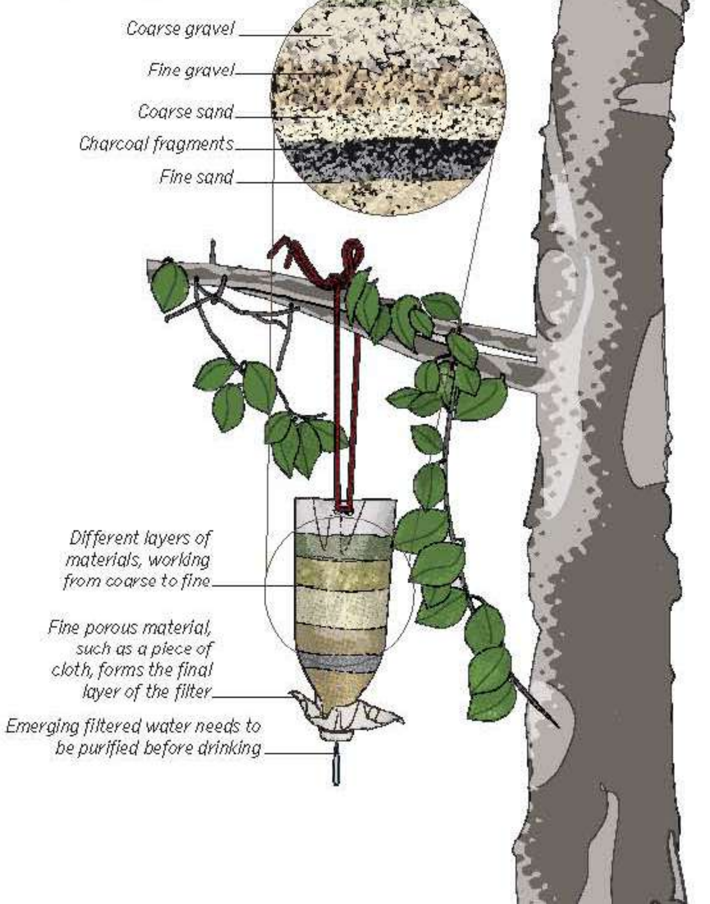
Coarse gravel
Fine gravel
Coarse sand
Charcoal fragments
Fine sand
Different layers of materials, working from coarse to fine
Fine porous material, such as a piece of cloth, forms the final layer of the filter
Emerging filtered water needs to be purified before drinking

DISINFECTING WATER

If you drink untreated water, you run the risk of becoming infected with a water-borne disease (see p. 187), so it's vital that you treat any water first. If you have the ability to start a fire, the most effective way of making water safe to drink is by boiling it. However, if you are unable to make a fire, there are several devices available that are capable of filtering and purifying water to make it safe for drinking.

MINI PORTABLE WATER PURIFIERS

These are specially designed units that filter the water and then purify it, by pumping the contaminated water through either microfilters, chemicals, or a combination of both. Sizes vary from small emergency pumps capable of purifying up to 13 gallons (50 liters) of water, to larger units that can filter huge quantities of water.

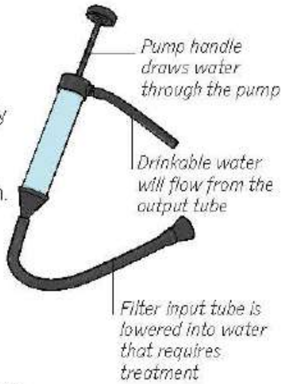

Pump handle draws water through the pump

Drinkable water will flow from the output tube

Filter input tube is lowered into water that requires treatment

OTHER METHODS

If you can't boil water, or if you don't have a water-purification device, you'll have to rely on non-mechanical techniques. The concentration and contact time required for some of these methods is dictated by the quality and temperature of the water being treated.

GRAVITY/PRESSURE FILTERS

These devices are incorporated within drinking bottles. The water is either allowed to flow naturally through the system via gravity, or is squeezed through by the operator. All bottles of this type usually employ: a filter to remove sediment and organic contaminants; a micron filter to remove protozoa; and a chemical that kills water-borne bacteria and viruses.

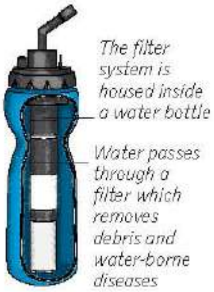

The filter system is housed inside a water bottle

Water passes through a filter which removes debris and water-borne diseases

SURVIVAL STRAWS

Compact emergency water purifiers, survival straws contain a filter system and employ either carbon or iodine resin systems to eliminate water-borne diseases and harmful chemicals. You need to get the water to a point where you can reach it with your straw. If you want to draw a supply of water to carry with you, you'll have to draw the water into your mouth and decant it into a container—a laborious process.

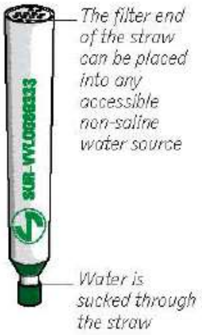

The filter end of the straw can be placed into any accessible non-saline water source

Water is sucked through the straw

"BIG BUBBLES, NO TROUBLES"

Microorganisms and virtually all intestinal pathogens are killed at temperatures well below boiling point. The process of bringing water to the boil is sufficient to disinfect it—continuing to boil it just wastes fuel, time, and water.

METHOD	DESCRIPTION
Iodine (Liquid and Tablets)	Iodine—which destroys bacteria, viruses, and cysts—can be used to disinfect water effectively and conveniently. Its action is dependent on its concentration, the water temperature, and duration of contact—a concentration of 8 mgs per liter at 68°F (20°C) will destroy all pathogens if left for ten minutes.
Chlorine tablets	Chlorine-based tablets will destroy most bacteria, but are less effective for viruses and cysts. They are more effective when used in combination with phosphoric acid and will destroy both Giardiasis and Cryptosporidium.
Potassium Permanganate	Potassium permanganate can be bought at most pharmacies. Mix a few granules with your water until it turns light pink. Leave for at least 30 minutes before drinking.
Bleach	Adding unscented household bleach is the cheapest way of adding chlorine to water (it contains 5 percent sodium hypochlorite). Be careful to add just one drop of bleach per liter of water—two if the water is cloudy—and leave for at least 30 minutes before drinking. This method is not always effective against Giardiasis and Cryptosporidium.
Ultraviolet (UV) Light	When many harmful microorganisms are exposed to UV light, the process of light absorption disrupts the cell's DNA, rendering the organism harmless. The quality of the water will affect the amount of exposure to UV light required: the cloudier the water the harder it will be for UV to penetrate.
UV Passive	Fill plastic bottles with water, replace the lids, and place them in direct sunlight, preferably on a dark surface. The sun's rays will kill the bacteria that cause common water-borne diseases.
UV Active (Steripen)	A small UV purifier that is placed in pre-filtered water and activated for a short time. Some models can purify up to 1 liter (2.1 pints) of water in as little as 48 seconds.

FIND AND PREPARE FOOD

Your body converts food into fuel, which provides you with heat and energy, and helps you to recover from hard work, injury, or sickness. If you are healthy, your body can survive for weeks without food by using the reserves stored in its tissues—although you will use approximately 70 calories per hour just breathing and up to 5,500 calories a day if laboring hard.

In a short-term survival situation, food should not be your major priority. You would probably have eaten recently and, if you'd prepared properly, you should have some basic emergency food in your pack. While you may go through food withdrawal symptoms—when your stomach complains because the food it's expecting doesn't arrive—you're not going to die of starvation within a few days. However, the body will react to its fuel not being replaced: hunger, a lack of energy, and a deterioration in coordination can be expected after a few days. If the opportunity to procure food arises, it should always be taken. Eat little and often but always make sure you have sufficient water to digest it.

In this section YOU WILL DISCOVER...

- how to **cook** with **hot rocks**...
- how to **make** a **fishing reel** from a **beverage can**...
- why a **snare** must have a **perfect end**...
- how to **lasso a lizard** and **snare a squirrel**...
- that there's **more than one way** to **skin a rabbit**...
- which **grubs taste** like **scrambled eggs**...
- how to **catch** a **bird** in a **bush**...

In a long-term survival situation, your survival priorities will change and the need for food in order to simply survive will become more important. There is a thin line between food not being your priority and then subsequently finding that you're in no physical condition to do anything about it when it does become a priority—you should regularly reassess your situation and alter your plans accordingly. It takes effort, skill, and a certain amount of luck to obtain food in the wild, especially if you're not in your natural environment.

When gathering food in the wild, always ensure that the energy gained from the food is more than the energy you expended in procuring it, otherwise it's a wasteful exercise.

FOOD THAT'S EASY TO FIND and gather should always be your first priority:

- **Plants** are easy to collect and, as long as they're readily available in the environment you're in, should be your first choice for food. However, make sure that you are absolutely sure that they are edible—mistakenly eating the wrong leaf or berry could cause vomiting and diarrhea, making your situation worse.
- **Fishing** requires little effort once the lines or traps have been set, and they will work for you around the clock. Fish is high in protein and relatively simple to prepare and cook.
- **Insects, reptiles, and amphibians** may also be available, but be careful that you don't expend more energy in catching them than you gain in nutrients from eating them. Remember also that many insects, snakes, and amphibians are poisonous, but they can still be used as bait for fish and mammals.
- **Birds and mammals** have their own survival mechanisms and are wary of humans, especially in remote areas where contact with humans is limited. Even if caught, the bird or mammal will need to be killed, plucked or skinned, and cooked.

> “Always ensure that the **energy gained** from the **food** is **more than** the **energy you expended** in **procuring it**, otherwise it's a **wasteful exercise**”

WILD COOKING

IN A SURVIVAL SITUATION, you must cook every item of food that you're not sure about in order to kill any parasites or harmful bacteria. While cooking reduces the nutritional value of food, it does help to improve the taste of many wild foods and to make them more digestible than if eaten raw.

FOUND MEAT

If you find a dead animal, you can eat the meat if you cut it into small pieces and boil it for 30 minutes. Don't touch it if you have cuts on your hands—you don't know how it died.

COOKING OVER AN OPEN FIRE

Use the fierce flames of a fire to boil water then, when the flames have died down, use the steady heat of the embers to cook on (see p. 121). The most basic method of cooking is roasting, using a spit made from green wood to suspend the food over a fire. Make sure the meat is thoroughly cooked.

BOILING OR STEWING

When you boil food, fat and natural juices are retained in the water. It is important to drink the water to obtain the maximum nutrition from the meal, unless you have been boiling toxic substances out of the food.

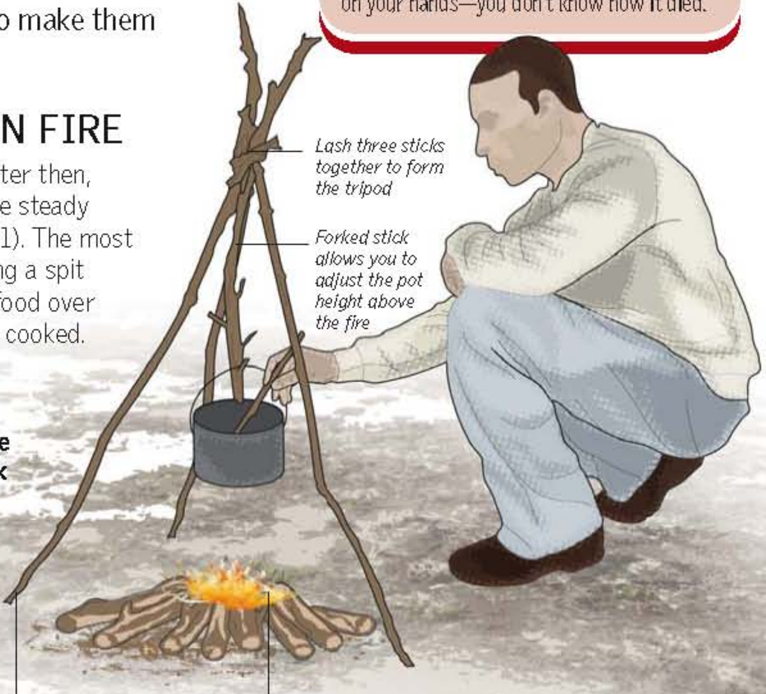

Lash three sticks together to form the tripod

Forked stick allows you to adjust the pot height above the fire

A tripod is a very stable and therefore safe way to suspend a cooking pot over a fire

Wait until flames die down a little before you begin cooking

DAMPER BREAD

This simple method for making yeast-free bread was developed by stockmen working out in the Australian bush.

① Mix flour and water (and a pinch of salt, if you have it) into a pliable dough, then roll it into a long, thick sausage shape.

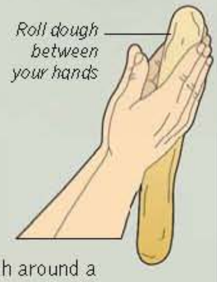

Roll dough between your hands

② Wind the dough around a stick, then hold it over the embers of a fire, turning regularly until it browns. It will slide easily off the stick when cooked.

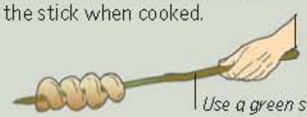

Use a green stick, which is less likely to burn

STEAMING FOOD

Steaming leaches fewer nutrients away than boiling and is a particularly good way of cooking fish and green vegetables—fresh leaves will be ready to eat in just a few minutes. To steam food, it must be suspended above boiling water in some way.

BAMBOO STEAMER

Bamboo stems are tough, hollow, and divided into sections. Use a sharp stick to make small holes in the walls dividing a three-section piece of cane, leaving the wall at the base intact. Pour water into the stem until it's just below the bottom section ring. Add food to the top section and cover with grass and a loose-fitting bamboo lid. Lean the stem over a fire, propping it against a forked stick.

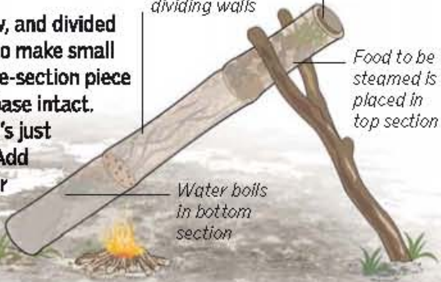

Steam rises up stem through holes in dividing walls

Place grass on top of food to retain heat and cover with a lid

Food to be steamed is placed in top section

Water boils in bottom section

CLAY BAKING

Baking food in clay requires no cooking utensils. Animals must be cleaned and gutted first, but can otherwise just be covered in soft clay. When the meat is cooked, skin, spines, or feathers will remain embedded in the clay. Cooking root vegetables or fish in this way removes their skins, losing valuable nutrients, so it's best to wrap them in leaves first.

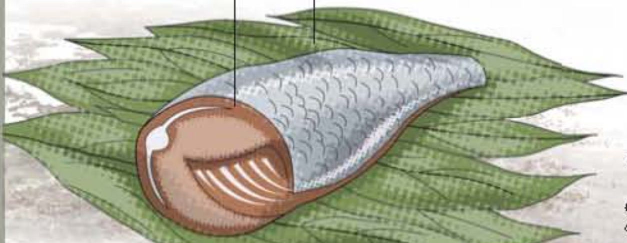
Protect nutritious skin of fish by wrapping in fresh leaves
Select long, wide green leaves

① Wrap the food in a bundle of fresh green leaves and tie them in place with some long strands of grass to make a secure parcel. Use only leaves from plants that you have identified as non-toxic.

② Cover the food parcel with an even layer of clay, making sure it's well sealed.

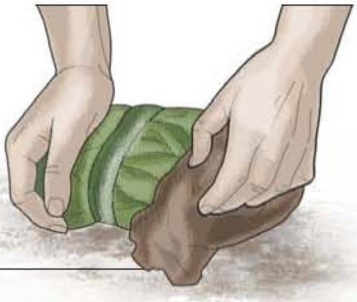
Layer of clay is 1 in (2.5 cm) thick

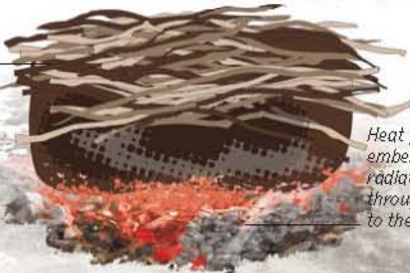
Build a fire on top of clay to increase the heat and decrease the cooking time
Heat from embers radiates through clay to the food

③ Place the parcel of clay in a bed of hot embers and build a fire up on top of it. Cooking time will be between 30 and 60 minutes, depending on the size of the food item. Break open the clay and remove the food.

COOKING WITH HOT ROCKS

Rocks take a while to heat up but they stay hot for a long time, allowing food to bake steadily on them. To reduce cooking time, cover the food with some birch bark or a flat piece of wood. When cooking on rocks, don't use slate or other layered rocks as they are likely to shatter when heated. Another method of cooking with hot rocks is to place them in a pit. The food is covered with leaves and placed on the rocks, then the pit is filled in to retain the heat. The food is left buried until it is ready to eat.

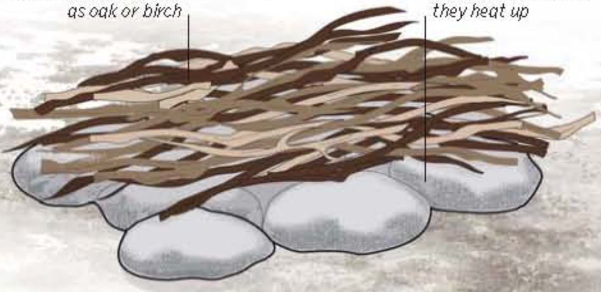
Use sticks from hard-wood trees such as oak or birch
Don't use wet rocks – they may explode as they heat up

① Place some large, reasonably flat rocks close together. Light a fire on top of the rocks using some tinder and dry sticks, then leave the fire to burn down to hot ashes while you prepare your food.

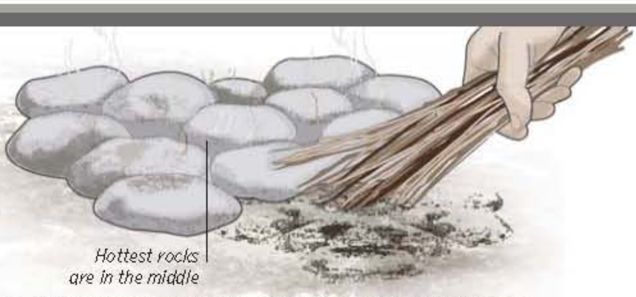
Hottest rocks are in the middle

② Brush the embers and ash off the rocks, taking care not to touch the rocks with your hands as they will now be extremely hot.

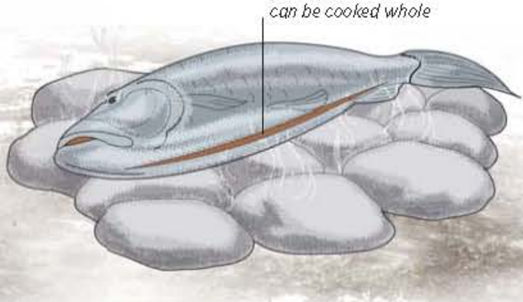
Fish should be gutted but otherwise can be cooked whole

③ Place food on top of the rocks to bake. The hottest rocks will be in the center, so items that must be cooked more slowly should be put near the edges. Keep adding more food to cook until the rocks have cooled.

EDIBLE PLANTS

IN A SHORT-TERM survival situation, food is not a priority—identifying edible plants uses energy, requires knowledge and skill, and the risks of getting it wrong far outweigh the benefits. In a long-term survival situation, however, your priorities would change, so being able to identify edible plants is advantageous.

GATHERING PLANTS

When foraging for plants, take a bag or can with you and be careful not to crush what you collect. Gather only a few species to lessen the chances of mixing in something inedible or poisonous. Don't assume that because birds or mammals have eaten a plant it's safe for you to eat too. Unless you are absolutely certain, you will need to carry out the following test to find out whether a plant is edible.

WARNING!

The Universal Edibility Test (see below) does not apply to mushrooms. For this reason alone, you should never eat any mushrooms that you have gathered (especially in a survival situation) unless you are 100 percent positive that you can identify each specific mushroom as being edible. In 2008, a British woman died after eating a poisonous mushroom that had been foraged by mistake with edible mushrooms.

DIGGING FOR ROOTS

Roots and tubers are a good source of carbohydrate and their skins contain vitamins. Take care when gathering in spring as some plants will only have small shoots and will be hard to identify.

1. Cut a stick from a hard-wood tree, such as oak, sharpen one end to a chisel-shape then harden it in a fire.

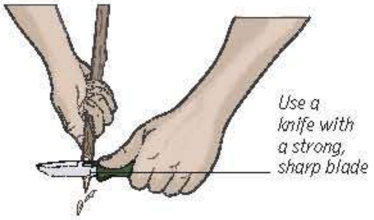
Use a knife with a strong, sharp blade

2. Choose a larger plant, then dig down at its side, loosening the earth around the root until it can be pulled out in one piece.

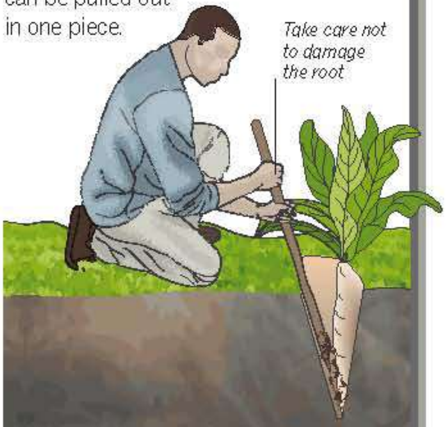
Take care not to damage the root

UNIVERSAL EDIBILITY TEST

This test enables you to check whether a plant is safe to eat. If you're in a group, only one of you should test the plant. Make sure that you have plenty of drinking water and firewood with you (see box, right), and eat nothing for eight hours beforehand. Test only one type of plant at a time, use only one part of that plant—roots, leaves, stalk, buds, fruit—at a time, and test it in the same state in which it will be eaten: either raw or cooked. Make sure that the plant you are testing is in plentiful supply (there's no point testing it otherwise) and avoid plants with milky or soapy sap, or bright colors—they are usually Nature's warning signal.

1. First, inspect the plant—it should look fresh and in good condition. Avoid anything slimy. Next, sniff the plant. Discard it if it smells bad. If it smells of peaches or bitter almonds, it may contain cyanide.

2. Take a small portion of the plant part you're testing and gently rub it on the sensitive skin inside your elbow or wrist. Wait for 15 minutes. If no irritation, stinging sensation, rash, or swelling develops, proceed to Step 3.

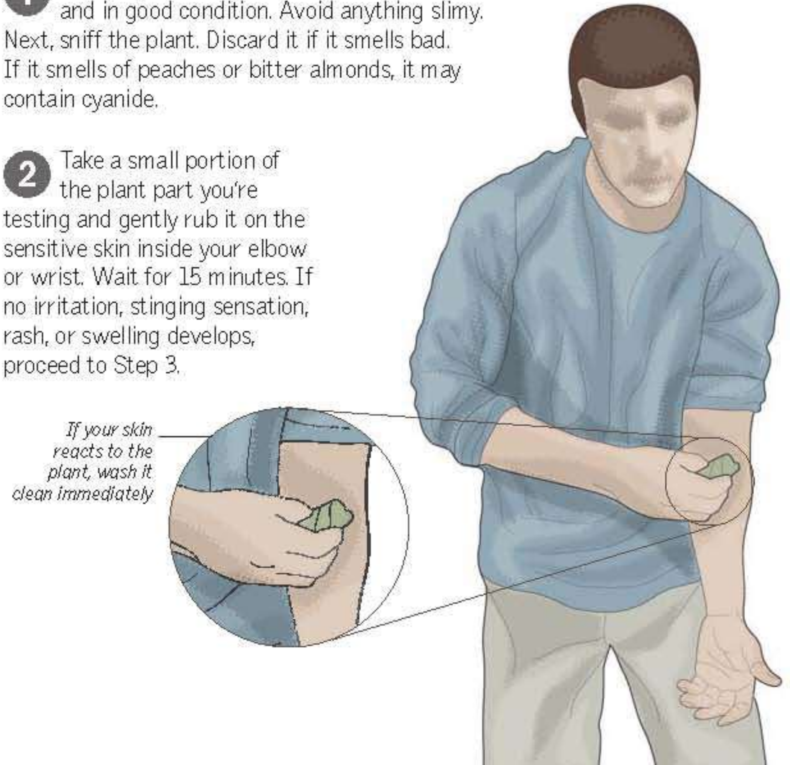
If your skin reacts to the plant, wash it clean immediately

NEEDLE TEA

The needles of evergreen trees such as pine and spruce are rich in vitamins A and C and can be used to make a refreshing—and potentially life-saving—tea.

1. For each cup of tea, collect two teaspoons of fresh green needles and bruise them with a stone.

2. Drop the needles into boiling water, then let them infuse for 10 minutes. Keep the pot warm and stir occasionally.

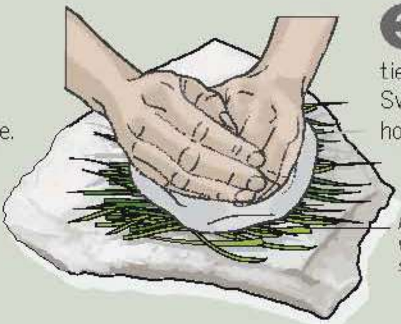

Rub needles with a large stone

3. Strain the tea through a cloth tied over a container. Sweeten with sugar or honey, if you have any.

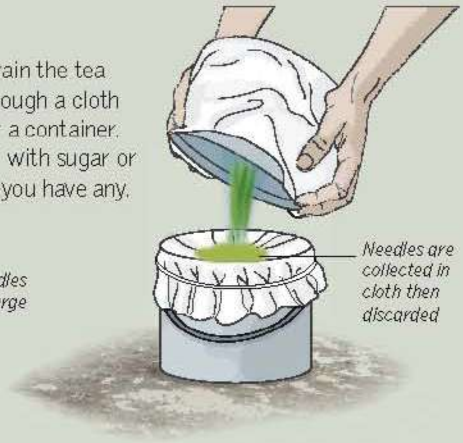

Needles are collected in cloth then discarded

3. Touch the corner of your mouth with the plant and wait 15 minutes. If no adverse reaction occurs, touch your lip and tongue with it. Wait another 15 minutes. If you suffer no ill effect, proceed to Step 4.

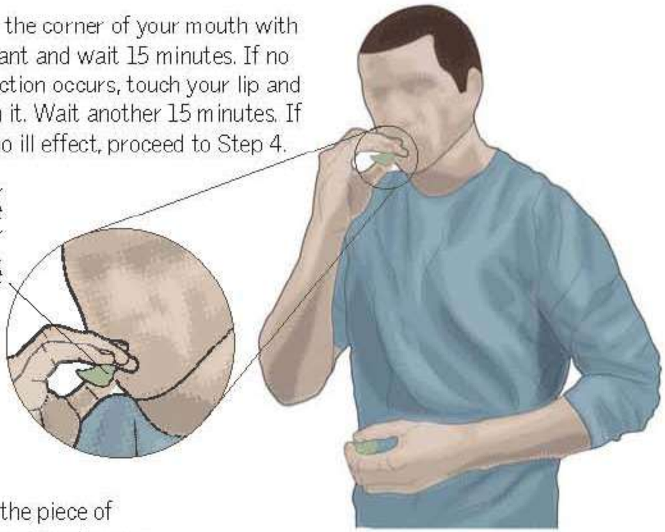

If your lips or tongue start to tingle or become numb, reject the plant

4. Place the piece of plant on your tongue, wait 15 minutes, then chew it and hold it in your mouth for 15 minutes more. If there are no ill effects, swallow the piece and wait eight hours. Eat nothing else but continue to drink water.

5. If no ill effect occurs, eat a quarter of a cup of the same part of the plant, prepared in the same way, then wait another eight hours. If there is no reaction, it is safe to eat. Eat little and often rather than gorging.

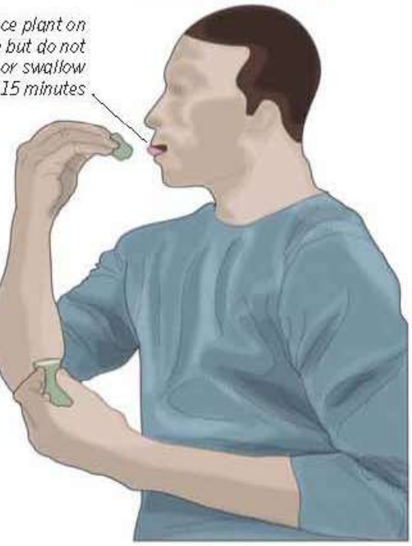

Place plant on tongue but do not chew or swallow it for 15 minutes

WHAT TO DO IF YOU HAVE AN ADVERSE REACTION

If at any stage you experience any type of adverse reaction to a plant during the edibility test, you should:

- Stop the test immediately.
- Avoid eating anything else until the symptoms cease.
- Wash the affected area thoroughly (if the reaction is external).
- Attempt to induce vomiting (if the reaction comes after eating the plant) by drinking salt water or by pushing a finger into the back of your throat.
- Drink lots of warm water (this is why you must ensure that you have a plentiful supply of purified water and firewood before you begin the test).
- Crush a teaspoonful of charcoal taken from a partially burned log and mix it into a paste with warm water, then swallow the mixture. It will either induce vomiting or, if you can keep it down, it may absorb the toxins.
- Try swallowing a paste made from white wood ash and water to reduce stomach pain.

If the tester is unconscious:
Do not induce vomiting if the tester is unconscious—get medical help and monitor the casualty.

CATCHING FISH

FISH CAN BE HOOKED, netted, trapped, speared, or even caught by hand if you're lucky. Observe their behavior—where and when they feed, and what they eat—to determine which method to try. It must be noted that some fishing techniques are illegal in some parts of the world and should only be used in a genuine survival situation.

BAIT AND LURES

Fish have a keen sense of smell to help them find their preferred food, so if one type of bait doesn't work, try another. Predatory fish are attracted to live bait by its movement—a worm wriggling on a hook, for example. You can make artificial bait out of shiny metal, cloth, or feathers. Try to imitate the fish's natural prey, such as an insect skimming across the water, to lure the fish into biting.

FISHING TACKLE

Fishing equipment can be made from all kinds of material. Your survival kit should have some fishing line and a few hooks, but if you don't have a kit you can improvise. You can use a stick as a makeshift rod, for example.

IMPROVISED HOOKS

Fishhooks can be fashioned from any piece of metal, such as a nail, pin, safety pin, needle, or some wire. You can also make them from natural materials, such as thorns, hard wood, coconut shells, bone, spines, or seashells. If you don't have any fishing line you can use parachute cord (see p. 136) or make cordage from plant fibers (see p. 138).

MAKING A FISHING REEL

You can make a reel with an empty beverage can. Tie one end of a line to the tab, then wrap it around the can until you have about 2 ft (60 cm) left. Attach a hook, float, and some sinkers. Hold the top of the can in one hand and the float in the other. Point the bottom of the can toward where you want the hook to land and throw the float. The rest of the line will unwind and follow. If a fish bites, give the line a jerk then wind it in around the can.

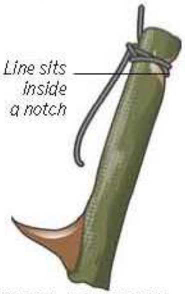
Line sits inside a notch

SINGLE THORN
Cut a 1 in (2.5 cm) length of bramble stem with a large, strong thorn. Tie a line to a notch at the other end.

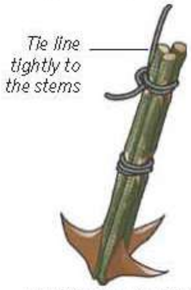
Tie line tightly to the stems

MANY THORNS
Tie three strong thorns together securely with some fishing line. Cut a notch in the stems and attach the line.

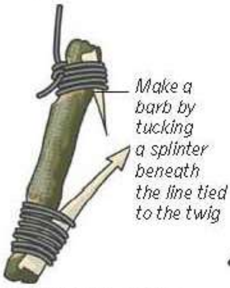
Make a barb by tucking a splinter beneath the line tied to the twig

WOOD OR BONE
Carve a splinter of bone into an arrow point or take a sliver of hard wood, like oak, and tie to a twig.

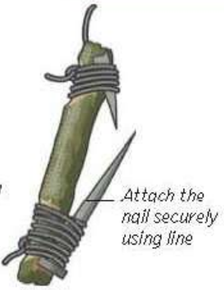
Attach the nail securely using line

NAIL
Cut a notch at one end of a small piece of hard wood. Place the head of a nail in the notch and tie it in place.

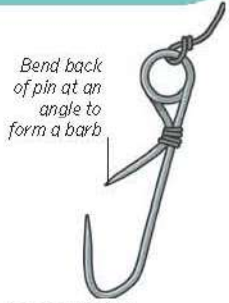
Bend back of pin at an angle to form a barb

SAFETY PIN
Remove the safety clasp. Bend the sharp point of the pin round to form a hook. Attach line to the wire loop.

REMOVING A FISHHOOK

You should never try to remove a fishhook that is deeply embedded in flesh. Cut the fishing line as close to the hook as possible, then put a pad around the exposed part of the hook and secure with a bandage. Seek medical help as soon as possible and monitor for signs of infection. However, in a survival situation, you can try to remove the hook by doing the following:

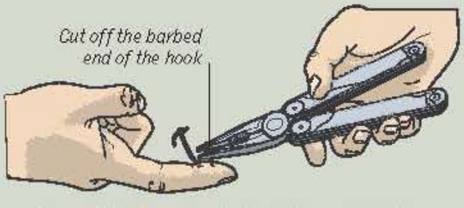
Cut off the barbed end of the hook

① If the barb is visible, cut it off with a pair of pliers. If you can't see the barb, firmly and quickly push the hook in further until the barb emerges.

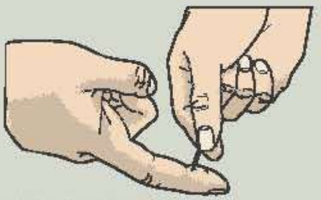

② Carefully ease the hook out by its eye. Clean the wound then cover it with a bandage.

TYPES OF BAIT

Live bait includes creatures such as:

- Worms and slugs
- Grasshoppers, crickets, and beetles
- Maggots and caterpillars
- Frogs (including large tadpoles)
- Small fish—to catch larger fish

Examples of inanimate bait include:

- Meat, guts, and reproductive organs recovered from animals
- Nuts and small fruits
- Bread, cheese, and pasta

FEATHER LURE

Make a hook, then attach it to a line. Tie a brightly colored feather just above the hook. Slowly move the lure across the water's surface to attract the fish.

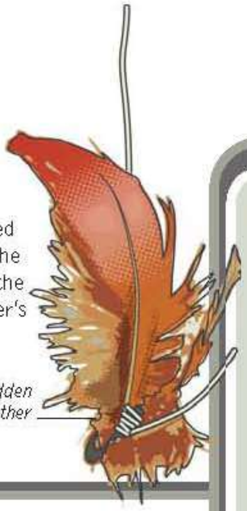

FLOATS AND SINKERS

A float keeps a baited hook at the best depth to attract the species of fish you're trying to catch. A sinker (weight) placed below the hook also helps to hold it in position. Your survival kit should include a few split-shot weights but, if it doesn't, you can tie small stones to the line instead. If the float bobs below the water's surface, you may have hooked a fish.

MAKING A FLOAT

You can use any natural material that floats, such as a piece of bark or a rose hip (see below). If you find a bird's flight feather, you can make a float by trimming the feather until you are left with just the hollow quill, then folding it in half and tying the two ends together.

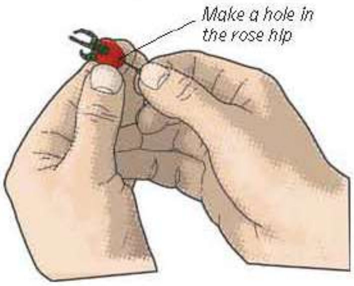
Make a hole in the rose hip

1 Using a sharp pin, a piece of wire, or a long thorn, pierce through the center of a rose hip.

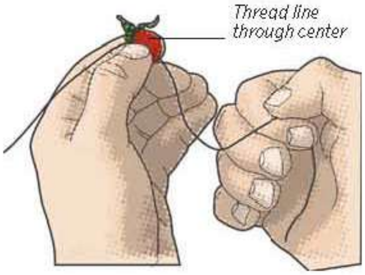

2 Thread fishing line through the hole. If you don't have line, use a length of fine cordage.

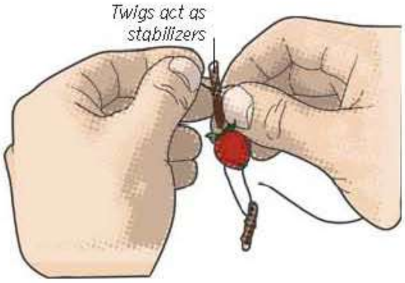

3 Tie small twigs above and below the rose hip. These will act as stabilizers and will also prevent the float from slipping up or down the line.

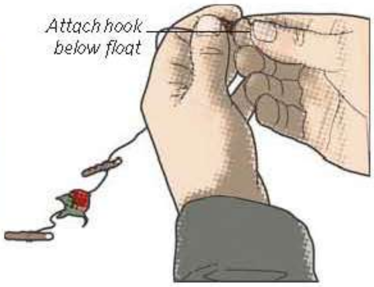

4 Attach the hook (in this example, a modified safety pin) to the line at the required depth below the float. If you have one, tie a sinker below the hook.

LINE FISHING

The more hooks you can get in the water, the greater your chances of catching a fish. All the methods shown below are passive—you simply set them and leave them to work for you.

SELF-STRIKING LINE

If a fish bites at the bait, the trigger will be jerked out of the catch and the bent rod will fly up. The line will be pulled tight, embedding the hook in the fish's mouth.

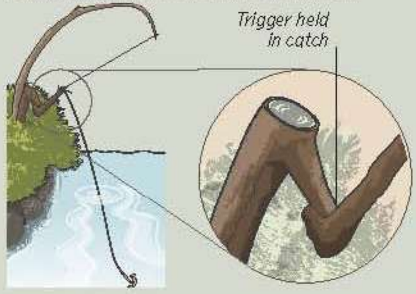

ICE FISHING

First check that the ice is at least 2 in (5 cm) thick and can bear your weight, then cut a hole about 1 ft (30 cm) in diameter over the deepest part of the lake.

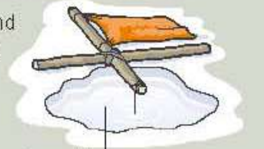

1 Tie a hooked and baited line to a stick with a flag at the other end and a central crosspiece.

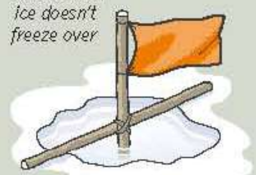
Ensure the ice doesn't freeze over

2 If a fish takes the bait, the crosspiece will be jerked over the hole and the flagpole pulled upright.

NIGHT LINES

To increase your chances of making a catch, you need to attract fish that live at all depths. Tie a rock to one end of a line and attach hooks with live bait at intervals along it. Tie the other end to a post at the edge of the bank and throw the line in the water, leaving it there overnight.

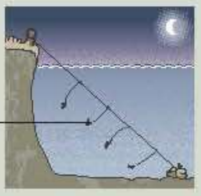
Keep hooklines short to avoid them getting tangled

NETS, TRAPS, AND HARPOONS

You can leave a hooked line in the water, but unless you regularly check it any fish you catch may be eaten by bigger fish. In a survival situation, nets and traps are more convenient methods of fishing because once set you can leave them and they work for you all the time.

WARNING!
Gill nets and traps are illegal fishing methods in many parts of the world, so use them only in an emergency.

MAKING A DIPPING NET

Fish that are too small to hook or harpoon may still be large enough to catch with a dipping net. Such fish are usually found at the edges of streams and lakes, and around rocks in pools. You can make a simple dipping net if you can split a branch or find a forked sapling and have a mosquito net—or a spare item of clothing, such as a sleeveless vest, T-shirt, or stockings.

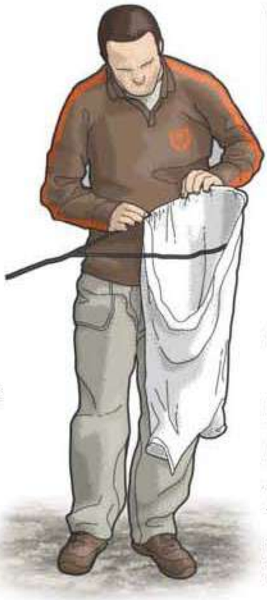

1 Make two small nicks in the hem of the mosquito net, vest, or stockings, then thread the forks of the sapling through them.

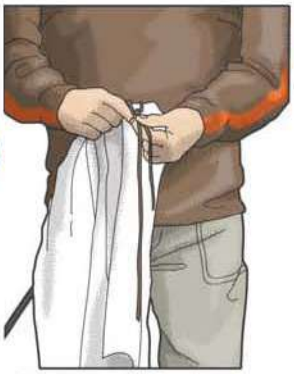

2 Cut another nick in the other side of the hem, where the forks meet, and pull them through. Tie the ends together with cordage, then push them back inside the hem.

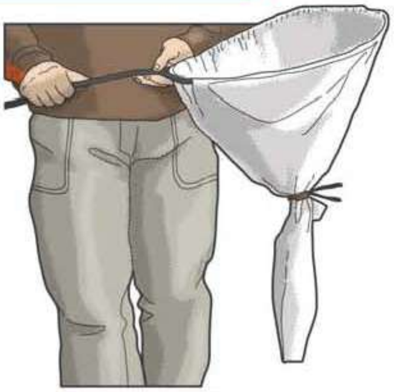

3 Tie off the net (or the vest above the armholes and neck), then either cut off the excess material or invert the net. This will prevent extra drag when you use the net.

MAKING A GILL NET

Once it has been set in a river, a gill net is a highly effective way of catching fish, whether they are swimming up- or downstream. However, a gill net should only be set for short periods of time as fish of all sizes can be entangled or injured. Make the lines out of paracord or natural cordage (see pp. 136–39).

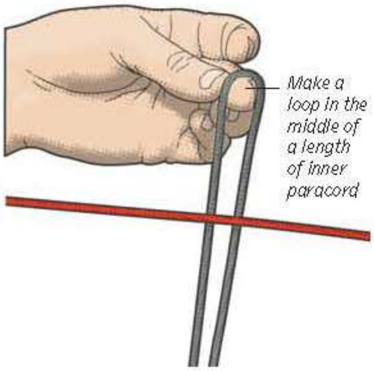
Make a loop in the middle of a length of inner paracord

1 Take a section of the outer sheath of some paracord and suspend it between two trees. Loop the inner paracord behind this line.

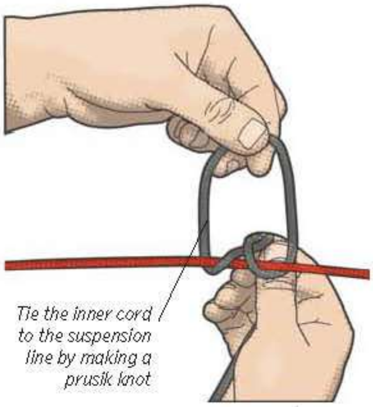
Tie the inner cord to the suspension line by making a prusik knot

2 Holding the two ends of the inner cord in your other hand, pass them around the suspended line and through the loop.

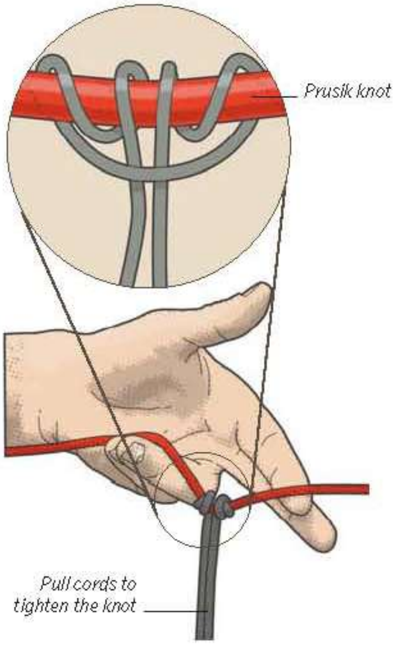
Prusik knot
Pull cords to tighten the knot

3 Pull the knot tight. Repeat the process every 1½ in (4 cm) along the suspended line for the required width of the net.

MAKING A HARPOON

Using a harpoon to spear fish takes time, patience, and a certain level of skill. Keep the point of your harpoon in the water to avoid splashing and scaring off the fish. Strike quickly when you see a potential catch, aiming just ahead of it. A spiked harpoon is the simplest to make.

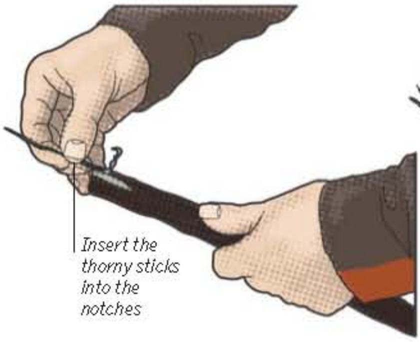
Insert the thorny sticks into the notches

1 Gather a few short, thorny sticks to form the barbs of the harpoon. Cut a long, thick, straight branch and make notches around one end.

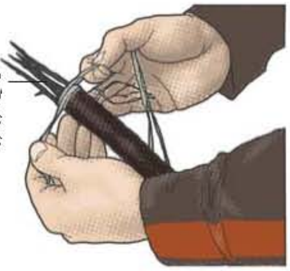
Wrap cordage around the sticks several times

2 Bind the thorny sticks tightly to the shaft with cordage. They will need to be able to support some of the weight of the fish so wrap them securely.

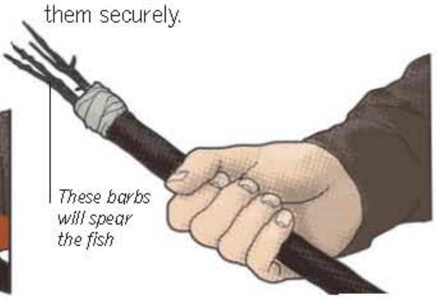
These barbs will spear the fish

3 When using the harpoon, take care not to damage the barbs by driving them against rocks or onto the riverbed. Use a dipping net, if you have one, to land the fish.

4 Make the mesh by tying a cord from one pair of cords to a cord from an adjacent pair, using a simple overhand knot (see p. 143).

- Continue alternating along and down the line. Tie stones to the ends of the net to hold it down once it is in the water.

5 To set the gill net, hang the suspension line between trees or poles on either side of the river, submerging the net in the water to a depth of about 6 in (15 cm).

- The net can also be used to catch birds (see p. 226).

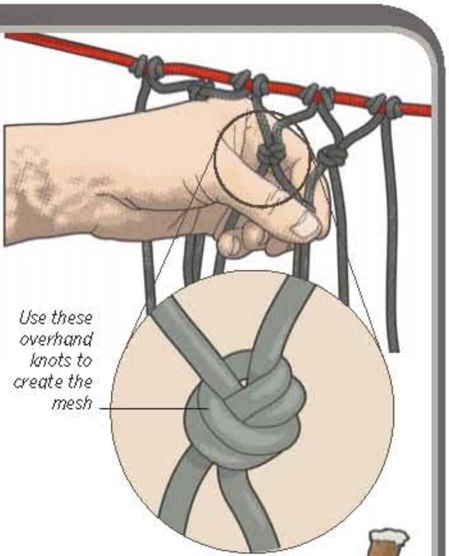
Use these overhand knots to create the mesh

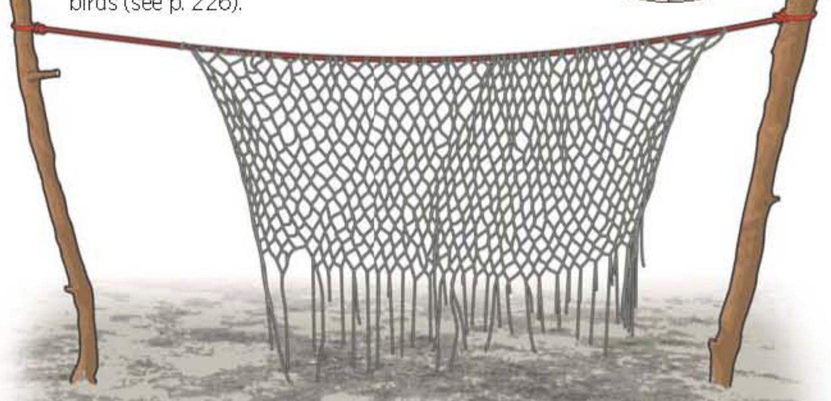

TRAPS

Fish traps can be made from a variety of materials. You can use rocks to build a wall or you can make a basket from sticks and cordage. You can even use man-made materials, such as bottles. If you only succeed in catching very small fish you can still use them as bait.

BOTTLE TRAP

Also known as a "minnow trap," this method uses a large plastic bottle to catch small fish. Cut the bottle in two just below the shoulder. Invert the neck and insert it into the cut end, then tie the two pieces together. Make holes in the plastic with a hot pin so the trap will sink. Bait the trap, then place it in a stream. Check it regularly to remove any fish caught and replace the bait.

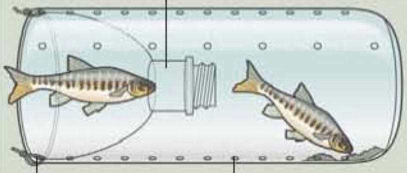
Fish smells bait and swims into bottle but can't find its way back out

Bottle neck tied to bottle base

Holes in bottle let water in

SINGLE-WALL TRAP

If your camp is near a tidal creek, you can build a curved wall of large rocks out from the bank. Pick the location at high tide and build a low wall at low tide. Fish may be trapped in the pool created between the wall and the bank as the tide recedes.

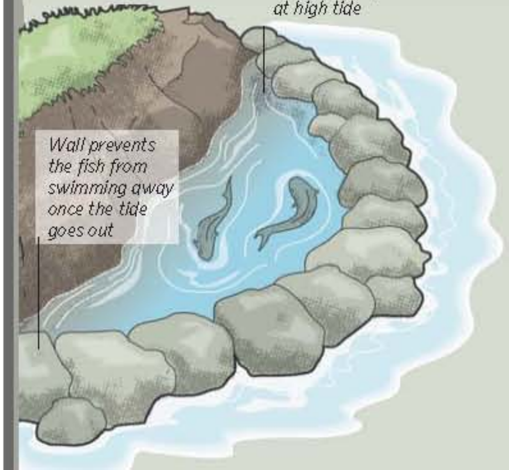
Water and fish enter the trap at high tide

Wall prevents the fish from swimming away once the tide goes out

PREPARING FISH

ALL FRESHWATER FISH are edible but they must be cooked before eating as they are often infested with parasites and harbor bacteria. Saltwater fish are safer to eat raw, if necessary, but will taste better cooked. Never eat a fish that has pale gills, sunken eyes, flabby skin or flesh, or that smells bad.

WARNING!
When handling fish be careful not to accidentally touch your eyes. The slime may contain bacteria that can cause a painful inflammation of the eye covering (conjunctivitis). Make sure you wash your hands clean of all slime. If you have any cuts on your hands, cover them before touching the fish.

FISH HYGIENE

As soon as you have landed a fish, it should be killed and gutted, then cooked and eaten as quickly as possible, particularly in a warm climate. It doesn't take long for fish to go bad because their slimy skin provides a breeding site for flies and bacteria. In a cold climate, you can delay filleting for up to 12 hours, which will make the job easier.

FILLETING A FISH

Filleting removes the parts of a fish that might quickly go bad, while leaving as much flesh as possible. With some species, you may find it easier to remove the bones after the fish has been cooked. Boil the bones and the head to make a nutritious stock. This must be done immediately and the stock should be kept in a cool place and drunk within a few hours.

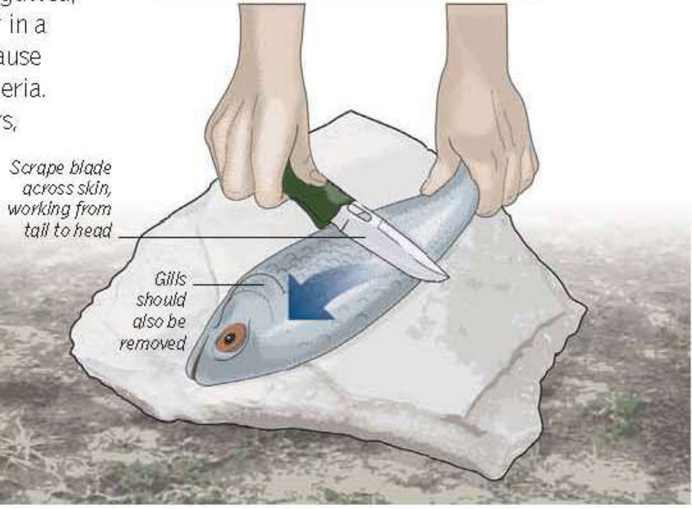
Scrape blade across skin, working from tail to head
Gills should also be removed

1 Once you have killed the fish, cut its throat to bleed it and cut out the gills. Wipe the slime off its skin to make it less slippery.

- Most fish don't need to be skinned—in fact, the skin is nutritious—but eels and catfish do (see panel, opposite).

2 Fish can be cooked with their scales on but, if you have time, it is best to remove them, particularly if they are large, as they can be a choking hazard.

- Hold the fish by the tail and scrape off the scales, holding the blade away from you and moving toward the head.

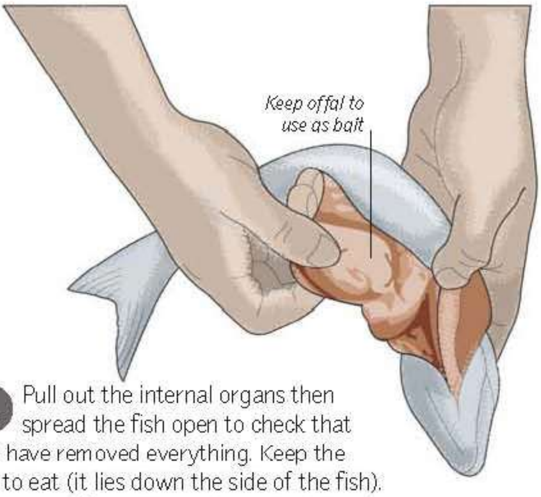
Keep offal to use as bait

4 Pull out the internal organs then spread the fish open to check that you have removed everything. Keep the roe to eat (it lies down the side of the fish).

- Wash the fish thoroughly, inside and out.

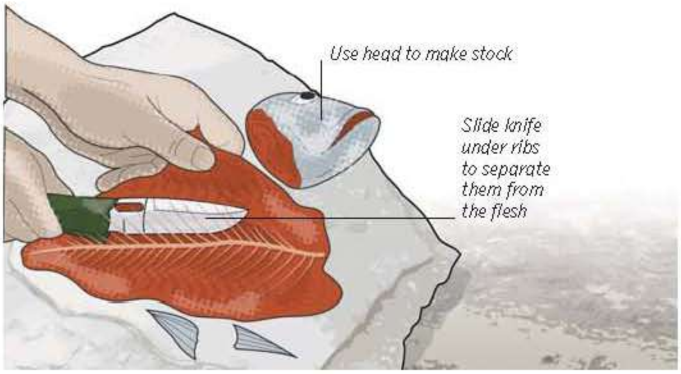
Use head to make stock
Slide knife under ribs to separate them from the flesh

5 Cut off the head, tail, and fins. Open out the body and slide the knife under the ribs to separate them from the body, working toward the head. Fish oils can make the knife slippery, so you can use your thumbs instead.

- Repeat for the ribs on the other side of the backbone.

SMALL FISH

Any fish that is less than 6 in (15 cm) long doesn't need to be filleted. After gutting, they can be fried, grilled, or roasted whole. Keeping the heads and tails on helps to prevent the fish from falling to pieces.

SKEWERING FISH

Small fish, such as perch, can be grilled over an ember fire on a skewer to make a simple, yet tasty and nutritious meal. Impale the fishes on a green stick and hold them close to the embers. They won't take long to cook.

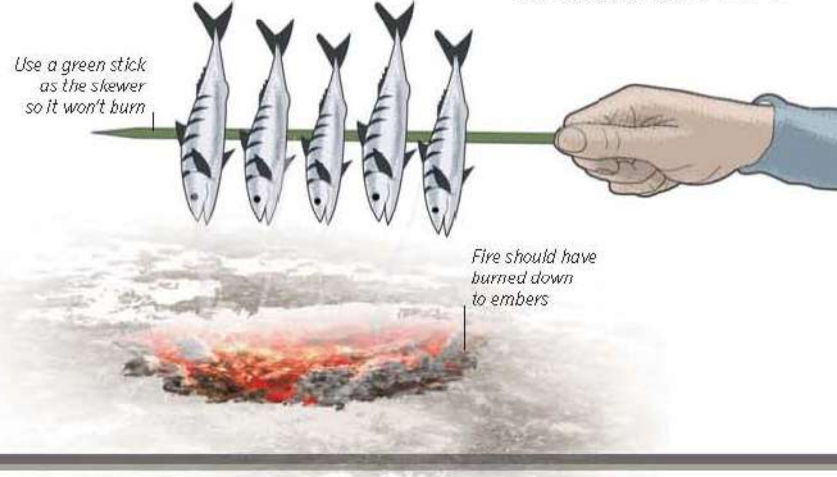
Use a green stick as the skewer so it won't burn

Fire should have burned down to embers

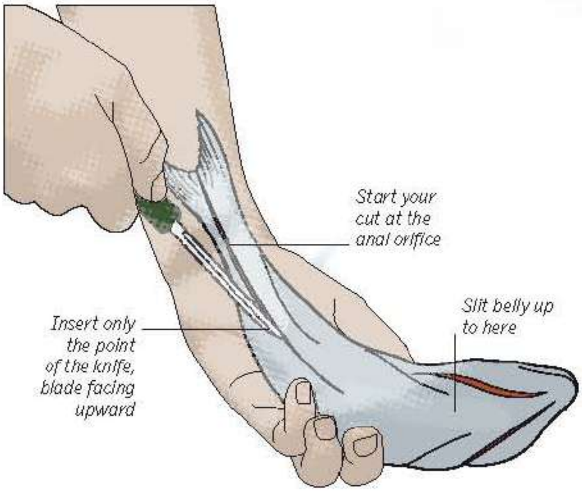
Start your cut at the anal orifice

Insert only the point of the knife, blade facing upward

Slit belly up to here

③ Holding the fish with its tail toward you, insert the point of a sharp knife, blade up, into the anal orifice and slit the fish open along its belly to its throat. This prevents you from puncturing the internal organs.

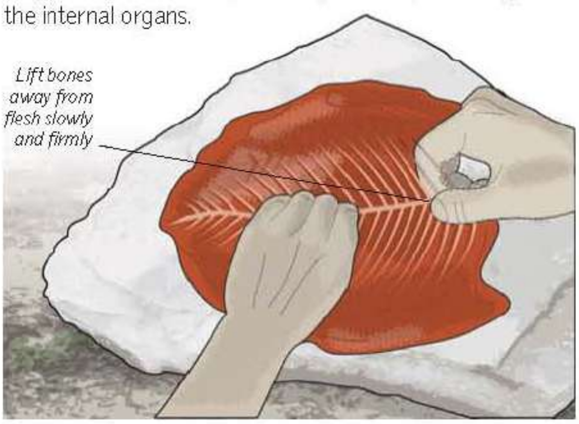
Lift bones away from flesh slowly and firmly

⑥ Separate the top and bottom of the backbone from the flesh with the tip of your knife. Carefully pull the backbone and ribs away from the flesh in one piece.

- If you have any tweezers or needle-nose pliers, use them to remove any small bones that remain.

SKINNING FISH

Eels and catfish are tasty but have to be skinned and gutted before cooking. To gut the fish, either use the filleting method shown opposite or Step 3 below. Catfish have a cartilage skeleton and can simply be cut crosswise into steaks.

① Pass a stake through the gills of the fish and support on two strong uprights.

- Cut around the skin below the head with a sharp knife.
- Cut around the fins.

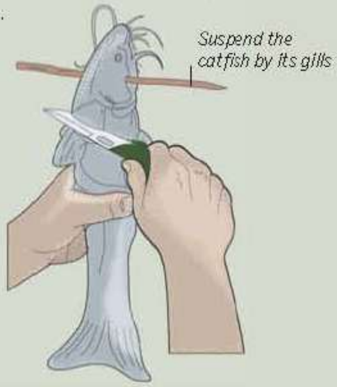
Suspend the catfish by its gills

② Separate the top of the skin from the flesh then peel the skin downward.

- You will need to use both hands to get a good grip, and a firm, steady action. If the fish is very large, you may need to slit the skin.

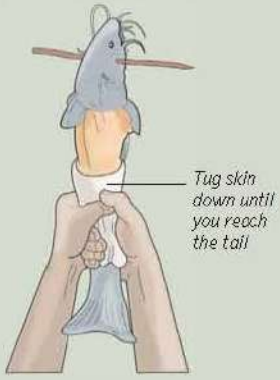
Tug skin down until you reach the tail

③ Remove the fish from the stake and break its backbone. When you pull the head off, the guts will come away with it.

- Remove tail and fins.

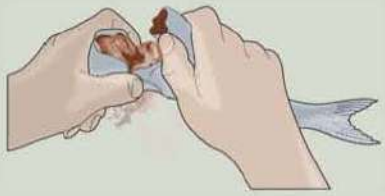

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—ADRIFT AT SEA

USEFUL EQUIPMENT

- Liferaft, lifejacket, survival suit
- Water catchment devices
- Solar still/reverse-osmosis pump
- Emergency Locator Beacon
- Marine VHF radio
- First-aid kit, sunscreen, sunglasses
- Flashlight and batteries
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

IN 1972, A FAMILY OF FIVE SURVIVED for 38 days adrift in the Pacific Ocean. The Robertson family were 18 months into a round-the-world sailing trip when their 43 ft (13 m) wooden schooner *Lucette* sank. They survived thanks to a combination of good seamanship, improvisation, and good fortune.

The family set off from Falmouth, England, on January 27, 1971—father Dougal at the helm, wife Lynn, and children Anne (who disembarked in the Bahamas), Douglas, and twins Neil and Sandy as crew—and safely navigated the Atlantic and Caribbean. On June 15, 1972, however, about 200 miles (320 km) west of the Galapagos, disaster struck—a pod of killer whales charged the boat, splintering the hull and holing her irreparably.

"A POD OF KILLER WHALES CHARGED THE BOAT...HOLING HER IRREPARABLY"

With *Lucette* sinking rapidly and no time to radio a distress call, the crew launched their inflatable liferaft and roped it to the boat's 10 ft (3 m) solid-hull dinghy, the *Ednamair*, which they used as a towboat after improvising a sail. Their supplies amounted to 2 gallons (10 liters) of water, a bag of onions, oranges and lemons, vitamin-fortified bread, glucose, four fishing hooks, a fishing line, a first-aid kit, a kitchen knife, and eight signal flares. They sailed north toward the Doldrums to find rain, which they collected with the use of a tarpaulin, and caught fish, eating some of the meat raw and drying the rest in the sun to be stored as rations.

After 16 days, the liferaft had deteriorated to such an extent that the family was forced to transfer to the dinghy, using the remnants of the raft as a canopy to provide shelter and aid in the collection of rainwater. They used the wind and currents to sail northeast toward Central America, and built up sufficient rations over the next three weeks to provide energy for the extra exertion of rowing toward the coast. Luckily for them, there was no need—their ordeal came to an end on July 23, when a Japanese fishing boat spotted a signal flare and picked them up.

WHAT TO DO

ARE YOU IN DANGER? +

NO YES

- If you are in a group, try to help any others who are in danger
- Prepare for the possibility of abandoning ship, and try to steer closer to either land or known shipping lanes
- Delegate responsibilities
- Ensure lifejackets and liferafts are ready for deployment

Get yourself out of it:
Sinking vessel—You need water, location aids, and protection from drowning and the elements
Animals—Try not to splash in the water as this will attract sharks
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO YES

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO YES

You are faced with surviving for an indefinite period—until you are located or you find help

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO YES

Abandon ship in a controlled manner and deploy all the liferafts. Always try to enter the liferaft dry. Use the rescue line to reach people in distress. Follow the "immediate actions" instructions printed inside the liferaft

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Deploy the drogue—this will lessen drifting
- Inventory all food, water, and equipment and start rationing
- Prepare water procurement devices, such as solar stills and reverse osmosis pumps
- Protect yourself from the elements—sun, wind, and salt spray
- If you have no liferaft, then huddle together in a group with children in the center. If alone, adopt the H.E.L.P. position
- Improvise flotation aids from anything that can float or hold trapped air. Good examples are plastic bags, plastic bottles, and knotted clothing

DON'T

- Cut the painter to the vessel until you know it will sink, as the vessel is what people will be looking for
- Drink seawater under any circumstances—this will only increase your rate of dehydration
- Eat unless you have sufficient water to digest the food—fish is high in protein and requires plenty of water to digest

YOU SHOULD STAY **

DO

- Ensure that the rescue services are kept aware of your situation and updated with relevant information as the situation dictates
- Prepare to abandon ship: pack lifejackets and suitable clothing
- Ensure you know how to operate your survival equipment and that you have basic aids to location such as a whistle, flashlight, or plastic water bottle on your person
- Clear the deck of loose objects and be prepared to drop sail should rescue by helicopter be attempted
- Make an inventory and ration supplies

DON'T

- Neglect to take anti-seasickness tablets. Vomiting will dehydrate you and the effects are very demotivating
- Forget to wear your survival equipment and ensure everyone knows how to operate it
- Try to second-guess the rescuers when they arrive. Do exactly as they say, they know what they are doing

+ Never abandon a vessel unless you have to—searchers will be trying to locate it. Use the vessel's equipment while you can, and before you leave collect essential items to aid your survival.

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.

** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

TRAPPING ANIMALS

ALTHOUGH YOU SHOULD be prepared to take down sitting prey if you get the opportunity, trapping small animals is easier than hunting them—it requires less skill and energy and leaves you free to carry on with other tasks. One of the simplest traps is a snare.

MAKING A SNARE

Stainless steel snares of various strengths can be bought ready-made, with a running loop, or eye, at one end and a securing loop at the other. However, most survival kits contain a length of single-strand brass wire that can be used to make a snare.

1. Decide on the strength of snare you want (see panel, below) and double or quadruple the wire accordingly.

- Pass the strands around a stick, place it on the ground, then loop the loose ends around a second stick.
- Rotate the second stick until the strands have entwined, forming a single, thick wire. Remove the sticks.

STRENGTHS OF SNARE

The single-strand wire found in most survival kits isn't strong enough to hold most animals you would want to catch. You'll need to double or quadruple the strands by winding them together to increase their strength.

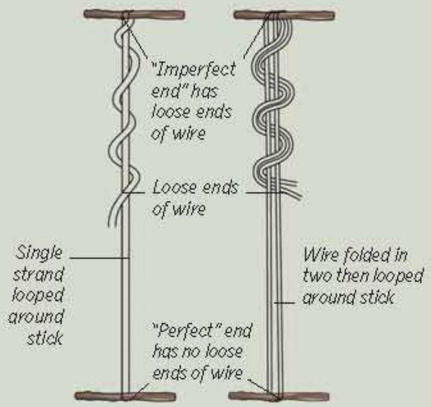
"Imperfect end" has loose ends of wire
Loose ends of wire
Single strand looped around stick
"Perfect" end has no loose ends of wire
Wire folded in two then looped around stick

TWO STRANDS
A two-strand snare will be sufficient to bear the weight of a squirrel-sized animal.

FOUR STRANDS
A four-strand snare will be strong enough to hold onto an animal the size of a rabbit.

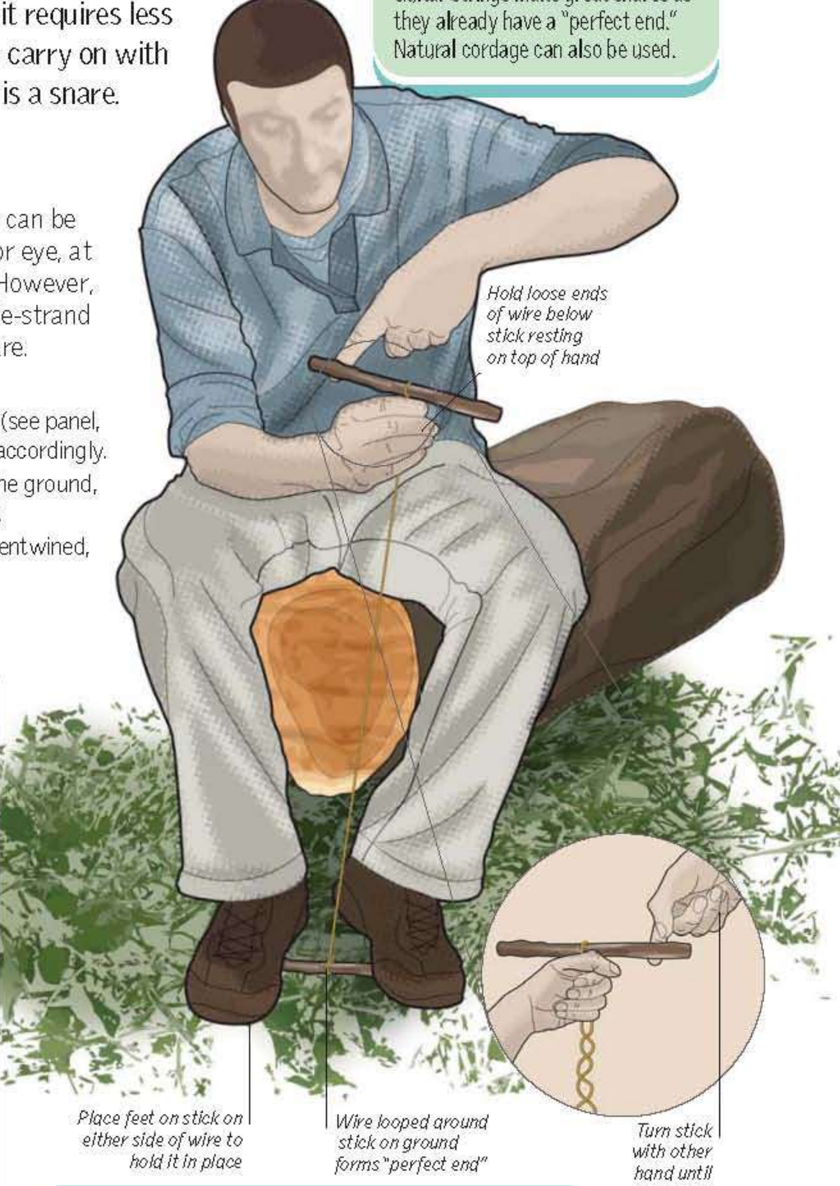
Hold loose ends of wire below stick resting on top of hand
Place feet on stick on either side of wire to hold it in place
Wire looped around stick on ground forms "perfect end"
Turn stick with other hand until the wire strands are intertwined

ALTERNATIVE MATERIALS

Guitar strings make great snares as they already have a "perfect end." Natural cordage can also be used.

KILLING A RABBIT

Unless you're an expert, breaking a rabbit's neck by holding its head and pulling on its legs can be difficult and could result in you dislocating its hips, causing it more pain. The best method is to strike the rabbit on the back of the neck with a solid stick, thus avoiding having to pick it up and risk being bitten or scratched. The eyes will glass over immediately when the animal is dead.

LOOK AND LISTEN

To be successful at hunting and trapping, you need to decide which types of animal you are going to try to catch. To do that, you need to find out what animals are in the vicinity and where exactly they are to be found. Prey animals will use their keen senses to avoid being caught—so you will have to use your senses to find them. Look for any signs of animals, both on the ground and in the trees. And use your ears—you may be able to hear an animal even when you can't see it.

SIGNS TO LOOK OUT FOR:

- Runs, trails, and tracks
- Droppings
- Chewed or rubbed vegetation
- Feeding and watering areas
- Lairs, dens, and resting sites

SQUIRREL POLE

Make several two-strand snares with nooses 3 in (8 cm) in diameter and stake-out cords 8 in (30 cm) in length. Place them around a long pole, with the lowest 2 ft (60 cm) off the ground. If you snare a squirrel, leave it, as others will come to investigate.

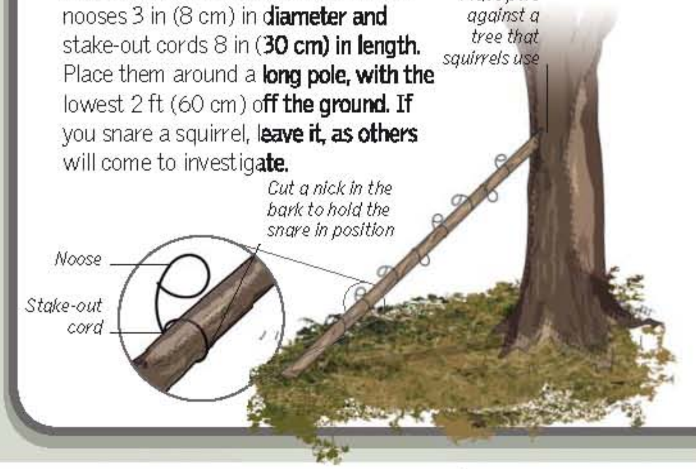
Place pole against a tree that squirrels use
Cut a nick in the bark to hold the snare in position
Noose
Stake-out cord

2 To make the noose, pass the imperfect end, or securing loop, through the perfect end, or running loop. The running loop has no loose ends of wire so it won't snag in an animal's fur and prevent the noose from tightening as it struggles.

- Use a single prusik knot (see p. 210) to tie a length of cord to the securing loop.

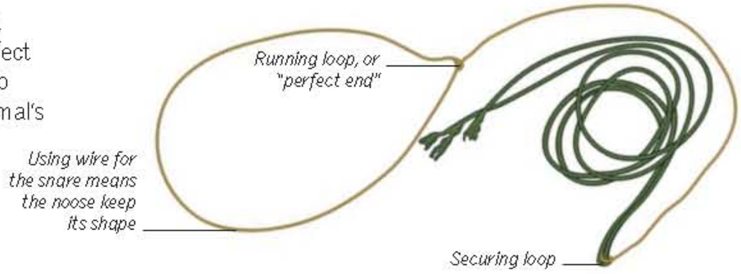
Running loop, or "perfect end"
Using wire for the snare means the noose keep its shape
Securing loop

3 Before setting the snare, either bury it in the ground for a few hours or pass it over a flame for a few seconds to remove any human scent and dull its surface.

- Don't waste snares; always set them on runs you know the animals are using and set as many as you can.
- Use sticks to make a support from which to suspend the snare, or bend saplings into an arch.

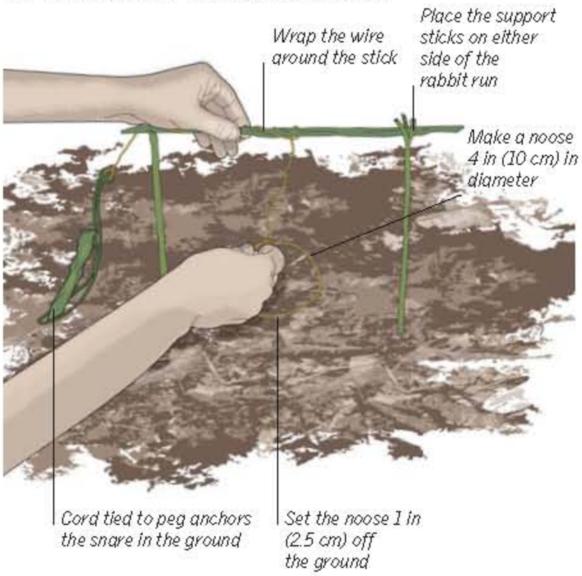
Wrap the wire around the stick
Place the support sticks on either side of the rabbit run
Make a noose 4 in (10 cm) in diameter
Cord tied to peg anchors the snare in the ground
Set the noose 1 in (2.5 cm) off the ground

4 Avoid disturbing or treading on the run, as the disturbance or your scent could alert the animal. Camouflage the frame with vegetation, which you can also use to funnel the animal toward the snare.

- Use natural vegetation, such as holly or other prickly shrubs, to form a funnel on either side of the snare.
- The funnel helps to ensure that the animal has no option but to pass through the snare.

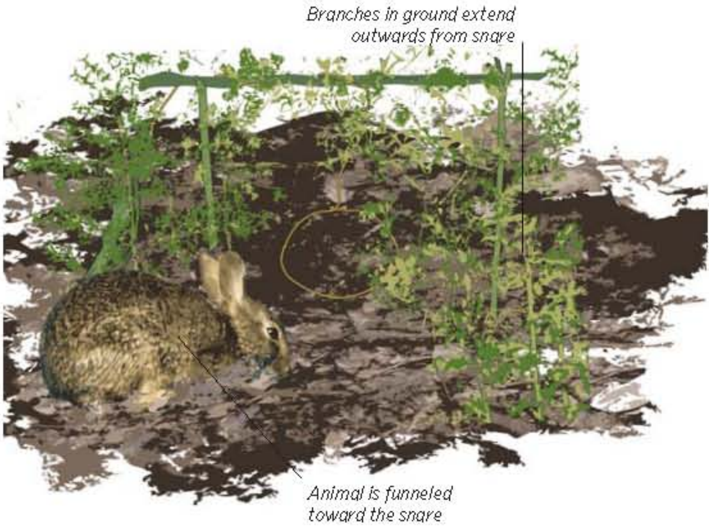
Branches in ground extend outwards from snare
Animal is funneled toward the snare

WEAPONS AND TECHNIQUES

A spear is the simplest weapon to make and use, but if you have the time and materials you can make more complicated weapons, although they require more skill to use. Learning other basic techniques, such as catching insects, is also invaluable when hunting.

THROWING STAR

With a throwing star you're four times more likely to hit prey than with a spear, and, if you don't succeed in wounding the animal with one of the sharp ends, the weight of the weapon will at least stun it. Keep it close at hand and ready to use, should the opportunity arise.

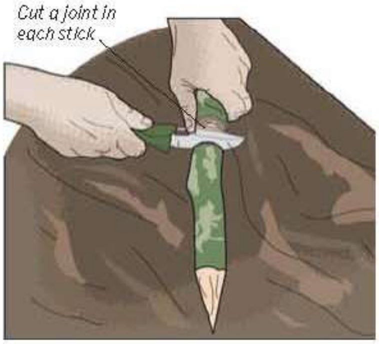
Cut a joint in each stick

① Find two sticks 2 in (5 cm) thick and 18 in (45 cm) long. Make a square-cut joint in the center of each, then sharpen each end.

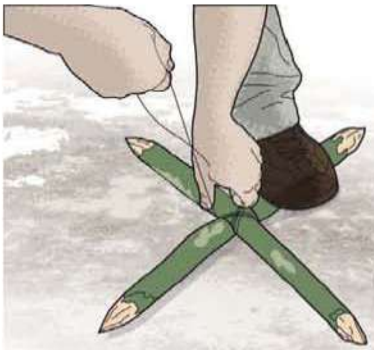

② Join the sticks together by overlapping the two square-cut joints and lashing them to each other with paracord or natural cordage.

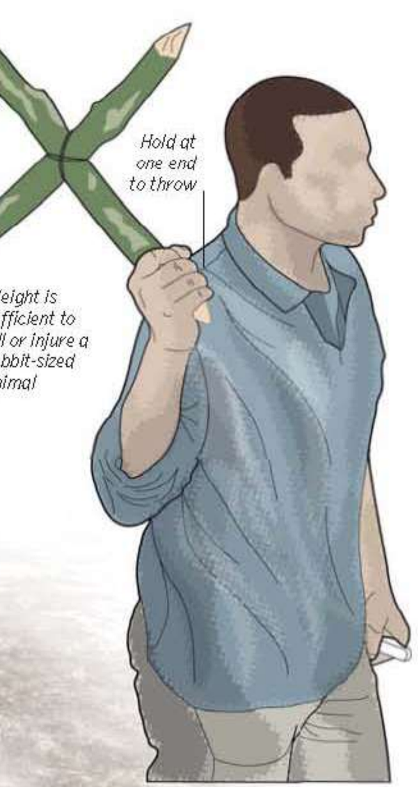
Hold at one end to throw

Weight is sufficient to kill or injure a rabbit-sized animal

③ You can exert more force if you're throwing from a standing position. Aim for the animal's legs.

RODENT SKEWER

A forked spear can be used to catch small mammals in their burrows. Thrust the pointed end into the hole. When you can feel the animal, twist the stick until you have snagged its fur, then carefully pull it out.

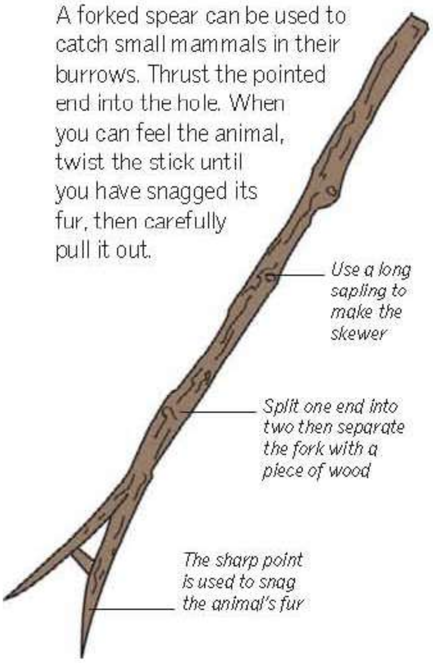
Use a long sapling to make the skewer

Split one end into two then separate the fork with a piece of wood

The sharp point is used to snag the animal's fur

SLINGSHOT

A slingshot can be used to kill small animals. Cut a strong forked branch and make a notch in each fork. Take the rubber tubing from your survival kit (see pp. 60–61) and thread it through a piece of leather or plastic. Tie the ends of the tubing to the notches. Place a pebble in the pouch, stretch back the sling, and take aim.

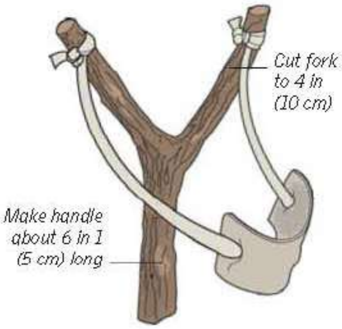
Cut fork to 4 in (10 cm)

Make handle about 6 in 1 (5 cm) long

NOOSE STICK

You can use a noose stick to lasso lizards, slow-moving mammals, or roosting birds, jerking the pole to tighten the noose around the animal's neck. Use a wire or cordage snare (see pp. 216–17). When stalking prey, move slowly and very quietly.

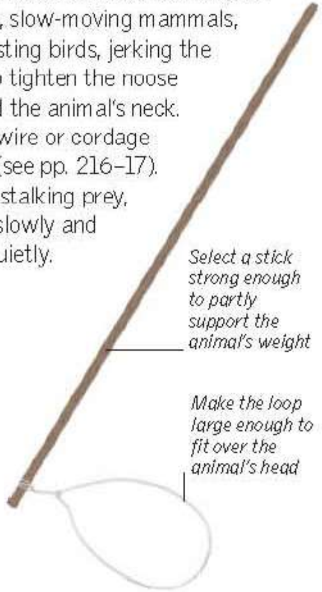
Select a stick strong enough to partly support the animal's weight

Make the loop large enough to fit over the animal's head

EASY PICKINGS

Ants are social insects, and most species are aggressive in defense of their nest. They have a stinging bite and some species will also squirt formic acid at their attacker. However, they make a nutritious meal if gathered with care. In summer months in northern temperate regions, you can eat the larvae of wood ants. And, if you use the technique shown below, you can get the ants to do all the work of collecting the larvae together for you. You'll have to vandalize the nest, but don't remove all of it.

ANT CUISINE

Ant larvae are best fried—they taste like shrimp. Adult ants also make a tasty snack—in Bogotá, Colombia, movie theaters sell roasted leaf-cutter ant abdomens instead of popcorn.

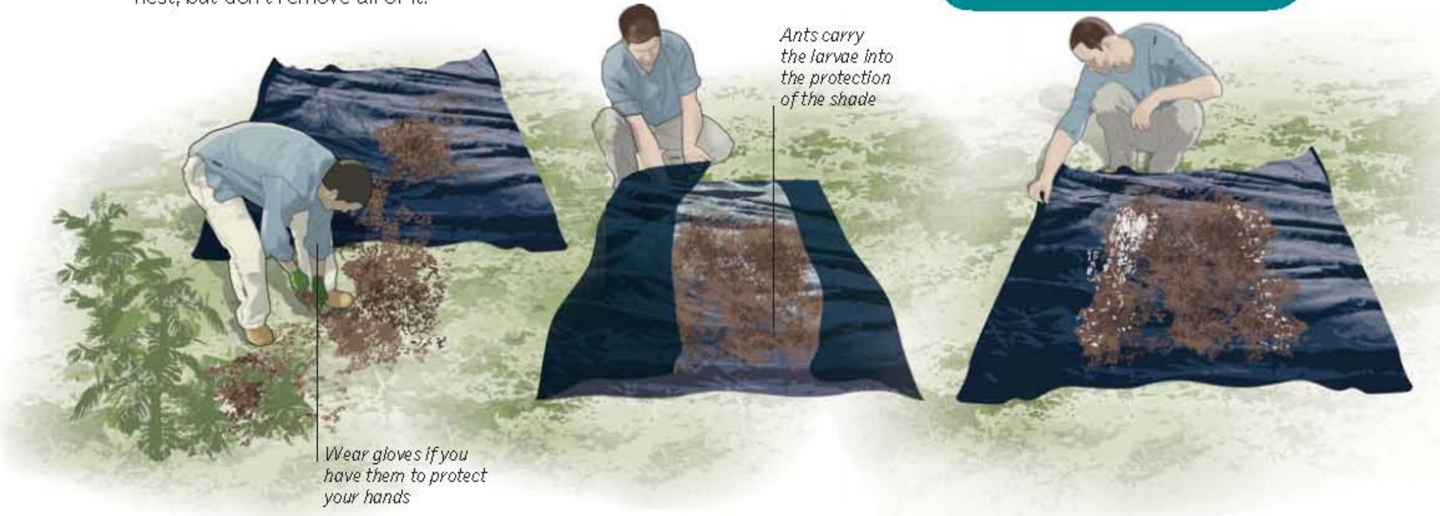
Ants carry the larvae into the protection of the shade
Wear gloves if you have them to protect your hands

1 Place a tarpaulin in a sunny patch next to the nest. Scoop the nest material, ants, and larvae into the middle of the tarpaulin.

2 Lay some sticks on the tarpaulin, close to the edge, then fold over the sides to create shade. The ants will carry the larvae into the shade.

3 After a while, throw back the folded sides of the tarpaulin and scoop up the larvae, which look like fat grains of white rice.

CATCHING INSECTS

Many flying insects can be eaten, but catching them takes a bit of ingenuity. Crawling insects, such as ants (see above) and termites, can provide a good meal if you can collect them in sufficient quantities.

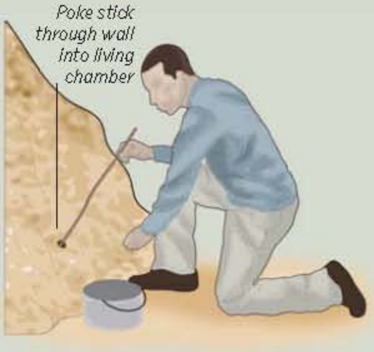
Poke stick through wall into living chamber

1 Cut a long, thin, straight stick then peel it until smooth. Push it slowly into the termite hill.

TERMITE FISHING

Termites are found in the tropics and subtropics. Some species live in vast numbers inside mounds, which they build out of mud and saliva. If a foreign object breaches the walls, the termites will attack it with their powerful jaws.

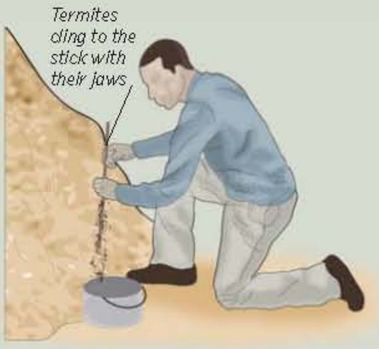
Termites cling to the stick with their jaws

2 Remove the stick and scrape off the attached termites into a container ready for frying or roasting.

NIGHT FLIERS

Nocturnal flying insects, such as moths, are attracted to light. Stretch a white sheet between two branches and peg the bottom taut above a bowl of water. Hang a flashlight behind the sheet. Insects will fly into the sheet and fall into the water.

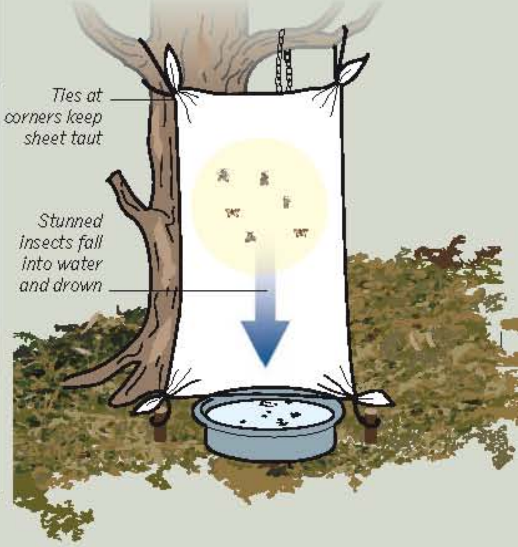
Ties at corners keep sheet taut
Stunned insects fall into water and drown

PREPARING SMALL MAMMALS

PARASITES SUCH AS FLEAS and lice will leave the body once a warm-blooded animal is dead so, if your circumstances permit, leave the carcass to cool before preparing it for cooking. In a hot climate, place the body in the shade. However, don't let it get too cold, as skin is more easily removed when a body is warm.

WARNING!
Rabbits, hares, and rodents may be infected with a bacteria called tularemia, which can be fatal to humans. Do not touch these animals with your bare hands if you have a scratch on your skin. If you have no gloves, cover your hands in soap lather before handling the animal, and wash your hands when you have finished. The germ is destroyed by heat, so cook the meat well. Myxamatosis, a viral disease that affects rabbits' mucous glands, is not harmful to humans.

RABBIT-SIZED MAMMALS

The method shown here applies to all furred mammals approximately the same size as a rabbit. Before gutting the animal, remove any urine by holding the body by the forelegs and progressively squeezing down from the chest toward the bowels, making sure that you direct the spray away from your own body. You'll need to remove any scent glands, which are usually inside the forelegs or around the anus, in case the musk taints the meat (take particular care when handling a skunk).

GUTTING

No part of a mammal carcass should be wasted. If your camp is near water, keep the guts for fish bait. The offal should be removed carefully and eaten if healthy—discard any organ that's pale or spotty. Liver is rich in essential vitamins and minerals and needs little cooking. It should be eaten as fresh as possible (removing the gall bladder from the middle first). Kidneys, which in most mammals are surrounded by fat, are also a valuable source of nutrition.

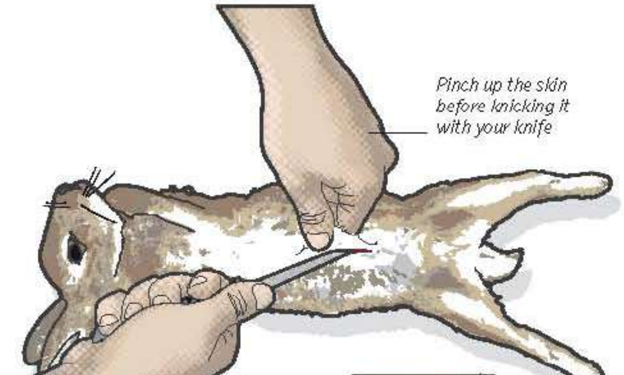

Pinch up the skin before knicking it with your knife

1 Place the animal on its back on clean ground, such as a bed of pine boughs, with its head pointing toward you. Cut a small hole in its belly with the point of your knife, taking care not to pierce the guts.

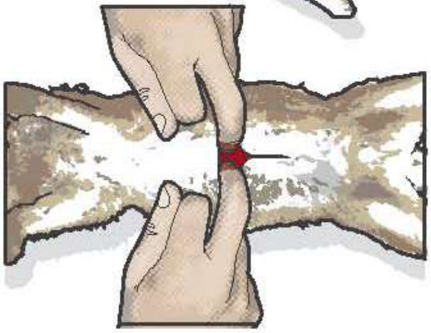

2 Pull the skin apart at the cut and insert a finger from each hand into the opening.

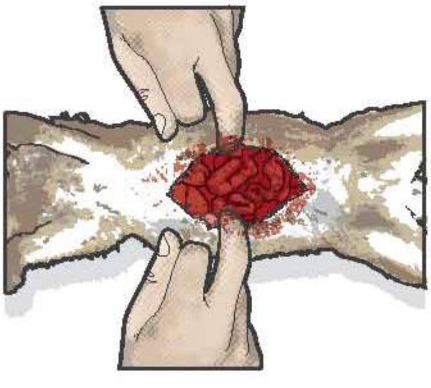

3 Prise open the belly to expose the guts; remove them along with the heart, liver, and kidneys. Wash your hands before moving on to skinning.

WARNING!
Rabbits make an easy meal, but their flesh lacks fat and vitamins so, if your diet consists entirely of rabbit, your body will use its own vitamins and minerals to digest the meat. These are then lost from your body in your feces and, if not replaced, you'll get diarrhea and so become hungry and weak. Eating more rabbit makes the condition worse and you will eventually die of starvation. You must supplement your diet with some vegetation and fatty food.

SKINNING

If you wish to keep the skin for use later (to make a pair of mittens, for example), you should remove it as carefully as possible. You'll also need to cure it; stretch the skin as tight as possible and leave it in the sun or hang it close to the heat of a fire to dry. Rubbing wood ash into the skin will help to speed up the process.

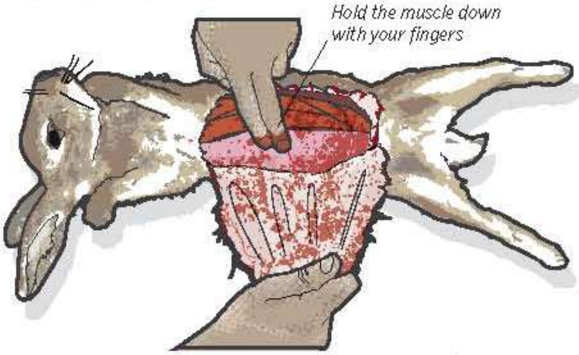
Hold the muscle down with your fingers

1 Beginning at the belly, separate the skin from the muscle surrounding the gut cavity. You'll find that it pulls away quite easily. When you have reached the back on one side, repeat the process on the other.

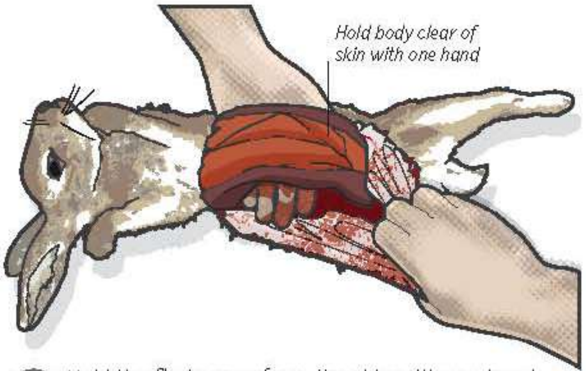
Hold body clear of skin with one hand

2 Hold the flesh away from the skin with one hand and pull the skin over the rear legs, one by one. The hindquarters are now free of skin and fur, except for the tail, which should be cut off and the scent glands removed.

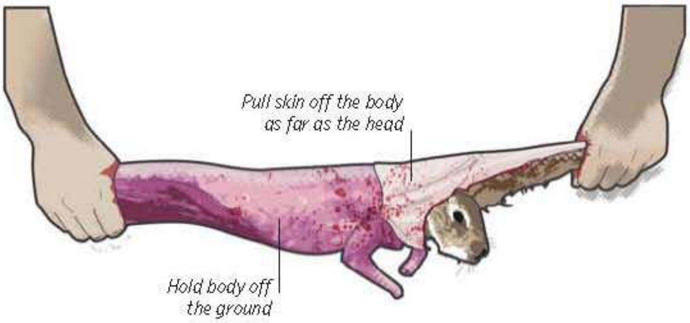
Pull skin off the body as far as the head

Hold body off the ground

3 Hold the animal by the rear legs and pull the skin forward, easing out both forelegs. Pull the skin over the neck and then cut off the head. Wash the body to remove any traces of fur.

GLOVE SKINNING

The preferred survival method of skinning and gutting a rabbit-sized mammal is to suspend it by its rear legs from a strong branch as this keeps the animal off the dirty ground.

- If you have no knife to hand, snap off the lower part of a foreleg and use the sharp edge of the broken bone to cut the skin.
- Gut the animal, then cut the skin around all four paws and between the rear legs.
- Tug the skin down toward the head.
- If you wish, pull the skin over the head. Otherwise just cut the head off.

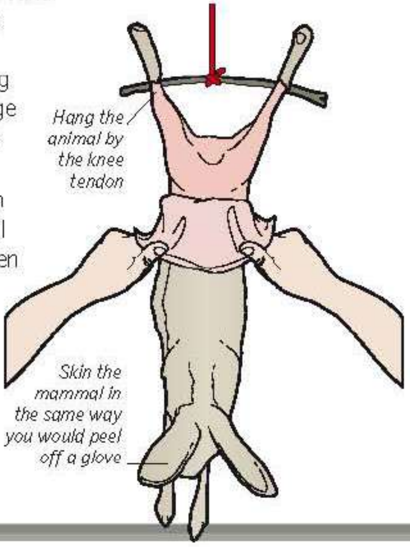
Hang the animal by the knee tendon

Skin the mammal in the same way you would peel off a glove

HOW TO SKIN A SQUIRREL

Squirrel flesh is tender and tasty. Once you have skinned the squirrel, you can simply skewer it on a spit and roast it whole over a fire. Alternatively, you can cut the meat into pieces and make a stew. A large squirrel will feed two people.

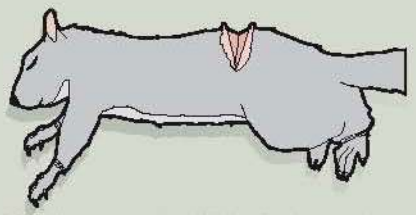

1 Gut the squirrel following the method shown opposite. Cut the skin around the paws, then cut through the skin across the middle of its back.

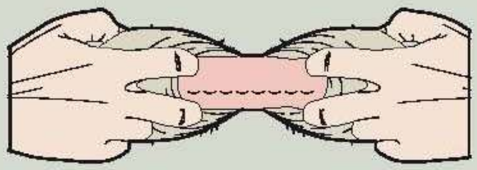

2 Insert two fingers under the skin on either side of the cut and pull the two pieces apart and off the body. Wash off any traces of fur.

PREPARING LARGE MAMMALS

ALL LARGE MAMMALS are edible, but you must not eat the livers of polar bears or bearded seals as they contain toxic levels of vitamin A. Wild game is sometimes infected with roundworm larvae and, if you eat the meat raw or undercooked, it will infect you with parasites, so always cook all wild game thoroughly.

WARNING!
Approach all game with caution, as the animal may still be alive and most wounded animals are extremely dangerous. To check, touch its eye with the tip of a long stick (or the muzzle of a gun). Even if the animal is unconscious, it will blink if it's alive. Always skin, gut, and butcher game at the site of the kill. You don't want the smell of fresh blood to attract predators or scavengers to your camp.

PREPARING THE KILL

Lifting large animals, such as deer, uses energy so, if you're alone, prepare the carcass on the ground. However, if you're in a group and circumstances permit, it's best to suspend the animal by its hocks (the tarsal joints of its hind legs) from a branch, as it will bleed better and will be easier to skin and gut. If you can't find a suitable tree from which to hang it, build a frame.

BLEEDING

Bleeding is essential to preserve the meat and stops it from tasting too gamy. It also helps to cool the carcass. Slit the animal's throat from ear to ear and let the blood drain out. Blood is rich in vitamins and minerals, including salt, so, if you're able to, collect it in a container to use later in a stew. Cover the container to protect the blood from flies and to keep it cool.

SKINNING

If you wish to preserve the hide, it's better to take the skin off before gutting. Cut the belly skin from throat to tail, cutting around the genitals. Cut along each leg from above the foot to the belly. Pull the hide off the carcass, severing connective tissue as necessary. If you're skinning on the ground and you don't want to keep the hide, you can use it to protect the meat—remove it fully only after jointing is complete.

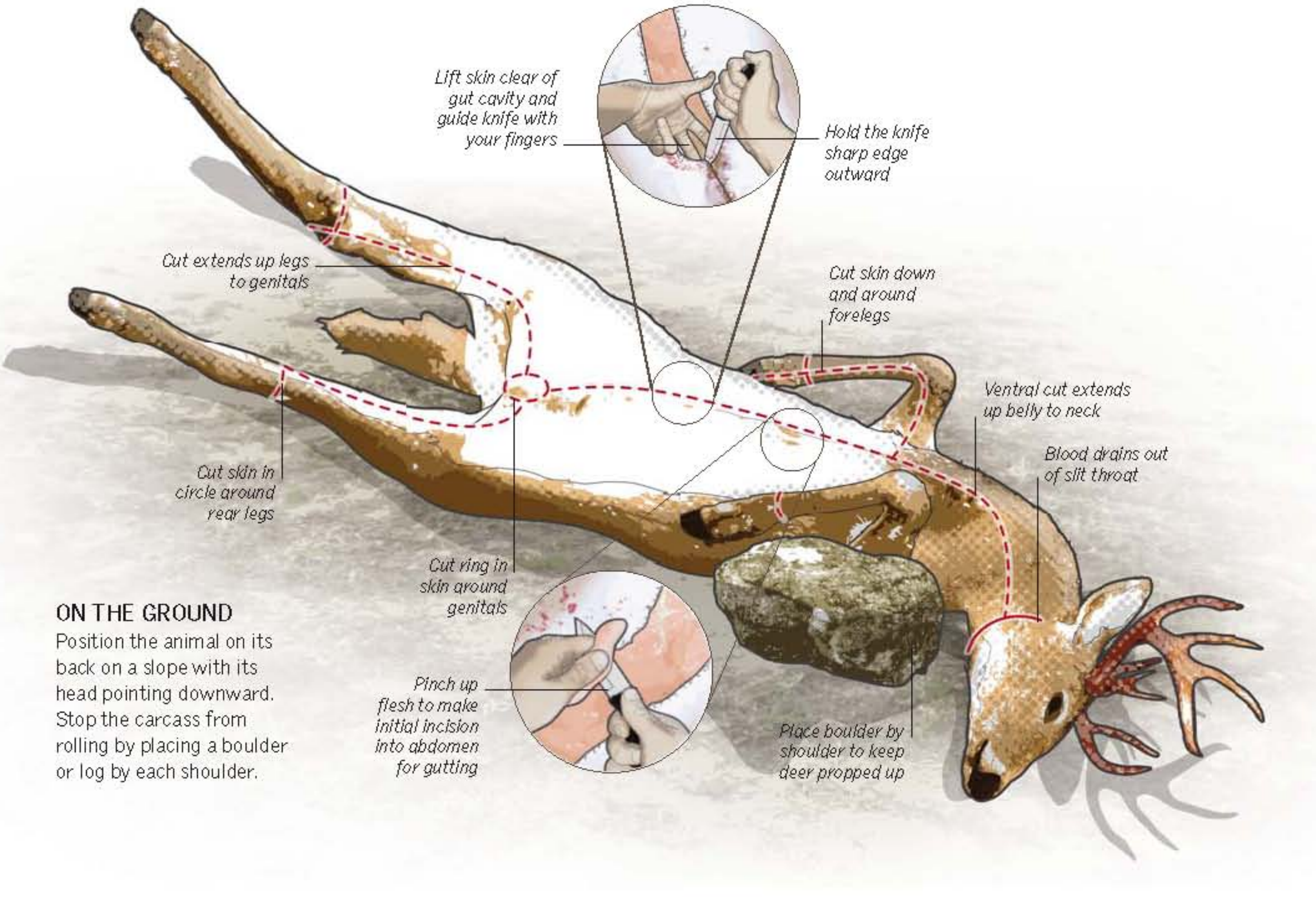
Lift skin clear of gut cavity and guide knife with your fingers
Hold the knife sharp edge outward
Cut extends up legs to genitals
Cut skin down and around forelegs
Ventral cut extends up belly to neck
Blood drains out of slit throat
Cut skin in circle around rear legs
Cut ring in skin around genitals
Pinch up flesh to make initial incision into abdomen for gutting
Place boulder by shoulder to keep deer propped up

ON THE GROUND

Position the animal on its back on a slope with its head pointing downward. Stop the carcass from rolling by placing a boulder or log by each shoulder.

GUTTING

The method shown here is for gutting a carcass on its back. When finished, inspect the heart, liver, and kidneys for signs of worms or other parasites and, if the organs are healthy, keep them to eat. If the liver is spotted, a sign of disease, discard all internal organs and boil the meat.

1 To avoid piercing the internal organs, pinch up the abdomen near the breastbone and make an incision big enough to insert two fingers.

- Use your fingers to guide the knife, cutting toward, then around, the anus with the sharp blade-edge upward.

2 Using a saw, if you have one, cut through the outer muscle and the breastbone to open up the chest cavity,

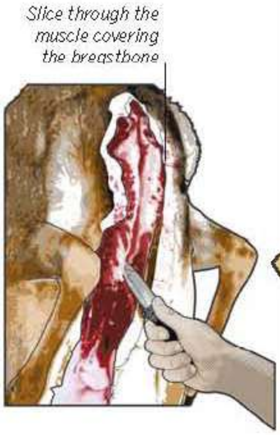
Slice through the muscle covering the breastbone

3 Using the knife, slice through the diaphragm muscle, which separates the chest cavity from the gut cavity, getting as close to the spine as you can.

- Remove the liver, taking care not to cut the gall bladder in the center.

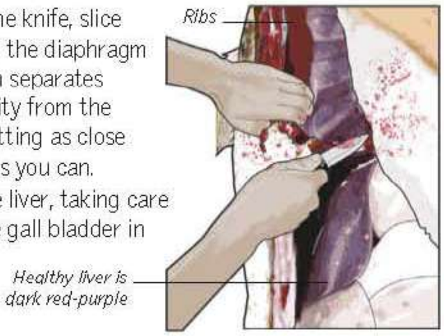
Ribs
Healthy liver is dark red-purple

4 Reach up through the chest cavity and cut the windpipe and esophagus. Hold them with one hand and pull out all the internal organs as one unit.

- Check that the anus is clear of feces, pushing a hand through if you can't see daylight.

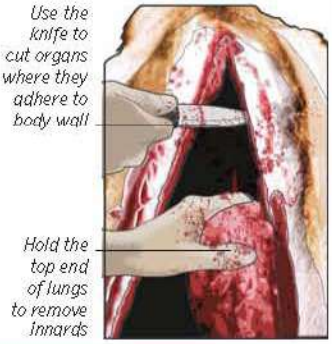
Use the knife to cut organs where they adhere to body wall
Hold the top end of lungs to remove innards

KEEP THE GUTS

If you're able to carry them, take the guts, reproductive organs, and glands back to your camp in a sealed container. You can use them as bait in traplines or for fishing. Keep the fat that surrounds the intestines for cooking.

JOINTING

Large game should be cut into manageable pieces that can be carried back to camp. Boning the meat helps to reduce the weight. Bag or wrap the meat as you remove it to keep it clean and free of flies, and keep it in the shade in hot climates. How a carcass is divided partly depends on the species, but the hindquarters contain the steaks and the best cuts.

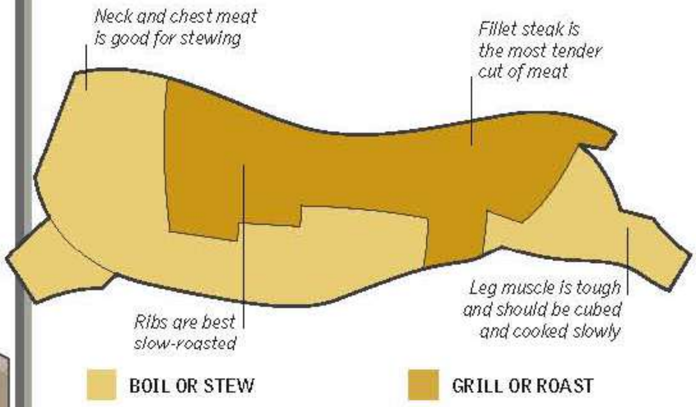
Neck and chest meat is good for stewing
Fillet steak is the most tender cut of meat
Ribs are best slow-roasted
Leg muscle is tough and should be cubed and cooked slowly
BOIL OR STEW
GRILL OR ROAST

USES FOR HIDES

Keep the skin if you can. When dried, it's light and can be used as a blanket, or it can be made into an item of clothing. Hide is one of the best materials for lashes. Sinew also makes good cordage (see p. 140).

TANNING THE HIDE

The best way to clean a skin is to stretch it on a frame and then carefully scrape off the fat and any remaining scraps of flesh, using a flint or a piece of bone. Every mammal, with the exception of buffalo, has enough brain matter for use in tanning its own hide. Mash the brains well in warm water, then apply the mixture to the hairless side of the skin and leave it to dry for 24 hours.

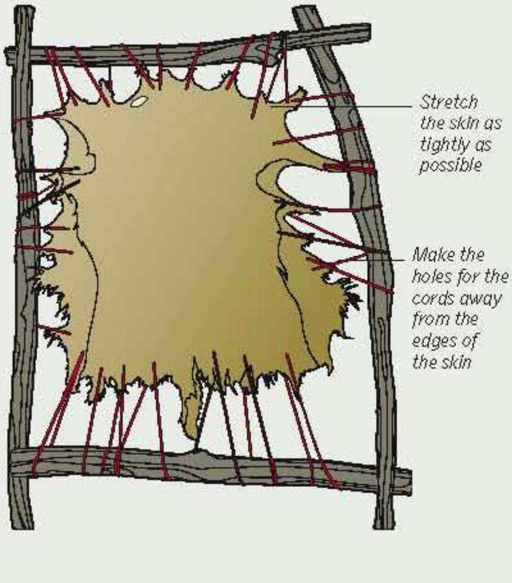
Stretch the skin as tightly as possible
Make the holes for the cords away from the edges of the skin

PREPARING OTHER ANIMALS

ALL TYPES OF ANIMAL should be considered when searching for your next meal. Invertebrates are far more plentiful and widespread than other animals, and take less energy to gather and prepare. Reptiles and amphibians can also provide a vital source of nutrition.

EATING FOR SURVIVAL

The thought of eating grubs or grasshoppers might not appeal, but experimentation is vital in a survival situation—don't forget that in some cultures such creatures are highly prized, if prepared properly. However, some animals are less palatable than others, in which case you should chop them finely and add to a stew.

SLUGS, SNAILS, AND WORMS

Slugs should be avoided as they often feed on poisonous fungi. To prepare snails, either starve them or feed them safe food such as wild garlic for 24 hours to purge their guts before cooking. Either boil them in water or bake them in their shells in hot embers until the juices bubble over. Avoid all marine snails and any terrestrial snail with a brightly colored shell as they may be toxic. All earthworms are edible. Place them in salty water until they are purged, then boil, though they can be eaten raw if necessary.

EDIBLE INSECTS

Insects are mostly protein and make good emergency food. Avoid any hairy or brightly colored insects or those that emit a foul smell (including larvae). As a general rule, avoid adults that sting or bite, but ants and honey bees can be eaten if collected carefully. Most insects, except those with a hard carapace, like beetles, can be eaten raw.

HOPPING INSECTS

Grasshoppers and crickets have large leg muscles, and most are quite tasty when cooked (avoid brightly colored ones). Remove the antennas, wings, and leg spurs, and roast to kill any parasites.

WITCHETTY GRUBS

Witchetty grubs are a highly nutritious Australian bush delicacy. They can be eaten raw but, if roasted quickly in hot ashes, they taste of scrambled eggs. Treat palm grubs in the same way.

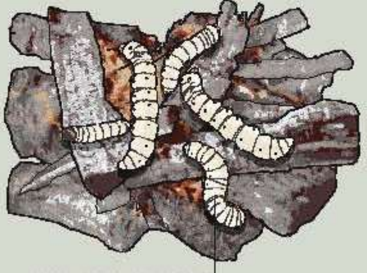

Large, white grubs high in protein and calcium

AMPHIBIANS AND REPTILES

All frogs in the *Rana* genus (see pp. 292–93) are edible. Skin and then boil or roast them. The best, most meaty part is the hind legs. Don't handle or eat toads or brightly colored tropical frogs as many have highly toxic skin secretions. Lizards and snakes are a good source of protein. They must be skinned and gutted, then either roasted on a stick if small, or cut into small pieces and boiled. All snake flesh is edible—and tastes like chicken—apart from the head, which must be removed.

1. Hold the dead snake firmly behind the head and cut it off about 6 in (15 cm) down the body.

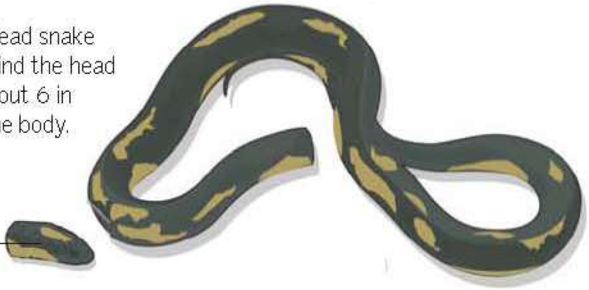

Bury the head if the snake is venomous

2. Slit the belly with a sharp knife and remove the entrails and other organs. Either throw them away or keep as bait.

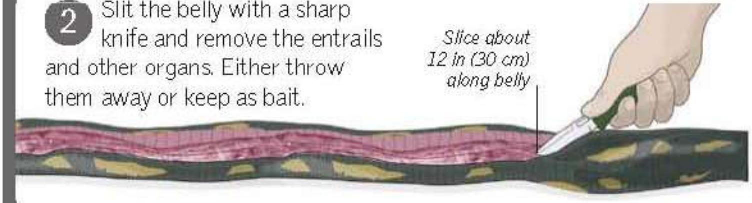

Slice about 12 in (30 cm) along belly

3. Peel the skin back, then grasp the body in one hand and the skin in the other and pull it off.

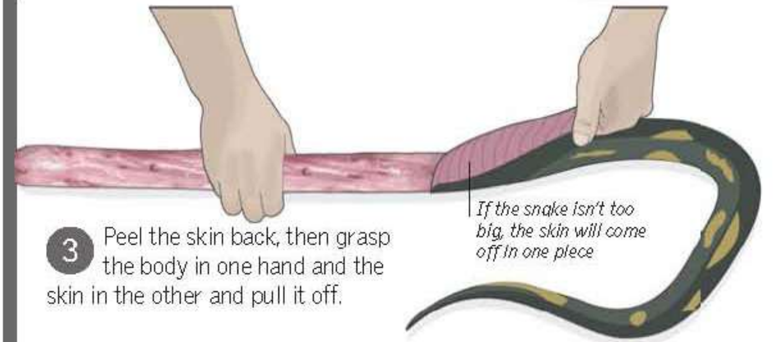

If the snake isn't too big, the skin will come off in one piece

PREPARING A CRAB

Avoid their strong claws (tie them up if you can). To kill a crab, plunge it in boiling water or stab it through the eye socket or the orifice under the shell flap on the body. Cook the crab for about 15 minutes.

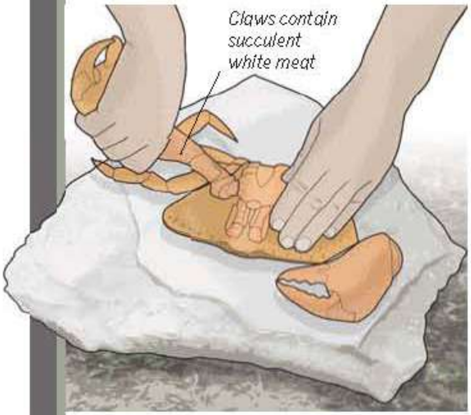
Claws contain succulent white meat

① Place the cooked crab on its back, then twist off its legs and the claws. They all contain edible meat, so crack them open with a rock.

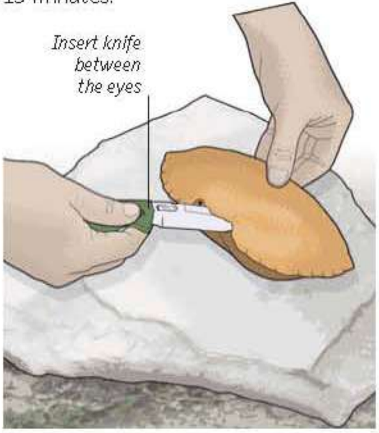
Insert knife between the eyes

② Open the shell by inserting the point of your knife between the two halves and twisting it. Lift back the lower shell.

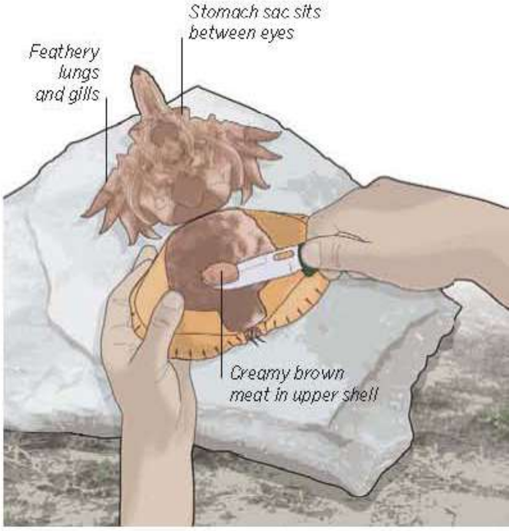
Stomach sac sits between eyes
Feathery lungs and gills
Creamy brown meat in upper shell

③ Remove and discard any green matter and the lungs, gills, and stomach as they're all poisonous. Scoop the meat out of the shell.

SHELLFISH

Shellfish can be found in streams and on lake and sea shores. Don't collect marine shellfish that are not covered at high tide or are near any source of pollution. Shellfish must be alive when you collect them—bivalves, such as mussels, will close their shell if you tap them, and univalves, such as limpets, will cling tightly to their rock—and you must cook and eat them immediately.

WARNING!

Tropical species of mussel should be avoided in the summer as they may be poisonous. The black mussel, which is found on Arctic and sub-Arctic coasts, may be poisonous at any time of the year. The symptoms of mussel poisoning are acute nausea, vomiting, diarrhea, abdominal pain, and fever.

HOW TO SHUCK AN OYSTER

Oysters are rich in vitamins and minerals. To eat an oyster raw, first open its shell by inserting a blunt knife into the hinge at the thicker, more pointed end of the oyster and twisting. However, in a survival situation, play safe and boil oysters in their shells for five minutes after their shells open. Do not eat any whose shells do not open during cooking.

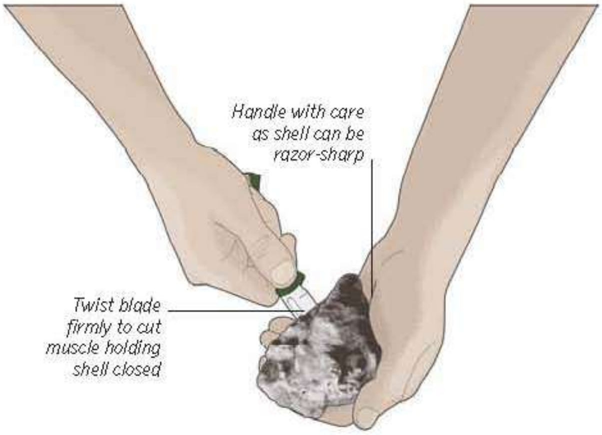
Handle with care as shell can be razor-sharp
Twist blade firmly to cut muscle holding shell closed

SHRIMPS, CRABS, AND LOBSTERS

Small crustaceans, such as shrimps and crawfish, should be cooked in boiling water for five minutes. Larger crustaceans, such as some crabs and lobsters, should be boiled for up to 20 minutes. A tasty alternative to boiling shrimps or prawns is to roast them on skewers made of green wood, placed in the embers of a fire.

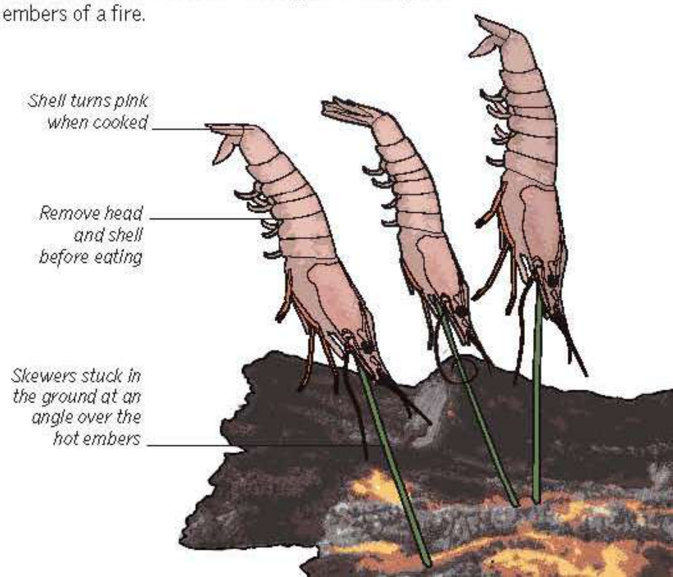
Shell turns pink when cooked
Remove head and shell before eating
Skewers stuck in the ground at an angle over the hot embers

CATCHING BIRDS

BIRDS CAN BE TRAPPED in a variety of ways—in snares, nets, tunnels, or cages—and they can be hunted with noose sticks or any projectile weapon. The trick is to observe their behavior and to match your technique—and bait—to the bird. You need to discover their regular roosts, feeding spots, and flight paths.

WARNING!

Eggs are easy to collect, particularly from ground-nesters, although gulls, for example, will attack you if you try. However, taking eggs from nests is illegal in many countries and should only be done if absolutely necessary.

FIGURE-4 CAGE TRAP

This trap enables you to catch ground-feeding birds (or rabbit-sized mammals) without killing or wounding them, so you can choose when you want to kill them and eat the meat fresh. The "Figure 4" is the trigger that's used to drop the cage over the animal.

MAKING THE TRIGGER

To construct the Figure 4 trigger you need to make three notched sticks (bait stick, upright stick, and release stick) and then link them together as shown below.

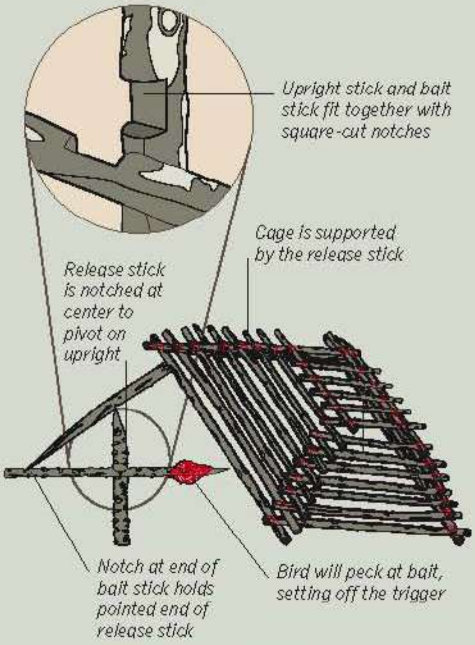
Upright stick and bait stick fit together with square-cut notches
Cage is supported by the release stick
Release stick is notched at center to pivot on upright
Notch at end of bait stick holds pointed end of release stick
Bird will peck at bait, setting off the trigger

MAKING THE CAGE

Create a pyramid of sticks by placing progressively shorter sticks across longer sticks and lashing them together at the ends. Balance the cage over the bait.

ON THE GROUND

Gamebirds, such as quail, and inquisitive scavengers, such as crows, can be lured into traps if you use the right bait. Migratory waterfowl, such as ducks and geese, molt in the late summer, which means they are easier to catch as they can't fly away.

WALK-IN TUNNEL TRAP

A tunnel trap can be used to catch gamebirds, which have stiff feathers that lie in one direction and bend only with difficulty. As a bird tries to retreat from the trap, its feathers will become wedged in the tunnel walls.

- Dig a near-horizontal, funnel-shaped tunnel close to the ground.
- Lay a trail of bait (seed or berries) leading to the rear of the tunnel.
- As it eats the bait, the bird will move deeper into the tunnel and will be unable to back out.

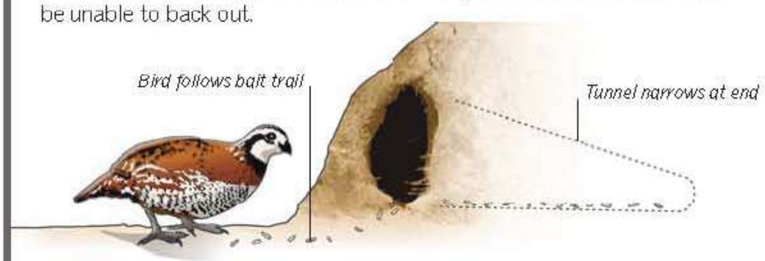
Bird follows bait trail
Tunnel narrows at end

ON THE WING

A fine net stretched between two trees across a flight path is an effective method of catching birds in flight. Throwing a bola requires more skill, and works by entangling the bird in the spinning ropes.

MAKING A BOLA

To make a bola, use an overhand knot (see p. 143) to tie three 3-ft (1-m) lengths of cord together about 3 in (8 cm) from one end. Find three rocks weighing about 7 oz (200 g), wrap each one in a piece of cloth, and tie it up with the free end of a cord.

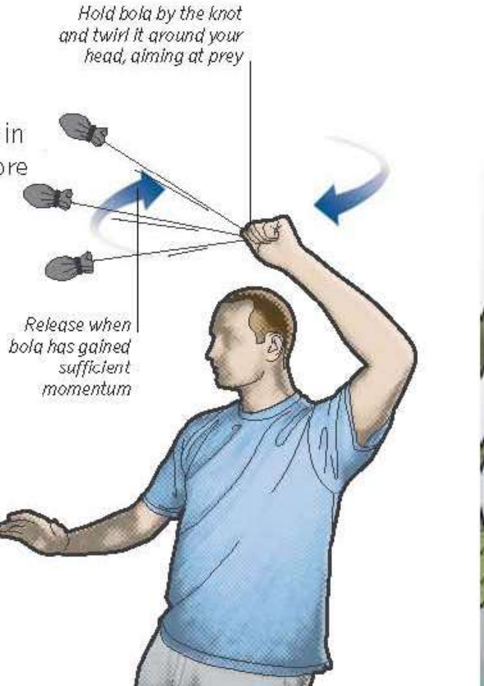
Hold bola by the knot and twirl it around your head, aiming at prey
Release when bola has gained sufficient momentum

ON THE NEST

Birds are creatures of habit and will usually roost in the same place every night, such as in the branches of a tree as protection against ground predators. They rarely move once it's dark. Many birds are more accessible when breeding as, once you've located the nest, you'll always know where to find them. However, you should only take nesting birds in a genuine survival situation.

MAKING A SNARE STICK

A snare stick is used to trap birds roosting in trees. Remove a bird once it has been snared, as its squawking and fluttering will alert other birds to the danger and scare them off.

- Make several single-strand snares (see pp. 216–27) with a loop diameter of about 1–2 in (2.5–5 cm), depending on the size of the bird.
- Place them close together along a stick, making notches in the stick to hold them in position with the loops uppermost.
- Tie the stick on top of a branch where you have seen birds roosting.

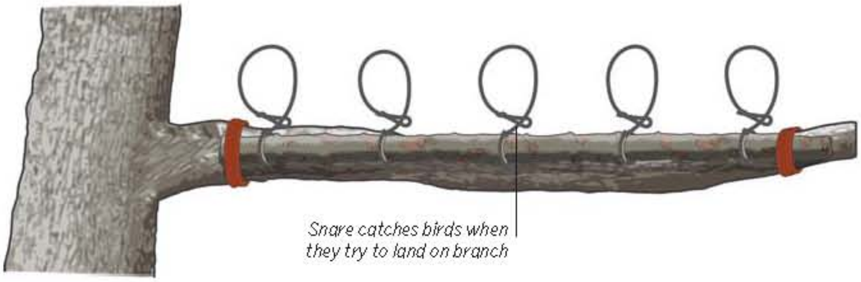
Snare catches birds when they try to land on branch

PHEASANT POLE

Pheasants roost in trees at night, usually returning to the same branch, and they won't leave once it's dark. To snare a sleeping pheasant, raise the pole up in front of it. The bird will wake and peck at the foil. As its head goes through the loop, pull the cordage and the snare will close around the bird's neck.

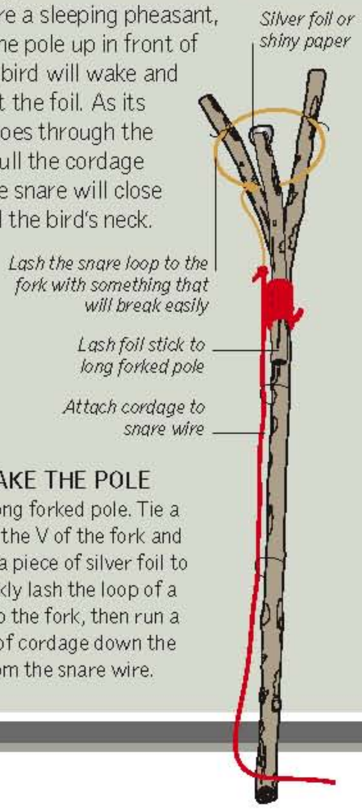
Silver foil or shiny paper
Lash the snare loop to the fork with something that will break easily
Lash foil stick to long forked pole
Attach cordage to snare wire

TO MAKE THE POLE

Cut a long forked pole. Tie a stick in the V of the fork and attach a piece of silver foil to it. Weakly lash the loop of a snare to the fork, then run a length of cordage down the pole from the snare wire.

BOUGH FENCE TRAP

This method is used to trap ducks, which often go ashore during the day to sun themselves, or at night to roost. They tend to favor small islands as protection against predators.

- Cut several saplings and make double-stranded snares (see p. 216).
- Place the boughs a few feet from the water's edge, and bend them into a series of overlapping hoops to form a fence around the island.
- **If you only have a few snares, place them over the ducks' inward and outward tracks. Placing logs between the snares will then force the birds to pass through the hoops in order to get onto the island.**
- **Suspend a snare from the center of each bough arch.**

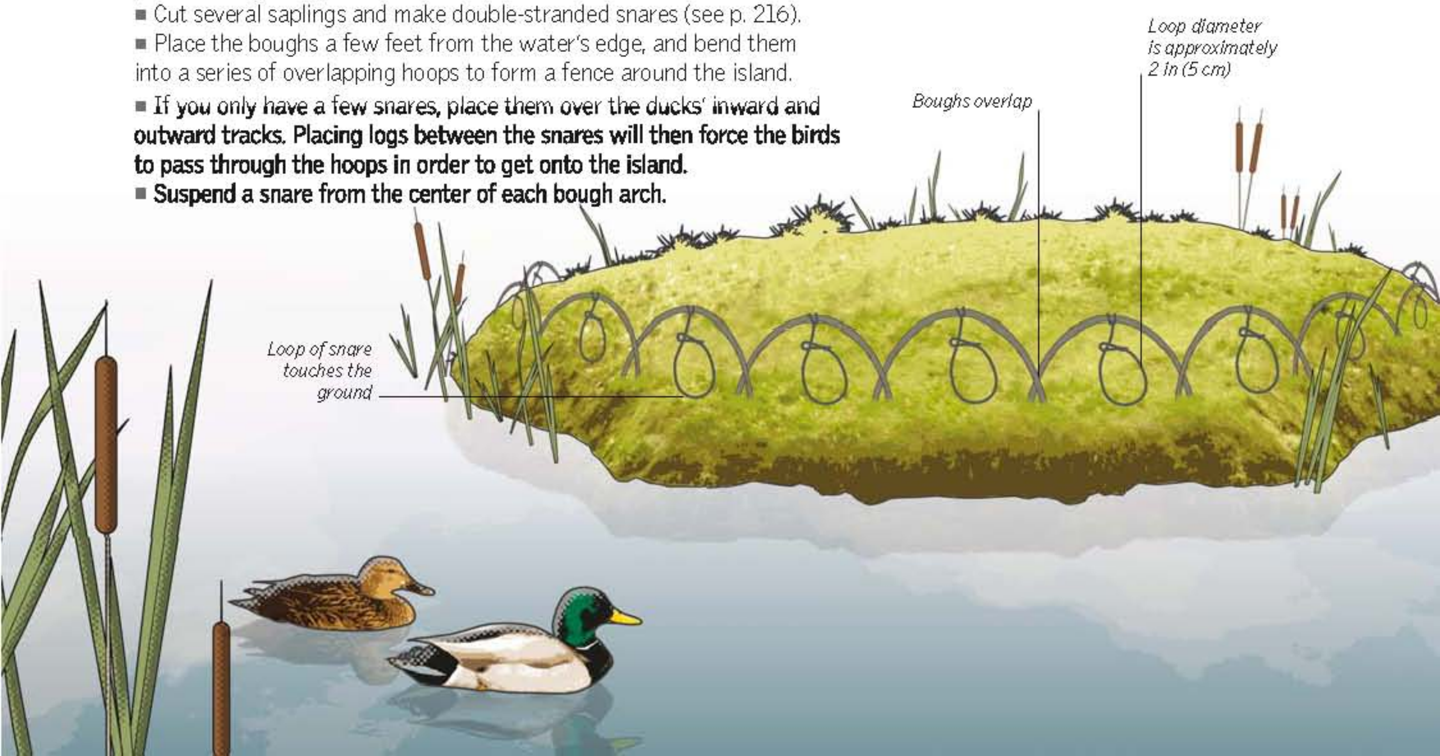
Loop diameter is approximately 2 in (5 cm)
Boughs overlap
Loop of snare touches the ground

PREPARING BIRDS

GAMEBIRDS AND WATERFOWL are the most commonly eaten wild birds, but all birds are edible if prepared correctly. Skinning a bird is the quickest way to get at the meat, but this removes the nutritional value of the skin, so always pluck a bird if you have the time to do so.

WARNING!
Handle scavengers and carrion eaters, such as crows, buzzards, and vultures, as little as possible, because they are more likely to be infested with lice and ticks, and are prone to infection.

- The meat of such birds must be boiled for at least 30 minutes to kill any infectious organisms. This will also tenderize the meat.
- Always wash your hands after handling any bird.

GETTING STARTED

If the bird hasn't already been killed by the method of capture you'll need to dispatch it—for example, by stretching its neck and cutting its throat (for the best way to kill gamebirds, see below). Whichever method you use to kill the bird, you must bleed it before plucking, but make sure you pluck it while the body is still warm.

KILLING GAMEBIRDS

To quickly and humanely dispatch a gamebird, such as a pheasant, fold its wings into the body and hold it under your arm. Cover its head with your jacket to calm it down (this also reduces the chances of you being injured by its claws or beak). Still holding the bird firmly, take a 2 ft (60 cm) long, thick stick and place it on the ground. Put the bird's head underneath the stick, place your feet on the stick either side of the head, and pull the bird up sharply by the legs—the bird's head will come off.

POACHER'S METHOD
The poacher's way to quickly get the meat from a dead pheasant that hasn't been decapitated is to place it face down on the ground with its head toward you, put a foot on each wing, take hold of the legs, and pull up sharply. This motion will tear the legs and breast meat away from the rest of the bird. Give it a sharp flick to detach the guts. In a survival situation, this is a good way to get meat off the bird if you don't have a knife or razor.

PLUCKING A BIRD

It's not essential, but scalding a bird in a bucket of hot (not boiling) water for a couple of minutes usually helps to loosen the feathers; the exceptions are waterfowl and seabirds—their feathers will tighten instead. Be careful not to overscald the bird or the skin will start to cook.

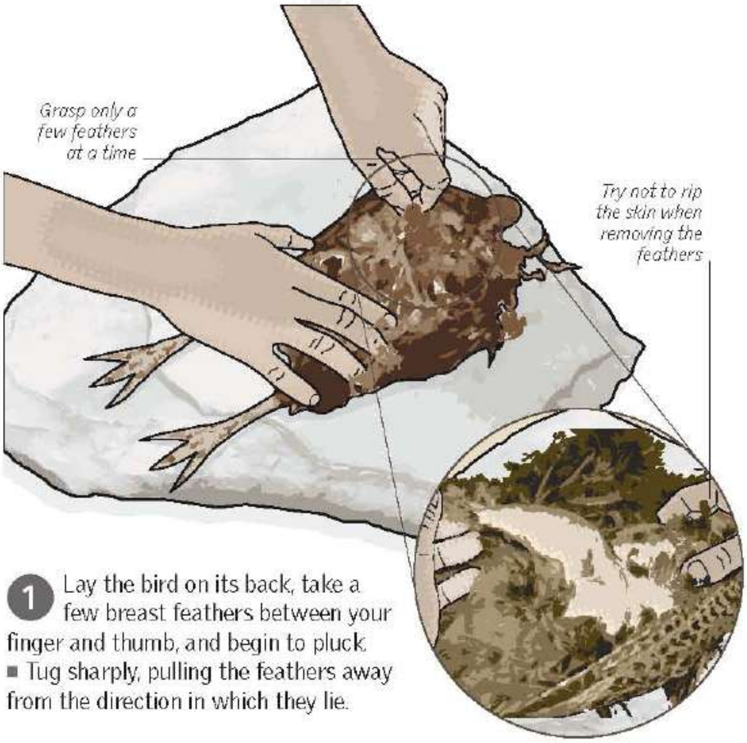
Grasp only a few feathers at a time
Try not to rip the skin when removing the feathers

1. Lay the bird on its back, take a few breast feathers between your finger and thumb, and begin to pluck.
   - Tug sharply, pulling the feathers away from the direction in which they lie.

2. Don't try to pluck too many feathers at once or you'll tear the skin. Work your way steadily around the front and back until you've plucked the whole body.
   - Keep the feathers (unless the bird is a scavenger, see box, above) to use for tinder, insulation, or fishing lures.

OPTIONAL EXTRAS

Keep the crop (esophagus) contents of a gamebird—you can use the seeds and berries as bait to catch other birds. If the bird is female and has eggs in the oviduct, keep them and eat them. You should also eat the heart and the liver if they are in good condition; discard them if they look old or spotty, or show signs of parasites.

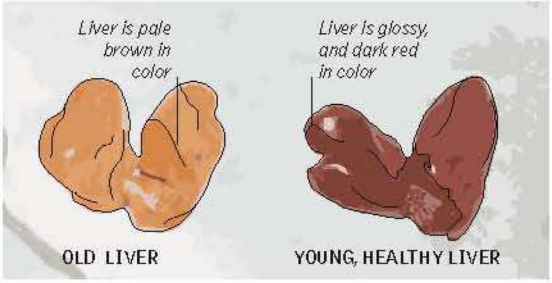
Liver is pale brown in color
Liver is glossy, and dark red in color
OLD LIVER
YOUNG, HEALTHY LIVER

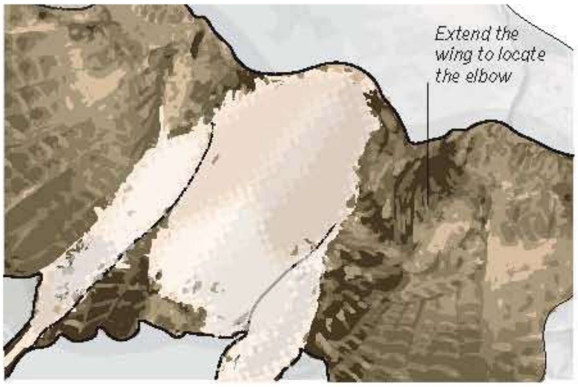
Extend the wing to locate the elbow

3 Pluck the legs in the same way, then turn to the wings. Locate the elbows and cut off the lower wings at this joint. Pluck the upper wings.

- If the bird's head is still attached, cut it off as close as possible to the body.

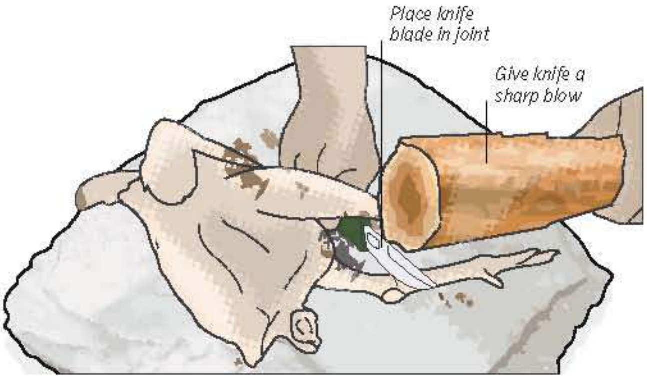
Place knife blade in joint
Give knife a sharp blow

4 To remove the feet, find the ankle, insert the blade of a sharp knife into the joint, and press down hard.

- If you're dealing with a large bird, such as a turkey, hit the back of the knife with a thick stick for extra force.

DRAWING METHODS

The usual way to remove the offal from a bird (known as "drawing") is to take the crop (esophagus) out of the neck end and the rest of the offal out of the rear end. The survival method is to make a cut from the throat to the tail with a sharp knife, reach in, and pull out all the offal (see below). Should you find yourself without a knife, you can still butcher a bird by pulling the skin apart with your fingers and working your way up the breast to the neck. The offal can be removed by ripping open the skin over the belly.

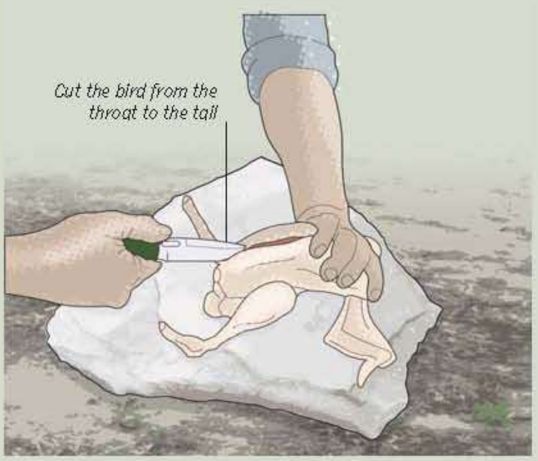
Cut the bird from the throat to the tail

1 Hold the bird in one hand, breast upward and with its rear end toward you. Insert the point of a sharp knife into the throat.

- Make a single incision down toward the tail, being careful to not pierce the guts.

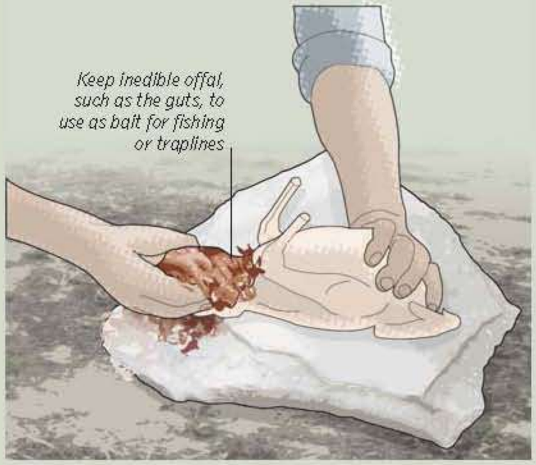
Keep inedible offal, such as the guts, to use as bait for fishing or traplines

2 Reach into the gut cavity and pull out the offal, taking care not to break any eggs. Keep the liver and heart if healthy (see above left).

- Wash the bird clean with cold water and wash your hands thoroughly before further handling.

6

IN AN EMERGENCY

IN AN EMERGENCY

You're only a survivor when you have been rescued. This means that you must be able to either get yourself out of the predicament you are in (self-rescue) or be rescued from that situation by others. Sometimes you will be able to choose whether to undertake self-rescue; at other times the decision will be out of your hands. There may be several reasons why you cannot achieve self-rescue: you could be utterly lost; local conditions such as flooding or bad weather could trap you; or injury to yourself or a group member could make movement impossible. In this situation, the onus is on you to attract attention. You must be able to make contact with rescuers, using location aids you have with you or those you can improvise. Never delay any form of rescue because of the embarrassment factor—the only important factor is the outcome, and I would always rather be embarrassed and alive than eventually found dead!

Preparation is key—informing people of your intentions and timeframes will at least have someone wondering why you are not back yet. Equally, taking the best

In this chapter YOU WILL DISCOVER...

- how to **recognize** the **enemies** of **survival**...
- why **LEOs** are so **important**...
- when to **set fire** to a **log cabin**...
- that **tinsel** isn't just for **Christmas**...
- what to **do** if a **hippo yawns**...
- that you **should go** with the **flow** in an **avalanche**...
- how to **inflate** your **pants**...

location aids for your environment and knowing how best to use them will increase the chances of you being found.

In many survival situations, a major decision will be whether you remain where you are or move to a location that offers a better chance of survival, rescue, or both. There are numerous factors that will dictate your best option but, in general, it's always best to stay where you are. It's all too easy to make a rash decision and attempt to walk out of a situation only to put yourself in even greater danger.

A location aid can make the difference between life and death. A comparatively recent invention is the multi-tasking Skystreme device.

SKYSTREME is a silver, inflatable foil kite, which weighs just 1½ oz (43 g) and packs down to a small and convenient size. It can perform four tasks in a survival situation:

Location aid The kite is orally inflated and its wedge shape means it can lift off from the ground unaided in 4 mph (6 kph) winds.

- The 165 ft (50 m) of line attached to the kite (this much cordage is invaluable in itself) allows it to rise above tree levels and reflect sunlight. It can be seen by the naked eye at a distance of 2 miles (3 km).
- At night, you can hang a small flashlight or lightstick underneath the kite to aid visibility.
- The metallic surface reflects radar and can be detected by aircraft at a distance of 10 miles (17 km). A UK Royal Marine on a polar crossing flew his Skystreme behind his sled so that support aircraft could find and track him.

Emergency first aid splint The kite can be inflated around a broken or sprained limb.

Thermal vest The kite can be inflated and put inside clothing to act as a body warmer.

Water storage Instead of being inflated with air, the kite can be used to carry and store water.

“You are not a **survivor** until you have **been rescued**”

ASSESSING YOUR SITUATION

ONCE YOU'RE OUT of immediate danger, assess your situation and plan accordingly. At this initial stage, it's crucial to think clearly—the decisions you make now could mean the difference between life and death. In most cases, remaining where you are is the preferred option, but no two situations are ever the same: the circumstances, environment, conditions, and you—the individual—all have a major impact on what can and cannot be achieved.

THE STRATEGY FOR STAYING ALIVE

In an emergency situation, think of the four priorities of survival: protection, location, water, and food. Your situation will determine which is the most important. In most cases, as long as you're in no further danger from injury or the elements, you should focus your efforts on establishing a safe location and getting yourself rescued. The strategy for staying alive, known by the acronym SURVIVAL (see below), gives you a framework to use and helps you remember what you need to do to remain alive and get rescued.

THE RULE OF THREES

A common understanding among those concerned with survival is that various key time limits have a numerical value linked to the number three. Remembering the "rule of threes" can help you focus your decision-making, particularly if you are injured, at risk of further injury, or in immediate danger from the elements. In most cases:

- **Three seconds** is the psychological reaction time for making a decision.
- **Three minutes** is the length of time your brain can do without oxygen before it suffers irreparable damage.
- **Three hours** is the critical time you can survive unprotected in extreme climates.
- **Three days** is the approximate length of time you can live without water.
- **Three weeks** is the approximate length of time you can live without food.

SIZE UP YOUR SITUATION	**U**SE ALL YOUR SENSES	**R**EMEMBER WHERE YOU ARE	**V**ANQUISH FEAR
▪ First, assess the particulars of your surroundings, physical condition, and equipment. ▪ If you're in a group, you can share tasks and responsibilities, but remember that everyone will be affected by the consequences of your decisions.	▪ Most people react to a true survival situation through either training (they automatically do what they've been trained to do in a particular situation) or by instinct (they automatically do what their mind and body tell them to do).	▪ In any survival situation, it always helps to know where you are so that you can make the best decisions about what to do and where to go next.	▪ Fear and panic can be formidable enemies, so it's imperative that you have the knowledge and training to counteract them and prevent them from making your situation worse.
▪ Your surroundings: every environment has its own idiosyncrasies—hot and dry, hot and wet, cold, exposed, or enclosed. Determine what you need to do in order to adapt to that environment. ▪ Your equipment: assess your equipment and consider how it can best be used in your particular situation. ▪ Your physical condition: remove yourself and others from danger. Check for injuries and administer first aid as necessary. Remember that the trauma and stress of a survival situation may cause you to overlook, or subconsciously ignore, injuries that you yourself may have received.	▪ Whatever has happened, approach your predicament in a calm and rational manner. The situation requires careful thought and planning. ▪ If you act in haste, you may overlook important factors, lose vital equipment, or simply make matters worse. The saying "Undue haste makes waste" is especially true in a survival situation. ▪ Listen to your subconscious survival senses and gut feelings, and learn to act on them when they send you warning signals.	▪ Knowing your precise location can clarify whether rescuers are likely to find you or if you'll have to rescue yourself. You'll get a good idea of the obstacles you may face, whether you should remain where you are, and where best to position your aids to location. ▪ If you have made an Emergency Plan of Action (see pp. 24–25), someone will know your approximate location, and when you're due back. ▪ Channel your efforts into making sure that when rescuers are looking for you, your aids to location are in place.	▪ If uncontrolled, fear and panic can destroy your ability to make intelligent and rational decisions. They can cause you to react to your feelings and imagination rather than to your actual situation and your abilities. Moreover, they can incapacitate you and drain your energy, thus triggering other negative emotions. ▪ If you're in a group, your responses can have a direct effect on others—positive responses are productive and can motivate, while negative responses can undermine confidence and morale.

THE ENEMIES OF SURVIVAL

In a survival situation, there are seven factors, known as "the enemies of survival," that can work against you. In many cases, you can deal with them by knowing what they are and understanding their effects. One way to memorize these is by using a mnemonic such as: "**B**e **P**repared **T**o **F**ace **T**hese **H**ostile **F**actors".

BOREDOM AND LONELINESS

When boredom sets in, you become inactive and lose the ability to deal with your situation effectively, so you need to keep busy. Loneliness makes you overwhelmed by what you need to achieve, leading to a feeling of helplessness.

PAIN

If you're injured, don't ignore the pain. Attend to a minor injury, as it could grow into a major problem that could impair your ability to survive. A positive mental attitude coupled with keeping busy helps to distract the mind from pain.

THIRST

Thirst is not a good indicator of the body's need for water. Your body can be dehydrated before you feel thirsty. Stay ahead of dehydration, rather than have to deal with it. Prioritize your need for water in a way that's relevant to your environment.

FATIGUE

Tiredness leads to mistakes that, at best, cause frustration and, at worst, may result in injury or death. In a survival situation, it's unlikely that you'll replace the energy you use effectively, so everything becomes harder to achieve. Never underestimate the importance of quality rest to your physical and mental wellbeing.

TEMPERATURE

Temperature is a major factor in any survival situation and it will be affected by wind, rain, and humidity. You should dress to suit the environment you're in and be aware of the signs and symptoms of temperature-related injuries, such as dehydration, hypothermia, heat stress, and heat stroke (see pp. 272–73).

HUNGER

In a short-term survival situation (one to five days), procuring food is not a high priority. You can offset your reduced energy and stamina levels by drinking water and pacing yourself so you work within your limits. However, take every opportunity to procure food without expending energy.

FEAR

Fear is one of our body's greatest survival tools, as it can stimulate you, so that you're ready to act—however, it can also debilitate. Fear is good as long as you have control over it, and the key to controlling fear in a survival situation is knowledge.

	IMPROVISE	VALUE LIVING AND LIFE	ACT LIKE THE LOCALS	LEARN BASIC SKILLS
	▪ The true skill of a survivor is to understand what's required and improvise solutions to particular problems. Do you have the skills and knowledge to keep yourself alive and in a condition to be proactive in your own rescue?	▪ Some people without training and equipment have survived the most horrendous situations. In many cases, this was simply because they had the will to live and refused to give up!	▪ Whatever environment you're trying to survive in, you can be sure that the local or indigenous people and the local wildlife have developed ways of adapting to it in order to survive.	▪ Learning basic skills increases your chances of survival. Without training your prospects of survival are down to luck, which is never the best place to start. There is a saying: "Luck favors those who are best prepared."
	▪ You may start out with all the right equipment, but it may get lost or broken, or simply wear out. Your ability to improvise may mean the difference between your continued struggle to survive in relative comfort, or absolute misery. ▪ Think laterally, like the climber who was stranded on the side of a mountain with no aids to location. He used the flash on his camera to signal his location to a rescue helicopter. Improvise and overcome!	▪ The stories of prisoners of war often reveal what kept them alive: religious beliefs, thoughts of family and friends, or a determination not to let the enemy win. While these alone may not always be enough, they're certainly a key factor in any survival situation. ▪ It helps to bear in mind that hardship means different things to different people and in different cultures. ▪ Survival is about dealing with hardships and having the will to live. If the will to live is not there, then just having the knowledge and equipment may not be enough.	▪ Look at how the local people dress and act: in hot countries they leave manual work until the coolest parts of the day and work in a slow and deliberate manner to reduce sweating and therefore conserve water. ▪ If you're in a desert, for example, watch where animals go to find shade: they're mostly nocturnal and spend the day underground. Learn their tricks to find water, like the darkling beetle in the Namib Desert that drinks water from the fog that condenses on its carapace. ▪ If you're in a jungle, pay attention when the animals go quiet or quickly leave an area—danger is usually around the corner.	▪ Prior preparation is the key to survival: discover what you need to know about the environment you're going to; familiarize yourself with all your equipment; and practice your basic skills until they become second nature. This thorough preparation will help you to combat the fear of the unknown and give you the self-confidence to meet the challenges of any survival situation you may be in.

ATTRACTING RESCUERS

TO BE A SURVIVOR you need to rescue yourself or be rescued by others. If you can't rescue yourself—perhaps you're injured, completely lost, or trapped by bad weather—you must attract attention with location aids that you've either brought with you or have improvised.

AIDS TO LOCATION

Location aids can save your life, so make sure you know how to use them effectively. A helicopter on a search pattern may make only one pass over an area before moving on, so you'll only have a few minutes to act decisively.

COMMUNICATING WITH SATELLITES

When activated, a Personal Locator Beacon (PLB) transmits a radio distress signal to two complementary satellite systems called LEOSAR and GEOSAR. Together, these form the "COSPAS-SARSAT" system. The signal is then relayed to a rescue coordination center closest to the beacon's location. PLBs mainly use 406Mhz; the military also use 243.5Mhz and 282.8Mhz.

ATTRACTING ATTENTION

When you're in a survival situation, you need to employ your location aids in the best way possible. There are three main principles for attracting attention:

- **Attract:** pick a place that maximizes your chances of attracting attention using your location aids, such as open or high ground. Place your aids carefully so they can be detected from as wide an area as possible. The more obvious the signal the better.
- **Hold:** you need to hold the attention of rescuers by maintaining the signal until they indicate they've seen you. Try to send some critical information (the type of assistance you require or the number of survivors and their condition, for example) via a May Day, or Help message (see pp. 237–41).
- **Direct:** regardless of which location aids you use, once rescuers have been alerted, do all you can to direct them to your present location. If you have voice communication, direct them to your position. If you leave notes before moving on from places, make sure they contain dates and precise details of your intentions.

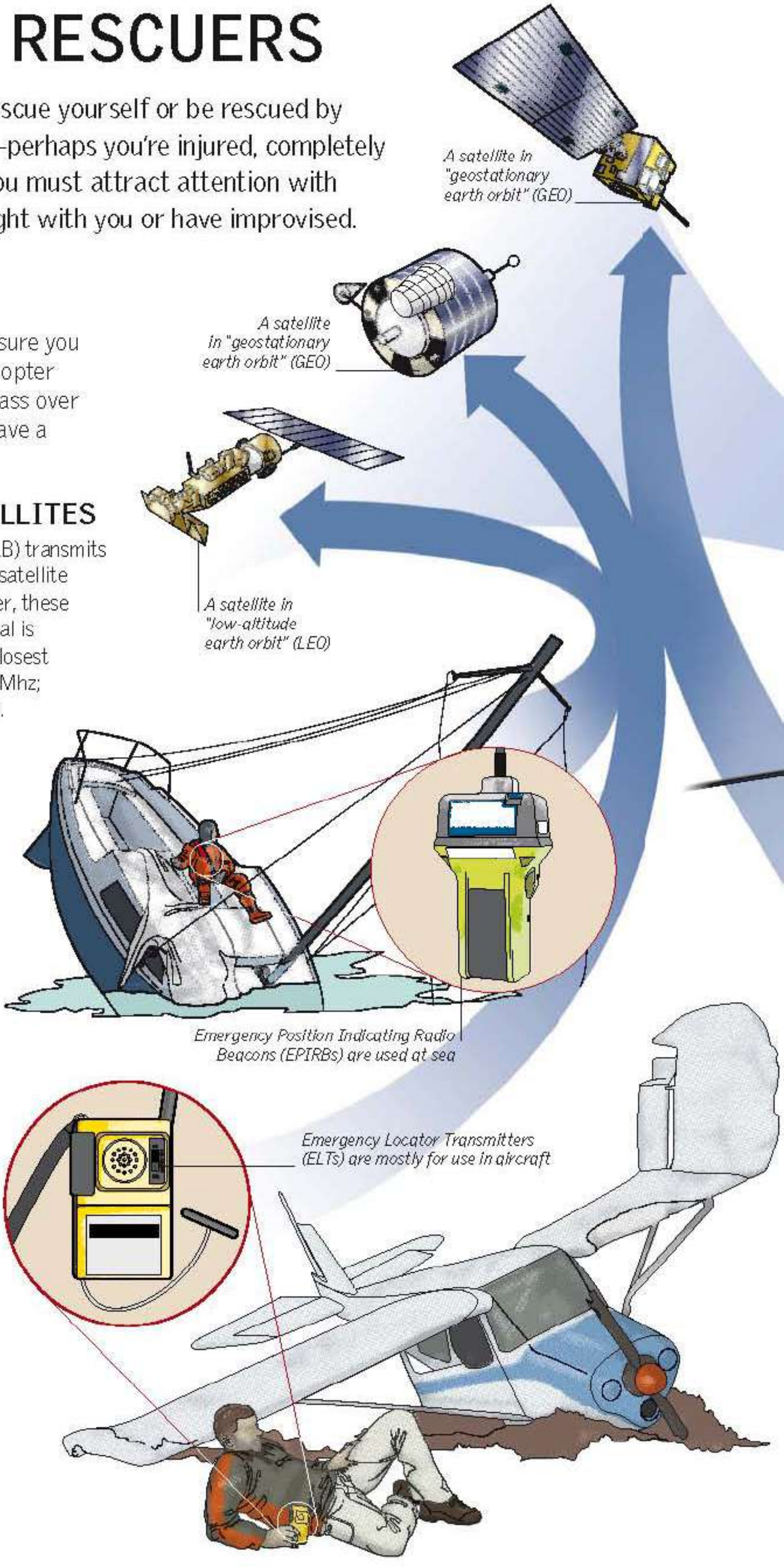

A satellite in "geostationary earth orbit" (GEO)

A satellite in "geostationary earth orbit" (GEO)

A satellite in "low-altitude earth orbit" (LEO)

Emergency Position Indicating Radio Beacons (EPIRBs) are used at sea

Emergency Locator Transmitters (ELTs) are mostly for use in aircraft

SHOULD I STAY OR SHOULD I GO?

You need to decide whether to stay where you are or to move to a location where you have a better chance of rescue or self-rescue. If you move on, remember the following:

- Keep your aids to location at hand. It's no good having your signal mirror at the bottom of your backpack when you may have only seconds to attract a passing vehicle or plane.
- Deploy your aids to location at the end of each day, even though it takes effort and you'll be moving in the morning.
- To show where you've been and where you're heading, leave markers, such as a note left in a visible position and visual clues on the ground or vegetation, to indicate your direction of travel.

MOBILE PHONES AND RADIOS

Always take a cell or satellite phone with you when traveling—they are essential in an emergency. On a sea trip, take a marine VHF radio to contact would-be rescuers.

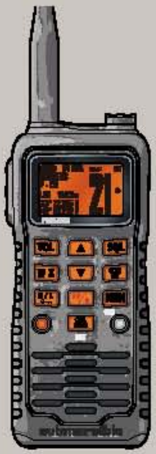

CELL AND SATELLITE PHONES

Wherever you are in the world, you can rent or buy a cell phone linked to local networks. Choose one with a GPS unit for fixing and racking your position and a camera so you can send pictures of your location and any injuries you've sustained. Alternatively, rent or buy a satellite phone that's connected to the Iridium Satellite Phone System, in which 66 low-earth orbiting (LEO) satellites provide complete coverage of Earth (including oceans, airways, and polar regions).

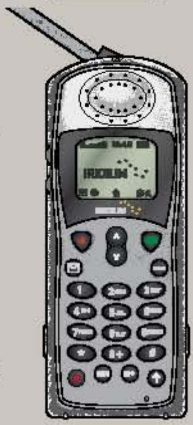

MARINE VHF RADIOS

All large ships and most motorized small craft are equipped with marine VHF radios. These hand-held units transmit and receive on frequencies between 156 to 174MHz—usually on Channel 16, the international calling and distress channel. Channel 9 can also be used in some places. Transmission power ranges between 1 and 25 watts, giving a maximum range of up to about 60 miles (110 km) between aerials mounted on tall ships and hills, and 5 miles (9 km) between aerials mounted on small boats at sea level. Your VHF radio should be waterproof, able to float, and kept on charge. Follow the instructions whenever transmitting and receiving on VHF.

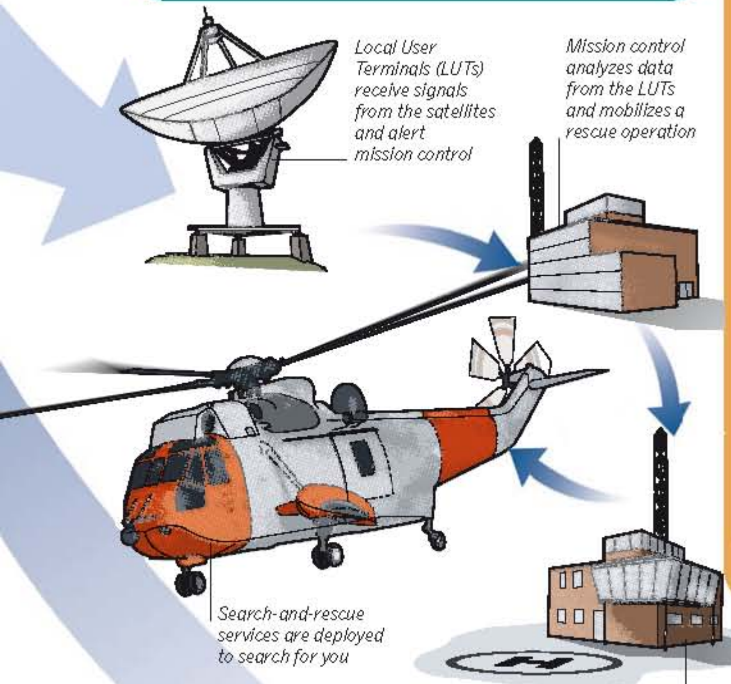

Local User Terminals (LUTs) receive signals from the satellites and alert mission control

Mission control analyzes data from the LUTs and mobilizes a rescue operation

Search-and-rescue services are deployed to search for you

A regional rescue coordination center alerts the search-and-rescue teams

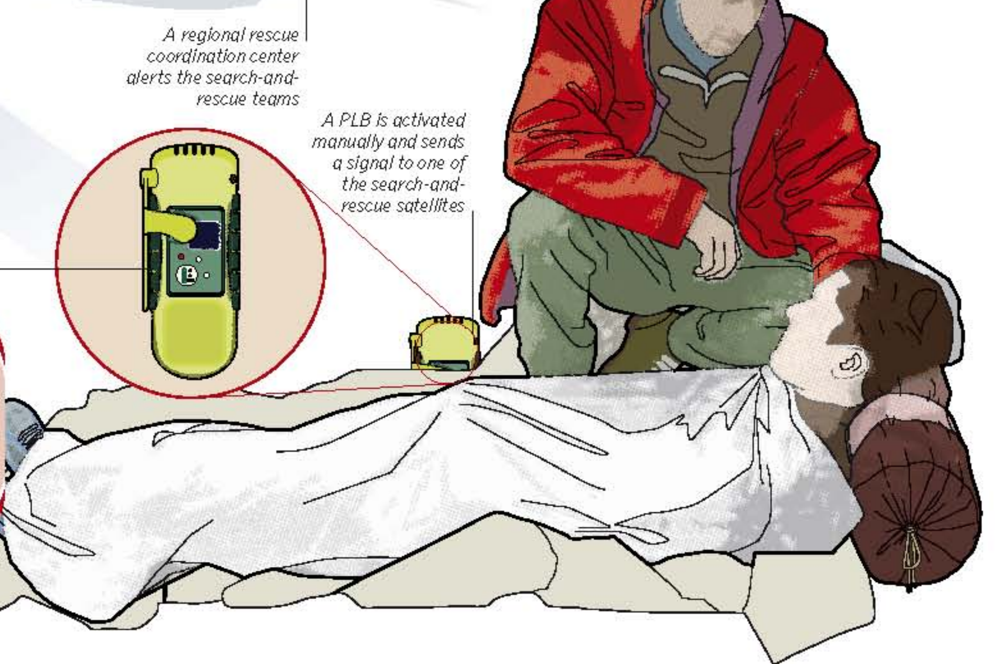

A PLB is activated manually and sends a signal to one of the search-and-rescue satellites

Personal Locator Beacons (PLBs) are for personal use and kept on the body

WARNING!

PLBs are for emergency use only and should always be used responsibly.

BUILDING SIGNAL FIRES

A signal fire is a very effective way of attracting the attention of rescuers, but you do need the right materials and building one requires some effort. A well-constructed signal fire located in a good position will generate a large amount of smoke that can be seen from a long way away. Two variations—the dome and the log cabin—are shown here.

TOOLS AND MATERIALS

- Knife and saw
- Poles and long saplings
- Green vegetation
- Tinder, kindling, and fuel
- Lighter or matches

SIGNAL FIRE ESSENTIALS

Whichever signal fire you make, there are several important principles to follow in order to make it as effective as possible.

FIRE FORMATION

If you can, prepare three signal fires in a recognized formation, such as a triangle or in a straight line. If the fires are random, they could be misread as a bush fire or a group of native fires. The distance between each fire should be at least 65 ft (20 m) but needs to be dictated by how quickly you can effectively light each one in turn.

FIRE COMPONENTS

Essentially, each signal fire is composed of a ready-to-light firebase made of dry tinder, kindling, and fuel. When you add the merest spark or flame to the tinder, the firebase immediately turns into a sustainable fire.

- Site the firebase off the ground to prevent damp from reaching it, and to allow enough airflow to help it ignite more effectively.
- Cover the top of the firebase with large amounts of green vegetation and anything else that produces smoke, such as tires. The covering keeps the firebase dry.
- Place a second stack of green vegetation near the firebase and add this as the original vegetation burns out.

AT THE READY

Once you have prepared the signal fire, you need to keep the following nearby, ready for use at a moment's notice:

- Dry tinder in a waterproof container underneath the firebase.
- Hexamine fuel tablets, stove fuel, gas, paper, or birch bark to guarantee the fire gets going again if it starts to dwindle.
- Something to create a spark or flame.
- A witch's broom, made of a cleft stick stuffed with kindling or bark, next to your camp fire, so you can quickly light the broom and transfer it to the signal fire.

MAKING A DOME SIGNAL FIRE

Prepare a large fire on a raised platform under a dome-shaped structure made from bent saplings. Fueled by air from below, the fire creates plumes of smoke from the green vegetation. If you have no saplings, use poles to a make a teepee shape.

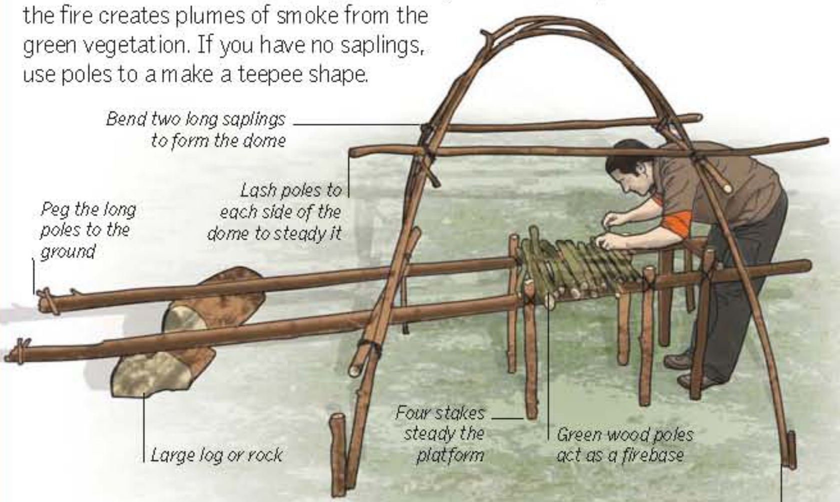
Bend two long saplings to form the dome
Peg the long poles to the ground
Lash poles to each side of the dome to steady it
Large log or rock
Four stakes steady the platform
Green wood poles act as a firebase
Stick the ends of the saplings firmly into the ground

1. Lay two long poles parallel to each other. Prop them up over a log or rock at one end, and tie them to four stakes at the other. Lay lengths of green wood side by side on the long poles under the dome, to act as a firebase.

- Bend two long saplings at 90 degrees to each other to form a dome.

4. Layer green vegetation on top of the dome to form a roof over the fire platform.

- Keep additional green vegetation nearby to add to the fire when it's alight.

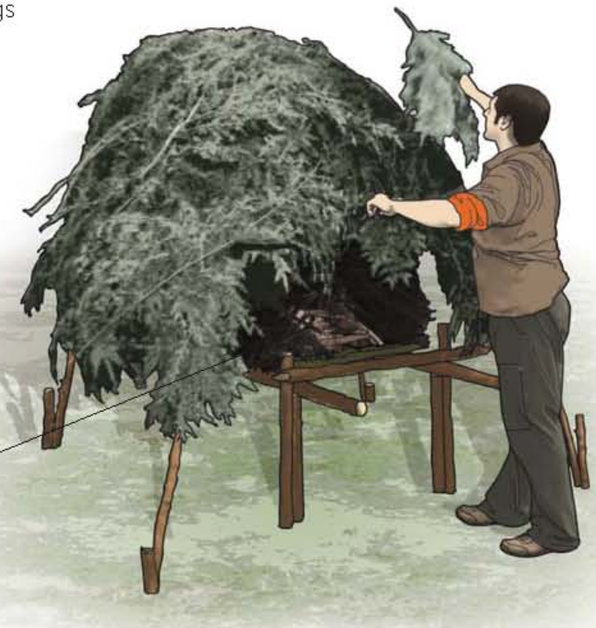
Leave an access point for lighting the fire, but keep it closed so the tinder, kindling, and fuel stays dry

MAKING A LOG CABIN SIGNAL FIRE

This signal fire is known as the "log cabin" because of the way the fuel wood is stacked.

- The framework for the fire is a stack of green wood, made up of pairs of poles arranged in alternating layers set at 90 degrees to each other.
- Place an additional supply of green vegetation nearby to be used as required, but not so close that it catches fire by accident.

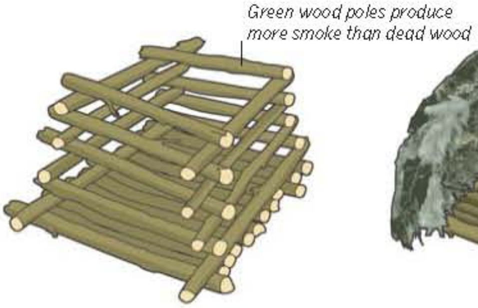
Green wood poles produce more smoke than dead wood

1 Lay a platform of green wood on the ground and build a firebase of dry tinder, kindling, and fuel on top.

- Build a "log cabin" over the firebase using green wood.

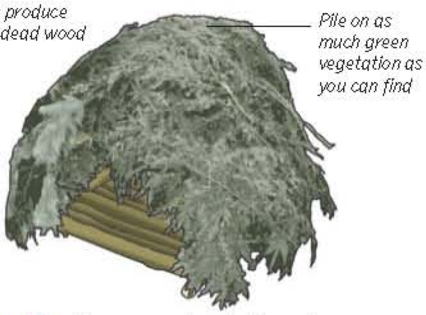
Pile on as much green vegetation as you can find

2 Heap green vegetation over the platform to generate thick plumes of smoke when the fire is lit.

- Leave an access point at the front to allow you to light the firebase.

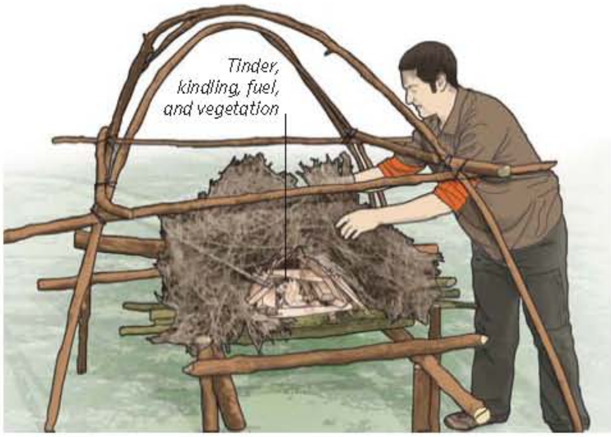
Tinder, kindling, fuel, and vegetation

2 Prepare a fire (see pp. 124–25) on the platform and load it with green vegetation.

- Lay the vegetation close enough to the fuel below to catch easily when the fire is lit—but don't smother it.

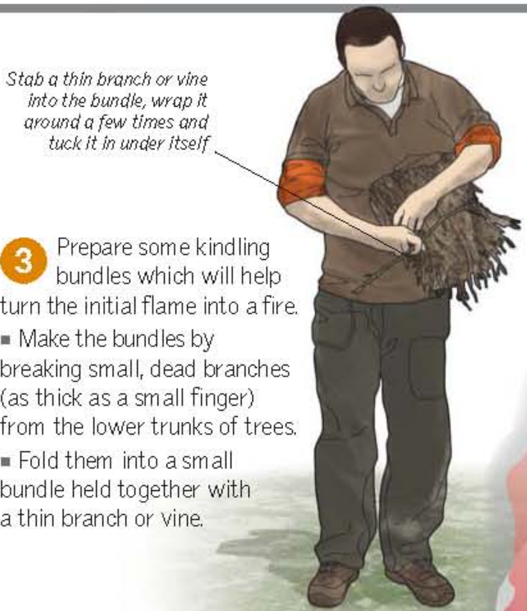
Stab a thin branch or vine into the bundle, wrap it around a few times and tuck it in under itself

3 Prepare some kindling bundles which will help turn the initial flame into a fire.

- Make the bundles by breaking small, dead branches (as thick as a small finger) from the lower trunks of trees.
- Fold them into a small bundle held together with a thin branch or vine.

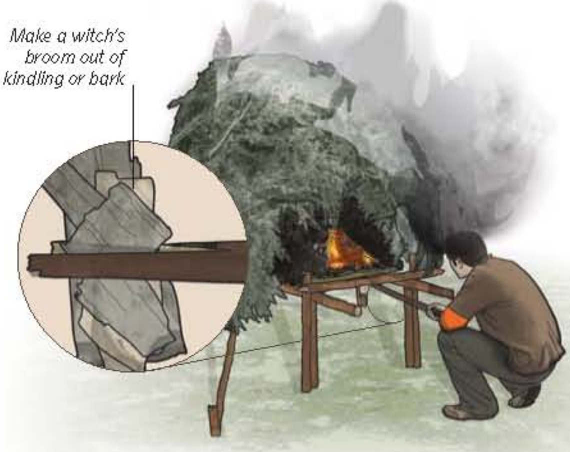
Make a witch's broom out of kindling or bark

5 Light a witch's broom (see panel, left) from your camp fire, take it to the dome, and light the fire.

- Use kindling bundles (see Step 3) to fuel the fire.

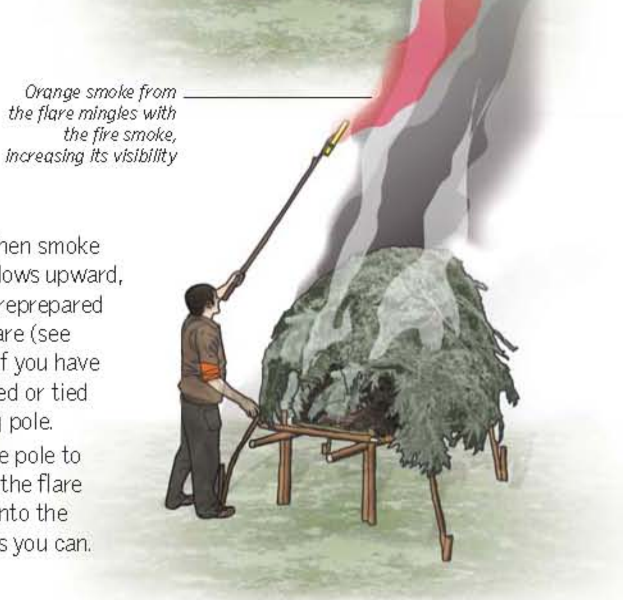
Orange smoke from the flare mingles with the fire smoke, increasing its visibility

6 When smoke billows upward, light a preprepared signal flare (see p. 241), if you have one, taped or tied to a long pole.

- Use the pole to position the flare as high into the smoke as you can.

OTHER RESCUE SIGNALS

As well as signal fires, there are other visual devices for attracting potential rescuers. Some are unusual and others, such as a whistle, quite obvious. Whatever device you use, you have to try to hold the rescuers' attention by persisting with your signal.

LIGHT WINDMILL

You can create a highly visible, illuminated "windmill" effect by whirling a chemical lightstick (cyalume) in front of you attached to the end of a cord 3 ft (1 m) long. Depending on local conditions it can be seen up to 2 miles (3 km) away—or further by an aircraft. Blow your whistle at the same time, using the International Distress Signal of six blasts over one minute, then one minute's silence. The reply is three short blasts.

CREATING A STROBE EFFECT

LED strobe lights are compact, robust, waterproof, and are sometimes incorporated within standard flashlights. They are very powerful and can be seen from a long way away. Many strobes can be set to produce a variety of sequenced flashes, including flashing SOS in Morse Code. Even if you don't have a dedicated strobe light, you can still switch your regular flashlight on and off repeatedly to attract attention.

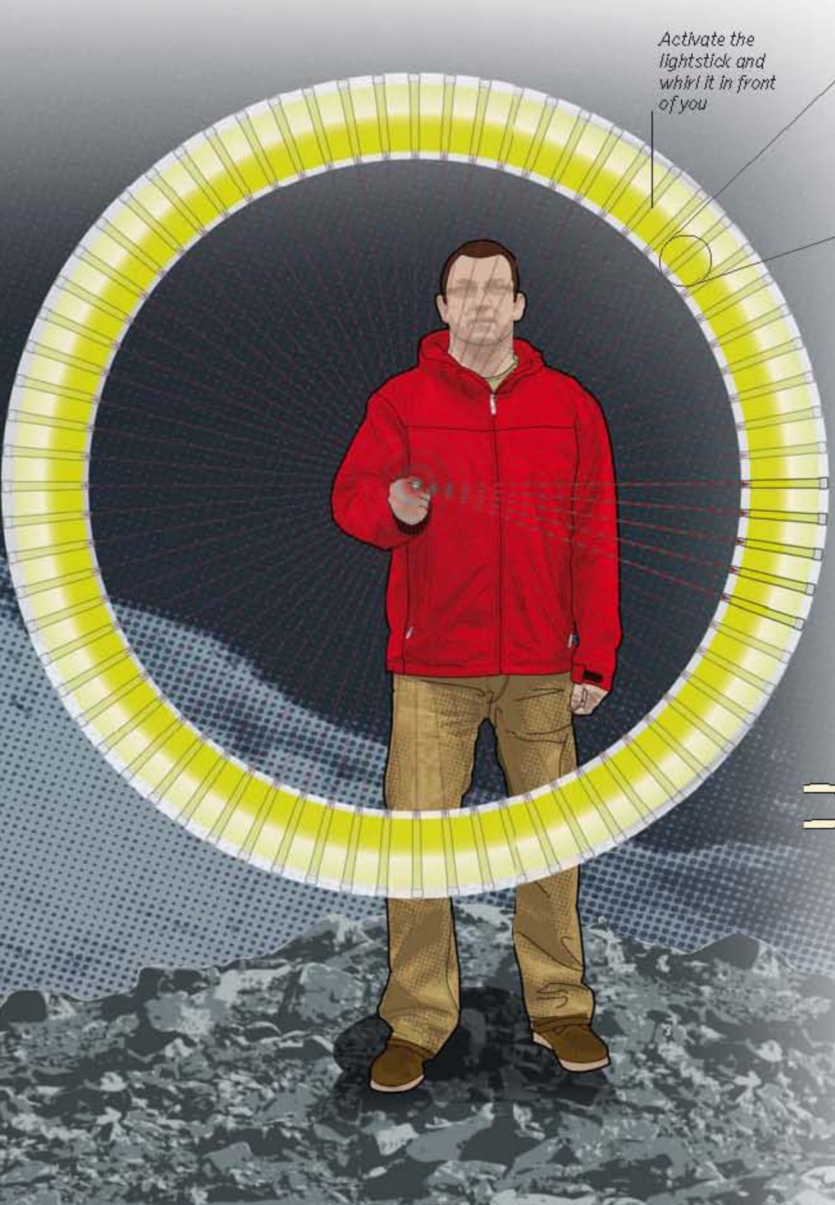

Activate the lightstick and whirl it in front of you

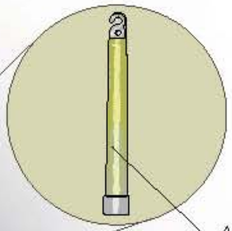

A lightstick is filled with a luminous substance known as cyalume

TINSEL TREE

The signaling device known as a "tinsel tree" is best used when you're remaining in one place for a while. It's effective only during daylight hours, so you'll need to find a suitable and easily visible tree or bush in a sunny position.

- Cut a silver survival blanket, a roll of aluminum foil, or any reflective material into strips.
- Attach the strips to the tree so that they move in the breeze, catching the sunlight and glinting like small mirrors.

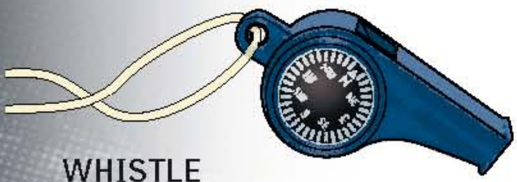

WHISTLE

Take a whistle with you when you venture out into the wilderness because sound, especially in remote areas where there's little or no noise pollution, travels incredible distances. A whistle is an item of first-line gear (see pp. 42–43), and should be kept on a cord around your neck.

SIGNALING DISTRESS

Two of the most recognized signs of distress are a little red star shooting up into the sky or a billowing cloud of orange smoke. There are many different types of signal flares and rockets, ranging from simple, handheld devices that fire flares into the sky to specialized kits designed to penetrate thick jungle canopy. Usually found in the emergency survival packs of aircraft and liferafts, flares are also available from specialized shops.

SIGNAL FLARES

One end of a signal flare has an orange-colored smoke signal for daytime use. The other end has a flare for night-time use (but it can be operated during the day, too). Remember the following:

- Follow the instructions on the outside of the device.
- Wear gloves to protect your hands.
- The heat from a flare can damage a liferaft, so keep your flare well clear when you light it.
- Don't discard flares unless you have used both ends.

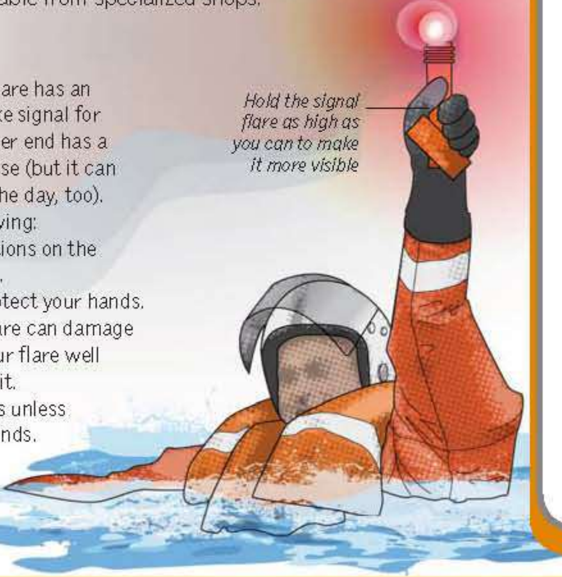

Hold the signal flare as high as you can to make it more visible

GROUND-TO-AIR MARKERS

You can improvise internationally recognized emergency signals on the ground that rescuers in the air can see.

GETTING NOTICED

To make a ground-to-air marker, use anything that contrasts with the ground, such as orange lifejackets, seaweed, clothing, rocks, branches of trees, or soil.

- Whatever you use, make sure the message is big and visible from all directions.
- Check on your markers regularly.
- When appropriate, use one of the emergency codes shown below. SOS or HELP written in big letters will attract attention too.

V	Require assistance
X	Require medical assistance
N	No/Negative
Y	Yes/Affirmative
↑	Proceed this way (arrow points)

SIGNAL MIRROR

A signal mirror, or "heliograph," has a shiny surface that reflects the sun and sends flashes over distances exceeding 30 miles (50 km), depending on the strength of the sun, the size of the mirror, and the clarity of the air. You can even use the light from the moon when it's full. By interrupting the flashes you can send messages in Morse Code.

IMPROVISING A SIGNAL MIRROR

If you don't have a heliograph or any other mirror, try using anything with a shiny, reflective surface, such as a foil food packet, a CD, or the bottom of a beverage can.

WARNING!

Don't dazzle your would-be rescuers by continuing to flash directly at them. Using your signal mirror intermittently is more effective for attracting attention.

① Polish the base of a can to a shine with a slightly abrasive paste. You can use charcoal and water, toothpaste, or even chocolate.

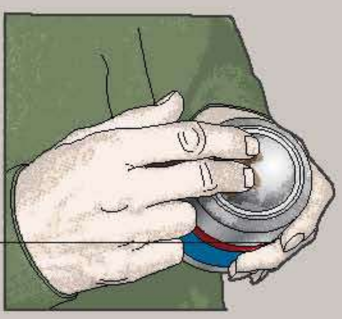

Polish the concave base—it will only take a couple of minutes

② Hold the can in front of your face with the polished base facing the sun.

- Direct the flash of reflected sunlight on to the palm of your other hand to practice controlling the light.
- Aim the can toward a rescuer and send a signal, either by moving your palm up and down to interrupt the flash or by directing the flash between a "V" made with your thumb and fingers.

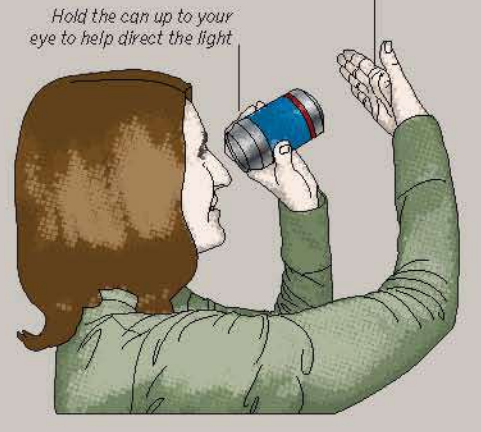

Hold the can up to your eye to help direct the light

The V-shape helps you direct the light

WILD ANIMALS

WALKING IN THE WILDS of Alaska and coming across a fresh deposit of bear feces is a sobering experience. It quickly dawns on you that you're potentially a part of the natural food chain. What do you have to do to protect yourself, and will it be effective or just provoke the animal? Fortunately, the survival instincts of wild animals are such that, with only a few exceptions, most animals, regardless of their size, will avoid confronting you. However, they will defend themselves by attacking you if provoked, cornered, or surprised, particularly when they have young.

FERAL ANIMALS

Wherever you go in the world there are animals that have escaped into the wild from zoos, wildlife parks, and domestic situations. These feral animals include pigs, cats, dogs, mink, and muntjacs. In many cases, they have become particularly aggressive and seem to have lost any fear of confrontation with humans. Try to find out which feral animals are common in the region you're traveling in and avoid them if at all possible.

	BEARS	BIG CATS	HIPPOS	ELEPHANTS	CROCODILIANS
ANIMAL FACTS	▪ Brown bears stand taller then black bears and weigh more. Polar bears are the largest land predator. All bears are powerful, with claws on their strong paws. Brown bears can run faster than humans. ▪ Bears are found in the Northern Hemisphere and parts of South America.	▪ Tigers are the largest cat, with lions a close second. All big cats have sharp teeth and claws. ▪ Other dangerous cats include cougars (also called pumas, mountain lions, and panthers), leopards, and jaguars. ▪ Big cats are found on all continents, except for Australasia.	▪ The hippopotamus is the third-largest animal in Africa after the elephant and white rhino. ▪ Hippos can weigh more than 3 tons. They have long, razor-sharp incisors and tusklike canines.	▪ African elephants have larger ears than Indian elephants and are more aggressive. ▪ Elephants stand up to 13 ft (4 m) tall and weigh up to 6 tons. ▪ They can reach speeds of 25–30 mph (40–48 kph).	▪ Crocodilians include alligators, gharials, and crocodiles. ▪ They can stay submerged for more than an hour at a time, can swim up to 20 mph (32 kph), and run as fast as 11 mph (17 kph) over short distances. ▪ Crocodilians live in many subtropical and tropical parts of the world.
PRECAUTIONS	▪ Look for signs of bears and check with locals about bear activity. ▪ Stow and cook food away from the campsite. Keep unused food and garbage out of reach. ▪ Avoid thickets and streams where bears rest and feed. ▪ Carry a stick, knife, or spray (see opposite).	▪ Avoid coming into contact with big cats. They are most unlikely to attack unless you provoke them or threaten their cubs. ▪ An exception is the cougar, which is responsible for an increasing number of unprovoked attacks in urban areas of North America.	▪ Avoid provoking a hippo. Many encounters are the result of hitting a partially submerged hippo with a canoe or boat. Almost all hippo attacks are fatal. ▪ Stay vigilant on or beside rivers that hippos are known to frequent. ▪ Don't get between a hippo and water.	▪ Keep away from places that elephants frequent, such as watering holes. ▪ If you do come across them, don't get too close, especially if there's a baby elephant nearby. ▪ Look for a safe place to retreat to, such as a vehicle, rocky outcrop, or a tree, before you need it.	▪ Stay away from waters and river banks where crocodilians live. Be vigilant at all times. ▪ If you need to go near the water, watch the area for at least 30 minutes first. ▪ Don't go to the same spot twice, as crocodiles may lie in wait the next time. ▪ Keep a defensive weapon, such as a knife, close by.
WHAT TO DO IF...	▪ Make a noise to let a bear know you're there. ▪ If you see a bear, stay calm. Make yourself look as big as possible by raising your arms. ▪ Walk slowly backward. Don't run. If the bear follows, stop and hold your ground. ▪ If the bear attacks, play dead, or fight back.	▪ If a big cat approaches, stare at its eyes, shout, and make a noise to confuse it. Use a spray if you need to (see opposite). ▪ Don't turn your back or run away. ▪ If a big cat sees you, do not crouch or bend down. An upright human makes less attractive prey than a four-legged animal.	▪ If you encounter a hippo, go back on your tracks and find another route. ▪ If a hippo yawns at you, he's not feeling sleepy—this is a threat. He's showing you his teeth, tusks, and jaws that can snap a canoe in half! Make every effort to escape.	▪ If an elephant squares up to you with flared ears, trumpeting, and kicking the dirt in front of it, then back away. ▪ If an elephant charges, run to your safe place (see above). ▪ As a last resort, play dead and hope the elephant loses interest.	▪ If you meet a crocodilian on land, run away. ▪ If a crocodilian does grab you, it may let go—if it does, run. ▪ If it drags you into the water, fight back. Stab under its throat with a knife, hit or poke its eyes, strike its nostrils hard, or bang the large valve at the back of its throat.

ARMING YOURSELF

If you're traveling in areas where encounters with wild animals are possible, be ready with a weapon such as a knife, large stick, or deterrent spray. Remember, avoidance is always better than confrontation.

PEPPER SPRAY

One type of deterrent spray contains the pepper ingredients capsaicin and capsaicinoids. It shoots a cloud or stream that irritates an animal's eyes. The animal usually withdraws, but take care until the danger has passed.

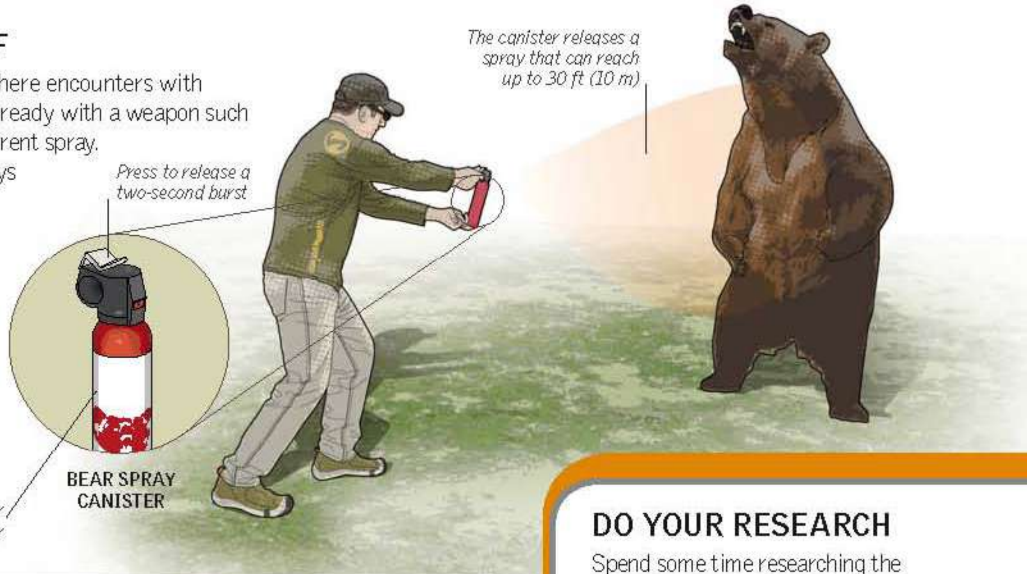

The canister releases a spray that can reach up to 30 ft (10 m)

Press to release a two-second burst

BEAR SPRAY CANISTER

The pepper spray is under pressure in a canister

	SHARKS	SNAKES	AMPHIBIANS
ANIMAL FACTS	▪ There are more than 450 species of shark, but only a few of them pose any danger to humans. These include great white, bull, and tiger sharks. ▪ Most attacks occur in coastal waters in the tropics and subtropics, especially where the water is murky or stirred up by the surf of breaking waves.	▪ Less than 15 percent of the 3000+ types of poisonous snake are regarded as being dangerous to humans. ▪ Depending on the snake, venom either affects the blood, the nervous system, or the heart. ▪ Snakes live in every part of the world except for very cold environments.	▪ Many frogs, toads, and salamanders secrete poison through their skin. The most poisonous frog known is the golden poison-dart frog from South America. ▪ Amphibians live where the climate is sufficiently warm and wet for them to breed.
PRECAUTIONS	▪ Avoid shark-infested waters or waters where sharks are known to feed or frequent. Ask the locals about recent sightings of shark activity. ▪ Carry something, such as a pole or spear, that you can use as a weapon.	▪ Wear long pants and boots, and cover your neck. ▪ As you walk, tap the ground in front of you with a stick to alert snakes to avoid you. ▪ Step onto logs rather than over them. ▪ At night, place your boots upside down on sticks. ▪ Don't put your hands into holes or cracks where snakes might be hiding.	▪ You can touch a poisonous frog or toad with no ill effect. The poison only takes effect if it enters your system via an open wound or your mouth or eyes. If this happens, get emergency help without delay.
WHAT TO DO IF...	▪ If you see a shark, stay calm and move to safety. ▪ If it makes a move toward you, swim away smoothly but keep watching it. ▪ If it rushes at you, hit it on the nose with whatever you have at hand. ▪ If it grabs you, aggressively strike its eyes or gills.	▪ If you come across a snake, try to remain completely still. Most snakes will instinctively move away and are more likely to attack a moving target. ▪ Don't panic. If you have a stick, slowly bring it to a position ready for use. ▪ If a snake attacks, hit it hard on the head.	▪ If contact with a poisonous amphibian is unavoidable, immediately wash the affected area. Keep the water you use away from open cuts or abrasions. ▪ Thoroughly wash your hands and don't put your fingers in your mouth or rub your eyes.

DO YOUR RESEARCH

Spend some time researching the area you'll be traveling in and find out which animals present the greatest dangers. Discover as much as you can about their habits and how best to avoid coming into contact with them.

SPOT THE SIGNS

Understand the habits of these dangerous creatures: when do they forage, sleep, and drink? What are their habits?

- Learn to recognize the natural signs that indicate their presence: tracks or prints; feces/droppings; territory markings, such as flattened vegetation where the animal rubs its scent on the ground; telltale signs, such as where bears have clawed the bark from trees.
- Be aware of other animals disappearing quickly from a watering hole, or a noisy jungle suddenly going quiet.
- Ask local people if there have been any reports of animals attacking people.
- Find out the best ways to ward off an attack or defend yourself should you have no option (see left). Adopting a certain stance or submissive posture, or avoiding or maintaining eye contact, could make the difference between safety or danger, life or death.

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—IN COLD CONDITIONS

USEFUL EQUIPMENT

- Waterproof, layered clothing
- Trekking pole or ice ax
- Collapsible snow shovel
- Light, whistle, avalanche beacon
- Tarpaulin
- Gore-Tex® gloves and gaitors
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

34-YEAR-OLD ICE HOCKEY PLAYER ERIC LEMARQUE became lost in the snowy wilderness of the Sierra Nevada mountains in California after snowboarding at Mammoth Mountain ski resort. Despite being under-equipped and ill-prepared, his improvisation helped him to survive for seven days.

LeMarque was snowboarding alone late on Friday, February 6, 2004, setting off down an unmarked run of virgin powder snow just as the lifts were shutting for the day. Coming to a stop on a flat section, and with visibility reduced to just 10 ft (3 m), he realized he had lost the trail. Wearing uninsulated ski trousers and jacket, and with just an MP3 player, a cell phone with a dead battery, and wet matches, he was ill-prepared for survival, so he began to search for the trail.

LeMarque chose the wrong direction and walked away from the ski resort. Realizing he faced a night in the open, he used his snowboard to dig a crude trench, lining it with bark scraped from a pine tree for insulation from the snow. He tried to light a fire with pine needles and lint from his clothing, but his matches were too wet. He also ate pine needles and bark.

"WITH VISIBILITY REDUCED TO JUST 10 FT (3 M), HE REALIZED HE HAD LOST THE TRAIL"

Over the next five days, LeMarque walked further into the wilderness, leaving scraps of clothing for rescuers and attempting to signal to passing planes with the blue LCD screen of his MP3 player. Fortunately for LeMarque, rescuers spotted his snowboard tracks and followed it for 24 hours, finding him on February 13. He was barely conscious, dehydrated, hypothermic, and malnourished—having lost 35 lb (16 kg) in body weight—and was suffering from severely frostbitten feet, which later required both legs to be amputated below the knees. But he was alive.

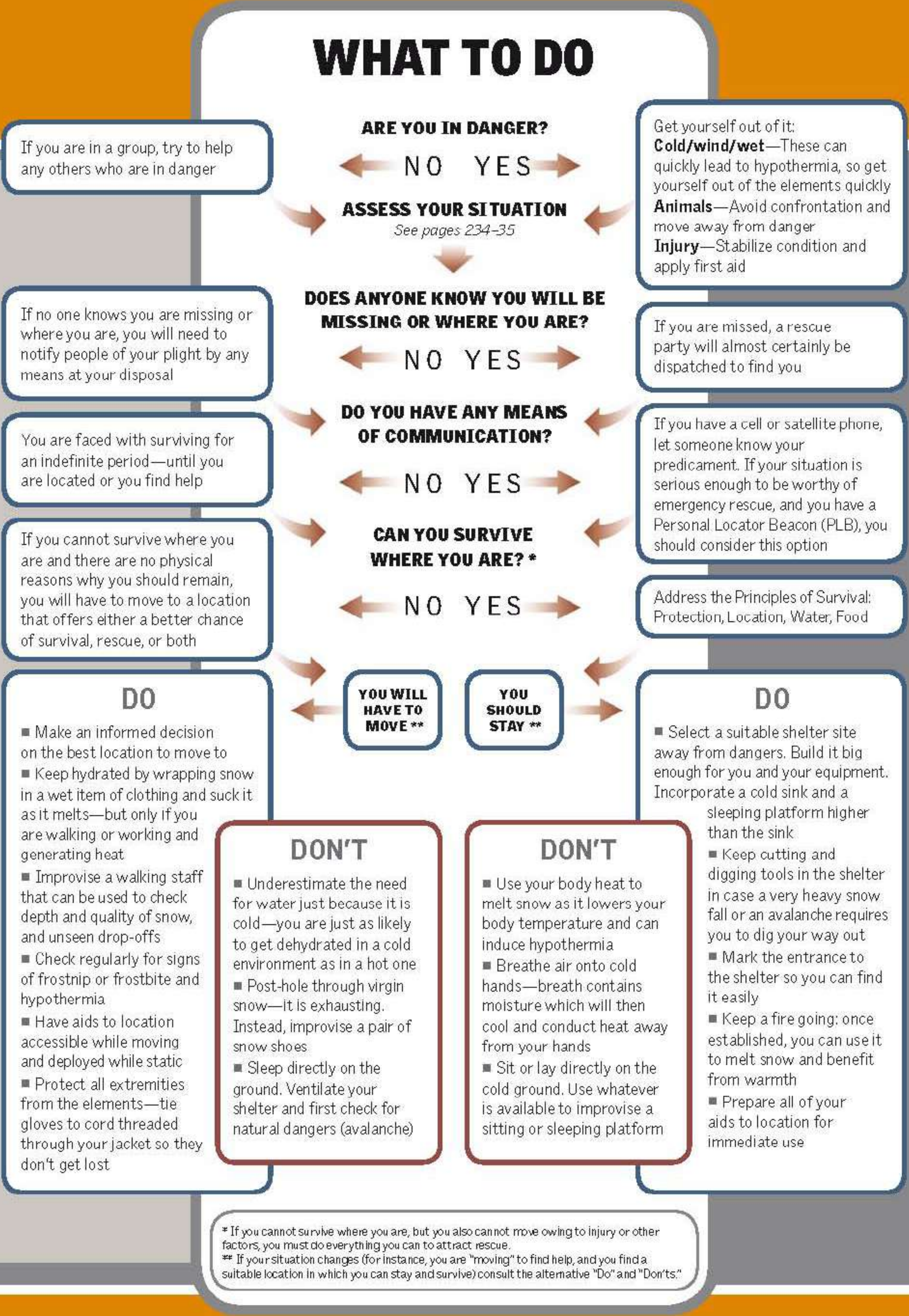

WHAT TO DO

ARE YOU IN DANGER?

NO YES

If you are in a group, try to help any others who are in danger

Get yourself out of it:
Cold/wind/wet—These can quickly lead to hypothermia, so get yourself out of the elements quickly
Animals—Avoid confrontation and move away from danger
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO YES

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO YES

You are faced with surviving for an indefinite period—until you are located or you find help

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO YES

If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers either a better chance of survival, rescue, or both

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Make an informed decision on the best location to move to
- Keep hydrated by wrapping snow in a wet item of clothing and suck it as it melts—but only if you are walking or working and generating heat
- Improvise a walking staff that can be used to check depth and quality of snow, and unseen drop-offs
- Check regularly for signs of frostnip or frostbite and hypothermia
- Have aids to location accessible while moving and deployed while static
- Protect all extremities from the elements—tie gloves to cord threaded through your jacket so they don't get lost

DON'T

- Underestimate the need for water just because it is cold—you are just as likely to get dehydrated in a cold environment as in a hot one
- Post-hole through virgin snow—it is exhausting. Instead, improvise a pair of snow shoes
- Sleep directly on the ground. Ventilate your shelter and first check for natural dangers (avalanche)

YOU SHOULD STAY **

DO

- Select a suitable shelter site away from dangers. Build it big enough for you and your equipment. Incorporate a cold sink and a sleeping platform higher than the sink
- Keep cutting and digging tools in the shelter in case a very heavy snow fall or an avalanche requires you to dig your way out
- Mark the entrance to the shelter so you can find it easily
- Keep a fire going: once established, you can use it to melt snow and benefit from warmth
- Prepare all of your aids to location for immediate use

DON'T

- Use your body heat to melt snow as it lowers your body temperature and can induce hypothermia
- Breathe air onto cold hands—breath contains moisture which will then cool and conduct heat away from your hands
- Sit or lay directly on the cold ground. Use whatever is available to improvise a sitting or sleeping platform

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.
** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

ENVIRONMENTAL HAZARDS

SOME PARTS OF THE WORLD have a reputation for sudden or extreme weather conditions, and unpredictable environmental hazards that cause havoc for everyone, not just travelers. Before you set off, make sure that you're aware of any such potential hazards in the areas you're visiting, and be ready for them, just in case.

BEING PREPARED

Many environmental hazards—from avalanches and volcanoes to forest fires, tornadoes, and hurricanes—can suddenly propel you into an emergency situation. While you can never be sure of escaping the worst excesses that Nature can throw at you, you can prepare yourself. A combination of the right equipment, local knowledge, and an awareness of evasive techniques will mean that you shouldn't have to find yourself in a survival situation thinking "if only".

COMMONSENSE ADVICE

No fail-safe guidelines will protect you from every environmental danger and hazard, but there is some commonsense advice that you will do well to heed before you venture into the wilderness:

- Check the prevailing weather conditions of your destination at the time of your journey. Log onto the internet or tune into local radio stations.
- Find out if you're heading into a particularly risky season of weather—for example, monsoons in India or hurricanes in the Caribbean.
- Take an emergency survival kit (see pp. 58–59).
- Learn the relevant survival techniques either before you go or as soon as you arrive at your destination—you may not get a second chance.
- Don't challenge Nature—it rarely loses.

EQUIPPING YOURSELF FOR AN AVALANCHE REGION

Before heading for the mountains, find out whether the prevailing conditions make avalanches more likely (see right).

- Take a collapsible shovel for moving snow and a probe for checking the depth of the snow.
- Keep some survival aids in your pockets in case you lose your backpack.
- Carry an avalanche transceiver—when activated, its signal can be detected by rescue services.

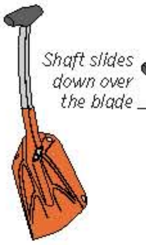
Shaft slides down over the blade

COLLAPSIBLE SHOVEL

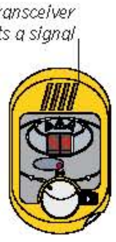

AVALANCHE TRANSCEIVER

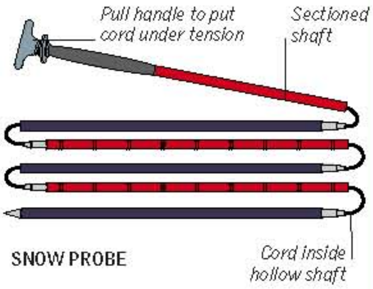

SNOW PROBE

SURVIVING AN AVALANCHE

There's always the danger of an avalanche on slopes that face away from the sun in the middle of winter, when a fresh layer of heavy snow sits on top of a weak layer of snow. An avalanche may be triggered when snow is disturbed by loud noises, an earth tremor, or the movement of skiers or snowboarders. Learn the warning signs, take the right equipment, and practice the emergency steps in case you get caught.

SPOTTING THE WARNING SIGNS

As you cross the snowbound mountains, hills, and valleys, look out for signs that warn of a possible avalanche:

- Convex slopes at an angle between 30 and 45 degrees.
- Slopes without trees or rocks.
- Loose, dry snow that doesn't settle.
- Soft, newly fallen snow that's more than 1 ft (30 cm) deep.
- Snow that sounds hollow.
- Snow that falls as crystals or pellets.
- Snow that falls at more than 1 in (2.5 cm) an hour.

TAKING EVASIVE ACTION

If you see or hear an avalanche, and you think it might be coming your way, take evasive action at once:

- Activate your avalanche transceiver in case you get caught in the snow.
- Try to take cover—for example, under a solid rock overhang if you can see one nearby (see below).
- If you can't find or reach suitable cover, try to sidestep the avalanche by skiing out of the way—at right angles to its potential path.

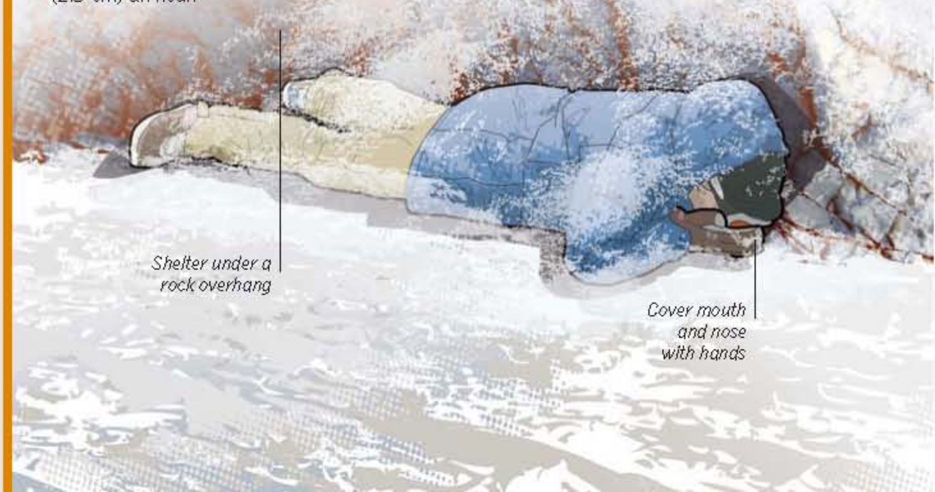
Shelter under a rock overhang
Cover mouth and nose with hands

HAZARDOUS GROUND

All kinds of local conditions can prove treacherous if you don't keep your wits about you. Walking through wetlands can be risky, as the ground can suddenly give way under foot and you can find yourself in a swamp or, worse still, a patch of quicksand.

SURVIVING A SWAMP

Freshwater swamps are found in low-lying inland areas and contain masses of thorny undergrowth, reeds, and grasses, and there may be dangerous animals, such as crocodilians and snakes. The water may be foul and mosquitoes, which can cause malaria, are often present. Moving through a swamp is difficult: you can be on solid ground one moment, and chest-deep in water the next. Where possible, try to build a raft or some kind of flotation aid to help you escape. Use the waterways to navigate your way to open water, where you have the best chance of either being seen by rescuers or finding your way back to civilization.

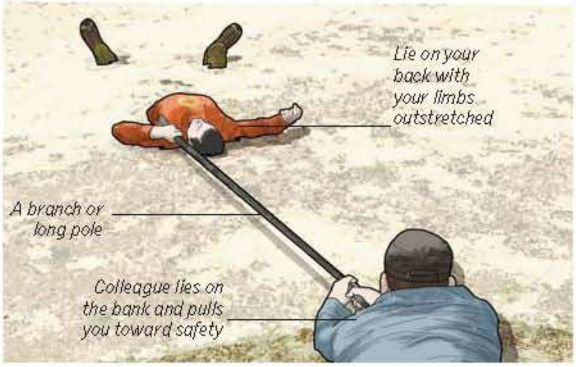
Lie on your back with your limbs outstretched

A branch or long pole

Colleague lies on the bank and pulls you toward safety

ESCAPING FROM QUICKSAND

If you step into quicksand, try to fall onto your back with your limbs outstretched to spread your weight. Try to get to the bank by paddling with your hands. Don't struggle as you will sink faster. If you're with someone, he or she needs to lie on firm ground and pull you to safety, reaching you with the help of a long pole, rope, or branch.

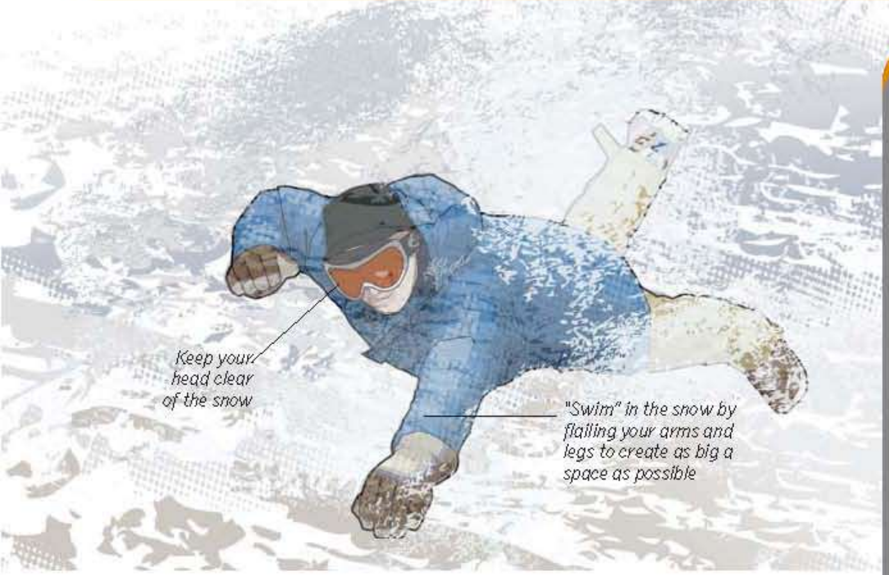
Keep your head clear of the snow

"Swim" in the snow by flailing your arms and legs to create as big a space as possible

ESCAPING FROM AN AVALANCHE

If you can't take evasive action and find yourself below an advancing avalanche with no prospect of escape, try to keep calm and remember the following advice:

- Remove the bindings from your skis and the loops from your ski poles. Put your rucksack on one shoulder, but ditch it rather than risk dislocating your shoulder.
- If you're overtaken by snow, "swim" with the flow (see above) to make a big space around you before the snow compacts and "sets". Use your hands or a collapsible shovel to clear the snow around your head and to create a space for you to breathe, as often people suffocate before they freeze.
- If you don't know which way is up, dribble saliva to find out which way is down—then dig in the opposite direction.
- Get out quickly, as time is of the essence, and shout when you hear potential rescuers to attract their attention.

ESCAPING A VOLCANIC ERUPTION

Find out if there's an active volcano in the region you're visiting. If there is, and its eruption is imminent, leave the area as soon as you can. It will pump enormous quantities of lava from the mantle below the crust and produce huge amounts of ash, toxic gases, debris, and mud. However, if you do get caught in an eruption:

- Be careful when driving to safety, as the ash and mud make roads slippery.
- Take cover to avoid the flying fiery debris, hot gases, and suffocating clouds of ash. Sulfur dioxide in the ash chokes your lungs and, when mixed with rain to form sulfuric acid, burns your skin.
- If you're caught in an ash cloud, cover your face with a wet cloth (or use a mask if you have one). Wash your skin afterward.

SURVIVING AN EARTHQUAKE

Find out if an earthquake is likely to happen in the region you're visiting. If you feel an earthquake coming:

- Get into the open, away from structures or trees that can fall on you. Then lie down.
- Stay in open ground until aftershocks and tremors have stopped altogether.
- If you're in a building, go to the lowest floor, and stay beside a wall or under a sturdy table. Turn off the gas if you have time.
- If you're in a vehicle, stop, but stay inside.

SURVIVING A FOREST FIRE

When vegetation on the ground is tinder dry, the slightest spark can start a fire that gets out of control and sweeps through a forest. A forest fire may be caused by lightning, a piece of glass focusing the sun's rays, a discarded cigarette, or a spark from a camp fire. Forest fires produce intense heat, thick smoke, and toxic gases—and they use up all the available oxygen in the air, making it impossible to breathe.

PREPARATION AND SPOTTING THE WARNING SIGNS

If you're going to a place where forest fires are a hazard, find out from local radio reports whether conditions are making them more likely. If you venture into a forest, make sure you carry a cell phone or some electronic means of signaling, and let others know your route and destination. The following warning signs may help you gain time before a fire is upon you:

- You'll probably smell a fire first, and you'll hear it crackling as it burns before you see it. If you smell a fire and you notice animals growing agitated, a forest fire may be close by.
- The smoke will help you establish how close the fire is. The direction of the smoke tells you which way the wind is blowing.

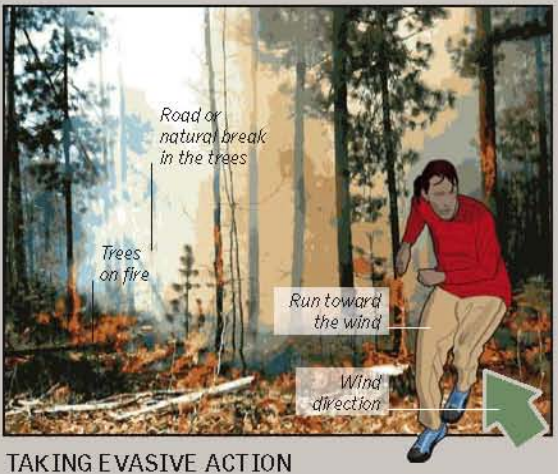

TAKING EVASIVE ACTION

If the wind is blowing toward the fire then move into it quickly. However, if the wind is behind the fire you could be in serious danger because the fire will be moving very fast.

- Try to find a river, lake, road, or natural break in the forest. Stay there until you're rescued or the fire has passed you by.
- Don't go up to high ground as fire is drawn faster uphill.
- Many forest fires spread on a wide front, so avoiding the fire by going around it may be impossible.
- If the fire is upon you, and the wall of flame is fragmented, the best course of action may be to run through the flames: take off man-made clothes (as they will melt onto you), cover as much of your skin as you can, and douse yourself in water if you have any. Take a deep breath, press a damp cloth to your mouth and nose, pick a spot where the wall of flames is thinnest, and run without stopping until you're through.
- If you're in a vehicle, stay inside. Park it as far away from the trees as possible, turn off the engine, close the windows, lie down on the floor, and cover yourself if you can.
- If escape is impossible, try to dig and bury yourself under soil. Dampen all clothing and get as low as possible.

EXTREME WEATHER

High winds and torrential rain are common seasonal occurrences across the world. From tornadoes and whirlwinds to tropical storms such as hurricanes and typhoons, these extreme weather patterns cause enormous damage and threaten many lives.

AVOIDING HURRICANES AND TORNADOES

Weather forecasts can predict approximate directions and paths of impending hurricanes and local emergency services will give advice on the predicted severity. If you're planning a trek in an area known for hurricanes and tornados, check long-range weather forecasts and be prepared to cancel or alter your plans if necessary—if you're caught in the open your chances of survival are extremely limited. If you're going to be in the wilderness during a "storm season" then take a battery- or solar-powered radio, check for weather updates, and be prepared to head back to civilization. If your home is in a tornado or hurricane area, follow the advice of local services, but most importantly be prepared with a plan that will ensure you don't get caught out.

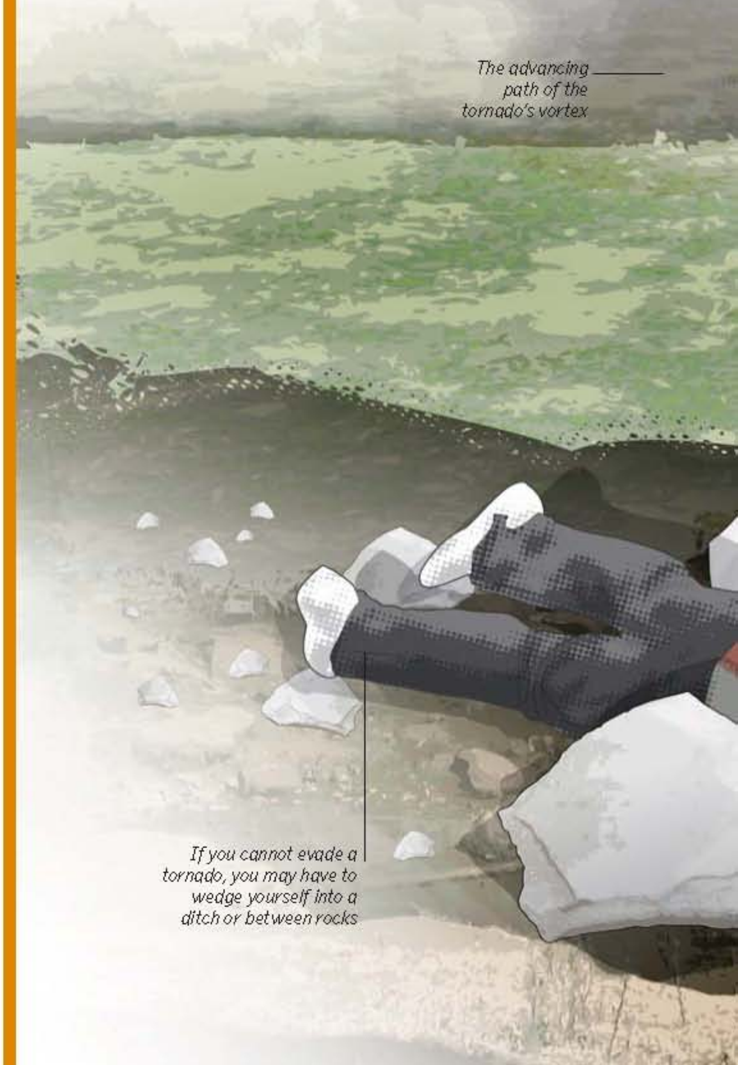
The advancing path of the tornado's vortex

If you cannot evade a tornado, you may have to wedge yourself into a ditch or between rocks

PREPARING TO EVACUATE

The arrival of a hurricane or tornado is a time when you can find yourself in a survival situation on your own home turf. The following advice can help you prepare to evacuate.

- Secure anything outside that could be picked up by winds and cause damage, such as rubbish bins and garden furniture.
- Prepare an evacuation plan (including pets if you have them). Find out where you will evacuate to and make several route plans to the location in case roads are unusable. Fuel your vehicle as soon as a hurricane warning is forecast.
- Put together a hurricane survival pack that contains what you need for a 72-hour period. Include drinking water, a change of clothing, non-perishable food, sleeping bags, radio, flashlights, and contact numbers for family and emergency services.
- Protect your house by securely boarding up all windows and doors. Turn off the water, gas, and electricity.

STAYING PUT

If you decide to stay put, you should still follow the guidelines above. As the storm approaches, move everyone and your supplies to an underground shelter or a room without windows.

- Monitor the radio reports and comply with the advice of the emergency services.
- Don't go outside until the "all-clear" has been given.

TAKING COVER OUTDOORS

If it appears that you are, or could be, in the path of a tornado or hurricane, do everything you can to move out of its way—choose a direction at a right angle to its path.

- If you're driving but can't get out of the way, try to park in an underpass. Leave the vehicle and wedge yourself into a secure position as high as you can under the underpass. If you're in the open, shelter upwind of your abandoned vehicle, so that it does not become a hazard to you.
- If you're out in the open, head for the lowest ground and find a ditch or depression (see below), in which you can shelter, or wedge yourself between rocks.

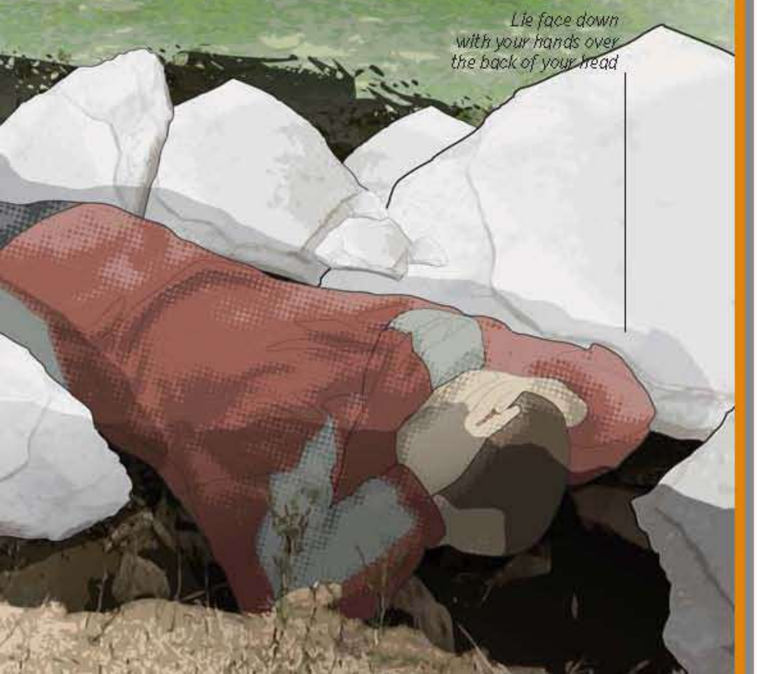
Lie face down with your hands over the back of your head

SURVIVING AN ELECTRICAL STORM

An electrical storm may develop when warm air rises and meets colder air. Electricity sparks between water droplets in the clouds, forming lightning that takes the shortest route to the ground. The following advice will help you avoid being struck by lightning, or minimize its effects if you are:

- Avoid getting caught in open ground—seek shelter but not under a lone tree: when lightning strikes a single tree a tremendous voltage fans out from its base.
- Remain in your vehicle if there is no other cover: it acts as a Faraday Cage (a metallic enclosure that prevents the entry or escape of an electromagnetic field).
- If caught in the open: make yourself as small as possible and limit the amount of area you cover. Do not lay down or stand up, crouch down low with feet together and hands off the ground.

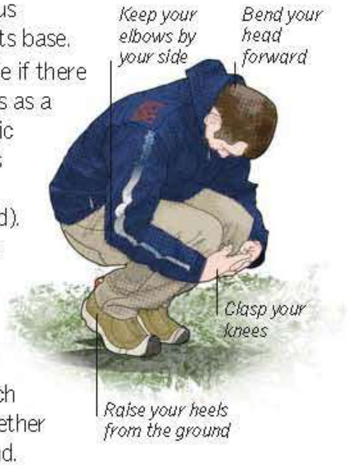
Keep your elbows by your side
Bend your head forward
Clasp your knees
Raise your heels from the ground

SURVIVING A FLASH FLOOD

A sudden deluge of rain doesn't always drain away quickly, instead flowing rampantly over the surface of the land in torrents, and causing flash floods. Soil, animals, vegetation, and even buildings can be quickly washed away. There may be landslides, too, and rivers may break their banks. The following advice can help in the event of a flash flood:

- If you're inside a building, move to the upper floors, taking with you essentials such as bedding, food, and matches. Unless the building is threatened, stay there until the waters have receded or you have been rescued.
- If you're outside, head for higher ground.
- Never walk or drive through a flood.
- Filter and boil water for drinking, as the sources of water around you may have been contaminated. Alternatively, collect rainwater to drink.

ESCAPING A SANDSTORM

If you can see a sandstorm coming, mark your direction of travel before it strikes and find somewhere safe to shelter—for example, behind some rocks. Face away from the direction of the wind and cover yourself as completely as you can, particularly your head, face, and neck.

SURVIVING AT SEA

THE SEA IS ARGUABLY THE HARSHEST environment of all and you could be thrown into a sea-survival situation for any number of reasons: bad weather, fire, mechanical failure, or collision. However, there's plenty of good advice to follow and equipment designed to protect you.

BE PREPARED

If you're venturing onto the open seas, you should plan for the worst and be prepared with knowledge, skills, and equipment.

KNOW YOUR VESSEL

Regardless of the size of your vessel, find out everything you can about what to do in the event of a major problem that requires you to abandon ship and take to the water. The points below are generic and may not apply to all situations or types of vessel.

LARGE PASSENGER VESSELS

If you're a passenger on a large vessel, such as a ferry, liner, or cruise ship, make sure you learn the safety procedures.

- Attend the "Abandon Ship" drills.
- Find out which emergency alarms indicate fire, collision, and abandon ship.
- Find out where the lifejackets are stowed, and how to put them on and operate them.
- Learn the escape routes.
- Locate the Emergency Lifeboat Stations and find out what your responsibilities are.
- If you have children or people with special needs with you, make sure you have a system for getting them on deck and providing them with suitable survival equipment.

SMALL VESSELS

These craft include yachts, small boats, canoes, and kayaks, so crew members need to agree on an emergency plan of action and delegate responsibilities relevant to the situation and the skills of each person.

- Know where the emergency equipment is stowed and how it's operated.
- If your vessel has an EPIRB (see p. 236), make sure everyone knows where it is and how it's operated. Those activated automatically are released hydrostatically from a bracket at a water depth of 3–10 ft (1–3 m). The buoyant EPIRB then floats to the surface and begins transmitting.
- Make sure your EPIRB has been registered so that, once the signal is detected, the rescue services know who it belongs to.
- Keep a grab bag to hand (see box, right).

SURVIVAL SUIT AND LIFEJACKET

A survival suit is designed to keep you warm and dry in extreme conditions and rough seas, and a lifejacket will keep you afloat with your head out of the water, even if you're unconscious.

PROTECTION

Survival suits and lifejackets are equipped with various protection features to help you survive in the sea.

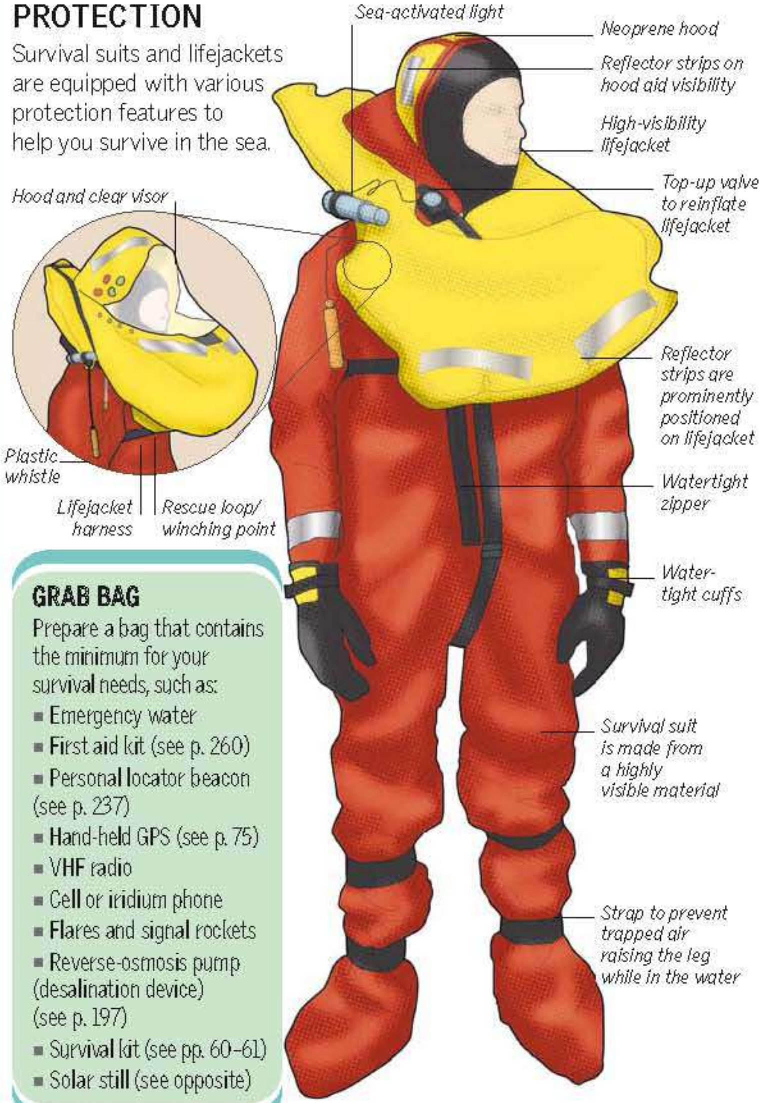
Sea-activated light
Neoprene hood
Reflector strips on hood aid visibility
High-visibility lifejacket
Top-up valve to reinflate lifejacket
Hood and clear visor
Reflector strips are prominently positioned on lifejacket
Watertight zipper
Plastic whistle
Lifejacket harness
Rescue loop/ winching point
Water-tight cuffs
Survival suit is made from a highly visible material
Strap to prevent trapped air raising the leg while in the water

GRAB BAG

Prepare a bag that contains the minimum for your survival needs, such as:

- Emergency water
- First aid kit (see p. 260)
- Personal locator beacon (see p. 237)
- Hand-held GPS (see p. 75)
- VHF radio
- Cell or iridium phone
- Flares and signal rockets
- Reverse-osmosis pump (desalination device) (see p. 197)
- Survival kit (see pp. 60–61)
- Solar still (see opposite)

TYPES OF LIFERAFT

Single-seater and multi-seater liferafts contain similar design features and survival aids. As well as being able to accommodate more people, multi-seaters carry larger quantities of fresh water and more anti-sea sickness tablets, for example.

MULTI-SEATER LIFERAFT

Many vessels carry multi-seater liferafts in a valise or a hard container. Multi-seaters can accommodate between four and 25 people and may be open-topped or covered. Many larger multi-seaters may also include locator beacons (see p. 236), paddles, and a solar still (see below).

SINGLE-SEAT LIFERAFT

Only one person can fit into a single-seat liferaft. You can inflate it with carbon dioxide, by pumping air in manually, or by oral inflation.

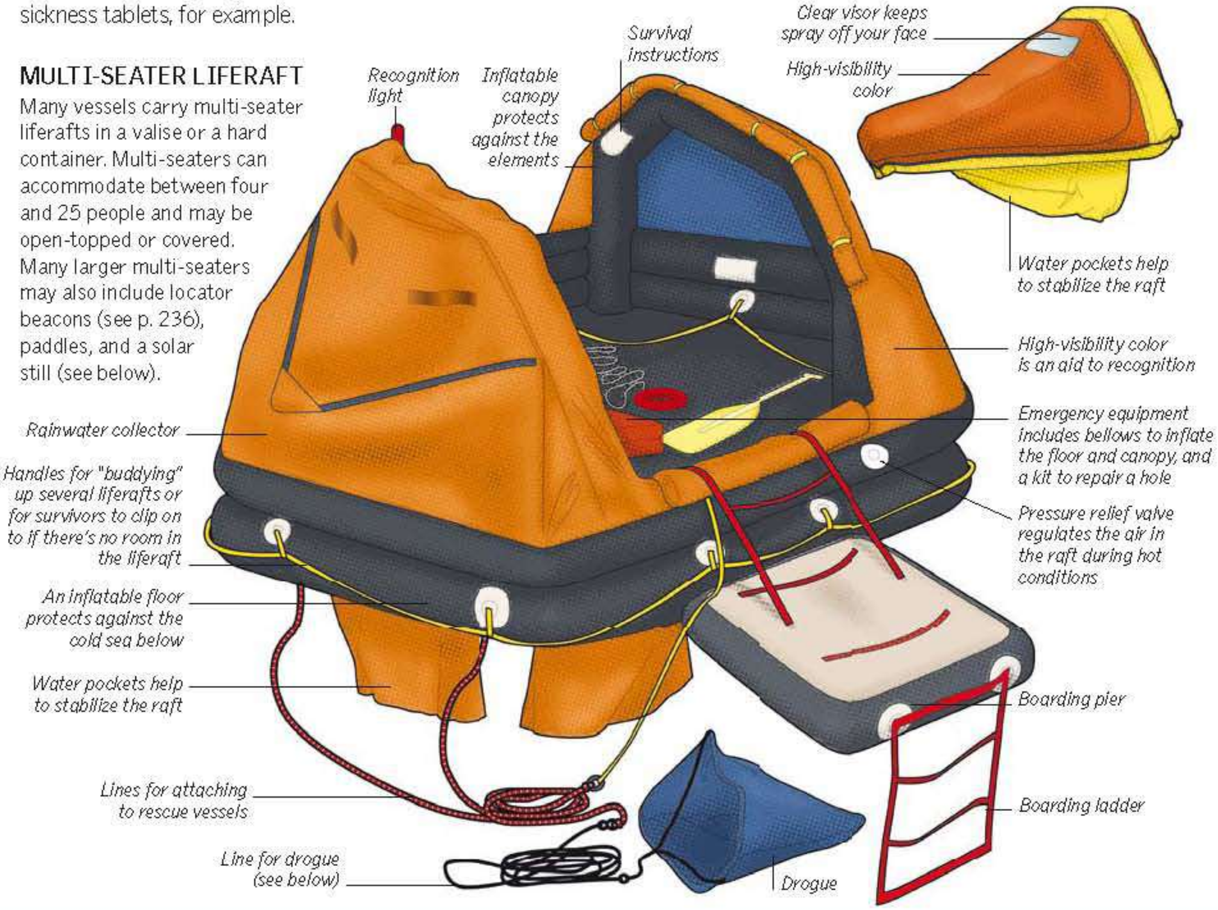
Clear visor keeps spray off your face
High-visibility color
Water pockets help to stabilize the raft
Survival instructions
Recognition light
Inflatable canopy protects against the elements
High-visibility color is an aid to recognition
Emergency equipment includes bellows to inflate the floor and canopy, and a kit to repair a hole
Rainwater collector
Handles for "buddying" up several liferafts or for survivors to clip on to if there's no room in the liferaft
Pressure relief valve regulates the air in the raft during hot conditions
An inflatable floor protects against the cold sea below
Water pockets help to stabilize the raft
Boarding pier
Lines for attaching to rescue vessels
Boarding ladder
Line for drogue (see below)
Drogue

SOLAR STILL

A solar still is a light, compact, and easy-to-use device for producing drinking water from sea water. Heat from the sun evaporates the salty water inside the still. Condensation on the walls collects in a channel around the rim and is directed to a store. Depending on the prevailing conditions and the availablilty of sunshine, a solar still can produce as much as 2 liters (3½ pints) of fresh drinking water a day.

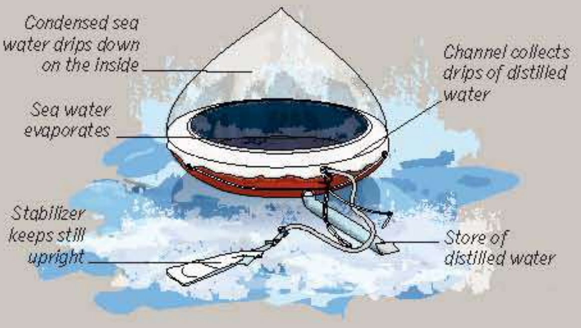
Condensed sea water drips down on the inside
Channel collects drips of distilled water
Sea water evaporates
Stabilizer keeps still upright
Store of distilled water

DROGUE

A sea anchor, known as a "drogue," is a critical piece of equipment because it helps to reduce drift and keeps the liferaft stable and seaworthy, particularly in heavy seas.

STEADYING A RAFT

A drogue stops a liferaft from overturning by creating drag that "anchors" the trailing side of the raft to the water.

- A drogue can position a liferaft either downwind or 90 degrees to the wind.
- It helps to keep the raft near the ditching location, thereby improving the chance of rescue. Even in a 2-knot current a liferaft can drift 50 miles (80 km) a day.

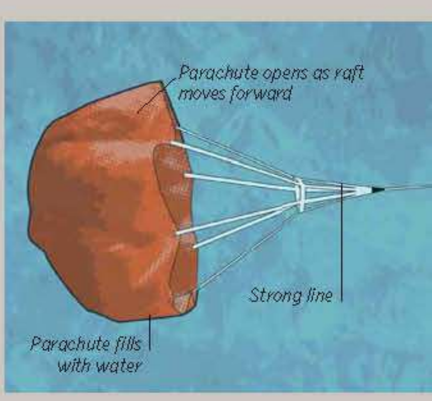
Parachute opens as raft moves forward
Strong line
Parachute fills with water

ABANDONING SHIP

Unless your vessel is an immediate danger to you, don't abandon it until you really need to. Even a badly damaged vessel can provide you with protection from the elements, equipment such as radios and flares, and provisions such as water and food. In addition, it is a big target for rescuers to spot. Make every effort to keep the vessel afloat.

THE DANGERS THAT AWAIT

When you do decide to abandon ship, you will be faced with the following dangers. If you have prepared yourself properly before heading out to sea, you will—in many cases, at least—be able to deal with them.

- Hypothermia brought on by inadequate clothing and exposure to wet, windy, and rainy conditions.
- Drowning because you have no lifejacket.
- Dehydration caused by a lack of water or an injury.
- Malnutrition due to a lack of food.
- Cold shock (see p. 254) due to sudden immersion in cold water.

BEFORE ABANDONING SHIP

Where time and circumstances permit, don't abandon your ship until absolutely necessary. Send a May Day signal with your name, position, group size, physical condition, and circumstances. Then carry out the following measures:

- Activate the 406 EPIRB (see p. 236).
- Ensure everyone on the vessel wears layers of clothing, and has a survival suit and lifejacket (see p. 250).
- Check the liferafts are ready for launch.
- Get the grab bag (see p. 250).
- Prepare to take the jerry-cans of water on board the liferafts.
- Fill spare containers with water.
- Gather up as much food as you can.

TAKING TO THE LIFERAFT

If you have to abandon ship, make every effort to launch all available liferafts. Even the craft that are not used will make it easier for search and rescue teams to detect the "footprint" of the survivors. They also contain additional supplies of water, flares, and other useful items.

- Many liferafts have a painter. This is a line that attaches the liferaft to the vessel to ensure it does not blow away when thrown overboard and inflated.
- Should the vessel sink, the painter has a weak link that breaks under pressure, or you can cut it.

PREVENTING HYPOTHERMIA

Getting wet greatly reduces your survival chances. At 41°F (5°C), a normally dressed person has only a 50 percent chance of surviving for one hour. You are six times more likely to survive in protective clothing.

- Climb into the liferaft carefully so that you don't get wet.
- Wear several layers of clothing to trap air. Even wet layers retain some heat around the body. Don't forget your head, hands, and feet.
- Wear a survival suit (see p. 250) to increase your survival time.

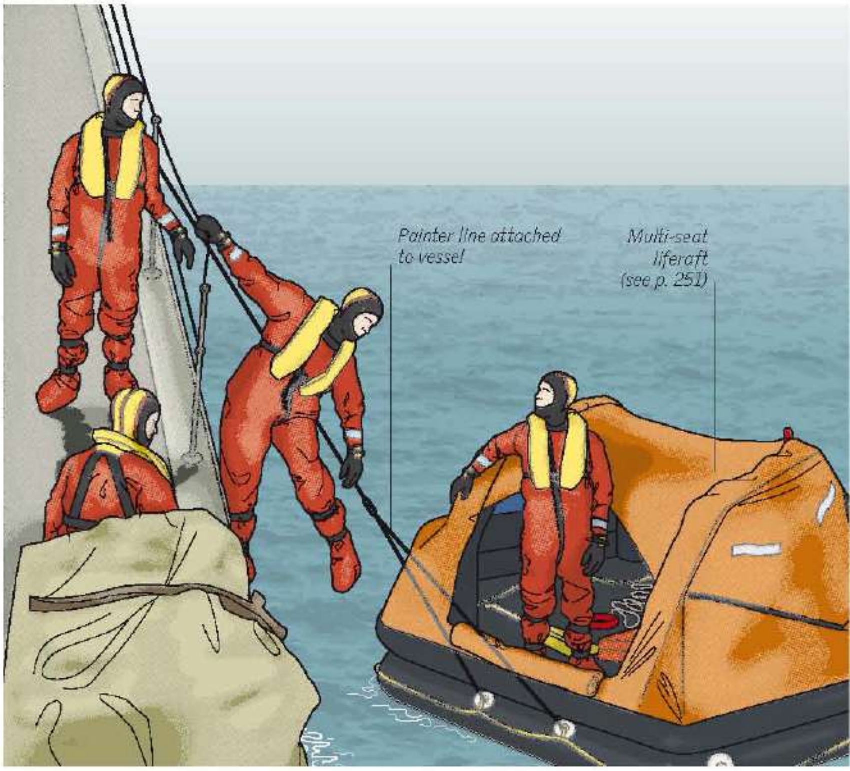
Painter line attached to vessel
Multi-seat liferaft (see p. 251)

① Move slowly and deliberately into the liferaft, keeping out of the water, and trying to stay as dry as possible. If you have to enter the water, climb down or lower yourself into it, rather than jumping.

- Look for the Day-Glo instruction patch inside the liferaft. Follow the recommended advice under "Immediate Actions" (see panel, opposite).

IN THE LIFERAFT

Once you have successfully negotiated the tricky procedure of abandoning ship and climbing into the liferaft, there are a number of actions you need to take. These are prioritized into immediate, secondary, and subsequent.

IMMEDIATE ACTIONS

- Inflate the floor of the liferaft with the bellows, while carrying out a roll call of your group to check for missing members.
- If the vessel is still afloat, keep the raft attached via the painter. Someone should be ready to cut the painter close to the vessel in case it starts to sink.
- Once clear of the vessel, set up the drogue (see p. 251).
- In difficult weather conditions, close the entrances of the raft in order to keep in heat, and keep out wind, rain, seawater, or spray.
- Bale out water, check for leaks, use the sponge to dry the liferaft, and use the leak stopper and clamps if necessary.

SECONDARY ACTIONS

- Treat the injured.
- Take anti-seasickness tablets.
- Post lookouts.
- Bring liferafts together.
- Warm up as best you can.

WARNING!

Never drink seawater, because the salt it contains will increase your rate of dehydration. If you're in hot conditions without any water, dehydration can set in within one hour.

SUBSEQUENT ACTIONS

- Delegate a leader based on experience.
- Find out who has useful skills, such as first aid and sea survival training.
- Ask everyone to look for sharp objects that could puncture the liferaft. At the same time, ask them if they have anything that could have a survival use.
- Find out what survival aids you have. Prepare aids to location (see p. 236) and show everyone how to operate them.
- Establish a routine and detail a watch system for inside and outside the liferaft.
- Delegate one member of the group to make repairs, another to keep the rations, and another to administer first aid.
- Establish how much water and food you have and start to ration in accordance with your particular situation.
- Procure water as soon as possible—don't wait until you need it. Deploy solar stills (see p. 251) and use the reverse-osmosis pump (see p. 197).

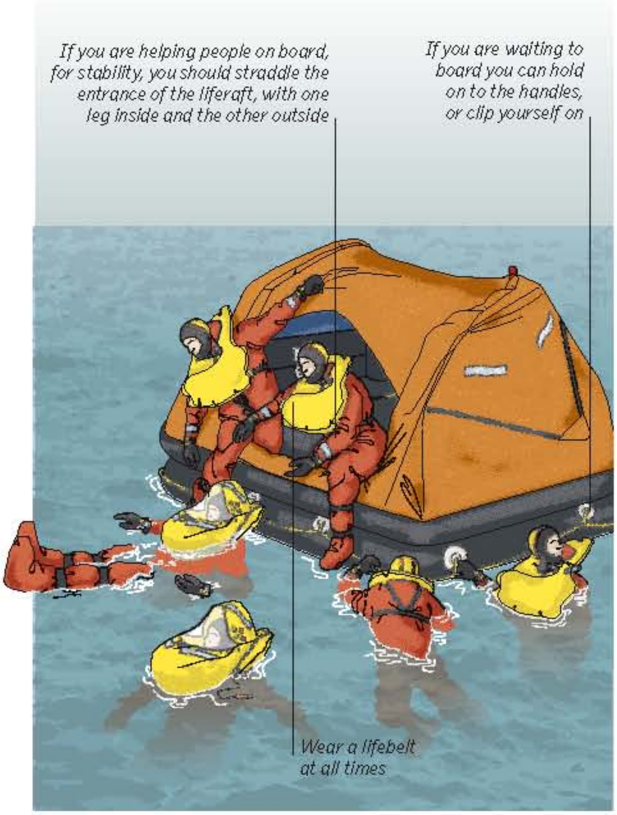
If you are helping people on board, for stability, you should straddle the entrance of the liferaft, with one leg inside and the other outside
If you are waiting to board you can hold on to the handles, or clip yourself on
Wear a lifebelt at all times

2 If your group does end up in the water, the first two people on board the liferaft should help others on, one at a time, lifting them under the arms.

- Don't overload the raft. Those least vulnerable can hang on to the raft's handles, or tie their lifejacket lifeline to one.

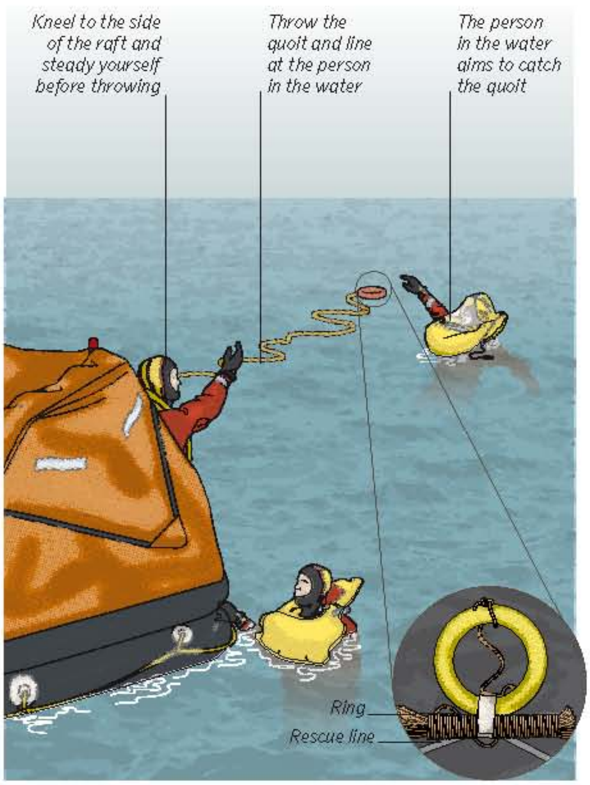
Kneel to the side of the raft and steady yourself before throwing
Throw the quoit and line at the person in the water
The person in the water aims to catch the quoit
Ring
Rescue line

3 Look for anyone who is struggling to make their way toward you. Use the rescue line and ring to pull them to safety in the liferaft.

- Avoid entering the sea unless you have to rescue an unconscious person from the water.

TAKING TO THE WATER

About two-thirds of people who drown in open water were within 10 ft (3 m) of a safe refuge and 60 percent of them were "good" swimmers. Here are some useful tips to prevent drowning.

JUMPING INTO THE WATER

Abandoning ship can be a dangerous procedure. If you have to enter the water, climb down using ropes or nets. If you have no option but to jump into the sea, these straightforward steps should improve your chances of survival. Jumping into the water is a last resort, as you will be prone to cold shock (see box, right) and at immediate risk of hypothermia and drowning. If you have to do it, first ensure your lifejacket is fitted correctly and check the surface of the water to find a safe entry point. Watch out for people, debris, and burning fuel. If you can't avoid the fuel, swim under it and, before coming up for air, poke your hand through the surface to ensure you're clear of it. As you come up, keep your face down to protect your nose, eyes, and mouth.

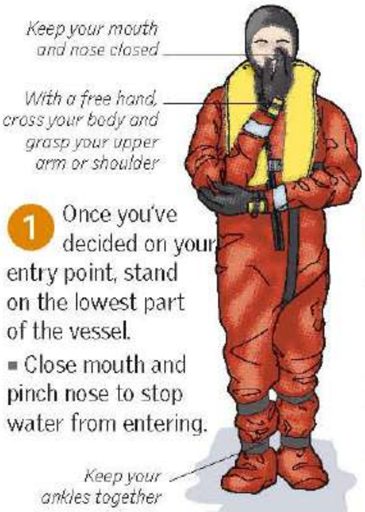
Keep your mouth and nose closed
With a free hand, cross your body and grasp your upper arm or shoulder
Keep your ankles together

1 Once you've decided on your entry point, stand on the lowest part of the vessel.

- Close mouth and pinch nose to stop water from entering.

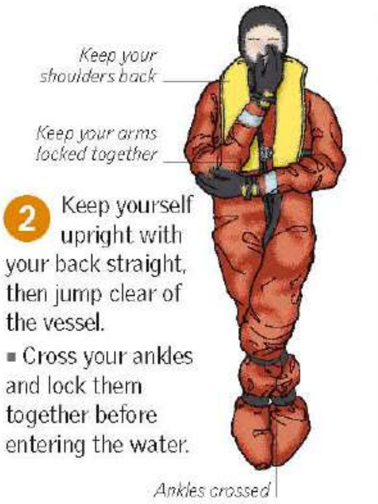
Keep your shoulders back
Keep your arms locked together
Ankles crossed

2 Keep yourself upright with your back straight, then jump clear of the vessel.

- Cross your ankles and lock them together before entering the water.

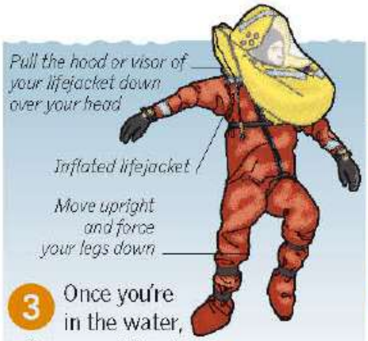
Pull the hood or visor of your lifejacket down over your head
Inflated lifejacket
Move upright and force your legs down

3 Once you're in the water, inflate your lifejacket.

- If you're in a survival suit, raise your arms and gently pull one seal away from your wrist to let the excess air escape.

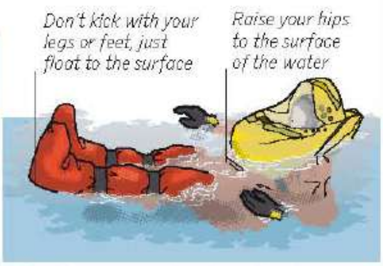
Don't kick with your legs or feet, just float to the surface
Raise your hips to the surface of the water

4 Lie on your back and swim as calmly as you can away from hazards, such as burning fuel.

- Push your head back onto the neck of the lifejacket to help raise your hips.
- Keep your feet and knees together, and use a backward butterfly stroke to swim toward the liferaft.

COLD SHOCK

Cold shock response is the body's reaction to immersion in very cold water—for example, if abandoning a ship in winter in the North Atlantic Ocean. It is a common cause of death in such circumstances. Symptoms include gasping for breath and hyperventilation, which can lead to the inhalation of water, disorientation, panic, and the possible onset of hypothermia. A sudden increase in blood pressure and heart rate can cause cardiac problems in some people. Prolonged immersion in water will make it hard to perform physical movements, making swimming, climbing aboard a liferaft, or firing a flare extremely difficult.

- Body type or mental conditioning can help some people survive swimming in icy water.
- Dressing in layers and/or wearing a survival suit improves your chances of survival.
- Avoid entering the water if you possibly can.

IF YOU HAVE NO LIFERAFT

Even without a liferaft, you stand a better chance of surviving at sea if you're in a group. More survivors create a larger target for rescuers to see, and being with others can be good for morale.

- Collect any floating debris before it disappears with the current, as this can increase your "footprint" so that rescuers can see you more easily.
- Determine what location aids and other equipment are available and prepare these for use.
- If there are children or injured among your group, place them in the center of the group and huddle close together.
- If you are alone, get into a position known as H.E.L.P. (see below), which stands for "Heat Escape Lessening Position." This will help to keep heat within the core of your body.

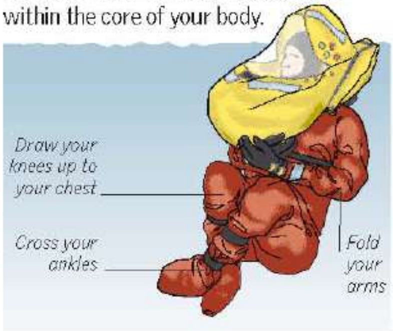
Draw your knees up to your chest
Cross your ankles
Fold your arms

FLOATING FACE DOWN

Don't panic if you have to enter the water without a lifejacket. Your body's natural buoyancy will keep at least the top of your head above water. To keep your face above water, too, make small windmill motions with your arms extended. However, if the water's rough, your only option may be to float face down.

STAY POSITIVE

It's very important to remain positive and to focus on your situation right now rather than on what might happen in a few hours or tomorrow.

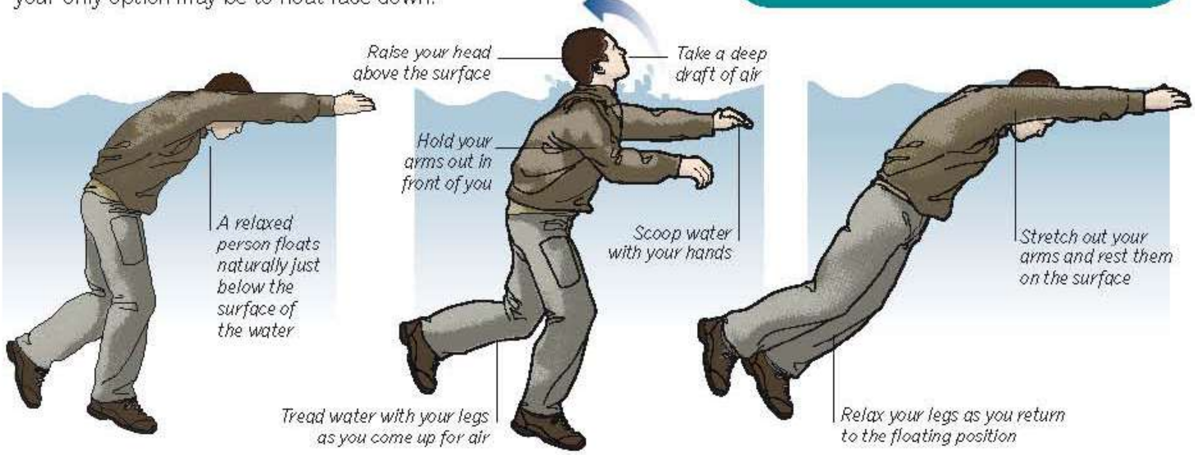
A relaxed person floats naturally just below the surface of the water
Raise your head above the surface
Take a deep draft of air
Hold your arms out in front of you
Scoop water with your hands
Tread water with your legs as you come up for air
Stretch out your arms and rest them on the surface
Relax your legs as you return to the floating position

1. It's important to relax, even though this seems hard when your life is in danger.
   - Let your face lie in the water, and put your arms out in front of you.

2. Begin to exhale into the water as you raise your head up.
   - Lift your head as it breaks through the surface and empty your lungs completely before inhaling more air.

3. With fresh air in your lungs, duck your face back into the water, keeping your mouth closed.
   - Let your body float again, before you repeat the sequence.

IMPROVISING A BUOYANCY AID

If you're in the water and wearing pants, you can improvise a buoyancy aid that will help you to keep your head above water. Initially, it may be awkward to organize, but the benefit outweighs the effort involved.

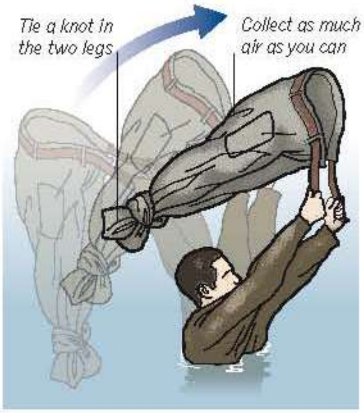
Tie a knot in the two legs
Collect as much air as you can

1. Take off your pants and tie the legs together near the bottom. Tighten the knot as much as you can with your teeth.
   - Flick the pants over your head from behind until they fill with air. Tread water as you do this.

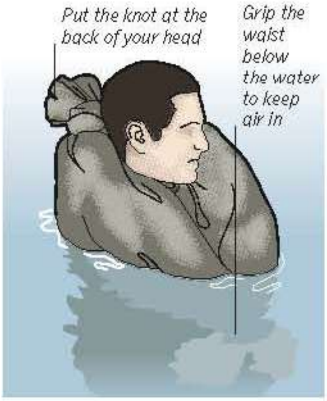
Put the knot at the back of your head
Grip the waist below the water to keep air in

2. Quickly catch the waist with your hands and grip it tight to hold air in.
   - Put your head between the trouser legs and float.
   - You will need to repeat the process regularly.

SECONDARY DROWNING

A person who nearly drowns inhales water that can cause potentially fatal chemical and biological changes in the lungs. This is known as "secondary drowning." It can be caused by inhaling fresh or salty water—even as little as 1fl oz (30 ml). Secondary drowning may occur 24–72 hours after immersion and, while uncommon, it's not rare. Take the following precautions to guard against secondary drowning:

- Monitor anyone who has nearly drowned or has spent some time in water.
- Ask the person to take a couple of deep breaths and check for any pain or discomfort that can't be accounted for.
- Look for symptoms such as coughing, breathing difficulties, chest pain, and saliva that looks like foam.
- Make sure the person is sitting in an upright position.
- If possible, provide oxygen—the blood levels of oxygen can fall quickly—and lots of rest and reassurance.

FIRST AID

FIRST AID

Regardless of the events that put you in a survival situation, a major factor that will dictate your options and actions will be whether anyone has been injured. Someone's survival may depend on the treatment they receive at the time of the incident, and during the ongoing survival ordeal, so it's essential that everyone should be proficient and confident in basic first aid techniques. In a true survival situation, the term "seek medical help" really means "deal with it yourself." While it would be impossible to be able to treat all injuries—a task even a paramedic would find daunting without the right equipment—the vast majority can be either dealt with or stabilized using a combination of basic first-aid techniques and common sense.

The "Protection" element of the survival principles (see p. 27) applies to your situation at all times—you should continually look at the consequences of every action in relation to protection against injury. Prevention is always a better option than cure. In hot climates, being able to recognize the first stages

In this chapter YOU WILL DISCOVER...

- that **maggots** have their **uses**...
- how to **improvise goggles** to prevent **snow blindness**...
- when to **stop**, **drop**, and **roll**...
- how to treat **snake bites** and **jellyfish stings**...
- when to **eat charcoal** or drink **bark tea**...
- how to prevent **frostnip** from becoming **frostbite**...
- when to **perform** a **firefighter's lift**...

of heat stress will allow you to take action with them before they develop into life-threatening dehydration and heatstroke. In cold climates, being able to recognize the first stages of frostnip will allow you to stop it from becoming frostbite. In many cases, further injury can be avoided by adopting the basic principles detailed throughout this book, such as being able to assess the best course of action in a given situation, plan a route and move safely over terrain, and protect yourself against the elements.

The will to survive is often the only factor that determines whether you live or die—regardless of your equipment, training, knowledge, and skills.

WHEN FACED WITH everything that man and nature can throw at you, when there appears to be no hope, you'll be faced with two choices: will you accept your situation and wait and see what hand fate deals you, or will you endure the pain and discomfort and fight for you own survival?

This determination was clearly shown by Aron Ralston during a hike in Blue John Canyon, Utah. After accidentally dislodging a 800 lb (363 kg) boulder, which pinned his right arm, he was faced with a bleak outlook. After five days, aware that no one knew he was missing and having run out of water, Ralston decided to amputate his own arm with his knife, applied a tourniquet, and hiked off to safety.

In another famous example, Simon Yates took the agonizing decision to cut the rope that held his injured climbing partner, Joe Simpson (who Yates thought had died), over a crevasse in the Peruvian Andes. Incredibly, Simpson survived the fall and chose to endure. He spent three days without food and only splashes of water from melting ice as he crawled and hopped the 5 miles (8 km) back over frozen mountainous terrain to reach their camp.

> “The **vast majority** of **injuries** can be either **dealt with** or **stabilized** using a **combination** of **basic first aid techniques** and **common sense**”

FIRST AID ESSENTIALS

SAFETY IS KEY on any expedition. Before you set off, make sure that you and everyone in your party has the necessary medical equipment—especially medications. If someone is injured, he or she should be treated immediately. If you can't contact the emergency services (and if there are enough of you), one person should stay with the casualty while two others go for help.

WARNING!
Protect yourself from danger at all times. You can't help anyone if you become a casualty as well. If the area is unsafe, don't approach the casualty—get emergency help and monitor the casualty's condition from a safe distance.

BASIC FIRST AID KIT

Keep your first aid kit (see panel, opposite) dry and readily accessible. Check the seals on sterile dressings; if they're not intact they're not sterile. Replace anything you use as soon as you can.

PRIORITIES AFTER AN INCIDENT

Assess a situation quickly and methodically. Find out what happened. Check casualties for life-threatening conditions such as unconsciousness or severe bleeding (see p. 264) and treat those first.

- **Response** Is the casualty conscious or unconscious? If the casualty is responding to you, he or she is conscious. Shake the shoulders gently if you're not sure.
- **Airway** Is it open and clear? If the casualty can talk, it is. If he or she is unconscious, open and clear it (see p. 276).
- **Breathing** Is it normal? Treat any difficulty such as asthma (see p. 275). If he or she is unconscious and not breathing, call for emergency help and begin CPR (see p. 277).
- **Circulation** Are there any signs of severe bleeding? If so, treat immediately. Once life-threatening conditions are under control, you can make a more detailed assessment. Examine the casualty methodically from head to toe. Ask how the incident occurred, as it can indicate likely injuries.

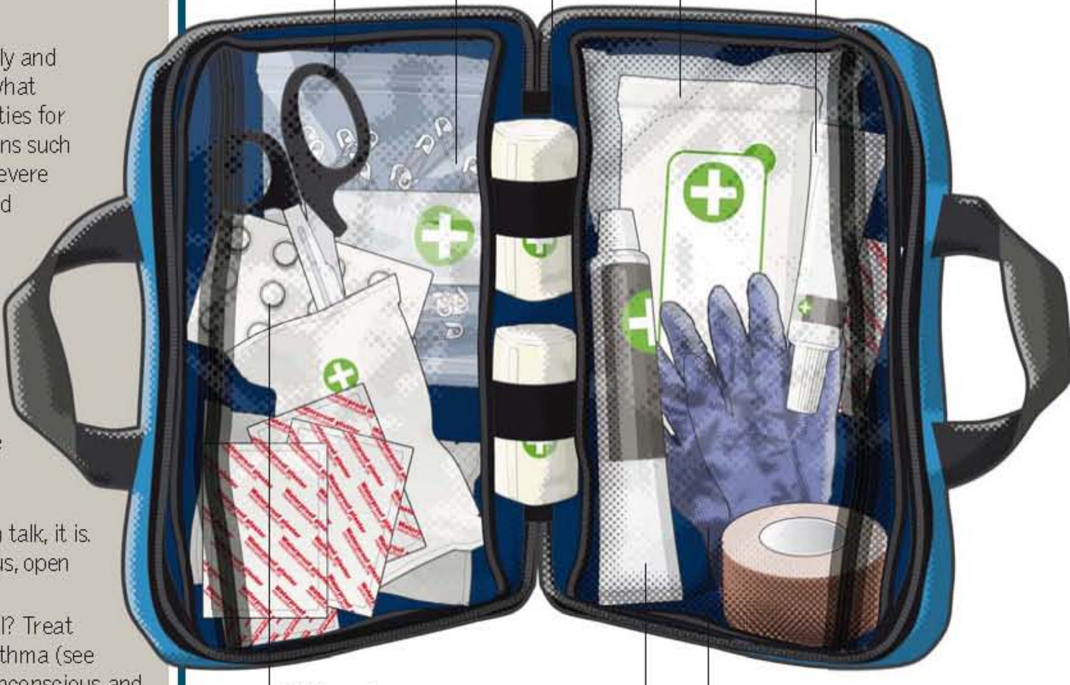
Lightweight carrier

Combined sterile dressings

Antibiotic eye ointment

Large scissors are useful for cutting clothing

Safety pins for securing bandages

Tablets such as antihistamine and paracetamol

Antiseptic cream

Disposable gloves

Take fabric, waterproof, and hypoallergenic plasters

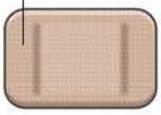

Gauze roller bandage to secure dressings—can be washed and reused

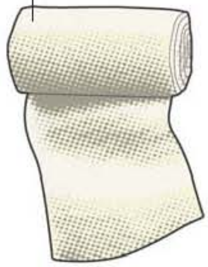

Zinc-oxide tape can be used to secure dressings

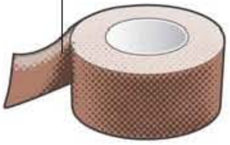

IMPROVISED SLINGS

Hand, arm, or shoulder injuries need to be immobilized and supported in a raised position. If you don't have a triangular bandage with you, use a piece of strong cloth about 3 ft (1 m) square, folded in half to form a triangle (see p. 270). You can also use your jacket or even rucksack straps. The jacket corner "sling" is the only one that provides enough support for a hand, wrist, or forearm injury. Ask the casualty to support the arm with his or her other hand while you secure the sling.

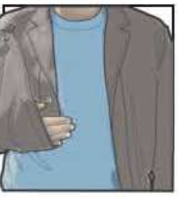

JACKET CORNER

To support an injured forearm or hand, fold the jacket up over the arm and pin it.

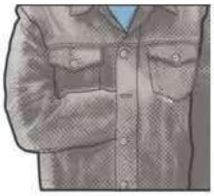

BUTTON-UP JACKET

Undo one of the buttons and slide the injured arm into the opening for support.

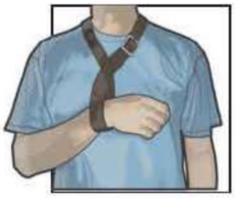

BELT SUPPORT

Support an upper arm injury in a raised position with a belt looped into a figure-eight.

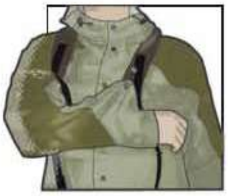

PINNED SLEEVE

Pin a sleeve to a jacket or the strap of a backpack for support.

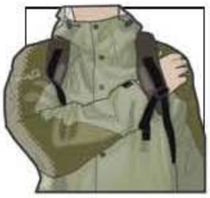

SHOULDER STRAP

Rest a sprain by tucking your hand in your backpack strap.

PROTECTING AGAINST INFECTION

Disposable gloves prevent cross-infection between you and the casualty; they must be latex-free, since contact with latex can cause an allergic reaction. Antiseptic wipes are also invaluable when cleaning wounds.

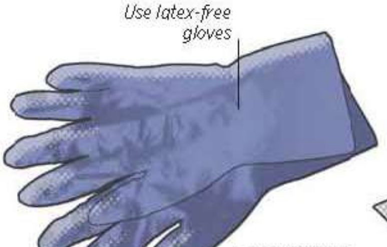

LATEX-FREE DISPOSABLE GLOVES

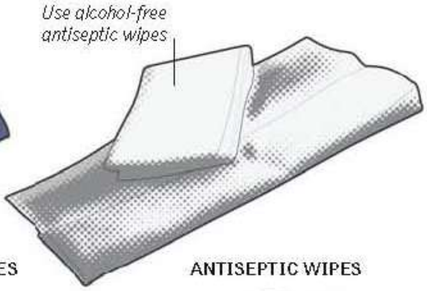

ANTISEPTIC WIPES

COMBINED STERILE DRESSING

This is a sealed dressing consisting of a pad attached to a bandage. It's easy to apply, and can be used as a sling. It should be taped on to your backpack strap for easy access in an emergency.

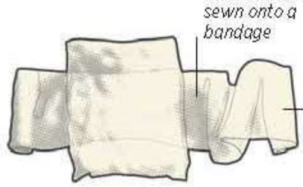
Sterile pad is sewn onto a bandage

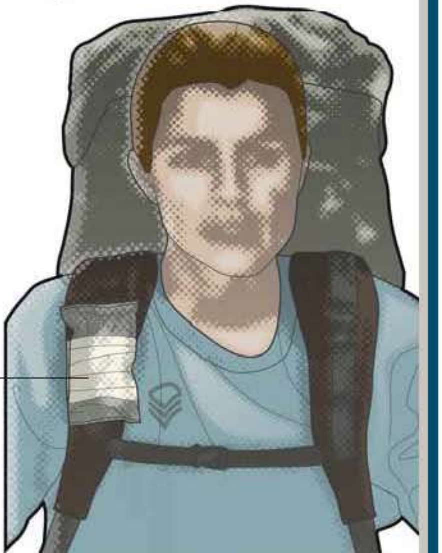

CHECK LIST

Make sure that your first aid kit and medicines are suited to the environment you will be visiting.

BASIC KIT

- Alcohol-free antiseptic wipes
- Latex-free disposable gloves
- Alcohol gel for handwashing
- Antiseptic cream
- Antibiotic eye ointment
- Adhesive dressings—fabric, waterproof, and hypoallergenic
- Gel blister bandages
- Combined sterile dressings, or sterile pads and bandages in assorted sizes
- Roller bandages—take self-adhesive for supporting joints and gauze for securing dressings
- Two triangular bandages
- Micropore or zinc oxide tape
- Scissors and tweezers
- Safety pins
- Disposable syringes

PERSONAL MEDICATION

- Painkillers
- Anti-inflammatories
- Medical alert bracelet/pendant
- Prescription medicines such as asthma inhaler and/or adrenaline (epinephrine) autoinjector
- Antihistamines
- Anti-diarrhea medicine
- Packets of oral rehydration salts
- Hydrocortisone cream

ENVIRONMENT-SPECIFIC EXTRAS

- Malaria tablets
- Mosquito repellent
- Anti-poison-ivy cream
- Sunblock
- Tick remover
- DEET powder for removing leeches

FLESH WOUNDS

ANY INJURY THAT BREAKS THE SKIN carries a risk of infection as germs can enter the body. These can come from the cause of injury, air, dirt, or clothing embedded in the wound. In the wilderness, keeping the injured area clean can be a real challenge, but it is essential. Tetanus is a potentially lethal infection caused by bacteria that live in soil. It can be prevented by immunization, so ensure that your vaccinations are up to date.

MAGGOT THERAPY

If a wound is seriously infected and you have no antibiotics, expose it to flies for one day, then cover. Maggots will develop and eat any dead tissue. Check daily, and flush the maggots out of the wound with sterile water before they start to eat healthy tissue.

BLISTERS AND BRUISING

A blister is a fluid-filled "bubble" of skin that occurs when skin is rubbed repeatedly against a surface (a friction burn). A bruise is bleeding into the skin and surrounding tissue from a blow that does not break the surface of the skin.

HOW TO TREAT BLISTERS

The ideal treatment for a blister is to rest and wait until it has healed, but this may not be possible in a survival situation. Cover it with a gel blister dressing if you have one or, if the blister is large, you may have to pierce it to enable you to continue walking. However, never pierce a blister caused by a burn—you risk infection.

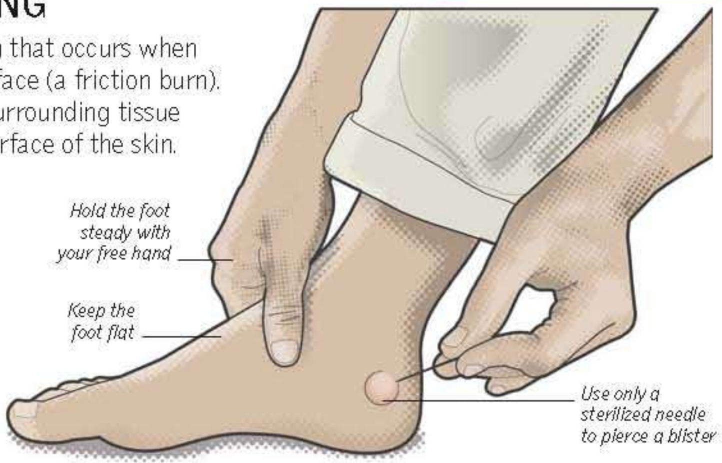
Hold the foot steady with your free hand
Keep the foot flat
Use only a sterilized needle to pierce a blister

1 If you have to burst a blister, first sterilize a needle by holding it in a flame until it's red hot, then letting it cool.

- Clean the area with water or wipes, pat dry, then pierce the edge of the blister.

2 Carefully apply pressure to the side of the blister opposite the hole made by the needle. Continue to apply pressure until all of the fluid has been squeezed out.

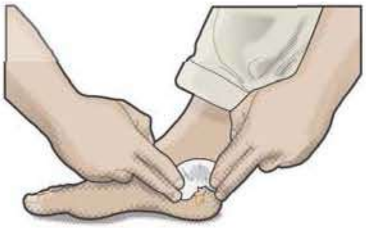

3 Gently clean and dry the blister, then protect the wound from infection by applying a dressing.

- Use a gel blister plaster or padded moleskin, if possible.
- Alternatively, use a pad held in place with zinc-oxide tape.

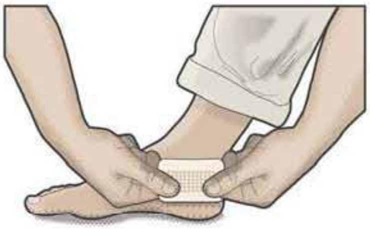

BLISTER PREVENTION

Following these simple rules when out hiking should prevent blisters from developing in the first place:

- Ensure boots or shoes fit properly and are well "broken in" before you set out.
- Always wear clean, dry, comfortable socks next to your skin. Avoid wearing two pairs as they may bunch up, increasing friction.
- Keep toenails cut short and straight.
- Remove boots and air your feet during rest periods.
- If you feel a "hot spot," treat it before it becomes a problem: stop immediately and apply moleskin, a gel pad, or zinc-oxide tape.

HOW TO TREAT BRUISING

To reduce the swelling and pain of a bruise, raise the injured area, and apply a cold compress. If you have access to cold running water, ice, or snow, soak a cloth and hold it against the bruised area for at least ten minutes. Severe bruising may indicate a more serious injury, such as a broken bone (see p. 271) or internal injury (see Shock, p. 274), which will require immediate treatment.

WOUNDS AND BLEEDING

Severe bleeding can be distressing for both you and the casualty, but it can usually be controlled by a combination of direct pressure and elevation of the injury. Remain calm and reassure the casualty while treating the wound. Treat for shock (see p. 274).

FOREIGN OBJECT IN A WOUND

Any loose foreign objects, such as pieces of dirt or gravel, should be removed from a wound, otherwise they may cause infection or delay healing. Either rinse them off with cold running water or carefully pick loose pieces off the wound with tweezers. However, if an object is embedded in a wound don't remove it, but treat as shown below.

MINOR CUTS OR GRAZES

Any break in the skin, however minor, needs to be cleaned and protected from infection. Rinse the wound with clean cold water and pat it dry. Then cover the injury; the dressing pad must be larger than the wound. For small grazes, use a plaster; for larger ones, use a sterile pad and a bandage.

NATURAL WOUND DRESSING

The birch polypore, or razorstrop fungus, which grows on the sides of old birch trees, can be used as a natural wound dressing. Cut a thin slice from the top of the fungus and secure it over the wound. If you are in any doubt about the identity of the fungus, do not use it.

① Don't try to remove the object as it may be plugging bleeding. Control bleeding by pressing firmly on either side of the wound.

- Push the edges of the cut together but take care not to press directly on the object.
- Raise the wound above the level of the heart.

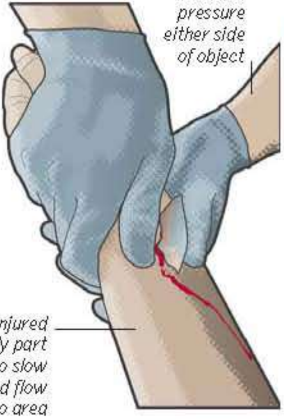

Maintain pressure either side of object

Keep injured body part raised to slow blood flow to area

② Place a piece of gauze over the object to protect it, then build up padding on either side (rolled bandages are ideal). Bandage over the pads and the object.

- Check the circulation beyond the bandage every ten minutes (see p. 264).

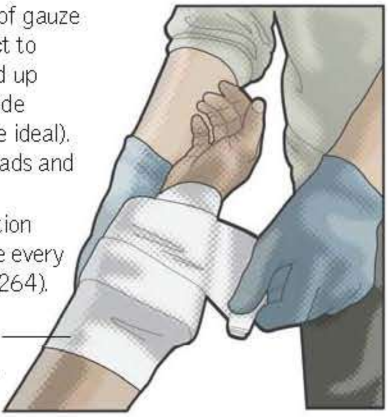

Bandage over the object to prevent further injury

TYPES OF WOUND

Different types of object and force produce different kinds of wound. It's useful to identify the type of wound you or the casualty has incurred so that the correct method of treatment can be applied.

BRUISE (CONTUSION)

A blunt blow will break blood vessels under the skin, causing blood to leak into the tissues. This results in a bruise: the skin is tender, swollen, and blue-black in color.

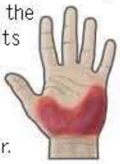

GRAZE (ABRASION)

A friction burn from a rope, or a sliding fall, will scrape off the top layers of skin, leaving a raw, tender area. Grazes often contain embedded foreign matter.

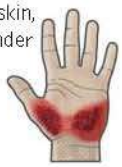

TEAR (LACERATION)

If the skin is torn open, the wound may not bleed as badly as an incised wound, but a larger area of tissue may be damaged and vulnerable to infection.

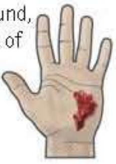

PUNCTURE WOUND

Sharp objects such as nails or sea urchin spines can puncture the skin. The entry hole will be small but the wound will be deep, with a high risk of infection.

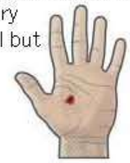

INCISED WOUND

If a sharp-edged object cuts across the skin, blood vessels will be sliced open and bleeding will be severe. Nerves or tendons may also be damaged.

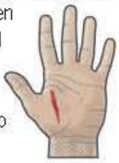

STAB WOUND

A penetrating wound from a long, bladed instrument is a very serious injury. A stab wound to the torso can damage vital organs and cause internal bleeding.

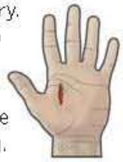

GUNSHOT WOUND

Check if the casualty has an exit wound. Treat entry and exit wounds separately.

ENTRY WOUND

Bullets drive deep into, or through, the body. They leave a small, neat entrance wound and cause serious internal damage and contamination.

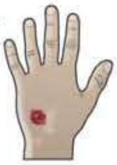

EXIT WOUND

If a bullet passes through the body, the exit wound will be large and ragged. Don't remove the bullet if there is no exit wound.

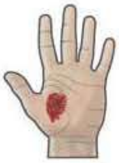

SEVERE EXTERNAL BLEEDING

Control the bleeding with direct pressure over a wound and, if possible, call for emergency help. Never use a tourniquet as this can cause severe tissue damage. Life-threatening shock is likely to develop if blood loss is severe (see p. 274).

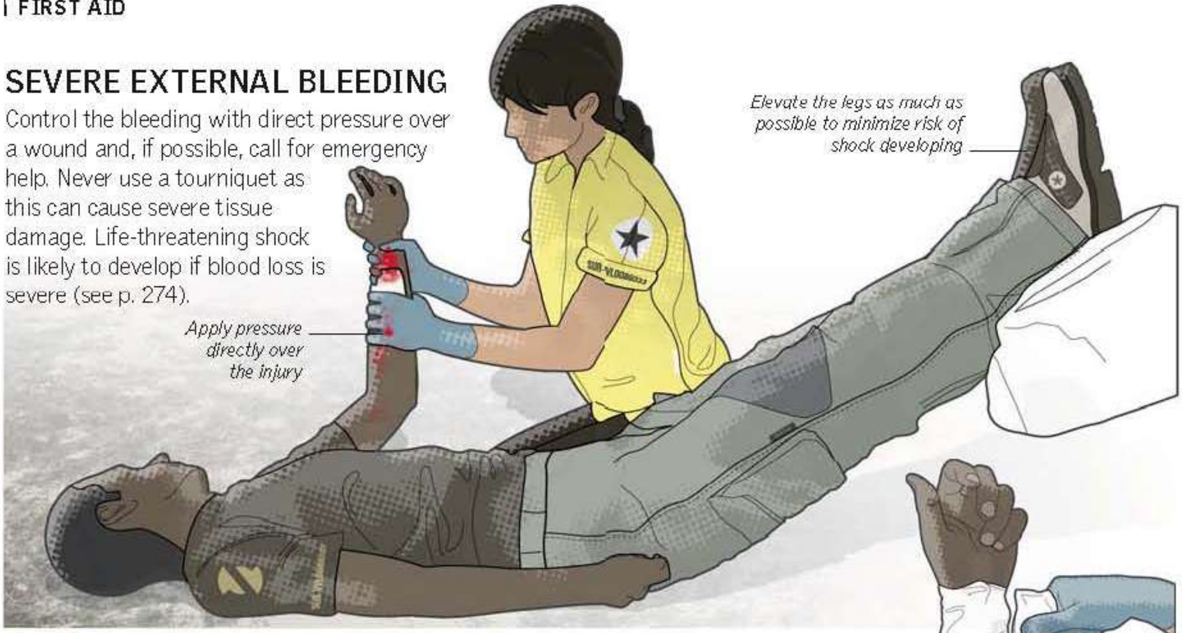

Apply pressure directly over the injury

Elevate the legs as much as possible to minimize risk of shock developing

1 Remove or cut away clothing to expose the wound. Apply direct pressure to the wound, over a sterile dressing or pad if you have one. Raise the injured limb above the level of the casualty's heart to reduce blood flow to the area. Help him to lie down and raise his legs.

2 Secure the dressing with a bandage. If blood soaks through, apply a second dressing on top of the first.

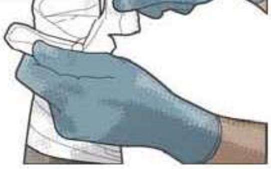

3 Every ten minutes, check the bandage is not too tight. Gently press a fingernail beyond the dressing. If the skin color does not return quickly, rebandage more loosely.

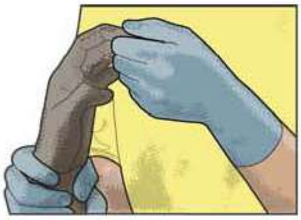

VARICOSE VEINS

When the one-way valves in veins fail, blood pools behind them, causing raised knobbly skin. The taut, thin-walled veins can easily be burst by a knock. Bleeding will be profuse.

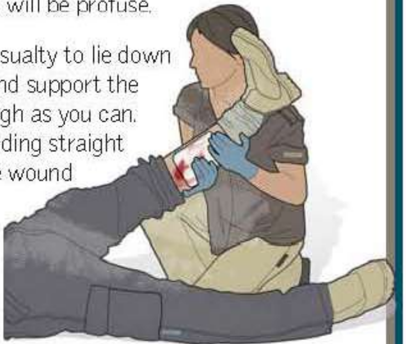

1 Help the casualty to lie down and raise and support the injured area as high as you can. This reduces bleeding straight away. Expose the wound and apply direct pressure over a sterile dressing or pad.

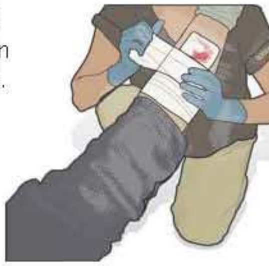

2 Bandage the pad firmly to maintain pressure on the wound. Keep the area raised. Check the bandage is not too tight (see step 3, above). If necessary, loosen it but maintain pressure.

SCALP WOUNDS

A scalp wound can bleed profusely, making it appear worse than it actually is. However, it may mask a more serious head injury. If a casualty becomes drowsy, has a headache, or double vision, get emergency help if possible.

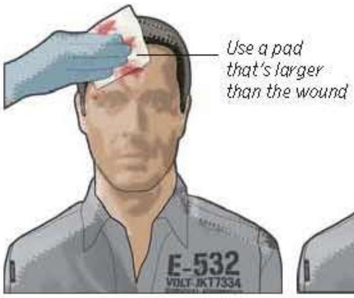

Use a pad that's larger than the wound

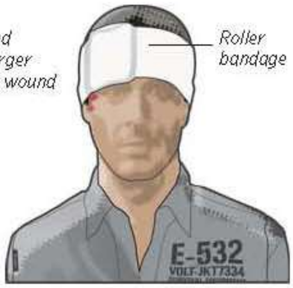

Roller bandage

1 Sit the casualty on the floor. Carefully replace loose flaps of skin, cover the wound with a sterile pad, and apply pressure.

2 Secure the dressing with a roller bandage. If the casualty doesn't quickly recover or if he deteriorates, get help.

EYE INJURIES

The eye can be seriously injured by a blow or contact with sharp objects, such as a tree branch, risking scarring, infection, or even loss of vision. Always wear eye protection when hiking through dense vegetation.

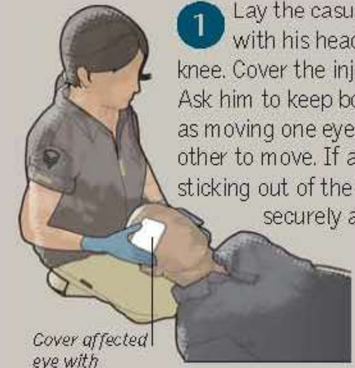
Cover affected eye with sterile pad

1 Lay the casualty down with his head on your knee. Cover the injured eye. Ask him to keep both eyes still, as moving one eye causes the other to move. If an object is sticking out of the eye, pad securely around it.

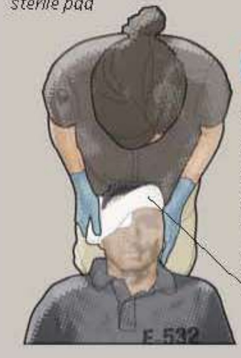
Secure pad with bandage

2 Secure the dressing with a bandage. If you're on your own, fix the dressing in place with plasters or tape and try not to move your eyes.

FOREIGN OBJECT IN THE EYE

If you can see an object on the surface of the eye, try lifting it off with the corner of a handkerchief, or wash it out with sterile water, washing away from the good eye. Don't remove anything that is sticking to the eye.

SNOW BLINDESS

Also known as "flash burn," this occurs if the surface of the eye is damaged through exposure to ultraviolet light, such as the glare from sun reflected off snow or water. Wearing sunglasses can prevent this. In an emergency, make your own goggles from cardboard or birch bark. If a person is affected by snow blindness, give him gauze pads to hold against his eyes. Bandage them in place if help is delayed.

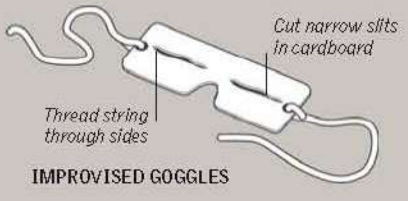
Cut narrow slits in cardboard

Thread string through sides

IMPROVISED GOGGLES

BURNS AND SCALDS

There is a serious risk of infection with all burns. Burns may affect only the outermost layer of skin, the upper layers, or the full thickness. Severe burns will affect all three layers. If the burn is larger than the casualty's hand it needs hospital treatment.

SMALL SUPERFICIAL BURNS

Cool the injury for ten minutes with cold running water or any cold, harmless liquid. Remove jewelry or watches from the affected area before it swells. Cover the burn to prevent infection.

LARGE OR DEEP BURNS

If a burn is extensive or deep, then fluid will be lost from the body and life-threatening shock is likely to develop (see p. 274). If the injury occured in a fire, the casualty could also have burns to his air passages so may have breathing difficulties. Don't burst any blisters as you will increase the risk of infection.

STOP, DROP, AND ROLL

If clothing is on fire:

- Stop moving
- Drop to the ground
- Roll over on the ground until flames are extinguished

1 Help the casualty lie down and protect him from the ground if possible. Cool the injury by dousing it with cold water for ten minutes; this also reduces swelling and relieves pain.

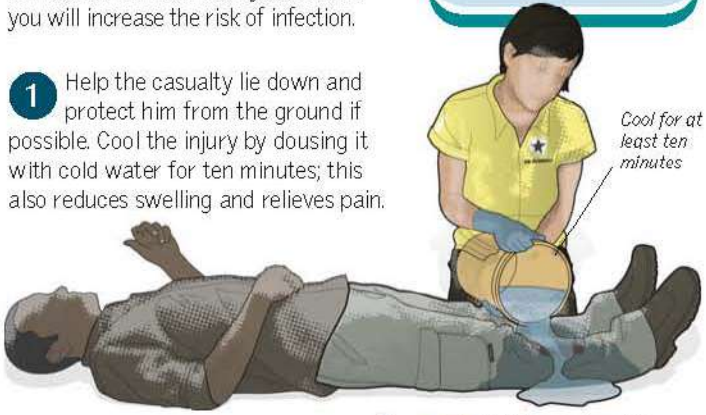
Cool for at least ten minutes

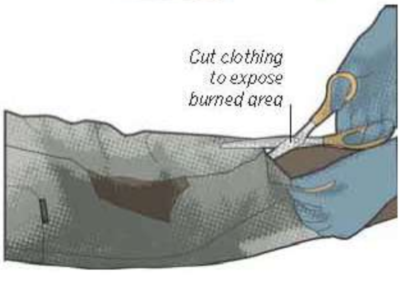
Cut clothing to expose burned area

2 While cooling the burn, remove or cut away clothing from around the burn. Don't touch or remove anything that's sticking to the burn.

3 Cover the entire area with plastic kitchen wrap, a clean plastic bag, or a clean, lint-free dressing. Get emergency help as soon as possible.

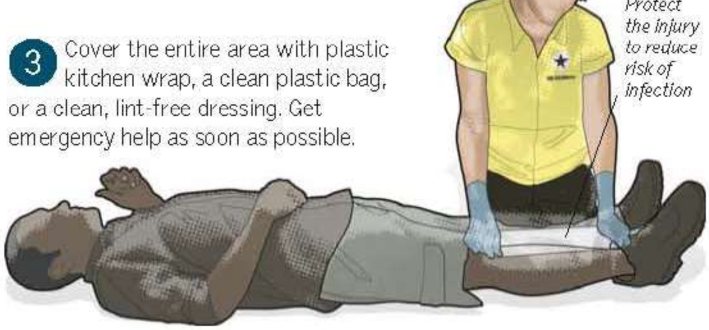
Protect the injury to reduce risk of infection

BITES AND STINGS

MANY BITES OR INSECT STINGS are painful, but most can be treated with simple first aid. However, there is a risk of an allergic reaction called anaphylactic shock. If a casualty develops a red blotchy rash, watery eyes or puffiness around the eyes, and/or breathing difficulties, get urgent medical help (see p. 274).

PORCUPINE QUILLS

If you are pierced by a porcupine's barbed quills, you must remove them or they will continue to work their way into the flesh and may pierce a vital organ.

- Cut off the ends of the quills, which are hollow, to allow them to deflate slightly.
- Pull the quills out, using pliers if you have them.
- Clean the puncture wounds and apply an antiseptic ointment.

TREATING BITES

An injury that breaks the skin carries a risk of infection. This is greatest with animal bites, since the mouth harbors so many germs. For a snake bite, you'll need to get emergency help, as the casualty needs to be transported on a stretcher in the treatment position to prevent the venom spreading through the body.

SNAKE BITES

Relatively few snakes are venomous, but it's generally safer to assume that all of them are. Most snakes cause a painful bite that leaves small puncture marks in the skin; however, the bite of a poisonous snake may in fact be painless. Symptoms also include nausea and vomiting, disturbed vision, and breathing difficulties.

1 Keep the casualty calm. Help her to lie down with her head, chest, and shoulders supported.

- Advise her to keep still.
- Make a note, if you can, of the time the bite occurred.

2 Apply a bandage around the site of the injury. Do not wash the area, remove any footwear or clothing, or attempt to suck the venom.

3 Tie another bandage around the affected limb that extends from the bite as far up the limb as possible.

- Check the circulation (see p. 264).
- Immobilize the affected limb.
- Move the casualty only with the use of a stretcher.

MAMMAL BITES

Bites from mammals—animals or humans—carry a serious risk of infection because the sharp teeth cause puncture wounds that transport bacteria deep into the tissue (see p. 263). As always, wear gloves to protect yourself. The bite can also crush surrounding tissue, and bleeding may be severe (see p. 264).

CLEAN WOUND

Raise the injury. Wash the wound with clean gauze and water. Pat dry, and cover it with a wound dressing.

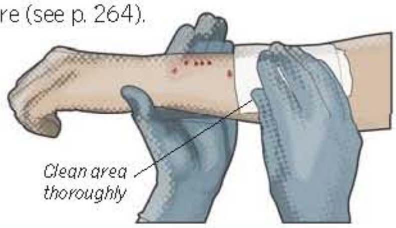

SNAKE IDENTIFICATION

Identify the snake if you can, since this will help medical teams find the correct antivenom. If you're not certain of the identity, make a note of its color or distinguishing features. If it's safe, put the snake in a secure container, but bear in mind that venom is active even if a snake is dead. Don't wash the venom off the bite; it can be used to identify the antivenom.

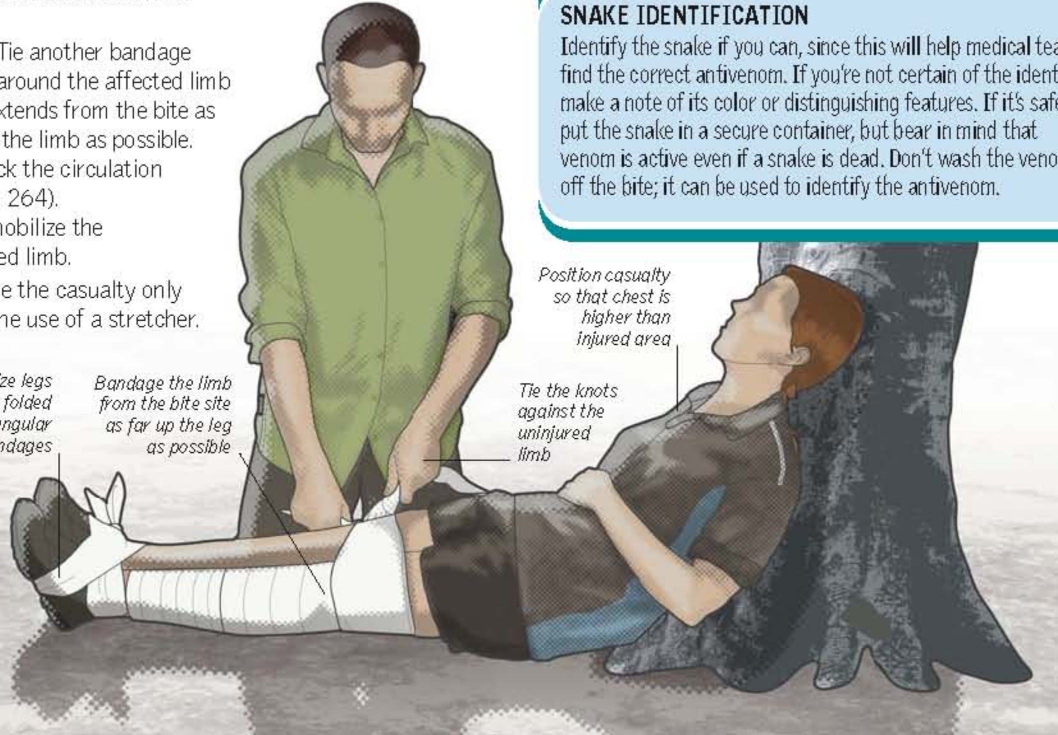
Immobilize legs with folded triangular bandages
Bandage the limb from the bite site as far up the leg as possible
Tie the knots against the uninjured limb
Position casualty so that chest is higher than injured area

TICK BITES

Ticks are tiny, spider-like parasites that live in grass or woodland. They feed on blood, attaching themselves to the skin with spiked mouthparts, and swell to the size of a pea. They carry Lyme Disease, so must be removed.

TRADITIONAL METHOD

Using tweezers, grasp the tick's head as close to the skin as you can. Pull the head upward using steady pressure (don't twist). Keep the tick in a container so it can be checked for Lyme Disease.

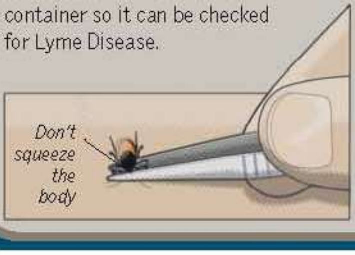

SPECIALIZED REMOVAL HOOK

Tick extractor tools that "unscrew" the mouthparts of the tick from the skin are available from pet shops. Slide the hook along the skin to grab the tick. Raise the hook very slightly and rotate it to lift the tick clear.

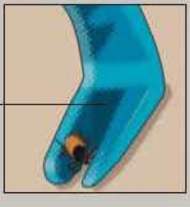
Slide the hook until it engages with the tick

HANDLING LEECHES

When in leech-infested territory, inspect your clothing and limbs every few minutes. Never pull a leech off if it's attached, as the jaws will remain in the skin and cause infection. To remove:

- Apply the juice of a raw lime, DEET, alcohol, or a dash of salt. If you're a smoker, put your cigarette butts in a piece of cloth, moisten the cloth, and squeeze nicotine onto the leech.
- Once the leech has dropped off, wash the area to remove the anticoagulant in and around the wound. Treat any bleeding (see p. 263) and cover the wound.

TREATING STINGS

Many stings are painful, but they are rarely life-threatening. Scorpion stings can be very painful and cause severe illness, and treatment should be as for snake bite (see opposite). Multiple insect stings of any type can produce a more serious reaction (see anaphylactic shock, p. 274).

WARNING!
A sting in the mouth or throat can cause swelling that blocks the airway. To minimize this, give the casualty cold water to sip.

INSECT STING

A bee, wasp, or hornet sting is often painful and followed by swelling and redness at the site. Some people have an allergy to stings, so monitor the casualty for signs of anaphylactic shock.

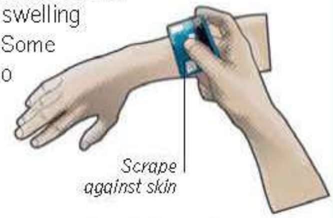

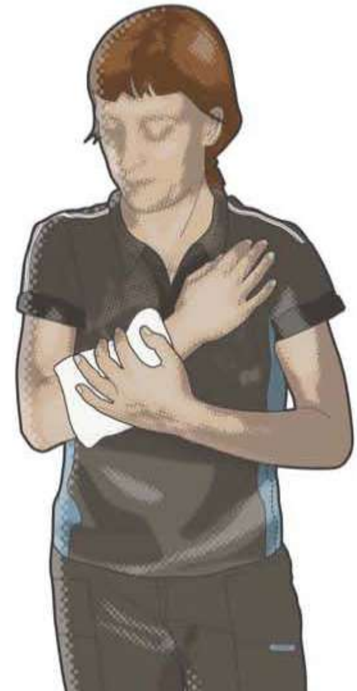

1 If the sting is visible, scrape it off sideways with the edge of a credit card, knife, or a fingernail.

- Don't squeeze the sac as you may squeeze more venom into the area.

2 Raise the affected part and place a cold pad against it for at least ten minutes to minimize swelling.

- Monitor the casualty for signs of allergy such as wheezing, or swelling around the face.

SEA CREATURES

When touched, sea anemones, corals, jellyfish, and Portuguese men-of-war release venomous cells that stick to the skin. Treat jellyfish stings as below; apply a cold compress to other stings to minimize swelling and relieve pain. Creatures such as weever fish have sharp spines that, if trodden on, become embedded, and may become infected.

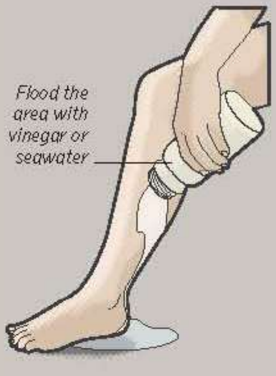
Flood the area with vinegar or seawater

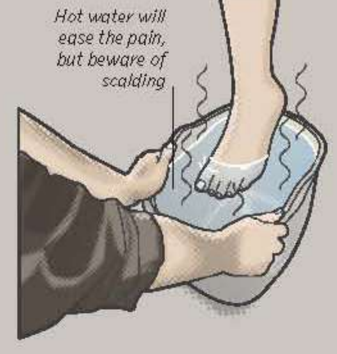
Hot water will ease the pain, but beware of scalding

JELLYFISH STING

Pour vinegar or seawater over the area to neutralize the sting. Help the casualty to sit down; treat as for snake bite (see opposite).

SEA URCHIN SPINES

Immerse the injured part in water as hot as the casualty can tolerate for about 30 minutes. Get medical help, as the spines must be removed.

POISONOUS PLANTS AND INTESTINAL PARASITES

IT'S ESSENTIAL TO MAINTAIN personal hygiene in camp, to purify all drinking water from natural sources (see pp. 198–201), and to observe the rules of food safety if you wish to remain healthy in the wilderness. You should also learn how to recognize and avoid the poisonous plants in the region you're traveling in.

CONTACT POISONING

Poisonous plants don't always have to be ingested to cause damage. If they come into contact with the skin, the result can be pain, swelling, redness, a rash, and itching. Get medical help fast and, in the meantime, rinse the skin with cold water for 20 minutes. If the water splashes an eye, rinse for ten minutes. Although plants are the most common cause of contact poisoning, chemicals such as camp fuel can also be harmful.

WARNING!

If you think you've handled a poisonous plant, don't touch especially sensitive parts of your body such as the eyes, mouth, or genitals until you have washed your hands thoroughly.

- If you develop blisters on your skin, don't scratch them—however much they itch. If you break the skin, you'll run the risk of getting an infection.
- As well as painful blistering, some people may have an extreme reaction to the toxin. They should be monitored and treated for shock (see p. 274). Get medical help as soon as possible.

POISONOUS PLANTS

Poison ivy, poison oak, and poison sumac are the most well known of the plants that contain urushiol, an irritant oil. If you damage the plant and get the oil on your skin, you must wash it off immediately with soap and cold water. You should also remove and wash any contaminated clothing to prevent the oil spreading. Apply an anti-poison-ivy cream as soon as possible. Many people will develop an itchy rash and often painful blistering within 4 to 24 hours of contact with the plant (see warning panel, above).

POISON SUMAC

Found in wet acid swamps in eastern North America, poison sumac can grow as tall as 18 ft (6 m).

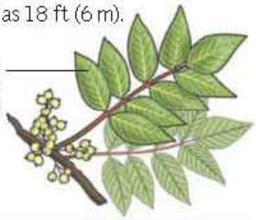
Oval leaflets grow in opposite pairs

POISON IVY

Native to wooded areas of North America, poison ivy is now found worldwide.

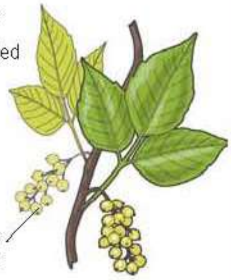
Berries are white when ripe

POISON OAK

Like poison ivy, the leaflets of poison oak grow in threes. This plant is found in wooded parts of North America.

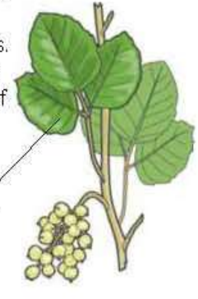
Leaflets are shaped like oak leaves

STINGING NETTLES

Found in many countries, stinging nettles usually cause only a temporary stinging sensation where the plant's hairs touch the skin. Apply a soothing cold compress or rub the affected area with a dock leaf if you can find one. Watch for an itchy red rash that indicates an allergic reaction.

NATURAL REMEDIES

Where nature causes a problem, she sometimes provides the solution. The following counter the effects of urushiol:

- Jewelweed has a sticky juice that will dry out blisters in a few days. Cut a piece of stem, split it down the middle, and rub the pieces on the affected skin.

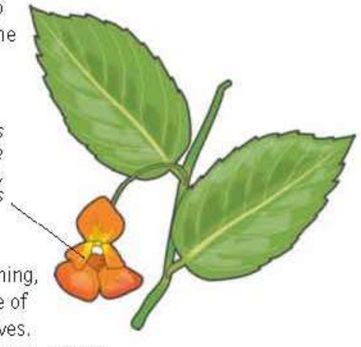
Jewelweed has distinctive pale yellow or orange, spotted flowers

- To soothe itching, make a poultice of witch hazel leaves. Mash up the leaves, adding water if too dry, then apply to the affected area.
- Try washing with a tannic acid solution. Tannic acid is found in tea, or you can make it from oak bark (see opposite).
- Tea tree oil, from the leaves of the melaleuca tree, is also said to counter the effects of urushiol. Apply directly to the affected area.

SWALLOWED POISON

Try to find out what the casualty ingested, how much, and when. If you suspect the casualty has swallowed a caustic chemical, such as camp fuel, do not induce vomiting as it will burn again on the way back up. Get medical help and monitor the casualty. If the lips are burnt, give frequent sips of cool milk or water. If you know the casualty has swallowed a poisonous plant or fungus and he or she is conscious, induce vomiting by tickling the back of the throat. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.

DIARRHEA AND VOMITING

In the wilderness, diarrhea and vomiting can kill, because they lead to dehydration (see p. 272) or even shock (see p. 274). The most likely causes are food poisoning or drinking contaminated water, although infectious diseases are a possibility, so get medical help if symptoms persist. You should rest, keep warm, and replace lost fluids. If you're hungry, you can eat small amounts of bland foods such as pasta for 24 hours.

NATURAL REMEDIES

In a survival situation, there are various natural remedies you can try to stop diarrhea or alleviate symptoms such as stomach pain, although some are more palatable than others:

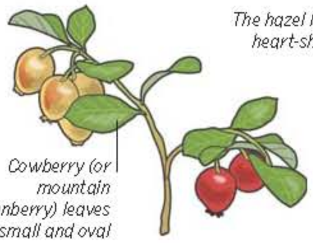

Cowberry (or mountain cranberry) leaves are small and oval

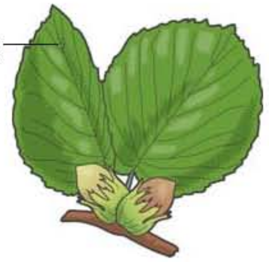

The hazel leaf is heart-shaped

- **Tea:** Drink tea made from hazel, cowberry, or cranberry leaves.
- **Bark:** Pull some bark off a tree (preferably oak), remove the inner bark and boil it for at least 12 hours, adding more water as necessary. The resulting black brew smells and tastes vile, but it contains tannic acid and will cure diarrhea. Drink one cup every two hours.
- **Charcoal:** Take a partially burned piece of wood, scrape off the char, then swallow about a handful with water.
- **Bones:** Burn to ashes then grind, or grind bones into a powder between two rocks. Make a paste with water, then swallow about a tablespoonful.
- **Chalk:** Grind into a powder, mix with water to a paste, then swallow about a tablespoonful.
- **Ash:** Make a paste of wood ash and water, then swallow. This will alleviate stomach pain.

REST AND REHYDRATE

Drink water to maintain your fluid levels. To help replace lost salts, dissolve a packet of rehydration salts in the water, or one teaspoon of salt in a liter of water, before drinking.

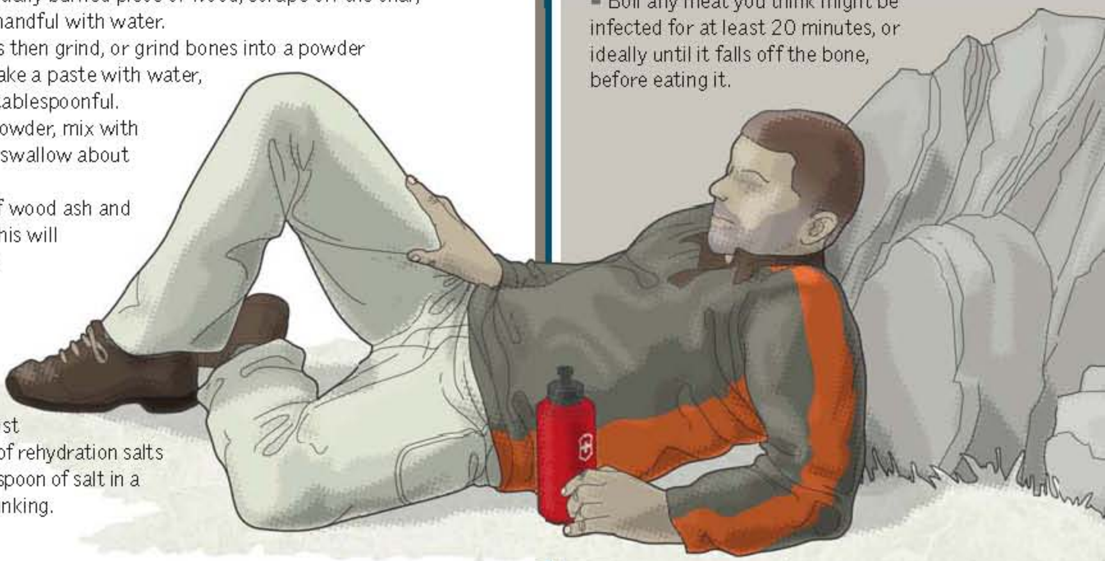

INTESTINAL PARASITES

There are two main types of intestinal parasite: helminths (tapeworms, pinworms, and roundworms) and protozoa (giardia, for example). The usual causes of infection are ingesting contaminated water or food, or poor personal hygiene. Symptoms include nausea or vomiting, diarrhea (see left), dysentry, bloating, stomach pain, weight loss, or a rash or itching around the rectum. Seek medical advice as soon as possible.

WILDERNESS CURE FOR WORMS

If you have passed a worm in your stool, swallow a couple of tablespoons of kerosene. While this method may make you sick, it will make the worms sicker. Gasoline will also work but is not as effective.

HOW TO AVOID BEING INFECTED

Prevention is far better than cure and to avoid being infected by parasites present in water or spread in feces you should always do the following in the wilderness:

- Boil all water, or use other reliable methods of purification (see pp. 198–201).
- Don't clean your teeth or rinse your mouth with water that isn't pure.
- Don't swim or stand in rivers or lakes in places where you could be at risk unless absolutely necessary.
- Cover any cuts or wounds on your skin.
- Maintain strict personal hygiene in camp (see pp. 116–17) and when handling food.
- Boil any meat you think might be infected for at least 20 minutes, or ideally until it falls off the bone, before eating it.

BONES, JOINTS, AND MUSCLES

CRAMP

This painful muscle spasm can be caused by dehydration and a reduction of body salts through perspiration—so make sure you have enough to drink when exercising. Sit down, rest, and stretch the affected muscles.

FOOT

Help the person stand on his good foot and stretch the muscles to reverse the spasm. Once the cramp has eased, massage the affected area of the foot.

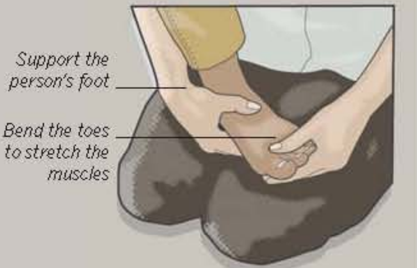
Support the person's foot
Bend the toes to stretch the muscles

CALF

Sit the person down and support the affected leg. Help him to straighten his leg, and flex his toes to reverse the spasm. Then massage the painful muscle.

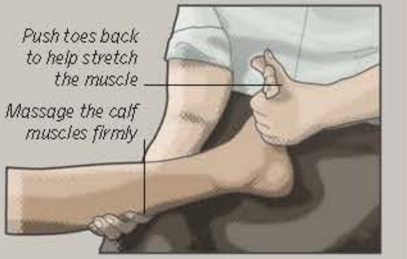
Push toes back to help stretch the muscle
Massage the calf muscles firmly

THIGH

If the cramp is in the back of the thigh, straighten the leg to stretch the muscle; if it's in the front of the thigh, bend the leg. Once the pain eases, massage the affected area.

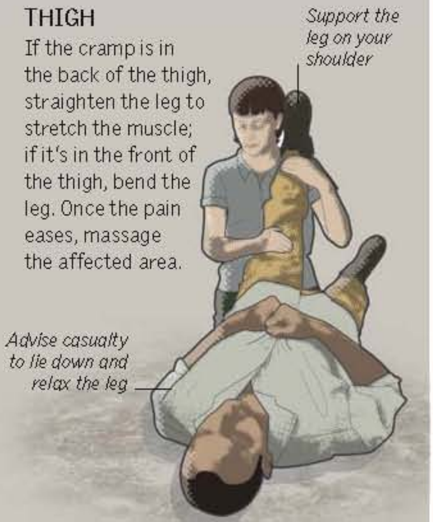
Support the leg on your shoulder
Advise casualty to lie down and relax the leg

IT CAN BE DIFFICULT TO TELL whether an injury is a sprain, a broken bone, or a dislocation. The ends of broken bones can move, damaging blood vessels or nerves nearby, so treat the casualty in the position found and immobilize the injured area before letting him or her move. Anyone with a spine injury or broken leg must be carried by stretcher.

SPRAINS AND STRAINS

A strain is a pulled muscle. A sprain occurs when ligaments that hold a joint together are damaged. The ideal treatment is to raise the injury, cool, and rest it. If you provide comfortable support, gentle movement can help the injury. If in any doubt, treat the injury as a broken bone (see opposite).

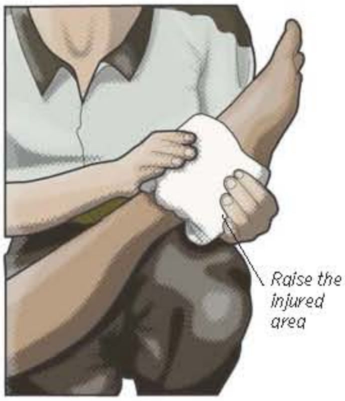
Raise the injured area

1. Rest and support the injury. Wrap a cold compress around it for at least 10 minutes to reduce swelling and bruising.

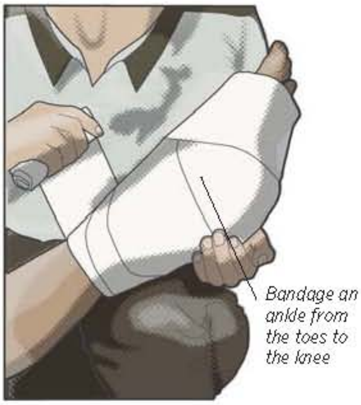
Bandage an ankle from the toes to the knee

2. Leave the compress in place, or wrap padding around the injury. Apply a crêpe bandage from below the injury to the next joint.

ARM INJURY

Falling onto an outstretched hand can cause a broken wrist, forearm, upper arm, or collar bone. Support the affected arm in a sling. If a casualty can't bend his arm, he may have injured his elbow, in which case don't use a sling. Instead wrap padding around the joint and secure the arm to the body with triangular bandages. To make sure that the bandage isn't too tight, check the wrist pulse.

USING A SLING

Slide a triangular bandage between the arm and the chest. Bring the front up over the arm and tie a reef knot on the uninjured side.

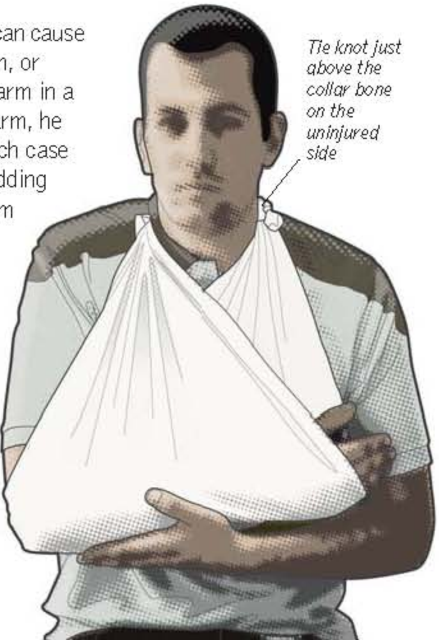
Tie knot just above the collar bone on the uninjured side

LEG INJURY

Injuries to the legs can be serious, as any fracture is likely to be unstable, which means that the ends of the bones can move easily, and could pierce one of the large blood vessels in the leg, resulting in severe bleeding. Don't move the casualty unless you have to, and, even then, only when the legs have been immobilized. If you see any signs of shock (see p. 274) ensure the head is low, but do not raise the legs.

HAND INJURY

Injuries to hands are often complicated by bruising or bleeding. Raise the injured hand and treat bleeding with direct pressure (see p. 264). Remove jewelry before the area starts to swell. Wrap the hand in padding and support it in a raised position with a sling.

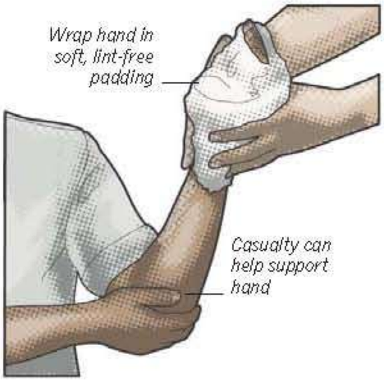

Wrap hand in soft, lint-free padding

Casualty can help support hand

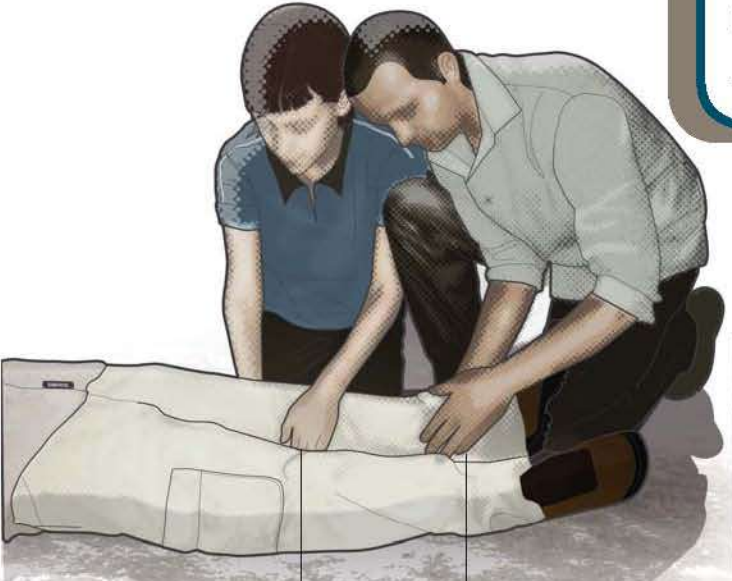

Support joints above and below the injury

Keep the injured leg as straight as possible

1 Lay the casualty down and support the injury to minimize further damage. Call for emergency help. If this is nearby just maintain this support. You can put rolled-up coats or blankets on either side for extra support.

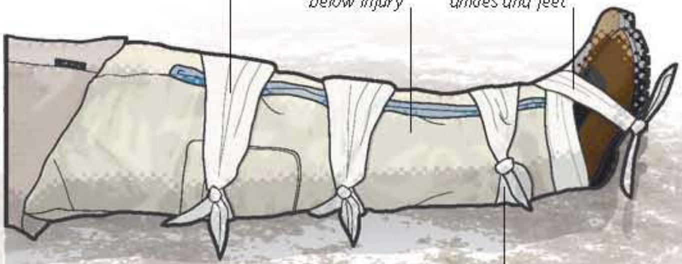

Put first bandage around the knees

Secure third and fourth bandages above and below injury

Tie the second bandage in a figure of eight around the ankles and feet

Secure bandages with reef knots on the injured side

2 If help is delayed or you need to transport the casualty, put bandages around the knees and ankles (and pelvis, if the thigh is injured), and above and below the injury. Place padding between the legs, then tie the bandages.

SPINAL INJURY

If a person falls and lands on his back, or falls from a height, it is best to assume that he has a spinal and, probably, a head injury. Don't move him—support his head and neck in line with the rest of his back. Moving him could damage the spinal cord, which may result in permanent loss of movement below the injured area. Call for emergency help, or send someone to get help while you stay with the casualty. If you have to move him because he is in danger, use the log-roll technique (see p. 279).

SUPPORT THE HEAD AND NECK

Kneel or lie behind the casualty's head. Keep your arms steady by resting your elbows on your thighs and place your hands on either side of his head to keep it in line with his body. Wait for help.

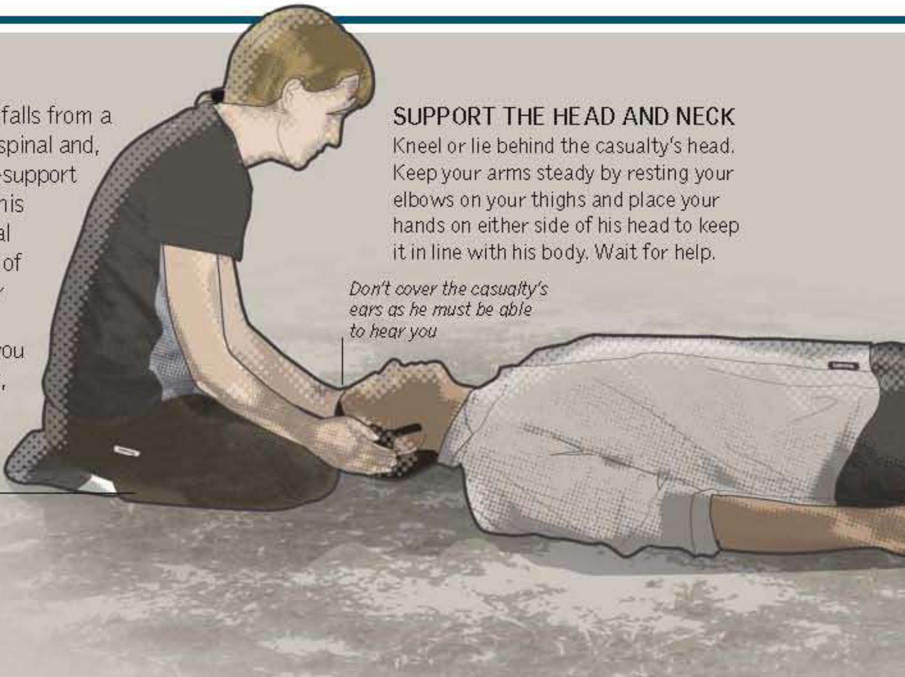

Don't cover the casualty's ears as he must be able to hear you

Make sure you are comfortable because you may have to wait for help to arrive

TREATING EXPOSURE

EXTREMES OF HEAT AND COLD can prevent the body's temperature-regulation mechanisms from functioning properly. Both extremes can cause life-threatening injures, so it's vital to act quickly. Don't leave a casualty—call for help or, if possible, send someone to seek help while you treat him or her.

WARNING!
If a casualty becomes unconscious, be prepared to give cardiopulmonary resuscitation (see p. 277).

HEAT INJURIES

In hot weather, wear a hat, reapply sunscreen frequently, and stay in the shade when you can to avoid heat injuries such as sunburn. You'll quickly become dehydrated if you don't drink enough water to replace fluids lost through sweating.

DEHYDRATION

Help the casualty to sit down and give him fluids to drink; water is usually sufficient, but rehydration salts mixed with water is best. If the casualty complains of cramp, help him stretch the affected muscles, then massage them firmly (see p. 270).

HEAT EXHAUSTION

If a casualty feels dizzy and starts to sweat profusely, but has cold, clammy skin, get him into the shade and give him fluids to drink. Help him to lie down then raise his legs—support his feet on a backpack—to help improve blood flow to the brain. Monitor him while he recovers.

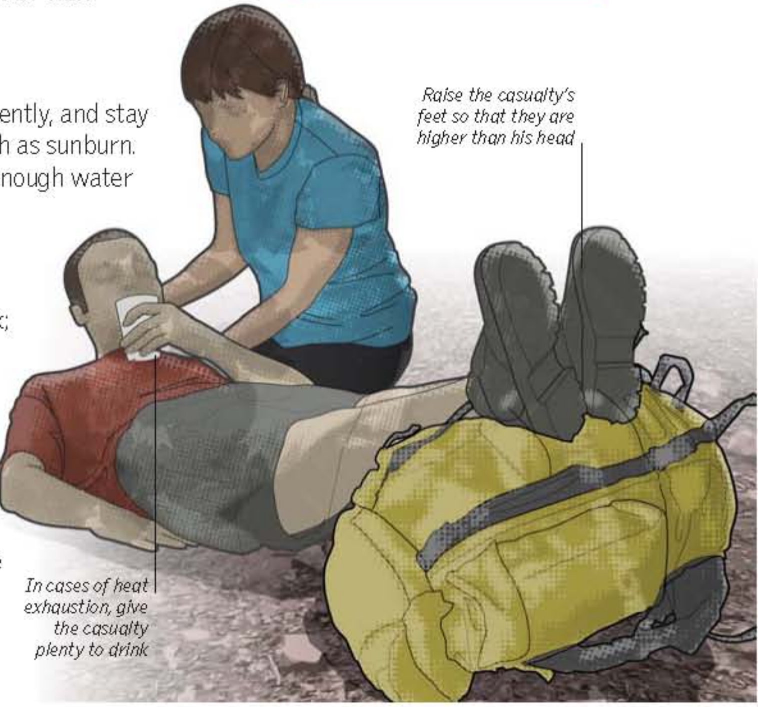

Raise the casualty's feet so that they are higher than his head

In cases of heat exhaustion, give the casualty plenty to drink

HEATSTROKE

This life-threatening condition may follow heat exhaustion or develop with no warning. Heatstroke causes the body's temperature-control mechanism to fail. If a person complains of headache, feels dizzy, has hot, dry skin, and begins losing consciousness, he or she may have heatstroke and will need urgent medical help.

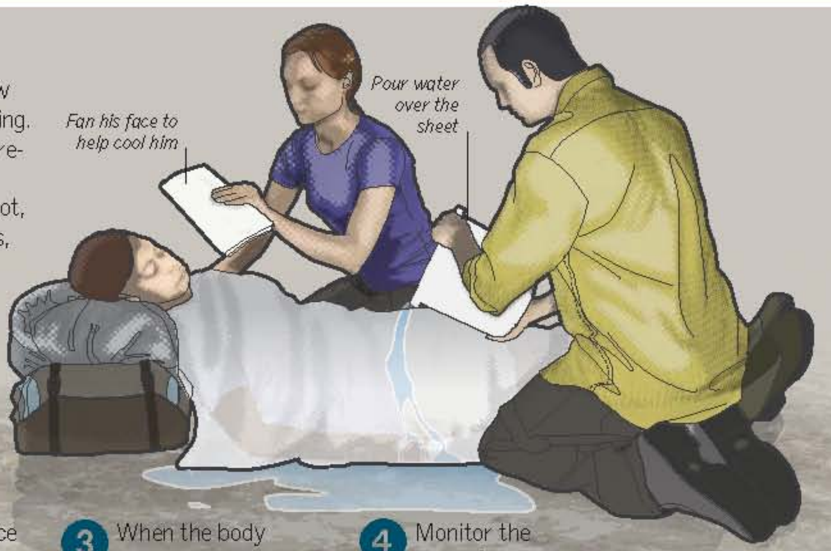

Fan his face to help cool him

Pour water over the sheet

1. Move the casualty to as cool a place as possible—out of the sun. Help him to sit or lie down with his head raised and remove all of his outer clothing.

2. Do anything possible to quickly reduce the body temperature. Ideally, wrap him in a cold, wet sheet. Keep the sheet cool by continually pouring water over it.

3. When the body temperature drops below 104°F (40°C), replace the wet sheet with a dry one.

4. Monitor the level of response, pulse, and breathing. If his temperature starts to rise again, repeat the treatment.

COLD INJURIES

Exposure to cold can result in parts of the body freezing (frostnip and frostbite) or the body's core temperature becoming dangerously low (hypothermia).

FROSTNIP AND FROSTBITE

Frostnip is the freezing of the top layer of skin, usually on the face and extremities. The skin turns numb, white, and hard. Untreated, it can lead to frostbite, which is much more serious—the deeper tissues, and even the bone, freeze. The skin turns white or blue and feels solidly frozen.

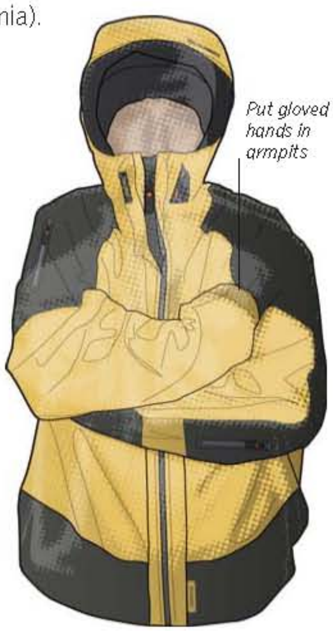

Put gloved hands in armpits

WARM GRADUALLY

Frostnip and frostbite can both be treated by warming the affected area, though frostbite requires more intensive treatment. Warm the affected area with body heat—place the casualty's hands in his own armpits, or place his feet in your armpits. Remove any rings and raise the injured part to reduce swelling. Ideally, place the affected area in warm water. Dress the injury in sterile dressings.

HYPOTHERMIA

This life-threatening condition develops if the body's core temperature falls below 95°F (35°C). Treatment aims to prevent further heat loss. It's vital that a casualty is warmed up gradually. If he or she is warmed up too quickly, blood is diverted away from vital organs, such as the heart and brain, to the skin, which can actually speed up cooling of the body. A casualty with hypothermia must be moved on a stretcher.

1. Help the casualty to a sheltered place where he should rest to prevent his body temperature falling further. Send someone else to get help.
2. If possible, and if there is no risk of further cold, remove any wet clothing and replace with warm, dry clothes—but don't give up your own clothes.
3. Put a thick layer of dry leaves underneath him to insulate him from the ground. Help him to lie down in a sleeping bag and, if you have one, cover him with a survival blanket.

Give the casualty warm (not hot) sweet drinks and high-energy foods such as chocolate

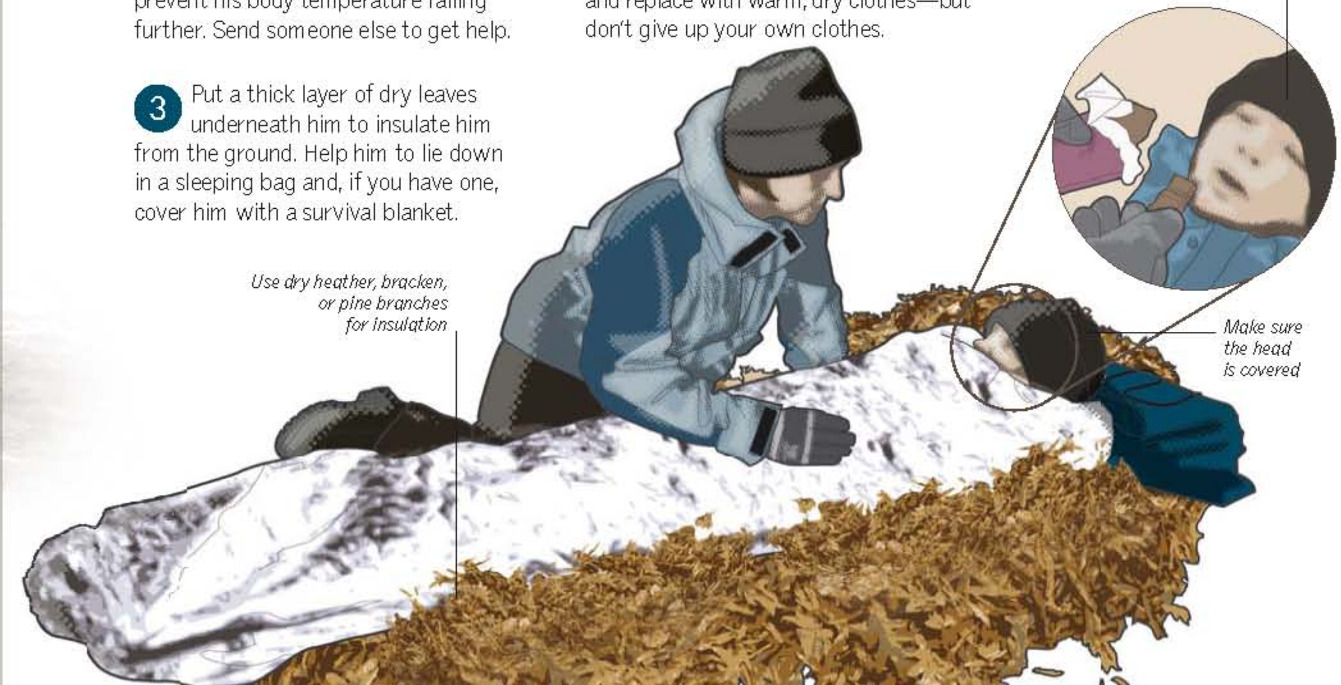

Use dry heather, bracken, or pine branches for insulation

Make sure the head is covered

DROWNING

If a person has been immersed in cold water, there is a high risk of hypothermia. In addition, the cold can cause the heart to stop, or throat spasms can block the airway. Water can also enter the lungs and cause secondary drowning (see p. 255) hours after the person appeared to have recovered.

RESCUING A CASUALTY

If you have rescued a person from water, help him or her to lie down with the head low. Replace wet clothes with dry ones. Treat for hypothermia (see left). If the person is unconscious and not breathing, give five rescue breaths before beginning chest compressions (see p. 277).

ALTITUDE SICKNESS

Symptoms of altitude sickness include nausea, loss of appetite, shortness of breath, and a headache that's not relieved by medication. The casualty may also have difficulty sleeping and will feel unwell. The only solution is to start the descent immediately and remain at a lower altitude for a few days. Severe cases will need to be carried.

TREATING SHOCK

SHOCK IS A LIFE-THREATENING CONDITION that occurs if the circulatory system fails. The most common cause is severe bleeding, but it can also be a consequence of burns. Initially, there will be a rapid pulse and pale clammy skin. As the condition progresses, breathing becomes rapid and shallow, pulse weakens, and skin becomes pale gray-blue. If untreated, unconsciousness results.

ANAPHYLACTIC SHOCK

Anaphylactic shock is a rare but severe allergic reaction affecting the whole body. People who know they are susceptible to this carry a special adrenaline (epinephrine) autoinjector for use in an emergency. If the casualty has one, but is too weak to use it, take off the safety cap and, holding it in your fist, press it against the casualty's thigh (through clothing if necessary).

DEALING WITH SHOCK

Do not give the casualty anything to eat or drink as an anaesthetic may be needed; moisten his lips with water if he's thirsty. Call for help; the casualty must be transported in the treatment position.

1. Treat the cause of shock, for example bleeding or burns (see p. 265). Suspect shock, too, if you notice any symptoms, yet can't see any obvious injury—it could be caused by internal bleeding.

2. Help the casualty to lie down; insulate him from the ground with blankets or bracken. Raise and support his legs as high as you can above the level of his heart.

3. Loosen tight clothing, for example at the neck, chest, and waist. Keep his head low; this may prevent him from losing consciousness. Keep the casualty warm; cover him with a blanket or sleeping bag if you have one.

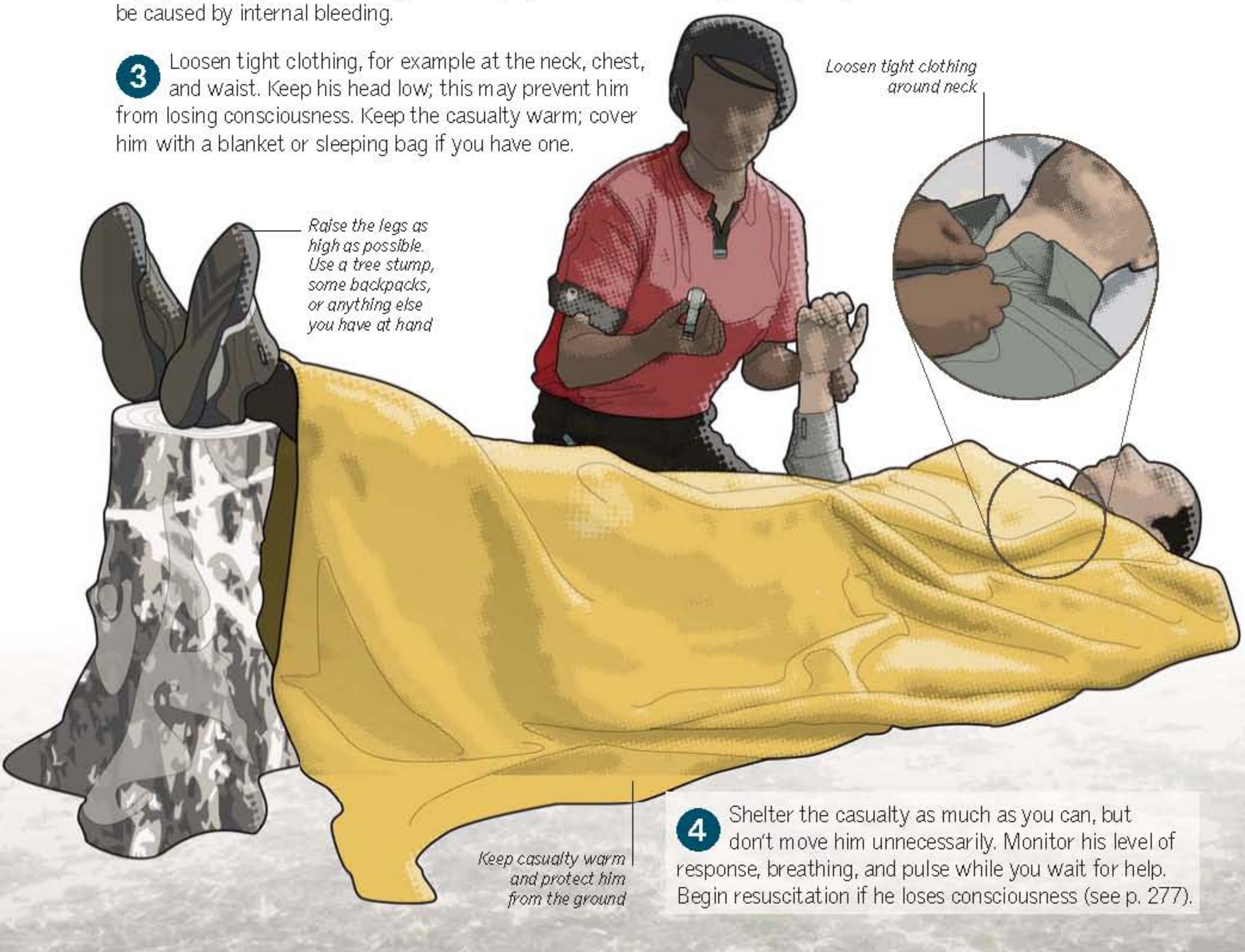

Raise the legs as high as possible. Use a tree stump, some backpacks, or anything else you have at hand

Loosen tight clothing around neck

Keep casualty warm and protect him from the ground

4. Shelter the casualty as much as you can, but don't move him unnecessarily. Monitor his level of response, breathing, and pulse while you wait for help. Begin resuscitation if he loses consciousness (see p. 277).

BREATHING DIFFICULTIES

RESPIRATORY PROBLEMS need prompt treatment because they can prevent sufficient oxygen reaching the body tissues. The cause may be temporary, for example choking, suffocation, or smoke inhalation, or it can be a long-term condition such as asthma that requires medication. You may need to get emergency help.

WARNING!
If at any stage the casualty loses consciousness, open the airway (see p. 276) and check breathing; the throat muscles may relax enough to allow breathing. If he or she is not breathing, begin resuscitation as this may dislodge the blockage (see p. 277).

CHOKING

When an object becomes stuck in the throat it can cause a muscular spasm that blocks the airway. Always ask the casualty if she is choking, to make sure. If she can speak, cough, or breathe, the obstruction is mild and she will probably be able to clear it herself.

1. If the casualty is breathing, tell her to continue coughing. If she can't speak or cough, help her to bend forward. Support her upper body and give up to five back blows in between her shoulder blades with the heel of your hand.

2. If back blows fail, stand behind her. Put your arms around her abdomen, clench one fist and grasp it with the other hand. Pull sharply inwards and upward up to five times. Check the mouth and remove any obvious blockage.

3. If the obstruction has still not cleared, repeat the five back slaps and five abdominal thrusts three times, then call for help if it's not already on the way. Continue until help arrives or the casualty becomes unconscious (see pp. 276–77).

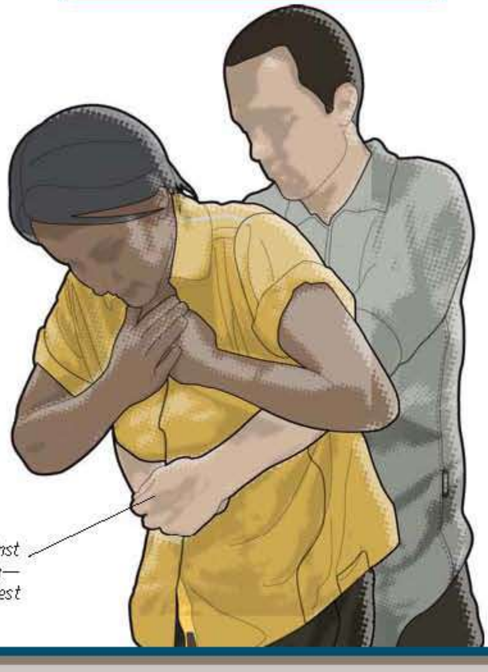

Pull against the abdomen—not the chest

ASTHMA

This is a condition in which breathing becomes difficult because the muscles in the air passages go into spasm. Most people who suffer from asthma carry inhalers with them. Many have two inhalers—a brown or white "preventer" inhaler and a blue "reliever" inhaler to use in an attack. If the person does not have any medication, sit him or her down and call for emergency help.

1. Sit the casualty down and advise her to take a puff from her reliever inhaler. Tell her to take slow, deep breaths and to sit as upright as she can. The attack should start to ease in a few minutes.

2. If the attack does not ease, tell her to take another dose from her inhaler and rest while she recovers. If the attack still does not ease, or is becoming worse, emergency help is needed as she could lose consciousness.

WHAT TO DO IN A SURVIVAL SITUATION

The best advice is prevention—know what triggers the asthma and avoid those things if at all possible. However, should someone have an asthma attack and not have an inhaler, try the following:

- Tell her to exhale as completely as she can. This expels the "stale" air, with little or no oxygen in it. It may be difficult—and will seem strange—to blow air out when the basic instinct is to gulp air in, but this does work.
- Then, tell her to inhale, slowly and steadily, and to close her eyes to help calm herself while doing so.

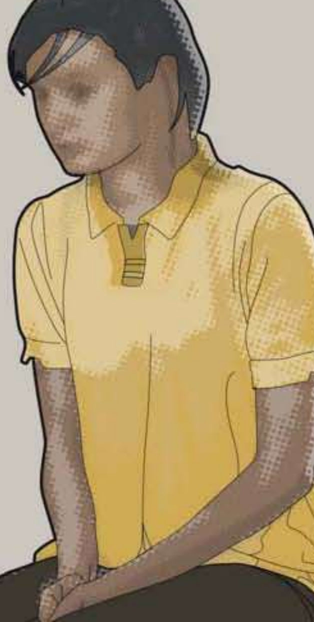

Take slow, deep breaths

UNCONSCIOUSNESS

IF SOMEONE FALLS UNCONSCIOUS, your priority is to make sure their airway is open so that they can breathe. Call for emergency help immediately (ideally ask someone to do this while you treat the casualty). Don't move the casualty and don't leave him or her alone unless you have to go and get help.

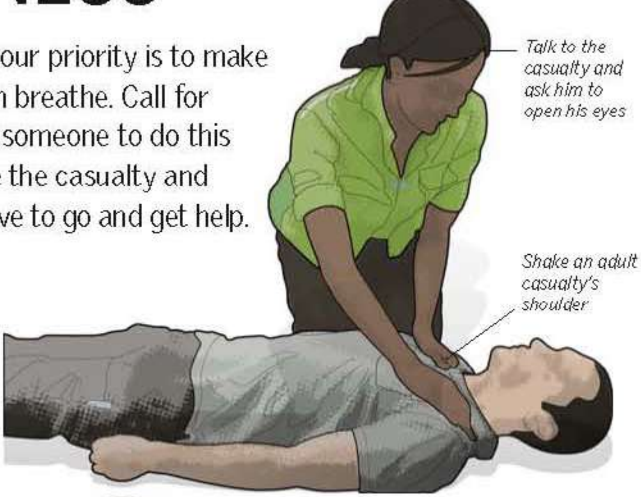
Talk to the casualty and ask him to open his eyes

Shake an adult casualty's shoulder

CHECK FOR RESPONSE

Gently shake the casualty's shoulders (if it's a child, tap the shoulders). Talk to him and watch for a response. If he's alert, he's conscious. If, for example, he reacts weakly, he may not be fully conscious—monitor him for any change (deterioration or improvement). If there's no response, he's unconscious.

OPEN THE AIRWAY

If an unconscious casualty is on his back, he's at risk of swallowing his tongue—thereby blocking his air passages. Tilting the head and lifting the chin will "lift" the tongue, clearing the air passage.

CHECK THE BREATHING

Tilt a casualty's head back with one hand and lift the chin with two fingers of your other hand; don't press on the soft tissues under the chin. Keep the airway open, and look, listen, and feel for normal breathing. If the casualty is breathing normally, place him in the recovery position (see below). If he's not breathing, begin chest compressions right away (see opposite).

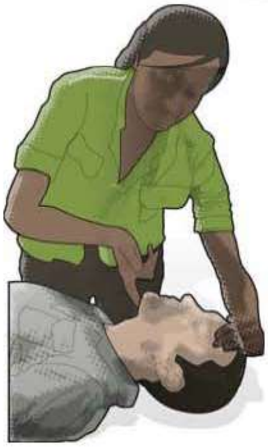
1 Place a hand on the forehead to tilt the head; lift the chin.

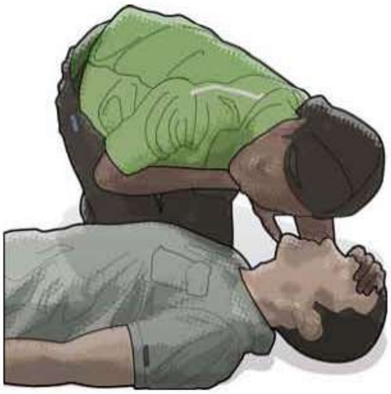
2 Look along the casualty's chest; listen, and feel, for breath against your cheek for no more than ten seconds.

RECOVERY POSITION

If an unconscious casualty is breathing, place him in the recovery position to keep his airway open and clear. Remove anything bulky from his pockets. Kneel beside him. Bend the arm nearest you at a right angle to his body then bring the other arm across his chest until his hand rests against his near cheek, and hold it there. Bend the far leg at the knee, and, still holding the knee, pull the casualty toward you until he is on his side.

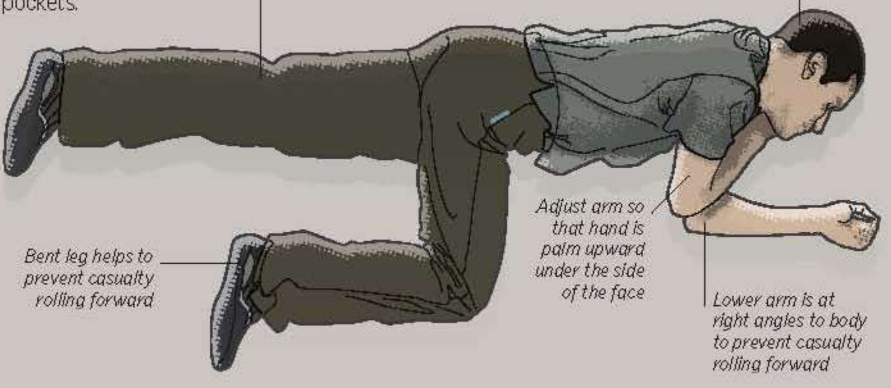
Keep lower leg straight, in line with the spine

Tilt the head back to keep airway open

Bent leg helps to prevent casualty rolling forward

Adjust arm so that hand is palm upward under the side of the face

Lower arm is at right angles to body to prevent casualty rolling forward

CARDIOPULMONARY RESUSCITATION (CPR)

If a casualty is not breathing, you must try to keep the body supplied with oxygen by using chest compressions and rescue breaths until emergency help arrives. This is known as cardiopulmonary resuscitation, or CPR. If an adult collapses, the cause is most likely to be a heart problem so treat as below. If you have rescued an unconscious casualty from water, start with rescue breaths as for a child (see right). If you are unable to achieve rescue breaths, you can give chest compressions alone.

CPR ON A CHILD

- Start by giving the child five rescue breaths.
- Give 30 compressions using the heel of one hand only and slightly less pressure than you would for an adult. Depress the chest by approximately one third of its depth.
- Continue with 30 compressions followed by two rescue breaths until the child recovers, help arrives, or you become exhausted.
- If you are on your own, give CPR for one minute before you try to call for help.

HOW TO GIVE CPR

Kneel beside the casualty, level with his chest so that you don't have to change position. If you have someone else with you, take it in turns to give CPR so you don't become too exhausted. Change over at the end of each two-minute cycle.

1. Put one hand on the center of the casualty's chest—make sure you don't press on the lower abdomen, the tip of the breastbone, or the ribs.

2. Place the heel of your other hand on top of the first and link your fingers together. Keep your fingers off the casualty's chest.

3. Begin chest compressions. Lean over and, keeping your arms straight, press straight down on the casualty's chest, depressing it by 1½–2 in (4–5 cm). Release the pressure and let the chest come back up, but don't move your hands. Repeat 30 times.

4. Tilt the casualty's head to open the airway and pinch his nose to close the nostrils. Let his mouth fall open slightly. Lift his chin with the fingers of your other hand.

5. To begin rescue breaths, take a normal breath and seal your lips over those of the casualty. Blow into his mouth until you see his chest rise, then lift your mouth away and watch his chest fall. If his chest doesn't rise, adjust his head and try again. Repeat to give a second breath, but don't make more than two attempts at giving rescue breaths before compressing again.

6. Continue the cycle of 30 compressions followed by two rescue breaths until the casualty recovers, help arrives, or you are too exhausted to keep going.
   - If at any stage the casualty starts breathing normally, place him in the recovery position (see opposite) and monitor his condition until help arrives.

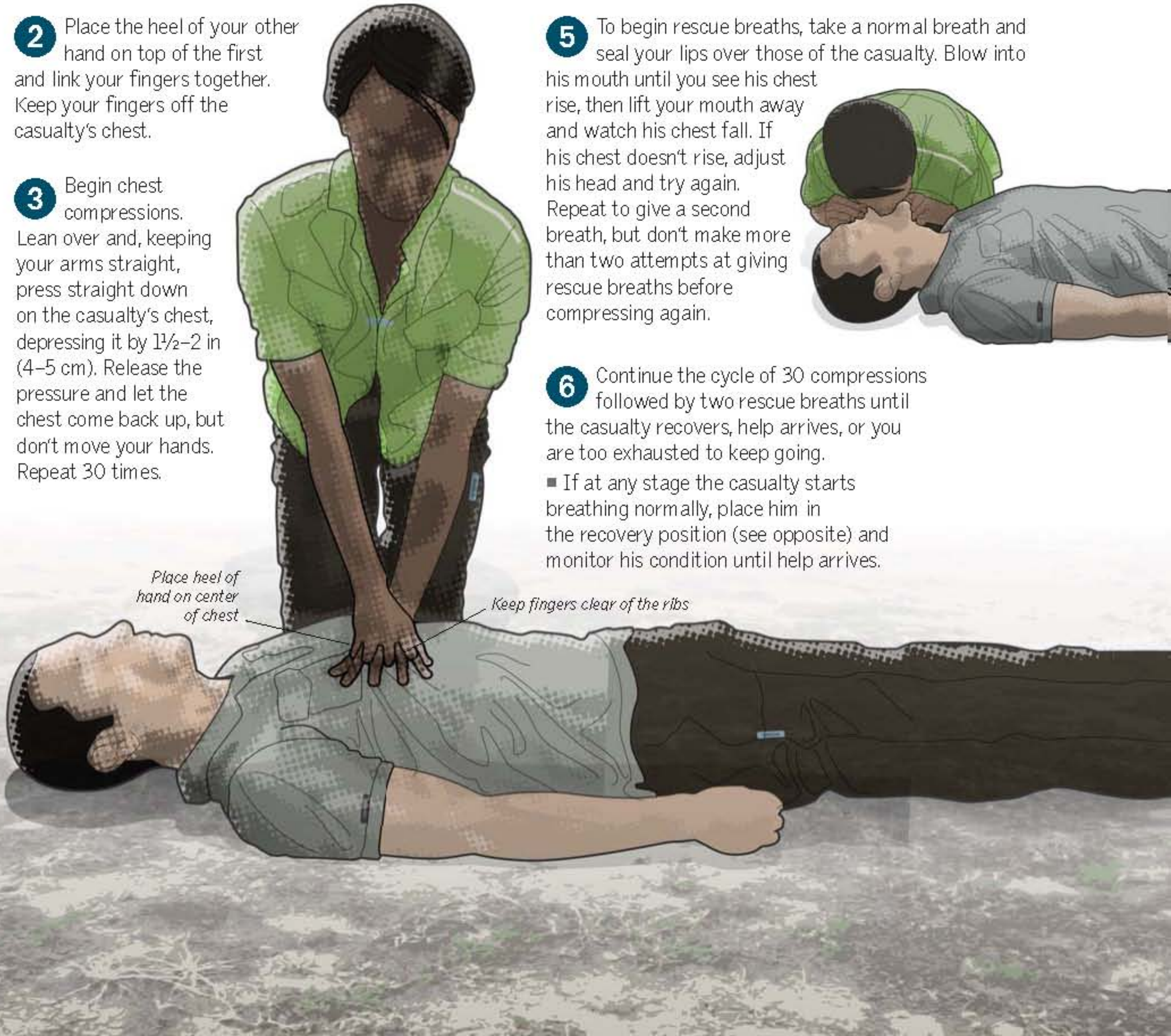
Place heel of hand on center of chest
Keep fingers clear of the ribs

MOVING A CASUALTY

IDEALLY, AN INJURED PERSON should be treated in the position in which he or she was found. Make the casualty as comfortable as possible and wait for help. In a survival situation, if you have to move someone, it's essential to immobilize an injury first to avoid aggravating it.

WARNING!
Move a casualty only if his life is in immediate danger if he is left where you found him. Even then, only move him if by doing so you will not injure yourself. Don't otherwise attempt to move a casualty, as it will make his injuries worse—stay with the casualty and send others for help.

PREPARING FOR A MOVE

Plan your move before you start. Choose a method appropriate to the injury. Don't attempt to move anyone by yourself if you have help. Encourage the casualty to do as much as possible himself.

FIREFIGHTER'S LIFT

This is a technique that can be used to move a conscious casualty if you're on your own and need to carry an injured person a short distance. It gets its name from the firefighters who originally used it. Don't use it if a casualty has head or facial injuries or a broken arm or leg. To prevent injuring your back, use your legs to power the move.

1 Support the casualty's injury with padding and bandages. Help him to stand up.

- Stand at right angles to the casualty, then squat down in front of him, staying as close to him as possible.
- Your shoulder should be level with the top of the casualty's legs.

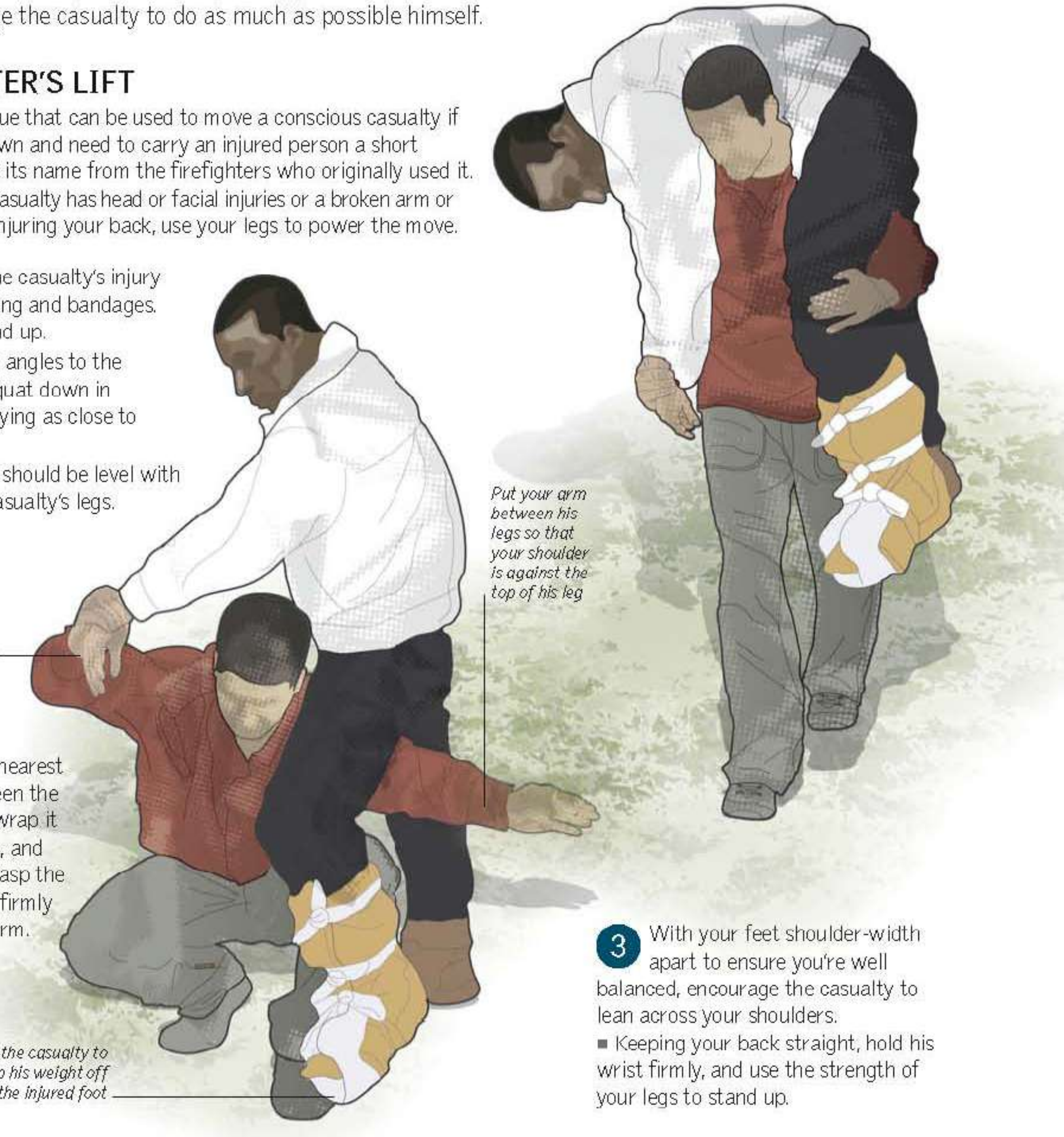

Put your arm between his legs so that your shoulder is against the top of his leg

Hold his wrist to help support him

Tell the casualty to keep his weight off the injured foot

2 Pass your nearest arm between the casualty's legs, wrap it around his thigh, and grasp his leg. Grasp the casualty's wrist firmly with the other arm.

3 With your feet shoulder-width apart to ensure you're well balanced, encourage the casualty to lean across your shoulders.

- Keeping your back straight, hold his wrist firmly, and use the strength of your legs to stand up.

IMPROVISING A STRETCHER

If you need to move a casualty who is unconscious, has a broken leg, or a spine injury, he must be carried by stretcher. Ideally, call the emergency services, and wait for them to arrive. If you can't contact them and you need to get the casualty to medical assistance, you may need to make a stretcher. Use a hurdle or gate, or make the stretcher from poles and coats.

WARNING!
Rescuers should not use their own clothing to make a stretcher if this might put them at risk of exposure.

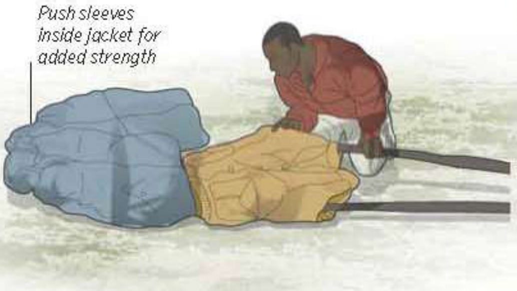
Push sleeves inside jacket for added strength

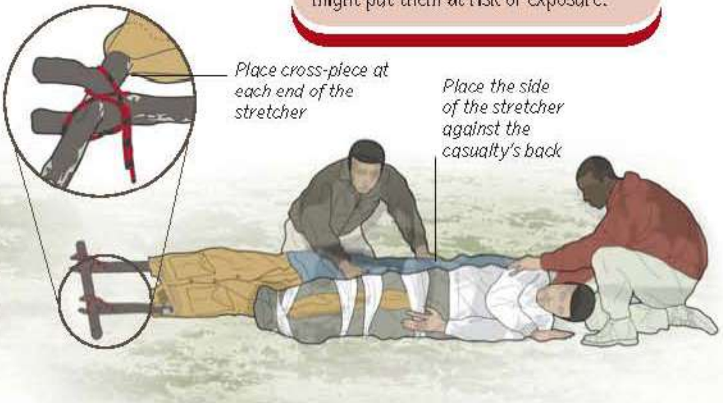
Place cross-piece at each end of the stretcher
Place the side of the stretcher against the casualty's back

1. Zip or button up two or three jackets. Cut two support poles, strong enough to bear the casualty's weight and 3 ft (1 m) longer than the casualty. Lash a short forked branch across each end (see pp. 166–67) to keep the support poles apart.

2. Immobilize any injury (see p. 271), and roll the casualty onto his uninjured side. One person should support his head while others help to keep his body straight, ideally one for the upper body and one for the legs. Slide the stretcher in place then roll him gently back on to it.

TWO-PERSON SEAT

If there are two of you, this method can be used to carry a conscious casualty who can't walk, but who can support himself with his arms. Nominate one person to be in charge of the move and give the instructions. Keep your backs straight at all times.

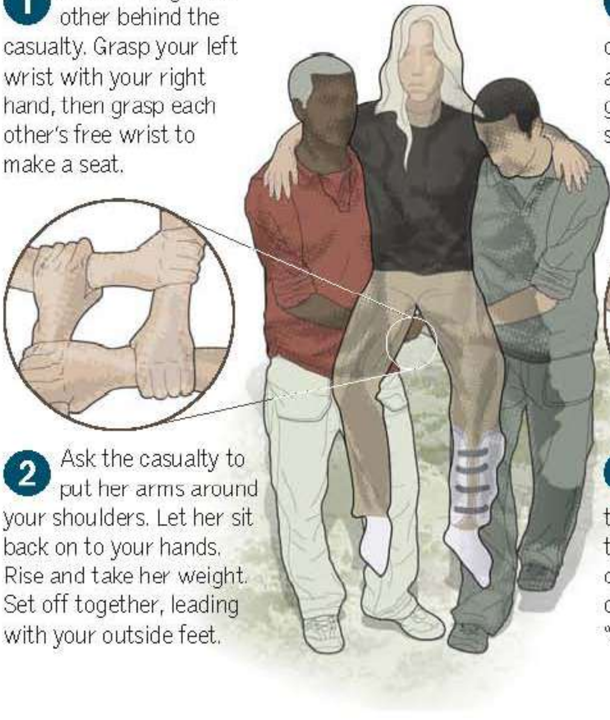

1. Stand facing each other behind the casualty. Grasp your left wrist with your right hand, then grasp each other's free wrist to make a seat.

2. Ask the casualty to put her arms around your shoulders. Let her sit back on to your hands. Rise and take her weight. Set off together, leading with your outside feet.

SUPPORTING THE UPPER BODY

If a casualty has an injured arm, the seat carry (see left) can be adapted to support the upper body. As before, one of you should direct the move and always move together. Keep your backs straight as you walk, and stop if it causes you or the casualty any discomfort.

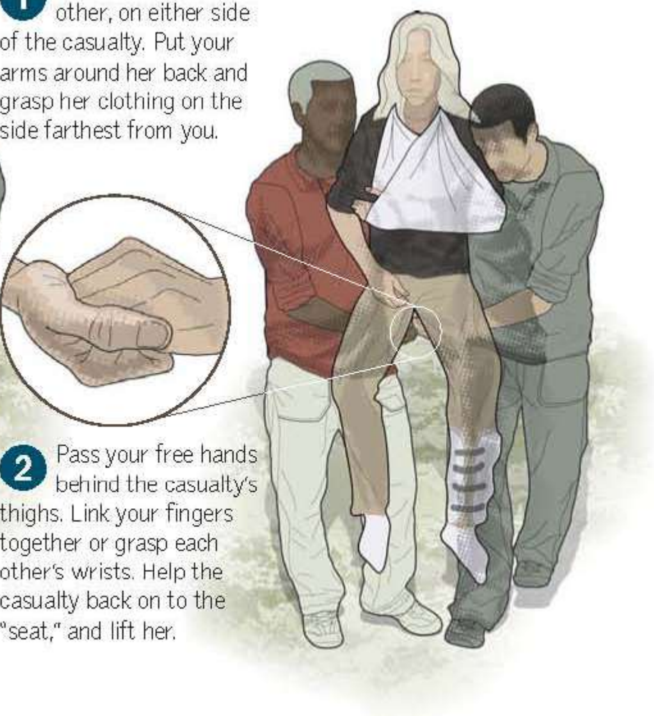

1. Stand facing each other, on either side of the casualty. Put your arms around her back and grasp her clothing on the side farthest from you.

2. Pass your free hands behind the casualty's thighs. Link your fingers together or grasp each other's wrists. Help the casualty back on to the "seat," and lift her.

WILD FOOD

MANY POISONOUS PLANTS in temperate regions resemble edible ones, so you should only eat plants that you can positively identify. If you're in any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07). Don't forget that some plants are edible only at certain stages of their growth. Never eat wild plants if you have any known

NAME	WHERE	IDENTIFICATION	EDIBLE PART
BISTORT/KNOTWEED *Polygonum* spp.	▪ Damp, grassy habitats, such as meadows, in North America and Eurasia.	▪ Perennial plant, 16–40 in (40–100 cm) tall. Green, triangular leaves. Small pink or white flowers form dense, cylindrical flowerheads at end of long, unbranched stem.	▪ Young shoots and leaves ▪ Roots
BRAMBLE/BLACKBERRY *Rubus fruticosus*	▪ In most temperate habitats worldwide, but particularly scrubland, hedges, and woodland.	▪ Scrambling deciduous shrub with thorny stems forming thickets. Green, toothed leaves. White or deep pink five-petaled flowers. Fruit ripens to blue-black in late summer.	▪ Fruit ▪ Young shoots ▪ Leaves
BUCKWHEAT *Fagopyrum esculentum*	▪ Open grassland worldwide.	▪ Red-stemmed plant up to 2 ft (60 cm) tall. Green, triangular leaves. Small pink, five-petaled flowers grow in clusters. Triangular seeds.	▪ Seeds
DANDELION *Taraxacum* spp.	▪ Widely distributed in temperate areas of Eurasia; introduced to the Americas, Australia, and New Zealand.	▪ Rosette of green, jagged-edged leaves grows close to the ground from a central taproot. Large, bright yellow flower head matures into a spherical "clock" of several seeds.	▪ Leaves ▪ Roots
DOG ROSE *Rosa canina*	▪ In hedgerows, woodland margins, and scrubland in Europe, northwest Africa, and western Asia; introduced to other regions.	▪ Scrambling deciduous shrub with spiny stems. Dark green, toothed leaves. Scentless, white, pink or deep pink five-petaled flowers. Fruit, or "hip," is orange-red and oval.	▪ Hips ▪ Buds and flowers ▪ Young leaves
STINGING/ COMMON NETTLE *Urtica dioica*	▪ By hedgerows and woodland margins in North America, Eurasia, and North Africa.	▪ Perennial herbaceous plant, 20–60 in (50–150 cm) tall. Green, oval, toothed leaves covered in stiff stinging hairs. Small green flowers sometimes have a red tinge.	▪ Young shoots and leaves
SUGAR MAPLE *Acer saccharum*	▪ Woods and forests in northeastern North America.	▪ Deciduous tree, usually 82–115 ft (25–35 m) tall. Gray-brown bark smooth on young trees, furrowed and flaky on old trees. Green, five-lobed leaves turn bright red in fall.	▪ Sap ▪ Inner bark
WALNUT *Juglans* spp.	▪ Temperate areas in Eurasia and North America. Common in the Himalayas.	▪ Large deciduous tree up to 82 ft (25 m) tall with distinctive gnarled bark. Green leaves have narrow leaflets. Nut has a thick, green husk, which rots away once it falls to the ground.	▪ Nuts
WATER CHESTNUT AND WATER CALTROP *Trapa natans* and *T. bicornis*	▪ In slow-moving water in warm temperate areas of Eurasia and Africa; introduced to North America and Australia.	▪ Green, triangular, saw-toothed leaves float on surface. White four-petaled flowers. Fruit is borne underwater and has four sharp spines and one hard gray seed.	▪ Seeds
WILD GRAPE *Vitis* spp.	▪ In most habitats in North America and Eurasia.	▪ Sprawling, high-climbing vine with large, lobed leaves. Fruit grows as hanging bunches of berries, which are amber or dark purple when ripe.	▪ Fruit ▪ Young leaves

TEMPERATE PLANTS

allergies, a pre-existing medical condition, or are pregnant. The UET doesn't apply to mushrooms, so never eat any mushrooms unless you can positively identify every single one as being edible. As part of your preparation for your trip, you should familiarize yourself with the edible plants in the region you're visiting, and their seasonal availability.

PREPARATION	SIMILAR PLANTS	BEWARE OF	EXTRA INFORMATION
▪ Young leaves and shoots can be eaten raw or boiled. ▪ Soak roots, then boil or roast.	▪ Alpine bistort (*P. viviparum*) is common on rocky soils in northern alpine regions and the Arctic.		▪ Roots are twice twisted and are said to resemble two snakes. They contain starch and can be used to make flour.
▪ Fruit and peeled young shoots are edible raw. ▪ Infuse the leaves to make tea.	▪ Wild raspberry (*R. idaeus*) fruit ripens to a bright red. Tea made from the leaves can be used to treat diarrhea.	▪ Catching yourself on thorns.	▪ Blackberries and raspberries are high in vitamin C and sugars. Blackberry leaf tea can be used to treat colds and coughs.
▪ Hull and grind the seeds to make buckwheat flour. ▪ Roast seeds and cook with water to make porridge.		▪ Buckwheat leaves are edible but can cause photosensitivity of the skin if eaten in large quantities.	
▪ Young leaves can be eaten raw; boil old leaves to remove the bitter taste (change the water once). ▪ Boil the roots.			▪ Roots can be roasted and ground to make a substitute for coffee. The leaves are high in vitamins A and C, iron, and calcium.
▪ Buds and flowers are edible raw. ▪ Chew the pulp of raw hips, or dry and eat later. ▪ Boil leaves in water to make tea.		▪ The hips of all wild roses (*Rosa* spp.) are edible, but do not eat the seeds if they are prickly as they will irritate the throat and stomach.	▪ Rose hips remain on the shrub all year. They are rich in vitamin C. Boil crushed hips in water then strain to make a nutritious syrup.
▪ Boil the plants for 10 minutes to destroy the formic acid and histamine in the stinging hairs.	▪ Dead nettles (*Lamium* spp.) are also edible if cooked. They have heart-shaped leaves and white or purple flowers. No stinging hairs.	▪ Gather and handle the plants carefully, wearing gloves if you have them, to avoid being stung by the hairs.	▪ Fresh nettles are rich in protein and vitamin K. The fibers in mature stems can be woven into cordage (see pp. 138–41).
▪ Boil sap into a high-energy syrup. ▪ Eat inner bark raw or boil into a gelatinous mass, then roast and grind into a flour.	▪ Red maple (*A. rubrum*) also produces sweet sap. ▪ White birch (*Betula* spp.) has edible inner bark.		▪ To collect the sap, cut a V-shape into the trunk, make a hole below the cut, and insert a leaf to guide the drips into a container below.
▪ Crack the shell and eat the ripe nut raw.			▪ Rich in fat, protein, and vitamins. ▪ Crushed green husks of black walnut (*J. nigra*) are toxic to fish. Use to poison and catch them.
▪ Seeds can be eaten raw or roasted.			▪ Water chestnuts are a source of carbohydrate.
▪ Ripe fruit is best eaten raw. ▪ Boil young leaves.		▪ Moonseed vines (*Menispermum* spp.) have grapelike fruits, which are poisonous. Unlike edible grapes, they only have one seed.	▪ Ripe grapes are rich in vitamin C and sugar. Water can be obtained from the vine stem (see p. 192).

WILD FOOD

SOME OF THE PLANTS listed below are good sources of water/liquid. In a survival situation in the desert, don't eat food if you don't have any water, as your body will use up its own water to digest it. Only eat plants that you can positively identify and, if you are in any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07).

NAME	WHERE	IDENTIFICATION
ACACIA/WATTLE *Acacia* spp.	▪ Africa, southern Asia, Australia, and the Americas.	▪ Thorny, medium-sized trees with gray-white bark. Green leaves divided into many small, oval leaflets. Small, ball-shaped yellow, white or pink flowers.
AGAVE *Agave* spp.	▪ Southern USA and Mexico, Central America, the Caribbean, and northeastern South America.	▪ Rosette of fleshy leaves with sharp tip and spiny margin. Central, tall flower stalk.
AMARANTH *Amaranthus* spp.	▪ The Americas, Africa, and Asia.	▪ Tall herb with alternating leaves and erect fleshy stems. Tiny brown/black seeds (one plant produces 40,000–60,000 seeds).
BAOBAB *Adansonia* spp.	▪ Africa and northern Australia.	▪ Large trees with swollen trunks, up to 30 ft (9 m) in diameter. Green leaves have 5–7 digit-like leaflets. Oblong fruit up to 8 in (20 cm) long.
CAROB TREE *Ceratonia siliqua*	▪ Mediterranean, northern Africa, Middle East, and India.	▪ Evergreen tree up to 50 ft (15 m) tall. Shiny green leaves up to 8 in (20 cm) long. Small red flowers. Flat leathery seed pods; dark-green/black when ripe. Hard brown seeds.
CARRION FLOWER *Hoodia* spp.	▪ Southwest Africa.	▪ Numerous thorny, succulent stems up to 6 ft (2 m) tall. Star-shaped or shallowly bell-shaped flowers give off the smell of rotting meat.
DATE PALM *Phoenix dactylifera*	▪ West and North Africa, Middle East, and India; introduced to Mexico and USA.	▪ Tall slender palms with a large crown of green leaves, divided into many narrow leaflets. Fruit reddish-brown when ripe.
PRICKLY PEAR CACTUS *Opuntia* spp.	▪ The Americas; introduced to the Caribbean, the Mediterranean, Africa, and Australia.	▪ Thick, jointed, flat, pad-like, green stems covered in spines. Red or yellow flowers. Fruit red when ripe.
WILD GOURDS *Cucurbitaceae*	▪ Kalahari Desert (Namibia), the Sahara, islands in the Mediterranean, the Middle East, and southeastern India.	▪ Ground-trailing green vine. Bright yellow flowers. Orange-sized green or yellow fruit.

DESERT PLANTS

Don't forget that some plants are edible only at certain stages of their growth. Don't eat wild plants if you have any known allergies, a preexisting medical condition, or are pregnant. When planning your trip, it's advisable to familiarize yourself with the edible plants in the region you are visiting and their seasonal availability.

EDIBLE PART	PREPARATION	BEWARE OF	EXTRA INFORMATION
▪ Seeds (from dark brown pods) ▪ Young leaves and shoots	▪ Roast seeds. ▪ Boil leaves and shoots.	▪ Catching yourself on thorns.	▪ Roots may be tapped for water.
▪ Stalk, before in flower ▪ Buds and flowers	▪ Roast the stalk. ▪ Boil buds and flowers.	▪ Juice from many species can cause acute contact dermatitis, with reddening and blistering lasting 1–2 weeks. Leaves have needle-like ends.	▪ Roasted stalk tastes sweet, like molasses.
▪ Young shoots and leaves ▪ Seeds	▪ Eat shoots and leaves raw or boiled or stir-fried. ▪ Remove chaff from seeds, then cook like popcorn or grind into flour to make bread.	▪ Amaranth leaves are high in oxalic acid, which can irritate the gut and cause kidney stones, so eat in moderation.	▪ Amaranth leaves are a common leaf vegetable in the tropics and warm temperate regions.
▪ Shoots and young leaves ▪ Fruit ▪ Seeds	▪ Eat the fruit pulp raw. ▪ Boil shoots and leaves. ▪ Roast seeds.		▪ Tap the trunk for water. Kalahari bushmen suck water through holes in the trunk using the hollow stems of grasses.
▪ Pods ▪ Seeds	▪ Eat the sweet, nutritious pulp of the pods raw. ▪ Grind the seeds into flour and make into porridge.		
▪ Stems	▪ Tap the stems for their water.		▪ The stems contain an appetite suppressant. Kalahari bushmen chew the stems of several *Hoodia* species before and during their long, arduous hunts.
▪ Fruit ▪ Growing tip of the palm (heart) ▪ Young leaves ▪ Sap	▪ Eat the fruit (dates) raw or sun-dried. ▪ Boil the leaves and heart. ▪ Boil the sap down to a syrup.		▪ Date palms always grow near water. The leaves can be used to thatch a shelter.
▪ Fruit ▪ Seeds ▪ Pads	▪ Peel fruit and eat raw. ▪ Roast and grind seeds to a flour. ▪ Boil or grill young pads, first removing thorns by peeling or scorching over a fire.	▪ Spines. Avoid any similar-looking plant that has milky sap as it may be poisonous.	▪ The non-milky sap from the pads is a good source of water.
▪ Fruit (Tsamma melon only) ▪ Flowers ▪ Seeds ▪ Young leaves and shoots	▪ Pound fruit to a pulp and eat raw. ▪ Eat the flowers raw. ▪ Roast or boil seeds. ▪ Chew leaves and shoots for water.	▪ The fruit of the wild desert gourd or colocynth (*Citrullus colocynthis*) is very bitter and is a strong laxative.	▪ The Tsamma melon (*C. lanatus*) is also known as the wild watermelon. The fruit pulp is a good source of water.

WILD FOOD

A GREAT NUMBER OF PLANTS flourish in the tropics, growing all year round in the warm and humid conditions—those listed below are just a few of the most common varieties. When you plan your trip, you should familiarize yourself with the edible plants in the region you are visiting. In tropical forests, most fruits are borne high in the canopy—out of reach,

NAME	WHERE	IDENTIFICATION	EDIBLE PART
BAMBOO *Bambuseae*	▪ Tropical and subtropical forests.	▪ Tree-like grasses with woody, segmented, hollow stems ranging from black to green to gold in color. Green blade-like leaves. Species range in height from 1⅔–11½ ft (0.5–3.5 m).	▪ Young shoots ▪ Seeds
BANANA AND PLANTAIN *Musa* spp.	▪ Native to Australia and Southeast Asia; introduced to other tropical regions.	▪ Tree-like herbaceous plants up to 33 ft (10 m) tall. Large, green, strap-like, split leaves. Flowers and fruit grow in dense hanging clusters.	▪ Fruit ▪ Buds and flowers ▪ Shoots and young stems ▪ Roots
BRAZIL NUT *Bertholletia excelsa*	▪ Along river banks in rain forests in South America.	▪ Large dry-season deciduous tree up to 150 ft (45 m) tall. Green, oblong, crinkly leaves. Yellow flowers. Fruit is the size of a coconut with a hard, woody shell and contains 8–24 seeds.	▪ Seeds or "nuts"
CEYLON/VINE SPINACH *Basella alba*	▪ Widespread in tropical forests.	▪ Trailing, vine-like plant reaching 100 ft (30 m) in length. Beet-red stems. Fleshy, oval or heart-shaped, green or red-purple leaves.	▪ Young leaves and stems
FIG *Ficus* spp.	▪ Variety of habitats in tropical and subtropical regions.	▪ Evergreen tree with long aerial roots growing from the trunk and branches. Leathery green leaves. Pear-shaped fruit grow directly from the trunk or branches.	▪ Ripe fruit
PAPAYA/PAWPAW *Carica papaya*	▪ Native to tropical rain forests in the Americas; introduced to other tropical and some temperate regions.	▪ Small tree up to 20 ft (6 m) high, with a soft hollow trunk. Large green leaves have seven lobes. Large melon-like fruit turn yellow or orange when ripe.	▪ Fruit ▪ Young flowers, leaves, and stems
PEANUT *Arachis hypogaea*	▪ Native to tropical rain forests in the Americas; introduced to other tropical and some temperate regions.	▪ Small, bushy plant up to 1⅔ ft (50 cm) tall. Paired, oval green leaves grow four to a stem. Yellow flowers. Underground legume (pod) contains 1–4 nuts.	▪ Seeds or "nuts"
SAGO PALM *Metroxylon sagu*	▪ Damp lowlands of tropical rain forests in Southeast Asia; introduced elsewhere.	▪ Spiny-trunked palm up to 33 ft (10 m) tall. Crown of pinnate (feather-like) green leaves.	▪ Pith ▪ Young shoots ▪ Young nuts
WATER LILY *Nymphaea*	▪ Lakes, ponds, and rivers in tropical and subtropical regions worldwide. Also found in temperate regions.	▪ Flat, green, heart-shaped leaves float on surface of water. Large, white, yellow, pink, or blue fragrant flowers.	▪ Tubers ▪ Stems ▪ Seeds

TROPICAL PLANTS

unless you can climb the tree. Only eat plants that you can positively identify. If you are in any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07). Don't forget that some plants are edible only at certain stages of their growth. Don't eat wild plants if you have a known allergy, a preexisting medical condition, or are pregnant.

PREPARATION	SIMILAR PLANTS	BEWARE OF	EXTRA INFORMATION
▪ Split the tough outer sheath of the shoots, and boil or steam. ▪ Boil the seeds or grind, mix with water, make into cakes, and bake.		▪ Shoots of the giant bamboo or volohosy (*Cathariostachys madagascariensis*) contain prussic acid (cyanide).	▪ Bamboo stems often hold water (see p. 190). The stems can be used for cooking (see p. 204) and building a shelter (see p. 169).
▪ Fruit and flowers are edible raw or cooked. ▪ Cook shoots, stems, and roots.			▪ Banana and plantain fruit is rich in potassium and vitamins A, B6, and C. Hard, unripe plantain fruit is only edible when cooked.
▪ Break fruit capsule open, crack the nut's shell, and eat raw.	▪ The sapucaia tree (*Lecythis* spp.) produces edible seeds in a similar fruit capsule. They are called paradise nuts.		▪ Ripe fruit capsules fall to the ground. Rodents and monkeys are attracted to the nuts, which are rich in fats and selenium.
▪ Steam, boil, stew, or stir-fry the leaves and stems.		▪ Don't eat too much, as the plant can have a mild laxative effect.	▪ Young leaves and stems are rich in vitamins.
▪ Eat the fruit raw or cooked.		▪ Don't eat any hard, woody fruit, or those covered in hairs.	▪ Edible figs are soft when ripe and green, red, or black in color.
▪ Flesh of ripe fruit is edible raw. Unripe fruit is edible cooked. ▪ Flowers, leaves, and stems must be boiled, and the water must be changed at least once.		▪ Don't get the milky juice of green, unripe fruit in your eyes as it will cause temporary blindness.	▪ Ripe fruit is rich in vitamin C. Unripe fruit will ripen quickly if placed in the sun.
▪ Remove the shell and eat the nuts raw.		▪ Don't handle or eat if you are allergic to peanuts.	▪ Peanuts are a good source of protein, B vitamins, and minerals.
▪ Crush the pith, knead it in water, strain into a pot, and let the fine sago settle. Squeeze and let it dry. Mix to a paste with boiling water.	▪ Many palms have edible parts. The sugar palm (*Arenga pinnata*) has an edible sap which can be boiled down to a thick syrup.	▪ Don't eat a palm's fruit unless you have positively identified it as an edible variety as the fruit of some species contain harmful crystals.	▪ A mature palm will yield 330–660 lb (150–300 kg) of sago starch—almost pure carbohydrate.
▪ Peel and slice the tubers, which can be eaten raw. ▪ Stems are best cooked. ▪ Dry and grind the seeds into flour.	▪ All parts of the lotus (*Nelumbo* spp.) are edible raw or cooked.		▪ The tubers are rich in starch. The seeds have a bitter taste but they are edible.

WILD FOOD

WHEN PREPARING FOR YOUR TRIP, you should familiarize yourself with the edible plants in the region you are visiting and their seasonal availability. Only eat plants that you can positively identify and, if you are in any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07). Don't forget that some plants are edible only

	NAME	WHERE	IDENTIFICATION
LICHENS	**ICELAND MOSS** *Cetraria islandica*	▪ Mountainous areas in Arctic, sub-Arctic, and cold temperate regions in North America and Europe; lava slopes and plains of Iceland.	▪ Mat-forming lichen, up to 4 in (10 cm) high. Gray-green to pale chestnut branches, rolled into tubes terminating in flattened lobes with fringed edges.
	REINDEER MOSS *Cladonia rangiferina*	▪ Tundra, bogs, and open woodlands in Arctic, sub-Arctic, and northern temperate regions.	▪ Mat-forming lichen, 2–4 in (5–10 cm) high. Gray, rounded branches resemble antlers.
	ROCK TRIPE *Umbilicaria* spp.	▪ On rocks in Arctic, sub-Arctic, and northern temperate regions.	▪ Rounded lichens with curling edges, usually gray or brown.
PLANTS	**ARCTIC WILLOW** *Salix arctica*	▪ Tundra in North America, Europe, and Asia; mountainous areas in some northern temperate regions.	▪ Mat-forming shrub, 1–2 ft (30–60 cm) high. Rounded, shiny green leaves. Flowers grow as yellow catkins.
	BEARBERRY *Arctostaphylos uva-ursi*	▪ Mountainous areas in Arctic and sub-Arctic regions.	▪ Dwarf evergreen shrub. Thick, leathery, club-shaped green leaves. Pink or white flowers. Bright red berries grow in clusters.
	BOG LABRADOR TEA *Rhododendron groenlandicum*	▪ Bogs and alpine areas in Arctic, sub-Arctic, and temperate regions in North America and Europe.	▪ Evergreen shrub, 1–3 ft (30–90 cm) high. Narrow, leathery, green leaves are hairy underneath and fragrant. Small, fragrant, white flowers grow in clusters and are sticky.
	CLOUDBERRY *Rubus chamaemorus*	▪ Bogs, marshes, and wet meadows in alpine and sub-Arctic regions of North America, Europe, and Asia.	▪ Perennial shrub, 4–10 in (10–25 cm) high. Soft, green leaves have 5–7 lobes. White flowers have five petals. Rasberry-sized, amber berries grow at the top of the plant.
	CROWBERRY *Empetrum nigrum*	▪ Tundra, moorlands, bogs, and spruce forests in Arctic, sub-Arctic, and northern temperate regions. Also found in the Andes.	▪ Dwarf, evergreen, mat-forming shrub. Short, pale green, needle-like leaves. Small, purple-red flowers. Black berries.
	KERGUELEN CABBAGE *Pringlea antiscorbutica*	▪ Rocky areas of sub-Antarctic islands in the Indian and Southern Oceans.	▪ Cabbage-like plant.

PLANTS IN COLD CLIMATES

at certain stages of their growth. Don't eat wild plants if you have any known allergies, a preexisting medical condition, or are pregnant. There are no poisonous types of lichen (althought see rock tripe entry, below), but they must all be soaked in water overnight and boiled well before they can be eaten.

EDIBLE PART	PREPARATION	BEWARE OF	EXTRA INFORMATION
▪ All	▪ Soak for several hours then boil well.	▪ All lichens contain an acid that will cause stomach irritation unless soaked and boiled before eating.	▪ Lichens are low in protein but high in carbohydrates (lichen-starch).
▪ All	▪ Soak for several hours then boil well.	▪ All lichens contain an acid that will cause stomach irritation unless soaked and boiled before eating.	▪ High in vitamins A and B. Important source of food for reindeer (caribou). Partially digested lichens from the reindeer's stomach are considered a delicacy by herders.
▪ All	▪ Soak for several hours then boil well.	▪ There have been some reports of rock tripe poisoning so apply the Universal Edibility Test first.	
▪ Young shoots ▪ Leaves ▪ Young roots	▪ Peel off outer bark and eat inner shoots raw. ▪ Eat the leaves raw. ▪ Peel roots and eat raw.		▪ Arctic willow leaves have 7–10 times more vitamin C than an orange.
▪ Berries	▪ Cook before eating.	▪ Bears feeding on the berries.	▪ Forms mats in Arctic regions.
▪ Leaves	▪ Infuse to make tea.		▪ Pick individual leaves rather than whole branches and harvest from several shrubs. Northern Labrador tea (*R. tomentosum*) leaves also make a fragrant tea. The plant grows in peaty soils and on tundra.
▪ Berries	▪ Edible raw.	▪ Bears feeding on the berries.	▪ Berries initially pink, ripening to amber in fall. High in vitamin C.
▪ Berries	▪ Eat raw or cooked.	▪ Bears feeding on the berries.	▪ Previous year's berries will last on plant until spring if not picked. Fresh berries can be dried for storage. Low in vitamins.
▪ Leaves	▪ Boil well.		▪ Leaves have a bitter taste so must be well-boiled. High in potassium and vitamin C. Antarctic explorers and whalers used to eat Kerguelen cabbage to ward off scurvy, hence its scientific name.

WILD FOOD

THERE ARE NO POISONOUS SEAWEEDS, but some may cause gastrointestinal upset. The seaweeds listed below are common and safe to eat if gathered when still growing, but should only be eaten in small quantities at first. Don't eat seaweed if you are short of fresh water to drink. Only eat plants that you can positively identify and, if you are in

	NAME	WHERE	IDENTIFICATION
SEAWEEDS	**GUTWEED** *Enteromorpha* spp.	- Saltmarshes and rock pools in cold-temperate waters worldwide.	- Pale or bright green, tubular, unbranched stems, 8–16 in (20–40 cm) long.
	KELP *Alaria* spp., *Laminaria* spp., and *Macrocystis* spp.	- Rocky coasts of the Atlantic and Pacific oceans.	- Very long, strap-like, olive-green to brown fronds. Giant kelp (*M. pyrifera*) is the largest seaweed on Earth, reaching lengths of 150 ft (45 m).
	LAVER *Poryphyra* spp.	- Rocky coasts worldwide.	- Very thin, irregularly shaped membranous frond, up to 20 in (50 cm) long. Color varies from olive-green to purple-brown or blackish.
SEASHORE PLANTS	**COCONUT PALM** *Cocos nucifera*	- Sandy, rocky, and coral coasts in tropics and subtropics worldwide.	- Woody perennial tree, up to 72 ft (22 m) tall, with a crown of pinnate (feather-like) leaves. Gray trunk is ringed with growth scars.
	GLASSWORT/ MARSH SAMPHIRE *Salicornia* spp.	- Salt marshes and coastal mudflats of western and eastern North America, western Europe, and the Mediterranean.	- Bright green, jointed, fleshy stems, 4–12 in (10–30 cm) high. Scale-like leaves and tiny flowers are sunk into the stems.
	ORACHE/SALTBUSH *Atriplex* spp.	- Sandy and shingle beaches worldwide.	- Sprawling plant with spikes of small, greenish flowers. Pale- or silvery-green leaves are either triangular or spear-shaped, sometimes lobed.
	SCREW PINE *Pandanus* spp.	- Tropical coasts from Madagascar to southern Asia and islands in the southwestern Pacific.	- Tree up to 30 ft (9 m) tall, supported by stilt-like aerial roots. Strap-like, saw-toothed leaves are grouped in spirals. Large, globular knobbly fruit ripens from green to orange or red.
	SCURVY-GRASS *Cochlearia* spp.	- Salt marshes and rocky coasts of northern North America and northern Europe and Asia.	- Creeping plant, 4–16 in (10–40 cm) high, with thick, fleshy, dark green, heart-shaped leaves. Small, white, four-petaled flowers.
	SEA BEET *Beta vulgaris maritima*	- Salt marshes, shingle beaches, and cliffs in Europe.	- Sprawling plant up to 40 in (1 m) high, with glossy green, stems and leaves often tinged red. Tiny green flowers grow in clusters on the stems.
	SEAROCKET *Cakile* spp.	- Sandy shores in North America, Europe, Asia, and Australia.	- Up to 16 in (40 cm) high with deeply lobed, fleshy, green leaves. Pale lavender, four-petaled flowers.

COASTAL PLANTS

any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07). Don't eat wild plants if you have any known allergies, a preexisting medical condition, or are pregnant. When planning your trip, it is advisable to familiarize yourself with the edible plants in the region you are visiting and their seasonal availability.

EDIBLE PART	PREPARATION	BEWARE OF	EXTRA INFORMATION
▪ All	▪ Eat raw or dry in the sun and grind to a powder, for ease of storage.		▪ Gutweed is a good indicator of where fresh water runs across a seashore as it is particularly abundant in such sites.
▪ Fronds	▪ Best boiled, though young fronds can be eaten raw.	▪ Large doses of iodine can be harmful.	▪ Sugarwrack (*L. saccharina*) tastes sweet, as its name implies, and can be eaten fried or boiled. Kelp is high in iodine, small amounts of which are needed for good health.
▪ All	▪ Boil, then mash.		▪ The Welsh make a cake of out of laver and oatmeal, called laverbread. The Japanese produce thousands of tons of paper-thin sheets of dried, processed laver, called nori.
▪ Seeds (coconuts)	▪ Drink the milk fresh. ▪ Eat the flesh raw or sun-dried.	▪ Mature coconut milk needs extra water to be digested. It's also a laxative, so don't drink too much.	▪ Green, unripe coconuts are a good source of water (see p. 190). Coconut milk is rich in sugar and vitamins and contains protein. The oily flesh is also nutritious.
▪ Stems	▪ Steam or boil.		▪ Also known as Poor Man's Asparagus. Rock samphire (*Crithmum maritimum*), which is found on cliffs and rocky shores, is unrelated but has edible succulent leaves.
▪ Leaves	▪ Young leaves can be eaten raw, older leaves should be boiled.		▪ *Atriplex* species are restricted to saline soils. As well as coasts, they are also found inland on the shores of alkaline lakes and in deserts. Gray saltbush (*A. cinerea*) is a type of bushfood in Australia.
▪ Fruit	▪ Inner flesh of ripe fruit can be eaten raw. Partly ripe fruit can be eaten if baked for two hours.	▪ The leaves have small, thorn-like spines at the edges, which can scrape and irritate the skin.	▪ Unripe, green fruit is inedible.
▪ Leaves	▪ Eat the leaves raw or pulp and drink.		▪ Scurvy-grass leaves are rich in vitamin C— sailors used to eat (or drink) them to prevent scurvy. They are very bitter, so are best leached in water, but you should eat them in a survival situation.
▪ Leaves	▪ Eat raw or boiled.		▪ Sea beet is the wild ancestor of beets, sugar beets, and Swiss chard.
▪ Leaves ▪ Young pods	▪ Leaves and young pods can be eaten raw.		▪ Searocket leaves have a peppery taste. One species of *Cakile* grows in the deserts of the Arabian Peninsula.

WILD FOOD

ALL MAMMALS ARE EDIBLE, but some species or groups of mammals are in danger of extinction and are protected by law; the mammal groups listed below contain many species that are common in their areas of distribution, and are not usually dangerous if approached and handled correctly. However, when preparing for your trip, you must

	NAME	WHERE	IDENTIFICATION
AERIAL	**BATS** *Chiroptera*	▪ Temperate and tropical regions worldwide.	▪ The only mammals with true wings and the ability to fly. Wingspans vary from 6 in (15 cm) to over 5 ft (1.5 m) in the large fruit bats, or flying foxes (*Megachiroptera*).
TREES / GROUND	**TREE SQUIRRELS** *Sciurus*	▪ North and South America, Europe, and temperate parts of Asia.	▪ Small or medium-sized rodents, with large, bushy tails.
	PORCUPINES *Hystricomorpha*	▪ North and South America, Africa, and tropical Asia.	▪ Large, rounded rodents, with a coat of sharp quills.
GROUND / BURROWS	**ANTELOPE AND DEER** *Bovidae* and *Cervidae*	▪ North and South America, Africa, and Eurasia.	▪ Hoofed mammals, with long legs and barrel-shaped bodies. Male antelope have permanent horns; male deer have antlers, which they shed and regrow each year.
	GUINEA PIGS *Cavia* spp.	▪ Northwestern, central, and southeastern South America.	▪ Small rodent, with dark, coarse fur, short legs, and no tail.
	HEDGEHOGS *Erinaceinae*	▪ Europe, Africa, and Asia; introduced to New Zealand.	▪ Small, short-legged insectivorous mammal, with spines on back and sides of body.
	KANGAROOS *Macropus* spp.	▪ Australia.	▪ The largest of the marsupials (pouched mammals), with long, strong tails, large back legs, and short forelegs.
	RABBITS AND HARES *Leporidae*	▪ Most habitats worldwide, from arctic tundra to semi-desert.	▪ Small or medium-sized herbivorous mammals, with large ears and small, round, furry tails. Fur is usually brown or dark gray; the Arctic hare (*Lepus arcticus*) has a winter coat: white with black ear tips.
GROUND / AQUATIC	**BEAVER** *Castor* spp.	▪ North America and northwestern Eurasia.	▪ Large semi-aquatic rodents, with coarse, brown fur, webbed feet, and a flat, scaly tail.
	CANE RATS *Thryonomys* spp.	▪ Sub-Saharan Africa.	▪ Large rodents, with coarse, brown fur and a long, naked tail. The greater cane rat (*T. swinderianus*) is semi-aquatic.

MAMMALS

familiarize yourself with those mammals that can be hunted in the region you are visiting and their seasonal movements. Some countries and/or states will require you to have a hunting license or permit. For trapping and preparation methods, see pp. 216–23, and for cooking suggestions, see pp. 204–05.

HOW TO FIND	BEWARE OF	EXTRA INFORMATION
▪ Active at night. Roost in colonies in caves and trees during the day. Temperate species hibernate in the winter.	▪ Sharp teeth. Some bats carry rabies or other diseases (see pp. 300–01), so cook the meat thoroughly. Bats are protected in Europe and some parts of North America.	▪ Large, plump, fruit-eating bats, such as the Indian flying-fox (*Pteropus giganteus*), are considered a delicacy in many tropical areas.
▪ Mostly active by day, feeding in branches and on the ground on shoots, nuts, and birds' eggs. Nest in trees in a range of woodland habitats.	▪ Sharp teeth and claws.	▪ Flying squirrels (*Pteromyini*) are nocturnal, tree-living members of the squirrel (*Sciuridae*) family. They glide from tree to tree rather than fly.
▪ Some New World porcupines climb trees to feed, but all Old World porcupines spend all their time on the ground across a range of habitats.	▪ Barbed quills (see p. 266 for how to remove).	▪ Slow-moving, so can easily be run down and speared. Unlike most small game, the meat is fatty rather than lean.
▪ Antelope are mainly found in savanna, or marshland habitats, and deer mainly in woodland, forests, or tundra.	▪ Horns or antlers.	▪ Most active at dawn and dusk. Usually live in groups.
▪ Feed mostly at dawn and dusk, in shrubby grassland in mountainous regions.	▪ Sharp teeth.	▪ Have communal feeding runways. Traps can be baited with leafy vegetables.
▪ Habitats range from woodland, hedgerows, and grassland to desert. Feed at night on small animals, such as worms and insects.	▪ Sharp spines. Usually infested with parasites, so handle carefully and cook well.	▪ Traditionally prepared for eating by covering with clay and baking in embers of a fire (see p. 205).
▪ Live in open savanna woodland, feeding mainly at night on vegetation. Gather at waterholes in times of drought.	▪ Sharp claws and a powerful kick.	▪ Kangaroos are protected in some Australian states, but can be hunted (if you have a permit) in others.
▪ Rabbits live in burrows, often in large numbers, coming above ground to feed on vegetation, using well-worn runs. Hares live above ground.	▪ Rabbits and hares may be infected with germs, so handle them carefully and cook well. Their meat is very lean, and must be eaten with green vegetables (see p. 220).	▪ In most areas, these are the first animals to try to trap. The European rabbit (*Oryctolagus cuniculus*) is an invasive pest species in Australia.
▪ Look for the distinctive mud-and-stick lodges in the middle of a pond or lake. Beavers leave the lodge at night to feed on water plants and trees along the riverbank.	▪ Strong, chisel-like teeth. The Eurasian beaver (*Castor fiber*) is protected in many of the countries in which it is found.	▪ Use regular runs along streams. The tail can be eaten as well as the flesh.
▪ Feed at night on reeds and grasses in marshland and on riverbanks, or on grasses in moist savanna, and on rocky hillsides.	▪ Sharp teeth.	▪ Cane rats are a valuable source of bush meat in West and Central Africa. "True" rats (*Rattus* spp.) are edible but often carry diseases, so handle with care.

WILD FOOD

ALL BIRDS ARE EDIBLE, although a few taste horrible (birds-of-paradise, for example), and pitohuis have poisonous feathers and skin. Some species or groups of birds are in danger of extinction and are protected by law; the bird groups listed below contain many species that are common in their areas of distribution. When planning your trip,

NAMES	WHERE	IDENTIFICATION
GAMEBIRDS *Galliformes*	▪ Worldwide, except Antarctica.	▪ Typically have plump bodies, small heads, and short, rounded wings. Range in size from quail, which can weigh as little as 2 oz (50 g), to the wild turkey (*Meleagris galloparvo*), which can be as heavy as 22 lb (10 kg).
OSTRICH *Struthio camelus*	▪ From West to East Africa south of the Sahara, and southern Africa.	▪ The world's tallest and heaviest bird, reaching 220 lb (100 kg), the ostrich is flightless. Long neck and long, strong legs, with two-toed feet. Males have black and white plumage, females brown.
OWLS *Strigiformes*	▪ Every continent, except Antarctica.	▪ Upright posture and a flattened face with large, forward-facing eyes. Strong, hooked bill and sharp talons (claws).
PIGEONS AND DOVES *Columbiformes*	▪ Worldwide, except the polar regions.	▪ Plump, full-breasted bodies, with a small head and bill. Thick, soft plumage, which can be brown or gray, or brightly colored in some tropical species.
STORKS AND HERONS *Ciconiiformes*	▪ Worldwide, except the polar regions.	▪ All have large bodies, with long legs, long necks, and bills. Range height from 10 in (25 cm) for the smallest bitterns to 5 ft (1.5 m) for the largest storks.
VULTURES *Accipitridae* and *Cathartidae*	▪ All continents, except Antarctica and Australasia.	▪ Large carrion-eating birds, with large wings, powerful feet with sharp talons, and a strongly hooked bill. Head and neck is often bald or sparsely feathered.
WADERS AND GULLS *Charadriiformes*	▪ Worldwide.	▪ Vary greatly in size and shape, but most have subdued brown, gray, black, or white plumage. Waders (or shorebirds) usually have lightweight bodies and long legs. Gulls have compact bodies.
WATERFOWL *Anseriformes*	▪ Worldwide, except Antarctica.	▪ Typically have plump bodies, powerful wings, and short legs with webbed feet. Most species have a broad, flattened bill. Ducks usually have short necks, while geese and swans have long necks.

BIRDS

you must familiarize yourself with those birds that can be hunted in the region you are visiting and their seasonal availability. Some countries and/or states will require you to have a hunting license or permit. For trapping and preparation methods, see pp. 226–29; for cooking suggestions, see pp. 204–05.

HOW TO FIND	BEWARE OF	EXTRA INFORMATION
▪ Wide range of terrestrial habitats, from high mountains, to tropical forests, to Arctic tundra.	▪ Short but sharp beak and claws. In some species (including pheasants), the males have sharp spurs on their ankles.	▪ Nearly all gamebirds feed and nest on the ground. At night, many gamebirds roost in trees.
▪ Open semi-arid plains (from desert to savanna) and open woodland.	▪ Can run fast and deliver a powerful kick. Very protective of eggs in the nest and the young birds.	▪ The eggs are the largest of any bird, weighing 3 lb (1.4 kg). The nests are communal and contain up to 40 eggs.
▪ In most habitats, from tundra to dense forest.	▪ Sharp bill and talons.	▪ Most species hunt at night and roost in trees during the day. Usually nest in holes in trees, under rocky overhangs, or, sometimes, in burrows.
▪ Variety of forest habitats, from open woodland to dense, tropical rain forest, and grassland and semi-arid areas.		▪ Often feed in flocks on the ground or in trees, and usually roost communally in trees. Can be taken by hand from the roost if you approach slowly and quietly.
▪ Freshwater habitats, ranging from wetlands, rivers, lakes, marshes, swamps, mangroves, and lagoons to tidal mudflats.	▪ Sharp bill.	▪ Usually feed alone at the water's edge, but often nest communally in trees.
▪ Open areas, including mountains, plains, deserts, and savannas.	▪ Sharp bill and talons. Handle as little as possible, as prone to parasites and infection. Boil meat for at least 30 minutes.	▪ Carrion eaters and scavengers, such as crows (*Corvus* spp.) and gulls (see below), can be lured into traps baited with meat.
▪ Tundra and a variety of wetland and coastal habitats, including tidal mudflats, beaches, and cliffs.	▪ Sharp bill. Gulls and terns will aggressively defend their nest sites.	▪ Most birds in this group nest on the ground or on rock ledges, laying about four well-camouflaged eggs. Often nest in large colonies.
▪ Arctic tundra, wetlands, rivers, and lakes. Ducks and swans usually feed on the water, while geese graze on the land.	▪ Geese and swans can be very aggressive, particularly in the breeding season. Mute swans (*Cygnus olor*) can weigh up to 26 lb (12 kg).	▪ Many species migrate between their breeding grounds in the Arctic and their wintering grounds further south. They molt their feathers in the late summer, rendering them flightless for a couple of weeks and thus easier to catch.

WILD FOOD

SOME AMPHIBIAN AND REPTILE SPECIES are at risk of extinction and are protected by law; the animals listed below are common in their areas of distribution, and are non-venomous. As part of preparation for your trip, it is advisable to familiarize yourself with the reptiles and amphibians in the region you are visiting. Many countries and/or states

	NAME	WHERE	IDENTIFICATION
FROGS	**AFRICAN CLAWED FROG** *Xenopus* spp.	▪ Sub-Saharan Africa; introduced to North America, South America, and Europe.	▪ Flattened, brown body, 2¼–5 in (6–13 cm) long, with a line of white "stitch marks" along the sides. Clawed toes on the muscular legs.
	AMERICAN BULLFROG *Rana catesbeiana*	▪ North America; introduced to Europe and Asia.	▪ Body is green with brown markings above and white below, 3½–8 in (9–20 cm) long. Large legs and large eardrums.
	SOUTH AMERICAN BULLFROG *Leptodactylus pentadactylus*	▪ Central and northern South America.	▪ Smooth yellow or pale brown body with some dark markings, 3–9 in (8–22 cm) long.
TURTLES	**AFRICAN HELMETED TURTLE** *Pelomedusa subrufa*	▪ Sub-Saharan Africa.	▪ Flattened, brown shell, 8–12½in (20–32 cm) long.
	ASIAN LEAF TURTLE *Cyclemus dentata*	▪ Southeast Asia.	▪ Oval, light to dark brown shell, 6–9½in (15–24 cm) long, with serrated edge near tail. Reddish brown head and legs.
	PAINTED TURTLE *Chrysemys picta*	▪ Southern Canada, USA, and northern Mexico.	▪ Flattened, smooth shell, 6–10 in (15–25 cm) long; brown above and yellow below, sometimes patterned. Yellow or red stripes on the neck.
SNAKES	**CARPET PYTHON** *Morelia spilota*	▪ Indonesia, New Guinea, and Australia.	▪ Several subspecies, all of which have a bold pattern of irregular markings, which can be reddish brown, brown, black or gray. Average length is 6½ft (2 m), but can reach 13 ft (4 m).
	COMMON BOA *Boa constrictor*	▪ Central America, South America, and some Caribbean islands.	▪ Several subspecies but all have characteristic dark saddle markings along the back and a dark stripe behind each eye. Narrow head and pointed snout.
	COMMON EGG-EATING SNAKE *Dasypeltis scabra*	▪ Sub-saharan Africa	▪ Reddish-brown or gray body with dark, angular markings. A slender snake, 28–39 in (70–100 cm) long. Rounded snout.
	COMMON RATSNAKE *Elaphe obsoleta*	▪ Southern Canada, and central and eastern USA	▪ Subspecies coloration varies from bright yellow-orange to pale gray with darker blotches. Length range is 4–6 ft (1.2–1.8 m). Long head with rounded snout.
LIZARDS	**COMMON/VIVIPAROUS LIZARD** *Zootoca vivipara*	▪ Europe, extending to the Arctic Circle, and central to eastern Asia, including Japan.	▪ Body is usually brown olive, sometimes black, with males having bright yellow or orange bellies and females creamy white ones. Length range is 4–7 in (10–18 cm).
	GREEN IGUANA *Iguana iguana*	▪ Central America and northern South America.	▪ Green or grayish body, 3¼–6½ft (1–2 m) long, including the long striped tail. Long, stout legs for climbing trees. Adults have a fleshy dewlap beneath the throat, which is large in males.
	RAINBOW LIZARD *Agama agama*	▪ West, Central, and East Africa.	▪ Males turn brightly colored in the sun, with an orange-red head and a blue or turquoise body; females and juvenile males remain gray. Length range is 12–16 in (30–40 cm).
	WESTERN FENCE LIZARD *Sceloporus occidentalis*	▪ Southwestern USA and northwestern Mexico.	▪ Brown body with raised, pointed scales. Blue patches on belly are most pronounced in males. Length range is 6–9 in (15–23 cm).

AMPHIBIANS AND REPTILES

will require you to have a hunting license or permit. See p. 218 for some trapping methods and p. 224 for how to prepare the animal for cooking. Avoid brightly colored tropical frogs as they are often highly toxic. In addition, don't eat box turtles, as they sometimes eat poisonous fungi and their flesh may be toxic.

HOW TO FIND	SIMILAR SPECIES	EXTRA INFORMATION
▪ Wetlands and waterholes in savannas and tropical and subtropical forests.		▪ Also known as the African clawed toad.
▪ Wetlands.	▪ All frogs in the *Rana* genus are edible. The genus includes New World true frogs and the European common frog (*R. temporaria*).	▪ American bullfrog legs can be the size of chicken drumsticks. All temperate species of frog hibernate in the winter.
▪ Tropical forests and rain forests.		▪ Active at night. If picked up, the South American bullfrog emits a loud scream to startle the predator into dropping it.
▪ Watering holes and rain pools in open country.		▪ In the rainy season, this turtle wanders from pool to pool to forage. In the dry season, it buries itself in mud.
▪ Shallow streams in mountains or lowlands.		▪ Very active on land and in water.
▪ Lakes, ponds, and slow-moving streams and rivers.		▪ Active by day. You will often find several painted turtles piled up together on a log in the water, basking in the sun.
▪ Wide range of habitats.		▪ The carpet python is non-venomous, but it can inflict a nasty bite. It is one of Australia's most widespread snakes.
▪ Wide range of habitats, from tropical forest to dry savanna.	▪ There are several species in the boa family, which includes the world's largest snake, the green anaconda (*Eunectes murinus*).	▪ Kills its prey by coiling around it and suffocating it. Don't tackle large common boas—they can be up to 13 ft (4 m) long.
▪ Desert and open habitats.		▪ This nocturnal African snake feeds on eggs during the bird-breeding season then fasts for the rest of the year. It has no teeth.
▪ Rocky hillsides with open woodland.	▪ Ratsnakes (*Elaphe* spp.), which include the brightly colored corn snake (*E. guttata*), are found in both the New and Old Worlds.	▪ Like the common boa (see above), all ratsnakes kill their prey by constriction.
▪ Wide range of terrestrial habitats.		▪ There's no point trying to catch a common lizard by its tail as, like many lizards, it can shed the tail and regrow it.
▪ Tropical forest and rain forest.		▪ Defends itself by lashing out with its tail and claws.
▪ Open habitats.		▪ Rainbow lizard can shed and regrow its tail.
▪ On rocks and other prominent places in temperate and coniferous woodlands.	▪ The eastern fence lizard (*S. undulatus*) is widely distributed across southeastern USA and northeastern Mexico.	▪ Fence lizard can shed and regrow its tail.

WILD FOOD

WHEN PLANNING YOUR TRIP, it is advisable to familiarize yourself with the species of fish in the region you are visiting, their seasonal availability, and the best method of catching them (see pp. 208–09 for how to make tackle and techniques to use in a survival situation). Many countries and/or states will require you to have a fishing

NAME	WHERE	IDENTIFICATION
BARRAMUNDI *Lates calcarifer*	▪ Western and eastern Indian Ocean, and northwestern and western Pacific Ocean.	▪ Up to 6½ft (2 m) long, with a rounded tail fin. Dark greenish-gray upper body fading to silver below.
COMMON/BRONZE BREAM *Abramis brama*	▪ Europe to Central Asia.	▪ Up to 2¾ft (82 cm) long, with a deep, narrow, bronze-colored body and a deeply forked tail.
COMMON/KING CARP *Cyprinus carpio*	▪ Europe and Asia; introduced in North America and Australasia.	▪ Up to 5 ft (1.5 m) long, with a dark gold upper body paling to silver below.
EUROPEAN TURBOT *Psetta maxima*	▪ Northeast Atlantic Ocean up to the Arctic Circle and the Mediterranean.	▪ Almost circular flatfish, up to 3¼ft (1 m) in diameter. Sandy brown with brown or black speckles on top.
NILE PERCH *Lates niloticus*	▪ Northern, central, and eastern Africa.	▪ Up to 6¼ft (1.9 m) long, with a rounded tail fin. Dark gray-blue upper body fading to silver below.
RUDD *Scardinius erythrophthalmus*	▪ Europe and Asia.	▪ Up to 17 in (45 cm) long, with a deeply forked tail. Dark greenish upper body paling to silver below. Red pelvic, pectoral, and anal fins.
ATLANTIC SALMON *Salmo salar*	▪ North Atlantic and Arctic oceans, and Baltic Sea and adjacent rivers; introduced to Argentina and Australasia.	▪ Up to 5 ft (1.5 m) long, with a powerful, streamlined, silver blue-green body.
TARPON *Megalops atlanticus*	▪ Eastern and western Atlantic Ocean, the Gulf of Mexico, and the Caribbean.	▪ Up to 8 ft (2.5 m) long, with a deeply forked tail. Bright silver body, with large, hard scales.
TENCH *Tinca tinca*	▪ Europe and Asia; introduced in North America.	▪ Olive green upper body, golden below, with a square tail fin. Small barbel at corners of mouth. Average length is 27 in (70 cm).
BROWN TROUT AND SEA TROUT *Salmo trutta*	▪ Temperate waters worldwide.	▪ Streamlined body with an average length of 3¼ft (1 m). The brown trout has a brownish body flecked with distinctive black and red spots. The sea-going form, the sea trout, is silver-blue with black spots.

FISH

license or permit, and some species of fish can only be caught at certain times of the year. Before fishing, you should also check with the locals that fish in the area are safe to eat—in some regions the waters may be contaminated. See pp. 212–13 for how to prepare a fish for cooking.

HOW TO FIND	SIMILAR FISH	EXTRA INFORMATION
- Tropical, slow-moving creeks and estuaries, hiding in mangrove roots and rocky outcrops.		- Beware of the sharp dorsal fin.
- Bottom of still and slow-moving waters, typically in lakes, rivers and ponds. Swim in large shoals.	- Silver bream (*Blicca bjoerkna*) and white-eye bream (*Abramis sapa*).	
- Still and slow-moving waters, often among vegetation near banks of lakes or ponds.	- Wild carp, mirrored carp, and leather carp are all varieties of the common carp.	
- Sandy, rocky, or mixed sea beds in shallow coastal and brackish water.	- Other flatfish (*Pleuronectiformes*) found in shallow water, such as sole, plaice, and flounder.	
- Lakes and large rivers.	- Other perch (*Perciformes*).	- Beware of the sharp dorsal fin and teeth (the Nile perch is a voracious predator). This species weighs up to 440 lb (200 kg).
- Still and slow-moving waters, often near banks of ponds and marshlands.	- Roach (*Rutilus rutilus*)	
- Cold, fast-flowing rivers and coastal waters.	- Pacific salmon (*Oncorhynchus* spp.)	- The salmon run (when the fish migrate from the sea to their birth rivers) is in the autumn, with spawning between November and January.
- Estuaries, lagoons, tidal flats and mangrove swamps. Feed in shoals.	- Indo-Pacific tarpon (*M. cyprinoides*), which is smaller and rarer.	- This species is hard to hook because of its bony mouth. It weighs up to 160kg (350lb).
- Still and slow-moving waters, in ponds, lakes and rivers. The best time of day to catch tench is at dawn, at the edges of dense vegetation.		
- Streams, rivers, lakes, and coastal waters.	- Rainbow trout and steelhead trout (*Oncorhynchus mykiss*) and the cutthroat trout (*O. clarki*).	- Sea-going forms migrate from the sea to rivers to spawn in the autumn. Beware of small bones when eating.

WILD FOOD

SOME SPECIES OF INVERTEBRATE are in danger of extinction and are protected by law; the animals listed below are common in their areas of distribution. When preparing for your trip, it is advisable to familiarize yourself with the edible invertebrates in the region you are visiting and the best way to catch them (see p. 219 for some techniques to use in a survival

	NAME	WHERE	IDENTIFICATION
MOLLUSCS	**COMMON/BLUE MUSSEL** *Mytilus edulis*	▪ Intertidal zones of estuaries and coasts of North and southeastern Atlantic and northeastern and southwestern Pacific.	▪ Dark-shelled bivalve, reaching 4–6 in (10–15 cm) in length.
	GIANT AFRICAN LAND SNAILS *Achatina* spp.	▪ Sub-tropical and tropical parts of Africa; introduced to the Asia-Pacific region.	▪ Large terrestrial snails, reaching up to 12 in (30 cm) in length when body is extended, and 4 in (10 cm) in diameter. Whorled, conical, brownish shell with darker bands.
ARTHROPODS	**GRASSHOPPERS** *Acrididae*	▪ Worldwide, on vegetation and the ground.	▪ Winged insects with powerful hindlegs, ⅖–3¼ in (1–8 cm) long, usually with camouflage coloring and patterning.
	HONEYPOT/HONEY ANTS *Myrmecocystus* spp.	▪ Semi-arid and arid regions of southwest USA, Mexico, Africa, and Australia.	▪ Segmented bodies with six legs and a pair of antennae, and a constricted waist. Range in colour and size, from ⅒–½in (2–12mm). Special members of the colony, called repletes, have abdomens that can swell to the size of a grape.
	WITCHETTY GRUB *Endoxyla leucomochla*	▪ Underground, in the roots of the witchetty bush (*Acacia kempeana*), central Australia.	▪ The larva of this species of cossid moth reaches about 2¾ in (7 cm) in length. It is white with a brown head.
CRUSTACEANS	**CRAYFISH/CRAWFISH/ YABBIES/KOURAS** *Astacidea*	▪ Worldwide, in freshwater streams.	▪ Segmented bodies, which can be sandy yellow, green, dark brown, or blue-gray in color. Ten legs, the front two of which are large claws. Average length is 4 in (8 cm), but some species grow much larger.
	WOODLICE *Oniscidea* and *Armadillidiidae*	▪ Worldwide, in damp terrestrial microhabitats, such as in rotting wood and leaf litter.	▪ Flattened, segmented bodies up to ¾ in (2 cm) in length. Gray or light brown or black in color. Pill millipedes will roll up into a ball when threatened.
WORMS	**COMMON EARTHWORM** *Lumbricus terrestris*	▪ Temperate regions of Europe; introduced to most parts of the world.	▪ Reddish worm, reaching 14 in (35cm) in length when extended.

INVERTEBRATES

situation). Many countries and/or states will require you to have a hunting license or permit, and some species of shellfish can only be caught at certain times of the year. Also check with the locals that the shellfish are safe to eat—in some regions the waters may be contaminated. See pp. 224–25 for some preparation and cooking methods.

SIMILAR INVERTEBRATES	BEWARE OF	EXTRA INFORMATION
▪ Most species of marine and freshwater bivalves and univalves are edible if alive and healthy when gathered.	▪ Don't collect marine shellfish that aren't covered at high tide. The black mussel (*Musculus niger*), found in Arctic waters, can be poisonous year-round. Mussels in tropical zones are poisonous during the summer.	▪ Healthy bivalves should close their shells when tapped and the shells should open when cooked. Healthy univalves, such as limpets, should be difficult to pry off the rocks on which they live.
▪ Many terrestrial and freshwater snails are edible. The winkle or common periwinkle (*Littorina littorea*) is an edible sea snail.	▪ Avoid any terrestrial snails with brightly colored shells and all sea snails unless you can positively identify them as an edible species, as they may be poisonous.	▪ Snails should be starved for 24 hours or fed a diet of edible green leaves to purge their guts before cooking. Boil for at least 10 minutes.
▪ Most species of cricket (*Gryllidae*) and katydid (*Tettigoniidae*) are edible.	▪ Avoid any brightly colored grasshoppers as they may be toxic.	▪ Remove the antennae, wings, and leg spurs, and roast to kill any parasites.
▪ Most species of ant are edible if gathered carefully. The larvae of wood ants (*Formica* spp.) in northern temperate regions make a nutritious meal in summer months. The abdomen of the green or weaver ant (*Oecophylla smaragdina*) found in southeast Asia and Australia tastes like citrus sherbert. Termites (*Isoptera*) are also edible. See p. 219 for catching methods.	▪ Most species of ant are aggressive in defense of their nest and have a stinging bite. Some will then squirt formic acid at the site of the bite.	▪ The swollen abdomens of the repletes contain a nutritious fluid. ▪ The repletes live deep underground in the ants' nest so you would have to dig them out.
▪ The term "witchetty grub" is also applied to the edible larvae of other cossid moths, ghost moths (*Hepialidae*), and longhorn beetles (*Cerambycidae*). Palm grubs are the edible larvae of the palm weevil (*Rhynchophorus* Spp.). They live in the trunks of sago palms (see pp. 284–285) in Southeast Asia.	▪ Do not eat any insect larvae that are already dead when you find them, or that look sick, smell bad, or that irritate the skin if handled.	▪ Should be eaten raw or roasted quickly in hot embers.
▪ Marine decapods such as crabs, lobsters, and prawns are also edible.	▪ May contain harmful parasites so cook well.	▪ Keep alive until ready to eat. See p. 225 for preparation and cooking methods.
		▪ Boil or fry gently. Will turn slightly pink, like shrimp, when cooked.
▪ All earthworms are edible.		▪ The common eathworm is unusual in that it feeds at the surface, making it easier to find and gather than other worms.

NATURAL DANGERS

FOR INFORMATION ON LARGER dangerous animals, such as bears, big cats, and sharks, and venomous animals, such as snakes, see pp. 242–43. When planning your trip, it is advisable to familiarize yourself with the potentially harmful wildlife in the region you are visiting, the nature of the threat, and how to avoid it if possible. In the case of suspected

	NAME	WHERE	IDENTIFICATION
MAMMALS	**VAMPIRE BATS** *Desmodontinae*	▪ In rain forests, deserts, and grasslands, from Mexico to northern and central South America.	▪ Small bats, with a body length of 2¾–3½ in (7–9 cm) and a wingspan of 14–16 in (35–40 cm). Dark brownish gray fur, paler on the belly. Razor-like upper incisors.
	RATS *Rattus* spp.	▪ Worldwide, except for the polar regions.	▪ Medium-sized rodents, with a minimum body length of 5 in (12 cm), and long, naked tails. Coarse brown, gray-brown, or black fur, paler on belly. Pink feet.
SNAKES	**SPITTING COBRAS AND RINKHALS** *Naja* spp. and *Hemachatus haemachatus*	▪ Warm temperate, subtropical, and tropical regions of Asia and Africa.	▪ Pale brown to black snakes up to 6 ft (2 m) long. When threatened, will rear up and spread their hoods, revealing striped patterns on their necks.
FISH	**ELECTRIC EEL** *Electrophorus electricus*	▪ Amazon and Orinoco river systems, South America, often in shallow water.	▪ Not a true eel, but an eel-like fish up to 8¼ ft (2.5 m) long and can be as thick as a human thigh. Continuous fin along lower body.
	STONEFISH *Synanceia verrucosa*	▪ In tropical coastal waters of northern Indian Ocean and southwestern Pacific Ocean.	▪ Up to 16 in (40 cm) long, this lumpy fish can match its color to the rock or sediment on which it lies while waiting for passing prey. Venomous spines in the dorsal fin.
OTHER ANIMALS	**BOX JELLYFISH/ SEA WASPS** *Cubomedusae*	▪ At the surface of tropical waters of southwest Pacific and eastern Indian Ocean.	▪ Box-shaped, transparent jellyfish, up to 10 in (25 cm) in diameter. Up to 15 long tentacles, which bear stinging cells, at each corner.
	CENTIPEDES *Chilopoda*	▪ In soil, leaf litter, cracks, and crevices in temperate, subtropical, and tropical regions worldwide.	▪ Elongated, flattened bodies divided into at least 16 segments, most of which have one pair of legs. Tropical species are typically brightly colored—yellow, red, orange, or green—with dark stripes. The world's largest centipede, *Scolpendra gigantea*, reaches 12 in (30 cm) in length.
	CONE SHELLS/ CONE SNAILS *Conidae*	▪ In inter-tidal zones and on coral reefs in warm temperate, subtropical, and tropical seas and oceans worldwide.	▪ Brightly colored and patterned, cone-shaped shells, up to 9 in (23 cm) long. These predatory snails have venomous, harpoon-like mouthparts.

WILD ANIMALS

poisoning, call the emergency services immediately. For further information regarding treatment for bites and stings, see pp. 266–67. If a casualty becomes unconscious, open the airway and check breathing (see p. 276). Be ready to begin CPR—chest compressions and rescue breaths (see p. 277), use a plastic face mask or face shield, if you have one.

DANGER	TREATMENT	HOW TO AVOID
▪ Many mammal species are infected with rabies, but because vampire bats feed by biting their victims, then lapping blood from the wound, the virus can be transmitted via their saliva. Symptoms include fever, headache, fear (especially of water), and seizures.	▪ If bitten by a vampire bat, or other potentially rabid mammal, clean the wound thoroughly and seek immediate medical help. Symptoms usually appear 2–8 weeks after infection, by which time the disease is almost always fatal.	▪ Vampire bats roost in hollow trees and caves, emerging at night to feed. Don't shelter in bat caves (bat dung can cause a respiratory disease). Keep covered at night, and sleep under a mosquito net. Get vaccinated before traveling to a high-risk area.
▪ Can be infested with parasites and can also carry infectious diseases. Leptospirosis is a water-borne bacterial disease spread by infected rats' urine; the acute form in humans is called Weil's disease.	▪ If you've been in contact with potentially contaminated water and develop flu-like symptoms (fever, headache, muscle pain) you must get your blood tested as soon as possible. Antibiotics are effective in the early stages of the disease.	▪ If handling a rat, beware of its teeth and claws, and wash thoroughly afterward. Cook the meat thoroughly. In rat-infested areas avoid getting water into any cuts; if you do, wash with soap and boiled water.
▪ Will eject a fine spray of venom from their fangs upward the eyes of an aggressor. Causes intense pain, tearing, and discharge from the eyes, spasm and swelling of the eyelids. Ulceration of the cornea can lead to infection resulting in blindness.	▪ Call the emergency services. Rinse the eyes with cool water for 10 minutes. Apply an antibiotic eye ointment.	▪ If you come across a spitting cobra, try to remain completely still and allow the snake to move away, if possible (see p. 243).
▪ Can generate shocks of up to 600 volts, used to stun or kill prey when hunting, or in self-defense. Such a shock can be discharged up to 8 hours after its death, and is potentially lethal to humans.	▪ Call the emergency services. If the casualty becomes unconscious, open the airway and check breathing (see p. 276). Be ready to begin CPR (see p. 277).	▪ Keep your boots on if crossing a tributary and take care when washing. These waters are also home to piranhas (*Serrasalminae*), which have razor-sharp teeth.
▪ The world's most venomous fish. If trodden on, the spines inject poison into the puncture wound. This is extremely painful and can be fatal. The injured body part will swell and muscular paralysis may set in.	▪ Call the emergency services—anti-venin is available. Immerse the injured part in water as hot as the casualty can stand for at least 30 minutes (see p. 267). Don't immobilize the injured limb. Be ready to begin CPR if the casualty loses consciousness (see pp. 276–77).	▪ Wear something on your feet when walking in shallow water. Weaverfish (*Trachinidae*) and scorpionfish (*Scorpaenidae*) also have venomous stings; the treatment is the same.
▪ Sting is extremely painful—it will damage the skin and can cause permanent scarring. In severe cases, box jellyfish (*Chironex fleckeri*) stings can lead to cardiac arrest and death within minutes.	▪ Call the emergency services—anti-venin is available. Flood the stung area with vinegar or sea water for at least 30 seconds to neutralize the stinging cells (see p. 267). Be ready to begin CPR if the casualty loses consciousness (see pp. 276–77).	▪ Never touch jellyfish, even when dead. All jellyfish, sea anemones, corals, and the Portuguese man-of-war (*Physalia physalis*) release venomous cells when touched. Some are more toxic than others.
▪ Have a pair of large, venomous claws on their heads, which in some species can be dangerous to humans. The sting is painful and leads to swelling. May result in anaphylactic shock.	▪ Call the emergency services. Apply a cold compress for at least 10 minutes to reduce pain and swelling. Monitor the casualty for anaphylactic shock (see p. 274) and be ready to begin CPR (see p. 277).	▪ As a general rule, avoid all multiple-legged arthropods. If you discover one on your body, gently brush it off in the direction in which it is walking.
▪ Will fire its harpoon in self-defense if touched. Sting of most species is painful; that of the larger tropical species can be fatal. Symptoms include swelling and numbness, and may be followed by severe breathing difficulties.	▪ Call the emergency services and treat as for snake bite (see p. 266). Monitor the casualty for shock (see p. 274) and be ready to begin CPR (see p. 277).	▪ Don't pick up any cone shells in rock pools or when snorkeling or diving. (There is no anti-venin: treatment entails providing life support until the venom is metabolized by the casualty's body.)

NATURAL DANGERS

IN THE CASE OF SUSPECTED POISONING from a spider bite or a scorpion sting, you must call the emergency services immediately. For further information regarding treatment for bites and stings, see pp. 266–67. If a casualty becomes unsconscious, open the airway and check breathing (see p. 276). Be ready to begin CPR—chest compressions and rescue

	NAME	WHERE	IDENTIFICATION
INSECTS	**BEES, WASPS, AND HORNETS** *Hymenoptera*	▪ In a variety of terrestrial habitats worldwide.	▪ Narrow-waisted bodies, often striped black and yellow. Bees are hairy, while wasps and hornets have hairless bodies. Up to 1½ in (3.5 cm) long. Bees have barbed stingers, wasps and hornet stingers are smooth.
	MOSQUITOES *Culicidae*	▪ In a variety of terrestrial habitats near water worldwide, especially in warmer regions.	▪ Narrow bodies and long, slender legs. Up to ¾ in (2 cm) long. Females have syringe-like mouthparts for piercing skin.
ARACHNIDS	**BROWN RECLUSE/ FIDDLEBACK SPIDER** *Loxosceles reclusa*	▪ Southern midwestern states of USA, south to the Gulf of Mexico.	▪ Small brown spider, about ½ in (1.25 cm) long) with a distinctive dark, violin-shaped mark on the back of its cephalothorax (head).
	FUNNEL-WEB SPIDERS *Atrax robustus* and *Hadronyche* spp.	▪ Moist, cool, sheltered habitats in eastern and southern Australia.	▪ Large spiders, up to 1½ in (4.5 cm) long, with glossy dark brown or black bodies and short legs.
	TARANTULAS *Theraphosidae*	▪ Deserts and forests in subtropical and tropical regions worldwide.	▪ Very large spiders with bodies up to 5 in (12 cm) long and a legspan of up to 11 in (28 cm). Bodies and legs are covered in bristly hairs. Pale brown to black in color, with pink, red, brown, or black markings.
	WIDOW SPIDERS *Latrodectus* spp.	▪ Warm temperate, subtropical, and tropical regions worldwide.	▪ Small spiders, with dark, glossy bodies and distinctive red, yellow, or white markings on the abdomen, which are shaped like an hourglass in some species.
	SCORPIONS *Scorpiones*	▪ Deserts, grasslands, woodland, and forests in warm temperate, subtropical, and tropical regions worldwide.	▪ Segmented bodies, with large, claw-like pedipalps and sting-bearing tails. Yellow, brown, or black bodies, 3–8 in (8–20 cm) long, depending on the species.
	TICKS *Ixodidae*	▪ Grassland and woodland habitats worldwide.	▪ Rounded bodies with no distinct divisions. Yellow to red- or black-brown in color. Up to ⅖ in (1 cm) long, but larger after feeding. They have spiked mouthparts, which they use to attach themselves to the skin of warm-blooded animals to feed on their blood.

INSECTS AND ARACHNIDS

breaths (see p. 277). If you need to give rescue breaths and there is poison in the casualty's mouth, or on his or her face, use a plastic face mask or face shield, if you have one. Before your trip, it is advisable to familiarize yourself with the harmful insects and arachnids (spiders, scorpions, and ticks) in the region you are visiting, and the nature of the threat.

DANGER	TREATMENT	HOW TO AVOID
▪ When a bee stings you, the stinger is ripped out of its body with the venom sac. Wasps and hornets can sting repeatedly. The stings are painful, followed by redness and swelling. A single sting may result in anaphylactic shock (see p. 274)	▪ Scrape the sting off sideways at the base of the sting if it is visible, using a fingernail or a credit card. Raise the affected part and place a cold compress against it for at least 10 minutes (see p. 267). Monitor for signs of allergic reaction.	▪ Watch out for flowers where bees may be feeding. Be careful when gathering fruit, or cleaning fish or game, as wasps or hornets will be attracted to the smell. Don't disturb their nests.
▪ Female mosquitoes feed on blood and their bites can irritate the skin. They are carriers of infectious diseases such as malaria, yellow fever, and dengue fever.	▪ Apply an antihistamine cream to the bites if they itch, and do not scratch them, as there is a risk of infection. If casualty develops a headache and fever, get medical help. Get casualty to rest, drink plenty of water, and take anti-malarials, if directed.	▪ Cover your arms and legs after sunset, use mosquito repellent, and sleep under a mosquito net. Take anti-malarial medication as directed. Get a yellow fever vaccination before traveling, where advised.
▪ Bite can cause fever, chills, vomiting, and, pain in the joints. In most cases the bite is minor, but in some cases the tissue around the wound will die, leaving a deep scar. Rarely fatal.	▪ Sit casualty down, treat as for snakebite (see p. 266), and call the emergency services. Apply a cold compress to the site of the sting. Monitor for signs of allergic reaction, such as wheezing, and anaphylactic shock (see p. 274).	▪ This spider likes to hide in dark places, so be careful where you put your hands and bare feet, and always check your bedding.
▪ Bite is very painful and cause profuse sweating, nausea, vomiting, tingling around the mouth and tongue, salivation, and weakness. Rarely fatal.	▪ Sit casualty down, treat as for snakebite (see p. 266), and call the emergency services—anti-venin is available. Monitor for signs of allergic reaction and anaphylactic shock (see p. 274).	▪ Take care when moving rocks and logs.
▪ Bite can be painful but is only mildly venomous. If provoked, tarantulas will flick irritating hairs from their abdomen which will cause itching and swelling of the skin and nasal passages, and watering of the eyes.	▪ Give pain relief and apply a cold compress to the site of the bite. Monitor for signs of allergic reaction, such as wheezing, and anaphylactic shock (see p. 274). Wash any hairs away with cool water.	▪ Tarantulas are nocturnal and hunt prey in trees and on the ground. Move away if you disturb one; don't touch it.
▪ Bite is painful and results in localized redness and swelling. More severe reactions include profuse sweating, abdominal and chest pains, nausea, and vomiting. Rarely fatal.	▪ Apply a cold compress (ice if available) and give pain relief. If symptoms worsen, call the emergency services – anti-venin is available. Monitor for signs of allergic reaction, such as wheezing, and anaphylactic shock (see p. 274).	▪ Widow spiders, which include Australia's redback spider (*L. hasselti*), build their webs in dry, sheltered spots, such as in shrubs or among rocks or logs, so take care when moving among them.
▪ Scorpion stings can be very painful and may cause severe illness. A few species have neurotoxins in their venom, resulting in temporary paralysis for 1-2 days. Rarely fatal and only in young children, the elderly, or the ill or infirm.	▪ If swelling occurs, sit casualty down, treat as for snake bite (see p. 266). Apply an ice pack or cold compress to the site of the sting. Call the emergency services, monitor for anaphylactic shock (see p. 274) and allergic reaction. Try to identify the scorpion.	▪ Scorpions are mostly nocturnal and take shelter during the day. Check your boots and clothing before putting them on and check your bedding at night. Take care when gathering firewood or turning over rocks.
▪ Ticks are carriers of infectious diseases such as Lyme disease and Rocky Mountain spotted fever. Sufferers of Lyme disease may develop a red, ringed "bull's-eye" skin lesion at the bite site.	▪ Carefully remove an attached tick with tweezers or a specialized tool (see p. 267). Clean and monitor the wound for symptoms. Get medical help if casualty develops a severe headache, stiff neck, or a fever.	▪ Wear light-shaded, tightly woven clothing and apply tick repellent. Check your skin and clothing for ticks at regular intervals.

NATURAL DANGERS

IN THE CASE OF SUSPECTED PLANT POISONING, you must call the emergency services immediately. If a casualty becomes unconscious, open the airway and check their breathing (see p. 276). Be ready to begin CPR—chest compressions and rescue breaths (see p. 277). If you need to give rescue breaths, use a plastic face mask or face shield, if you have one, to

	NAME	WHERE	IDENTIFICATION	POISONOUS PART
CONTACT POISONING	**COWHAGE/COWAGE/ COWITCH** *Mucana pruriens*	▪ Scrub and light woodland in the the tropics and USA.	▪ Vine-like plant with green, oval leaflets growing in threes. Spikes of hairy, dull purplish flowers. Hairy brown seedpods.	▪ Flowers ▪ Pods
	POISON IVY *Toxicodendron radicans*	▪ Wooded areas in North America; widely introduced elsewhere.	▪ See p. 268 for illustration.	▪ All
	WHITE/GRAY MANGROVE *Avicennia marina*	▪ Mangrove swamps from tropical Africa to Indonesia and Australasia.	▪ Slender tree with pale bark and long aerial roots. Green, oblong, leathery leaves. Yellow flowers and fleshy, pale orange seed capsules.	▪ Sap
INGESTION POISONING	**BANEBERRY** *Actaea* spp.	▪ Temperate woodland in the northern hemisphere.	▪ Low shrub with, toothed green leaves. Small white flowers. White, black, or red berries grow in clusters.	▪ All, but the berries are the most poisonous.
	CASTOR OIL PLANT/ CASTOR BEAN *Ricinus communis*	▪ Throughout the tropics in scrub and wasteland; introduced to temperate regions.	▪ Tall shrub with glossy dark green or dark purple, star-like leaves. Spikes of yellow flowers. Spiky red pods.	▪ All, but the seeds contain particularly high levels of the poison ricin.
	DEATH CAMAS/ DEATH LILY *Zigadenus* spp.	▪ Grassy, wooded, and rocky areas in temperate regions of North America.	▪ Long grass-like leaves arise from a bulb. Flowers have six white petals with a green heart at base.	▪ All
	PANGI *Pangium edule*	▪ Tropical forests in Southeast Asia.	▪ Tall tree with green, heart-shaped leaves. Green flower spikes. Large brownish pear-shaped fruits grow in clusters.	▪ All, but especially the seeds, which contain prussic acid (cyanide).
	PHYSIC NUT *Jatropha curcas*	▪ Woodland throughout the tropics and southern USA.	▪ Large shrub with large, green, ivy-like leaves. Small greenish-yellow flowers. Apple-sized yellow fruit has large seeds.	▪ All, but particularly the seeds.
	POISON HEMLOCK *Conium maculatum*	▪ Wet or moist ground in temperate parts of Eurasia; introduced to North America and Australia.	▪ Tall herb with branching purple-spotted stem; toothed leaves; small clusters of white flowers grown in flat umbrellas.	▪ All, but the root has the highest level of toxins.
	STRYCHNINE TREE *Strychnos nux-vomica*	▪ Tropical and subtropical forests in India, Southeast Asia, and Australia.	▪ Evergreen tree with paired oval leaves. Small clusters of greenish flowers grow at end of branches. Large orange-red berries.	▪ All, but the seeds contain lethal levels of strychnine.

POISONOUS PLANTS

protect yourself from any poison on the casualty's mouth. When planning your trip, it is advisable to familiarize yourself with the poisonous plants in the region you are visiting. For further information regarding plant contact poisoning and ingestion poisoning, and their treatment in a survival situation, see pp. 268–69.

EFFECT	TREATMENT	ADDITIONAL INFO
▪ Skin irritation. Contact with the eyes can cause blindness.	▪ Call the emergency services. Rinse the skin with cold water for 20 minutes, the eye(s) for 10 minutes. Don't let the water collect under the casualty or you.	
▪ Reddening, itching, swelling, and blistering of skin within 4–24 hours of contact with the plant. Some people may suffer an allergic reaction called anaphylactic shock (see p. 274).	▪ Wash the affected area with soap and cold water. Apply anti-poison ivy cream. Monitor and treat for shock (see p. 274) if necessary. For natural remedies, see p. 268.	▪ Western poison oak (*T. diversilobum*), Atlantic poison oak (*T. pubescens*), and poison sumac (*T. vernix*) also cause contact poisoning (see p. 268).
▪ The sap blisters the skin and will cause temporary blindness if it gets in the eyes.	▪ Call the emergency services. Rinse the skin with cold water for 20 minutes, the eye(s) for 10 minutes. Monitor casualty and treat for shock (see p. 274).	▪ Blinding or milky mangrove (*Exoecaria agallocha*) also has poisonous sap.
▪ Can be fatal—ingestion of berries can lead to cardiac arrest and death. Other symptoms include dizziness, vomiting, and severe internal irritation.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	
▪ Can be fatal—ingestion of raw seeds causes severe diarrhea.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	▪ The seedpods explode when ripe, scattering the large, oval, glossy, brown-mottled seeds, which can be mistaken for beans.
▪ Deadly poisonous. Symptoms include profuse salivation, vomiting, diarrhea, confusion, slow, irregular heartbeat, low temperature, difficulty breathing, and unconsciousness, leading to death.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	▪ This plant looks similar to wild onions, which are edible, but the death camas does not have the onion smell.
▪ Deadly poisonous. Causes unconsciousness and severe breathing difficulties, which, if untreated, may lead to death.	▪ Call the emergency services (antidotes are available). Induce vomiting if casualty is still conscious. Dilute the poison (see Death Camas, above). Keep casualty warm and at rest.	▪ If the casualty is unconscious and stops breathing, give rescue breaths, using a face shield. Alternatively, use an oxygen bag and a mask, if possible.
▪ Severe diarrhea and vomiting.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	▪ The seeds look like betel nuts and taste sweet. Several species in the *Jatropha* genus are poisonous, for example, the aptly named bellyache bush (*J. gossypifolia*).
▪ Deadly poisonous. A small amount can result in death by respiratory failure from muscular paralysis. Early symptoms include nausea, vomiting, and rapid heartbeat.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison (see Physic Nut, above). If the casualty is unconscious and stops breathing, begin CPR, using a face shield.	▪ Water hemlock or spotted cowbane (*Cicuta* spp.), which looks similar, is also deadly poisonous. Don't mistake these plants for wild carrots or wild parsnips.
▪ Deadly poisonous. Muscle spasms occur 10–20 minutes after exposure, leading to seizures followed by death from severe breathing difficulties or exhaustion within 2–3 hours.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	▪ Warning: strychnine can also enter the system through contact with the eyes.

RESOURCES

SCOUTS AND SCOUTING

Boy Scouts of America

1325 West Walnut Hill Lane
P.O. Box 152079, Irving,
Texas 75015-2079

Phone: 972-580-2000

Fax: 972-580-2413

www.scouting.org

Boy Scouts of America prepares young people to make ethical and moral choices over their lifetimes by instilling in them the values of the Scout Oath and Law.

Girl Scouts of the USA

420 Fifth Avenue
New York, New York 10018-2798

Phone: 800-GSUSA-4-U [800-478-7248] or 212-852-8000

www.girlscouts.org

Girl Scouts is one of the largest organizations dedicated specifically to girls. It encourages them to develop skills and values such as leadership, social conscience, and self-worth that will stand them in good stead in the wider world.

National Scouting Museum

Boy Scouts of America, S505,
1329 West Walnut Hill Lane, Irving,
Texas 75038

Phone: 800-303-3047 and 972-580-2100

e-mail: nsmuseum@netbsa.org

www.bsamuseum.org

This museum is a tribute to the rich history of the Boy Scouts of America and features a Norman Rockwell art gallery, virtual reality adventures, hands-on learning experiences, and a collection that traces the Scouting movement from its beginnings.

WEATHER AND ENVIRONMENTAL HAZARDS

National Oceanic and Atmospheric Administration (NOAA)

1401 Constitution Avenue, NW,
Room 6217, Washington, DC 20230

Phone: 202-482-6090

Fax: 202-482-3154

www.noaawatch.gov

The NOAA has several departments, such as the National Severe Storms Laboratory (NSSL), the Storm Prediction Center (SPC), which is part of the National Weather Service (NWS), and the National Centers for Environmental Prediction (NCEP). The SPC provides regular, timely, and accurate storm forecasts, and watches out for severe thunderstorms, tornadoes, heavy rain, heavy snow, and fire weather events over the contiguous United States.

National Weather Service

1325 East West Highway, Silver Spring,
Maryland 20910

www.weather.gov/om/severeweather/index

Provides weather, hydrologic, and climate forecasts and observations for the whole country and the waters and oceans around it. The service offers a national information database for the use of public, private, and global communities.

NOAA Weather Radio All Hazards (NWR)

www.weather.gov/nwr

A nationwide network of radio stations broadcasts continuous weather information directly from the nearest National Weather Service office. NWR broadcasts official Weather Service warnings, watches, forecasts, and other hazard information 24 hours a day, seven days a week.

National Interagency Coordination Center (NICC)

www.nifc.gov/nicc/index.htm

The NICC serves as a focal point for coordinating the national mobilization of resources, such as heavy airtankers, helicopters, smokejumpers, area command teams, and Remote Automated Weather Stations (RAWS), for wildland fire and other incidents throughout the United States.

The Forest Fire Lookout Association (FFLA)

2590 W. Versailles Drive, Coeur d'Alene,
Idaho 83815

Phone: 208-765-1714 and 800-GRN-TREE

Email: weblookout@imbris.net

www.firelookout.org

The FFLA investigates former forest-fire lookout sites, ground cabins, and early forest fire detection methods to encourage public groups and others in restoring forest-fire lookouts.

U.S. Geological Service

Western Distribution Branch,
P.O. Box 25286, Denver, Colorado 80225

Phone: 303-236-7477

www.usgs.gov

This multi-disciplinary science organization focuses on biology, geography, geology, geospatial information, and water, and is dedicated to the timely, relevant, and impartial study of the US landscape, and natural resources and hazards.

DesertUSA

www.desertusa.com

This website is a comprehensive resource for North American deserts and Southwest destinations. It includes information about desert biomes, and the ways plants and animals adapt to them, and about national and state parks.

HEALTH AND MEDICAL

Centers for Disease Control and Prevention

1600 Clifton Rd, Atlanta, Georgia 30333

Phone: 800-232-4636 and 888-232-6348

Email: cdcinfo@cdc.gov

www.cdc.gov

Provides the expertise, information, advice, and tools that individuals and communities need to protect their health, including the promotion of healthy living and the prevention of disease, injury, and disability.

U.S. Department of Health and Human Services

200 Independence Avenue, S.W.,
Washington, D.C. 20201

Phone: 202-619-0257

Toll Free: 1-877-696-6775

www.hhs.gov/disasters/index.shtml

For information and advice on coping with disasters and related medical emergencies.

American Medical Association

515 N. State Street, Chicago, Illinois 60610

Phone: 800-621-8335

www.ama-assn.org/ama/pub/category/6206.html

The AMA's Center for Public Health Preparedness and Disaster Response is an important national educational resource for enhancing the disaster preparedness and response capabilities of both civilian and military providers.

American Red Cross

National Headquarters, 2025 E Street NW
Washington, DC 20006
Phone: 800-733-2767
www.redcross.org/services

The nation's premier response organization that provides relief for communities during natural and manmade emergencies. It belongs to a worldwide movement that offers neutral humanitarian care and assistance to victims of war and devastating natural disasters such as earthquakes.

National Capital Poison Center

3201 New Mexico Ave, Suite 310,
Washington D.C. 20016
Emergency Phone: 800 222 1222
www.poison.org

Affiliated with The George Washington University Medical Center, the Poison Center is committed to prevent poisonings, save lives, and limit injury from poisoning.

SEARCH AND RESCUE

National Association for Search & Rescue (NASAR)

PO Box 232020, Centerville,
Virginia 20120-2020
Phone: 703-222-6277
Toll Free: 877-893-0702
Email: info@nasar.org
www.nasar.org/nasar

A not-for-profit association dedicated to the advancement of professional, literary, and scientific knowledge in fields related to search and rescue throughout the United States and around the world.

National Institute for Urban Search & Rescue

PO Box 91648, Santa Barbara,
California 93190
Phone: 800-767-0093 and 805-569-5066
Cell: 805-798-0169
Email: niusr@cox.net
www.niusr.org

A non-profit organization dedicated to finding ways of saving lives by improving disaster readiness and response through public awareness, collaboration, research, and engineering.

Mountain Rescue Association

www.mra.org

Represents highly skilled and active mountain rescue teams throughout the country, who work for local government authorities and who are dedicated to saving lives through rescue and mountain safety education.

International Association of Dive Rescue Specialists

201 North Link Lane, Fort Collins,
Colorado 80524-2712
Phone (toll free): 800-IADRS-911
Phone (international): 1-970-482-1562
Fax (international): 970-482-0893
www.iadrs.org

An association of public safety divers and water rescue personnel who, over the years, have educated and assisted professionals in all 50 states and 15 foreign countries.

Association of Air Medical Services

526 King Street, Suite 415,
Alexandria, Virgina 22314
Phone: 703-836-8732
Fax: 703-836-8920
www.aams.org

An international not-for-profit organization that represents providers of air and surface medical transport systems, and aims to ensure that everyone has access to quality air medical and critical care support.

FEDERAL

Environmental Protection Agency (EPA)

EPA Headquarters, Ariel Rios Building
1200 Pennsylvania Avenue, N.W.,
Washington, DC 20460
EPA Hotlines: Various phone numbers and websites for assistance.
www.epa.gov/ebtpages/emeremergencypreparedness.html

The EPA develops, implements, and coordinates preparations for chemical and other emergencies in order to be able to respond quickly and effectively to environmental crises and to keep the public informed about hazards in their community.

Federal Emergency Management Agency (FEMA)

500 C Street S.W., Washington, D.C. 20472
Phone: 800-621-FEMA-3362
TDD: TTY users can dial 1-800-462-7585 to use the Federal Relay Service.
www.fema.gov

FEMA aims to reduce the loss of life and property, and protect the country from all hazards, including natural disasters, acts of terrorism, and other manmade disasters.

U.S. Dept of the Interior National Parks Pass

1849 C Street, NW, Washington, DC 20240
Phone: 202-208-3100
Email: webteam@ios.doi.gov
www.doi.gov/parkpass.html

You can buy an annual pass for entry into national parks from the US Geological Survey store and through the government's federal lands recreation web portal at www.recreation.gov.

CANADA

National Search and Rescue Secretariat (NSS)

400-275 Slater Street, Ottawa,
Ontario K1A 0K2
Phone: (toll-free) 1-800-727-9414
Email: inquiry@nss.gc.ca
www.nss.gc.ca

An independent government agency that coordinates the search-and-rescue requirements of the military, coastguard, police, transport, parks, and meteorological services.

Search and Rescue Volunteer Association of Canada

8 Paradise Road, Paradise,
Newfoundland and Labrador A1L 3B4
Phone (toll free): 1-866-9SARVAC
(1-866-972-7822)
Phone (office): 709-368-5533
Fax: 709-368-1298
Email: info@sarvac.ca
www.sarvac.ca

A not-for-profit organization that provides a national search-and-rescue service, and is committed to fostering the exchange of information between other rescue services.

Meteorological Service of Canada

www.msc-smc.ec.gc.ca

Provides information and conducts research on water, climate, weather, atmospheric science, and other environmental issues.

Environment Canada

Inquiry Center, 351 St. Joseph Boulevard
Place Vincent Massey, 8th Floor
Gatineau, Quebec K1A 0H3
Phone: 1-800-668-6767
or 819-997-2800
Fax: 819-994-1412
TTY: 819-994-0736
Email: enviroinfo@ec.gc.ca
www.ec.gc.ca

Government department committed to preserving and enhancing the country's environment, and to providing weather forecasts.

GLOSSARY

A-frame A type of shelter design in which the roof is suspended by a single support running from the front to the back, resting on an "A"-shaped framework at either end.

Air mass A large body of air in a weather system that has its own temperature and humidity.

Altitude sickness An illness brought on by low air pressure at high altitudes.

Anabatic winds Daytime, upslope winds.

Anaphylactic shock A dangerous allergic reaction brought on by a bite or sting.

Attack point See *Stand-off.*

Avalanche transceiver A personal beacon that can be activated to alert rescuers by someone caught in an avalanche.

Base layer The bottom layer of clothing that sits directly against the skin and is designed to "wick" sweat away from the body.

Base rate The heart rate, as measured when the body is at rest.

Beaufort scale A way of describing wind speed on land and at sea, using a scale from 0 (calm) to 12 (hurricane).

Belaying A rope technique used by climbing partners to safeguard each other.

Bivi pup tent A basic one-man tent that uses a bivi bag as a covering.

Bivi bag A large waterproof sack suitable for use as a basic shelter; smaller and lighter than a tent.

Blazes A general term for markers on a trail.

Bola A hunting device made of a rope with weights attached, designed to be spun and thrown.

Boot skiing A technique for descending snow-covered slopes without skis.

Bothy bag A large waterproof bag that provides shelter from the wind, rain or sun.

Bow-and-drill set One of the oldest methods of creating a glowing ember (coal) for use in lighting a fire.

Boxing an object See *Detouring.*

Breathable clothing Clothing that allows sweat in the form of water vapour to be carried away from the body but is waterproof from the outside so keeps liquids (rain or snow) out.

Bungee A type of elasticated cord, often with hooks at either end.

Bushcraft knife A versatile survival knife used primarily for carving, cutting, and splitting small logs.

Button compass A small, simple compass, ideal as a back-up compass.

Button tie A method of attaching cordage to fabric that has no hoops or grommets; useful in that it avoids ripping a hole in the fabric, maintaining any waterproof qualities of the fabric, and allows both fabric and cordage to be reused subsequently for other purposes.

Cairn Trail marker consisting of a pile of rocks of varying sizes, designed to be visible in fog.

Canoe An open-deck paddle boat, for one or more people, and equipment.

Carabina A metal clip used for joining ropes or attachment.

Carbon monoxide A colourless, odourless, tasteless, but highly toxic gas.

Cardiopulmonary resuscitation (CPR) An emergency system of chest compressions and rescue breaths designed to restart a casualty's heart and lungs.

Cerebral oedema Accumulation of excess water in the brain that causes reduced brain function and potentially fatal swelling.

Char cloth Cotton cloth that has been combusted in the absence of oxygen. Used for lighting fires.

Chimneying A climbing technique used for climbing up the inside of large rock clefts.

Cirrus A wispy cloud that forms high in the sky and is made from ice crystals.

Cold shock An involuntary gasp reflex followed by hyperventilation, caused by sudden immersion in very cold water. In some circumstances cold shock response can lead to the inhalation of water, disorientation, panic, hypothermia, cardiac problems, and death.

Commando saw A small saw consisting of a serrated wire blade with a ring at each end.

Compacted snow trench A shelter dug into compacted snow; the excavated snow can be cut into blocks and used to form a roof.

Compass baseplate A plate to which some compasses are fixed, containing additional markings used for orientation and navigation.

Compass scale Scaled markings on a compass baseplate that measure distances on a map and help in working out grid references.

Compass An instrument used for orientation and navigation, using a freely rotating needle that indicates the direction of magnetic north.

Coniferous tree A needle-leaved, cone-bearing, mostly evergreen tree, such as pine, spruce, or fir.

Continental air mass Air that has tracked over land and carries comparatively less moisture than maritime air mass.

Contour lines Lines on a map that show points of equal height above sea level, thus detailing the changing height of natural features.

Contour navigation Walking at a constant height around a high natural obstacle.

Contouring See *Contour navigation.*

Contusion A bruise.

Cordage A type of light rope – an essential piece of survival equipment.

COSPAS–SARSAT system A satellite system, made up of low-altitude earth-orbiting (LEOSAR) and geostationary earth-orbiting (GEOSAR) satellites, that picks up signals from distress beacons.

Cramp Painful muscle spasm, often a result of dehydration.

Crampons Spiked metal plates that attach to boots to provide grip on icy surfaces.

Cumulonimbus A type of cloud that starts low in the sky and builds upwards, producing short, heavy downpours or thunderstorms.

Cumulus A billowing, puffy cloud that is generally small and develops on bright, sunny days, indicating fine weather.

Damper bread Yeast-free bread, suitable for making on a campfire.

Deadfall A trap in which a weight falls onto the prey. Also a mass of fallen dead timber.

Defensive swimming A type of swimming technique designed to keep the swimmer protected from obstacles in the water.

Degree A unit of latitude or longitude, equal to $^{1}/_{360}$ of a circle.

Dehydration A low level of water in the body. A very dangerous condition if not reversed.

Deliberate off-set (aiming off) Deliberately aiming left or right of a known feature.

Desalting kit An emergency desalination kit for turning saltwater into freshwater.

Detouring Walking around an obstacle as a method of clearing it, using a point in the distance as a reference.

Dew point The temperature at which water vapour held in the air cools to become liquid.

Dew trap A method of collecting dew to form usable water.

Digging stick A sturdy piece of wood with a pointed end.

Dipping net A fishing net used for scooping up fish that are too small to catch on a line.

Drill and flywheel A friction-based method of starting a fire.

Drip rag A method of collecting rainwater dripping down a tree using cloth.

Drogue A sea anchor used to stabilize a vessel.

Dysentery A water-borne disease that causes severe diarrhoea.

Eastings Vertical grid lines on a map; they increase in value the further east they are.

Emergency Locator Transmitter (ELT) A beacon that can send a distress signal to an orbiting satellite to initiate a rescue; mostly for use in aircraft.

Emergency plan of action (EPA) A document left with relevant authorities that contains vital details about you and your intended route, should you need to be rescued.

Emergency Position-Indicating Radio Beacon (EPIRB) A beacon that can send a distress signal to an orbiting satellite to initiate a rescue; mostly used at sea.

Emergency rendezvous (ERV) A pre-determined point at which all members of a group should meet in the event of separation.

Equator An imaginary circle running around the Earth's diameter that is equidistant from the North and South Poles at all points.

Eskimo roll A technique for righting a capsized kayak.

Faggot A bundle of kindling.

Feather stick A stick that has been feathered to use as kindling and fuel.

Fighter trench A type of snow trench dug into soft snow. Ideally uses a tarpaulin for the roof.

Finnish marshmallow A method of melting chunks of ice over a fire then collecting the water in a container.

Fire tin A tin filled with waxed cardboard, used for lighting a campfire in difficult conditions, or as a basic cooking device.

Firebase A stable, non-flammable base suitable for building a campfire on.

Firefighter's lift A method of moving a casualty by carrying them over your shoulder.

Fire set A set of components necessary to generate a glowing ember that can start a fire.

Firesteel A metal bar which, when hit with a striker, produces a spark for lighting campfires.

Flash burn See *Snow blindness*.

Föhn effect A dry downslope wind which occurs on the lee-side of a mountain or hill.

Four-wheel drive (4WD) A system on a vehicle in which drive is provided by all four wheels, rather than just two; results in superior traction.

Frostbite A serious condition caused by excessive exposure to extreme cold, in which body tissues actually freeze, sometimes to the bone. Often starts with the extremities.

Frostnip The freezing of the top layer of skin, usually on the face and extremities, caused by exposure to extreme cold. Can develop into frostbite if left untreated.

Gaiters Protective fabric that wraps around the lower leg and ankle, to keep out water and protect against sharp objects.

Gill net A fishing net in the form of a mesh stretched across a stream.

Global positioning system (GPS) A hand-held unit that uses orbiting satellites to determine the user's position accurately.

Gourd The shell of hollowed-out, dried fruit often used to carry water.

Grab bag A pre-prepared bag of essential survival equipment suitable for quickly picking up in an emergency at sea.

Gravity filter An effective method of filtering and purifying water, incorporated into a bottle.

Grid bearing A horizontal direction expressed in degrees east or west of north or south.

Grid magnetic angle (GMA) The difference between magnetic north and grid north, expressed as an angle.

Grid north A northerly direction that runs parallel to vertical gridlines on a map. Different from true north because a map is flat.

Grid reference A method of pinpointing the location of a place or object anywhere on a map, using coordinates provided by a numbered grid system printed on the map.

Grommet A metal-ringed eyelet in clothing.

Groundsheet A waterproof sheet.

Gypsy well A method of using a hole in the ground to filter non-potable water.

Hand jam A crack in a rockface suitable for wedging your whole hand into when climbing.

Handrailing A method of navigating using long linear features on the landscape that run in the general direction of your travel, such as rivers, roads, or paths.

Hank A small coil of cordage.

Heat exhaustion (heat stress) Dehydration that results in dizziness and profuse sweating; often a precursor of heatstroke.

Heatstroke A life-threatening condition brought on by severe overheating.

HELP (Heat Escape Lessening Posture) A posture that reduces heat-loss from the body when floating in water.

Hexamine stove A stove that uses hexamine solid fuel instead of gas or liquid fuel.

Hot platform A method of melting ice over a fire and collecting the liquid in a container.

Hyperthermia An abnormally high body temperature; potentially fatal.

Hyponatraemia An accumulation of excess water in the body, leading to a dangerously low concentration of sodium.

Hypothermia A life-threatening condition caused by exposure to cold. Symptoms usually appear in the following order: uncontrollable shivering (which may stop as the body temperature lowers); irrational or out-of-character behaviour; confusion, mood swings, and withdrawal; pale, cold skin; and slurred speech and/or stumbling. Watching out for and counteracting the early symptoms may help prevent them developing into a full-blown potentially fatal condition.

Hypoxia Inadequate levels of oxygen in the blood or body tissues. Potentially fatal.

Igloo A shelter constructed from cut blocks of snow.

Iodine A chemical used to purify water.

Isobar contours Lines on a weather chart that connect points at which the barometric pressure is the same.

Katabatic winds Cool, light downslope winds that form on clear nights.

Kayak A closed-deck paddle boat for one or more people.

Kindling Small pieces of fuel for a fire, added to burning tinder.

Kukri A large-bladed knife from Nepal, used mainly for chopping.

Land breeze A night-time wind that blows from land to sea.

Lanyard A small cord or rope used for securing or suspending an item.

Latitude The angular distance north or south of the Equator, in degrees along a meridian.

Layering Wearing several light layers of clothes rather than just one or two thick ones. Layers can be added or taken off to maintain a consistent body temperature.

Lean-to A shelter with a sloping roof that leans against a horizontal ridgepole.

Lee-side The side that is sheltered from the wind.

Lensatic compass A type of compass used for precise navigational work.

Longitude The angular distance east or west across the Earth between one meridian and the prime meridian at Greenwich, England.

Machete A large-bladed knife used mainly for chopping, approximately 45–50cm (18–20in) long.

Magnetic north The direction indicated by a magnetic compass.

Magnetic variation The difference between magnetic north and grid north.

Mantelling A technique used for climbing overhanging rock.

Maritime air mass Air that has tracked over the sea, typically containing moisture.

Mayday The internationally recognized distress signal for a grave emergency.

Matchless fire set Small kit that includes everything you need to start a fire – tinder, fuel, and spark device.

Melting sack An improvised sack that can hold snow or ice and is positioned close to a fire to provide water.

Meridian An imaginary arc on the Earth's surface from the North to the South Pole, connecting all locations running along it.

Mid-layer An insulating layer of clothing that sits between the base layer and the outer layer.

Motor impairment A limitation or loss of muscle control or movement.

Naismith's Rule A method of calculating the time it will take to arrive at a destination, taking into account distance and topography.

Natural hollow A natural dip in the ground that can be used as an emergency shelter.

New World Non-Eurasian and non-African regions of the world – specifically the Americas and Australia.

Noggin A piece of wood used to help support a structure, reducing the amount of cordage required.

Non-potable water Water that is not suitable for drinking.

Northings The horizontal grid lines on a map; increase in value the further north they are.

Occluded front A point at which two air masses meet in a weather system.

Old World Europe, Asia, and Africa.

Outer layer Outer clothing; ideally it should keep out rain but let sweat (as water vapour) escape.

Pace counting A method of determining distance; involves knowing how many paces you take to cover a set distance.

Pack animal An animal such as a mule or horse used for carrying heavy loads.

Paracord A useful type of cordage originally developed as rigging lines for parachutes. Usually comprises an outer sheath over an inner section of yarn.

Parang A large-bladed knife from Malaysia, used mainly for chopping.

Pathogen A micro-organism, especially bacteria or fungi, that causes disease.

Permafrost Permanently frozen soil.

Personal Locator Beacon (PLB) A beacon that can transmit a distress signal to an orbiting satellite to initiate a rescue.

Poncho A multi-purpose waterproof outer garment made from one sheet of fabric; also useful for building a basic shelter or bed.

Post-holing A technique for walking in deep snow without snow shoes.

Potable water Water suitable for drinking.

Potassium permanganate A chemical used for a variety of tasks, from lighting fires to purifying water.

Precipitation Moisture in a cloud that may fall as rain, snow, or hail.

Prismatic compass See *Lensatic compass*.

Protractor A device for measuring angles.

Psychogenic shock A very high level of psychological and emotional stress brought on by a sudden disaster situation.

Pulk A simple plastic sled used for carrying equipment over snow.

Pulmonary oedema Swelling and fluid accumulation in the lungs that causes potentially fatal breathing problems.

Pygmy hut A domed hut made from a circle of bent saplings or limber poles and thatched with natural materials.

Pygmy roll A method of rolling fibres together to make cordage.

Quicksand A bed of loose, wet sand that yields easily to pressure and can engulf anything on its surface.

Quinzhee A basic dome-shaped, snow-covered overnight shelter.

Ranger flint and steel A rod of flint which, when hit with a striker, produces a spark for lighting campfires.

Recovery position The position an unconscious person should be placed in to minimize further injury and help recovery.

Reflector A device for directing heat from a campfire towards shelter.

Reverse-osmosis pump A survival pump system that turns saltwater into freshwater.

Ridgepole A long horizontal pole that forms the apex of a roof.

Romer measure See *Compass scale*.

Salting A salt marsh.

Scrambling Climbing without ropes.

Scrape A type of basic shelter built in a natural hollow or depression in the ground.

Sea breeze A daytime wind that blows from sea to land.

Secondary drowning Potentially fatal biological changes that can take place in the lungs after a person nearly drowns, caused by the body's response to inhaling water.

Shock A life-threatening condition that occurs if the circulatory system fails, often triggered by severe bleeding, burns, or sudden cold.

Show-stopper A failure in planning that could cause delays to, or stop a trip or expedition.

Sighting mirror An advanced component on a high-end compass used for helping in orientation and navigation.

Signal fire A fire specifically designed to produce lots of smoke to attract attention.

Signal flare A hand-held distress flare that gives off orange smoke in daylight and a bright light at night.

Signal mirror A simple signalling device that works by reflecting sunlight.

Silva compass A type of basic compass useful for hiking.

Sip well A method of extracting water trapped under rocks by sandy ground.

Skidoo A motorized vehicle used for travelling over snow and ice.

Slingshot A hunting weapon that fires small stones and is made from a Y-shaped stick with an elastic strip stretched between the prongs.

Snare A wire noose used for trapping animals.

Snow blindness Eye damage caused by ultraviolet light reflected off snow or water.

Snow cave An effective cave-like shelter dug into compacted snow on the lee-side of a hill.

Snowshoes Wide-soled strap-on shoes designed to stop you sinking into snow.

Soak A water source found close to rivers and creeks in areas that are usually lower-lying than the existing water table.

Solar still A method of extracting drinking water from any source of moisture by a process of evaporation and condensation.

Space blanket A plastic blanket covered in aluminum foil that can be used as a shelter or a reflective signalling device; to carry, store, and heat water; or to cook in.

Stand-off A technique for navigating to a point that may be hard to locate, by orientating from a nearby prominent feature.

Standing deadwood Wood from a tree that has died but is still standing. Hazardous, as branches or the whole tree may fall, but makes excellent firewood.

Straddling A climbing technique used for ascending wide chimneys.

Stratus Dense, grey cloud that forms a sheet. Rain can fall for long periods of time if the cloud has depth.

Strobe light An effective signalling device in the form of a rapidly flashing LED light.

Survival blanket See *Space blanket*.

Survival kit Essential items for survival carried on your person.

Survival straw A compact emergency water purifier.

Survival suit A waterproof, buoyant suit designed for survival in open water.

Survival tin An essential, compact container carried on the person that addresses basic survival needs: protection, location, water, and food.

Taiga A subpolar region characterized by coniferous forest.

Tarp Shortened form of "tarpaulin", a type of sturdy waterproof sheet.

Temperate climate A climate without extremes of temperature or rainfall.

Tetanus A potentially lethal infection caused by bacteria that live in soil.

Throwing star An improvised four-pointed hunting weapon used for throwing at prey.

Tinder A dry, light, combustible material that is the first fuel used when lighting a fire.

Topographic map A map that shows the main features on a landscape, and elevation via use of contour lines.

Transpiration bag A method of evaporating, condensing, and siphoning off water from foliage using a plastic bag and sunlight.

Triangulation A method of determining your location by taking bearings with a compass from recognizable landscape features.

True north The direction of a meridian of longitude that converges on the North Pole.

Tundra A polar region characterized by permafrost and stunted vegetation.

Universal Edibility Test (UET) A test that checks whether a plant is safe to eat via a methodical process of testing small amounts on the skin, in the mouth, and in the stomach.

UV active purification Purifying water via a battery-operated ultraviolet (UV) purifier.

UV passive purification Purifying water by leaving filled bottles in strong direct sunlight. The UV rays eventually kill most pathogens.

Vegetation bag See *Transpiration bag*.

VHF radio Very high frequency radio transmitter; usually used at sea.

Water balance The difference between water lost from the body through sweat and the water taken in through drinking.

Water purifier A device that filters and purifies water by pumping it through micro filters, chemicals, or a combination of both.

Weather chart A map of major weather systems and their predicted direction of travel.

Wicking material A material that moves moisture away from your body to evaporate.

Wickiup A small hut made from straight poles lashed together at the top, with an interwoven framework covered with animal hides or grass.

Witch's broom A cleft stick stuffed with thin, dry bark suitable for lighting a fire.

Withies The strong, flexible stems of plants, such as willow, birch, ash, and hazel.

Zig-zag route Method of climbing a steep slope using a zig-zag path to reduce the effort required to achieve the climb. Also effective when walking down steep slopes.

INDEX

A

- A-frames
  - forest 164–65
  - front opening 165
  - jungle 168
  - poncho shelters 159, 167, 175
- abrasions (grazes) 263
- abandoning ship 250, 252–55
- ABC (airway/breathing/circulation), first aid 260, 276–77
- aerobic exercise 19
- agaves 193, 282–83
- air quality, caves 157
- aiming off 73
- air masses 79
- allergies and anaphylactic shock 266, 267, 274
- altitude sickness 33, 273
- altocumulus 80, 81
- altostratus 80, 81
- amoebic dysentery 187
- amphibians and reptiles
  - catching and eating 203, 218, 224, 294–95
  - dangers from 243, 266, 300–01
  - *see also* snakes
- anabatic winds 82
- analogue watches, navigating using 77
- anaphylactic shock 266, 267, 274
- anchor points, winching 103
- anchors, sea 39, 251
- animals
  - dangers from 115, 157, 189, 242–43, 300–3
  - edible 290–99
  - locating water 189
  - navigation aids 77
  - traveling with 42, 96–97, 98–99
  - weather forecasting 83
  - *see also* specific animals (e.g. bears); specific types (e.g. mammals)
- ants 219, 298–99
- arbor knot 142
- arm injuries 270
- ash
  - for soap 117
  - for stomach pain 269
- asthma 275
- automatic fires 124
- avalanches 32, 246–47

B

- bacillary dysentery 187
- backpacks 41, 42, 44–45, 108
- bacterial infections 185, 187, 212, 220, 262, 267, 301, 303
- baits and lures 208–09, 223, 229
- baking food 204, 205
- bamboo 284–85
  - fire saw 133
  - lean-tos 169
  - rafts 107
  - steamer 204
  - water sources and collection 192, 199
- bark
  - cordage 138
  - tea 269
  - tinder and kindling 120, 121, 238
- bats 157, 290–91, 300–01
- beaches 39, 177
- beacons, phones, and radios 39, 43, 61, 232–33, 236–37, 246, 250, 252
- bearings, compass 64–65, 68–71
- bears 28, 36, 157, 242
- Beaufort scale 83
- beds and bedding 41, 52–53, 156, 166–67
- bees 224, 267, 302–3
- belaying 91
- berms 177
- big cats 32, 242
- bilharzia 187
- birds
  - catching and eating 203, 204, 205, 218, 226–29, 292–93
  - locating water 189
- bites and stings 116, 266–67, 268, 300–03
- bivy sacks 53
  - bivy pup tent 160–61
  - hooped bivy 41, 54
- blankets, space 53, 61, 155, 175
- blazes (trail markers) 87
- bleeding
  - birds, for food 228
  - first aid 263–64
  - mammals, for food 222
- blended fuel stoves 58
- blisters 262, 268
- boats *see* ships and boats
- body fat 56
- boiling
  - food 204, 212
  - water 118, 185, 200
- bolas 226
- bones, joints, and muscles injuries 233, 270–1
- boot skiing 94
- boots 46, 49, 114
- bothy bags 161
- bottle filters 200
- bottle traps, fishing 211
- bough fence traps 227
- bow drills 130–33
- boxing an object (detouring) 73
- branches
  - as snowshoes 95
  - vehicle recovery 102
- bread, damper 204
- breathable waterproof fabric 48
- breathing difficulties
  - first aid 275
  - in shelters 157, 181
- breezes and winds 38, 79, 82–83
- brown bears 28, 242
- bruising 262, 263
- brushwood rafts 106–07
- buddy-buddy sleeping system 163
- buoyancy aids (floats) *see* lifejackets
- burns and scalds 265
- bushcraft knives 146–49
- butane stoves 58
- button compasses 68
- button ties 161

C

- cacti 193, 282–83
- cage traps, figure 4 226
- cairns 87
- calorie needs 56, 202
- camels 42, 99
- camping soap 117
- camping stoves 58–59
- camps
  - organization and administration 112–13, 114–15
  - *see also* specific areas (e.g. latrines)
- canoeing and kayaking 42, 104–05
- capacity, packs and bags 41, 42, 44
- capsizing 104, 105
- carbohydrates 56
- carbon dioxide 157
- carbon monoxide poisoning 181
- cardinal points 68, 76, 77
- cardiopulmonary resuscitation (CPR) 277
- cats, big 32, 242
- caves 157, 181
- cell phones 43, 237
- chainsaws 60, 150
- char cloths 122–23
- char test, softwoods and hardwoods 131
- charts
  - maps and map reading 43, 64–65, 66–71
  - weather 78
- chemical flames 127
- chemical lightsticks 61, 240
- chimneying 91
- chlorine tablets 201
- choking 275
- cholera 187
- cirrostratus 80, 81
- cirrus 80, 81
- clay baking, food 205
- climate 78
  - *see also* specific environments (e.g. temperate environment)
- climbing and scrambling 90–91
- clothing 46–49, 94
  - cleanliness 116
  - cultural differences 22
  - on fire 265
  - hats 43, 46, 47, 48
- clothing, improvising with
  - buoyancy aids 255
  - cordage 137
  - dipping nets 210
  - slings 261
  - stretchers 279
  - water collection and treatment 185, 190, 192, 195, 199, 200

clouds 78, 79, 80–81
coastal caves 157
coasts *see* marine environments
cold environments 36–37
  clothing 47, 94
  cold shock 254
  dos and don'ts 51, 245
  edible plants 286–87
  hypothermia and frostbite 21, 29, 33, 37, 252, 273
  tents 55
  water sources 194–95
  see also ice; snow
cold fronts, weather 78, 79
cold shock 254
cold-weather clothing 47, 94
commando saws 151
commando-style match striking 126
communications equipment 43, 61, 232–33, 236–37
compacted snow trench 178
compasses and bearings 43, 64–65, 68–71, 74–75
components of fire 118, 238
condoms 61
conjunctivitis 212
constellations, navigating by 77
continental air masses 79
contour lines 66, 67
contouring 73
contusions (bruising) 262, 263
cooking
  fires, rocks, and pits 119, 204–05
  fish preparation 212–13
  invertebrates 219, 224–25
  mammals preparation 220–23
  stoves 58–59
coping strategies 20–21, 113, 234–35, 259
cordage 136–41
cordless scrape shelters 174
cougars 32, 242
CPR (cardiopulmonary resuscitation) 277
cardiovascular fitness 18, 19
crabs 225
cramp 270
crickets and grasshoppers 224, 298–99
crocodilians 189, 242
crossing rivers 88–89, 101, 108–09
crustaceans 225
cryptosporidium 187
cumulonimbus 80
cumulus 80, 81
cuts and grazes 263
cutting tools 60, 146–51
cyalumes 61, 240

D

Dakota hole fires 125
damper bread 204
day packs 42, 44
deadfalls 115
deadwood, standing 115, 121, 131
defensive swimming 108
dehydration and rehydration 187, 190, 193, 196, 253, 269, 272
deliberate off-set 73
denatured alcohol stoves 59
desalting kits 197
desert environments 34–35
  clothing 47
  dos and don'ts 129
  edible plants 282–83
  four-wheel driving 35, 101
  night walking 86
  shelters 174–75
  water sources 34, 35, 193
detouring (boxing an object) 73
dew harvesting and traps 190, 196
dew point 80
diarrhea and vomiting 187, 203, 269
digging
  ponds 199
  roots and tubers 206
  underground water 193
  vehicle recovery 102
dipping nets 210
disaster situations and emergencies
  causes 16–17, 26, 84
  dos and don'ts 27, 51, 93, 129, 173, 215, 232–33, 245
  emergency plans of action (EPAs) 24–25, 250
  emergency rendezvous points (ERVs) 72
  environmental hazards 246–49
  four-wheel driving (4WDs) 35, 103
  gear priorities 43
  mental preparation and responses 20–21, 234–35, 259
  PLWF (protection, location, water, and food) 11, 27, 258
  rescue services 24
  rescue signaling 39, 43, 232–33, 236–41, 250, 252
  survival at sea 16–17, 39, 196–97, 214–15, 250–55
  survival kits 60–61
  SURVIVAL strategies 234–35
  water sources and rationing 39, 187, 188–97, 199, 251
  see also first aid
diseases *see* infections
distance measuring 67, 72
distilling water 185
distress signals 39, 232–33, 236–41, 250, 252
dogs 17, 96–97, 99
dome signal fires 238–39
dos and don'ts
  disaster situations and emergencies 27, 51, 93, 129, 173, 215, 232–33, 245
  fires 118, 119, 124
double overhand knot 143
double sheet bend 144
down sleeping bags 52–53
downhill walking 86, 87
drawing birds (food preparation) 229
driftwood shelters 176–77
drip rags 192
driving techniques, 4WDs 35, 100–03
drogues 39, 251
drowning 254, 255, 273
drum rafts 107
duck traps 227
dunes, sand 35, 177, 189
dysentery 187

E

E. coliform infection 187
earthquakes 247
earthworms 224, 298–99
eastings *67*
edible plants 203, 206–07, 280–89
eggs 226
electric current, magnetizing compass needles 75
electrical storms 81, 249
elements of fire *see* fuel; kindling; tinder
elephants 99, 189, 242
ember pans 132
emergencies *see* disaster situations and emergencies
emergency signaling codes 240, 241, 252
emergency transmitters 39, 232–33, 236–37, 246, 250, 252
emergency water packs 199
emotional responses
  disaster situations 20–21, 113, 234–35, 259
  team dynamics 23
environmental hazards 246–49
environments
  understanding and preparing for 23, 26
  see also specific environments (e.g. cold environments)
EPAs (emergency plans of action) 24–25, 250
EPIRBs (emergency position indicating radio beacons) 236, 250, 252
equipment *see* gear and equipment
ERVs (emergency rendezvous points) 72
eskimo rolls 104
Evenk knot 144
exercise and fitness 18–19
exposure, first aid 272–73
external-frame packs 44
extreme weather 246, 248–49
eyes
  care 116
  problems and injuries 212, 265

F

fallen trees, shelters 163
fatigue 21, 235
fats 56
fear 234, 235
feather lures 209
feather sticks 120–21
features, use in navigation 70–71, 72, 73
feelings *see* emotional responses
felling trees 148–51
feral animals 242
fibre (nutrient) 56
fighter trench 179
figs 171, 284–85
filleting fish 212–13
filtering water 184, 200, 201
Finnish marshmallow 195
fire saws, bamboo 133
fire tins 123
fire wads 123
firefighter's lift 278
fires
  benefits 27, 113
  burns, first aid 265
  in caves 157
  clothing on fire 265

AF REGULATION 64-4
VOLUME I

Search and Rescue
SURVIVAL TRAINING

SURVIVE • RETURN • HONOR

15 JULY 1985

DEPARTMENT OF THE AIR FORCE

DEPARTMENT OF THE AIR FORCE
Headquarters US Air Force
Washington DC 20330-5000

AF REGULATION 64-4
VOLUME I
15 July 1985

Search and Rescue

SURVIVAL TRAINING

This regulation describes the various environmental conditions affecting human survival, and describes individual activities necessary to enable that survival. This regulation is for instructor and student use in formal and USAF survival and survival continuation training. This regulation also applies to US Air Force Reserve and Air National Guard units and members. Sources used to compile this regulation are listed in the bibliography, attachment 1.

Paragraph Page

Part One — The Elements of Surviving
Chapter 1 — Mission

Introduction 1–1 22
Aircrew Mission 1–2 22
Goals 1–3 23
Survival 1–4 23
Decisions 1–5 23
Elements 1–6 23

Chapter 2 — Conditions Affecting Survival

Introduction 2–1 24
Environmental Conditions 2–2 24
The Survivors Condition 2–3 24
Duration - The Time Condition 2–4 27
Sociopolitical Condition 2–5 27
Induced Conditions 2–6 27

Chapter 3 – The Survivor's Needs

Introduction 3–1 30
Maintaining Life 3–2 30
Maintaining Honor 3–3 31
Returning 3–4 32

Part Two — Psychological Aspects of Survival
Chapter 4 — Contributing Factors

Introduction 4–1 36
Survival Stresses 4–2 36
Pain 4–3 37
Thirst and Dehydration 4–4 37
Cold and Heat 4–5 38
Hunger 4–6 39
Frustration 4–7 40
Fatigue 4–8 40

Supersedes AFM 64-3, 15 August 1969. (See signature page for summary of changes.)
No. of Printed Pages: 579
OPR: ATC/DONZ (MSgt R.J. Paetz)
Approved by: ATC/DON (Col W.W. Miller)
Editor: M.M. Rastellini
Distribution: F

This publication includes copyrighted material.

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402-5000

Paragraph Page
Sleep Deprivation 4–9 42
Isolation 4–10 42
Insecurity 4–11 42
Loss of Self–Esteem 4–12 43
Loss of Self–Determination 4–13 44
Depression 4–14 44

Chapter 5 — Emotional Reactions

Introduction 5–1 45
Fear 5–2 45
Anxiety 5–3 46
Panic 5–4 46
Hate 5–5 47
Resentment 5–6 47
Anger 5–7 47
Impatience 5–8 48
Dependence 5–9 48
Loneliness 5–10 49
Boredom 5–11 49
Hopelessness 5–12 49
Summary 5–13 50

Chapter 6 — The Will to Survive

Introduction 6–1 51
Overcoming Stress 6–2 51
Crisis Period 6–3 51
The Coping Period 6–4 52
Attitude 6–5 52
Optimism 6–6 52
Summary 6–7 52

Part Three — Basic Survival Medicine
Chapter 7 — Survival Medicine

Introduction 7–1 53
Procedures and Expedients 7–2 53
Hygiene 7–3 54
Care of the Mouth and Teeth 7–4 54
Care of the Feet 7–5 55
Clothing and Bedding 7–6 55
Rest 7–7 55
Rules for Avoiding Illness 7–8 56
General Management of Injuries 7–9 56
Illnesses 7–10 60

Chapter 8 - PW Medicine

Introduction 8–1 61
History 8–2 61
Trauma 8–3 62
Gastrointestinal Problems 8–4 62
Hepatitis 8–5 63
Nutritional Deficiencies 8–6 63
Communicable Diseases 8–7 64
Skin Diseases 8–8 64
Fungal Infection 8–9 64
Dental Problems 8–10 65
Burns 8–11 65
Lacerations and Infection 8–12 65
Fraction/Sprains 8–13 66

Paragraph Page

Summary 8–14 67
Conclusion 8–15 68

Part Four — Facts and Conditions Affecting a Survivor
Chapter 9 — Weather

Introduction 9–1 69
Knowledge of Weather 9–2 69
Elements Affecting Weather 9–3 69
Circulation of the Atmosphere 9–4 71
Storms 9–5 76
Weather Forecasting 9–6 77
Summary 9–7 78

Chapter 10 — Geographic Principles

Introduction 10–1 80
Components of the Environment 10–2 80
Effects of Climate on Terrain 10–3 83
Effects of Terrain on Climate 10–4 84
Effects of Climate and Terrain on Life Forms 10–5 85

Chapter 11 — Environmental Characteristics

Introduction 11–1 88
The Köppen-Geiger System 11–2 88
Tropical Climates 11–3 88
Dry Climates 11–4 100
Warm Temperature Climate 11–5 105
Snow Climates 11–6 116
Ice Climates 11–7 120
Open Seas 11–8 124
Ocean Currents 11–9 126
Climates Conditions 11–10 126
Life Forms 11–11 127

Chapter 12 — Local People

Introduction 12–1 138
Contact with People 12–2 138
Survivor's Behavior 12–3 138
Changing Political Allegiance 12–4 139

Part Five — Personal Protection
Chapter 13 — Proper Body Temperature

Introduction 13–1 140
Body Temperatures 13–2 141
Heat Transfer 13–3 141

Chapter 14 — Clothing

Introduction 14–1 143
Protection 14–2 143
Clothing Materials 14–3 143
Natural Materials 14–4 143
Synthetic Materials 14–5 143
Types of Insulation 14–6 143
Insulation Measurement 14–7 144
Clothing Wear in Snow and Ice Areas 14–8 147
Care of the Feet 14–9 149
Clothing in the Summer Arctic 14–10 149
Clothing at Sea 14–11 150
Antiexposure Garments 14–12 151
Warm Oceans 14–13 153

Paragraph Page
Tropical Climates 14–14 153
Dry Climates 14–15 153

Chapter 15 — Shelter

Introduction 15–1 155
Shelter Considerations 15–2 155
Location 15–3 156
Immediate Action Shelters 15–4 157
Improvised Shelters 15–5 157
Shelters for Warm Temperature Areas 15–6 157
Maintenance and Improvements 15–7 159
Construction of Specific Shelters 15–8 159
Shelter for Tropical Areas 15–9 161
Specific Shelter for Tropical Environments 15–10 165
Shelter for Dry Climates 15–11 165
Principles of Desert Shelters 15–12 168
Shelter for Snow and Ice Areas 15–13 168
General Construction Techniques 15–14 171
Shelter Living 15–15 171
Shelter Considerations for Arctic and Arctic–like Areas 15–16 175
Shelter for Open Seas 15–17 175

Chapter 16 — Firecraft

Introduction 16–1 177
Elements of Fire 16–2 177
Fire Location 16–3 178
Fire Site Preparation 16–4 178
Firemaking with Matches (or Lighter) 16–5 179
Heat Sources 16–6 179
Firemaking with Special Equipment 16–7 183
Burning Aircraft Fuel 16–8 184
Useful Firecraft Hints 16–9 184
Fire Lays 16–10 185

Chapter 17 — Equipment

Introduction 17–1 188
Types of Kits 17–2 188
Issued Equipment 17–3 188
Improvised Equipment 17–4 192
Miscellaneous Improvised Equipment 17–5 195
Ropes and Knots 17–6 200
Personal Survival Kits 17–7 217

Part Six — Sustenance

Chapter 18 — Food

Introduction 18–1 219
Nutrition 18–2 219
Food 18–3 220
Animal Food 18–4 221
Plant Food 18–5 234
Food in Tropical Climates 18–6 240
Food in Dry Climates 18–7 244
Food in Snow and Ice Climates 18–8 246
Food in Open Seas 18–9 251
Preparing Animal Food 18–10 255
Cooking 18–11 264
Preparing Food in Enemy Areas 18–12 267
Preserving Food 18–13 269

	Paragraph	Page
Preparing Plant Food	18–14	272

Chapter 19 — Water

	Paragraph	Page
Introduction	19–1	275
Water Requirements	19–2	275
Water Sources	19–3	277
Water in Snow and Ice Areas	19–4	279
Water on the Open Seas	19–5	280
Water in Tropical Areas	19–6	281
Water in Dry Areas	19–7	283
Preparation of Water for Consumption	19–8	287

Part Seven — Travel

Chapter 20 — Land Navigation

	Paragraph	Page
Introduction	20–1	290
Maps	20–2	290
Aeronautical Charts	20–3	291
Information Contained in Margin	20–4	298
Topographical Map Symbols and Colors	20–5	307
Coordinate Systems	20–6	307
Elevation and Relief	20–7	318
Representative Fraction (RF)	20–8	320
Graphic (Bar) Scales	20–9	321
Using a Map and Compass, and Expressing Direction	20–10	323

Chapter 21 – Land Travel

	Paragraph	Page
Introduction	21–1	352
Decision to Stay or Travel	21–2	352
Travel	21–3	353
Land Travel Techniques	21–4	353
Mountain Walking Techniques	21–5	362
Burden Carrying	21–6	363

Chapter 22 — Rough Land Travel and Evacuation Techniques

	Paragraph	Page
Introduction	22–1	369
Safety Coordination	22–2	369
Climbing Ropes	22–3	369
Specialized Knots for Climbing and Evacuation	22–4	370
Special Equipment	22–5	373
The Successful Climb	22–6	379
Belaying	22–7	383
Communication	22–8	385
Anchors	22–9	386
Climbing	22–10	388
Rappelling	22–11	392
Overland Snow Travel	22–2	395
Snow and Ice Climbing Procedures and Techniques	22–13	398
Glaciers and Glacial Travel	22–14	405
Evacuation Principles and Techniques	22–15	410

Chapter 23 — Water Travel

	Paragraph	Page
Introduction	23–1	427
River Travel	23–2	427
Using Safe Judgment and Rules for River Travel	23–3	427
River Hydraulics	23–4	433
Emergency Situations	23–5	442
Improvised Rafts	23–6	445
Fording Streams	23–7	445

Paragraph Page
Traveling on Open Seas 23–8 446
Physical Considerations 23–9 450
Life Preserver Use 23–10 451
Raft Procedures 23–11 451
Making a Landfall 23–12 456
Methods of Getting Ashore 23–13 456

Part Eight — Signaling and Recovery
Chapter 24 — Signaling

Introduction 24–1 460
Manufactured Signals 24–2 461
Improvised Signals 24–3 465

Chapter 25 — Recovery Principles

Introduction 25–1 474
National Search and Rescue (SAR) Plan 25–2 474
Survivor Responsibilities 25–3 474
Recovery Site 25–4 475
Recovery Procedures 25–5 475
Pickup Devices 25–6 477
Preparation for Open Sea Recovery 25–7 480

Part Nine — Evasion
Chapter 26 — Legal and Moral Obligations

Introduction 26–1 482
Definitions 26–2 482
Military Drive 26–3 482

Chapter 27 — Factors of Successful Evasion

Basic Principles 27–1 485
Preparation 27–2 485
Opportunity 27–3 486
Motivation 27–4 487
Evasion Principles 27–5 487

Chapter 28 — Camouflage

Introduction 28–1 491
Types of Observation 28–2 491
Comparison of Direct and Indirect Observation 28–3 492
Preventing Recognition 28–4 492
Factors of Recognition 28–5 493
Principles and Methods of Camouflage 28–6 495
Individual Camouflage 28–7 499
Concealment in Various Geographic Areas 28–8 503
Concealment Factor for Areas other than Temperate 28–9 506
Camouflage Techniques for Shelter Areas 28–10 507
Firecraft Under Evasion Conditions 28–11 509
Sustenance for Evasion 28–12 512
Encampment Security Systems 28–13 513
Evasion Movement 28–14 518
Searching Terrain 28–15 519
Movement Techniques Which Limit the Potential for Detection of an Evader (Single) 28–16 520
Barriers to Evasion Movement 28–17 528
Evasion Aids 28–18 539
Combat Recovery 28–19 543

Paragraph Page

Part Ten — Induced Conditions — Nuclear, Biological, and Chemical (NBC)
Chapter 29 — Nuclear Conditions

Introduction 29–1 547
Effects of Nuclear Weapons 29–2 547
Types of Nuclear Bursts 29–3 549
Injuries 29–4 550
Types of Residual Radiation 29–5 552
Effects of Fallout on the Individual 29–6 552
Radiation Detection 29–7 553
Body Reactions to Radiation 29–8 554
Countermeasures Against Penetrating External Radiation 29–9 556
Shelter 29–10 558
Water 29–11 561
Food 29–12 562

Chapter 30 — Biological Conditions

Introduction 30–1 568
Detection 30–2 568
Climate 30–3 568
Terrain 30–4 569
Protection 30–5 569
Tips for a Survivor 30–6 569

Chapter 31 — Chemical Conditions

Introduction 31–1 570
Chemical Groups 31–2 570
Detection 31–3 572
Protection 31–4 573
Avoiding Chemical Agents 31–5 574
Decontamination of Chemical Agents 31–6 575
Tips for a Survivor 31–7 575

Attachment

1. Bibliography 577

Figures

1-1. Elements of Surviving 22
1-2. Survival Triangle 23
2-1. Five Basic Conditions 24
2-2. Cold Front Occlusion 25
2-3. Terrain 26
2-4. Life Forms 26
2-5. Survivor's Condition 27
2-6. Code of Conduct 28
2-7. Induced Conditions 29
3-1. Survivor's Needs 30
3-2. Personal Protection 31
3-3. Shelters 32
3-4. Fire 32
3-5. Water 33
3 6. Food 33
3-7. Self-Aid 33
3-8. Health and Morale 34
3-9. Recovery 34
3-10. Travel 35
3-11. Burden Carrying 35
4-1. Psychological Aspects 36

Part One

THE ELEMENTS OF SURVIVING

Chapter 1

MISSION

1-1. Introduction. An ejection sequence, a bailout, or crash landing ends one mission for the crew but starts another—to successfully return from a survival situation. Are they prepared? Can they handle the new mission, not knowing what it entails? Unfortunately, many aircrew members are not fully aware of their new mission or are not fully prepared to carry it out. All instructors teaching aircrew survival must prepare the aircrew member to face and successfully complete this new mission. (Figure 1-1 shows situations a member might encounter.)

1-2. Aircrew Mission. The moment an aircrew member leaves the aircraft and encounters a survival situation, the assigned mission is to: "return to friendly control without giving aid or comfort to the enemy, to return early and in good physical and mental condition."

a. On first impressions, "friendly control" seems to relate to a combat situation. Even in peacetime, however, the environment may be quite hostile. Imagine parachuting into the arctic when it's minus 40°F. Would an aircrew member consider this "friendly?" No. If the aircraft is forced to crash-land in the desert where temperatures may soar above 120°F, would this be agreeable? Hardly. The possibilities for encountering hostile conditions affecting human survival are endless. Crewmembers who egress an aircraft may confront situations difficult to endure.

b. The second segment of the mission, "without giving aid or comfort to the enemy," is directly related to a combat environment. This part of the mission may be most effectively fulfilled by following the moral guide—the Code of Conduct. Remember, however, that the Code of Conduct is useful to a survivor at all times and in all situations. Moral obligations apply to the peacetime situation as well as to the wartime situation.

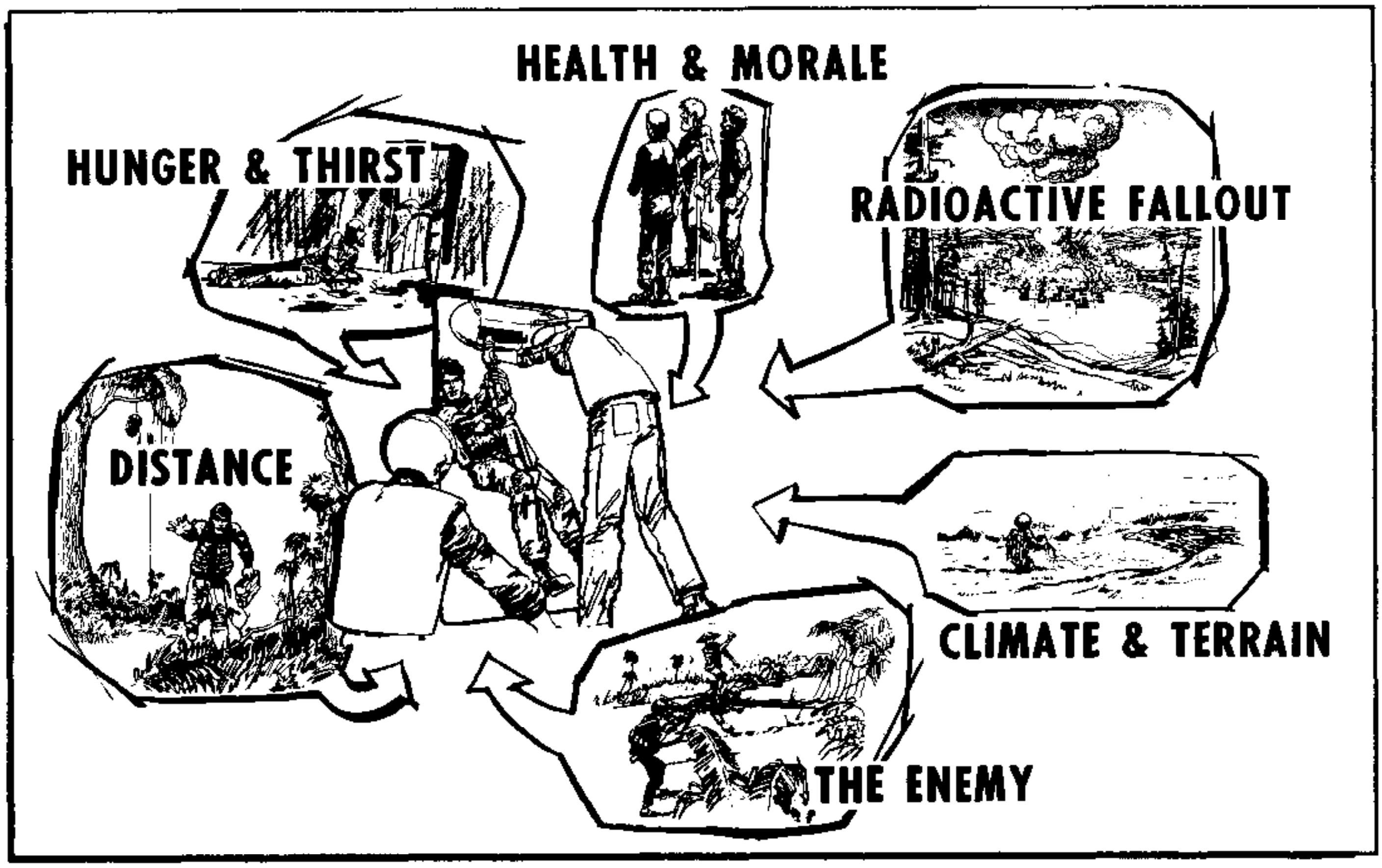

Figure 1-1. Elements of Surviving.

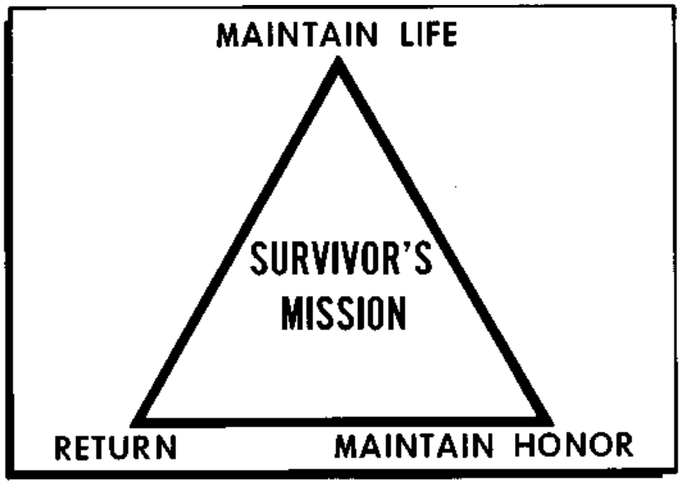

Figure 1-2. Survival Triangle.

c. The final phase of the mission is "to return early and in good physical and mental condition." A key factor in successful completion of this part of the mission may be the *will to survive.* This will is present, in varying degrees, in all human beings. Although successful survival is based on many factors, those who maintain this important attribute will increase their chance of success.

1-3. Goals. Categorizing this mission into organizational components, the three goals or duties of a survivor are to maintain life, maintain honor, and return. Survival training instructors and formal survival training courses provide training in the skills, knowledge, and attitudes necessary for an aircrew member to successfully perform the fundamental survival duties shown in figure 1-2.

1-4. Survival. Surviving is extremely stressful and difficult. The survivor may be constantly faced with hazardous and difficult situations. The stresses, hardships, and hazards (typical of a survival episode) are caused by the cumulative effects of existing conditions. (See chapter 2 pertaining to conditions affecting survival.) Maintaining life and honor and returning, regardless of the conditions, may make surviving difficult or unpleasant. The survivor's mission forms the basis for identifying and organizing the major needs of a survivor. (See survivor's needs in chapter 3.)

1-5. Decisions. The decisions survivors make and the actions taken in order to survive determine their prognosis for surviving.

1-6. Elements. The three primary elements of the survivor's mission are: the conditions affecting survival, the survivor's needs, and the means for surviving.

Chapter 2

CONDITIONS AFFECTING SURVIVAL

2-1. Introduction. Five basic conditions affect every survival situation (figure 2-1). These conditions may vary in importance or degree of influence from one situation to another and from individual to individual. At the onset, these conditions can be considered to be neutral—being neither for nor against the survivor, and should be looked upon as neither an advantage nor a disadvantage. The aircrew member may succumb to their effects—or use them to best advantage. These conditions exist in each survival episode, and they will have great bearing on the survivor's every need, decision, and action.

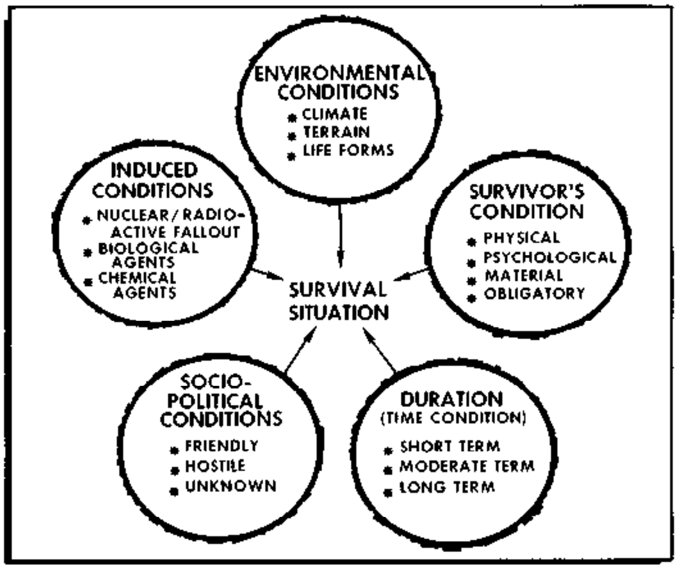

Figure 2-1. Five Basic Conditions.

2-2. Environmental Conditions. Climate, terrain, and life forms are the basic components of all environments. These components can present special problems for the survivor. Each component can be used to the survivor's advantage. Knowledge of these conditions may very well contribute to the success of the survival mission.

a. Climate. Temperature, moisture, and wind are the basic climatic elements. Extreme cold or hot temperatures, complicated by moisture (rain, humidity, dew, snow, etc.) or lack of moisture, and the possibility of wind, may have a life threatening impact on the survivor's needs, decisions, and actions. The primary concern, resulting from the effects of climate, is the need for personal protection. Climatic conditions also have a significant impact on other aspects of survival (for example, the availability of water and food, the need and ability to travel, recovery capabilities, physical and psychological problems, etc.) (figure 2-2).

b. Terrain. Mountains, prairies, hills, and lowlands, are only a few examples of the infinite variety of land forms which describe "terrain." Each of the land forms have a different effect on a survivor's needs, decisions, and actions. A survivor may find a combination of several terrain forms in a given situation. The existing terrain will affect the survivor's needs and activities in such areas as travel, recovery, sustenance, and, to a lesser extent, personal protection. Depending on its form, terrain may afford security and concealment for an evader; cause travel to be easy or difficult; provide protection from cold, heat, moisture, wind, or nuclear, biological, chemical (NBC) conditions; or make surviving a *seemingly* impossible task (figure 2-3).

c. Life Forms. For survival and survival training purposes, there are two basic life forms—plant life and animal life (other than human). NOTE: The special relationship and effects of people on the survival episode are covered separately. Geographic areas are often identified in terms of the abundance of life (or lack thereof). For example, the barren arctic or desert, primary (or secondary) forests, the tropical rain forest, the polar ice cap, etc., all produce images regarding the quantities of life forms. These examples can have special meaning not only in terms of the hazards or needs they create, but also in how a survivor can use available life forms (figure 2-4).

(1) Plant Life. There are hundreds of thousands of different types and species of plant life. In some instances, geographic areas are identified by the dominant types of plant life within that area. Examples of this are savannas, tundra, deciduous forests, etc. Some species of plant life can be used advantageously by a survivor—if not for the food or the water, then for improvising camouflage, shelter, or providing for other needs.

(2) Animal Life. Reptiles, amphibians, birds, fish, insects, and mammals are life forms which directly affect a survivor. These creatures affect the survivor by posing hazards (which must be taken into consideration), or by satisfying needs.

2-3. The Survivor's Condition. The survivor's condition and the influence it has in each survival episode is often overlooked. The primary factors which constitute the survivor's condition can best be described by the four categories shown in figure 2-5. Aircrew members must prepare themselves in each of these areas before each mission, and be in a state of "constant readiness" for the possibility of a "survival mission." Crewmembers must be aware of the role a survivor's condition plays both before and during the survival episode.

a. Physical. The physical condition and the fitness level of the survivor are major factors affecting

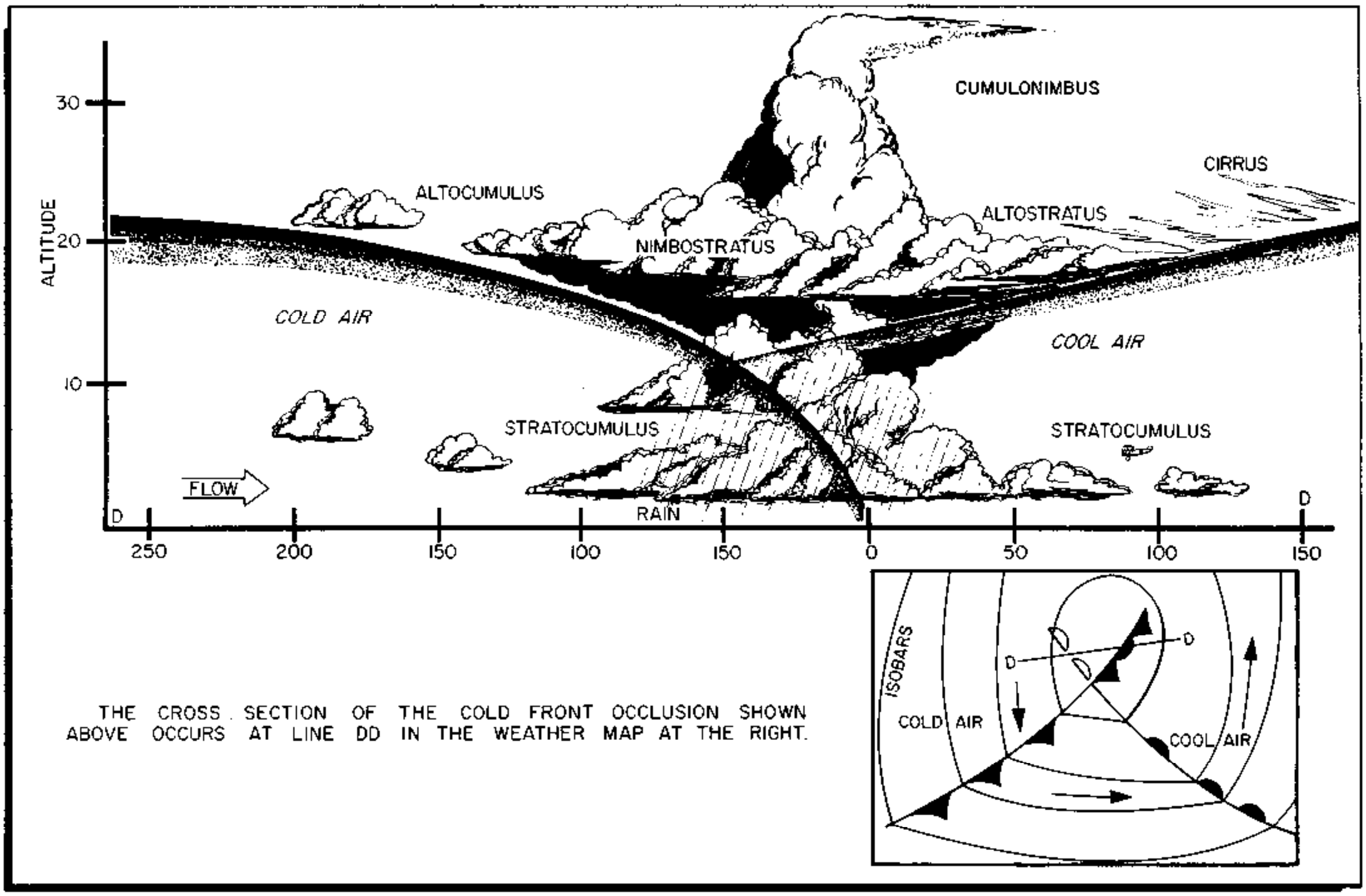

THE CROSS SECTION OF THE COLD FRONT OCCLUSION SHOWN ABOVE OCCURS AT LINE DD IN THE WEATHER MAP AT THE RIGHT.

Figure 2-2. Cold Front Occlusion.

survivability. Aircrew members who are physically fit will be better prepared to face survival episodes than those who are not. Further, a survivor's physical condition (injured or uninjured) during the initial phase of a survival episode will be a direct result of circumstances surrounding the ejection, bailout, parachute landing, or crash landing. In short, high levels of physical fitness and good post-egress physical condition will enhance a survivor's ability to cope with such diverse variables as: (1) temperature extremes, (2) rest or lack of it, (3) water availability, (4) food availability, and (5) extended survival episodes. In the last instance, physical weakness may increase as a result of nutritional deficiencies, disease, etc.

b. Psychological. Survivors' psychological state greatly influences their ability to successfully return from a survival situation.

(1) Psychological effectiveness in a survival episode (including captivity) results from effectively coping with the following factors:

(a) Initial shock - Finding oneself in a survival situation following the stress of ejection, bailout, or crash landing.

(b) Pain - Naturally occurring or induced by coercive manipulation.

(c) Hunger - Naturally occurring or induced by coercive manipulation.

(d) Thirst - Naturally occurring or induced by coercive manipulation.

(e) Cold or Heat - Naturally occurring or induced by coercive manipulation.

(f) Frustration - Naturally occurring or induced by coercive manipulation.

(g) Fatigue (including Sleep Deprivation). - Naturally occurring or induced by coercive manipulation.

(h) Isolation - Includes forced (captivity) and the extended duration of any episode.

(i) Insecurity - Induced by anxiety and self-doubts.

(j) Loss of self-esteem - Most often induced by coercive manipulation.

(k) Loss of self-determination - Most often induced by coercive manipulation.

(l) Depression - Mental "lows."

(2) A survivor may experience emotional reactions during a survival episode due to the previously stated

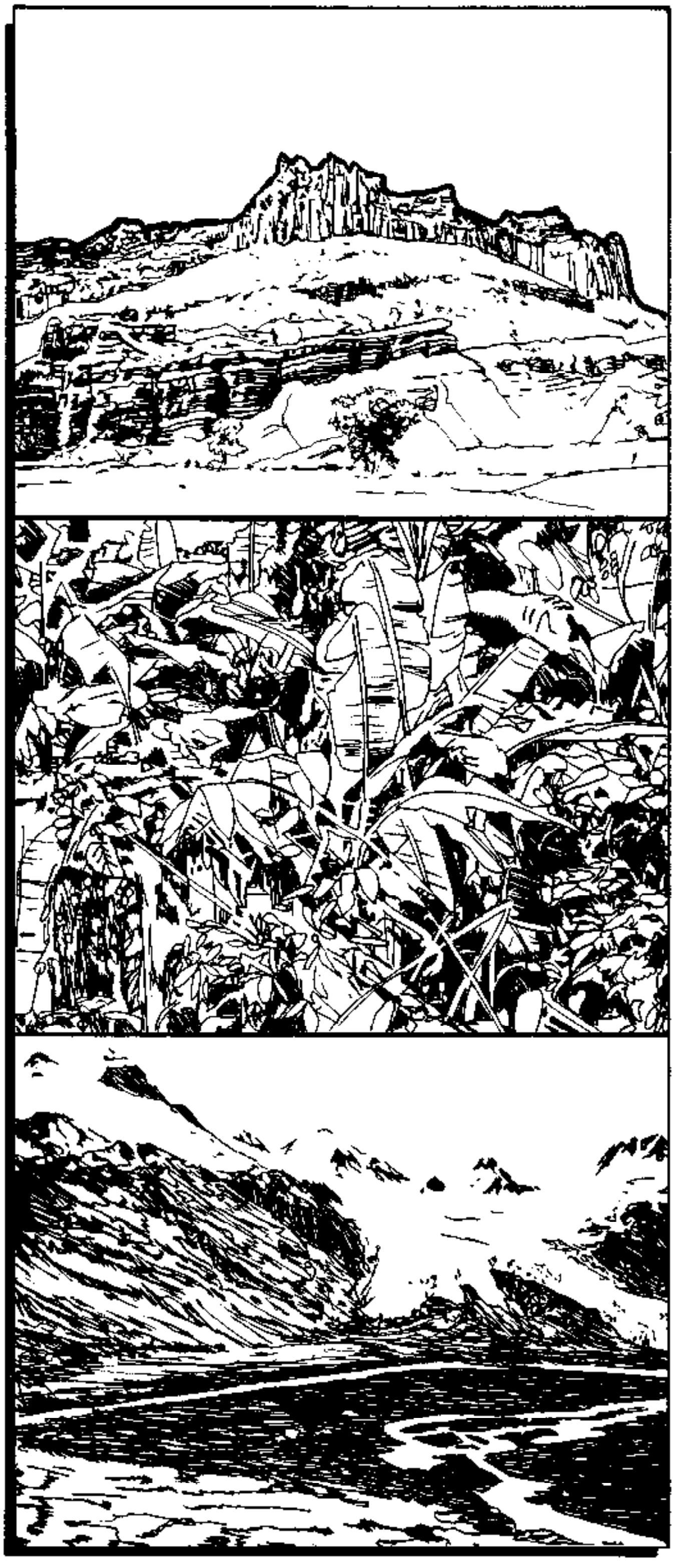

Figure 2-3. Terrain.

factors, previous (life) experiences (including training) and the survivor's psychological tendencies. Emotional reactions commonly occurring in survival (including captivity) situations are:

(a) Boredom - sometimes combined with loneliness.

(b) Loneliness.

(c) Impatience.

(d) Dependency.

(e) Humiliation.

(f) Resentment.

(g) Anger - sometimes included as a subelement of hate.

(h) Hate.

(i) Anxiety.

(j) Fear - often included as a part of panic or anxiety.

(k) Panic.

Figure 2-4. Life Forms.

(3) Psychologically survival episodes may be divided into "crisis" phases and "coping" phases. The initial crisis period will occur at the onset of the survival situation. During this initial period, "thinking" as well as "emotional control" may be disorganized. Judgment is impaired, and behavior may be irrational (possibly to the point of panic). Once the initial crisis is under control, the coping phase begins and the survivor is able to respond positively to the situation. Crisis periods may well recur, especially during extended situations (captivity). A survivor must strive to control if avoidance is impossible.

(4) The most important psychological tool that will affect the outcome of a survival situation is the *will to survive.* Without it, the survivor is surely doomed to failure—a strong will is the best assurance of survival.

c. Material. At the beginning of a survival episode, the clothing and equipment in the aircrew member's possession, the contents of available survival kits, and salvageable resources from the parachute or aircraft are the sum total of the survivor's material assets. Adequate premission preparations are required (must be stressed during training). Once the survival episode has started, special attention must be given to the care, use, and storage of all materials to ensure they continue to be

SURVIVOR'S CONDITION

- ○ PHYSICAL
- ○ PSYCHOLOGICAL
- ○ MATERIAL
- ○ OBLIGATORY

Figure 2-5. Survivor's Condition.

serviceable and available. Items of clothing and equipment should be selectively augmented with improvised items.

(1) Clothing appropriate to anticipated environmental conditions (on the ground) should be worn or carried as aircraft space and mission permit.

(2) The equipment available to a survivor affects all decisions, needs, and actions. The survivor's ability to improvise may provide ways to meet some needs.

d. Legal and Moral Obligations. A survivor has both legal and moral obligations or responsibilities. Whether in peacetime or combat, the survivor's responsibilities as a member of the military service continues. Legal obligations are expressly identified in the Geneva Conventions, Uniform Code of Military Justice (UCMJ), and Air Force directives and policies. Moral obligations are expressed in the Code of Conduct. (See figure 2-6.)

(1) Other responsibilities influence behavior during survival episodes and influence the *will to survive.* Examples include feelings of obligation or responsibilities to family, self, and(or) spiritual beliefs.

(2) A survivor's individual perception of responsibilities influence survival needs, and affect the psychological state of the individual both during and after the survival episode. These perceptions will be reconciled either consciously through rational thought or subconsciously through attitude changes. Training specifically structured to foster and maintain positive attitudes provides a key asset to survival.

2-4. Duration—The Time Condition. The duration of the survival episode has a major effect upon the aircrew member's needs. Every decision and action will be driven in part by an assessment of when recovery or return is probable. Air superiority, rescue capabilities, the distances involved, climatic conditions, the ability to locate the survivor, or captivity are major factors which directly influence the duration (time condition) of the survival episode. A survivor can never be certain that rescue is imminent.

2-5. Sociopolitical Condition. The people a survivor contacts, their social customs, cultural heritage, and political attitudes will affect the survivor's status. Warfare is one type of sociopolitical condition, and people of different cultures are another. Due to these sociopolitical differences, the interpersonal relationship between the survivor and any people with whom contact is established is crucial to surviving. To a survivor, the attitude of the people contacted will be friendly, hostile, or unknown.

a. Friendly People. The survivor who comes into contact with friendly people, or at least those willing (to some degree) to provide aid, is indeed fortunate. Immediate return to home, family, or home station, however, may be delayed. When in direct association with even the friendliest of people, it is essential to maintain their friendship. These people may be of a completely different culture in which a commonplace American habit may be a gross and serious insult. In other instances, the friendly people may be active insurgents in their country and constantly in fear of discovery. Every survivor action, in these instances, must be appropriate and acceptable to ensure continued assistance.

b. Hostile People. A state of war need not exist for a survivor to encounter hostility in people. With few exceptions, any contact with hostile people must be avoided. If captured, regardless of the political or social reasons, the survivor must make all efforts to adhere to the Code of Conduct and the legal obligations of the UCMJ, the Geneva Conventions, and USAF policy.

c. Unknown People. The survivor should consider all factors before contacting unknown people. Some primitive cultures and closed societies still exist in which outsiders are considered a threat. In other areas of the world, differing political and social attitudes can place a survivor "at risk" in contacting unknown people.

2-6. Induced Conditions. Any form of warlike activity results in "induced conditions." Three comparatively new induced conditions may occur during combat operations. Nuclear warfare and the resultant residual radiation, biological warfare, and chemical warfare (NBC) create life-threatening conditions from which a survivor needs immediate protection. The longevity of NBC conditions further complicates a survivor's other needs, decisions, and actions (figure 2-7).

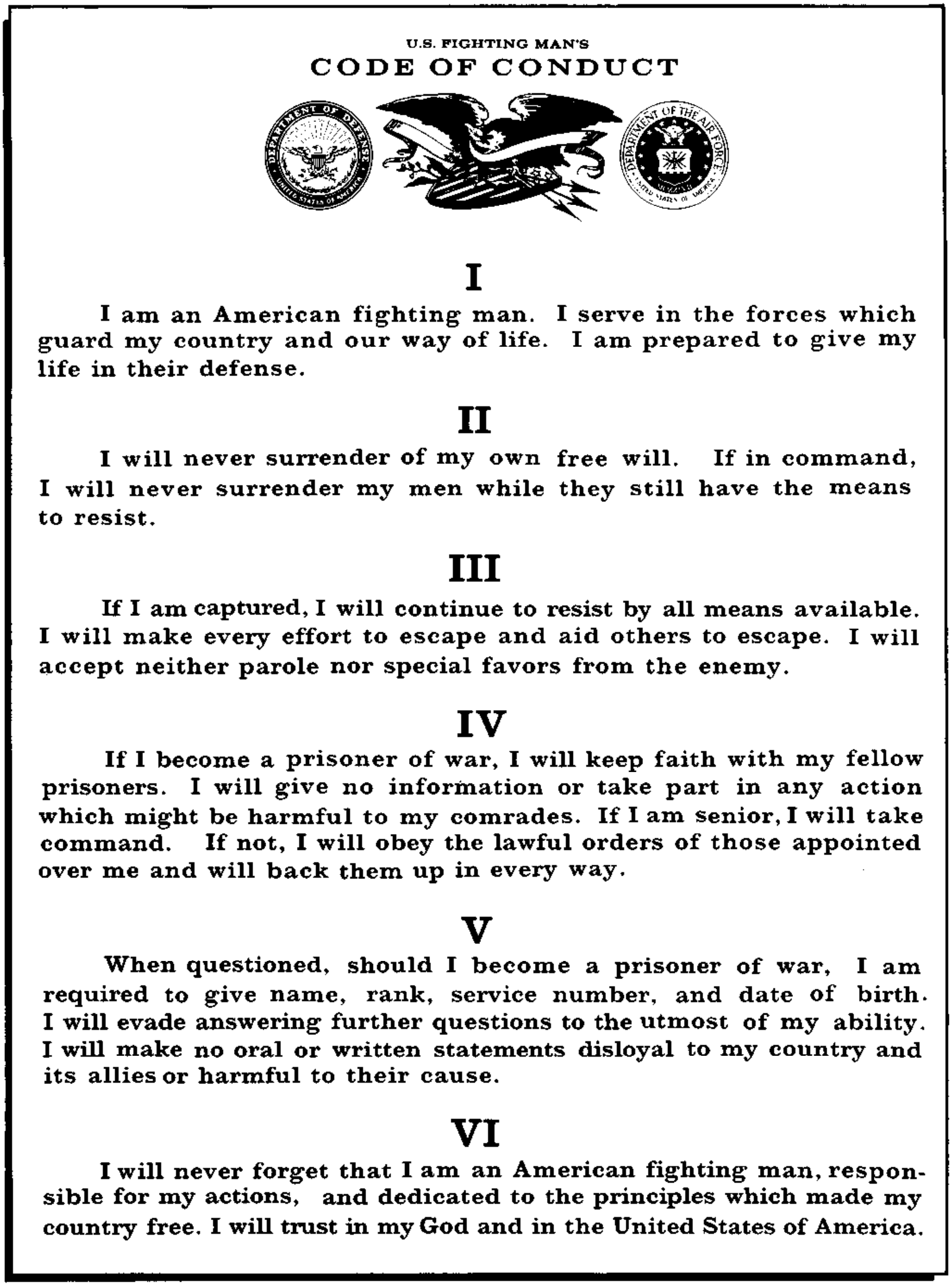

U.S. FIGHTING MAN'S

CODE OF CONDUCT

I

I am an American fighting man. I serve in the forces which guard my country and our way of life. I am prepared to give my life in their defense.

II

I will never surrender of my own free will. If in command, I will never surrender my men while they still have the means to resist.

III

If I am captured, I will continue to resist by all means available. I will make every effort to escape and aid others to escape. I will accept neither parole nor special favors from the enemy.

IV

If I become a prisoner of war, I will keep faith with my fellow prisoners. I will give no information or take part in any action which might be harmful to my comrades. If I am senior, I will take command. If not, I will obey the lawful orders of those appointed over me and will back them up in every way.

V

When questioned, should I become a prisoner of war, I am required to give name, rank, service number, and date of birth. I will evade answering further questions to the utmost of my ability. I will make no oral or written statements disloyal to my country and its allies or harmful to their cause.

VI

I will never forget that I am an American fighting man, responsible for my actions, and dedicated to the principles which made my country free. I will trust in my God and in the United States of America.

Figure 2-6. Code of Conduct.

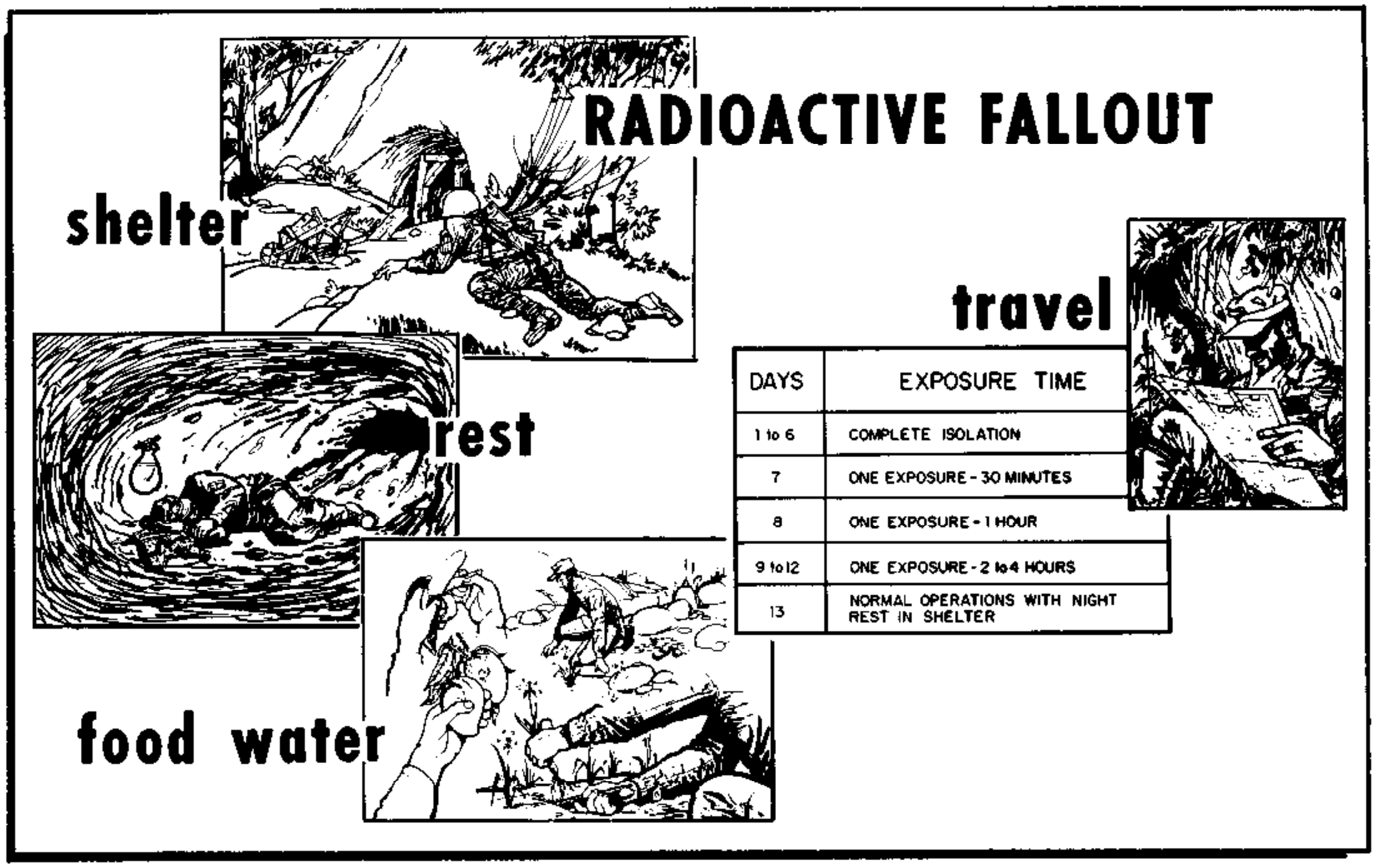
RADIOACTIVE FALLOUT
shelter
travel
rest
food water

DAYS	EXPOSURE TIME
1 to 6	COMPLETE ISOLATION
7	ONE EXPOSURE - 30 MINUTES
8	ONE EXPOSURE - 1 HOUR
9 to 12	ONE EXPOSURE - 2 to 4 HOURS
13	NORMAL OPERATIONS WITH NIGHT REST IN SHELTER

Figure 2-7. Induced Conditions.

Chapter 3

THE SURVIVOR'S NEEDS

3-1. Introduction. The three fundamental goals of a survivor—to maintain life, maintain honor, and return—may be further divided into eight basic needs. In a noncombatant situation, these needs include: personal protection, sustenance, health, travel, and communications (signaling for recovery). During combat, additional needs must be fulfilled. They are: evasion, resistance if captured, and escape if captured. Meeting the individual's needs during the survival episode is essential to achieving the survivor's fundamental goals (figure 3-1).

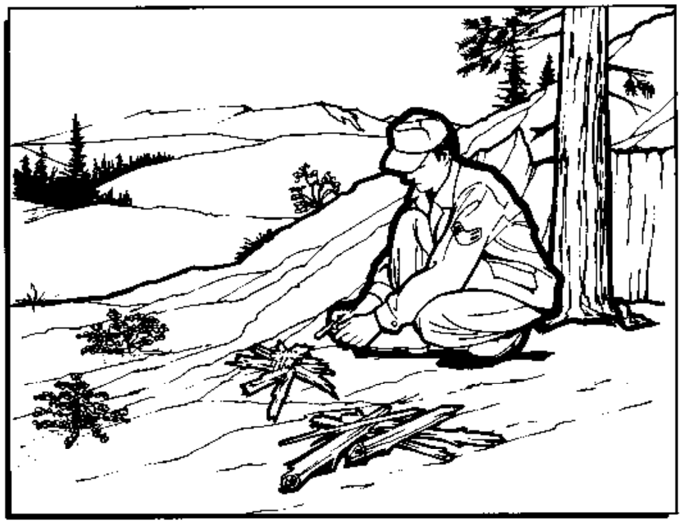

Figure 3-1. Survivor's Needs.

3-2. Maintaining Life. Three elementary needs of a survivor in any situation which are categorized as the integral components of maintaining life are: personal protection, sustenance, and health.

a. Personal Protection. The human body is comparatively fragile. Without protection, the effects of environmental conditions (climate, terrain, and life forms) and of induced conditions (radiological, biological agents, and chemical agents) may be fatal. The survivor's primary defenses against the effects of the environment are clothing, equipment, shelter, and fire. Additionally, clothing, equipment, and shelter are the primary defenses against some of the effects of induced conditions (figure 3-2).

(1) The need for adequate clothing and its proper care and use cannot be overemphasized. The human body's tolerance for temperature extremes is very limited. However, its ability to regulate heating and cooling is extraordinary. The availability of clothing and its proper use is extremely important to a survivor in using these abilities of the body. Clothing also provides excellent protection against the external effects of alpha and beta radiation, and may serve as a shield against the external effects of some chemical or biological agents.

(2) Survival equipment is designed to aid survivors throughout their episode. It must be cared for to maintain its effectiveness. Items found in a survival kit or aircraft can be used to help satisfy the eight basic needs. Quite often, however, a survivor must improvise to overcome an equipment shortage or deficiency.

(3) The survivor's need for shelter is twofold—as a place to rest and for protection from the effects of the environmental and(or) induced conditions (NBC). The duration of the survival episode will have some effect on shelter choice. In areas that are warm and dry, the survivor's need is easily satisfied using natural resting places. In cold climates, the criticality of shelter can be measured in minutes, and rest is of little immediate concern. Similarly, in areas of residual radiation, the criticality of shelter may also be measured in minutes (figure 3-3).

(4) Fire serves many survivor needs: purifying water, cooking and preserving food, signaling, and providing a source of heat to warm the body and dry clothing (figure 3-4).

b. Sustenance. Survivors need food and water to maintain normal body functions and to provide strength, energy, and endurance to overcome the physical stresses of survival.

(1) Water. The survivor must be constantly aware of the body's continuing need for water (figure 3-5).

(2) Food. During the first hours of a survival situation, the need for food receives little attention. During the first 2 or 3 days, hunger becomes a nagging aggravation which a survivor can overcome. The first major food crisis occurs when the loss of energy, stamina, and strength begin to affect the survivor's physical capabilities. The second major food crisis is more insidious. A marked increase in irritability and other attitudes may occur as the starvation process continues. Early and continuous attention must be given to obtaining and using any and all available food. Most people have food preferences. The natural tendency to avoid certain types of food is a major problem which must be overcome early in the survival situation. The starvation process ultimately overcomes all food aversions. The successful survivor overcomes these aversions before physical or psychological deterioration sets in (figure 3-6).

c. Health (Physical and Psychological). The survivor must be the doctor, nurse, corpsman, psychologist, and cheerleader. Self-aid is the survivor's sole recourse.

(1) Prevention. The need for preventive medicine and safety cannot be overemphasized. Attention to sani-

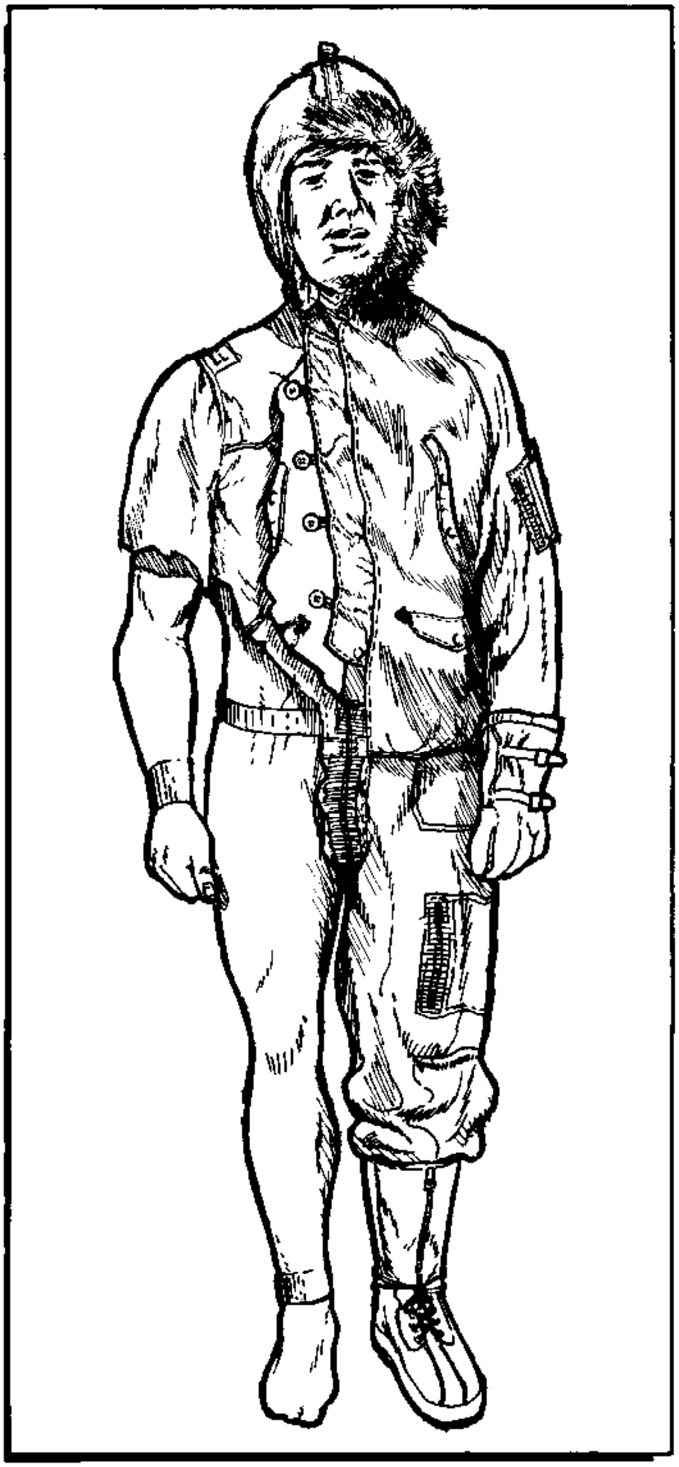

Figure 3-2. Personal Protection.

tation and personal hygiene is a major factor in preventing physical, morale, and attitudinal problems.

(a) The need for cleanliness in the treatment of injuries and illness is self-evident. The prisoner of war (PW) who used maggots to eat away rotting flesh caused by infection is a dramatic example. Prevention is much more preferred than such drastic procedures.

(b) Safety must be foremost in the mind of the survivor; carelessness is caused by ignorance and(or) poor judgment or bad luck. One miscalculation with a knife or ax can result in self-inflicted injury or death.

(2) Self-Aid:

(a) Injuries frequently occur during ejection, bailout, parachute landing, or ditching. Other post-egress factors may also cause injury. In the event of injury, the survivor's existence may depend on the ability to perform self-aid. In many instances, common first aid procedures will suffice; in others, more primitive techniques will be required (figure 3-7).

(b) Illness and the need to treat it is more commonly associated with long-term situations such as an extended evasion episode or captivity. When preventive techniques have failed, the survivor must treat symptoms of disease in the absence of professional medical care.

(3) Psychological Health. Perhaps the survivor's greatest need is the need for emotional stability and a positive, optimistic attitude. An individual's ability to cope with psychological stresses will enhance successful survival. Optimism, determination, dedication, and humor, as well as many other psychological attributes, are all helpful for a survivor to overcome psychological stresses (figure 3-8).

3-3. Maintaining Honor. Three elementary needs which a survivor may experience during combat survival situations are categorized as integral components of maintaining honor. These three elementary needs are: (a) avoiding capture or evading, (b) resisting (if captured), and (c) escaping (if captured).

a. Avoiding Capture. Evasion will be one of the most difficult and hazardous situations a survivor will face. However difficult and hazardous evasion may be, captivity is always worse. During an evasion episode, the survivor has two fundamental tasks. The first is to use concealment techniques. The second is to use evasion movement techniques. The effective use of camouflage is common to both of these activities.

(1) Hiding oneself and all signs of presence are the evader's greatest needs. Experience indicates that the survivor who uses effective concealment techniques has a better chance of evading capture. Capture results most frequently when the evader is moving.

(2) The evader's need to move depends on a variety of needs such as recovery, food, water, better shelter, etc. Evasion movement is more successful when proven techniques are used.

b. Resisting. The PW's need to resist is self-evident. This need is both a moral and a legal obligation. Resistance is much more than refusing to divulge some bit of classified information. Fundamentally, resisting is two distinctly separate behaviors expected of the prisoner:

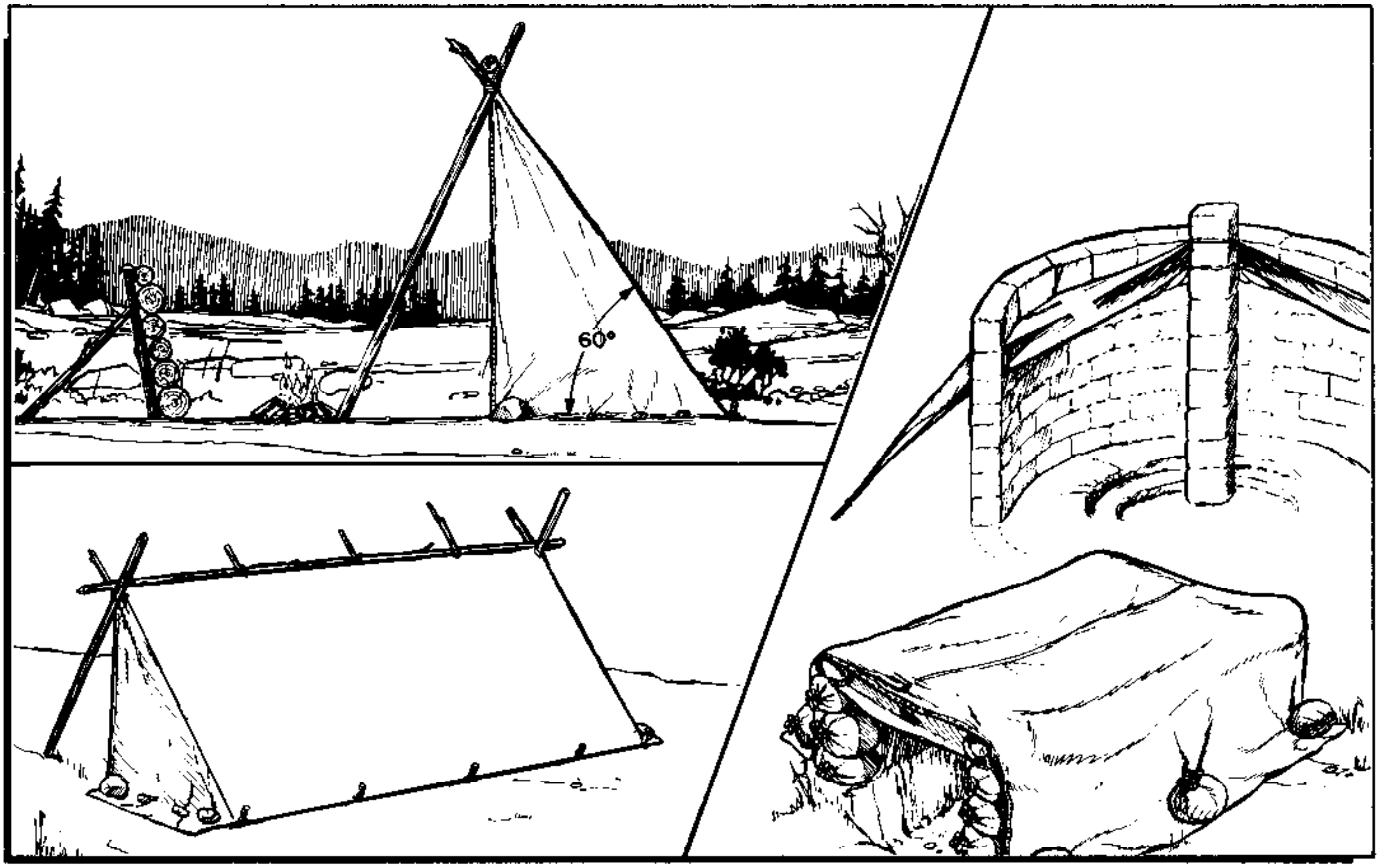

Figure 3-3. Shelters.

(1) Complying with legal and authorized requirements only.

(2) Disrupting enemy activity through resisting, subtle harrassment, and tying up enemy guards who could be used on the front lines.

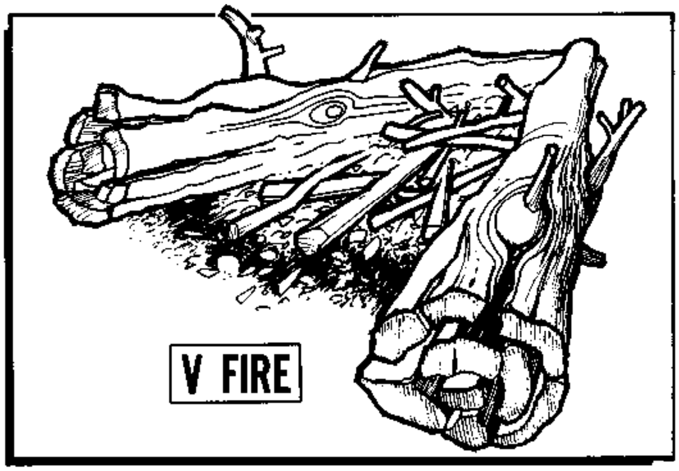

Figure 3-4. Fire.

c. Escaping (When Possible and Authorized). Escape is neither easy nor without danger. The Code of Conduct states a survivor should make every effort to escape and aid others to escape.

3-4. Returning. The need to return is satisfied by successful completion of one or both of the basic tasks confronting the survivor: aiding with recovery and traveling (on land or water).

a. Aiding With Recovery. For survivors or evaders to effectively aid in recovery, they must be able to make their position and the situation on the ground known. This is done either electronically, visually, or both (figure 3-9).

(1) Electronic signaling covers a wide spectrum of techniques. As problems such as security and safety during combat become significant factors, procedures for using electronic signaling to facilitate recovery become increasingly complex.

(2) Visual signaling is primarily the technique for attracting attention and pinpointing an exact location for rescuers. Simple messages or information may also be transmitted with visual signals.

b. Travel On Land. A survivor may need to move on land for a variety of reasons, ranging from going for

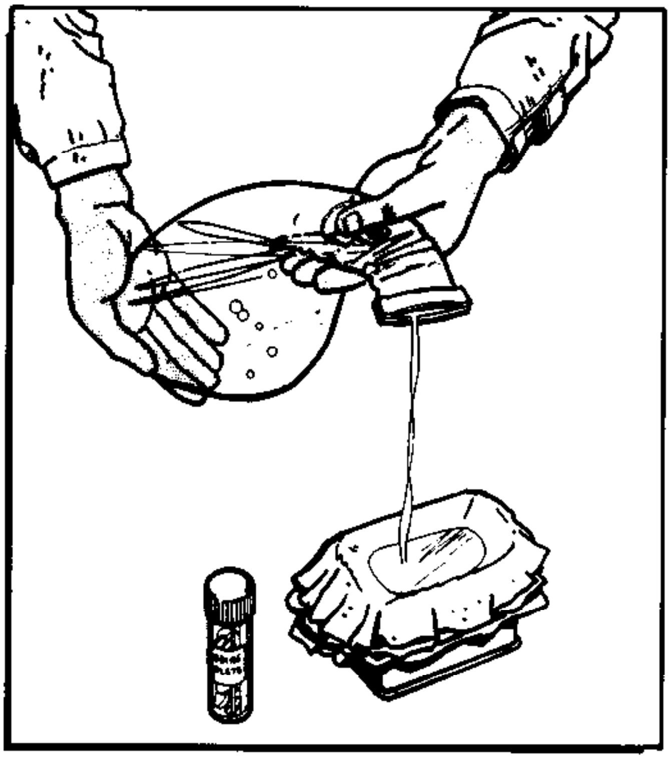

Figure 3-5. Water.

water to attempting to walk out of the situation. In any survival episode, the survivor must weigh the need to travel against capabilities and(or) safety (figure 3-10). Factors to consider may include:

Figure 3-6. Food.

(1) The ability to walk or traverse existing terrain. In a nonsurvival situation, a twisted or sprained ankle is an inconvenience accompanied by some temporary pain and restricted activity. A survivor who loses the mobility, due to injury, to obtain food, water, and shelter, can face death. There is a safe and effective way to travel across almost any type of terrain.

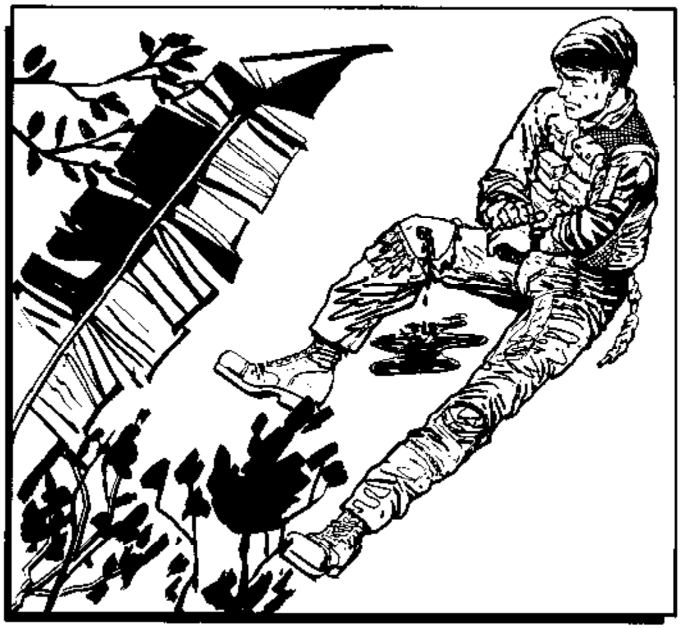

Figure 3-7. Self-Aid.

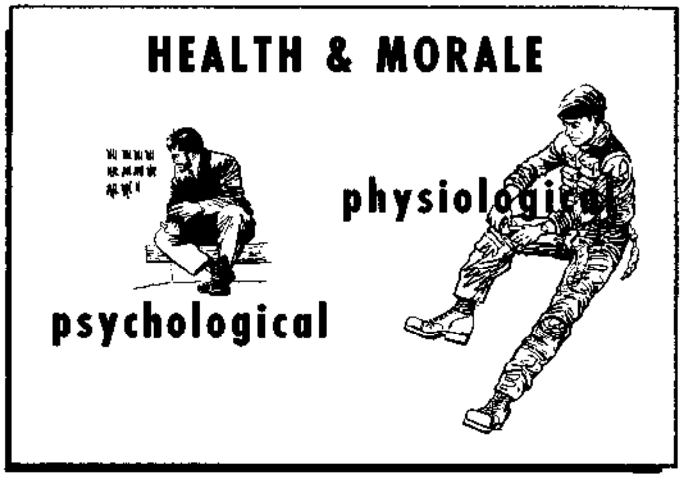

Figure 3-8. Health and Morale.

(2) The need to transport personal possessions (burden carrying). There are numerous documented instances of survivors abandoning equipment and clothing simply because carrying was a bother. Later, the abandoned materials were not available when needed to save life, limb, or aid in rescue. Burden carrying need not be difficult or physically stressful. There are many simple ways for a survivor to carry the necessities of life (figure 3-11).

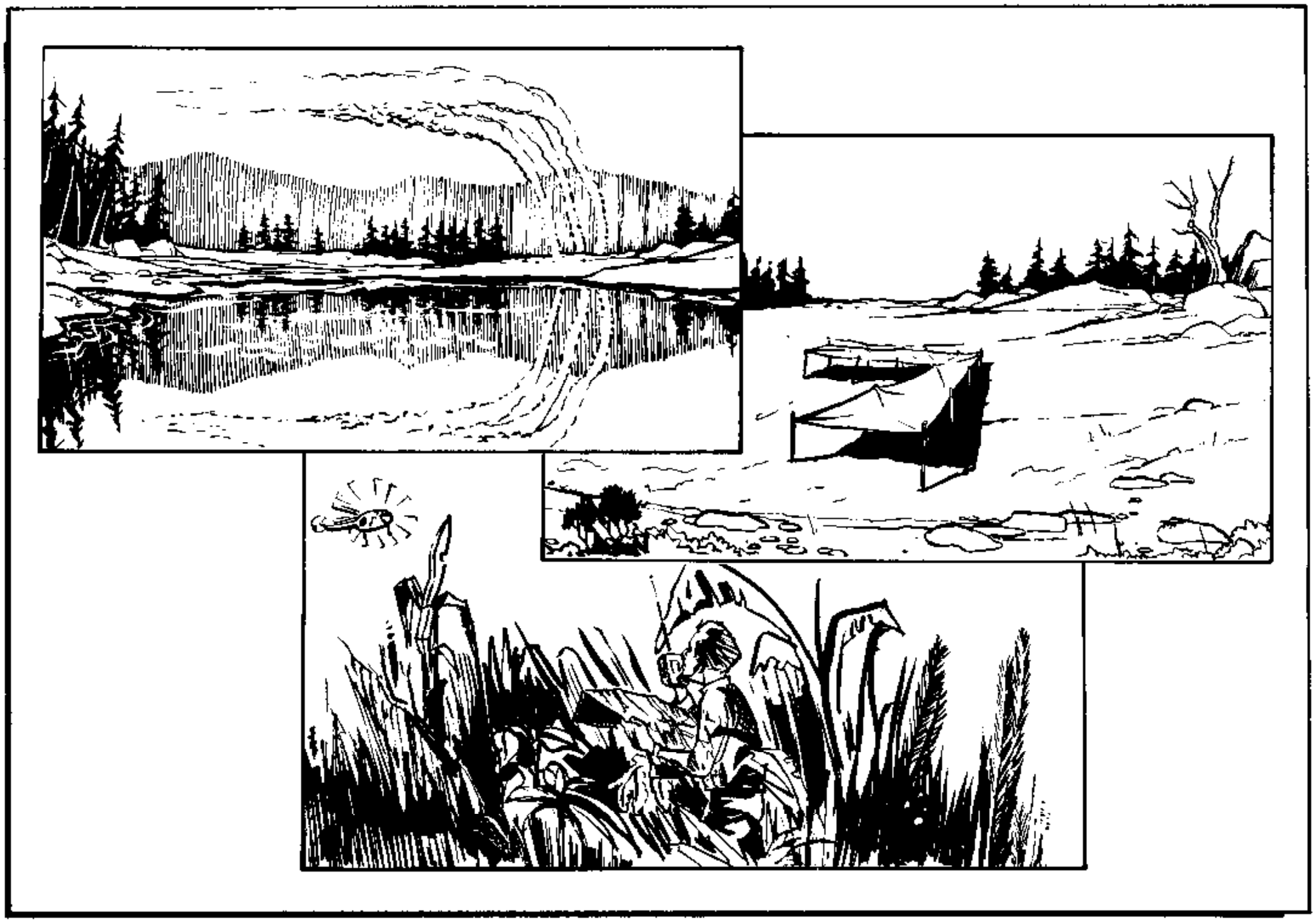

Figure 3-9. Recovery.

(3) The ability to determine present position. Maps, compasses, star charts, Weems plotters, etc., permit accurate determination of position during extended travel. Yet, the knowledgeable, skillful, and alert survivor can do well without a full complement of these aids. Constant awareness, logic, and training in nature's clues to navigation may allow a survivor to determine general location even in the absence of detailed navigation aids.

(4) Restrictions or limitations to select and maintain a course of travel. The tools used in determining position are the tools used to maintain a course of travel. A straight line course to a destination is usually the simplest, but may not always be the best course for travel. Travel courses may need to be varied for diverse reasons, such as to get food or water, to enhance covert travels or to avoid hazardous or impassable obstacles or terrain. Careful planning and route selection before and during travel is essential.

c. Travel On Water. Two differing circumstances may require survivors to travel on water. First, those who crash-land or parachute into the open sea are confronted with one type of situation. Second, survivors who find a river or stream which leads in a desirable direction are faced with a different situation. In each instance, however, a common element is to stay afloat.

(1) The survivor's initial problems on the open sea are often directly related to the winds and size of the waves. Simply getting into a liferaft and staying there are often very difficult tasks. On the open sea, the winds and ocean currents have a significant effect on the direction of travel. As the survivor comes closer to shore, the direction in which the tide is flowing also becomes a factor. There are some techniques a survivor can use to aid with stabilizing the raft, controlling the direction and rate of travel, and increasing safety.

Figure 3-10. Travel.

(2) Survivors using rivers or streams for travel face both hazards and advantages as compared to overland travel. First, floating with the current is far less difficult than traveling overland. An abundance of food and water are usually readily available. Even in densely forested areas, effective signaling sites are generally available along streambeds. A survivor must use care and caution to avoid drowning, the most serious hazard associated with river travel.

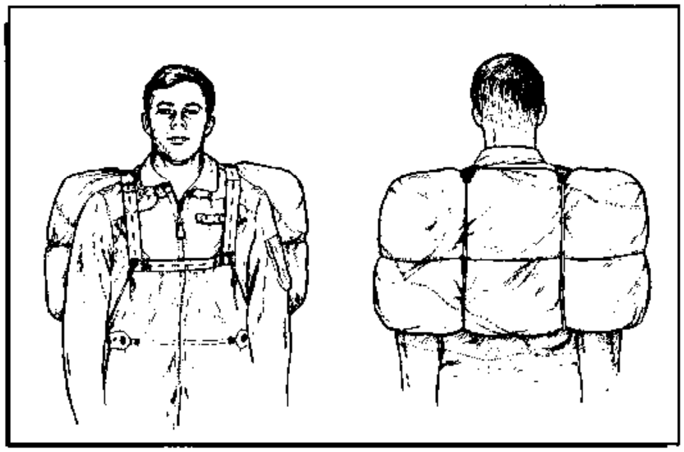

Figure 3-11. Burden Carrying.

Part Two

PSYCHOLOGICAL ASPECTS OF SURVIVAL

Chapter 4

CONTRIBUTING FACTORS

4-1. Introduction. Aircrew members in a survival situation must recognize that coping with the psychological aspects of survival are at least as important as handling the environmental factors. In virtually any survival episode, the aircrew will be in an environment that can support human life. The survivors' problems will be compounded because they never really expected to bail out or crash-land in the jungle, over the ocean, or anywhere else. No matter how well prepared, aircrews probably will never completely convince themselves that "it can happen to them." However, the records show it can happen. Before aircrew members learn about the physical aspects of survival, they must first understand that psychological problems may occur and that solutions to those problems must be found if the survival episode is to reach a successful conclusion (figure 4-1).

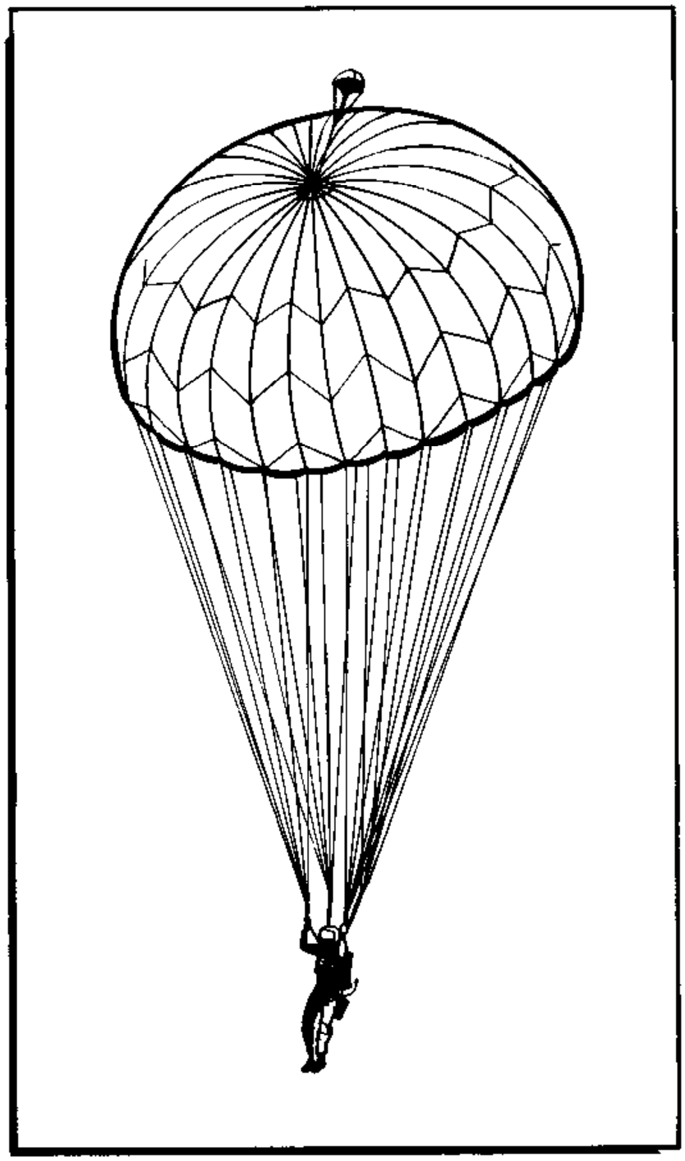

Figure 4-1. Psychological Aspects.

4-2. Survival Stresses:

a. The emotional aspects associated with survival must be completely understood just as survival conditions and equipment are understood. An important factor bearing on success or failure in a survival episode is the individual's psychological state. Maintaining an even, positive psychological state or outlook depends on the individual's ability to cope with many factors. Some include:

(1) Understanding how various physiological and emotional signs, feelings, and expressions affect one's bodily needs and mental attitude.

(2) Managing physical and emotional reactions to stressful situations.

(3) Knowing individual tolerance limits, both psychological and physical.

(4) Exerting a positive influence on companions.

b. Nature has endowed everyone with biological mechanisms which aid in adapting to stress. The bodily changes resulting from fear and anger, for example, tend to increase alertness and provide extra energy to either run away or fight. These and other mechanisms can hinder a person under survival conditions. For instance, a survivor in a raft could cast aside reason and drink sea water to quench a thrist; or, evaders in enemy territory, driven by hunger pangs, could expose themselves to capture when searching for food. These examples illustrate how "normal" reactions to stress could create problems for a survivor.

c. Two of the gravest threats to successful survival are concessions to comfort and apathy. Both threats represent attitudes which must be avoided. To survive, a person must focus planning and effort on fundamental needs.

(1) Many people consider comfort their greatest need. Yet, comfort is not essential to human survival. Survivors must value life more than comfort, and be willing to tolerate heat, hunger, dirt, itching, pain, and any other discomfort. Recognizing discomfort as tem-

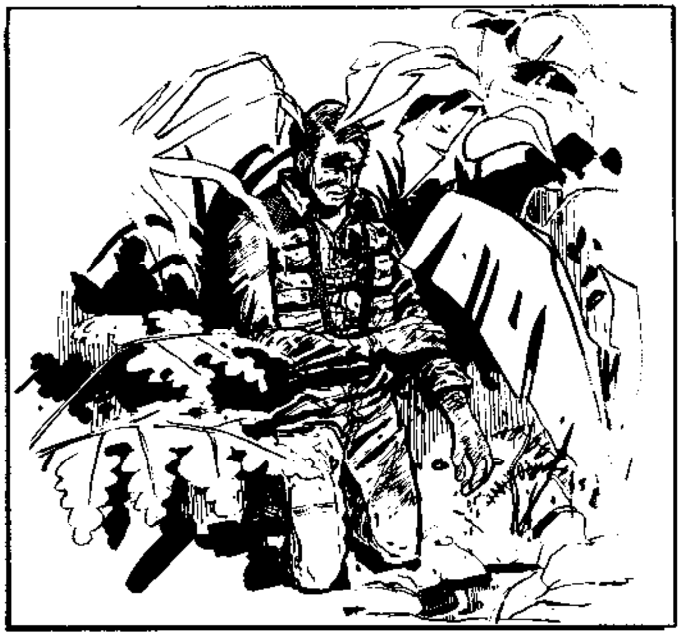

Figure 4-2. Pain.

porary will help survivors concentrate on effective action.

(2) As the will to keep trying lessens, drowsiness, mental numbness, and indifference will result in apathy. This apathy usually builds on slowly, but ultimately takes over and leaves a survivor helpless. Physical factors can contribute to apathy. Exhaustion due to prolonged exposure to the elements, loss of body fluids (dehydration), fatigue, weakness, or injury are all conditions which can contribute to apathy. Proper planning and sound decisions can help a survivor avoid these conditions. Finally, survivors must watch for signs of apathy in companions and help prevent it. The first signs are resignation, quietness, lack of communication, loss of appetite, and withdrawal from the group. Preventive measures could include maintaining group morale by planning, activity, and getting the organized participation of all members.

d. Many common stresses cause reactions which can be recognized and dealt with appropriately in survival situations. A survivor must understand that stresses and reactions often occur at the same time. Although survivors will face many stresses, the following common stresses will occur in virtually all survival episodes: pain, thirst, cold and heat, hunger, frustration, fatigue, sleep deprivation, isolation, insecurity, loss of self-esteem, loss of self-determination, and depression.

4-3. Pain:

a. Pain, like fever, is a warning signal calling attention to an injury or damage to some part of the body. Pain is discomforting but is not, in itself, harmful or dangerous. Pain can be controlled, and in an extremely grave situation, survival must take priority over giving in to pain (figure 4-2).

b. The biological function of pain is to protect an injured part by warning the individual to rest it or avoid using it. In a survival situation, the normal pain warnings may have to be ignored in order to meet more critical needs. People have been known to complete a fight with a fractured hand, to run on a fractured or sprained ankle, to land an aircraft despite severely burned hands, and to ignore pain during periods of intense concentration and determined effort. Concentration and intense effort can actually stop or reduce feelings of pain. Sometimes this concentration may be all that is needed to survive.

c. A survivor must understand the following facts about pain:

(1) Despite pain, a survivor can move in order to live.

(2) Pain can be reduced by:

(a) Understanding its source and nature.

(b) Recognizing pain as a discomfort to be tolerated.

(c) Concentrating on necessities like thinking, planning, and keeping busy.

(d) Developing confidence and self-respect. When personal goals are maintaining life, honor, and returning, and these goals are valued highly enough, a survivor can tolerate almost anything.

4-4. Thirst and Dehydration:

a. The lack of water and its accompanying problems of thirst and dehydration are among the most critical problems facing survivors. Thirst, like fear and pain, can be tolerated if the will to carry on, supported by calm, purposeful activity, is strong. Although thirst indicates the body's need for water, it does not indicate how much water is needed. If a person drinks only enough to satisfy thirst, it is still possible to slowly dehydrate. Prevention of thirst and the more debilitating dehydration is possible if survivors drink plenty of water any time it is available, and especially when eating (figure 4-3).

b. When the body's water balance is not maintained, thirst and discomfort result. Ultimately, a water imbalance will result in dehydration. The need for water may be increased if the person:

(1) Has a fever.

(2) Is fearful.

(3) Perspires unnecessarily.

(4) Rations water rather than sweat.

c. Dehydration decreases the body's efficiency or ability to function. Minor degrees of dehydration may not noticeably affect a survivor's performance, but as it becomes more severe, body functioning will become increasingly impaired. Slight dehydration and thirst can

Figure 4-3. Thirst.

also cause irrational behavior. One survivor described it:

> "The next thing I remember was being awakened by an unforgettable sensation of thirst. I began to move about aimlessly and finally found a pool of water."
>
> "We finally found water. In the water were two dead deer with horns locked. We went down to the water and drank away. It was the best damned drink of water I ever had in my life. I didn't taste the stench of the deer at all."

While prevention is the best way to avoid dehydration, virtually any degree of dehydration is reversible simply by drinking water.

4-5. Cold and Heat. The average normal body temperature for a person is 98.6°F. Victims have survived a body temperature as low as 20°F below normal, but consciousness is clouded and thinking numbed at a much smaller drop. An increase of 6 to 8 degrees above normal for any prolonged period may prove fatal. Any deviation from normal temperature, even as little as 1 or 2 degrees, reduces efficiency.

a. Cold is a serious stress since even in mild degrees it lowers efficiency. Extreme cold numbs the mind and dulls the will to do anything except get warm again. Cold numbs the body by lowering the flow of blood to the extremities, and results in sleepiness. Survivors have endured prolonged cold and dampness through exercise, proper hygiene procedures, shelter, and food. Wearing proper clothing and having the proper climatic survival equipment when flying in cold weather areas are essential to enhance survivability (figure 4-4).

(1) One survivor described cold and its effect:

> "Because of the cold water, my energy was going rapidly and all I could do was to hook my left arm over one side of the raft, hang on, and watch the low flying planes as they buzzed me...As time progressed, the numbing increased...and even seemed to impair my thinking."

(2) Another survivor remembered survival training and acted accordingly:

> "About this time, my feet began getting cold. I remembered part of the briefing I had received about feet freezing so I immediately took action. I thought about my shoes, and with my jack knife, cut off the bottom of my Mark II immersion suit and put them over my shoes. My feet immediately felt warmer and the rubber feet of the immersion suit kept the soles of my shoes dry."

b. Just as "numbness" is the principal symptom of cold, "weakness" is the principle symptom of heat. Most people can adjust to high temperatures, whether in the hold of a ship or in a harvest field on the Kansas prairie. It may take from 2 days to a week before circulation, breathing, heart action, and sweat glands are all adjusted to a hot climate. Heat stress also accentuates dehydration, which was discussed earlier. In addition to the problem of water, there are many other sources of discomfort and impaired efficiency which are directly attributable to heat or to the environmental conditions in hot climates. Extreme temperature changes, from extremely hot days to very cold nights, are experienced in desert and plains areas. Proper use of clothing and shelters can decrease the adverse effects of such extremes (figure 4-5).

c. Bright sun has a tremendous effect on eyes and any exposed skin. Direct sunlight or rays reflecting off the terrain require dark glasses or improvised eye protec-

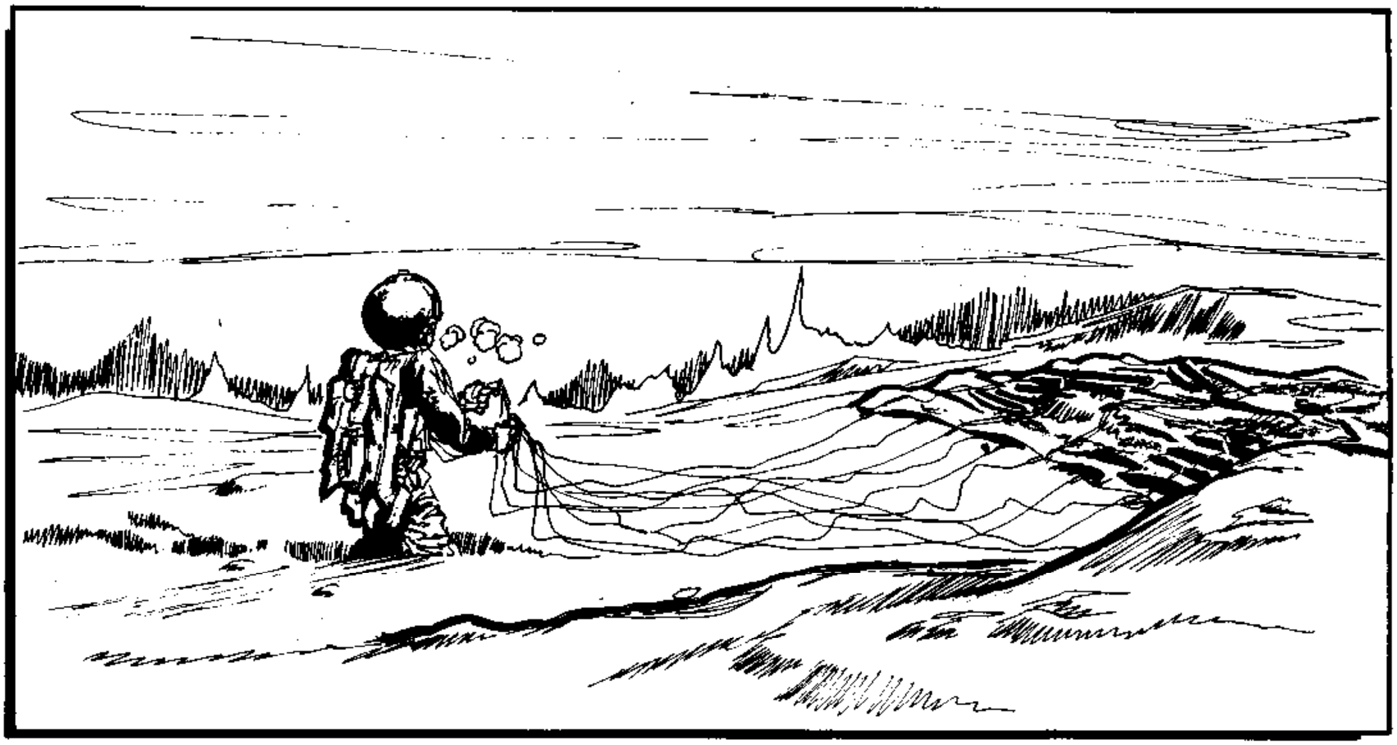

Figure 4-4. Cold.

tors. Previous suntanning provides little protection; protective clothing is important.

d. Blowing wind, in hot summer, has been reported to get on some survivors' nerves. Wind can constitute an additional source of discomfort and difficulty in desert areas when it carries particles of sand and dirt. Protection against sand and dirt can be provided by tying a cloth around the head after cutting slits for vision.

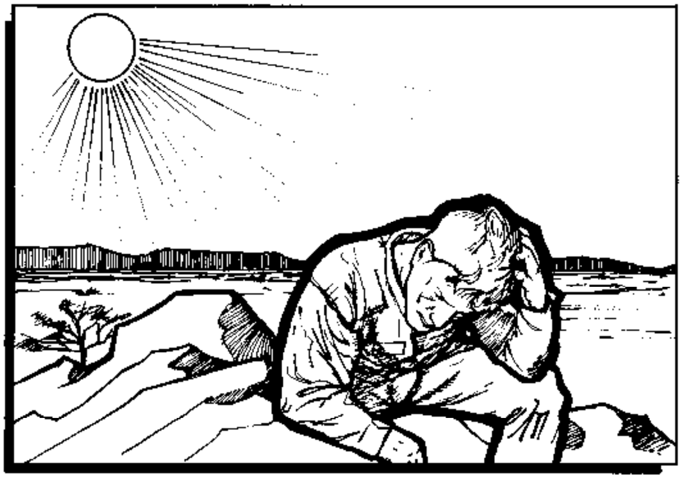

Figure 4-5. Heat.

e. Acute fear has been experienced among survivors in sandstorms and snowstorms. This fear results from both the terrific impact of the storm itself and its obliteration of landmarks showing direction of travel. Finding or improving shelter for protection from the storm itself is important.

f. Loss of moisture, drying of the mouth and mucous membranes, and accelerate dehydration can be caused by breathing through the mouth and talking. Survivors must learn to keep their mouths shut in desert winds as well as in cold weather.

g. Mirages and illusions of many kinds are common in desert areas. These illusions not only distort visual perception but sometimes account for serious incidents. In the desert, distances are usually greater than they appear and, under certain conditions, mirages obstruct accurate vision. Inverted reflections are a common occurrence.

4-6. Hunger. A considerable amount of edible material (which survivors may not initially regard as food) may be available under survival conditions. Hunger and semistarvation are more commonly experienced among survivors than thirst and dehydration. Research has revealed no evidence of permanent damage nor any decrease in mental efficiency from short periods of total fasting (figure 4-6).

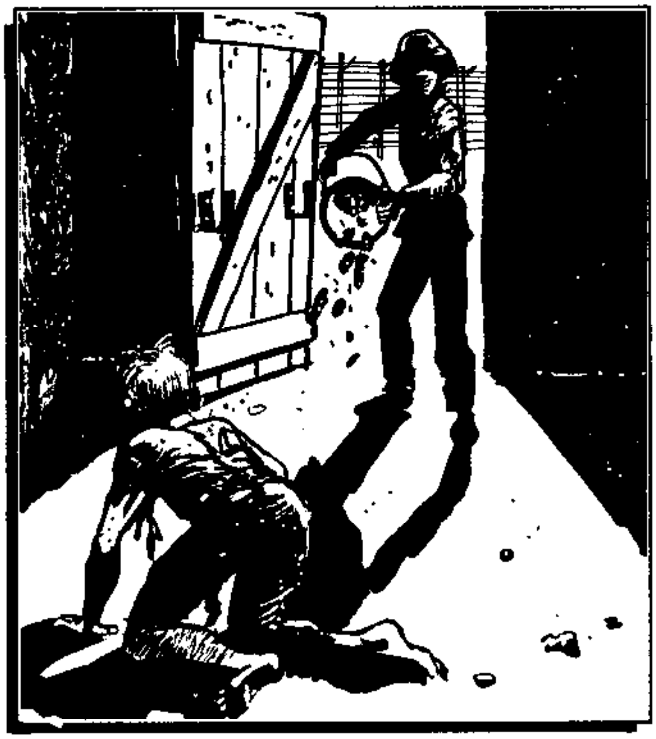

Figure 4-6. Hunger.

a. The prolonged and rigorous Minnesota semistarvation studies during World War II revealed the following behavioral changes:

(1) Dominance of the hunger drive over other drives.

(2) Lack of spontaneous activity.

(3) Tired and weak feeling.

(4) Inability to do physical tasks.

(5) Dislike of being touched or caressed in any way.

(6) Quick susceptibility to cold.

(7) Dullness of all emotional responses (fear, shame, love, etc.)

(8) Lack of interest in others—apathy.

(9) Dullness and boredom.

(10) Limited patience and self-control.

(11) Lack of a sense of humor.

(12) Moodiness—reaction of resignation.

b. Frequently, in the excitement of some survival, evasion, and escape episodes, hunger is forgotten. Survivors have gone for considerable lengths of time without food or awareness of hunger pains. An early effort should be made to procure and consume food to reduce the stresses brought on by food deprivation. Both the physical and psychological effects described are reversed when food and a protective environment are restored. Return to normal is slow and the time necessary for the return increases with the severity of starvation. If food deprivation is complete and only water is ingested, the pangs of hunger disappear in a few days, but even then the mood changes of depression and irritability occur. The individual tendency is still to search for food to prevent starvation and such efforts might continue as long as strength and self-control permit. When the food supply is limited, even strong friendships are threatened.

c. Food aversion may result in hunger. Adverse group opinion may discourage those who might try foods unfamiliar to them. In some groups, the barrier would be broken by someone eating the particular food rather than starving. The solitary individual has only personal prejudices to overcome and will often try strange foods.

d. Controlling hunger during survival episodes is relatively easy if the survivor can adjust to discomfort and adapt to primitive conditions. This man would rather survive than be fussy:

> "Some men would almost starve before eating the food. There was a soup made of lamb's head with the lamb's eyes floating around in it....When there was a new prisoner, I would try to find a seat next to him so I could eat the food he refused."

4-7. Frustration. Frustration occurs when one's efforts are stopped, either by obstacles blocking progress toward a goal or by not having a realistic goal. It can also occur if the feeling of self-worth or self-respect is lost (figure 4-7).

a. A wide range of obstacles, both environmental and internal, can lead to frustration. Frustrating conditions often create anger, accompanied by a tendency to attack and remove the obstacles to goals.

b. Frustration must be controlled by channeling energies into a positive and worthwhile obtainable goal. The survivor should complete the easier tasks before attempting more challenging ones. This will not only instill self-confidence, but also relieve frustration.

4-8. Fatigue. In a survival episode, a survivor must continually cope with fatigue and avoid the accompanying strain and loss of efficiency. A survivor must be aware of the dangers of over-exertion. In many cases, a survivor may already be experiencing strain and reduced efficiency as a result of other stresses such as heat or cold, dehydration, hunger, or fear. A survivor must judge capacity to walk, carry, lift, or do necessary work, and plan and act accordingly. During an emergency, considerable exertion may be necessary to cope with the situation. If an individual understands fatigue and the attitudes and feelings generated by various kinds of effort, that individual should be able to call on available reserves of energy when they are needed (figure 4-8).

a. A survivor must avoid complete exhaustion which may lead to physical and psychological changes. A survivor should be able to distinguish between exhaustion and being uncomfortably tired. Although a person

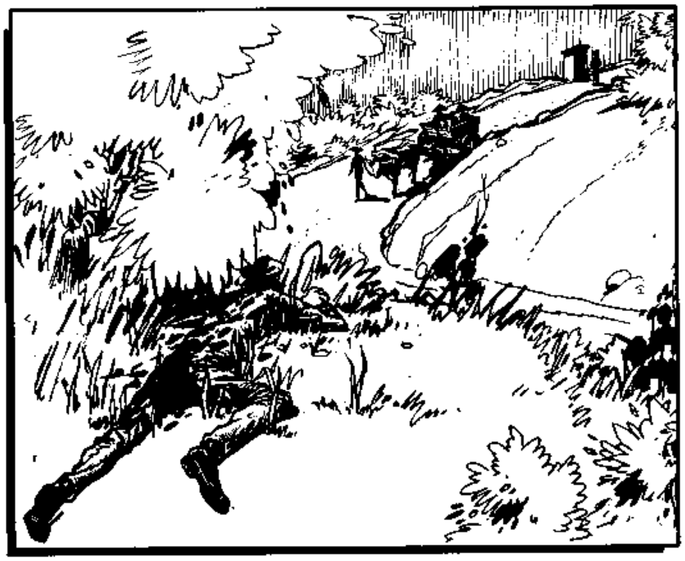

Figure 4-7. Frustration.

should avoid working to complete exhaustion, in emergencies certain tasks must be done in spite of fatigue.

(1) Rest is a basic factor for recovery from fatigue and is also important in resisting further fatigue. It is essential that the rest (following fatiguing effort) be sufficient to permit complete recovery; otherwise, the residual fatigue will accumulate and require longer periods of rest to recover from subsequent effort. During the early stages of fatigue proper rest provides a rapid recovery. This is true of muscular fatigue as well as mental fatigue. Sleep is the most complete form of rest available and is basic to recovery from fatigue.

Figure 4-8. Fatigue.

(2) Short rest breaks during extended stress periods can improve total output. There are five ways in which rest breaks are beneficial:

(a) They provide opportunities for partial recovery from fatigue.

(b) They help reduce energy expenditure.

(c) They increase efficiency by enabling a person to take maximum advantage of planned rest.

(d) They relieve boredom by breaking up the uniformity and monotony of the task.

(e) They increase morale and motivation.

(3) Survivors should rest before output shows a definite decline. If rest breaks are longer, fewer may be required. When efforts are highly strenuous or monotonous, rest breaks should be more frequent. Rest breaks providing relaxation are the most effective. In mental work, mild exercise may be more relaxing. When work is monotonous, changes of activity, conversation, and humor are effective relaxants. In deciding on the amount and frequency of rest periods, the loss of efficiency resulting from longer hours of effort must be weighed against the absolute requirements of the survival situation.

(4) Fatigue can be reduced by working "smarter." A survivor can do this in two practical ways:

(a) Adjust the pace of the effort. Balance the load, the rate, and the time period. For example, walking at a normal rate is a more economical effort than fast walking.

(b) Adjust the technique of work. The way in which work is done has a great bearing on reducing fatigue. Economy of effort is most important. Rhythmic movements suited to the task are best.

(5) Mutual group support, cooperation, and competent leadership are important factors in maintaining group morale and efficiency, thereby reducing stress and fatigue. A survivor usually feels tired and weary before the physiological limit is reached. In addition, other stresses experienced at the same time; such as cold, hunger, fear, or despair, can intensify fatigue. The feeling of fatigue involves not only the physical reaction to effort, but also subtle changes in attitudes and motivation. Remember, a person has reserves of energy to cope with an important emergency even when feeling very tired.

b. As in the case of other stresses, even a moderate amount of fatigue reduces efficiency. To control fatigue, it is wise to observe a program of periodic rest. Because the main objective—to establish contact with friendly forces—survivors may overestimate their strength and risk exhaustion. On the other hand, neither an isolated individual nor a group leader should underestimate the capacity of the individual or the group on the basis of fatigue. The only sound basis for judgment must be gained from training and past experience. In training, a person should form an opinion of individual capacity based on actual experience. Likewise, a group leader must form an opinion of the capacities of fellow aircrew members. This group didn't think:

> "By nightfall, we were completely bushed ... We decided to wrap ourselves in the 'chute instead of making a shelter. We were too tired even to build a fire. We just cut some pine boughs, rolled ourselves in the nylon and went to sleep...and so, of course, it rained, and not lightly. We stood it until we were soaked, and then we struggled out and made a shelter. Since it was pitch dark, we didn't get the sags out of the canopy, so the water didn't all run off. Just a hell of a lot of it came through. Our hip and leg joints ached as though we had acute rheumatism. Being wet and cold accentuated the pain. We changed positions every 10 minutes, after gritting our teeth to stay put that long."

4-9. Sleep Deprivation. The effects of sleep loss are closely related to those of fatigue. Sleeping at unaccustomed times, sleeping under strange circumstances (in a strange place, in noise, in light, or in other distractions), or missing part or all of the accustomed amount of sleep will cause a person to react with feelings of weariness, irritability, emotion tension, and some loss of efficiency. The extent of an individual's reaction depends on the amount of disturbance and on other stress factors which may be present at the same time (figure 4-9).

a. Strong motivation is one of the principal factors in helping to compensate for the impairing effects of sleep loss. Superior physic and mental conditioning, opportunities to rest, food and water, and companions help in enduring sleep deprivation. If a person is in reasonably good physical and mental condition, sleep deprivation can be endured 5 days or more without damage, although efficiency during the latter stages may be poor. A person must learn to get as much sleep and rest as possible. Restorative effects of sleep are felt even after "catnaps." In some instances, survivors may need to stay awake. Activity, movement, conversation, eating, and drinking are some of the ways a person can stimulate the body to stay awake.

b. When one is deprived of sleep, sleepiness usually comes in waves. A person may suddenly be sleepy immediately after a period of feeling wide awake. If this can be controlled, the feeling will soon pass and the person will be wide awake again until the next wave appears. As the duration of sleep deprivation increases, these periods between waves of sleepiness become shorter. The need to sleep may be so strong in some people after a long period of deprivation that they become desperate and do careless or dangerous things in order to escape this stress.

4-10. Isolation. Loneliness, helplessness, and despair which are experienced by survivors when they are isolated are among the most severe survival stresses. People often take their associations with family, friends, military collegues, and others for granted. But survivors soon begin to miss the daily interaction with other people. However, these, like the other stresses already discussed, can be conquered. Isolation can be controlled and overcome by knowledge, understanding, deliberate countermeasures, and a determined will to resist it (figure 4-10).

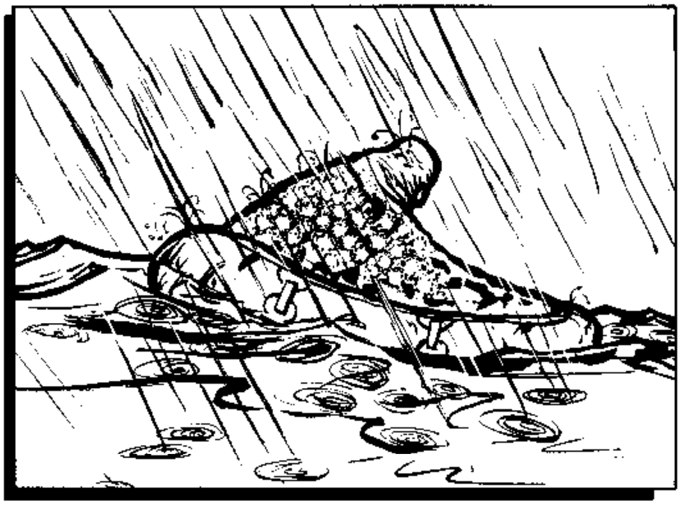

Figure 4-9. Sleep Deprivation.

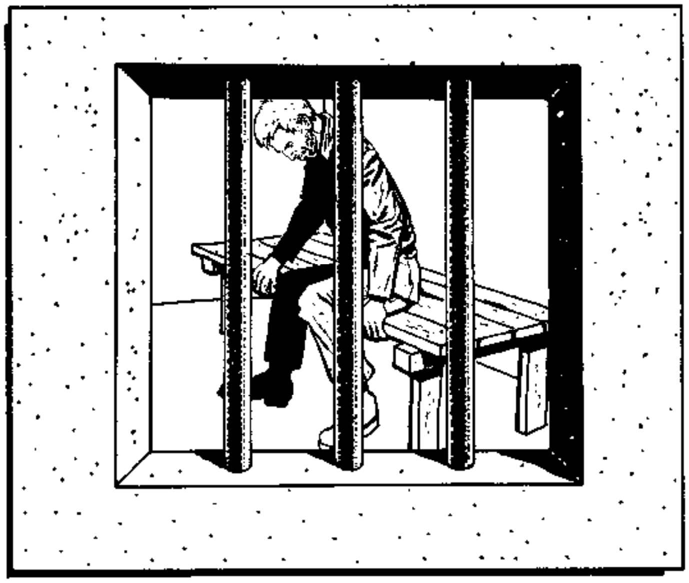

Figure 4-10. Isolation.

4-11. Insecurity. Insecurity is the survivor's feeling of helplessness or inadequacy resulting from varied stresses and anxieties. These anxieties may be caused by uncertainty regarding individual goals, abilities, and the future in a survival situation. Feelings of insecurity may

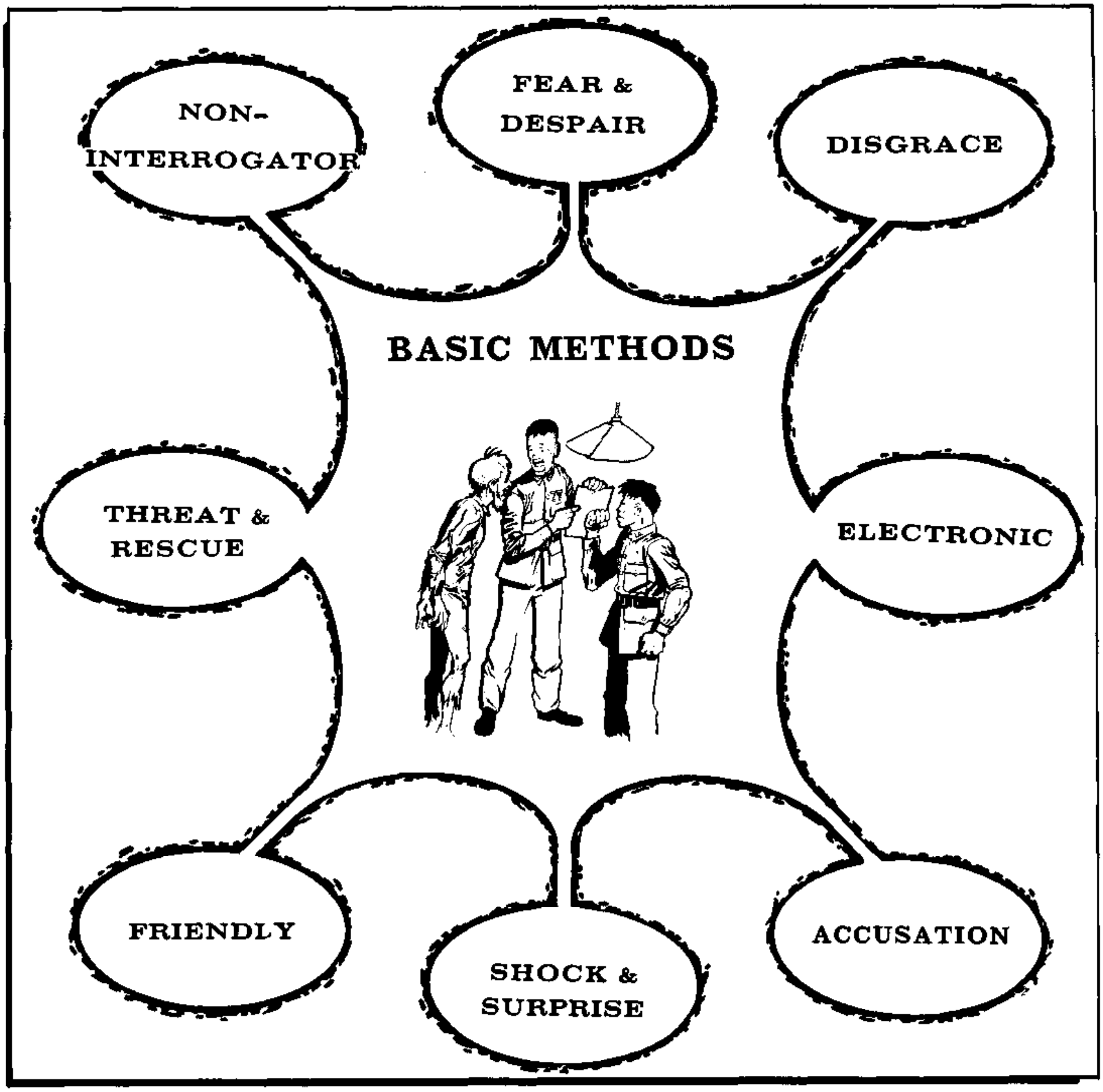
NON-INTERROGATOR
FEAR & DESPAIR
DISGRACE
BASIC METHODS
THREAT & RESCUE
ELECTRONIC
FRIENDLY
SHOCK & SURPRISE
ACCUSATION

Figure 4-11. Loss of Self-Esteem.

have widely different effects on the survivor's behavior. A survivor should establish challenging but attainable goals. The better a survivor feels about individual abilities to achieve goals and adequately meet personal needs, the less insecure the survivor will feel.

4-12. Loss of Self-Esteem. Self-esteem is the state or quality of having personal self-respect and pride. Lack of (or loss of) self-esteem in a survivor may bring on depression and a change in perspective and goals. A loss of self-esteem may occur in individuals in captivity. Humiliation and other factors brought on by the captor may cause them to doubt their own worth. Humiliation comes from the feeling of losing pride or self-respect by being disgraced or dishonored, and is associated with the loss of self-esteem. Prisoners must maintain their pride and not become ashamed either because they are PWs or because of the things that happen to them as a result of being a PW. The survivor who "loses face" (both personally and with the enemy) becomes more vulnerable to captor exploitation attempts. To solve this problem, survivors should try to maintain proper perspective about both the situation and themselves. Their feelings of self-worth may be bolstered if they recall the implied commitment in the Code of Conduct—PWs will not be forgotten (figure 4-11).

4-13. Loss of Self-Determination. A self-determined person is relatively free from external controls or influences over his or her actions. In everyday society, these "controls and influences" are the laws and customs of our society and of the self-imposed elements of our personalities. In a survival situation, the "controls and influences" can be very different. Survivors may feel as if events, circumstances, and (in some cases) other people, are in control of the situation. Some factors which may cause individuals to feel they have lost the power of self-determination are a harsh captor, captivity, bad weather, or rescue forces that make time or movement demands. This lack of self-determination is more perceived than actual. Survivors must decide how unpleasant factors will be allowed to affect their mental state. They must have the self-confidence, fostered by experience and training, to live with their feelings and decisions, and to accept responsibility for both the way they feel and how they let those feelings affect them.

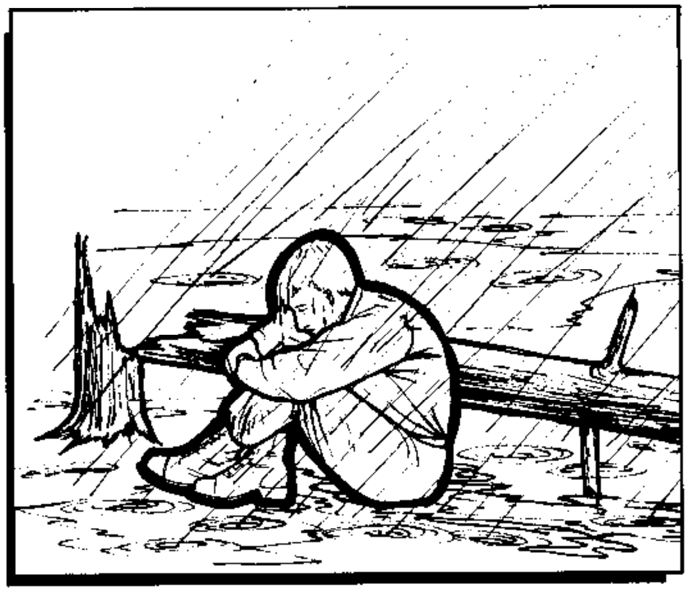

Figure 4-12. Depression.

4-14. Depression. As a survivor, depression is the biggest psychological problem that has to be conquered. It should be acknowledged that everyone has mental "highs" as well as mental "lows." People experiencing long periods of sadness or other negative feelings are suffering from depression. A normal mood associated with the sadness, grief, disappointment, or loneliness that everyone experiences at times is also described as depression. Most of the emotional changes in mood are temporary and do not become chronic. Depressed survivors may feel fearful, guilty, or helpless. They may lose interest in the basic needs of life. Many cases of depression also involve pain, fatigue, loss of appetite, or other physical ailments. Some depressed survivors try to injure or kill themselves (figure 4-12).

a. Psychiatrists have several theories as to the cause of depression. Some feel a person who, in everyday life and under normal conditions, experiences many periods of depression would probably have a difficult time in a survival situation. The main reason depression is a most difficult problem is that it can affect a wide range of psychological responses. The factors can become mutually reinforcing. For example, fatigue may lead to a feeling of depression. Depression may increase the feeling of fatigue, and this, in turn, leads to deeper depression and so on.

b. Depression usually begins after a survivor has met the basic needs for sustaining life, such as water, shelter, and food. Once the survivor's basic needs are met, there is often too much time for that person to dwell on the past, the present predicament, and on future problems. The survivor must be aware of the necessity to keep the mind and body active to eliminate the feeling of depression. One way to keep busy (daily) is by checking and improving shelters, signals, and food supply.

Chapter 5

EMOTIONAL REACTIONS

5-1. Introduction. Survivors may depend more upon their emotional reactions to a situation than upon calm, careful analysis of potential danger—the enemy, the weather, the terrain, the nature of the in-flight emergency, etc. Whether they will panic from fear, or use it as a stimulant for greater sharpness, is more dependent on the survivor's reactions to the situation than on the situation itself. Although there are many reactions to stress, the following are the most common and will be discussed in detail: fear, anxiety, panic, hate, resentment, anger, impatience, dependency, loneliness, boredom, and hopelessness.

5-2. Fear. Fear can SAVE A LIFE—or it can COST ONE. Some people are at their best when they are scared. Many downed fliers faced with survival emergencies have been surprised at how well they remembered their training, how quickly they could think and react, and what strength they had. The experience gave them a new confidence in themselves. On the other hand, some people become paralyzed when faced with the simplest survival situation. Some of them have been able "to snap themselves out of it" before it was too late. In other cases, a fellow aircrew member was on hand to assist them. However, others have not been so fortunate. They are not listed among the survivors (figure 5-1).

Figure 5-1. Fear.

a. How a person will react to fear depends more upon the individual than it does upon the situation. This has been demonstrated both in actual survival situations and in laboratory experiments. It isn't always the physically strong or the happy-go-lucky people who handle fear most effectively. Timid and anxious people have met emergencies with remarkable coolness and strength.

b. Anyone who faces life-threatening emergencies experiences fear. Fear is conscious when it results from a recognized situation (such as an immediate prospect of bailout) or when experienced as apprehension of impending disaster. Fear also occurs at a subconscious level and creates feelings of uneasiness, general discomfort, worry, or depression. Fear may vary widely in intensity, duration, and frequency of occurrence, and affect behavior across the spectrum from mild uneasiness to complete disorganization and panic. People have many fears; some are learned through personal experiences, and others are deliberately taught to them. Fear in children is directed through negative learning, as they are taught to be afraid of the dark, of animals, of noise, or of teachers. These fears may control behavior, and a survivor may react to feelings and imagination rather than to the problem causing fear.

c. When fantasy distorts a moderate danger into a major catastrophe, or vice versa, behavior can become abnormal. There is a general tendency to underestimate and this leads to reckless, foolhardy behavior. The principal means of fighting fear (in this case) is to pretend that it does not exist. There are no sharp lines between recklessness and bravery. It is necessary to check behavior constantly to maintain proper control.

d. One or more of the following signs or symptoms may occur in those who are afraid. However, they may also appear in circumstances other than fear.

(1) Quickening of pulse; trembling.

(2) Dilation of pupils.

(3) Increased muscular tension and fatigue.

(4) Perspiration of palms of hands, soles of feet, and armpits.

(5) Dryness of mouth and throat; higher pitch of voice; stammering.

(6) Feeling of "butterflies in the stomach," emptiness of the stomach, faintness, and nausea.

e. Accompanying these physical symptoms are the following common psychological symptoms:

(1) Irritability; increased hostility.

(2) Talkativeness in early stages, leading finally to speechlessness.

(3) Confusion, forgetfulness, and inability to concentrate.

(4) Feelings of unreality, flight, panic, or stupor.

f. Throughout military history, many people have coped successfully with the most strenuous odds. In adapting to fear, they have found support in previous training and experience. There is no limit to human control of fear. Survivors must take action to control fear. They cannot run away from fear. Appropriate actions should be to:

(1) Understand fear.

(2) Admit that it exists.

(3) Accept fear as reality.

g. Training can help survivors recognize what individual reactions may be. Using prior training, survivors should learn to think, plan, and act logically, even when afraid.

h. To effectively cope with fear, a survivor must:

(1) Develop confidence. Use training opportunities; increase capabilities by keeping physically and mentally fit; know what equipment is available and how to use it; learn as much as possible about all aspects of survival.

(2) Be prepared. Accept the possibility that "it can happen to me." Be properly equipped and clothed at all times; have a plan ready. Hope for the best, but be prepared to cope with the worst.

(3) Keep informed. Listen carefully and pay attention to all briefings. Know when danger threatens and be prepared if it comes; increase knowledge of survival environments to reduce the "unknown."

(4) Keep busy at all times. Prevent hunger, thirst, fatigue, idleness, and ignorance about the situation, since these increase fear.

(5) Know how fellow crewmembers react to fear. Learn to work together in emergencies—to live, work, plan, and help each other as a team.

(6) Practice religion. Don't be ashamed of having spiritual faith.

(7) Cultivate "good" survival attitudes. Keep the mind on a main goal and keep everything else in perspective. Learn to tolerate discomfort. Don't exert energy to satisfy minor desires which may conflict with the overall goal—to survive.

(8) Cultivate mutual support. The greatest support under severe stress may come from a tightly knit group. Teamwork reduces fear while making the efforts of every person more effective.

(9) Exercise leadership. The most important test of leadership and perhaps its greatest value lies in the stress situation.

(10) Practice discipline. Attitudes and habits of discipline developed in training carry over into other situations. A disciplined group has a better chance of survival than an undisciplined group.

(11) Lead by example. Calm behavior and demonstration of control are contagious. Both reduce fear and inspire courage.

i. Every person has goals and desires. The greatest values exercise the greatest influence. Because of strong religious, moral, or patriotic values, people have been known to face torture and death calmly rather than reveal information or compromise a principle. Fear can kill or it can save lives. It is a normal reaction to danger. By understanding and controlling fear through training, knowledge, and effective group action, fear can be overcome.

5-3. Anxiety:

a. Anxiety is a universal human reaction. Its presence can be felt when changes occur which affect an individual's safety, plans, or methods of living. It is generally felt when individuals perceive something bad is about to happen. A common description of anxiety is "butterflies in the stomach." Anxiety creates feelings of uneasiness, general discomfort, worry, or depression. Anxiety and fear differ mainly in intensity. Anxiety is a milder reaction and the specific cause(s) may not be readily apparent, whereas fear is a strong reaction to a specific, known cause. Common characteristics of anxiety are: fear of the future, indecision, feelings of helplessness, resentment (figure 5-2).

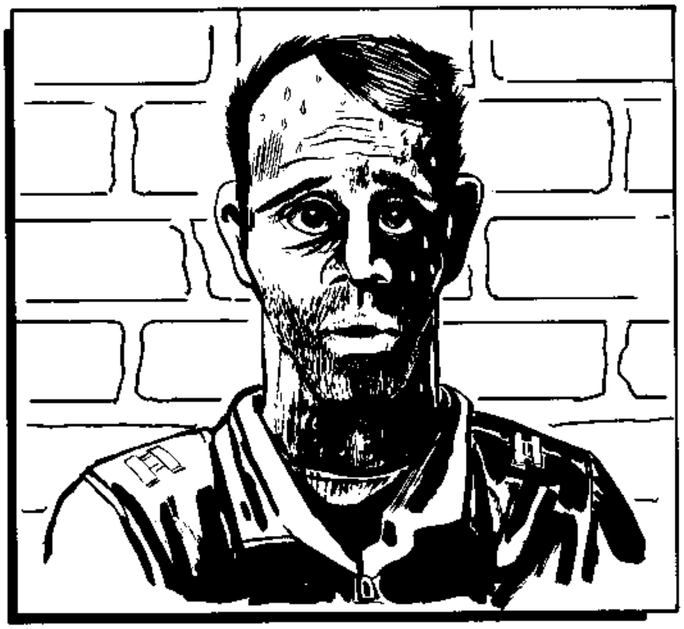

Figure 5-2. Anxiety.

b. To overcome anxiety, the individual must take positive action by adopting a simple plan. It is essential to keep your mind off of your injuries and do something constructive. For example, one PW began to try and teach English to the Chinese and to learn Chinese from them.

5-4. Panic. In the face of danger, a person may panic or "freeze" and cease to function in an organized manner. A person experiencing panic may have no conscious control over individual actions. Uncontrollable, irrational behavior is common in emergency situations. Anybody can panic, but some people go to pieces more easily than others. Panic is brought on by a sudden overwhelming fear, and can often spread quickly through a group of people. Every effort must be made to bolster morale and calm the panic with leadership and discipline. Panic has the same signs as fear and should be controlled in the same manner as fear. This survivor allowed pain to panic him.

> "His parachute caught in the tree, and he found himself suspended about five feet above the ground...one leg strap was released while he balanced in this aerial position and he immediately slipped toward the ground. In doing so, his left leg caught in the webbing and he was suspended by one leg with his head down. Unfortunately, the pilot's head touched an ant hill and biting ants immediately swarmed over him. Apparently, in desperation, the flier pulled his gun and fired five rounds into the webbing holding his foot. When he did not succeed in breaking the harness by shooting at it, he placed the last shot in his head and thus took his own life. It was obvious from the discoverer's report that if the pilot had even tried to turn around or to swing himself from his inverted position, he could have reached either the aerial roots or the latticed trunk of the tree. With these branches, he should have been able to pull himself from the harness...The fact that his head was in a nest of stinging ants only added to his panic, which led to the action that took his life." (Figure 5-3)

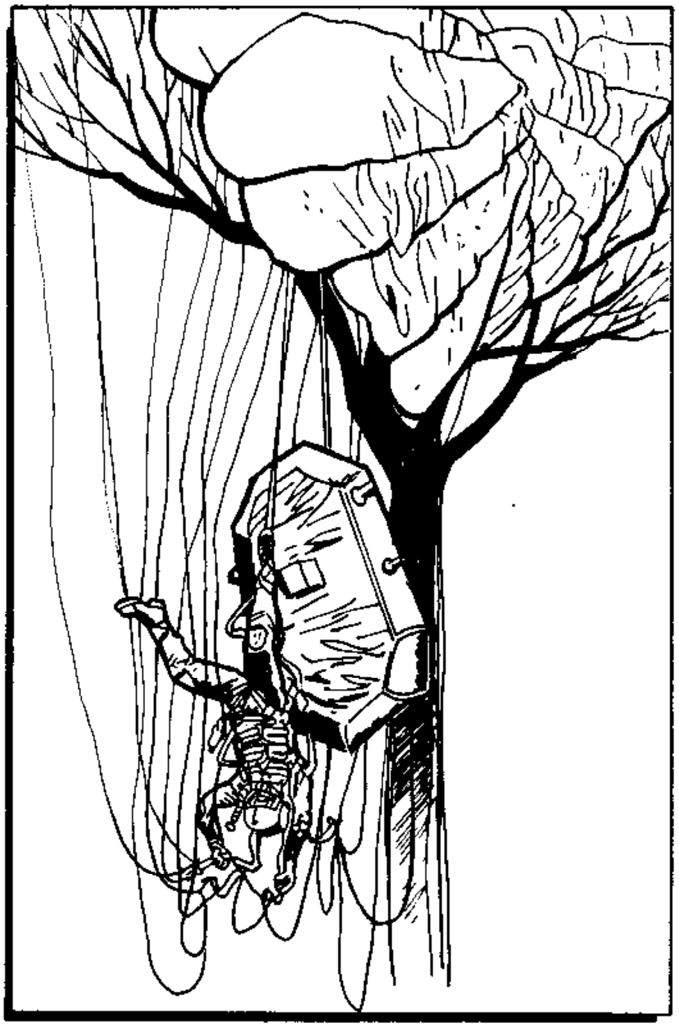

Figure 5-3. Panic.

5-5. Hate. Hate—feelings of intense dislike, extreme aversion, or hostility—is a powerful emotion which can have both positive and negative effects on a survivor. An understanding of the emotion and its causes is the key to learning to control it. Hate is an acquired emotion rooted in a person's knowledge or perceptions. The accuracy or inaccuracy of the information is irrelevant to learning to hate.

a. Any person, any object, or anything that may be understood intellectually, such as political concepts or religious dogma, can promote feelings of hate. Feelings of hate (usually accompanied with a desire for vengence, revenge, or retribution) have sustained former prisoners of war through their harsh ordeals. If an individual loses perspective while under the influence of hate and reacts emotionally, rational solutions to problems may be overlooked, and the survivor may be endangered.

b. To effectively deal with this emotional reaction, the survivor must first examine the reasons why the feeling of hate is present. Once that has been determined, survivors should then decide what to do about those feelings. Whatever approach is selected, it should be as constructive as possible. Survivors must not allow hate to control them.

5-6. Resentment. Resentment is the experiencing of an emotional state of displeasure or indignation toward some act, remark, or person that has been regarded as causing personal insult or injury. Luck and fate may play a role in any survival situation. A hapless survivor may feel jealous resentment toward a fellow PW, travel partner, etc., if that other person is perceived to be enjoying a success or advantage not presently experienced by the observer. The survivor must understand that events cannot always go as expected. It is detrimental to morale and could affect survival chances if feelings of resentment over another's attainments become too strong. Imagined slights or insults are common. The survivor should try to maintain a sense of humor and perspective about ongoing events and realize that stress and lack of self-confidence play roles in bringing on feelings of resentment.

5-7. Anger. Anger is a strong feeling of displeasure and belligerence aroused by a real or supposed wrong. People become angry when they cannot fulfill a basic need or desire which seems important to them. When anger is not relieved, it may turn into a more enduring attitude of hostility, characterized by a desire to hurt or destroy the person or thing causing the frustration. When anger is intense, the survivor loses control over the situation, resulting in impulsive behavior which may be destructive in nature. Anger is a normal response which can serve a useful purpose when carefully controlled. If the situation warrants and there is no threat to survival, one could yell or scream, take a walk, do some vigorous exercise, or just get away from the source of the anger, even if only for a few minutes. Here is a man who couldn't hold it.

> "I tried patiently to operate it (radio) in every way I had been shown. Growing more angry and disappointed at its failure, I tore the aerial off, threw the cord away, beat the battery on the rocks, then threw the pieces all over the hillside. I was sure disappointed." (See figure 5-4.)

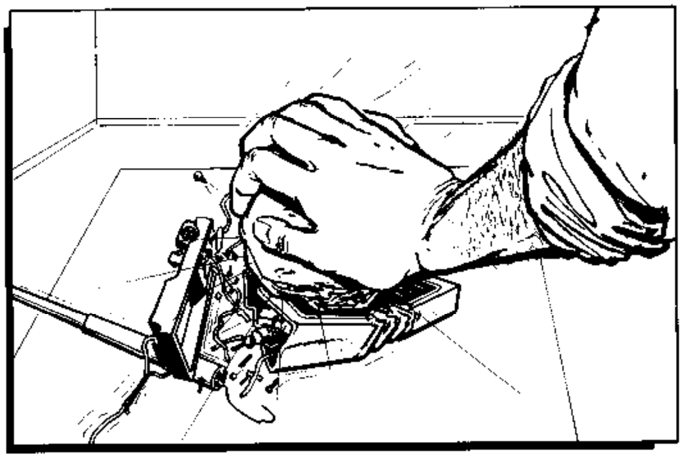

Figure 5-4. Anger.

5-8. Impatience:

a. The psychological stresses brought about by feelings of impatience can quickly manifest themselves in physical ways. Internally, the effects of impatience can cause changes in physical and mental well-being. Survivors who allow impatience to control their behavior may find that their efforts prove to be counterproductive and possibly dangerous. For example, evaders who don't have the ability or willingness to suppress annoyance when confronted with delay may expose themselves to capture or injury.

b. Potential survivors must understand they have to bear pain, misfortune, and annoyance without complaint. In the past, many survivors have displayed tremendous endurance, both mental and physical, in times of distress or misfortune. While not every survivor will be able to display such strength of character in all situations, each person should learn to recognize the things which may make them impatient to avoid acting unwisely. This survivor couldn't wait:

> "I became very impatient. I had planned to wait until night to travel but I just couldn't wait. I left the ditch about noon and walked for about two hours until I was caught."

5-9. Dependence. The captivity environment is the prime area where a survivor may experience feelings of dependency. The captor will try to develop in prisoners feelings of need, support, and trust for the captor. By regulating the availability of basic needs like food, water, clothing, social contact, and medical care, captors show their power and control over the prisoners' fate. Through emphasis on the prisoners' inability to meet their own basic needs, captors seek to establish strong feelings of prisoner dependency. This dependency can make prisoners extremely vulnerable to captor exploitation—a major captor objective. PW recognition of this captor tactic is key to countering it. Survivors must understand that, despite captor controls, they do control their own lives. Meeting even one physical or mental need can provide a PW with a "victory" and provide the foundation for continued resistance against exploitation (figure 5-5).

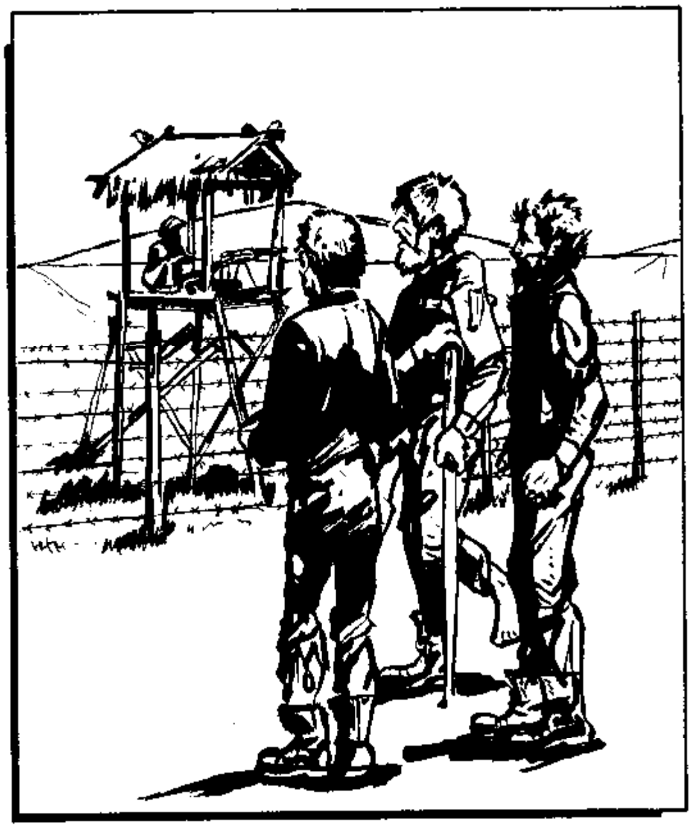

Figure 5-5. Dependence.

5-10. Loneliness. Loneliness can be very debilitating during a survival episode. Some people learn to control and manipulate their environment and become more self-sufficient while adapting to changes. Others rely on protective persons, routines, and familiarity of surroundings to function and obtain satisfaction (figure 5-6).

a. The ability to combat feelings of loneliness during a survival episode must be developed long before the episode occurs. Self-confidence and self-sufficiency are key factors in coping with loneliness. People develop these attributes by developing and demonstrating competence in performing tasks. As the degree of competence increases, so does self-confidence and self-sufficiency. Military training, more specifically survival training, is designed to provide individuals with the competence and self-sufficiency to cope with and adapt to survival living.

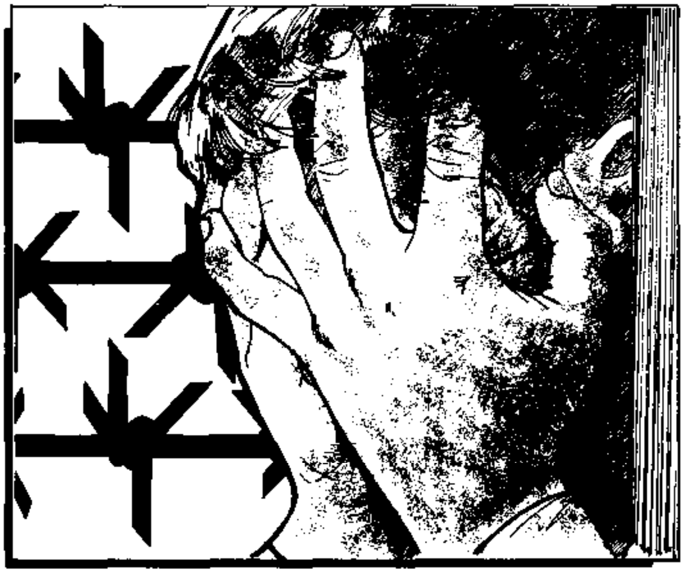

Figure 5-6. Loneliness.

b. In a survival situation, the countermeasure to conquer loneliness is to be active, to plan and think purposely. Development of self-sufficiency is the primary protection since all countermeasures in survival require the survivor to have the ability to practice self-control.

5-11. Boredom. Boredom and fatigue are related and frequently confused. Boredom is accompanied by a lack of interest and may include feelings of strain, anxiety, or depression, particularly when no relief is in sight and the person is frustrated. Relief from boredom must be based on correction of the two basic sources, repetitiveness and uniformity. Boredom can be relieved by a variation of methods—rotation of duties, broadening the scope of a particular task or job, taking rest breaks, or other techniques of diversification which may actually interfere with efficient performance of the job. The ungratifying nature of a task can be counteracted by clearing up its meaning, objectives, and, in some cases, its relation to the total plan.

a. This survivor couldn't think of anything to do:

> "The underground representative took me to a house to wait for another member of the underground to pick me up. This was the worst part of the whole experience—this waiting. I just sat in the house and waited for two weeks. I thought I would go mad." (See figure 5-7.)

b. This survivor invented something to do:

> "Not knowing what to do, I decided to kill all the bugs. There were a lot of spiders, the big ones that do not hurt humans, so I killed the flies and gave them to the spiders to eat."

Figure 5-7. Boredom.

5-12. Hopelessness. Hopelessness stems from negative feelings—regardless of actions taken, success is impossible, or the certainty that future events will turn out for the worst no matter what a person tries to do. Feelings of hopelessness can occur at virtually any time during a survival episode. Survivors have experienced loss of hope in trying to maintain health due to an inability to care for sickness, broken bones, or injuries; considering their chances of returning home alive; seeing their loved ones again; or believing in their physical or mental ability to deal with the situation; for example, evade long distances or not give information to an interrogator (figure 5-8).

Figure 5-8. Hopelessness.

a. During situations where physical exhaustion or exposure to the elements affects the mind, a person may begin to lose hope. The term "give-up-itis" was coined in Korea to describe the feeling of "hopelessness." During captivity, deaths occurred for no apparent cause. These individuals actually willed themselves to die or at least did not will themselves to live. The original premise (in the minds of such people) is that they are going to die. To them, the situation seemed totally futile and they had passively abandoned themselves to fate. It was possible to follow the process step by step. The people who died withdrew themselves from the group, became despondent, then lay down and gave up. In some cases, death followed rapidly.

b. One way to treat hopelessness is to eliminate the cause of the stress. Rest, comfort, and morale building activities can help eliminate this psychological problem. Another method used in Korea was to make the person so angry that the person wanted to get up and attack the tormentors. A positive attitude has a powerful influence on morale and combating the feeling of hopelessness.

c. Since many stress situations cannot be dealt with successfully by either withdrawal or direct attack, it may be necessary to work out a compromise solution. The action may entail changing a survivor's method of operation or accepting substitute goals.

d. Evaders faced with starvation may compromise with their conscience and steal "just this one time." They may ignore their food aversion and eat worms, bugs, or even human flesh. A related form of compromise is acceptance of substitute means to achieve the same goals.

5-13. Summary. All the psychological factors may be overcome by survivors if they can recognize the problem, work out alternative solutions, decide on an appropriate course of action, take action, and evaluate the results. Perhaps the most difficult step in this sequence is deciding on an appropriate course of action. Survivors may face either one or several psychological problems. These problems are quite dangerous and must be effectively controlled or countered for survival to continue.

Chapter 6

THE WILL TO SURVIVE

6-1. Introduction. The will to survive is defined as the desire to live despite seemingly insurmountable mental and(or) physical obstacles. The tools for survival are furnished by the military, the individual, and the environment. The training for survival comes from survival training publications, instruction, and the individual's own efforts. But tools and training are not enough without a *will to survive.* In fact, the records prove that "will" alone has been the deciding factor in many survival cases. While these accounts are not classic examples of "how to survive," they illustrate that a single-minded survivor with a powerful *will to survive* can overcome most hardships. There are cases where people have eaten their belts for nourishment, boiled water in their boots to drink as broth, or have eaten human flesh—though this certainly wasn't their cultural instinct.

a. One incident where the *will to survive* was the deciding factor between life and death involved a man stranded in the Arizona desert for 8 days without food and water. He traveled more than 150 miles during searing daytime temperatures, losing 25 percent of his body weight due to the lack of water (usually 10 percent loss causes death). His blood became so thick that the lacerations he received could not bleed until he had been rescued and received large quantities of water. When he started on that journey, something must have clicked in his mind telling him to live, regardless of any obstacles which might confront him. And live he did—on guts and will alone! (See figure 6-1.)

b. Let's flip a coin and check the other side of "will." Our location is the Canadian wilderness. A pilot ran into engine trouble and chose to deadstick his plane onto a frozen lake rather than punch out. He did a beautiful job and slid to a stop in the middle of the lake. He left the aircraft and examined it for damage. After surveying the area, he noticed a wooded shoreline only 200 yards away where food and shelter could be provided—he decided to go there. Approximately halfway there, he changed his mind and returned to the cockpit of his aircraft where he smoked a cigar, took out his pistol, and blew his brains out. Less than 24 hours later, a rescue team found him. Why did he give up? Why was he unable to survive? Why did he take his own life? On the other hand, why do people eat their belts or drink broth from their boots? No one really knows, but it's all related to the *will to survive.*

6-2. Overcoming Stress. The ability of the mind to overcome stress and hardship becomes most apparent when there appears to be little chance of a person surviving. When there appears to be no escape from the situation, the "will" enables a person to begin to win "the battle of the mind." This mental attitude can bridge the gap between the crisis period and the coping period.

Figure 6-1. Will to Survive.

6-3. Crisis Period:

a. The crisis period is the point at which the person realizes the gravity of the situation and understands that the problem will not go away. At this stage, action is needed. Most people will experience shock in this stage as a result of not being ready to face this new challenge. Most will recover control of their faculties, especially if they have been prepared through knowledge and training.

b. Shock during a crisis is normally a response to being overcome with anxiety. Thinking will be disorganized. At this stage, direction will be required because the individual is being controlled by the environment. The person's center of control is external. In a group survival episode, a natural leader may appear who will direct and reassure the others. But if the situation continues to control the individual or the group, the response may be panic, behavior may be irrational, and judgment is impaired. In a lone-survivor episode, the individual must gain control of the situation and respond constructively. In either case, survivors must evaluate the situation and develop a plan of action.

During the evaluation, the survivor must determine the most critical needs to improve the chance of living and being rescued.

6-4. The Coping Period. The coping period begins after the survivor recognizes the gravity of the situation and resolves to endure it rather than succumb. The survivor must tolerate the effects of physical and emotional stresses. These stresses can cause anxiety which becomes the greatest obstacle to self-control and solving problems. Coping with the situation requires considerable internal control. For example, the survivor must often subdue urgent desires to travel when that would be counterproductive and dangerous. A person must have patience to sit in an emergency action shelter while confronted with an empty stomach, aching muscles, numb toes, and suppressed feelings of depression and hopelessness. Those who fail to think constructively may panic. This could begin a series of mistakes which result in further exhaustion, injury, and sometimes death. Death comes not from hunger pains but from the inability to manage or control emotions and thought processes.

6-5. Attitude. The survivor's attitude is the most important element of the *will to survive.* With the proper attitude, almost anything is possible. The desire to live is sometimes based on the feelings toward another person and(or) thing. Love and hatred are two emotional extremes which have moved people to do exceptional things physically and mentally. The lack of a *will to survive* can sometimes be identified by the individual's motivation to meet essential survival needs, emotional control resulting in reckless, paniclike behavior, and self-esteem.

a. It is essential to strengthen the *will to survive* during an emergency. The first step is to avoid a tendency to panic or "fly off the handle." Sit down, relax, and analyze the situation rationally. Once thoughts are collected and thinking is clear, the next step is to make decisions. In normal living, people can avoid decisions and let others do their planning. But in a survival situation, this will seldom work. Failure to decide on a course of action is actually a decision for inaction. This lack of decisionmaking may even result in death. However, decisiveness must be tempered with flexibility and planning for unforeseen circumstances. As an example, an aircrew member down in an arctic nontactical situation decides to construct a shelter for protection from the elements. The planning and actions must allow sufficient flexibility so the aircrew can monitor the area for indications of rescuers and be prepared to make contact—visually, electronically, etc.—with potential rescuers.

b. Tolerance is the next topic of concern. A survivor or evader will have to deal with many physical and psychological discomforts, such as unfamiliar animals, insects, loneliness, and depression. Aircrew members are trained to tolerate uncomfortable situations. That training must be applied to deal with the stress of environments.

c. Survivors in both tactical and nontactical situations must face and overcome fears to strengthen the *will to survive.* These fears may be founded or unfounded, be generated by the survivor's uncertainty or lack of confidence, or be based on the proximity of enemy forces. Indeed, fear may be caused by a wide variety of real and imagined dangers. Despite the source of the fear, survivors must recognize fear and make a conscious effort to overcome it.

6-6. Optimism. One of a survivor's key assets is optimism—hope and faith. Survivors must maintain a positive, optimistic outlook on their circumstance and how well they are doing. Prayer or meditation can be helpful. How a survivor maintains optimism is not so important as its use.

6-7. Summary. Survivors do not choose or welcome their fate and would escape it if they could. They are trapped in a world of seemingly total domination—a world hostile to life and any sign of dignity or resistance. The survival mission is not an easy one, but it is one in which success can be achieved. This has been an introduction to the concepts and ideas that can help an aircrew member return. Having the *will to survive* is what it's all about!

Part Three

BASIC SURVIVAL MEDICINE

Chapter 7

SURVIVAL MEDICINE

7-1. Introduction:

a. Foremost, among the many things which can compromise a survivor's ability to return are medical problems encountered during ejection, parachute descent, and(or) parachute landing. In the Southeast Asian conflict, some 30 percent of approximately 1,000 US Air Force survivors, including 322 returned PWs, were injured by the time they disentangled themselves from their parachutes. The most frequently reported injuries were fractures, strains, sprains, and dislocations, as well as burns and other types of wounds (figure 7-1).

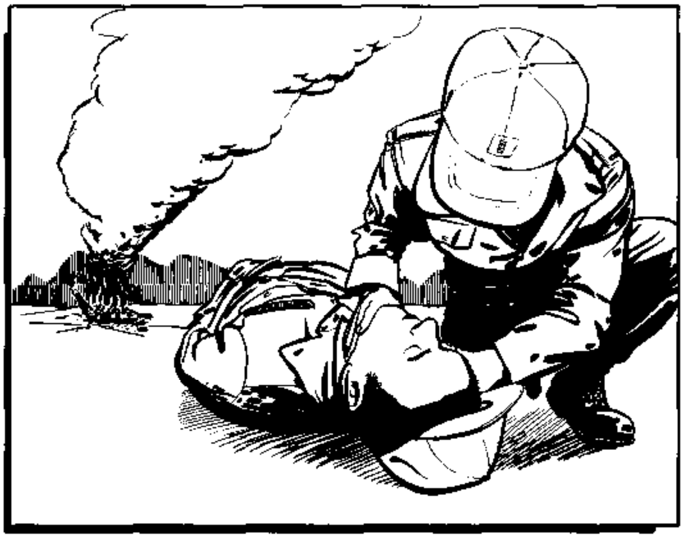

Figure 7-1. Survival Medicine.

b. Injuries and illnesses peculiar to certain environments can reduce survival expectancy. In cold climates, and often in an open sea survival situation, exposure to extreme cold can produce serious tissue trauma, such as frostbite, or death from hypothermia. Exposure to heat in warm climates, and in certain areas on the open seas, can produce heat cramps, heat exhaustion, or life-threatening heatstroke.

c. Illnesses contracted during evasion or in a captivity environment can interfere with successful survival. Among these are gastrointestinal disorders, respiratory diseases, skin infections and infestations, malaria, typhus, cholera, etc.

d. A review of the survival experiences from World War II, Korea, and Southeast Asia indicates that, while US military personnel generally knew how to administer first aid to others, there was a marked inability to administer self-aid. Further, only the most basic medical care had been taught to most military people. Lastly, it was repeatedly emphasized that even minor injuries or ailments, when ignored, became major problems in a survival situation. Thus, prompt attention to the most minor medical problem is essential in a survival episode. Applying principles of survival medicine should enable military members to maintain health and well-being in a hostile or nonhostile environment until rescued and returned to friendly control.

e. Information in this chapter and chapter 8 is a basic reference to self-aid techniques used by PWs in captivity and techniques found in folk medicine. The information describes procedures which can maintain health in medically austere situations. It includes items used to prevent and treat injuries and illnesses. Because there is no "typical" survival situation, the approach to self-aid must be flexible, placing emphasis on using what is available to treat the injury or illness. Further, survivors recognize that medical treatment offered by people of other cultures may be far different from our own. For example, in the rural areas of Vietnam, a poultice of python meat was and is used to treat internal lower back pain. Such treatment may be repugnant to some US military personnel; however, medical aid offered to survivors in non-US cultures may be the best available in the given circumstance.

f. The procedures in this chapter and chapter 8 must be viewed in the reality of a true survival situation. The results of treatment may be substandard compared with present medical standards. However, these procedures will not compromise professional medical care which becomes available following rescue. Moreover, in the context of a survival situation, they may represent the best available treatment to extend the individual's survival expectancy.

7-2. Procedures and Expedients. Survival medicine encompasses procedures and expedients that are:

a. Required and available for the preservation of health and the prevention, improvement, or treatment of injuries and illnesses encountered during survival.

b. Suitable for application by nonmedical personnel to themselves or comrades in the circumstances of the survival situation.

(1) Survival medicine is more than first aid in the conventional sense. It approaches final definitive treatment in that it is not dependent upon the availability of technical medical assistance within a reasonable period of time.

(2) To avoid duplication of information generally available, the basic principles of first aid will not be repeated, nor will the psychological factors affecting survival which were covered in part two.

7-3. Hygiene. In a survival situation, cleanliness is essential to prevent infection. Adequate personal cleanliness will not only protect against disease germs that are present in the individual's surroundings but will also protect the group by reducing the spread of these germs (figure 7-2).

Figure 7-2. Hygiene.

a. Washing, particularly the face, hands, and feet, reduces the chances of infection from small scratches and abrasions. A daily bath or shower with hot water and soap is ideal. If no tub or shower is available, the body should be cleaned with a cloth and soapy water, paying particular attention to the body creases (armpits, groin, etc.), face, ears, hands, and feet. After this type of "bath," the body should be rinsed thoroughly with clear water to remove all traces of soap which could cause irritation.

b. Soap, although an aid, is not essential to keeping clean. Ashes, sand, loamy soil, and other expedients may be used to clean the body and cooking utensils.

c. When water is in short supply, the survivor should take an "air bath." All clothing should be removed and the body simply exposed to the air. Exposure to sunshine is ideal, but even on an overcast day or indoors, a 2-hour exposure of the naked body to the air will refresh the body. Care should be taken to avoid sunburn when bathing in this manner. Exposure in the shade, shelter, sleeping bag, etc., will help if the weather conditions do not permit direct exposure.

d. Hair should be kept trimmed, preferably 2 inches or less in length, and the face should be clean-shaven. Hair provides a surface for the attachment of parasites and the growth of bacteria. Keeping the hair short and the face clean-shaven will provide less habitat for these organisms. At least once a week, the hair should be washed with soap and water. When water is in short supply, the hair should be combed or brushed thoroughly and covered to keep it clean. It should be inspected weekly for fleas, lice, and other parasites. When parasites are discovered, they should be removed.

e. The principal means of infecting food and open wounds is contact with unclean hands. Hands should be washed with soap and water, if available, after handling any material which is likely to carry germs. This is especially important after each visit to the latrine, when caring for the sick and injured, and before handling food, food utensils, or drinking water. The fingers should be kept out of the mouth and the fingernails kept closely trimmed and clean. A scratch from a long fingernail could develop into a serious infection.

7-4. Care of the Mouth and Teeth. Application of the following fundamentals of oral hygiene will prevent tooth decay and gum disease:

a. The mouth and teeth should be cleansed thoroughly with a toothbrush and dentifrice at least once each day. When a toothbrush is not available, a "chewing stick" can be fashioned from a twig. The twig is washed, then chewed on one end until it is frayed and brushlike. The teeth can then be brushed very thoroughly with the stick, taking care to clean all tooth surfaces. If necessary, a clean strip of cloth can be wrapped around the finger and rubbed on the teeth to wipe away food particles which have collected on them. When neither toothpaste nor toothpowder are available, salt, soap, or baking soda can be used as substitute dentifrices. Parachute inner core can be used by separating the filaments of the inner core and using this as a dental floss. Gargling with willow bark tea will help protect the teeth.

b. Food debris which has accumulated between the teeth should be removed by using dental floss or toothpicks. The latter can be fashioned from small twigs.

c. Gum tissues should be stimulated by rubbing them vigorously with a clean finger each day.

d. Use as much care cleaning dentures and other dental appliances, removable or fixed, as when cleaning natural teeth. Dentures and removable bridges should be removed and cleaned with a denture brush or "chew stick" at least once each day. The tissue under the dentures should be brushed or rubbed regularly for proper stimulation. Removable dental applicances should be removed at night or for a 2- to 3-hour period during the day.

7-5. Care of the Feet. Proper care of the feet is of utmost importance in a survival situation, especially if the survivor has to travel. Serious foot trouble can be prevented by observing the following simple rules:

a. The feet should be washed, dried thoroughly, and massaged each day. If water is in short supply, the feet should be "air cleaned" along with the rest of the body (figure 7-3).

b. Toenails should be trimmed straight across to prevent the development of ingrown toenails.

c. Boots should be broken in before wearing them on *any* mission. They should fit properly, neither so tight that they bind and cause pressure spots nor so loose that they permit the foot to slide forward and backward when walking. Insoles should be improvised to reduce any friction spots inside the shoes.

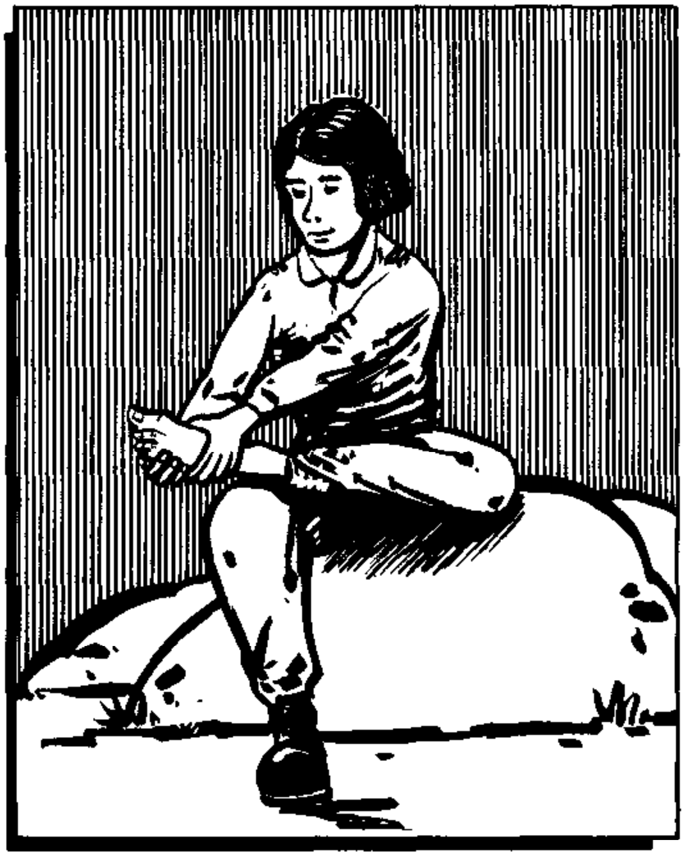

Figure 7-3. Care of Feet.

d. Socks should be large enough to allow the toes to move freely but not so loose that they wrinkle. Wool socks should be at least one size larger than cotton socks to allow for shrinkage. Socks with holes should be properly darned before they are worn. Wearing socks with holes or socks that are poorly repaired may cause blisters. Clots of wool on the inside and outside should be removed from wool socks because they may cause blisters. Socks should be changed and washed thoroughly with soap and water each day. Woolen socks should be washed in cool water to lessen shrinkage. In camp, freshly laundered socks should be stretched to facilitate drying by hanging in the sun or in an air current. While traveling, a damp pair of socks can be dried by placing them inside layers of clothing or hanging them on the outside of the pack. If socks become damp, they should be exchanged for dry ones at the first opportunity.

e. When traveling, the feet should be examined regularly to see if there are any red spots or blisters. If detected in the early stages of development, tender areas should be covered with adhesive tape to prevent blister formation.

7-6. Clothing and Bedding. Clothing and bedding become contaminated with any disease germs which may be present on the skin, in the stool, in the urine, or in secretions of the nose and throat. Therefore, keeping clothing and bedding as clean as possible will decrease the chances of skin infection and decrease the possibility of parasite infestation. Outer clothing should be washed with soap and water when it becomes soiled. Under clothing and socks should be changed daily. If water is in short supply, clothing should be "air cleaned." For air cleaning, the clothing is shaken out of doors, then aired and sunned for 2 hours. Clothing cleaned in this manner should be worn in rotation. Sleeping bags should be turned inside out, fluffed, and aired after each use. Bed linen should be changed at least once a week, and the blankets, pillows, and mattresses should be aired and sunned (figure 7-4).

7-7. Rest. Rest is necessary for the survivor because it not only restores physical and mental vigor, but also promotes healing during an illness or after an injury.

a. In the initial stage of the survival episode, rest is particularly important. After those tasks requiring immediate attention are done, the survivor should inventory available resources, decide upon a plan of action, and even have a meal. This "planning session" will provide a rest period without the survivor having a feeling of "doing nothing."

b. If possible, regular rest periods should be planned in each day's activities. The amount of time allotted for rest will depend on a number of factors, including the survivor's physical condition, the presence of hostile forces, etc., but usually, 10 minutes each hour is sufficient. During these rest periods, the survivor should change either from physical activity to complete rest or from mental activity to physical activity as the case may be. The survivor must learn to become comfortable and to rest under less than ideal conditions.

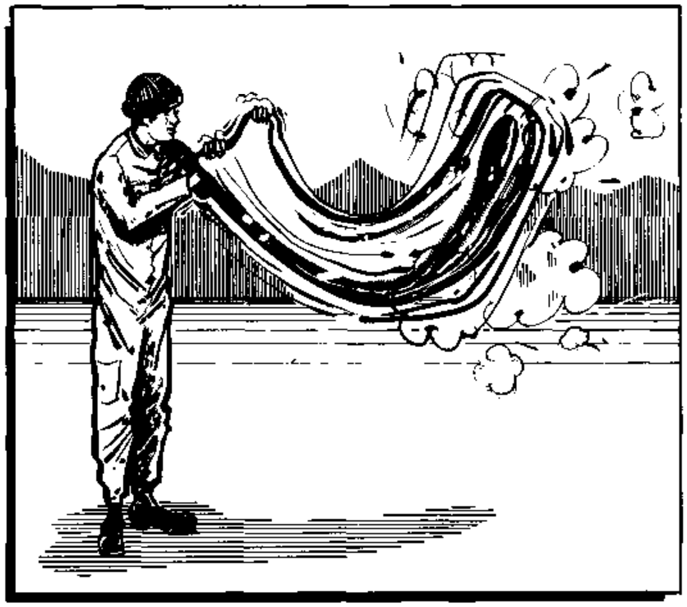

Figure 7-4. Bedding.

7-8. Rules for Avoiding Illness. In a survival situation, whether short-term or long-term, the dangers of disease are multiplied. Application of the following simple guidelines regarding personal hygiene will enable the survivor to safeguard personal health and the health of others:

a. ALL water obtained from natural sources should be purified before consumption.

b. The ground in the camp area should not be soiled with urine or feces. Latrines should be used, if available. When no latrines are available, individuals should dig "cat holes" and cover their waste.

c. Fingers and other contaminated objects should never be put into the mouth. Hands should be washed before handling any food or drinking water, before using the fingers in the care of the mouth and teeth, before and after caring for the sick and injured, and after handling any material likely to carry disease germs.

d. After each meal, all eating utensils should be cleaned and disinfected in boiling water.

e. The mouth and teeth should be cleansed thoroughly at least once each day. Most dental problems associated with long-term survival episodes can be prevented by using a toothbrush and toothpaste to remove accumulated food debris. If necessary, devices for cleaning the teeth should be improvised.

f. Bites and insects can be avoided by keeping the body clean, by wearing proper protective clothing, and by using head net, improvised bed nets, and insect repellants.

g. Wet clothing should be exchanged for dry clothing as soon as possible to avoid unnecessary body heat loss.

h. Personal items such as canteens, pipes, towels, toothbrushes, handkerchiefs, and shaving items should not be shared with others.

i. All food scraps, cans, and refuse should be removed from the camp area and buried.

j. If possible, a survivor should get 7 or 8 hours of sleep each night.

k. Aircrew members should keep all immunization "shots" current.

7-9. General Management of Injuries:

a. Bleeding. Control of bleeding is most important in survival situations where replacement transfusions are not possible. Immediate steps should be taken to stop the flow of blood, regardless of its source. The method used should be commensurate with the type and degree of bleeding. The tourniquet, when required and properly used, will save life. If improperly used, it may cost the life of the survivor. The basic characteristics of a tourniquet and the methods of its use are well covered in standard first aid texts; however, certain points merit emphasis in the survival situation. A tourniquet should be used only after every alternate method has been attempted. If unable to get to medical aid within 2 hours, after 20 minutes, gradually loosen the tourniquet. If bleeding has stopped, remove the tourniquet; if bleeding continues, reapply and leave in place. The tourniquet should be applied as near the site of the bleeding as possible, between the wound and the heart, to reduce the amount of tissue lost.

b. Pain:

(1) Control of Pain. The control of pain accompanying disease or injury under survival situations is both difficult and essential. In addition to its morale-breaking discomfort, pain contributes to shock and makes the survivor more vulnerable to enemy influences. Ideally, pain should be eliminated by the removal of the cause. However, this is not always immediately possible, hence measures for the control of pain are beneficial.

(2) Position, Heat, and Cold. The part of the body that is hurting should be put at rest, or at least its activity restricted as much as possible. The position selected should be the one giving the most comfort, and be the easiest to maintain. Splints and bandages may be necessary to maintain the immobilization. Elevation of the injured part, with immobilization, is particularly beneficial in the throbbing type pain such as is typical of the "mashed" finger. Open wounds should be cleansed, foreign bodies removed, and a clean dressing applied to protect the wound from the air and chance contacts with environmental objects. Generally, the application of warmth reduces pain—toothache, bursitis, etc. However, in some conditions, application of cold has the same effect—strains and sprains. Warmth or cold is best applied by using water due to its high specific heat, and the survivor can try both to determine which is most beneficial.

(3) Pain Killers. Drugs are very effective in reducing pain; however, they are not likely to be available in the survival situation. Hence, the importance of the above "natural" procedures. Aspirin, APCs, and such tablets are primarily intended to combat the discomforts of colds and upper respiratory diseases, and, at best, will just take the edge off severe pain. They should be taken, however, if available. If no aspirin is available, there are some parts of vegetation which can be used. For example, most of the willows have been used for their pain-relieving and fever-lowering properties for hundreds of years. The fresh bark contains salicin, which probably decomposes into salicylic acid in the human body. Wintergreen, also known as checkerberry, was used by some Indians for body aches and pains. The leaves are made into a tea. The boiled bark of the magnolia tree helps relieve internal pains and fever, and has been known to stop dysentery. To be really effective in control of pain, stronger narcotic drugs such as codeine and morphine are required. During active hostilities, morphine may be available in aircraft and individual first aid kits.

c. Shock:

(1) Circulatory Reaction. Shock in some degree accompanies all injuries to the body, and frequently it is the most serious consequence of the injury. In essence, shock is a circulatory reaction of the body (as a whole) to an injury (mechanical or emotional). While the changes to the circulatory system initially favor body resistance to the injury (by ensuring adequate blood supply to vital structures), they may progress to the point of circulatory failure and death. All aircrew members should be familiar with the signs and symptoms of shock so that the condition may be anticipated, recognized, and dealt with effectively. However, the best survival approach is to treat ALL moderate and severe injuries for shock. No harm will be done, and such treatment will speed recovery.

(2) Fluids. Normally, fluids administered by mouth are generally prohibited in the treatment of shock following severe injury. Such fluids are poorly absorbed when given by mouth, and they may interfere with later administration of anesthesia for surgery. In survival medicine, however, the situation is different in that the treatment being given is the final treatment. Survivors cannot be deprived of water for long periods just because they have been injured; in fact, their recovery depends upon adequate hydration. Small amounts of warm water, warm tea, or warm coffee given frequently early in shock are beneficial if the patient is conscious, can swallow, and has no internal injuries. In later shock, fluids by mouth are less effective as they are not absorbed from the intestines. Burns, particularly, require large amounts of water to replace fluid lost from injured areas. Alcohol should never be given to a person in shock or who may go into shock.

(3) Psychogenic Shock. Psychogenic shock is frequently noted during the period immediately following an emergency; for example, bailout. Psychogenic shock, which occurs even without injury, requires attention to limit it, both in degree and duration. The degree of this post-impact shock varies widely among individuals but its occurrence is almost universal. In reality, the survivor has passed through two major emergencies almost simultaneously; the aircraft incident leading to the survival situation, and the situation itself. Should the survivor be injured (and the majority of them are), a third emergency is added. It is not uncommon, then, that some psychogenic reaction with circulatory implications occurs. Resistance to this type of shock depends upon the individual's personality and the amount of training previously received. Treatment consists of stopping all activities (when possible), relaxing, evaluating the situation, and formulating a plan of action before the survival situation begins.

d. Fractures:

(1) Proper immobilization of fractures, dislocations, and sprains is even more important in survival medicine than in conventional first aid. Rather than merely making the patient comfortable during transport to eventual treatment, in survival medicine, the initial immobilization is part of the ultimate treatment. Immobilizing body parts to help control pain was discussed earlier. In addition, immobilization in proper position hastens healing of fractures and improves the ultimate functional result. In the survival situation, the immobilization must suffice for a relatively long period of time and permit the patient to maintain a fairly high degree of mobility. Materials for splinting and bandaging are available in most survival situations, and proper techniques are detailed in most first aid manuals.

(2) The reduction of fractures is normally beyond the scope of first aid; however, in the prolonged survival situation, the correction of bone deformities is necessary to hasten healing and obtain the greatest functional result. The best time for manipulation of a fracture is in the period immediately following the injury, before painful muscle spasms ensue. Traction is applied until overriding fragments of bone are brought into line, (check by the other limb) and the extremity is firmly immobilized. Frequently, it is advantageous to continue traction after reduction to ensure the proper alignment of the bones.

(3) As plaster casts are not available in the survival situation, improvising an immobilization device is necessary. This may be done by using several parallel, pliable willow branches, woven together with vines or parachute lines. Use care so that the extremity is not constricted when swelling follows the injury. In an escape and evasion situation, it may be necessary to preserve the mobility of the survivor after reduction of the fracture. This is difficult in fractures of the lower extremities, although tree limbs may be improvised as

crutches. With companions, the use of improvised litters may be possible.

(4) Reduction of dislocated joints is done similar to that of fractures. Gentle, but firm, traction is applied and the extremity is manipulated until it "snaps" back into place. If the survivor is alone, the problem is complicated but not impossible. Traction can still be applied by using gravity. The distal portion of the extremity is tied to (or wedged) into the fork of a tree or similar point of fixation. The weight of the body is then allowed to exert the necessary traction, with the joint being manipulated until the dislocation is reduced.

e. Infection:

(1) Infection is a serious threat to the survivor. The inevitable delay in definite medical treatment and the reality of the survival situation increases the chances of wound infection. Antibiotics may not be available in sufficient amounts in the survival situation. In survival medicine, one must place more emphasis on the prevention and control of infection by applying techniques used before the advent of antibiotics.

(2) Unfortunately, survivors have little control over the amount and type of infection introduced at the time of injury. However, they can help control the infection by wearing clean clothes. Use care to prevent additional infection into wounds. Wounds, regardless of the type or severity, should not be touched with dirty hands or objects. One exception to this rule is the essential control of arterial bleeding. Clothing should be removed from wounds to avoid contamination surrounding skin areas.

(3) All wounds should be promptly cleansed. Water is the most universally available cleaning agent, and should be (preferably) sterile. At sea level, sterilize water, by placing it in a covered container and boiling it for 10 minutes. Above 3,000 feet, water should be boiled for 1 hour (in a covered container) to ensure adequate sterilization. The water will remain sterile and can be stored indefinitely as long as it is covered.

(a) Irrigate wounds rather than scrubbing to minimize additional damage to the tissue. Foreign material should be washed from the wound to remove sources of continued infection. The skin adjacent to wounds should be washed thoroughly before bandaging. When water is not available for cleaning wounds, the survivor should consider the use of urine. Urine may well be the most nearly sterile of all fluids available and, in some cultures, is preferred for cleaning wounds. Survivors should use urine from the midstream of the urine flow.

(b) While soap is not essential to clean wounds, a bar of medicated soap placed in a personal survival kit and used routinely would do much to prevent the infection of seemingly inconsequential injuries. External antiseptics are best used for cleaning abrasions, scratches, and the skin areas adjacent to lacerations. Used in deep, larger wounds, antiseptics produce further tissue damage.

(c) Nature also provides antiseptics which can be used for wound care. The American mountain ash is found from Newfoundland south to North Carolina and its inner bark has antiseptic properties. The red berries contain ascorbic acid and have been eaten to cure scurvy. The Sweet Gum bark is still officially recognized as being an antiseptic agent. Water from boiled Sweet Gum leaves can also be used as antiseptic for wounds.

f. The "Open Treatment" Method. This is the only safe way to manage survival wounds. No effort should be made to close open wounds by suturing or by other procedure. In fact, it may be necessary to open the wound even more to avoid entrapment or infection and to promote drainage. The term "open" does not mean that dressings should not be used. Good surgery requires that although wounds are not "closed," nerves, bone, and blood vessels should be covered with tissue. Such judgment may be beyond the capability of the aircrew member, but protection of vital structures will aid in the recovery and ultimate function. A notable exception to "open treatment" is the early closure of facial wounds which interfere with breathing, eating, or drinking. Wounds, left open, heal by formation of infection resistant granulation tissue (proud flesh). This tissue is easily recognized by its moist red granular appearance, a good sign in any wound.

g. Dressings and Bandages. After cleansing, all wounds should be covered with a clean dressing. The dressing should be sterile; however, in the survival situation, any clean cloth will help to protect the wound from further infection. A proper bandage will anchor the dressing to the wound and afford further protection. Bandages should be snug enough to prevent slippage, yet not constrictive. Slight pressure will reduce discomfort in most wounds and help stop bleeding. Once in place, dressings should not be changed too frequently unless required. External soiling does not reduce the effectiveness of a dressing, and pain and some tissue damage will accompany any removal. In addition, changing dressings increases the danger of infection.

h. Physiological "Logistics." Despite all precautions, some degree of infection is almost universal in survival wounds. This is the primary reason for the "open" treatment advocated above. The human body has a tremendous capacity for combating infections if it is permitted to do so. The importance of proper rest and nutrition to wound healing and control of infection has been mentioned. In addition, the "logistics" of the injured part should be improved. The injury should be immobilized in a position to favor adequate circulation, both to and from the wound. Avoid constrictive clothing or bandages. Applying heat to an infected wound further aids in mobilizing local body defense measures. Lukewarm saltwater soaks will help draw out infection and promote oozing of fluids from the wound, thereby removing toxic products. Poultices, made of clean clay, shredded bark of most trees, ground grass seed, etc., do the same thing.

i. Drainage. Adequate natural drainage of infected areas promotes healing. Generally, wicks or drains are unnecessary. On occasion, however, it may be better to remove an accumulation of pus (abscess) and insert light, loose packing to ensure continuous drainage. The knife or other instrument used in making the incision for drainage must be sterilized to avoid introducing other types of organisms. The best way to sterilize in the field is with heat, dry or moist.

j. Antibiotics. Antibiotics, when available, should be taken for the control of infection. Consensus is that the drug should be of the so-called "broad spectrum type;" that is, be effective against any micro-organisms rather than specific for just one or two types. The exact amount to be included in survival kits will vary with the drug and basic assumptions as to the number and types of infections to be expected. Remember that antibiotics are potency-dated items (shelf-life about 4 years), and including them in survival kits requires kit inspection and drug replacement with active medical stocks.

k. Debridement. (The surgical removal of lacerated, devitalized, or contaminated tissue.) The debridement of severe wounds may be necessary to minimize infection (particularly of the gas gangrene type) and to reduce septic (toxic) shock. In essence, debridement is the removal of foreign material and dead or dying tissue. The procedure requires skill and should only be done by nonmedical personnel in case of dire emergency. If required, follow these general rules. Dead skin must be cut away. Muscle may be trimmed back to a point where bleeding starts and gross discoloration ceases. Fat which is damaged tends to die and should be cut away. Bone and nerves should be conserved where possible and protected from further damage. Provide ample natural drainage for the potentially infected wound and delay final closure of the wound.

l. Burns:

(1) Burns, frequently encountered in aircraft accidents and subsequent survival episodes, pose serious problems. Burns cause severe pain, increase the probability of shock and infection, and offer an avenue for the loss of considerable body fluids and salts. Direct initial treatment toward relieving pain and preventing infection. Covering the wound with a clean dressing of any type reduces the pain and chance for infection. Further, such protection enhances the mobility of the patient and the capability for performing other vital survival functions. In burns about the face and neck, ensure the victim has an open airway. If necessary, cricothyroidotemy should be done before the patient develops extreme difficulties. Burns of the face and hands are particularly serious in a survival situation as they interfere with the capability of survivors to meet their own needs. Soaking certain barks (willow, oak, maple) in water soothes and protects burns by astringent action. This is a function of the acid content of the bark used.

(2) Maintenance of body fluids and salts is essential to recover from burns. The only way to administer fluids in a survival situation is by mouth; hence the casualty should ingest sufficient water early before the nausea and vomiting of toxicity intervenes. Consuming the eyes and blood (both cooked) of animals can help restore electrolyte levels if salt tablets are not available. NOTE: The survivor may also pack salt in personal survival kits to replace electrolytes (¼ teaspoon per quart of water).

m. Lacerations: Lacerations (cuts) are best left open due to the probability of infection. Clean thoroughly, remove foreign material, and apply a protective dressing. Frequently, immobilization will hasten the healing of major lacerations. On occasion (tactical), it may be necessary to close (cover) the wound, despite the danger of infection, in order to control bleeding or increase the mobility of the patient. If a needle is available, thread may be procured from parachute lines, fabric, or clothing, and the wound closed by "suturing." If suturing is required, place the stitches individually, and far enough apart to permit drainage of underlying parts. Do not worry about the cosmetic effect; just approximate the tissue. For scalp wounds, hair may be used to close after the wound is cleansed. Infection is less a danger in this area due to the rich blood supply.

n. Head Injuries. Injuries to the head pose additional problems related to brain damage as well as interfering with breathing and eating. Bleeding is more profuse in the face and head area, but infections have more difficulty in taking hold. This makes it somewhat safer to close such wounds earlier to maintain function. Cricothyroidotemy may be necessary if breathing becomes difficult due to obstruction of the upper airways. In the event of unconsciousness, watch the patient closely and keep him or her still. Even in the face of mild or impending shock, keep the head level or even slightly elevated if there is reason to expect brain damage. Do not give fluids or morphine to unconscious persons.

o. Abdominal Wounds. Wounds of the abdomen are particularly serious in the survival situation. Such wounds, without immediate and adequate surgery, have an extremely high mortality rate and render patients totally unable to care for themselves. If intestines are not extruded through the wound, a secure bandage should be applied to keep this from occurring. If intestine is extruded, do not replace it due to the almost certain threat of fatal peritonitis. Cover the extruded bowel with a large dressing and keep the dressing wet with any fluid that is fit to drink, or urine. The patient should lie on the back and avoid any motions that increase intra-abdominal pressure which might extrude more bowel. Keep the survivor in an immobile state or move on a litter. "Nature" will eventually take care of the problem; either through death, or walling-off of the damaged area.

p. Chest Injuries. Injuries of the chest are common, painful and disabling. Severe bruises of the chest or fractures of the ribs require that the chest be immobilized to prevent large painful movements of the chest wall. The bandage is applied while the patient deeply exhales. In the survival situation, it may be necessary for survivors to wrap their own chest. This is more difficult but can be done by attaching one end of the long bandage (parachute material) to a tree or other fixed object, holding the other end in the hand, and slowly rolling body toward the tree, keeping enough counterpressure on the bandage to ensure a tight fit.

q. Sucking Chest Wounds. These wounds are easily recognized by the sucking noise and appearance of foam or bubbles in the wound. These wounds must be closed immediately before serious respiratory and circulatory complications occur. Ideally, the patient should attempt to exhale while holding the mouth and nose closed (Valsalva) as the wound is closed. This inflates the lungs and reduces the air trapped in the pleural cavity. Frequently, a taped, airtight dressing is all that is needed, but sometimes it is necessary to put in a stitch or two to make sure the wound is closed.

r. Eye Injuries. Eye injuries are quite serious in a survival situation due to pain and interference with other survival functions. The techniques for removing foreign bodies and for treating snow blindness are covered in standard first aid manuals. More serious eye injuries involving disruption of the contents of the orbit may require that the lids of the affected eye be taped closed or covered to prevent infection.

s. Thorns and Splinters. Thorns and splinters are frequently encountered in survival situations. Reduce their danger by wearing gloves and proper footgear. Their prompt removal is quite important to prevent infection. Wounds made by these agents are quite deep compared to their width which increases chances of infection by those organisms (such as tetanus) which grow best in the absence of oxygen. Removal of splinters is aided by the availability of a sharp instrument (needle or knife), needle nose pliers, or tweezers. Take care to get all of the foreign body out; sometimes it is best to open the wound sufficiently to properly cleanse it and to allow air to enter the wound. When cleaned, treat as any other wound.

t. Blisters and Abrasions. Care for blisters and abrasions promptly. Foot care is extremely important in the survival situation. If redness or pain is noted, the survivor should stop (if at all possible) to find and correct the cause. Frequently, a protective dressing or bandage and(or) adhesive will be sufficient to prevent a blister. If a blister occurs, do not remove the top. Apply a sterile (or clean) dressing. Small abrasions should receive attention to prevent infection. Using soap with a mild antiseptic will minimize the infection of small abrasions which may not come to the attention of the survivor.

u. Insect Bites. Bites of insects, leeches, ticks, chiggers, etc., pose several hazards. Many of these organisms transmit diseases, and the bite itself is likely to become infected, especially if it itches and the survivor scratches it. The body should be inspected frequently for ticks, leeches, etc., and these should be removed immediately. If appropriate and possible, the survivor should avoid infested areas. These parasites can best be removed by applying heat or other irritant to them to encourage a relaxation of their hold on the host. Then the entire organism may be gently detached from the skin, without leaving parts of the head imbedded. Treat such wounds as any other wound. Applying cold wet dressings will reduce itching, scratching, and swelling.

7-10. Illnesses. Many illnesses which are minor in a normal medical environment become major in a survival situation when the individual is alone without medications or medical care. Survivors should use standard methods (treat symptoms) to prevent expected diseases since treatment in a survival situation is so difficult. Key preventive methods are to maintain a current immunization record, maintain a proper diet, and exercise.

a. Food Poisoning. Food poisoning is a significant threat to survivors. Due to sporadic food availability, excess foods must be preserved and saved for future consumption. Methods for food preservation vary with the global area and situation. Bacterial contamination of food sources has historically caused much more difficulty in survival situations than the ingestion of so-called poisonous plants and animals. Similarly, dysentery or water-borne diseases can be controlled by proper sanitation and personal hygiene.

b. Treatment of Food Poisoning. If the food poisoning is due to preformed toxin; staphylococcus, botulism, etc. (acute symptoms of nausea, vomiting, and diarrhea soon after ingestion of the contaminated food), supportive treatment is best. Keep the patient quiet and lying down, and ensure the patient drinks substantial quantities of water. If the poisoning is due to ingestion of bacteria which grow within the body (delayed gradual onset of same symptoms), take antibiotics (if available). In both cases, symptoms may be alleviated by frequently eating small amounts of fine, clean charcoal. In PW situations, if chalk is available, reduce it to powder, and eat to coat and soothe the intestines. Proper sanitation and personal hygiene will help prevent spreading infection to others in the party or continuing reinfection of the patient.

Chapter 8

PW MEDICINE

8-1. Introduction:

a. Imprisoned PWs are, in the physical sense at least, under the control of their captors. Thus, the application of survival medicine principles will depend on the amount of medical service and supplies the captors can, and will, give to their prisoners. An enemy may both withhold supplies and confiscate survivor's supplies. Some potential enemies (even if they wanted to provide PW medical support) have such low standards of medical practice that their best efforts could jeopardize the recovery of the patient (figure 8-1).

b. An interesting and important sociological problem arises in getting medical care for PWs. How far should prisoners go in their efforts to get adequate rations and medicines for themselves or those for whom they are responsible? The Code of Conduct is quite specific concerning consorting with the enemy. Individuals must use considerable judgment in deciding whether to forget the welfare of fellow prisoners in order to follow the letter of the Code. Even more questionable is the individual who will offer such a justification for personal actions. Again, these questions involve more than purely medical consideration. In combat, there are apt to be frequent situations in which medical considerations are outweighed by more important ones.

8-2. History:

a. As in past wars, there were professional medical personnel among the captives in North Vietnam; however, these personnel were not allowed to care for the sick and injured as in the past. Medical care and assistance from the captors were limited and generally below comparable standards of the United States. Yet 566 men returned, most in good physical and psychological condition, having relied to a large extent on their own ingenuity, knowledge, and common sense in treating wounds and diseases. They were able to recall childhood first aid, to learn by trial and error, and to use available resources. Despite their measures of success in this respect, many released personnel felt that with some prior training, considerable improvement in self-help techniques was possible even in the most primitive conditions.

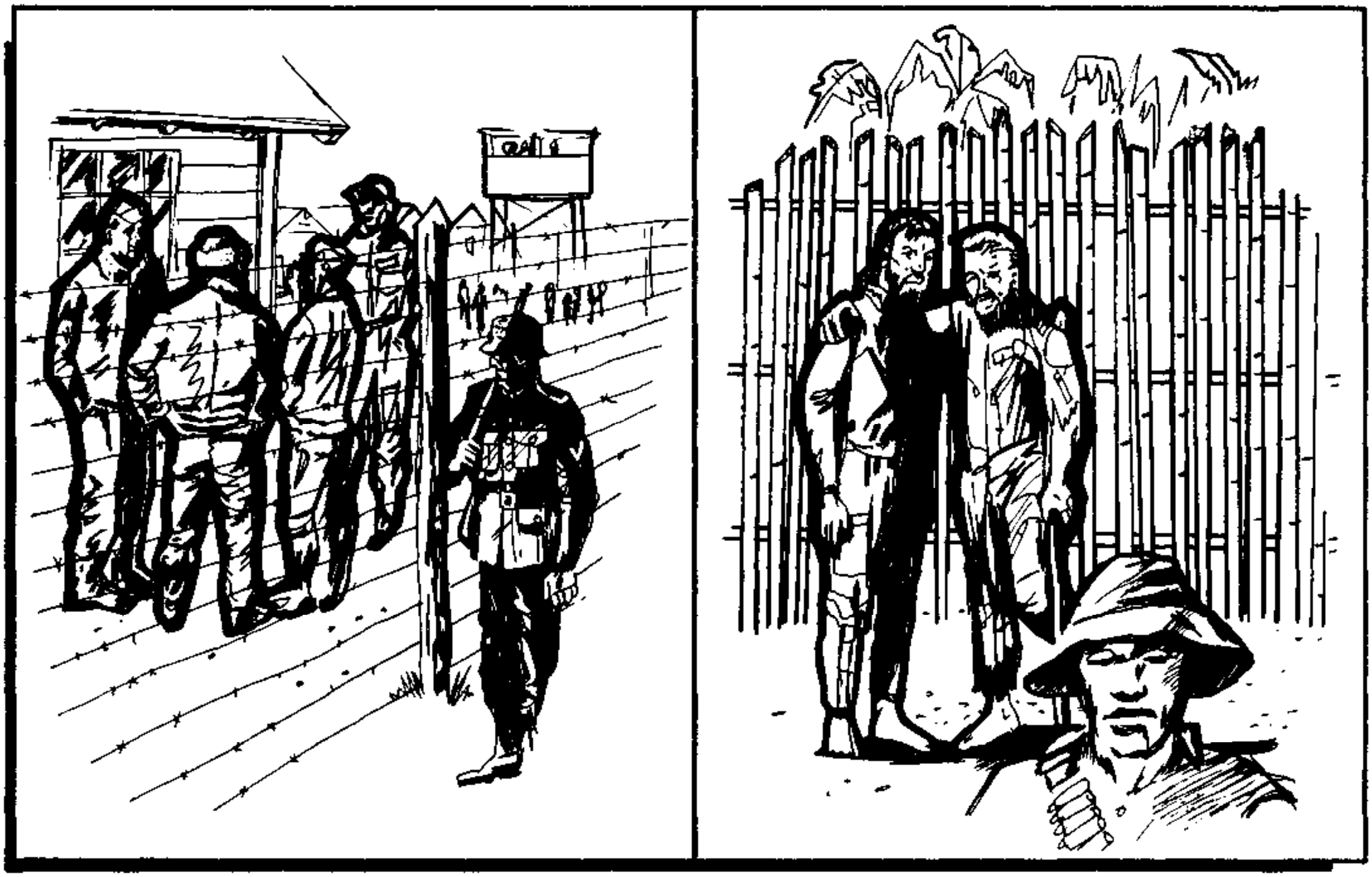

Figure 8-1. PW Medicine.

b. To determine how the services could help and to assist future PWs to care for themselves if the situation required it, the Medical Section of the Air Force Intelligence Service, with the Surgeon General of the Air Force, sponsored a 5-day seminar to examine the pertinent medical experiences of captivity and to recommend appropriate additions and changes in training techniques. As a basis for seminar discussion, Air Training Command provided data on the major diseases, wounds, and ailments, and the treatment methods used by the captives in Southeast Asia. Transcripts (325) of debriefing material were screened for medical data. Significant disease categories were established for analysis simplification based on the freqency of the problems encountered.

PROBLEMS	MAJOR CATEGORIES ESTABLISHED
Dysentery	Trauma (lacerations, burns, fractures)
Fungus	Gastrointestinal problems
Dental problems	Communicable diseases
Intestinal problems	Nutritional diseases
Fractures	Dermatological ailments
Lacerations	Dental problems
Respiratory ailments	
Burns	

In examining these major categories, attention was focused on those medical problems considered significant by the prisoners themselves in evaluating their primitive practices (self-help).

8-3. Trauma:

a. Most of the prisoners began their captivity experience with precapture injuries—burns, wounds, fractures, and lacerations. Other injuries were the result of physical abuse while a prisoner. Most of these individuals, upon their return, expressed a need to know more about managing their injuries in captivity and also what to expect about the long-term effects of injuries. It was not evident to them that the practice of a few simple rules will generally lead to acceptable results in wound treatment, and that much can be done after repatriation to correct cosmetic and functional defects.

b. The groundwork for management of injuries should begin well before an individual enters the captivity or survival environment. The treatment of injuries in survival or captivity depends primarily on providing the body the best possible circumstances to "repair itself." It is vital, therefore, to have the body in the best possible physical condition before exposure to survival or captivity. This means good cardiovascular conditioning, good muscle strength and tone, and good nutritional status. Physiological and nutritional status will markedly influence the rate and degree of healing in response to injury. The opportunity for maintaining the best possible physical conditioning and nutritional status in captivity will be greatly reduced. (Once in a captivity or survival setting, it is important to do everything possible to maintain a good physiological and nutritional status.)

8-4. Gastrointestinal Problems:

a. Diarrhea. This was a common ailment in the prison environment, not only in Vietnam, but also in WW II and Korea. It plagued the forces of North Vietnam and the allied forces. This was the second most frequent malady afflicting the Viet Cong forces. The causative factors of this almost epidemic state were varied. A variety of infectious agents gaining access to the body by use of contaminated food and water certainly contributed to the problem. Equally important as causative agents were the low level of sanitation and hygiene practices within the camps. Psychogenic responses to unappetizing diet, nutritional disturbances, and viral manifestations also contributed.

(1) Captor Therapy. This consisted primarily of local or imported antidiarrheal agents, antibiotics, and vitamins. Appropriate diet therapy was instituted.

(2) Captive Self-Therapy. After instituting diet restrictions (solid food denial and increased liquid intake), afflicted personnel were administered "concoctions" of banana skins, charcoal, chalk, or tree bark tea.

(3) Treatment Evaluation. The accepted therapy for diarrhea focuses on the causative factors which in the captivity experience were largely neglected. From a symptomatic perspective, the principles of self-treatment are simple to master: restrict intake to nonirritating foods (avoid vegetables and fruits), establish hygienic standards, increase fluid intake, and, when available, use antidiarrheal agents. The prisoners often resorted to a more exotic therapeutic regimen consisting of banana skins, charcoal, chalk, salt restriction, rice, or coffee. Charcoal, chalk, and the juice of tree barks have a scientific basis for their therapeutic success. Inasmuch as diarrhea was a source of concern and a disability for the North Vietnamese as well as the captives, therapy was often offered on request and was appropriate and successful.

(4) Conclusion. Diarrhea was frequent among PWs during captivity. Seldom fatal, it was disabling and a source of concern to those afflicted. Most captives were treated on demand and improved. This condition lends itself to some form of self-therapy through an understanding of its physiological derangements. The PW re-

sponded with intelligence, common sense, and a reasonably effective self-help regimen.

b. Dysentery. From a symptomatic perspective, dysentery is a severe form of diarrhea with passage of mucous and blood. Treatment and conclusions are similar to those for diarrhea.

c. Worms and Intestinal Parasites. Worms were extremely common among the captives. Twenty-eight percent of the released prisoners indicated worms as a significant medical problem during captivity. Worms often caused gastrointestinal problems similar to those resulting from a variety of other causes. The pin worm appears to have been the primary cause. This is not surprising, as its distribution is worldwide and the most common cause of helminthic infection of people in the United States. It requires no intermediate host; hence, infection is more rapidly acquired under poor hygienic conditions so commonly found in warm climates and conditions similar to the captivity environment. Seldom fatal, worms are significant, as they can lower the general resistance of the patient and may have an adverse affect on any intercurrent illness.

(1) Captor Therapy. This consisted of antihelminthic agents (worm medicine) dispensed without regularity, but with satisfactory results.

(2) Captive Therapy. The nuisance and irritating aspects of worms led to severe rectal itching, insomnia, and restlessness. This motivated the prisoner to find some form of successful self-therapy. Prevention was a simple and readily obtainable goal. Shoes were worn when possible; hands were washed after defecation; and fingernails were trimmed close and frequently. Peppers, popular throughout the centuries in medicine, contain certain substances chemically similar to morphine. They are effective as a counter-irritant for decreasing bowel activity. Other "house remedies" popular among the captives included drinking saltwater (a glass of water with 4 tablespoons of salt added), eating tobacco from cigarettes (chewing up to two or three cigarettes and swallowing them), and infrequently drinking various amounts of kerosene. All of these remedies have some degree of therapeutic effectiveness, but are not without danger and therefore deserve further comment. Saltwater alters the environment in the gastrointestinal tract and can cause diarrhea and vomiting. Too large an amount can have harmful effects on body fluid mechanisms and can lead to respiratory complications and death. Tobacco contains nicotine and historically was popular in the 19th century as an emetic expectorant and was used for the treatment of intestinal parasites. Nicotine is, however, one of the most toxic of all drugs and can cause death when more than 60 mg is ingested. A single cigarette contains about 30 mg of nicotine, so the captives who ate two or more cigarettes had been using a cure more dangerous than the disease. Kerosene is also toxic with 3 to 4 ounces capable of causing death. It is particularly destructive to the lungs and if through vomiting it were to make its way into the trachea and eventually the lungs, the complications would then again be far worse than the presence of worms.

(3) Treatment Evaluation. The antihelminthics therapy used by the captors was extremely effective. The problem during confinement was the nonavailability of such medication on demand. In addition, the inability to practice proper hygienic standards assured the continuation of, and reinfection with, worms.

(4) Conclusion. Worm infection in confinement is common and expected. It is seldom fatal, but contributes to general disability and mental depression due to its nuisance symptoms. Under certain circumstances, worms can assist in the spread of other diseases. The principle to follow in self-care is simple—use as high a hygienic standard as possible, and use medication causing bowel paristalsis and worm expulsion. Substances which interfere with the environment of the worms will aid in their expulsion. The toxic "house remedies" must be weighed against their possible complications.

8-5. Hepatitis. Infection of the liver was fairly common in some camps and present among the prison population throughout the captivity experience. Diagnosis was usually made on the basis of change of skin color to yellow (jaundice).

a. Captor Treatment. The Vietnamese seemed to have followed the standard therapy of rest, dietary management, and vitamin supplementation. They also displayed a heightened fear of the disease and avoided direct contact, when possible, with those afflicted.

b. Captive Therapy. For the most part, it parallels the therapy of the captors. This disease allows for little ingenuity or inventiveness of therapy.

c. Comments. Hepatitis is worldwide. Presumably most cases of hepatitis in captivity were viral in origin and easily disseminated to fellow prisoners. Conditions of poor sanitation and hygiene with close communal living foster its spread. Prevention through proper hygienic practices is the most effective tool. Equally important is an understanding of the disease characteristics. The majority of the cases recovered completely and less than 1 percent succumbed to this disease.

8-6. Nutritional Deficiencies. Symptoms attributed to malnutrition were frequent in the early years of confinement and continued up through 1969. The use of polished rice and the lack of fresh fruits and vegetables contributed to vitamin and protein deficiencies. From 1969 through 1973, food supplements were provided, and by release time, few obvious manifestations of diseases were present among those returning. The primary problems during the early years were vitamin deficiencies.

a. Vitamin B Deficiency (Beri-Beri). Presumably present among several PWs (especially those confined to the Briarpatch (Xom Ap Lo) about 15 miles west of

Sontay), it was rarely diagnosed on return. Its primary manifestation was pain in the feet described by the captives as "like a minor frostbite that turned to shooting pains."

(1) Captor Therapy. Prisoners were treated with vitamin injections and increased caloric content.

(2) Captive Therapy. Increasing caloric intake by eating anything of value. No specific self-care program existed for this malady.

(3) Comments. Beri-beri is a nutritional disease resulting from a deficiency of vitamin B (Thiamine). It is widespread in the Orient and in tropical areas where polished rice is a basic dietary staple. Of the various forms of the disease, dry beri-beri would seem most important to the confinement condition. Early signs and symptoms of the disease include muscle weakness and atrophy, loss of vibratory sensation over parts of the extremities, numbness, and tingling in the feet. From the comments of the PWs, it is difficult to formulate a diagnosis. Modern therapy consists of vitamin B or sources of the vitamin in food (such as green peas, cereal grains, and unpolished rice).

(4) Conclusion. In Vietnam, the possible early onset of the dry form of beri-beri was encountered. This is supported by the symptoms described and by the existence of dietary shortages of vitamin B and other nutritional deficiency.

b. Vitamin A Deficiency. There were several reported cases of decreased vision (primarily at night) attributed to vitamin A deficiency. This problem usually occurred during periods of punishment or politically provoked action when food was withheld as part of the discipline. The condition responded well to increased caloric intake and deserves little special mention. An understanding of the transient nature of this problem and its remedial response to therapy is important.

8-7. Communicable Diseases. Some communicable diseases were endemic in North Vietnam and certainly responsible for large scale disability among the personnel of the enemy forces. Plague, cholera, and malaria are frequent and a serious public health menace. Thanks to the immunization practices of the American forces, these diseases were of little concern to Americans during their captivity.

8-8. Skin Diseases:

a. Lesions. Dermatological lesions were common to the various prison experiences. Their importance lies not in their lethality (as they apparently did not cause any deaths), but for their irritant quality and the debilitating and grating effect on morale and mental health. Boils, fungi, heat rash, and insect bites appeared frequently and remained a problem throughout the captivity experience.

b. Boils and Blisters. A deep-seated infection usually involves the hair follicles and adjacent subcutaneous tissue, especially parts exposed to constant irritation.

(1) Captor Treatment. Prisoner complaints about the presence of boils usually brought about some action by the captors. Treatment varied considerably and obviously depended on the knowledge of medics, doctors treating their prisoners, the availability of medical supplies, and the current camp policy. For the most part, systemic antibiotics, sulfa, and tetracycline were administered. In other instances, the boils were lanced or excised and treated with topical astringents.

(2) Captive Treatment. As the medics normally responded to pleas about boils, self-treatment was practiced primarily when there was distrust of captor techniques. Prisoners would attempt to lance the boil with any sharp instrument such as needle, wire, splinters, etc., and exude their contents by applying pressure. The area was then covered with toothpaste and, when available, iodine.

(3) Comments. As noted above, the boil is an infection of hair follicles. It is more frequent in warm weather and aggravated by sweat which provides ideal conditions for the bacteria. Boils seldom appear singularly. Once present, they are disseminated by fingers, clothing, and discharges from the nose, throat, and groin. Modern therapy consists of hot compresses to hasten localization, and then conservative incision and drainage. Topical antibiotics and systemic antibiotics are then used. Boils increase in frequency with a decrease in resistance as seen in malnutrition and exhaustion states in a tropical environment. This almost mimics the prison conditions.

(4) Conclusions. Self-help treatment is limited. Of importance here is sterility when handling the boils, cleanliness, exposure to sunlight, keeping the skin dry, and getting adequate nutrition. The disease is self-limiting and not fatal. The application of any material or medication with a detergent effect may be used (soaks in saline, soap, iodine, and topical antibiotics).

8-9. Fungal Infections. Fungal infections were also a common skin problem for those in Southeast Asia captivity. As with other skin lesions, they are significant for their noxious characteristics and weakening effect on morale and mental health. Superficial fungal infections of the skin are widespread throughout the world. Their frequency among PWs reflects the favorable circumstances of captivity for cultivating fungal infections.

a. Captor Therapy. Treatment consisted of medication described by many PWs as iodine and the occasional use of sulfa powder.

b. Captive Therapy. Treatment (often the result of memory of childhood experiences and trial and error observations) consisted of the removal of body hair (to prevent or improve symptoms in the case of heat rash), exposure to sunlight to dry out fungal lesions, and development of effective techniques to foster body cooling and to decrease heat generation. Considerable effort was directed at keeping the body clean.

c. Comments. Superficial Dermatoses (skin lesions) due to fungi were common. Their invasive powers are at best uniformly weak, and because of this, infections are limited to the superficial portions of the skin and seldom by themselves fatal. Modern therapy since 1958 has relied heavily on an oral antifungal agent effective against many superficial fungi. This drug is expensive and not available in many parts of the world. Several lotions and emulsions can be used with some success. Elemental iodine is widely used as a germicide and fungicide. It is an effective antiseptic and obviously found favor in North Vietnam because of its availability. Without professional therapy, self-help, although limited in scope, can be effective. The principle of wet soaks for dry lesions and dry soaks for wet lesions is a fairly reliable guide. The use of the Sun as a drying agent can also be very effective.

d. Conclusions. Skin problems are common to the captivity environment. More importantly, extreme personal discomfort, accompanied by infection, was detrimental to the physical and mental well-being of the prisoner.

8-10. Dental Problems. These were common among all captives, not only during confinement, but also before capture. They were secondary to facial injury during egress, or caused by physical abuse during interrogation. Periodontitis (inflammation of tissue surrounding the tooth), pyorrhea (discharge of pus), and damage to teeth consistent with poor hygiene and "wear and tear" were also present.

a. Specific Complaints. Pain associated with the common toothache represented one of the most distressing problems faced by the PW. It affected the PW's nutrition and robbed the PW of the physical pleasure of eating (a highlight of isolated captivity). The inability of the PW to adequately deal with this problem caused persistent anxiety and decreased the ability to practice successful resistance techniques. In a few isolated instances, PWs actually considered collaboration with the captors in exchange for treatment and relief from tooth pain.

b. Captor Treatment. Treatment varied considerably and was no doubt influenced by political considerations. "Dentists" were infrequently available in camps before 1969. Cavities were filled, although usually inadequately, with subsequent loss of the filling. Use of local anesthesia also varied depending on the dentist providing care.

c. Captive Therapy. The PWs often chose to treat themselves rather than seek or accept prison dentistry when it was available. Abscesses were lanced with sharp instruments made locally out of wood, bamboo, or whatever was available. Brushing was excessive, again using whatever was available; chew sticks common to Asia were widely used. Aspirin (ASA), when available, was applied directly to the tooth or cavity.

d. Commentary. The self-help practices noted above had many positive aspects. The basic principle of maintaining a well-planned cleaning program using fiber, brushes, or branches certainly contributed to the relatively low incidence of cavities and infection among the prisoners. The lancing of abscesses using bamboo sticks, although not a professional maneuver, has merit insofar as the pressure is relieved and the tendency to develop into cellulitis (widespread infection) decreased. The application of aspirin directly into the cavity should be discouraged as might the application of any other substance not directly produced for this purpose.

e. Conclusions. The most effective tool against dental complications in captivity is proper preventive dentistry. The present program of the three services, if adhered to, is adequate to ensure a high state of dental hygiene while captive.

8-11. Burns. Burns were an extremely frequent injury among PWs. Severity ranged from first through third degree and occurred frequently on hands and arms.

a. Captor Treatment. For the most part, burns were treated by captors by cleaning the burns and applying antiseptics and bandages. The results obtained were, by and large, inadequate, with frequent infections and long-term debilitation.

b. Captive Treatment. No specific treatment was developed among the PWs for burns. Reliance for some form of therapy was almost completely left to the captor.

c. Commentary. Burns are extremely painful and can severely interfere with the ability to escape or to survive in captivity. The basic principle here is prevention.

d. Prevention. Adequate protection of exposed surfaces while flying (flame-retardant suits, gloves, boots, and helmet with visor down) is the best preventive action.

8-12. Lacerations and Infections:

a. Treatment. Captor treatment for lacerations and infections reflected the medical standards in North Vietnam and their domestic priorities. Wound and infection treatment varied considerably from being adequate to substandard and malpracticed. Obviously, the availability of trained physicians, a changing political climate, and difficulty in obtaining sophisticated medical supplies and equipment dictated and influenced the quality of the care delivered. The prisoners could do little professionally with this type of injury. As with diseases, the maintenance of good nutritional standards, cleanliness, and "buddy self-care" were the basic treatments.

b. Comments. When soft tissue is split, torn, or cut, there are three primary concerns—bleeding, infection, and healing of the wound.

(1) Bleeding is the first concern and must be controlled as soon as possible. Most bleeding can be controlled by direct pressure on the wound and that should be the first treatment used. If that fails, the next line of defense would be the use of classic pressure points to stop hemorrhaging. And the last method for controlling hemorrhage would be the tourniquet. The tourniquet should be used only as a last resort. Even in more favorable circumstances where the tourniquet can be applied as a first aid measure and left in place until trained medical personnel remove it, the tourniquet may result in the loss of the limb. The tourniquet should be used only when all other measures have failed, and it is a life and death matter. To control bleeding by direct pressure on the wound, sufficient pressure must be exerted to stop the bleeding, and that pressure must be maintained long enough to "seal off" the bleeding surfaces. Alternate pressing and then releasing to see if the wound is still bleeding is not desirable. It is best to apply the pressure and keep it in place for up to 20 minutes. Oozing blood from a wound of an extremity can be slowed or stopped by elevating the wound above the level of the heart.

(2) The next concern is infection. In survival or captivity, consider all breaks in the skin due to mechanical trauma contaminated, and treat appropriately. Even superficial scratches should be cleaned with soap and water and treated with antiseptics, if available. Antiseptics should generally not be used in wounds which go beneath the skin's surface since they may produce tissue damage which will delay healing. Open wounds must be thoroughly cleansed with boiled water. Bits of debris such as clothing, plant materials, etc., should be rinsed out of wounds by pouring large amounts of water into the wounds and ensuring that even the deepest parts are clean. In a fresh wound where bleeding has been a problem, care must be taken not to irrigate so vigorously that clots are washed away and the bleeding resumes. Allow a period of an hour or so after the bleeding has been stopped before beginning irrigation with the boiled water. Begin gently at first, removing unhealthy tissue, increasing the vigor of the irrigation over a period of time. If the wound must be cleaned, use great care to avoid doing additional damage to the wound. The wound should be left open to promote cleansing and drainage of infection. In captivity, frequently deep open wounds will become infested with maggots. The natural tendency is to remove these maggots, but actually, they do a good job of cleansing a wound by removing dead tissue. Maggots may, however, damage healthy tissue when the dead tissue is removed. So the maggots should be removed if they start to affect healthy tissue. Remember that it is imperative that the wound be left open and allowed to drain.

(3) An open wound will heal by a process known as secondary intention or granulation. During the healing phase, the wound should be kept as clean and dry as possible. For protection, the wound may be covered with clean dressings to absorb the drainage and to prevent additional trauma to the wound. These dressings may be loosely held in place with bandages (clean parachute material may be used for dressings and bandages). The bandages should not be tight enough to close the wound or to impair circulation. At the time of dressing change, boiled water may be used to gently rinse the wound. The wound may then be air dried and a clean dressing applied. (The old dressing may be boiled, dried, and reused.) Nutritional status is interrelated with the healing process, and it is important to consume all foods available to provide the best possible opportunity for healing.

c. Conclusions. Obviously the PW is at a distinct disadvantage in treating wounds, lacerations, and infections without modern medicine. Yet, knowledge of the basic principles mentioned above, locally available equipment and resources, and optimism and common sense can help a survivor to maintain life.

8-13. Fractures and Sprains. Fractures and sprains often occurred during shootdown and(or) egress from the aircraft. They also occurred during evasion attempts.

a. Captor Treatment. As with other treatment, treatment of sprains and fractures varied considerably depending on the severity of the injury and the resources available for treatment. Even after immediate treatment or surgical procedures, there was little followup therapy. Prisoners were usually returned to camp to care for themselves or to rely on the help of fellow prisoners.

b. Captive Treatment. Captive therapy was primarily that of helping each other to exercise or immobilize the injured area, and in severe cases, to provide nursing care.

c. Comments:

(1) An acute nonpenetrating injury to a muscle or joint can best be managed by applying cold as soon as possible after the injury. Icepacks or cold compresses should be used intermittently for up to 48 hours following the injury. This will minimize hemorrhage and disability. Be careful not to use snow or ice to the point where frostbite or cold injury occurs. As the injured part begins to become numb, the ice should be removed to permit rewarming of the tissues. Then the ice can be reapplied. Following a period of 48 to 72 hours, the cold treatment can be replaced by warm packs to the affected part. A "sprain or strain" may involve a wide variety of damage ranging from a simple bruise to deep hemorrhage or actual tearing of muscle fibers, ligaments, or tendons. While it is difficult to establish specific guidelines for treatment in the absence of a specific diagnosis, in general, injuries of this type require some period of rest (immobilization) to allow healing. The period of rest is followed by a period of rehabilitation (massage and exercise) to restore function. For what appears to be a simple superficial muscle problem, a period of 5 to 10

days rest followed by a gradual progressive increase in exercise is desirable. Pain should be a limiting factor. If exercise produces significant pain, the exercise program should be reduced or discontinued. In captivity, it is probably safest to treat severe injuries to a major joint like a fracture with immobilization (splint, cast) for a period of 4 to 6 weeks before beginning movement of the joint.

(2) Bone fractures are of two general types, open and closed. The open fracture is associated with a break in the skin over the fracture site which may range all the way from a broken bone protruding through the skin to a simple puncture from a bone splinter. The general goals of fracture management are: restore the fracture to a functional alignment; immobilize the fracture to permit healing of the bone; and rehabilitation. Restoring or reducing the fracture simply means realigning the pieces of bones, putting the broken ends together as close to the original position as possible. The natural ability of the body to heal a broken bone is remarkable and it is not necessary that an extremity fracture be completely straight for satisfactory healing to occur. In general, however, it is better if the broken bone ends are approximated so that they do not override. Fractures are almost always associated with muscle spasms which become stronger with time. The force of these muscle spasms tends to cause the ends of the broken bones to override one another, so the fracture should be reduced as soon as possible. To overcome the muscle spasm, force must be exerted to reestablish the length of the extremity. Once the ends of the bone are realigned, the force of the muscle spasm tends to hold the bones together. At this point, closed fractures are ready to be immobilized, but open fractures require treatment of the soft tissue injury in the manner outlined earlier. In other words, the wound must be cleansed and dressed, then the extremity should be immobilized. The immobilization preserves the alignment of the fracture and prevents movement of the fractured parts which would delay healing. For fractures of long bones of the body, it becomes important to immobilize the joints above and below the fracture site to prevent movement of the bone ends. In a fracture of the mid forearm, for example, both the wrist and the elbow should be immobilized. In immobilizing a joint, it should be fixed in a "neutral" or functional position. That is, neither completely straight nor completely flexed or bent, but in a position about midway between. In splinting a finger, for example, the finger should be curved to about the same position the finger would naturally assume at rest.

(3) A splint of any rigid material such as boards, branches, bamboo, metal boot insoles, or even tightly rolled newspaper may be almost as effective as plaster or mud casts. In conditions such as continuous exposure to wetness, the splint can be cared for more effectively than the plaster or mud cast. In cases where there is a soft tissue wound in close proximity to the fracture, the splint method of immobilization is more desirable than a closed cast because it permits change of dressing, cleaning, and monitoring of the soft tissue injury. The fracture site should be loosely wrapped with parachute cloth or soft plant fibers; then the splints can be tied in place extending at least the entire length of the broken bone and preferably fashioned in such a way as to immobilize the joint above and below the fracture site. The splints should not be fastened so tightly to the extremity that circulation is impaired. Since swelling is likely to occur, the bindings of the splint will have to be loosened periodically to prevent the shutting off of the blood supply.

(4) The time required for immobilization to ensure complete healing is very difficult to estimate. In captivity, it must be assumed that healing time will, in general, be prolonged. This means that for a fracture of the upper extremity of a "nonweight-bearing bone," immobilization might have to be maintained for 8 weeks or more to ensure complete healing. For a fracture of the lower extremity or a "weight-bearing bone," it might require 10 or more weeks of immobilization.

(5) Following the period of immobilization and fracture healing, a program of rehabilitation is required to restore normal functioning. Muscle tone must be reestablished and the range of motion of immobilized joints must be restored. In cases where joints have been immobilized, the rehabilitation program should be started with "passive range of motion exercises." This means moving the joint through a range of motion without using the muscles which are normally used to move that joint. For example, if the left wrist has been immobilized, a person would begin the rehabilitation program by using the right hand to passively move the left wrist through a range of motion which can be tolerated without pain. When some freedom of motion of the joint has been achieved, the individual should begin actively increasing that range of motion using the muscles of the joint involved. Do not be overly forceful in the exercise program—use pain as a guideline—the exercise should not produce more than minimal discomfort. Over a period of time, the joint movement should get progressively greater until the full range of motion is restored. Also, exercises should be started to restore the tone and strength of muscles which have been immobilized. Again, pain should be the limiting point of the program and progression should not be so rapid as to produce more than a minimal amount of discomfort.

8-14. Summary. Common sense and basic understanding of the type of injuries are most helpful in avoiding complication and debilitation. Adequate nourishment and maintenance of physical condition will materially assist healing of burns, fractures, lacerations, and other injuries—the body will repair itself.

8-15. Conclusions. In the management of trauma and burns in captivity or survival, remember that the body will do the healing or repair, and the purpose of the "treater" is to provide the body with the best possible atmosphere to conduct that self-repair. Some general principles are:

a. Be in the best possible physical, emotional, and nutritional status before being exposed to the potential survival or captivity setting.

b. Minimize the risk of injury at the time of survival or captivity by following appropriate safety procedures and properly using protective equipment.

c. Maintain the best possible nutritional status while in captivity or the survival setting.

d. Don't overtreat!!! Overly vigorous treatment can do more harm than good.

e. Use cold applications for relief of pain and to minimize disability from burns and soft tissue strains or sprains.

f. Clean all wounds by gentle irrigation with large amounts of the cleanest water available.

g. Leave wounds open.

h. Splint fractures in a functional position.

i. After the bone has healed, begin an exercise program to restore function.

j. Remember that even improperly healed wounds or fractures may be improved by cosmetic or rehabilitative surgery and treatment upon rescue or repatriation.

Part Four

FACTS AND CONDITIONS AFFECTING A SURVIVOR

Chapter 9

WEATHER

9-1. Introduction. History records many attempts by people of ancient cultures to understand the heavens. In the primitive past, the ability to predict weather was of primary importance. It was by observing the stars and other celestial bodies that these early societies could predict the coming of the seasons and therefore the weather patterns that played such a large role in their survivability. When people today are forced to live under the primitive conditions of a survival situation, they are no different from those who have struggled before them against those same conditions.

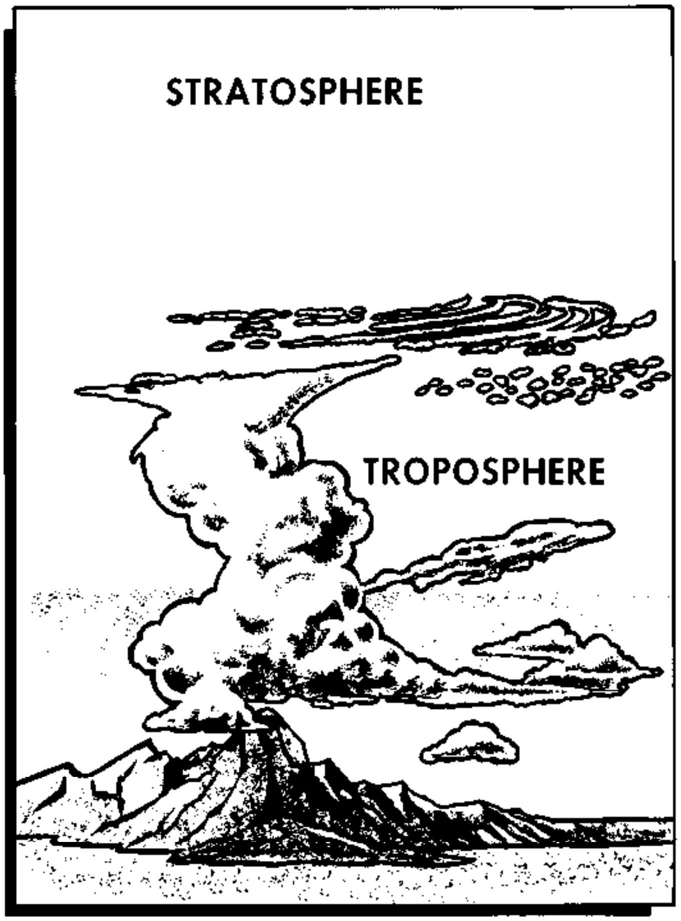

Figure 9-1. Structure of the Atmosphere.

9-2. Knowledge of Weather. However, today's participants in the age old struggle against nature must (out of necessity) still be concerned with the effect of weather. It is still true that weather cannot be controlled, but the person who is prepared (through knowledge) will be more successful.

a. Weather is not the same as climate. Weather is the state of the atmosphere, with respect to wind, temperature, cloudiness, moisture, pressure, humidity, etc. Climate, on the other hand, is the type of weather condition generally prevailing over a region throughout the year, averaged over a series of years.

b. The atmosphere extends upward from the surface of the Earth for a great many miles, gradually thinning as it approaches its upper limit. Near the Earth's surface, the air is relatively warm due to contact with the Earth. As altitude increases, the temperature decreases by about 3.5°F for every 1,000 feet until air temperature reaches about 67°F below zero at 7 miles above the Earth.

c. To understand where the weather patterns originate, a brief familiarization of the "layers" or structure of the atmosphere is needed (figure 9-1). The atmosphere is divided into two layers. The upper layer is the "stratosphere" where the temperature remains constant. The lower layer is the "troposphere" where the temperature changes. Nearly all weather occurs in this lower layer which begins at the Earth's surface and extends upwards for 6 to 10 miles.

9-3. Elements Affecting Weather. Weather conditions in the troposphere and on the Earth are affected by four elements: temperature, air pressure, wind, and moisture.

a. Temperature is the measure of the warmth or coldness of an object or substance and, for this discussion, the various parts of the atmosphere. The sunlight entering the atmosphere reaches the Earth's surface and warms both the ground and the seas. Heat from the ground and the seas then warms the atmosphere. The atmosphere absorbs the heat and prevents it from escaping into space. This process is called the greenhouse effect because it resembles the way a greenhouse works. Once the Sun sets, the ground cools more slowly than the air because it is a better conductor of heat. At night the ground is warmer than the air, especially under a clear dry sky. The ground cools more slowly than the humid nights. Temperature changes near the ground for other reasons. Dark surfaces are warmer than light-colored surfaces. Evening air settles in low areas and valleys creating spots colder than higher elevations.

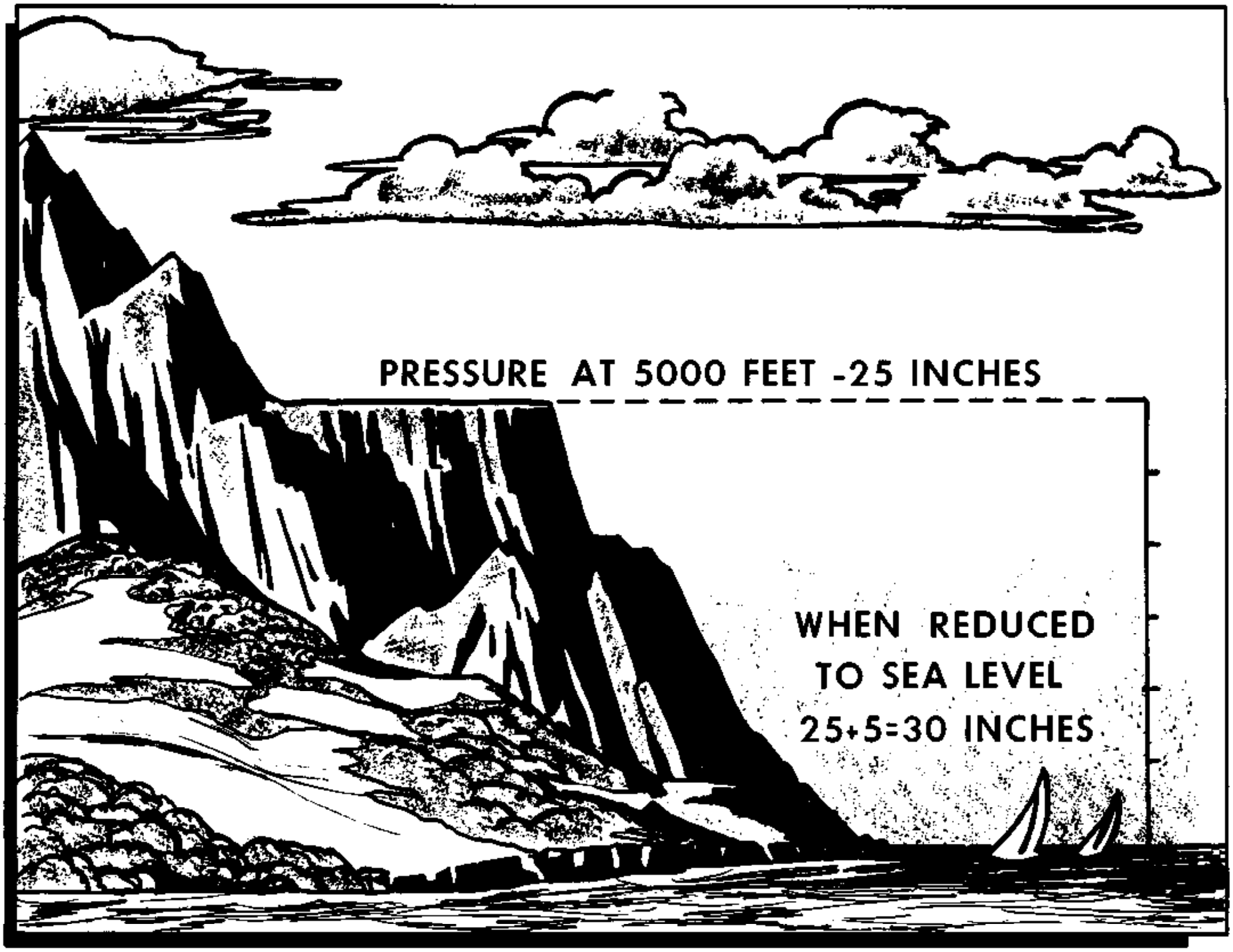
PRESSURE AT 5000 FEET -25 INCHES
WHEN REDUCED TO SEA LEVEL 25+5=30 INCHES

Figure 9-2. Air Pressure.

Seas, lakes, and ponds retain heat and create warmer temperatures at night near shore. The opposite is true during the day, especially in the spring when lakes are cold. On the beach, daytime temperatures will be cooler than the temperatures on land further from shore. Knowing this will help determine where a survivor should build a shelter.

b. Air pressure is the force of the atmosphere pushing on the Earth. The air pressure is greatly affected by temperature. Cool air weighs more than warm air. As a result, warm air puts less pressure on the Earth than does cool air. A low-pressure area is formed by warm air whereas cool air forms a high-pressure area (figure 9-2).

c. Wind is the movement of air from a high-pressure area to a low-pressure area. The larger the difference in pressure, the stronger the wind. On a global scale, the air around the Equator is replaced by the colder air around the poles (figure 9-3). This same convection of air on a smaller scale causes valley winds to blow upslope during the day and down the mountainside at night. Cool air blows in from the ocean during the day due to the heating and rising of the air above the land and reverses at night (figures 9-4 and 9-5). This movement of air creates winds throughout the world. When cool air moves into a low-pressure area, it forces the air that was already there to move upward. The rising air expands and cools.

d. Moisture enters the atmosphere in the form of water vapor. Great quantities of water evaporate each day from the land and oceans causing vapor in the air called humidity. The higher the humidity, the higher the moisture content in the air. Air holding as much moisture as possible is saturated. The temperature at which the air becomes saturated is called the dew point. When the temperature falls below the dew point, moisture in the air condenses into drops of water. Low clouds called fog may develop when warm, moist air near the ground is cooled to its dew point. A cooling of the air may also

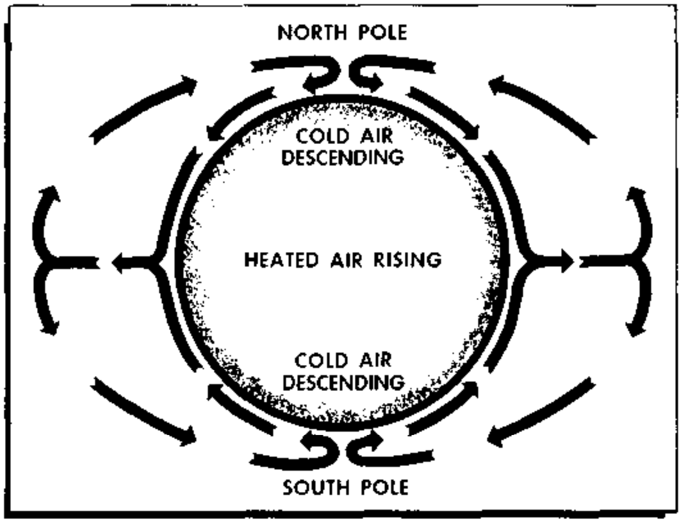
NORTH POLE
COLD AIR DESCENDING
HEATED AIR RISING
COLD AIR DESCENDING
SOUTH POLE

Figure 9-3. Wind Movement.

cause moisture to fall to the Earth as precipitation (rain, snow, sleet, or hail).

9-4. Circulation of the Atmosphere. If the Earth did not rotate, wind would move directly from the high-pressure areas of the poles to the low-pressure areas of the Equator. The movement of air between the poles and the Equator would go on constantly.

a. The rotation of the Earth prevents winds from the poles and the Equator from moving directly north or south. The Earth rotates from west to east, and as a result, winds moving toward the Equator seem to curve toward the west. Winds moving away from the Equator seem to curve toward the east. This is known as the Coriolis effect (figure 9-6). This effect results in winds circling the Earth in wide bands. These prevailing winds are divided into six belts which are known as the trade winds, the prevailing westerlies, and the polar easterlies; all three are found in both the Northern and Southern Hemispheres (figure 9-7).

(1) The winds blowing toward the Equator are known as the trade winds. The air above the Equator is so hot it is always rising. The north and south trade winds move in to take the place of the rising air. The Coriolis effect makes the trade winds appear to move from the east. The weather in the region of the trade winds moves from east to west because of the Earth's rotation. The doldrums is the region where the trade winds from the north and south meet near the Equator. The doldrums is usually calm, but it is quite rainy and may have periods of gusty winds.

(2) The prevailing westerlies blow away from the Equator. They occur north of the trade winds in the Northern Hemisphere and south of the trade winds in the Southern Hemisphere. The prevailing westerlies seem to move from the west because of the Coriolis effect. The weather in the region of these winds blows

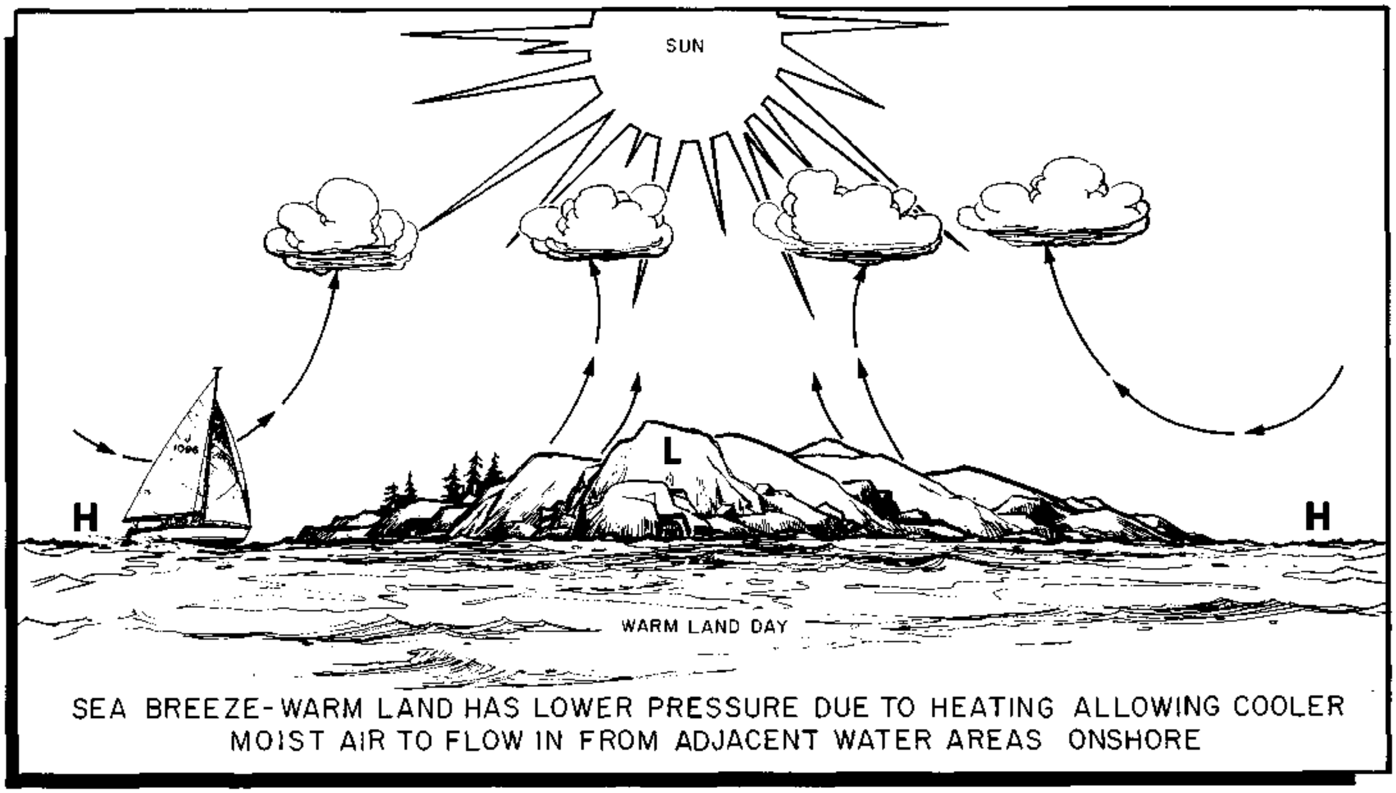
SUN
L
H
H
WARM LAND DAY
SEA BREEZE-WARM LAND HAS LOWER PRESSURE DUE TO HEATING ALLOWING COOLER MOIST AIR TO FLOW IN FROM ADJACENT WATER AREAS ONSHORE

Figure 9-4. Air Transfer (Daytime).

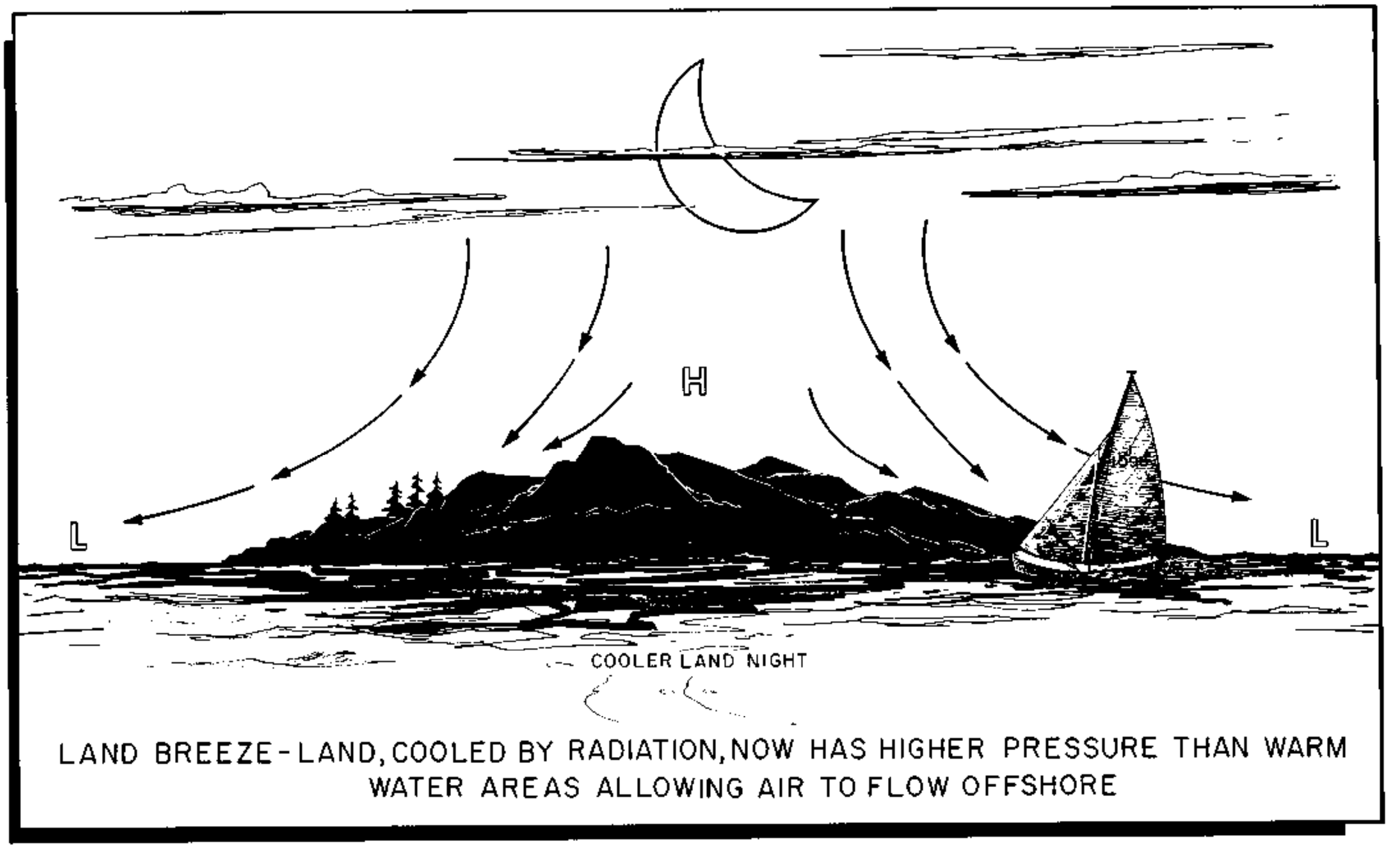
H
L
L
COOLER LAND NIGHT
LAND BREEZE - LAND, COOLED BY RADIATION, NOW HAS HIGHER PRESSURE THAN WARM WATER AREAS ALLOWING AIR TO FLOW OFFSHORE

Figure 9-5. Air Transfer (Nighttime).

from west to east. The prevailing westerlies move across most of the United States and Canada, and are divided from the trade winds in a region called the Horse Latitudes. The air in the Horse Latitudes blows downward to fill the space which was left between the prevailing westerlies and the trade winds. The winds are very light in the Horse Latitudes.

(3) The winds from the North and South Poles are known as the polar easterlies. Because the air is so cold, making it heavy, the air above the poles sinks downward. The air spreads out when it reaches the ground and moves toward the Equator. The weather in the region of the polar easterlies moves from east to west with the Coriolis effect making the winds seem to blow from the east. The polar front is the meeting place of the polar easterlies and the prevailing westerlies and is a cloudy, rainy region. Above the polar front is a band of west winds called the jet stream. The jet stream occurs about 5 to 7 miles above the ground. Its winds may exceed 200 miles per hour.

b. Pressure systems are highs or lows covering areas as big as 1 million square miles. Most pressure systems found in the United States and Canada develop along the polar front. There, the cold winds of the polar easterlies and the warmer winds of the prevailing westerlies move past one another and create swirling winds called eddies. These eddies are carried eastward across the United States and Canada by the prevailing westerlies. There are two kinds of eddies: cyclones and anticyclones (figure 9-8).

(1) Cyclones formed by eddies are not the same as the storms known as cyclones. The winds of the eddies that create cyclones swirl inward toward a center of low pressure. A low-pressure system is formed by the cyclone and its low-pressure region. Because of the rotation of the Earth, cyclones that build north of the Equator blow in a counterclockwise direction. Cyclones that form south of the Equator move in a clockwise direction. Cyclones in North America generally approach on brisk winds, bringing cloudy skies and usually rain or snow.

(2) Anticyclones swirl outward around a center of high pressure, forming a high-pressure system. Anticyclones move in a clockwise direction north of the Equator and counterclockwise south of the Equator. Anticyclones come after cyclones, bringing dry, clearing weather and light winds.

c. Airmasses depend largely on the temperature and moisture of the areas in which they originate. Airmasses may cover 5 million square miles. As they move away from their source regions and pass over land and sea, the airmasses are constantly being modified through

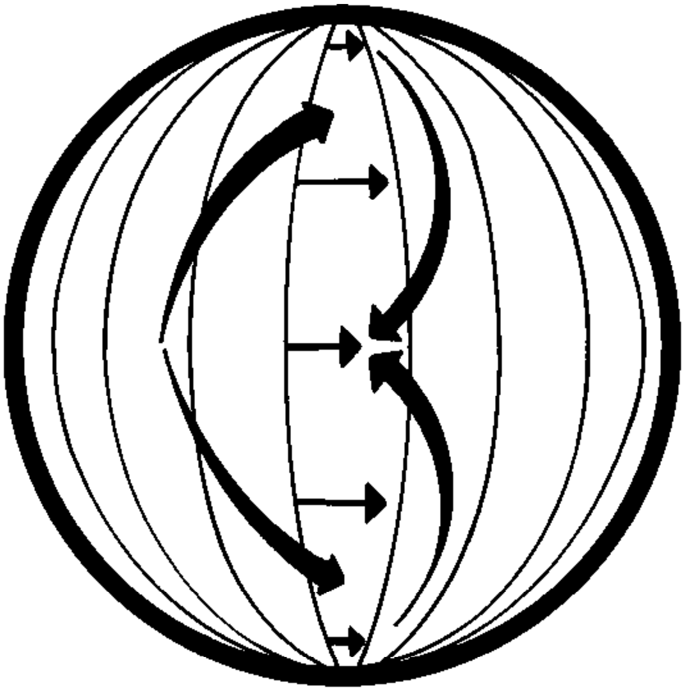

Figure 9-6. Coriolis Effect.

heating or cooling from below, lifting or subsiding, absorbing or losing moisture. In general, however, they retain some of their original characteristics and can be recognized and identified.

(1) There are four major types of airmasses:

(a) Continental polar - cold and dry.

(b) Continental tropical - hot and dry.

(c) Maritime polar - cool and moist.

(d) Maritime tropical - warm and moist.

(2) In North America, the continental polar airmass over northern Canada blows cold, dry air into southern Canada and the United States. Maritime polar airmasses off the northeast and northwest coasts of North America bring cool, damp weather to the continent. Maritime tropical airmasses from the southeast and southwest coasts bring warm, muggy weather. The polar airmasses are strongest in the winter, and the tropical airmasses are strongest in the summer. During the winter, a cold arctic airmass from the North Pole also influences the weather of North America. A continental tropical airmass forms over the southwest United States during the warm months but disappears in the winter.

d. When two different airmasses meet, they do not ordinarily mix (unless their temperatures, pressures, and relative humidities happen to be very similar). Instead, they set up boundaries called frontal zones, or "fronts." The colder airmass moves under the warmer airmass in the form of a wedge. If the boundary is not moving, it is termed a stationary front. Usually, however, the boundary moves along the Earth's surface, and as one airmass withdraws from a given area, it is replaced by another airmass. This action creates a moving front. If warmer air is replacing colder air, the front is called "warm;" if colder air is replacing warmer air, the front is called "cold." Most changes in the weather occur along fronts. The movement of fronts depends on the formation of pressure systems. Cyclones push fronts along at speeds of 20 to 30 miles per hour. Anticyclones blow into an area after a front has passed.

(1) When a warm front moves forward, the warm air slides up over the wedge of colder air lying ahead of it (figure 9-9). This warm air usually has high humidity. As this warm air is lifted, its temperature is lowered. As the lifting process continues, condensation occurs, low nimbostratus and stratus clouds form from which rain develops. The rain falls through the cooler air below, increasing its moisture content. Any reduction of temperature in the colder air, which might be caused by upslope motion or cooling of the ground after sunset, may result in extensive fog. As the warm air progresses up the slope, with constantly falling temperature, clouds appear at increasing heights in the form of altostratus and cirrostratus, if the warm air is stable. If the warm air is unstable, cumulonimbus clouds and altocumulus clouds will form and frequently produce thunderstorms. Finally, the air is forced up near the stratosphere and in the freezing temperatures at that level, the condensation appears as thin wisps of cirrus clouds. The upslope movement is very gradual, rising about 1,000 feet every 20 miles. Thus, the cirrus clouds, forming at perhaps 25,000 feet altitude, may appear as far as 500 miles in advance of the point on the ground which marks the position of the front. Warm fronts produce more gradual changes in the weather than do cold fronts. The changes depend chiefly on the humidity of the advancing warm airmass. If the air is dry, cirrus clouds may form and there will be little or no precipitation. If the air is humid, light, steady rain or snow may fall for several days. Warm fronts usually have light winds. The passing of a warm front brings a sharp rise in temperature, clearing skies, and an increase in humidity.

(2) When the cold front moves forward, it acts like a snowplow, sliding under the warmer air and tossing it aloft. This causes sudden changes in the weather. In fast-moving cold fronts, friction retards the front near the ground, which brings about a steeper frontal surface. This steep frontal surface results in a narrower band of weather concentrated along the forward edge of the front. If the warm air is stable, an overcast sky may occur for some distance ahead of the front, accompanied by general rain. If the warm air is conditionally unstable, scattered thunderstorms and showers may form in the warm air. In some cases, an almost continuous line of thunderstorms is formed and called a "squall line." Behind the fast moving cold front there is usually rapid clearing, with gusty and turbulent surface winds and colder temperatures. The slope of a cold front is much steeper than that of a warm front and the progress

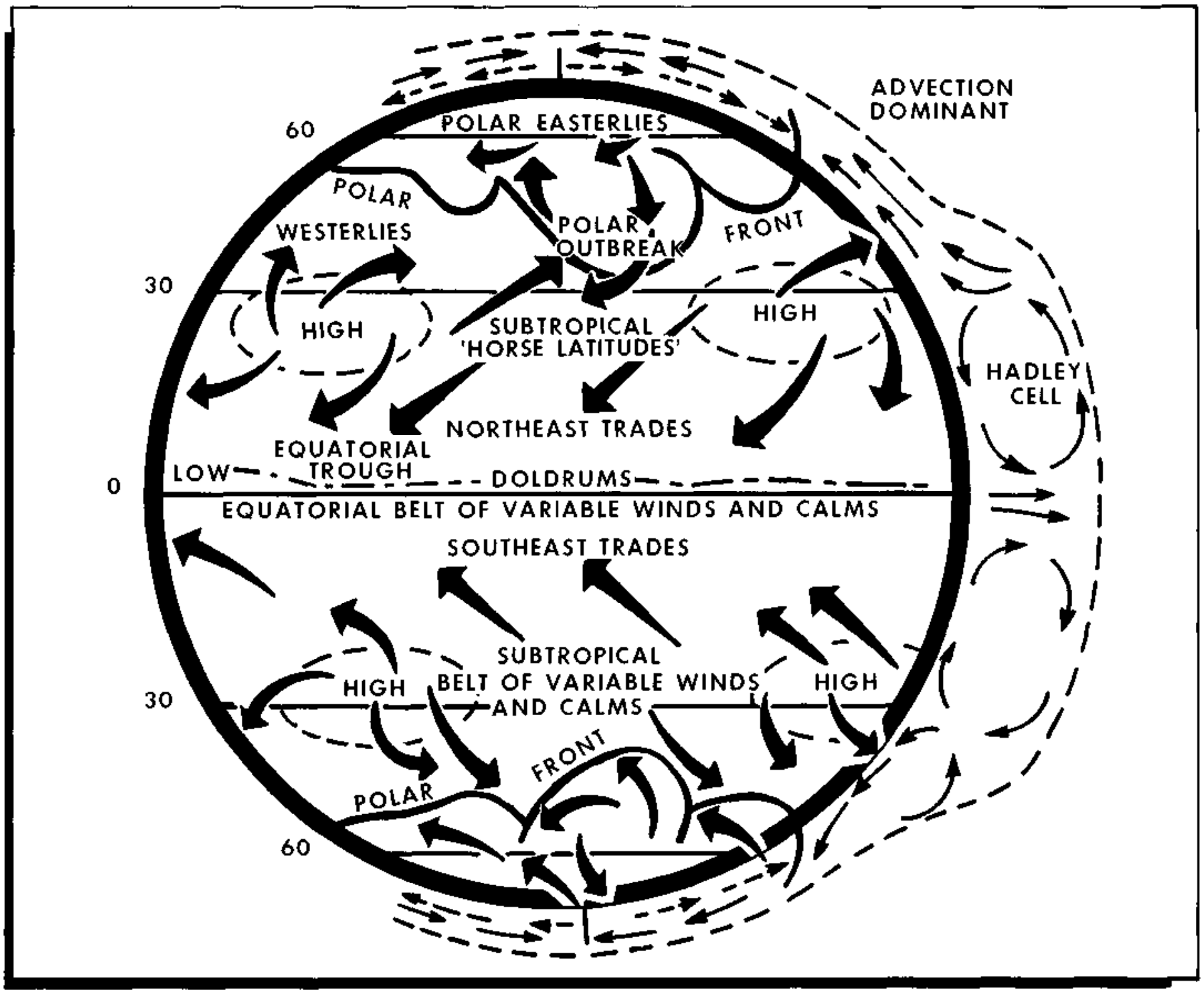
ADVECTION DOMINANT
60
POLAR EASTERLIES
POLAR WESTERLIES
POLAR OUTBREAK
FRONT
30
HIGH
SUBTROPICAL 'HORSE LATITUDES'
HIGH
HADLEY CELL
NORTHEAST TRADES
EQUATORIAL LOW TROUGH
0
DOLDRUMS
EQUATORIAL BELT OF VARIABLE WINDS AND CALMS
SOUTHEAST TRADES
SUBTROPICAL BELT OF VARIABLE WINDS AND CALMS
HIGH
HIGH
30
FRONT
POLAR
60

Figure 9-7. Atmosphere Circulation.

is generally more rapid—usually from 20 to 35 miles an hour, although in extreme cases, cold fronts have been known to move at 60 miles per hour (figure 9-10). Weather activity is more violent and usually takes place directly at the front instead of in advance of the front. However, especially in late afternoon during the warm season, a squall line will frequently develop as much as 50 to 200 miles in advance of the actual cold front. Whereas warm front dangers lie in low ceilings and visibilities, cold front dangers lie chiefly in sudden storms with high and gusty winds. Unlike the warm front, the cold front arrives almost unannounced, makes a complete change in the weather within the space of a few hours, and passes on. The squall line is ordinarily quite narrow—50 to 100 miles in width—but is likely to extend for hundreds of miles in length, frequently lying across the entire United States in a line running from northeast to southwest. Altostratus clouds sometimes form slightly ahead of the front, but these are seldom more than 100 miles in advance. After the front has passed, the weather clears rapidly with cooler, drier air.

(3) One other form of front with which the survivor should become familiar is the "occluded front" (figure 9-11). Cold fronts travel about twice as fast as warm fronts. As a result, cold fronts often catch up to warm fronts. When a cold front reaches a warm front, an occluded front develops. Meteorologists subdivide occlusions into two types: cold-front occlusions and warm-front occlusions. In a cold-front occlusion, the air behind the cold front is colder than the air ahead of the warm front. The weather of a cold-front occlusion re-

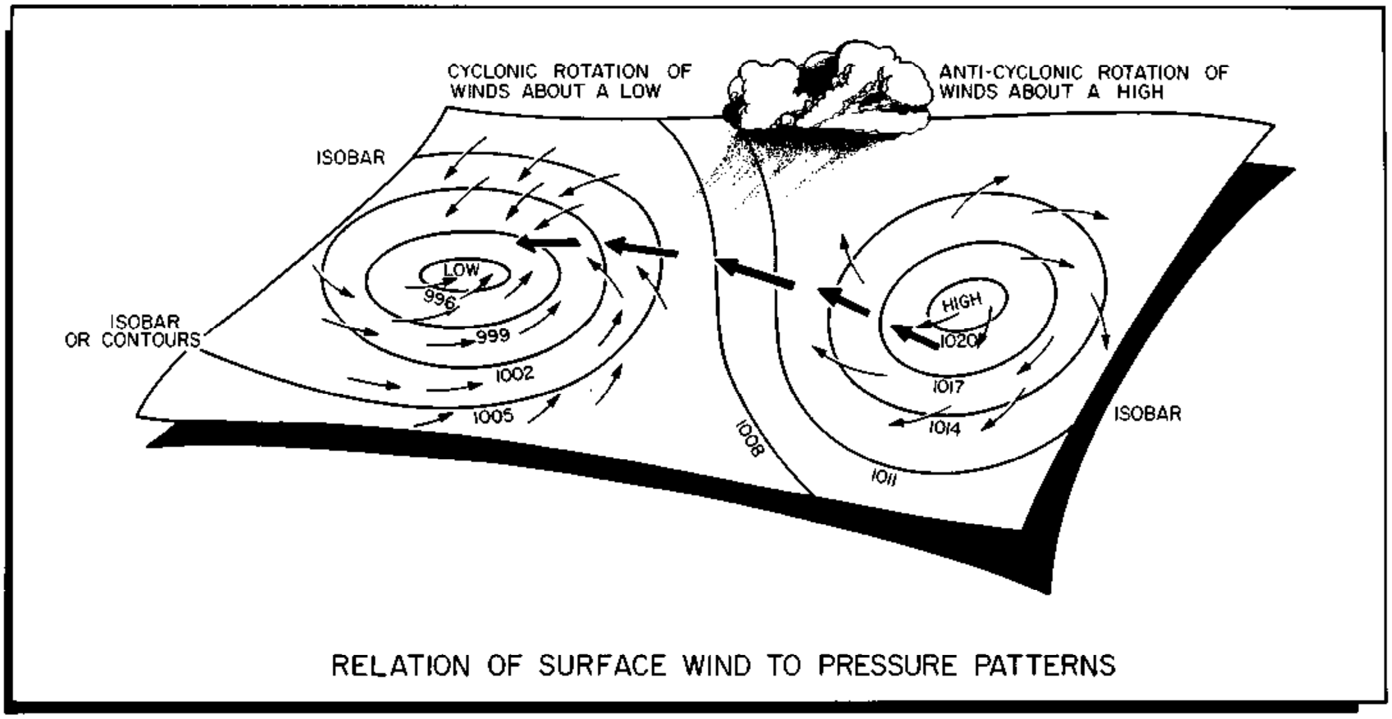
CYCLONIC ROTATION OF WINDS ABOUT A LOW
ANTI-CYCLONIC ROTATION OF WINDS ABOUT A HIGH
ISOBAR
LOW
996
999
1002
1005
ISOBAR OR CONTOURS
1008
HIGH
1020
1017
1014
1011
ISOBAR
RELATION OF SURFACE WIND TO PRESSURE PATTERNS

Figure 9-8. Cyclonic and Anticyclonic Rotation.

sembles that of a cold front. When the air behind the cold front is warmer than the air ahead of the warm front, it is known as a warm-front occlusion. Warm-front occlusion weather is similar to a warm front. These fronts produce milder weather than do cold or warm fronts.

(4) Stationary fronts are another type of front which occurs when airmasses meet but move very slow-

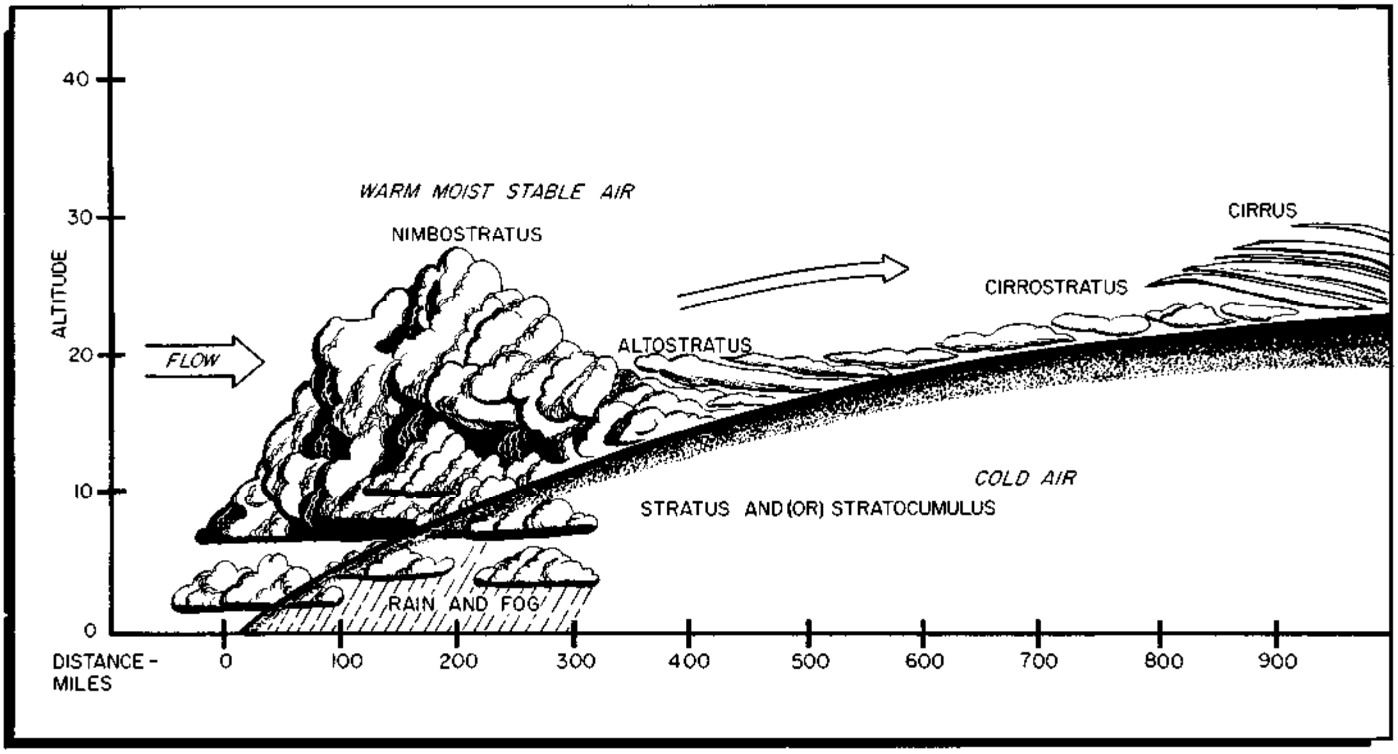
40
30
20
10
0
ALTITUDE
WARM MOIST STABLE AIR
NIMBOSTRATUS
FLOW
ALTOSTRATUS
CIRROSTRATUS
CIRRUS
COLD AIR
STRATUS AND (OR) STRATOCUMULUS
RAIN AND FOG
DISTANCE - MILES
0
100
200
300
400
500
600
700
800
900

Figure 9-9. Stable Air Warm Front.

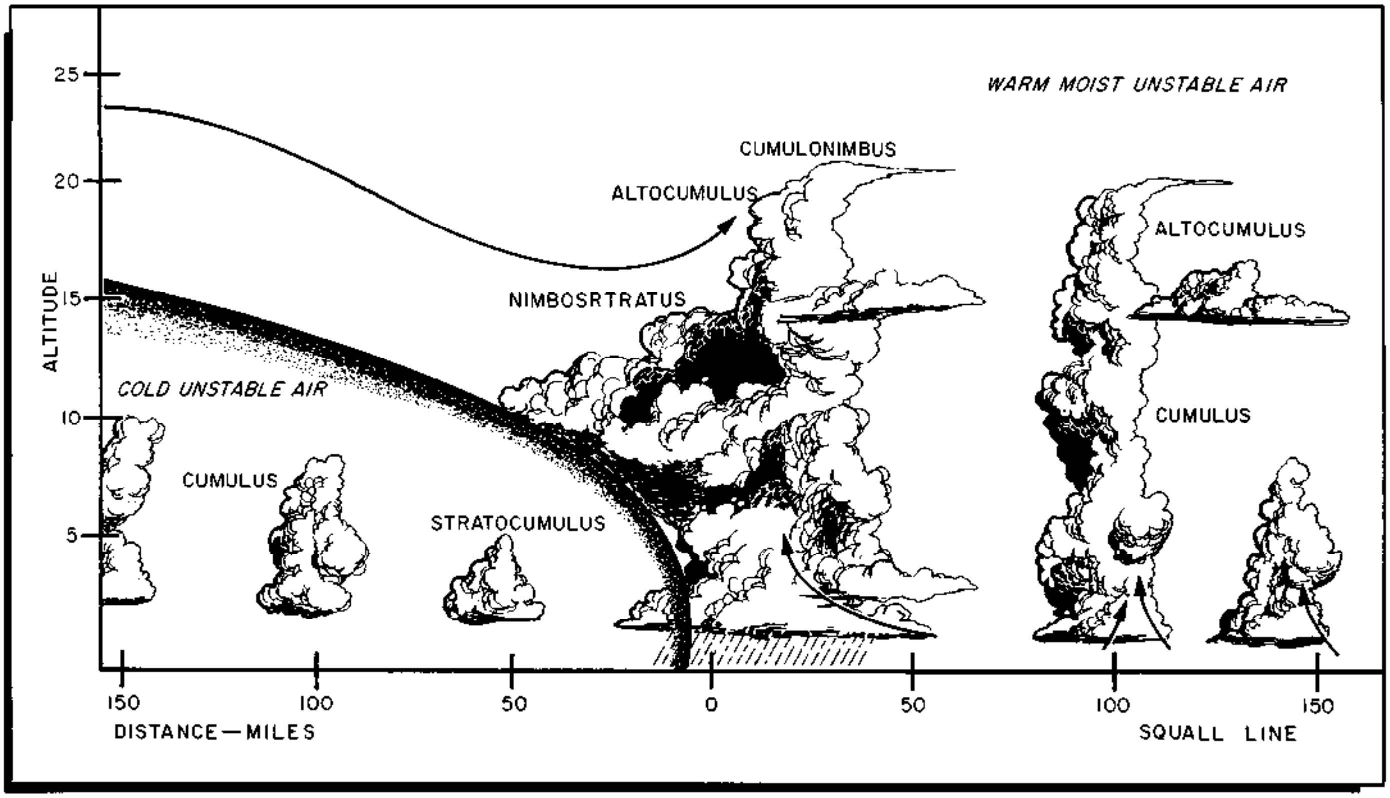

Figure 9-10. Fast-Moving Cold Front.

ly. It may remain over an area for several days bringing moderate weather.

9-5. Storms. The four main types of violent weather a person should be familiar with are thunderstorms, winter storms, tornadoes, and hurricanes.

a. Thunderstorms are the most frequent kinds of storms. As many as 50,000 thunderstorms occur throughout the world each day. Under some conditions, the rapid lifting of moist, warm air results in thunderstorms and dramatic cloud formations (figure 9-12). They develop from tall, puffy cumulonimbus clouds. Clouds may tower 5 to 10 miles high during hot, humid days. The temperatures inside the clouds are well below freezing. The air currents inside the clouds move up and down as fast as 5,000 feet per minute. Heavy rain is common because water vapor condenses rapidly in the air. Lightning and thunder occur during the life of a thunderstorm. When the sound of thunder is heard, a survivor should seek shelter immediately. Lightning causes more fatalities than any other type of weather phenomenon. In the United States alone more than 200 lightning deaths occur each year. Another reason for seeking shelter immediately is to escape the hail which sometimes accompanies the thunderstorm. Hail, which can grow as large as baseballs, is most noted for damaging crops, but a powerful storm can bring injuries, even fatalities, to survivors if shelter is not available.

b. Tornadoes are the most violent form of thunderstorms. Under certain conditions, violent thunderstorms will generate winds swirling in a funnel shape with rotational speeds of up to 400 miles per hour which extends out of the bottom of the thunderstorm. When this funnel-shaped cloud touches the surface, it can cause major destruction. The path of a tornado is narrow, usually not more than a couple of hundred yards wide. Tornadoes form in advance of a cold front and are usually accompanied by heavy rain and thun- southern areas of the United States.

c. Winter storms include ice storms and blizzards. An ice storm may occur when the temperature is just below freezing. During this storm, precipitation falls as rain but freezes on contact with the ground. A coating of ice forms on the ground and makes it very hazardous to the traveler. Snowstorms with high winds and low temperatures are called blizzards. The wind blows at 35 miles per hour or more during a blizzard, and the temperture may be 10°F or less. Blowing snow makes it impossible to travel because of low visibility and drifting.

d. A hurricane or typhoon, the most feared of storms, has a far more widespread pattern than a tornado. The storm forms near the Equator over the oceans and is a

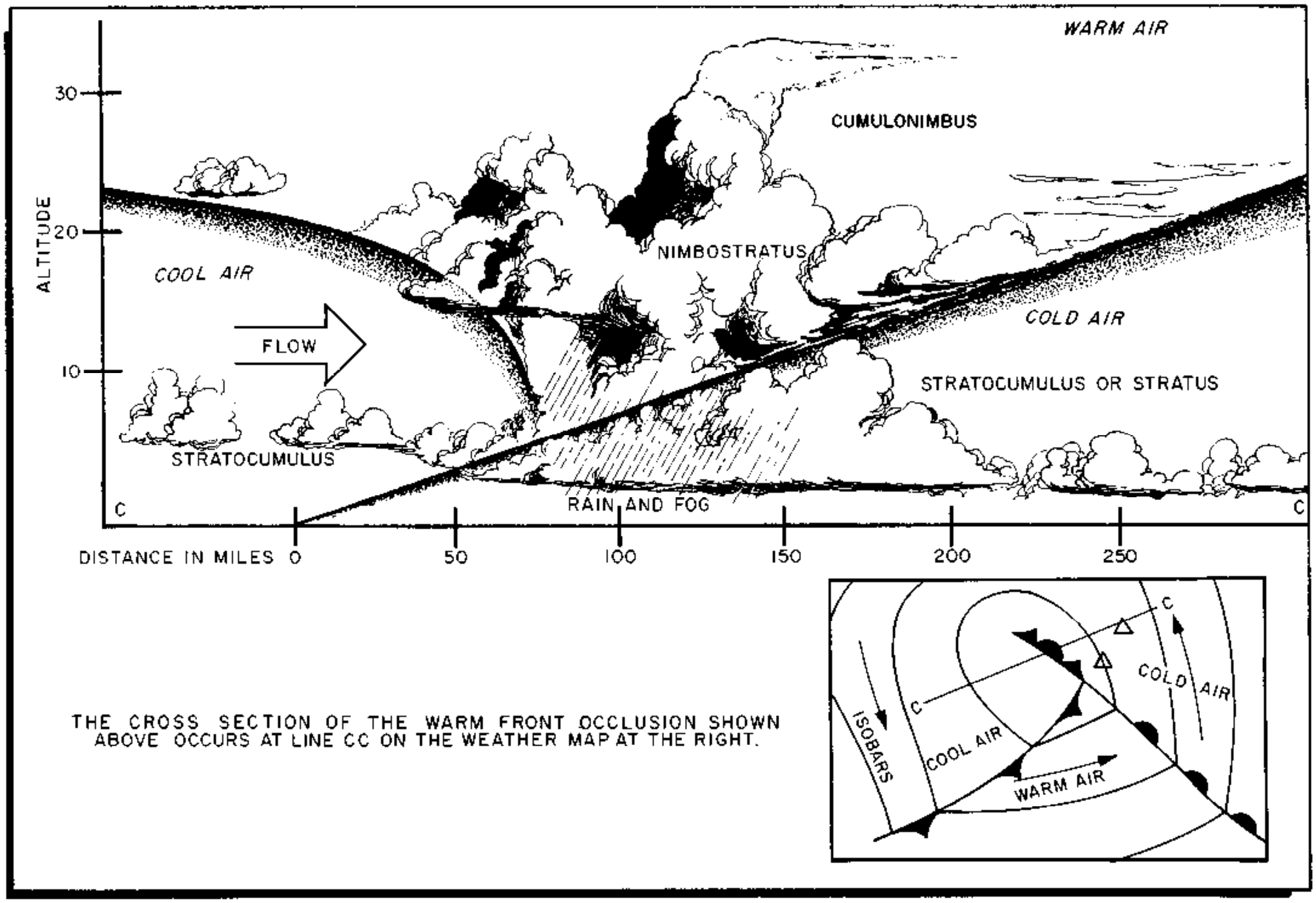

THE CROSS SECTION OF THE WARM FRONT OCCLUSION SHOWN ABOVE OCCURS AT LINE CC ON THE WEATHER MAP AT THE RIGHT.

Figure 9-11. A Warm-Front Occlusion.

large low-pressure area, about 500 miles in diameter. Winds swirl around the center (eye) of the storm at speeds over 75 miles per hour and can reach 190 miles per hour. Hurricanes break up over land and often bring destructive winds and floods. Thunderstorms often form within hurricanes and can produce tornadoes. Most hurricanes occurring in the United States sweep over the West Indies and strike the southeastern coast of the country. An early indication of a hurricane is a wind from an unusual direction, like the replacement of the normal flow of the trade winds from an easterly direction. The arrival of high waves and swells at sea coming from an unusual direction may also give some warning. The high waves and swells are moving faster than the storm and may give several days warning.

9-6. Weather Forecasting. Weather forecasting enables survivors to make plans based on probable changes in the weather. Forecasts help survivors decide what clothes to wear and type of shelter to build. During an evasion situation, it may help survivors determine when to travel. While accurate weather prediction or forecasting normally requires special instruments, an awareness of changing weather patterns and attention to existing conditions can help a survivor or evader prepare for and, when appropriate, use changing weather conditions to enhance their survivability. The following are some elementary weather indicators which could help predict the weather and help save lives.

a. Clouds which move higher are good signs of fair weather. Lower clouds indicate an increase in humidity, which in all probability means precipitation (figure 9-13).

b. The Moon, Sun, and stars are all weather indicators. A ring around the Moon or Sun means rain (figure 9-14). The ring is created when tiny ice particles in fine cirrus clouds scatter the light of the Moon and the Sun in different directions. When stars appear to twinkle, it indicates that strong winds are not far off, and will become strong surface winds within a few hours. Also, a large number of stars in the heavens show clear visibility with a good chance of frost or dew.

c. "Low-hanging" clouds over mountains mean a weather change (figure 9-13). If they get larger during

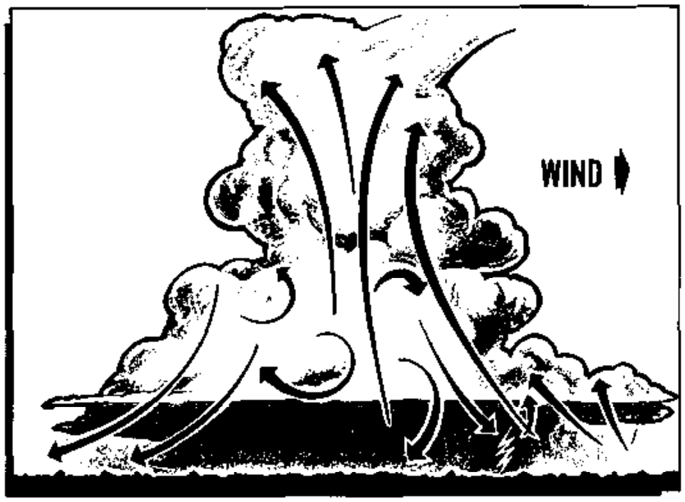

Figure 9-12. Thunderstorms.

the daytime, bad weather will arrive shortly. Diminishing clouds mean dry weather is on its way. Storms are often preceded by high thin cirrus clouds arriving from the west. When these thicken and are obscured by lower clouds, the chances increase for the arrival of rain or snow.

d. The old saying "red skies at night, sailor's delight; red skies at morning, sailors take warning," has validity. The morning Sun turning the eastern sky crimson often signals the arrival of stormy weather. As the storm moves east, clouds may turn red as a clearing western sky opens for the setting Sun.

e. "The farther the sight, the nearer the rain," is a seaman's chant. When bad weather is near, the air pressure goes down and the atmosphere becomes clearer. High atmospheric pressure with stable and dusty air means fair weather.

f. A cold front arriving in the mountains during the summer usually means several hours of rain and thunderstorms. However, the passing of a cold front means several days of clear, dry weather.

g. A morning rainbow is often followed by a squall. An afternoon rainbow means unsettled weather, while an evening rainbow marks a passing storm. A faint rainbow around the Sun precedes colder weather.

h. Stormy weather will probably follow within hours when flowers seem to be much more fragrant.

i. People say "when sounds are clear, rain is near," because sound travels farther before storms.

j. Even birds can help predict the weather. Water birds fly low across the water when a storm is approaching. Birds will huddle close together before a storm.

k. The flowers of many plants, like the dandelion, will close as humidity increases and rain is on the horizon.

l. As humidity increases, the rocks in high mountain areas will "sweat" and provide an indication of forthcoming rain.

m. Lightning can tell survivors something by noting the color and compass direction. If the lightning looks white when seen through clear air and is located in the west or northwest, survivors would know the storm is headed toward them. Storms to the south or east will normally pass to the east. Red or colored lightning is seen at a distance in storms that will pass to the north or south.

n. Smoke, rising from a fire then sinking low to the ground, can indicate that a storm is approaching.

9-7. Summary. Even with the modern equipment available, forecasting tomorrow's weather is often difficult. This chapter provided background information and tips to use to teach survivors to predict weather. By understanding the basic characteristics and actions of weather, the survivor can better prepare for its effects.

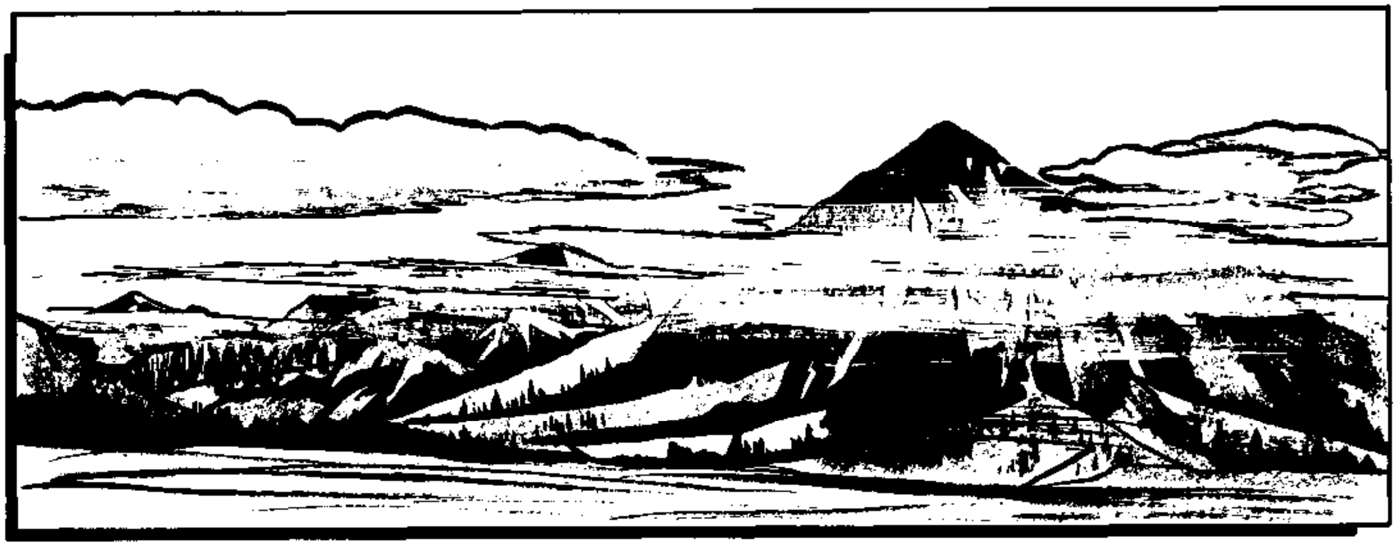

Figure 9-13. Low Clouds.

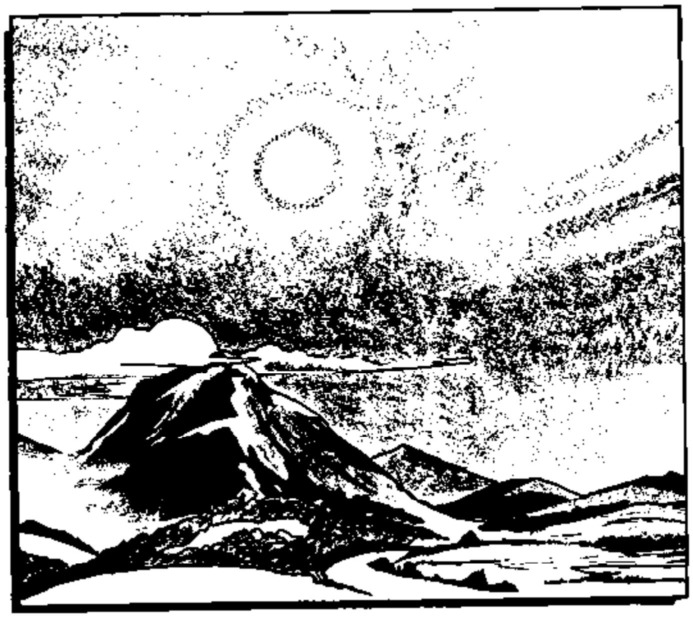

Figure 9-14. Rings Around Moon or Sun.

Chapter 10

GEOGRAPHIC PRINCIPLES

10-1. Introduction. Geographic principles bring together the three major components of the environment (terrain, life forms, and climate) for the purpose of understanding the relationship between survivors and the physical environment around them. The more survivors know about these environmental components, the better they can help themselves in a survival situation. This chapter provides a brief introduction to these complex topics.

10-2. Components of the Environment:

a. Terrain is defined as a geographic area consisting of land and its features. The landmass of the Earth is covered with a variety of topography, including mountains, valleys, plateaus, and plains.

(1) The mountains vary greatly in size, structure, and steepness of slopes. For example, there is as much contrast between the large volcanic Cascade Mountains and those of the Rocky Mountains as there is between the Rocky Mountains and the Appalachian Mountains. Most major mountain systems will have corresponding foothills (figure 10-1).

(2) With two exceptions, valleys are formed as mountains are pushed up. The exceptions are massive gorges formed by glacial action and valleys carved out by wind and water erosion (figure 10-2).

(3) Plateaus are elevated and comparatively large, level expanses of land. Throughout the southwest, examples of the typical plateau can be seen. These plateaus were formed when a volcano deposited either lava or ash over a softer sedimentary area. Through years of erosion, the volcanic "cap" broke loose in places and allowed the softer ground to be carried away. This type of plateau is the least common; however, it is the largest. The Columbia Plateau of Washington State is one example which covers 200,000 square miles (figure 10-3).

(4) The water forms of the Earth include oceans, seas, lakes, rivers, streams, ponds, and ice.

(a) Oceans comprise approximately 70 percent of the Earth's surface. The major oceans include the Pacific, Atlantic, Indian, and Arctic. Oceans have an enormous effect on land, not only in their physical contact but in their effect on weather. In most cases, lakes today are descended from much larger lakes or seas.

(b) Ice covers 10 percent of the Earth's surface. This permanent ice is found in two forms—pack ice and glaciers. Pack ice (normally 7 to 15 feet thick) is frozen sea water and may be as much as 150-feet thick. Those pieces which break off form ice islands. The two permanent icepacks on Earth are found near the North and South Poles—Arctic and Antarctic. The polar regions, which are thousands of feet thick, partially, but never completely, thaw. An icecap is a combination of pack ice and ice sheets. The term is usually applied to an ice plate limited to high mountain and plateau areas. During glacial periods, an icecap will spread over the surrounding lowlands (figure 10-4).

b. Life forms can best be described in terms of vegetation and animal life, with special emphasis on humans (which are covered later).

(1) There are hundreds of plant species on Earth. An in-depth study is obviously impossible. To understand the plant kingdom better, it is important to understand basic plant functions and adaptations they have made to exist in diverse environments. Vegetation will be categorized into either trees or plants.

(a) Of all the variety in species and types, trees can be divided simply into two types: coniferous or deciduous. Conifers are generally considered to be cone-bearing, evergreen trees. Some examples of conifers are pine, fir, and spruce. Deciduous trees are those which lose their leaves in winter and are generally considered as "hardwood." Some examples are maple, aspen, oak, and alder.

(b) For discussion, we will divide plants into two categories: annuals and perennials. Annuals complete their life cycle in 1 year. They produce many seeds and regenerate from seed. Climatic conditions may not be conducive for growth the following year, so seeds may remain dormant for many years. A classic example is the 1977 desert bloom in Death Valley. Plants bloomed for the first time in 80 years. Perennials are plants which last year after year without regeneration from seed.

(2) As with plants, the discussion of animal life has to be limited. Animals will be classified as either warmblooded or coldblooded. Using this division as a basic, it will be easier to describe animal adaptations to extreme climatic conditions. Warmblooded animals are generally recognized as cold-adapted animals and include all birds and mammals. Obviously, humans are a part of this classification because they are cold-adapted. Coldblooded animals gain heat from the environment. These are animals adapted for life in warm or moderate climates (lizards, snakes, etc.)

c. Climate can be described as an average condition of the weather at any given place. However, this description must be expanded to include the seasonal variations and extremes as well as the averages in terms of the climatic elements. In some areas, the climate is so domineering that the corresponding biome is named either in part or as a whole by the climate. Examples are the cloud forests and rain forests. The climate can only be described in terms of its various elements—temperature, moisture, and wind.

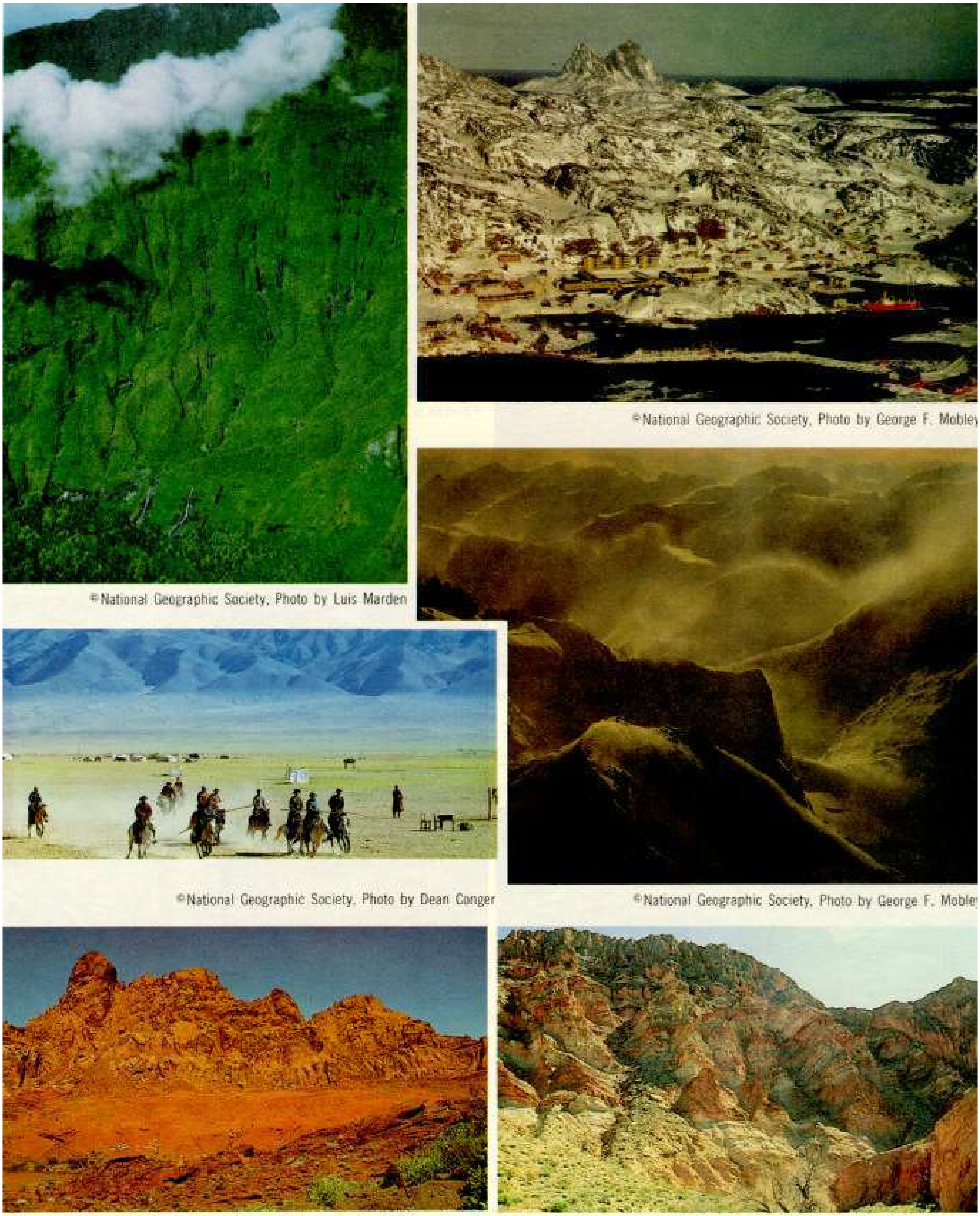

©National Geographic Society, Photo by George F. Mobley

©National Geographic Society, Photo by Luis Marden

©National Geographic Society, Photo by Dean Conger

©National Geographic Society, Photo by George F. Mobley

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowsk

Figure 10-1. Composite of Mountains.

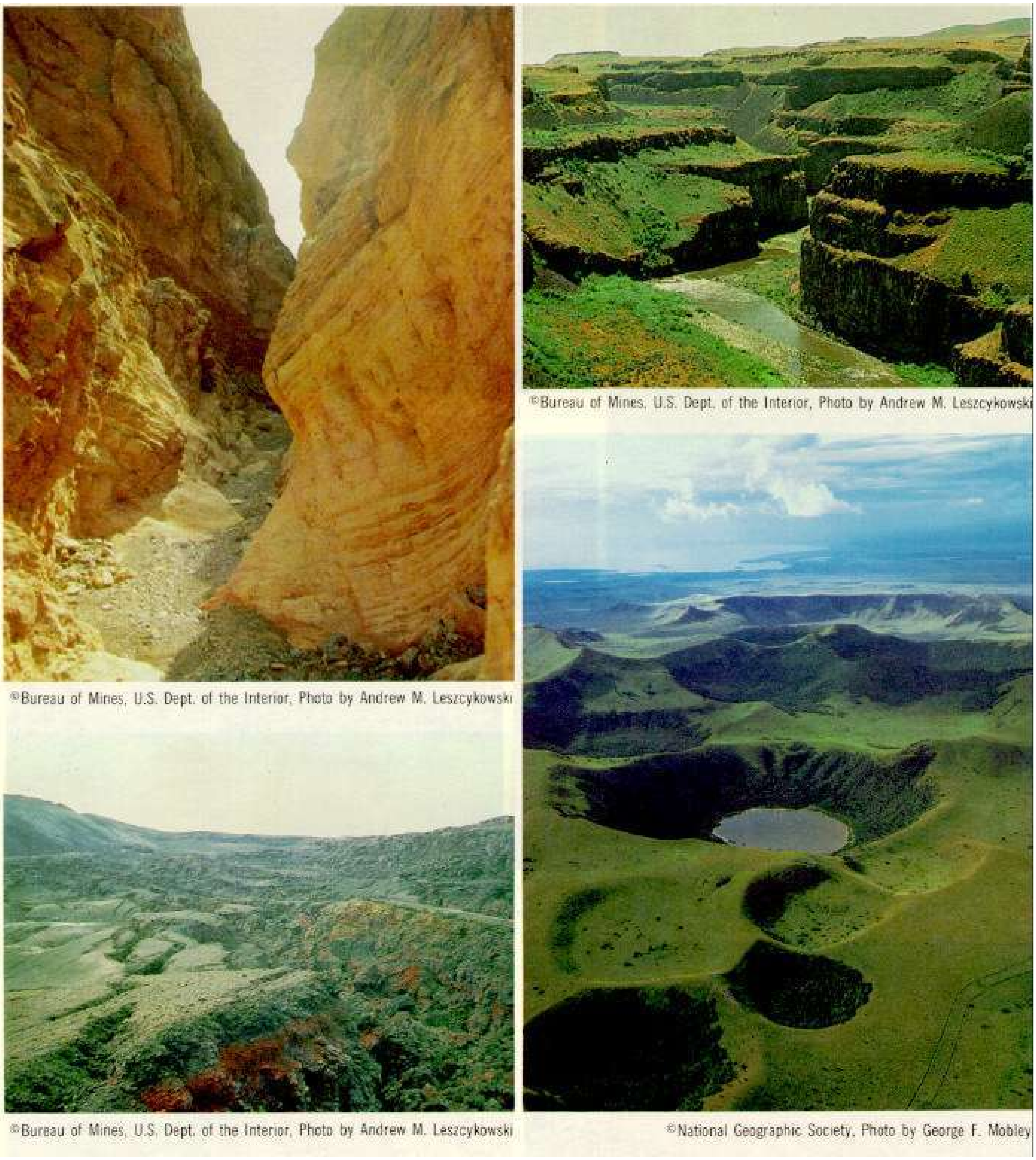

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

©National Geographic Society, Photo by George F. Mobley

Figure 10-2. Gorges and Valleys.

(1) The atmosphere gains only about 20 percent of its temperature from the direct rays of the Sun. Most of the atmospheric heat gain comes from the Earth radiating that heat (energy) back into the atmosphere and being trapped. This is known as the greenhouse effect.

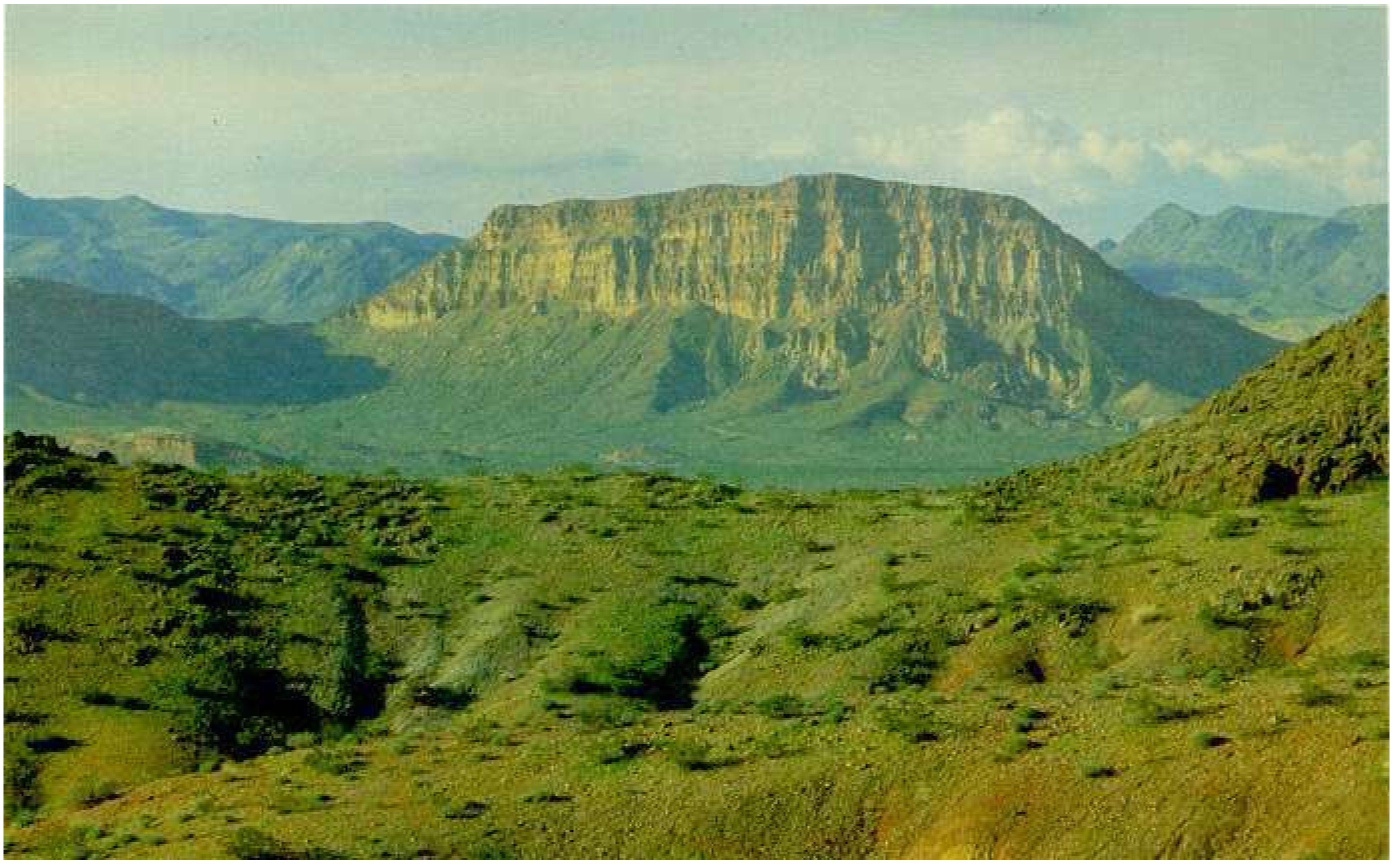

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

Figure 10-3. Plateau.

(2) Thinking of the atmosphere as a greenhouse, it is easier to understand the relationship water has in this "closed system." As water evaporates, the amount of water vapor the atmosphere can absorb depends solely upon temperature. The dew point is achieved when the amount of water vapor in the air equals the maximum volume the air will hold at a given air temperature. Lowering of air temperature creates condensation. Condensation appears in the form of clouds, fog, and dew. Any additional temperature reduction results in precipitation, such as rain. If the temperature of the dew point is below freezing, precipitation may appear in the form of hail or snow.

(3) Variation in air pressure is the primary cause of wind. When air is heated, it creates an area of lower pressure. As air cools, the pressure increases. Air movement occurs as the pressure tries to equalize, thus creating wind. Because wind is also a control of climate, people need to know why and how it affects climate. Let's look at wind in two aspects: localized wind (low altitude) and upper-air wind (high altitude). Localized wind is formed at low altitude, occurring due to dynamic topographical features and fluctuating air temperature and pressure. High-altitude winds surrounding the Earth are bands of stable high- and low-pressure areas (cells). Predictable winds move off these cells which are referred to as jetstreams. These high-altitude winds control weather.

10-3. Effects of Climate on Terrain. The major effect climate has on terrain is erosion. Erosion can occur directly from heavy precipitation or indirectly by the accumulation of snow on snowpacks and glaciers. Wind and temperature both have erosion potential.

a. Heavy precipitation or melting water from icepacks and glaciers can create deep ravines by cutting into mountainous areas. Broad flood basins along major rivers can also aid in the development of river deltas in lakes, oceans, and deep fjords. The action of glaciers throughout the years has carved out deep, broad valleys with steep valley walls (figure 10-5).

b. The effects of wind erosion are greatest in barren, dry areas. The Great Arches National Park has some of the most dramatic examples of the effect of wind erosion. This type of erosion is caused by the wind driving sand and dust particles against an exposed rock or soil

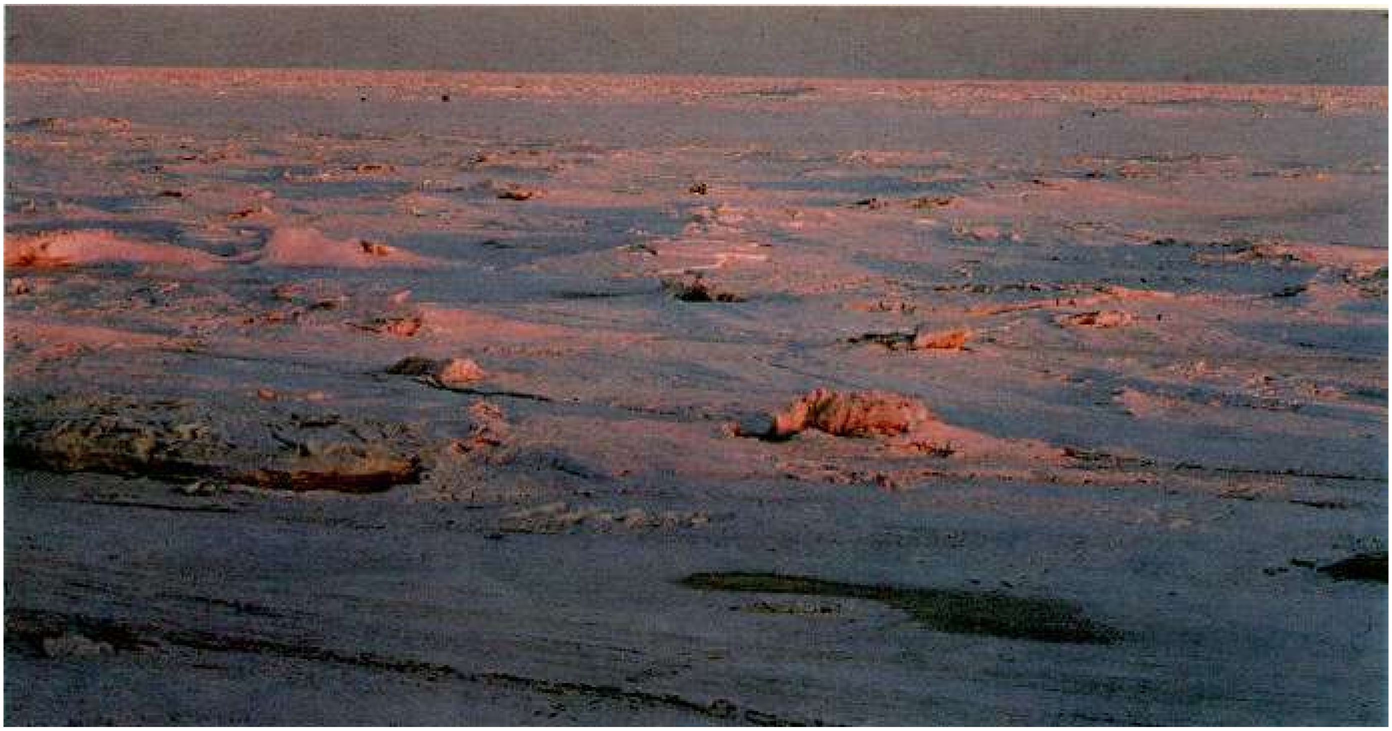

Figure 10-4. Pack Ice.

surface, causing it to be worn away by the impact of the particles in an abrasive action (figure 10-6). Another form of wind erosion involves the movement of loose particles lying upon the ground surface which may be lifted into the air or rolled along the ground. Dry river beds, beaches, areas of recently formed glacial deposits, and dry areas of sandy or rocky ground are highly susceptible to this type of erosion. Sand dunes are attributed to this phenomenon (figure 10-7).

c. Frost action will have a weathering or eroding effect on rock land formations and ground surfaces. The frost action is the repeated growth and melting of ice crystals in the pore space or fractures of soil and rock. The tremendous force of growing ice crystals can exert a pressure great enough to pry apart rock. Many scree and talus slopes are caused by this action. Where soil water freezes, it tends to form ice layers parallel with the ground surface, heaving the soil upward unevenly. The peat moss mounds of the tundra are an example of this action. The net effect of frost action will be dependent on the amount of surface moisture.

10-4. Effects of Terrain on Climate. The effect of terrain on climate is not nearly as subtle as the effect of climate on terrain. Three major factors exist which must be considered when studying the effects of terrain on weather.

a. Moisture for most major weather systems comes from the evaporation of the oceans of the world. The temperature, location, and flow of ocean currents, combined with the prevailing winds will affect how much water will evaporate into the atmosphere. The warmer the ocean and corresponding current, the greater the rate of evaporation. Since the currents are deflected by landmasses, many warm currents flow parallel to major continents. When this moisture is blown inland by the prevailing winds, the net effect is the creation of a wet maritime climate, such as that found along the west coasts of Canada, Washington, Oregon, and Central Europe. If the temperature of the ocean and currents is cold, very little moisture will be yielded to the atmosphere. Examples of this occur along the Pacific coastline of Peru and Chile and along the Atlantic coastline of Angola and Southwestern Africa.

b. The interior of large continent masses are dry because of the distance which isolates them from the effects of maritime climates. The large continents of the Northern Hemisphere create dry, high-pressure cells which isolate the interiors from the lower pressure moist air cells and keep them from having much effect. The climate is referred to as the Continental Climate. The concept will be explained in the next chapter—Environmental Characteristics.

c. Mountains serve as moisture barriers, separating the maritime influenced climates from the continental influenced climates. The barrier effect of mountains on weather will be dependent on the height, length, and width of the range and the severity of the weather

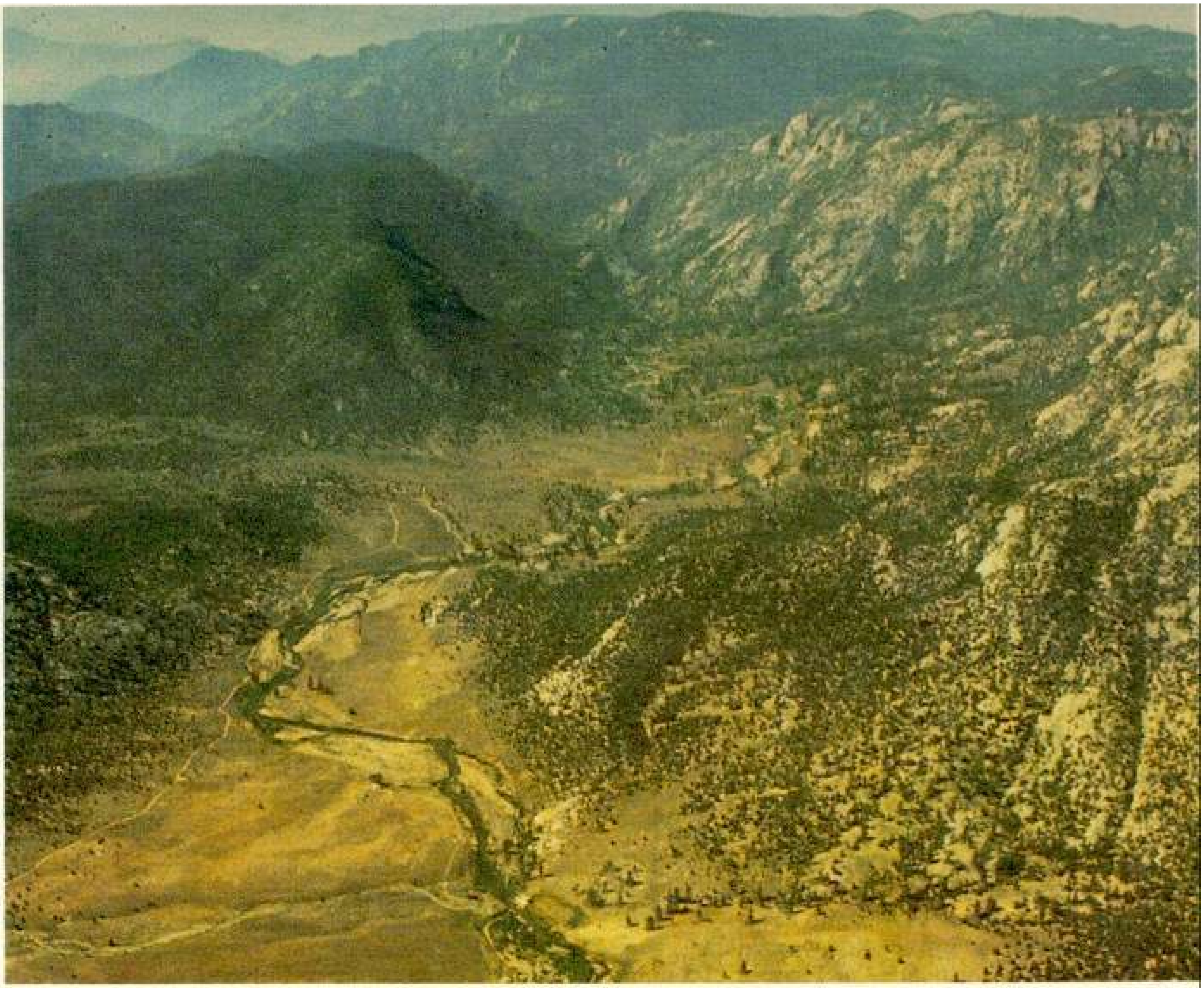

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

Figure 10-5. Valleys.

fronts. In many cases, a lack of precipitation will extend for several hundred miles beyond the mountains. An example of this phenomenon occurs in the western states. The Cascade and Sierra Mountains block a great deal of Pacific Ocean moisture from the the inland deserts of Washington, Oregon, and Nevada. The Rocky Mountains further block most of the moisture which is left in the atmosphere. Only the high cirrus clouds escape the barrier effect of these mountains. Another example can be seen in Asia. The Himalayan Mountains serve as a very effective barrier, blocking the Asiatic monsoon from central interior Asia, which helps create the Gobi Desert.

10-5. Effects of Climate and Terrain on Life Forms:

a. Since plants require water and light, climate will greatly affect the type and number of plants in an area.

(1) In areas with a great deal of rainfall, plants will be plentiful. In these areas, plants must compete for available sunlight. In areas where the primary vegetation has been knocked down (by clear cutting, landslides, or along flood basins of rivers), a thick secondary growth will occur. In time, the secondary growth, if undisturbed, will become a climax forest. Some of the trees in these areas may grow to 300 feet. Because of the shade, vines and shade-tolerant perennials may sparsely cover the ground.

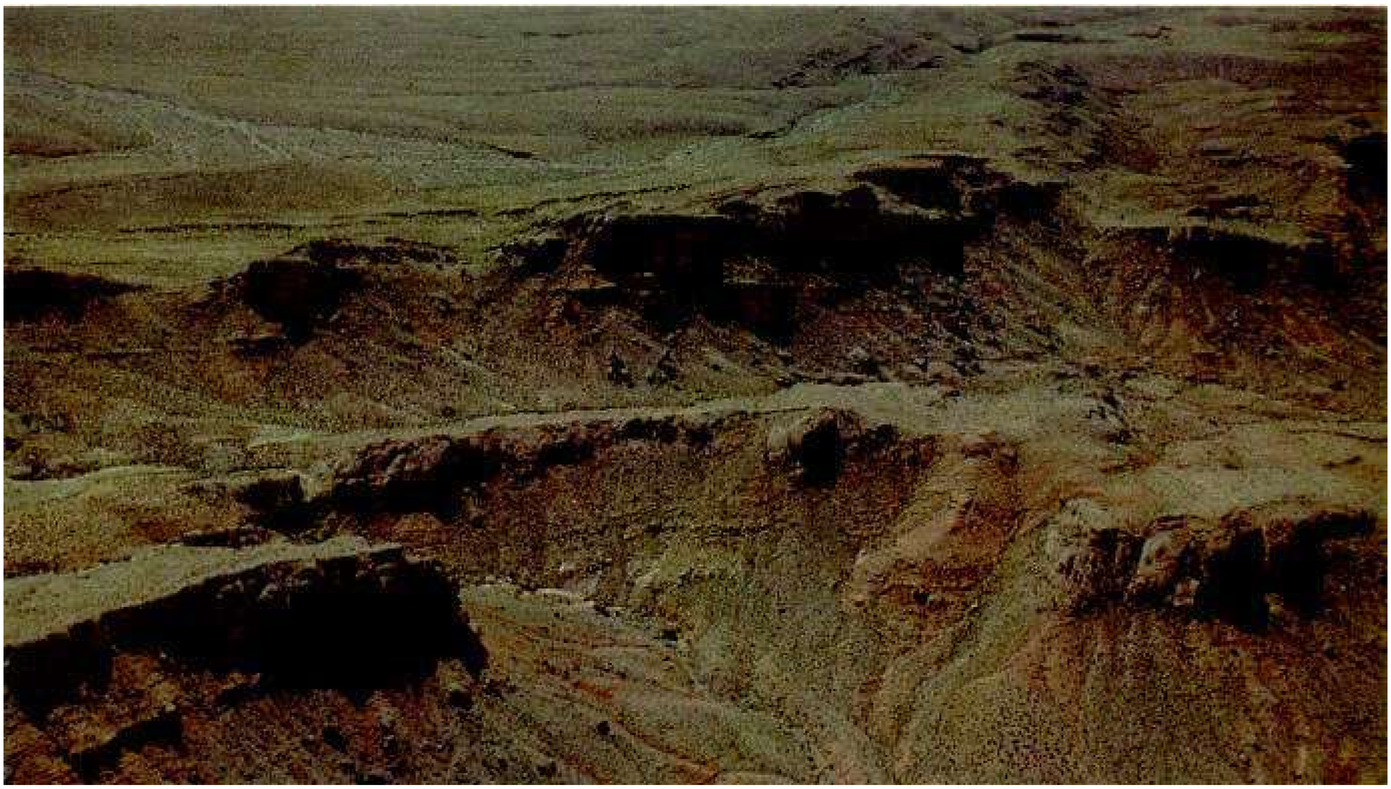

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

Figure 10-6. Erosion.

(2) In contrast, in areas where the amount of rainfall is limited, the plants must compete for the available water. The number of plants will also be sparse. Due to the harsh climatic and soil conditions, plant life is typically hardy. Through millions of years of evolution, plants have developed the following survival characteristics:

(a) Production of many seeds which germinate when water does come.

(b) Shallow root systems gather water quickly when they can.

(c) Ability to store water (cacti and other succulents).

(d) Rough, textured leaves (transpiration).

(e) Production of toxins (kill off competing plants).

b. Vegetation is also affected by the terrain. In mountainous regions, the clouds begin to lose moisture as they pass over the tops of the mountains. The result is more water is available for the growth of vegetation. However, with any increase in elevation, the temperature becomes colder. This exposure to colder temperature has a drastic effect on plant life.

c. Generally, animal life is mostly dependent on two factors: water and vegetation. The greater the rainfall, the greater the number of animals. Conversely, the drier an area, the less vegetation there will be to support animals. The location of small animals is determined by the secondary growth and gound cover, used for protection.

d. Temperature also affects the habits of animals. For example, animals may burrow to protect themselves from extreme heat or extreme cold or will be more active at night in hot, dry regions. Animals also respond physiologically to temperature extremes. During extreme cold, some species of mammals enter into a state of winter dormancy (hibernation). It is a special case of temperature regulation in which animals lower the setting of their "thermostats" to maintain lower than normal body temperatures in order to save energy while maintaining minimum body functions essential for survival. This is important since their normal food supply is not always available during the winter. During periods of excessive heat, some species of fish, reptiles, mammals, and amphibians will enter into a "summer sleep" called estivation. Estivation is a state in which the animal's body functions and activities are greatly reduced. Estivation and hibernation are not merely a result of temperature regulation but rather are methods by which the organisms survive unfavorable periods.

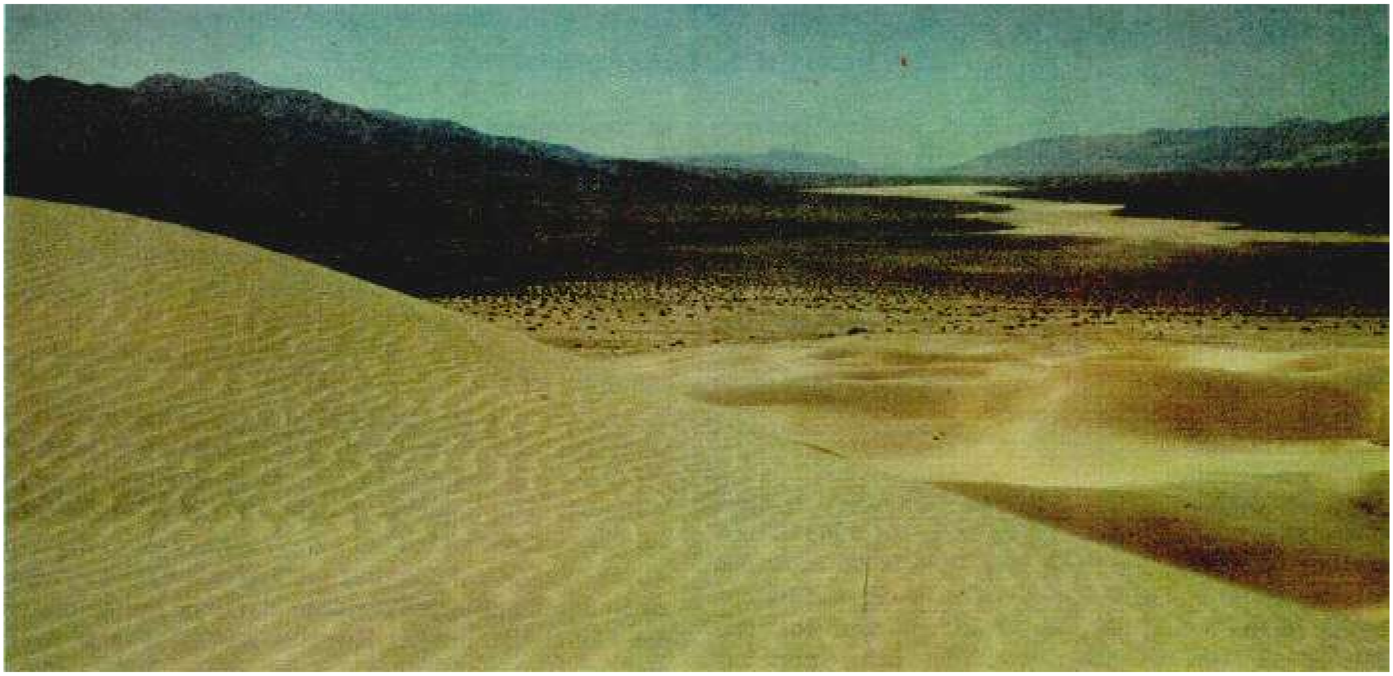

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

Figure 10-7. Sand Dunes.

Chapter 11

ENVIRONMENTAL CHARACTERISTICS

11-1. Introduction. Most survivors will not have a choice as to where they have to survive. The ease or difficulty in maintaining life and honor and returning are dependent on the types and extremes of the climate, terrain, and life forms in the immediate area. The Köppen-Geiger System of Climate Classification will be used in this chapter as the basis for organizing the discussion of environmental characteristics. This system has become the most widely used of climatic classifications for geographical purposes.

11-2. The Köppen-Geiger System. This system defines each climate according to fixed values of temperature and precipitation, computed according to the averages of the year or of individual months. A climate system based on these data has a great advantage in that the area covered by each subtype of climate can be outlined for large parts of the world. The five major climate groups are designated as: (See figure 11-1.)

a. Tropical Climates. Average temperature of each month is above 64.4°F. These climates have no winter season. Annual rainfall is large and exceeds annual evaporation.

b. Dry Climates. Potential evaporation exceeds precipitation on the average throughout the year. No water surplus; hence, no permanent streams originate in dry climate zones.

c. Warm Temperate Climates. Coldest month has an average temperature under 64.4°F, but above 26.6°F. The warm temperate climates thus have both a summer and winter season.

d. Snow Climates. Coldest month average temperature under 26.6°F. Average temperature of warmest month above 50°F.

e. Ice Climates. Average temperature of warmest month below 50°F. These climates have no true summer.

11-3. Tropical Climates:

a. Tropics. Some people think of the tropics as an enormous and forbidding tropical rain forest through which every step taken must be hacked out and where every inch of the way is crawling with danger. Actually, much of the tropics is not rain forest. What rain forest there is must be traveled with some labor and difficulty. The tropical area may be rain forest, mangrove or other swamps, open grassy plains, or semi-dry brushland. The tropical area may also have deserts or cold mountainous districts. There is in fact, a variety of tropical climates. Each region, while subject to the general climatic condition of its own zone, may show special modifications locally. Each general climate is a whole range of basic minor climates. In all their diversity, the climates of the tropics have the following in common:

(1) An almost constant length of day and night, a length that varies by no more than half an hour at the Equator to 1 hour at the limits of the tropics. The plant life thus has an evenly distributed period of daylight throughout the year.

(2) Temperature variation throughout the tropics is minimal—9°F to 18°F.

(3) There is no systematic pattern of major tropical landforms. There are high rugged mountains; such as the Andes of South America, karst formations as in Southeast Asia, plateaus like the Deccan of India, hilly lands like those which back the Republic of Guinea in Africa, and both large and small plains like the extensive one of the upper Amazon River or the restricted plain of the Irrawaddy River in Burma. The arrangement of all these landforms is part of the pattern of the larger landmasses, not of the tropics alone.

b. Vegetation:

(1) The jungles in South America, Asia, and Africa are more correctly called tropical rain forests. These forests form a belt around the entire globe, bisected somewhat equally by the Equator. However, the tropical rain forest belt is not a continuous one, even in any of the various regions in which it occurs. Usually it is broken by mountain ranges, plateaus, and even by small semi-desert areas, according to the irregular pattern of climate which regulates the actual distribution of rain forest.

(2) Some of the leading characteristics of the tropical rain forest common to those areas in South America, in Asia, and in Africa, are:

(a) Temperatures average close to 80°F for every month.

(b) Vegetation consists of three stories.

(c) High rainfall (80 inches or more) distributed fairly evenly throughout the year.

(d) Areas of occurrence lie between 23.5 North and 23.5 South Latitudes.

(e) Evergreen trees predominate; many of large girth up to 10 feet in diameter, with thick leathery leaves.

(f) Vines (lianas) and air plants (epiphytes) are abundant.

(g) Herbs, grasses, and bushes are rare in the understory.

(h) Uniformity.

(i) Tree bark thin, green, smooth, and usually lacking fissures.

(3) The majority of plants that grow in the forest of the rainy tropics are woody and of the dimensions of trees. Trees form the principal elements of the vegeta-

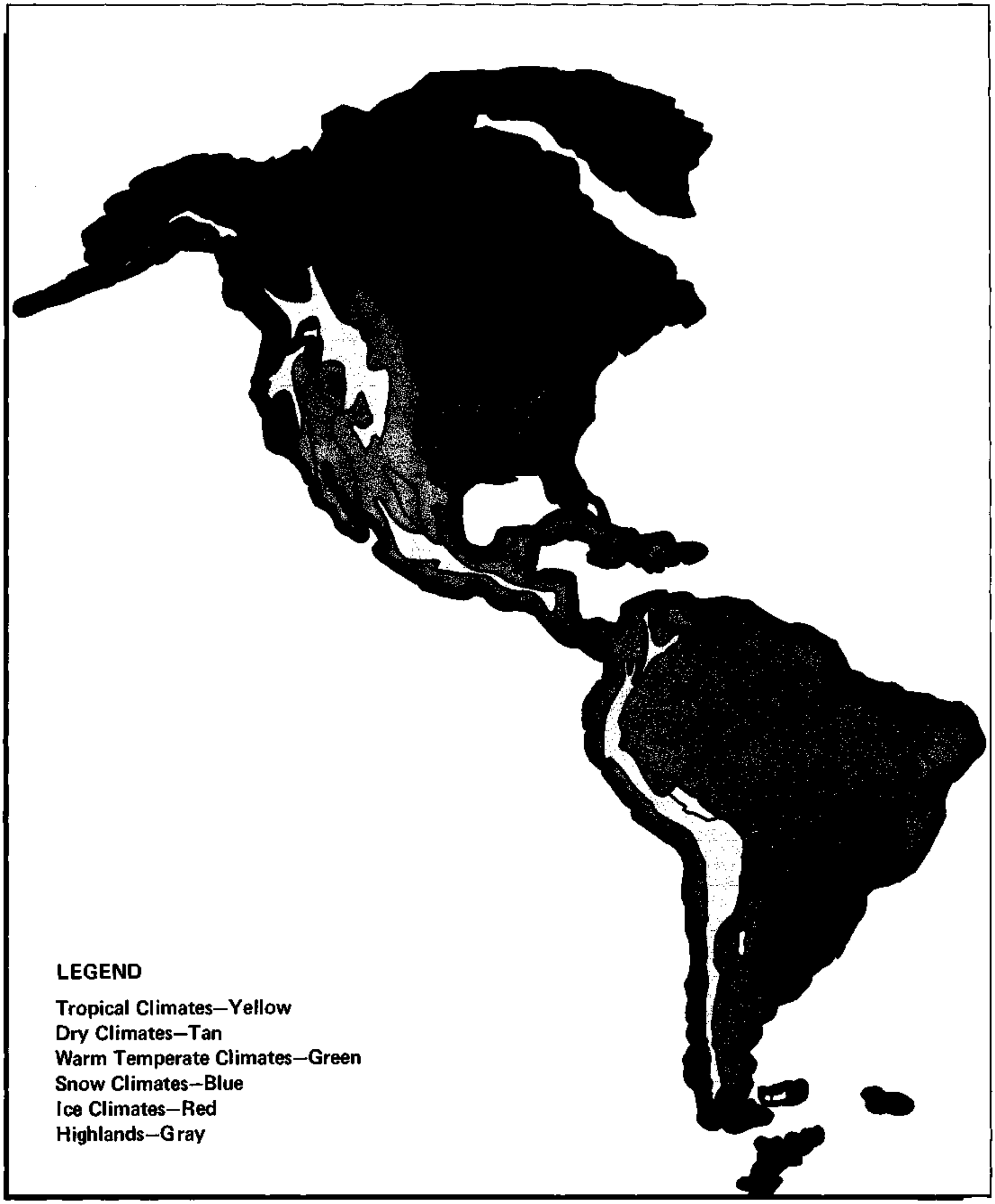
LEGEND
Tropical Climates—Yellow
Dry Climates—Tan
Warm Temperate Climates—Green
Snow Climates—Blue
Ice Climates—Red
Highlands—Gray

Figure 11-1. Köppen-Geiger System.

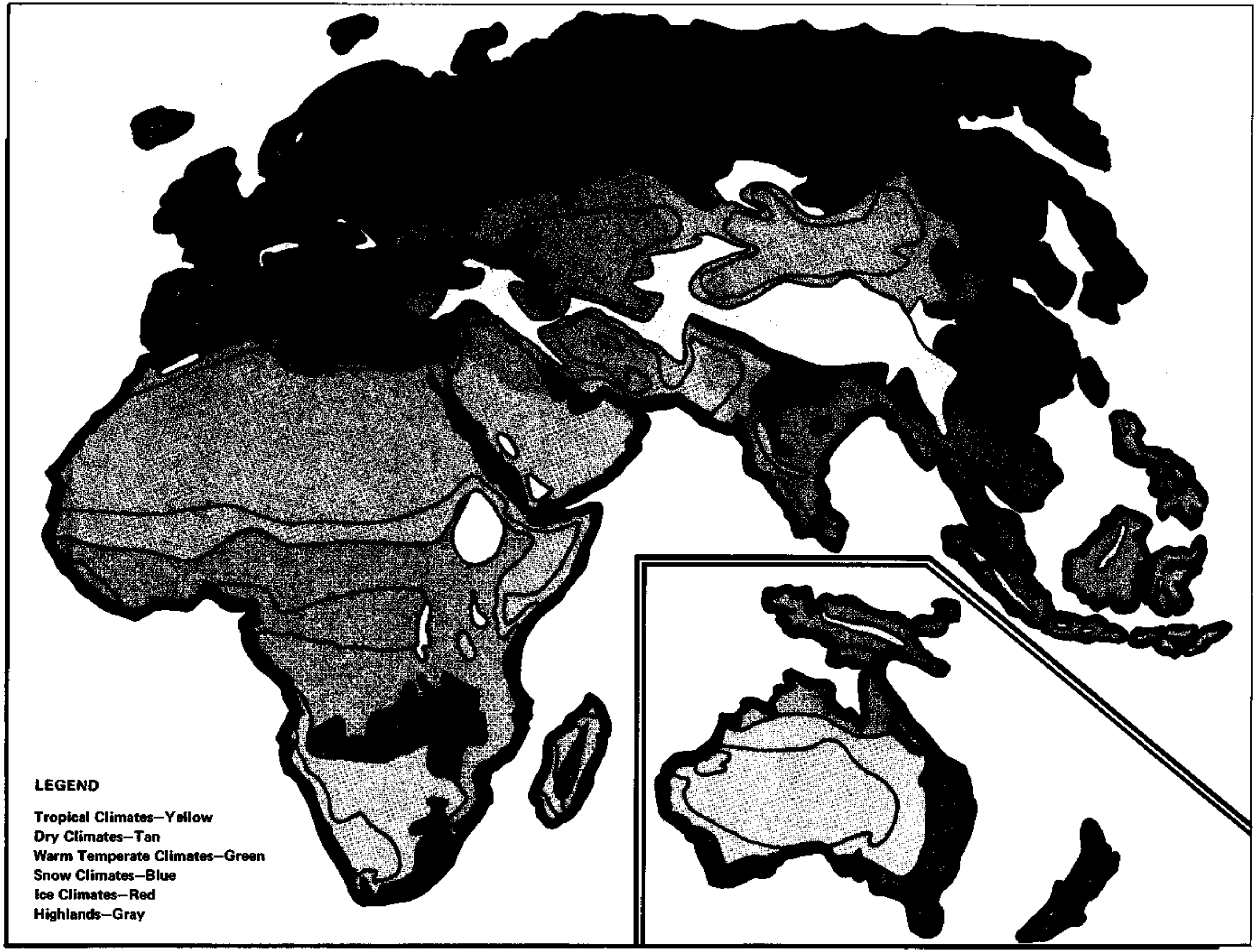
LEGEND
Tropical Climates—Yellow
Dry Climates—Tan
Warm Temperate Climates—Green
Snow Climates—Blue
Ice Climates—Red
Highlands—Gray

Figure 11-1. Köppen-Geiger System. (cont)

tion. The vines and air plants that grow on the trunks and branches of trees are woody. Grasses and herbs, which are common in the temperate woods of the United States, are rare in the tropical rain forest. The undergrowth consists of woody plants—seedling and sapling trees, shrubs, and young woody climbers. The bamboos, which are really grasses, grow to giant proportions, 20 to 80 feet high in some cases. Bamboo thickets in parts of some rain forests are very difficult to penetrate. The plants that produce edible parts in the jungle are often scattered, and require searching to find several of the same kind. A tropical rain forest (figure 11-2) has a wider variety of trees than any other area in the world. Scientists have counted 179 species in one 8.5-acre area in South America. An area this size in a forest in the United States would have fewer than seven species of trees.

(4) The average height of the taller trees in the rain forest is rarely more than 150 to 180 feet. Old giants of the tropical rain forest attain 300 feet in height, but this is extremely rare. Trees more than 10 feet in diameter are also rare in the jungle. The trunks are, as a rule, straight and slender and do not branch until near the top. The base of many trees is provided with plank buttresses, flag-like outgrowths which are common in all tropical forests. The majority of mature tropical trees have large, leathery, dark-green leaves which resemble laurel leaves in size, shape, and texture. The general appearance is monotonous, and large and strikingly colored flowers are uncommon. Most of the trees and shrubs have inconspicuous flowers, often greenish or whitish.

(5) Travel books often give a misleading impression of the density of tropical forests. On riverbanks or in clearings, where much light reaches the ground, there is a dense growth which is often quite impenetrable. But in the interior of an old undisturbed forest, it is not difficult to walk in any direction. Photographs give an exaggerated notion of the density of the undergrowth. It is usually possible to see another person at least 60 feet away.

(6) The abundance of climbing plants is one of the characteristic features of rain forest vegetation. The great majority of these climbers are woody and many have stems of great length and thickness. Stems as thick as a man's thigh are not uncommon. Some lianas cling closely to the trees that support them, but most ascend to the forest canopy like cables or hang down in loops or

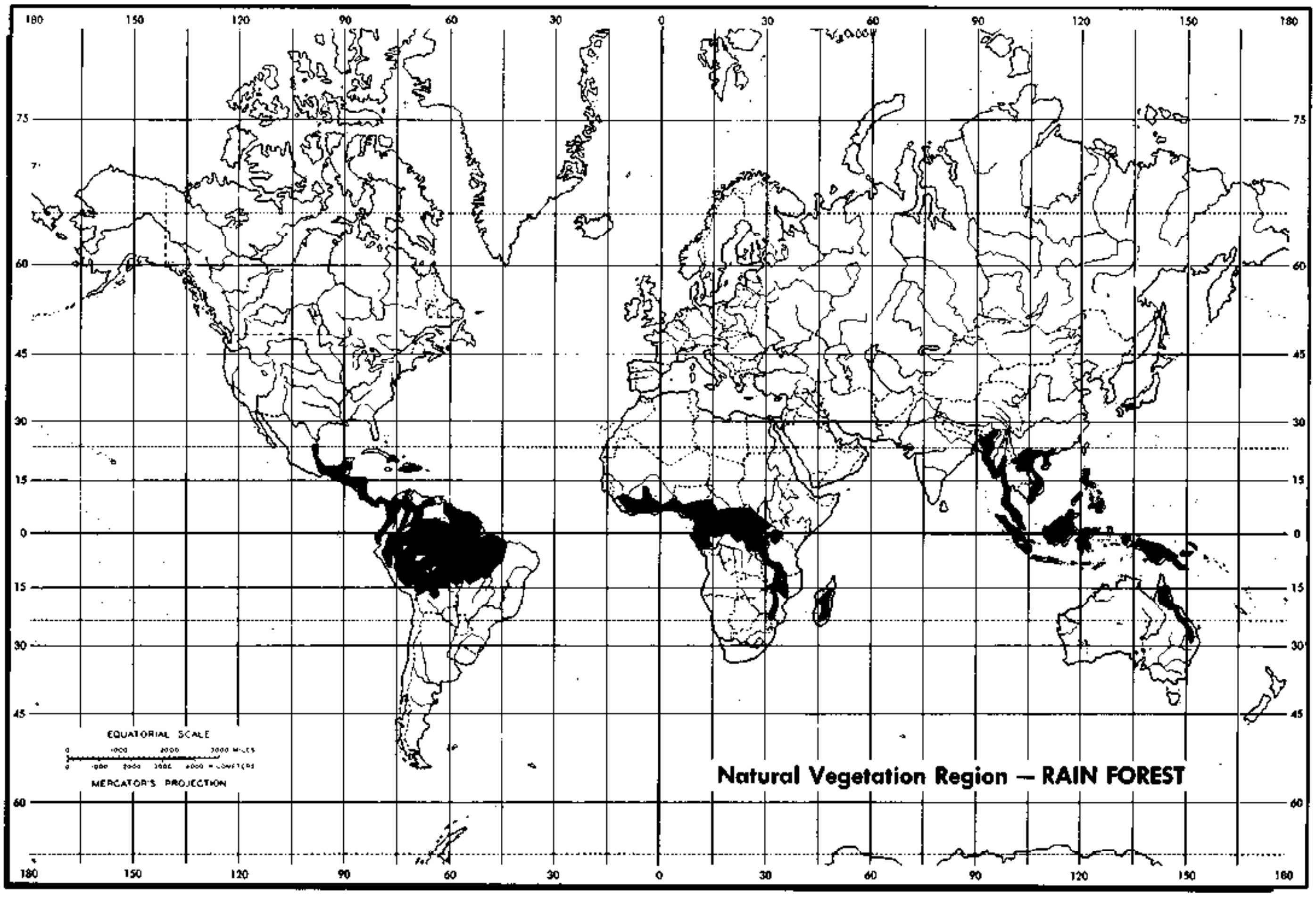

Figure 11-2. Rain Forest.

festoons. In the rain forest, there is no winter or spring, only perpetual midsummer. The appearance of the vegetation is much the same at any time of year. There are seasons of maximum flowering during which more species bloom than at any other time, and also seasons of maximum production of young leaves, plant growth, and reproduction is continuous and some flowers can be found at any time. The margins of a tropical rain forest clearing and areas around abandoned dwellings abound in edible plants. However, in the center of the virgin rain forest, trying to find food is more difficult due to accessibility. The lofty trees are so tall that fruits and nuts are generally out of reach.

c. Distribution of Tropical Rain Forests:

(1) In the Americas, the largest continuous mass of rain forest is found in the basin of the Amazon River. This extends west to the lower slopes of the Andes and east to the Atlantic coast of the continent; it is broken only by relatively small areas of savanna and deciduous forest. This great South American rain forest extends south into the region of the Gran Chaco (south-central South America) and north along the eastern side of Central America into southern Mexico and into the Antilles chain of the West Indies. In the extreme northwest of South America (Ecuador, Colombia), there is a narrow belt of rain forest, separated from the Amazonian forest by a wide expanse of deciduous forest, extending from about latitude 6 degrees South to a little beyond the Tropic of Capricorn. The distribution of rain forest in Central America is perhaps less well known than any other major tropical region. The main areas are below the 500-foot elevation (figure 11-3).

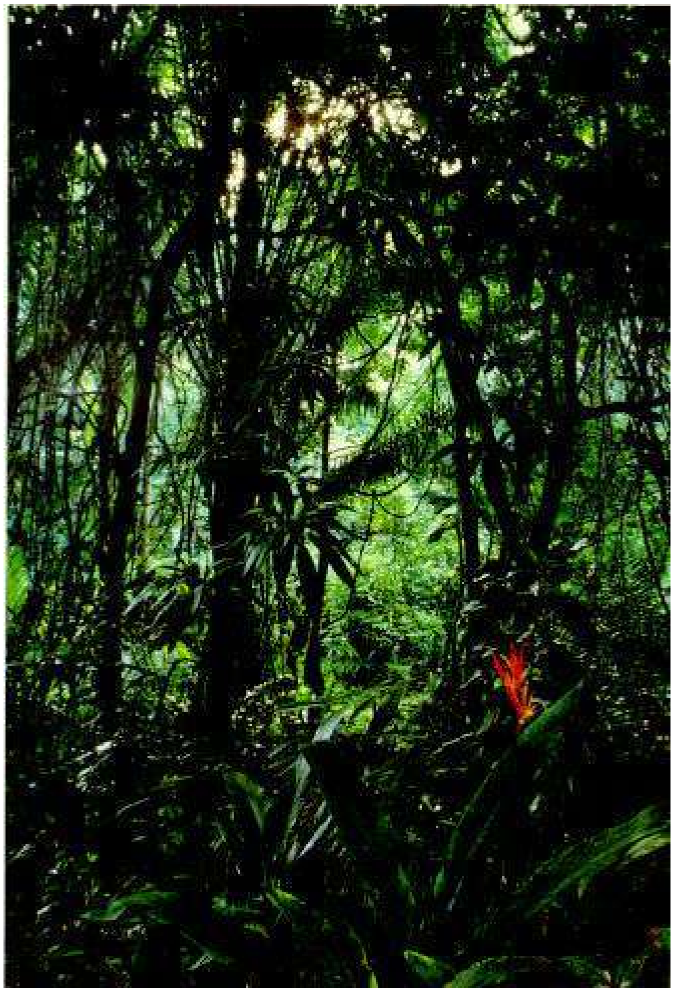

©National Geographic Society, Photo by James Stanfield

Figure 11-3. Rain Forest (El Salvador).

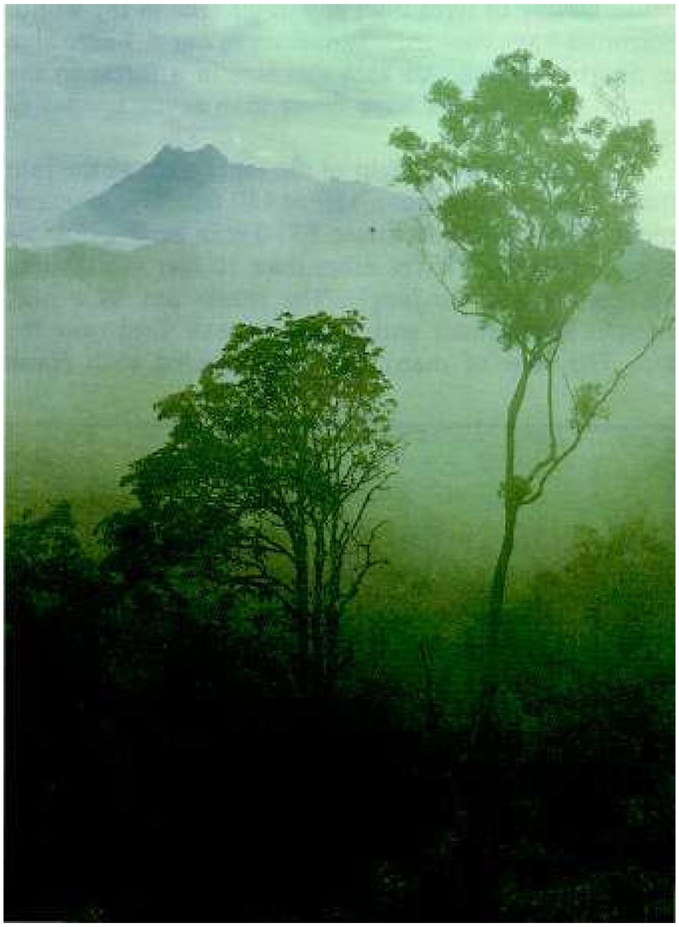

©National Geographic Society, Photo by Winfield Parks

Figure 11-4. Rain Forest (Borneo).

(2) In Africa, the largest area of rain forest lies in the Congo basin and extends westward into the Republic of Cameroon. As a narrow strip, the forest continues still farther west, parallel to the Gulf of Guinea, through Nigeria and the Gold Coast of Liberia and Guinea. Southward from the Congo basin, the forest extends towards Rhodesia.

(3) In the eastern tropics, the rain forest extends from Ceylon and Western India to Southeast Asia and the Philippines, as well as through the Malay Archipela-

go to New Guinea. The largest continuous areas are in New Guinea, the Malay Peninsula, and the adjoining islands of Sumatra and Borneo, where the Indo-Malayan rain forest reaches its greatest luxuriance and floral wealth (figure 11-4).

(a) In India, the area of rain forest is not large, but it is found locally in the western and eastern Ghats (coastal ranges) and, more extensively, in the lower part of the eastern Himalayas, the Khasia hills, and Assam.

(b) In Burma and Southeast Asia, the rain forest is developed only locally; the principal vegetation being the monsoon forest. The monsoon forest is a tropical type which is partly leafless at certain seasons.

(c) In the eastern Sunda Islands from western Java to New Guinea, the seasonal drought (due to the dry east monsoon from Australia) is too severe for the development of a rain forest, except in locally favorable situations (figure 11-5).

(d) In Australia, the tropical rain forest of Indo-Malaya is continued south as a narrow strip along the eastern coast of Queensland. Rain forest also extends into the islands of the western Pacific (Solomons, New Hebrides, Fiji, Samoa, etc.). (See figure 11-6.)

d. Food Plants. Some of the available food plants in the rain forest include:

(1) Indian or Tropical Almond.
(2) Rose Apple.
(3) East Indian Arrowroot.
(4) Bullock's Heart.
(5) Sugar Cane.
(6) Cattail.
(7) Bael Fruit.
(8) Water Chestnut (Trapa Nut).
(9) Bamboo.
(10) Chufa (Nut Grass).
(11) Goa Bean.
(12) Luffa Sponge (Wild Gourd).
(13) Yam Bean.
(14) Wild Fig.
(15) Bignay.
(16) Wild Grape.
(17) Lotus Lilly.
(18) Water Lettuce.
(19) Breadfruit.
(20) Canna Lily.
(21) Bracken (Fern).
(22) Sego Palm.
(23) Tree Fern.
(24) Palm Sugar.
(25) Mango.
(26) Papaya.
(27) Italian Millet.
(28) Screw Pine.
(29) Pearl Millet.
(30) Plantain.
(31) Mulberry.
(32) Batolo Plum.
(33) Cashew Nut.
(34) Pokeweed.
(35) Buri Palm.
(36) Polypody.
(37) Fishtail Palm.
(38) Air Potato (ubi tuber).
(39) Coconut Palm.
(40) Purslane.
(41) Nipa Palm.
(42) Rice.
(43) Rattan Palm.
(44) Soursop.
(45) Ceylon Spinach.
(46) Water Lilly.
(47) Sterculia.
(48) Sweetsop.
(49) Tamarind.
(50) Taro.
(51) Ti Plant.
(52) Horseradish Tree.
(53) Tropical Yam.

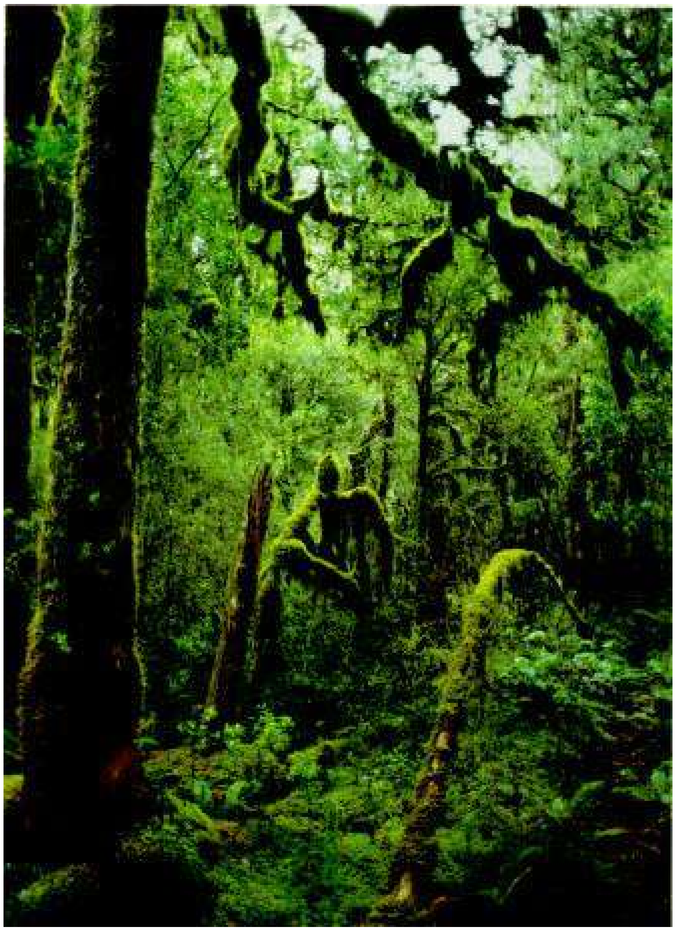

©National Geographic Society, Photo by Robert E. Gilka

Figure 11-5. Rain Forest (New Zealand).

©National Geographic Society, Photo by Luis Marden

Figure 11-6. Rain Forest (Tahiti).

e. Semi-Evergreen Seasonal Forest:

(1) In character, the semi-evergreen seasonal forest in Central and South America and Africa corresponds essentially to the monsoon forest of Asia. Characteristics of the semi-evergreen forest are:

(a) Two stories of tree strata—upper story 60 to 80 feet high; lower story 20 to 45 feet high.

(b) Large trees are rare; average diameter about 2 feet.

(c) Seasonal drought causes leaf fall; more in dry years.

(2) The peculiar distribution of the rainy season and the dry season which occurs in the countries bordering the Bay of Bengal in southeastern Asia brings on the monsoon climate. The monsoons of India, Burma, and Southeast Asia are of two types. The dry monsoon occurs from November to April, when the dry northern winds from central Asia bring long periods of clear weather with only intermittent rain. The wet monsoon occurs from May to October, when the southern winds from the Bay of Bengal bring rain, usually in torrents, that lasts for days and often weeks at a time. During the dry season, most leaves drop completely off, giving the landscape a wintry appearance, but as soon as the monsoon rains begin, the foliage reappears immediately.

f. Plant Foods of the Semi-Evergreen Seasonal Forest. These foods include:

(1) Agave (Century Plant).
(2) Amaranth.
(3) Bael Fruit.
(4) Banana.
(5) Tropical or Indian Almond.
(6) Rose Apple.
(7) Bamboo.
(8) Goa Bean.
(9) Yam Bean.
(10) Mango.
(11) Purslane.
(12) Mulberry.
(13) Bignay.
(14) Italian Millet.
(15) Soursop.
(16) Cashew Nut.
(17) Breadfruit.
(18) Pearl Millet.
(19) Ceylon Spinach.
(20) Sterculia.
(21) Sugar Cane.
(22) English Acorns (Oak).
(23) Luffa Sponge (Wild Gourd).
(24) Cattail.
(25) Buri Palm.
(26) Sweetsop.
(27) Chestnut.
(28) Water Chestnut (Trapa Nut).
(29) Rattan Palm.
(30) Tamarind.
(31) Chufa (Nut Grass).
(32) Papaya.
(33) Taro.
(34) Ti Plant.
(35) Wild Fig.
(36) Screw Pine.
(37) Horseradish Tree.
(38) Tree Fern.
(39) Plaintain.
(40) Tropical Yam.
(41) Wild Grape.
(42) Pokeweed.
(43) Water Lettuce.
(44) Polypody.
(45) Canna Lily.
(46) Air Potato (ubi tuber).
(47) Lilly Lotus.
(48) Water Lily.

g. Tropical Scrub and Thorn Forest (figure 11-7):

(1) Chief Characteristics of the Tropical Scrub and Thorn Forest:

(a) Definite dry season, with wet season varying in length from year to year. Rains appear mainly as downpours from thunderstorms.

(b) Trees are leafless during dry season; average height is 20 to 30 feet with tangled undergrowth in places (figure 11-8).

(c) Ground is bare except for a few tufted plants in bunches; grasses are not common.

(d) Plants with thorns are predominate.

(e) Fires occur at intervals.

(2) Food Plants:

(a) Within the tropical scrub and thorn forest areas, the survivor will find it difficult to get food plants in the dry season. During the height of the drought period, the primary kinds of foods come from the following plant parts:

-1. Tubers.
-2. Bulbs.
-3. Pitch.
-4. Nuts.
-5. Rootstalks.
-6. Corms.
-7. Gums and Resins.
-8. Seeds and Grains.

(b) During the rainy season in the tropical scrub and thorn forest, plant food is considerably more abundant. At this time, the survivor should look for the following edible plants:

-1.Sweet Acacia.
-2.Wild Chicory.
-3.St. John's Bread.
-4.Wild Caper.
-5.Agave (Century Plant).
-6.Wild Fig.
-7.Almond.
-8.Cashew Nut.
-9.Baobab.
-10.Juniper.
-11.Tamarind.
-12.Tropical Yam.
-13.Sea Orach.
-14.Prickly Pear.
-15.Wild Pistachio.
-16.Air Potato (ubi tuber).

h. Tropical Savanna. (See figure 11-9.)

(1) General Characteristics of the Savanna:

(a) Savannas lie wholly within the tropical zone in South America and Africa.

(b) The savanna looks like a broad, grassy meadow with trees spaced at wide intervals.

(c) The grasses of the tropical savanna often exceed the height of a man. However, none of the savanna grasses are sod-forming in the manner of lawn grasses, but are bunch grasses with a definite space between each grass plant.

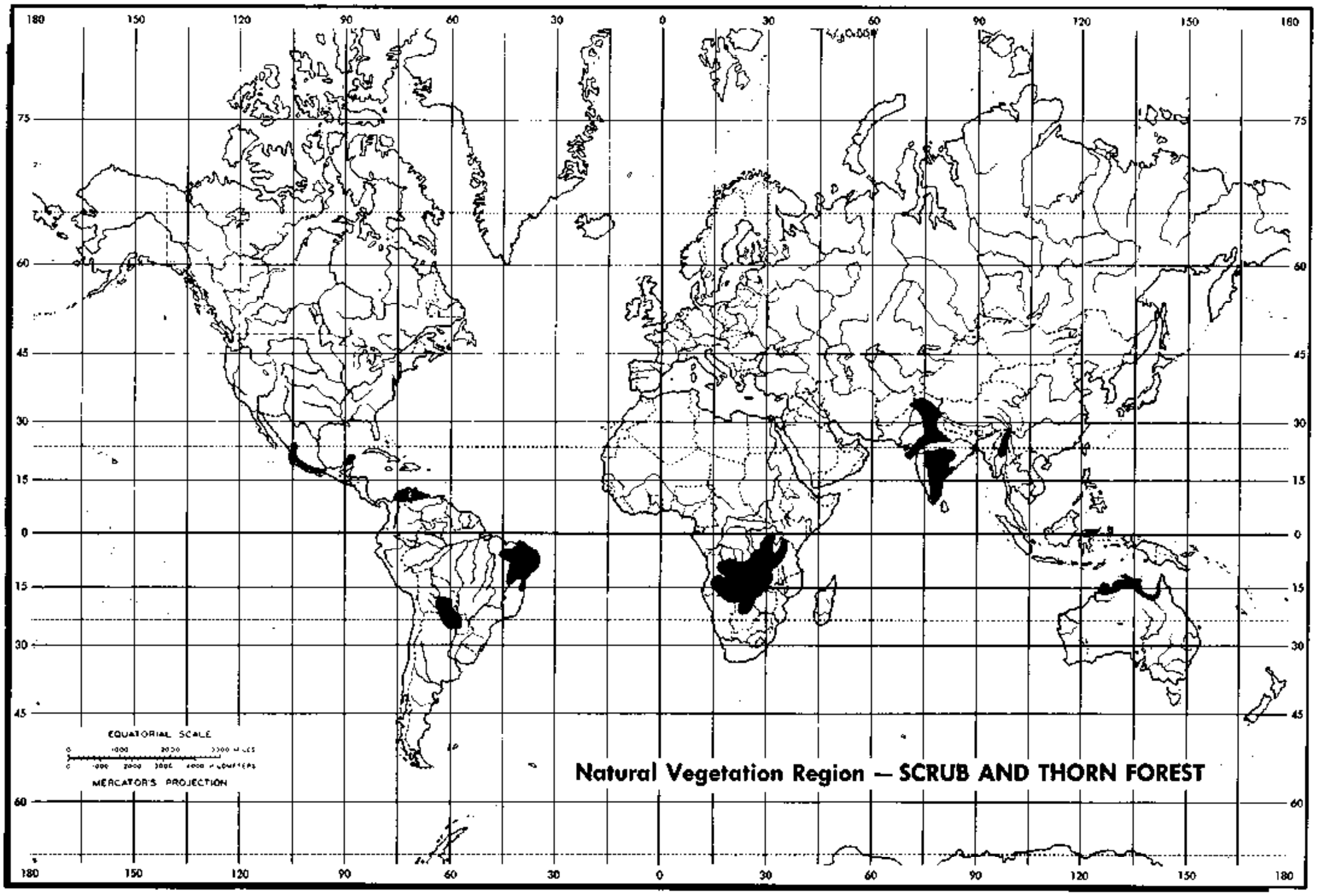

Figure 11-7. Shrub and Thorn Forest.

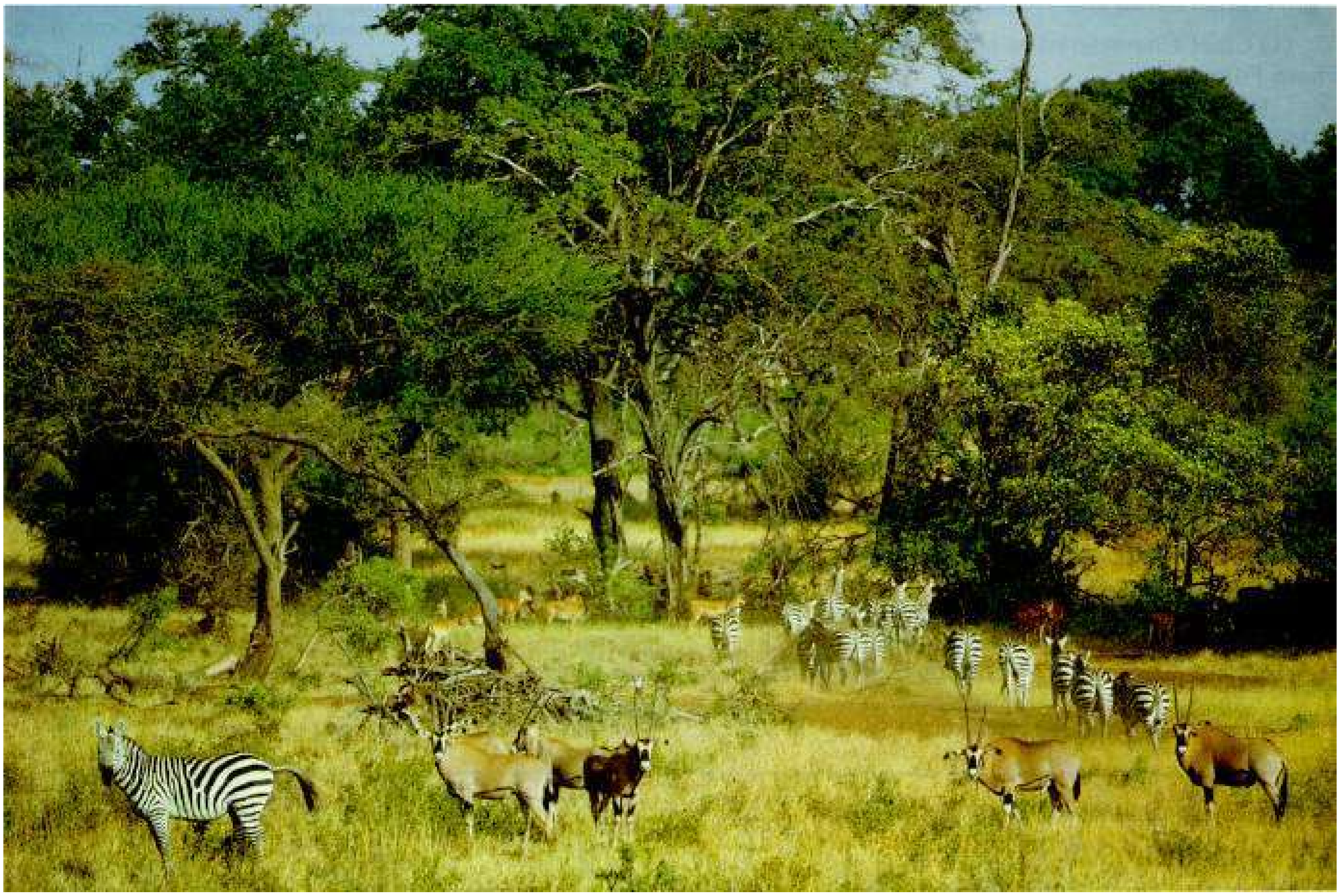

©National Geographic Society, Photo by Gilbert M. Grosvenor

Figure 11-8. Tropical Shrub or Thorn Forest (Kenya).

(d) The soil in the savanna is frequently red.

(e) The scattered trees usually appear stunted and gnarled like old apple

(f) Palms may be found on savannas.

(2) Savanna of South America. For the most part, the vegetation is of the bunch-grass type. A long, dry season alternates with a rainy season. In these areas, both high and low grasses are present. Bright colored flowers appear between the grass bunches during the rainy season. The grains from the numerous grasses are useful as survival food, as well as the underground parts of the many seasonal plants that appear with and following the rains.

(3) Savanna in Africa. The high grass tropical savanna of Africa is dominated by very tall, coarse grasses which grow from 5 to 15 feet high. Unless the natives burn the grass during the dry season, the savanna becomes almost impenetrable. This type of savanna occurs in a broad belt surrounding the tropical rain forest and extends from western Africa eastward beyond the Nile River. From the Nile, it extends southward and westward. The tropical bunch grass savanna comprises the greatest part of the African savanna consisting of grasses about 3 feet tall. The African savanna has both dwarf and large trees. The most renowned of these large trees is the monkeybread or baobab (figure 11-10).

(4) Food Plants. The food plants found on the savanna are also found in other vegetation areas.

(a) Amaranth.
(b) Wild Crabapple.
(c) Purslane.
(d) Wild Apple.
(e) Wild Dock.
(f) Wild Sorrel.
(g) Wild Chicory.
(h) Wild Fig.
(i) Tamarind.
(j) Chufa (Nut Grass).
(k) Water Plaintain.
(l) Water Lily.

i. Animal Life. The tropics abound in animal life. The tremendous varieties of animal species found in tropical areas throughout the world preclude discussions of each animal. It is essential that the survivor realize that just as people have an inherent fear of some animals, most animals also fear people. With some exceptions, animals of the tropics will withdraw from any encounter with humans. Being primarily nocturnal animals, most

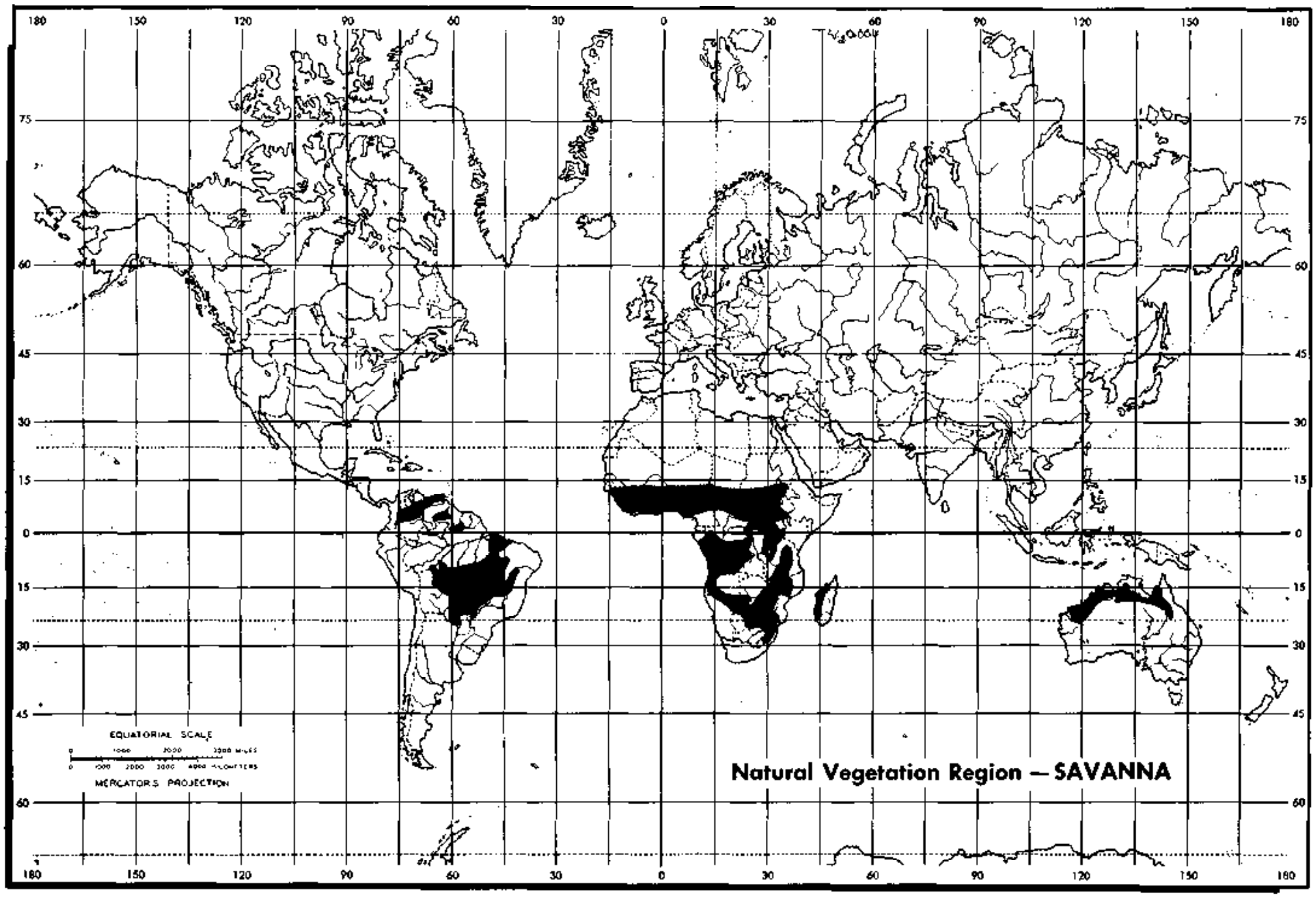

Figure 11-9. Savanna.

will never be seen by the survivor. By becoming familiar with the wild inhabitants of the tropics, the survivor will better understand this type of environment and will respect, not fear, the surroundings in which survival takes place.

(1) All tropic areas have members of the pig family. By habit, pigs are gregarious and are omnivorous in diet. They will eat any small animals they can kill, although they feed mainly on roots, tubers, and other vegetable substances. The most common species found in the Old World tropics are the peccary, the Indian wild boar, the Babirussa of Celebes, and the Central African Giant Forest Hog. In Central and South American tropics, peccaries are common. These pigs are represented by two species, the "white-lipped" peccary and the "collared" peccary. Both are grizzled black color, distinguishable by markings from which they derive their names. The white-lipped peccary, the larger of the two (height of approximately 18 inches), is black with white under the snout, and has the reputation of being the more ferocious. The collared peccary, reaching a height of 14 inches, is identified by the white or gray band around the body where the neck joins the shoulder. The collared peccary often travels in groups of 5 to 15. While alone, they are not particularly dangerous, but a pack can effectively repel any enemy and can make short work of a jaguar, cougar, or human. Both types of peccaries have musk glands which are located 4 inches up from the tail on the spine. This gland must be removed soon after the animal is killed, otherwise the flesh will become tainted and unfit for consumption.

(2) Tropic areas harbor many species of reptiles and amphibians. Most of them are edible when skinned and cooked. Hazards from these animals are mostly imagined; however, some are venomous or dangerous if encountered. Individual species of crocodylidae family (alligators, crocodiles, caiman, and gavials) are usually only abundant in remote areas away from humans. Most dangerous are the saltwater crocodiles of the Far East and the Nile crocodiles in Africa. Poisonous snakes, while numerous in the tropics, are rarely seen and pose little danger to the wary survivor. There are no known poisonous lizards in the tropics. Several species of frogs and toads contain poisonous skin secretions.

PLAINS OF TANGANYIKA

©National Geographic Society, Photo by W. Robert Moore

KENYA

©National Geographic Society, Photo by Melvin M. Payne

NAIROBI, KENYA

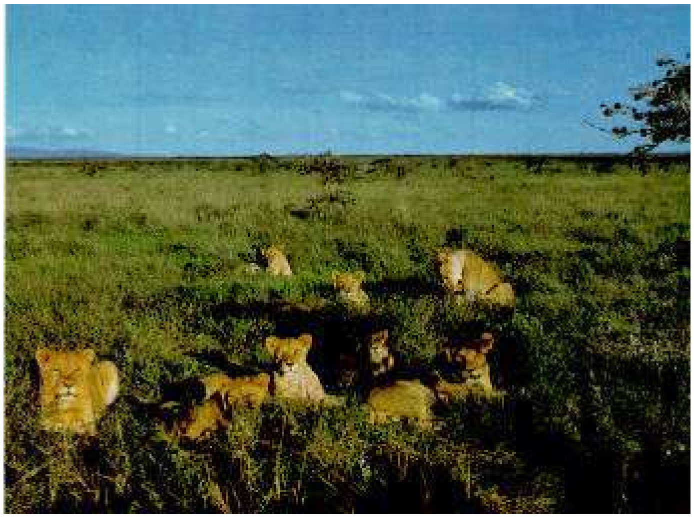

©National Geographic Society, Photo by W. Robert Moore

IHOSY, MADAGASCAR

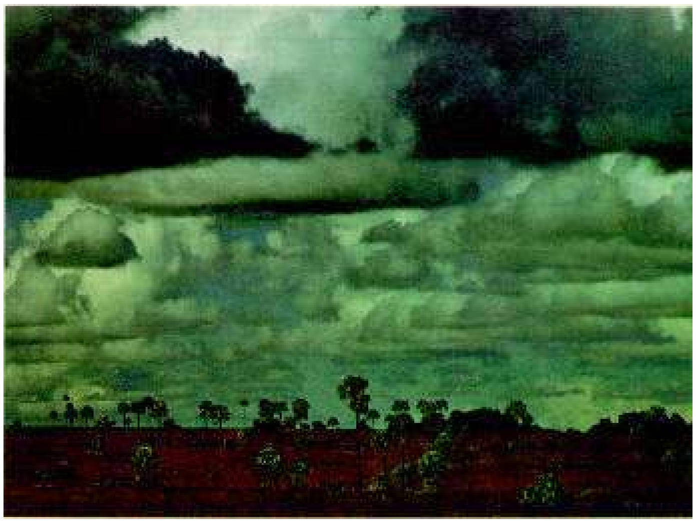

©National Geographic Society, Photo by Al Moldvay

Figure 11-10. Savanna (Ihosy, Madagascar; Plains of Tanganyika; Kenya; and Nairobi, Kenya).

The large, pan-tropic toad, Bufo Marimum, exudes a particularly irritating secretion if handled roughly. Aside from skin irritations, these amphibians pose little danger to humans, unless the secretion gets into the eyes, where it may cause blurred vision, intense burning, and possible blindness.

(3) All tropical areas of the world abound in monkeys. They are very curious animals and this fact may be used to the survivor's advantage in trying to procure one for food. Only the very large species of monkey, such as the mandrill or baboon, could constitute any danger to humans.

(4) Tropic areas also have a large number of the various species of mice, rats, squirrels, and rabbits.

(5) Members of the cat family are found in jungles throughout the world. The ocelot abounds in the jungles of Central and South America. It is a small, lean, savage cat whose coloring closely resembles that of the jaguar and will attain a weight of approximately 40 pounds and a length of 3 feet when fully grown. Cats such as the leopard are found in the tropics of the old world. The leopard is one of the most wary of beasts, becoming a powerful fighter when wounded or cornered. Unlike lions and tigers, the leopard can climb trees with ease; therefore, caution should be used when hunting this animal.

(6) Deer are found in most jungle areas; however, their population is normally small. In the Asian jungles,

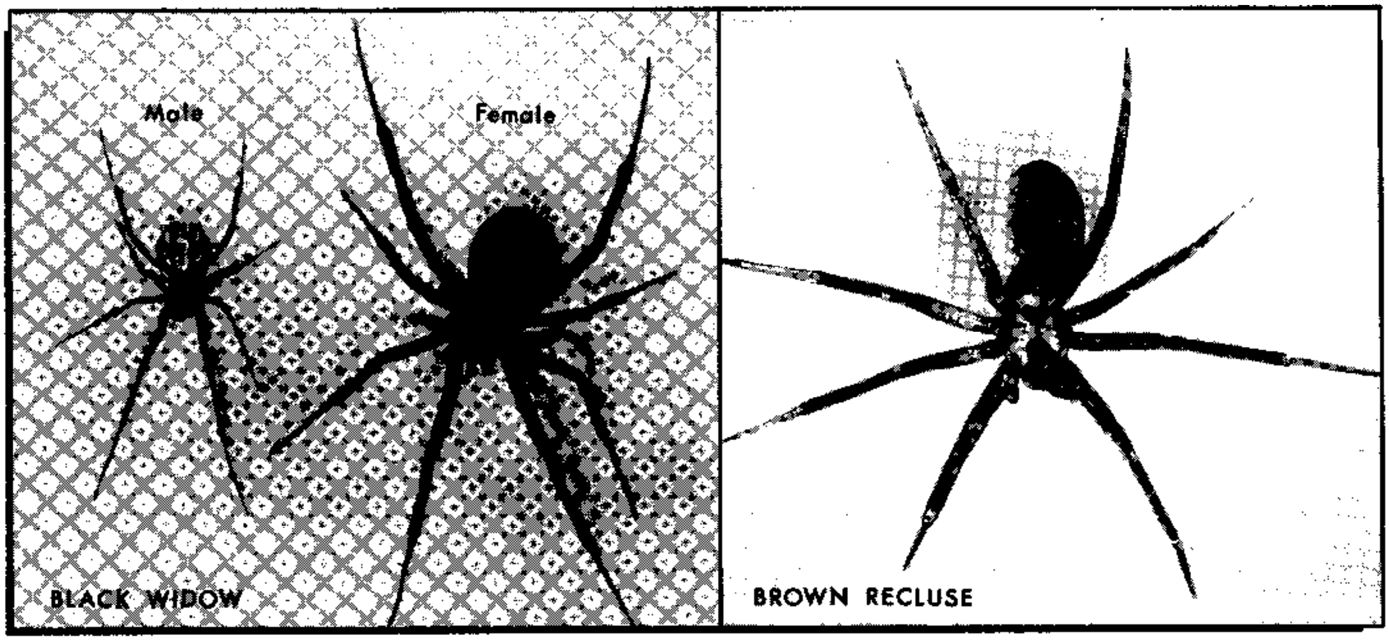

Figure 11-11. Black Widow and Brown Recluse Spiders.

several species of deer frequent the low, marshy areas adjacent to rivers. In the Central and South American jungles, two species of deer are most common. The jungle species is found in thick upland forests. It is much smaller than the North American species and seldom attains a weight of more than 80 pounds. Another deer found in the Central and South American jungles is the "brocket" or "jungle deer." This small reddish-brown deer, which attains a height of about 23 inches, is extremely shy and is found mostly in dense cover since it has no defense against other animals.

(7) The real dangers lie in the insects located in the jungle, which can pass on diseases or parasites.

(a) Malaria may be the worst enemy. It is transmitted by mosquitoes, which are normally encountered from late afternoon until early morning. Guard against bites by camping away from swamps on high land and sleep under mosquito netting, if available; otherwise, use mud on the face as a protection against insects. Wear full clothing, especially at night, and tuck pants into the tops of socks or shoes. Wear mosquito head net and gloves. Take antimalaria tablets (if available) according to directions.

(b) The greatest number of ant species is found in the jungle regions of the world. Nesting sites may be in the ground or in the trees. Ants can be a considerable nuisance especially if near a campsite. They inflict pain by biting, stinging, or squirting a spray of formic acid. Before selecting a campsite, a close check of the area should be made for any nests or trails of ants.

(c) Ticks may be numerous, especially in grassy places. Use a protected area and undress often, inspecting all parts of the body for ticks, leeches, bed bugs, and other pests. If there are several people in the group, examine each other.

(d) Fleas are common in dry, dusty buildings. The females will burrow under the toenails or into the skin to lay their eggs. Remove them as soon as possible. In India and southern China, bubonic plague is a constant threat. Rat fleas carry this disease and discovery of dead rats usually means a plague epidemic in the rat population. Fleas may also transmit typhus fever and in many parts of the tropics, rats also carry parasites which cause jaundice and other fevers. Keep food in rat-proof containers or in rodent-proof caches. Do not sleep with any food in the shelter!

(e) In many parts of the Far East, a type of typhus fever is carried by tiny red mites. These mites resemble the chiggers of southern and southwestern United States. They live in the soil and are common in tall grass, cut-over jungle, or stream banks. When a person lies or sits on the ground, the mites emerge from the soil, crawl through clothes, and bite. Usually people don't know they have been bitten, as the bite is painless and does not itch. Mite typhus is a serious disease and the survivor should take preventive measures to avoid this pest. The survivor should clear the camping ground and burn it off, sleep above the ground, and treat clothing with insect repellant.

(f) Leeches are primarily aquatic and their dependence on moisture largely determines their distribution. The aquatic leeches are normally found in still, freshwater lakes, ponds, and waterholes. They are attracted by disturbances in the water and by a chemical sense.

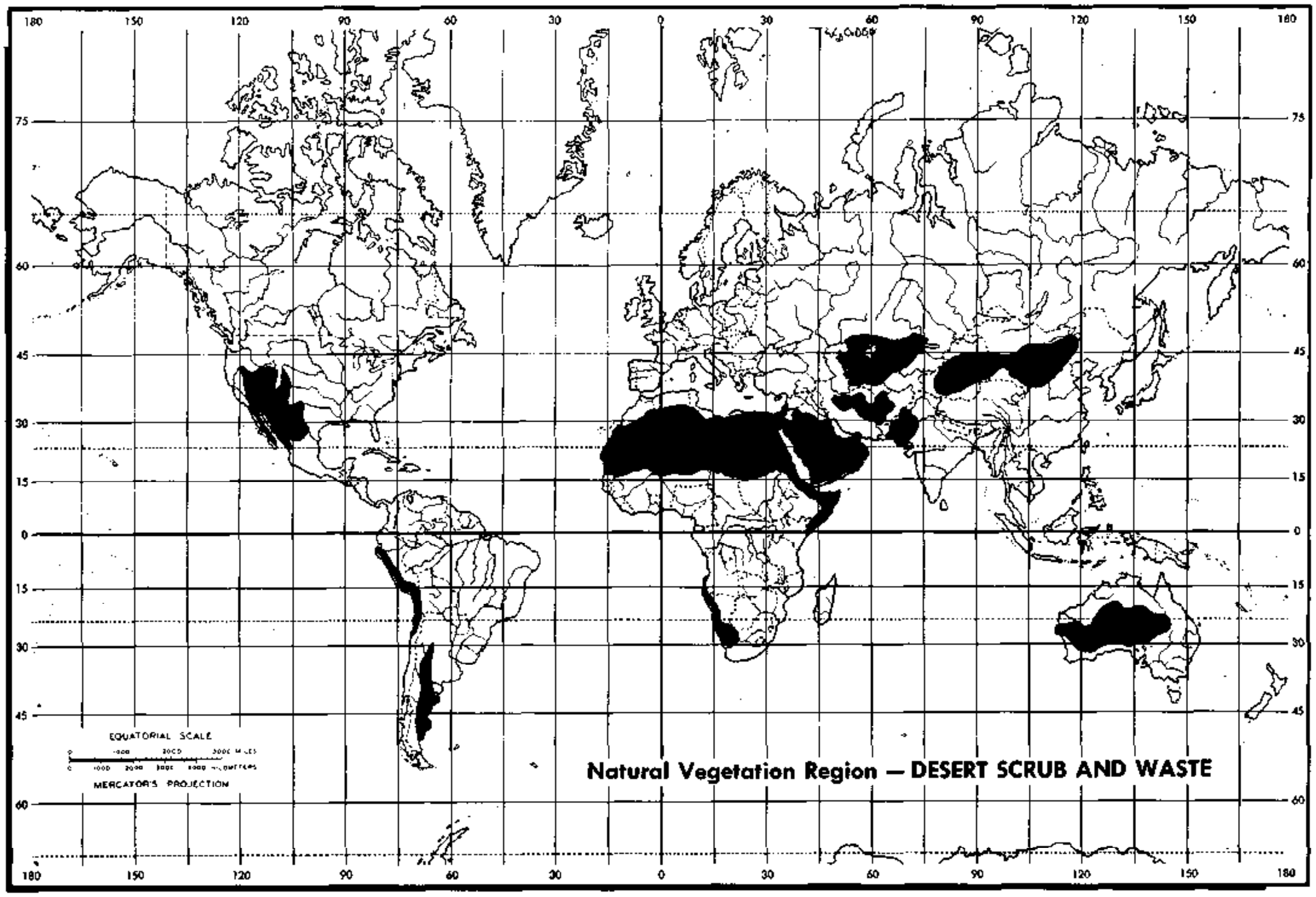

Figure 11-12. Desert Scrub and Waste Areas.

Land leeches are quite bloodthirsty and easily aroused by a combination of odor, light, temperature, and mechanical sense. These leeches are the most feared of all since they may enter air passages from which they cannot escape once they have fed and become distended. Normally, there is little pain when leeches attach themselves, and after they fill with blood, they drop off unnoticed. Some leeches, living in springs and wells, may enter the mouth or nostrils when drinking and may cause bleeding and obstruction.

(g) Spiders, scorpions, hairy caterpillers, and centipedes are often abundant. The survivor should shake out shoes, socks, and clothing and inspect bedding morning and evening. A few spiders have poisonous bites which may cause severe pain. The black widow and the brown recluse spiders are venomous and should be considered very dangerous (figure 11-11). The large spiders called tarantulas rarely bite, but if touched, the short, hard hairs which cover them may come off and irritate the skin. Centipedes bite if touched and their bite is like that of a wasp's sting. Avoid all types of many-legged insects. Scorpions are real pests as they like to hide in clothing, bedding, or shoes and strike without being touched. Their sting can cause illness or death.

j. Population. Density of human population varies with the climate. Cultivation is difficult in areas of tropical rain forests along the Equator. The torrential rains leach out the soil and weeds grow rapidly. Consequently, cultivated food sources must be supplemented by game and other products of the forest. Villages are usually scattered along rivers since movement is easier by water than through the dense forest. Numerous people are also located along coastal areas where farming takes place and people can obtain food from the sea.

11-4. Dry Climates. Dry climates are generally thought of as hot, barren areas that receive scanty rainfall. Rainfall is limited but dry climates are not barren wastelands and many kinds of plants and animals thrive (figure 11-12).

a. Deserts:

(1) Most deserts are located between the latitudes of 15 and 35 degrees on each side of the Equator and are

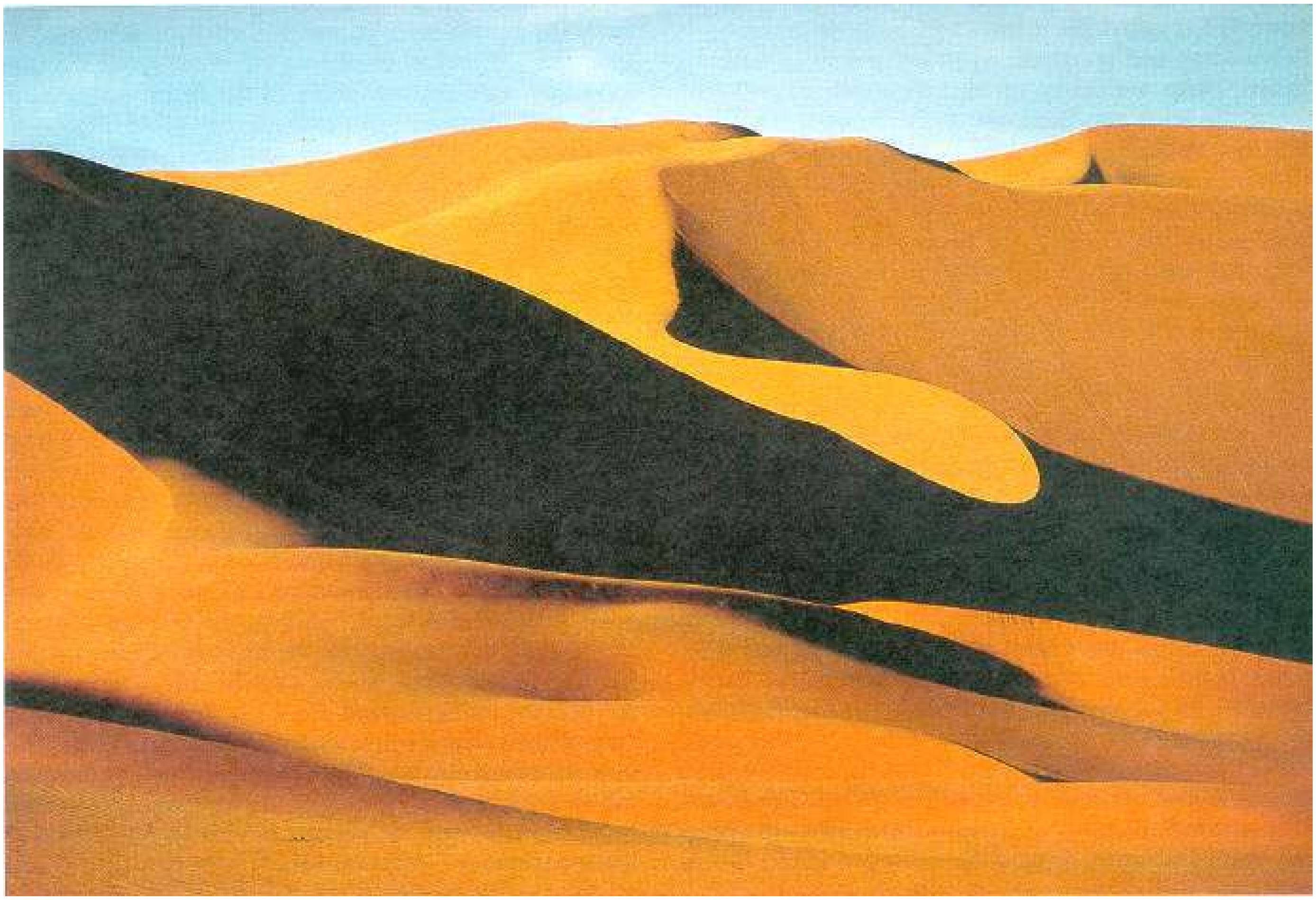

©National Geographic Society, Photo by Thomas J. Abercrombie

Figure 11-13. The Sahara.

dry regions where the annual evaporation rate exceeds the annual precipitation rate (generally less than 10 inches of rain annually). Extremes of temperature are as characteristic of deserts as is lack of rain and great distances. Hot days and cool nights are usual. A daily low-high of 45°F in the Sahara Desert and a 25°F to 35°F difference in the Gobi Desert is the rule. The difference between summer and winter temperatures is also extreme.

(2) Deserts occupy nearly 20 percent of the Earth's land surface, but only about 4 percent of the world's population lives there. The term "desert" is applied to a variety of areas. There are alkali deserts, rock deserts, and sand deserts. Some are barren gravel plains without a spear of grass, a bush, or cactus for a hundred miles. In other deserts, there are grasses and thorny bushes where camels, goats, or even sheep find a subsistent diet. Anywhere they are found, deserts are places of extremes. They can be extremely dry, hot, cold, and often devoid of plants, trees, lakes, or rivers. Most important to the survivor is the extremely long time (distance) between water sources. The vast deserts of North Africa, Arabia, Iran, West Pakistan, the Mojave Desert of the southwestern United States and northern Mexico, the interior Kalahari Desert of South Africa, and the Australian Desert are major examples of this type.

(3) Desert areas and climatic characteristics and seasonal variations of world deserts are as follows:

(a) Sahara Desert. The Sahara is the largest desert in the world. It stretches across North Africa from the Atlantic Ocean to the Red Sea and from the Mediterranean and the Sahara Atlas Mountains in the north to the Niger River in tropical Africa. It consists of 3 million square miles of level plains and jagged mountains, rocky plateaus, and graceful sand dunes. There are thousands of square miles where there is not a spear of grass, not a bush or tree, nor a sign of any vegetation. But Sahara oases—low spots in the desert where water can be reached for irrigation—are among the most densely populated areas in the world. Date groves and garden patches supporting 1,000 people per square mile are surrounded by barren plains devoid of life. Only 10 percent of the Sahara is sandy. The greater part of the desert is flat gravel plain from which the sand has been blown away and accumulated in limited areas forming dunes. There are rocky mountains rising 11,000 feet above sea level and there are a few depressions 50 to 100 feet below sea level. The change from plain to mountain is abrupt in the Sahara. Mountains generally go straight up from the plain like jagged skyscrapers from a city street. Sharp rising mountains on a level plain are especially noticeable in many desert landscapes because there is no vegetation to modify that abruptness. The lack of trees or bushes makes even occasional foothills appear more abrupt than in temperate climates (figure 11-13).

(b) Arabian Desert. Some geographers consider the Arabian Desert as a continuation of the Sahara. Half a million square miles in area, the Arabian Desert covers most of the Arabian Peninsula except for fertile fringes along the Mediterranean Sea, Red Sea, Arabian Sea, and the valleys of the Tigris-Euphrates Rivers. Along much of the Arabian coastline, the desert meets the sea. There is more sand in the Arabian Desert than in the Sahara and there are fewer date grove oases. These are on the east side of the desert at Gatif, Hofuf, and Medina. Also, there is some rain in Arabia each year, in contrast to the decades in the Sahara that pass without a drop. Arabia has more widespread vegetation, but nomads find scanty pasture for their flocks of sheep and goats and must depend on wells for water. Oil is carried across the desert in pipelines which are regularly patrolled by aircraft. Pumping stations are located at intervals. All these evidences of modern civilization have increased the well-being of the desert people and, as a result, chances for a safe journey afoot. However, the desert of Arabia is rugged, and native Arabs still get lost and die from dehydration.

(c) Gobi Desert - "Waterless Place." As used here, "Gobi" means only the 125,000 square mile basin or saucer-like plateau north of China which includes Inner and Outer Mongolia. On all sides of the Gobi, mountains form the rim of the basin. The basin itself slopes so gently that much of it appears to be a level plain. The Gobi has rocks, buttes, and numerous badlands, or deeply gullied areas (figure 11-14). For a hundred miles or so around the rim of the Gobi, there is a band of grassland. In average years, the Chinese find this to be a productive farmland. In drought years, agriculture retreats. Moving toward the center of the Gobi, there is less and less rainfall; soil becomes thinner, and grass grows in scattered bunches. This is the home of the Mongol herdsman. His wealth is chiefly horses, but he also raises sheep, goats, camels, and a few cattle. Beyond the rich grassland, the Gobi floor is a mosiac of tiny pebbles which often glisten in sunlight. These pebbles were once mixed with the sand and soil of the area, but in the course of centuries, the soil has been washed or blown away and the pebbles have been left behind as loose pavement. What rain there is in the Gobi drains

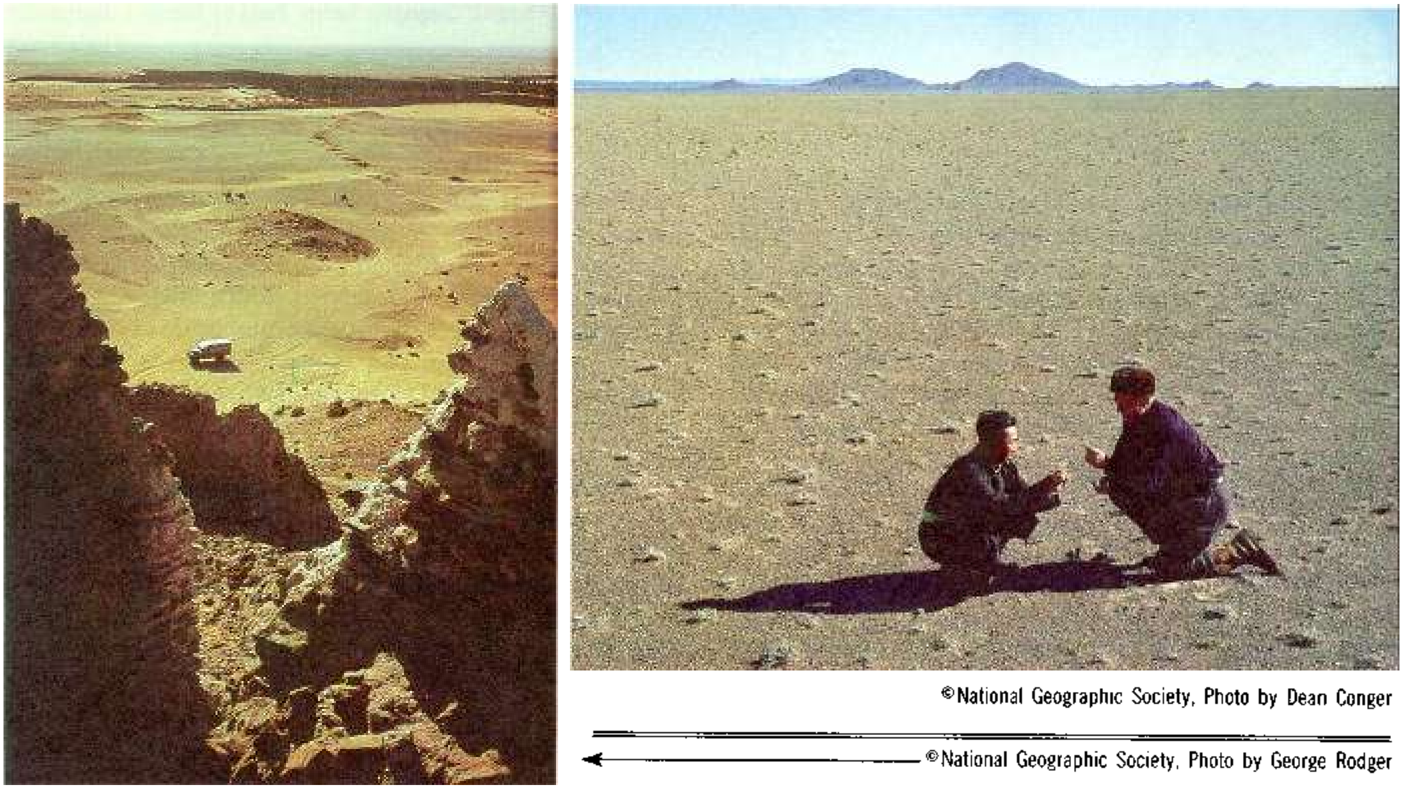

©National Geographic Society, Photo by Dean Conger

← ©National Geographic Society, Photo by George Rodger

Figure 11-14. The Gobi.

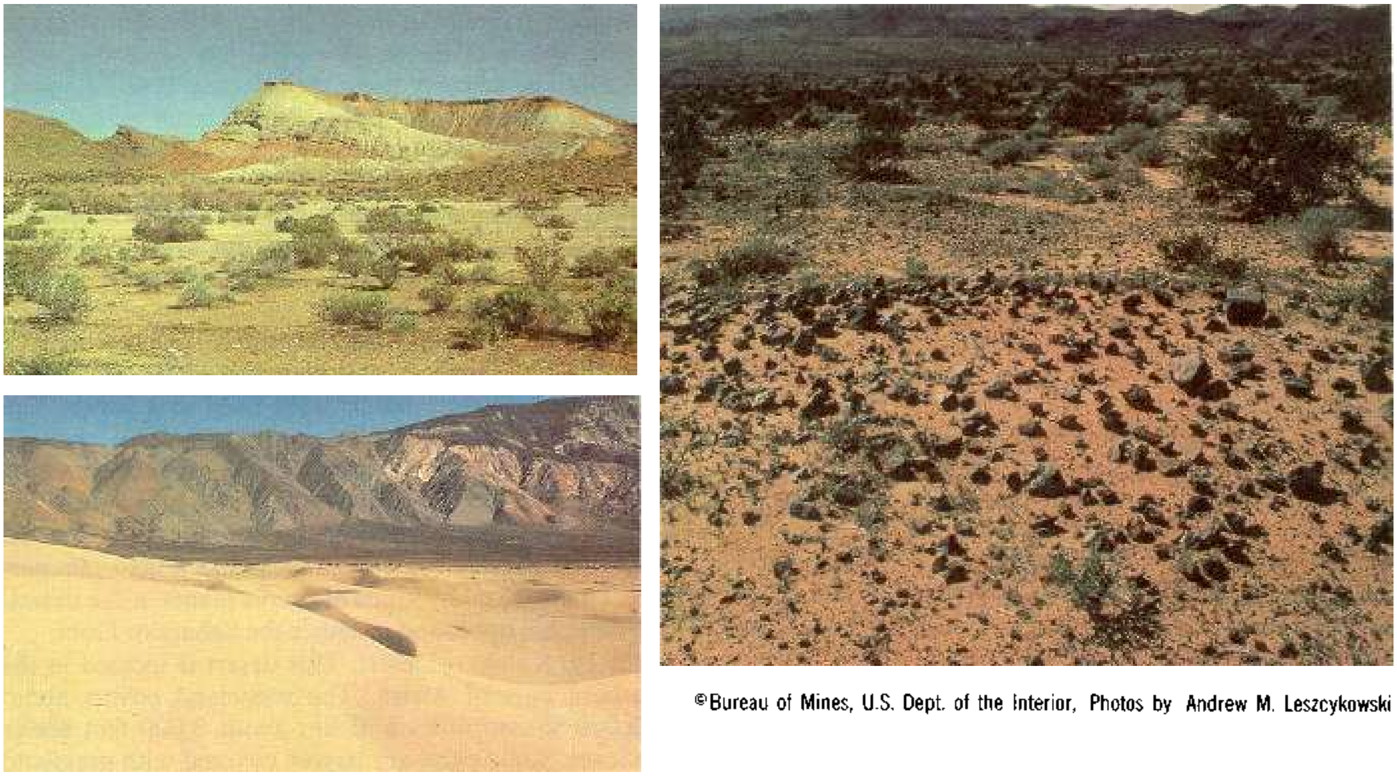

©Bureau of Mines, U.S. Dept. of the Interior, Photos by Andrew M. Leszcykowski

Figure 11-15. Composite of American Deserts.

toward the basin; almost none of it cuts through the mountain rim to the ocean. There are some distinct and well-channeled watercourses, but these are usually dry. Many are remnants of prehistoric drainage systems. In the east, numerous shallow salt lakes are scattered over the plain. They vary in size and number with the changes of rainfall in the area. Sand dunes are found in the eastern and western Gobi, but these features are not as pronounced as they are in certain sections of the Sahara Desert. The Gobi is not a starkly barren wasteland like the great African Desert. Grass grows everywhere, although it is often scanty. Mongols live in collective type farm systems and habitations instead of being concentrated in oases (figure 11-14).

(d) Australian Desert. More than one-third of Australia's total area is desert. Rainfall in the area is unpredictable with an average of less than 10 inches per year. There are three connecting deserts which occupy western Australia and one desert located in the center of the continent. They are the Great Sandy Desert, Great Victorian Desert, Gibson Desert, and Simpon or Arunta Desert. The three largest deserts, Sandy, Victoria, and Arunta are of the sandy type, held in place by vegetation. The Gibson in the western portion, is a stone-type desert. Most of the deserts of Australia have elevations of 1,000 to 2,000 feet.

(e) Atacama-Peruvian Desert of South America. Generally, there are two regions of desert in South America. The first, and by far the largest, is along the west coast, beginning in the southern part of Ecuador, extending the entire coastline of Peru, and reaching nearly as far south as Valparaiso in Chile. This region, of about 2,000 miles in length and approximately 100 miles in width, is classified as true desert. Even so, along the shoreline of Peru as far south as Africa and inland a few miles, there is often a low-cloud or misty-fog layer. The layer is approximately 1,000 feet thick and produces a fine drizzle. Because of this frequent cloud cover and other phenomena, the temperature along the coastal desert is remarkably cool, averaging about 72°F in the summer daytime and about 55°F in the winter daytime. From about 30 degrees south, the cloud cover does not exist and this region may truly be called rainless. The rare and uncertain showers are valueless for cultivated vegetation. Behind the coastal ranges in the higher elevations, the dryness is at a maximum. In the nitrate fields of the Great Atcama Desert, the air is very dry and the slightest shower is very rare. Here the summer daytime temperatures are from 85°F to 90°F. The second desert region is entirely in Argentina (east of the Andes) extending in a finger-like strip from about 30 degrees south, southwest to about 50 degrees south. This region is approximately 1,200 miles long and 100 miles wide. In this highly dissected plateau region, the temperature ranges from a yearly average of 63°F in the north to 47°F in the south. The average annual rainfall pattern is from about 4 inches in the north to about 6 inches in the south.

(f) Southwest Deserts of the United States and Mexico:

-1. These desert areas have four major subdivisions:

-a. Great Basin—the basin between the Rocky Mountains and the Sierra Nevada-Cascade Ranges of southern Nevada and western Utah.

-b. Mojave Desert— Southwestern California.

-c. Sonoran Desert—Southeastern California across southern Arizona into the southwest corner of New Mexico and from Sonora and Baja, California, into Mexico.

-d. Chihuahuan Desert—Lies to the east of the Great Sierra Madre Occidental, spreading north into southwest Texas, southern New Mexico, and southeast corner of Arizona (figure 11-15).

-2. The flat plains with scanty vegetation and abruptly rising buttes of our Southwest are reminders of both the Gobi and Sahara. But the spectacular rock-walled canyons found in the Southwest have few counterparts in the deserts of Africa and Asia. The gullied badlands of the Gobi resemble similar formations in both the Southwest and the Dakotas, but our desert rivers—the lower Colorado, lower Rio Grande, and tributaries such as the Gila and the Pecos—have a more regular supply of water than is found in Old World deserts. The Nile and Niger are, in part, desert rivers but get their water from tropical Africa. They are desert immigrant rivers (like the Colorado, which collects the melting snows of the southern Rockies) and gain sufficient volume to carry them through the desert country. In general, the southwest deserts have more varied vegetation, greater variety of scenery, and more rugged landscape than either the Gobi or the Sahara. In all three areas, it is often a long time (distance) between water sources. Death Valley, a part of the Mojave lying in southern California, probably has more waterholes and more vegetation than exist in vast stretches of the Sahara. The evil reputation of the Valley appears to have been started by unwise travelers who were too terrified to make intelligent searches for food and water. The dryness of the Death Valley atmosphere is unquestioned, but it lacks the vast barren plains stretching from horizon to horizon in the Sahara. Compared to the Sahara, the desert country of southwestern United States and Mexico sometimes looks like a luxuriant garden. There are many kinds of cactus plants in the desert, but these are not found in either the Sahara or Gobi.

(g) Kalahari Desert. This desert is located in the southern part of Africa. The wasteland covers about 200,000 square miles and lies about 3,000 feet above sea level. Some parts are largely covered with grassland and scrubby trees. The climate is similar to that of the Atacama-Peruvian Desert of South America.

b. Vegetation. The following are the more common xerophytic plants (those plants that can live with a limited water supply) which are found in the major deserts of the world.

(1) Cactus Family (Cactacae). Most cactus fruits and leaves have spines to protect them from birds and

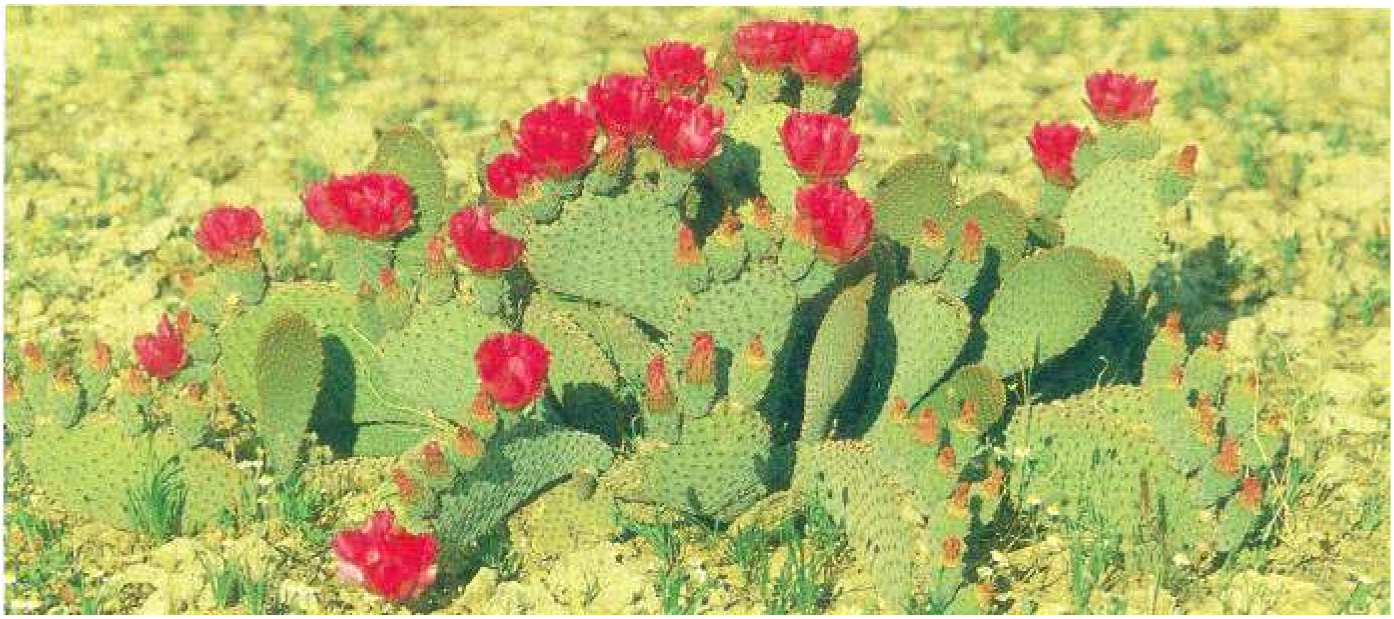

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

Figure 11-16. Prickly Pear.

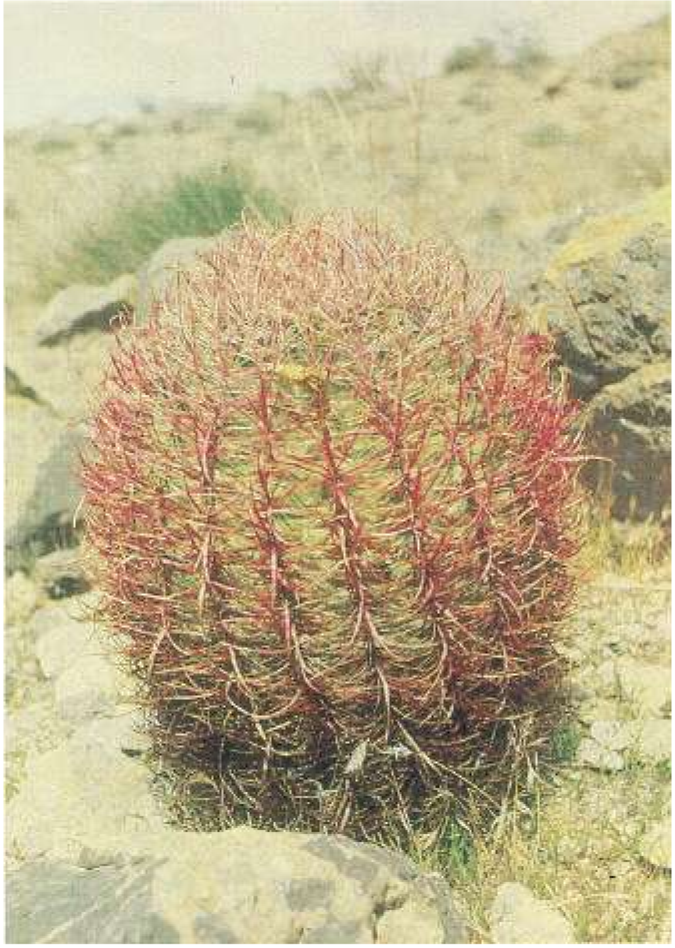

©Bureau of Mines, U.S. Dept. of the Interior, Photo by Andrew M. Leszcykowski

Figure 11-17. Barrel Cactus.

animals seeking water stored in the stems and leaves. Flat cactus leaves can be boiled and eaten and the flowering fruit is edible.

(a) Prickly Pear. Is native and most abundant in the American deserts, but has been introduced into the Gobi, Sahara, and Australian Deserts, and other parts of the world (figure 11-16).

(b) Barrel Cactus. Found in many places, but is only native to the North American deserts. It grows to 5 or 6 feet high (figure 11-17).

(c) Suguaro (Giant) Cactus. Abundant in southern Arizona and in Sonora, Mexico. Can grow up to 50 feet tall.

(2) Wild Onion. Found in the Great Basin of the Southwest and in the Gobi Desert. The bulbs are edible if they look, smell, and taste like an onion or garlic.

(3) Wild Tulip. Found in the Gobi and Sahara Deserts. The bulbs can be eaten.

(4) Shrubs:

(a) Abal. Grows to about 4 feet tall in sandy deserts. The fresh flowers can be eaten. The dry twigs can be crushed and used as a tea substitute. It is found in the Sahara and Arabian Deserts.

(b) Acacia. Most common in the Sahara, Gobi, and Australian desert regions, and in the warmer and drier parts of America. The beans can be crushed and cooked as porridge. It is spiny with many branches and grows to 10 feet tall; roots yield water 4 to 5 feet from the tree trunk (figure 11-18).

(c) Saxaul. Found on the salt deserts of the Gobi Desert. The bark acts as water storage and is a good water source.

(d) St John's Bread. Found along the border of the Mediterranean coast of the Sahara and across the Arabian Desert. Grows 40 to 50 feet tall and seeds can be pulverized and cooked as a porridge (most nutritious plant food in the Middle East).

(e) Juniper. Found in the mountainous areas of the American deserts.

(f) Vines. Wild desert gourds are found in the Sahara and Arabian Deserts. They have a vine which grows from 8 to 30 feet long, and produce a melon-like poisonous fruit. The seed can be eaten when roasted or boiled. Flowers are also edible (figure 11-19).

(g) Succulent Plants. They are filled with juices and store moisture to survive. The surface is covered with a layer of wax or a blanket of fine hairs for protection against the heat. The moisture is contained in a tough cellulos that is not digestible and must be manually broken down to release the water.

(h) Creosote Bush. This is the most widespread and successful plant of the American deserts (height from 2 to 10 feet).

(5) Dates. Occur in groves around desert oases of the lower areas of the Sahara Desert.

c. Animal Life. There are over 5,000 species of birds, reptiles, mammals, and insects found in desert areas. The raven, dove, woodpecker, owl, and hawk are common bird species. Reptiles such as lizards and snakes are numerous due to their adaptation to desert areas and ability to conserve body fluids. Many types of mammals live in desert areas and are primarily found near water sources.

d. Human Population. Humans are greatly influenced by the presence of water and they live close to rivers, wells, cisterns, or oases. For example, in the Sahara Desert, the 2 million people inhabiting this area are located near about 50 desert oases and small coastal cities. In the Gobi Desert, the Mongols live in scattered camps and move from one well to the next as they travel. In the southwest deserts of America, the population is greater along the Colorado and Rio Grande rivers.

11-5. Warm Temperate Climates:

a. The temperate zone is the area or region between the Tropic of Cancer and the Artic Circle and between the Tropic of Capricorn and the Antarctic Circle. The

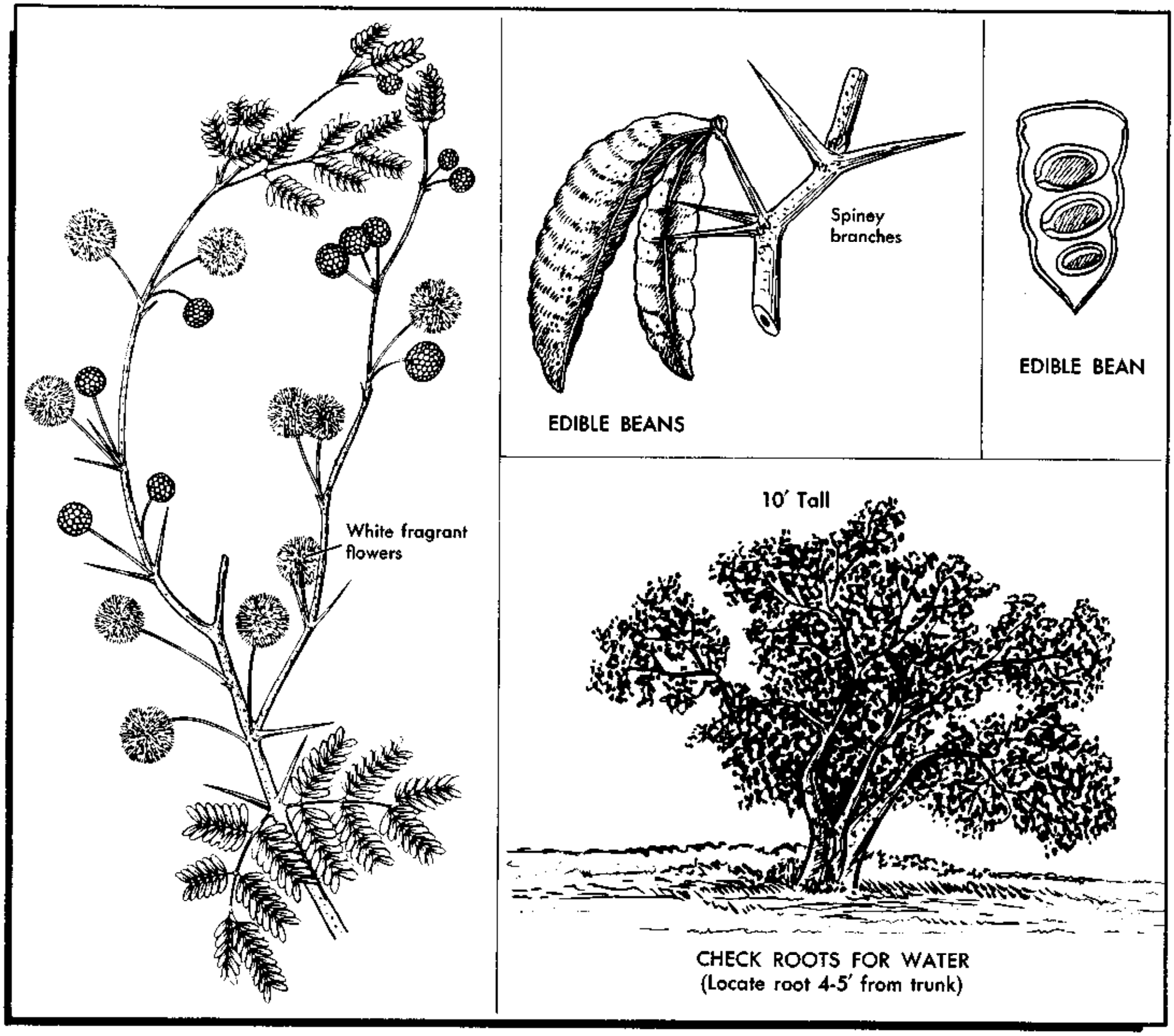

Figure 11-18. Acacia.

latitudes which comprise the temperate zone are 23½° north latitude to 66½° north latitude and 23½° south latitude to 66½° south latitude.

b. There are two main types of climate which comprise the temperate group—mild type, dominated by oceanic or marine climate; and a more severe one called continental climate.

(1) The temperate oceanic climate is the result of warm ocean currents where the westerly winds carrying moisture have a warming effect on the landmass. This oceanic type climate cannot develop over an extensive area on the eastern or leeward side of large continents in the middle latitudes. The extended effect of the ocean climate can be limited by mountain ranges. Such is the case with the Olympic, Cascade, and Rocky Mountains. As the oceanic weather system moves across the Olympic Mountains, it drops nearly 300 inches of precipitation annually. On the windward side of the Cascades, the annual precipitation ranges from 80 to 120 inches annually. In contrast, the region from the leeward side of the Cascades is a relatively dry area, receiving between 10 to 20 inches precipitation annually. As the system moves across the Rocky Mountains, most of the remaining moisture is lost.

(2) The temperate continental climate is a land-controlled climate which is a product of broad middle latitude continents. Because of this, the continental climate is not found in the Southern Hemisphere. This type of climate is very characteristic of the leeward side of mountain barriers and eastern North America and

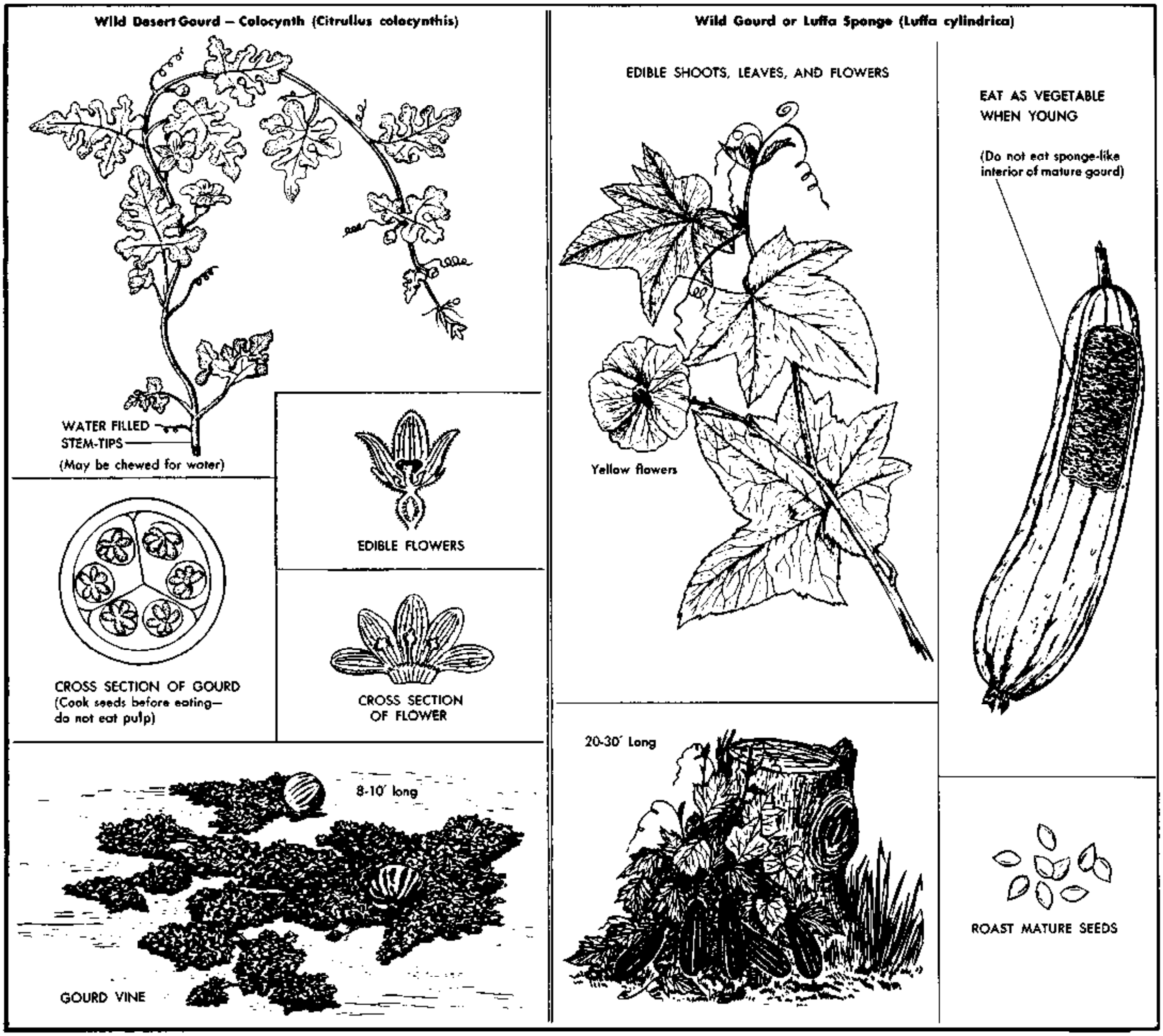
Wild Desert Gourd – Colocynth (Citrullus colocynthis)
WATER FILLED STEM-TIPS
(May be chewed for water)
EDIBLE FLOWERS
CROSS SECTION OF GOURD
(Cook seeds before eating—do not eat pulp)
CROSS SECTION OF FLOWER
8-10' long
GOURD VINE
Wild Gourd or Luffa Sponge (Luffa cylindrica)
EDIBLE SHOOTS, LEAVES, AND FLOWERS
Yellow flowers
EAT AS VEGETABLE WHEN YOUNG
(Do not eat sponge-like interior of mature gourd)
20-30' Long
ROAST MATURE SEEDS

Figure 11-19. Wild Desert Gourd.

Asia. These areas are associated with dry interiors since there are few major warm-water sources available for formation of water systems. The average temperature in the winter and summer are not only extreme but also variable from one year to the next. The severe winter temperature is caused by the polar airflow toward the Equator, and neither winter nor summer temperatures are moderated by the effects of large water masses (oceans).

c. The climate within the temperate zone varies greatly in temperature, precipitation, and wind. The temperate (midlatitude) zone is divided into four major climate zones which are controlled by both tropic and polar airmasses.

(1) The humid subtropical zone is located generally between 20 and 30 degrees north and south latitude. This climate also tends to occur on the east coast of the continents which are at these latitudes. An example of this zone in the United States is the area between Missouri to lower New York and east Texas to Florida. The temperature ranges from 75°F to 80°F in the summer months to 27°F to 50°F in the winter months. The total average precipitation is 30 to 60 inches or more. During the summer months, convectional rainfall is common and thunderstorms frequent. In the winter, the rain is

©National Park Service, U.S. Dept. of the Interior, Photo by M. Woodbridge Williams

Figure 11-20. Marine West Coast Climate (Olympic Peninsula).

more widespread and is usually associated with passing midlatitude cyclones. The wind has a great influence in this area. The area is affected by both the prevailing westerlies and eastern trade winds. During the summer, the winds are influenced by eastern moist maritime airmass flows. Winters are influenced by westerly continental polar airflows. The weather is also influenced by low latitudes. The equatorial current which turns poleward forms warm currents (Gulf Stream and Japanese and Brazilian) that parallel the coasts.

(2) The marine west coast climate (figure 11-20) is sometimes referred to as the temperate oceanic climate. This climate is generally between 40 and 60 degrees north and south latitudes, on the west side of the continent. Examples are the west coasts of Washington to Alaska, Chile, nearly all of Europe, and New Zealand. The summer months are cool with average temperatures of 60°F to 70°F. The winter months are mild with temperatures averaging 27°F to 50°F. The total average rainfall ranges from 20 to 200 inches. Since the maritime climates are under the influence of the westerly winds all year, rainfall is nearly uniform from season to season. These climates are probably more cloudy than any other. They are characterized by widespread stratus and nimbostratus clouds and frequent fog. One of the main reasons for the tremendous rainfall in these climatic areas is the warm ocean currents. These currents yield moisture to the air which is blown inland by the westerly winds (figures 11-21 and 11-22).

(3) Middle latitude desert and steppe climates of complex origins are found generally between latitudes 35 to 50 degrees and in the interior of Asia and North America. Mountain ranges serve as barriers to the moist maritime airmasses, thus resulting in low levels of precipitation. In summer, these interiors generate tropical airmasses, while in winter they are overrun by polar air masses originating in Canada and Siberia. Deserts are also characterized by considerable differences between the average summer and winter temperatures. Of greater importance are the vast semi-dry steppes. Their annual precipitation of 10 to 20 inches supports shortgrass vegetation. They comprise the great sheep and cattle ranges of the world; for example, the veldt of South Africa and the American Great Plains support vast numbers of animals.

(4) The Mediterranean climate is sometimes referred to as subtropical dry summer climate. It is generally located from 30 to 45 degrees north and south lati-

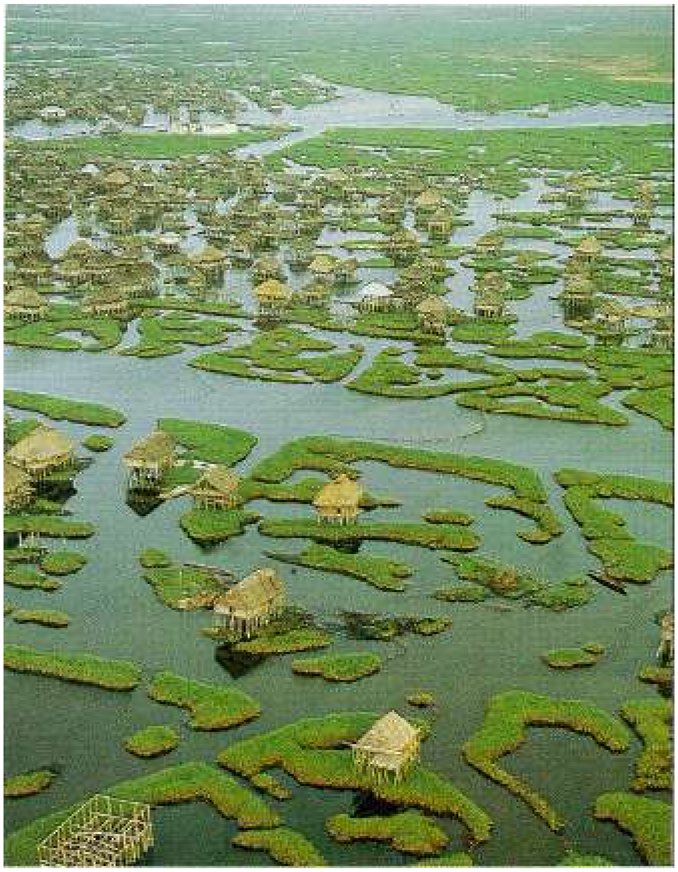

©National Geographic Society, Photo by John Scofield

Figure 11-21. Maritime Climate (Southern Dahomey, Africa).

tudes. Examples of this climate occur in the Mediterranean region, most notably Spain, Italy, and Greece. Summer temperatures usually average 75°F to 80°F; but in coastal locations near cool currents, the average is 5°F to 10°F lower. Typical temperature averages for the coldest months are 45°F to 55°F. Coastal locations are usually somewhat warmer in the winter than inland locations. Total annual rainfall is normally 15 to 30 inches along the equatorial margins and increases poleward. This climate is a transitional zone between the dry west coast desert and the wet west coast climate. The westerly winds and cold ocean currents are the controlling influences of the Mediterranean climate. An example of a cold current which affects climates are the Humbolt Current (Peru Current) along the coast of Chile, Peru, and California.

d. Major topographical characteristics found in temperate regions are:

(1) Mountains. Areas of steep slopes with local relief of more than 2,000 feet. Examples of this land form are the Rocky Mountains of North America, the Andes Mountains of South America, and the Himalayan Mountains of Asia.

(2) High Tablelands. Upland surfaces over 5,000 feet in elevation and having local relief of less than 1,000 feet, except where cut by widely separated canyons such as the High Tableland of the Wyoming Basin.

(3) Hills and Low Tablelands. Hill areas having local relief of more than 325 feet but less than 2,000 feet. At the ocean shoreland, however, local relief may be as low as 200 feet. A low tableland is an area less than 5,000 feet in elevation with local relief less than 325 feet, but which (unlike plains) either does not reach the sea or where it does, terminates in a bluff overlooking a low coastal plain. Examples of this terrain can be found in the Appalachian Mountains, Quebec, Southern Argentina.

(4) Plains. Surfaces with local relief of less than 325 feet. On the marine side, the surface slopes gently to the sea. Plains rising continuously inland may attain elevations of high plains—over 2,000 feet. The greatest expanses of plains occur in the center of the North American Continent, Eastern Europe, and Western Asia (figure 11-23).

(5) Depressions. Basins surrounded by mountains, hills, or tablelands which abruptly outline the basins. Examples of depressions can be found in the southwestern United States.

e. There are several biomes of plants and animal life within the temperature zone, and the characteristic life forms are dependent upon climatic characteristics within a specific area. The biomes are named for the plants most plentiful in the area.

(1) Coniferous Forests (figure 11-24). These occur in a broad band across the northern portions of the continents of North America, Europe, and Asia. The northern boundary is the tundra and the southern limits are generally around 50 degrees north latitude. However, this zone extends down to 35 degrees north latitude in the mountainous regions of the western North American continent. This biome corresponds with the humid continental climate, except in the mountainous portions of North America below 50 degrees north latitude. The main life forms in this zone are the conifers or needle leaf, cone-producing trees, such as pines, firs, spruces, and hemlock. In these areas, the trees may grow closer together, not being severely limited by a need for sunlight, and are subject to frequent fires caused by lightning. When this occurs, the ecological succession is reversed, allowing low shrubs to spring up in the burned-over areas. Although the conifer is the predominate tree, there is more subclimax or secondary growth in these biomes than in climax forests (mature or primary forests). In these areas, the pines, alders, aspen, and poplars are the dominant trees. The dominant shrubs are heather, small maples, and yews. If forced to survive in these areas for long periods, especially in winter, the survivor will find that edible food plants are scarce (figure 11-25).

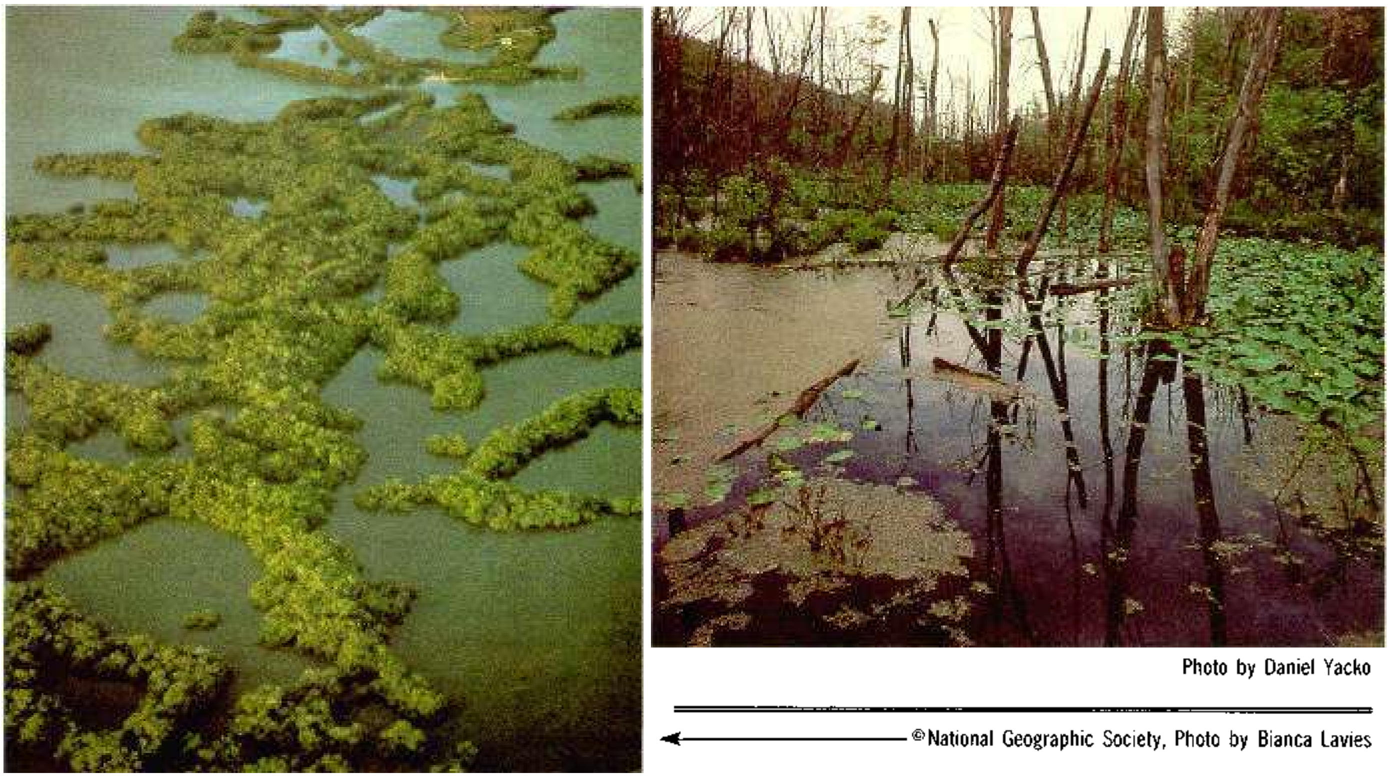

Photo by Daniel Yacko

◄ ©National Geographic Society, Photo by Bianca Lavies

Figure 11-22. Red Mangrove and Swamps.

(a) In winter, the primary edible food plants available are:

-1.Rootstalks.
-2.Bulbs.
-3.Roots.
-4.Seeds.
-5.Resins (from pines).
-6.Infusion (teas) from evergreen needles.
-7.Bark (inner part).

(b) During the summer months, many more plants are available for food, including:

-1.Nuts.
-2.Sweet Acacia.
-3.Water Plantain.
-4.Shoots (potherbs).
-5.Wild Apple.
-6.Polypody.
-7.Leaves (potherbs).
-8.Baobab.
-9.Wild Rhubarb.
-10.Pollen (cattail).
-11.Beechnut.
-12.Flowering Bush.
-13.Flowers.
-14.Braken (Fern).
-15.Wild Sorrel.
-16.Fruits (dessert).
-17.Wild Calla (water Arum).
-18.Cattail.
-19.Fiddleheads (Ferns).
-20.Chufa (Nut Grass).
-21.Chestnut.
-22.Wild Crabapple.
-23.Wild Dock.
-24.Chicory.
-25.Wild Filbert (Hazelnut).
-26.Wild Grape.
-27.Juniper.
-28.Common Jujube.
-29.Pine Nuts.
-30.Spreading Wood Fern.
-31.Wild Lily.
-32.English Oak (Acorn).
-33.Tree Fern.
-34.Water Lily (Temperate Zone).

(c) In many places, the ground is thickly covered with mosses and there may be a few varieties of early flowering plants and many berry-bearing shrubs which invite birds and mammals into the open areas. Some of the largest herbivores (plant eaters) live in these evergreen forests—caribou, reindeer, moose, and deer. The small herbivores may include porcupines, several species of squirrels, mice, and rabbits. The carnivores (flesh eaters) which feed upon the smaller animals include black bear, gray wolf, lynx, wolverine, red fox, and weasel. Multitudes of insects provide food for the birds. (These insects may also present a menace to the survivor). A large variety of birds feed not only on insects but also on plants.

(2) Deciduous Forests. Decodipis (broad leaved) forests are found extensively in the eastern portion of the United States; in Europe, between 40 to 50 degrees north latitude; and also in eastern portions of the USSR, China, Korea, and Japan from 35 to 50 degrees north latitude. This biome corresponds with the subtropic and humid continental climatic zones; the area in which any deciduous forest group determines the predominant trees or climax vegetation found there. Here are a few examples: In north central United States, Beech and Maple trees assume the dominant role; in Wisconsin and Minnesota, it is Basswood and Maple; in the eastern and southern regions, the dominant trees are

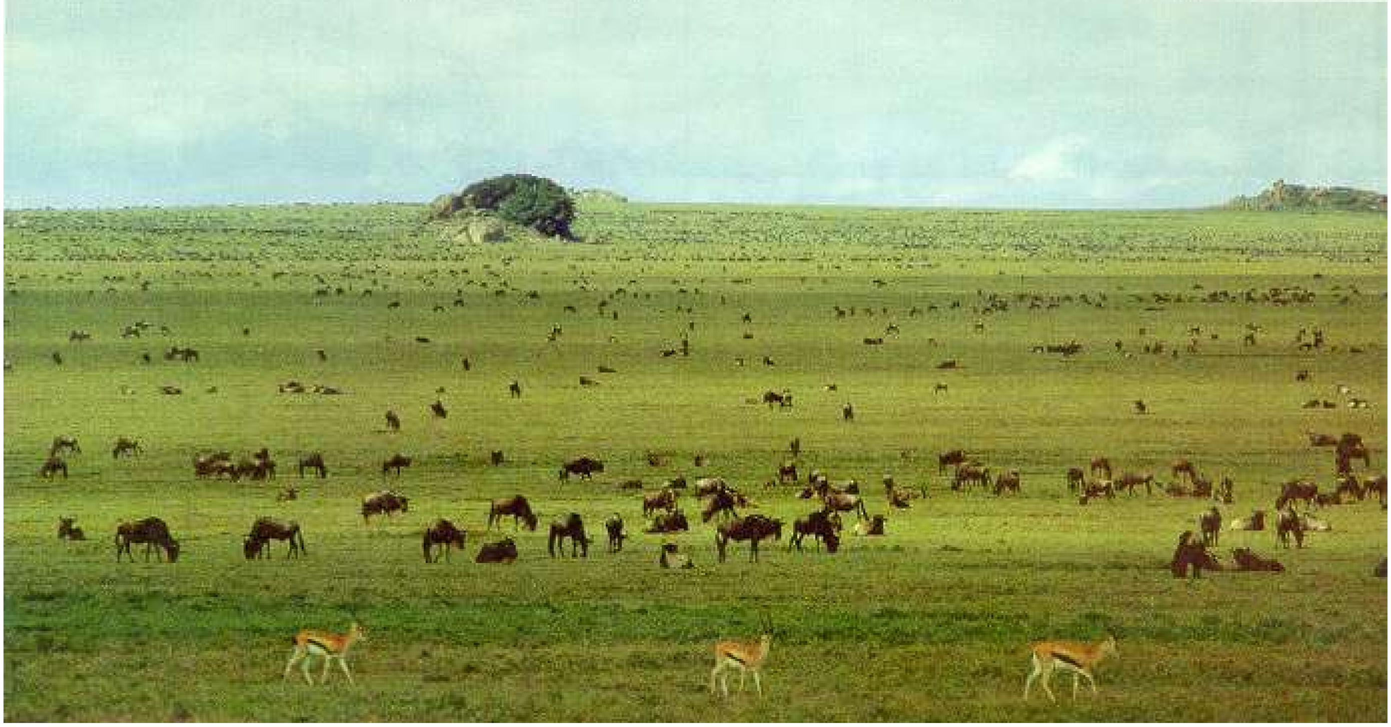

©National Geographic Society, Photo by George B. Schaller

Figure 11-23. Plains.

Oak and Hickory (figure 11-26). There are also spots in this biome where pines and broadleaf evergreens grow.

(3) Deciduous and Mixed Deciduous-Coniferous Forest:

(a) Deciduous and mixed deciduous-coniferous forests manifest the following characteristics:

-1. Warm summer with rain; winters cold and drier; short drought periods.

-2. Only three stories of vegetation (trees, scrubs, herbs).

-3. Broadleaf trees without leaves in winter.

-4. Mature trees, uniform in height.

-5. Unimpeded view into interior of forest.

-6. Few herbs, ferns, mosses in summer, and abundance of edible fungi in spring and autumn.

-7. Trunks of trees covered with thick-fissured, dark-colored bark.

-8. Resting buds enclosed in hard scaly protecting leaves frequently covered with gum or resin.

-9. For the most part, leaves are thin and delicate, rarely thick and leathery like those of tropical rain forest trees.

(b) The deciduous and mixed deciduous-coniferous forests that predominate over much of eastern United States are typical of this vegetation type. The deciduous forest is wholly temperate in character. By contrast with the tropical evergreen forest with its richly shaded but chiefly dark glossy green canopy, the broadleaved temperate forest extends in a uniformly bright green expanse. The temperate deciduous and mixed deciduous-coniferous forest vegetation type occupies extensive areas in several parts of the world (figure 11-27).

-1. North America. Eastern United States.

-2. South America. Southern Chile, southeastern Brazil.

-3. Europe. Western and northern Europe, southern Scandanavia, southeastern Europe (Balkans).

-4. Asia. South central Siberia, southeastern Siberia and part of Manchuria, Korea (throughout), Japan (throughout), China (throughout except the extreme south and extreme north).

-5. Oceania. New Zealand.

(c) A general characteristic of a climax forest is the stratification of layers of plant growth similar to the canopy systems in the tropical rain forest. In a climax forest, there are usually a limited number of flowering plants, ferns, and shrubs for ground cover. The edible food plants in the vegetation zone are numerous, and a large array of edible species are available and include:

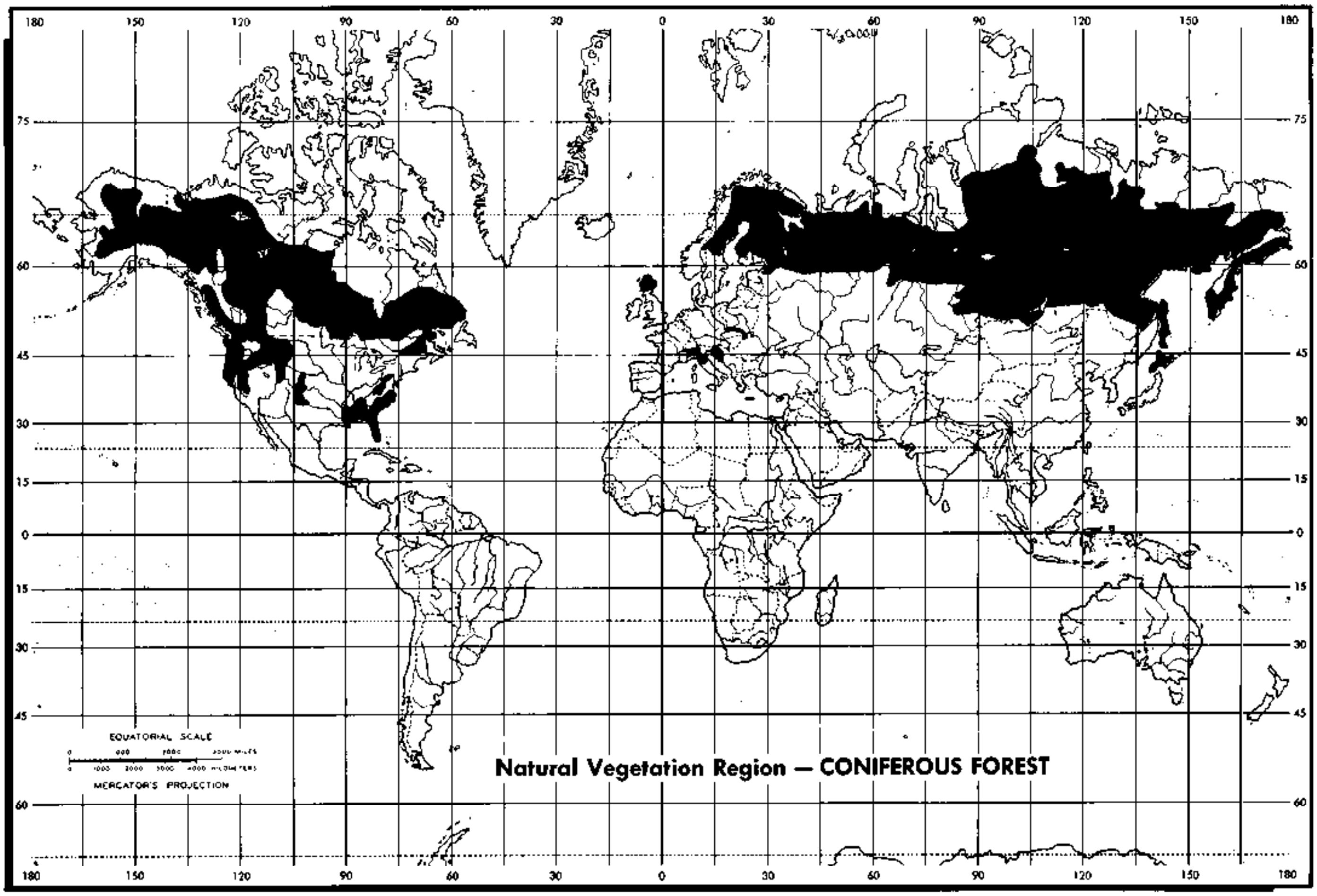

Figure 11-24. Coniferous Forest.

-1.Amarath.
-2.Chestnut, Water (Trapa Nut).
-3.Wild Lily.
-4.Wild Apple.
-5.Chicory.
-6.Lotus Lily.
-7.Beechnut.
-8.Chufa (Nut Grass).
-9.Mulberry.
-10.Braken (Fern).
-11.Wild Filbert (Hazelnut).
-12.English Oak (Acorn).
-13.Wild Calla (Water Arum).
-14.Wild Grape.
-15.Wild Onion.
-16.Cattail.
-17.Common Jujube.
-18.Pine Nuts.
-19.Chestnut.
-20.Juniper.
-21.Polypody.
-22.Air Potato (ubi Tuber).
-23.Tree Fern.
-24.Wild Dock.
-25.Purslane.
-26.Wild Tulip.
-27.Water Plantain.
-28.Wild Rhubarb.
-29.Walnut.
-30.Pokeweed.
-31.Flowering Rush.
-32.Water Lily.
-33.Tropical Yam.
-34.Wild Sorrel.
-35.Wild Crabapple.

(d) Animal life associated with deciduous forests is more varied and plentiful than in evergreen forests, though some animals such as certain species of deer, squirrels, martins, lynx, and wildcats are common in both areas. Wolves, foxes, and other small carnivores (flesh-eating animals) feed mainly on small rodents. Some forest dwellers, such as rodents, dig their dens below the ground while other dens are dug near streams where food and shelter are found. In the aquatic environment, the beaver builds dams for food and shelter. Muskrat, otter, and mink also seek the water's edge, while snakes, turtles, and frogs are found in the streams or lakes.

(4) Steppes and Prairies: (See figure 11-28.)

(a) The part of Russia extending from the Volga River through central Asia to the Gobi Desert has been referred to as the steppes. However, as a vegetation type, the steppe grasslands occur in many other parts of the world. The rainfall in steppe areas averages 15 to 30 inches per year, as compared to prairie areas which average 30 to 40 inches per year. The general aspect of a steppe area, like the prairie, is a broad treeless expanse of open countryside which may be quite rolling in places. The principal steppe areas are:

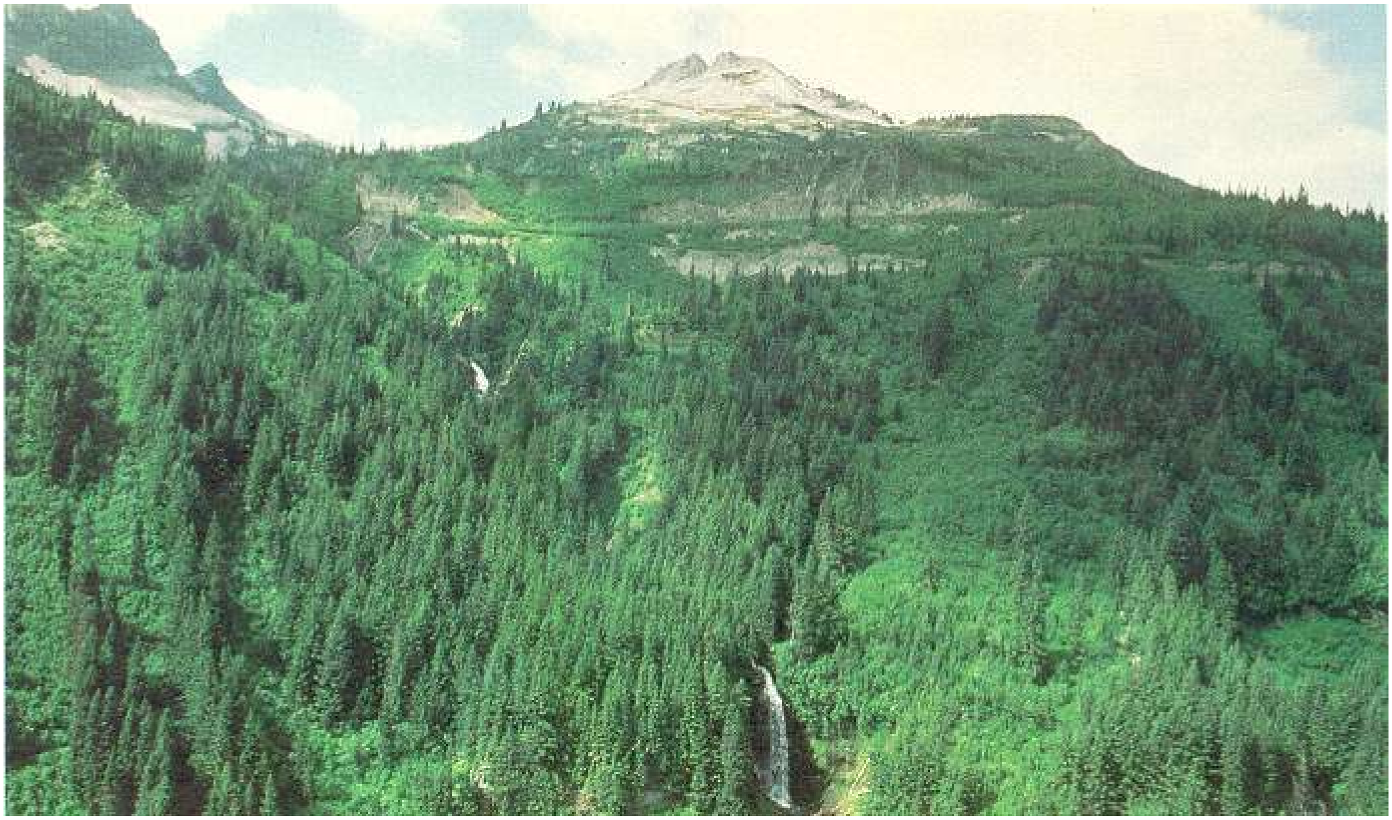

Photo by Daniel Yacko

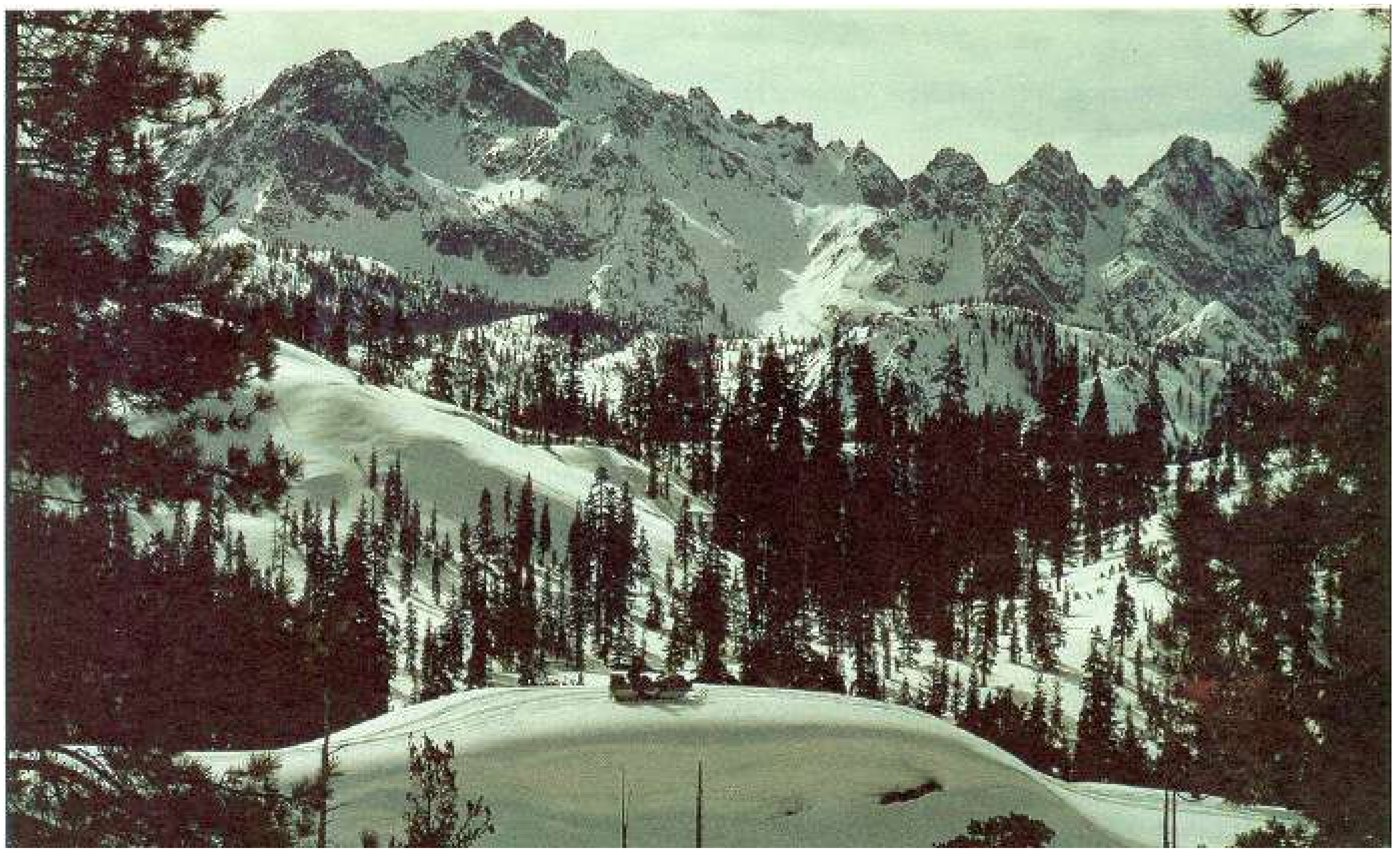

©National Geographic Society, Photo by David Hiser

Figure 11-25. Coniferous Forest.

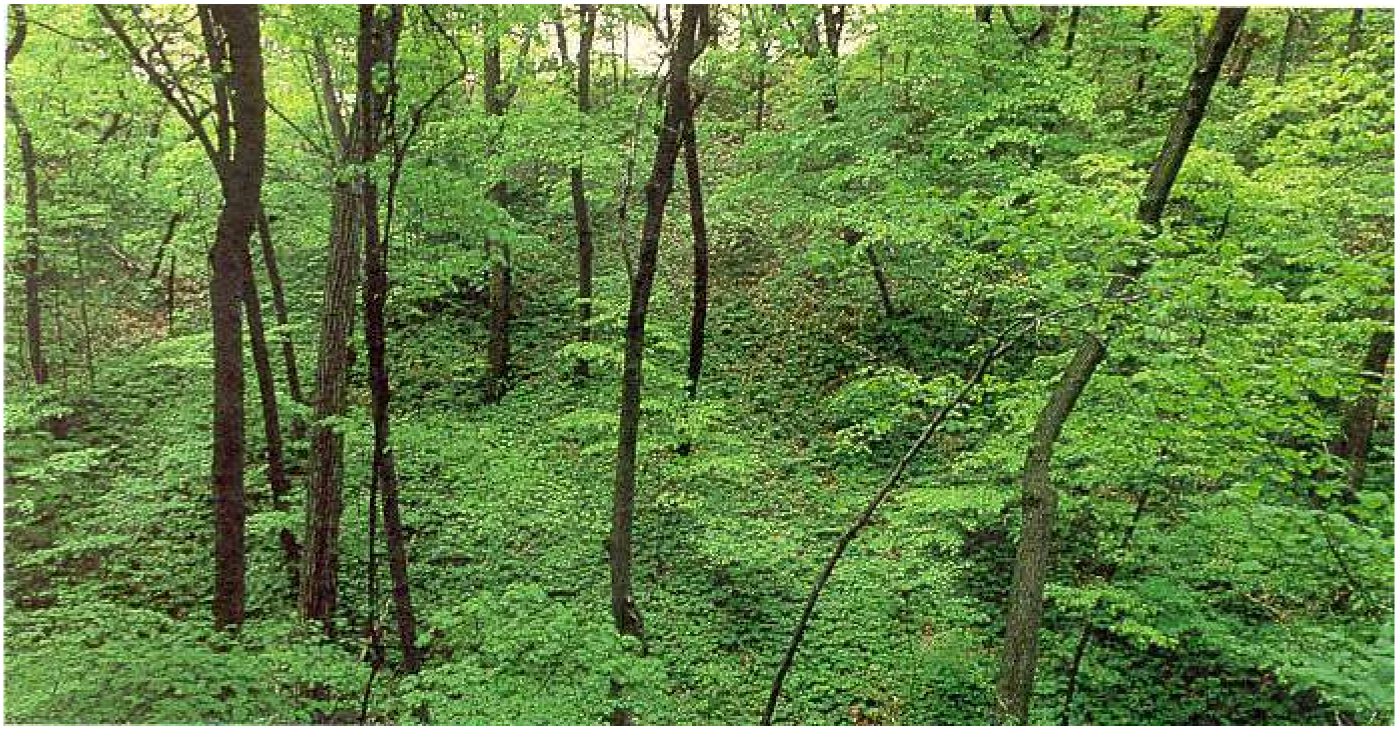

Photo by Daniel Yacko

Figure 11-26. Deciduous Forest.

-1. North America. Western Great Plains of the United States.

-2. South America. Argentina.

-3. Africa. Narrow belt extending across Africa on the southern rim of the Sahara, and parts of Ethiopia and Kenya.

-4. Europe. Southeastern Russia.

-5. Asia, Turkey, Iran, Baluchistan, Pakistan, Turkestan, and a broad belt through central Asia.

-6. Australia. Fringes of the great central desert, especially in eastern Australia.

(b) The prairie and steppe areas are very closely related. However, the true prairie supports a somewhat different flora than the steppe areas, and for this reason, it is important that they be discussed separately. The chief distinction between prairie and steppe is the seasonal distribution of rainfall.

	PRAIRIE	STEPPE
Rainfall per year	30-40 inches	15-30 inches
Subsoil	Permanently moist	Permanently dry

In both, the precipitation comes during the short growing season (spring). Summers are hot with intermittent showers. The primary prairie regions of the world are:

-1. North America. South central Canada, and east central United States.

-2. South America. Northeastern Argentina, Uruguay, Paraguay, and Brazil.

-3. Africa. Union of South Africa.

-4. Europe. Parts of Hungary, Rumania, and Russia (Ukraine and in a belt extending through central Russia to the Urals).

-5. Asia. Manchuria.

(c) The main plants in these biomes are grasses. Due to different conditions various characteristic grasses grow in specific areas on the prairies. The tall grasses are found near the edges of deciduous forests where larger amounts of water are available. The midgrasses grow farther west, close to the Great Basin within the United States with short grasses growing in the rain shadows of the mountains. Wild flowers and other annuals are found throughout these regions. On the fringes of the desert, desert plants may have moved into the grasslands. The following are food plants of the steppes:

-1.Sweet Acacia.
-2.Wild Chicory.
-3.Wild Rose.
-4.Amaranth.
-5.Chufa (Nut Grass).
-6.Wild Sorrel.
-7.Baobab.
-8.Wild Dock.
-9.Wild Tulip.
-10.Wild Calla (Water Arum).
-11.Wild Onion.
-12.Water Lily.
-13.Cattail.
-14.Sea Orach.

(d) Common herbivores of the prairies are ground squirrels, prairie dogs, rabbits, gophers, and a great many species of mice. These are preyed upon by

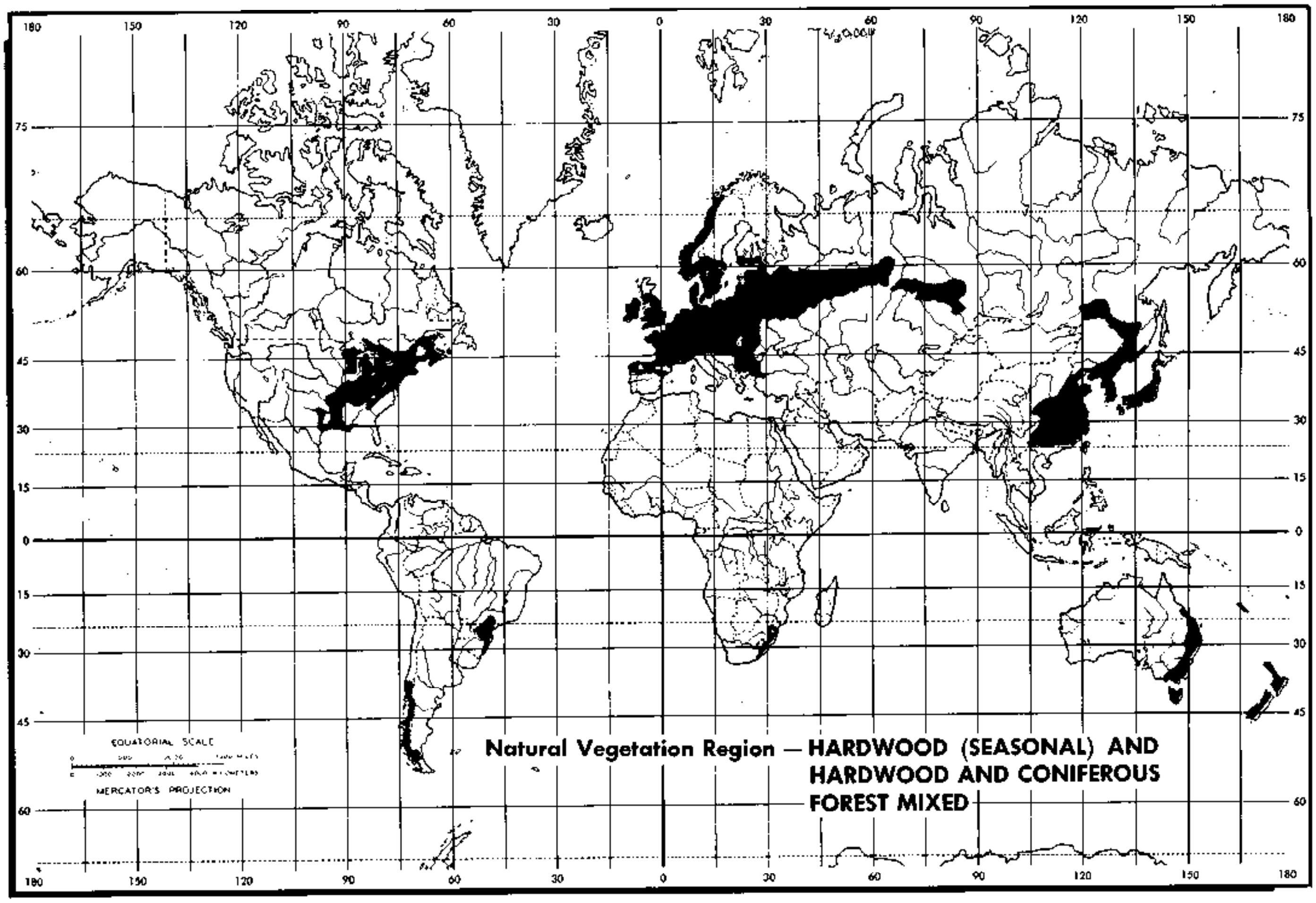
Natural Vegetation Region — HARDWOOD (SEASONAL) AND HARDWOOD AND CONIFEROUS FOREST MIXED
EQUATORIAL SCALE
MERCATOR'S PROJECTION
180
150
120
90
60
30
0
30
60
90
120
150
180
75
60
45
30
15
0
15
30
45
60

Figure 11-27. Hardwood (Seasonal).

badgers, coyotes, foxes, skunks, and hawks. Prairie animals travel in packs or herds which serve both to protect their individual members and assist in hunting prey. They typically have excellent vision and sense of smell, but their hearing, though keen, may be impaired by the noise of the pack or herd.

(e) A number of birds nest among the grasses. These include the meadowlark, prairie chicken, and grouse. During the dry season, some of these birds migrate to places better suited to raising their young.

(f) Insects like grasshoppers are well adapted to a grassland environment. The natural enemies of such insects are birds and reptiles which in turn become the prey of owls and hawks.

(5) Evergreen Scrub Forests. These biomes occur in southern California, in countries around the Mediterranean Sea, and in southern portions of Australia and correspond with the Mediterranean climate (figure 11-29).

(a) The major life form in this area is vegetation composed of broad-leaved evergreen shrubs, bushes, and trees usually less than 8 feet tall. This vegetation generally forms thickets. Sage and evergreen oaks are the dominant plants in North America in areas with rainfall between 20 and 30 inches. Areas with less rainfall or poorer soil have fewer, more drought-resistant shrubs such as manzanita. Scrub forest vegetation becomes extremely dry by late summer. The hot, quick fires that commonly occur during this period are necessary for germination of many shrub seeds and also serve to clear away dense ground cover. This ground cover is difficult for the survivor to penetrate. The branches are tough, wiry, and difficult to bend. Trees are usually widely scattered, except where they occur in groves near a stream. Usually, both trees and shrubs have undivided leaves. Grasses and brightly colored spring-flowering bulbs and other flowers may also be found. The survivor will find relatively few kinds of edible plant food within the scrub forest. During the growing season—usually only the spring months—the following kinds of plant foods are available:

-1.Agave (Century Plant). (See figure 11-30.)

-2.Wild Dock.

-3.Wild Pistachio.

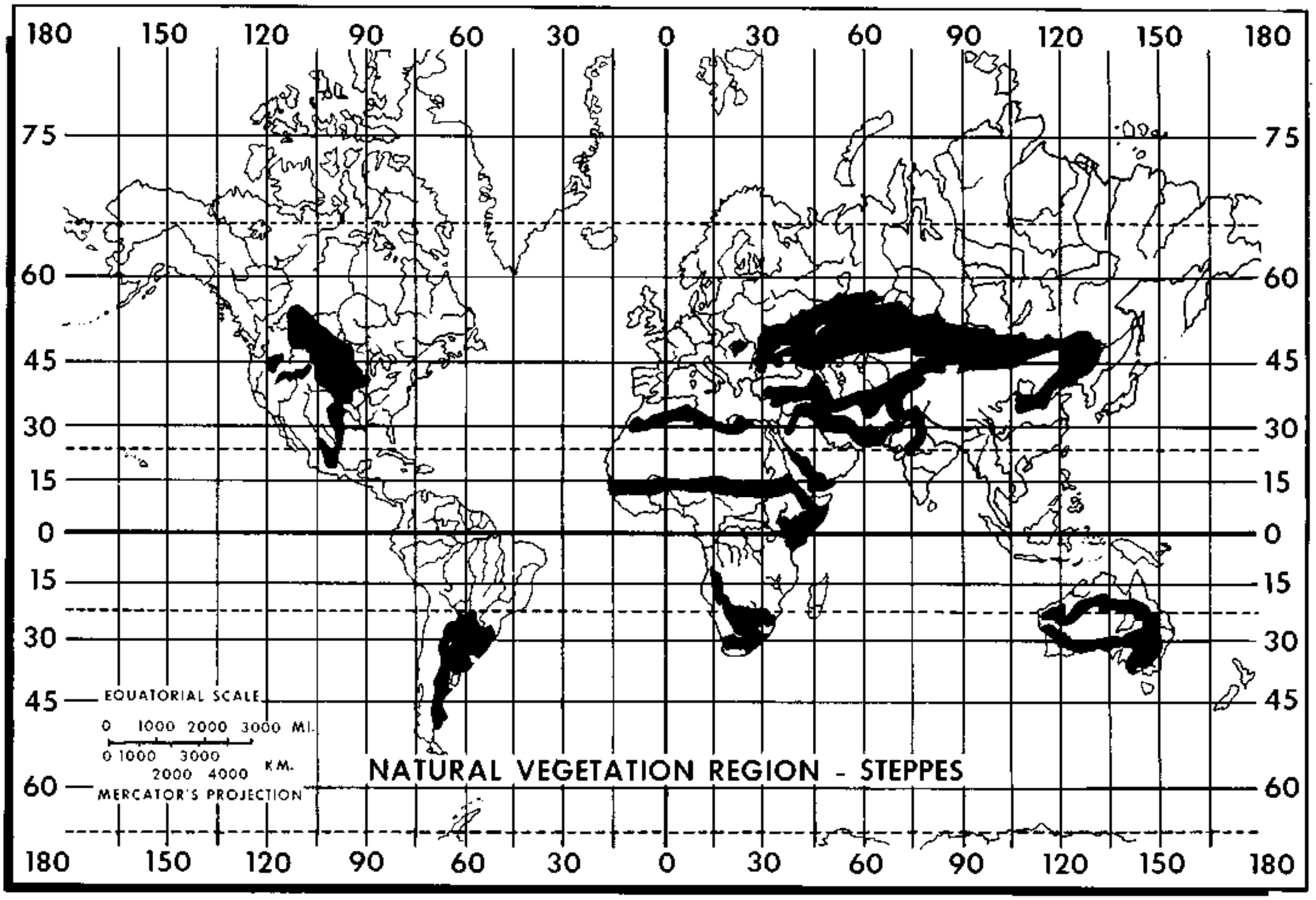

Figure 11-28. Steppes.

-4.Almond.
-5.Wild Grape.
-6.Wild Sorrel.
-7.Wild Apple.
-8.Juniper.
-9.Wild Tulip.
-10.Chicory.
-11.English Oak (Acorn).
-12.Walnut.
-13.Wild Crabapple.
-14.Wild Onion.

(b) Deer and birds usually inhabit these forests only during the wet season, which is the growth period for most scrub forest plants. Small dull colored animals such as lizards, rabbits, chipmunks, and quail are year-round residents.

11-6. Snow Climates. "Snow climates" are defined as the interior continental areas of the two great landmasses of North America and Eurasia that lie between 35 and 70 degrees north latitude. The tree line provides the best natural boundary for a topographical description of the snow climate areas. There are definite differences between the forest area to the south and the tundra to the north in snow-cover characteristics, wind conditions, animal types, and vegetation. Snow climates are comprised of two separate climate types: continental subarctic and humid continental.

a. The continental subarctic climate is one of vast extremes. The temperature may range from -108°F to 110°F. Temperatures may also fluctuate 40 to 50 degrees within a few hours. This area includes several climate subtypes. The largest areas run from Alaska to Labrador and Scandinavia to Siberia. They are cold, snowy forest climates, moist all year, with cool, short summers. A colder climate is found in northern Siberia which has very cold winters with an average cold temperature of -36°F. Another area is found in northeastern Asia where the climate is a cold, snowy forest climate with dry winters. Winter is the dominant season of the continental subarctic climate. Because freezing temperatures occur for 6 to 7 months, all moisture in the ground is frozen to a depth of many feet.

b. The humid continental climates are generally located between 35 and 60 degrees north latitude. For the most part, these climates are located in central and eastern parts of continents of the middle latitudes. These climates are a battle zone of polar and tropical

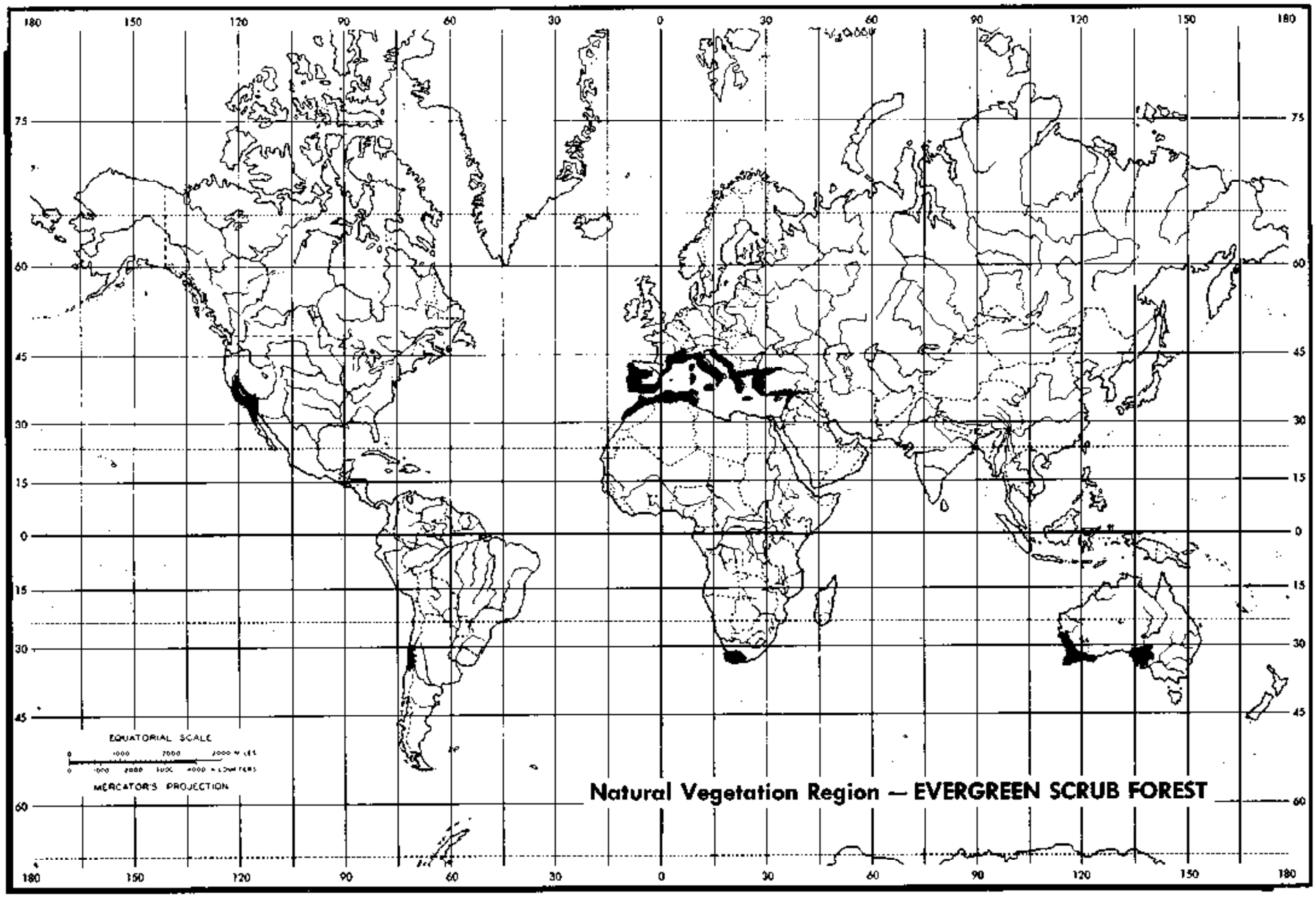

Figure 11-29. Evergreen Scrub Forest.

airmasses. Seasonal contrasts are strong and the weather is highly variable. In North America, this climate extends from New England westward beyond the Great Lakes region into the Great Plains and into the prairie provinces of Canada. This climate can also be found in central Asia. The summers are cooler and shorter than in any other climate in the temperate zone with the exception of the highland (Alpine) subarctic climate. The summer temperatures range from 60°F to 70°F. The winter temperatures range from -15°F to 26°F. The precipitation for the year varies from 10 to 40 inches. A higher percentage of the precipitation is snow, with less snow occurring in areas along the coasts. The weather is influenced by the polar easterly winds and the subtropical westerly winds. The effect of ocean currents on this continental climate is minimal. This climate is dominated by the high- or low-pressure cells centered in interiors of the continent.

c. Both climate regions have seasonal extremes of daylight and darkness resulting from the tilt of the Earth's axis (figure 11-31). Snow climate nights are long, even continuous in winter; conversely, north of the Arctic Circle, the Sun is visible at midnight at least once a year. Darkness presents a number of problems to the survivor. No heat is received directly from the Sun in midwinter, thus the cold reaches extremes. Outside activities are limited to necessity, although the light from the Moon, stars, and auroras, shining on a light ground surface, is of some help. Confinement to cramped quarters adds boredom to discomfort, and depression becomes a dominant mood as time drags on. Fortunately, the period of complete darkness does not last long.

d. The terrain of the snow climate areas coincides with a great belt of needle-leaf forests. This region is found in the higher middle latitudes. Its poleward side usually borders on tundra and its southern margin usually adjoins continental temperate climates. This area is like the tundra because it has poor drainage. As a result, there are an abundance of lakes and swamps. The coastlines vary from gentle plains sweeping down to the ocean to steep, rugged cliffs. Glaciers are a predominate feature of the high altitudes (6,000-feet elevation or above). These glaciers flow down to lower elevations or terminate at the ocean.

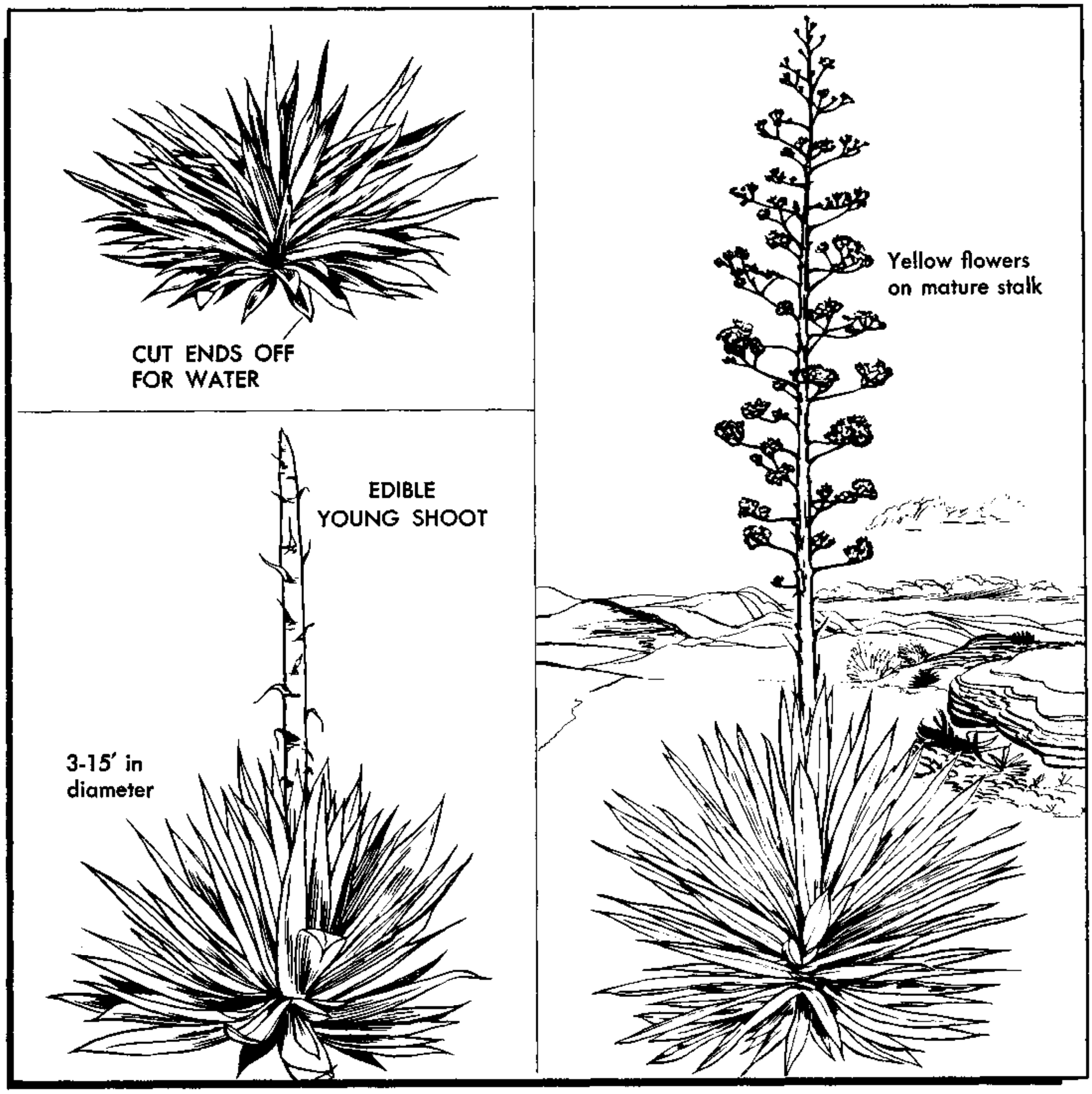

Figure 11-30. Agave (Century Plant).

e. The vegetation is similar to that found in more temperate zones; however, the cold temperatures have caused variations in the physical appearance of the plants. Dark evergreen forests thrive south of the tree line. They consist mainly of cedar, spruce, fir, and pine, mingled with birch. These subarctic forests are called taigas. A transitional zone lies between the taiga and the tundra. In this zone, the trees are sparse and seldom grow over 40 feet tall. Dwarf willow, birch, and alder mix with evergreens, and reindeer moss sometimes forms a thick carpet (figure 11-32).

f. Depending on the time of year and the place, chances for obtaining animal food vary considerably. Shorelines are normally scraped clean of all animals and plants by winter ice. Inland animals are migratory.

(1) Large Arctic Game. Caribou and reindeer migrate throughout northern Canada and Alaska (figure 11-33). In northern Siberia, they migrate inland to al-

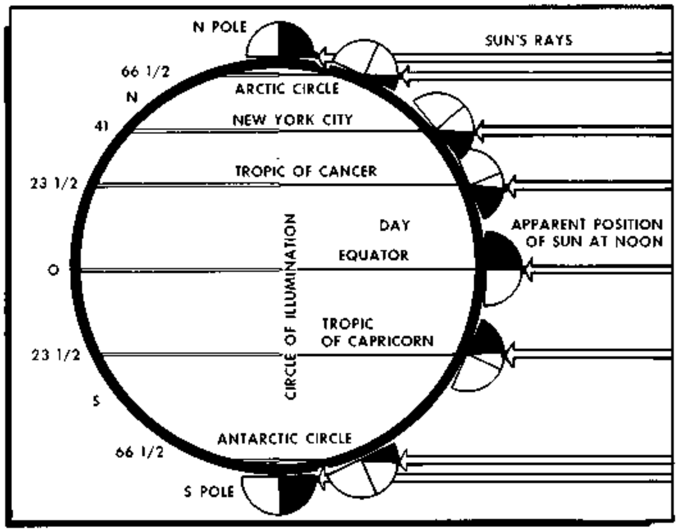
N POLE
SUN'S RAYS
66 1/2
N
ARCTIC CIRCLE
41
NEW YORK CITY
TROPIC OF CANCER
23 1/2
CIRCLE OF ILLUMINATION
DAY
APPARENT POSITION OF SUN AT NOON
EQUATOR
0
TROPIC OF CAPRICORN
23 1/2
S
ANTARCTIC CIRCLE
66 1/2
S POLE

Figure 11-31. Circle of Illumination.

most 50 degrees north latitude. Some are found in west Greenland. All move close to the sea or into the high mountains in summer. In winter, they feed on the tundra. Musk oxen may be found in northern Greenland and on the islands of the Canadian archipelago. Sheep descend to lower elevations and to valley-feeding grounds in the winter. Wolves usually run in pairs or groups. Foxes are solitary and are seen most frequently when mice and lemming are abundant. Bears are dangerous, especially when wounded, startled, or with their young. They generally shun areas of human habitation.

(2) Small Land Game. Tundra animals include snowshoe and arctic hare, lemming, mice, and ground squirrels. They may be trapped or shot in the winter or summer. Most prefer some cover and can be found in shallow ravines or in groves of short willows. Ground squirrels and marmots hibernate in the winter. In summer, ground squirrels are abundant along sandy banks of large streams. Marmots live in the mountains, among the rocks, usually near the edge of a meadow or in deep soil—much like woodchucks. To find the burrow in rocky areas, look for a large patch of orange-colored lichen on rocks. This plant grows best on animal or bird dung, and the marmot always seeks relief in the same spot, not far from a well-hidden entrance.

g. The arctic is the breeding ground for many birds. In summer, ducks, geese, loons, and swans build their nests near ponds on the coastal plains or bordering lakes

Photo by Daniel Yacko

Figure 11-32. Subarctic Forest.

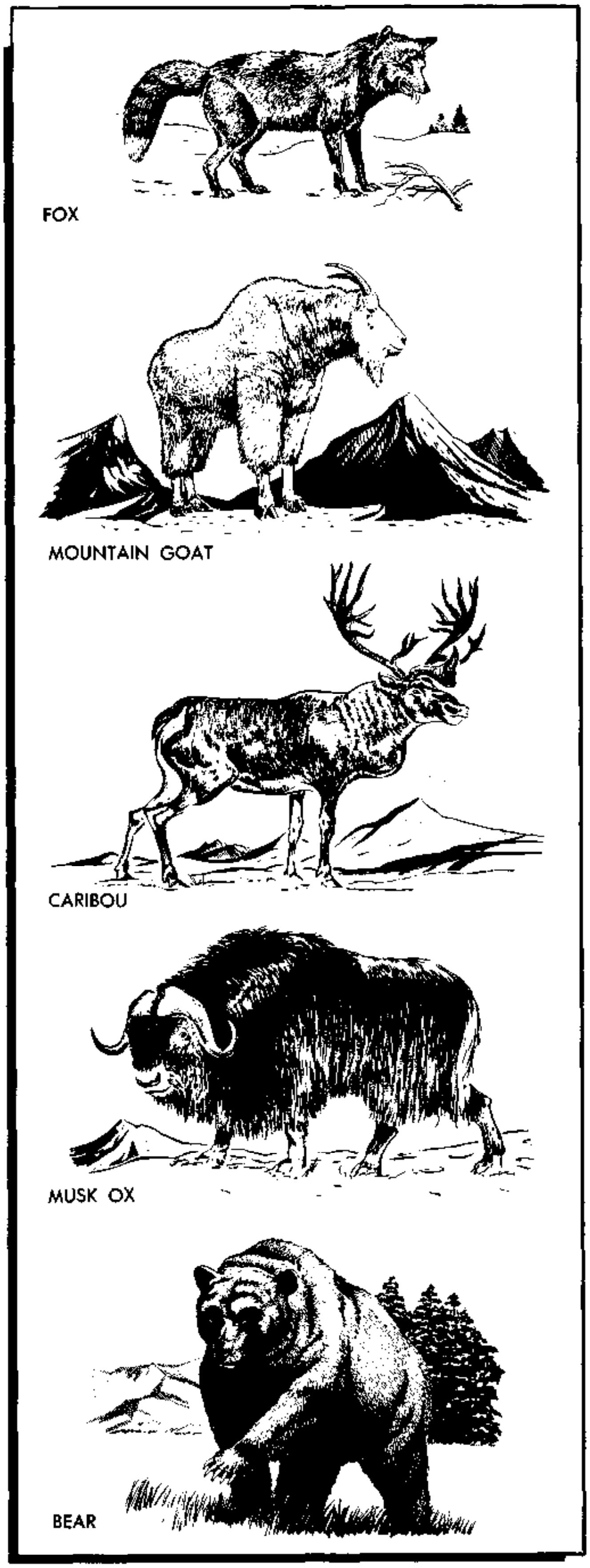

Figure 11-33. Arctic Game.

or rivers of the low tundra. A few ducks on a small pond usually indicates that setting birds may be found and flushed from the surrounding shores. Swans and loons normally nest on small, grassy islands in the lakes. Geese crowd together near large rivers or lakes. Smaller wading birds customarily fly from pond to pond. Grouse and ptarmigan, are common in the swampy forest regions of Siberia. Sea birds may be found on cliffs or small islands off the coast. Their nesting areas can often be located by their flights to and from their feeding grounds. Jaeger gulls are common over the tundra, and frequently rest on higher hillocks. In the winter, fewer birds are available because of migratory patterns. Ravens, grouse, ptarmigan, and owls are the primary birds available. Ptarmigan are seen in pairs or flocks, feeding along grassy or willow-covered slopes.

h. Arctic and tom cod, sculpin, eelpout, and other fish may be caught in the ocean. The inland lakes and rivers of the surrounding coastal tundra generally have plenty of fish which are easily caught during the warmer season. In the North Pacific and in the North Atlantic extending slightly northward into the Arctic Sea, the coastal waters are rich in all seafoods. Varieties include fish, crawfish, snails, clams, and oysters, and one of the world's largest and meatiest crabs—the king crab of the Aleutian Islands and Bering Sea areas. In the spring (breeding season), this crab comes close to shore and may be caught on fish lines set in deep water or by lowering baited lines through holes cut in the ice. Do not eat shellfish that are not covered at high tide. Never eat any type of shellfish that is dead when found, or any that do not close tightly when touched. Poisonous fish are rarer in the arctic than in the tropics. Some fish, such as sculpins, lay poisonous eggs; but eggs of the salmon, herring, or freshwater sturgeon are safe to eat. In arctic or subarctic areas, the black mussel may be very poisonous. If mussels are the only available food, select only those in deep inlets far from the coast. Remove the black meat (liver) and eat the white meat. Arctic shark meat is also poisonous (high concentration of vitamin A).

11-7. Ice Climates. There are three separate climates in the category of ice climates: marine subarctic climate, tundra climate, and icecap climate (figure 11-34).

a. Marine Subarctic Climate. Key characteristics of this climate are the persistence of cloudy skies and strong winds (sometimes in excess of 100 miles per hour) and a high percentage of days with precipitation. The region lies between 50 and 60 degrees north latitude and 45 to 60 degrees south latitude. The marine subarctic climate is found on the windward coasts, on islands, and over wide expanses of ocean in the Bering Sea and the North Atlantic, touching points of Greenland, Iceland, and Norway. In the Southern Hemisphere, this climate is found on small landmasses.

b. Tundra Climate. The tundra region lies north of 55 degrees north latitude and south of 50 degrees south latitude. The average temperature of the warmest month is below 50°F. Proximity to the ocean and persistent cloud cover keep summer air temperatures down despite abundant solar energy at this latitude near the summer solstice (figures 11-35 and 11-36).

c. Icecap Climate. There are three vast regions of ice on the Earth. They are Greenland and Antarctic continental icecaps and the larger area of floating sea ice in the Arctic Ocean. The continental icecaps differ in various ways, both physically and climatically, from the polar sea ice and can be treated separately (figure 11-37).

(1) Greenland. The largest island in the world is Greenland. Most of the island lies north of the Arctic Circle and ice covers about 85 percent of it. The warmest region of the island is in the southwestern coast. The average summer temperature is 50°F. The coldest region is the center of the icecap. The temperature there averages -53°F in the winter.

(2) The Antarctic. The Antarctic lies in a unique triangle formed by South America, Africa, and Australia. Surrounding the continent are portions of the Atlantic, Pacific, and Indian Oceans. The area is almost entirely enclosed by the Antarctic Circle. The climate is considered as one of the harshest in the world. The average temperature remains below 0°F all year. In the winter months, the mean temperature is from -40°F to -80°F. Winter temperatures inland often drop below -100°F. Great storms and blizzards (with accompanying high winds) range over the entire area due to both the continent's great elevation and by being completely surrounded by warm ocean water.

(3) Sea Ice on the Arctic Ocean. Ice on the Arctic Ocean includes frozen sea water and icebergs that have broken off glaciers. This ice remains frozen near the North Pole year around. Near the coast, the sea ice melts during the summer. Currents, tides, and winds may cause it to fold and form high ridges called pressure ridges. One piece of ice may slide over another causing a formation called rafted ice. When the ice breaks into sections separated by water, these sections are called leads. Great explosions and rolling thunder are caused by the breaking and folding of the ice.

d. Terrain. The terrain of the true ice climates encompasses nearly every variation known. Much of the landmass is composed of tundra. In its true form, the tundra

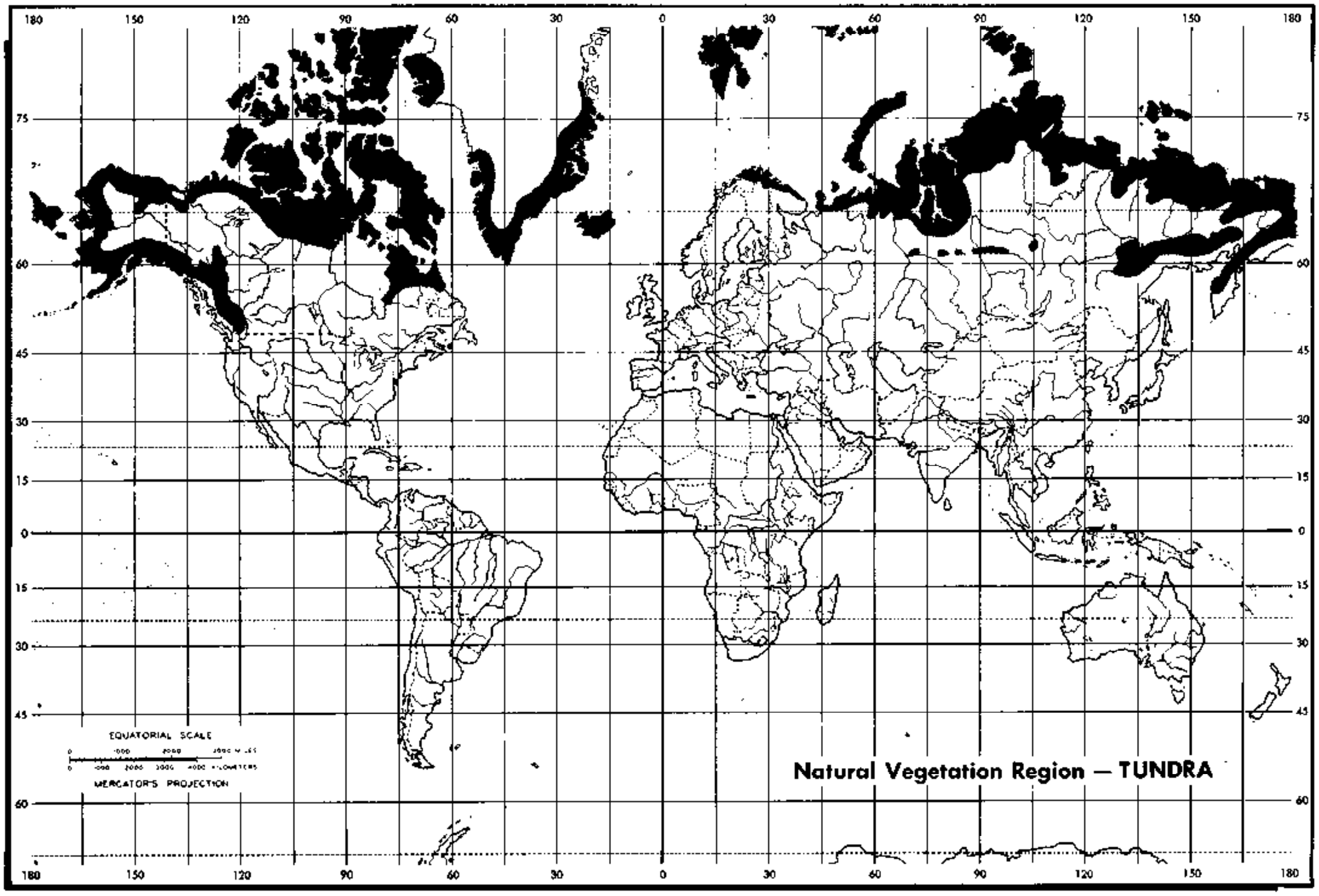

Figure 11-34. Tundra.

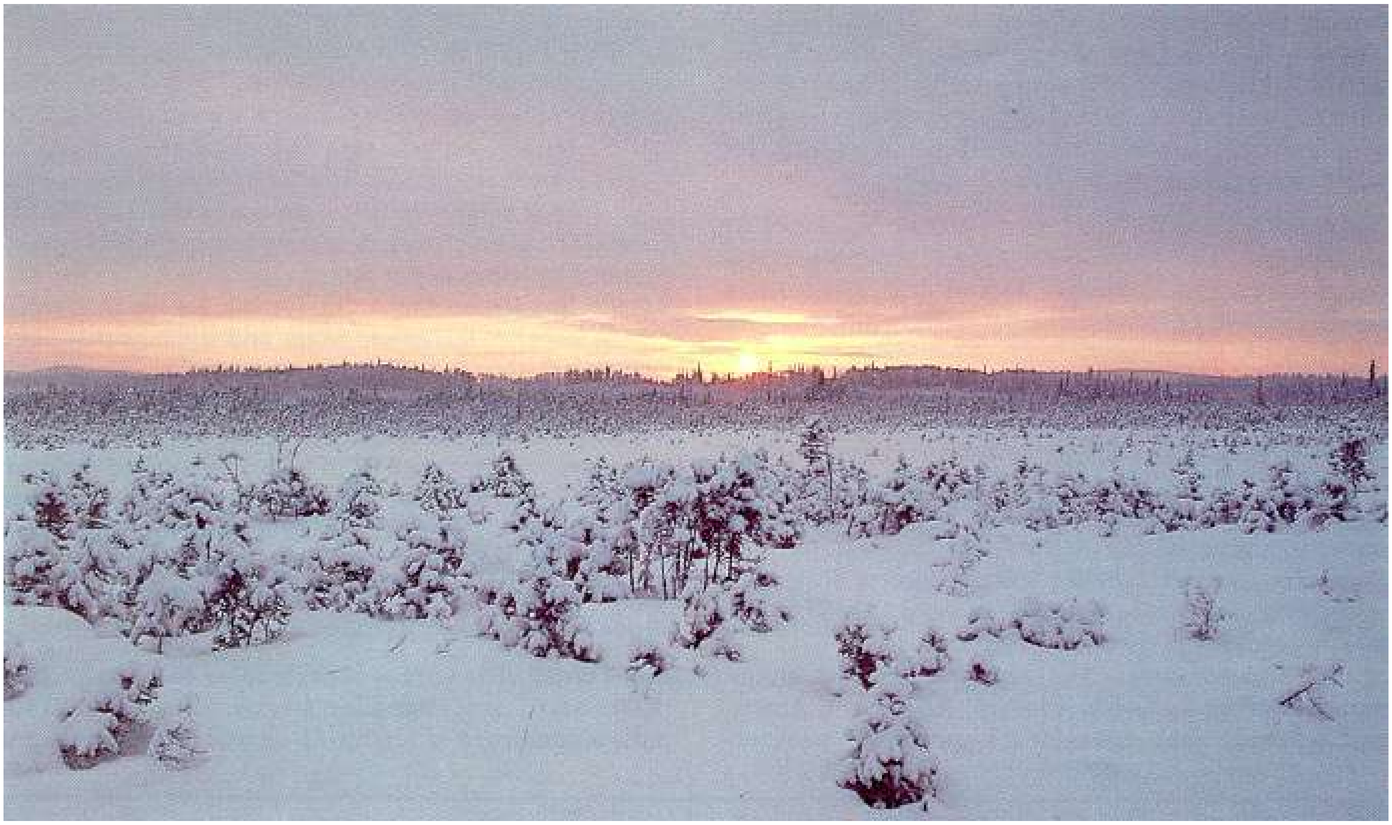

Photo by James Tourtillotte

Figure 11-35. Winter Tundra.

is treeless. Vast rugged mountain ranges are found in the area and rise several thousand feet above the surrounding areas. Steep terrain, snow and ice fields, glaciers, and very high wind conditions make this area a very desolate place. Continental glaciers such as the icecaps covering Greenland and the Antarctic continent are large expanses of wind-swept ice moving slowly toward the sea. Ice thickness in continental glacier areas can exceed 10,000 feet.

e. Vegetation:

(1) Shrub Tundra. In Russia, the area surrounding the Lena River is known as a typical shrub tundra environment. Shrubs, herbs, and mosses occur in this zone. Arctic birch predominates but other shrubs occur and several may be useful as supplementary food such as the crystal tea ledum (Labrador tea), willows, and the bog bilberry. In this same shrub zone a lower herbaceous layer occurs which is composed of black crowberry, several grasses, and the cowberry. On the ground, mosses and lichens are present in abundance. The shrubs on the open tundra reach a height of only 3 to 4 feet but in valleys and along the rivers, the same shrubs may reach the height of a person.

(2) Wooded Tundra. The region immediately adjoining the treeless tundra is an extension of the coniferous areas of the south. These subarctic wooded areas include a variety of tree species of which the genus Picea (Spruce) predominates. On the Kola Peninsula of northeastern Scandinavia, these northernmost forests are birch. Siberian spruce occurs between the White Sea and the Urals. Siberian larch occurs between the Urals and the Pyasina River. Dahurian larch occurs between the Pyasina River and the upper reaches of the Anadyr River. In extreme northeastern Asia, Mongolian poplar, Korean willow, and birch are found along the rivers. The trees extending into the tundra are distinguished by their stunted growth (except in river valleys where they reach 18 to 24 feet) and sparseness. Permanent ground frost, or permafrost, penetrates most parts of the true tundra and the northern limits of the forest belt closely coincide with the southern limits of permafrost. A few different plants will cover very large areas so that extensive stands of a single variety of plant are common in the arctic tundra. All tundra plants are small in stature compared to the plants in the warmer climates of more southerly latitudes. The arctic willow and birch, for instance, spread along the ground in the tundra to form

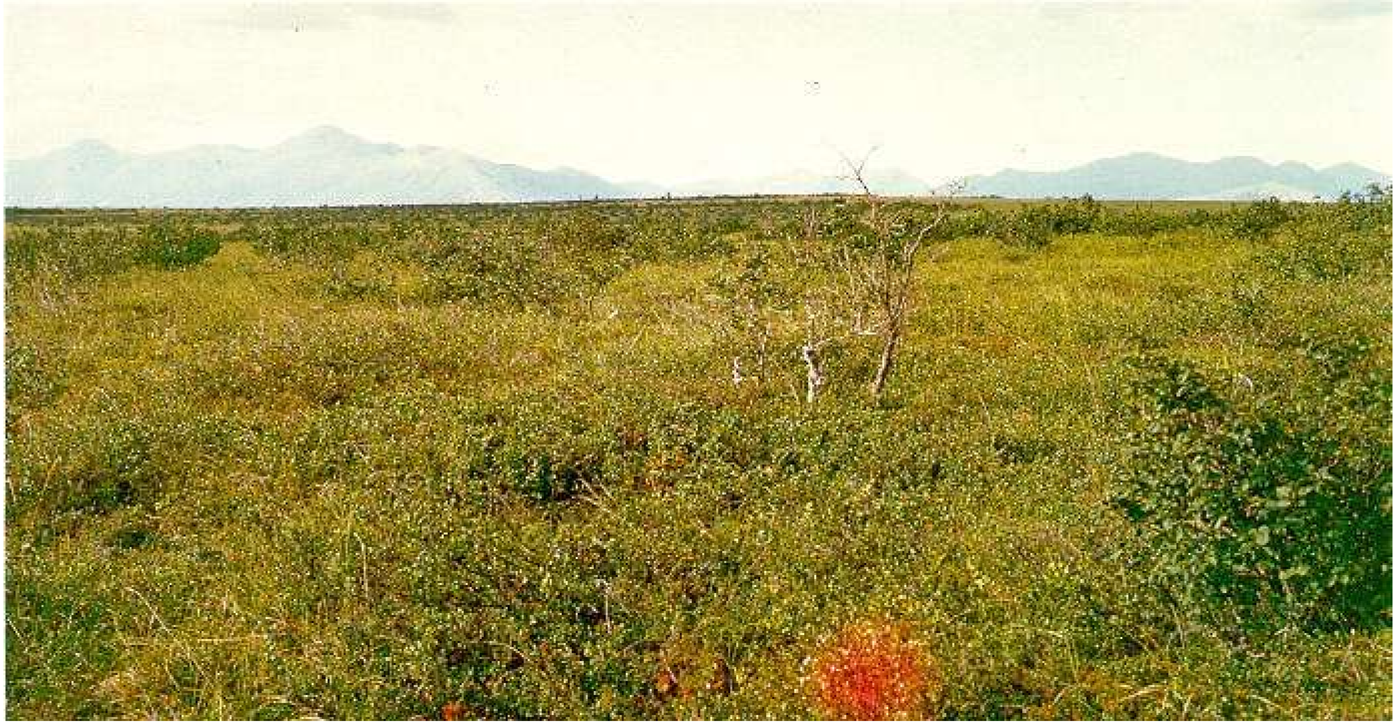

Figure 11-36. Summer Tundra.

large mats. Stunted growth in all the woody plants is the rule, although there are many evergreen plants and hardy bulbous or tuberous plants. Lichens, especially reindeer moss, are widespread in the tundra. As mentioned before, the plant life of the tundra is remarkably uniform in its distribution. Some species are common to all three areas, but other species are more restricted in their distribution. The tundra also contains many species of vegetation found in the forest regions to the south (figure 11-38).

(3) Bogs. The tundra has often been classified as a continuous bog, but this is far from the truth. Many bogs do exist. There are also many hilly and even mountainous areas with considerably drier soil. The moss, or sphagnum bog is less common than the sedge bog. A characteristic of more southern tundras is the development of large peat mounds 9 to 15 feet high and 15 to 75 feet in diameter. These mounds have been formed by ground upheavals caused by freezing water. Many edible plants grow on these bog mounds, such as the cloudberry, dwarf arctic birch, bog bilberry, black crowberry, crystal tea ledum, sheathed cotton sedge, cowberry, and others.

f. Animal Life. Compared to other parts of the world, animal life is poor in species but rich in numbers. Large animals such as caribou, reindeer, and musk oxen migrate through the tundra areas. Carnivores—wolves, foxes, lynx, wolverines, and bears—range through the landmass area and polar bears, seals, walruses, and foxes are found far out on the sea ice. Small animals are the most abundant animal life found and include hares, lemmings, marmots, mink, fishers, and porcupines.

(1) Bird life is very limited during the winter months, mainly owls and ptarmigan, but during the summer months millions of migratory waterfowl nest in the arctic tundra. Species include ducks, geese, cranes, loons, and swans, nesting in and around the swamps, bogs, and lakes of the tundra. The coastal areas are home for many species of sea birds during the summer months. The coastal waters and iceflows are rich in a variety of marine life such as seals, walruses, whales, crustaceans, and fishes.

(2) The freshwater rivers, lakes, and streams are teeming with many varieties of fish—salmon, trout, and grayling. Due to the amount of surface water in the tundra area, there are a large variety of insects. Some 40 to 60 species of mosquitoes, flies, and gnats inhabit the area.

(3) In the Antarctic, animals are virtually nonexistent. Only the lowest forms of animal life can live mainly on mosses and lichens. Marine animals, particularly whales, seals, and penguins, are found along the coastal regions. Sea birds are abundant in the summer and nest on the coastal regions and the islands. There are a few species of wingless insects, lice, ticks, mites, etc., which live off the bird population.

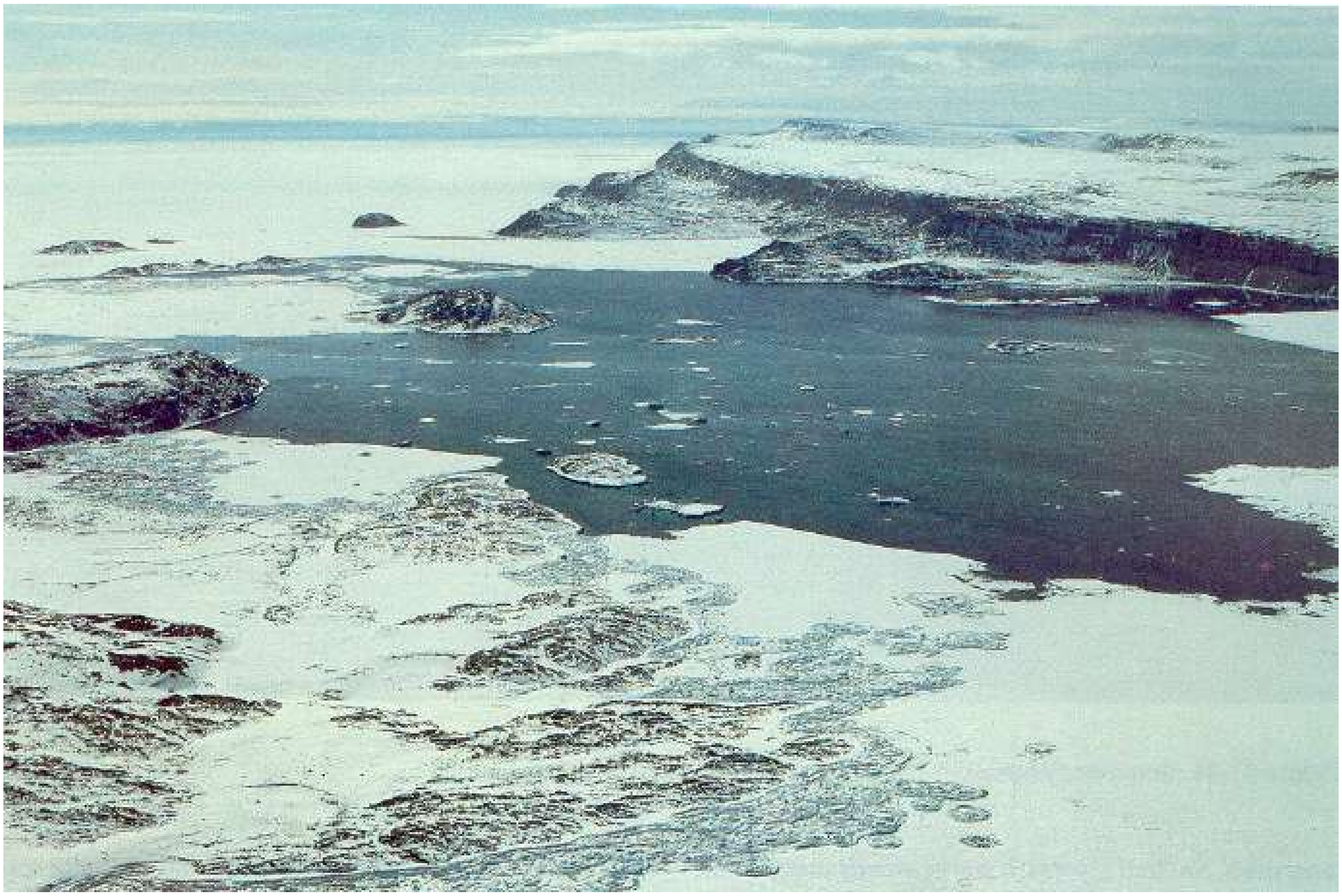

©National Geographic Society, Photo by Peter Schledermann

Figure 11-37. Ice Climate.

11-8. Open Seas. The Köppen-Geiger system of climate classification has been used to describe the environmental characteristics of the landmasses. However, this system is not used to categorize the largest area of the world—the oceans. They are simply divided by their names and locations. All limits of oceans, seas, etc., are arbitrary, as there is only one global sea. The terms "sea" and "ocean" are often used interchangeably in reference to saltwater. However, from a geographic point of view, a sea is a body of water that is substantially smaller than an ocean or is part of an ocean.

a. The seas cover 70.8 percent of the Earth's surface. The waters are not evenly distributed, covering 61 percent of the surface in the Northern Hemisphere and 81 percent in the Southern Hemisphere. Traditionally, the seas are divided into four oceans: Atlantic, Pacific, Indian, and Arctic. These, with their fringing gulfs and smaller seas, make up the world's seas. If the land features of the Earth were smoothed out, the seas would cover the entire globe to a depth of 12,000 feet. Mount Everest, the tallest mountain peak (29,028 feet), would disappear without a trace in the 37,800-foot deep Marianas Trench in the western Pacific Ocean. The sea floor is made up of mountains, valleys, great plains, and deep trenches. The deepest trenches and tallest mountains are found in the north Pacific. The sea floor features do, to some extent, influence the surface properties of the seas; that is, currents, waves, and tides.

b. The average salinity of the seas is usually taken as 3.5 percent. Higher values occur at or near the surface in areas where high temperatures and strong, dry winds favor evaporation. The highest salinities occur in semilandlocked seas at mid-latitudes such as the Red Sea, the Persian Gulf, and the Mediterranean Sea. The Pacific Ocean is the largest and is over twice the size of the Atlantic or Indian Oceans. Its size allows for greater climatic variations and more widespread influence. Due to similar latitudinal references the Atlantic and Pacific Oceans have many similar characteristics. The Indian Ocean is slightly smaller than the Atlantic, but is more significantly influenced by a continental landmass than

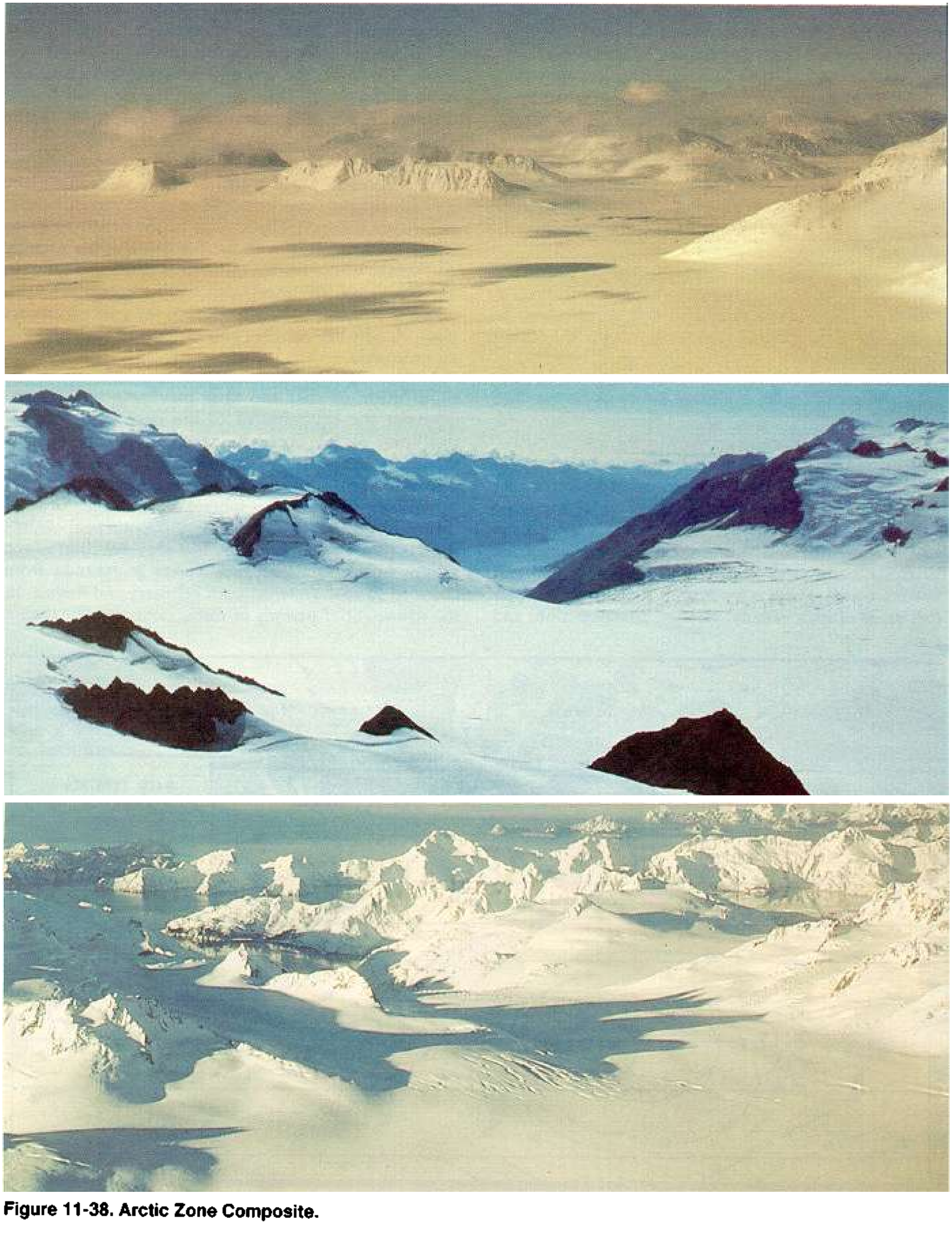

Figure 11-38. Arctic Zone Composite.

any other ocean. The Arctic Ocean is generally recognized as that body of water which lies north of 75 degrees latitude and is nearly enclosed by landmasses.

c. Within each of these four major oceans, numerous subdivisions known as seas, may be geographically aligned along indistinct boundaries (island chains; geography of ocean floor). Examples are:

(1) The Coral Sea is an arm of the South Pacific Ocean lying east of Queensland, Australia, and west of New Hebrides and New Caledonia. It extends from the Solomom Islands on the north to the Chesterfield Islands on the south.

(2) The Bering Sea is located between Alaska and Eastern Siberia, with its southern boundary formed by the arc of the Alaskan Peninsula and the Aleutian Islands. The Bering Strait connects it with the Arctic Ocean to the North.

d. Many water bodies are partially enclosed by land and are known as gulfs. An example would be the Gulf of Mexico.

11-9. Ocean Currents. The ocean has a complex circulation system made up of a variety of currents and countercurrents. These currents move at a rate from barely measureable to about 5.75 miles per hour. They may be relatively cold or warm currents and influence the climate and environment that exists on land and over the ocean. There is a constant movement of water from areas of high density, salinity, concentration, and pressure to areas of low density, salinity, concentration, and pressure in an attempt to establish an equilibrium. These factors influence the movement of ocean currents. However, the primary influence on ocean currents is the wind. They also may be diverted by the Coriolis Force and Continental Deflection (figure 11-39).

11-10. Climatic Conditions. To fully understand the general climatic conditions and seasonal variations that exist over the global sea, each major ocean must be examined separately, with the exception of the Atlantic and Pacific whose similar latitudinal references result in like characteristics (exceptions will be noted). The two physical phenomenons which have the greatest impact upon climate are currents and systems of high and low air pressures.

a. Currents with their basic characteristics of being either warm or cold and their inevitable convergence influence the environment of the open seas. Equally significant as the ocean influence on typical weather sequences (for example, temperature, wind, precipitation, and storms) are semi- and quasi-permanent centers of high and low atmospheric pressures. To observe their effect on climate, imagine a hypothetical voyage from the Pole to the Equator. The southern limit of the solidly frozen arctic icepack varies in latitude from about 65 to 75 degrees between February and August. In the winter, brief periods of calm, clear weather with a

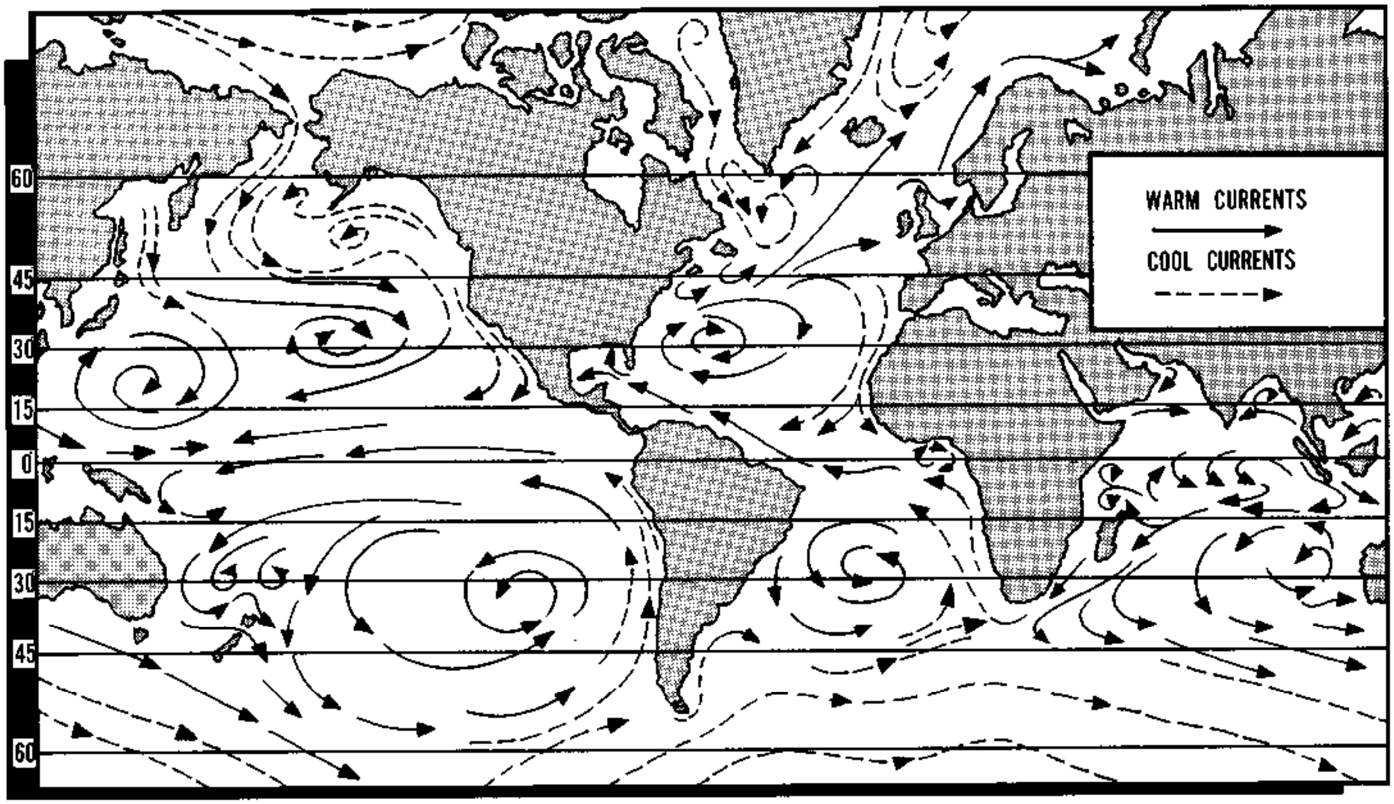

Figure 11-39. Ocean Currents.

mean temperature of -5°F are interspersed between passages of cyclonic storms characterized by snow, winds 30 to 40 mph, temperatures from -20°F to -30°F, and gale-force winds 30 percent of the time. In the summer, frequent periods of several days of calm or light variable winds with temperatures in the mid-40's may be experienced. Skies are uniformly overcast with layers of stratus or nimbostratus clouds. Dense fog banks are prevalent during calms. Rain or drizzle may continue for weeks at a time. One of the most stormy regions in the hemisphere is in the middle of the prevailing westerlies at 50 degrees north latitude. In the winter, calms are rare with winds of 15 to 20 knots and temperature near freezing. Every 2 to 3 days, a pale sun and scattered clouds give way to cumulostratus clouds and rain squalls. Wind intensity may reach 50 to 60 knots with temperatures dropping to -10°F to -15°F as the rain turns to sleet, soft snow, or hail. In the summer, protracted periods of fog, low stratus clouds, and drizzle exist with moderate breezes. The weather improves in the fall with a week or so of calm, clear weather in late September. As we move south to 40 degrees north latitude and the horse latitudes, the semipermanent high-pressure centers result in generally fair, clear weather with a tendency toward dryness. In the winter, temperatures hover near 50°F and summer brings temperatures into the 70's with calms existing one-fourth of the time. Below 25 degrees north latitude, in the heart of the trade wind belt, winds of 5 to 15 mph are normal. Endless bands of cumulus clouds and clear sky exist with little difference between summer and winter. Daytime temperatures range from 70°F to 80°F.

b. In the Atlantic, Pacific, and Indian Oceans between 5 degrees north latitude and 5 degrees south latitude, an equatorial trough of low pressure forms a belt where no prevailing surface winds exist and is known as the doldrums. Instead, the lack of extreme pressure gradients result in shifting winds and calms which exist as much as one-third of the time. Intense solar heating results in violent thunderstorms associated with strong squall winds. The convergence of these equatorial winds and trade winds from the intertropical front can be seen at a great distance because of towering cumulus clouds rising to 30,000 feet.

c. In the vicinity of the intertropical front, heavy convective showers are quite common. Across the Atlantic and Eastern Pacific, the front is usually north of the Equator. Over the western Pacific, west of 180-degrees longitude, the doldrum belt oscillates considerably. Areas north of the Equator receive their heavy rainfall from June to September. Areas south of the Equator receive their heaviest precipitation between December and March. The meteorological sequence described above may be interrupted by periods of extreme weather centered around low pressure.

d. Waterspouts are the marine equivalent of tornadoes attached to the base of a cumulus or cumulonimbus cloud. They are common off the Atlantic and Gulf coasts of the United States and along the coasts of Japan and China during any season. They are usually seen around noon when solor heating is the greatest. They are small in diameter (10 to 100 yards) and short in duration (10 to 15 minutes). Waterspouts generally exhibit less intensity than overland tornadoes.

e. Hurricanes and typhoons are synonyms for tropical cyclones whose maximum winds exceed 75 mph. They occur in the warm western sectors of all oceans during summer and fall. Winds may reach 170 to 230 mph. The lifespan of tropical cyclones ranges from 1 to 2 weeks. In the middle and high latitudes, extratropical cyclones contrast with tropical cyclones in several ways. There is no warm, clear eye, but rather, a cold region of heavy precipitation. Sustained winds are more moderate (70 to 80 mph). Extratropical cyclones may persist for 2 to 3 days at a fixed location.

f. All ocean currents have a profound influence on climate since the properties of the surface largely determine the properties of the various airmasses. The following are a few examples.

(1) The cold water of the Peru or Humbolt currents has a tremendous affect on the climate of Peru and Chile. The cold air that lies over the current is warmed as it reaches land, increasing its capacity to hold moisture. The warm air does not give up the moisture until it passes over the high Andes Mountains. This accounts for the dry climate of the coast of Chili and Peru and a more temperate climate toward the Equator than is usually found in the lower latitudes.

(2) Where the Labrador current contacts the warm gulf stream, fog prevails and steep temperature gradients are present. The northeast coast of North America has much colder climates than the west coast of Europe at the same latitude.

(3) The warm gulf stream current accounts for the continually warm and pleasant weather in the Caribbean Sea and the Gulf of Mexico.

(4) The winds blowing off the warm water of the Norwegian and east and west Greenland currents account for the unusually mild climates in northern Europe. At the same latitude elsewhere, the temperatures are usually much colder.

11-11. Life Forms. Life forms in the seas range from one-celled animals (protozoan) to complex aquatic mammals. The fish and aquatic mammals rule the sea and are of the most concern to anyone in a survival situation on the open seas. The majority of fishes and mammals can be used as food sources, but some must be considered as a hazard to life; such as, sharks, whales, barracudas, eels, sea snakes, rays, and jellyfish.

a. Sharks. (See figure 11-40.)

(1) Most sharks are scavengers, continuously on the move for food. If none is available, they lose interest and swim on. Even in warm oceans where most attacks

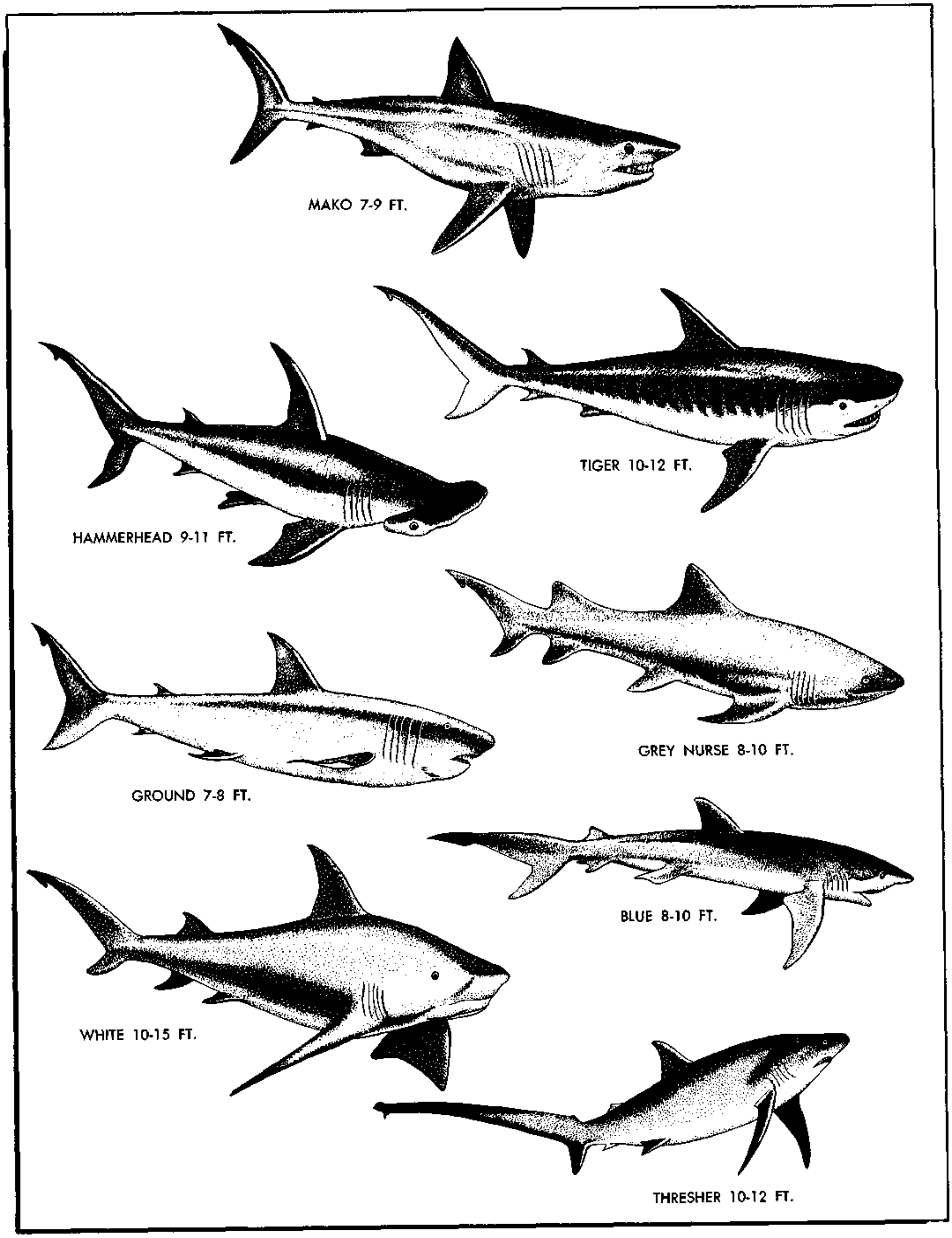
MAKO 7-9 FT.
TIGER 10-12 FT.
HAMMERHEAD 9-11 FT.
GREY NURSE 8-10 FT.
GROUND 7-8 FT.
BLUE 8-10 FT.
WHITE 10-15 FT.
THRESHER 10-12 FT.

Figure 11-40. Sharks.

occur, the risk can be reduced by knowing what to do and how to do it. Sharks live in almost all oceans, seas, and in river mouths. Normally, there isn't a shark problem in areas of colder water due to the temperature of the water decreasing swim activities. Sharks vary greatly in size, but there is no close relationship between the size of a shark and the risk of attack.

(2) Hungry sharks sometimes follow fish up to the surface and into shallow waters along the shore. When sharks explore such waters, they are more likely to come in contact with people. Sharks seem to feed most actively during the night and particularly at dusk and dawn. After dark, they show an increased tendency to move toward the surface and into shore waters. Evidence indicates that a shark first locates food by smell or sound. Such things as garbage, body wastes, and blood probably stimulate the desire for food. A shark is also attracted by weak fluttery movements similar to those of a wounded fish. While a shark will investigate any large floating object as a possible food source, it probably will not attack a human unless it is hungry. Often the shark will swim away after investigating. At other times, it may approach and circle the object once or twice, or it may swim close and nudge the object with its snout. When swimming, a shark cannot stop suddenly or turn quickly in a tight circle. A shark rarely jumps out of the water to take food; however, it may grasp its prey near the surface. For this reason, people on rafts are relatively safe unless they dangle their hands, arms, feet or legs in the water.

(3) Individuals on or in the water must keep a sharp lookout for sharks. Clothing and shoes should be worn. If sharks have been noticed, survivors must be especially careful of the methods in which body wastes are eliminated and must avoid dumping blood and garbage. Vomiting, when it cannot be prevented should be done into a container or hand and thrown as far away as possible.

(a) If a group in the water is threatened or attacked by a shark, they should bunch together, form a tight circle, and face outward so an approaching shark can be seen. Ward off attack by kicking or stiff-arming the shark. Striking with the bare hand should be used only as a last resort; instead, survivors should use a hard and heavy object.

(b) Individuals should stay as quiet as possible and float to save energy. If it is necessary to swim, they should use strong, regular strokes, not frantic irregular movements.

(c) When alone, swimmers should stay away from schools of fish. If a single shark threatens at close range, the swimmer should use strong, regular swimming movements. Feinting toward the shark may scare it away.

(d) The survivor should not swim directly away from the shark, but face the shark and swim to one side, with strong rhythmic movements.

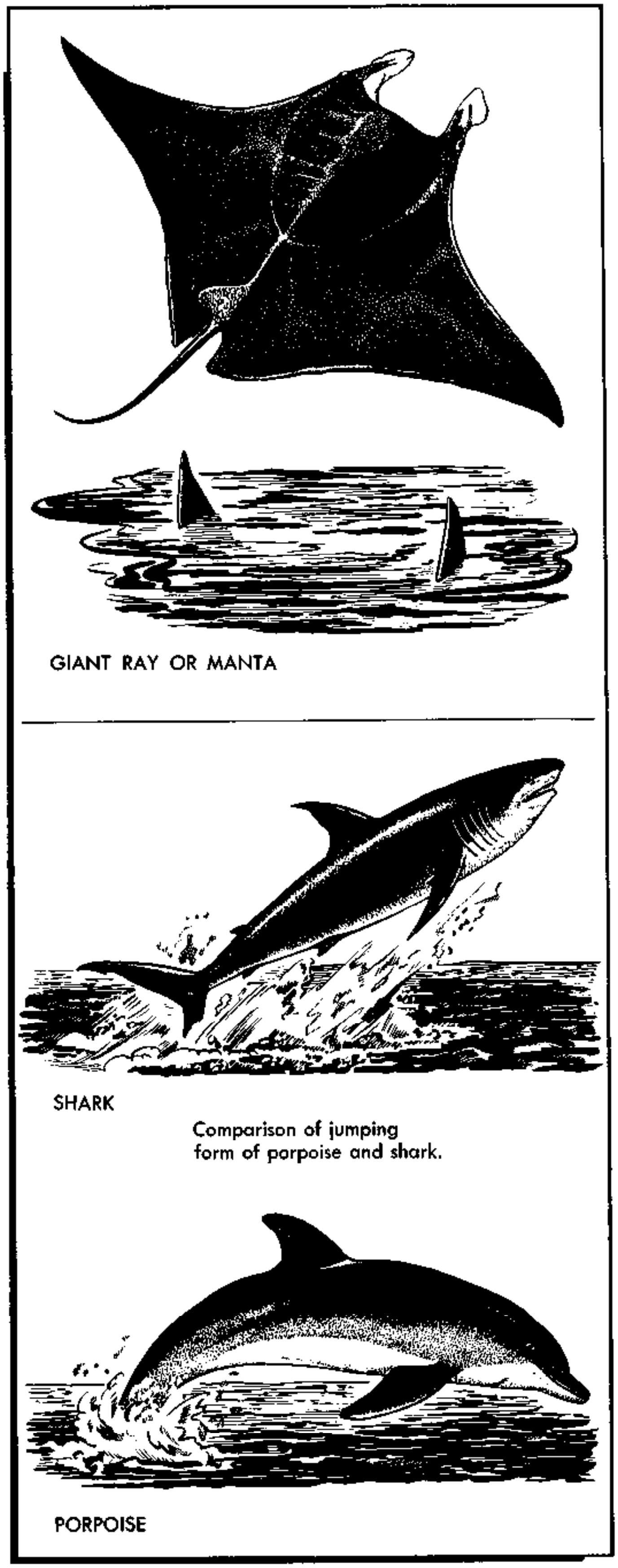
GIANT RAY OR MANTA
SHARK
Comparison of jumping form of porpoise and shark.
PORPOISE

Figure 11-41. Animals Sometimes Mistaken for Sharks.

(e) If a shark threatens to attack or damage a raft, jabbing the snout or gills with an oar may discourage it. Check for sharks around and under the raft before going into the water.

(4) Other animals are sometimes mistaken for sharks.

(a) A school of porpoises or dolphins gracefully breaking the surface, blowing and grunting, may look alarming. Actually, it should be a reassuring sight, because porpoises and dolphins are enemies of sharks. Porpoises and dolphins are harmless to humans (figure 11-41).

(b) Giant rays or mantas, which also appear in tropical waters, may be mistaken for sharks. A swimming ray curls up the tips of its fins, and when seen from water level, the fins somewhat resemble the fins on the backs of two sharks swimming side by side. In deep water, all rays are harmless to swimmers; however, some are dangerous if stepped on in shallow waters (figure 11-41).

b. Grouper or Sea Bass. These fish do not constitute the same degree of hazard as sharks; however, these carnivorous fish are curious, bold, and have a never-ending appetite. Sea bass are most commonly found around rocks, caverns, old wrecks, and caves. Stay away from these areas.

c. Killer Whales. The killer whale has the reputation of being a fearless, ruthless, and ferocious creature. These fast swimmers are found in all oceans and seas, from the tropics to both polar regions. If encountered, a survivor can be assured there are others nearby since they hunt in packs of up to 40 creatures. They have been known to attack anything that swims or floats. If an initial attack is survived, get out of the water. The raft may afford some protection, but they have been known to come up under iceflows and knock other animals into the water. Stay out of the water. On thin ice, do not stand near seals, etc., as the whale may mistake the human form for a seal. However, the probability of being attacked by a killer whale is slim. If an aircrew member is attacked, it will probably be due to the fact that this intelligent whale simply mistook the person for its regular diet.

d. Barracuda. There are 20-odd species of barracuda; some are more feared in certain parts of the world than are sharks. If survivors come down in any tropical or subtropical sea, they may encounter this fish. Barracuda are attracted by anything which enters the water and they seem to be particularly curious about bright objects. Accordingly, survivors should avoid dangling dogtags or other shiny pieces of equipment in the water. Dark colored clothing is also best to wear in the water if no raft is available.

e. Moray Eels. If attacked by some species of moray eel, the survivor may have to cut off its head since some eels will retain their sharp crushing grip until dead. The knife used to do this should be very sharp since their skin is tough and difficult to cut. Their bodies are very slippery and hard to hold. A survivor is most likely to come into contact with a moray eel when poking into holes and crevices around or under coral reefs. Use caution in these areas.

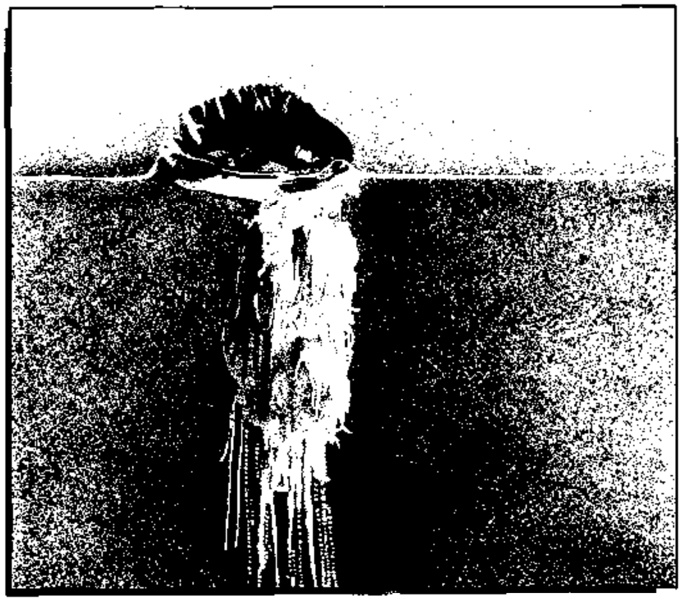

Figure 11-42. Portuguese Man-of-War.

f. Poisonous and Venomous Marine Animals (Invertebrates). There are many marine animals that have no backbone and can inflict injuries by stinging. Three major categories of invertebrates are important to the survivor.

(1) Coelenterates. This group includes jellyfish, hydroids, sea anemones, and corals. Coelenterates are all simple, many-celled organisms. They all possess tentacles equipped with stinging cells or nematocysts in addition to other technical characteristics. The family of coelenterates is divided into three major classes.

(a) Hydrozoan Class. Two of the more common members of this class are:

-1. Stinging or Fire Coral. This false coral can be found in areas of true coral reefs in warm waters.

-2. Portuguese Man-of-War or Blue Bottle. This hydroid is frequently mistaken for a true jellyfish. It is almost always found floating at the surface of the water (figure 11-42). Its stinging tentacles may extend several yards below the surface. Their float is 5 to 10 inches in length. Each tentacle may contain thousands of stinging cells. When one considers the large number of such tentacles, it is apparent that the fishing filaments of the Blue Bottle are quite a formidable venom apparatus.

(b) Anthozoa Class:

-1. Corals. Elkhorn coral and stony coral are very adaptable and have a real immunity to predators.

This helps explain why they tend to dominate reef communities. Corals are carnivores and with the use of small tentacles, capture and consume living zooplankton. Survivors should treat coral cuts by thoroughly cleaning the wound and removing any coral particles. Some coral cuts have been helped by painting them with an antiseptic solution of tincture of iodine.

-2. Sea Anemones. The sea anemone is one of the most plentiful marine creatures, with well over 1,000 species. They can be found from tide level to depths of more than 7,900 fathoms in all seas. Their size ranges from very small, (less than an inch) to over 2 feet in diameter. They eat fish, mollusks, crustaceans, and other invertebrates. Most of the stinging cells of the sea anemones are located on the outer ring of the tentacles.

(c) Scyphozoa Class—Jellyfish. There are many and varied species of jellyfish distributed throughout all seas. Their size ranges from extremely small to a diameter of 6 feet with tentacles hanging below to a depth of 100 feet. All are carnivorous. Some are transparent and glassy while others are brilliantly colored. Regardless of their size and color, they are very fragile creatures which, for the most part, depend on wind and tidal currents to help them move. Most adults can swim but this ability is weak. Whether they stay on the surface or under the surface, and to what depth, varies with each species. The stinging cells of jellyfish are located in the tentacles.

(d) Venom Apparatus of Coelenterates:

-1. All of the coelenterates have stinging cells or nemotocysts located on the tentacles. Each of these cells is like a capsule. If the survivor comes into contact with the capsule, part of it springs open and a very sharp, extremely small "thread"-type tube appears. The sharp tip of the tube penetrates the skin and the venom is injected. When coming in contact with the tentacles of any coelenterate, the survivor brushes up against literally thousands of these small stinging organs.

-2. The symptoms produced by coelenterate stings will vary according to species, where the sting is located, and the physical condition of the survivor. In general, though, the sting caused by hydroids and hydroid corals is primarily skin irritations of a local nature. Stings of the Portuguese Man-of-War may be very painful. True corals and sea anemones produce a similar reaction. Some of the sting of these organisms may be hardly noticeable, while others may cause death in 3-8 minutes. Symptoms common to all of these may vary from an immediate mild prickly or stinging sensation, like that of touching a nettle, to a burning, throbbing, shooting-type pain which may cause the survivor to become unconscious. In some cases, the pain may be localized, while in others, it may spread to the groin, armpits, or abdomen. The area in which contact was made will usually become red, followed by severe inflammation, rash, swelling, blistering, skin hemorrhages, and sometimes ulceration. In severe cases of reaction, in addition to shock, the person may experience one or more of the following: muscular cramps, lack of touch and temperature sensations, nausea, vomiting, backache, loss of speech, constriction of the throat, frothing at the mouth, delirium, paralysis, convulsions, and death. Since some of these traits appear quickly, the victim should try to get out of the water if at all possible to avoid drowning.

-3. One of the most deadly jellyfish is the sea wasp (uncommon creature which is found in tropical southern Pacific waters). This animal can cause death anywhere from 30 seconds to 3 hours after contact. Most deaths take place within 15 minutes. The pain is said to be excruciating. The sea wasp can be recognized by the long tentacles that hang down from the four corners of its squarish body.

-a. Relieve pain. Tentacles or other matter on the skin should be removed immediately. This is important because as long as this matter is on the skin, additional stinging cells may be discharged. Use clothing, seaweed, or any other available material to remove the matter. Morphine is effective in relieving pain. *DO NOT* rub the wound with anything, especially sand, as this may cause the stinging cells to be activated. *DO NOT* suck the wound.

-b. Alleviate poison effects. Suntan lotion, oil, and alcohol should be applied to the area to stop further stinging. The following local remedies have been used in various parts of the world with varying degrees of success: papain (protein destroying enzyme), sodium bicarbonate, olive oil, sugar, soap, vinegar, lemon juice, diluted ammonia solution, papaya latex, plant juices, boric acid solution, flour, baking powder, etc. (Urine—with its ammonia content—may be the only source of relief available to a survivor).

-c. Artificial respiration and cardiopulmonary resuscitation may be required. There are no known specific antidotes for most coelenterate stings; however, there is one antivenin for the sea wasp which is papain, a proteolytic enzyme in the juice of the green fruit of the papaya. Even if the survivor is in an area where the antivenin is available, it may be too late to obtain and use it. The venom acts so quickly that medical help is often too late.

-4. Jellyfish should be given a wide berth since in some species the tentacles may trail 50 feet or more from the body. After a storm in tropical areas where large numbers of jellyfish are present, the survivor may be injured by pieces of floating tentacles that have been removed from the animals during the storm. Jellyfish washed up on the beach may appear dead, but can still, in some cases, inflict painful injuries. The best prevention is to stay out of the water by getting into a raft or onto shore. If in a raft, do not let arms and legs trail over the side. The clothing (antiexposure suit) that the survivor wears should cover as much of the body as possible. Flight clothing items currently available should provide adequate protection.

(2) Mollusks. Octopus, squid, and univalve shellfish are in this category. Mollusks make up the largest single group of biotoxic marine invertebrates of direct importance to the survivor. The phylum of mollusks is generally divided into five classes. Stinging or venomous mollusks which concern the survivor fall mainly into two categories:

(a) Gastropoda (Stomache Footers):

-1. Mollusks. These in general are unsegmented invertebrates. Sometimes their soft bodies will secrete a calcareous shell. They have a muscular foot which serves a variety of functions. Some breathe by means of a type of siphon while others use gills. Some types have jaws. In those that don't have jaws, food is obtained by a rasp-like device called a radula. In the cone shells, the radula is a barb or tooth more like a hollow, needle-like structure.

-2. Gastropods. These univalves include marine snails, slugs, as well as land and freshwater snails. It is estimated that there are over 33,000 living species of gastropods; however, only members of the genus conus are of concern to the survivor. Of these cone shells, there are over 400 species, but they will only be discussed in general terms with the emphasis placed on the more dangerous species. With few exceptions, these attractive shellfish are located in tropical or subtropical areas. All of these shells have a very highly developed venom apparatus designed for vertebrate or invertebrate creatures and are found from shallow tidal areas to depths of many hundreds of feet. The area in which the survivor may come into contact with these shellfish is in coral reefs and sandy or rubble habitat. All cone-shaped shells in these areas should be avoided. Cone shells are usually nocturnal. During the daytime, they burrow and hide in the sand, rocks, or coral; they feed at night on worms, octopus, other gastropods, and small fish. Several of these shells have caused death in humans. The venom apparatus lies within a body cavity of the animal and the animal is capable of thrusting and injecting the poison via the barb into the flesh of the victim. The cone shell is able to inflict its wound only when the head of the animal is out of the shell.

-a. Complications. The sting made by a cone shell is a puncture-type wound. The area around the wound may exhibit one or more of the following: turn blue, swelling, numbness, stinging, or burning sensation. The amount of pain will vary from person to person. Some say the pain is like a bee sting, while others find it unbearable. The numbness and tingling sensations around the site of the wound may spread rapidly, involving the whole body, especially around the lips and mouth. Complete general muscle paralysis may occur. Coma may ensue and death is usually the result of cardiac failure.

-b. Treatment. The pain comes from the injection of venom, slime, and other irritating foreign matter into the wound site. The treatment is primarily symptomatic because there is no specific treatment. Applying hot towels or soaking the affected area in hot water may relieve some of the pain. Artificial respiration may be needed.

(b) Cephalpods. This group includes the nautilis, squid, cuttlefish, and octopus. Since the octopus is the marine animal most likely to be encountered by a survivor, it is the only one that will be discussed. The head of this animal is large and contains well-developed eyes. The mouth is surrounded by eight legs equipped with many suckers. It can move rapidly by expelling water from its body cavity, though it usually glides or creeps over the bottom. Most octopuses live in water ranging from very shallow to depths of over 100 fathoms. All are carnivorous and feed on crabs, and other mollusks. Octopuses like to hide in holes or underwater caves—avoid these areas.

-1. Complications. The sharp parrot-like beak of the octopus makes two small puncture wounds into which a toxic solution or venom is injected. Pain is usually felt immediately in the form of a burning, itching, or stinging sensation. Bleeding from the wound is usually very profuse which may indicate the venom contains an anticoagulant. The area around the wound, and in some cases the entire appendage, may swell, turn red, and feel hot. There has been one report of a fatal octopus bite. This death was attributed to the blue ringed octopus (Octopus Maculosus) (figure 11-43). This small octopus is usually only 3 or 4 inches across although some may be slightly larger. Found throughout the Indo-Pacific area, this octopus is not aggressive toward humans. Because its bite is so dangerous, it should not be handled at any time. When this animal is disturbed the intensity of its blue rings varies rapidly on a light yellow or cream to brown background.

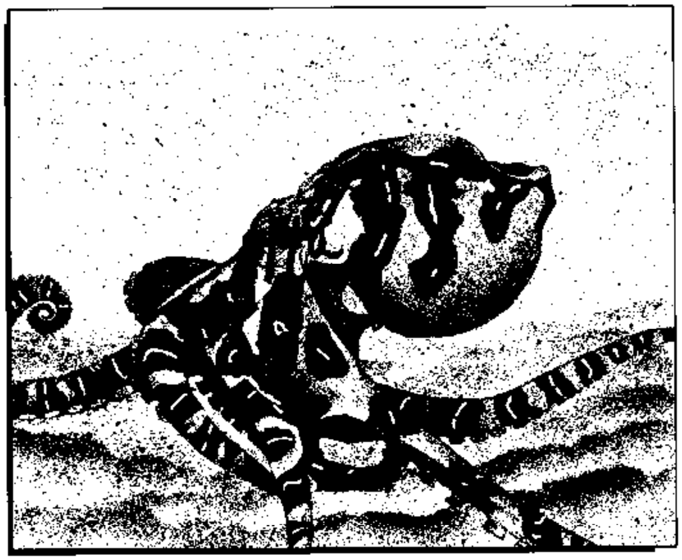

Figure 11-43. Blue Ringed Octopus.

-2. Treatment. Treat for shock, stop bleeding, clean the wound area since more venomous saliva could be in the area, and treat symptoms as they arise. There is no known cure for the venom of the Blue Ringed Octopus.

(3) Echinoderms. Sea cucumbers, starfish, and sea urchins are members of this group. Sea urchins comprise the most dangerous type of echinoderms. Sea urchins have rounded, egg-shaped, or flattened bodies. They have hard shells that carry spines. In some species, the spines are venomous and present a hazard if stepped on or handled. Some urchins are nocturnal. They all tend to be omnivorous, eating algae, mollusks, and other small organisms. They can be found in tidal pools or in areas of great depth in many parts of the world. Sea urchins are not good food sources. At certain times of the year, certain species can be poisonous.

(a) Complications. The needle-sharp points of sea urchin spines are able to penetrate the flesh easily. These spines are also very brittle and tend to break off while still attached to the wound and are very difficult to withdraw. Stepping on one of these spines produces an immediate and very intense burning sensation. The area of pain will also swell, turn red, and ache. Numbness and muscular paralysis, swelling of the face, and a change in the pulse have also been reported. Secondary infection usually sets in. While some deaths have been reported, other victims have experienced loss of speech, respiratory distress, and paralysis. The paralysis will last from 15 minutes to 6 hours.

(b) Treatment. Spines (pedicellaria) that are detached from the animal will continue to secrete venom into the wound. The spines of some species will be easily dislodged whereas others must be surgically removed. There will also be some discoloration due to a dye the animal secretes—do not be disturbed by this. Some experts say to apply grease to allow the spines to be scraped off. Others advise leaving them alone since some of the spines will dissolve in the wound within 24 to 48 hours. Still other experts say to apply citrus juice, if available, or soak the area in vinegar several times a day to dissolve them.

(c) Prevention. No sea urchin should be handled. The spines can penetrate leather and canvas with ease.

g. Venomous and Poisonous Marine Animals (Vertebrates). These fish can be divided into two general groups—fish that sting and fish that are poisonous to eat (figure 11-44).

(1) Venomous Spine Fish (Fish That Sting):

(a) Types of fish in this group are:

-1. Spiny dog fish.
-2. Stingrays. Includes whiprays, batrays, butterfly rays, cow-nosed rays, and round stingrays.
-3. Rat fish.
-4. Weever fish.
-5. Catfish.
-6. Toad fish.
-7. Scorpion fish.
-8. Surgeon fish.
-9. Rabbit fish.
-10. Star gazers.

NOTE: For all wounds from these types of fish, aid should be directed to three areas: alleviating the pain of the sting, trying to halt the effects of the venom, and preventing infection.

(b) Certain types of these fish have up to 18 spines. The pain caused by the sting of one of these spines is so great in some species that the victim may scream and thrash about wildly. In one case, a man stung in the face by a weever fish begged for bystanders to shoot him, even after two shots of morphine sulfate. Many of these fish are bottom dwellers who will not move out of the way when being approached by humans. Instead, they will lie quietly camouflaged, put up their spines, and simply wait for the unlucky individual to step on them. Other people have been injured by them while trying to remove them from fishing nets and fishing lines. In cases where humans are stung by stingrays, the barbs on the sharp spines may cause severe lacerations as well as introduce poison. These wounds should be irrigated without delay. Puncture wounds from the fish are small and make removal of the poison a difficult process. It may be necessary to remove the barb. A procedure which is fairly successful is to make a small cut across the wound (debride) and then apply suction. Even if no incision is made, suction should be tried since it is important to remove as much of the venom as possible. The more poison removed, the better. Morphine does not relieve the pain of some of these venoms. Most doctors agree that the injured part should be soaked in hot water from 30 minutes to 1 hour. The

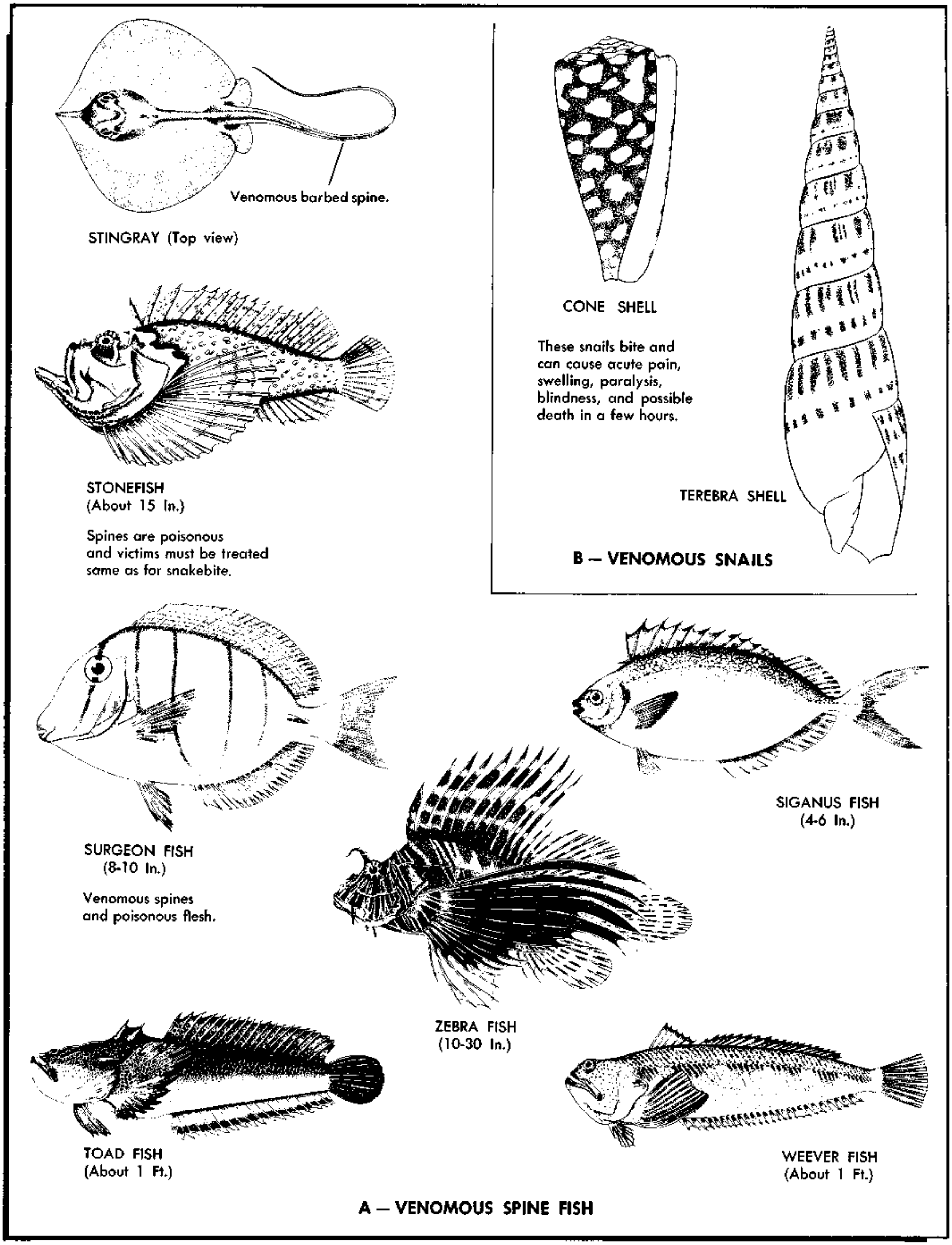
Venomous barbed spine.
STINGRAY (Top view)
STONEFISH
(About 15 In.)
CONE SHELL
TEREBRA SHELL
SURGEON FISH
(8-10 In.)
SIGANUS FISH
(4-6 In.)
ZEBRA FISH
(10-30 In.)
TOAD FISH
(About 1 Ft.)
WEEVER FISH
(About 1 Ft.)

Spines are poisonous and victims must be treated same as for snakebite.

These snails bite and can cause acute pain, swelling, paralysis, blindness, and possible death in a few hours.

Venomous spines and poisonous flesh.

B — VENOMOUS SNAILS

A — VENOMOUS SPINE FISH

Figure 11-44. Venomous Spine Fish and Snails.

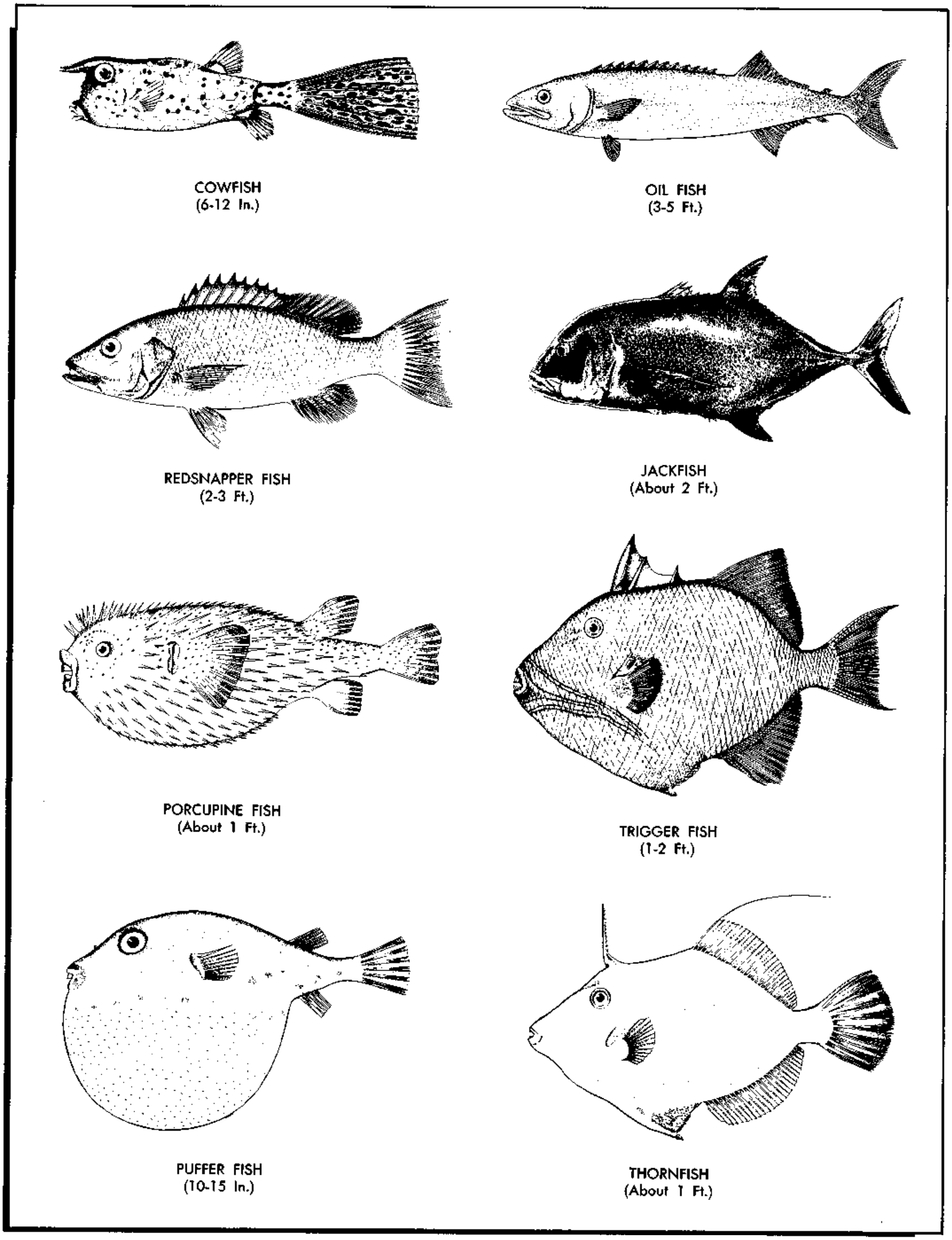

COWFISH
(6-12 In.)

OIL FISH
(3-5 Ft.)

REDSNAPPER FISH
(2-3 Ft.)

JACKFISH
(About 2 Ft.)

PORCUPINE FISH
(About 1 Ft.)

TRIGGER FISH
(1-2 Ft.)

PUFFER FISH
(10-15 In.)

THORNFISH
(About 1 Ft.)

Figure 11-45. Fish With Poisonous Flesh.

temperature of the water should be as hot as the patient can stand without injury. If the wound is on the face or body, hot moist cloth compresses can be used. The use of heat in this manner may weaken the effect of the poison in some cases. After soaking the wound, clean it again, if necessary. Cover the area of the wound with antiseptic and a clean sterile dressing. If antibiotics are available, it may be advisable to use them to help prevent infection. Treatment for shock is wise. Artificial respiration may be needed since some venoms may cause cardiac failure, convulsions, or respiratory distress.

(c) For fish that are poisonous to eat, see figure 11-45.

-1. There is no known way to detect a poisonous fish merely by its appearance. Fish that are poisonous in one area may be safe to eat in another. In general, bottom dwellers and feeders, especially those associated with coral reefs, should be suspect. Also, unusually large predator-type fish should be eaten with caution. The internal organs and roe of all tropical marine fish should never be eaten, as those parts contain a higher concentration of poison.

-2. Under certain conditions, where the survivor may be required to eat questionable fish, rules should be followed. A fish will be safer if it can be caught away from reefs or entrances to lagoons. Once the fish has been secured, the "marine animal edibility test" should be used. The fish should be cut into thin strips and boiled in successive changes of water for an hour or more. This may help since some, but not all, of the toxins are water soluble. Further, it should be noted, that normal cooking techniques and temperatures will not weaken or destroy poisons.

-3. If boiling is not possible, cut the meat into thin strips and soak in changes of sea water for an hour or so, squeezing the meat juices out as thoroughly as possible. A survivor should eat only a small portion of the flesh and wait 12 hours to see if any symptoms arise (if the fish will not spoil). Remember that the degree of poisoning is directly related to how much fish is eaten. If in doubt, do not eat it. The advice of native people on eating tropical marine fish may not be valid. In many instances they check edibility by first feeding fish portions to their dogs and cats.

-4. Treatment. As soon as any symptoms arise, vomiting should be induced by administering warm saltwater or the whites of eggs. If these procedures don't work, try sticking a finger down the person's throat. A laxative should also be given to the victim if one is available. The victim may have to be protected from injury during convulsions. If the victim starts to foam at the mouth and exhibits signs of respiratory distress, a cricothyroidotemy may have to be performed. Morphine may help relieve pain in some cases. If the victim complains of severe itching, cool showers may give some relief. Treat any other symptoms as they arise.

(2) Poisonous Marine Turtles:

(a) Species. There are over 265 species of marine turtles. Of these, only five have been reported as poisonous and dangerous to the survivor. Many of these species are commonly eaten, but for some unknown reason, these same turtles become extremely toxic under certain conditions. Basically, the main species to be concerned with are the green, the hawksbill, and the leatherback turtles. These turtles are found mainly in tropical and subtropical seas but can also be found in temperate waters.

(b) Origin. The origin of turtle poison is unknown but some investigators suggest it comes from the poisonous marine algae eaten by the turtles. It should be noted that a species of turtle may be safe to eat in one area but deadly in another. There is absolutely no way a survivor can distinguish between a poisonous and nonpoisonous sea turtle just by looking at it or by examining any part of it. Toxicity may occur at any time of the year; however, the most dangerous months appear to be the warmer months. The degree of freshness also has nothing to do with how poisonous the turtle is.

(c) Complications. The symptoms will vary with the amount of turtle ingested. Symptoms will develop within a few hours to a few days after eating the food. These symptoms include nausea, vomiting, diarrhea, pain, sweating, coldness in the extremities, vertigo, dry and burning lips and tongue, tightness of the chest, drooling, and difficulty in swallowing. Other victims reported a heavy feeling of the head, a white coating on the tongue, diminished reflexes, coma, and sleepiness. About 44 percent of the victims poisoned by marine turtles die.

(d) Treatment. There is no known antidote for this kind of poisoning. There is no specific treatment—treat symptomatically.

(e) Prevention. If there is the slightest suspicion about the edibility of a marine turtle, it should not be eaten, or at least the marine animal edibility test should be used. Turtle liver is especially dangerous to eat because of its high vitamin A content.

h. Birds. There are roughly 260 species of sea birds. Most of the birds travel only a few miles out to sea but the albatross ranges across the seas far from any landmasses.

i. Red Tide. Red tide is a name used to describe the reddish or brownish coloration in saltwater, resulting from tiny plants and organisms called plankton, which suddenly increase tremendously in numbers. Red tides appear in waters worldwide. In the United States, they are most common off the coasts of Florida, Texas, and southern California. Although most red tides are harmless, some may kill fish and other water creatures. Still other types of red tides do not kill sea life, but cause the shellfish feeding on them to be poisonous. Some of these creatures secrete poisons which can paralyze and kill fish, or can kill fish by using nearly all of the oxygen in the water. Although the exact reason for the sudden increase of the plankton is unknown, there is evidence that shows favorable food, temperature, sunlight, water currents, and salt in the water will increase the population. It is not unusual for it to remain from a few hours to several months. A survivor should not eat any fish that are found dead.

Chapter 12

LOCAL PEOPLE

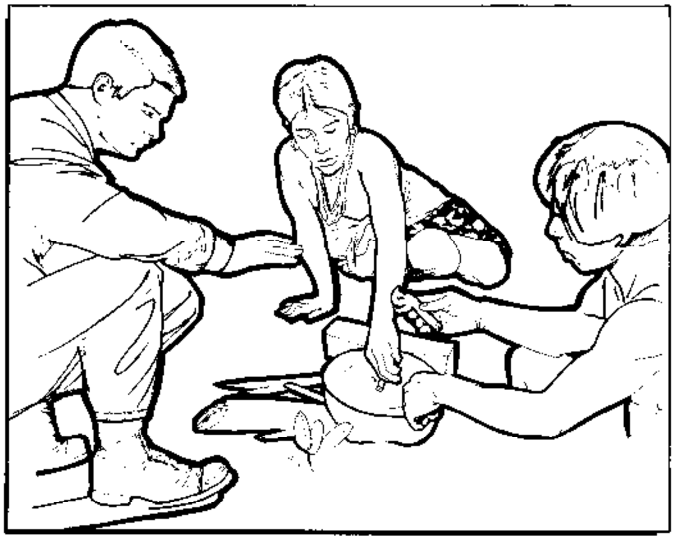

Figure 12-1. Local People.

12-1. Introduction. One evader concluded with the following advice: "My advice is, 'When in Rome, do as the Romans do!' Show interest in their country, and they will go overboard to help you!" One of the most frequently given bits of advice is to accept, respect, and adapt to the ways of the people among whom survivors find themselves. This is good advice, but there are a number of important problems involved in putting this advice into practice (figure 12-1).

12-2. Contact With People. The survivor must give serious consideration to people. Are they people with a primitive culture? Are they farmers, fishermen, friendly people, or enemies? To the survivor, "cross-cultural contact" can vary quite radically in scope. It could mean interpersonal relationships with people of an extremely different (primative) culture, or contacts with people who are culturally modern by our standards. A culture is identified by standards of behavior that are considered proper and acceptable for the members and may or may not conform to our idea of propriety. Regardless of who these people are, the survivor can expect they will have different laws, social and economic values, and political and religious beliefs.

a. People will be friendly, unfriendly, or choose to ignore the survivor. Their attitude may be unknown. If the people are known to be friendly, the survivor must make every attempt to keep them that way by being courteous and respecting the religion, politics, social customs, habits, and all other aspects of their culture. If the people are known to be enemies or are unknowns, the survivor should make every effort to avoid any contact and leave no sign of presence. Therefore, a basic knowledge of the daily habits of the local people can be extremely important in this attempt. An exception might be, if after careful and covert observation, it is determined an unknown people are friendly, contact might be made if assistance is absolutely necessary.

b. Generally, there is little to fear and everything to gain from thoughtful contact with the local peoples of friendly or neutral countries. Familiarity with local customs, displaying common decency, and most importantly, showing respect for their customs should help a survivor avoid trouble and possibly gain needed assistance. To make contact, a survivor should wait until only one person is near and, if possible, let that person make the initial approach. Most people will be willing to help a survivor who appears to be in need; however, political attitudes and training or propaganda efforts can change the attitudes of otherwise friendly people. Conversely, in nominally unfriendly countries, many people, particularly in remote areas, may feel abused or ignored by their politicians, and may be more friendly toward outsiders.

c. The key to successful contact with local peoples is to be friendly, courteous, and patient. Displaying fear, displaying weapons, and making sudden or threatening movements can cause a local person to fear a survivor which can, in turn, prompt a hostile response. When attempting contact, smile frequently. Many local peoples may be shy and seem unapproachable or they may ignore the survivor. Approach them slowly and don't rush matters.

12-3. Survivor's Behavior:

a. Salt, tobacco, silver money, and similar items should be used discreetly in trade. Paper money is well known worldwide. Don't overpay; it may lead later to embarrassment and even danger. Treat people with respect and do not laugh at or bully them.

b. Sign language or acting out needs or questions can be very effective. Many people are accustomed to it and communicate using nonverbal sign language. Aircrew members should learn a few words and phrases of the local language in and around their area of operations. Attempting to speak someone's language is an excellent way to show respect for their culture. Since English is widely used, some of the local people may understand a few words of English.

c. Certain areas may be taboo. They range from religious or sacred places to diseased or danger areas. In some areas, certain animals must not be killed. A survivor must learn what the rules are and follow them. The survivor must be observant and learn as much as possible. This will not only help in strengthening relations, but new knowledge and skills may be very important later. The downed aircrew member should seek advice on local hazards and find out from friendly people where there are hostile people. Keep in mind though, that frequently, people, as in our culture, insist others are hostile because they also do not understand different cultures and distant peoples. The people that generally can be trusted, in their opinion, are their immediate neighbors—much the same as in our own neighborhood. Local people, like ourselves, suffer from diseases which are contagious. The survivor should build a separate dwelling, if possible, and avoid physical contact without seeming to do so. Personal preparation of food and drink is desirable if it can be done without giving offense. Frequently, the use of "personal or religious custom" as an explanation for isolationist behavior will be accepted by the local people.

d. Trading or barter is common in more primitive societies. Hard coin is usually good, whether for its exchange value or as jewelry or trinkets. In isolated places, matches, tobacco, salt, razor blades, empty containers, or cloth may be worth more than any form of money.

e. The survivor must be very cautious when touching people. Many people consider "touching" taboo and such actions may be dangerous. Sexual contact should be avoided.

f. Hospitality among some people is such a strong cultural trait they may seriously reduce their own supplies to make certain a stranger or visitor is fed. What is offered should be accepted and shared equally with all present. The survivor should eat in the same way they eat and, most importantly, attempt to eat all that is offered. If any promises are made, they must be kept. Personal property and local customs and manners, even if they seem odd, must be respected. Some kind of payment for food, supplies, etc., should be made.

g. Privacy must be respected and a survivor should not enter a house unless invited.

12-4. Changing Political Allegiance. In today's world of fast-paced international politics and "shuttle diplomacy," political attitudes and commitments within nations are subject to rapid change. The population of many countries, especially politically hostile countries, must not be considered friendly just because they do not demonstrate open hostility. Unless briefed to the contrary, avoid all contact with such people.

Part Five

PERSONAL PROTECTION

Chapter 13

PROPER BODY TEMPERATURE

13-1. Introduction. In a survival situation the two key requirements for personal protection are maintenance of proper body temperature and prevention of injury. The means for providing personal protection are many and varied. They include the following general categories: clothing, shelter, equipment, and fire. These individual items are not necessary for survival in every situation; however, all four will be essential in some

WIND SPEED		COOLING POWER OF WIND EXPRESSED AS "EQUIVALENT CHILL TEMPERATURE"																				
KNOTS	MPH	TEMPERATURE (°F)																				
Calm	Calm	40	35	30	25	20	15	10	5	0	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50	-55	-60
		EQUIVALENT CHILL TEMPERATURE																				
3 - 6	5	35	30	25	20	15	10	5	0	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50	-55	-65	-70
7 - 10	10	30	20	15	10	5	0	-10	-15	-20	-25	-35	-40	-45	-50	-60	-65	-70	-75	-80	-90	-95
11 - 15	15	25	15	10	0	-5	-10	-20	-25	-30	-40	-45	-50	-60	-65	-70	-80	-85	-90	-100	-105	-110
16 - 19	20	20	10	5	0	-10	-15	-25	-30	-35	-45	-50	-60	-65	-75	-80	-85	-95	-100	-110	-115	-120
20 - 23	25	15	10	0	-5	-15	-20	-30	-35	-45	-50	-60	-65	-75	-80	-90	-95	-105	-110	-120	-125	-135
24 - 28	30	10	5	0	-10	-20	-25	-30	-40	-50	-55	-65	-70	-80	-85	-95	-100	-110	-115	-125	-130	-140
29 - 32	35	10	5	-5	-10	-20	-30	-35	-40	-50	-60	-65	-75	-80	-90	-100	-105	-115	-120	-130	-135	-145
33 - 36	40	10	0	-5	-15	-20	-30	-35	-45	-55	-60	-70	-75	-85	-95	-100	-110	-115	-125	-130	-140	-150
WINDS ABOVE 40 HAVE LITTLE ADDITIONAL EFFECT.		LITTLE DANGER					INCREASING DANGER (Flesh may freeze within 1 min)							GREAT DANGER (Flesh may freeze within 30 seconds)								
		DANGER OF FREEZING EXPOSED FLESH FOR PROPERLY CLOTHED PERSONS																				

INSTRUCTIONS

MEASURE LOCAL TEMPERATURE AND WIND SPEED IF POSSIBLE; IF NOT, ESTIMATE. ENTER TABLE AT CLOSEST 5°F INTERVAL ALONG THE TOP AND WITH APPROPRIATE WIND SPEED ALONG LEFT SIDE. INTERSECTION GIVES APPROXIMATE EQUIVALENT CHILL TEMPERATURE. THAT IS, THE TEMPERATURE THAT WOULD CAUSE THE SAME RATE OF COOLING UNDER CALM CONDITIONS.

NOTES

WIND

1. THIS TABLE WAS CONSTRUCTED USING MILES PER HOUR (MPH), HOWEVER, A SCALE GIVING THE EQUIVALENT RANGE IN KNOTS HAS BEEN INCLUDED ON THE CHART TO FACILITATE ITS USE WITH EITHER UNIT.
2. WIND MAY BE CALM BUT FREEZING DANGER GREAT IF PERSON IS EXPOSED IN MOVING VEHICLE, UNDER HELICOPTER ROTORS, IN PROPELLOR BLAST, ETC. IT IS THE RATE OF RELATIVE AIR MOVEMENTS THAT COUNTS AND THE COOLING EFFECT IS THE SAME WHETHER YOU ARE MOVING THROUGH THE AIR OR IT IS BLOWING PAST YOU
3. EFFECT OF WIND WILL BE LESS IF PERSON HAS EVEN SLIGHT PROTECTION FOR EXPOSED PARTS. LIGHT GLOVES ON HANDS , PARKA HOOD SHIELDING FACE , ETC.

ACTIVITY DANGER IS LESS IF SUBJECT IS ACTIVE. A PERSON PRODUCES ABOUT 100 WATTS (341 BTUs) OF HEAT STANDING STILL BUT UP TO 1000 WATTS (3413 BTUs) IN VIGOROUS ACTIVITY LIKE CROSS-COUNTRY SKIING.

PROPER USE OF CLOTHING AND **ADEQUATE DIET** ARE BOTH IMPORTANT.

COMMON SENSE THERE IS NO SUBSTITUTE FOR IT. THE TABLE SERVES ONLY AS A GUIDE TO THE COOLING EFFECT OF THE WIND ON BARE FLESH WHEN THE PERSON IS FIRST EXPOSED. GENERAL BODY COOLING AND MANY OTHER FACTORS AFFECT THE RISK OF FREEZING INJURY.

Figure 13-1. Windchill Chart.

environments. In this part of the regulation, the conditions which affect the body temperature, the physical principles of heat transfer, and the methods of coping with these conditions will be covered.

13-2. Body Temperature. The body functions best when core temperatures range from 96°F to 102°F. Preventing too much heat loss or gain should be a primary concern for survivors. Factors causing changes in body core temperature (excluding illness) are the climatic conditions of temperature, wind, and moisture.

a. Temperature. As a general rule, exposure to extreme temperatures can result in substantial decreases in physical efficiency. In the worst case, incapacitation and death can result.

b. Wind. Wind increases the chill effect (figure 13-1), causes dissipation of heat, and accelerates loss of body moisture.

c. Moisture — Precipitation, Ground Moisture, or Immersion. Water provides an extremely effective way to transfer heat to and from the body. When a person is hot, the whole body may be immersed in a stream or other body of water to be cooled. On the other hand, in the winter, a hot bath can be used to warm the body. When water is around the body, it tends to bring the "body" to the temperature of the liquid. An example is when a hand is burned and then placed in cold water to dissipate the heat. One way to lower body temperature is by applying water to clothing and exposing the clothed body to the wind. This action causes the heat to leave the body 25 times faster than when wearing dry clothing. This rapid heat transfer is the reason survivors must always guard against getting wet in cold environments. Consider the result of a body totally submerged in water at a temperature of 50°F and determine how long a person could survive (figures 13-2 and 13-3).

13-3. Heat Transfer. There are five ways body heat can be transferred. They are radiation, conduction, convection, evaporation, and respiration.

a. Radiation. Radiation is the primary cause of heat loss. It is defined as the transfer of heat waves from the body to the environment and (or) from the environment back to the body. For example, at a temperature of 50°F, 50 percent of the body's total heat loss can occur through an exposed head and neck. As the temperature drops, the situation gets worse. At 5°F, the loss can be 75 percent under the same circumstances. Not only can heat be lost from the head, but also from the other extremities of the body. The hands and feet radiate heat at a phenomenal rate due to the large number of capillaries present at the surface of the skin. These three areas

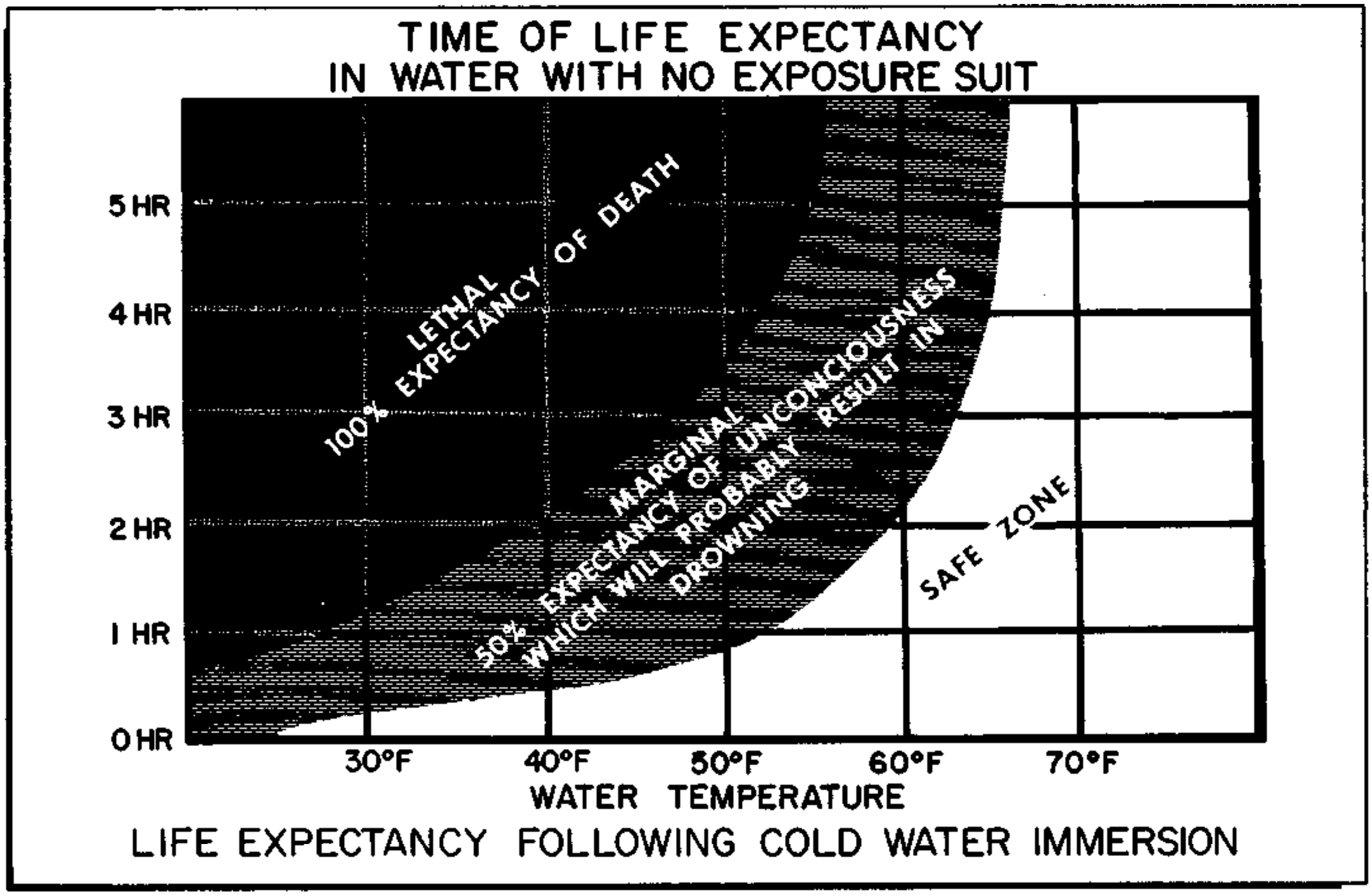

Figure 13-2. Life Expectancy Following Cold-Water Immersion.

of the body must be given particular attention during all periods of exposure to temperature extremes.

b. Conduction:

(1) Conduction is defined as the movement of heat from one molecule to another molecule within a solid object. Extreme examples of how heat is lost and gained quickly are deep frostbite and third-degree burns, both gained from touching the same piece of metal at opposite extremes of cold and heat. Heat is also lost from the body in this manner by touching objects in the cold with bare hands, by sitting on a cold log, or by kneeling on snow to build a shelter. These are practices which survivors should avoid since they can lead to overchilling the body.

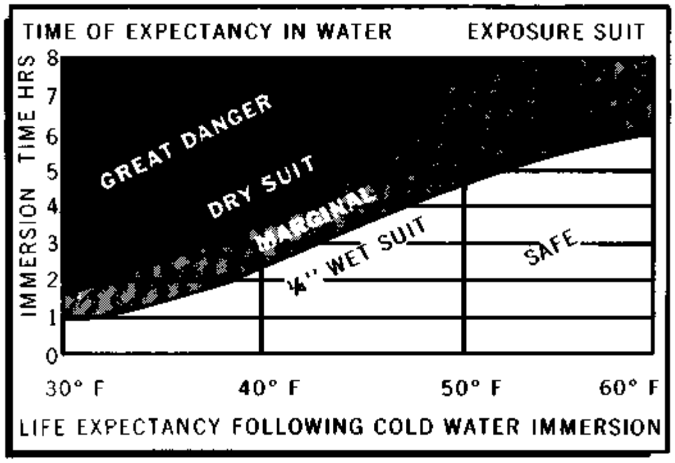

Figure 13-3. Life Expectancy Following Cold-Water Immersion (Exposure Suit.)

(2) Especially dangerous is the handling of liquid fuel at low temperatures. Unlike water which freezes at 32°F, fuel exposed to the outside temperatures will reach the same temperature as the air. The temperature of the fuel may be 10°F to 30°F below zero or colder. Spilling the fluid on exposed skin will cause instant frostbite, not only from the conduction of heat by the cold fluid, but by the further cooling effects of rapid evaporation of the liquid as it hits the skin.

c. Convection. Heat movement by means of air or wind to or from an object or body is known as convection. The human body is always warming a thin layer of air next to the skin by radiation and conduction. The temperature of this layer of air is nearly equal to that of the skin. The body stays warm when this layer of warm air remains close to the body. However, when this warm layer of air is removed by convection, the body cools down. A major function of clothing is to keep the warm layer of air close to the body; however, by removing or disturbing this warm air layer, wind can reduce body temperature. Therefore, wind can provide beneficial cooling in dry, hot conditions, or be a hazard in cold, wet conditions.

d. Evaporation. Evaporation is a process by which liquid changes into vapor, and during this process, heat within the liquid escapes to the environment. An example of this process is how a "desert water bag" works on the front of a jeep while driving in the hot desert. The wind created by the jeep helps to accelerate evaporation and causes the water in the bag to be cooled. The body also uses this method to regulate core temperature when it perspires and air circulates around the body. The evaporation method works any time the body perspires regardless of the climate. For this reason, it is essential that people wear fabrics that "breathe" in cold climates. If water vapor cannot evaporate through the clothing, it will condense, freeze, and reduce the insulation value of the clothing and cause the body temperature to go down.

e. Respiration. The respiration of air in the lungs is also a way of transferring heat. It works on the combined processes of convection, evaporation, and radiation. When breathing, the air inhaled is rarely the same temperature as the lungs. Consequently, heat is either inhaled or expelled with each breath. A person's breath can be seen in the cold as heat is lost to the outside. Because this method is so efficient at transferring heat, warm, moist oxygen is used to treat hypothermia patients in a clinical environment. Understanding how heat is transferred and the methods by which that transfer can be controlled can help survivors keep the body's core temperature in the 96°F to 102°F range. (See figure 13-4.)

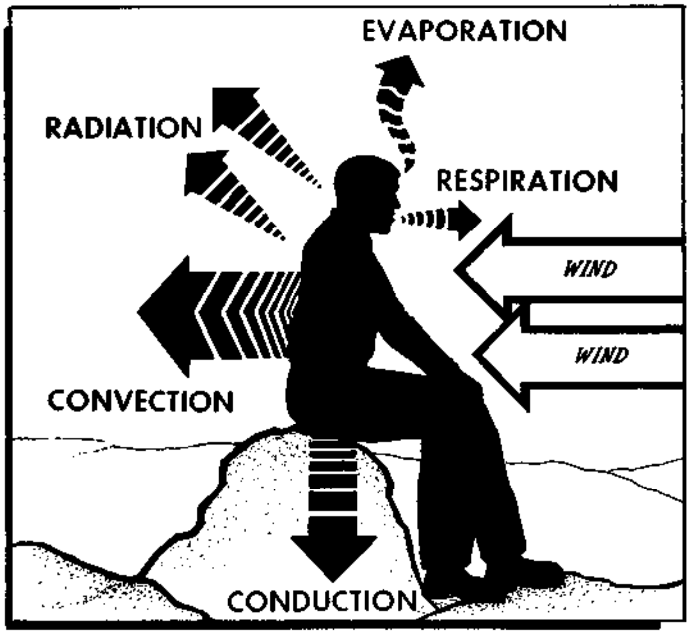

Figure 13-4. Heat Transfer.

Chapter 14

CLOTHING

14-1. Introduction. Every time people go outside they probably neglect to think about one of the most important survival-oriented assets—clothing. Clothing is often taken for granted; people tend to neglect those things which should be the most familiar to them. Clothing is an important asset to survivors and is the most immediate form of shelter. Clothing is important in staying alive, especially if food, water, shelter, and fire are limited or unobtainable. This is especially true in the first stages of an emergency situation because survivors must work to satisfy other needs. If survivors are not properly clothed, they may not survive long enough to build a fire or shelter, to find food, or to be rescued.

14-2. Protection:

a. People have worn clothing for protection since they first put on animal skins, feathers, or other coverings. In most parts of the world, people need clothing for protection from harsh climates. In snow or ice climates, people wear clothing made of fur, wool, or closely woven fabrics. They also wear warm footwear.

b. In dry climates, people wear clothing made of lightweight materials, such as cotton or linen, which have an open weave. These materials absorb perspiration and allow air to circulate around the body. People in dry climates sometimes wear white or light-colored clothes to reflect the Sun's rays. They may also wear sandals, which are cooler and more comfortable than shoes. To protect the head and neck, people wear hats as sunshades.

c. Clothing also provides protection from physical injuries caused by vegetation, terrain features, and animal life which may cause bites, stings, and cuts.

14-3. Clothing Materials:

a. Clothing is made from a variety of materials such as nylon, wool, cotton, etc. The type of material used has a significant effect on protection. Potential survivors must be aware of both the environmental conditions and the effectiveness of these different materials in order to select the best type of clothing for a particular region.

b. Clothing materials include many natural and synthetic fibers. As material is woven together, a "dead air" space is created between the material fibers. When two or three layers of material are worn, a layer of air is trapped between each layer of material creating another layer of "dead air" or insulation. The ability of these different fibers to hold "dead air" is responsible for differing insulation values.

14-4. Natural Materials. They include fur, leather, and cloth made from plant and animal fibers.

a. Fur and leather are made into some of the warmest and most durable clothing. Fur is used mainly for coats and coat linings. Leather has to be treated to make it soft and flexible and to prevent it from rotting.

b. Wool is somewhat different because it contains natural lanolin oils. Although wool is somewhat absorbent, it retains most of its insulating qualities when wet.

c. Cotton is a common plant fiber widely used to manufacture clothing. It absorbs moisture quickly and, with heat radiated from the body, will allow the moisture to pass away from the body. It does not offer much insulation when wet. It's used as an inner layer against the skin and as an outer layer with insulation (for example, wool, Dacron pile, synthetic batting) sandwiched between. The cotton protects the insulation and, therefore, provides warmth.

14-5. Synthetic Materials. Clothing manufacturers are using more and more of these materials. Many synthetic materials are stronger, more shrink-resistant, and less expensive than natural materials. Most synthetic fibers are derived from petroleum in the form of long fibers which consist of different lengths, diameters, and strengths, and sometimes have hollow cores. These fibers, woven into materials such as nylon, Dacron, and polyester, make very strong long-lasting clothing, tarps, tents, etc. Some fibers are spun into a batting type material with air space between the fibers, providing excellent insulation used inside clothing.

a. Many fabrics are blends of natural and synthetic fibers. For example, fabrics could be a mixture of cotton and polyester or wool and nylon. Nylon covered with rubber is durable and waterproof but is also heavy. There are other coverings on nylon which are waterproof but somewhat lighter and less durable. However, most coated nylon has one drawback — it will not allow for the evaporation of perspiration. Therefore, individuals may have to change the design of the garment to permit adequate ventilation (for example, wearing the garment partially unzipped).

b. Synthetic fibers are generally lighter in weight than most natural materials and have much the same insulating qualities. They work well when partially wet and dry out easily; however, they generally do not compress as well as down.

14-6. Types of Insulation:

a. Natural:

(1) Down is the soft plumage found between the skin and the contour feathers of birds. Ducks and geese are good sources for down. If used as insulation in clothing, remember that down will absorb moisture (either precipitation or perspiration) quite readily. Because of

the light weight and compressibility of down, it has wide application in cold-weather clothing and equipment. It is one of the warmest natural materials available when kept clean and dry. It provides excellent protection in cold environments; however, if the down gets wet it tends to get lumpy and loses its insulating value.

(2) Cattail plants have a worldwide distribution, with the exception of the forested regions of the far north. The cattail is a marshland plant found along lakes, ponds, and the backwaters of rivers. The fuzz on the tops of the stalks forms dead-air spaces and makes a good down-like insulation when placed between two pieces of material.

(3) Leaves from deciduous trees (those that lose their leaves each autumn) also make good insulation. To create dead-air space, leaves should be placed between two layers of material.

(4) Grasses, mosses, and other natural materials can also be used as insulation when placed between two pieces of material.

b. Synthetic:

(1) Synthetic filaments such as polyesters and acrylics absorb very little water and dry quickly. Spun synthetic filament is lighter then an equal thickness of wool and unlike down does not collapse when wet, it is also an excellent replacement for down in clothing.

(2) The nylon material in a parachute insulates well if used in the layer system because of the dead-air space. Survivors must use caution when using the parachute in cold climates. Nylon may become "cold soaked;" that is, the nylon will take on the temperature of the surrounding air. People have been known to receive frostbite when placing cold nylon against bare skin.

14-7. Insulation Measurement:

a. The next area to be considered is how well these fibers insulate from the heat or cold. The most scientific way to consider the insulating value of these fibers is to use an established criterion. The commonly accepted measurement used is a comfort level of clothing, called a "CLo" factor.

b. The CLo factor is defined as the amount of insulation which maintains normal skin temperature when the outside ambient air temperature is 70°F with a light breeze. However, the CLo factor alone is not sufficient to determine the amount of clothing required. Such variables as metabolic rate, wind conditions, and the physical makeup of the individual must be considered.

c. The body's rate of burning or metabolizing food and to produce heat varies among individuals. Therefore, some may need more insulation than others even though food intake is equal, and consequently the required CLo value must be increased. Physical activity also causes an increase in the metabolic rate and the rate of blood circulation through the body. When a person is physically active, less clothing or insulation is needed than when standing still or sitting. The effect of the wind, as shown on the windchill chart, must be considered (figure 13-1). When the combination of temperature and wind drops the chill factor to minus 100°F or lower, the prescribed CLo for protecting the body may be inapplicable (over a long period of time) without relief from the wind. For example, when the temperature is minus 60°F, the wind is blowing 60 to 70 miles per hour, and the resultant chill factor exceeds minus 150°F, clothing alone is inadequate to sustain life. Shelter is essential.

d. The physical build of a person also affects the amount of heat and cold that can be endured. For example, a very thin person will not be able to endure as low a temperature as one who has a layer of fat below the skin. Conversely, heavy people will not be able to endure extreme heat as effectively as thinner people.

e. In the Air Force clothing inventory, there are many items which fulfill the need for insulating the body. They are made of the different fibers previously mentioned, and when worn in layers, provide varying degrees of insulative CLo value. The following average zone temperature chart is a guide in determining the best combination of clothing to wear.

TEMPERATURE RANGE	CLo REQUIRED
86 to 68°F	1 - Lightweight
68 to 50°F	2 - Intermediate Weight
50 to 32°F	3 - Intermediate Weight
32 to 14°F	3.5 - Heavyweight
14 to -4°F	4.0 - Heavyweight
-4 to -40°F	4.0 - Heavyweight

The amount of CLo value per layer of fabric is determined by the loft (distance between the inner and outer surfaces) and the amount of dead air held within the fabric. Some examples of the CLo factors and some items of clothing are:

LAYERS:	1 - Aramid underwear (1 layer)	0.6 CLo
	2 - Aramid underwear (2 layers)	1.5 CLo
	3 - Quilted liners	1.9 CLo
	4 - Nomex coveralls	.6 CLo
	5 - Winter coveralls	1.2 CLo
	6 - Nomex jacket	1.9 CLo

This total amount of insulation should keep the average person warm at a low temperature. When comparing items one and two in the above example, it shows when doubling the layer of underwear, the CLo value more than doubles. This is true not only on the number one item but between all layers of any clothing system. Therefore, one gains added protection by using several very thin layers of insulation rather than two thick layers. The air held between these thin layers increases the insulation value.

f. The use of many thin layers also provides (through removal of desired number of layers) the ability to closely regulate the amount of heat retained inside the clothing. The ability to regulate body temperature helps to alleviate the problem of overheating and sweating, and preserves the effectiveness of the insulation.

g. The principle of using many thin layers of clothing can also be applied to the "sleeping system" (sleeping bag, liner, and bed). This system uses many layers of synthetic material, one inside the other, to form the amount of dead air needed to keep warm. To improve this system, a survivor should wear clean and dry clothing in layers (the layer system) in cold climates. While discussing the layer system, it is important to define the "COLDER" principle. This acronym is used to aid in remembering how to use and take care of clothing.

C - Keep clothing *C*lean.
O - Avoid *O*verheating.
L - Wear clothing *L*oose and in *L*ayers.
D - Keep clothing *D*ry.
E - *E*xamine clothing for defects or wear.
R - Keep clothing *R*epaired.

(1) Clean. Dirt and other materials inside fabrics will cause the insulation to be ineffective, abrade and cut the fibers which make up the fabric, and cause holes. Washing clothing in the field may be impractical; therefore, survivors should concentrate on using proper techniques to prevent soiling clothing.

(2) Overheating. Clothing best serves the purpose of preserving body heat when worn in layers as follows: absorbent material next to the body, insulating layers, and outer garments to protect against wind and rain. Because of the rapid change in temperature, wind, and physical exertion, garments should allow donning and removal quickly and easily. Ventilation is essential when working because enclosing the body in an airtight layer system results in perspiration which wets clothing, thus reducing its insulating qualities.

(3) Loose. Garments should be loose fitting to avoid reducing blood circulation and restricting body movement. Additionally, the garment should overhang the waist, wrists, ankles, and neck to reduce body heat loss.

(4) Dry. Keep clothing dry since a small amount of moisture in the insulation fibers will cause heat losses up to 25 times faster than dry clothing. Internally produced moisture is as damaging as is externally dampened clothing. The outer layer should protect the inner layers from moisture as well as from abrasion of fibers; for example, wool rubbing on logs or rocks, etc. The outer shell keeps dirt and other contaminants out of the clothing. Clothing can be dried in many ways. Fires are often used; however, take care to avoid burning the items. The "bare hand" test is very effective. Place one hand near the fire in the approximate place the wet items will be and count to three slowly. If this can be done without feeling excessive heat, it should be safe to dry items there. Never leave any item unattended while it is drying. Leather boots, gloves, and mitten shells require extreme care to prevent shrinkage, stiffening, and cracking. The best way to dry boots is upright beside the fire (not upside down on sticks because the moisture does not escape the boot) or simply walk them dry in the milder climates. The Sun and wind can be used to dry clothing with little supervision except for checking occasionally on the incoming weather and to make sure the article is secure. Freeze-drying is used in subzero temperatures with great success. Survivors let water freeze on or inside the item and then shake, bend, or beat it to cause the ice particles to fall free from the material. Tightly woven materials work better with this method than do open fibers.

(5) Examine. All clothing items should be inspected regularly for signs of damage or soil.

(6) Repair. Eskimos set an excellent example in the meticulous care they provide for their clothing. When damage is detected, immediately repair it.

h. The neck, head, hands, armpits, groin, and feet lose more heat than other parts of the body and require greater protection. Work with infrared film shows tremendous heat loss in those areas when not properly clothed. Survivors in a cold environment are in a real emergency situation without proper clothing. Figure 14-1 shows some examples of how military clothing works to hold body heat.

i. Models wearing samples of aircrew attire appear as spectral figures in a thermogram, an image revealing differences in infrared heat radiated from their clothing and exposed skin. White is warmest; red, yellow, green, blue, and magenta form a declining temperature scale spanning about 15 degrees; while black represents all lower temperatures. Almost the entire scale is seen on the model in boxer shorts. Warm, white spots appear on the underarm and neck. Only the shorts block radiation from the groin. Temperatures cool along the arm to dark blue fingertips far from the heat-producing torso. The addition of the next layer of clothing (Aramid long underwear) prevents heat loss except where it is tight against the body. As more layers are added, it is easy to see the areas of greatest concern are the head, hands, and feet. These areas are difficult for crewmembers to properly insulate while flying an aircraft. Mittens are ineffective due to the degraded manual dexterity. Likewise, it is difficult to feel the rudder pedal action while wearing bulky warm boots. These problems require inclusion of warm hats, mittens, and footgear (mukluk type) in survival kits during cold weather operation. Research has shown when a CLo value of 10 is used to insulate the head, hands, and feet and the rest of the body is only protected by one CLo, the average individual can be exposed to low temperatures (-10°F) comfortably for a reasonable period of time (30 to 40 minutes). When the amount of CLo value placed on the

GROUP ONE - REMOVING CLOTHES

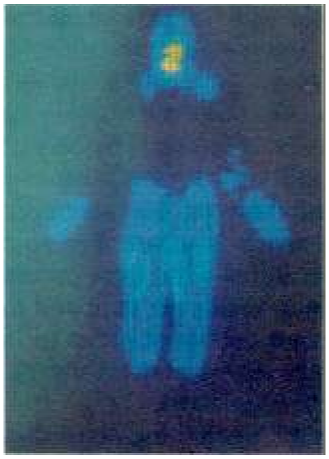

1-1: Fully clothed

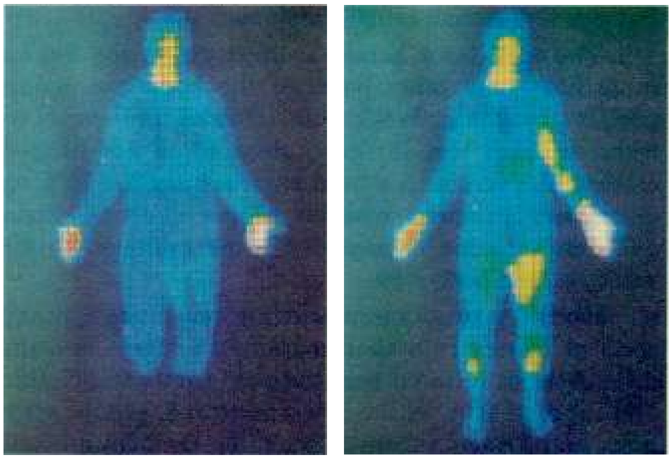

1-2: Flight jacket, wool cap and mittens, and leather shell

1-3: Flight suit

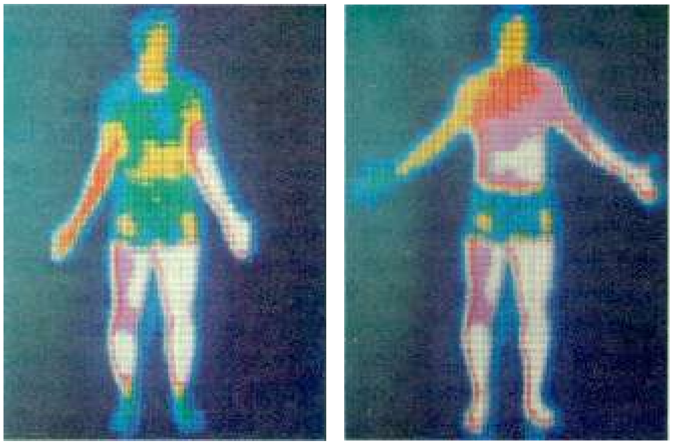

1-4: Thermals

1-5: "T" shirt and two pair cotton socks

GROUP TWO - DONNING CLOTHES

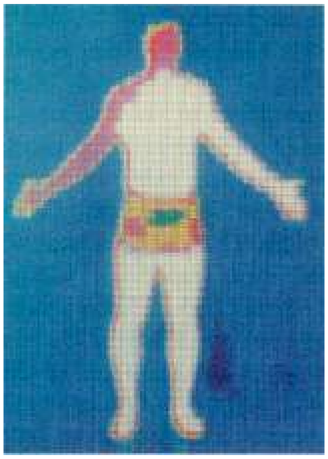

2-1: Unclothed

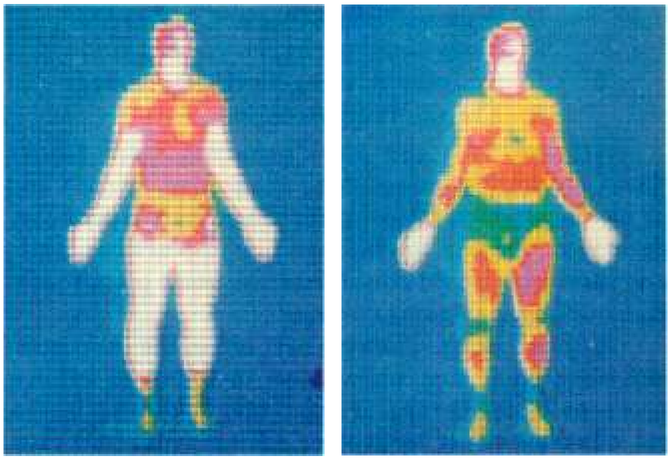

2-2: "T" shirt and two pair cotton socks

2-3: Thermals

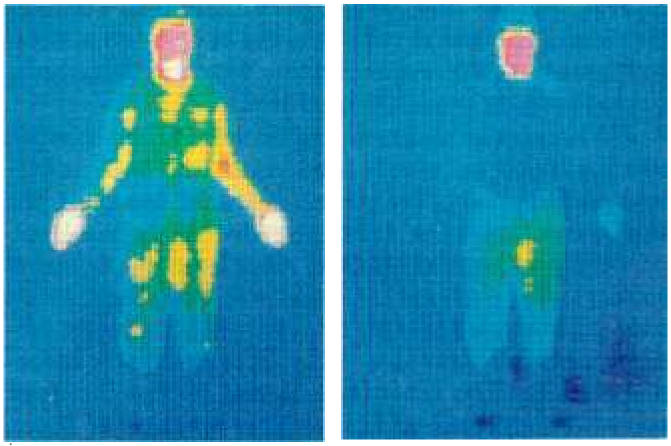

2-4: Flight suit

2-5: Flight jacket, wool cap and mittens, and leather shell

NOTE: Dark blue indicates no heat loss; the lighter the color, the greater the heat loss.

Figure 14-1. Thermogram of Body Heat Loss.

individual is reversed, the amount of time a survivor can spend in cold weather is greatly reduced due to the heat loss from their extremities. This same principle works in reverse in hot parts of the world if one submerges the head, hands, or feet in cold water, it lets the most vascular parts of the body lose heat quickly.

14-8. Clothing Wear in Snow and Ice Areas:

a. The survivor should:

(1) Avoid restricting the circulation. Clothing should not be worn so tight that it restricts the flow of blood which distributes the body heat and helps prevent frostbite. When wearing more than one pair of socks or gloves, ensure that each succeeding pair is large enough to fit comfortably over the other. Don't wear three or four pairs of socks in a shoe fitted for only one or two pairs. Release any restriction caused by twisted clothing or a tight parachute harness.

(2) Keep the head and ears covered. Survivors will lose as much as 50 percent of their total body heat from an unprotected head at 50°F.

(3) When exerting the body, prevent perspiration by opening clothing at the neck and wrists and loosening it at the waist. If the body is still warm, comfort can be obtained by taking off outer layers of clothing, one layer at a time. When work stops, the individual should put the clothing on again to prevent chilling.

(4) If boots are big enough, use dry grass, moss, or other material for added insulation around the feet. Footgear can be improved by wrapping parachute cloth or other fabric lined with dry grass or moss for insulation.

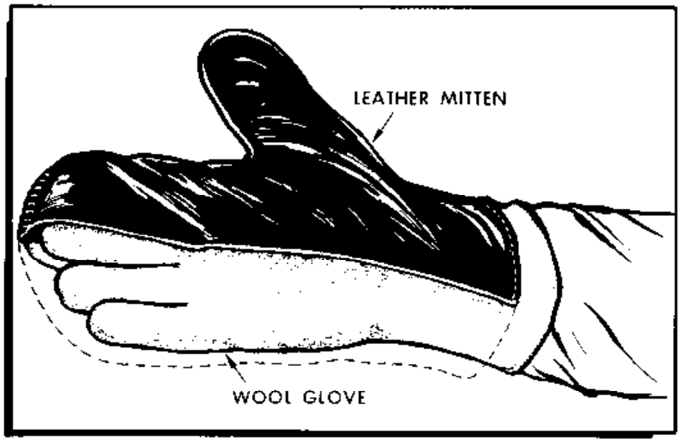

Figure 14-2. Layer System for Hands.

b. Felt booties and mukluks with the proper socks and insoles are best for dry, cold weather. Rubber-bottomed boot shoepacs with leather tops are best for wet weather. Mukluks should not be worn in wet weather. The vapor-barrier rubber boots can be worn under both conditions and are best at extremely low temperatures. The air release valve should be closed at ground level. These valves are designed to release pressure when airborne. Air should not be blown into the valves as the moisture could decrease insulation.

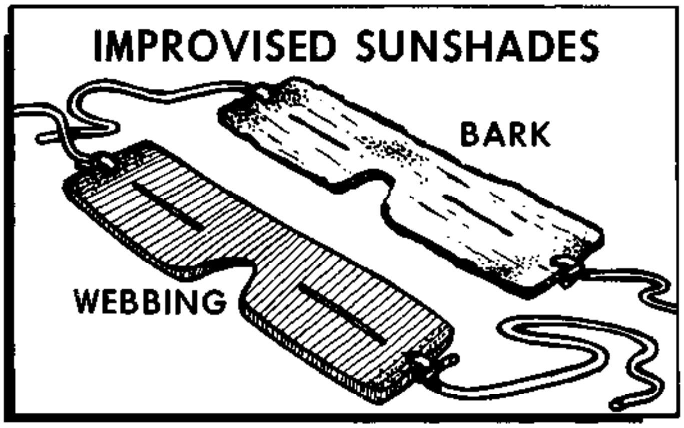

Figure 14-3. Improvised Goggles.

c. Clothing should be kept as dry as possible. Snow must be brushed from clothing before entering a shelter or going near a fire. The survivors should beat the frost out of garments before warming them, and dry them on a rack near a fire. Socks should be dried thoroughly.

d. One or two pairs of wool gloves and (or) mittens should be worn inside a waterproof shell (figure 14-2). If survivors have to expose their hands, they should warm them inside their clothing.

e. To help prevent sun or snow blindness, a survivor should wear sun or snow goggles or improvise a shield with a small horizontal slit opening (figure 14-3).

f. In strong wind or extreme cold, as a last resort, a survivor should wrap up in parachute material, if available, and get into some type of shelter or behind a windbreak. Extreme care should be taken with hard materials, such as synthetics, as they may become cold soaked and require more time to warm.

g. At night, survivors should arrange dry spare clothing loosely around and under the shoulders and hips to help keep the body warm. Wet clothes should never be worn into the sleeping bag. The moisture destroys the insulation value of the bag.

h. If survivors fall into water, they should roll in dry snow to blot up moisture, brush off the snow, and roll again until most of the water is absorbed. They should not remove footwear until they are in a shelter or beside a fire.

i. All clothing made of wool offers good protection when used as an inner layer. When wool is used next to the face and neck, survivors should be cautioned that moisture from the breath will condense on the surface and cause the insulating value to decrease. The use of a wool scarf wrapped around the mouth and nose is an excellent way to prevent cold injury, but it needs to be

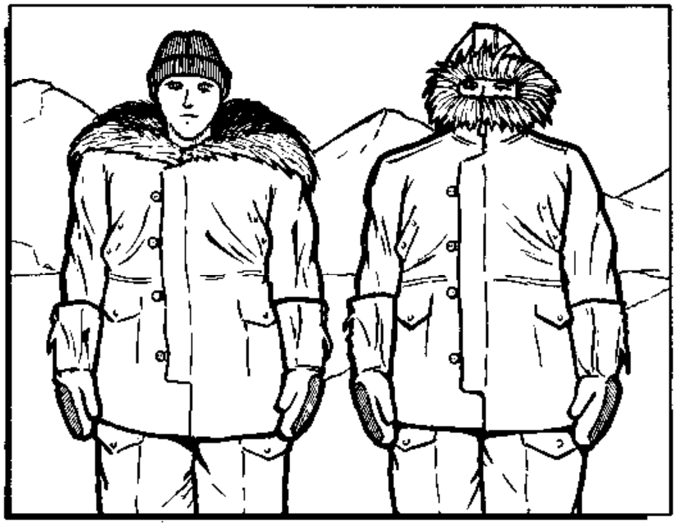

Figure 14-4. Proper Wear of Parka.

de-iced on a regular basis to prevent freezing flesh adjacent to it. An extra shell is generally worn over the warming layers to protect them and to act as a windbreak.

j. Other headgear includes the pile cap and hood. These items are most effective when used with a covering for the face in extreme cold. The pile cap is extremely warm where it is insulated, but it offers little protection for the face and back of the neck.

k. The hood is designed to funnel the radiant heat rising from the rest of the body and to recycle it to keep the neck, head, and face warm (figure 14-4). The individual's ability to tolerate cold should dictate the size of the front opening of the hood. The "tunnel" of a parka hood is usually lined with fur of some kind to act as a protecting device for the face. This same fur also helps to protect the hood from the moisture expelled during breathing. The closed tunnel holds heat close to the face longer; the open one allows the heat to escape more freely. As the frost settles on the hair of the fur, it should be shaken from time to time to keep it free of ice buildup.

l. Sleeping systems (sleeping bag, liner, and bed) are the transition "clothing" used between normal daytime activities and sleep (figure 14-7).

m. The insulating material in the sleeping bag may be synthetic or it may be down and feathers. (Feathers and down lining require extra protection from moisture). However, the covering is nylon. Survivors must realize that sleeping bags are compressed when packed and must be fluffed before use to restore insulation value. Clean and dry socks, mittens, and other clothing can be used to provide additional insulation.

n. Footgear is critical in a survival situation because walking is the only means of mobility. Therefore, care of footgear is essential both before and during a survival situation. Recommendations for care are:

(1) Ensure footgear is properly "broken-in" before flying.

(2) "Treat" footgear to ensure water-repellency (follow manufacturer's recommendations).

(3) Keep leather boots as dry as possible.

o. Mukluks have been around for thousands of years and have proven their worth in extremely cold weather. The Air Force mukluks are made of cotton duck with rubber-cleated soles and heels. (See figure 14-6.) They have slide fasteners from instep to collar, laces at instep and collar, and are 18 inches high. They are used by flying and ground personnel operating under *dry*, cold conditions in temperatures below +15°F. Survivors should change liners daily when possible.

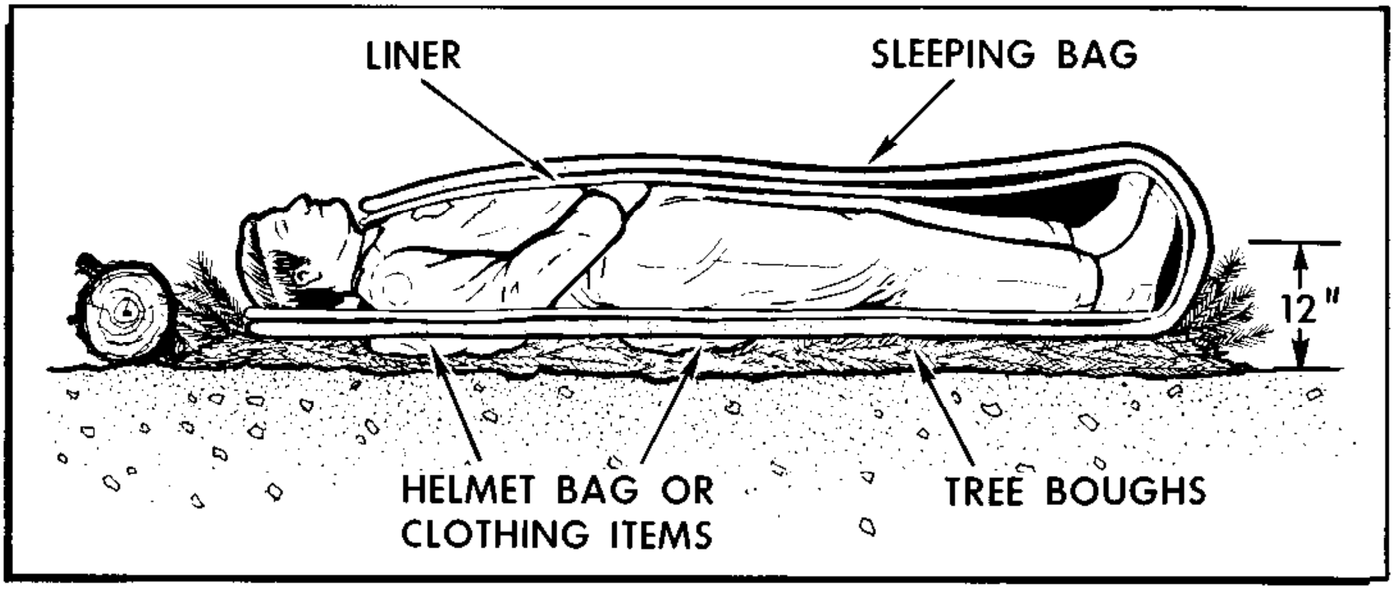

Figure 14-5. Sleeping System.

Figure 14-6. Issued Mukluks.

14-9. Care of the Feet. Foot care is critical in a survival situation. Improvising foot gear may be essential to caring for feet.

a. Moose Hock Shoe. The hock skin of a moose or caribou will provide a suitable pair of shoes (figure 14-7). Cut skin around leg at A and B. Separate from the leg and pull it over the hoof. Shape and sew up small end C. Slit skin from A to B; bore holes on each side of cut for lacing; turn inside out, and lace with rawhide, suspension line, or other suitable material.

b. Grass Insoles. Used extensively by northern natives to construct inner soles. Grass is a good insulator and will collect moisture from the feet. The survivor should use the following procedure to prepare grass for use as inner soles: Grasp a sheaf of tall grass, about one-half inch in diameter, with both hands. Rotate the hands in opposite directions. The grass will break up or "fluff" into a soft mass. Form this fluff into oblong shapes and spread it evenly throughout the shoes. The inner soles should be about an inch thick. Remove these inner soles at night and make new ones the following day.

c. Hudson Bay Duffel. A triangular piece of material used as a foot covering. To improvise this foot covering, a survivor can use the following procedures:

(1) Cut two to four layers of parachute cloth into a 30-inch square.

(2) Fold this square to form a triangle.

(3) Place the foot on this triangle with the toes pointing at one corner.

(4) Fold the front cover up over the toes.

(5) Fold the side corners, one at a time, over the instep. This completes the foot wrap. (See figure 14-8.)

d. Gaiters. Made from parachute cloth, webbing, or canvas. Gaiters help keep sand and snow out of shoes and protect the legs from bites and scratches (figure 14-9).

e. Double Socks. Cushion padding, feathers, dry grass, or fur stuffed between layers of socks. Wrap parachute or aircraft fabric around the feet and tie above the ankles. A combination of two or more types of improvised footwear may be more desirable and more efficient than any single type (figure 14-10).

14-10. Clothing in the Summer Arctic:

a. In the summer arctic, there are clouds of mosquitoes and black flies so thick a person can scarcely see through them. Survivors can protect themselves by wearing proper clothing to ensure no bare skin is exposed. A good head net and gloves should be worn.

b. Head nets must stand out from the face so they won't touch the skin. Issued head nets are either black or green. If one needs to be improvised they can be sewn to the brim of the hat or can be attached with an elastic band that fits around the crown. Black is the best color, as it can be seen through more easily than green or white. A heavy tape encasing a drawstring should be attached to the bottom of the head net for tying snugly at the collar. Hoops of wire fastened on the inside will make the net stand out from the face and at the same time allow it to be packed flat. The larger they are, the better the ventilation. But very large nets will not be as effective in wooded country where they may become snagged on brush.

c. Gloves are hot, but are a necessity where flies are found in swamps. Kid gloves with a 6-inch gauntlet closing the gap at the wrist and ending with an elastic band halfway to the elbow are best. For fine work, kid gloves with the fingers cut off are good. Cotton/Nomex work gloves are better than no protection at all, but mosquitoes will bite through them. Treating the gloves with insect repellent will help. Smoky clothing may also help to keep insects away. (See figure 14-11.)

d. A survivor should remember that mosquitoes do not often bite through two layers of cloth; therefore, a lightweight undershirt and long underwear will help. To

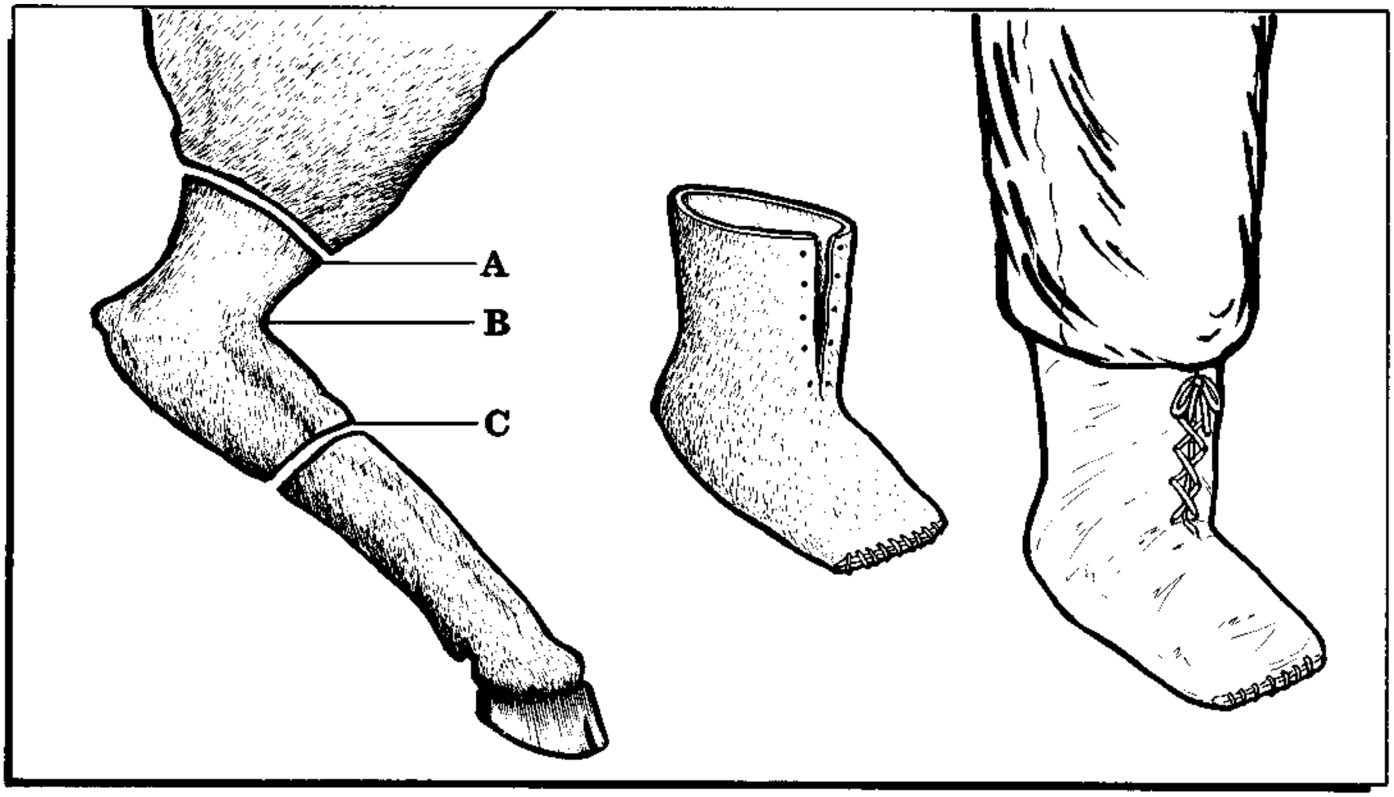

Figure 14-7. Moose Hock Shoes.

protect ankles, blouse the bottoms of trousers around boots, or wear some type of leggings (gaiters).

e. If the head net is lost or none is available, make the best of a bad situation by wearing sunglasses with improvised screened sides, plugging ears lightly with cotton, and tying a handkerchief around the neck. Treat clothing with insect repellent at night.

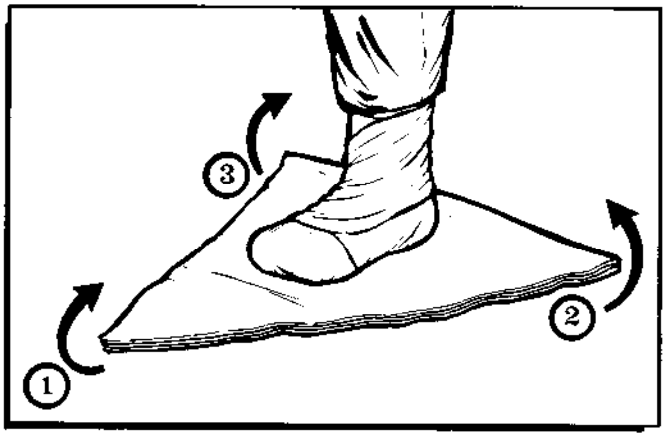

Figure 14-8. Hudson Bay Duffel.

14-11. Clothing at Sea. In cold oceans, survivors must try to stay dry and keep warm. If wet, they should use a wind screen to decrease the cooling effects of the wind. They should also remove, wring out, and replace outer

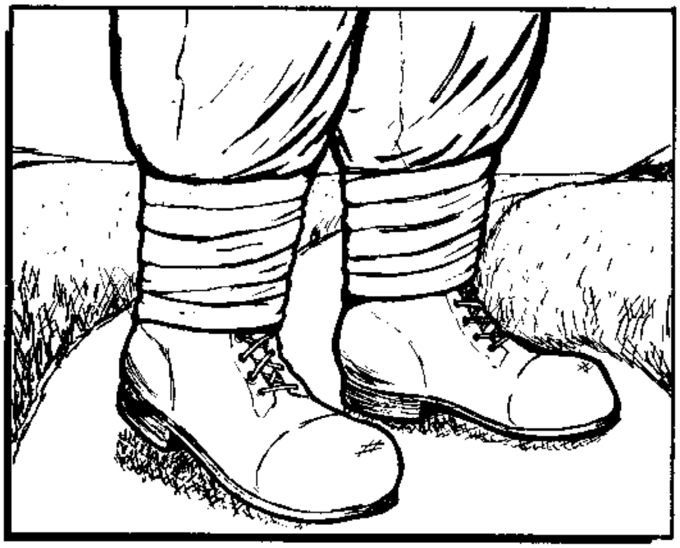

Figure 14-9. Gaiters.

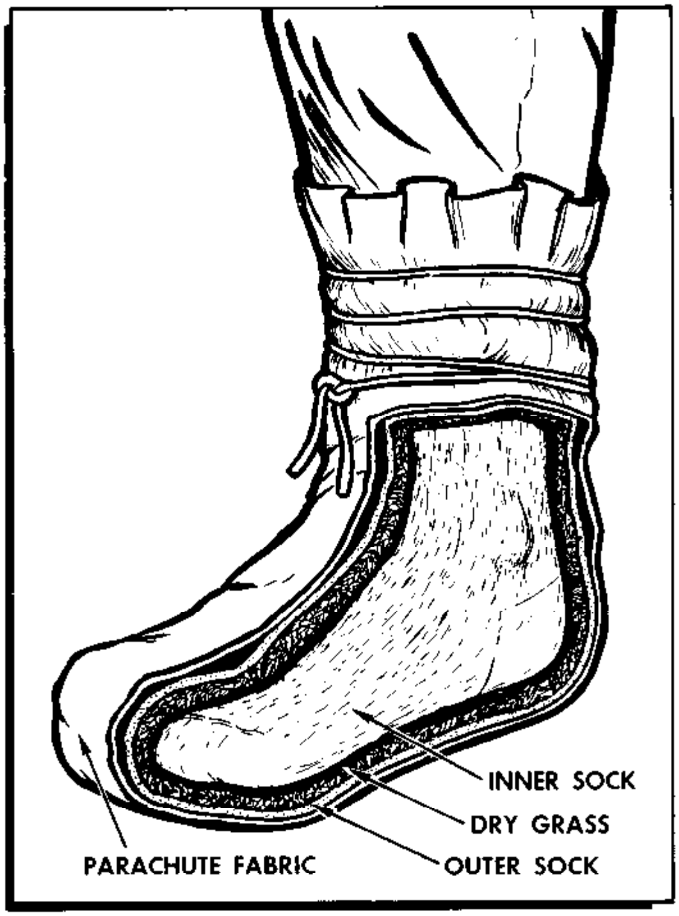

Figure 14-10. Double Socks.

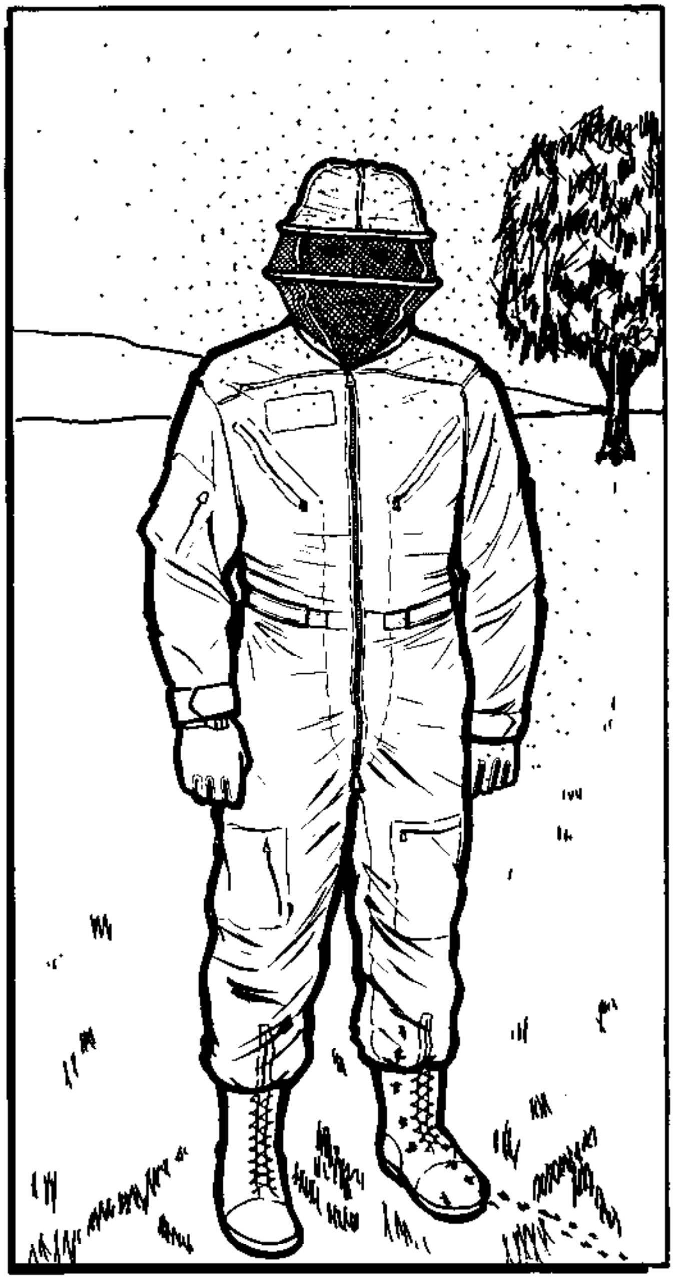

Figure 14-11. Insect Protection.

garments or change into dry clothing. Hats, socks, and gloves should also be dried. If any survivors are dry, they should share extra clothes with those who are wet. Wet personnel should be given the most sheltered positions in the raft. Let them warm their hands and feet against those who are dry. Survivors should put on any extra clothing available. If no anti-exposure suits are provided, they can drape extra clothing around their shoulders and over their heads. Clothes should be loose and comfortable. Also, survivors should attempt to keep the floor of the raft dry. For insulation, covering the floor with any available material will help. Survivors should huddle together on the floor of the raft and spread extra tarpaulin, sail, or parachute material over the group. If in a 20- or 25-man raft, canopy sides can be lowered. Performing mild exercises to restore circulation may be helpful. Survivors should exercise fingers, toes, shoulders, and buttock muscles. Mild exercise will help keep the body warm, stave off muscle spasms, and possibly prevent medical problems. Survivors should warm hands under armpits and periodically raise feet slightly and hold them up for a minute or two. They should also move face muscles frequently to prevent frostbite. Shivering is the body's way of quickly generating heat and is considered normal. However, persistent shivering may lead to uncontrollable muscle spasms. They can be avoided by exercising muscles. If water is available, additional rations should be given to those suffering from exposure to cold. Survivors should eat small amounts frequently rather than one large meal.

14-12. Antiexposure Garments:

a. Assemblies. The antiexposure assemblies, both quick donning and constant wear, are designed for personnel participating in over-water flights where unprotected or prolonged exposure to the climatic conditions

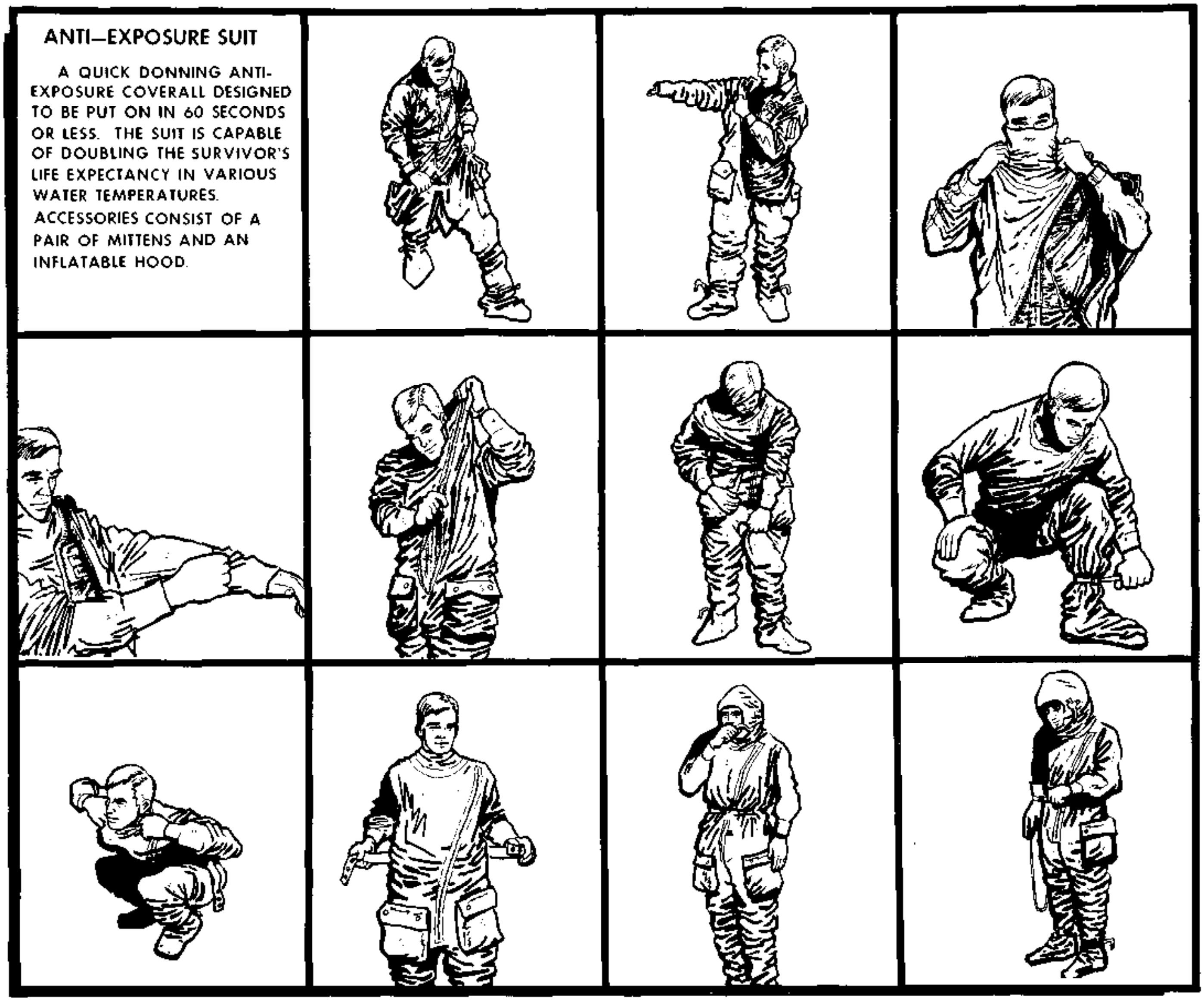

ANTI—EXPOSURE SUIT

A QUICK DONNING ANTI-EXPOSURE COVERALL DESIGNED TO BE PUT ON IN 60 SECONDS OR LESS. THE SUIT IS CAPABLE OF DOUBLING THE SURVIVOR'S LIFE EXPECTANCY IN VARIOUS WATER TEMPERATURES. ACCESSORIES CONSIST OF A PAIR OF MITTENS AND AN INFLATABLE HOOD.

Figure 14-12. Donning Antiexposure Suit.

of cold air and (or) cold water (as a result of ditching or abandoning an aircraft) would be dangerous or could prove fatal. The suit provides protection from the wind and insulation against the chill of the ocean. The result of exposure in the water is illustrated in figures 13-2 and 13-3. Exposure time varies depending on the particular antiexposure assembly worn, the cold sensitiveness of the person, and survival procedures used.

b. Quick-Donning Antiexposure Flying Coverall. Some antiexposure coveralls are designed for quick donning (approximately 1 minute) before emergency ditching. After ditching the aircraft, the coverall protects the wearer from exposure while swimming in cold water, and from exposure to wind, spray, and rain when adrift in a liferaft.

(1) The coverall is a one-size garment made from chloroprene-coated nylon cloth. It has two expandable-type patch pockets, an adjustable waist belt, and attached boots with adjustable ankle straps. One pair of insulated, adjustable wrist strap mittens, each with a strap attached to a pocket, is provided. A hood, also attached with a strap, is in the left pocket. A carrying case with instructions and a snap fastener closure is furnished for stowing in the aircraft.

(2) To use the coverall, personnel should wear it over regular flight clothing. It is large enough to wear over the usual flight gear. The gloves and hood are stowed in the pockets of the coverall and are normally worn after boarding the liferaft.

(3) The survivor should be extremely careful when donning the coverall to prevent damage by snagging,

tearing, or puncturing it on projecting objects. After donning the coverall, the waist band and boot ankle straps should be adjusted to take up fullness. If possible, crewmembers should stoop while pulling the neck seal to expel air trapped in the suit. When jumping into the water, they should leap feet first with hands and arms close to sides or brought together above the head (figure 14-12). Note there is a constant wear exposure suit designed to be worn continuously during overwater flights where the water temperature is 60 degrees or below. The Command may waiver it to 51 degrees.

14-13. Warm Oceans. Protection against the Sun and securing drinking water are the most important problems. A survivor should keep the body covered as much as possible to avoid sunburn. A sunshade can be improvised out of any materials available or the canopy provided with the raft may be used. If the heat becomes too intense, survivors may dampen clothing with sea water to promote evaporation and cooling. The use of sunburn preventive cream or a Chapstick is advisable. Remember, the body must be kept covered completely. Exposure to the Sun increases thirst, wastes precious water, reduces the body's water content, and causes serious burns. Survivors should roll down their sleeves, pull up their socks, close their collars, wear a hat or improvised headgear, use a piece of cloth as a shield for the back of the neck, and wear sunglasses or improvise eye covers.

14-14. Tropical Climates:

a. In tropical areas, the body should be kept covered for prevention of insect bites, scratches, and sunburn.

b. When moving through vegetation, survivors should roll down their sleeves, wear gloves, and blouse the legs of their pants or tie them over their boot tops. Improvised puttees (gaiters) can be made from parachute material or any available fabric. This will protect legs from ticks and leeches.

c. Loosely worn clothing will keep survivors cooler, especially when subjected to the direct rays of the Sun.

d. Survivors should wear a head net or tie material around the head for protection against insects. The most active time for insects is at dawn and dusk. An insect repellent should be used at these times.

e. In open country or in high grass, survivors should wear a neck cloth or improvised head covering for protection from sunburn and (or) dust. They should also move carefully through tall grass, as some sharp-edged grasses can cut clothing to shreds. Survivors should dry clothing before nightfall. If an extra change of clothing is available, effort should be made to keep it clean and dry.

14-15. Dry Climates:

a. In the dry climates of the world, clothing will be needed for protection against sunburn, heat, sand, and insects. Survivors should not discard any clothing. They should keep their head and body covered and blouse the legs of pants over the tops of footwear during the day. Survivors should not roll up sleeves, but keep them rolled down and loose at the cuff to stay cool.

b. Survivors should keep in mind that the people who live in the hot dry areas of the world usually wear heavy white flowing robes which protect almost every inch of their bodies. The only areas open to the Sun are the face and the eyes. This produces an area of higher humidity between the body and the clothing, which helps keep them cooler and conserves their perspiration (figure 14-13). The white clothing also reflects the sunlight.

c. Survivors should wear a cloth neckpiece to cover the back of the neck and protect it from the Sun. A T-

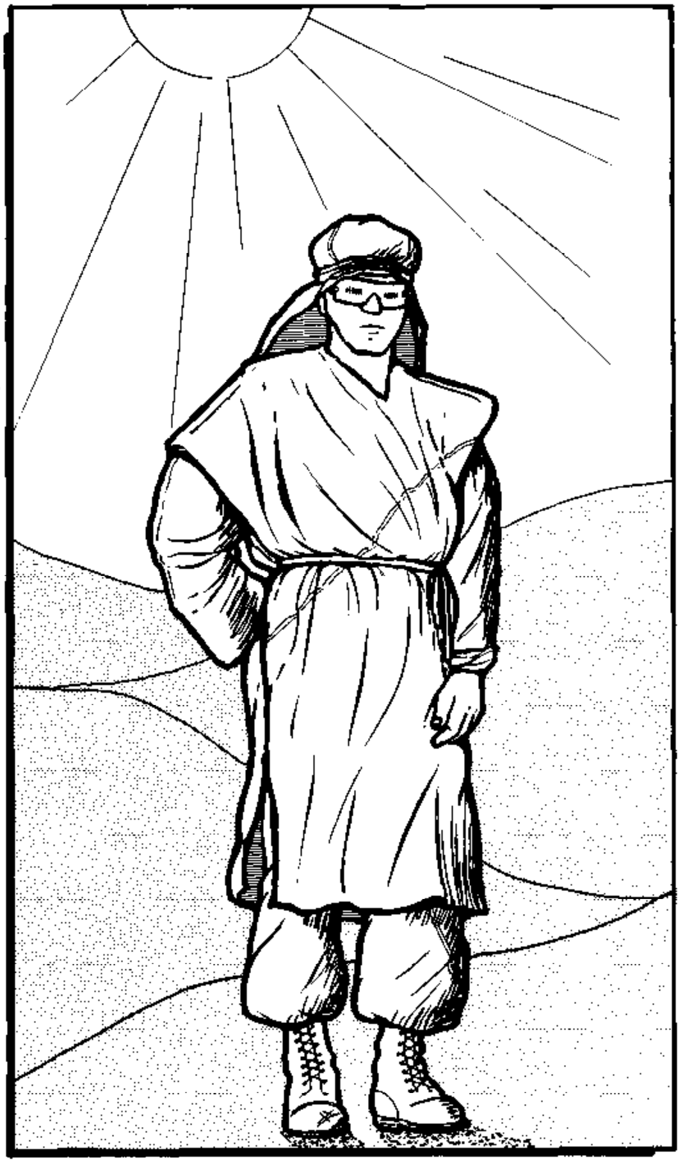

Figure 14-13. Protective Desert Clothing.

shirt makes an excellent neck drape, with the extra material used as padding under the cap. If hats are not available, survivors can make headpieces like those worn by the Arabs, as shown in figure 14-13. During dust storms, they should wear a covering for the mouth and nose; parachute cloth will work.

d. If shoes are lost or if they wear out, survivors can improvise footgear. One example of this is the "Russian Sock." Parachute material can be used to improvise these socks. The parachute material is cut into strips approximately 2 feet long and 4 inches wide. These strips are wrapped bandage fashion around the feet and ankles. Socks made in this fashion will provide comfort and protection for the feet.

Chapter 15

SHELTER

15-1. Introduction. Shelter is anything that protects a survivor from the environmental hazards. The information in this chapter describes how the environment influences shelter site selection and factors which survivors must consider before constructing an adequate shelter. The techniques and procedures for constructing shelters for various types of protection are also presented.

15-2. Shelter Considerations. The location and type of shelter built by survivors vary with each survival situation. There are many things to consider when picking a site. Survivors should consider the time and energy required to establish an adequate camp, weather conditions, life forms (human, plant, and animal), terrain, and time of day. Every effort should be made to use as little energy as possible and yet attain maximum protection from the environment.

a. Time. Late afternoon is not the best time to look for a site which will meet that day's shelter requirements. If survivors wait until the last minute, they may be forced to use poor materials in unfavorable conditions. They must constantly be thinking of ways to satisfy their needs for protection from environmental hazards.

b. Weather. Weather conditions are a key consideration when selecting a shelter site. Failure to consider the weather could have disastrous results. Some major weather factors which can influence the survivor's choice of shelter type and site selection are temperature, wind, and precipitation.

(1) Temperature. Temperatures can vary considerably within a given area. Situating a campsite in low areas such as a valley in cold regions can expose survivors to low night temperatures and windchill factors. Colder temperatures are found along valley floors which are sometimes referred to as "cold air sumps." It may be advantageous to situate campsites to take advantage of the Sun. Survivors could place their shelters in open areas during the colder months for added warmth, and in shaded areas for protection from the Sun during periods of hotter weather. In some areas a compromise may have to be made. For example, in many deserts the daytime temperatures can be very high while low temperatures at night can turn water to ice. Protection from both heat and cold are needed in these areas. Shelter type and location should be chosen to provide protection from the existing temperature conditions.

(2) Wind. Wind can be either an advantage or a disadvantage depending upon the temperature of the area and the velocity of the wind. During the summer or on warm days, survivors can take advantage of the cool breezes and protection the wind provides from insects by locating their camps on knolls or spits of land. Conversely, wind can become an annoyance or even a hazard as blowing sand, dust, or snow can cause skin and eye irritation and damage to clothing and equipment. On cold days or during winter months, survivors should seek shelter sites which are protected from the effects of windchill and drifting snow.

(3) Precipitation. The many forms of precipitation (rain, sleet, hail, or snow) can also present problems for survivors. Shelter sites should be out of major drainages and other low areas to provide protection from flash floods or mud slides resulting from heavy rains. Snow can also be a great danger if shelters are placed in potential avalanche areas.

c. Life Forms. All life forms (plant, human, and animal) must be considered when selecting the campsite and the type of shelter that will be used. The "human" factor may mean the enemy or other groups from whom survivors wish to remain undetected. Information regarding this aspect of shelters and shelter site selection is in part nine of this regulation (Evasion). For a shelter to be adequate, certain factors must be considered, especially if extended survival is expected.

(1) Insect life can cause personal discomfort, disease, and injury. By locating shelters on knolls, ridges, or any other area that has a breeze or steady wind, survivors can reduce the number of flying insects in their area. Staying away from standing water sources will help to avoid mosquitoes, bees, wasps, and hornets. Ants can be a major problem; some species will vigorously defend their territories with painful stings or bites or particularly distressing pungent odors.

(2) Large and small animals can also be a problem, especially if the camp is situated near their trails or waterholes.

(3) Dead trees that are standing, and trees with dead branches should be avoided. Wind may cause them to fall, causing injuries or death. Poisonous plants, such as poison oak or poison ivy, must also be avoided when locating a shelter.

d. Terrain. Terrain hazards may not be as apparent as weather and animal life hazards, but they can be many times more dangerous. Avalanche, rock, dry streambeds, or mud-slide areas should be avoided. These areas can be recognized by either a clear path or a path of secondary vegetation, such as 1- to 15-foot tall vegetation or other new growth which extends from the top to the bottom of a hill or mountain. Survivors should not choose shelter sites at the bottom of steep slopes which may be prone to slides. Likewise, there is a danger in camping at the bottom of steep scree or talus slopes. Additionally, rock overhang must be checked for safety before using it as a shelter.

15-3. Location:

a. Four prerequisites must be satisfied when selecting a shelter location.

(1) The first is being near water, food, fuel, and a signal or recovery site.

(2) The second is that the area be safe, providing natural protection from environmental hazards.

(3) The third is that sufficient materials be available to construct the shelter. In some cases, the "shelter" may already be present. Survivors seriously limit themselves if they assume shelters *must* be a fabricated framework having predetermined dimensions and a cover of parachute material or a signal paulin. More appropriately, survivors should consider using sheltered *places* already in existence in the immediate area. This does not rule out shelters with a fabricated framework and parachute or other manufactured material covering; it simply enlarges the scope of what can be used as a survival shelter.

(4) Finally, the area chosen must be both large enough and level enough for the survivor to lie down. Personal comfort is an important fundamental for survivors to consider. An adequate shelter provides physical and mental well-being for sound rest. Adequate rest is extremely vital if survivors are to make sound decisions. Their need for rest becomes more critical as time passes and rescue or return is delayed. Before actually constructing a shelter, survivors must determine the specific purpose of the shelter. The following factors influence the type of shelter to be fabricated.

(a) Rain or other precipitation.

(b) Cold.

(c) Heat.

(d) Insects.

(e) Available materials nearby (manufactured or natural).

(f) Length of expected stay.

(g) Enemy presence in the area—evasion "shelters" are covered in part nine of the regulation (Evasion).

(h) Number and physical condition of survivors.

b. If possible, survivors should try to find a shelter which needs little work to be adequate. Using what is already there, so that complete construction of a shelter is not necessary, saves time and energy. For example, rock overhangs, caves, large crevices, fallen logs, root buttresses, or snow banks can all be modified to provide adequate shelter. Modifications may include adding snow blocks to finish off an existing tree well shelter, increasing the insulation of the shelter by using vegetation or parachute material, etc., or building a reflector fire in front of a rock overhang or cave. Survivors must consider the amount of energy required to build the shelter. It is not really wise to spend a great deal of time and energy in constructing a shelter if nature has provided a natural shelter nearby which will satisfy the survivor's needs. See figure 15-1 for examples of naturally occurring shelters.

c. The size limitations of a shelter are important only if there is either a lack of material on hand or if it is cold. Otherwise, the shelter should be large enough to be comfortable yet not so large as to cause an excessive amount of work. Any shelter, naturally occurring or otherwise, in which a fire is to be built must have a ventilation system which will provide fresh air and allow smoke and carbon monoxide to escape. Even if a fire does not produce visible smoke (such as heat tabs), the shelter must still be vented. See figure 15-27 for placement of ventilation holes in a snow cave. If a fire is to be placed outside the shelter, the opening of the shelter should be placed 90 degrees to the prevailing wind. This will reduce the chances of sparks and smoke being blown into the shelter if the wind should reverse direction in the morning and evening. This frequently occurs

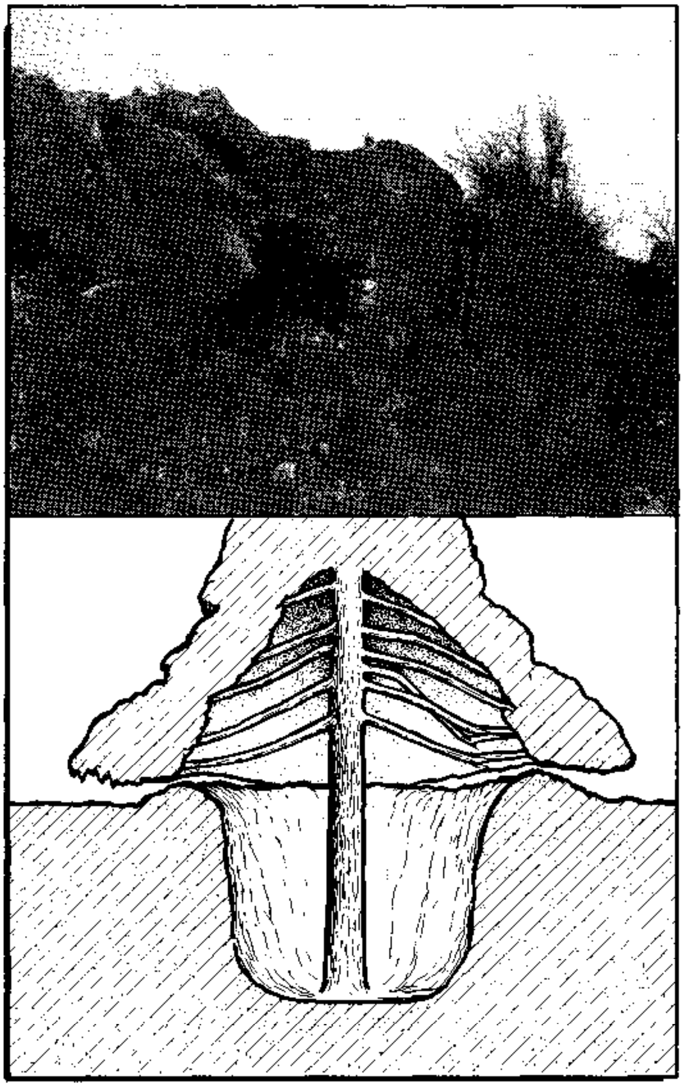

Figure 15-1. Natural Shelter.

in mountainous areas. The best fire to shelter distance is approximately 3 feet. One place where it would not be wise to build a fire is near the aircraft wreckage, especially if it is being used as a shelter. The possibility of igniting spilled lubricants or fuels is great. Survivors may decide instead to use materials from the aircraft to add to a shelter located a safe distance from the crash site.

15-4. Immediate Action Shelters. The first type of shelter that survivors may consider using, or the first type they may be forced to use, is an immediate action shelter. An immediate action shelter is one which can be erected quickly with minimum effort; for example, raft, aircraft parts, parachutes, paulin, and plastic bag. Natural formations can also shield survivors from the elements immediately, to include overhanging ledges, fallen logs, caves, and tree wells (figure 15-2). It isn't necessary to be concerned with exact shelter dimensions. Survivors should remember that if shelter is needed, use an existing shelter if at all possible. They should improvise on natural shelters or construct new shelters only if necessary. Regardless of type, the shelter must provide whatever protection is needed and, with a little ingenuity, it should be possible for survivors to protect themselves and do so quickly. In many instances, the immediate action shelters may have to serve as permanent shelters for aircrew members. For example, many aircrew members fly without parachutes, large cutting implements (axes), and entrenching tools; therefore, multiperson liferafts may be the only immediate or long-term shelter available. In this situation, multiperson liferafts must be deployed in the quickest manner possible to ensure maximum advantages are attained from the following shelter principles:

a. Set up in areas which afford maximum protection from precipitation and wind and use the basic shelter principle in paragraphs 15-2 and 15-3.

b. Anchor the raft for retention during high winds.

c. Use additional boughs, grasses, etc., for ground insulation.

15-5. Improvised Shelters. Shelters of this type should be easy to construct and (or) dismantle in a short period of time. However, these shelters usually require more time to construct then an immediate action shelter. For this reason, survivors should only consider this type of shelter when they aren't immediately concerned with getting out of the elements. Shelters of this type include the following:

a. The "A frame" design is adaptable to all environments as it can be easily modified; for example, tropical para-hammock, temperate area "A frame," arctic thermal "A frame," and fighter trench.

b. Simple shade shelter; these are useful in dry areas.

c. Various paratepees.

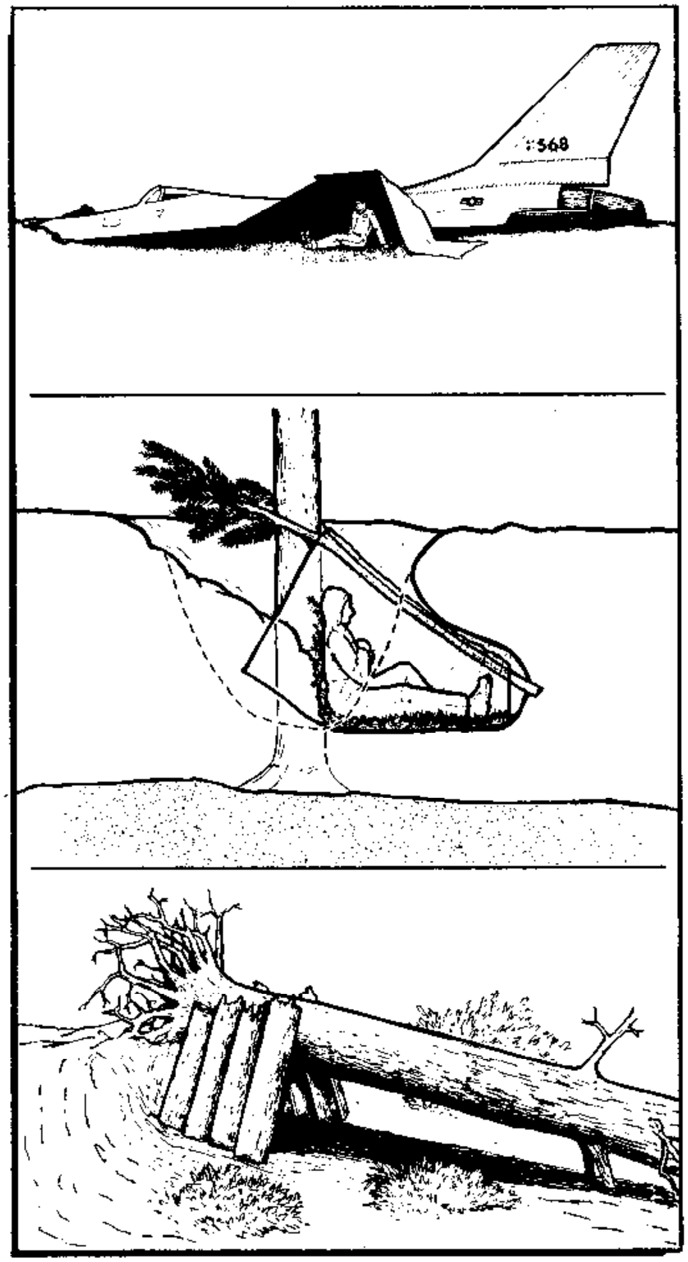

Figure 15-2. Immediate Action Shelters.

d. Snow shelters; includes tree-pit shelters.

e. All other variations of the above shelter types; sod shelters, etc.

15-6. Shelters for Warm Temperature Areas:

a. If survivors are to use parachute material, they should remember that "pitch and tightness" apply to shelters designed to shed rain or snow. Parachute material is porous and will not shed moisture unless it is stretched tightly at an angle of sufficient pitch which will encourage run-off instead of penetration. An angle of 40 to 60 degrees is recommended for the "pitch" of the shelter. The material stretched over the framework should be wrinkle-free and tight. Survivors should not touch the material when water is running over it as this will break the surface tension at that point and allow water to drip into the shelter. Two layers of parachute material, 4 to 6 inches apart, will create a more effective water repellent covering. Even during hard rain, the outer layer only lets a mist penetrate if it is pulled tight. The inner layer will then channel off any moisture which may penetrate. This layering of parachute material also creates a dead-air space that covers the shelter. This is especially beneficial in cold areas when the shelter is enclosed. Adequate insulation can also be provided by boughs, aircraft parts, snow, etc. These will be discussed in more depth in the area of cold climate shelters. A double layering of parachute material helps to trap body heat, radiating heat from the Earth's surface, and other heating sources.

b. The first step is deciding the type of shelter required. No matter which shelter is selected, the building or improvising process should be planned and orderly, following proven procedures and techniques. The second step is to select, collect, and prepare all materials needed before the actual construction; this includes framework, covering, bedding, or insulation, and implements used to secure the shelter ("dead-men," lines, stakes, etc.).

(1) For shelters that use a wooden framework, the poles or wood selected should have all the rough edges and stubs removed. Not only will this reduce the chances of the parachute fabric being ripped, but it will eliminate the chances of injury to survivors.

(2) On the outer side of a tree selected as natural shelter, some or all of the branches may be left in place as they will make a good support structure for the rest of the shelter parts.

(3) In addition to the parachute, there are many other materials which can be used as framework cover-

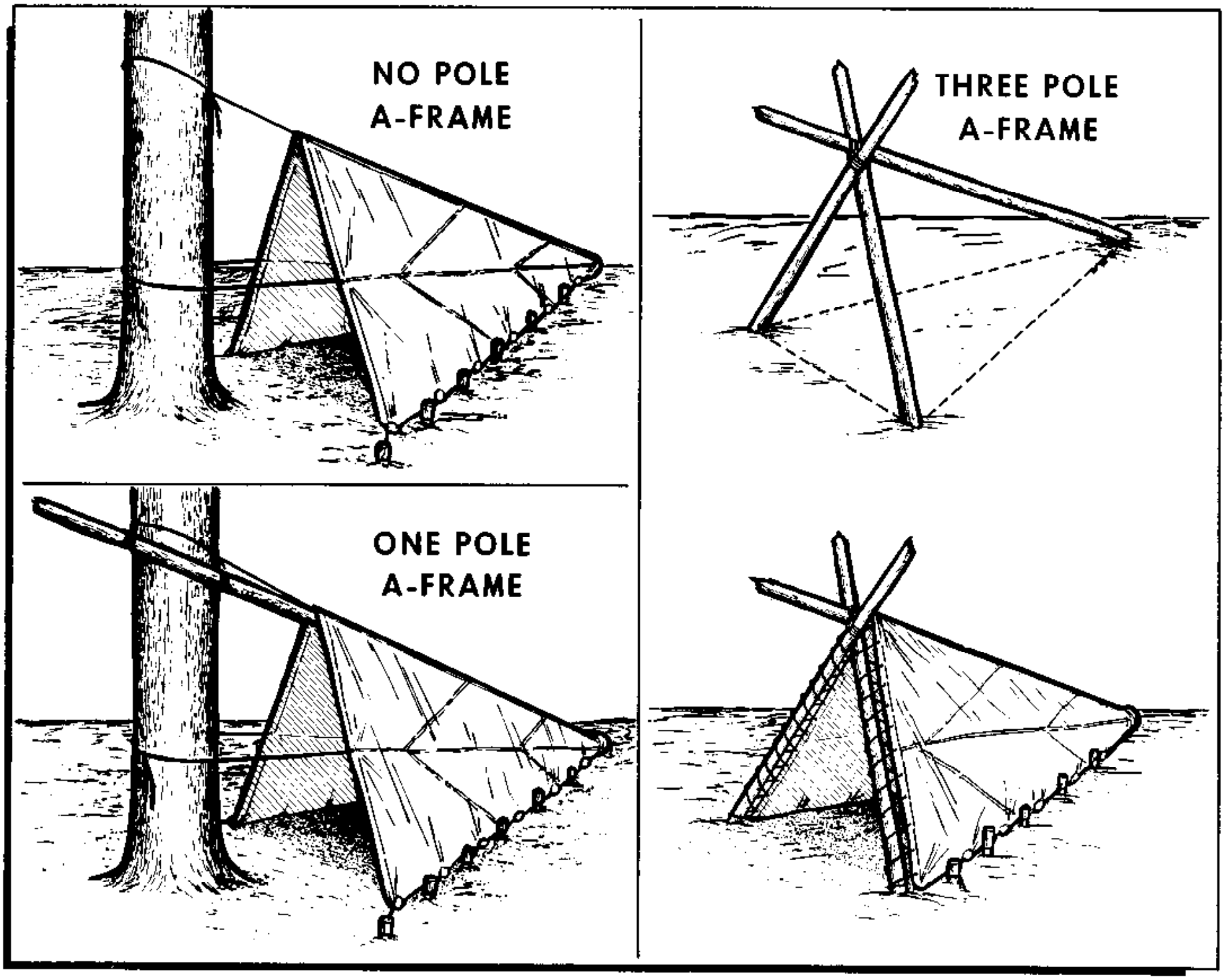

Figure 15-3. A-Frame Shelters.

ings. Some of the following are both framework and covering all in one:

(a) Bark peeled off dead trees.

(b) Boughs cut off trees.

(c) Bamboo, palm, grasses, and other vegetation cut or woven into desired patterns.

(4) If parachute material is to be used alone or in combination with natural materials, it must be changed slightly. Survivors should remove all of the lines from the parachute and then cut it to size. This will eliminate bunching and wrinkling and reduce leakage.

c. The third step in the process of shelter construction is site preparation. This includes brushing away rocks and twigs from the sleeping area and cutting back overhanging vegetation.

d. The fourth step is to actually construct the shelter, beginning with the framework. The framework is very important. It must be strong enough to support the weight of the covering and precipitation buildup of snow. It must also be sturdy enough to resist strong wind gusts.

(1) Construct the framework in one of two ways. For natural shelters, branches may be securely placed against trees or other natural objects. For parachute shelters, poles may be lashed to trees or to other poles. The support poles or branches can then be layed and (or) attached depending on their function.

(2) The pitch of the shelter is determined by the framework. A 60-degree pitch is optimum for shedding precipitation and providing shelter room.

(3) The size of the shelter is controlled by the framework. The shelter should be large enough for survivors to sit up, with adequate room to lie down and to store all personal equipment.

(4) After the basic framework has been completed, survivors can apply and secure the framework covering. The care and techniques used to apply the covering will determine the effectiveness of the shelter in shedding precipitation.

(5) When using parachute material on shelters, survivors should remove all suspension line from the material. (Excess line can be used for lashing, sewing, etc.) Next, stretch the center seam tight; then work from the back of the shelter to the front, alternating sides and securing the material to stakes or framework by using buttons and lines. When stretching the material tight, survivors should pull the material 90 degrees to the wrinkles. If material is not stretched tight, any moisture will pool in the wrinkles and leak into the shelter.

(6) If natural materials are to be used for the covering, the shingle method should be used. Starting at the bottom and working toward the top of the shelter, the bottom of each piece should overlap the top of the preceding piece. This will allow water to drain off. The material should be placed on the shelter in sufficient quantity so that survivors in the shelter cannot see through it.

15-7. Maintenance and Improvements. Once a shelter is constructed, it must be maintained. Additional modifications may make the shelter more effective and comfortable. Indian lacing (lacing the front of the shelter to the bipod) will tighten the shelter. A door may help block the wind and keep insects out. Other modifications may include a fire reflector, porch or work area, or another whole addition such as an opposing lean-to.

15-8. Construction of Specific Shelters:

a. A-Frame. The following is one way to build an A-frame shelter in a warm temperate environment using parachute material for the covering. There are as many variations of this shelter as there are builders. The procedures here will, if followed carefully, result in the completion of a safe shelter that will meet survivors' needs. For an example of this and other A-frame shelters, see figure 15-3.

(1) Materials Needed:

(a) One 12 to 18 foot long sturdy ridge pole with all projections cleaned off.

(b) Two bipod poles, approximately 7 feet long.

(c) Parachute material, normally 5 or 6 gores.

(d) Suspension lines.

(e) "Buttons," small objects placed behind gathers of material to provide a secure way of affixing suspension line to the parachute material.

(f) Approximately 14 stakes, approximately 10 inches long.

(2) Assembling the Framework:

(a) Lash (See chapter 17 — Equipment.) the two bipod poles together at eye-level height.

(b) Place the ridge pole, with the large end on the ground, into the bipod formed by the poles and secure with a square lash.

(c) The bipod structure should be 90 degrees to the ridge pole and the bipod poles should be spread out to an approximate equilateral triangle of a 60-degree pitch. A piece of line can be used to measure this.

(3) Application of Fabric:

(a) Tie off about 2 feet of the apex in a knot and tuck this under the butt end of the ridge pole. Use half hitches and clove hitches to secure the material to the base of the pole.

(b) Place the center radial seam of the parachute piece (or the center of the fabric) on the ridge pole. After pulling the material taut, use half hitches and clove hitches to secure the fabric to the front of the ridge pole.

(c) Scribe or draw a line on the ground from the butt of the ridge pole to each of bipod poles. Stake the fabric down, starting at the rear of the shelter and alternately staking from side to side to the shelter front. Use a sufficient number of stakes to ensure the parachute material is wrinkle-free.

(d) Stakes should be slanted or inclined away from the direction of pull. When tying off with a clove hitch, the line should pass in front of the stake first and then pass under itself to allow the button and line to be pulled 90 degrees to the wrinkle.

(e) Indian lacing is the sewing or lacing of the lower lateral band with inner core or line which is secured to the bipod poles. This will remove the remaining wrinkles and further tighten the material.

(f) A rain fly, bed, and other refinements can now be added.

b. Lean-To:

(1) Materials Needed:

(a) A sturdy, smooth ridge pole (longer than the builder's body) long enough to span the distance between two sturdy trees.

(b) Support poles, 10 feet long.

(c) Stakes, suspension lines, and buttons.

(d) Parachute material (minimum of four gores).

(2) Assembling the Framework:

(a) Lash the ridge pole (between two suitable trees) about chest or shoulder high.

(b) Lay the roof support poles on the ridge pole so the roof support poles and the ground are at approximately a 60-degree angle. Lash the roof support poles to the ridge pole.

(3) Application of Fabric:

(a) Place the middle seam of the fabric on the middle support pole with lower lateral band along the ridge pole.

(b) Tie-off the middle and both sides of the lower lateral band approximately 8 to 10 inches from the ridge pole.

(c) Stake the middle of the rear of the shelter first, then alternate from side to side.

(d) The stakes that go up the sides to the front should point to the front of the shelter.

(e) Pull the lower lateral band closer to the ridge pole by indian lacing.

(f) Add bed and other refinements (reflector fire, bed logs, rain fly, etc.). See figure 15-4 for lean-to examples.

c. Paratepee, 9-Pole. The paratepee is an excellent shelter for protection from wind, rain, cold, and insects. Cooking, eating, sleeping, resting, signaling, and washing can all be done without going outdoors. The paratepee, whether 9-pole, 1-pole, or no-pole, is the only improvised shelter that provides adequate ventilation to build an inside fire. With a small fire inside, the shelter also serves as a signal at night.

(1) Materials Needed:

(a) Suspension line.

(b) Parachute material, normally 14 gores are suitable.

-1. Spread out the 14-gore section of parachute and cut off all lines at the lower lateral band, leaving about 18 inches of line attached. All other suspension lines should be stripped from the parachute.

-2. Sew two smoke flaps, made from two large panels of parachute material, at the apex of the 14-gore section on the outside seams. Attach suspension line with a bowline in the end to each smoke flap. The ends of the smoke flap poles will be inserted in these (see figure 15-5).

(c) Stakes.

(d) Although any number of poles may be used, 11 poles, smoothed off, each about 20 feet long, will normally provide adequate support.

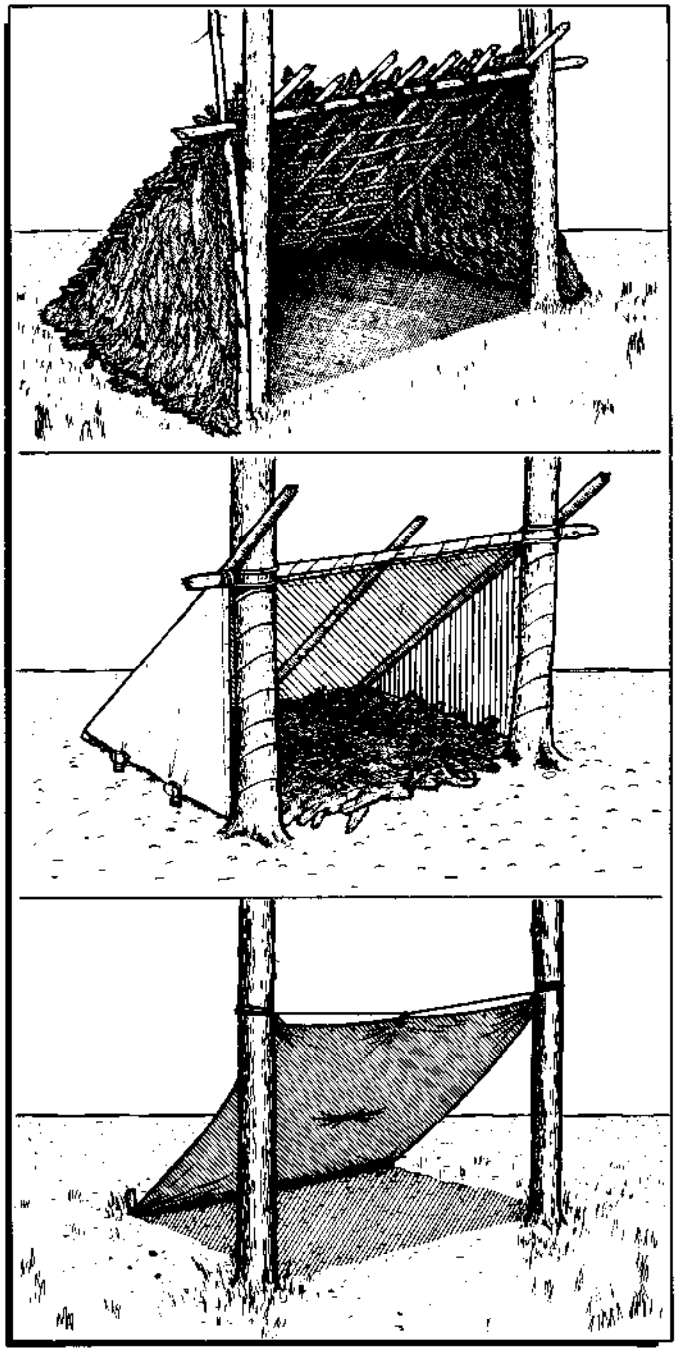

Figure 15-4. Lean-To Shelters.

(2) Assembling the Framework. (Assume 11 poles are used. Adjust instructions if different numbers are used.)

(a) Lay three poles on the ground with the butts even. Stretch the canopy along the poles. The lower lateral band should be 4 to 6 inches from the bottoms of the poles before the stretching takes place. Mark one of the poles at the apex point.

(b) Lash the three poles together, 5 to 10 inches above the marked area. (A shear lash is effective for this purpose.) These poles will form the tripod (figure 15-5).

(c) Scribe a circle approximately 12 feet in diameter in the shelter area and set the tripod so the butts of the poles are evenly spaced on the circle. Five of the remaining eight poles should be placed so the butts are evenly spaced around the 12-foot circle and the tops are laid in the apex of the tripod to form the smallest apex possible (figure 15-5).

(3) Application of Fabric:

(a) Stretch the parachute material along the tie pole. Using the suspension line attached to the middle radial seam, tie the lower lateral band to the tie pole 6 inches from the butt end. Stretch the parachute material along the middle radial seam and tie it to the tie pole using the suspension line at the apex. Lay the tie pole onto the shelter frame with the butt along the 12-foot circle and the top in the apex formed by the other poles. The tie pole should be placed directly opposite the proposed door.

(b) Move the canopy material (both sides of it) from the tie pole around the framework and tie the lower lateral band together and stake it at the door. The front can now be sewn or pegged closed, leaving 3 to 4 feet for a door. (A sewing "ladder" can be made by lashing steps up the front of the tepee (figure 15-5).

(c) Enter the shelter and move the butts of the poles outward to form a more perfect circle and until the fabric is relatively tight and smooth.

(d) Tighten the fabric and remove remaining wrinkles. Start staking directly opposite the door, and alternate from side to side, pulling the material down and to the front of the shelter. Use clove hitches or similar knots to secure material to the stakes.

(e) Insert the final two poles into the loops on the smoke flaps. The paratepee is now finished (figure 15-5).

(f) One improvement which could be made to the paratepee is the installation of a liner. This will allow a draft for a fire without making the occupants cold, since there may be a slight gap between the lower lateral band and the ground. A liner can be affixed to the inside of the paratepee by taking the remaining 14-gore piece of material and firmly staking the lower lateral band directly to the ground all the way around, leaving room for the door. The area where the liner and door meet may be sewn up. The rest of the material is brought up the inside walls and affixed to the poles with buttons (figure 15-5).

d. Paratepee, 1-Pole:

(1) Materials Needed:

(a) Normally use a 14-gore section of canopy, strip the shroud lines leaving 16- to 18-inch lengths at the lower lateral band.

(b) Stakes.

(c) Inner core and needle.

(2) Construction of the 1-Pole Paratepee:

(a) Select a shelter site and scribe a circle about 14 feet in diameter on the ground.

(b) The parachute material is staked to the ground using the lines attached at the lower lateral band. After deciding where the shelter door will be located, stake the first line (from the lower band) down securely. Proceed around the scribed line and stake down all the lines from the lateral band, making sure the parachute material is stretched taut before the line is staked down.

(c) Once all the lines are staked down, loosely attach the center pole, and, through trial and error, determine the point at which the parachute material will be pulled tight once the center pole is placed upright—securely attach the material at this point.

(d) Using a suspension line (or innercore), sew the end gores together leaving 3 or 4 feet for a door (figure 15-6).

e. Paratepee, No-Pole. For this shelter, the 14 gores of material are prepared the same way. A line is attached to the apex and thrown over a tree limb, etc., and tied off. The lower lateral band is then staked down starting opposite the door around a 12- to 14-foot circle. (See figure 15-7 for paratepee example.)

f. Sod Shelter. A framework covered with sod provides a shelter which is warm in cold weather and one that is easily made waterproof and insect-proof in the summer. The framework for a sod shelter must be strong, and it can be made of driftwood, poles, willow, etc. (Some natives use whale bones.) Sod, with a heavy growth of grass or weeds, should be used since the roots tend to hold the soil together. Cutting about 2 inches of soil along with the grass is sufficient. The size of the blocks are determined by the strength of the individual. A sod house is strong and fireproof.

15-9. Shelter for Tropical Areas. Basic considerations for shelter in tropical areas are as follows:

a. In tropical areas, especially moist tropical areas, the major environmental factors influencing both site selection and shelter types are:

(1) Moisture and dampness.

(2) Rain.

(3) Wet ground.

(4) Heat.

(5) Mud-slide areas.

(6) Dead standing trees and limbs.

(7) Insects.

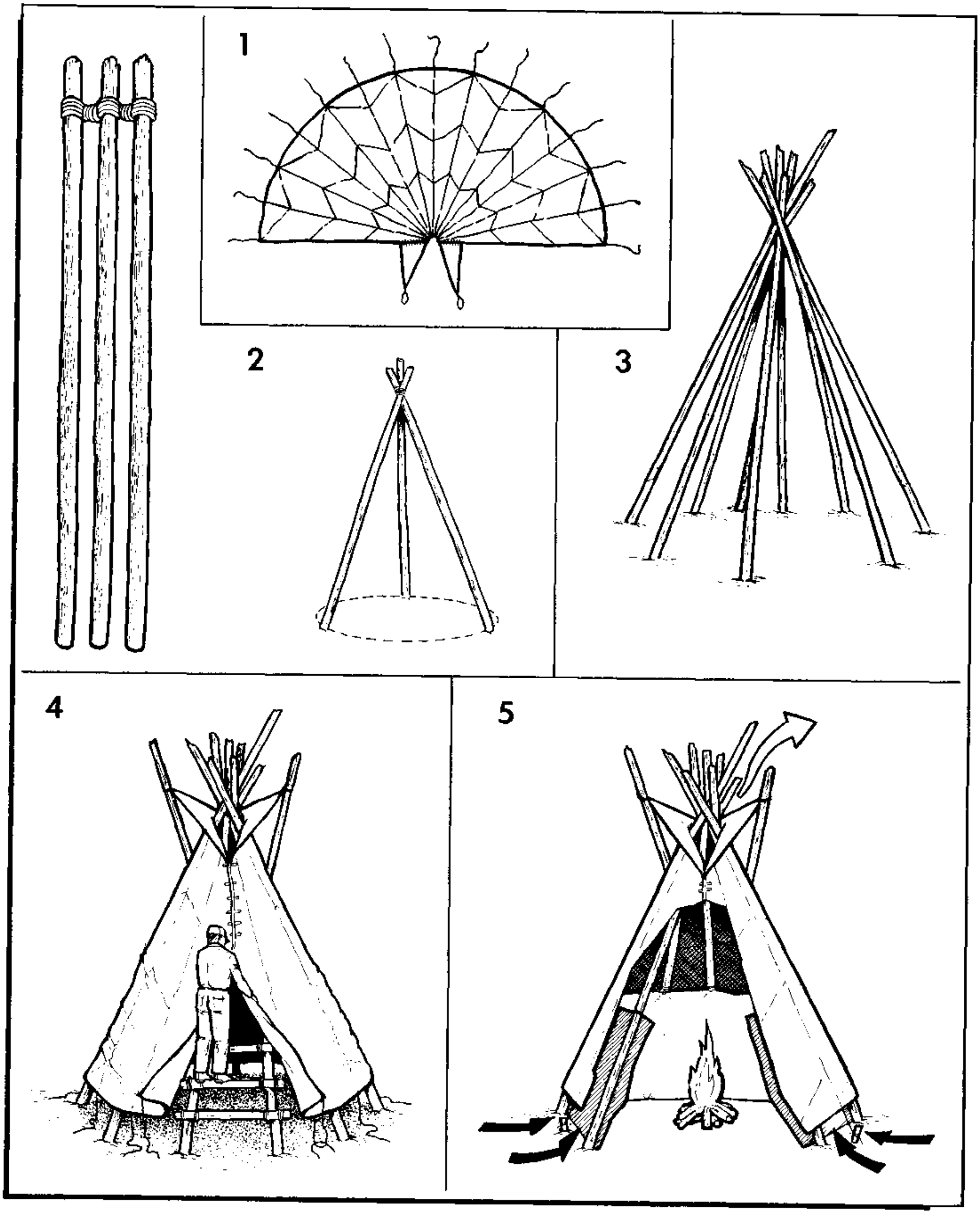

Figure 15-5. 9-Pole Tepee.

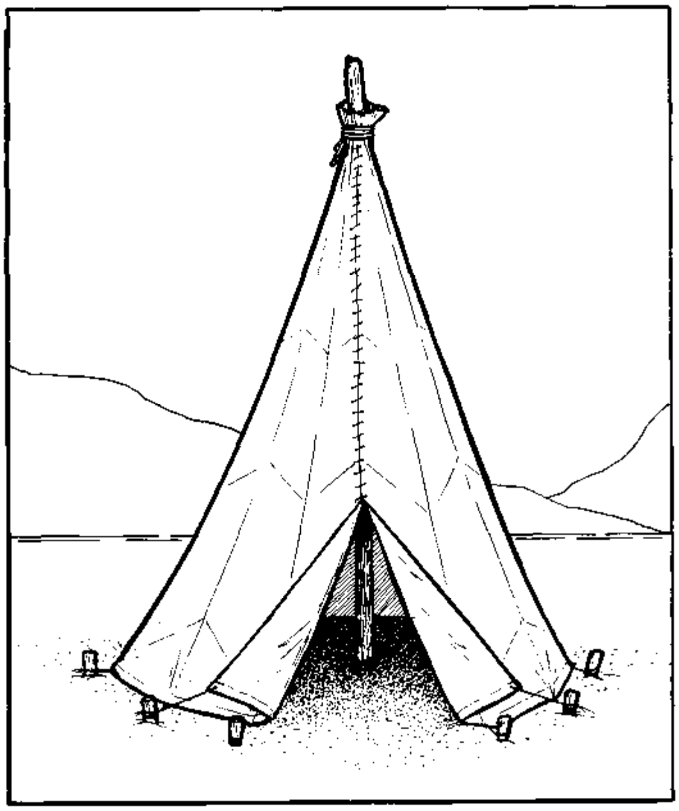

Figure 15-6. 1-Pole Tepee.

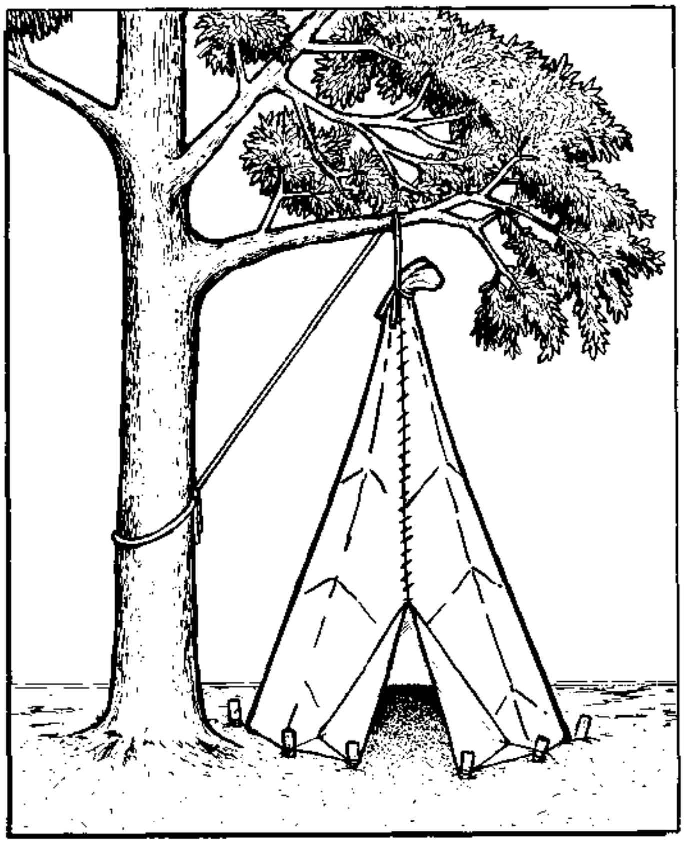

Figure 15-7. No-Pole Tepee.

b. Survivors should establish a campsite on a knoll or high spot in an open area well back from any swamps or marshy areas. The ground in these areas is drier, and there may be a breeze which will result in fewer insects.

c. Underbrush and dead vegetation should be cleared from the shelter site. Crawling insects will not be able to approach survivors as easily due to lack of cover.

d. A thick bamboo clump or matted canopy of vines for cover reflects the smoke from the campfire and discourages insects. This cover will also keep the extremely heavy early morning dew off the bedding.

e. The easiest improvised shelter is made by draping a parachute, tarpaulin, or poncho over a rope or vine stretched between two trees. One end of the canopy should be kept higher than the other; insects are discouraged by few openings in shelters and smudge fires. A hammock made from parachute material will keep the survivor off the ground and discourage ants, spiders, leeches, scorpions, and other pests.

f. In the wet jungle, survivors need shelter from dampness. If they stay with the aircraft, it should be used for shelter. They should try to make it mosquito-proof by covering openings with netting or parachute cloth.

g. A good rain shelter can be made by constructing an A-type framework and shingling it with a good thickness of palm or other broad leaf plants, pieces of bark, and mats of grass (figure 15-8).

h. Nights are cold in some mountainous tropical areas. Survivors should try to stay out of the wind and build a fire. Reflecting the heat off a rock pile or other barrier is a good idea. Some natural materials which can

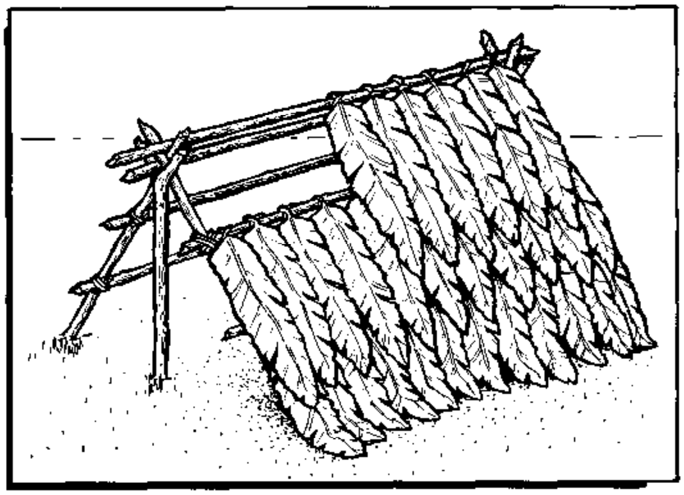

Figure 15-8. Banana Leaf A-Frame.

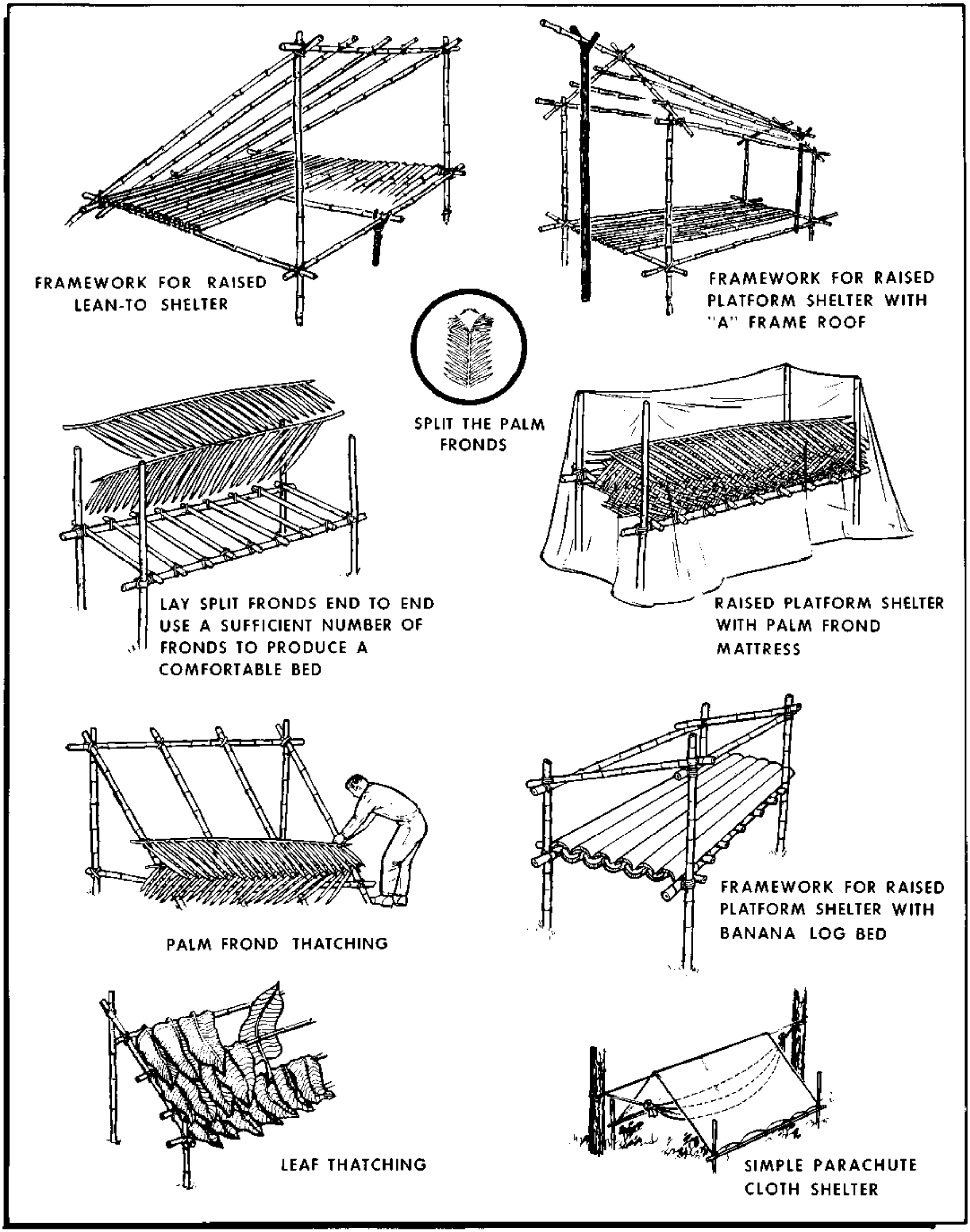
FRAMEWORK FOR RAISED LEAN-TO SHELTER
FRAMEWORK FOR RAISED PLATFORM SHELTER WITH "A" FRAME ROOF
SPLIT THE PALM FRONDS
LAY SPLIT FRONDS END TO END USE A SUFFICIENT NUMBER OF FRONDS TO PRODUCE A COMFORTABLE BED
RAISED PLATFORM SHELTER WITH PALM FROND MATTRESS
PALM FROND THATCHING
FRAMEWORK FOR RAISED PLATFORM SHELTER WITH BANANA LOG BED
LEAF THATCHING
SIMPLE PARACHUTE CLOTH SHELTER

Figure 15-9. Raised Platform shelter.

be used in the shelters are green wood (dead wood may be too rotten), bamboo, and palm leaves. Vines can be used in place of suspension line for thatching roofs or floors, etc. Banana plant sections can be separated from the banana plant and fashioned to provide a mattress effect.

15-10. Specific Shelters for Tropical Environments:

a. Raised Platform Shelter (figure 15-9). This shelter has many variations. One example is four trees or vertical poles in a rectangular pattern which is a little longer and a little wider than the survivor, keeping in mind the survivor will also need protection for equipment. Two long, sturdy poles are then square lashed between the trees or vertical poles, one on each side of the intended shelter. Cross pieces can then be secured across the two horizontal poles at 6- to 12-inch intervals. This forms the platform on which a natural mattress may be constructed. Parachute material can be used as an insect net and a roof can be built over the structure using A-frame building techniques. The roof should be waterproofed with thatching laid bottom to top in a thick shingle fashion. See figure 15-9 for examples of this and other platform shelters. These shelters can also be built using three trees in a triangular pattern. At the foot of the shelter, two poles are joined to one tree.

b. Variation of Platform Shelter. A variation of the platform-type shelter is the paraplatform. A quick and comfortable bed is made by simply wrapping material around the two "frame" poles. Another method is to roll poles in the material in the same manner as for an improvised stretcher (figure 15-10).

c. Hammocks. Various parahammocks can also be made. They are more involved than a simple parachute wrapped framework and not quite as comfortable (figure 15-11).

d. Hobo Shelter. On tropical coasts and other coastal environments, if a more permanent shelter is desired as opposed to a simple shade shelter, survivors should build a "hobo" shelter. To build this shelter:

(1) Dig into the lee side of a sand dune to protect the shelter from the wind. Clear a level area large enough to lie down in and store equipment.

(2) After the area has been cleared, build a heavy driftwood framework which will support the sand.

(3) Wall sides and top with strong material (boards, driftwood, etc.) that will support the sand; leave a door opening.

(4) Slope the roof to equal the slope of the sand dune. Cover the entire shelter with parachute material to keep sand from sifting through small holes in the walls and roof.

(5) Cover with 6 to 12 inches of sand to provide protection from wind and moisture.

(6) Construct a door for the shelter (figure 15-12).

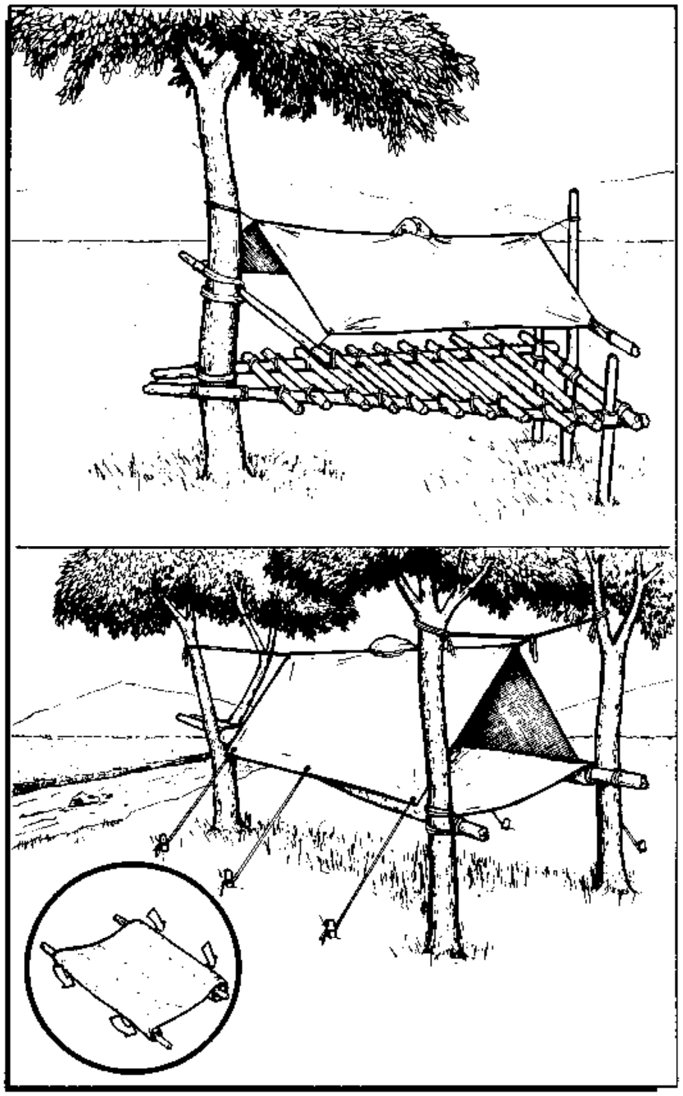

Figure 15-10. Raised Paraplatform Shelter.

15-11. Shelters for Dry Climates:

a. Natives of hot, dry areas make use of light-proof shelters with sides rolled up to take advantage of any breeze. Survivors should emulate these shade-type shelters if forced to survive in these areas. The extremes of heat *and* cold must be considered in hot areas, as most can become very cold during the night. The major problem for survivors will be escaping the heat and Sun rays.

b. Natural shelters in these areas are often limited to the shade of cliffs and the lee sides of hills, dunes, or rock formations. In some desert mountains, it is possible to find good rock shelters or cave-like protection under tumbled blocks of rocks which have fallen from cliffs. Use care to ensure that these blocks are in areas void of future rock falling activity and free from animal hazards.

c. Vegetation, if any exists, is usually stunted and armed with thorns. It may be possible to stay in the shade by moving around the vegetation as the Sun moves. The hottest part of the day may offer few shadows because the Sun is directly overhead. Parachute material draped over bushes or rocks will provide some shade.

d. Materials which can be used in the construction of desert shelters include:

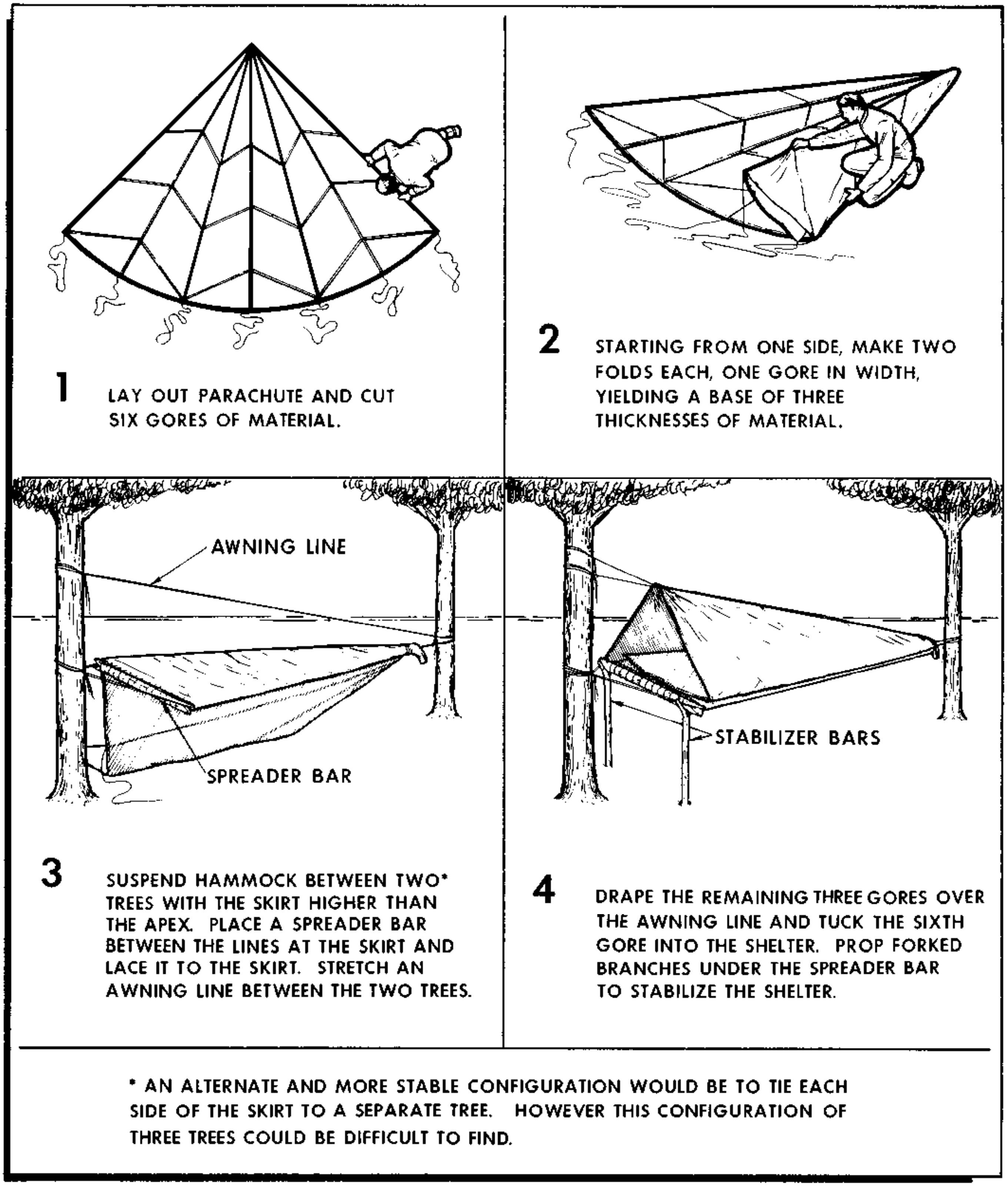

1 LAY OUT PARACHUTE AND CUT SIX GORES OF MATERIAL.

2 STARTING FROM ONE SIDE, MAKE TWO FOLDS EACH, ONE GORE IN WIDTH, YIELDING A BASE OF THREE THICKNESSES OF MATERIAL.

3 SUSPEND HAMMOCK BETWEEN TWO* TREES WITH THE SKIRT HIGHER THAN THE APEX. PLACE A SPREADER BAR BETWEEN THE LINES AT THE SKIRT AND LACE IT TO THE SKIRT. STRETCH AN AWNING LINE BETWEEN THE TWO TREES.

4 DRAPE THE REMAINING THREE GORES OVER THE AWNING LINE AND TUCK THE SIXTH GORE INTO THE SHELTER. PROP FORKED BRANCHES UNDER THE SPREADER BAR TO STABILIZE THE SHELTER.

* AN ALTERNATE AND MORE STABLE CONFIGURATION WOULD BE TO TIE EACH SIDE OF THE SKIRT TO A SEPARATE TREE. HOWEVER THIS CONFIGURATION OF THREE TREES COULD BE DIFFICULT TO FIND.

Figure 15-11. Parahammock.

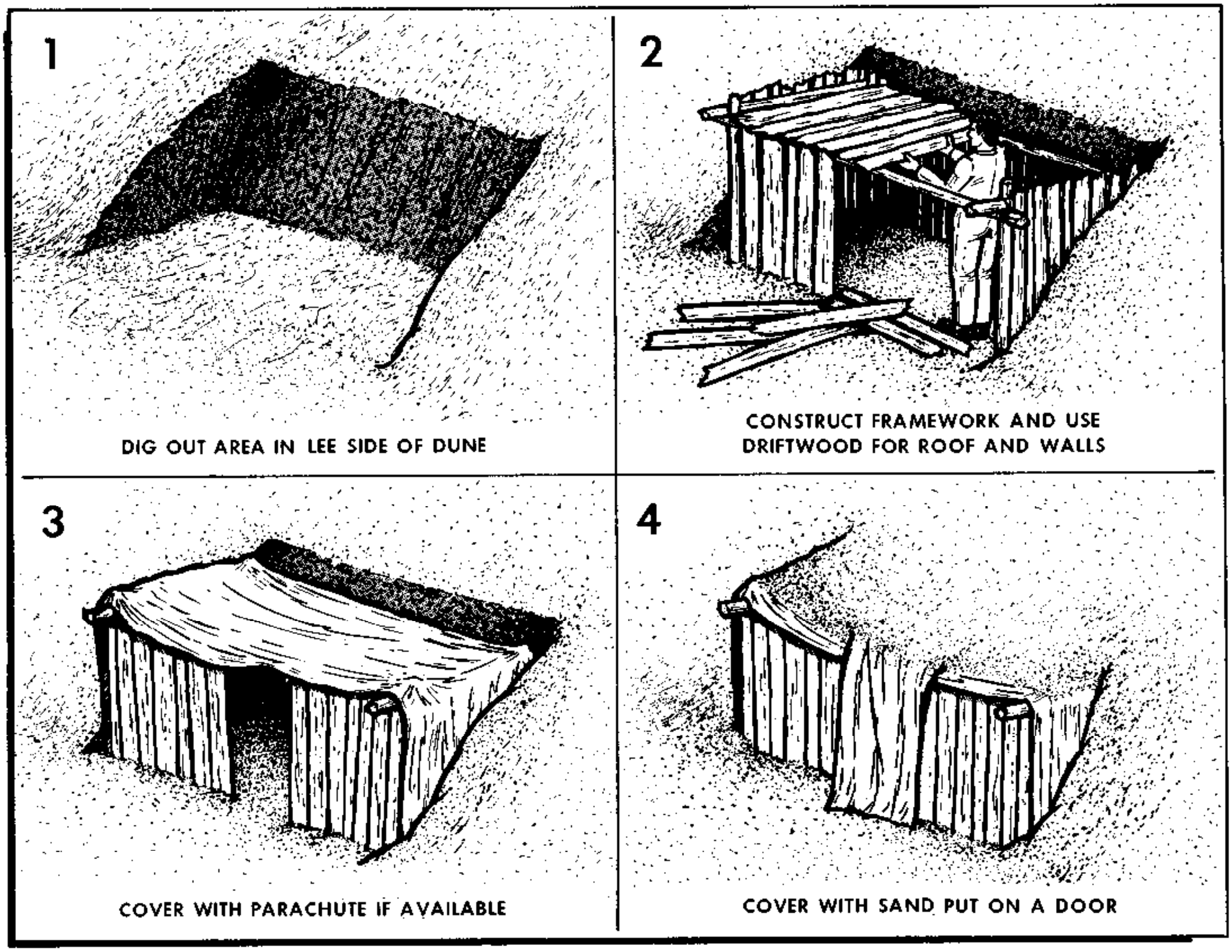
1
DIG OUT AREA IN LEE SIDE OF DUNE
2
CONSTRUCT FRAMEWORK AND USE DRIFTWOOD FOR ROOF AND WALLS
3
COVER WITH PARACHUTE IF AVAILABLE
4
COVER WITH SAND PUT ON A DOOR

Figure 15-12. Hobo Shelter.

(1) Sand, though difficult to work with when loose, may be made into pillars by using sandbags made from parachute or any available cloth.

(2) Rock can be used in shelter construction.

(3) Vegetation such as sage brush, creosote bushes, juniper trees, and desert gourd vines are valuable building materials.

(4) Parachute canopy and suspension lines. These are perhaps the most versatile building materials available for use by survivors. When used in layers, parachute material protects survivors from the Sun's rays.

(a) The shelter should be made of dense material or have numerous layers to reduce or stop dangerous ultraviolet rays. The colors of the parachute materials used make a difference as to how much protection is provided from ultraviolet radiation. As a general rule, the order of preference should be to use as many layers as practical in the order of orange, green, tan, and white.

ULTRAVIOLET TESTS ON PARACHUTE CANOPY MATERIAL

% Ultraviolet (Short Wave 2537 A° Sunburn Rays) Blocked as compared to Direct Exposure

	1 Layer	2 Layers	3 Layers
Orange	78.2%	96.2%	99.36%
Sage Green	79.5%	96.2%	98.7%
Tan	64.1%	84.6%	93.6%
White	47.5%	61.6%	70.5%

% Ultraviolet (Long Wave 3660 A°) Blocked as Compared to Direct Exposure

	1 Layer	2 Layers	3 Layers
Orange	63.4%	92.3%	97.8%
Sage Green	60.0%	88.9%	97.8%
Tan	38.9%	66.7%	82.3%
White	28.9%	47.8%	58.9%

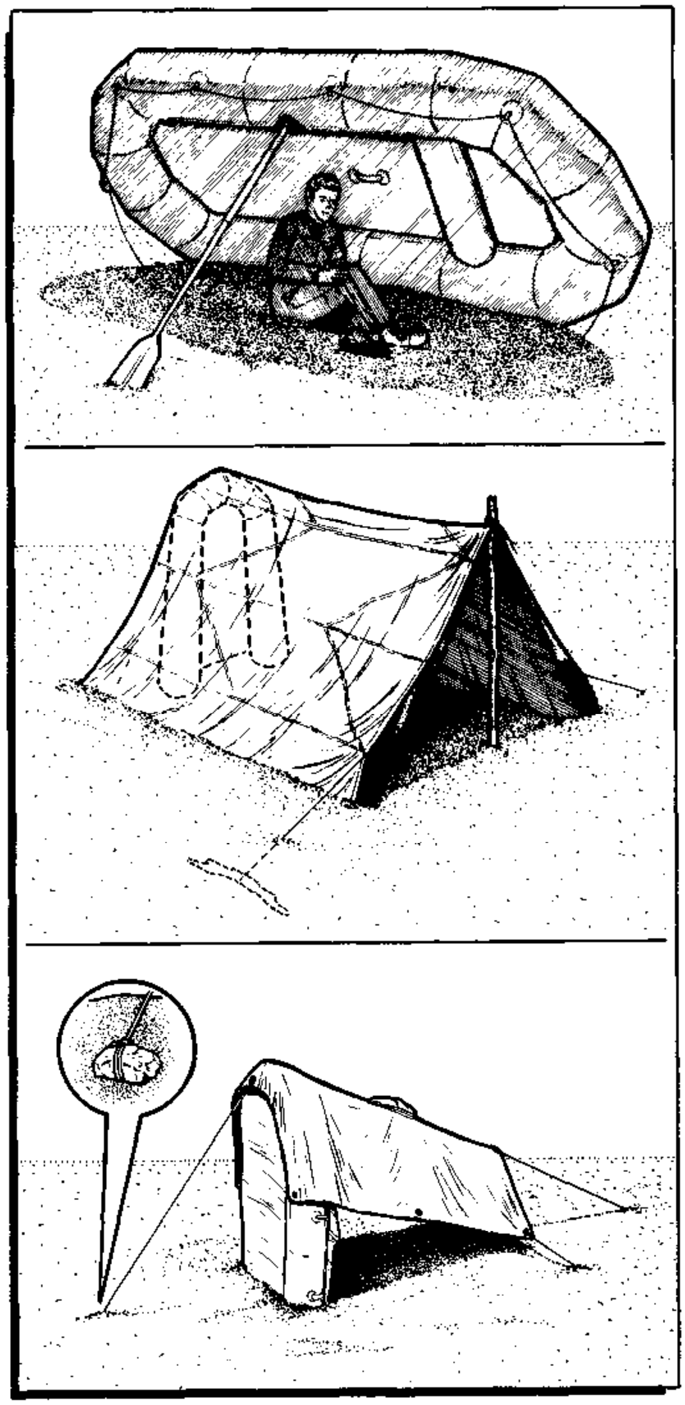

Figure 15-13. Improvised Natural Shade Shelters.

(b) The material should be kept approximately 12 to 18 inches above the individual. This allows the air to cool the underside of the material.

(c) Aircraft parts and liferafts can also be used for shade shelters. Survivors may use sections of the wing, tail, or fuselage to provide shade. However, the interior of the aircraft will quickly become superheated and should be avoided as a shelter. An inflatable raft can be tilted against a raft paddle or natural object such as a bush or rock to provide relief from the Sun (figure 15-13).

15-12. Principles of Desert Shelters:

a. The roof of a desert shelter should be multilayered so the resulting airspace reduces the inside temperature of the shelter. The layers should be separated 12 to 18 inches apart (figure 15-14).

b. Survivors should place the floor of the shelter about 18 inches above or below the desert surface to increase the cooling effect.

c. In warmer deserts, white parachute material should be used as an outer layer. Orange or sage green material should be used as an inner layer for protection from ultraviolet rays.

d. In cooler areas, multiple layers of parachute material should be used with sage green or orange material as the outer layer to absorb heat.

e. The sides of shelters should be movable in order to protect survivors during cold and (or) windy periods and to allow for ventilation during hot periods.

f. In a hot desert, shelters should be built away from large rocks which store heat during the day. Survivors may need to move to the rocky areas during the evening to take advantage of the warmth heated rocks radiate.

g. Survivors should:

(1) Build shelters on the windward sides of dunes for cooling breezes.

(2) Build shelters during early morning, late evening, or at night. However, potential survivors should recall that survivors who come down in a desert area during daylight hours must be immediately concerned with protection from the Sun and loss of water. In this case, parachute canopy material can be draped over liferaft, vegetation, or a natural terrain feature for quick shelter.

15-13. Shelters for Snow and Ice Areas:

a. The differences in arctic and arctic-like environments create the need for different shelters. Basically, there are two types of environments which may require special shelter characteristics or building principles before survivors will have adequate shelter. They are:

(1) Barren lands which include some seacoasts, icecaps, sea ice areas, and areas above the tree line.

(2) Tree-line areas.

b. Barren lands offer a limited variety of materials for shelter construction. These are snow, small shrubs, and grasses. Ridges formed by drifting or wind-packed snow may be used for wind protection (survivors should build on the lee side). In some areas, such as sea ice, windy conditions usually exist and cause the ice to shift forming pressure ridges. These areas of unstable ice and snow should be avoided at all times. Shelters which are suitable for barren-type areas include:

(1) Molded dome (figure 15-15).

(2) Snow cave (figure 15-16).

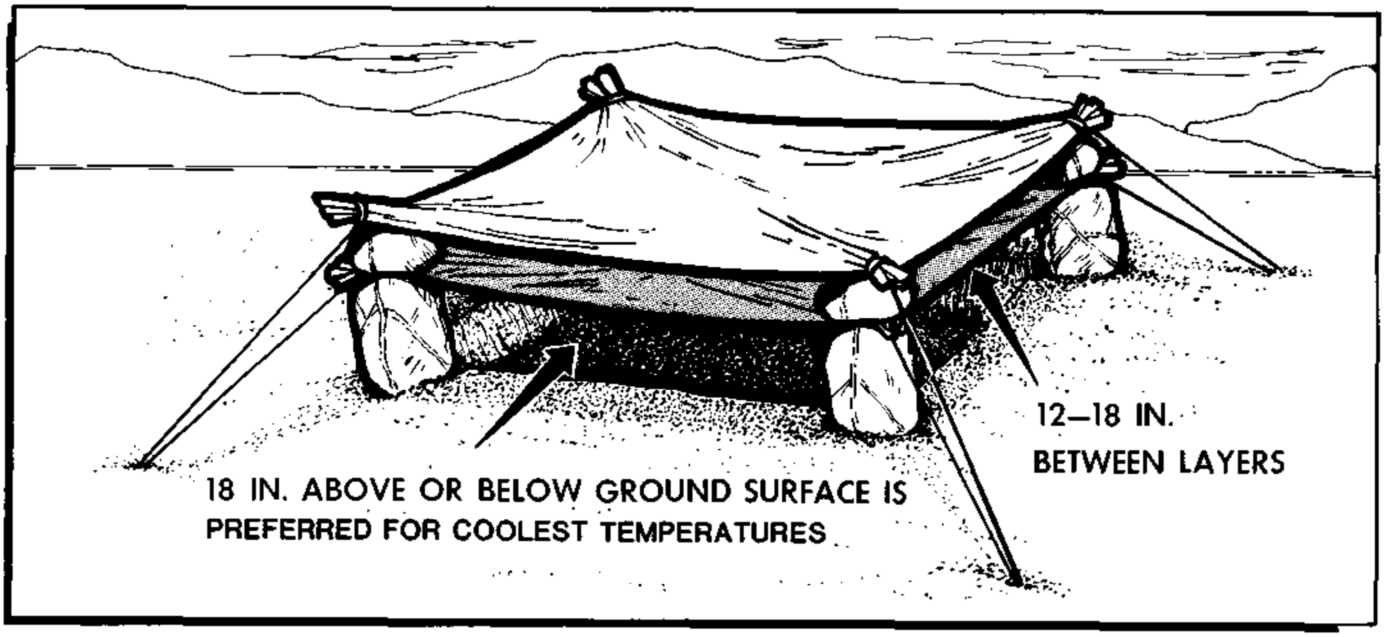
12—18 IN.
BETWEEN LAYERS
18 IN. ABOVE OR BELOW GROUND SURFACE IS
PREFERRED FOR COOLEST TEMPERATURES

Figure 15-14. Parachute Shade Shelter.

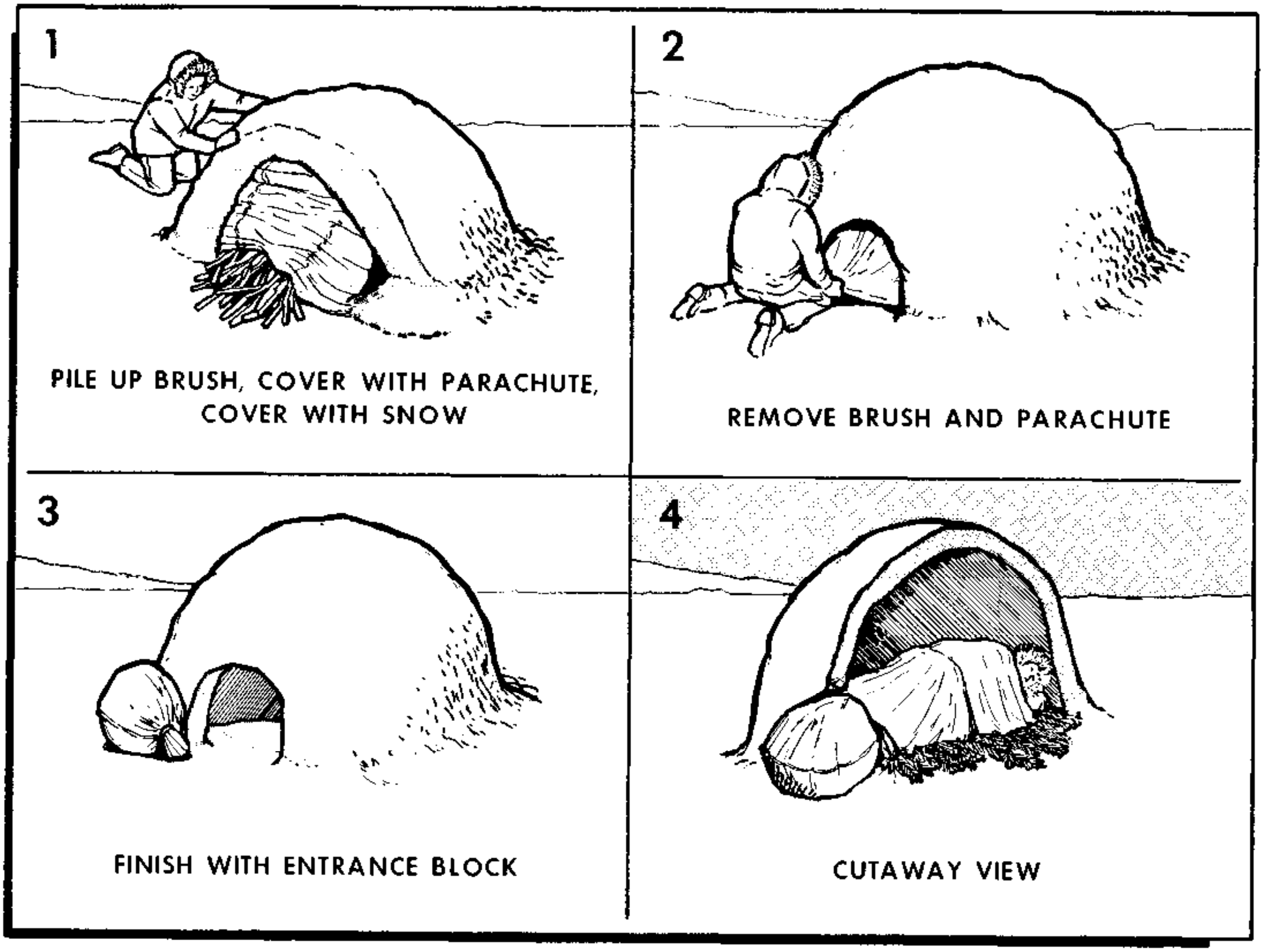
1
PILE UP BRUSH, COVER WITH PARACHUTE,
COVER WITH SNOW
2
REMOVE BRUSH AND PARACHUTE
3
FINISH WITH ENTRANCE BLOCK
4
CUTAWAY VIEW

Figure 15-15. Molded Dome Shelter.

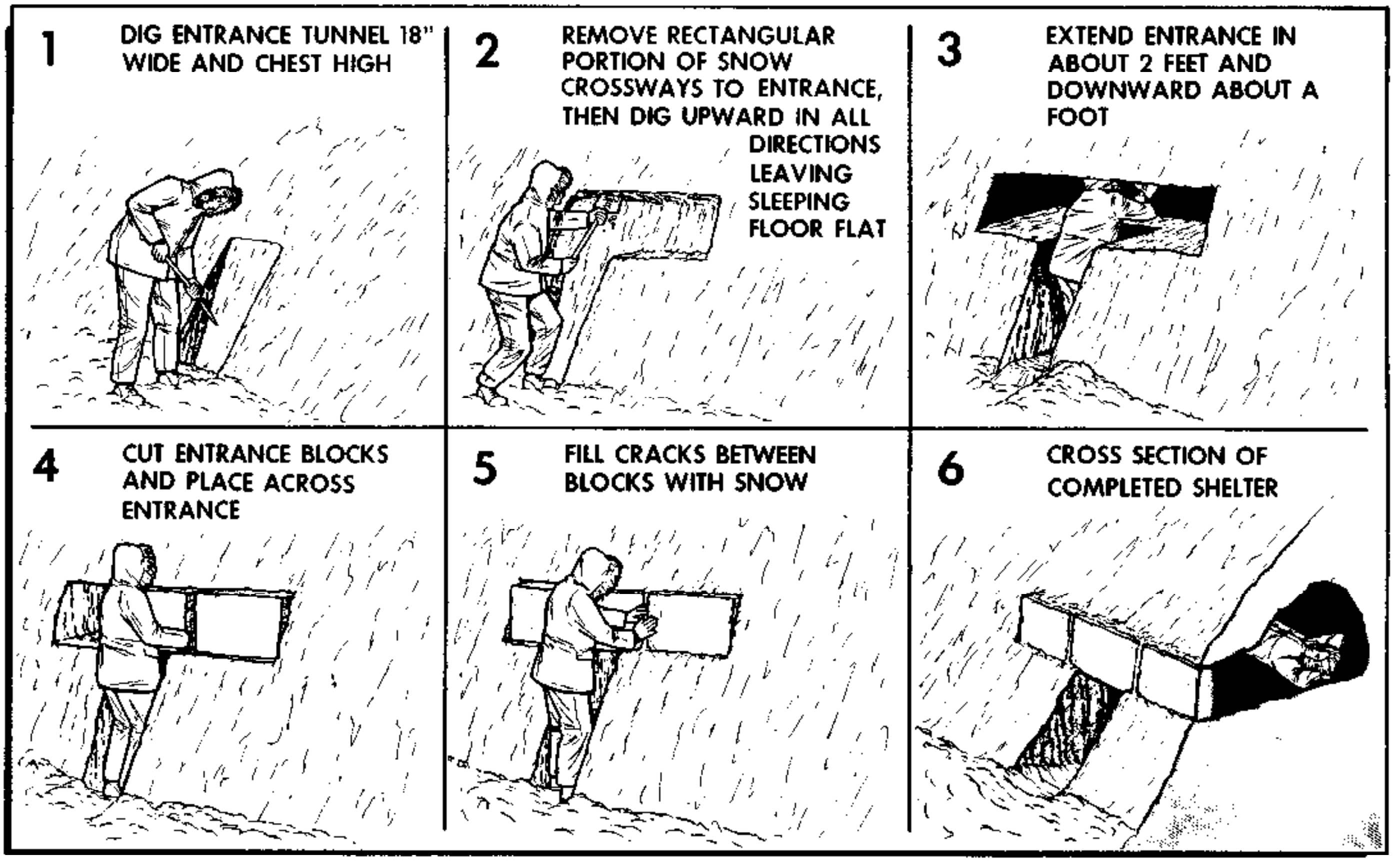
1
DIG ENTRANCE TUNNEL 18" WIDE AND CHEST HIGH
2
REMOVE RECTANGULAR PORTION OF SNOW CROSSWAYS TO ENTRANCE, THEN DIG UPWARD IN ALL DIRECTIONS LEAVING SLEEPING FLOOR FLAT
3
EXTEND ENTRANCE IN ABOUT 2 FEET AND DOWNWARD ABOUT A FOOT
4
CUT ENTRANCE BLOCKS AND PLACE ACROSS ENTRANCE
5
FILL CRACKS BETWEEN BLOCKS WITH SNOW
6
CROSS SECTION OF COMPLETED SHELTER

Figure 15-16. Snow Cave.

(3) Fighter trench (figure 15-17).

(4) Igloo (figure 15-18).

(5) Para-snow house (Figure 15-19).

NOTE: Of these, the ones that are quick to construct and require minimum effort and energy are the molded dome, snow cave, and fighter trench. It is important to know which of these shelters is the easiest to build since reducing or eliminating the effect of the windchill factor is essential to remaining alive.

c. In tree-covered areas, sufficient natural shelter building materials are normally available. Caution is required. Shelters built near rivers and streams may get caught in the overflow.

d. Tree-line area shelter types include:

(1) Thermal A-Frame construction (figure 15-20).

(2) Lean-to or wedge (figure 15-21).

(3) Double lean-to (figure 15-22).

(4) Fan (figure 15-23).

(5) Willow frame (figure 15-24).

(6) Tree well (figure 15-25).

e. Regardless of the type of shelter used, the use of thermal principles and insulation in arctic shelters is required. Heat radiates from bare ground and from ice masses over water. This means that shelter areas on land should be dug down to bare earth if possible (figure 15-26). A minimum of 8 inches of insulation above survivors is needed to retain heat. All openings except ventilation holes should be sealed to avoid heat loss. Leaving vent holes open is especially important if heat producing devices are used. Candles, sterno, or small oil lamps produce carbon monoxide. In addition to the ventilation hole through the roof, another may be required at the door to ensure adequate circulation of the air. (As a general rule, unless persons can see their breath, the snow shelter is too warm and should be cooled down to proclude melting and dripping.)

f. Regardless of how cold it may get outside, the temperature inside a small well-constructed snow cave will probably not be lower than -10°F. Body heat alone can raise the temperature of a snow cave 45 degrees above the outside air. A burning candle will raise the temperature 4 degrees. Burning Sterno (small size, 2⅝ oz) will raise the cave temperature about 28 degrees. However, since they cannot be heated many degrees above freezing, snow shelters provide a rather rugged life. Once the inside of the shelter "glazes" over with ice, this layer of ice should be removed by chipping it off or a new shelter built since ice reduces the insulating quality of a shelter. Maintain the old shelter until the new one is constructed. It will provide protection from the wind.

g. The aircraft should not be used as a shelter when temperatures are below freezing except in high wind conditions. Even then a thermal shelter should be constructed as soon as the conditions improve. The aircraft will not provide adequate insulation, and the floor will usually become icy and hazardous.

15-14. General Construction Techniques:

a. All thermal shelters use a layering system consisting of the frame, parachute (if available), boughs or shrubs, and snow. The framework must be sturdy enough to support the cover and insulation. A door block should be used to minimize heat loss. Insulation should be added on sleeping areas.

b. If a barren land-type shelter is being built with snow as the only material, a long knife or digging tool is a necessity. It normally takes 2 to 3 hours of hard work to dig a snow cave, and much longer for the novice to build an igloo.

c. Survivors should dress lightly while digging and working; they can easily become overheated and dampen their clothing with perspiration which will rapidly turn to ice.

d. If possible, all shelter types should have their openings 90 degrees to the prevailing wind. The entrance to the shelter should also be screened with snowblocks stacked in a L-shape.

e. Snow on the sea ice, suitable for cutting into blocks, will usually be found in the lee of pressure ridges or ice hummocks. The packed snow is often so shallow that the snowblocks have to be cut out horizontally.

f. No matter which shelter is used, survivors should take a digging tool into the shelter at night to cope with the great amount of snow which may block the door during the night.

15-15. Shelter Living:

a. Survivors should limit the number of shelter entrances to conserve heat. Fuel is generally scarce in the

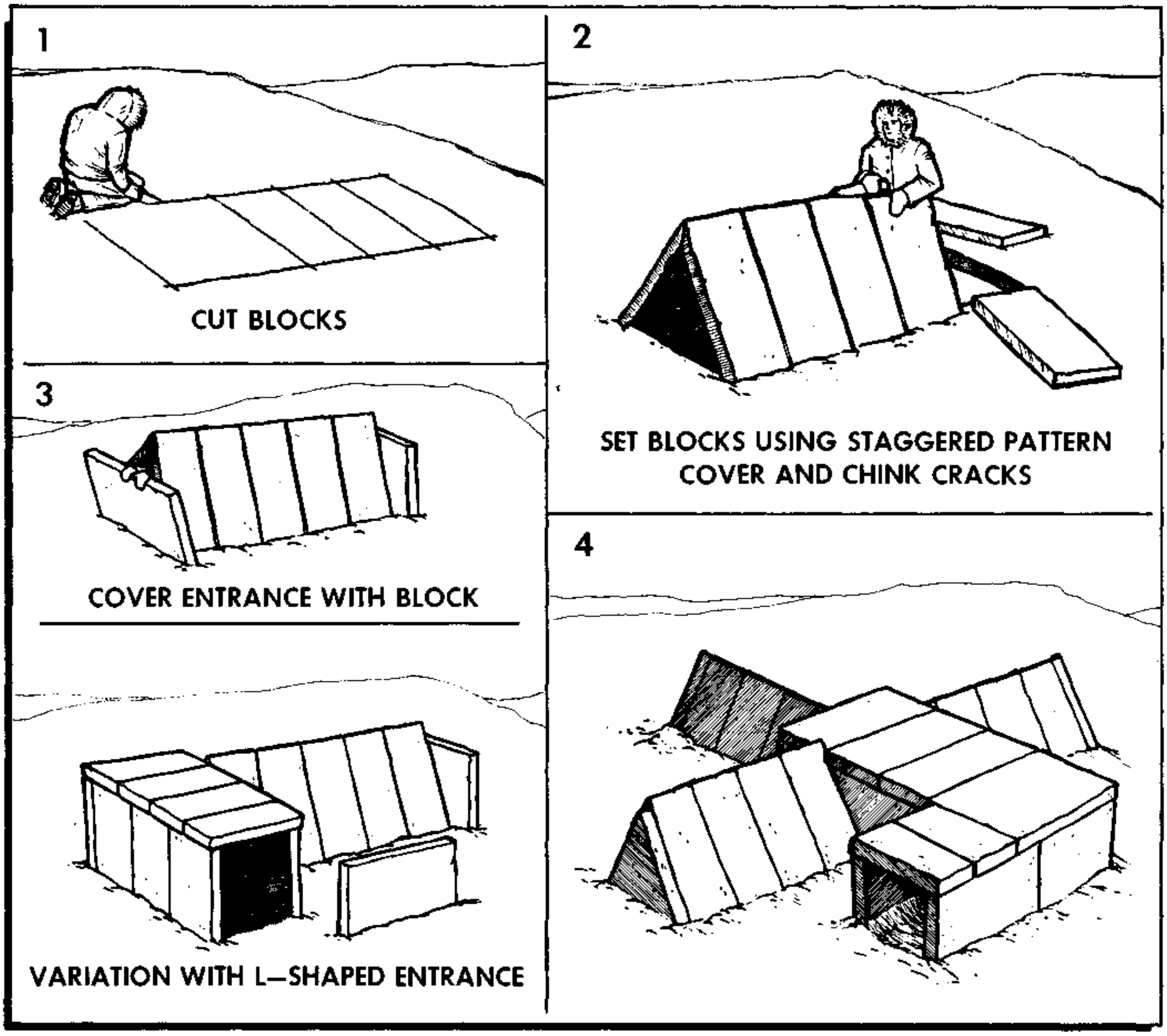
1
CUT BLOCKS
2
SET BLOCKS USING STAGGERED PATTERN
COVER AND CHINK CRACKS
3
COVER ENTRANCE WITH BLOCK
VARIATION WITH L—SHAPED ENTRANCE
4

Figure 15-17. Fighter Trench.

arctic. To conserve fuel, it is important to keep the shelter entrance sealed as much as possible (figure 15-27). When it is necessary to go outside the shelter, activities such as gathering fuel, snow or ice for melting, etc., should be done. To expedite matters, a trash receptacle may be kept inside the door, and equipment may be stored in the entry way. Necessities which cannot be stored inside may be kept just outside the door. Any firearms (guns) the survivor may have must be stored outside the shelter to prevent condensation building which could cause them to malfunction.

b. A standard practice in snow shelter living is for people to relieve themselves indoors when possible. This practice conserves body heat. If the snowdrift is large enough to dig connecting snow caves, one may be used as a toilet room. If not, tin cans may be used for urinals, and snowblocks for solid waste (fecal) matter.

c. Survivors should use thick insulation under themselves when sleeping or resting even if they have a sleeping bag. They can use a thick bough bed in shingle-fashion, seat cushions, parachute, or an inverted and inflated rubber raft.

d. Outer clothing makes good mattress material. A parka makes a good footbag. The shirt and inner trousers may be rolled up for a pillow. Socks and insoles can be separated and aired in the shelter. Drying may be completed in the sleeping bag by stowing around the hips. This drying method should only be used as a last resort.

e. Keeping the sleeping bag clean, *dry*, and fluffed will give maximum warmth. To dry the bag, it should be turned inside out, frost beaten out, and warmed before the fire—taking care that it doesn't burn.

f. To keep moisture (from breath) from wetting the sleeping bag, a moisture cloth should be improvised from a piece of clothing, a towel, or parachute fabric. It can then be lightly wrapped around the head in such a way that the breath is trapped inside the cloth. A piece of fabric dries easier than a sleeping bag. If cold is experienced during the night, survivors should exercise by fluttering their feet up and down or by beating the inside of the bag with their hands. Food or hot liquids can be helpful.

g. Snow remaining in clothing will melt in a warm shelter. When the clothing is again taken outside, the water formed will turn to ice and reduce the CLo value. Brush clothes before entering the shelter. Under living conditions where drying clothing is difficult, it is easier to keep clothing from getting wet than having to dry it out later.

h. If all the snow cannot be eliminated from outer clothing, survivors should remove the clothing and store it in the entry way or on the floor away from the source of heat so it remains cold. If ice should form in clothing, it may be beaten out with a stick.

i. In the cramped quarters of any small emergency shelter, pots of food or drink can be accidentally kicked over. The cooking area, even if it is only a Sterno stove, should be located out of the way, possibly in a snow alcove.

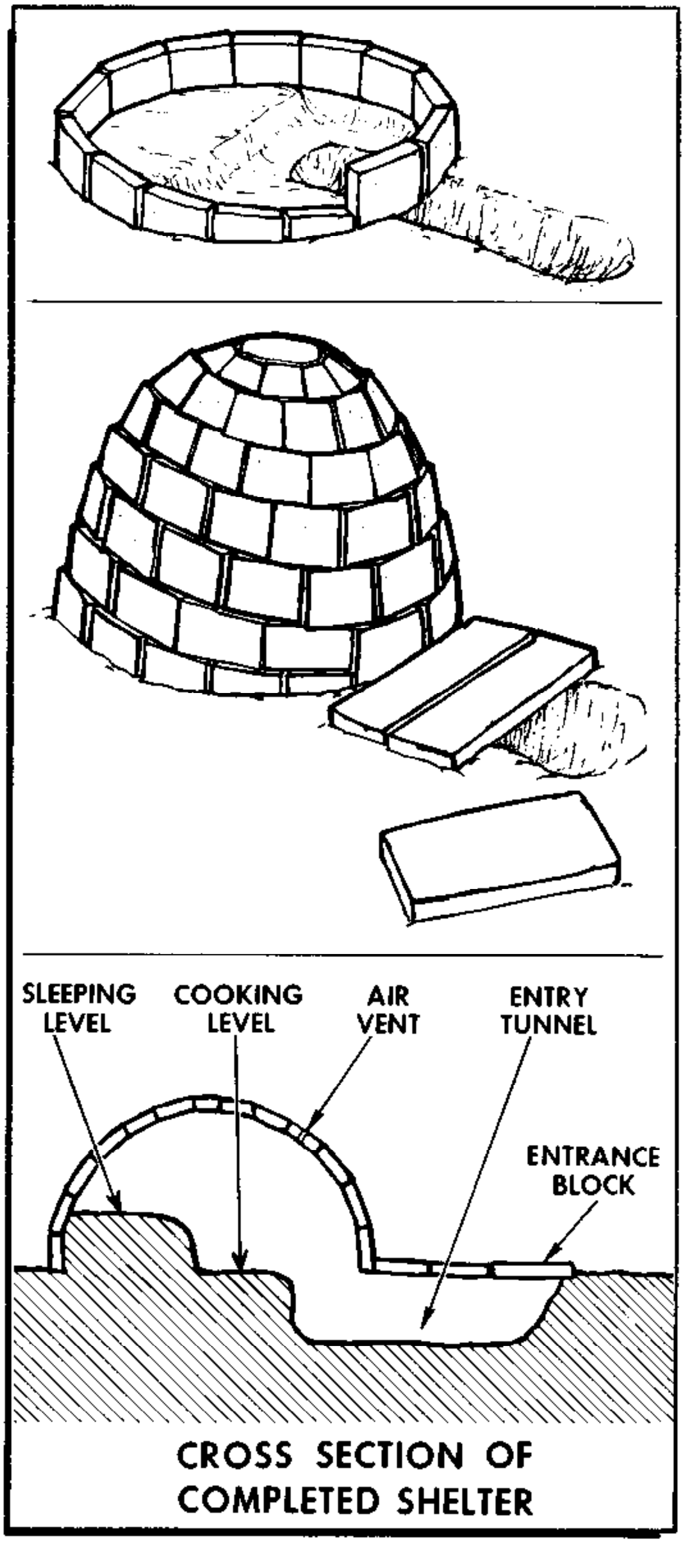

Figure 15-18. Igloo.

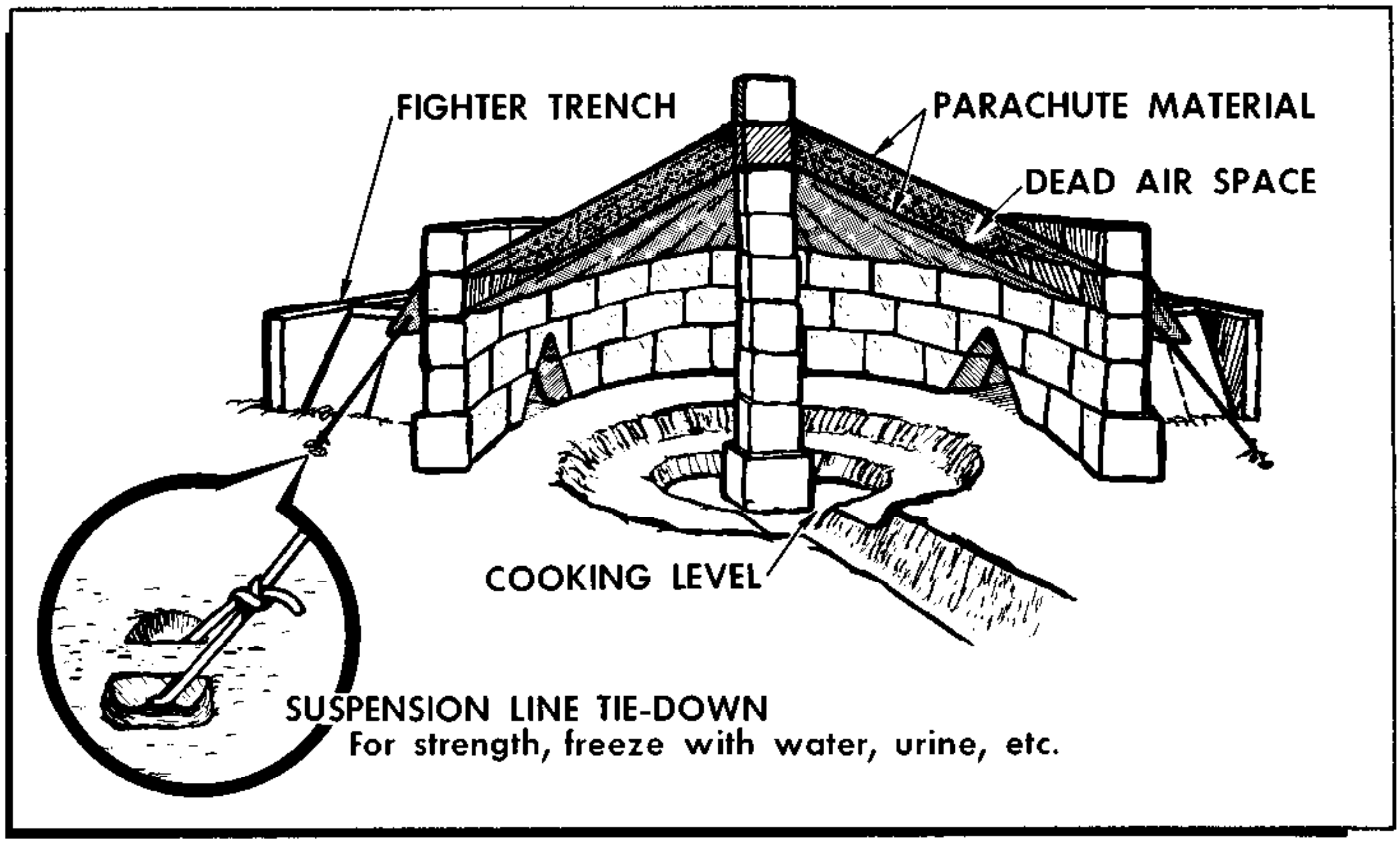
FIGHTER TRENCH
PARACHUTE MATERIAL
DEAD AIR SPACE
COOKING LEVEL
SUSPENSION LINE TIE-DOWN
For strength, freeze with water, urine, etc.

Figure 15-19. Para-Snow House.

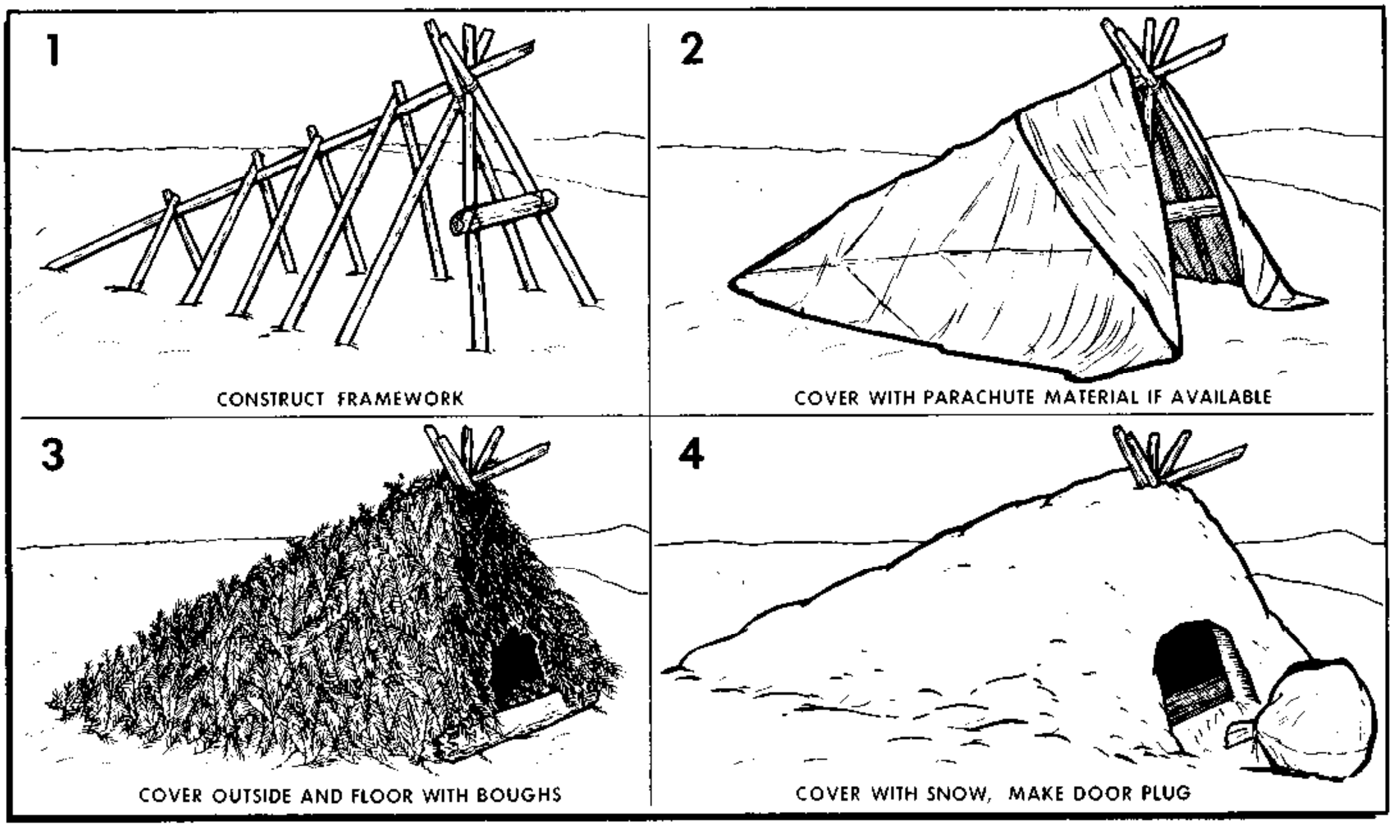
1
CONSTRUCT FRAMEWORK
2
COVER WITH PARACHUTE MATERIAL IF AVAILABLE
3
COVER OUTSIDE AND FLOOR WITH BOUGHS
4
COVER WITH SNOW, MAKE DOOR PLUG

Figure 15-20. Thermal A-Frame.

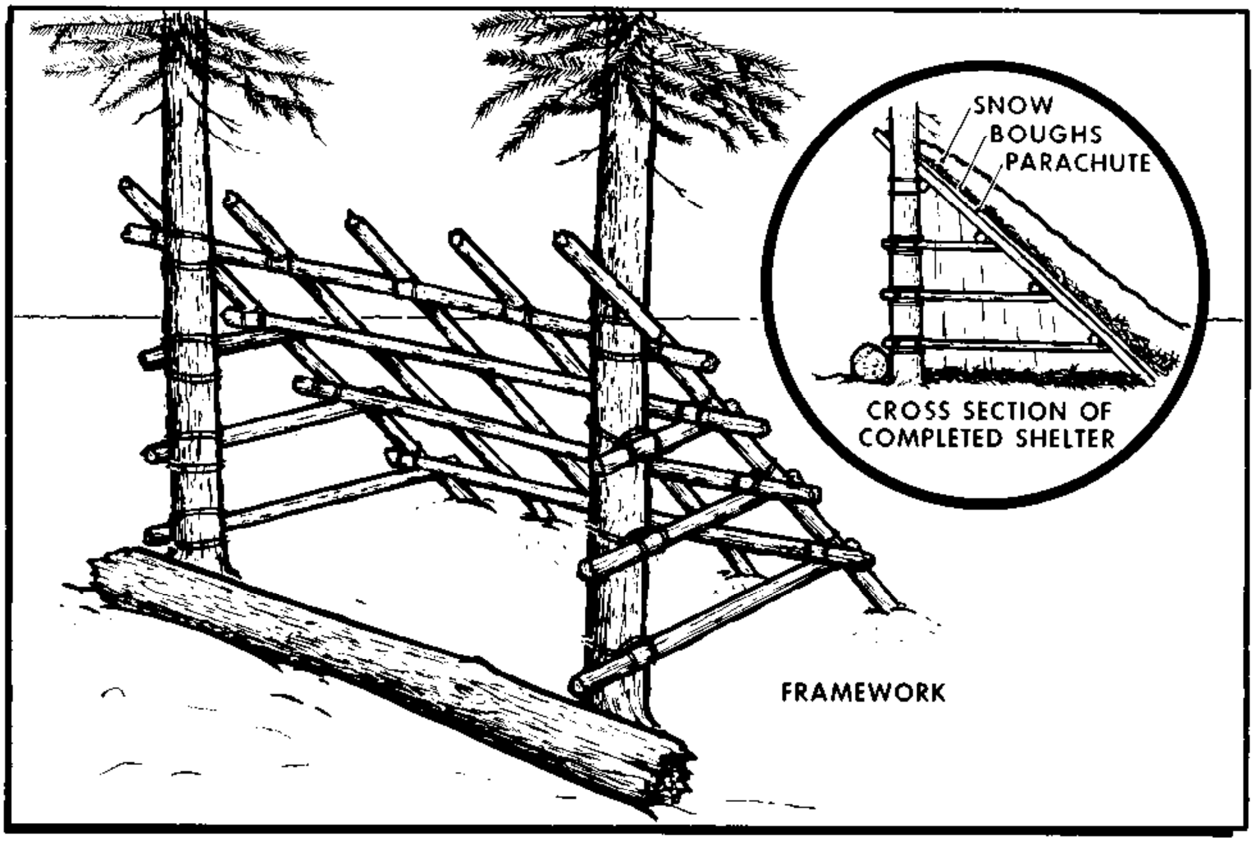
SNOW
BOUGHS
PARACHUTE
CROSS SECTION OF COMPLETED SHELTER
FRAMEWORK

Figure 15-21. Lean-To or Wedge.

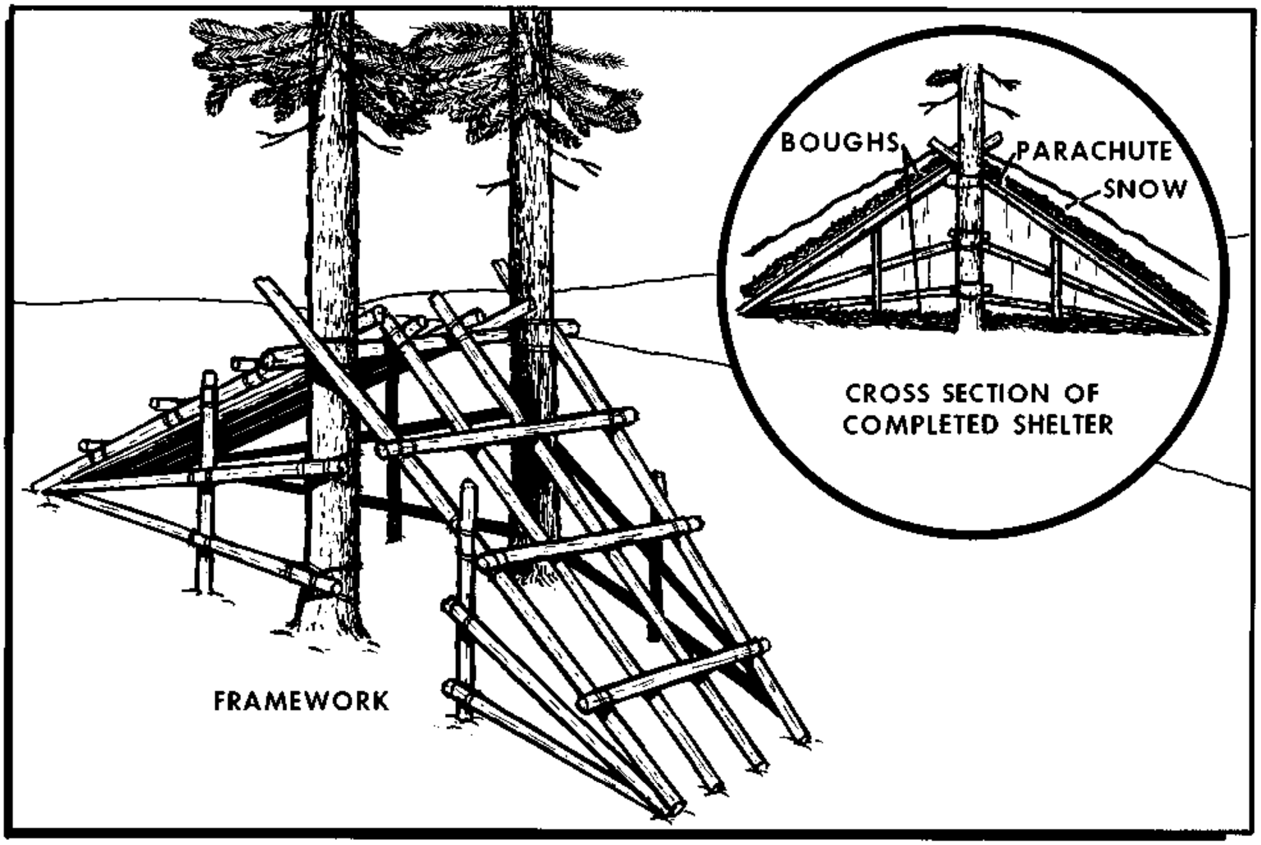
BOUGHS
PARACHUTE
SNOW
CROSS SECTION OF COMPLETED SHELTER
FRAMEWORK

Figure 15-22. Double Lean-To.

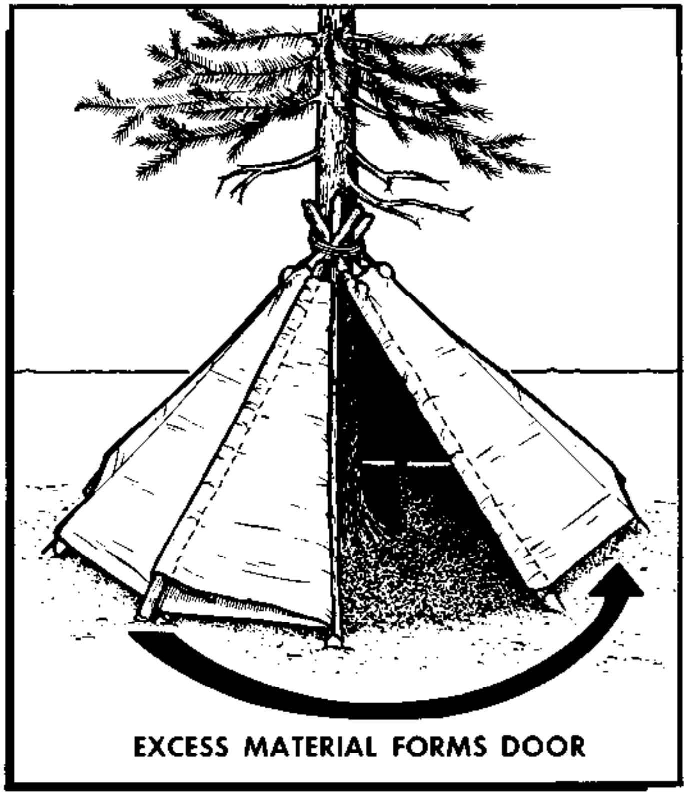

Figure 15-23. Fan Shelter.

15-16. Summer Considerations for Arctic and Arctic-Like Areas:

a. Survivors need shelter against rain and insects. They should choose a campsite near water but on high, dry ground if possible. Survivors should also stay away from thick vegetation, as mosquitoes and flies will make life miserable. A good campsite is a ridge top, cold lake shore, or a spot which gets a breeze.

b. If survivors stay with the aircraft, it can be used for shelter during the summer. They should cover openings with netting or parachute cloth to keep insects out and cook outside to avoid carbon monoxide poisoning. Fires must be built a safe distance from the aircraft.

c. Many temperate area shelters are suitable for summer arctic conditions. The paratepee (of the 1- or no-pole variety) is especially good. It will protect from precipitation and keep insects out.

15-17. Shelter for Open Seas. Personal protection from the elements is just as important on the seas as it is anywhere else. Some rafts come equipped with insulated floors, spray shields, and canopies to protect survivors from heat, cold, and water. If rafts are not so equipped or the equipment has been lost, survivors should try to improvise these items using parachute material, clothing, or other equipment.

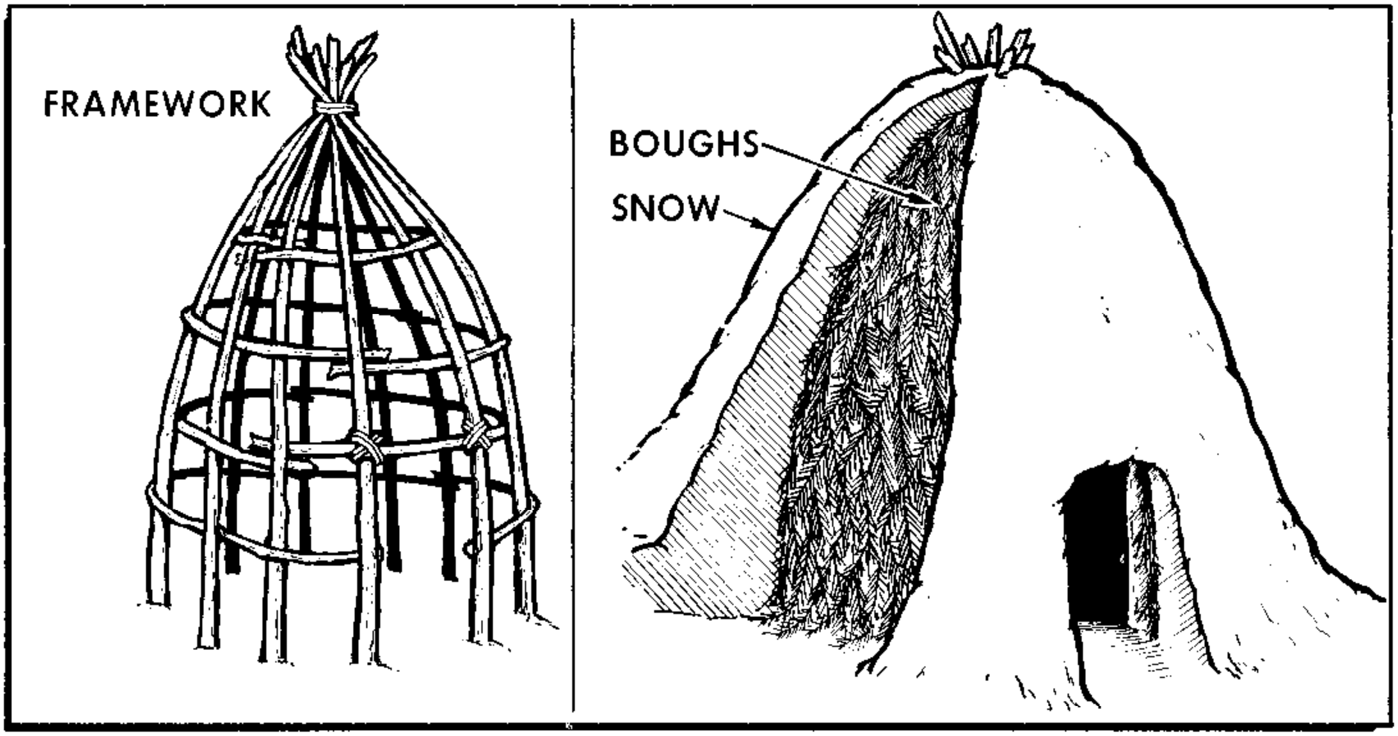

Figure 15-24. Willow Frame Shelter.

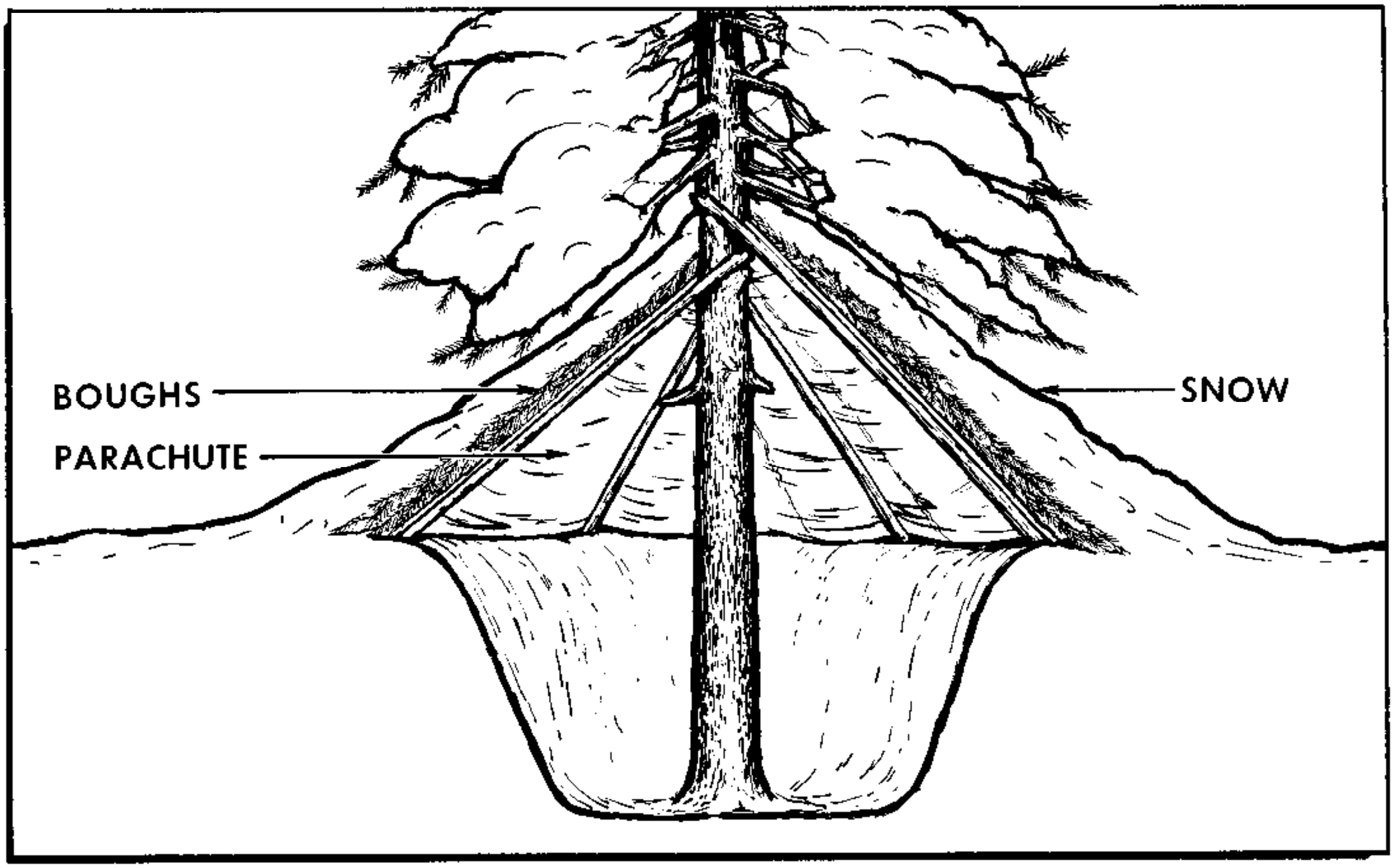
BOUGHS
PARACHUTE
SNOW

Figure 15-25. Tree Well Shelter.

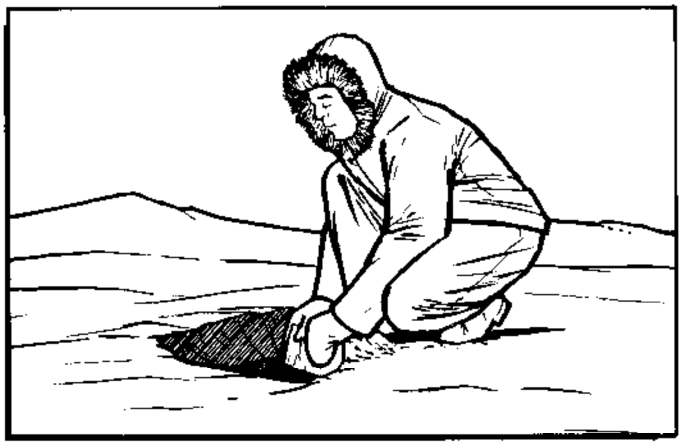

Figure 15-26. Scraping Snow to Bare Earth.

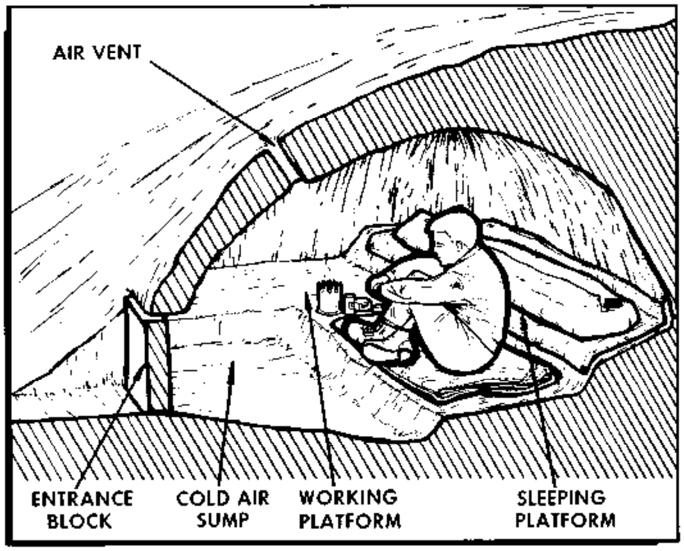
AIR VENT
ENTRANCE BLOCK
COLD AIR SUMP
WORKING PLATFORM
SLEEPING PLATFORM

Figure 15-27. Snow Cave Shelter Living.

Chapter 16

FIRECRAFT

16-1. Introduction:

a. The need for a fire should be placed high on the list of priorities. Fire is used for warmth, light, drying clothes, signaling, making tools, cooking, and water purification. When using fire for warmth, the body uses less calories for heat and consequently requires less food. Just having a fire to sit by is a morale booster. Smoke from a fire can be used to discourage insects.

b. Avoid building a very large fire. Small fires require less fuel, are easier to control, and their heat can be concentrated. Never leave a fire unattended unless it is banked or contained. Banking a fire is done by scraping cold ashes and dry earth onto the fire, leaving enough air coming through the dirt at the top to keep the fuel smoldering. This will keep the fire safe and allow it to be rekindled from the saved coals.

16-2. Elements of Fire:

a. The three essential elements for successful fire building are fuel, heat, and oxygen. These combined elements are referred to as the "fire triangle." By limiting fuel, only a small fire is produced. If the fire is not fed properly, there is too much or too little fire. Green fuel is difficult to ignite, and the fire must be burning well before it is used for fuel. Oxygen and heat must be accessible to ignite any fuel.

b. The survivor must take time and prepare well! Preparing all of the stages of fuel and all of the parts of the fire starting apparatus is the key. To be successful at firecraft, one needs to practice and be patient.

c. The fuels used in building a fire normally fall into three categories (figure 16-1), relating to their size and flash point: tinder, kindling, and fuel.

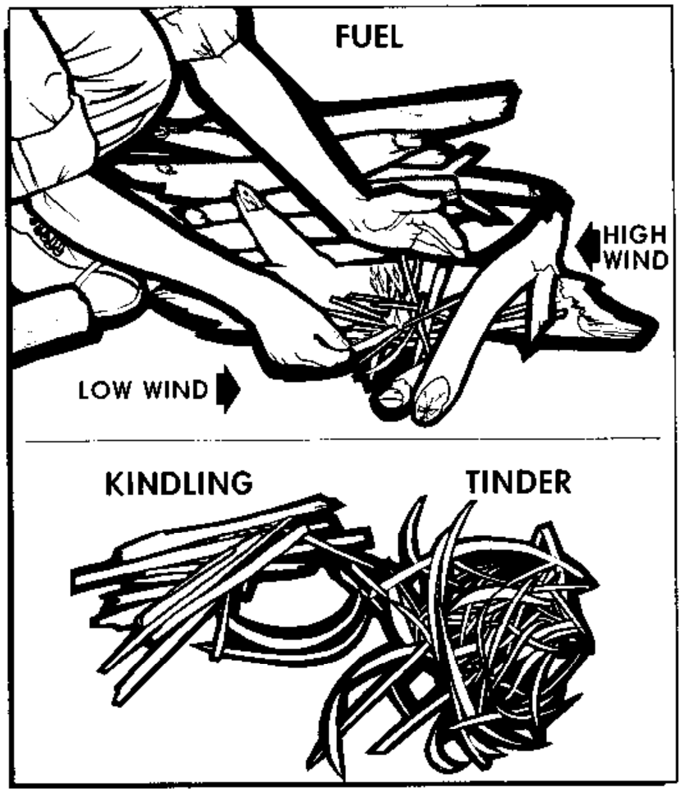

Figure 16-1. Stages of a Fire.

(1) Tinder is any type of small material having a low flash point. It is easily ignited with a minimum of heat, even a spark. Tinder must be arranged to allow air (oxygen) between the hair-like, bone-dry fibers. The preparation of tinder for fire is one of the most important parts of firecraft. Dry tinder is so critical that pioneers used extreme care to have some in a waterproof "tinder box" at all times. It may be necessary to have two or three stages of tinder to get the flame to a useful size. Tinders include:

(a) The shredded bark from some trees and bushes.

(b) Cedar, birch bark, or palm fiber.

(c) Crushed fibers from dead plants.

(d) Fine, dry woodshavings, and straw/grasses.

(e) Resinous sawdust.

(f) Very fine pitch woodshavings (resinous wood from pine or sappy conifers).

(g) Bird or rodent nest linings.

(h) Seed down (milkweed, cattail, thistle).

(i) Charred cloth.

(j) Cotton balls or lint.

(k) Steel wool.

(l) Dry powdered sap from the pine tree family (also known as pitch).

(m) Paper.

(n) Foam rubber.

(2) Kindling is the next larger stage of fuel material. It should also have a high combustible point. It is added to, or arranged over, the tinder in such a way that it ignites when the flame from the tinder reaches it. Kindling is used to bring the burning temperature up to the point where larger and less combustible fuel material can be used. Kindling includes:

(a) Dead dry small twigs or plant fibers.

(b) Dead dry thinly shaved pieces of wood, bamboo, or cane (always split bamboo as sections can explode).

(c) Coniferous seed cones and needles.

(d) "Squaw wood" from the underside of coniferous trees; dead, small branches next to the ground sheltered by the upper live part of the tree.

(e) Pieces of wood removed from the insides of larger pieces.

(f) Some plastics such as the spoon from an inflight ration.

(g) Wood which has been soaked or doused with flammable materials; that is, wax, insect repellent, petroleum fuels, and oil.

(h) Strips of petrolatum gauze from a first aid kit.

(i) Dry split wood burns readily because it is drier inside. Also the angular portions of the wood burn easier than the bark-covered round pieces because it exposes more surface to the flame. The splitting of all fuels will cause them to burn more readily.

(3) Fuel, unlike tinder and kindling, does not have to be kept completely dry as long as there is enough kindling to raise the fuel to a combustible temperature. It is recommended that all fine materials be protected from moisture to prevent excessive smoke production. (Highly flammable liquids should not be poured on an existing fire. Even a smoldering fire can cause the liquids to explode and cause serious burns.) The type of fuel used will determine the amount of heat and light the fire will produce. Dry split hardwood trees (oak, hickory, monkey pod, ash) are less likely to produce excessive smoke and will usually provide more heat than soft woods. They may also be more difficult to break into usable sizes. Pine and other conifers are fast-burning and produce smoke unless a large flame is maintained. Rotten wood is of little value since it smolders and smokes. The weather plays an important role when selecting fuel. Standing or leaning wood is usually dry inside even if it is raining. In tropical areas, avoid selecting wood from trees that grow in swampy areas or those covered with mosses. Tropical soft woods are not usually a good fuel source. Trial and error is sometimes the best method to determine which fuel is best. After identifying the burning properties of available fuel, a selection can be made of the type needed. Recommended fuel sources are:

(a) Dry standing dead wood and dry dead branches (those that snap when broken). Dead wood is easy to split and break. It can be pounded on a rock or wedged between other objects and bent until it breaks.

(b) The insides of fallen trees and large branches may be dry even if the outside is wet. The heart wood is usually the last to rot.

(c) Green wood which can be made to burn is found almost anywhere, especially if finely split and mixed evenly with dry dead wood.

(d) In treeless areas, other natural fuels can be found. Dry grasses can be twisted into bunches. Dead cactus and other plants are available in deserts. Dry peat moss can be found along the surface of undercut streambanks. Dried animal dung, animal fats, and sometimes even coal can be found on the surface. Oil impregnated sand can also be used when available.

16-3. Fire Location. The location of a fire should be carefully selected. An old story is told of a mountain man who used his last match to light a fire built under a snow-covered tree. The heat from the fire melted the snow and it slid off the tree and put out the fire. For a survivor, this type of accident can be very demoralizing or even deadly. Locate and prepare the fire carefully.

16-4. Fire Site Preparation:

a. After a site is located, twigs, moss, grass, or duff should be cleaned away. Scrape at least a 3-foot diameter area down to bare soil for even a small fire. Larger fires require a larger area. If the fire must be built on snow, ice, or wet ground, survivors should build a platform of green logs or rocks. (Beware of wet or porous rocks, they may explode when heated.)

b. There is no need to dig a hole or make a circle of rocks in preparation for fire building. Rocks may be placed in a circle and filled with dirt, sand, or gravel to raise the fire above the moisture from wet ground. The purpose of these rocks is to hold the platform only.

c. To get the most warmth from the fire, it should be built against a rock or log reflector (figure 16-2). This will direct the heat into the shelter. Cooking fires can be walled-in by logs or stones. This will provide a platform for cooking utensils and serve as a windbreak to help keep the heat confined.

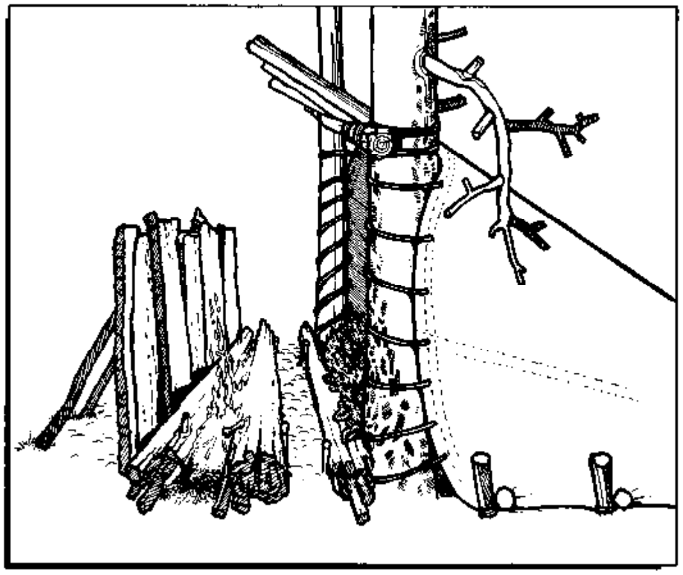

Figure 16-2. Fire Reflector.

d. After preparing the fire, all materials should be placed together and arranged by size (tinder, kindling, and fuel). As a rule of thumb, survivors should have three times the amount of tinder and kindling than is necessary for one fire. It is to their advantage to have too much than not enough. Having plenty of material on hand will prevent the possibility of the fire going out while additional material is gathered.

16-5. Firemaking With Matches (or Lighter):

a. Survivors should arrange a small amount of kindling in a low pyramid, close enough together so flames can jump from one piece to another. A small opening should be left for lighting and air circulation.

b. Matches can be conserved by using a "shave stick," or by using a loosely tied fagot of thin, dry twigs. The match must be shielded from wind while igniting the shave stick. The stick can then be applied to the lower windward side of the kindling.

c. Small pieces of wood or other fuel can be laid gently on the kindling before lighting or can be added as the kindling begins to burn. The survivors can then place smaller pieces first, adding larger pieces of fuel as the fire begins to burn. They should avoid smothering the fire by crushing the kindling with heavy wood.

d. Survivors only have a limited number of matches or other instant fire-starting devices. In a long-term situation, they should use these devices sparingly or carry fire with them when possible. Many primitive cultures carry fire (fire bundles) by using dry punk or fiberous barks (cedar) encased in a bark. Others use torches. Natural fire bundles also work well for holding the fire (figure 16-3).

e. The amount of oxygen must be just enough to keep the coals inside the dry punk burning slowly. This requires constant vigilance to control the rate of the burning process. The natural fire bundle is constructed in a cross section as shown in figure 16-3.

16-6. Heat Sources. A supply of matches, lighters, and other such devices will only last a limited time. Once the supply is depleted, they cannot be used again. If possible, before the need arises, survivors should become skilled at starting fires with more primitive means, such as friction, heat, or a sparking device. It is essential that they continually practice these procedures. The need to start a fire may arise at the most inopportune times. One of the greatest aids a survivor can have for rapid fire starting is the "tinder box" previously mentioned. Using friction, heat, and sparks are very reliable methods for those who use them on a regular basis. Therefore, survivors must practice these methods. Survivors must be aware of the problems associated with the use of primitive heat sources. If the humidity is high in the immediate area, a fire may be difficult to ignite even if all other conditions are favorable. For primitive methods to be successful, the materials must be BONE DRY. The primitive people who use these ignition methods take great care to keep their tinder, kindling, and other fuels dry, even to the point of wrapping many layers of waterproof materials around it. *PREPARATION, PRACTICE,* and *PATIENCE* in the use of primitive fire-building techniques cannot be over emphasized. A key point in all primitive methods is to ensure that the tinder is not disturbed.

a. Flint and Steel:

(1) Flint and steel is one way to produce fire without matches.

(a) To use this method, survivors must hold a piece of flint in one hand above the tinder.

(b) Grasp the steel in the other hand and strike the flint with the edge of the steel in a downward glancing blow (figure 16-4).

(2) True flint is not necessary to produce sparks. Iron pyrite and quartz will also give off sparks even if they are struck against each other. Check the area and select the best spark-producing stone as a backup for the

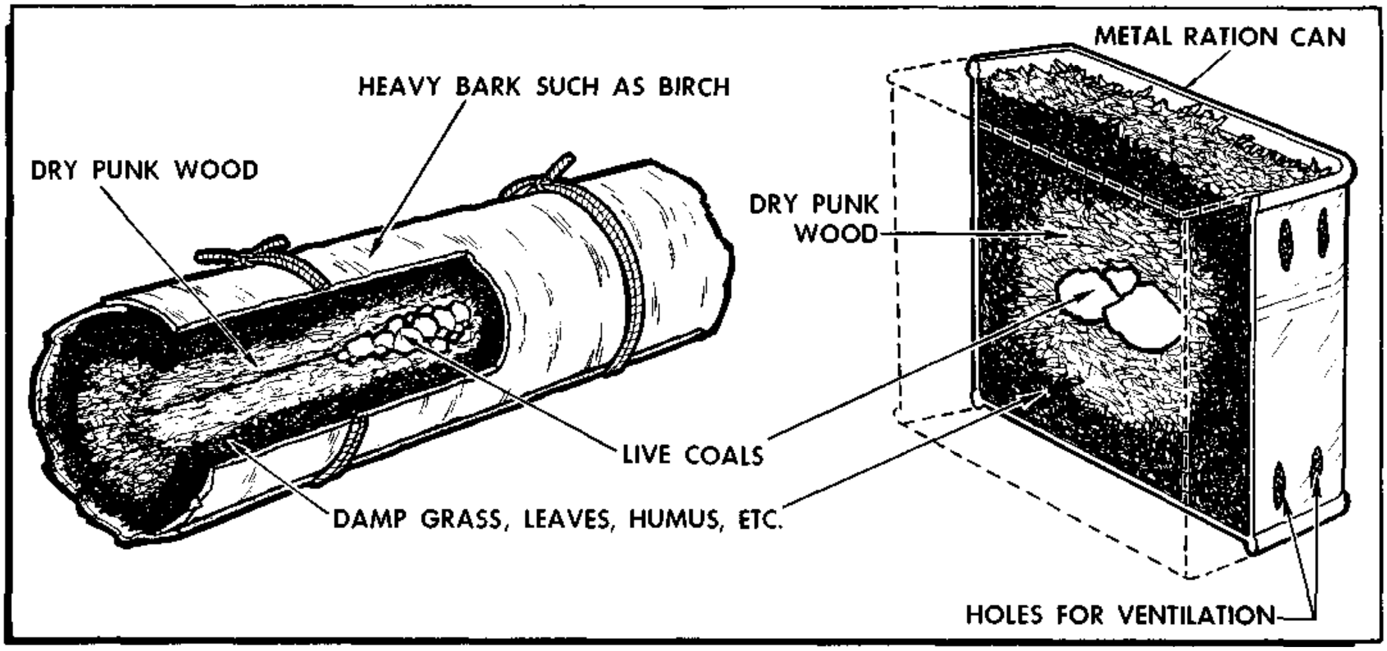

Figure 16-3. Fire Bundles.

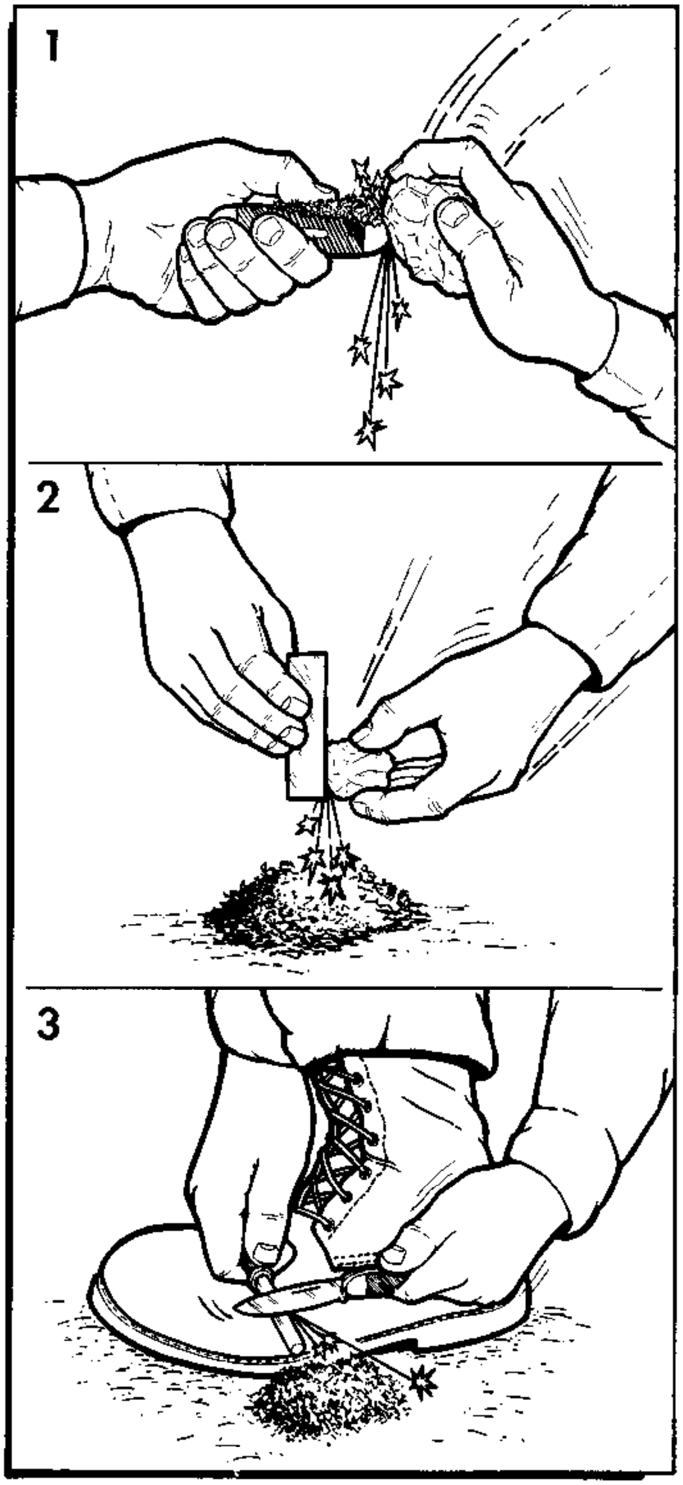

Figure 16-4. Fire Starting With Flint and Steel.

available matches. The sparks must fall on the tinder and then be blown or fanned to produce a coal and subsequent flame.

(3) Synthetic flint, such as the so-called metal match, consists of the same type material used for flints in commercial cigarette lighters. Some contain magnesium which can be scraped into tinder and into which the spark is struck. The residue from the "match" burns hot and fast and will compensate for some moisture in tinder. If issued survival kits do not contain this item and the survivors choose to make one rather than buy it, lighter flints can be glued into a groove in a small piece of wood or plastic. The survivors can then practice striking a spark by scratching the flint with a knife blade. A 90-degree angle between the blade and flint works best. The device must be held close enough for the sparks to hit the tinder, but enough distance must be allowed to avoid accidentally extinguishing the fire. Cotton balls dipped in petroleum jelly make excellent tinder with flint and steel. When the tinder ignites, additional tinder, kindling, and fuel can be added.

b. Batteries:

(1) Another method of producing fire is to use the battery of the aircraft, vehicle, storage batteries, etc. Using two insulated wires, connect one end of a wire to the positive post of the battery and the end of the other wire to the negative post. Touch the two remaining ends to the ends of a piece of noninsulated wire. This will cause a short in the electrical circuit and the noninsulated wire will begin to glow and get hot. Material coming into contact with this hot wire will ignite. Survivors should use caution when attempting to start a fire with a battery. They should ensure that sparks or flames are not produced near the battery because explosive hydrogen gas is produced and can result in serious injury (figure 16-5).

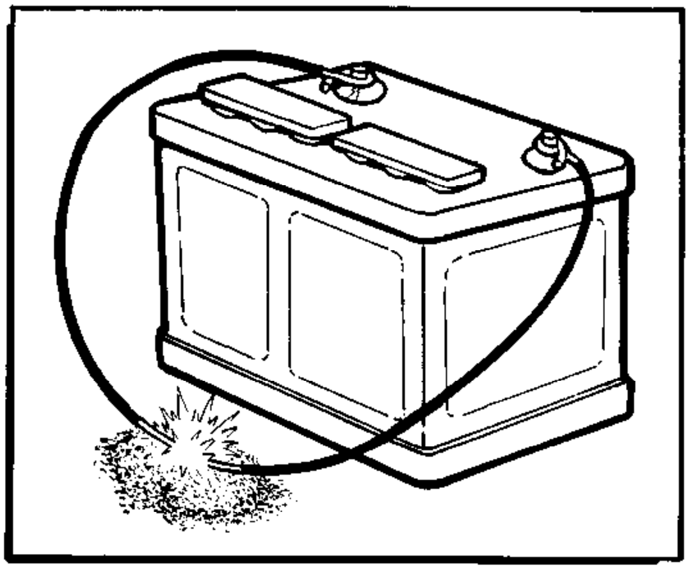

Figure 16-5. Fire Starting With Batteries.

(2) If fine grade steel wool is available, a fire may be started by stretching it between the positive and negative posts until the wire itself makes a red coal.

c. Burning Glass. If survivors have sunlight and a burning glass, a fire can be started with very little physical effort (figure 16-6). Concentrate the rays of the Sun on tinder by using the lens of a lensatic compass, a camera lens, or the lens of a flashlight which magnifies; even a convex piece of bottle glass may work. Hold the lens so that the brightest and smallest spot of concentrated light falls on the tinder. Once a whisp of smoke is produced, the tinder should be fanned or blown upon until the smoking coal becomes a flame. Powdered charcoal in the tinder will decrease the ignition time. Add kindling carefully as in any other type of fire. Practice will reduce the time it takes to light the tinder.

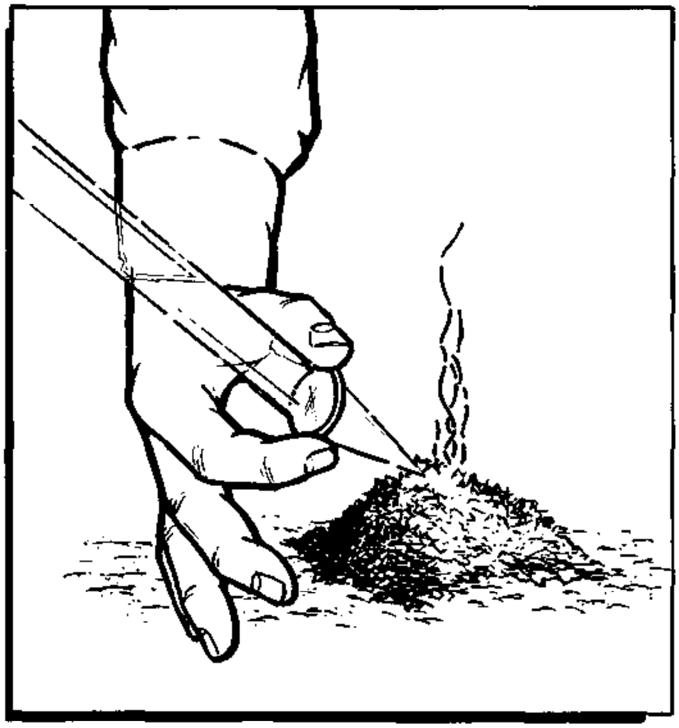

Figure 16-6. Fire Starting With Burning Glass.

d. Flashlight Reflector. A flashlight reflector can also be used to start a fire (figure 16-7). Place the tinder in the center of the reflector where the bulb is usually located. Push it up from the back of the hole until the hottest light is concentrated on the end and smoke results. If a cigarette is available, use it as a tinder for this method.

e. Bamboo Fire Saw:

(1) The bamboo fire saw is constructed from a section of dry bamboo with both end joints cut off. The section of bamboo, about 12 inches in length, is split in half lengthwise. The inner wall of one of the halves (called the "running board") is scraped or shaved thin. This is done in the middle of the running board. A notch to serve as a guide is cut in the outer sheath opposite the scraped area of the inner wall. This notch runs across the running board at a 90-degree angle (figure 16-8).

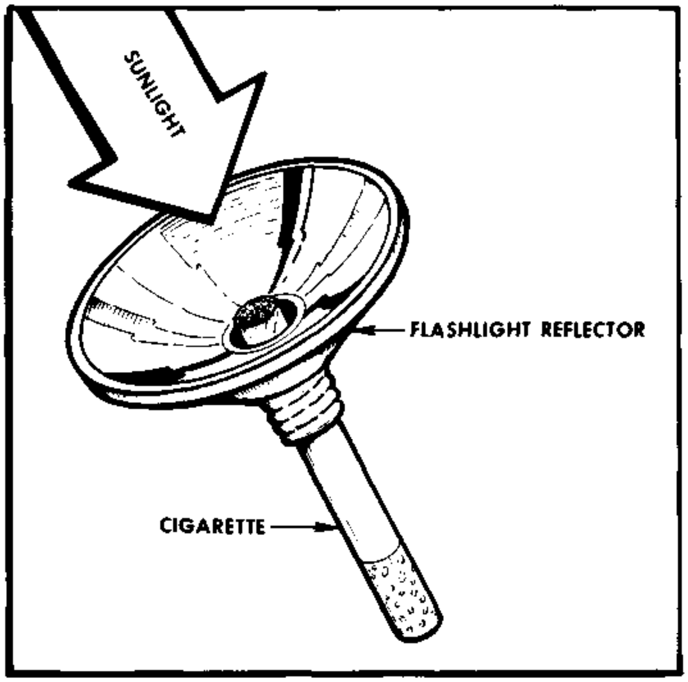

Figure 16-7. Fire Starting With Flashlight Reflector.

(2) The other half of the bamboo joint is further split in half lengthwise, and one of the resultant quarters is used as a "baseboard." One edge of the baseboard is shaved down to make a tapered cutting edge. The baseboard is then firmly secured with the cutting edge up. This may be done by staking it to the ground in any manner which does not allow it to move (figure 16-8).

(3) Tinder is made by scraping the outer sheath of the remaining quarter piece of the bamboo section. The scrapings (approximately a large handful) are then rubbed between the palms of the hands until all of the wood fibers are broken down and dust-like material no longer falls from the tinder. The ball of scrapings is then fluffed to allow maximum circulation of oxygen through the mass (figure 16-8).

(4) The finely shredded and fluffed tinder is placed in the running board directly over the shaved area, opposite the outside notch. Thin strips of bamboo should be placed lengthwise in the running board to hold the tinder in place. These strips are held stationary by the hands when grasping the ends of the running board (figure 16-8).

(5) A long, very thin sliver of bamboo (called the "pick") should be prepared for future use. One end of the running board is grasped in each hand, making sure the thin strips of bamboo are held securely in place. The running board is placed over the baseboard at a right angle, so that the cutting edge of the baseboard fits into

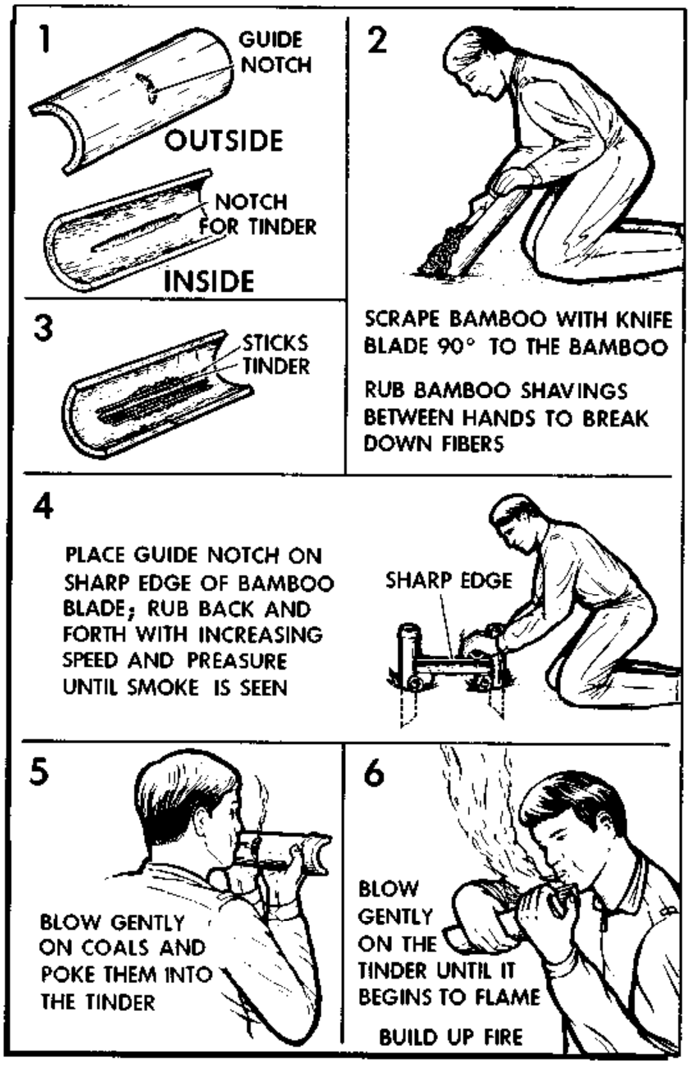
1
GUIDE NOTCH
OUTSIDE
NOTCH FOR TINDER
INSIDE
2
3
STICKS
TINDER
SCRAPE BAMBOO WITH KNIFE BLADE 90° TO THE BAMBOO
RUB BAMBOO SHAVINGS BETWEEN HANDS TO BREAK DOWN FIBERS
4
PLACE GUIDE NOTCH ON SHARP EDGE OF BAMBOO BLADE; RUB BACK AND FORTH WITH INCREASING SPEED AND PREASURE UNTIL SMOKE IS SEEN
SHARP EDGE
5
BLOW GENTLY ON COALS AND POKE THEM INTO THE TINDER
6
BLOW GENTLY ON THE TINDER UNTIL IT BEGINS TO FLAME
BUILD UP FIRE

Figure 16-8. Bamboo Fire Saw.

the notch in the outer sheath of the running board. The running board is then slid back and forth as rapidly as possible over the cutting edge of the baseboard, with sufficient downward pressure to ensure enough friction to produce heat.

(6) As soon as "billows" of smoke rise from the tinder, the running board is picked up. The pick is used to push the glowing embers from the bottom of the running board into the mass of tinder. While the embers are being pushed into the tinder, they are gently blown upon until the tinder bursts into flame.

(7) As soon as the tinder bursts into flame, slowly add kindling in small pieces to avoid smothering the fire. Fuel is gradually added to produce the desired size fire. If the tinder is removed from the running board as soon as it flames, the running board can be reused by cutting a notch in the outer sheath next to the original notch and directly under the scraped area of the inner wall.

f. Bow and Drill:

(1) This is a friction method which has been used successfully for thousands of years. A spindle of yucca, elm, basswood, or any other straight grainwood (not softwood) should be made. The survivors should make sure that the wood is not too hard or it will create a glazed surface when friction is applied. The spindle should be 12 to 18 inches long and three-fourths inch in diameter. The sides should be octagonal, rather than round, to help create friction when spinning. Round one end and work the other end into a blunt point. The round end goes to the top upon which the socket is placed. The socket is made from a piece of hardwood large enough to hold comfortably in the palm of the hand with the curved part up and the flat side down to hold the top of the spindle. Carve or drill a hole in this side and make it smooth so it will not cause undue friction and heat production. Grease or soap can be placed in this hole to prevent friction (figure 16-9).

(2) The bow is made from a stiff branch about 3 feet long and about 1 inch in diameter. This piece should have sufficient flexibility to bend. It is similar to a bow used to shoot arrows. Tie a piece of suspension line or leather thong to both ends so that it has the same tension as that of a bow. There should be enough tension for the spindle to twist comfortably.

(3) The fireboard is made of the softwood and is about 12 inches long, three-fourths inch thick, and 3 to 6 inches wide. A small hollow should be carved in the fireboard. A V-shaped cut can then be made in from the edge of the board. This V-shape should extend into the center of the hollow where the spindle will make the hollow deeper. The object of this "V" cut is to create an angle which cuts off the edge of the spindle as it gets hot and turns to charcoal dust. This is the critical part of the fireboard and must be held steady during the operation of spinning the spindle.

(4) While kneeling on one knee, the other foot can be placed on the fireboard as shown in figure 16-9 and the tinder placed under the fireboard just beneath the V-cut. Care should be taken to avoid crushing the tinder under the fireboard. Space can be obtained by using a small, three-fourths inch diameter stick to hold up the fireboard. This allows air into the tinder where the hot powder (spindle charcoal dust) is collected.

(5) The bow string should be twisted once around the spindle. The spindle can then be placed upright into the spindle hollow (socket). The survivor may press the socket down on the spindle and fireboard. The entire apparatus must be held steady with the hand on the socket braced against the leg or knee. The spindle should begin spinning with long even slow strokes of the bow until heavy smoke is produced. The spinning should become faster until the smoke is very thick. At this point, hot powder, that can be blown into a glowing ember, has been successfully produced. The bow and spindle can then be removed from the fireboard and the

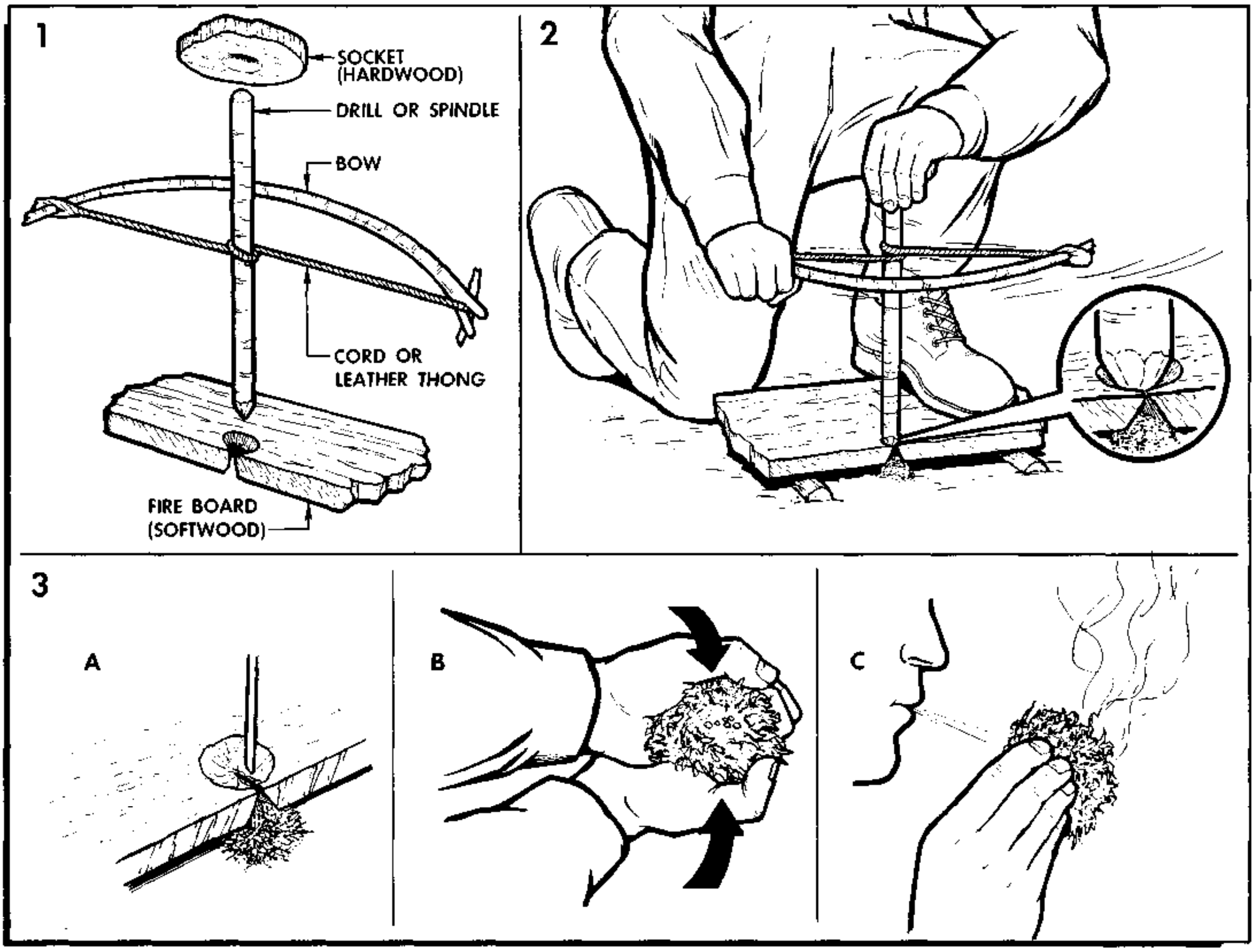

Figure 16-9. Bow and Drill.

tinder can be placed next to the glowing ember making sure not to extinguish it. The tinder must then be rolled gently around the burning ember, and blow into the embers, starting the tinder to burn. This part of the fire is most critical and should be done with care and planning.

(6) The burning tinder is then placed into the waiting fire "lay" containing more tinder and small kindling. At no time in this process should the survivor break concentration or change sequence. The successful use of these primitive methods of fire starting will require a great deal of patience. Success demands dedication and practice.

g. The Fire Thong. The fire thong, another friction method, is used in only those tropical regions where rattan is found. The system is simple and consists of a twisted rattan thong or other strong plant fiber, 4 to 6 feet long, less than 1 inch in diameter, and a 4-foot length of dry wood which is softer than rattan (dedicuous wood) (figure 16-10). Rub with a steady but increasing rhythm.

h. The Plow. The plow is a method used by some primitives and basically follows the principles of other friction methods. The wood used must not glaze with heat applied and must be able to produce powder with friction.

i. Ground Stake. Another variation can be constructed by driving a stake into the ground as shown in figure 16-11.

16-7. Firemaking With Special Equipment:

a. The night end of the day-night flare can be used as a fire starter. This means, however, that survivors must weigh the importance of a fire against the loss of a night flare.

b. Some emergency kits contain small fire starters, cans of special fuels, windproof matches, and other aids.

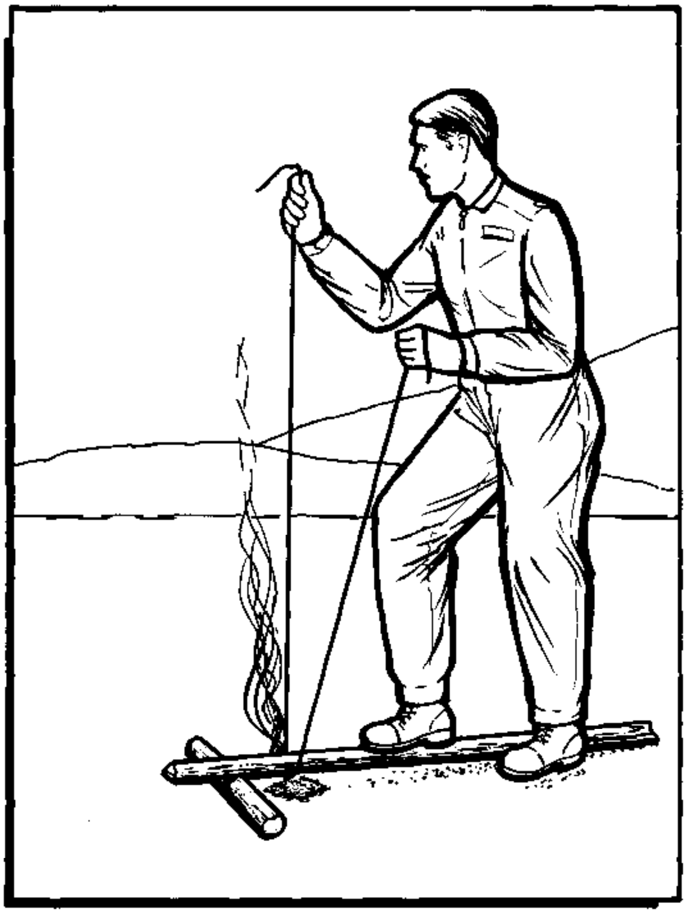

Figure 16-10. Fire Thong.

Survivors should save the fire starters for use in extreme cold and damp (moist) weather conditions.

c. The white plastic spoon (packed in various in-flight rations) may be the type that burns readily. The handle should be pushed deep enough into the ground to support the spoon in an upright position. Light the tip of the spoon. It will burn for about 10 minutes (long enough to dry out and ignite small tinder and kindling).

d. If a candle is available, it should be ignited to start a fire and thus prevent using more than one match. As soon as the fire is burning, the candle can be extinguished and saved for future use.

e. Tinder can be made more combustible by adding a few drops of flammable fuel/material. An example of this would be mixing the powder from an ammunition cartridge with the tinder. After preparing tinder in this manner, it should be stored in a waterproof container for future use. Care must be used in handling this mixture because the flash at ignition could burn the skin and clothing.

f. For thousands of years, the Eskimos and other northern peoples have relied heavily upon oils from animals to heat their homes. A fat stove or "Koodlik" is used by the Eskimos to burn this fuel.

g. Survivors can improvise a stove from a ration can and burn any flammable oil-type liquid or animal fats available. Here again, survivors should keep in mind that if there is only a *limited* amount of animal fat, it should be eaten to produce heat inside the body.

16-8. Burning Aircraft Fuel. On barren lands in the arctic, aircraft fuel may be the only material survivors have available for fire.

a. A stove can be improvised to burn fuel, lubricating oil, or a mixture of both (figure 16-12). The survivor should place 1 or 2 inches of sand or fine gravel in the bottom of a can or other container and add fuel. *Care should be used when lighting the fuel because it may explode.* Slots should be cut into the top of the can to let flame and smoke out, and holes punched just above the level of the sand to provide a draft. A mixture of fuel and oil will make the fire burn longer. If no can is available, a hole can be dug and filled with sand. Fuel is then poured on the sand and ignited. The survivor should not allow the fuel to collect in puddles.

b. Lubricating oil can be burned as fuel by using a wick arrangement. The wick can be made of string, rope, rag, sphagnum moss, or even a cigarette and should be placed on the edge of a receptacle filled with oil. Rags, paper, wood, or other fuel can be soaked in oil and thrown on the fire.

c. A stove can be made of any empty waxed carton by cutting off one end and punching a hole in each side near the unopened end. Survivors can stand the carton on the closed end and loosely place the fuel inside the carton. The stove can then be lit using fuel material left hanging over the end. The stove will burn from the top down.

d. Seal blubber makes a satisfactory fire without a container if gasoline or heat tablets are available to provide an initial hot flame (figure 16-13). The heat source should be ignited on the raw side of the blubber while the fur side is on the ice. A square foot of blubber burns for several hours. Once the blubber catches fire, the heat tablets can be recovered. Eskimos light a small piece of blubber and use it to kindle increasingly larger pieces. The smoke from a blubber fire is dirty, black, and heavy. The flame is very bright and can be seen for several miles. The smoke will penetrate clothing and blacken the skin.

16-9. Useful Firecraft Hints:

a. Conserve matches by only using them on properly prepared fires. They should never be used to light cigarettes or for starting unnecessary fires.

b. Carry some dry tinder in a waterproof container. It should be exposed to the Sun on dry days. Adding a little powdered charcoal will improve it. Cotton cloth is good tinder, especially if scorched or charred. It works well with a burning glass or flint and steel.

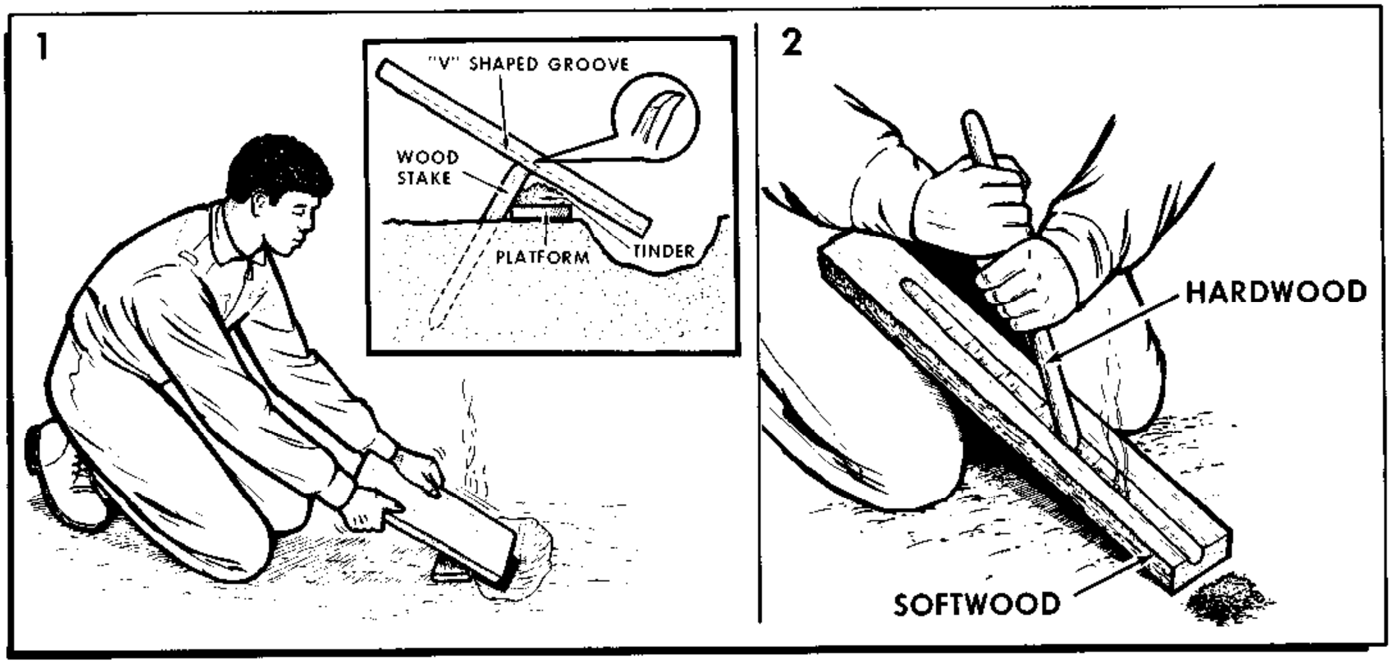

Figure 16-11. Fire Plow.

c. Remember that firemaking can be a difficult job in an arctic environment. The main problem is the availability of firemaking materials. Making a fire starts *WELL* before the match is lit. The fire must be protected from the wind. In wooded areas, standing timber and brush usually make a good windbreak but in open areas, some type of windbreak may have to be constructed. A row of snowblocks, the shelter of a ridge, or a pile of brush will work as a windbreak. It must be high enough to shield the fire from the wind. It may also act as a heat reflector if it is of solid material.

d. Remember, a platform will be required to prevent the fire from melting down through the deep snow and extinguishing it. A platform is also needed if the ground is moist or swampy. The platform can be made of green logs, metal, or any material that will not burn through very readily. Care must be taken when selecting an area for fire building. If the area has a large accumulation of humus material and (or) peat, a platform is needed to avoid igniting the material as it will tend to smolder long after the flames of the fire are extinguished. A smoldering peat fire is almost impossible to put out and may burn for years.

e. In forested areas, the debris on the ground and the lichen mat should be cleared away to mineral soil, if possible, to prevent the fire from spreading.

f. The ignition source used to ignite the fire must be quick and easily operated with hand protection such as mittens. Any number of devices will work well—matches, candle, lighter, fire starter, metal matches, etc.

16-10. Fire Lays. Most fires are built to meet specific needs or uses, either heat, light, or preparing food and water. The following configurations are the most commonly used for fires and serve one or more needs (figure 16-14).

a. Tepee:

(1) The tepee fire can be used as a light source and has a concentrated heat point directly above the apex of the tepee which is ideal for boiling water. To build:

(a) Place a large handful of tinder on the ground in the middle of the fire site.

(b) Push a stick into the ground, slanting over the tinder.

(c) Then lean a circle of kindling sticks against the slanting stick, like a tepee, with an opening toward the windward side for draft.

(2) To light the fire:

(a) Crouch in front of the fire lay with the back to the wind.

(b) Feed the fire from the downwind side, first with thin pieces of fuel, then gradually with thicker pieces.

(c) Continue feeding until the fire has reached the desired size. The tepee fire has one big drawback. It tends to fall over easily. However, it serves as an excellent starter fire.

b. Log Cabin. As the name implies, this lay looks similar to a log cabin. Log cabin fires give off a great amount of light and heat primarily because of the amount of oxygen which enters the fire. The log cabin

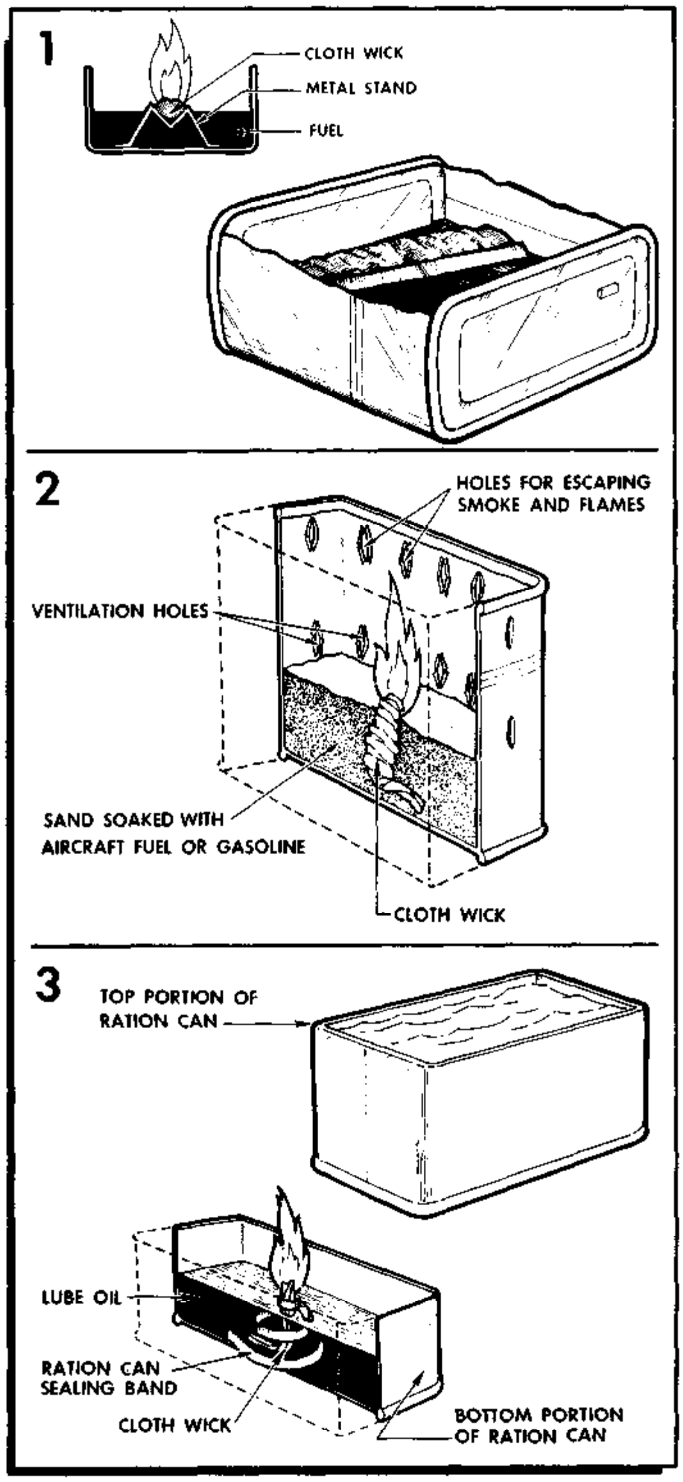
1
CLOTH WICK
METAL STAND
FUEL
2
HOLES FOR ESCAPING SMOKE AND FLAMES
VENTILATION HOLES
SAND SOAKED WITH AIRCRAFT FUEL OR GASOLINE
CLOTH WICK
3
TOP PORTION OF RATION CAN
LUBE OIL
RATION CAN SEALING BAND
CLOTH WICK
BOTTOM PORTION OF RATION CAN

Figure 16-12. Fat and Oil Stoves.

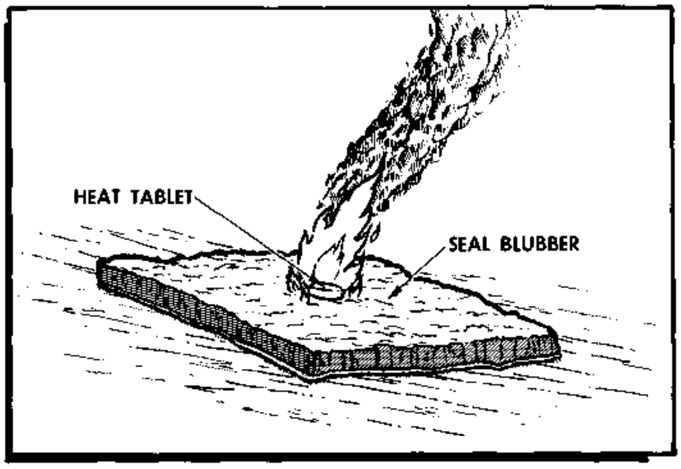
HEAT TABLET
SEAL BLUBBER

Figure 16-13. Heat Tablet/Seal Blubber

fire creates a quick and large bed of coals and can be used for cooking or as the basis for a signal fire. If one person or a group of people are going to use the coals for cooking, the log cabin can be modified into a long fire or a keyhole fire.

c. Long Fire. The long fire begins as a trench, the length of which is layed to take advantage of existing wind. The long fire can also be built above ground by using two parallel green logs to hold the coals together. These logs should be at least 6 inches in diameter and situated so the cooking utensils will rest upon the logs. Two 1-inch thick sticks can be placed under both logs, one at each end of the long fire. This is done to allow the coals to receive more air.

d. Keyhole Fire. To construct a keyhole fire, a hole is dug in the shape of an old style keyhole and does the same thing as the long fire.

e. Pyramid Fire. The pyramid fire looks similar to a log cabin fire except there are layers of fuel in place of a hollow framework. The advantage of a pyramid fire is that it burns for a long time resulting in a large bed of coals. This fire could possibly be used as an overnight fire when placed in front of a shelter opening.

f. Star Fire. This fire is used when conservation of fuel is necessary or a small fire is desired. It burns at the center of the "wheel" and must be constantly tended. Hardwood fuels work best with this type of fire.

g. "T" Fire. Used for large group cooking. The size of this lay may be adjusted to meet the group's cooking needs. In the top part of the "T," the fire is constructed and maintained as long as needed to provide hot coals for cooking in the bottom part of the "T" fire lay. The number of hot coals may be adjusted in the lower part of the "T" fire lay to regulate the cooking temperature.

h. "V" Fire. This fire lay is a modification of the long fire. The configuration allows a survivor to either block strong winds, or take advantage of light breezes. During high wind conditions, the vertex of the "V"—formed by the two outside logs—is placed in the direction from which the winds are coming, thereby sheltering the tinder (kindling) for ignition. Reversing the "lay" will funnel light breezes into the tinder (kindling) thereby facilitating ease of ignition (figure 16-1).

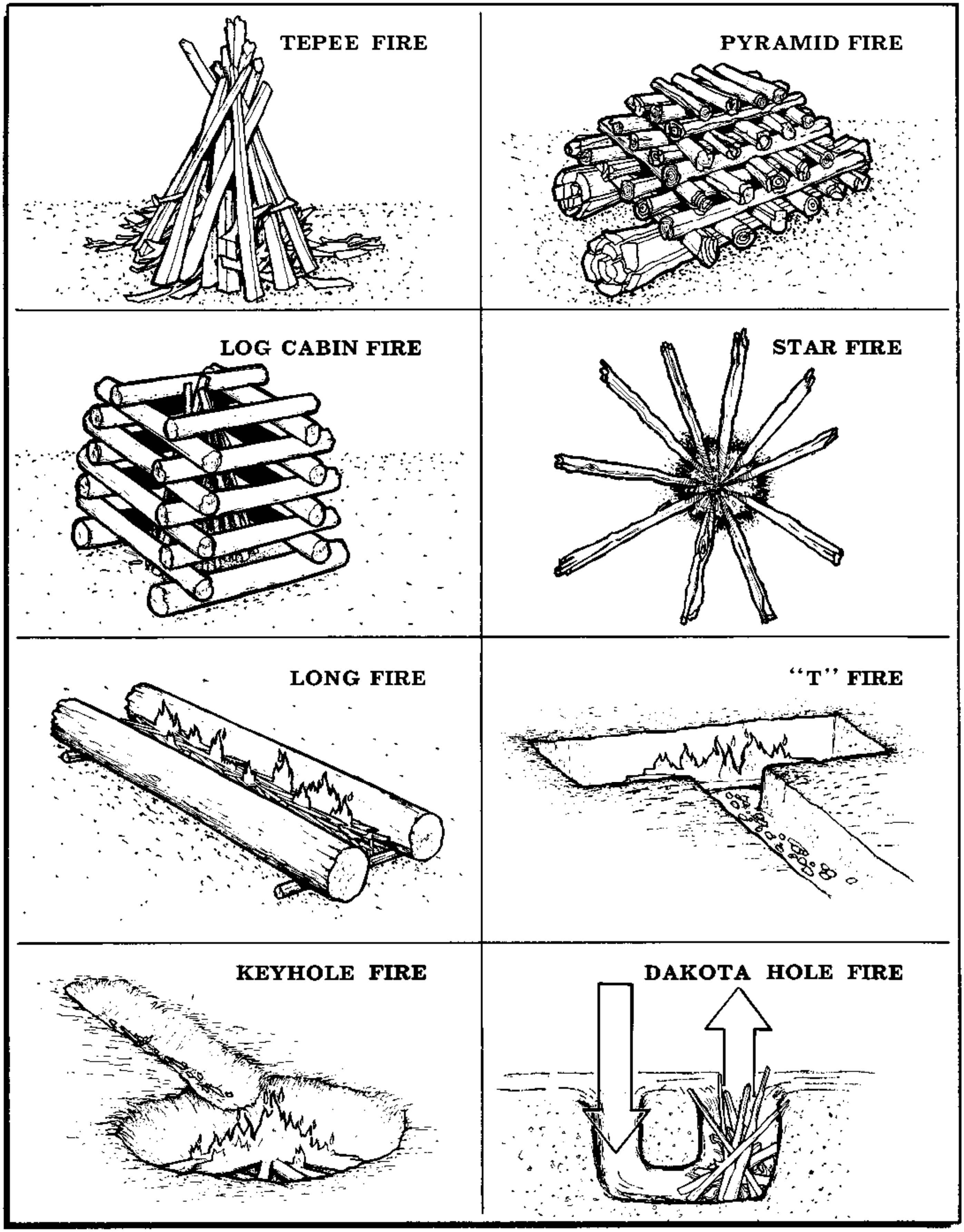
TEPEE FIRE
PYRAMID FIRE
LOG CABIN FIRE
STAR FIRE
LONG FIRE
"T" FIRE
KEYHOLE FIRE
DAKOTA HOLE FIRE

Figure 16-14. Fire Lays.

Chapter 17

EQUIPMENT

17-1. Introduction. Survivors in a survival situation have needs which must be met—food, water, clothing, shelters, etc. The survival kit contains equipment which can be used to satisfy these needs. Quite often, however, this equipment may not be available due to damage or loss. This chapter will address the care and use of issued equipment and improvising the needed equipment when not available. The uses of some issued items are covered in appropriate places throughout this regulation. The care and use of equipment (not covered elsewhere) will be addressed here.

17-2. Types of Kits:

a. All survival kits contain two types of equipment—mandatory and optional. The mandatory equipment for survival kits are:

(1) One-man liferaft (1 each).
(2) Compass (1 each).
(3) Smoke and illumination flares (2 each).
(4) Signal mirror (1 each).
(5) Hand-held launched flare (1 each).
(6) First aid kit (1 each).
(7) Survival radio (1 each).

NOTE: These items of equipment may not be mandatory for raft kit.

b. Optional items are authorized by the major air commands; this authority is delegated to subordinate commanders. These optional items are directly related to climatic conditions and the type of terrain which is being flown over. There are over 40 optional items. Here are a few examples:

(1) Sleeping bag.
(2) Strobelight with lenses.
(3) Wire saw.
(4) Water container.
(5) Survival shovel.
(6) Matchbox container.

17-3. Issued Equipment. Survival equipment is designed to aid survivors throughout their survival episode. To maintain its effectiveness, the equipment must be well cared for.

a. Electronic Equipment:

(1) Electronic signaling devices are by far the survivors' most important signaling devices. Therefore, it is important for survivors to properly care for them to ensure their continued effectiveness. In cold temperatures, the electronic signaling devices must be kept warm to prevent the batteries from becoming cold soaked.

(2) In a cold environment, if survivors speak directly into the microphone, the moisture from their breath may condense and freeze on the microphone, creating communication problems.

(3) Caution must be used when using the survival radios in a cold environment. If the radio is placed against the side of the face to communicate, frostbite could result.

(4) In a wet environment, survivors should make every effort to keep their electronic signaling devices dry. In an open-sea environment, the only recourse may be to shake the water out of the microphone before transmitting.

b. Firearms:

(1) A firearm is a precision tool. It will continue functioning only as long as it is cared for. Saltwater, perspiration, dew, and humidity can all corrode or rust a firearm until it is inoperable. If immersed in saltwater, the survivor should wash the parts in freshwater and then dry and oil them. As an expedient, one way to dry the firearm is to place it in boiling water and after removal, wipe off the excess moisture. The residual heat will evaporate most of the remaining moisture. Survivors should not use uncontrolled heat to dry the firearm as heat over 250°F can remove the temper from the springs in a short time and weaken the action.

(2) Any petroleum-based lubricants used in cold environments will stiffen or freeze causing the firearm to become inoperative. It would be better to thoroughly clean the firearm and remove all lubricant. Metal becomes brittle from cold and is, therefore, prone to breakage.

(3) A firearm was not intended for use as a club, hammer, or pry bar. To use it for any purpose other than for which it was designed, would only result in damage to the firearm.

c. Cutting Tools:

(1) A file and sharpening stone are often packed in a survival kit. The file is normally used for axes, and the stone is normally used for knives.

(2) An old axiom states that a sharp cutting tool is a safe cutting tool. Control of a cutting tool is easier to maintain if it is sharp, and the possibility of accidental injury is reduced.

(3) One of the most valuable items in any survival situation is a knife, since it has a large number of uses. Unless the knife is kept sharp, however, it falls short of its potential.

(4) A knife should be sharpened only with a stone as repeated use of a file rapidly removes steel from the blade. In some cases, it may be necessary to use a file to remove plating from the blade before using the stone.

(5) One of two methods should be used to sharpen a knife. One method is to push the blade down the stone

in a slicing motion. Then turn the blade over and draw the blade toward the body (figure 17-1).

(6) The other method is to use a circular motion the entire length of the blade; turn the blade over and repeat the process. What is done to one side of the cutting edge should also be done to the other to maintain an even cutting edge (figure 17-2).

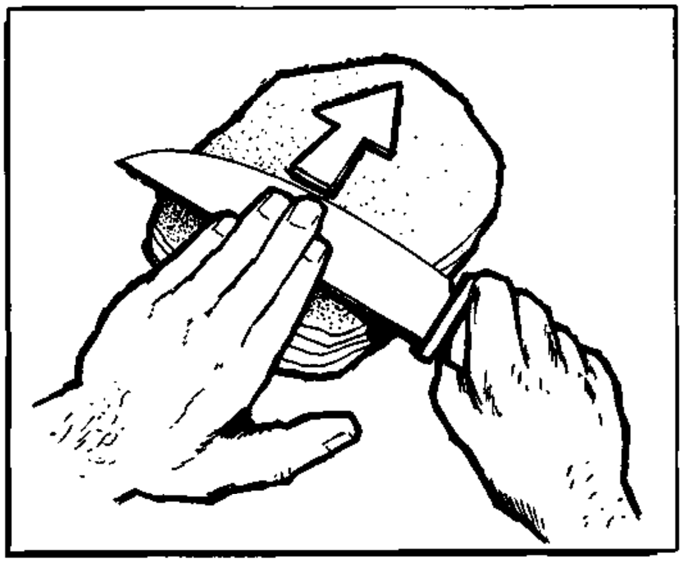

Figure 17-1. Knife Sharpening (Draw).

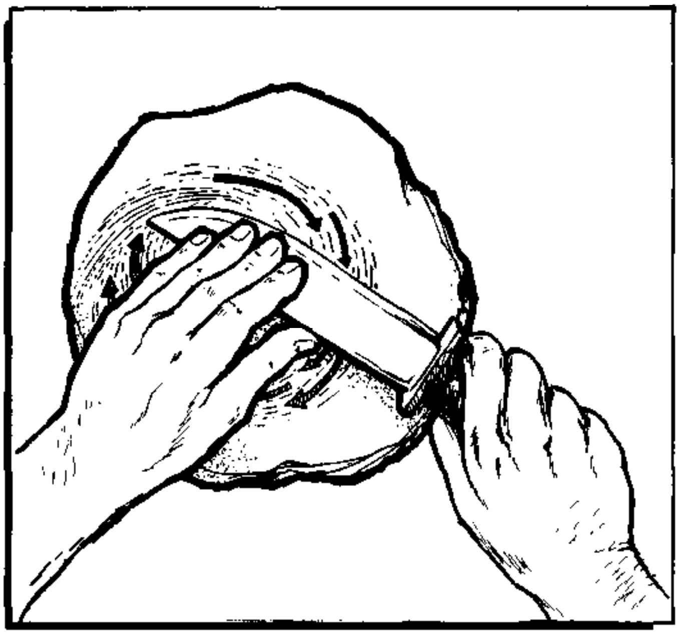

Figure 17-2. Knife Sharpening (Circular).

(7) Most sharpening stones available to survivors will be whetstones. Water should be applied to these stones. The water will help to float away the metal removed by sharpening and make cleaning of the stone easier.

(8) If a commercial whetstone is not available, a natural whetstone can be used. Any standstone will sharpen tools, but a gray, clay-like sandstone gives better results. Quartzite should be avoided. Survivors can recognize quartzite instantly by scratching the knife blade with it—the quartz crystals will bite into steel. If no sandstone is available, granite or crystalline rock can be used. If granite is used, two pieces of the stone should be rubbed together to smooth the surface before use.

(9) As with a knife, a sharp axe will save time and energy and be much safer.

(10) A file should be used on an axe or hatchet. Survivors should file away from the cutting edge to prevent injury if the file should slip. The file should be worked from one end of the cutting edge to the other. The opposite side should be worked to the same degree. This will ensure that the cutting edge is even. After using a file, the stone may be used to hone the axe blade (figure 17-3).

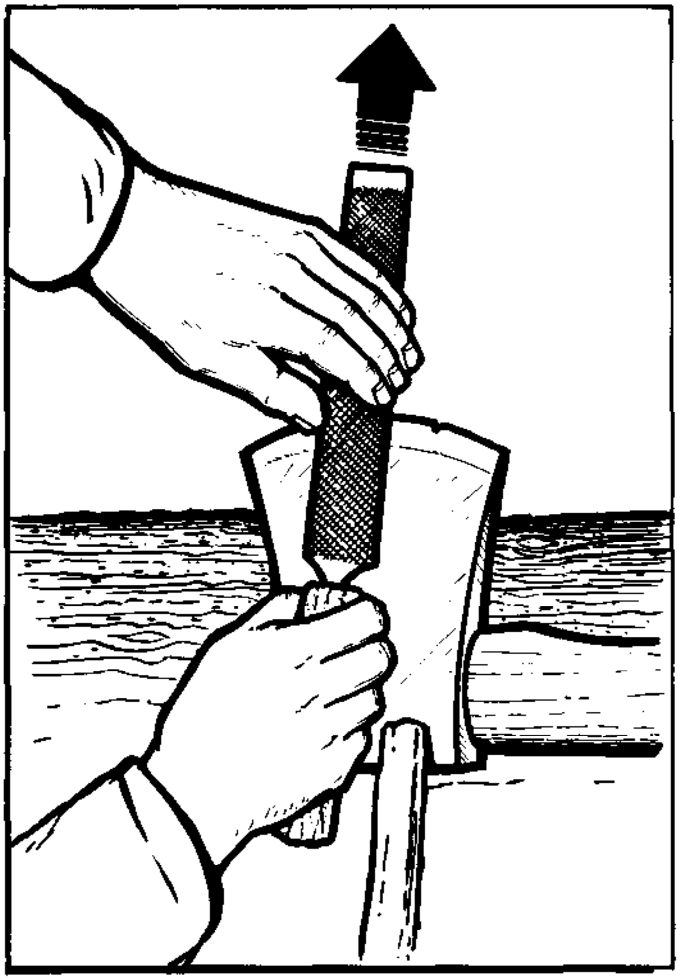

Figure 17-3. Sharpening Axe.

(11) When using an axe, don't try to cut through a tree with one blow. Rhythm and aim are more important than force. Too much power behind a swing interferes with aim. When the axe is swung properly, its weight provides all the power needed.

(12) Carving a new axe handle and mounting the axe head takes a great deal of time and effort. For this reason, a survivor should avoid actions which would require the handle to be changed. Using aim and paying attention to where the axe falls will prevent misses which could result in a cracked or broken handle. Survivors should not use an axe as a pry bar and should avoid leaving the axe out in cold weather where the handle may become brittle.

(13) A broken handle is difficult to remove from the head of the axe. Usually the most convenient way is to burn it out (figure 17-4). For a single-bit axe, bury the bit in the ground up to the handle, and build a fire over it. For a double-bit, a survivor should dig a small trench, lay the middle of the axe head over it, cover both "bits" with earth, and build the fire. The covering of earth keeps the flame from the cutting edge of the axe and saves its temper. A little water added to the earth will further ensure this protection.

(14) When improvising a new handle, a survivor can save time and trouble by making a straight handle instead of a curved one like the original. Survivors should use a young, straight piece of hardwood without knots. The wood should be whittled roughly into shape and finished by shaving. A slot should be cut into the axe-head end of the handle. After it is fitted, a thin, dry wooden wedge can then be pounded into the slot. Survivors should use the axe awhile, pound the wedge in again, then trim it off flush with the axe. The handle must be smoothed to remove splinters. The new handle can be seasoned to prevent shrinkage by "scorching" it in the fire.

d. Whittling:

(1) Whittle means to cut, trim, or shape (a stick or piece of wood) by taking off bits with a knife. Survivors should be able to use the techniques of whittling to help save time, energy, and materials as well as to prevent injuries. They will find that whittling is a necessity in constructing triggers for traps and snares, shuttles and spacers, and other improvised equipment.

(2) When whittling, survivors must hold the knife firmly and cut away from the body (figure 17-5). Wood should be cut with the grain. Branches should be trimmed as shown in figure 17-6.

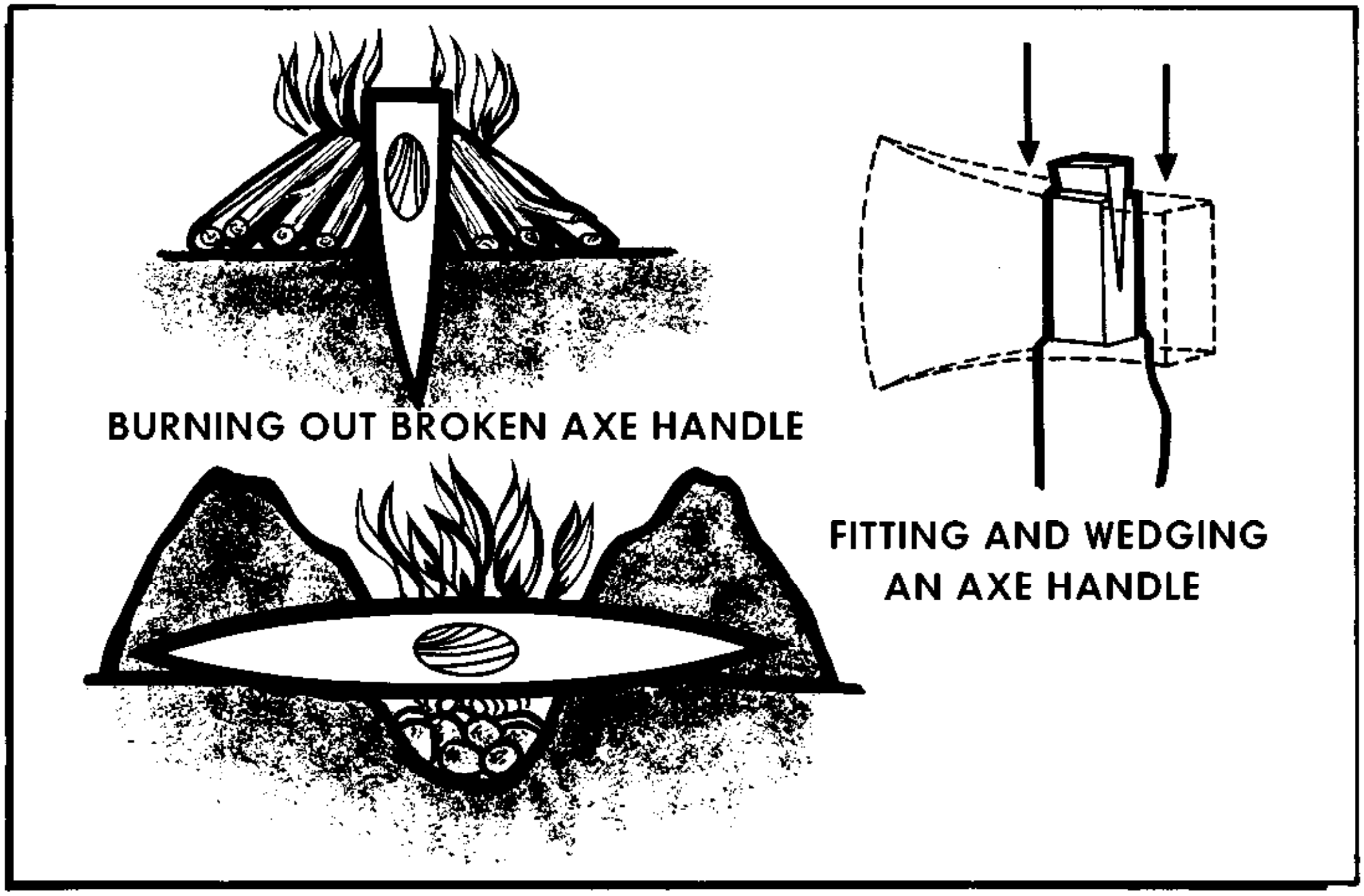
BURNING OUT BROKEN AXE HANDLE
FITTING AND WEDGING AN AXE HANDLE

Figure 17-4. Removing Broken Axe Handle.

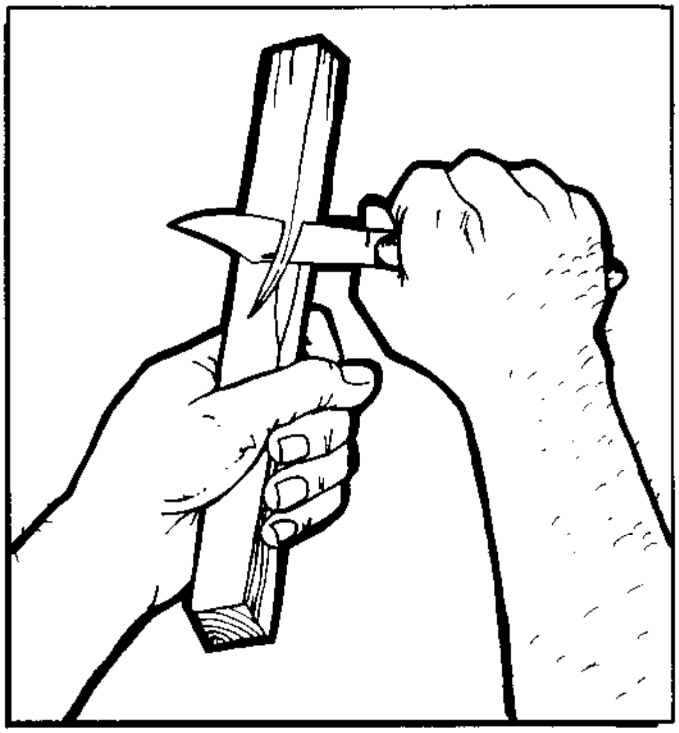

Figure 17-5. Whittling.

(3) To cut completely through a piece of wood, a series of V-cuts should be made all the way around as in figure 17-7. Once the piece of wood has been severed, the pointed end can then be trimmed.

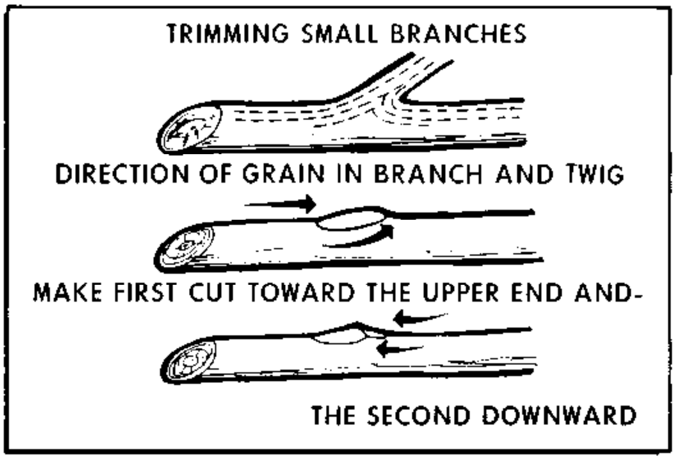
TRIMMING SMALL BRANCHES
DIRECTION OF GRAIN IN BRANCH AND TWIG
MAKE FIRST CUT TOWARD THE UPPER END AND-
THE SECOND DOWNWARD

Figure 17-6. Trimming Branches.

(4) The thumb can be used to help steady the hand. Be sure and keep the thumb clear of the blade. To maintain good control of the knife, the right hand is steadied with the right thumb while the left thumb pushes the blade forward (figure 17-8). This method is very good for trimming.

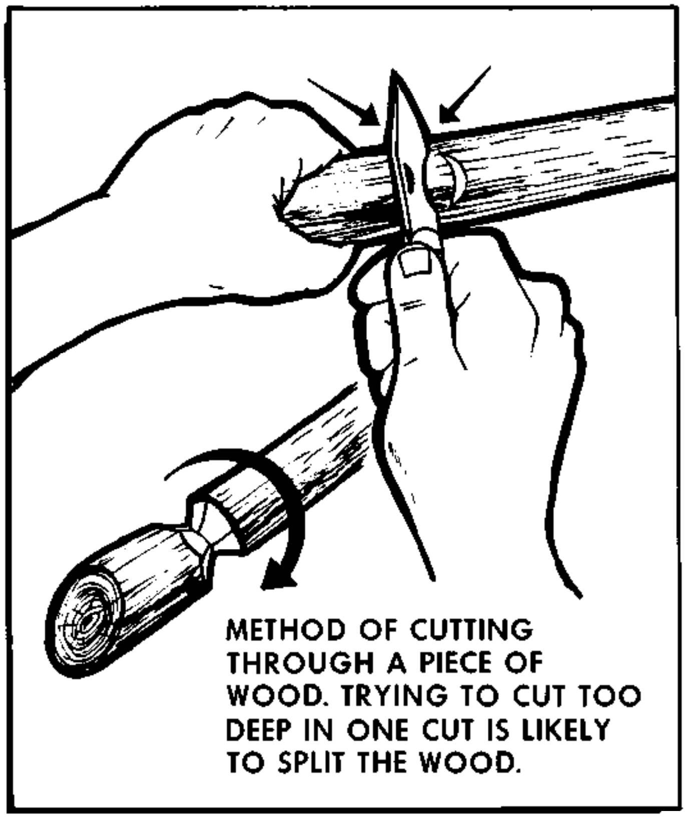

METHOD OF CUTTING THROUGH A PIECE OF WOOD. TRYING TO CUT TOO DEEP IN ONE CUT IS LIKELY TO SPLIT THE WOOD.

Figure 17-7. Cutting Through a Piece of Wood.

e. Felling Trees:

(1) To fell a tree, the survivor must first determine the direction in which the tree is to fall. It is best to fell the tree in the direction in which it is leaning. The lean of the tree can be found by using the axe as a plumb line

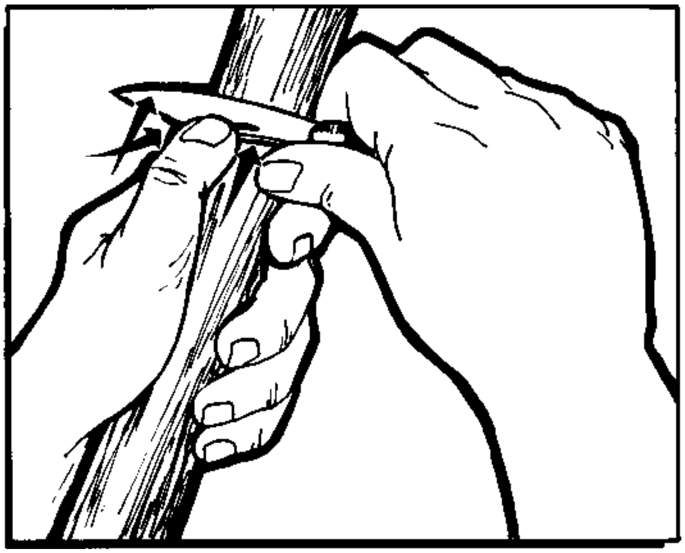

Figure 17-8. Fine Trimming.

(figure 17-9). The survivor should then clear the area around the tree from underbrush and overhanging branches to prevent injury (figure 17-10).

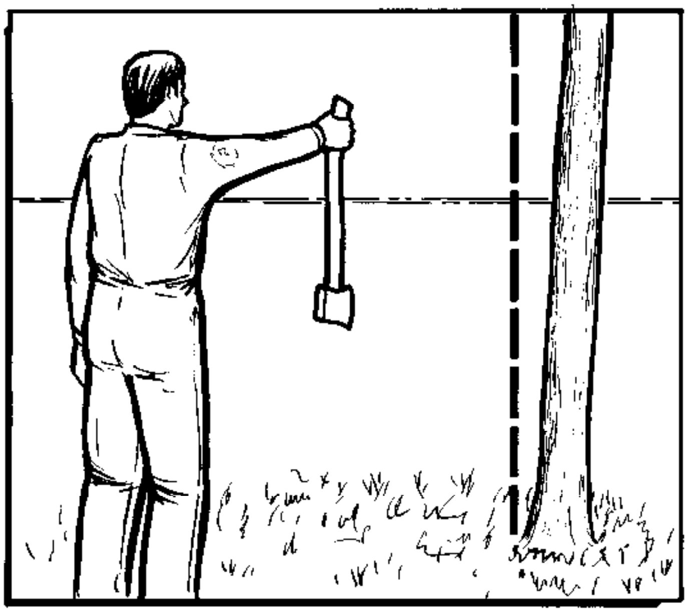

Figure 17-9. Using the Axe as a Plumb Line.

(2) The survivor should make two cuts. The first cut should be on the leaning side of the tree and close to the ground and the second cut on the opposite side and a little higher than the first cut (figure 17-11).

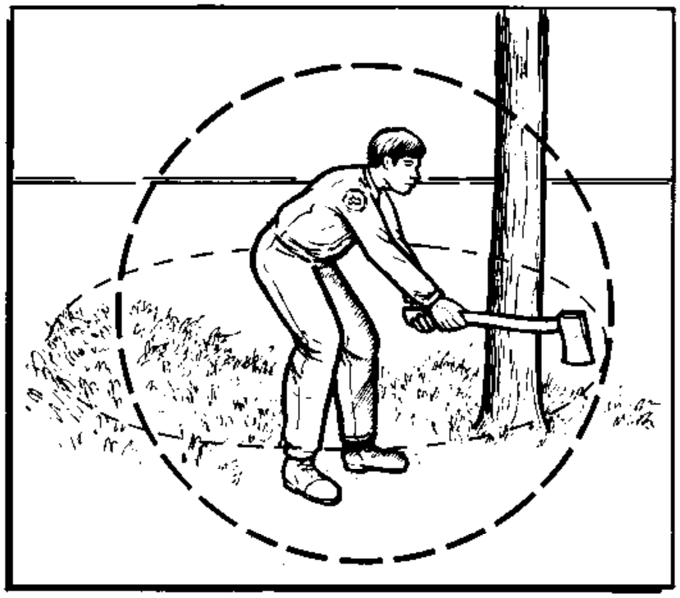

Figure 17-10. Clearing Brush from Cutting Area.

(3) Falling trees often kick back and can cause serious injury (figure 17-12), so survivors must ensure they have a clear escape route. When limbing a tree, start at the base of the tree and cut toward the top. This procedure will allow for easier limb removal and results in a smoother cut. For safety, the survivor should stand on one side of the trunk with the limb on the other.

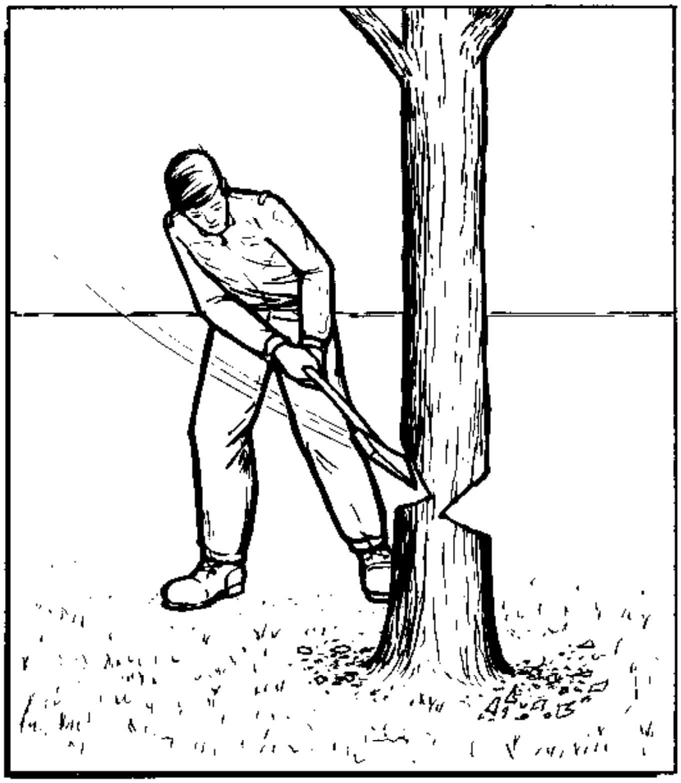

Figure 17-11. Felling Cuts.

(4) To prevent damage to the axe head and possible physical injury, any splitting of wood should be done on a log as in figure 17-13. The log can also be used for cutting sticks and poles (figure 17-14).

(5) To make cutting of a sapling easier, bend it over with one hand, straining grain. A slanting blow close to the ground will cut the sapling (figure 17-15).

17-4. Improvised Equipment:

a. If issued equipment is inoperative, insufficient, or nonexistent, survivors will have to rely upon their ingenuity to manufacture the needed equipment. Survivors must determine whether the need for the item outweighs the work involved to manufacture it. They will also have to evaluate their capabilities. If they have injuries, will the injuries prevent them from manufacturing the item(s)?

b. Undue haste may not only waste materials, but also waste the survivors' time and energy. Before manufacturing equipment, they should have a plan in mind.

c. The survivors' equipment needs may be met in two different ways. They may alter an existing piece of equipment to serve more than one function, or they may also construct a new piece of equipment from

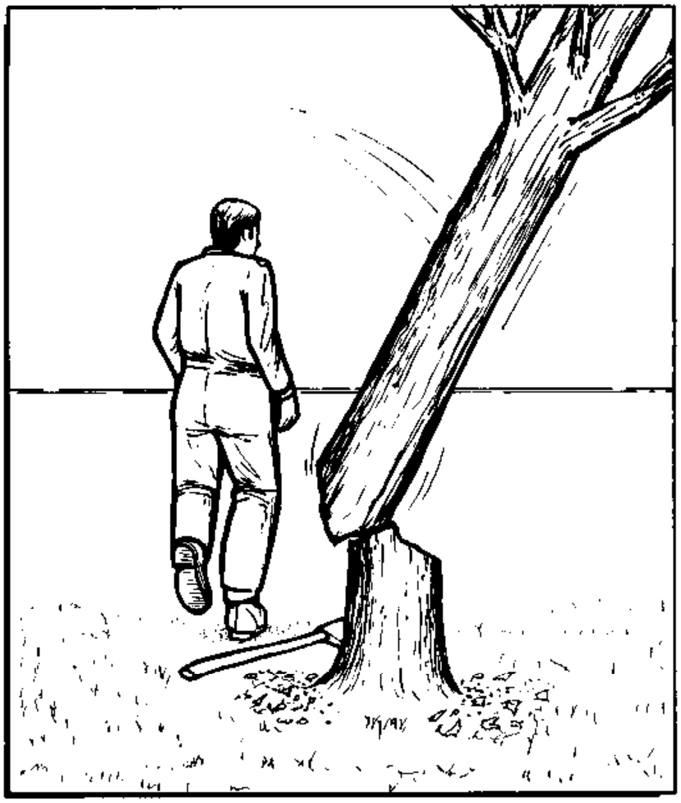

Figure 17-12. Tree Kickbacks.

available materials. Since the items survivors can improvise are limited only by their ingenuity, all improvised items cannot be covered in this regulation.

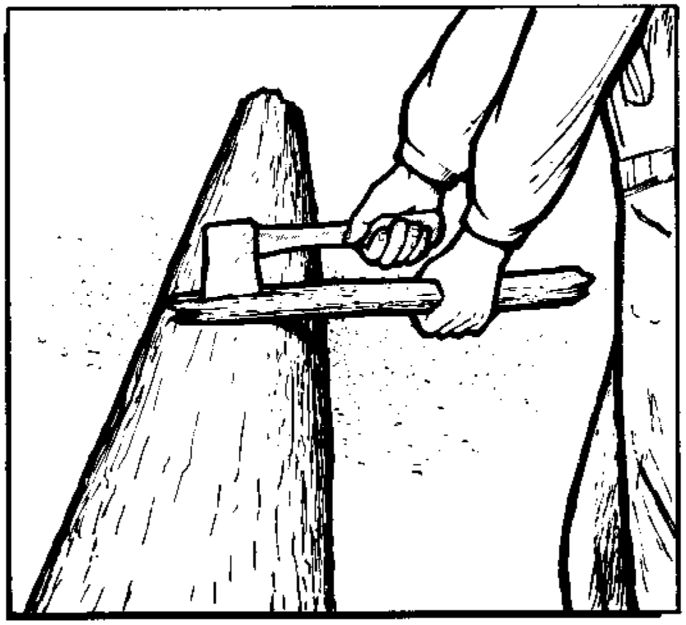

Figure 17-13. Splitting Wood.

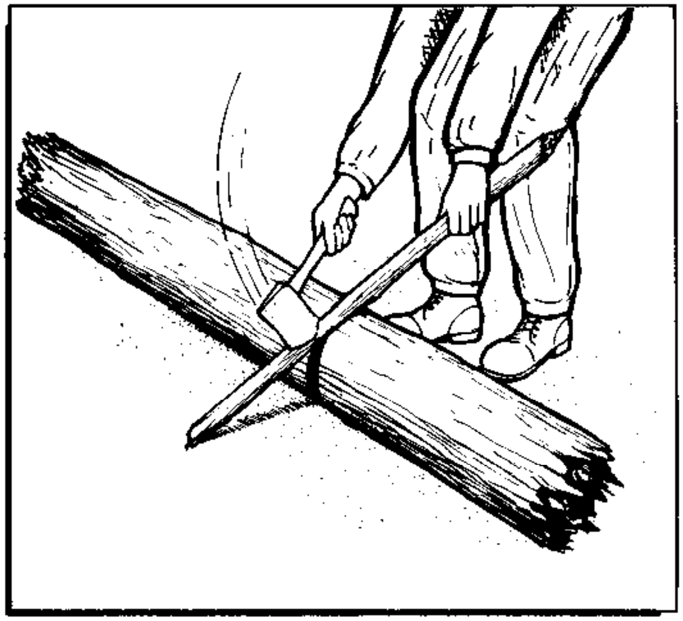

Figure 17-14. Cutting Poles.

d. The methods of manufacturing the equipment referred to in this regulation are only ideas and do not have to be strictly adhered to. Many Air Force survivors have a parachute. This device can be used to improvise a variety of needed equipment items.

e. The parachute consists of (figure 17-16):

(1) The pilot chute which deploys first and pulls the rest of the parachute out.

(2) The parachute canopy which consists of the apex (top) and the skirt or lower lateral band. The canopy material is divided by radial seams into 28 sections called gores. Each gore measures about 3 feet at the skirt and tapers to the apex. Each gore is further subdivided into four sections called panels. The canopy is normally divided into four colors. These colored areas are intended to aid the survivor in shelter construction, signaling, and camouflage.

(3) Fourteen suspension lines connect the canopy material to the harness assembly. Each piece of suspension line is 72 feet long from riser to riser and 22 feet long from riser to skirt and 14 feet from skirt to apex. The tensile strength of each piece of suspension line is 550 pounds. Each piece of suspension line contains seven to nine pieces of innercore with a tensile strength of 35 pounds. The harness assembly contains risers and webbing, buckles, snaps, "D" rings, and other hardware which can be used in improvisation.

f. The whole parachute assembly should be considered as a resource. Every piece of material and hardware can be used.

(1) To obtain the suspension lines, a survivor should cut them at the risers or, if time and conditions permit, consider disassembling the connector links. Cut

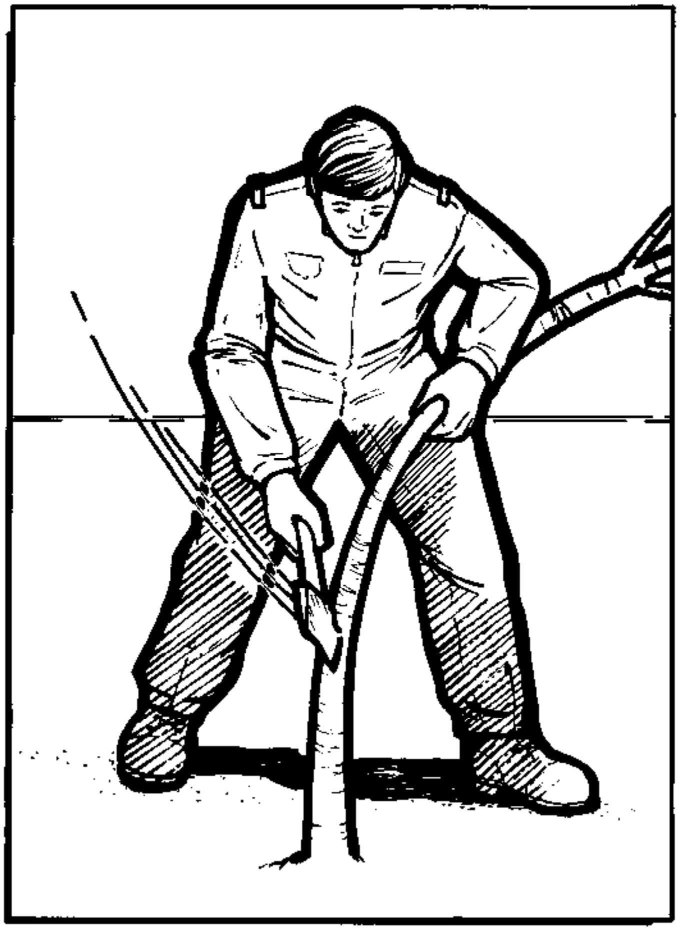

Figure 17-15. Cutting Saplings.

the suspension lines about 2 feet from the skirt of the canopy. When cutting suspension lines or dismantling the canopy/pack assembly, it will be necessary to maintain a sharp knife for safety and ease of cutting.

(2) Survivors should obtain all available suspension line due to its many uses. Even the line within the radial seams of the canopy should be stripped for possible use. The suspension line should be cut above the radial seam stitches next to the skirt end of the canopy (two places). The cut should not go all of the way through the radial seam (figure 17-17). At the apex of the canopy, and just below the radial seam stitching, a horizontal cut can be made and the suspension line extracted. The line can then be cut.

(3) For maximum use of the canopy, survivors must plan its disassembly. The quantity requirements for shelter, signaling, etc., should be thought out and planned for. Once these needs have been determined, the canopy may be cut up. The radial seam must be stretched tightly for ease of cutting. The radial seam can then be cut by holding the knife at an angle and following the center of the seam. With proper tension and the gentle pushing (or pulling) of a sharp blade there will be a controlled splitting of the canopy at the seam (figure 17-17). It helps to secure the apex either to another individual or to an immobile object such as a tree.

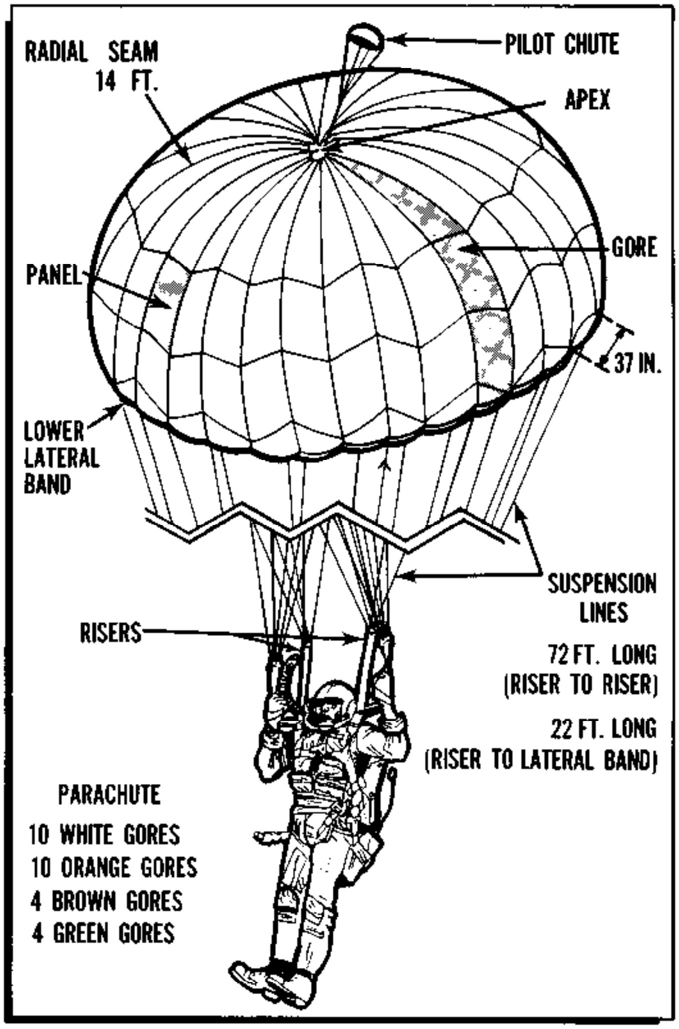

Figure 17-16. Parachute Diagram.

(4) When stripping the harness assembly, the seams of the webbing should be split so the maximum usable webbing is obtained. The harness material and webbing should not be randomly cut as it will waste much needed materials.

g. One requirement in improvising is having available material. Parachute fabric, harness, suspension lines, etc., can be used for clothing. Needles are helpful for making any type of emergency clothing. Wise survivors should always have extra sewing needles hidden somewhere on their person. A good needle or sewing awl can be made from the can-opening key from the ration tin (figure 17-18) or, as the Eskimos do, from a sliver of bone. Thread is usually available in the form of innercore. It will be to the survivor's benefit to collect small objects which may "come in handy." Wire, nails, buttons, a piece of canvas, or animal skin should not be

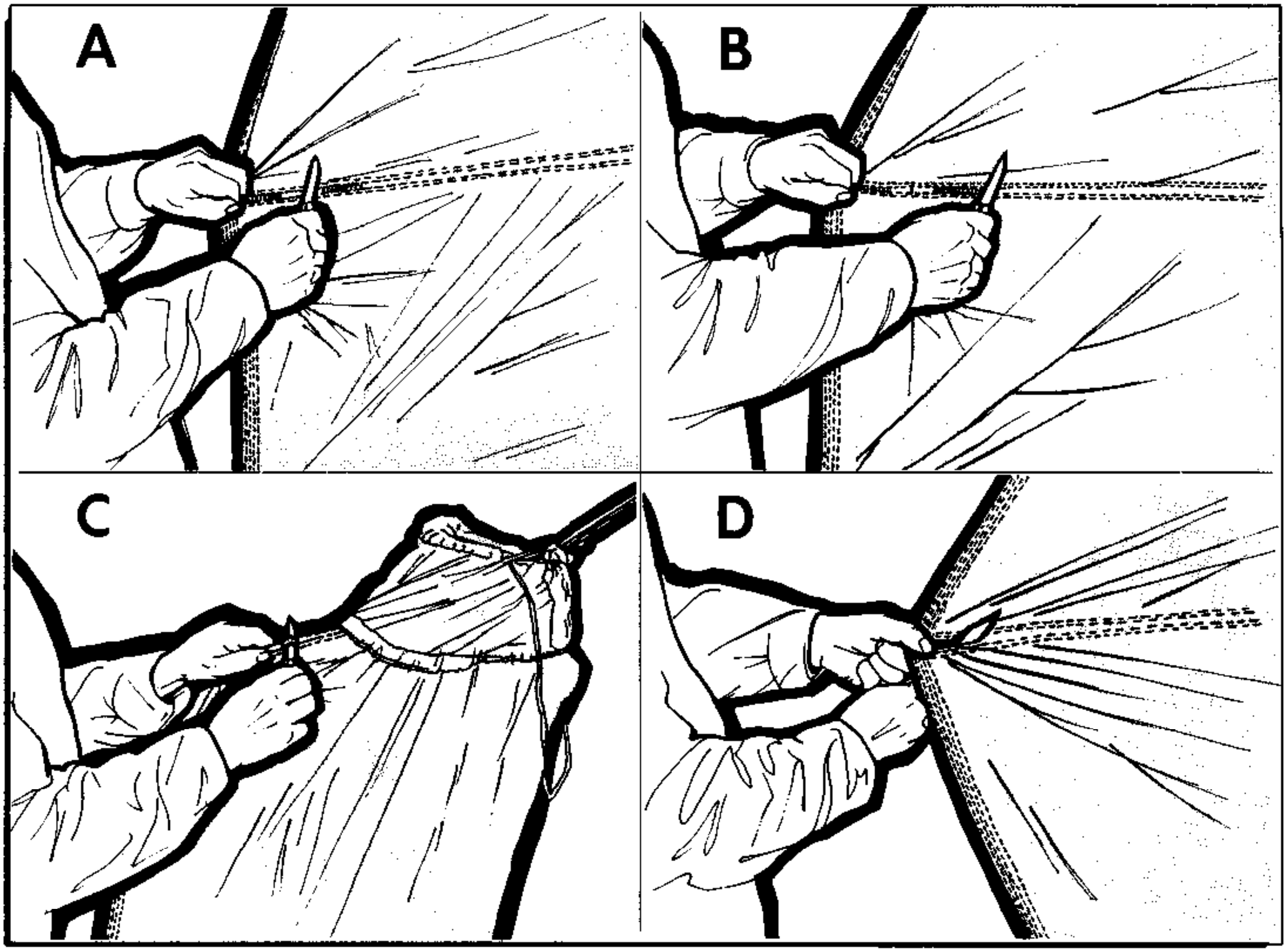

Figure 17-17. Cutting the Parachute.

discarded. Any such object may be worth its weight in gold when placed in a hip pocket or a sewing kit. Any kind of animal skin can be used for making clothing such as gloves or mittens or making a ground cover to keep the sleeping bag dry and clean. Small skins can be used for mending and for boot insoles. Mending and cleaning clothes when possible will pay dividends in health, comfort, and safety.

h. The improvised equipment survivors may need to make will probably involve sewing. The material to be sewn may be quite thick and hard to sew, and to keep from stabbing fingers and hands, a palm-type thimble can be improvised (figure 17-19). A piece of webbing, leather, or other heavy material, with a hole for the thumb, is used. A flat rock, metal, or wood is used as the thimble and this is held in place by a doughnut-shaped piece of material sewn onto the palm piece. To use, the end of the needle with the eye is placed on the thimble and the thimble is then used to push the needle through the material to be sewn.

17-5. Miscellaneous Improvised Equipment:

a. Improvised Trail-Type Snowshoes. The snowshoe frame can be made from a sapling 1 inch in diameter and 5 feet long. The sapling should be bent and spread to 12 inches at the widest point. The survivor can then include the webbing of suspension lines (figure 17-20). The foot harness, for attaching the snowshoe to the boot, is also fashioned from suspension line.

b. Improvised Bear Paw-Type Snowshoes. A sapling can be held over a heat source and bent to the shape shown in figure 17-21. Wire from the aircraft or parachute suspension line can be used for lashing and for making webbing. Snowshoes can also be quickly improvised by cutting a few pine boughs and lashing them together at the cut ends. The lashed boughs positioned with the cut ends forward can then be tied to the feet (figure 17-22).

(1) Survivors should guard against frostbite and blistering while snowshoeing. Due to the design of the

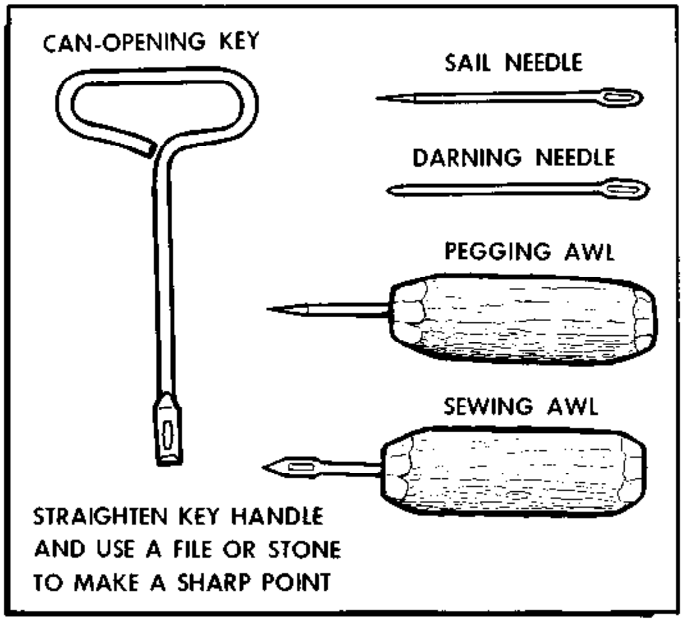
CAN-OPENING KEY
SAIL NEEDLE
DARNING NEEDLE
PEGGING AWL
SEWING AWL
STRAIGHTEN KEY HANDLE AND USE A FILE OR STONE TO MAKE A SHARP POINT

Figure 17-18. Needle and Sewing Awl.

harness, the circulation of the toes is usually restricted, and the hazard of frostbite is greater. They should check the feet carefully, stop often, take off the harness, and massage the feet when they seem to be getting cold.

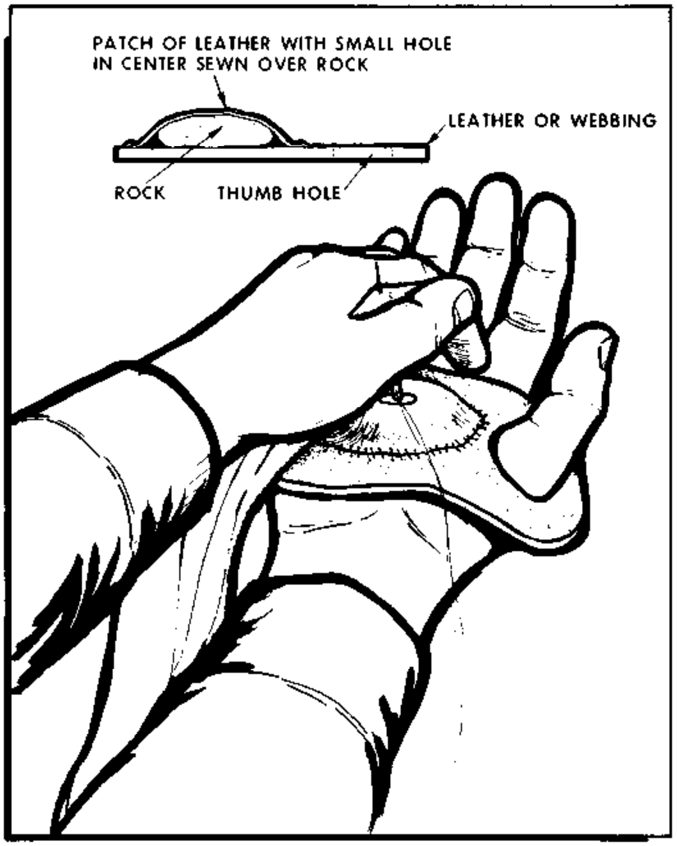
PATCH OF LEATHER WITH SMALL HOLE IN CENTER SEWN OVER ROCK
LEATHER OR WEBBING
ROCK
THUMB HOLE

Figure 17-19. Palm Thimble.

(2) Blistering between the toes or on the ball of the foot is sometimes unavoidable in a "tenderfoot" if much snowshoeing is done. To make blisters less likely, the survivor should keep socks and insoles dry and change them regularly.

c. Sleeping Bag. Immediate action should be to use the whole parachute until conditions allow for improvising. A sleeping bag can be improvised by using four gores of parachute material or an equivalent amount of other materials (figure 17-23). The material should be folded in half lengthwise and sewn at the foot. To measure the length, the survivor should allow an extra 6 to 10 inches in addition to the individual's height. The two raw edges can then be sewn together. The two sections of the bag can be filled with cattail down, goat's beard lichen, dry grass, insulation from aircraft walls, etc. The

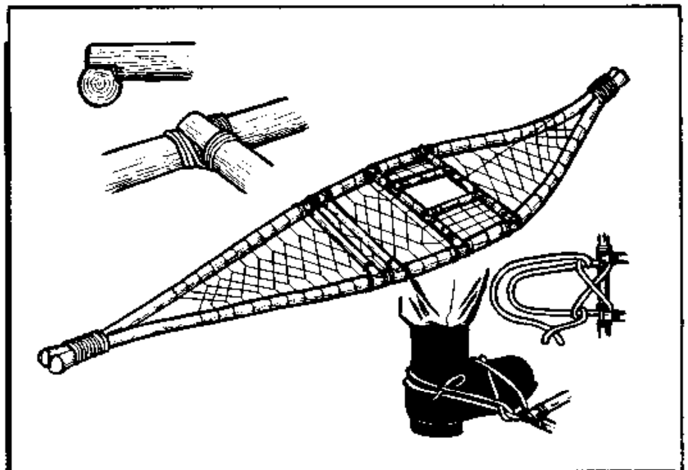

CANADIAN EMERGENCY SNOWSHOES

1. Select 6 poles 6 feet long (individuals height), ¾ inch (thumb size) at the base, ¼ inch (little finger size) at the tip. Cut 6 sticks approximately 10 inches long and ¾ inches wide and tie them in the following manner:

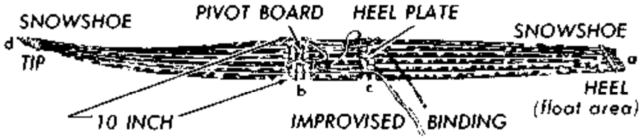
SNOWSHOE
PIVOT BOARD
HEEL PLATE
SNOWSHOE
d
TIP
a
HEEL (float area)
b
c
10 INCH
IMPROVISED BINDING

a. Lash one stick to the snowshoe float area (cut off excess).

b. Lash three sticks forward of the center of the shoe to form the pivot board. This position of the pivot board allows the float to remain on the snow and causes the tip to rise when walking.

c. Lash two sticks where your heel strikes the snowshoe to form the heel plate.

d. Tie the snowshoe tips together.

2. The snowshoe binding must be secured to the snowshoe so that the survivor's foot can pivot when walking.

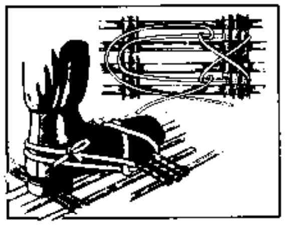

Binding — make as shown from continuous length of split harness webbing or from suspension lines (braided lines preferred).

Figure 17-20. Improvised Trail Snowshoes.

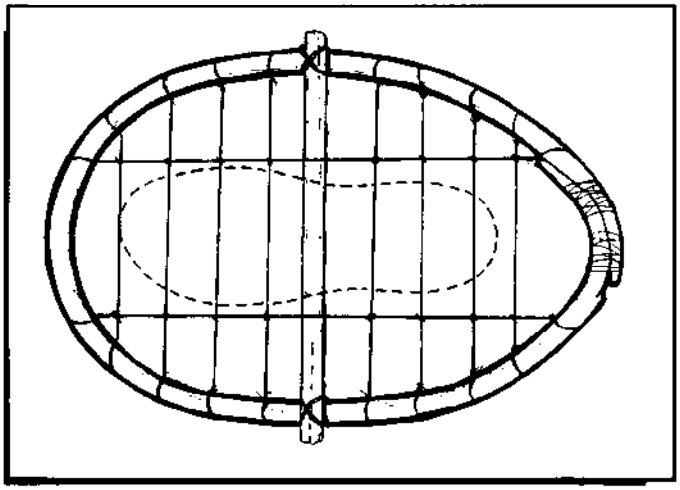

Figure 17-21. Improvised Bear Paws.

stuffed sleeping bag should then be quilted to keep the insulation from shifting. The bag can be folded in half lengthwise and the foot and open edges sewn. The length and width can be adjusted for the individual.

d. Insulating Bed:

(1) In addition to the sleeping bag, some form of ground insulation is advisable. An insulation mat will help insulate the survivor from ground moisture and the cold. Any nonpoisonous plants such as ferns and grasses will suffice. Leaves from a deciduous tree make a comfortable bed. If available, extra clothing, seat cushions, aircraft insulation, rafts, and parachute material may be used. In a coniferous forest, boughs from the trees would do well if the bed is constructed properly.

(2) The survivor should start at the foot of the proposed bed and stick the cut ends in the ground at about a 45-degree angle and very close together. The completed bed should be slightly wider and longer than the body. If the ground is frozen, a layer of dead branches can be used on the ground with the green boughs placed in the dead branches, similar to sticking them in the ground.

(3) A bough bed should be a minimum of 12 inches thick before use. This will allow sufficient insulation between the survivor and the ground once the bed is compressed. The bough bed should be fluffed up and boughs added daily to maintain its comfort and insulation capabilities.

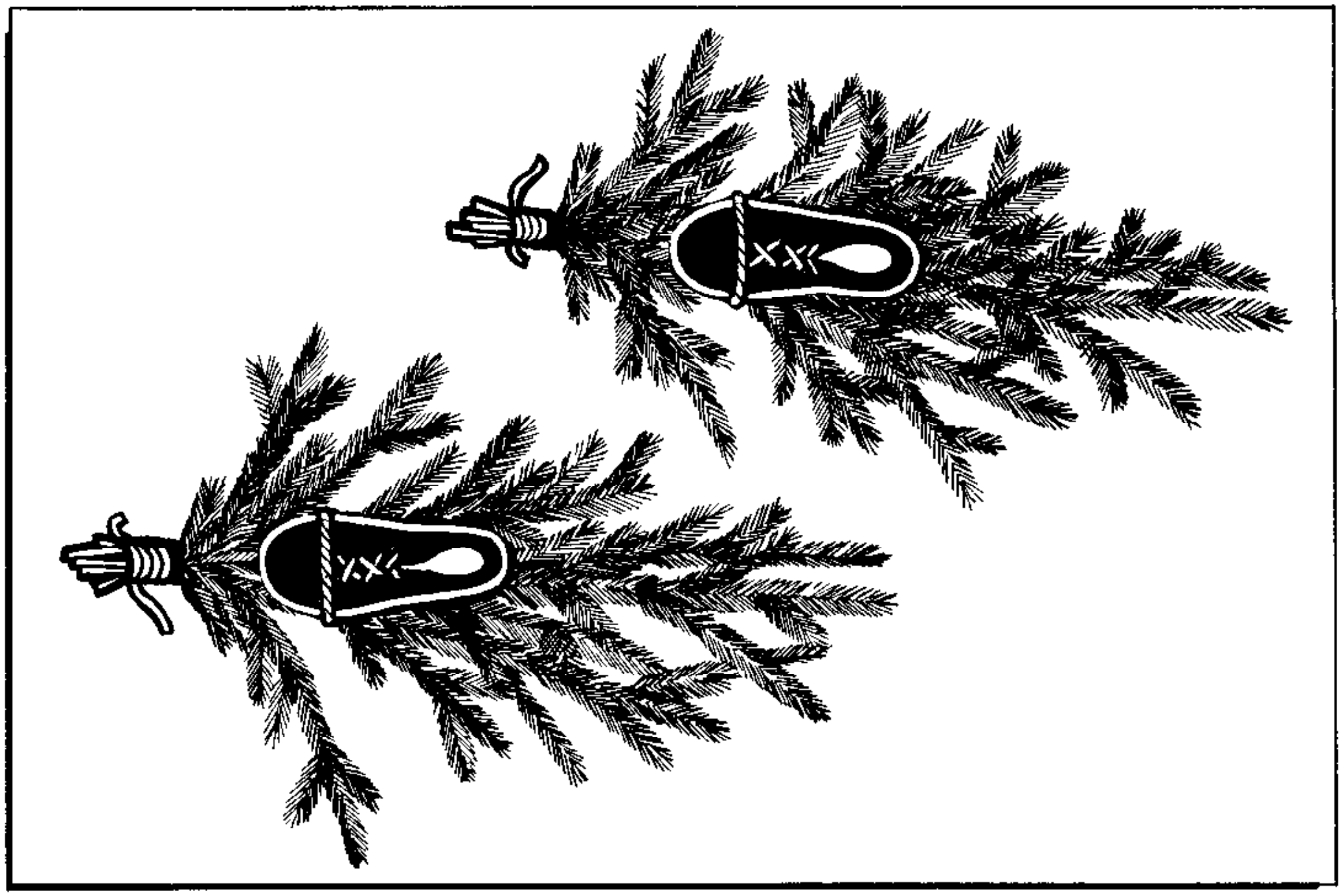

Figure 17-22. Bough Snowshoes.

(4) Spruce boughs have many sharp needles and can cause some discomfort. Also the needles on various types of pines are generally located on the ends of the boughs, and it would take an abundance of pine boughs to provide comfort and insulation. Fir boughs on the other hand, have an abundance of needles all along the boughs and the needles are rounded. These boughs are excellent for beds, providing comfort and insulation (figure 17-24).

e. Rawhide. Rawhide is a very useful material which can be made from any animal hide. Processing it is time consuming but the material obtained is strong and very durable. It can be used for making sheaths for cutting tools, lashing materials, ropes, etc.

(1) The first step in making rawhide is to remove all of the fat and muscle tissue from the hide. The large pieces can be cut off and the remainder scraped off with a dull knife or similar instrument.

(2) The next step is to remove the hair. This can be done by applying a thick layer of wood ashes to the hair side. Ashes from a hardwood fire work best. Thoroughly sprinkle water all over the ashes. This causes lye to leach out of the ash. The lye will remove the hair. The hide should be rolled with the hair side in and stored in a cool place for several days. When the hair begins to slip (check by pulling on the hair), the hide should be unrolled and placed over a log. Remove the hair by scraping it off with a dull knife. Once the hair is removed, the

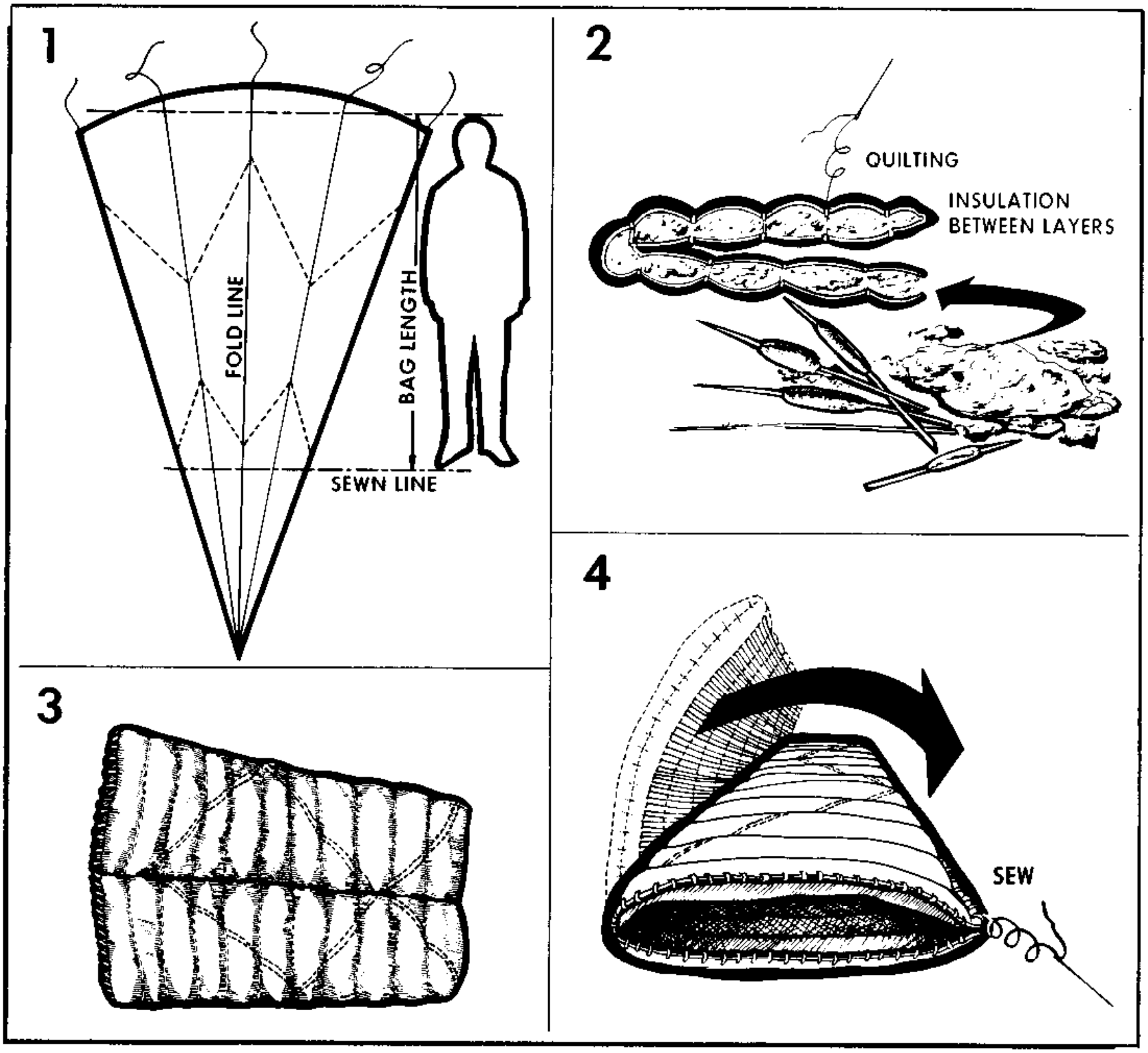

Figure 17-23. Improvised Sleeping Bag.

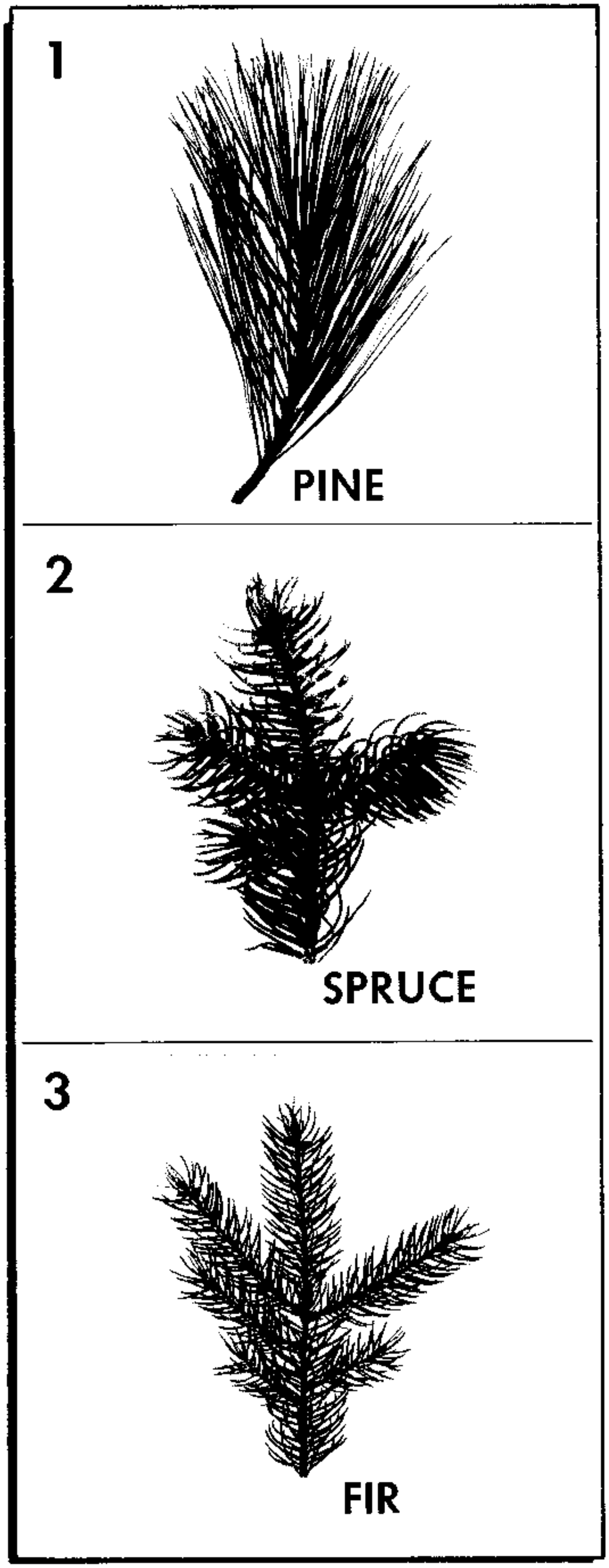

Figure 17-24. Boughs.

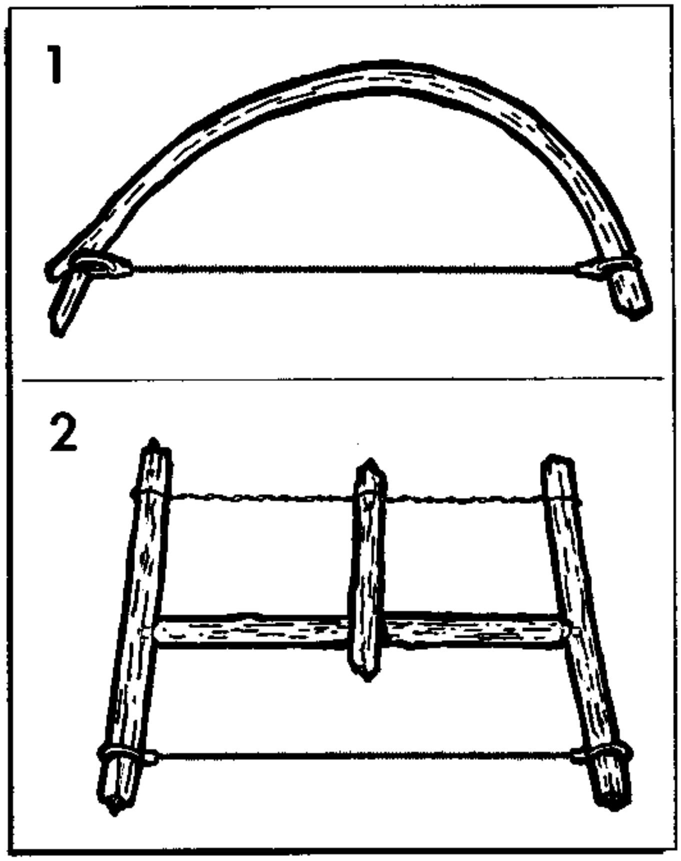

Figure 17-25. Bow Saw and Buck Saw.

hide should be thoroughly washed, stretched inside a frame, and allowed to dry slowly in the shade. When dry, rawhide is extremely hard. It can be softened by soaking in water.

f. Wire Saws. Wire or pieces of metal can be used to replace broken issued saws. With minor modifications, the survivor can construct a usable saw. A bow-saw

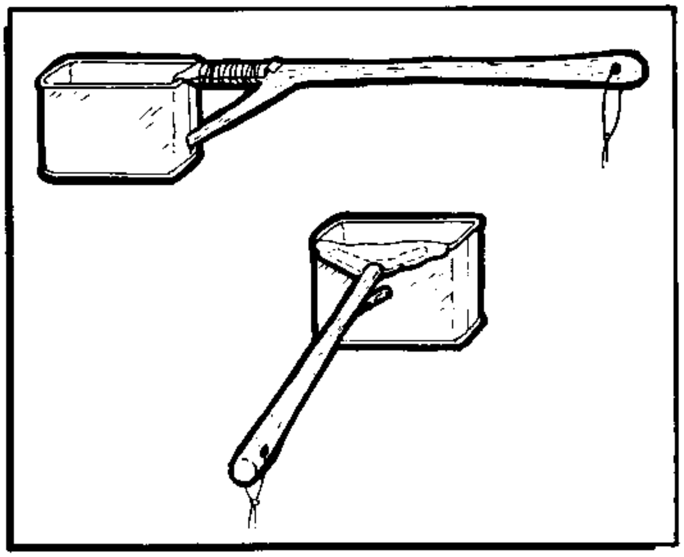

Figure 17-26. Cooking Utensils.

arrangement will help to prevent the blade from flexing. A green sapling may be used for the bow as shown in figure 17-25. If a more durable saw is required and time permits, a bucksaw may be improvised (figure 17-25). Blade tension can be maintained by use of a tightening device known as a "windlass" (figure 17-25).

g. Cooking Utensils. Ration tins can serve as adequate cooking utensils. If the end has been left intact as in figure 17-26, use a green stick long enough to prevent burning the hand while cooking. If the side has been left intact, a forked stick may be used to add support to the container (figure 17-26).

17-6. Ropes and Knots:

a. Basic Knowledge of Tying a Knot. A basic knowledge of correct rope and knot procedures will aid the survivor to do many necessary actions. Such actions as improvising equipment, building shelters, assembling packs, and providing safety devices require the use of proven techniques. Tying a knot incorrectly could result in ineffective improvised equipment, injury, or death.

b. Rope Terminology: (See figure 17-27.)

(1) Bend. A bend (called a knot in this regulation) is used to fasten two ropes together or to fasten a rope to a ring or loop.

(2) Bight. A bight is a bend or U-shaped curve in a rope.

(3) Hitch. A hitch is used to tie a rope around a timber, pipe, or post so that it will hold temporarily but can be readily untied.

(4) Knot. A knot is an interlacement of the parts of bodies, as cordage, forming a lump or knot or any tie or fastening formed with a cord, rope, or line, including bends, hitches, and splices. It is often used as a stopper to prevent a rope from passing through an opening.

(5) Line. A line (sometimes called a rope) is a single thread, string, or cord.

(6) Loop. A loop is a fold or doubling of the rope through which another rope can be passed. A temporary loop is made by a knot or a hitch. A permanent loop is made by a splice or some other permanent means.

(7) Overhand Turn or Loop. An overhand loop is made when the running end passes over the standing part.

(8) Rope. A rope (often called a line) is made of strands of fiber twisted or braided together.

(9) Round Turn. A round turn is the same as a turn, with running end leaving the circle in the same general direction as the standing part.

(10) Running End. The running end is the free or working end of a rope.

(11) Standing End. The standing end is the balance of the rope, excluding the running end.

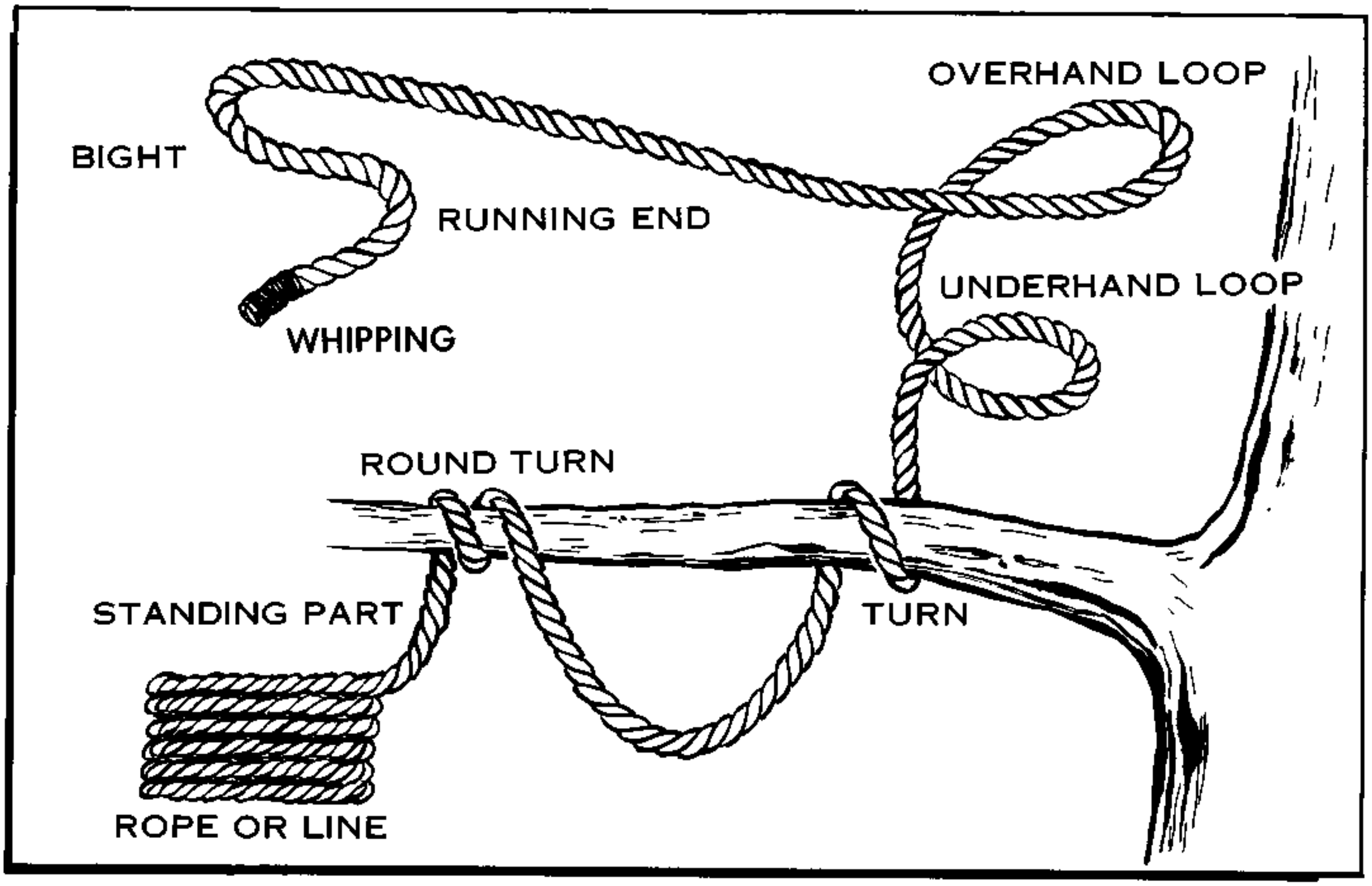

Figure 17-27. Elements of Ropes and Knots.

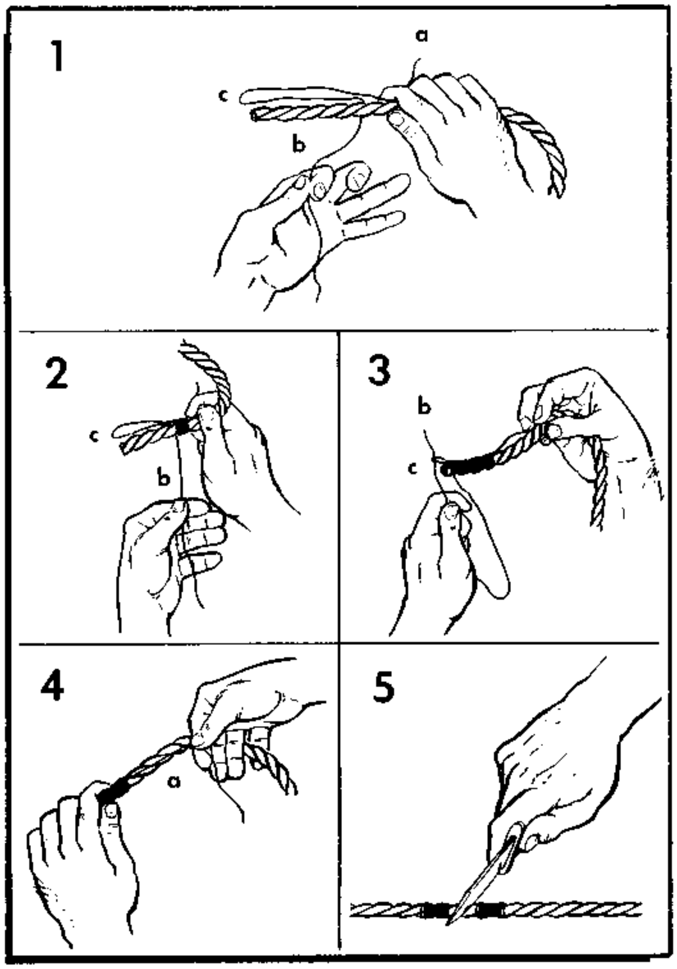

Figure 17-28. Whipping the End of a Rope.

(12) Turn. A turn describes the placing of a rope around a specific object such as a post, rail, or ring with the running end continuing in the opposite direction from the standing end.

(13) Underhand Turn or Loop. An underhand turn or loop is made when the running end passes under the standing part.

c. Whipping the Ends of a Rope. The raw, cut end of a rope has a tendency to untwist and should always be knotted or fastened in some manner. Whipping is one method of fastening the end of the rope. This method is particularly satisfactory because it does not increase the size of the rope. The whipped end of a rope will still thread through blocks or other openings. Before cutting a rope, place two whippings on the rope 1 or 2 inches apart and make the cut between the whippings (figure 17-28-5). This will prevent the cut ends from untwisting immediately after they are cut. A rope is whipped by wrapping the end tightly with a small cord. Make a bight near one end of the cord and lay both ends of the small cord along one side of the rope (figure 17-28-1). The bight should project beyond the end of the rope about one-half inch. The running end (b) of the cord should be wrapped tightly around the rope and cord (figure 17-28-2) starting at the end of the whipping which will be farthest from the end of the rope. The wrap should be in the same direction as the twist of the rope strands. Continue wrapping the cord around the rope, keeping it tight, to within about one-half inch of the end. At this point, slip the running end (b) through the bight of the cord (figure 17-28-3). The standing part of the cord (a) can then be pulled until the bight of the cord is pulled under the whipping and cord (b) is tightened (figure 17-28-4). The ends of cord (a and b) should be cut at the edge of the whipping, leaving the rope end whipped.

d. Knots at End of the Rope:

(1) Overhand Knot. The overhand knot (figure 17-29) is the most commonly used and the simplest of all knots. An overhand knot may be used to prevent the end of a rope from untwisting, to form a knot at the end of a rope, or as a part of another knot. To tie an overhand knot, make a loop near the end of the rope and pass the running end through the loop, pulling it tight.

(2) Figure-Eight Knot. The figure-eight knot (figure 17-30) is used to form a larger knot than would be formed by an overhand knot at the end of a rope. A figure-eight knot is used in the end of a rope to prevent the ends from slipping through a fastening or loop in another rope. To make the figure-eight knot, make a loop in the standing part, pass the running end around

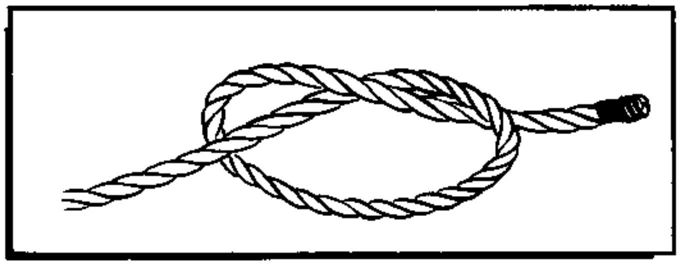

Figure 17-29. Overhand Knot.

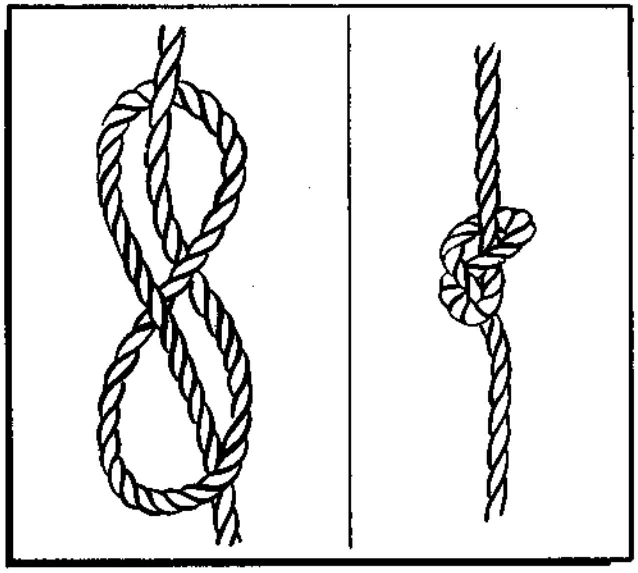

Figure 17-30. Figure-Eight Knot.

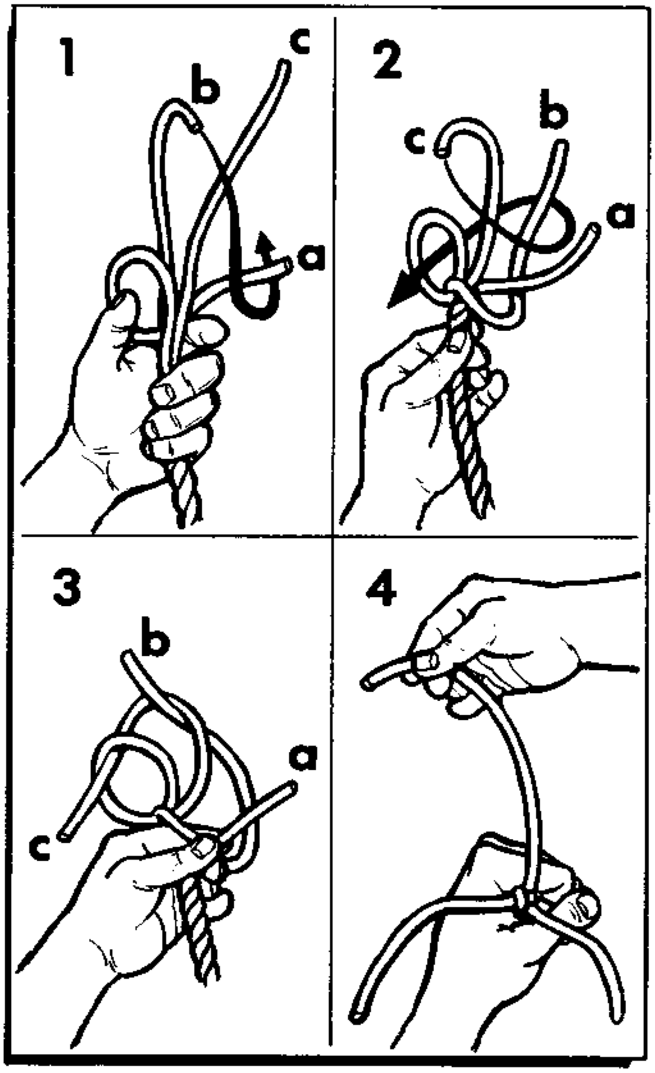

Figure 17-31. Wall Knot.

the standing part back over one side of the loop, and down through the loop. The running end can then be pulled tight.

(3) Wall Knot. The wall knot (figure 17-31) with a crown is used to prevent the end of a rope from untwisting when an enlargement is not objectionable. It also makes a desirable knot to prevent the end of the rope from slipping through small openings, as when rope handles are used on boxes. The crown or the wall knots may be used separately. To make the wall knot, untwist the strands for about five turns of the rope. A loop in strand "a" (figure 17-32-1) should be used and strand "b" brought down (figure 17-31-2) and around strand "a." Strand "c" (figure 17-31-3) can then be brought around strand "b" and through the loop in strand "a." The knot can then be tightened (figure 17-31-4) by grasping the rope in one hand and pulling each strand tight. The strands point up or away from the rope. To make a neat, round knot, the wall knot should be crowned.

(4) Crown on Wall Knot. To crown a wall knot, the end of strand "a" (figure 17-32-1) should be moved between strands "b" and "c." Next strand "c" is passed (figure 17-32-2) between strand "b" and the loop in strand "a." Line "b" is then passed over line "a" and through the bight formed by line "c" (figure 17-32-3). The knots can then be drawn tight and the loose strands cut. When the crown is finished, strands should point down or back along the rope.

e. Knots for Joining Two Ropes:

(1) Square Knot. The square knot (figure 17-33) is used for tying two ropes of equal diameter together to prevent slippage. To tie the square knot, lay the running end of each rope together but pointing in opposite directions. The running end of one rope can be passed under

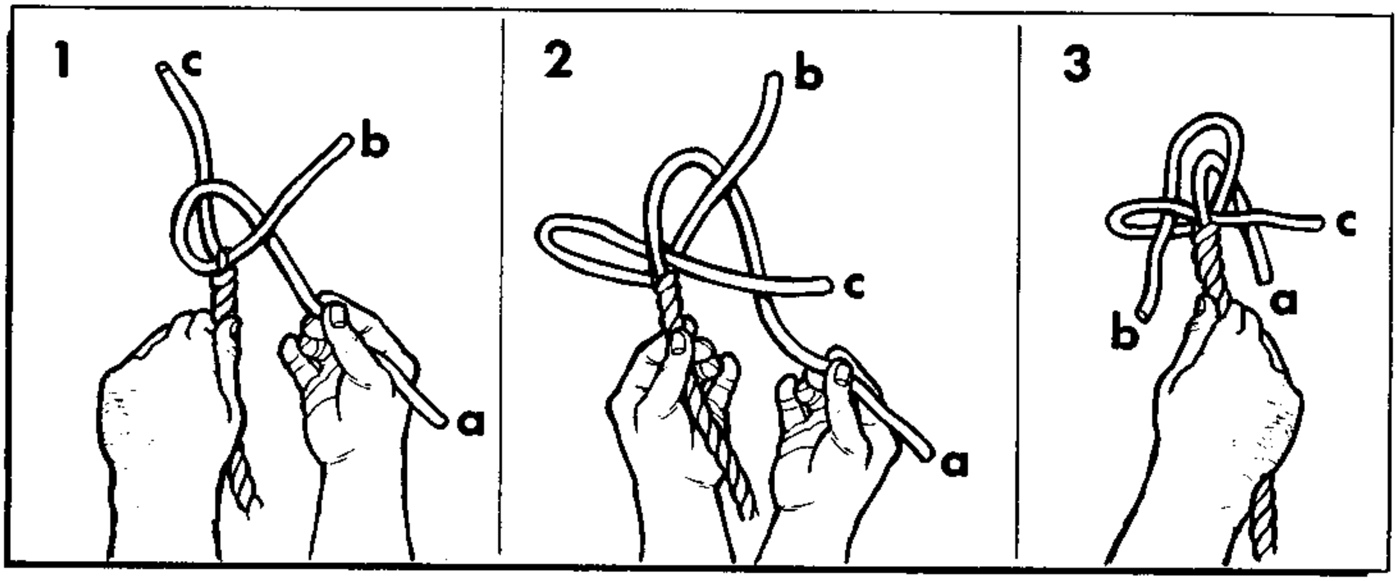

Figure 17-32. Crown on Wall Knot.

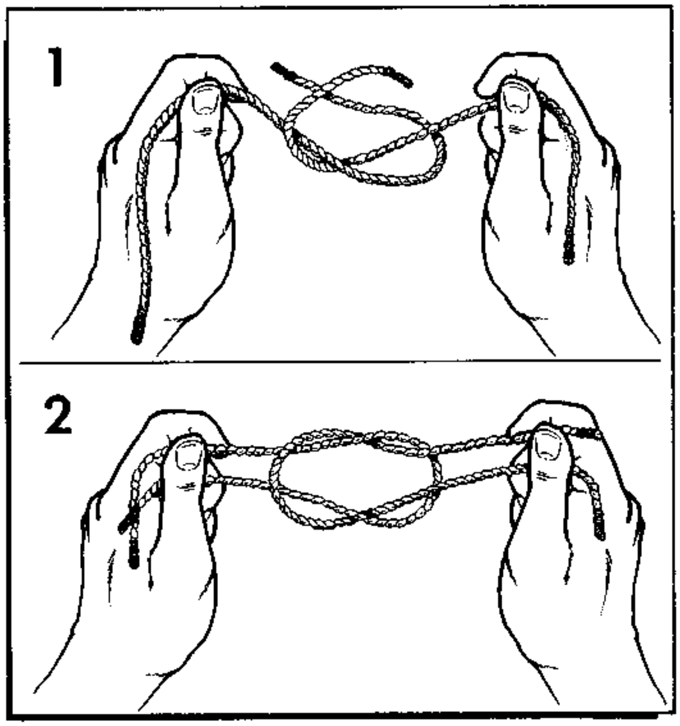

Figure 17-33. Square Knot.

the standing part of the other rope. Bring the two running ends up away from the point where they cross and crossed again (figure 17-33-1). Once each running end is parallel to its own standing part (figure 17-33-2), the two ends can be pulled tight. If each running end does not come parallel to the standing part of its own rope, the knot is called a "granny knot" (figure 17-34-1). Because it will slip under strain, the granny knot should not be used. A square knot can also be tied by making a bight in the end of one rope and feeding the running end of the other rope through and around this bight. The running end of the second rope is routed from the standing side of the bight. If the procedure is reversed, the resulting knot will have a running end parallel to each standing part but the two running ends will not be opposite each other. This knot is called a "thief" knot (figure 17-34-2). It will slip under strain and is difficult to untie. A true square knot will draw tighter under

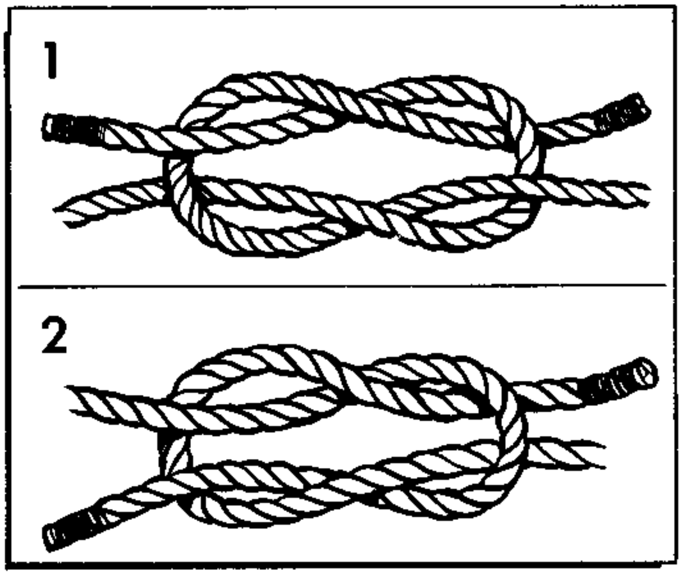

Figure 17-34. Granny and Thief Knots.

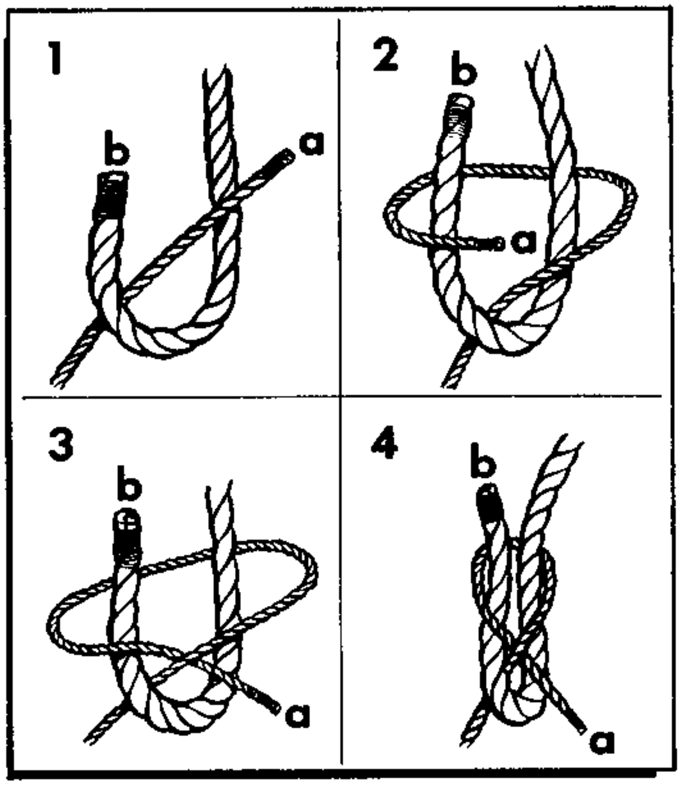

Figure 17-35. Single Sheet Bend.

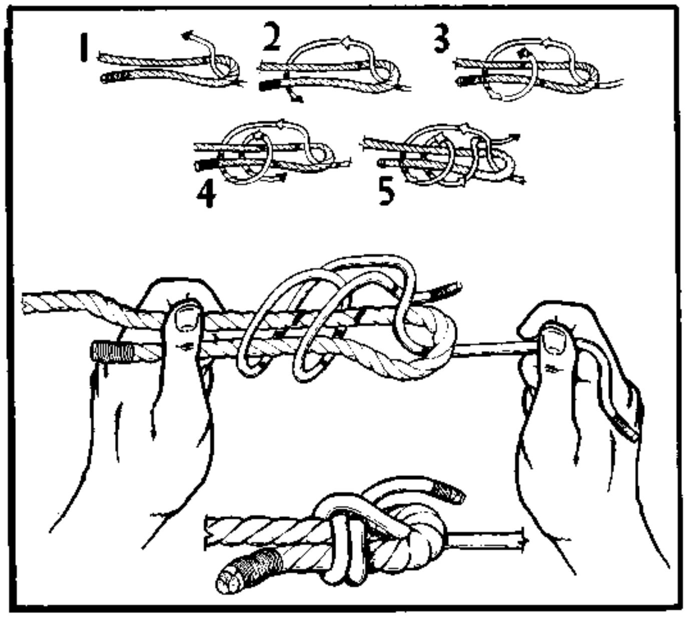

Figure 17-36. Double Sheet Bend.

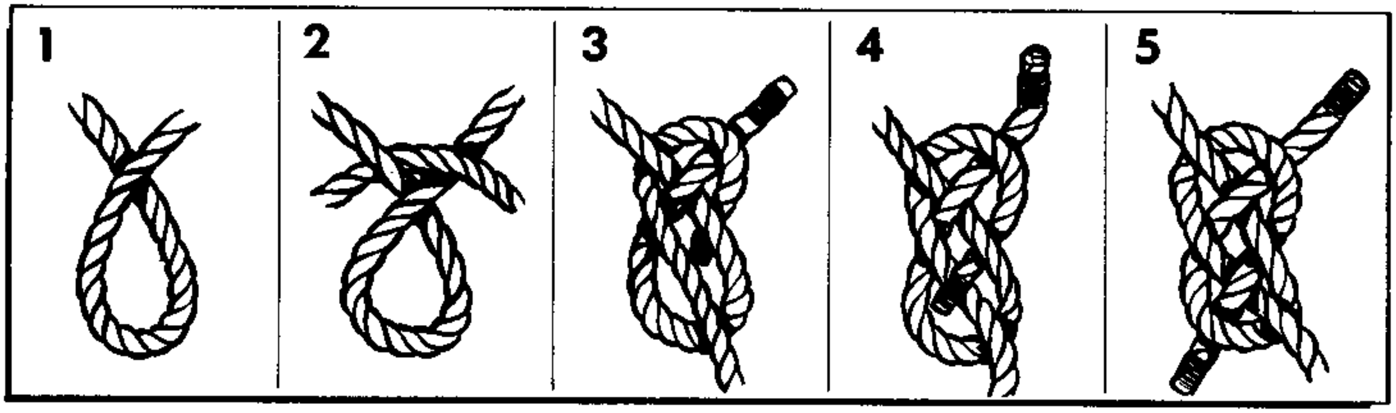

Figure 17-37. Carrick Bend.

strain. A square knot can be untied easily by grasping the bends of the two bights and pulling the knot apart.

(2) Single Sheet Bend. The use of a single sheet bend (figure 17-35), sometimes called a weaver's knot, is limited to tying together two dry ropes of unequal size. To tie the single sheet bend, the running end (a) (figure 17-35-1) of the smaller rope should pass through a bight (b) in the larger rope. The running end should continue around both parts of the larger rope (figure 17-35-2), and back under the smaller rope (figure 17-35-3). The running end can then be pulled tight (figure 17-35-4). This knot will draw tight under light loads but may loosen or slip when the tension is released.

(3) Double Sheet Bend. The double sheet bend (figure 17-36) works better than the single sheet bend for joining ropes of equal or unequal diameter, joining wet ropes, or for tying a rope to an eye. It will not slip or draw tight under heavy loads. To tie a double sheet bend, a single sheet bend is tied first. However, the running end is not pulled tight. One extra turn is taken around both sides of the bight in the larger rope with the running end for the smaller rope. Then tighten the knot.

(4) Carrick Bend. The carrick bend (figure 17-37) is used for heavy loads and for joining thin cable or heavy rope. It will not draw tight under a heavy load. To tie a carrick bend, a loop is formed (figure 17-37-1) in one rope. The running end of the other rope is passed behind the standing part (figure 17-37-2) and in front of the running part of the rope in which the loop has been formed. The running end should then be woven under one side of the loop (figure 17-37-3), through the loop, over the standing part of its own rope (figure 17-37-4), down through the loop, and under the remaining side of the loop (figure 17-37-5).

f. Knots for Making Loops:

(1) Bowline. The bowline (figure 17-38) is a useful knot for forming a loop in the end of a rope. It is also easy to untie. To tie the bowline, the running end (a) of the rope passes through the object to be affixed to the bowline and forms a loop (b) (figure 17-38-1) in the standing part of the rope. The running end (a) is then passed through the loop (figure 17-38-2) from underneath and around the standing part (figure 17-38-3) of the rope, and back through the loop from the top (figure 17-38-4). The running end passes down through the loop parallel to the rope coming up through the loop. The knot is then pulled tight.

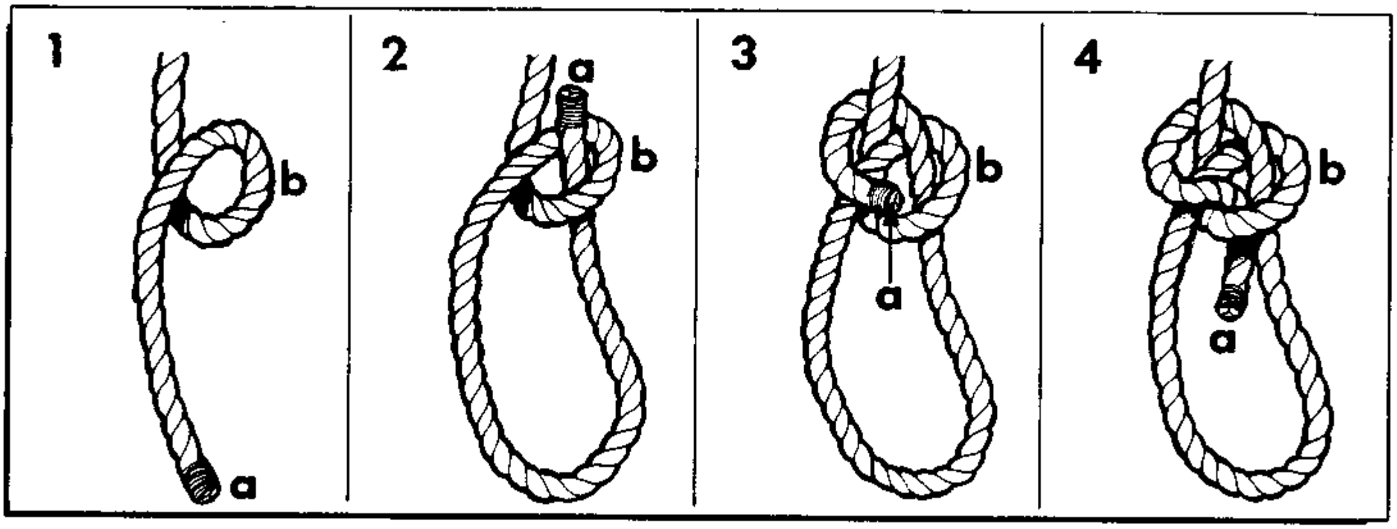

Figure 17-38. Bowline.

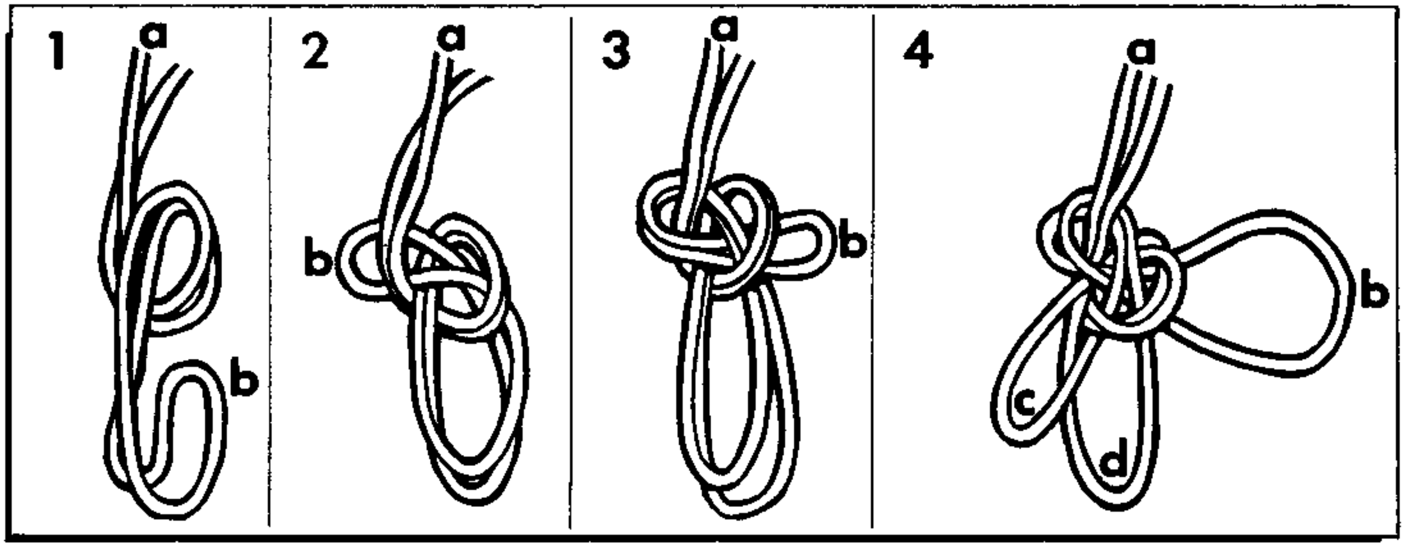

Figure 17-39. Double Bowline.

(2) Double Bowline. The double bowline (figure 17-39) with a slip knot is a rigging used by tree surgeons who work alone in trees for extended periods. It can be made and operated by one person and is comfortable as a sling or boatswain's chair (figure 17-40). A small board with notches as a seat adds to the personal comfort of the user. To tie a double bowline, the running end (a) (figure 17-39) of a line should be bent back about 10 feet along the standing part. The bight (b) is formed as the new running end and a bowline tied as described and illustrated in figure 17-38. The new running end (b) (figure 17-39) or loop is used to support the back and the remaining two loops (c) and (d) support the legs.

(3) Rolling or Magnus Hitch (figure 17-41). A rolling or Magnus hitch is a safety knot designed to make a running end fast to a suspension line with a nonslip grip yet it can be released by hand pressure bending the knot downward. The running end (a) (figure 17-41-1) is passed around the suspension line (b) twice, making two full turns downward (figure 17-41-2). The running end (a) (figure 17-41-3) is then turned upward over the two turns, again around the suspension line, and under itself (figure 17-41-3). This knot is excellent for fastening a rope to itself, a larger rope, a cable, a timber, or a post.

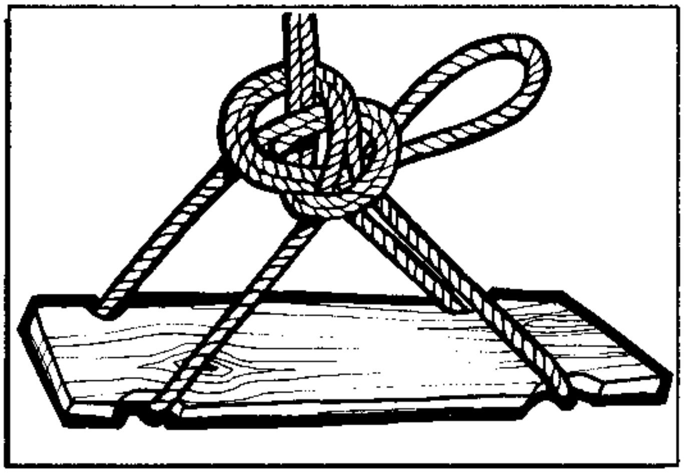

Figure 17-40. Boatswain's Chair.

(4) Running Bowline. The running bowline (figure 17-42) is the basic air transport rigging knot. It provides a sling of the choker type at the end of a single line and is generally used in rigging. To tie a running bowline, make a bight (b) (figure 17-42-1) with an overhand loop (c) made in the running end (a). The running end (a) is passed around the standing part, through the loop (c) (figure 17-42-2), under, then back over the side of the bight, and back through the loop (c) (figure 17-42-3).

(5) Bowline on a Bight. It is sometimes desirable to form a loop at some point in a rope other than at the end. The bowline on a bight (figure 17-43) can be used

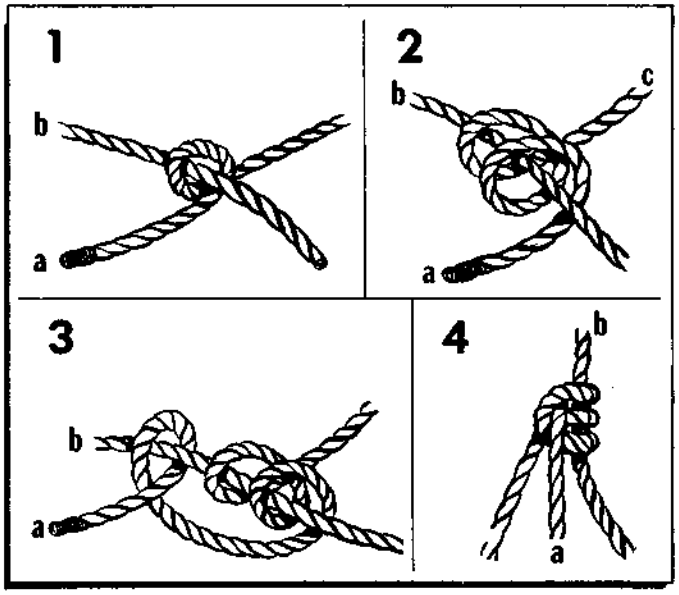

Figure 17-41. Rolling or Magnus Hitch.

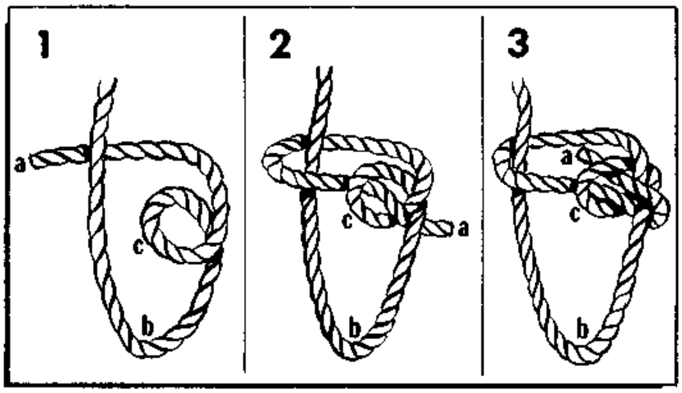

Figure 17-42. Running Bowline.

for this purpose. It is easily untied and will not slip. The same knot can be tied at the end of the rope by doubling the rope for a short section. A doubled portion of the rope is used to form a loop (b) (figure 17-43-1) as in the case of the bowline. The bight end (a) of the doubled portion is passed up through the loop (b), back down (figure 17-43-2), up around the entire knot (figure 17-43-3), and tightened (figure 17-43-4).

(6) Spanish Bowline. A Spanish bowline (figure 17-44) can be tied at any point in a rope, either at a place where the line is doubled or at an end which has been doubled back. The Spanish bowline is used in rescue work or to give a two-fold grip for lifting a pipe or other round object in a sling. To tie the Spanish bowline, a doubled portion of the rope is held in the left hand with the loop up and the center of the loop is turned back against the standing parts to form two loops (figure 17-44-1) or "rabbit ears." The two rabbit ears (c) and (d) (figure 17-44-2) are moved until they partly overlap each other. The top of the loop nearest the person is brought down toward the thumb of the left hand, being sure it is rolled over as it is brought down. The thumb is placed over this loop (figure 17-44-5) to hold it in position. The top of the remaining loop is grasped and brought down, rolling it over and placing it under the thumb. There are now four small loops, (c, d, e, and f) in the rope. The lower left-hand loop (c) is turned one-half turn and inserted from front to back of the upper left-hand loop (e). The lower right-hand loop (d) is turned (figure 17-44-4) and inserted through the upper right-hand loop (f). The two loops (c and d) which have been passed through are grasped and the rope pulled tight (figure 17-44-5).

(7) French Bowline. The French bowline (figure 17-45) is sometimes used as a sling for lifting injured

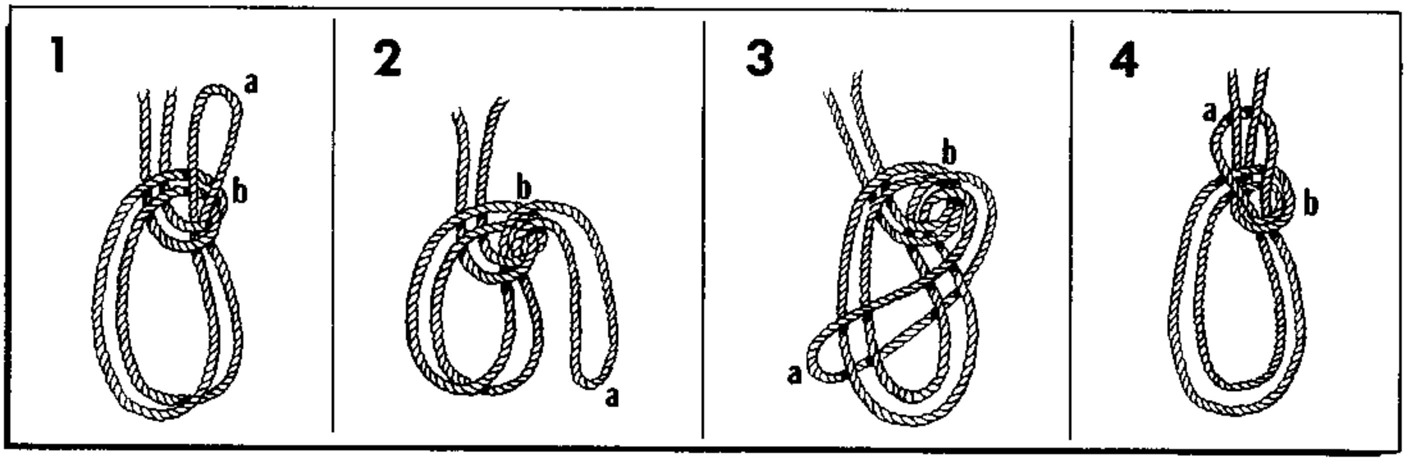

Figure 17-43. Bowline on a Bight.

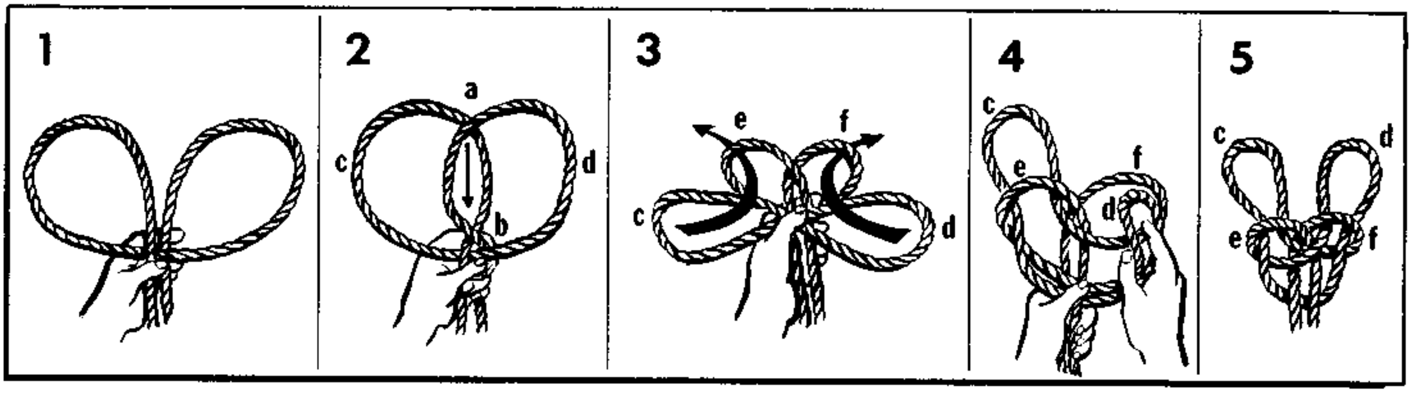

Figure 17-44. Spanish Bowline.

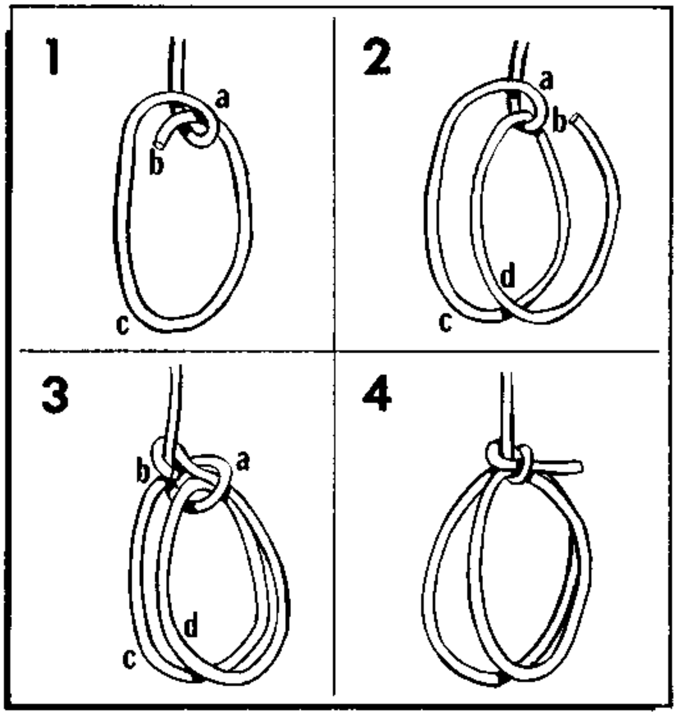

Figure 17-45. French Bowline.

people. When used in this manner, one loop is used as a seat and the other loop is used around the body under the arms. The weight of the injured person keeps the two loops tight so that the victim cannot fall out and for this reason, it is particularly useful as a sling for someone who is unconscious. The French bowline is started in the same way as the simple bowline. Make a loop (a) (figure 17-45-1) in the standing part of the rope. The running end (b) is passed through the loop from underneath and a separate loop (c) is made. The running end (b) is passed through the loop (a), again from underneath (figure 17-45-3), around the back of the standing part and back through the loop (a) so that it comes out parallel to the looped portion. The standing part of the rope is pulled to tighten the knot (figure 17-45-4), leaving two loops (c and d).

(8) Harness Hitch. The harness hitch (figure 17-46) is used to form a nonslipping loop in a rope. To make the harness hitch, form a bight (a) (figure 17-46-1) in the running end of the rope. Hold this bight in the left hand and form a second bight (b) in the standing part of the rope. The right hand is used to pass bight (b) over bight (a) (figure 17-46-2). Holding all loops in place with the left hand, the right hand is inserted through bight (a) behind the upper part of bight (b) (figure 17-46-3). The bottom (c) of the first loop is grasped and pulled up through the entire knot (figure 17-46-4), pulling it tight.

g. Hitches:

(1) Half Hitch. The half hitch (figure 17-47-1) is used to tie a rope to a timber or to another larger rope. It is not a very secure knot or hitch and is used for temporarily securing the free end of a rope. To tie a half hitch, the rope is passed around the timber, bringing the running end around the standing part, and back under itself.

(2) Timber Hitch. The timber hitch (figure 17-47-2) is used for moving heavy timbers or poles. To make the timber hitch, a half hitch is made and similarly the running end is turned about itself at least another time. These turns must be taken around the running end itself or the knot will not tighten against the pull.

(3) Timber Hitch and Half Hitch. To get a tighter hold on heavy poles for lifting or dragging a timber hitch and half hitch are combined (figure 17-47-3). The running end is passed around the timber and back under the standing part to form a half hitch. Further along the timber, a timber hitch is tied with the running end. The strain will come on the half hitch and the timber hitch will prevent the half hitch from slipping.

(4) Clove Hitch. A clove hitch (figure 17-47-4) is used to fasten a rope to a timber, pipe, or post. It can be tied at any point in a rope. To tie a clove hitch in the center of the rope, two turns are made in the rope close together. They are twisted so that the two loops lay back-to-back. These two loops are slipped over the timber or pipe to form the knot. To tie the clove hitch at the end of a rope, the rope is passed around the timber in two turns so that the first turn crosses the standing part and the running end comes up under itself on the second turn.

(5) Two Half Hitches. A quick method for tying a rope to a timber or pole is the use of two half hitches.

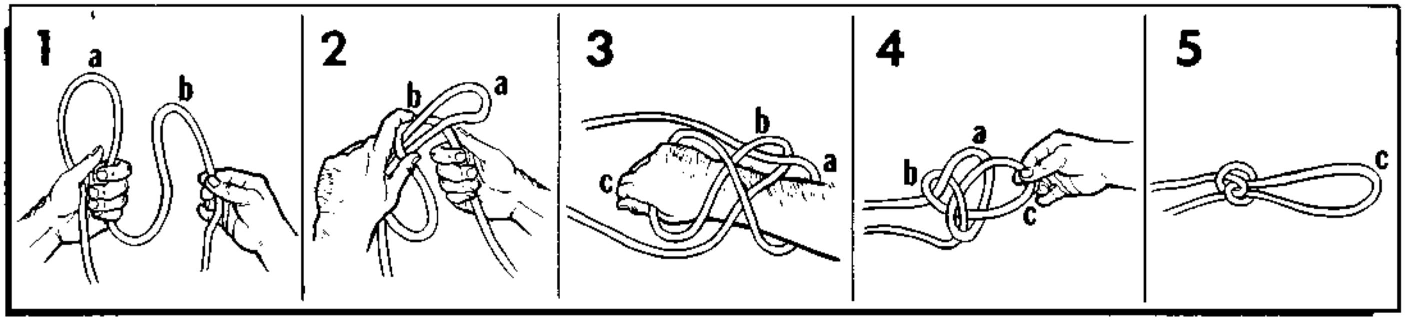

Figure 17-46. Harness Hitch.

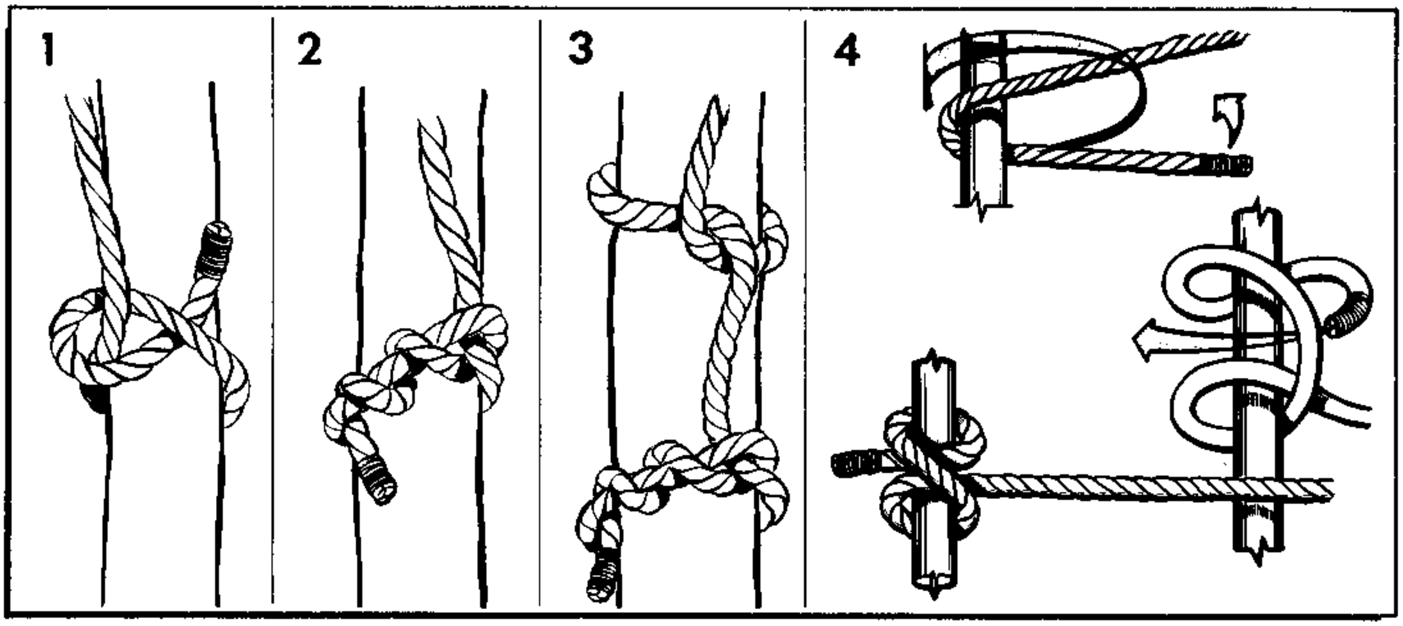

Figure 17-47. Half Hitch, Timber Hitch, and Clove Hitch.

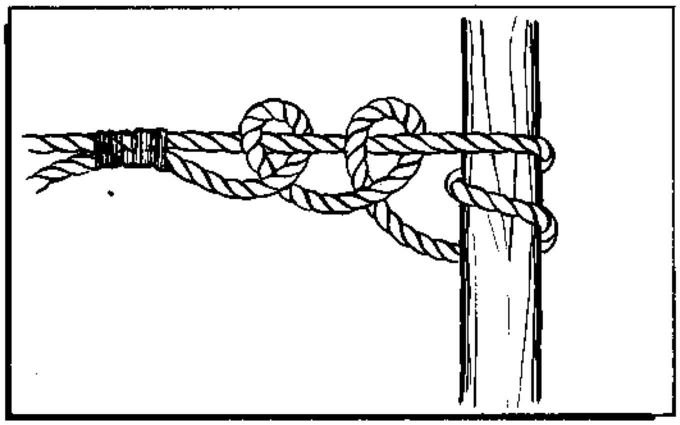

Figure 17-48. Round Turn and Two Half Hitches.

The running end of the rope is passed around the pole or timber, and a turn is taken around the standing part and under the running end. This is one half hitch. The running end is passed around the standing part of the rope and back under itself again.

(6) Round Turn and Two Half Hitches. Another hitch used for fastening a rope to a pole, timber, or spar is the round turn and two half hitches (figure 17-48). The running end of the rope is passed around the pole or spar in two complete turns, and the running end is brought around the standing part and back under itself to make a half hitch. A second half hitch is made. For greater security, the running end of the rope should be secured to the standing part.

(7) Fisherman's Bend. The fisherman's bend (figure 17-49) is used to fasten a cable or rope to an anchor, or for use where there will be a slackening and tightening motion in the rope. To make this bend, the running end of the rope is passed in two complete turns through the

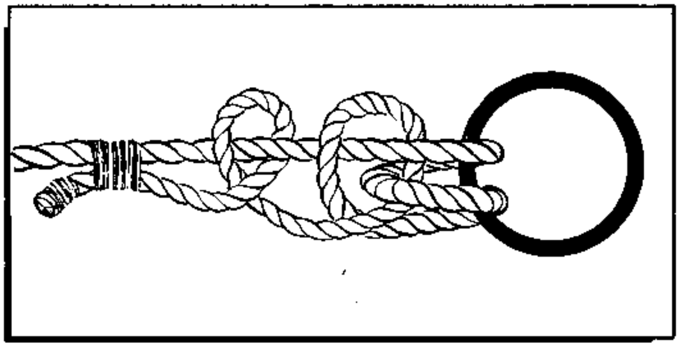

Figure 17-49. Fisherman's Bend.

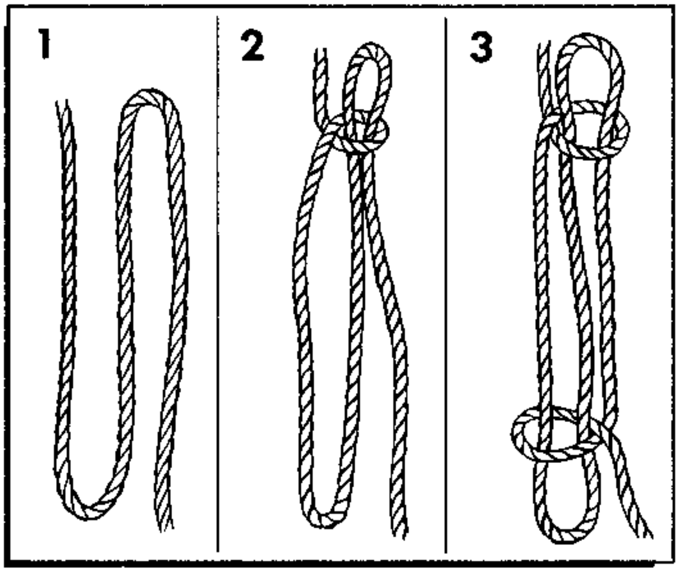

Figure 17-50. Sheep Shank.

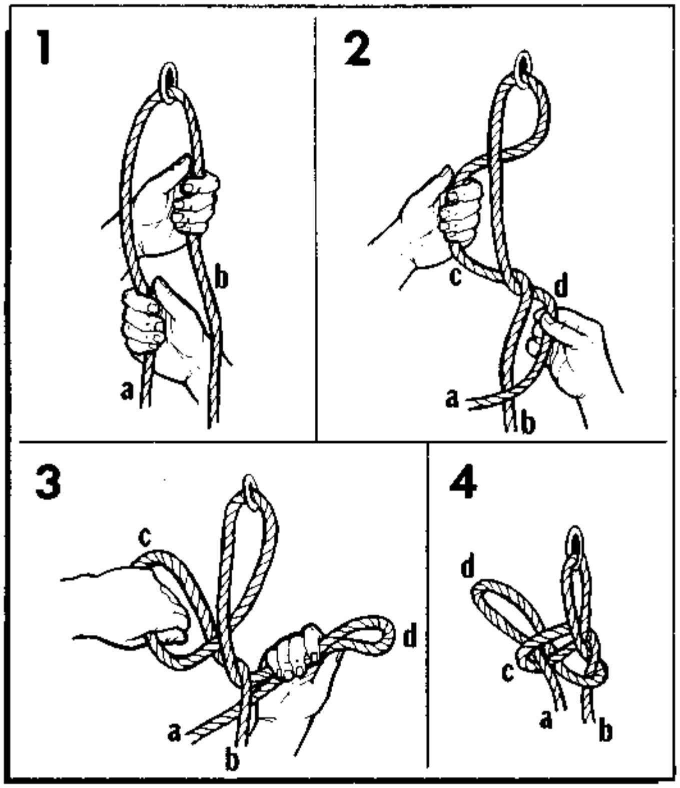

Figure 17-51. Speir Knot.

ring or object to which it is to be secured. The running end is passed around the standing part of the rope and through the loop which has just been formed around the ring. The running end is then passed around the standing part in a half hitch. The running end should be secured to the standing part.

(8) Sheepshank. A sheepshank (figure 17-50) is a method of shortening a rope, but it may also be used to take the load off a weak spot in the rope. To make the sheepshank (which is never made at the end of a rope), two bights are made in the rope so that three parts of the rope are parallel. A half hitch is made in the standing part over the end of the bight at each end.

(9) Speir Knot. A Speir knot (figure 17-51) is used when a fixed loop, a nonslip knot, and a quick release are required. It can be tied quickly and released by a pull on the running end. To tie the Speir knot, the running end (a) is passed through a ring (figure 17-51-1) or around a pipe or post and brought back on the left side of the standing part (b). Both hands are placed, palms up, under both parts of the rope with the left hand higher than the right hand; grasping the standing part (b) with the left hand and the running end (a) with the right hand. The left hand is moved to the left and the right hand to the right (figure 17-51-3) to form two bights (c and d). The left hand is twisted a half turn toward the body so that bight (c) is twisted into a loop (figure 17-51-3). Pass bight (d) over the rope and down through the loop (c). The Speir knot is tightened by pulling on the bight (d) and the standing part (b) (figure 17-51-4).

(10) Rolling Hitch (Pipe or Pole). The rolling hitch (pipe or pole) (figure 17-52) is used to secure a rope to a pipe or pole so that the rope will not slip. The standing part (a) of the rope is placed along the pipe or pole (figure 17-52-1) extending in the direction opposite to the direction the pipe or pole will be moved. Two turns (b) are taken with the running end around the standing part (a) and the pole (figure 17-52-3). The standing part (a) of the rope is reversed so that it is leading off in the direction in which the pole will be moved (figure 17-52-3) and two turns taken (c) (figure 17-52-4) with the running end (d). On the second turn around, the running end (d) is passed under the first turn (c) to secure it. To make this knot secure, a half hitch (e) (figure 17-52-6) is tied with the standing part of the rope 1 or 2 feet above the rolling hitch.

(11) Blackwall Hitch. The blackwall hitch (figure 17-53) is used for fastening a rope to a hook. To make the blackwall hitch, a bight of the rope is placed behind the hook. The running end (a) and standing part (b) are crossed through the hook so that the running end comes out at the opposite side of the hook and under the standing part.

(12) Catspaw. A catspaw can be made at the end of a rope (figure 17-54) for fastening the rope to a hook. Grasp the running end (a) of the rope in the left hand and make two bights (c and d) in the standing part (b). Hold these two bights in place with the left hand and take two turns about the junction of the two bights with the standing part of the rope. Slip the two loops (c and d) so formed over the hook.

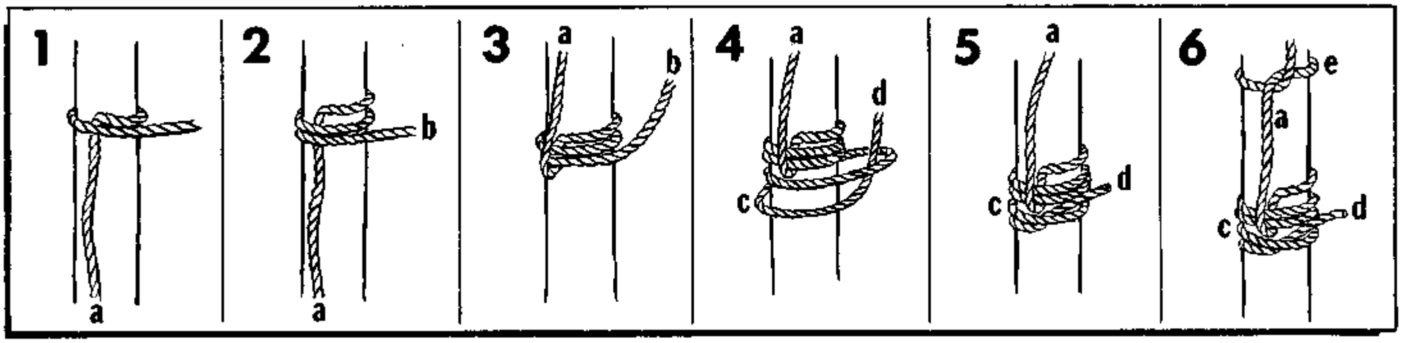

Figure 17-52. Rolling Hitch.

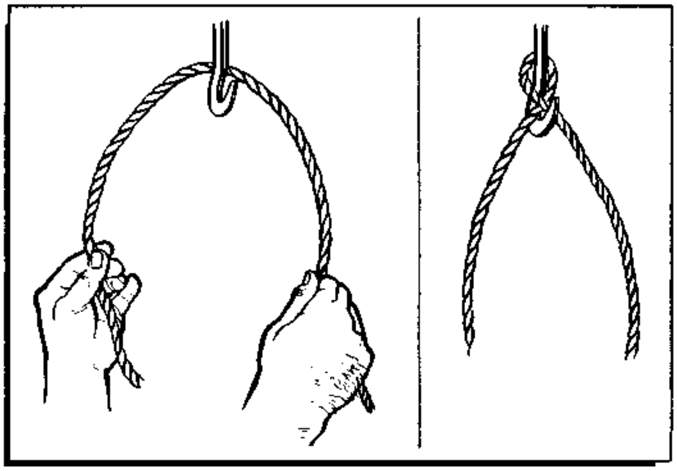

Figure 17-53. Blackwall Hitch.

(13) Scaffold Hitch. The scaffold hitch (figure 17-55) is used to support the end of a scaffold plank with a single rope. To make the scaffold hitch, the running end of the rope is layed across the top and around the plank, then up and over the standing part (figure 17-55-1). A doubled portion of the running end is brought back under the plank (figure 17-55-2) to form a bight (b) at the opposite side of the plank. The running end is taken back across the top of the plank (figure 17-55-3) until it can be passed through the bight (b). A loop is made (c) in the standing part (figure 17-55-4) above the plank. The running end is passed through the loop (c) around the standing part, and back through the loop (c).

(14) Barrel Slings. Barrel slings can be made to hold barrels horizontally or vertically. To sling a barrel horizontally (figure 17-56), a bowline is made with a long bight. The rope at the bottom of the bight is brought up over the sides of the bight. The two "ears" are thus moved foward over the end of the barrel. To sling a barrel vertically (figure 17-57) the rope is passed under the barrel and up to the top. An overhand knot is made (a) on top (figure 17-57-1). With a slight tension on the rope, the two parts (figure 17-57-2) of the overhand knot are grasped, separated and pulled down to the center of the barrel (b and c). The rope is pulled snug and a bowline tied (d) over the top of the barrel (figure 17-57-3).

h. Lashing. There are numerous items which require lashings for construction; for example, shelters, equipment racks, and smoke generators. Three types of lash-

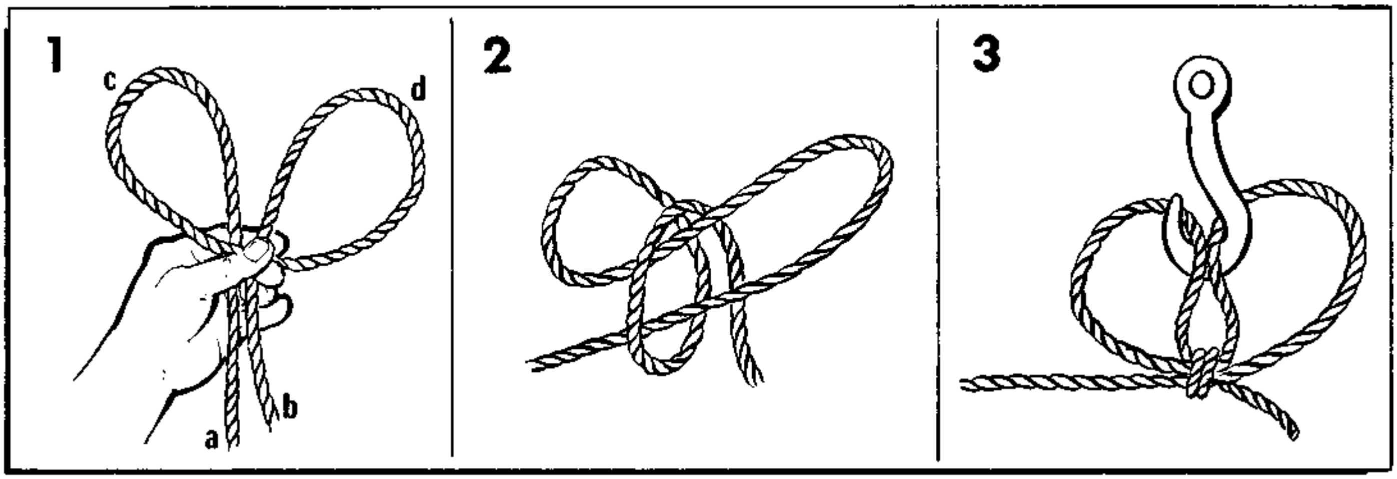

Figure 17-54. Catspaw.

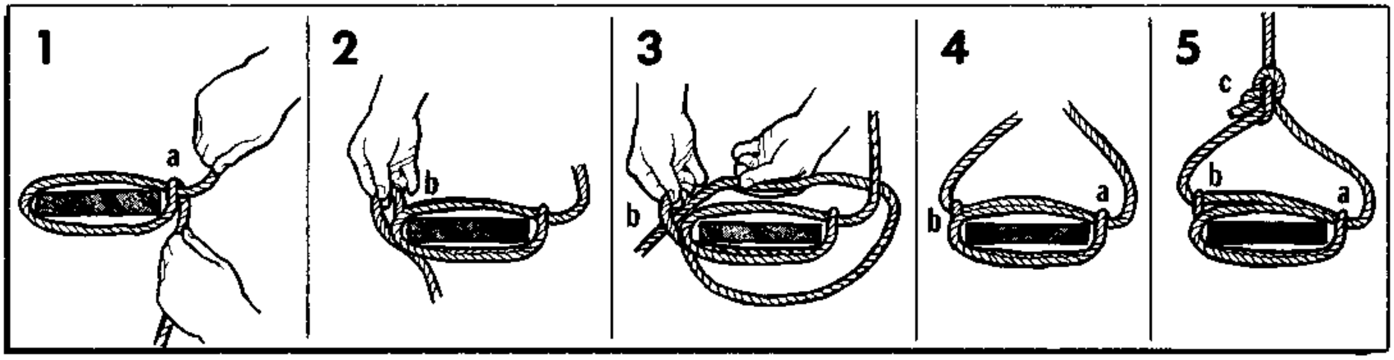

Figure 17-55. Scaffold Hitch.

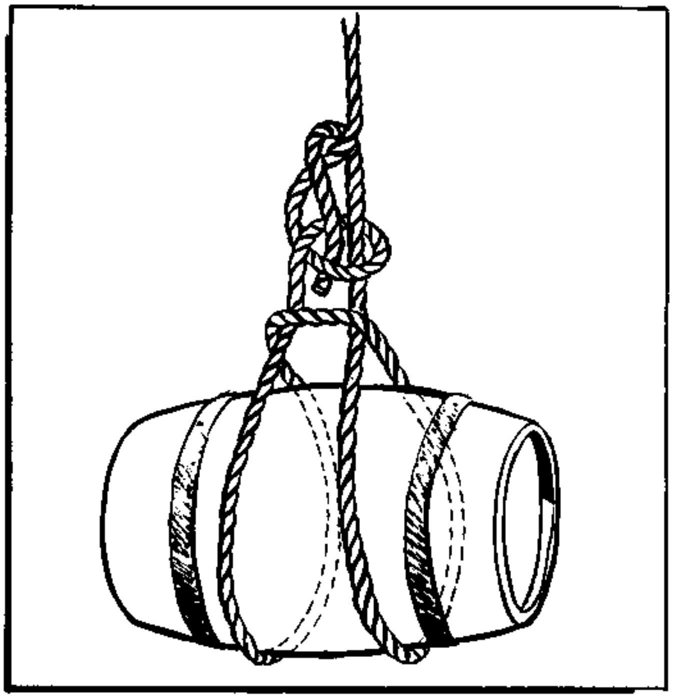

Figure 17-56. Barrel Slung Horizontally.

ings will be discussed here—the square lash, the diagonal lash, and the shear lash.

(1) Square Lash. Square lashing is started with a clove hitch around the log, immediately under the place where the crosspiece is to be located (figure 17-58-1). In laying the turns, the rope goes on the outside of the previous turn around the crosspiece, and on the inside of the previous turn around the log. The rope should be kept tight (figure 17-58-2). Three or four turns are necessary. Two or three "frapping" turns are made between the crosspieces (figure 17-58-3). The rope is pulled tight; this will bind the crosspiece tightly together. It is finished with a clove hitch around the same piece that the lashing was started on (figure 17-58-4). The square lash is used to secure one pole at right angles to another pole. Another lash that can be used for the same purpose is the diagonal lash.

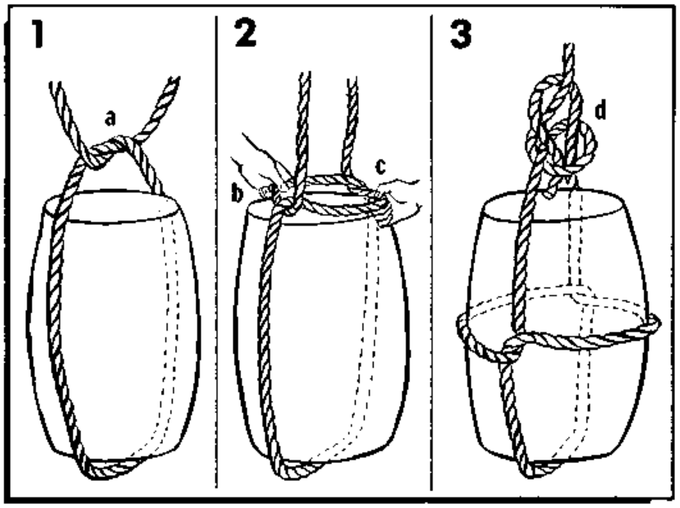

Figure 17-57. Barrel Slung Vertically.

(2) Diagonal Lash. The diagonal lash is started with a clove hitch around the two poles at the point of crossing. Three turns are taken around the two poles (figure 17-59-1). The turns lie beside each other, not on top of each other. Three more turns are made around the two poles, this time crosswise over the previous turns. The turns are pulled tight. A couple of frapping turns are made between the two poles, around the lashing turns, making sure they are tight (figure 17-59-2). The lashing is finished with a clove hitch around the same pole the lash was started on (figure 17-59-3).

(3) Shear Lash. The shear lash is used for lashing two or more poles in a series. The desired number of poles are placed parallel to each other and the lash is started with a clove hitch on an outer pole (figure 17-60-1). The poles are then lashed together, using seven or eight turns of the rope laid loosely beside each other (figure 17-60-2). Make frapping turns between each pole (figure 17-60-3). The lashing is finished with a clove hitch on the pole opposite that on which the lash was started (figure 17-60-4).

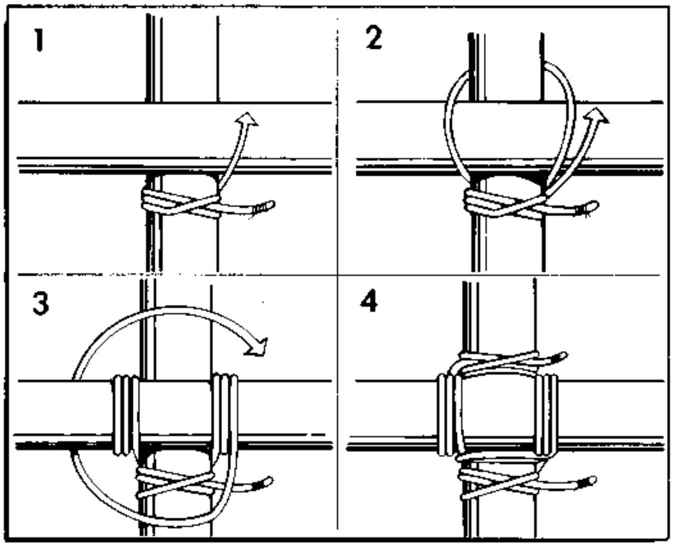

Figure 17-58. Square Lash.

i. Making Ropes and Cords. Almost any natural fibrous material can be spun into good serviceable rope or cord, and many materials which have a length of 12 to 24 inches or more can be braided. Ropes up to 3 and 4 inches in diameter can be "laid" by four people, and tensile strength for bush-made rope of 1-inch diameter range from 100 pounds to as high as 3,000 pounds.

(1) Tensile Strength. Using a three-lay rope of 1-inch diameter as standard, the following table of tensile strengths may serve to illustrate general strengths of various materials. For safety's sake, the lowest figure should always be regarded as the tensile strength.

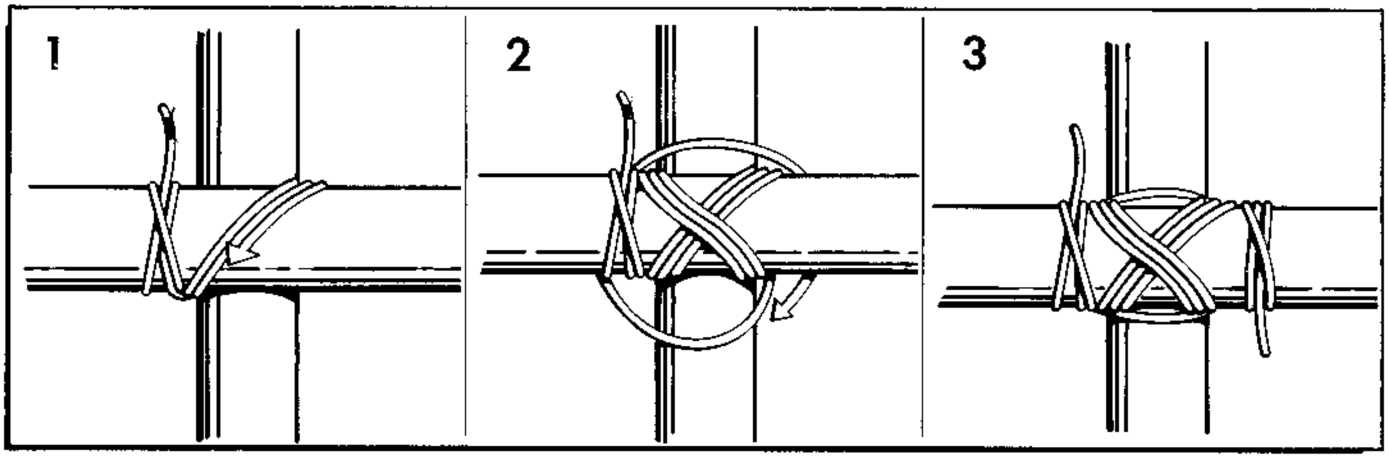

Figure 17-59. Diagonal Lash.

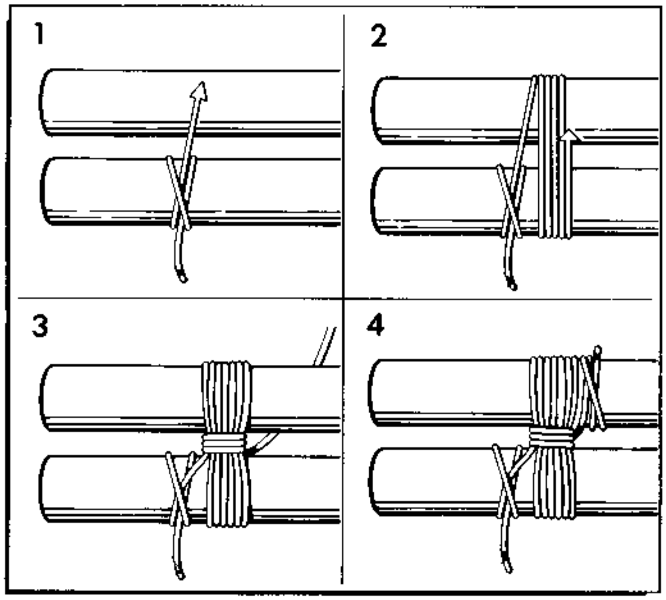

Figure 17-60. Shear Lash.

Green Grass	100 lbs to 250 lbs
Bark Fiber	500 lbs to 1,500 lbs
Palm Fiber	650 lbs to 2,000 lbs
Sedges	2,000 lbs to 2,500 lbs
Monkey Rope (Lianas)	560 lbs to 700 lbs
Lawyer Vine (Calamus)	½-inch diam, 1,200 lbs

NOTE: Doubling the diameter quadruples the tensile strength half the diameter reduces the tensile strength to one-fourth.

(2) Principles of Ropemaking Materials. To discover whether a material is suitable for rope making, it must have four qualities:

(a) It must be reasonably long in the fiber.

(b) It must have "strength."

(c) It must be pliable.

(d) It must have "grip" so the fibers will "bite" onto one another.

(3) Determining Suitability of Material. There are simple tests to determine if a material is suitable:

(a) First, pull on a length of the material to test for strength.

(b) Second, twist it between the fingers and "roll" the fibers together; if it will withstand this and not "snap" apart, an overhand knot is tied and gently tightened. If the material does not cut upon itself, but allows the knot to be pulled taut, it is suitable for ropemaking if the material will "bite" together and is not smooth or slippery.

(4) Where to Find Suitable Material. These qualities can be found in various types of plants, in ground vines, in most of the longer grasses, in some of the water reeds and rushes, in the inner barks of many trees and shrubs, and in the long hair or wool of many animals.

(5) Obtaining Fibers for Making Ropes. Some green freshly gathered materials may be "stiff" or unyielding. When this is the case, it should be passed through hot flames for a few moments. The heat treatment should cause the sap to burst through some of the cell structure, and the material thus becomes pliable. Fibers for rope making may be obtained from many sources such as:

(a) Surface roots of many shrubs and trees have strong fibrous bark.

(b) Dead inner bark of fallen branches of some species of trees and in the new growth of many trees such as willows.

(c) The fibrous material of many water and swamp growing plants and rushes.

(d) Many species of grass and weeds.

(e) Some seaweeds.

(f) Fibrous material from leaves, stalks, and trunks of many palms.

(g) Many fibrous-leaved plants such as the aloes.

(6) Gathering and Preparing Materials. There may be a high content of vegetable gum in some plants. This can often be removed by soaking the plants in water, by

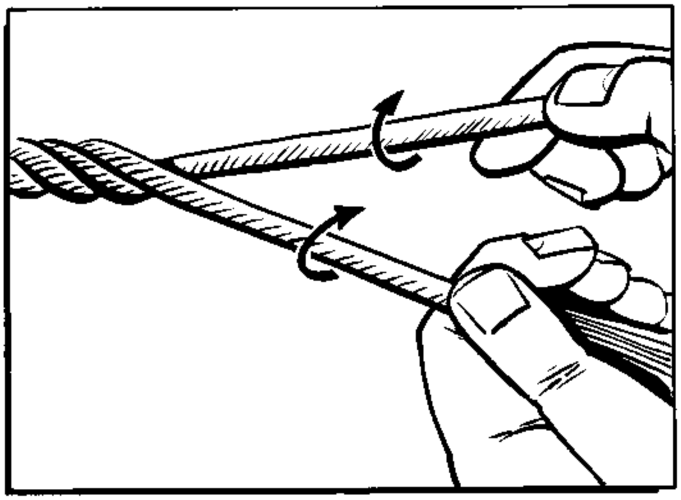

Figure 17-61. Twisting Fibers.

boiling, or by drying the material and "teasing" it into thin strips.

(a) Some of the materials have to be used green if any strength is required. The materials that should be green include the sedges, water rushes, grasses, and lianas.

(b) Palm fiber is harvested in tropical or subtropical regions. It is found at the junction of the leaf and the palm trunk, or it will be found lying on the ground beneath many palms. Palm fiber is a "natural" for making ropes and cords.

(c) Fibrous matter from the inner bark of trees and shrubs is generally more easily used if the plant is dead or half dead. Much of the natural gum will have dried out and when the material is being teased, prior to spinning, the gum or resin will fall out in fine powder.

(7) Making a Cord by Spinning with the Fingers:

(a) Use any material with long strong threads or fibers which have been previously tested for strength and pliability. The fibers are gathered into loosely held strands of even thickness. Each of these strands is twisted clockwise. The twist will hold the fibers together. The strands should be formed one-eighth inch diameter. As a general rule, there should be about 15 to 20 fibers to a strand. Two, three, or four of these strands are later twisted together, and this twisting together or "laying" is done with a counterclockwise twist, while at the same time, the separate strands which have not yet been laid up are twisted clockwise. Each strand must be of equal twist and thickness.

(b) Figure 17-61 shows the general direction of twist and the method whereby the fibers are bonded into strands. In a similar manner, the twisted strands are put together into lays, and the lays into ropes.

(c) The person who twists the strands together is called the "layer" and must see that the twisting is even, the strands are uniform, and the tension on each strand is equal. In "laying," care must be taken to ensure each of the strands is evenly "laid up;" that is, one strand does not twist around the other one.

(d) When spinning fine cords for fishing lines, snares, etc., considerable care must be taken to keep the strands uniform and the lay even. Fine thin cords of no more than 1/32-inch thickness can be spun with the fingers and are capable of taking a breaking strain of 20 to 30 pounds or more.

(e) Normally two or more people are required to spin and lay up the strands for cord. However, many native people spin cord unaided. They twist the material by running the flat of the hand along the thigh, with the fibrous material between hand and thigh; and with the free hand, they feed in fiber for the next "spin." Using this technique, one person can make long lengths of single strands. This method of making cord or rope with the fingers is slow if any considerable length of cord is required.

(f) An easier and simpler way to rapidly make lengths of rope from 50 to 100 yards or more in length is to make a rope machine and set up multiple spinners in the form of cranks. Figure 17-62 shows the details of rope spinning.

(g) To use a rope machine, each feeder holds the material under one arm and with one free hand feeds it into the strand which is being spun by the crank. The other hand lightly holds the fibers together till they are spun. As the lightly spun strands are increased in length, they must be supported on crossbars. They should not be allowed to lie on the ground. Spin strands from 20 to 100 yards before laying up. The material should not be spun in too thickly. Thick strands do not help strength in any way, rather, they tend to make a weaker rope.

(8) Setting Up a Rope Machine:

(a) When spinning ropes of 10 yards or longer, it is necessary to set crossbars every 2 or 3 yards to carry the strands as they are spun. If crossbars are not set up, the strands or rope will sag to the ground, and some of the fibers will tangle up with grass, twigs, or dirt on the ground. Also, the twisting of the free end may either be stopped or interrupted and the strand will be unevenly twisted.

(b) The easiest way to set up crossbars for the rope machine is to drive pairs of stakes into the ground about 6 feet apart and at intervals of about 6 to 10 feet. The crossbars must be smooth and free from twigs and loose portions of bark that might twist in with the spinning strands.

(c) The crossbar (a) is supported by two uprights and pierced to take the cranks (b). These cranks can be made out of natural sticks, morticed slab, and pegs, or if available, bent wire. The connecting rod (c) enables one person to turn all cranks clockwise simultaneously. Crossbars supporting the strands as they are spun are shown (d). A similar crank handle to the previous ones (b) is supported on a forked stick at the end of the rope

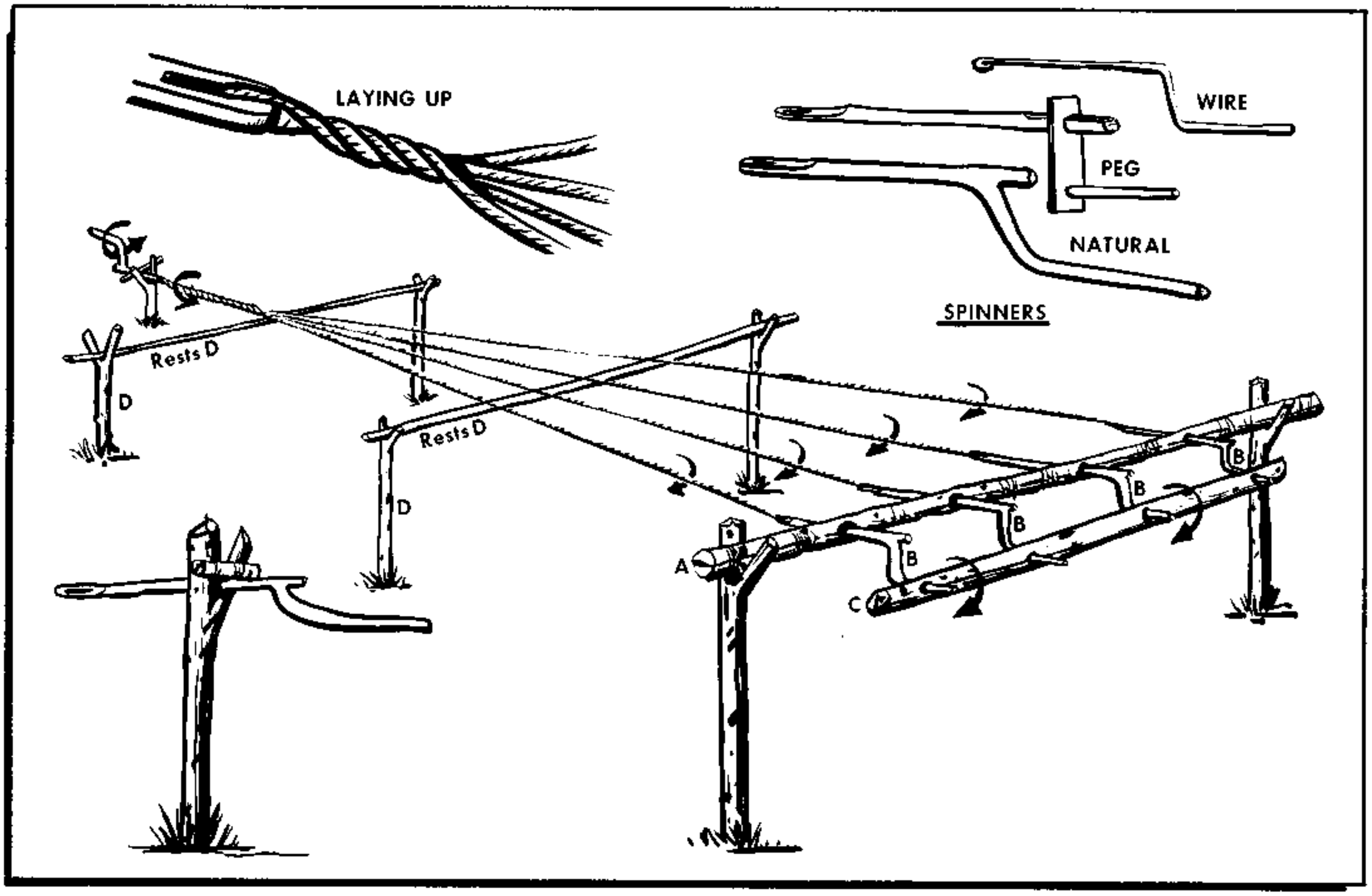

Figure 17-62. Rope Machine.

machine. This handle is turned in reverse (counterclockwise) to the cranks (c) to twist the connected strands together. These are "laid up" by one or more of the feeders.

(d) The first strand should be turned clockwise, then the laying up of the strands will be done counterclockwise and the next laying will again be clockwise. Proof that the rope is well made is that the individual fibers lay lengthways along the rope.

(e) In the process of laying up the strands, the actual twisting together or laying will take some of the original "twist" out of the strand which has not yet been laid. Therefore, it is necessary to keep twisting the strands while laying together.

(f) When making a rope too long to be spun and laid in one piece, a section is laid up and coiled on the ground at the end of the rope walk farthest from the cranks. Strands for a second length are spun, and these strands are married or spliced into the strands of the first section and then the laying up of the second section continues the rope.

(g) The actual "marrying" of the strands is done only in the last lay, which makes the rope when completed. The ends where the strands are married should be staggered in different places. By this means, rope can be made and extended in sections to a great length.

(h) After a complete length of rope is laid up, it should be passed through the fire to burn off the loose ends and fibers. This will make the rope smooth and more professional looking.

(9) Laying the Strands:

(a) The strands lie on the crossbar as they are spun. When the strands have been spun to the required length, which should not be more than about a hundred feet, they are joined together by being held at the far end. They are then ready for laying together. The turner, who is facing the cranks, twists the ends together counterclockwise, at the same time keeping full weight on the rope which is being layed up. The layer advances placing the strands side by side as they turn.

(b) It is important to learn to feed the material evenly, and lay up slowly, thereby getting a smooth even rope (figure 17-63). Do not try to rush the ropemaking. Speed in ropemaking only comes with practice. At first it will take a team of three or four up to 2 hours or more to make a 50-yard length of rope of three lays, each of three strands; that is, nine strands for a rope with a finished diameter of about 1 inch. With practice, the same three or four people will make the same rope in 15 to 20 minutes. These times do not include time for gathering material.

(c) In feeding the free ends of the strands, twist in the loose material fed in by the feeder. As the feeders move backward, they must keep a slight tension on the strands.

(10) Making Rope with a Single Spinner:

(a) Using a Single Crank. Two people can make a rope, using a single crank. A portion of the material is fastened to the eye of the crank, as with the multiple crank. Supporting crossbars, as used in a ropewalk, are required when a length of more than 20 or 30 feet is being spun.

(b) Feeding:

-1. If the feeder is holding material under the left arm, the right hand is engaged in continuously pulling material forward to the left hand which feeds it into the turning strand. These actions, done together as the feeder walks backward, govern the thickness of the strands. The left hand, lightly closed over the loose turning material, must "feel" the fibers "biting" or twisting together.

-2. When the free end of the turning strand, which is against the loose material under the arm, takes in too thick a tuft of material, the left hand is closed, and so arrests the twist of the material between the left hand and the bundle. This allows teasing out the overall "bite," with the right hand, thus maintaining a uniform thickness of the spinning strand.

(c) Thickness of Strands. Equal thickness and twist for each of the strands throughout their length are important. The thickness should not be greater than is necessary with the material being used. For a grass rope, the strand should not be more than one-fourth inch diameter; for coarse bark or palm, not more than one-eighth or three-sixteenth inch; and for fine bark, hair, or sisal fiber, not more than one-eighth inch.

(d) Common Errors in Ropemaking:

-1. There is a tendency with beginners to feed unevenly. Thin wispy sections of strand are followed by thick portions. Such feeding degrades the quality of rope. Rope made from such strands will break with less than one-fourth of the tensile strain on the material.

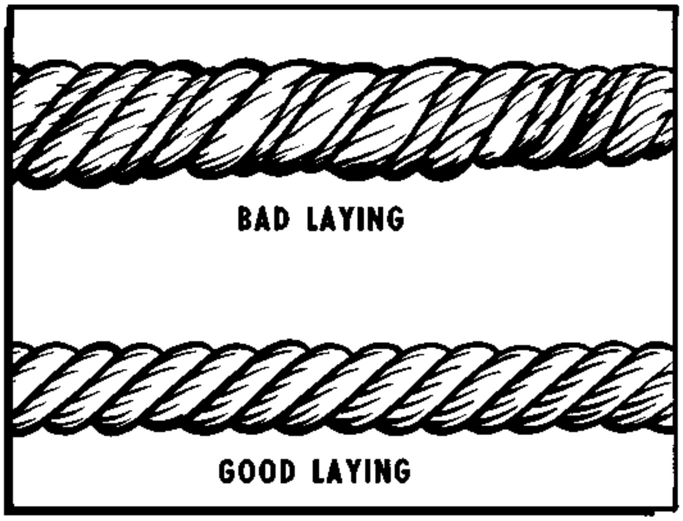

Figure 17-63. Rope Laying.

-2. Beginners are wise to twist and feed slowly. Speed, with uniformity of twist and thickness, comes with practice.

-3. Thick strands do not help. It is useless to try and spin a rope from strands an inch or more in thickness. Such a rope will break with less than half the tensile strain on the material. Spinning "thick" strands does not save time in ropemaking.

(e) Lianas, Vines, and Canes. Lianas and ground vines are natural ropes, and grow in subtropical and tropical scrub and jungle. Many are of great strength and useful for braiding, tree climbing, and other purposes. The smaller ground vines, when "braided", give great strength and flexibility. Canes and stalks of palms provide excellent material if used properly. Only the outer skin is tough and strong, and this skin will split off easily if the main stalk is bent away from the skin. This principle also applies to the splitting of lawyer cane (calamus), palm leaf stalks, and all green material. If the split starts to run off, bend the material away from the thin side, and it will gradually gain in size and come back to an even thickness with the other split side.

(f) Bark Fibers:

-1. The fibers in many barks which are suitable for "ropemaking" are located near the innermost layers. This is the bark next to the sap wood. When seeking suitable barks of green timber, cut a small section about 3 inches long and 1 inch wide. Cut this portion from the wood to the outer skin of the bark.

-2. The specimen should be peeled and the different layers tested. Green bark fibers are generally difficult to spin because of "gum" and it is better to search around for windfall dead branches and try the inner bark of these. The gum probably has leached out, and the fibers should separate easily.

-3. Many shrubs have excellent bark fiber, and here it is advisable to cut the end of a branch and peel off a strip of bark for testing. Thin bark from green shrubs is sometimes difficult to spin into fine cord and is easier to use as braid for small cords.

-4. Where it is necessary to use green bark fiber for rope spinning the gum will generally wash out when the bark is teased and soaked in water for a day or so. After removing from the water, the bark strips should be allowed to dry before shredding and teasing into fiber.

(11) Braiding. One person may require a length of rope. If there is no help available to spin materials, it is necessary to find reasonably long material. With this material, one person can braid and make suitable rope. The usual three-strand braid makes a flat rope, and while quite good, it does not have finish or shape, nor is it as "tight" as the four-strand braid. On other occa-

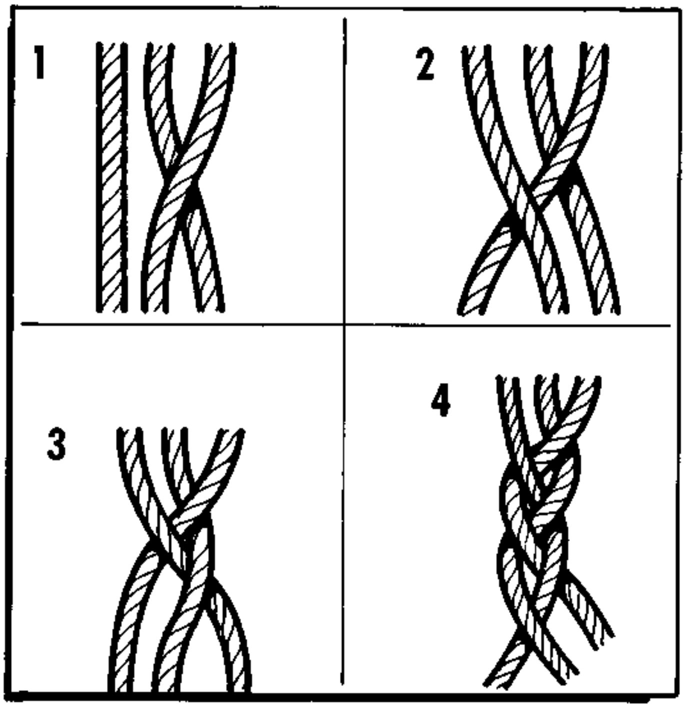

Figure 17-64. Three-Strand Braid.

sions, it may be necessary to braid broad bands for belts or for shoulder straps. There are many fancy braids which can be developed from these, but these three are basic, and essential for practical woodcraft work. A general rule for all braids is to work from the outside into the center.

(a) Three Plait:

-1. The right-hand strand is passed over the strand to the left.

-2. The left-hand strand is passed over the strand to the right.

-3. This is repeated alternately from left to right (figure 17-64).

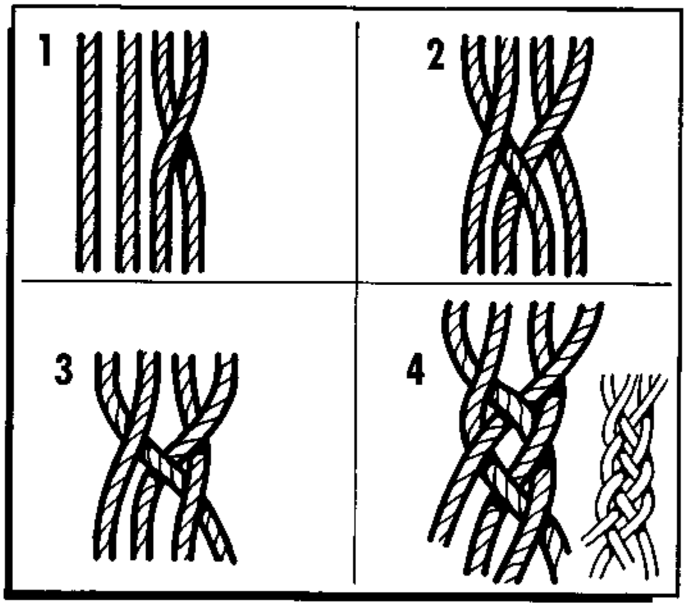

Figure 17-65. Four-Strand Braid.

(b) Flat Four-Strand Braid:

-1. The four strands are placed side by side. The right-hand strand is taken (figure 17-65-1) and placed over the strand to the left.

-2. The outside left-hand strand (figure 17-65-2) is laid under the next strand to itself and over what was the first strand.

-3. The outside right-hand strand is laid over the first strand to its left (figure 17-65-3).

-4. The outside left strand is placed under and over the next two strands, respectively, moving toward the right.

-5. Thereafter, the right-hand strand goes over one strand to the left, and the left-hand strand under and over to the right (figure 17-65-4).

(c) Broad Braid. Six or more strands are held flat and together.

-1. A strand in the center is passed over the next strand to the left, as in figure 17-66-1.

-2. The second strand to the left of center is passed toward the right and over the first strand so that it points toward the right (figure 17-66-2).

-3. The strand next to the first one is taken and woven under and over (figure 17-66-3).

-4. The next strands are woven from left and right alternately towards the center (figure 17-66-4 through 6). The finished braid should be tight and close (figure 17-66-7).

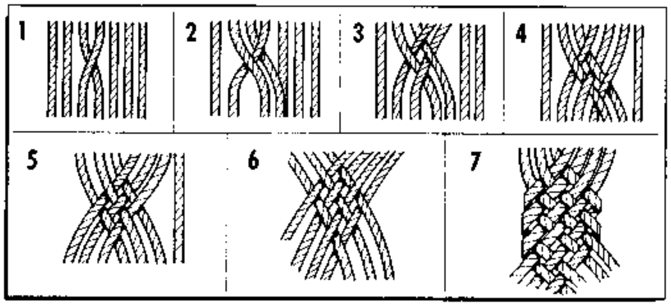

Figure 17-66. Broad Braid.

-5. To finish the broad braid:

-a. One of the center strands is laid back upon itself (figure 17-67-1).

-b. Now take the first strand which it enclosed in being folded back, and weave this back upon itself (figure 17-67-2).

-c. Strand from the opposite side is laid back and woven between the strands already braided (figure 17-67-3).

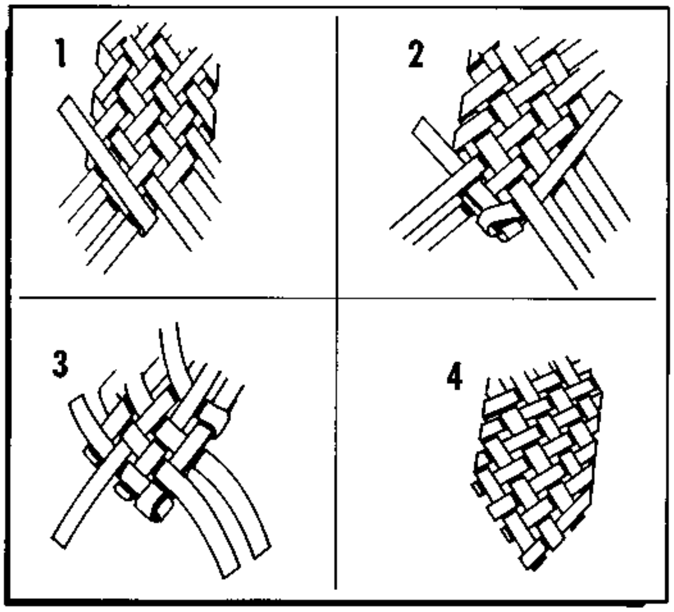

Figure 17-67. Finishing the Broad Braid.

-d. All the strands should be so woven back that no strands show an uneven pattern, and there should be a regular under-over-under of the alternating weaves (figure 17-67-4).

-e. If the braid is tight, there may be a difficulty in working the loose ends between the plaited strands.

-f. This can be done easily by sharpening a thin piece of wood to a chisel edge to open the strands sufficiently to allow the ends being finished to pass between the woven strands.

-g. It should be rolled under a bottle or other round object and made smooth for final finishing.

17-7. Personal Survival Kit:

a. Even though a survival kit may be available, aircrew members should consider assembling and carrying personal survival kits. Survival experiences have occurred where survivors hit the ground running, and because of shock and fear left their survival kits behind. If survivors have a personal survival kit in a pocket, it may improve their survival chances considerably.

b. A great deal of thought should go into preparing personal survival kits. The potential needs of the survivors must be a consideration, such as the impact of the environmental elements, type of mission to be flown (tactical or nontactical), availability of rescue, and how far to friendly forces (figure 17-68).

c. There are two basic ways to carry a personal survival kit. One way is to pack all items into one or two waterproof containers. The other way is to scatter the items throughout personal clothing. Any type of small container can be used to encase the contents of the personal survival kit. Plastic cigarette cases, soap dishes, and Band-aid boxes are excellent containers.

d. Examples of items which can be packed into a small container are:

(1) Matches.
(2) Safety pins (varied sizes).
(3) Fishhooks.
(4) Knife (small, multibladed).
(5) Button compass.
(6) Prophylactic (for water container).
(7) Bouillon cubes.
(8) Salt.
(9) Snare wire.
(10) Water purification tablets.
(11) Signal mirror.
(12) Needles.
(13) Band-aids.
(14) Aluminum Foil.
(15) Insect repellent stick.
(16) Chapstick.
(17) Soap (Antiseptic).

NOTE: All kits carried aboard the aircraft should be approved by the unit life support officer.

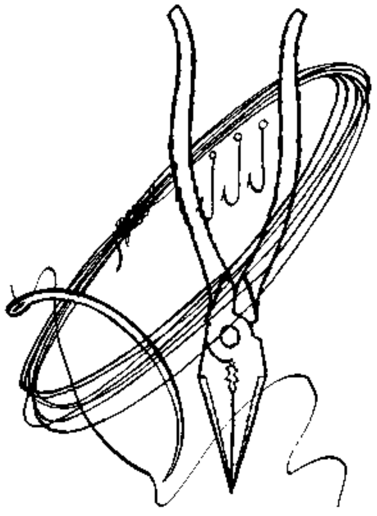

MINIMUM ESSENTIAL ITEMS

High quality pocket knife with at least two cutting blades.

Pocket compass.

Match safe with matches.

- Plastic or metallic container.
- Waterproof kitchen-type matches (cushion heads against friction), or
- Waterproof matches rolled in paraffin-soaked muslin in an easily opened container such as small soap box, toothbrush case, etc.

Needles — sailmakers, surgeons, and darning — at least one of each.

Assorted fishhooks in heavy foil, tin, or plastic holders.

Snare wire — small hank.

Needle-nosed pliers with side cutters; high quality.

Bar surgical soap or hand soap containing physohex.

Small fire starter of pyrophoric metal (some plastic match cases have a strip of the metal anchored on the bottom outside of the case).

Personal medicines.

Water purification tablets.

"Bandaids."

Insect repellent stick.

Chapstick.

GOOD TO HAVE ITEMS

*Pen-gun and flares.

*Colored cloth or scarf for signaling.

Stick-type skin dye (for camouflage).

Plastic water bottle.

*Flexible saw (wire saw).

*Sharpening stone.

Safety pins (several sizes).

Travel razor.

Small steel mirror.

6" flat bastard file.

Aluminum foil.

ADDITIONAL SUGGESTIONS

Toothbrush — small type.

Surgical tape.

Prophylactics (make good waterproof containers or canteens).

*Penlight with batteries.

Fishline.

*Fishline monofilament.

*Clear plastic bags.

Emergency ration can opener (can be taped shut and strung on dog tag chain).

Split shot — for fishing sinkers.

Gill net.

Small, high quality candles.

INDIVIDUAL MEDICAL KIT

Sterile gauze compress bandage.

Antibiotic ointment (Neomycin polymycin bacitracin opthalmic ointment is good).

Tincture of zephrine — skin antiseptic.

Aspirin tablets.

Salt tablets.

Additional medications may be desirable, depending upon nature of the mission and an individual's particular personal needs.

This should be discussed with and procured from your local flight surgeon.

*Especially valuable.

Figure 17-68. Personal Survival Kit Items.

Part Six

SUSTENANCE

Chapter 18

FOOD

18-1. Introduction. Except for the water they drink and the oxygen they breathe, survivors must meet their body needs through the intake of food. This chapter will explore the relationship of proper nutrition to physical and mental efficiency. It is extremely important that survivors maintain a proper diet at all times. A nutritionally sound body stands a much better chance of surviving. Improper diet over a long period of time may lead to a lack of stamina, slower reactions, less resistance to illness, and reduced mental alertness, all of which can cost survivors their lives in a survival situation. A knowledge of the body's nutritional requirements will help survivors select foods to supplement their rations.

18-2. Nutrition. Survivors and evaders expend much more energy in survival situations than they would in the course of their normal everyday jobs and life. Basal metabolism is the amount of energy expended by the body when it is in a resting state. The rate of basal metabolism will vary slightly with regard to the sex, age, weight, height, and race of a person. The basic energy expended, or number of calories consumed by the hour will change as a person's activity level changes. A person who is simply sitting in a warm shelter, for example, may consume anywhere from 20 to 100 calories an hour, while that same person evading through thick undergrowth with a heavy pack, would expand a greater amount of energy. In a survival situation, proper food can make the difference between success and failure.

a. The three major constituents of food are carbohydrates, fats, and proteins. Vitamins and minerals are also important as they keep certain essential body processes in good working order. It is also necessary for survivors to maintain proper water and salt levels in their bodies, as they aid in preventing certain heat disorders.

(1) Carbohydrates. Carbohydrates are composed of very simple molecules which are easily digested. Carbohydrates lose little of their energy to the process of digestion and are therefore efficient energy suppliers. Because carbohydrates supply easily used energy, many nutritionists recommend that, if possible, survivors should try to use them for up to half of their calorie intake. Examples of carbohydrates are: starches, sugars, and cellulose. These can be found in fruits, vegetables, candy, milk, cereals, legumes, and baked goods. Cellulose cannot be digested by humans, but it does provide needed roughage for the diet.

(2) Fats. Fats are more complex than carbohydrates. The energy contained in fats is more slowly released than the energy in carbohydrates. Because of this, it is a longer lasting form of energy. Fats supply certain fat-soluble vitamins. Sources of these fats and vitamins are butter, cheese, oils, nuts, egg yolks, margarine, and animal fats. If survivors eat fats before sleeping, they will sleep warmer. If fats aren't included in the diet of survivors, they can become run down and irritable. This can lead to both physical and psychological breakdown.

(3) Protein. The digestive process breaks protein down into various amino acids. These amino acids are formed into new body tissue protein, such as muscles. Some protein gives the body the exact amino acids required to rebuild itself. These proteins are referred to as "complete." Protein that lacks one or more of these essential amino acids is referred to as "incomplete." Incomplete protein examples are cheese, milk, cereal grains, and legumes. Incomplete protein, when eaten in combination with milk and beans for example, can supply an assortment of amino acids needed by the body. Some complete protein is found in fish, meat, poultry, and blood. No matter which type of protein is consumed, it will contain the most complex molecules of any food type listed.

(a) If possible, the recommended daily allowance of 2½ to 3 ounces complete protein should be consumed by each survivor each day. If only the incomplete protein is available, two, three, or even four types of foods may need to be eaten in combination so that enough amino acids are combined to form complete protein.

(b) If amino acids are introduced into the body in great numbers and some of them are not used for the rebuilding of muscle, they are changed into fuel or stored in the body as fat. Because protein contains the more complex molecules, over fats or carbohydrates, they supply energy after those forms of energy have been used up. A lack of protein causes malnutrition, skin and hair disorders, and muscle atrophy.

b. Vitamins occur in small quantities in many foods, and are essential for normal growth and health. Their chief function is to regulate the body processes. Vitamins can generally be placed into two groups: fat-soluble and water-soluble. The body only stores slight amounts of the water-soluble type. In a long survival episode where a routinely balanced diet is not available, survivors must overcome food aversions and eat as much of a variety of vitamin-rich foods as possible. Often one or more of the four basic food groups (meat, fish, poultry; vegetables and fruits; grain and cereal; milk and milk products) are not available in the form of familiar foods, and vitamin deficiencies such as beriberi or scurvy result. If the survivor can overcome aversions to local foods high in vitamins, these diseases as well as signs and symptoms such as depression and irritability can be warded off.

c. Adequate minerals can also be provided by a balanced diet. Minerals build and (or) repair the skeletal system and regulate normal body functions. Minerals needed by the body include iodine, calcium, iron, and salt, to name but a few. A lack of minerals can cause problems with muscle coordination, nerves, water retention, and the ability to form or maintain healthy red blood cells.

d. For survivors to maintain their efficiency, the following number of calories per day is recommended. These figures will change because of individual differences in basal metabolism, weight, etc. During warm weather survivors should consume anywhere from 3,000 to 5,000 calories per day. In cold weather the calorie intake should rise from 4,000 to 6,000 calories per day. A familiarity with the calorie and fat amounts in foods is important for survivors to meet their nutritional needs. For example, it would take quite a few mussels and dandelion greens to meet those requirements. Survivors should attempt to be familiar enough with foods that they can select or find foods that provide a high calorie intake (figure 18-1).

(1) Survivors should also be familiar with the number of calories supplied by the food in issued rations. In most situations, rations will have to be supplemented with other foods procured by survivors. If possible, survivors should limit their activities to save energy. Rationing food is a good idea since survivors never know when their ordeal will end. They should eat when they can, keeping in mind that they should maintain at least a minimum calorie intake to satisfy their basic activity needs.

(2) Caloric and fat values of selected foods are shown in the chart, and unless otherwise specified, the foods listed are raw. Depending on how survivors cook the food, the usable food value can be increased or decreased.

FOOD	CALORIES	FAT
WHOLE LARGE DUCK EGG	177	12.0
SMALL OR LARGE MOUTH BASS – 3 TO 4 OZ.	109	3.6
CLAMS – 4 TO 5 LARGE	88	.2
FRESHWATER CRAYFISH – 3 TO 4 OZ.	75	.6
EEL – 3 TO 5 OZ.	240	20.0
OCTOPUS – 3 TO 4 OZ.	76	.9
ATLANTIC SALMON – 4 OZ.	220	14.0
RAINBOW TROUT – 4 OZ.	200	11.8
BANANA – ONE SMALL	87	.3
BREADFRUIT – 3 TO 4 OZ.	105	.5
GUAVA – ONE MEDIUM	64	.7
MANGO – ONE SMALL	68	.5
WILD DUCK – 4 OZ.	230	16.0
BAKED OPOSSUM – 4 OZ.	235	10.6
WILD RABBIT – 4 OZ.	124	4.0
VENISON – 4 OZ.	128	3.1
DANDELION GREENS – ONE CUP COOKED	70	1.4
POTATO – MEDIUM	78	.2
PRICKLY PEAR – 4 OZ.	43	.2

Figure 18-1. Food and Calorie Diagram.

18-3. Food. Survivors should be able to find something to eat wherever they are. One of the best places to find food is along the seacoast, between the high and low watermark. Other likely spots are the areas between the beach and a coral reef; the marshes, mud flats, or mangrove swamps where a river flows into the ocean or into a larger river; riverbanks, inland waterholes, shores of ponds and lakes, margins of forests, natural meadows, protected mountain slopes, and abandoned cultivated fields.

a. Rations placed in survival kits have been developed especially to provide some of the proper sustenance needed during survival emergencies. When eaten as directed on the package, it will keep the survivor relatively efficient. If enough other food can be found, rations should be conserved for emergency use.

b. Consideration must be given to available food and water and how long the survival episode may last. Environmental conditions must also be considered. If a survivor is in a cold environment, more of the *proper* food will be required to provide necessary body heat. Rescue may vary from a few hours to several months, depending on the environment, operational commitments, and availability of rescue resources in that area. Available food must be rationed based on the estimated time which will elapse before being able to supplement issued rations with natural foods. If it is decided that some of the survivors should go for help, each traveler should be given twice as much food as those remaining behind. In this way, the survivors resting at the encampment and those walking out will stay in about the same physical condition for about the same length of time.

c. If available water is less than a quart a day, avoid dry, starchy, and highly seasoned foods and meat. Keep in mind that eating increases thirst. For water conservation, the best foods to eat are those with high carbohydrate content, such as hard candy and fruit. All work requires additional food and water. When work is being performed, the survivor must increase food and water consumption to maintain physical efficiency. If food is available, it is alright to nibble throughout the day. It is preferable though to have at least two meals a day, with one being hot. Cooking usually makes food safer, more digestible, and palatable. The time spent cooking will provide a good rest period. On the other hand, some food such as sapodilla, star apple, and soursop, are not palatable unless eaten raw.

d. Native foods may be more appetizing if they are eaten by themselves. Rations and native foods usually do not mix well. In many countries, vegetables are often contaminated by human feces which the natives use as fertilizer. Dysentary is transmitted in this way. If possible, survivors should try to select and prepare their own meals. If necessary to avoid offending the natives, indicate that religious beliefs or taboos require self-preparation of food.

e. Learn to overcome food prejudices. Foods that may not look good to the survivor are often a part of the natives regular diet. Wild foods are high in mineral and vitamin content. With a few exceptions, all animals are edible when freshly killed. Avoid strange looking fish and fish with flesh that remains indented when depressed as it is probably becoming spoiled and should not be eaten. With knowledge and the ability to overcome food prejudices, a survivor can eat and sustain life in strange or hostile environment.

18-4. Animal Food. Animal food gives the most food value per pound. Anything that creeps, crawls, swims, or flies is a possible source of food. People eat grasshoppers, hairless caterpillars, wood-boring bettle larvae and pupae, ant eggs, spider bodies, and termites. Such insects are high in fat and should be cooked until dried. Everyone has probably eaten insects contained in flour, cornmeal, rice, beans, fruits, and greens in their daily foods.

a. Man as a Predator. To become successful in hunting, the hunter must go through a behavioral change and reorganize personal priorities. This means the one and only goal for the present is to kill an animal to eat. To kill this animal, the hunter must mentally become a predator. The hunter must be prepared to undergo stress in order to hunt down and kill an animal. Because of the type of weapons survivors are likely to have, it will be necessary to get very close to the animal to immobilize or kill it. This is going to require all the stealth and cunning survivors can muster. In addition to stealth and cunning, knowledge of the animal being hunted is very important. If in an unfamiliar area, survivors may learn much about the animal life of the area by studying signs such as trails, droppings, and bedding areas.

b. Animal Sign. The survivor should establish the general characteristics of the animals. The size of the tracks will give a good idea of the size of the animal. The depth of the tracks will indicate the weight of the animal. The animal dung can tell the hunter much. For example, if it is still warm or slimy, it was made very recently; if there is a large amount scattered around the area, it could well be a feeding or bedding area. The droppings may indicate what the animal feeds upon. Carnivores often have hair and bone in the dung; herbivores have coarse portions of the plants they have eaten. Many animals mark their territory by urinating or scraping areas on the ground or trees. These signs could indicate good trap or ambush sites. Following the signs (tracks, droppings, etc.) may reveal the feeding, watering, and resting areas. Well worn trails will often lead to the animal's watering place. Having made a careful study of all the signs of the animal, the hunter is in a much better position to procure it, whether electing to stalk, trap, or snare it, or lie in wait to shoot it.

c. Hunting. If survivors elect to hunt, there are some basic techniques which will be helpful and improve chances of success. Wild animals rely entirely upon their senses for their preservation. These senses are smell, vision, and hearing. Humans have lost the keenness of some of their senses like smelling, hearing, etc. To overcome this disadvantage, they have the ability to reason. As an example, some animals have a fantastic sense of smell, but this can be overcome by approaching the quarry from a downwind direction. The best times to hunt are at dawn and dusk as animals are either leaving or returning to their bedding areas. Both diurnal and nocturnal animals are active at this time. There are five basic methods of hunting:

(1) Still or Stand. This is the best method for inexperienced hunters as it involves less skill. The main principle of this method is to wait in ambush along a well-used game trail, until the quarry approaches within killing range. Morning and evening are usually the best times to still hunt. Care should be taken not to disturb the area; always wait downwind. Patience and self-control are necessary to remain motionless for long periods of time.

(2) Stalking. "Stalking" refers to the stealthily approach toward game. This method is normally used when an animal has been sighted and the hunter then proceeds to close the distance using all available cover. Stalking must be done slowly so that minimum noise is made; quick movement is easily detected by the animal. Always approach from the downwind side and move when the animal's head is down eating, drinking, or looking in another direction. The same techniques are

used in blind stalking as in the regular stalk, the main difference being that the hunter is stalking a position where the animal is expected to be while the animal is not in sight.

(3) Tracking. Tracking is very difficult unless conditions are ideal. This method involves reading all of the signs left behind by the animal, interpreting what the animal is doing, and how it can best be killed. The most common signs are trails, beds, urine, droppings, blood, tracks, and feeding signs.

(4) Driving. Some wild animals can be scared or driven in a direction where other hunters or traps have been set. This method is normally used where the game can be funneled; a valley or canyon is a good place to make a drive. More than one person is usually necessary to make a drive.

(5) Calling. Small predators may be called in by imitating an injured animal. Ducks and geese can be attracted by imitating their feeding calls. These noises can be made by sucking on the hand, blowing on a blade of grass or paper, sucking the lip, or using specially designed devices. Survivors should not call animals unless they know what they are doing as strange noises may "spook" the animal.

d. Killing Implements. It is difficult to kill animals of any size without using some type of tool or weapon. As our technology has increased in complexity, so have our killing tools. If a firearm is available, a basic knowledge of shooting and hunting techniques is necessary.

(1) Learning to become proficient with primitive weapons is important. Many primitive tribes of the world are still effectively using spears, clubs, bows and arrows, sling shots, etc., to provide food for their families. One of the limiting factors in the use of firearms is the amount of ammunition on hand. Therefore, a survivor cannot afford to waste ammunition on moving game or game which is beyond the effective range of the firearm being used. Wait for a pause in the animal's motions. The shot must be placed in a vital area with any firearm. Aim for the brain, spine, lungs, or heart (figure 18-2). A hit in these areas is usually fatal.

(2) A full-jacketed bullet often won't immediately down a larger animal hit in a vital area such as the lungs or heart. The alternative to losing the animal is tracking it to where it falls. Often it's better to wait awhile before pursuing the animal. If not pursued, it may lay down and stiffen or perhaps bleed to death. Follow the blood trail to where the game has gone down and kill it if it is

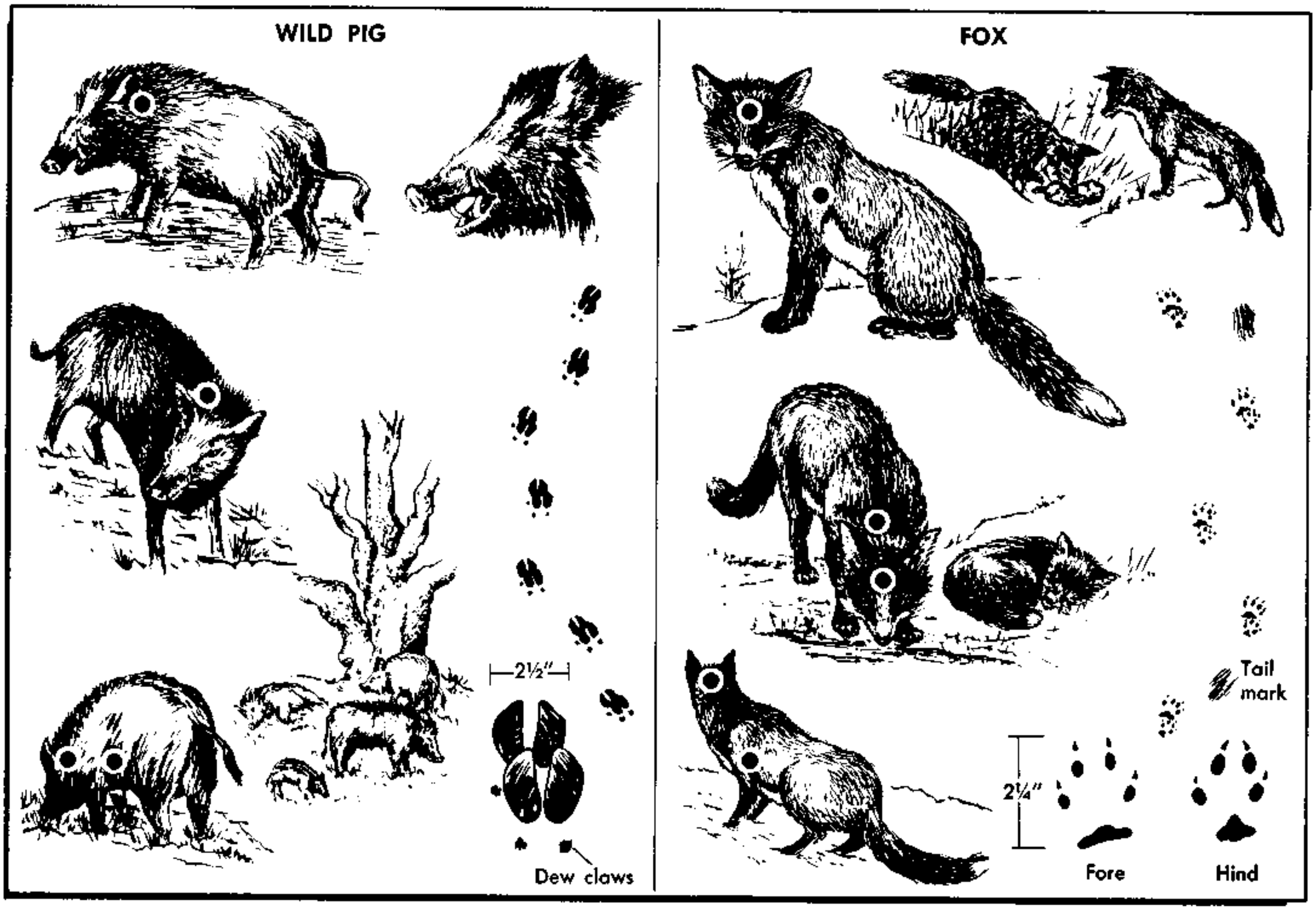

Figure 18-2. Shooting Game.

still alive. Even though ammunition might be limited, small game may be more productive than large game. Although they present smaller targets and have less meat, they are less wary, more numerous, and travel less distance to escape if wounded. A large amount of edible meat on small game can be destroyed from a bullet wound. On rodents, most of the meat is on the hindquarters and frontquarters; birds, it is the breast and legs. The survivor should try to hit a vital spot that spoils the least meat.

(3) Night hunting is usually best, since most animals move at night. A flashlight or torch may be used to shine in the animal's eyes. It will be partly blinded by the light and a survivor can get much closer than in the daytime. If no gun is available, the animal can be killed with a club or a sharpened stick used as a spear.

(4) Remember that large animals, when wounded, cornered, or with their young, can be dangerous. Be sure the animal is dead, not just wounded, unconscious, or playing "possum." Animals usually die with their eyes open and glazed-over. Poke all "dead" animals in the eye with a long sharp stick before approaching them.

(5) Small freshwater turtles can often be found sunning themselves along rivers and lakeshores. If they dash into shallow water, they can still be procured with nets, clubs etc; watch out for mouth and claws. Frogs and snakes also sun and feed along streams. Use both hands to catch a frog—one to attract it and keep it busy while grabbing it with the other. Bright cloth on a fishhook also works. All snakes are good eating and can be killed with a long stick. Both marine and dry-land lizards are edible. A noose, small fishhook baited with a bright cloth lure, slingshot, or club can be used. A slingshot can be made with a forked stick and the elastic from the parachute pack or surgical tubing found in some survival kits (figure 18-3). With practice, the slingshot can be very effective for killing small animals.

e. Snaring and Trapping. Snaring and trapping animals are ways survivors can procure animal food to supplement issued rations. Since small animals are usually more abundant than large animals, they will probably be the survivor's main source of food. Snares should be set out on a 15:1 ratio; 15 snares should be set out for every one animal expected to be caught.

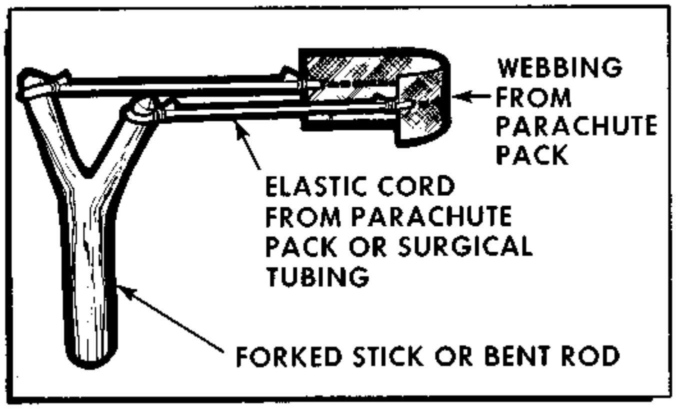
WEBBING FROM PARACHUTE PACK
ELASTIC CORD FROM PARACHUTE PACK OR SURGICAL TUBING
FORKED STICK OR BENT ROD

Figure 18-3. Slingshot.

(1) Using traps and snares are more advantageous then going out on foot and physically hunting the animal. The most important advantage being that traps work 24 hours a day with no assistance from the hunter. A large area can be effectively strapped with the possibility of catching many animals within the same period of time. Survivors (generally) use much less energy maintaining a trapline than is used by hunting. This means less food is required because less energy is expended.

(2) The traps or snares should be set in areas where the game is known to live or travel. Look for signs such as tracks, droppings, feeding signs, or actual sightings of the animal. If snares are used, they should be set up to catch the animal around the neck. Therefore, the loop must allow the head to pass through but not the body. Loops will vary in size from one animal to another. When placing snares, try to find a narrow area of the game trail where the animal has no choice but to enter the loop. If a narrow area cannot be found, brush or other obstacles can be arranged to funnel the animal into the snare (figure 18-4). Do not overdo the funneling; use as little as possible. Avoid disturbing the natural surroundings if possible. Do not walk on game trails, but approach 90 degrees to the trail, set the snare, and back away. Snares may also be set over holes or burrows. All snares and traps should be set during the midday because most animals are nocturnal in nature. Check snares and traps twice daily. If possible, check after sunup and before sunset. The checks should be made from a distance so any animals moving at the time of checking will not be disturbed or frightened away.

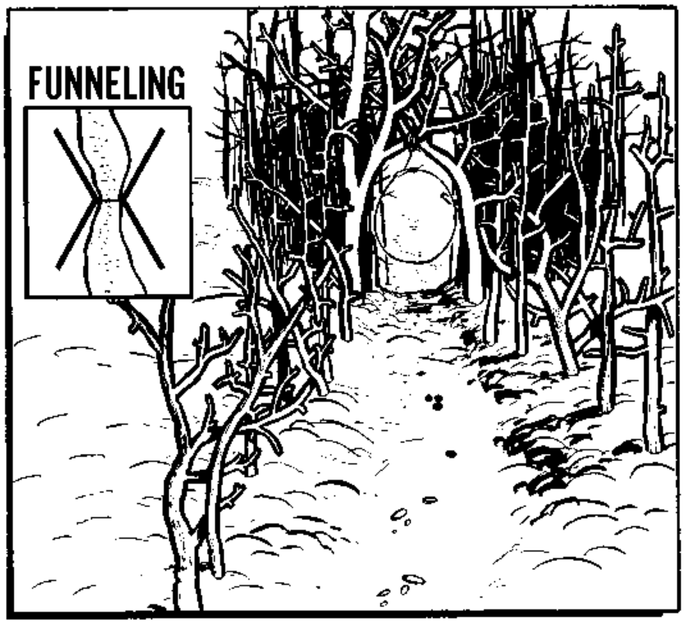

Figure 18-4. Funneling.

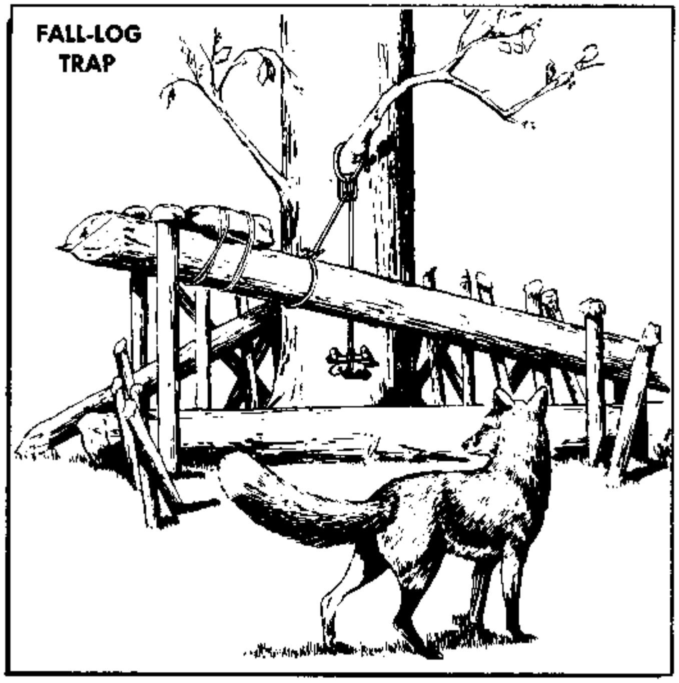

Figure 18-5. Mangle.

(3) There are three ways to immobilize or trap animals.

(a) Strangle. This is done by simply using a free-sliding noose which, when tightened around the neck, will restrict circulation of air and blood. The materials should be strong enough to hold the animal; for example, suspension line, string, wire, cable, or rawhide.

(b) Mangle. Mangle traps use a weight which is suspended over the animal's trail or over bait. When the animal trips the trigger, the weight (log) will descend and mangle the animal (figure 18-5).

(c) Hold. Any means of impeding the animal and detaining its progress would be considered a hold-type trap.

(4) The apache foot snare is an example of a hold-type trap. It is used for large browsers and grazers like deer (figure 18-6). It should be located along game trails where an obstruction, such as a log, blocks the trail. When animals jump over this obstruction, a very shallow depression is formed where their hooves land. The apache foot snare should be placed at this depression. The box trap for birds is another example of hold-type traps (figure 18-7).

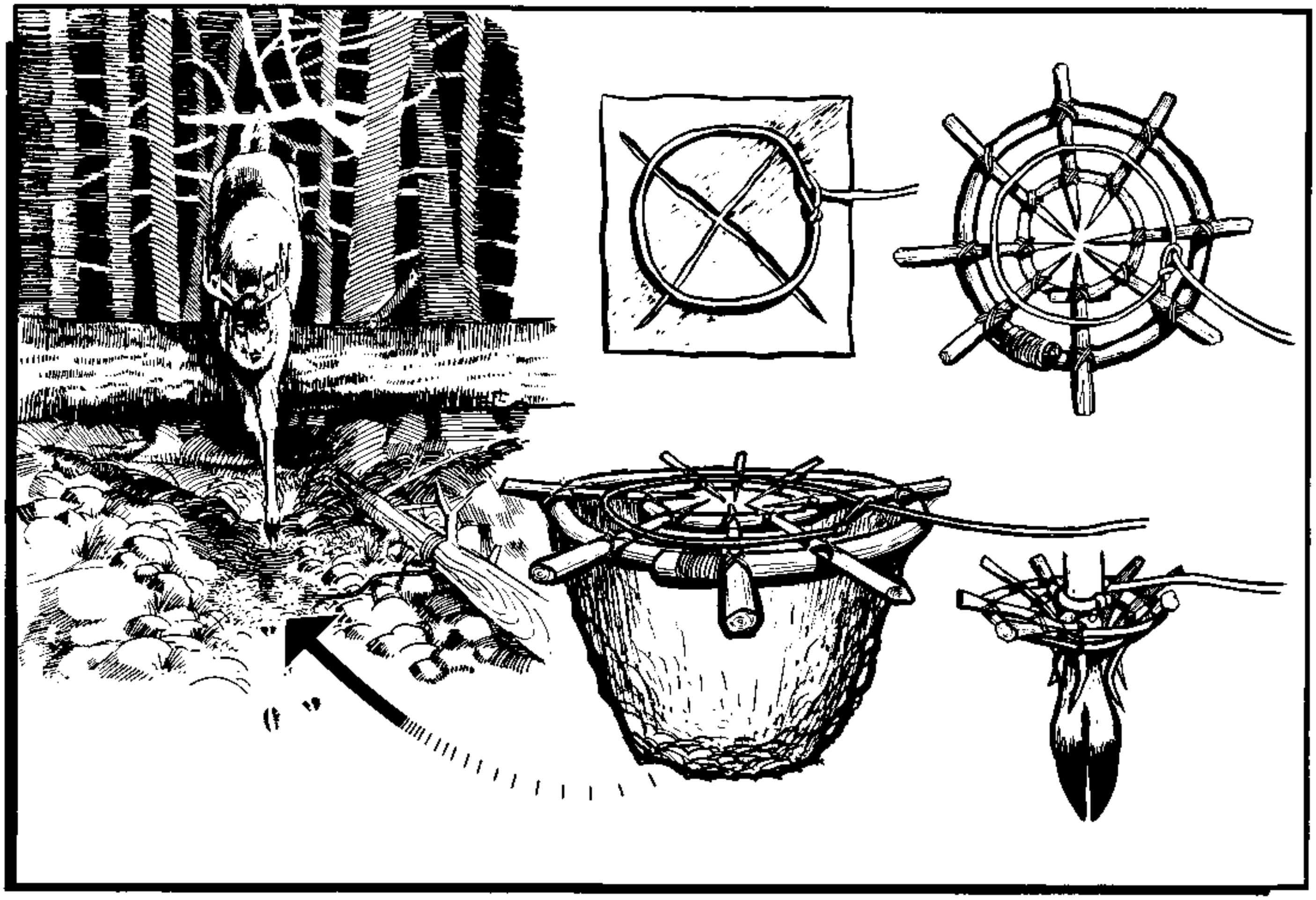

Figure 18-6. Apache Foot Snare.

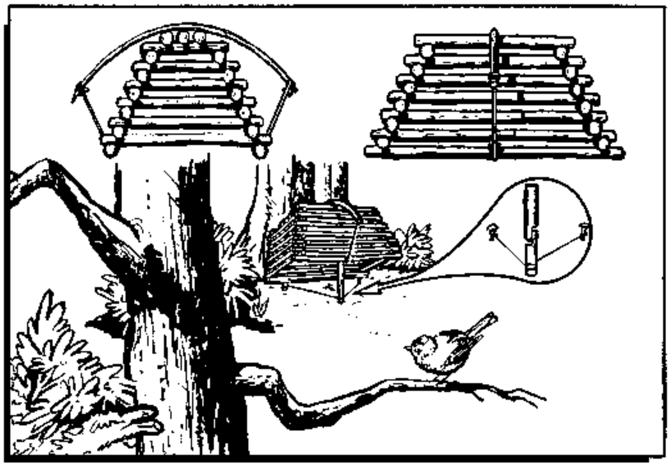

Figure 18-7. Box Trap.

(5) The simple loop is the quickest snare to construct. All snares and traps should be simple in construction with as few moving parts as possible. This loop can be constructed from any type of bare wire, suspension line, inner core, vines, long strips of green bark, clothing strips or belt, and any other material that will not break under the strain of holding the animal. If wire is being used for snares, a figure "8" or locking loop should be used (figure 18-8). Once tightened around the animal, the wire is locked into place by the figure "8" which prevents the loop from opening again. A simple loop snare is generally placed in the opening of a den, with the end of the snare anchored to a stake or similar object (figure 18-8). The simple loop snare can also be used when making a squirrel pole (figure 18-9) or with some types of trigger devices.

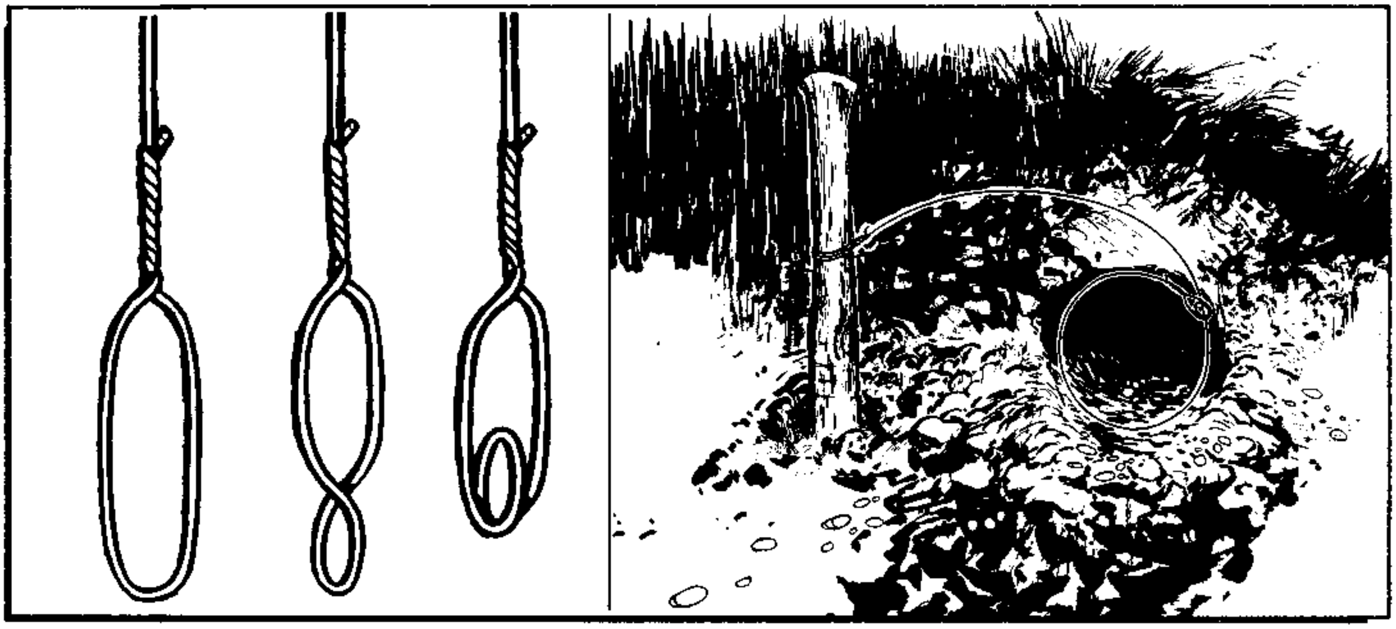

Figure 18-8. Locking Loop and Setting Noose.

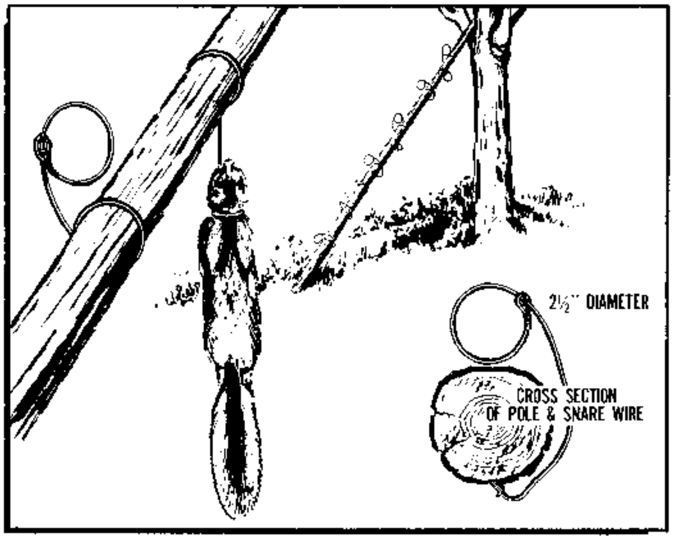
2½" DIAMETER
CROSS SECTION OF POLE & SNARE WIRE

Figure 18-9. Squirrel Pole.

f. Triggers. Triggers may be used with traps. The purpose of the trigger is to set the device in motion, which will eventually strangle, mangle, or hold the animal. There are many triggers. Some of the more common ones are:

(1) Two-pin toggle with a counterweight for small to medium animals which are lifted out of the reach of predators (figure 18-10).

(2) Figure "H" with wire snare for small mammals and rodents (figure 18-11).

(3) Canadian ace for predators such as bobcat, coyote, etc., (figure 18-12).

(4) Three-pin toggle with deadfall for medium to large animals (figure 18-13). Medium and large animals can be captured using deadfalls, but this type of trap is recommended only when big game exists in large quantities to justify the great expense of time and effort spent in constructing the trap.

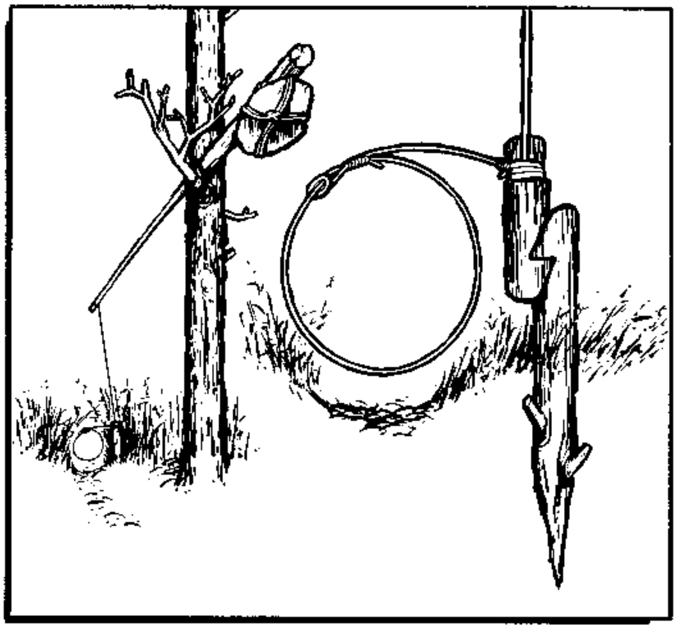

Figure 18-10. Two-Pin Toggle.

(5) The twitch-up snare which incorporates the simple loop, can be used to catch small animals (figure 18-14). When the animal is caught, the sapling jerks it up into the air and keeps the carcass out of the reach of predators. This type of snare will not work well in cold climates, since the bent sapling will freeze in position and not spring up when released.

(6) A long forked stick can be used as a twist stick to procure ground squirrels, rabbits, etc. A den that has signs of activity must be located. Using the long forked stick, the survivor probes the hole with the forked end until something soft is felt then twisting the stick will entangle the animal's hide in the stick and the animal can be extracted (figure 18-18).

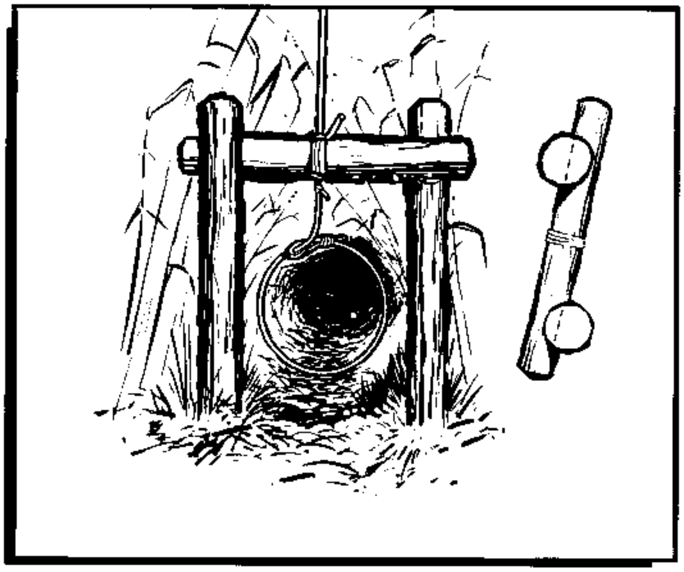

Figure 18-11. Figure H.

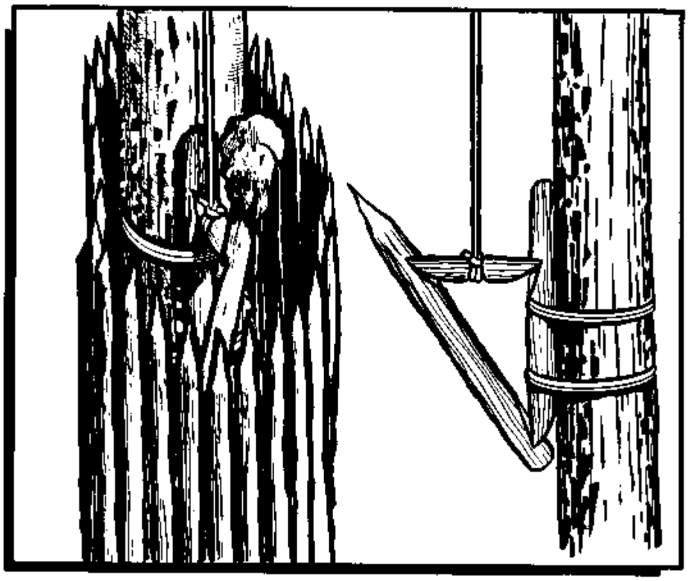

Figure 18-12. Canadian Ace.

g. Birds. Birds can be caught with a gill net. The net should be set up at night vertically to the ground in some natural flyway, such as an opening in dense foliage. A small gill net on a wooden frame with a disjointed stick for a trigger can also be used. A gill net can be made by using inner core from parachute suspension line (figure 18-20).

(1) Birds can be caught on baited fishhooks (figure 18-15) or simple slipping loop snares. Bird's nest can be a source of food. All bird eggs are edible when fresh. Large wading birds such as cranes and herons often nest in mangrove swamps or in high trees near water.

(2) During molting season, birds cannot fly because of the loss of their "flight" feathers; they can be procured by clubbing or netting.

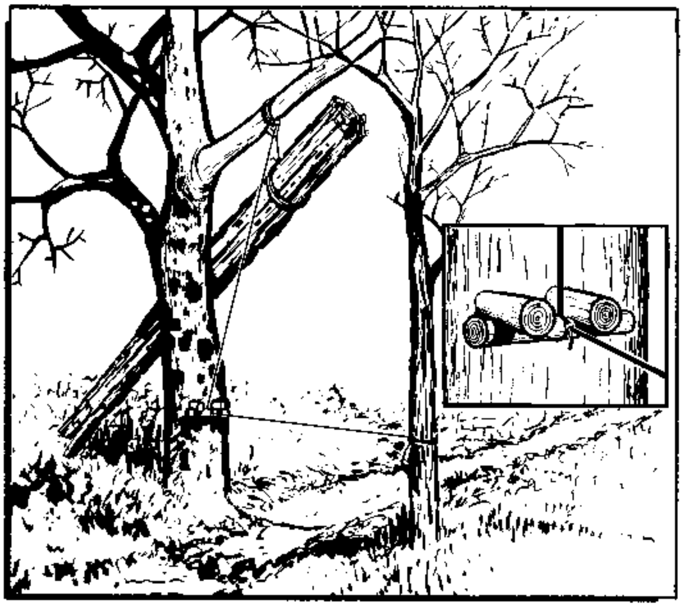

Figure 18-13. Three-Pin Toggle.

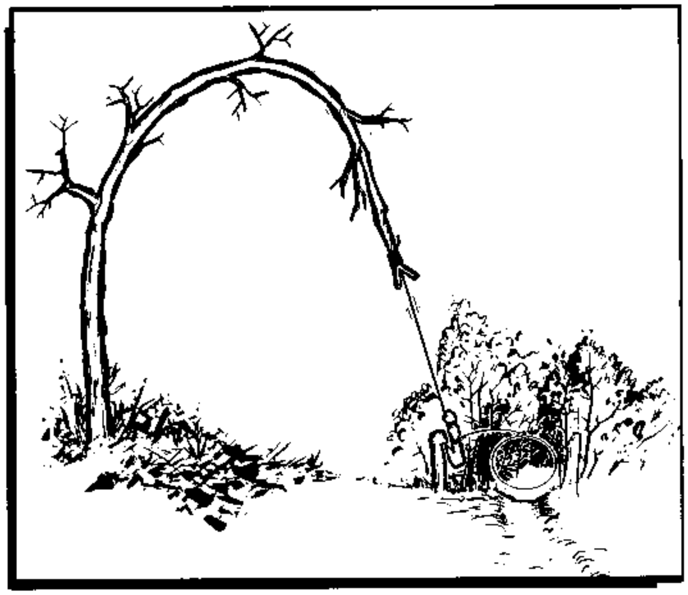

Figure 18-14. Twitch Up.

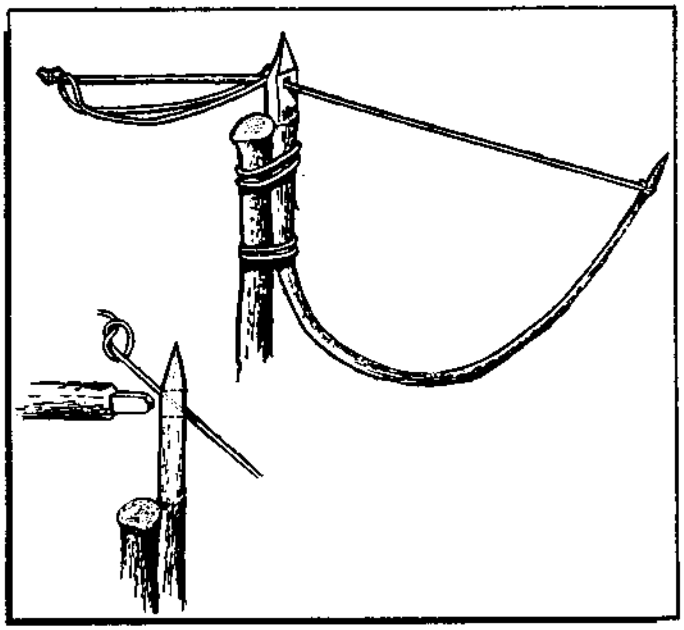

Figure 18-16. Ojibwa Bird Snare.

(3) Birds can be also caught in an Ojibway snare. This snare is made by cutting a 1- or 2-inch thick sapling at a height of 4½ to 5 feet above the ground (figure 18-16). A springy branch is then whittled flat at the butt end and a rectangular hole is cut through the flattened end. One end of a ½-inch thick stick, 15 inches long, is then whittled to fit slightly loose in the hole and the top corner of the whittled end is rounded off so the stick will easily drop away from the hole. The branch is then tied by its butt end to the top of the sapling. A length of inner core from suspension line is tied to the bottom end of the branch and the branch is bent into a bow with the line passing through the hole in the butt end. A knot is tied in the line and the 15-inch stick is then placed in the hole to lock the line in place (just behind the knot). An 8-inch loop is made at the end of the line and laid out on the 15-inch stick (spread out as well as possible). A piece of bait is placed on top of the sapling, and when a bird comes to settle on the 15-inch stick, the stick drops from the hole causing the loop to tighten around the bird's legs.

(4) When many birds frequent a particular type of bush, some simple loop snares may be set up throughout the bush. Make the snares as large as necessary for the particular type of birds that come to perch, feed, or roost there (figure 18-17).

(5) In wild, wooded areas, many larger species of birds such as spruce grouse and ptarmigan may be approached. The spruce grouse, which has merited the name of "fools hen," can be approached and killed with

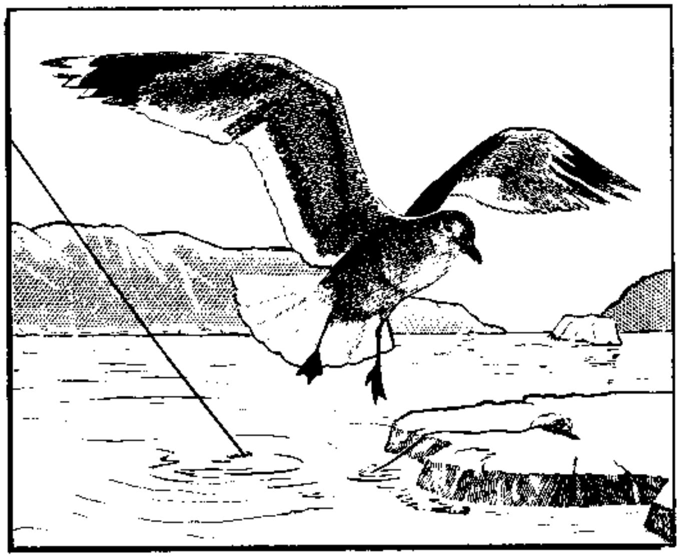

Figure 18-15. Baited Fishhook.

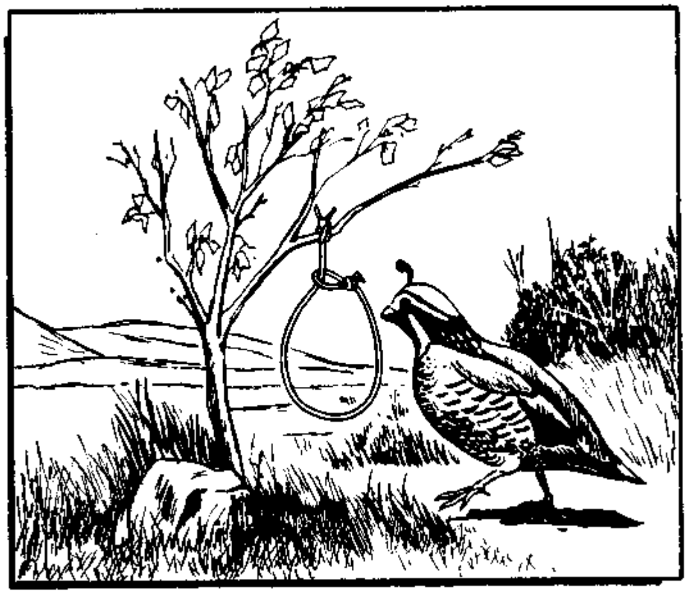

Figure 18-17. Ptarmigan or Small Game Snare.

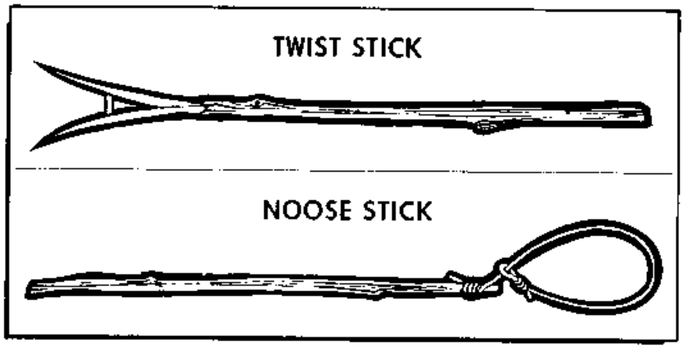

Figure 18-18. Twist Stick and Noose Stick.

a stick with little trouble. It often sits on the lower branches of trees and can be easily caught with a long stick with a loop at the end (figure 18-18).

(6) Ground feeding birds (Quail, Hungarian Partridge, Chukar) can be trapped in a trench dug into the ground. The trench should be just wide enough for the bird to walk into, so survivors must first observe the type of ground feeding birds in the area. The trench should be 2 to 3 feet long and about 10 to 12 inches deep at the deep end. The other end of the trench should be ramped down from the surface level. Bait is scattered along the surface into the pit, and after having pecked the last piece of bait the bird will not be able to get out of the pit because it can't fly out or climb out, its feathers keep it from backing out, and it can't turn around to walk out.

(7) Perching birds may be captured by using bird lime. Bird lime is a term applied to any sticky or gluey substance which is rubbed on a branch to prevent the flight of a bird which has landed on it or has flapped a wing against it. Bird lime is usually made from the sap of plants in the Euphorbia family. The common names of some of these plants are spotted spurge, cypress spurge, snow-on-the-mountain, and poinsettias. The Euphorbias have a wide range in North and Central America. The milky sap is poisonous and may cause blisters on the skin and should be handled with care. Bird lime is most effective in the desert and jungle, but

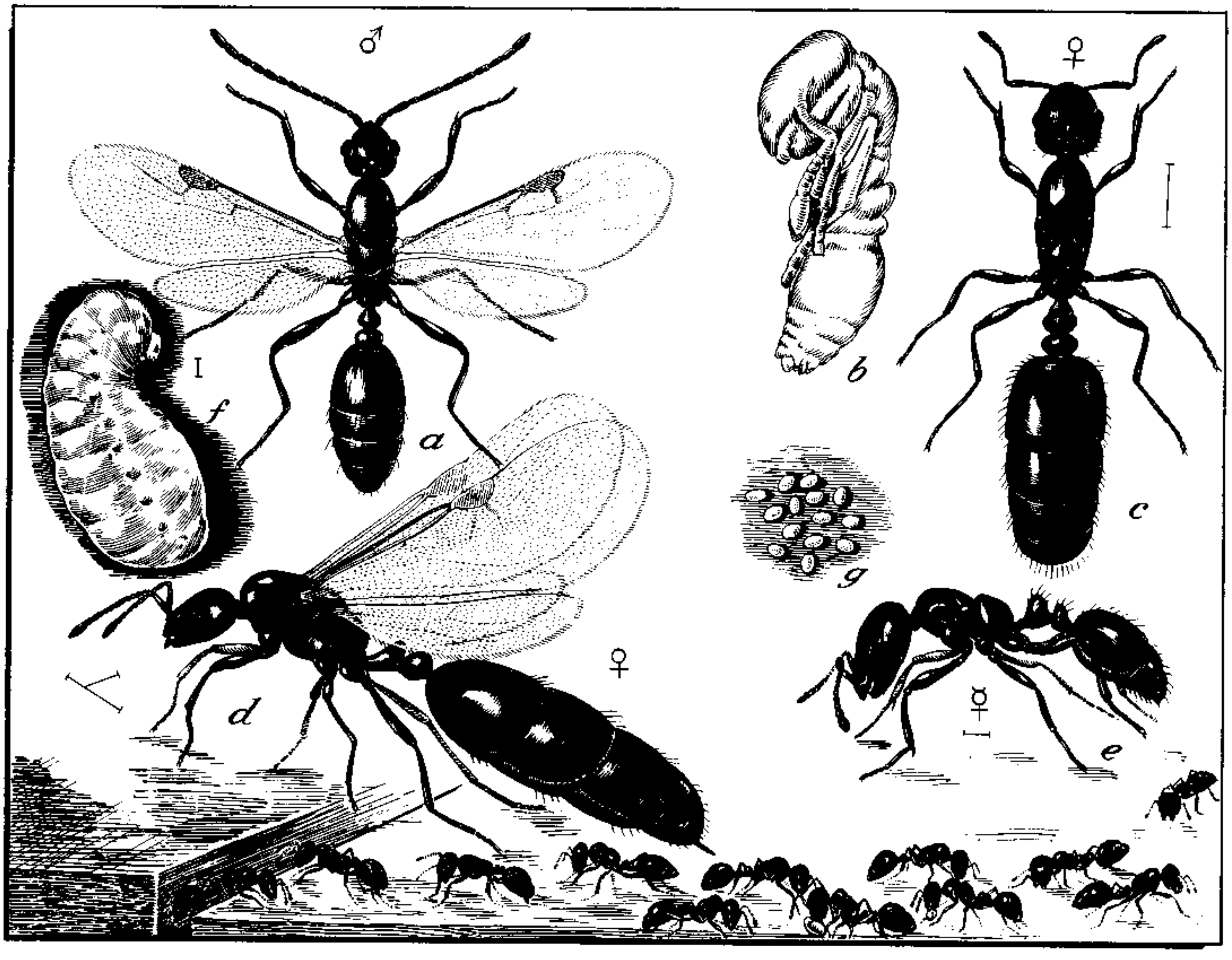

Figure 18-19. Ants.

it will not work in cold weather. Dust will make bird lime ineffective, so it should be used in spots where dust is not prevalent. The sap of the breadfruit tree makes excellent bird lime as it swells and become glutinous upon contact with air.

h. Insects. If there *ever* is a time when food aversions must be overcome, it is when survivors turn to insects as a food source.

(1) Primitive peoples eat insects and consider them great delicacies. When food is limited and insects are available, they can become a valuable food source. In some places, locusts and grasshoppers, cicadas, and crickets are eaten regularly; occasionally termites, ants, and a few species of stonefly larvae are consumed. Big beetles such as the Goliath Beetle of Africa, the Giant Water Beetles, and the big Long Horns are relished the world over. Clusters, like those of the Snipefly Atherix (that overhang the water), and the windrows of Brinefly puparia are eaten. Aquatic water bugs of Mexico are grown especially for food. All stages of growth can be eaten, including the eggs but, the large insects must be cooked to kill internal parasites.

(2) Termites and white ants are also an important food source. Strangely enough, these are closely related to cockroaches. The reason they are eaten so extensively in Africa is the fact that they occur in enormous numbers and are easily collected both from their nests and during flight. They are sometimes attracted to light in unbelievable numbers and the natives become greatly excited when the large species appear.

(3) Many American Indian tribes made a habit of eating the large carpenter ants that are sometimes pests in houses. These were eaten both raw and cooked. Even today the practice of eating them has not entirely disappeared, although they do not form an essential part of the diet of any of the inhabitants of this country (figure 18-19).

(4) It is not at all unnatural that the American Indians should have relished the honey ants in all parts of the continent where they occur. These ants are peculiar in that some of the workers become veritable storehouses for honey, their abdomens becoming more or less spherical and so greatly enlarged that they are scarcely able to move. They cluster on the ceilings and walls of their nests and disgorge part of their stored food to other inhabitants. The Indians discovered the sweetness stored in these insects and made full use of it. At first they ate the ants alive, later gathering them in quantity and crushing them so that they formed an enticing dish—one which was considered a delicacy and served to guests of distinction as a special favor. The next step in the use of the honey ant was the extraction of the pure honey by crushing the insects and straining the juices. After the honey was extracted, it was allowed to ferment, forming what is said to be a highly flavored wine.

(5) Indians of the American tropics, with a much larger ant fauna from which to choose, select the queens of the famous leafcutting or so-called umbrella ants upon which to feed, eating only the abdomens, either raw or cooked.

(6) It is natural that caterpillers, the larvae of moths and butterflies, should form a very substantial part of the food of primitive peoples because these are often of large size or occur in great abundance. In Africa, many tribes consider caterpillers choice morsels of food, and much time is spent in collecting them. Some of the native tribes recognize 20 or more different kinds of caterpillars that are edible, and are sufficiently well acquainted with the life history of the insects to know the plants upon which they feed and the time of year when they have reached the proper stage of development for collecting. Caterpillars with hairs should be avoided. If eaten, the hairs may become lodged in the throat causing irritation or infection. Today it is known that insects have nutritional or medicinal value. The praying mantis, for example, contains 58 percent protein, 12 percent fat, 3 percent ash, vitamin B complex, and vitamin A. The insect's outer skeleton is an interesting compound of sugar and amino acids.

(7) Bee larvae were eaten by the ancient Chinese. Some Chinese today eat locusts, dragonflies, and bumblebees. Cockroaches and locusts are a favorite dish in Szechuan. In Kwangtun, grasshoppers, golden June beetles, crickets, wasp larvae, and silkworm larvae are used for food.

(8) Stinging insects should have their stinging apparatus removed before they are eaten.

(9) As can be seen, insects have been used as a food source for thousands of years and will undoubtedly continue to be used. If survivors cannot overcome their aversion to insects as a food source, they will miss out on a valuable and plentiful supply of food.

i. Fishing. Fishing is one way to get food throughout the year wherever water is found. There are many ways to catch fish which include hook and line, gill nets, poisons, traps, and spearing.

(1) If an emergency fishing kit is available, there will be a hook and line in it, but if a kit is not available a hook and line will have to be procured elsewhere or improvised. Hooks can be made from wire or carved from bone or wood. The line can be made by unraveling parachute suspension line or by twisting threads from clothing or plant fibers. A piece of wire between the fishing line and the hook will help prevent the fish from biting through the line. Insects, smaller fish, shellfish, worms, or meat can be used as bait. Bait can be selected by observing what the fish are eating. Artificial lures can be made from pieces of brightly colored cloth, feathers, or bits of bright metal or foil tied to a hook. If the fish will not take the bait, try to snag or hook them in any part of the body as they swim by. In freshwater, the deepest water is usually the best place to fish. In shallow streams, the best places are pools below falls, at the foot of rapids, or behind rocks. The best time to fish is usual-

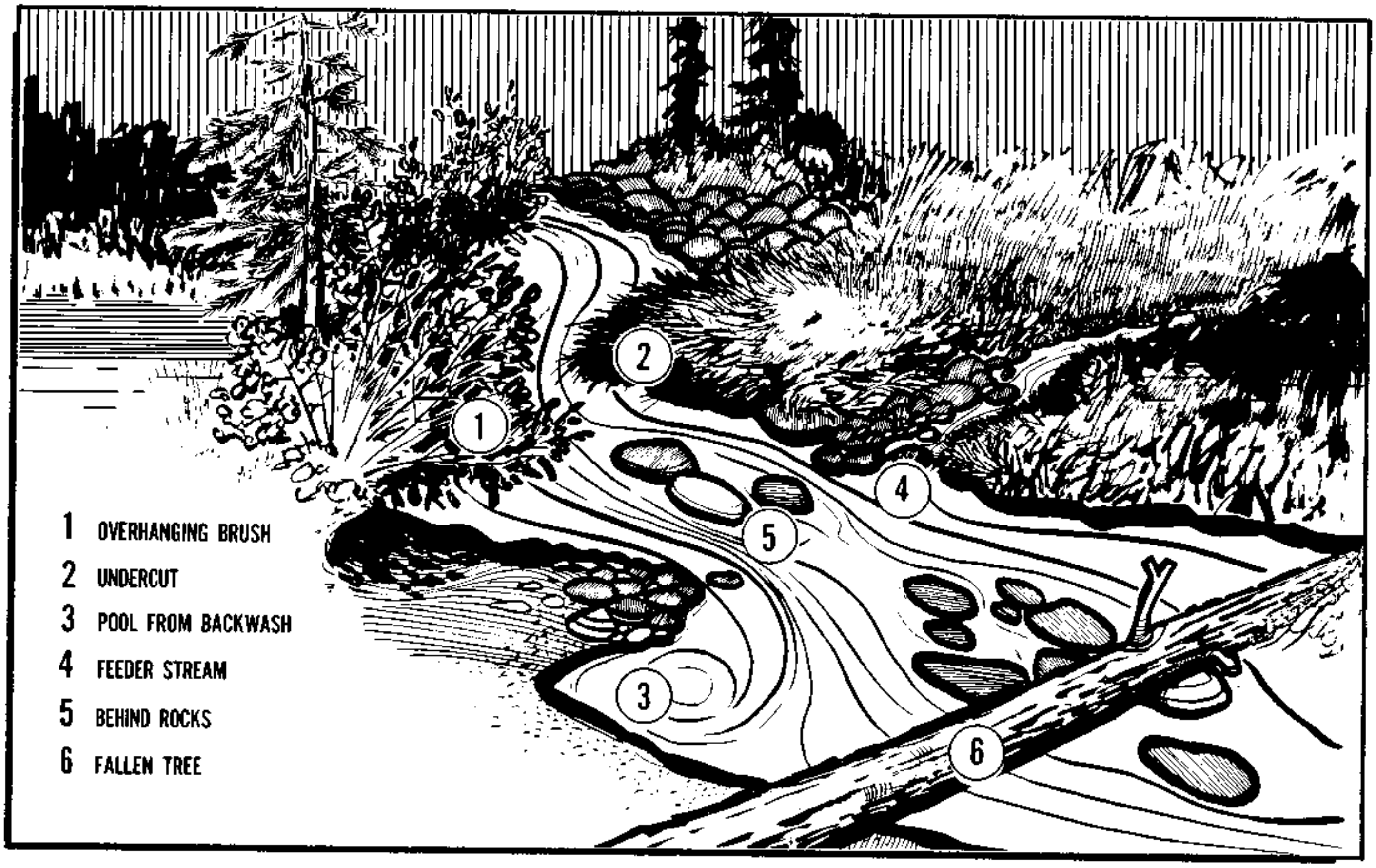

Figure 18-20. Fishing Places.

ly early morning or late evening (figure 18-20). Sometimes fishing is best at night, especially in moonlight or if a light is available to attract the fish. The survivor should be patient and fish at different depths in all kinds of water. Fishing at different times of the day and changing bait often is rewarding.

(2) The most effective fishing method is a net because it will catch fish without having to be attended (figures 18-21 and 18-22). If a gill net is used, stones can be used for anchors and wood for floats. The net should be set at a slight angle to the current to clear itself of any floating refuse that comes down the stream. The net should be checked at least twice daily (figure 18-23). A net with poles attached to each end works effectively if moved up or down a stream as rapidly as possible while moving stones and threshing the bottom or edges of the streambanks. The net should be checked every few moments so the fish cannot escape.

(3) Shrimp (prawns) live on or near the sea bottom and may be scraped up. They may be lured to the surface with light at night. A hand net made from parachute cloth or other material is excellent for catching shrimp. Lobsters are creeping crustaceans found on or near the sea bottom. A lobster trap, jig, baited hook, or dip net can be used to catch lobster. Crabs will creep, climb, and burrow and are easily caught in shallow water with a dip net or in traps baited with fish heads or animal viscera.

(4) Fishtraps (figure 18-24) are very useful for catching both freshwater and saltwater fish, especially those that move in schools. In lakes or large streams, fish tend to approach the banks and shallows in the morning and evening. Sea fish, traveling in large schools, regularly approach the shore with the incoming tide, often moving parallel to the shore guided by obstruction in the water.

(a) A fishtrap is basically an enclosure with a blind opening where two fence-like walls extend out, like a funnel, from the entrance. The time and effort put into building a fishtrap should depend on the need for food and the length of time survivors plan to stay in one spot.

(b) The trap location should be selected at high tide and the trap built at low tide. One to 2 hours of work should do the job. Consider the location, and try to adapt natural features to reduce the labors. Natural rock pools should be used on rock shores. Natural pools on the surface of reefs should be used on coral islands by blocking the opening as the tide recedes. Sandbars, and the ditches they enclose, can be used on sandy shores. The best fishing off sandy beaches is the lee side of offshore sandbars. By watching the swimming habits of fish, a simple dam can be built which extends out into the water forming an angle with the shore. This will trap fish as they swim in their natural path. When planning a more complex brush dam, select protected bays or inlets using the narrowest area and extending one arm almost to the shore.

(c) In small, shallow streams, the fishtraps can be made with stakes or brush set into the stream bottom or weighted down with stones so that the stream is blocked except for a small narrow opening into a stone or brush pen or shallow water. Wade into the stream, herding the fish into the trap, and catch or club them when they get in shallow water. Mud-bottom streams can be trampled until cloudy and then netted. The fish are blinded and cannot avoid the nets. Freshwater crawfish and snails can be found under rocks, logs, overhanging bushes, or in mud bottoms.

(5) Fish may be confined in properly built enclosures and kept for days. In many cases, it may be advantageous to keep them alive until needed and thus ensure there is a fresh supply without danger of spoilage. Mangrove swamps are often good fishing grounds. At low tide, clusters of oysters and mussles are exposed on the mangrove "knees" or lower branches. Clams can be found in the mud at the base of trees. Crabs are very active among branches or roots and in the mud. Fish can be caught at high tide. Snails are found on mud and clinging to roots. Shellfish which are not covered at high tide or those from a colony containing diseased members should not be eaten. Some indications of diseased shellfish are shells gaping open at low tide, foul odor, and (or) milky juice.

(6) Throughout the warm regions of the world, there are various plants which the natives use for poisoning fish. The active poison in these plants is harmful only to cold-blooded animals. Survivors can eat fish killed by this poison without ill effects.

(a) In Southeast Asia, the derris plant is widely

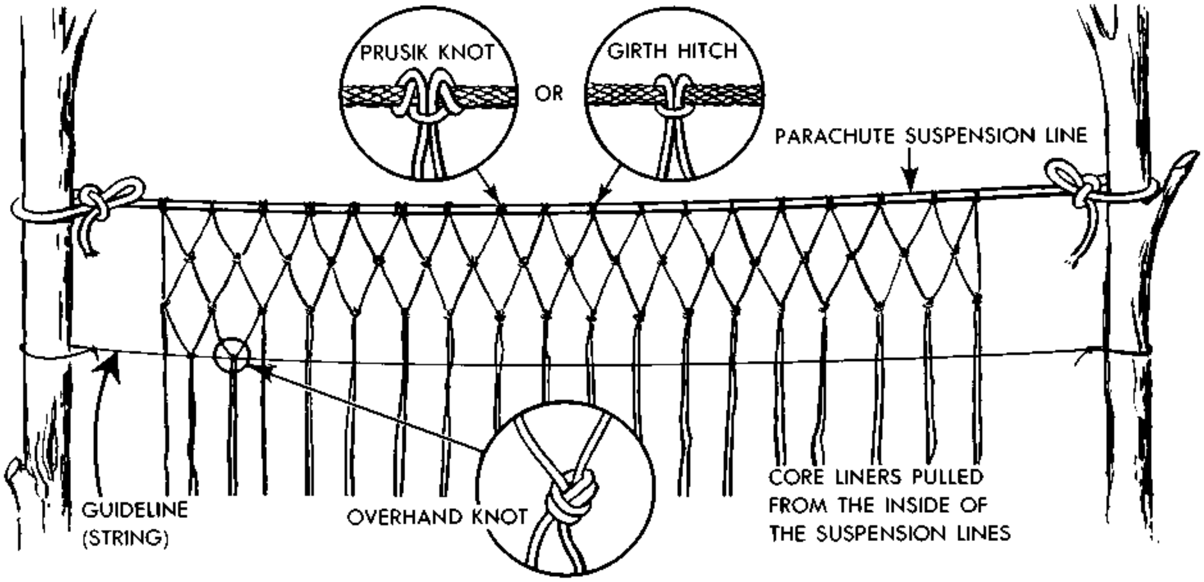
PRUSIK KNOT
OR
GIRTH HITCH
PARACHUTE SUSPENSION LINE
GUIDELINE (STRING)
OVERHAND KNOT
CORE LINERS PULLED FROM THE INSIDE OF THE SUSPENSION LINES

1. Suspend a suspension line casing (from which the core liners have been pulled) between two uprights, approximately at eye level.
2. Hang core liners (an even number) from the line suspended as in 1, above. These lines should be attached with a Prusik knot or girth hitch and spaced in accordance with the mesh you desire. One-inch spacing will result in a 1-inch mesh, etc. The number of lines used will be in accord with the width of the net desired. If more than one man is going to work on the net, the length of the net should be stretched between the uprights, thus providing room for more than one man to work. If only one man is to make up the net, the depth of the net should be stretched between the uprights and step 8, below, followed.
3. Start at left or right. Skip the first line and tie the second and third lines together with an overhand knot. Space according to mesh desired. Then tie fourth and fifth, sixth and seventh, etc. One line will remain at the end.
4. On the second row, tie the first and second, third and fourth, fifth and sixth, etc., to the end.
5. Third row, skip the first line and repeat step 3 above.
6. Repeat step 4, and so on.
7. You may want to use a guide line which can be moved down for each row of knots to ensure equal mesh. Guide line should run across the net on the side opposite the one you are working from so that it will be out of your way.
8. When you have stretched the depth between the uprights and get close to ground level, move the net up by rolling it on a stick and continue until the net is the desired length.
9. String suspension line casing along the sides when net is completed to strengthen it and make the net easier to set.

Figure 18-21. Making a Gill Net.

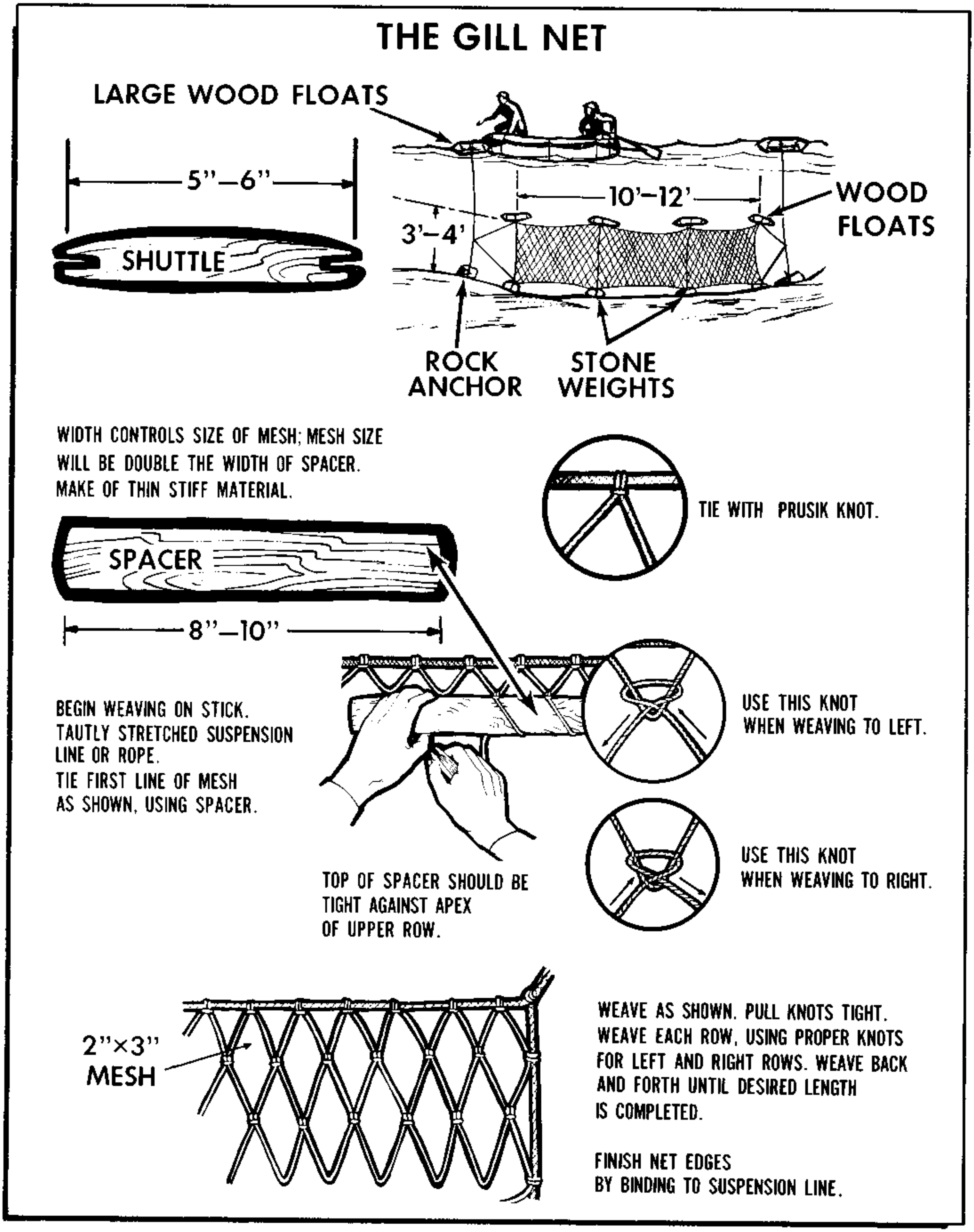
THE GILL NET
LARGE WOOD FLOATS
5"–6"
SHUTTLE
10'–12'
3'–4'
WOOD FLOATS
ROCK ANCHOR
STONE WEIGHTS
SPACER
8"–10"
2"×3" MESH

WIDTH CONTROLS SIZE OF MESH; MESH SIZE WILL BE DOUBLE THE WIDTH OF SPACER. MAKE OF THIN STIFF MATERIAL.

TIE WITH PRUSIK KNOT.

BEGIN WEAVING ON STICK. TAUTLY STRETCHED SUSPENSION LINE OR ROPE.
TIE FIRST LINE OF MESH AS SHOWN, USING SPACER.

USE THIS KNOT WHEN WEAVING TO LEFT.

USE THIS KNOT WHEN WEAVING TO RIGHT.

TOP OF SPACER SHOULD BE TIGHT AGAINST APEX OF UPPER ROW.

WEAVE AS SHOWN. PULL KNOTS TIGHT. WEAVE EACH ROW, USING PROPER KNOTS FOR LEFT AND RIGHT ROWS. WEAVE BACK AND FORTH UNTIL DESIRED LENGTH IS COMPLETED.

FINISH NET EDGES BY BINDING TO SUSPENSION LINE.

Figure 18-22. Making a Gill Net With Shuttle and Spacer.

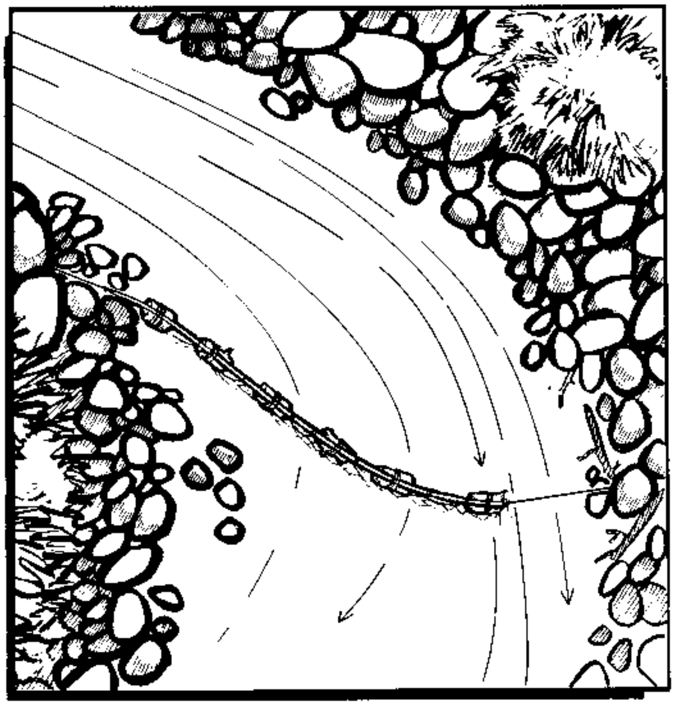

Figure 18-23. Setting the Gill Net.

used as a source of fish poison. The derris plant, a large woody vine, is also used to produce a commercial fish poison called rotenone. Commercial rotenone can be used in the same manner as crushed derris roots; it causes respiratory failure in fish, but has no ill effects on humans. However, rotenone has no effect if dusted over the surface of a pond. It should be mixed to a malted-milk consistency with a little water, and then distributed in the water. If the concentration is strong, it takes effect within 2 minutes in warm water, or it may take an hour in colder water. Fish sick enough to turn over on their backs will eventually die. An ounce of 12 percent rotenone can kill every fish for a half mile down a slow-moving stream that is about 25 feet wide. A few facts to remember about the use of rotenone are:

-1. It is very swift acting in warm water at 70°F and above.

-2. It works more slowly in cold water and is not practical in water below 55°F.

-3. It is best applied in small ponds, streams, or tidal pools.

-4. Excess usage will be wasted. However, too little will not be effective.

(b) A small container of 12 percent rotenone (one-half ounce) is a valuable addition to any emergen-

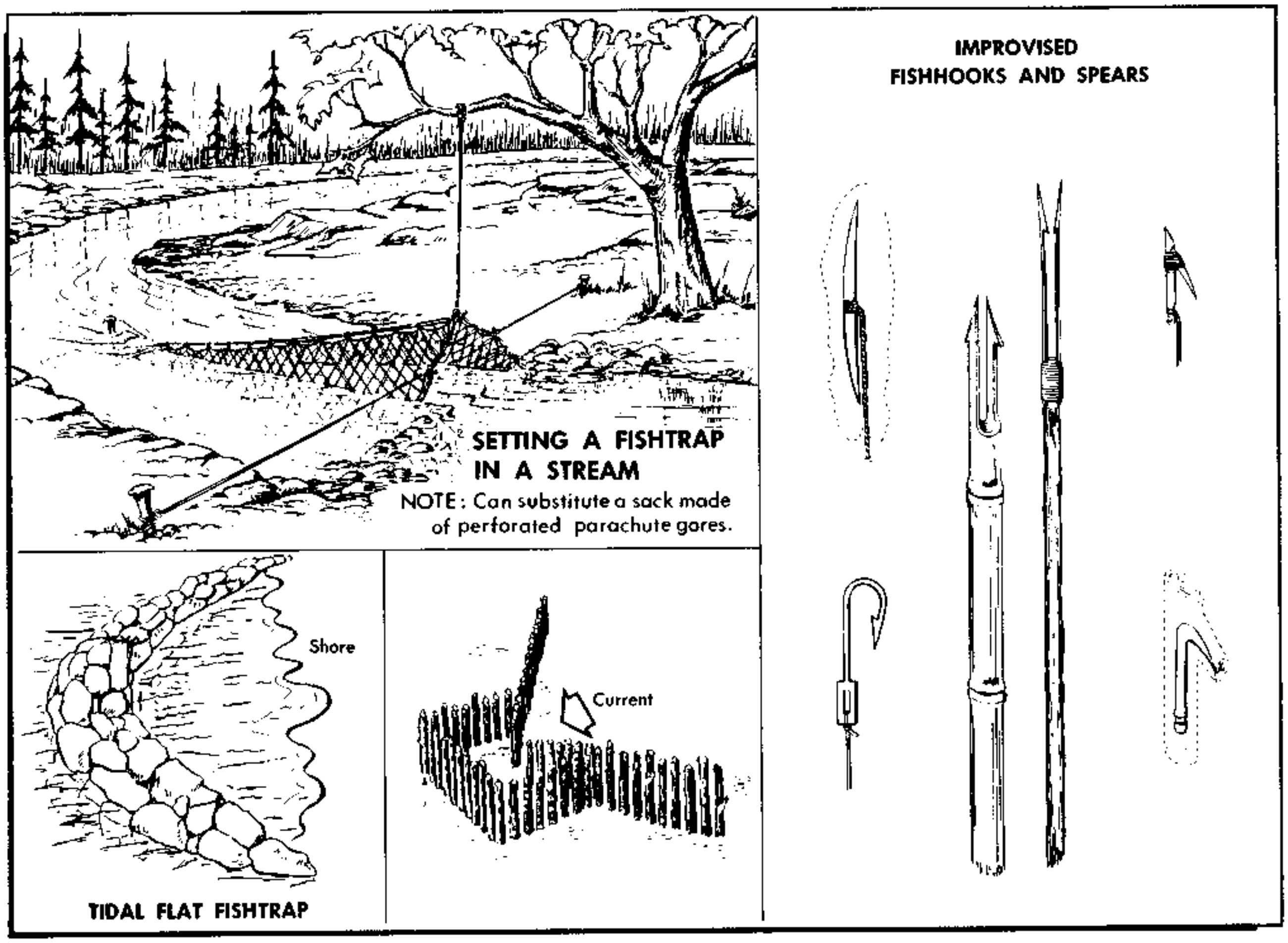

IMPROVISED FISHHOOKS AND SPEARS

SETTING A FISHTRAP IN A STREAM

NOTE: Can substitute a sack made of perforated parachute gores.

TIDAL FLAT FISHTRAP

Figure 18-24. Maze-type Fishtraps.

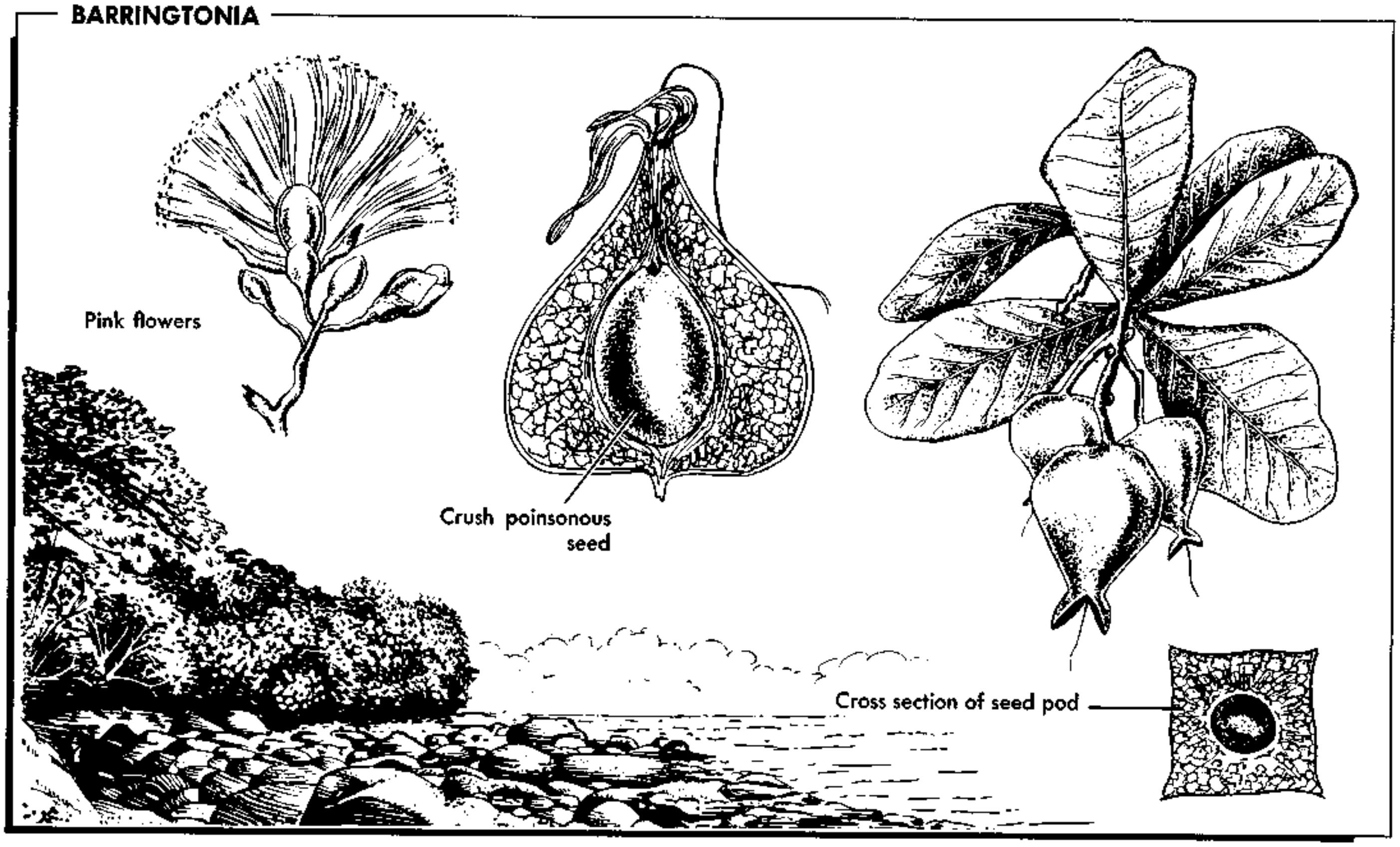

Figure 18-25. Barringtonia Plant for Poisoning Fish.

cy kit. Do not expose it unnecessarily to air or light; it retains its toxicity best if kept in a dark-colored vial. Lime thrown in a small pond or tidal pool will kill fish in the pool. Lime can be obtained by burning coral and seashells.

(c) The most common method of using fish-poison plants is to crush the plant parts (most often the roots) and mix them with water. Drop large quantities of the crushed plant into pools or the headwaters of small streams containing fish. Within a short time, the fish will rise in a helpless state to the surface. After putting in the poison, follow slowly down stream and pick up the fish as they come to the surface, sink to the bottom, or swim crazily to the bank. A stick dam or obstruction will aid in collecting fish as they float downstream. The husk of "green" black walnuts can be crushed and sprinkled into small sluggish streams and pools to act as a fish stupefying agent. In the southwest Pacific, the seeds and bark from the barringtonia tree (figure 18-25) are commonly used as a source of fish poison. The barringtonia tree usually grows along the seashore.

(7) Tickling can be effective in small streams with undercut banks or in shallow ponds left by receding flood waters. Place hands in the water and reach under the bank slowly, keeping the hands close to the bottom if possible. Move the fingers slightly until they make contact with a fish. Then work hands gently along its belly until reaching its gills. Grasp the fish firmly just behind the gills and scoop it onto land. In the tropics, this type of fishing can be dangerous due to hazardous marine life in the water such as piranhas, eels, and snakes.

18-5. Plant Food. The thought of having a diet consisting only of plant food is often distressing to stranded aircrew members. This is not the case if the survival episode is entered into with the confidence and intelligence based on knowledge or experience. If the survivors know what to look for, can identify it, and know how to prepare it properly for eating, there is no reason why they can't find sustenance. In many isolated regions, survivors who have had some previous training in plant identification can enjoy wild plant food.

a. Plants provide carbohydrates, which provide body energy and calories. Carbohydrates keep weight and energy up, and include important starches and sugars.

b. A documented and authoritative example of the value of a strictly plant diet in survival can be cited in the case of a Chinese botanist who had been drafted into the Japanese Army during World War II. Isolated with his company in a remote section of the Philippines, the Chinese botanist kept 60 of his fellow soldiers alive for 16 months by finding wild plants and preparing them properly. He selected six men to assist him, and then found 25 examples of edible plants in the vicinity of their camp. He acquainted the men with these samples, showing them what parts of the plants could be used for food. He then sent the men out to look for similar plants and had them separate the new plants according to the original examples to avoid any poisonous plant mingling with the edible ones. The result of this effort was impressive. Though all the men had a natural desire for ordinary food, none suffered physically from the plant food diet. The report was especially valuable because the botanist kept a careful record of all the food used, the results, and the comments of the men. This case history reflects the same opinions as those found in questionnaires directed to American survivors during World War II.

c. Another advantage of a plant diet is availability. In many instances, a situation may present itself in which procuring animal food is out of the question because of injury, being unarmed, being in enemy territory, exhaustion, or being in an area which lacks wildlife. If convinced that vegetation can be depended upon for daily food needs, the next question is "where to get what and how."

(1) Experts estimate there are about 300,000 classified plants growing on the surface of the Earth, including many which thrive on mountain tops and on the floors of the oceans. There are two considerations that survivors must keep in mind when procuring plant foods. The first consideration, of course, is the plant be edible, and preferably, palatable. Next, it must be fairly abundant in the areas in which it is found. If it includes an inedible or poisonous variety in its family, the edible plant must be distinguishable to the average eye from the poisonous one. Usually a plant is selected because one special part is edible, such as the stalk, the fruit, or the nut.

(2) To aid in determining plant edibility, there are general rules which should be observed and an edibility test that should be performed. In selecting plant foods, the following should be considered. Select plants resembling those cultivated by people. It is risky to rely upon a plant (or parts thereof) being edible for human consumption simply because animals have been seen eating it (for example, horses eat leaves from poison ivy; some rodents eat poisonous mushroom). Monkeys will put poisonous plants and fruits in pouches of their mouths and spit them out later. When selecting an unknown plant as a possible food source, apply the following general rules:

(a) Mushrooms and fungi should not be selected. Fungi have toxic peptides, a protein-base poison which has no taste. There is no field test other than eating to determine whether an unknown mushroom is edible. Anyone gathering wild mushrooms for eating must be absolutely certain of the identity of every specimen picked. Some species of wild mushrooms are difficult for an expert to identify. Because of the potential for poisoning, relying on mushrooms as a viable food source is not worth the risk.

(b) Plants with umbrella-shaped flowers are to be completely avoided, although carrots, celery, dill, and parsley are members of this family. One of the most poisonous plants, poison water hemlock, is also a member of this family (figure 18-26).

(c) All of the legume family should be avoided (beans and peas). They absorb minerals from the soil and cause problems. The most common mineral absorbed is selenium. Selenium is what has given locoweed its fame. (Locoweed is a vetch.)

(d) As a general rule, all bulbs should be avoided. Examples of poisonous bulbs are tulips and death camas.

(e) White and yellow berries are to be avoided as they are almost always poisonous. Approximately one-half of all red berries are poisonous. Blue or black berries are generally safe for consumption.

(f) Aggregated fruits and berries are always edible (for example, thimbleberry, raspberry, salmonberry, and blackberry).

(g) Single fruits on a stem are generally considered safe to eat.

(h) Plants with shiny leaves are considered to be poisonous and caution should be used.

(i) A milky sap indicates a poisonous plant.

(j) Plants that are irritants to the skin should not be eaten, such as poison ivy.

(k) A plant that grows in sufficient quantity within the local area should be selected to justify the edibility test and provide a lasting source of food if the plant proves edible.

(l) Plants growing in the water or moist soil are often the most palatable.

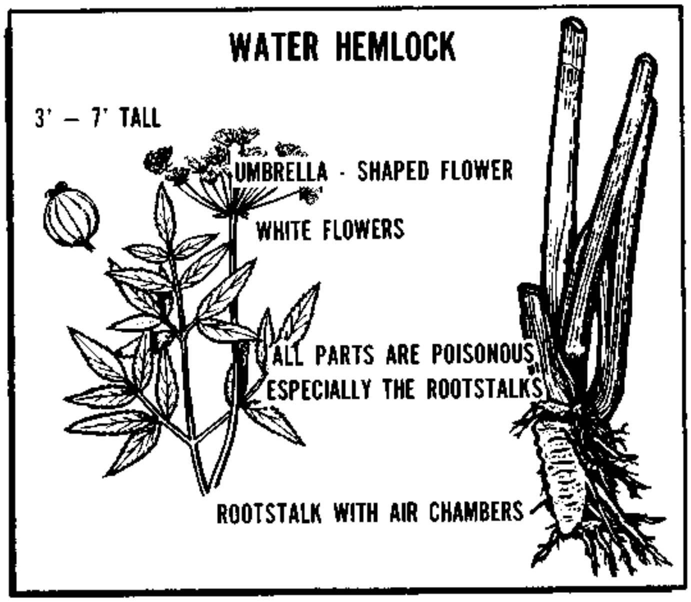

Figure 18-26. Water Hemlock.

(m) Plants are less bitter when growing in shaded areas.

(3) The previously mentioned information concerning plants is general. There are exceptions to every rule, but when selecting unknown plants for consumption, plants with these characteristics should be avoided. Plants that do not have these characteristics should be considered as possible food sources. Apply the edibility test to only one plant at a time so if some abnormality does occur, it will be obvious which plant caused the problem. Once a plant has been selected for the edibility test, proceed as follows:

(a) Crush or break part of the plant to determine the color of its sap. If the sap is clear, proceed to the next step.

(b) Touch the plant's sap or juice to the inner forearm or tip of the tongue. (A small taste of a poisonous plant will not do serious harm.) If there are no ill effects, such as a rash or burning sensation to the skin, bitterness to the taste, or numbing sensation of the tongue or lips, then proceed with the rest of the steps. (NOTE: Sometimes heavy smokers are unable to taste various poisons, such as alkaloids).

(c) Prepare the plant or plant part for consumption by boiling in two changes of water. The toxic properties of many plants are water soluble or destroyed by heat; cooking and discarding in two changes of water lessens the amount of poisonous material or removes it completely. Parboiling is a process of boiling the individual plant parts in repeated changes of water to remove bitter elements. This boiling period should last about 5 minutes.

(d) Place about 1 teaspoonful of the prepared plant food in the mouth for 5 minutes and chew but do not swallow it. A burning, nauseating, or bitter taste is a warning of possible danger. If any of these ill effects occur, remove the material from the mouth at once and discard that plant as a food source. However, if no burning sensation or other unpleasant effect occurs, swallow the plant material and wait 8 hours.

(e) If after this 8 hours there are no ill effects, such as nausea, cramps, or diarrhea, eat about 2 tablespoonfuls and wait an additional 8 hours.

(f) If no ill effects occur at the end of this 8-hour period, the plant may be considered edible.

(g) Keep in mind that any new or strange food should be eaten with restraint until the body system has become accustomed to it. The plant may be slightly toxic and harmful when large quantities are eaten.

(4) If cooking facilities are not available, survivors will not be able to boil the plant before consumption. In this case, plant food may be prepared as follows:

(a) Leach the plant by crushing the plant material and placing it in a container. Pour large quantities of cold water over it (rinse the plant parts). Leaching removes some of the bitter elements of nontoxic plants.

EDIBLE PARTS OF PLANTS	
Underground Parts	Tubers Roots and Rootstalks Bulbs
Stems and Leaves (potherbs)	Shoots and Stems Leaves Pith Bark
Flower Parts	Flowers Pollen
Fruits	Fleshy Fruits (dessert and vegetable) Seeds and Grains Nuts Seed Pods Pulps
Gums and Resins	
Saps	

Figure 18-27. Edible Parts of Plants.

(b) If leaching is not possible, survivors should follow the steps they can in the edibility test.

d. The survivor will find some plants which are completely edible, but many plants which they may find will have only one or more identifiable parts having food and thirst-quenching value. The variety of plant component parts which might contain substance of food value is shown in figure 18-27.

(1) Underground Parts:

(a) Tubers. The potato is an example of an edible tuber. Many other kinds of plants produce tubers such as the tropical yam, the Eskimo potato, and tropical water lilies. Tubers are usually found below the ground. Tubers are rich in starch and should be cooked by roasting in an earth oven or by boiling to break down the starch for ease in digestion. The following are some of the plants with edible tubers.

-1.Arrowroot, East Indian.
-2.Taro.
-3.Cassava (Tapioca).
-4.Bean, Yam.
-5.Chufa (Nut Grass).
-6.Water Lily (Tropical).
-7.Sweet Potato (Kamote).
-8.Yam Tropical.

(b) Roots and Rootstalks. Many plants produce roots which may be eaten. Edible roots are often several feet in length. In comparison, edible rootstalks are underground portions of the plant which have become thickened, and are relatively short and jointed. Both true roots and rootstalks are storage organs rich in stored starch. The following are some of the plants with edible roots or rootstalks (rhizomes):

-1.Baobab.
-2.Pine, Screw.
-3.Bean, Goa.
-4.Plantain, Water.
-5.Bracken.
-6.Reindeer Moss.
-7.Calla, Wild (Water Arum).
-8.Rock Tripe.
-9.Pollypody.
-10.Canna Lily.
-11.Rush, Flowering.
-12.Cattail.
-13.Spinach, Ceylon.
-14.Chicory.
-15.Ti Plant.
-16.Horseradish.
-17.Tree Fern.
-18.Lotus Lily.
-19.Water Lily (Temperate Zone).
-20.Manioc.

(c) Bulbs. The most common edible bulb is the wild onion, which can easily be detected by its characteristic odor. Wild onions may be eaten uncooked, but other kinds of bulbs are more palatable if cooked. In Turkey and Central Asia, the bulb of the wild tulip may be eaten. All bulbs contain a high percentage of starch. (Some bulbs are poisonous, such as the death camas which has white or yellow flowers.) The following are some of the plants with edible bulbs:

-1.Lily, Wild.
-2.Tulip, Wild.
-3.Onion, Wild.
-4.Blue Camas.
-5.Tiger Lily.

(2) Shoots and Leaves:

(a) Shoots (Stems). All edible shoots grow in much the same fashion as asparagus. The young shoots of ferns (fiddleheads) and especially those of bamboo and numerous kinds of palms are desirable for food. Some kinds of shoots may be eaten raw, but most are better if first boiled for 5 to 10 minutes, the water drained off, and the shoots reboiled until they are sufficiently cooked for eating (parboiled). (See figure 18-28).

-1.Agave (Century Plant).
-2.Palm, Coconut.
-3.Purslane.
-4.Reindeer Moss.
-5.Bamboo.
-6.Palm, Fishtail.
-7.Bean, Goa.
-8.Palm, Nipa.
-9.Bracken.
-10.Palm, Rattan.
-11.Rhubarb, Wild.
-12.Cattail.
-13.Palm, Sago.
-14.Spinach, Ceylon.
-15.Rock Tripe.
-16.Colocynth.
-17.Palm, Sugar.
-18.Papaya.
-19.Sugar Cane.
-20.Lotus Lily.
-21.Pokeweed (poisonous roots).
-22.Sweet Potato-Kamote.
-23.Luffa Sponge.
-24.Water Lily (Tropical).
-25.Polypody.
-26.Palm, Buri.
-27.Willow, Arctic.

(b) Leaves. The leaves of spinach-type plants (potherbs), such as wild mustard, wild lettuce, and lamb quarters, may be eaten either raw or cooked. Prolonged cooking, however, destroys most of the vitamins. Plants which produce edible leaves are perhaps the most numerous of all edible plants. The young tender leaves of nearly all nonpoisonous plants are edible. The follow-

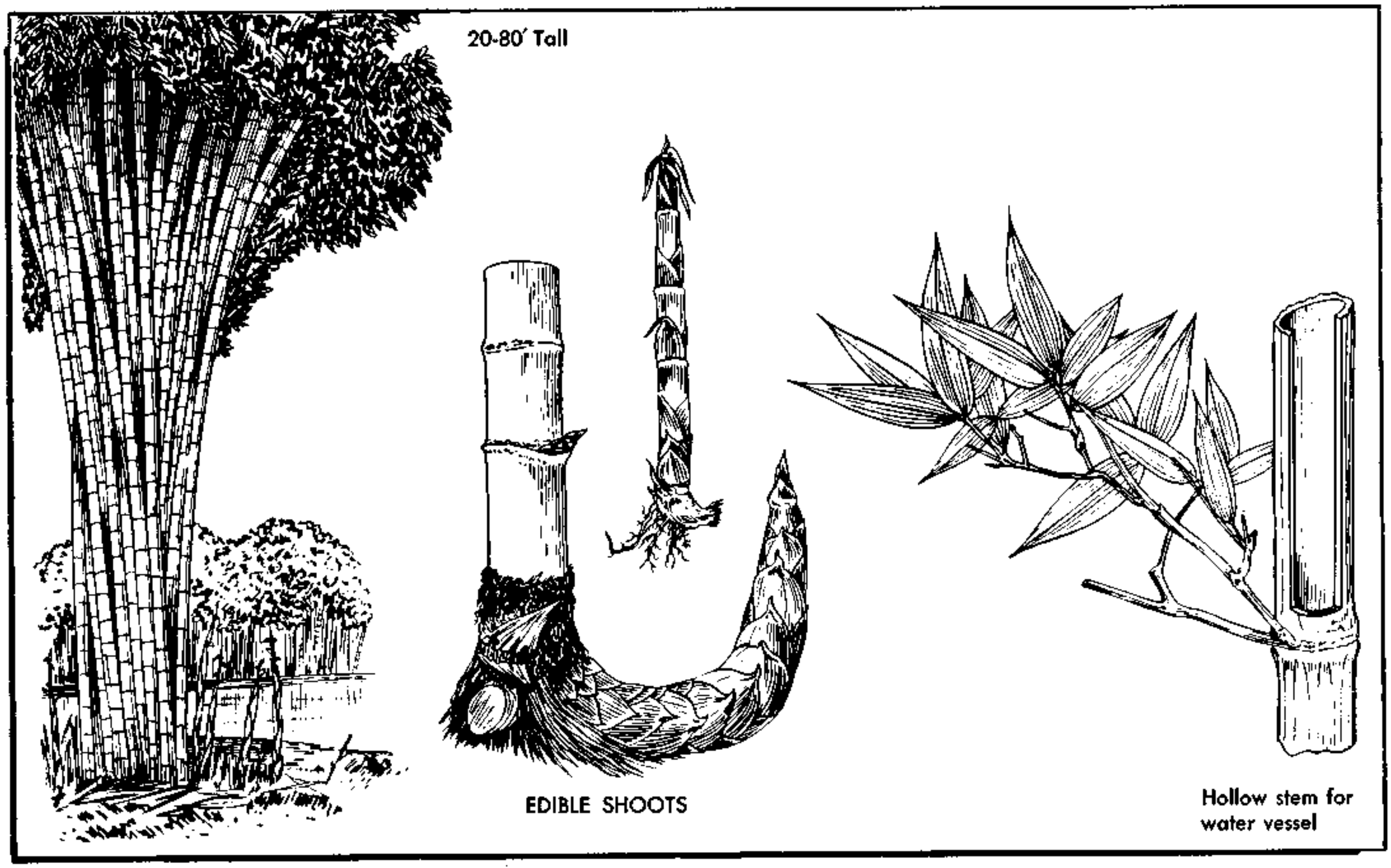
20-80' Tall
EDIBLE SHOOTS
Hollow stem for water vessel

Figure 18-28. Bamboo.

ing are only some of the plants with edible leaves:

-1.Amarath.
-2.Luffa Sponge.
-3.Rock Tripe.
-4.Avocado.
-5.Mango.
-6.Sorrel, Wild.
-7.Baobab.
-8.Orach, Sea.
-9.Bean, Goa.
-10.Papaya.
-11.Spinach, Ceylon.
-12.Cassava.
-13.Chickory.
-14.Pine, Screw.
-15.Spreading Wood Fern.
-16.Dock.
-17.Plantain.
-18.Pokeweed (poisonous roots).
-19.Sweet Potato-Kamote.
-20.Tamarind.
-21.Horseradish.
-22.Prickly Pear.
-23.Taro (only after cooking).
-24.Lettuce, Water.
-25.Purslane.
-26.Ti Plant.
-27.Willow, Arctic.
-28.Lotus Lily.
-29.Reindeer Moss.

(c) Pith. Some plants have an edible pith in the center of the stem. The pith of some kinds of tropical plants is quite large. Pith of the sago palm is particularly valuable because of its high food value. The following are some of the palms with edible pith (starch):

-1.Buri.
-2.Fishtail.
-3.Sago.
-4.Coconut.
-5.Rattan.
-6.Sugar.

(d) Bark. The inner bark of a tree—the layer next to the wood—may be eaten raw or cooked. It is possible in northern areas to make flour from the inner bark of such trees as the cottonwood, aspen, birch, willow, and pine. The outer bark should be avoided in all cases because this part contains large amounts of bitter tannin. Pine bark is high in vitamin C. The outer bark of pines can be cut away and the inner bark stripped from the trunk and eaten fresh, dried, or cooked, or it may be pulverized into flour. Bark is most palatable when newly formed in spring. As food, bark is most useful in the arctic regions, where plant food is often scarce.

(3) Flower Parts:

(a) Flowers and Buds. Fresh flowers may be eaten as part of a salad or to supplement a stew. The hibiscus flower is commonly eaten throughout the southwest Pacific area. In South America, the people of the Andes eat nasturtium flowers. In India, it is common to eat the flowers of many kinds of plants as part of a vegetable curry. Flowers of desert plants may also be eaten. The

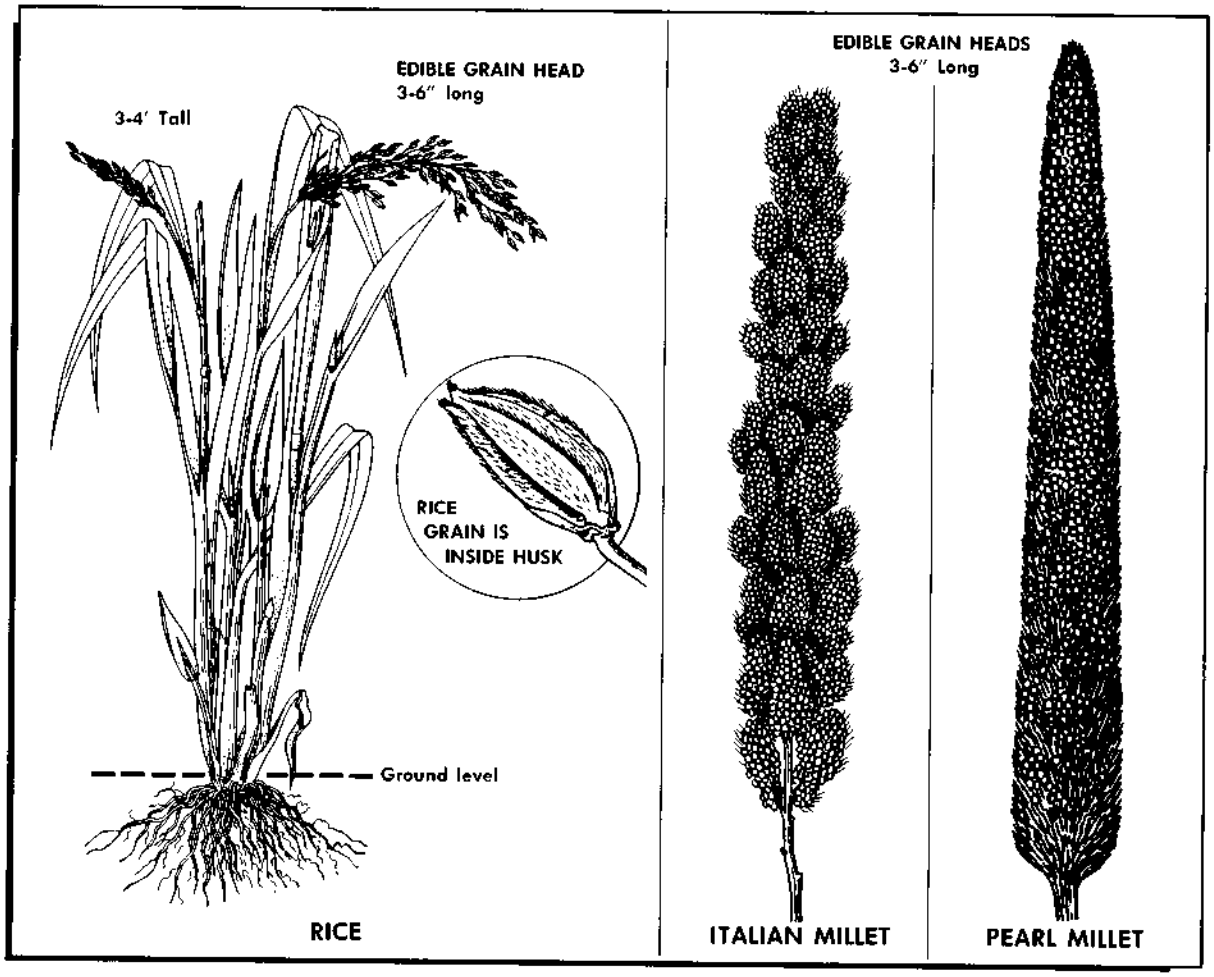

Figure 18-29. Grains.

following are plants with edible flowers:

-1.Abal.
-2.Colocynth.
-3.Papaya.
-4.Banana.
-5.Horseradish.
-6.Caper, Wild.
-7.Luffa Sponge.

(b) Pollen. Pollen looks like yellow dust. All pollen is high in food value and in some plants, especially the cattail. Quantities of pollen may easily be collected and eaten as a kind of gruel.

(4) Fruits. Edible fruits can be divided into sweet and nonsweet (vegetable) types. Both are the seed bearing parts of the plant. Sweet fruits are often plentiful in all areas of the world where plants grow. For instance, in the far north, there are blueberries and crowberries; in the temperate zones, cherries, plums, and apples; and in the American deserts, fleshy cactus fruits. Tropical areas have more kinds of edible fruit than other areas, and a list would be endless. Sweet fruits may be cooked, or for maximum vitamin content, left uncooked. Common vegetable fruits include the tomato, cucumber, and pepper.

(a) Fleshy Fruits (Sweet). The following are plants with edible fruits:

-1.Apple, Wild.
-2.Bael Fruit.
-3.Banana.
-4.Bignay.
-5.Blueberry, Wild.
-6.Bullocks Heart.
-7.Cloudberry.
-8.Crabapple.
-9.Cranberry.
-10.Fig, Wild.
-11.Grape, Wild.
-12.Huckleberry.
-13.Jackfruit.
-14.Jujube, Common.
-15.Mango.
-16.Mulberry.
-17.Papaya.
-18.Plum, Batako.
-19.Pokeberry.
-20.Prickly Pear.
-21.Rose Apple.
-22.Soursop.
-23.Sweetsop.

(b) Fleshy Fruits (Vegetables). The following are plants with edible fruits (vegetables):

-1.Breadfruit.
-2.Horseradish.
-3.Plantain.
-4.Caper, Wild.
-5.Luffa Sponge.

(c) Seeds and Grains. Seeds of many plants, such as buckwheat, ragweed, amaranth, and goosefoot, contain oils and are rich in protein. The grains of all cereals and many other grasses, including millet, are also extremely valuable sources of plant protein. They may either be ground between stones, mixed with water and cooked to make porridge, parched or roasted over hot stones. In this state, they are still wholesome and may be kept for long periods without further preparation (figure 18-29). The following are some of the plants with edible seeds and grains:

-1.Amaranth.
-2.Millet, Italian.
-3.Rice.
-4.Bamboo.
-5.Millet, Pearl.
-6.Palm, Nipa.
-7.Tamarind.
-8.Pine, Screw.
-9.Coloynth.
-10.Water Lily (Tropical).
-11.Sterculia.
-12.Baobab.
-13.Orach, Sea.
-14.St. John's Bread.
-15.Bean, Goa.
-16.Lotus Lily.
-17.Purslane.
-18.Water Lily (Temperate).
-19.Luffa Sponge.

(d) Nuts. Nuts are among the most nutritious of all raw plant foods and contain an abundance of valuable protein. Plants bearing edible nuts occur in all the climatic zones of the world and in all continents except in the arctic regions. Inhabitants of the temperate zones are familiar with walnuts, filberts, almonds, hickory nuts, acorns, hazelnuts, beechnuts, and pine nuts, to mention just a few. Tropical zones produce coconuts and other palm nuts, brazil nuts, cashew nuts, and macadamia nuts (figure 18-30). Most nuts can be eaten raw but some such as acorns, are better when cooked. The following are some of the plants with edible nuts:

-1.Almond.
-2.Chestnut, Water (Trapa Nut).
-3.Palm, Buri
-4.Almond, Indian or Tropical.
-5.Chestnut, Mountain.
-6.Palm, Coconut.
-7.Beechnut.
-8.Filbert (Hazelnut).
-9.Palm, Fishtail.
-10.Jackfruit Seeds.
-11.Oak, English (Acorn).
-12.Palm, Sago.
-13.Palm, Sugar.
-14.Pine.
-15.Pistachio, Wild.
-16.Walnut.

(e) Pulps. The pulp around the seeds of many fruits is the only part that can be eaten. Some fruits produce sweet pulp; others have a tasteless or even bitter pulp. Plants that produce edible pulp include the custard apple, inga pod, breadfruit, and tamarind. The pulp of breadfruit must be cooked, whereas in other plants, the pulp may be eaten uncooked. Use the edibility rules in all cases of doubt.

(5) Gums and Resins. Gum and resin are sap that collects and hardens on the outside surface of the plant. It is called gum if it is soft and soluble, and resin if it is hard and not soluble. Most people are familiar with the gum which exudes from cherry trees and the resin which seeps from the pine trees. These plant byproducts are edible and are a good source of nutritious food which should not be overlooked.

(6) Saps. Vines or other plant parts may be tapped as potential sources of usable liquid. The liquid is obtained by cutting the flower stalk and letting the fluid drain into some sort of container such as a bamboo section. Palm sap with its high-sugar content is highly nutritious. The following are some plants with edible sap and drinking water:

(a) Acacia, Sweet (water).
(b) Colocynth (water).
(c) Palm, Coconut (sap).
(d) Palm, Fishtail (sap).
(e) Agave (water).
(f) Cuipo Tree (water).
(g) Saxual (water).
(h) Palm, Nipa (sap).
(i) Palm, Rattan (water).
(j) Cactus (water).
(k) Grape (water).
(l) Banana (water).
(m) Palm, Sago (sap).
(n) Palm, Sugar (sap).
(o) Palm, Buri (sap).

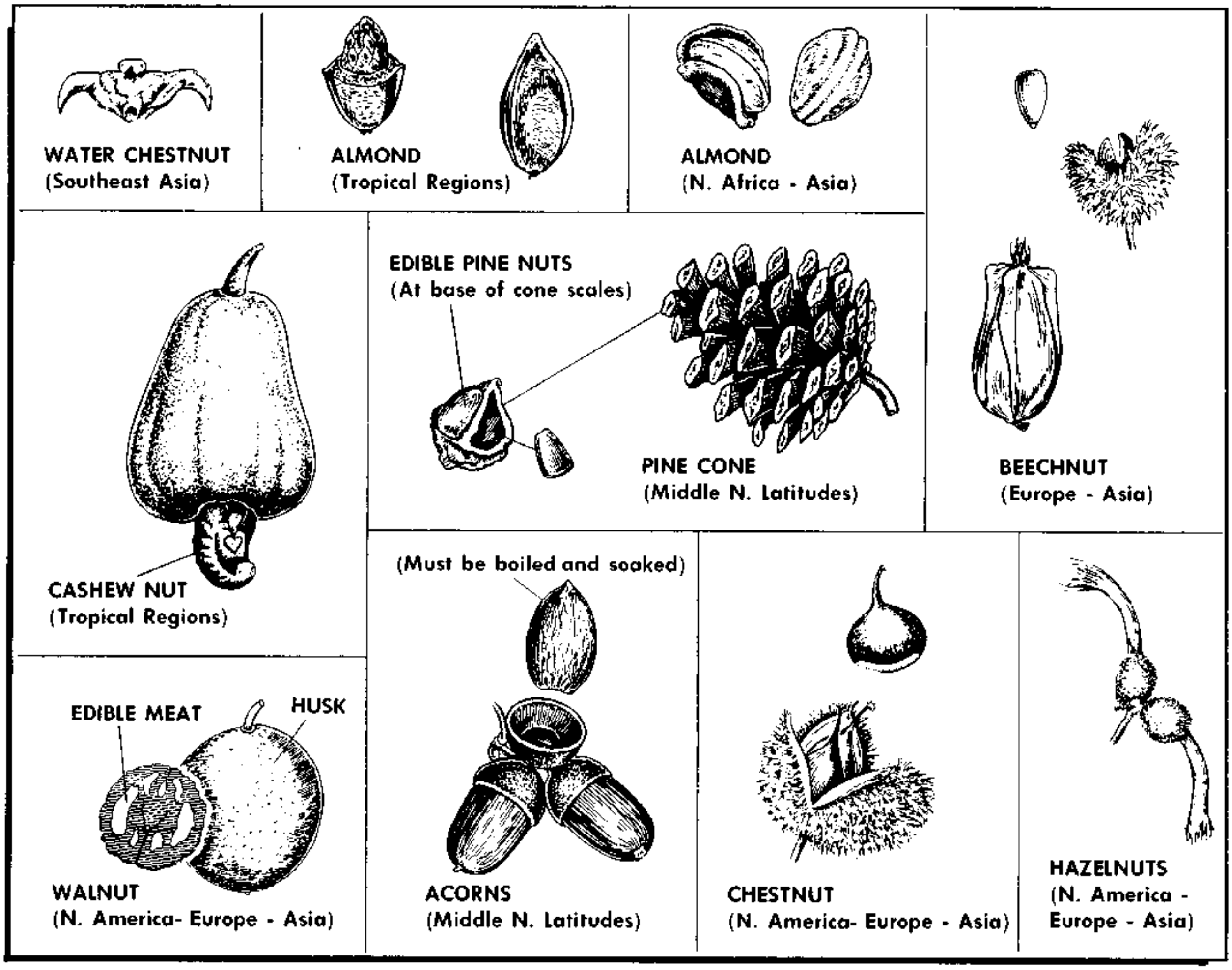
WATER CHESTNUT
(Southeast Asia)
ALMOND
(Tropical Regions)
ALMOND
(N. Africa - Asia)
EDIBLE PINE NUTS
(At base of cone scales)
PINE CONE
(Middle N. Latitudes)
BEECHNUT
(Europe - Asia)
CASHEW NUT
(Tropical Regions)
(Must be boiled and soaked)
EDIBLE MEAT
HUSK
WALNUT
(N. America- Europe - Asia)
ACORNS
(Middle N. Latitudes)
CHESTNUT
(N. America- Europe - Asia)
HAZELNUTS
(N. America - Europe - Asia)

Figure 18-30. Edible Nuts.

18-6. Food in Tropical Climate. There are more types of animals in the jungles of the world than in any other region. A jungle visitor who is unaware of the life style and eating habits of these animals would not observe the presence of a large number of the animals.

a. Game trails are the normal routes along which animals travel through a jungle. Some of the animals used as food are hedgehogs, porcupines, anteaters, mice, wild pigs, deer, wild cattle, bats, squirrels, rats, monkeys, snakes, and lizards.

(1) Reptiles are located in all jungles and should not be overlooked as a food source. All snakes should be considered poisonous and extreme caution used when killing the animal for a food source. All cobras should be avoided since the spitting cobra aims for the eyes; the venom can blind if not washed out immediately. Lizards are good food, but may be difficult to capture since they can be extremely fast. A good blow to the head of a reptile will usually kill it. Crocodiles and caimans are extremely dangerous on land as well as in the water.

(2) Frogs can be poisonous; all brilliantly colored frogs should be totally avoided. Some frogs and toads in the tropics secrete substances through the skin which has a pungent odor. These frogs are often poisonous.

(3) The larger, more dangerous animals such as tigers, rhinocerous, water buffalo, and elephants are rarely seen and should be left alone. These larger animals are usually located in the open grasslands.

b. Seafood such as fish, crabs, lobsters, crayfish, and small octopi can be poked out of holes, crevices, or rock pools (figure 18-31). Survivors should be ready to spear them before they move off into deep water. If they are in deeper water, they can be teased shoreward with a baited hook, or a stick.

(1) A small heap of empty oysters shells near a hole may indicate the presence of an octopus. A baited hook placed in the hole will often catch the octopus. The survivor should allow the octopus to surround the hook and line before lifting. Octopi are not scavengers like sharks, but they are hunters, fond of spiny lobster and other crab-like fish. At night, they come into shallow water and can be easily seen and speared.

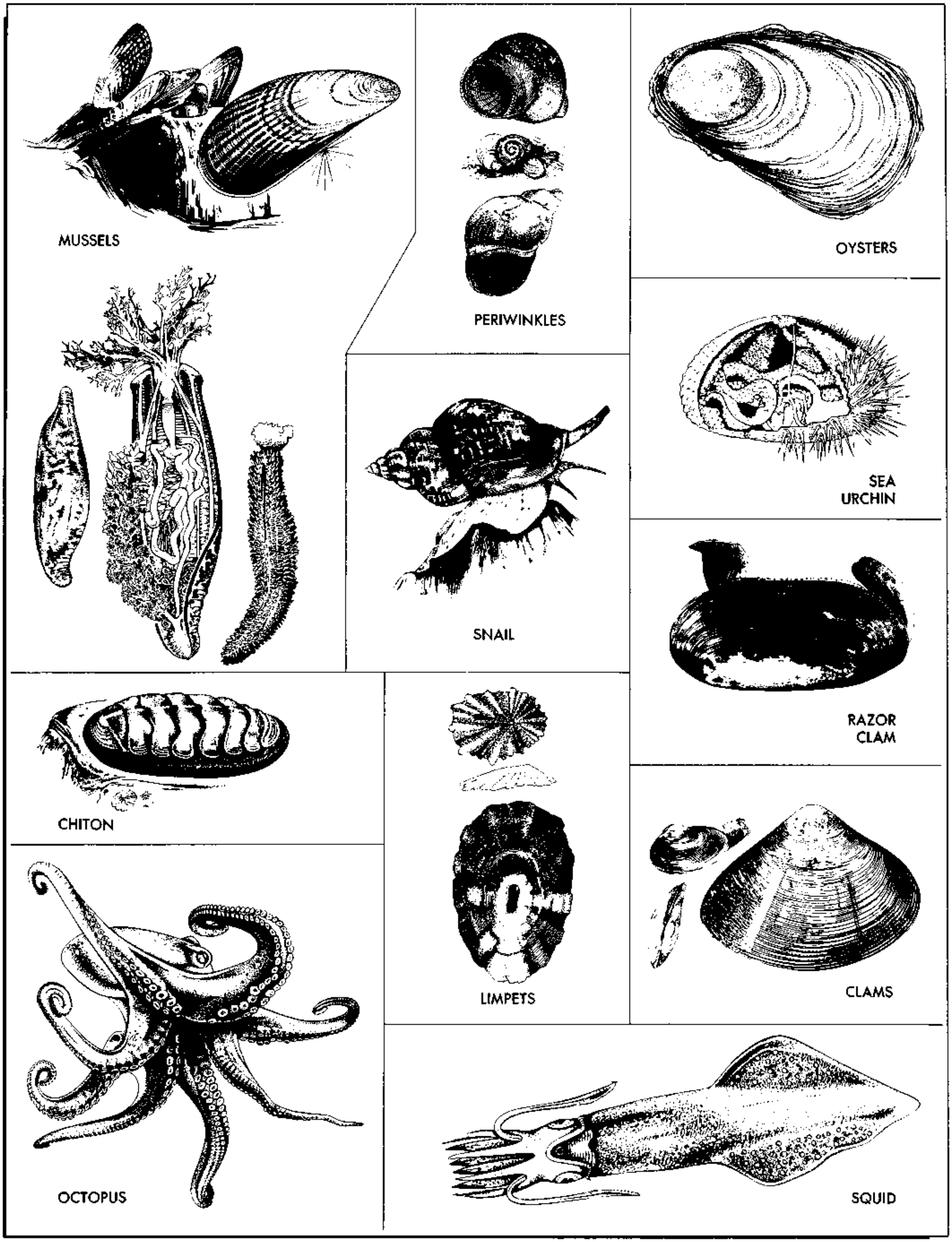
MUSSELS
PERIWINKLES
OYSTERS
SEA URCHIN
SNAIL
RAZOR CLAM
CHITON
LIMPETS
CLAMS
OCTOPUS
SQUID

Figure 18-31. Edible Invertebrates.

(2) Snails and limpets cling to rocks and seaweed from the low-water mark up. Large snails called chitons adhere tightly to rocks just above the surf line.

(3) Mussels usually form dense colonies in rock pools, on logs, or at the bases of boulders. *Mussels are poisonous in tropical zones during the summer,*especially when seas are highly phosphorescent or reddish.

(4) Sluggish sea cucumbers and conchs (large snails) live in deep water. The sea cucumber will shoot out its stomach when excited. The stomach is not edible. The skin and the five strips of muscle can be eaten after boiling. Conches can be boiled out of their shells and have very firm flesh. Use care when picking conches up. The bottom of their "foot" has a boney covering which can severely cut the survivor who procures it.

(5) The safest fish to eat are those from the open sea or deep water beyond the reef. Silvery fishes, river eels, butterfly fishes, and flounders from bays and rivers are good to eat.

(6) Land crabs are common on tropical islands and are often found in coconut groves. An open coconut can be used for bait.

(7) A number of methods can be used for procuring fish.

(a) Hook-and-Line Fishing. This type of fishing on a rocky coast requires a lot of care to keep the line from becoming entangled or cut on sharp edges. Most shallow-water fish are nibblers. Unless the bait is well placed and hooked and the barb of the hook offset by bending, the bait may be lost without catching a fish. Use hermit crabs, snails, or the tough muscle of a shellfish as bait. Take the cracked shells and any other animal remains and drop them into the area to be fished. This brings the fish to the area and provides a better procurement opportunity. Examine stomach contents of the first fish caught to determine what the fish are feeding on.

(b) Jigging. A baited or spooned hook dipped repeatedly beneath the surface of the water is sometimes effective. This method may be used at night.

(c) Spearing. This method is difficult except when the stream is small and the fish are large and numerous during the spawning season, or when the fish congregate in pools. Make a spear by sharpening a long piece of wood, lashing two long thorns on a stick, or fashioning a bone spear point, and take a position on a rock over a fish run. Wait patiently and quietly for a fish to swim by.

(d) Chop Fishing. Chop fishing is effective at night during low tide. This method requires a torch and a machete. The fish are attracted by the light of the torch, and then they may be stunned by slashing at them with the *back* of the machete blade. Care should be taken when swinging the machete (figure 18-32).

c. The jungle environment has a uniquely favorable condition for plant and animal life. The variety and richness of plant growth in these areas are paralleled nowhere else on the Earth. Because the rainfall is distributed throughout the year and there is a lack of cold seasons, plants in the humid regions can grow, produce leaves, and flower the year round. Some plants grow very rapidly. For example, the stem of the giant bamboo may grow more than 22 inches in a single day.

(1) A survivor in search of plant food should apply some basic principles to the search. A survivor is lucky to find a plant that can readily be identified as edible. If a plant resembles a known plant, it is very likely to be of the same family and can be used. If a plant cannot be identified, the edibility test should be applied. A survivor will find many edible plants in the tropical forest, but chances of finding them in abundance are better in an area that has been cultivated in the past (secondary growth).

(2) Some plants a survivor might find:

(a) Citrus fruit trees may be found in uncultivated areas, but are primarily limited to areas of secondary growth. The many varieties of citrus fruit trees and shrubs have leaves 2 to 4 inches long alternately arranged. The leaves are leathery, shiny, and evergreen. The leaf stem is often winged. Small (usually green) spines are often present by the side of the bud. The flowers are small and white to purple in color. The fruit has a leathery rind with numerous glands and is round and fleshy with several cells (fruit sections or slices) and many seeds. The great number of wild and cultivated fruits (oranges, limes, lemons, etc.) native to the tropics are eaten raw or used in beverages.

(b) Taro can be found in both secondary growth and in virgin areas. It is usually found in the damp, swampy areas in the wild, but certain varieties can be found in the forest. It can be identified by its large heart-shaped or arrowhead-shaped leaves growing at the top of a vertical stem. The stem and leaves are usually green and rise a foot or more from a tuber at the base of the stem. Taro leaf tips point down; poisonous elephant ear points up. All varieties of taro must be cooked to break down the irritating crystals in the plant.

(c) Wild pineapple can be found in the wild, and common pineapples may be found in secondary growth areas. The wild pineapple is a coarse plant with long clustered, sword-shaped leaves with sawtoothed edges. The leaves are spirally arranged in a rosette. Flowers are violet or reddish. The wild pineapple fruit will not be as fully developed in the wild state as when cultivated. The seeds from the flower of the plant are edible as well as the fruit. The ripe fruit may be eaten raw, but the green fruit must be cooked to avoid irritation. (The leaf fibers make excellent lashing material and ropes can be manufactured from it.)

(d) Yams may be found cultivated or wild. There are many varieties of yam, but the most common has a vine with square-shaped cross section and two rows of heart-shaped leaves growing on opposite sides of the vine. The vine can be followed to the ground to locate

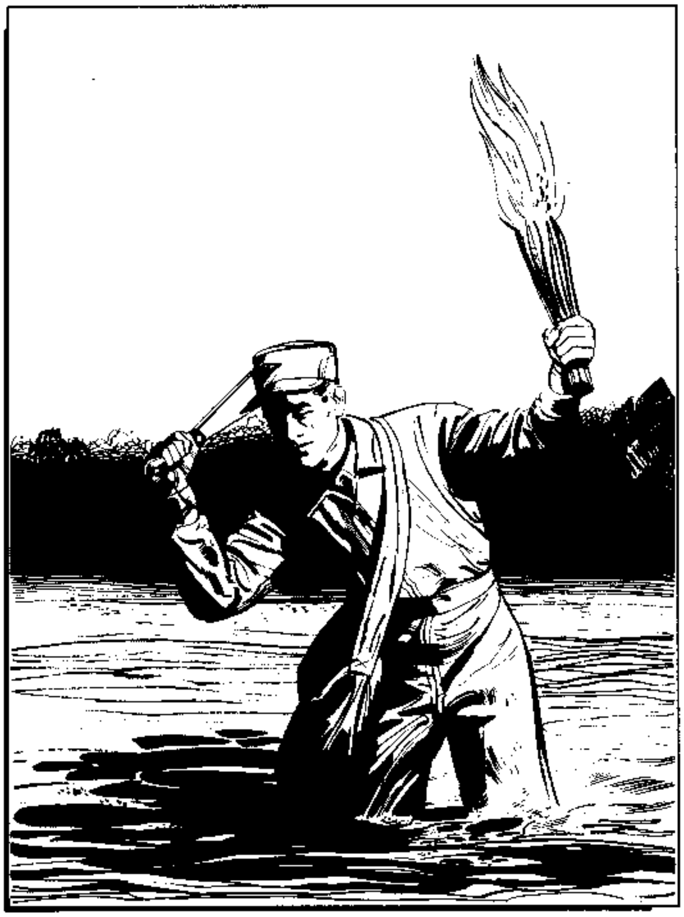

Figure 18-32. Chop Fishing.

the tuber. The tubers should be cooked to destroy the poisonous properties of the plant (figure 18-33).

(e) Ginger grows in the tropical forest and is a good source of flavoring for food. It is found in shaded areas of the primary forest. The ginger plant grows 5 to 6 feet high. It has seasonal white snapdragon-type flowers, some variations have red flowers. The leaves when crushed produce a very sweet odor and are used for seasoning or tea. The tea is used by primitive people to treat colds and fever.

(f) The coconut palm is found wild on the seacoast and in farmed areas inland. It is a tree 50 to 100 feet high, either straight or curved, marked with ringlike leaf scars. The base of the tree is swollen and surrounded by a mass of rootlets. The leaves are leathery and reach a length of 15 to 20 feet. (The leaves make excellent sheathing for shelter.) The fruit grows in clusters at the top of the tree. Each nut is covered with a fibered hard shell. The "heart" of the coconut palm is edible and is found at the top. (The new leaves grow out of the heart.) Cut the tree down and remove the leaves to gain access to the heart. The flower of the coconut tree is also edible and is best used as a cooked vegetable. The germinating nut is filled with a meat that can be eaten raw or cooked. There are many other varieties of palm found in the tropics which have edible hearts and fruits (figure 18-34).

(g) The papaya is an excellent source of food and can be found in secondary growth areas. The tree grows to a height of 6 to 20 feet. The large, dark green, many fingered, rough-edged leaves are clustered at the top of the plant. The fruit grows on the stem clustered under the leaves. The fruit is small in the wild state, but cultivated varieties may grow to 15 pounds. The peeled fruit can be eaten raw or cooked. The peeling should never be eaten. The green fruit is usually cooked. The milky sap of the green fruit is used as a meat tenderizer; care should be taken not to get it in the eyes. Always wash the hands after handling fresh green papayas. If some of the sap does get in the eyes, they should be washed immediately (figure 18-35).

(h) Cassava (tapioca) can be found in secondary growth areas. It can be identified by its stalk-like leaves which are deeply divided into numerous pointed sections or fingers. The woody (red) stem of the plant is slender and at points appears to be sectioned. When found growing wild in secondary growth areas, pull the trunks to find where a root grows. When one is found, a tuber can be dug. Tubers have been found growing around a portion of the stem that was covered with vegetation. The brown tuber of the plant is white inside and must be boiled or roasted. The tuber must also be peeled before boiling. (The green-stemmed species of cassava is poisonous and must be cooked in several changes of water before eating it.)

(i) Ferns can be found in the virgin tropical forest or in secondary growth areas. The new leaves (fiddle heads) at the top are the edible parts. They are covered with fuzzy hair which is easily removed by rubbing or washing. Some can be eaten raw, but as a rule, should be cooked as a vegetable (figure 18-36).

(j) Sweet sops can be found in the tropical forest. It is a small tree with simple, oblong leaves. The fruit is shaped like a blunt pine cone with thick grey-green or yellow, brittle spines. The fruit is easily split or broken when ripe, exposing numerous dark brown seeds imbedded in the cream colored, very sweet pulp.

(k) The star apple is common in the tropical forests. The tree grows up to a height of 60 feet and can be identified by the leaves which have shiny, silky, brown hairs on the bottom. The fruit looks like a small apple or plum with a smooth greenish or purple skin. The meat is greenish in color and milky in texture. When cut through the center, the brown, elongated seeds make a figure like a 6- or 10-pointed star. The fruit is sweet and eaten only when fresh. When cut, the rind will, like other parts of the tree, emit a white sticky juice or latex which is not poisonous (an exception to the milky sap rule).

(3) Of the 300,000 different kinds of wild plants in the world, a large number of them are found in the

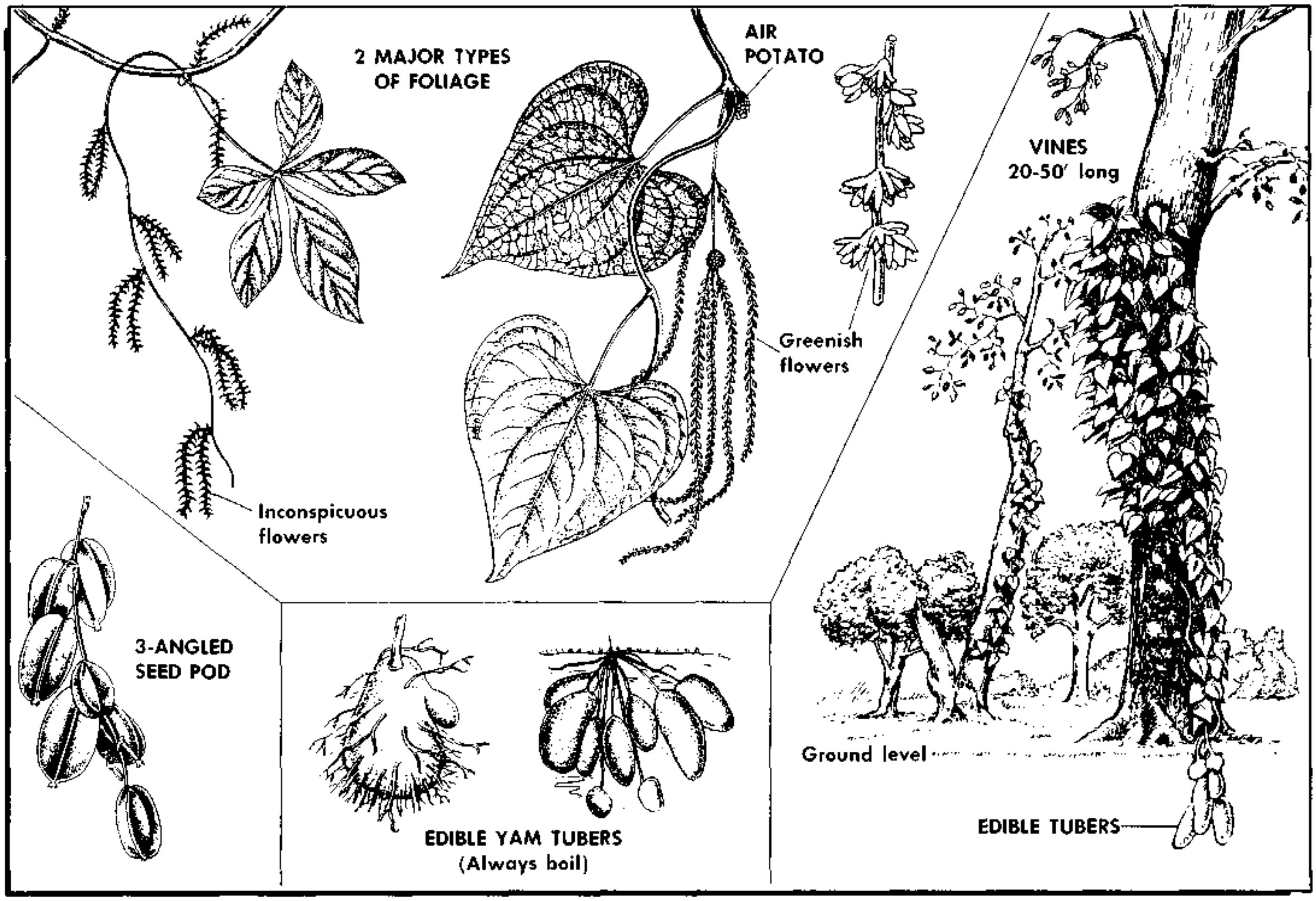

Figure 18-33. Yams.

tropics and many of them are potentially edible. Very few are deadly when eaten in small quantities. Those which are poisonous may be detected by using the edibility rules. Only a small number of jungle plants have been discussed. It would be of great benefit to anyone flying over or passing through a tropical environment to study the plant foods available in this type of environment.

18-7. Food in Dry Climates. Although not as readily available as in the tropical climate, food is available and obtainable.

a. Plant life in the desert is varied due to the different geographical areas. It must be remembered, therefore, that available plants will depend on the actual desert, the time of year, and if there has been any recent rainfall. The aircrew member should be familiar with plants in the area to be flown over.

(1) Date palms are located in most deserts and are cultivated by the native people around oases and irrigation ditches. They bear a nutritious, oblong, black fruit (when ripe).

(2) Fig trees are normally located in tropical and subtropical zones; however, a few species can be found in the deserts of Syria and Europe. Many kinds are cultivated. The fruit can be eaten when ripe. Most figs resemble a top or a small pear somewhat squashed in shape. Ripe figs vary greatly as to palatability. Many are hard, woody, covered with irritating hairs, and worthless as survival food. The edible varieties are soft, delectable, and almost hairless. They are green, red, or black when ripe.

(3) Millet, a grain bearing plant, is grown by natives around oases and other water sources in the Middle East deserts.

(4) The fruit of all cacti are edible. Some fruits are red, some yellow, but all are soft when ripe. Any of the flat leaf variety, such as the prickly pear, can be boiled and eaten as greens (like spinach) if the spines are first removed. During severe droughts, cattlemen burn off the spines and use the thick leaves for fodder. Although the cactus originates in the American deserts, the prickly pear has been introduced to the desert edges in Asia,

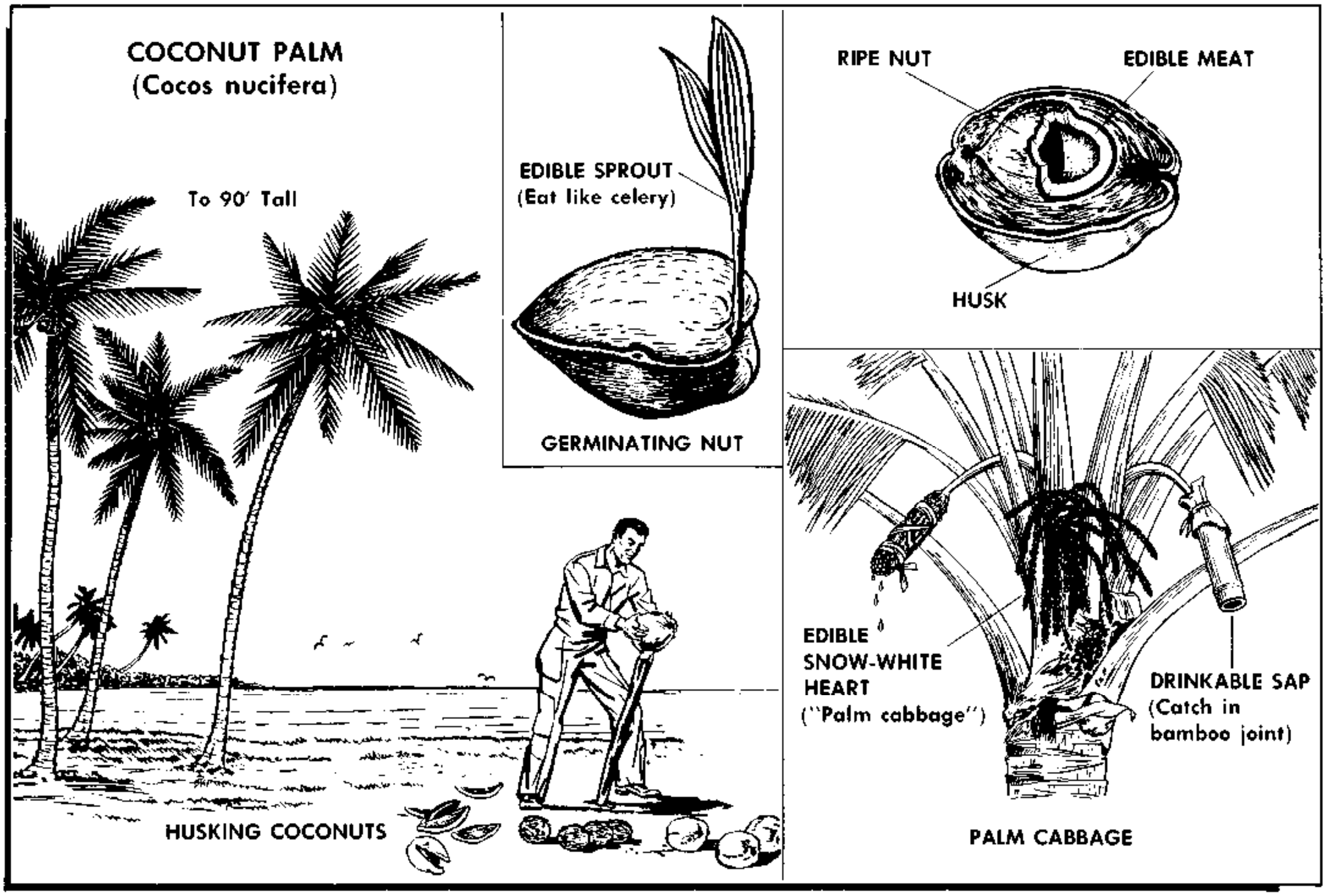

Figure 18-34. Coconut Palm.

Africa, the Near East, and Australia, where it grows profusely. Natives eat the fruit as fast as it ripens.

(5) There are two types of onions in the Gobi desert. A hot, strong, scallion-type grows in the late summer. It will improve the taste of food, but should not be used as a primary food. The highland onions grow 2 to 2.5 inches in diameter. These can be eaten like apples and the greens can also be eaten raw or cooked.

(6) All desert flowers can be eaten except those with milky or colored sap.

(7) All grasses are edible. Usually the best part is the whitish tender end that shows when the grass stalk is pulled from the ground. All grass seeds are edible.

b. Animal food sources may be used to supplement diets and provide needed protein and fats. When looking at a desert area, it is sometimes difficult to visualize an abundance of animal life existing in it. There is, however, a great quantity of animal life present. Most are edible, but some may be hazardous to a survivor during the procurement stage. Some of the abundant animal life includes:

(1) At the peak of seasonal plant growth, the desert crawls and buzzes with an enormous number and variety of beetles, ants, wasps, moths, and bugs. They appear with the first good rains and generally feed during nighttime. The Ute Indians of North America have harvested crickets, and peoples of the Middle East have roasted locusts. The human diet in Mexico and the American Indians of the Southwest frequently includes grasshoppers and caterpillars.

(2) On the Playas of the Sonora and Chichuahua deserts, several species of freshwater shrimp appear every summer in warm temporary ponds. In the Mohave Desert, where summer rains are rare, they may appear only a few times in a century.

(3) Snakes, lizards, tortoises, etc., have adapted well to the desert environment. Care must be observed when procuring them as some are hazardous, such as the Gila monster and rattlesnake. The desert tortoise, about a foot long when full grown, lives in some of the harshest regions of the Mohave and Sonora deserts. It is club footed, herbivorous, and can crawl about 20 feet per minute. The tortoise converts some of its food into water which is stored for the hot months in two sacs under the upper shell. A pint of water lasts the dry season. In spring and fall, the tortoise browses in broad

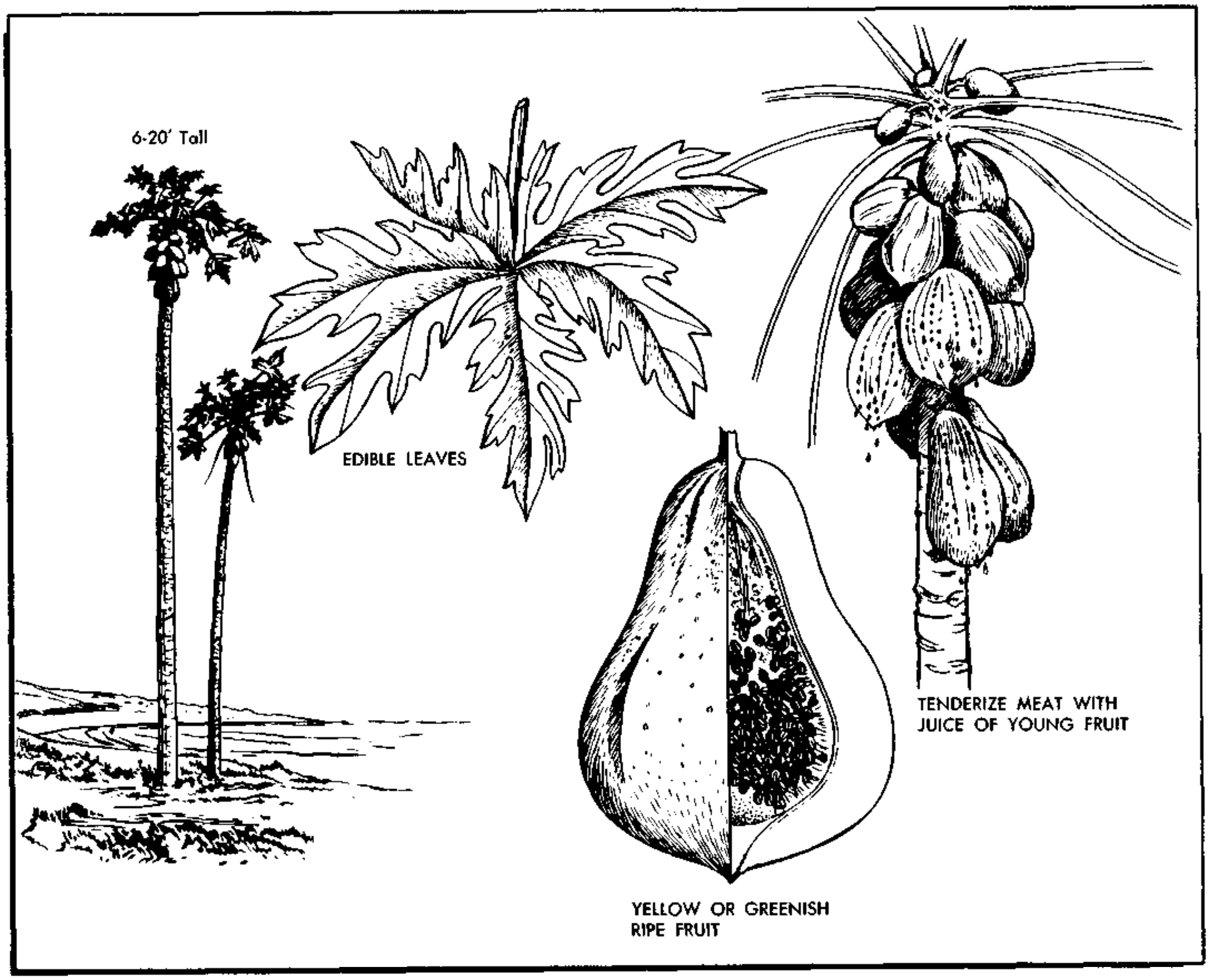
6-20' Tall
EDIBLE LEAVES
TENDERIZE MEAT WITH JUICE OF YOUNG FRUIT
YELLOW OR GREENISH RIPE FRUIT

Figure 18-35. Papaya.

daylight, becoming livelier as the day warms up. In the heat of the summer, it comes out of its shallow burrow in the early morning, the late evening, or not at all.

(4) In general, desert birds stay in areas of heavier vegetation and many need water daily; therefore, most will be found within short flights of some type of water source. Many birds will migrate during the drought season. If an abundance of birds is seen, insects, vegetation, and a water source will normally be nearby.

(5) Rabbits, prairie dogs, and rats have learned to live in deserts. They remain in the shade or burrow into the ground protecting themselves from the direct sun and heated air as well as from the hot desert surface.

(6) Larger mammals are also found in the desert. This group consists of gazelles, antelope, deer, foxes, small cats, badgers, dingos, hyenas, etc., and are amazingly abundant. Most are nocturnal and generally avoid humans. They roam at night eating smaller game and insects; a few eat plants; and a few can be hazardous to a survivor. Any of these mammals should be approached with caution.

(7) Only a few of the available animals and plants have been discussed. If the possibility of having to survive in a desert area exists, the aircrew member should try to become familiar with the food source available in that area.

18-8. Food in Snow and Ice Climates. In the snow and ice climates, food is more difficult to find than water. Animal life is normally more abundant during the warm months, but it can still be found in the cold months. Fish are available in most waters during the warmer months but they congregate in deep waters, large rivers, and lakes during the cold months. Some edible plant life can be found throughout the year in most areas of the arctic.

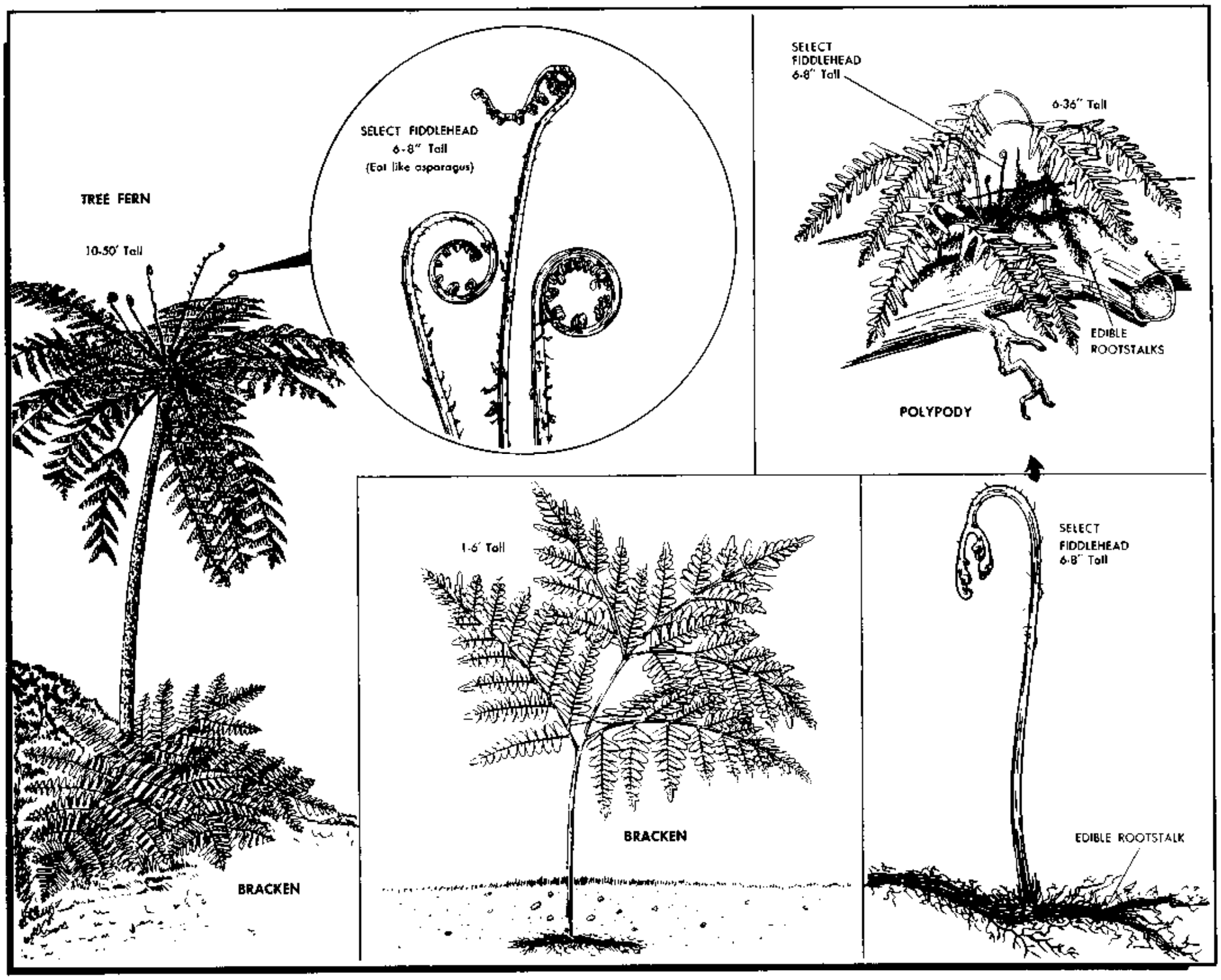
TREE FERN
10-50' Tall
SELECT FIDDLEHEAD
6-8" Tall
(Eat like asparagus)
BRACKEN
SELECT
FIDDLEHEAD
6-8" Tall
6-36" Tall
EDIBLE
ROOTSTALKS
POLYPODY
1-6' Tall
BRACKEN
SELECT
FIDDLEHEAD
6-8" Tall
EDIBLE ROOTSTALK

Figure 18-36. Edible Ferns.

a. All animals in the arctic regions are edible, but the livers of seals and polar bears must not be eaten because of the high concentration of vitamin A. Death could result from ingesting large quantities of the liver. On the open sea ice, game animals such as seal, walrus, polar bear, and fox are available. Many types of birds can be found during the warmer months. Fish can be caught throughout the year.

(1) Seal will probably be the main source of food when stranded on the open sea ice. They can be found in open leads, areas of thin ice, or where snow has drifted over a pressure ridge forming a cave which could have open water or very thin ice. These areas may also house polar bears which feed primarily on seals. Polar bears should be avoided.

(a) Newborn seals have trouble staying afloat or swimming and will be found on the ice in the early summer. The seal cubs can be easily killed with a club, spear, knife, or firearm and make an excellent source of food. The meat, blubber (fat), and coagulated milk in their stomachs are edible. When killing a cub, it is best to keep a lookout for the mother. She tends to protect offspring in any way possible.

(b) Seals must surface periodically to breathe. When the icepack is thin, the seals poke their noses through the ice and take a breath of air in a lead or in open water. In thick ice, the seal will chew and (or) claw a breathing hole through the ice. Normally most seals will have more than one breathing hole. In hunting seals, it is best to take a position beside a breathing hole and wait until a seal comes up to breathe, then spear or strike it on the head with a club. Seals are very sensitive to blows on or about the nose. They will often lose consciousness but not die. A hook can be suspended through the breathing hole so it hangs down at least 6 inches below the ice. When a seal comes to breathe, it can become hooked when it tries to depart the breathing hole. Seals can be recovered by gaffing or grabbing by hand, but in some cases, the breathing hole might have to be enlarged to pull the body through. If the seal is killed in open water, a "manak" or "grapple hook" can be used to retrieve it. All seals killed in open water or those that fall into open water should be recovered immediately. During the cold months, they will float for quite awhile, but during the warm months or when a female is nursing young, they sink rapidly. This is due to the loss of body fat (figure 18-37).

(2) Birds are plentiful during the summer months and can be procured by spearing, clubbing, catching with a baited fishhook, or use of a weapon.

(3) On tundra areas, there are large game, small game, and birds available as a food source.

(a) The large game consists of caribou, musk oxen, sheep, wolf, and bears (figure 18-38). Even though the large game animals can be a food source, they will be difficult to procure if a firearm is not available. Therefore, they should be considered a hazard to a survivor without a firearm. In the spring, bears tend to congregate along rivers and streams due to the amount of food available—normally salmon. During the fall, bears will be found feeding at berry patches. During certain seasons of the year, these areas should be avoided.

(b) Small game animals of the tundra include hares, lemmings, mice, ground squirrels, marmots, and foxes. They may be trapped or killed the entire year. When snaring, it is best to use a simple loop made of strong line or wire. The wire must be a two-strand twisted wire since metal becomes brittle in the cold and breaks very easily. Other snares and triggers will be less effective in the cold climate. A gill net can be used as a snare by spreading it across a trail so that the animal will entangle itself.

(c) Surface water is generally plentiful due to the number of lakes, ponds, bogs, and marshes. Water fowl and birds are very abundant during the warm months and include ducks, terns, geese, gulls, owls, and ptarmigan. The eggs and young birds are an excellent food source and can be easily procured (figure 18-39).

(4) As in the tundra areas, the forested areas in the arctic and arctic-like areas abound in wildlife.

(a) The large game species include moose, deer, caribou, and bear.

Figure 18-37. Walruses.

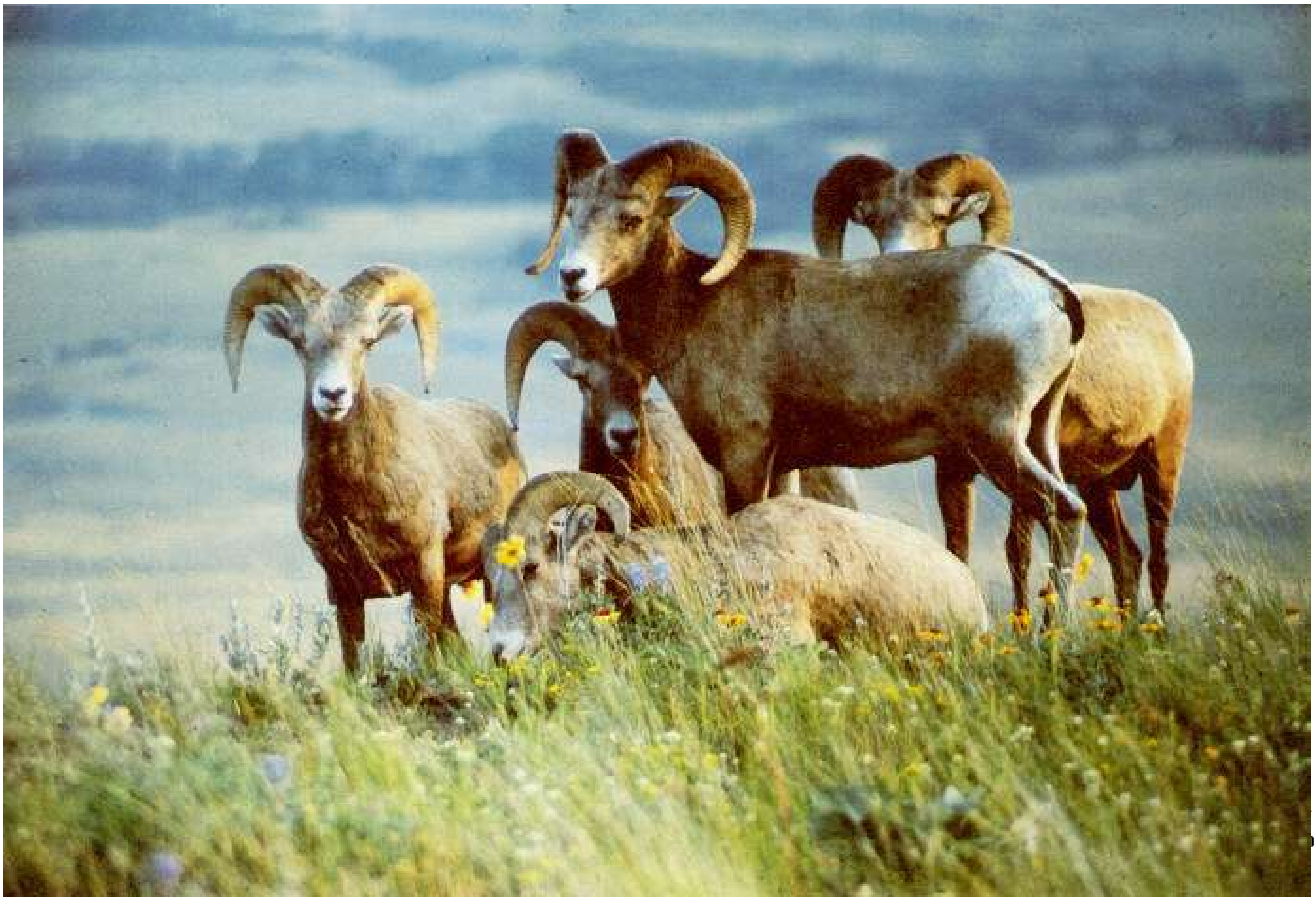

Figure 18-38. Big Horn Sheep.

(b) Small game of the forests includes hares, squirrels, porcupine, muskrat, and beaver. They can be snared or trapped easily in winter or summer. Small animal trails can be found in the winter with great ease. Most animals do not like to travel in deep snow so they tend to travel the same trail most of the time and this trail will look like a small superhighway — the snow packed down well below the normal snow level. Most trails will also be located in heavy cover or undergrowth or parallel to roads and open areas. The same trails will normally be used during the summer.

(5) During the summer months, the open water provides an excellent opportunity to procure all types of fish, both freshwater and saltwater, and freshwater mussels. The mussels can be handpicked off the bottom, while the fish can be netted, speared, clubbed, or caught with a hook and line. After freezeup, fishing is still possible through the ice. Shallow lakes, rivers, or ponds can freeze completely killing off all fish life. Fish tend to congregate in the deepest water possible. A hole should be cut through the ice at the estimated deepest point. Other good locations are at outlets or where tributaries flow into lakes or ponds. The ice is normally thinner over rapid moving water and at the edges of deep streams or rivers with snowdrifts extending out from the banks. Open water is often marked by a mist or fog formed over the area by vaporizing water. All methods of procuring fish in the summer will work in the winter (figure 18-40).

(6) The ocean shores are rich hunting grounds for edible sea life such as clams, mussels, scallops, snails, limpets, sea urchins, chitons, and sea cucumbers. They can be procured most of the year wherever there is open water. Tidal pools usually contain a great number of both fish and mollusks. The fish can be netted, speared, or hand caught. All sea life can be eaten raw, but cooking usually makes it more palatable (figure 18-41).

b. The plant life of the arctic regions is generally small and stunted due to the effects of permafrost, low mean temperatures, and a short growing season.

Figure 18-39. Swans.

(1) On the barren tundra areas, a wide variety of small edible plants and shrubs exist. During the short summer months on the tundra, Labrador tea, fireweed, coltsfoot, dwarf arctic birch, willow, and numerous other plants and berries can be found. During the winter, roots, rootstalks, and frozen berries can be found beneath the snow. Lichens and mosses are abundant but should be selected carefully as some species are poisonous.

(2) In bog or swamp areas, many types of water sedge, cattail, dwarf birch, and berries are available. During spring and summer, many young shoots from these plants are easily collected.

(3) The wooded areas of the arctic contain a variety of trees (birch, spruce, poplar, aspen, and others). Many berry-producing plants can be found, such as blueberries, cranberries, raspberries, cloud berries, and crow berries. Wild rose hips, Labrador tea, alder, and other shrubs are very abundant. Many wild edible plants are highly nutritious. Greens are particularly rich in carotene (vitamin A). Leafy greens, many berries, and rose hips are all rich in ascorbic acid (vitamin C). Many roots and rootstalks contain starch and can be used as a potato substitute in stews and soups.

(4) Although there are several types of edible mushrooms, fungi, and puff–balls in the arctic, a person should avoid ingesting them because it is difficult to identify the poisonous and nonpoisonous species. During the growing season, the physical characteristics can change considerably making positive identification even more difficult.

(5) There are many poisonous plants and a few poisonous berries in the arctic. Very few cause death; many will cause extreme nausea, dizziness, abdominal pain, and diarrhea. Contact poisonous plants, such as poison ivy, are not found in the arctic. The more common poisonous plants are shown in figures 18-42 through 18-49.

(6) When selecting edible plants, select young shoots when possible as these will be the most tender. Plants should be eaten raw to obtain the most nutritive value. Some of the more common edible plants are:

(a) Dandelions generally grow with grasses but may be scattered over rather barren areas. Both leaves

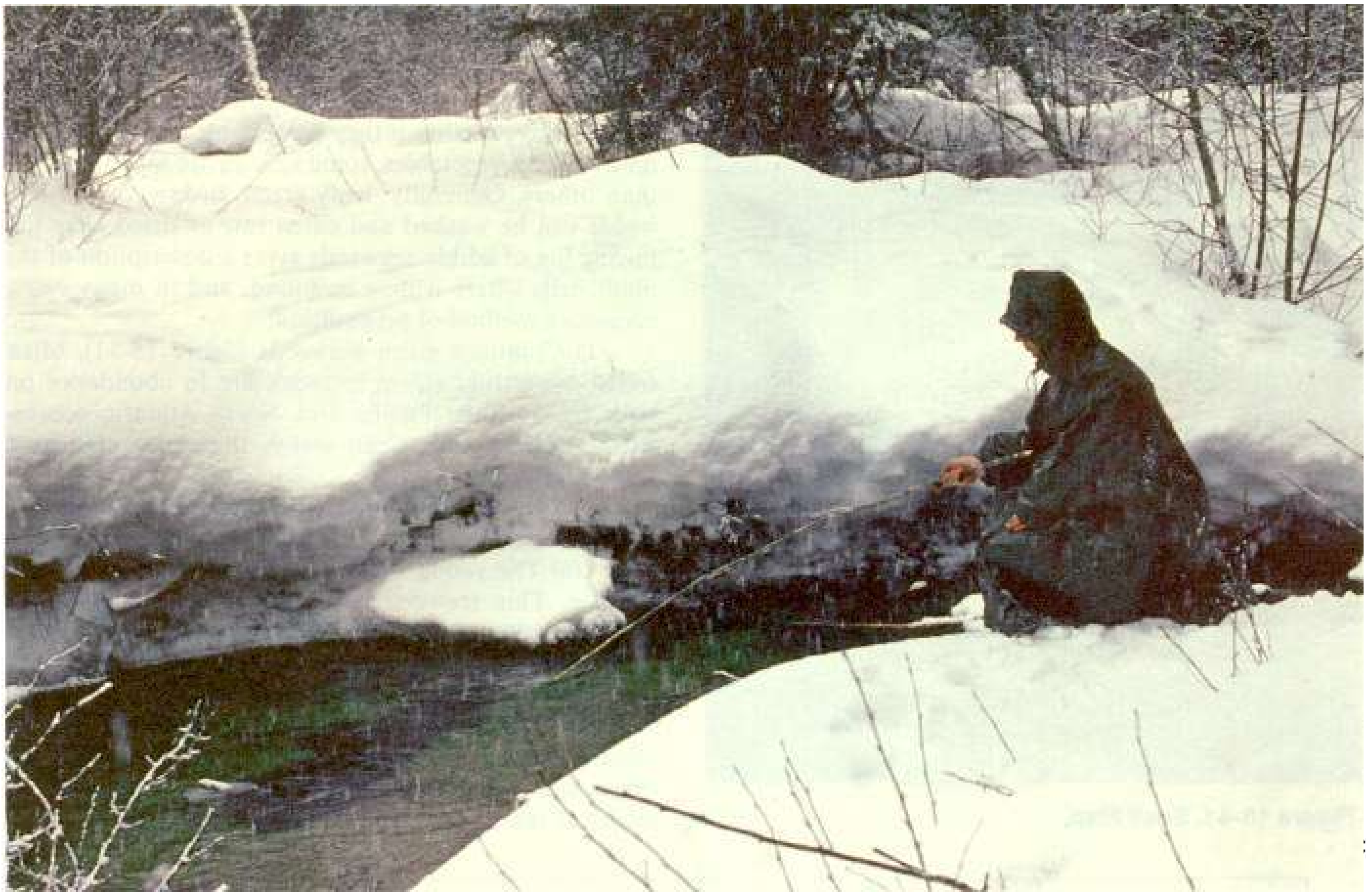

Figure 18-40. Winter Fishing.

and roots are edible raw or cooked. The young leaves make good greens; the roots (when roasted) are used as a substitute for coffee.

(b) Black and white spruce are generally the northern most evergreens. These trees have short, stiff needles that grow singularly rather than in clusters like pine needles. The cones are small and have thin scales. Although the buds, needles, and stems have a strong resinous flavor, they provide essential vitamin C by chewing them raw. In spring and early summer, the inner bark can be used for food.

(c) The dwarf arctic birch is a shrub with thin tooth-edged leaves and bark which peels off in sheets. The fresh green leaves and buds are rich in vitamin C. The inner bark may also be eaten.

(d) There are many different species of willow in the arctic. Young tender shoots may be eaten as greens and the bark of the roots is also edible. They have a decidedly sour taste but contain a large amount of vitamin C (figure 18-50).

(7) Lichens are abundant and widespread in the far North and can be used as a source of emergency food. Many species are edible and rich in starch-like substances, including Iceland moss, peat moss and reindeer lichen. Beard lichen growing on trees has been used as food by Indians. However, some of it contains a bitter acid which causes irritation of the digestive tract. If lichens are boiled, dried, and powdered, this acid is removed and the powder can then be used as flour or made into a thick soup.

18-9. Food on Open Seas. Almost all sea life is not only edible, but is also an excellent source of nutrients essential to humans. The protein is complete because it contains all the essential amino acids, and the fats are similar to those of vegetables. Seafoods are high in minerals and vitamins. The majority of life in the sea (fish, birds, plants, and aquatic animals) is edible.

a. Most seaweeds are edible and are a good source of food, especially for vitamins and minerals. Some seaweeds contain as much as 25 percent protein, while others are composed of over 50 percent carbohydrates. At least 75 different species are used for food by seacoast residents around the world. For many people, es-

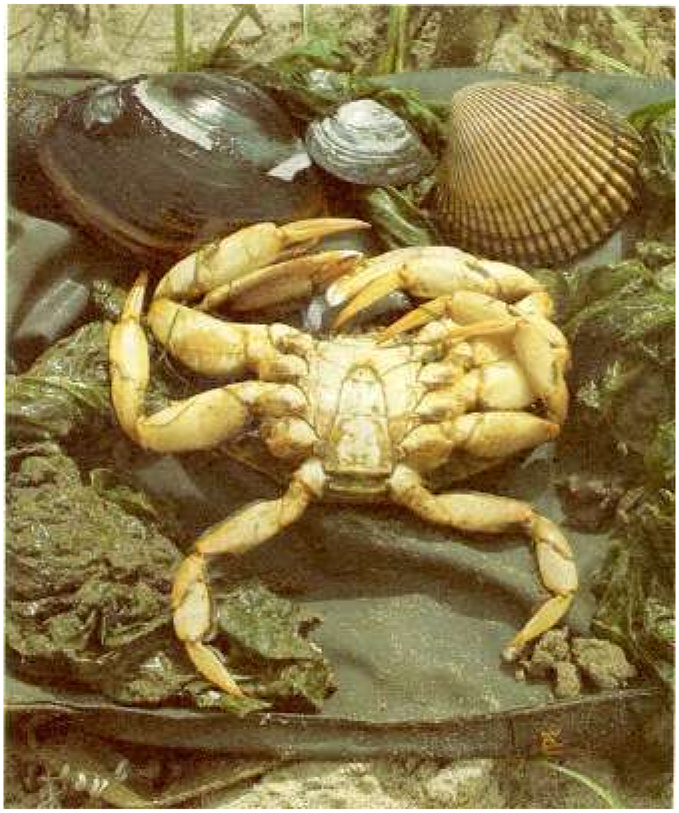

Figure 18-41. Shell Fish.

pecially the Japanese, seaweeds are an essential part of the diet, and the most popular varieties have been successfully farmed for hundreds of years. The high cellulose content may require gradual adaptation because of their laxative quality if they comprise a large part of the diet. As with vegetables, some species are more flavorful than others. Generally, leafy green, brown, or red seaweeds can be washed and eaten raw or dried. The following list of edible seaweeds gives a description of the plant, tells where it may be found, and in many cases, suggests a method of preparation:

(1) Common green seaweeds (figure 18-51), often called sea lettuce (Ulva lactuca), are in abundance on both sides of the Pacific and North Atlantic oceans. After washing it in clean water, it can be used as a garden lettuce.

(2) The most common edible brown seaweeds are the sugar wrack, kelp, and Irish moss (figure 18-52).

(a) The young stalks of the sugar wrack are sweet to taste. This seaweed is found on both sides of the Atlantic and on the coasts of China and Japan.

(b) Edible kelp has a short cylindrical stem and thin, wavy olive-green or brown fronds one to several feet in length. It is found in the Atlantic and Pacific oceans, usually below the high-tide line on submerged ledges and rocky bottoms. Kelp should be boiled before eating. It can be mixed with vegetables or soup.

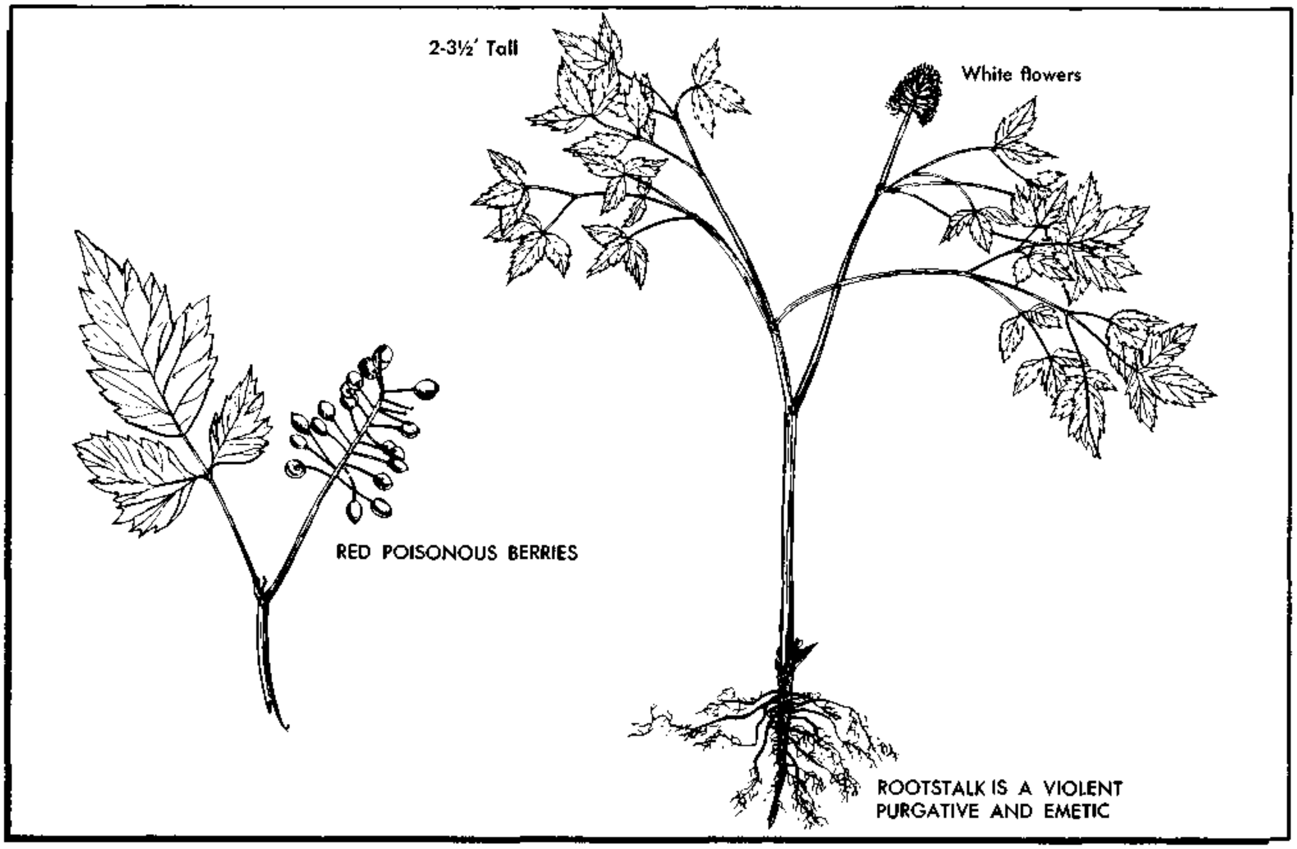

Figure 18-42. Baneberry.

(c) Irish moss, a variety of brown seaweed, is quite edible, and is often sold in market places. It is found on both sides of the Atlantic Ocean and can be identified by its tough, elastic, and leathery texture; however, when dried, it becomes crisp and shrunken. It should be boiled before eating. It can be found at or just below the high-tide line. It is sometimes found cast upon the shore.

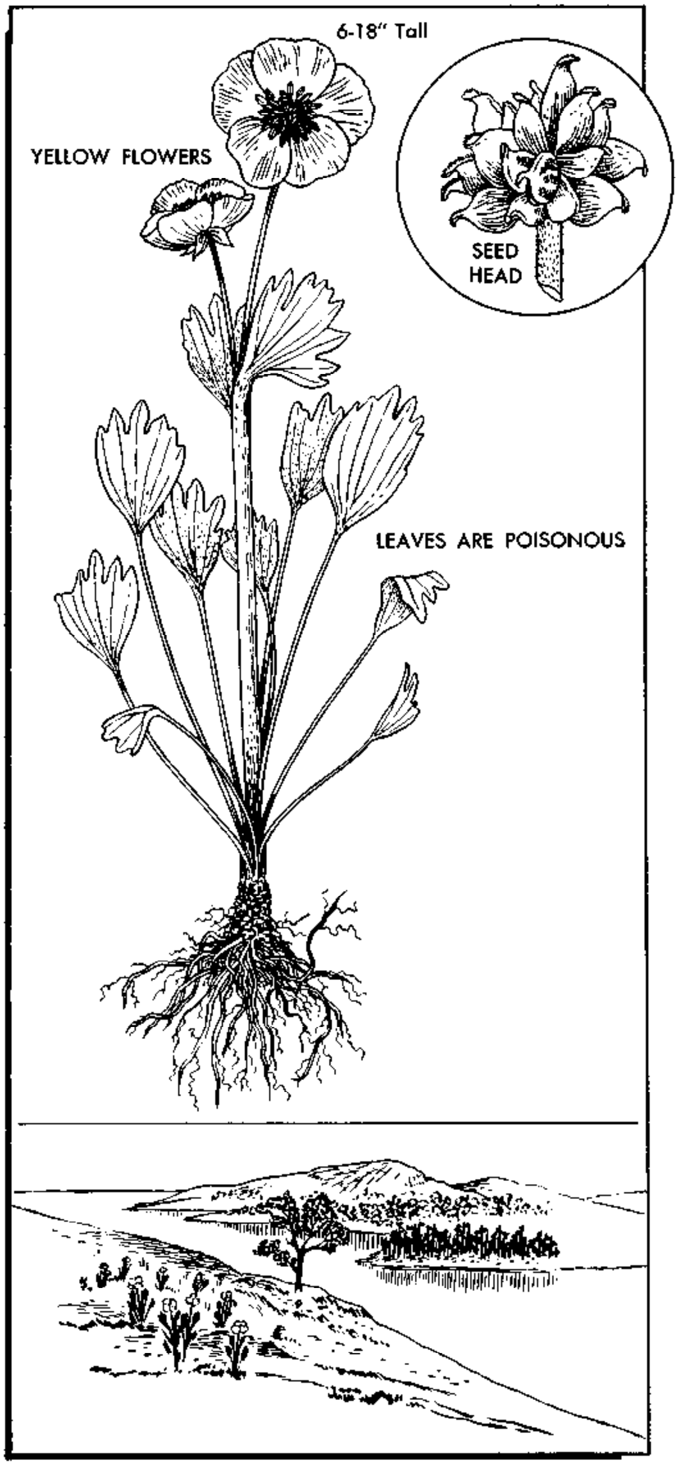
6-18" Tall
YELLOW FLOWERS
SEED HEAD
LEAVES ARE POISONOUS

Figure 18-43. Buttercup.

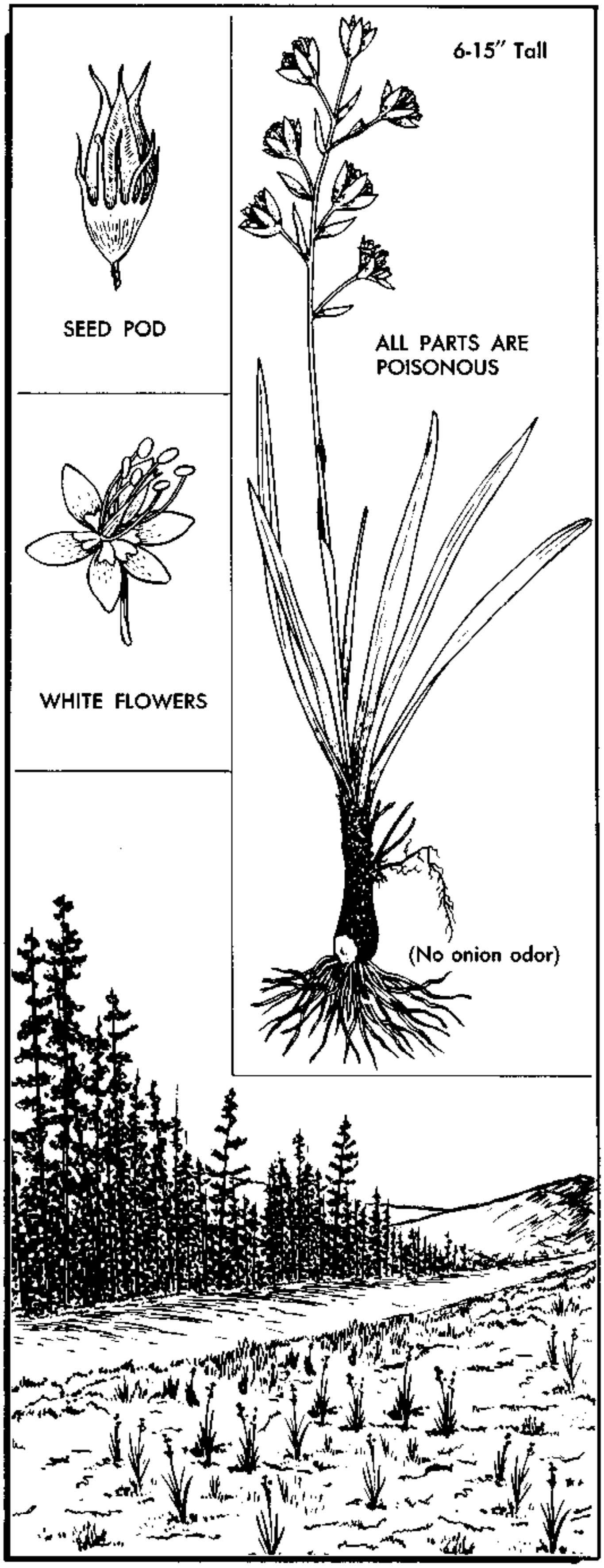
SEED POD
6-15" Tall
ALL PARTS ARE POISONOUS
WHITE FLOWERS
(No onion odor)

Figure 18-44. Death Camas.

(3) Red seaweeds can usually be identified by their characteristic reddish tint, especially the edible varieties. The most common and edible red seaweeds include the dulse, laver, and other warm-water varieties (figure 18-53).

(a) Dulse has a very short stem which quickly broadens into a thin, broad, fan-shaped expanse which is dark red and divided by several clefts into short, round-tipped lobes. The entire plant is from a few inches to a foot in length. It is found attached to rocks or coarser seaweeds, usually at the low-tide level, on both sides of the Atlantic Ocean and in the Mediterranean. Dulse is leathery in consistency and is sweet to the taste. If dried and rolled, it can be used as a substitute for tobacco.

(b) Laver is usually red, dark purple, or purplish-brown, and has a satiny sheen or filmy luster. Common to both the Atlantic and Pacific oceans, it has been used as food for centuries. This seaweed is used as a relish, or is cleaned and then boiled gently until tender. It can also be pulverized and added to crushed grains and fried in the form of flatcakes. During World War II, laver was chewed for its thirst-quenching value by New Zealand troops. Laver is usually found on the beach at the low-tide level.

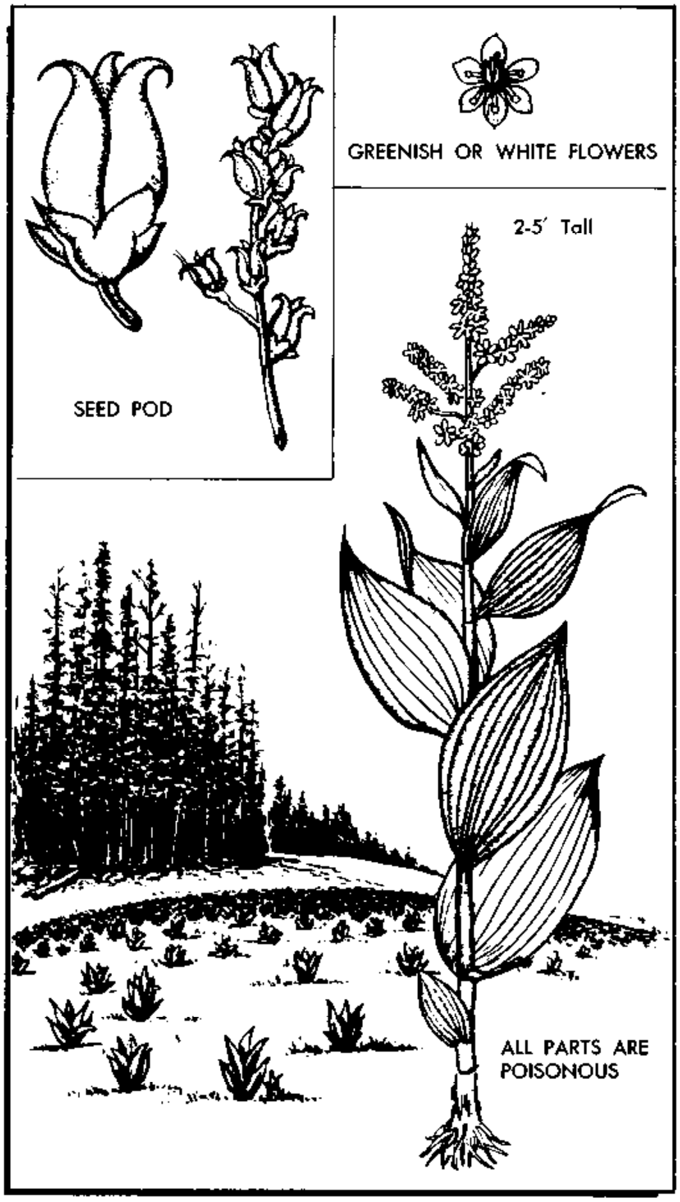

Figure 18-45. False Hellebore.

(c) A great variety of red, warm-water seaweed is found in the South Pacific area. This seaweed accounts for a large portion of the native diet. When found on the open sea, bits of floating seaweed may not only be edible but often contains tiny animals that can be used for food. The small fish and crabs can be dislodged by shaking the clump of seaweed over a container.

b. Plankton includes both minute plants and animals that drift about or swim weakly in the ocean. These basic organisms in the marine food chain are generally more common near land since their occurrence depends upon the nutrients dissolved in the water. Plankton can be caught by dragging a net through the water. The taste of the plankton will depend upon the types of organisms predominant in the area. If the population is mostly fish larvae, the plankton will taste like fish. If the population is mostly crab or shellfish larvae, the plankton will taste like crab or shellfish. Plankton contains valuable protein, carbohydrates and fats. Because of its high chiton and cellulose content, however, plankton cannot be immediately digested in large quantities. Therefore, anyone subsisting primarily on a plankton diet must gradually increase the quantities consumed. Most of the planktonic algae (phytoplankton) are smaller than the planktonic animals (zooplankton) and, although edible, are less palatable. Some plankton algae, for example, those dinoflagellates that cause "red tides" and paralytic shellfish poisoning, are toxic to humans.

(1) If a survivor is going to use plankton as a food source, there must be a sufficient supply of freshwater for drinking. Each plankton catch should be examined to remove all stinging tentacles broken from jellyfish or Portuguese man-of-war. The primarily gelatinous species may also be selectively discarded since their tissues are predominately composed of saltwater. When the plankton is found in subtropical waters during the summer months, and the presence of poisonous dinoflagellates is suspected (due to discoloration or high luminescence of the ocean), the edibility test should be applied before eating.

(2) The final precaution which a survivor may wish to take before ingesting plankton is to feel or touch the plankton to check for species that are especially spiney. The catch should be sorted (visually) or dried and crushed before eating if it contains large numbers of these spiney species.

c. If a fishing kit is available, the task of fishing will be made much easier. Small fish will usually gather under the shadow of the raft or in clumps of floating seaweed.

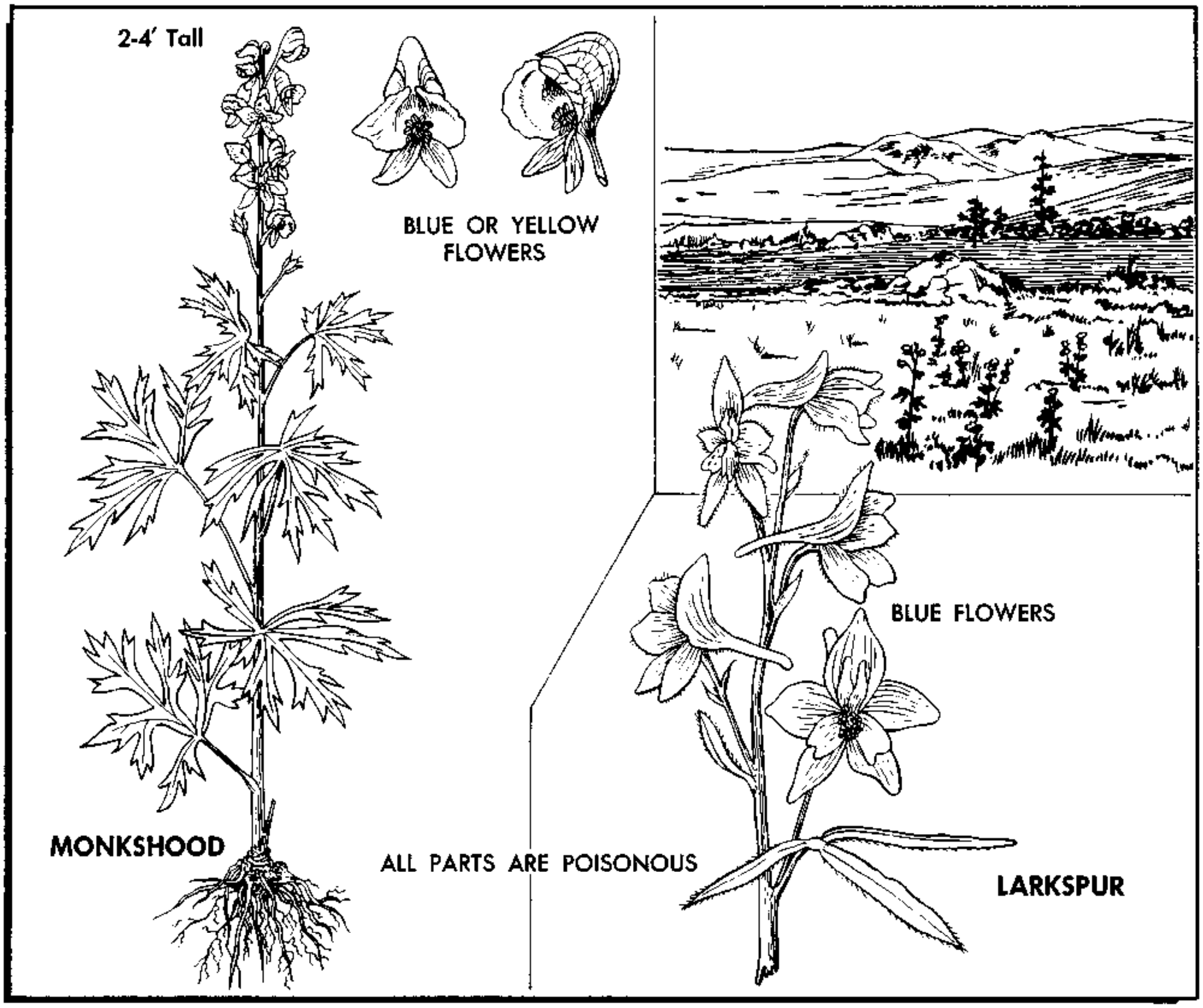

Figure 18-46. Monkshood and Larkspur.

These fish can be eaten or used as bait for larger fish. A net can also be used to procure most all sea life. Light attracts some types of fish. A flashlight or reflected moonlight can be used. It is not advisable to secure fishing lines to the body or the raft because a large fish may pull a person out of the raft or damage the raft. Fish, bait, or bright objects dangling in the water can attract large dangerous fish. All large fish should be killed outside the raft by a blow to the head or by cutting off the head.

d. Sea birds have proven to be a useful food source which may be more easily caught than fish. Survivors have reported capturing birds by using baited hooks, by grabbing, and by shooting. Freshly killed birds should be skinned, rather than plucked, to remove the oil glands. They can be eaten raw or cooked. The gullet contents can be a good food source. The flesh should be eaten or preserved immediately after cleaning. The viscera, along with any other unused parts, make good fish bait.

e. Marine mammals are rarely encountered by a person in the water, although they may be seen from a distance. Any large mammal is capable of inflicting injuries, but unless such mammals are pursued, they will generally avoid people. The killer whale (Orca) is rarely seen and, although large enough to feed on humans, has never been known to do so. Almost all sea mammals are a good source of food but difficult to obtain. The liver, especially that of any arctic or cold-water mammal, should not be eaten because of toxic concentrations of vitamin A.

f. All sea life must be cleaned, cut up, and eaten as soon as possible to avoid spoilage. Any meat left over can be preserved by sun-drying or smoking. The internal parts can be used as bait. If any doubt exists as to the edibility of a seafood, apply the "marine animal edibility test" found in chapter 11.

18-10. Preparing Animal Food. Survivors must know how to use the meat of game and fish to their advantage and how to do this with the least effort and physical exertion. Many people have died from starvation because they had failed to take full advantage of a game carcass. They abandoned the carcass on the mistaken theory that they could get more game when needed.

a. If the animal is large, the first impulse is usually to pack the meat to camp. In some cases, it might be easier to move the camp to the meat. A procedure often advocated for transporting the kill is to use the skin as a sled for dragging the meat. When the entire animal is dragged, this method may prove satisfactory only on frozen lakes or rivers or over very smooth snow-covered terrain. In rough or brush-covered country, however, it is generally more difficult to use this method, although it will work. Large mountain animals can sometimes be dragged down a snow-filled gully to the base of the mountain. If meat is the only consideration, and the survivors do not care about the condition of the skin, mountain game can sometimes be rolled for long distances. Before transporting a whole animal, it should be gutted and the incision closed. Once the bottom of the hill is reached, almost invariably the method is either to backpack the meat to camp, making several trips if no other survivors are present, or to pack the camp to the animal. Under survival conditions, home is on the back. When the weight of the meat proves excessive and moving the camp is not practical, some of the meat could be eaten at the scene. The heart, liver, and kidneys should be eaten as soon as possible to avoid spoilage.

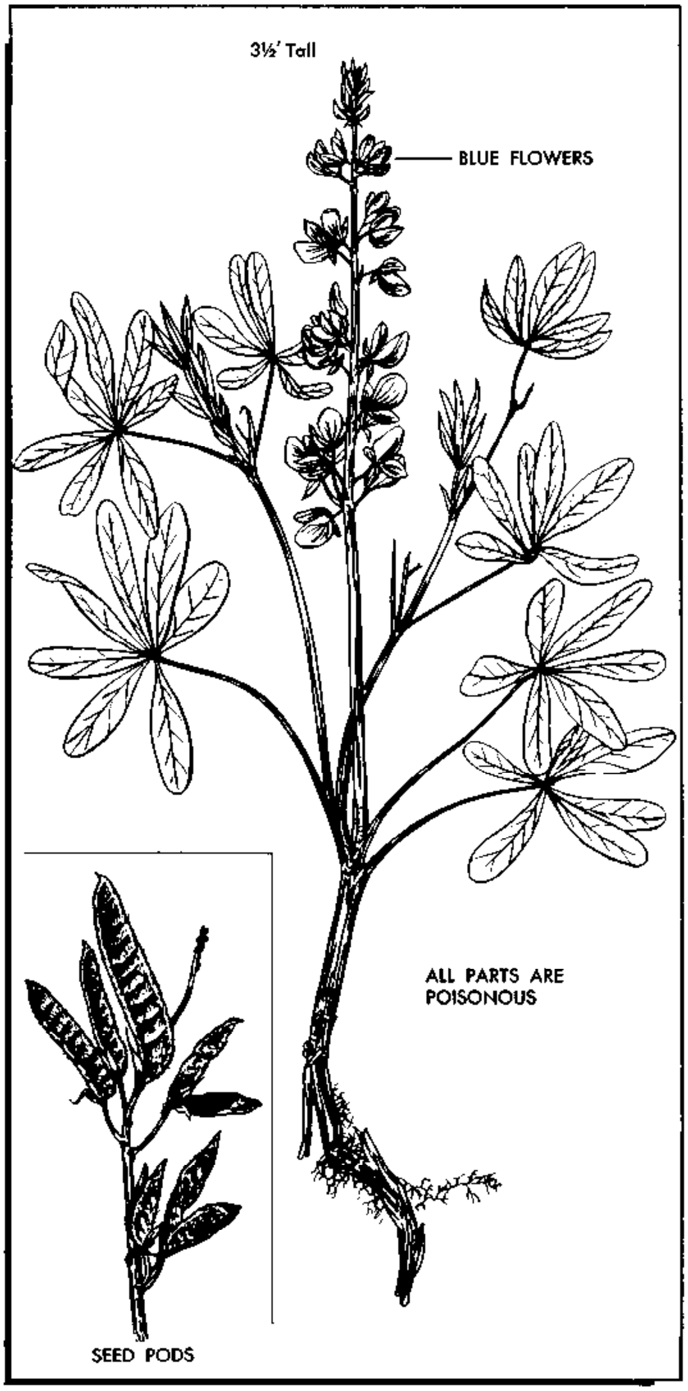

Figure 18-47. Lupine.

(1) Under survival conditions, skinning and butchering must be done carefully so that all edible meat can be saved. When the decision is made to discard the skin, a rough job can be done. However, considerations should be given to possible uses of the skin. A square of fresh skin, long enough to reach from the head to the knees, will not weigh much less when it is dried, and is an excellent ground cloth for use under a sleeping bag on frozen ground or snow. The best time to skin and butcher an animal is immediately after the kill. However, if an animal is killed late in the day, it can be gutted immediately and the other work done the next morning. An effort to keep the carcass secure from predators should be made.

(2) When preparing meat under survival conditions, all edible fat should be saved. This is especially important in cold climates, as the diet may consist almost entirely of lean meat. Fat must be eaten in order to provide a complete diet. Rabbits lack fat, and the fact that a person will die after an extended diet consisting only of rabbit meat indicates the importance of fat in a primitive diet. The same is true of birds, such as the ptarmigan.

(3) Birds should be handled in the same manner as other animals. They should be cleaned after killing and protected from flies. Birds, with the exception of sea birds, should be plucked and cooked with the skin on. Carrion-eating birds, such as vultures, must be boiled for at least 20 minutes to kill parasites before further cooking and eating. Fish-eating birds have a strong, fish-oil flavor. This may be lessened by baking them in mud or by skinning them before cooking.

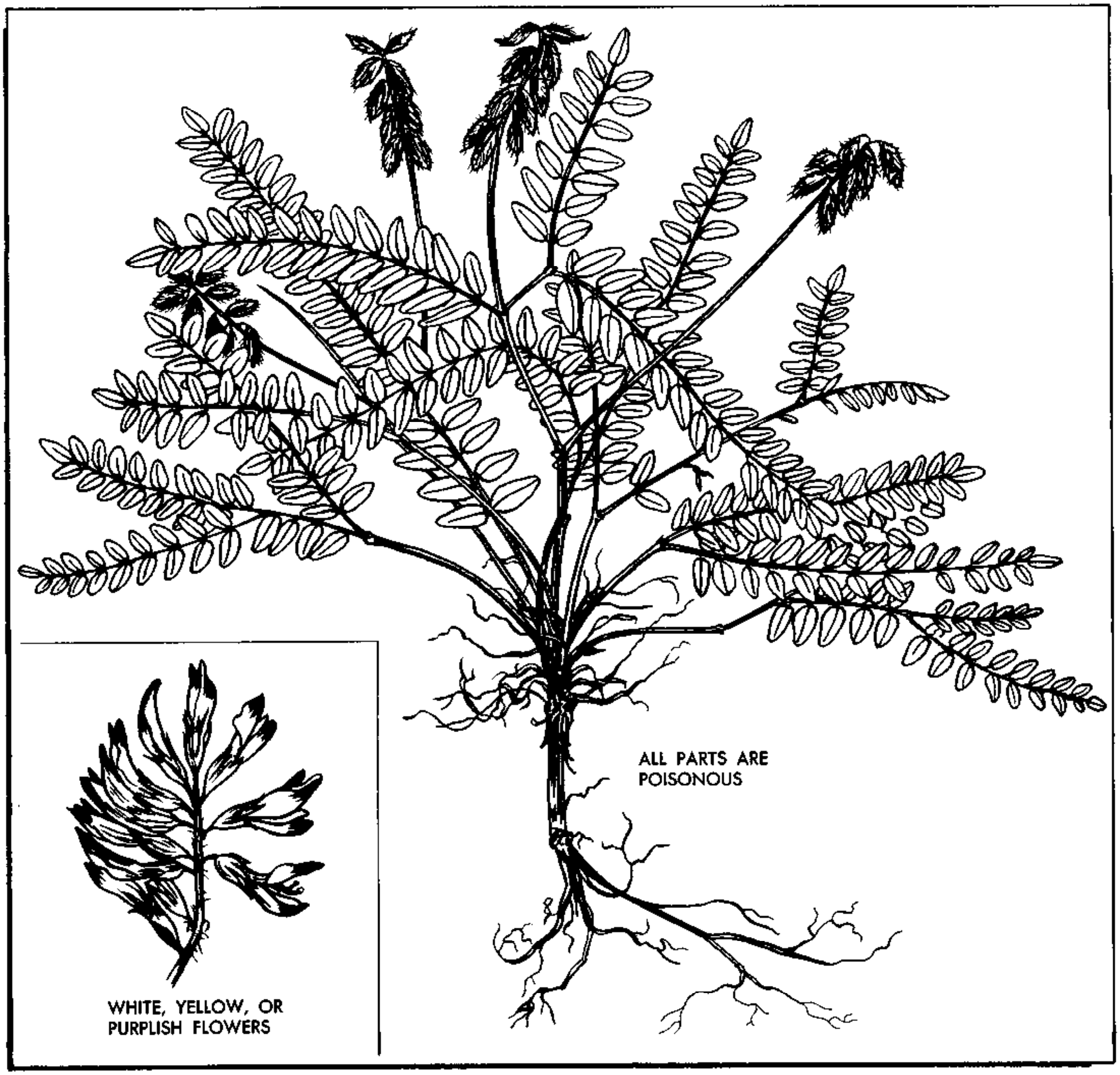
ALL PARTS ARE POISONOUS
WHITE, YELLOW, OR PURPLISH FLOWERS

Figure 18-48. Vetch and Locoweed.

b. There are two general ways to skin animals depending upon the size: the big game method, or the glove skinning method.

(1) Survivors should use the big game method when skinning and butchering large game.

(a) The first step in skinning is to turn the animal on its back and with a sharp knife, cut through the skin on a straight line from the tail bone to a point under its neck as illustrated in figure 18-54. In making this cut, pass around the anus and, with great care, press the skin open until the first two fingers can be inserted between the skin and the thin membrane enclosing the entrails. When the fingers can be forced forward, place the blade of the knife between the fingers, blade up, with knife held firmly. While forcing the fingers forward, palm upward, follow with the knife blade, cutting the skin but not cutting the membrane.

(b) If the animal is a male, cut the skin parallel to, but not touching the penis. If the tube leading from the bladder is accidentally cut, a messy job and unclean meat will result. If the gall or urine bladders are broken, washing will help clean the meat. Otherwise, it is best

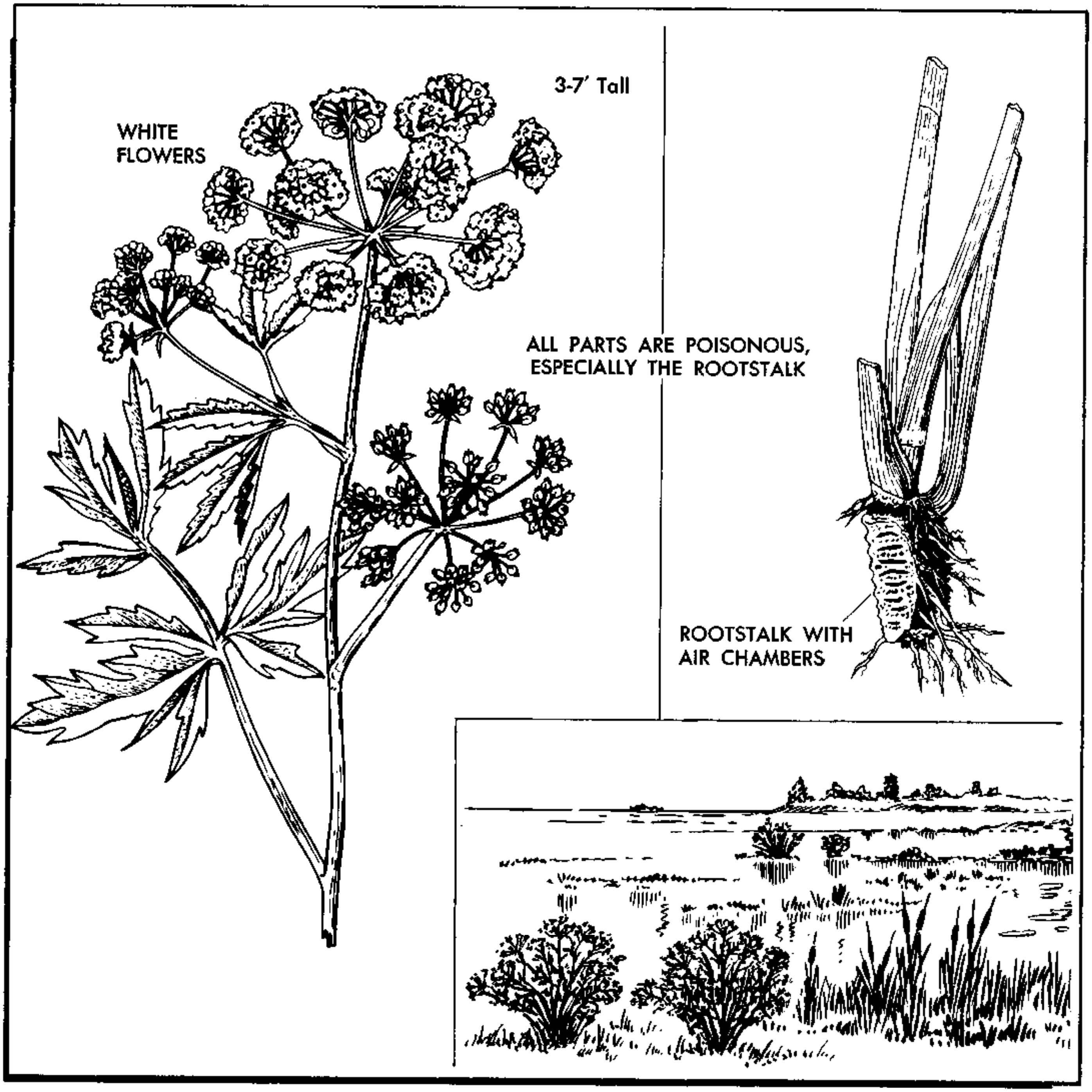
3-7′ Tall
WHITE FLOWERS
ALL PARTS ARE POISONOUS, ESPECIALLY THE ROOTSTALK
ROOTSTALK WITH AIR CHAMBERS

Figure 18-49. Water Hemlock.

not to wash the meat but to allow it to form a protective glaze.

(c) On reaching the ribs, it is no longer possible to force the fingers forward, because the skin adheres more strongly to flesh and bone. Furthermore, care is no longer necessary. The cut to point C can be quickly completed by alternately forcing the knife under the skin and lifting it. With the central cut completed, make side cuts consisting of incisions through the skin, running from the central cut (A-C) up the inside of each leg to the knee and hock joints. Then make cuts around the front legs just above the knees and around the hind legs above the hocks. Make the final cross cut at point C, and then cut completely around the neck and in back of the ears. Now is the time to begin skinning.

(d) On a small or medium-sized animal, one person can skin on each side. The easiest method is to begin at the corners where the cuts meet. When the animal is large, three people can skin at the same time. However, one should remember that when it is getting

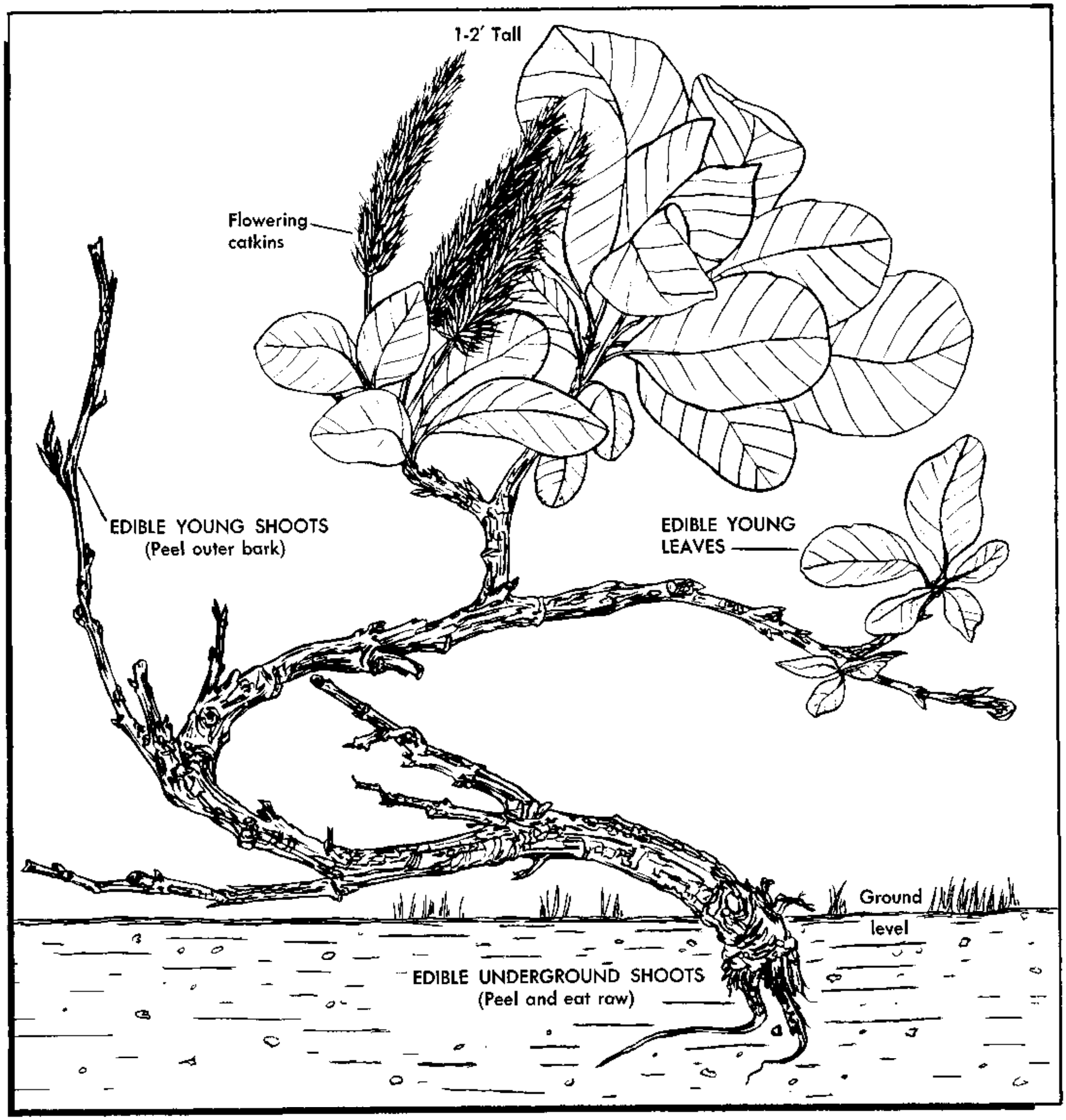
1-2′ Tall
Flowering catkins
EDIBLE YOUNG SHOOTS (Peel outer bark)
EDIBLE YOUNG LEAVES
Ground level
EDIBLE UNDERGROUND SHOOTS (Peel and eat raw)

Figure 18-50. Arctic Willow.

dark and hands are clumsy because of the cold, a sharp skinning knife can make a deep wound. After skinning down the animal's side as far as possible, roll the carcass on its side to skin the back. Then spread out the loose skin to prevent the meat from touching the ground and turn the animal on the skinned side. Follow the same procedure on the opposite side until the skin is free.

(e) In opening the membrane which encloses the entrails, follow the same procedure used in cutting the skin by using the fingers of one hand as a guard for the knife and separating the intestines from membrane. This thin membrane along the ribs and sides can be cut away in order to see better. Be careful to avoid cutting the intestines or bladder. The large intestine passes through an aperture in the pelvis. This tube must be separated from the bone surrounding it with a knife. Tie

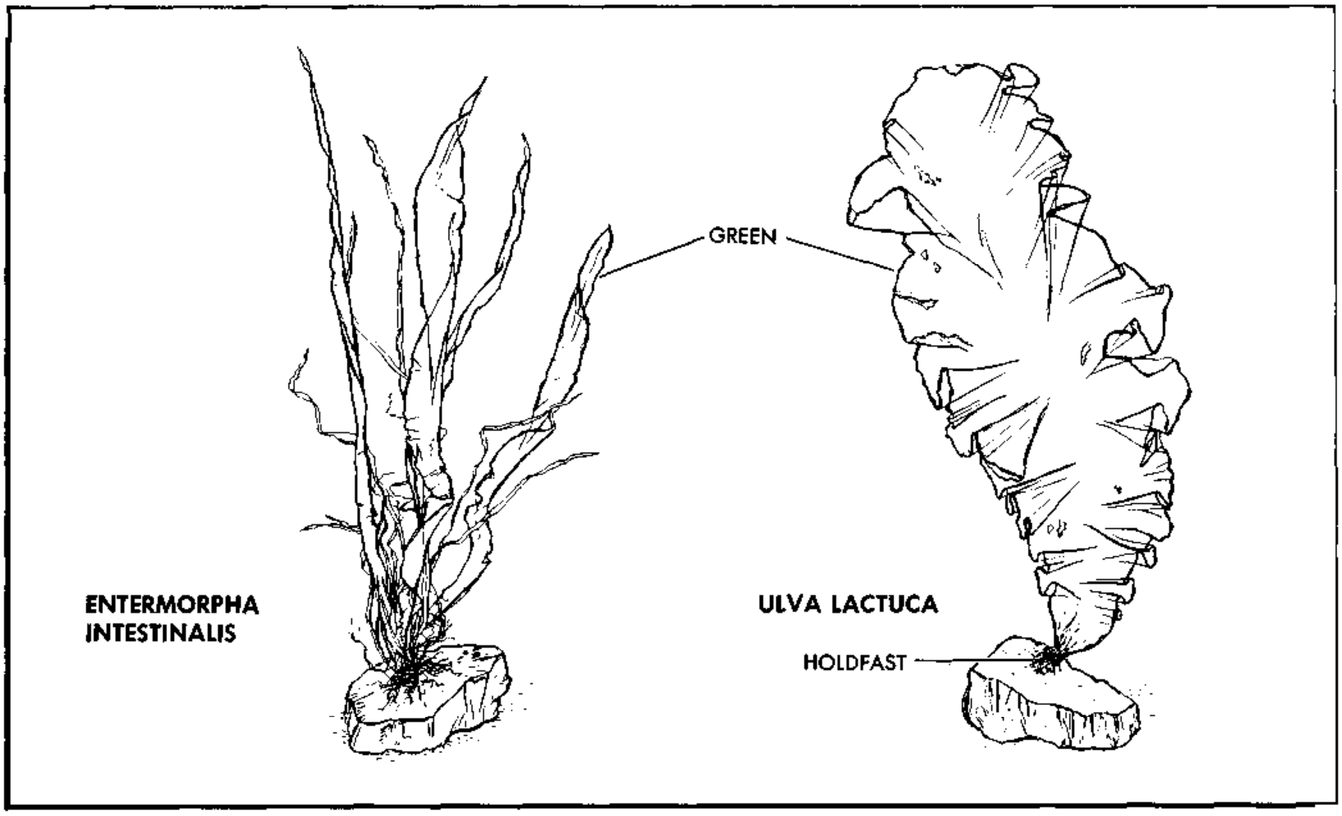
GREEN
ENTERMORPHA INTESTINALIS
ULVA LACTUCA
HOLDFAST

Figure 18-51. Edible Green Seaweeds.

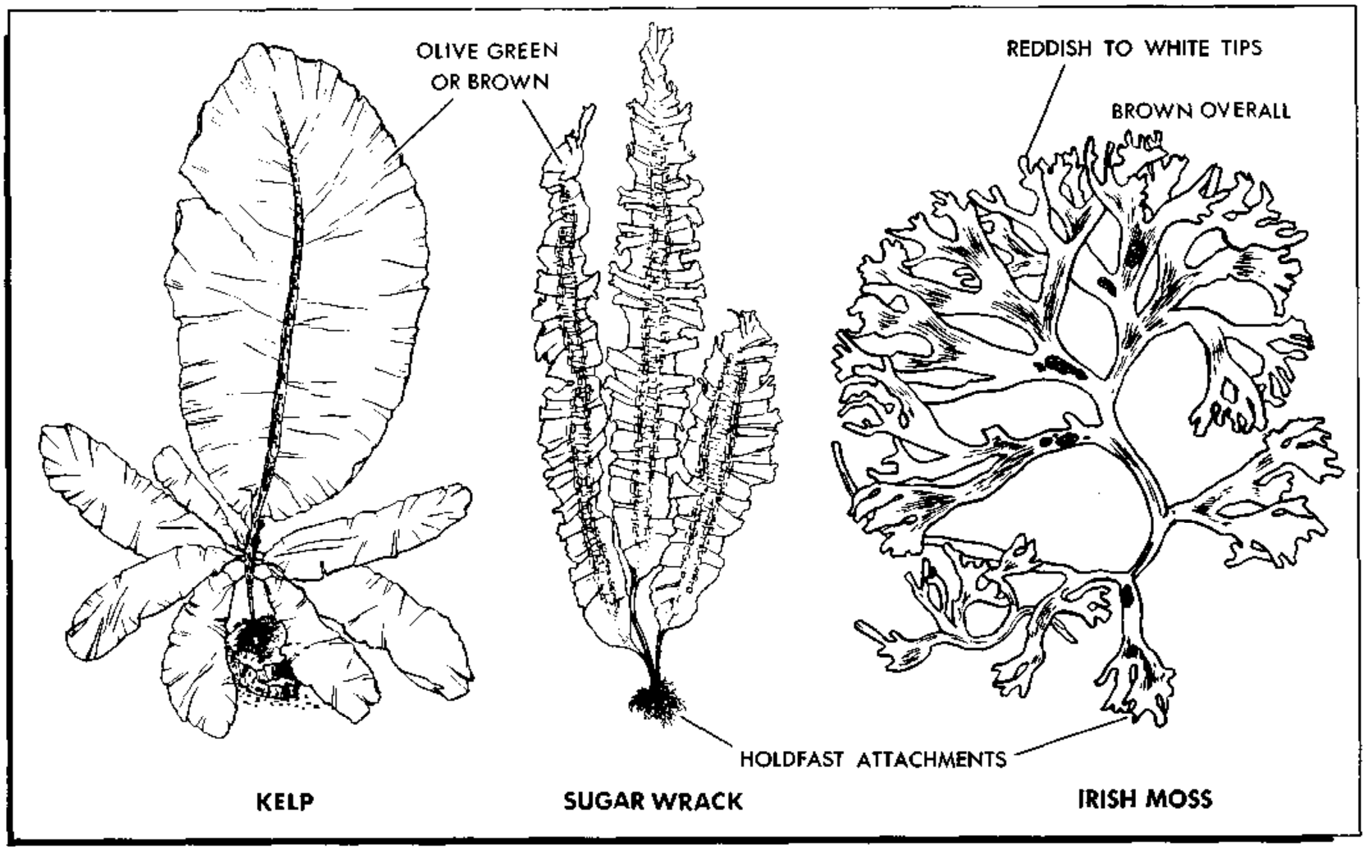
OLIVE GREEN OR BROWN
REDDISH TO WHITE TIPS
BROWN OVERALL
HOLDFAST ATTACHMENTS
KELP
SUGAR WRACK
IRISH MOSS

Figure 18-52. Edible Brown Seaweeds.

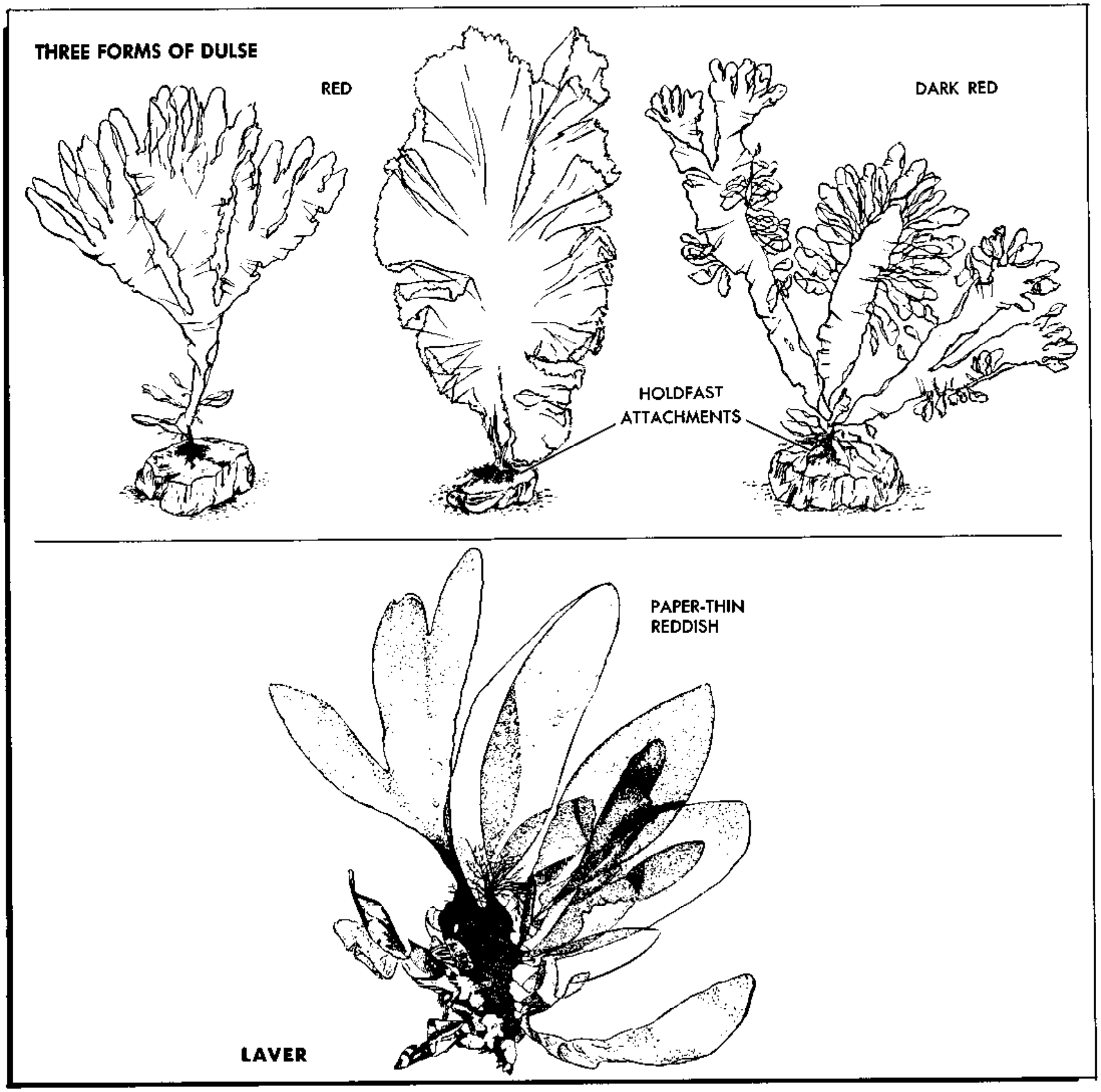

Figure 18-53. Edible Red Seaweeds.

a knot in the bladder tube to prevent the escape of urine. With these steps completed, the entrails can be easily disengaged from the back and removed from the carcass. Another method of gutting or field dressing is shown in figure 18-55. After gutting is completed, it may be advisable to hang the animal. Figure 18-56 shows two methods. (NOTE: If it is hot, gut the animal before skinning it.)

(f) The intestines of a well-conditioned animal are covered with a lace-like layer of fat which can be lifted off and placed on nearby bushes to dry for later use. The gall bladder which is attached to the liver of some animals should be carefully removed. If it should happen to rupture, the bile will taint anything it touches. Be sure to clean the knife if necessary. The kidneys are imbedded in the back, forward of the pelvis, and are covered with fat. Running forward from the kidneys on each side of the backbone are two long strips of chop-meat or muscle called tenderloin or backstrap. Eat this after the liver, heart, and kidneys as it is usually very tender. Edible meat can also be removed from the head, brisket, ribs, backbone, and pelvis.

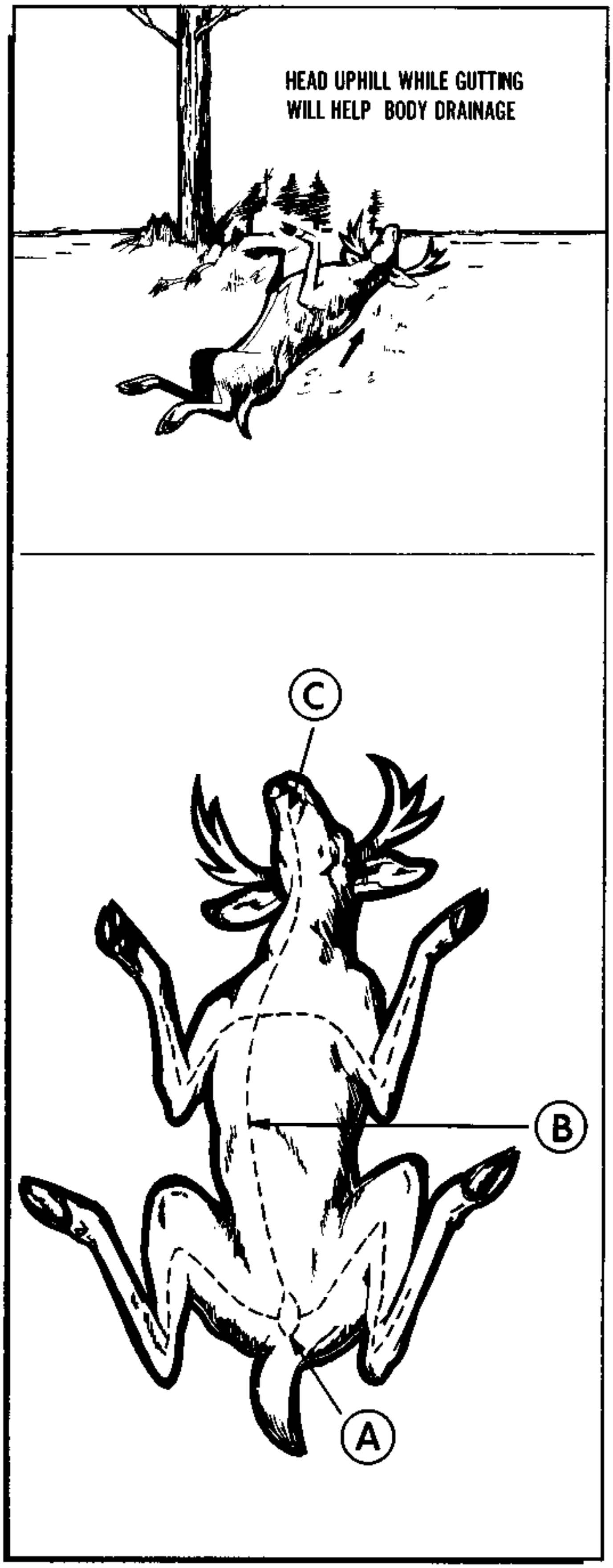

Figure 18-54. Big Game Skinning.

(g) Large animals should be quartered. To do this, cut down between the first and second rib and then sever the backbone with an axe or machete. Cut through the brisket of the front half and then chop lengthwise through the backbone to produce the front quarters. On the rear half, cut through the pelvic bone and lengthwise through the backbone. To make the load lighter and easier to transport, a knife could be used to bone the animal, thereby eliminating the weight of the bones. Butchering is the final step and is simplified for survival purposes. The main purpose is to cut the meat in manageable size portions (figure 18-57).

(2) Glove skinning is usually performed on small game (figure 18-58).

(a) The initial cuts are made down the insides of the back legs. The skin is then peeled back so that the hindquarters are bare and the tail is severed. To remove the remaining skin, pull it down over the body in much the same way a pullover sweater is removed. The head and front feet are severed to remove the skin from the body. For one-cut skinning of small game, cut across the lower back and insert two fingers under each side of the slit. By pulling quickly in opposite directions, the hide will be easily removed (figure 18-59).

(b) To remove the internal organs, a cut should be made into the abdominal cavity without puncturing the organs. This cut must run from the anus to the neck. There are muscles which connect the internal organs to the trunk and they must be severed to allow the viscera to be removed. A rabbit may be gutted by using a knifeless method with no mess and little time lost. Squeeze the entrails toward the rear resulting in a tight bulging abdomen. Raise the rabbit over the head and sling it down hard striking the forearms against the thighs. The momentum will expel the entrails through a tear in the vent (figure 18-60). Save the internal organs such as heart, liver, and kidneys, as they are nutritious. The liver should be checked for any white blotches and discarded if affected as these indicate tularemia (also known as rabbit fever). The disease is transmitted by rodents but also infects humans.

c. Cold-blooded animals are generally easy to clean and prepare.

(1) Snakes and lizards are very similar in taste and they have similar skin. Like the mammals, the skin and viscera should be removed. The easiest way to do this is to sever the head and (or) legs. In the case of a lizard, peel back enough skin so that it may be grasped securely and simply pull it down the length of the body turning the skin inside out as it goes. If the skin does not come away easily, a cut down the length of the animal can be made. This will allow the skin to part from the body more easily. The entrails are then removed and the animal is ready to cook.

(2) Except for the larger amphibians such as the bullfrog, the hind legs are the largest portion of the animal worth saving. To remove the hindquarters, sim-

ply cut through the backbone with a knife, leaving the abdomen and upper body. Pull the skin from the legs and they are ready to cook. With the bullfrogs and larger amphibians, the whole body can be eaten. The head, the skin, and viscera should be removed and discarded (use as bait to catch something else).

d. Most fish need little preparation before they are eaten. Scaling the fish before cooking is not necessary. A cut from the anus to the gills will expose the internal organs which should be removed. The gills should also be removed before cooking. The black line along the inside of the backbone is the kidney and should be removed by running a thumbnail from the tail to the head. There is some meat on the head and should not be discarded. See figure 18-61 for one method of filleting a fish.

e. All birds have feathers which can be removed in two ways: by plucking or pulling out the feathers, and by skinning. The gizzard, heart, and liver should be retained. The gizzard should be split open as it contains partially digested food and stones which must be discarded before being eaten.

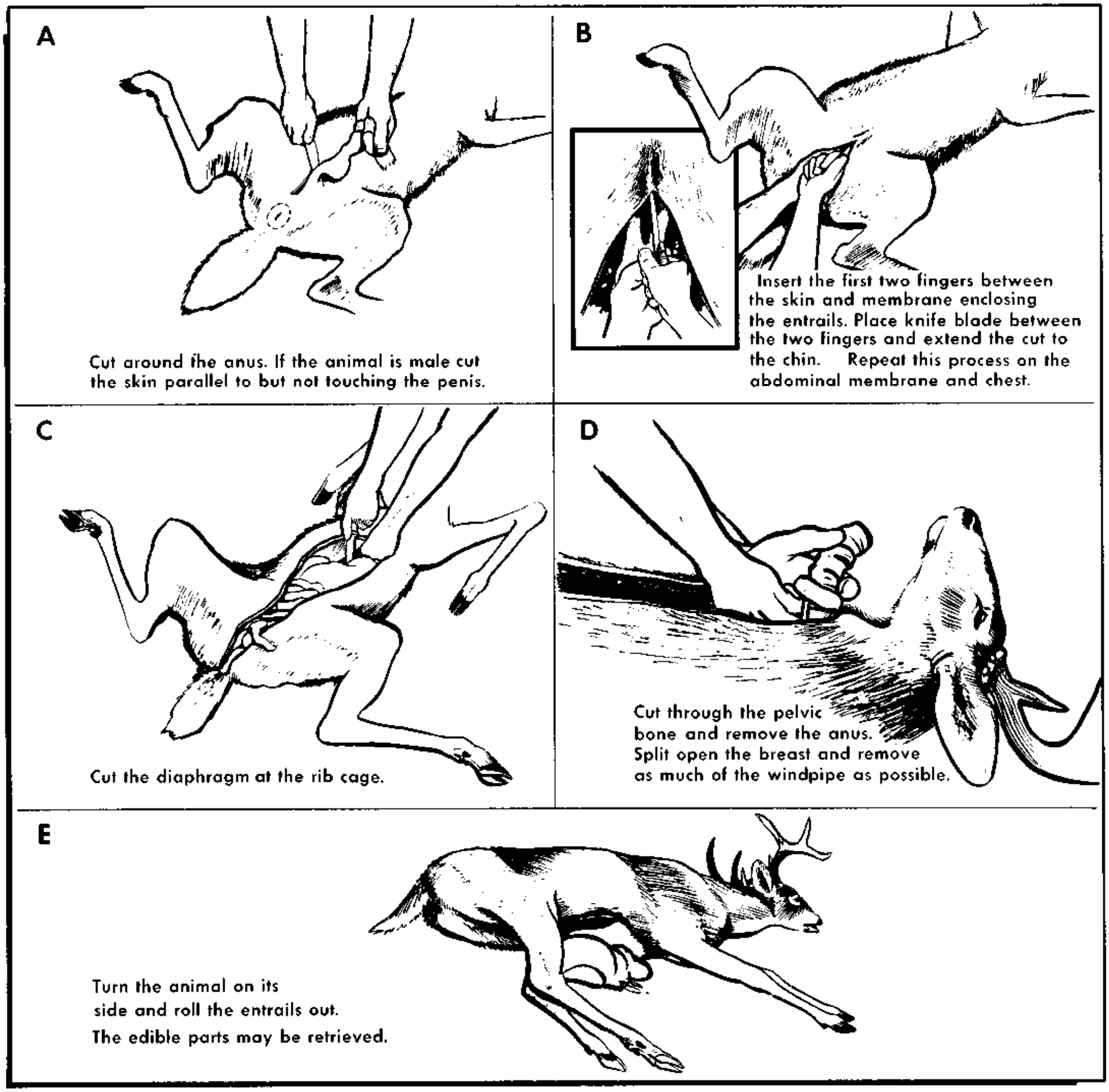

Cut around the anus. If the animal is male cut the skin parallel to but not touching the penis.

Insert the first two fingers between the skin and membrane enclosing the entrails. Place knife blade between the two fingers and extend the cut to the chin. Repeat this process on the abdominal membrane and chest.

Cut the diaphragm at the rib cage.

Cut through the pelvic bone and remove the anus. Split open the breast and remove as much of the windpipe as possible.

Turn the animal on its side and roll the entrails out. The edible parts may be retrieved.

Figure 18-55. Field Dressing.

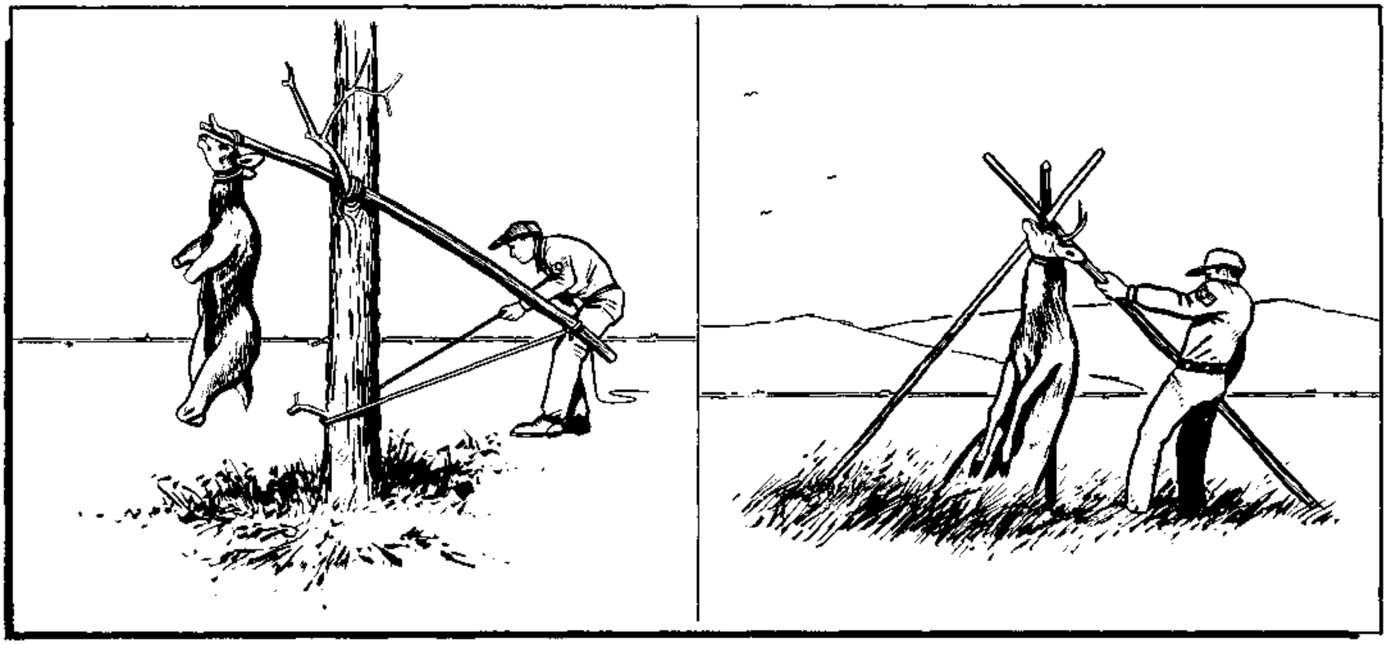

Figure 18-56. Hanging Game.

f. Insects are an excellent food source and they require little or no preparation. The main point to remember is to remove all hard portions such as the hind legs of a grasshopper and the hard wing covers of beetles. The rest is edible.

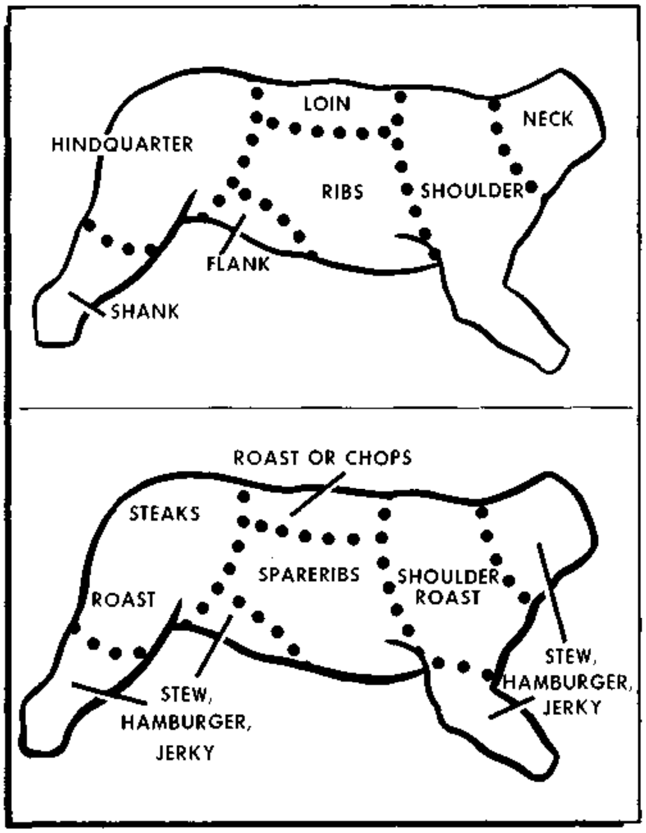

Figure 18-57. Butchering.

18-11. Cooking. All wild game, large insects (grasshoppers), freshwater fish, clams, mussels, snails, and crawfish must be thoroughly cooked to kill internal parasites. Mussels and large snails may have to be minced to make them tender.

a. Boiling is the most nutritious, simplest, and safest method of cooking (figure 18-62). Numerous containers can be used for boiling; for example, a metal container suspended above, or set beside, a heat source to boil foods. Green bamboo makes an excellent cooking container. Stone boiling is a method of boiling using super-heated rocks and a container that holds water but cannot be suspended over an open flame. Example of containers are survival kit containers, flying helmet, a hole in the ground lined with waterproof material, or a hollow log. The container is filled with food and water and then heated with super-hot stones until the water boils. Stones from a stream or damp area should not be used. The moisture in the stones may turn to steam and cause the stone to explode while the stones are being heated in the fire. The container should be covered and new stones added as the water stops boiling. The rocks can be removed with the aid of a wire secured to the rock before being put into the container or two sticks used in a chopstick fashion.

b. Baking is a good method of cooking as it is slow and is usually done by putting food into a container and cooking it slowly. Baking is often used with various types of ovens. Foods may be wrapped in wet leaves (figure 18-63) (avoid using a type of plant that will give an unpleasant flavor to what is being cooked), placed inside a metal container, or they may be packed with mud or clay and placed directly on the coals. Fish and

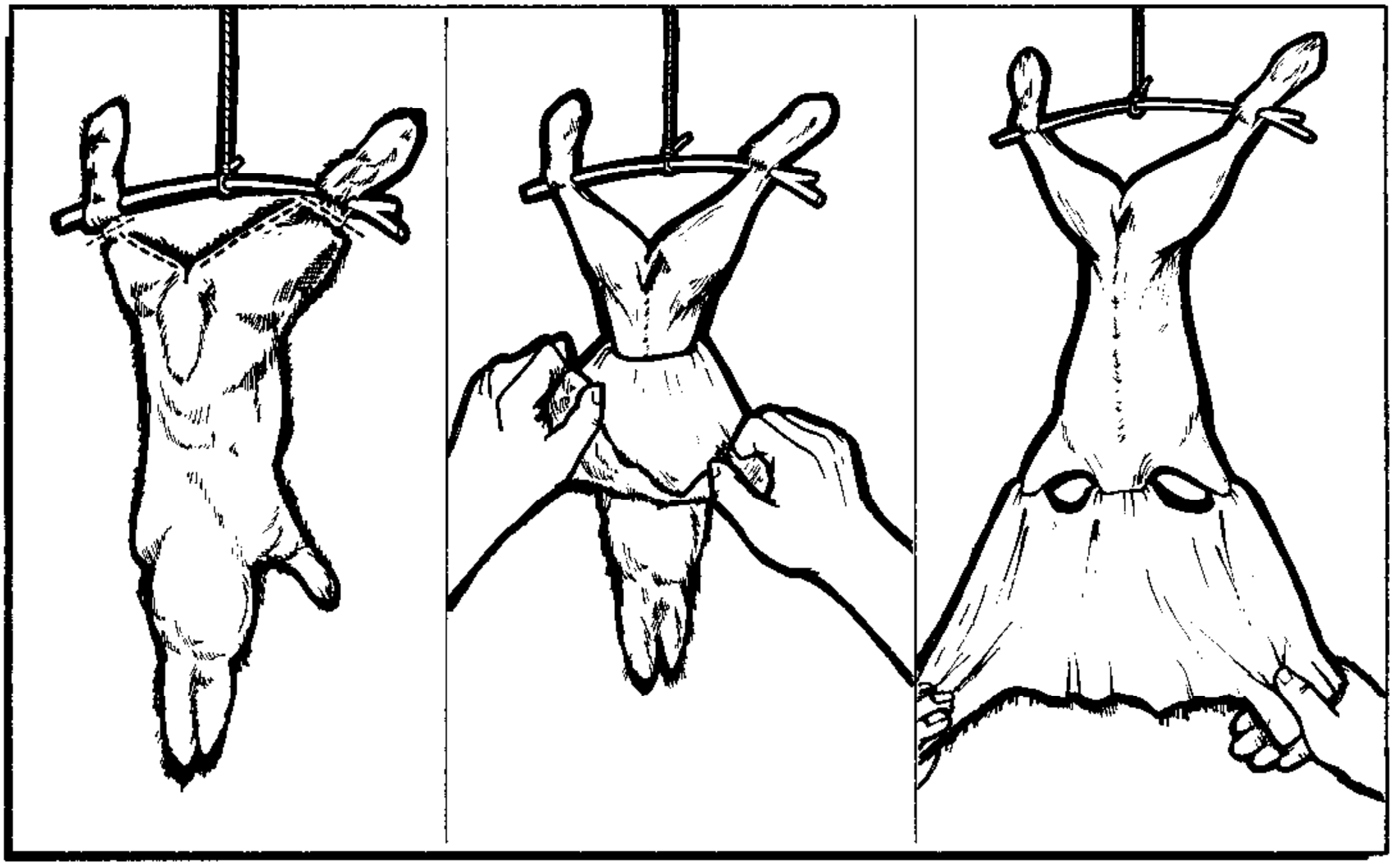

Figure 18-58. Glove Skinning.

birds packed in mud and baked must not be skinned because the scales, skin, or feathers will come off the animal when the mud or clay is removed. Clambake-style baking is done by heating a number of stones in a fire and allowing the fire to burn down to coals. A layer of wet seaweed or leaves is then placed over the hot rocks. Food such as mussels and clams in their shells are placed on the wet seaweed and (or) leaves (figure 18-64). More wet seaweed and (or) leaves and soil is used as a cover. When thoroughly steamed in their own juices, clam, oyster, and mussel shells will open and may be eaten without further preparation.

c. Any type of food can be cooked in the ground in a rock oven (figure 18-65). First, a hole is dug about 2 feet deep and 2 or 3 feet square, depending on the amount of food to be cooked. The sides and bottom are then lined with rock. Next, procure several green trees about 6 inches in diameter and long enough to bridge the hole. Firewood and grass or leaves for insulation should also be gathered. A fire is started in the hole. Two or three green trees are placed over the hole and several rocks are placed on the trees. The fire must be maintained until the green trees burn through. This indicates the fire has burned long enough to thoroughly heat the rocks and the oven is ready. The fallen rocks, fire, and ash are removed from the hole and a thin layer of dirt is spread over the bottom. The insulating material (grass, leaves, moss, etc.) is placed over the soil, then the food more insulating material on top and around the food, another thin layer of soil, and the extra hot rocks are placed on top. The hole is then filled with soil up to ground level.

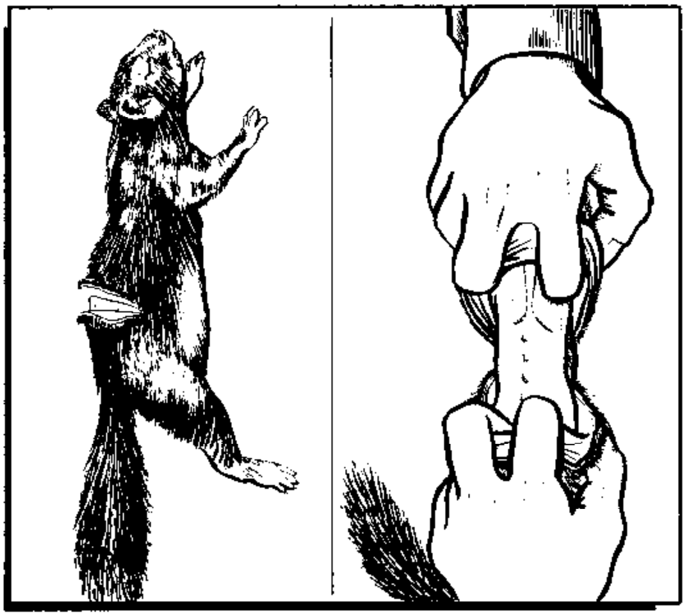

Figure 18-59. Small Animal Skinning.

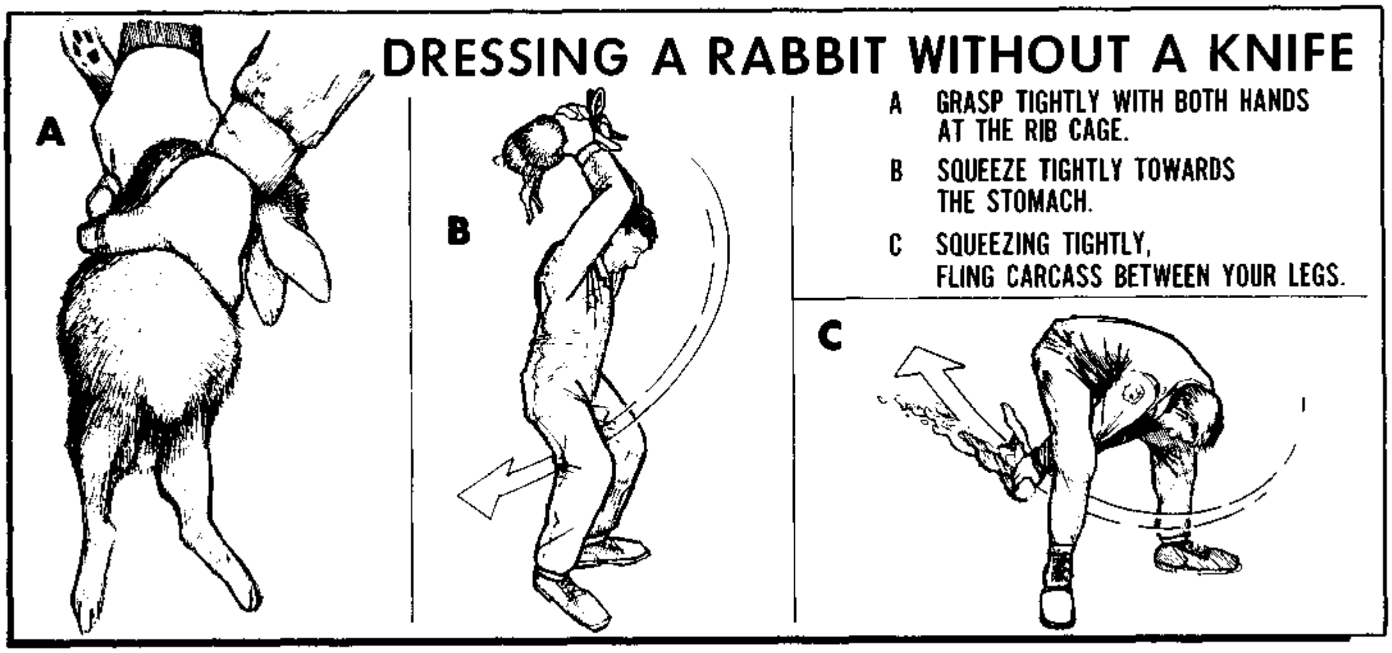
DRESSING A RABBIT WITHOUT A KNIFE
A GRASP TIGHTLY WITH BOTH HANDS AT THE RIB CAGE.
B SQUEEZE TIGHTLY TOWARDS THE STOMACH.
C SQUEEZING TIGHTLY, FLING CARCASS BETWEEN YOUR LEGS.
A
B
C

Figure 18-60. Dressing a Rabbit Without a Knife.

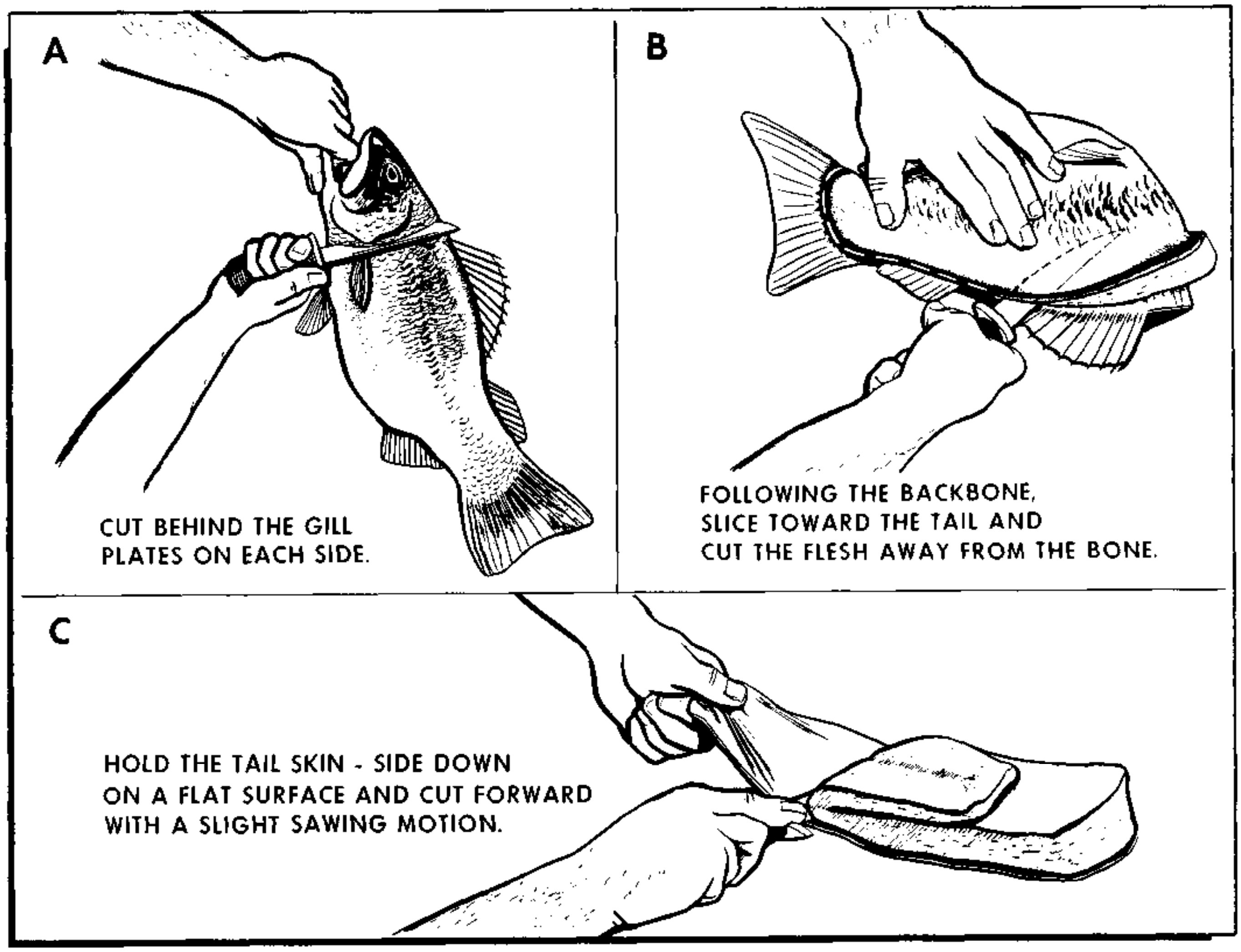
A
CUT BEHIND THE GILL PLATES ON EACH SIDE.
B
FOLLOWING THE BACKBONE, SLICE TOWARD THE TAIL AND CUT THE FLESH AWAY FROM THE BONE.
C
HOLD THE TAIL SKIN - SIDE DOWN ON A FLAT SURFACE AND CUT FORWARD WITH A SLIGHT SAWING MOTION.

Figure 18-61. Filleting a Fish.

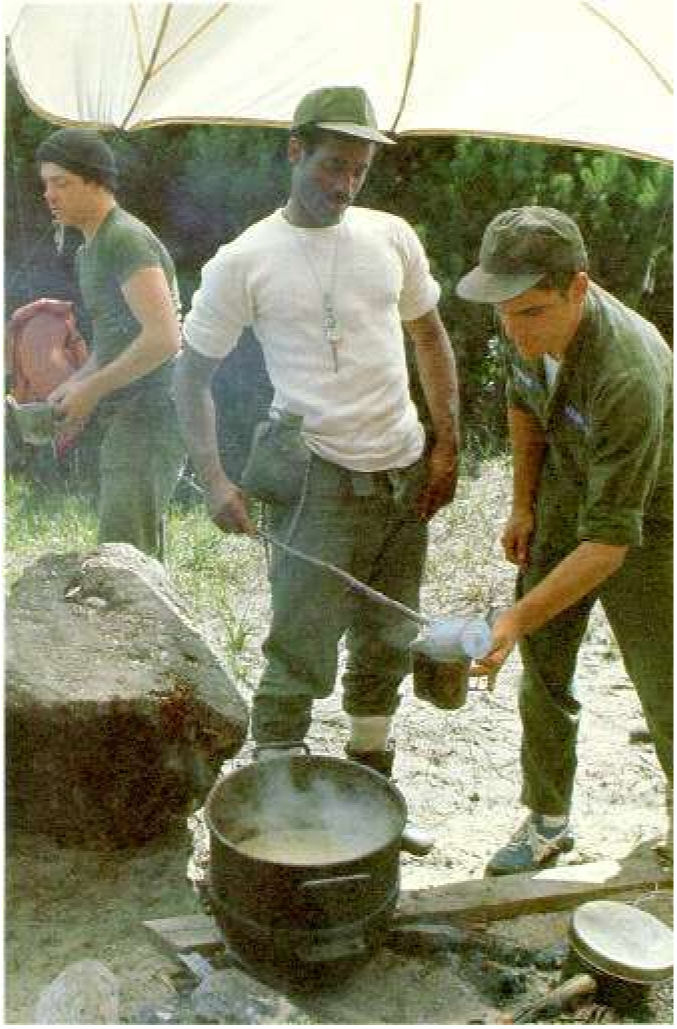

Figure 18-62. Boiling.

Small pieces of meat (steaks, chops, etc.) cook in 1½ to 2 hours and large pieces take 5 to 6 hours.

d. Roasting is less desirable as it involves exposing the food to direct heat which quickly destroys the nutritional properties (figure 18-66). Putting a piece of meat on a stick and holding it over the fire is considered roasting.

e. Broiling is the quickest way to prepare fish. A rock broiler may be made by placing a layer of small stones on top of hot coals, and laying the fish on the top. Scaling the fish before cooking is not necessary, and small fish need not be cleaned. Cooked in this manner, fish have a moist and delicious flavor. Crabs and lobsters may also be placed on the stones and broiled.

f. Meat may be cooked by laying it on a flat board or stone (planking) which is propped up close to the fire (figure 18-67). The meat will have to be turned over at least once to allow thorough cooking. The cooking time depends on how close the meat is to the fire.

g. Frying is by far the least favorable method of preparing food. It tends to make the meat tough because most all of the natural juices are cooked out of the meat. Some of the nutritional value of the meat will also be destroyed. Frying can be done on any nonporous surface which can be heated. Examples are unpainted aircraft parts, turtle shells, large seashells, flat rocks, and some survivial kit parts.

18-12. Preparing Food in Enemy Areas. The problem of preparing food in a hostile area becomes acute when a fire, even a small cooking fire, can bring about capture. After finding food in a hostile area, the problem of preparing the food in a manner which will not compromise the survivors presence must be resolved. Of course, it

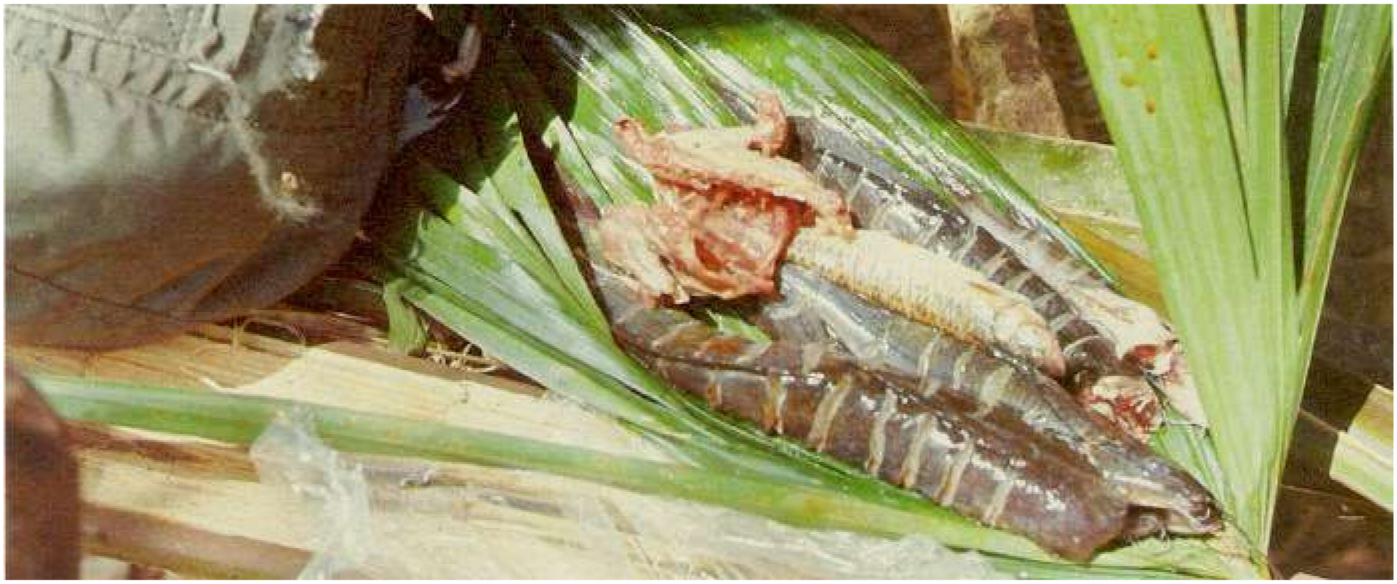

Figure 18-63. Baking.

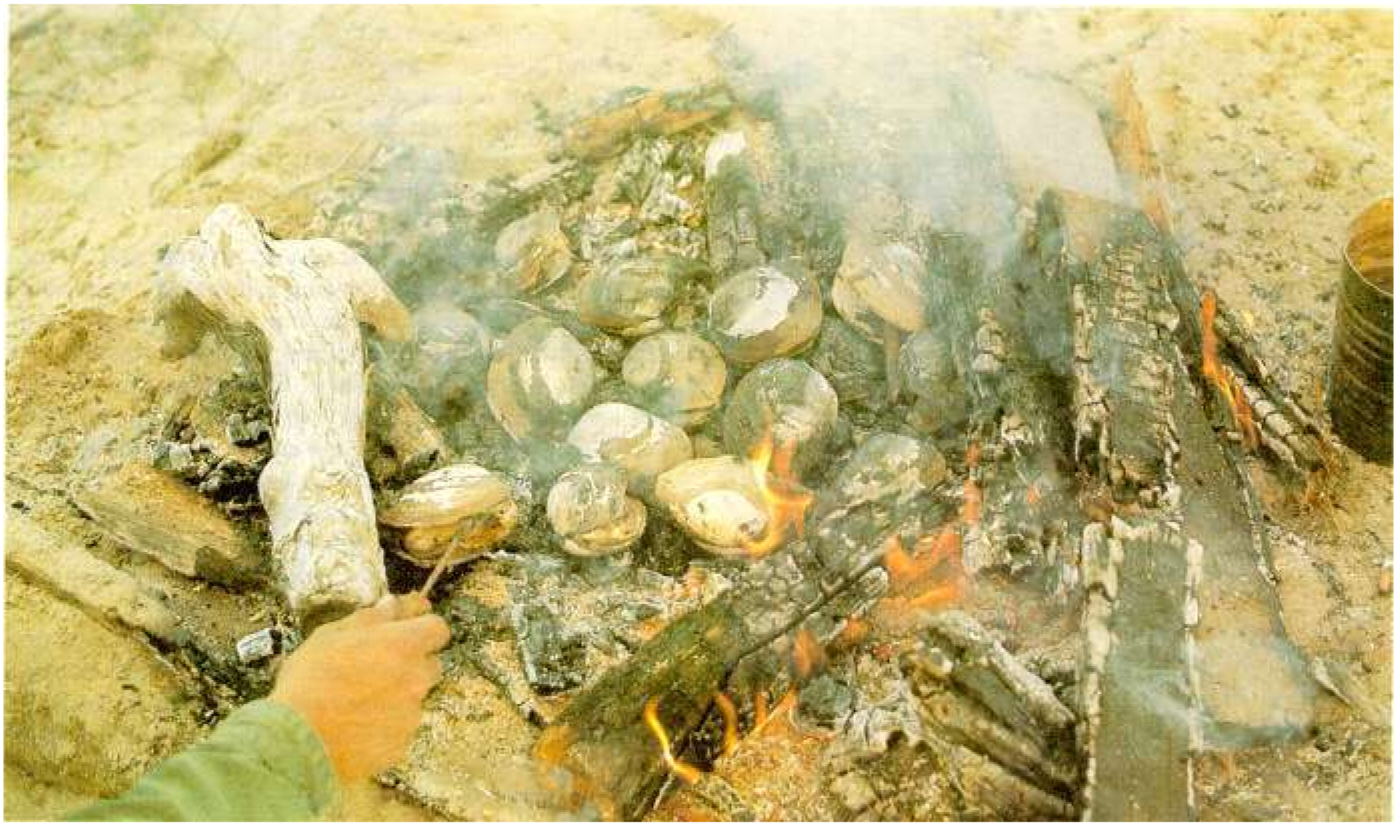

Figure 18-64. Clam Baking.

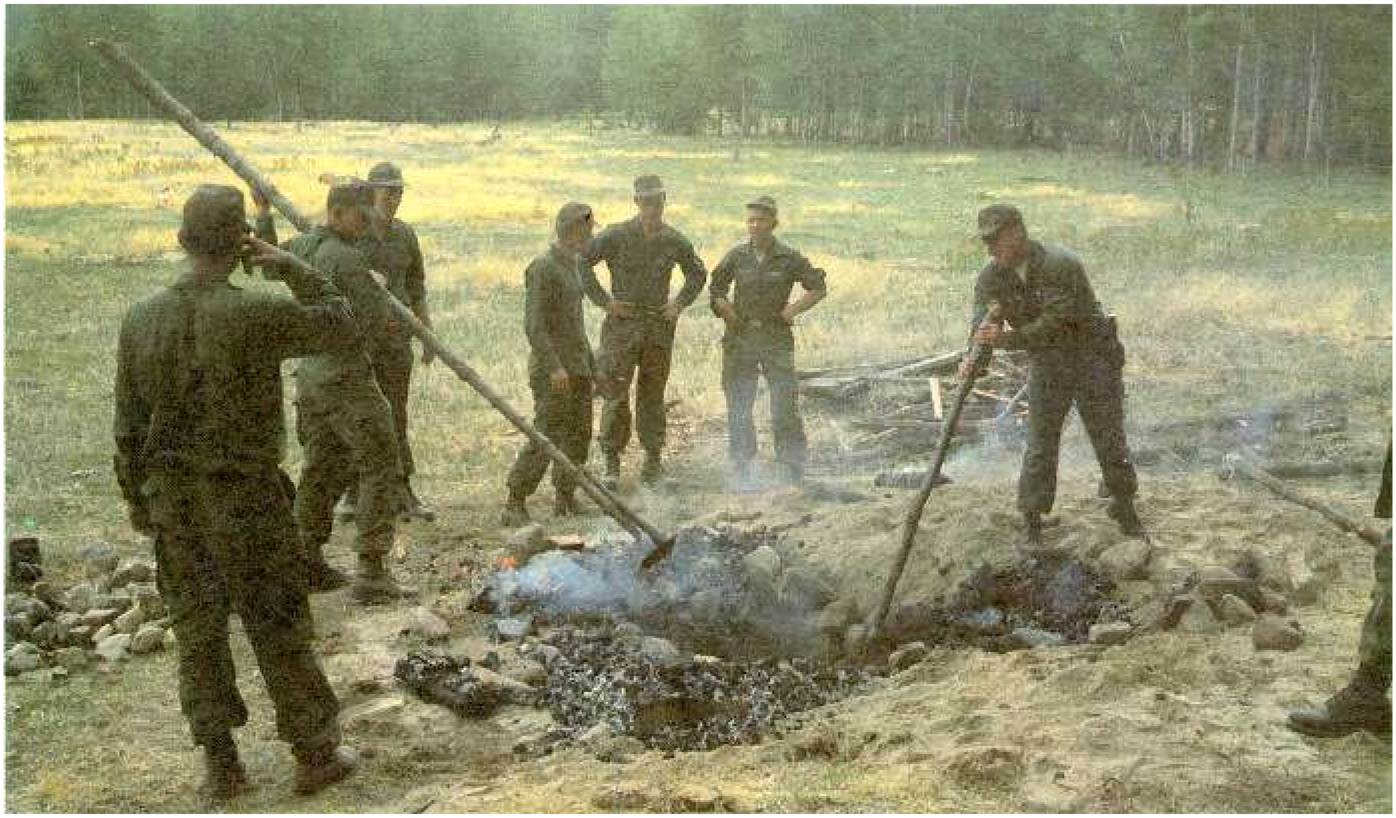

Figure 18-65. Rock Oven.

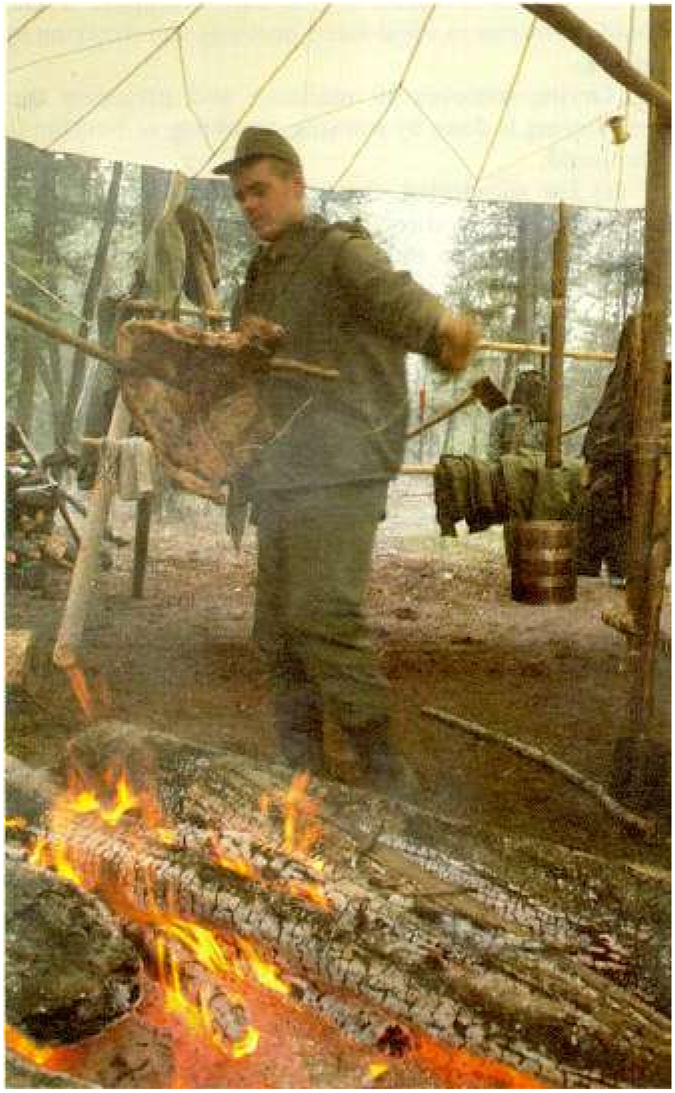

Figure 18-66. Broiling and Roasting.

would be simple to state that the best solution would be to eat the food without cooking it.

a. In some respects, this would be a more reasonable solution than it might initially seem to be. From the standpoint of palatability, it is mostly a matter of adjusting the "frame of mind." Animal foods are recognized as being palatable when cooked to a very minor degree. The need for food cannot be ignored and the situation may demand that it be eaten partially cooked or even uncooked.

b. With regard to the health considerations involved, many of the reasons for cooking are recognized as a means of destroying organisms that may be present in the food and can cause sickness or ill effects if they enter into the body. Under survival conditions in a hostile area, one may be forced to forego thorough cooking and accept the risk involved until their return to friendly forces where professional treatment is available.

c. Assuming that there will have to be a way to prepare food under hostile conditions, a survivor should be aware of some of the ways in order to achieve some degree of safety, and at the same time, improve palatability. Parasites and other organisms living in the flesh of the animals depend upon the body temperature of the animals, the moisture within the flesh of the animals, and other factors to support their life. Any action that modifies these conditions (for example, freezing or thorough drying of the meat) and kills some parasites may improve the palatability.

d. If cooking is considered necessary, use extreme care in selecting the site for a fire and ensure that security considerations are favorable. The food should be prepared in very small quantities in order to keep the size of the fire as small as possible. The use of the "Dakota Hole" configuration is more appropriate for cooking food during a tactical situation (figure 16-14).

18-13. Preserving Food. Finding natural foods is an uncertain aspect of survival. The survivor must make the best use of the available food. Food, especially meat, has a tendency to spoil within a short period of time unless it is preserved. There are many ways to preserve food; some of the most common are cooking, refrigeration, freezing, and dehydration.

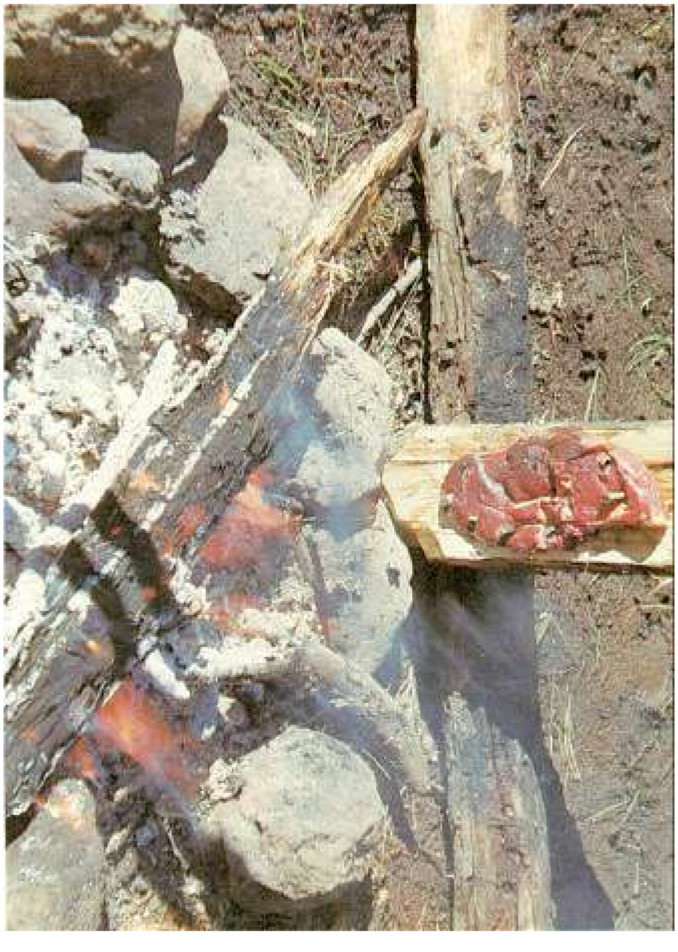

Figure 18-67. Planking.

a. Cooking will slow down the decomposition of food but will not eliminate it. This is because many bacteria are present which work to break it down. Cooking methods which are the best for immediate consumption, such as boiling, are the least effective for preserving food. Food should be recooked every day until all is consumed.

b. Cooling is an effective method of storing food for short periods of time. Heat tends to accelerate the decomposition process where cooling retards decomposition. The colder food becomes, the less the likelihood of deterioration until freezing eliminates decomposition. Cooling devices available to a survivor are:

(1) Food items buried in snow will maintain a temperature of approximately 32°F.

(2) Food wrapped in waterproof material and placed in streams will remain cool in summer months. Care should be taken to ensure food is secured.

(3) Earth, below the surface, particularly in shady areas or along streams, remains cooler than the surface. A hole may be dug, lined with grass, and covered to form an effective cool storage area much the same as a root celler.

(4) When water evaporates, it tends to cool down the surrounding area. Using this fact, articles of food may be wrapped in an absorbent material such as cotton or burlap and rewetted as the water evaporates.

c. Once food is frozen, it will not decompose. Food should be frozen in meal-sized portions so refreezing is avoided.

d. Drying removes all moisture and preserves the food. Drying is done by sunning, smoking, or burying it in hot sand.

(1) For sun-drying, the food should be sliced very thin and placed in direct sunlight. Meat should be cut across the grain to improve tenderness and decrease drying time. If salt is available, it should be added to improve flavor and accelerate the drying process (figure 18-68).

(2) Smoking is a process done through the use of nonresinous wood such as willow or aspen and is used to produce smoke which adds flavor and dries the meat. A smoke rack is also necessary to contain the smoke (figure 18-69). The following are the procedures for drying meat using smoke:

(a) Cut meat very thin and across the grain. If the meat is warm and difficult to slice thin, cut the meat in 1 or 2-inch cubes and beat it thin with a clean wooden mallet (improvised).

(b) Remove fat.

(c) Hang the meat on a rack so each piece is separate.

(d) Elevate meat no less than 2 feet above coals.

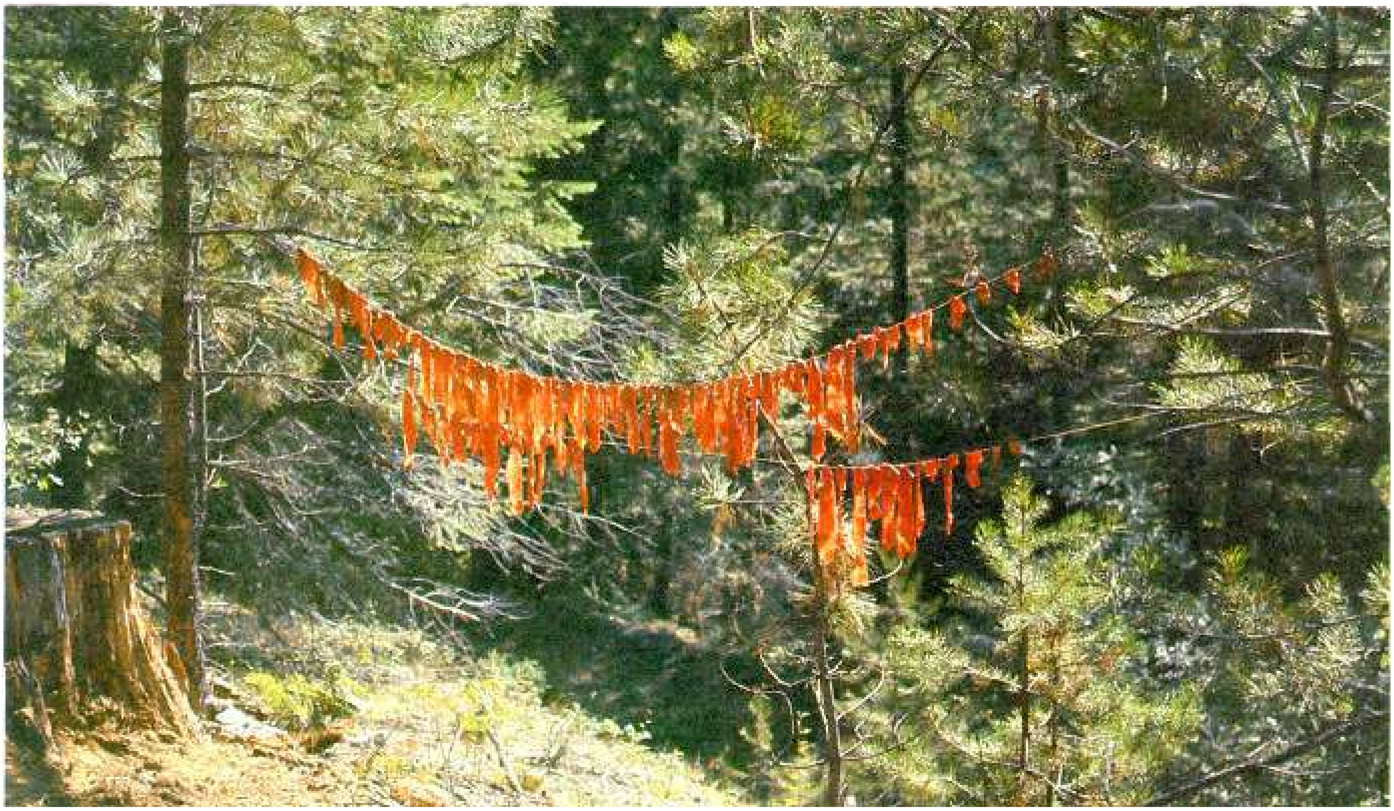

Figure 18-68. Sun-Drying.

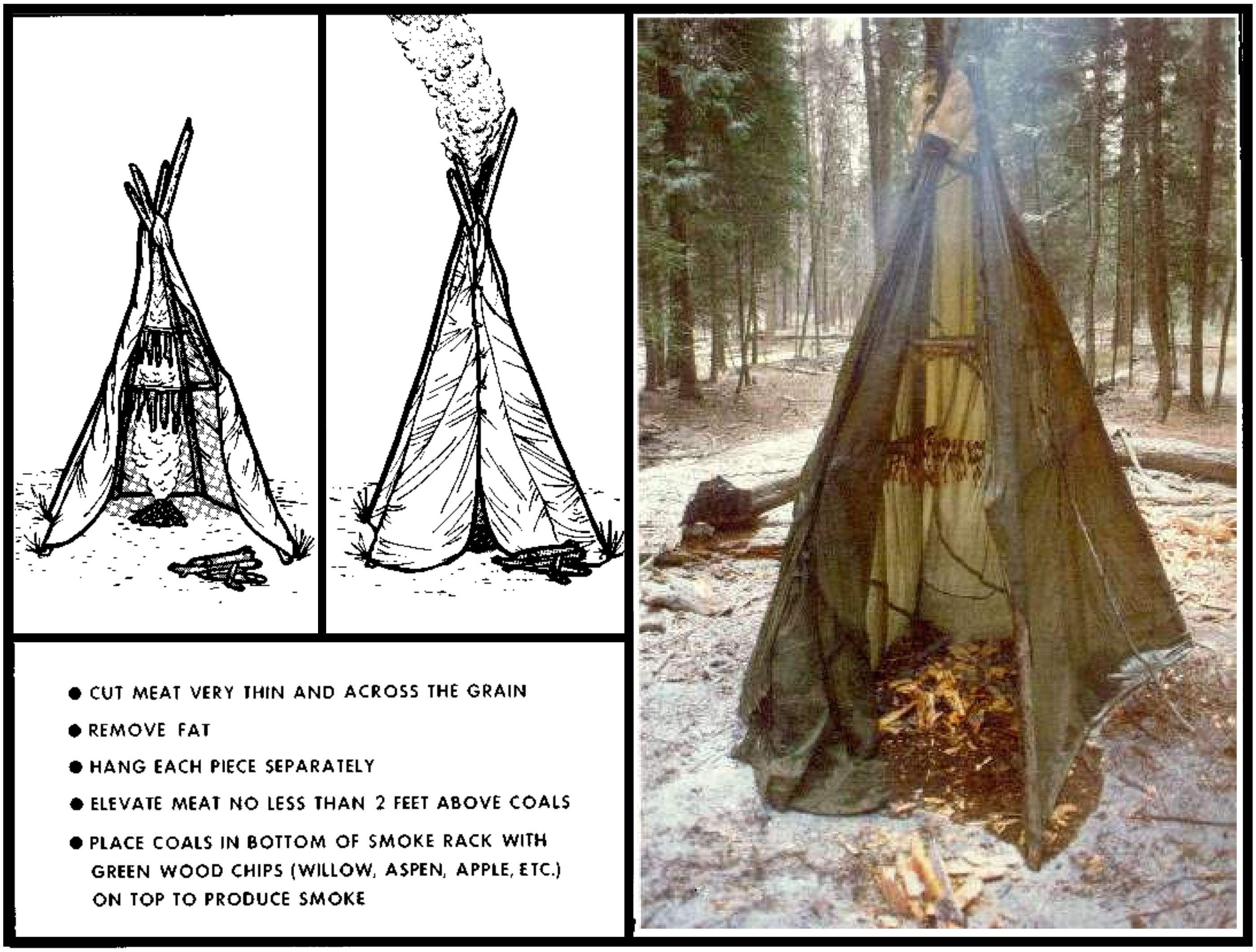

- CUT MEAT VERY THIN AND ACROSS THE GRAIN
- REMOVE FAT
- HANG EACH PIECE SEPARATELY
- ELEVATE MEAT NO LESS THAN 2 FEET ABOVE COALS
- PLACE COALS IN BOTTOM OF SMOKE RACK WITH GREEN WOOD CHIPS (WILLOW, ASPEN, APPLE, ETC.) ON TOP TO PRODUCE SMOKE

Figure 18-69. Smoke-Drying.

(e) Coals are placed in the bottom of a smoke rack with green woodchips on top to produce smoke.

e. The method used to preserve fish through warm weather is similar to that used in preserving meat (figure 18-70). When there is no danger of predatory animals disturbing the fish, the fish should be placed on available fabric and allowed to cool during the night. Early the next morning, before the air gets warm, the fish should be rolled in moist fabric (and leaves). This bundle can be placed inside the survivor's pack. During the rest periods, or when the pack is removed, it should be placed in a cool location out of the Sun's rays.

(1) Fish may be dried in the same manner described for smoking meat. To prepare fish for smoking, the heads and backbone are removed and the fish are spread flat on a grill. Thin willow branches with bark removed make skewers.

(2) Fish may also be dried in the Sun. They can be suspended from branches or spread on hot rocks. When the meat has dried, sea water or salt should be used on the outside, if available.

f. In survival environments, there are many animals and insects that will devour a survivor's food if it is not correctly stored. Protecting food from insects and birds is done by wrapping it in parachute material, wrapping and tying brush around the bundle, and finally, wrapping it with another layer of material. This creates "dead air" space making it more difficult for insects and birds to get to the food. If the outer layer is wetted, evaporation will also cool the food to some degree. In most cases, if the food is stored several feet off the ground, it will be out of reach of most animals. This can be done by hanging the food or putting it into a "cache". If the food is dehydrated, the container must be completely waterproof to prevent reabsorption. Frozen food will remain frozen only if the outside temperature remains below freezing. Burying food is a good way to store as long as scavengers are not in the area to uncover it. Insects and small animals should also be

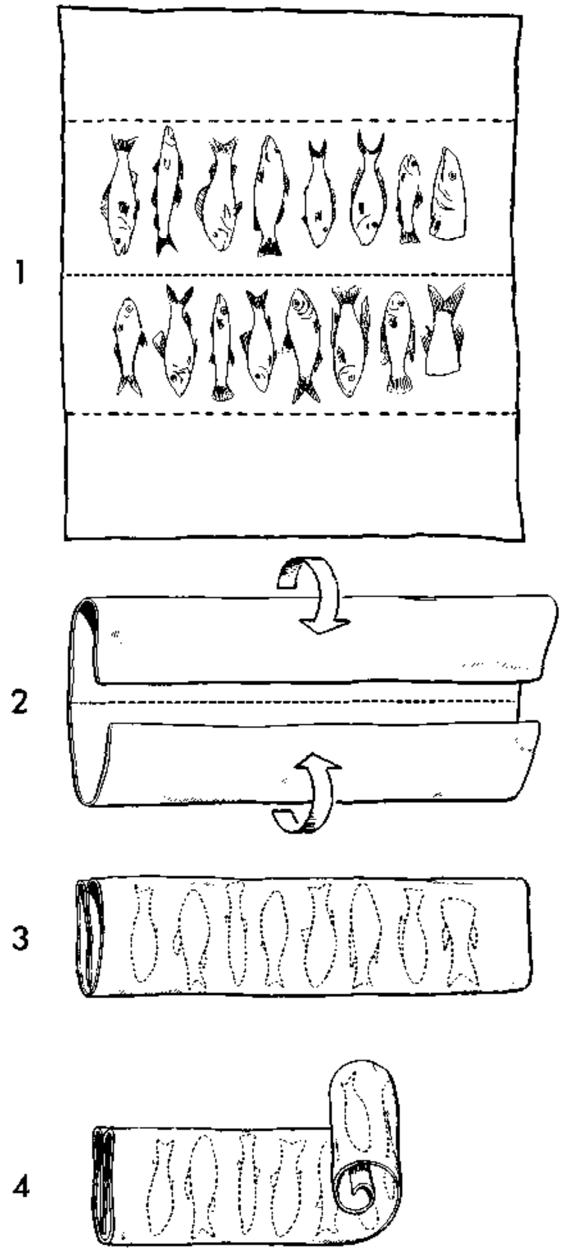

1. Arrange fish on available fabric.

2. Turn down the upper edge of wrap over the top line of fish and turn up the lower edge over the lower line.

3. Fold in the center as shown.

4. Then begin on the edge and roll the wrap. You will have a rounded roll of protected fish. This roll should be securely, but not tightly, tied and wrapped in a sleeping bag, parachute fabric or clothing, as you would do with meat.

Figure 18-70. Preserving Fish.

remembered when burying the food. Food should never be stored in the shelter as this may attract wild animals and could be hazardous to the survivors.

18-14. Preparing Plant Food. Preparing plant foods can be more involved than preparing animal life.

a. Some plant foods, such as acorns and tree bark may be bitter because of tannin. These plants will require leaching by chopping up the plant parts, and pouring several changes of fresh water over them. This will help wash out the tannin, making the plant more palatable. Other plants such as cassava and green papaya must be cooked before eating to break down the harmful enzymes and chemical crystals within them and make them safe to eat. Plants such as skunk cabbage must undergo this cooking process several times before it is safe to eat.

b. All starchy foods must be cooked since raw starch is difficult to digest. They are boiled, steamed, roasted, or fried and are eaten plain, or mixed with other wild foods. The manioc (cassava) is best cooked, because the bitter form (green stem) is poisonous when eaten raw. Starch is removed from sago palm, cycads, and other starch-producing trunks by splitting the trunk and pounding the soft, whitish inner parts with a pointed club. This pulp is washed with water and the white sago (pure starch) is drained into a container. It is washed a second time, and then it may be used directly as a flour. One trunk of the sago palm will supply a survivor's starch needs for many weeks.

c. The fiddleheads of all ferns are the curled, young succulent fronds which have the same food value as cabbage or asparagus. Practically all types of fiddleheads are covered with hair which makes them bitter. The hair can be removed by washing the fiddleheads in water. If fiddleheads are especially bitter, they should be boiled for 10 minutes and then reboiled in fresh water for 30 to 40 minutes. Wild bird eggs or meat may be cooked with the fiddleheads to form a stew.

d. Wild grasses have an abundance of seeds, which may be eaten boiled or roasted after separating the chaff from the seeds by rubbing. No known grass is poisonous. If the kernels are still soft and do not have large stiff barbs attached, they may be used for porridge. If brown or black rust is present, the seeds should not be eaten (Ergot Poisoning). To gather grass seeds, a cloth is placed on the ground and the grass heads beaten with sticks.

e. Plants that grow in wet places along margins of rivers, lakes, and ponds, and those growing directly in water are of potential value as survival food. The succulent underground parts and stems are most frequently eaten. Poisonous water plants are rare. In temperate climates, the water hemlock is the most poisonous plant found around marshes and ponds. In the tropics, the various members of the calla lily family often grow in very wet places. The leaves of the calla lily look like

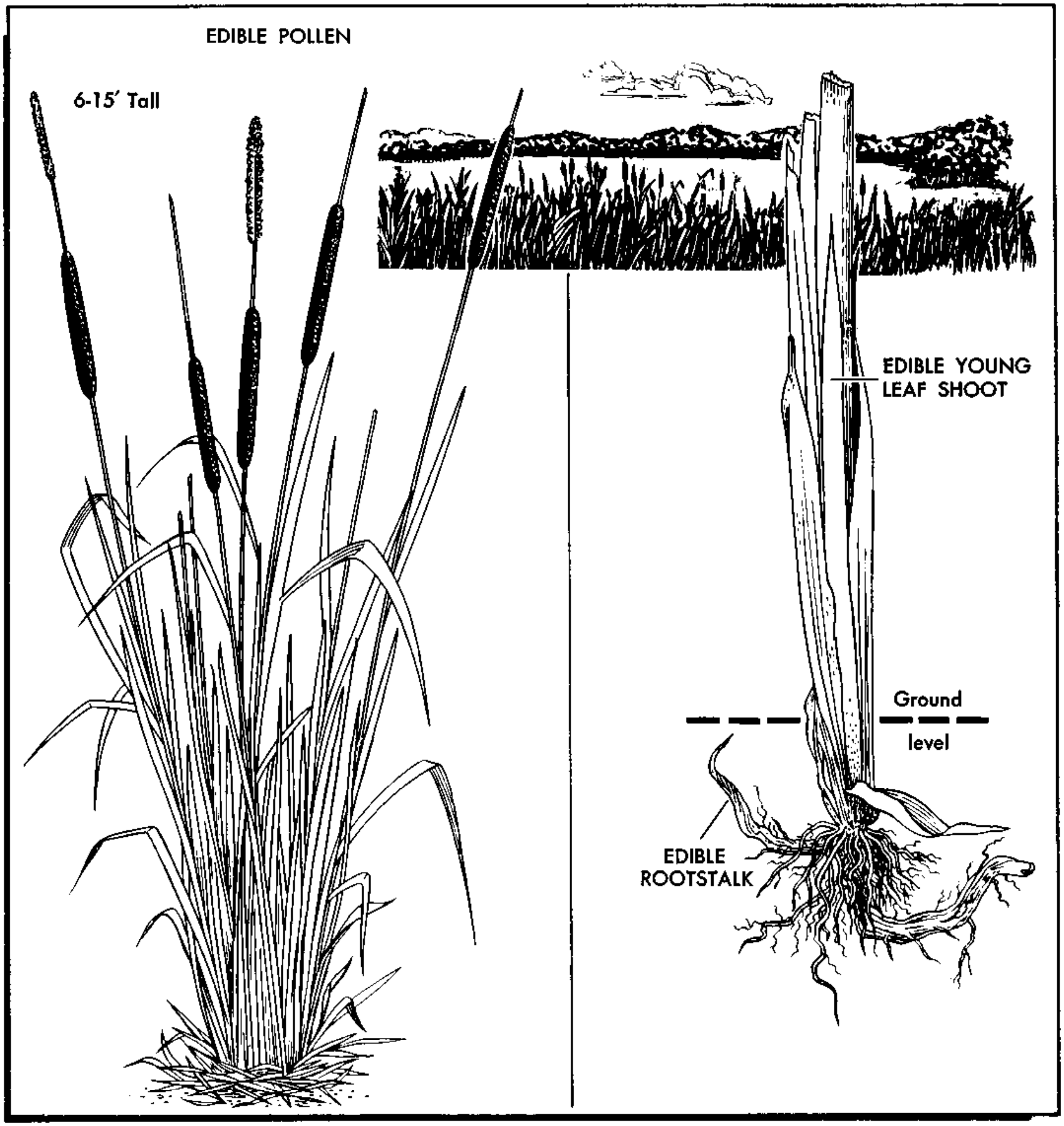

Figure 18-71. Cattails.

arrowheads. Jack-in-the-pulpit, calla lily, and sweet flag are members of the Arum family. To be eaten, the members of this plant family must be cooked in frequent changes of water to destroy the irritant crystals in the stems. Two kinds of marsh and water plants are the cattail and the water lily.

(1) The cattail (Typha) is found worldwide except in tundra regions of the far north (figure 18-71). Cattails can be found in the more moist places in desert areas of all continents as well as in the moist tropic and temperate zones of both hemispheres. The young shoots taste like asparagus. The spikes can be boiled or steamed when green and then eaten. The rootstalks, without the outer covering, are eaten boiled or raw. Cattail roots can be cut into thin strips, dried, and then ground into flour. They are 46 percent starch, 11 percent sugar, and the rest is fiber. While the plant is in flower, the yellow pollen is very abundant; this may be mixed with water

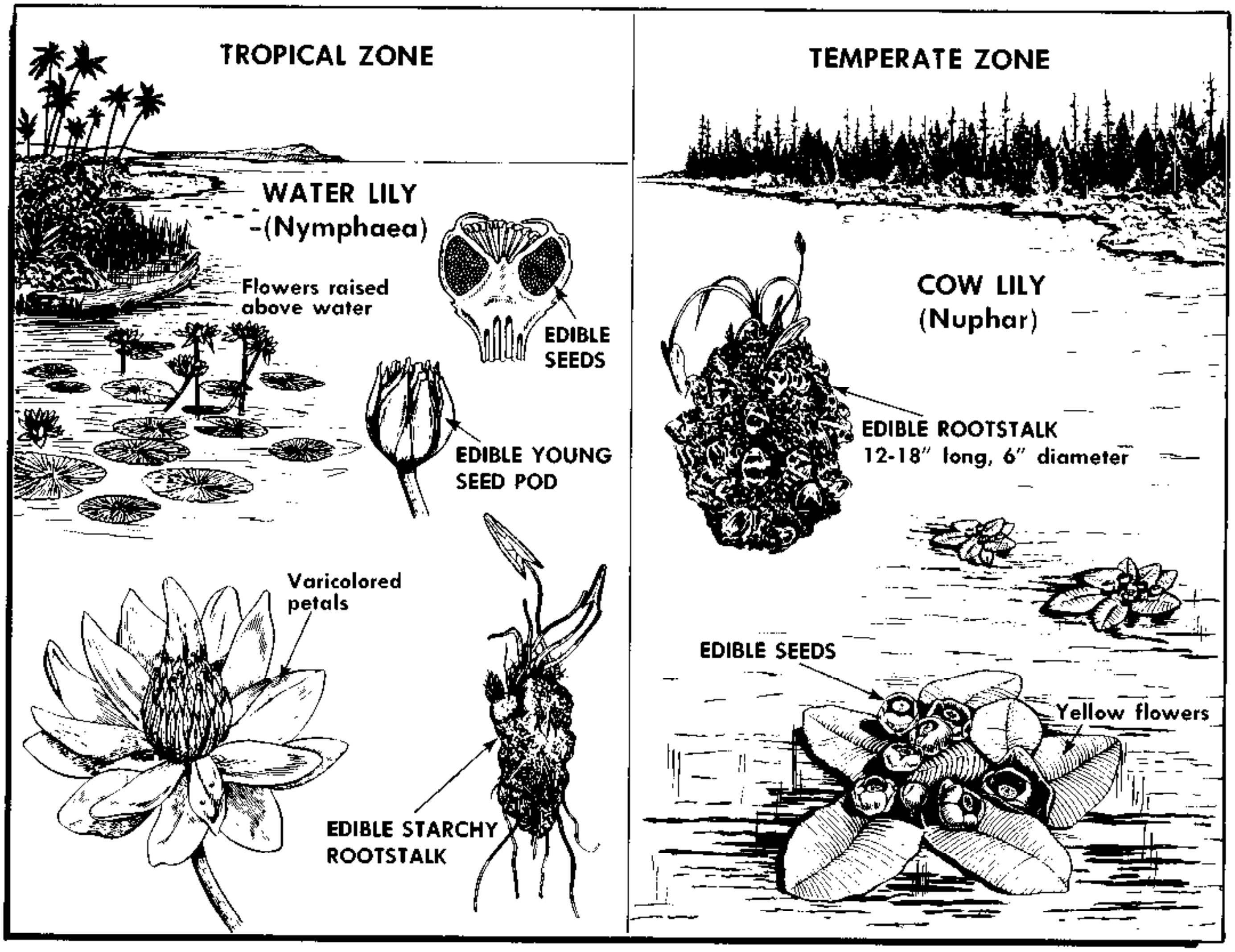
TROPICAL ZONE
WATER LILY
(Nymphaea)
Flowers raised above water
EDIBLE SEEDS
EDIBLE YOUNG SEED POD
Varicolored petals
EDIBLE STARCHY ROOTSTALK
TEMPERATE ZONE
COW LILY
(Nuphar)
EDIBLE ROOTSTALK
12-18" long, 6" diameter
EDIBLE SEEDS
Yellow flowers

Figure 18-72. Water Lilies.

and made into small cakes and steamed as a substitute for bread.

(2) Water lilies (Nymphaea and Nuphar) occur on all the continents, but principally in southern Asia, Africa, North America, and South America (figure 18-72). Two main types are:

(a) Temperate water lilies produce enormous rootstalks and yellow or white flowers which float on the water.

(b) Tropical water lilies produce large edible tubers and flowers which are elevated above the water surface.

(3) Rootstalks or tubers may be difficult to obtain because of deep water. They are starchy and high in food value. They can be eaten either raw or boiled. Stems may be cooked in a stew. Young seed pods may be sliced and eaten as a vegetable. Seeds may be bitter, but are very nourishing. They may be parched and rubbed between stones as flour. The water lily is considered an important food source by native peoples in many parts of the world.

f. Nuts are very high in nutritional value and usually can be eaten raw. Nuts may be roasted in the fire or roasted by shaking them in a container with hot coals from the fire. They may then be ground to make a flour. If a survivor does not wish to eat a plant or plant part raw, it can be cooked using the same methods used in cooking meat—by boiling, roasting, baking, broiling, or frying.

g. If survivors have been able to procure more plant foods than can be eaten, the excess can be preserved in the same manner as animal foods. Plant foods can be dried by wind, air, sun, or fire, with or without smoke. A combination of these methods can be used. The main object is to remove the moisture. Most wild fruits can be dried. If the plant part is large, such as some tubers, it should be sliced, and then dried. Some type of protection may be necessary to prevent consumption and (or) contamination by insects. Extra fruits or berries can be carried with the survivor by wrapping them in leaves or moss.

Chapter 19

WATER

19-1. Introduction. Nearly every survival account details the need survivors had for water. Many ingenious methods of locating, procuring, purifying, and storing water are included in the recorded experiences of downed aircrew members. If survivors are located in temperate, tropic, or dry climates, water may be their first and most important need. The priority of finding water over that of obtaining food must be emphasized to potential survivors. An individual may be able to live for weeks without food, depending on the temperature and amount of energy being exerted. A person who has no water can be expected to die within days. Even in cold climate areas or places where water is abundant, survivors should attempt to keep their body fluids at a level that will maintain them in the best possible state of health. Even in relatively cold climates, the body needs 2 quarts of water per day to remain efficient (figure 19-1).

19-2. Water Requirements. Normally, with atmospheric temperature of about 68°F, the average adult requires 2 to 3 quarts of water daily.

a. This water is necessary to replace that lost daily in the following ways:

(1) Urine. Approximately 1.4 quart of water is lost in the urine.

(2) Sweat. About 0.1 quart of water is lost in the sweat.

(3) Feces. Approximately 0.2 quart of water is lost in the feces.

(4) Insensible Water Loss. When the individual is unaware water loss is actually occurring, it is referred to as insensible water loss. Insensible water loss occurs by the following mechanisms:

(a) Diffusion through the skin. Water loss through the skin occurs as a result of the actual diffusion of water molecules through the cells of the skin. The average loss of water in this manner is approximately 0.3 to 0.4 quart. Fortunately, loss of greater quantities of water by diffusion is prevented by the outermost layer of the skin, the epidermis, which acts as a barrier to this type of water loss.

(b) Evaporation through the lungs. Inhaled air initially contains very little water vapor. However, as soon as it enters the respiratory passages, the air is exposed to the fluids covering the respiratory surfaces. By the time this air enters the lungs, it has become totally saturated with moisture from these surfaces. When the air is exhaled, it is still saturated with moisture and water is lost from the body.

b. Larger quantities of water are required when water loss is increased in any one of the following circumstances:

(1) Heat Exposure. When an individual is exposed to very high temperatures, water lost in the sweat can be increased to as much as 3.5 quarts an hour. Water loss at this increased rate can deplete the body fluids in short time.

(2) Exercise. Physical activity increases the loss of water in two ways as follows:

(a) The increased respiration rate causes increased water loss by evaporation through the lungs.

(b) The increased body heat causes excessive sweating.

(3) Cold Exposure. As the temperature decreases, the amount of water vapor in the air also decreases. Therefore, breathing cold air results in increased water loss by evaporation from the lungs.

(4) High Altitude. At high altitudes, increased water loss by evaporation through the lungs occurs not only as a result of breathing cooler air but also as a result of the increased respiratory efforts required.

(5) Burns. After extensive burns, the outermost layer of the skin is destroyed. When this layer is gone, there is no longer a barrier to water loss by diffusion, and the rate of water loss in this manner can increase up to 5 quarts each day.

(6) Illness. Severe vomiting or prolonged diarrhea can lead to serious water depletion.

c. Dehydration (body fluid depletion) can occur when required body fluids are not replaced.

(1) Dehydration is accompanied by the following symptoms:

(a) Thirst.
(b) Weakness.
(c) Fatigue.
(d) Dizziness.
(e) Headache.
(f) Fever.
(g) Inelastic abdominal skin.
(h) Dry mucous membranes, that is, dry mouth and nasal passages.
(i) Infrequent urination and reduced volume. The urine is concentrated so that it is very dark in color. In severe cases, urination may be quite painful.

(2) Companions will observe the following behavioral changes in individuals suffering from dehydration:

(a) Loss of appetite.
(b) Lagging pace.
(c) Impatience.
(d) Sleepiness.
(e) Apathy.
(f) Emotional instability.
(g) Indistinct speech.
(h) Mental confusion.

(3) Dehydration is a complication which causes decreased efficiency in the performance of even the simplest task. It also predisposes survivors to the development of severe shock following minor injuries. Constriction of blood vessels in the skin as a result of dehydration increases the danger of cold injury during

NO WALKING AT ALL

MAXIMUM DAILY TEMPERATURE (°F) IN SHADE	AVAILABLE WATER PER MAN, US QUARTS					
	0	1 Qt	2 Qts	4 Qts	10 Qts	20 Qts
	DAYS OF EXPECTED SURVIVAL					
120°	2	2	2	2.5	3	4.5
110	3	3	3.5	4	5	7
100	5	5.5	6	7	9.5	13.5
90	7	8	9	10.5	15	23
80	9	10	11	13	19	29
70	10	11	12	14	20.5	32
60	10	11	12	14	21	32
50	10	11	12	14.5	21	32

WALKING AT NIGHT UNTIL EXHAUSTED AND RESTING THEREAFTER

MAXIMUM DAILY TEMPERATURE (°F) IN SHADE	AVAILABLE WATER PER MAN, US QUARTS					
	0	1 Qt	2 Qts	4 Qts	10 Qts	20 Qts
	DAYS OF EXPECTED SURVIVAL					
120°	1	2	2	2.5	3	
110	2	2	2.5	3	3.5	
100	3	3.5	3.5	4.5	5.5	
90	5	5.5	5.5	6.5	8	
80	7	7.5	8	9.5	11.5	
70	7.5	8	9	10.5	13.5	
60	8	8.5	9	11	14	
50	8	8.5	9	11	14	

Figure 19-1. Water Requirements.

cold exposure. Failure to replace body fluids ultimately results in death.

(a) Proper treatment for dehydration is to replace lost body fluids. The oral intake of water is the most readily available means of correcting this deficiency. A severely dehydrated person will have little appetite. This person must be encouraged to drink small quantities of water at frequent intervals to replenish the body's fluid volume. Cold water should be warmed so the system will accept it easier.

(b) To prevent dehydration, water loss must be replaced by periodic intake of small quantities of water throughout the day. As activities or conditions intensify, the water intake should be increased accordingly. Water intake should be sufficient to maintain a minimum urinary output of 1 pint every 24 hours. Thirst is not an adequate stimulus for water intake, and a person often dehydrates when water is available. Therefore, water intake should be encouraged when the person is not thirsty. Humans cannot adjust to decreased water intake for prolonged periods of time. When water is in short supply, any available water should be consumed sensibly. If sugar is available, it should be mixed with the water, and efforts should be made to find a local water source. Until a suitable water source is located, individual water losses should be limited in the following ways:

-1. Physical activity should be limited to the absolute minimum required for survival activities. All tasks should be performed slowly and deliberately with minimal expenditure of energy. Frequent rest periods should be included in the daily schedule.

-2. In hot climates, essential activity should be conducted at night or during the cooler part of the day.

-3. In hot climates, clothing should be worn at all times because it reduces the quantity of water loss by sweating. Sweat is absorbed into the clothing evaporated from its surface in the same manner it evaporates from the body. This evaporation cools the air trapped between the clothing and the skin, causing a decrease in the activity of the sweat glands and a subsequent reduction in water loss.

-4. In hot weather, light-colored clothing should be worn rather than dark-colored clothing. Dark-colored clothing absorbs the Sun's light rays and converts them into heat. This heat causes an increase in body temperatures which activates the sweat glands and increases water loss through sweating. Light-colored clothing, however, reflects the Sun's light rays, minimizing the increase in body temperature and subsequent water loss.

19-3. Water Sources. Survivors should be aware of both the water sources available to them and the resources at their disposal for producing water.

a. Survivors may obtain water from solar stills, desalter kits, or canned water packed in various survival kits. It would be wise for personnel, who may one day have to use these methods of procuring water, to be knowledgeable of their operating instructions and the amount of water they produce.

(1) Canned water provides 10 ounces per can.

(2) Desalter kits are limited to 1 pint per chemical bar—kits contain eight chemical bars.

(3) A "sea solar still" can produce as much as 2½ pints per day.

(4) "Land solar stills" produce varied amounts of water. This amount is directly proportionate to the amount of water available in the soil or placed into the still (vegetation, entrails, contaminated water, etc.), and the ambient temperature.

b. Aircrew members would be wise to carry water during their missions. Besides the fact that the initial shock of the survival experience sometimes produces feelings of thirst, having an additional water container can benefit survivors. The issued items (canned water, desalter kits, and solar stills) should be kept by survivors for times when no natural sources of freshwater are available.

c. Naturally occurring indicators of water are:

(1) Surface water, including streams, lakes, springs, ice, and snow.

(2) Precipitation, such as rain, snow, dew, sleet, etc.

(3) Subsurface water, which may not be as readily accessible as wells, cisterns, and underground springs and streams, can be difficult for survivors to locate and use.

d. Several indicators of possible water are:

(1) Presence of abundant vegetation of a different variety, such as deciduous growth in a coniferous area.

(2) Drainages and low-lying areas.

(3) Large clumps of plush grass.

(4) Animal trails which may lead to water. The "V" formed by intersecting trails often point toward water sources.

e. Survivors may locate and procure water as follows:

(1) Precipitation may be procured by laying a piece of nonporous material such as a poncho, piece of canvas, plastic, or metal material on the ground. If rain or snow is being collected, it may be more efficient to create a bag or funnel shape with the material so the water can be easily gathered. Dew can be collected by wiping it up with a sponge or cloth first, and then wringing it into a container (figure 19-2). Consideration should be given to the possibility of contaminating the water with dyes, preservatives, or oils on the surfaces of the objects used to collect the precipitation. Ice will yield more water per given volume than snow and requires less heat to do so. If the Sun is shining, snow or ice may be placed on a dark surface to melt (dark surfaces absorb heat, whereas light surfaces reflect heat). Ice can be found in the form of icicles on plants and trees, sheet ice on rivers, ponds, and lakes, or sea ice. If snow must be used, survivors should use snow closest to the ground. This snow is packed and will provide more water for the amount of snow than will the upper layers. When snow is to be melted for water, place a small amount of snow in the bottom of the container being used and place it over or near a fire. Snow can be added a little at a time. Survivors should allow water in the container bottom to become warm so that when more snow is added, the mixture remains slushy. This will prevent burning the bottom out of the container. Snow absorbs water, and if packed, forms an insulating airspace at the bottom of the container. When this happens, the bottom may burn out.

(2) Several things may help survivors locate ground water, such as rivers, lakes, and streams.

(a) The presence of swarming insects indicates water is near. In some places, survivors should look for signs of animal presence. For example, in damp places, animals may have scratched depressions into the ground to obtain water; insects may also hover over these areas.

(b) In the Libyan Sahara, donut-shaped mounds of camel dung often surround wells or other water sources. Bird flights can indicate direction to or from water. Pigeons and doves make their way to water regularly. They fly from water in the morning and to it in the evening. Large flocks of birds may also congregate around or at areas of water.

(c) The presence of people will indicate water. The location of this water can take many forms—stored water in containers that are carried with people who are traveling, wells, irrigation systems, pools, etc. Survivors who are evaders should be extremely cautious when approaching any water source, especially if they are in dry areas; these places may be guarded or inhabited.

(3) When no surface water is available, survivors may have to tap the Earth's supply of ground water. Access to this depends upon the type of ground—rock or loose material, clay, gravel, or sand.

(a) In rocky ground, survivors should look for springs and seepages. Limestone and lava rocks will have more and larger springs than any other rocks. Most

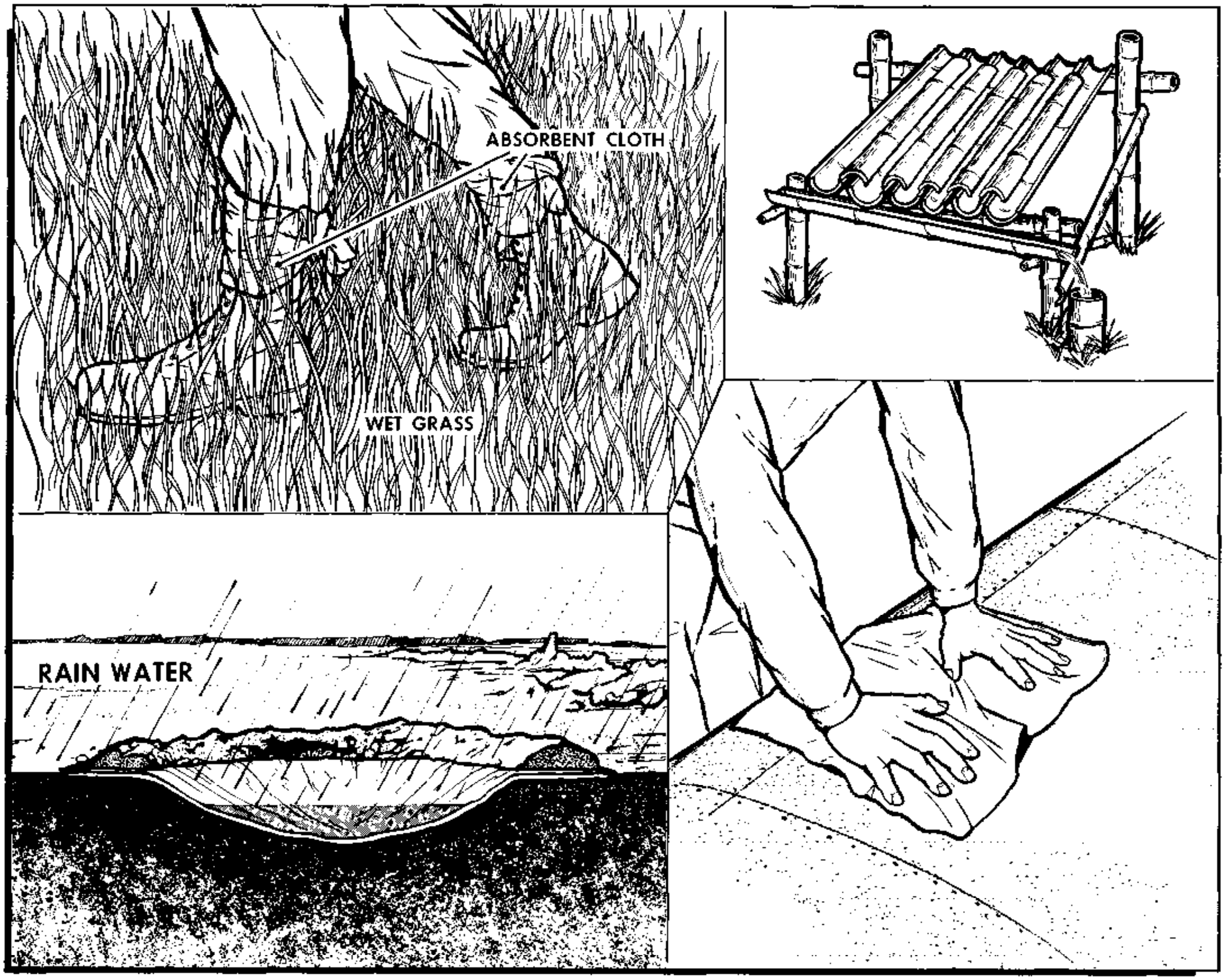

Figure 19-2. Methods of Procuring Water.

lava rocks contain millions of bubble holes; ground water may seep through them. Survivors can also look for springs along the walls of valleys that cross a lava flow. Some flows will have no bubbles but do have "organ pipe" joints—vertical cracks that part the rocks into columns a foot or more thick and 20 feet or more high. At the foot of these joints, survivors may find water creeping out as seepage, or pouring out in springs.

(b) Most common rocks, like granite, contain water only in irregular cracks. A crack in a rock with bird dung around the outside may indicate a water source that can be reached by a piece of surgical hose used as a straw or siphon.

(c) Water is more abundant and easier to find in loose sediments than in rocks. Springs are sometimes found along valley floors or down along their sloping sides. The flat benches or terraces of land above river valleys usually yield springs or seepages along their bases, even when the stream is dry. Survivors shouldn't waste time digging for water unless there are signs that water is available. Digging in the floor of a valley under a steep slope, especially if the bluff is cut in a terrace, can produce a water source. A lush green spot where a spring has been during the wet season is a good place to dig for water. Water moves slowly through clay, but many clays contain strips of sand which may yield springs. Survivors should look for a wet place on the surface of clay bluffs and try digging it out.

(d) Along coasts, water may be found by digging beach wells (figure 19-3). Locate the wells behind the first or second pressure ridge. Wells can be dug 3 to 5 feet deep and should be lined with driftwood to prevent sand from refilling the hole. Rocks should be used to line the bottom of the well to prevent stirring up sand when procuring the water. The average well may take as long as 2 hours to produce 4 to 5 gallons of water. (Do not be discouraged if the first try is unsuccessful—dig another.)

19-4. Water in Snow and Ice Areas. Due to the extreme cold of arctic areas, water requirements are greatly increased. Increased body metabolism, respiration of cold air, and extremely low humidity play important roles in reducing the body's water content. The processes of heat production and digestion in the body also increase the need for water in colder climatic zones. The constructing of shelters and signals and the obtaining of firewood are extremely demanding tasks for survivors. Physical exertion and heat production in extreme cold place the water requirements of a survivor close to 5 or 6 quarts per day to maintain proper hydration levels. The diet of survivors will often be dehydrated rations and high protein food sources. For the body to digest and use these food sources effectively, increased water intake is essential.

a. Obtaining water need not be a serious problem in the arctic because an abundant supply of water is available from streams, lakes, ponds, snow, and ice. All surface water should be purified by some means. In the summer, surface water may be discolored but is drinkable when purified. Water obtained from glacier-fed rivers and streams may contain high concentrations of dirt or silt. By letting the water stand for a period of time, most silt will settle to the bottom; the remaining water can be strained through porous material for further filtration.

b. A "water machine" can be constructed which will produce water while the survivors are doing other tasks. It can be made by placing snow on any porous material (such as parachute or cotton), gathering up the edges, and suspending the "bag" of snow from any support near the fire. Radiant heat will melt the snow and the water will drip from the lowest point on the bag. A container should be placed below this point to catch the water (figure 19-4).

c. In some arctic areas, there may be little or no fuel supply with which to melt ice and snow for water. In this case, body heat can be used to do the job. The ice or snow can be placed in a waterproof container like a waterbag and placed between clothing layers next to the body. This cold substance should not be placed directly

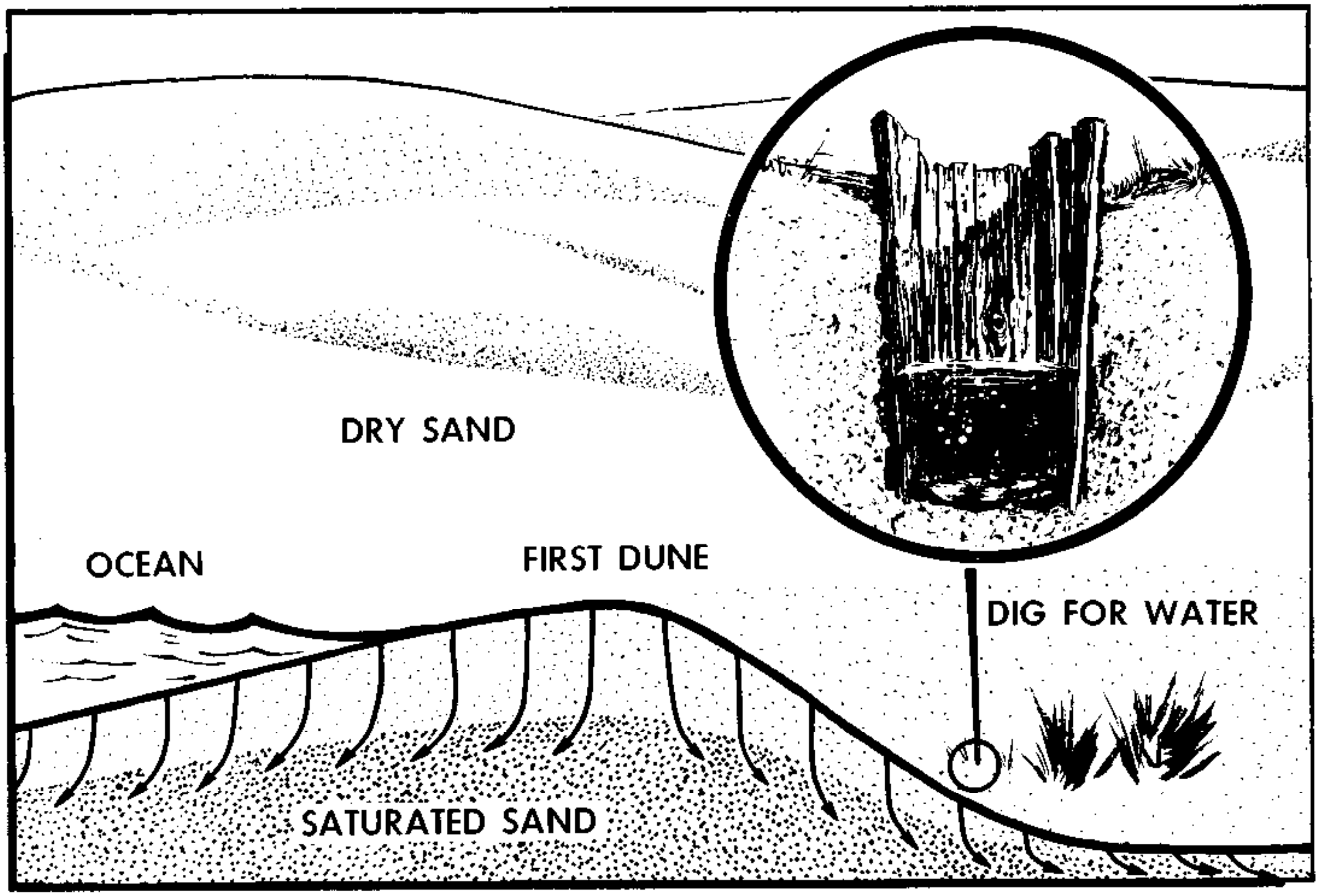

Figure 19-3. Beach Well.

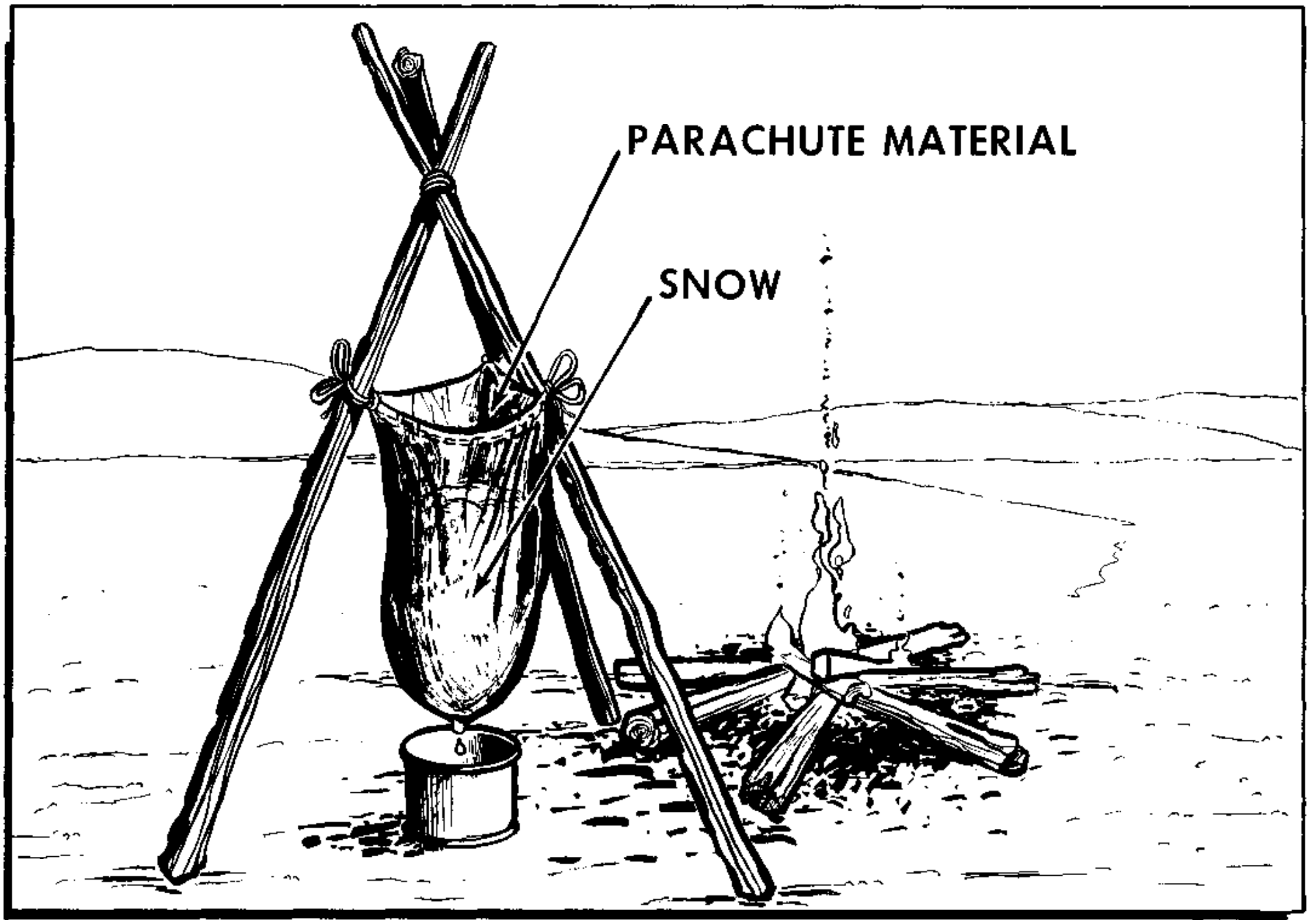

Figure 19-4. Water Machine.

next to the skin; it causes chilling and lowering of the body temperature.

d. Since icebergs are composed of freshwater, they can be a readily available source of drinking water. Survivors should use extreme caution when trying to obtain water from this source. Even large icebergs can suddenly roll over and dump survivors into the frigid sea water. If sea ice is the primary source of water, survivors should recall that like seawater itself, saltwater ice should never be ingested. To obtain water in polar regions or sea ice areas, survivors should select old sea ice, a bluish or blackish ice which shatters easily and generally has rounded corners. This ice will be almost salt-free. New sea ice is milky or gray colored with sharp edges and angles. This type of ice will not shatter or break easily. Snow and ice may be saturated with salt from blowing spray; if it tastes salty, survivors should select different snow or ice sources.

e. The ingesting of unmelted snow or ice is not recommended. Eating snow or ice lowers the body's temperature, induces dehydration, and causes minor cold injury to lips and mouth membranes. Water consumed in cold areas should be in the form of warm or hot fluids. The ingestion of cold fluids or foods increases the body's need for water and requires more body heat to warm the substance.

19-5. Water on the Open Seas. The lack of drinkable water could be a major problem on the open seas. Seawater should never be ingested in its natural state. It will cause an individual to become violently ill in a very short period of time. When water is limited and cannot be replaced by chemical or mechanical means, it must be used efficiently. As in the desert, conserving sweat not water, is the rule. Survivors should keep in the shade as much as possible and dampen clothing with seawater to keep cool. They should not over exert but relax and sleep as much as possible.

a. If it rains, survivors can collect rainwater in available containers and store it for later use. Storage containers could be cans, plastic bags, or the bladder of a life preserver. Drinking as much rainwater as possible while it is raining is advisable. If the freshwater should become contaminated with small amounts of seawater or salt spray, it will remain safe for drinking (figure 19-5). At night and on foggy days, survivors should try to collect dew for drinking water by using a sponge, chamois, handkerchief, etc.

b. Solar stills will provide a drinkable source of water. Survivors should read the instructions immediately and set them up, using as many stills as available. (Be sure to attach them to the raft.) Desalter kits, if available, should probably be saved for the time when no other means of procuring drinking water is available. Instructions on how to use the desalter kit are on the container.

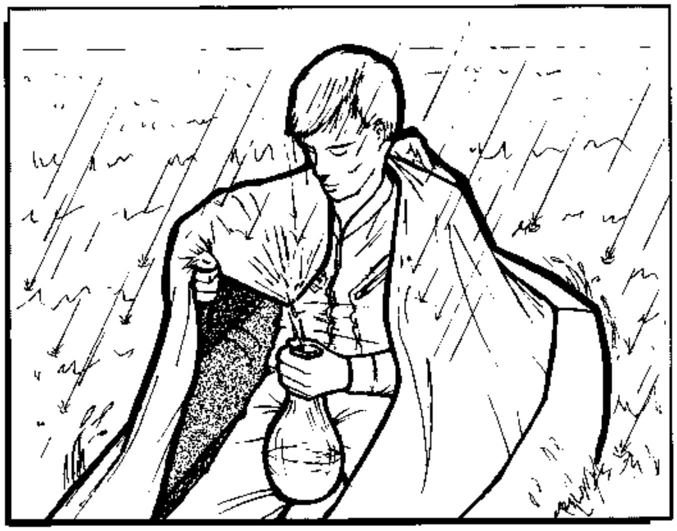

Figure 19-5. Collecting Water from Spray Shield.

c. Only water in its conventional sense should be consumed. The so-called "water substitutes" do little for the survivor, and may do much more harm than not consuming any water at all. There is no substitute for water. Fish juices and other animal fluids are of doubtful value in preventing dehyration. Fish juices contain protein which requires large amounts of water to be digested and the waste products must be excreted in the urine which increases water loss. Survivors should never drink urine—urine is body waste material and only serves to concentrate waste materials in the body and require more water to eliminate the additional waste.

19-6. Water in Tropical Area. Depending on the time of the year and type of jungle, water in the tropical climates can be plentiful; however, it is necessary to know where to look and procure it. Surface water is normally available in the form of streams, ponds, rivers, and swamps. In the savannas during the dry season, it may be necessary for the survivor to resort to digging for water in the places previously mentioned. Water obtained from these sources may need filtration and should be purified. Jungle plants can also provide survivors with water.

a. Many plants have hollow portions which can collect rainfall, dew, etc. (figure 19-6). Since there is no absolute way to tell whether this water is pure, it should be purified. The stems or the leaves of some plants have a hollow section where the stem meets the trunk. Look for water collected here. This includes any Y-shaped plants (palms or air plants). The branches of large trees often support air plants (relatives of the pineapple) whose overlapping, thickly growing leaves may hold a considerable amount of rainwater. Trees may also catch and store rainwater in natural receptacles such as cracks or hollows.

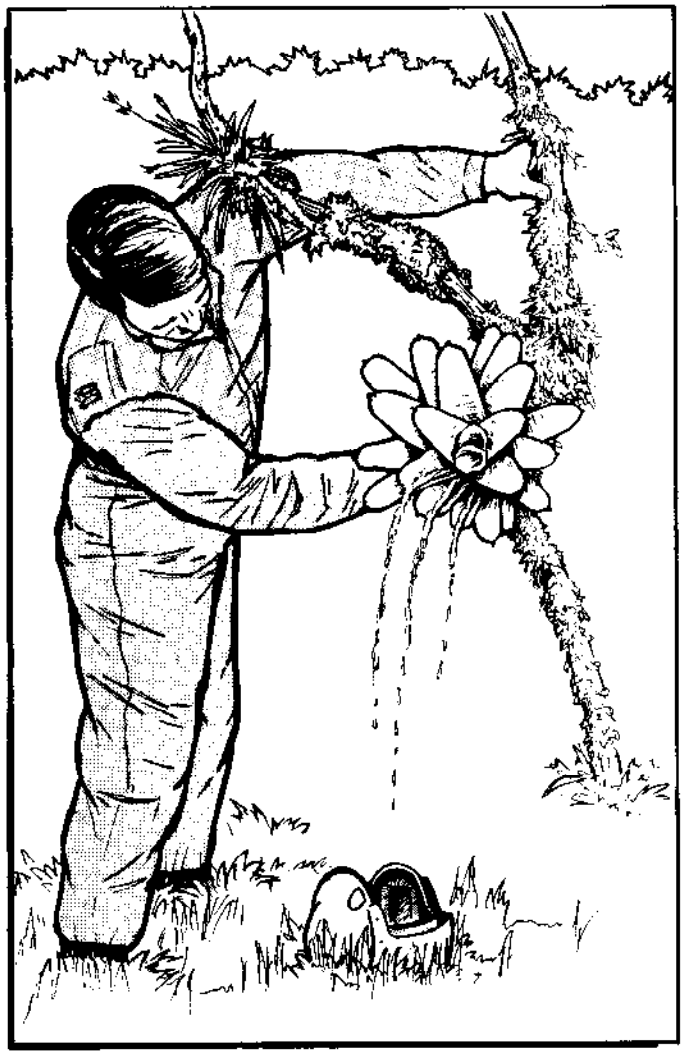

Figure 19-6. Water Collectors.

b. Pure freshwater needing no purification can be obtained from numerous plant sources. There are many varieties of vines which are potential water sources. The vines are from 50 feet to several hundred feet in length and 1 to 6 inches in diameter. They also grow like a hose along the ground and up into the trees. The leaf structure of the vine is generally high in the trees. Water

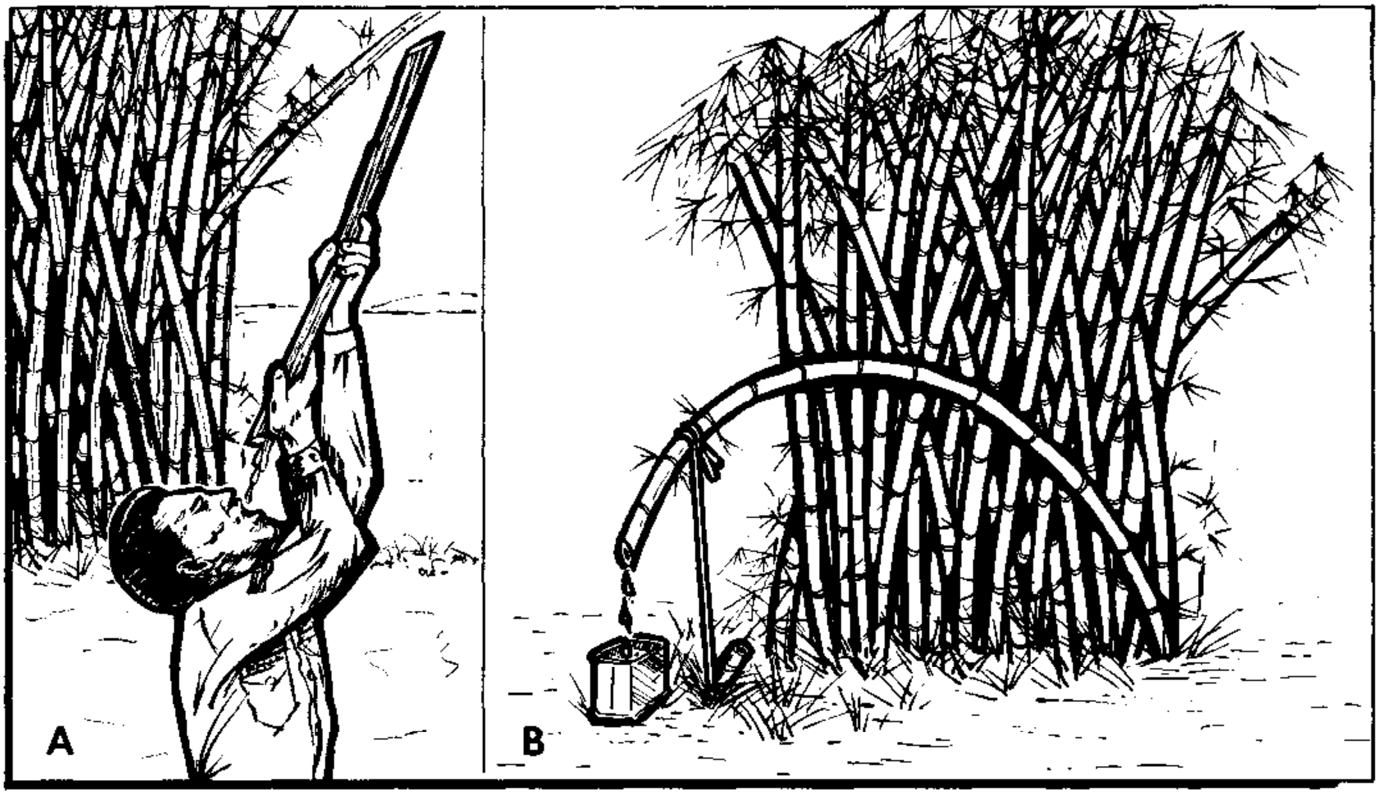

Figure 19-7. Water Vines and Bamboo.

vines are usually soft and easily cut. The smaller species may be twisted or bent easily and are usually heavy because of the water content. The water from these vines should be tested for potability. The first step in testing the water from vines is for survivors to nick the vine and watch for sap running from the cut. If milky sap is seen, the vine should be discarded; if no milky sap is observed, the vine may be a safe water vine. Survivors should cut out a section of the vine, hold that piece vertically, and observe the liquid as it flows out. If it is clear and colorless, it may be a drinkable source. If it is cloudy or milky-colored, they should discard the vine. They should let some of the liquid flow into the palm of the hand and observe it. If the liquid does not change color, they can now taste it. If it tastes like water or has a woody or sweet taste, it should be safe for drinking. Liquid with a sour or bitter taste should be avoided. Water trapped within a vine is easily obtained by cutting out a section of the vine. The vine should first be cut high above the ground and then near the ground. This will provide a long length of vine and, in addition, will tend to hide evidence of the cuts if the survivors are in an evasion situation. When drinking from the vine, it should not touch the mouth as the bark may contain irritants which could affect the lips and mouth (figure 19-7). The pores in the upper end of the section of vine may reclose, stopping the flow of water. If this occurs, survivors should cut off the end of the vine opposite the drinking end. This will reopen the pores allowing the water to flow.

c. Water from the rattan palm and spiny bamboo may be obtained in the same manner as from vines. It is not necessary to test the water if positive identification of the plant can be made. The slender stem (runner) of the rattan palm is an excellent water source. The joints are overlapping in appearance, as if one section is fitted inside the next.

d. Water may be trapped within sections of green bamboo. To determine if water is trapped within a section of bamboo, it should be shaken. If it contains water, a sloshing sound can be heard. An opening may be made in the section by making two 45-degree angle cuts, both on the same side of the section, and prying loose a piece of the section wall. The end of the section may be cut off and the water drunk or poured from the open end. The inside of the bamboo should be examined before consuming the water. If the inside walls are clean and white, the water will be safe to drink. If there are brown or black spots, fungus growth, or any discoloration, the water should be purified before consumption. Sometimes water can also be obtained by cutting the top off certain types of green bamboo, bending it over, and staking it to the ground (figure 19-7). A water container should be placed under it to catch the dripping water. This method has also proven effective on some vines and the rattan palm.

e. Water can also be obtained from banana plants in a couple of different ways, neither of which is satisfactory in a tactical situation. First, survivors should cut a banana plant down, then a long section should be cut off which can be easily handled. The section is taken apart by slitting from one end to the other and pulling off the layers one at a time. A strip 3 inches wide, the length of the section, and just deep enough to expose the cells should be removed from the convex side. This section is folded toward the convex side to force the water from the cells of the plant. The layer must be squeezed gently to avoid forcing out any tannin into the water. Another technique for obtaining water from the banana plant is by making a "banana-well." This is done by making a bowl out of the plant stump, fairly close to the ground, by cutting out and removing the inner section of the stump (figure 19-8). Water which first enters the bowl may contain a concentration of tannin (an astringent which has the same effect as alum). A leaf from the banana plant or other plant should be placed over the bowl while it is filling to prevent contamination by insects, etc.

f. Water trees can also be a valuable source of water in some jungles. They can be identified by their blotched bark which is fairly thin and smooth. The leaves are large, leathery, fuzzy, and evergreen, and may grow as large as 8 or 9 inches. The trunks may have short outgrowths with fig-like fruit on them or long tendrils with round fruit comprised of corn kernel-shaped nuggets. In a nontactical situation, the tree can be tapped in the same manner as a rubber tree, with either a diagonal slash or a "V." When the bark is cut into, it will exude a white sap which if ingested causes temporary irritation of the urinary tract. This sap dries up quite rapidly and can easily be removed. The cut should be continued into the tree with a spigot (bamboo, knife, etc.) at the bottom of the tap to direct the water into a container. The water flows from the leaves back into the roots after sundown, so water can be procured from this source only after sundown or on overcast (cloudy) days. If survivors are in a tactical situation, they can obtain water from the tree and still conceal the procurement location. If the long tendrils are growing thickly, they can be separated and a hole bored into the tree. The white sap should be scraped off and a spigot placed below the tap with a water container to catch the water. Moving the tendrils back into place will conceal the container. Instead of boring into the tree, a couple of tendrils can be cut off or snapped off if no knife is available. The white sap should be allowed to dry and then be removed. The ends of the tendrils should be placed in a water container and the container concealed.

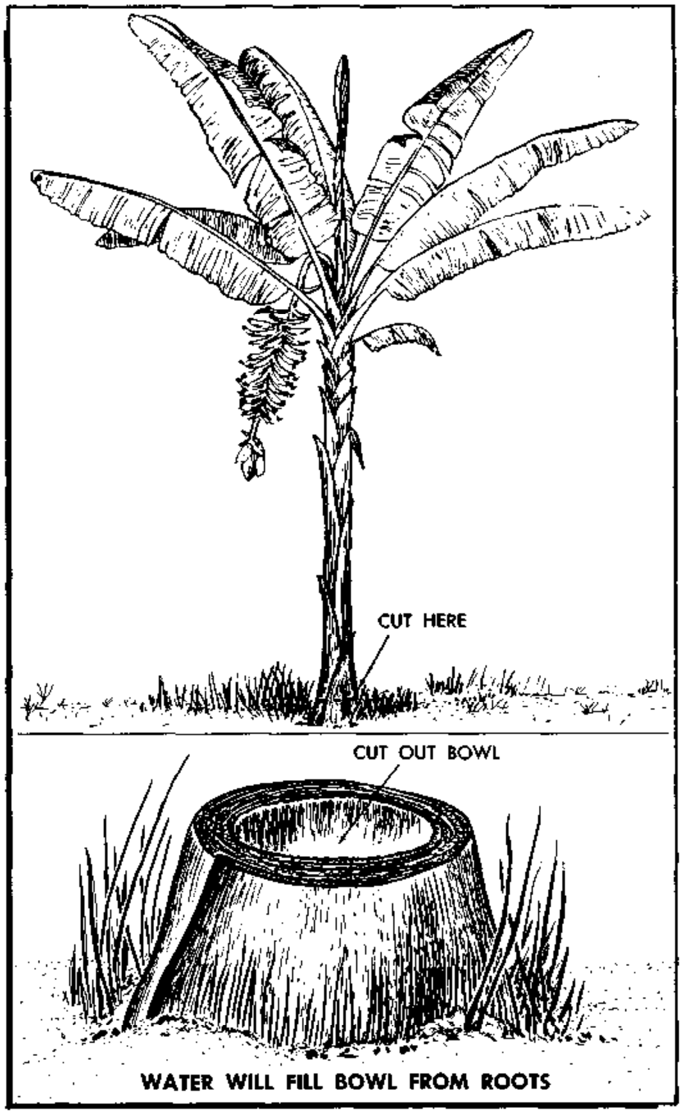
CUT HERE
CUT OUT BOWL
WATER WILL FILL BOWL FROM ROOTS

Figure 19-8. Water from Banana Plant.

g. Coconuts contain a refreshing fluid. Where coconuts are available, they may be used as a water source. The fluid from a mature coconut contains oil, which when consumed in excess can cause diarrhea. There is little problem if used in moderation or with a meal and not on an empty stomach. Green unripe coconuts about the size of a grapefruit are the best for use because the fluid can be taken in large quantities without harmful effects. There is more fluid and less oils so there is less possibility of diarrhea.

h. Water can also be obtained from liquid mud. Mud can be filtered through a piece of cloth. Water taken by this method must be purified. Rainwater can be collected from a tree by wrapping a cloth around a slanted tree and arranging the bottom end of the cloth to drip into a container (figure 19-9).

19-7. Water in Dry Areas. Locating and procuring water in a dry environment can be a formidable task. Some of the ways to find water in this environment have been

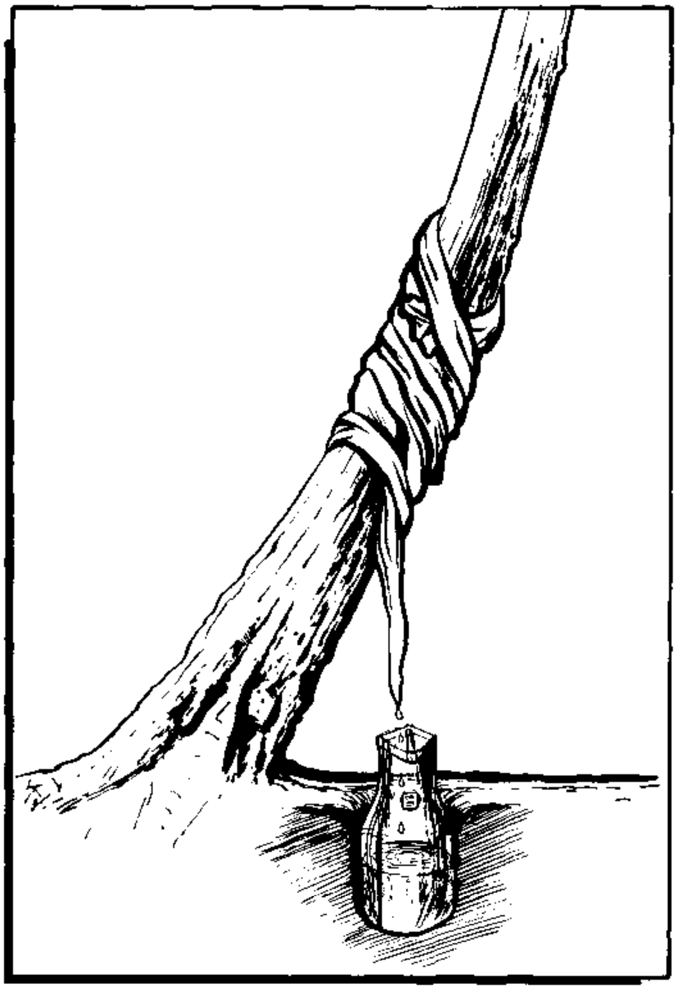

Figure 19-9. Collecting Water from Slanted Tree.

explored, such as locating a concave bend in a dry riverbed and digging for water (figure 19-10). If there is any water within a few feet of the surface, the sand will become slightly damp. Dig until water is obtained.

a. Some deserts become humid at night. The humidity may be collected in the form of dew. This dew can be collected by digging a shallow basin in the ground about 3 feet in diameter and lining it with a piece of canvas, plastic, or other suitable material. A pyramid of stones taken from a minimum of 1 foot below the surface should then be built in this basin. Dew will collect on and between the stones and trickle down onto the lining material where it can be collected and placed in a container.

b. Plants and trees having roots near the surface may be a source of water in dry areas. Water trees of dry Australia are a source of water, as their roots run out 40 to 80 feet at a depth of 2 to 9 inches under the surface. Survivors may obtain water from these roots by locating a root 4 to 5 feet from the trunk and cutting the root into 2- or 3-foot lengths. The bark can then be peeled off and the liquid from each section of root drained into a container. The liquid can also be sucked out. The trees growing in hollows or depressions will have the most water in their roots. Roots that are 1 to 2 inches thick are an ideal size. Water can be carried in these roots by plugging one end with clay.

c. Cactus-like or succulent plants may be sources of water for survivors, but they should recall that no plants should be used for water procurement which have a milky sap. The barrel cactus of the United States provides a water source. To obtain it, survivors should first cut off the top of the plant. The pulpy inside portions of the plant should then be mashed to form a watery pulp. Water may ooze out and collect in the bowl; if not, the pulp may be squeezed through a cloth directly into the mouth.

d. The solar still is a method of obtaining water that uses both vegetation and ground moisture to produce water (figure 19-11). A solar still can be made from a sheet of clear plastic stretched over a hole in the ground. The moisture in the soil and from plant parts (fleshy stems and leaves) will be extracted and collected by this emergency device. Obviously, where the soil is extremely dry and no fleshy plants are available, little, if any, water can be obtained from the still. The still may also be used to purify polluted water.

(1) The parts for the still are a piece of plastic about 6 feet square, a water collector-container or any waterproof material from which a collector-container can be fashioned, and a piece of plastic tubing about one-fourth inch in diameter and 4 to 6 feet long. The tubing is not absolutely essential but makes the still easier to use. A container can be made from such materials as plastic, aluminum foil, poncho, emergency ration tins, or a flight helmet. The tubing, when available, is fastened to the bottom of the inside of the container and used to remove drinking water from the container without disturbing the plastic. Some plastics work better than others, although any clear plastic should work if it is strong.

(2) If plants are available or if polluted water is to be purified, the still can be constructed in any convenient spot where it will receive direct sunlight throughout the day. Ease of digging will be the main consideration. If soil moisture is to be the only source of water, some sites will be better than others. Although sand generally does not retain as much moisture as clay, a wet sand will work very well. Along the seacoast or in any inland areas where brackish or polluted water is available, any wet soil, even sand, produces usable amounts of water. On cloudy days, the yield will be reduced because direct sunlight is necessary if the still is to operate at full efficiency.

(3) Certain precautions must be kept in mind. If polluted water is used, survivors should make sure that none is spilled near the rim of the hole where the plastic touches the soil and that none comes in contact with the

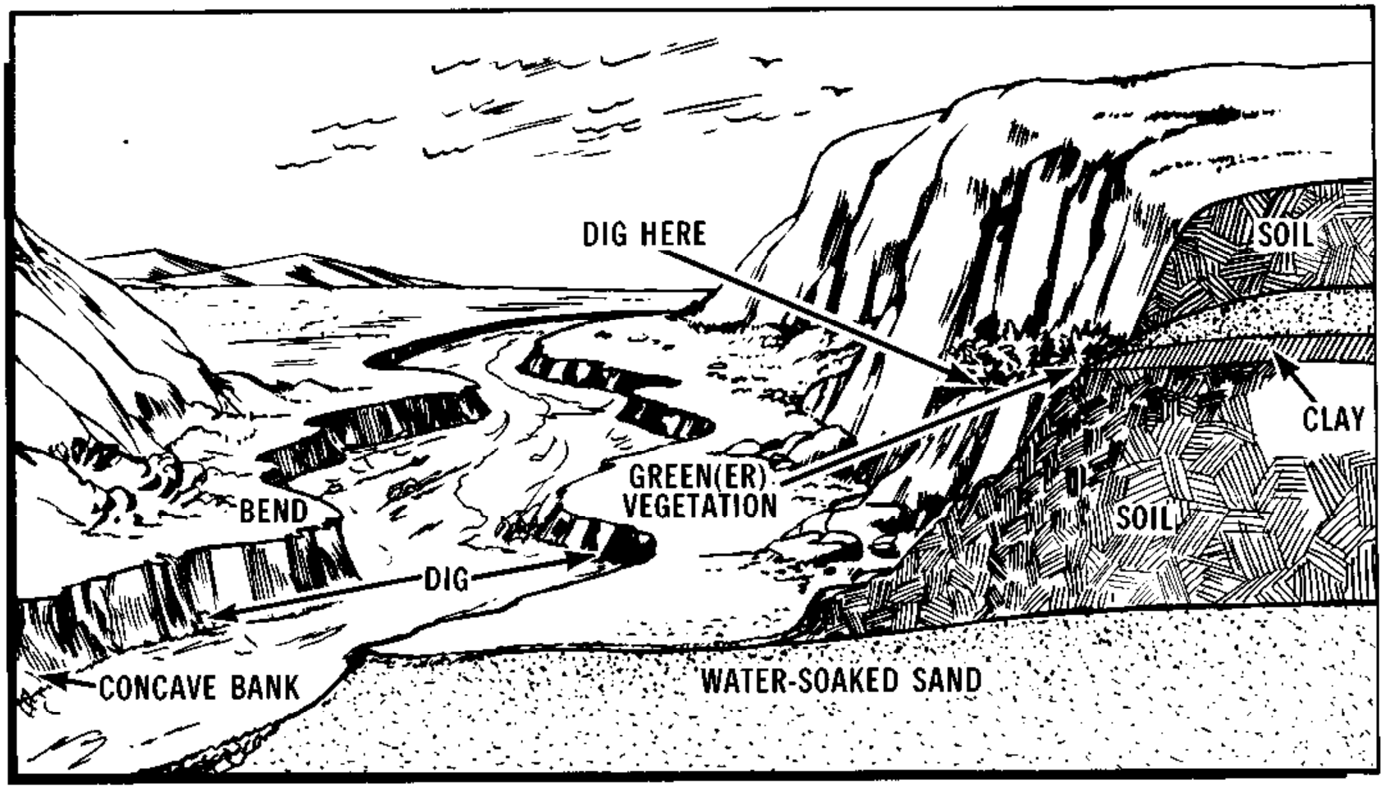

Figure 19-10. Dry Stream Bed.

container to prevent the freshly distilled water becoming contaminated. Survivors should not disturb the plastic sheet during daylight "working hours" unless it is absolutely necessary. If a plastic drinking tube is not available, raise the plastic sheet and remove the container as few times as possible during daylight hours. It takes one-half hour for the air in the still to become resaturated and the collection of water to begin after the plastic has been disturbed. Even when placed on fairly damp soil and in an area where 8 hours of light per day is directed on the solar still, the average yield is only about 1 cup per day per still. Due to the low yields obtained from this device, survivors must give consideration to the possible danger of excessive dehydration brought about by constructing the solar still. In certain circumstances, solar still returns, even over 2- or 3-day periods, will not equal the amount of body fluid lost in construction and will actually hasten dehydration.

(4) Steps survivors should follow when constructing a solar still are: Dig a bowl-shaped hole in the soil about 40 inches in diameter and 20 inches deep. Add a smaller, deeper sump in the center bottom of the hole to accommodate the container. If polluted waters are to be purified, a small trough can be dug around the side of the hole about halfway down from the top. The trough ensures that the soil wetted by the polluted water will be exposed to the sunlight and at the same time that the polluted water is prevented from running into the container. If plant material is used, line the sides of the hole with pieces of plant or its fleshy stems and leaves. Place the plastic over the hole and put soil on the edges to hold it in place. Place a rock no larger than a plum in the center of the plastic until it is about 15 inches below ground level. The plastic will now have the shape of a cone. Put more soil on the plastic around the rim of the hole to hold the cone securely in place and to prevent water-vapor loss. Straighten the plastic to form a neat cone with an angle of about 30 degrees so the water drops will run down and fall into the container. It takes about 1 hour for the air to become saturated and start condensing on the underside of the plastic cone.

e. The vegetation bag is a simpler method of water procurement. This method involves cutting foliage from trees or herbaceous plants, sealing it in a large clear plastic bag, and allowing the heat of the Sun to extract the fluids contained within. A large, heavy-duty clear plastic bag should be used. The bag should be filled with about 1 cubic yard of foliage, sealed, and exposed to the Sun. The average yield for one bag tested was 320 ml/bag 5-hour day. This method is simple to set up. The vegetation bag method of water procurement does have one primary drawback. The water produced is normally bitter to taste, caused by biological breakdown of the leaves as they lay in the water produced and super-

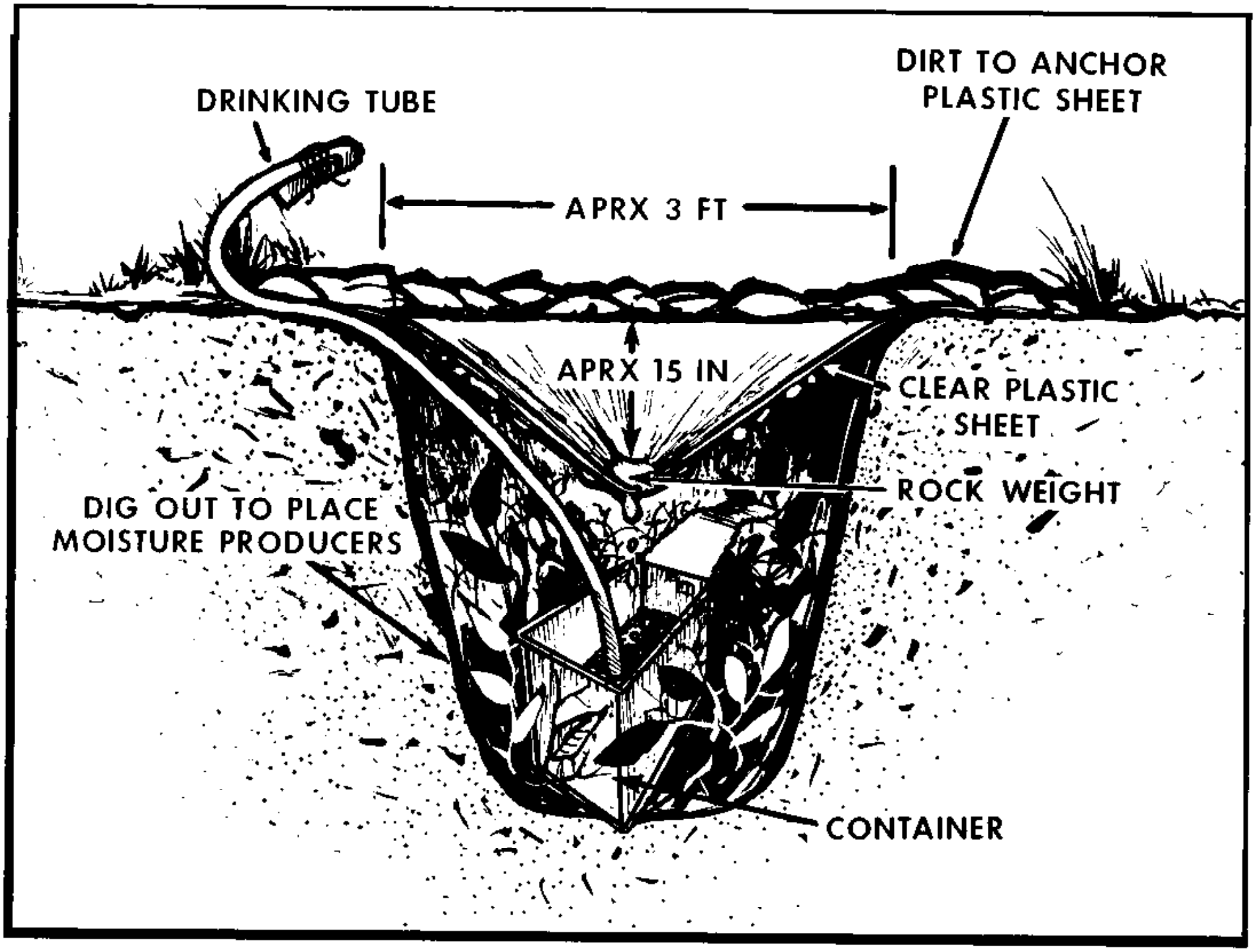

Figure 19-11. Solar Still.

heated in the moist "hothouse" environment. This method can be readily used in a survival situation, but before the water produced by certain vegetation is consumed, it should undergo the taste test. This is to guard against ingestion of cyanide-producing substances and other harmful toxins, such as plant alkaloids. (See figure 19-12.)

f. One more method of water procurement is the water transpiration bag, a method that is simple to use and has great potential for enhancing survival. This method is the vegetation bag process taken one step further. A large plastic bag is placed over a living limb of a medium-size tree or large shrub. The bag opening is sealed at the branch, and the limb is then tied down to allow collected water to flow to the corner of the bag. For a diagram of the water transpiration method, see figure 19-13.

(1) The amount of water yielded by this method will depend on the species of trees and shrubs available. During one test of this method, a transpiration bag produced approximately a gallon per day for 3 days with a plastic bag on the same limb, and with no major deterioration of the branch. Other branches yielded the same amount. Transpired water has a variety of tastes depending on whether or not the vegetation species is allowed to contact the water.

(2) The effort expended in setting up water transpiration collectors is minimal. It takes about 5 minutes' work and requires no special skills once the method has been described or demonstrated. Collecting the water in a survival situation would necessitate survivors dismantling the plastic bag at the end of the day, draining the contents and setting it up again the following day. The same branch may be reused (in some cases with almost similar yields); however, as a general rule, when vegetation abounds, a new branch should be used each day.

(3) Without a doubt, the water transpiration bag method surpasses other methods (solar stills, vegetation bag, cutting roots, barrel cactus) in yield, ease of assembly, and in most cases, taste. The benefits of having a simple plastic bag can't be over-emphasized. As a water procurer, in dry, semi-dry, or desert environments

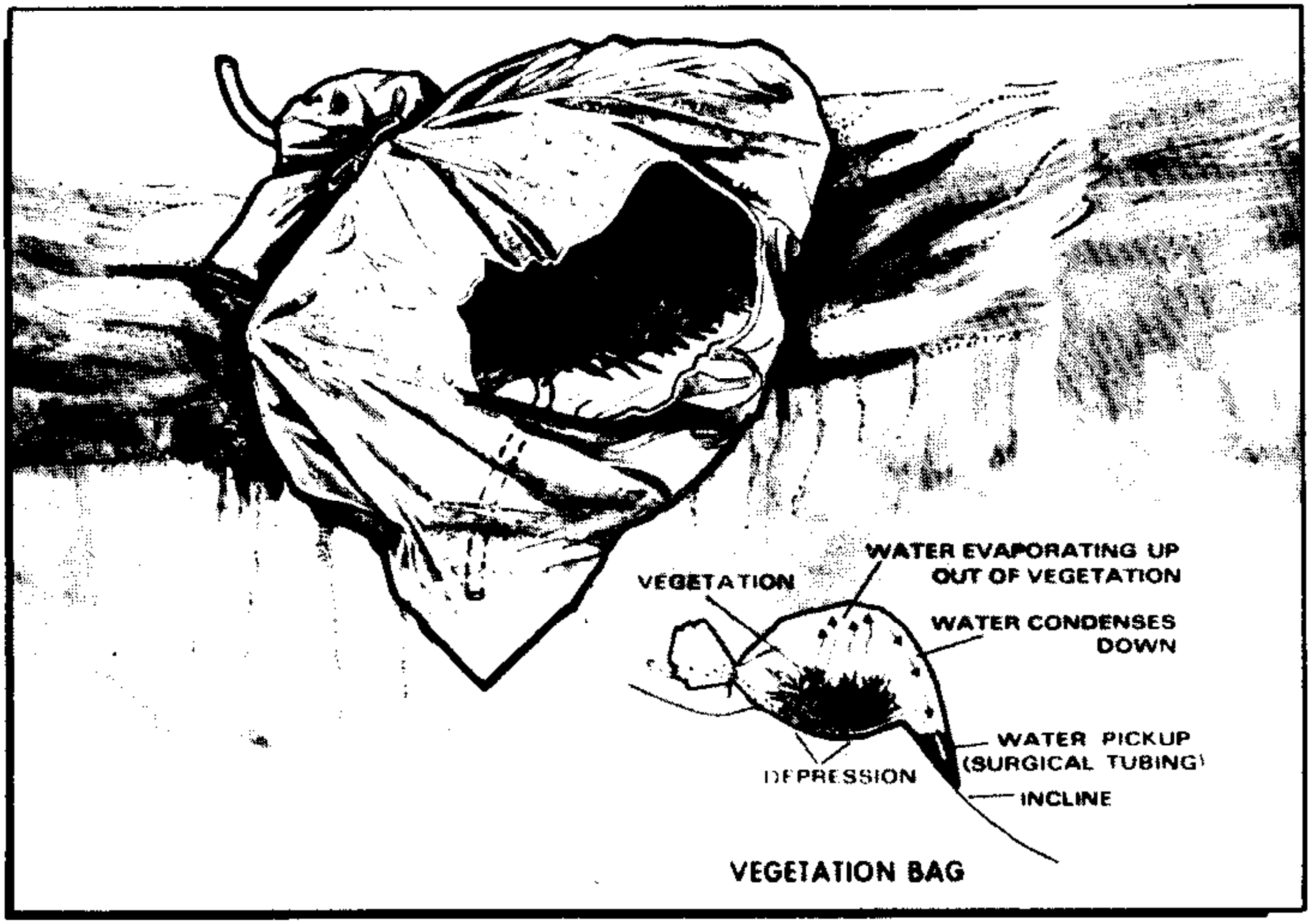
VEGETATION
WATER EVAPORATING UP OUT OF VEGETATION
WATER CONDENSES DOWN
WATER PICKUP (SURGICAL TUBING)
DEPRESSION
INCLINE
VEGETATION BAG

Figure 19-12. Vegetation Bag.

where low woodlands predominate, it can be used as a water transpirator; in scrubland, steppes, or treeless plains, as a vegetation bag; in sandy areas without vegetation, it can be cut up and improvised into solar stills. Up to three large, heavy-duty bags may be needed to sustain one survivor in certain situations.

19-8. Preparation of Water for Consumption:

a. The following are ways survivors can possibly determine the presence of harmful agents in the water:

(1) Strong odors, foam, or bubbles in the water.

(2) Discoloration or turbid (muddy with sediment).

(3) Water from lakes found in desert areas are sometimes salty because they have been without an outlet for extended periods of time. Magnesium or alkali salts may produce a laxative effect; if not too strong, it is drinkable.

(4) If the water gags survivors or causes gastric disturbances, drinking should be discontinued.

(5) The lack of healthy green plants growing around any water source.

b. Because of survivors' potential aversion to water from natural sources, it should be rendered as potable as possible through filtration. Filtration only removes the solid particles from water—it does not purify it. One simple and quick way of filtering is to dig a sediment hole or seepage basin along a water source and allow the soil to filter the water (figure 19-14). The seepage hole should be covered while not in use. Another way is to construct a filter—layers of parachute material stretched across a tripod (figure 19-15). Charcoal is used to eliminate bad odors and foreign materials from the water. Activated charcoal (obtained from freshly burned wood is used to filter the water). If a solid container is available for making a filter, use layers of fine-to-coarse sand and gravel along with charcoal and grass.

c. Purification of water may be done a variety of ways. The method used will be dictated by the situation (such as tactical or nontactical).

(1) Boil the water for at least 10 minutes.

(2) To use purification tablets survivors should follow instructions on the bottle. One tablet per quart of clear water; two tablets if water is cloudy. Let water

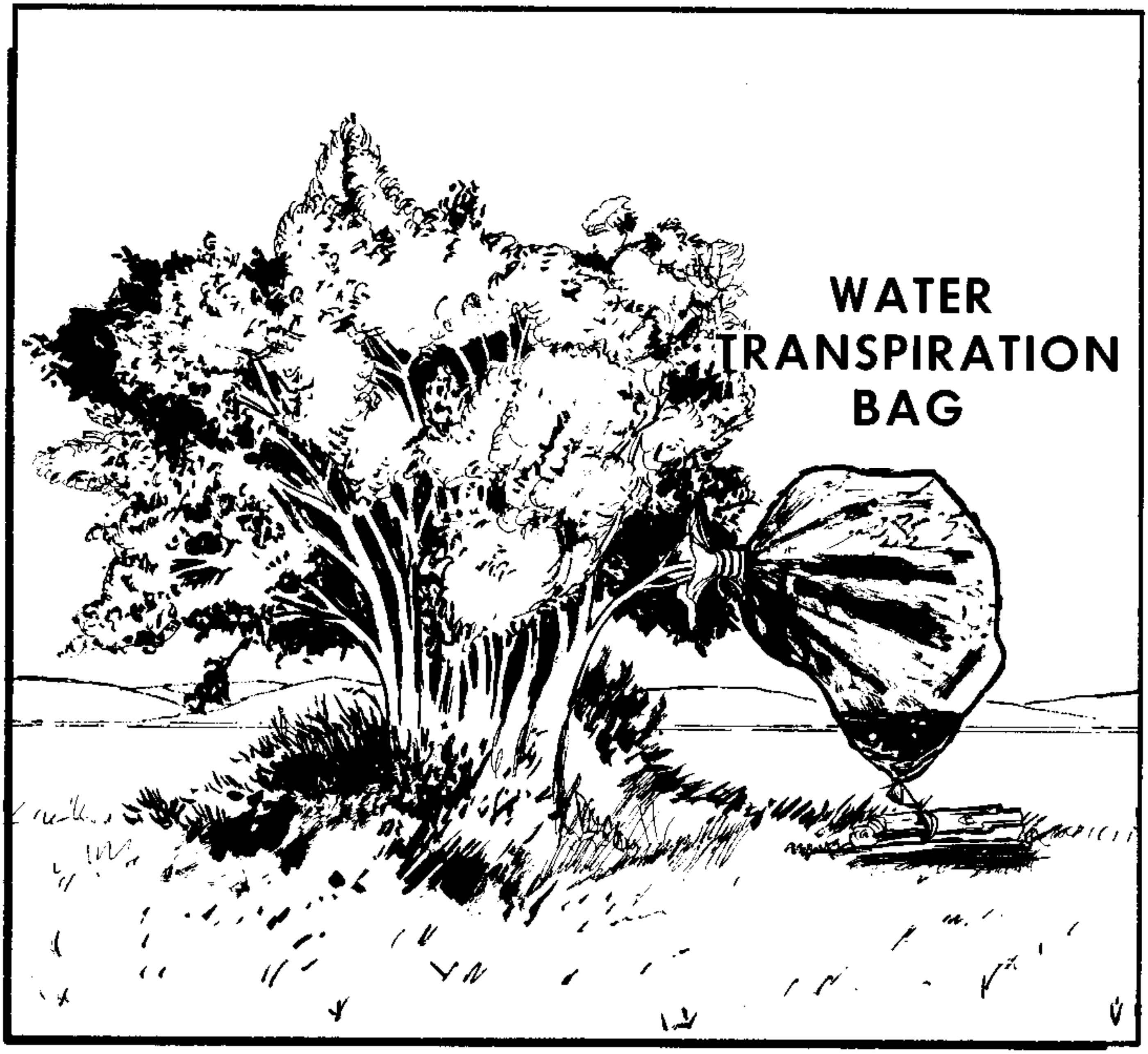

Figure 19-13. Transpiration Bag.

stand for 5 minutes (allowing the tablet time to dissolve), then shake and allow to stand for 15 minutes. Survivors should remember to turn the canteen over and allow a small amount of water to seep out and cover the neck part of the canteen. In an evasion situation, water purification tablets should be used for purifying water. If these are not available, plant sources or non-stagnant, running water obtained from a location upstream from habitatio should be consumed.

(3) Eight drops of 2½-percent iodine per quart—stir or shake and let stand for at least 10 minutes.

d. After water is found and purified, survivors may wish to store it for later consumption. The following make good containers:

(1) Waterbag.

(2) Canteen.

(3) Prophylactic inside a sock for protection of bladder.

(4) Segment of bamboo.

(5) Birch bark and pitch canteen.

(6) LPU bladder.

(7) Hood from antiexposure suit.

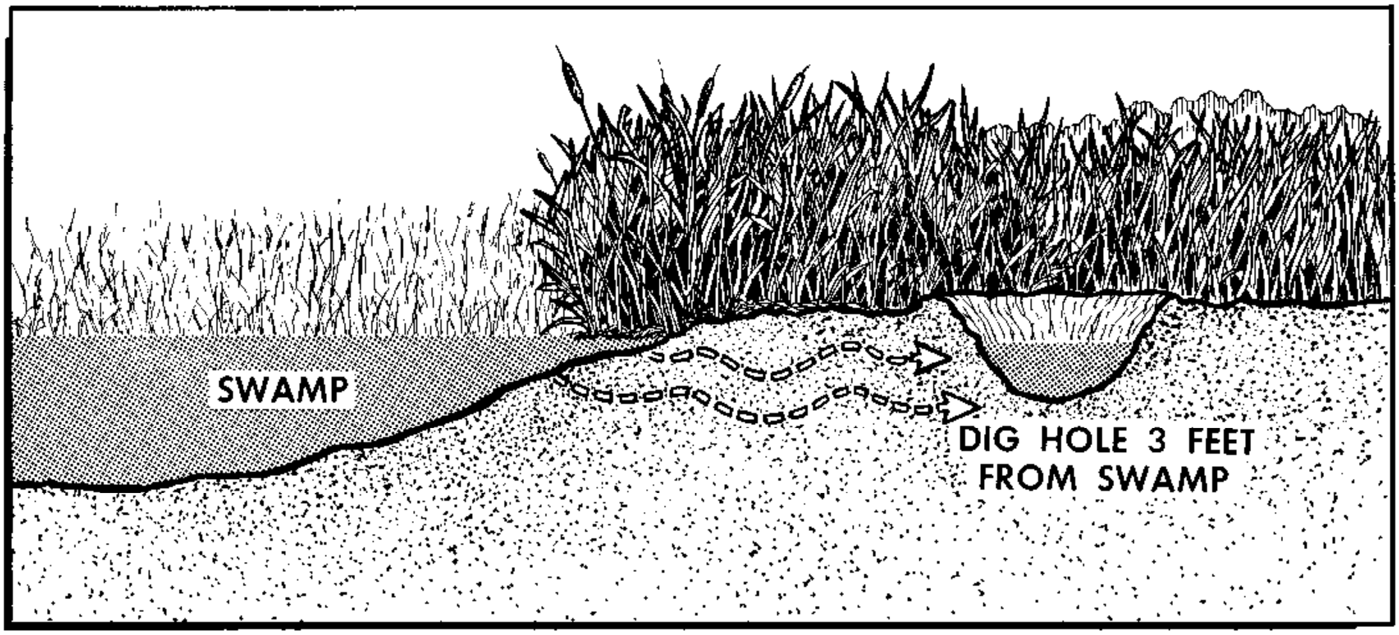
SWAMP
DIG HOLE 3 FEET FROM SWAMP

Figure 19-14. Sediment Hole.

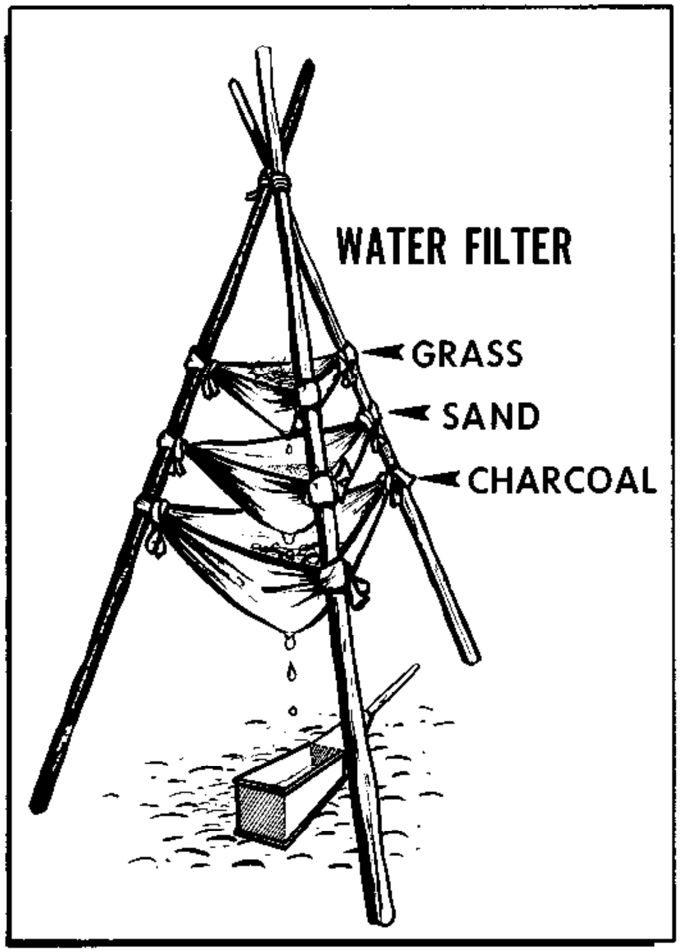
WATER FILTER
GRASS
SAND
CHARCOAL

Figure 19-15. Water Filter.

Part Seven

TRAVEL

Chapter 20

LAND NAVIGATION

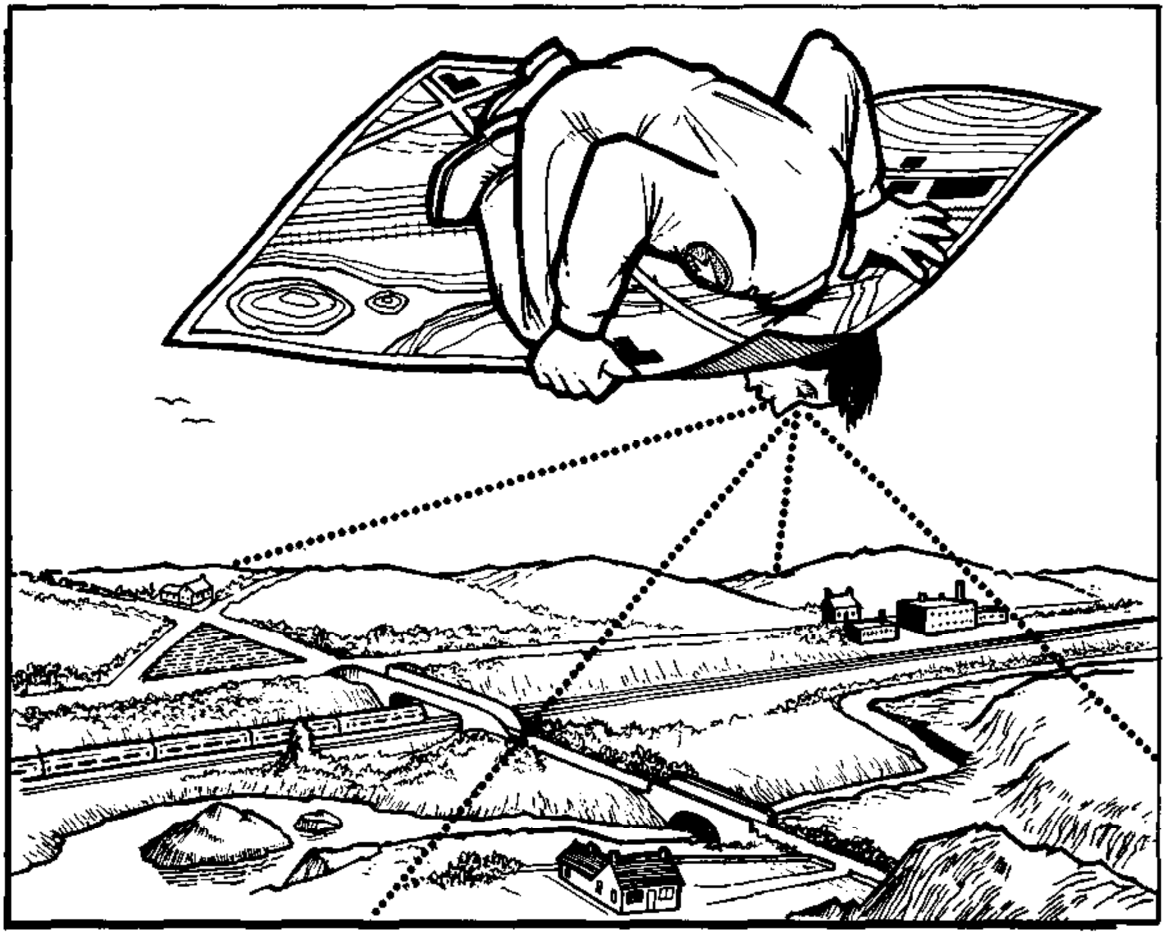

Figure 20-1. Land Nav.

20-1. Introduction:

a. Survivors must know their location in order to intelligently decide if they should wait for rescue or if they should determine a destination and (or) route to travel. If the decision is to stay, the survivors need to know their location in order to radio the information to rescue personnel. If the decision is to travel, survivors must be able to use a map to determine the best routes of travel, location of possible food and water, and hazardous areas which they should avoid.

b. This chapter provides background information in the use of the map and compass (figure 20-1).

20-2. Maps:

a. A map is a pictorial representation of the Earth's surface drawn to scale and reproduced in two dimensions. Every map should have a title, legend, scale, north arrow, grid system, and contour lines. With these components, survivors can determine the portion of the Earth's surface the map covers. Survivors should be able to understand all of the markings on the map and use them to advantage. They should also be able to determine the distance between any two points on the map and be able to align the map with true north so it conforms to the actual features on the ground.

b. A map is a conceptionalized picture of the Earth's surface as seen from above, simplified to bring out important details and lettered for added identification. A map represents what is known about the Earth rather than what can be seen by an observer. However, a map is selective in that only the information which is necessary for its intended use is included on any one map. Maps also include features which are not visible on Earth, such as parallels, meridians, and political boundaries.

c. Since it is impossible to accurately portray a round object, such as the Earth, on a flat surface, all maps have some elements of distortion. Depending on the intended use, some maps sacrifice constant scale for accuracy in measurement of angles, while others sacrifice accurate measurement of angles for a constant scale. However, most maps used for ground navigation use a compromise projection in which a slight amount of distortion is introduced into the elements which a map portrays, but in which a fairly true picture is given.

d. A planimetric map presents only the horizontal positions for the features represented. It is distinguished from a topographic map by the omission of relief in a measurable form.

e. A topographic map (figure 20-2) portrays terrain and landforms in a measurable form and the horizontal positions of the features represented. The vertical positions, or relief, are normally represented by contours. On maps showing relief, the elevations and contours are measured from a specified vertical datum plane and usually mean sea level.

f. A plastic relief map is a reproduction of an aerial photograph or a photomosaic made from a series of aerial photographs upon which grid lines, marginal data, place names, route numbers, important elevations, boundaries, approximate scale, and approximate direction have been added.

g. A PICTOMAP (figure 20-3) is the acronym for photographic image conversion by tonal masking procedures. It is a map on which the photographic imagery of a standard photomap has been converted into interpretable colors and symbols.

h. A photomosaic is an assembly of aerial photographs and is commonly called a mosiac in topographic usage. Mosiacs are useful when time does not permit the compilation of a more accurate map. The accuracy of a mosiac depends on the method used in its preparation and may vary from simply a good pictorial effect of the ground to that of a planimetric map.

i. Military city map is a topographic map, usually 1:12,500 scale, of a city, outlining streets and showing street names, important buildings, and other urban elements of military importance which are compatible with the scale of the map. The scales of military city maps can vary from 1:25,000 to 1:5,000, depending on the importance and size of the city, density of detail, and available intelligence information.

j. Special maps are for special purposes such as trafficability, communications, and assault. These are usually overprinted maps of scales smaller than 1:100,000 but larger than 1:1,000,000. Other types of special maps are those made from organosol or materials other than paper to meet the requirements of special climatic conditions.

k. A terrain model is a scale model of the terrain showing landforms, and in large scale models, industrial and cultural shapes. It is designed to provide a means for visualizing the terrain for planning or indoctrination purposes and for briefing on assault landings.

l. A special purpose map is one that has been designed or modified to give information not covered on a standard map or to elaborate on standard map data. Special purpose maps are usually in the form of an overprint. Overprints may be in the form of individual sheets or combined and bound into a study of an area. A few of the subjects covered are:

(1) Landform.
(2) Drainage characteristics.
(3) Vegetation.
(4) Climate.
(5) Coast and landing beaches.
(6) Railroads.
(7) Airfields.
(8) Urban areas.
(9) Electric power.
(10) Fuels.
(11) Surface water resources.
(12) Ground water resources.
(13) Natural construction materials.
(14) Cross-country movement.
(15) Suitability for airfield construction.
(16) Airborne operations.

20-3. Aeronautical Charts. Air navigation and planning charts are used for flight planning. Each different series of charts is constructed at a different scale and format to meet the needs of a particular type of air navigation. The air navigation and planning charts are smaller in scale and less detailed than Army maps or air target materials. The control of positional error is less critical. The following list includes the charts most commonly used in intelligence operations. They are available through the Defense Mapping Agency (DMA) Officer of Distribution Services, Washington DC. A description of each chart follows the listing:

CHART	SCALE	CODE
USAF Global Navigation and Planning Chart	1:5,000,000	GNC
USAF Jet Navigation Chart	1:3,000,000	JNC-A
USAF Operational Navigation Chart	1:1,000,000	ONC
USAF Tactical Pilotage Chart	1:500,000	TPC
USAF Jet Navigation Chart	1:2,000,000	JN
Joint Operations Graphic	1:250,000	JOG

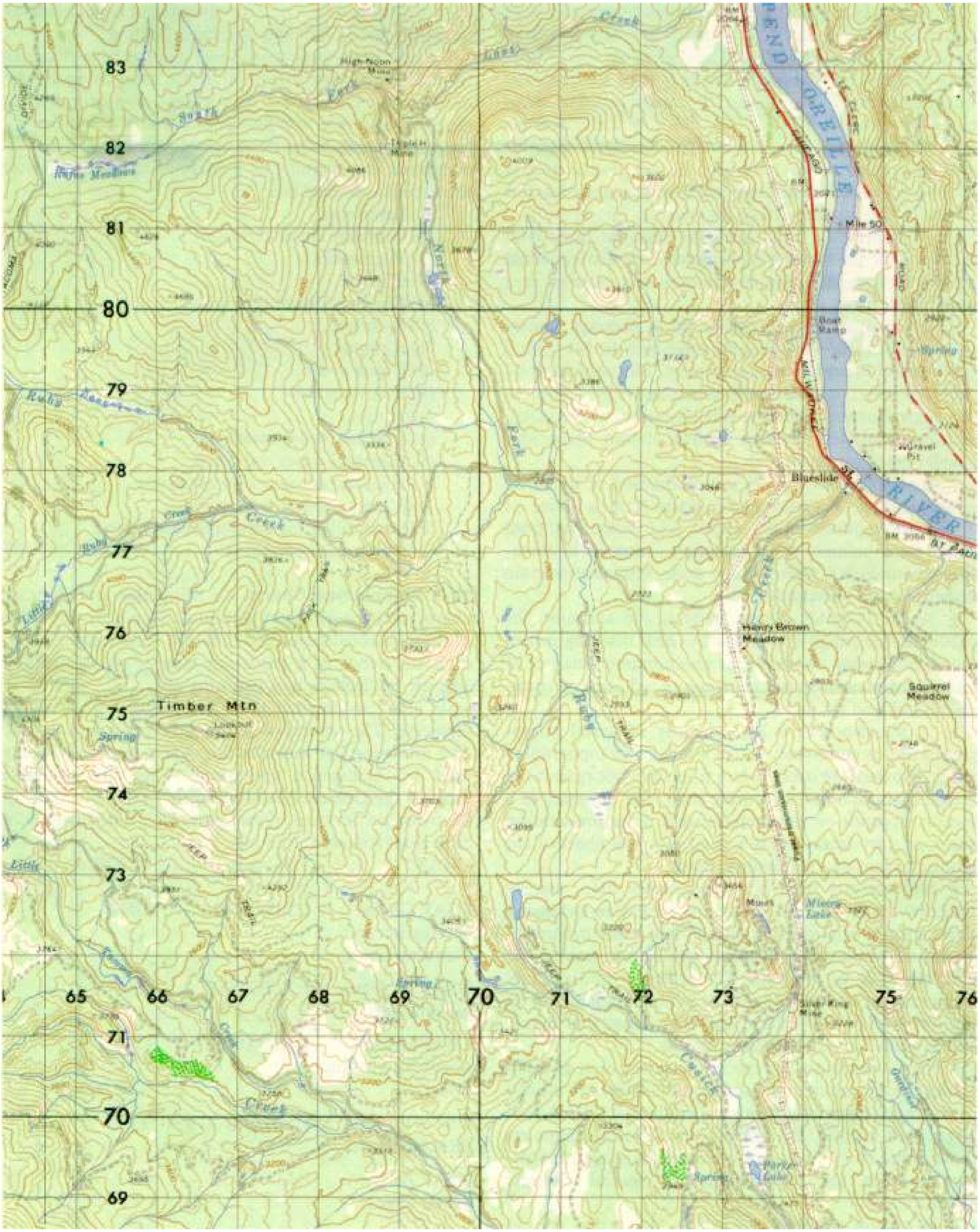
83
82
81
80
79
78
77
76
75
74
73
71
70
69
65
66
67
68
69
70
71
72
73
75
76
High Noon Mine
Triplett Mine
South Fork
North Fork
Ruby Creek
Little Ruby
Creek
PEND OREILLE RIVER
Mile 50
Boat Ramp
Gravel Pit
Blueslide
Spring
Henry Brown Meadow
Squirrel Meadow
Timber Mtn
Lookout
Spring
Jeep Trail
Mines
Miners Lake
Silver King Mine
Parker Lake
Spring

Figure 20-2. Topographic Map.

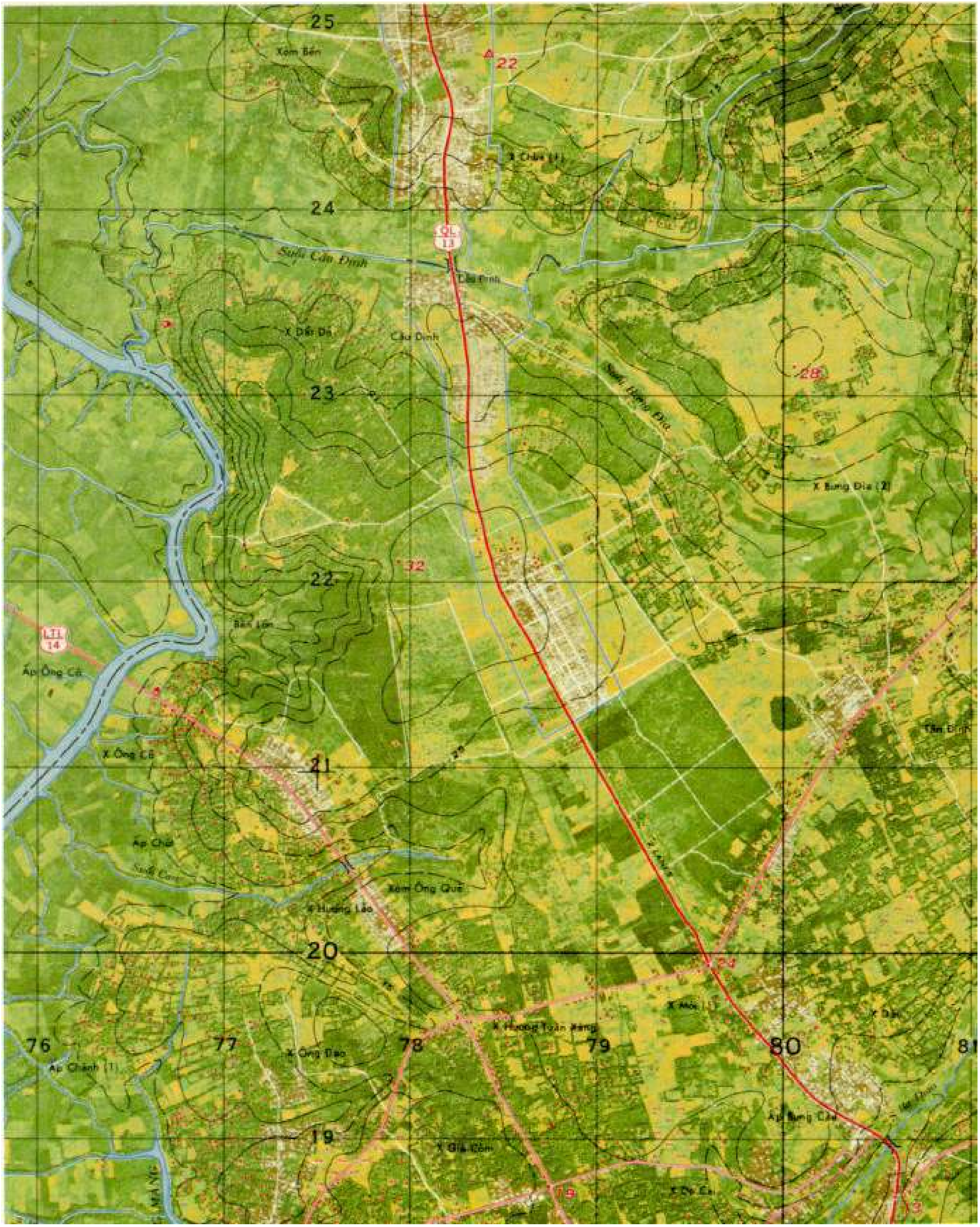
25
Xom Bến
22
X Châu (1)
24
QL 13
Suối Cầu Định
Cầu Định
X Đất Đỏ
Cầu Định
23
Suối Đồng Đá
28
X Bưng Địa (2)
32
22
Bến Lớn
LTL 14
Ấp Ông Cả
X Ông Cả
21
Ấp Chữ
Suối Cạn
Xóm Ông Quế
X Hương Lão
20
24
X Mới (1)
X Đại
Tân Định
76
77
X Ông Đạo
78
X Hương Tuấn Xang
79
80
81
Ấp Chánh (1)
Ấp Bưng Cầu
19
X Gia Cồm
19
X Đỗ Cả
13

Figure 20-3. Pictomap.

a. Global Navigation and Planning Chart (GNC) (figure 20-4). This chart is designed for general planning purposes where large areas of interest and long-distance operations are involved. It serves as a navigation chart for long-range, high-altitude, and high-speed aircraft since sheet lines have been selected on the basis of primary areas of strategic interest. Several other general planning charts are available through the DMA. Some of these charts are produced on selected areas of strategic interest; others provide wide coverage. All general planning charts are produced at a small or very small scale which provides extensive area coverage on a single sheet.

b. USAF Jet Navigation Chart (JN/JNC-A) (figure 20-5). The basic JNC is produced at a scale of 1:2,000,000. The JNC-A is produced on the north polar area and in the United States at a scale of 1:3,000,000. Both jet navigation charts are printed on 41½- by 57⅜-inch sheets.

(1) The JN chart is used for preflight planning and en route navigation by long-range jet aircraft with dead reckoning, radar, celestial, and grid navigational capabilities. The charts are designed so they can be joined to produce a strip chart which provides the necessary navigational information for any intended course. Relief is indicated through the use of contours, spot elevations, and gradient tints. Large, level terrain areas are indicated by a symbol that consists of narrow, parallel lines with the elevation annotated within the symbol.

(2) Principal cities and towns and principal roads and rail networks are shown on the JN chart. The transportation network is shown in the immediate area of populated places. Lakes and principal drainage patterns are also pictured. The elevations of major lakes are indicated so that the altitude may be determined by using the aircraft radar altimeter.

c. USAF Operational Navigation Chart (ONC) (figure 20-6):

(1) The ONC was developed to meet military requirements for a chart adaptable to low-altitude navigation. The ONC is used for preflight planning and en route navigation. It is also used for operational planning, intelligence briefing and plotting, and flight planning displays.

(2) This chart covers an area of 8° of latitude and 12° of longitude. ONC sheets are identified by combining a letter and a number (figure 20-7). Letters identify 8° bands of latitude, starting at the North Pole and progressing southward. Numbers identify 12° sections of longitude from the prime meridian eastward. The successful execution of low-altitude mission depends upon visual and radar identification of ground features used as checkpoints and a rapid visual association of these features with their chart counterparts. The ONC portrays, by conventional signs and symbols, cultural features which have low-altitude checkpoint significance. Powerlines are shown (except on cities) and are indicated by the usual line and pole symbol.

(3) For certain circumstances, operational requirements may be more effectively satisfied by pictorial illustrations than by the conventional symbolization of such structures as prominent buildings, bridges, dams, towers, holding or storage tanks, stadiums, and related features. For these reasons, significant landmarks are depicted on ONCs by pictorial symbols.

(4) The ONC portrays relief in perspective so that the user gets instantaneous appreciation of relative heights, slope gradients, and the forms of ground patterns. Topographic expression, illustrated basically with contours and spot elevations, is emphasized by the use of shaded relief and terrain characteristic tints defining the overall elevation levels. ONC contour intervals and terrain characteristic tints are selected regionally. This captures the relative significance of ground forms as a complete picture, and this feature aids preflight planning and in-flight identification.

d. USAF Tactical Pilotage Chart (TPC) (figure 20-8):

(1) The TPC is produced in a coordinated series at a scale of 1:500,000. Sheet sizes are the same dimensions as the ONC sheets; however, a TPC covers only one-fourth as much area as an ONC sheet. The TPC breakdown on the ONC is illustrated in figure 20-9. A TPC is identified by the ONC identification and the letter "A," "B," "C," or "D."

(2) The TPC is used for detailed preflight planning and mission analysis. In designing the TPC, emphasis was placed on ground features which are significant for low-level, high-speed navigation, using visual and radar means. The selected ground features also permit immediate ground-chart orientation at predetermined checkpoints.

(3) Relief on the TPC is displayed by contours (intervals may vary between 100 feet and 1,000 feet), spot elevations, relief shading, and terrain characteristic tints. Cultural features such as towns and cities, principal roads, railroads, power transmission lines, boundaries, and other features of value for low-altitude visual missions are included on the TPC. Pictorial symbols are used for features which provide the best checkpoints. Other features of the TPC which enhance its tactical air navigation qualities are as follows:

(a) UTM grid overprint.

(b) Vegetation color and symbol code.

(c) Enlarged vertical obstruction symbols.

(d) Enlarged road and railroad symbols.

(e) Emphasized radio aid to navigation symbols.

(f) Foreign place name glossary.

(g) Airdrome runway patterns to scale when information is available.

(h) Spot elevation, gradient tints, and shaded relief depicted for all elevations.

(i) The highest elevation for each 15-minute quadrangle is shown in thousands and hundreds of feet.

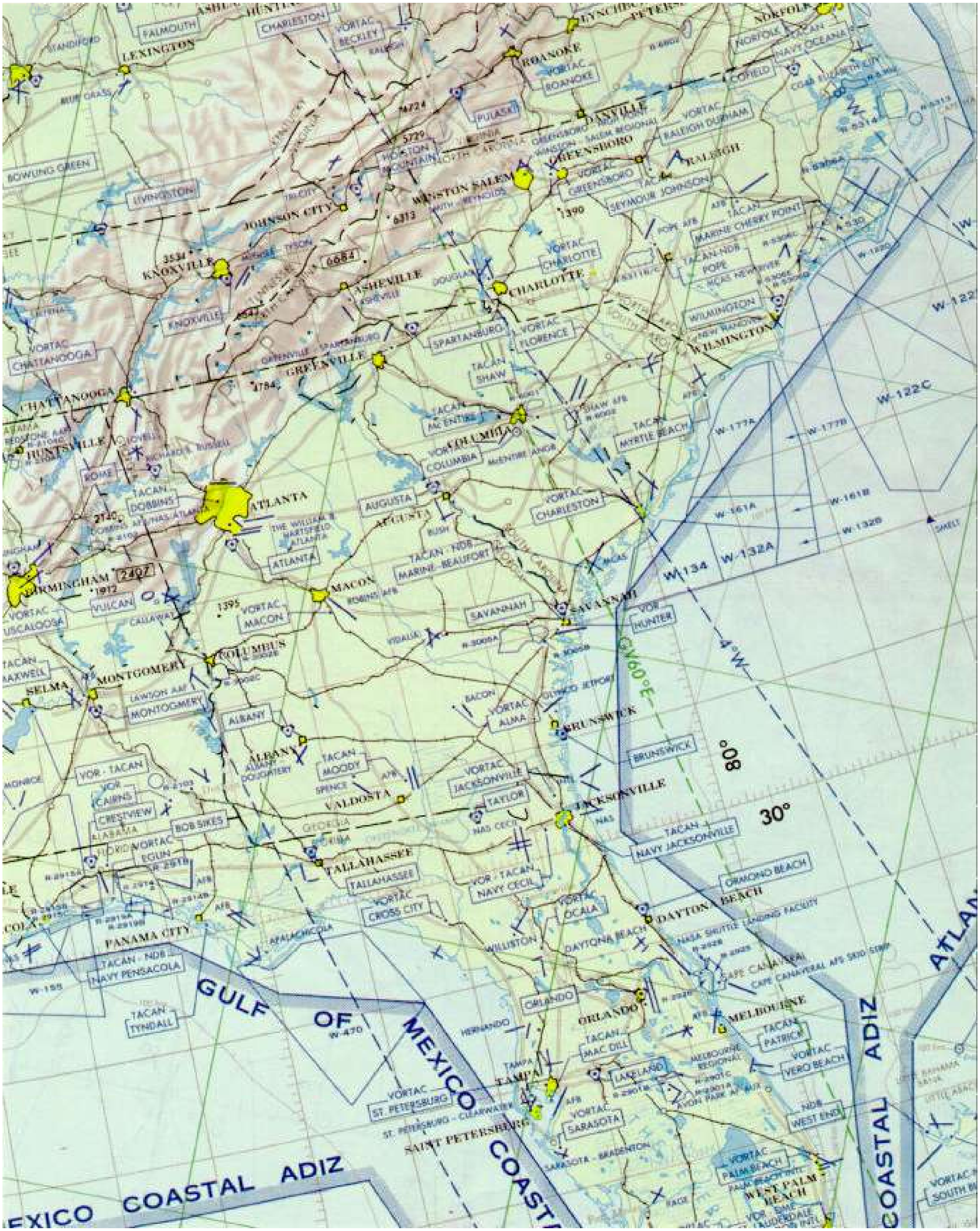
FALMOUTH
CHARLESTON
VORTAC BECKLEY
LYNCHBURG
NORFOLK
LEXINGTON
ROANOKE
VORTAC ROANOKE
NAVY OCEANA
PULASKI
DANVILLE
VORTAC RALEIGH DURHAM
HOLSTON MOUNTAIN
GREENSBORO
RALEIGH
BOWLING GREEN
WINSTON SALEM
VORTAC GREENSBORO
SEYMOUR JOHNSON
LIVINGSTON
JOHNSON CITY
MARINE CHERRY POINT
KNOXVILLE
VORTAC CHARLOTTE
CHARLOTTE
ASHEVILLE
POPE
KNOXVILLE
WILMINGTON
VORTAC CHATTANOOGA
SPARTANBURG
VORTAC FLORENCE
WILMINGTON
GREENVILLE
TACAN SHAW
W-122
CHATTANOOGA
TACAN MC ENTIRE
COLUMBIA
W-122C
HUNTSVILLE
TACAN MYRTLE BEACH
W-177A
W-177B
COLUMBIA
ROME
TACAN DOBBINS
ATLANTA
AUGUSTA
AUGUSTA
VORTAC CHARLESTON
W-161A
W-161B
W-132B
THE WILLIAM B HARTSFIELD ATLANTA
ATLANTA
TACAN - NDB MARINE BEAUFORT
W-134
W-132A
BIRMINGHAM
VULCAN
MACON
VORTAC TUSCALOOSA
VORTAC MACON
SAVANNAH
SAVANNAH
VOR HUNTER
COLUMBUS
TACAN MAXWELL
MONTGOMERY
SELMA
LAWSON AAF
MONTGOMERY
3°00'E
4°W
ALBANY
VORTAC ALMA
BRUNSWICK
BRUNSWICK
ALBANY
TACAN MOODY
80°
VOR - TACAN
VOR CAIRNS
CRESTVIEW
VALDOSTA
VORTAC JACKSONVILLE
TAYLOR
JACKSONVILLE
30°
BOB SIKES
VORTAC EGLIN
TACAN NAVY JACKSONVILLE
TALLAHASSEE
TALLAHASSEE
VOR - TACAN NAVY CECIL
ORMOND BEACH
VORTAC CROSS CITY
VORTAC OCALA
DAYTONA BEACH
PANAMA CITY
DAYTONA BEACH
TACAN - NDB NAVY PENSACOLA
GULF OF MEXICO
W-155
W-470
CAPE CANAVERAL
TACAN TYNDALL
ORLANDO
ORLANDO
MELBOURNE
TACAN MAC DILL
TACAN PATRICK
TAMPA
LAKELAND
VORTAC VERO BEACH
VORTAC ST. PETERSBURG
ST. PETERSBURG - CLEARWATER
SAINT PETERSBURG
VORTAC SARASOTA
NDB WEST END
SARASOTA - BRADENTON
VORTAC PALM BEACH
WEST PALM BEACH
MEXICO COASTAL ADIZ
COAST
COASTAL ADIZ
ATLAN
VORTAC SOUTH B

Figure 20-4. GNC Map.

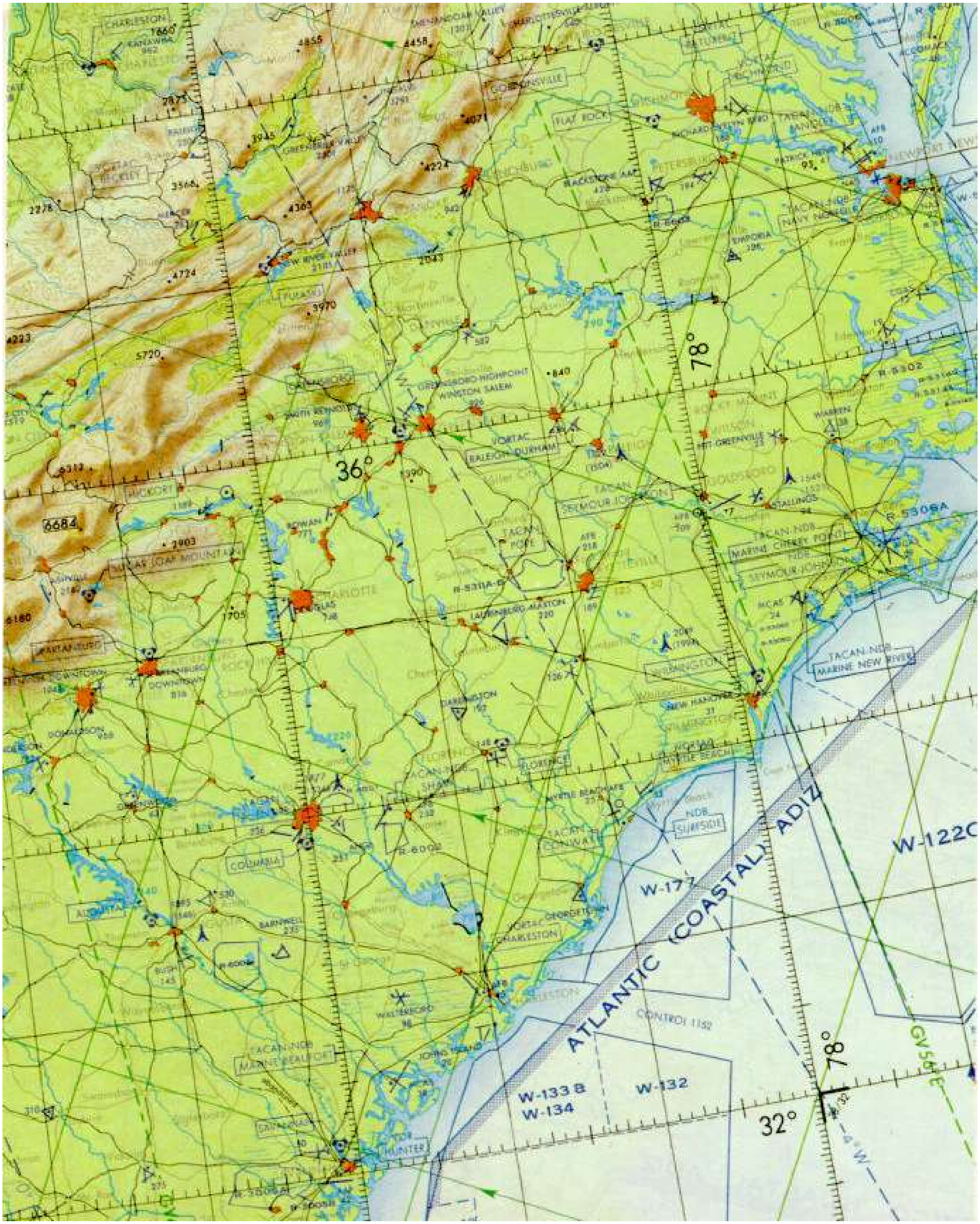
ATLANTIC (COASTAL) ADIZ
W-122C
W-177
W-133 B
W-134
W-132
36°
78°
32°
CHARLOTTE
RALEIGH-DURHAM
GREENSBORO-HIGHPOINT WINSTON SALEM
FLORENCE
COLUMBIA
WILMINGTON
MYRTLE BEACH
CHARLESTON
RICHMOND
PETERSBURG
SEYMOUR JOHNSON
MARINE CHERRY POINT
MARINE NEW RIVER
MARINE BEAUFORT
HUNTER
SUMTER
SHAW
POPE
LYNCHBURG
GORDONSVILLE
BECKLEY
PULASKI
HICKORY
SPARTANBURG
AUGUSTA
SAVANNAH
CONWAY
GEORGETOWN
LAURINBURG-MAXTON
FLAT ROCK
BLACKSTONE AAF
NAVY NORFOLK
NEWPORT NEWS

Figure 20-5. JNC Map.

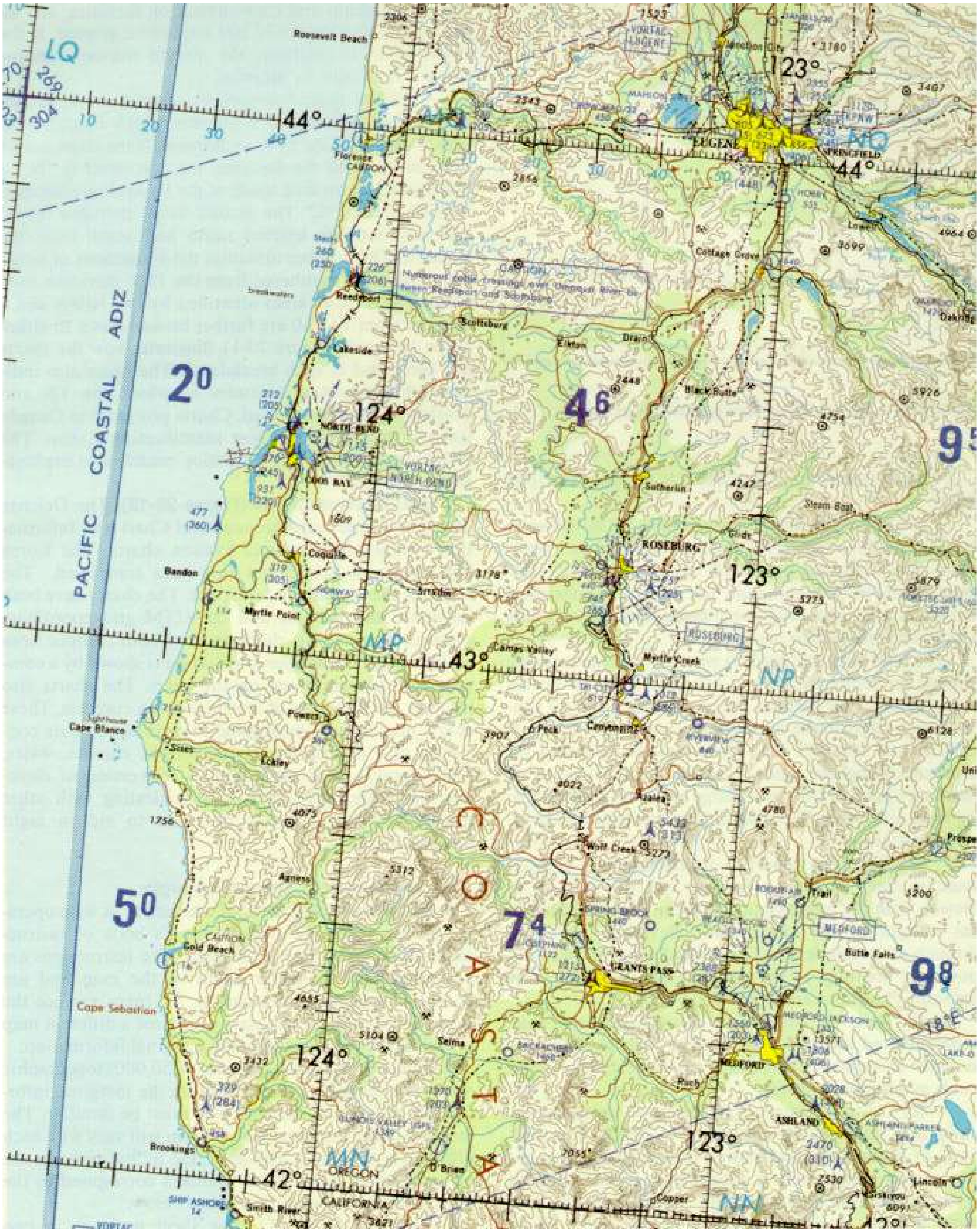
LQ
PACIFIC COASTAL ADIZ
44°
124°
123°
43°
42°
2⁰
4⁶
5⁰
7⁴
9⁸
COAST
EUGENE
SPRINGFIELD
Florence
Reedsport
Lakeside
NORTH BEND
COOS BAY
Bandon
Coquille
Myrtle Point
Cape Blanco
Gold Beach
Cape Sebastian
Brookings
ROSEBURG
Sutherlin
Drain
Elkton
Scottsburg
Cottage Grove
Myrtle Creek
Canyonville
GRANTS PASS
MEDFORD
ASHLAND
OREGON
CALIFORNIA
Smith River
Numerous cable crossings over Umpqua River between Reedsport and Scottsburg
MP
NP
MN
NN

Figure 20-6. ONC Map.

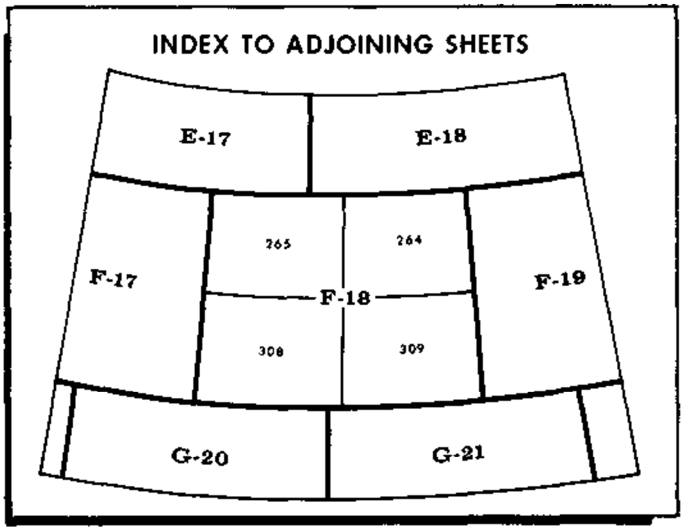

Figure 20-7. Operational Navigation Chart Index.

e. Joint Operations Graphic (JOG) (Series 1501 AIR):

(1) JOGs (figure 20-10) are series of 1:250,000 scale military maps designed for joint ground and air operations. The maps are published in ground and air editions. Both series emphasize the air-landing facilities but the air series has additional symbols to identify aids and obstructions to air navigation.

(2) JOG was designed to provide a common-scale graphic for Army, Navy, and Air Force use. Air Forces make use of it for tactical air operations, close air support, and interdiction by medium- and high-speed aircraft at low altitudes. The chart may also be used for dead reckoning and visual pilotage for short-range en route navigation. Due to its large scale, it is unsuitable for local area command planning for strategic and tactical operations.

(a) Relief on the JOG is indicated by contour lines (in feet). In some areas, the intervals may be in meters, with the approximate value in feet indicated in the margin of the chart. Spot elevations are used through all terrain levels. The ground series show elevations and contours in meters while the air series show the same elevations and contours in feet.

(b) Relief is also shown through gradient tints, supplemented by shaded relief. The highest elevations in each 15-minute quadrangle are indicated in thousands and hundreds of feet.

(c) Cultural features, such as cities, towns, roads, trails, and railroads are illustrated in detail. The locations of boundaries and power transmission lines are also shown. Vegetation is shown by symbol. Detailed drainage patterns and water tint are used to illustrate water features, such as coastlines, oceans, lakes, rivers and streams, canals, swamps, and reefs. The JOG includes aeronautical information such as airfields, fixed radio navigation and communication facilities, and all known obstructions over 200 feet above ground. If the information is available, the airfield runway patterns are shown to scale by diagram.

(d) The basic numbering system of the JOG consists of two letters and a number which identifies an area 6° in longitude by 4° in latitude. If the chart covers an area north of the Equator, the first letter is "N;" a chart covering an area south of the Equator is identified with an initial "S." The second letter identifies the 4° bands of latitude lettered north and south from the Equator. The number identifies the 6° sections of longitude which are numbered from the 180° meridian eastward. The 6° x 4° areas identified by two letters and a number from 1 to 60 are further broken down to either 12 or 16 sheets. Figure 20-11 illustrates how the sheets are numbered in each breakdown. The figure also indicates the respective latitudes at which the 12- and 16-sheet breakdown is used. Charts produced in Canada use a slightly different sheet identification system. The DOD Aeronautical Chart Catalog contains an explanation of the system.

f. DOD Evasion Charts (Figure 20-12). The Defense Mapping Agency and Aeronautical Chart and Information Center prepare DOD evasion charts. The Korea and Southeast Asia charts have been completed. The scale for these charts is 1:250,000. The charts have both longitude and latitude and the UTM grid coordinate systems. The relief is duplicated by both contour lines and shading. The magnetic variation is shown by a compass rose superimposed on the chart. The charts also indicate the direction of seasonal ocean currents. These charts may include geographic environmental data consisting of a description of the people, climate, water, food, hazards, and vegetation. A conversion of elevation bar scale may aid in communicating with other forces. The star chart is provided to aid in night navigation.

20-4. Information Contained in Margin:

a. Before using any piece of equipment, a wise operator always reads the manufacturer's book of instructions. This is also true with maps. The instructions are placed around the outer edges of the map and are known as marginal information. All maps are not the same, so it becomes necessary each time a different map is used to carefully examine the marginal information.

b. Figure 20-13 is a large-scale (1:50,000) topographic map. The circled numbers indicate the marginal information with which the map user must be familiar. The location of the marginal information will vary with each different type of map. However, the following items are on most maps. The circled numbers correspond to the item numbers listed and described below.

(1) Sheet Name (1). The sheet name is in two places; the center of the upper margin and the right side of the lower margin. Generally, a map is named after its

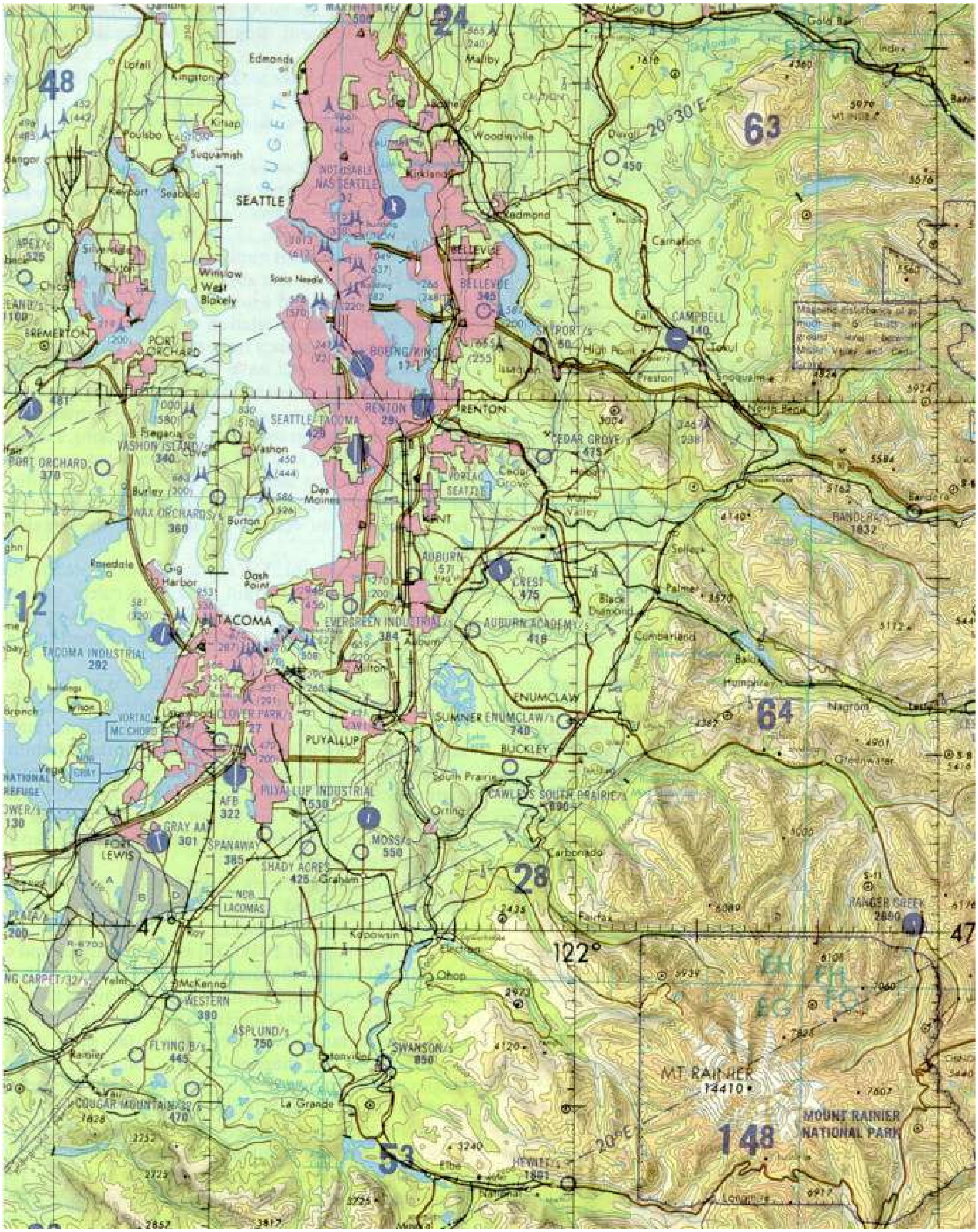
48
24
63
12
64
28
53
148
Edmonds
Kingston
Poulsbo
Suquamish
Bangor
PUGET
SEATTLE
NAS SEATTLE
Kirkland
Woodinville
Redmond
Duvall
Carnation
BELLEVUE
Space Needle
Winslow
West Blakely
BREMERTON
PORT ORCHARD
BOEING/KING
RENTON
SEATTLE TACOMA
Fall City
CAMPBELL
Snoqualmie
North Bend
Preston
Issaquah
CEDAR GROVE
VASHON ISLAND
Vashon
Des Moines
Burton
Burley
KENT
AUBURN
CREST
Gig Harbor
Dash Point
TACOMA
TACOMA INDUSTRIAL
EVERGREEN INDUSTRIAL
AUBURN ACADEMY
Black Diamond
Cumberland
Palmer
Selleck
BANDERA
CLOVER PARK
PUYALLUP
SUMNER
ENUMCLAW
BUCKLEY
Greenwater
South Prairie
Orting
AFB
PUYALLUP INDUSTRIAL
GRAY AAF
FORT LEWIS
SPANAWAY
SHADY ACRES
Graham
MOSS
Carbonado
RANGER CREEK
Fairfax
Kapowsin
Electron
Ohop
122°
47°
WESTERN
McKenna
Yelm
ASPLUND
FLYING B
SWANSON
Eatonville
Rainier
COUGAR MOUNTAIN
La Grande
MT RAINIER
14410
MOUNT RAINIER NATIONAL PARK
HEWITT
Elbe
Longmire
20°30'E
20°E
Magnetic disturbance of as much as 6° exists at ground level between Miller Valley and Cedar Grove

Figure 20-8. TPC Map.

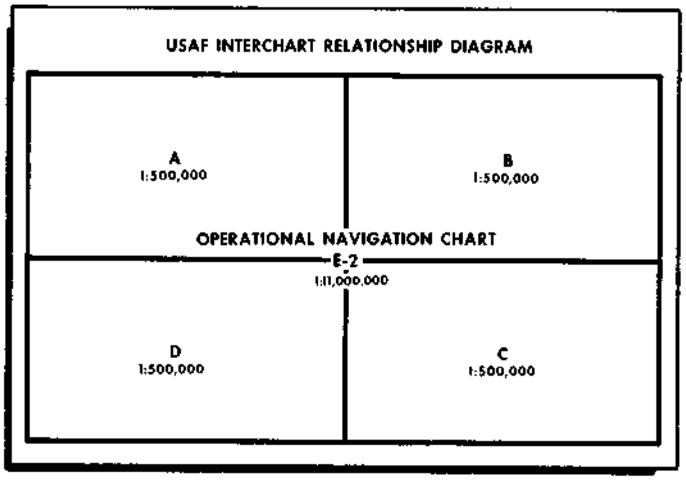
USAF INTERCHART RELATIONSHIP DIAGRAM
A
1:500,000
B
1:500,000
OPERATIONAL NAVIGATION CHART
E-2
1:1,000,000
D
1:500,000
C
1:500,000

Figure 20-9. Relationship of TPC and ONC.

outstanding cultural or geographic feature. When possible, the name of the largest city on the map is used (not shown).

(2) Sheet Number (2). The sheet number is in the upper right margin and is used as a reference number for that map sheet. For maps at 1:100,000 scale and larger, sheet numbers are based on an arbitrary system which makes possible the ready orientation of maps at scales of 1:100,000, 1:50,000, and 1:25,000 (figures 20-14 and 20-15).

(3) Series Name and Scale (3):

(a) The map series name is in the upper left margin. A map series usually comprises a group of similar maps at the same scale and on the same sheet lines or format designed to cover a particular geographic area. It may also be a group of maps which serve a common purpose, such as military city maps. The name given a series is of the most prominent area. The scale note is a representative fraction which gives the ratio of map distance to the corresponding distance on the Earth's surface. For example, the scale note 1:50,000 indicates that one unit of measure on the map equals 50,000 units of the same measure on the ground.

(b) Scale. The scale is expressed as a fraction and gives the ratio of map distance to ground distance. The terms "small scale," "medium scale," and "large scale" may be confusing when read with the numbers. However, if the number is viewed as a fraction, it quickly becomes apparent the 1:600,000 of something is smaller than 1:75,000 of the same thing. Hence, the larger the number after 1:, the smaller the scale of the map.

-1. Small Scale. Maps at scales of 1:600,000 and smaller are used for general planning and strategical studies at the high echelons. The standard small scale is 1:1,000,000.

-2. Medium Scale. Maps at scales larger than 1:600,000 but smaller than 1:75,000 are used for planning operations, including the movement and concentration of troops and supplies. The standard medium scale is 1:250,000.

-3. Large Scale. Maps at scales of 1:75,000 and larger are used to meet the tactical, technical, and administrative needs of field units. The standard large scale is 1:50,000.

(4) Series Number (4). The series number appears in the upper right margin and the lower left margin. It is a comprehensive reference expressed either as a four-digit numeral (example, 1125), or as a letter, followed by a three- or four-digit numeral (example, V7915).

(5) Edition Number (5). The edition number is in the upper margin and lower left margin. It represents the age of the map in relation to other editions of the same map and the agency responsible for its production. The latest edition will have the highest number. EDITION 1 DMATC indicates the first edition prepared by the Defense Mapping Agency Topological Center. Edition numbers run consecutively; a map bearing a higher edition number is assumed to contain more recent information than the same map bearing a lower edition number. Advancement of the edition number constitutes authority to rescind or supersede the previous edition.

(6) Bar Scales (6). The bar scales are located in the center of the lower margin. They are rulers used to convert map distance to ground distance. Maps normally have three or more bar scales, each a different unit of measure.

(7) Credit Note (7). The credit note is in the lower left margin. It lists the producer, dates, and general methods of preparation or revision. This information is important to the map user in evaluating the reliability of the map as it indicates when and how the map information was obtained. On some recent 1:50,000 scale maps, the map credits are shown in tabular form in the lower margin, with reliability information presented in a coverage diagram.

(8) Adjoining Sheets Diagram (8) (not shown). Maps at all standard scales contain a diagram which illustrates the adjoining sheets.

(a) On maps at 1:100,000 and larger scales and at 1:1,000,000 scales, the diagram is called the Index to Adjoining Sheets, and consists of as many rectangles, representing adjoining sheets, as are necessary to surround the rectangle which represents the sheet under consideration. The diagram usually contains nine rectangles, but the number or names may vary depending on the location of the adjoining sheets. All represented sheets are identified by their sheet numbers. Sheets of an adjoining series, whether published or planned, that are the same scale are represented by dashed lines. The series number of the adjoining series is indicated along the appropriate side of the division line between the series (figure 20-16).

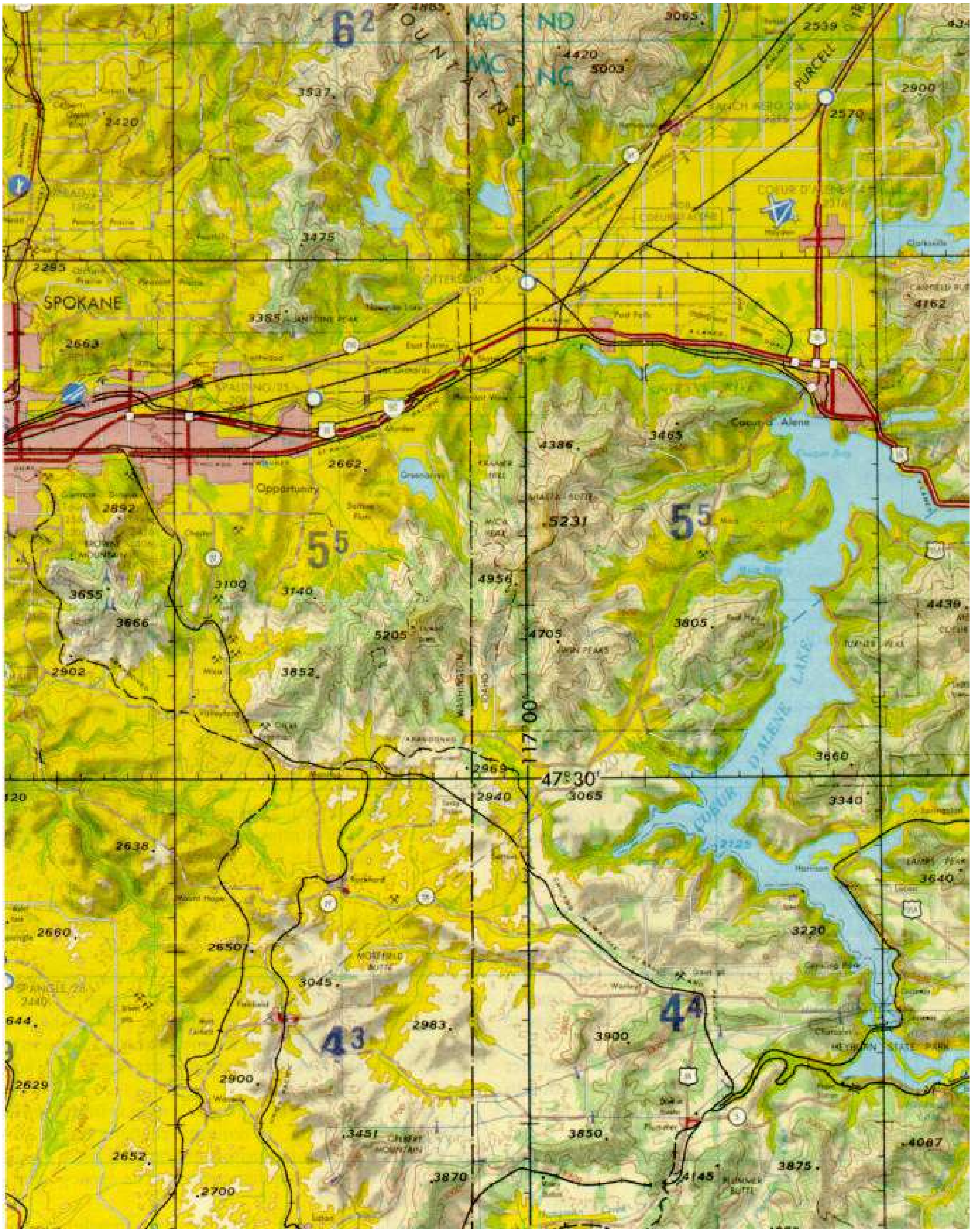
SPOKANE
Opportunity
Coeur d'Alene
COEUR D'ALENE LAKE
WASHINGTON
IDAHO
117°00'
47°30'
62
55
55
43
44
5231
5205
4956
4705
4386
3465
3805
3852
3655
3666
3100
3140
2662
2652
2285
2663
3385
3475
3537
2420
4420
5003
3065
2539
2570
2900
4162
4439
3660
3340
3640
3220
3900
3850
3875
4087
4145
3870
3451
2983
3045
2900
2650
2638
2660
2629
2652
2700
2902
2969
2940
3065

Figure 20-10. Joint Operations Graphics Map.

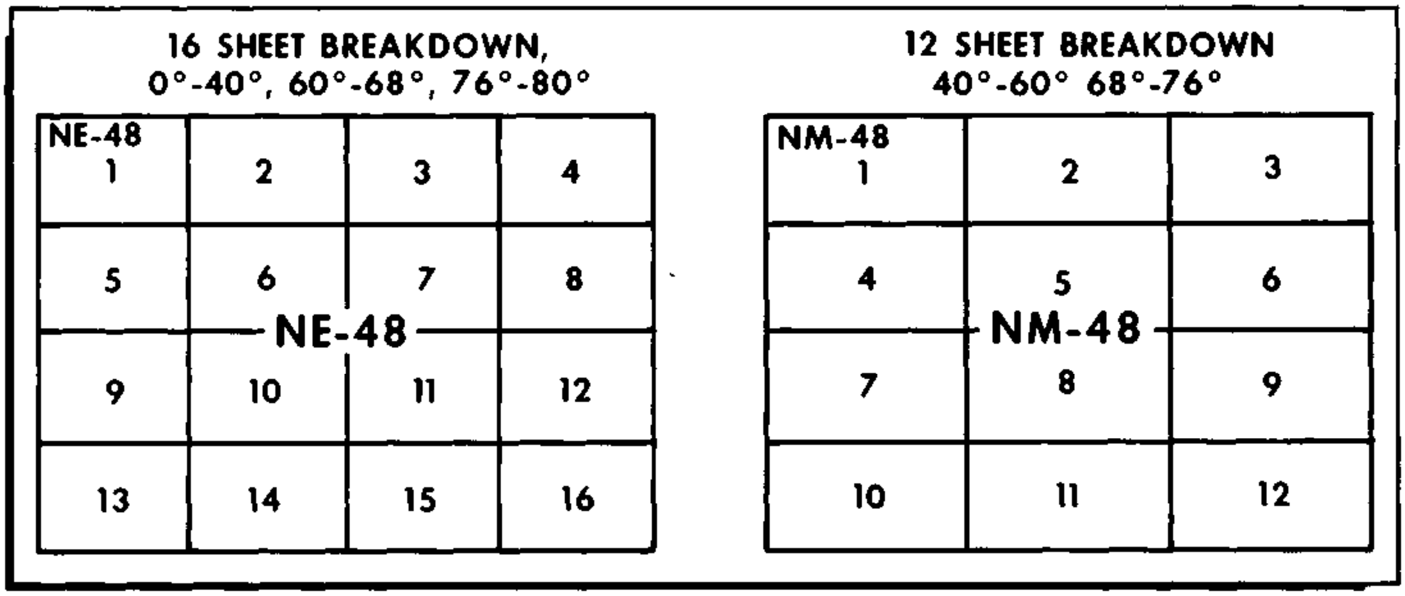

Figure 20-11. JOG Sheet Numbering System.

(b) On 1:50,000 scale maps, the sheet number and series number of the 1:250,000 scale map of the area are shown below the Index to Adjoining Sheets.

(c) On maps at 1:250,000 scale, the adjoining sheets are shown in the location diagram. Usually, the diagram consists of 25 rectangles, but the number may vary with the locations of the adjoining sheets.

(9) Index to Boundaries (9). The index to boundaries diagram appears in the lower right margin of all sheets 1:100,000 scale or larger, and 1:1,000,000 scale. This diagram, which is a miniature of the map, shows the boundaries which occur within the map area, such as county lines and state boundaries. On 1:250,000 scale maps, the boundary information is included in the location diagram.

(10) Projection Note (10). The projection system is the framework of the map. For maps, this framework is the conformal type; that is, small areas of the surface of the Earth retain their true shapes on the projection, measured angles closely approximate true values, and the scale factor is the same in all directions from a point. The projection is identified on the map by a note in the lower margin.

(11) Grid Note (11). The grid note is in the center of the lower margin. It gives information pertaining to the grid system used, the interval of grid lines, and the number of digits omitted from the grid values. Notes pertaining to overlapping or secondary grids are also included when appropriate.

(12) Grid Reference Box (12). The grid reference box has instructions for composing a grid reference.

(13) Vertical Datum Note (13). This note is in the center of the lower margin. It designates the basis for all vertical control stations, contours, and elevations appearing on the map. On JOGs at 1:250,000 scale, the vertical datum note may appear in the reliability diagram.

(14) Horizontal Datum Note (14). This note is located in the center of the lower margin. It indicates the basis for all horizontal control stations appearing on the map. This network of stations controls the horizontal positions of all mapped features. On JOGs at 1:250,000 scale, the horizontal datum note may appear in the reliability diagram.

(15) Legend (15). The legend is located in the lower left margin. It illustrates and identifies the topographic symbols used to depict some of the more prominent features on the map. The symbols are not always the same on every map. To avoid error in the interpretation of symbols, the legend must always be referred to when a map is read.

(16) Declination Diagram (16). The declination diagram is usually located in the lower margin of large-scale maps and indicates the angular relationships of true north, grid north, and magnetic north. On maps at 1:250,000 scale, this information is expressed as a note in the lower margin.

(17) User's Note (17). A user's note is in the center of the lower margin. It requests cooperation in correcting errors or omissions on the map. Errors should be marked and the map forwarded to the agency identified in the note.

(18) Unit Imprint (18). The unit imprint, in the lower left margin, identifies the agency which printed the map and the printing date. The printing date should

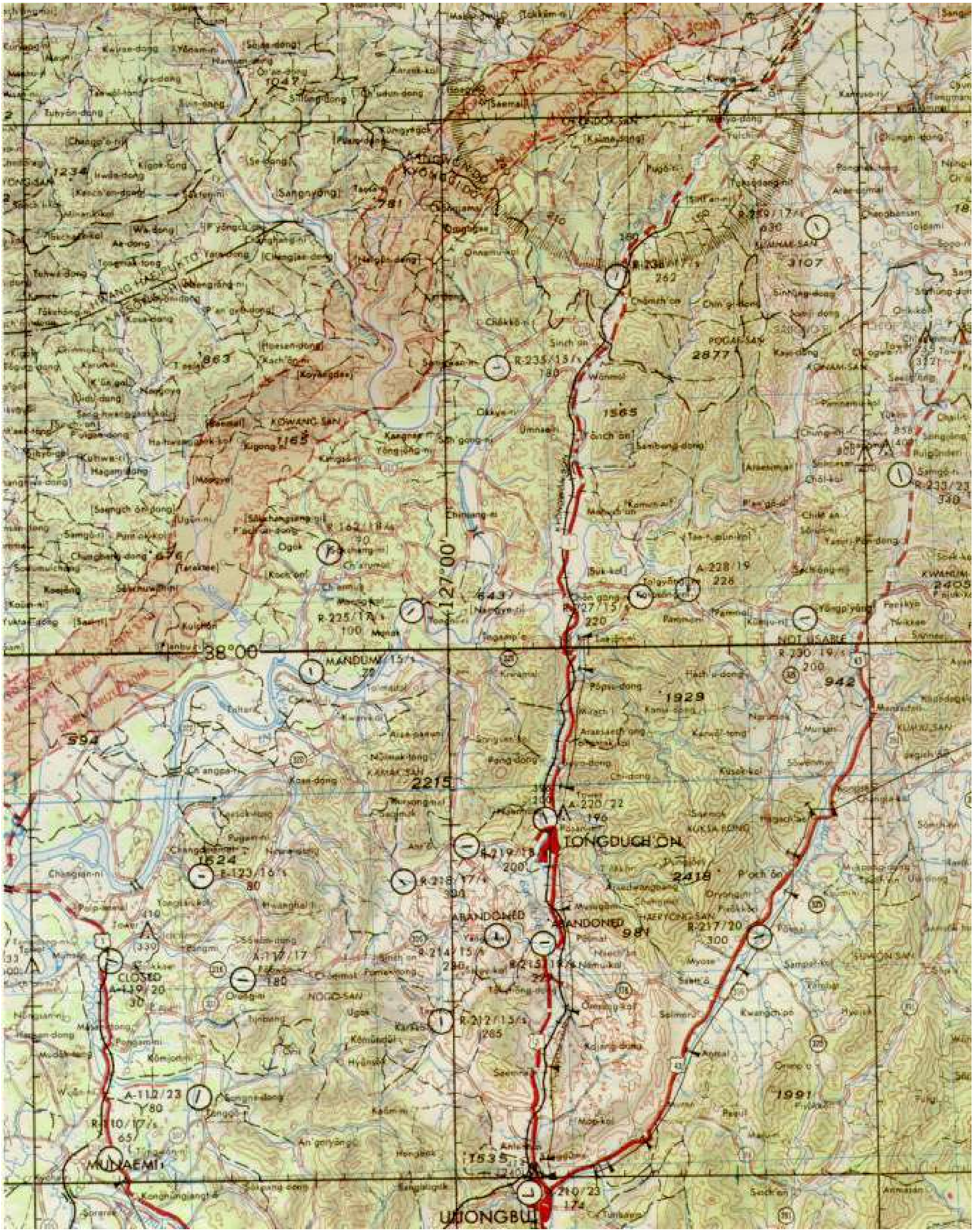

38°00'
127°00'
TONGDUCH'ON
UIJONGBU
MUNAEMI
ABANDONED
NDT UNUSABLE
CLOSED

Figure 20-12. DOD Evasion Chart.

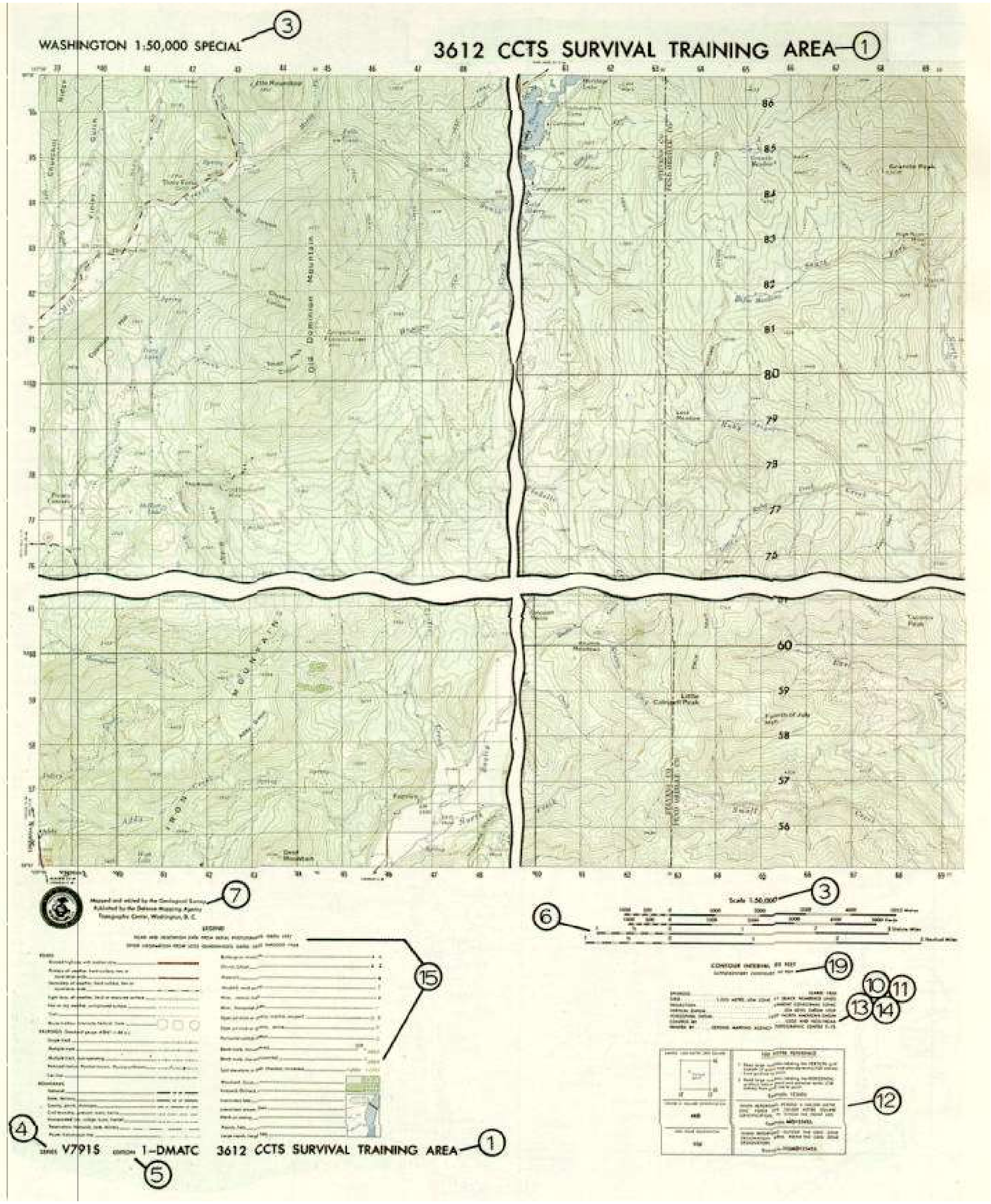
WASHINGTON 1:50,000 SPECIAL
③
3612 CCTS SURVIVAL TRAINING AREA
①
Old Dominion Mountain
MOUNTAINS
IRON
Little Collegiate Peak
Granite Peak
86
85
84
83
82
81
80
79
78
77
76
61
60
59
58
57
56
⑦
Scale 1:50,000
③
⑥
CONTOUR INTERVAL 20 FEET
⑲
⑩
⑪
⑬
⑭
⑫
⑮
④
SERIES V791S
EDITION 1-DMATC
⑤
3612 CCTS SURVIVAL TRAINING AREA
①

Figure 20-13. 1:50,000 Topographic Map.

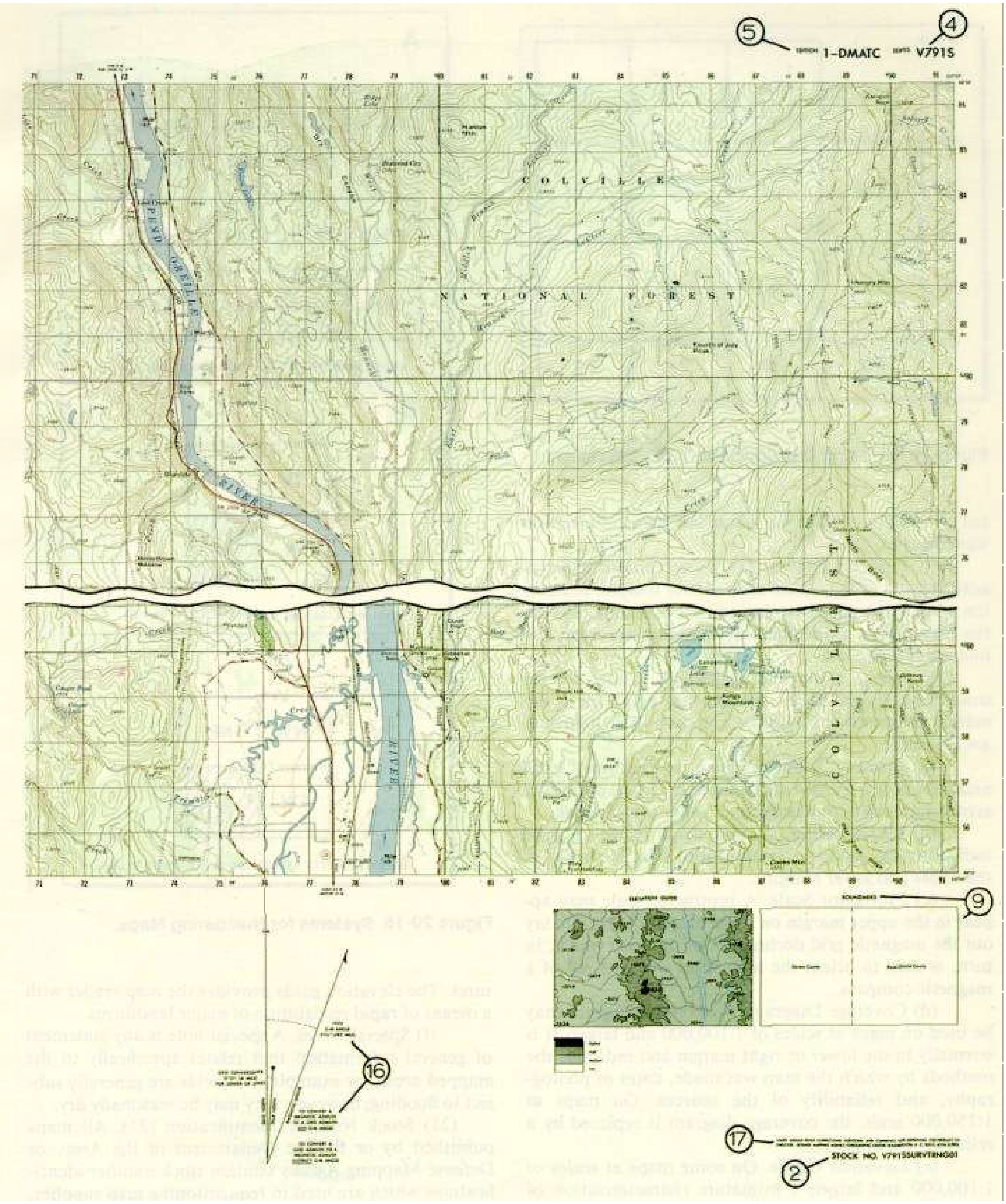
⑤
④
EDITION 1-DMATC
SERIES V791S
C O L V I L L E
N A T I O N A L F O R E S T
PEND OREILLE
RIVER
RIVER
COLVILLE
⑨
⑯
⑰
STOCK NO. V791SSURVTRNG01
②

Figure 20-13. 1:50,000 Topographic Map. (continued)

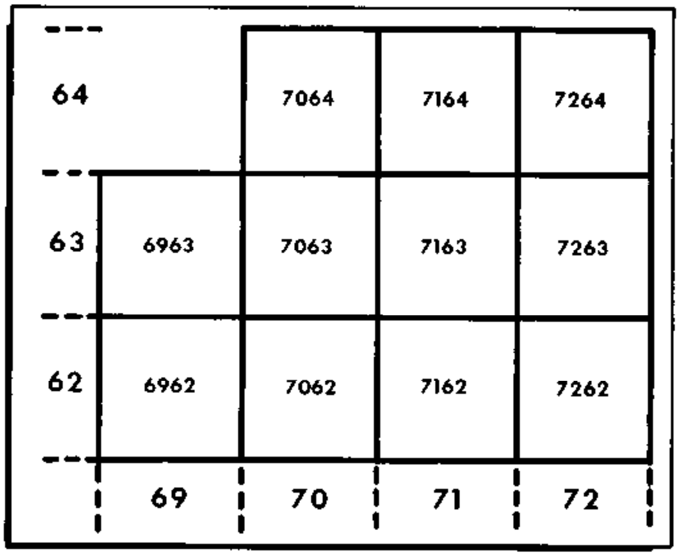
64
7064
7164
7264
63
6963
7063
7163
7263
62
6962
7062
7162
7262
69
70
71
72

Figure 20-14. Basic Development, 1:100,000 Scale.

not be used to determine when the map information was obtained.

(19) Contour Interval (19). The contour interval note appears in the center of the lower margin. It states the vertical distance between adjacent contour lines on the map. When supplementary contours are used, the interval is indicated.

(20) Special Notes and Scales (20). Under certain conditions, special notes or scales may be added to the margin information to aid the map user. The following are examples:

(a) Glossary. A glossary is an explanation of technical terms or a translation of terms on maps of foreign areas where the native language is other than English.

(b) Classification. Certain maps require a note indicating the security classification. This is shown in the upper and lower margins.

(c) Protractor Scale. A protractor scale may appear in the upper margin on some maps. It is used to lay out the magnetic grid declination of the map which, in turn, is used to orient the map sheet with the aid of a magnetic compass.

(d) Coverage Diagram. A coverage diagram may be used on maps at scales of 1:100,000 and larger. It is normally in the lower or right margin and indicates the methods by which the map was made, dates of photography, and reliability of the sources. On maps at 1:250,000 scale, the coverage diagram is replaced by a reliabilty diagram.

(e) Elevation Guide. On some maps at scales of 1:100,000 and larger, a miniature characterization of the terrain is shown by a diagram in the lower right margin of the map. The terrain is represented by bands of elevation, spot elevations, and major drainage features. The elevation guide provides the map reader with a means of rapid recognition of major landforms.

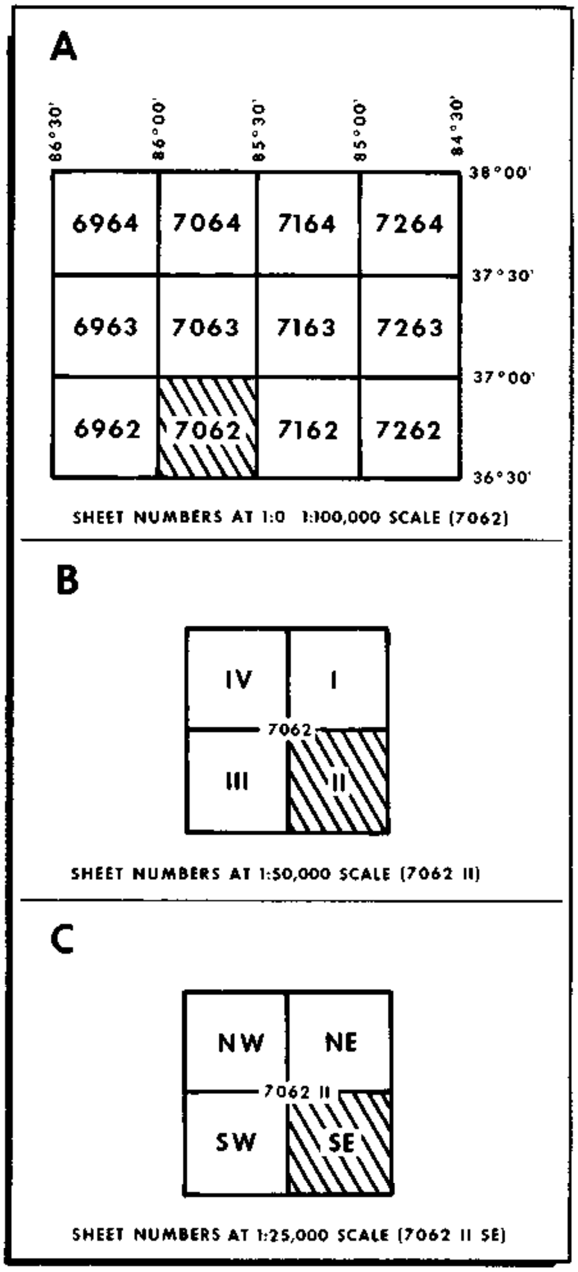
A
86°30'
86°00'
85°30'
85°00'
84°30'
38°00'
37°30'
37°00'
36°30'
6964
7064
7164
7264
6963
7063
7163
7263
6962
7062
7162
7262
SHEET NUMBERS AT 1:0 1:100,000 SCALE (7062)
B
IV
I
7062
III
II
SHEET NUMBERS AT 1:50,000 SCALE (7062 II)
C
NW
NE
7062 II
SW
SE
SHEET NUMBERS AT 1:25,000 SCALE (7062 II SE)

Figure 20-15. Systems for Numbering Maps.

(f) Special Notes. A special note is any statement of general information that relates specifically to the mapped area. For example, rice fields are generally subject to flooding; however, they may be seasonally dry.

(21) Stock Number Identification (21). All maps published by or for the Department of the Army or Defense Mapping Agency contain stock number identifications which are used in requisitioning map supplies. The identification consists of the words "STOCK NO." followed by a unique designation which is composed of the series number, the sheet number of the individual

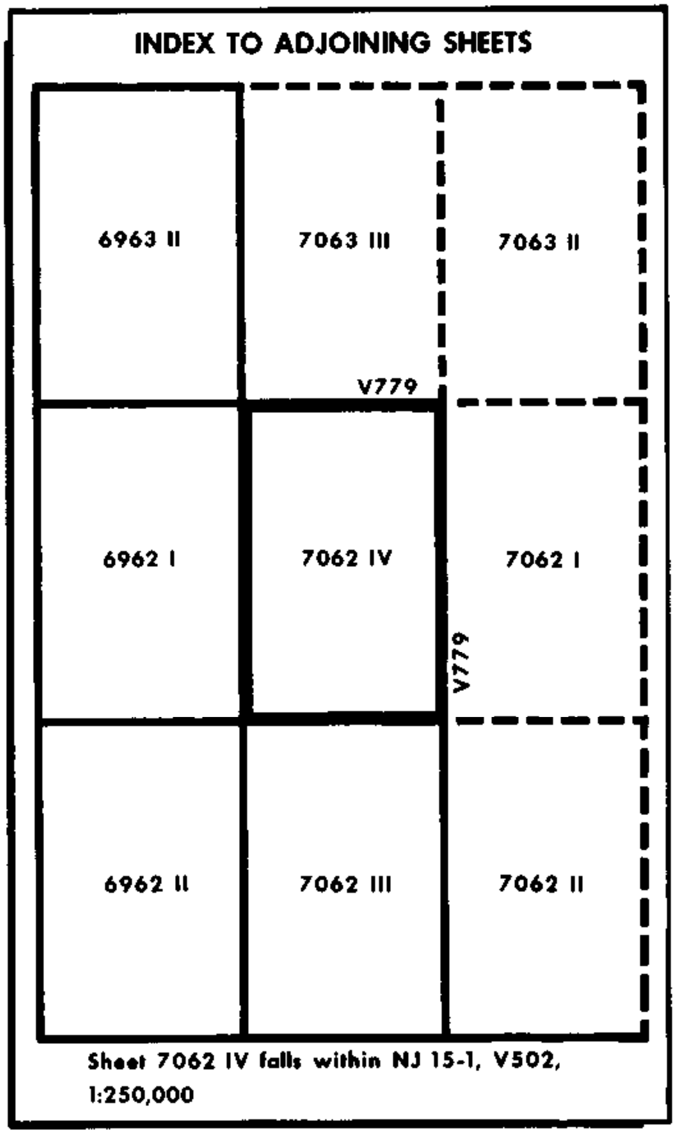
INDEX TO ADJOINING SHEETS
6963 II
7063 III
7063 II
V779
6962 I
7062 IV
7062 I
V779
6962 II
7062 III
7062 II
Sheet 7062 IV falls within NJ 15-1, V502, 1:250,000

Figure 20-16. Index to Adjoining Sheets.

map, and on recently printed sheets, the edition number.

20-5. Topographic Map Symbols and Colors:

a. The purpose of a map is to permit one to visualize an area of the Earth's surface with pertinent features properly positioned. Ideally, all the features within an area would appear on the map in their true proportion, position, and shape. This, however, is not practical because many of the features would be unimportant and others would be unrecognizable because of their reduction in size. The mapmaker has been forced to use symbols to represent the natural and manmade features of the Earth's surface. These symbols resemble, as closely as possible, the actual features as viewed from above (figures 20-17 and 20-18).

b. To facilitate identification of features on the map by providing more natural appearance and contrast, the topographic symbols are usually printed in different colors, with each color identifying a class of features. The colors vary with different types of maps, but on a standard large-scale topographic map, the colors used and the features represented are:

(1) Black—the majority of cultural or manmade features.

(2) Blue—water features such as lakes, rivers, and swamps.

(3) Green—vegetation such as woods, orchards, and vineyards.

(4) Brown—all relief features such as contours.

(5) Red—main roads, built-up areas, and special features.

(6) Occasionally, other colors may be used to show special information. (These, as a rule, are indicated in the marginal information. For example, aeronautical symbols and related information for air-ground operations are shown in purple on JOGs.)

c. In the process of making a map, everything must be reduced from its size on the ground to the size which appears on the map. For purposes of clarity, this requires some of the symbols to be exaggerated. They are positioned so that the center of the symbol remains in its true location. An exception to this would be the position of a feature adjacent to a major road. If the width of the road has been exaggerated, then the feature is moved from its true position to preserve its relation to the road.

d. Army Field Manual 21-31 gives a description of topographic symbols and abbreviations authorized for use on US military maps. Figure 20-19 illustrates several of the symbols used on maps.

20-6. Coordinate Systems. The intersections of reference lines help to locate specific points on the Earth's surface. Three of the primary reference line systems are the geographic coordinate system, the reference (GEOREF) system, and the universal transverse mercator grid system (UTM). Knowing how to use these plotting systems should help a survivor to determine point locations.

a. Coordinates. Quantities that give position with respect to two reference lines are called coordinates. Thus, the intersection of F Street and 4th Avenue (figure 20-20) is the coordinate location of the Gridville Public Library. The coordinates of the local theater are D Street and 6th Avenue. One can see from this simplified example that coordinates are read at intersections of vertical and horizontal lines. The basic coordinate system used on maps and charts is the geographic military grid. The structure and use of the geographic coordinate system, the world geographic reference system, and the military grid reference system will be discussed and illustrated.

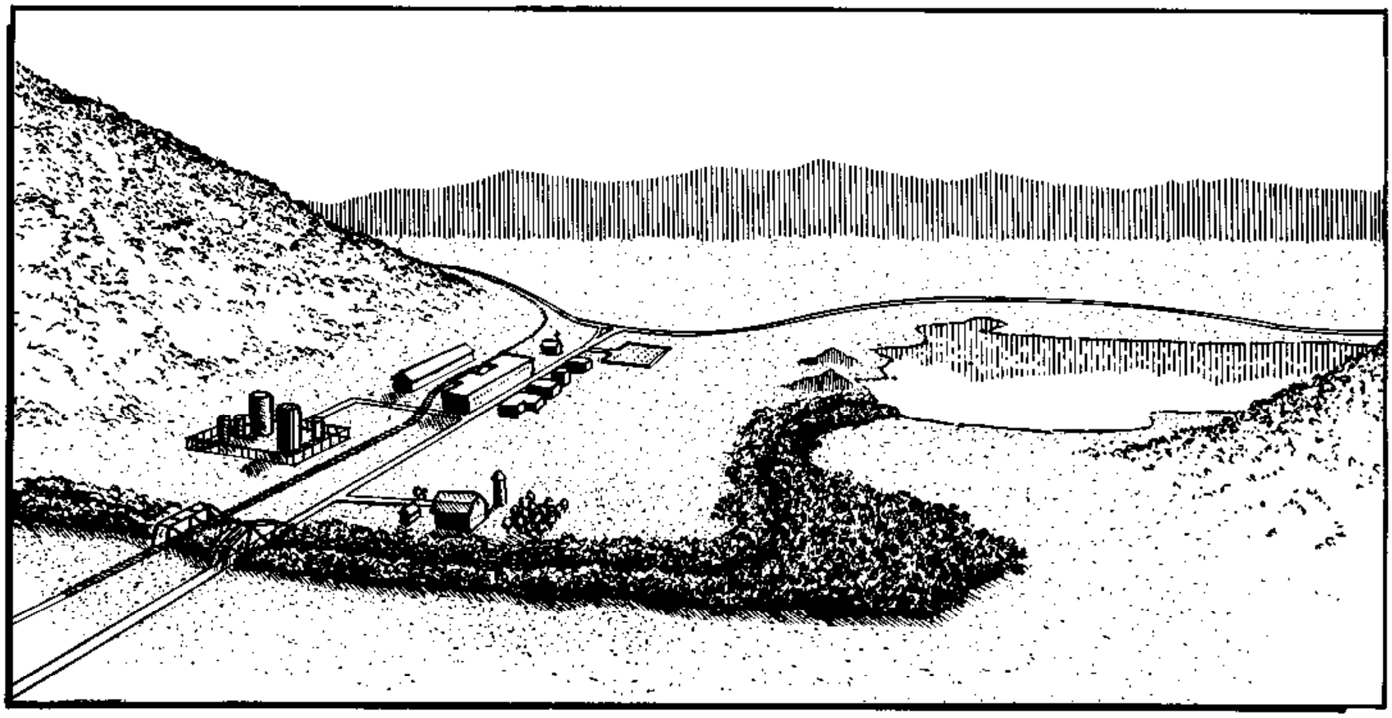

Figure 20-17. Area Viewed from Ground Position.

(1) Geographic Coordinates. The geographic coordinate system is a network of imaginary lines that circle the Earth. They are used to express Earth position or location. There are north-south lines called meridians of longitude and east-west lines named parallels of latitude. The location of any point on the Earth can be expressed in terms of the intersection of the line of latitude and the line of longitude passing through the point.

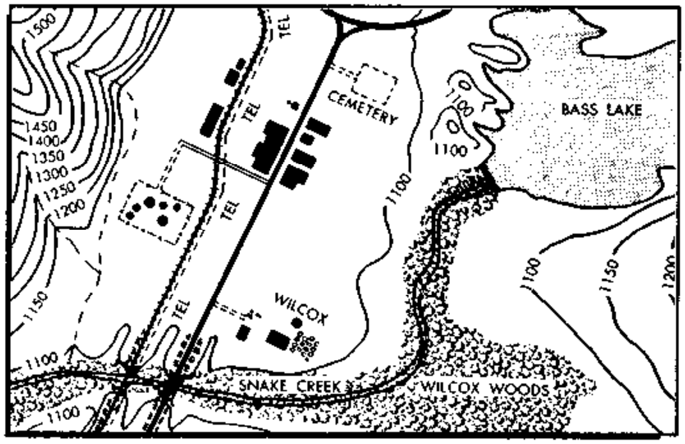

Figure 20-18. Area Viewed from Ground Position—Map.

(2) Meridians of Longitude. The lines of latitude and longitude are actually great and small circles formed by imaginary planes cutting the Earth. A great circle divides the Earth into two equal parts (halves); whereas, a small circle divides the Earth into two unequal parts. Study figure 20-21 and note that: (1) each north-south line is a great circle, and (2) each great circle passes through both the North and South Poles. Each half of each of these great circles from one pole, in either direction, to the other pole is called a meridian of longitude. The other half of the same great circle is a second meridian of longitude.

(a). Meridian is derived from the Latin word "meridianum," which means "lines that pass through the highest point on their course" (in this case, both the North and South Poles). Any angular distance measured east or west of the meridian is called longitudinal distance; hence, the term "meridian of longitude." It is necessary, of course, to assign values to the meridians to make them meaningful. The most appropriate values to use for circles are degrees (°), minutes (′), and seconds (″). Circles are customarily divided into 360° per circle, 60′ per degree, and 60″ per minute.

(b). All meridians are equal in value; hence, one of them must be assigned the value of 0° (the starting point). The meridian passing through Greenwich, England, is zero degrees (0°). This meridian is also called the prime meridian (figure 20-22). The other half of the great circle on which the prime meridian is located is designated the 180th meridian. Portions of this meridian are also called the international dateline.

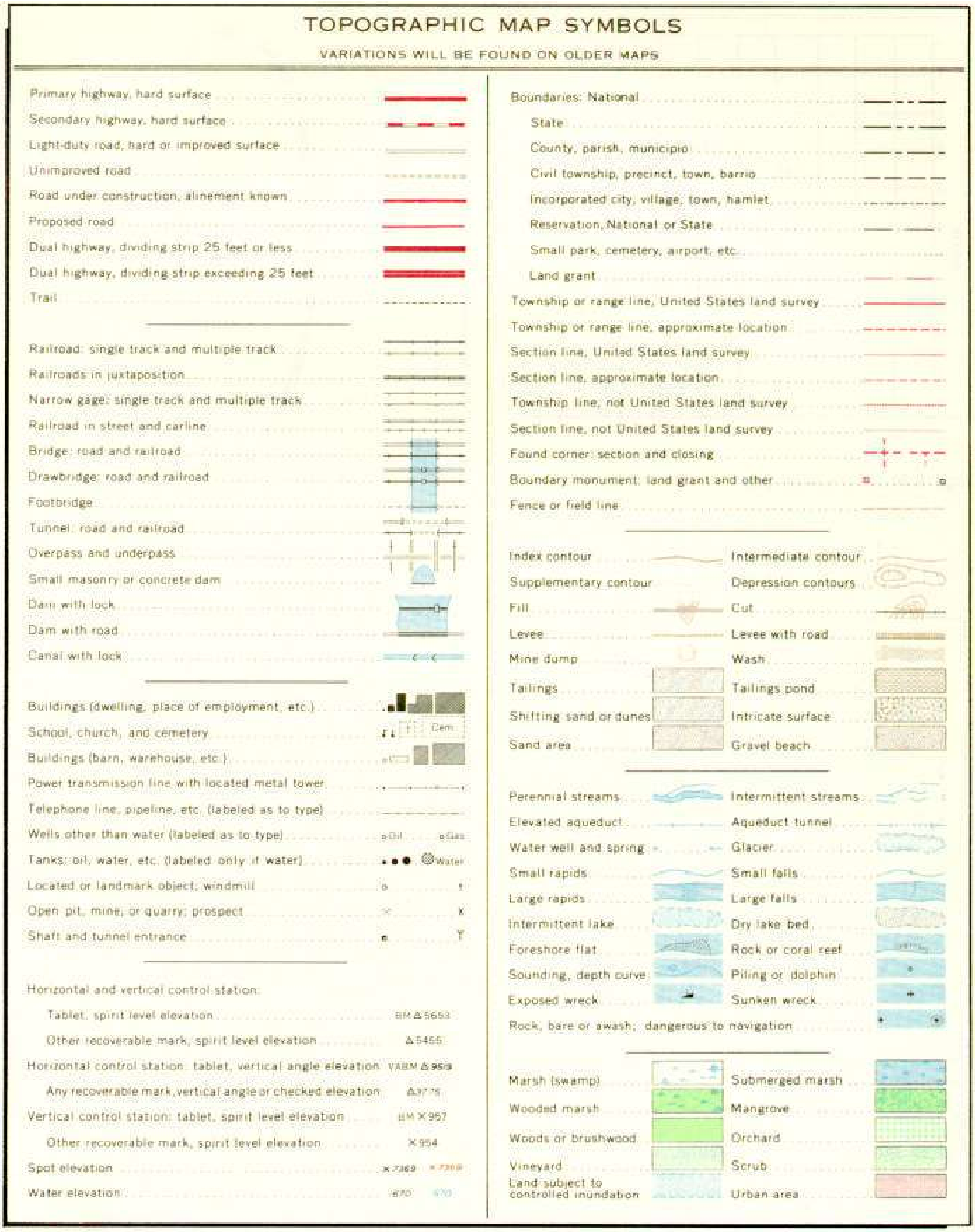
TOPOGRAPHIC MAP SYMBOLS
VARIATIONS WILL BE FOUND ON OLDER MAPS
Primary highway, hard surface
Secondary highway, hard surface
Light-duty road, hard or improved surface
Unimproved road
Road under construction, alinement known
Proposed road
Dual highway, dividing strip 25 feet or less
Dual highway, dividing strip exceeding 25 feet
Trail
Railroad: single track and multiple track
Railroads in juxtaposition
Narrow gage: single track and multiple track
Railroad in street and carline
Bridge: road and railroad
Drawbridge: road and railroad
Footbridge
Tunnel: road and railroad
Overpass and underpass
Small masonry or concrete dam
Dam with lock
Dam with road
Canal with lock
Buildings (dwelling, place of employment, etc.)
School, church, and cemetery
Cem
Buildings (barn, warehouse, etc.)
Power transmission line with located metal tower
Telephone line, pipeline, etc. (labeled as to type)
Wells other than water (labeled as to type)
Oil
Gas
Tanks: oil, water, etc. (labeled only if water)
Water
Located or landmark object; windmill
Open pit, mine, or quarry; prospect
Shaft and tunnel entrance
Horizontal and vertical control station:
Tablet, spirit level elevation
BM Δ5653
Other recoverable mark, spirit level elevation
Δ5455
Horizontal control station: tablet, vertical angle elevation
VABM Δ9519
Any recoverable mark, vertical angle or checked elevation
Δ3775
Vertical control station: tablet, spirit level elevation
BM ×957
Other recoverable mark, spirit level elevation
×954
Spot elevation
×7369
×7369
Water elevation
670
670
Boundaries: National
State
County, parish, municipio
Civil township, precinct, town, barrio
Incorporated city, village, town, hamlet
Reservation, National or State
Small park, cemetery, airport, etc.
Land grant
Township or range line, United States land survey
Township or range line, approximate location
Section line, United States land survey
Section line, approximate location
Township line, not United States land survey
Section line, not United States land survey
Found corner: section and closing
Boundary monument: land grant and other
Fence or field line
Index contour
Intermediate contour
Supplementary contour
Depression contours
Fill
Cut
Levee
Levee with road
Mine dump
Wash
Tailings
Tailings pond
Shifting sand or dunes
Intricate surface
Sand area
Gravel beach
Perennial streams
Intermittent streams
Elevated aqueduct
Aqueduct tunnel
Water well and spring
Glacier
Small rapids
Small falls
Large rapids
Large falls
Intermittent lake
Dry lake bed
Foreshore flat
Rock or coral reef
Sounding, depth curve
Piling or dolphin
Exposed wreck
Sunken wreck
Rock, bare or awash; dangerous to navigation
Marsh (swamp)
Submerged marsh
Wooded marsh
Mangrove
Woods or brushwood
Orchard
Vineyard
Scrub
Land subject to controlled inundation
Urban area

Figure 20-19. Topographic Map Symbols.

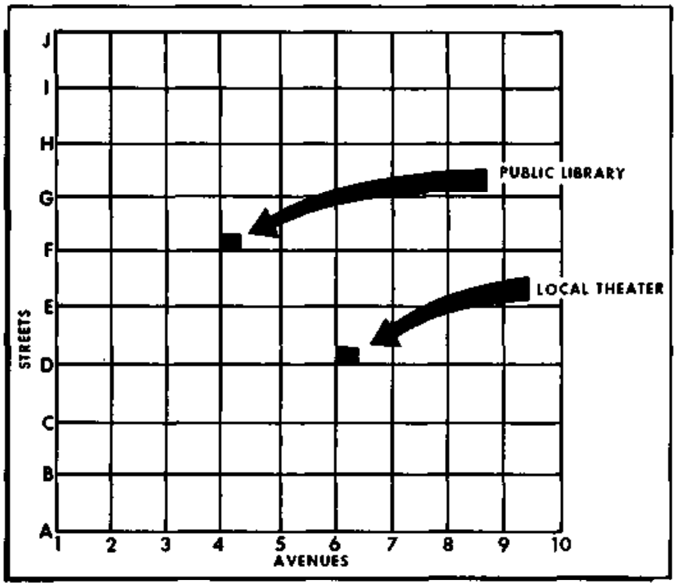
J
I
H
G
F
E
D
C
B
A
STREETS
PUBLIC LIBRARY
LOCAL THEATER
1
2
3
4
5
6
7
8
9
10
AVENUES

Figure 20-20. Gridville City.

(c). From the prime meridian east of the international dateline, meridians are assigned values of 0° through 180° east. Similarly, from the prime meridian west to the international dateline, meridians are assigned values of 0°through 180° west. The 0° meridian together with the 180° meridian forms a great circle which divides the Earth into east and west longitude (or hemispheres). There are 180° of east longitude plus 180° of west longitude for 360° of longitude.

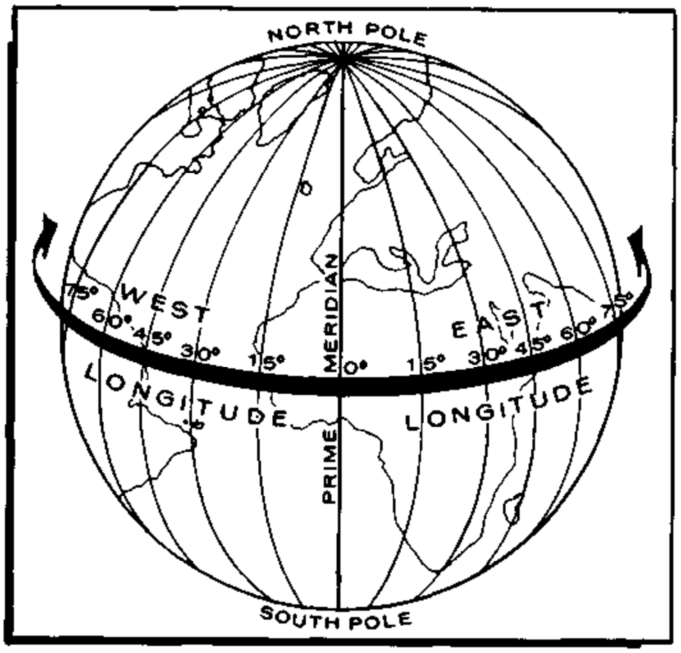
NORTH POLE
WEST
75°
60°
45°
30°
15°
0°
15°
30°
45°
60°
75°
EAST
LONGITUDE
LONGITUDE
MERIDIAN
PRIME
SOUTH POLE

Figure 20-21. Meridian of Longitude.

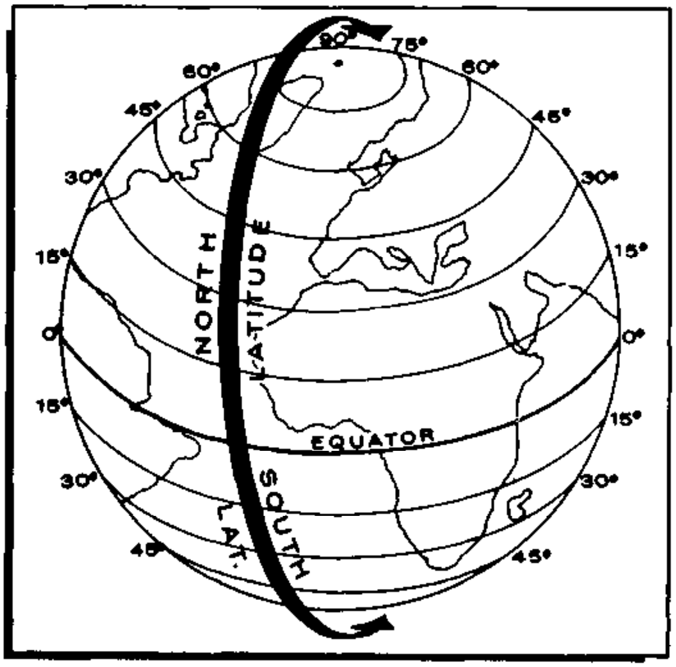
90°
75°
60°
60°
45°
45°
30°
30°
15°
15°
0°
0°
15°
15°
30°
30°
45°
45°
NORTH
LATITUDE
EQUATOR
SOUTH
LAT.

Figure 20-22. Parallels of Latitude.

b. Parallels of Latitude. Notice in figure 20-22 that the circles running in an east-west direction are of varying diameters (sizes). Only the circle designated "Equator" is a great circle. All others are small circles. Note that all circles are parallel to the Equator and run laterally around the Earth. Hence, each circle is called a parallel of latitude. Unlike meridians, which extend only halfway around the Earth, a parallel of latitude extends all the way around the Earth; for the record, the Equator is also a parallel of latitude. Since the Equator is the only great circle of latitude, it is a natural starting point for the 0° value of latitude. The North and South Poles are designated 90° north latitude and 90° south latitude, respectively. Parallels between the Equator and North Pole carry values between 0° and 90°north; parallels between the Equator and the South Pole are assigned values between 0° and 90° south.

(1) Figure 20-23 combines the lines of latitude and longitude. Lines 0°through 90° north or south latitude and 0° through 180° east or west longitude form the grid of the geographic coordinate system. Study the positions of Points A and B in figure 20-23. Determine the geographic coordinates of each in degrees. Note that point A is positioned 32° north of the Equator and 35° east of the prime meridian. The geographic position of point A, therefore, is 32°north 35° east. Point B is located 25° south of the Equator and 40° west of the prime meridian. Hence, the geographic position of point B is 25° south 40°west.

(2) Just as any point within the city of Gridville (figure 20-20) can be referenced by the intersection of

two imaginary lines, any point on the Earth's surface can be referenced by the intersection of the imaginary lines of latitude and longitude.

c. Writing Geographic Coordinates. To illustrate the proper way to write geographic coordinates, let's assume that a person needs to write the coordinates of a target. The target is located 30°20′ north of the Equator and 135°06′ east of the prime meridian. Thus, the position is located at 30°20′ north latitude and 135°06′ east longitude. By combining latitude and longitude, the position of the geographic location can be expressed as 30°20′N 135°06′E. To write these coordinates in the correct military form, eliminate the degree (°) and minute (′) symbols. Thus, the coordinates would be written 302000N1350600E.

(1) Writing geographic coordinates in the military form is necessary for wire and radio transmission of geographic coordinates. Why? The transmission equipment does not include the degree (°), minute (′), and second (″) characters in its keyboards. Coordinates are also stored in automated data processing computers which are programmed to handle coordinates in military characters or spaces. If the sequence of numbers and letters fed into a computer is less than 15 spaces, or in error, the resulting printout will be meaningless.

(2) When a position is located that is less than 10° latitude, a zero is added to the left of the degree number. For example, 7° of latitude is written as 07. Likewise, two digits always designate minutes and two digits for seconds. Thus, 7°N becomes 07N; 7°6′N becomes 0706N; and 7°6′5″N becomes 070605N. In expressing longitude, three digits are required to indicate degrees, two digits for minutes, and two digits for seconds. Thus 8°E becomes 008E; 8°5′E becomes 00805E; and 8°5′4″E becomes 0080504E.

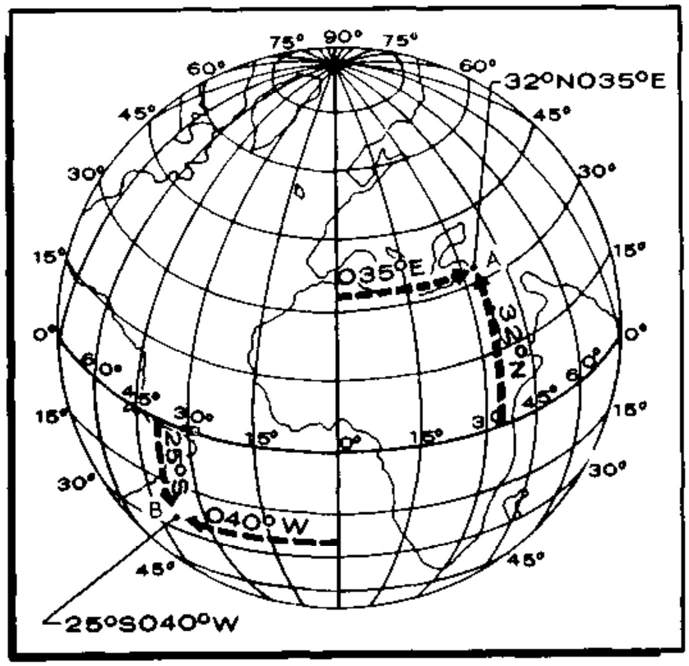

Figure 20-23. Latitudes and Longitudes.

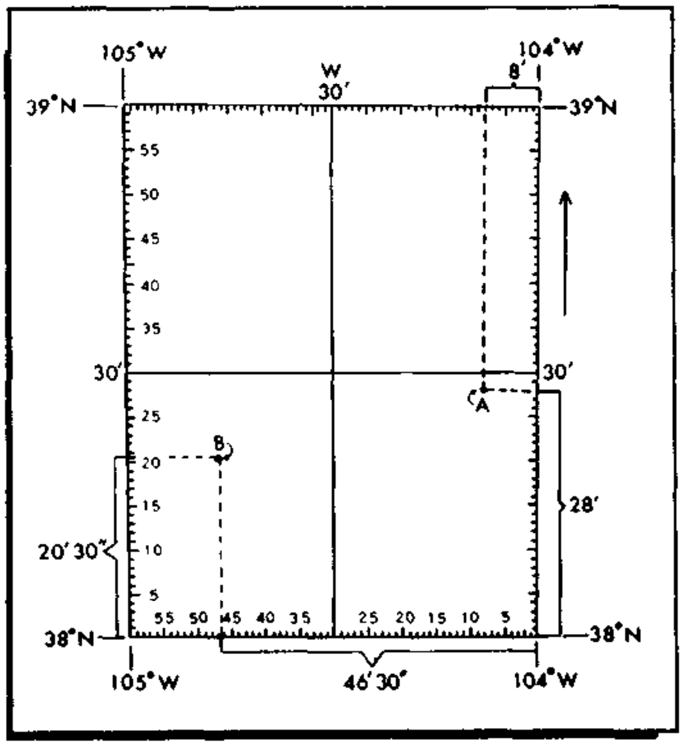

Figure 20-24. Plotting Geographic Coordinates.

(3) In general, there are five rules to follow in correctly writing geographic coordinates:

(a) Write latitude first, followed by longitude.

(b) Use an even number of digits for latitude and an odd number of digits for longitude.

(c) Do not use a dash or leave a space between latitude and longitude.

(d) Use single upper case letter to indicate direction from the Equator and prime meridians.

(e) Omit the symbols for degrees, minutes, and seconds.

d. Plotting Geographic Coordinates:

(1) One can probably read the coordinates of point A and B in figure 20-24 rather easily; however, plotting points on maps from given coordinates must also be done. To do this, first get acquainted with the map being used. Assume that figure 20-24 is the map being used. Note that it covers an area from 38N to 39N and from 104W to 105W, an area of 1° by 1°. Also note that latitude and longitude are subdivided by 30′ division lines and then with tick marks into 5- and 1-minute subdivisions.

(2) Assume that the coordinates of the point which must be plotted are 382800N1040800W. Next, follow the general procedure listed below to plot the point on the map:

(a) Locate the parallel of latitude for degrees (38°N).

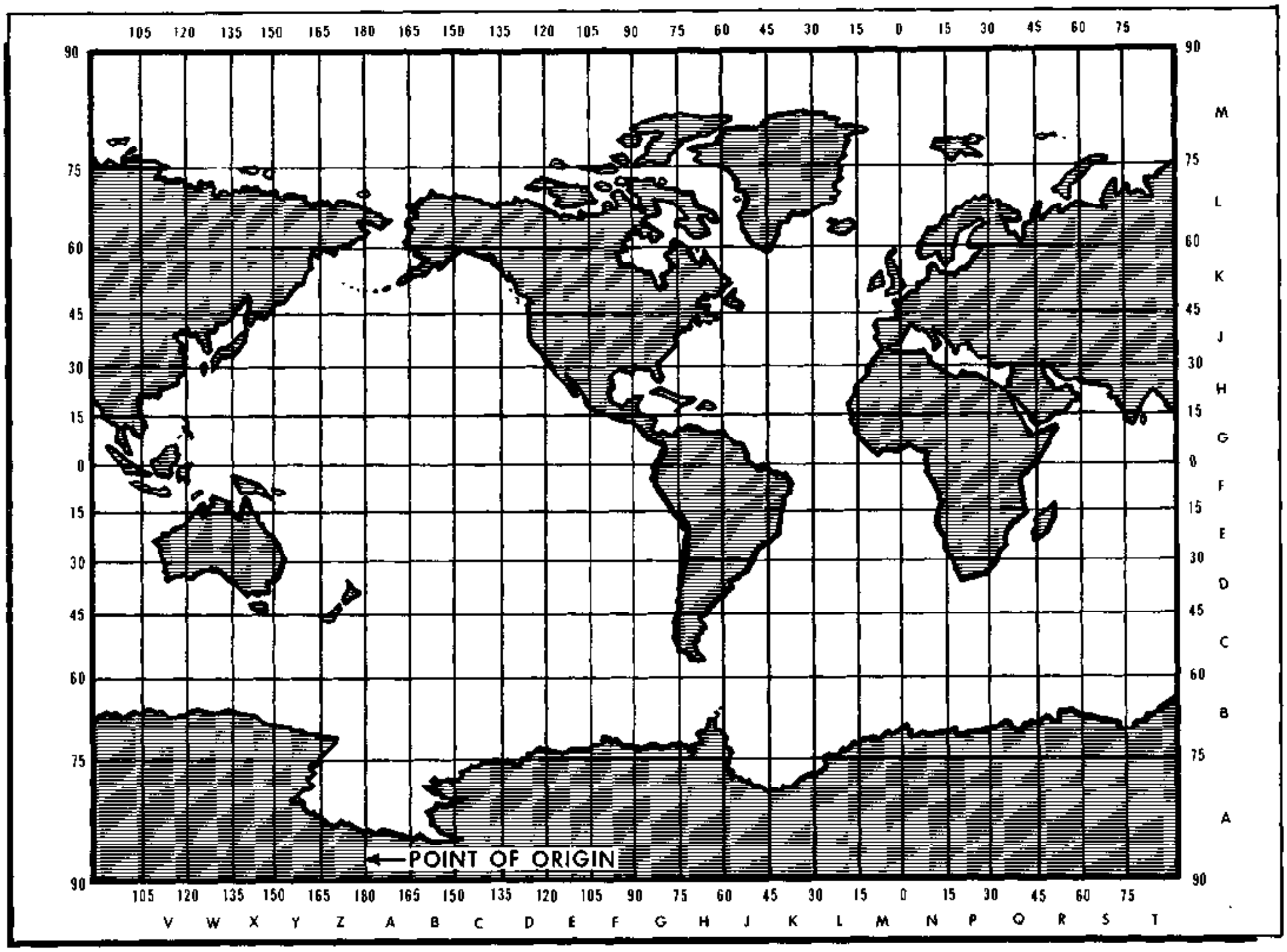

Figure 20-25. GEOREF 15-Degree Quadrants.

(b) Find the meridian of longitude for degrees (104°W).

(c) Move to the meridian (usually a tick mark) for minutes (08°W).

(d) Move to the parallel (usually a tick mark) for minutes (28'N).

(e) Plot the point on the map (point A in figure 20-24; plot at 382800N1040800W.

(3) Recovery points, rally points, and destination positions may be plotted or identified on a map or chart to enable rescue personnel, the survivors, and evaders to locate these positions. Seconds are not shown between the 1-minute tick marks on maps and charts; they must be estimated. It is easy to estimate halfway tick marks (30 seconds); one-fourth (15 seconds) and three-fourths (45 seconds) are also reasonably easy to estimate. Then, as experience is gained, people will find that on large-scale maps they can estimate the sixths (10 seconds) and eights (about 8 seconds). They cannot, however, accurately estimate to sixths or eights at the scale shown in figure 20-24.

(4) To write geographic coordinates more precisely than minutes, merely carry the coordinates out to include seconds. In the previous example, the coordinates of a target located 30°20′ north of the Equator and 135°06′ east of the prime meridian were written as 302000N1350600E. A more exact position of the target might be 30°20′05″N latitude and 135°06′16″E longitude. This more precise position is correctly written as 302005N1350616E.

e. World Geographic Reference System (GEOREF). The geographic coordinate system has several shortcomings when it is used in military operations. One objectionable feature is the large number of characters necessary to identify a location. To specify a location within 300 yards, a coordinate reading such as 241412NO141512W is necessary, with a total of 15 characters. Another objectionable feature is the diversity of directions used in applying the grid numbering system. Any particular point on a geographic grid can be north and east, north and west, south and east, or south and west. This means there are four different ways to proceed when reading various geographic coordinates.

Such a system obviously promotes errors. To overcome the disadvantages and promote speed in position reporting, other grid systems are used. We shall now examine one of these systems—that which is commonly called GEOREF. Air Force uses the GEOREF system as a reference in the control and direction of forces engaged in large area operations and operations of a global nature.

(1) GEOREF System Structure. The geographic coordinate grid serves as the base for the GEOREF system. The grid originates at the 180° meridian and the South Pole. Starting at the 180° meridian, it proceeds right, or eastward, around the world, and back to the 180° starting point. From the South Pole, it proceeds northward to the North Pole (figure 20-25).

(a) Notice in figure 20-25 that the basic layout is subdivided into 24 east-west zones and 12 north-south zones. This forms 288 quadrangles that measure 15°X15°. The 24 east-west zones are lettered "A" through "Z" omitting "I" and "O." The 12 north-south zones are lettered "A" through "M" (omitting "I"). Each quadrangle is identified by two letters and is located by reading right and up. For example, the southern tip of Florida is located in GEOREF quadrangle G-H (figure 20-25).

(b) Each of the 15° quadrangles is divided into 1° quadrangles (figure 20-26). First, they are divided to the right into 15 zones lettered "A" through "Q" (omitting "I" and "O"), then up into 15 zones lettered "A" through "Q" ("I" and "O" omitted).

(c) This system makes it possible to identify any quadrangle by four letters; for example, WGAN. The two letters designate the 15° grid zone, and the other two letters identify a 1° quadrangle within the 15° grid zone. In figure 20-27, WGAN refers to the quadrangle situated between 120° east longitude and 12° and 13°north latitude. Notice that the 1° quadrangle WGAN is further divided by 30-minute division lines, and then with tick marks into 5- and 1-minute subdivisions.

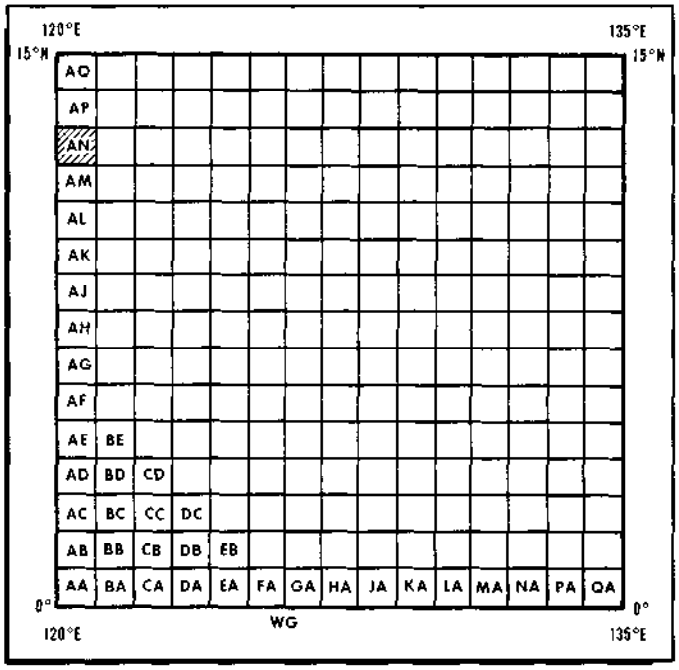
120°E
135°E
15°N
15°N
AQ
AP
AN
AM
AL
AK
AJ
AH
AG
AF
AE
BE
AD
BD
CD
AC
BC
CC
DC
AB
BB
CB
DB
EB
AA
BA
CA
DA
EA
FA
GA
HA
JA
KA
LA
MA
NA
PA
QA
0°
0°
WG
120°E
135°E

Figure 20-26. GEOREF 1-Degree Quadrants.

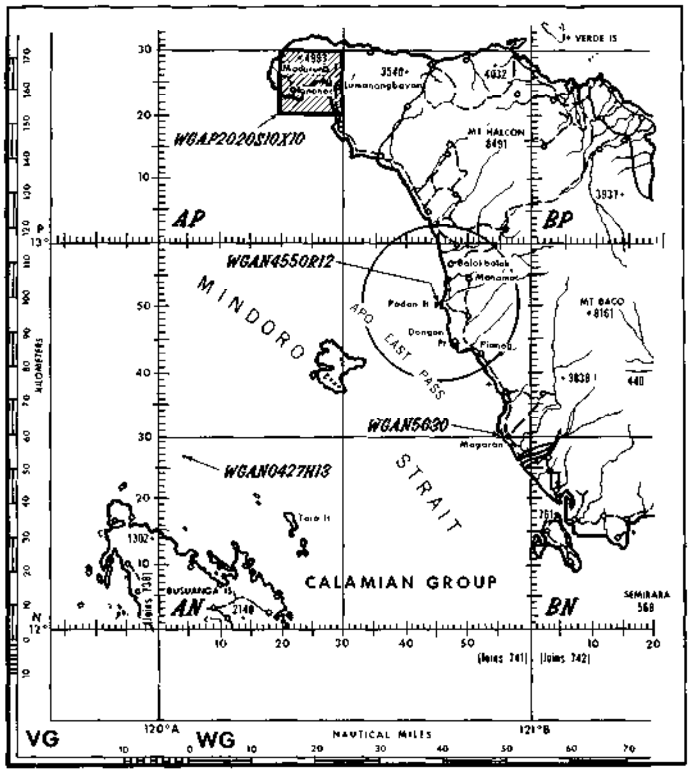
VERDE IS
WGAP2020S10X10
MT HALCON
AP
BP
13°
WGAN4550R12
MINDORO
APO EAST PASS
MT BACO
WGAN5630
Magaran
WGAN0427H13
STRAIT
CALAMIAN GROUP
BUSUANGA IS
AN
BN
SEMIRARA
12°
10
20
30
40
50
10
20
VG
120°A
WG
NAUTICAL MILES
121°B
KILOMETERS

Figure 20-27. GEOREF 1-Degree Quadrants WGAN.

(2) GEOREF Coordinates:

(a) Any feature within a 1° quadrangle can be located by reading the number of minutes to the right and the number of minutes up. For example, the city of Magaran (figure 20-27) can be located by proceeding as follows:

-1. 15° X 15° quadrangle identification	WG
-2. 1° X 1° quadrangle identification	WGAN
-3. Minutes to the right	WGAN 56
-4. Minutes up	WGAN 5630
-5. Full GEOREF coordinate	WGAN 5630

(b) If a reference of greater accuracy than 1 minute is required, the 1-minute tick marks may be divided into decimal values (tens or hundreds). By doing this, it is possible to locate a point within one-tenth of a minute with four letters and six numbers and within one-hundredth of a minute by four letters and eight numbers.

(3) GEOREF Special References. Another real advantage of the GEOREF system is the simplicity with

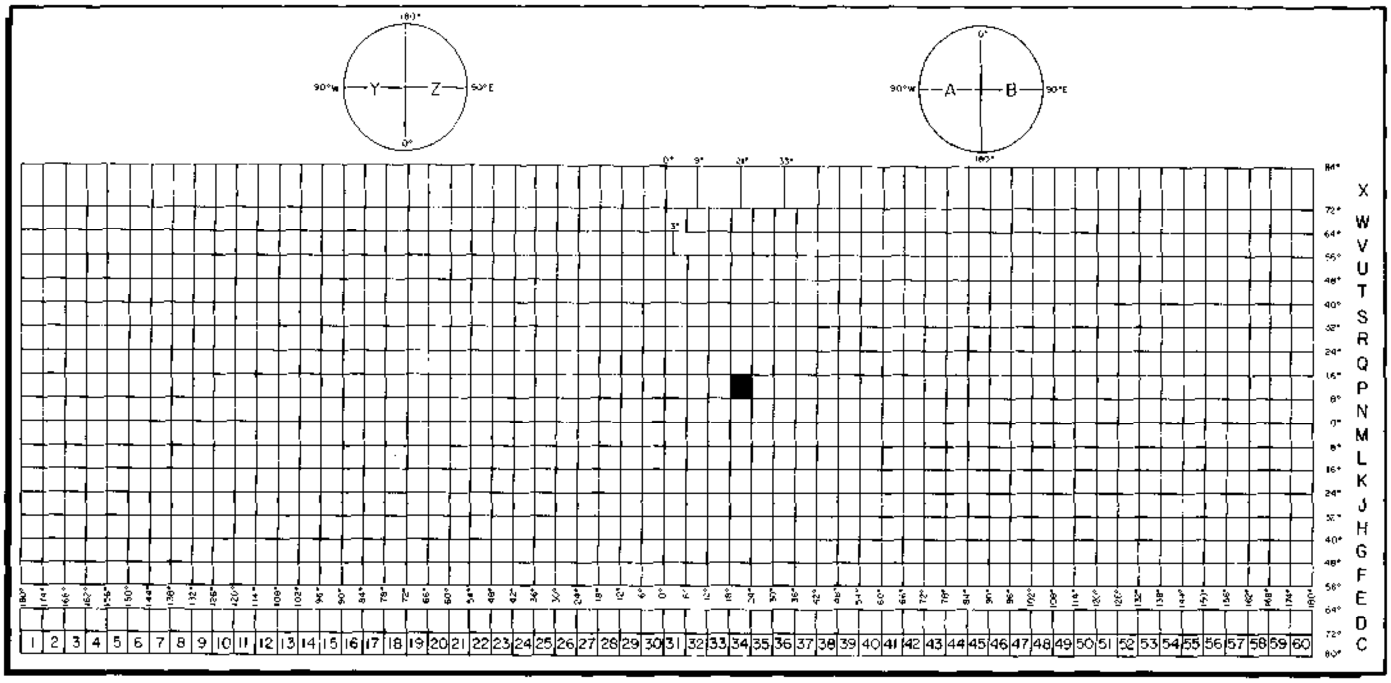
180°
90°W
Y
Z
90°E
0°
0°
90°W
A
B
90°E
180°
0°
9°
21°
33°
3°
84°
72°
64°
56°
48°
40°
32°
24°
16°
8°
0°
8°
16°
24°
32°
40°
48°
56°
64°
72°
80°
X
W
V
U
T
S
R
Q
P
N
M
L
K
J
H
G
F
E
D
C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

Figure 20-28. UPS Grid Zones.

which it allows a person to designate an area or indicate the elevation of a point. To designate the location and dimensions of a rectangular area, first read the GEOREF coordinates of the southwest corner of the area. Then add and "S," which denotes "side," and digits denoting the number of nautical miles that the area extends to the east. Then add an "X," denoting "by," and digits denoting the number of nautical miles the area extends to the north. An example of such a reference is WGAP2020S10X10 (figure 20-27). Circular areas are designated in much the same manner. First, read the GEOREF coordinates of the center of the area. Then add an "R," denoting radius, and digits defining the radius in nautical miles. This is also illustrated in figure 20-27 as WGAN4550R12.

(4) Military Grid Reference System. A grid is a rectangular coordinate system superimposed on a map. It consists of two sets of equally spaced parallel lines that are mutually perpendicular and form a pattern of squares. Some maps carry more than one grid. In such cases, each grid is shown in a different color or is otherwise distinguished. The military grid reference system is comprised of two grid systems. The US Army adopted the universal transverse mercator (UTM) grid for areas between 80° south latitude and 84°north latitude. For the polar caps, areas below 80° south latitude and above 84° north latitude, the universal polar stereographic (UPS) grid was adopted. The unit of measurement for the UTM and UPS grids are the meter, but the interval at which the grid lines are shown on the maps depends upon the scale.

(5) The UTM Grid System. In the UTM system, the surface of the Earth is divided into large quadrilateral grid zones (figure 20-28). Beginning at the 180th meridian, 6° columns are numbered 1 through 60 eastward with each column broken down into rows. From 80° south through 72° north, each row is 8° south-north.

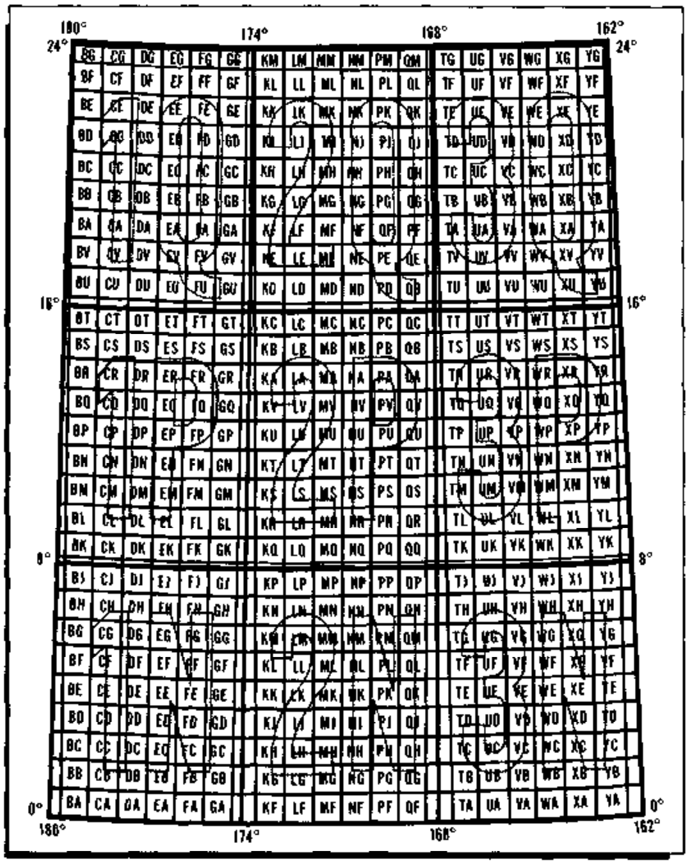
180°
174°
168°
162°
24°
24°
16°
16°
8°
8°
0°
0°
180°
174°
168°
162°

Figure 20-29. UTM Grid Zones.

The top row, 72° to 84° north is 12° south-north. The south-north rows are lettered "C" through "X" (omitting "I" and "O") as shown in figure 20-28. The grid zones are located by reading right and up. For instance, right to column 34 and up to "P" locates grid zone 34P, which is the shaded grid zone of figure 20-28. The UPS grid zones covering the polar areas are designated by a single "A," "B," "Y," or "Z" (figure 20-28).

(a) Each UTM grid zone is divided into columns and rows to form small squares measuring 100,000 meters on each side and are called 100,000-meter squares. Each square is identified with two letters. The columns are lettered "A" through "Z" (omitting "I" and "O"), starting at the 180th meridian and progressing eastward around. The 24 letters are repeated every 18° (figure 20-29). Starting at the Equator, the horizontal row 100,000-meter squares are lettered "A" through "V" (omitting "I" and "O") northward. From the Equator southward, the designation "V" through "A" is used. The letters are repeated periodically (figure 20-29).

(b) The Earth's meridians converge toward the poles. Therefore, the grid zones are not square or rectangular. The actual width of each grid zone decreases toward the poles. This condition causes partial squares to occur along the grid zones. In the far north and south latitudes, the grid zones become so narrow that 100,000-meter square designations may disappear completely. However, each full or partial 100,000-meter square within a grid zone is referenced with two letters. The first letter refers to the vertical column (left to right), and the second letter identifies the horizontal row (bottom to top). Thus, a grid zone designation plus two letters identifies or designates an area 100,000 meters on each side. Furthermore, as the UTM system is set up, no two squares with the same designation are included in a grid zone or on the same map sheet.

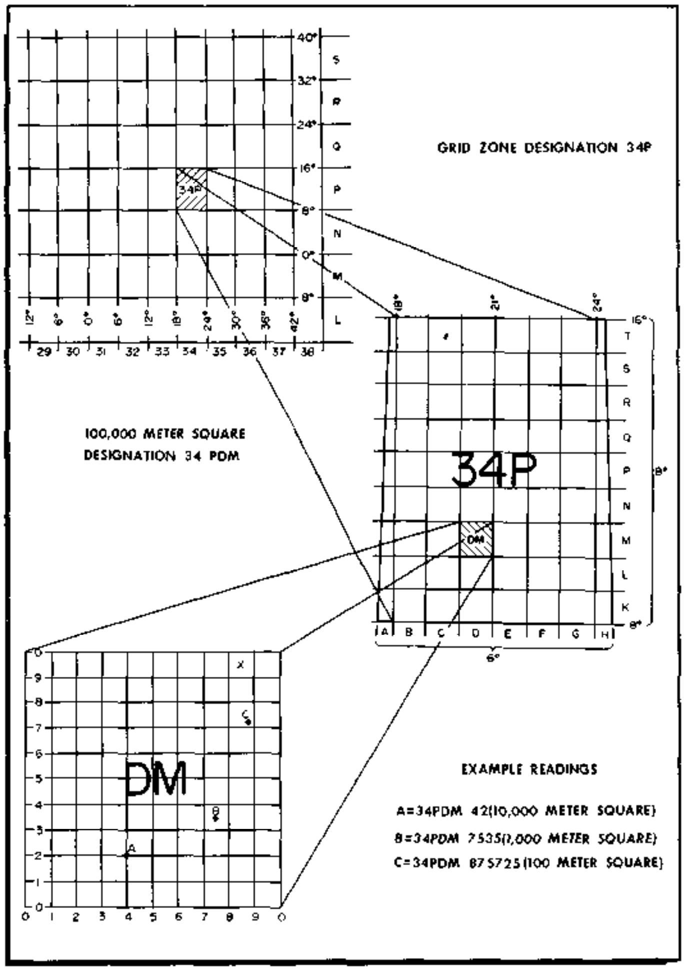

Figure 20-30. Plotting UTM Grid Coordinates.

(c) Observe grid zone 34P which is expanded in figure 20-30 to show the 100,000-meter squares. For grid zone 34P, the columns are designated "A" through "H," and the rows are designated "K" through "T" (omitting "O"). The left column begins with "A" because, as stated earlier, columns repeat the alphabet each 18°. The bottom row begins with "K" because "A" through "J" (omitting "I") was used up in the previous 8° of north latitude.

(d) Next, note the partial squares along the left and right sides of grid zone 34P (figure 20-30). Partial squares occur because the distance east and west of the central meridian of each grid zone does not contain an even number of 100,000-meter squares. The last squares, therefore, must terminate at the meridian junctions. In the north-south direction, partial 100,000-meter squares seldom occur.

(e) Figure 20-31 shows the grid reference box for a map or chart. Note the statement in the upper left corner of the grid reference box. It identifies the grid zones that are represented on the map sheet—52S and 53S. Thus, the full UTM coordinates of any point within the map area begins with either 52S or 53S.

(f) Still, it is not clear which area is 52S and which is 53S. Therefore, study the 100,000-meter square block identification located directly below the grid zone designation. From the diagram a person can see that everything to the left of the center meridian is grid zone 52S and everything to the right of the same meridian is 53S.

(g) Other important information given in the grid reference block includes: (1) sample reference and, (2) step-by-step procedures for locating or writing coordinates. Each time a new map is used, identify the sample point and write its UTM coordinates to ensure the grid breakdown for that map is understood.

(h) A troublesome and sometimes confusing situation exists where 100,000-meter squares fuse together along meridians separating grid zones. Remember, this happens every 6° around the world. Notice in figure 20-32 that the 100,000-meter squares GP and KJ are only partial squares, fusing along the center meridian (so are GQ, GN, KH, and KK). There are then full and

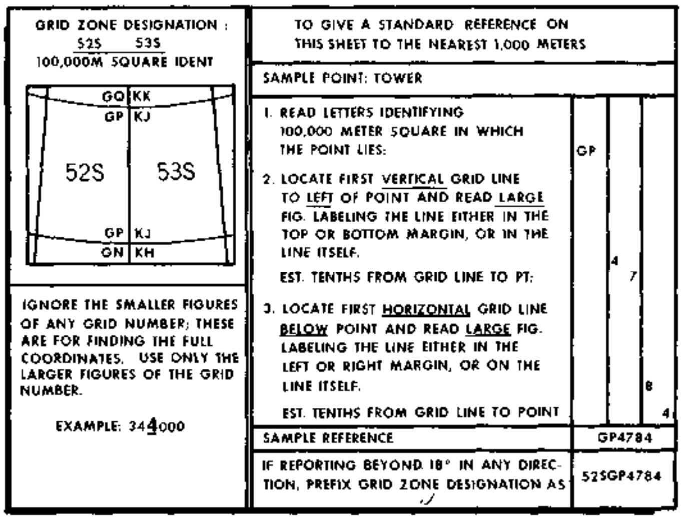
GQ
KK
GP
KJ
52S
53S
GP
KJ
GN
KH

GRID ZONE DESIGNATION :
52S 53S
100,000M SQUARE IDENT

IGNORE THE SMALLER FIGURES OF ANY GRID NUMBER; THESE ARE FOR FINDING THE FULL COORDINATES. USE ONLY THE LARGER FIGURES OF THE GRID NUMBER.

EXAMPLE: 344000

TO GIVE A STANDARD REFERENCE ON THIS SHEET TO THE NEAREST 1,000 METERS			
SAMPLE POINT: TOWER			
1. READ LETTERS IDENTIFYING 100,000 METER SQUARE IN WHICH THE POINT LIES:	GP		
2. LOCATE FIRST VERTICAL GRID LINE TO LEFT OF POINT AND READ LARGE FIG. LABELING THE LINE EITHER IN THE TOP OR BOTTOM MARGIN, OR IN THE LINE ITSELF.		4	
EST. TENTHS FROM GRID LINE TO PT:		7	
3. LOCATE FIRST HORIZONTAL GRID LINE BELOW POINT AND READ LARGE FIG. LABELING THE LINE EITHER IN THE LEFT OR RIGHT MARGIN, OR ON THE LINE ITSELF.			8
EST. TENTHS FROM GRID LINE TO POINT			4
SAMPLE REFERENCE	GP4784		
IF REPORTING BEYOND 18° IN ANY DIRECTION, PREFIX GRID ZONE DESIGNATION AS	52SGP4784		

Figure 20-31. UTM Grid Reference Box.

partial 10,000 meter-squares within GP and KJ. Column 7 of the GP is comprised of partial 10,000-meter squares; columns 8 and 9 are missing because of the forced fusing along the meridian; similarly, column 2 of KJ is partial; columns O and I are missing. There is no problem in reading coordinates with full 10,000-meter squares. The tower in GP (sample point in figure 20-32) is 47 right and 84 up (omitting the grid zone and 100,000 meter square designation. All partial 100,000-meter squares are full sized in a north-south dimension). Therefore, distances up are referenced as full squares. However, partial squares, which occur in an east-west dimension, are something less than 10,000 meters long. Points within such partial squares are referenced as if the omitted part were present. That is, each partial square is imaginarily expanded into a full-sized square for reference purposes.

(i) The city of Bergen is 40 up; Celle is 35 up. Celle is three-tenths of the horizontal distance between grid line 7 and grid line 8—if there were a grid line 8. Thus, Celle is 73 right, and its full coordinates are 52SGP7335. Bergen is eight-tenths of the distance (reading left to right) between grid line 2—if there were a grid line 2—and grid line 3. Thus Bergen is 28 right, its full coordinates are 53SKJ2840.

(j) Figure 20-32 depicts a UTM grid breakdown as it normally appears on 1:250,000 scale charts. The smallest physical square is 10,000 meters on each side. However, larger scale maps with grid squares of 1,000 or even 100 meters on each side may be used. If so, add the values for the smaller grid squares. As additional digits are added, more precise points on the Earth's surface can be located.

(6) The UPS Grid System. The UPS reference system is the companion system to the UTM system. It covers the area of the world above 84° north latitude and below 80° south latitude. The UPS has a similar divisional breakdown and is read or written like the UTM system.

(a) Figure 20-33 shows the arbitrarily assigned designations for the UPS system in the North and South Pole regions. Note that from the small circles of the figure, the polar area is divided into two grid zone divisions by the 180° and 0° meridians. The west longitude half is designated "grid zone A or Y." Also notice that no numbers are used with the letters to identify the grid zones.

(b) The two grid zones "A" and "B" of the South Pole are divided into 100,000-meter squares, as shown in the large circle of figure 20-33. Each square is identified by a two-letter designation, which is assigned so no duplication exists between the two grid zones. The letters "I" and "O" are omitted, and to avoid confusion with 100,000-meter squares in adjoining UTM zones, the letters D, E, M, N, V, and W are also omitted.

(c) The UPS system is also read right and then up. Thus, the shaded 100,000-meter square at 10 o'clock in figure 20-33 is identified as AQR. (Remember, no numbers are used in identifying the grid zone.) The shaded 100,000-meter square near the South Pole of the same drawing is identified as BBM.

(d) The UPS breakdown of the North Pole region is similar to the South Pole region. Conversion of figure 20-33 to fit the North Pole would require the following changes: Substitute grid zone letters "Y" for "A" and "Z" for "B," and interchange the 0° and 180° positions on the common parallel at 80° South latitude. Designation of the 100,000-meter squares for the North Pole region is shown in figure 20-33.

(e) If the map scale is sufficiently large, the 100,000-meter squares can be subdivided into smaller squares of 10,000 meters on each side. Then, the 10,000-meter squares can be divided into 1,000-or even 100-meter squares. However, there is rarely a requirement in the polar regions for such a large-scale chart.

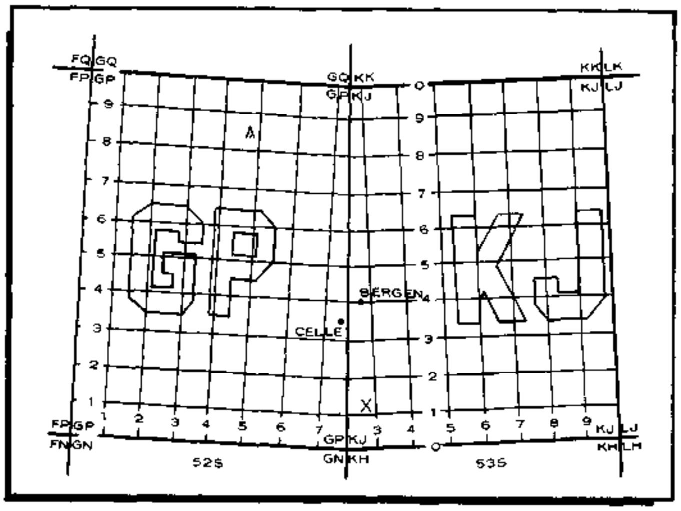

Figure 20-32. Fusion of Grid Zones 52S and 53S.

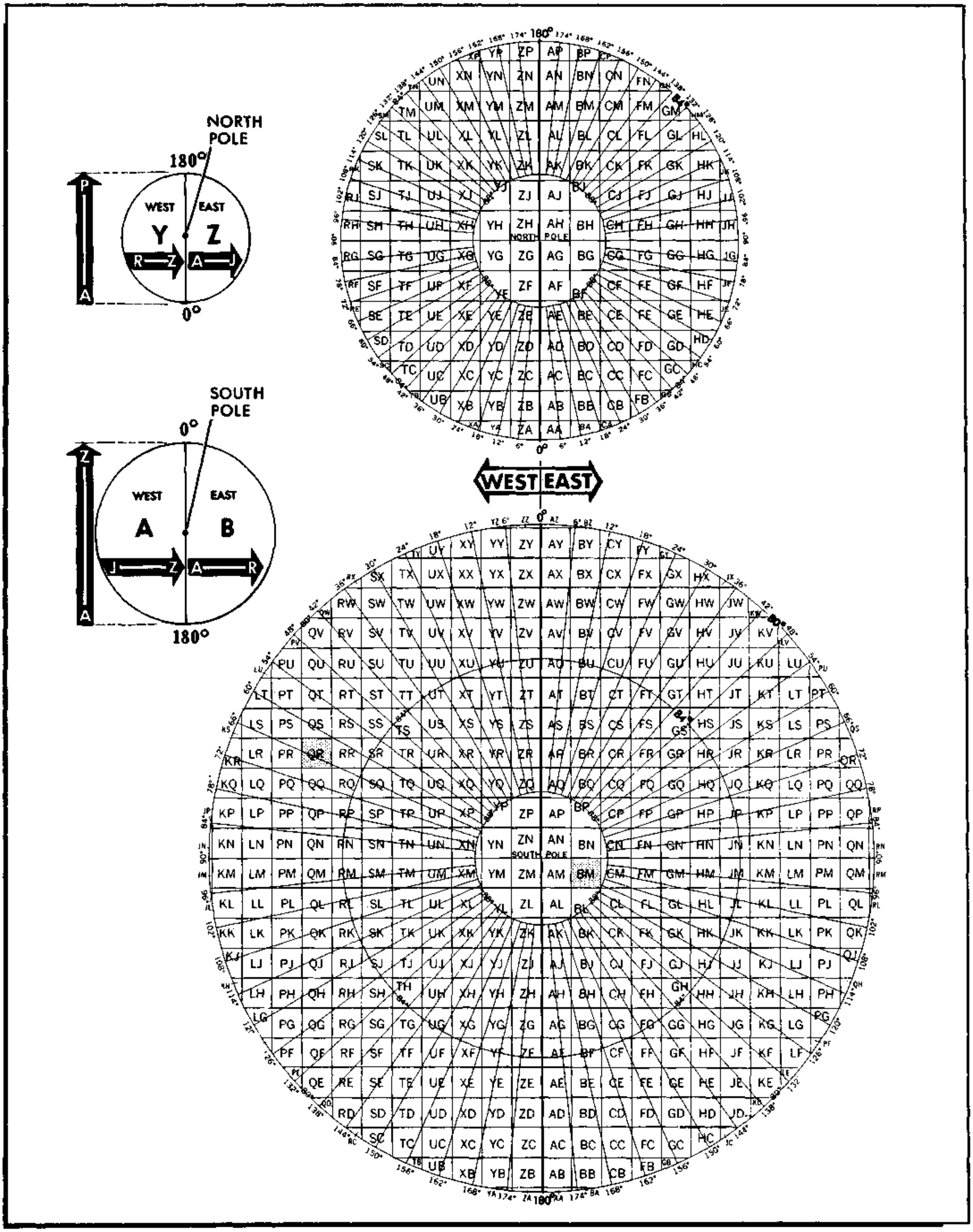
NORTH POLE
180°
0°
WEST
EAST
Y
Z
P
A
R—Z
A—J
SOUTH POLE
0°
180°
WEST
EAST
A
B
Z
A
J—Z
A—R
WEST
EAST
NORTH POLE
SOUTH POLE

Figure 20-33. UPS Grid Zones.

Generally, a person can expect to work with small-scale charts with the grid broken down no further than 100,000-meter squares.

(7) Public Land Survey. In the western part of the United States or in areas which were not settled before the Federal Government was formed, all land is laid out in rectangular survey as established by the Government. This public land survey is based on all land being divided in relation to true north. Public land surveys all originate from six or seven initial points which are exact locations of even latitude and longitude lines which have been established astronomically.

(a) From any one of the initial points a true north-south line, referred to as the principal meridian, is established. From the same point a true east-west line, referred to as the baseline, is established. Along this principal meridian and baseline are laid out 6-mile squares or townships. Each of these townships are numbered in relation to the initial point of survey. To the east and west of the initial point, the townships are designated by range numbers; to the north and south of the initial point, the townships are designated by township numbers. Therefore, township 2 north, range 3 east, would lie between 6 and 12 miles north of the initial point and between 12 and 18 miles east of the initial point.

(b) Each township contains 36 square miles and is divided into 36 sections. A section is 1 square mile or 640 acres. The section layout on townships is the same throughout the Public Land Survey. The sections are numbered in rows back and forth beginning in the upper right-hand corner of the township and ending in the lower right-hand corner (figure 20-34).

(c) Each section is divided into quarters or quarter sections of 160 acres each. These quarter sections are named by the compass location in relation to the section. The upper right-hand quarter section is referred to as the northeast ¼, the lower right-hand quarter is the southeast ¼, the lower left-hand quarter is the southwest ¼, and the upper left-hand quarter is the northwest ¼.

(d) Each quarter section is further subdivided into quarters or four blocks of 40 acres each known as forties. The forties are also located by the compass directions. In locating a particular piece of property, the 40-compass quadrant is given first, followed by the quarter section quadrant. Thus the SW-SE means the southwest 40 of the southeast quarter section. This is the basic unit of land management and, therefore, one should become familiar with the Public Land Survey and the means of locating specific pieces of property.

20-7. Elevation and Relief. A knowledge of map symbols, grids, scale, and distance gives enough information to identify two points, locate them, measure between them, and determine how long it would take to travel between them. But what happens if there is an obstacle between the two points? The map user must become proficient in recognizing various landforms and irregularities of the Earth's surface and be able to determine the elevation and differences in height of all terrain features.

a. Datum Plane. This is the reference used for vertical measurements. The datum plane for most maps is mean or average sea level.

b. Elevation. This is defined as the height (vertical distance) of an object above or below a datum plane.

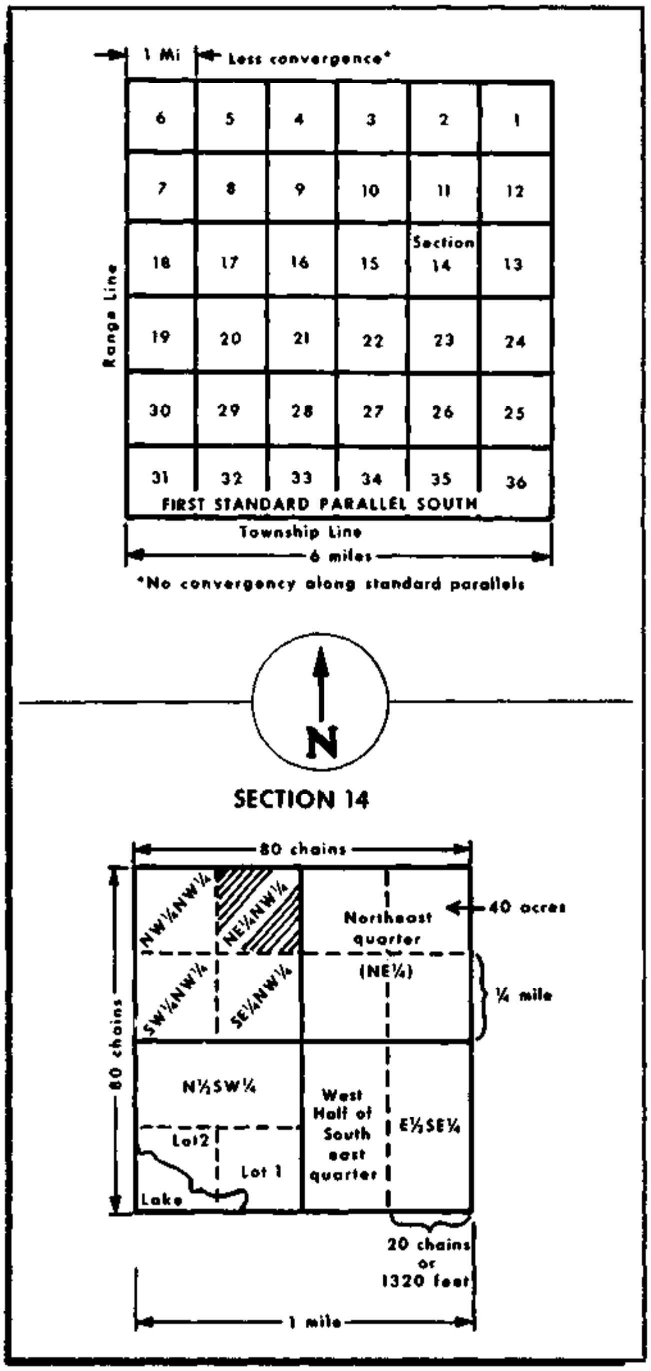
1 Mi
Less convergence*
6 5 4 3 2 1
7 8 9 10 11 12
18 17 16 15 Section 14 13
19 20 21 22 23 24
30 29 28 27 26 25
31 32 33 34 35 36
Range Line
FIRST STANDARD PARALLEL SOUTH
Township Line
6 miles
*No convergency along standard parallels
N
SECTION 14
80 chains
NW¼NW¼
NE¼NW¼
SW¼NW¼
SE¼NW¼
Northeast quarter (NE¼)
40 acres
¼ mile
80 chains
N½SW¼
Lot2
Lot 1
Lake
West Half of South east quarter
E½SE¼
20 chains or 1320 feet
1 mile

Figure 20-34. Section 14.

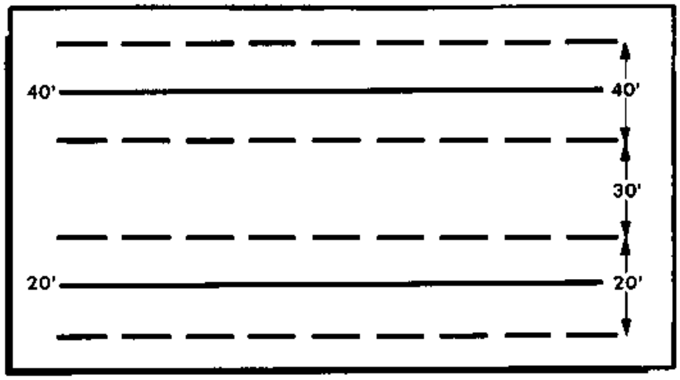

Figure 20-35. Estimating Elevation and Contour Lines

c. Relief. Relief is the representation of the shape and height of landforms and characteristic of the Earth's surface.

d. Contour Lines:

(1) There are several ways of indicating elevation and relief on maps. The most common way is by contour lines. A contour line is an imaginary line connecting points of equal elevation. Contour lines indicate a vertical distance above or below a datum plane. Starting at sea level, each contour line represents an elevation above sea level. The vertical distance between adjacent contour lines is known as the contour interval. The amount of contour interval is given in the marginal information. On most maps, the contour lines are printed in brown. Starting at zero elevation, every fifth contour line is drawn with a heavier line. These are known as index contours and somewhere along each index contour, the line is broken and its elevation is given. The contour lines falling between index contours are called intermediate contours. They are drawn with a finer line than the index contours and usually do not have their elevations given.

(2) Using the contour lines on a map, the elevation of any point may be determined by:

(a) Finding the contour interval of the map from the marginal information, and noting the amount and unit of measure.

(b) Finding the numbered contour line (or other given elevation) nearest the point for which elevation is being sought.

(c) Determining the direction of slope from the numbered contour line to the desired point.

(d) Counting the number of contour lines that must be crossed to go from the numbered line to the desired point and noting the direction—up or down. The number of lines crossed multiplied by the contour interval is the distance above or below the starting value. If the desired point is on a contour line, its elevation is that of the contour; for a point between contours,

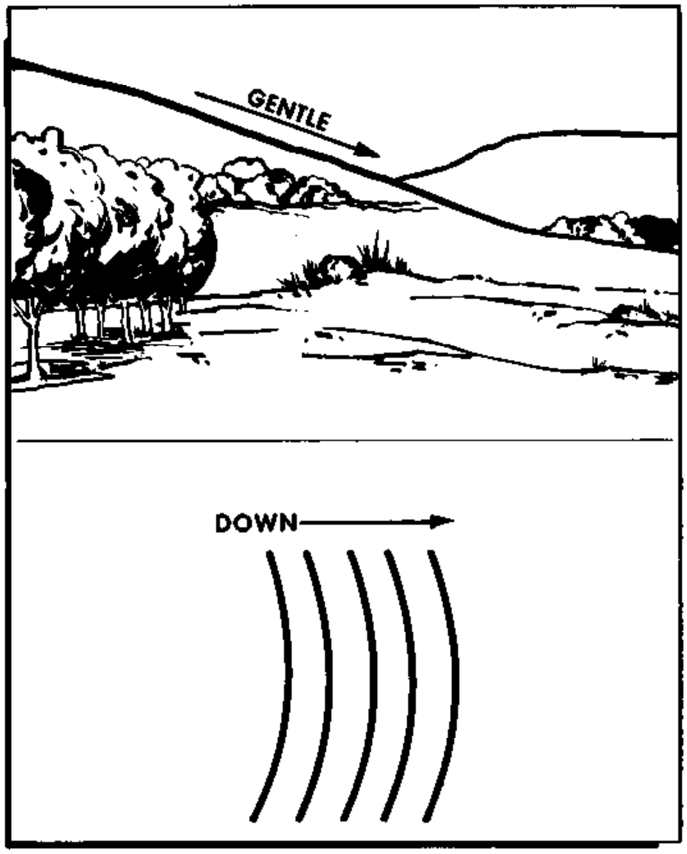

Figure 20-36. Uniform Gentle Slope.

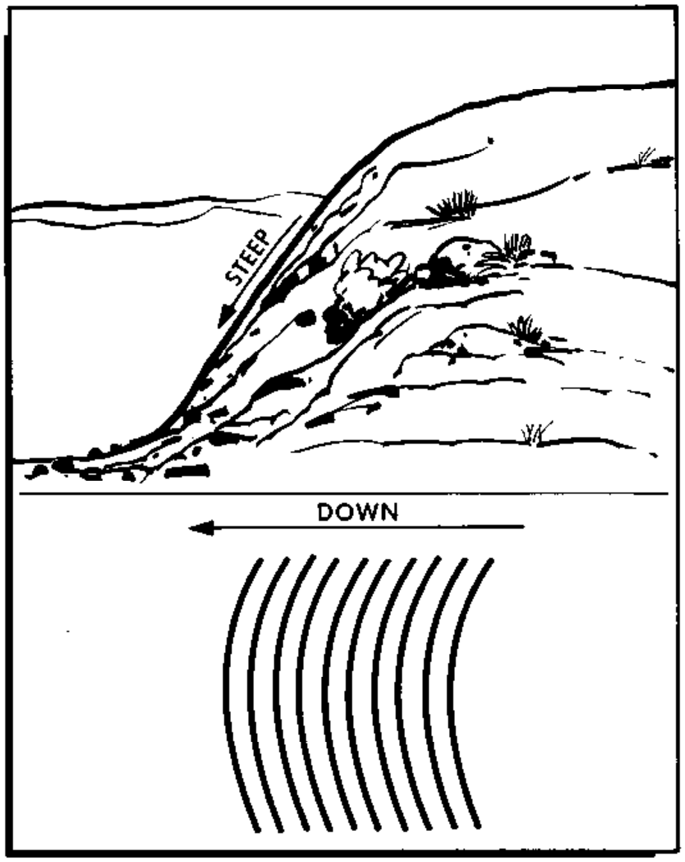

Figure 20-37. Uniform Steep Slope.

most military needs are satisfied by estimating the elevation to an accuracy of one-half the contour interval. All points less than one-fourth the distance between the lines are considered to be at the same elevation as the line. All points one-fourth to three-fourths the distance from the lower line are considered to be at an elevation one-half the contour interval above the lower line (figure 20-35).

(e) To estimate the elevation of the top of an unmarked hill, add half the contour interval to the elevation of the highest contour line around the hill. To estimate the elevation of the bottom of a depression, subtract half the contour interval from the value of the lowest contour around the depression.

(f) On maps where the index and intermediate contour lines do not show the elevation and relief in as much detail as may be needed, supplementary contour may be used. These contour lines are dashed brown lines, usually at one-half the contour interval for the map. A note in the marginal information indicates the interval used. They are used exactly as are the solid contour lines.

(g) On some maps contour lines may not meet the standards of accuracy but are sufficiently accurate in both value and interval to be shown as contour rather than as form lines. In such cases, the contours are considered as approximate and are shown with a dashed symbol; elevation values are given at intervals along the heavier (index contour) dashed lines. The contour note in the map margin identifies them as approximate contours.

(h) In addition to the contour lines, bench marks and spot elevations are used to indicate points of known elevation on the map. Bench marks, the more accurate of the two, are symbolized by a black X, as X BM 124. The elevation value shown in black refers to the center of the X. Spot elevations shown in brown generally are located at road junctions, on hilltops, and other prominent landforms. The symbol designates an accurate horizontal control point. When a bench mark and a horizontal control point are located at the same point, the symbol BM is used.

(i) The spacing of the contour lines indicates the nature of the slope. Contour lines evenly spaced and wide apart indicate a uniform, gentle slope (figure 20-36). Contour lines evenly spaced and close together indicate a uniform, steep slope. The closer the contour lines to each other, the steeper the slope (figure 20-37). Contour lines closely spaced at the top and widely spaced at the bottom indicate a concave slope (figure 20-38). Contour lines widely spaced at the top and closely spaced at the bottom indicate a convex slope (figure 20-39).

(j) To show the relationship of land formations to each other and how they are symbolized on a contour map, stylized panoramic sketches of the major relief

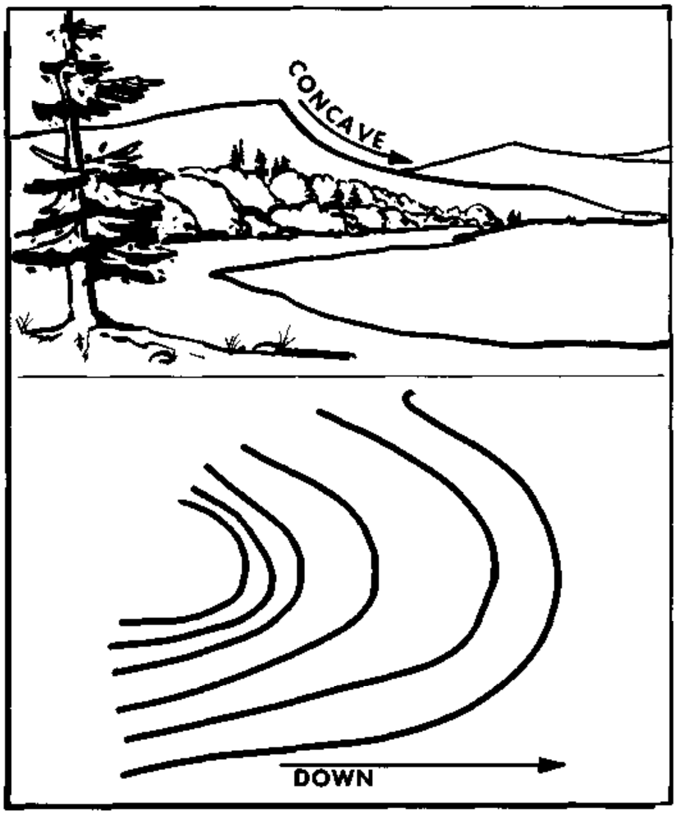

Figure 20-38. Concave Slope.

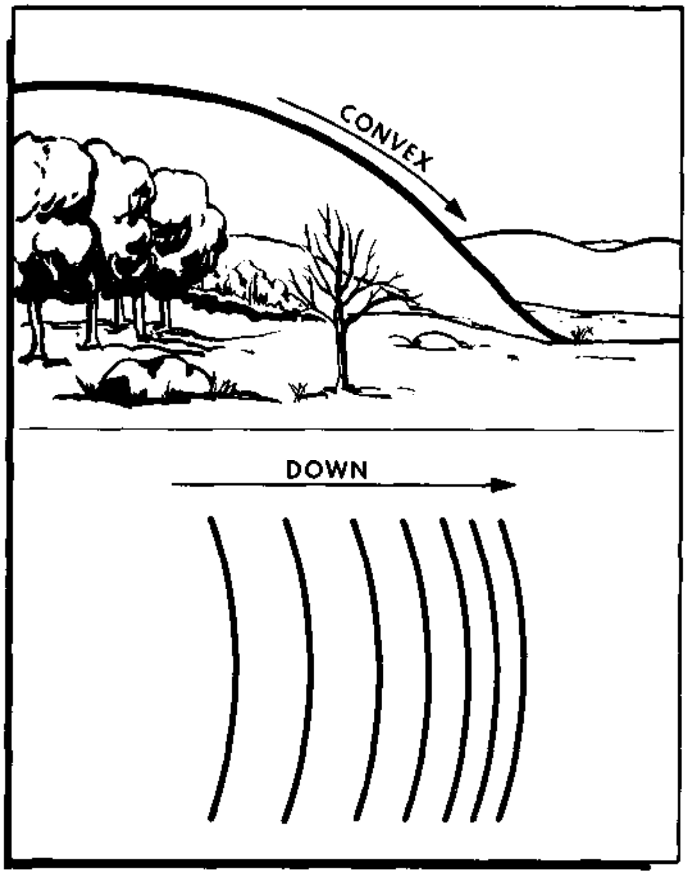

Figure 20-39. Convex Slope.

formations were drawn and a contour map of each sketch developed. Each figure (figure 20-40 through 20-46) shows a sketch and a map with a different relief feature and its characteristic contour pattern.

(3) Hill. A point or small area of high ground (figure 20-40). When one is located on a hilltop, the ground slopes down in all directions.

(4) Valley. Usually a stream course which has at least a limited extent of reasonably level ground bordered on the sides by higher ground (figure 20-41A). The valley generally has maneuvering room within its confines. Contours indicating a valley are U-shaped and tend to parallel a major stream before crossing it. The more gradual the fall of a stream, the farther each contour inner part. The curve of the contour crossing always points upstream.

(5) Drainage. A less-developed stream course in which there is essentially no level ground and, therefore, little or no maneuvering room within its confines (figure 20-41B). The ground slopes upward on each side and toward the head of the drainage. Drainages occur frequently along the sides of ridges, at right angles to the valleys between the ridges. Contours indicating a drainage are V-shaped, with the point of the "V" toward the head of the drainage.

(6) Ridge. A range of hills or mountains with normally minor variations along its crest (figure 20-42A). The ridge is not simply a line of hills; all points of the ridge crest are appreciably higher than the ground on both sides of the ridge.

(7) Finger Ridge. A ridge or line of elevation projecting from or subordinate to the main body of a mountain or mountain range (figure 20-42B). A finger ridge is often formed by two roughly parallel streams cutting drainages down the side of a ridge.

(8) Saddle. A dip or low point along the crest of a ridge. A saddle is not necessarily the lower ground between two hilltops; it may simply be a dip or break along an otherwise level ridge crest (figure 20-43).

(9) Depression. A low point or sinkhole surrounded on all sides by higher ground (figure 20-44).

(10) Cuts and Fills. Manmade features by which the bed of a road or railroad is graded or leveled off by

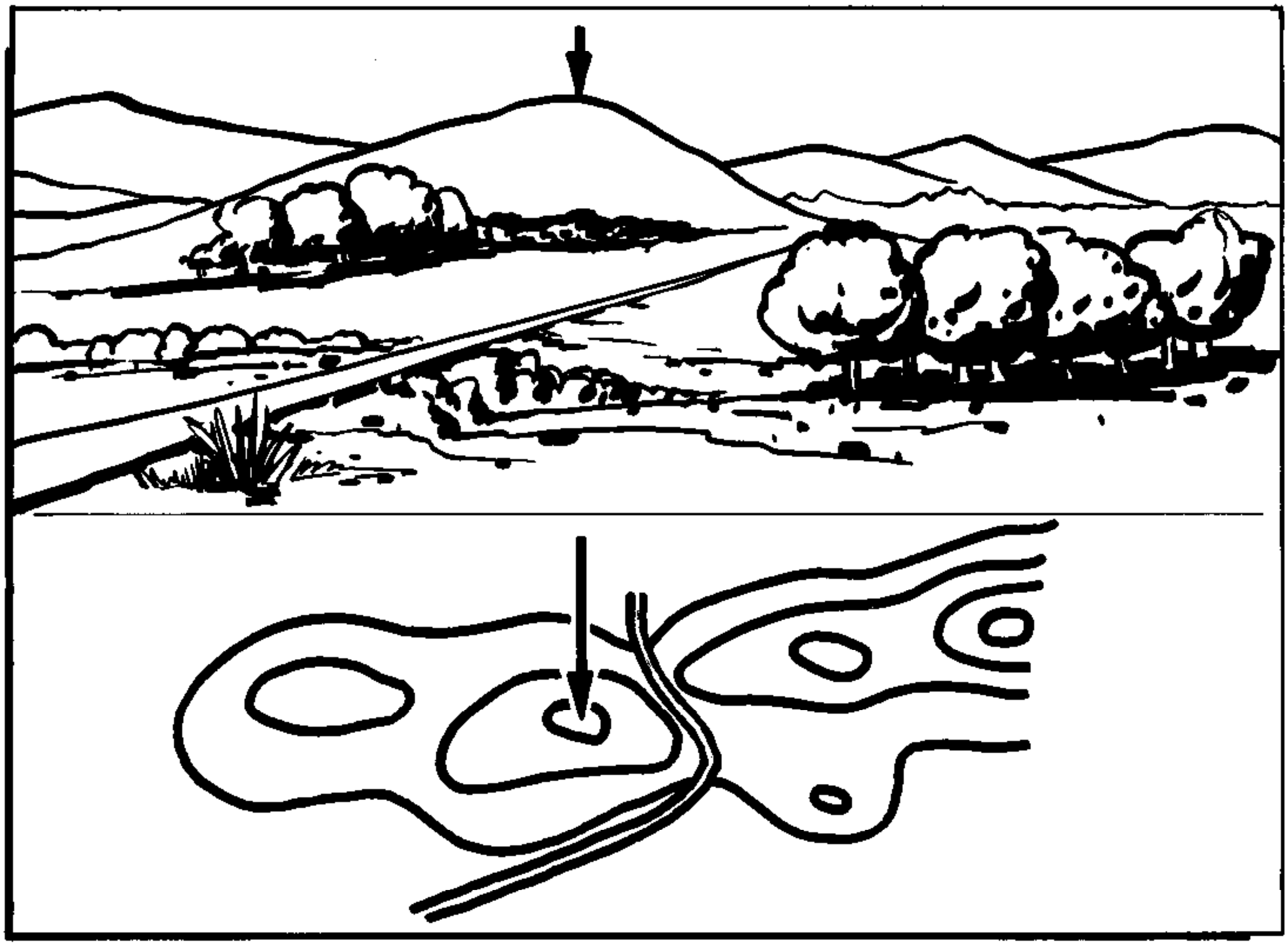

Figure 20-40. Hill.

cutting through high areas (figure 20-45A) and filling in low areas (figure 20-45B) along the right-of-way.

(11) Cliff. A vertical or near vertical slope (figure 20-46). When a slope is so steep that it cannot be shown at the contour interval without the contours fusing, it is shown by a ticked "carrying" contour(s). The ticks always point toward lower ground.

20-8. Representative Fraction (RF):

a. The numerical scale of a map expresses the ratio of horizontal distance on the map to the corresponding horizontal distance on the ground. It usually is written as a fraction, called the representative fraction (RF). The representative fraction is always written with the map distance as one (1). It is independent of any unit of measure. An RF of 1/50,000 or 1:50,000 means that one (1) unit of measure on the map is equal to 50,000 of the same units of measure on the ground.

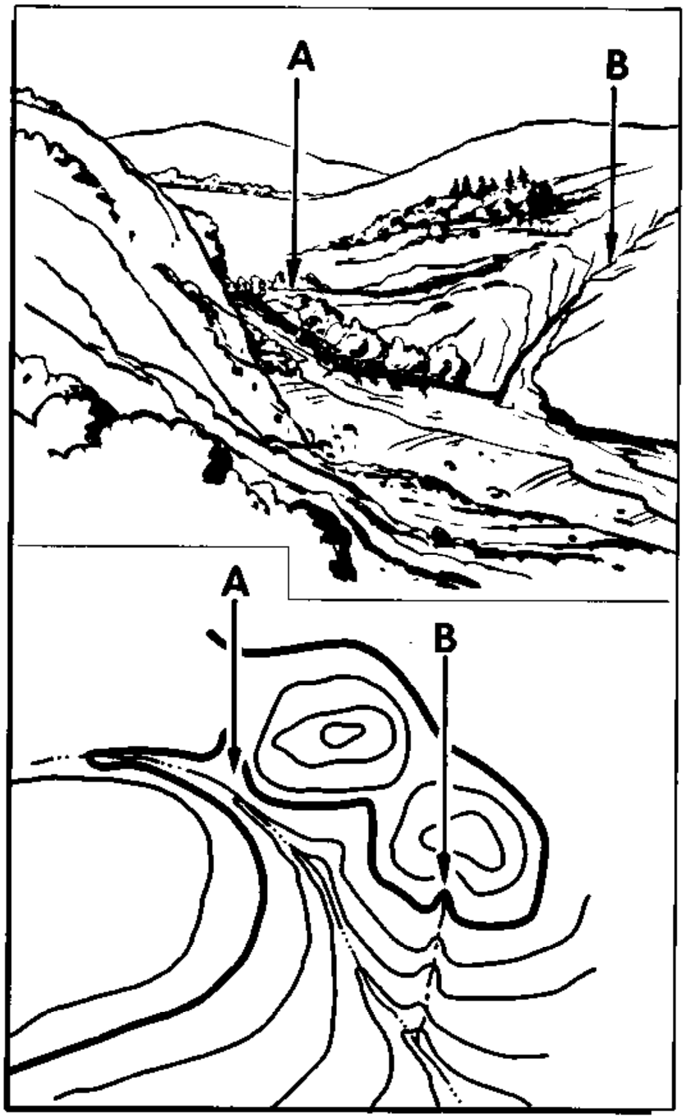

Figure 20-41. (A) Valley (B) Drainage.

b. The ground distance between two points is determined by measuring between the points on the map and multiplying the map measurement by the denominator of the RF.

Example: RF = 1:50,000 or $\frac{1}{50{,}000}$
Map distance = 5 units (CM)
5 X 50,000 - 250,000 units (CM) of
ground distance (figure 20-47).

c. When determining ground distance from a map, the scale of the map affects the accuracy. As the scale becomes smaller, the accuracy of measurement decreases because some of the features on the map must be exaggerated so that they may be readily identified.

20-9. Graphic (Bar) Scales:

a. On most military maps, there is another method of determining ground distance. It is by means of the graphic (bar) scales. A graphic scale is a ruler printed on the map on which distances on the map may be measured as actual ground distances. To the right of the zero (0), the scale is marked in full units of measure and is called the primary scale. The part to the left of zero (0) is divided into tenths of a unit and is called the extension scale. Most maps have three or more graphic scales, each of which measures distance in a different unit of measure (figure 20-48).

b. To determine a straight-line ground distance between two points on a map, lay a straight-edged piece of paper on the map so that the edge of the paper touches both points. Mark the edge of the paper at each point. Move the paper down to the graphic scale and read the ground distance between the points. Be sure to use the scale that measures in the unit of measure desired (figure 20-49).

c. To measure distance along a winding road, stream, or any other curved line, the straightedge of a piece of paper is used again. Mark one end of the paper and place it at the point from which the curved line is to be measured. Align the edge of the paper along a straight portion and mark both the map and the paper at the end of the aligned portion. Keeping both marks together, place the point of the pencil on the mark on the paper to hold it in place. Pivot the paper until another approximately straight portion is aligned and again mark on the map and the paper. Continue in this manner until measurement is complete. Then place the paper on the graphic scale and read the ground distance (figure 20-50).

d. Often, marginal notes give the road distance from the edge of the map to a town, highway, or junction of the map. If the road distance is desired from a point on the map to such a point off the map, measure the distance to the edge of the map and add the distance speci-

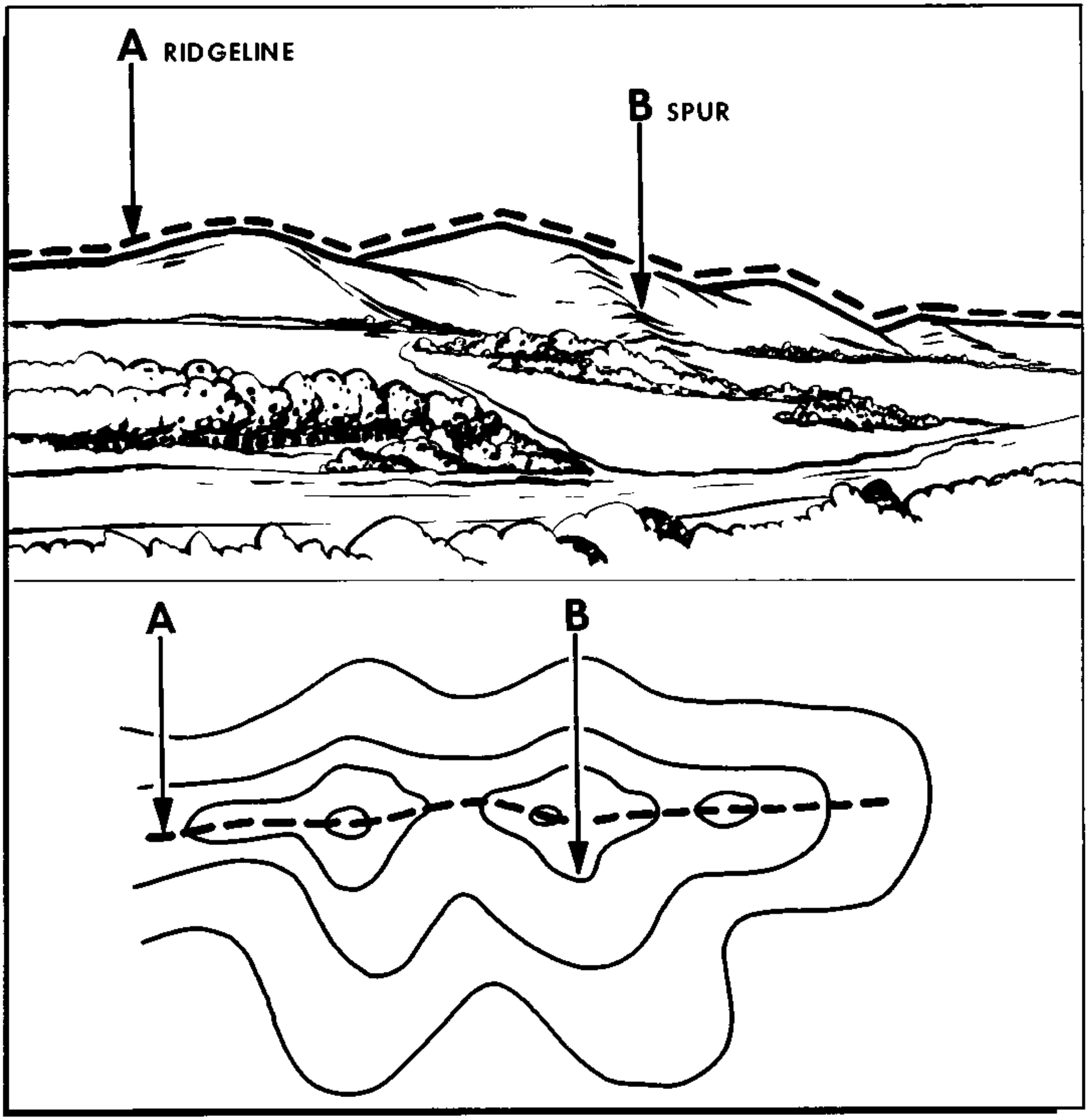

Figure 20-42. (a) Ridge Line (b) Finger Ridge.

fied in the marginal note to that measurement. Be sure the unit of measure is the same (figure 20-51).

20-10. Using a Map and Compass, and Expressing Direction:

a. To use a map, the map must correspond to the lay of the land, and the user must have a knowledge of direction and how the map relates to the cardinal directions. In essence, to use a map for land navigation, the map must be "oriented" to the lay of the land. This is usually done with a compass. On most maps, either a declination diagram, compass rose, and lines of map magnetic variations are provided to inform the user of the difference between magnetic north and true north.

b. Directions are expressed in everyday life as right, left, straight ahead, etc.; but the question arises, "to the right of what?" Military personnel require a method of expressing direction which is accurate, adaptable for use

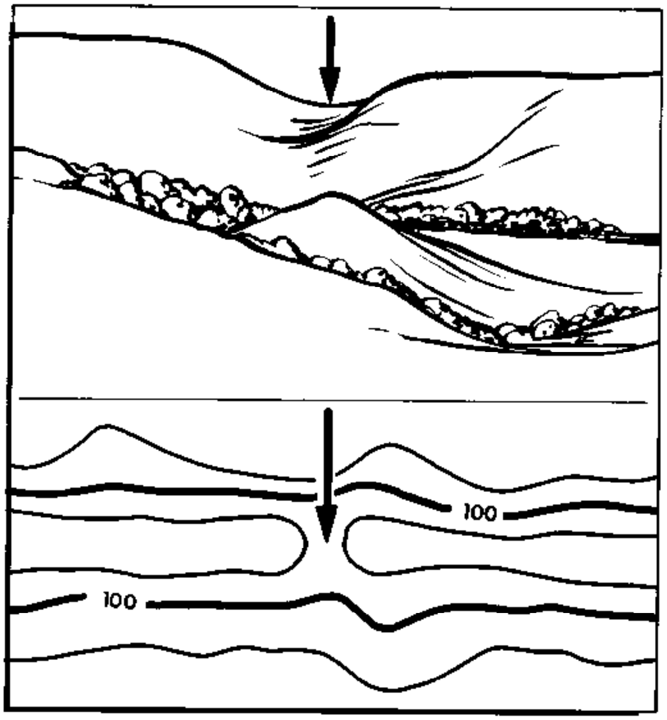

Figure 20-43. Saddle.

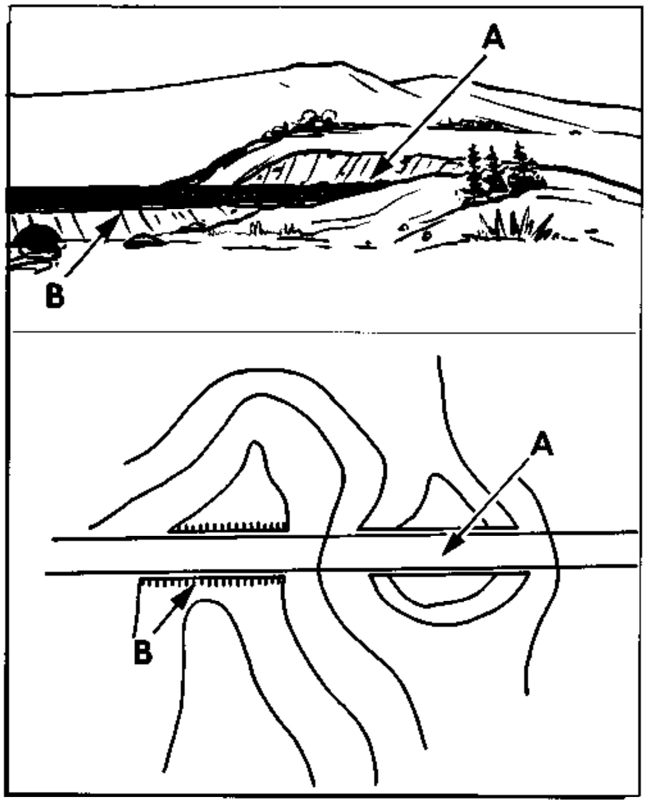

Figure 20-45. (A) Cut (B) Fill.

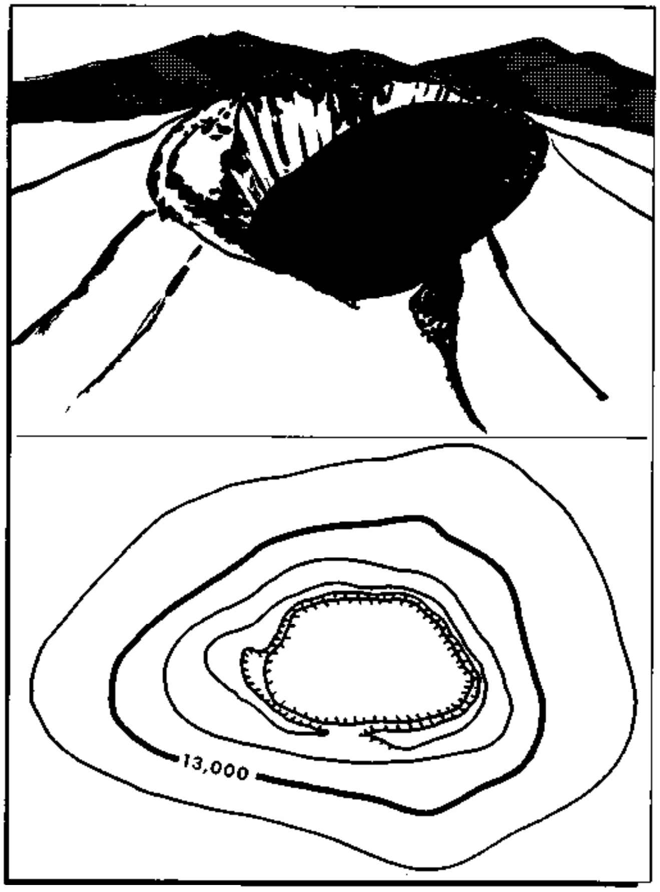

Figure 20-44. Depression.

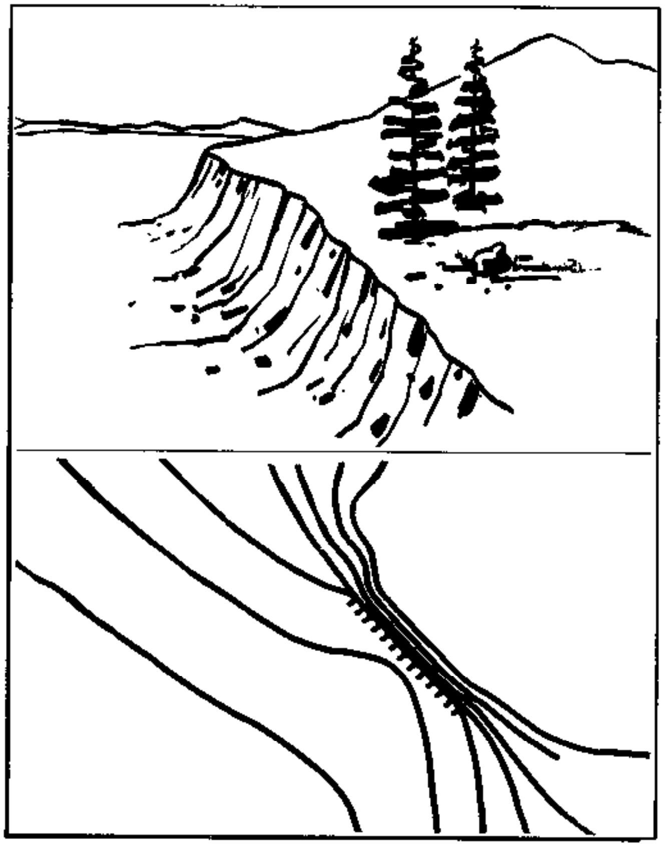

Figure 20-46. Cliff.

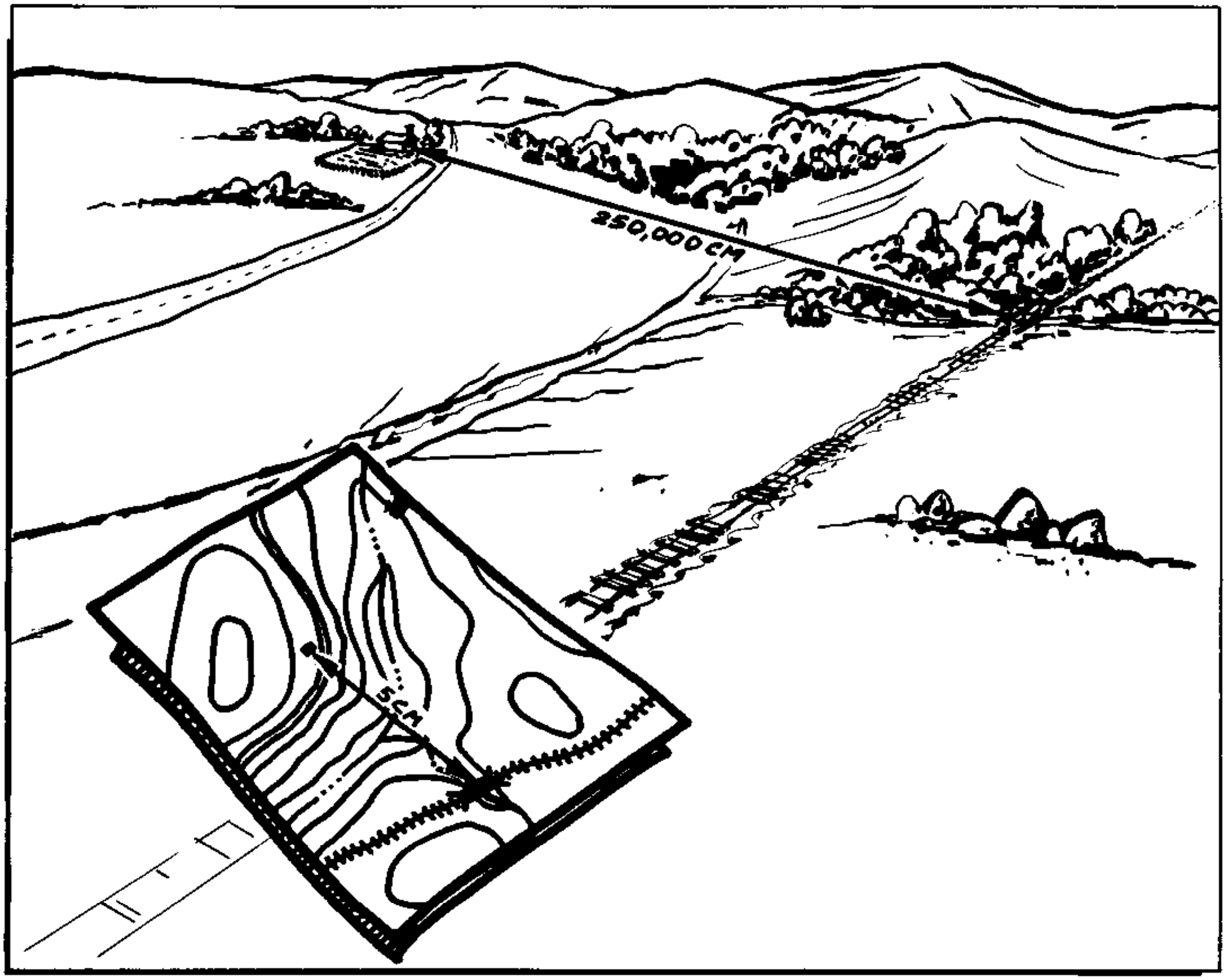
250,000 CM
5CM

Figure 20-47. Ground Distance.

in any area of the world, and has a common unit of measure. Directions are expressed as units of angular measure. The most commonly used unit of angular measure is the degree with its subdivisions of minutes and seconds.

(1) Baselines. To measure anything, there must always be a starting point or zero measurement. To express a direction as a unit of angular measure, there must be a starting point or zero measure and a point of reference. These two points designate the base or refer-

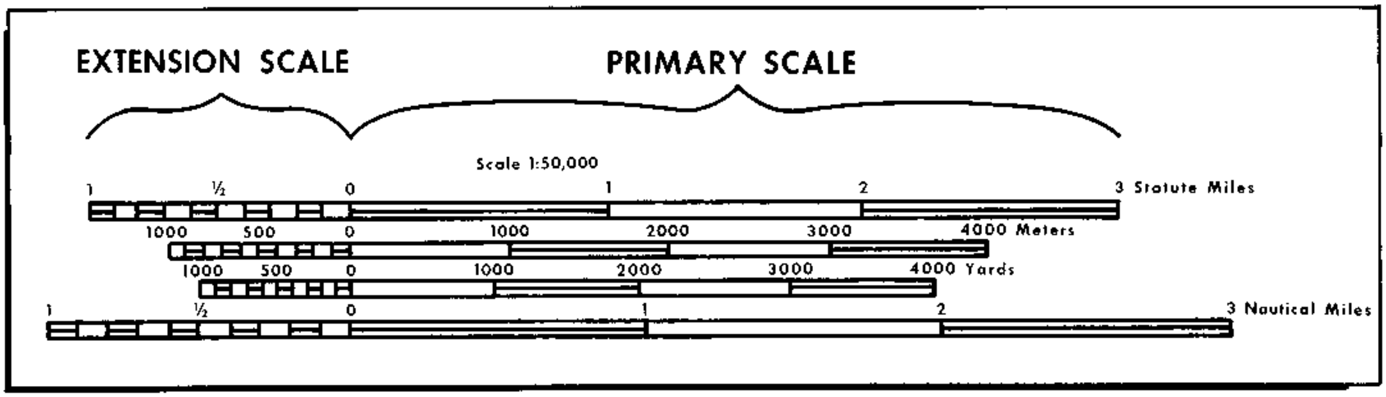
EXTENSION SCALE
PRIMARY SCALE
Scale 1:50,000
1 ½ 0 1 2 3 Statute Miles
1000 500 0 1000 2000 3000 4000 Meters
1000 500 0 1000 2000 3000 4000 Yards
1 ½ 0 1 2 3 Nautical Miles

Figure 20-48. Graphic Bar Scale.

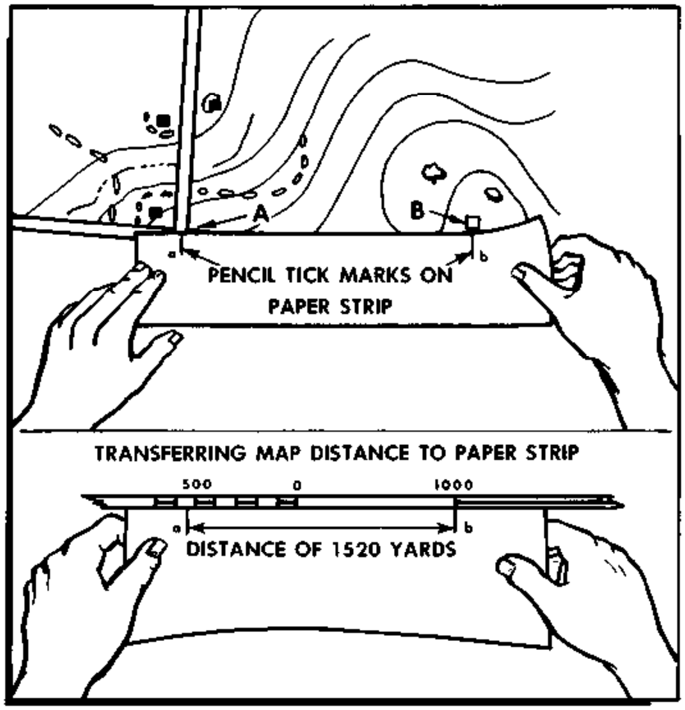
A
B
a
b
PENCIL TICK MARKS ON PAPER STRIP
TRANSFERRING MAP DISTANCE TO PAPER STRIP
500
0
1000
a
b
DISTANCE OF 1520 YARDS

Figure 20-49. Measuring Straight Line Map Distances.

ence line. There are three baselines—true north, magnetic north, and grid north. Those most commonly used are magnetic and grid—the magnetic when working with a compass, and the grid when working with a military map.

(a) True north—a line from any position on the Earth's surface to the North Pole. All lines of longitude are true north lines. True north is usually symbolized by a star (figure 20-52).

(b) Magnetic north—the direction to the north magnetic pole, as indicated by the north-seeking needle of a magnetic instrument. Magnetic north is usually symbolized by a half arrowhead (figure 20-52).

(c) Grid north—the north established by the vertical grid lines on the map. Grid north may be symbolized by the letters GN or the letter Y.

(2) Azimuth and Back Azimuth:

(a) The most common method used by the military for expressing a direction is azimuths. An azimuth is defined as a horizontal angle, measured in a clockwise manner from a north baseline. When the azimuth between two points on a map is desired, the points are joined by a straight line and a protractor is used to measure the angle between north and the drawn line. This measured angle is the azimuth of the drawn line (figure 20-53). When using an azimuth, the point from which the azimuth originates is imagined to be the center of the azimuth circle (figure 20-54). Azimuths take their name from the baseline from which they are measured; true azimuths from true north, magnetic azimuths from magnetic north, and grid azimuths from grid north (figure 20-52). Therefore, any given direction can be expressed in three different ways: a grid azimuth if measured on a military map, a magnetic azimuth if measured by a compass, or a true azimuth if measured from a meridian of longitude.

(b) A back azimuth is the reverse direction of an azimuth. It is comparable to doing an "about face." To obtain a back azimuth from an azimuth, add 180°if the azimuth is 180° or less, or subtract 180° if the azimuth is 180° or more (figure 20-55). The back azimuth of 180° may be stated as either 000° or 360°.

(3) Declination Diagram. A declination diagram is placed on most large-scale maps to enable the user to properly orient the map. The diagram shows the interrelationship of magnetic north, grid north, and true north (figure 20-56). On medium-scale maps, declination information is shown by a note in the map margin.

(a) Declination is the angular difference between true north and magnetic or grid north. There are two declinations, a magnetic declination (figure 20-57) and a grid declination.

(b) Grid-Magnetic (G-M) Angle is an arc indicated by a dashed line, which connects the grid north and the magnetic north prongs. The value of this arc (G-M ANGLE) states the size of the angle between grid north and magnetic north and the year it was prepared. This value is expressed to the nearest ½°, with mil equivalents shown to the nearest 10 mils.

(c) Grid Convergence is an arc, indicated by a dashed line, which connects the prongs for true north and grid north. The value of the angle for the center of the sheet is given to the nearest full minute with its equivalent to the nearest mil. These data are shown in the form of a grid convergence note.

(d) Conversion notes may also appear with the diagram explaining the use of the G-M angle. One note

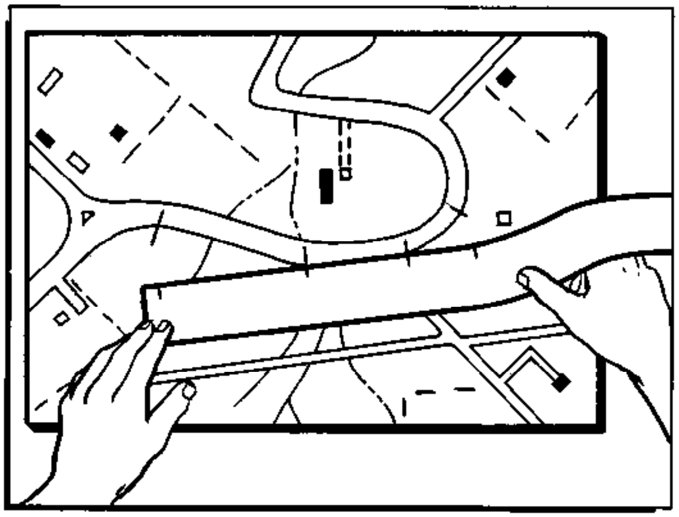

Figure 20-50. Measuring Curved Line Distances.

B–5. Conversion Factors

One	*Inches*	*Feet*	*Yards*	*Statute miles*	*Nautical miles*	*Millimeters*
Inch	1	0.0833	0.0277			25.40
Foot	12	1	0.333			304.8
Yard	36	3	1	0.00056		914.4
Statute Mile	63,360	5,280	1,760	1	0.8684	
Nautical Mile	72,963	6,080	2,020	1.1516	1	
Millimeter	0.0394	0.0033	0.0011			1
Centimeter	0.3937	0.0328	0.0109			10
Decimeter	3.937	0.328	0.1093			100
Meter	39.37	3.2808	1.0936	0.0006	0.0005	1,000
Decameter	393.7	32.81	10.94	0.0062	0.0054	10,000
Hectometer	3,937	328.1	109.4	0.0621	0.0539	100,000
Kilometer	39,370	3,281	1,094	0.6214	0.5396	1,000,000
Myriameter	393,700	32,808	10,936	6.2137	5.3959	10,000,000

One	*cm*	*dm*	*m*	*dkm*	*hm*	*km*	*mym*
Inch	2.540	0.2540	0.0254	0.0025	0.0003		
Foot	30.48	3.048	0.3048	0.0305	0.0030	0.0003	
Yard	91.44	9.144	0.9144	0.0914	0.0091	0.0009	
Statute Mile	160,930	16,093	1,609	160.9	16.09	1.6093	0.1609
Nautical Mile	185,325	18,532	1,853	185.3	18.53	1.8532	0.1853
Millimeter	0.1	0.01	0.001	0.0001			
Centimeter	1	0.1	0.01	0.001	0.0001		
Decimeter	10	1	0.1	0.01	0.001	0.0001	
Meter	100	10	1	0.1	0.01	0.001	0.0001
Decameter	1,000	100	10	1	0.1	0.01	0.001
Hectometer	10,000	1,000	100	10	1	0.1	0.01
Kilometer	100,000	10,000	1,000	100	10	1	0.1
Myriameter	1,000,000	100,000	10,000	1,000	100	10	1

Example I

Problem: Reduce 76 centimeters to (?) inches

76 cm × 0.3937 = 29 inches.

Answer: There are 29 inches in 76 centimeters.

Example II

Problem: How many feet are there in 2.74 meters?

$$\frac{2.74}{.3048} = 9 \text{ feet}$$

Answer: There are approximately 9 feet in 2.74 meters.

Figure 20-51. Conversion Factors.

provides instructions for converting magnetic azimuth to grid azimuth; the other note provides for converting grid azimuth to magnetic azimuth. The conversion (add or subtract) is governed by the direction of the magnetic north prong relative to the grid north prong.

(e) The grid north prong is aligned with the easting grid lines on the map, and on most maps is formed by an extension of an easting grid line into the margin. The angles between the prongs are seldom plotted exactly. The relative position of the directions is obtained from the diagram, but the numerical value should not be measured from it. For example, if the amount of declination from grid north to magnetic north is 1°, the arc shown in the diagram may be exaggerated if mea-

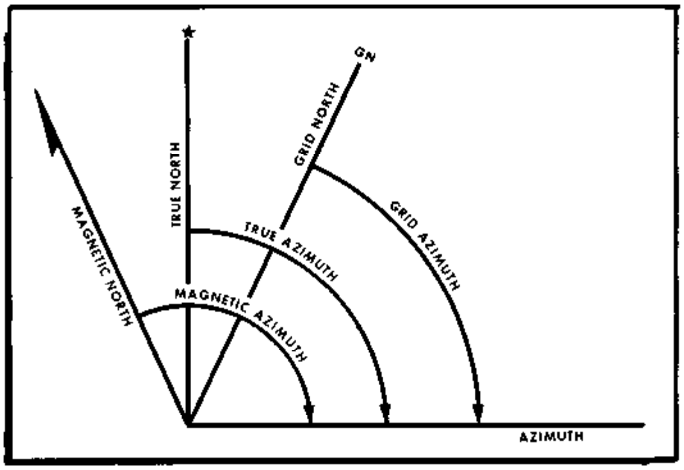

Figure 20-52. True, Grid and Magnetic Azimuths.

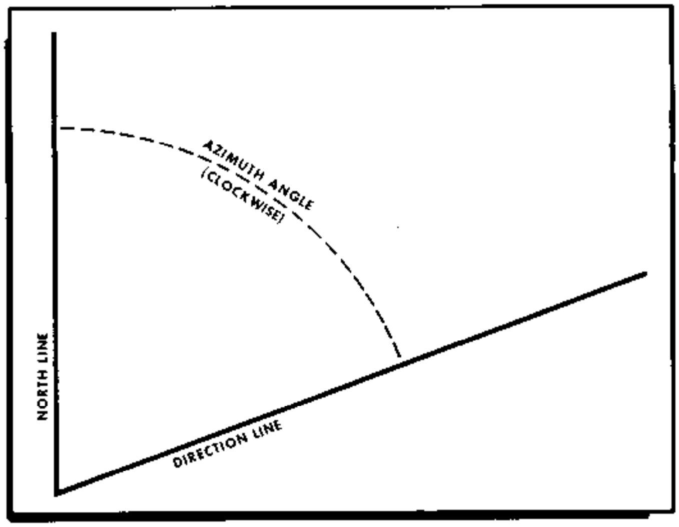

Figure 20-53. Azimuth Angle.

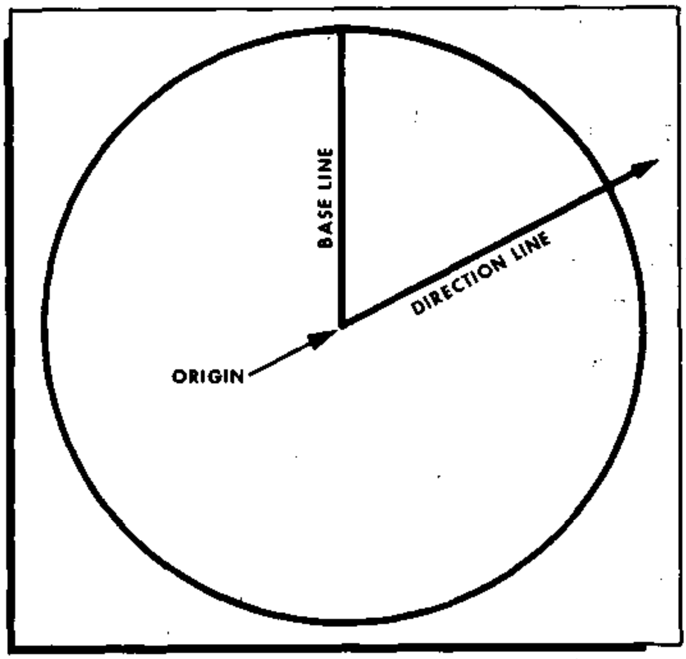

Figure 20-54. Origin of Azimuth Circle.

sured, having an actual value of 5°. The position of the three prongs in relation to each other varies according to the declination data for each map.

(f) Some older maps have a note under the declination diagram which gives the magnetic declination for a certain year and the amount of annual change. The annual change is so small when compared to the ½° value of the G-M angle that it is no longer shown on standard large scale maps.

(4) Protractors. Protractors come in several forms—full circle, half circle, square, and rectangular (figure 20-58). All of them divide a circle into units of angular measure, and regardless of their shape, consist of a scale around the outer edge and an index mark. The index mark is the center of the protractor circle from which all the direction lines radiate.

(a) To determine the grid azimuth of a line from one point to another on the map from (A to B or C to D) (figure 20-59), draw a line connecting the two points.

-1. Place the index of the protractor at that point where the line crosses a vertical (north-south) grid line.

-2. Keeping the index at this point, align the 0° - 180° line of the protractor on the vertical grid line.

-3. Read the value of the angle from the scale; this is the grid azimuth to the point.

(b) To plot a direction line from a known point on a map (figure 20-60):

-1. Construct a north-south grid line through the known point:

-a. Generally, align the 0° - 180° line of the protractor in a north-south direction through the known point.

-b. Holding the 0° - 180° line of the protractor on the known point, slide the protractor in the north-south direction until the horizontal line of the protractor (connecting the protractor index and the 90° tick mark) is aligned on an east-west grid line.

-c. Then draw a line connecting 0°, the known point, and 180°.

-2. Holding the 0° - 180° line on the north-south line, slide the protractor index to the known point.

-3. Make a mark on the map at the required angle. (In an evasion situation, do not mark on the map.)

-4. Draw a line from the known point through the mark made on the map. This is the grid direction line.

(5) The Compass and Its Uses:

(a) The magnetic compass is the most commonly used and simplest instrument for measuring directions and angles in the field. The lensatic compass (figure 20-61) is the standard magnetic compass for military use today.

(b) The lensatic compass must always be held level and firm when sighting on an object and reading an azimuth (figure 20-62). There are several techniques for holding the compass and sighting. One way is to align the sighting slot with the hairline on the front sight

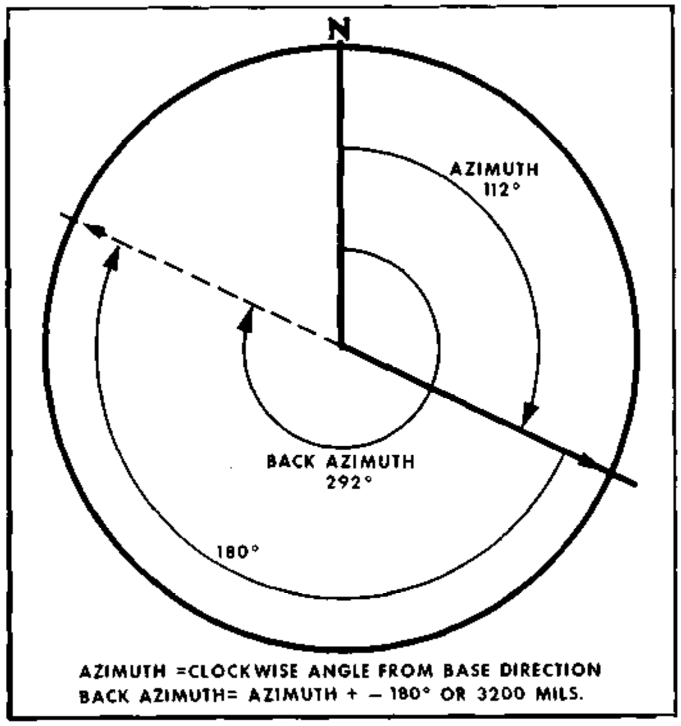
N
AZIMUTH
112°
BACK AZIMUTH
292°
180°
AZIMUTH =CLOCKWISE ANGLE FROM BASE DIRECTION
BACK AZIMUTH= AZIMUTH + – 180° OR 3200 MILS.

Figure 20-55. Azimuth and Back Azimuth.

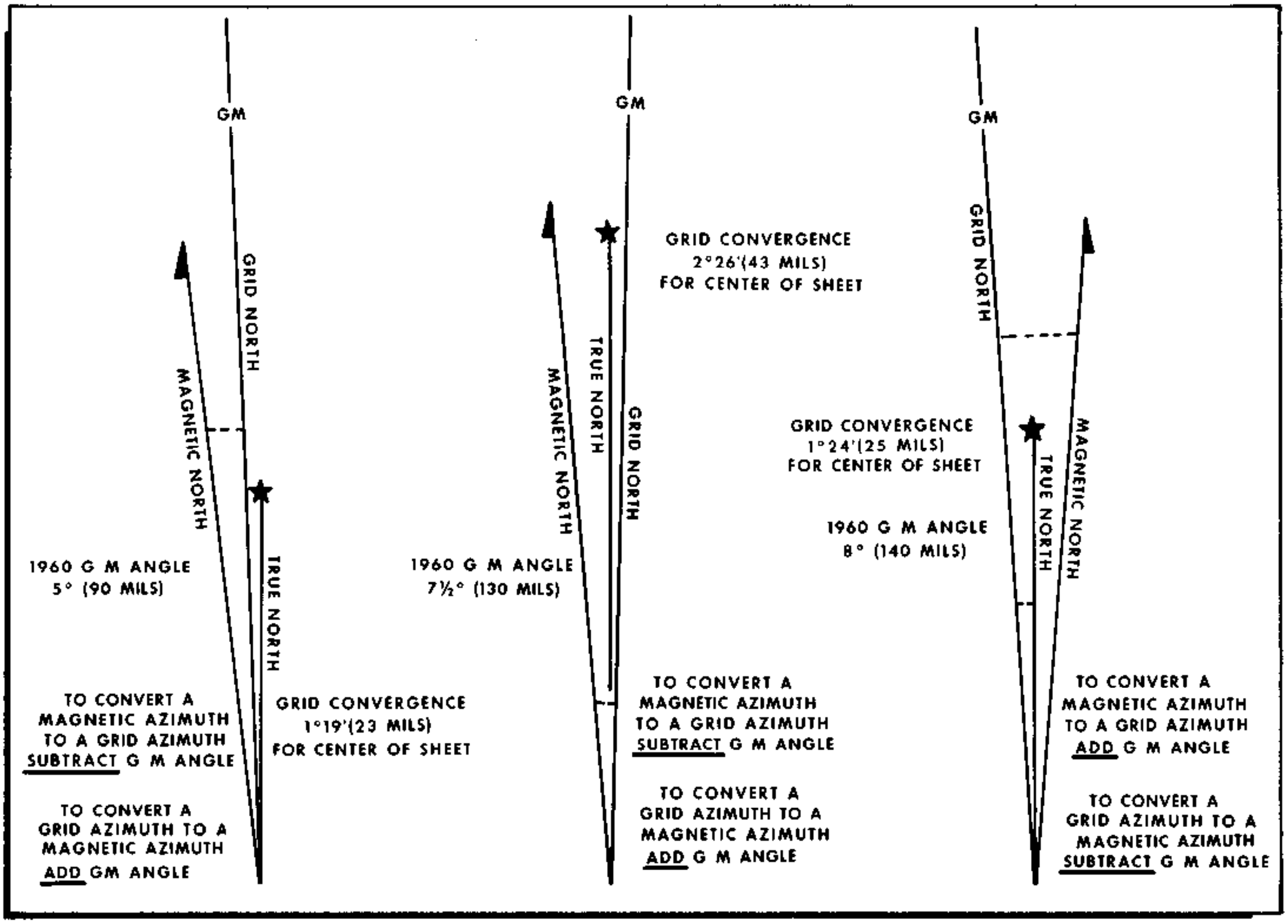
GM
GRID NORTH
MAGNETIC NORTH
TRUE NORTH
1960 G M ANGLE
5° (90 MILS)
TO CONVERT A MAGNETIC AZIMUTH TO A GRID AZIMUTH SUBTRACT G M ANGLE
TO CONVERT A GRID AZIMUTH TO A MAGNETIC AZIMUTH ADD GM ANGLE
GRID CONVERGENCE
1°19'(23 MILS)
FOR CENTER OF SHEET
GM
MAGNETIC NORTH
TRUE NORTH
GRID NORTH
GRID CONVERGENCE
2°26'(43 MILS)
FOR CENTER OF SHEET
1960 G M ANGLE
7½° (130 MILS)
TO CONVERT A MAGNETIC AZIMUTH TO A GRID AZIMUTH SUBTRACT G M ANGLE
TO CONVERT A GRID AZIMUTH TO A MAGNETIC AZIMUTH ADD G M ANGLE
GM
GRID NORTH
TRUE NORTH
MAGNETIC NORTH
GRID CONVERGENCE
1°24'(25 MILS)
FOR CENTER OF SHEET
1960 G M ANGLE
8° (140 MILS)
TO CONVERT A MAGNETIC AZIMUTH TO A GRID AZIMUTH ADD G M ANGLE
TO CONVERT A GRID AZIMUTH TO A MAGNETIC AZIMUTH SUBTRACT G M ANGLE

Figure 20-56. Declination Diagram (East and West).

in the cover and the target. The azimuth can then be read by glancing down at the dial through the lens. This technique provides a reading precise enough to use.

(6) Night Use of the Compass. For night use, special features of the compass include the luminous markings, the bezel ring, and two luminous sighting dots. Turning the bezel ring counterclockwise causes an increase in azimuth, while turning it clockwise causes a decrease. The bezel ring has a stop and spring which allows turns at 3° intervals per click and holds it at any desired position. One accepted method for determining compass directions at night is:

(a) Rotate the bezel ring until the luminous line is over the black index line.

(b) Hold the compass with one hand and rotate the bezel ring in a counterclockwise direction with the other hand to the number of clicks required. The number of clicks is determined by dividing the value of the required azimuth by 3. For example, for an azimuth of 51°, the bezel ring would be rotated 17 clicks counterclockwise (figure 20-63).

(c) Turn the compass until the north arrow is directly under the luminous line on the bezel.

(d) Hold the compass open and level in the palm of the left hand with the thumb along the side of the compass. In this manner, the compass can be held consistently in the same position. Position the compass approximately halfway between the chin and the belt, pointing to the direct front. (Practice in daylight will make a person proficient in pointing the compass the same way every time.) Looking directly down into the compass, turn the body until the north arrow is under the luminous line. Then proceed forward in the direction of the luminous sighting dots (figure 20-61). When the compass is to be used in darkness, an initial azimuth should be set while light is still available. With this initial azimuth as a base, any other azimuth which is a multiple of 3° can be established through use of the clicking feature of the bezel ring. The magnetic compass is a delicate instrument, especially the dial balance. The survivor should take care in its use. Compass readings should never be taken near visible masses of iron or electrical circuits.

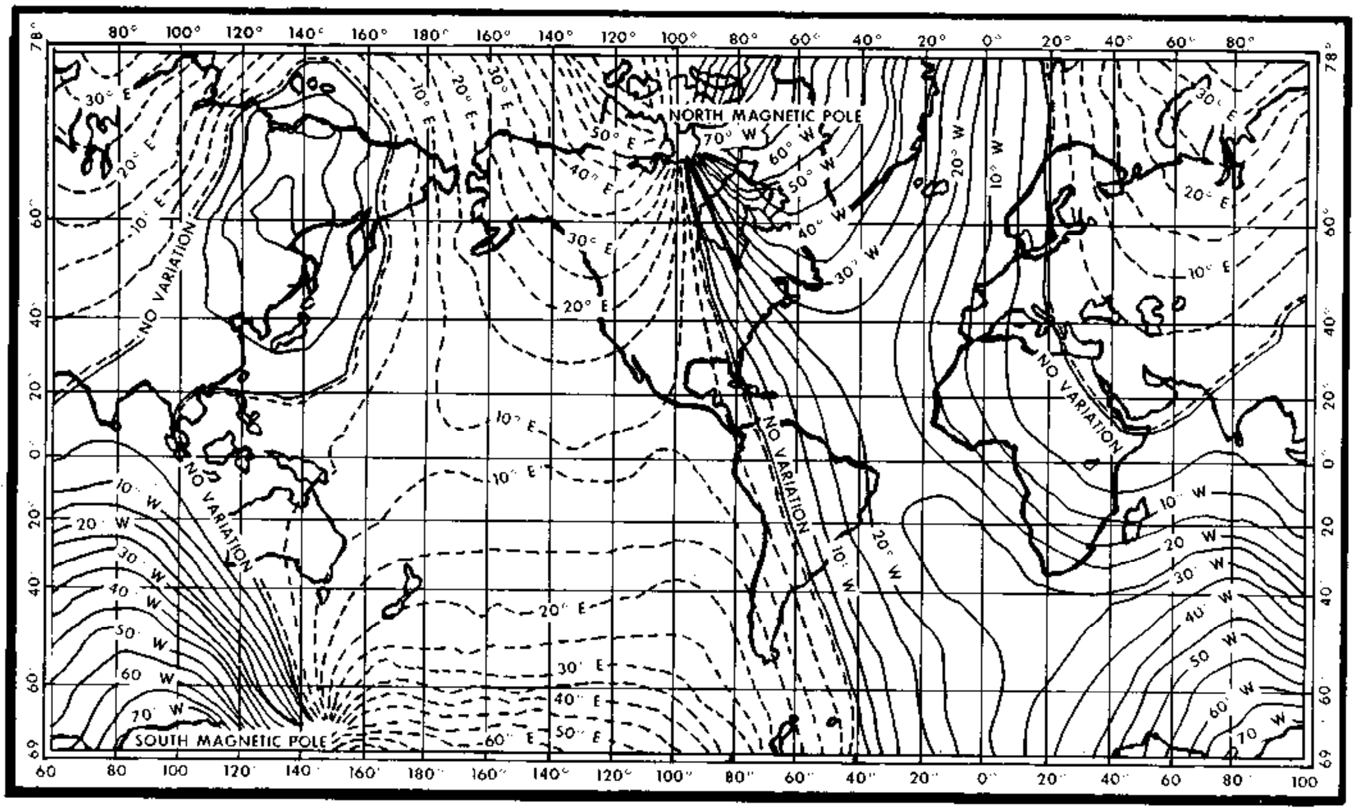
NORTH MAGNETIC POLE
SOUTH MAGNETIC POLE
NO VARIATION
NO VARIATION
NO VARIATION
NO VARIATION
10° E
20° E
30° E
40° E
50° E
60° E
10° W
20° W
30° W
40° W
50° W
60° W
70° W

Figure 20-57. Lines of Magnetic Variation.

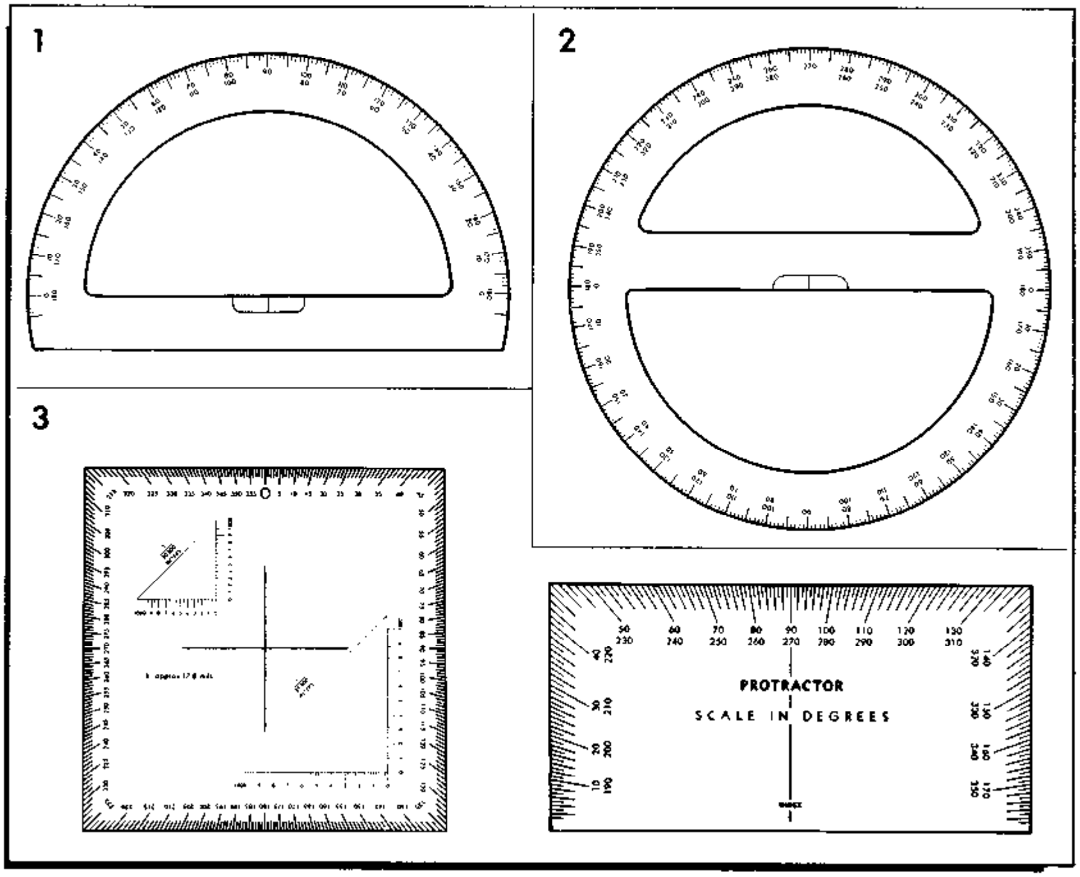
1
2
3
PROTRACTOR
SCALE IN DEGREES
INDEX

Figure 20-58. Types of Protractors.

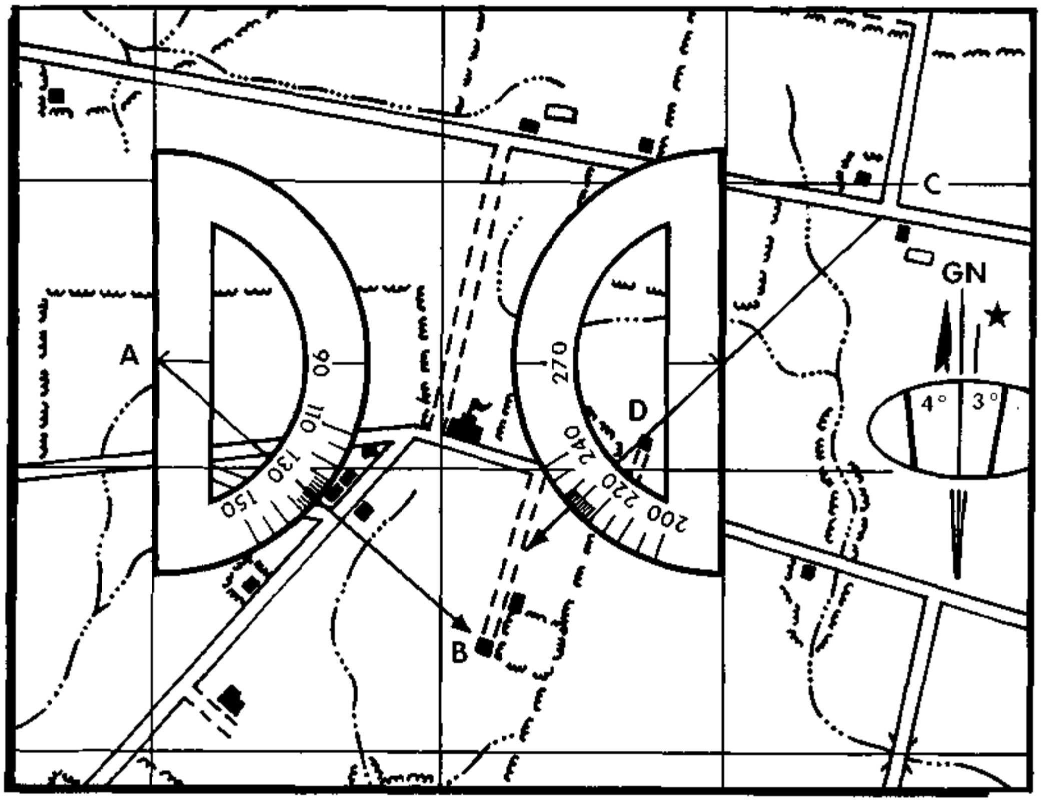
A
B
C
D
GN
4°
3°
90
110
130
150
270
240
220
200

Figure 20-59. Measuring an Azimuth on a Map.

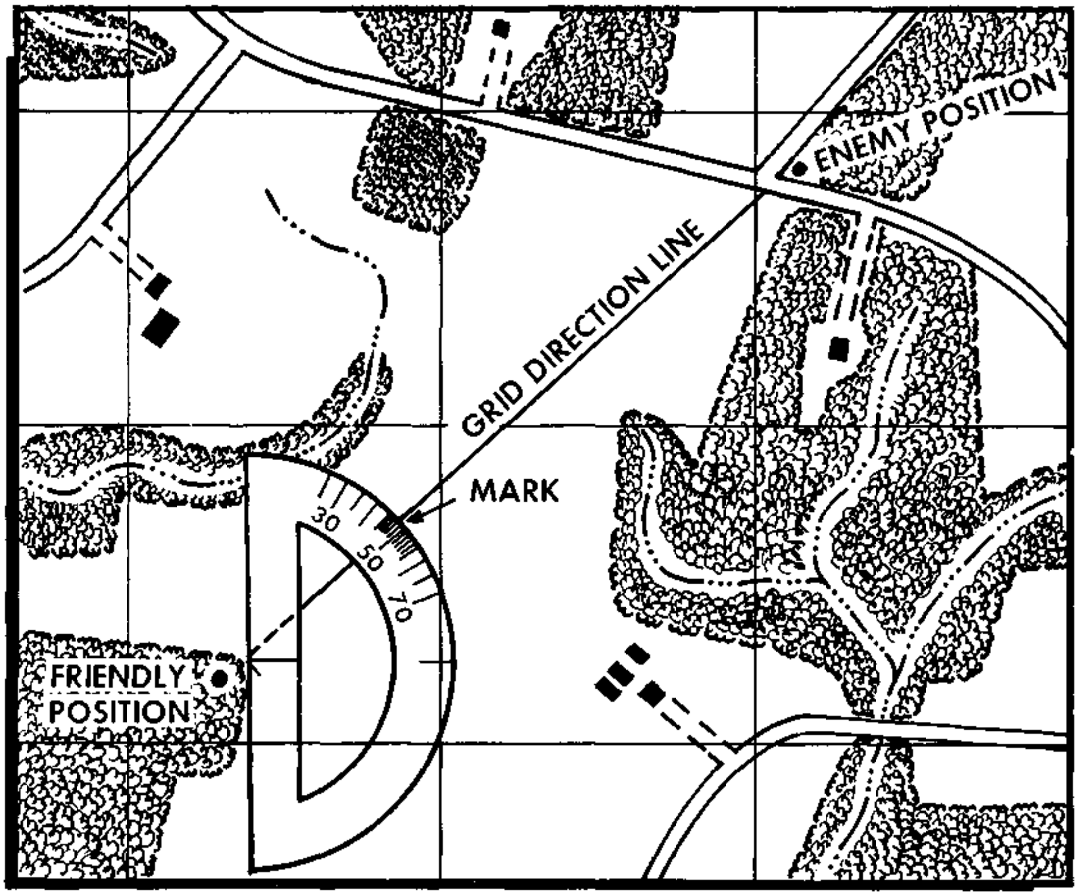
ENEMY POSITION
GRID DIRECTION LINE
MARK
30
50
70
FRIENDLY POSITION

Figure 20-60. Plotting an Azimuth on a Map.

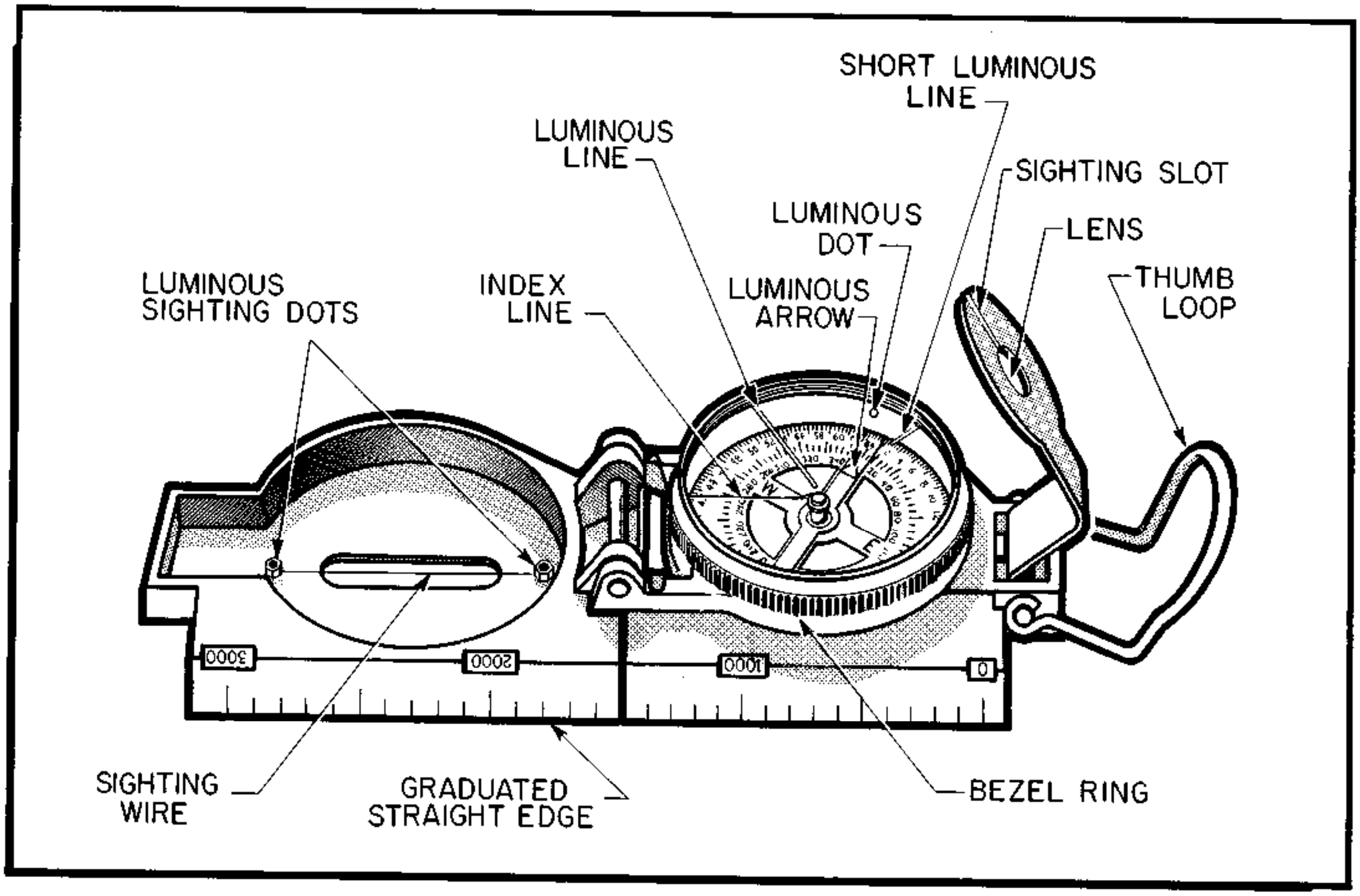
SHORT LUMINOUS LINE
LUMINOUS LINE
SIGHTING SLOT
LUMINOUS DOT
LENS
LUMINOUS SIGHTING DOTS
INDEX LINE
LUMINOUS ARROW
THUMB LOOP
3000
2000
1000
0
SIGHTING WIRE
GRADUATED STRAIGHT EDGE
BEZEL RING

Figure 20-61. Lensatic Compass.

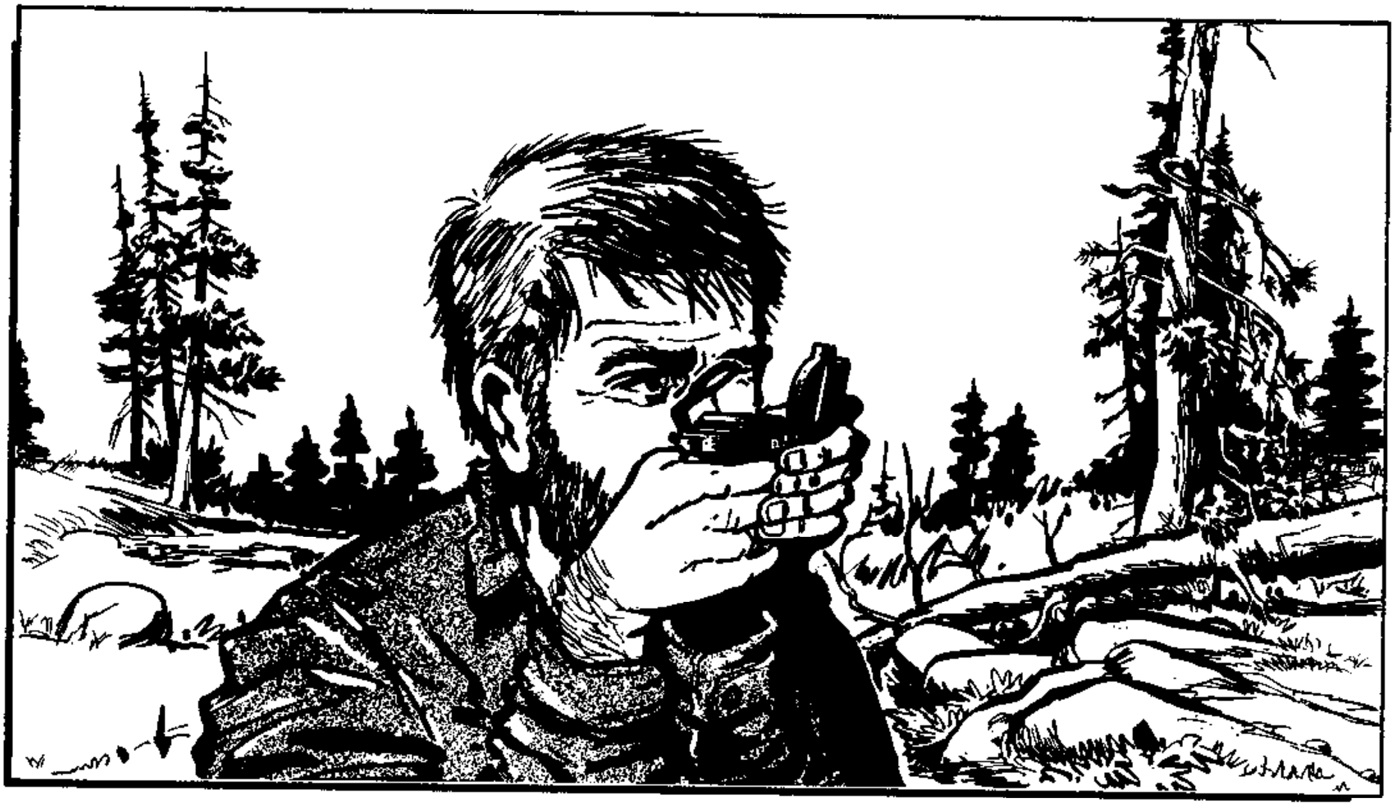

Figure 20-62. Holding the Compass.

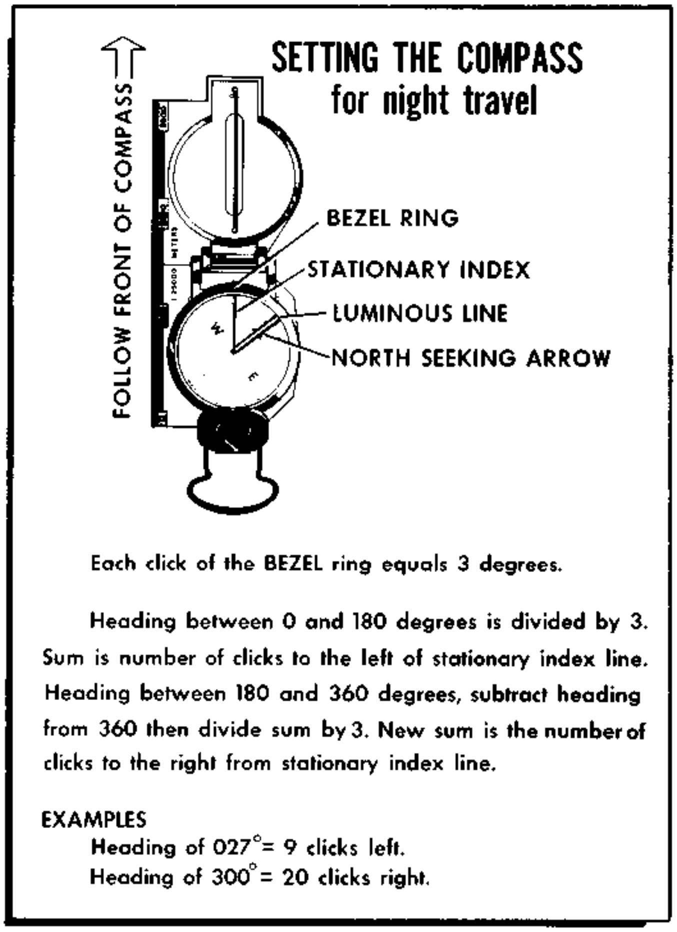
SETTING THE COMPASS
for night travel
FOLLOW FRONT OF COMPASS
BEZEL RING
STATIONARY INDEX
LUMINOUS LINE
NORTH SEEKING ARROW

Each click of the BEZEL ring equals 3 degrees.

Heading between 0 and 180 degrees is divided by 3. Sum is number of clicks to the left of stationary index line. Heading between 180 and 360 degrees, subtract heading from 360 then divide sum by 3. New sum is the number of clicks to the right from stationary index line.

EXAMPLES
Heading of 027°= 9 clicks left.
Heading of 300°= 20 clicks right.

Figure 20-63. Night Travel.

(7) Map Orientation:

(a) A map is oriented when it is in a horizontal position with its north and south corresponding to north and south on the ground. The best way to orient a map is with a compass. (NOTE: Caution should be used to ensure nothing (metal, mine ore, etc.) in the area will alter the compass reading.)

(b) With the map in a horizontal position, the lensatic compass is placed parallel to a north-south grid line with the cover side of the compass pointing toward the top of the map. This will place the black index line on the dial of the compass parallel to grid north. Since the needle on the compass points to magnetic north, a declination diagram is (on the face of the compass) formed by the index line and the compass needle.

(c) Rotate the map and compass until the directions of the declination diagram formed by the black index line and compass needle match the directions shown on the declination diagram printed in the margin of the map. The map is then oriented (grid north).

(d) If the magnetic north arrow on the map is to the left of grid north, the compass reading will equal the G-M angle (given in the declination diagram). If the magnetic north is to the right of grid north, the compass reading will equal 360° minus the G-M angle. In figure 20-64, the declination diagram illustrates a magnetic north to the right of grid north and the compass reading will be 360° minus 21½° or 338½°.

(e) Remember to point the compass north arrow in the same direction as the magnetic north arrow, and the compass reading (equal to the G-M angle or the 360° minus G-M angle) will be quite apparent.

(f) In summary, if the variation is to the east of true north or the magnetic north arrow of the declination diagram is to the east (right) of the grid north line, subtract the degrees of variation from 360°. If it is to the left (west), add to 000°. East is least and west is best.

(g) If a grid line is not used, a true north-south line can be used. True north-south lines are longitudinal lines or lines formed by the vertical lines on a tick map (assuming the top of the map is north). The same procedure is used if magnetic variation is figured from true north—not grid north.

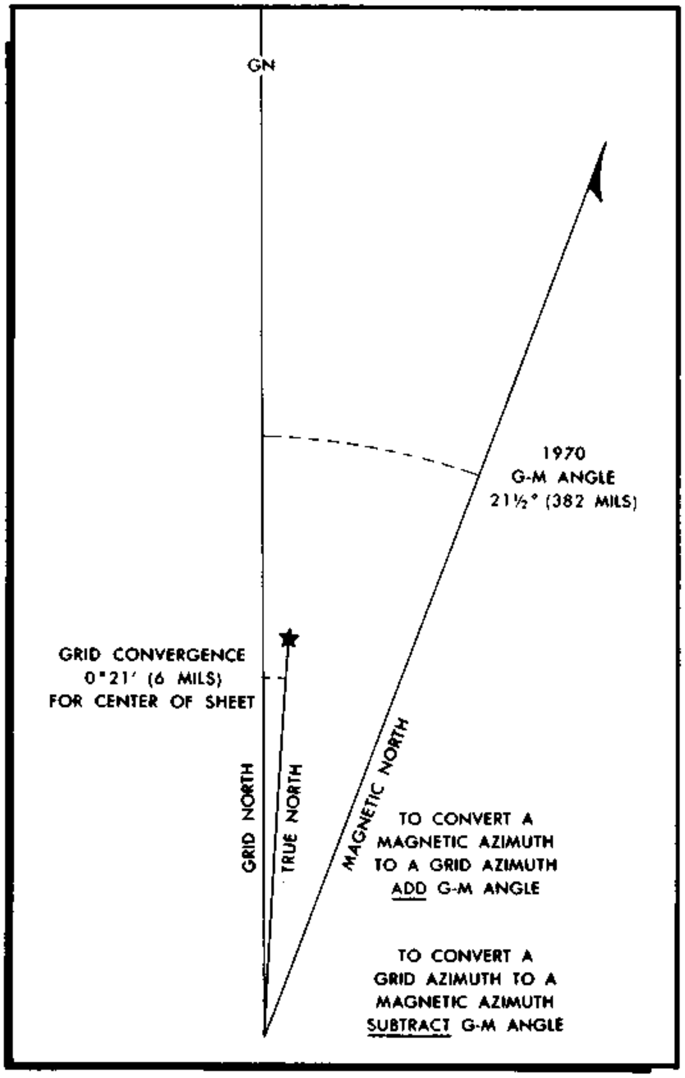
GN
1970
G-M ANGLE
21½° (382 MILS)
GRID CONVERGENCE
0°21′ (6 MILS)
FOR CENTER OF SHEET
GRID NORTH
TRUE NORTH
MAGNETIC NORTH
TO CONVERT A MAGNETIC AZIMUTH TO A GRID AZIMUTH ADD G-M ANGLE
TO CONVERT A GRID AZIMUTH TO A MAGNETIC AZIMUTH SUBTRACT G-M ANGLE

Figure 20-64. Declination Diagram.

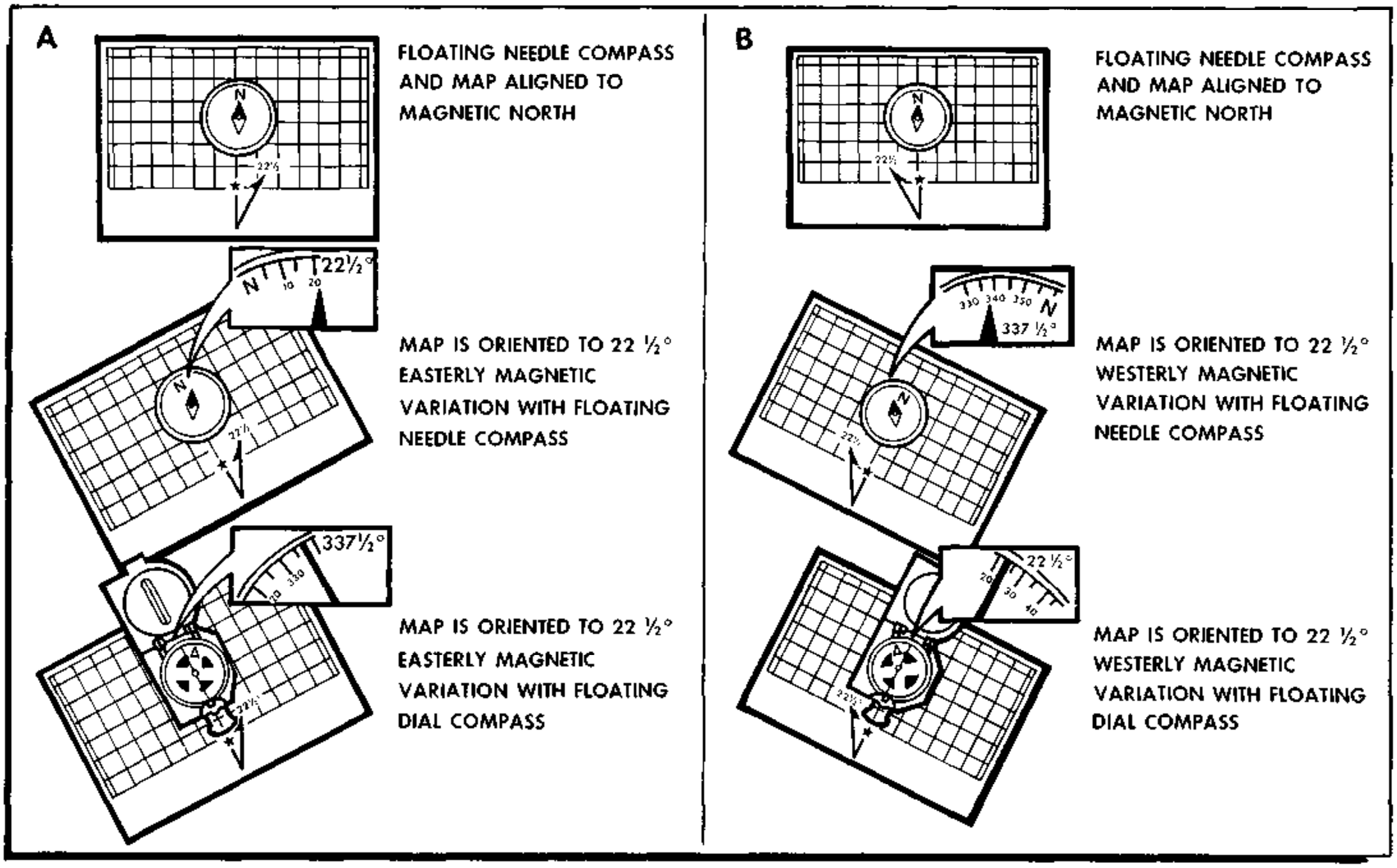

FLOATING NEEDLE COMPASS AND MAP ALIGNED TO MAGNETIC NORTH

MAP IS ORIENTED TO 22 ½° EASTERLY MAGNETIC VARIATION WITH FLOATING NEEDLE COMPASS

MAP IS ORIENTED TO 22 ½° EASTERLY MAGNETIC VARIATION WITH FLOATING DIAL COMPASS

FLOATING NEEDLE COMPASS AND MAP ALIGNED TO MAGNETIC NORTH

MAP IS ORIENTED TO 22 ½° WESTERLY MAGNETIC VARIATION WITH FLOATING NEEDLE COMPASS

MAP IS ORIENTED TO 22 ½° WESTERLY MAGNETIC VARIATION WITH FLOATING DIAL COMPASS

Figure 20-65. Floating Needle Compass.

(h) A floating needle compass (figures 20-65a and 20-65b) has a needle with a north direction marked on it. The degree and direction marks are stationary on the bottom inside of the compass. The button and wrist compasses may be floating dial or floating needle. To determine the heading, line up the north-seeking arrow over 360° by rotating the compass. Then read the desired heading. Orienting a map with a floating needle compass is similar to the method used with the floating dial. The only exception is with the adjustment for magnetic variation. If magnetic variation is to the east, turn the map and the compass to the left (the north axis of the compass should be aligned with the map north) so that the magnetic north-seeking arrow is pointing at the number of degrees on the compass which corresponds with the angle of declination.

(i) When a compass is not available, map orientation requires a careful examination of the map and the ground to find linear features common to both, such as roads, railroads, fence lines, power lines, etc. By aligning the feature on the map with the same feature on the ground (figure 20-66), the map is oriented. Orientation by this method must be checked to prevent the reversal of directions which may occur if only one linear feature is used. This reversal may be prevented by aligning two or more map features (terrain or manmade). If no second linear feature is visible but the map user's position is known, a prominent object may be used. With the

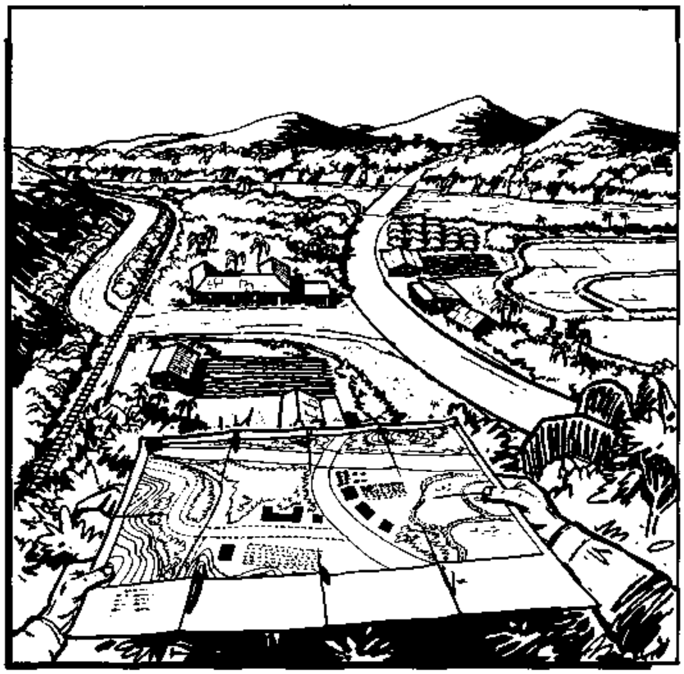

Figure 20-66. Map Orientation by Inspection.

prominent object and the user's position connected with a straight line on the map, the map is rotated until the line points toward the feature.

(j) If two prominent objects are visible and plotted on the map and the position is not known, move to one of the plotted and known positions, place the straightedge or protractor on the line between the plotted positions, turn the protractor and the map until the other plotted and visible point is seen along the edge. The map is then oriented.

(k) When a compass is not available and there are no recognizable prominent landforms or other features, a map may be oriented by any of the field expedient methods we will now discuss.

(8) Determining Cardinal Directions Using Field Expedients:

*(a) Shadow tip method of determining direction and time. This simple method of finding direction by the Sun consists of only three basic steps (figure 20-67).

-1. Step 1. Place a stick or branch into the ground at a fairly level spot where a distinct shadow will be cast. Mark the shadow tip with a stone, twig, or other means.

-2. Step 2. Wait until the shadow tip moves a few inches. If a 4-foot stick is being used, about 10 minutes should be sufficient. Mark the new position of the shadow tip in the same way as the first.

-3. Step 3. Draw a straight line through the two marks to obtain an approximate east-west line. If uncertain which direction is east and which is west, observe this simple rule: The Sun "rises in the east and sets in the west" (but rarely DUE east and DUE west). The shadow tip moves in just the opposite direction. Therefore, the first shadow-tip mark is always in the west direction, and the second mark in the east direction, anyplace on Earth.

(b) A line drawn at right angles to the east-west line at any point is the approximate north-south line, which will help orient a person to any desired direction of travel.

(c) Inclining the stick to obtain a more convenient shadow does not impair the accuracy of the shadow-tip method. Therefore, a traveler on sloping ground or in highly vegetated terrain need not waste valuable time looking for a large level area. A flat spot, the size of the hand, is all that is necessary for shadow-tip markings and the base of the stick can be either above, below, or to one side of it. Also, any stationary object (the end of a tree limb or the notch where branches are jointed) serves just as well as an implanted stick because only

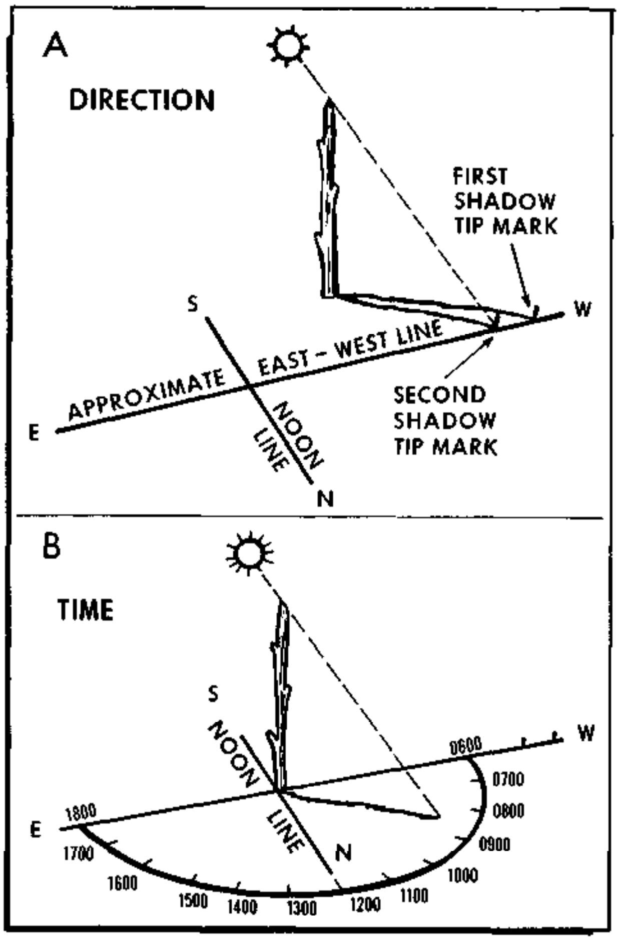

Figure 20-67. Determining Time and Direction by Shadow.

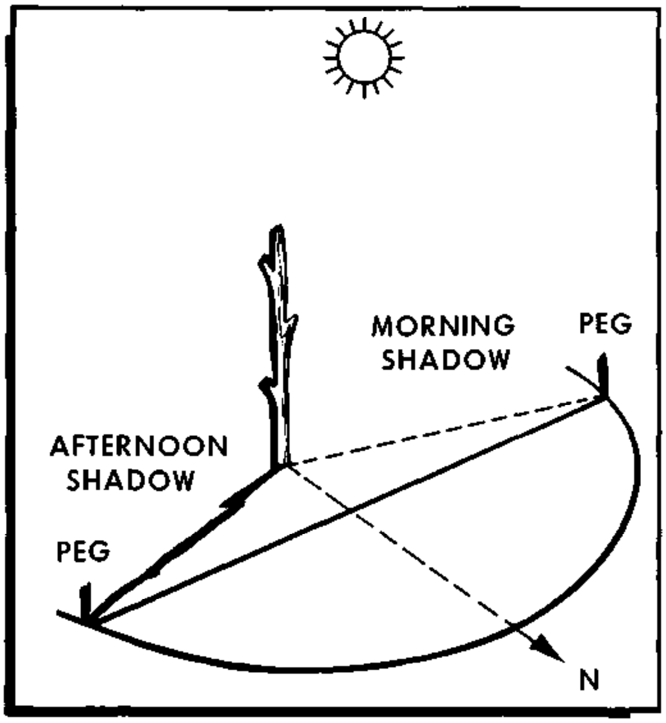

Figure 20-68. Equal Shadow Method of Determining Direction.

---

*From *Better Ways of Pathfinding*, by Robert S. Owendoff, 153 Cundry Drive, Falls Church VA 22046. 1964§ by Stackpole Company, Harrisburg PA. All rights reserved by copyright owner (author).

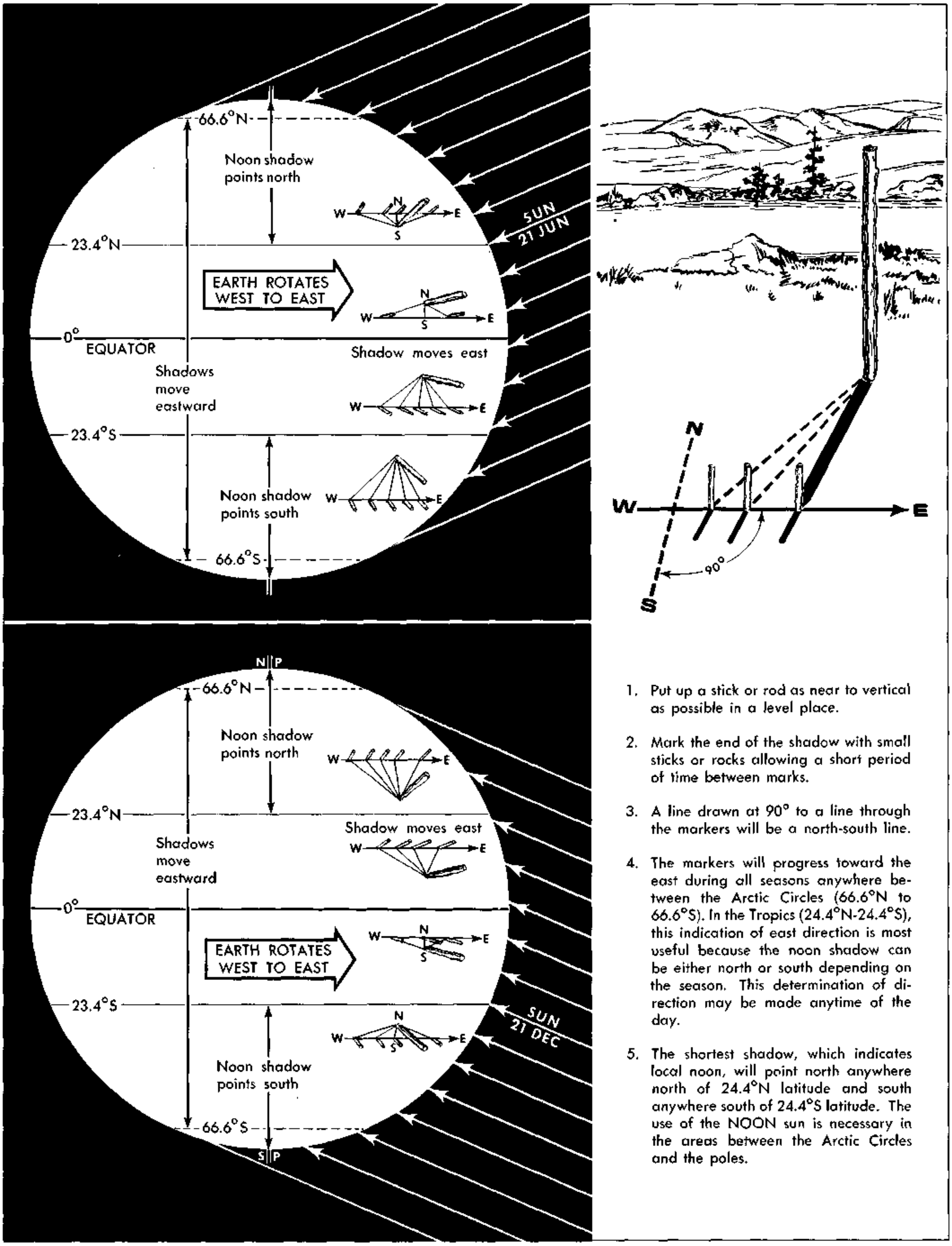
66.6°N
Noon shadow points north
W
N
S
E
23.4°N
SUN 21 JUN
EARTH ROTATES WEST TO EAST
N
W
S
E
0°
EQUATOR
Shadow moves east
Shadows move eastward
W
E
23.4°S
Noon shadow points south
W
E
66.6°S
N
W
E
S
90°
N P
66.6°N
Noon shadow points north
W
E
23.4°N
Shadow moves east
Shadows move eastward
W
E
0°
EQUATOR
W
N
E
S
EARTH ROTATES WEST TO EAST
23.4°S
N
W
S
E
SUN 21 DEC
Noon shadow points south
66.6°S
S P

1. Put up a stick or rod as near to vertical as possible in a level place.
2. Mark the end of the shadow with small sticks or rocks allowing a short period of time between marks.
3. A line drawn at 90° to a line through the markers will be a north-south line.
4. The markers will progress toward the east during all seasons anywhere between the Arctic Circles (66.6°N to 66.6°S). In the Tropics (24.4°N-24.4°S), this indication of east direction is most useful because the noon shadow can be either north or south depending on the season. This determination of direction may be made anytime of the day.
5. The shortest shadow, which indicates local noon, will point north anywhere north of 24.4°N latitude and south anywhere south of 24.4°S latitude. The use of the NOON sun is necessary in the areas between the Arctic Circles and the poles.

Figure 20-69. Stick and Shadow Method of Determining Direction.

the shadow tip is marked.

(d) The shadow-tip method can also be used to find the approximate time of day (figure 20-67B).

-1. To find the time of day, move the stick to the intersection of the east-west line and the north-south line, and set it vertically in the ground. The west part of the east-west line indicates the time is 0600 and the east part is 1800, ANYWHERE on Earth, because the basic rule always applies.

-2. The north-south line now becomes the noon line. The shadow of the stick is an hour hand in the shadow clock and with it the time can be estimated using the noon line and 6 o'clock line as the guides. Depending on the location and the season, the shadow may move either clockwise or counterclockwise, but this does not alter the manner of reading the shadow clock.

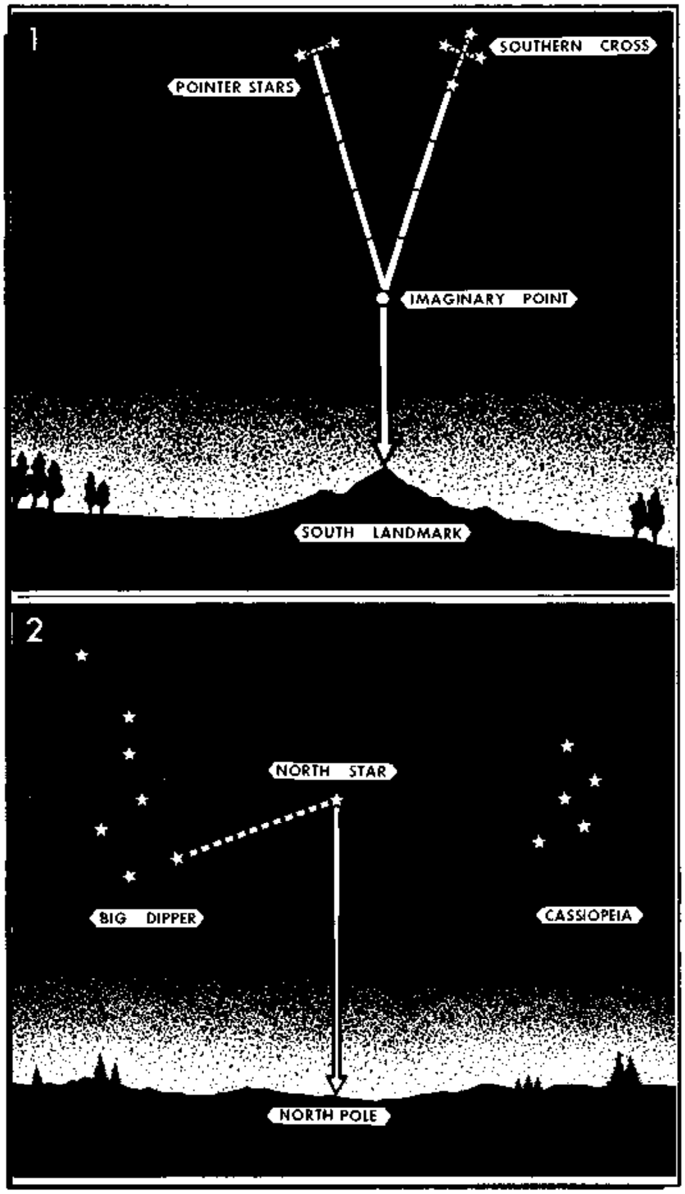

Figure 20-70. Determination of Direction by Using the Stars.

-3. The shadow clock is not a timepiece in the ordinary sense. It always reads 0600 at sunrise and 1800 at sunset. However, it does provide a satisfactory means of telling time in the absence of properly set watches. Being able to establish the time of day is important for such purposes as keeping a rendezvous, prearranged concerted action by separated persons or groups, and estimating the remaining duration of daylight. Shadow-clock time is closest to conventional clock time at mid-day, but the spacing of the other hours, compared to conventional time, varies somewhat with the locality and date.

(e) The shadow-tip system is ineffective for use beyond 66½° latitude in either hemisphere due to the position of the Sun above the horizon. Whether the Sun is north or south of a survivor at mid-day depends on the latitude. North of 23.4°N, the Sun is always due south at local noon and the shadow points north. South of 23.4°S, the Sun is always due north at local noon and the shadow points south. In the tropics, the Sun can be either north or south at noon, depending on the date and location but the shadow progresses to the east regardless of the date.

(f) Equal-shadow method of determining direction (Figures 20-68 and 20-69). This variation of the shadow-tip method is more accurate and may be used at all latitudes less than 66° at all times of the year.

-1. Step 1. Place a stick or branch into the ground vertically at a level spot where a shadow at least 12 inches long will be cast. Mark the shadow tip with a stone, twig, or other means. This must be done 5 to 10 minutes before noon (when the Sun is at its highest point (zenith)).

-2. Step 2. Trace an arc using the shadow as the radius and the base of the stick as the center. A piece of string, shoelace, or a second stick may be used to do this.

-3. Step 3. As noon is approached, the shadow becomes shorter. After noon, the shadow lengthens until it crosses the arc. Mark the spot as soon as the shadow tip touches the arc a second time.

-4. Step 4. Draw a straight line through the two marks to obtain an east-west line.

(g) Although this is the most accurate version of the shadow-tip method, it must be performed around noon. It requires the observer to watch the shadow and complete step 3 at the exact time the shadow tip touches the arc.

(h) At night, the stars may be used to determine the north line in the northern hemisphere or the south line in the southern hemisphere. Figure 20-70 shows how this is done.

(i) A watch can be used to determine the approximate true north or south (figure 20-71). In the northern hemisphere, the hour hand is pointed toward the Sun. A south line can be found midway between the hour hand

and 1200 standard time. During daylight savings time, the north-south line is midway between the hour hand and 1300. If there is any doubt as to which end of the line is north, remember that the Sun is in the east before noon and in the west in the afternoon.

(j) The watch may also be used to determine direction in the Southern Hemisphere; however, the method is different. The 1200-hour dial is pointed toward the Sun, and halfway between 1200 and the hour hand will be a north line. During daylight savings time, the north line lies midway between the hour hand and 1300.

(k) The watch method can be in error, especially in the extreme latitudes, and may cause "circling." To avoid this, make a shadow clock and set the watch to the time indicated. After traveling for an hour, take another shadow-clock reading.

(9) Determining Specific Position. When using a map and compass, the map must be oriented using the method described earlier in this chapter. Next, locate two or three known positions on the ground and the map. Using the compass, shoot an azimuth to one of the known positions (figure 20-72). Once the azimuth is determined, recheck the orientation of the map and plot the azimuth on the map. To plot the azimuth, place the front corner of the straightedge of the compass on the corresponding point on the map. Rotate the compass until the determined azimuth is directly beneath the stationary index line. Then draw a line along the straightedge of the compass and extend the line past the estimated position on the map. Repeat this procedure for the second point (figure 20-72). If only two azimuths are used, the technique is referred to as biangulation (figure 20-72). If a third azimuth is plotted to check the accuracy of the first two, the technique is called triangulation (figure 20-72). When using three lines, a triangle of error may be formed. If the triangle is large, the work should be checked. However, if a small triangle is formed, the user should evaluate the terrain to determine the actual position. One azimuth may be used with a linear land feature such as a river, road, railroad, etc., to determine specific position (figure 20-72).

(10) Determining Specific Location Without a Compass. A true north-south line determined by the stick and shadow, Sun and watch, or celestial constella-

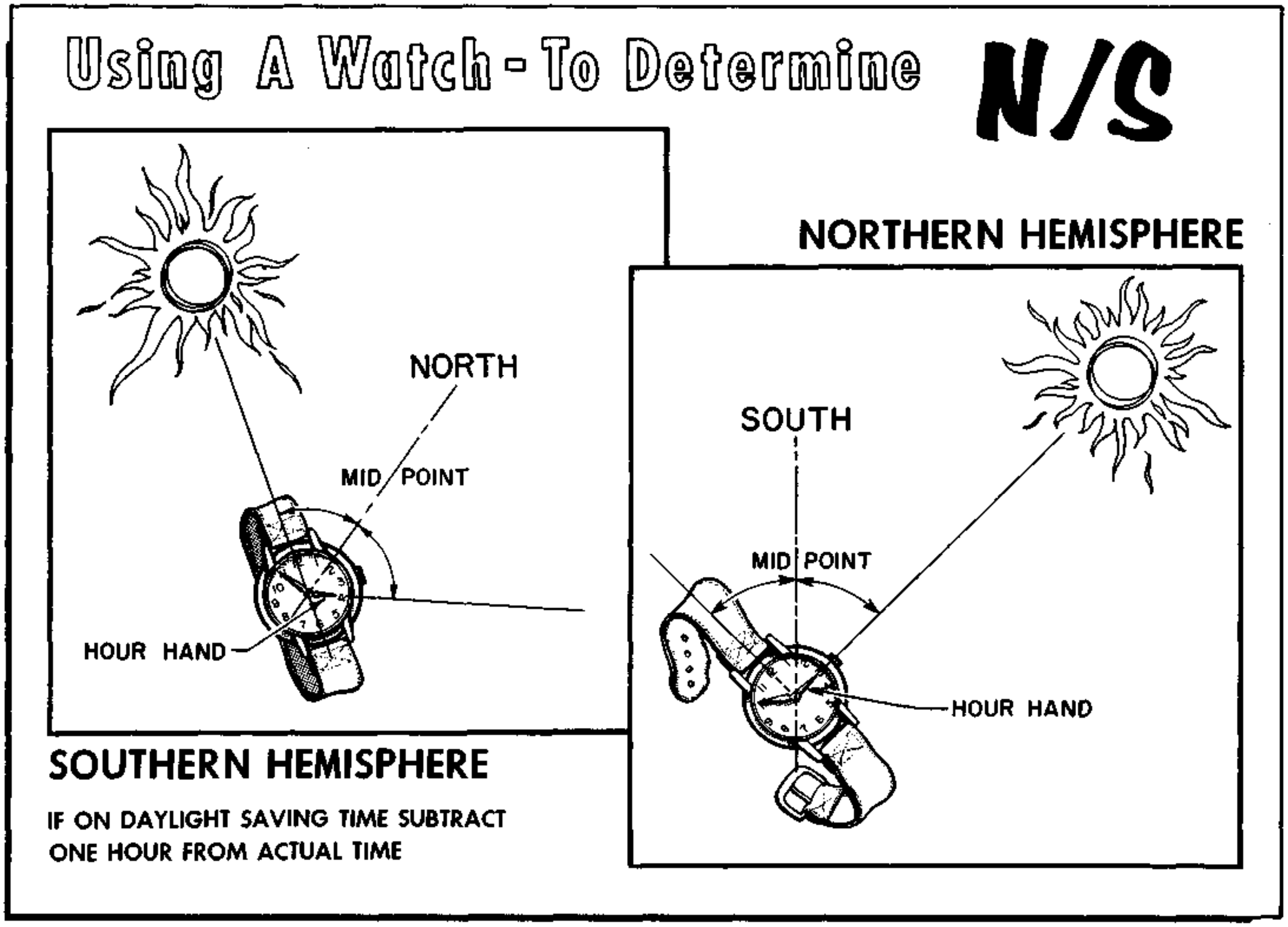
Using A Watch - To Determine N/S
NORTHERN HEMISPHERE
NORTH
MID POINT
HOUR HAND
SOUTH
MID POINT
HOUR HAND
SOUTHERN HEMISPHERE
IF ON DAYLIGHT SAVING TIME SUBTRACT
ONE HOUR FROM ACTUAL TIME

Figure 20-71. Directions Using a Watch.

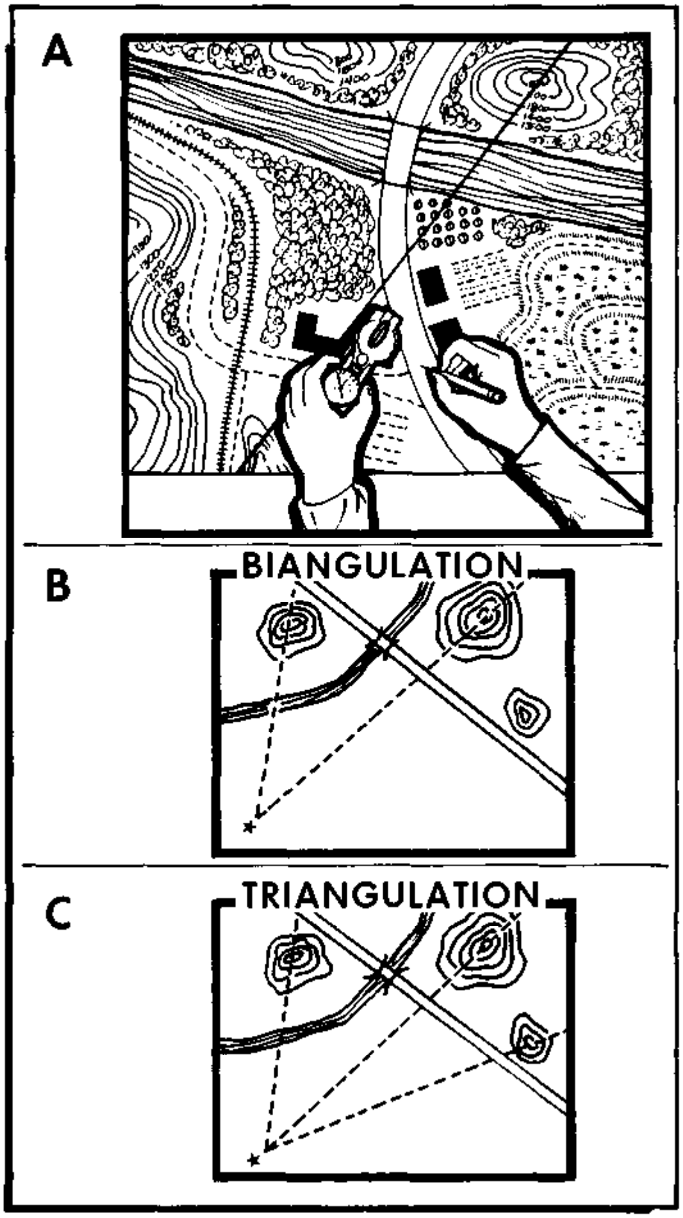

Figure 20-72. Azimuth, Biangulation, and Triangulation.

tion method may be used to orient the map without a compass. However, visible major land features can be used to orient the map to the lay of the land. Once the map is oriented, identify two or three landmarks and mark them on the map. Lay a straightedge on the map with the center of the straightedge at a known position as a pivot point and rotate the straightedge until the known position of the map is aligned with present position, and draw a line. Repeat this for the second and third position. Each time a line of position is plotted, the map must still be aligned with true north and south. If three lines of position are plotted and form a small triangle, use terrain evaluation to determine present position. If they form a large triangle, recheck calculations for errors.

(11) Dead Reckoning:

(a) Dead reckoning is the process of locating one's position by plotting the course and distance from the last known location. In areas where maps exist, even poor ones, travel is guided by them. It is a matter of knowing one's position at all times by associating the map features with the ground features. A great portion of the globe is unmapped or only small scale maps are available. The survivor may be required to travel in these areas without a usable map. Although these areas could be anywhere, they are more likely to be found in frozen wastelands and deserts.

(b) For many centuries, mariners used dead reckoning to navigate their ships when they were out of sight of land or during bad weather, and it is just as applicable to navigation on land. Movement on land must be carefully planned. In military movement, the starting location and destination are known, and if a map is available, they are carefully plotted along with any known intermediate features along the route. These intermediate features, if clearly recognizable on the ground, serve as checkpoints. If a map is not available, the plotting is done on a blank sheet of paper. A scale is selected so the entire route will fit on one sheet. A north direction is clearly established. The starting point and destination are then plotted in relationship to each other. If the terrain and enemy situations permit, the ideal course is a straight line from starting point to destination. This is seldom possible or practicable. The route of

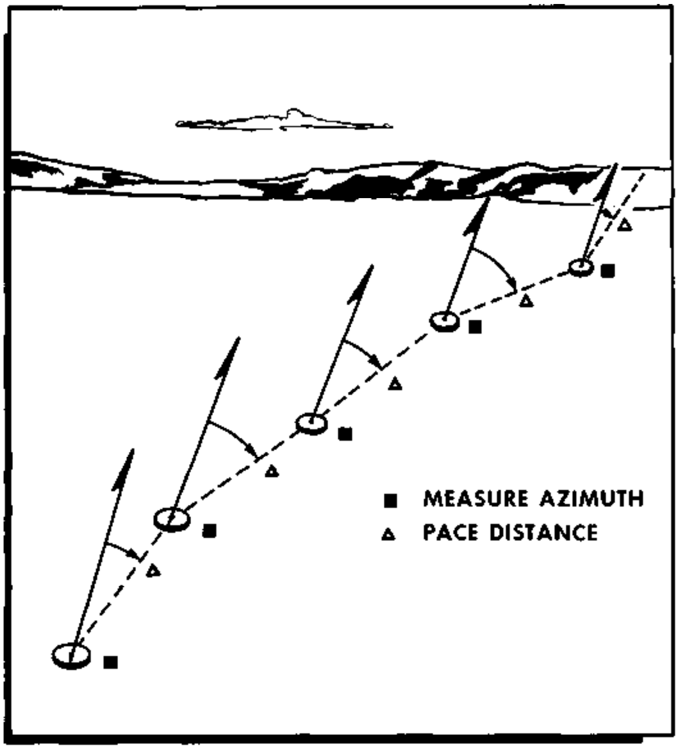

Figure 20-73. Compass Navigation on Foot.

1	2	3	4	5	6	7
ODOMETER READING AT START OF EACH COURSE	DISTANCE IN MILES	FORWARD AZIMUTH (MAGNETIC)	DECLINATION CORRECTION	DEVIATION CORRECTION	TRUE AZIMUTH	NOTES
A 4750						
	6	17°	13°	+3°	33°	
B 4756						
	9	358°	"	+2°	13°	
C 4765						
	8	341°	"	+1°	355°	
D 4773						
	1	314°	"	0°	327°	
E 4774						
	1.5	341°	"	+1°	355°	
F 4775	5					
	1.5	322°	"	0°	335°	
G 4777						
	1	312°	"	0°	325°	
H 4778						
	12	300°	12°	–1°	311°	
I 4790						
	6	341°	"	+1°	354°	
J 4796						
	6	302°	"	–1°	313°	
K 4802						
	20	319°	"	0°	331°	
4810						Wahoo River
						Crossing
4814						Cut 2 miles
L 4824			Base Camp (data)			

Figure 20-74. Sample Log.

travel usually consists of several courses, with an azimuth established at the starting point for the first course to be followed. Distance measurement begins with the departure and continues through the first course until a change in direction is made. A new azimuth is established for the second course and the distance is measured until a second change of direction is made, and so on. Records of all data are kept and all positions are plotted.

(c) A pace (for our purposes) is equal to the distance covered every time the same foot touches the ground (surveyor's paces). To measure distance, count the number of paces in a given course and convert to the map unit. Usually, paces are counted in hundreds, and hundreds can be kept track of in many ways: making notes in a record book; counting individual fingers; placing small objects such as pebbles into an empty pocket; tying knots in a string; or using a mechanical hand counter. Distances measured this way are only approximate, but with practice can become very accurate. It is important that each person who uses dead reckoning navigation establish the length of an average pace. This is done by pacing a measured course many times and computing the mean (figure 20-73). In the field, an average pace must often be adjusted because of the following conditions:

-1. Slopes. The pace lengthens on a downgrade and shortens on an upgrade.

-2. Winds. A headwind shortens the pace while a tailwind increases it.

-3. Surfaces. Sand, gravel, mud and similar surface materials tend to shorten the pace.

-4. Elements. Snow, rain, or ice reduces the length of the pace.

-5. Clothing. Excess weight of clothing shortens the pace while the type of shoes affects traction and therefore the pace length.

(d) A log (figure 20-74) should be used for navigation, by dead reckoning, to record all of the distances and azimuths while traveling. Often, relatively short stretches of travel cannot be traversed in a straight course because of some natural features such as a river, or a steep, rugged slope. This break in normal navigation is shown on the log to ensure proper plotting.

(e) The course of travel may be plotted directly on the face of the map or on a separate piece of paper at the same scale as the map. If the latter method is chosen, the complete plot can be transferred to the map sheet, if at least one point of the plot is also shown on the map. The actual plotting can be done by protractor and scale. The degree of accuracy obtained depends upon the quality of draftmanship, the environmental conditions, and the care taken in obtaining data while en route. Figure 20-75 illustrates a paper plot of the data obtained for the log sample in figure 20-74. It should be noted that four of the courses from A to H are short and have been plotted as a single course, equal to the sum of the four distances and using a mean azimuth of the four. This is recommended because it saves time without a loss of accuracy. If possible, a plot should be tied into at least one known intermediate point along the route. This is done by directing the route to pass near or over a point. If the plotted position of the intermediate point differs from its known location, discard the previous plot and start a new plot from the true location. The previous plot may be inspected to see if there is a detectable constant error applicable to future plots; otherwise, it is of no further use.

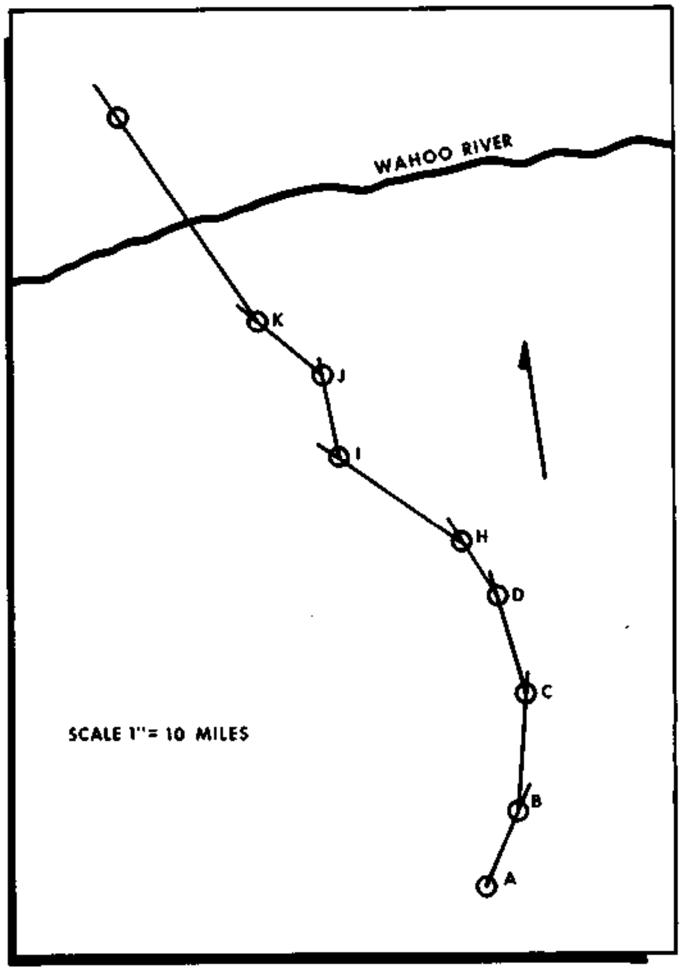

Figure 20-75. Separate Paper Plot.

(f) An offset is a planned magnetic deviation to the right or left of an azimuth to an objective. It is used when approaching a linear feature from the side, and a point along the linear feature (such as a road junction) is the objective. Because of errors in the compass, or in map reading, one may reach the linear feature and not know whether the objective lies to the right or left. A deliberate offset by a known number of degrees in a known direction compensates for possible errors and ensures that, upon reaching the linear feature, the user knows whether to go right or left to reach the objective.

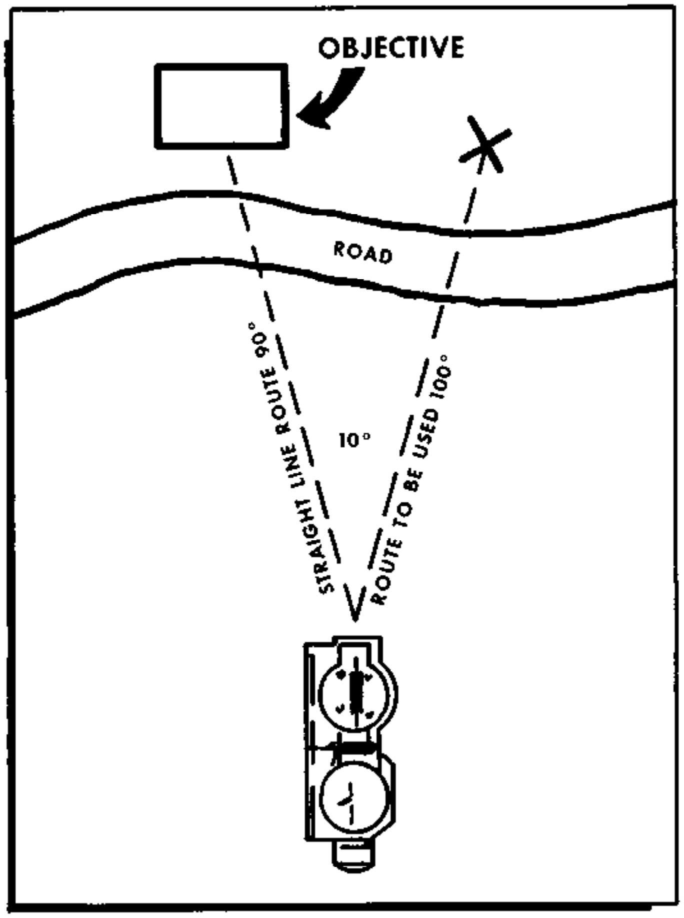

Figure 20-76. Deliberate Offset.

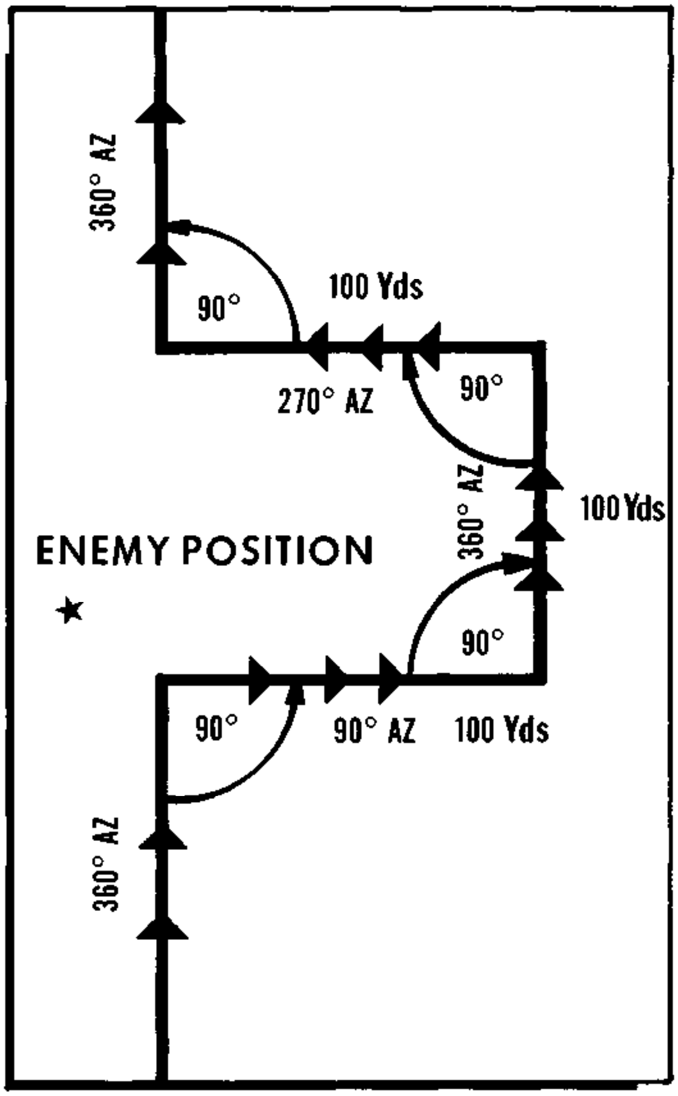

Figure 20-77. Detour Around Energy Position.

Figure 20-76 shows an example of the use of offset to approach an objective. It should be remembered that the distance from "X" to the objective will vary directly with the distance to be traveled and the number of degrees offset. Each degree offset will move the course about 20 feet to the right or left for each 1,000 feet traveled. For example: In figure 20-76, the number of degrees offset is 10 to the right. If the distance traveled to "X" is 1,000 feet, then "X" is located about 200 feet to the right of the objective.

(g) Figure 20-77 shows an example of how to bypass enemy positions or obstacles by detouring around them and maintaining orientation by moving at right angles for specified distances; for example, moving on an azimuth 360°and wish to bypass an obstacle or position. Change direction to 90° and travel for 100 yards, change direction back to 360° and travel for 100 yards, change direction to 270° and travel for 100 yards, then change direction to 360°, and back on the original azimuth. Bypassing an unexpected obstacle at night is done in the same way.

(12) Polar Coordinates:

(a) A point on the map may be determined or plotted from a known point by giving a direction and a distance along the direction line. This method of point location uses polar coordinates (figure 20-78). The reference direction is normally expressed as an azimuth, and the distance is determined by any convenient unit of measurement such as meters or yards.

(b) Polar coordinates are especially useful in the field because magnetic azimuth is determined from the compass and the distance can be estimated.

(13) Position Determination:

(a) Determining Latitude. (From the Sun at sunrise and sunset), Figure 20-79 shows the true azimuth of the rising Sun and the relative bearing of the setting Sun for all of the months in the year in the Northern and Southern Hemispheres (the table assumes a level horizon and is inaccurate in mountainous terrain).

-1. Latitude can be determined by using a compass to find the angle of the Sun at sunrise or sunset (subtracting or adding magnetic variation) and the date. According to the chart in figure 20-79, on January 26th, the azimuth of the rising Sun will be 120° to the left when facing the Sun in the Northern Hemisphere (it would be 120° to the right for setting Sun); therefore, the latitude would be 50°. If in the Southern Hemisphere, the direction of the Sun would be the opposite.

-2. The table does not list every day of the year, nor does it list every degree of latitude. If accuracy is desired within 1° of azimuth, interpolation may be nec-

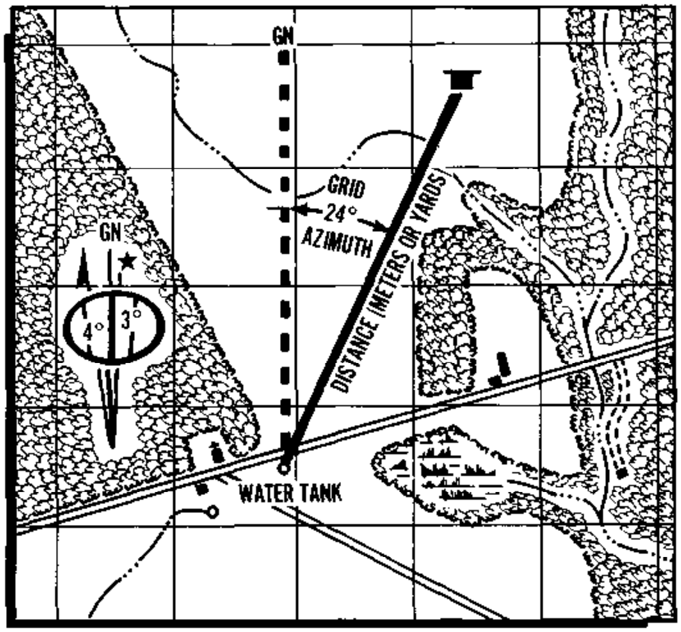

Figure 20-78. Polar Coordinates Used to Designate Position on Map.

DATE		Angle to North from the rising or setting Sun (level terrain)												
		LATITUDE												
		0°	5°	10°	15°	20°	25°	30°	35°	40°	45°	50°	55°	60°
JANUARY	1	113	113	113	114	115	116	117	118	121	124	127	133	141
	6	112	113	113	113	114	115	116	118	120	123	127	132	140
	11	112	112	112	113	113	114	115	117	119	122	125	130	138
	16	111	111	111	112	112	113	114	116	118	120	124	129	136
	21	110	110	110	111	111	112	113	115	117	119	122	127	133
	26	109	109	109	109	110	111	112	113	115	117	120	124	130
FEBRUARY	1	107	107	108	108	108	109	110	111	113	115	117	121	126
	6	106	106	106	106	107	107	108	109	111	113	115	118	123
	11	104	104	105	105	105	106	107	108	109	110	112	116	120
	16	103	103	103	103	103	104	105	106	107	108	110	112	116
	21	101	101	101	101	101	102	102	103	104	105	107	109	112
	26	99	99	99	99	100	100	100	101	102	103	104	106	108
MARCH	1	98	98	98	98	99	99	99	100	100	101	102	104	106
	6	96	96	96	96	96	97	97	97	98	98	99	100	102
	11	94	94	94	94	94	94	95	95	95	96	96	97	98
	16	92	92	92	92	92	92	92	92	93	93	93	93	94
	21	90	90	90	90	90	90	90	90	90	90	90	90	90
	26	88	88	88	88	88	88	88	88	87	87	87	87	86
APRIL	1	86	86	86	86	85	85	85	85	84	84	83	82	81
	6	84	84	84	83	83	83	83	82	82	81	80	79	77
	11	82	82	82	82	81	81	81	80	80	79	77	76	74
	16	80	80	80	80	79	79	78	78	77	76	74	72	70
	21	78	78	78	78	78	77	76	76	75	73	72	69	66
	26	77	77	76	76	76	75	75	74	72	71	69	66	63
MAY	1	75	75	75	74	74	73	73	72	70	69	66	63	59
	6	74	74	73	73	73	72	71	70	68	67	64	61	56
	11	72	72	72	72	71	70	69	68	67	64	62	58	52
	16	71	71	71	70	70	69	68	67	65	63	60	55	49
	21	70	70	70	69	69	68	67	65	63	61	58	53	47
	26	69	69	69	68	68	67	66	64	62	60	56	51	44
JUNE	1	68	68	68	67	66	66	64	63	61	58	54	49	41
	6	67	67	67	67	66	65	64	62	60	57	53	48	40
	11	67	67	67	66	65	64	63	62	59	56	53	47	39
	16	67	67	67	66	65	64	63	62	59	56	53	47	39
	21	67	67	67	66	65	64	63	62	59	56	53	47	39
	26	67	67	67	66	65	64	63	62	59	56	53	47	39
JULY	1	67	67	67	66	65	64	63	62	59	56	53	47	39
	6	67	67	67	66	66	65	64	62	60	57	53	48	40
	11	68	68	68	67	66	65	64	63	61	58	54	49	41
	16	69	68	68	68	67	66	65	64	62	59	55	50	43
	21	69	69	69	69	68	67	66	65	63	60	57	52	45
	26	70	70	70	70	69	68	67	66	64	62	59	54	48
AUGUST	1	72	72	72	71	71	70	69	68	66	64	61	57	51
	6	73	73	73	73	72	71	71	69	68	66	63	60	55
	11	75	75	74	74	74	73	72	71	70	68	66	63	58
	16	76	76	76	76	75	75	74	73	72	70	68	65	61
	21	78	78	77	77	77	76	76	75	74	72	71	68	65
	26	79	79	79	79	79	78	78	77	76	75	73	71	68
SEPTEMBER	1	82	82	82	81	81	81	80	80	79	78	77	75	73
	6	83	83	83	83	83	83	82	82	81	81	80	78	77
	11	85	85	85	85	85	85	85	84	84	83	83	82	81
	16	87	87	87	87	87	87	87	86	86	86	85	85	84
	21	89	89	89	89	89	89	89	89	89	89	88	88	88
	26	91	91	91	91	91	91	91	91	91	91	92	92	92
OCTOBER	1	93	93	93	93	93	93	93	94	94	94	95	95	96
	6	95	95	95	95	95	96	96	96	97	97	98	99	100
	11	97	97	97	97	97	98	98	99	99	100	101	102	104
	16	99	99	99	99	99	100	100	101	101	102	104	105	108
	21	101	101	101	101	101	102	102	103	104	105	107	109	112
	26	102	102	103	103	103	104	104	105	106	108	109	112	115
NOVEMBER	1	104	104	105	105	105	106	107	108	109	110	113	116	120
	6	106	106	106	107	107	108	109	110	111	113	115	119	123
	11	107	107	108	108	108	109	110	111	113	115	117	121	126
	16	109	109	109	109	110	111	112	113	115	117	120	124	130
	21	110	110	110	111	111	112	113	114	116	119	122	126	133
	26	111	111	111	112	112	113	114	116	118	120	124	128	135
DECEMBER	1	112	112	112	113	113	114	115	117	119	122	125	130	138
	6	112	112	113	113	114	115	116	118	120	123	126	132	140
	11	113	113	113	114	115	116	117	118	121	124	127	133	141
	16	113	113	113	114	115	116	117	118	121	124	127	133	141
	21	113	113	113	114	115	116	117	118	121	124	127	133	141
	26	113	113	113	114	115	116	117	118	121	124	127	133	141

NOTE: When the Sun is rising, the angle is reckoned from East to North.
When the Sun is setting, the angle is reckoned from West to North.

Figure 20-79. Finding Direction from the Rising or Setting Sun.

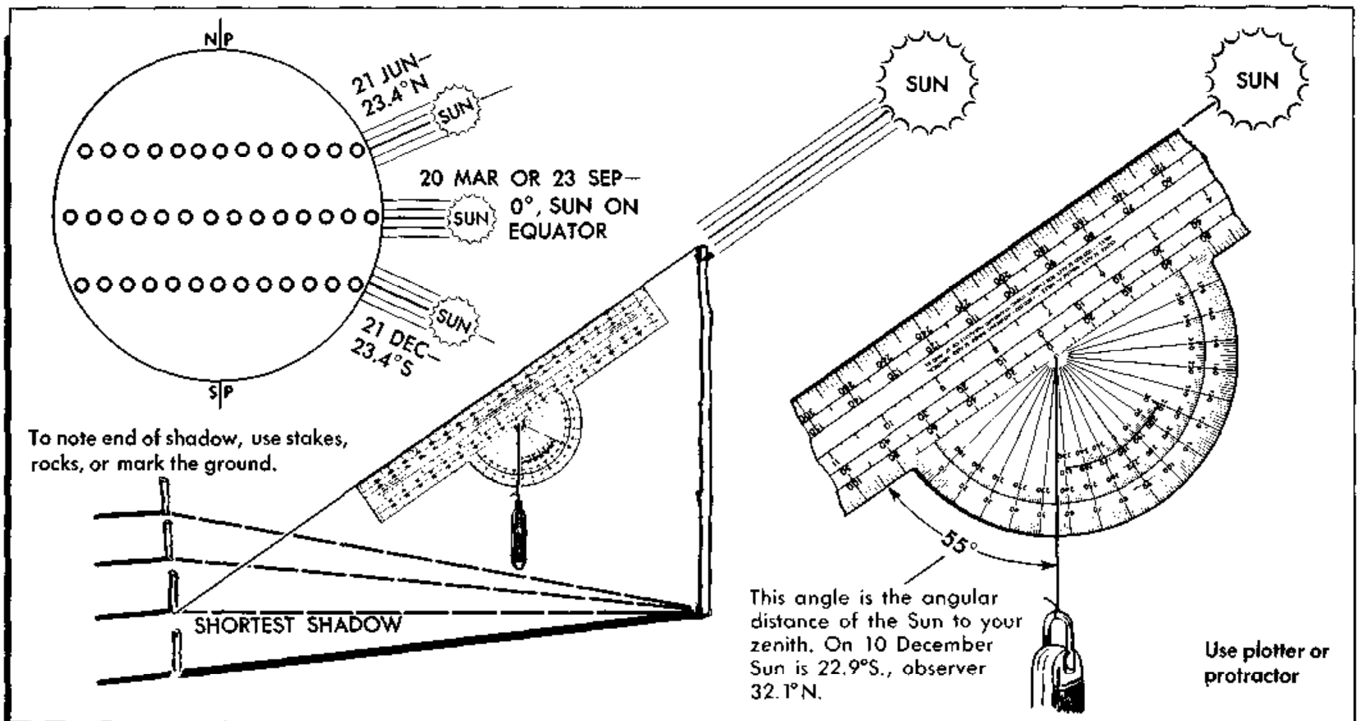

To note end of shadow, use stakes, rocks, or mark the ground.

This angle is the angular distance of the Sun to your zenith. On 10 December Sun is 22.9°S., observer 32.1°N.

Use plotter or protractor

DECLINATION OF SUN

(IN DEGREES AND TENTHS OF A DEGREE)

Declination is tabulated to the nearest tenth of a degree rather than to the nearest minute of arc. To convert 1/10° (0.1°) to minutes, multiply by 6. (ie. 27.9°=27° 54′)

DAY	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1	S 23.1	S 17.5	S 7.7	N 4.4	N 15.0	N 22.0	N 23.1	N 18.1	N 8.4	S 3.1	S 14.3	S 21.8
2	23.0	17.2	7.3	4.8	15.3	22.1	23.1	17.9	8.1	3.4	14.6	21.9
3	22.9	16.9	6.9	5.2	15.6	22.3	23.0	17.6	7.7	3.8	15.0	22.1
4	22.9	16.6	6.6	5.6	15.9	22.4	22.9	17.3	7.3	4.2	15.3	22.2
5	22.8	16.3	6.2	5.9	16.2	22.5	22.8	17.1	7.0	4.6	15.6	22.3
6	S 22.7	S 16.0	S 5.8	N 6.3	N 16.4	N 22.6	N 22.7	N 16.8	N 6.6	S 5.0	S 15.9	S 22.5
7	22.5	15.7	5.4	6.7	16.7	22.7	22.6	16.5	6.2	5.4	16.2	22.6
8	22.4	15.4	5.0	7.1	17.0	22.8	22.5	16.3	5.8	5.7	16.5	22.7
9	22.3	15.1	4.6	7.4	17.3	22.9	22.4	16.0	5.5	6.1	16.8	22.8
10	22.2	14.8	4.2	7.8	17.5	23.0	22.3	15.7	5.1	6.5	17.1	22.9
11	S 22.0	S 14.5	S 3.8	N 8.2	N 17.8	N 23.1	N 22.2	N 15.4	N 4.7	S 6.9	S 17.3	S 23.0
12	21.9	14.1	3.5	8.6	18.0	23.1	22.0	15.1	4.3	7.3	17.6	23.1
13	21.7	13.8	3.1	8.9	18.3	23.2	21.9	14.8	3.9	7.6	17.9	23.1
14	21.5	13.5	2.7	9.3	18.5	23.2	21.7	14.5	3.6	8.0	18.1	23.2
15	21.4	13.1	2.3	9.6	18.8	23.3	21.6	14.2	3.2	8.4	18.4	23.3
16	S 21.2	S 12.8	S 1.9	N 10.0	N 19.0	N 23.3	N 21.4	N 13.9	N 2.8	S 8.8	S 18.7	S 23.3
17	21.0	12.4	1.5	10.4	19.2	23.4	21.3	13.5	2.4	9.1	18.9	23.3
18	20.8	12.1	1.1	10.7	19.5	23.4	21.1	13.2	2.0	9.5	19.1	23.4
19	20.6	11.7	0.7	11.1	19.7	23.4	20.9	12.9	1.6	9.9	19.4	23.4
20	20.4	11.4	0.3	11.4	19.9	23.4	20.7	12.6	1.2	10.2	19.6	23.4
21	S 20.2	S 11.0	N 0.1	N 11.7	N 20.1	N 23.4	N 20.5	N 12.2	N 0.8	S 10.6	S 19.8	S 23.4
22	20.0	10.7	0.5	12.1	20.3	23.4	20.4	11.9	0.5	10.9	20.1	23.4
23	19.8	10.3	0.9	12.4	20.5	23.4	20.2	11.6	N 0.1	11.3	20.3	23.4
24	19.5	9.9	1.3	12.7	20.7	23.4	20.0	11.2	S 0.3	11.6	20.5	23.4
25	19.3	9.6	1.7	13.1	20.9	23.4	19.7	10.9	0.7	12.0	20.7	23.4
26	S 19.0	S 9.2	N 2.1	N 13.4	N 21.1	N 23.4	N 19.5	N 10.5	S 1.1	S 12.3	S 20.9	S 23.4
27	18.8	8.8	2.5	13.7	21.2	23.3	19.3	10.2	1.5	12.7	21.1	23.3
28	18.5	8.5	2.9	14.0	21.4	23.3	19.1	9.8	1.9	13.0	21.3	23.3
29	18.3	8.1	3.2	14.4	21.6	23.3	18.8	9.5	2.3	13.3	21.4	23.3
30	18.0	. . .	3.6	14.7	21.7	23.2	18.6	9.1	2.7	13.7	21.6	23.2
31	S 17.7	. . .	N 4.0	. . .	N 21.9	. . .	N 18.4	N 8.8	. . .	S 14.0	. . .	S 23.1

EXAMPLE: On 10 December the declination of the Sun is 22.9°S., so observers who measure the zenith distance as 0° would know that they are at latitude 22.9°S. If they measure a zenith distance of 5° with the Sun south of this zenith, they are 5° north of 22.9°S, or at a latitude 17.9°S; and if the Sun is north, they are 5° south of 22.9°S, or latitude 27.9°S.

Figure 20-80. Determining Latitude by Noon Sun.

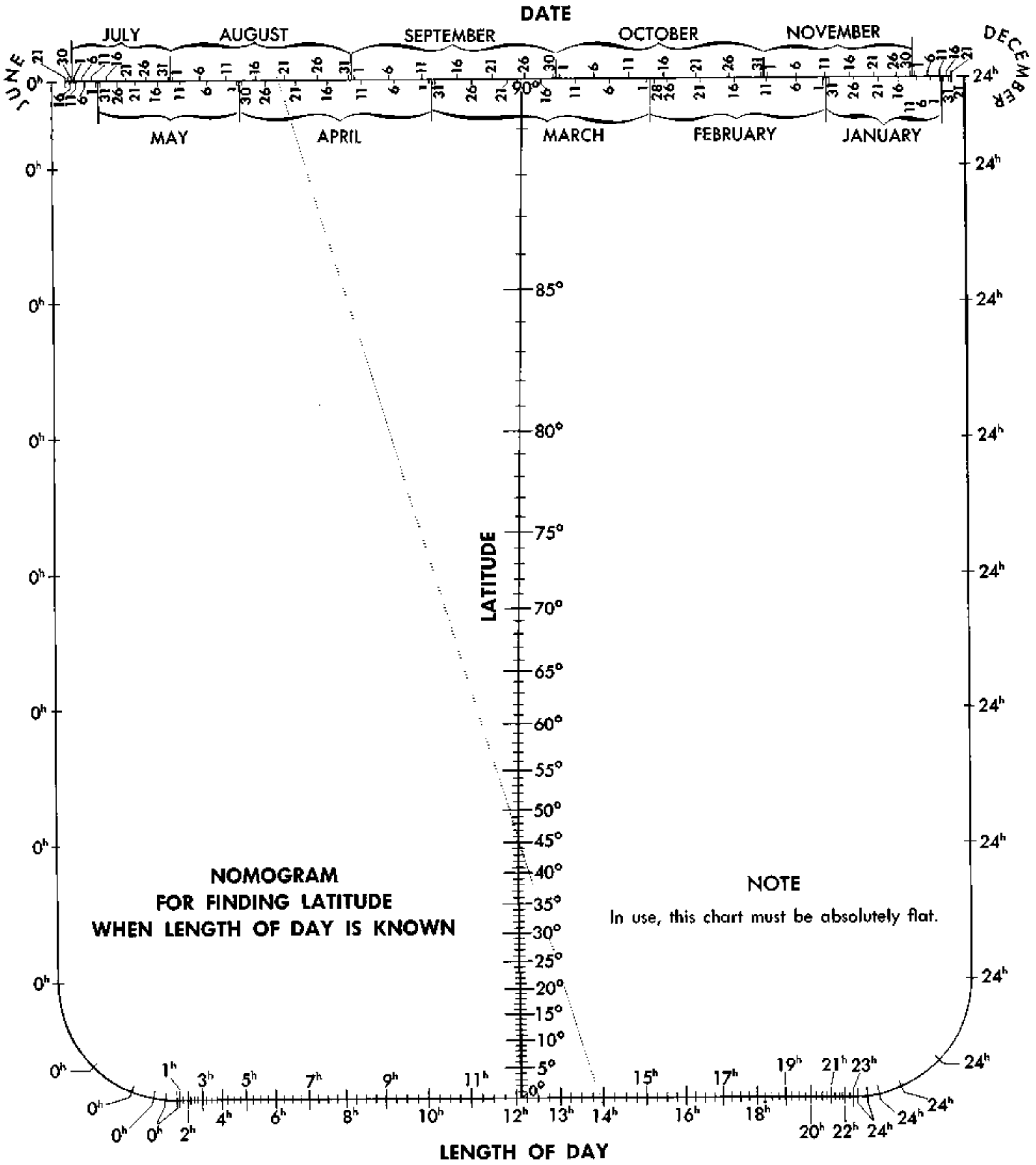

NOMOGRAM FOR FINDING LATITUDE WHEN LENGTH OF DAY IS KNOWN

NOTE

In use, this chart must be absolutely flat.

INSTRUCTIONS

To find your latitude:

In Northern Latitudes:

1. Find length of the day from the instant the top of the Sun appears above the ocean horizon to the instant it disappears below the horizon. This instant is often marked by a green flash.
2. Lay a straight edge or stretch across the Nomogram, connecting the observed length of the day on the Length of Day Scale with the date on the Date Scale.
3. Read your latitude on the Latitude Scale.

EXAMPLE: On August 20, observed length of the day is 13 hours and 54 minutes. Latitude is 45°30′N.

In Southern Latitudes:

Add 6 months to the date and proceed as in northern latitudes.

EXAMPLE: On 11 May observed length of day is 10 hours and 4 minutes. Adding 6 months gives 11 November. Latitude is 41°30′S.

Figure 20-81. Nomogram.

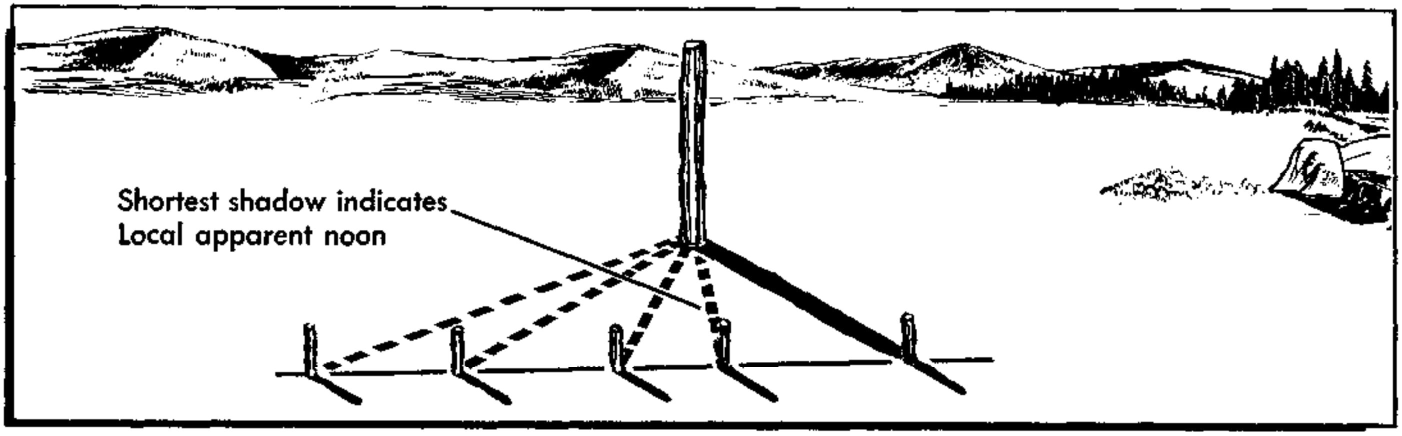

Figure 20-82. Stick and Shadow.

essary (split the difference) between the values given in the table. For example, between 45° latitude and 50° latitude is 5°. The difference in latitudes (5°) and the difference in azimuths (3°) split (5⁄3) is 1°⅔′(1°40′), so the more accurate degree of latitude would be 46°40′ latitude.

(b) Latitude by Noon Altitude of the Sun. On any given day, there is only one latitude on Earth where the Sun passes directly overhead or through the zenith at noon. In all latitudes north of this, the Sun passes to the south of the zenith; and in those south of it, the Sun passes to the north. For each 1° change of latitude, the zenith distance also changes by 1 degree. Figure 20-80 gives the latitude for each day of the year where the Sun is in the zenith at noon. If a Weems plotter or other protractor is available, maximum altitude of the Sun should be used to find latitude by measuring the angular distance of the Sun from the zenith at noon. Local noon can be found using the methods described earlier. Stretch a string from the top of a stick to the point where the end of the noon shadow rested, place the plotter along the string and drop a plumb line from the center of the plotter. The intersection of the plumb line with the outer scale of the plotter shows the angular distance of the Sun from the zenith.

(c) Latitudes by Length of Day. This method is used most effectively while on open seas. When in any latitude between 60°N and 60°S, the exact latitude within 30 nautical miles (½°) can be determined if the length of the day within 1 minute is known. This is true throughout the year except for about 10 days before and 10 after the equinoxes—approximately 11-31 March and 12 September-2 October. During these two periods, the day is about the same length at all latitudes. A level horizon is required to time sunrise and sunset accurately. Find the length of day from the instant the top of the Sun first appears above the ocean horizon to the instant it disappears below the horizon. This instant is often marked by a green flash. Write down the times of sunrise and sunset. Don't count on remembering them.

Date	Eq. of Time*	Date	Eq. of Time*	Date	Eq. of Time*	Date	Eq. of Time*	Date	Eq. of Time*	Date	Eq. of Time*
Jan. 1	−3.5 min.	Mar. 4	−12.0	May 2	+3.0 min.	Aug. 4	−6.0	Oct. 1	+10.0 min.	Dec. 1	+11.0
2	−4.0	8	−11.0	14	+3.8	12	−5.0	4	+11.0	4	+10.0
4	−5.0	12	−10.0	May 28	+3.0	17	−4.0	7	+12.0	6	+9.0
7	−6.0	16	−9.0	June 4	+2.0	22	−3.0	11	+13.0	9	+8.0
9	−7.0	19	−8.0	9	+1.0	26	−2.0	15	+14.0	11	+7.0
12	−8.0	22	−7.0	14	0.0	Aug. 29	−1.0	20	+15.0	13	+6.0
14	−9.0	26	−6.0	19	−1.0	Sept. 1	0.0	Oct. 27	+16.0	15	+5.0
17	−10.0	Mar. 29	−5.0	23	−2.0	5	+1.0	Nov. 4	+16.4	17	+4.0
20	−11.0	Apr. 1	−4.0	June 28	−3.0	8	+2.0	11	+16.0	19	+3.0
24	−12.0	5	−3.0	July 3	−4.0	10	+3.0	17	+15.0	21	+2.0
Jan. 28	−13.0	8	−2.0	9	−5.0	13	+4.0	22	+14.0	23	+1.0
Feb. 4	−14.0	12	−1.0	18	−6.0	16	+5.0	25	+13.0	25	0.0
13	−14.3	16	0.0	July 27	−6.6	19	+6.0	Nov. 28	+12.0	27	−1.0
19	−14.0	20	+1.0			22	+7.0			29	−2.0
Feb. 28	−13.0	Apr. 25	+2.0			25	+8.0			Dec. 31	−3.0
						Sep. 28	+9.0				

* Add plus values to mean time and subtract minus values from mean time to get apparent time.

Figure 20-83. Equation of Time.

Note that only the length of day counts in the determination of latitude; a watch may have an unknown error and yet serve to determine this factor. If only one water horizon is available, as on a seacoast, find local noon by the stick and shadow method. The length of day is twice the interval from sunrise to noon or from noon to sunset. Knowing the length of day, latitude can be found by using the nomogram shown in figure 20-81.

(d) Longitude from Local Apparent Noon. To find longitude, a survivor must know the correct time and the rate at which a watch gains or loses time. If this rate and the time the watch was last set is known, the correct time can be computed by adding or subtracting the gain or loss. Correct the zone time on the watch to Greenwich time; for example, if the watch is on eastern standard time, add 5 hours to get Greenwich time. Longitude can be determined by timing the moment a celestial body passes the meridian. The easiest body to use is the Sun. Use one of the following methods:

-1. Stick and Shadow. Put up a stick or rod (figure 20-82) as nearly vertical as possible in a level place. Check the alignment of the stick by sighting along the line of a makeshift plumb bob. (To make a plumb bob, tie any heavy object to a string and let it hang free. The line of the string indicates the vertical.) Sometime before midday, begin marking the position of the end of the stick's shadow. Note the time for each mark. Continue marking until the shadow definitely lengthens. The time of the shortest shadow is the time when the Sun passed the local meridian or local apparent noon. A survivor will probably have to estimate the position of the shortest shadow by finding a line midway between two shadows of equal length, one before noon and one after. If the times of sunrise and sunset are accurately determined on a water horizon, local noon will be midway between these times.

-2. Double Plumb Bob. Erect two plumb bobs about 1 foot apart so that both strings line up on Polaris, much the same as a gun sight. Plumb bobs should be set up when Polaris is on the meridian and has no east-west correction. The next day, when the shadows of the two plumb bobs coincide, they will indicate local apparent noon.

-3. Mark Down the Greenwich Time of Local Apparent Noon. The next step is to correct this observed time of meridian passage for the equation of time; that is, the number of minutes the real Sun is ahead of or behind the mean Sun. (The mean Sun was invented by astronomers to simplify the problems of measuring time. Mean Sun rolls along the Equator at a constant rate of 15° per hour. The real Sun is not so considerate; it changes its angular rate of travel around the Earth with the seasons.) Figure 20-83 gives the value in minutes of time to be added to or subtracted from mean (watch) time to get apparent (Sun) time.

-4. After computing the Greenwich time of local noon, the difference of longitude between the survivor's position and Greenwich can be found by converting the interval between 1200 Greenwich and the local noon from time to arc. Remember that 1 hour equals 15° of longitude, 4 minutes equal 1° of longitude, and 4 seconds equal 1′ of longitude. Example: The

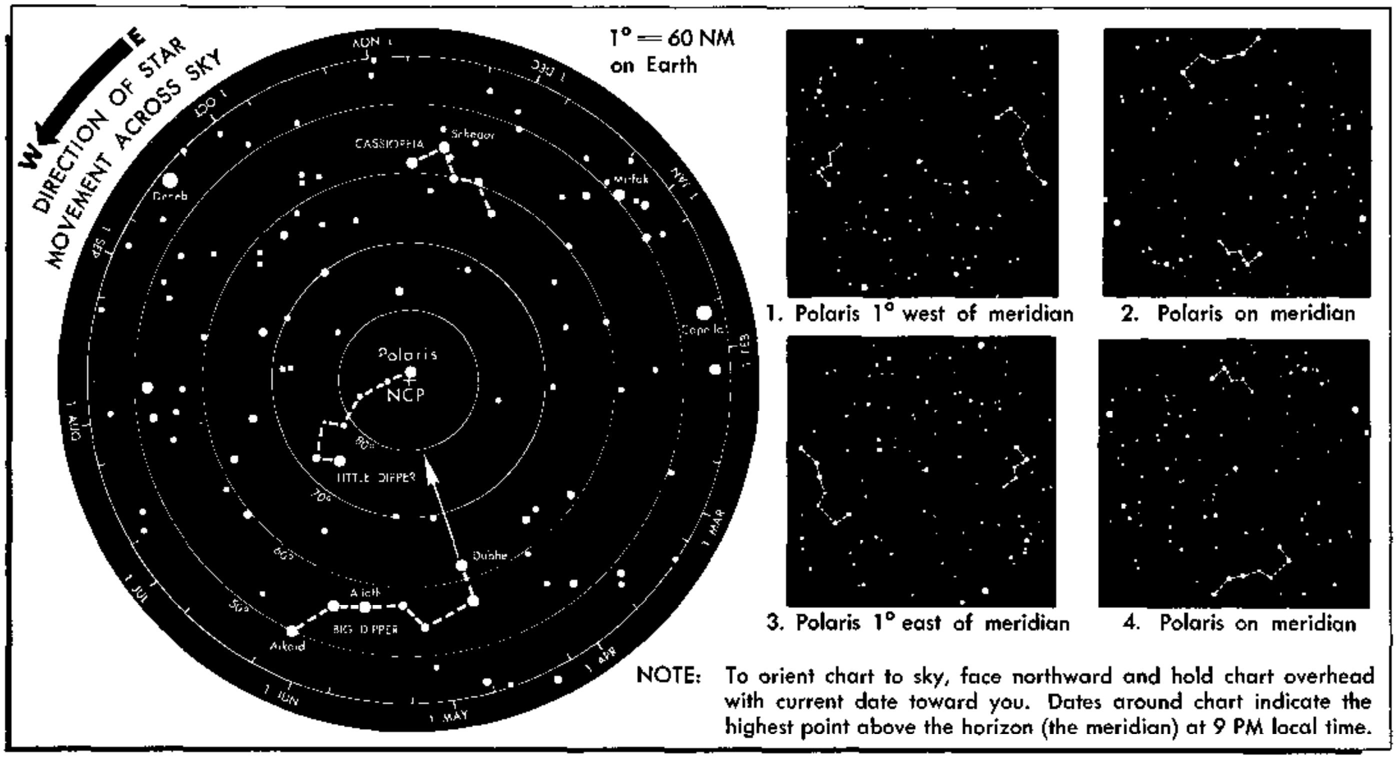

1. Polaris 1° west of meridian

2. Polaris on meridian

3. Polaris 1° east of meridian

4. Polaris on meridian

NOTE: To orient chart to sky, face northward and hold chart overhead with current date toward you. Dates around chart indicate the highest point above the horizon (the meridian) at 9 PM local time.

Figure 20-84. Finding Direction from Polaris.

survivor's watch is on eastern standard time, and it normally loses 30 seconds a day. It hasn't been set for 4 days. The local noon is timed at 15:08 on the watch on 4 February. Watch correction is 4 X 30 seconds, or plus 2 minutes. Zone time correction is plus 5 hours. Greenwich time is 15:08 plus 2 minutes plus 5 hours or 20:10. The equation of time for 4 February is minus 14 minutes. Local noon is 20:10 minus 14 minutes or 19:56 Greenwich. The difference in time between Greenwich and present position is 19:56 minus 12:00 or 7:56. A time of 7:56 equals 119°of longitude. Since local noon is later than Greenwich noon, the survivor is west of Greenwich, longitude is 119°W.

(e) Direction and Position Finding at Night:

-1. Direction from Polaris. In the Northern Hemisphere, one star, Polaris (the Pole Star), is never more than approximately 1° from the North Celestial Pole. In other words, the line from any observer in the Northern Hemisphere to the Pole Star is never more than 1° away from true north. Find the Pole Star by locating the Big Dipper or Cassiopeia, two groups of stars which are very close to the North Celestial Pole. The two stars on the outer edge of the Big Dipper are called pointers because they point almost directly to Polaris. If the pointers are obscured by clouds, Polaris can be identified by its relationship to the constellation Cassiopeia. Figure 20-84 indicates the relation between the Big Dipper, Polaris, and Cassiopeia.

-2. Direction from the Southern Cross. In the Southern Hemisphere, Polaris is not visible. There the Southern Cross is the most distinctive constellation. When flying south, the Southern Cross appears shortly before Polaris drops from sight astern. An imaginary line through the long axis of the Southern Cross, or True Cross, points toward the South Pole. The True Cross should not be confused with a larger cross nearby known as the False Cross, which is less bright and more widely spaced. Two of the four stars in the True Cross are among the brightest stars in the heavens; they are the stars on the southern and eastern arms. Those of the northern and western arms are not as conspicuous but are bright.

-3. There is no conspicuous star above the South Pole to correspond to Polaris above the North Pole. In fact, the point where such a star would be, if one existed, lies in a region devoid of stars. This point is so dark in comparison with the rest of the sky that it is known as the Coalsack. Figure 20-85 shows the True Cross and—to the west of it—the False Cross.

(f) Finding Due East and West by Equatorial Stars. Due to the altitude of Polaris above the horizon, it may sometimes be difficult to use as a bearing. To use a point directly on the horizon may be more convenient.

-1. The celestial equator, which is a projection of the Earth's equator onto the imaginary celestial sphere, always intersects the horizon line at the due east and west points of the compass. Therefore, any star on the celestial equator rises due east and sets due west (disallowing a small error because of atmospheric refraction). This holds true for all latitudes except those of the North and South Poles, where the celestial equator

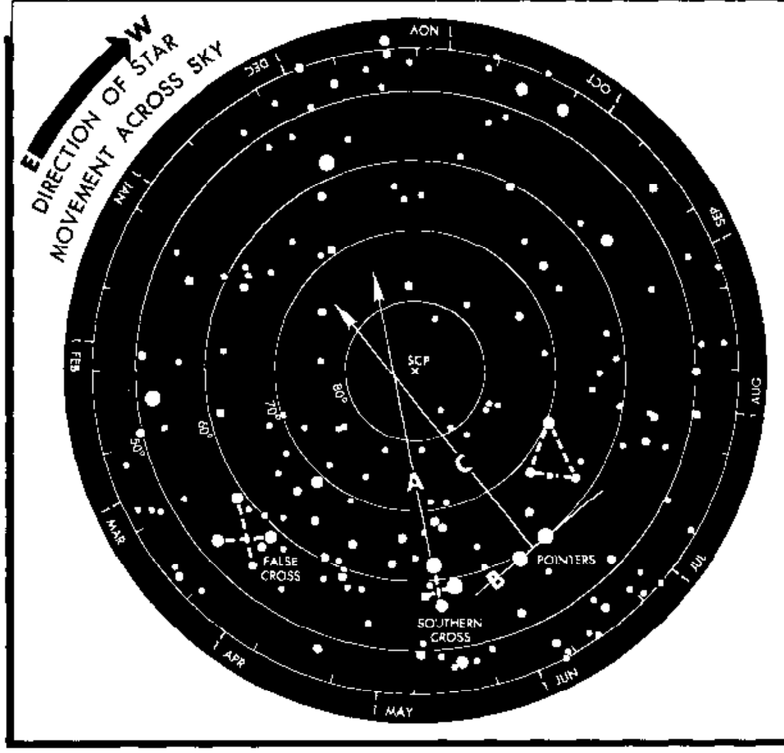

TO LOCATE THE SOUTH CELESTIAL POLE:

1. Extend an imaginary line (A) along the long axis of the True Cross to the south.
2. Join the two bright stars to the east of the Cross with an imaginary line (B). Bisect this line with one at right angles (C) and let it extend southward to intersect line (A).
3. The intersection of line (C) with the line through the Cross (A) is a few degrees from the South Celestial Pole (approximately 5 or 6 full moon widths).

NOTE: To orient chart to sky, face southward and hold chart overhead with current date toward you. Dates around chart indicate the highest point above the horizon (the meridian) at 9 PM local time.

Figure 20-85. Finding Direction from Southern Cross.

and the horizon have a common plane. However, if a survivor is at the North or South Pole, it will probably be known, so this technique can be assumed to be of universal use.

-2. Certain difficulties arise in the practical use of this technique. Unless a survivor is quite familiar with the constellations, it may be difficult to spot a specific rising star as it first appears above the eastern horizon. It will probably be simpler to depend upon the identification of an equatorial star before it sets in the west.

-3. Another problem is caused by atmospheric extinction. As stars near the horizon, they grow fainter in brightness because the line of sight between the observer's eyes and the star passes through a constantly thickening atmosphere. Therefore, faint stars disappear from view before they actually set. However, a fairly accurate estimate of the setting point of a star can be made some time before it actally sets. The atmospheric conditions of the area have a great effect on obstructing a star's light as it sets. Atmospheric haze, for example, is much less a problem on deserts than along temperate zone coastal strips.

-4. Figure 20-86 shows the brighter stars and some prominent star groups which lie along the celestial equator. There are few bright stars actually on the celestial equator. However, there are a number of stars which lie quite near it, so an approximation within a degree or so can be made. Also, a rough knowledge of the more conspicuous equatorial constellations will give the survivor a continuing checkpoint for maintaining orientation.

(g) Finding Latitude from Polaris. A survivor can find the latitude in the Northern Hemisphere north of 10°N by measuring the angular altitude of Polaris above the horizon, as shown in figure 20-87.

(h) Finding Direction (North) from Overhead Stars that are not in the General Location of the Celestial Poles:

-1. At times, survivors may not be able to locate Polaris (the North Star) due to partial cloud cover, or its position below the observer's horizon. In this situation, it would seem that survivors would be unable to locate direction. Fortunately, survivors who wish to initially find direction or who wish to check a course of

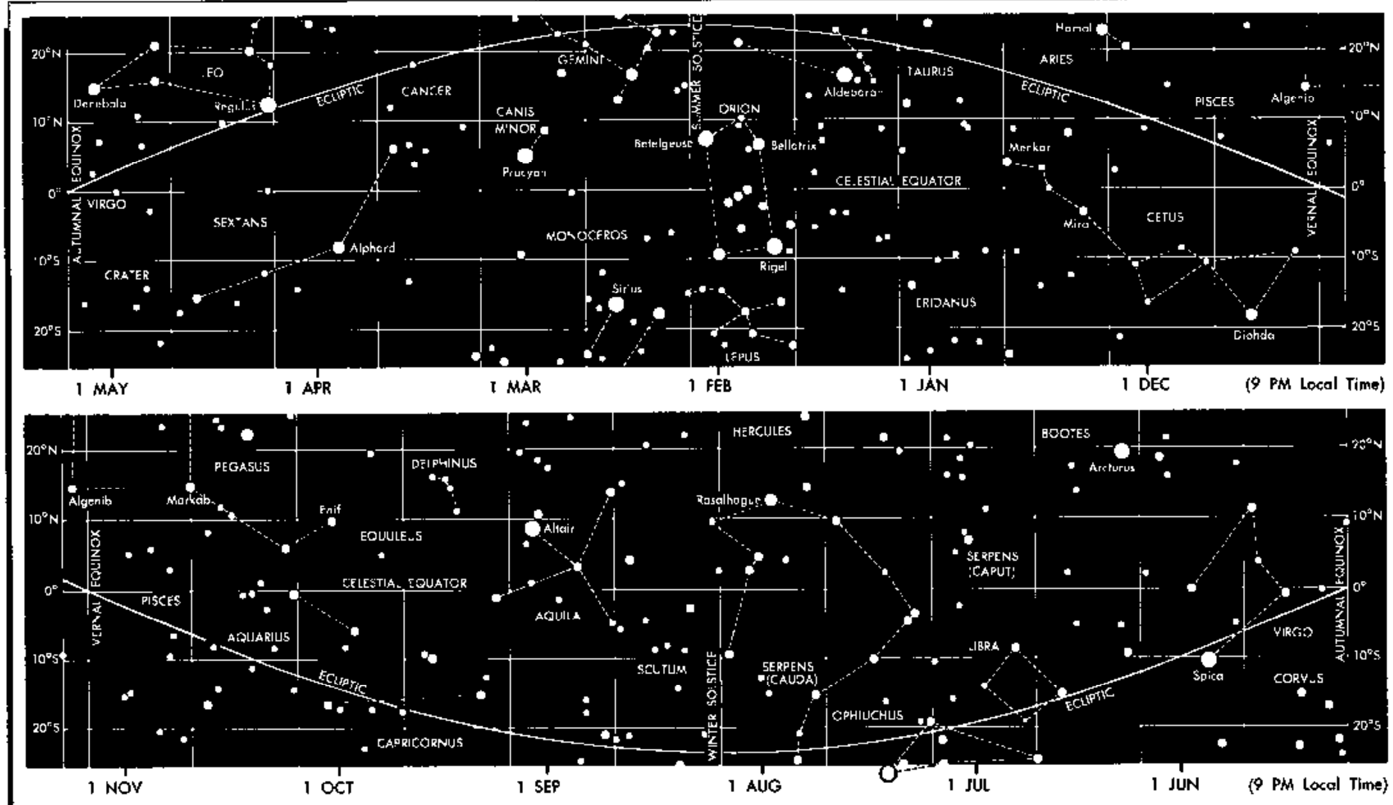
20°N
10°N
0°
10°S
20°S
LEO
Denebola
Regulus
ECLIPTIC
CANCER
GEMINI
CANIS MINOR
Procyon
AUTUMNAL EQUINOX
VIRGO
SEXTANS
Alphard
CRATER
MONOCEROS
Sirius
SUMMER SOLSTICE
ORION
Betelgeuse
Bellatrix
Rigel
LEPUS
Aldebaran
TAURUS
CELESTIAL EQUATOR
ERIDANUS
Hamal
ARIES
ECLIPTIC
PISCES
Algenib
Menkar
Mira
CETUS
Diphda
VERNAL EQUINOX
1 MAY
1 APR
1 MAR
1 FEB
1 JAN
1 DEC
(9 PM Local Time)
PEGASUS
Algenib
Markab
Enif
DELPHINUS
EQUULEUS
CELESTIAL EQUATOR
VERNAL EQUINOX
PISCES
AQUARIUS
ECLIPTIC
CAPRICORNUS
Altair
AQUILA
SCUTUM
WINTER SOLSTICE
HERCULES
Rasalhague
SERPENS (CAPUT)
SERPENS (CAUDA)
OPHIUCHUS
LIBRA
BOOTES
Arcturus
Spica
VIRGO
CORVUS
AUTUMNAL EQUINOX
1 NOV
1 OCT
1 SEP
1 AUG
1 JUL
1 JUN
(9 PM Local Time)

The star charts on this page show only those star-groups which lie near the celestial equator. Each map covers a strip approximately 50° wide (north and south) and 180° wide (east and west). Most of the more conspicuous naked eye stars are shown, the brighter stars being indicated by larger dots. Some star-groups have been joined by dashed lines to aid in identification.

Below each chart will be found the months of the year. On the first of each month the stars shown above the month name will be on the celestial meridian (highest point of their path above the horizon) at 9 p.m. local time.

The celestial equator makes a 45° angle with the southern horizon at latitude 45°N. At 60°N latitude it makes a 30° angle. In other words, degrees of latitude subtracted from 90° gives the angle formed by the celestial equator and the southern horizon.

The preceding information should enable the survivor to identify the equatorial stars to be used in determining due east or west. When observing stars rising in the east, the charts should be slanted down and to the left; when observing stars setting in the west, they should be slanted down and to the right.

Figure 20-86. Charts of Equatorial Stars.

HOW TO USE WEEMS PLOTTER TO MEASURE ALTITUDE OF POLARIS

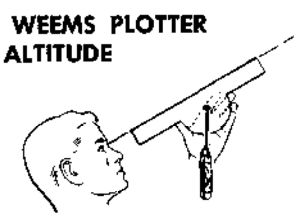

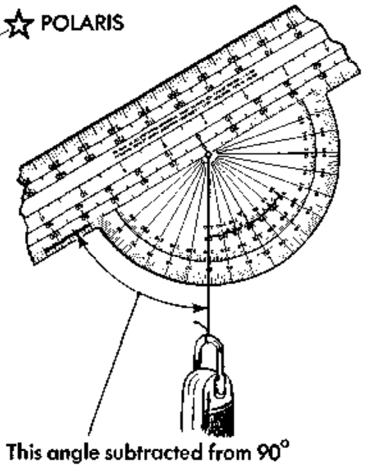

1. Attach string with weight to center hole of plotter.
2. After sighting, grasp string at outer edge of plotter and hold firmly against surface.
3. Holding string firmly against face of plotter, roll thumb under the edge of plotter and take reading.

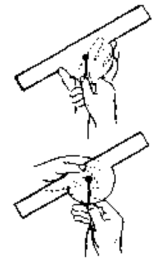

This angle subtracted from 90° is the altitude of Polaris.

HOW TO USE ANY PROTRACTOR TO MEASURE ALTITUDE OF POLARIS

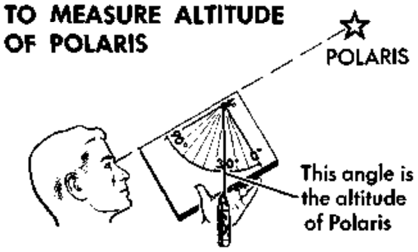

This angle is the altitude of Polaris

Any protractor (such as the one printed in AFM 64-5) can be used to find latitude.

1. Attach string with weight to Point "P".
2. After sighting, grasp string at edge of protractor.
3. Hold string in this position and take reading. This gives the altitude of Polaris.

CORRECTION FOR OBSERVED ALTITUDE OF POLARIS

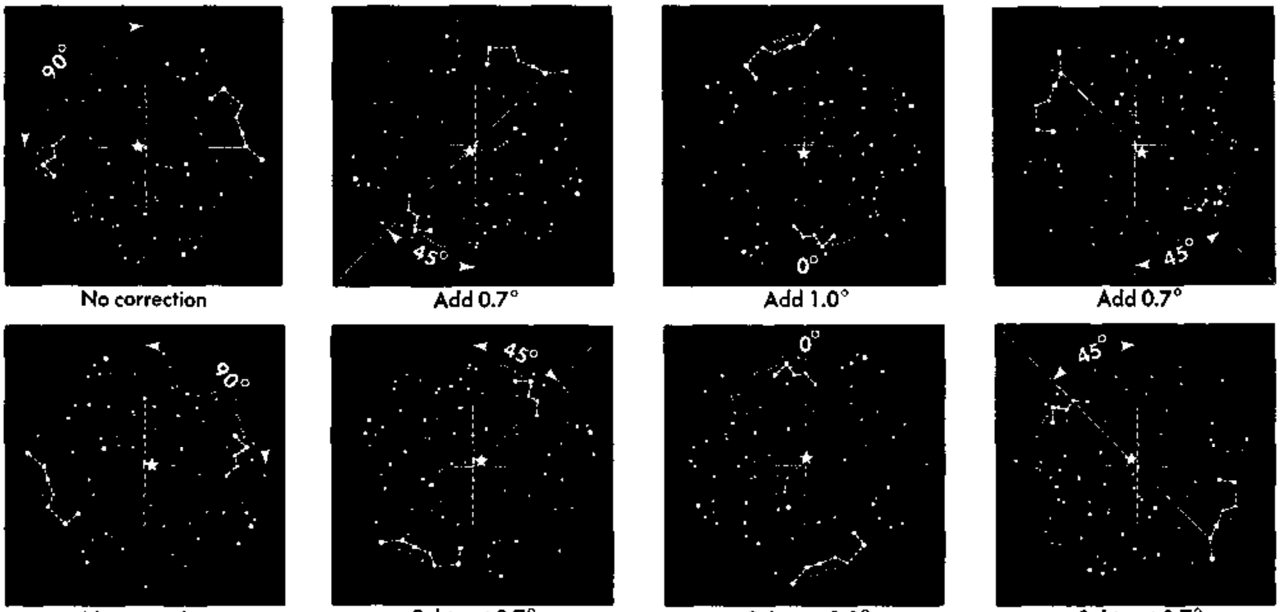

No correction — Add 0.7° — Add 1.0° — Add 0.7°

No correction — Subtract 0.7° — Subtract 1.0° — Subtract 0.7°

The star diagrams above are drawn for angles of 0°, 45°, and 90° between the vertical dotted line thru the pole and the line thru Cassiopeia and the Big Dipper (Ursa Major). For intermediate positions the angle may be estimated and the correction taken from the Correction Table. Subtract corrections given in the table when Polaris is above the horizontal line thru the pole and add corrections when Polaris is below this line.

Note that the correction changes very slowly near the time when the correction is greatest and hence an error in estimation of the position has little effect at this time.

CORRECTION TABLE

ANGLE	CORRECTION	ANGLE	CORRECTION
0°	1.0°	50°	0.6°
10°	1.0°	60°	0.5°
20°	0.9°	70°	0.3°
30°	0.9°	80°	0.2°
40°	0.8°	90°	No correction

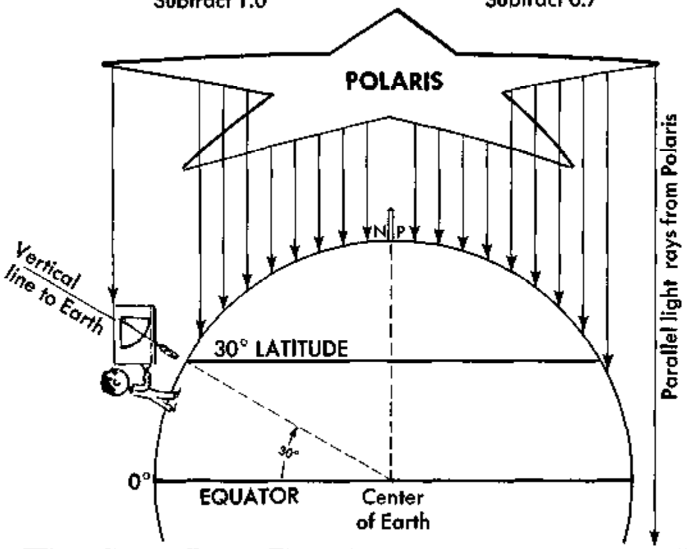

Figure 20-87. Finding Latitude by Polaris.

travel during the night need not worry about being lost or unable to travel if Polaris cannot be identified.

-2. The following is an adaptation of the stick and shadow method of direction finding. This method is based on the principles that all the heavenly bodies (Sun, Moon, planets, and stars) rise (generally) in the east and set (generally) in the west. This technique can be used anywhere on Earth with any stars except those which are circumpolar. Circumpolar stars are those which appear to travel around Polaris instead of apparently "moving" from east to west.

-3. To use this technique, survivors should keep in mind that they may use *any star other than a circumpolar one.*

-4. Survivors who wish to know general direction must prepare a device to aid them. This can be done by placing a stick (about 5 feet in length) at a slight angle in the ground in an open area (figure 20-88). Thin material (suspension line, string, vine, braided cloth, etc.) is then attached to the tip of the stick. This material should be longer than what is required to reach the ground (figure 20-88).

-5. The survivor should lie on the back with the head next to this hanging line and pull the cord up to the temple area and hold it tautly.

-6. Next, the survivor moves around on the ground until the taut line is pointing directly at the selected, bright, noncircumpolar star (or planet).

-7. The taut line is now in position to simulate the star's (or planet's) shadow. Survivors should remember that this method of finding direction is an adaptation of the Sun, stick, and shadow approach. Here the more distant stars and (or) planets take the place of our Sun. Since these objects are too distant from the Earth to create shadow, the string represents the shadow.

-8. With the taut line simulating the star's shadow, survivors should mark the point on the ground where the line touches with a stick, stone, etc. The survivor should repeat this sighting on the same star (or planet) after about 15 to 20 minutes (marking the spot at which the line "shadow" touches the ground). A line scribed on the ground which connects these two points will run west-east (as the stars and planets move from east to west, the "shadow" will move in the opposite direction). The first mark will be in the west. Drawing a line perpendicular to the west-east line, a survivor will have a north-south line and be able to travel.

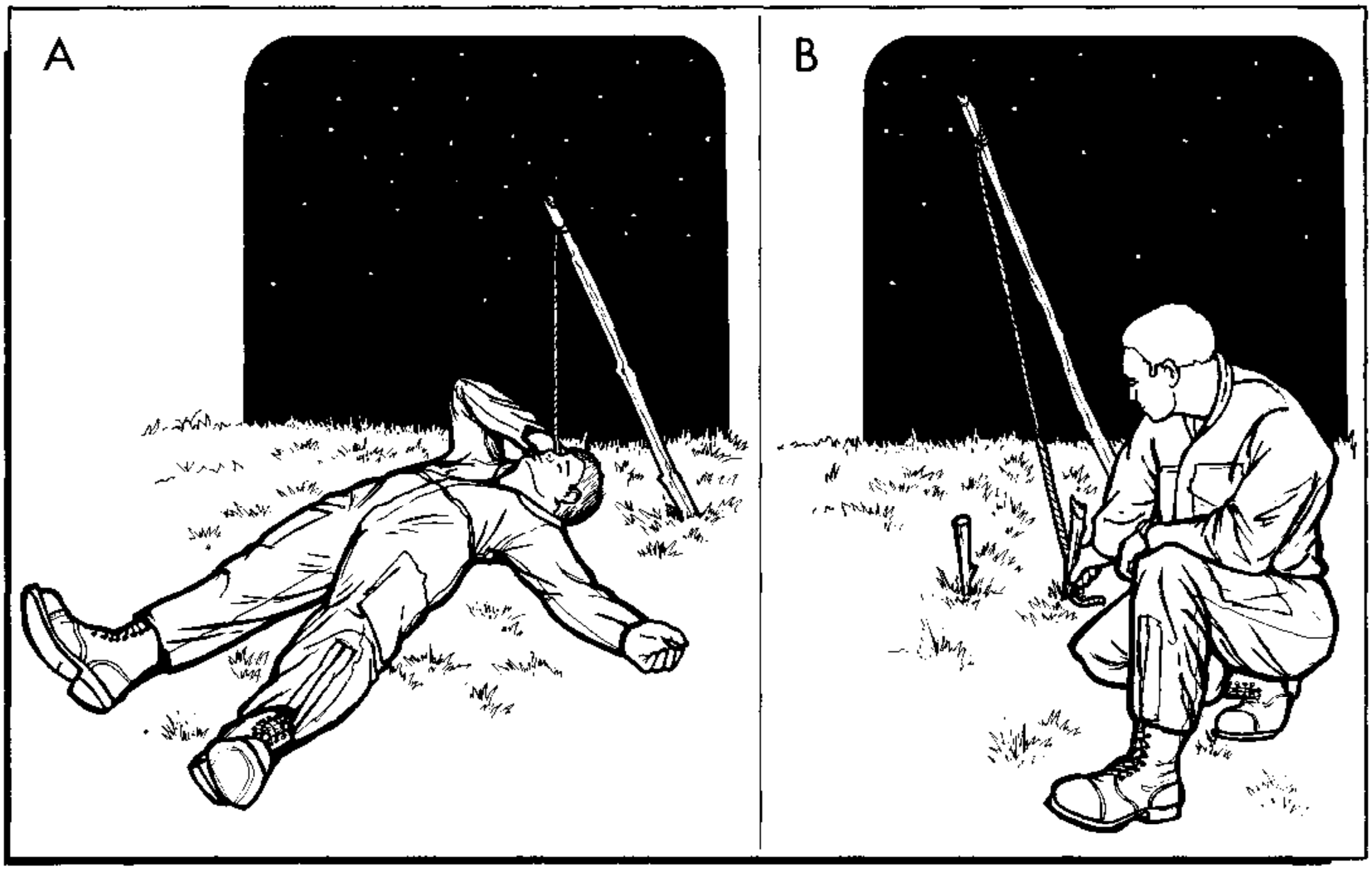

Figure 20-88. Stick and String Direction Finding.

Chapter 21

LAND TRAVEL

21-1. Introduction:

a. In any survival situation following an aircraft emergency, a decision must be made to either move or remain as close as possible to the parachute or crash site. In this chapter, land travel will be discussed and the various considerations that survivors should address before determining if travel is or is not a necessity.

b. Survivors may need to carry supplies and equipment while traveling to sustain life. For this reason, the techniques of backpacking and improvised packing are discussed to help a person do this task.

c. As a survivor, the ability to walk effectively is important in conserving energy and safety. Additionally, in rough terrain, travel may need to be done with the aid of a rope. The techniques of ascending and descending steep terrain are fundamental to understanding and performing rescue from rough terrain. These techniques, as well as techniques for snow travel, are covered. Travel may not be easy, but a knowledgeable traveler can travel safely and effectively while saving time and energy.

21-2. Decision to Stay or Travel. In hostile areas, the decision to travel is normally automatic. To stay in the vicinity of the crash or parachute landing can lead to capture. In friendly areas, a choice exists. The best advice is to stay with the aircraft. Most rescues have been made when downed aircrews remained with the aircraft.

a. Survivors should only leave the area when they are certain of their location and know that water, shelter, food, and help can be reached, or after having waited several days, they are convinced that rescue is not coming and they are equipped to travel.

b. Before making any decision, survivors should consider their personal physical condition and the condition of others in the party when estimating their ability to sustain travel. If people are injured, they should try to get help. If travel for help is required, they should send the people who are in the best physical and mental condition. Send two people if possible. To travel alone is dangerous. Before any decision is made, survivors should consider all of the facts.

(1) If the decision is to stay, these problems should be considered:

(a) Environmental conditions.

(b) Health and body care; camp sanitation.

(c) Rest and shelter.

(d) Water supplies.

(e) Food.

(2) If the decision is to travel: In addition to the primary survival problems of providing food, water, and shelter, the following must be considered:

(a) Direction of travel and why.

(b) Travel plan.

(c) Equipment required.

c. Before departing the site, survivors should leave information at their aircraft (nontactical situation only) stating departure time, destination, route of travel, personal condition, and available supplies.

d. From the air, it is easier to spot the aircraft than it is to spot people traveling on the ground. Someone may have seen the aircraft crash and investigate. The aircraft or parts from it can provide shelter, signaling aids, and other equipment (cowling for reflecting signals, tubing for shelter framework, gasoline and oil for fires, etc.). Avoiding the hazards and difficulties of travel is another reason to stay with the aircraft. Rescue chances are good if survivors made radio contact, landing was made on course or near a traveled air route, and weather and air observation conditions are good.

e. Present location must be known to decide intelligently whether to wait for rescue or to determine a destination and route of travel. The survivors should try to locate their position by studying maps, landmarks, and flight data, or by taking celestial observations. Downed personnel should try to determine the nearest rescue point, the distance to it, the possible difficulties and hazards of travel, and the probable facilities and supplies en route and at the destination.

f. There are a number of other factors that should be considered when deciding to travel.

(1) The equipment and materials required for cross-country travel should be analyzed. Travel is extremely risky unless the necessities of survival are available to provide support during travel. Survivors should have sufficient water to reach the next probable water source indicated on a map or chart and enough food to last until they can procure additional food. To leave shelter to travel in adverse weather conditions is foolhardy unless in an escape and evasion situation.

(2) In addition to the basic requirements, the physical condition of the survivor must be considered in any decision to travel. If in good condition, the survivor should be able to move an appreciable distance, but if the survivor is not in good condition or is injured, the ability to travel extended distances may be reduced. Analyze all injuries received during the emergency. For example, if a leg or ankle injury occurred during landing, this must be considered before traveling.

(3) If possible, survivors should avoid making any decision immediately after the emergency. They should wait a period of time to allow for recovery from the mental—if not the physical—shock resulting from the emergency. When shock has subsided survivors can then evaluate the situation, analyze the factors involved, and make valid decisions.

21-3. Travel. Once the survivors decide to travel, there are several considerations that apply regardless of the circumstances.

a. The ranking person must assume leadership, and the party must work as a team to ensure that all tasks are done in an equitable manner. Full use should be made of any survival experience or knowledge possessed by members of the group, and the leader is responsible for ensuring that the talents of all survivors are used.

b. Survivors should keep the body's energy output at a steady rate to reduce the effects of unaccustomed physical demands.

(1) A realistic pace should be maintained to save energy. It increases durability and keeps body temperature stable because it reduces the practice of quick starts and lengthy rests. More importantly, a moderate, realistic pace is essential in high altitudes in avoiding the risks of lapse of judgment and hallucinations due to lack of oxygen (hypoxia). Travel speed should provide for each survivor's physical condition and daily needs, and the group pace should be governed by the pace of the slowest group member. Additionally, rhythmic breathing should be practiced to prevent headache, nausea, lack of appetite, and irritability.

(2) Rest stops should be short since it requires added energy to begin again after cooling off. Survivors should wear their clothing in layers (layer system) and make adjustments to provide for climate, temperature, and precipitation. It is better to start with extra clothing and stop and shed a layer when beginning to warm up.

(3) Wearing loose clothing provides for air circulation, allows body moisture to evaporate, and retains body heat. Loose clothing also allows freedom of movement.

(4) Travelers should keep in mind when planning travel time and distance that the larger the group, the slower the progress will be. Time must be added for those survivors who must acclimate themselves to the climate, altitudes, and the task of backpacking. Survivors should also allow time for unexpected obstacles and problems which could occur.

(5) Proper nutrition and water are essential to building and preserving energy and strength. Several small meals a day are preferred to a couple of large ones so that calories and fluids are constantly available to keep the body and mind in the best possible condition. Survivors should try to have water and a snack available while trekking, and they should eat and drink often to restore energy and prevent chills in cold temperatures. This also applies at night.

21-4. Land Travel Techniques. Land travel techniques are based largely on experience, which is acquired through performance. However, experience can be partially replaced by the intelligent application of specialized practices that can be learned through instruction and observation. For example, travel routes may be established by observing the direction of a bird's flight, the actions of wild animals, the way a tree grows, or even the shape of a snowdrift. Bearings read from a compass, the Sun, or stars will improve on these observations and confirm original headings. All observations are influenced by the location and physical characteristics of the area where they are made and by the season of the year.

a. Route Finding. The novice should follow a compass line, whereas the experienced person follows lines of least resistance by realizing that a curved route may be faster and easier under certain circumstances. Use game trails when they follow a projected course only. For example, trails made by migrating caribou are frequently extensive and useful. On scree or rockslides, mountain sheep trails may be helpful. Game trails offer varying prospects, such as the chance of securing game or locating waterholes. Successful land travel requires knowledge beyond mere travel techniques. Survivors should have at least a general idea of the location of their starting place and their ultimate destination. They should also have knowledge of the people and terrain through which they will travel. If the population is hostile, they must adapt their entire method of travel and mode of living to this condition.

b. Wilderness. Wilderness travel requires constant awareness. A novice views a landscape from the top of a hill with care and interest, and says, "let's go." The experienced person carefully surveys the surrounding countryside. A distant blur may be mist or smoke; a faint, winding line on a far-off hill may be manmade or an animal trail; a blur in the lowlands may be a herd of caribou or cattle. People should plan travel only after carefully surveying the terrain. Study distant landmarks for characteristics that can be recognized from other locations or angles. Careful and intelligent observation will help survivors to correctly interpret the things they see, distant landmarks, or a broken twig at their feet. Before leaving a place, travelers should study their backtrail carefully. Survivors should know the route forward and backward. An error in route planning may make it necessary to backtrack in order to take a new course. For this reason, all trails should be marked (figure 21-1).

c. Mountain Ranges. Mountain ranges frequently affect the climate of a region and the climate in turn influences the vegetation, wildlife, and the character and number of people living in the region. For example, the oceanside of mountains has more fog, rain, and snow than the inland side of a range. Forests may grow on the oceanside, while inland, it may be semi-dry. Therefore, a complete change of survival techniques may be necessary when crossing a mountain range.

(1) Travel in mountainous country is simplified by conspicuous drainage landmarks, but it is complicated by the roughness of the terrain. A mountain traveler can

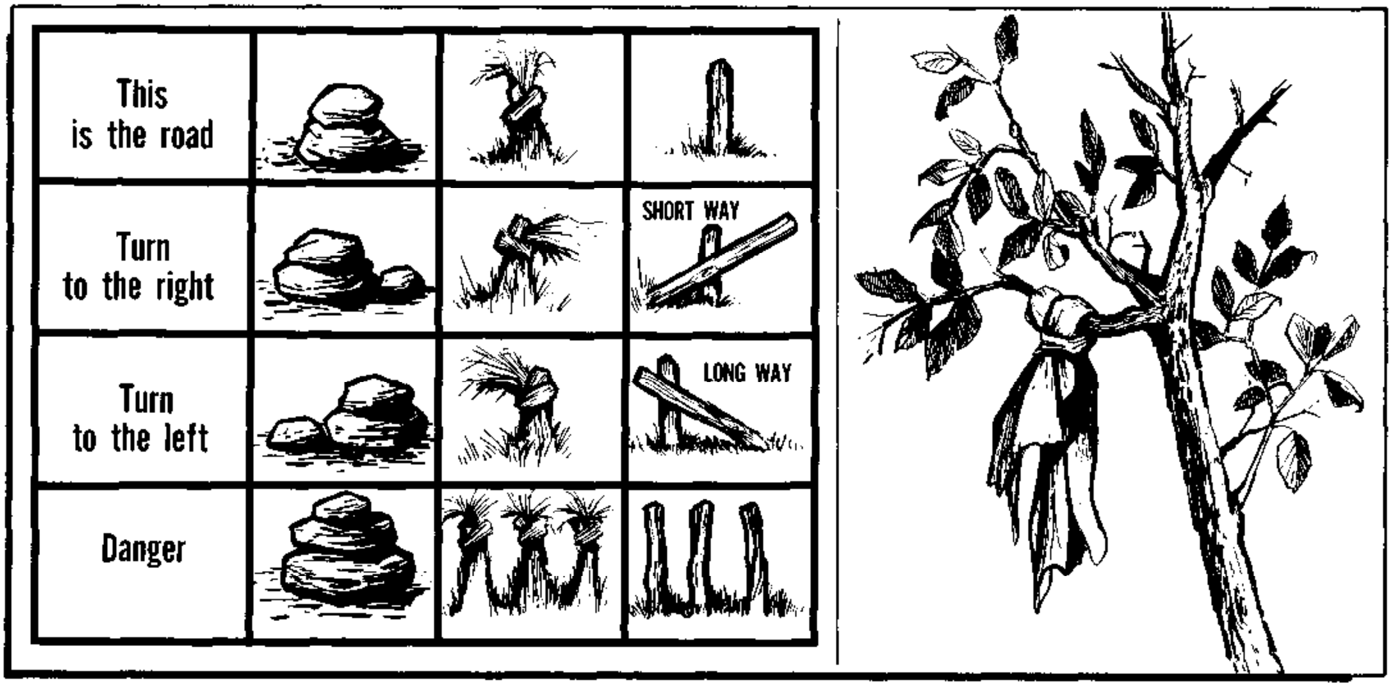

Figure 21-1. Marking a Trail.

readily determine the direction in which rivers or streams flow; however, surveying is necessary to determine if a river is safe for rafting, or if a snowfield or mountainside can be traversed safely. Mountain travel differs from travel through rolling or level country, and certain cardinal rules govern climbing methods. A group descending into a valley, where descent becomes increasingly steep and walls progressively more perpendicular, may be obliged to climb up again in order to follow a ridge until an easier descent is possible. In such a situation, rappeling with a parachute line rope may save many weary miles of travel. In mountains, travelers must avoid possible avalanches of earth, rock, and snow, as well as crevasses (deep cracks in the ice) in ice fields.

(2) In mountainous country, it may be better to travel on ridges—the snow surface is probably firmer and there is a better view of the route from above. Survivors should watch for snow and ice overhanging steep slopes. Avalanches are a hazard on steep snow-covered slopes, especially on warm days and after heavy snowfalls.

(3) Snow avalanches occur most commonly and freqently in mountainous country during wintertime, but they also occur with the warm temperatures and rainfalls of springtime. Both small and large avalanches are a serious threat to survivors traveling during winter as they have tremendous force. The natural phenomena of snow avalances is complex. It is difficult to definitely predict impending avalanches, but knowing general behaviors of avalanches and how to identify them can help people avoid avalanche hazard areas.

(a) Snow or Sluff Avalanche. The loose snow or sluff avalanche is one kind of avalanche that starts over a small area or in one specific spot. It begins small and builds up in size as it descends. As the quantity of snow increases, the avalanche moves downward as a shapeless mass with little cohesion.

(b) Terrain Factors Affecting Avalanches:

-1. Steepness. Most commonly, avalanches occur on slopes ranging from 30 to 45 degrees (60 to 100 percent grade), but large avalanches do occur on slopes ranging from 25 to 60 degrees (40 to 173 percent grade). (See figure 21-2.)

-2. Profile. The dangerous slab avalanches have more chance of occurring on convex slopes because of the angle and the gravitational pull. Concave slopes cause a danger from slides that originate at the upper, steep part of the slope (figure 21-3).

-3. Slopes. Midwinter snowslides usually occur on north-facing slopes. This is because the north slopes do not receive the required sunlight which would heat and stabilize the snowpack. South-facing slope slides occur on sunny, spring days when sufficient warmth melts the snow crystals causing them to change into wet, watery, slippery nowslidessnow. Leeward slopes are dangerous because the wind blows the snow into well packed drifts just below the crest. If the drifts have not adhered to the snow underneath, a slab avalanche can occur. Windward slopes generally have less snow and are compact. It is usually strong enough to resist move-

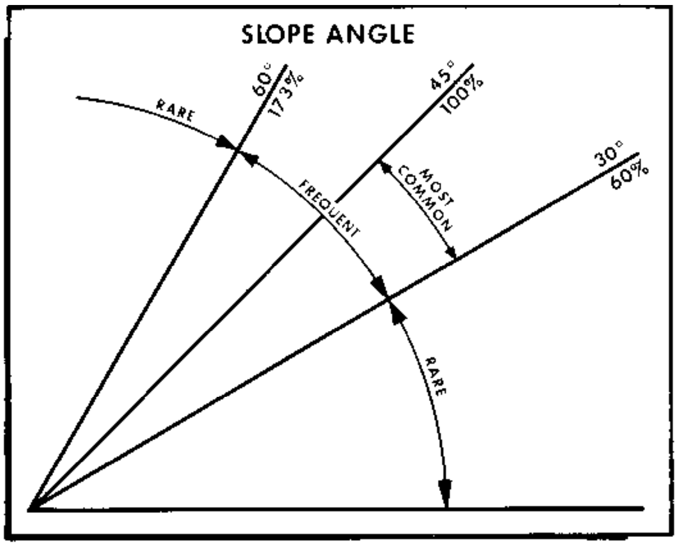

Figure 21-2. Slope Angle.

ment, but avalanches may still occur with warm temperature and moisture.

-4. Surface Features. Most avalanches are common on smooth, grassy slopes that offer no resistance. Brush, trees, and large rocks bind and anchor the snow, but avalanches can still occur in tree areas (figure 21-4).

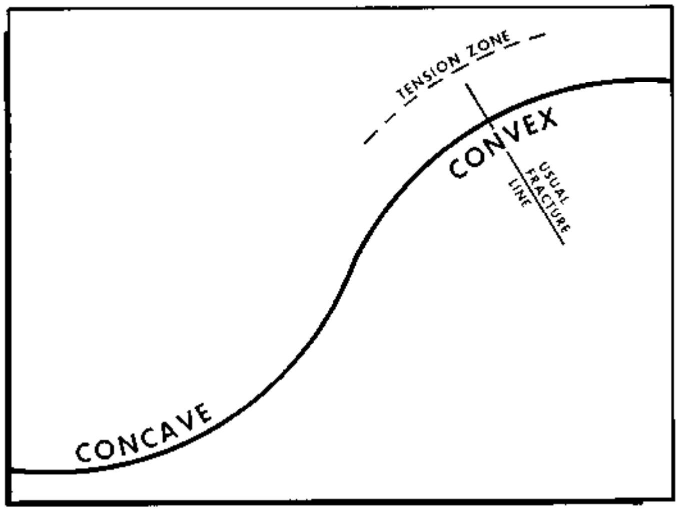

Figure 21-3. Profile of Slope.

(c) Weather Factors:

-1. Old snow depth covers up natural anchors (rocks, brush, and fallen trees) so that the new snowfall slides easier. The type of old snow surface is important. Sun crests or smooth surface snows are unstable; whereas a rough, jagged surface would offer stability and an anchorage. A loose snow layer underneath is far more hazardous than a compacted one as the upper layer of snow will slide more easily with no rough texture to restrain it. Travelers should check the underlying snow by using a rod or stick.

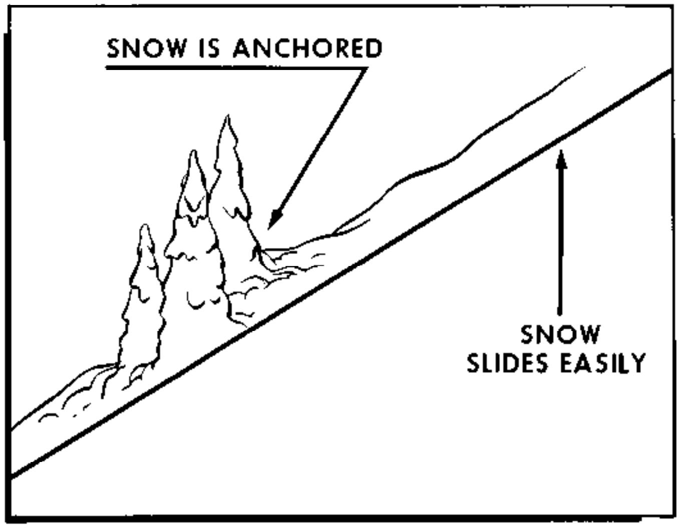

Figure 21-4. Snowslides.

-2. Winds, 15 miles per hour or more, cause the danger of avalanches to develop rapidly. Leeward slopes will collect snow that has been blown from the windward sides, forming slabs or sluffs, depending upon the temperature and moisture. Snow plumes or cornices indicate this condition (figure 21-5).

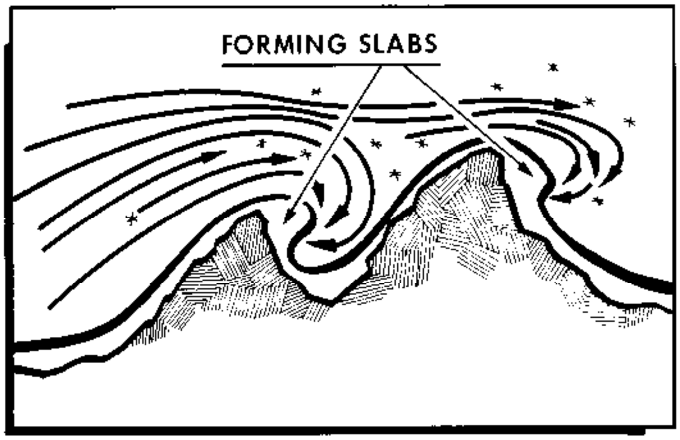

Figure 21-5. Forming Slides.

-3. A high percentage of all avalanches occur during or shortly after storms. Layers of different types of snow from different storms will cause unstable snow because the bond between layers will vary in strength. The rate of snowfall also has a significant effect on stability. A heavy snowfall spread out over several days is not as dangerous as a heavy snowfall in a few hours because slow accumulation allows time for the snow to settle and stabilize. A large amount of snow over a short period of time results in the snow constantly changing and building up, giving it little time to settle and stabilize. If the snow is light and dry, little settling or cohesion occurs, resulting in instability.

-4. Under extremely cold temperatures, snow is unstable. In temperatures around freezing or just above, the snow tends to settle and stabilize quickly. Storms, starting at low temperature with light, dry snow which are followed by rising temperature, cause the top layer of snow to be moist and heavy, providing opportune conditions for avalanching. The light, dry snow underneath lacks the strength and elastic bondage necessary to hold the heavier, wetter snow deposited on top; therefore, the upper layer slides off. Also, extreme temperature differences between night and day cause the same problems. Rapid changes in weather conditions cause adjustments and movement within the snowpack. Survivors should be alert to rapid changes in winds, temperatures, and snowfall which may affect snow stability.

-5. Avalanches of wet snow are more likely to occur on south slopes. Sun, rainstorms, or warmer temperatures brought by spring weather are absorbed by the snow causing it to become less stable.

(d) Warning Signs. Avalanches generally occur in the same area. After a path has been smoothed, it's easier for another avalanche to occur. Steep, open gullies and slopes, pushed over small trees, limbs broken off, and tumbled rocks, are signs of slide paths. Snowballs tumbling downhill or sliding snow is an indication of an avalanche area on leeward slopes. If the snow echoes or sounds hollow, conditions are dangerous. If the snow cracks and the cracks persist or run, the danger of a slab avalanche is imminent. The deeper the snow, the more the terrain features will be obscured. Knowledge of common terrain features can help survivors visualize what they may be up against, what to avoid, and the safest areas to travel. Knowing the general weather pattern for the area is helpful. Survivors should try to determine what kind of weather will be coming by observing and knowing the signs that indicate certain weather conditions.

(e) Route Selection. If avoiding the mountains and avalanche danger areas is impossible, there are precautions survivors should take when confronting dangerous slopes. They should decide which slopes will be the safest by analyzing the factors that determine what makes one slope safe and another deadly. Study the slope terrain and keep in mind why avalanches occur.

-1. When survivors decide to cross a slope, one person at a time should cross. If all go together, they should not tie together since there is no way one person can hold another against an avalanche. Instead, they should tie a contrasting color line about 100 feet long (using suspension line or PLD tape) to each person. If they should get caught in an avalanche, the line will help identify their position if it is exposed. Survivors should select escape routes before and throughout the climb and keep these routes in mind at all times. They should also stay to the fall line (natural path an object will follow when falling or sliding down the slope) when climbing and not zig-zag or climb a different route because it seems easier. Staying to the fall line will prevent making fractures and at the same time, compact the snow, making it more stable for others who follow. If traversing, they should travel above the danger area. Survivors should travel quickly and quietly to avoid extended exposure to the probable danger of avalanches.

-2. If caught in an avalanche, any equipment that weighs survivors down should be dropped. The pack, snowshoes, and any other articles should be jettisoned.

-3. The standard rule is to use swimming motions to try to move towards the snow surface. Further, survivors should go for the sides and not try to out swim the avalanche. If near the surface, they should try to keep one arm or hand above the surface to mark their position. If buried, a person should inhale deeply (nose down) before the snow stops moving to make room for their chest. Trapped persons should try to make breathing space around their faces. They shouldn't struggle, but should relax and conserve their energy and oxygen. Only when fellow survivors or rescuers are nearby should the trapped individual shout. Rescue should be done as quickly as possible. Avalanche victims generally have a 50 percent chance of surviving after 30 minutes have passed because the snow will set up (harden).

-4. Glaciers may offer emergency travel routes across mountain ranges. Glacier crossing demands special knowledge, techniques, and equipment such as the use of a lifeline and poles for locating crevasses. There are many places where mountain ranges can be negotiated on foot in a single day by following glaciers. Survivors must be especially careful on glaciers and watch for crevasses covered by snow. If traveling in groups of three persons or more, they should be roped together at intervals of 30 to 40 feet. Every step should be probed with a pole. Snow-bridged crevasses should be crossed at right angles to their course. The strongest part of the bridge can be found by probing. When crossing a bridged crevasse, weight should be distributed over a large area by crawling or by wearing snowshoes.

-5. When forested areas are dense, river travel and ridges usually afford the easiest travel routes. In open forests, land travel is easier and offers a better selection of travel routes; however, such forests may not offer sufficient cover or concealment for escape and evasion travel.

-6. After a fire, windstorm, or logging operation, second-growth timber usually grows thick. It is worse after it grows about 20 feet high since any space between the trees is filled in by branches and the overhead timber isn't (yet) thick enough to cause the lower branches to die from lack of sunlight.

-a. Deciduous brush also contributes to the overgrowth. Blowdowns, avalanche fans, and logging slush are difficult to negotiate. Such obstructions, even a few hundred feet may require major changes to the original travel route.

-b. Scrub cedar (subalpine fir) is hard to penetrate. There are tactics that can be used to make travel easier. Survivors can use fallen trees as walkways to provide a short route of travel through the scrub cedar to a clear area. Gloves should be worn when penetrating thorny vegetation. Overlaying bushes can be separated to allow passage. When land is steep, brush can be used to provide handholds if it is strong and anchored well.

-c. Brush can be dangerous. Survivors should be aware of the possibility of slipping while going downhill. Therefore, they should ensure each step is firmly placed. Survivors should be aware of travel difficulties presented by cliffs, boulders, and ravines which are covered by brush.

-7. Do not travel through dense brush if it can be avoided.

-a. Travel on trails rather than taking shortcuts through the brush. Brush is frequently easier to travel through (over) during the winter season when it is covered by snow and when snowshoes are available (improvised).

-b. During the summer, avoid avalanche tracks because the debris may be difficult to penetrate. Traveling on the "timber cones" between avalanche paths is best when climbing a valley.

-c. The heaviest timber is the best area to travel because little or no brush will be growing on the forest floor.

-d. Try to avoid areas near creeks and valley floors because they have more brush and trees than the valley walls and ridges. However, traveling in the stream channels may be preferable when the area is covered with dense brush and vegetation. Survivors may have to wade, but the stream may offer the best route through the brush.

-e. Traveling high above the brush at the timberline may be worthwhile if the bottom and sides of the valley look futile.

d. Snow and Ice Areas. Travel in snow and ice areas is not recommended except to move from an unsafe to a safe area, or from an area that has few natural resources to an area of greater resources (shelter material, food, and signaling area).

(1) Before traveling to a possible rescue site, town, village, or cabin, travelers should know their approximate position and the location of the desired site. The greatest hazard in snow and ice areas is the intense cold and high winds. Judging distance is difficult due to the lack of landmarks and the clear arctic air. Image distortion is a common phenomena. "White-out" conditions exist and the survivor should not travel during this time. A white-out condition occurs when there is complete snow cover and the clouds are so thick and uniform that light reflected by the snow is about the same intensity as that from the sky. If traveling during bad weather, great care must be taken to avoid becoming disoriented or falling into crevasses, over cliffs or high snow ridges, or walking into open leads. A walking stick is very useful to probe the area in the line of travel.

(2) Strong winds often sweep unchecked across tundra areas (due to the lack of vegetation) causing white-out conditions. Because of blowing snow, fog, and lack of landmarks, a compass is a must for travel, yet it is still difficult to navigate a true course since the magnetic variation in the high latitudes (polar areas) is often extreme.

(3) During the summer months, the area is a mass of bogs, swamps, and standing water. Crossing these areas will be difficult at best. Rain and fog are common. Insects such as mosquitoes, midges, and black flies can and will cause the survivor physical discomfort and may cause travel problems. If the body is not completely covered with clothing, or if survivors do not use a head net or insect repellant, insect bites may be severe and infection can set in.

(4) In mountainous country, it is often best to travel along ridge lines because it provides a firmer walking surface and there is usually less vegetation to contend with. High winds make travel impractical if not impossible at times. Glaciers have many hidden dangers. Glacial streams may run just under the surface of the snow or ice, creating weak spots, or they may run on the surface and cause slick ice. Crevasses which run across the glacier can be a few feet to several hundred feet deep. Quite often crevasses are covered over with a thin layer of snow, making them practically invisible. Survivors could fall into crevasses and sustain severe injuries or death. If glacier travel is required, it is best to use a probe pole to test the footing ahead.

(5) Summer travel in timbered areas should not present any major problems; however, travel on ridges is preferred since the terrain is drier and there are usually fewer insects. During the cold months, snow may be deep and travel will be difficult without some type of snowshoes or skis. Travel is generally easier on frozen rivers, streams, and lakes since there is less snow or wind-packed snow and they are easier to walk on.

(6) River and stream travel can be hazardous. Rivers comparatively straight are that way because of the volume of water flow and extremely fast currents. These rivers tend to have very thin ice in the winter (cold climates), especially where snowbanks extend out over the water. If an object protrudes through the ice, the immediate area will be weak and should be avoided if possible. Where two rivers and streams come together, the current is swift and the ice will be weaker than the ice on the rest of the river. Very often after freezeup, the

source of the river or stream dries up so rapidly that air pockets are formed under the ice and can be dangerous if fallen into. During the runoff months (spring and summer), rivers and streams usually have a large volume of water which is very cold and can cause cold injuries. Wading across or down rivers and streams should be done with proper footwear and exposure protection due to the depth, swiftness, unsure footing, and coldness of the water. Generally, streams are too small and shallow for rafting. Streams are often bordered by high cliffs or banks at the headwaters. As a stream progresses, its banks are often choked with alder, devil's club, and other thick vegetation making traveling very slow and difficult. Many smaller streams will simply lead the traveler to a bog or swamp where they end, causing more problems for the survivor.

(7) Sea ice conditions vary greatly from place to place and season to season. During the winter months, there is generally little open water except between the edges of floes. Crossing from one floe to another can be done by jumping across the open-water area, but footing may be dangerous. When large floes are touching each other, the ice between is usually ground into brash ice by the action of the floes against each other and this ground-up ice will not support a survivor's weight. Pressure ridges are long ridges in sea ice caused by the horizontal pressure of two ice floes coming together. Pressure ridges may be 100 feet high and several miles long; they may occur in a gulf or bay, or on polar seas. They must be crossed with great care because of the ruggedness of ice formations, weak ice in the area, and possibility of open water covered with a thin layer of snow or ground-up ice. During summer months, the ice surface becomes very rough and covered with water. The ice also becomes soft and honeycombed (candlestick ice) even though the air temperature may be below freezing. Traveling over sea ice in the summer months is very dangerous.

(8) Icebergs are great masses of ice and are driven by currents and winds; about two-thirds of the iceberg is below the surface of the sea. Icebergs in open seas are always dangerous because the ice under the water will melt faster than the surface exposed to the air, upsetting the equilibrium and toppling them over. The resulting waves can throw small pieces of ice in all directions. Avoid pinnacle-shaped icebergs—low, flat-topped icebergs are safer.

e. Dry Climates. Before traveling in the desert, the decision to travel must be weighed against the environmental factors of terrain and climate, condition of survivors, possibility of rescue, and the amount of water and food required.

(1) The time of day for traveling is greatly dependent on two significant factors; the first and most apparent is temperature, and the second is type of terrain. For example, in rocky or mountainous deserts, the eroded drainages and canyons may not be seen at night and could result in a serious fall. Additionally, manmade features such as mining shafts or pits and irrigation channels could cause similar problems. If the temperature is not conducive to day travel, survivors should travel during the cooler parts of the day (in early morning or late evening). Traveling on moonlit nights is another possibility; however, survivors must be aware that moonlight can cast deceiving shadows. This problem can be decreased by scanning the ground to allow the night sensitive portions of the eye time to pick up the slight differences in lighting. In hot desert areas where these hazards do not occur, traveling at night is a very practical solution. During the winter in the midlatitude deserts, the cold temperatures make day travel most sensible.

(2) There are three types of deserts: mountain, rocky plateau, and sandy or dune deserts. These deserts can present difficult travel problems.

(a) Mountain deserts are characterized by scattered ranges or areas of barren hills or mountains, separated by dry, flat basins. High ground may rise gradually or abruptly from flat areas to a height of several thousand feet above sea level. Most desert rainfall occurs at high elevations and the rapid runoff causes flash floods, eroding deep gullies or ravines and depositing sand and gravel around the edges of the basins. These floods are a problem on high and low grounds. The flood waters rapidly evaporate, leaving the land as barren as before, except for plush vegetation which rapidly becomes dormant. Basins without shallow lakes will have alkaline flats which can cause problems with chemical burns and can destroy clothing and equipment.

(b) Rocky plateau deserts have relief interspersed by extensive flat areas solid or broken rock at or near the surface. They may be cut by dry, steep-walled, eroded valleys, known as wadis in the Middle East and arroyos or canyons in the United States and Mexico. The narrower of these valleys can be extremely dangerous to humans and material due to flash flooding. Travel in these valleys may present another problem: a survivor can lose site of reference points and travel farther than intended. The Golan Heights in Israel is an example of a rocky plateau desert.

(c) Sandy (dune) deserts are extensive flat areas covered with sand or gravel. They are the product of ancient and modern wind erosion. "Flat" is relative in this case, as some areas contain sand dunes that are over 1,000 feet high and 10 to 15 miles long. Travel in such terrain depends on windward or leeward gradients of the dunes and texture of sand. These dunes help a survivor determine general direction. Longitudinal dunes are continuous banks of sand at even heights that lie parallel with the dominant wind. Other areas, however, may be totally flat for distances in excess of 2 miles. Plant life varies from none to scrub reaching over 6 feet. Examples of this desert include the ergs of the Sahara Desert, Empty Quarter of the Arabian Desert, areas of California and New Mexico, and the Kalahari Desert in South Africa. A seif dune has forms similar to a drift behind a rock. Its length lies in the direction of the prevailing winds. Additionally, the horseshoe-shaped crescent dune has a hollow portion that faces downward. Ripples caused by wind in the sand may indicate the direction of the prevailing winds. These ripples generally lie perpendicular to the prevailing winds. In deserts, it is easier to travel on the windward side of the tops of dunes. Even though these ridges may not lay in a straight line and may wander, they offer a better route of travel than traveling in straight lines. A great deal of energy and time can be expended walking up and down dunes, especially in the loose sand on the leeward side of dunes.

(3) People who travel the desert at night orient themselves by the stars and Moon. Those who traveled by day use compasses, when available, and the Sun. Survivors should use all directional aids during emergency travel and each aid should be frequently cross-checked against each other. For desert travel, a compass is a valuable piece of equipment.

(a) Without a compass, landmarks must be used for local navigation. This can lead to difficulties. Mirages cause considerable trouble. Ground haze throughout the day may obscure vision. Distances are deceptive in the deserts and survivors have reported difficulty in estimating distances and the size of objects. In southern Egypt, one survivor reported large boulders always appeared smaller than they were and in other cases small obstacles appeared insurmountable. Survivors in Saudi Arabia and in Tunisia warned that it is difficult to maintain a single landmark in navigation. Several groups reported they found it necessary to take turns keeping an eye on a specific mountain, peak, or object which was their goal. Objects have a way of vanishing in some cases when the eye is moved for an instant, and in other cases, many peaks or hills looked alike and caused difficulties in determining the original object. In Tunisia, twin peaks are not reliable landmarks because of their frequent occurrence. (Survivors have found after a short time of traveling they may have up to a dozen twin peaks for reference in the same vicinity.) The Great Sand Sea (Egypt and Libya) was the emergency landing site of several groups of survivors and caused navigational difficulties. In these rolling sand hills, it is impossible to keep one object in view, and even footmarks fail to provide a reliable backtrail for determining travel directions. The extreme flatness of other stretches of desert terrain in North Africa also makes navigation difficult. With no landmarks to follow, no objectives to sight, survivors may walk in circles or large arcs before realizing their difficulties.

(b) A Marine pilot who made an emergency landing in the Arizona desert took the precaution of immediately spreading his parachute on the ground and putting rocks on the edges to ensure maximum visibility from the air. Then he decided to walk to his crashed plane, a distance he estimated to be 500 yards from his landing spot. He reached the plane and found it gutted by fire, and spent 5 days wandering the flat desert trying to find his parachute.

(c) Navigational difficulties of a different type may be experienced in Ethiopia, Kenya, and Somalia. Here the density of the thorn brush, even though it was primarily acacia with small leaves, makes it extremely hard to navigate from one point to another. In this area, survivors should follow animal trails and hope they lead to rivers or waterholes. Elephant trails seem to offer the best and clearest route.

(d) In the Sinai Desert area and in portions of Egypt, travel routes may be used; survivors can "stay put" on the trails. One survivor, who made a trail, encountered a camel caravan almost immediately, athough he reported it bothered him that he had not seen them approach, as they suddenly appeared out of a mirage. Another commented it was awfully hard to be alone in his section of the desert, for in every direction, he saw wandering tribes, camel herds, or people watching him. Two survivors independently suggested that survivors pay attention to the wind as an aid in navigation. One survivor, on the Arabian Peninsula, noted the wind blew consistently from the same direction. The other, in the Libyan Desert, made the same comment and said he was able to judge his direction of travel by the angle at which the wind blew his clothes or struck his body. Survivors in certain areas may orient themselves to the prevailing winds once it is established that these are consistent.

(4) People who walked across the North African deserts had much to say about the local environment and little of it was complimentary. The extreme temperatures bothered them most. It was extremely hot during the day and often bitter cold at night, especially during January and February.

(a) The bright sunlight was hard on their eyes, extremities, and exposed skin. The blinding effects of the Sun reflecting off the terrain caused many persons to express concern regarding sunglasses. Several built fires and smoked their goggles to obtain protection against the glare. Lenses alone do not adequately protect the eyes from glare. Sunglasses may be improvised to reduce this problem. Light-skinned individuals tended to sunburn faster and more severely than darker-skinned individuals. Some reported that no amount of previous suntanning seemed to make any difference. The heat affected their feet and hands. Survivors reported that the surface became so hot their feet were blistered through their shoes. Exposure of bare hands to the sunlight resulted in painful burns. Placing sunburned hands in bare armpits gave considerable relief since the armpits were one of the few places on the body where a person could find continuous perspiration to aid in cooling.

(b) The persistent winds of desert areas seem to provide no cooling effect, and several survivors found the constant blowing of the wind "got on their nerves." More significant is the fact that the constant winds usually carried an amount of sand or dirt particles. These particles got in eyes, ears, nostrils, and mouth and caused irritations which were often severe. Additionally, this persistent wind also caused earaches. One survivor reported that the abrasions of the eyes by the particles of dust reached a point where first the man's eyes watered so much that he could not see, and eventually the watering stopped and "emery cloth eyelids remained," making life miserable for him.

(c) Extreme winds blew sandstorms which lasted from a few minutes to months. Generally, survivors reported they could see the approach of such storms and were able to take proper precautions; however, sandstorms completely surprised a few groups and they had difficulty navigating. None of the survivors who experienced sandstorms in the northern desert underestimated the power and danger of such storms. Protection from the storms was uppermost in their minds. Most survivors used rock cairns, natural ledges, boulders, depressions, or wells for shelter. Survivors had time to dig depressions and rig a shelter from blankets, parachutes, or tarpaulins. A few wrapped themselves in their parachutes and endured the storm in a prone position.

(d) Nearly all of these people had some comment to make on orientation before, during, and after a sandstorm. They warned specifically that it is necessary to adequately mark the direction of travel before the storm. A few survivors said when the storm was over they had no idea which way they had been traveling and all their landmarks were forgotten, obliterated, or indistinguishable. The general plan for marking travel routes before a sandstorm is to place a stick to indicate direction. One survivor oriented himself with one rock a few feet in front of his position. He commented after the storm, that one point was not adequate and recommended using a row of stones, sticks, or heavy gear about 10 yards in length to give adequate direction following such an emergency.

(e) Several survivors reported they learned the hard way to keep their mouths shut in the desert. This meant breathing through the mouth caused drying and talking not only got on the other's nerves but also caused excessive drying of the mucous membranes.

(5) Mirages are common in desert areas. They are optical phenomena due to refraction of light by uneven density distribution in the lower layer of the air. The most common desert mirage occurs during the heat of the day when the air close to the ground is much warmer than the air aloft. Under this condition, atmospheric refraction is less than normal and the image of the distant low sky appears on the ground looking like a sheet of water. Distant objects may appear to be reflected in the "water." When the air close to the ground is much colder than the air aloft, as in the early morning under a clear sky, atmospheric refraction is greater than normal. When this condition occurs, distant objects appear larger and closer than they are and objects below the normal horizon are visible. Unless the density distribution in the lower layers is such that the light rays from an object reach the observer along two or more paths, they will see a distorted image or multiple images of the object.

(a) Reports of mirages were very common in the survival episodes examined. In most cases, they were recognized as mirages and only caused minor difficulties. No survivor reported these mirages actually represented bodies of water. While traveling, the survivor experienced problems as a result of mirages. Distances could not be judged because intermediate terrain was obscured by the mirage. Mirages hampered vision and navigation since it concealed objects. Additionally, mirages "magnified some objects and concealed others." One man hunting in the heat of the day reported that when he sighted an animal, it ran into or hid in a mirage. The lower layer of hot air which causes the mirage, commonly called desert haze, hampers vision and distorts objects. Signaling difficulties resulted from this since sighting a reflection on an object was apparently very difficult due to the low haze on the desert.

(b) Several survivors reported cases of imaginary illusions which were due to the haze or mirages. One group looked for a hill, for a viewpoint, so long that the entire party began to see hills in all directions. They finally held a conference to iron out their difficulties and all settled on one hill which the group should approach. Everyone in the party saw the hill, and the group walked an estimated 9 miles looking for the hill which never existed and which eventually disappeared into the desert flatness. Dawn and dusk illusions also occurred and were reported in the survivors' stories. One group was severely troubled with the false-dawn spectral light on the western horizon. The fact that the Sun at first appeared to be rising in the west caused anxious moments. Another party had one person who claimed he saw a flashing beacon on the evening horizon. The pilot explained the illusion as the occasional refraction of bright starlight near the horizon, through the residual heat waves of the cliff before them. But one crewmember was so convinced it was a beacon that he started walking to investigate this object and was never seen again.

(6) The following are manmade characteristics of the desert:

(a) Roads and trails are scarce in the desert, as complex road systems are not needed. Most existing road systems have been used for centuries to connect centers of commerce or important religious shrines, such as Mecca and Medina in Saudi Arabia. These roads are now supplemented by routes for transporting oil or other mineral deposits to collection points. Rudimentary trails exist in many deserts for caravans and nomadic tribesmen. These trails have wells or oases approximately every 20 to 40 miles, although there are waterless stretches of over 100 miles. These trails vary in width from a few yards to over 800 yards. Vehicle travel in mountainous desert terrain may be severely restricted. Available routes may be easily blocked by hostile people or climatic conditions. Passes may be blocked by snow in the winter. The travel distance on foot or by animal between two points in the mountains may be less than a tenth of the distance required if vehicles are used to make the trip.

(b) Apart from nomadic tribesmen who live in tents, desert inhabitants live in thick-walled structures with small windows, usually built of masonry or a mud and straw (adobe) mixture. The ruins of earlier civilizations are scattered across the deserts. Ancient posts and forts invariably command important avenues of approach and frequently dominate the only available passes in difficult terrain. Control of these positions may be imperative for forces intent on dominating the immediate area.

(c) Exploration for and exploitation of natural resources, of which oil is the best known, occurs in many desert areas, especially the Middle East. Wells, pipelines, refineries, quarries, and crushing plants may lead a survivor to rescue or captivity. Additionally, pipelines are often elevated, making them visible from a distance.

(d) Many desert areas are irrigated for agricultural and habitation purposes. Agriculture and irrigation canals are signs which can lead a survivor to people.

f. Tropical Climates. The inexperienced person's view of jungle travel may range from difficult to nearly impossible. However, with patience and good planning, the best and least difficult route can be selected. In some cases, the easiest routes are rivers, trails, and ridge lines. However, there may be hazards associated with these routes.

(1) Rivers and streams may be overgrown making them difficult to reach and impossible to raft. These waterways may also be infested with leeches. Trails may have traps or animal pits set on them. Trails can also lead to a dead end or into thick brush or swamps. Ridges may end abruptly at a cliff. The vegetation along a ridge may also conceal crevices or extend out past cliffs, making the cliff unnoticed until it's too late.

(2) The machete is the best aid to survival in the jungle. However, survivors should not use it unless there is no other way. They should part the brush rather than cut it if possible. If the machete must be used, cut at a down-and-out angle, instead of flat and level, as this method requires less effort.

(3) Survivors should take their time and not hurry. This allows them to observe their surroundings and gives better insight as to the best route of travel. Watch the ground for the best footing as some areas may be slippery or give way easily. Avoid grabbing bushes or plants when traveling. Falling may be a painful experience as many plants have sharp edges, thorns, or hooks. (NOTE: Wear gloves and fully button clothes for personal protection.)

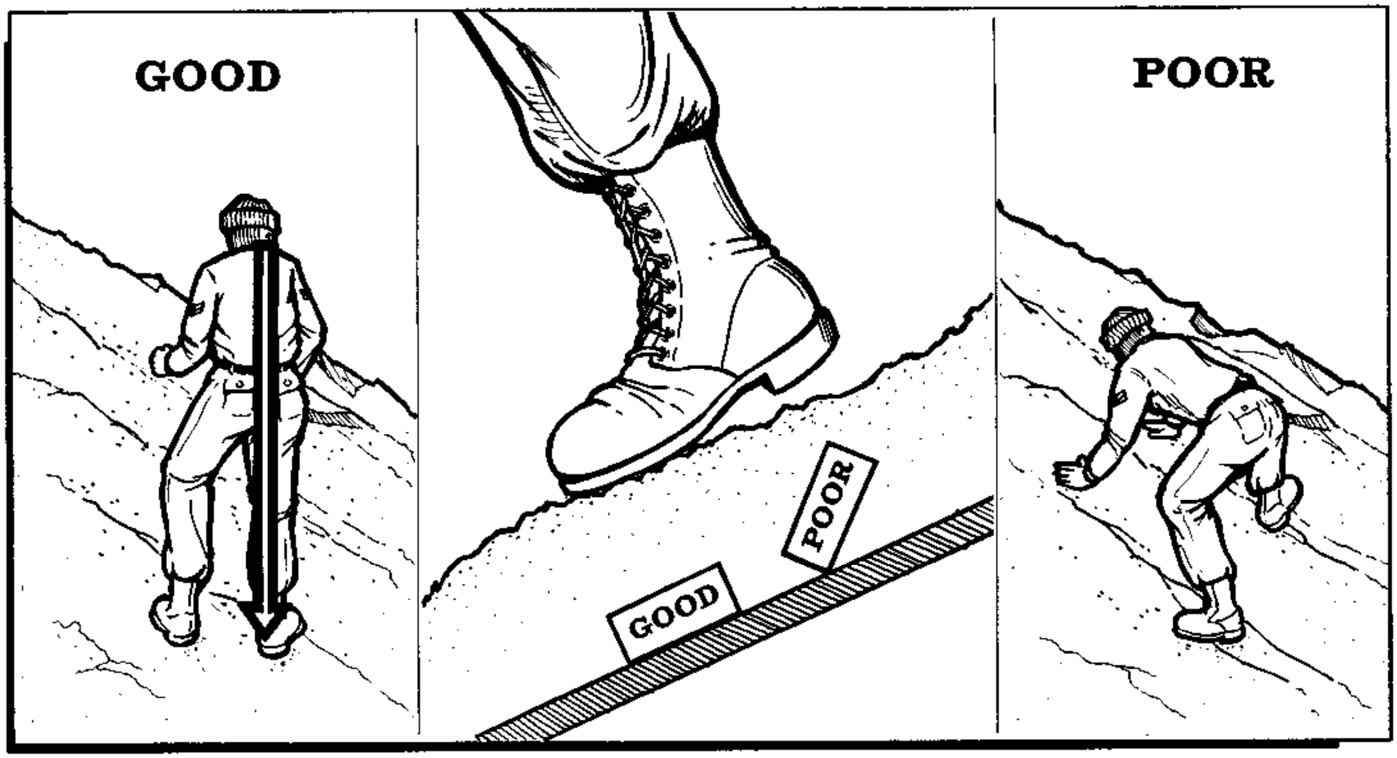

Figure 21-6. Weight of Body Over Feet.

(4) Quicksand can be a problem. In appearance, quicksand looks just like the surrounding area with an absence of vegetation. It is usually located near the mouths of large rivers and on flat shores. The simplest description of quicksand is a natural water tank filled with sand and supplied with water. The bottom consists of clay or other substances capable of holding water. The sand grains are rounded, as opposed to normal sharper-edged sand. This is caused by water movement which also prevents it from settling and stabilizing. The density of this sand-water solution will support a person's body weight. The danger if a survivor panics may be drowning. In quicksand, the survivor should use the spread-eagle position to help disperse the body weight to keep from sinking and a swimming technique to return to solid ground. (NOTE: Remember to avoid panicking and struggling, and spread out and swim or pull along the surface.)

21-5. Mountain Walking Techniques. Depending upon the terrain formation, mountain walking can be divided into four different techniques—walking on hard ground, walking on grassy slopes, walking on scree slopes, and walking on talus slopes. Included in all of these techniques are two fundamental rules which must be mastered in order to expend a minimum of energy and time. These fundamentals are: The weight of the body must be kept directly over the feet (figure 21-6), and the sole of the boot must be placed flat on the ground (figure 21-6). These fundamentals are most easily accomplished by taking small steps at a slow steady pace. An angle of descent which is too steep should be avoided and any indentations or protrusions of the ground, however small, should be used to advantage.

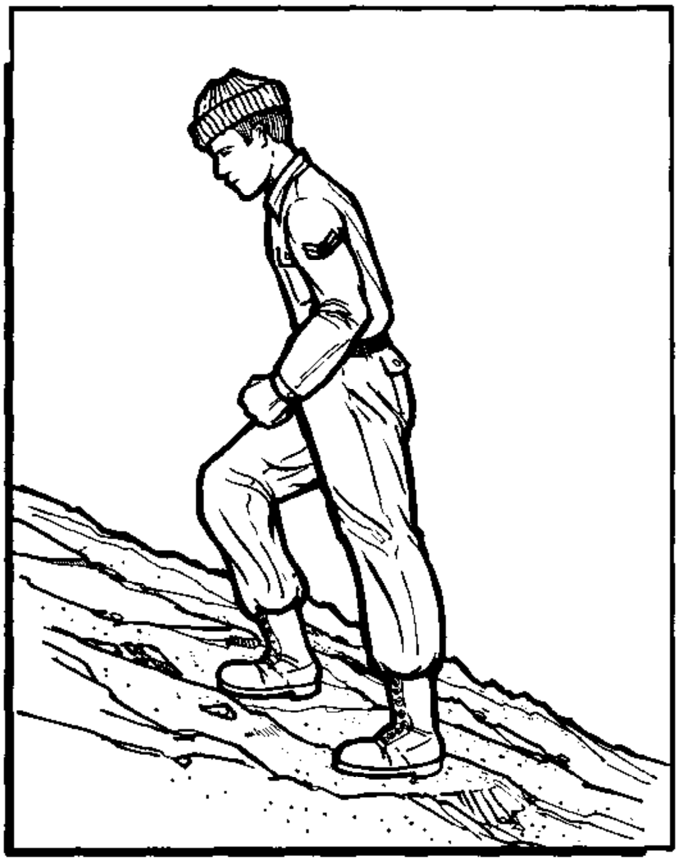

Figure 21-7. Locking Knees.

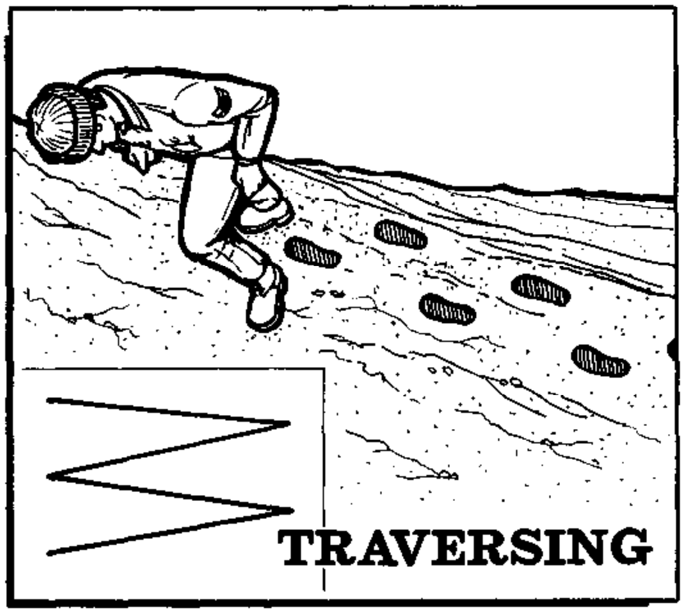

Figure 21-8. Traversing.

a. Hard ground is usually considered to be firmly packed dirt which will not give way under the weight of a person's step. When ascending, the above fundamentals should be applied with the addition of locking the knees on every step in order to rest the leg muscles (figure 21-7). When steep slopes are encountered, they can be traversed easier than climbed straight up. Turning at the end of each traverse should be done by stepping off in the new direction with the uphill foot (figure 21-8). This prevents crossing the feet and possible loss of balance. In traversing, the full sole principle is done by rolling the ankle away from the hill on each step. For narrow stretches, the herringbone step may be used; that is, ascending straight up a slope with toes pointed out (figure 21-9) and using the principles stated above. Descending is usually easiest by coming straight down a slope without traversing. The back must be kept straight and the knees bent (figure 21-10) in such a manner that they take up the slack of each step. Again, remember the weight must be directly over the feet, and the full sole must be placed on the ground with every step.

b. Grassy slopes are usually made up of small tussocks of growth rather than one continuous field. In ascending, the techniques previously mentioned are applicable; however, it is better to step on the upper side of each tussock (figure 21-11) where the ground is more level than on the lower side. Descending is best done by traversing.

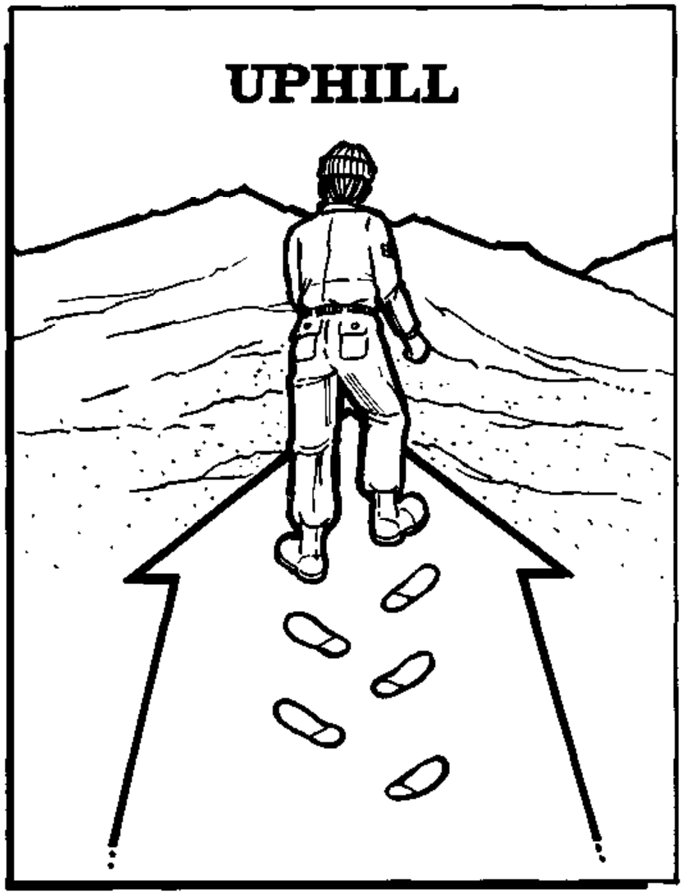

Figure 21-9. Walking Uphill.

c. Scree slopes consist of small rocks and gravel which have collected below rock ridges and cliffs. The size of the scree varies from small particles to the size of a fist. Occasionally, it is a mixture of all size rocks, but normally scree slopes will be made up of rocks the same size. Ascending scree slopes is difficult, tiring, and should be avoided when possible. All principles of ascending hard ground apply, but each step must be picked carefully so the foot will not slide down when weight is placed on it. This is best done by kicking in with the toe of the upper foot so a step is formed in the scree (figure 21-12). After determining the step is stable, carefully transfer weight from the lower foot to upper and repeat the process. The best method for descending scree is to come straight down the slope with feet in a slight pigeon-toed position using a short shuffling step with the knees bent and back straight (figure 21-13). When several climbers descend a scree slope together, they should be as close together as possible, one behind the other, to prevent injury from dislodged rocks. Since there is a tendency to run down scree slopes, care must be taken to ensure that this is avoided and control is not lost. By leaning forward, one can obtain greater control. When a scree slope must be traversed with no gain or loss of altitude, use the hop-skip method. This is a hopping motion in which the lower foot takes all the weight and the upper foot is used for balance.

d. Talus slopes are similar in makeup to the scree slopes, except the rock pieces are larger. The technique of walking on talus is to step on top of, and on the uphill side of, the rocks (figure 21-14). This prevents them from tilting and rolling downhill. All other previously mentioned fundamentals are applicable. Usually, talus is easier to ascend and traverse, while scree is a more desirable avenue of descent.

21-6. Burden Carrying. Backpacking is essential when heavy loads must be carried for distances. Using a suitable harness and following certain approved packing and carrying techniques can eliminate unnecessary hardships and help in transporting the load with greater safety and comfort. Carrying a burden initially creates mental irritation and fatigue, either of which can lower morale. Survivors should keep their minds occupied with other thoughts when packing a heavy load. Adjustments should be made during each rest stop to improve the fit and comfort of the pack. Additionally, the rate of travel should be adjusted to the weight of the pack and the environmental characteristics of the terrain being crossed.

a. Burden carrying is a task that must be done in most survival environments. Often survivors must quickly

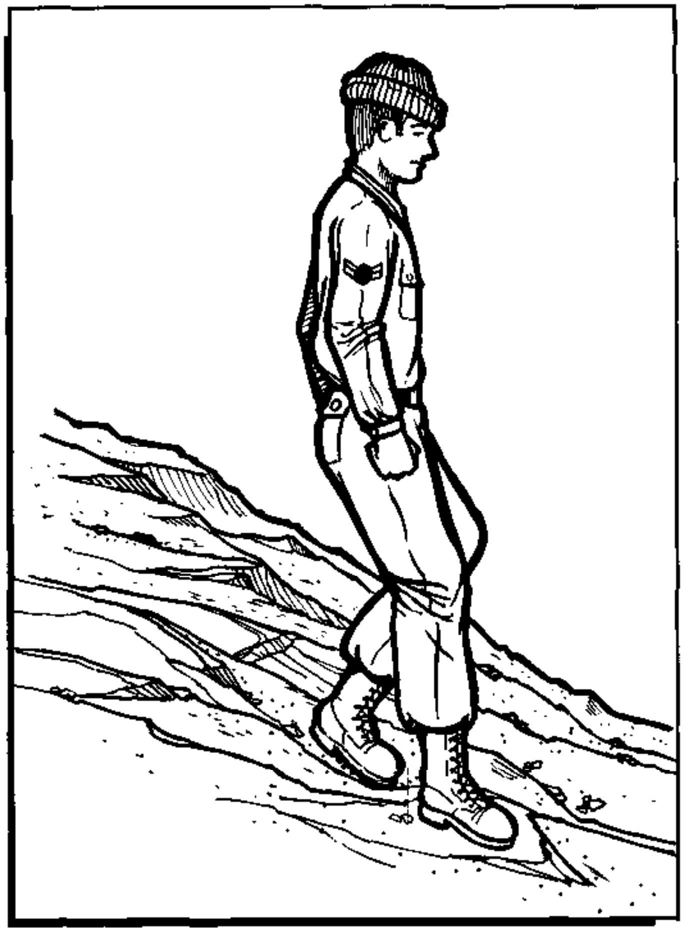

Figure 21-10. Walking Downhill.

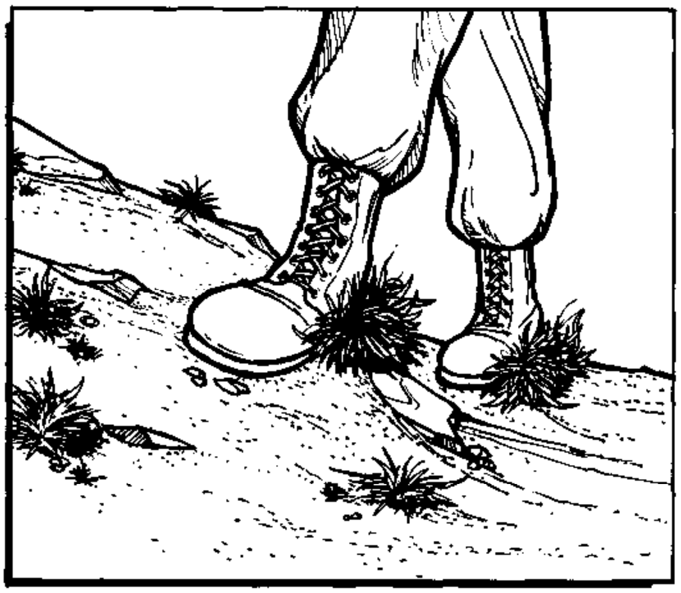

Figure 21-11. Walking on Grassy Slopes.

gather their equipment and move out without the assistance of a good pack. The gear may have to be carried in the arms while rapidly leaving the area. In such an instance, it would be better to fashion a roll of the gear and wear it over the shoulder, time permitting. When time is not a factor, it may be desirable to make a semirigid pack such as as square pack. The convenience of being able to keep track of equipment, particularly small items, can be critical in survival situations.

(1) Packsack. A packsack can be fashioned from available survival kit containers or several layers of the fabric from the parachute canopy. The sack is sewed with inner core and a needle.

(2) Square Pack:

(a) The following instructions explain how to improvise a square pack (figure 21-15). Lay a rectangle of

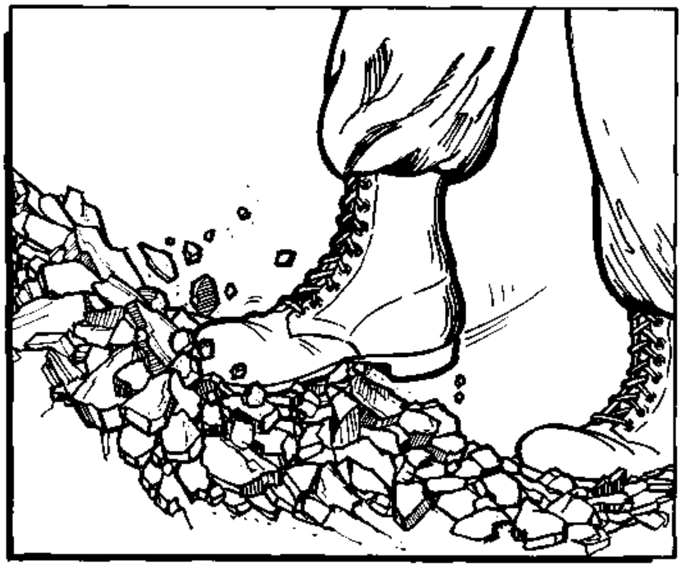

Figure 21-12. Walking on Scree Slopes.

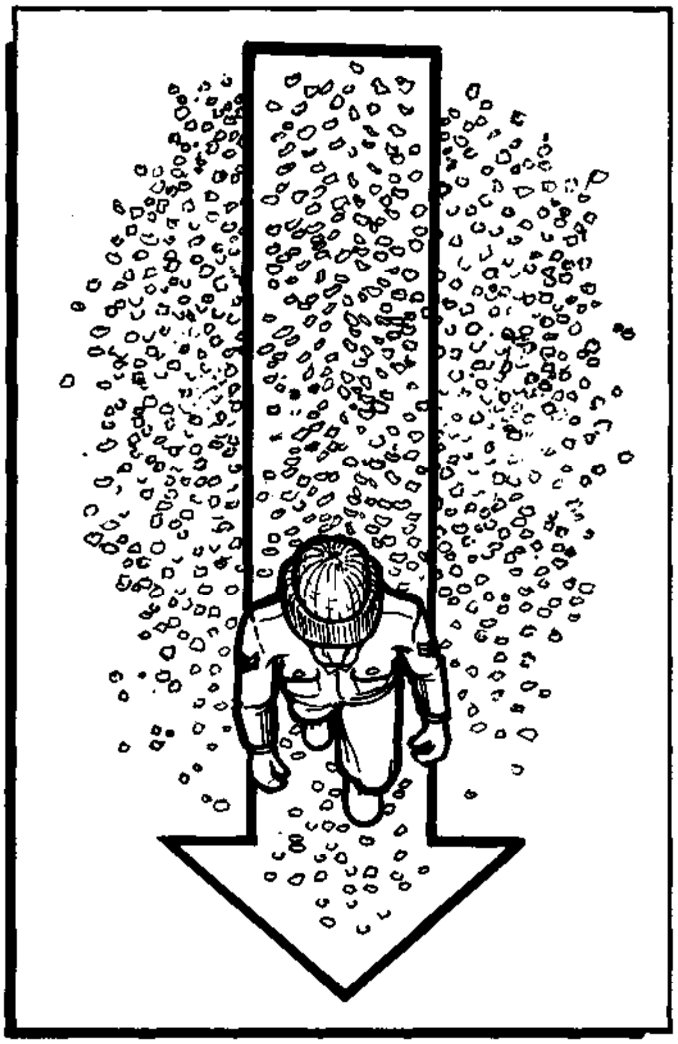

Figure 21-13. Walking Down Scree Slopes.

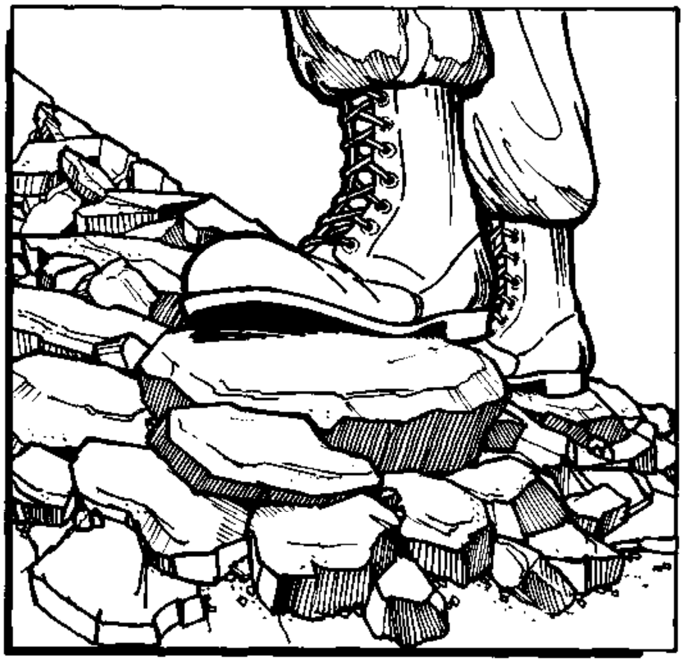

Figure 21-14. Walking on Talus Slopes.

material, waterproof if available, (5 feet by 5 feet minimum size) flat on the ground (examples are plastic, tarp, poncho, paulin 77, etc.) (illustration A). Visualize the material divided into squares like a tic-tac-toe board. The largest piece of soft, bulky equipment (sleeping bag, parachute canopy, etc.) is placed in the center square in an "S" fold. This places the softest item in the pack against the tarp which rests on the back while traveling. If using a poncho, place the sleeping bag just below the hood opening (illustration B). Place hard, heavy objects between the top layer and the middle layer of the "S" fold near the top of the pack. Soft items can be placed between the middle and bottom layer (illustration C).

(b) After all desired items are inside the folds, tie the inner pack in the fashion shown in illustration D. Start with a 1-inch diameter loop in the end of a long piece of parachute suspension line or other suitable line and loop it around the "S" fold laterally. Standing at the bottom of the pack, divide it into thirds and secure the running end of the line to the loop with a trucker's hitch. Both of these hitches should be at the intersection of the thirds so as to divide the pack vertically into thirds. Wrap the running ends around the pack at 90 degrees (working toward the center) to the line and when it crosses another line, use a jam (postal) hitch to secure it and pull both ways to ensure tightness in all directions. When returning to the original Figure 22-42 starting position, use the loop of the tied trucker's hitch to secure another trucker's hitch and the inner part is complete (illustration D). The waterproof materials are then folded around the inner pack as shown in illustrations E and F. Tie the "outer" pack in the same manner ensuring that it is waterproof with all edges folded in securely. If a poncho is used, the head portion may be used to get into the pack if necessary. However, it must be properly secured to ensure that the inner items are protected. With a square pack constructed in this manner, there is no reason why the equipment should get wet. (NOTE: With practice, an excellent pack can be

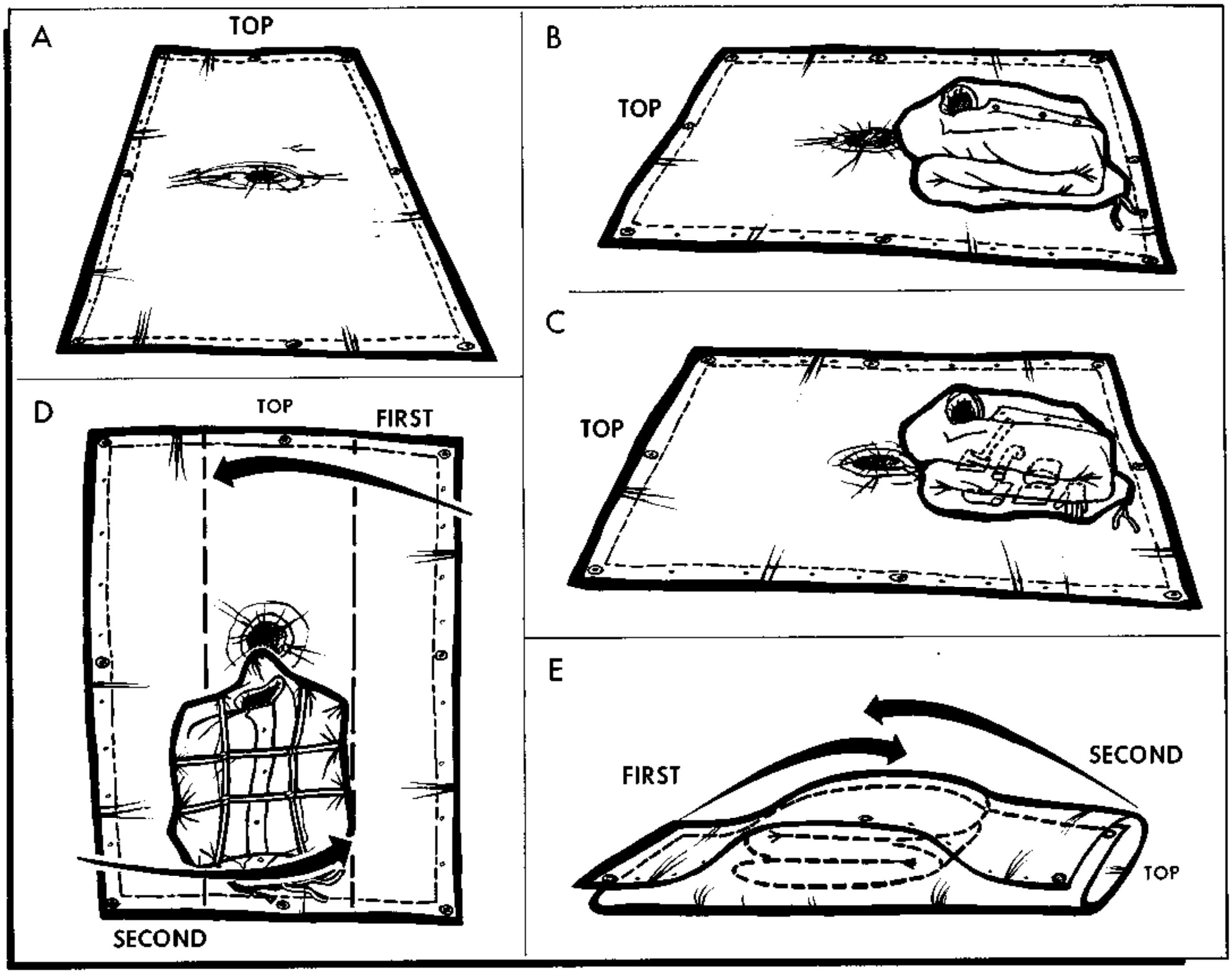

Figure 21-15. Square Pack.

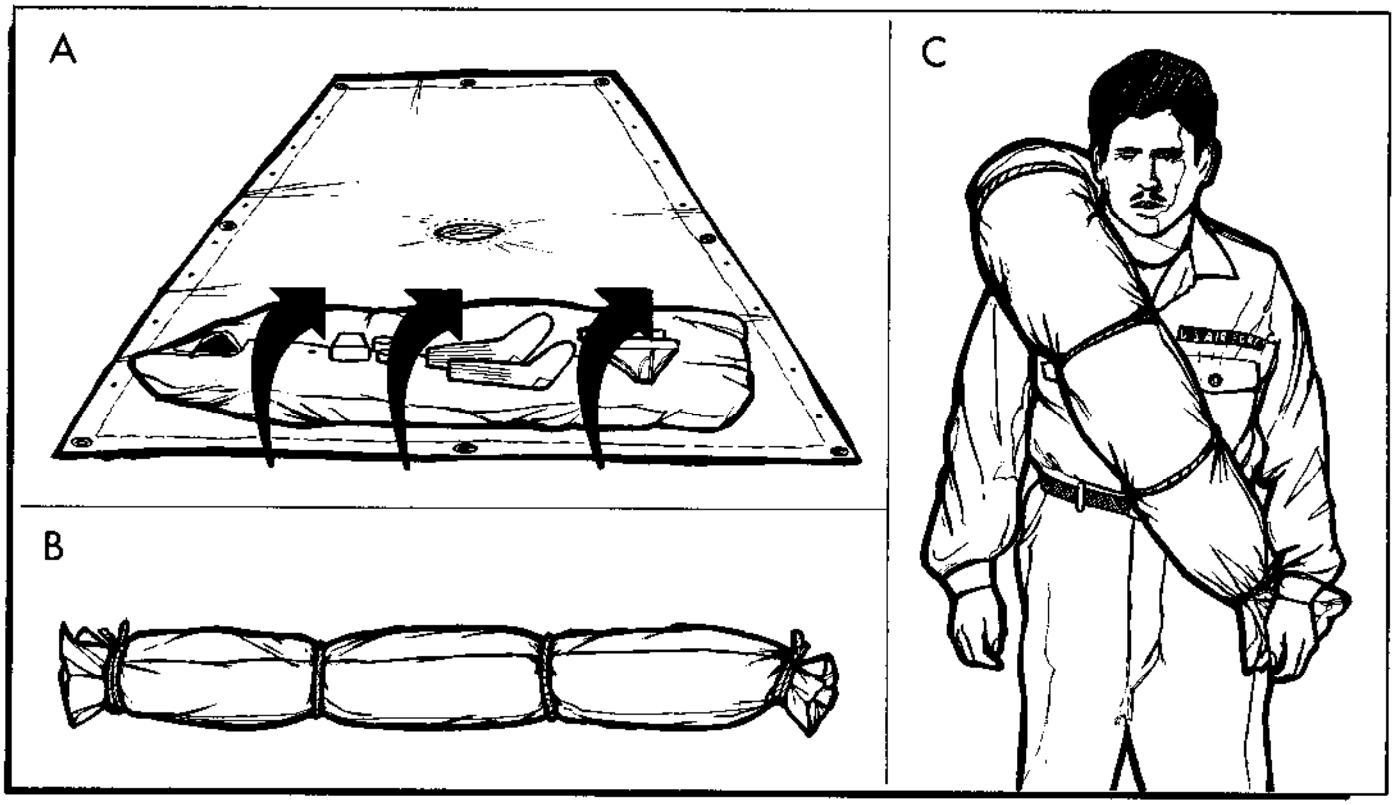

Figure 21-16. Horseshoe Pack.

constructed by tying the inner pack and outer cover simultaneously.)

(3) Horseshoe Pack. The horseshoe pack has been referred to by many names in history including the "Johnny Reb Pack," "cigarette pack," "bed roll," and others. It is simple to make and use and relatively comfortable to carry over one shoulder. It is constructed as follows (figure 21-16): Lay available square-shaped (preferably 5 feet by 5 feet) material (waterproof if available) flat on the ground (illustration A) and place all gear on the long edge of the material, leaving about 6 inches at each end. All hard items should be padded. Roll the material with the gear to the opposite edge of the square (illustration B). Tie each end securely. Place at least two or three evenly spaced ties around the roll. Bring both tied ends together and secure. This pack is compact and comfortable if all hard, heavy items are packed well inside the padding of the soft gear. If one's shoulder is injured, the pack can be carried on the other shoulder. It is easy to put on and remove (illustration C).

b. The most widely used improvised packstrap is called an Alaskan packstrap (figure 21-17). This type of packstrap can be fashioned out of any pliable and strong material. Some suitable materials for constructing the packstrap are animal skins, canvas, and parachute harness webbing. The pack should be worn so it can be released from the strap with a single pull of the cord in the event of an emergency, such as falling into water. The knot securing the pack should be made with an end readily available which can be pulled to drop the pack quickly; for example, a trucker's hitch with safety for normal terrain travel and with the safety removed when in areas of danger, such as water or rough terrain.

(1) Some advantages of the Alaskan packstrap are:

(a) Small in bulk and light in weight.

(b) Easily carried in a pocket while traveling.

(c) Quickly released in an emergency.

(d) Can be adjusted to efficiently pack items of a variety of shapes and sizes.

(e) Can be used with a tumpline to help distribute the weight of the pack over the shoulders, neck, and chest, thereby eliminating sore muscles and chafed areas.

(2) Some disadvantages of the Alaskan packstrap are:

(a) Difficult to put on (without practice).

(b) Experience and ingenuity are necessary to use it with maximum efficiency.

c. The following principles should be considered when packing and carrying a pack:

(1) The pack or burden-carrying device should be adequate for the intended job.

(2) The pack or burden may be adaptable to a pack frame. The pack frame could have a belly band to distribute the weight between the shoulders and hips and prevent undue swaying of the pack. Pack frames are also used to carry other burdens such as meat, brush, and firewood.

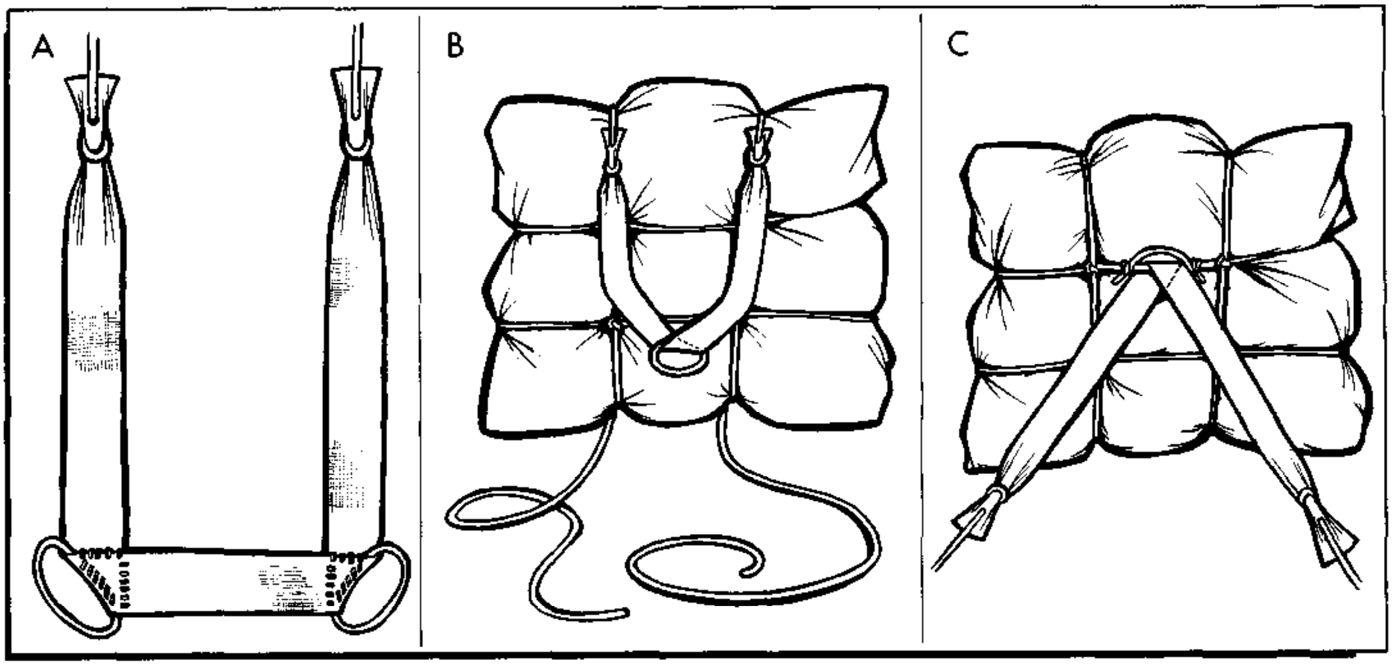

Figure 21-17. Alaskan Packstrap.

(3) Proper weight distribution is achieved by ensuring that the weight is equally apportioned on each side of the pack and as close as possible to the body's center of gravity. This enhances balance and the ability to walk in an upright position. If heavy objects are attached to the outside of the pack, the body will be forced to lean forward. A pack bundle without a frame or packboard should be carried high on the back or shoulders. For travel on level terrain, weight can be carried high. When traveling on rough terrain, weight should be carried low or midway on the back to help maintain balance and footing.

(4) Emergency and other essential items (extra and (or) protective clothing, first-aid kit, radio, flashlight, etc.) should be readily available by being placed in the top of the pack.

(5) Fragile items are protected by padding them with extra clothing or some soft material and placing them in the pack where they won't shift or bounce around. Hard and (or) sharp objects cannot damage the pack or other items if cutting edges are properly sheathed, padded, and not pointed toward the bearer. Items outside the pack should be firmly secured but not protruding where they could snag on branches and rocks.

(6) Adjust and carry the pack so that overloading or straining of muscles or muscle groups is avoided. When using a pack, the straps should be adjusted so they ride comfortably on the trapezius muscles and avoid movement when walking. Back support should be tight and placed to ensure good ventilation and support. During breaks on the trail, rest using the proper position to ease the weight of the pack and take the strain off muscles. (See figure 21-18 for methods of resting.) A comfortable pack is adjustable to the physique of the person. A waistband will support 80 to 90 percent of the weight and is fitted relatively tight. The waistband should be cinched down around the pelvic girdle/crest area to avoid constricting circulation or restricting muscle movement.

d. A tumpline is an excellent aid to burden carrying since it transfers part of the weight of the pack to the skeletal system (figure 21-19).

(1) Tumpline Construction:

(a) Attach a soft band, which rests on the upper forehead, to the pack by using light line. Make the band of any strong, soft material, such as animal skin with hair, tanned skin, an old sock, or parachute cloth. Make the band long enough to reach over the forehead and down to a point opposite each ear. A tumpline does not require any sewing.

(b) Adjust the tumpline to fit the head by making loops at the ends. It is difficult to reach down and behind to make necessary adjustments of the pack, but a person can easily reach up and adjust the pack by using the loops on either side of the forehead.

(c) Make mainstrings from rawhide or parachute suspension lines. Tie them to the lower corners of the pack, bring them up to the loops at the ends of the tumpline, and tie them with a slipknot. Experience is needed to estimate proper adjustments before putting on the pack; however, adjustments can always be made after the packstraps are adjusted.

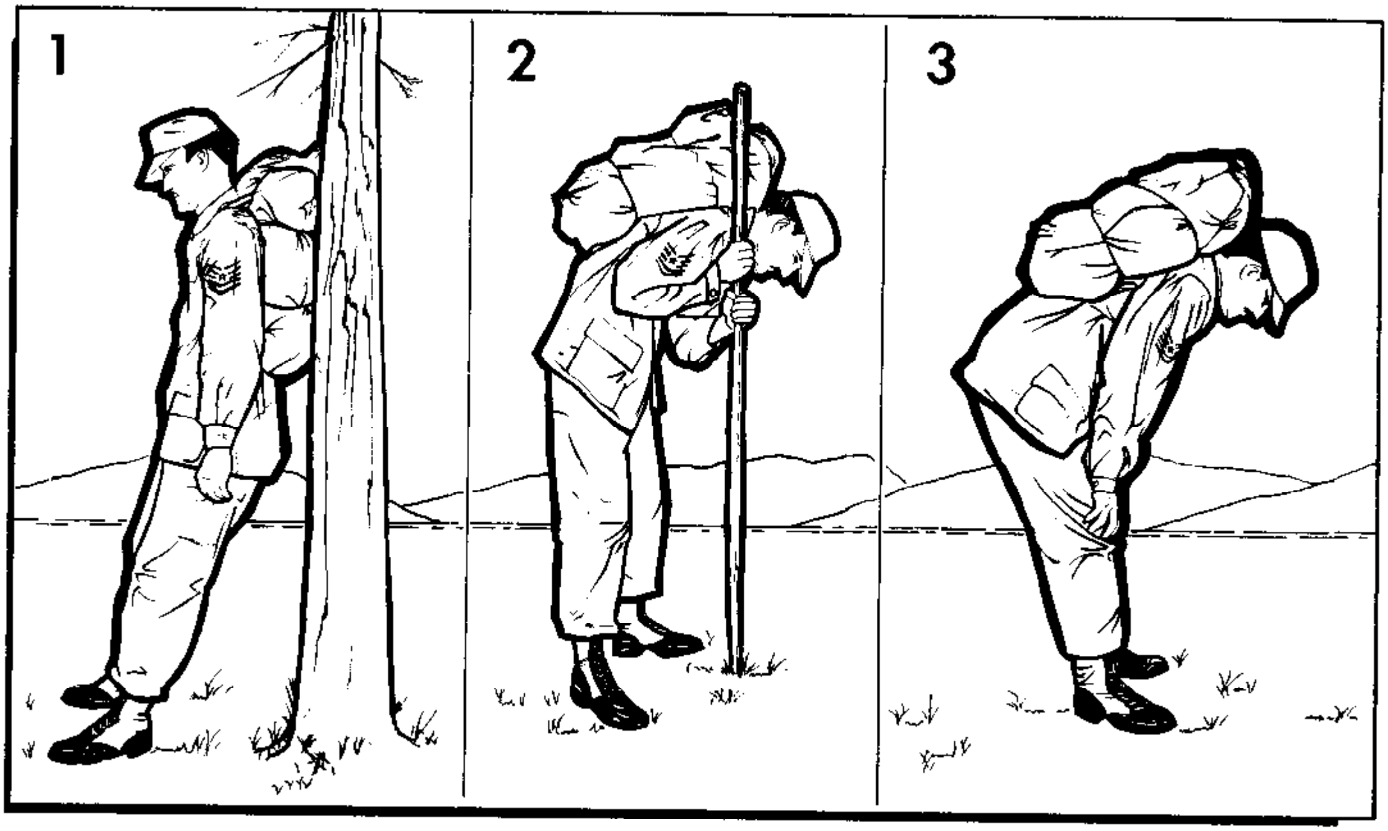

Figure 21-18. Methods of Resting.

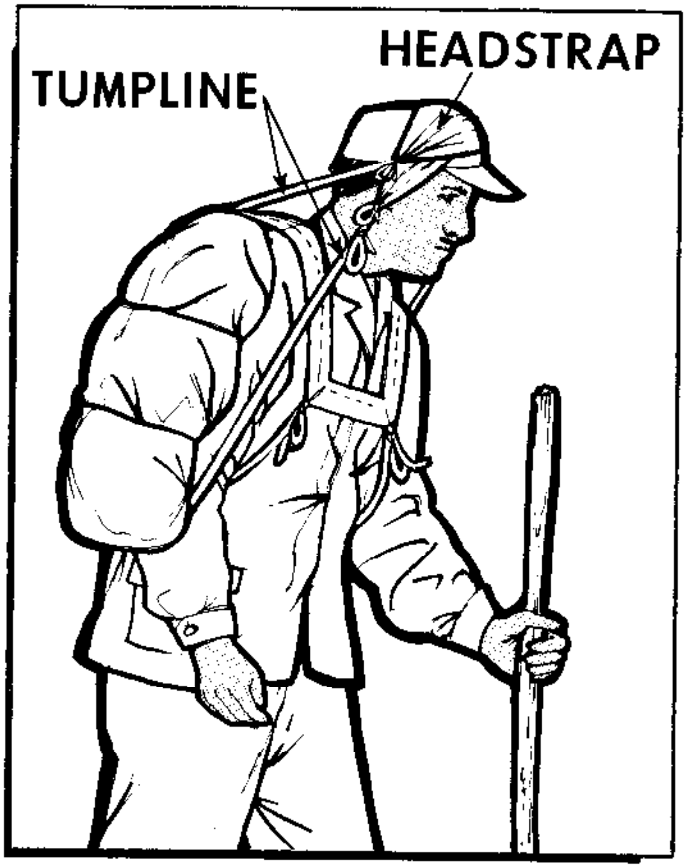

Figure 21-19. Tumpline.

(2) Tumpline Use. The tumpline should be tight enough to transfer about half of the weight of the pack/burden from the shoulders to the head. Occasionally, a heavy pack may cut off the blood circulation to the shoulders and arms. In such cases, a tumpline is of great value. By slight adjustments, most of the weight can be transferred to the head and neck, thereby loosening the shoulder straps and permitting circulation to return to numb arms. A tumpline may cause the neck muscles to be sore for the first few days due to the unusual strain placed on them; however, this discomfort soon disappears. With practice, heavy weights can be supported with only the neck and head. A tumpline can also be used to pack animal carcasses, firewood, or equipment. Since it can be rolled up and carried in a pocket, it can be a real aid to survival.

Chapter 22

ROUGH LAND TRAVEL AND EVACUATION TECHNIQUES

22-1. Introduction. There are survival situations where traveling over rough terrain is required. However, if it is necessary to travel in such areas, specialized skills, knowledge, and equipment are required. These skills may include rope management, specialized knots, belaying and climbing techniques, prusik climbing, rope bridges, rappelling, litter evacuation, etc. The amount and type of equipment needed will depend on the type of operation (climbing, evacuation, etc.).

22-2. Safety Considerations. A safety rope must be used when there is danger of the rescue team or climber falling. The rescue team leader will identify the members of the team who will climb together. Any member of the team may request to be roped during any mountain operation. The rope will be maintained as long as requested by any member. The rescue team leader will decide whether or not the entire team will rope up. Environmental factors may require a rapid retreat. During these circumstances, when speed is critical, it may be desirable to unrope. The rescue team leader's decision to unrope may be overridden by any team member who desires to remain roped. This option pertains to areas such as ridge climbs and low-grade climbs when roping is not required. Solo climbing in severe terrain is not recommended.

22-3. Climbing Ropes. Climbing ropes are made of synthetic fibers, such as nylon. The two essential qualities of a climbing rope are tensile strength and the ability to absorb the shock of a fall. This combination of elongation plus strength varies with the type of rope construction.

a. There are two basic types. The tensile strengths average 5,500 pounds, and the construction designs are called the Kernmantle and the Hauser Lay.

(1) Kernmantle rope consists of a core of braided or twisted strands protected by a braided sheath. They have a 6 to 8 percent stretch factor. Depending upon use, Kernmantle ropes are 9 to 11 millimeter in diameter; standard lengths are 150 feet and 165 feet. Ropes for wet conditions are also available.

(2) Mountain-Lay (three strand hard lay) ropes consist of a right-hand lay of three main strands twisted around each other. These ropes have a 9 to 13 percent stretch (twice as much as Kernmantle ropes) factor. They are constructed in 120-, 150-, and 165-foot lengths and are three-eights to seven-sixteenths inches in diameter.

b. Climbing ropes and anchor slings require special care. Stepping on the ropes will grind abrasive dirt particles into them which will cut the fibers of the rope. Contact between the rope and sharp corners or edges of rocks should be avoided. This may cut the rope. If the rope must run over a sharp edge, it should be padded (figure 22-1). The ropes should be kept dry because wet ropes are not as strong and collect more dirt. Do not hang ropes on sharp edges. Do not let one rope rub against another during use. Nylon rubbing on nylon will create friction which can burn through the rope(s). This is called weld abrasion. Smoking around ropes should be prohibited and they should not contact any source of excessive heat. Petroleum products will accelerate deterioration of nylon. When not in use, protect the rope from the ultraviolet rays of the Sun which also have a deteriorating effect on nylon.

c. Before constructing any rope system, the ropes should be backcoiled or layed out. This is done by simply removing the rope from the coil and, starting with one end, arranging the rope into neat overlapping loops on the ground. This inspection ensures the rope is free of knots or kinks (figure 22-1).

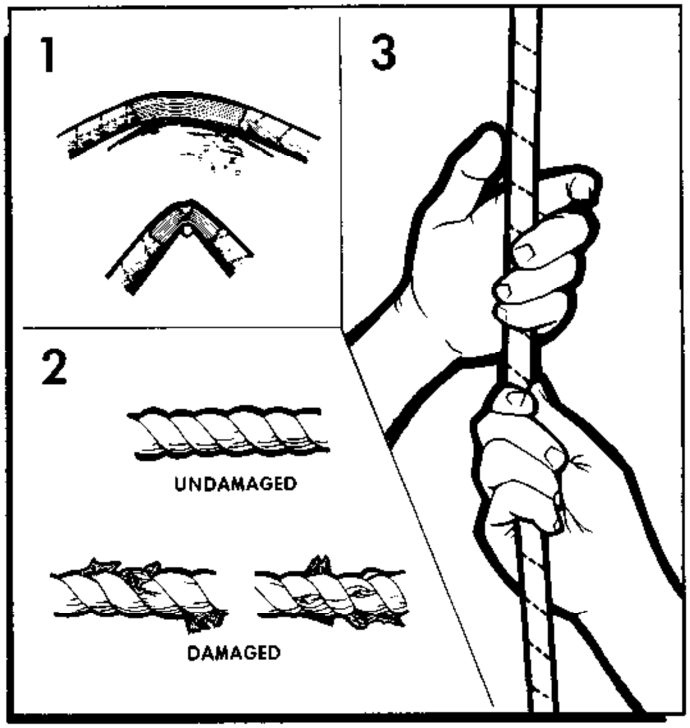

Figure 22-1. Damaged Rope.

(1) Before throwing the rope, one end should be secured. This end should be the standing end on the backcoil. Next, several small loops are taken from the working end and placed in one hand, several more loops are placed in the other hand. The throw is done by raising the hand closest to the rope end (running end)

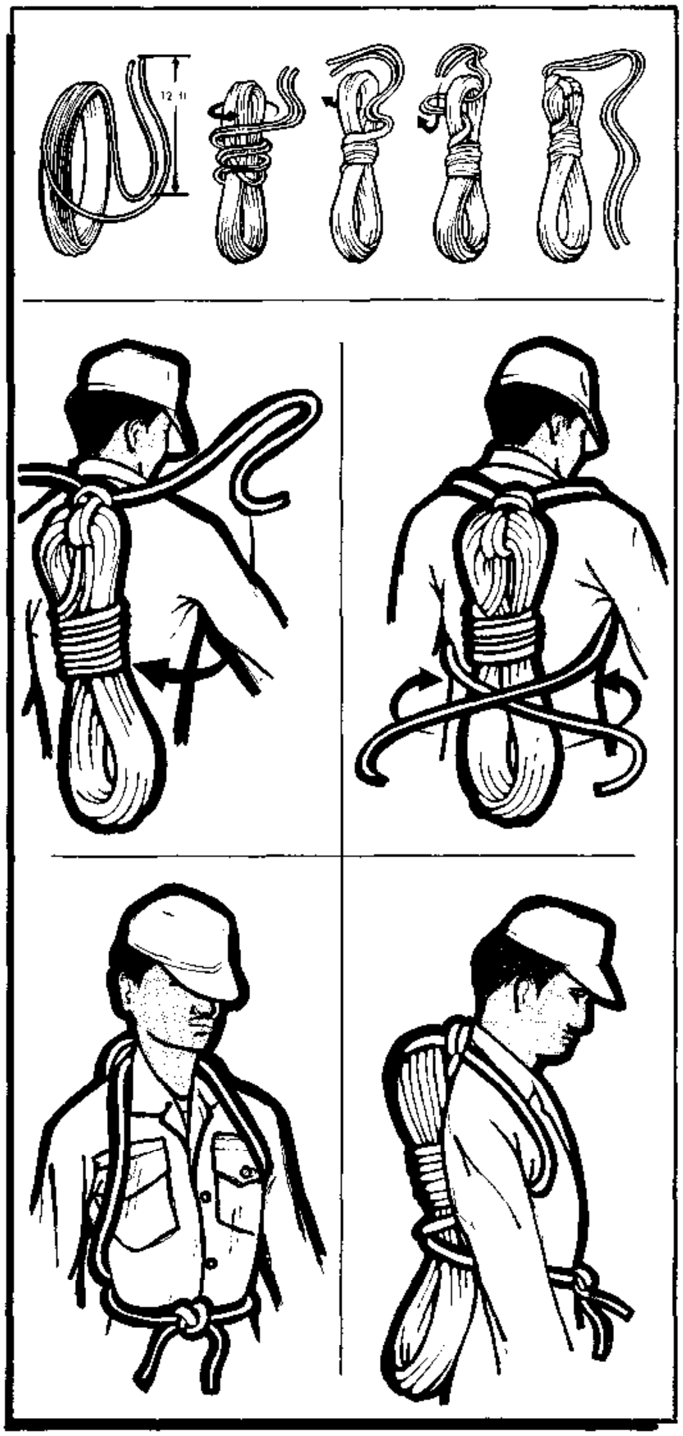

Figure 22-2. The European Coil.

over the head and throwing the loops out and down in the intended direction. The loops in the other hand are allowed to drop immediately after the running end is thrown. Properly done, the remaining rope will feed out. Before throwing the rope, the term "rope" should be sounded. This will alert anyone below to the danger of falling rope. Wait for the response "clear" if there is someone below. Under normal conditions, there should be no knots in the rope or any equipment attached to the rope due to the possibility of the rope becoming entangled if it should snag on a rock, brush, etc.

(2) Two common methods are used to coil rope:

(a) The European coil is done by starting in the center of the rope, coiling the rope into 3-foot diameter loops until there is about 12 feet of rope tail left on the two ends. The ends are then wrapped three to four times around the coil, a bight is passed through the center of the coil, and the tails passed through the bight. This is then tightened down with two remaining tails of about 6 feet. The coil can now be carried by placing the coil on the back, passing one end over each shoulder, between the arm and the chest, and passing around the back to return to the front of the body. The two ends are now secured at waist level (figure 22-2).

(b) To form a mountain coil, one end of the rope is taken in one hand while the other hand is run along the rope until both hands are outstretched. The hands are then brought together forming a loop which is transferred to one hand. This is repeated, forming uniform coils, until the entire rope is coiled. To secure the coil, a 12-inch bight is made in the standing end of the rope and laid on top of the coil. Drop the last loop on the coil and take this length of rope and wrap it around the coil and the bight. The first wrap is made at the open end of the bight to lock. Continue to wrap toward the closed end of the bight until just enough rope remains to be run through the inside of the bight. Pull the running end of the bight to secure the coil (figure 22-3). The coil may now be carried draped over a pack or over one shoulder and beneath the other.

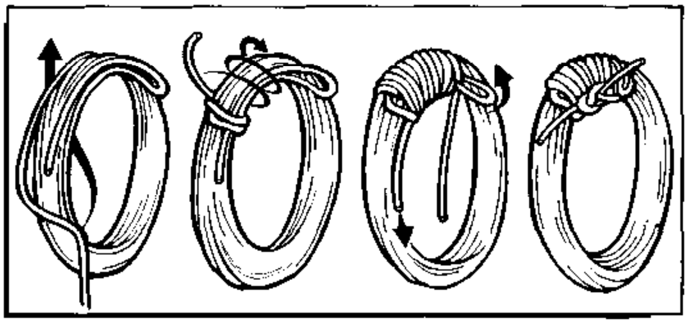

Figure 22-3. The Mountain Coil.

22-4. Specialized Knots for Climbing and Evacuation. Each of the following knots have a specific purpose. These knots have survived the test of time and are used in maintaining operations. They are designed to have the least effect on the fiber of a rope lock without slipping, and they are easy to untie when wet and icy. All knots reduce the strength of ropes; however, these knots reduce the strength of the rope as little as possible. Most

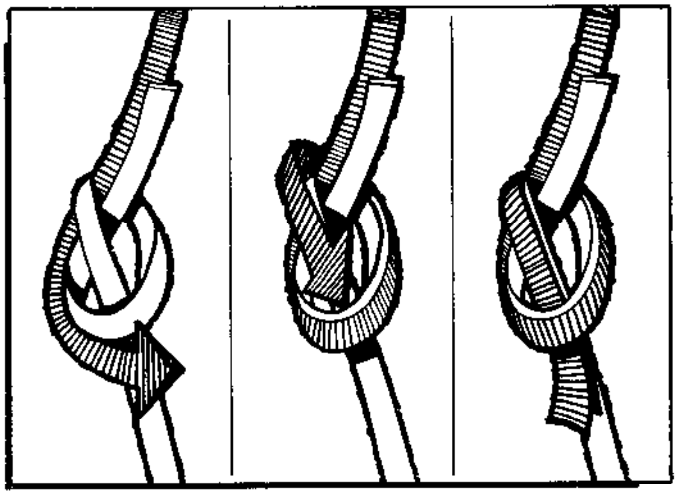

Figure 22-4. The Water Knot.

knots should be safetied with an overhand knot or two half hitches. A knot does not have to be safetied if the knot is designed for the middle of a line or if it is 15 feet or more from the end of the rope.

a. The water knot or right bend is used for joining nylon webbing (figure 22-4).

b. The double fisherman's bend is used to securely join Kernmantle or hard lay lines (figure 22-5).

c. The double figure-eight knot is used for temporarily joining Kernmantle or hard lay ropes (rappels, Tyrolean traverses) (figure 22-6A). Carabiners should be placed between lines within the knot to prevent the knot from clinching down when loaded. The figure-eight loop may be tied at the end or in the middle of a line establishing a fixed loop (figure 22-6B). If the loop is tied at the end of the line, then an overhand or single fisherman's safety knot must be used. The figure eight on a bight is used to provide two loops from a single knot to attach to two different anchors (figure 22-7).

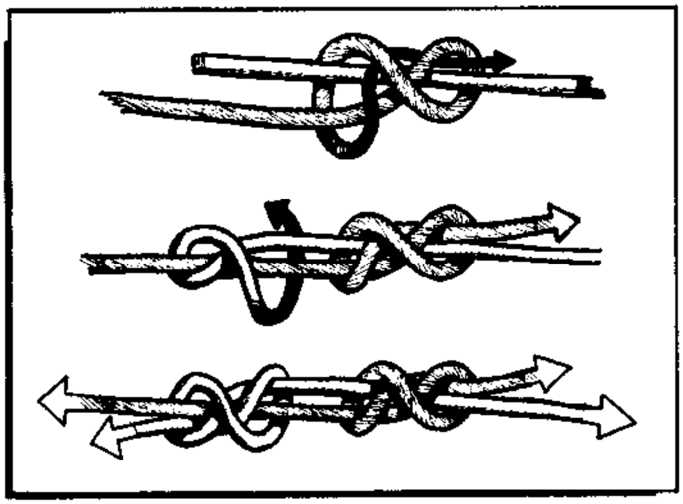

Figure 22-5. Double Fisherman's Bend.

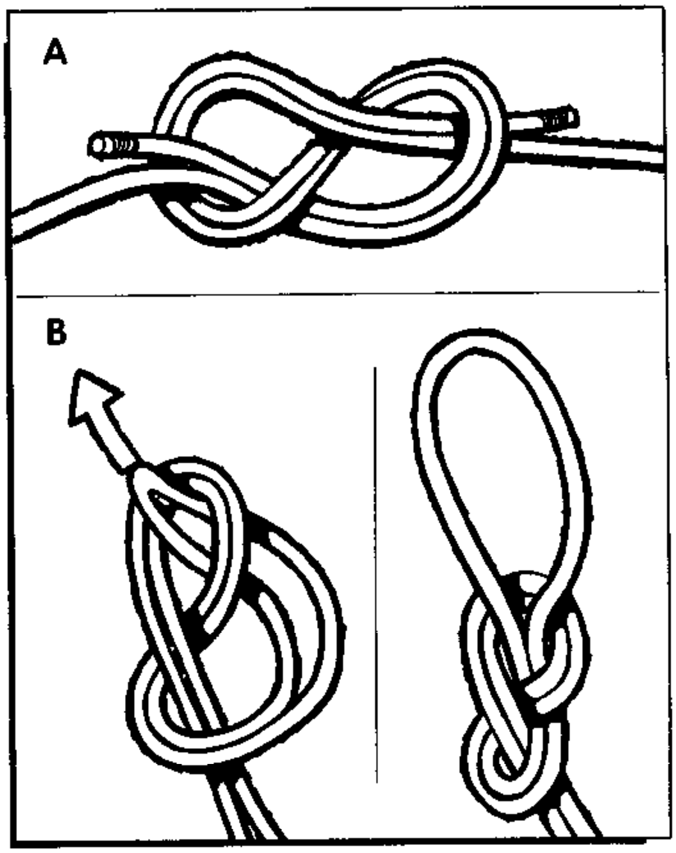

Figure 22-6. Figure-Eight Loop.

d. The butterfly is used to make a fixed loop in the middle of a line where the direction is between 120- to 180-degree angles (figure 22-8). For angles less than 120 degrees, a figure eight in a loop will suffice. For an angle of pull greater than 120 degrees, the figure eight will become weakened and begin to split.

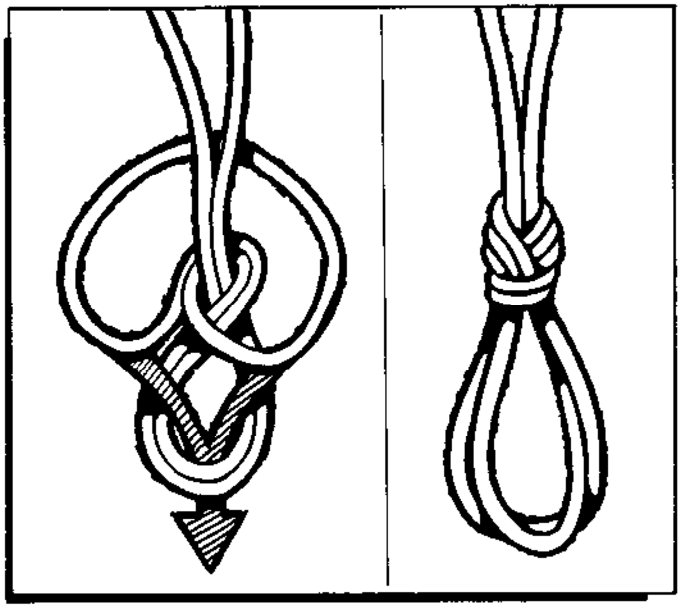

Figure 22-7. Figure-Eight with Two Loops.

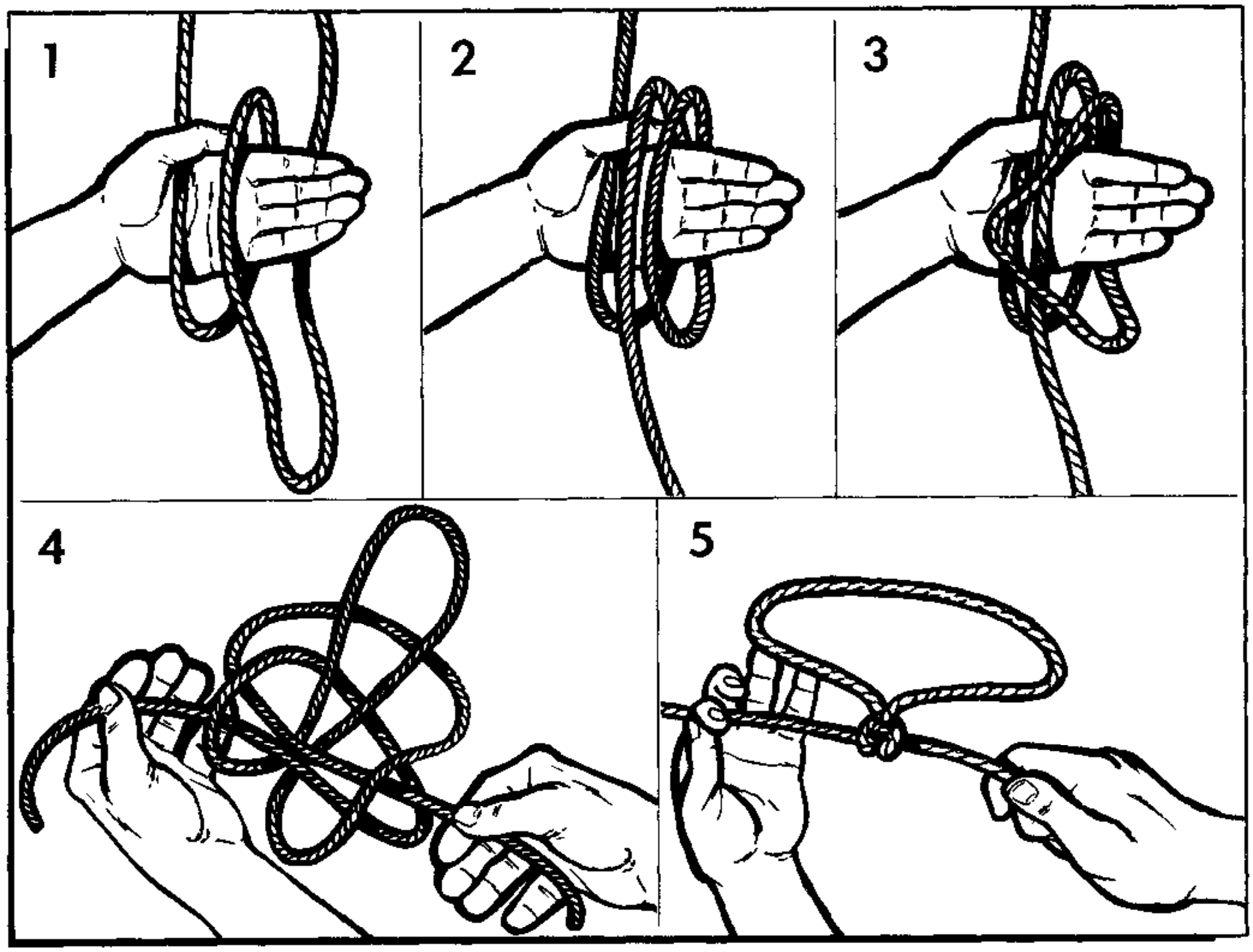

Figure 22-8. Butterfly Knot.

e. The prusik knot may be used to ascend a fixed line or safety a rappel (figure 22-9).

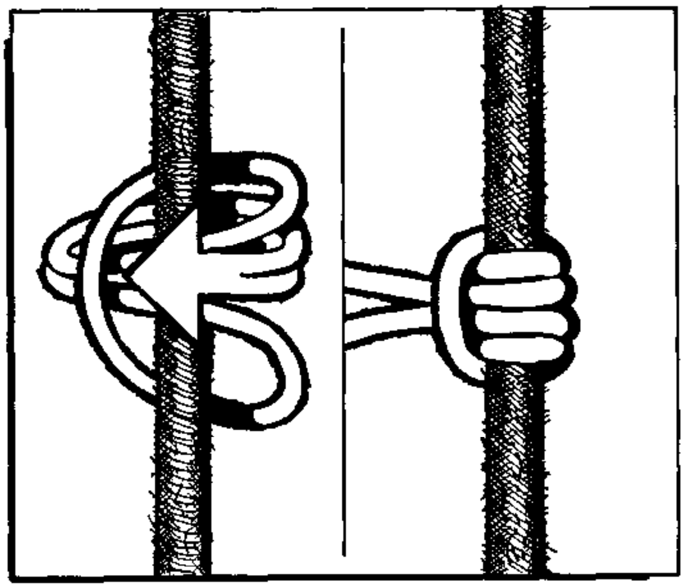

Figure 22-9. Prusik Knot.

f. A three-loop bowline is a variation of the bowline on a bight. It is used for three anchor points or as an improvised harness (figure 22-10).

g. The mariner's knot is a combination of two knots tied with a sling and a carabiner (figure 22-11) and used

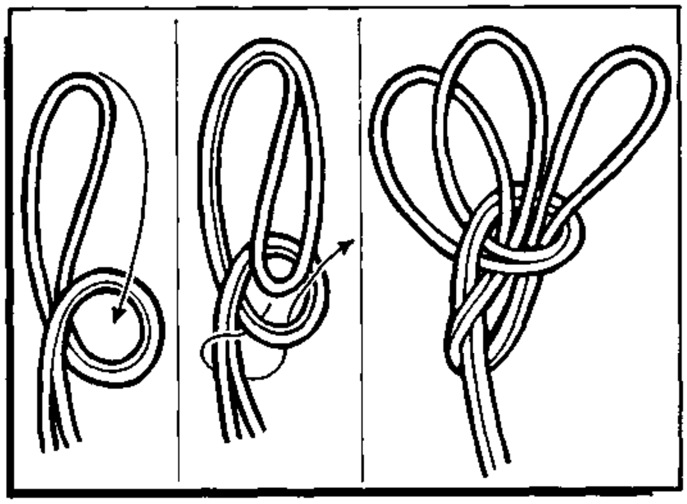

Figure 22-10. Three-Loop Bowline.

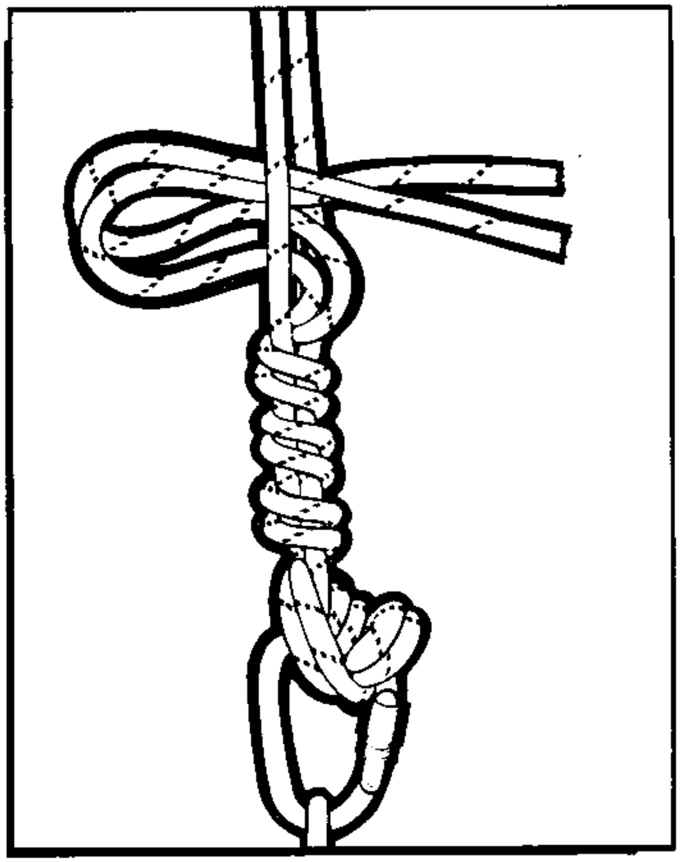

Figure 22-11. Mariner's Knot.

to transfer loads from one system to another. With a sling, a prusik is placed on the main rope (load rope). The remaining portion of the sling is then wrapped three times through the anchor carabiner and then three times around itself next to the carabiner. The remaining tail is tucked between the two sides of the sling. Tension may be gradually released by removing the wraps.

22-5. Special Equipment. Without proper equipment, safety is jeopardized and travel is impossible in severe terrain.

a. Seat Harness. The seat harness is a safety sling which is used to attach the rope to the climber or rappeller. It must be tied correctly for safety and comfort reasons. An improvised seat harness can be made of 1-inch tubular nylon tape. The tape is placed across the back so the midpoint (center) is on the hip opposite the hand that will be used for braking during belaying or rappelling. Keep the midpoint on the appropriate hip, cross the ends of the tape in front of the body, and tie half of a surgeon's knot (three or four overhand wraps) where the tapes cross (figure 22-12). The ends of the tape are brought between the legs (front to rear), around the legs, and then secured with a jam hitch to tape around the waist on both sides. The tapes are tightened by pulling down on the running ends of the tape. This must be done to prevent the tape from crossing between the legs (figure 22-12). Bring both ends around to the front and across the tape again. Then bring the tape to the opposite side of the intended brake hand and tie a square knot with an overhand knot or two half-hitch safety knots on either side of the square knot. The safety knots should be passed around as much of the tape as possible (figure 22-12). Once the seat harness has been properly tied, attach a single locking carabiner to the harness by clipping all of the web around the waist and the web of the half surgeon knot together. The gate of the carabiner should open on top and away from the climber.

b. Commercial Seat Harness. A commercial seat harness is a climbing harness which is worn around the pelvic girdle and is used with climbing ropes for descending and ascending devices and for rescue evacuations. The harness is easily donned, comfortable, allows freedom of movement, and evenly distributes the body

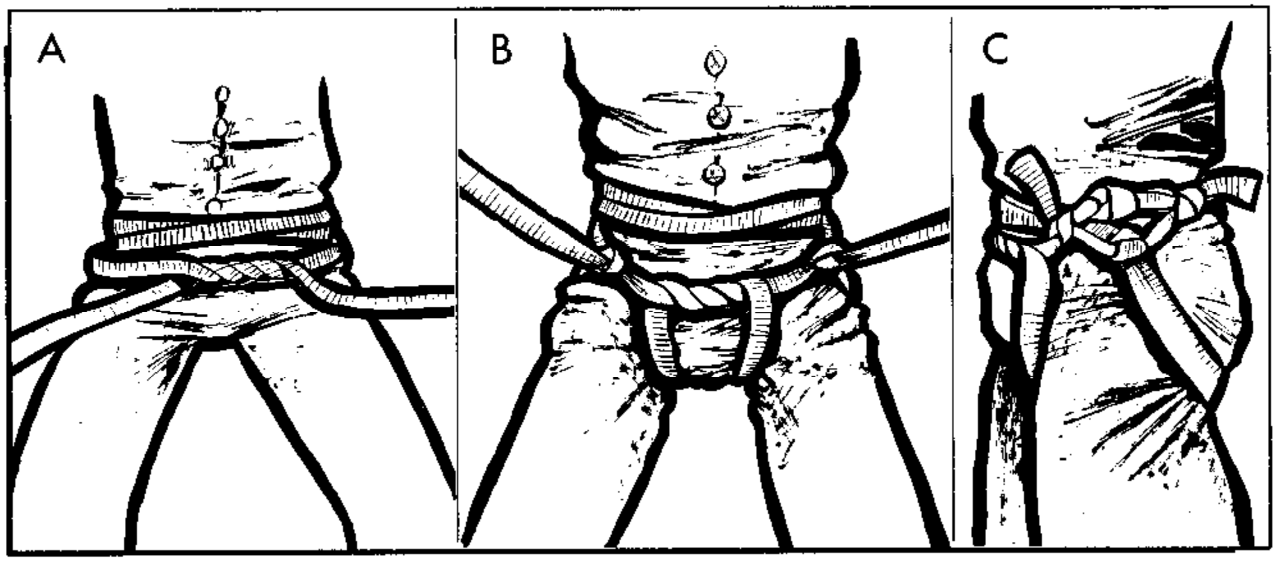

Figure 22-12. Improvised Seat Harness.

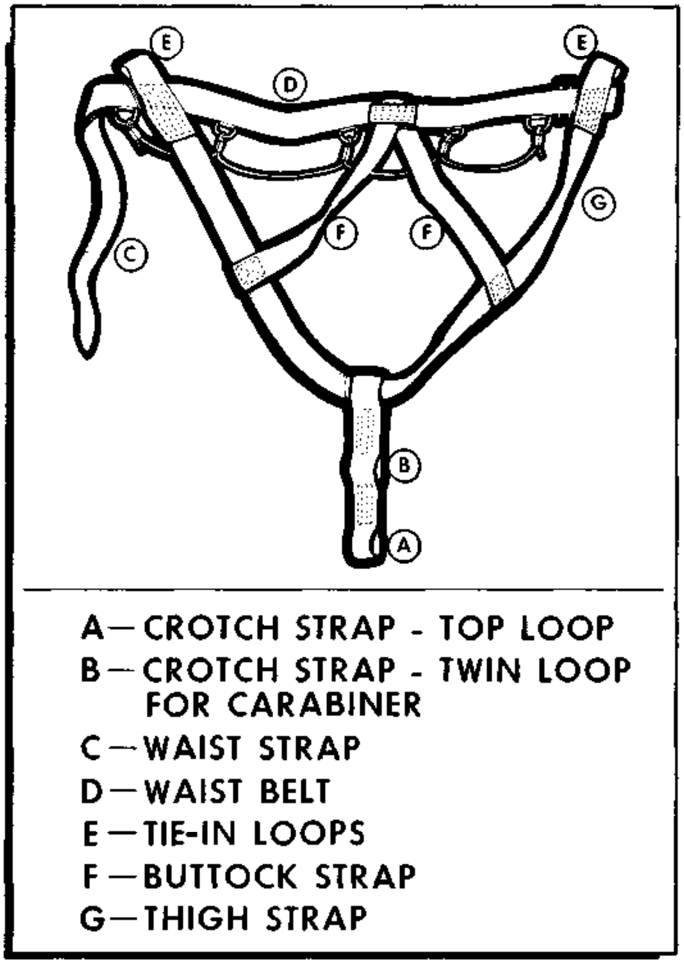

Figure 22-13. Sit Harness.

weight to the pelvic region. Figure 22-13 identifies the parts of the harness. The seat harness webbing should be cared for in the same manner as all webbing. The plastic buckle should not be subject to direct jars since it may fracture or establish a weak point on the buckle which can fail when used. It is primarily constructed of nylon webbing with box stitching at juncture points. Small metal rings are sewn on the lower half of the waist belt and connected with 300-pound line for equipment storage. The harness should have tie-in loops. Webbing should not have tears, rips, burns, or have been subjected to chemical exposure. The plastic buckle should not have fracture marks. All seams and thread should be intact. Harnesses showing signs of excessive wear, tear, rips, burns, or exposure to chemicals should be removed from service. Fractured buckles should be replaced.

c. Improvised Chest Harness. The Parisian bandolier or chest harness is a secondary safety sling used to attach to a safety sling at the top of climbs or for top belaying on high-angle rappels. A Parisian bandolier is made from a continuous loop of webbing. The loop should be 3 to 4 feet long. The knot connecting the two ends of tape together is a water knot with an overhand knot or half hitch as safeties. To don the bandolier, place one arm through the loop and bring the running end of the loop behind the back and under the opposite arm. A sheet bend is tied in the center of the chest by using the center of the tape which was passed under the climber's arm and inserted through the portion of bight formed by the tape that goes over the shoulder (figure 22-14).

d. Climbing Helmet. Climbers should wear this hard shell helmet which is designed to minimize injury dur-

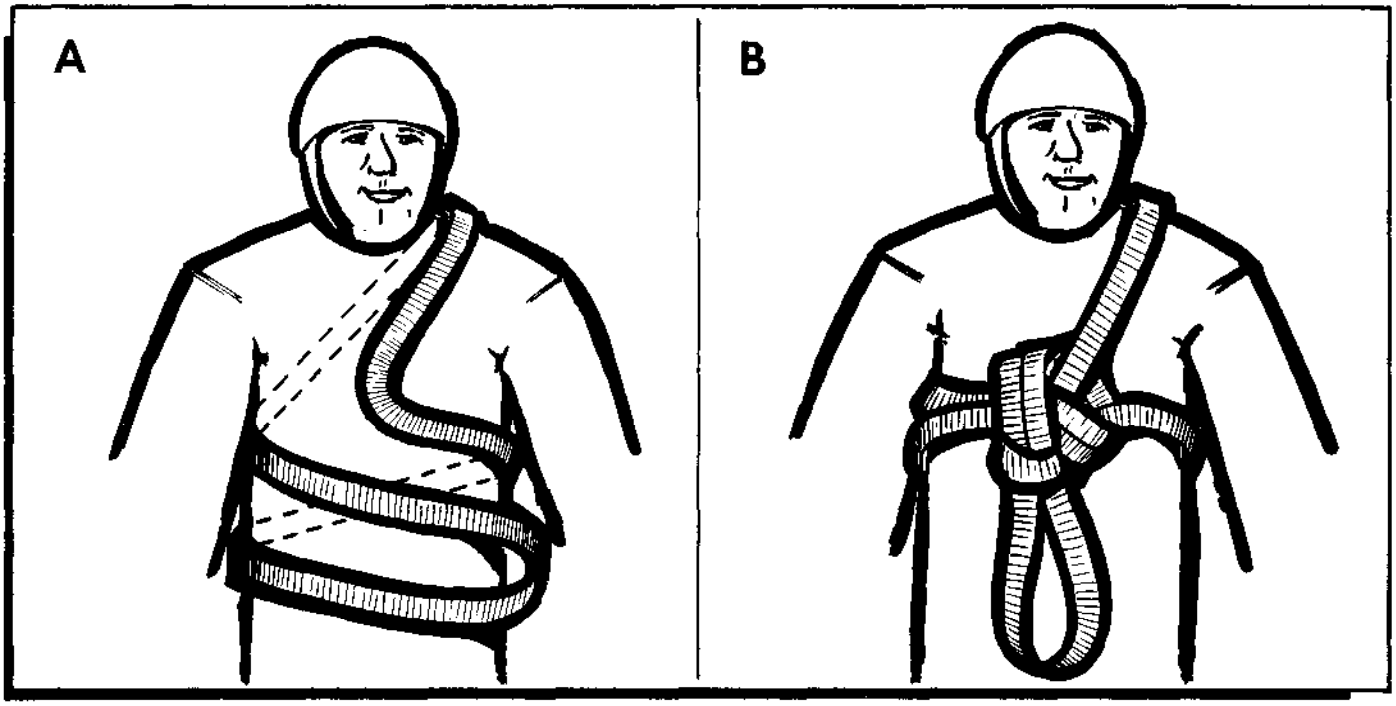

Figure 22-14. Parisian Bandolier.

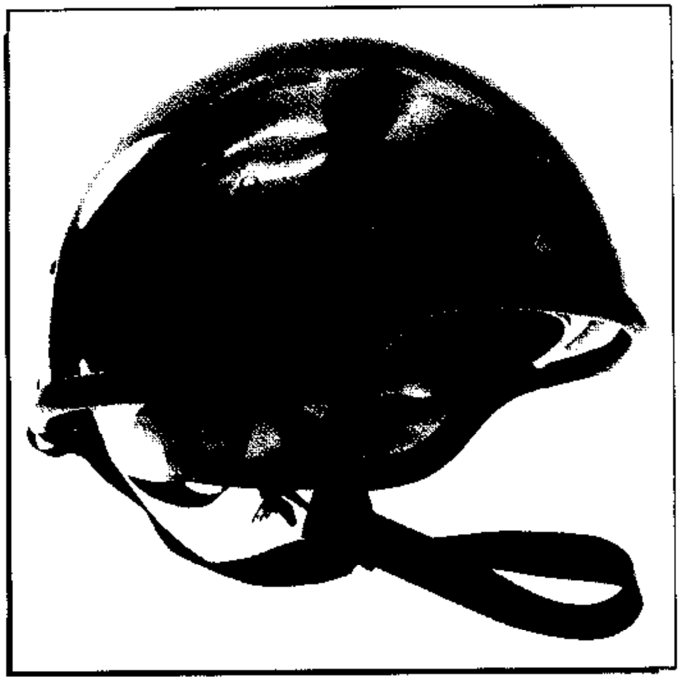

Figure 22-15. Climbing Helmet.

ing falls or when struck by a small object falling from above their position. The helmet consists of a hard shell which is held away from the head by a suspension system. The suspension system absorbs a portion of the blow to the top of the helmet. Helmet fit should be such that the side of the head is protected. A "Y" style strap system of lightweight webbing retains the helmet on the head better than a single strap. The headband is adjustable to allow donning over a watch cap in cold weather (figure 22-15). The hard shell should not have cracks or breaks. The suspension system must be securely riveted in place. The straps must be free of cuts, frays, and securely fastened to the helmet.

e. Gloves. Gloves should be worn when belaying or rappelling. Since these techniques are performed frequently, the gloves should be attached to the climber (rescuer) at all times.

f. Slings. Slings have many uses in adverse terrain work—construction of anchor systems, stirrups to aid movement, attachment to chocks, improvised harness-

SIZE	MOUNTAIN-LAY	KERNMANTLE	NYLON WEB
5/16" OR 7 MM WIDE	2700	2300	
3/8" OR 9 MM WIDE	3700	3200	
7/16" OR 11 MM	5500	4800	
1/2" WIDE			1100
9/16" WIDE			1700
1" WIDE			4000

Figure 22-16. Breaking Strength of Rope.

CARABINER BREAKING STRENGTH		
TYPE	MINIMUM	MAXIMUM
STANDARD OVAL, ALLOY	3890	4210
STANDARD D, ALLOY	4065	5515
LOCKING D, ALLOY	4960	6310
LOCKING D, STEEL	GREATER THAN 11,000 POUNDS	

Figure 22-17. Breaking Strength of Carabiners.

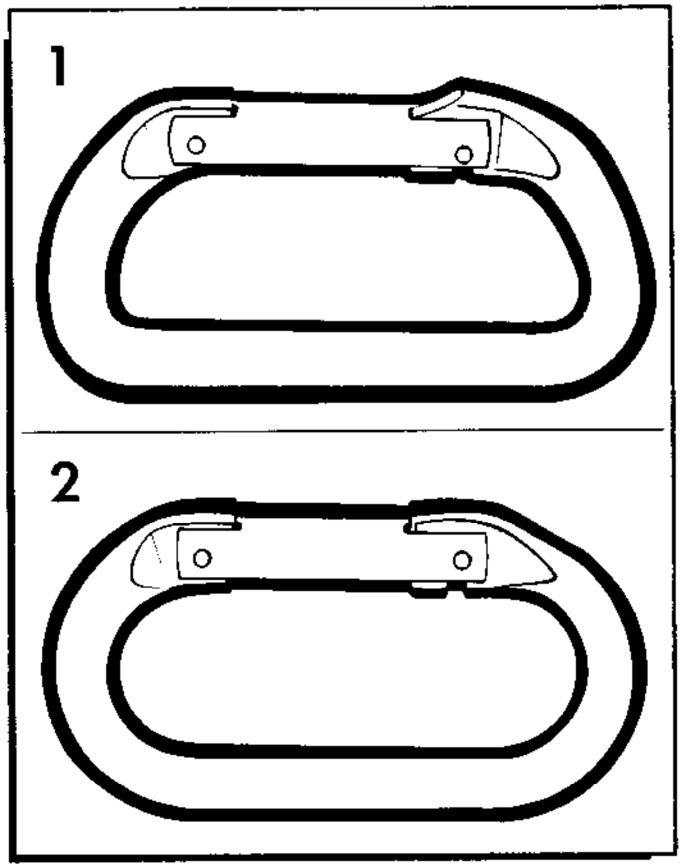

Figure 22-18. Types of Carabiners.

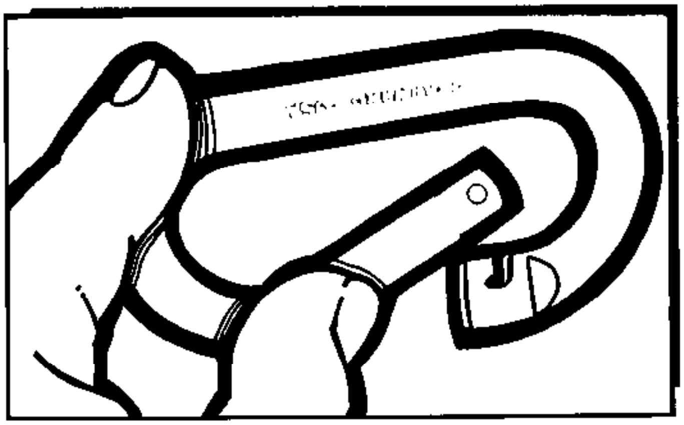

Figure 22-19. Nonlocking Carabiners.

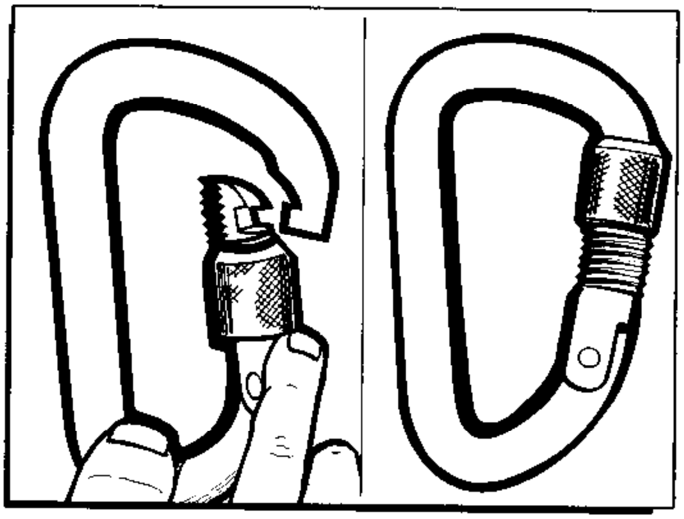

Figure 22-20. Large Locking Carabiner.

es, and to attach equipment to rope systems. Slings may be constructed from Mountain-Lay, Kernmantle, and flat or tubular nylon. The standard length for long slings is 13 feet with smaller slings cut to the individual task. For climbing, slings about 4 feet long, giving a circumference of 2 feet, are acceptable. Sling rope diameters for Mountain-Lay or Kernmantle are 5 mm, 7 mm, and 9 mm while tape slings are usually one-half inch or 1 inch wide. During construction of any system or slinging of chocks, the largest possible rope diameter or tape width should be used. Mountain-Lay and Kernmantle ropes used for slings are the same as climbing ropes. Nylon webbing is constructed of multiple strands of nylon sewn in a weave pattern throughout the length. It should be constructed in flat or tubular shapes. Specific strengths are noted in figure 22-16. Nylon webbing which is burned, torn, cut, or frayed should be removed from service. If nylon webbing is suspected to be inferior due to shock, loading, or abuse, it should not be used.

g. Carabiners. The carabiners used by climbers (rescuers) fall into two material designs—aluminum alloy and chrome vanadium steel. Both material designs may be either locking or nonlocking. Alloy metals provide the maximum strength for adverse terrain work. As with ropes, carabiners are rated differently depending on the manufacturer. Figure 22-17 depicts the breaking strengths for the various types of carabiners. Aluminum alloy and chrome vanadium steel carabiners will be oval-shaped or D-shaped (figure 22-18).

(1) Alloy standard oval (nonlocking) is shaped in an oval with a gate to allow access into the center of the oval. The gate contains a locking pin which fits into the locking notch when the gate is closed (figure 22-19). The alloy D-shape (nonlocking) is shaped in the fashion of a D which allows greater distribution of weight applied to the longer side. Features and operation are the same as the standard oval.

(2) Alloy locking D or oval are both constructed the same as previously discussed but are machined with threads on the gate and a sleeve which, when screwed clockwise, will cover the locking notch and pin for a positive lock. The steel locking D is the same construction as the alloy locking D, but is made of steel and is normally used for mountain rescue work (figure 22-20).

(3) Carabiners should not be dropped or used for other than the designed purpose since small fracture lines may develop and weaken the structure. Carabiners should not be used as a hammer nor loaded (stressed)

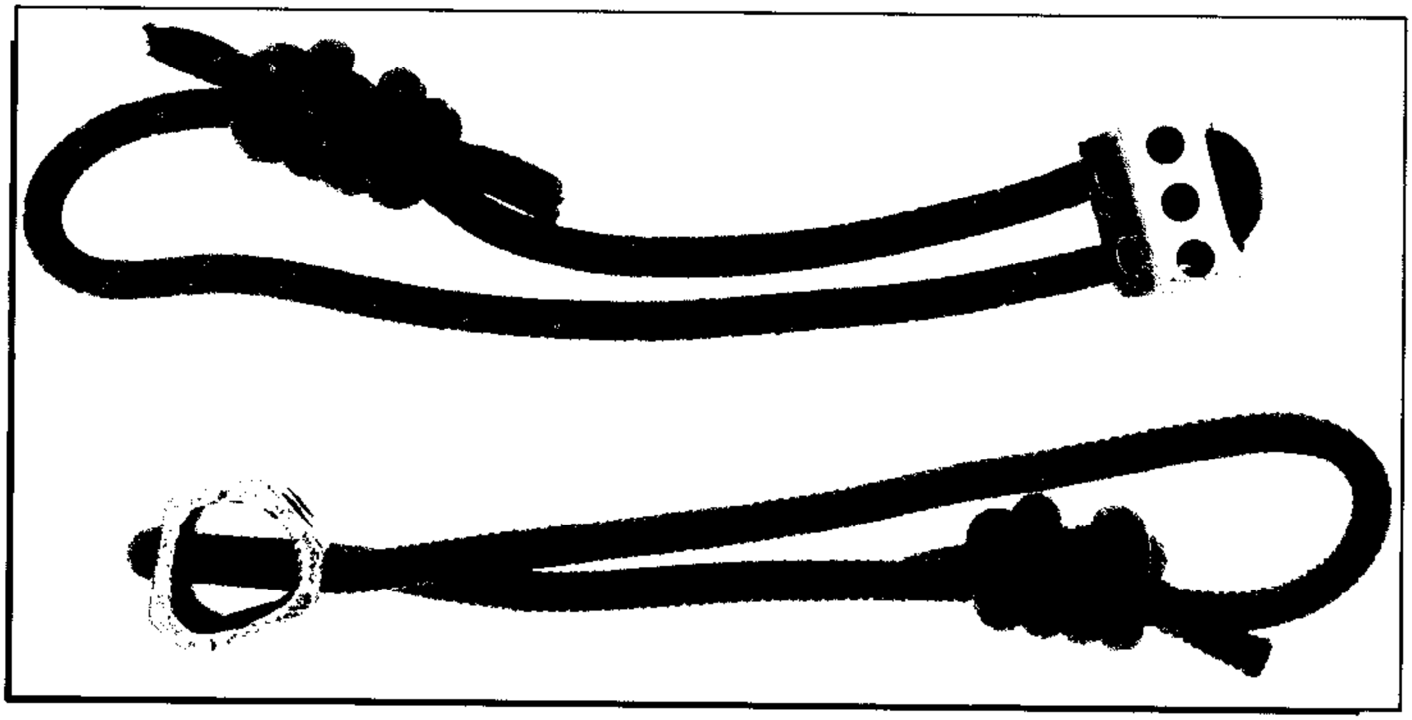

Figure 22-21. Hexcentric Chock with Kermantle Sling.

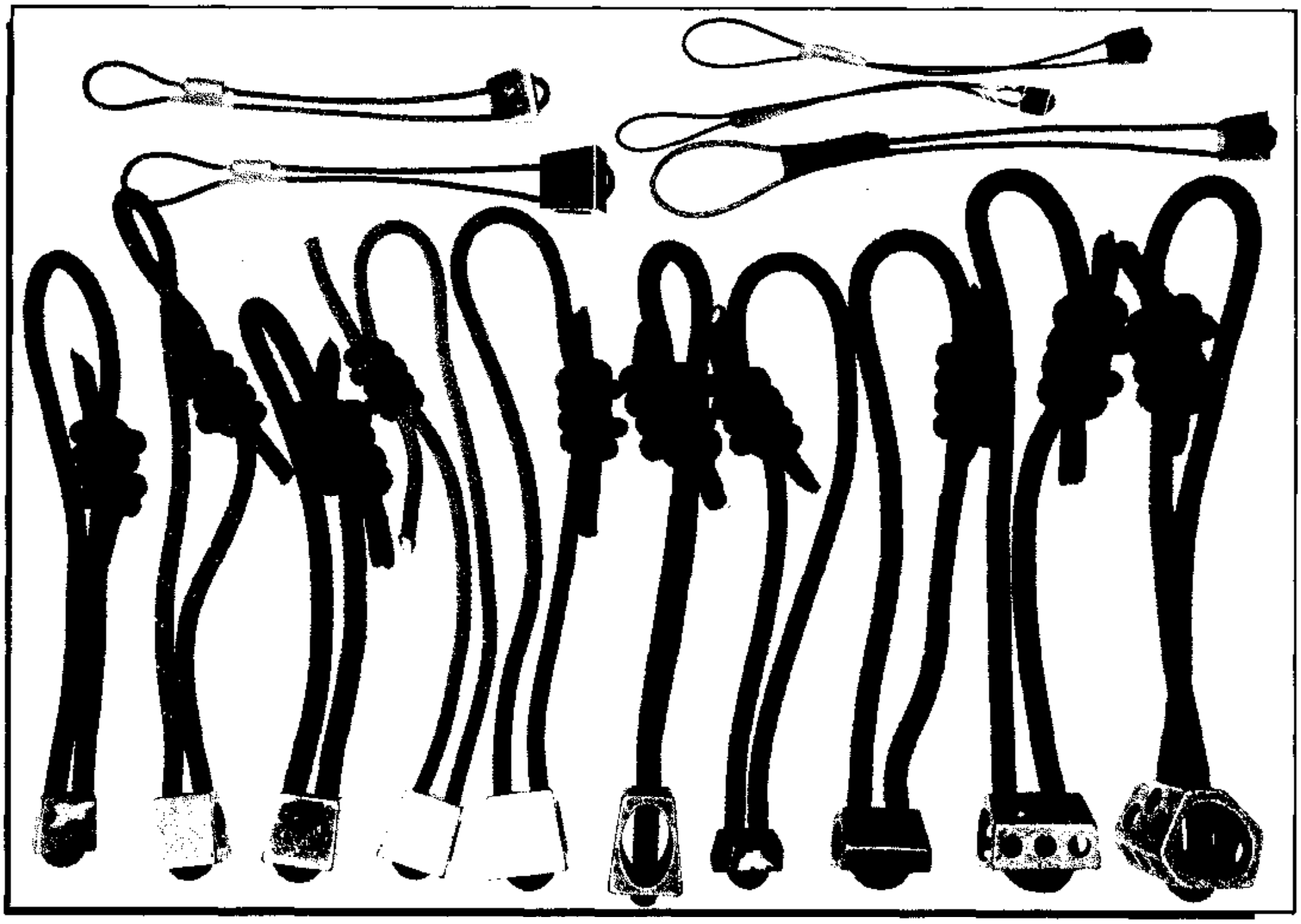

Figure 22-22. Chocks.

beyond their maximum breaking strength. The moving parts, hinge and sleeve, of locking carabiners should be kept clean for free movement. If a carabiner "binds," do not oil it — discard it! Carabiners should not be filed, stamped, or marked with an engraving tool. Colored tape or Teflon paint may be used to identify carabiners. All moving parts (gate, locking sleeve) should operate freely and the locking pin must properly align with the locking notch. Obvious fractures, regardless of size, are cause for condemning a carabiner.

h. Chocks:

(1) Chocks are metal alloy or copper shapes with unequal sides which are placed within cracks in rocks to serve as anchor points or parts of a protection system. Figures 22-21 and 22-22 show the common chocks used. Each is designed to fit within a variety of cracks in rocks. Sizes vary from one-sixteeth inch thick by one-fourth inch wide to several inches thick and wide. The various shapes with uneven sides allow one size to fit many rock openings. Carrying different sizes of chocks allows a person to choose the most suitable chock for the job. Depending on the style, chocks are constructed differently and require a sling for use.

(a) Copperhead. A cylindrical copper chock ranging from three-sixteeths inch in diameter by one-half inch long to one-half inch in diameter by 1 inch long. The relatively soft head bites well into rock and has

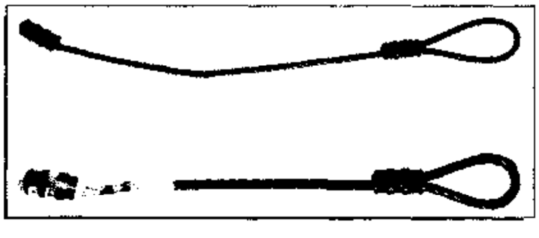

Figure 22-23. Copperheads.

little rotation after placement. It is manufactured with a cable sling attached (figure 22-23).

(b) Hexentric. A metal alloy chock constructed with six sides. Each side is a different length than the opposing side (figure 22-21). The two ends are tapered, gradually getting smaller from the back to the front of the chock. The front of the hexentric chock is the narrowest of the faces. Small hexentric chocks are manufactured with wire slings and larger ones have two holes bored through the front to back for threading of a Kernmantle sling. Kernmantle slings vary with the size of the chock, from 5 mm to 9 mm. The largest sling size possible will be used with each chock.

(c) Wedge. A solid alloy metal chock shaped in the form of a wedge. All four sides decrease in size from back to front. As with small hexentrics, small wedges are manufactured with wire slings while larger wedges require Kernmantle slings.

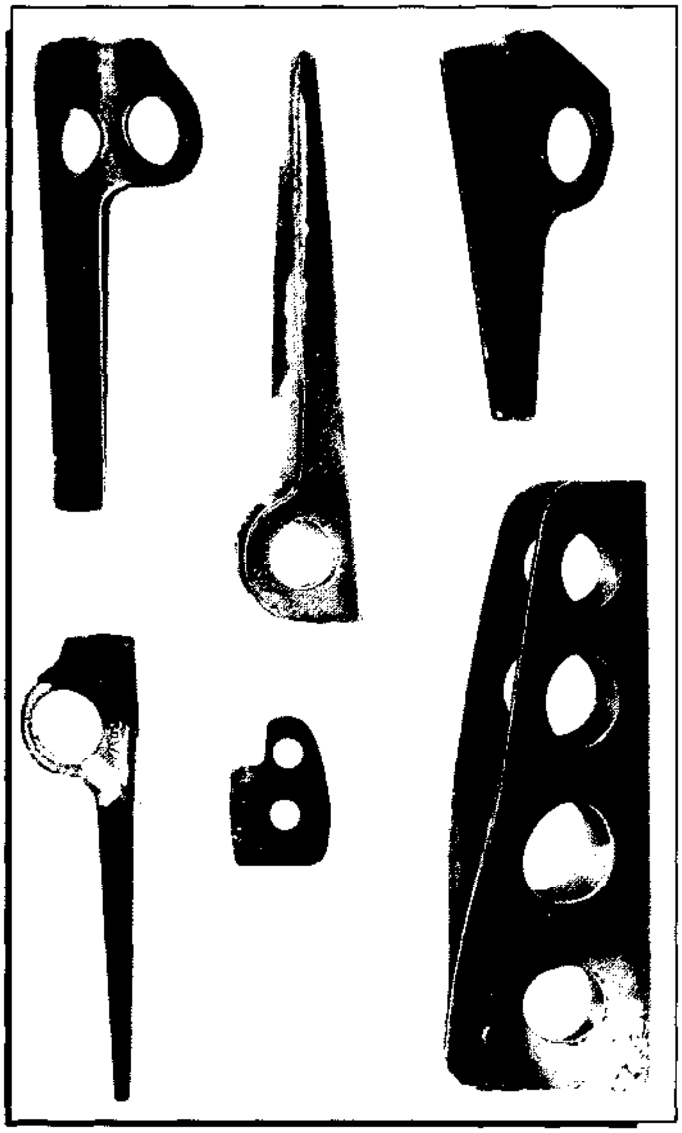

Figure 22-24. Pitons.

(2) All chocks should be treated the same as carabiners. However, chocks will take considerable abuse since they are wedged in rock, and at times, the leverage of a hammer may be required to remove them. Chocks manufactured with wire slings and wedges should be closely monitored for security. If there is doubt about the condition of a wedge, it must not be used. Rope or tape slings on chocks must meet inspection requirements. Cracks in any surface of the chock is cause to remove it from service. Small scratches or grooves on the outer surfaces are not cause for removal from service.

i. Pitons. Pitons are metal alloy spikes of various shapes, widths, and lengths which are hammered into cracks in rock surfaces for anchor points or as part of a protection system. The different shapes are identified by name; each name has the same shape but comes in various sizes in length and width. Figure 22-24 depicts the types of pitons. Pitons are divided into two different categories—blades, whose holding power results from being wedged into cracks; and angles, which hold by wedging and blade compression.

(1) Blades comprise those pitons which have flat surfaces and range in size from knife-blade thickness to one-half inch.

(2) Angles are those which have rounded or V-shaped blades and allow the sharp edges to cut into the edge of a crack. Sizes vary from a short "shallow angle" of one-half inch thick and 2½ inches long to "bongs" (large angles of 4 inches in width). The leeper, a special Z-shaped angle, is designed to give greater holding power due to its cutting ability and blade compression.

(3) Pitons are made of chrome-molybdenum alloy metal which has a high strength versus weight ratio. Multiple holes further reduce the weight but do not compromise their strength. Pitons are virtually indestructible. However, they should only be used for the designed purpose. Damage may occur during the placement of a piton by ineffective or off-center hammering. Pitons must not have cracks or be bent from the shape of manufacture. Any piton which is suspected of cracks or bends should be removed from service.

j. Figure-Eight Clog. A mechanical friction device used for belaying, rappelling, and breaking while lowering personnel and equipment. The device is formed in a figure-eight (figure 22-25) with two different size openings comprising the inner holes. The rope is passed through the larger hole of the "8" and over the connecting portion of the device. The smaller hole is attached to a carabiner. The figure-eight clog is manufactured in various sizes.

k. Protection System (Belay System). This system is commonly referred to as a belay, securing of a climber tied to the end of a rope by a stationary second climber (figure 22-26A and 22-26B).

(1) The major components of the belay system are (figure 22-27):

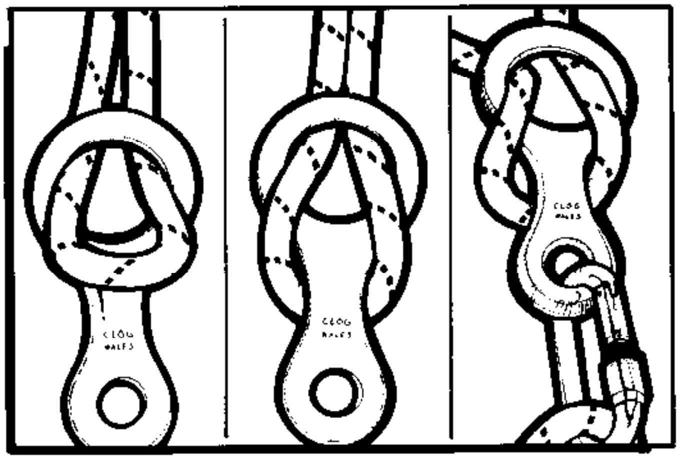
CLOG WALES
CLOG WALES
CLOG WALES

Figure 22-25. Figure-Eight Clog Direct Belay Installation.

(a) An anchor—secure point to which the belayer is attached.

(b) Belayer—individual responsible for the security of the climber.

(c) Intermediate protection point(s)—placement of slings, chocks, or pitons along the route of climb by the lead climber. The climbing rope is threaded through carabiners attached to the devices and the climbing rope clipped into the carabiner.

(d) Climbing rope size—11 mm Kernmantle or seven-sixteenths inch Goldline.

(2) The minimum equipment for a belay system is divided between rescue team and personal equipment. Rescue team equipment is comprised of the items necessary for the climbers to reach the objective. These are the climbing rope and climbing hardware, including chocks, pitons, slings, and carabiners. Individual equipment is comprised of items which allow each climber to perform belaying and climb. These equipment items are climbing helmet, seat harness, climbing boots, and gloves (worn while belaying or rappelling only).

22-6. The Successful Climb. The successful accomplishment of a climb is based on the strict application of basic principles and techniques.

a. Route Selection. Route selection can be the deciding factor in planning a climb. A direct line is seldom the proper route from a given point to the area of the survivor. Time spent at the beginning of the climbing operation in proper route selection may save a large amount of time once the operation has started. The entire route must be planned before it is carried out, with the safest route selected. Natural hazards present,

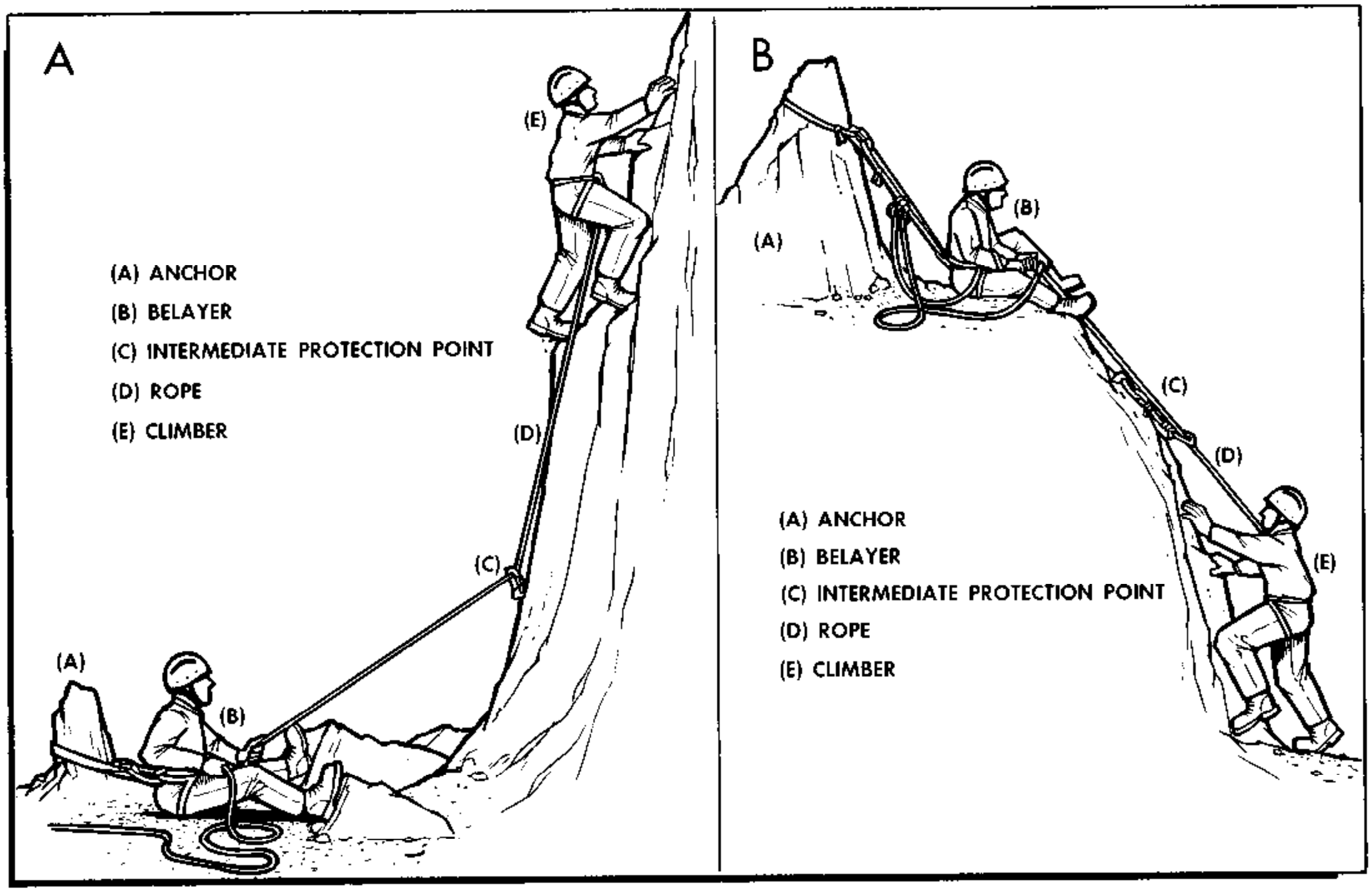
A
(E)
(A) ANCHOR
(B) BELAYER
(C) INTERMEDIATE PROTECTION POINT
(D) ROPE
(E) CLIMBER
(D)
(C)
(A)
(B)
B
(B)
(A)
(C)
(D)
(A) ANCHOR
(B) BELAYER
(C) INTERMEDIATE PROTECTION POINT
(D) ROPE
(E) CLIMBER
(E)

Figure 22-26. Belay System (A) Bottom (B) Top.

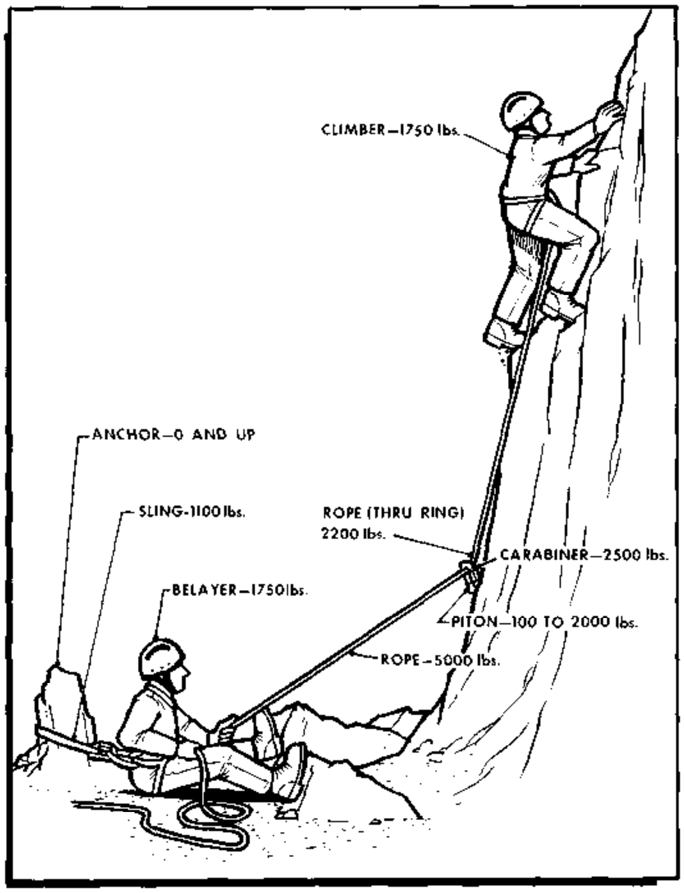
CLIMBER—1750 lbs.
ANCHOR—0 AND UP
SLING-1100 lbs.
ROPE (THRU RING) 2200 lbs.
CARABINER—2500 lbs.
BELAYER—1750lbs.
PITON—100 TO 2000 lbs.
ROPE—5000 lbs.

Figure 22-27. Belay Chain.

retreat routes available, time involved to perform the climb, and logistics will be major influencing factors in selecting the route.

b. Lead Climber Techniques. The lead climber is responsible for following the preplanned route of climb and altering the route as necessary. If the route of climb must be altered, consideration must be given to the experience of the climber(s). Climbing moves which are considered easy for the leader may be difficult for the second climber. Additionally, lead climbers must provide for their own personal protection during climbs to minimize the distance of a possible fall. Each intermediate protection point placed must be secure and follow as direct a line as possible.

c. Protection Placement. The basic principle to placement of protection is to decrease the distance of a fall. If the protection is not placed at frequent intervals, the climber's descent during a fall will equal twice the distance from the climber to the belayer. Slack in the belay system and rope stretch will slightly increase the distance. If a leader climbed 90 feet above the belayer without adding protection, the resultant fall will be 180 feet plus slack and stretch (figure 22-28). The belayer is helpless to stop the leader's fall in this situation. However, providing protection at intervals (figure 22-28) of 15, 20, and 25 feet, for example, will decrease the distance of a leader's fall to a little over 30 feet, making it easier for the belayer to stop the fall. Preferably, protection should be placed at about 10-foot intervals during all climbing operations to minimize the distance of a fall.

(1) The leader can reduce the drag on the rope by climbing in as straight a line as possible through protection points. The zigzagging of the climbing rope through protection points widely spaced or at abrupt angles (figure 22-29) will increase rope drag. Ideally, when points of protection are separated, a sling should be added to the anchor to keep the climbing rope in a straight line (figure 22-29). The attachment of a sling to an intermediate protection device is called a runner.

(2) Chocks on wire should always have a runner placed on the wire as shown in figure 22-30. The runner reduces the direct rope movement of the wired chock, thereby reducing possible dislodgment. For all wired and roped chocks, runners should be clipped into a carabiner at the chock, and the climbing rope clipped into an additional carabiner.

(3) Pitons may also be extended in a similar manner to that used for chocks. If sufficient runners are not available on long climbs, a chock sling may be used. The climbing leader must correctly thread the climbing rope through the carabiners attached to the piton. The carabiner should open either down and out or toward the belayer. Further, the rope running through the carabiner should run from the inside to the outside to prevent binding or the carabiner gate from opening.

d. Leader Belay. The length of the climbing rope or the length of the route climbed will determine when a belay will be established by the leader. If the route is longer than a rope length, a belay must be established by the leader to protect the first belayer's ascent to the leader's position. The belay chain is established as shown in figure 22-31. The only alteration necessary to the belay chain would be if the second climber continued the climb as the leader once the belayer's (first leader's) position is reached. This leap frogging of the leader is referred to as "climbing through". The sequence for the climbing through method begins when the first leader reaches the end of the climbing rope length. The leader selects a suitable belay position when the belayer (second climber) states that approximately 20 feet of rope remain. With a suitable site selected, one that has anchor and an area for a standing or sitting belay, the lead climber belays. The second climber then climbs up to the lead climber. Once reaching the lead climber, the second climber assumes the lead climber's role and continues the climb.

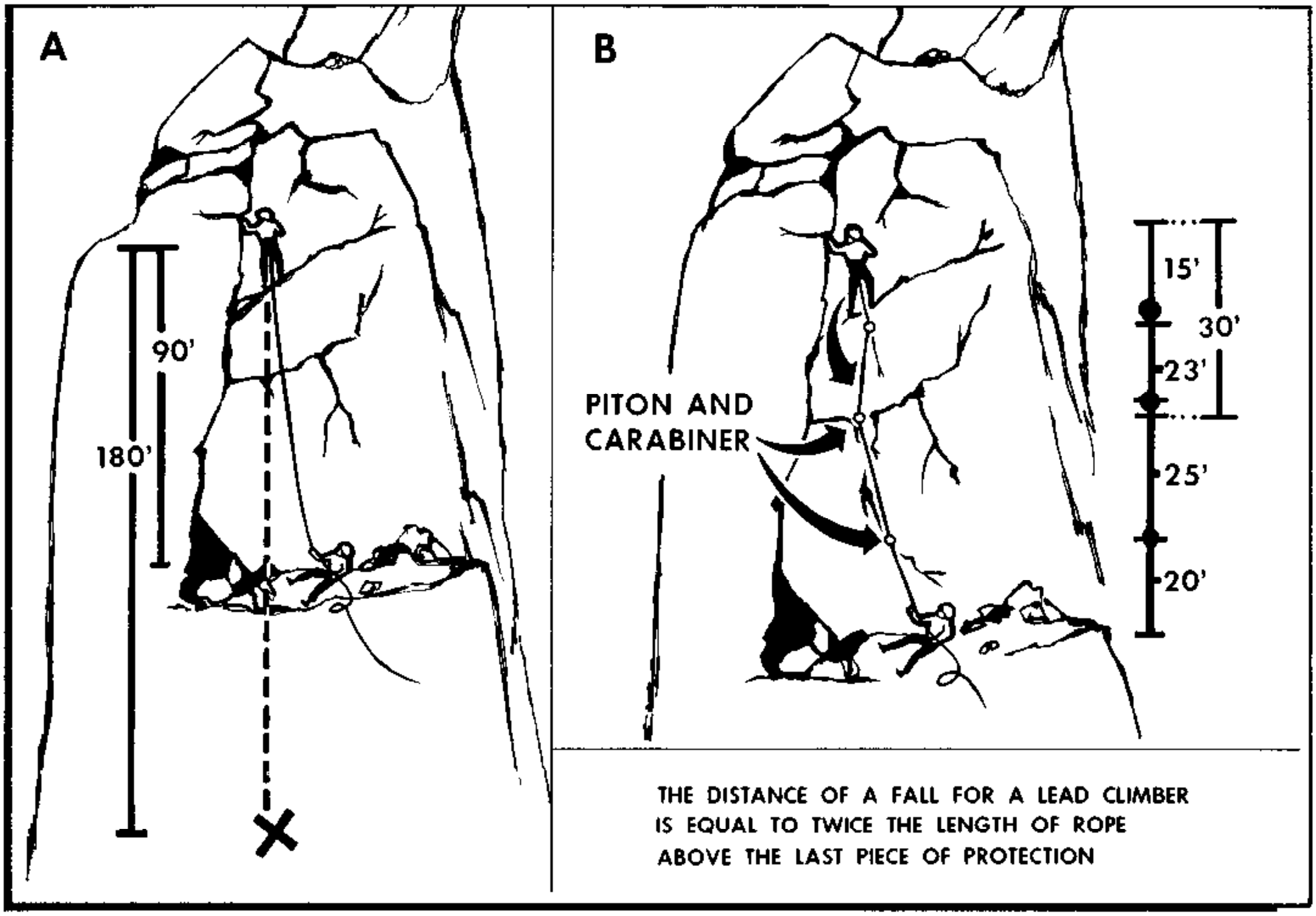
A
90'
180'
B
PITON AND
CARABINER
15'
30'
23'
25'
20'

THE DISTANCE OF A FALL FOR A LEAD CLIMBER IS EQUAL TO TWICE THE LENGTH OF ROPE ABOVE THE LAST PIECE OF PROTECTION

Figure 22-28. Protection Placement.

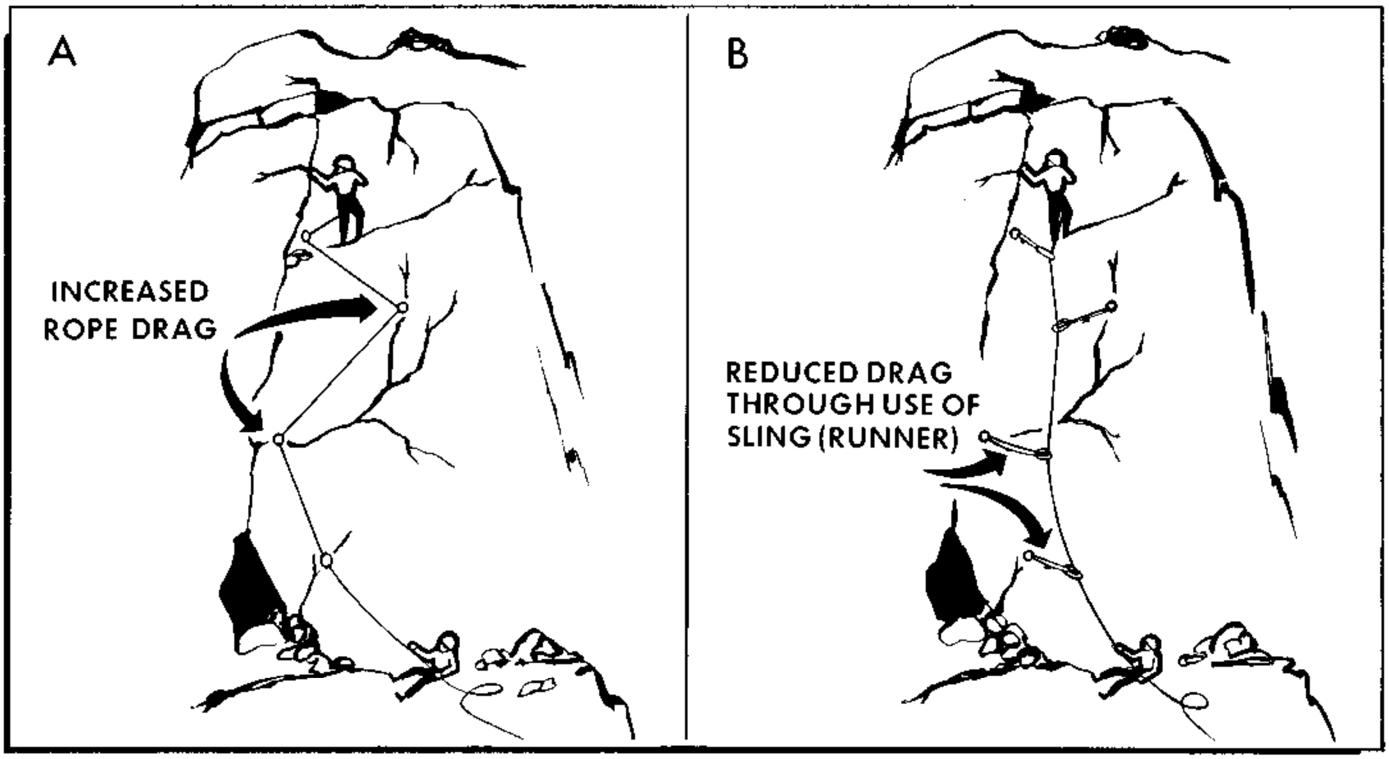
A
INCREASED
ROPE DRAG
B
REDUCED DRAG
THROUGH USE OF
SLING (RUNNER)

Figure 22-29. Direct Placement.

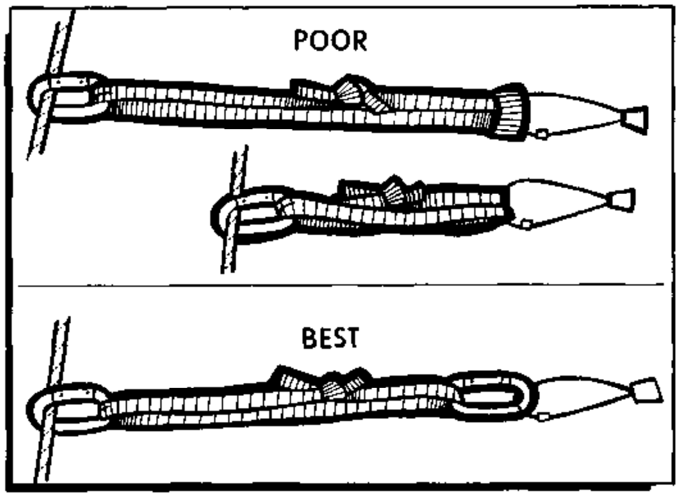

Figure 22-30. Runner Placement on Chocks.

(1) Terrain must be analyzed to find an efficient route of travel. The rescuer (climber) must make a detailed reconnaissance, noting each rock obstacle, the best approach, height, angle, type of rock, difficulty, distance between belay positions, amount of equipment, and number of trained rescuers needed to accomplish the mission on or beyond the rocks. If the strata dips toward the rescuer, holds will be difficult as the slope will be the wrong way. However, strata sloping away from the rescuer and toward the mountain mass provides natural stairs with good holds and ledges.

(2) At least two vantage points should be used so a three-dimensional understanding of the climb can be attained. Use of early morning or late afternoon light, with its longer shadows, is helpful in this respect. Actual ground reconnaissance should be made, if possible.

e. Dangers to Avoid:

(1) On long routes, changing weather will be an important consideration. Wet or icy rock can make an otherwise easy route almost impassable; cold may reduce climbing efficiency; snow may cover holds. A weather forecast should be obtained if possible. Smooth rock slabs are treacherous, especially when wet or iced after freezing rain. Ledges should then be sought. Rocks overgrown with moss, lichens, or grass become treacherous when wet. Under these conditions, cleated boots are by far better than composition soles.

(2) Tufts of grass and small bushes that appear firm may be growing from loosely packed and unanchored soil, all of which may give way if the grass or bush is pulled upon. Grass and bushes should be used only for balance by touch or as push holds—never as pull holds. Gently inclined but smooth slopes of rock may be covered with pebbles that may roll treacherously underfoot.

(3) Ridges can be free of loose rock, but topped with unstable blocks. A route along the side of a ridge just below the top is usually best. Gullies provide the best protection and often the easiest routes, but are more subject to rockfalls. The side of the gully is relatively free from this danger. Climbing slopes of talus, moraines, or other loose rock are not only tiring to the individual but dangerous because of the hazards of rolling rocks to others in the party. Rescuers should close up intervals when climbing simultaneously. In electrical

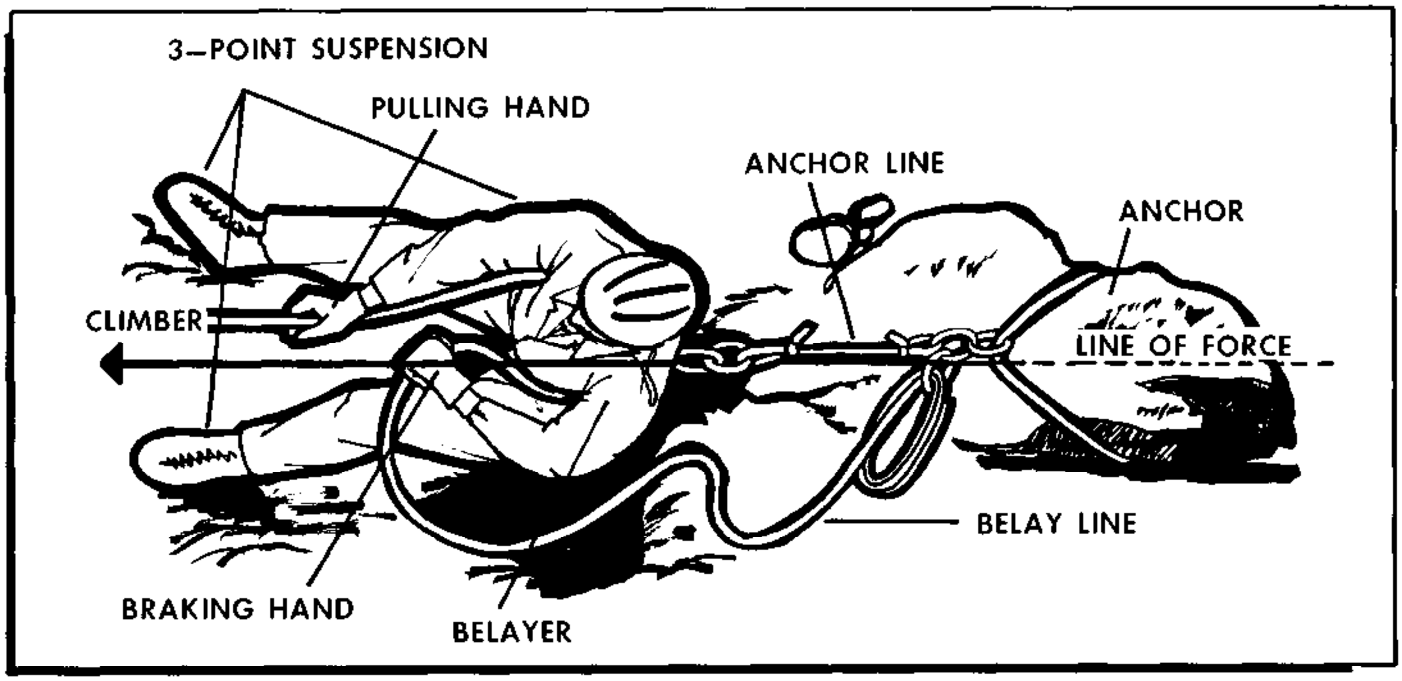

Figure 22-31. Belay Chain.

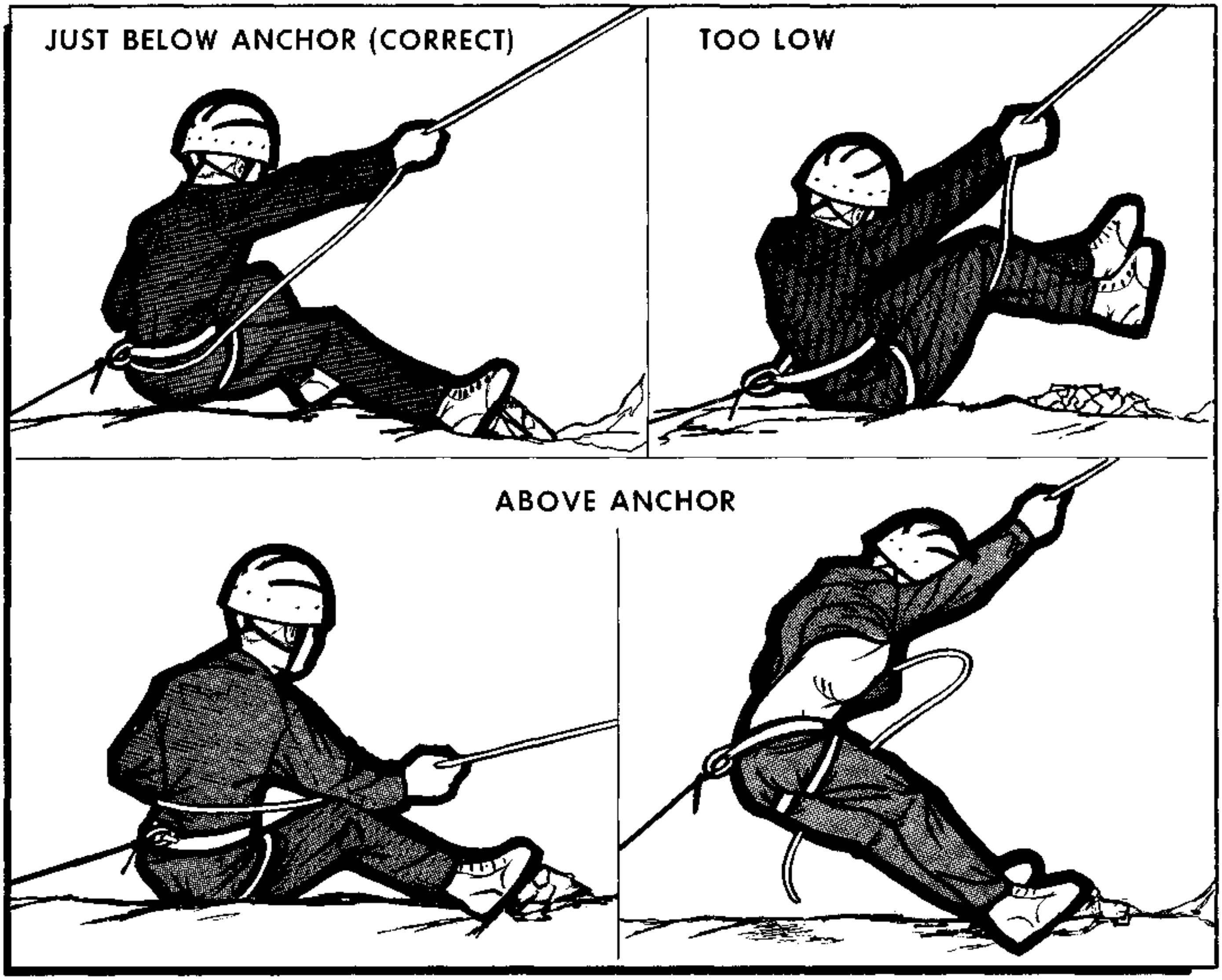

Figure 22-32. Sitting Belay.

storms, lightning can endanger the climber. Peaks, ridges, pinnacles, and lone trees should be avoided.

(4) Rockfalls are the most common mountaineering danger. The most frequent causes of rockfalls are other climbers, heavy rain and extreme temperature changes in high mountains, and resultant splitting action caused by intermittent freezing and thawing. Warning of a rockfall may be the cry "ROCK," a whistling sound, a grating sound, a thunderous crashing, or sparks where the rocks strike at night. A rockfall can be a single rock or a rockslide covering a relatively large area. Rockfalls occur on all steep slopes, particularly in gullies and chutes. Areas of frequent rockfalls may be indicated by fresh scars on the rock walls, fine dust on the talus piles, or lines, grooves, and rock-strewn areas on snow beneath cliffs. Immediate action is to seek cover, if possible. If there is not enough time to avoid the rockfall, the climber should lean into the slope to minimize exposure. Danger from falling rock can be minimized by careful climbing and route selection. The route selected must be commensurate with the ability of the least experienced team member. (NOTE: Yell "rock" when equipment is dropped.)

22-7. Belaying:

a. Belaying provides the safety factor or tension, which enables the party to climb with greater security. Without belaying skill, the use of rope in party climbing is a hazard. When climbing, a climber is belayed from above or below by another rescue team member.

(1) The belayer must run the rope through the guiding hand, which is the hand on the rope running to the climber or rescuer, and around their body to the brake hand, making certain that it will slide readily. The belayer must ensure that the remainder of the rope is laid out so it will run freely through the braking hand.

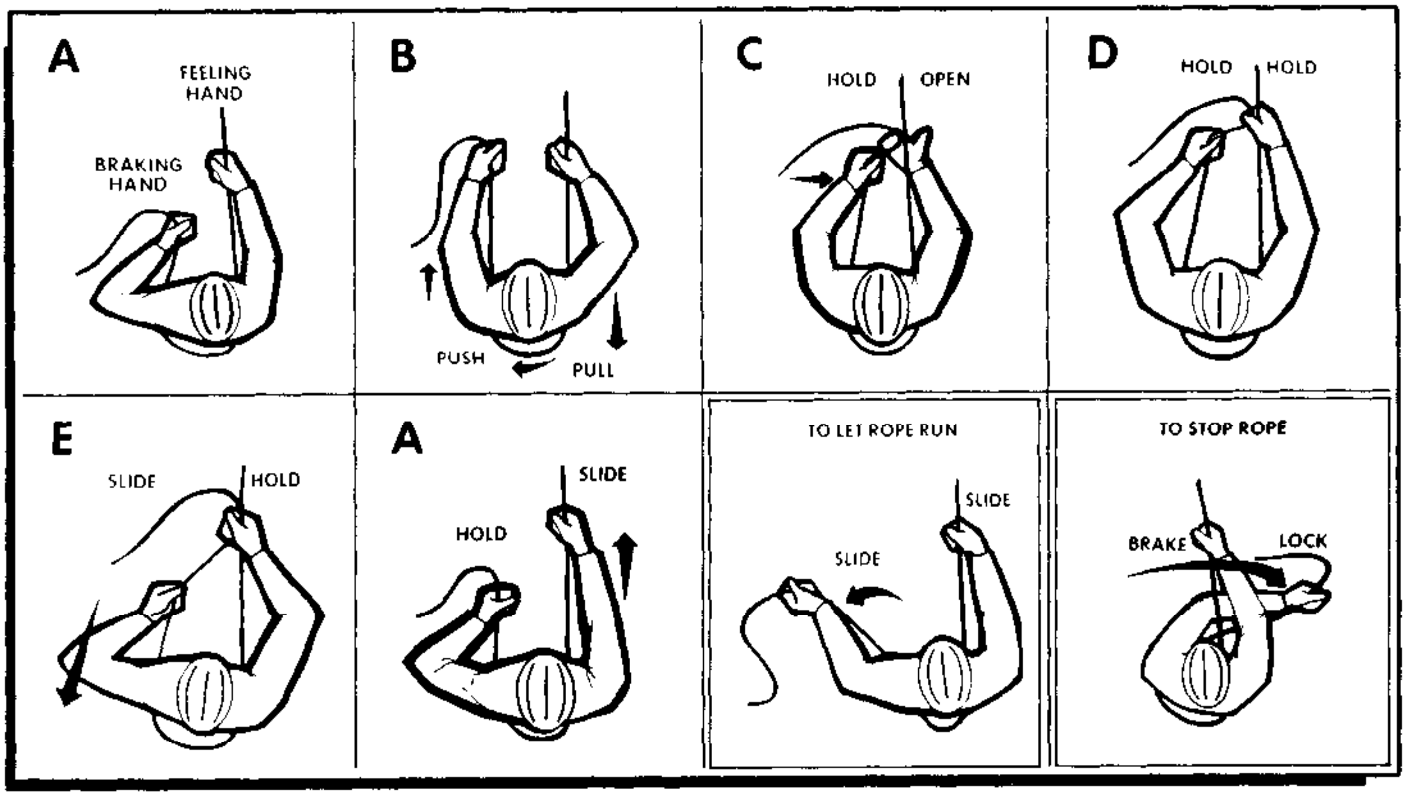

Figure 22-33. The Belay Sequence.

(2) The belayer must constantly be alert to the climber's movements in order to anticipate any needs. Avoid letting too much slack develop in the rope through constant use of the guiding hand. Keep all slack out of the rope leading to the rescuer, thus sensing any movement. If belaying a lead climber, the climber will need a constant flow of slack while climbing. If the rope is fed too slow or fast, the climber must communicate with the belayer to adjust the rate. Avoid taking up slack too suddenly to prevent throwing the climber off balance. When taking up slack, the braking hand is not brought in front of the guiding hand, but just behind the guiding hand. This allows the braking hand to slide back and to remain constantly on the rope. The braking hand is never removed from the rope during a belay.

(3) The belayer should brace well for the expected direction of pull in a fall so the force of the pull will, when possible, pull the belayer more firmly into position. A climber should neither trust nor assume a belay position which has not been personally tested.

b. The sitting belay is normally the most secure and preferred position (figure 22-32).

(1) The belayer sits and attempts to get good triangular bracing position with the legs and buttocks. Legs should be straight when possible, and the guiding hand must be on the side of the better braced leg. The rope should run around the hips. If the belay spot is back from the edge of a cliff, friction of the rope will be greater and will simplify the holding of a fall, but the direction of pull on the belayer will be directly outward. The rope must not pass over sharp edges.

(2) Even with a good belay stance, if the rope is too high or too low on the back of a belayer, the belayer may be unable to hold a falling climber (figure 22-32). With the rope too high on the back of the belayer, on a bottom belay, the rope will ride up in the belayer's armpits. (NOTE: Rope should run under the anchor.) This will pull the belayer forward and off balance if a climber were to fall. If the rope is too low on the back of a belayer, the rope will be pulled under the buttocks of the belayer. (NOTE: Rope should run on top of the anchor.) This will force the belayer to attempt to stop the fall by trying to hold onto the rope.

(3) When necessary, seek a belay position that offers cover from a rockfall.

(4) If the climber falls, the belayer should be able to perform the following movements automatically.

(a) Relax the guiding hand.

(b) Apply immediate braking action. This is done by bringing the braking hand across the chest or in front of the body (figure 22-33).

c. When a falling climber has been brought to a stop, the belayer must hold until the situation is relieved.

There are different ways to do this. If the climber is alright and can safely climb onto the rock or can be lowered to a secure ledge, all is well. However, the climber may be injured, or there may be no place to which the climber can be lowered. The belayer cannot continue holding the climber. The belayer must be relieved to assist the climber in the next step. This can be done by using a prusik sling. If the belayer is alone, the braking hand will hold the static climber while the other hand is free to anchor the belay line. This particular method is only good when using a belay anchor. Follow the procedure shown in figure 22-34. Climbers who are in good condition can either prusik up or pendulum across to a ledge.

22-8. Communications. Because constant communication between the belayer and climber is essential for safety, a standard group of climbing commands must be mastered.

a. Communications while climbing must be as simple as possible and single words are preferred. Communications in the form of commands are necessary when climbing. Commands which sound alike should be avoided. The command system follows a set pattern which leads to safe climbing and ensures that belays are used. Commands should be clear, specific, and given in a loud voice. The sequence of the system should not be broken. A review of the command system should be made before any climb. The European command has all the necessary commands to ensure safety. Only one command has a reply; the other commands are answered by the next command in the system. Sounding off the next command is not done until the requirements of the prior command are complete.

b. Commands are generally climber initiated. The belayer, in each instance, acknowledges when commands are heard and understood. If the command is not understood, the belayer should not say anything. The silence indicates to the climber that the belayer doesn't know what is happening. The command must be repeated.

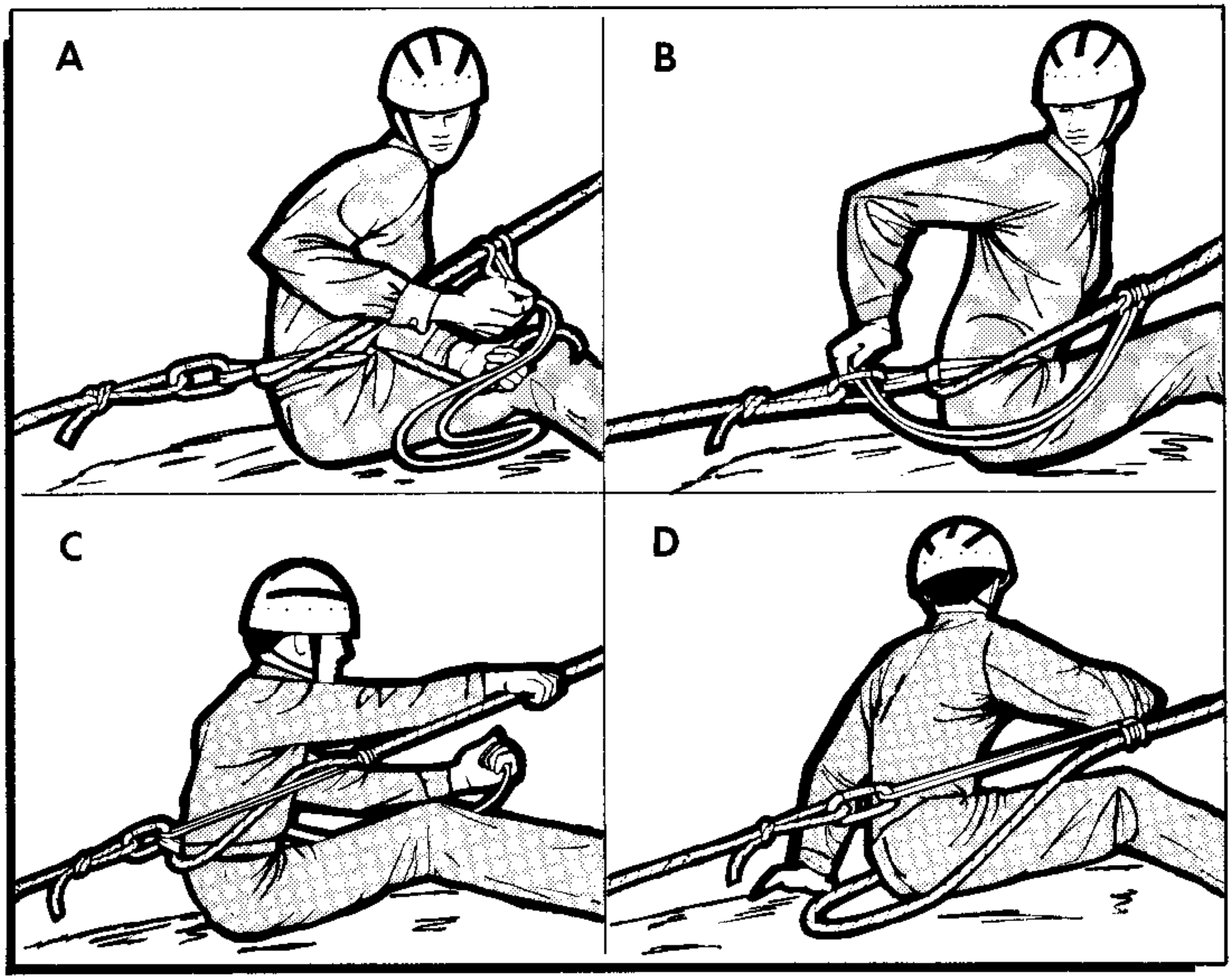

Figure 22-34. Safety Off the Belay Line.

(1) The following are the standard commands used:

(a) Initiation of a Climbing Sequence:

-1. "ON BELAY"—a signal indicating the climber is secured on the end of the rope and asks if belayer is ready.

-2. "BELAY ON"—(the belayer acknowledges the climber) means the belayer is in position and ready.

-3. "BELAY TEST"—the climber is asking to test the belay.

-4. "TEST"—signifies the belayer is ready for the climber to gradually apply weight on the rope. Releasing of the tension will indicate the test is finished.

-5. "UP ROPE"—the climber is directing the belayer to take up the slack.

-6. Belayer response: "ROPE UP."

-7. "THAT'S ME"—the climber signifies to the belayer that the tension felt on the rope is the climber.

-8. Belayer response: "THANK YOU."

-9. "CLIMBING"—the climber signifies the climber has chosen a route and is ready to climb.

-10. "CLIMB ON"—the belayer is acknowledging being ready for the climber to begin climbing. The climber will not start climbing before the command is heard.

(b) General Commands used while Climbing:

-1. "TENSION"—the climber is telling the belayer to take up all of the slack in the rope. This command is given if the climber needs assistance or there is an impending fall.

-2. Belayer responds with: "TENSION ON."

-3. "RESTING"—the climber is in a position for rest; belayer will lock brake across waist.

-4. Belayer responds with: "REST ON."

-5. When the climber begins again the command of "CLIMBING" will be given and "CLIMB ON" is given when the belayer is ready.

-6. "SLACK ___ FEET"—the climber will use this command when slack is needed in the rope to get over a difficult section or traverse. The climber should say "SLACK" and then the estimated number of feet the climber needs. Belayer will respond by repeating the number of feet of slack given. The climber will indicate that enough slack has been given by saying "THANK YOU."

-7. The command "FALLING" should be given when the climber is going to fall. This provides the belayer adequate time to apply the brake. Before continuing the climb, the climber will give the command "CLIMBING."

(c) Commands to Terminate the Climbing Sequence:

-1. "OFF BELAY"—the climber uses this command to indicate the climber is secure and through climbing.

-2. "BELAY OFF"—the belayer uses this command to indicate the brake hand is off the rope and the climber is no longer on belay.

22-9. Anchors. Anchors are secure points in the belay chain providing protection for the belayer and the climber. Anchor systems must be able to withstand high loads. The basis for any type anchor is strong, secure points for attachment.

a. Anchor Points. Anchor systems may be simple and consist of a single anchor point, or complex and consist of multiple anchor points. Anchor points are divided into two classes—natural and artificial. Anchor systems may be constructed entirely of one class, or a combination of the two.

(1) Natural Anchor Points (figure 22-35):

(a) Spike. A spike is a vertical projection of rock. For use as an anchor point, a sling is placed around the spike.

(b) Rock Bollard. A rock bollard is a large rock or portion of such a rock which has an angular surface enabling a sling or rope to be placed around it in such a manner that will not allow it to slip off. Care must be taken to ensure the bollard will not be pulled loose when subjected to a sudden load.

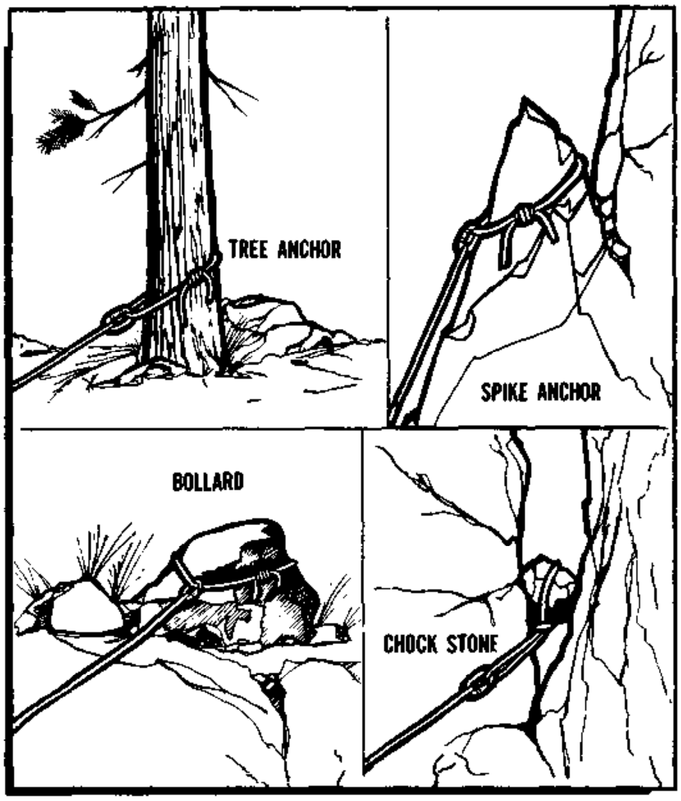

Figure 22-35. Natural Anchor Points.

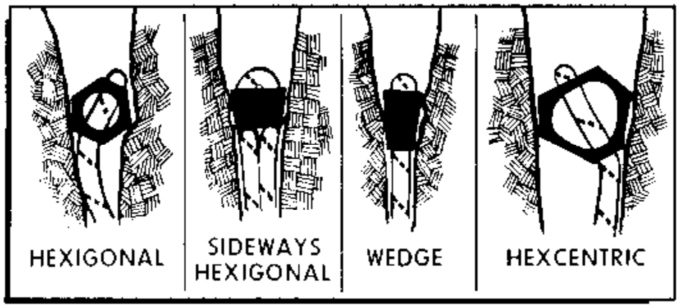
HEXIGONAL
SIDEWAYS HEXIGONAL
WEDGE
HEXCENTRIC

Figure 22-36. Placement of Chocks.

(c) Chockstone. A natural chockstone is a securely wedged rock providing an anchor point for a sling. In most cases, the rock is wedged within a crack.

(d) Tree. Trees often make very secure anchor points. A rope can be tied directly to the tree (or a sling doubled around the tree and connected by a carabiner). In loose or rocky soil, trees should be carefully watched and avoided if other anchor points are available.

(2) Artificial Anchor Points. Artificial anchor points are those constructed from equipment carried by the team. These are usually the chocks or pitons placed in cracks or bolts drilled in the rock.

(a) Chock Placement. The basic principle is to wedge the selected chock into a crack so that a pull in the direction of fall will not pull the chock out. The proper method is to select a crack suitable for an anchor point and select a chock to fit that crack. The chock chosen should closely fit the widest portion of the crack. Work the chock into the crack until it is securely seated. It should be seated so that the load will come on the entire chock without rotating it out of position. When the chock is in place, jerk hard on the attached sling in the direction of fall to ensure it is well seated. (Ensure the chock cannot continue to work downward to a larger area of the crack and become dislodged). A chock may also be placed so that its pull is in one direction while the second chock has a pull in the opposing direction. One sling is passed through the loop of the other chock. A force exerted downward will pull the two chocks toward each other along the axis of the crack. In vertical cracks, the lower sling should be passed through the upper for best results. The two chocks should be placed far enough apart so neither sling will reach the other chock. Figure 22-36 shows the various types of chocks

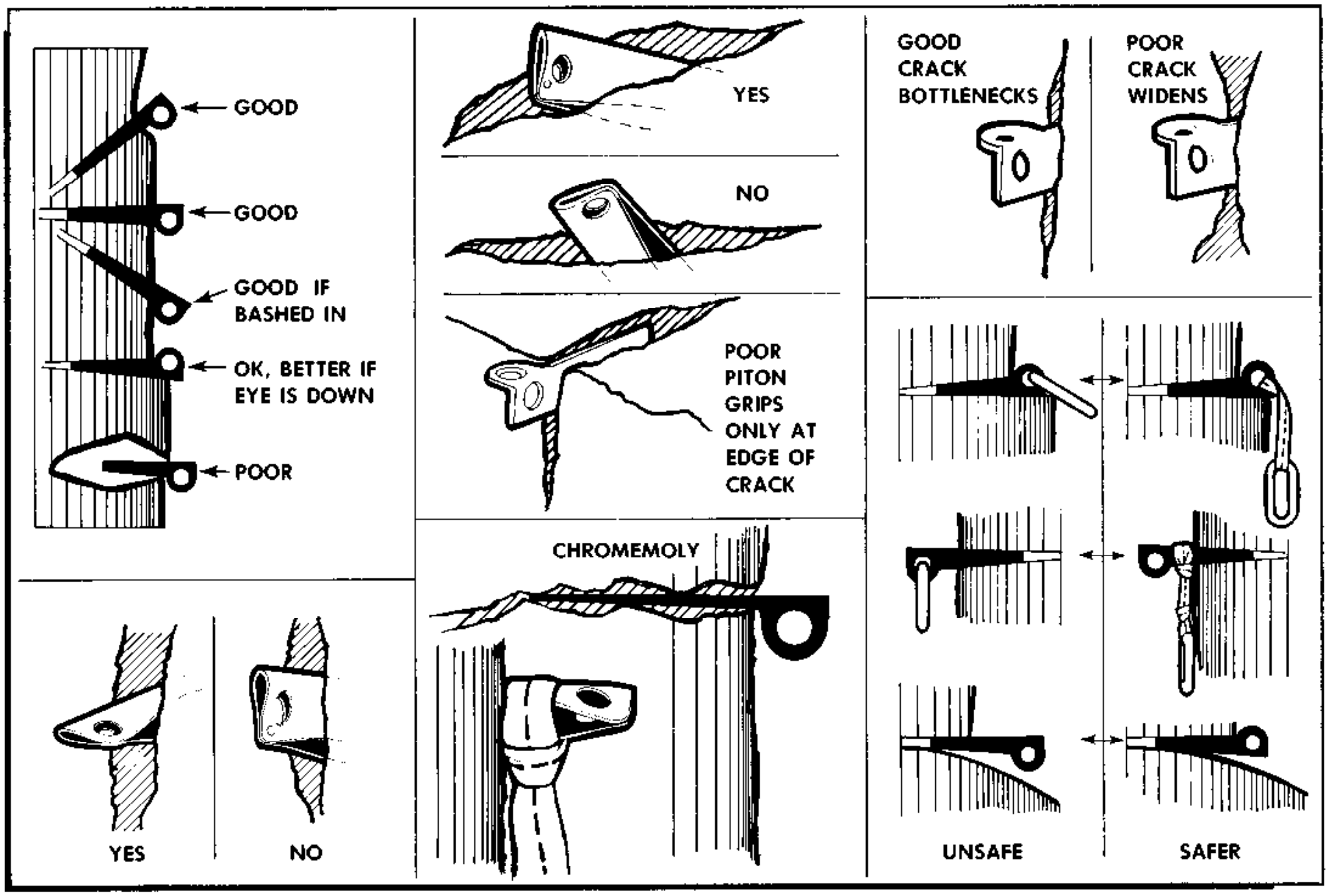
GOOD
GOOD
GOOD IF BASHED IN
OK, BETTER IF EYE IS DOWN
POOR
YES
NO
YES
NO
POOR PITON GRIPS ONLY AT EDGE OF CRACK
CHROMEMOLY
GOOD CRACK BOTTLENECKS
POOR CRACK WIDENS
UNSAFE
SAFER

Figure 22-37. Placement of Pitons.

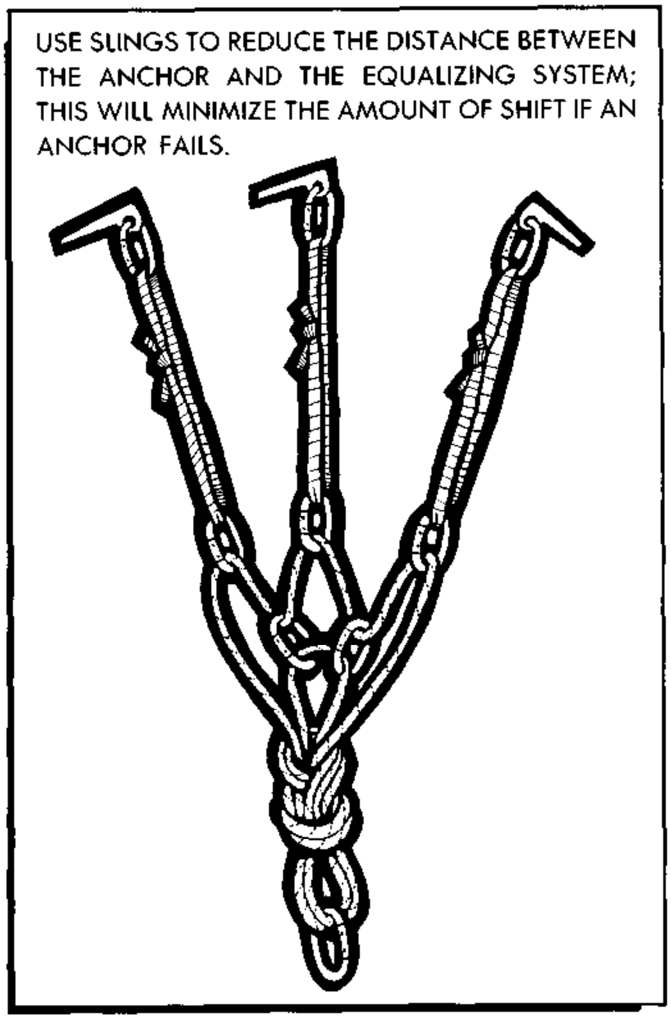

USE SLINGS TO REDUCE THE DISTANCE BETWEEN THE ANCHOR AND THE EQUALIZING SYSTEM; THIS WILL MINIMIZE THE AMOUNT OF SHIFT IF AN ANCHOR FAILS.

Figure 22-38. Equalizing Anchor.

and their placement. (NOTE: Chocks are preferred over pitons and bolts since they do not deface the rock.)

(b) Piton Placement. Basically, the use of pitons is a matter of locating a crack, selecting a piton which fits, driving the piton in, clipping a carabiner to the eye, and attaching the rope system to the carabiner.

-1. First look at the crack to decide the best position for driving the piton. The piton should be driven into the wider portion of the crack to reduce the likelihood of shifting or rotating under pressure. The crack must not widen or flare internally and not have a change in direction of more than 15 to 20 degrees.

-2. Before driving, the piton should fit one-half to two-thirds of the blade length into the crack. Drive until only the eye protrudes or the piton meets resistance. Do not attempt to overdrive the piton because it may fracture. Lightly tap the piton to test for movement or improper seating. If movement is noted, remove the piton and replace it with the next larger size or locate another anchor position. The proper placement of pitons in horizontal cracks is with the eye down. If the piton cannot be driven into the rock until only the eye protrudes, it may be tied off by placing a short sling around the piton with a girth or clove hitch as close to the rock as possible. In this case, the sling should be used as the attachment point and not the eye of the piton. Figure 22-37 shows the proper placement of pitons.

b. Anchor System. The purpose of an anchor system is to unite weak anchor points into a strong anchor system. Systems are divided into two classes—equalizing and nonequalizing.

(1) Equalizing systems are constructed so that if there is a change in direction of the load, the stress will be equally distributed to all anchor points. One major problem with this system is that if one point fails, the remaining points will be shock loaded (figure 22-38).

(2) Nonequalizing systems are constructed when a change of direction is not expected. The major advantage is the entire load is shared by the anchor points equally and would require the entire system to fail before coming loose. Any multipoint anchor system tied together so the attaching rope does not slip would be considered nonequalizing.

22-10. Climbing:

a. Balance Climbing. Balance climbing is the type of movement used to climb rock faces. It is a combination of the balance movement of a tightrope walker and the unbalanced climbing of a person ascending a tree or ladder. During the process of route selection, the climber should mentally climb the route to know what is expected. Climbers should not wear gloves when balance climbing.

(1) Body Position. The climber must keep good balance when climbing (the weight placed over the feet during movement). (See figure 22-39.) The feet, not the hands, should carry the weight (except on the steepest cliffs). The hands are for balance. The feet do not provide proper traction when the climber leans in toward the rock. With the body in balance, the climber moves with a slow, rhythmic motion. Three points of support, such as two feet and one hand, are used when possible. The preferred handholds are waist to shoulder high. Resting is necessary when climbing because tense muscles tire quickly. When resting, the arms should be kept low where circulation is not impaired. Use of small intermediate holds is preferable to stretching and clinging to widely separated big holds. A spread-eagle position, where a climber stretches too far (and cannot let go), should be avoided.

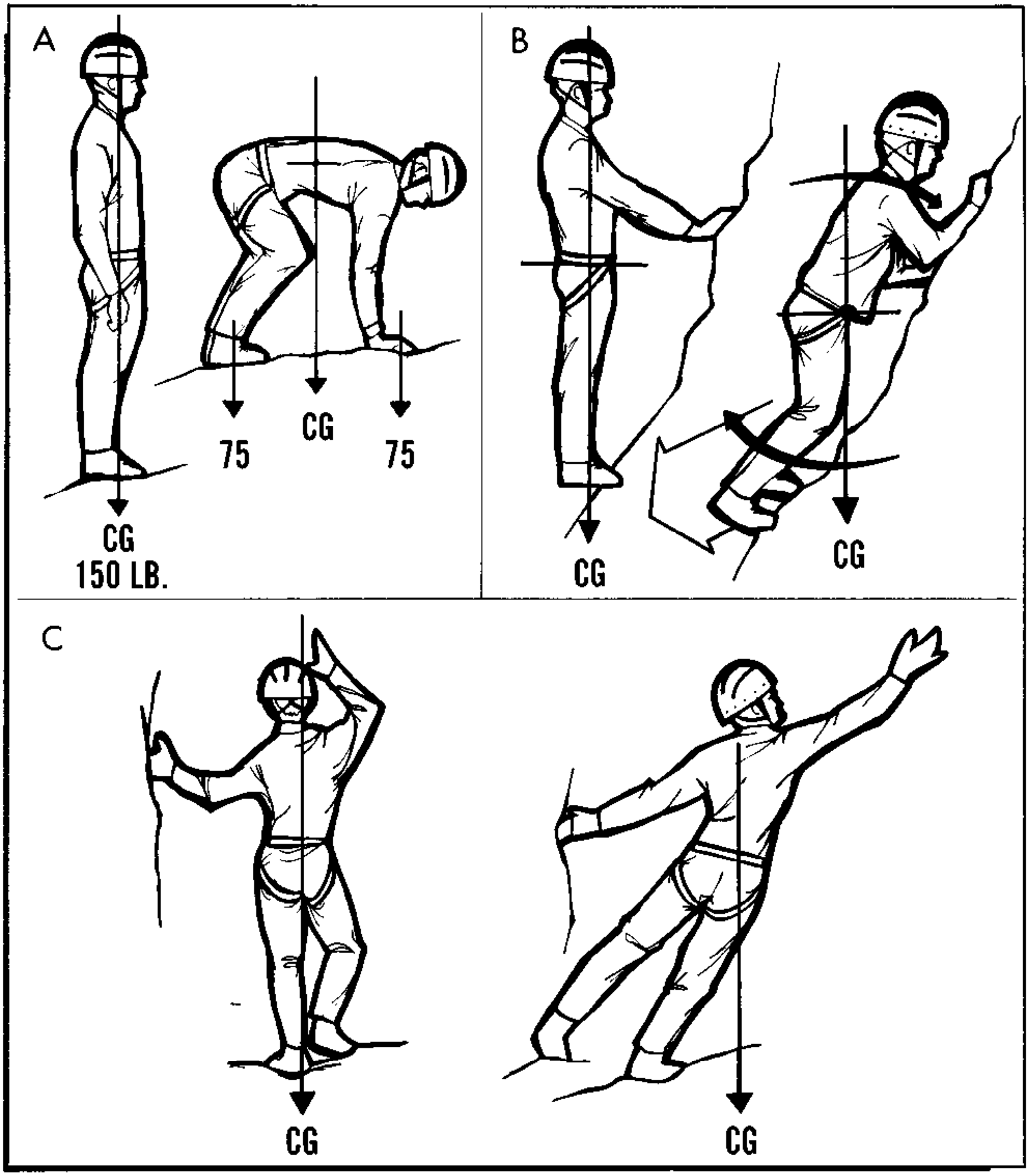

Figure 22-39. Body Position.

(2) Types of Holds:

(a) Push Holds. Push holds are desirable because they (figure 22-40) help the climber keep the arms low; however, they are more difficult to hold onto in case of a slip. A push hold is often used to advantage in combination with a pull hold.

(b) Pull Holds. Pull holds (figure 22-41) are those that are pulled down upon and are the easiest holds to use. They are also the most likely to break out.

(c) Jam Holds. Jam holds (figure 22-42) involve jamming any part of the body or extremity into a crack. This is done by putting the hand into the crack and clenching it into a fist or by placing the arm into the

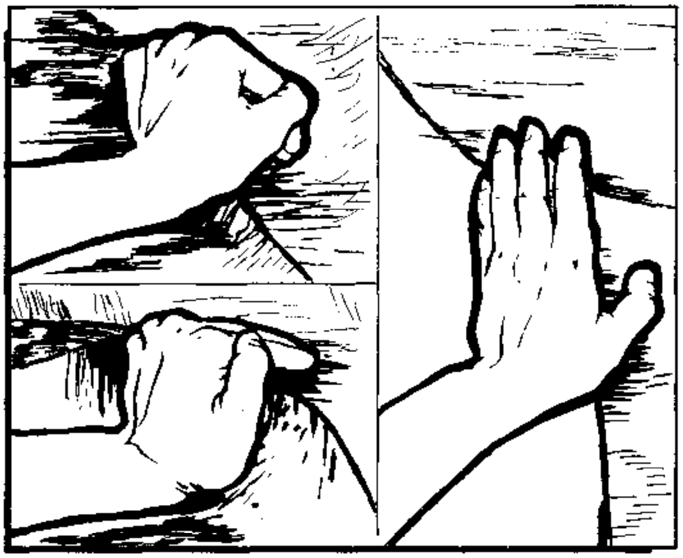

Figure 22-40. Push Holds.

crack and twisting the elbow against one side and the hand against the other side. When using the foot in a jam hold, care should be taken to ensure the boot is placed so it can be removed easily when climbing is continued.

(d) Combination Holds. The holds previously mentioned are considered basic and from these any number of combinations and variations can be used. The number of these variations depends only on the limit of the individual's imagination. Following are a few of the more common ones:

-1. The counterforce (figure 22-43) is attained by pinching a protruding part between the thumb and fingers and pulling outward or pressing inward with the arms.

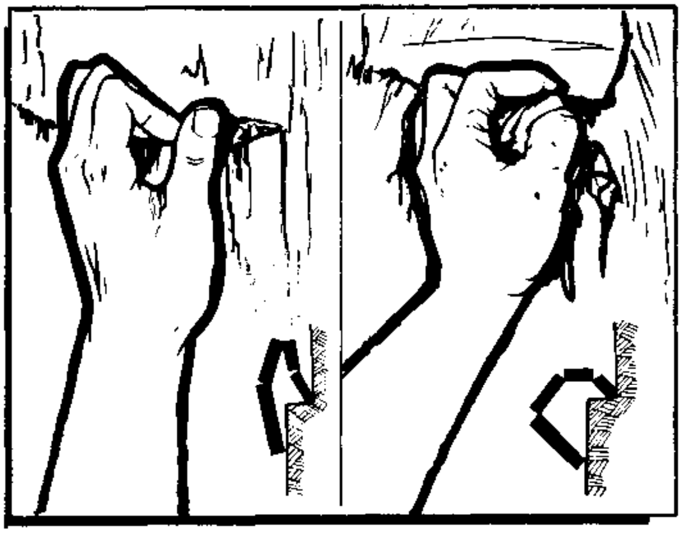

Figure 22-41. Pull Holds.

-2. The lay-back (figure 22-44) is done by leaning to one side of an offset crack with the hands pulling and the feet pushing against the offset side. Lay-backing is a classic form of force or counterforce where the hands and feet pull and push in opposite directions enabling the climber to move up in a series of shifting moves. It is very strenuous.

-3. Underclings (figure 22-45) permit cross pressure between hands and feet.

-4. Mantleshelving, or mantling, takes advantage of down pressure exerted by one or both hands on a slab or shelf. By straightening and locking the arm, the body is raised, allowing a leg to be placed on a higher hold (figure 22-46).

(e) Chimney Climb. This is a body-jam hold used in very wide cracks (figure 22-47). The arms and legs are used to apply pressure against the opposite faces of the

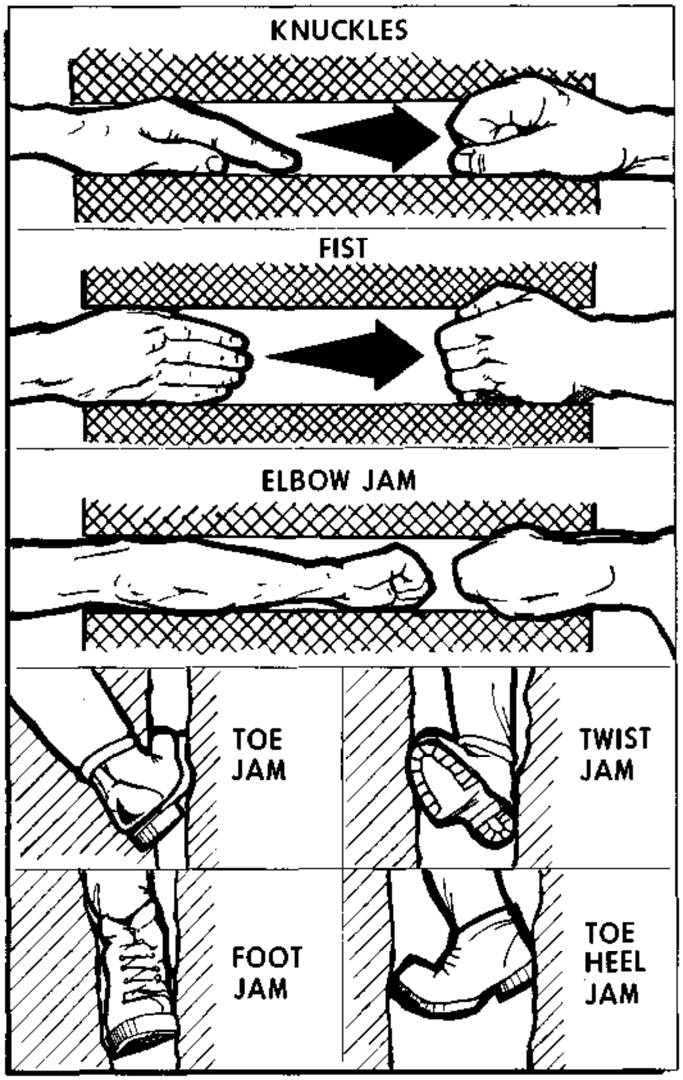

Figure 22-42. Jam Holds.

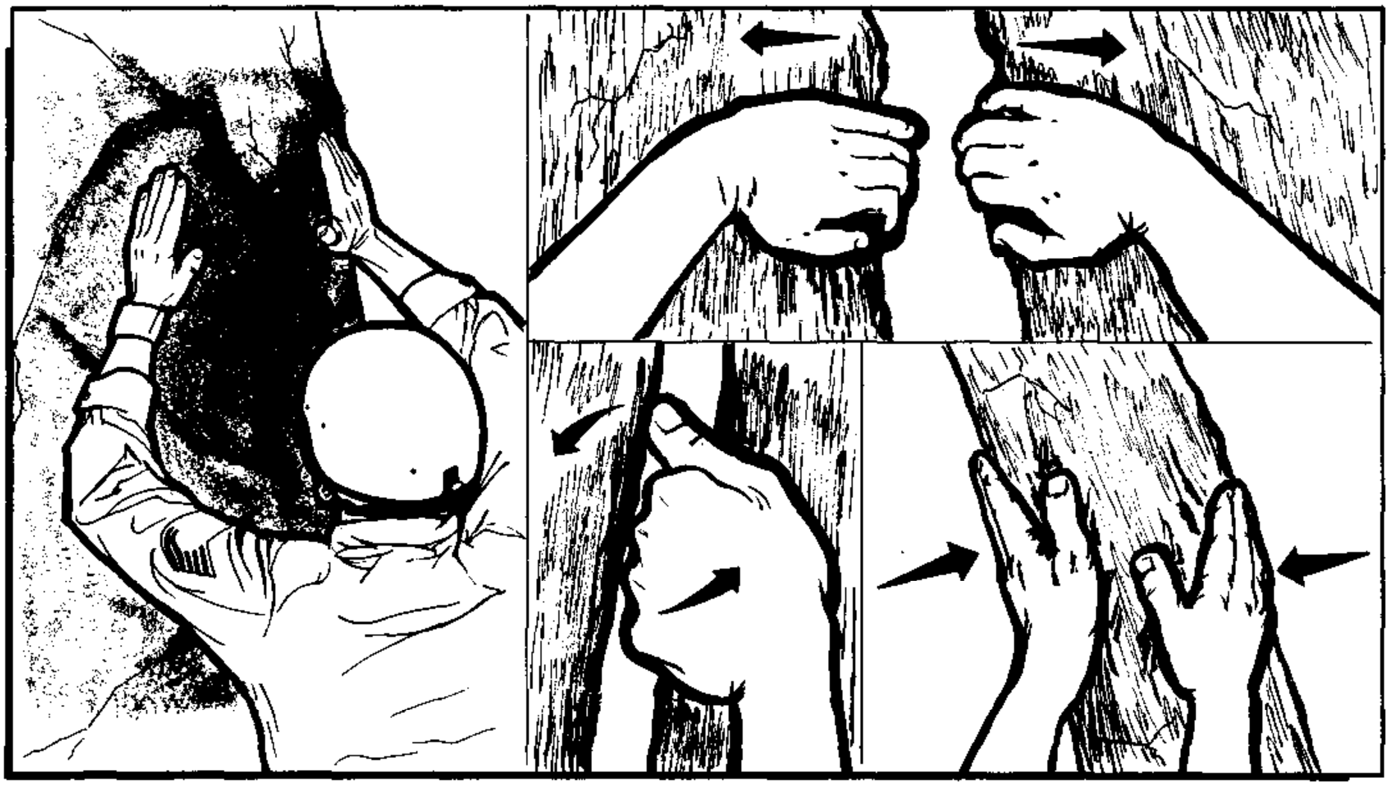

Figure 22-43. Combination Holds.

rock in a counterforce move. The outstretched hands hold the body while the legs are drawn as high as possible. The legs are flexed forcing the body up. This procedure is continued as necessary. Another method is to place the back against one wall and the legs and arms against the other and "worm" upward (figure 22-47).

b. Friction Climbing:

(1) A slab is a relatively smooth portion of rock lying at an angle. When traversing, the lower foot is pointed slightly downhill to increase balance and friction of the foot. All irregularities in the slope should be used for additional friction. On steep slabs, it may be necessary to squat with the body weight well over the feet with hands used alongside for added friction. This position may be used for ascending, traversing, or descending. A slip may result if the climber leans back or lets the buttocks down. Wet, icy, mossy, or a scree-covered slab is the most dangerous.

(2) Friction holds (figure 22-48) depend solely on the friction of hands or feet against a relatively smooth surface with a shallow hold. They are difficult to use because they give a feeling of insecurity which the inexperienced climber tries to correct by leaning close to the rock, thereby increasing the insecurity. They often serve well as intermediate holds, giving needed support while the climber moves over them; however, they would not hold if the climber decided to stop.

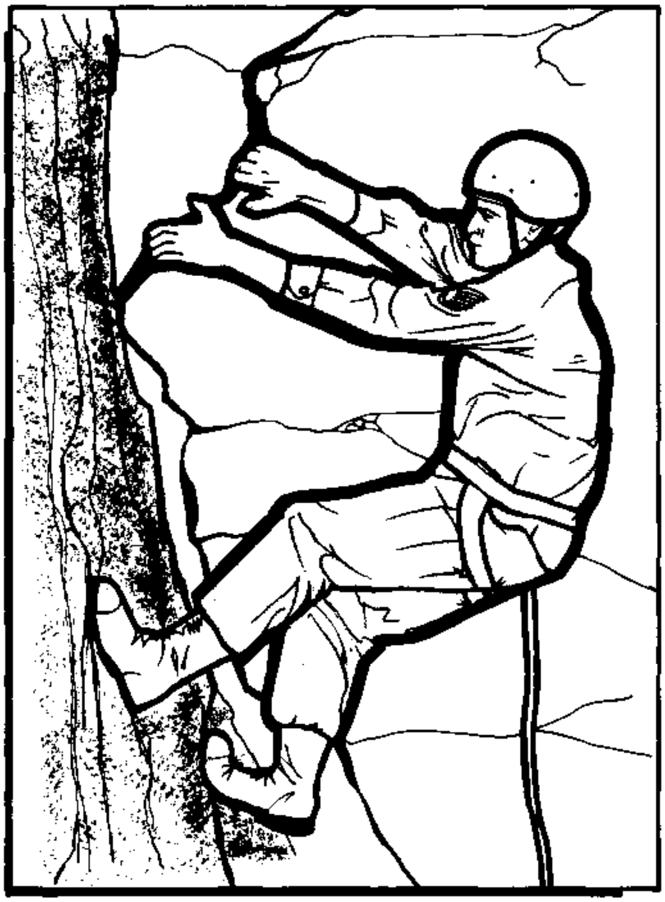

Figure 22-44. Lay-Back.

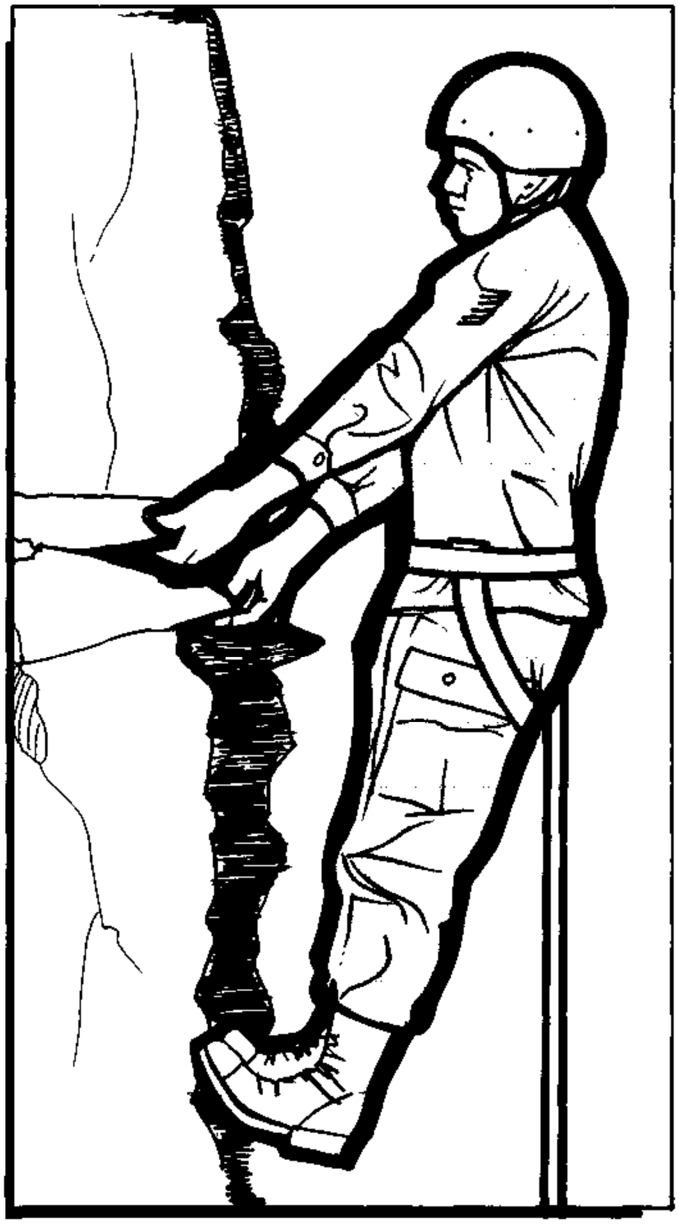

Figure 22-45. Underclings.

c. Prusik Climbing. Prusiking is a method of ascension using a fixed rope (figure 22-49). This method does not require a belay. With two prusik slings, a chest prusik sling, and a carabiner, a person can climb the length of a fixed rope. The slings are attached with prusik knots that grip tightly when loaded yet slide when the load is removed. Prusiking is strenuous and requires the use of both hands and both feet, hence it is no system for a badly injured person. Prusik slings are usually made from lengths of 7 mm Kernmantle or Mountain-Lay rope. Smaller diametet rope is less bulky and grips the climbing rope better, but in practice the knots are very difficult to work with gloved hands and is not recommended. To allow the climber to make maximum steps up with each foot in turn, the longer foot sling should extend to about nose level and the other to several inches below the waist loop. Both prusicks should have loops just big enough for the foot. The chest sling is just long enough to keep the climber from toppling backwards when rigged. Standing upright with all the weight bearing on the foot slings, the climber lifts one foot and raises its unweighted knot. Now, stepping up and shifting weight onto this sling, it is repeated with the other foot. The climb is less tiring when the foot slings are of different lengths, allowing the climber to take equal steps with both feet. There is some difficulty in sliding the knots when the rope is wet or the slings are made from laid rope. Spinning is also a potential problem under overhangs where the wall cannot be touched to maintain stability, particularly when climbing a laid rope. While a rescuer (climber) is ascending via the prusik method, companions have little to do except guard the anchors and prepare to lift the climber over the edge.

22-11. Rappelling. The climber with a rope can descend quickly by sliding down a rope which has been doubled around such anchor points as a tree, a projecting rock, or several artificial anchors secured to each other with sling rope.

a. Establishing a Rappel. In selecting the route, the climber should be sure the rope reaches the bottom or a place from which further rappels or climbing can be done. The rappel point should be carefully tested, and inspected to ensure the rope will run around it when one end is pulled from below and the area is clear of loose rocks. If a sling rope is used for a rappel point, it should be tied twice to form two separate loops. The first person down chooses a smooth route for the rope which is free of sharp rocks. Place loose rocks, which the rope might later dislodge, far enough back on ledges to be out of the way. The rappeler should ensure the rope runs freely around the rappel point when pulled from below. Each person down will give the signal "OFF RAPPEL," straighten the rope, and ensure the rope runs freely around its anchor. When silence is needed, a prearranged signal of pulling on the rope is substituted for the vocal signal. Recover the rope when the last person is down. The rope should be pulled smoothly to prevent the rising end from whipping around the rope. Climbers should stand clear of falling rope and rocks which may be dislodged. The rope should be inspected frequently if a large number of people are rappelling. Rappellers should wear gloves during rappels to protect the palms from rope burns.

b. Types of Rappel. The type of rappel is determined by the steepness of the terrain. The hasty rappel is used only on moderate pitch slopes. The body rappel may be used on moderate to severely pitched slopes, but never

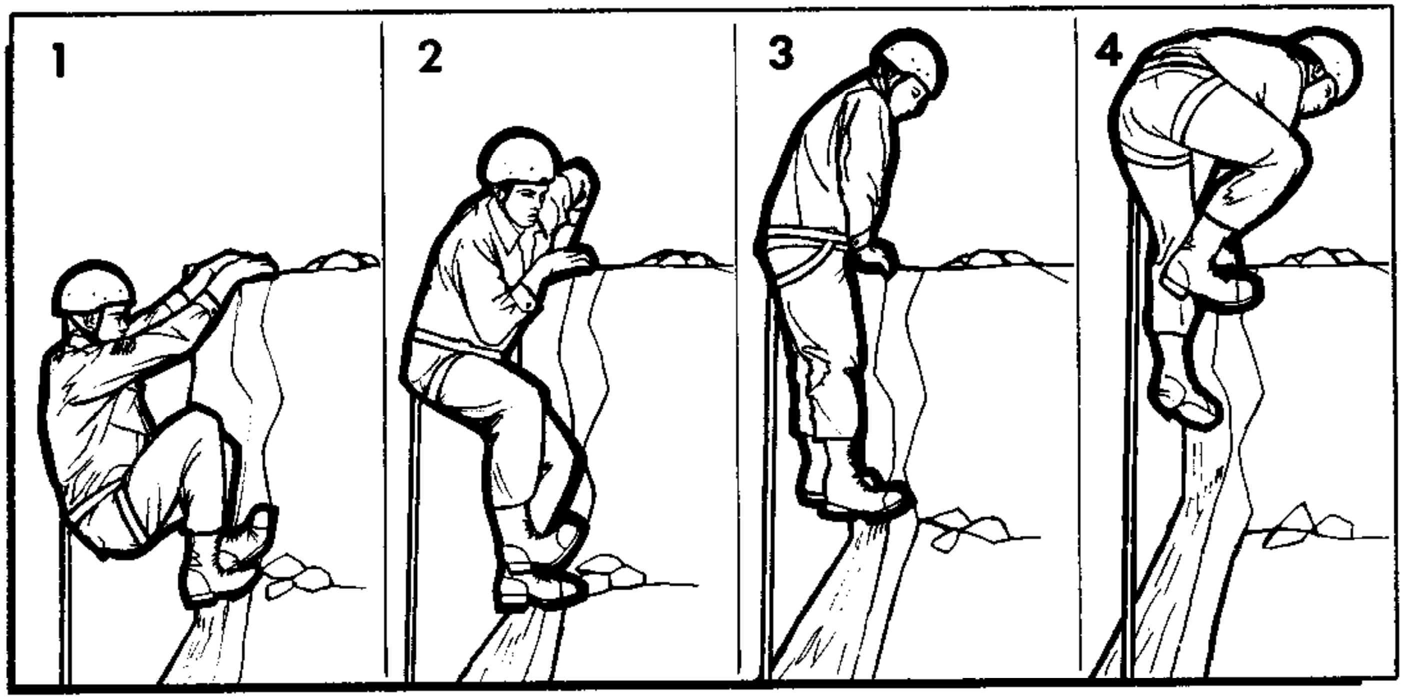

Figure 22-46. Mantleshelving.

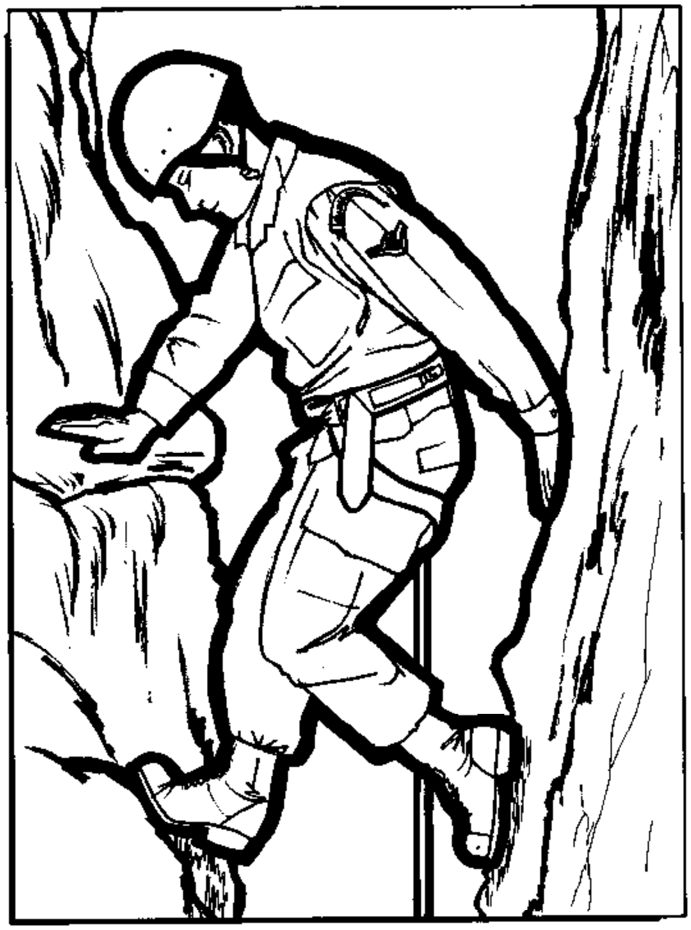

Figure 22-47. Chimney Climbing.

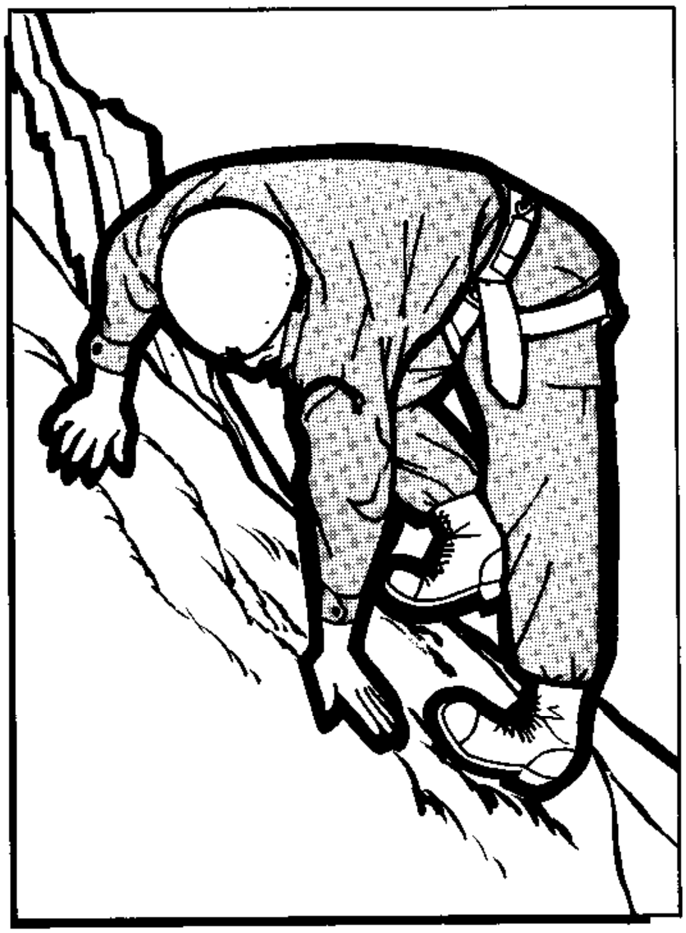

Figure 22-48. Friction Holds.

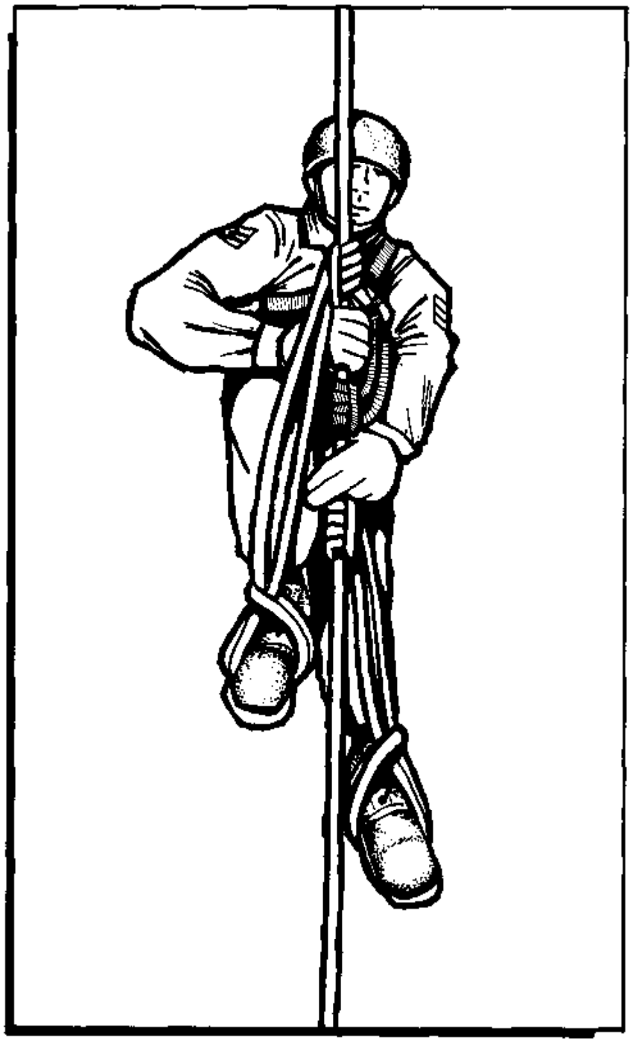

Figure 22-49. Prusiking.

used on overhangs. The seat rappel is used on very steep pitches, including overhangs.

(1) The Hasty Rappel (figure 22-50). Facing slightly sideways to the anchor, the climber places the ropes across the back. The hand nearest the anchor is the guiding hand. To stop, the rappeller brings the braking hand across in front of the body, and at the same time turns to face the anchor point.

(2) The Body Rappel (figure 22-51). The climber faces the anchor point and straddles the rope; then pulls the rope from behind, runs it around either hip, diagonally across the chest, and back over the opposite shoulder. From there the rope runs to the braking hand which is on the same side of the hip that the rope crosses (for example, the right hip to the left shoulder to the right hand). The climber should lead with the braking hand down and should face slightly sideways. The foot corresponding to the braking hand should precede the other at all times. The guiding hand should be used only to guide and not to brake. To rappel, lean out at a 45-degree angle to the rock. Keep the legs well spread and relatively straight for lateral stability, and the back straight since this reduces unnecessary friction. The collar should be turned up to prevent rope burns on the neck. Gloves should be worn and any other articles of clothing may be used as padding for the shoulders and buttocks. To brake, lean back and face directly into the rock so the feet are flat on the rock.

(3) Four-Carabiner Seat Rappel. Seat rappels differ from the body and hasty rappels in that the friction is primarily absorbed by a carabiner in the sling rope seat worn by the rappeller. The rappeller stands to one side of the rope (when braking with the right hand on the left and when braking with the left hand on the right). Some slack between the carabiner and the anchor point is taken up and brought through two carabiners which are attached to the harness and are horizontal and reversed and opposed (figure 22-52). Two additional carabiners are clipped from opposite sides, gates down, at 90 degrees to the original carabiners to form a friction brake bar (figure 22-53).

c. Body Position. The body must be perpendicular to the face of the rock, and the feet must be about shoulder width apart and flat on the rock (figure 22-54). To de-

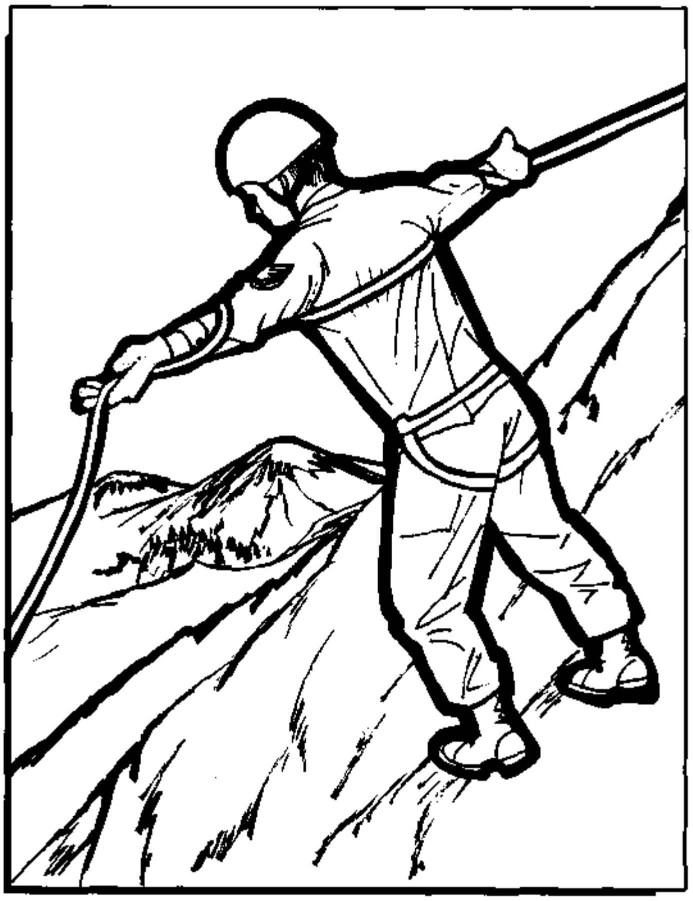

Figure 22-50. Hasty Rappel.

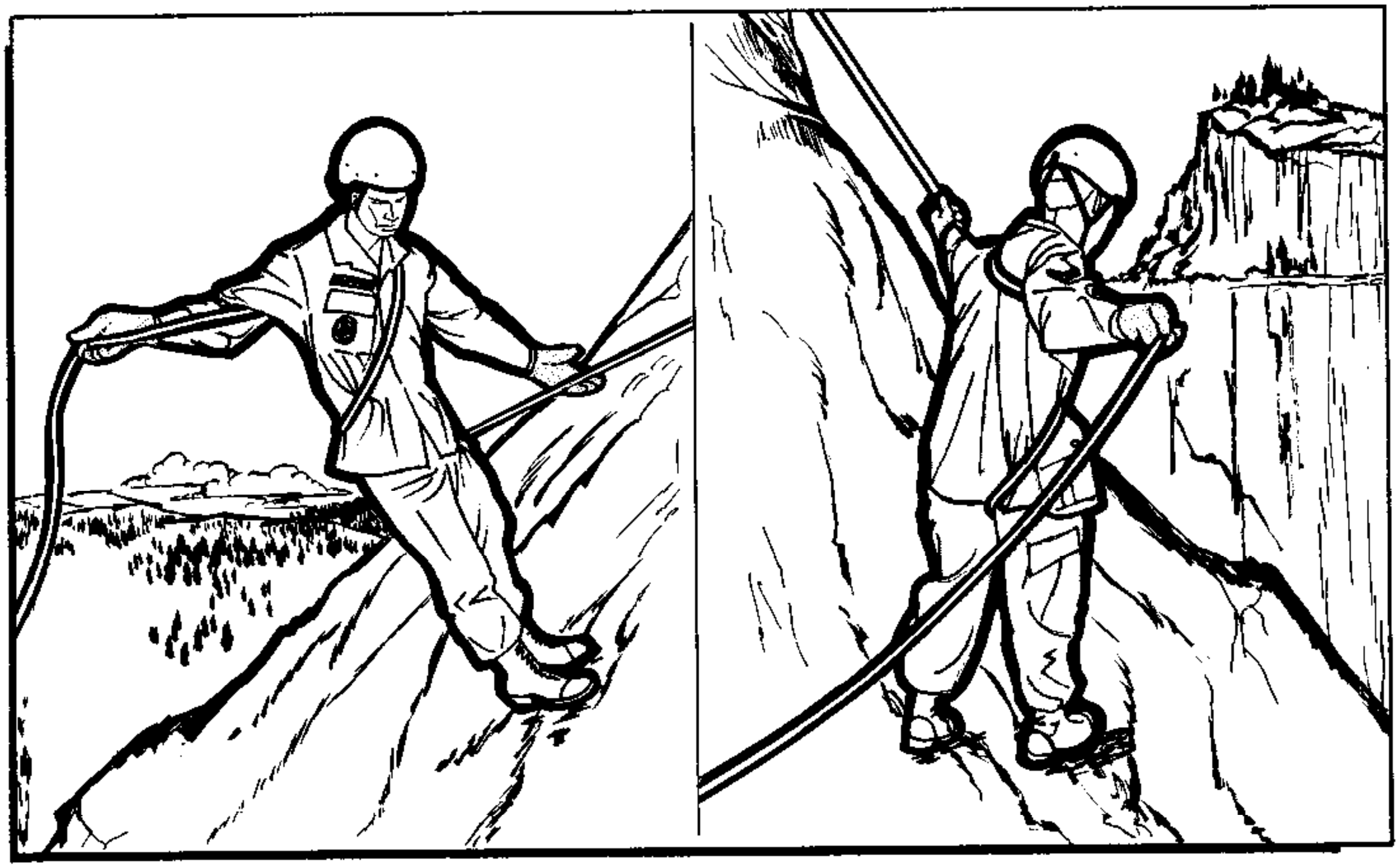

Figure 22-51. Body Rappel.

scend, hold the brake out to the side to reduce friction. To brake, cinch down on the rope and move the brake hand to the small of the back.

22-12. Overland Snow Travel. Cold-weather operations conducted in snow-covered regions magnify the difficulties of reaching a survivor and effecting an extraction. Routes of travel surveyed from the air may not be possible from the ground. A straight line from the insertion point to the objective is the most desired route; however, inherent dangers (avalanches, collapsing cornices, etc.) may necessitate an alternate route entailing a longer trek. During all operations, safety, and not ease of travel, will be the primary concern.

a. Snow Conditions. Travel time varies from hour to hour. Certain indicators may assist in the direction of travel; that is, the best snow condition is one which supports a person on or near the surface when wearing boots and the second best is calf-deep snow conditions. If possible avoid traveling in thigh- or waist-deep snow. Snowshoes should be worn when conditions dictate.

(1) South and west slopes offer hard surfaces late in the day after exposure to the Sun and the surface is refrozen. East and north slopes tend to remain soft and unstable. Walking on one side of a ridge, gully, clump of

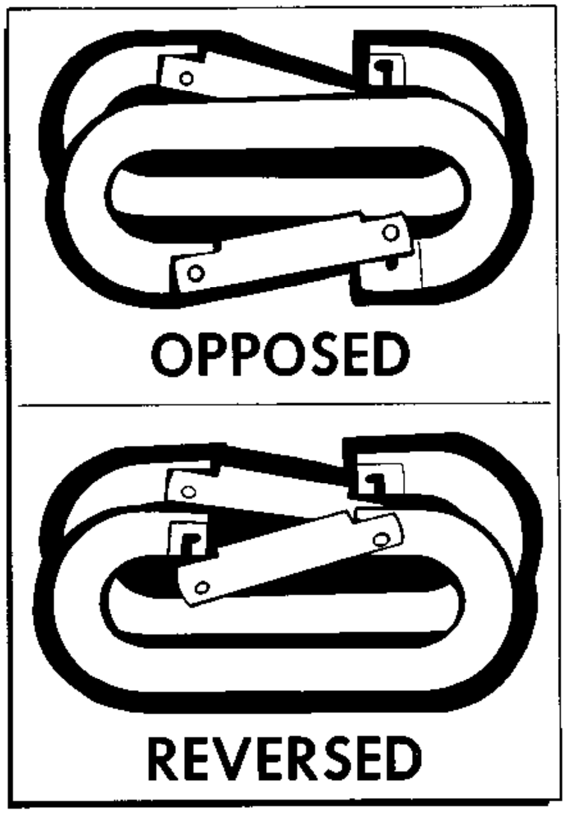

Figure 22-52. Seat Rappel (Carabiners).

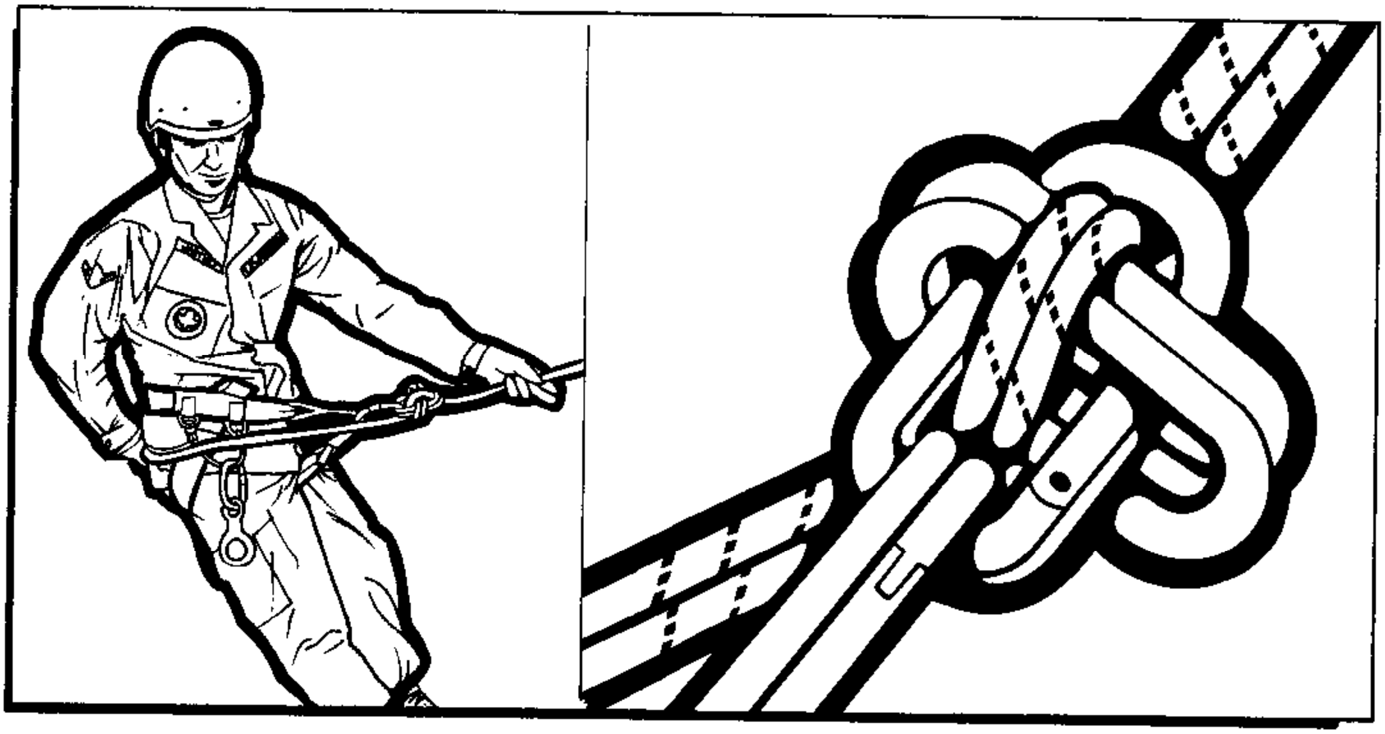

Figure 22-53. Four-Carabiner Brake System.

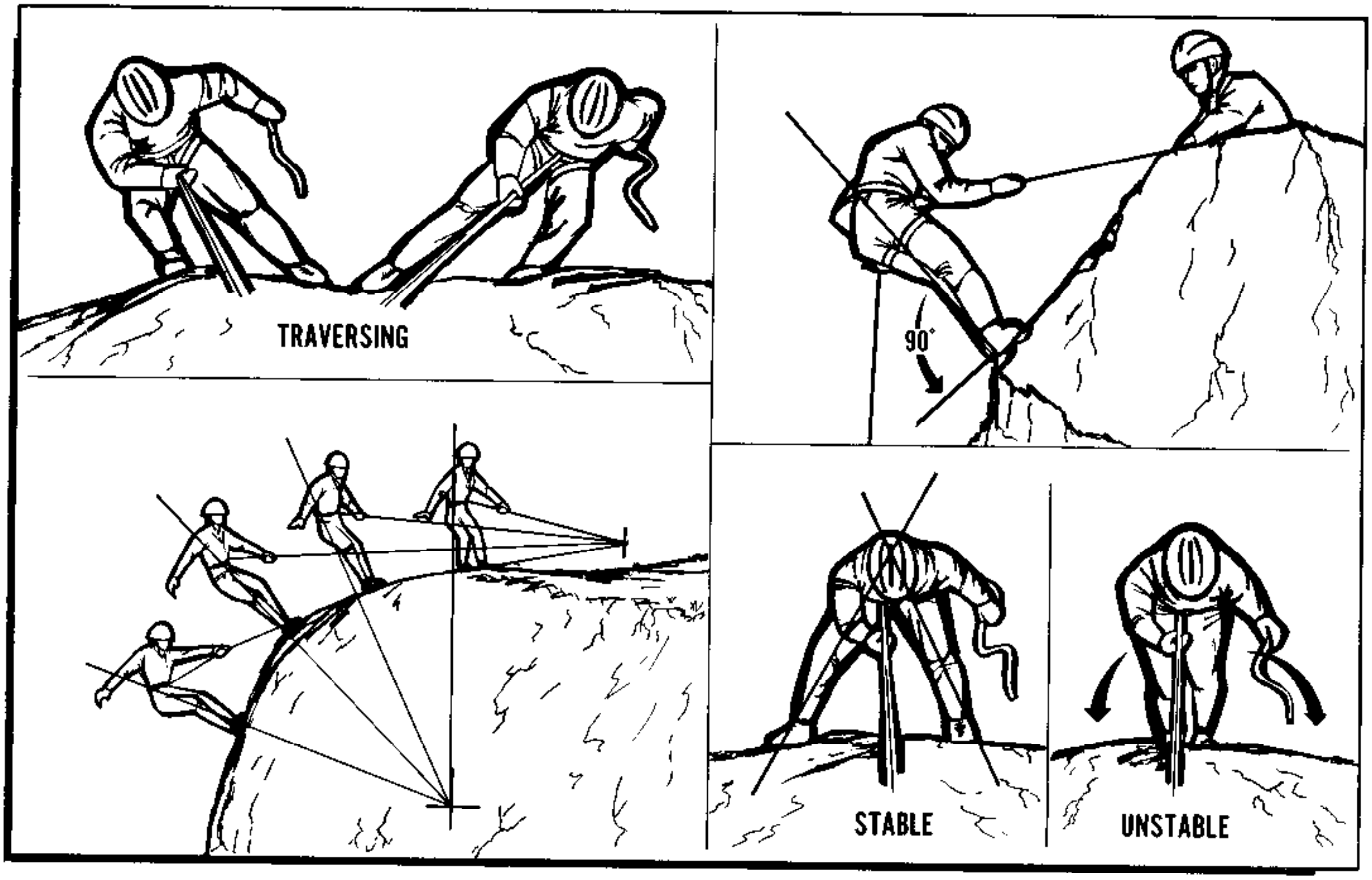
TRAVERSING
90°
STABLE
UNSTABLE

Figure 22-54. Body Position for Rappel.

trees, or large boulders is often more solid than the other side. Dirty snow absorbs more heat than clean snow; slopes darkened by rocks, dust, or uprooted vegetation usually provide more solid footing. Travel should be done in the early morning after a cold night to take advantage of stable snow conditions. Since sunlight affects the stability of snow, travel should be concentrated in shaded areas where footing should remain stable.

(2) In areas covered by early seasonal snowfall, travel between deep snow, and clear ground must be done cautiously. Snow on slopes tends to slip away from rocks on the downhill side, forming openings. These openings, called moats, are filled by subsequent snowfalls. During the snow season, moats below large rocks or cliffs may become extremely wide and deep, presenting a hazard to the rescue team.

b. Travel Speed. An over-zealous drive to reach an objective may be too fast for the endurance of the team. Fast starts at the point of insertion usually result in frequent stops for recuperation. The best way to reach an objective is to start with a steady pace and continue that pace throughout. Movement at reasonable speeds, with rest stops as required, will help prevent team "burnout." The following considerations further ensure steady advancement with minimal degradation to the team. A steady pace helps maintain an even rate of breathing. After the initial period of travel; that is, one-half hour, a shakedown rest should be initiated to adjust boots, snowshoes, crampons, packs, etc., or to remove or add layers of clothing.

c. Snowshoe Technique. A striding technique is used for movement with snowshoes. In taking a stride, the toe of the snowshoe is lifted upward to clear the snow and thrust forward. Energy is conserved by lifting the snowshoe no higher than is necessary to clear the snow. If the front of the snowshoe catches, the foot is pulled back to free it and then lifted before proceeding with the stride. The best and least exertive method of travel is a loose-kneed rocking gait in a normal rhythmic stride. Care should be taken not to step on or catch the other snowshoe.

(1) On gentle slopes, ascent is made by climbing straight upward. (Traction is generally very poor on hard-packed or crusty snow.) Steeper terrain is ascended by traversing and packing a trail similar to a shelf. When climbing, the snowshoe is placed horizontally in the snow. On hard snow, the snowshoe is placed flat on the surface with the toe of the upper one diagonally uphill to get more traction. If the snow will support the weight of a person, it is better to remove the snowshoes and temporarily proceed on foot. In turning, the best method is to swing the leg up and turn the snowshoe in the new direction of travel.

(2) Obstacles such as logs, tree stumps, ditches, and small streams should be stepped over. Care must be taken not to place too much strain on the snowshoe ends by bridging a gap, since the frame may break. In shallow snow, there is danger of catching and tearing the webbing on tree stumps or snags. Wet snow will frequently ball up under the feet, making walking uncomfortable. This snow should be knocked off with a stick or pole.

(3) Generally, ski poles are not used in snowshoeing; however, one or two poles are desirable when carrying heavy loads, especially in mountainous terrain. The bindings must not be fastened too tightly or circulation will be impaired and frostbite can occur. During stops, bindings should be checked for fit and possible readjustment.

d. Uphill Travel. Maximum altitude may be obtained with less effort by traversing a slope. A zigzag or switchback route used to traverse steep slopes places body weight over the entire foot as opposed to the balls of the feet as in a straight line uphill climb. An additional advantage to zigzagging or switchbacking is alternating the stress and strain placed on the feet, ankles, legs, and arms when a change in direction is made.

(1) When a change in direction is made, the body is temporarily out of balance. The proper method for turning on the steep slope is to pivot on the outside foot (the one away from the slope). With the upper slope on the right side, the left foot (pivot foot) is kicked directly into the slope. The body weight is transferred onto the left foot while pivoting toward the slope. The slope is then positioned on the left side and the right foot is on the outside.

(2) In soft snow on steep slopes, pit steps must be stamped in for solid footing. On hard snow, the surface is solid but slippery, and level pit steps must be made. In both cases, the steps are made by swinging the entire leg in toward the slope, not by merely pushing the boot into the snow. In hard snow, when one or two blows do not suffice, crampons should be used. Space steps evenly and close together to facilitate ease of travel and balance. Additionally, the lead climber must consider the other team members, especially those who have a shorter stride.

(3) The team should travel in single file when ascending, permitting the leader to establish the route. The physical exertion of the climbing leader is greater than that of any other team member. The climbing leader must remain alert to safeguard other team members while choosing the best route of travel. The lead function should be changed frequently to prevent exhaustion of any one individual. Team members following the leader should use the same leg swing technique to establish foot positions, improving each step as they climb. Each foot must be firmly kicked into place, securely positioning the boot in the step. In compact

snow, the kick should be somewhat low, shaving off snow during each step, thus enlarging the hole by deepening. In very soft snow, it is usually easier to bring the boot down from above, dragging a layer of snow into the step to strengthen and decrease the depth of it.

(4) When it is necessary to traverse a slope without an increase in elevation, the heels rather than the toes form the step. During the stride, the climber twists the leading leg so that the boot heel strikes the slope first, carrying most of the weight into the step. The toe is pointed up and out. Similar to the plunge step, the heel makes the platform secure by compacting the snow more effectively than the toe.

e. Descending. The route down a slope may be different from the route up a slope. Route variations may be required for descending different sides of a mountain or moving just a few feet from icy shadows onto sun-softened slopes. A good surface snow condition is ideal for descending rapidly since it yields comfortably underfoot. The primary techniques for descending snow-covered slopes are plunge stepping and descending step by step.

(1) The plunge step makes extensive use of the heels of the feet (figure 22-55) and is applicable on scree as well as snow. Ideally, the plunging route should be at an angle, one that is within the capabilities of the team and affords a safe descent. The angle at which the heel should enter the surface varies with the surface hardness. On soft snow slopes, almost any angle suffices; however, if the person leans too far forward, there is a risk of lodging the foot in a rut and inflicting injuries. On hard snow, the heel will not penetrate the surface unless it has sufficient force behind it. Failure to firmly drive the heel into the snow can cause a slip and subsequent slide. The quickest way to check a slip is to shift the weight onto the other heel, making several short, stiff-legged stomps. This technique is not intended to replace "the ice arrest" technique which is usually more effective. When roped, plunging requires coordination and awareness of all team members' progress. Speed of the team must be limited to the slowest member. Plunging is unsatisfactory when wearing crampons due to the snow compacting and sticking to them.

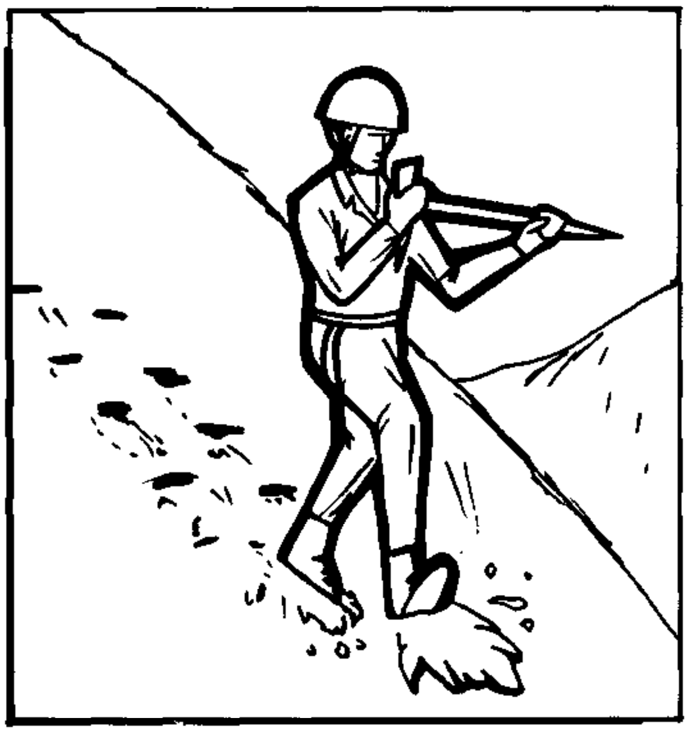

Figure 22-55. Plunge Step.

(2) The technique of step-by-step descending is used when the terrain is extremely steep, snow significantly deep, or circumstances dictate a slower pace. On near-vertical walls, it is necessary to face the slope and cautiously lower oneself step by step, thrusting the toe of the boot into the snow while maintaining an anchor or handhold with the axe. Once the new foothold withstands the body's full weight, the technique is repeated. On moderately angled terrain, the team can face away from the slope and descend by step-kicking with the heels.

22-13. Snow and Ice Climbing Procedures and Techniques. Snow and ice climbing differs from rock climbing, yet many of the procedures and techniques are the same. Belay tie-in commands, principles of runner placement, straight-line climbing, and placement of protection are common to snow and ice as well as rock. As expected, however, there are major differences from rock climbing.

a. Ice Axe Techniques (figure 22-56). The axe is the most important tool a climber carries. It can be used for braking assistance when a climber begins sliding down a steep snow-covered incline.

(1) Each rescuer (climber) must practice the self-arrest technique before venturing onto steep grades. Since the ice axe arrest requires the use of the ice axe, the climber must hang onto it at all times. The ice axe, whether sharp or not, is a lethal weapon when flying about on the attached cord. Physically, the climber rigs

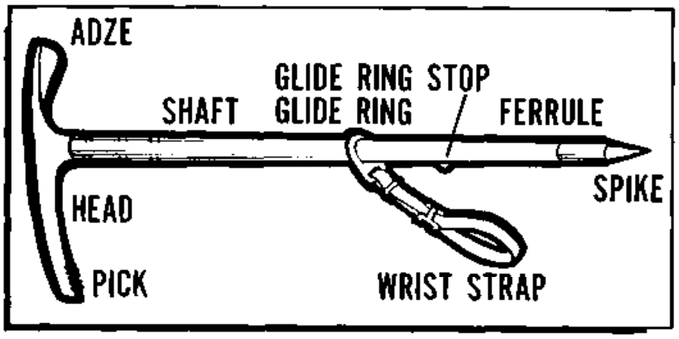

Figure 22-56. Ice Axe.

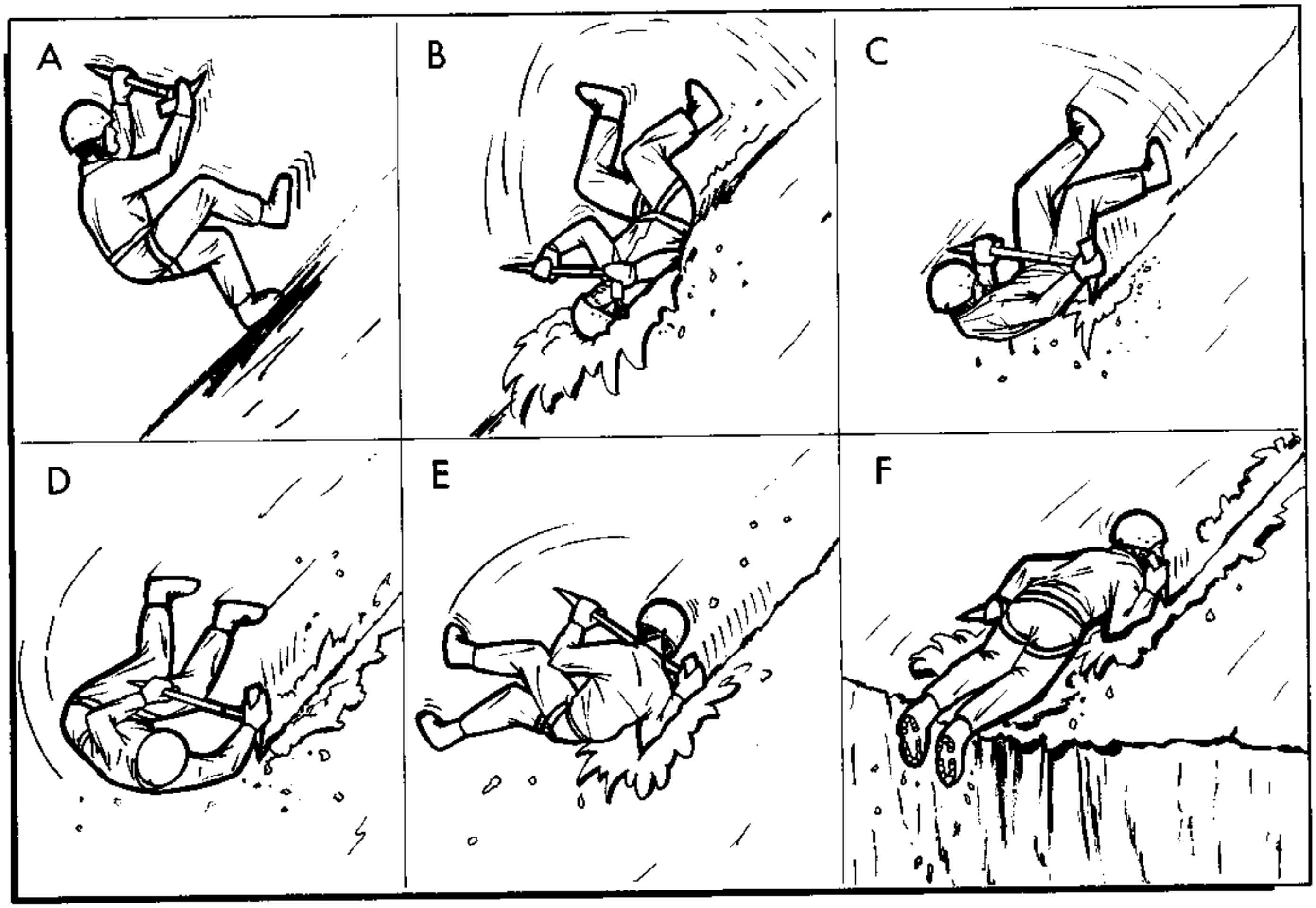

Figure 22-57. Ice Axe Self-Arrest.

for arrest by rolling down shirt sleeves, putting on mittens, securing loose gear, and most important of all, making certain the axe is held correctly. Mentally prepare by recognizing the importance of instantaneous application. A quick arrest, before the fall picks up speed, has a better chance of success than a slow arrest. Preparation for an ice axe arrest should be taken when traveling on terrain which could result in a fall.

(2) The proper method of holding an ice axe for self-arrest (see figure 22-57) is to place one hand on the head of the axe with the thumb under the adze and fingers over the pick. The other hand is placed on the shaft next to the spike. The pick is pressed into the slope just above the shoulder so the adze is near the angle formed by the neck and shoulders. The shaft should cross the chest diagonally with the spike held firmly close to the opposite hip. A short axe is held in the same position, although the spike will not reach the opposite hip. Chest and shoulders should press strongly on the shaft and the spine should be arched slightly to distribute weight primarily at the shoulders and toes. The legs should be stiff and spread apart, toes digging in (if wearing crampons, keep the toes off the surface until almost stopped), and hang on to the axe!

b. Team Arrest. The team arrest is intermediate between self-arrest and belays. When there is doubt that a person could arrest a fall, such as on crevassed glaciers and steep snowfields, and conditions are not so extreme as to make belaying necessary, the party ropes up and travels in unison. If any member falls, arrest is made by two or three axes. The rope between the climbers must be fully extended except for minimum slack carried by the second and subsequent persons to allow them to flip the rope out of the track (steps). This also allows easy compensation for pace variations. However, slack is minimized to bring the second and subsequent axes into action at the moment of need. A roped climber who falls should immediately yell "FALLING." It is not advisable to delay the alarm to see how "self-arrest" will develop because team members may hear the signal after they have been pulled into their own falls, decreasing their ability to help. When roped climber(s) hear the cry "falling," they immediately drop into self-arrest position.

c. Boot Axe Belay (figure 22-58). The boot axe belay can be set up rapidly and used when a team is moving together and belaying is only required at a few spots. The boot axe belay should be practiced until a sweep

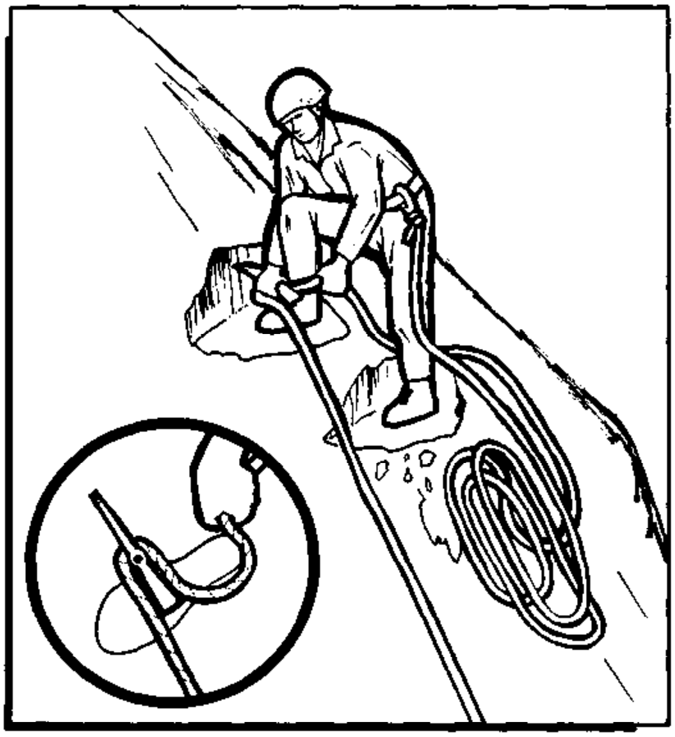

Figure 22-58. Boot Axe Belay.

and jab of the ice axe can set up the stance within a couple of seconds. The axe provides an anchor to the slope and the slope and the boot braces the axe. Both give a friction surface over which the run of rope is controlled.

(1) To prepare a boot axe belay, a firm platform, large enough for the axe and uphill boot, is stamped out in the snow. The ice axe shaft is jammed as deeply as possible, at a slight uphill angle (against the anticipated fall) into the snow at the rear of the platform. The pick is parallel to the fall line, pointing uphill, thus applying the strongest dimension of the shaft against the force of a fall. The length of the pick prevents the rope from escaping over the top of the shaft.

(2) The belayer stands below the axe, facing at a right angle to the fall line. The uphill boot is stamped into the slope against the downhill side of the shaft at a right angle to the fall line, bracing the shaft against downhill pull. The downhill boot is in a firmly compacted step below the uphill boot so that the leg is straight, stiffly bracing the belayer. The uphill hand is on the axe head in arrest grasp, bracing the shaft against downhill and lateral stress. From below, the rope crosses the toe of the boot, preventing the rope from trenching into the snow. The rope bends around the uphill side of the shaft, then down across the instep of the bracing boot, and is controlled by the downhill hand. To apply braking through greater friction, the downhill or braking hand brings the rope uphill around the heel, forming an "S" bend.

d. Crampon Techniques:

(1) Donning. When attaching the harness, the buckles should be positioned to cinch on the outward sides of the boots. Special care must be taken to strap the crampons tightly to the boots, running the strap through each attachment prong or ring. If crampons do not have heel loops, ankle straps should be long enough to be crossed behind the boot before being secured to prevent boots from sliding backward out of the crampons. Many crampons have been lost because this precaution was not taken. When trimming new straps, allowance must be made for gaiters which sometimes cover the instep of the boot. Donning is best done by laying each crampon on the snow or ice with all rings and straps outward; then place the boot on the crampon and tighten the straps. Even modern neoprene-coated nylon straps should be checked from time to time to make sure they are tight, have not been cut, and are not trailing loop strap ends which could cause the wearer to trip.

(a) If it is believed crampons may be needed, they must be carried. Conditions change rapidly; an east-facing slope may be mushy enough for step-kicking during the morning, but can become a sheet of smooth white ice in the afternoon shade. Furthermore, cramponing may contribute directly to the team's safety by enabling it to negotiate stretches of ice faster and with less fatigue than having to chop steps. The decision of whether or not to wear crampons is determined by the situation. Wearing crampons should not be considered mandatory because of venturing onto a glacier; neither should a team attempt to save time by never wearing crampons on steep, exposed icy patches just because they are fairly short. Another important guideline is to don crampons before they are needed to avoid donning them while teetering in ice steps. On mixed rock and ice climbs, constant donning and removing of crampons takes so much time that the objective may be lost.

(b) Crampons should be worn throughout the entire climb if the terrain is 50 percent or more suitable for crampons (crampons may skid or be broken on rock surfaces). Crampons are not required if the snow or ice patches are fairly short, good belays are available, and rock predominates. These alternatives are suggestions and the team leader's decision must be based on the conditions at hand.

(c) Crampons should be taken off when the snow begins to ball up badly in them and no improvement in snow conditions is anticipated. On the ascent, it may be possible to clear away the soft surface snow and climb on the ice below, but this is usually impractical and futile on the descent. Occasionally the climber should kick the crampons free of accumulated snow. The time-worn practice of striking the ice axe shaft against the crampons to knock out the snow is effective, but hard on the axe and perhaps the ankle. In situations where the crampons must be worn even though the snow balls up in them, shuffling the feet through the snow instead of stepping over the surface tends to force the accumulated snow through the back points. The normal kicking motion of the foot generally keeps the crampons snow-free on the ascent and while traversing.

(d) On the descent, drive the toe of the boot under the surface of the snow ahead of the heel, walking on the ball of the foot. Keep the weight well forward and use short skating steps allowing the foot to slide forward and penetrate the harder sublayers.

(2) Flatfooting. Flatfooting involves a logical and natural progression of coordinated body and ice axe positions to allow the climber to move steadily and in balance while keeping all vertical points of the crampons biting into the ice. The weight is carried directly over the feet, the crampon points stamped firmly into the ice with each step, with the ankles and knees flexed to allow boot soles to remain parallel to the slope.

(a) On gentle slopes, the climber walks straight up the hill. Normally, the feet are naturally flat to the slope and the axe is used as a cane. If pointing the toes uphill becomes awkward, they may be turned outward in duck-fashion. As the slope steepens, the body is turned to face across the slope rather than up it. The feet may also point across the slope, but additional flexibility and greater security are gained by pointing the lower foot downhill. The axe is used only to maintain balance and may be carried in the cane position or the arrest grasp with either the pick or point touching the slope. (Movement is diagonal rather than straight upward and the climber takes advantage of terrain irregularities and graded slopes. Changes in direction are done as in step-kicking on snow by planting the downhill foot, turning the body toward the slope to face the opposite direction, and stepping off with the new downhill foot.)

(b) On gentler slopes, the flat-footed approach is used throughout, but it is more secure and easier on steeper slopes to initiate the turn by kicking the front points and briefly front-pointing through the turn. At some point in the turn, the grip on the axe must be reversed. The exact moment for this depends on the climber and the specific situation. However, the climber's stance must be secure when the third point of support—the axe is temporarily relinquished.

(c) On steep slopes, which approach the limit of practical use of this style of ascent, the climber relies on the axe for security of a hold as well as for balance. The axe is held in the arrest grasp with one hand on the head and the other on the shaft, above the point. The well-sharpened pick is planted firmly in the ice at about shoulder height to provide one point of suspension while a foot moves forward and the crampons are stamped in (figure 22-59).

(d) Descent follows the same general progression of foot and axe positions; descend the fall line, gradually turning the toes out as the slope gets steeper. As the slope steepens, widen the stance, flex the knees, and lean forward to keep weight over the feet, and finally, face sideways and descend with the support of the axe in the arrest position. On very steep or hard ice, it may be necessary to face the slope and front-point downwards. When flat-footing downhill, all crampon points should be stamped firmly into the ice. It may be necessary to strive to take small steps which allow the climber to maintain balance during moves; long steps require major weight shifts to adjust balance.

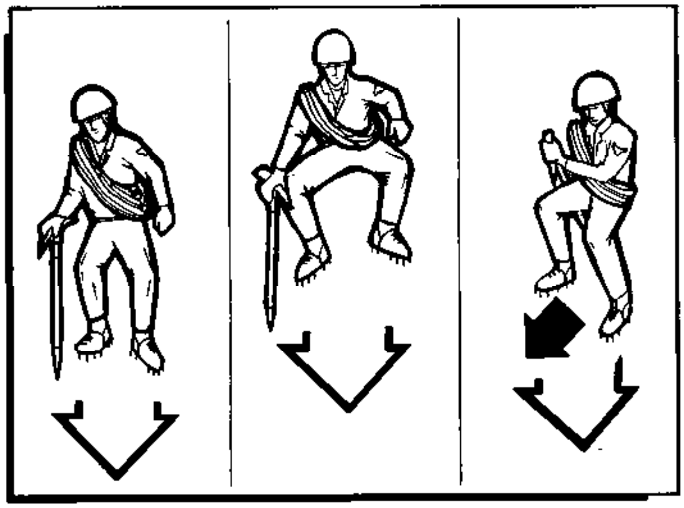

Figure 22-59. Descent with Crampons.

e. Anchors. Snow and ice conditions require the use of special devices for establishing belay anchors or placement of intermediate protection during a climb.

(1) Snow Pickets. Three- to four-foot lengths of aluminum "T" or tubular sections perform as long pitons and are suited for belaying. They must always be used in pairs or greater numbers, one anchoring the other (figure 22-60).

(2) Snow Fluke. A 12-inch piece of metal is buried with a runner coming to the surface; this can pull out if snow conditions are not just right. Better resistance to pull-out is gained with a large flat piece of metal driven into the snow surface at an angle, acting in the same manner as the fluke of an old-fashioned anchor. There is no danger of the runner being cut or weakened from wet conditions with the attachment of a wire cable. The softer the snow, the larger the size of the plate. When using flukes, it is very important that the proper angle with the surface be maintained; otherwise, wire, instead of becoming stronger (going deeper) when pulled, will become weaker (surface). Additionally, the cable may

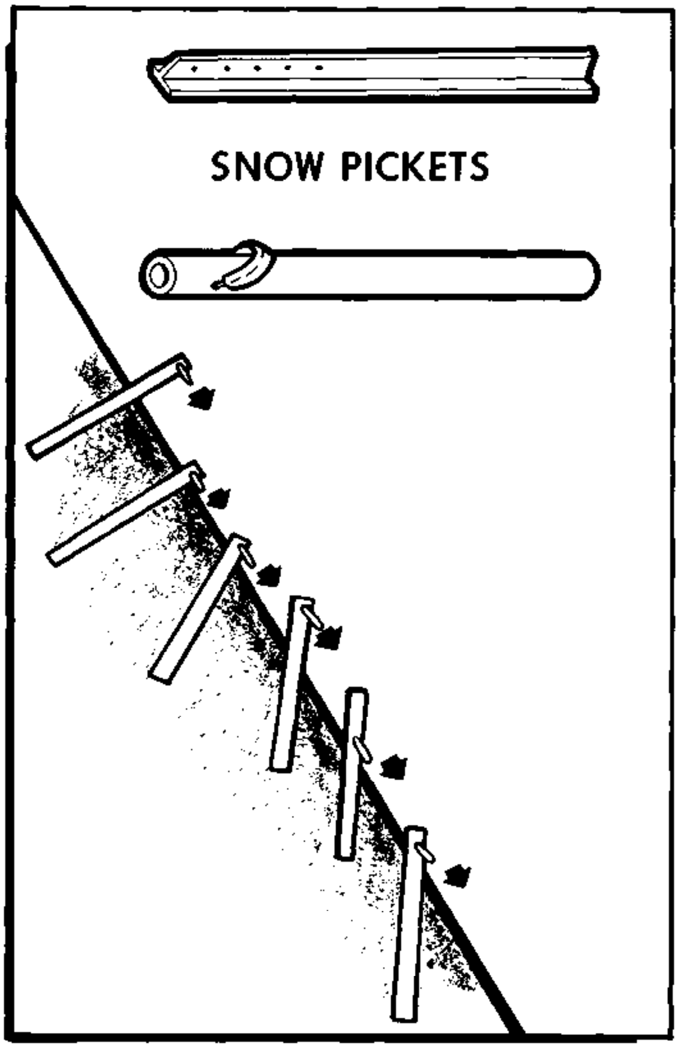

Figure 22-60. Snow Pickets.

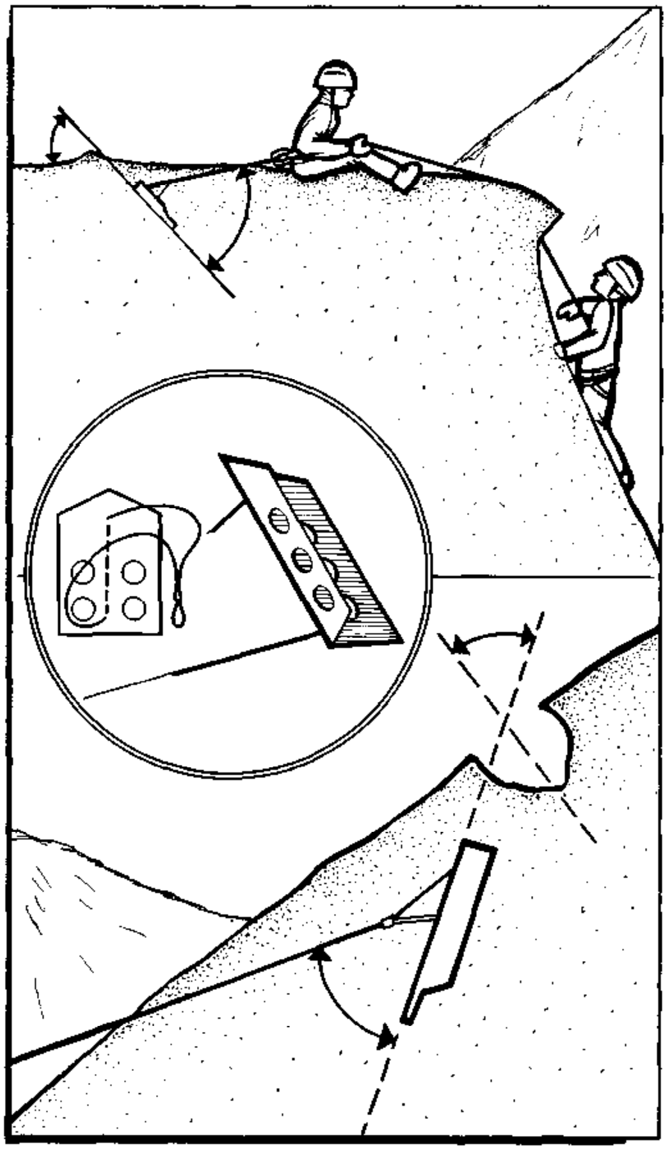

Figure 22-61. Snow Flukes.

act as a lever arm on hard snow, causing the fluke to pop out. This can be prevented by carefully cutting a channel in the snow for cables so that the pull comes directly at the plate. If attention is paid to placement, snow flukes will provide great security as belay and rappel anchors. A properly placed fluke is secure for a sitting or standing hop belay on snow (figure 22-61).

(3) Ice Screws:

(a) Tubular screws are very strong and are the most reliable (figure 22-62). They are difficult to place in hard or water ice since they tend to clog and have a large cross section. Their main advantage is that they minimize "spalling" (a crater-like splintering of the ice around the shaft of the screw) by allowing the displaced ice to work itself out through the core of the screw. If the core of the ice remaining in the screw is frozen in place, it jams the screw in subsequent placements. The ice may be removed by pushing with a length of wire or by heating with a cigarette lighter. This type of screw requires both hands for placement; however, once it is started, the pick of an ice hammer or axe inserted in the eye allows the climber to gain the advantage of leverage. Removal is easy and melt-out is slow due to the large cross section.

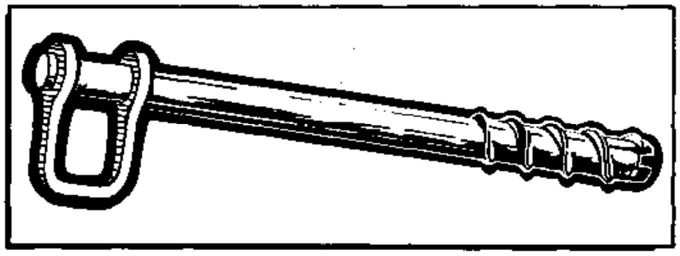

Figure 22-62. Tubular Ice Screw.

(b) Heavier "coathanger" type screws can be relied upon to stop a fall (figure 22-63). They are easier to start in hard ice than tubular screws and can often be placed with one hand, although it may be necessary to tap them while twisting as they are started. Their holding power is less than tubular screws as they tend to fracture hard ice and, under heavy loads, tend to shear through the ice because of their small cross section.

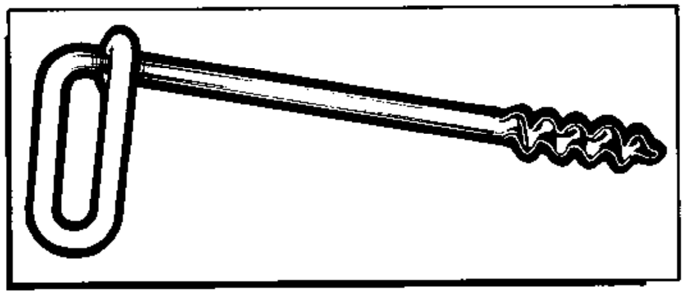

Figure 22-63. Coathanger Ice Screw.

(c) Developed as an attempt to make an easy-to-place and easy-to-remove screw, the solid screws are driven in like a piton and screwed out (figure 22-64). They offer excellent protection in water ice but are less effective in other ice forms. Melt-out is sometimes rapid because of limited thread displacement and, under load, they tend to shear through the ice as do coathanger screws.

(d) Before placement of ice screws or pitons, any soft snow or loose ice should be scraped or chopped away until a hard and trustworthy surface is reached. A small starting hole punched out with the pick or spike of the axe or hammer facilitates a good grip for the starting threads or teeth. The screw is pressed firmly into the ice and twisted in at the same time, angled slightly uphill against the anticipated direction of pull. Ice pitons are, of course, driven straight in, but must also be angled against the pull that would result from a fall (figure 22-65). If any spalling or splintering of the ice occurs, the screw should be removed and another placement tried 1 or 2 feet away. Some glacier ice will spall near the surface but by continuing to place the screw and gently chopping out the shattered ice, a deep, safe placement may be obtained. As a general rule, short screws or pitons should be used in hard ice and long ones in softer ice. They should always be placed in the ice until the eye is flush with the surface. When removing ice hardware, take care not to bend it since this diminishes its effectiveness in future use.

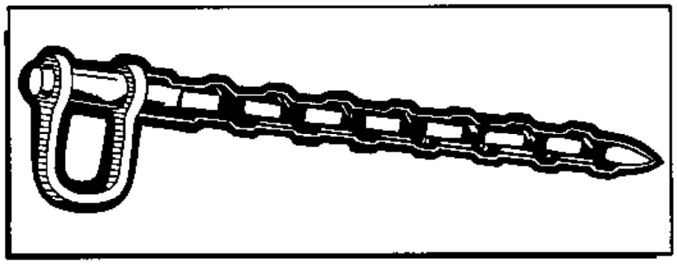

Figure 22-64. Solid Ice Screw.

(4) Ice/Snow Bollard. Although not a natural anchor in itself, an ice or snow bollard is easily made from natural materials. A semicircular trench is dug in the snow or ice. The trench should be 3 to 4 feet across and 6 to 12 inches deep. Allow a larger size for poor snow or ice conditions. The rope can be positioned in the trench to provide a downward belay (figure 22-66).

f. Glissading. Glissading is a means of rapidly descending a slope. Consisting of two basic positions, glissading offers a speedy means of travel with less energy exerted than using the descending step-by-step or plunging techniques.

(1) When snow conditions permit, the sitting glissade position is the easiest way to descend. The climber simply sits in the snow and slides down the slope while holding the axe in an arrest position (figure 22-67). Any tendency of the body to pivot head downwards may be checked by running the spike of the axe rudder-like along the surface of the snow. Speed is increased by lying on the back to spread the body weight over a greater area and by lifting the feet in the air. Sitting back up and returning the feet to the snow surface reduces speed. On crusted or firmly consolidated snow, sit fairly erect with the heels drawn up against the buttocks and the boot soles skimming along the surface. Turns are nearly impossible in a sitting glissade; however, the spike, dragged as a rudder and assisted by body contortions, can effect a change in direction of several degrees. Obstructions on the slope are best avoided by rising into a standing glissade (figure 22-67) for the turn, and then returning to the sitting position. Speed is decreased by dragging the spike and increasing pressure on it. After the momentum has been checked by the spike, the heels are dug in for the final halt but not while sliding at a fast rate as the result is likely to be a somersault. Emergency stops at high speeds are made by arresting.

(2) The standing glissade is similar to skiing. Positioned in a semicrouch stance with the knees bent as if sitting in a chair (figure 22-67), the legs are spread laterally for stability, and one foot is advanced slightly to anticipate bumps and ruts. For additional stability, the spike of the axe can be skimmed along the surface, the shaft held alongside the knee in the arrest grasp, with the pick pointing down or to the outside away from the body. Stability is increased by widening the spread of

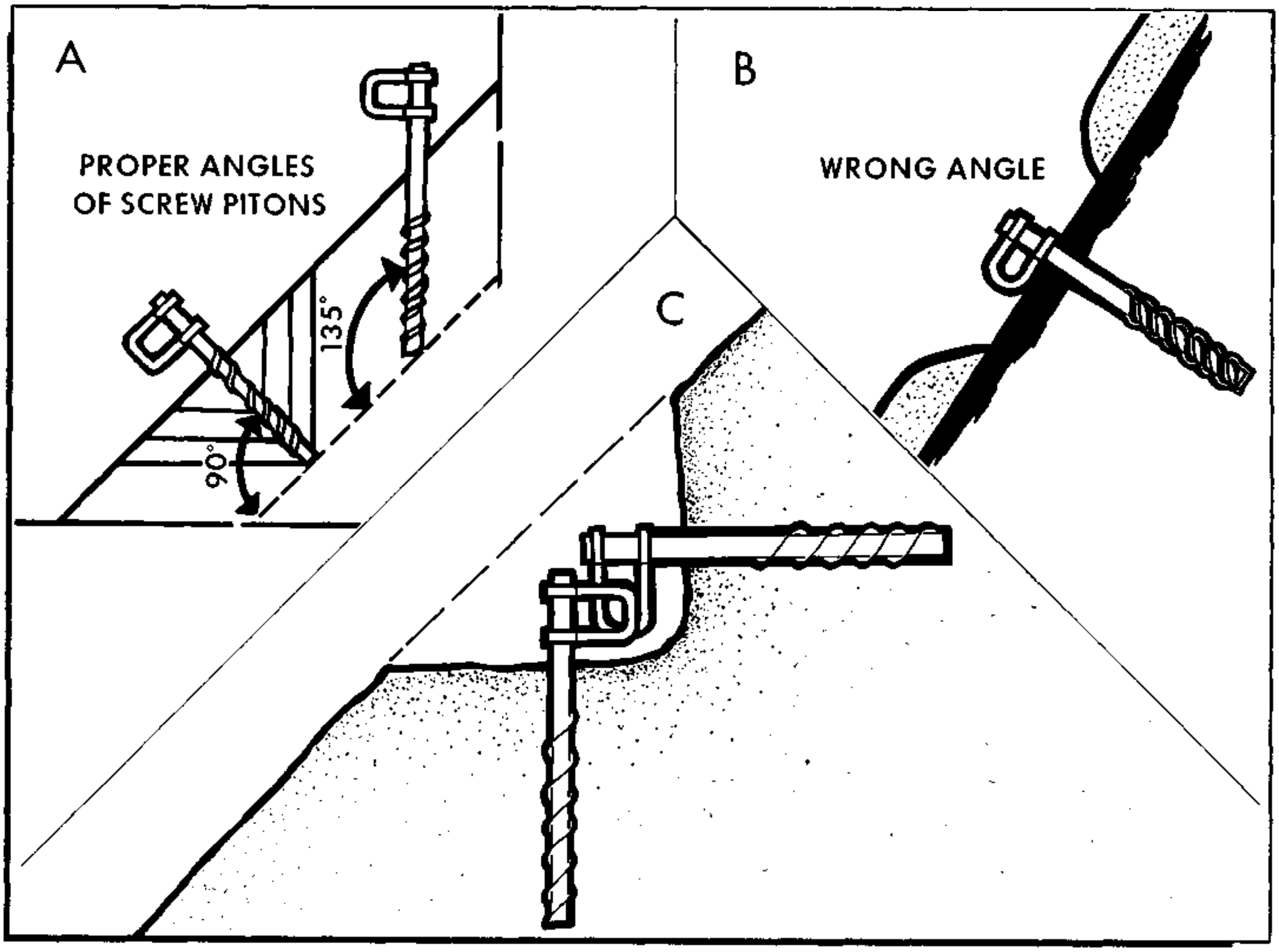

Figure 22-65. Placement of Ice Screw and Piton.

the legs, deepening the crouch, and putting more weight on the spike. A decrease in speed increases muscular strain and the technique becomes awkward and trying, although safe. Speed is increased by bringing the feet close together, reducing weight on the spike, and leaning forward until the boot soles are running flat along the surface like short skis. If the slide is too shallow, a long skating stride helps.

(3) A glissade should be made only when there is a safe runout. Unless a view of the entire run can be obtained beforehand, the first person down the run must use extreme caution, stopping frequently to study the terrain ahead. Equipment must be adjusted before beginning the descent. Crampons and other hardware must be properly stowed. Never attempt to glissade while wearing crampons as it is extremely easy to snag a crampon and be thrown down the slope. Mittens or gloves are worn to protect the hands and to maintain control of the axe. Heavy waterproof pants provide protection to the buttocks. Gaiters are also helpful for all glissading. Glissades should never be attempted in terrain where the axe safety cord is required. The hazards

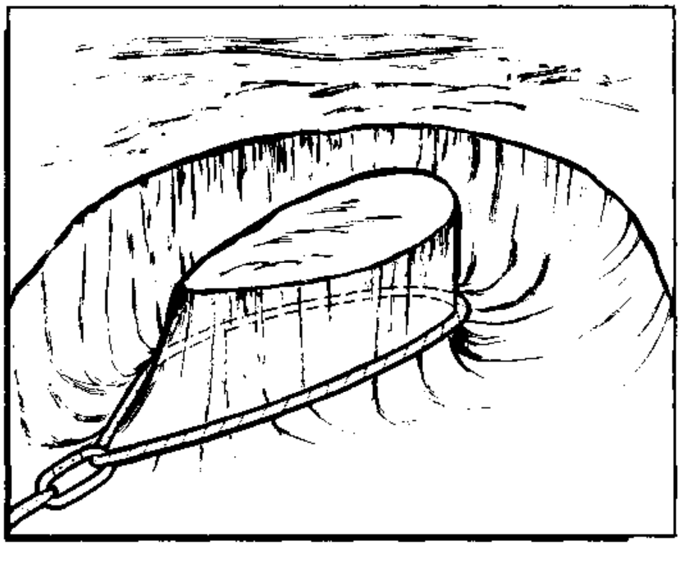

Figure 22-66. Ice/Snow Bollard.

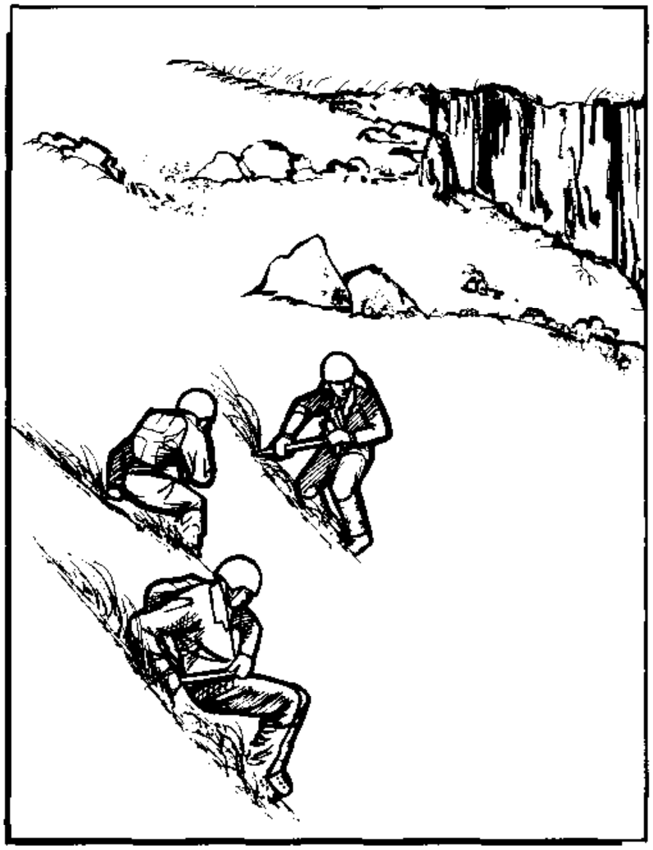

Figure 22-67. Glissading.

of a flailing ice axe should never be risked during a glissade.

22-14. Glaciers and Glacial Travel:

a. Features. To cope with the problems which can arise in using glaciers as avenues of travel, it is important to understand something of the nature and composition of glaciers.

(1) A valley glacier is essentially a river of ice and it flows at a rate of speed that depends largely on its mass and the slope of its bed. A glacier consists of two parts:

(a) The lower glacier, which has an ice surface void of snow during the summer.

(b) The upper glacier, where the ice is covered, even in summer with layers of accumulated snow that changes into glacier ice.

(2) To these two integral parts of a glacier may be added two others which, although not a part of the glacier proper, are generally adjacent to it and are of similar composition. These adjacent features, the ice and snow slopes, are immobile since they are anchored to underlying rock slopes. A large crevasse separates such slopes from the glacier proper and defines the boundary between moving and anchored ice.

(3) Ice is plastic-like near the surface, but not smooth enough to prevent cracking as the ice moves forward over irregularities in its bed. Fractures in a glacier surface, called crevasses, vary in width and depth from only a few inches to many feet. Crevasses form at right angles to the direction of greatest tension and due to a limited area, tension is usually in the same direction. Crevasses in any given area tend to be roughly parallel to each other. Generally, crevasses develop across a slope. Therefore, when traveling up the middle of a glacier, people usually encounter only transverse crevasses (crossing at right angles to the main direction of the glacier). Near the margins or edges of a glacier, the ice moves more slowly than it does in midstream. This speed differential causes the formation of crevasses diagonally upstream away from the margins or sides. While crevasses are almost certain to be encountered along the margins of a glacier and in areas where a steepening in gradient occurs, the gentlest slopes may also contain crevasses.

(4) An icefall forms where an abrupt steeping of slope occurs in the course of a glacier. These stresses are set up in many directions. As a result, the icefall consists of a varied mass of iceblocks and troughs with no well-defined trend to the many crevasses.

(5) As a glacier moves forward, debris from the valley slopes on either side is deposited on its surface. Shrinkage of the glacier from subsequent melting causes this debris to be deposited along the receding margins of the glacier. Such ridges are called lateral (side) moraines. Where two glaciers join and flow as a single river of ice, the debri on the adjoining lateral margins of the glaciers also unites and flows with the major ice stream, forming a medial (middle) moraine. (By examining the lower part of a glacier, it is often possible to tell how many tributaries have joined to form the lower trunk of the glacier.) Terminal (end) moraine is usually found where the frontage of the glacier has pushed forward as far as it can go; that is, to the point at which the rate of melting equals the speed of advance of the ice mass. This moraine may be formed of debri pushed forward by the advancing edge or it may be formed by a combination of this and other processes.

(a) Lateral and medial moraines may provide excellent avenues of travel. When the glacier is heavily crevassed, moraines may be the only practical routes. Ease of progress along moraines depends upon the stability of the debris composition. If the material consists of small rocks, pebbles, and earth, the moraine is usually loose and unstable and the crest may break away at each footstep. If large blocks compose the moraine, they have probably settled into a compact mass and progress may be easy.

(b) On moraine travel, it is best either to proceed along the crest or, in the case of lateral moraines, to follow the trough which separates it from the mountainside. Since the slopes of moraines are usually unstable, there is a great risk of spraining an ankle on them. Medial moraines are usually less pronounced than lateral moraines because a large part of their material is transported within the ice. Travel on them is usually easy but should not be relied upon as routes for long distances since they may disappear beneath the glacier surface. Only rarely is it necessary for a party traveling along or across moraines to be roped together (figure 22-68).

(6) Glacial rivers are varied in type and present numerous problems to those who must cross or navigate them. Wherever mountains and highlands exist in the arctic regions, melting snows produce concentrations of water pouring downward in a series of falls and swift chutes. Rivers flowing from icecaps, hanging piedmonts (lake-like), or serpentine (winding or valley) glaciers are all notoriously treacherous. Northern glaciers may be vast in size and the heat of the summer sun can release vast quantities of water from them. Glacier ice is extremely unpredictable. An ice field may look innocent from above, but countless subglacial streams and water reservoirs may be under its smooth surface. These reservoirs are either draining or temporarily blocked. Mile-long lakes may lie under the upper snowfield, waiting only for a slight movement in the glacier to liberate them sending their waters into the valleys below. Because of variations in the amounts of water released by the Sun's heat, all glacial rivers fluctuate in water level. The peak of the flood water usually occurs in the afternoon as a result of the noonday heat of the Sun on the ice. For some time after the peak has passed, rivers which drain glaciers may not be fordable or even navigable. However, by midnight or the following morning, the water may recede so fording is both safe and easy. When following a glacial river broken up into many shifting channels, choose routes next to the bank rather than taking a chance on getting caught between two dangerous channels.

(7) Glaciers from which torrents of water descend are called flooding glaciers. Two basic causes of such glaciers are the violent release of water which the glacier carried on its surface as lakes, or the violent release of large lakes which have been dammed up in tributary glaciers because of the blocking of the tributary valley by the main glacier. This release is caused by a crevasse or a break in the moving glacial dam; the water then roars down in an all-enveloping flood. Flooding glaciers can be recognized from above by the flood-swept character of the lower valleys. The influence of such glaciers is sometimes felt for many miles below. Prospectors have lost their lives while rafting otherwise safe rivers because a sudden flood entered by a side tributary and descended as a wall of white, rushing water.

(8) On those portions of a glacier where melting occurs, runoff water cuts deep channels in the ice surface and forms surface streams. Many such channels exceed 20 feet in depth and width. They usually have smooth sides and undercut banks. Many of these streams terminate at the margins of the glacier where in summer they contribute to the torrent that constantly flows between the ice and the lateral moraine. Size increases greatly as the heat of the day moves to an end. The greatest caution must be taken in crossing a glacial surface stream since the bed and undercut banks are usually hard, smooth ice which offers no secure footing.

(9) Some streams disappear into crevasses or into round holes known as glacial mills, and then flow as subglacial streams. Glacial mills are cut into the ice by the churning action of water. They vary in diameter. Glacial mills differ from crevasses, not only in shape but also in origin, since they do not develop as a result of moving ice. In places, the depth of a glacial mill may equal the thickness of the glacier.

b. Glacier Operations. The principal dangers and obstacles to operations in glacier areas are crevasses and icefalls. Hidden crevasses present unique problems and situations since their presence is often difficult to detect. When one is detected, often it is due to a team member having fallen through the unstable surface cover. The following techniques and procedures should be followed when performing glacier operations.

(1) Equipment Preparation. The prevention of hypothermia should be of primary importance when performing glacier operations. Sufficient protective clothing must be worn or carried to cover all climatic temperature variations. Climbers trapped in crevasses have died of hypothermia while their team members, helpless to assist from their position on the glacier surface above, were sweltering in sunshine. Backpacks should be equipped with a lanyard consisting of a 6-foot piece of line with a figure-eight and nonlocking carabiner at one end. The free end of the lanyard is attached to the pack and the unlocking carabiner is snapped into the buttock strap of the seat harness. If the climber falls into a crevasse and is suspended upside down by the weight of the pack, the pack can be released with the lanyard and the person can return to an upright position.

(2) Team Composition. The first law of glacial travel is to rope-in during travel. The principal consideration is to avoid crevasses. When stepping onto a known glacier or onto a snowfield of unknown stability,

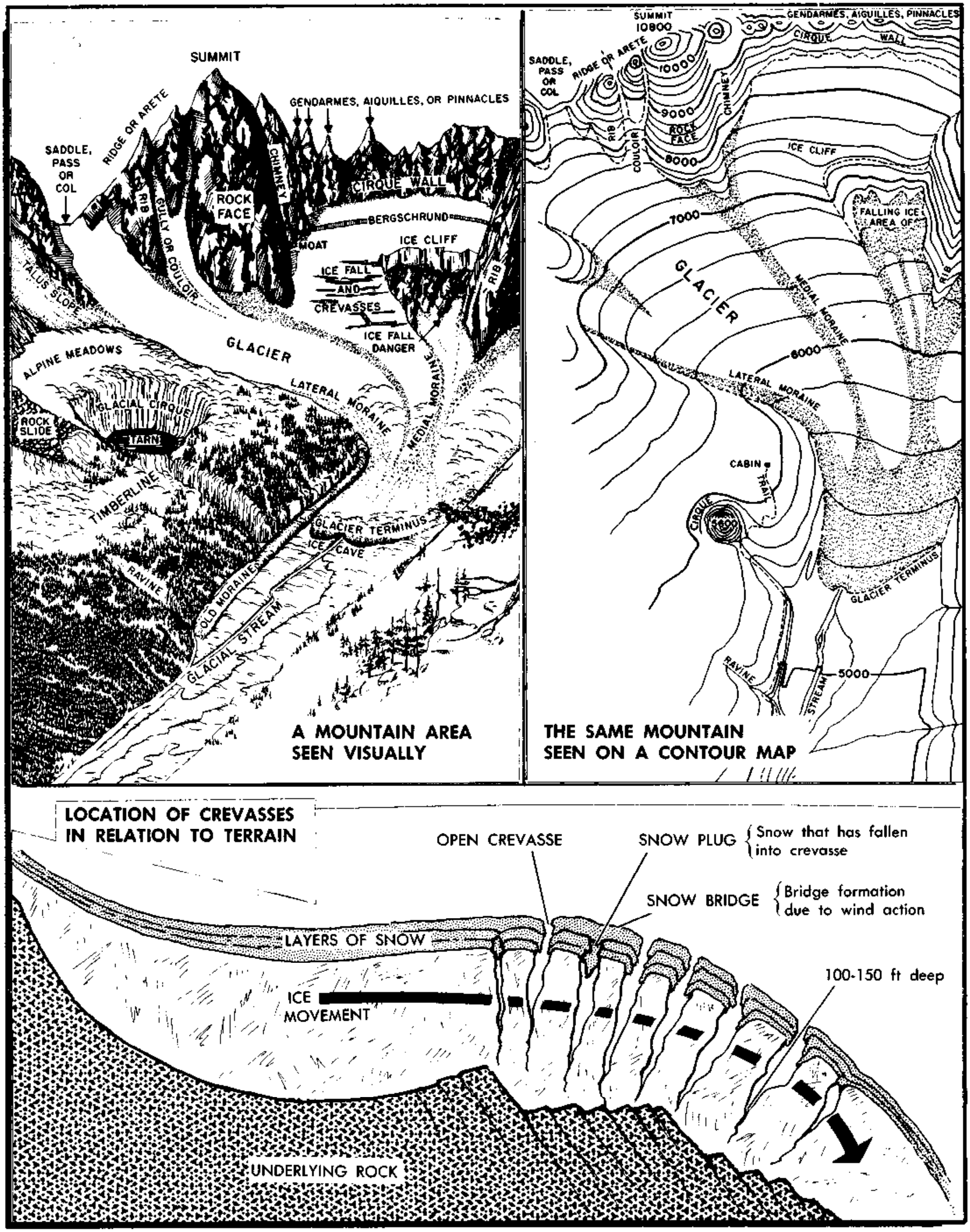

A MOUNTAIN AREA SEEN VISUALLY

THE SAME MOUNTAIN SEEN ON A CONTOUR MAP

LOCATION OF CREVASSES IN RELATION TO TERRAIN

SUMMIT
RIDGE OR ARETE
SADDLE, PASS OR COL
GENDARMES, AIQUILLES, OR PINNACLES
RIB
GULLY OR COULOIR
ROCK FACE
CHIMNEY
CIRQUE WALL
BERGSCHRUND
MOAT
ICE CLIFF
ICE FALL AND CREVASSES
ICE FALL DANGER
RIB
TALUS SLOPE
GLACIER
ALPINE MEADOWS
LATERAL MORAINE
MEDIAL MORAINE
ROCK SLIDE
GLACIAL CIRQUE
TARN
TIMBERLINE
GLACIER TERMINUS
ICE CAVE
RAVINE
OLD MORAINE
GLACIAL STREAM
SUMMIT 10800
GENDARMES, AIGUILLES, PINNACLES
CIRQUE
WALL
SADDLE, PASS OR COL
RIDGE OR ARETE
10000
9000
ROCK FACE
CHIMNEY
RIB
COULOIR
8000
ICE CLIFF
7000
FALLING ICE AREA OF
RIB
GLACIER
MEDIAL MORAINE
6000
LATERAL MORAINE
CABIN
TRAIL
CIRQUE
5600
GLACIER TERMINUS
RAVINE
STREAM
5000
OPEN CREVASSE
SNOW PLUG { Snow that has fallen into crevasse
SNOW BRIDGE { Bridge formation due to wind action
LAYERS OF SNOW
100-150 ft deep
ICE MOVEMENT
UNDERLYING ROCK

Figure 22-68. Typical Glacier Construction.

whether crevasses are visible or not, the law of roping-in remains. The only variable to this law is when avalanches present a greater hazard than the threat of crevasses. The most experienced climber in glacial travel should be the lead climber; however, if crevasses are completely masked, the lightest climber may lead. During moderate climbs, three climbers tied into a 165-foot rope is ideal. During severe climbs requiring belay, a 120-foot rope with only two climbers is recommended. If a two-person climbing team falls, the team must be arrested by a single axe. If a three-person climbing team is roped in, the rope is usually so shortened that if one climber falls, the others are often dragged in before they have time to react.

(3) Roping-In. Climbers are roped together by constructing figure-eight knots at the ends and middle of the rope. The rope is attached by passing a locking carabiner through the figure-eight and the crotch strap of the seat harness. Associated climbing equipment such as ice axes, slings, and packs are donned. When completely roped in and prepared for travel, there should not be less than 50 feet of rope between each of the climbers. The more rope between the climbers, the better the chance for a successful arrest.

c. Glacier Travel. Due to the difficulty of crevasse rescue, two or more rope teams are recommended for glacier travel since a single team is sometimes pinned down in the arrest position and members are unable to free themselves to begin rescue. Rope teams must travel close together to lend assistance to each other; however, not so close as to fall into the same crevasse. During extended periods on a glacier, skis and snowshoes are often of great value. This footgear will distribute the weight more widely than boots alone and place less strain on snow bridges. Neither skis nor snowshoes are substitutes for the rope, but may be used for easy travel.

(1) Operations in the mountains have certain limitations imposed by nature in glacial movement. Access to the end portion of a glacier may be difficult due to abruptness of the ice and possible presence of crevasses. Additional obstacles of mounting a glacier may be swift glacial streams or abrupt mountain terrain bordering the glacier ice. The same obstacles may also have to be negotiated when dismounting or mounting a valley glacier at any place along its course. Further considerations to movement on a glacier are steep sections, heavily crevassed regions, and icefalls. The use of up-to-date aerial photographs, when available, with aerial reconnaissance is a valuable means of gathering advance information about a particular glacier. The photos, however, only supplement and do not negate the advantages of surface reconnaissance conducted from available vantage points.

(2) Trail wands are used to mark the route and crevasses. The wands, especially essential to safety during periods of adverse weather, are placed every 150 feet along the route and can be used during day or night. A climbing team should not cluster close together during rest stops. If areas of safety cannot be found, the rope must be kept extended during rests just as during travel. A party establishing camp on a snow-covered glacier similarly remains roped-in for as long a period as required to safely inspect the area by stomping and probing the surface thoroughly before placing trust in the site. Hidden crevasses should always be assumed to exist in the area.

(3) Normally a team will travel in single file, stepping in the leader's footsteps or in echelon formation (figure 22-69). If a crevasse pinches out, an end run must be made (figure 22-70) even if it involves traveling half a mile to gain a few dozen feet of forward progress. The time taken to walk around is generally much less than in forcing a direct crossing. Important to remember in an end run is the possible hidden extension of a visible crevasse. A frequent error is aiming at the visible end. Unless the true or subsurface end is clearly visible during the approach, it is best to make a wide swing around the end.

(a) In late summer, the visible end is often the true end due to surface snow and ice having melted. When end runs are impractical because of the distance involved or because the end of one crevasse is adjacent to another, snow bridges may provide a crossing point. One kind consists of remnant snow cover sagging over an inner open space. Another kind, with a foundation which extends downward into the body of the glacier, is less a bridge than a solid area between two crevasses.

(b) Any bridge should be closely and completely examined before use. If overhanging snow obscures the bridge, the lead climber must explore at closer range by probing the depth and smashing at the sides while walking delicately, ready for an arrest or sudden drop. The second climber establishes a belay (figure 22-71) anchored by the third climber who is also prepared to initiate rescue if the leader falls. An excessively narrow or weak bridge may be crossed by straddling or even slithering on the stomach, thereby lowering the center of gravity and distributing the weight over a broader area. When there is doubt about the integrity of a bridge, but it is the only possible route, the lightest climber in the team should be the first across, with the following climbers walking with light steps and taking care to step exactly in the same tracks.

(c) Bridges vary in strength with changes in temperatures. In the cold of winter or early morning, the thinnest and most fragile of bridges may have incredible structural strength. However, when the ice crystals melt

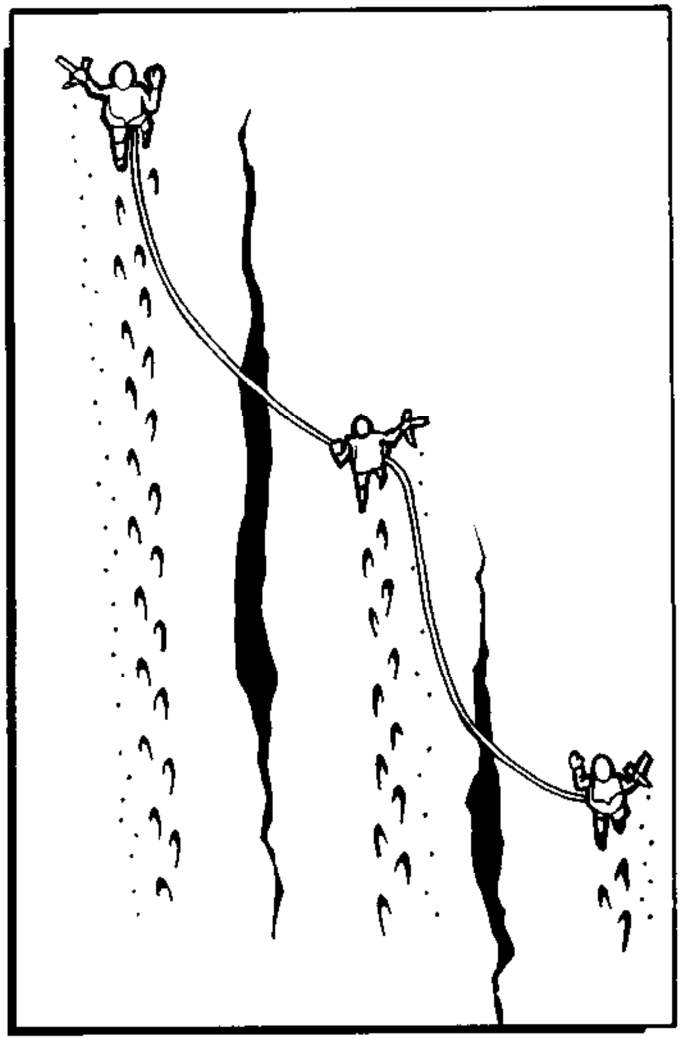

Figure 22-69. Echelon Formation.

in the afternoon temperature, even the largest bridge may suddenly collapse. Each bridge must be tested with care, being neither abandoned nor trusted until its worth is determined (figure 22-71).

(d) Narrow cracks in a bridge can be stepped across, but wider crevasses require jumping. If the jump is so long that a run is required, the approach should be carefully packed. A running jump (figure 22-72) can carry the climber further than a standing jump, although running jumps are not often practical. Most jumps are made with only two or three lead-up steps. In any case, care must be taken to locate the precise edge of the crevasse before any attempt is made to jump. Encumbering clothing and equipment must be removed before the jump, although the jumper must bear in mind the low temperature which often exists within crevasses.

d. Crevasse Rescue. Each climber must be able to effect a crevasse rescue if a team member falls into a crevasse. When a climber falls, the remaining team members must drop into a self-arrest position and stabilize their positions. All climbers should never be dragged into the crevasse. If a climber falls, the remaining team members must support the weight until one of them can establish a reliable anchor point or until the second team arrives to help. If the fallen climber is able to assist the recovery, self-extraction from the crevasse may be performed by using prusiks.

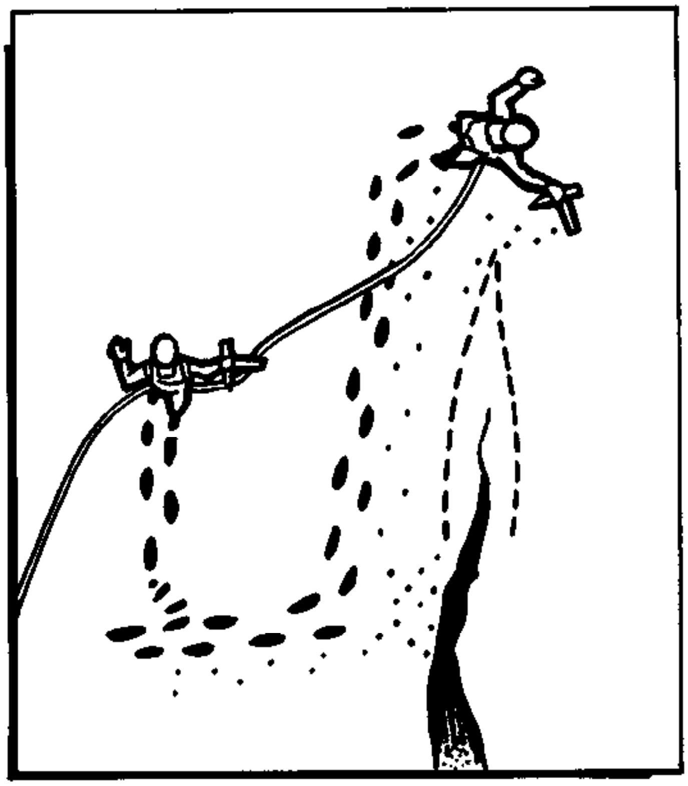

Figure 22-70. End Run.

(1) A problem inherent to crevasse rescue is the imbedding of the rope (caused by the fallen climber's weight) in the ice and snow. Unless the rope is buffered with an ice axe during the climb out, it will tend to entrench itself deeper into the ice, eventually creating a deep groove in which it will be extremely difficult to use and retrieve, or it will freeze in place, rendering it useless. Corrective actions are to travel down along the rope, taking care not to drop debris on the climber, and free it from the ice. An additional method is to drop a spare rope down to the climber who shifts weight off the imbedded rope until it can be freed.

(2) If the climber is using prusiks and the action of the climb seesaws the rope into the tip of the crevasse, it will be extremely difficult to ascend the few remaining feet. A procedure to overcome this situation is for the climber to tie into the rope near the prusik. The climber then strikes the figure-eight knot from the harness and sends the end to the team above via a retrieving line.

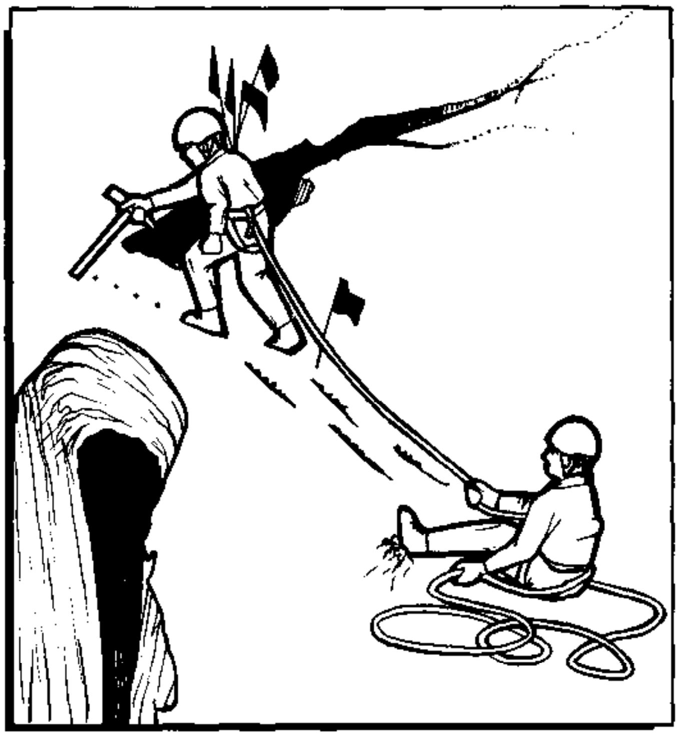

Figure 22-71. Crossing Bridged Crevasse.

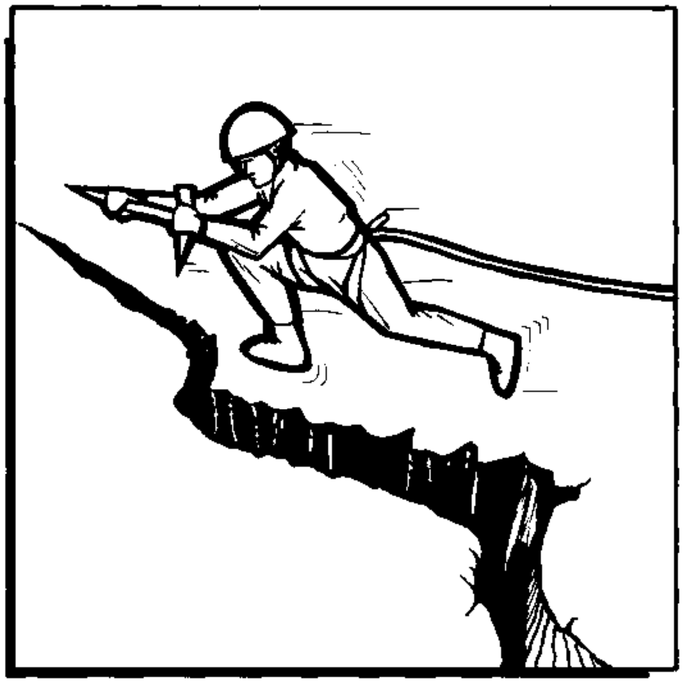

Figure 22-72. Jumping a Crevasse.

Once firmly anchored in place, the rope affords a viable route of ascent. Negotiating the final few feet of a crevasse is usually difficult due to the pressure of the rope against the lip or side of the wall. Prusiks tend to compress against or gouge into the wall rendering them nonfunctional. In most cases, the final few feet are overcome by brute strength. If a second rope is available, an alternate method can be used (figure 22-73).

(3) If a fallen climber is unable to help in the recovery, another climber may be required to enter the crevasse. Before the team member is lowered, all assurances must be determined that the assistance will enhance the outcome of the operation and not compound it. The rescuer should administer medical treatment as required, paying special attention to preventing or treating cold weather injuries as the interior of the crevasse can become extremely cold. Warm protective clothing must be used if the medical situation does not permit immediate extraction.

22-15. Evacuation Principles and Techniques. The performance of mountain rescue is not only physically demanding, but also mentally challenging. Hard and fast procedures to fit all circumstances for mountain rescue cannot be established. The team's ability to innovate will, in most instances, allow the adaptation of the basic principles, techniques, and procedures into a system suitable to effect the rescue. A normal rescue system will use anchors, belays, and various specialized systems. There are basically two methods of rescue: bringing the victim (patient) up to the rescuer's position, or evacuating the victim down from the position.

a. Safety. The establishment of rescue systems must be thoroughly tested prior to use. One missed step in setting up a rescue system may result in further injury of the victim and (or) injury to the rescuers.

b. Evacuations. Evacuation of a victim from the position in a downhill direction is an easier task than establishing a mechanical leverage for pulling a victim to the top of a hill. The victim's medical condition will dictate the method of evacuation and equipment used. The primary litter used is the Stokes litter. This is a tubular frame litter with a wire basket. The tubular main bar provides a very strong framework for mountain operations. The patient may be secured in the litter by means of several cross-body straps, or by the interlacing of slings. When an evacuation team arrives on the scene, there are two activities which should take place at the same time: (1) The patient should be treated and prepared for transporting in the Stokes litter, and (2) the anchor and mechanical brake system must be established. Within obvious restraints of time and distance, low-angle evacuations are always preferable to high-angle evacuations. Low-angle work eliminates many hazards and requires a lesser degree of knowledge and skill. Using the Stokes litter eliminates excessive knot tying and lashing essential on other litters.

c. Preparing the Braking System. The following describes the preparation of the brake system for a low-

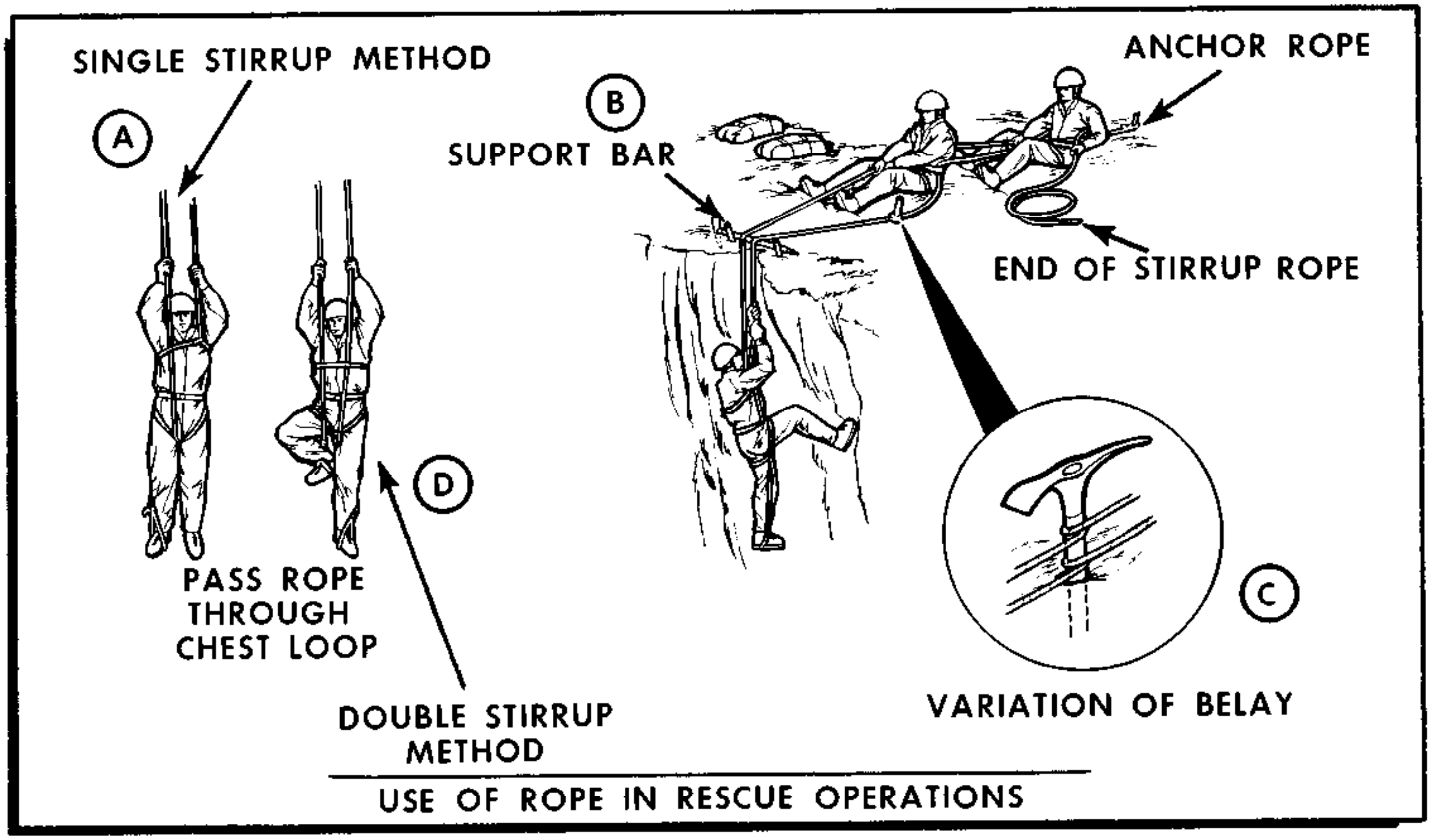

Figure 22-73. Two-Rope Crevasse Rescue.

angle litter evacuation. Establish a very sound anchor, which will be in the direction of pull for the first pitch. If a sound anchor is not available, establish an anchor system. If the terrain does not allow the brake operator ample room to safely and effectively perform the required tasks, than an 11 mm (seven-sixteenth inch) rope sling or double tape sling can be used to adjust for the distance from the anchor to the area where the brake operator will work. The mechanical braking device is securely attached to the anchor or sling. This braking device should either be a figure-eight rappel ring, or a four-carabiner brake system. The rope to be used for the litter descension should be backcoiled. If the rope is to be used for ascending, the rope should lay out along the route of ascension. The head of the litter will be attached to the figure-eight at the end of the rope with a steel-locking carabiner. If locking carabiners are not available, the rope should be tied directly to the litter using several round turns on the outer rail at the head of the Stokes litter and tied off with a bowline and a safety knot. The rope is then properly locked into the mechanical braking system.

d. Preparing the Patient for Transport. While the rope and brake system are being prepared, the litter and patient should also be prepared. The litter must be secured to prevent its loss or further injury to the patient. Additionally, the litter may be padded or insulated (blankets or foam pads) for protection. The ties for securing the feet and pelvis should be attached to the litter. Before evacuating, all emergency medical treatment appropriate to the situation should be performed (splinting fractures, maintaining an open airway, etc.). The patient should be insulated from environmental conditions such as cold, wind, or rain. The person in charge of the patient's medical condition should ensure that the patient's condition is stable enough for transporting. In mountainous terrain, the patient should be protected from further injury due to rockfall by wearing a helmet at all times. A litter team generally consists of four to six people. Fewer than six cannot withstand the fatigue of frequent or long trips while carrying an injured person.

e. Three- or Four-Man Lift. Three bearers take up positions on one side of the victim, one at the shoulder, one at the hip, and one at the knees. If one side is injured, the three bearers should be on the uninjured side. A fourth bearer, if available, takes a position on the opposite side, at the victim's hip.

(1) The bearers should kneel next to the victim. Then, simultaneously, the bearer at the victim's shoulder puts one arm under the victim's head, neck, and

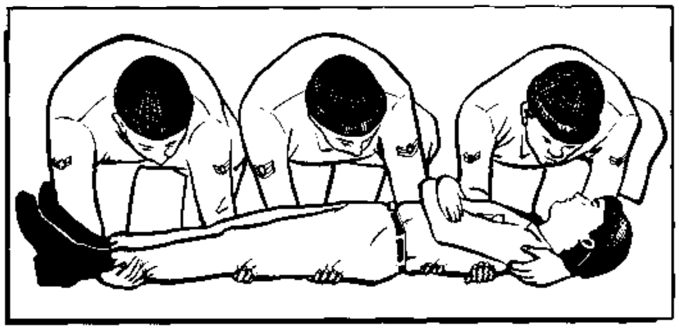

Figure 22-74. Lifting the Patient.

shoulder, and the other under the upper part of the victim's back. Each bearer at the victim's hips places one arm under the victim's back and the other under the victim's thighs. The bearer at the victim's knees places one arm under the victim's knees and the other under the ankles (figure 22-74).

(2) The person at the victim's head gives all the commands. The command "prepare to lift!" is followed by the command "lift." Immediately, all the bearers lift simultaneously and place the victim in line on their knees. If the victim needs to be moved any distance to the litter, move as shown in figure 22-75.

(3) The fourth bearer, if available, places a stretcher under the victim and against the toes of the three kneeling bearers. The command "Prepare to lower!" is followed by the command "Lower!" and the victim is gently lowered to the litter. Once properly positioned in the litter, the victim must be secured in a manner to prevent further injury. The victim may be secured to the litter in a variety of ways depending upon the evacuation route and the victim's condition.

f. Securing Patient in the Litter:

(1) The tape sling used to secure the feet is tied to the framework of the Stokes litter which separates the legs near the groin area. The tape sling should be tied with a clove hitch in the middle of the tape in a manner to prevent the tape from sliding down to the feet when pulled tight. The feet are secured by running the tape across the legs to the window on the outside of the litter then across the patient's legs to the feet. An overhand knot is made in each tape which can be passed over the corresponding foot. When the feet are secured, there should be ample room to apply tension to the head if needed. The tape is then tied at the foot of the litter to a major support bar on the inside of the litter frame. The ties on a Stokes litter should never be made on any outside rail as they are subject to abrasion. The tape slings should be tied off with a two-round turn and two half hitches. If the two-round turn does not hold tension, then a clove hitch can be used in its place.

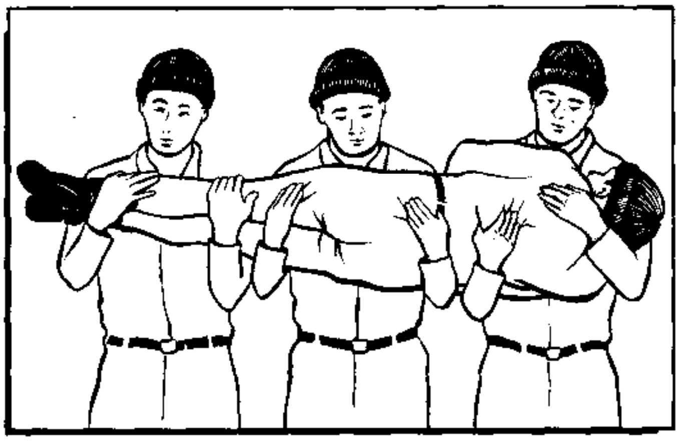

Figure 22-75. Moving the Patient.

(2) The tape used to secure the pelvis should be tied just above the tape used to secure the feet, and secured in the same fashion. Each end of the rope is passed over the leg to the larger upright cross-member of the Stokes litter between the outer rail and inner basket rail. This cross-member corresponds with the side of the hip. The tape is secured with a two-round turn and two half hitches or a clove hitch and two half hitches. The ends of the tapes are then tied together at the middle of the patient's waistline with a square knot and two half hitches on either side of the knot.

(3) The upper torso is secured by placing the middle of the tape in the center of the patient's chest and the two ends of the tape are secured to the large upright cross-member. The running ends of the tape are then passed diagonally across the patient to the cross-member which is next to the abdomen. The tape is secured again and the ends are tied at the midline of the body. The head is secured by running a tape sling over the helmet and securing the tape at the corresponding cross-members. The helmet can be used with a tape sling to provide traction; however, it is not a substitute for the neck collar (figure 22-76).

(4) Once the patient and the system are ready for the low-angle evacuation, the entire system must be doublechecked. Once the litter and patient are prepared as described, ascent or descent is made through a team of litter bearers and a belay point. The minimum essential members for a low-angle evacuation are five—a belayer and four litter bearers (figure 22-77).

g. Low-Angle Evacuation Descent. On the descent, the belay rope attached to the head of the litter runs through a mechanical belay brake system. One team member acts as the belayer. This secures the litter and aids in lowering. Another person may assist the belayer with the rope. Litter bearers take their positions on the litter, ideally three on each side. The chief medic and crew boss should be among these to ensure the patient is

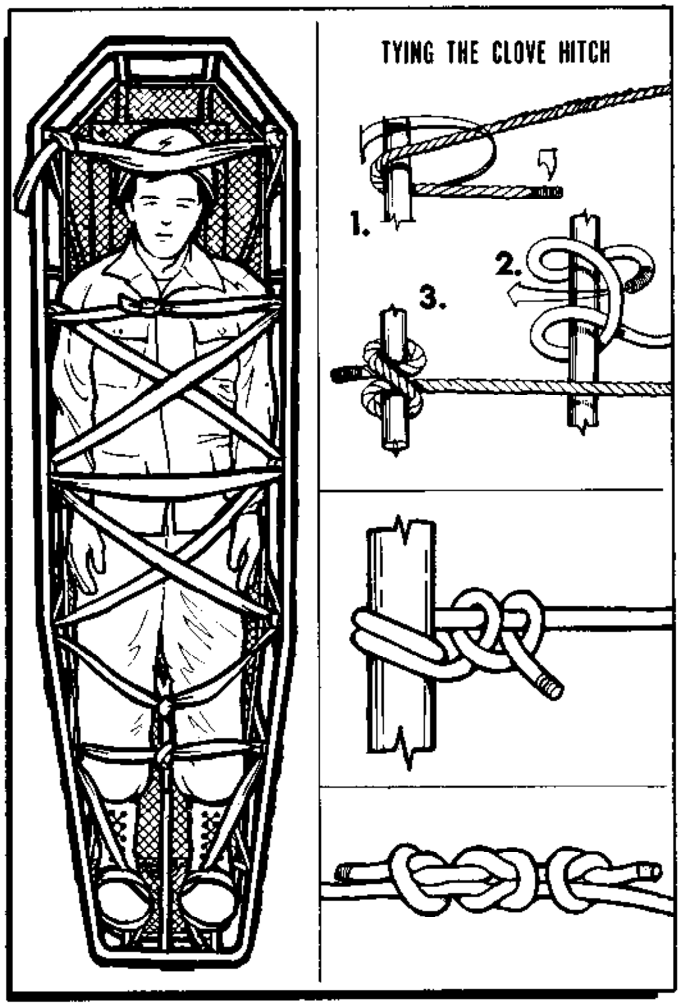

Figure 22-76. Tying the Patient into a Stokes Litter.

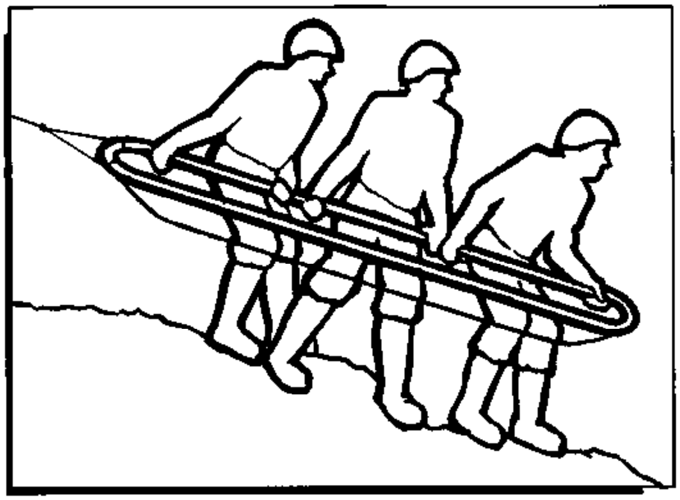

Figure 22-77. Carrying the Litter.

monitored 100 percent of the time and that effective communication exists between the litter bearers and the belayer.

(1) In descending, the most direct, practical passage which takes advantage of available trees and rocks for belay points should be used. Communication is made through a series of commands. As litter bearing is rapidly exhausting, team members should alternate roles. Additionally, a sling attached with a girth hitch to the litter may be used to transfer some of the weight from the arm to the skeletal system via the shoulder. It is also advantageous to use the belay system brake by leaning forward, thus reducing the amount of lifting required.

(2) The scout may precede the team to pick a trail, make the passage more negotiable, or make a reconnaissance so the team need not retrace its course if an impasse is encountered. The scout can also select the site for and secure the next anchor. The scout must remember that the anchors and belay stations must be less than 140 feet apart (with 150-foot rope). Any time the route of descent or ascent changes course more than 90 degrees, a new anchor and brake system must be established or a runner used to change the direction of pull from the belay to the litter. In addition, if the rope is binding against vegetation or rock formations, a sling with a carabiner should be used to pull the rope away from the obstacle.

(3) Once all rescuers have been assigned positions and understand their responsibilities, the litter can be moved. The crew boss will count to three and give the command to lift. The crew boss should coordinate the litter bearers' activities while moving over rough terrain. Standardized commands between the crew boss and belayer are used to control the rate of descent. The following are the commands which should be used:

(a) "ON BELAY?" - Crew boss.

(b) "BELAY ON" - Belayer.

(c) "ROPE" - The crew boss is telling the belayer to feed the rope out in feet.

(d) "SLOW ROPE" - The crew boss is telling the belayer to feed the rope out in inches.

(e) "BRAKE" - Can be said by anyone to avoid a fall or obstacles.

(f) "LITTER SECURE" - The crew boss is telling the belayer that the litter will not be moving and it is attached to the anchor.

(g) "OFF BELAY" - The crew boss is telling the belayer to break the belay system and begin moving equipment to the next station.

(4) When ascending a steep slope, the procedures described in the descent are generally reversed. Additional manpower is required to pull the belay rope through a four-carabiner brake system or figure-eight clog while the litter bearers lift and slowly climb. One person is required to operate the brake system. As this procedure is considerably more fatiguing than descending, the litter bearers should not try to do all the work.

However, as they are ascending the slope, they should assist and hold the litter off the ground.

h. The Buddy System. The buddy system is an evacuation for a slightly injured patient or a patient who is incapable of getting off a precipice. Equipment required is two climbing ropes, four slings, chest, and seat harness for the rescuer, optional seat harness for the patient, and sufficient equipment to construct an anchor system. The following steps are necessary:

(1) An anchor system is set up with a belay device. The first rope is backcoiled as a belay line. A figure-eight knot is tied in the end of the rope with a fisherman's safety.

(2) The second rope is coiled with large enough coils to fit around the shoulders of the patient and the rescuer (figure 22-78). Once coiled, the loops are divided into two groups so that a figure eight is formed. The patient steps into the divided coil so that each leg is through one-half of the figure eight. The knot securing the coil should be in the small of the patient's back, and the coils should be beneath the patient's arms.

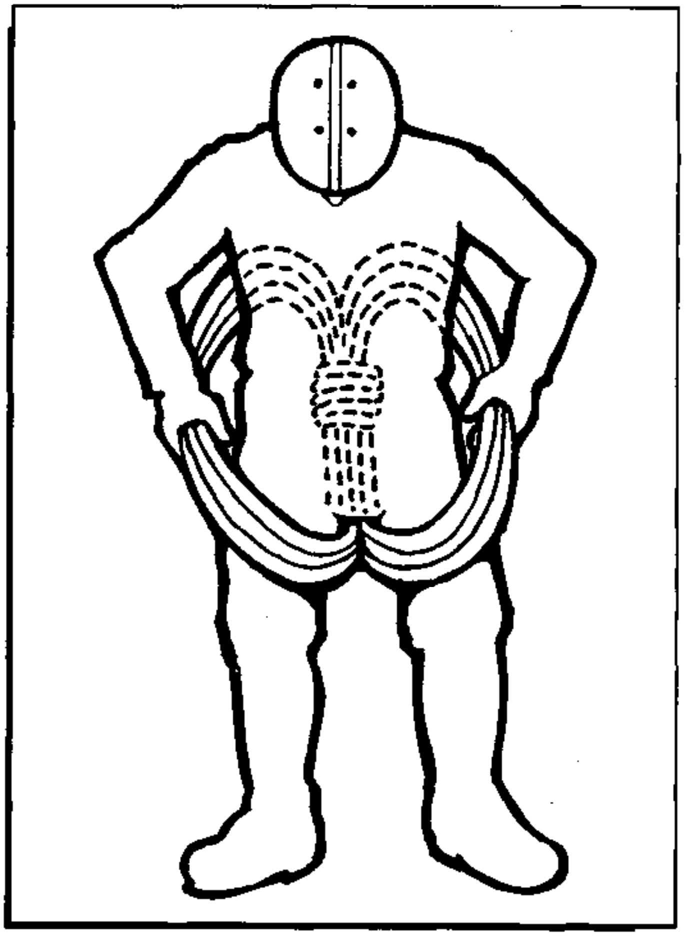

Figure 22-78. Preparing Patient for Buddy Evacuation System.

(3) The rescuer then stands in front of the patient and places each arm through the loops of the coil, half over the right shoulder and the other half over the left shoulder (figure 22-79).

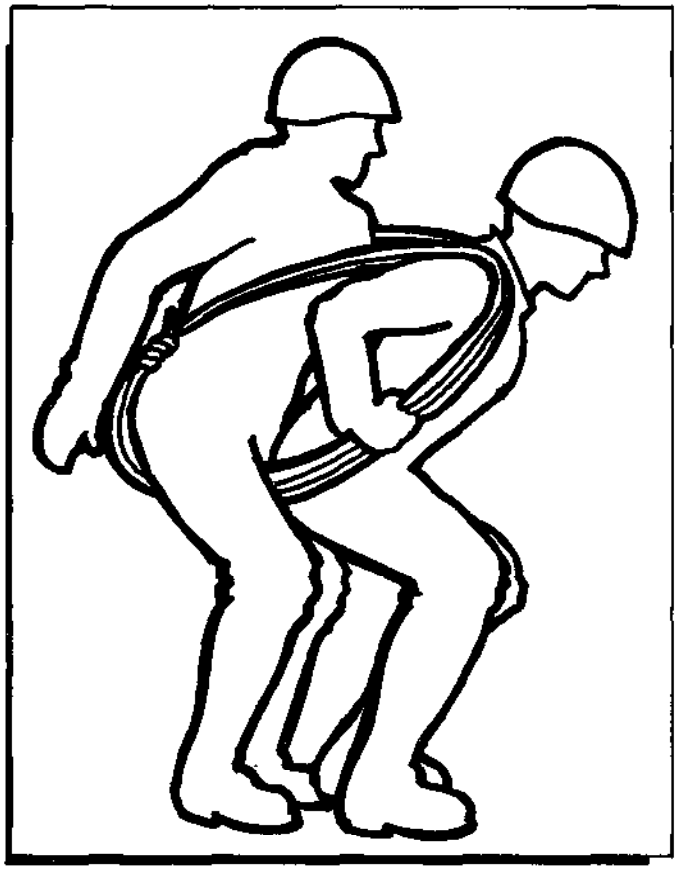

Figure 22-79. Rescuer Donning Rope and Patient.

(4) A sling is passed around the back of the patient, passing under the arms and over the shoulders of the rescuer. The sling is then wrapped around the coil as it passes over the front of the rescuer's shoulders (figure 22-80). The working ends of the sling should pass over the top of the loop formed in the wrap. Ensure the wrap is made low enough not to cross the rescuer's neck or interfere with breathing. The two ends are tied together in the center of the rescuer's chest using a square knot. The tails are taken down to the rescuer's seat harness, one tail is passed through each side, tied in a loop of the harness, and back up to the square knot. Here they are secured to the rope between the coil and square knot by a clove hitch knot on each tail (figure 22-81).

(5) To attach the party to the belay rope, the figure-eight knot on the end of the belay rope is attached to the carabiner on the crotch strap of the seat harness of the rescuer. The two tie end loops are secured with a carabiner sling; another sling is connected to the rescuer's chest harness and the belay rope with a prusik knot (figure 22-82). This sling is used to support the addition-

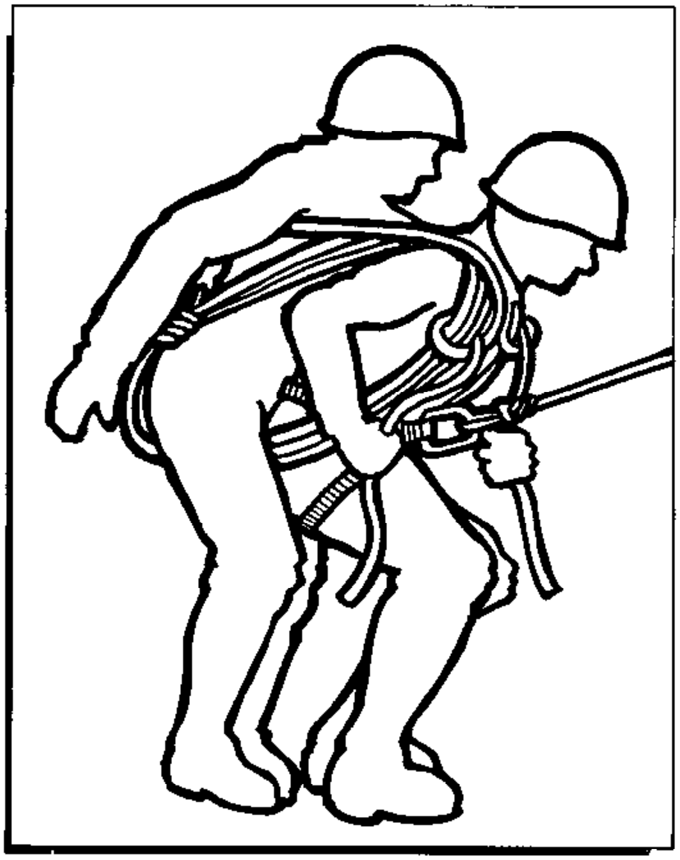

Figure 22-80. Securing the Patient for Buddy Evacuation.

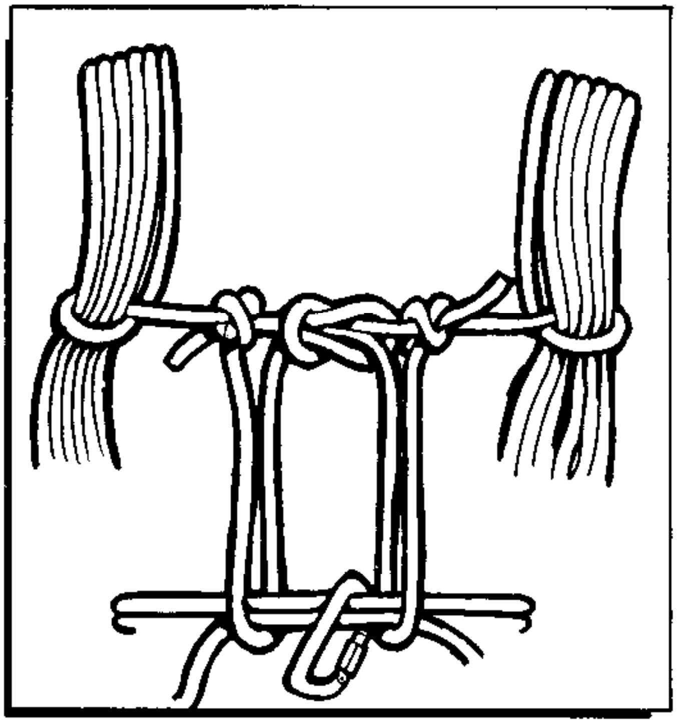

Figure 22-81. Tie-In Procedures.

al weight of the patient. This will enable the rescuer to remain perpendicular to the surface of the rock. Without this sling, the rescuer would fall over backward. The sling should be short enough so the rescuer can adjust the sling during the evacuation. The two people are then belayed down using a mechanical braking device (four-carabiner brake). (See figure 22-82.)

(6) The commands are:

(a) "ON BELAY?" - Given by rappeller.

(b) "BELAY ON" - Given by belayer.

(c) "BRAKE" - The belayer will stop the descent.

(d) "ROPE" - Belayer repeats "Rope." The belayer will feed the rope out in inches.

(e) "PATIENT SECURE" - Given by rappeller.

(f) "THANK YOU" - Given by belayer. The patient and the rescuer are in a safe location from rockfall and will not fall.

(g) "OFF BELAY" - Given by rappeller.

(h) "BELAY OFF" - Given by belayer. The rappeller is no longer secured by the belayer and the belayer can disconnect from the system.

i. Vertical Litter Evacuation. The vertical litter evacuation is used in areas where a horizontal descent is not possible. The patient is secured to the litter in the same manner as for a horizontal evacuation. Two ropes are used for the vertical evacuation. The belay rope to the litter is configured with a figure-eight knot and a locking steel carabiner. A double-tape sling is passed between the two steel frame bars forming the main body at the head of the litter (figure 22-83). The sling should pass outside of the four joining bars between the frame bars. Both ends of the sling are connected into the belay rope carabiner. Two carabiners are connected to the windows on one side of the litter. The second rope runs down the side of the litter through these carabiners to finish at the foot of the litter. A figure-eight knot is tied in the end of the rope. The seat harness of the barrelman attaches to this knot. A figure-eight knot is tied in the end of the sling and a carabiner clipped into the knot and connected to the foot of the litter (figure 22-84). A knot is added to the other end of the sling. A carabiner is added to the mariner's loop and connected to the barrelman's chest harness. As in the horizontal evacuation, the commands are given by the barrelman. The commands are the same as in the high angle horizontal litter evacuation (figure 22-85).

j. Horizontal High Angle Litter Evacuation. The horizontal litter evacuation is the preferred position for an injured patient. This position allows for easier medical treatment and is required for shock prevention. The danger of this position is the patient's exposure to rockfall. Because of its complexity, it is also the most

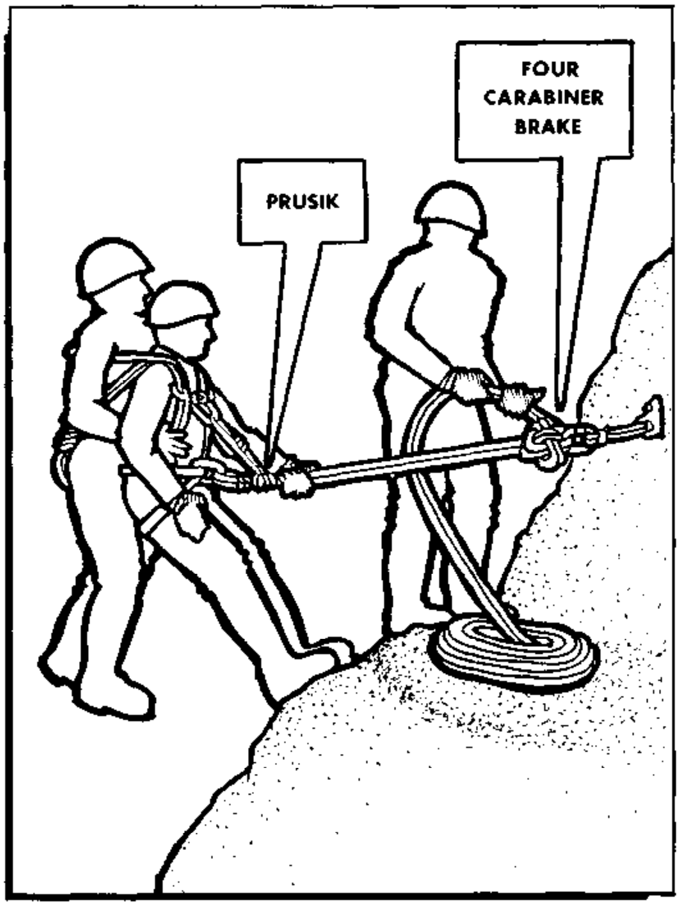

Figure 22-82. Hookup for Buddy Evacuation.

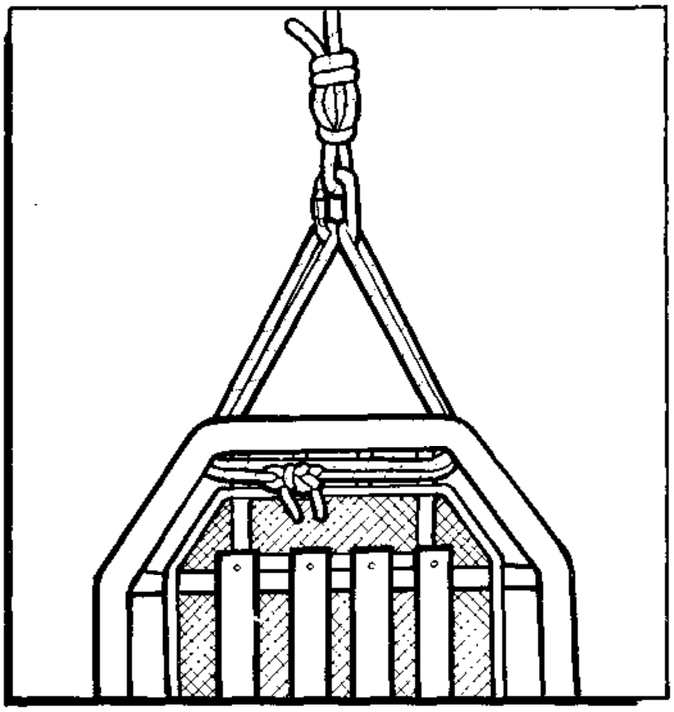

Figure 22-83. Attaching Rope to Stokes Litter.

potentially dangerous evacuation and should not be done unless no alternative exists.

(1) Team Composition. This method requires a minimum of three highly qualified team members. It also requires a great deal of equipment. It should not be attempted without the required equipment and qualified team members. Because the relative nature of the high-angle litter evacuation is potentially dangerous, the margin of safety must far exceed the stress on equipment. The safety factor is greatly increased by using equalizing anchors, two ropes for lowering a belayer, a rope handler, and a mechanical brake system. The high-angle evacuation on vertical walls is performed by only one litter bearer with the litter fully supported by the lowering ropes. However, on walls which are not vertical, it requires a great deal of strength to pick up the litter and walk it down. When an injured victim is on a steep wall, a second person will administer first aid and then assist in loading the victim into the litter. In fact, this "second person" may perform the technically most difficult part of the evacuation in loading the injured person. In some cases, a "third person" can help load a severely injured patient. The job of the barrelman is to get the litter to the victim, to see that the victim is properly loaded and secured, to give the victim a smooth evacuation by preventing the loaded litter from knocking against the rock, and to administer first aid to the victim if necessary. For first aid purposes, the litter should normally be horizontal. Proper operation of the brake controlling the litter is essential.

(2) Rigging the Litter. Two $^{7}/_{16}$-inch Mountain Lay ropes are used for lowering, each terminating in a figure-eight knot with a 4- to 5-foot tail, attached to a large locking steel carabiner. Each tail will terminate with a figure-eight knot and safety knot. One end will be attached to the seat harness of the barrelman and the other will be attached to the patient. These ropes should be the same length for ease in rope changes, and ideally the ropes will also be matched in elasticity and of similar wear. The spiders are the nylon slings, between 30 to 36 inches long, which attach the litter to the rope. These spiders are attached to the litter with large locking carabiners. These large carabiners are clipped over the outer rail of the Stokes litter and should have their gates facing in and locked (figure 22-86). If these carabiners are not available, the spiders can be attached to the major upright supports in the window of the litter with a double round turn and two or more half hitches. There are four spiders. One group of two spiders is attached to the windows at the sides of the litter at the head. The head group of two spiders is attached to the window at the sides of the foot of the litter. The spider groups should be connected with the key locking carabiner. This carabiner is attached to both of the ropes (figure 22-87). The

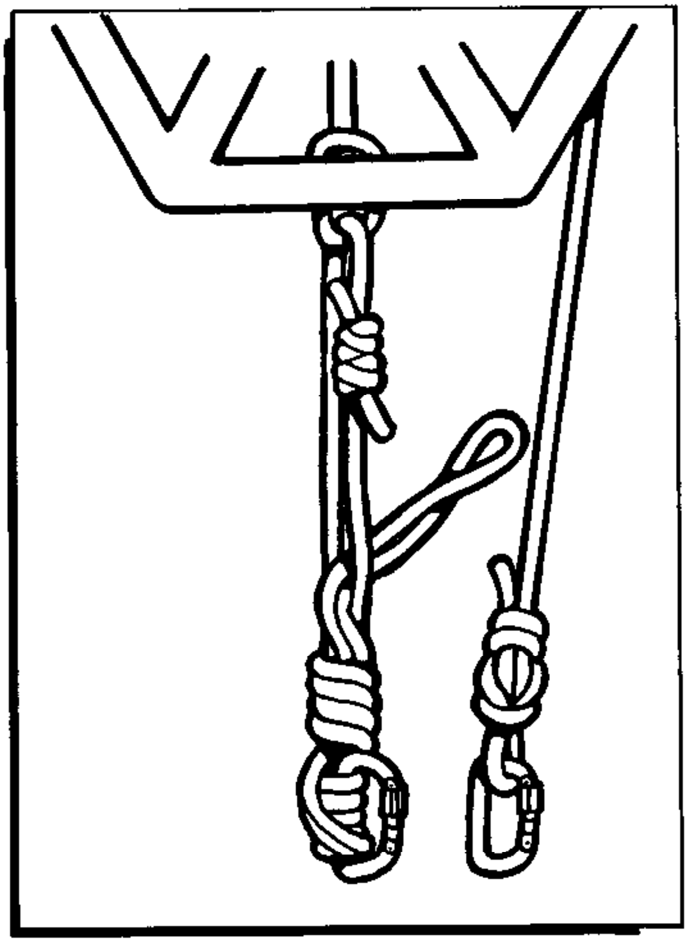

Figure 22-84. Tie-Ins for Foot of Litter.

spiders should be adjusted so the litter is horizontal or with the head only slightly elevated. The patient is tied into the litter in the same way as for the low-angle litter evacuation. However, the patient should have a chest or seat harness on. The figure-eight knot from one of the lowering ropes must be attached to the chest harness with a standard locking carabiner.

(3) The Barrelman. The barrelman is attached to the other lowering ropes by the figure-eight knot at the end of this rope. This rope is attached to the barrelman's seat harness with a standard locking carabiner. This is a safety line only. The majority of the barrelman's weight is supported with a 7 mm line which attaches the barrelman's seat harness with a standard locking carabiner to the key character. This line is secured with a prusick or Bachman's knot. The mariner's knot can be adjusted so the rescuer's feet are flat on the rock below the Stokes litter. The rescuer should be able to shield the patient with the upper body.

(4) Anchors. The anchor location should be directly above the victim if possible. For an evacuation on a large face when visibility is good, a spotter on the ground with a plumb bob can direct an accurate alignment of the anchor over the victim by communicating to the people on top. If the cliff is not vertical, the observer must observe the cliff "hang-on" and must be in the vertical plane containing the fall line through the victim. If the cliff is irregular, the plumb bob technique may be unreliable, but the observer may still be the best source of information in locating the anchor. If the anchor is placed in such a way that the rope will have to pass over a sharp corner of rock at the lip of the face,

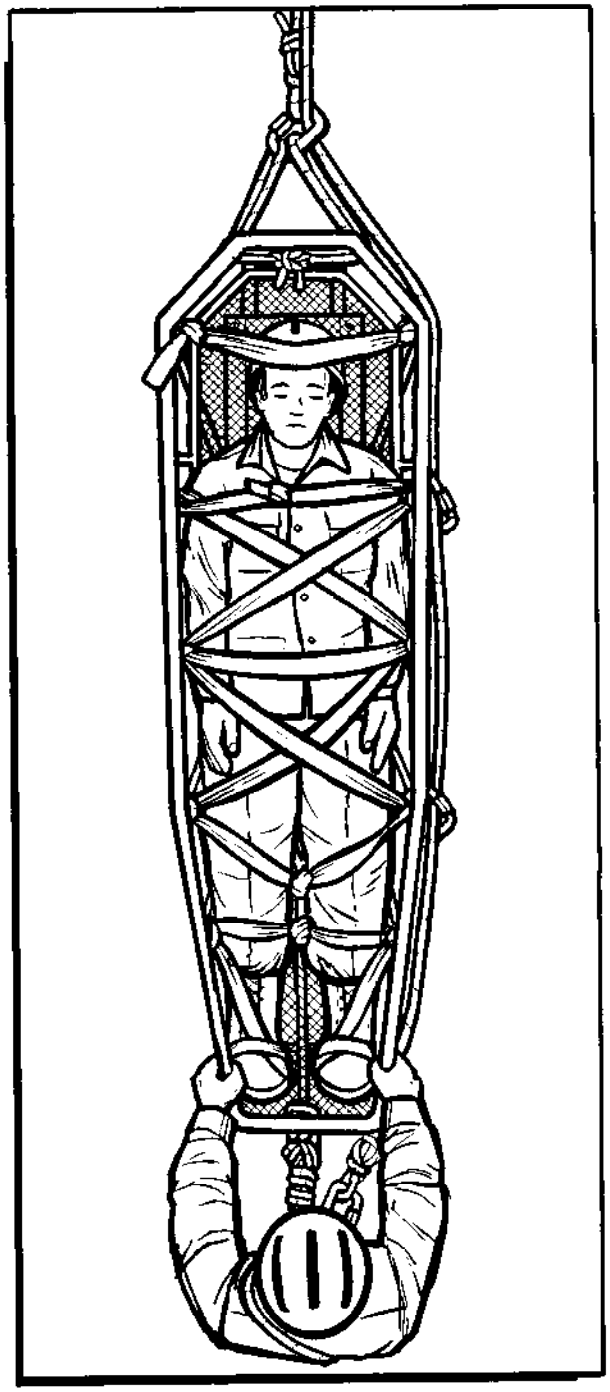

Figure 22-85. Horizontal Descent of Litter.

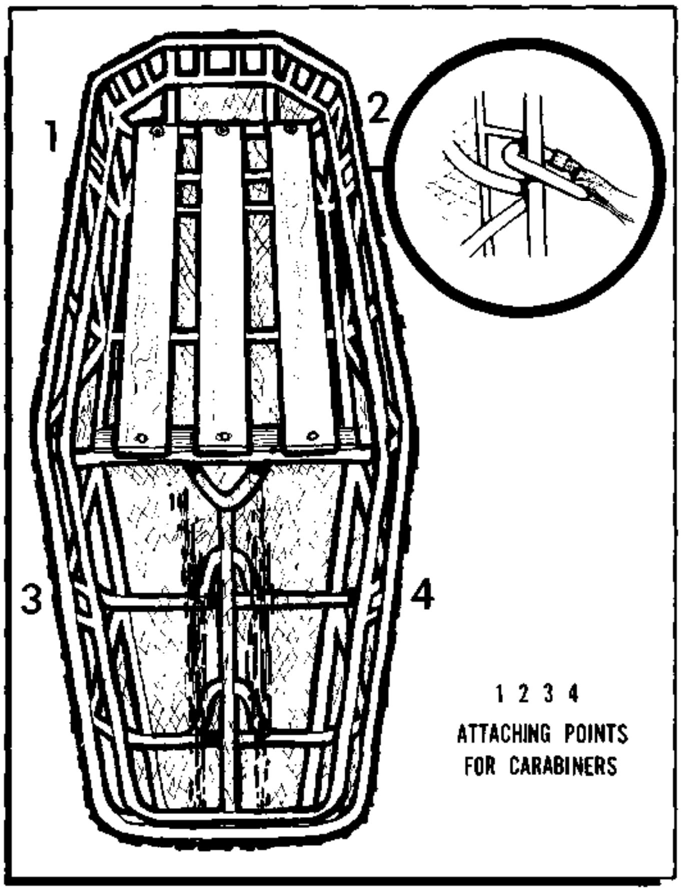

Figure 22-86. Attaching Points for Carabiners.

the corner should be "softened" by breaking the edge with a hammer and by securing padding. If possible, an equalizing anchor system (figure 22-88) should be used for anchoring the top brake systems. This is the safest anchor system to use because it will adjust for a shift in position and the shock load will be equally distributed if one of the individual anchors fail, the litter bearer falls, or the braking is of a fast descent.

(5) Braking System. The four-carabiner or figure-eight clog brake system can be used; however, the four-carabiner system is superior. This brake system will have a brake operator assigned. The brake operator obviously plays an important part in a successful evacuation, and should use gloves and avoid touching a hot braking device. To maintain full control while lowering, the brake operator should never let go of the ropes until tied off or "off belay." The operator should keep the descent smooth and steady, feeding the ropes equally into the brakes and never allowing a kink to form. A kink may jam in the brake and be difficult to release. Smoothness is easily produced if the loading by the litter is steady and has sufficient weight to require the brake operator to maintain a moderate grip on the rope and let the litter do the pulling. The litter should normally remain horizontal; thus the two lines lowering the ropes should be fed equally into the brake. The ropes may be grasped together by one hand and fed equally, even with unbalanced loads. It is often helpful to the brake operator to have a rope handler pull the rope which is not feeding rapidly enough through the brakes. This is particularly useful when the litter is lightly loaded.

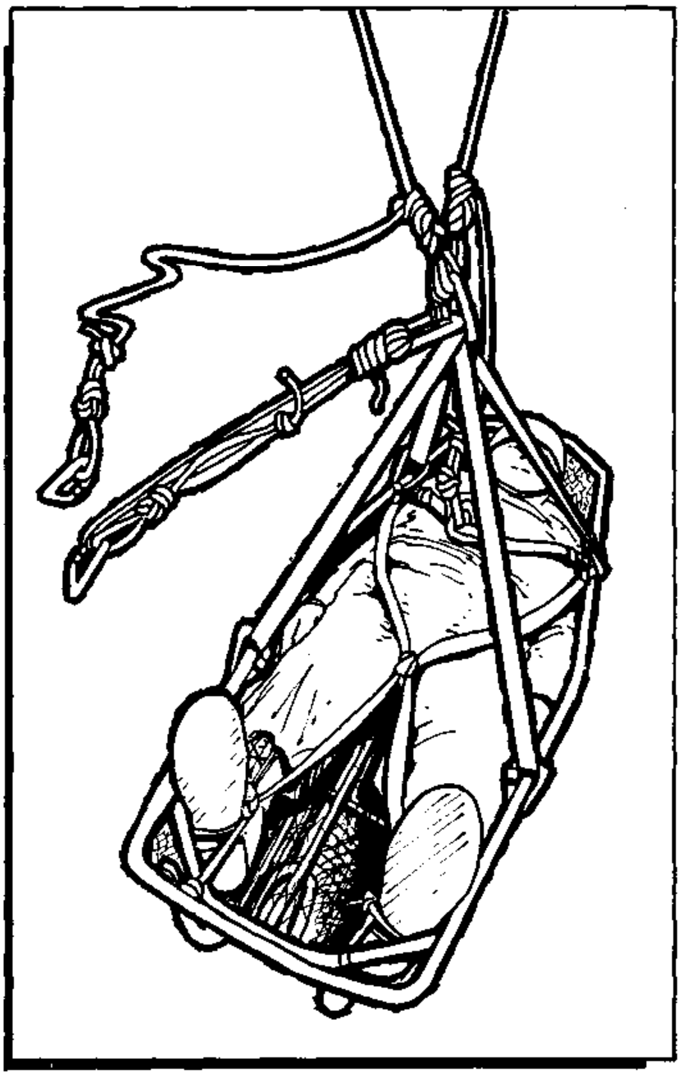

Figure 22-87. Rigging the Stokes Litter for Horizontal Evacuation.

(6) Rope Management. The proper management of the rope by the team at the top is essential for efficient evacuations. In fact, the overall job on top usually involves more thought and skill than is required by the litter bearer. The operations on top are best done by at least two people. The brake operator directly controls the speed of descent. A rope handler provides slack rope with no kinks to the operator and assists on the brake when required. The rope handler's tasks are crucial since the evacuation comes to a rapid halt if the ropes

become snarled. The ideal situation on top is to be able to lay the lowering ropes out in a straight line in their entire length. The rope handler will then have an easy time assisting the operator. If space is limited, the ropes can be stacked separately and neatly. There may be a tendency for the rope to kink at the brake, in which case the rope handler will stay busy twisting and spinning the slack rope.

(7) Handling the Litter. The barrelman will always wear a hardhat, pack, and gloves. The reason for this is that in case small rocks fall toward the helpless victim, the barrelman can lean over the victim providing a shield; and the packs and hardhats protect the barrelman. The victim will also wear a hardhat (when possible). A redundant barrelman tie-in is used for safety, and the locking of doubled carabiners should be used to prevent the barrelman tie-in from being accidentally released. Smooth handling of the litter by a skilled barrelman is the best way to ensure a safe ride down for the victim; banging the litter against the rock is very hard on the victim, and poor handling of the litter may result in injury to the barrelman. In starting the evacuation, the barrelman may have to carry the litter a short distance at the top of the face taking care that ropes are positioned over a smooth rounded edge of the cliff and the spider carabiners have gates facing in. Once their weight is on the lowering ropes, the barrelman's main job is to hold the litter away from the rock. The feet are used against the rock as they are in rappelling to maintain footing. The barrelman's position on a nearly vertical wall is to be hanging from the tie-ins which may be a comfortable seat sling. The barrelman (figure 22-88) is normally on the outside of the litter (away from the rock), gripping the outer or an underneath rail to hold the litter still farther off the rock to clear projections, etc., or to level the litter if loading is uneven. Occasionally, the rail at the cliff face is grasped, but care must be taken to avoid getting crushed hands.

(a) The descent is easy on flat, smooth, vertical walls. On a wall with obstructions, footwork techniques are the same as in rappelling. The legs should be perpendicular to the wall and spread comfortably apart. The Prusick or Bachman's knot should be adjusted so the barrelman moves the litter to the edge of a vertical face and may require readjusting so the space between the litter and the barrelman can be maintained. The barrelman will be bent at a 90-degree angle at the waist which will enable the barrelman to bend over to protect the patient or administer first aid.

(b) The litter must be secured to an anchor to keep it from sliding and falling. The lowering ropes should be removed from the spiders and the knots untied to allow the ropes to be hauled back to the summit without a knot jamming. The spiders themselves should not be dragged up the cliff because of the danger of jamming. Normally, the ropes will be pulled up and carried down rather than dropped because of the possibility of snagging or damage. After the evacuation is complete, the litter should be moved to an area sheltered from falling rock.

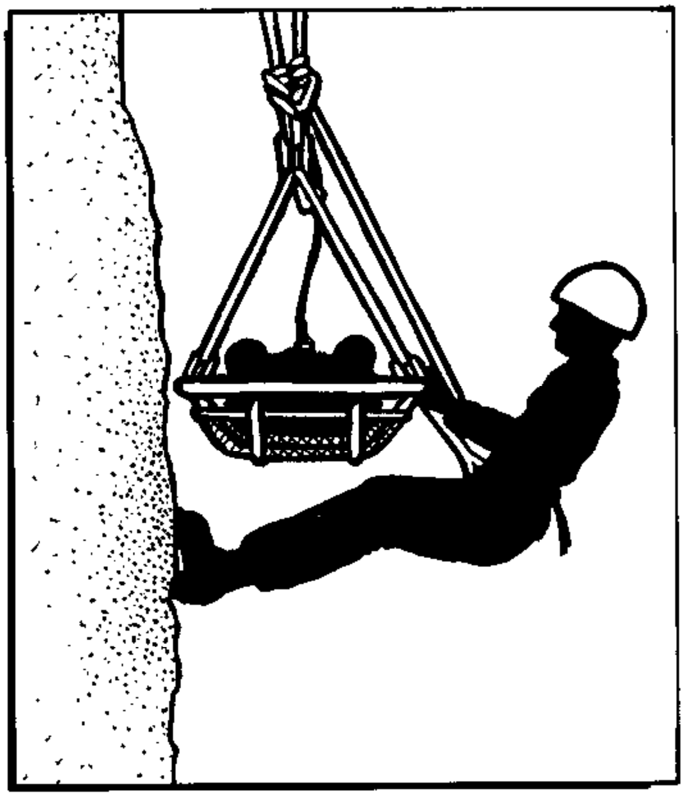

Figure 22-88. Body Position for Rescuer.

(8) Communication. The general goal of the brake operator is to lower the litter smoothy and safely as directed by the barrelman. The commands to be used are as follows:

(a) "ON BELAY" - Has the usual meaning and is used before the evacuation begins.

(b) "ROPE" - Means for the brake operator to feed the lowering ropes equally through the brake at a moderate speed.

(c) "SLOW ROPE" - Means for the brake operator to feed the lowering ropes equally through the brakes at a slow speed.

(d) "BRAKE" - Means for the brake operator to brake both lowering ropes simultaneously.

(e) "SECURE" - Means to maintain the brake until the litter is anchored or in a position where it will not fall.

(f) "UP ROPE" - Means the litter is detached and all knots and slings have been removed. The rope can be pulled up without becoming caught on rock outcroppings or in cracks.

k. Horizontal Hauling Lines and Tyrolean Traverse. A suspension system used to transport a load off the ground along a taut (static) line may be useful when crossing streams, gorges, or other difficult terrain. These systems can be used for team movement or for transporting victims. The two-rope bridge is time consuming to set up, but in spite of the complexity the techniques can be useful in rescue. However, time should not be wasted waiting for the system to be set up. Sometimes the rigging may be done while the victim is being transported to the crossing site. Since system stresses are often great, the margin of safety is smaller than usual or desirable in rescue operations, and the rigging should be supervised by someone experienced in the techniques.

(1) Anchors. A major requirement for any suspension system is the availability of secure anchors at each end. Anchors have to be as strong for this use as in any mountain rescue. To withstand the high-line stress, the anchors may have to be a combination of several anchor points rigged to be self-equalizing. For reasonable safety, each anchor system should be able to hold at least 10 times the load to be transported.

(a) The anchors must be suitably placed in addition to being strong. To prevent the load from "bottoming out" or to minimize line tension by increasing sag, the anchors must be placed in a high location. If the rope contacts any sharp rock, the rock edges must be padded with suitable material such as leather gloves, pack, tree branches, etc., since a taut rope can be easily cut. The anchors should be 6 to 10 feet back from the edge of the stream or cliff to provide for loading and unloading. The rope should be 3 feet or more off the ground at these positions, and if good anchors do not allow this, A-frames may be improvised from sturdy tree trunks.

(b) If possible, two ropes or a double rope should be used to increase the safety margin. However, one rope will work although the strength of the system is decreased. It is very important to understand the stresses involved with this system. For example, if the rope span is 40 feet and the sag in the middle of the rope is 1 foot with a 200-pound load, the rope tension is approximately 2,500 pounds. However, if the sag is 4 feet in a 40-foot span, then the tension is 650 pounds.

(2) Establishing the Two-Rope Bridge. Connect the center of the rope through the anchor. Do not tie, but simply pass through a carabiner. If the span requires two ropes to be tied together, a double figure eight should be used with single fisherman safeties. A carabiner must be placed in the double figure eight; if not, the double figure eight will be difficult to untie. One team member will cross the obstacle carrying the two ends of the rope plus the hauling rope. Once across, an anchor is established and one end of the rope connected to it (figure 22-89). Remove as much slack as possible so the knot will be taut against a 4-inch pulley or carabiner on the opposite side when tied off to the anchor system. The other end of the rope is then put through a 3:1 mechanical advantage system.

(a) To use the mechanical advantage, a safety prusik is attached to the rope. This safety prusik must be attached to a stout anchor and be as long as practical

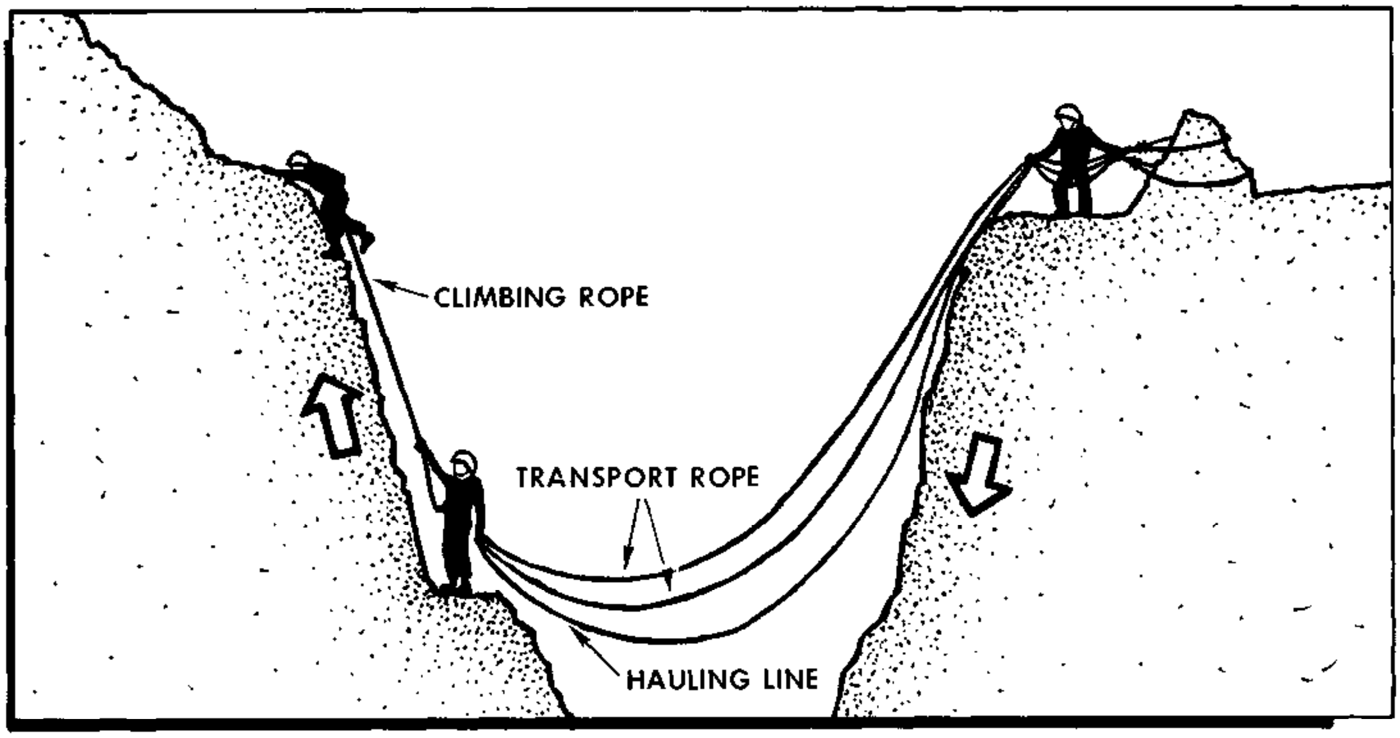

Figure 22-89. Horizontal Hand Evacuation System.

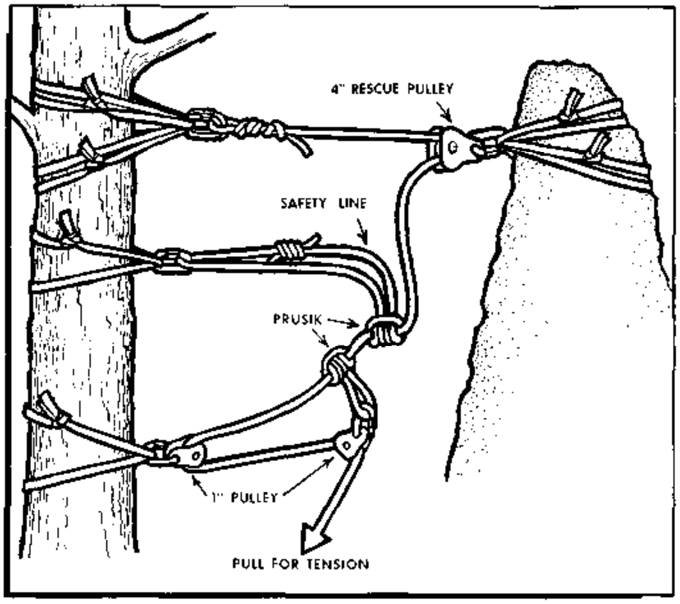

Figure 22-90. Rigging Hand Evacuation System.

to hold the tension on the rope and to slide forward as the slack is pulled out of the rope. The rope is then passed through a carabiner which is attached to an anchor. A smaller prusik sling is tied on the rope behind the first prusik. A carabiner is attached to the prusik sling. The rope is then passed through this carabiner. The rope should resemble a "Z" when the running end is pulled toward the anchor. This is called a "three to one mechanical advantage" because there are three moving lines. The force of pull on the running end of the rope will be increased (ideally) three times. This does not account for the friction of the rope passing over the carabiners or pullies.

(b) Using the mechanical advantage, the following steps should be used to tighten the rope. Pull on the hauling line or the running end of the rope until the snap links come together. Have another person keep this safety prusik slide out as far forward as possible. Then let the safety prusik take up the load, while slowly letting slack out of the running end of the rope. Then slide the second prusik out as far as possible, and repeat the entire maneuver. When the rope is tight enough, tie the rope off (figure 22-90).

(3) Crossing the Two-Rope Bridge. Half the team will now cross the rope by means of the Tyrolean traverse. This is a means of crossing a suspended rope while being secured to the rope. A pulley or carabiner is attached to the rope and clipped into the seat harness. The person then hangs under the rope and pulls hand over hand to the other side. The load is placed on the rope by pulleys or carabiners. The end of the second rope is tied to the load and used to pull the load across the rope to the other side (figure 22-91). Once the patient is secured in the litter, the slings to hold the litter horizontally can be attached. There are six slings in all, one set of three at the foot of the litter. In each of the

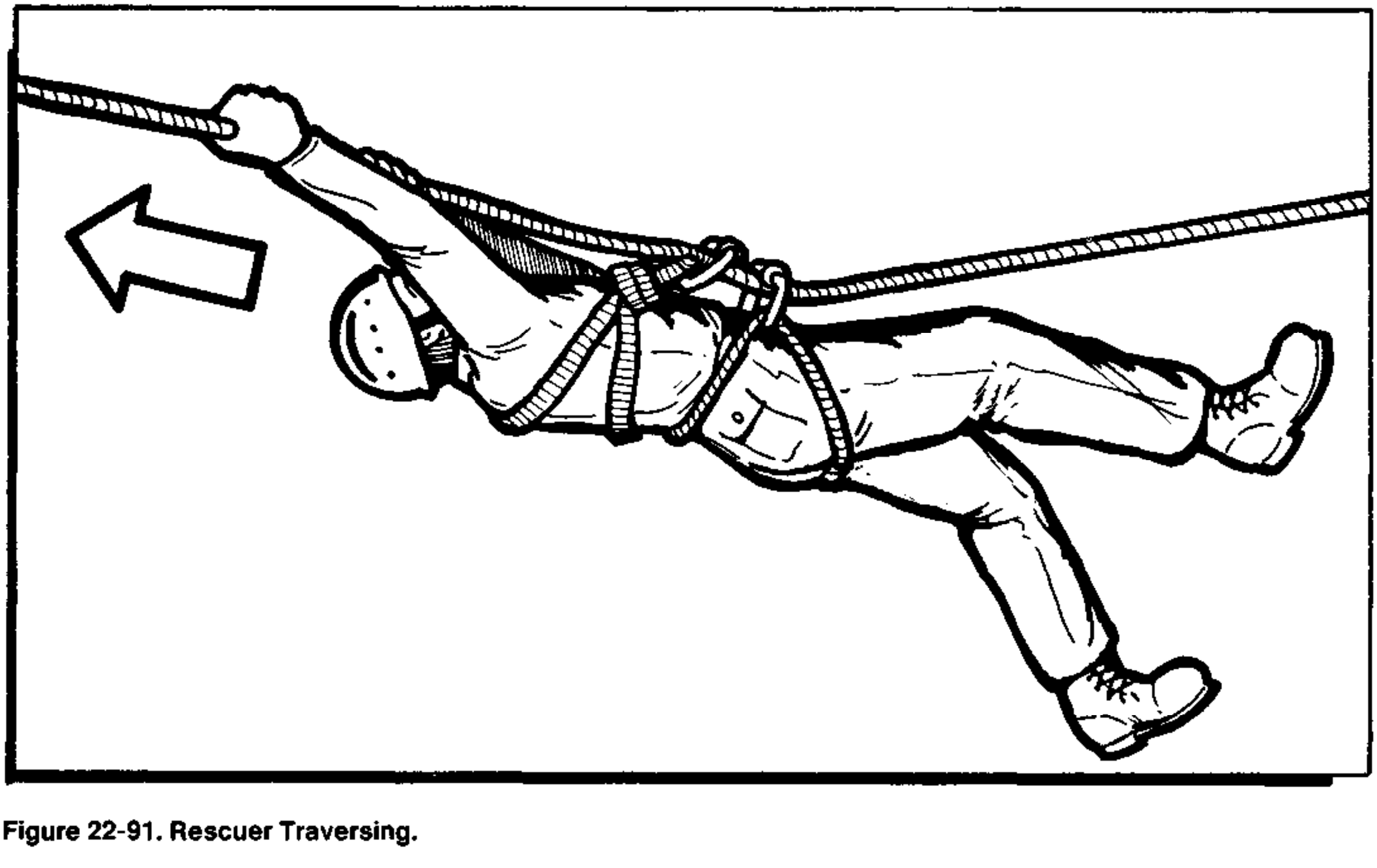

Figure 22-91. Rescuer Traversing.

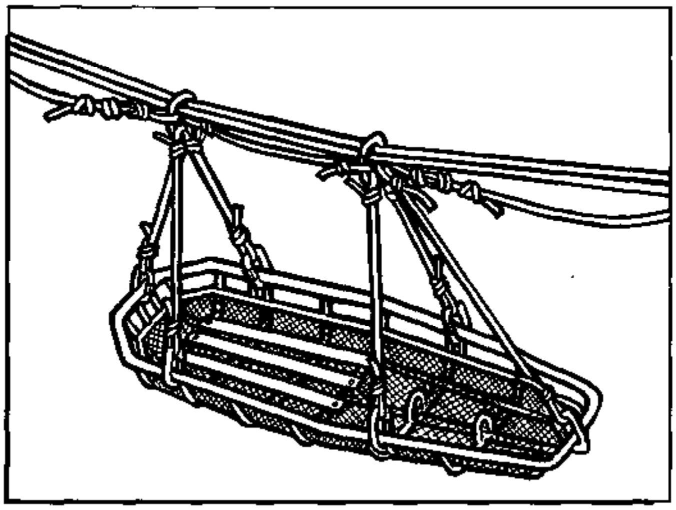

Figure 22-92. Rigging the Horizontal Hauling System.

sets, one is attached to the end of the litter and the other two to the sides of the litter. The slings are about 20 inches long and attached to the inner rail of the Stokes litter with a round turn and two half hitches (figure 22-92). The remaining team then crosses the rope. Each set of slings is attached to a steel "D" locking carabiner. A short length of rope will join the two carabiners for equal pull during the traverse. An in-haul line should be attached to the litter to control the descent or to haul across the open expanse (figure 22-93). To retrieve the rope, the anchors are removed and the rope simply pulled across the obstacle. If an artificial anchor was used for the midrope anchor, it is not retrievable (figure 22-94).

(4) Crevasse or Ravine Recoveries. The ability of a team to recover a victim below them, as in a crevasse, requires the use of a mechanical advantage. A mechanical advantage is the arrangement of ropes and use of pulleys to arrange a system which will allow a force directed on the system to produce a lifting power in multiples of the force. The primary use of a mechanical advantage is to perform vertical lifts as in crevasse or cliff rescue and to apply tension for tightening the system used in construction of hauling lines (figure 22-95). The mechanical advantage can be in different ratios depending on the number of pulleys and changes of direction. Pulleys should be used when a rope passes through a carabiner. If pulleys are not available, a carabiner can be used; however, a larger amount of friction will be present causing a loss of some of the advantage in the system. Rope placement is important in the construction of a mechanical advantage. The ropes should run side by side with no twist around each other. Two basic types are illustrated. The Z-pulley method uses a portion of the load-bearing rope to make the mechanical advantage and, therefore, is a one-rope system (figure 22-96). The second type uses a second rope to establish the mechanical advantage on the load-bearing rope (figure 22-97).

(5) Z-Pulley System. This system of mechanical advantage is used to lift a fallen climber or stranded individual out of a crevasse, up a cliff face, etc. If the fallen climber is attached to a rope, this rope may be used for the system. If the climber is not attached, it will be necessary for a member of the rescue team to rappel or be belayed down to the climber. The rope should be connected to a seat harness and a chest harness. An improvised harness may be required (figure 22-97). Pull all rope slack up from the crevasse. Near the edge, anchor a prusik sling and attach it to the rope. This keeps the rope from slipping back down over the edge.

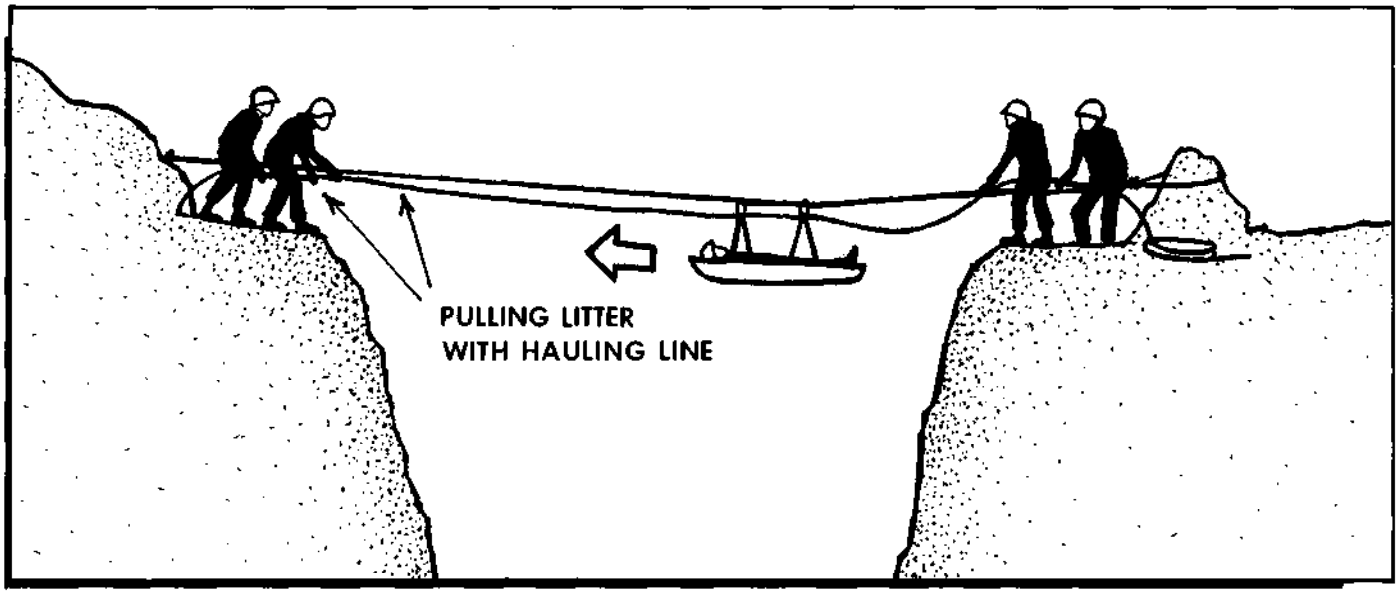

Figure 22-93. Pulling the Litter.

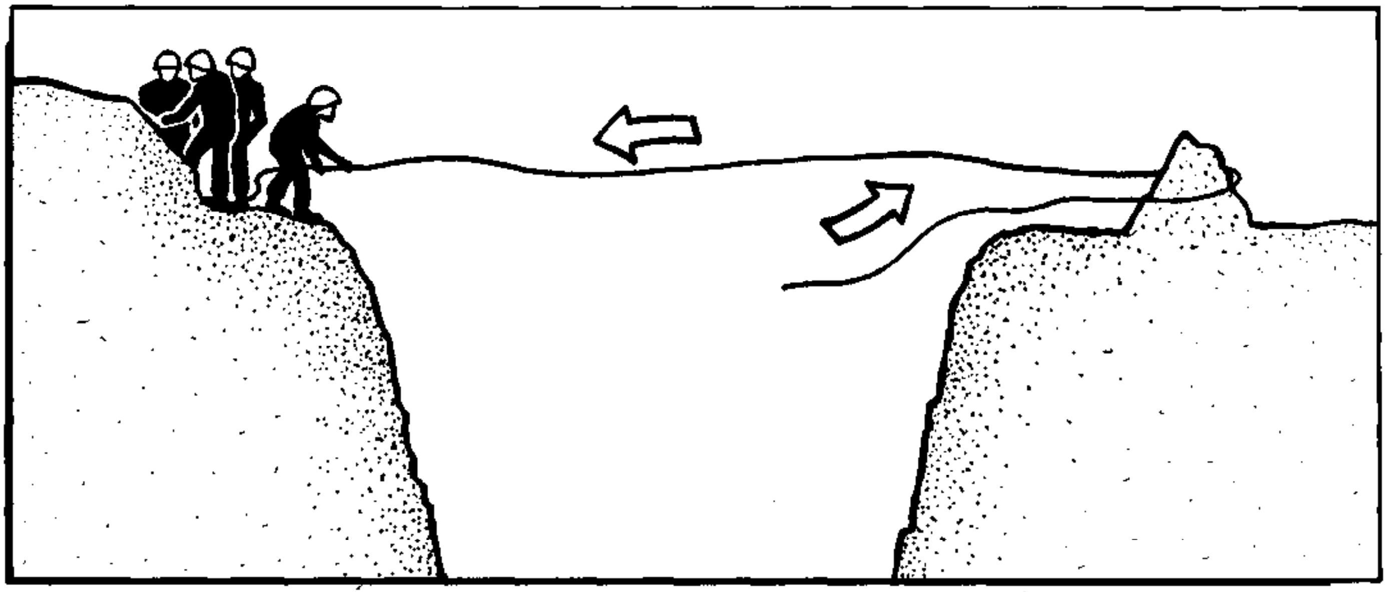

Figure 22-94. Retrieving the Rope.

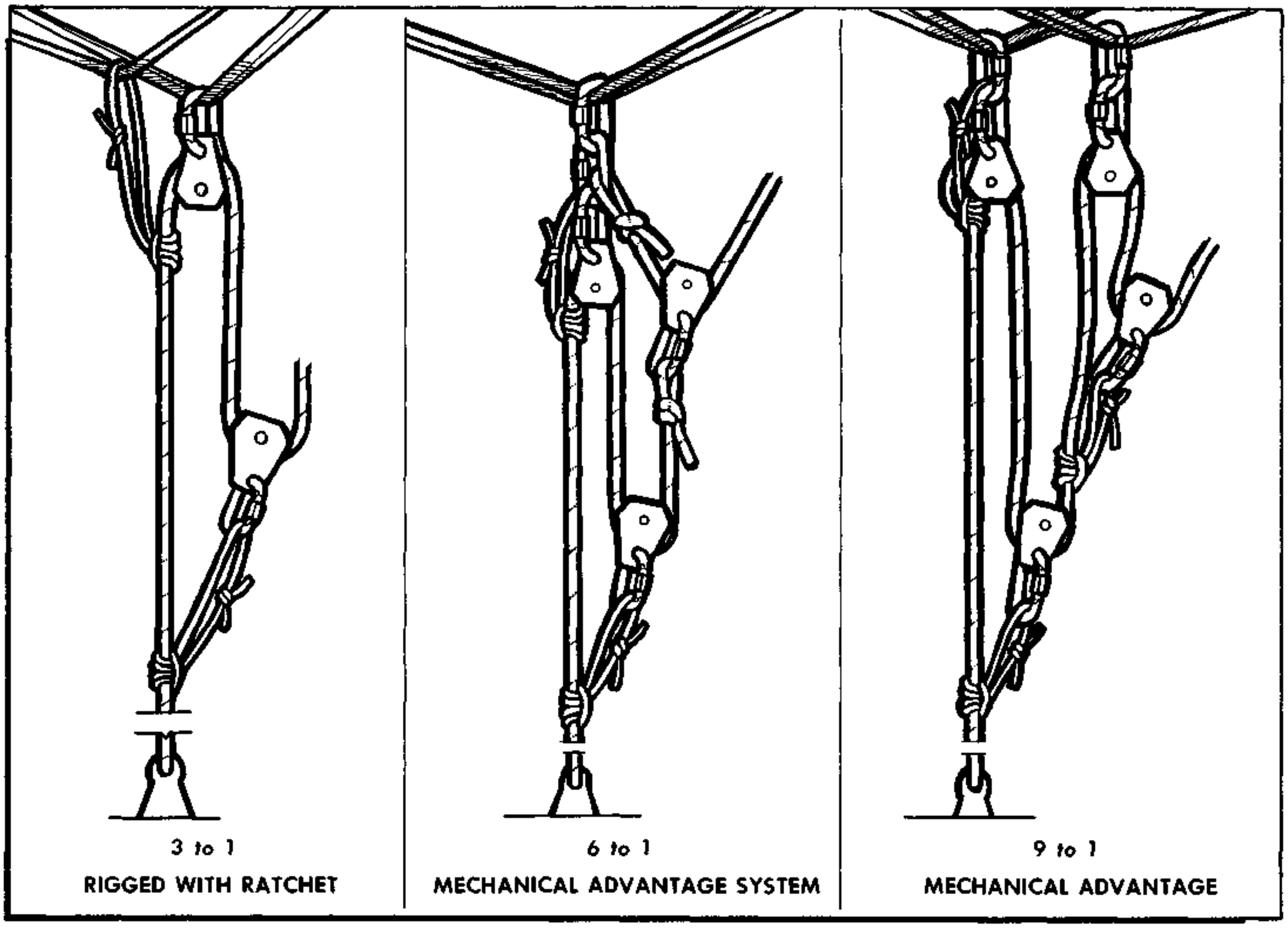
3 to 1
RIGGED WITH RATCHET
6 to 1
MECHANICAL ADVANTAGE SYSTEM
9 to 1
MECHANICAL ADVANTAGE

Figure 22-95. Mechanical Advantage Systems.

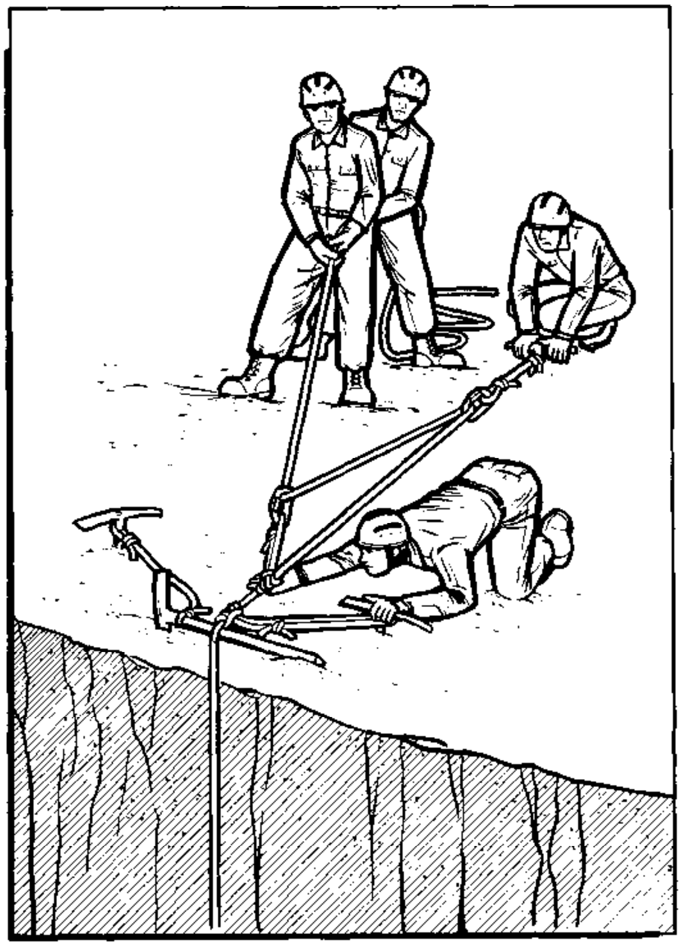

Figure 22-96. Two-Pulley System.

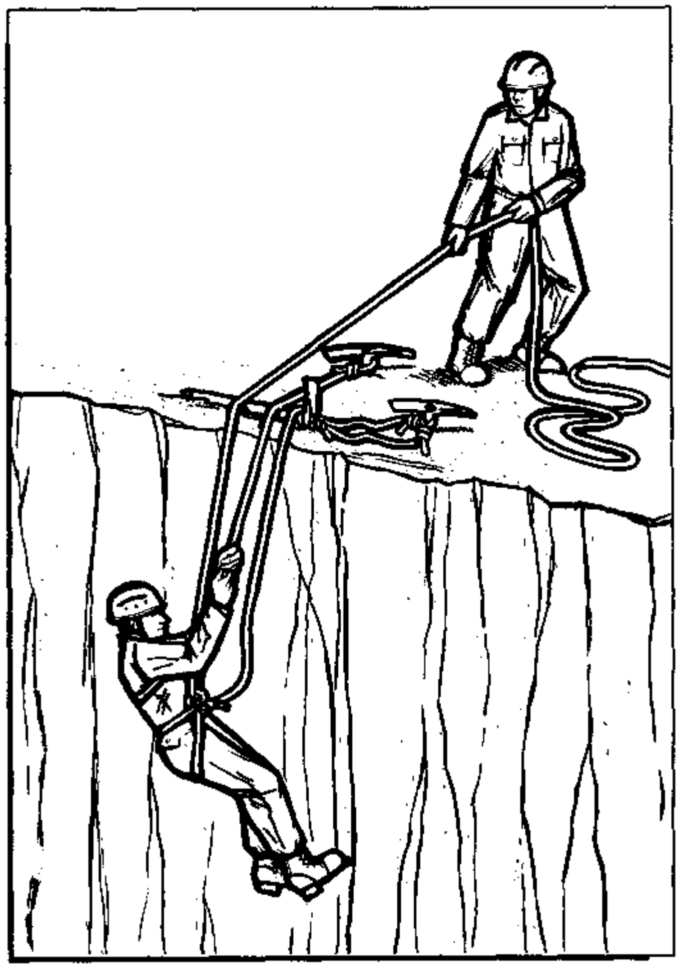

Figure 22-97. Second Rope Added.

An anchor system should be established back from the crevasse and a pulley attached to the anchor. The rope is routed from the crevasse through this pulley. The rope comes out of the pulley and runs back to the crevasse edge. Here another prusik sling is added to the rope as it comes out of the crevasse and a pulley attached to it. The rope from the anchor pulley is passed through this pulley. The system is now complete. The operation is conducted by pulling on the standing end of the rope until the second prusik sling is pulled to the anchor pulley. The prusik at the edge of the crevasse takes the load while the other prusik is reset in the original position. Continue with the lift until the climber is free of the crevasse.

(6) Added Rope Mechanical Advantage. This system requires an additional rope to form the system. As in the Z-pulley system, a rope from the fallen climber is required. Pull all rope slack up from the crevasse. Establish an anchor and connect a prusik sling to the anchor. Secure the belay rope with this prusik sling. Establish two separate anchors and anchor one end of a long sling to the anchor nearest the belay rope. Attach another prusik sling to the belay rope between the anchor prusik and crevasse edge. On the standing end of this sling, attach a carabiner using a figure-eight knot. Attach another carabiner through the body of the figure-eight knot. Place a pulley on each of these carabiners and on the remaining anchor point. Run the long sling through the pulley at the end of the sling, back through the pulley on the second anchor, and through the remaining pulley in the body of the figure-eight knot. The system is ready for operation (figure 22-96). Pull on the long sling end until the prusik with the two carabiners is pulled into the anchor system for the long sling. Allow the anchor prusik to hold the weight and reset the long sling in the initial way. Continue with this procedure until the climber is raised from the crevasse.

(7) Gorge Lift Rescue System. One rescuer rappels to the climber while the other goes around the crevasse or gorge to the other side. An anchor is established on both sides. The end of one rope is thrown across to the far side where it is attached to the anchor. The center of the rope is allowed to drop in the crevasse where the climber is attached to it by a pulley. The end of a second

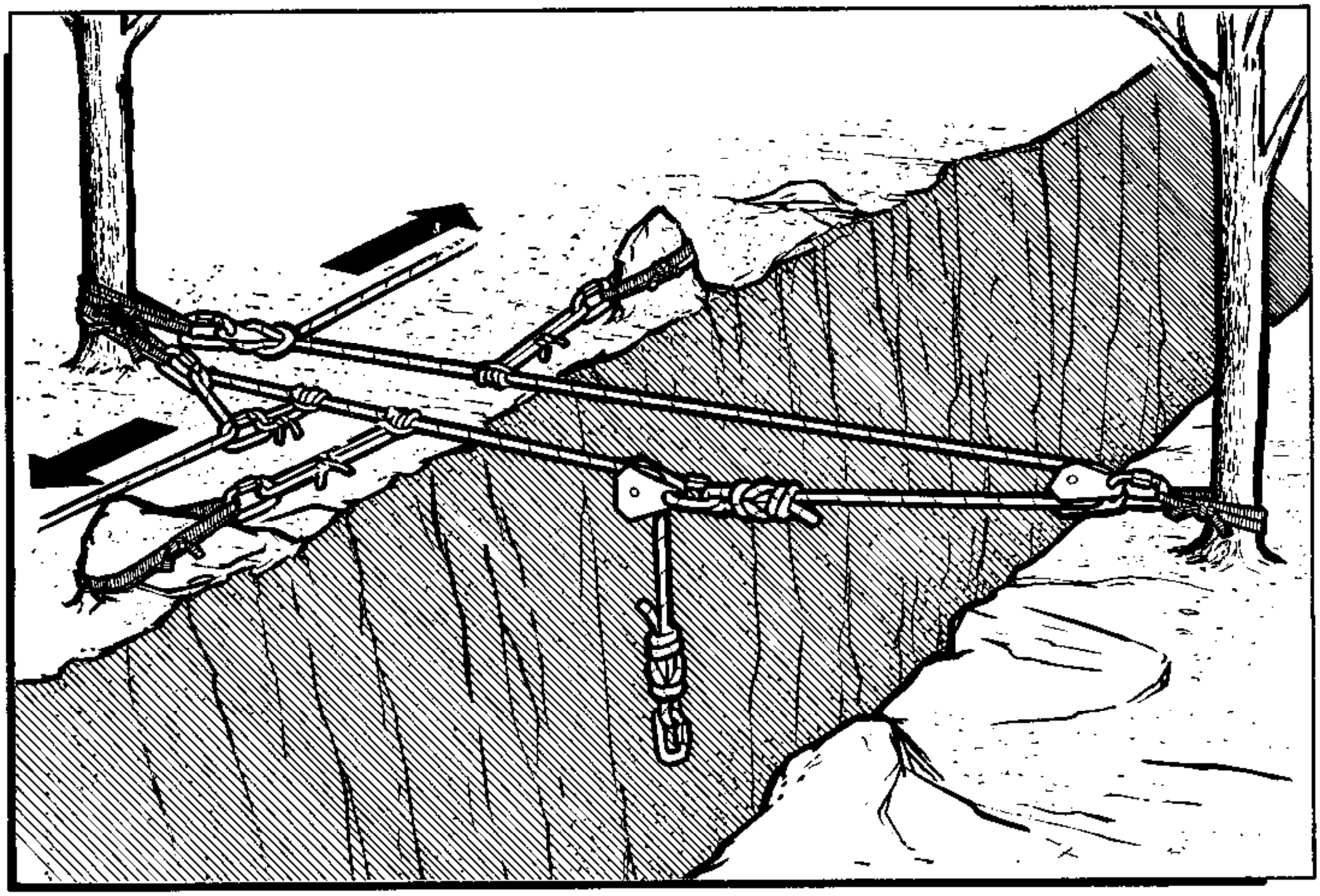

Figure 22-98. George Lift.

rope is lowered to the climber and anchored to the person. The free end of the first rope is put into the mechanical advantage system. As tension is applied to the rope, the climber is lifted. When the rope is at an approximate 40-degree angle, the climber is pulled toward the edge by the ropes. As the climber nears the edge, the weight of the climber will sharpen the angle of the rope. It will be necessary to apply more tension to the rope. Continue until the climber is rescued (figure 22-98).

(8) Rope/Knot Bypass While on Belay. When using multiple ropes or adding extra ropes to a long descent, it is necessary to pass the joining knots through the belay system. The knots will not pass through the system so a method is used which will secure the rope while the belay device is removed and the knot passed. As the knot approaches the belay system, apply a brake and secure the system. Place a prusik sling on the rope below the belay device. Run this sling to the anchor and tie a mariner's knot (figure 22-99). Take the brake off the belay and allow enough rope through the belay until the

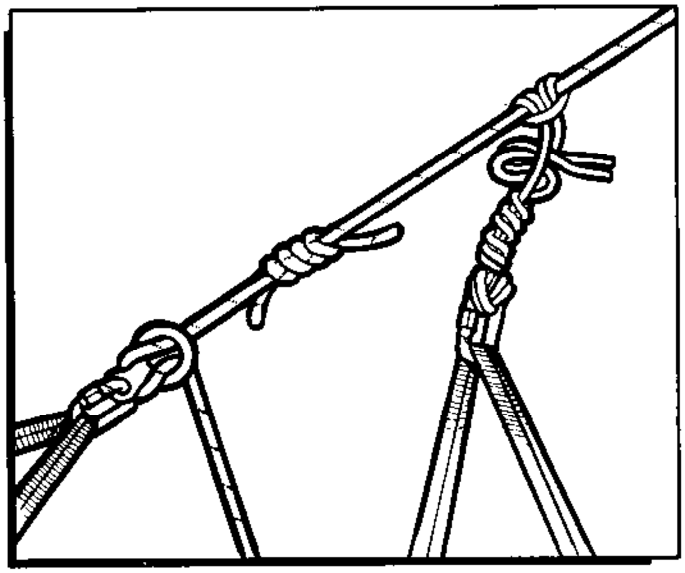

Figure 22-99. Multiple Ropes.

tension is taken by the mariner's knot. Replace the rope through the belay immediately behind the knot and secure the brake. Slowly remove the mariner's knot until the tension is again on the belay system (figure 22-100). Continue with the descent and repeat procedures as necessary.

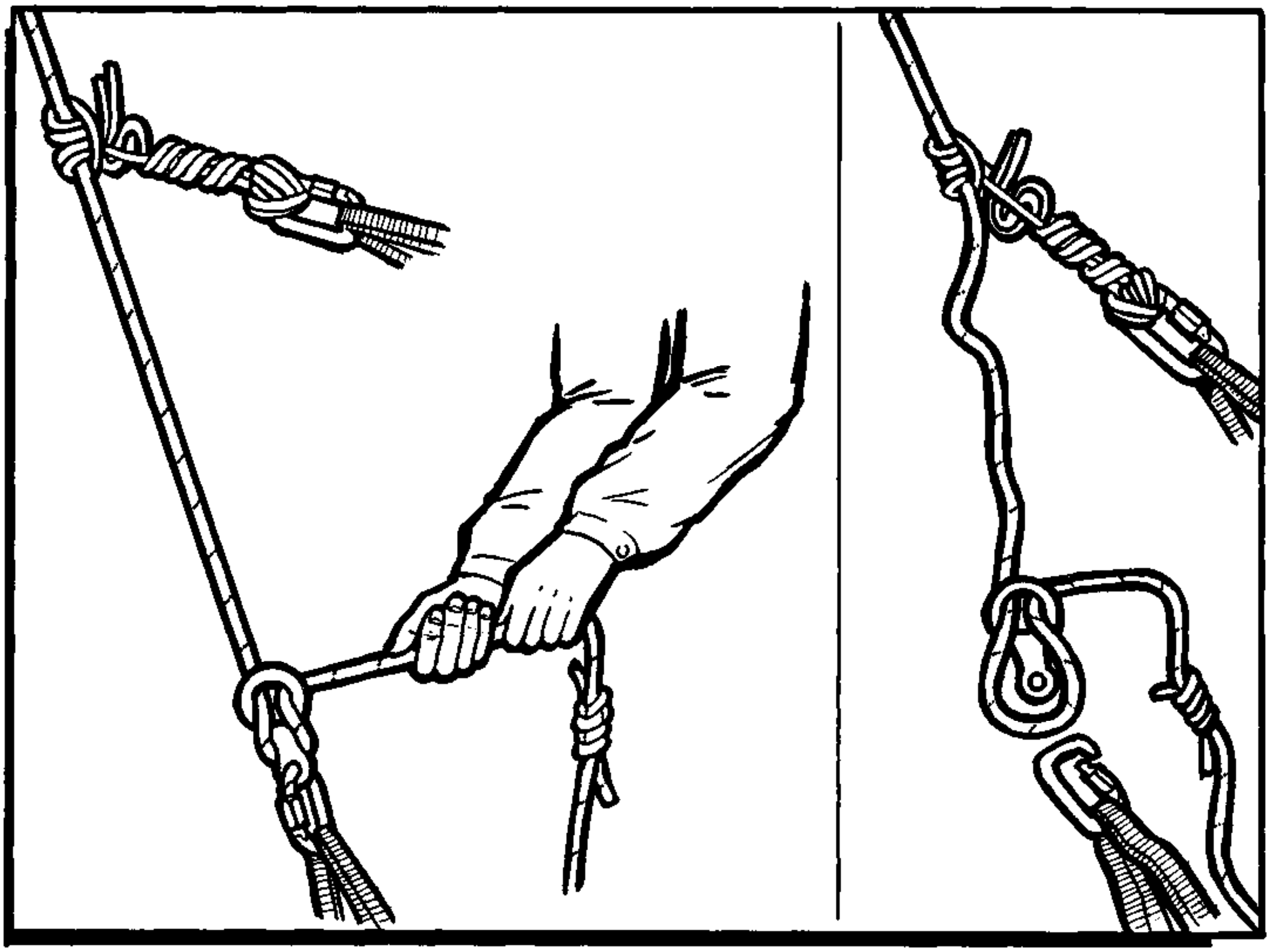

Figure 22-100. Rope Bypass.

Chapter 23

WATER TRAVEL

23-1. Introduction. In this chapter, the travel techniques which can be used on 71 percent of the Earth's surface, the water environment, will be addressed. The techniques of river and open sea travel may be adapted to other water features such as swamps and lakes. In the paragraphs concerning open sea travel, the environmental factors of oceans of the world will be considered as they relate to travel. The techniques for individual water rescue and swimming are important for recovering injured survivors and equipment. Additionally, the problem of submersion must be considered as well as how the antiexposure suit, life preservers, and liferafts can extend a survivor's life expectancy. This is why a survivor must be familiar with the individual rafts and raft procedures. The ultimate goal of a survivor on the open seas is to be rescued; however, a second goal, if rescue is not made, would be to make it to land. A thorough knowledge of water travel techniques will greatly increase the survivor's chances of reaching landfall. If done correctly, landfall can be reached with minimum loss of equipment or injury.

23-2. River Travel:

a. Rivers have been used for centuries in all environments as a safe means of travel and is the reason why most of the cities of the world are located on rivers. It is not uncommon for a river to flow at 4 or 5 knots per hour. A survivor could travel 20 to 25 miles in 5 hours of travel. This may contrast greatly with the rate of travel on land. The amount of energy required to carry survivors' equipment and other supplies and to travel 20 to 25 miles on land is much greater.

b. Each major continent has thousands of miles of navigable rivers. Some rivers such as the Nile, Amazon, Mississippi, Lena, and Mackenzie have hundreds of miles of navigable water with seldom a ripple. These navigable sections are generally found flowing through the flatlands, plains, tundras, and basins of the world. In these areas, only the temperatures of the water and the plant and animal life may present hazards. In contrast, the headwaters of rivers, like the Mackenzie, Yangtze, and Ganges, are so rough that they would best be categorized as a threat to life. This would also be true of the Snake, Salmon, and Rogue Rivers of the Northwestern United States. These rivers, although traveled by white water rafters, pose an unreasonable hazard to survivors. Survivors must take into account individual or group skills, injuries, type and severity of rapids, the temperature of the water, and direction the river flows in making the decision to travel. Even if a portage of several miles is required, the energy saved by floating on a river might warrant river travel. However, once the energy expended for portage exceeds the energy conserved by floating a section of a river, the river as a mode of travel should be abandoned.

c. In a nontactical situation, rivers will most likely carry survivors to indigenous people who could aid the survivor in meeting the basic needs for sustaining life and effecting rescue. Even if some form of civilization is not encountered, the survivors would most likely reach a lakeshore or even the seacoast. These environments, particularly the seacoast, provide transition zones from land to water which are rich in food and other survival resources. In these areas, the resources could improve the chances of survival. It is much easier to spot signs of survivors along a shore versus the interior of a landmass.

23-3. Using Safe Judgment and Rules for River Travel:

a. There are certain safety rules and guidelines that must be followed to reduce the dangers associated with river travel. Respect for these rules and guidelines is necessary to reduce the potential dangers.

b. The most important safety rule is personal preparation. Preparation should begin by thoroughly scouting the river. The conditions of the river will determine the intermittent stops. High ridges along river edges provide needed visibility to plan each leg of travel. If there are numerous bends and poor lookout points to view the river, stops are frequent. Sound judgment must be used when planning routes. Patience in planning each leg of travel helps prevent disaster. All survivors must prevents know the plans and be able to handle the route safely, considering their skills and strength. Survivors should be aware of and avoid river hazards and have alternate routes and communication signals in case flow conditions suddenly change, making the run more difficult. All rapids which cannot be seen clearly from the river should be scouted. The route should be discussed with the crew. The skills, knowledge, and abilities of the aircrew members must be considered, including swimming abilities and physical condition. Areas of high risk should not be attempted. Before reaching an area of suspected great difficulty, rafts should be beached and carried to the next point of travel; this is called portaging.

c. Before entering the raft, survivors should don life preservers and suitable clothing for adequate protection. The equipment should be tested to ensure it is serviceable. Bulkiness is not advisable due to the possibility of the raft capsizing and being weighted down with water. The antiexposure suit (if available) should be worn. Items that might absorb water should be packed in a waterproof container. The survivor should ensure that:

(1) All medical supplies, repair kit, and survival kit are in the raft.

(2) Survival kit is checked and inventoried.

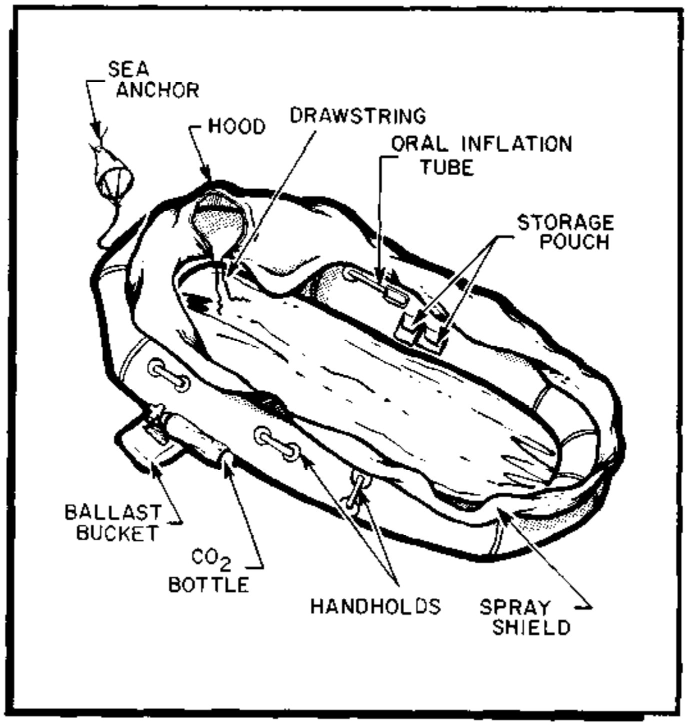

Figure 23-1. One-Man Raft.

(3) Extra efforts should be made to keep supplies and equipment in good condition.

(4) All items are secured to the raft to prevent loss and (or) injuries.

(5) Before use, the raft is checked for leaks and necessary repairs are made.

d. When using a one-man raft for river travel (figure 23-1), it may be advisable to tie or cut off the ballast bucket, fasten the spray shield in the opened position, and remove the sea anchor to prevent problems with swamping or entanglement with subsurface obstacles. Without the ballast bucket, the raft can be easily maneuvered by paddling with the backstroke or for slight adjustment, with a front stroke. When using either the backstroke or the front stroke, the survivor will find it easier if the two underarm cells of the life preserver underarms (LPU) are disconnected in front and the cells placed behind the back (figure 23-2A). This gives the survivor a full range of motion. When rough water is encountered, the survivors should fasten the LPU and face downstream.

e. One of the primary methods of avoiding hazards on the river is to slow the speed of the raft and move across the river, avoiding a collision with the obstacle. This ferry position should be initiated early to avoid large rocks and reversals in the river. If the collision obstacles are to be avoided, point the bow of the raft toward them and backstroke against the current to slow the speed of the raft's downstream progress and move it across the river. Usually, the best angle is about 45 degrees to the current. The greater the angle, the quicker the movement across the river, but this also increases the downstream speed of the raft. Decreasing the angle will slow downstream speed, but movement across the width of the river will also be decreased (figure 23-2B). The raft will be more maneuverable if it is well inflated. If the raft should pass over a rock, arch the back up to prevent injury to the buttocks or back.

f. When using multiplace rafts, the boarding ladder and sea anchor should be removed to prevent entanglement. If available, about 50 feet of line should be tied on the bow and stern of the raft to be used for tie-offs. An additional 200 feet of line should be coiled and tucked away for emergency and rescue work; one end is secured to the raft while the other end has a fixed loop. An improvised suspension line rope may be used for this; for example, three-strand braid can support about 1,000 to 1,500 pounds of pressure, or a two-strand twist strengthens the line to support about 700 to 900 pounds of pressure (figure 23-3).

g. Proper placement of equipment and personnel should equalize weight distribution to ensure stable control; overloading should be avoided. Assign personnel crew positions and responsibilities in the raft; captain (person in charge), stern paddler (maneuvers raft), and side paddlers. Twilight and night rafting should be avoided (nontactical) as poor visibility increases danger.

h. Two ways to steer a multiplace raft are:

(1) To steer a raft by using sweeps (long oar) and poles. A pole is more efficient in fairly shallow water, but a sweep is preferable in deep water. Poles and sweeps from both ends of the raft are used. The person in the bow (front) can see any obstructions ahead, and the one in the stern (rear) can follow directions for steering. Poles are also useful for pushing a raft in quiet water.

(2) Paddle techniques are used to maneuver the raft. When paddling, there are three possible body positions on a raft. The best way is to sit on the upper buoyancy tube with both legs angled to the inside of the raft. The body should be perpendicular to the sides of the raft, enabling the rafter to paddle. Another way is to sit cowboy style, straddling the upper bouyancy tube of the raft with one leg on either side, and folding at the knee with each leg back. However, the outside knee may collide with obstacles and cause injury. The third way is normally used in calmer waters because it consists of partially straddling the upper tube, with legs comfortably extended. In a smaller raft, the survivor may be able to sit down inside the raft and reach over the buoyancy tube. The following strokes can be done from these positions. Knowing the parts of paddles will help in explaining the different paddling strokes (figure 23-4).

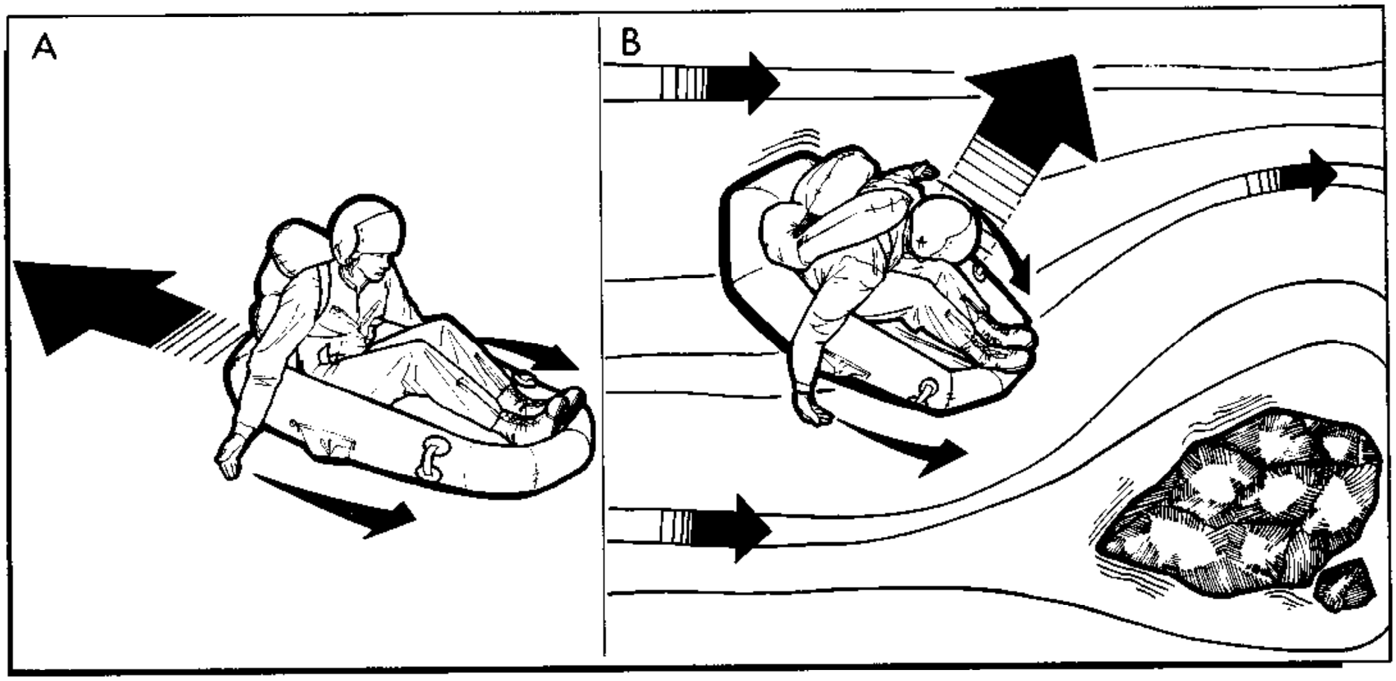

Figure 23-2. One-Man Raft (Paddling).

(a) One of the easiest is the forward stroke which is done in smooth continuous movements using these techniques:

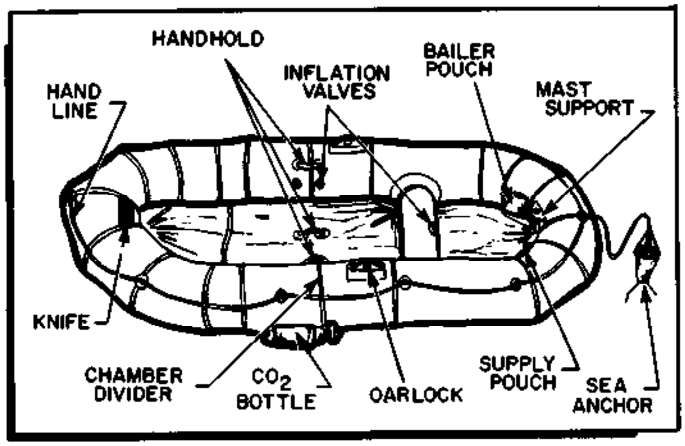

Figure 23-3. Seven-Man Raft.

-1. Thrust the blade of the paddle forward using the outboard arm, then momentarily keeping the outboard arm stiff and away from the raft, push the grip. The inboard hand is then moved forward to cut the blade deeply into the water. Continue the stroke by pushing on the grip and pulling on the shaft keeping the blade at a 90-degree angle to the raft. Stop the motion as the blade comes slightly past the hip, because a full followthrough provides little forward power and wastes valuable energy. Slide the blade out of the water by pushing down on the grip and swinging it toward the inboard hip and turning the blade at a parallel angle to the water once it has cleared the water. By paralleling the blade, it cuts wind and wave resistance and saves time and energy. This cycle is repeated until the strokes are changed.

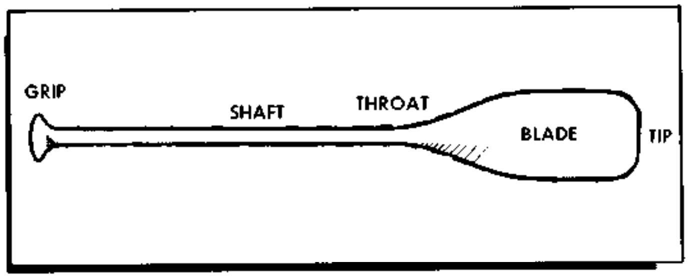

Figure 23-4. Paddle.

-2. In mild water, there is no need to over reach or excessively twist the upper trunk of the body. When extra speed is needed, lean deeply into the strokes which brings the entire body into play. Position the inboard hand across the tip of the paddle grip and the outboard hand halfway to three-fourths of the way down the shaft.

(b) The opposite of the forward stroke is the backward stroke. The blade is thrust into the water just behind the hip, and pressure applied by simultaneously pushing forward on the shaft and pulling back on the grip. End the stroke where the forward stroke would

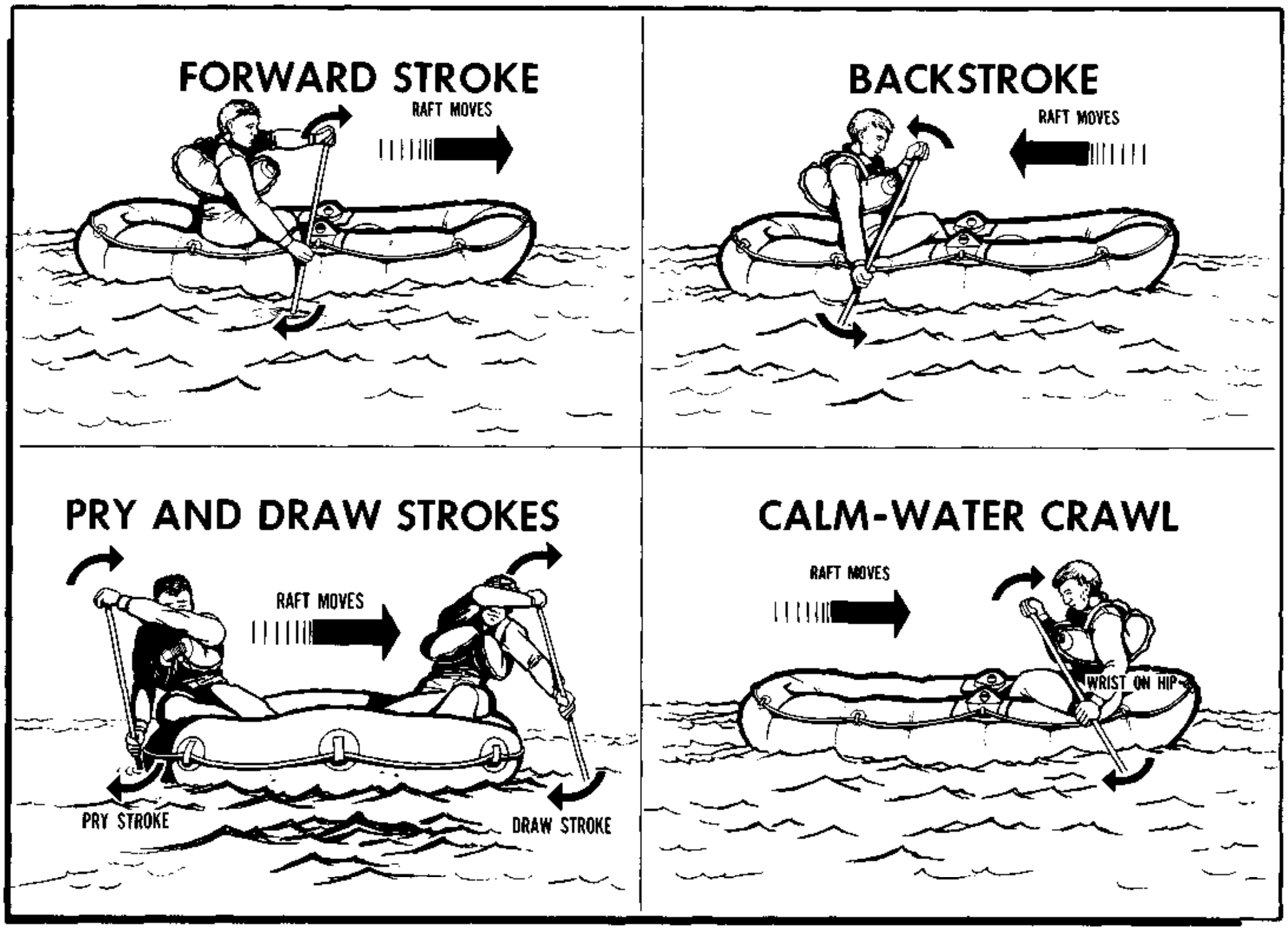

Figure 23-5. The Paddle Strokes.

begin, and again angle the blade out of the water back to the beginning of the backward stroke.

(c) The draw and pry strokes are opposite sideways strokes. These strokes are good for small sideways maneuvers and for turning the raft when used from the front or rear of the raft.

-1. Drawstroke. Reach out from the raft, dip the blade in parallel to the raft, and pull on the shaft while pushing on the grip. Pull the blade flat to the side of the raft. Pull the paddle out and repeat.

-2. Pry stroke. Dip the blade in close to the raft, and push out on the shaft while pulling in on the grip.

(d) The fifth stroke, called the calm water crawl, is used alternately with the forward stroke when paddling through long calms. Sit cowboy fashion while facing the stern and hold the paddle diagonally in front with the shaft which is held by the outboard hand against the outboard hip and the grip held by the inboard hand in front of the inboard shoulder. Extend the inboard arm to swing the blade behind, dip the blade in the water, and pull back on the grip, prying it forcefully, using the hip as a fulcrum (the point of support on which a lever works). Using the shoulder, hip, and hand as assisters, the crawl is easy yet powerful (figure 23-5).

(e) The ferry is a basic paddle maneuvering technique used to navigate bends and to sidestep obstacles in swift currents. The ferry is essentially paddling upstream at an angle to move the raft sideways in the current. Paddle rafts can ferry either with the bow (front) angled upstream or downstream. The bow-upstream ferry is stronger because it uses the more powerful and easier forward stroke. It is carried out by placing the raft at a 45-degree angle to the current with the bow angled upstream and the side toward the desired direction. The bow-downstream ferry is weaker because it uses the less powerful backstroke, but it does offer certain advantages. It enables paddlers to look ahead without straining their necks, and makes it easy to put the bow into waves (figures 23-6 and 23-7). It is carried out by backstroking with the stern (back) angled upstream at a 45-degree angle and the side facing the desired direction.

(f) There may be times when the only way for a heavy raft to enter a small or violent eddy is with a reverse ferry (figure 23-8). The following steps may be used for an oar or paddle raft, except the paddle raft approaches the eddy bow first and finishes in a bow-upstream position:

-1. Raft approaches sideways.

-2. Raft turns around to angle its bow downstream.

-3. With careful timing, the captain should have the crew begin to pull powerfully on the paddle. The angle of the raft to the current can be close to 90 degrees, but is best at about 45 degrees.

-4. While aiming for the eddy, the crew should continue with the front stroke and gain momentum.

-5. With the crew still using the front stroke, the raft breaks through the eddy fence.

-6. With the bow in the upstream eddy current and the stern still in the downstream current, the raft is spun into a normal ferry angle. The crew continues with the front stroke while making the necessary turn to bring the boat entirely into the eddy.

-7. The raft rides easily in the eddy. (NOTE: The reverse ferry and eddy turns are not only used to enter eddies, but can also be used to dodge through tight places. The reverse ferry (or sometimes an extreme ferry) scoots the raft sideways, the eddy turn snaps the bow into a bow-downstream position, and the raft, rather than entering the eddy, rides the eddy fence past a major obstruction or hole.)

(g) The straight forward paddle is used in calm and moderate waters where there is ample maneuvering time. Simply point the bow in the desired direction and follow the forward stroke method of paddling. The back paddle is performed the exact opposite of the forward paddle. Point the stern (back) in the desired direction and follow the backstroke method of paddling (figure 23-5).

(h) To make a left turn, the left side of the raft will backpaddle, while the right side paddles forward. It's just the opposite to make the raft turn right. The right side on the raft backpaddles while the left side paddles forward—both performing the paddling maneuvers at the same time (figure 23-9).

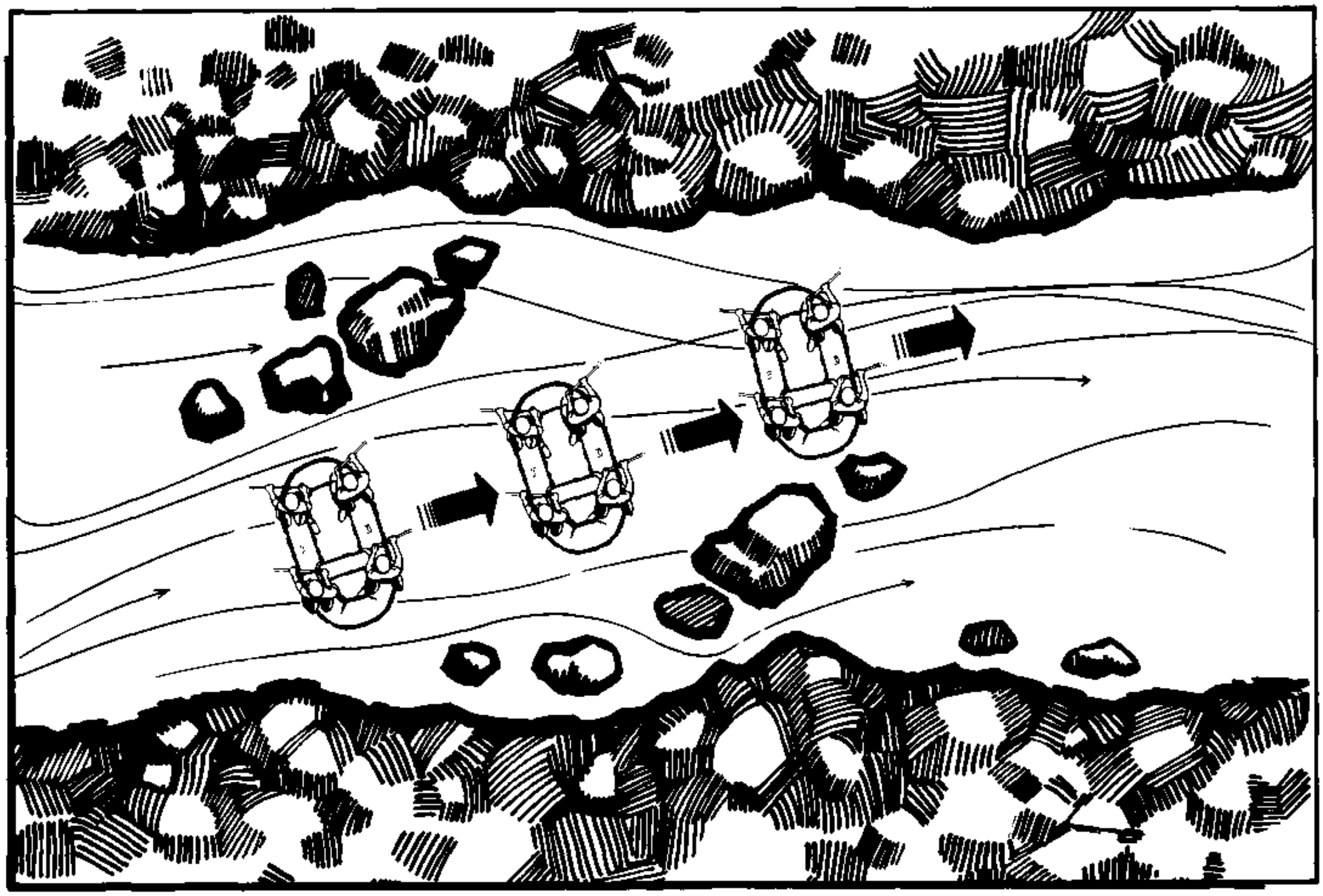

Figure 23-6. The Paddle Ferry.

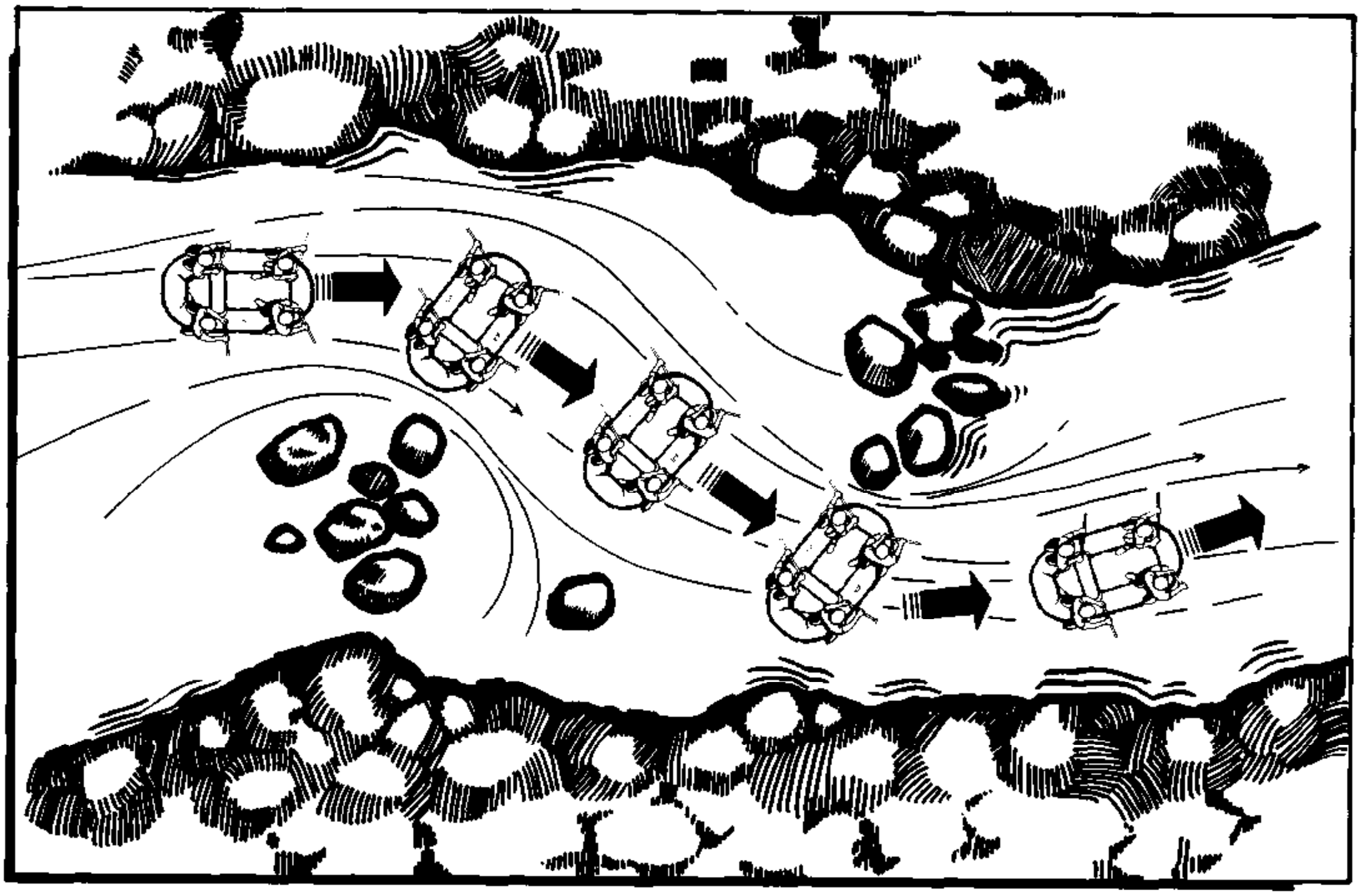

Figure 23-7. Bow Downstream Ferry.

(i) The pry and draw strokes are used to move the raft sideways (figure 23-10).

(j) Stern maneuvers are used to increase the maneuverability of the raft. The paddle used at the stern of the boat is basically a rudder which controls direction. To turn right, the paddle blade is held to the right, square against the direction of the current. To turn left, the paddle is held to the left.

(k) Other strokes, such as a forward stroke or draw stroke, used at the stern of the raft, will cause it to turn or move faster. If stroking is done slightly to the side (either right or left) of the raft, it will help move the raft in the opposite direction (figure 23-11).

(l) River travel requires fast, decisive action. Therefore, a paddle raft needs a captain to coordinate the crew's actions by the use of commands or signals. Communications between captain and crew are crucial; all members must agree on a set of short, clear commands. The following are suggested commands:

CAPTAIN'S COMMANDS:	CREW RESPONSE:
Forward	Crew paddles forward.
Backpaddle	Crew does backstroke.
Turn right	Left side paddles forward, right side does the backstroke.
Turn left	Right side paddles forward, left side does the backstroke.
Draw right	Right side uses draw stroke, left side uses pry stroke.
Draw left	Left side uses draw stroke, right side uses pry stroke.
Stop	Paddlers relax.

-1. Commands must be carried out immediately, so the crews should practice until they can snap through all the commands without hesitation. The captain controls both the direction and speed of the raft with a specific tone of voice and commands. This control, and the captain's ability to anticipate how the water ahead will affect the raft, will help avoid undercompensation and overcompensation of maneuvers through obstacles. Good captains think well ahead and

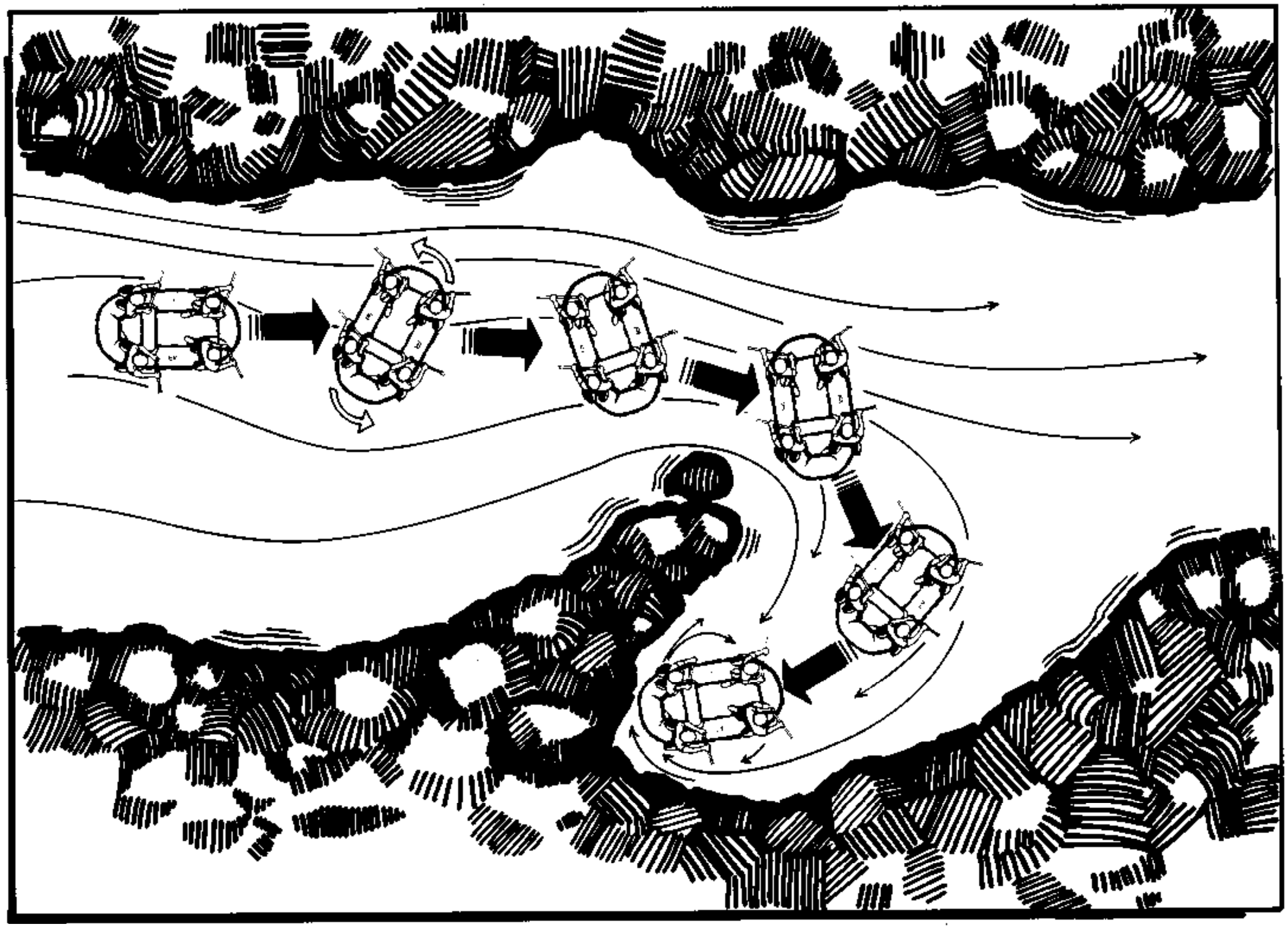

Figure 23-8. Entering an Eddy with a Reverse Ferry.

move with the river, issuing commands precisely and sparingly, working their crews as little as possible.

-2. When using commands and maneuvering the boat in harmony with the river's currents, paddling can be easy and effective, even fun. When instant action is necessary, the captain may say, "paddle at will." When time permits, the captain should introduce commands with a preparatory statement such as, "we're going to ferry to the right of that big rock. OK—(gives command)." This gives the crew time to prepare for the next response. If a command is not heard or understood, it should be repeated with zest until it is understood. If a raft member spots a better way through a rapid or channel, a fully extended arm is used to point it out. This signal, like the others agreed upon, should be repeated until it is understood.

23-4. River Hydraulics. An understanding of river hydraulics is important to the survivor. A knowledge of the types of obstacles and why they should be avoided or overcome is necessary for a safe river journey.

a. Laminar Flow. The drag produced when moving water flows over or past various types of objects and surfaces is called a laminar flow (figure 23-12). The laminar flow principle is that various layers or channels of water move at different speeds. The lower layer of the river moves more slowly than the top layer. This is due to the friction on the bottom and sides of the river which is caused by soil, vegetation, or contours of the riverbank. The layers next to the bottom and sides are the slowest; each subsequent layer will increase in speed. The top layer of the river is only affected by the air. The fastest part of the flow on smooth straight stretches of water will be between 5 and 15 percent of the river depth below the surface. Even straight running riverbeds are not smooth; they have jutting and receding banks on the sides which affect the laminar flow. The friction caused by the banks causes the sides of the flow to be slower than the midstream. The areas near the banks are more shallow and have fewer layers. When the river travels at 4 to 5 knots, turbulence begins to develop which interferes with the regular flow of the current. When this rate of riverflow is achieved, the

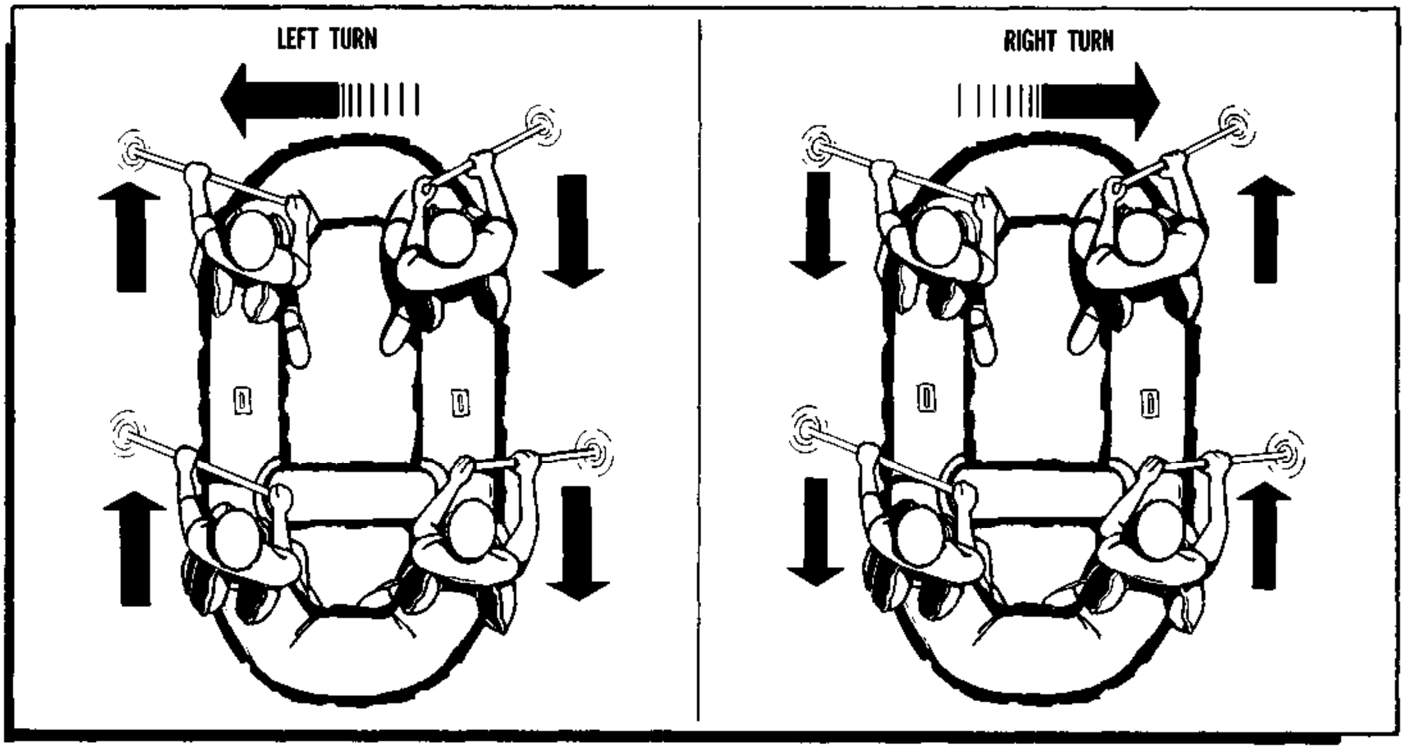
LEFT TURN
RIGHT TURN

Figure 23-9. Turning the Raft to the Right or Left.

friction between the layers of water will cause whirling and spinning actions which agitates the smooth flow of water, creating more resistance.

b. River Currents:

(1) When a current of a river is deflected by obstructions, the overall downward flow of a river will

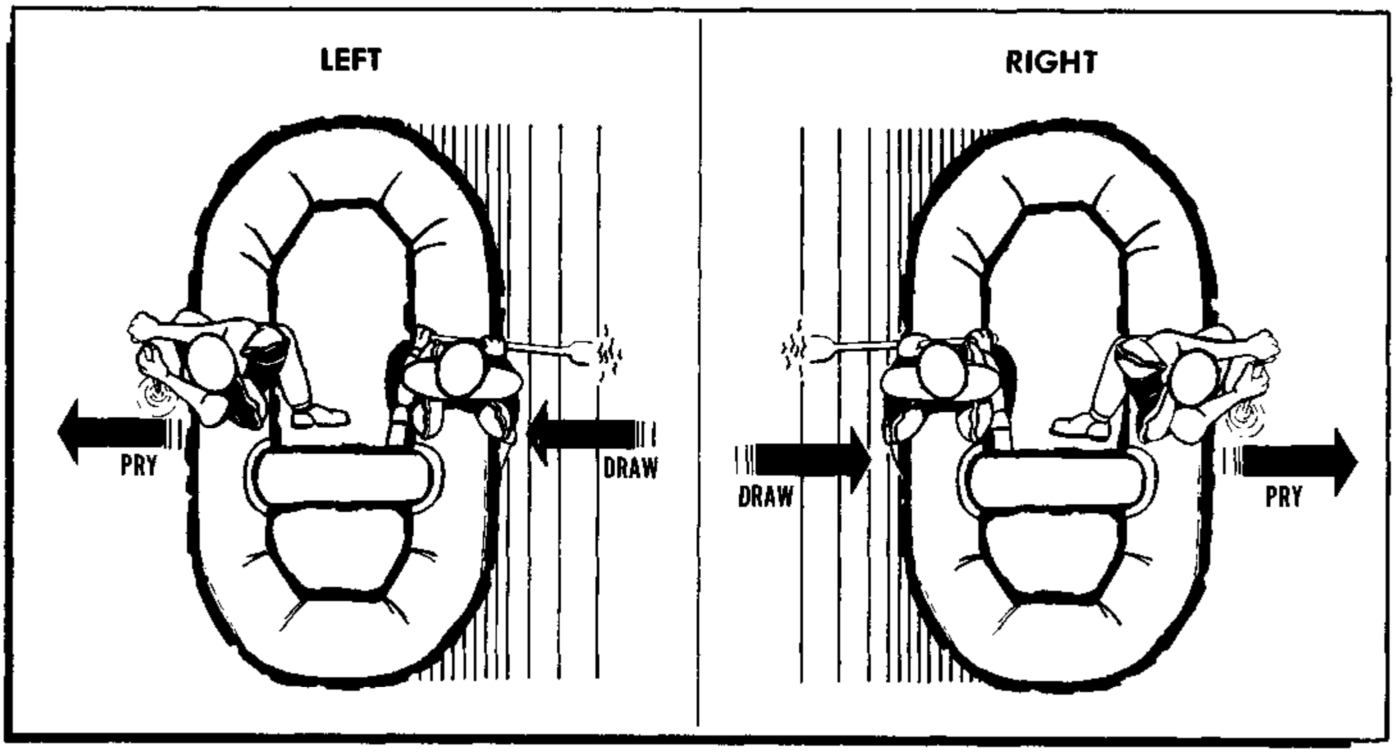
LEFT
RIGHT
PRY
DRAW
DRAW
PRY

Figure 23-10. Making the Raft Move Left or Right with Pry and Draw Stroke.

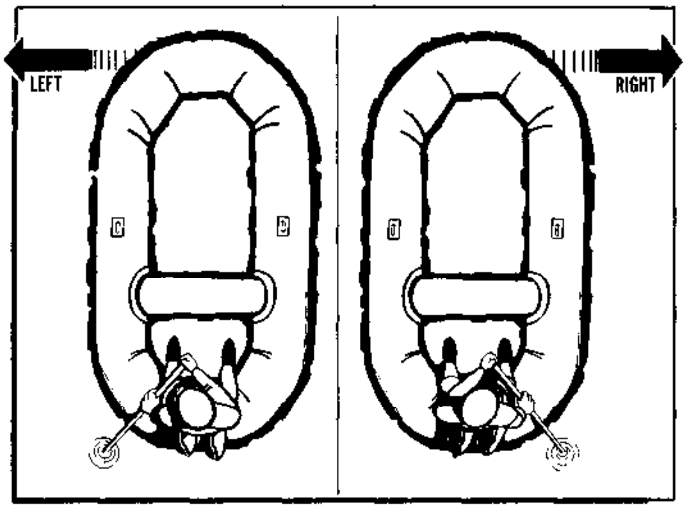

Figure 23-11. Stern Maneuvers.

respond. These responses vary from mild to radical deflection, creating direction and speed changes of waterflow. These changes are called reflex current. The reflex current responds to an obstruction such as bends or submerged rocks.

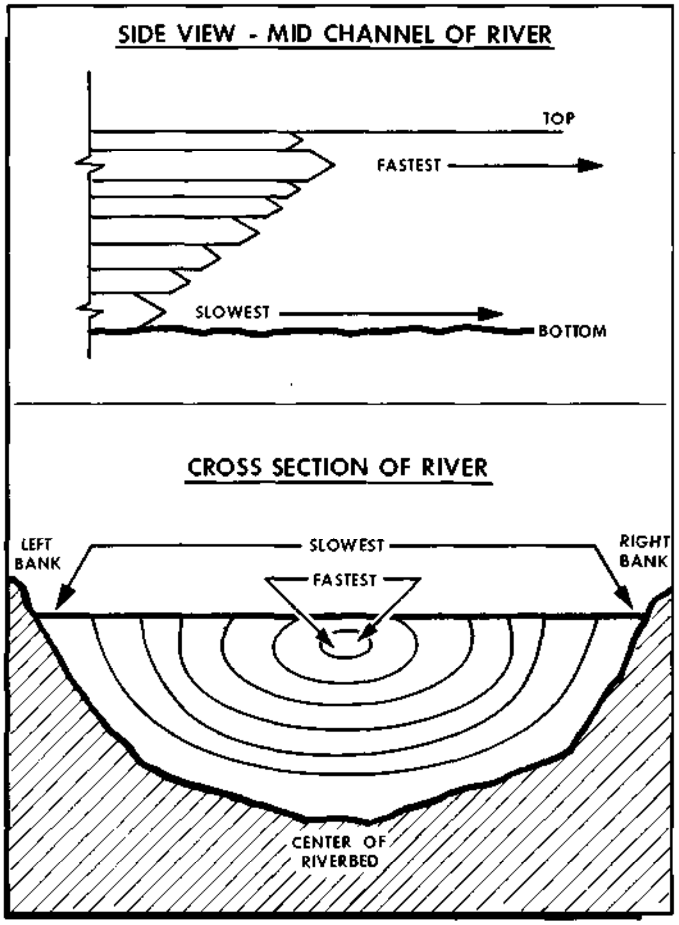
SIDE VIEW - MID CHANNEL OF RIVER
TOP
FASTEST
SLOWEST
BOTTOM
CROSS SECTION OF RIVER
LEFT BANK
SLOWEST
RIGHT BANK
FASTEST
CENTER OF RIVERBED

Figure 23-12. Laminar Flow.

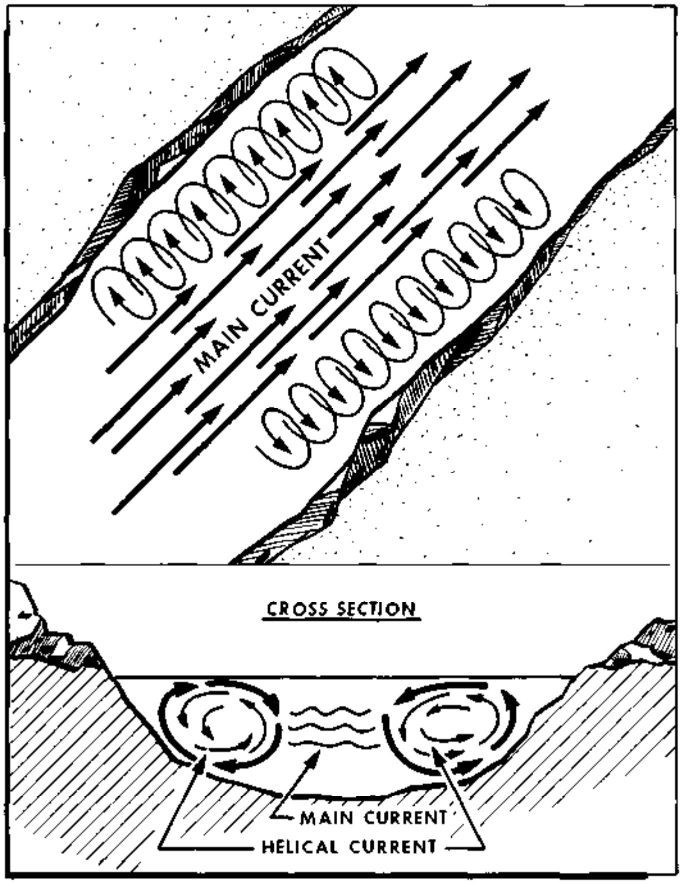
MAIN CURRENT
CROSS SECTION
MAIN CURRENT
HELICAL CURRENT

Figure 23-13. Helical Flow.

(2) One response to the laminar flow is a spiraling, coil-springing flow called a helical current (figure 23-13) which corkscrews as a result of the friction with the riverbank. Going downriver, on the left side of the supposed straight-line river, the helical flow turns clockwise to the main current, and on the right side the helical flow is counterclockwise. This results from friction and drag caused by shallowing banks combined with the strong force of the main current flowing down. The helical flow and the mainstream create a circular, whirling secondary current which travels down along a line near the point of maximum flow. Helical current flow starts along the bottom of the river going out toward the riverbank, surfacing, and then spiraling back into the mainstream at a downward angle. This flow causes floating objects around the edges to be pulled into the mainstream and held there. By understanding where the fast water is and how to observe the characteristics which show the current, a survivor can maneuver the raft to take advantage of the faster water to increase the rate of travel. Even at the quietest edge of a flow, particles are still drawn into the strongest part of the current. Laminar and helical flows are always present in fast-flowing rivers.

(3) The main channel is the deepest part of the river and can wander from bank to bank. The turbulence

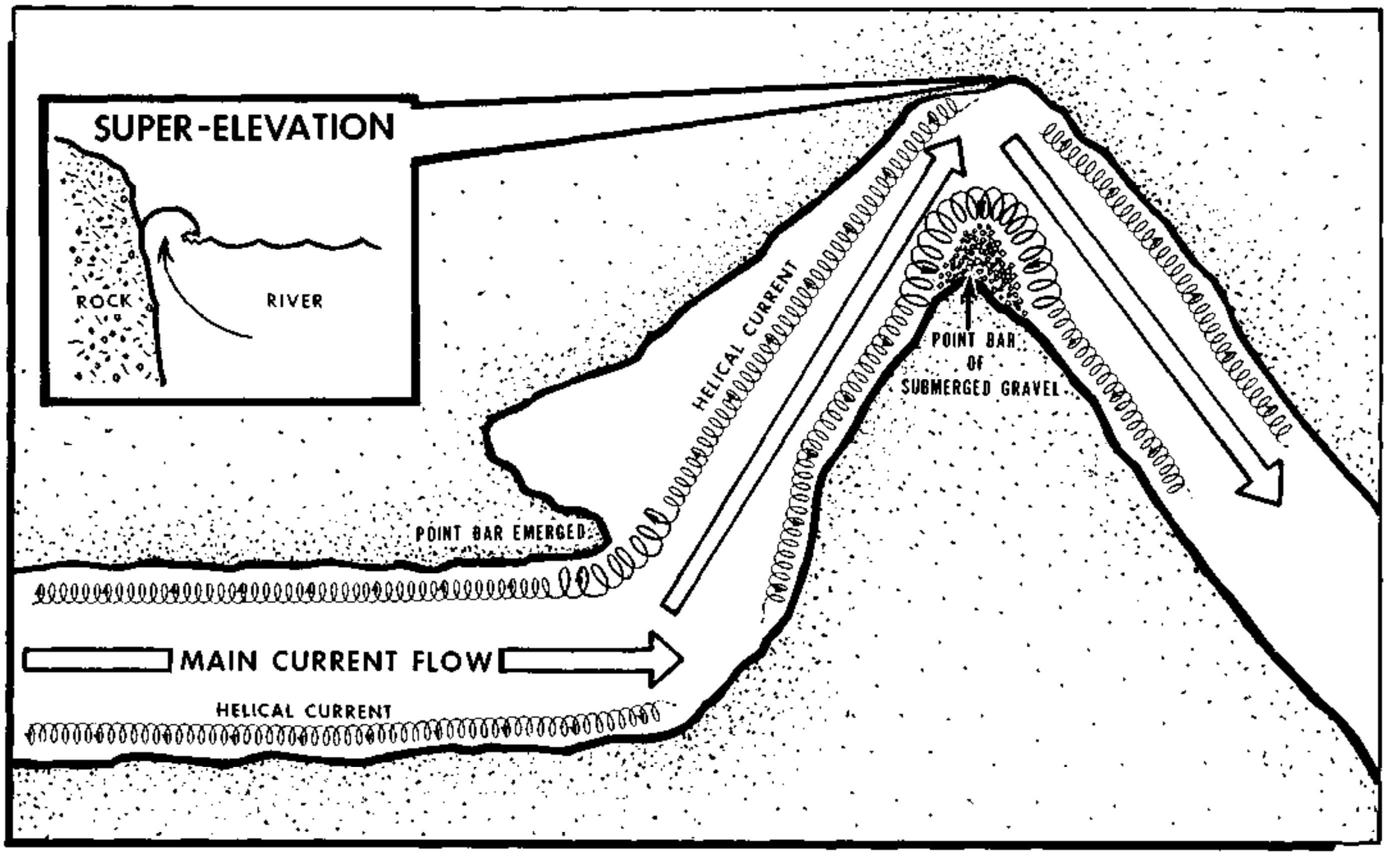

Figure 23-14. Macroturbulence.

caused by the wandering main current erodes wide curves into sharper, more defined bends, creating indirect courses.

(4) When the river makes sharp turns, the current is affected by centrifugal force swinging it wide into the outside bank. The helical current diminishes, being smothered by the laminal flow, thereby increasing the corkscrew effect on the inside of the curve. The surface water is being whirled hard in the direction of the outside curve of the bank; the faster the waterflow, the stronger the push. Floating objects are forced, with the surface water, to the outside of the curve and into the banks, usually getting lodged against and onto the shore.

(a) A powerful helical flow not only pushes the surface outward, but as it swirls up from the bottom it carries sediment up with it. The sediment and other debris is deposited at the highest point of the inside bank of the bend. The sediment is then dropped during high water, and when the water recedes, a point bar (made of sand and gravel) is revealed. The point bar generally sticks out far enough to funnel floating objects into the swiftest part of the river during high waters, avoiding the sandbar.

(b) Super-elevation is a feature where the water is being increased in volume, intensity, and height. When both the stream volume and movement are high, centrifugal force exerts another type of influence on flow characteristics. The river surface water tends to curve in a dish shape towards the outer bend, like a banked turn on a racetrack. The dished inside curve is the easiest and safest route to travel through. If maneuvered correctly, the slight rise of the water and the force of the current around the curve will cause the raft to slip gently off the wave and into the quiet pools of water below. But if the raft was maneuvered across the line of the currents, the raft may either be sucked under by the dominant helical flow, or the power and force of the river on the outside of the dish on the curve could smash and pin any floating device against the outside bank.

(5) Macroturbulence is any extreme, unpredictable turbulence (figure 23-14). It is an especially dangerous phenomenon caused by a drop or decline in the river bottom. The phenomenon also occurs when the water comes in contact with the river bend or rocks. The extreme amounts of froth created from the turbulence and the gravitational pull cause rafts to spin. Raft control is extremely limited because the lack of water viscosity causes resistance against paddles and a lack of buoyancy, making it difficult to float or maneuver. This type of

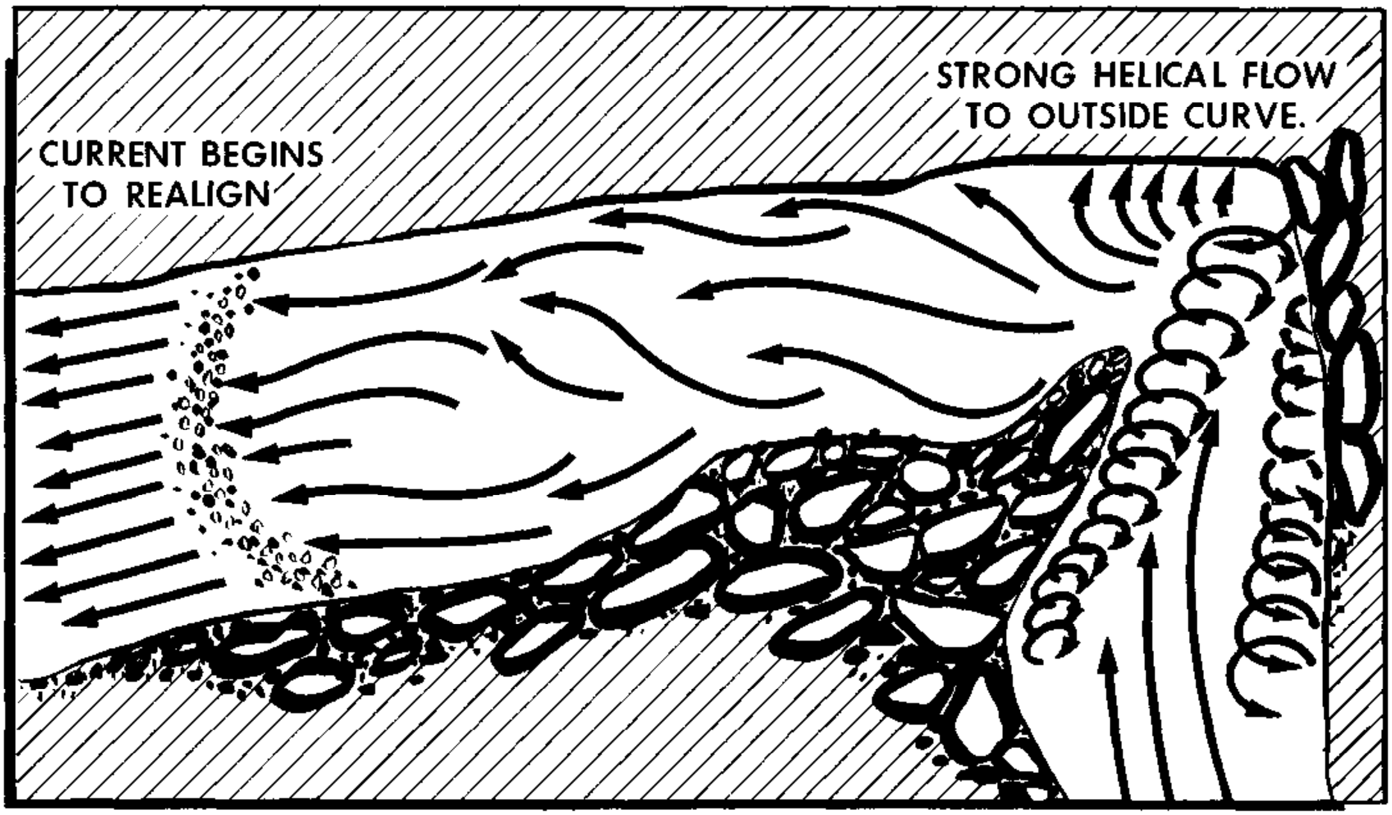
STRONG HELICAL FLOW TO OUTSIDE CURVE.
CURRENT BEGINS TO REALIGN

Figure 23-15. Anatomy of a River Bend.

white water can be impassable, depending upon how extreme the dip and amount of waterflow.

(6) Coming out of a sharp bend, the river currents are mixed, but the dangerous movement is still pulling. The result of the laminar flow shooting into a bank creates a helical flow effect immediately below the turn, where the river is still trying to assume a natural "straight" flow. Being a liquid, water cannot resist stress and it responds to a variety of obstacles (most common are submerged boulders) (figure 23-15).

(7) When water flows over obstructions, such as submerged boulders, the character of the laminar flow is changed. As the water flows over the top of the rock, the layers of the laminar flow increase in speed. This is known as a venturi effect. The hydraulic area is a type of "vacuum" formed as water flows around the rock. Created directly below the obstruction are confused and disordered currents which accelerate the layers of the laminar flow (figure 23-16).

(8) One type of hydraulic is the surge, which usually occurs when the current is slow and the water is deep. This hydraulic is formed downstream from an obstruction with a surge in the water volume. When obstacles no longer have the ability to hold the water back, the pressure is released. Surges present few problems if the boulder or obstruction is covered with enough waterflow to prevent contact when floating over the top. Survivors should be aware of obstructions (known as "sleepers") if the water does not sufficiently cover them (figure 23-17). Failing to recognize a sleeper can result in raft destruction and severe bodily injuries. With large sleepers, the water flows over the top creating a powerful current. This powerful, secondary current is trying to fill the vacuum created by the hydraulic downstream action.

(9) Another form of large sleeper is referred to as a dribbling fall. These are caused by minimal water flowing over submerged obstacles with considerable drop below. This type of sleeper causes a bumpy ride, reducing speed, and can capsize the raft.

(10) Breaking holes occur where a large quantity of water flows over a sleeper and the drop is not steep enough to create a suction hole (figure 23-18). A wave of standing water, much like an ocean breaker, is found downstream. This wave is stationary and can vary from 1 to 10 feet high, and even though it lacks the strong upriver flow of a suction hole, it can be a trap for rafts too small to climb up and over the crest. The size of the breaking hole and the survivor's seamanship must be considered before tackling this obstacle.

c. Suction Hole or Three-Dimensional Eddy:

(1) The vacuum created by a suction hole is strong enough to pull a survivor wearing a life preserver be-

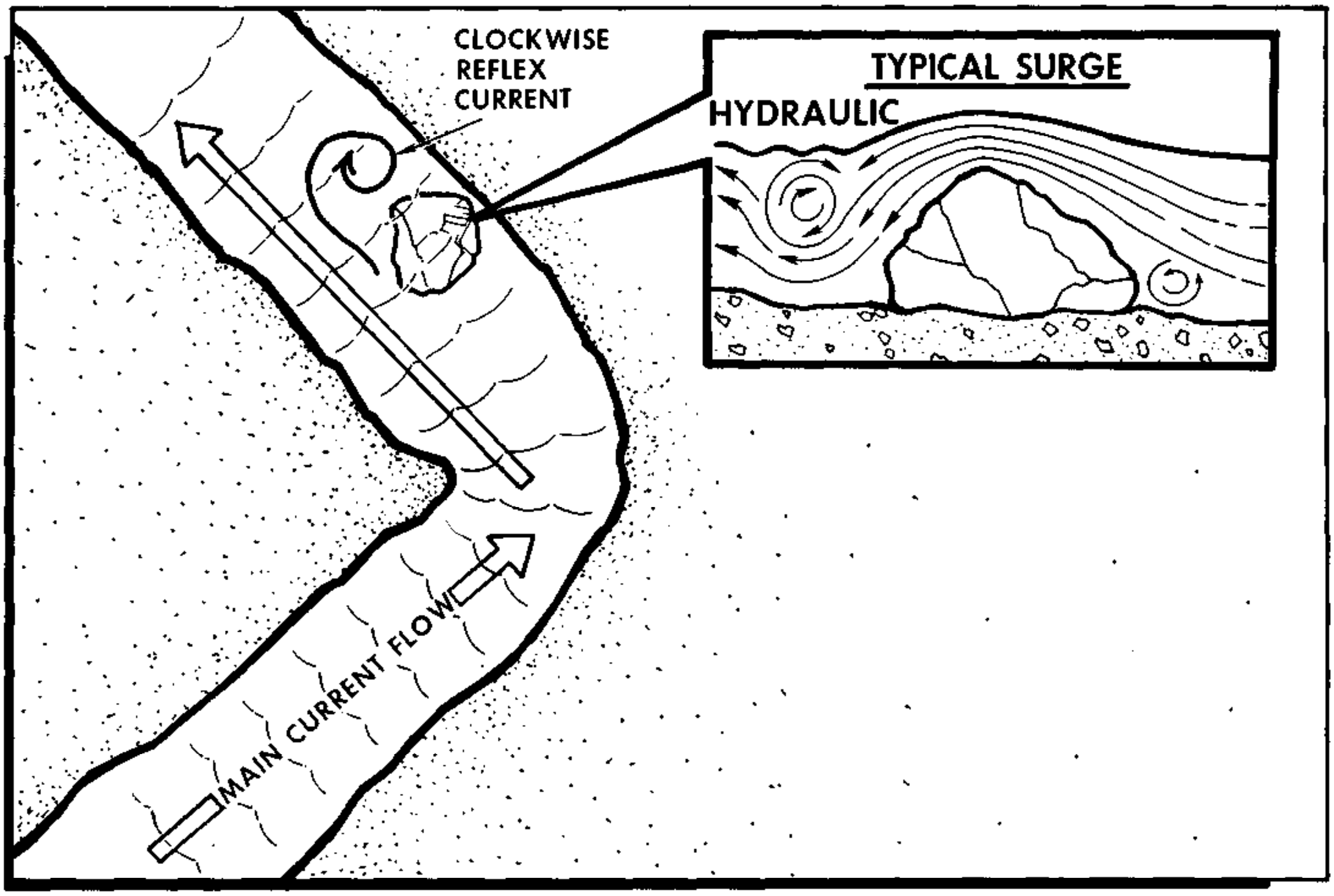
CLOCKWISE REFLEX CURRENT
TYPICAL SURGE
HYDRAULIC
MAIN CURRENT FLOW

Figure 23-16. Response to a Submerged Rock.

neath the surface. If pulled down into a suction hole, the survivor will normally be whirled to the surface downriver and returned to the suction hole by the upriver flow, to be pushed under once again. Objects too buoyant to sink normally remain trapped. It's usually difficult to identify a suction hole because there is no frothing, no obvious curling water, and little noise. Extreme caution should be used when a large bulge appears in the water (figure 23-19). There are three possible ways of surviving and avoiding serious injury in suction holes. One way is to find the layer of water below the surface which is moving in the same desired direction. The second way is to reach down with a paddle or hand and feel for a current which is moving out of the hole. However, in a large suction hole the downstream flow will be too deep to reach. The survivor should attempt to cut across through the side of the eddy into the water rushing by. The final and best solution is to scout ahead and try to identify the location of suction holes and avoid them.

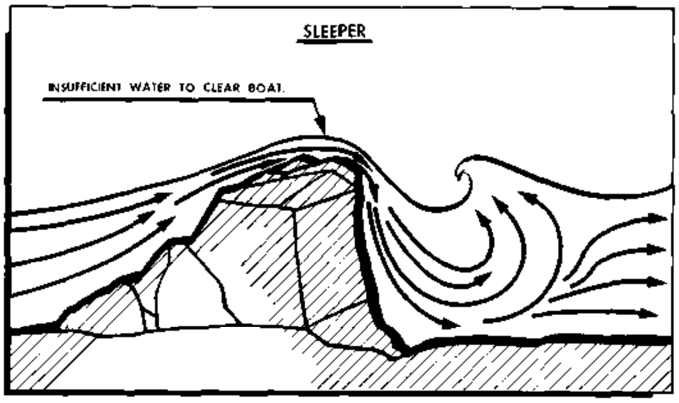
SLEEPER
INSUFFICIENT WATER TO CLEAR BOAT.

Figure 23-17. Sleeper.

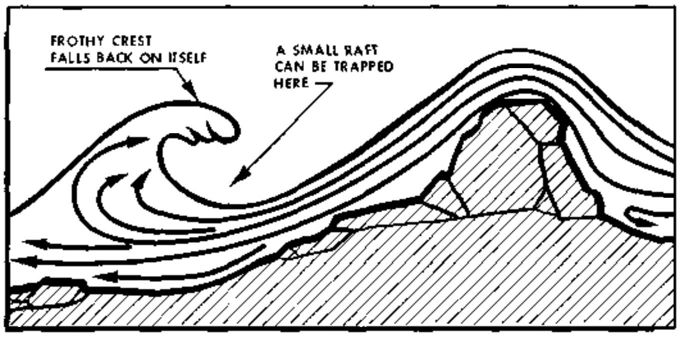
FROTHY CREST FALLS BACK ON ITSELF
A SMALL RAFT CAN BE TRAPPED HERE

Figure 23-18. Breaking Hole.

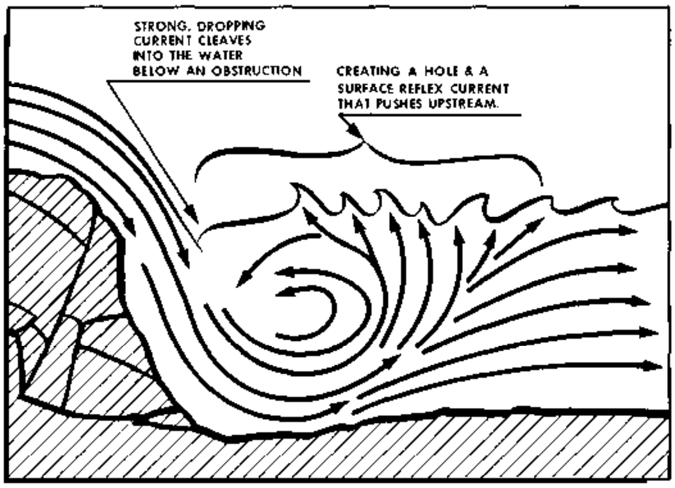
STRONG, DROPPING CURRENT CLEAVES INTO THE WATER BELOW AN OBSTRUCTION
CREATING A HOLE & A SURFACE REFLEX CURRENT THAT PUSHES UPSTREAM.

Figure 23-19. Suction Hole.

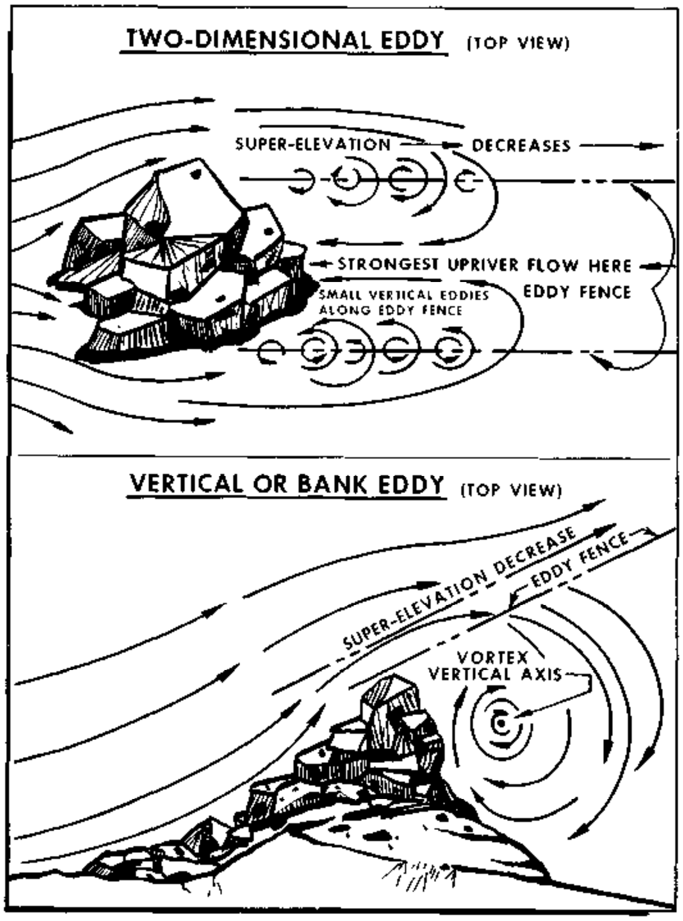
TWO-DIMENSIONAL EDDY (TOP VIEW)

VERTICAL OR BANK EDDY (TOP VIEW)
SUPER-ELEVATION — DECREASES
STRONGEST UPRIVER FLOW HERE
SMALL VERTICAL EDDIES ALONG EDDY FENCE
EDDY FENCE
SUPER-ELEVATION DECREASE
EDDY FENCE
VORTEX VERTICAL AXIS

Figure 23-21. Two-Dimensional Eddies.

(2) An eddy (figure 23-20) is a reaction to an obstruction. The type of eddy which occurs next to the bank is caused by portions of the main current being deflected and forced to flow back upriver where it again joins the mainstream. These areas are usually associated with quiet and slow-flowing water. They are also associated with areas where the river widens or just above or below a bend in the river. An eddy has two distinct currents: the upstream current, and the downstream current. The dividing line between the two is called an eddy fence. It is a line of small whirlpools spun off the upriver current by the power of the downstream current.

(3) A two-dimensional eddy is when the tip of an obstruction is slightly above the water and causes a two-dimensional flow around the obstacle. Because of the speed and power of the current, the water is super-elevated and is significantly higher than the level of water directly behind the obstacle. This creates a hole which is filled by the flow of water around the obstacle (figure 23-21). Two-dimensional eddies which occur in midstream will create two eddy fences, one on each side of the obstacle. The water will enter the depression from both sides and will travel in a circular motion, clockwise from the right bank, counterclockwise from the left bank, and back upstream directly behind the obstacle. If the projection is large enough and a strong circular motion is created, it becomes a whirlpool. The outer reaches of the swirling water are super-elevated by centrifugal force and a suction is created, similar to a drain in a bathtub. These vortexes are very rare and usually occur on huge rivers. Survivors may stop and rest where the eddies occur since there should be no strong swift currents in the eddy. If the obstacle is huge, it may be impossible to paddle fast enough to cross the eddy fence without being spun around in a pinwheeling manner.

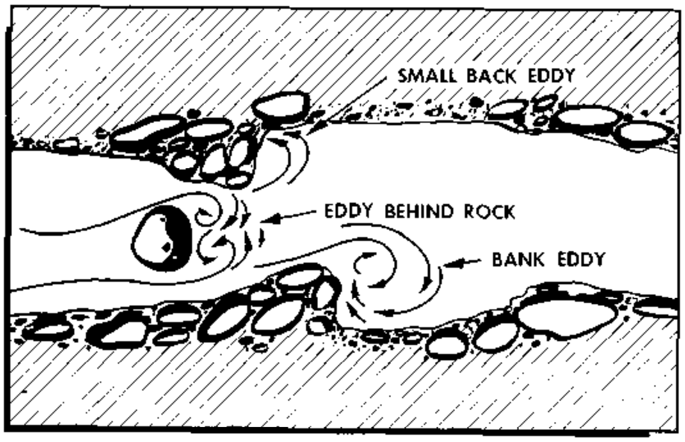
SMALL BACK EDDY
EDDY BEHIND ROCK
BANK EDDY

Figure 23-20. Eddies.

d. Falls. In most falls, there are two reflex currents or suction holes forming, both whirling on crosscurrent axes into the falls (figure 23-22). One current falls behind the crashing main stream of water while the other falls in front. It's not too dangerous on a 2- or 3-foot drop, but when heights of 6 or more feet are present, it could be a death trap. The suction holes formed below these falls are inherently inescapable because of their power. The foam and froth formed at the bottom of falls

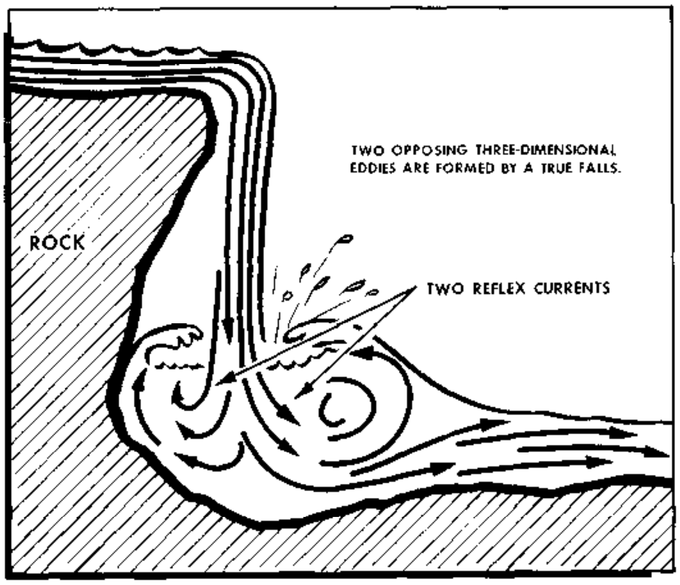
TWO OPPOSING THREE-DIMENSIONAL EDDIES ARE FORMED BY A TRUE FALLS.
ROCK
TWO REFLEX CURRENTS

Figure 23-22. Falls.

can be dangerous. Jagged boulders and other hazards may be hidden.

e. Boils. A boil may occur below a fall or sleeper, downstream of the curling or reverse-current suction. This appears as a dome or mound-shaped water formation. Boils are the result of layers of flows hitting bottom, aimed upward, and reaching the surface parting into a flowerlike flow. The water billowing out in boils is super-oxygenated, taking away the resistance needed to push with the paddles or to suspend a survivor in a life preserver.

f. Rollers. Another difficulty found when traveling on fast rivers are rollers. Rollers are large, cresting waves caused by a variety of situations. A wave seen below a breaking hole is one type of roller. Velocity waves are another type which occur on straight stretches of fast dropping waters and caused by the drag of sandy banks and submerged sandbars. They may be large enough to overturn a raft, but are easily recognized, and are usually regular and easy to navigate by keeping the raft direction of travel in line with the crest of the wave (figure 23-23).

g. Tail Waves. Tail waves are quiet waves which are a reflex from the current hitting small rocks along the bed of the river and deflecting the current toward the surface. They are usually so calm that going over them is not noticeable.

h. Bank Rollers. Bank rollers are similar in appearance to crested tail waves and occur when there is a sharp bend in the riverbed which turns so sharply the water can't readily turn in the bend. The water slams into the outside bank and super-elevates, falling back upon itself. Small bank rollers cause few problems, but large ones cresting 5 feet or higher can capsize a raft (figure 23-24). There are three terms used to specify the severity or height of rollers; one being washboard. Washboard rollers are a series of swells which gently ripple and are safe and easy to ride. The next stage of rollers is called standing water (cresting rollers), where the speed of contours on the bottom are such that the tops of the swells fall back onto themselves. The most dangerous and insurmountable rollers are referred to as haystacks or roosters (figure 23-25). They resemble haystacks of water coming down from every direction—giant frothing bulges. They may be so high a raft cannot go over them. It would be easy to become trapped in a deep trough (depression) and be buried under tons of frothing water. These troughs also hide dangers such as sharp rocks which could tear a raft or break through the bottom of a boat.

i. Chutes. River chutes are good, logical travel routes, but they may harbor dangers. A chute (or tongue) is a

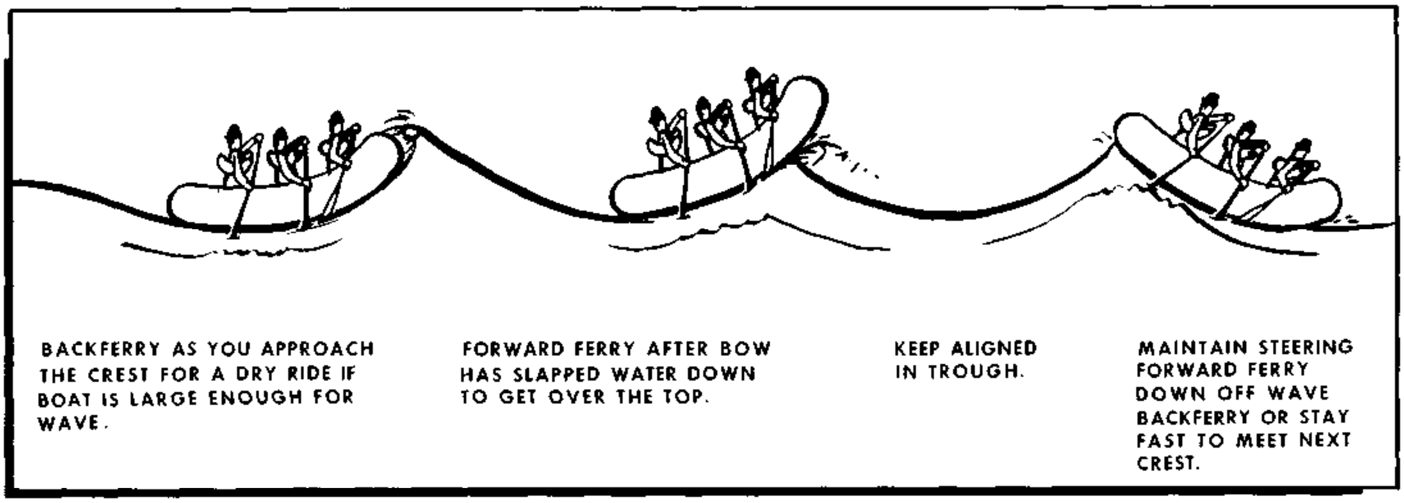

BACKFERRY AS YOU APPROACH THE CREST FOR A DRY RIDE IF BOAT IS LARGE ENOUGH FOR WAVE.

FORWARD FERRY AFTER BOW HAS SLAPPED WATER DOWN TO GET OVER THE TOP.

KEEP ALIGNED IN TROUGH.

MAINTAIN STEERING FORWARD FERRY DOWN OFF WAVE BACKFERRY OR STAY FAST TO MEET NEXT CREST.

Figure 23-23. Tail Waves and Rollers.

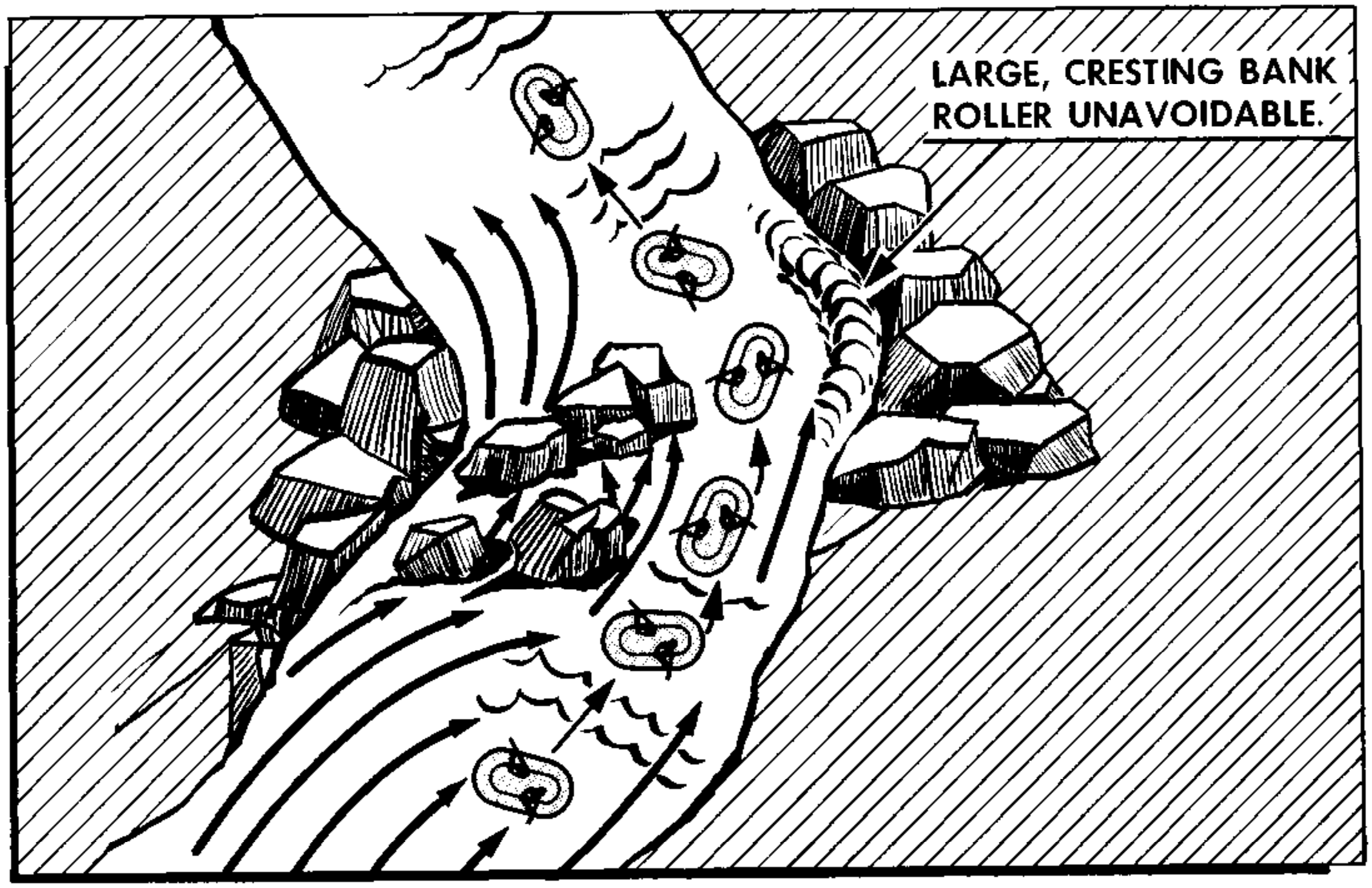
LARGE, CRESTING BANK ROLLER UNAVOIDABLE.

Figure 23-24. Bank Roller.

swift-running narrow passage between river obstructions caused by a damming effect. An example would be water forced between two large boulders. Because the waterflow is restricted, it accelerates and a powerful current rushes through. Because of the water velocity, there may be a suction hole on either side of the chute (figure 23-26).

j. Log Jams (Tongue of the Rapids). Log jams are extremely dangerous. They consist of logs, brush, and debris collected from high waters that become lodged across the current. They remain stationary while the river flows through them. If a craft should be swept up against the stationary logs, it will be pinned in place by the current. Should the craft be tipped and swamped, it could be swept under the log jam. If the current is strong

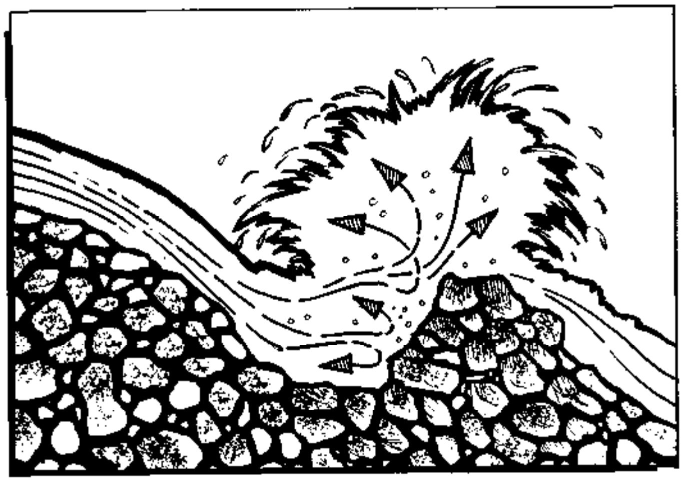

Figure 23-25. Haystack.

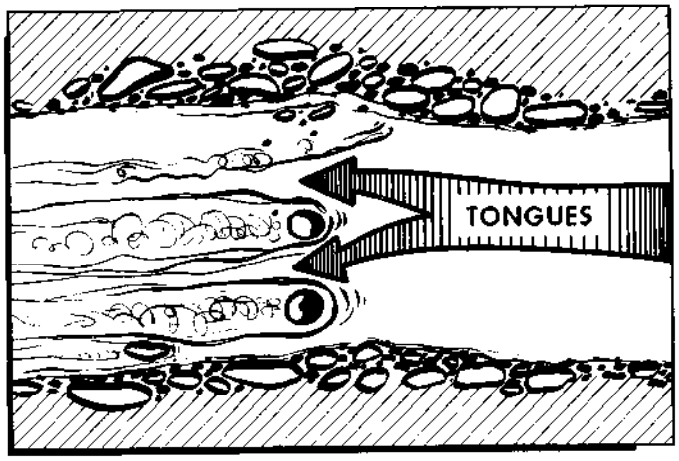

Figure 23-26. Tongue of the Rapids.

enough to do this, the occupants may also be pinned underneath the jam.

k. Sweepers. Sweepers can be the most dangerous obstructions in rivers. A sweeper is a large tree growing on a riverbank which has fallen over and is resting at or near the surface of the water. It may bounce up and down with the current. Survivors may be suddenly confronted with a sweeper which blocks the channel while rounding a bend in a river. The survivors are relatively helpless when it encounters sweepers in swift water. The only precautionary measure is to land above a bend in order to study the river ahead. (NOTE: Many people have met disaster by hitting sweepers.)

23-5. Emergency Situations:

a. Rock Collisions. Collisions with rocks above the surface of the water are common occurrences on a river. If the collision is unavoidable, survivors should spin the raft powerfully just before contact, or hit the rock bow-on. If the survivor is able to spin the raft, it will usually turn the raft off and around the rock. If they hit the rock bow-on, it will stop the raft momentarily, giving time to manipulate a spin-off with a few turn strokes. When the bow-on method is used, occupants in the stern (aft) should move to the center of the raft before impact. This allows the stern to raise and the rushing current to slide under the raft. If the stern is low, the water will pile against it causing the raft to be swamped. If survivors are colliding broadside with a rock, the entire crew must immediately jump to the side of the raft nearest the rock—always being the raft's downstream side (figure 23-27). This should be done before contact. If not, the river will flow over and suck down the raft's upstream tube. The raft will be flooded and the powerful force of the current will wrap it around the rock, possibly trapping some or all of the crew between the rock and the raft. Once it is unswamped, the broached (broadsided) raft on a rock is usually freed easily (figure 23-28). If two people push with both feet on the rock in the direction of the current, the raft will swing or slide into the pull of the current. The rest of the crew should shift to the end which is swinging into the current. These methods rarely fail; however, if the raft refuses to budge, it can be freed using the enormous power of the current. Large gear bags or sea anchors are securely tied to a long rope and secured to the end of the raft expected to swing downstream. The sea anchors (gear bags) are then tossed downstream into the current. (NOTE: A safety line should be used.)

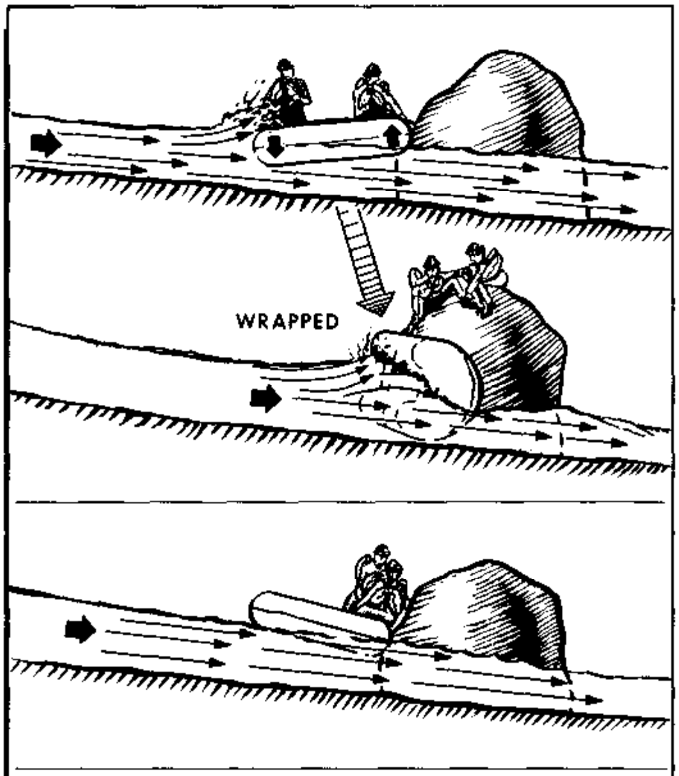

HANDLING A BROADSIDE COLLISION WITH A ROCK

JUST BEFORE A RAFT BROADSIDES ON A ROCK, THE CREW SHOULD JUMP TO THE SIDE OF THE RAFT NEAREST THE ROCK. THIS LIFTS THE UPSTREAM TUBE AND ALLOWS THE CURRENT TO SLIDE EASILY UNDER THE RAFT. IF THE CREW IS NOT QUICK TO HOP TOWARD THE ROCK, SWIFT WATER WILL SPEEDILY SWAMP THE RAFT AND WRAP IT FLAT AROUND THE ROCK'S UPSTREAM FACE.

Figure 23-27. Handling Broadside Collision with Rock.

b. Freeing a Wrapped Raft. Sometimes the powerful force of swift water may pin and wrap a raft around a rock. It is unusual for a raft to be equally balanced around the rock, so it will move more easily one way than it will the other. The part of the raft with more weight and bulk should be moved toward the flow of the current. Lines are attached to at least two points on the raft so, when pulled, the force is equally distributed. One of these points is on the far end of the raft, around the tube and is called the hauling line. (NOTE and CAUTION: A small hole may need to be cut in the floor of the raft to pass the line through and around the tube

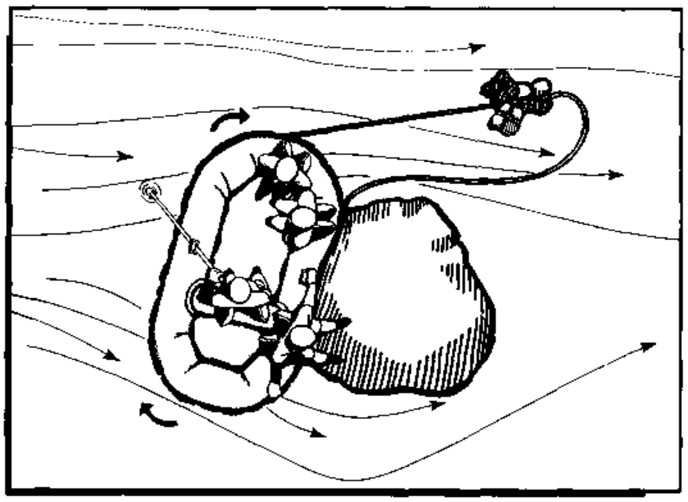

Figure 23-28. Freeing Unswamped Raft.

if there is no ring to pass it through.) The second tie-off point can be a cross tube. One person should hold the line attached to the stern of the raft. The raft should be moved over the rock into the pull of the current. Once it is freed, the person holding the stern line can move the raft to a safe position.

c. Raft Flips. When a raft is about to flip, there is little time to react to the situation. If the raft is diving into a big hole, the primary danger is being violently thrown forward into a solid object in the raft. Survivors should protect themselves by dropping low and flattening themselves against the backside of the baggage or cross tube. If the raft is being upset by a rock, fallen tree, or other obstacles, members should jump clear of the raft to prevent being crushed against the obstruction or struck by the falling raft. If the raft is pinned flat against an obstruction, the members should stay with the raft and try to safely climb up the obstruction (figure 23-29).

d. Lining Unrunnable Rafts. Lining a raft through rapids is basically letting it run through rapids with a crew on shore controlling it by attached lines (figure 23-30). The raft should be moved slowly by maintaining tight control of the lines attached to the bow and stern. If no strong eddies or steep narrow chutes are present, one member should walk along the shore and control the raft with a strong, long line. The lines running to shore should be long enough to allow the raft full travel through the rapids.

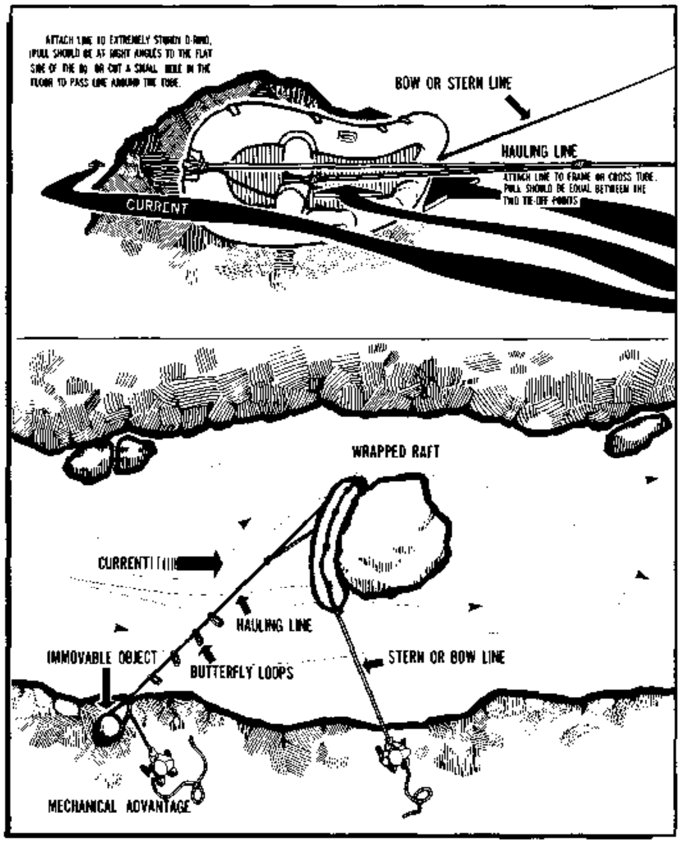
ATTACH LINE TO EXTREMELY STURDY D-RING. (PULL SHOULD BE AT RIGHT ANGLES TO THE FLAT SIDE OF THE D) OR CUT A SMALL HOLE IN THE FLOOR TO PASS LINE AROUND THE TUBE.
BOW OR STERN LINE
HAULING LINE
ATTACH LINE TO FRAME OR CROSS TUBE. PULL SHOULD BE EQUAL BETWEEN THE TWO TIE-OFF POINTS
CURRENT
WRAPPED RAFT
CURRENT
HAULING LINE
IMMOVABLE OBJECT
BUTTERFLY LOOPS
STERN OR BOW LINE
MECHANICAL ADVANTAGE

Figure 23-29. Freeing a Wrapped Raft.

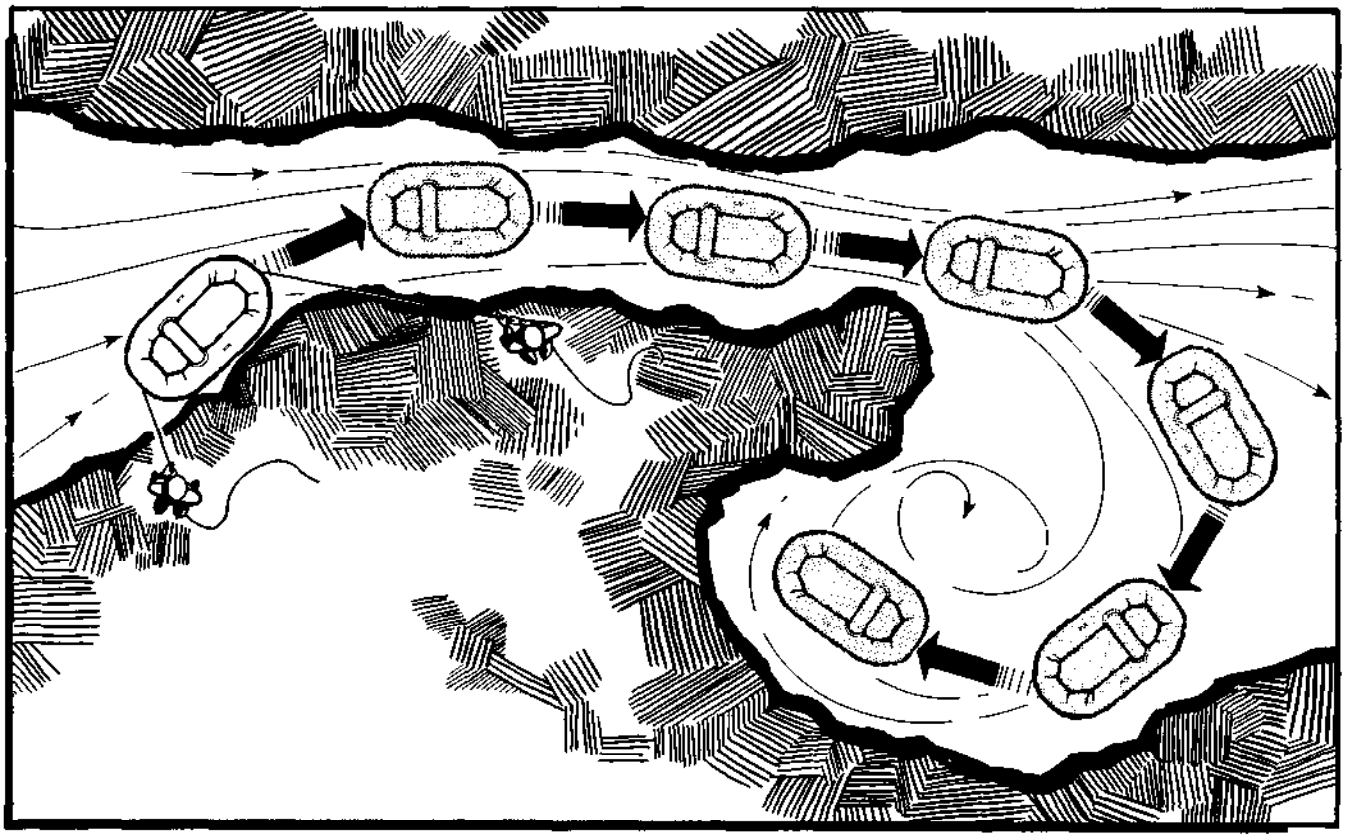

Figure 23-30. Lining a Chute.

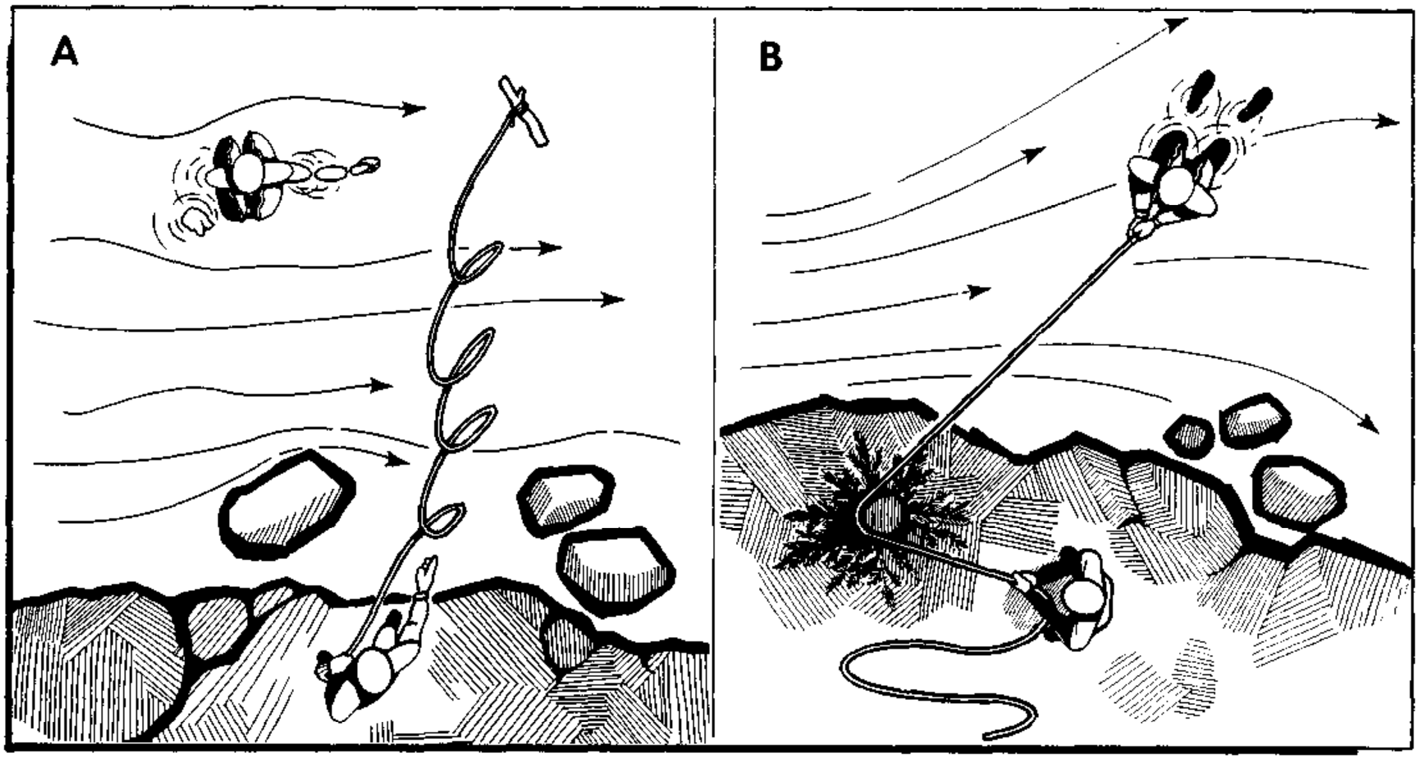

Figure 23-31. Making a Rescue.

e. Rescuing a Swimmer from Shore. The rescuer should carefully choose the right spot where the coil of rope thrown to the swimmer will not cross hazardous areas or obstacles, and yet be near a rock or tree which can be used for belaying (securing without being tied) the end of the line. The person throwing the line should make sure the line has a flotation device (life preserver) at the end before it is thrown to the swimmer. The rope should be coiled in a manner which will allow it to flow smoothly, without entanglement, to full extension. One hand holds one-half to one-third of the coil while the other throws the remainder of the coil out. The weighted coil should be thrown to a spot where the swimmer will drift, which is usually downstream, in front of the swimmer. As the rope travels out, all of the line except the last 10 or 12 feet should be uncoiled. The member

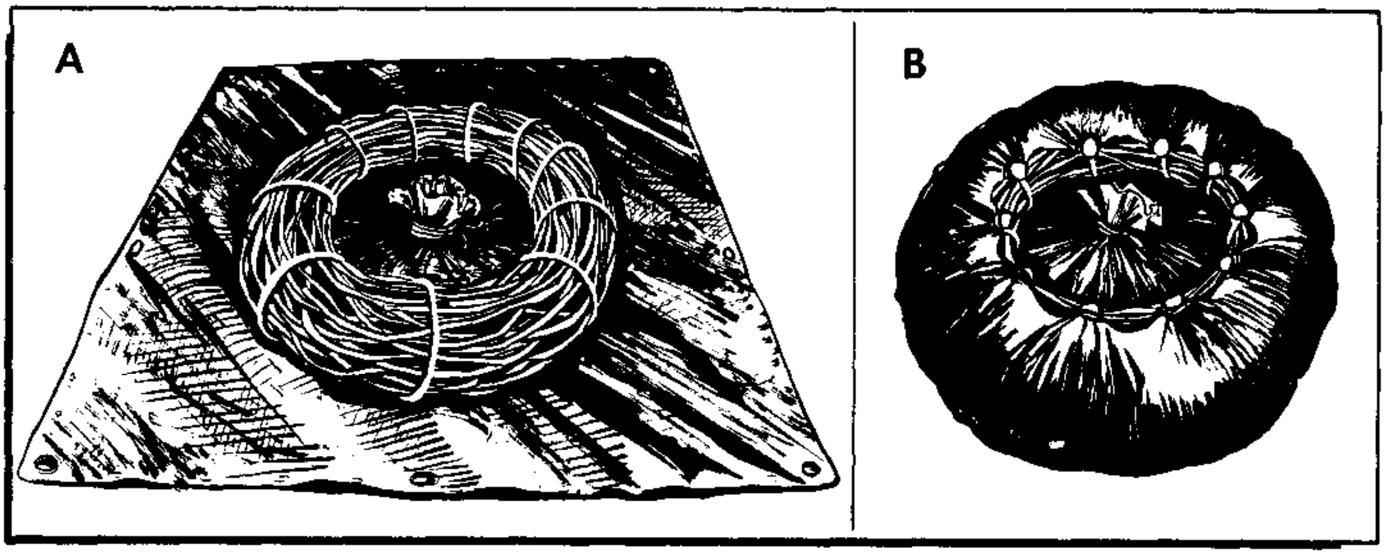

Figure 23-32. Poncho Ring Raft.

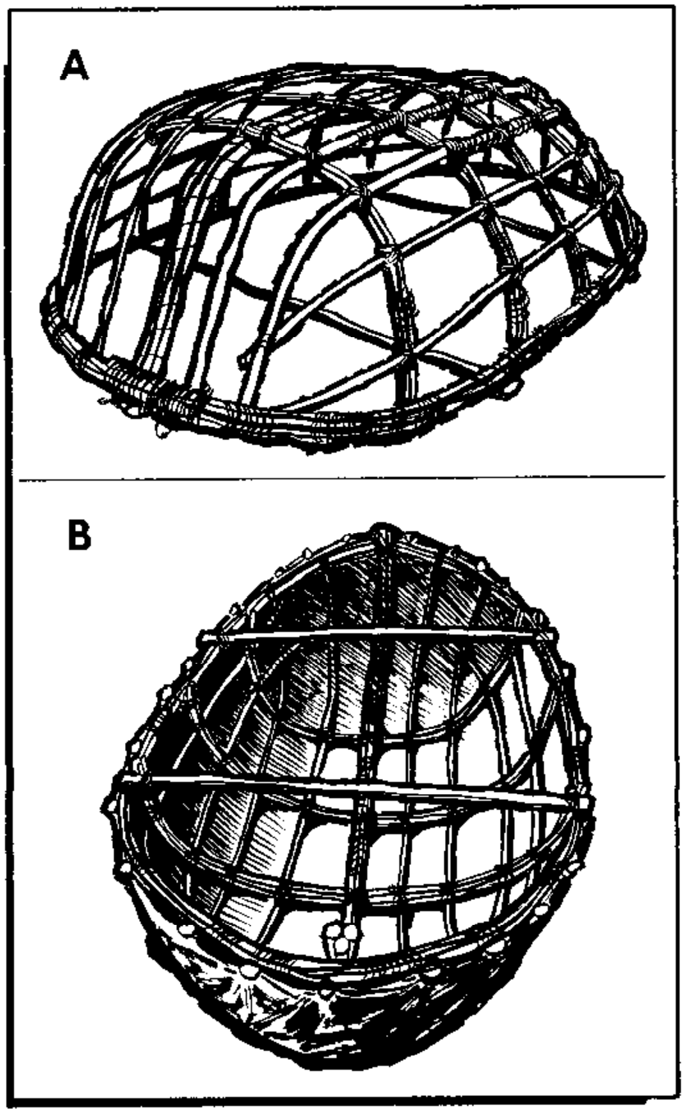

Figure 23-33. Bull Boat.

pulling the survivor should be braced or have the line around a rock or tree to hold the swimmer once the line has been reached by the survivor.

f. The Swimmer's Responsibility. The swimmer should be aware of the rescuer's location and face downstream when waiting for the hauling line. When the line is thrown, the swimmer will normally be required to swim to reach it. Once holding the line, the swimmer should be prepared for a very strong pull from the current and line. The line should be held tightly, but not wrapped around a wrist or hand! Entanglement must be avoided. The swimmer should pull, hand-over-hand, until reaching shallow water, and then use the rope for steadiness while walking to shore (figure 23-31).

23-6. Improvised Rafts:

a. Types of Flotation Devices. There are various types of flotation devices which may be improvised and used as rafts for equipment and personnel.

(1) The Poncho (donut) raft (figure 23-32) can be used for transporting equipment, but is not a good vehicle for people. The raft is constructed by using saplings or pliable willows and a waterproof cover. A hoop-shaped framework of saplings or pliable willow is constructed within a circle of stakes. The hoop is tied with cordage or suspension line and removed from the circle of stakes and placed on the waterproof cover to which it will be attached. Clothing and (or) equipment is then placed in the raft and the survivor swims, pushing the raft.

(2) The bull boat (figure 23-33) is a shallow-draft skin boat shaped like a tub and formerly used by Indians in the Great Plains area. The survivors should construct an oval frame, similar to a canoe, of willow or other pliable materials and cover the framework with waterproof material such as a signal paulin or skins. This makes a craft which is suitable for transporting equipment across a river with the survivor propelling it from behind.

(3) An emergency boat can be made by stretching a tarpaulin or light canvas cover over a skillfully shaped framework of willows and adding a well-framed keel of green wood, such as slender pieces of spruce. Gunwales (sides) of slender saplings are attached at both ends and the spreaders or thwarts are attached as in a canoe. Ribs of strong willows are tied to the keel. The ends of the ribs are bent upward and tied to the gunwales. The inside of the frame is closely covered with willows to form a deck upon which to stand. Such a boat is easy to handle and is buoyant, but lacks the strength necessary for long journeys. This boat is entirely satisfactory for ferrying a group across a broad, quiet stretch of river. When such a boat has served its purpose, the cover should be removed for later use.

(4) The vegetation raft is built of small vegetation which will float and is placed within clothing or parachute to form a raft for a survivor and (or) equipment. Plants such as water hyacinth or cattail may be used (figure 23-34).

(5) A good floating device for the single survivor can be fabricated by using two balsa logs or other lightweight wood. The logs should be placed about 2 feet apart and tied together. The survivor sits on the lines and travels with the current (figure 23-35).

(6) A dugout canoe is good transportation, but difficult to construct. One method is to build a long fire on the side to be dug out and chop away the burned material when the fire is out. Repeat this procedure as often as necessary.

b. Building a Raft. The greatest problem in raft construction (figure 23-36) is being able to construct a craft strong enough to withstand the buffeting it may have to take from rocks and swift water. Even if 6- to 8-inch

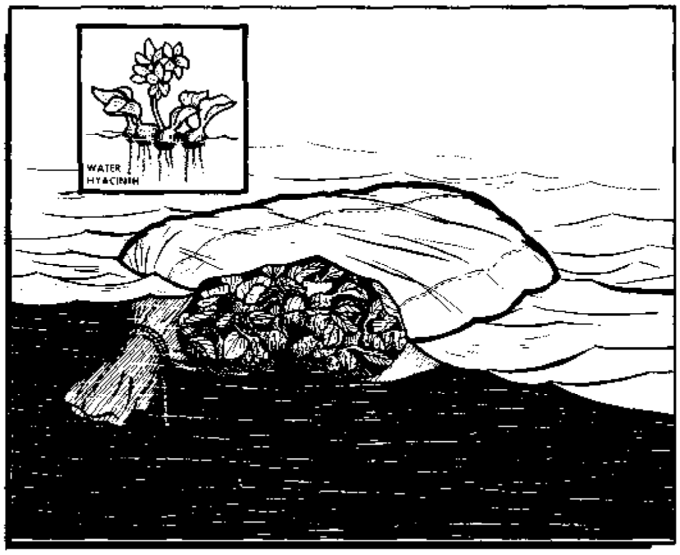

Figure 23-34. Vegetation Bag.

spikes are available, they are not satisfactory since they pull or twist out easily. Rope quickly wears out from frequent, rough contact with rocks and gravel. Northern woodsmen have evolved a construction method (figure 23-37) which requires neither spikes nor rope, yet produces a raft superior in strength. The only material required are logs, although rope is sometimes useful; the only tools needed are an axe and a sheath knife.

23-7. Fording Streams:

a. Survivors traveling on foot through wilderness areas may have to ford some streams. These can range from small, ankle-deep brooks to large rivers. Rivers are often so swift a survivor can hear boulders on the bottom being crashed together by the current. If these streams are of glacial origin, the survivor should wait for them to decrease in strength during the night hours before attempting to ford.

b. Careful study is required to find a place to safely ford a stream. If there is a high vantage point beside the river, the survivor should climb the rise and look over the river. Finding a safe crossing area may be easy if the river breaks into a number of small channels. The area on the opposite bank should be surveyed to make sure travel will be easier after crossing. When selecting a fording site, the survivor should:

(1) When possible, select a travel course which leads across the current at about a 45-degree angle downstream.

(2) Never attempt to ford a stream directly above, or close to, a deep or rapid waterfall or a deep channel. The stream should be crossed where the opposite side is comprised of shallow banks or sandbars.

(3) Avoid rocky places, since a fall may cause serious injury. However, an occasional rock which breaks the current may be of some assistance. The depth of the water is not necessarily a deterrent. Deep water may run more slowly and be safer than shallow water.

(4) Before entering the water, the survivors should have a plan of action for making the crossing. Use all possible precautions, and if the stream appears treacherous, take the steps shown in figure 23-38.

23-8. Traveling on Open Seas. Four-fifths of the Earth's surface is covered by open water. Although accounts of sea survival incidents are often gloomy, successful survival is possible; however, the raft is at the mercy of the currents and winds. There are many currents, both warm and cold, throughout the seas.

a. Currents. Sea currents flow in a clockwise direction in the Northern Hemisphere and counterclockwise in the Southern Hemisphere. This is caused by three factors: the Sun's heat, the winds, and the Earth's rotation (Coriolis Effect). Most sea currents travel at speeds of less than 5 miles per hour. Using currents as a mode of travel can be done by putting out the sea anchor and letting the current pull the raft along. Survivors should

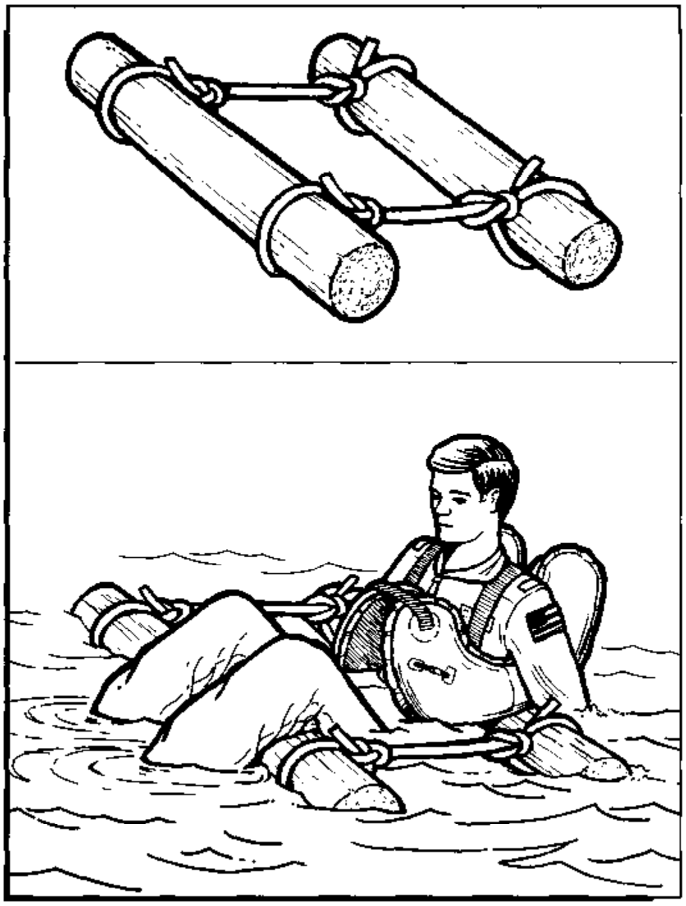

Figure 23-35. Log Flotation.

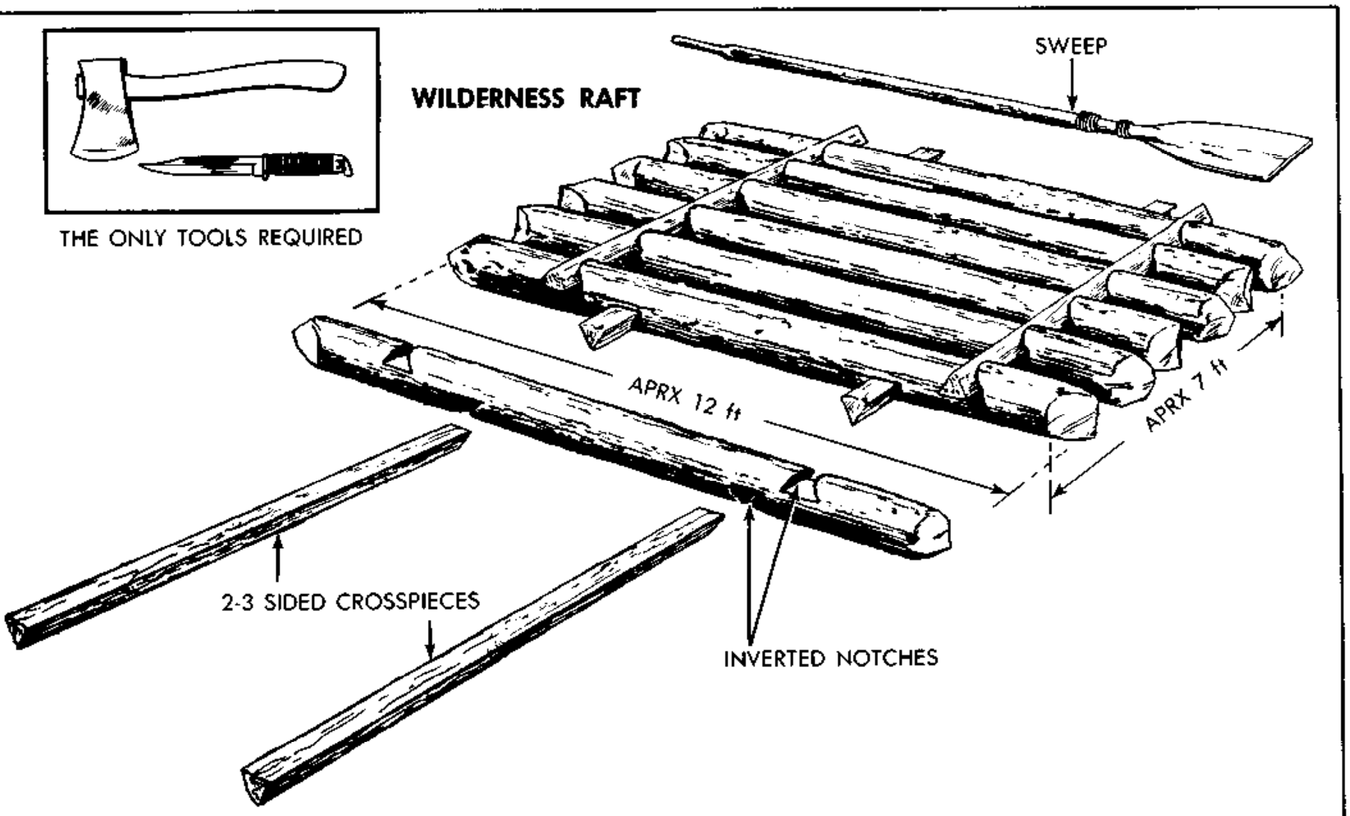

RAFT CONSTRUCTION

A raft for three persons should be about 12 feet long and 6 feet wide, depending on the size of the logs used. The logs should be 12 to 14 inches in diameter and so well matched in size that notches you make in them are level when crosspieces are driven into place.

Build the raft on two skid logs placed so that they slope downward to the bank. Smooth the logs with an ax so that the raft logs lie evenly on them. Cut two sets of slightly offset inverted notches, one in the top and bottom of both ends of each log. Make the notches broader at the base than at the outer edge of the log, as shown in the illustration. Use small poles with straight edges or a string pulled taut to make the notches. A three-sided wooden crosspiece about a foot longer than the total width of the raft is to be driven through each end of the four sets of notches.

Complete the notches on all logs at the top of the logs. Turn the logs over and drive a 3-sided crosspiece through both sets of notches on the underside of the raft. Then complete the top set of notches and drive through the two additional sets of crosspieces.

You can lash together the overhanging ends of the two crosspieces at each end of the raft to give it added strength; however, when the crosspieces are immersed in water they swell and tightly bind the raft logs together.

If the crosspieces fit too loosely, wedge them with thin, boardlike pieces of wood split from a dead log. When the raft is in water, the wood swells, and the crosspieces become very tight and strong.

Make a deck of light poles on top of the raft to keep packs and other gear dry.

Figure 23-36. Raft Construction.

use caution when traveling through areas where warm and cold currents meet. It can be a storm forming area with dense fog and high winds and waves.

b. Winds. Winds also aid raft travel. In tropical areas, the winds are easterly blowing (trade winds). In higher latitudes, they blow from the west (westerlies). To use the winds as a mode of travel, the sea anchor should be pulled into the raft and, if available, a sail should be erected.

c. Waves. Waves can be both an asset and a hazard to raft travel. Waves are normally formed by the wind. The severity of the wind determines the size of the waves. On open seas, waves range from a few inches to over 100 feet in height. Under normal conditions, waves alone will move a liferaft only a few inches at a time; therefore, using waves as a mode of propulsion is not practical. Waves are a great help in finding land or shallow areas in the sea. Ocean waves always break when they enter shallow water or when they encounter an obstruction. The force of energy of the wave depends on how abruptly the water depth decreases and the size of the waves. Breaking waves can be used as an aid to

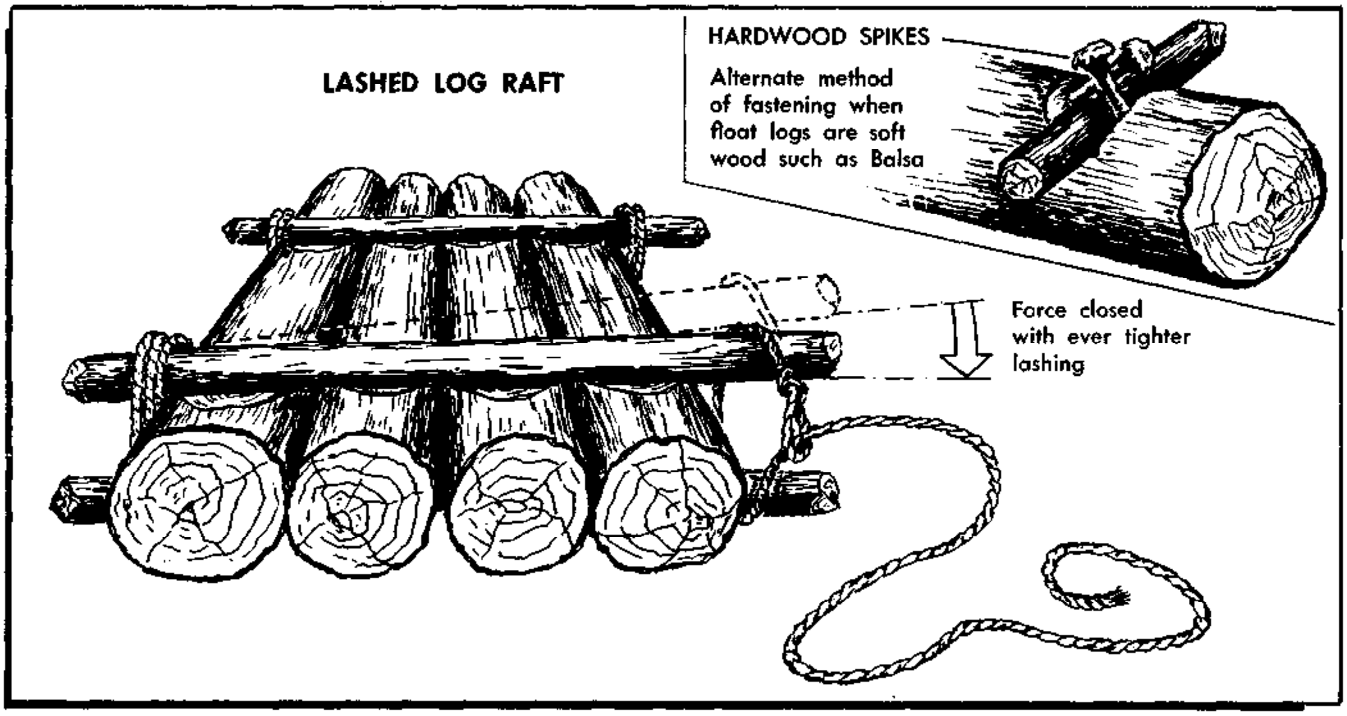
LASHED LOG RAFT
HARDWOOD SPIKES
Alternate method of fastening when float logs are soft wood such as Balsa
Force closed with ever tighter lashing

Figure 23-37. Lashed Log Raft.

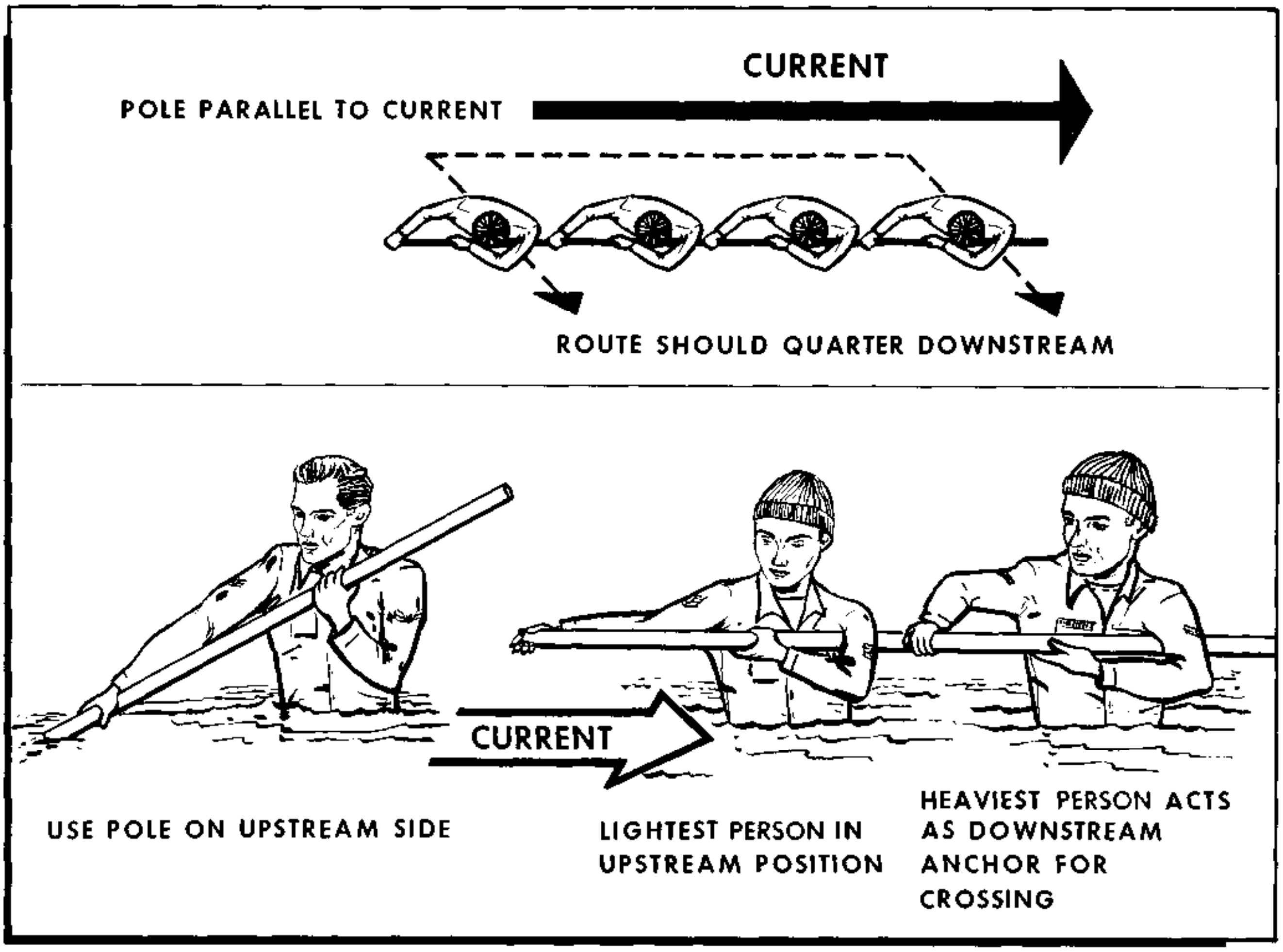
CURRENT
POLE PARALLEL TO CURRENT
ROUTE SHOULD QUARTER DOWNSTREAM
CURRENT
USE POLE ON UPSTREAM SIDE
LIGHTEST PERSON IN UPSTREAM POSITION
HEAVIEST PERSON ACTS AS DOWNSTREAM ANCHOR FOR CROSSING

Figure 23-38. Fording a Treacherous Stream.

make a landfall. Storms at sea are probably the greatest hazard to survivors in rafts. Aside from the waves created by a storm, the wind and rain can make life in a raft very difficult. The waves and wind can capsize a raft, or throw a person out of the raft, and then will constantly fill the raft with water. Seasickness can result from gentle to severe wave action. Additionally, rescue efforts may be severely hampered by large waves.

d. Tides. Tides are another form of wave but they are very predictable. They occur twice daily and usually cause no problems for anyone in a raft. Tides may range from one to 40 feet in height depending on the area of the world. They should be considered when planning a landing. When the tide is going in, it will help propel the raft to shore. The action of the water going away from shore when the tide is going out makes landing difficult.

e. Hazards. Certain marine life must be considered a hazard to survival. Sharks, jellyfish, eels, and most reef fishes can cause serious injuries if encountered. Waste materials should be disposed of when these creatures are not present.

(1) The survivor should be aware of the saltwater (estuarine) crocodile. It is found throughout the Southeast Asian shoreline. It is a well known maneater, and is almost always found in salt or brackish water. It is more commonly found near river mouths and along the coasts; however, it has been known to swim as much as 40 miles out into the sea. Females with nests are likely to be vicious and aggressive. They will grow to a length of 30 feet but most specimens are less than 15 feet in length. Survivors should watch for this reptile while landing their raft or fishing.

(2) The survivor will normally encounter reef fishes during the landing process by stepping on them. They may also be caught while fishing. Clothing and footgear must be worn at all times whether landing raft or fishing.

(3) Coral is normally found in warm waters, along the shores of islands and mainlands. There are many different types of coral. They should be avoided since all can destroy a raft or severely injure a survivor. It is best to stay in the raft when coral is encountered. If the survivor must wade to shore, footgear and pants should be worn for protection. Moving slowly and watching every step may prevent serious injuries. Coral does not exist where freshwater enters the sea.

(4) Ships can be a welcome sight to the survivor, but they can also be a hazard. Since the raft is a small object in a very large sea, the survivor must constantly be aware that at night or during inclement weather, the raft will be difficult to see and could be struck by a large ship.

f. Early Considerations. Survivors should stay upwind and clear of the aircraft (out of fuel covered waters) but in the vicinity of the crash until the aircraft sinks. A search for survivors is usually activated around the entire area of and near the crash site. Missing personnel may be unconscious and floating low in the water. Rescue procedures are illustrated in figure 23-39.

(1) The best technique for rescuing aircrew members from the water is to throw them a line with a life preserver attached. The second is to send a swimmer (rescuer) from the raft with a line using a flotation device which will support the weight of a rescuer. This will help to conserve energy while recovering the survivor. The least acceptable technique is to send an attached swimmer without floatable devices to retreive a survivor. In all cases, the rescuer should wear a life preserver. The strength of a person in a state of panic in the water should not be underestimated. A careful approach can prevent injury to the rescuer.

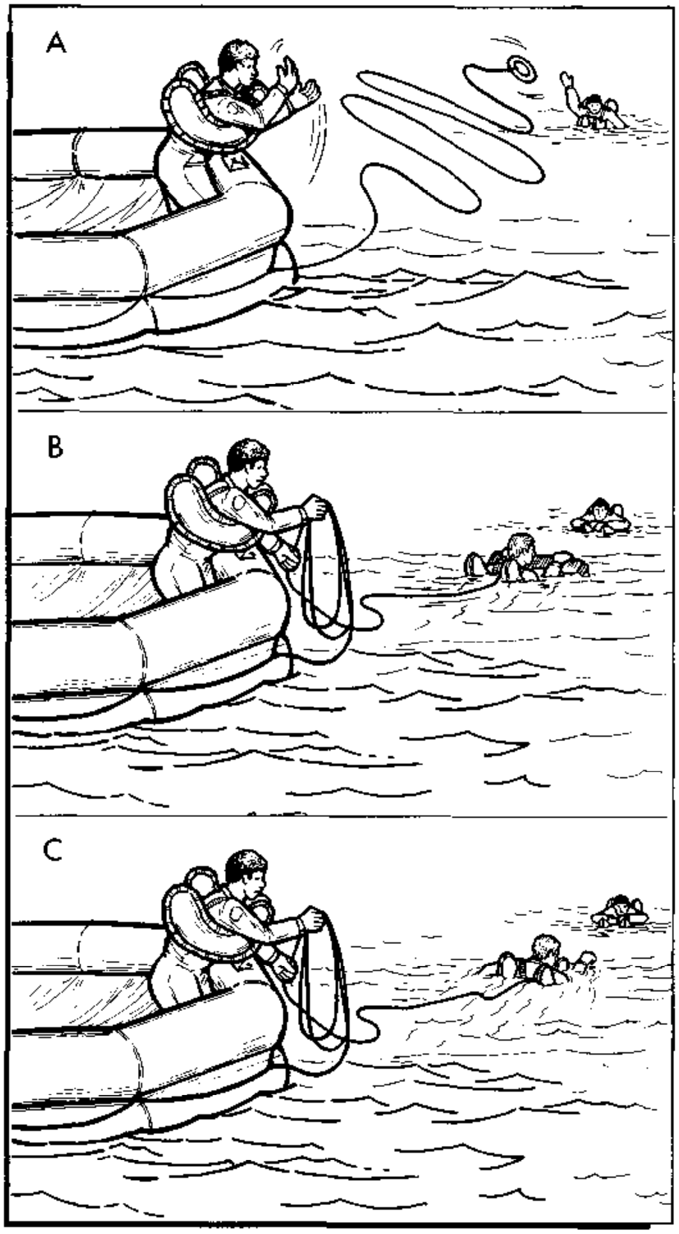

Figure 23-39. Rescue from Water.

(2) When the rescuer is approaching a survivor in trouble from behind, there is little danger of being kicked, scratched, or grabbed. The rescuer should swim to a point directly behind the survivor and grasp the backstrap of the life preserver. A sidestroke may then be used to drag the survivor to the raft.

(3) All debris from the aircraft should be inspected and salvaged (rations, canteens, thermos and other containers, parachutes, seat cushions, extra clothing, maps, etc). Secure equipment to the raft to prevent loss. Special precaution should be taken with flashlights and signaling equipment to keep them dry so they will function when needed.

(4) Rafts should be checked for inflation. Leaks, and points of possible chafing should be repaired, as required. All water should be removed from inside the raft. Care should be taken to avoid snagging the raft with shoes or sharp objects. Placing the sea anchor out will slow the rate of drift. If there is more than one raft, they should be connected with at least 25 feet of line. The lifeline attached to the outer periphery of the raft is to be used for the connection. Donning the antiexposure suit is essential in cold climates. Erecting wind breaks, spray shields, and canopies will protect survivors from the elements. Survivors should huddle together and exercise regularly to maintain body heat.

(5) Monitoring the physical condition of survivors and administering first aid to survivors is essential. If available, seasickness pills will help prevent vomiting and resulting dehydration.

(6) Survivors should prepare all available signaling equipment for immediate use.

(7) Compasses, watches, matches, and lighters will become worthless unless they are kept dry.

(8) The raft repair plugs should be attached to the raft for easy access as soon as possible.

(9) All areas of the body should be protected from the Sun. Precautions should be taken to prevent sunburn on the eyelids, under the chin, and on the backs of the ears. Sunburn cream and Chapstick will protect these areas.

(10) The leader should calmly analyze the situation and plan a course of action to include duty assignments (watch duty, procuring and rationing food and water, etc.). All survivors, except those who are badly injured or completely exhausted, are expected to perform watch duty, which should not exceed 2 hours. The survivor on watch should be looking for signs of land, passing vessels or aircraft, wreckage, seaweed, schools of fish, birds, and signs of chafing or leaking of the raft.

(11) Food and water can be conserved by saving energy. Survivors should remain calm.

(12) Maintaining a sense of humor will help keep morale high.

(13) The survivor(s) should remember that rescue at sea is a cooperative effort. Search aircraft contacts are limited by the visibility of survivors based on the availability of visual or electronic signaling devices. Visual and electronic communications can be increased by using all available signaling devices (signal mirrors, radios, signal panels, dye marker, and other available devices) when an aircraft is in the area.

(14) A log should be maintained with a record of the navigator's last fix, time of ditching, names and physical condition of survivors, ration schedule, winds, weather, direction of swells, times of sunrise and sunset, and other navigation data.

23-9. Physical Considerations:

a. The greatest problem a survivor is faced with when submerged in cold water is death due to hypothermia. When a survivor is immersed in cold water, hypothermia occurs rapidly due to the decreased insulating quality of wet clothing and the fact that water displaces the layer of still air which normally surrounds the body. Water causes a rate of heat exchange approximately 25 times greater than air at the same temperature. The following lists life expectancy times:

Temperature of Water	*Time*
70° - 60°	12 hours
60° - 50°	6 hours
50° - 40°	1 hour
40° - below	-1 hour

NOTE: These times may be increased with the wearing of an antiexposure suit.

b. The best protection for a survivor against the effects of cold water is to get into the liferaft, stay dry, and insulate the body from the cold surface of the bottom of the liferaft. If this is not possible, wear of the antiexposure suits will extend a survivor's life expectancy considerably. It's important to keep the head and neck out of the water and well insulated from the cold water effects when the temperature is below 66°F. The wearing of life preservers increases the predicted survival time just as the body position in the water increases the probability of survival. The following table shows predicted survival times for an average person in 50°F water:

Situation	*Predicted Survival Time (Hours)*
No Flotation	
Drownproofing	1.5
Treading Water	2.0
With Flotation	
Swimming	2.0
Holding Still	2.7
Help	4.0
Huddle	4.0

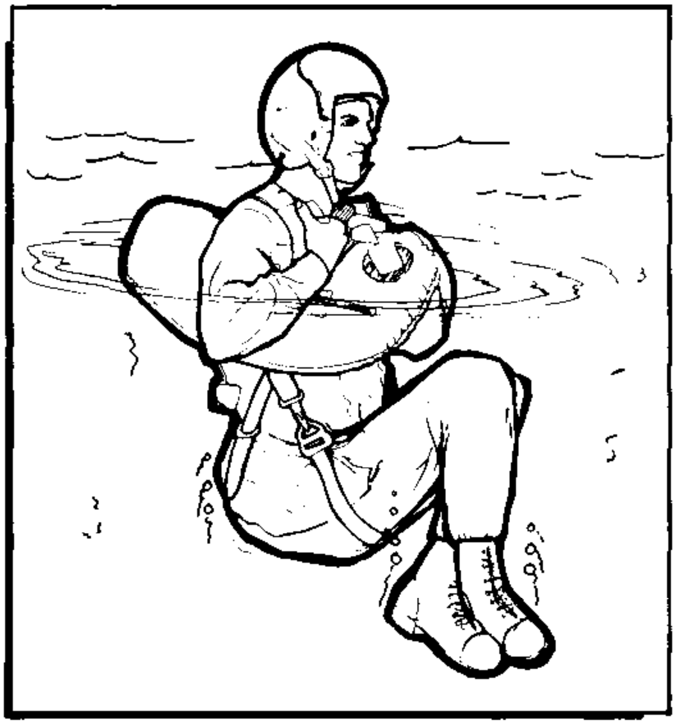

Figure 23-40. HELP Position.

(1) Help Body Position. Remaining still and assuming the fetal position, or heat escape lessening posture (HELP) (figure 23-40), will increase the downed crewmember's survival time. About 50 percent of the heat is lost from the head. It is therefore important to keep the head out of the water. Other areas of high heat loss are the neck, the sides, and the groin.

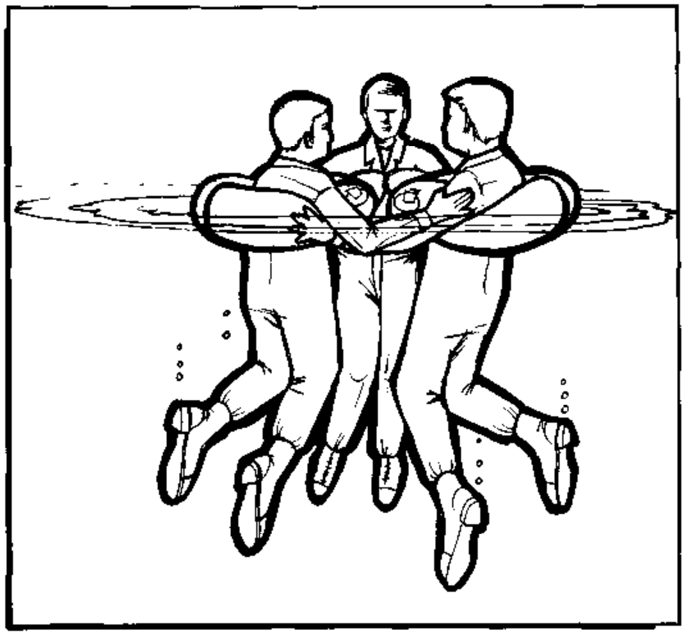

Figure 23-41. Huddling for Temperature Conservation.

(2) Huddling. If there are several survivors in the water, huddling close, side to side in a circle, body heat will be preserved (figure 23-41).

23-10. Life Preserver Use:

a. Survival Swimming Without a Life Preserver. A survivor who knows how to relax in the water is in little danger of drowning, especially in saltwater where the body is of lower density than the water. Trapped air in clothing will help buoy the survivor in the water. If in the water for long periods, the survivor will have to rest from treading water. The survivor may best do this by floating on the back. If this is not possible, the following technique should be used: Rest erect in the water and inhale; put the head face-down in the water and stroke with the arms; rest in this facedown position until there is a need to breathe again; raise the head and exhale; support the body by kicking arms and legs and inhaling; then repeat the cycle.

b. Swimming With a Life Preserver. The bulkiness of clothing, equipment, and (or) any personal injuries will necessitate the immediate need for flotation. Normally, a life preserver will be available for donning before entering the water.

(1) Proper inflation of the life preserver must be done after clearing the aircraft but, preferably, before entering the water. Upon entering the water, the two cells of the life preserver should be fastened together. Limited swimming may be done with the life preserver inflated by cupping the hands and taking strong strokes deep into the water. The life preserver may be slightly deflated to permit better arm movement.

(2) The backstroke should be used to conserve energy when traveling long distances. If aiding an injured or unconscious person, the sidestroke may have to be used. When approaching an object, it is best to use the breaststroke. If a group must swim, they should try to have the strongest swimmer in the lead with any injured persons intermingled within the group. It is best to swim in a single file.

23-11. Raft Procedures. There are three needs which can be satisfied by most of the rafts; they are personal protection, mode of travel, and evasion and camouflage.

a. One-Man Raft:

(1) The one-man raft has a main cell inflation. If the CO_2 bottle should malfunction and not inflate the raft or if the raft develops a leak, it can be inflated orally. The spray shield acts as a shelter from the cold, wind, and water. In some cases, this shield serves as insulation. The insulated bottom plays a significant role in the survivor's protection from hypothermia by limiting the conduction of the cold through the bottom of the raft (figure 23-42).

(2) Travel is more effectively made by inflating or deflating the raft to take advantage of the wind or cur-

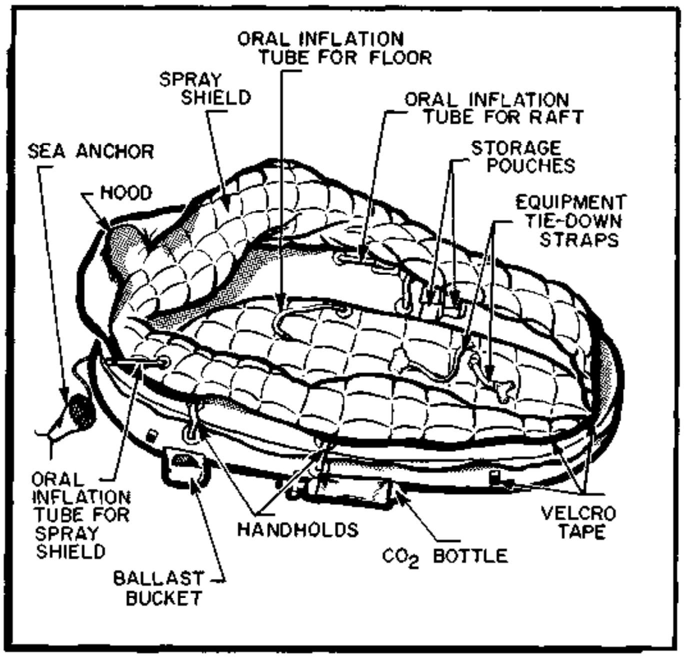
ORAL INFLATION TUBE FOR FLOOR
SPRAY SHIELD
ORAL INFLATION TUBE FOR RAFT
SEA ANCHOR
STORAGE POUCHES
HOOD
EQUIPMENT TIE-DOWN STRAPS
ORAL INFLATION TUBE FOR SPRAY SHIELD
HANDHOLDS
VELCRO TAPE
CO_2 BOTTLE
BALLAST BUCKET

Figure 23-42. One-Man Raft with Spray Shield.

rent. The spray shield can be used as a sail while the ballast buckets serve to increase raft drag in the water. The last device which may be used to control the speed and direction of the raft is the sea anchor. (NOTE: The primary purpose of the sea anchor is to stabilize the raft.)

(3) Black rafts have been developed for use in tactical areas. These rafts blend with the background of the sea. The raft can be further modified for evasion by partially deflating which provides a low profile.

(4) The one-man raft is connected to the aircrew member by a lanyard parachuting to the water. Survivors should not swim to the raft, but pull it to their position via the lanyard. The parachute J-1 releases should be closed and the life preserver separated before boarding the raft (figure 23-43). The raft may hit the water upside down, but may be righted by approaching the bottle side and flipping it over. The spray shield must be in the raft to expose the boarding handles.

(5) If the survivor has an arm injury, boarding is best done by turning the back to the small end of the raft, pushing the raft under the buttocks and lying back (figure 23-44). Another method of boarding is to push down on the small end until one knee is inside and lie forward (figure 23-44).

(6) In rough seas, it may be easier for the survivor to grasp the small end of the raft, and, in a prone position, kick and pull into the raft. Once in the raft, lying face down, the sea anchor should be deployed and adjusted. To sit upright in the raft, one side of the seat kit might have to be disconnected and the survivor should roll to that side. The spray shield is then adjusted. There are two variations of the one-man raft, with the improved model incorporating an inflatable spray shield and floor for additional insulation. The spray shield is designed to help keep the survivor dry and warm in cold oceans, and protect them from the Sun in the hot climates (figure 23-45).

(7) The sea anchor can be adjusted to either act as a drag by slowing down the rate of travel with the current or as a means of traveling with the current. This is done by opening or closing the apex of the sea anchor. When opened, the sea anchor (figure 23-46) will act as a drag and the survivor will stay in the general area. When the sea anchor is closed (figure 23-46), it will form a pocket for the current to strike and propel the raft in the direction of the current. Additionally, the sea anchor should be adjusted so that when the raft is on the crest of a wave, the sea anchor is in the trough of the wave (figure 23-47).

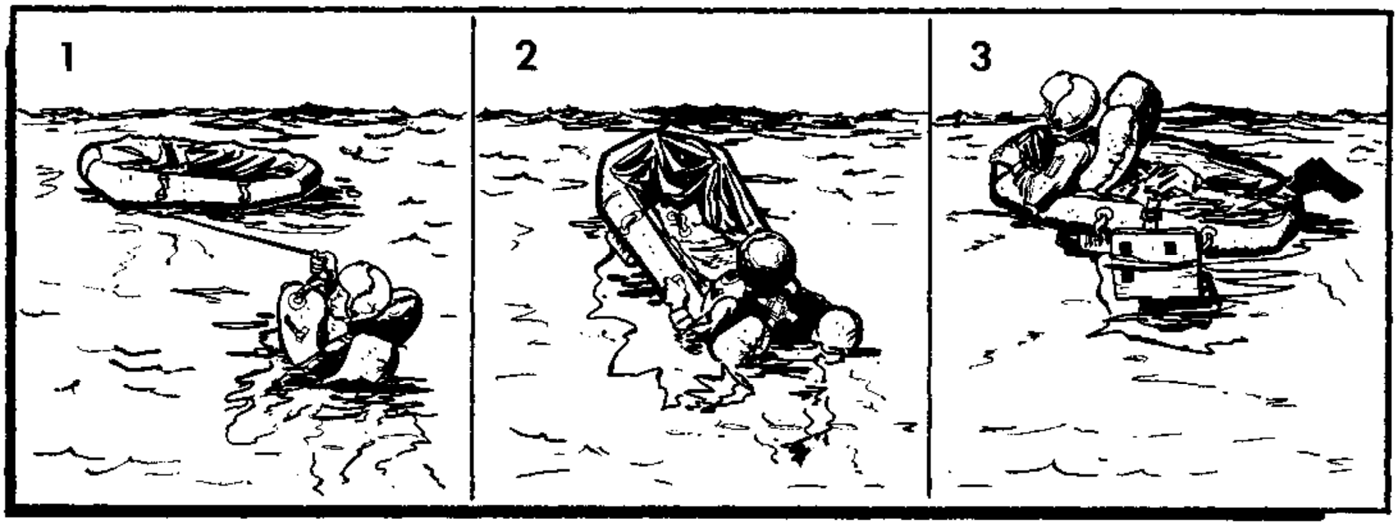

Figure 23-43. Boarding One-Man Raft.

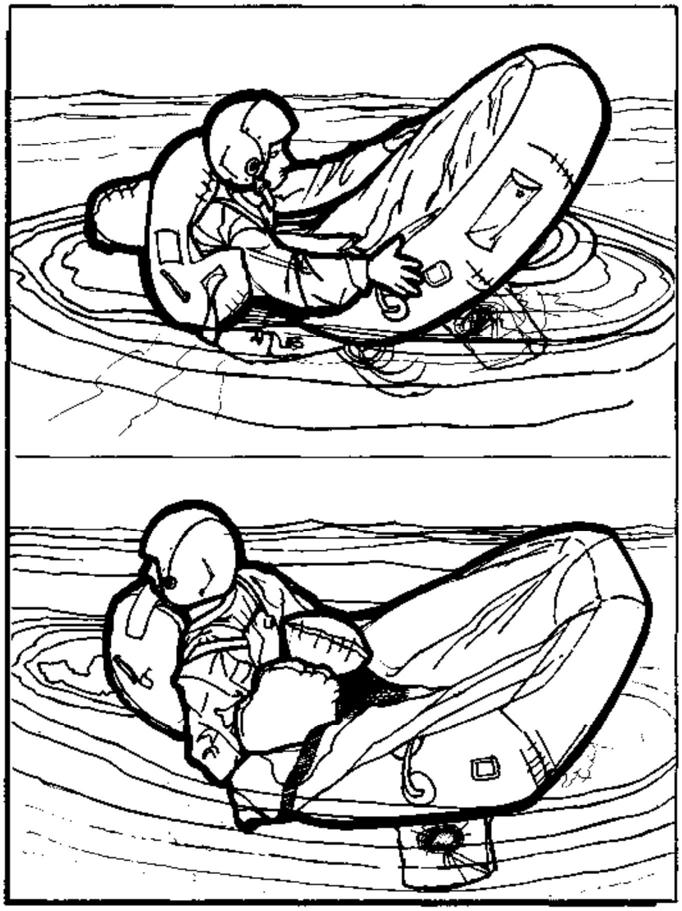

Figure 23-44. Boarding One-Man Raft (Other Methods).

b. Seven-Man Raft:

(1) The seven-man raft is found on some multiplace aircraft. It can also be found in the survival drop kit (MA-1 Kit, figure 23-48). This type of raft may inflate upside down and may, therefore, require the survivor to right the raft before boarding. The J-1 releases should be closed before boarding. The survivor should always work from the bottle side to prevent injury if the raft turns over. Facing into the wind provides additional assistance in righting the raft. The handles on the inside bottom of the raft are used for boarding (figure 23-49).

(2) The boarding ladder is used to board if someone assists in holding down the opposite side. If no assistance is available, the survivor should again work from the bottle side with the wind at the back to help hold down the raft. The survivor should separate the life preserver, grasp an oarlock and boarding handle, kick the legs to get the body prone on the water, and then kick and pull into the raft. If the survivor is weak or injured, the raft may be partially deflated to make boarding easier (figure 23-50).

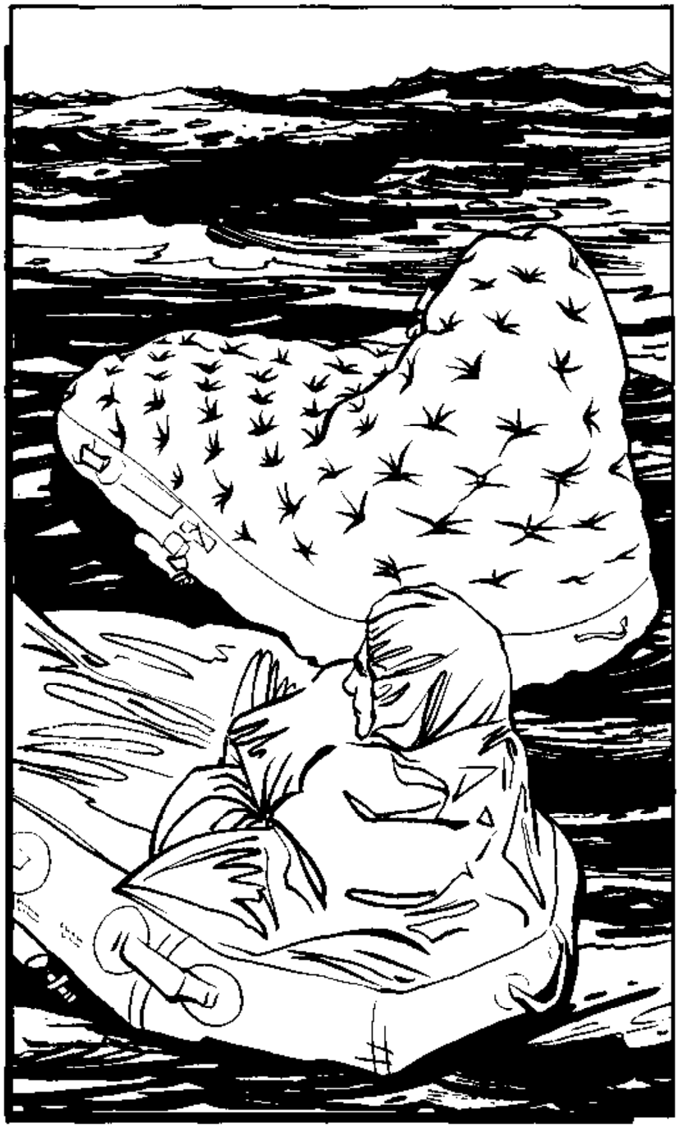

Figure 23-45. One-Man Raft with Spray Shield Inflated.

(3) Manual inflation can be done by using the pump to keep buoyancy chambers and cross-seat firm, but the raft should not be over inflated. The buoyancy chambers and cross-seat should be rounded but not drum tight. Hot air expands, so on hot days, some air may be released while air may be added on cold days (figure 23-51).

c. Sailing a Raft into the Wind. Rafts are not equipped with keels, so they cannot be sailed into the wind. However, anyone can sail a raft downwind, and multiplace (except 20/25-man) rafts can be successfully sailed 10 degrees off from the direction of the wind. An attempt to sail the raft should not be made unless land is near. If the decision to sail is made and the wind is blowing toward a desired destination, survivors should fully inflate the raft, sit high, take in the sea anchor, rig a sail, and use an oar as a rudder as shown in figure 23-52.

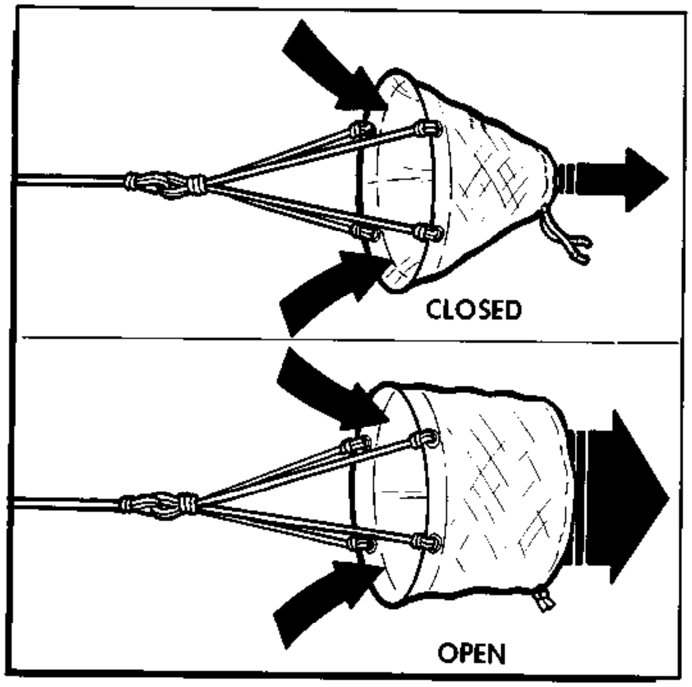
CLOSED
OPEN

Figure 23-46. Sea Anchor.

d. Multiplace Raft. In a multiplace (except 20/25-man) raft, a square sail should be erected in the bow using oars with their extensions as the mast and crossbar (figure 23-52). A waterproof tarpaulin or parachute material may be used for the sail. If the raft has no regular mast socket and step, the mast may be erected by tying it securely to the front cross-seat using braces. The bottom of the mast must be padded to prevent it from chafing or punching a hole through the floor whether or not a socket is provided. The heel of a shoe, with the toe wedged under the seat, makes a good improvised mast step. The corners of the lower edge of the sail should not be secured. The lines attached to the corners are held with the hands so that a gust of wind will not rip the sail, break the mast, or capsize the raft. Every precaution must be taken to prevent the raft from

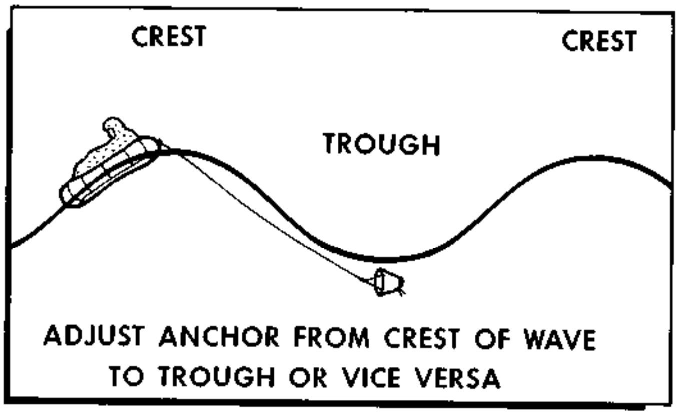
CREST
CREST
TROUGH
ADJUST ANCHOR FROM CREST OF WAVE
TO TROUGH OR VICE VERSA

Figure 23-47. Deployment of the Sea Anchor.

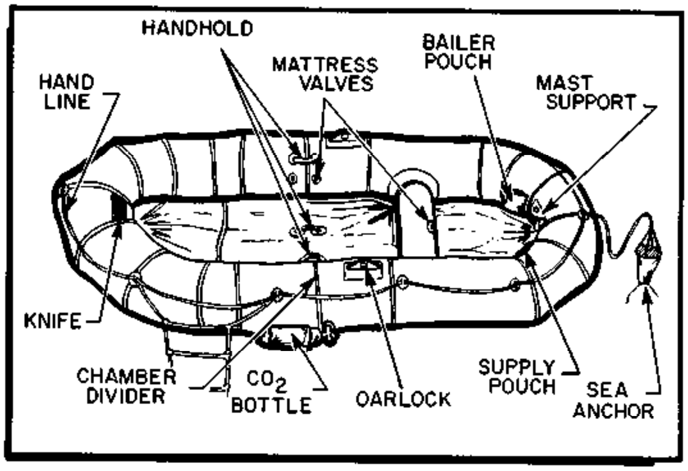
HANDHOLD
HAND LINE
MATTRESS VALVES
BAILER POUCH
MAST SUPPORT
KNIFE
CHAMBER DIVIDER
CO_2 BOTTLE
OARLOCK
SUPPLY POUCH
SEA ANCHOR

Figure 23-48. Seven-Man Raft.

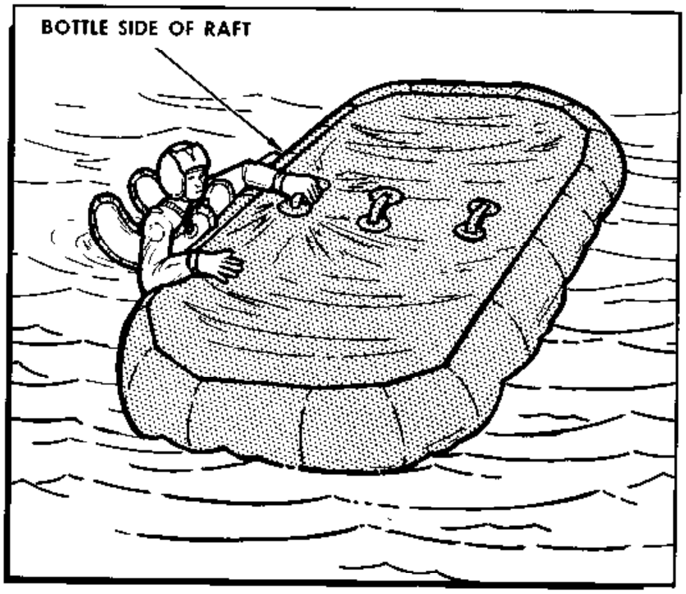
BOTTLE SIDE OF RAFT

Figure 23-49. Method of Righting Raft.

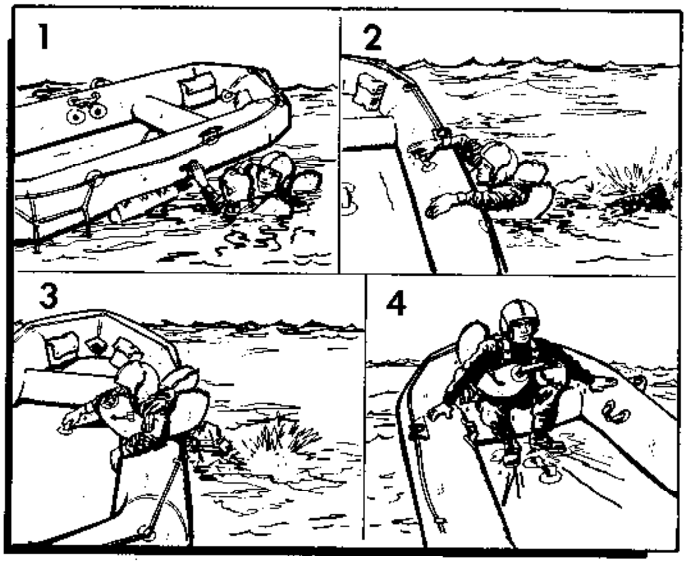

Figure 23-50. Method of Boarding Seven-Man Raft.

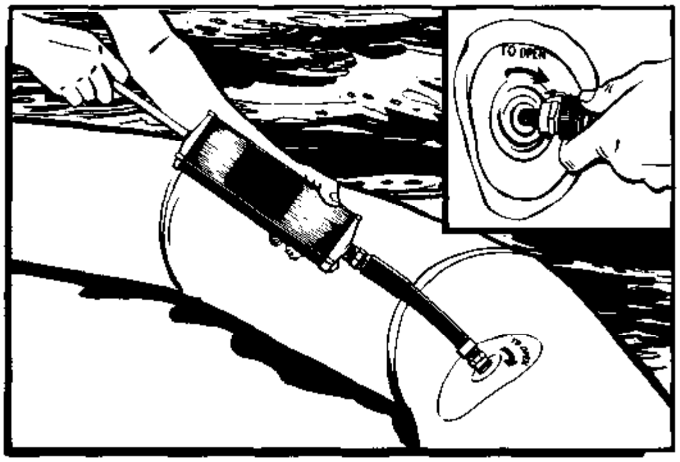

Figure 23-51. Inflating the Raft.

turning over. In rough weather, the sea anchor is kept out away from the bow. The passengers should sit low in the raft with their weight distributed to hold the upwind side down. They should also avoid sitting on the sides of the raft or standing up to prevent falling out. Sudden movements (without warning the other passengers) should be avoided. When the sea anchor is not in use, it should be tied to the raft and stowed in such a manner that it will hold immediately if the raft capsizes.

e. Twenty to Twenty Five-Man Rafts. The 20/25-man rafts may be found in multiplace aircraft (figures 23-53 and 23-54). They will be found in accessible areas of the fuselage or in raft compartments. Some may be automatically deployed from the cockpit, while others may need manual deployment. No matter how the raft lands in the water, it's ready for boarding. The accessory kit is attached by a lanyard and is retrieved by hand. The center chamber must be inflated manually with the hand pump. The 20/25-man raft should be boarded from the aircraft if possible; if not, the following steps should be taken:

(1) Approach lower boarding ramp.

(2) Separate the life preserver.

(3) Grasp the boarding handles and kick the legs to get the body into a prone position on the water's surface; then kick and pull until inside the raft.

(4) If for any reason the raft is not completely inflated, boarding will be made easier by approaching the intersection of the raft and ramp, grasping the upper boarding handle, and swinging one leg onto the center of the ramp as in mounting a horse (figure 23-55).

(5) The equalizer tube should be clamped immediately upon entering the raft to prevent deflating the entire raft, in case of puncture (figure 23-56).

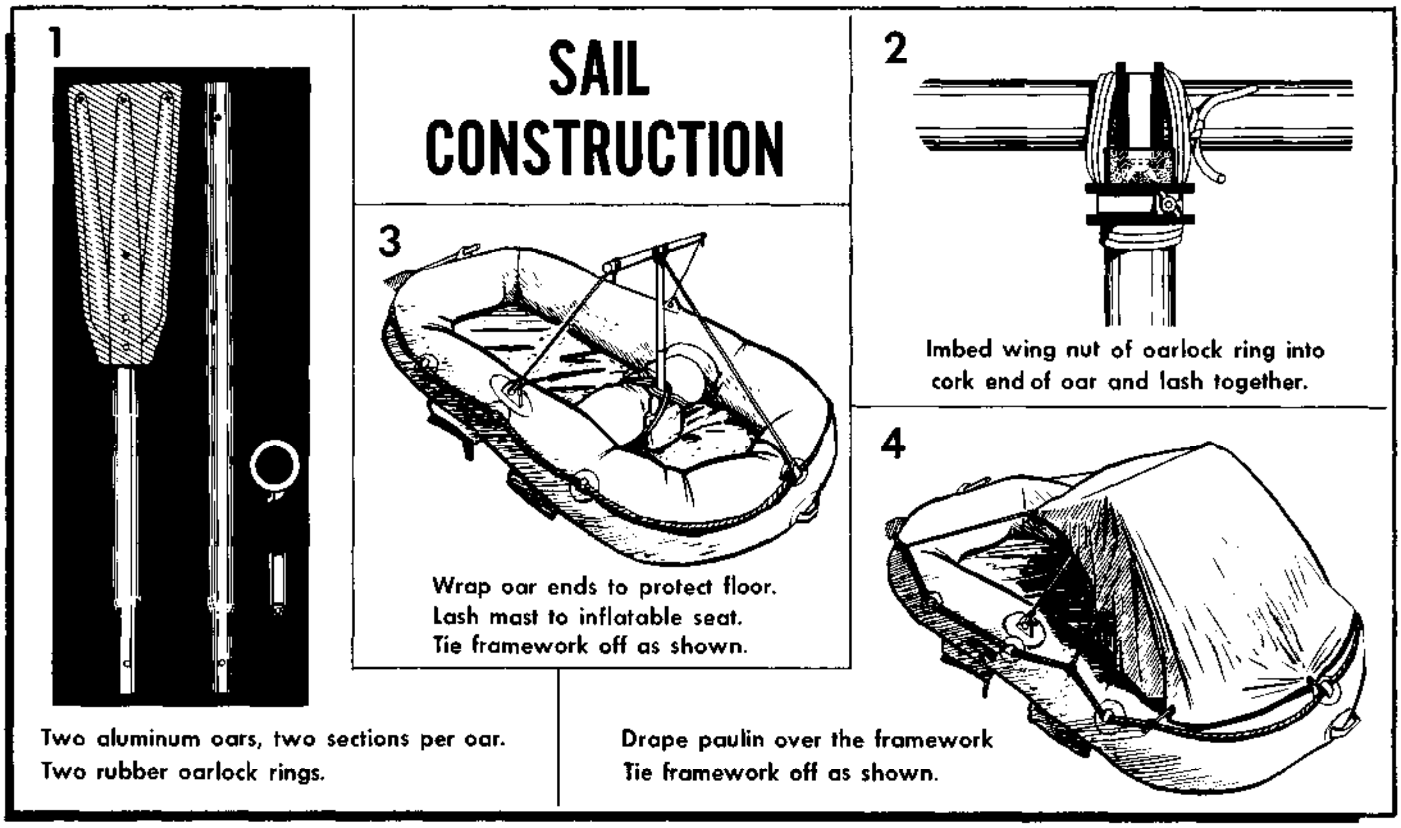
SAIL CONSTRUCTION
1
Two aluminum oars, two sections per oar.
Two rubber oarlock rings.
2
Imbed wing nut of oarlock ring into cork end of oar and lash together.
3
Wrap oar ends to protect floor.
Lash mast to inflatable seat.
Tie framework off as shown.
4
Drape paulin over the framework
Tie framework off as shown.

Figure 23-52. Sail Construction.

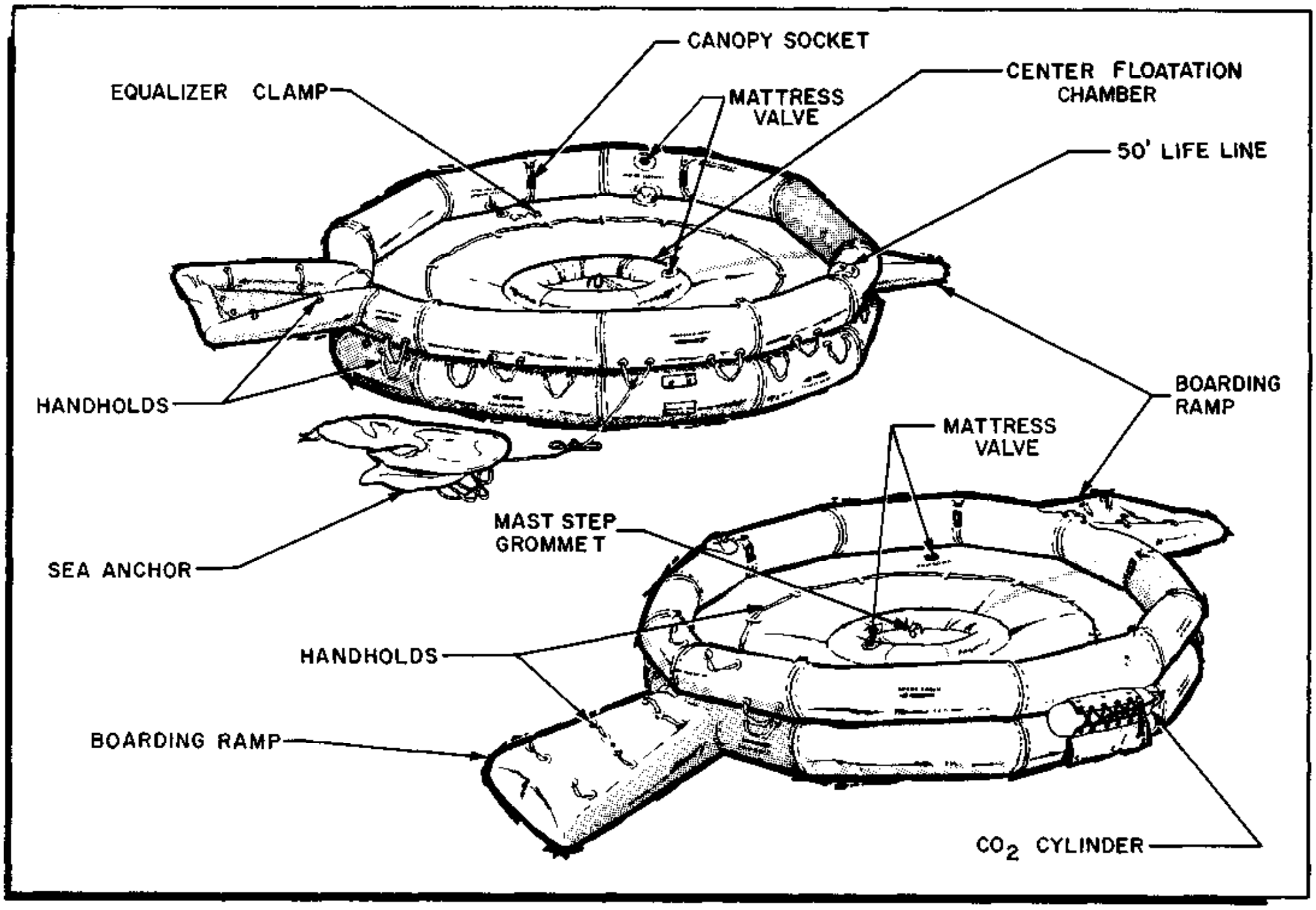

Figure 23-53. 20-Man Raft.

(6) The 20/25-man raft can be inflated by using the pump to keep the chambers and center ring firm. They should be well rounded but not drum tight (figure 23-57).

23-12. Making a Landfall. The lookout should watch carefully for signs of land. Some indications of land are:

a. A fixed cumulus cloud in a clear sky, or in a sky where all other clouds are moving, often hovers over or slightly downwind from an island.

b. In the tropics, a greenish tint in the sky is often caused by the reflection of sunlight from the shallow lagoons or shelves of coral reefs.

c. In the arctic, ice fields or snow-covered land are often indicated by light-colored reflections on clouds, quite different from the darkish gray reflection caused by open water.

d. Deep water is dark green or dark blue. Lighter color indicates shallow water, which may mean land is near.

e. In fog, mist, rain, or at night, when drifting past a nearby shore, land may be detected by characteristic odors and sounds. The musty odor of mangrove swamps and mudflats and the smell of burning wood carries a long way. The roar of surf is heard long before the surf is seen. Continued cries of sea birds from one direction indicate their roosting place on nearby land.

f. Birds are usually more abundant near land than over the open sea. The direction from which flocks fly at dawn and to which they fly at dusk may indicate the direction of land. During the day, birds are searching for food and the direction of flight has no significance unless there is a storm approaching.

g. Land may be detected by the pattern of the waves, which are refracted as they approach land. Figure 23-58 shows the form the waves assume. Land should be located by observing this pattern and turning parallel to the slightly turbulent area marked "X" on the illustration and following its direction.

23-13. Methods of Getting Ashore:

a. Swimming Ashore. This is a most difficult decision. It depends on many things. Some good swimmers

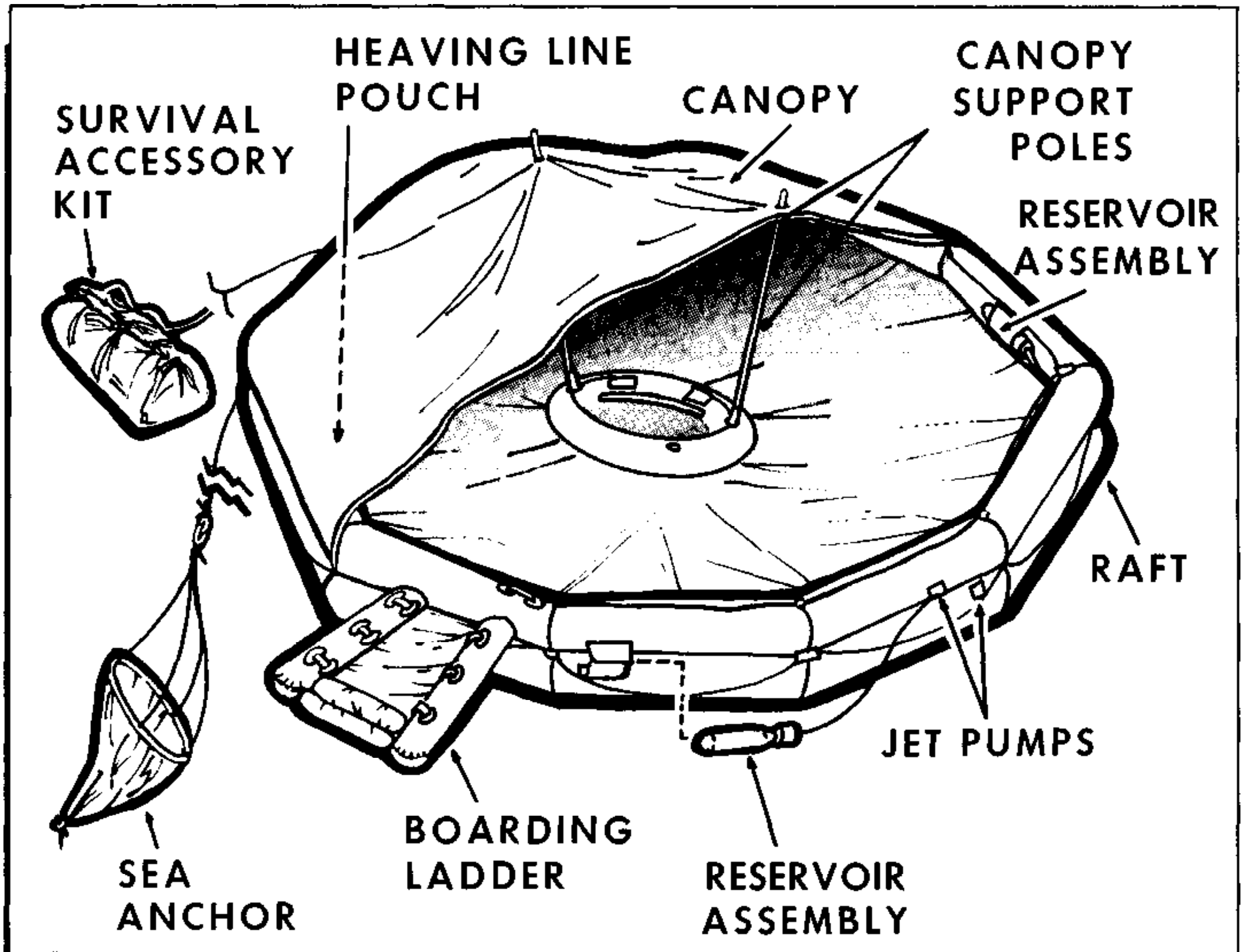

- CHECK RAFT FOR INFLATION AND LEAKS
- BAIL OUT THE RAFT
- INVENTORY AND STOW ALL GEAR
- DEPLOY AND ADJUST SEA ANCHOR
- RIG SPRAY SHIELD FOR PROTECTION FROM THE ELEMENTS
- KEEP CALM, ESTIMATE YOUR SITUATION
- PREPARE FOR PICKUP

Figure 23-54. 25-Man Raft.

have been able to swim eight-tenths of a mile in 50°F water before being overcome by hypothermia. Others have not been able to swim 100 yards. Furthermore, distances on the water are very deceptive. In most instances, staying with the raft is the best course of action. If the decision is made to swim, a life preserver or other flotation aid should be used. Shoes and at least one thickness of clothing should be worn. The side or breast stroke will help conserve strength.

(1) If surf is moderate the survivor can ride in on the back of a small wave by swimming forward with it and making a shallow dive to end the ride just before the wave breaks. The swimmer should stay in the trough between waves in high surf, facing the seaward wave and submerging when the wave approaches. After the wave passes, the swimmer should work shoreward in the next trough.

(2) If the swimmer is caught in the undertow of a large wave, push off the bottom and swim to the surface and proceed shoreward. A place where the waves rush up onto the rocks should be selected if it is necessary to land on rocky shores and avoid places where the waves explode with a high white spray. After selecting the landing point, the swimmer should advance behind a

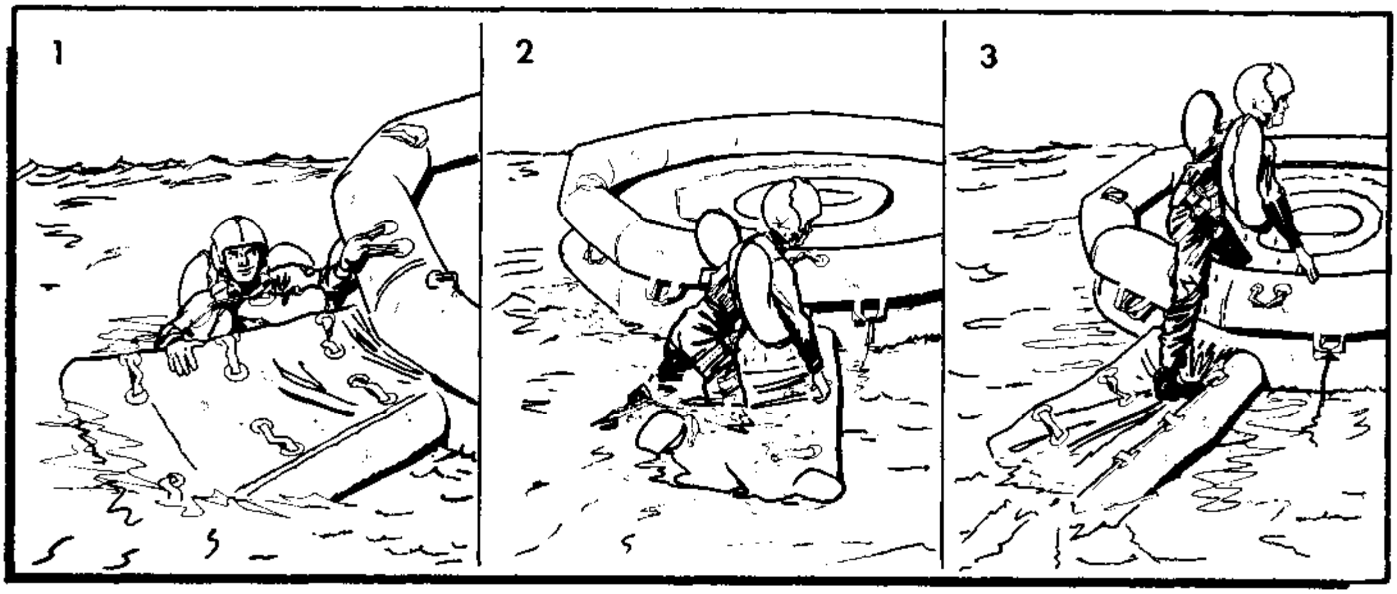

Figure 23-55. Boarding 20-Man Raft.

large wave into the breakers. The swimmer should face shoreward and take a sitting position with the feet in front, 2 or 3 feet lower than the head, so the knees are bent and the feet will absorb shocks when landing or striking submerged boulders or reefs. If the shore is not reached the first time, the survivor should swim with hands and arms only. As the next wave approaches, the sitting position with the feet forward should be repeated until a landing is made.

(3) Water is quieter in the lee of a heavy growth of seaweed. This growth can be very helpful. The swimmer should crawl over the top by grasping the vegetation with overhand movements.

(4) A rocky reef should be crossed in the same way as landing on a rocky shore. The feet should be close together with knees slightly bent in a relaxed sitting posture to cushion blows against coral.

b. Rafting Ashore. In most cases, the one-man raft can be used to make a shore landing with no danger. Going ashore in a strong surf is dangerous. The time should be taken to sail around and look for a sloping beach where the surf is gentle. The landing point should be carefully selected. Landing when the Sun is low and straight in front is not recommended. The survivor

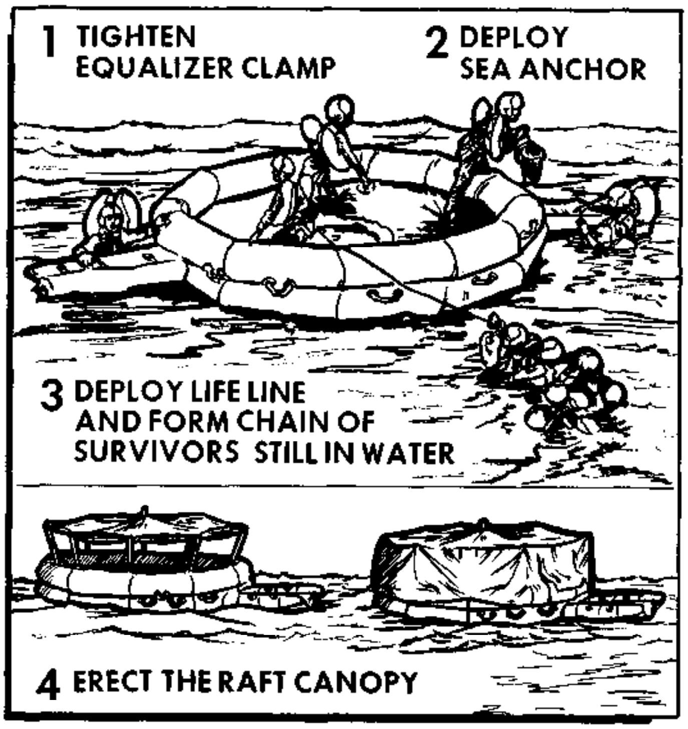
1 TIGHTEN EQUALIZER CLAMP
2 DEPLOY SEA ANCHOR
3 DEPLOY LIFE LINE AND FORM CHAIN OF SURVIVORS STILL IN WATER
4 ERECT THE RAFT CANOPY

Figure 23-56. Immediate Action—Multiplace Raft.

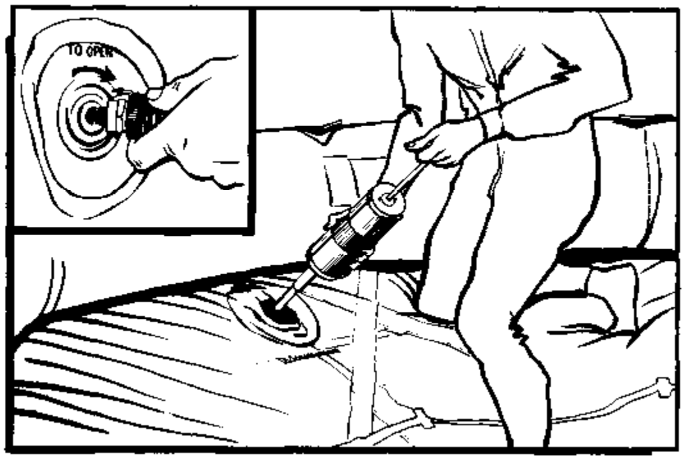

Figure 23-57. Inflating the 20-Man Raft.

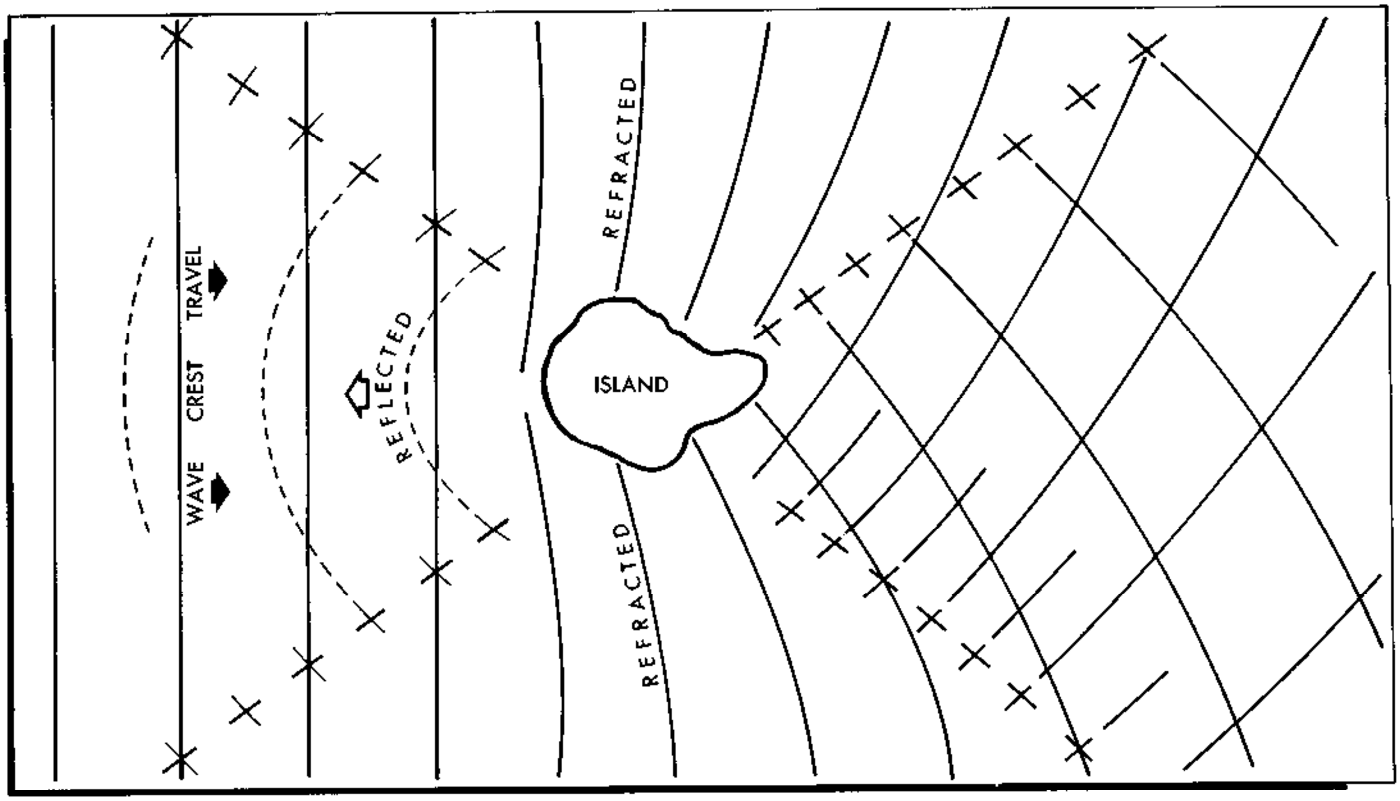

Figure 23-58. Diagram of Wave Patterns About an Island.

should look for gaps in the surf line and head for them while avoiding coral reefs and rocky cliffs. These reefs don't occur near the mouths of freshwater streams. Avoid rip currents or strong tidal currents which may carry the raft far out to sea.

(1) When going through surf, the survivor should:

(a) Take down the mast.
(b) Don clothing and shoes to avoid injuries.
(c) Adjust and fasten life preserver.
(d) Stow equipment.
(e) Use paddles to maintain control.

(f) Ensure the sea anchor is deployed to help prevent the sea from throwing the stern of the raft around and capsizing it. CAUTION: The sea anchor should not be deployed when traveling through coral.

(2) In medium surf with no wind, survivors should keep the raft from passing over a wave so rapidly that it drops suddenly after topping the crest. If the raft turns over in the surf, every effort should be made to grab hold.

(3) The survivor should ride the crest of a large wave as the raft nears the beach, staying inside until it has grounded. If there is a choice, a night landing should not be attempted. If signs of people are noted, it might be advantageous to wait for assistance.

(4) Sea-ice landings should be made on large stable floes only. Icebergs, small floes, and disintegrating floes could cause serious problems. The edge of the ice can cut, and the raft deflated. Use paddles and hands to keep the raft away from the sharp edges of the iceberg. The raft should be stored a considerable distance from the ice edge. It should be fully inflated and ready for use in case the floe breaks up.

Part Eight

SIGNALING AND RECOVERY

Chapter 24

SIGNALING

24-1. Introduction:

a. Most successful recoveries have resulted primarily because survivors were able to *assist* in their own recovery. Many rescue efforts failed because survivors lacked the knowledge and ability necessary to assist. When needed, this knowledge and ability could have made the difference between life or death—freedom or captivity (figure 24-1).

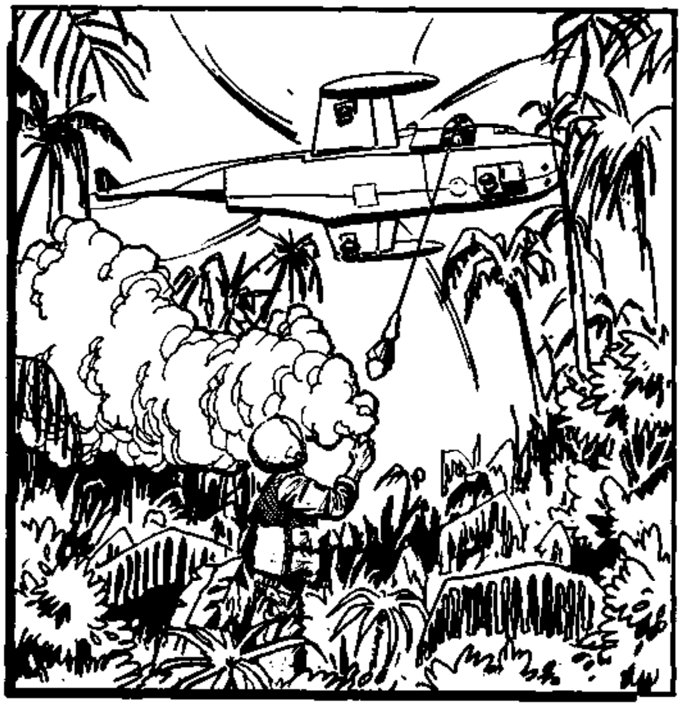

Figure 24-1. Signaling and Recovery.

b. What can survivors do to assist in their own recovery? First, they need to know what is being done to find them. Next, they need to know how to operate the communications equipment in the survival kit and when to put each item into use. Survivors should also be able to improvise signals to improve their chances of being sighted and to supplement the issued equipment.

c. It is not easy to spot one survivor, a group of survivors, or even an aircraft from the air, especially when visibility is limited. Emergency signaling equipment is designed to make a person easier to find. Emergency equipment may be used to provide rescue personnel with information about survivors' conditions, plans, position, or the availability of a rescue site where recovery vehicles might reach them (figure 24-2).

d. A part of a survivor's plan of action should be to visualize how emergencies will develop, recognize them, and, at the appropriate time, let friendly forces know about the problem. The length of time before survivors are rescued often depends on the effectiveness of emergency signals and the speed with which they can be used. Signal sites should be carefully selected. These sites should enhance the signal and have natural or manufactured materials readily available for immediate use. Survivors should avoid using pyrotechnic signals wastefully as they may be needed to enhance rescue efforts. Signals used correctly can hasten recovery and eliminate the possibility of a long, hard survival episode. Survivors should:

(1) Know how to use their emergency signals.

(2) Know when to use their signal.

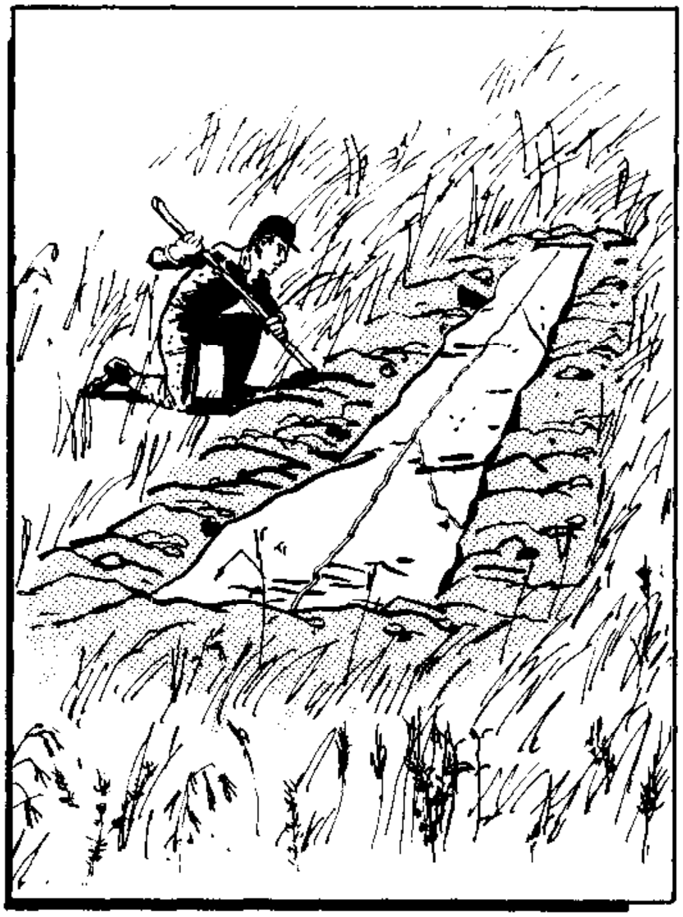

Figure 24-2. Signaling.

(3) Be able to use their signals on short notice.

(4) Use signals in a manner which will not jeopardize individual safety.

e. The situation on the ground governs the type of information which survivors can furnish the rescue team, and will govern the type of signaling they should use. In nontactical survival situations, there are no limitations on the ways and means survivors may use to furnish information.

f. In hostile areas, limitations on the use of signals should be expected. The use of some signaling devices will pinpoint survivors to the enemy as well as to friendly personnel. Remember the signal enhances the visibility of the survivors.

24-2. Manufactured Signals:

a. Electronic Signals:

(1) Current line-of-sight electronic signaling devices fall into two categories. One is the transceiver type; the other is the personal locator beacon type. The transceiver type is equipped for transmitting tone or voice and receiving tone or voice. The personal locator beacon is equipped to transmit tone only. The ranges of the different radios vary depending on the altitude of the receiving aircraft, terrain factors, forest density, weather, battery strength, type(s) of radios and interference. Interference is a very important aspect of the use of these radios. If a personal locator beacon is transmitting, it will interfere with incoming and outgoing signals of the transceivers.

(2) Before using survival radios, a few basic precautions should be observed. These will help in obtaining maximum performance from the radios in survival situations.

(a) The survival radios are line-of-sight communication devices; therefore, the best transmission range will be obtained when operating in clear, unobstructed terrain.

(b) Extending from the top and bottom of the radio antenna is an area referred to as the "cone of silence." To avoid the "cone of silence" problem, keep the radio/beacon antenna orthogonal to (at a right angle to) the path of the rescue aircraft.

(c) Since the radios have the capability of transmitting a tone (beacon) without being hand-held, they can be placed upright on a flat elevated surface allowing the operator to perform other tasks.

(d) Never allow the radio antenna to ground itself on clothing, body, foliage or the ground. This will severely decrease the effective range of the signals.

(e) Conserve battery power by turning the radio off when not in use. Do not transmit or receive constantly. Use the locator beacon to supplement the radio when transmitting is done. In tactical environments, the radio should be used as stated in the premission briefing.

(f) Survival radios are designed to operate in extreme heat or cold. The life expectancy of a battery decreases as the temperature drops below freezing and exposure to extreme heat or shorting out of the battery can cause an explosion. During cold weather, the battery should be kept warm by placing it between the layers of clothing to absorb body heat, or wrapped in some type of protective material when it is not being used.

(g) Survival radios are designed to be waterproof. However, precautions should be taken to keep them out of water.

(3) Presently, a satellite monitoring system has been developed to assist in locating survivors. To activate this system (SARSAT), the transmitter is "keyed" for a minimum of 30 seconds. In a nontactical situation, leave the beacon on until rescue is heard or sighted.

b. Pyrotechnics. Care should be used when operating around flammable materials.

(1) A device containing chemicals for producing smoke or light is known as a pyrotechnic. Hand-held flares are in this category. Survivors may be required to use a variety of flares. They must know the types of flares stored in their survival kits and (or) aircraft. Aircrew members should learn how to use each type of flare before they face an emergency. Flares are designed to be used during the day or night. Day flares produce a unique bright-colored smoke which stands out very clearly against most backgrounds. Night flares are extremely bright and may be seen for miles by air, ground, or naval recovery forces.

(2) The hand-held launched flares also fall in the pyrotechnic category. They were designed to overcome the problems of terrain masking and climatic conditions. For example, a person may be faced with multilayer vegetation or atmospheric conditions known as an inversion which keeps the smoke next to the ground.

(3) Flares must be fired at the right time to be of maximum use. Smoke flares, for example, take a second or two after activation before they produce a full volume of smoke. Therefore, the flare should be ignited just before the time it can be seen by rescue personnel. These signals should not be used in tactical environments unless directed to do so.

(4) Tracer ammunition is another pyrotechnic which may be issued to aircrew members. When fired, the projectile appears as an orange-red flash the size of a golf ball. According to specifications, these tracers have a range of 1,300 feet. Tracer bullets have been detected from a distance of 6 miles, but there is usually difficulty in pinpointing the survivor. A survivor should only use

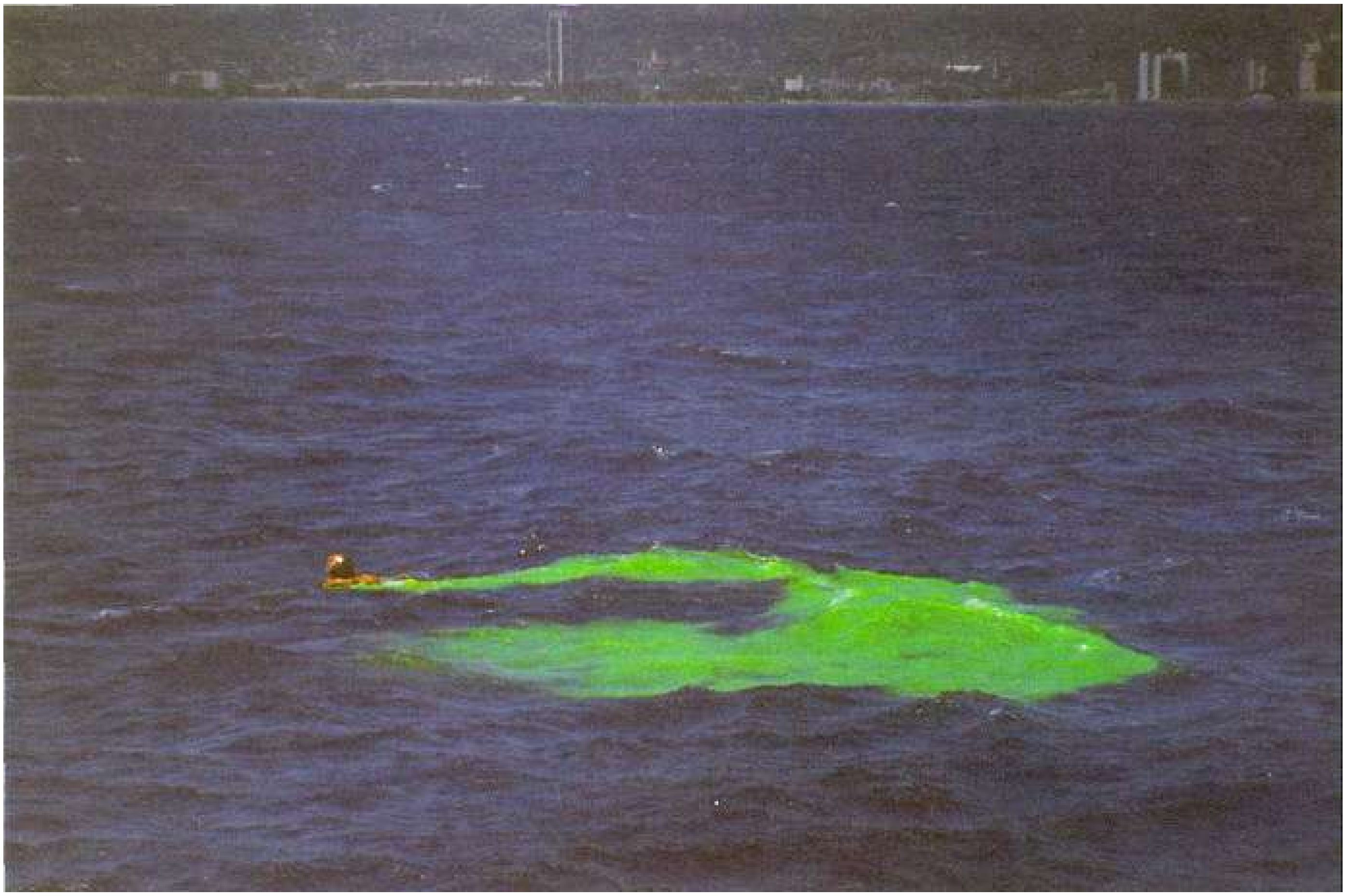

Figure 24-3. Sea Marker Dye.

this signaling device when rescue forces can be seen or heard. Do not direct this device *at* the aircraft.

(5) Because of the rapid changing technology in pyrotechnic signaling devices, an aircrew member should check regularly for new and improved models, making special note of the firing procedures and safety precautions necessary for their operations.

c. Sea Marker:

(1) Of the many dyes and metallic powders tested at various times for marking the sea, the most successful is the fluorescent, water-soluble, orange powder. When released in the sea, a highly visible, light green, fluorescent cast is produced. Sea marker dye has rapid dispersion power; a packet spreads into a slick about 150 feet in diameter and lasts an hour or more in calm weather. Rough seas will stream it into a long streak, which may disperse in 20 minutes (figure 24-3).

(2) Under ideal weather conditions, the dye can be sighted at 5 miles with the aircraft operating at 1,000 feet. The dye has also been spotted at 7 miles away from an aircraft operating at 2,000 feet.

(3) Sea marker dye should be used in friendly areas during daytime and only when there is a chance of being sighted (aircraft seen or heard in the immediate area). It is not effective in heavy fog, solid overcast and storms with high winds and waves. The release tab on the packet of dye is pulled to open for use. In calm water, the dye can be dispersed more rapidly stirring the water with paddles or hands.

(4) If left open in the raft, the escaping powder penetrates clothing, stains hands, face, and hair, and eventually may contaminate food and water. To avoid the inevitable messiness, some survivors have tied the sea marker dye to the sea anchor. Others have dipped the packet over the side, letting it drain off the side into the sea. After using the dye, it should be rewrapped to conserve the remainder of the packet.

d. Paulin Signals. The paulin is a conventional signaling device used to send specific messages to aircraft. It may be packed with some sustenance kits and multiplace liferaft accessory kits. The paulin is constructed of rubberized nylon material and is blue on one side and yellow on the other. These colors contrast against each other so when one side is folded over the other, the designs are easily distinguished (figure 24-4). The size is 7 feet by 11 feet which is a disadvantage when folded because it makes a small signal. The paulin has numerous uses. It can be used as a camouflage cloth, sunshade, tent, or sail, or it can be used to catch drinking water. The space blanket, used as a substitute for the sleeping bag in some survival kits, can be used in the same manner as the signal paulin because it is highly reflective (silver on one side and various colors on the other side).

e. Audio Signals. Sounds carry far over water under ideal conditions; however, they are easily distorted and deadened by the wind, rain, or snow. On land, heavy foliage cuts down on the distance sound will travel. Shouting and whistling signals have been effective at short ranges for summoning rescue forces. Most contacts using these means were made at less than 200 yards, although a few reports claim success at ranges of up to a mile. A weapon can be used to attract attention by firing shots in a series of three. The number of available rounds determine whether this is practical. Survivors have used a multitude of devices to produce sound. Some examples are: striking two poles together, striking one pole against a hollow tree or log, and improvising whistles out of wood, metal, and grass.

f. Light Signals. When tested away from other manufactured lights, aircraft lights have been seen up to 85 miles. At night, a survivor should use any type of light to attract attention. A signal with a flashlight, or a light or fire in a parachute shelter, can be seen from a long distance. A flashing light (strobelight) is in most survival kits.

g. Signal Mirror.

(1) The signal mirror is probably the most underrated signaling device found in the survival kit. It is the most valuable daytime means of visual signaling. A mirror flash has been visible up to 100 miles under ideal conditions, but its value is significantly decreased unless it is used correctly. It also works on overcast days. Practice is the key to effective use of the signal mirror. Whether the mirror is factory manufactured or improvised, aim it so the beam of light reflected from its surface hits the overflying aircraft.

(2) The signal mirror's effectiveness is its greatest weakness if the survivor is in enemy territory. It is just as bright to the enemy as to the rescuer; use it wisely! Survivors should understand that even if the mirror flash is directly on the aircraft (especially if the aircraft is using terrain masking techniques), that same flash may be visible to others (possibly the enemy) who are located at the proper angle in regard to the survivors' position.

(3) In a hostile environment, the exact location of the flash is extremely important. The signal mirror should be covered when not in use. One of the easiest

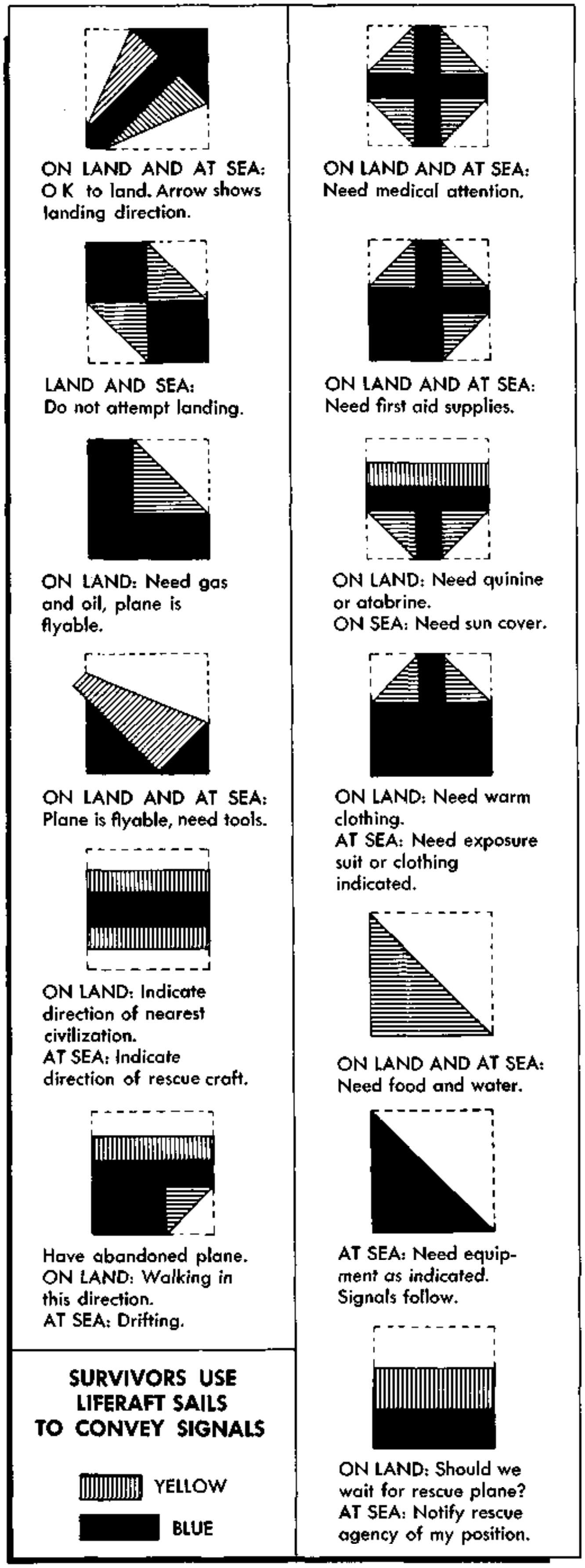
ON LAND AND AT SEA: O K to land. Arrow shows landing direction.
LAND AND SEA: Do not attempt landing.
ON LAND: Need gas and oil, plane is flyable.
ON LAND AND AT SEA: Plane is flyable, need tools.
ON LAND: Indicate direction of nearest civilization. AT SEA: Indicate direction of rescue craft.
Have abandoned plane. ON LAND: Walking in this direction. AT SEA: Drifting.
SURVIVORS USE LIFERAFT SAILS TO CONVEY SIGNALS
YELLOW
BLUE
ON LAND AND AT SEA: Need medical attention.
ON LAND AND AT SEA: Need first aid supplies.
ON LAND: Need quinine or atabrine. ON SEA: Need sun cover.
ON LAND: Need warm clothing. AT SEA: Need exposure suit or clothing indicated.
ON LAND AND AT SEA: Need food and water.
AT SEA: Need equipment as indicated. Signals follow.
ON LAND: Should we wait for rescue plane? AT SEA: Notify rescue agency of my position.

Figure 24-4. Paulin Signals.

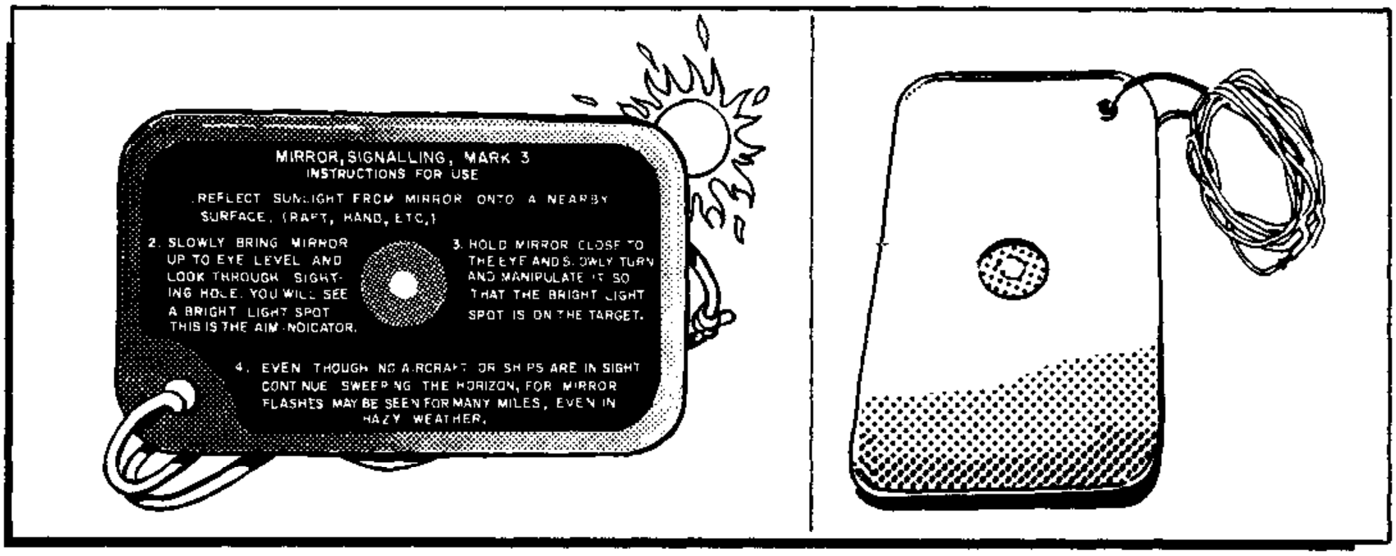

MIRROR, SIGNALLING, MARK 3
INSTRUCTIONS FOR USE

.REFLECT SUNLIGHT FROM MIRROR ONTO A NEARBY SURFACE. (RAFT, HAND, ETC.)

2. SLOWLY BRING MIRROR UP TO EYE LEVEL AND LOOK THROUGH SIGHTING HOLE. YOU WILL SEE A BRIGHT LIGHT SPOT THIS IS THE AIM INDICATOR.

3. HOLD MIRROR CLOSE TO THE EYE AND SLOWLY TURN AND MANIPULATE IT SO THAT THE BRIGHT LIGHT SPOT IS ON THE TARGET.

4. EVEN THOUGH NO AIRCRAFT OR SHIPS ARE IN SIGHT CONTINUE SWEEPING THE HORIZON, FOR MIRROR FLASHES MAY BE SEEN FOR MANY MILES, EVEN IN HAZY WEATHER.

Figure 24-5. Signal Mirrors.

methods is to tie the string from the mirror around the neck and tuck the mirror in the shirt or flight suit. When the mirror is removed from inside the clothing, the hand should be placed over the mirror surface to prevent accidental flashing. The covered mirror may then be raised toward the sky and the hand withdrawn. The flash can then be directed onto the free hand and the aiming indicator (sunspot) located. This minimizes the indiscriminate flashing of surrounding terrain. When putting the mirror away, the survivor should remember to cover the mirror to prevent a flash.

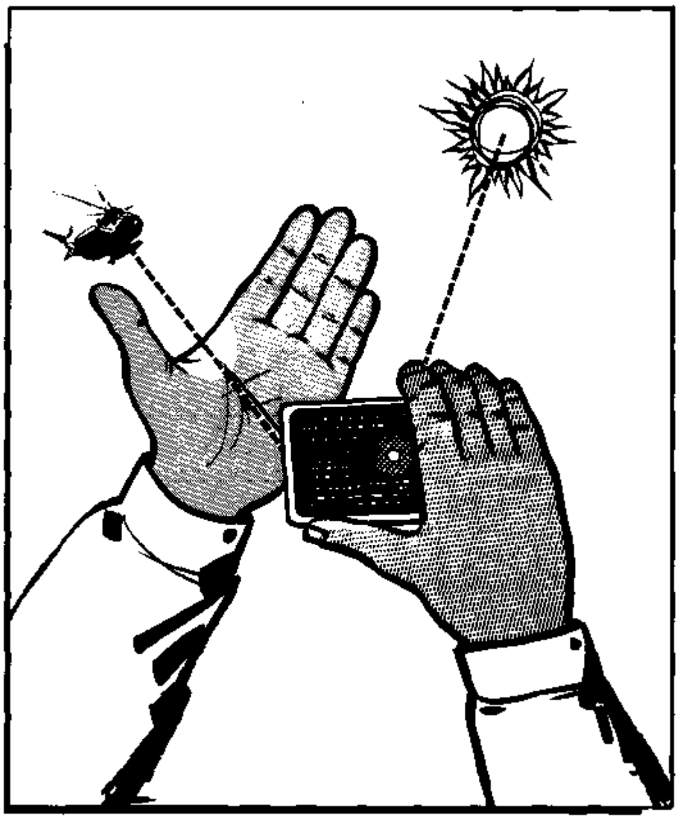

Figure 24-6. Aiming Signal Mirror.

h. Aiming Manufactured Mirrors. Instructions are printed on the back of the mirror. Survivors should:

(1) Reflect sunlight from the mirror onto a nearby surface—raft, hand, etc.

(2) Slowly bring the mirror up to eye-level and look through the sighting hole where a bright spot of light will be seen. This is the aim indicator.

(3) Hold mirror near the eye and slowly turn and manipulate it so the bright spot of light is on the target.

(4) In friendly areas, where rescue by friendly forces is anticipated, free use of the mirror is recommended. Survivors should continue to sweep the horizon even though no aircraft or ships are in sight (figure 24-5).

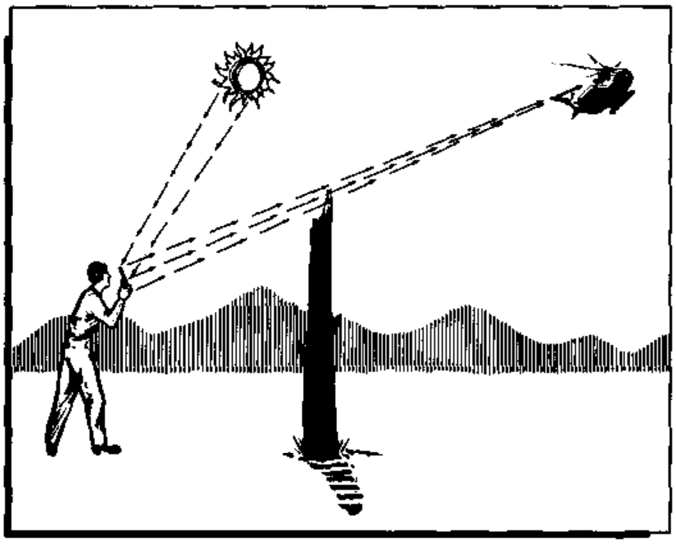

Figure 24-7. Aiming Signal Mirror-Stationary Object.

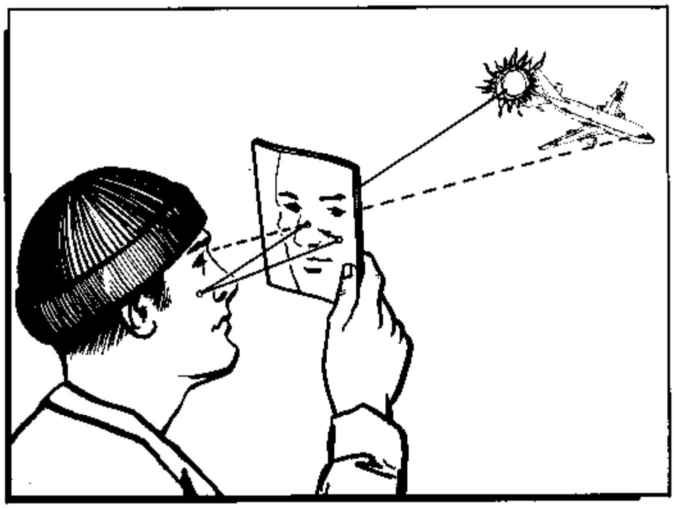

Figure 24-8. Aiming Signal Mirror—Double-Faced Mirror.

24-3. Improvised Signals:

a. Signal Mirrors. Improvised signal mirrors can be made from ration tins, parts from an aircraft, polished aluminum, glass, or the foil from rations or cigarette packs. However, the mirror must be accurately aimed if the reflection of the Sun in the mirror is to be seen by the pilot of a passing aircraft or the crew of a ship.

b. Aiming Improvised Mirrors:

(1) The simple way to aim an improvised mirror is to place one hand out in front of the mirror at arm's length and form a "V" with two fingers. With the target in the "V" the mirror can be manipulated so that the majority of light reflected passes through the "V" (figure 24-6). This method can be used with all mirrors. Another method is to use an aiming stake as shown in figure 24-7. Any object 4 to 5 feet high can serve as the point of reference.

(2) Survivors should hold the mirror so they can sight along its upper edge. Changing their position until the top of the stick and target line up, they should adjust the angle of the mirror until the beam of reflected light hits the top of the stick. If stick and target are then kept in the sighting line, the reflection will be visible to the rescue vehicle.

(3) Another method is to improvise a double-faced mirror (shiny on both sides). A sighting hole can be made in the center of the mirror.

(a) When trying to attract the attention of a friendly rescue vehicle that is no more than 90 degrees from the Sun, proceed as shown in figure 24-8.

(b) The survivor's first step will be to hold the double-faced mirror about 3 to 6 inches away from the face and sight at the rescue target through the hole in the center of the mirror. The light from the Sun shining through the hole will form a spot of light on the survivor's face. This spot will be reflected in the rear surface of the mirror. Then, aiming at the rescue vehicle through the hole, the survivor can adjust the angle of the mirror until the reflection of the spot on the face in the rear surface of the mirror lines up with, and disappears, into the sighting hole.

(c) When the reflected spot disappears and the rescue vehicle is still visible through the hole, the survivor can be sure the reflected light from the Sun is accurately aimed. The survivor may also "shimmer" the mirror by moving it rapidly over the target. This ensures that the part of the bright flash the rescuers see coincides with the position of the survivor. This (shimmering) is especially useful on a moving target.

(d) When the angle between the target and the Sun is more than 90 degrees (when the survivor is between the rescue vehicle and the Sun) a different method may be used for aiming. The survivor should adjust the angle of the mirror until the spot made by the

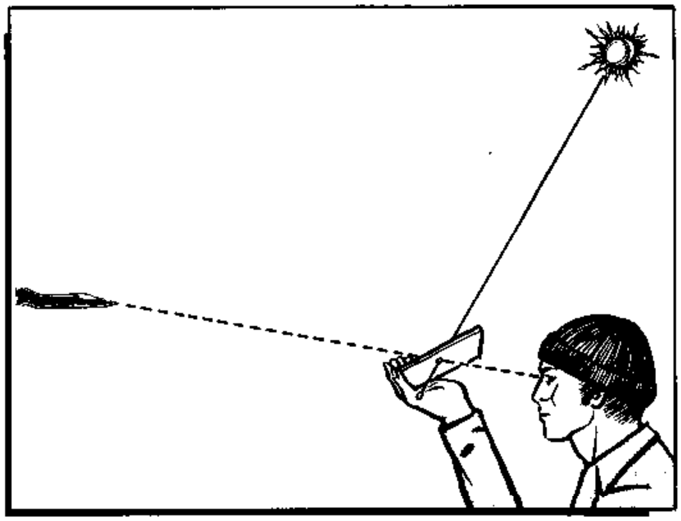

Figure 24-9. Aiming Signal Mirror—Angle Greater than 180 Degrees.

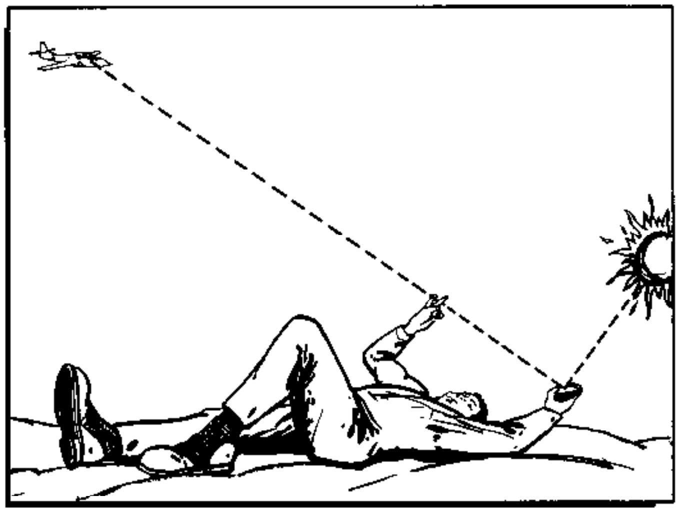

Figure 24-10. Aiming Signal Mirror—Angle Greater than 180 Degrees (Another Method).

Figure 24-11. Shelter as a Signal.

Sun's rays passing through the hole in the mirror lands on the hand instead of on the face. The reflection in the back of the mirror that comes off the hand may then be manipulated in the same way (figure 24-9).

(e) Another method used when the angle is greater than 90 degrees is to lie on the ground in a large clearing and aim the mirror using one of the methods previously discussed (figure 24-10).

c. Fire and Smoke Signals:

(1) Fire and smoke can be used to attract the attention of recovery forces. Three evenly spaced fires, 100 feet apart, arranged in a triangle or in a straight line, serve as an international distress signal. One signal fire will usually work for a survivor. During the night, the flames should be as bright as possible, and during the day, as much smoke as possible should be produced.

(2) Smoke signals are most effective on clear and calm days. They have been sighted from up to 50 miles away. High winds, rain, or snow tend to disperse the smoke and lessen the chances of it being seen. Smoke signals are not dependable when used in heavily wooded areas.

(3) The smoke produced should contrast with its background. Against snow, dark smoke is most effective. Likewise, against a dark background, white smoke is best. Smoke can be darkened by rags soaked in oil, pieces of rubber, matting, or electrical insulation, or plastic being added to the fire. Green leaves, moss, ferns, or water produce white smoke.

(4) To increase its effectiveness, the signal fire must be prepared before the recovery vehicle enters the area.

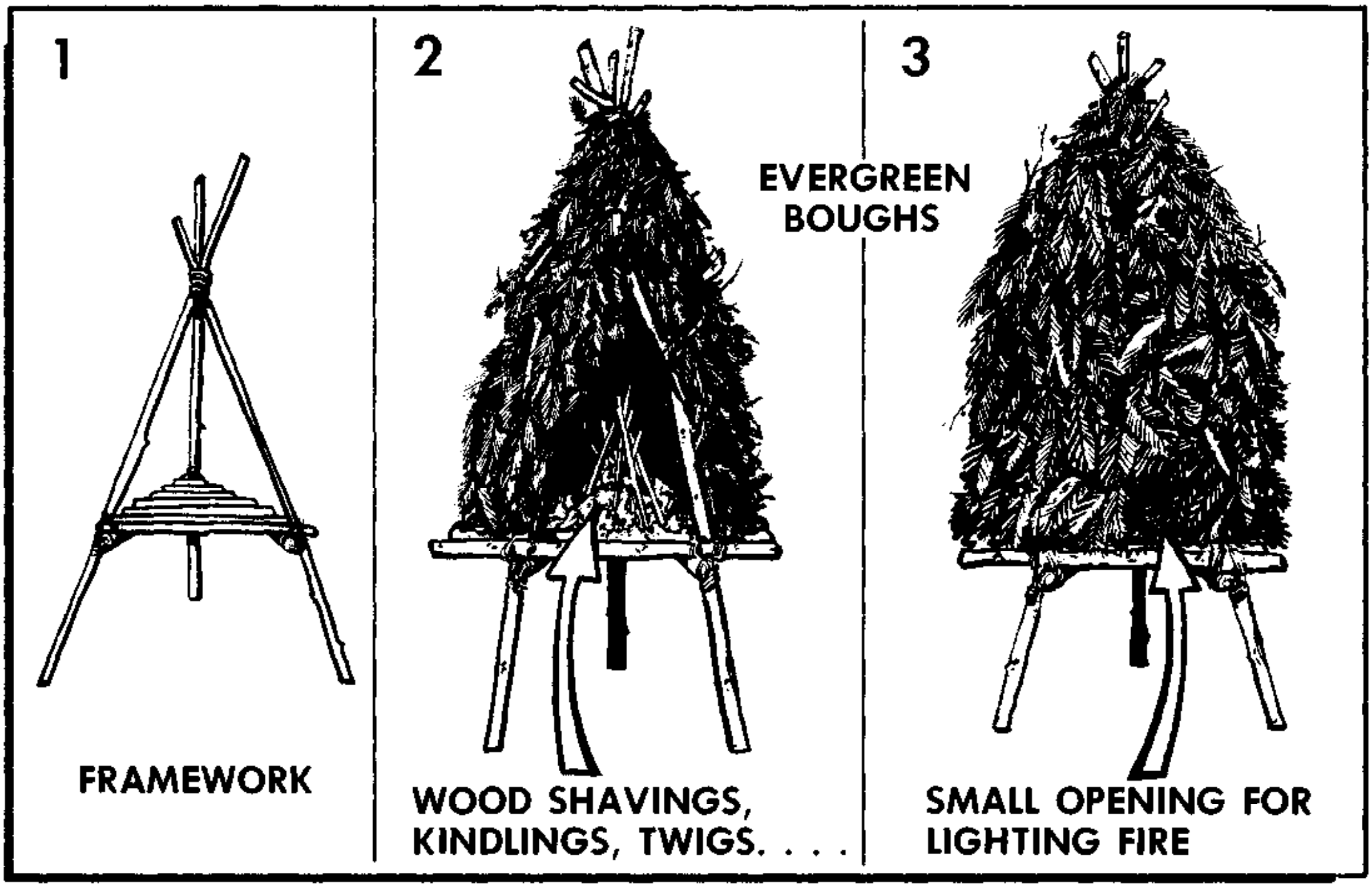
1
2
3
EVERGREEN BOUGHS
FRAMEWORK
WOOD SHAVINGS, KINDLINGS, TWIGS. . . .
SMALL OPENING FOR LIGHTING FIRE

Figure 24-12. Smoke Generator—Platform.

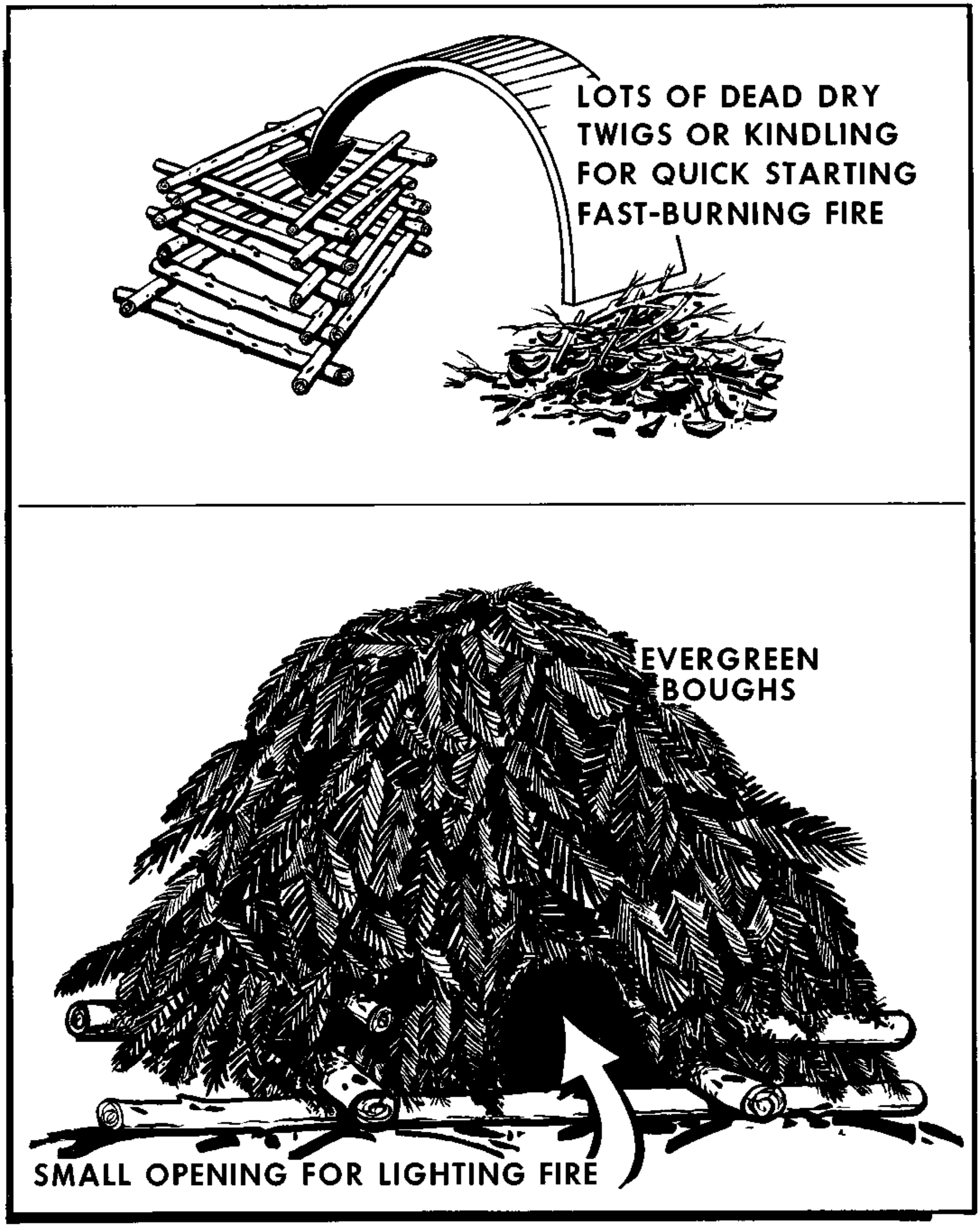
LOTS OF DEAD DRY TWIGS OR KINDLING FOR QUICK STARTING FAST-BURNING FIRE
EVERGREEN BOUGHS
SMALL OPENING FOR LIGHTING FIRE

Figure 24-13. Smoke Generator—Ground.

The fires used by survivors for heat and cooking may be used as signal fires as long as the necessary materials are available in the immediate vicinity. Survivors should supplement the fire to provide the desired signal (figure 24-11).

(5) Smoke Generators:

(a) Raised Platform Generator (figure 24-12). The survivor should:

-1. Build a raised platform above wet ground or snow.

-2. Place highly combustible materials on the platform.

-3. Then place smoke-producing materials over the platform and light when search aircraft is in the immediate vicinity.

(b) Ground Smoke Generator (figure 24-13). The survivor should:

-1. Build a large log cabin fire configuration on the ground. This provides good ventilation and supports the green boughs used for producing smoke.

-2. Place smoke-producing materials over the fire lay; ignite when a search aircraft is in the immediate vicinity.

(c) Tree Torch Smoke Generator (figure 24-14). To build this device, the survivor should:

-1. Locate a tree in a clearing to prevent a forest fire hazard.

-2. Add additional smoke-producing materials.

-3. Add igniter.

-4. Light when a search aircraft is in the immediate vicinity.

(d) Fuel Smoke Generator. If survivors are with the aircraft, they can improvise a generator by burning aircraft fuels, lubricating oil, or a mixture of both. One to 2 inches of sand or fine gravel should be placed in the bottom of a container and saturated with fuel. Care should be used when lighting the fuel as an explosion may occur initially. If there is no container available, a hole can be dug in the ground, filled with sand or gravel, saturated with fuel, and ignited. Care should be taken to protect the hands and face.

d. Pattern Signals. The construction and use of pattern signals must take many factors into account. Size, ratio, angularity, contrast, location, and meaning are each important if the survivors' signals are to be effec-

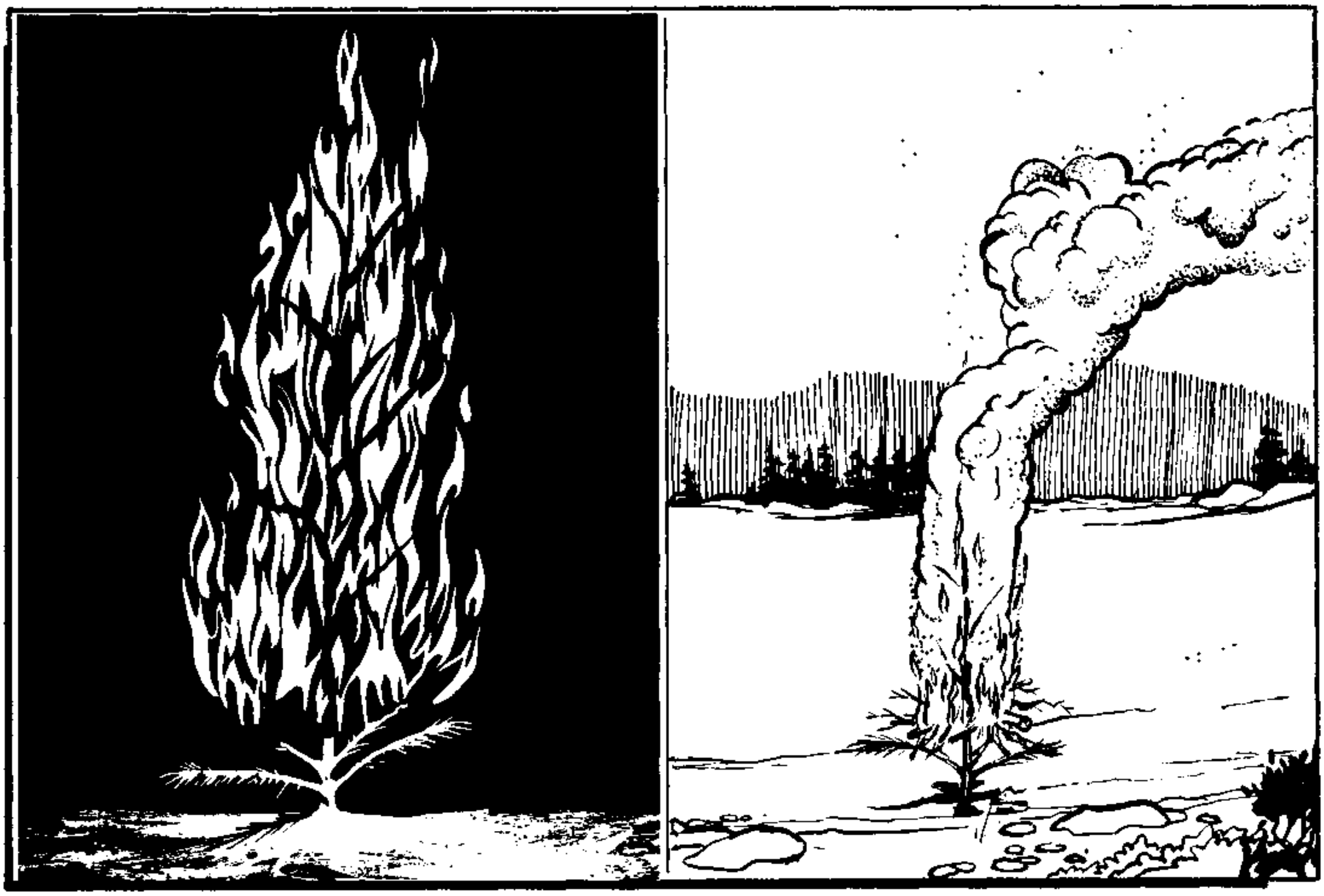

Figure 24-14. Tree Torch.

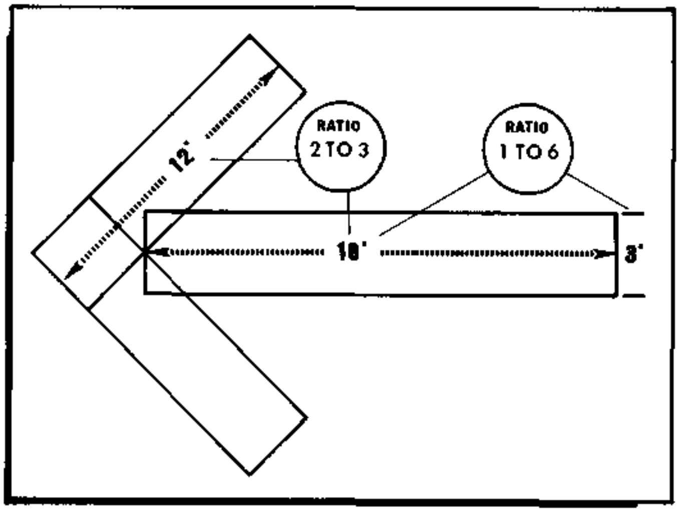

Figure 24-15. Pattern Signal Sizes.

tive. The type of signal constructed will depend on the material available to survivors. Not every crewmember will have a parachute, so ingenuity plays an important role in the construction of the signal. Survivors should remember to judge their signals from the standpoint of aircrew members who are flying over their location searching for them.

(1) Size. The signal should be as large as possible. To be most effective, the signal should have "lines" no less than 3 feet wide and 18 feet long (1:6) (figure 24-15).

(2) Ratio. Proper proportion should also be remembered. For example, if the baseline of an "L" is 18 feet long, then the vertical line of the "L" must be longer (27 feet), a 2 to 3 ratio, to keep the letter in proper proportion.

(3) Angularity. Straight lines and square corners are not found in nature. For this reason, survivors should make all pattern signals with straight lines and square corners.

(4) Contrast. The signal should stand out sharply against the background. The idea is to make the signal look "larger." On snow, the fluorescent sea dye available in the liferaft accessory kit can be used to add contrast around the signal. The survivor should do everything

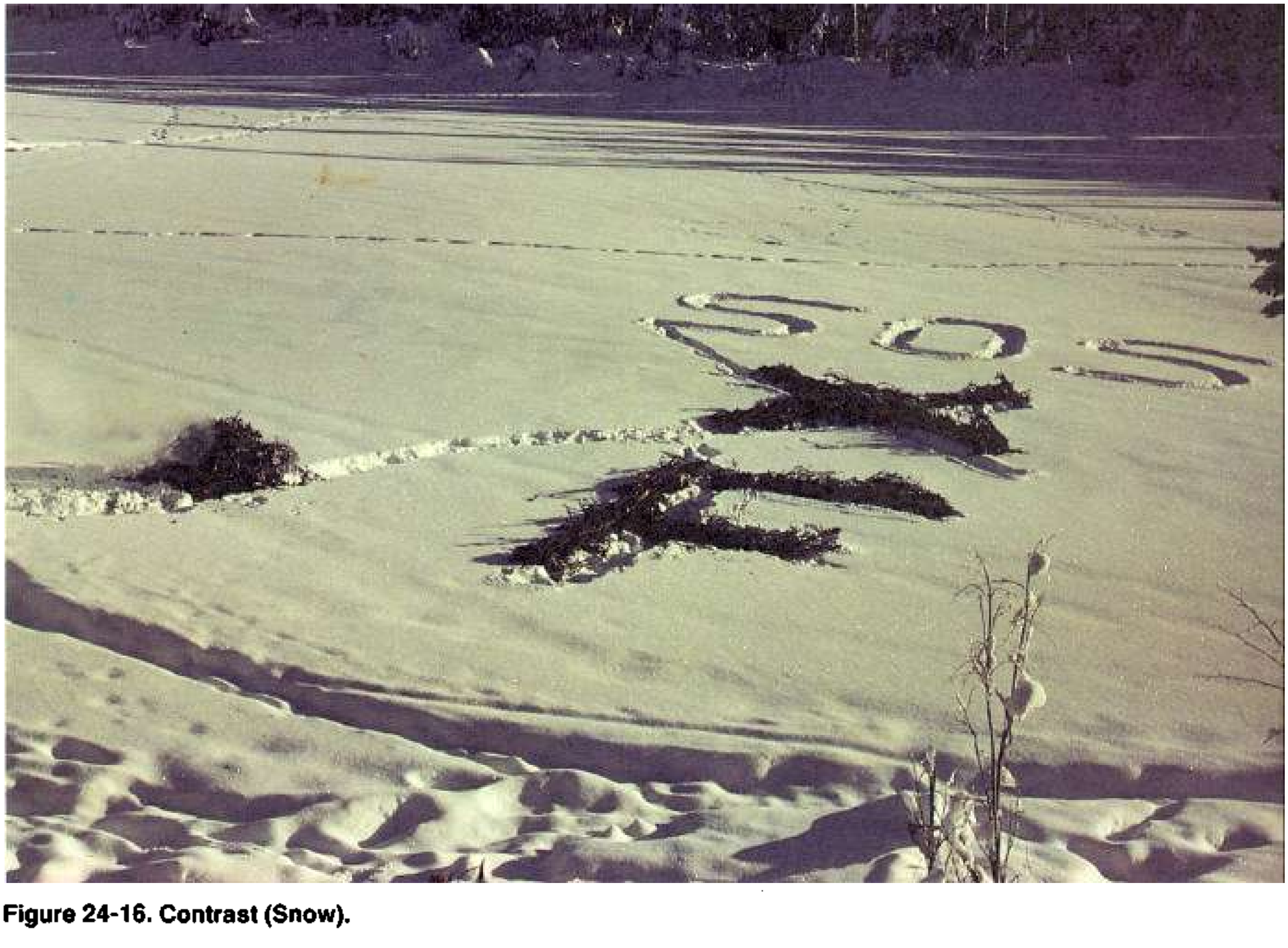

Figure 24-16. Contrast (Snow).

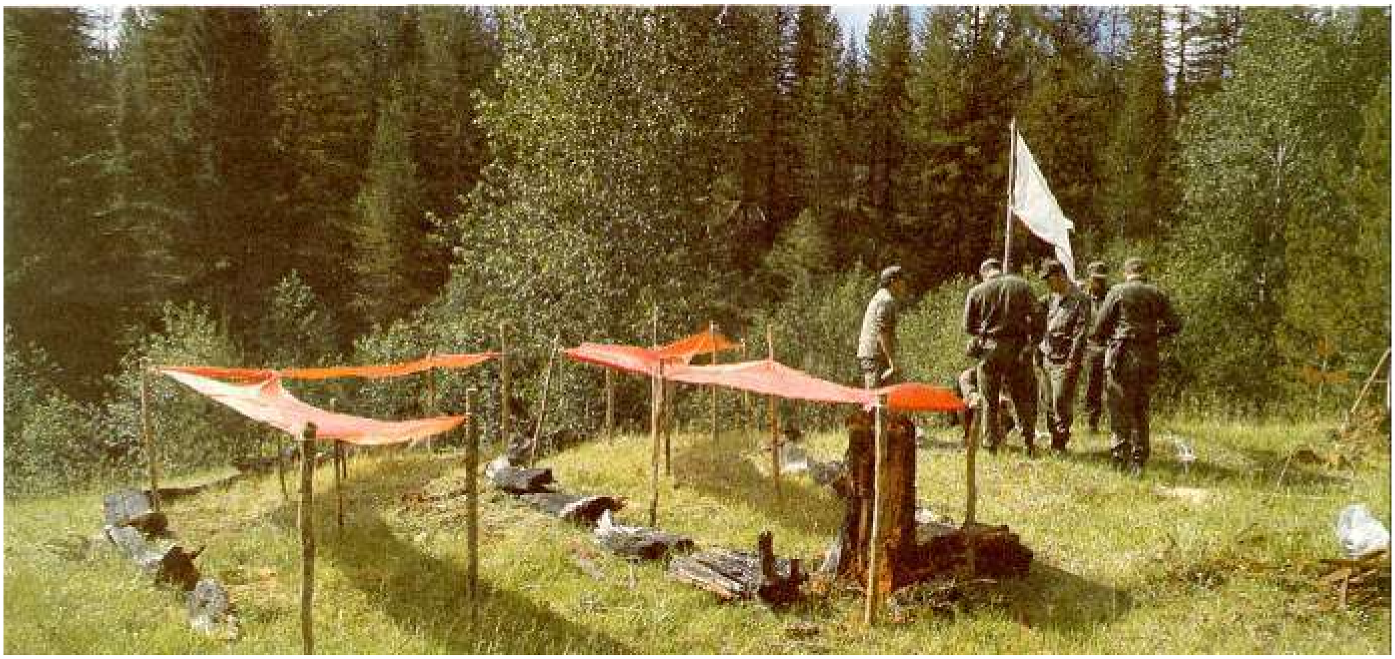

Figure 24-17. Contrast.

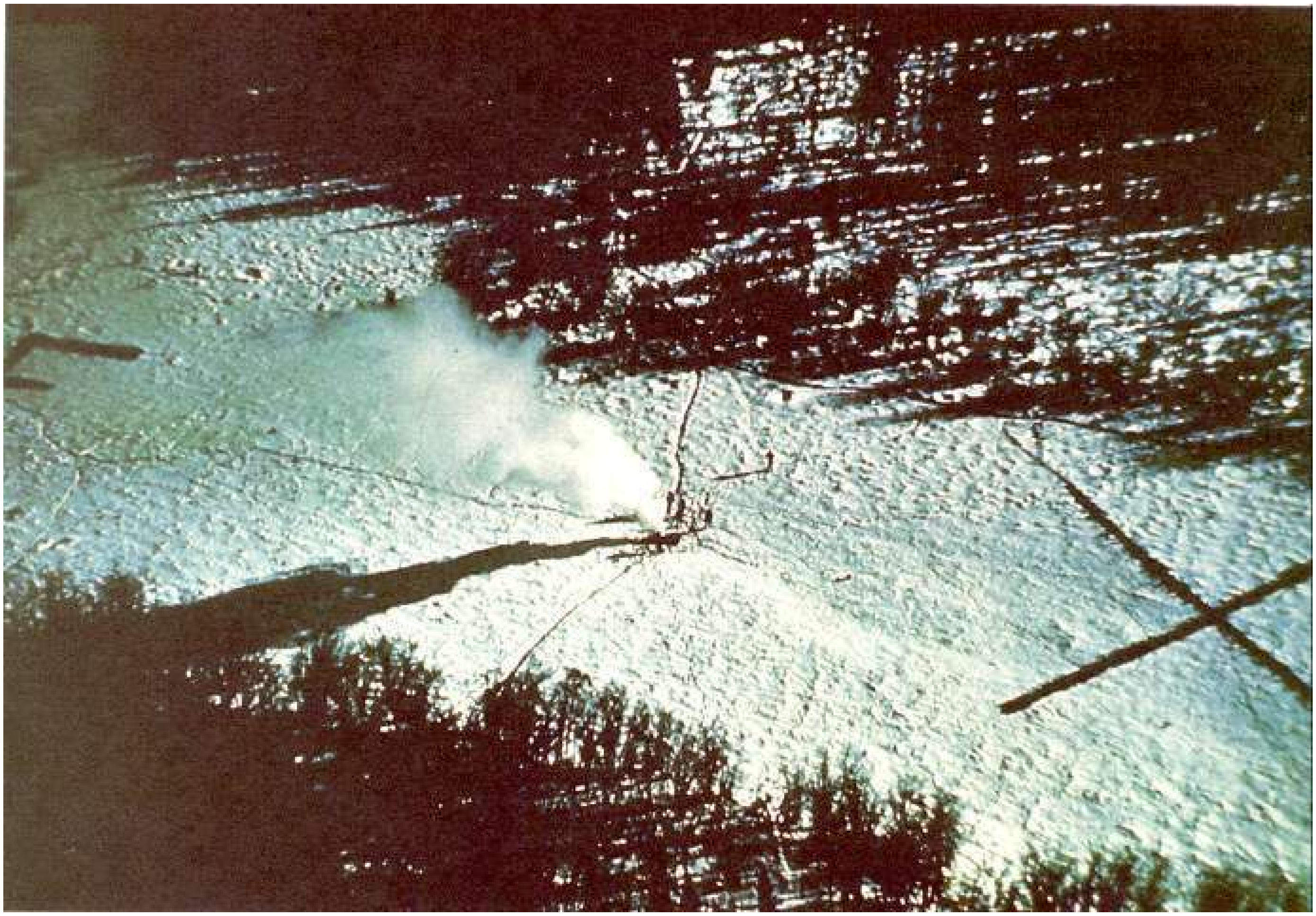

Figure 24-18. Location.

NO.	MESSAGE	CODE SYMBOL
1	REQUIRE ASSISTANCE	V
2	REQUIRE MEDICAL ASSISTANCE	X
3	NO or NEGATIVE	N
4	YES or AFFIRMATIVE	Y
5	PROCEEDING IN THIS DIRECTION	↑

Figure 24-19. Signal Key.

possible to disturb the natural look of the ground. In grass and scrubland, the grass should be stamped down or turned over to allow the signal to be easily seen from the air. A burned grass pattern is also effective. When in snow, a trampled out signal is very effective. Survivors should use only one path to and from the signal to avoid disrupting the signal pattern. Avoid using orange parachute material on a green or brown background as it has a tendency to blend in (figure 24-16). Contrast can be improved by outlining the signal with green boughs, piling brush and rocks to produce shadows, or raising the panel on sticks to cast its own shadow (figure 24-17).

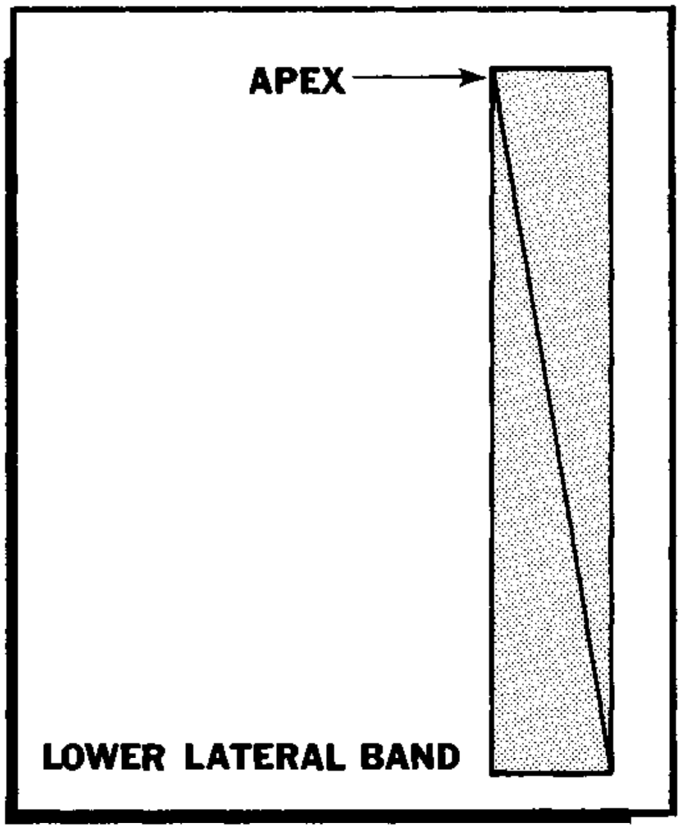

Figure 24-20. Parachute Strips.

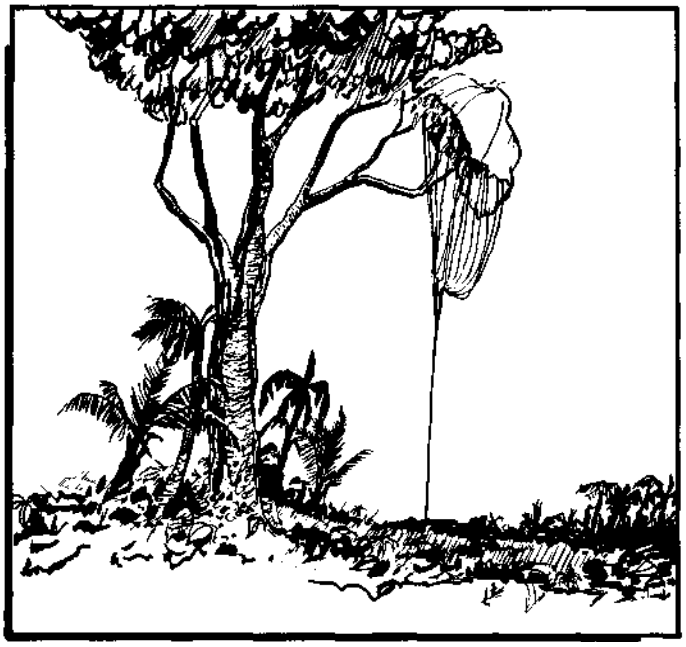

Figure 24-21. Parachute in Tree.

(5) Location. The signal should be located so it can be seen from all directions. Survivors should make sure the signal is located away from shadows and overhangs. A large high open area is preferable. It can serve a dual function—one for signaling and the other for rescue aircraft to land (figure 24-18).

(6) Meaning. If possible, the signal should tell the rescue forces something pertaining to the situation. For example: "require medical assistance," or a coded symbol used during evasion, etc. Figure 24-19 shows the internationally accepted symbols.

e. Parachute Signals:

(1) Parachute material can be used effectively to construct pattern signals. A rectangular section of parachute material can be formed as shown in figure 24-20. When making a pattern signal, survivors should ensure the edges are staked down so the wind will not blow the panels away.

(2) A parachute caught in a tree will also serve as a signal. Survivors should try to spread the material over the tree to provide the maximum amount of signal (figure 24-21).

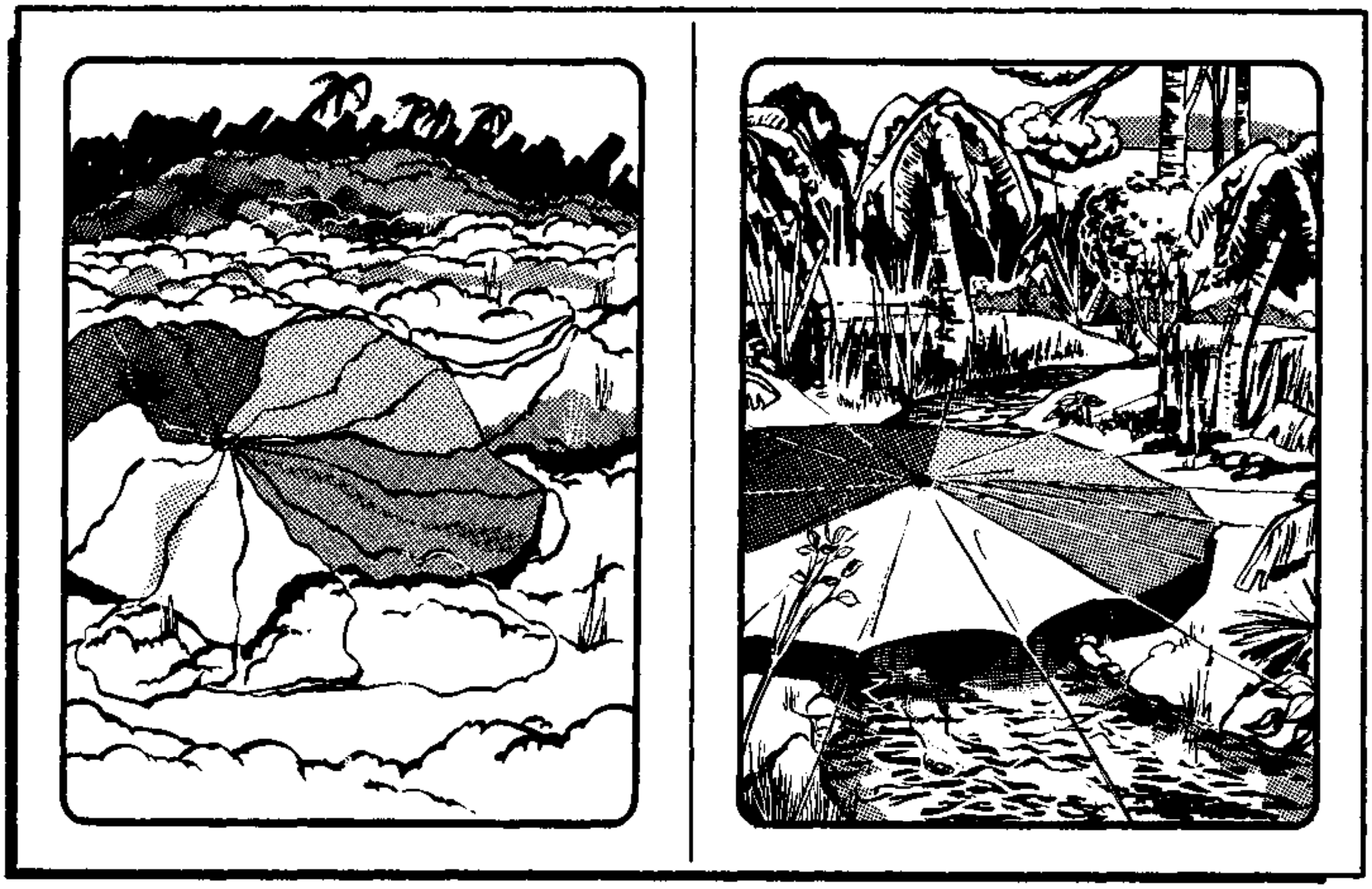

Figure 24-22. Chute Over Trees or Streams.

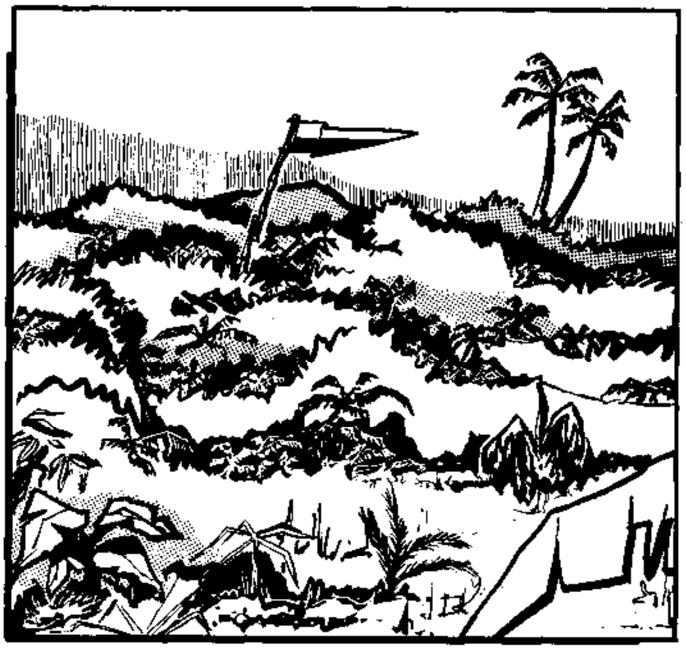

Figure 24-23. Pennants and Banners.

(3) When open areas are not available, survivors should stretch the chute over low trees and brush or across small streams (figures 24-22 and 24-23).

f. Shadow Signals. If no other means are available, survivors may have to construct mounds which will use the Sun to cast shadows. These mounds should be constructed in one of the international distress patterns. Brush, foliage, rocks, or snowblocks may be used to cast shadows. To be effective, these shadow signals must be oriented to the Sun to produce the best shadow. In areas close to the Equator, a north-south line gives a shadow at any time except noon. Areas farther north or south require the use of an east-west line or some point of the compass in between to give the best results.

g. Acknowledgements:

(1) Rescue personnel will normally inform the survivors they have been sighted by:

(a) Flying low with landing lights on (figure 24-24), and (or) rocking the wings.

(b) Emergency radio.

(2) Figure 24-25 depicts the standard body signals which can be used if electronic signaling devices are not available.

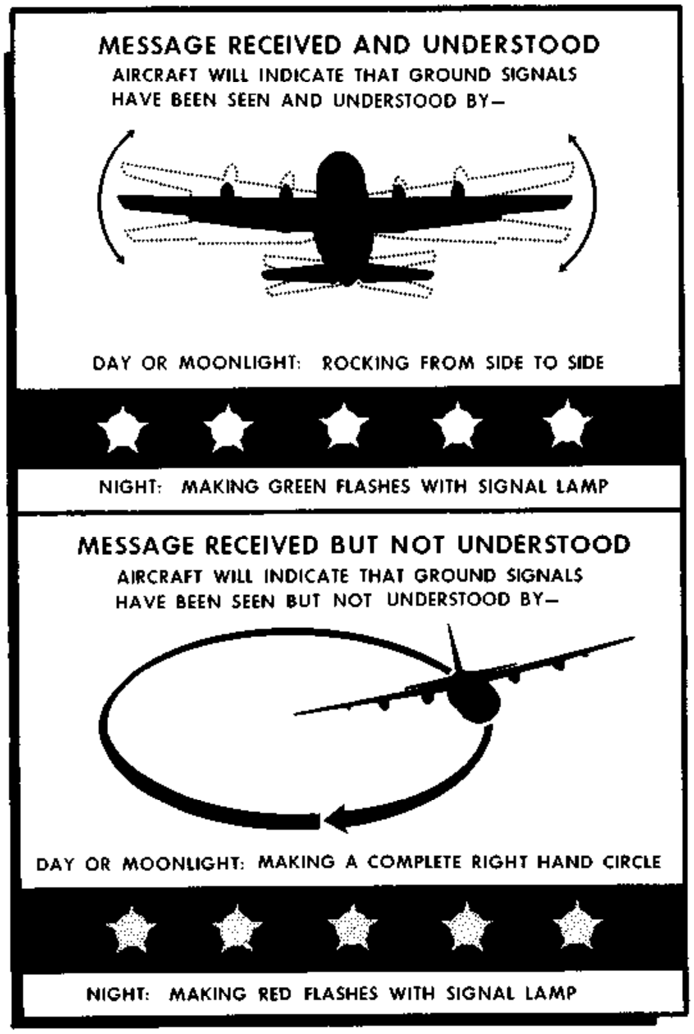
MESSAGE RECEIVED AND UNDERSTOOD
AIRCRAFT WILL INDICATE THAT GROUND SIGNALS HAVE BEEN SEEN AND UNDERSTOOD BY—
DAY OR MOONLIGHT: ROCKING FROM SIDE TO SIDE
NIGHT: MAKING GREEN FLASHES WITH SIGNAL LAMP
MESSAGE RECEIVED BUT NOT UNDERSTOOD
AIRCRAFT WILL INDICATE THAT GROUND SIGNALS HAVE BEEN SEEN BUT NOT UNDERSTOOD BY—
DAY OR MOONLIGHT: MAKING A COMPLETE RIGHT HAND CIRCLE
NIGHT: MAKING RED FLASHES WITH SIGNAL LAMP

Figure 24-24. Standard Aircraft Acknowledgements.

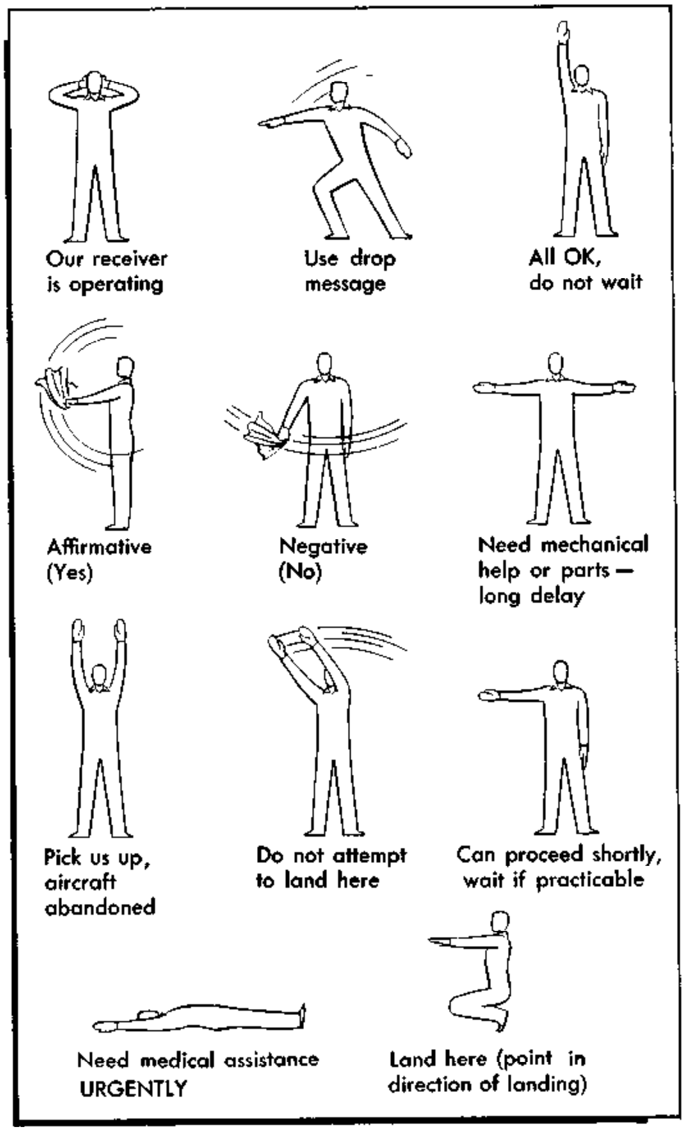
Our receiver is operating
Use drop message
All OK, do not wait
Affirmative (Yes)
Negative (No)
Need mechanical help or parts — long delay
Pick us up, aircraft abandoned
Do not attempt to land here
Can proceed shortly, wait if practicable
Need medical assistance URGENTLY
Land here (point in direction of landing)

Figure 24-25. Close-In Visual Signals.

Chapter 25

RECOVERY PRINCIPLES

25-1. Introduction:

a. Receipt of a distress call sets a highly trained and well-equipped organization into operation; however, prompt and safe recovery is by no means ensured. The success of the rescue effort depends on many factors. Such factors as the availability of rescue forces, the proximity of enemy forces, and weather conditions can affect the success of the rescue. Above all, the survivors' knowledge of what to do in the rescue effort may make the difference between success and failure (figure 25-1).

b. The role of survivors in effecting their rescue changes continuously as aircraft and rescue equipment become more sophisticated. The probability of a downed aircrew member applying long-term survival training concepts under noncombat conditions continues to decrease while increasing under combat conditions.

c. There are several independent organizations engaged in search and rescue (SAR) operations or influencing the SAR system. The organizations may be international, federal, state, county, or local governmental, commercial, or private organizations. Survivors are responsible for being familiar with procedures used by international SAR systems in order to assist in rescue efforts. Some international organizations are:

(1) International Civil Aviation Organization (ICAO).

(2) Intergovernmental Maritime Consultive Organization (IMCO).

(3) Automated Mutual-Assistance Vessel Rescue (AMVER) System.

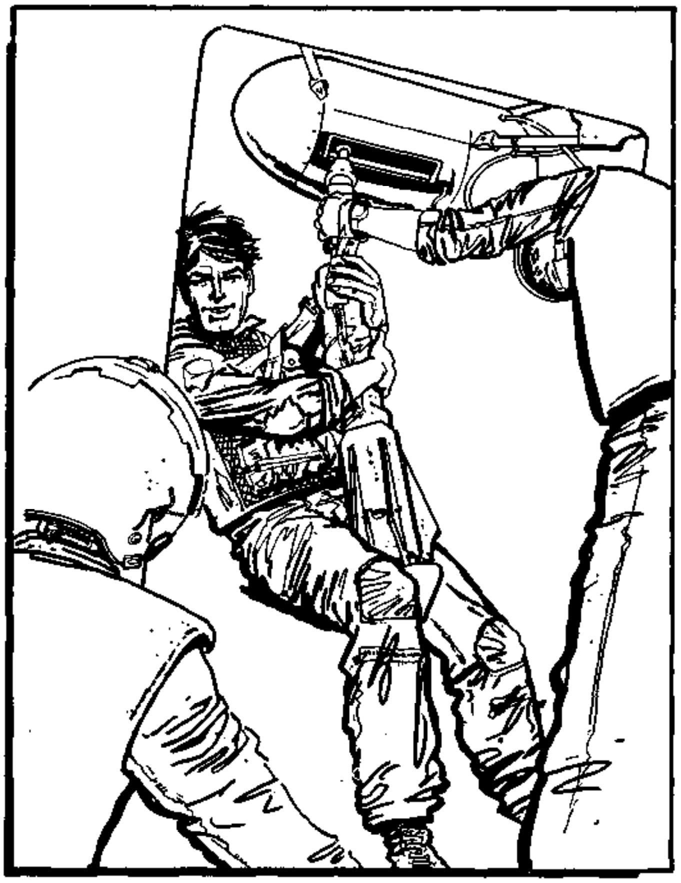

Figure 25-1. Recovery.

25-2. National Search and Rescue (SAR) Plan:

a. The National SAR Plan is implemented the instant an aircraft is known to be down. There are three primary SAR regions; they are the Inland Region, the Maritime Region, and the Overseas Region.

b. The Air Force is the SAR coordinator for the Inland Region, which encompasses the continental United States. The Coast Guard is the SAR coordinator for the Maritime Region, which includes the Carribean Area and Hawaii. The Third National Region is the Overseas Region. The Secretary of Defense designates certain Defense Department officers as United Commanders of specified areas where US Forces are operating. Wherever such commands are established, the Unified Commander is the Regional SAR Coordinator. Overseas regions are normally served by the Joint Rescue Coordination Center, operated under the Unified Action Armed Forces. Under the terms of the National SAR Plan, the "inland" area of Alaska is considered a part of the Overseas Region.

c. The National SAR Manual, Air Force Manual 64-2, provides a long-range rescue plan which personnel should study for additional information.

25-3. Survivors' Responsibilities:

a. The survivors' responsibilities begin at the onset of the emergency, with the dispatching of an immediate radio message. The radio message should include position, course, altitude, groundspeed, and actions planned. This information is essential for initiating efficient recovery operations.

b. Once recovery operations have been initiated, survivors have a continuing responsibility to furnish information. Both ground and radio signals should be immediate considerations.

c. If a group of survivors should become separated, each group member should, when contacted by rescue forces, provide information surrounding the dispersal of the group.

d. The greatest responsibility of aircrew members is to follow all instructions to the letter. The intelligence officer will brief aircrew members on procedures for tactical situations. These instructions must be followed explicitly since it could mean the difference between life and death. When rescue personnel tell the survivor to unhook from the raft—it should be done immediately! If instructions are not followed, survivors could be responsible for causing their own death and (or) the death of rescue personnel.

25-4. Recovery Site:

a. Consideration must be given to a recovery site. The survivor's major considerations are the type of recovery vehicle carrying out the recovery and the effects of the weather and terrain on the rescue aircraft, such as updrafts and downdrafts, heat, wind, etc. Survivors should try to pick the highest terrain possible in the immediate area for pickup. When locating this rescue site, they should watch for obstacles such as trees, cliffs, etc., which could limit the aircraft's ability to maneuver. Overhangs, cliffs, or sides of steep slopes should be avoided. Such terrain features restrict the approach and maneuverability of the rescue vehicle and require an increase in rescue time.

b. Even though survivors should select a recovery site, it is the ultimate responsibility of rescue personnel to decide if the selected site is suitable.

25-5. Recovery Procedures:

a. Knowing Current Procedures. Since procedures involving recovery vary with changes in equipment and rescue capability, survivors must always know the current procedures and techniques. This is particularly true of the procedures used for wartime recovery, which are in AFR 64-3.

(1) In deciding whether or not supplies should be dropped, rescue forces consider such factors as the relative locations of the distress site to rescue unit bases, the lapse expected before rescue is initiated, and the danger of exposure. If a delay is expected, supplies are usually dropped to survivors to help sustain and protect them while they await rescue. The mobility of survivors on the land generally makes it possible to recover equipment dropped some distance away, but airdrops at sea must be accurate.

(2) Aircraft with internal aerial delivery systems, such as the HC-130, are the most suitable for delivery of supplies to survivors. Aircraft having bomb bays or exterior racks capable of carrying droppable containers or packages of survival requisites are the next most suitable for dropping supplies. However, these aircraft are not always available for supply dropping operations, so aircraft not specifically designed for this function may have to be used.

b. Rescue by Helicopter:

(1) Helicopters make rescues by landing or hoisting. Landings are usually required at high altitudes due to limitations of helicopter power for maintaining a hover. Hoist recovery is the preferred method for effecting a water rescue. Helicopter landings are made for all rescues when a suitable landing site is available, and danger from enemy forces is not a problem. Hovering the helicopters and hoisting the survivor aboard requires

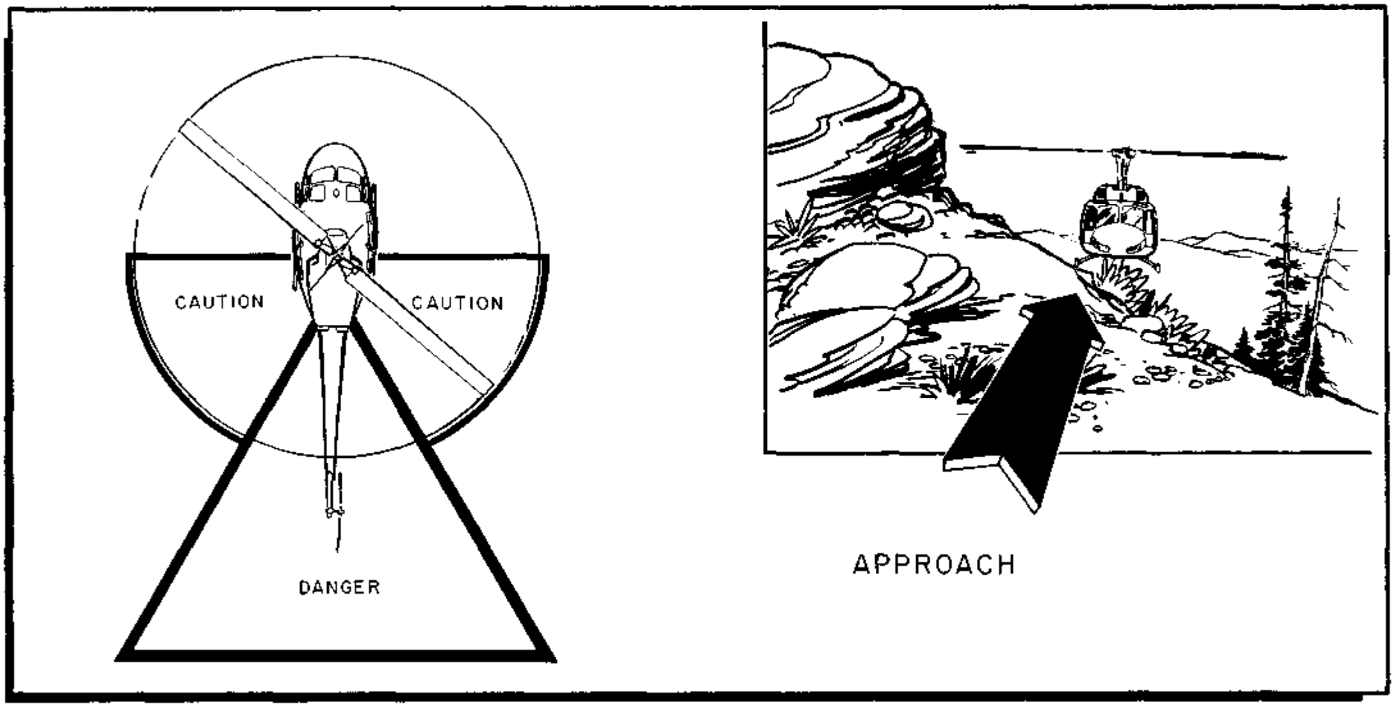

Figure 25-2. Approaching Helicopter.

more helicopter power than landing and presents a hazard to both the aircraft and the survivor. There is a danger if helicopters are operated close to collapsed parachutes. Parachute inflation by rotor downwash can cause the parachute to be sucked into the rotor blades of the helicopter.

(2) After landing, a crewmember will usually depart the aircraft. If for some reason this cannot be done, as in combat, the survivor should approach the helicopter from the 3 o'clock to 9 o'clock position relative to the nose of the helicopter and follow instructions (figure 25-2).

c. Rescue by Fixed-Wing Aircraft on Land:

(1) The most significant role played by fixed-wing aircraft in rescue operations is providing immediate assistance to survivors and serving as the "eyes" of approaching rescue units. This is done by pinpointing the survivors' position, orbiting the survivors, and dropping survival equipment. This type operation improves the morale of the survivors, fixes the survivors' location to prevent additional searching, and saves valuable time in getting the pickup unit on the scene.

(2) The role of fixed-wing aircraft in actually performing a rescue is limited to instances where there is a suitable runway near the survivor or where the aircraft is designed to operate from rough and improvised strips. Fixed-wing aircraft rescues have often been made in extremely cold climates where the aircraft have either used frozen lakes or rivers as runways or, when fitted with skis, have operated from snow-covered surfaces and glaciers. However, landing in unknown terrain under what appears to be ideal conditions is extremely hazardous.

d. Rescue by Ship:

(1) When a distress craft or survivors are a considerable distance from shore, rescue will normally be by long-range ships (specialized SAR ships, warships, or merchant ships). The rescue methods used by these ships vary considerably according to their displacement and whether the rescue is made in midocean or close to land. Weather, tides, currents, sea conditions, shallow water, reefs, daylight, or darkness may be important factors.

(2) Although it appears obvious that a marine craft should be used for rescue operations, it may be advisable to initiate an alternate method of recovery. For example, helicopters may be used to evacuate survivors picked up by marine craft in order to speed their delivery to an emergency care center.

(3) Removal of survivors from the water, liferafts, lifeboats, or other vessels to the safety of the rescue vessel deck may be the most difficult phase of a maritime search and rescue mission. In most cases, survivors will have to be assisted aboard. For this reason, all SAR vessels are usually equipped and prepared to lift survivors from the water without help from the survivors. There are numerous methods for rescuing survivors which may be used by SAR vessels. The most commonly used methods are listed in this chapter and are generally grouped as rescue of survivors in the water and rescue of survivors directly from their distressed vessel.

(a) When rescuing people from water, the following methods are generally used:

-1. Ship alongside/swimmer.
-2. Ship alongside/line thrower.
-3. Ship alongside/small boat.
-4. Ship circle/trail line.

(b) The most commonly used methods for rescuing personnel who are aboard distressed vessels are:

-1. Ship to ship/direct.
-2. Ship to ship/raft haul.
-3. Ship to ship/raft drift.
-4. Ship to ship/small boat.
-5. Ship to ship/haulaway line.

e. Rescue by Boat:

(1) When survivors are located on lakes, sheltered waters, rivers, or coastal areas, rescue will often be made by fast boats of limited range based close to the survivors or by private boats operating in the vicinity.

(2) Rescue boats are usually small and may not be able to take all survivors on board at one time; therefore, a sufficient number of boats to offset the rescue should be dispatched to the distress scene. When this is not possible, each boat should deploy its rafts so that those survivors who cannot be taken aboard immediately can be towed ashore or kept afloat while they are waiting. The boat crew should make sure any survivors who must be left behind are made as secure as circumstances permit.

(3) Assistance to an aircraft that has crashed or ditched on the water will usually consist of transferring personnel from plane to boat and picking up survivors from the water or liferafts. It may also include towing of an aircraft which is disabled on the water.

f. Coordinated Helicopter/Boat Rescues:

(1) Occasionally, boats and helicopters will be dispatched for a rescue operation. Generally the first rescue unit to arrive in the vicinity of the survivors will attempt the first rescue.

(2) If the helicopter arrives first, the boat will take a position upwind of the helicopter in the 2 o'clock position at a safe distance and stand by as a backup during the rescue attempt.

(3) If the helicopter must abort the rescue attempt, the pilot will depart the immediate area of the survivor and signal for the boat to move in and make its rescue attempt. Additionally, the helicopters may turn out the anticollision rotating beacon to indicate they require

boat assistance or are unable to complete the rescue. In certain operations where helicopter and boat coordinated rescue can be foreseen, specific signals should be prearranged.

(4) If the boat arrives first and makes the rescue, it will transfer the survivor to the helicopter to effect a rapid delivery to medical facilities.

25-6. Pickup Devices:

a. Assistance. When rescue forces are in the immediate area of survivors, they will, if conditions permit, deploy pararescue personnel to assist the survivors. Unfortunately, conditions may not always permit this, so survivors should know how to use different types of pickup devices.

b. Common Factors. Some common factors concerning all pickup devices are:

(1) The device should be allowed to ground to discharge static electricity before donning.

(2) To ensure stability, survivors should sit or kneel when donning a pickup device. Do not straddle the device.

(3) If no audio is available, survivors should visually signal the hoist operator when ready for lift-off—"thumbs up" or vigorously shake the cable from side to side.

(4) Most devices can be used as a sling (strop).

(5) Survivors must remember to follow all instructions provided by the rescue crew. When lifted to the door of the helicopter, survivors should not attempt to grab the door or assist the hoist operator in any way. They must not try to get out of the pickup device. The hoist operator will remove the device after the survivor is well inside the aircraft.

c. Rescue Sling. Before donning the rescue sling (strop), the survivor should face the drop cable and make sure the cable has touched the water or ground and lost its charge of static electricity.

(1) The most commonly accepted method for donning the rescue sling (strop) is the same as putting on a coat. After connecting the ring to form the sling (strop), the survivor's arms should be inserted one by one into the sling (strop) as it swings behind. The sling (strop) loop should be against the survivor's back with an arm around each side of the strop. The webbing under the metal ring can be held until tension is put on the cable. The survivor's hands may then be interlocked and rested on the chest. This tends to lock the survivor into the sling (strop) as upward pressure is applied (figure 25-3).

(2) Another way to enter the strop is to grasp the strop with both hands and lift it over the head to bring it down under the arms and around the body. Regardless of the method used, the survivor should remember the webbing and metal hardware of the device should be directly in front of the face.

d. Basket. If a basket is used, it will probably be accompanied to the water or ground by a member of the helicopter crew. The crewmember will assist survivors into the basket. There are two types of baskets: The litter type in which the person lies flat, and the seat type that survivors enter and sit down in as they would in a chair (figure 25-4).

e. Forest Penetrator:

(1) The forest penetrator rescue seat is designed to make its way through interlacing tree branches and dense jungle growth. It can also be used in open terrain or over water. The device is equipped with three seats

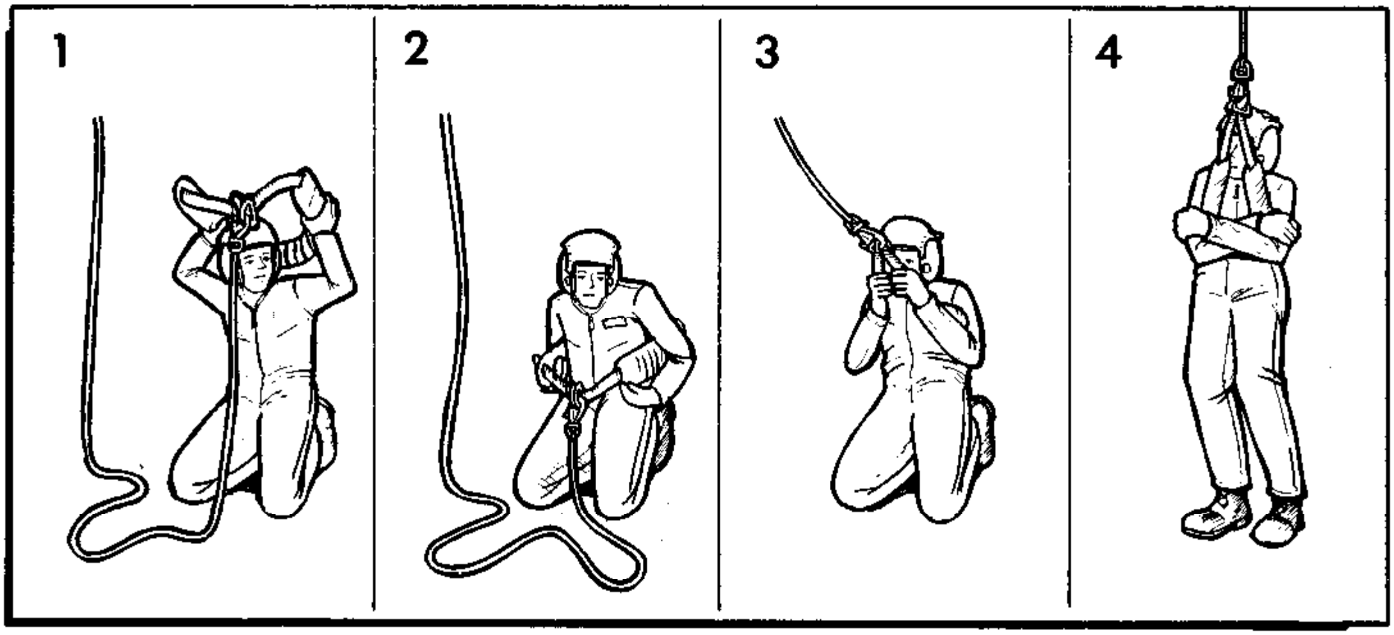

Figure 25-3. Horse Collar.

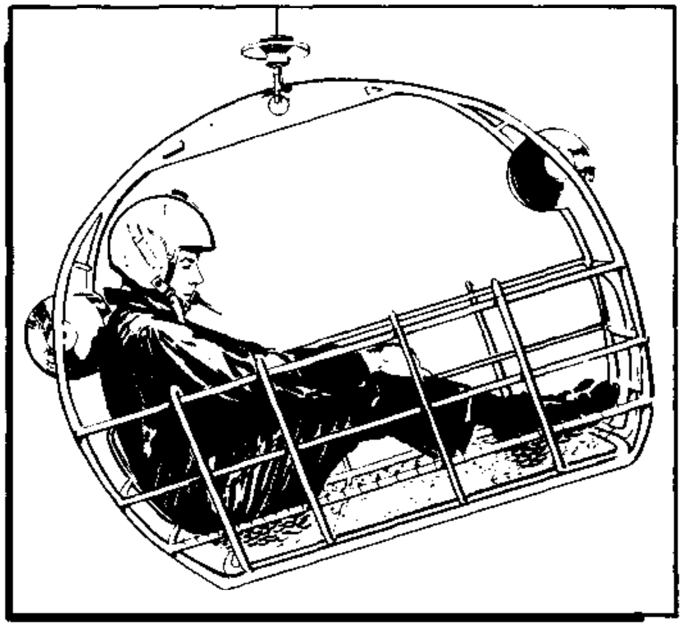

Figure 25-4. Basket.

which are spring-loaded in a folded position against the body or main shaft and must be pulled down to the locked position for use. On the main shaft of the tube, above the seats, there is a zippered fabric storage pouch for the safety (body) straps which are stowed when lowered to the survivor for a land pickup. The penetrator may also be equipped with a flotation collar. (NOTE: If the forest penetration is used for water pickup, it will be equipped with the flotation collar which enables the device to float with the upper one-third (approximately) of the device protruding above the water. Additionally, for water pickups, one strap will be removed from the stowed position, and one seat will be locked in the down position to assist the survivor in using the penetrator.)

(2) The safety strap is pulled from the storage pouch and placed around the body to hold the person on the penetrator seat. The strap should not be unhooked unless there is no other way to fasten it around the body. The survivor must make certain the safety strap does not become fouled in the hoist cable. After the strap is in place, the seat should be pulled down sharply to engage the hook which holds it in the extended position. The survivor can then place the seat between the legs. Then the survivor should pull the safety strap as tight as possible ensuring the device fits snugly against the body. The survivor must always keep the arms down, elbows locked against the body, and not attempt to grab the cable or weighted snap link above the device. After making certain the body is not entangled in the hoist cable, the signal to be lifted can be given (figure 25-5).

(3) In a combat area, under fire, survivors may be lifted out of the area with the cable suspended before

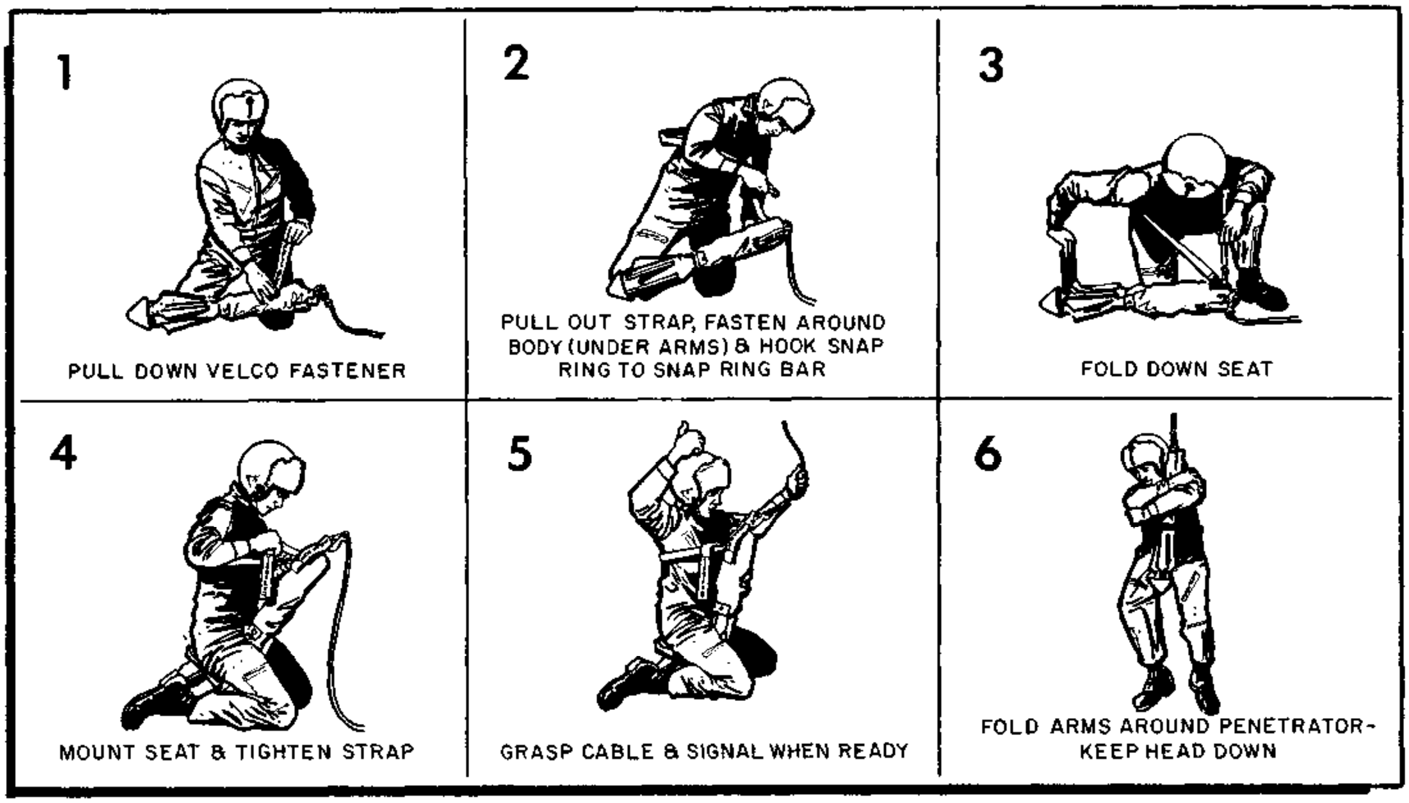
1
PULL DOWN VELCO FASTENER
2
PULL OUT STRAP, FASTEN AROUND BODY (UNDER ARMS) & HOOK SNAP RING TO SNAP RING BAR
3
FOLD DOWN SEAT
4
MOUNT SEAT & TIGHTEN STRAP
5
GRASP CABLE & SIGNAL WHEN READY
6
FOLD ARMS AROUND PENETRATOR-KEEP HEAD DOWN

Figure 25-5. Forest Penetrator.

being brought into the helicopter. It is important to be correctly and securely positioned on the pickup device. The seat should always be held tightly against the crotch to prevent injury when slack in the cable is taken up. The hands should be kept below and away from the swivel on the cable with the arms around the body of the penetrator. Survivors should keep their head close to the body of the penetrator so that tree branches or other obstructions will not come between the body and the hoist cable.

(4) When survivors reach a position level with the helicopter door, the hoist operator will turn them so they face away from the helicopter and then pull them inside. The crewmember will disconnect the survivors from the penetrator once the device is safely inside the helicopter.

(5) The forest penetrator is designed to lift as many as three persons. When two or three survivors are picked up, heads should be kept tucked in and each individual's safety strap drawn tight. The penetrator can be used to lower a paramedic or crewmember to assist injured personnel, and both (survivor and paramedic) can be hoisted to the helicopter. If the forest penetrator seat blades have been lowered in a tree area, and if for any reason the pickup cannot be made, the blades should be returned to the folded position to prevent possible hangup on tree limbs or other objects while the device is being retracted.

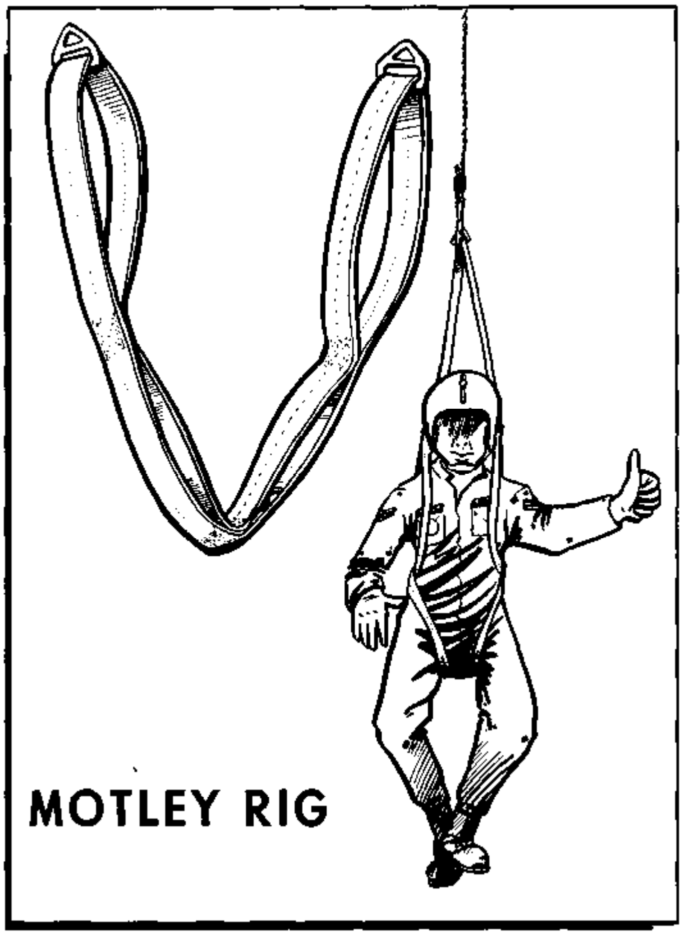

Figure 25-6. Motley Rig.

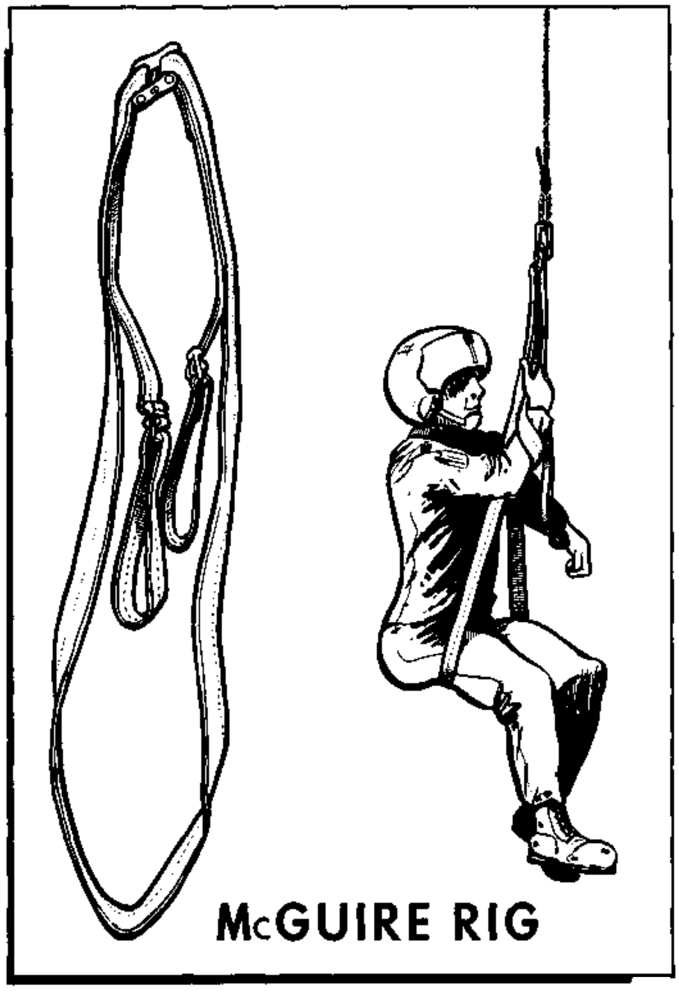

Figure 25-7. McGuire Rig.

(6) With all types of devices, it is necessary to watch the device as it is lowered. The devices weigh about 23 pounds. If the device were to hit a survivor, it could cause a serious injury or death.

f. Other Devices. There are other devices which could be used to pick up survivors. Some of them are the Motley and McGuire rigs (figures 25-6 and 25-7), the Swiss Seat and Stabo rig (figures 25-8 and 25-9), and the Rope Ladder (figure 25-10).

(1) Motley and McGuire Rigs. These devices may be carried by Army helicopters either designated as the recovery aircraft in assault or for use to insert or extract special ground forces. The device is normally packed in a weighted canvas container and dropped by rope. The device is dropped to the survivor, who is allowed time

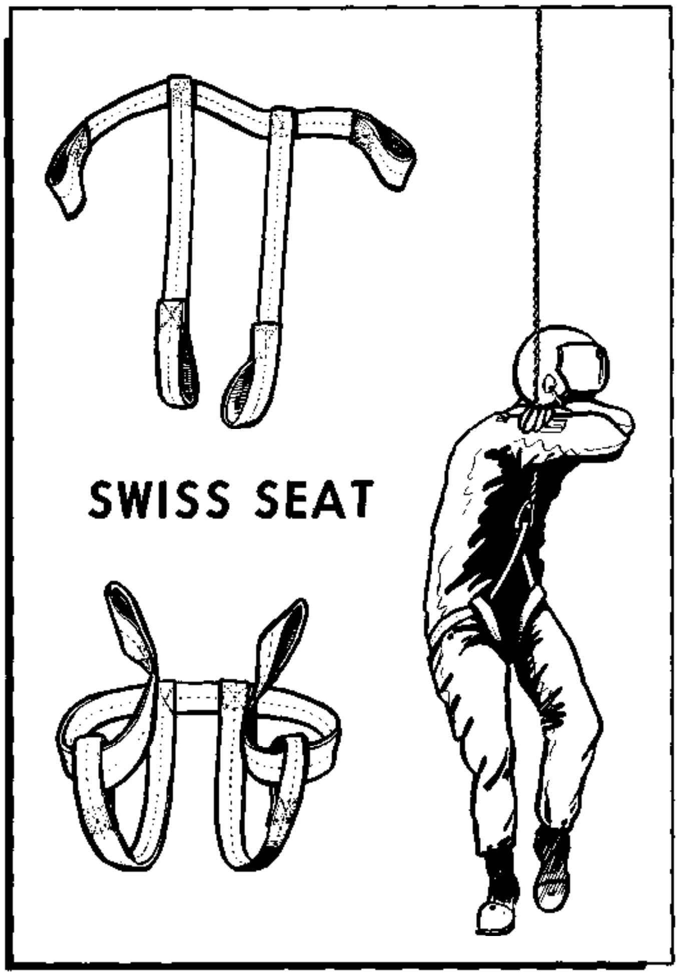

Figure 25-8. Swiss Seat.

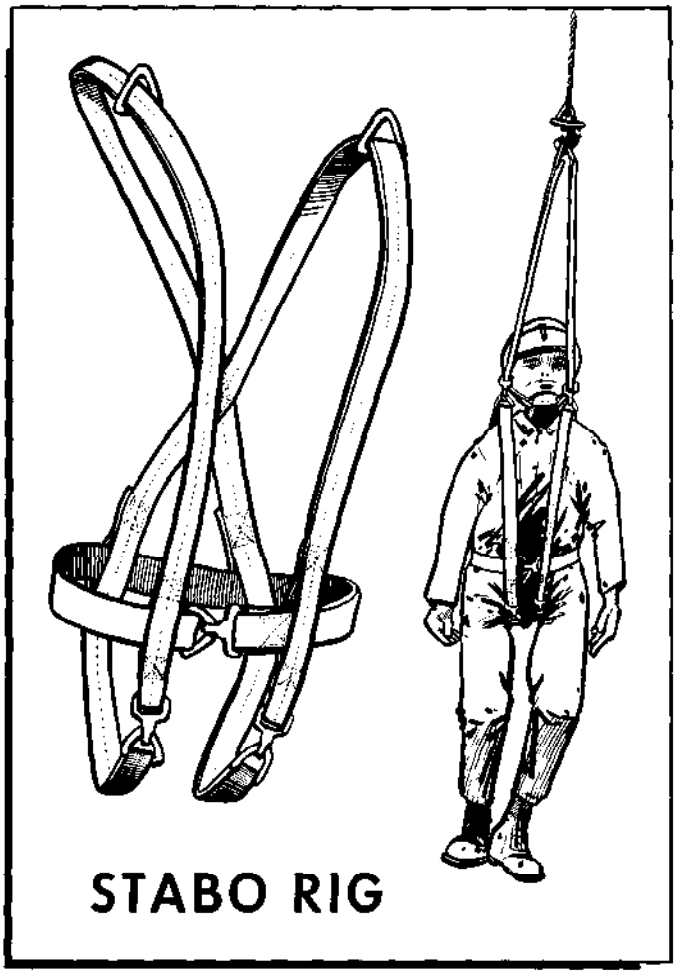

Figure 25-9. Stabo Rig.

for donning. The helicopter then returns trailing a rope which is then fastened to the device for pickup. Generally, the survivor is not hoisted into the helicopter; therefore, all safety straps should be securely fastened.

(2) Swiss Seat and Stabo Rig (figures 25-8 and 25-9). These devices are carried by special ground forces who may require instant extraction by helicopter. Special ground forces put their devices on and wait for the helicopter to drop ropes which are snapped into the devices for rapid extraction. Although not normally carried aboard the aircraft, the Army helicopter may supply one of these devices to the survivor. Again, the survivor would not be hoisted into the helicopter.

(3) Rope Ladder. This device is used primarily by the Army and special ground forces. If this device is used, it should be approached from the side and not the front. The survivor should climb up a few rungs, sit down on a rung, intertwine the body with rungs (figure 25-10). The survivor should not try to climb up the ladder and into the helicopter.

25-7. Preparations for Open Seas Recovery:

a. On sighting rescue craft approaching for pickup, (boat, ship, conventional aircraft, or helicopter), survivors must quickly clear any lines (fishing lines, desalting kit lines, etc.) or other gear which could cause entanglement during rescue. All loose items should be secured in the raft. Canopies and sails should be taken down to ensure a safer pickup. After all items are secure, the survivor should put on the helmet (if available). The life preserver should be fully inflated with the oral valve locking nut tight against the mouthpiece. Survivors should remain in the raft, unless otherwise instructed, and disengage all gear except the preservers. If possible, rescue personnel will be lowered into the water to assist survivors. The survivors should remember to follow all instructions given by rescue personnel.

b. If helicopter recovery is unassisted, the survivor will be expected to do the following before to pickup:

(1) Secure all loose equipment in raft, accessory bag, or in pockets.

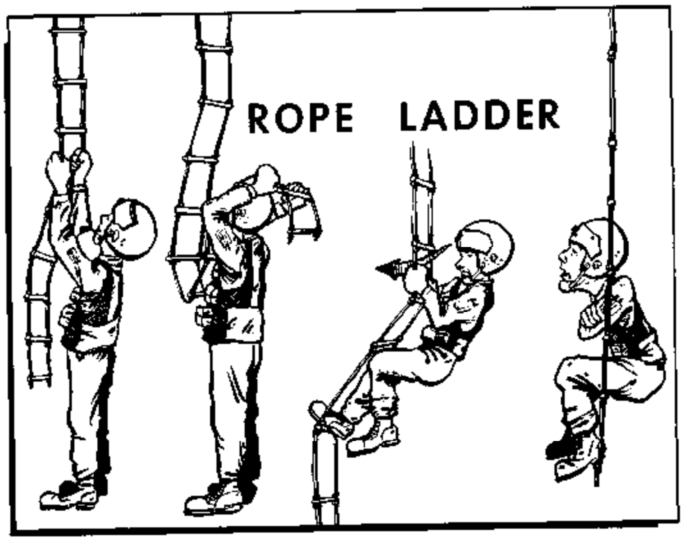

Figure 25-10. Rope Ladder.

(2) Deploy sea anchor, stability bags, and accessory bag.

(3) Partially deflate raft and fill with water.

(4) Unsnap survival kit container from parachute harness.

(5) Grasp raft handhold and roll out of raft.

(6) Allow recovery device and (or) cable to ground out on water surface.

(7) Maintain handhold until recovery device is in the other hand.

(8) Mount recovery device (avoid raft lanyard entanglement).

(9) Signal hoist operator for pickup.

Part Nine

EVASION

Chapter 26

LEGAL AND MORAL OBLIGATIONS

26-1. Introduction:

a. Aerial combat in the future, as in the past, will expose aircrews to possible ejection, bailout, or forced landings into enemy controlled territory. The aircrews who encounter such traumatic circumstances must be prepared to survive and evade the enemy to return to friendly control. An active commitment to solving problems and to individual survival (the "will to survive") is essential. Aircrew members must be prepared to exert extreme effort, both mental and physical, to successfully evade capture. In an excerpt from a debrief, one survivor describes what it was like coming to terms with these problems. "I thought to myself...'Well, I'm in a hell of a spot, what am I going to do about it?' The situation was such that I didn't do anything about it for quite awhile. I just sat there in the rain where I had landed and stared at the ground. It became colder, and after what seemed like hours, I lifted my gaze and sat staring now into the forest, looking at it without even really seeing it. I started thinking about the chain of events that had preceded my landing in this God-forsaken forest so far from my lines. My thoughts were as jumbled and unreal as the fog into which I had dived. Soon, it started to snow and I realized that I was in danger of being discovered or of freezing to death where I sat, and so, with a huge effort of will I forced myself to think coherently. I stood, and so began my journey." (See figure 26-1.)

b. This person had no survival training but nevertheless managed to successfully evade for miles back to his own lines with little more than a strong will to survive and common sense. Potential evaders must understand that evading capture presents a difficult challenge. They will meet and have to overcome a succession of obstacles, both manmade and natural. Knowledge gained from the experience of others, their own training, and prior preparation and planning will help them to overcome those obstacles.

c. This part addresses covert survival; that is, evasion. Areas to be covered include:

(1) Evader's moral obligations and legal status.

(2) Principles and techniques of evasion, camouflage, and travel (assisted or unassisted).

(3) The special aspects of food and water procurement.

(4) Combat signaling and recovery.

26-2. Definitions:

a. Evader. An "evader" (JCS Pub 1) is "a person who through training, preparation, and application of natural intelligence avoids contact with, and capture by, hostiles, both military and civilian."

b. Evasion. Evasion means all the processes involved in living off the land and, at the same time, avoiding capture while returning to friendly control. As used here, it includes all the techniques of evasion employed by those on foot in enemy territory.

26-3. Military Drive:

a. A crewmember becomes an evader when isolated in hostile areas, unable to continue the assigned mission, and when prevented from rejoining friendly forces. Definitions are useful but will not contribute to success in evading capture unless potential evaders understand what factors give direction and guidance to their efforts.

b. If the opportunity exists, evaders must be motivated to take advantage of it. The motivation to make a total effort to adhere to every evasion principle 24 hours a day may be personal or military. This strong central drive will give the survivor the necessary push to make these efforts.

c. Even if evasion is unsuccessful and the evader is captured, every hour spent eluding the enemy ties up enemy forces and lessens the evader's intelligence value.

d. In addition to the above reasons for evading capture, survivors also have moral and legal obligations to fulfull.

(1) Moral obligation is implied throughout the Articles of the Code of Conduct, specifically Article II. Article II states: "I will never surrender of my own free will. If in command, I will never surrender my men while they still have the means to resist." This Article of the Code should guide an evader's behavior during evasion just as it does in any other combat situation.

(2) The UCMJ continues to apply to evader's conduct during evasion or captivity. Particularly applicable are Article 99, Misbehavior Before the Enemy, Article 104, Aiding the Enemy, and Article 92, Failure to Obey a Lawful Order. Thus, one can be tried for misconduct as a combatant or as a noncombatant. A combatant is defined in AFP 110-31 as "a person who engages in hostile acts in an armed conflict on behalf of a party to the conflict." The combatant must conform to the standards established under international law for combatants, be authorized by his or her country to so act, and be recognizable as a combatant by uniform, insignia, or

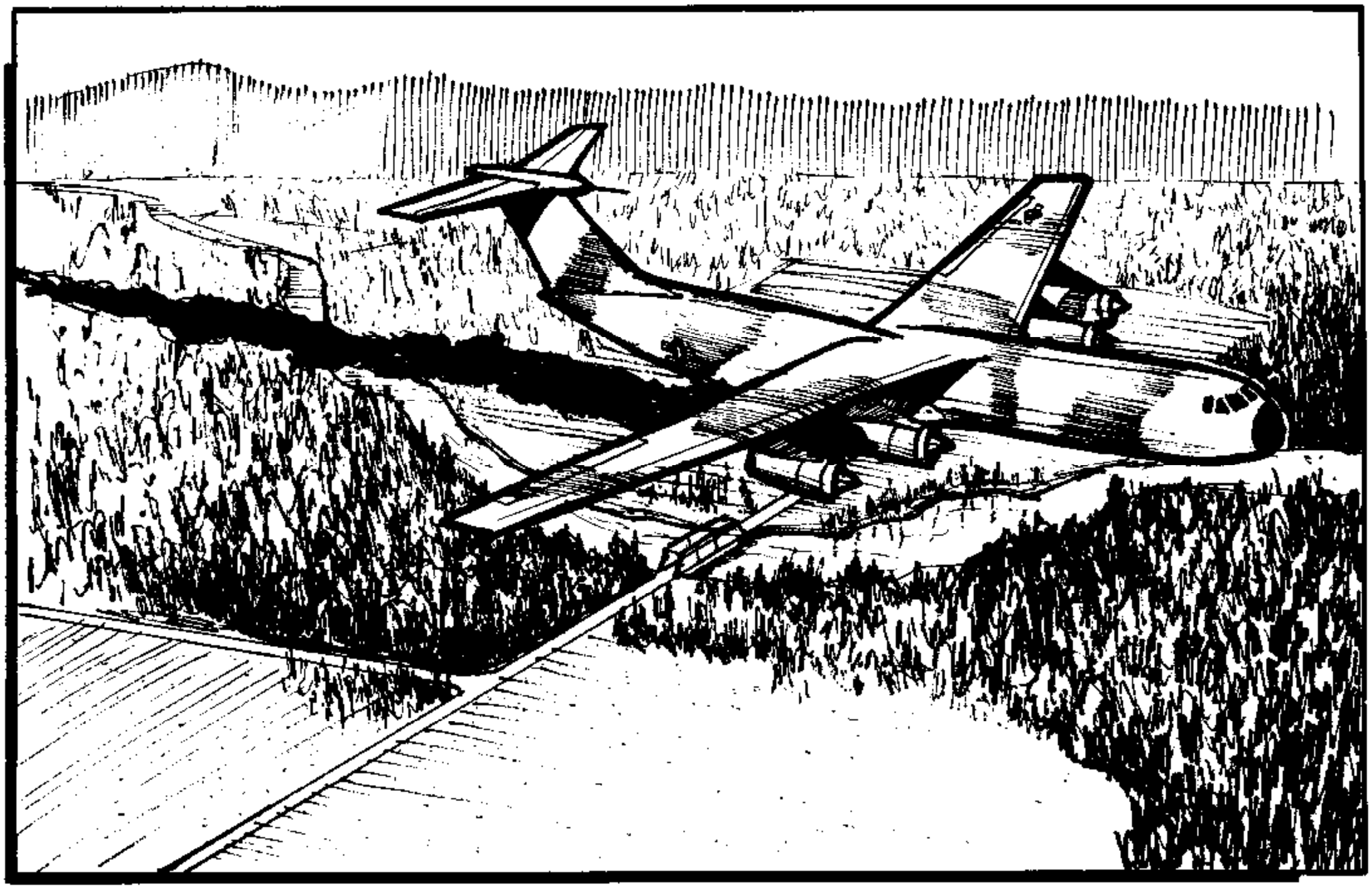

Figure 26-1. Evasion.

other sign. A noncombatant includes a wide variety of persons, including civilians, prisoners of war, sick and wounded persons, chaplains, medics, and other similar persons (AFP 110-31). Various countries around the world have developed written and unwritten laws of war. Four Geneva treaties were entered into by the United States and 60 other countries in 1949 and these treaties, as since amended, are in AFP 110-20.

(3) An "evader" is defined in international law as "any person who has become isolated in hostile or unfriendly territory and who eludes capture." An evader is a combatant and retains this status as a fighting person under arms according to international law until captured. Evaders are considered instruments of their government, under orders to evade capture, and never to surrender of their own free will. The evader is still militarily effective and may take such steps as necessary, within the rules of warfare, to accomplish the mission, which includes returning after striking an enemy target. While in combatant status, the evader may continue to strike legitimate military targets and enemy troops without being held liable to prosecution after capture for violation of the local criminal law. To do so while evading capture is the legal function of a combatant.

(4) Once captured, an evader becomes a noncombatant and will occupy the status of a prisoner of war (PW). A PW who kills or wounds enemy personnel of the Detaining Power in an attempt to escape or evade may be tried and punished for the offense. International law provides certain rights to a PW and requires issuance of an identity card showing name, rank, serial number, and date of birth. When questioned, a PW is bound only to provide this information. A PW who escapes remains a noncombatant until he or she rejoins the armed forces of his or her country or the armed forces of a friendly power. Once escape is completed, combatant status is regained and no punishment may be imposed by the Detaining Power in the event of subsequent recapture. A PW who attempts to escape and is recaptured before rejoining his or her armed forces is liable under the Geneva Convention only for disciplinary punishment in respect of this act, even if it is a repeated offense. However, special surveillance may be imposed.

(5) Disguise is a lawful means of evading enemy forces so long as a means of military identification is retained on the person. However, an evader while in disguise may not participate in or commit hostile acts

involving destruction of life or property. A person involved in such an offense will be classified as an "unlawful combatant," will not be entitled to PW status, and may be tried by the enemy and sentenced to imprisonment or execution under certain circumstances. A disguise also cannot be used by a military person for the purpose of gathering enemy military information or for waging war, and to do so will result in loss of PW status if captured.

(6) Evasion can be classified as either assisted or unassisted. Assistance can be defined as any help which is offered to an American aircrew member by any person. This help may include food, clothing, medicine, shelter, money, and even such a small item as a shoelace. Evaders should, in fact, consider they have been assisted even if, while evading through hostile territory, their presence has been ignored by indigenous personnel. The term unassisted evasion would then relate to the situation where the survivor, as an evader, must rely solely on his own knowledge and abilities to successfully emerge from an enemy-held or hostile area to areas under friendly control. The process of emerging may include aerial recovery, water recovery, or assisted evasion, but it is primarily an unassisted individual effort. Certain principles apply to all evasion situations; certain procedures should be followed; and certain techniques have widespread application. The beginning of an evasion experience is often the most critical phase. This is particularly true for a person who bails out over enemy territory during daylight hours and in sight of enemy personnel. The downed aircrew member can count on the enemy to make a determined effort to capture him if seen. Eluding the enemy is also a matter of effort and luck. Luck plays its part initially in establishing where a crewmember lands in relation to the location of any enemy personnel who may have seen the descent. For example, the downed crewmember who lands in a heavily populated area (city, military installation or combat area) may be taken prisoner immediately. Here the problem becomes one of early escape rather than one of evasion.

(7) An evasion situation should not be categorized in terms of length. History has proven that predicting the length of any specific evasion situation is practically impossible. All crewmembers should be prepared to evade until rescued, no matter how long the evasion experience might last. Emphasizing the advantages of "short-term" evasion over "long-term" may cause an overly optimistic, possibly even foolhardy, attitude toward evasion planning. Or, evaders may decide, if they are not rescured in a "short" period of time, it is no longer worth the effort; thereby, taking on a defeatist attitude.

Chapter 27

FACTORS OF SUCCESSFUL EVASION

27-1. Basic Principles. All potential evaders must have three things in their favor. These are the same three things needed by a potential escapee. The three factors which increase chances of successful evasion are preparation, opportunity, and motivation.

27-2. Preparation:

a. Preparation is one of the most important factors for successful evasion. The actions crewmembers take before the evasion episode can make the difference between being able to evade or being captured. In a hostile area the survivors should remember evasion is an integral part of their mission and plan accordingly. The enemy may make mistakes of every conceivable form and not suffer more than indignation, anger, and fatigue. The evader, on the other hand, must constantly guard against mistakes of any sort. Being seen is the greatest mistake an evader can make. The evader must prepare for this task (figure 27-1).

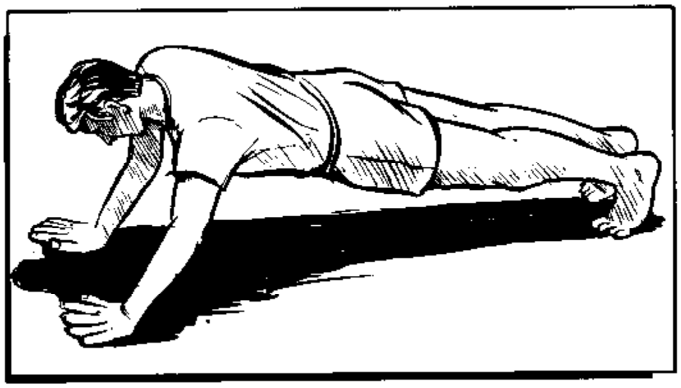

Figure 27-1. Preparation.

b. Three basic problems during evasion are:

(1) Evading the enemy.

(2) Surviving.

(3) Returning to friendly control.

c. Chances for a successful evasion are improved if evaders:

(1) Observe the elementary rules of movement, camouflage, and concealment.

(2) Have a definite plan of action.

(3) Be patient, especially while traveling. Hurrying increases fatigue and decreases alertness. Patience, preparation, and determination are key words in evasion.

(4) Conserve food.

(5) Conserve as much strength as possible for critical periods.

(6) Rest and sleep as much as possible.

(7) Maintain a highly developed "will to survive" and "can do" attitude. Evasion may require living off the land for extended periods of time and traveling on foot over difficult terrain, often during inclement weather.

(8) Study the physical features of the land. They should note the location of mountains, swamps, plains, deserts, or forests, type of vegetation, and availability of water.

(9) Consider the climate. Aircrew members should know the climatic characteristics and typical weather conditions of the area which may be flown over.

(10) Study ethnic briefs and survival, evasion, resistance, and escape (SERE) contingency guides before a mission and learn some of the customs and habits of the local people. Such knowledge will aid in planning missions and evasion plans of action. For example, it may give the evader the ability to avoid hostile people or groups or to identify and deal with "friendlies." This knowledge may also allow for blending into the local populace (figure 27-2).

(11) Know the equipment well! One must know the location of each item in the kit, its operation, and its value. An evader must preplan which equipment should be retained and which should be left behind.

d. Once in the evasion situation, planning for travel will be a consideration for evaders. They must have a definite objective and be confident in their approach and ability to achieve it. They will normally have several options with variations to choose from in selecting a plan of action or destination. The enemy force deployment, search procedures, terrain, population distribution, climate, distance, and environment (that is, NBC) will influence destination selection. Examples of options and destinations:

(1) Await SAR forces.

(2) Evade to a SAFE area.

(3) Evade to a neutral country. (NOTE: Border areas not disrupted by combat may have a security system intact.)

(4) If evaders are in the forward edge of the battle area (FEBA) and feel sure that friendly forces are moving in their direction, they should seek concealment and allow the FEBA to overrun their position. Evaders' attempts at penetrating the FEBA should be avoided. Evaders face stiff opposition from both sides.

e. The chances that one of these destinations may be close by will be determined by many things including the time and location of the bailout. Other determining factors are: the location and direction of movement of the FEBA, the presence or absence of willing assisters, and the knowledge of the evader's whereabouts possessed by rescue personnel. If the survivor does not land

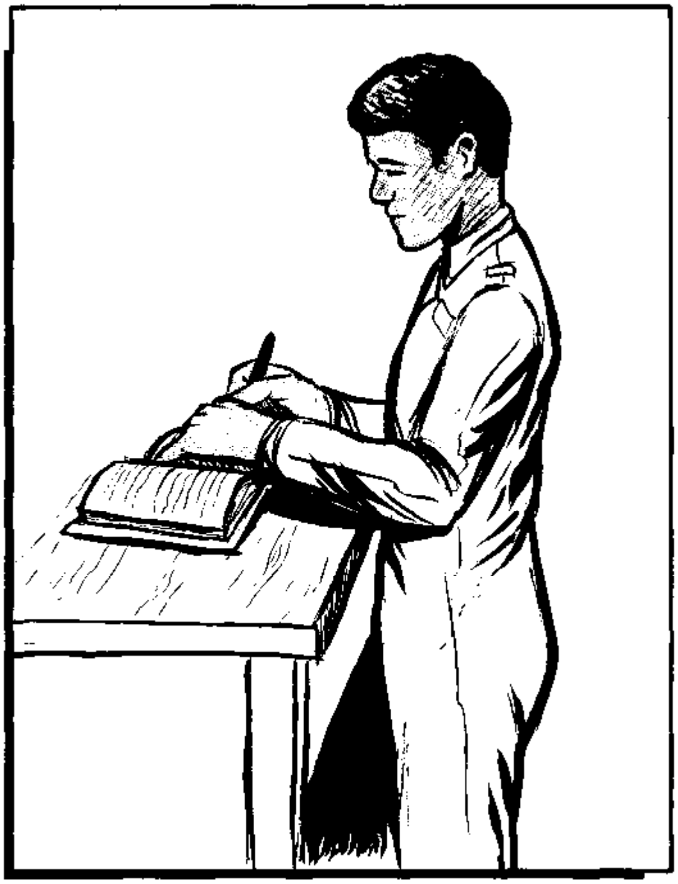

Figure 27-2. Study Ethnic Briefs.

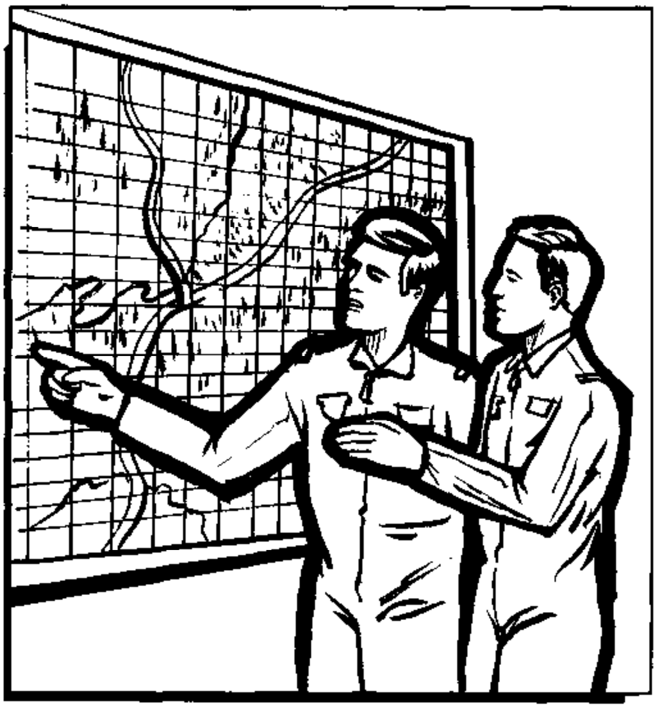

Figure 27-3. Planning for Travel.

close to one of the above areas or if the previously mentioned factors do not favor immediate air pickup, the survivor may have to travel some distance to reach one of the destinations (figure 27-3).

f. One consideration in choosing destination and direction of travel after bailout is whether one of these suitable areas for a pickup or contact with friendly forces exists and, if so, its location. Some preplanning should have been done before the mission. Information upon which to base a decision is derived from command area briefings, area studies, SERE contingency guides, and premission intelligence briefings.

g. Another consideration is physical condition. One's physical condition is the responsibility of the potential evader and has a great effect on the evader's ability to survive. Once on the ground, it is too late to get in shape. One more aspect is an aircrew member's personal habits. Upon first consideration, personal grooming habits might not be considered an important premission briefing item. However, using aftershave lotion, hair dressing, or cologne could add to the problems of an evader. The odor can carry for great distances and give away the evader's presence.

27-3. Opportunity. Potential evaders must take advantage of any and all opportunities to evade. This starts in the aircraft when an emergency is declared. Following current, approved emergency in-flight procedures for the theater of operations (when ejection, bailout, or ditching appears imminent), the aircraft commander will attempt to establish radio contact by first calling on the secure frequency of last contact; second, on an established common secure frequency; and third, on the international emergency frequency. When communication is established, the tactical call sign, type of aircraft, position, course, speed, altitude, nature of difficulties, and intentions will be transmitted. The identification friend or foe (IFF) should be set to the emergency position. When possible, ejection or bailout should be attempted over or near a SAFE area, lifeguard station, or submarine pickup point. This minimizes threat involvement for evaders and SAR forces alike. After ejection or bailout and during descent, the aircrew member must remain alert and steer the parachute away from potential threats (populated areas, gun emplacements, troop concentrations, etc.) or out to sea (feet wet). Once on the ground, the evader must be proficient in the use of the survival/evasion equipment to facilitate evasion (for example, use of the compass in conjunction with the survival radio to call in airstrikes on enemy forces threatening the evader). In addition to the opportunity to evade, motivation is essential to the evader's success.

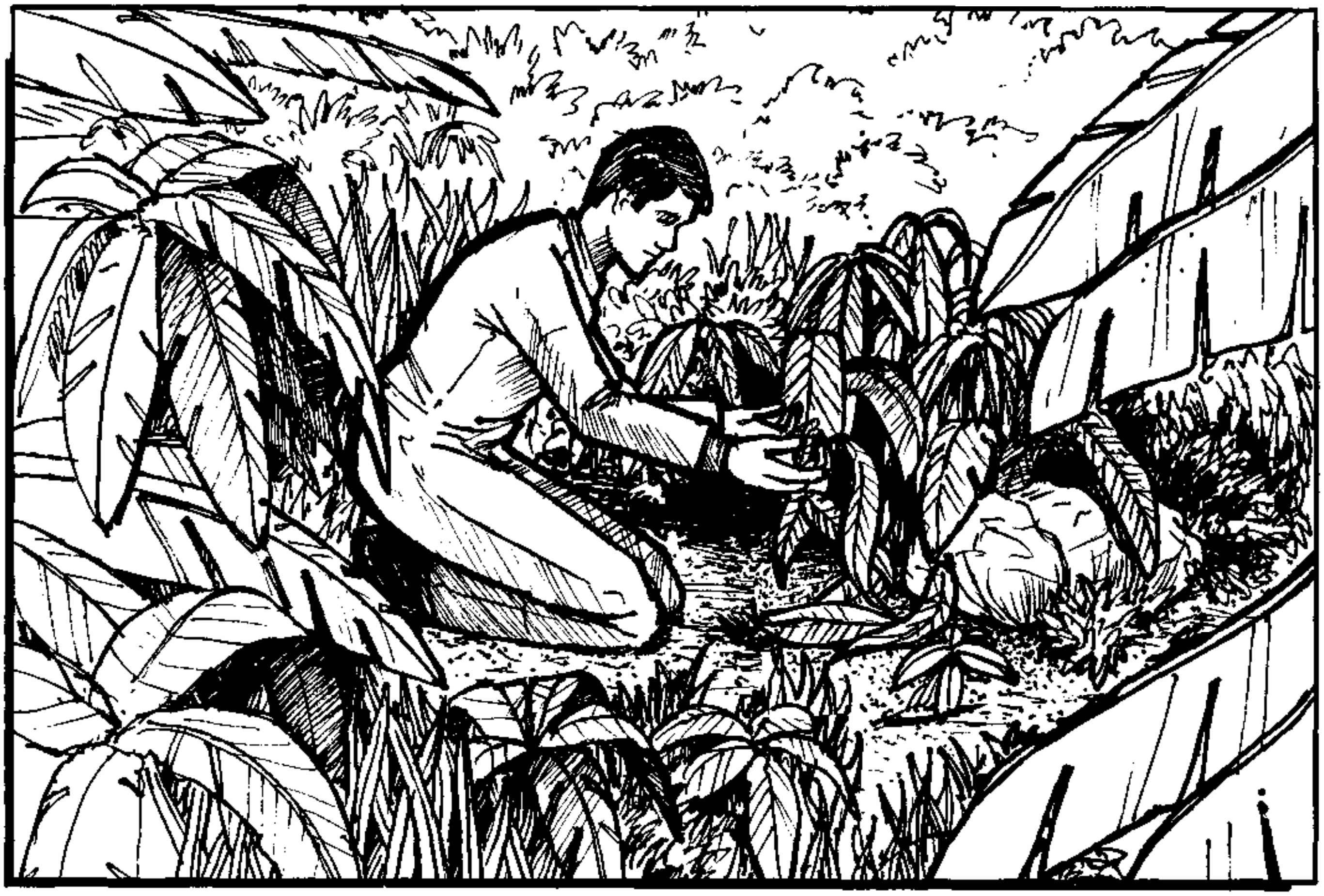

Figure 27-4. Avoiding Detection.

27-4. Motivation. A strong, central drive will give the evader the necessary push to succeed. It may be personal, ranging in nature from a frame of reference gained through training to a desire to return to a family or loved ones. Motivation may be strictly military, involving one or all of the following reasons applicable to all military men:

a. To return and fight again.

b. To deny the enemy a source of military information.

c. To deny the enemy a source of propaganda.

d. To deny the enemy a source of forced labor.

e. To tie up enemy forces, transportation, and communications that otherwise might be committed to the war effort.

f. Return with intelligence information.

g. In addition, the Code of Conduct calls upon the military members not to surrender of their own free will.

h. Additionally, it is suggested that other personal reasons for being motivated to evade include: fear of death, pain, suffering, humiliation, degradation, disease, illness, torture, uncertainty, and fear of the unknown. From the evader's point of view, *evasion is far more desirable than captivity or death.*

27-5. Evasion Principles:

a. Besides the preparation, opportunity, and motivation factors important to evasion, there are other important principles. The evader should try to recall any previous briefings, standard operating procedures, or training. A course of action should then be chosen which has the greatest likelihood of resulting in the return to friendly forces.

b. Evader actions should be flexible. Flexibility is one of the most important keys to successful evasion. The evader, basically, must never be so firmly set in a course of action that a change is out of the question. The best thing an evader can do is to stay open to new ideas, suggestions, and changes of events. Having several backup plans of action can give the evader organized flexibility. If one plan of action is upset by enemy activity, the evader could rapidly switch to a backup plan without panic.

c. The evader is primarily interested in avoiding detection. Each evader should remember that people catch people. If the evader avoids detection, success is almost assured. Evaders should:

(1) Observe and listen for sounds of enemy fire and vehicle activity during parachute descent and moving away from those enemy positions once on the ground. Flyers downed during daylight should assume they were seen during descent and expect a search to center on their likely point of landing.

(2) Be patient and determined while traveling.

(3) Use poor weather conditions as an aid in evading.

(4) Circumstances permitting, select times, routes, and methods of travel to avoid detection.

(5) Avoid lines of communication (waterways, roads, etc. (figure 27-4)).

d. The evaders' main objective is immediate recovery. In hostile areas or situations, survivors must sanitize all evidence of presence and direction of travel (figure 27-5.) Survivors may never be certain that rescue is imminent.

e. Although evaders would not normally move too far if rescue is imminent, in many situations they will have to leave the landing area quickly and travel as far as practical before selecting a hiding place. They should leave no sign which indicates the direction or presence of travel. All hiding places should be chosen with extreme care. The time evaders will remain in the first location is governed by enemy activity in the area, their physical condition, availability of water and food, rescue capabilities, and patience. It is in this place of initial concealment that the evaders should regain strength, examine the current situation, and plan for the evasion problems ahead (figure 27-6).

f. Once in a place of concealment, evaders should make use of all available navigation aids to orient themselves. After finding their location, evaders should also select an ultimate destination and any necessary alternate destinations. The best possible route of travel should then be decided upon. When the time comes to move, they should have a primary plan and alternate plans for travel that cover eventualities they may encounter.

(1) Evasion in a forward area has one great advantage which is not present further to the rear. Assistance may be close at hand. This assistance may come from serveral sources, each of which, under particular circumstances, may prove to be the most effective. These sources may be air cover by tactical fighter flights, helicopter recovery, and rescue by ground forces. Contact with friendly forces in forward areas requires extreme caution. Do not surprise them or move suddenly. They may mistake the evader as the enemy.

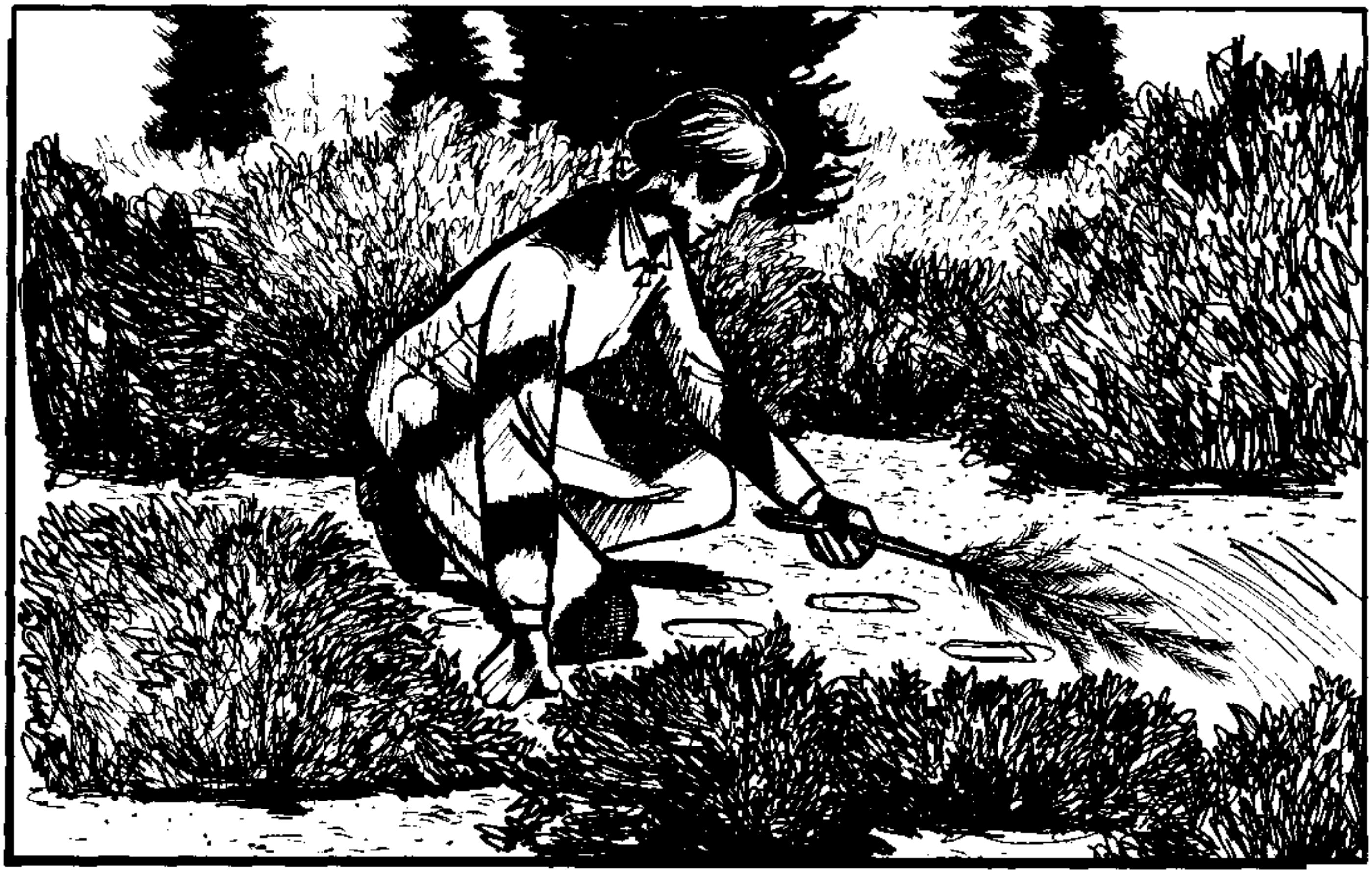

Figure 27-5. Sanitize Area.

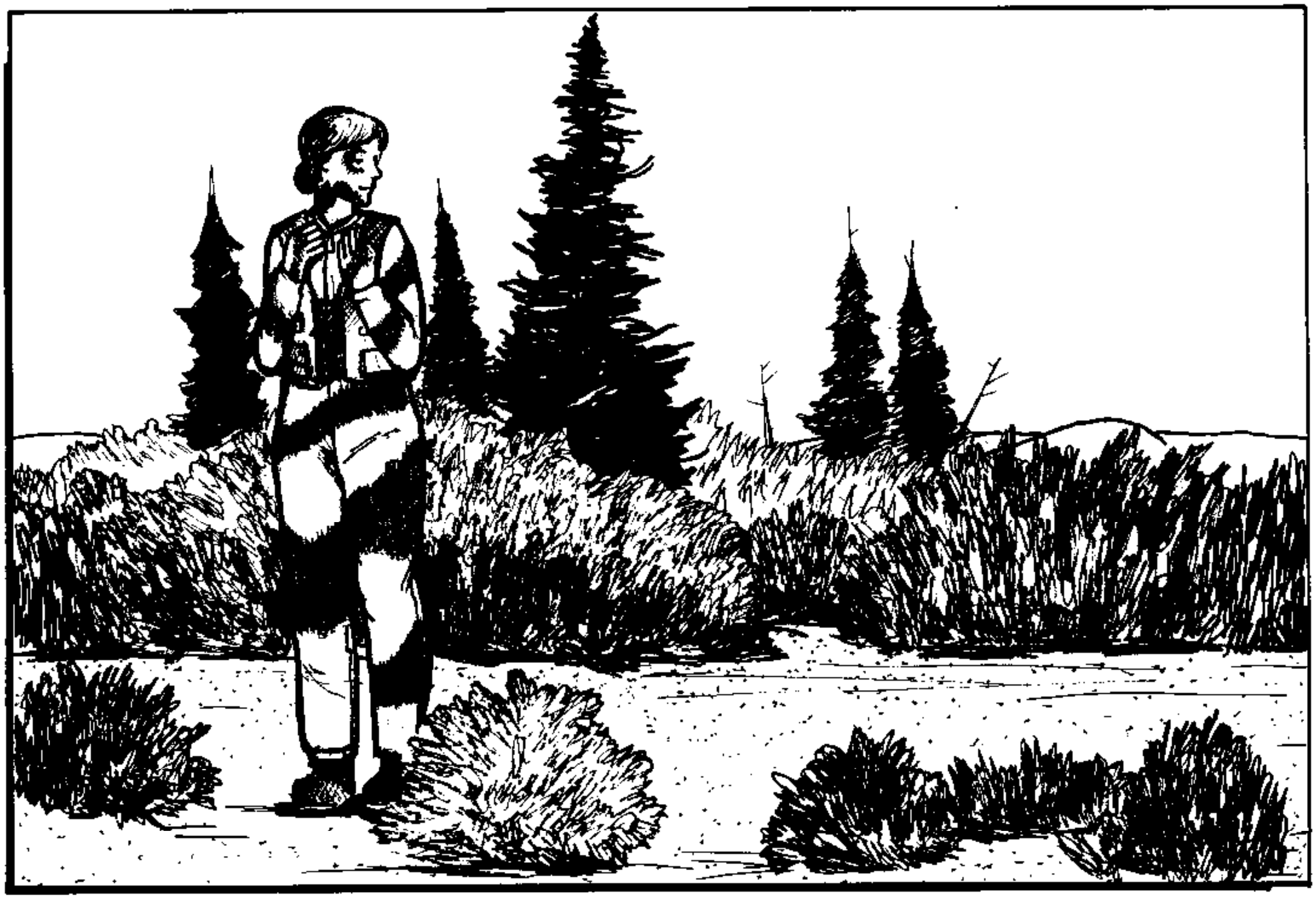

Figure 27-6. Movement from the Area.

(2) The situation at the time of the emergency will determine the evaders' best course of action. High ground is normally the best position from which to await rescue; evaders may expect the best results from signaling devices, may observe the surrounding terrain, and may be kept under observation by friendly air cover. Whatever position is chosen, it must be clear of obstacles that would prevent a successful rescue.

(3) If not rescued immediately, the situation may compel evaders to move. Evaders must plan a course of action before leaving their position. When the evaders are certain their position is known to friendly elements, they might expect ground forces to attempt a rescue. They should remember their position may be detected as the enemy search parties approach. They must be prepared to evade to a new position.

(4) Evaders should remember that when traveling they are probably more vulnerable to capture. Once past the danger of an immediate search, evaders must avoid people. Inhabited areas should be bypassed rather than penetrated, even if it means miles of added travel. Many evaders have been captured because they followed the easiest and shortest route, or failed to employ simple techniques such as scouting, patrolling, camouflage, and concealment. As a rule, the safest route avoids major roads and populated areas, even if it takes more time and energy. Unaccompanied evasion requires self-reliance and independent action (figure 27-7).

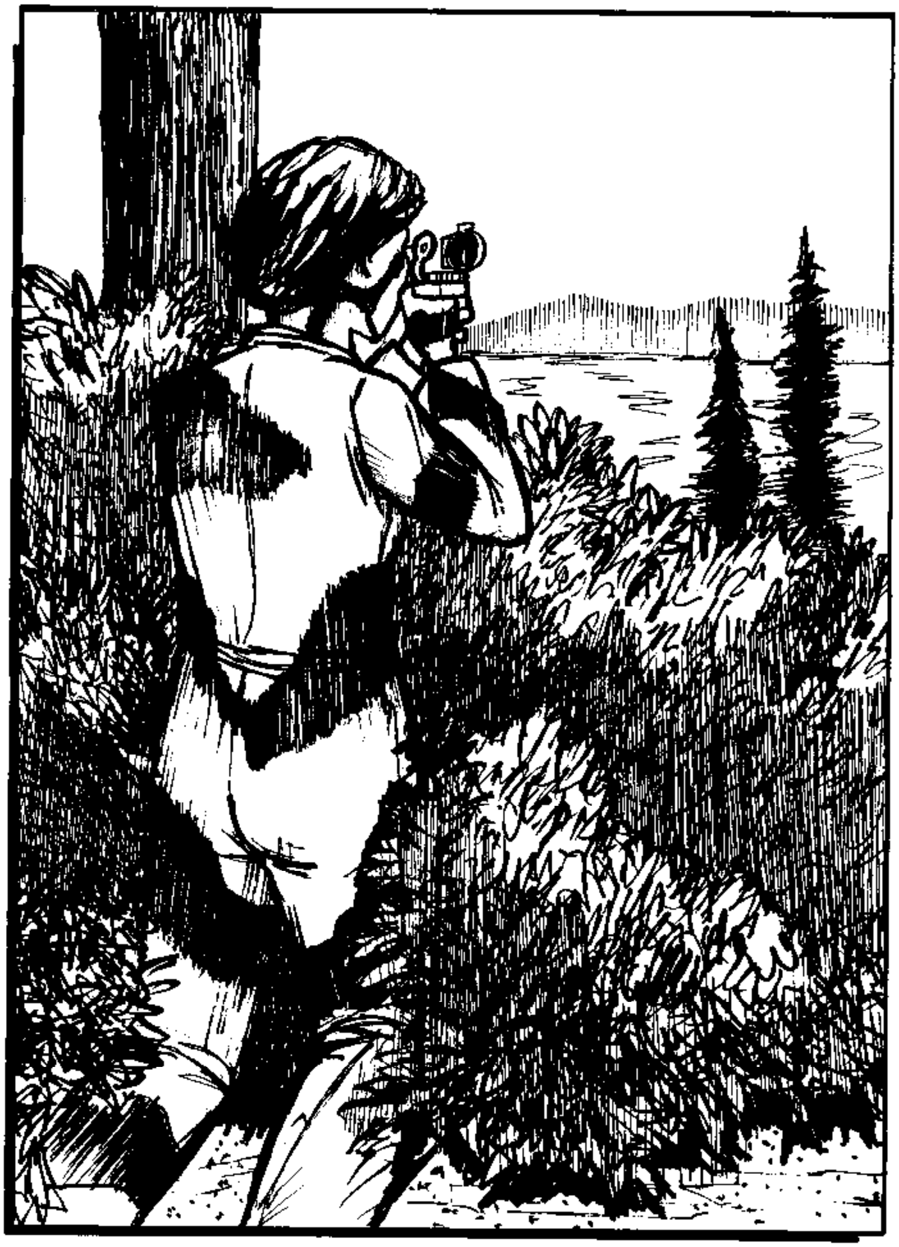

Figure 27-7. Avoiding Populated Areas.

Chapter 28

CAMOUFLAGE

28-1. Introduction. Presence of evaders in an area controlled by the enemy may require the evaders to adopt and maintain camouflage to avoid observation. Camouflage consists of those measures evaders use to conceal their presence from the enemy. Camouflage is a French word meaning disguise, and it is used to describe action taken to mislead the enemy by misrepresenting the true identity of an installation, an activity, an item of equipment, or an evader. As a tool for evasion, it enables evaders to carry out life supporting activities and to travel unseen, undetected, and free to return to friendly control. Camouflage allows them to see without being seen. They should try to blend in with the surrounding environment. Effective individual concealment often depends primarily on the choice of background and its proper use. Background is that portion of the surroundings against which an evader will be seen from the ground and the air. It may consist of a barren rocky desert, a farm yard, or a city street. It is the controlling element in individual camouflage and governs every concealment measure. At all times, camouflage is the responsibility of the individual evader. In the event of group evasion, the group leader and each individual are responsible for the camouflage of the group. Evaders should remember that camouflage is a continuous, never-ending process if they want to protect themselves from enemy observation and capture (figure 28-1).

28-2. Types of Observation:

a. Of the five senses, sight is by far the most useful to the enemy, hearing is second, while smell is of only occasional importance. But these same senses can be of equal value to the evader and observer.

b. How useful these senses are depends primarily on range. For this reason, basic camouflage stresses visual concealment which is relatively long range. Most people are accustomed to looking from one position on the ground to another position on the ground.

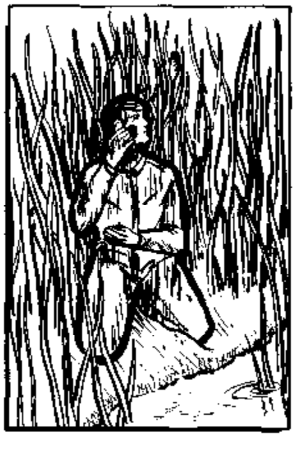

Figure 28-1. Camouflaging.

c. Before evaders can conceal themselves from aerial observation, they should become familiar with what their activities look like from the air, both in an aerial photograph and from direct observation. The evaders must also have an understanding of the types of observation used by the enemy. There are two categories of observation—direct and indirect.

(1) Direct Observation:

(a) Direct observation refers to the process whereby the observer looks directly at the object itself without the use of telescopes, field glasses, or sniperscopes. Direct observation may be made from the ground or from the air. Direct aerial observation becomes more and more important because of the rapid changes in weapons and in tactical situations due to greater mobility of troops. Reconnaissance aircraft over enemy lines report locations of troops, vehicles, and installations (or shelter areas) as seen from the air-to-ground control stations. Reported targets can be immediately fired upon or troops can be sent in to investigate shelter areas or other suspicious areas (figure 28-2).

(b) The enemy may also use dogs, foot patrols, and mechanized units to patrol a given area. Such teams could physically search an area for signs of the passage of strangers, such as footprints, old campfires, discarded or lost equipment, and other "telltale" signs which would indicate that someone had been in the area.

(c) Observation by the local populace is also a possibility. Upon seeing an evader or "telltale" signs an

Figure 28-2. Direct Observation.

evader left behind, they may contact the local authorities, who initiate organized searches.

(2) Indirect Observation:

(a) Indirect observation refers to the study of a photograph or an image of the subject via photography, radar, or television. This form of observation is becoming increasingly more varied and widespread, and may be used from either manned or unmanned positions.

(b) Views from the ground are familiar, but views from the air are usually quite unfamiliar. In modern warfare, the enemy may put emphasis on aerial photographs for information. It is important to become familiar with the "bird's-eye view" of the terrain as well as the ground view in order to learn how to guard against both kinds of observation.

28-3. Comparison of Direct and Indirect Observation:

a. The main advantage of direct observation is that observers see movement of an evader without camouflage. An observation can be maintained over relatively long periods of time. The main disadvantage lies in human frailty. For example, the observer's attention may be diverted to another area, or the observer may be fatigued and unable to concentrate.

b. Indirect observation has many advantages. Indirect observation can be far-reaching, cover large areas, and be very accurate. It also produces a record of the area observed so that the recorded picture can be studied in detail, compared, and evaluated. The principal disadvantage is a photograph covers a very short period of time, making detection of movement difficult. This disadvantage can be partially overcome by taking pictures of the same area at different intervals and comparing them for changes (figure 28-3).

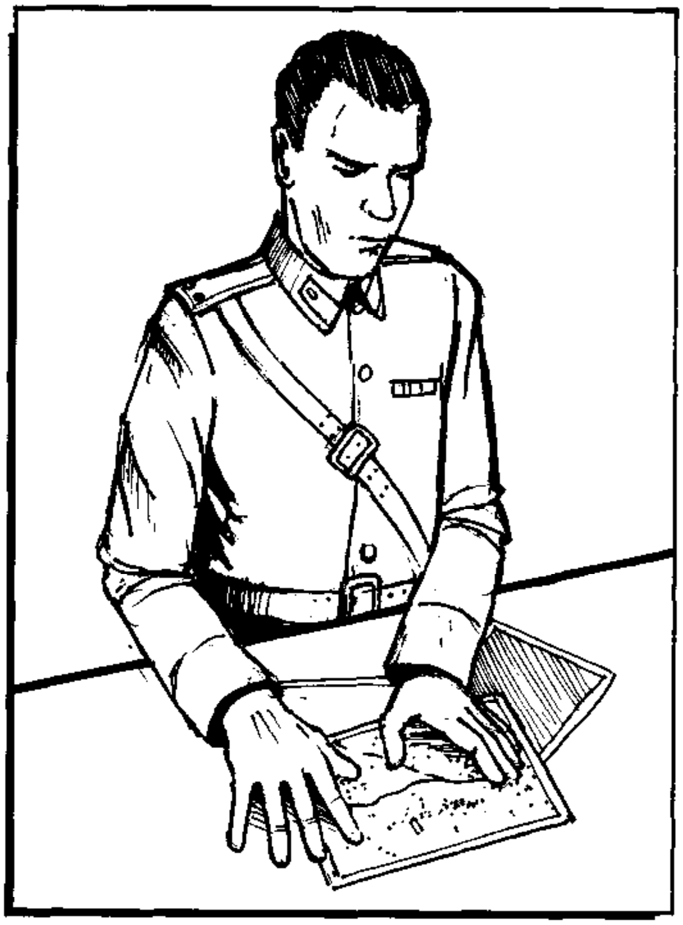

Figure 28-3. Indirect Observations.

28-4. Preventing Recognition:

a. Recognition is the determination (through appearance, behavior, or movement of the hostile or friendly nature) of objects or persons. One objective of camouflage concealment is to prevent recognition. Another objective is to deceive or induce false recognition. This implies that camouflage is not always designed to be a "cloak of invisibility." In some instances, camouflage is used to allow deception. The camouflaged object or person is then seen as a natural feature of the landscape.

b. Recognition through appearance is the result of conclusions drawn by the observer from the position, shape, shadow, texture, or color of the objects or persons. Recognition through behavior or movement includes deductions made from the actual movements themselves or from the record left by tracks of persons or vehicles or by other violations of camouflage discipline. Camouflage disciplines are those actions which contribute to an evader's ability to remain undetected. Proper use of camouflage discipline avoids any activity that changes an area or reveals objects to an enemy. Examples of common breaches of camouflage discipline include reflections from brightly shining objects (watches, glasses, rings, etc.) (figure 28-4), overcamouflaging, or using camouflage materials which are foreign to the area presently occupied by an evader. Evaders must also watch for signs that may reveal enemy camouflage efforts. Inadvertently walking into a camouflaged enemy position may result in capture.

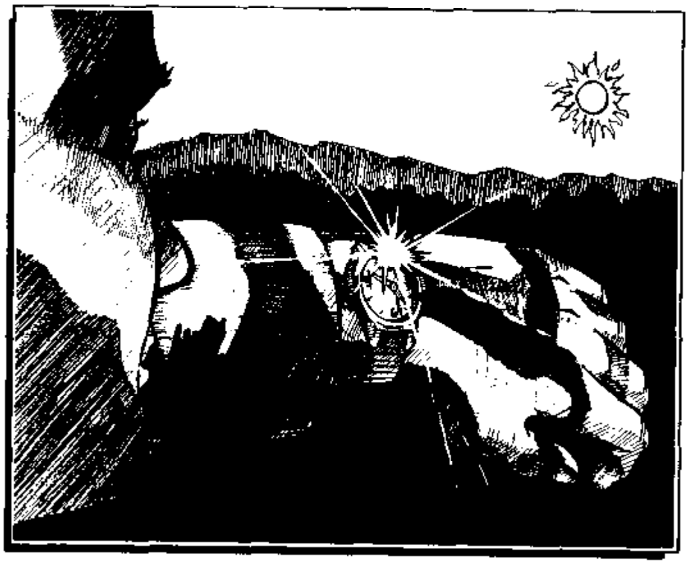

Figure 28-4. Reflections.

28-5. Factors of Recognition. Regardless of the type of observation, there are certain factors which help to identify an object. They are called the factors of recognition and are the elements which determine how quickly an object will be seen or how long it will remain unobserved. The eight factors of recognition are position, shape, shadow, texture, color, tone, movement, and shine. These factors must be considered in camouflage to ensure that one or more of these factors do not reveal the location of the evaders.

a. Position. Position is the relation of an object or person to its background. When choosing a position for concealment, a background should be chosen which will virtually absorb the evader (figure 28-5).

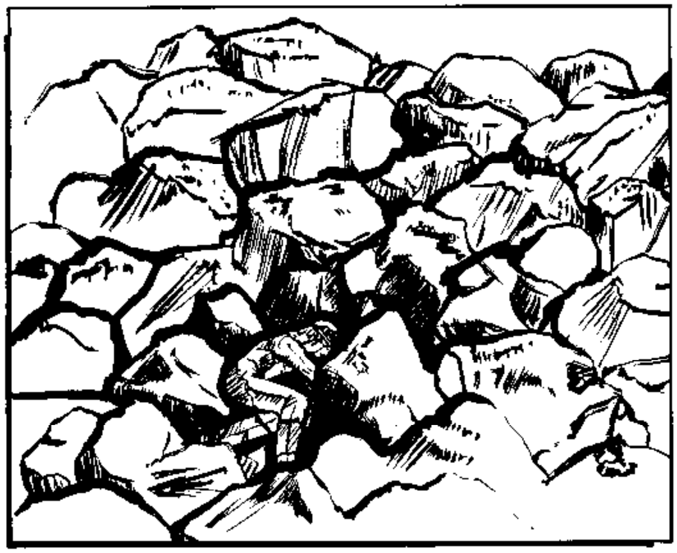

Figure 28-5. Position.

b. Shape. Shape is the outward or visible form of an object or person as distinguished from its surface characteristics and color. Shape refers to outline or form. Color or texture is not considered. At a distance, the forms or outlines of objects can be recognized before the observer can make out details in their appearance. For this reason, camouflage should disrupt the normal shape of an object or person (figure 28-6).

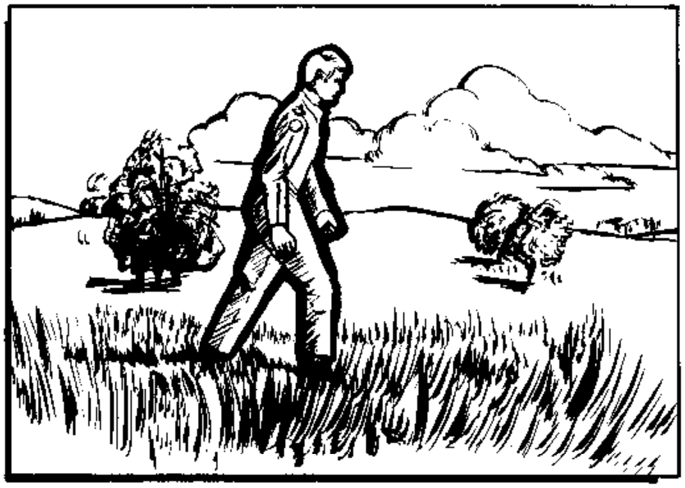

Figure 28-6. Shape.

c. Shadow. A shadow may be more revealing than the object itself, especially when seen from the air. Objects such as factory chimneys, utility poles, vehicles, and tents (or people) have distinctive shadows. Conversely, shadows may sometimes assist in concealment. Objects in the shadow of another object are more likely to be overlooked. As with shape, it is more important to disrupt the shadow pattern than to totally conceal the object or person. The identifiable shadows can be broken up by the addition of natural vegetation at various points on the body. Wearing "shapeless" garments will also disrupt the outline. For example, a soft and shapeless field cap can be used instead of a helmet or flight cap (figures 28-7 and 28-8).

d. Texture. Texture is a term used to describe the relative characteristics of a surface, whether that surface is a part of an object or an area of terrain. Texture affects the tone and apparent coloration of things because of its absorption and scattering of light. Highly textured surfaces tend to appear dark and remain constant in tone regardless of the direction of view and lighting, whereas relatively smooth surfaces change from dark to light with a change in direction of viewing or lighting. The application of texture to an object often has the added quality of disrupting its shape and the shape of its shadow, making it more difficult to detect and identify as something foreign to the surroundings in which it exists. As an example, a surface having the same color but with heavy "nap" or texture is tall grass. Each separate blade is capable of casting a shadow upon itself and its surroundings. The light reflecting properties have been cut to a minimum. It will look and photograph dark gray. Looking straight down, the aerial observer sees all of the shadows, whereas a person on the ground may not. The textured surface may look light at ground level, but to the aerial observer the same surface produces an effect of relative darkness. The material used to conceal a person or an object must approximate the texture of the terrain in order to blend in with the terrain. Personnel walking or vehicles moving across the terrain will change the texture by mashing down the growth. Therefore, this will show up clearly from the air as vehicle tracks or foot paths.

e. Color:

(1) Pronounced color differences at close range distinguish one object from another. The contrast between the color of the object and the color of its background

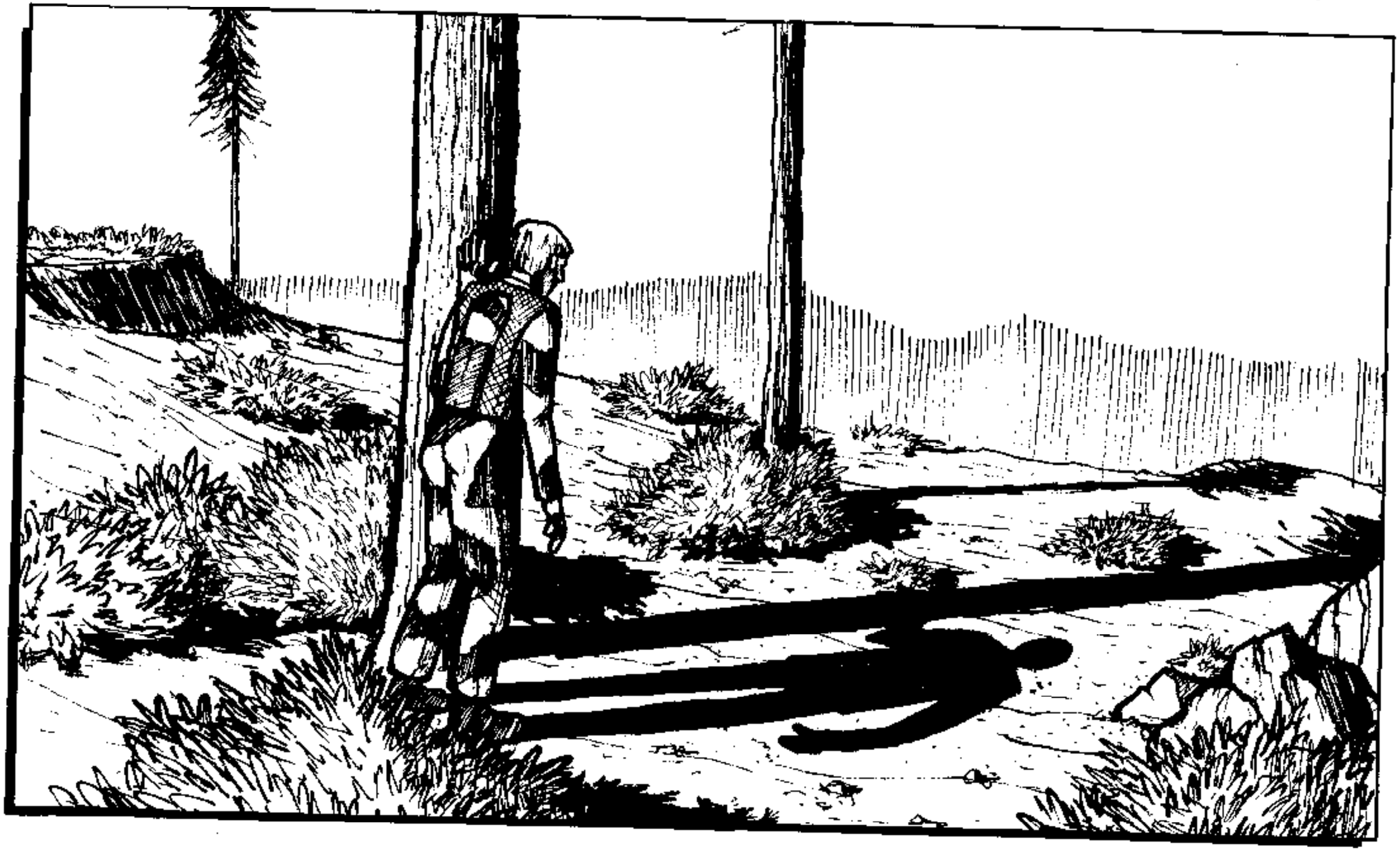

Figure 28-7. Shadow.

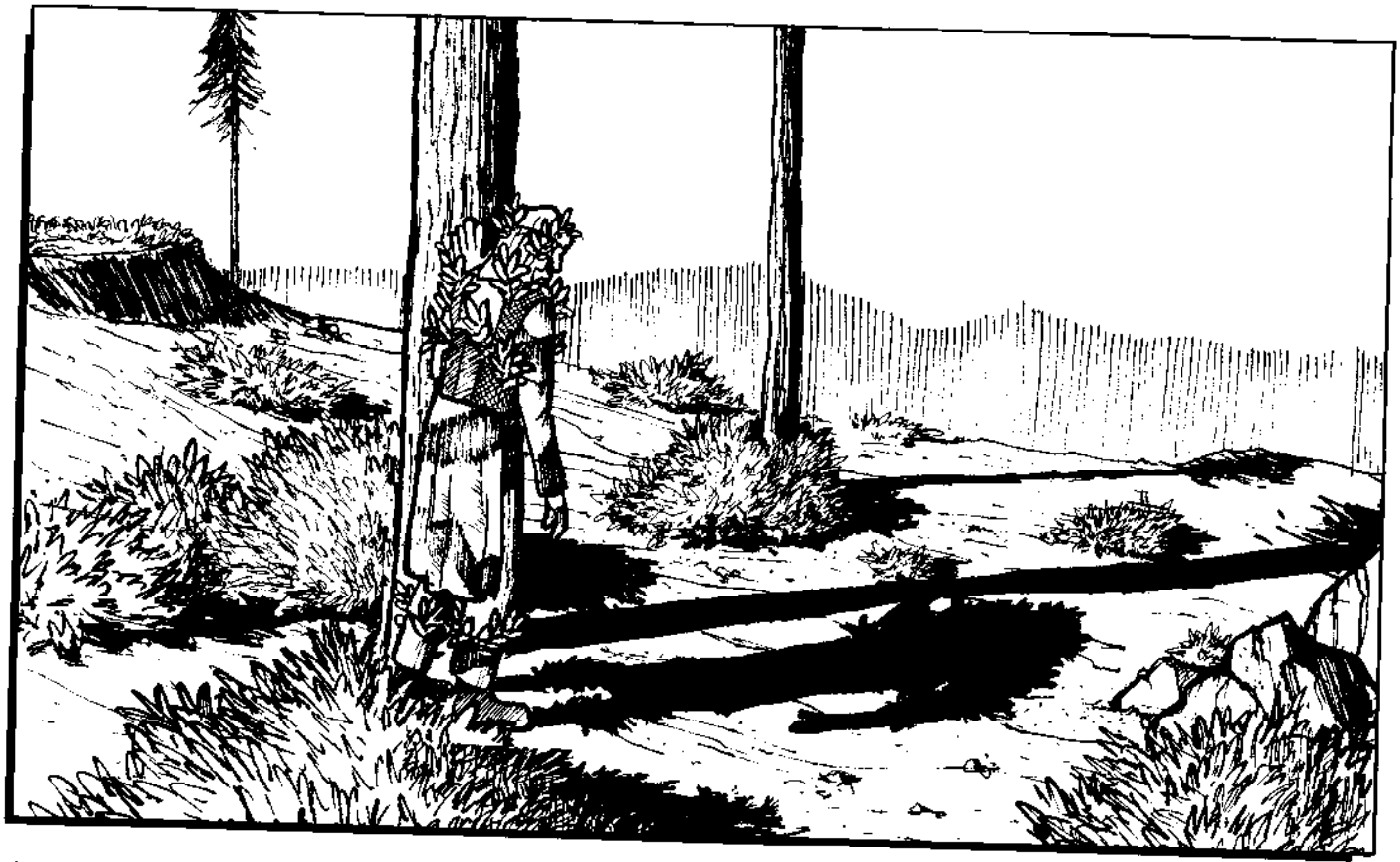

Figure 28-8. Shadow Breakup.

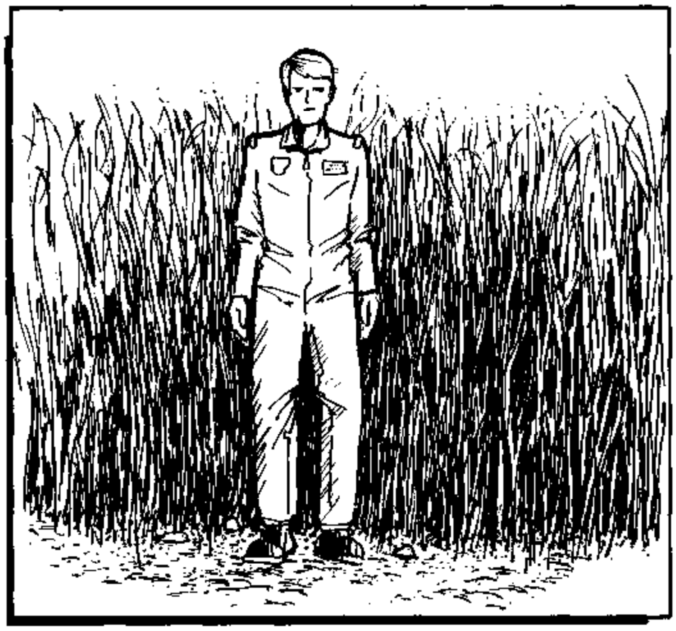

Figure 28-9. Contrast.

can be an aid to enemy observers. The greater the contrast in color, the more visible the object appears (figure 28-9).

(2) Color differences or differences in hue, such as red and green-yellow, become increasingly difficult to distinguish as the viewing range is increased. This happens because of atmospheric effects. Colors in nature, except for certain floral and tropical animal life, are not brilliant. The impression of the vividness of nature's colors results from the large areas of like colors involved and contrast of these areas with each other. The principal contrast is in their dark and light qualities. However, the dark and light color contrast does not fade out quickly and is distinguishable at greater distances. Therefore, as a first general principle, the camouflage should match the darker and lighter qualities of the background and be increasingly concerned with the colors involved as the viewing range is decreased or the size of the object or installation becomes larger. A second general rule to follow is to avoid contrasts of hues. This is especially true in areas with heavy vegetation. Light-toned colors, such as leaf bottoms, should be avoided as they tend to attract attention.

f. Tone. Tone is the amount of contrast between variations of the same color. It is the effect achieved by the combination of light, shade, and color. In a black-and-white photograph, the shades of gray in which an object appears is known as tone (figure 28-10). By adding texturing material to a smooth or shiny surface, the surface can be made to produce a darker tone in a photograph, because the textured surface now absorbs more light rays. Objects become identifiable as such because of contrasts between them and their background. Camouflage blending is the process of eliminating or reducing these contrasts. The principal contrast is that of tone; that is, the dark and light relationship existing between an object and its background. The two principal means available for reducing tone contrast are the application of matching or neutral coloration and the use of texturing to form disruptive patterns. Poorly chosen, disruptive patterns tend to make the object more conspicuous instead of concealing it.

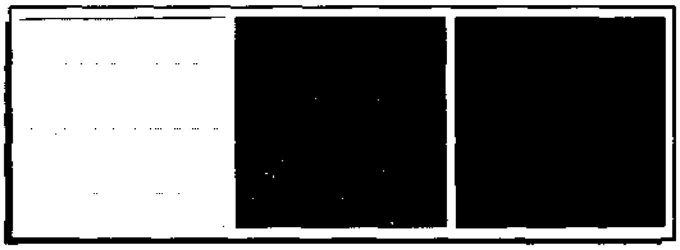

Figure 28-10. Three Half-Tone Blocks.

g. Movement. Of the eight factors of recognition, movement is the quickest and easiest to detect. The eye is very quick to notice any movement in an otherwise still scene. The aerial camera can record the fact that something has moved when two photographs of the same area are taken at different times. If an object has moved, the changed position is apparent when the two photographs are compared (figure 28-11).

h. Shine:

(1) Shine is a particularly revealing signal to an observer. In undisturbed, natural surroundings, there are comparatively few objects which cause a reflected shine. Skin, clean clothing, metallic insignia, rings, glasses, watches, buckles, identification bracelets, and similar items produce "shine." When light strikes smooth surfaces such as these, it may be reflected directly into the observer's eye or the camera lens with striking emphasis (figure 28-12).

(2) Such items must be neutralized by staining, covering, or removing to prevent their shine from revealing the location of evaders. This is especially true at night.

28-6. Principles and Methods of Camouflage:

a. No matter how applied, camouflage can be successful only by observing three fundamental principles. These basic principles of camouflage are choice of position, camouflage discipline, and camouflage construction.

b. When these factors have been considered, the evader is ready to begin application of various methods of deceiving the enemy. These methods are:

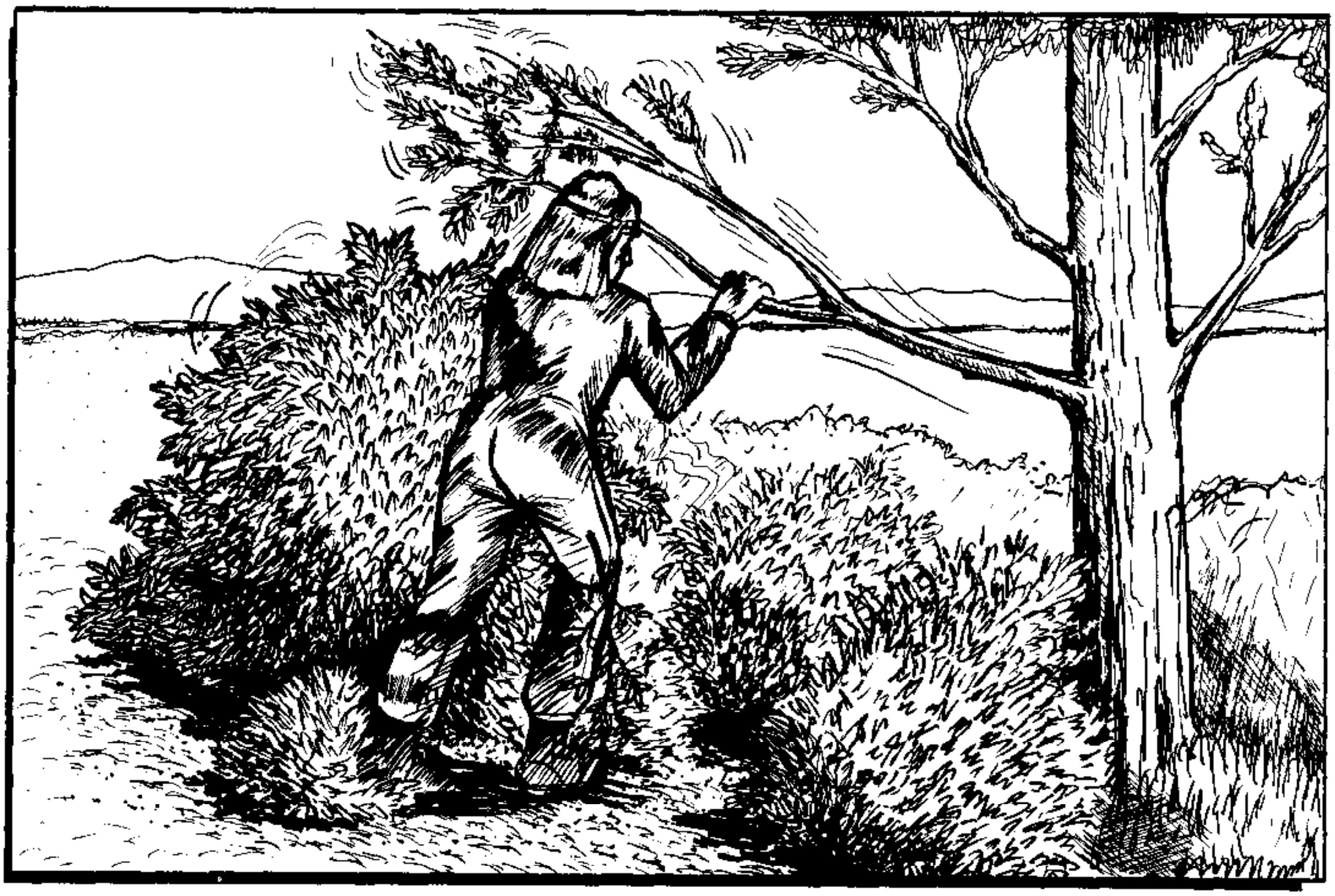

Figure 28-11. Movement.

(1) Hiding. The complete concealment of a person or object by physical screening.

(2) Disguising. Changing the physical characteristics of an object or person in such a manner as to fool the enemy.

(3) Blending. The arrangement of camouflage material on or about an object in such a manner as to make the object appear to be part of the background. To properly use these methods, three simple rules should be followed:

(a) First, the background should be changed as little as possible. When choosing a position to gain concealment, a background should be chosen that will visually absorb the elements of the position. Evaders should use a "natural" position if available. They should look for an existing position which can be used almost as is, such as a cave or thicket if there are many like it in the area. Isolated landmarks such as individual trees, haystacks, or houses should be avoided. They tend to attract attention and are likely to be searched first because they are so obvious. At times, by making use of background, complete concealment against visual and photographic detection may be gained with no construction. In terrain where natural cover is plentiful, this is a simple

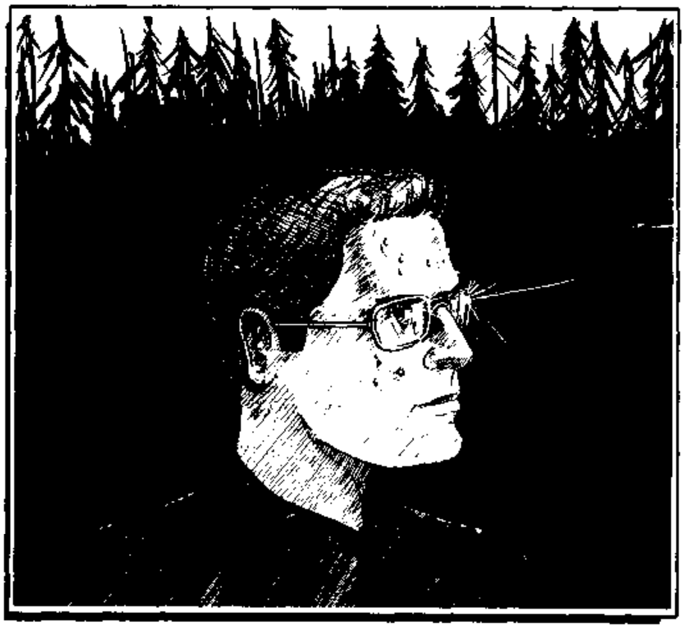

Figure 28-12. Reflected Shine.

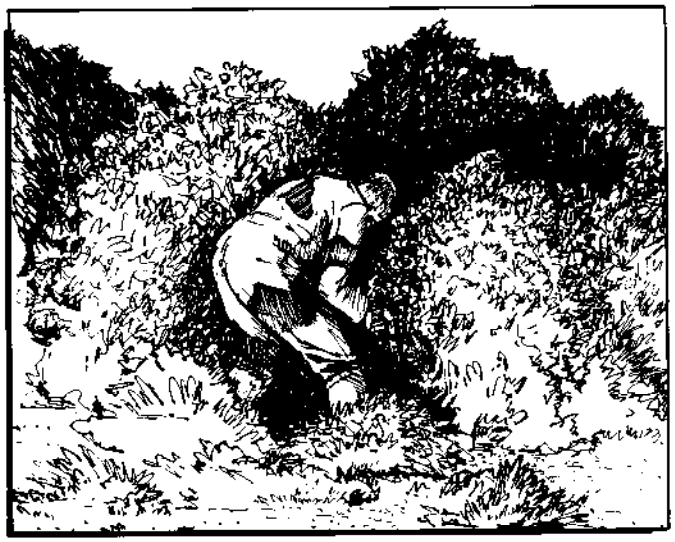

Figure 28-13. Background.

task. Even in areas where natural cover is scarce, concealment may be achieved through use of terrain irregularities. Regardless of the activity involved, evaders must always be mentally aware of their positions (figure 28-13).

(b) Secondly, the evader should use camouflage discipline. This means all of the factors of concealment are continuously applied.

-1. Daytime. Camouflage discipline is the avoidance of activity which changes the appearance of an area or reveals military objects to the enemy. A well-camouflaged position is only secure as long as it is well maintained. Concealment is worthless if obvious tracks point like directional arrows to the heart of the location or if signs of occupancy are permitted to appear in the vicinity. Tracks, debris, and terrain disturbances, are the most common signs of activity. Therefore, natural lines in the terrain should be used. If practical, exposed tracks should be camouflaged by brushing or beating them out. If leaving tracks is unavoidable, they should be placed where they will be least noticed and partially concealed (along logs, under bushes, in deep grass, in shadows, etc.). If tracks cannot be concealed, brushing them out will help them disintegrate quickly. Tying rags or brush to the feet will disguise boot prints and may help disguise them as refugee tracks (figure 28-14).

-2. Nighttime. Visual concealment at night is less necessary than in the daytime; however, noises at night are more noticeable. A simple act such as snoring may prove fatal. Calling to one another, talking, and even whispering should be kept to a minimum (figure

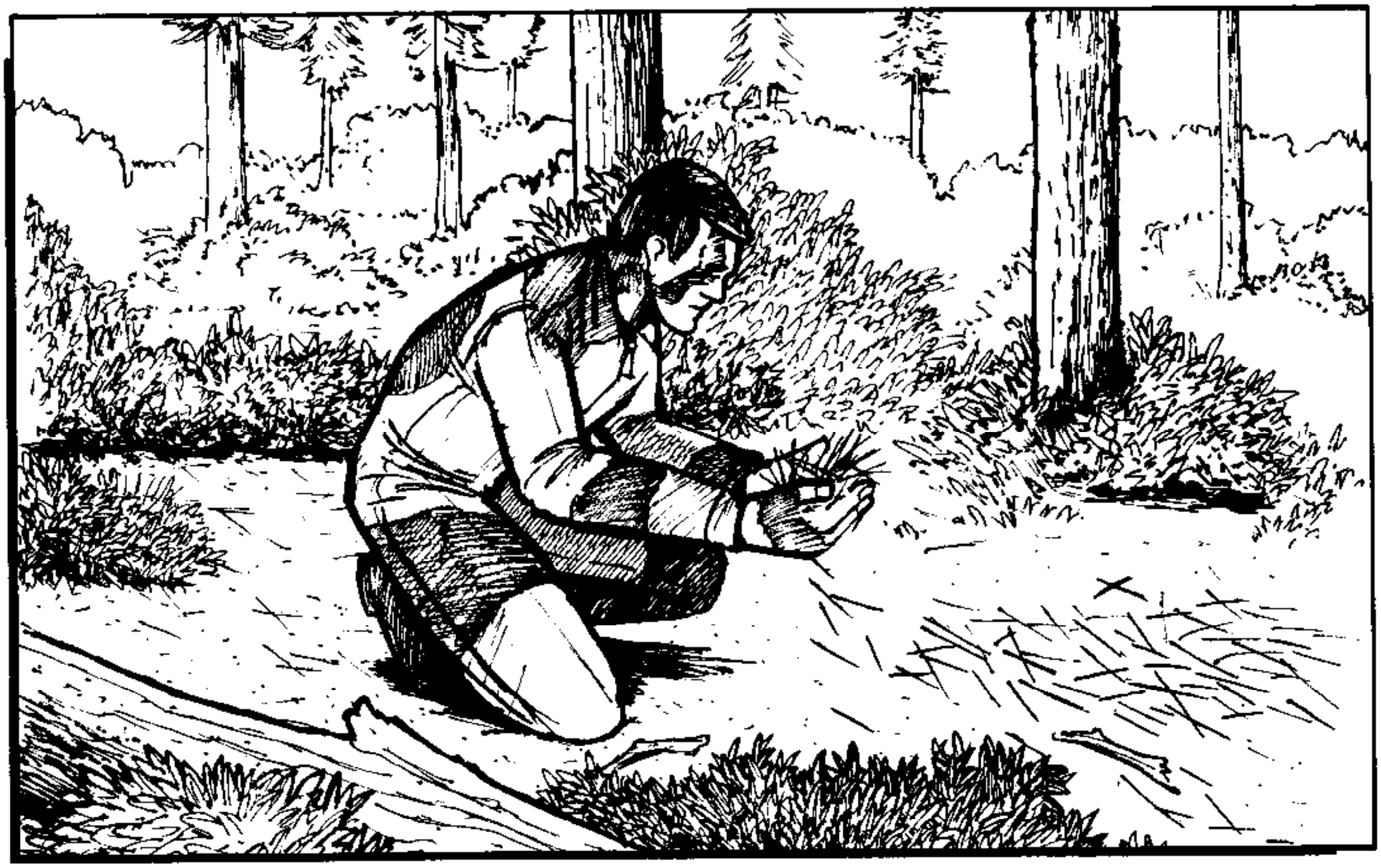

Figure 28-14. Camouflaging Tracks.

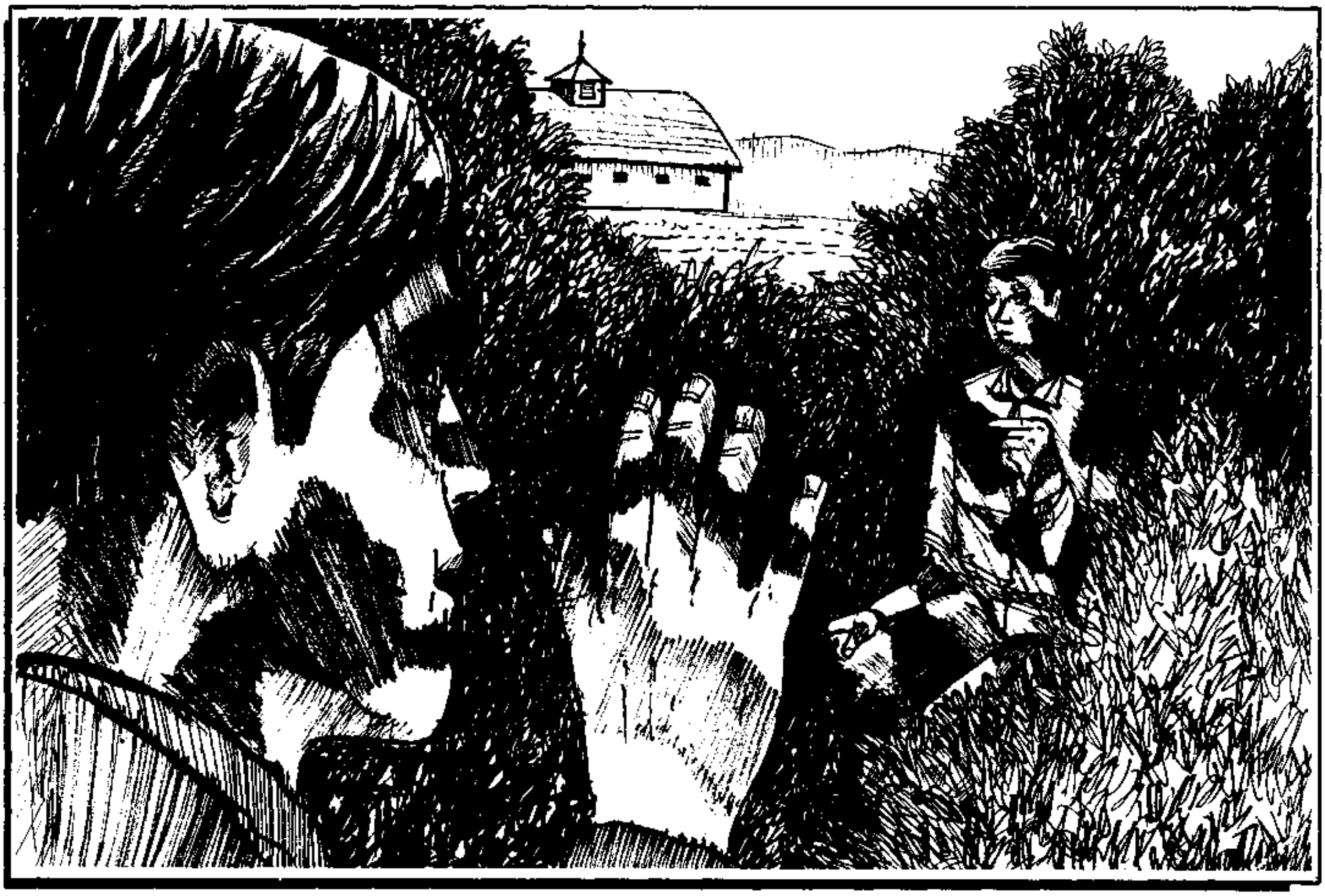

Figure 28-15. Sound.

28-15). But by far, the most important apsect of night discipline is light discipline. Lights at night not only disclose the evaders' position but also hinder the evaders' ability to detect the enemy. Even on the darkest nights, eyes grow accustomed to the lack of light in approximately 30 minutes. Everytime a match is lit or a flashlight is used, the eyes must go through the complete process of getting adjusted to the darkness again. Smoking and lights should be prohibited at night in areas in close proximity to the enemy because the light is impossible to conceal. Additionally, a cigarette light aggravates the situation by creating a reflection which completely illuminates the face. The smell of the evader's foreign tobacco would stand out even if the enemy is smoking.

-3. Evaders can lessen the effects of sound by simply taking precautions against sound production. They should avoid any sound-producing activity. Walking on hard surfaces should be avoided and full use should be made of soft ground for digging. Hand signals or signs should be used when possible during group travel. Individual equipment should be padded and fastened in such a manner as to prevent banging noises.

c. The evader should consider the following points regarding the use of camouflage:

(1) Take advantage of all natural concealment.

(2) Don't over-camouflage. Too much is as obvious as too little.

(3) When using natural camouflage, remember that it fades and wilts, so change it regularly.

(4) If taking advantage of shadows and shade, remember they shift with the Sun.

(5) Above all, avoid unnecessary movement.

(6) When moving, keep off the skyline; use the military crest (three-quarters' way up the hill).

(7) Do not expose anything that may shine.

(8) Break up outlines of manmade objects.

(9) When observing an area, do so from a prone position, while in cover.

(10) Match vegetation used as camouflage with that in the immediate locale, and when moving from position to position, change camouflage to blend with the new area's vegetation types.

28-7. Individual Camouflage:

a. At this point, with some of the general information about camouflage presented, it is time for a more detailed examination concerning individual camouflage.

b. Generally, individual camouflage is that personal concealment which evaders must use to deceive the enemy. Evaders must know how to use the terrain for effective concealment. Evaders must dress for the best concealment and carefully select their routes to provide for as much concealment as possible. All of the methods and techniques of camouflage addressed in this section have been successfully used by past evaders. If this information is learned and practiced by today's aircrew members (tomorrow's possible evaders), they will be more successful in evading and have a greater chance of returning to friendly forces.

c. Evaders should remember that in some areas they may have to engage in camouflage activities designed to deceive two types of enemy observation—ground and air. Many objects which are concealed from ground observation, may be seen from the air. This means the evader should camouflage for both types of observation.

d. Form is basic shape (body outline) and height. Three things which give an evader away in terms of form are to reveal outline of head and shoulders, to present straight lines of sides, and to allow the inverted "V" of the crotch and legs to be distinguishable. If staying in shadows, blending with background, adopting body positions other than standing erect, and other behavioral procedures are inadequate. They can be camouflaged by using "add-ons" such as branches or twigs to break up the lines. This addition of vegetation will also help an evader blend in with the background.

e. Effective concealment of evaders depends largely on the choice and proper use of background. Background varies widely in appearance, and evaders may find themselves in a jungle setting, in a barren or desert area, in a farmyard, or in a city street. Each location will require individual treatment because location governs every concealment measure taken by the individual. Clothing which blends with the predominant color of the background is desirable. There will be occasions when the uniform color must be altered to blend with a specific background. The color of the skin must receive individual attention and be toned to blend with the background.

f. There are certain general aspects of individual body and equipment camouflage techniques which apply almost anywhere. The evaders should take each of the following areas under consideration.

(1) Exposed Skin. The contrast in tone between the skin of face and hands and that of the surrounding foliage and other background must be reduced. The skin is to be made lighter or darker, as the case may be, to blend with the surrounding natural tones. The shine areas are the forehead, the cheekbones, nose, and chin.

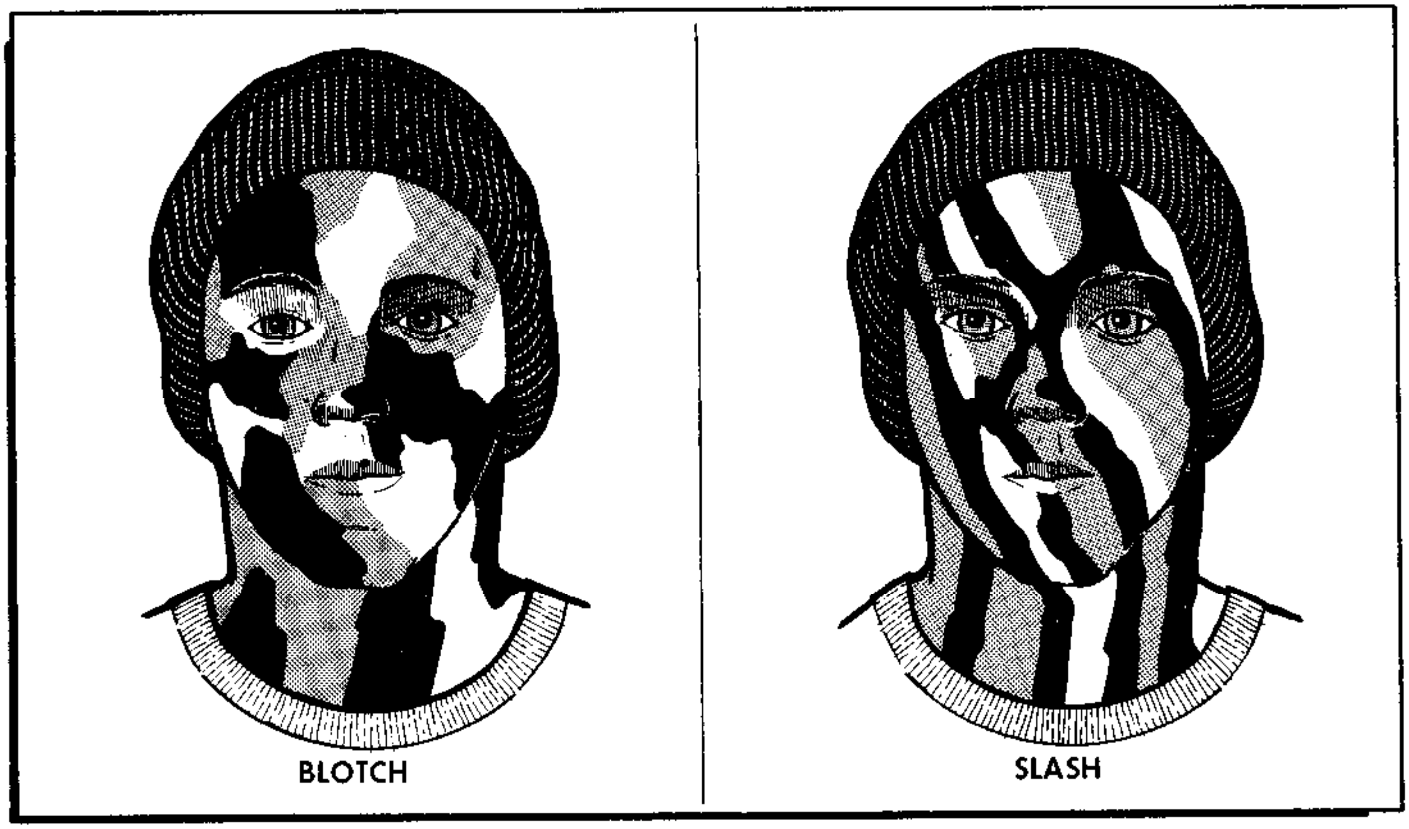

Figure 28-16. Camouflaged Faces.

These areas should have a dark color. The shadow areas such as around the eyes, under the nose and under the chin should have a light color. The hands, arms, and any other exposed areas of skin must also be toned down to blend with the surroundings. Burnt cork, charcoal, lampblack, mud, camouflage stick, berry stains, carbon paper, and green vegetation can all be used as toning materials.

(a) A mesh mosquito face net, properly toned down, is an effective method of breaking up the outlines of the face and ears.

(b) Two primary methods of facial camouflage have been found to be successful patterns. They are the "blotch" method for use in deciduous forests, and the "slash" method for use in coniferous forests (figure 28-16).

(c) Application of these two patterns are simply modified appropriately to whatever environment the evader is in. In the jungle, a broader slash method would be used to cover exposed skin; in the desert, a thinner slash; in barren snow, a wide blotch; and in grass areas, a thin type slash. To further break up the outline of the facial features, a flop hat or other loosely fitting hat may help. A beard that is not neatly trimmed may also aid the evader.

(d) When toning down the skin, evaders should not neglect to pattern all of the skin; for example, the back of the neck, the insides and backs of the ears, and the eyelids. Covering these areas may help somewhat, especially if there is a lack of other material to tone down the skin. Vegetation hung from the hat, collar buttoned and turned up, a scarf, or even earflaps may help. To cover the hands, evaders may use flight gloves, mittens, or loose cloth if unable to tone down wrists, backs of hands, and between fingers sufficiently. Evaders should watch for protruding white undergarments, T-shirt, long underwear sleeves, etc. They should also tone down these areas.

(e) Lack of hair or light colored hair requires some type of camouflage. This could include those applied to the skin or an appropriate hat, scarf, or mosquito netting.

(f) Odors in a natural environment stand out and may give evaders away. Americans are continually surrounded by artificial odors and are not usually aware of them. Human body odor would have to be very strong to be detected by ground troops (searchers) that have been in the field for long periods. The following odors should be of concern to the evader.

-1. Soaps and Shampoos. In combat areas, personnel should always use unscented toilet articles. Shaving cream, after shave lotion, perfume, and other cosmetics are to be avoided. The potential evaders should also realize that insect repellent is scented. They should try to use headnets, but if forced to use a repellent, the camouflage stick which has repellent in it is the least scented. Tobacco should not be used. The stain and odor should be removed from the body and clothing. Gum or candy may have strong or sweet smells—evaders should take care to rinse out their mouths after use. These odors, especially tobacco, can be detected at great distances.

-2. Smoke odors from campfires may permeate clothing, but if the potential searchers use fires for cooking and heat, they probably can't detect it on evaders.

(2) Clothing and Personal Items. These items require attention both before assignment to a combat mission and again if forced to evade. Prior preparation for the survivor may include:

(a) Ensuring that flight clothing does not smell of laundry products and is in good repair and not worn to the point where it shines or is faded.

(b) Checking zippers for shine and function.

(c) Checking rank insignia and patches for light reflection and color. Remove name tag, branch of service tag, and rank (whether they are stripes or metal insignia, bright unit patches, etc.) from the uniforms and place them either in the pack or in a secure pocket. Underwear should also be subdued for camouflage in the event that the outer layer is torn. Boots should be black but not shiny. Shiny eyelets should be repainted. Squeaky boots should be fixed or replaced. Sanitize pocket or wallet contents. Remove items which might aid in enemy exploitation attempts of the individual PW or PWs as a group; for example, credit cards, photographs, money, and addresses. Evaders should carry only those necessary pieces of identification which will prove a person is a US military member.

(3) Additional Clothing. Additional clothing may be desirable and located in flight clothing, such as hat, socks, scarf, and gloves. In an evasion environment, clothing and equipment need quick camouflaging attention. Anything to be discarded should be hidden at the initial landing site.

(4) Sanitizing Clothing. Clothing should be sanitized by removing anything bright or shiny. Evaders should consider camouflaging their clothing (to include boots) just as they would their skin. This is detrimental to the insulative qualities of their clothes but not as much as bullets or prison barbed wire if they are seen.

(5) Camouflaging Clothing:

(a) One principle of camouflage is to disrupt or conceal uniform color, straight lines, and squares—things rarely found in natural features. If ground-to-air signals are designed to exploit these visual characteristics, then evaders should certainly want to eliminate them.

(b) Evaders should reduce the tone of all equipment by smearing it with camouflage stick, mud, etc., or with whatever is available in a mottled pattern. In some instances equipment will have to be lighter in tone, in others, darker.

(c) To maintain the functional capability of clothing and equipment, it must be kept clean. In some areas, however, these items may be the only natural camouflage material survivors have to work with (such as desert regions). But, in areas where they have access to vegetation and the various dyes which can be made from vegetation, the vegetation should be used. In contrast to substances which soil the material and actually break down the fibers, dyes derived from grasses or plant sap (banana trees, ash trees, etc.) will offer the evader the toning material necessary to break up the solid green of a uniform and leave the fabric grit-free and still able to "breathe." The same saps which produce stains for cloth can be used to discolor metal objects. Banana tree sap, when left on the metal blade of a knife, will produce a blue-black stain which is a permanent discoloration. Trappers still boil their traps in ash tree chips and water to produce the blue-black, rust-inhibiting coloration to the tools of their trade.

g. All principles and techniques for care and use of clothing and equipment cannot be forgotten or ignored in an evasion situation, although some modifications may be necessary. A number of variables will influence what changes or omissions will be necessary.

(1) All cutting tools must be kept sharp. Evaders should try to coincide these noisy, yet essential, tasks with natural noise in the area (a downpour of rain for instance) or in a protected, noise-dampening area.

(2) Clothing must be kept clean if it is to protect a survivor from a harsh environment. Dirt-clogged, perspiration-soaked fibers will not give the insulating qualities of clean cloth. Clothing can be washed during the downpour of rain or possibly under the cover of darkness in a stream. Convenient, secluded puddles of water may afford the opportunity a survivor needs to clean clothing and equipment properly.

(3) Cooking and eating utensils must be kept clean on the inside to prevent dysentery and diarrhea. Simultaneously, the outsides can be toned down with soil, mud, etc., to camouflage them.

(4) Where metal pieces come into contact with one another, there should be padding between them so they will not inadvertently "clank" together. Evaders should place all items needed for environmental protection in the top of the pack where they will be most readily available. The rest of the gear can be used as padding around metal objects. In this manner, with everything stored inside the bundle (pack), it is secure from loss, damage, and enemy observation, as well as being readily available when needed. Evaders should also remove jewelry, watches, exposed pens, and glasses if possible. If glasses are required, hat netting or mask may help reduce shine.

(5) An evader's pockets should be secured and all equipment, including dog tags, arranged so that no jingle or rattle sounds are made. This can be done with cloth, vegetation, padding, or tape.

(6) Evaders should minimize the sound of clothing brushing together when the body moves. Moving in a careful manner can decrease this sound. Evaders should remember that camouflaged clothing and equipment alone won't conceal, but it must be used intelligently in accordance with the other principles of camouflage and movement. As one example, even if evaders are perfectly camouflaged for the arctic, there could still be problems. Because snow country is not all white, shadows and dark objects appear darker than usual. A snowsuit cannot conceal the small patches of shadow caused by the human figure, but that is not necessary if the background contains numerous dark areas. If the background does not contain numerous dark areas, maximum use is to be made of snowdrifts and folds in the ground to aid in individual concealment (figure 28-17).

(7) The concept of blending in with the background is indeed an important one for the evader to understand. One major point in blending with the background is not to show a body silhouette.

(8) Losing the body silhouette is done by making use of the shadows in the background. Evaders should be constantly aware of two factors—silhouette and shadow. From a concealment point of view, backgrounds consist of terrain, vegetation, artificial objects, sunlight, shadows, and color. The terrain may be flat and smooth, or it may be wrinkled with gullies, mounds, or rock outcroppings. Vegetation may be dense jungle growth or no more than small patches of desert scrub growth. The size of artificial objects may range from a signpost to a whole city block. There may be many colors in a single background, and they may vary from the almost black of a deep woods to the sand pink of some desert valleys. Blending simply means the matching with as many of these backgrounds as possible and avoiding contrast. If it is necessary for evaders to be positioned in front of a contrasting or fixed background, they must be aware of their position and take cover in the shortest possible time. The next point to which they will move for concealment must be selected in advance and reached as quickly as possible.

(9) As in the daytime, silhouette and background at night are still the vital elements in concealment (figure 28-18). A silhouette is always black against a night sky, and care must be taken at night to keep off the skyline. On moonlit nights, the same precautions must be taken as in daylight. It should be remembered that the position of the enemy observer, and not the topographic crest, fixes the skyline. At night, sound is an amplified, revealing signal. Movement must be careful, quiet, and

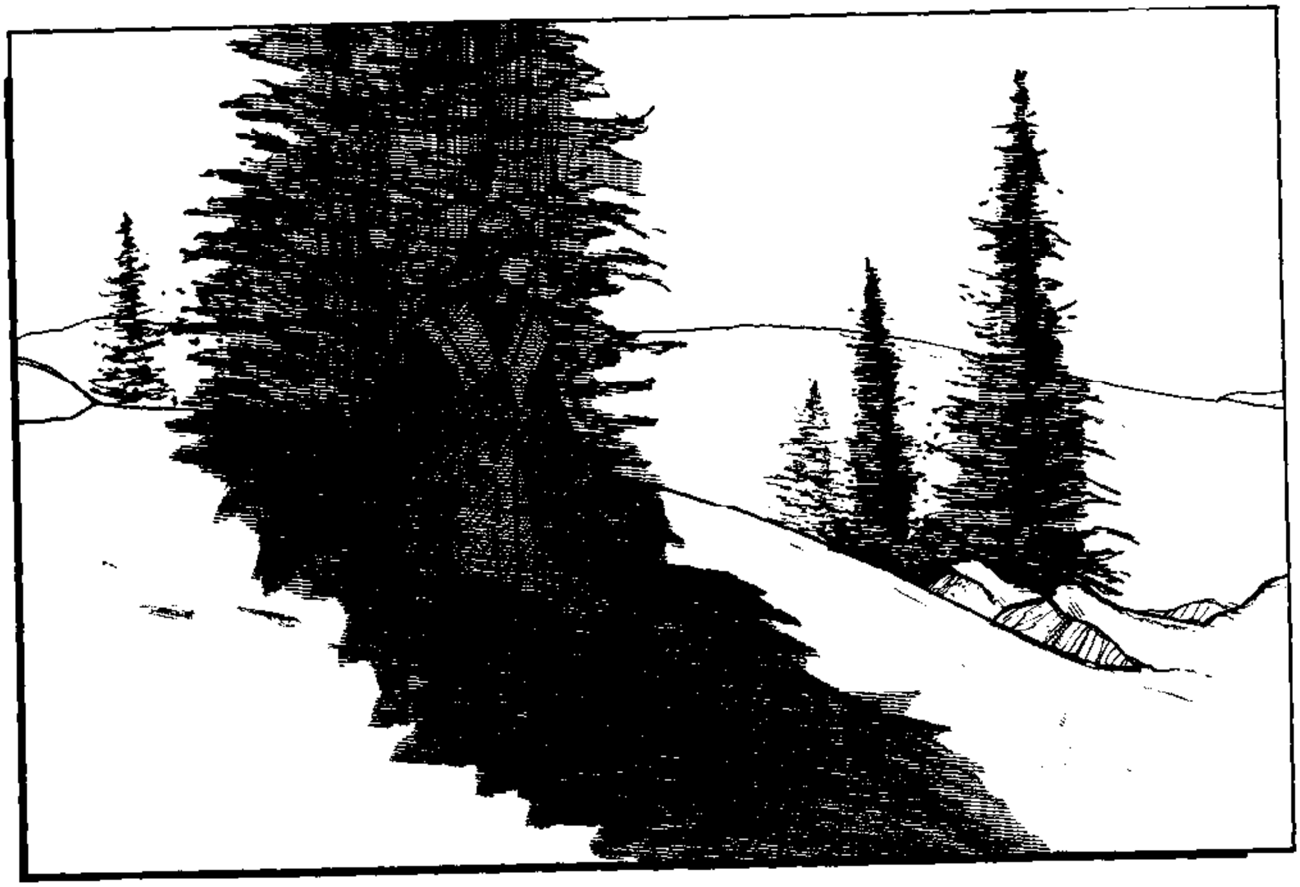

Figure 28-17. Arctic Travel.

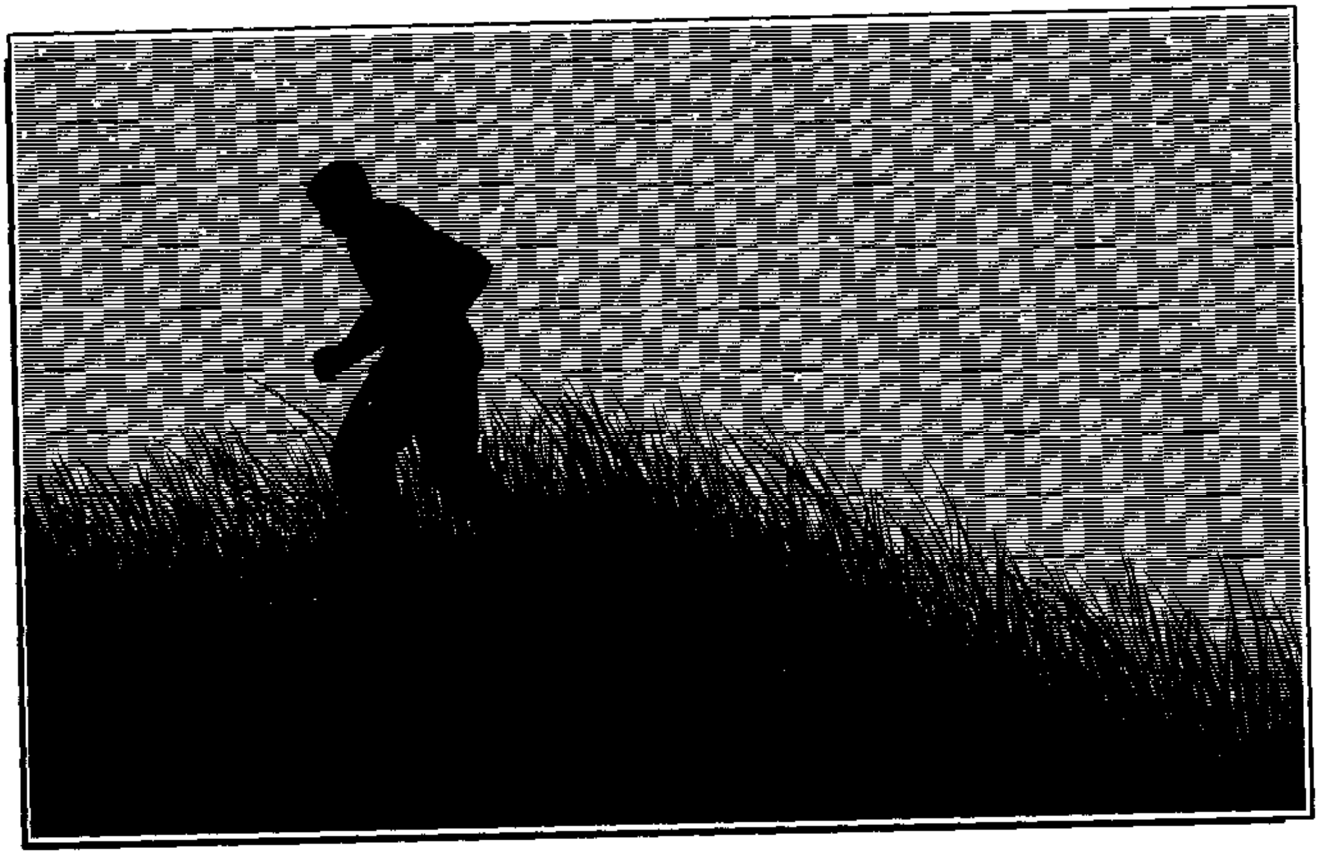

Figure 28-18. Silhouetting.

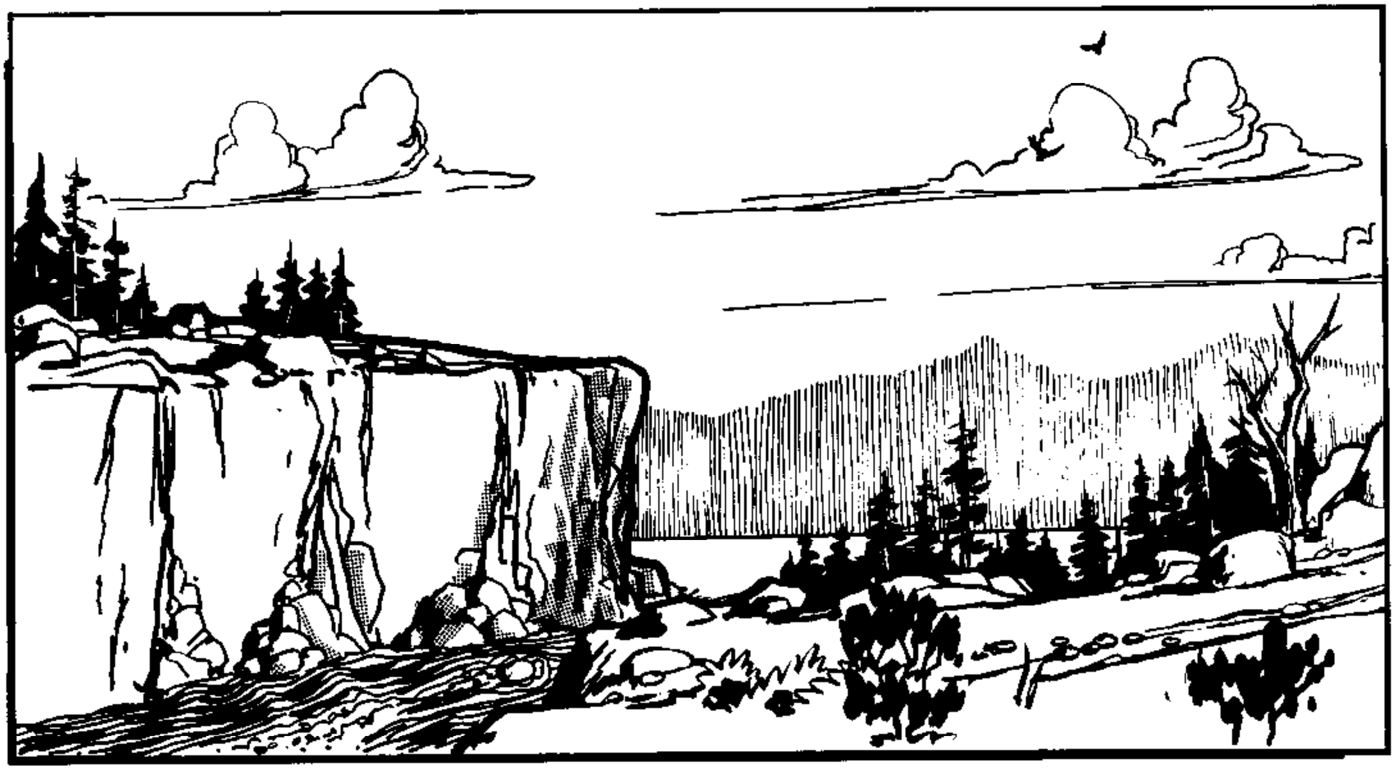

Figure 28-19. Natural Materials.

close to the ground. If the pop of a flare is heard before the illuminating burst, evaders must drop to the ground instantly and remain motionless. If they are surprised by the light, they must freeze in place with their faces down.

28-8. Concealment in Various Geographic Areas:

a. When not otherwise specified, temperate zone terrain is to be assumed in this section. Desert, snow, and ice areas are mostly barren and concealment may require considerable effort. Jungle and semitropical areas usually afford excellent concealment if the evader employs proper evasion techniques.

b. First, some general observations and rules regarding the addition of vegetation to the uniform and equipment. The cycles of the seasons bring marked changes in vegetation, coloring, and terrain pattern requiring corresponding changes in camouflage. Concealment which is provided in wooded areas during the summer is lost when leaves fall in the autumn. This will create a need for additional camouflage construction. Also, vegetation must be of the variety in the evaders' immediate location. It must be changed if it wilts or the evaders move into a different vegetation zone. Evidence of discarding the old and picking the new should be hidden. The vegetation should not be cut, this will give evidence of human presence.

c. Any type of material indigenous to the locality of the evaders may be classified as natural material. Natural materials consist of foliage, grasses, debris, and earth. These materials match local colors and textures and when properly used are an aid against both direct and indirect observation. The use of natural materials provides the best type of concealment. The chief disadvantage of natural foilage is that it cannot be prepared ahead of time, is not always available in usable types and quantities, wilts after gathering, and must be replaced periodically. Foliage of coniferous trees (evergreens) retains its camouflage qualities for considerable periods, but foliage that sheds leaves will wilt in a day or less, depending on the climate and type of vegetation (figure 28-19).

(1) The principal advantage in using live vegetation is its ability to reflect infrared waves and to blend in with surrounding terrain. When vegetation is used as garnishing or screening, it must be replaced with fresh materials before it has wilted sufficiently to change the color or the texture. If vegetation is not maintained, it is ineffective. Thorn bushes, cacti, and other varieties of desert growth retain growing characteristics for long periods after being gathered.

(2) The arrangement of foliage is important. The upper sides of leaves are dark and waxy; the undersides are lighter. In camouflage, therefore, foliage must be placed as it appears in its natural growing state, top sides of leaves up and tips of branches toward the outside of the leaves (figure 28-20).

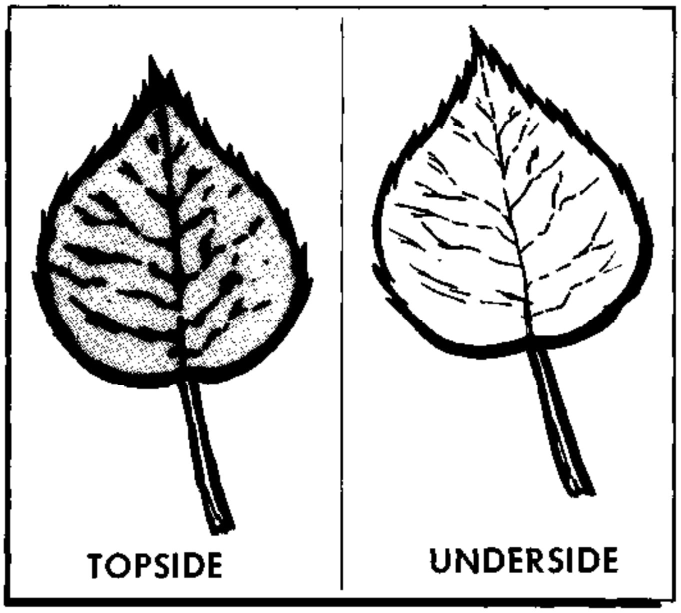

Figure 28-20. Dark and Light Leaves.

(3) Foliage gathered by survivors must be matched to existing foliage. For example, foliage from trees that shed leaves must not be used in an area where only evergreens are growing. Foliage with leaves that feel leathery and tough should be chosen. Branches grow in irregular bunches and, when used for camouflage, must be placed in the same way. When branches are placed to break up the regular, straight lines of an object, only enough branches to do this should be used. The evader must adopt principles that apply and know that the enemy also applies these principles.

(4) When vegetation is applied to the body or equipment of evaders, it must be secured to the clothing or equipment in such a way that:

(a) Any inadvertent movement of material will not attract attention.

(b) It appears to be part of the natural growth of the area; that is, when the evaders stop to hide, the light undersides of the vegetation are only visible from beneath. After evaders complete their camouflage, they should inspect it from the enemy's point of view. If it does not look natural, it should be rearranged or replaced.

(c) It does not fall off at the wrong moment and leave evaders exposed, or it should not show evidence of the evaders' passage through the area.

(5) Too much vegetation can give evaders away.

(6) If cloth material is used like vegetation to break up shape and outline and to help blend in with the environment, there are some points evaders should be aware of. Cloth can be used successfully when wrapped around equipment or designed into loose, irregular shaped clothing or accessories (figure 28-21).

(a) Some of the materials evaders may use are:

-1. The colors (green, brown, white) of parachute materials plus the harness.

-2. Excess clothing the evader may have prepacked (scarf, bandana, etc.).

-3. Burlap, when found. It is used in battle areas in the form of sandbags.

(b) Most artificial material is versatile, but it can have drawbacks:

-1. Parachute material, for example, tends to shine and the unraveled edges may leave fine filaments of nylon on the ground as evidence of evaders in an area. Parachute material is very lightweight. A sudden breeze might cause it to move when movement is not desirable.

-2. A white suit of parachute material is excellent for winter evasion over snow and ice environments. The time required to fabricate this suit should be considered. (NOTE: Shadows cast on the snow cannot be

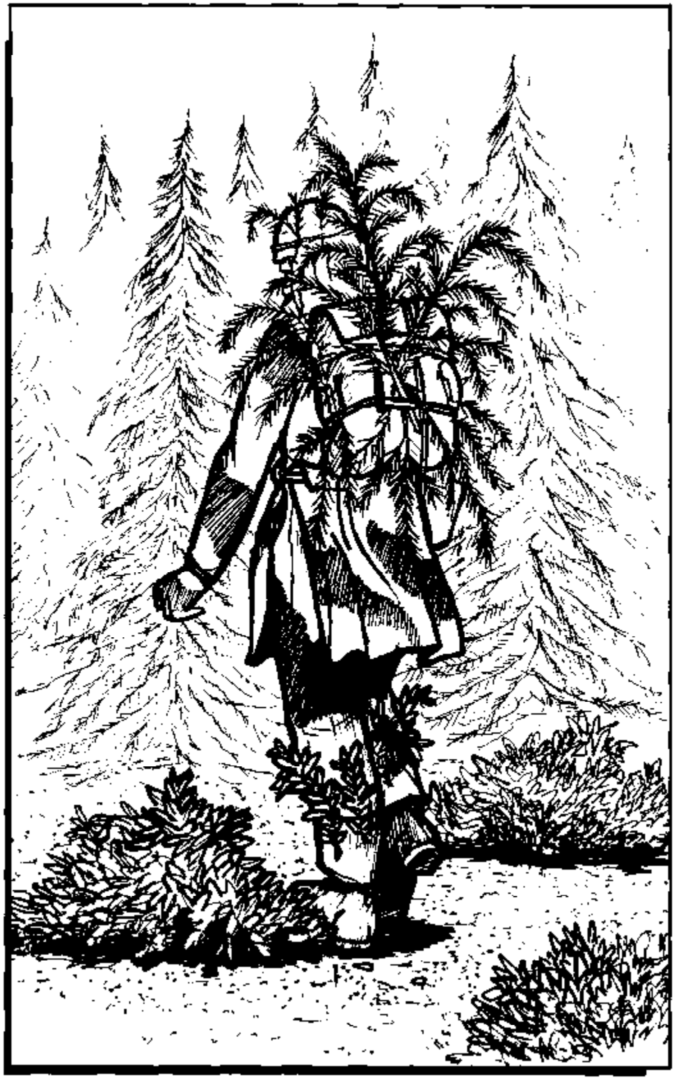

Figure 28-21. Breakup of Shape and Outline.

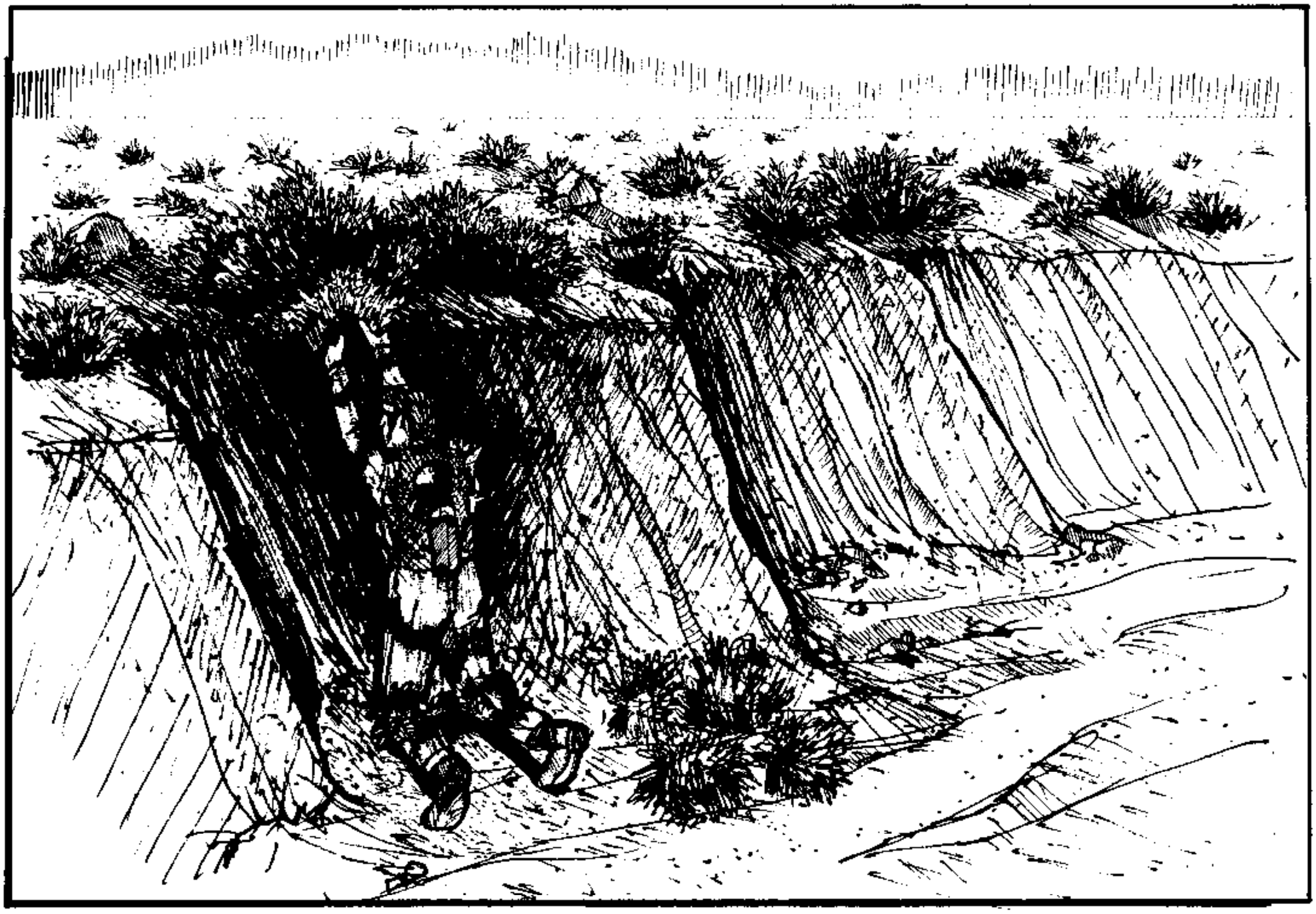

Figure 28-22. Concealment by Shadows.

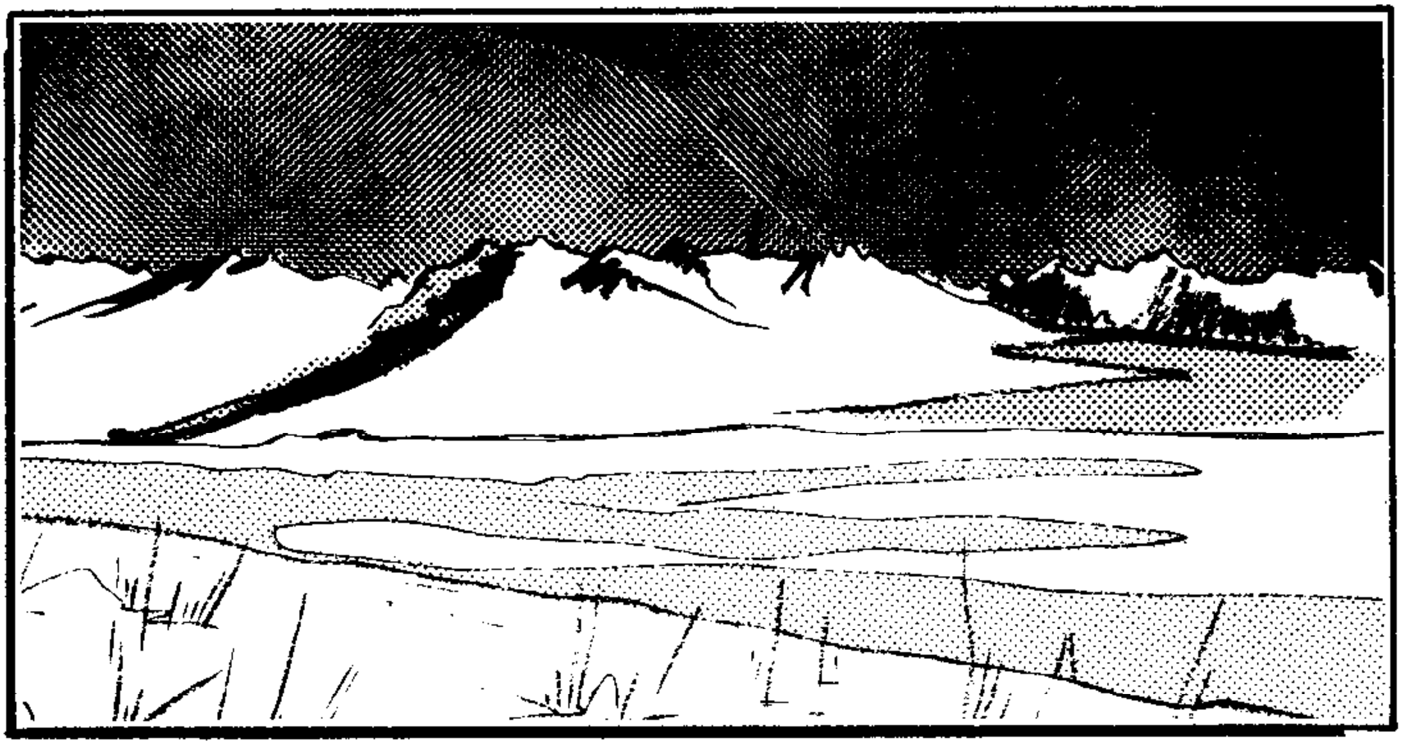

Figure 28-23. Above Timberline.

camouflaged; they must be masked by terrain or other shadows.)

28-9. Concealment Factor for Areas Other Than Temperate:

a. Desert. Lack of natural concealment, high visibility, and bright tone (smooth texture) all emphasize the need for careful selection of a position for a campsite. Deep shadows in the desert, strict observance of camouflage discipline, and the skillful use of deception and camouflage materials aid in concealing evaders in a desert area.

(1) Deserts are not always flat, single-toned areas. They are sometimes characterized by strong shadows with heavy broken terrain lines and sometimes by a mottled pattern. Each type of desert terrain presents its own problems. When evaders and their shelters are located in the desert, their shadows are inky black and in strong contrast to their surroundings and are extremely conspicuous. To minimize the effect of these shadows when possible, use concealment that is afforded by the shadows of deep gullies, scrub growth, and rocks (figure 28-22).

(2) Many objects which cannot be concealed from the air can be effectively viewed from the ground. Even though these objects are observed from the air, lack of reference points in the terrain will make them difficult to locate on a map.

b. Snow and Ice. From the air, snow-covered terrain is an irregular pattern of white, spotted with dark tones produced by objects projecting above the snow, their shadows, and irregularities in the snow-covered surface such as valleys, hummocks, ruts, and tracks. It is necessary, therefore, to make sure dark objects have dark backgrounds for concealment to control the making of tracks in the snow and to maintain the snow cover on camouflaged objects.

(1) Mountain Areas Above Timberline and Arctic Areas. Common characteristics include an almost complete snow cover with a minimum of opportunities for concealment. Only a few dark objects protrude above the snow except for rugged mountain peaks (figure 28-23).

(2) Mountain Areas Below the Timberline and Subarctic Areas. Common characteristics of these areas are forests, rivers, lakes, and artificial features such as trails and buildings. The appearance of the area is irregular in pattern and variable in tone and texture (figure 28-24).

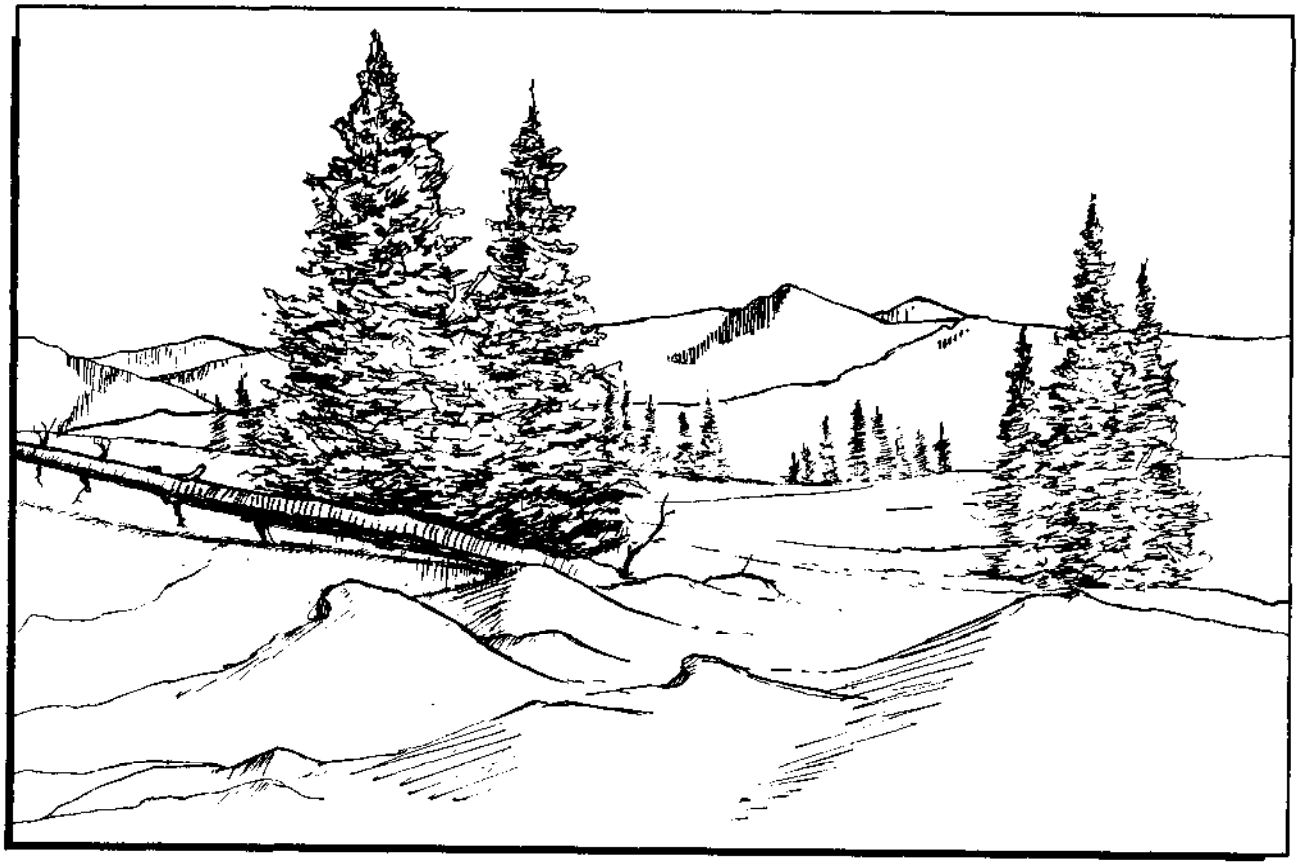

Figure 28-24. Below Timberline.

(3) Areas Between the Subarctic Zone and the Southern Boundary of the Temperate Zone. These have the same characteristics as mountain areas below the timberline and subarctic areas.

c. Blending With Background in Snow and Ice Terrains. No practical artificial material has yet been developed which will reproduce the texture of snow sufficiently well to be a protection against recognition by aerial observers. Concealment from direct ground observation is relatively successful with the use of white "snowsuits," white pants, and whitewash; these measures offer some protection against aerial detection. (White parachute cloth should also be considered.)

(1) People evading in snow-covered frozen areas should wear a complete white camouflage outfit. A white poncho-like cape can be made easily from parachute material (figure 28-25).

(2) A pair of white pants will normally be sufficient in a heavily wooded area. However, following or during a heavy snowfall when the trees are well covered with snow, the wearing of a complete white camouflage suit is necessary to blend in with the background. Other equipment, such as packs, should also be covered with white material.

d. Camouflage Checklist. The following checklist can be used to remember ideas concerning camouflage and to determine the completeness of individual camouflage application:

(1) Effective concealment is protection from hostile observation from the ground as well as from the air.

(2) Natural terrain lines are to be used for help in concealing the evader when possible.

(3) Every possible feature of the terrain should be used for concealment.

(4) A silhouette against the sky should be avoided.

(5) Every effort should be made to reduce tone contrast and eliminate shine.

(6) Evaders should be especially careful at night due to infrared and low light detection equipment which may be used by the enemy. Keeping close to the ground and using terrain masking for concealment provide the best protection.

28-10. Camouflage Techniques for Shelter Areas:

a. As used in this section, the word shelter refers to the concept of personal protection synonymous with the terms refuge, haven, or retreat. Readers should not visualize the word "shelter" when mentioned in this text to mean a dwelling traditionally occupied by survivors in nontactical situations; such as, tents, cabins, or other such places of habitation. While it is true that these structures can and do provide safety and relief, it must

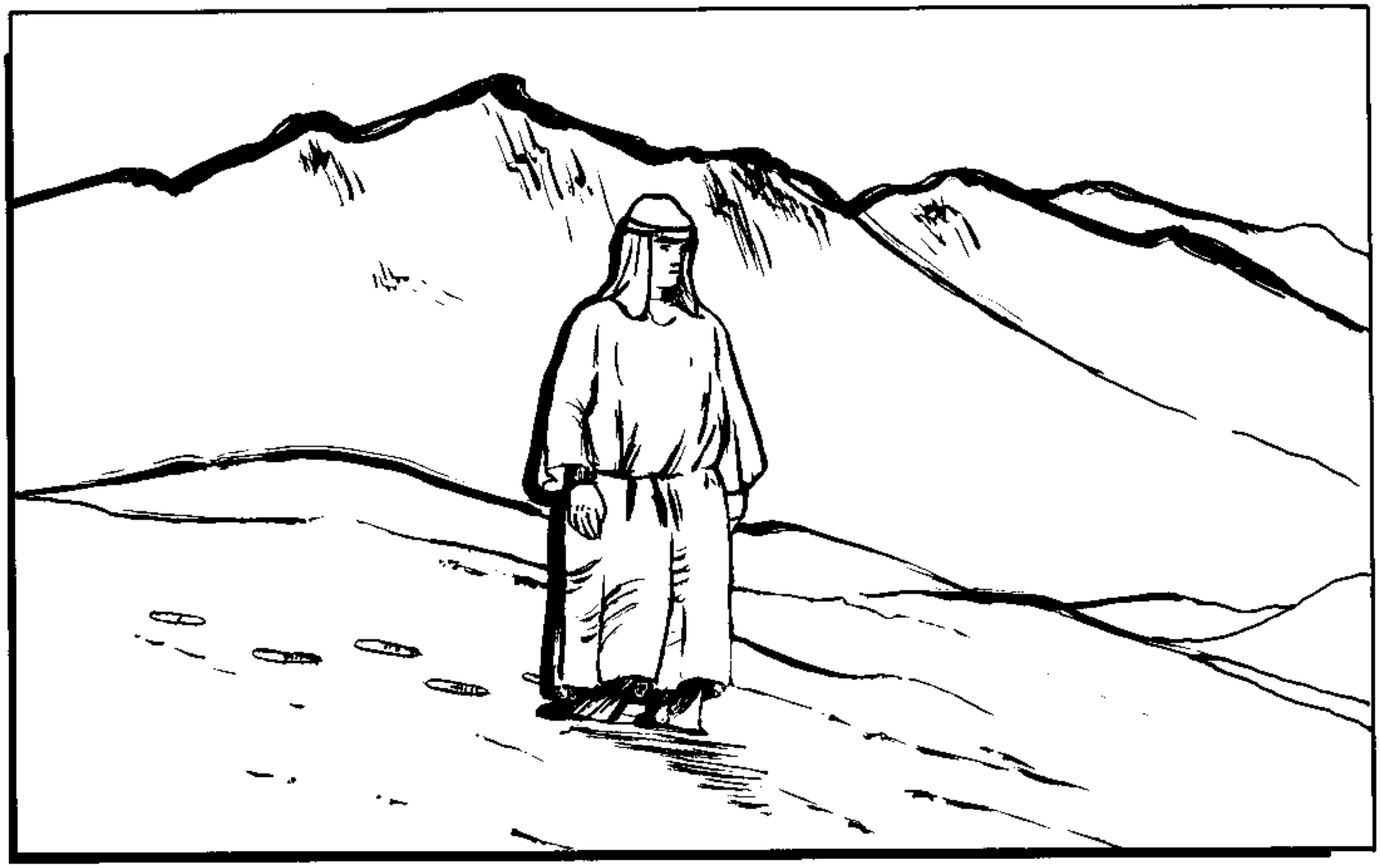

Figure 28-25. Evading in Snow-Covered Areas.

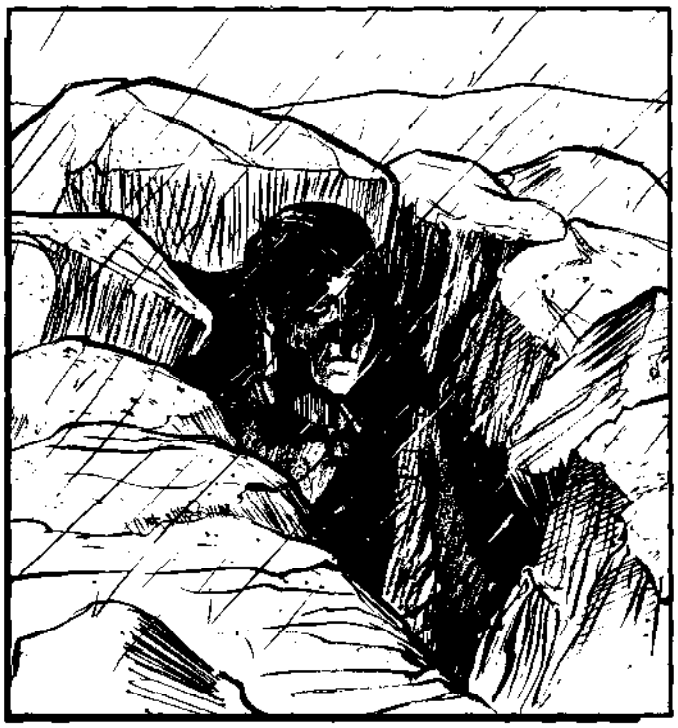

Figure 28-26. Shelter.

be understood that where an evasion situation is concerned, concealment, not personal comfort or convenience, will be the primary concern for the evader who seeks "shelter" or sanctuary (figure 28-26).

b. Besides resting or sleeping, there are a number of reasons why survivors may need to conceal themselves for varying amounts of time. They may need to take care of problems concerning personal hygiene, adjustment of clothing, maintenance or alteration of camouflage, triangulation for position determination, food and water procurement, etc. Concealment cannot be overstressed in respect to the areas survivors (evaders) may select for shelter. One important reason evaders must be able to select secure areas for refuge is to avoid and prevent detection by the enemy. This is especially important if the haven selected is to be used for resting or sleeping. Anyone who is resting or sleeping will not be totally alert, and added precautions may be necessary to maintain security. Another factor to consider, since evaders are also survivors, is to protect themselves from the elements as much as possible.

c. At no time will evaders be able to safely assume they are free from the threat of either ground or aerial observation. Therefore, not only is the shelter area and type determined by the needs of the moment (enemy

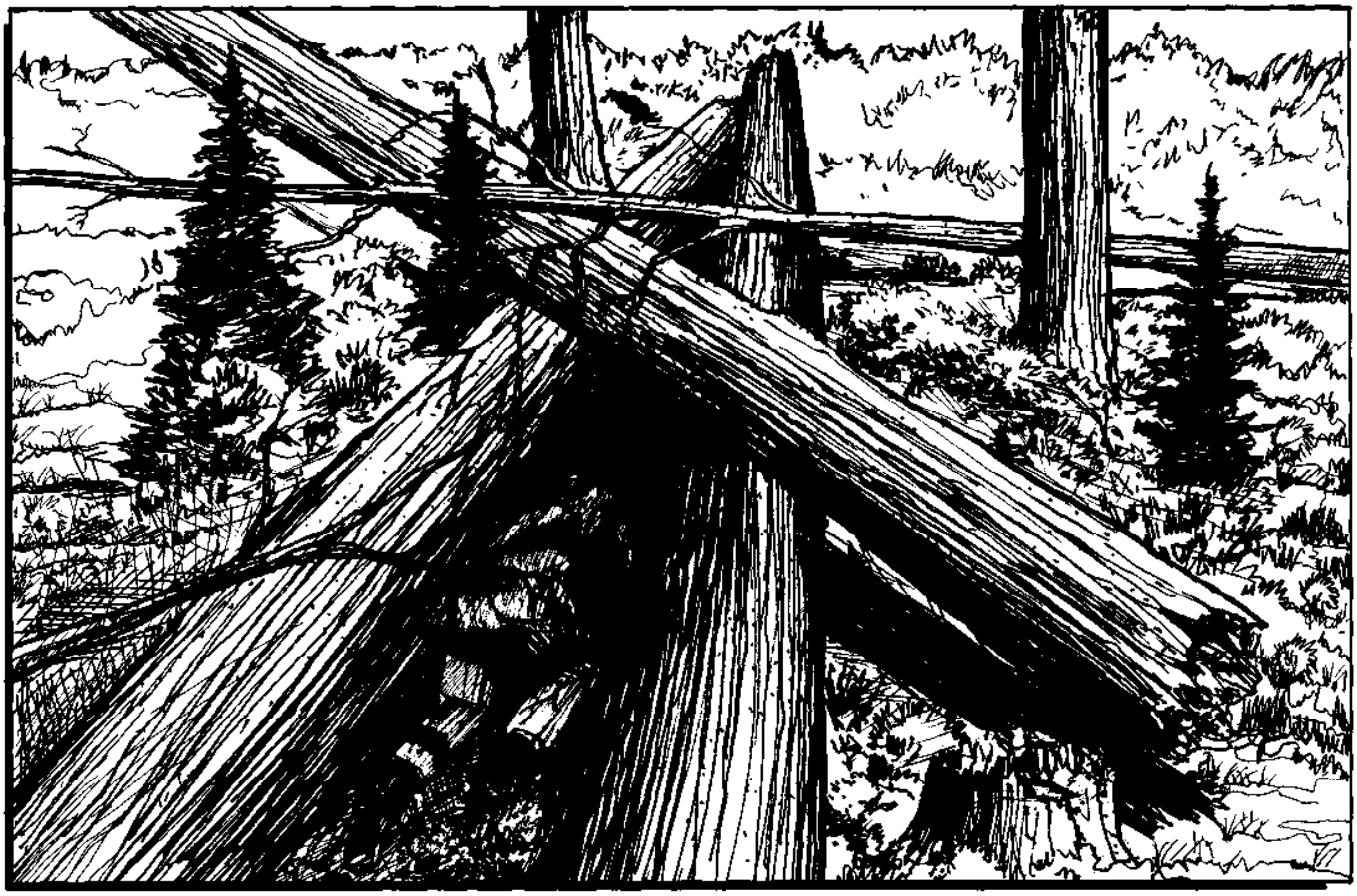

Figure 28-27. Natural Shelter.

presence, etc.), but consideration must also be given to the terrain and climatic conditions of the area. Evaders must constantly be aware of how long they may have to remain in the area and, most importantly, what type of enemy observation may be employed.

d. The shelters may be naturally present, or they may be those which are "assembled" and camouflaged by the evaders. Full use must be made of concealment and camouflage no matter what types of shelter areas are selected. The use of the natural concealment afforded by darkness, wooded areas, trees, bushes, and terrain features are recommended; however, any method used for disguise or hiding from view will increase the chances for success. There is much for evaders to consider concerning the many facets of evasion shelter site selection if they expect to establish and maintain the security of their area (figure 28-27).

e. Evaders should locate their shelter areas carefully. They should choose areas which are least likely to be searched. They must be in the least obvious locations. The chosen areas should look typical of the whole environment at a distance. They should not be near prominent landmarks. Areas that look bland get a cursory glance. The areas should also be those least likely to be searched; for example, rough terrain and thickly vegetated areas. The shelter sites should also be situated so that in the event of impending discovery, the evaders will be able to depart the area via at least one concealed escape route. The shelter areas should never be in areas which may trap them if the enemy discovers the places of concealment.

f. Evaders should choose natural concealment areas—a "natural shelter." Examples include small, concealed caves, hollow logs, holes or depressions, clumps of trees, or other thick vegetation (tall grass, bamboo, etc.). The site should have as much natural camouflage as possible. There should be cover on all sides; this includes natural formations or vegetation which can also protect evaders from aerial observations. The site should be as concealed as possible with a minimum of work. Sites chosen this way will make concealment easier and require less activity and movement. This is most important if the evaders are close to population centers or if the enemy is present.

g. The evaders should attempt to stay as high as possible and to select concealment sites near the military crest of a hill if cover is available. Noises from ridge to ridge tend to dissipate. Whispers or other sounds made in a valley tend to magnify as listeners get further up a hill. Shelter areas located on a slope are subject to higher daytime and lower nighttime winds, thereby minimizing the chance of detection through the sense of smell.

h. If possible, evaders should be in such a position in the shelter area that shadows will fall over the side of the area throughout the day. This can best be done in heavy brush and timber.

i. Evaders should try to locate alternate entrance and exit routes along small ridges or bumps, ditches, and rocks to keep the ground around the shelter area from becoming worn and forming "paths" to the site. They should avoid staying in one area so long that it has the appearance of being "lived in." Evaders should try to stay away from and out of sight of any open areas; examples are roads and meadows. Several miles distance from those may be desirable.

j. Waterways such as lakes, large rivers, and streams, especially at the junctions, are dangerous places. Power and fence lines or any prominent landmarks may indicate places where people may be. Evaders will want to stay clear of these areas. The enemy may patrol bridges frequently. Evaders should avoid any areas close to population centers. The evaders should be able to observe the enemy and their movements and the surrounding country from this hiding area if at all possible. If any assemblage of camouflage is necessary at the shelter site, evaders should keep in mind that they should always "construct" to blend. They should match the shelter area with natural cover and foliage, remembering that over-camouflage is as bad as no camouflage. Natural materials should be taken from areas of thick growth. Any place from which materials have been taken should be camouflaged. The following is an easy to remember acronym (BLISS) for evasion shelter principles:

B - Blend.
L - Low silhouette.
I - Irregular shape (outline).
S - Small.
S - Secluded location.

k. Other facilities evaders may use, such as latrines, caches, garbage pits, etc., must be located and camouflaged in the same manner as the shelter sites. Evaders should avoid forming a line of installations which lead from point to point to their location. They should dogleg through ground cover to use concealment to its best advantage. A dry, level sleeping spot is ideal, but the ideal spot to provide nonvisibility and comfort may be difficult to find. Evaders must have patience and perseverance to stay hidden until danger has passed or until they are prepared and rested enough to safely move on. They must be constantly on the watch for shelter areas which need little or no improvement for camouflage, protection from the elements, or security.

28-11. Firecraft Under Evasion Conditions:

a. Whether or not to build a fire under evasion conditions is indeed a difficult decision evaders must make at times. Basically, fire should only be used when it is absolutely necessary in a life or death situation. Potential evaders must understand that the use of fire can greatly increase the probability of discovery and subsequent capture.

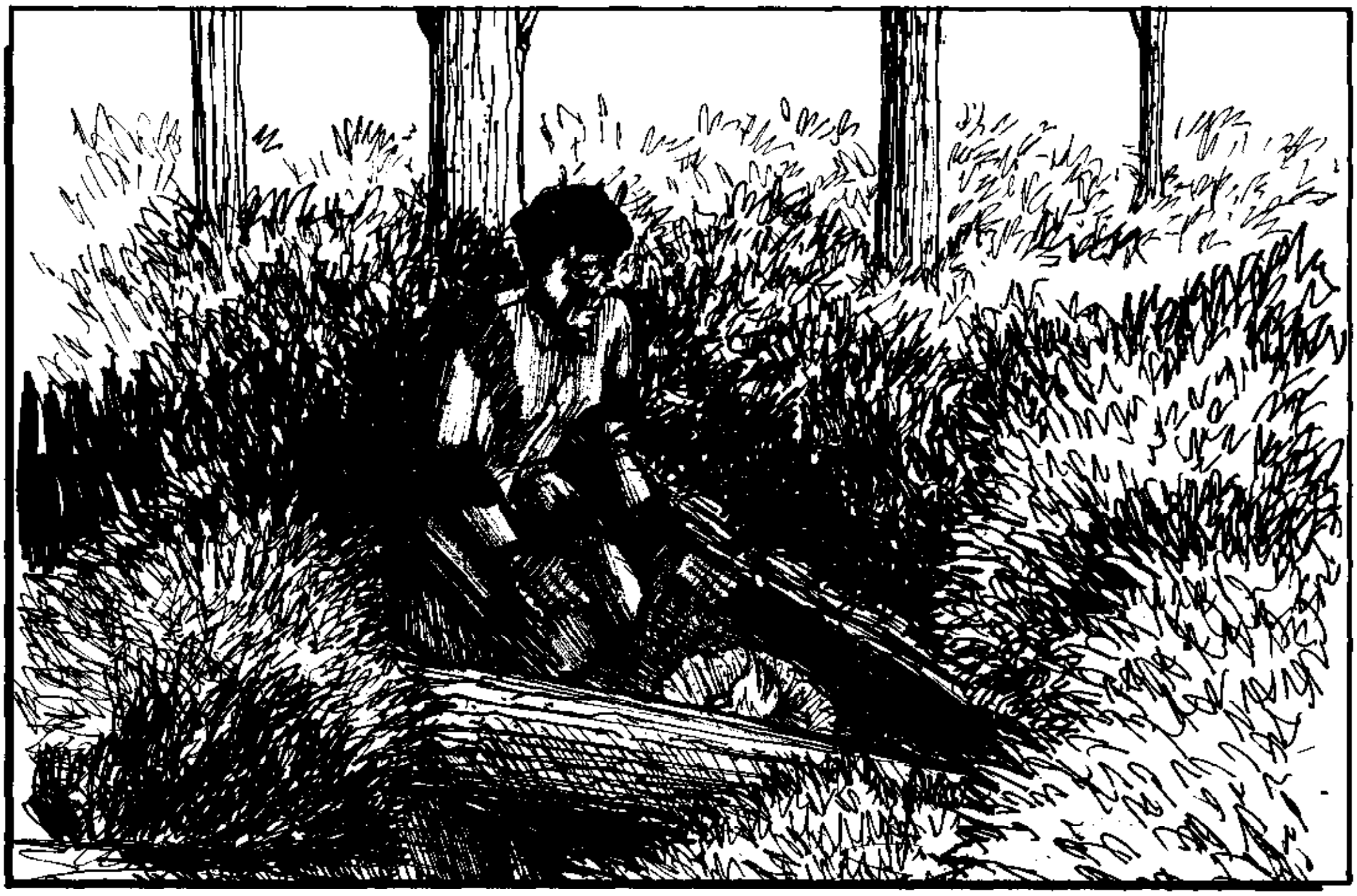

Figure 28-28. Fire.

b. If a fire is required, location, time, selection of tinder, kindling, and fuel, and construction should be major considerations.

(1) Evaders should keep the fire as inconspicuous as possible. The location of an evasion fire is of primary importance. All of the small evasion-type fires must be built in an area where the enemy is least likely to see them. If possible, in hilly terrain with cover, the fire should be built on the side of a ridge (military crest). No matter where the fire is built, it should be as small and smokeless as possible (figure 28-28).

(2) Fires are easier to disguise and will blend in better during the times of dawn and dusk and during times of bad weather. At these times, there is a haze or vapor trap that hinges in and around hills and depressions and is prevalent on the horizon. Any smoke from the fire will be masked by this haze in the early morning and at sunset. This is the time when the local populace is most likely to have their cooking or heating fires lit. Another method of disguising the smoke from a fire is to build it under a tree. The smoke will tend to dissipate as it rises up through the branches, especially if there is thick growth or the boughs are low hanging. If this is not possible, it will be helpful to camouflage the fire with earthen walls, stone fences, bark, brush, or snow mounds to block the light rays and help disperse the smoke.

(3) The best wood to use on an evasion fire is dry, dead hardwood no larger than a pencil with all the bark removed. This wood will produce more heat and burn cleaner with less smoke. Wood that is wet, heavy with pitch, or green will produce large amounts of smoke. When the wood has been gathered, evaders should select small pieces or make small pieces out of the wood collected. Small pieces of wood will burn more rapidly and cleanly thus reducing the chance of smoldering and creating smoke. The wood selected should be stacked so the fire gets plenty of air as ventilation will make the fire burn faster with less smoke.

c. One type of evasion fire which has the capability of being inconspicuous is called the Dakota Hole Fire (figure 28-29). After selecting a site for the fire hole, a "fireplace" must be prepared. This is done by digging two holes in the ground, one for air or ventilation and the other to actually lay the fire in. These holes should be roughly 8 to 12 inches deep and about 12 inches apart with a wide tunnel dug to connect the bottoms of the holes. The depth of the holes depends on the intended use of the fire. Place dirt on a piece of cloth so it can be used to rapidly extinguish the fire and conceal the fire

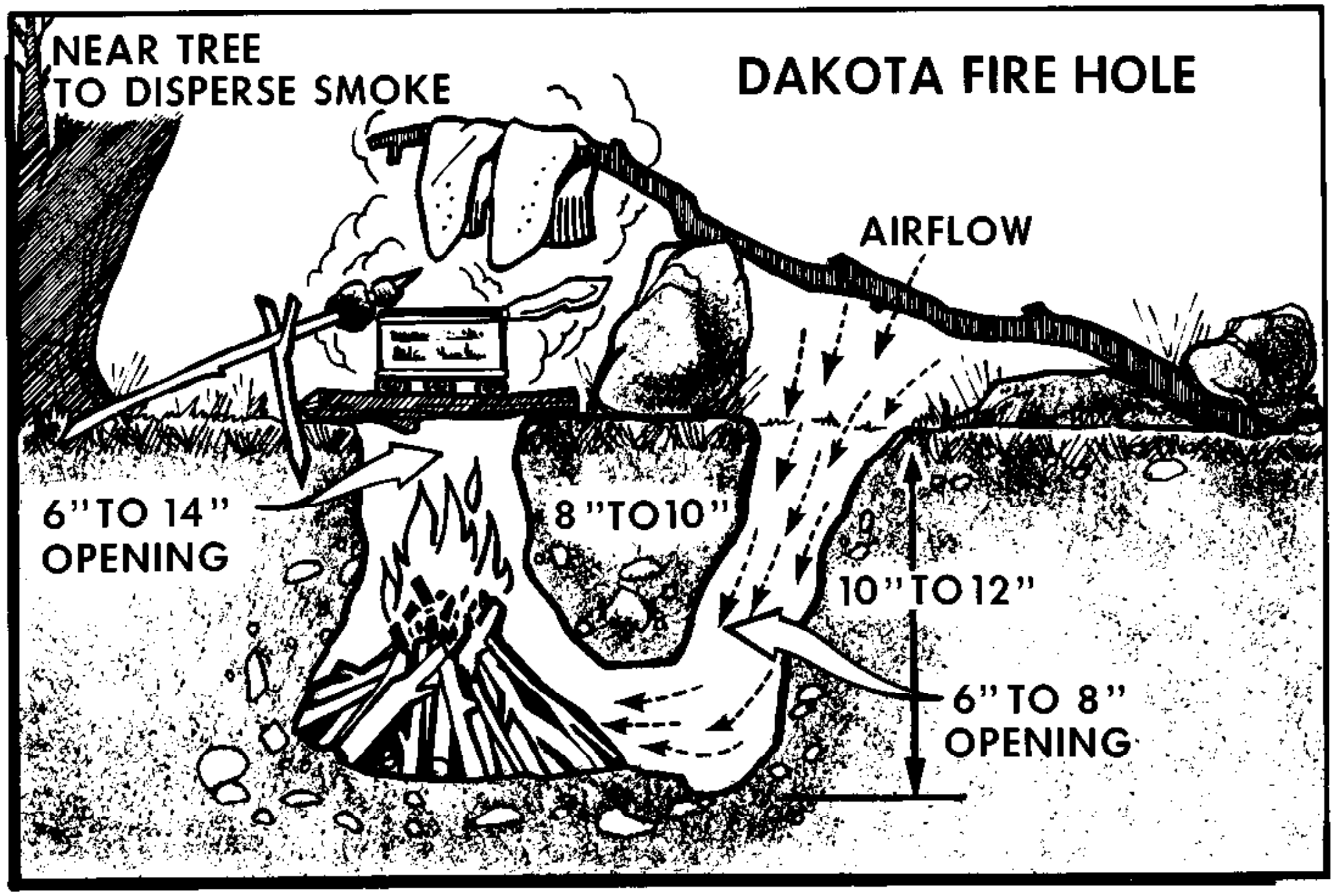

Figure 28-29. Dakota Hole.

site. Evasion fires should be small. Evaders who build these fires should strive to keep the flame under the surface of the ground. Initially the fire may appear to be smoking due to the moisture in the ground. At night the area may glow, but there should be no visible flame. A fire built in a hole this way will burn fast as all the heat is concentrated in a small area. This is a good type of fire for a single evader as opposed to many persons taking turns cooking over this one hole. Everything about the evasion situation will have to be examined before deciding which fire configuration would be most useful if a fire must be built.

d. Another good evasion fire is the trench fire. This fire is built by digging below the earth's surface 8 to 12 inches in an elongated pattern. The length depends on how many people need to use it. This is a fire more suited to meeting the needs of a small group. The fire should not crowd either end of the excavation, as it must be able to "draw" an adequate amount of air to help it burn hot and eliminate smoke (figure 28-30).

e. An evasion fire can also be built just below the ground cover (figure 28-31). Here the emphasis is on quick concealment of the area. It should be kept in mind, however, that some type of screen should be built

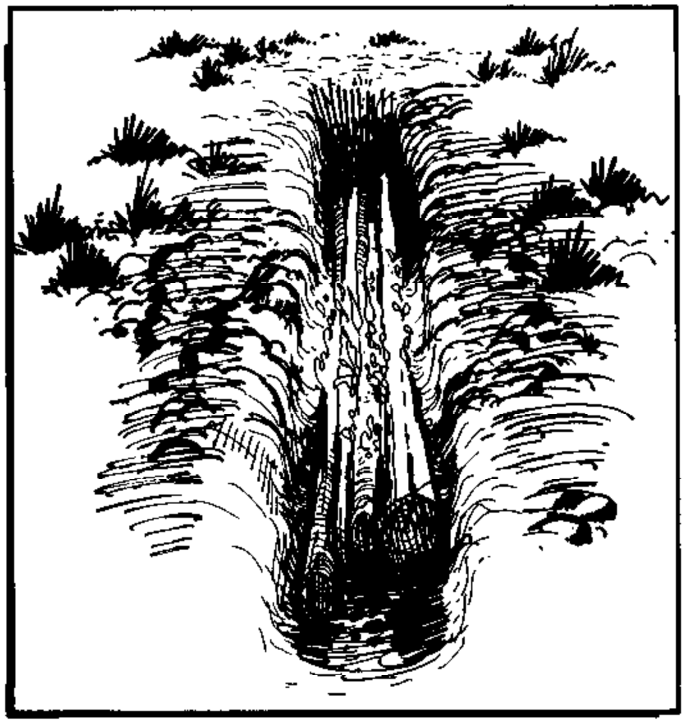

Figure 28-30. Trench Fire.

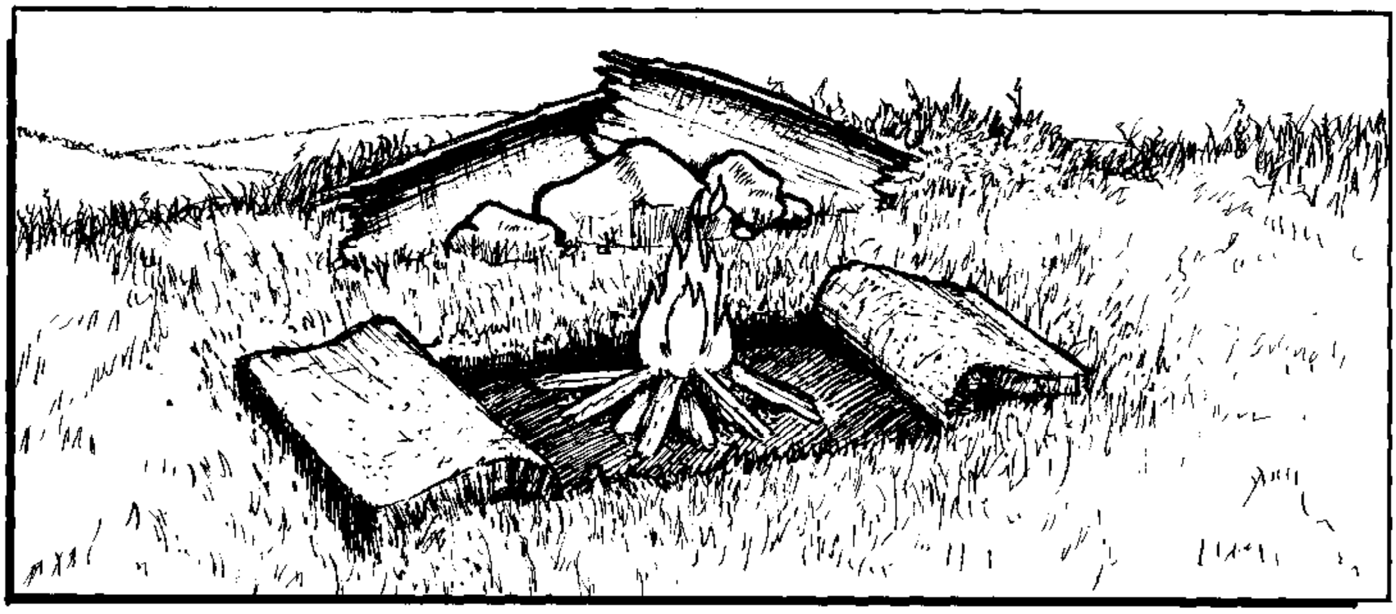

Figure 28-31. Fire Concealment.

to hide the flames. The small fire is built on the bare ground after a layer of sod or earth has been sliced and rolled back. After use, evaders simply scatter the fire remnants over the bare area and roll or fold the piece of ground back into place.

f. If evaders are in areas where holes can't be dug or sod can't be lifted, they will have to settle for some type of screen around the fire. They should also keep the fire small and finish using it quickly.

g. All traces of the fire should be removed. Unburned firewood should be buried. Holes should be totally filled in. Placing the soil on a holder, such as a map or piece of equipment, will aid in replacing it. This way there will be no leftover dirt patches on the ground after the holes were filled in. Once evaders feel all available measures have been taken to obliterate any leftover evidence, they should move out of the area if possible. Since evaders can never be positive the fire wasn't detected, they must assume it was spotted and take all necessary precautions.

28-12. Sustenance for Evasion:

a. As previously stated in this section, not only do evaders face the problem of remaining undetected by the enemy, they must also have the knowledge which will enable them to "live off the land" as they evade. They must be prepared to use a wide variety of both "wild" and domestic food sources, obtain water from different sources, and use many methods for preparing and preserving this food and water (figure 28-32).

b. No matter what the circumstances of the evasion situation are, evaders should never miss an opportunity to obtain food. Ordinarily food will be obtained from wild plant and animal sources. If possible, evaders should stay away from domestic plants (crops) and animals. Using wild animals and plant sources for food will reduce the probability of capture.

c. Animal foods are a prime source of sustenance for evaders, having more nutritional value for their weight than do plant foods. Evaders may obtain enough animal food in one place to last for several days while they travel.

d. There are several ways which evaders may procure animal foods (trapping or snaring, fishing, hunting, and poisoning). A few modifications should be considered regarding the use of traps and snares in a tactical situation even though the same basic principles apply in both tactical and nontactical environments.

(1) Because small game is more abundant than is large game in most areas, evaders should confine traps and snares to the pursuit of small game. There are other advantages to restricting trap and snare size. Evaders will find it is easier to conceal a small trap from the enemy. Small animals make less noise and create less disturbance of the area when caught.

(2) Conversely, there are also a few disadvantages pertaining to the use of snares or traps during evasion episodes. Two disadvantages are: Evaders must remain in one place while the snares are working, and there may be some disturbance of the area where materials have been removed.

e. Fishing is another effective means of procuring animal food. Fish are normally easy for evaders to catch, and they are easy to cool.

(1) There are several methods which evaders may use to catch fish. A simple hook and line is one of these methods; another is a "trotline." Evaders may construct a trotline by fixing numerous hooks to a pole and by

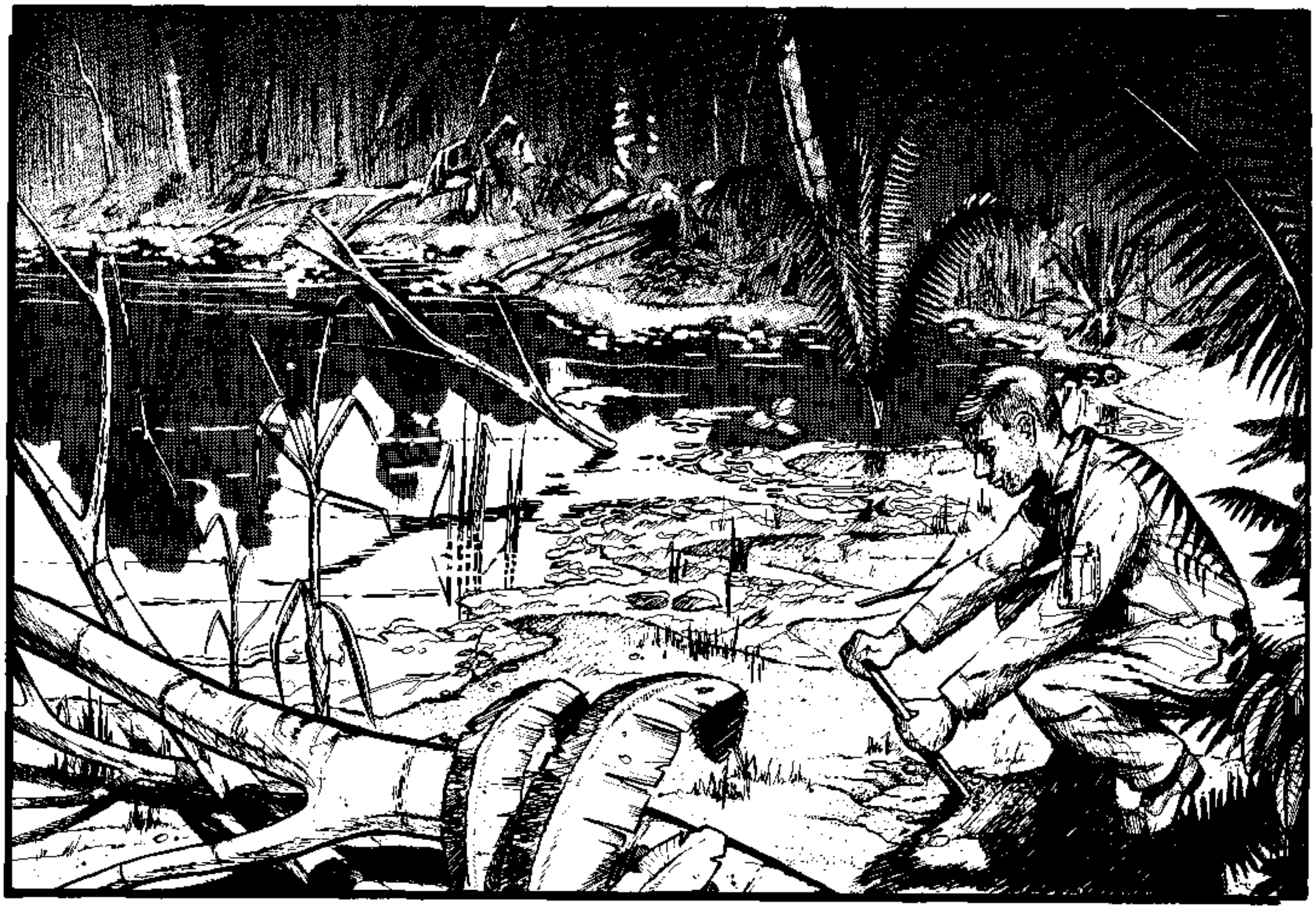

Figure 28-32. Sustenance for Evasion.

sliding it into the water from a place of concealment (figure 28-33).

(2) Nets and traps may also be used; however, they should be set below the water line to avoid detection. Spearing is another option. Here again, exposure in open waterways can be very dangerous to the evader.

(3) "Tickling" the fish (figures 28-34, 28-35, and 28-36) is also effective if evaders can remain concealed. This method requires no equipment to be successful. The main disadvantage to fishing is people live by water bodies and travel on them. This greatly increases the chances of being detected. (NOTE: Caution should be used when tickling fish in areas with carnivorous fish or reptiles.)

f. Weapons may be used by evaders to procure animals. The best weapons are those which can be operated silently, such as a blowgun, slingshot, bow and arrow, rock, club or spear. These should be used primarily against small game. One major advantage of using weapons is game can be taken while evaders travel. Because of noise, firearms should never be used in an evasion situation.

g. Plant foods are very abundant in many areas of the world. Evaders may be able to procure plant food types that require no cooking. One advantage of procuring plant foods during evasion is that by collecting natural fruits and nuts, evaders can remain deep in unpopulated areas. In some areas, it is possible to find old garden plots where vegetables may be obtained. When possible, select foods which can be eaten raw. (Refer to chapter 18 - Food.)

(1) The disadvantage of plant food procurement is that evaders may not be the only ones looking for food. The natives of the country could also be out looking for food. If natives know of a good area, they may visit that place many times. If evaders have been in the area, their presence could be discovered.

(2) Some other considerations and methods concerning plant food procurement are as follows. Evaders should:

(a) Never take all the plants or fruits from one area.

(b) Pick only a few berries off of any one bush.

(c) When digging plants, take only one plant, then move on some distance before digging up another.

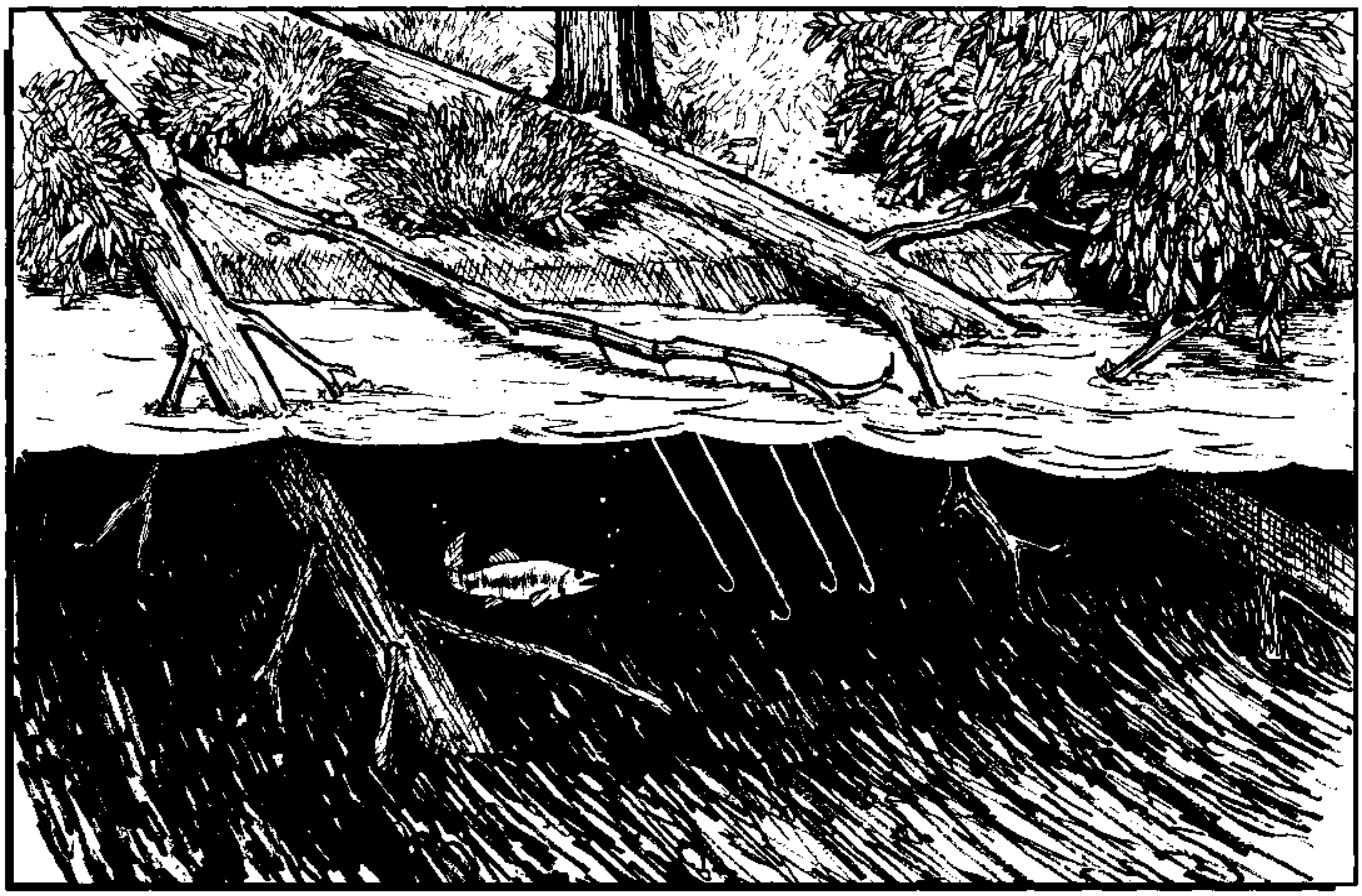

Figure 28-33. Fishing.

(d) When digging plants from old garden plots, make sure the plot is old. In many countries the people plant their crops and do not return to the plot until harvest time.

(e) Camouflage all signs of presence.

(3) Most domestic foods must be procured by theft which is very dangerous. However, if proper methods are used and the opportunity presents itself, plants and animals may be stolen. The main reason thieves are captured is the boldness they display after several successful thefts. The basic rules of theft are:

(a) If at all possible, the theft should take place at night.

(b) Evaders should thoroughly observe the area of intended theft from a safe vantage point.

(c) Evaders should find the vantage points just before dusk and look the place over to make sure everything is the same as it was the last time a check was made.

(d) Evaders should check for dogs, which could be a big hazard. Barking draws attention; also, some dogs are vicious and can harm evaders. Besides dogs, other animals or fowl can alert the enemy to the evaders' position.

(e) Evaders should never return twice to the scene of a theft.

(f) Every theft should be planned, and after its accomplishment, evaders should leave no evidence of either their presence or the theft itself (figure 28-37).

(g) Only small amounts should be taken (figure 28-38).

(4) When evaders find it necessary to take cultivated plant foods, they should never take the complete plant. Taking plants from the inside of the field, not the edge, and leaving the top of plants in place may help conceal the theft.

h. The rules of theft also apply to taking domesticated animals. The evader should concentrate on animals that don't make much noise. If a choice has to be made as to which animal to steal, evaders should take the smallest one.

i. Water is very essential, but it can be difficult to acquire (figure 28-39).

(1) When procuring water, evaders should try to find small springs or streams well away from populated areas. The enemy knows evaders need water and may check all known water sources. No matter where water is procured, evaders should try to remain completely

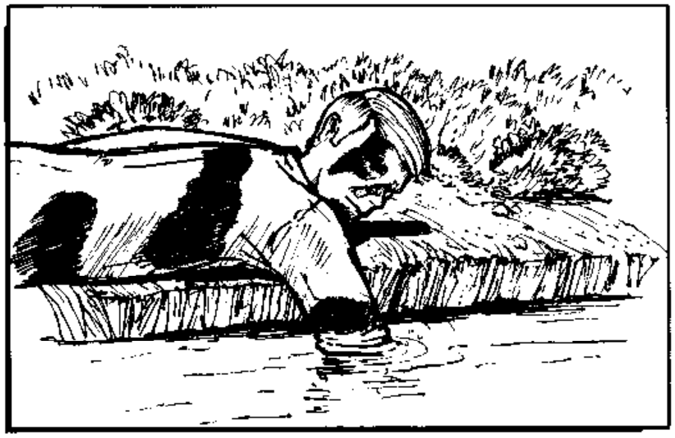

Figure 28-34. Fish Tickling - A.

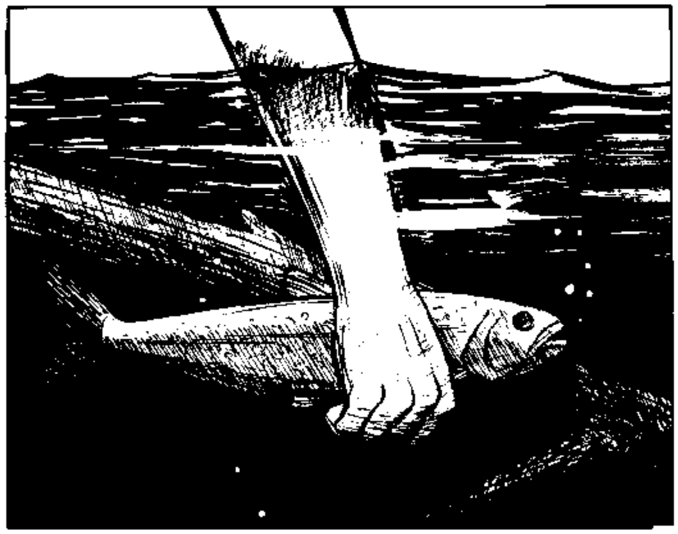

Figure 28-35. Fish Tickling - B.

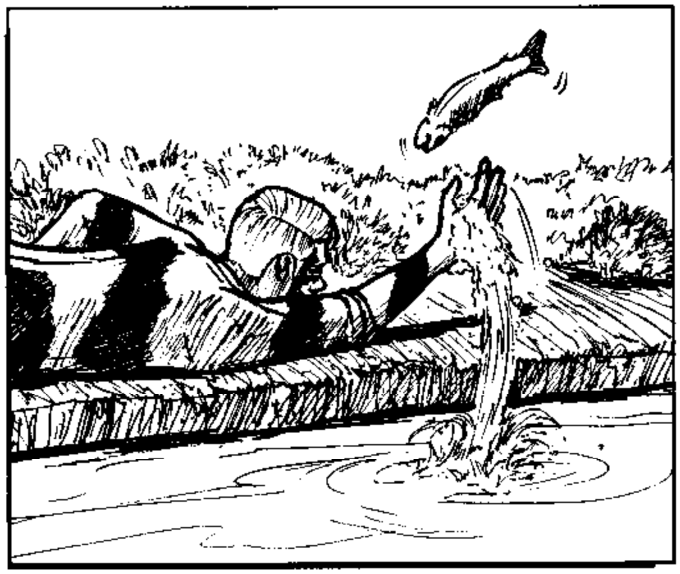

Figure 28-36. Fish Tickling - C.

concealed when doing so. The area around the water source should be observed to make sure it isn't patrolled or watched. While obtaining water and when leaving the area, evaders should conceal any evidence of their presence. Good sources of water include trapped rainwater in holes or depressions and plants which contain water.

(2) The preparation and purification of food and water by cooking is a precaution which should be weighed against the possibility of capture. It might be necessary to eat raw plant and animal foods at times. Some plant foods will require cutting off thorns, peeling off the outer layer, or scraping off fuzz before consumption. Raw animal foods may contain parasites and micro-organisms which may not effect evaders for days, weeks, or months, at which time, hopefully, they will be under competent medical care. If the environment in which the parasites live is altered by cooking, cooling, or drying, the organisms may be killed. Meat can be dried and cooled by cutting into thin strips and air-drying. Salting makes the meat more palatable. If the meat must be cooked, small pieces should be cooked over a small hot fire built in an unpopulated area. The best methods of preserving food during evasion are drying or freezing.

(3) In an evasion situation, boiling water for purification should not be used as a method except as a last resort. Iodine tablets are the best method of purification. If evaders do not have purification tablets, and the danger of the enemy detecting their fires is too great, evaders may have to forego purification. The only problem with this is if water is not purified, it may cause vomiting and (or) diarrhea. These ailments will slow down the evaders and make them susceptible to dehydration. Aeration and filtration may help to some degree and are better than nothing. If water cannot be purified, evaders should at least try to use water sources which are clean, cold, and clear. Rain, snow, or ice should be used if available.

28-13. Encampment Security Systems:

a. When evading for extended periods in enemy-held territories, it becomes essential for evaders to rest. To rest safely, especially if in a group, it is essential to devise and use some sort of early warning system to prevent detection and unexpected enemy infiltration. When establishing an evasion "shelter" area, there are certain things which should be done (day or night) for security purposes (figure 28-40).

b. Evaders should scout the area around their encampment for signs of people. They should pay particular attention to crushed grass, broken branches, footprints, cigarette butts, and other discarded trash. These signs may reveal identity, size, direction of travel, and time of passage of an enemy force. If large numbers of these signs are present, the evader should consider moving to a more secure area.

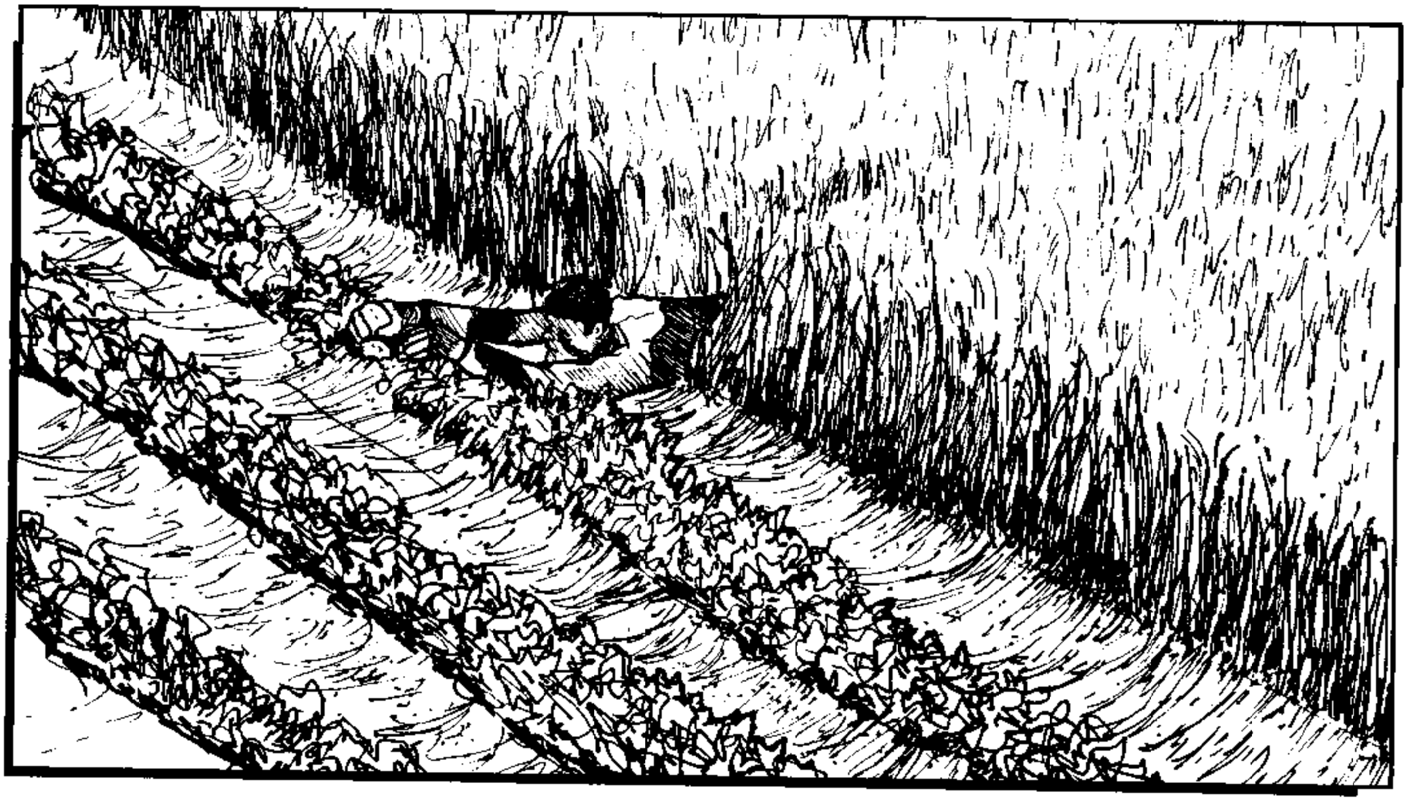

Figure 28-37. Stealing Vegetables.

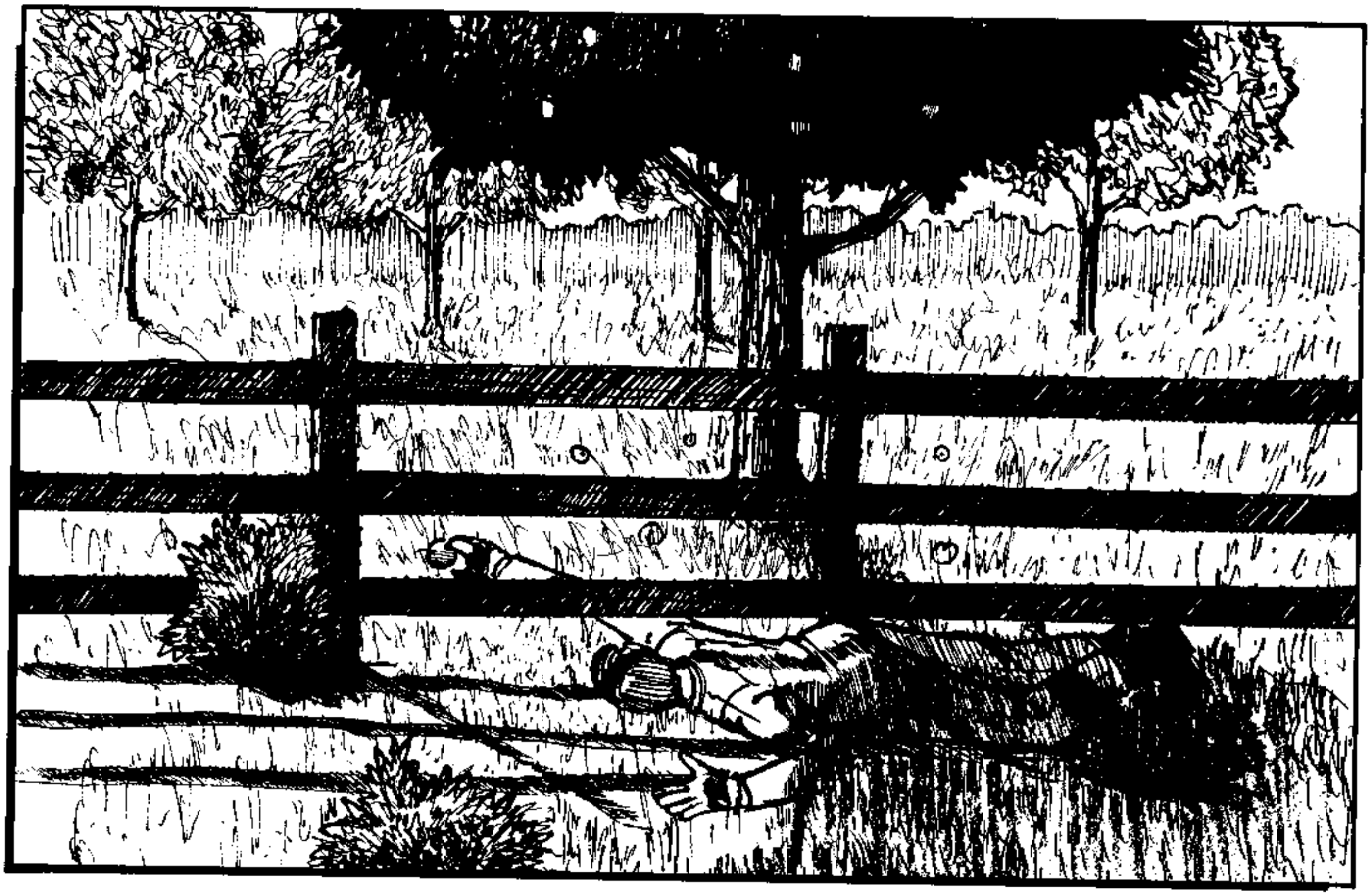

Figure 28-38. Stealing Small Amounts of Food.

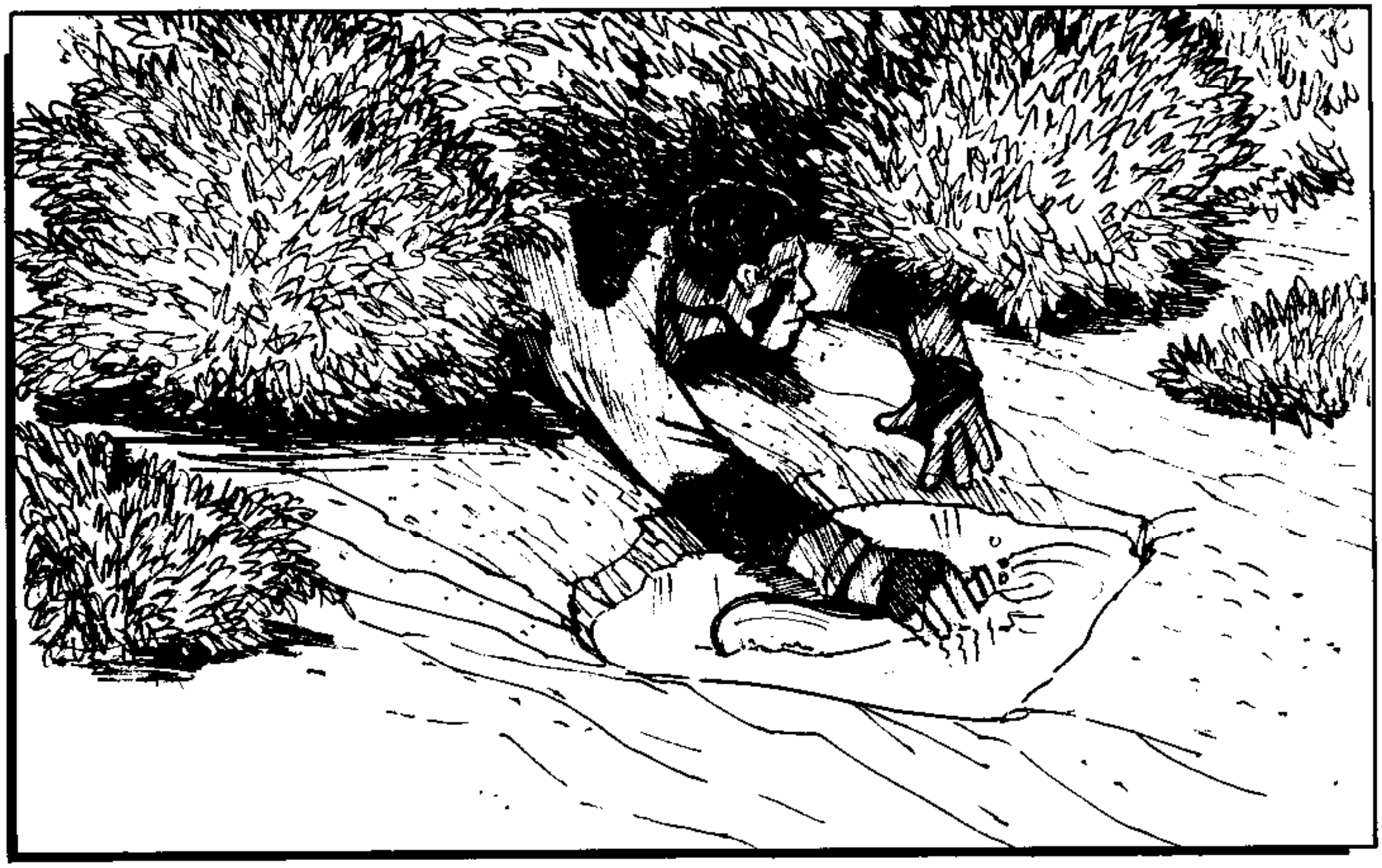

Figure 28-39. Procuring Water.

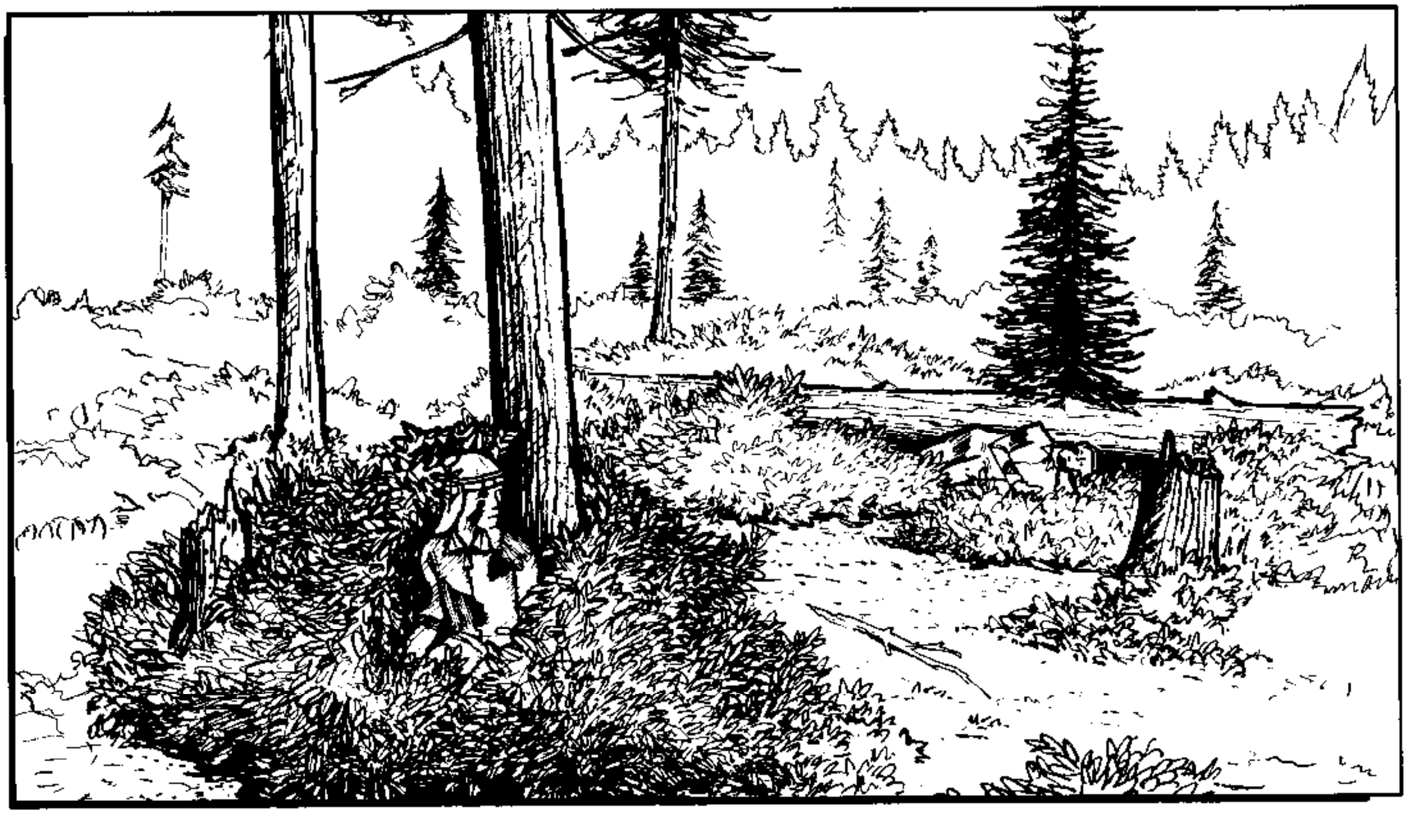

Figure 28-40. Rest Safely.

c. Once the camp area has been determined to be fairly secure, some type of alarm system must be devised. For a lone evader, this may consist of actually constructing wire or line with sound-producing devices attached. However, this system works for the enemy as well and may prove to alert enemy forces to the evader's presence. A lone evader should use the natural alarm system available. Disturbances in animal life around an evader may indicate enemy activity in the area. Group situations may allow for more security. Two or more evaders may use lookout(s) or "scout(s)" at observation posts strategically located around the encampment.

d. Readiness is another aspect of security. The evader should be aware that, at any time, the shelter area may be overrun, ambushed, or security compromised, making it necessary to vacate the area. If evaders are in a group and future group travel is desired, it is essential that everyone in the group knows and memorizes certain things; such as, compass headings or direction of travel, routes of travel, destination descriptions, and rally points (locations where evaders regroup after separation). Alternate points must be designated in case the original cannot be reached or if it is compromised by enemy activity.

(1) Once everyone in the evader's group has reached the final destination, alternate point, or rally point, a new emergency evacuation and rendezvous plan must be established.

(2) Evaders should always be aware of the next rally point, its location, and direction. These places, which provide concealment and cover, should be designated along the route in case an enemy raid or ambush scatters the group. There should be a rally point for every stage of the journey. Even when approaching the supposed "final destination" of the day, evaders should have an evacuation plan ready.

e. Maintaining silence is a very important aspect of security. It is essential to be able to communicate with individual group members and scouts so that everyone is aware of what is going on at all times during evasion. Hand signals are the best method of communication during evasion as they are silent and easily understood. Instructions and commands which must be conveyed throughout the entire group are: (See figure 28-41 for examples of hand signals.)

(1) Freeze.
(2) Listen.
(3) Take cover.
(4) Enemy in sight.
(5) Rally.
(6) All clear.
(7) Right.
(8) Left.

28-14. Evasion Movement:

a. During evasion travel, the evader is probably most susceptible to capture. Many evaders have been captured as a direct result of their failure to use proper evasion movement techniques. Evasion movement is the action of a person who, through training, preparation, and application of natural intelligence, avoids capture and contact with hostiles, both military and civilian. Not only is total avoidance of the enemy desirable for evaders, it is equally important for evaders not to have their presence in any enemy controlled area even suspected. A fleeting shadow, an inopportune movement or sound, and an improper route selection are among a number of things which can compromise security, reveal the presence of evaders, and lead to capture.

b. One evasion situation will not be identical to another. There are, however, general rules which apply to most circumstances. These rules, carefully observed,

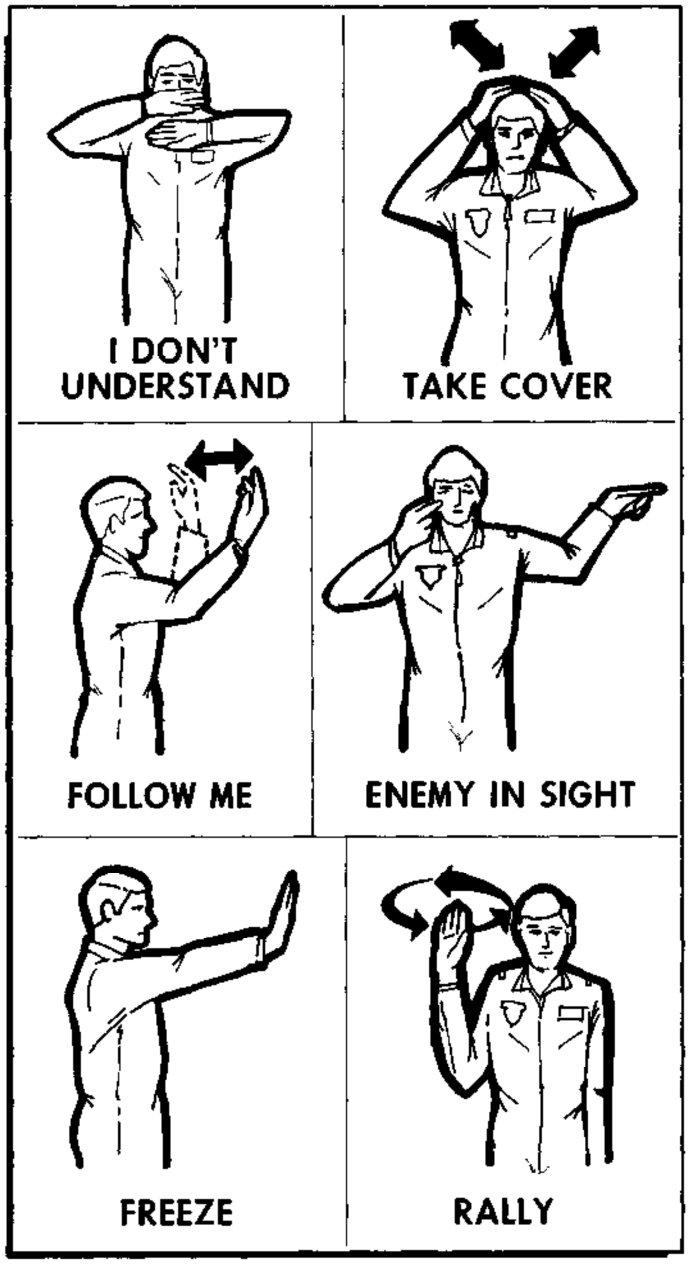

Figure 28-41. Hand Signals.

will enhance the evaders' chances of returning to friendly control.

(1) Evasion begins even before a crewmember leaves an aircraft over enemy territory. Two factors which are essential to successful evasion and return are opportuntiy and motivation. Pre-mission preparation and knowledge of areas of concern are very important to a crewmember. Pre-mission knowledge gained must be based on the most current information available through command area briefings, area studies, and intelligence briefings.

(2) Some areas of interest to the prospective evader are:

(a) Topography and Terrain. An aircrew member should know the physical features and characteristics, possible barriers, best areas for travel, availability of rescue, and the type of air or ground recovery possible. A future evader should also know the requirements for long-term unassisted survival in the area of operation.

(b) Climate. The typical weather conditions and variations should be known to aid in evasion efforts.

(c) People. A very critical consideration may be knowing the people in the area. From ethnic and cultural briefs read before the mission, crewmembers should familiarize themselves with the behavior, character, customs, and habits of the people. It may, at times during the evasion episode, be necessary to emulate the natives in these respects (figure 28-42).

(d) Equipment. Aircrew members who would hope to be successful evaders should be thoroughly familiar with all of their equipment and know where it is located. They should also preplan what equipment should be retained and what should be left behind under certain conditions of an evasion travel situation.

(3) Before addressing evasion movement techniques, it is appropriate to go over some of the factors which influence an evader's decision to travel.

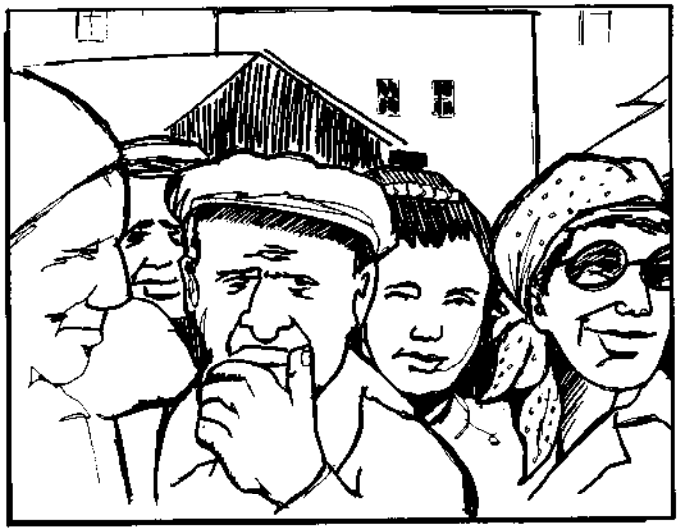

Figure 28-42. Knowing the People.

(a) The first few minutes after entering enemy or unfriendly territory is usually the most critical period for the evader. The evader must avoid panic and not take any action without thinking. In these circumstances, the evader must try to recall any previous briefings, standard operating procedures, or training and choose a course of action which will most likely result in return to friendly territory.

(b) In those first few minutes, there is a great deal for the downed crewmember to think about. Quick consideration must be given to landmarks, bearings, and distance to friendly forces and from enemy forces, likely location for helicopter landing or pickup, and the initial direction to take for evasion. Knowledge of what to expect is important because when circumstances arise which have been considered in advance, they can be carried out more quickly and easily. Evaders should try to adapt this knowledge and any skills they have to their particular situation. Flexibility is most important, as there are no hard and fast rules governing what may happen in an evasion experience.

(c) In most evasion situations, evaders will be required to move if for no other reason than to leave the immediate landing area when pickup is not imminent. Because any movement has the inherent risk of exposure, some specific principles and practices must be observed. Periods of travel are the phases of evasion when evaders are most vulnerable. Many evaders have been captured because they followed the easier or shorter route and failed to employ simple techniques such as watching and listening frequently and seeking concealment sites.

28-15. Searching Terrain:

a. Evaders should visually survey the surrounding terrain from an area of concealment to determine if the route of travel is a safe one. Evaders should first make a quick overall search for obvious signs of any presence such as unnatural colors, outlines, or movement. This can be done by first looking straight down the center of the area they are observing, starting just in front of their position, and then raising their eyes quickly to the maximum distance they wish to observe. If the area is a wide one, evaders may wish to subdivide it as shown in figure 28-43. Now all areas may be covered as follows: First, by searching the ground next to them. A strip about 6 feet deep should be looked at first. They may search it by looking from right to left parallel to their front. Secondly, by searching from left to right over a second strip farther out, but overlapping the first strip. Searching the terrain in this manner should continue until the entire area has been studied. When a suspicious spot has been located, evaders should stop and search it thoroughly (figure 28-44).

b. The evader must question the movement:

(1) Is the enemy searching for the evader?

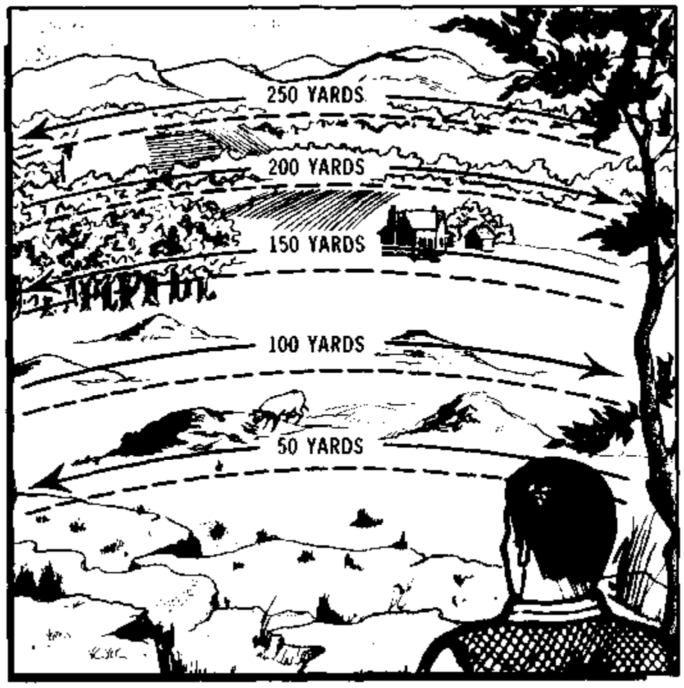

Figure 28-43. Viewing Terrain - A.

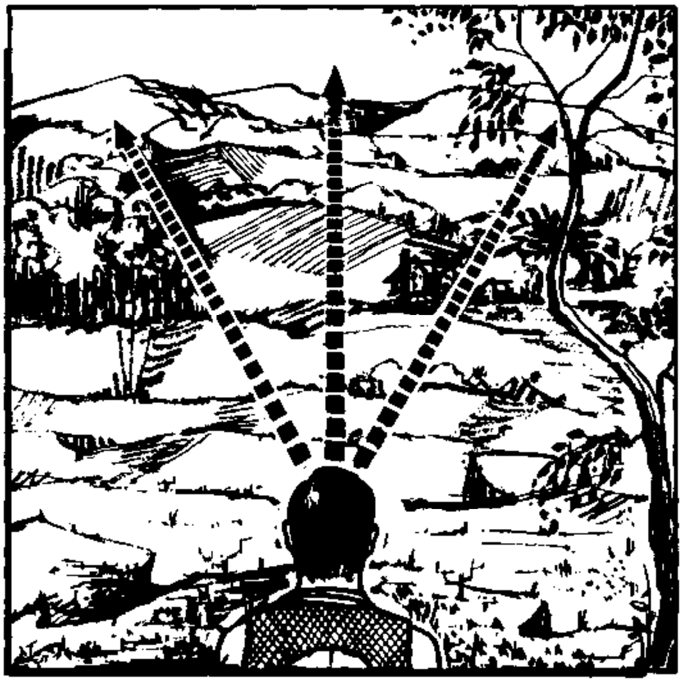

Figure 28-44. Viewing Terrain - B.

(2) What is the evader's present location?

(3) Are chances for rescue better in some other place?

(4) What type of concealment can be afforded in the present location?

(5) Where is the enemy located relative to the evader's position?

c. Having considered the necessity and risks of travel, evaders must:

(1) Orient themselves.

(2) Select a destination, alternates, and the best route.

(3) Have an alternate plan to cover all foreseeable events.

d. Cautious execution of plans cannot be overemphasized since capture of evaders has generally been due to one or more of the following reasons:

(1) Unfamiliarity with emergency equipment.

(2) Walking on roads or paths.

(3) Inefficient or insufficient camouflage.

(4) Lack of patience when pinned down.

(5) Noise or movement or reflection of equipment.

(6) Failure to have plans if surprised by the enemy.

(7) Failure to read signs of enemy presence.

(8) Failure to check and recheck course.

(9) Failure to stop, look, and listen frequently.

(10) Neglecting safety measures when crossing roads, fences, and streams.

(11) Leaving tracking signs behind.

(12) Underestimating time required to cover distance under varying conditions.

e. Evaders should understand progress on the ground is measured in stopover points reached. Speed and distance are of secondary importance. Evaders should not let failure to meet a precise schedule inhibit their use of a plan.

28-16. Movement Techniques Which Limit the Potential for Detection of an Evader (Single).

a. Evaders should constantly be on the lookout for signs of enemy presence. They should look for signs of passage of groups, such as crushed grass, broken branches, footprints, and cigarette butts or other discarded trash. These may reveal identity, size, direction of travel, and time of passage (figure 28-45).

(1) Workers in fields and normal activities in villages may indicate absence of the enemy.

(2) The absence of children in a village is an indication they may have been hidden to protect them from action which may be about to take place.

(3) The absence of young men in a village is an indication the village may be controlled by the enemy.

b. A knowledge of enemy signaling devices is very helpful. Those listed below are examples of some used by communist guerrillas in Southeast Asia.

(1) A farm cart moving at night shows one lantern to indicate no government troops are close behind.

(2) A worker in the fields stops to put on or take off a shirt. Either act can signal the approach of government troops. This is relayed by other informers.

(3) A villager fishing in a rice paddy holds a fishing pole out straight to signal all clear; up at an angle to signal the troops are approaching.

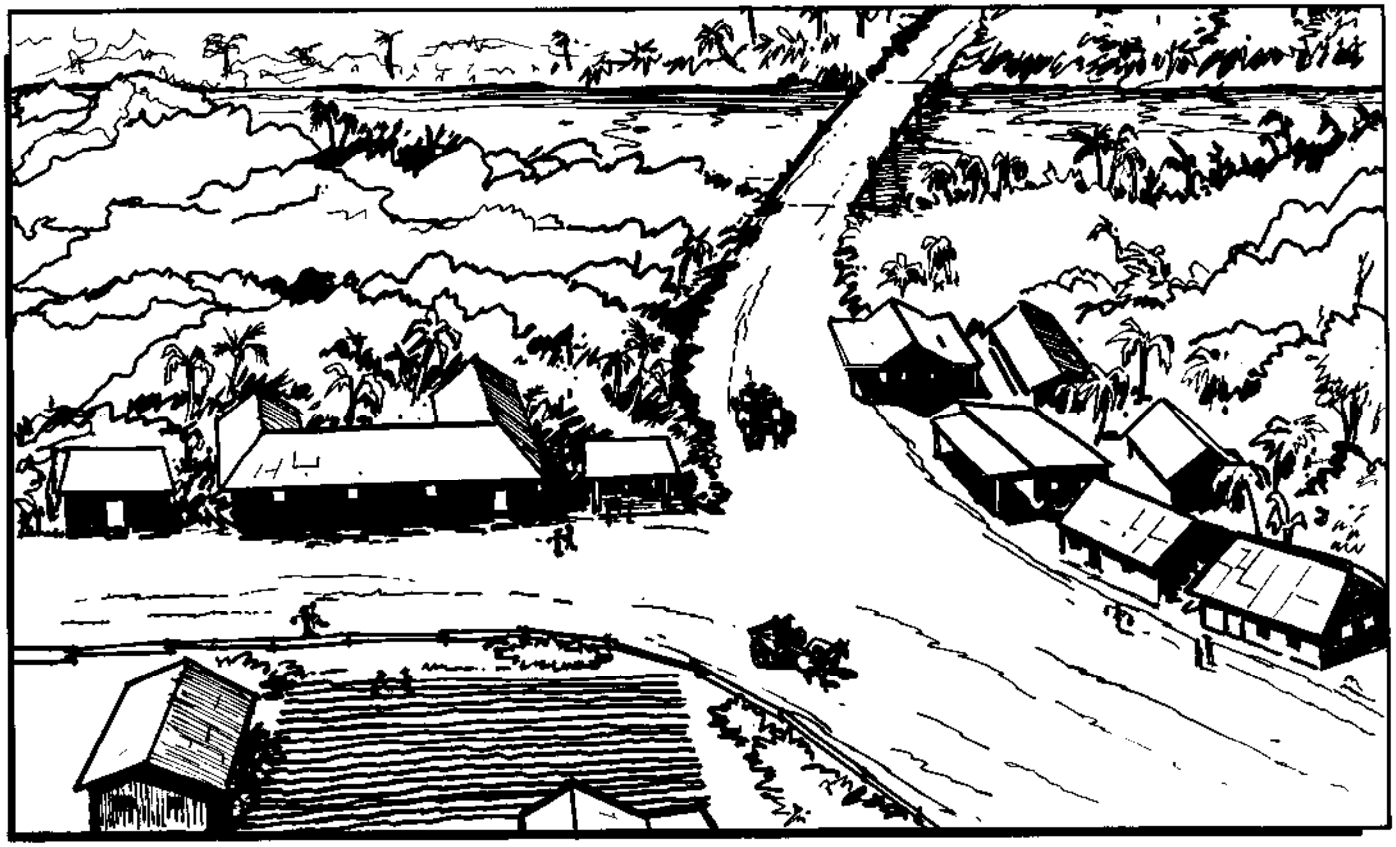

Figure 28-45. Signs of Passage.

c. The times evaders choose to travel are as critical as the routes they select. If possible, evaders should try to make use of the cover of darkness. Darkness provides concealment and in some cases there is also less enemy traffic. If, out of military necessity, the enemy is active during the hours of darkness, evaders may then find it wise to move in the early morning or late afternoon. Night movement is slower, more demanding, and more detailed than daylight movement; but it can be done. The alternative to night movement might be capture, imprisonment, and death. Evaders should then consider traveling under the cover of darkness first. However, if the enemy knows the evaders' position, or if the other factors dictate (terrain, vegetation, navigation consideration, etc.), other choices may have to be made. If travel is to be done during darkness, the terrain to be traversed should be observed during daylight if possible. While observing the area to be traveled, evaders should give attention to areas offering possible concealment as well as the location of obstacles they may encounter on their route. If the evader has a map, a detailed study of it should be made. However, it should be remembered that natural terrain features change with time of day and time of year. Certain features (ditches, roads, burned-off areas, etc.) may not be on the evaders' map if they are new. Such pretravel reconnoitering of an area will give evaders a head start on knowing how to adapt their travel movements and camouflage necessary from point to point (figure 28-46).

d. Evaders should try to memorize the routes they will take and the compass headings to their destinations. This information should not be written down on the map or on other pieces of material.

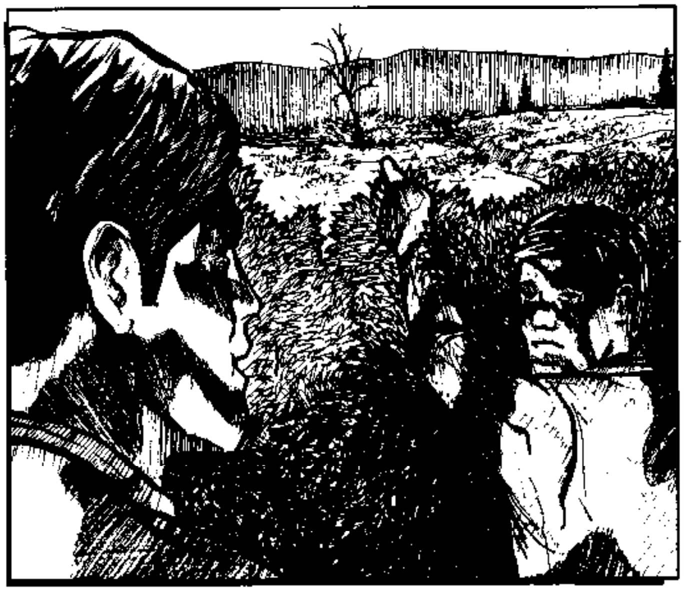

Figure 28-46. Pretravel Reconnoitering.

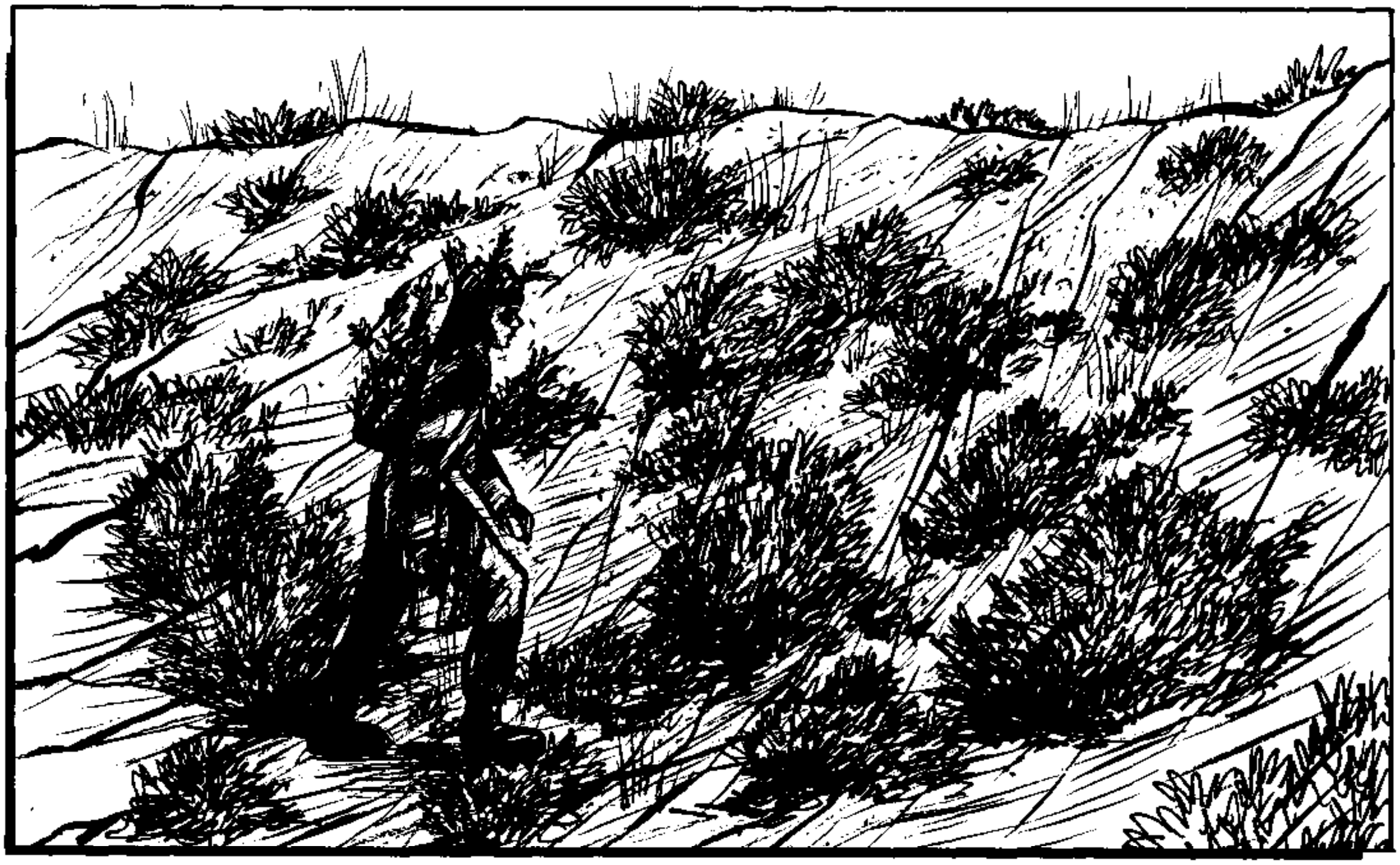

Figure 28-47. Military Crest.

e. If traveling through hilly country that provides cover, the military crest should be used as it may be the safest route. An evader's route should avoid game trails and human paths on the tops of ridges. The chance of encountering the enemy is greater on the tops of ridges. The risk of silhouetting during both day and night is also increased (figure 28-47).

(1) When it is necessary to cross the skyline at a high point in the terrain, an evader should crawl to it and approach the crest slowly using all natural concealment possible. How the skyline is to be crossed depends on whether it is likely the skyline at that point is under hostile observation. Evaders may never be certain any area is not under observation. When a choice of position is possible, the skyline is to be crossed at a point of irregular shapes such as rocks, debris, bushes, fence lines, etc.

(2) Another important point about moving along the military crest is that it is easier to evade the enemy on the side of the ridge than it is to lose sight of the enemy on the top of a ridge or in the valley below. Evasion along the side of a hill will afford a better chance for evaders to reach sites which are suitable for air recovery.

f. Evaders should move slowly, stopping and listening every few paces. Additionally, they should not make noise and should take advantage of all cover to avoid revealing themselves. If spotted, they should leave the area quickly by moving in a zigzag route to their goal if at all possible. If the enemy finds evidence of the evaders' passage, their route may help confuse the pursuer as to direction and goal. Background noise can be either a help or hindrance to those who are trying to move quietly—both the evaders and the enemy. Sudden bird and animal cries, or their absence, may alert evaders to the presence of the enemy, but those same signals can also warn the enemy of an approaching or fleeing evader. Sudden movement of birds or animals is also something to look out for.

g. The following are some techniques of limiting or concealing evidence of travel. Evaders should:

(1) Avoid disturbing any vegetation above knee level. Evaders should not grab at or break off branches, leaves, or tall grass.

(2) Glide gently through tall grass or brush. Avoid using thrashing movements. A walking stick may be used to part the vegetation in front, and then it can be

used behind to push the vegetation back to its original position. The best time to move is when the wind is blowing the grass.

(3) Realize that grabbing small trees or brush as handholds may scuff the bark at eye level. The movement of trees can be spotted very easily from a distance. In snow country it may mark a path of snowless vegetation that can be spotted when tracks cannot be.

(4) Select firm footing and place the feet lightly but squarely on the surface avoiding the following:

(a) Overturning duff, rocks, and sticks.

(b) Scuffing bark on logs and sticks.

(c) Making noise by breaking sticks.

(d) Slipping, which may make the noise of falling.

(e) Mangling of low grass and bushes that would normally spring back.

h. Evaders can mask their tracks in snow by:

(1) Using a zigzag route from one point of concealment to the next and, when possible, placing the unavoidable tracks in the shadows of trees, brush, and along downed logs and snowdrifts.

(2) Restricting movement before and during snowfall so tracks will fill in. This may be the only safe time to cross roads, frozen rivers, etc.

(3) Traveling during and after windy periods when wind blows clumps of snow from trees, creating a pockmarked surface which may hide footprints.

(4) Remembering that snowshoes leave bigger tracks, but they are shallower tracks which are more likely to fill in during heavy snowfall. Branches or bough snowshoes make less discernible prints.

i. Evaders' tracks in sand, dust, or loose soil should be avoided or else marked by:

(1) Using a zigzag route, making use of shadows, rocks, or vegetation to walk on to mask or prevent tracks.

(2) Wrapping cloth material loosely about the feet, this makes tracks less obvious.

j. By moving before or during wet or windy weather, evaders may find that their tracks are obliterated or worn down by the elements. Along roadways, evaders should be particularly cautious about leaving their tracks in the soft soil to the side of the road. They should step on sticks, rocks, or vegetation. They shouldn't leave tracks unless there are already tracks of natives on the road and their tracks can be made to look like the existing ones (small bare feet, tire sandals, or enemy footgear). Rolling across the road is a method of avoiding tracks. Walking in wheel ruts with the toes parallel to the road will help conceal tracks. If the road surface is dry, sand, dust, or soil tracks may be "brushed" out to make them look old or will help the wind erase them more quickly. This must be done lightly, however, so the tracks do not look as if they have been deliberately swept over. Mud will retain footprints unless the mud is shallow and there is a heavy rain. Evaders should try to go around these areas.

k. Many principles and techniques which work for the individual are also appropriate for use in groups of evaders. While it is not true that the only safe way to evade is individually, there is a certain danger in moving with a group.

(1) It is generally not advisable to travel in a group larger than three. If possible, the senior person should divide the group into pairs. Paired individuals must be compatible since any disagreement may prove fatal during the evasion process. Group travel can be advantageous because supplementary assistance is available in case of injury, in defense against hostile elements, in travel over rough terrain, and it provides moral support.

(2) In group travel, movement becomes critical. Movement attracts attention.

(3) The intervals and distances between individuals in the group should be made according to the terrain and the time of day. Intervals will probably be greater during the day. The natural but extremely dangerous tendency to "bunch up" is to be avoided when traveling in a group (figure 28-48).

(4) Under favorable conditions it is possible for the enemy to see 100 yards into open woods. If the undergrowth is light, the route must be farther into the woods, and the interval between evaders must be greater. Added consideration must be taken in deciding whether to travel at night or during the day. The leader will direct and guide the group to and from the best positions. All communications within the group should be made with silent signals only. Security in group evasion is of paramount importance. All members should stay alert. Security posts, lookouts, or guards should be designated for periods of rest or stopping.

(5) Various formations are available for use by the evading group during periods of travel. The group must be flexible and able to adapt to changes in the conditions of the situation. The type of formation may also change with the route. In choosing a formation, the following points should be considered:

(a) Group control and intercommunications.

(b) Security.

(c) Terrain.

(d) Speed in movement.

(e) Visibility.

(f) Weather.

(g) Enemy placement.

(h) The need for dispersion.

(i) Flexibility of change in speed and direction of travel.

(6) A "formation" is merely the formal arrangement of individuals within a group. This formal arrangement is designed to give the greatest dispersion consistent with adequate intercommunication, ease and speed of movement, and flexibility to change direction and speed of travel at a moment's notice; that is, close control. Any arrangement which provides the above advantages

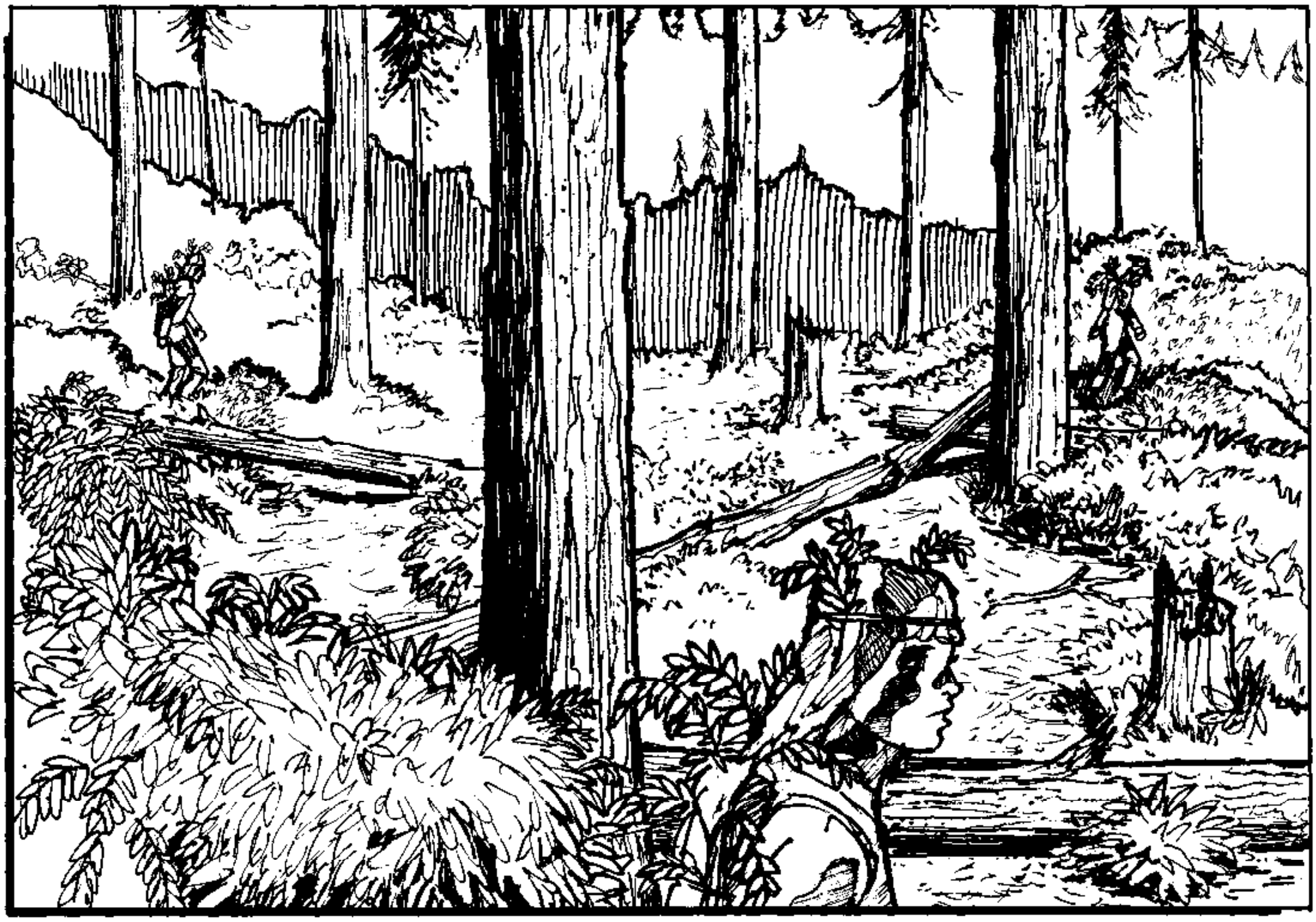

Figure 28-48. Group Travel.

is satisfactory. The Army, which constantly moves groups of men of various sizes on various missions, has found the squad file, squad column, and squad line to be satisfactory.

(a) In a squad file, the personnel are arranged in a single file, or one directly behind the other, at different distances. The distance will vary depending on need for security, terrain, visibility, group control, etc. It is primarily used when moving over terrain which is so restrictive that the squad cannot adopt any other formation. It is also used when visibility is poor and squad control becomes difficult. When people are in a squad file, it is easy to control the group and provide maximum observation of the flanks. This is a fast way to travel, especially in the snow.

-1. However, there are disadvantages to this type of movement. A major one is that this formation is visually eye-catching. All the noise of the group is also concentrated in one place. This type of formation is easily defined and infiltrated. Depending on how many people are in the group, the area they walk through can become very packed down and easily detected by the enemy.

-2. If the group in the squad file is small, some members may have to double up on jobs; however, the "point" or lookout (scout) in the lead should only perform that job (figure 28-49).

(b) In the squad column, personnel are arranged in two files. The personnel are more closely controlled and yet maneuverable in any direction. There is greater dispersion with reasonable control for all-round security. It is used when terrain and visibility are less restrictive because it provides the best means of moving armed personnel into dispersed all-round security. It is easy to change into either the file or the line.

-1. There are many advantages to this type of formation, a major one being a greater dispersal of personnel. Visually, this way of moving is less eye-catching, and the sounds the group makes are less concentrated. With this formation, the rate of travel is reasonable, yet there is no well trampled corridor for enemy trackers to follow.

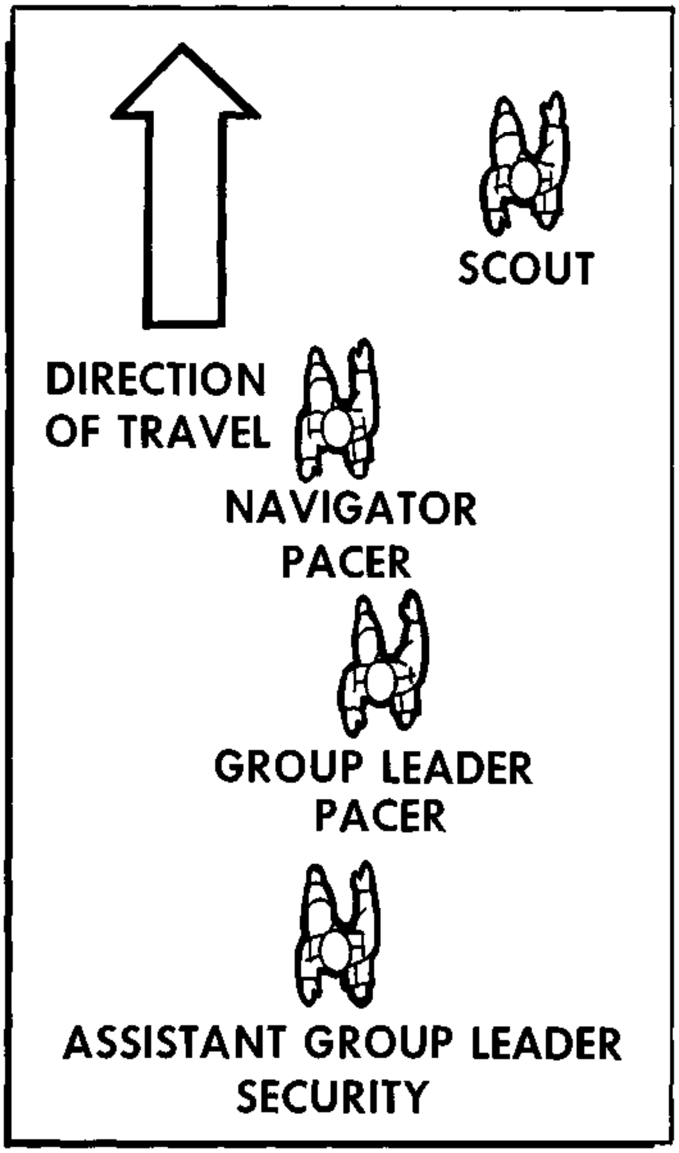
SCOUT
DIRECTION OF TRAVEL
NAVIGATOR
PACER
GROUP LEADER
PACER
ASSISTANT GROUP LEADER
SECURITY

Figure 28-49. File Formation.

-2. There are a few disadvantages to movement in this manner. This formation of travel is hard to control in areas of dense vegetation with poor visibility and makes straying from the group likely. Although the paths may be faint, this mode of travel will also form a large number of trails. The rate of travel will be slower overall when traveling this way. An example of how personnel may be dispersed using the squad column methods is shown in figure 28-50.

(c) In the squad line, all personnel are arranged in a line. This formation is used primarily by the Army as an assault formation because it is best for short, fast movements.

-1. The advantage of this formation is that it is the quickest way to cross such obstacles as roads, fences, and small open spaces. It provides for tight control of individual movement while providing security for short-distance moves.

-2. The disadvantages are that there are extreme communication and control problems. Some personnel in this formation may be forced to traverse undesirable rough terrain in contrast to the other two formations. Figure 28-51 illustrates the organization of personnel in this type formation. The speed at which these formations will progress will vary with terrain, light, cover, enemy presence, health of personnel, etc.

l. Regardless of the formation used, the evader should pay particular attention to the technique used to travel. Knowing how to walk or crawl may make the difference between success and failure.

(1) Correct walking techniques can provide safety and security to the evader. Solid footing can be maintained by keeping the weight totally on one foot when stepping, raising the other foot high to clear brush,

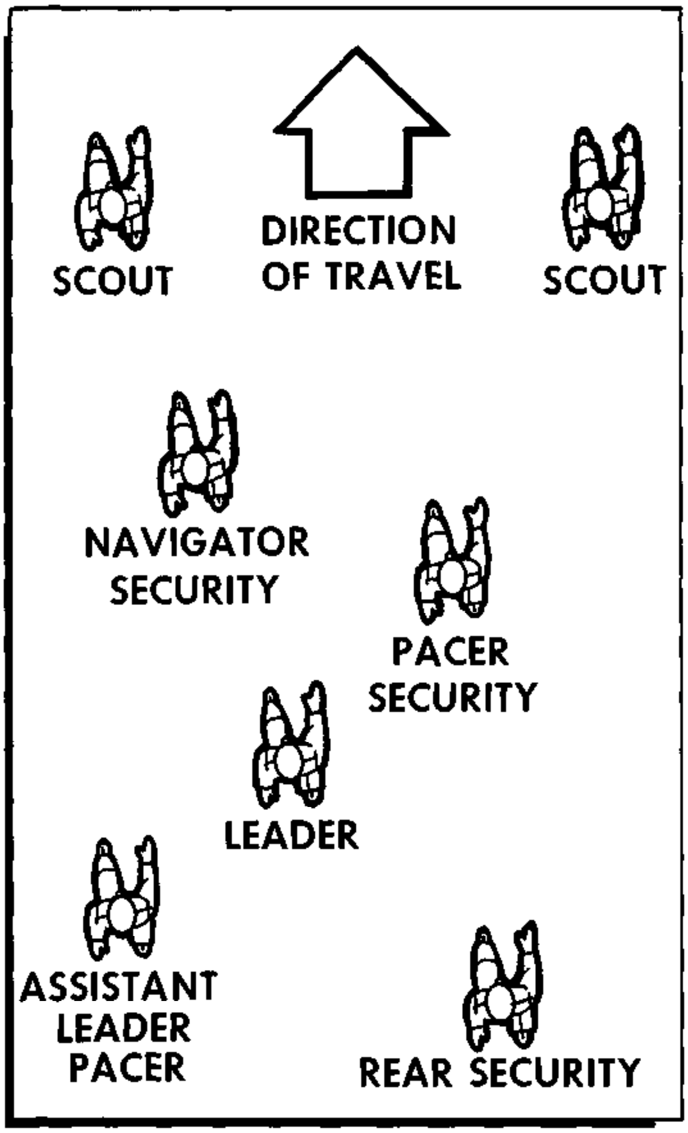
SCOUT
DIRECTION OF TRAVEL
SCOUT
NAVIGATOR
SECURITY
PACER
SECURITY
LEADER
ASSISTANT LEADER
PACER
REAR SECURITY

Figure 28-50. Squad Formation.

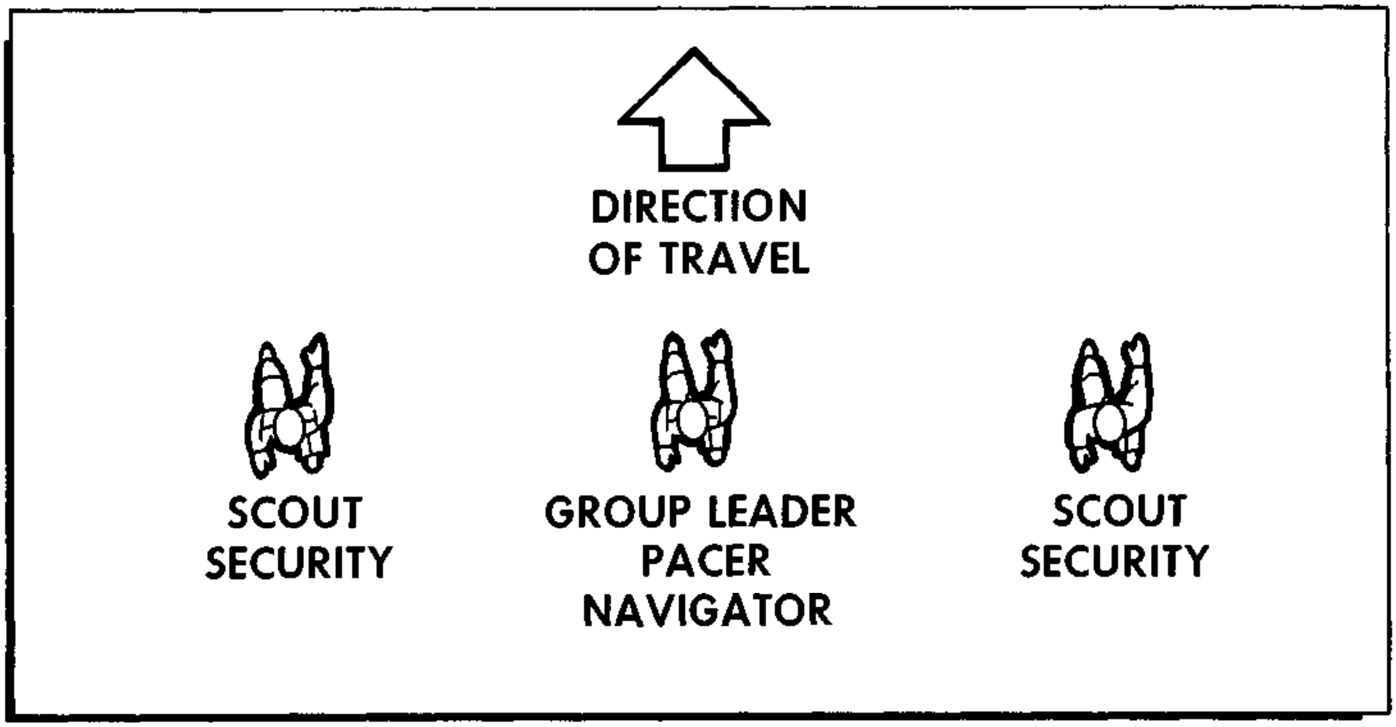

Figure 28-51. Squad Line.

grass, or other natural obstacles, and gently letting the foot down, toe first. Feel with the toe to pick a good spot—solid and free of noisy materials. Lower the heel only after finding a solid spot with the toe. Shift the weight and balance in front of the lowered foot and continue. Take short steps to avoid losing balance. When vision is impaired at night, a wand probe, or staff, should be used. By moving these aids in a figure-eight motion from near the ground to head height, obstructions may be felt.

(2) Another method of movement is by crawling. Crawling is useful when a low silhouette is required. There are times when evaders must move with their bodies close to the ground to avoid being seen and to penetrate some obstacles. There are three ways to do this: the low crawl, the high crawl, and the hands-and-knees position. Evaders should use the method best suited to the condition of visibility, ground cover, concealment available, and speed required.

(a) The Low Crawl. This can be done either on the stomach or back, depending on the requirement. The body is kept flat and movement is made by moving the arms and legs over the ground (figure 28-52).

(b) The High Crawl. This is a position of higher silhouette than the low crawl position, but lower than the hands and knees position. The body is free of the ground with the weight of the body resting on the forearms and lower legs. Movement is made by alternately advancing the right elbow and left knee, left elbow and right knee, elbows and knees laid flat on their inside surfaces (figure 28-53).

(c) Controlled Movement. The low crawl and high crawl are not always suitable when very near an enemy. They sometimes cause the evader to make a shuffling noise which is easily heard. On the other hand, carefully controlled movement can be made to be silent, and these techniques present the lowest possible silhouettes.

(d) The Hands-and-Knees Crawl. This position is used when near an enemy. Noise must be avoided, and a relatively high silhouette is permissible. It should only be used when there is enough ground cover and concealment to hide the higher silhouette involved (figure 28-54).

m. If evaders are moving as a group and are forced to disperse, being able to account for everyone after regrouping is important. Likely locations for rallying points are selected during map study or reconnaissance.

(1) Selecting a Rallying Point. The group leader must:

(a) Always select an initial rallying point. If a suitable area for this point is not found during map study or reconnaissance, the leader can select it by grid coordinates or in relation to terrain features.

(b) Select likely locations for rallying points en route.

(c) Plan for the selection and designation of additional rallying points en route as the patrol reaches suitable locations.

(d) Plan for the selection of rallying points on both near and far sides of danger areas which cannot be bypassed, such as trails and streams. This may be done

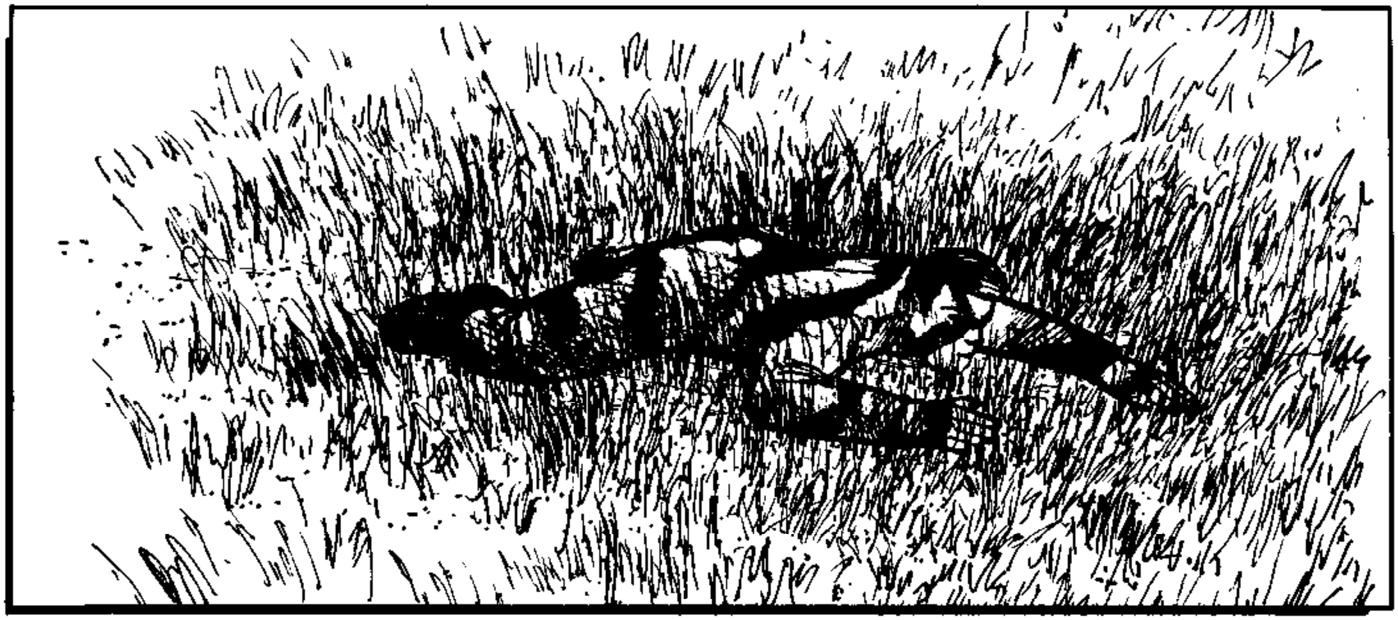

Figure 28-52. Low Crawl.

by planning that, if good locations are not available, rallying points will be designated in relation to the danger area; for example, "...50 yards this side of the trail," or "...50 yards beyond the stream."

(2) Use of Rallying Points. If dispersed by enemy activity or through accidental separation, each evader in the group should be prepared to evade, individually, to the regrouping (rallying) point to arrive at a predesignated time. If this is not possible, the individual will become a "lone evader." The group should not make any effort to locate someone not reaching the rally point. The group should formulate a new plan with new rallying points and clear the area.

(a) Rallying points should be changed or updated as they are passed. Points should not be directly on the line (route) of travel. By selecting points off line, the job of searchers (trackers) is made more difficult and the chance of being "headed off" or "blind stalked" by the enemy is reduced.

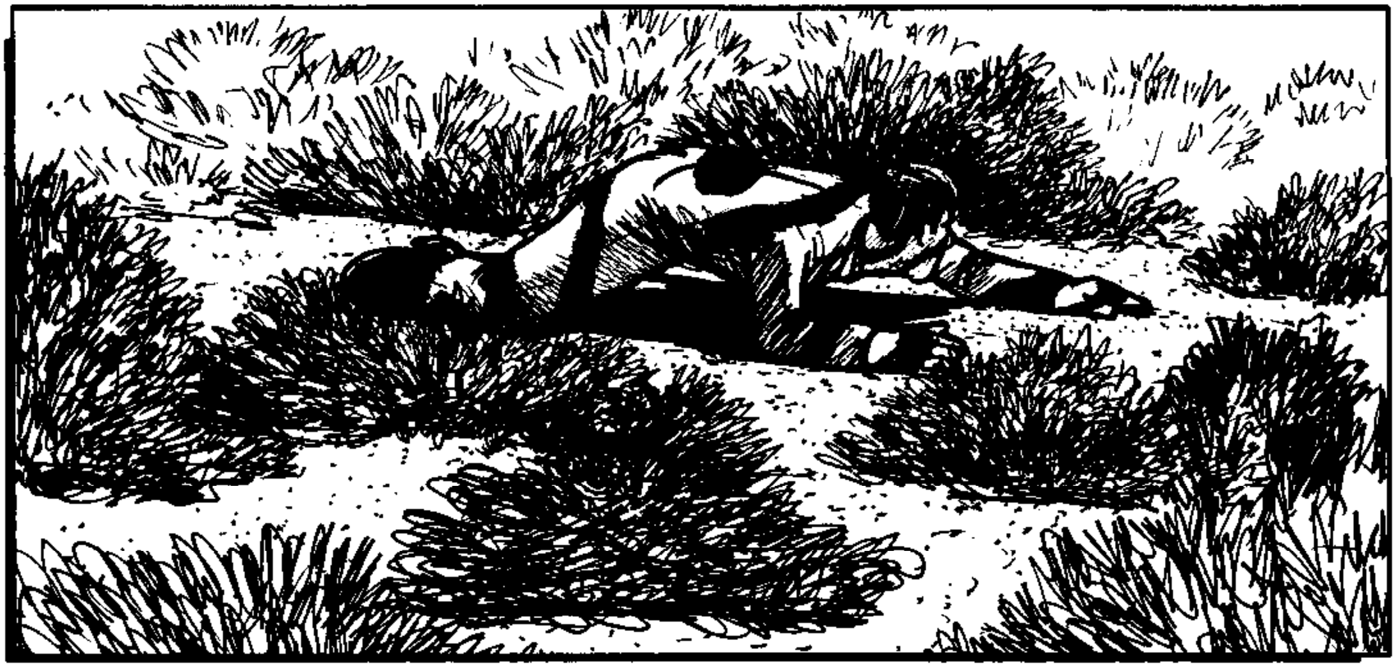

Figure 28-53. High Crawl.

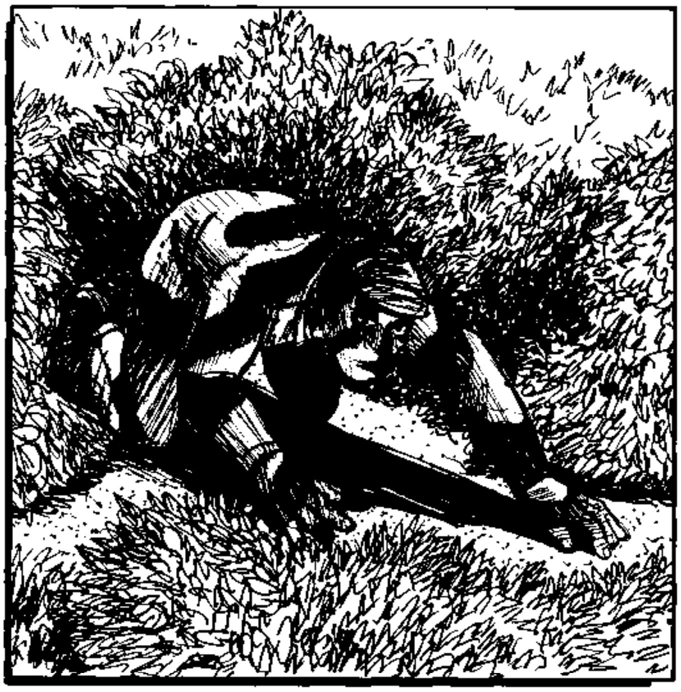

Figure 28-54. Hands-and-Knees Crawl.

(b) If the group is dispersed between rallying points en route, the group rallies at the last rallying point or at the next selected rallying point. The group leader announces the decision at each rallying point as to which point the group will rally.

(3) Actions at Rallying Points. Actions to be taken at rallying points must be planned in detail. Plans for actions at the initial rallying point and rallying points en route must provide for the continuation of the group as long as there is a reasonable chance to evade as a group. An example of a plan would be for the group to wait for a specified period, after which the senior person present will determine actions to be taken based on personnel and equipment present. Even during movement phases, it is important to be able to check on the presence and status of all group members. A low toned, actual head count starting at the rear of the formation might be one way to do this; hand signals is another. One reason for keeping in touch with everyone is to establish new plans or adjust old ones while moving.

28-17. Barriers to Evasion Movement:

a. Obstacles. Throughout the evasion episode, many obstacles may be encountered which may impede evaders or influence the selection of travel routes. These barriers or obstacles can be divided into natural ones, such as rivers or mountains, and human ones, such as border guards or manmade fences or roads. Some of these obstacles may be helpful while others might be a hinderance.

(1) Rivers and Streams. When crossing rivers and streams, bridges and ferries can seldom be used since the enemy normally establishes checkpoints at these locations. This leaves a choice of fording, swimming, crossing by boat, or using some improvised method (figure 28-55).

(2) Mountains. In mountainous areas, survival may be the primary concern. It may be necessary for evaders

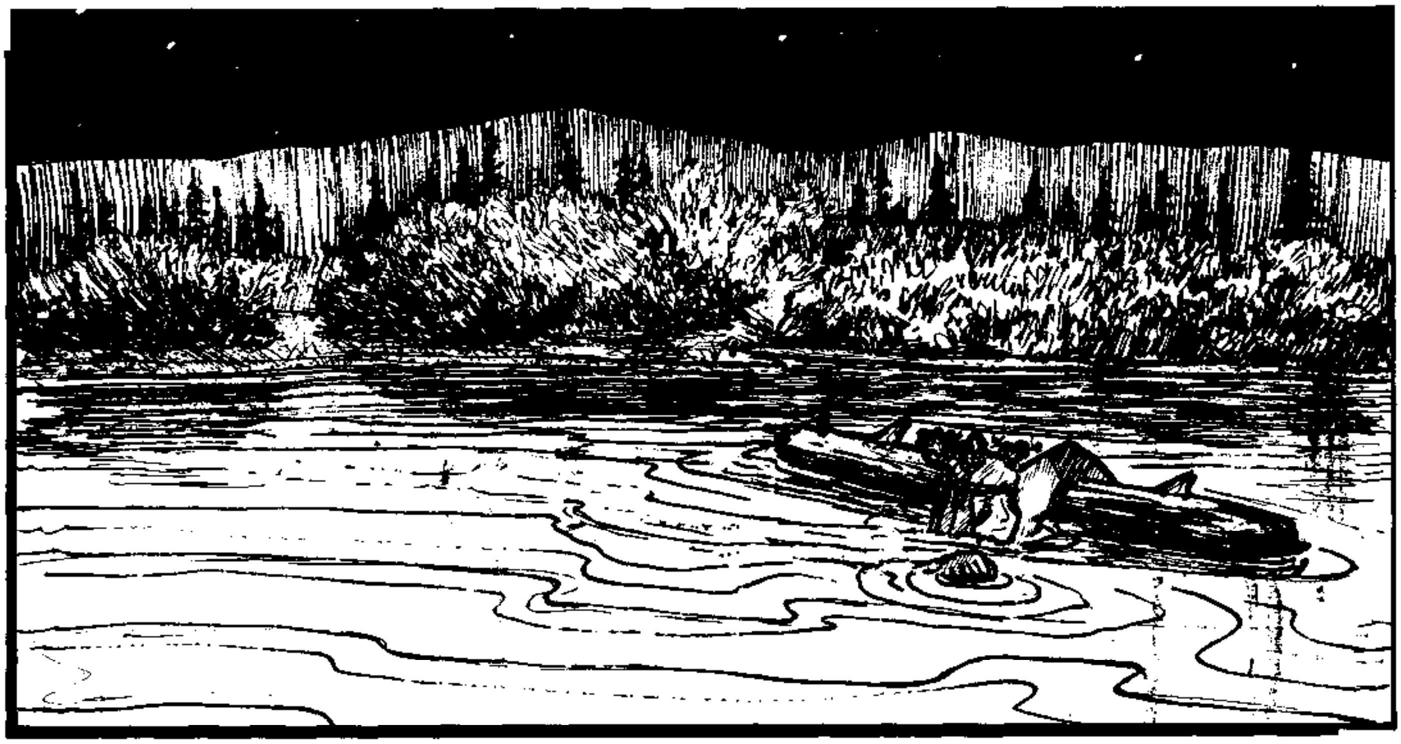

Figure 28-55. River Crossing.

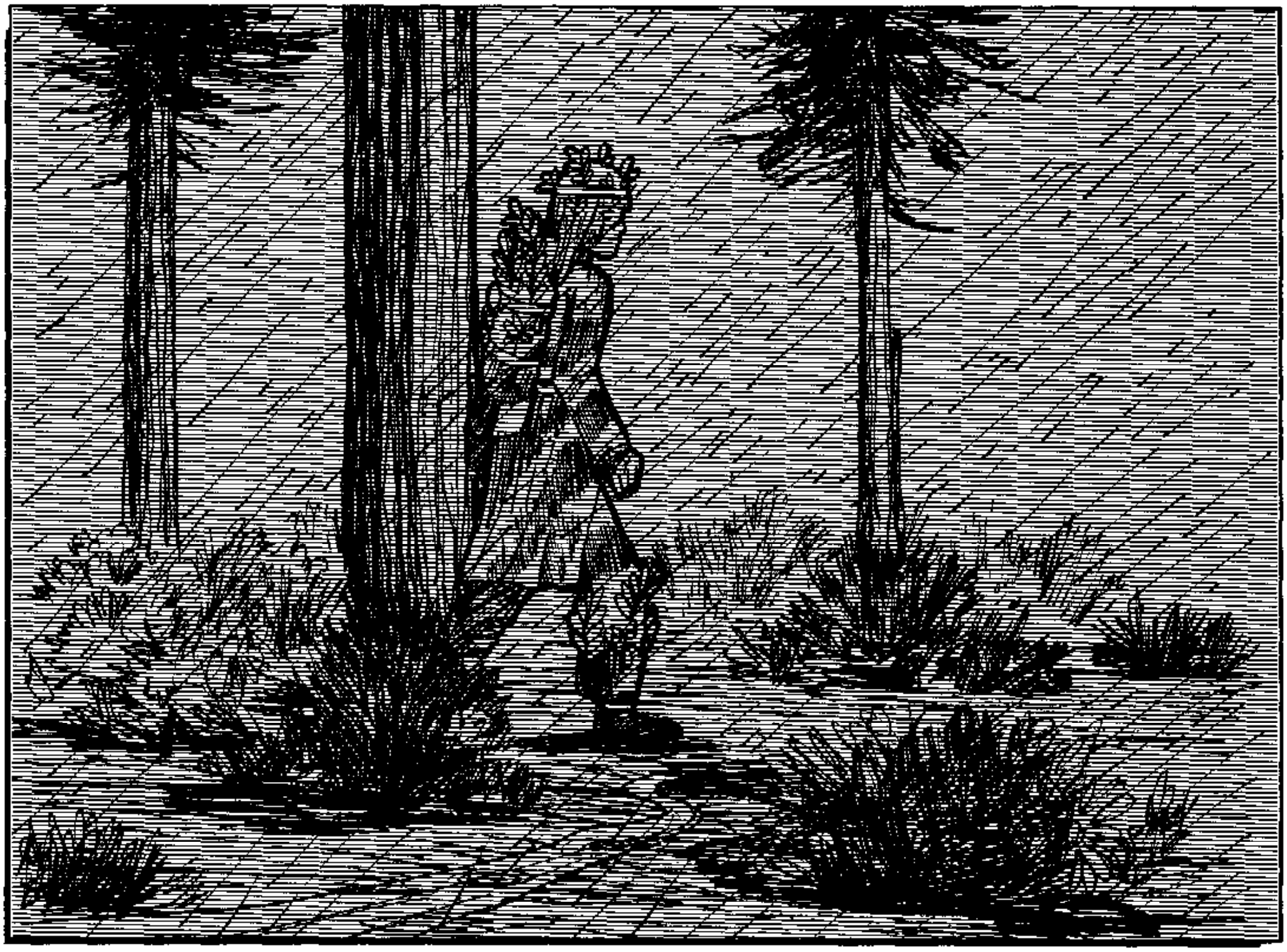

Figure 28-56. Weather.

to remain in one location for an extended period of time, perhaps even waiting for the coming of spring before attempting travel. Many mountainous areas, however, are havens which afford cover, water, food, and low population densities. Also, the chances of receiving assistance from people in areas where homes and farms are separated by great distances are more likely. When traveling in mountainous regions, evaders should not forget to use the military crest if concealment is available. In plains areas, evaders should use depressions, drainages, or other low spots to conceal their movements. Route selection should be planned with the utmost care to avoid unnecessary delays caused by cliffs, large bodies of water, and flat areas.

(3) Vegetation. Some swamps, drainage areas, and thickets may be too thick for evaders to penetrate, and may require that a detour or alternate route be used. If the vegetation can be moved through, evaders should take care not to leave evidence of passage by disturbing the growth.

(4) Weather. Weather can sometimes be used to screen evaders from the enemy. Certain weather conditions mask the noise made by traveling (figure 28-56). Moving during a rainstorm may erase the footprints left by an evader; but after the rain the soft soil will leave definite signs of passage. Thunder may mask the sounds evaders make, but lightning may cause them to be seen. Snowstorms may be used to cover evaders' signs and sounds, but once the storm is over, evaders must use extreme care not to leave a trail.

b. Artificial Obstacles. Evaders may also encounter a wide variety of artificial obstacles while traveling within enemy territory or when attempting to leave a controlled area. As a general rule, evaders should *not* attempt to penetrate these obstacles if they can be bypassed. If an analysis of the situation reveals the obstacles cannot be bypassed, evaders must be skilled in the methods and techniques for dealing with specific artificial barriers to evasion. If possible, move to a less fortified (controlled) area or find a better damaged area in the barrier.

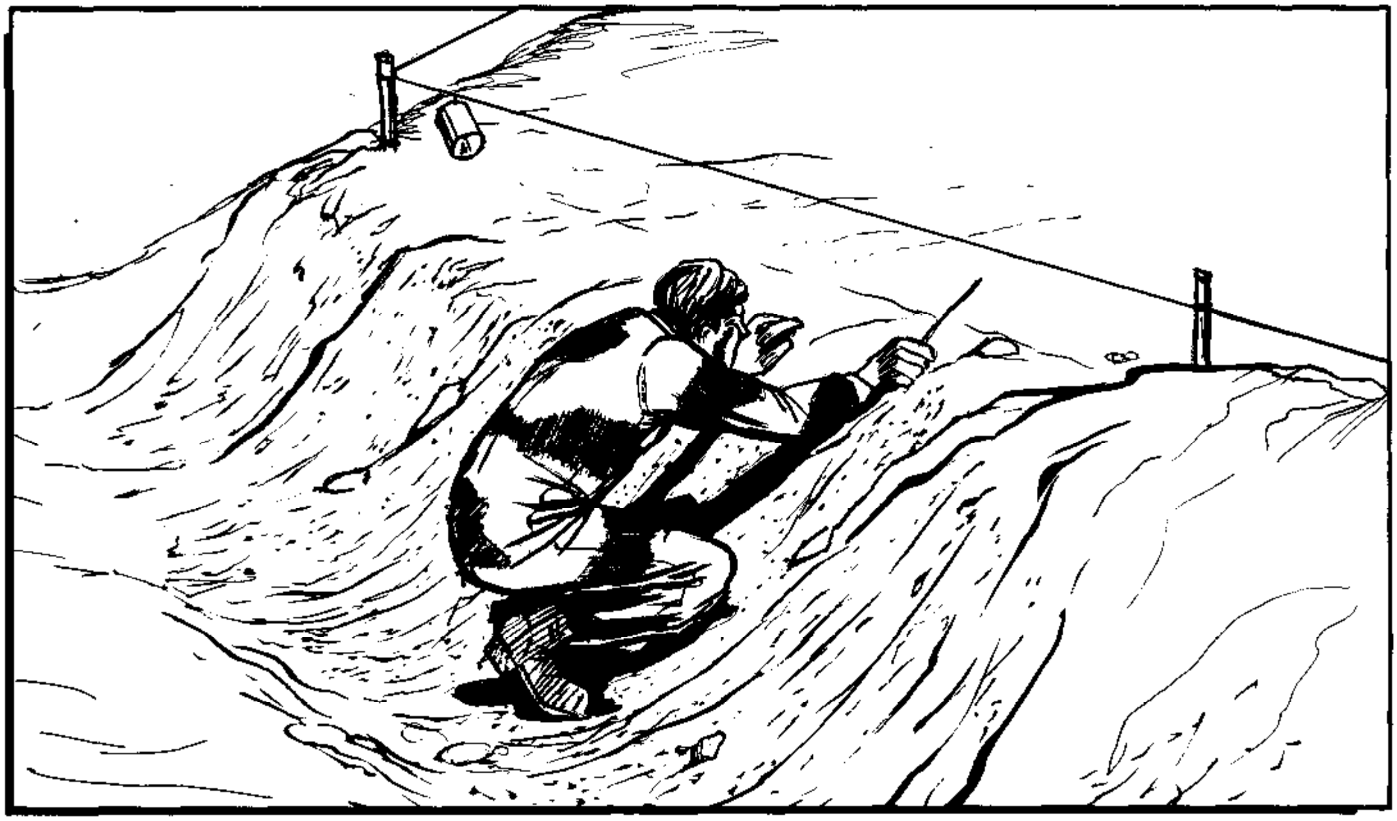

Figure 28-57. Trip Wires.

(1) Evaders may encounter trip wires. These wires may be attached to pyrotechnics, booby traps, sensor devices, mines, etc. These wires are normally thin, olive green (or other colors that blend with the environment), strong, and extremely difficult to see. A supple piece of wood can be improvised and used as a wand to detect these wires (figure 28-57).

(a) A tripwire may be set up to be from 1 inch to a number of feet above the ground and to extend any number of feet from the device to which it is attached.

(b) The pressure necessary to activate a sensing device, mine, pyrotechnic device, or other trap-associated device to which the wire may be connected can vary from a few ounces to several pounds. This means that the evader must be very careful when attempting to determine the presence of these devices.

(c) Once a tripwire has been detected, evaders should move around the wire if possible. If not possible, they should go either over or under it. They should not tamper with or cut the wire. If one device is discovered, be alert for "backup" devices.

(d) A number of devices activated by tripwires have a combination pressure-release arming mechanism. Cutting the wire or releasing the tension in the wire may activate the device. Some devices are electrically activated when there is a change in the current flowing through the attached wire—either because the wire is cut, in some way grounded, or otherwise altered. Evaders should take extreme care to avoid touching tripwires, but if contact is made, they must try to avoid sufficient pressure for activation.

(2) Illumination flares may also present a problem to evaders. These, of course, can be activated by evaders themselves by a tripwire, by the enemy in the form of electronically activated ground flares, or by flares dropped from an overflying aircraft. Other overhead flares may be fired by mortars, rifles, artillery, and hand projectors.

(a) The illumination flares may burn as bright as 20,000 candlepower and illuminate up to a 300-foot radius in case of a ground flare, or a much larger area in the case of an overhead flare which is lofted or dropped and burns high above the ground.

(b) If evaders hear the launching burst of an overhead flare, they should, if possible, get down while it is rising and remain motionless. If caught in the light of a flare when they blend well with the background, they should freeze in position and not move until the flare goes out. The shadow of a tree will provide some protection. If caught in an open area, they may elect to crouch low or lie on the ground and, as a general rule, should not move after the area is illuminated.

(c) However, if they are caught in the light of a ground flare and their position is such that the risk of

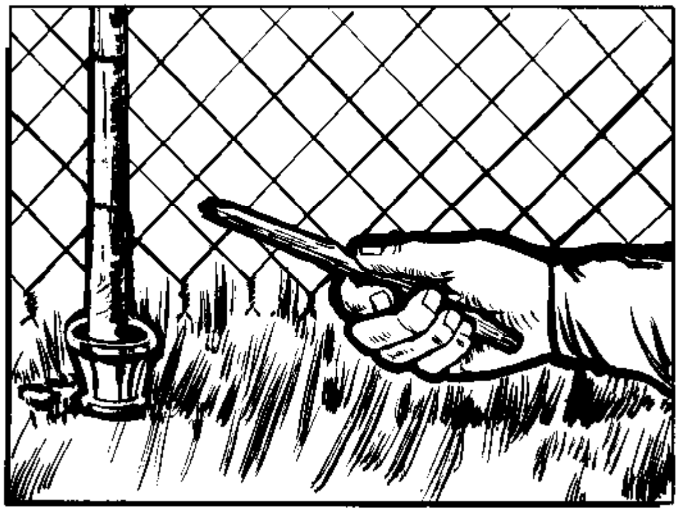

Figure 28-58. Electrified Fences.

remaining is greater than that of moving, they should move quickly out of the area. If within a series of obstacles or an obstacle system, evaders must remember running can be extremely dangerous because of the obstacles in the area and the fact that movement, especially fast movement, catches the eyes of an observer. If it is determined they cannot quickly move out of the area because of possible serious injuries due to existing obstacles or because they may be observed by the enemy, evaders should drop to the ground and conceal themselves as much as possible.

(d) Evaders should remember the light of a flare (either ground or overhead bursts) is temporarily blinding and the eyes should be covered to conserve night vision.

(e) If caught by a flare when actually penetrating an obstacle such as a concertina wire, evaders should get as low as possible, stay still, and cover their eyes. The light of a flare can act to an evader's advantage because the searching enemy will lose its night vision which may add to the evader's chance for success in departing the area after the flare has burned out. The light of a flare also creates very dark shadows which, under some circumstances, can afford good concealment from enemy observation.

(3) Various types of chain and wire fences may hinder the progress of evaders who are moving to the safety of friendly areas.

(a) For indications of electrical fences, evaders should watch for dead animals, insulators, flashes from wires during storms, and short circuits causing sparks (figure 29-58). A quick simple test can be conducted to determine if a wire is electrified. This test is made by carefully approaching the wire holding a stem of grass or a small, damp stick on the wire. If the wire is charged, a mild shock will result but will not cause injury.

(b) Evaders should use a wand to check for booby traps between strands of multi-strand barbed wire. Generally, they should penetrate the fence under the wire

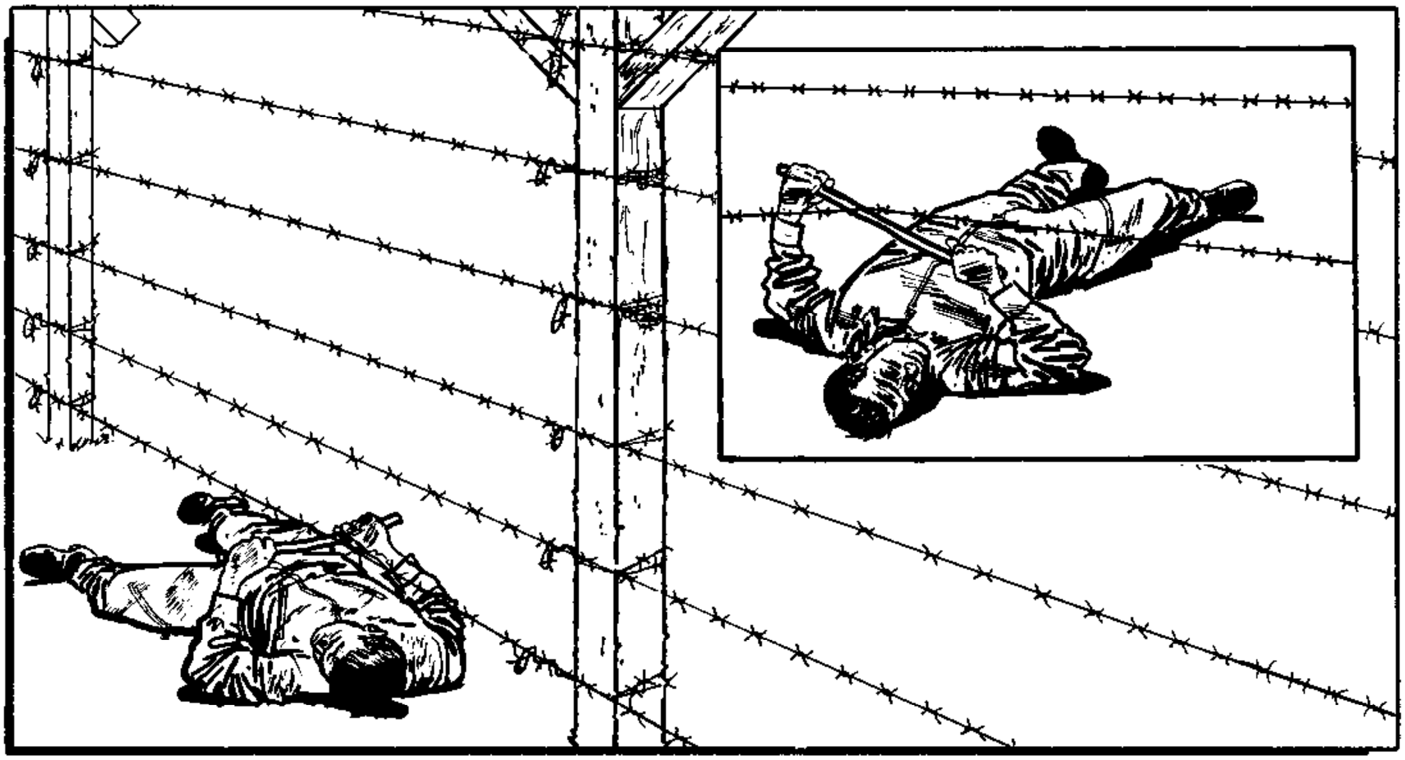

Figure 28-59. Penetrating Wire.

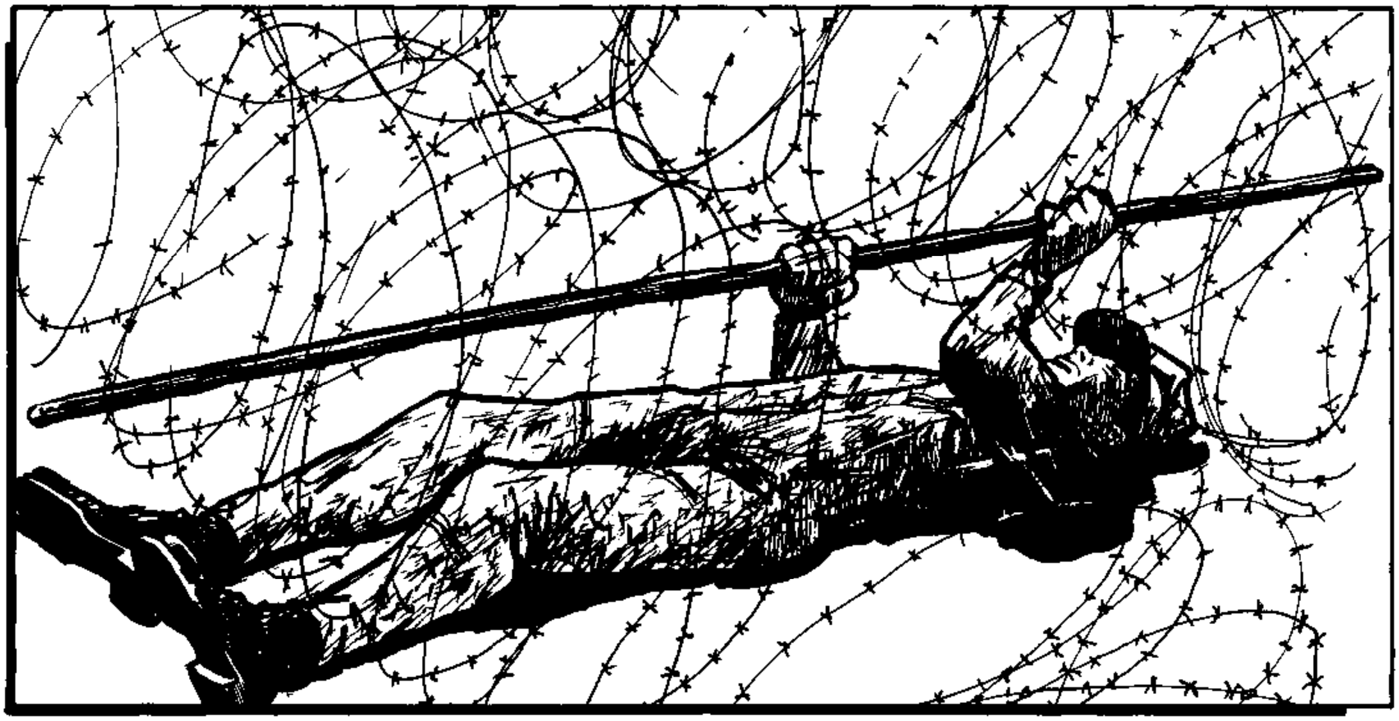

Figure 28-60. Penetrating Concertina Wire.

closest to the ground with the body parallel or perpendicular to the wire, depending on circumstances (figure 28-59). They should lie flat on their backs both to project the lowest possible silhouette and to provide good visibility of the wire against the sky. The probe can sometimes be used to lift the wire. If the lowest wire is close to the ground and is tight, evaders may have to modify their approach to the problem.

(c) The apron fence is penetrated in the same manner as any multi-strand fence—one wire at a time. Evaders should check the area between wires before proceeding.

(d) Concertina wire is penetrated with the body perpendicular to the wire using a probe to lift the wire if it is not secured to the ground (figure 28-60). If the wire is secured to the ground, the evader can crawl between the loops. If two loops are not separated enough, they may be tied apart using shoe laces, string, suspension line, or strips of cloth. After passing through, the ties should be removed for future use and to erase evidence of travel.

(e) Chain link fences should be avoided completely if at all possible. These fences are usually found only in highly sensitive zones. This means the area is probably more highly guarded and patrolled than other areas. There also may be other traps or devices installed. The fence may also be electrified. If the fence must be penetrated, the evader should go under it if possible (figure 28-61). If digging is required, the soil should be placed on the opposite side so it can be re-

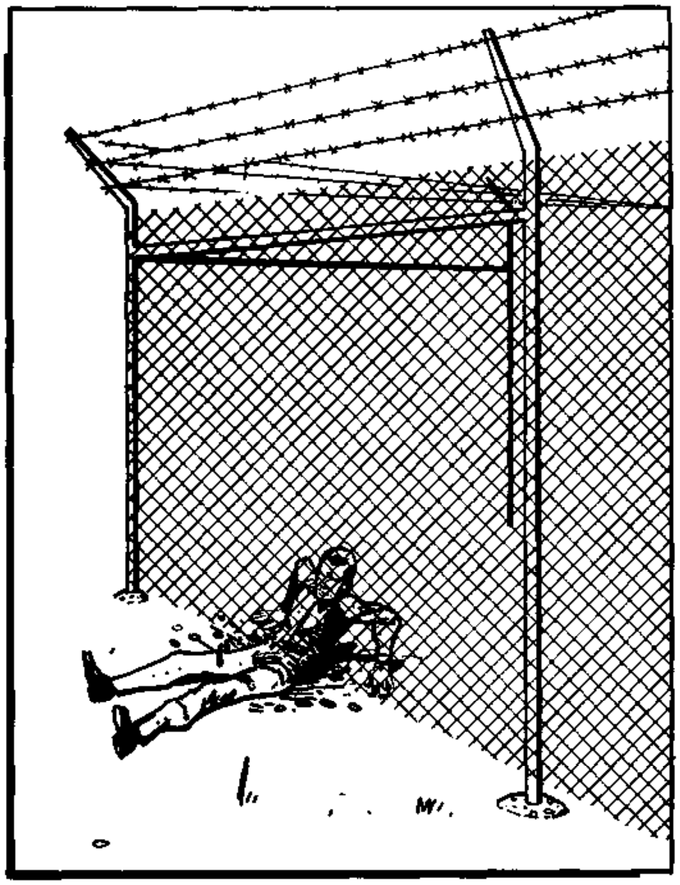

Figure 28-61. Penetrating Chain Link Fence.

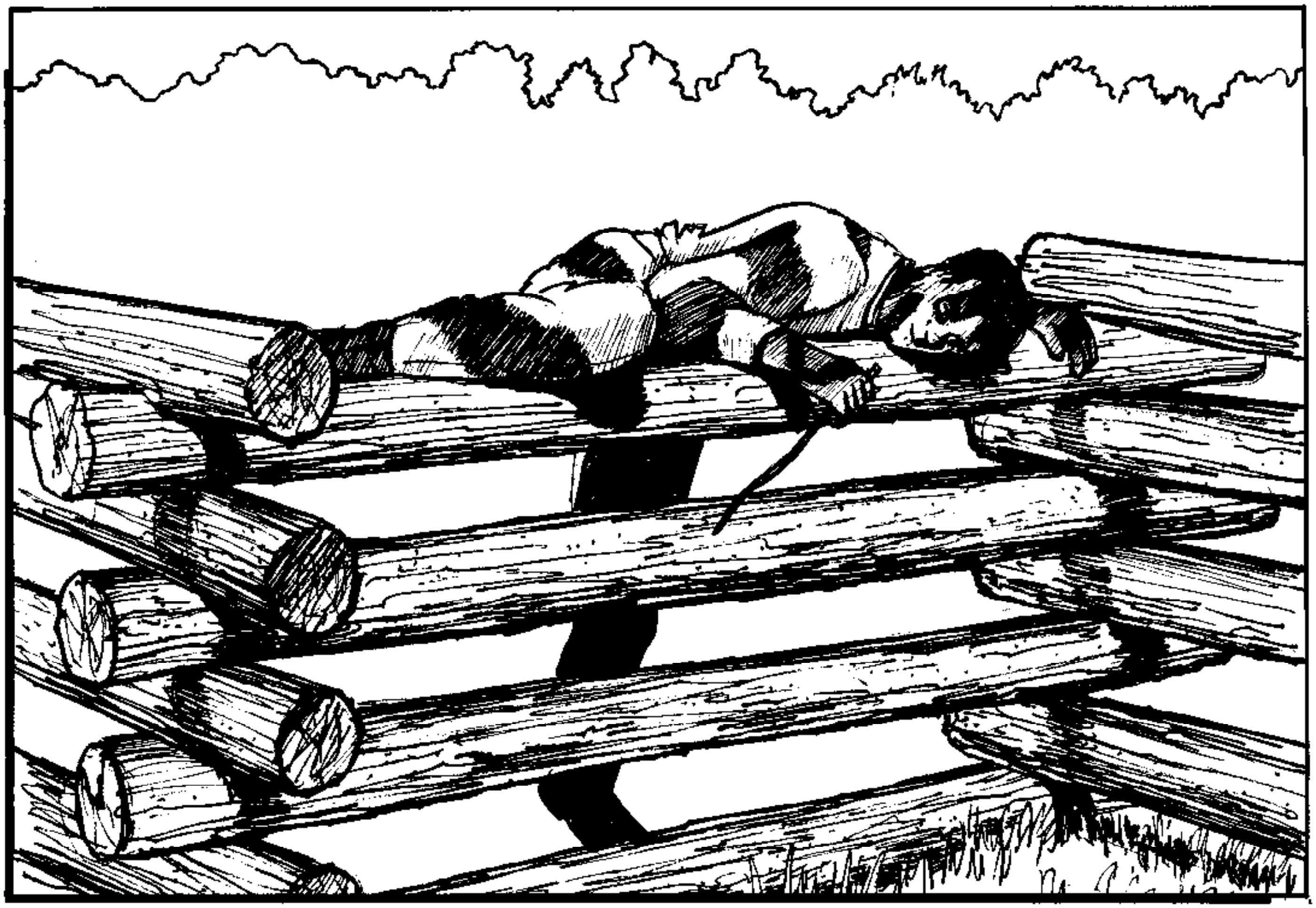

Figure 28-62. Penetrating Log Fence.

placed to remove evidence. Climbing the fence is recommended only as a last resort.

(f) Evaders may encounter rail and split-rail fences while evading or escaping. The fences are penetrated by going under or between lower rails if possible. If not, evaders should go over at the lowest point, projecting as low a silhouette as possible (figure 28-62). They should check between the rails and on the other side of the fence to detect tripwires or booby traps. Firmness of the ground should be checked on both sides of the fence. The body should be parallel to the fence before penetration.

(4) Raked or plowed areas may be found in areas of both low and high density security. If such an area is encountered, evaders should roll across the area, after making sure it is not a mine field, to avoid leaving footprints; or they may side-step, walk backwards, or brush out footprints. Any of the above may be done when it is a requirement not to leave clear-cut evidence of the direction of movement.

(5) Roads are common barriers to evasion and escape. When roads are encountered, evaders should closely observe the road from concealment to determine enemy travel patterns (figure 28-63). Crossing from points offering best concealment such as bushes, shadows, etc., is best (figure 28-64). Evaders should cross at straight stretches of road in open country and on the inside of curves in hilly or wooded areas. This allows the evader to see in both directions so the chances of being spotted or surprised in the open is minimized. Avoid leaving tracks both in the road and on the shoulder of the road.

(6) Culverts and drains offer excellent means of crossing a road unobserved (figure 28-65).

(7) Railroad tracks often lie in the path of evaders. If so, evaders should use the same procedures for observation as for roads. If it is determined that tracks are patrolled, a check should be made for booby traps and tripwires between tracks. Aligning the body parallel to

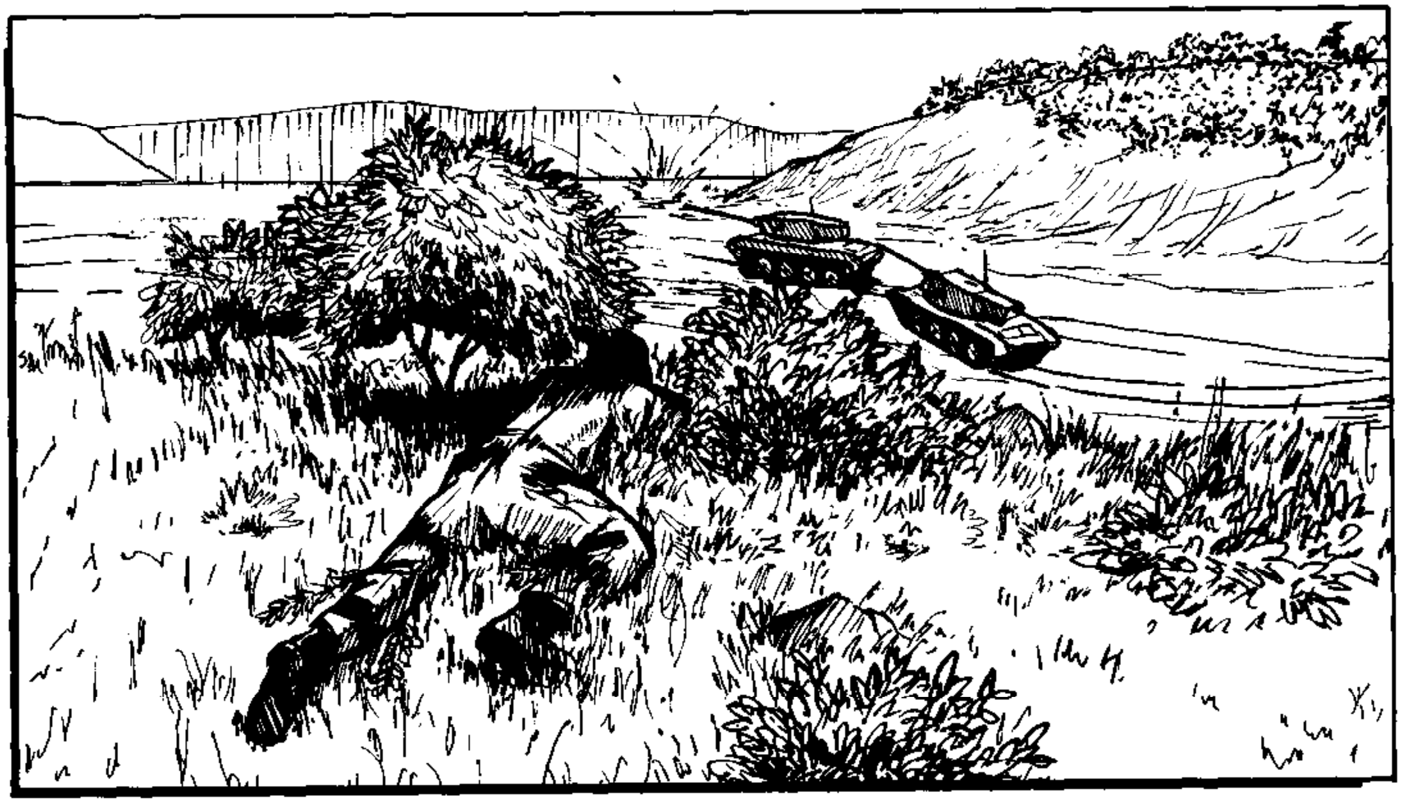

Figure 28-63. Common Barriers.

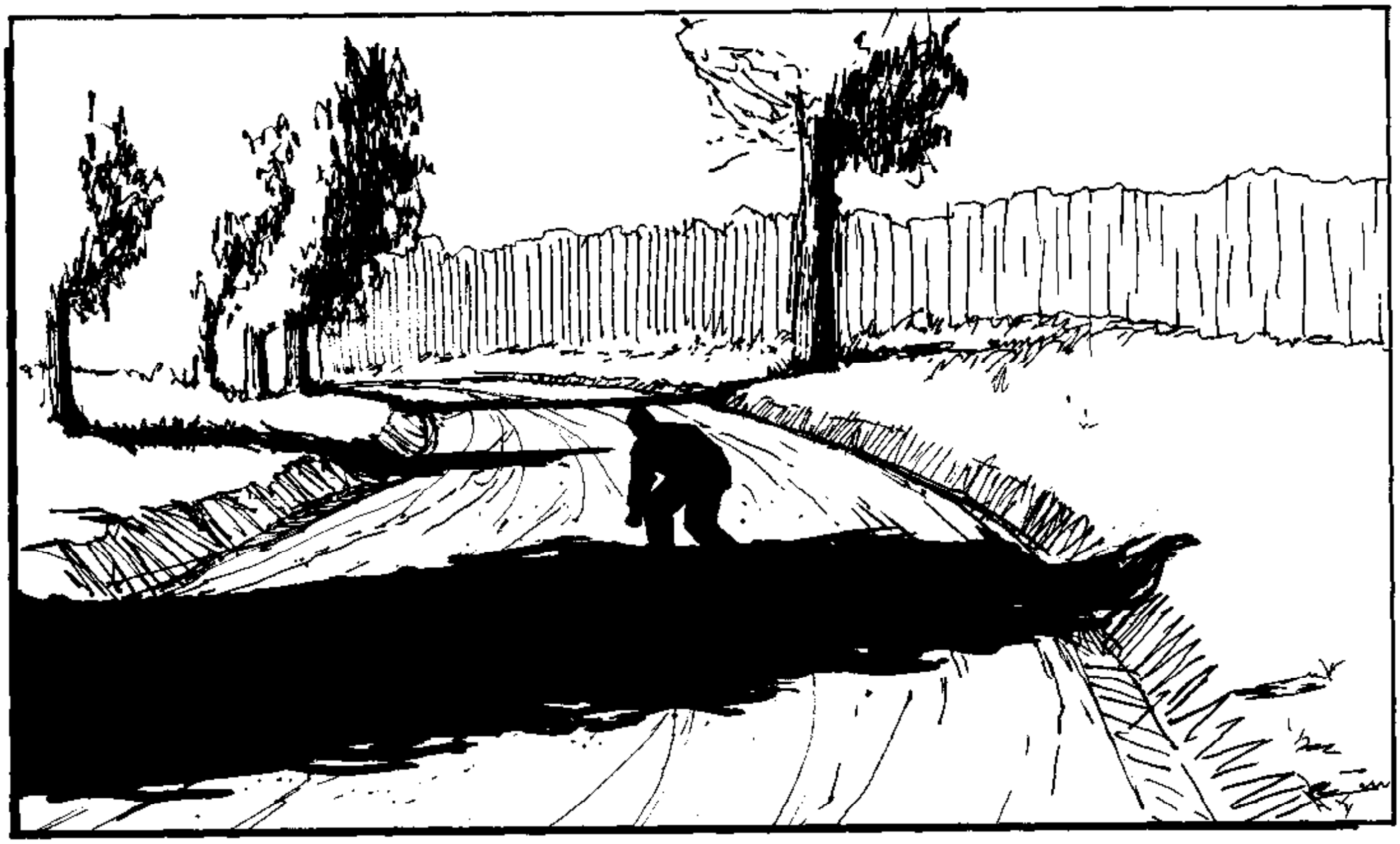

Figure 28-64. Crossing Roads.

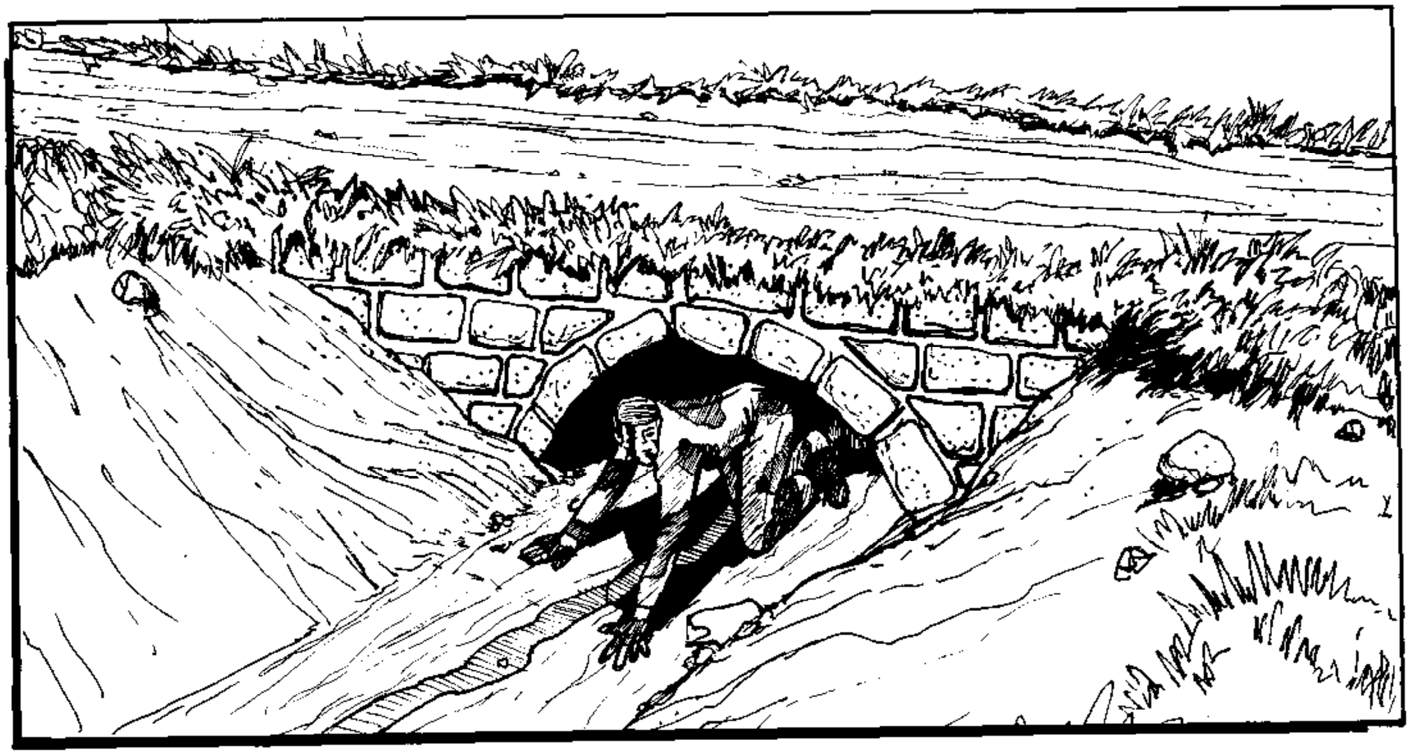

Figure 28-65. Crossing at Culverts.

the tracks with face down next to the first track, evaders should carefully move across the first track in a semi-pushup motion, repeating for the second track and subsequent sets of tracks (figure 28-66). If there is a third rail, they should avoid touching it as it could be electrified. Sound detectors can also be attached to the tracks and can reveal any crossing if a track is touched. If determined from observation that the tracks are not

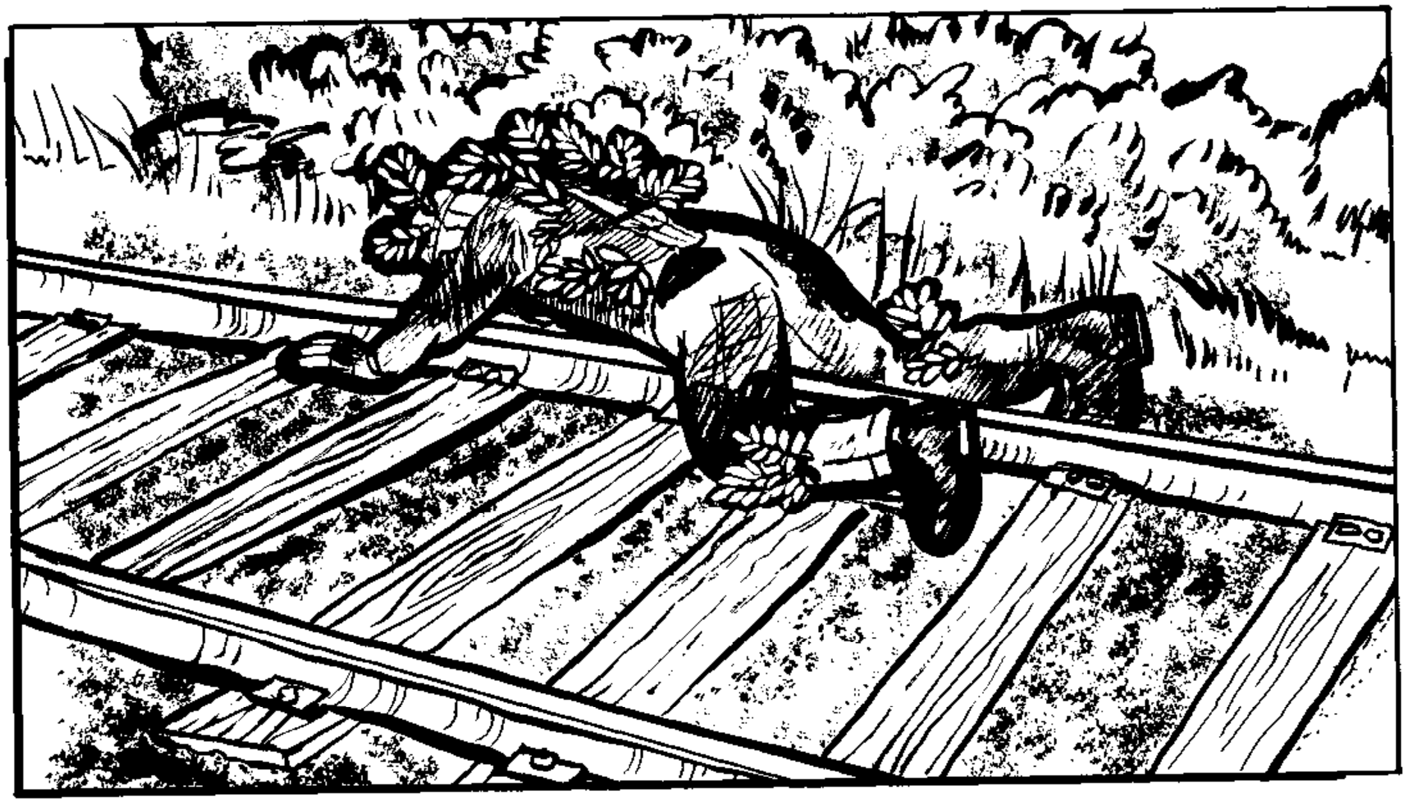

Figure 28-66. Crawling Over Railroad Tracks.

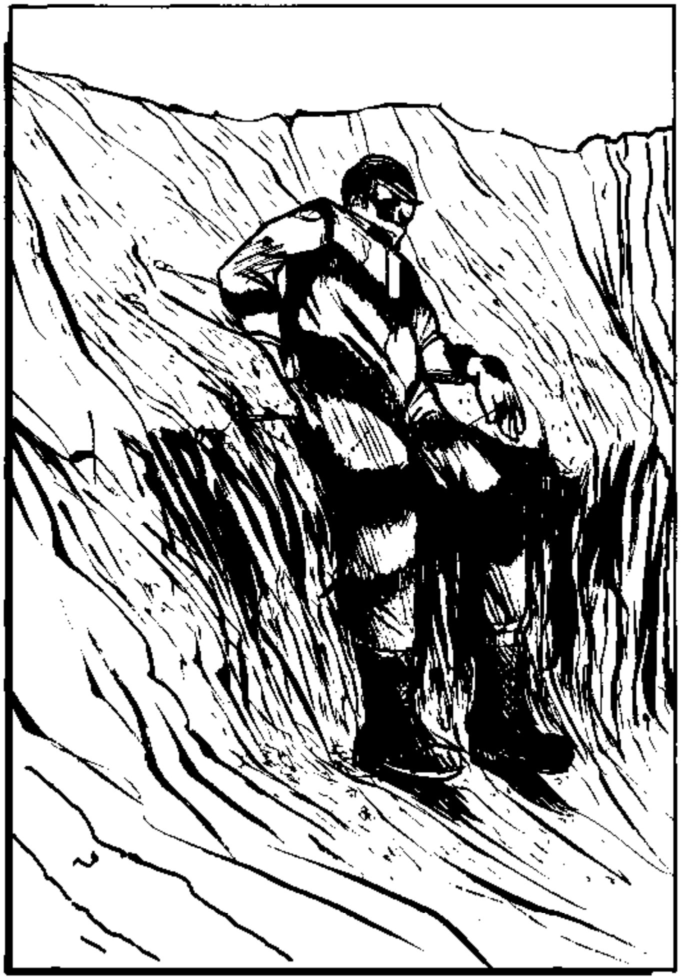

Figure 28-67. Climbing Down Cliff.

patrolled, evaders should cross in a normal walking or hands-and-knees manner, attempting to attract as little attention as possible. Evaders should try to keep their hands and feet on the railroad ties to prevent leaving foot or hand prints in the adjacent soil or gravel.

(8) Deep ditches (such as tank traps or natural drainages) may be obstacles with which evaders must deal. Ditches should be entered feet first to avoid injury to the head or upper torso should there be large rocks, barbed wire, or other hazards at the bottom (figure 28-67). Using a wand to detect tripwires and booby traps in the ditch and on the sides is highly recommended. Maintaining a low silhouette upon exiting the ditch is imperative.

(9) Open terrain complicated by guard towers or walking patrols is a definite hazard to evaders. These areas should be avoided if possible. If it becomes necessary to traverse open terrain or come near guard towers, evaders should stay low to the ground and, when possible, travel at night or during inclement weather. Use terrain masking since night vision devices may be used near border areas (figure 28-68).

(10) The problem of crossing areas which have been contaminated as a result of enemy or friendly NBC operations may arise. Chemical contamination should be suspected when the following are observed (figure 28-69):

(a) Shell craters with liquid in the bottom.
(b) Liquid droplets on vegetation.
(c) Water with "film" on the surface.
(d) Unexplained dead animals.
(e) Unseasonal discoloration of vegetation.

NOTE: Without protective clothing, mask, and accessories, evaders should bypass contaminated areas if possible.

(11) CAUTION: Stay away from borders unless absolutely necessary. The crossing of one or more borders presents a major problem. These areas may be located in any type of terrain.

(a) In areas where there is no well-defined terrain feature to indicate the border, artificial obstacles such as electrified or barbed wire fences, augmented with tripwires, anti-personnel mines, or flares may be encountered. Open areas may be patrolled by humans or dogs, or both, particularly during the hours of darkness. The enemy may also use floodlights and plowed strips as aids to detecting evaders (figure 28-70).

(b) The plan to cross a border must be deliberate and must be designed to take advantage of unusually bad weather (as a major distraction to the enemy force) or areas where security forces are overextended. These areas are usually found where there are natural obstacles.

(c) Crossings should be made at night, when possible, in battle-damaged areas. If it is necessary to cross during daylight hours, evaders should select a crossing point which offers the best protection and cover. They should then keep the area under close observation for several days to determine:

-1. The number of guards in the area.
-2. The manner of their posting.
-3. Aerial patrols and their frequency.
-4. The limits of the areas they patrol.
-5. Location of mines, flares, or tripwires.

(12) A difficult task in any situation is the attempt to cross the forward edge of the battle area. If unable to determine the general direction to friendly lines, evaders should remain in position and observe the movement of enemy military forces or supplies, the noise and flashes of the battle area, or the orientation of enemy artillery. After arriving in the combat zone, evaders should select a concealed position from which as much of the battle area as possible may be observed. They can then select a route and critical terrain features on which they can guide when infiltrating back to friendly positions under the cover of darkness. Several alternate

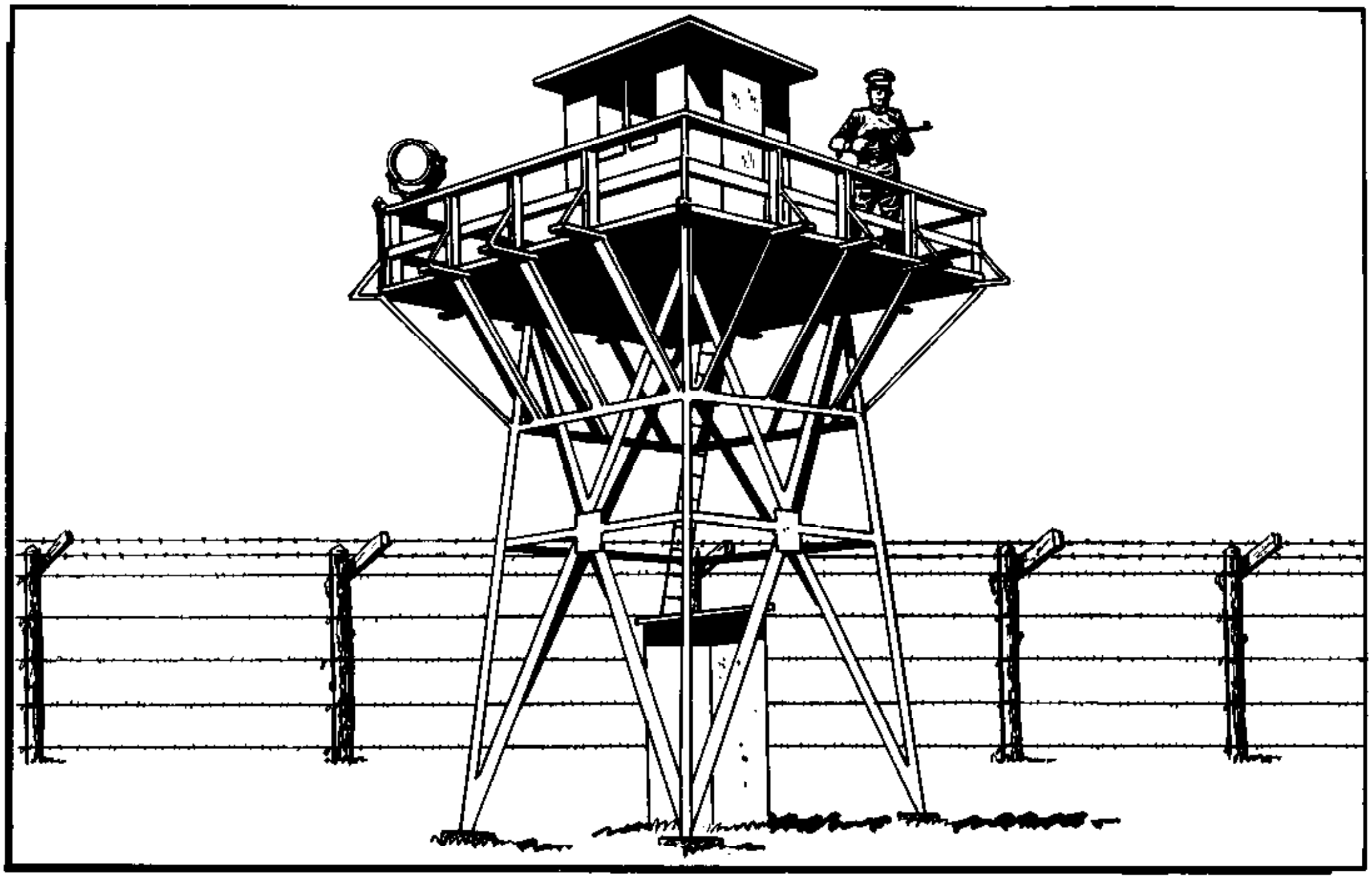

Figure 28-68. Open Terrain.

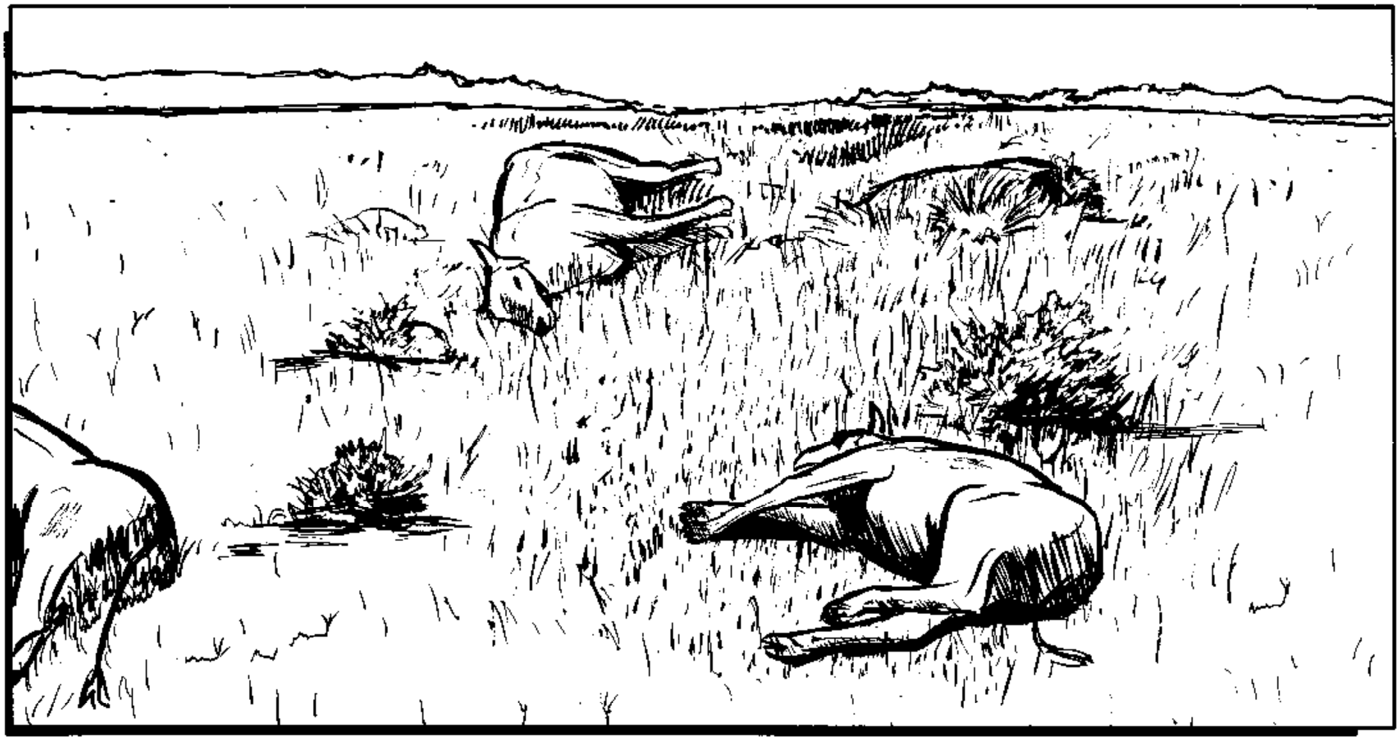

Figure 28-69. Contaminated Areas.

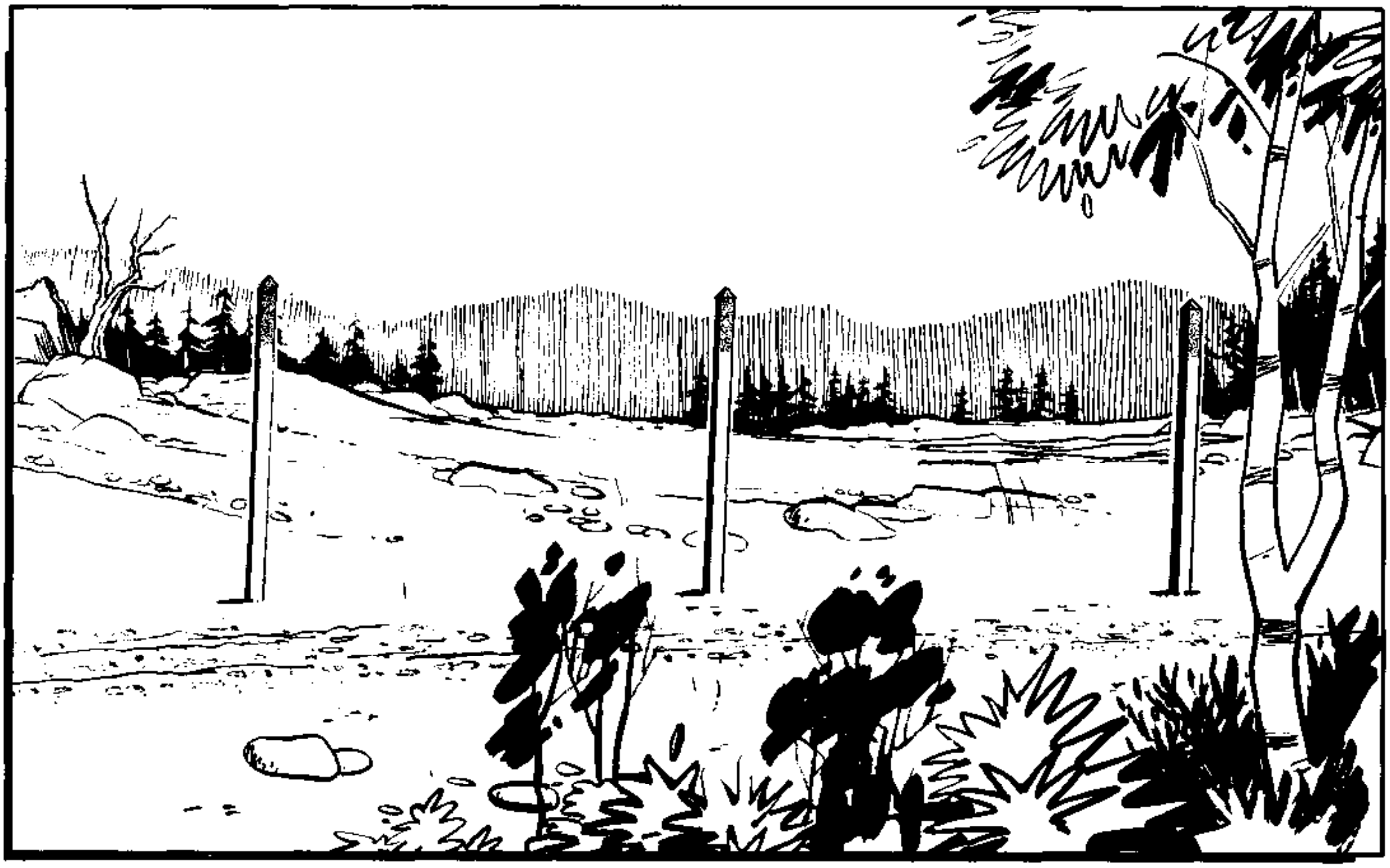

Figure 28-70. Border Crossings.

routes should be selected with care to avoid "easy" approaches to friendly lines which are more likely to be covered by friendly fire and enemy patrols. If in uniform, select exposure time during daylight hours and be close enough to be easily recognized by friendly troops.

(13) Evaders should also watch out for friendly patrols. If a patrol is spotted, evaders should remain in position and allow the patrol to approach them. When the patrol is close enough to recognize them, evaders should have a white cloth displayed before the patrol gets close enough to see the movement. A patrol may fire at any movement. Shouting or calling out to them jeopardizes both the patrol and the evader. Evaders should stand silently with hands over their head and legs apart so their silhouettes are not threatening. If evaders elect not to make contact with a patrol, they should, if possible, observe their route and approach friendly lines at approximately the same location. This may enable them to avoid mine fields and booby traps. (NOTE: The practice of following any patrol is extremely dangerous. The last persons in line are charged with security, and anyone following them would be considered hostile and eliminated.)

(14) If unable to contact a friendly patrol, the only alternative for evaders may be to make a direct approach of front-line positions. This will require them to crawl through the enemy's forward position near forward friendly elements. This action should be done during the hours of darkness. Once near friendly lines, however, evaders should not attempt to make contact until there is sufficient light for them to be recognized.

(15) In dogs, the ability to perceive odors is much greater than that of humans.

(a) In this portion, the term dog is meant to describe only the animals specifically trained in the areas of patrolling, guarding, and searching. Since dogs are basically odor-seeking animals, anything developed to work against their odor-seeking capabilities is worthy of experimentation for survivors (evaders).

(b) A problem which must be considered is evaders will not be working solely against a dog, but against a dog handler as well. There is no simple, sure method of evading a dog. Some possible means which could be tried by evaders are:

-1. Dogs detect the fatty acids in dead-skin cells that humans shed by the thousands every day. Scenting dogs may be distracted by scattering an irritant such as pepper behind the evader or traveling across an asphalt road on a hot day.

-2. Dogs track better when the weather is humid and the air is still—there is less evaporation and dissipation of odor.

-3. If evaders know they are being followed by dogs, either from the landing site or as a result from an escape, they should try to use water to conceal their tracks and eliminate their scent.

-4. If there is a choice of terrain and it is possible to travel on a hard surface, evaders should do so rather than travel on soft ground.

-5. Evaders should always attempt to move downwind of a dog. This should be attempted when they are traveling in open country or penetrating obstacles such as dog guard posts or border areas. If penetration of obstacles or escape is planned (after careful location and study of guard and dog areas of responsibility and their methods), evaders should select a time for movement when noise will distract the dog to a point away from the planned maneuver.

-6. If evaders are physically capable, they should attempt to maintain the maximum distance possible from dogs. Moving fast through rugged terrain will slow and probably defeat the handlers of dogs. Here evaders must choose between making mistakes in travel techniques while evading or being caught by dogs if they don't move fast enough.

28-18. Evasion Aids:

a. Survival Kits. Personnel may sometimes find it practical to devise and carry compact personal survival kits to complement issued survival equipment. If E&E kits are provided, potential evaders should be familiar with them, their uses, and their limitations.

b. Maps. Any maps of the area in an evader's possession should not be marked. A marked map in enemy hands can lead to the compromise of people and locations where assistance was given. Evaders should be wary of even accidentally marking a map; for example, soiled fingers will mark a map just as plainly as a pencil.

c. Pointee-Talkee. The "pointee-talkee" is a language aid which contains selected phrases in English on one side of the page and foreign language translations on the other side. To use it, evaders determine the question and statement to be used in the English text and then point to its foreign language counterpart. In reply, the natives will point to the applicable phrase in their own language; evaders then read the English translation.

(1) The major limitation of the "pointee-talkee" is in trying to communicate with illiterates. In many countries the illiteracy rate can be astoundingly high, and personnel have to resort to pantomime and sign language which have been relatively effective in the past.

(2) "Pointee-talkee" phrases are presented under the following eight subheadings:

(a) Finding an interpreter.
(b) Courtesy phrases.
(c) Food and drink.
(d) Comfort and lodging.
(e) Communications.
(f) Injury.
(g) Hostile territory.
(h) Other military personnel.

d. Barter Kits. Barter kits may be available in some commands. If not, crewmembers may elect to develop their own. Items for consideration should be selected from area studies. Some items to consider might be rings, watches, knives, local currency, coins, and lighters. Items should have no markings of personal significance or military value. Military items packed in kits may be considered if not essential to the evader. Flashing large amounts of cash or valuables can have negative results in a depressed, war-torn area. Show only small amounts and drive a hard bargain.

e. Other Evasion Aids. Information on other evasion aids and tools is available from unit intelligence officers.

f. Assisted Evasion. There may be people in a hostile nation or in an enemy-occupied country who are dissatisfied with existing conditions.

(1) History has revealed that in every major conflict there are groups of people in every country who will aid a representative of their government's enemies. The motivating force behind these groups may vary from purely monetary considerations to idealistic concepts of government reform. In many cases, their real objective will be the political advancement of their particular group. During WW II, many underground or resistance groups and movements were established in occupied countries. One of the major purposes or functions of these groups was the aiding of downed allied aircrew personnel. In most cases, the driving force behind these movements was patriotism and desire for political recognition for their cause.

(2) These circumstances favor active resistance movements. One of the functions of such movements may be the operation of escape and evasion (E&E) systems for the purpose of returning evaders to friendly territory.

(3) US Special Forces may also organize and operate E&E mechanisms (figure 28-71).

(a) E&E Organizations. Assistance may range from that rendered by a sympathetic individual to elaborate E&E nets organized by local inhabitants. E&E organizations may be limited in nature, such as providing assistance to reach a national frontier, or they may be linked to larger organizations capable of returning to friendly control.

(b) Acts of Mercy. These are usually isolated events during which evaders may be provided food, shelter, or medical attention for a brief period of time. Local people may find an exhausted or incapacitated

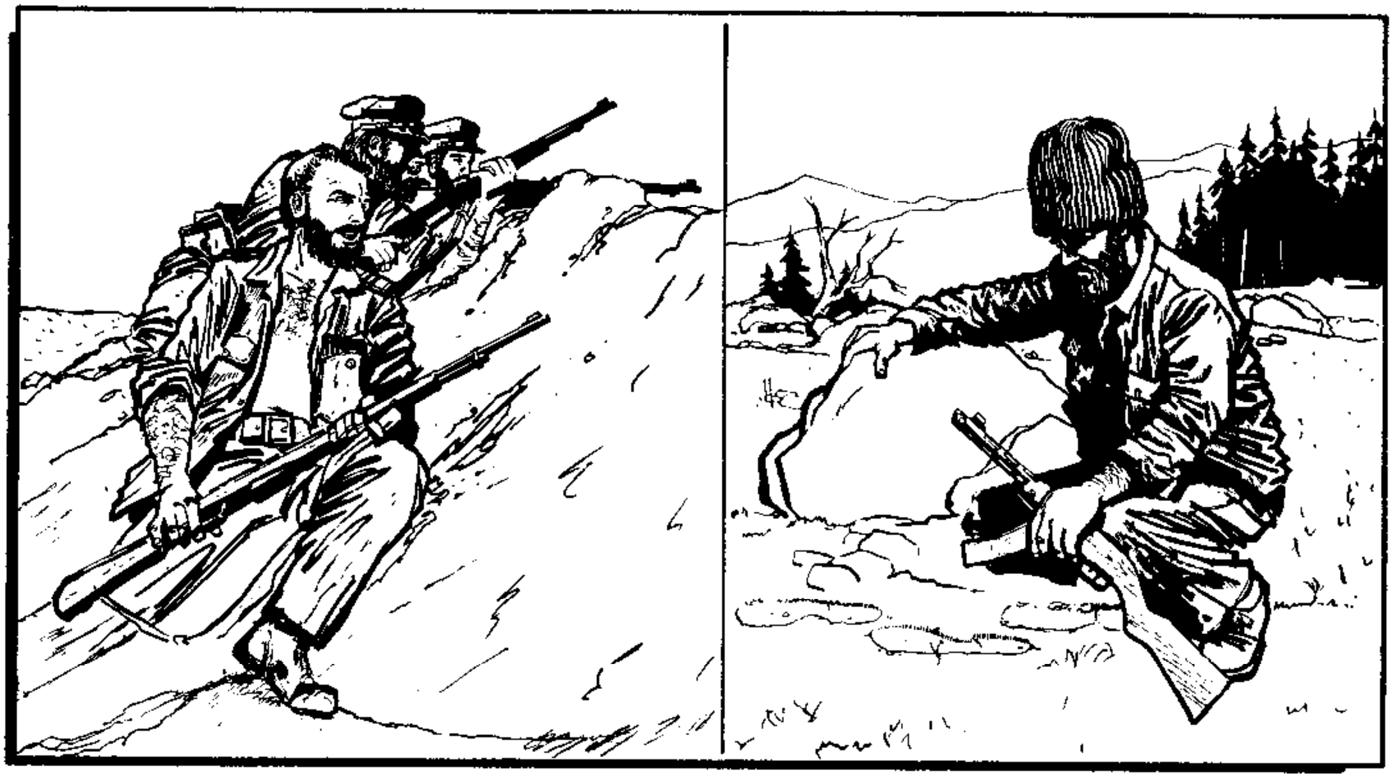

Figure 28-71. E&E Organizations.

evader and provide that evader with limited sustenance. This type of assistance is frequently offered with reluctance or under fear of reprisal because an act of providing comfort to the enemy would mean punishment. Unless an evader is in immediate need of medical attention, acts of mercy may consist of only an offer of food followed by an urgent plea that the evader leave the area immediately. If an evader is physically able to depart with a reasonable chance of evading capture, he or she should do so. An evader should not insist on receiving additional aid other than what is offered by the person who renders assistance only through human impulses (figure 28-72).

(4) Evaders must understand when dealing with any indigenous personnel while in enemy territory, their own actions will govern the treatment they will receive at the hands of these people. How evaders conduct themselves may also have much to do with getting back to their own forces should they fall into the hands of irregulars friendly to their (the evaders) own cause. The following list of suggestions may be a useful guide in dealing with these people or groups.

(a) Evaders should understand that failure to cooperate or obey may result in death.

(b) Evaders should try to avoid making any promises they cannot personally keep.

(c) Evaders should remember that the four conditions called for by the rules of land warfare must be met before members of an irregular group can be recognized as qualifying for PW status in the event of capture. These same rules will also apply to evaders who are involved with these groups.

(d) Prisoners of war, according to the current Geneva Convention, are persons belonging to one of the

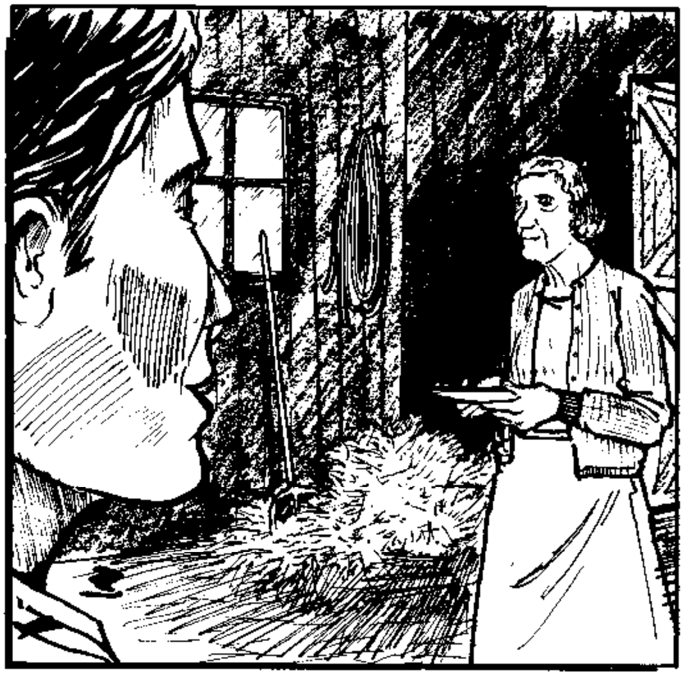

Figure 28-72. Receiving Aid.

following categories, who have been captured by the enemy. Members of other militias and volunteer corps, including those of organized resistance movements, belonging to a conflicting Party and operating in any territory, even if this territory is occupied, provided that such groups fulfill the following four conditions of:

-1. Being commanded by a person responsible for subordinates.

-2. Having a fixed distinctive sign recognizable at a distance.

-3. Carrying arms openly.

-4. Conducting their operations in accordance with the laws and customs of war.

(5) If evaders join such a unit, the closest thing to a guarantee of treatment as a military person will be their uniforms if they are captured.

(6) If evaders have the opportunity to influence the group they are with, they should try to encourage them to abide by the four conditions mentioned.

(7) Evaders must avoid becoming associated in any way with atrocities these groups may commit against civilians, prisoners, or enemy soldiers.

(8) Evaders should not become involved in their political or religious discussions, take sides in their arguments, or become involved with the opposite sex.

(9) Evaders should show consideration for being allowed to share food and supplies. It may also be helpful if evaders understand and show interest in the assister's customs and habits.

(10) The overall best and safest course for evaders to follow is to exercise self-discipline, display military courtesy, and be polite, sincere, and honest. Such qualities are recognized by any group of people throughout the world. The impression left can influence the aid provided to future evaders.

g. E&E Lines. An E&E line is a system of one or more secret nets organized to contact, secure, and when possible, evacuate friendly personnel. Well-organized and supported lines normally can be expected to provide the following assistance:

(1) Temporary shelter, food, and equipment for the next phase of the journey.

(2) Clothing and credentials acceptable in the area to be traveled.

(3) Information concerning enemy security measures along the evasion route.

(4) Local currency and transportation.

(5) Medical treatment.

(6) Available native guides.

h. Conduct of E&E Lines. The success of an E&E organization depends almost entirely upon its security. The organization of a line includes much planning and work carried out under dangerous conditions. The security of the system often depends upon the evader's cooperation and working knowledge of how it functions, how it may be contacted, and what rules of personal conduct are expected of the evader. The following paragraphs summarize the major aspects of the operation of an E&E line.

(1) Contacting the Line. Premission briefings may inform evaders where to go and what actions to take to make contact with an E&E mechanism. After being picked up by an evasion and escape mechanism, evaders will be moved under the control of this mechanism to territory under friendly control or to a removal area, and arrangements may be made for air or sea rescue. The organizer of a line in friendly but enemy-occupied territory normally will have arranged a network of spotters who will be especially active when evaders are in the immediate area, but so will the enemy police and counterintelligence organizations. For this reason, certain precautions must be observed when making contact.

(2) Approach. Help may be refused by a person simply because he or she thinks someone else has seen the evaders approach to seek assistance. If captured with a local helper, an evader will become a prisoner, but the helper and perhaps an entire family may be more severely punished.

(3) Making Contact. Contacts with the natives are discouraged unless observation shows they are dissatisfied with the local governing authority, or previous intelligence has indicated the populace is friendly. Evaders should proceed to, and remain in, the nearest SAFE area where arrangements for contact can be developed. If the E&E system is operating successfully, the spotter will know evaders are present and will search the immediate area, making frequent visits to designated contact points. Identification signs and countersigns, if used, may be included in the preoperational briefing. It is seldom advisable to seek first contact in a village or town. Strangers are conspicuous by day, and there may be curfews or other security measures during the hours of darkness. The time of contact should be at the end of the daylight period or shortly thereafter. Darkness will add to the chance of escape, if the contact proves to be unfriendly, and may be advantageous to the contact in providing further assistance.

(4) Procedure After Contact. If contact is made, evaders may be told to remain in the vicinity where spotted, or, more likely, they will be taken to a house or other structure used by the E&E net as a holding area. It must be decided at this time whether or not to trust the contact. If there is any doubt, plans should be made to leave at once. It is also possible that the house may not belong to the E&E organization but rather to someone who will look after evaders until arrangements can be made for the line to identify and accept them in E&E net.

(5) Establishing Identity. Verification of identity will be required before anyone is accepted as a bona fide evader. The constant danger facing the operators of an

escape line is the penetration of the E&E system by enemy agents pretending to be evaders or escapees. Evaders should be prepared to furnish proof of identity or nationality. Since it may lead to later difficulties of identification, they should never give a false name—just their name, grade, service number, and date of birth. It is best for them to avoid talking as much as possible.

(6) Awaiting Movement on the Line. Delays can be expected while proceeding along the escape line. If the period of waiting is prolonged, frustration and impatience may become unbearable, leading to a desire to leave the holding area. This must not be done, because if seen by other people, the lives of the assisting personnel and the existence of the entire line itself may be endangered.

(a) Evaders should follow the orders of those assisting them. If kept indoors for any length of time they can keep fit by moderate physical exercise.

(b) The host should have a plan for rapid evacuation of the area if enemy personnel should raid the holding area; if not, evaders should have a personal plan to include measures for moving all traces of having occupied the area. If the net is being overrun and capture is imminent, evaders must be prepared to fend for themselves. The evader is the only one who knows more than one part of the net. The assisters may attempt to eliminate the security risk to the net.

(7) Traveling the Line. It would be a grave breach of faith and security for evaders to discuss, at any point on the line, the earlier stages of the journey. Evaders might be tested to see if they are trustworthy—they should discuss nothing of the net. For security reasons and to protect the compartmentalization of the line, no information should be revealed. It is also useless to ask where a line leads or how it will eventually reach friendly territory. Evaders should not try to learn or memorize names and addresses and, above all, they shouldn't put these facts or any other information in writing. Evaders should give the impression of having received no assistance from local inhabitants.

(8) Fellow Evaders. Caution is required in the case of fellow evaders on an escape line unless they are personally known. Even when it has been satisfactorily determined that another person is a genuine evader, no information should be given.

(9) Travel with Guides. If under escort, this fact should not be apparent to outsiders. In a public vehicle, for example, evaders should never talk to their guide or appear to be associated with the person in any way unless told to do so. This will lessen the possibility of both the evader and the guide being apprehended if one should arouse suspicion. It should always be possible for the guide to disown an evader if the guide gets into difficulty. When evaders are escorted, they should follow the guide at a safe distance, rather then walk right next to the person, unless instructions indicate the latter action is required (figure 28-73).

(10) Speaking to Strangers. Evaders should never speak to a stranger if it can be avoided. As a last resort, they should pretend to be deaf and dumb or even half-witted. This technique has often been successful. To discourage conversation in a public conveyance, they can also pretend to read or sleep.

(11) Personal Articles and Habits. Evaders should not produce articles in public which might show their national origin. This pertains to items such as pipes, cigarettes, tobacco, matches, fountain pens, pencils, and wristwatches. Evaders should also ensure their personal habits do not give them away; for example, they should not hum or whistle popular tunes or utter involuntary oaths. Again, in restaurants, imitating local customs in the use of knives and forks and other table manners is advisable. Study of the area before the mission may help evaders avoid making mistakes.

(12) Payment to Helpers. On an escape line, evaders should not offer to pay for board, lodging, or other services rendered. These matters will be settled afterwards by those who are directing and financing the line. If in possession of escape kits or survival packs, evaders should keep them as reserves for emergencies. If they

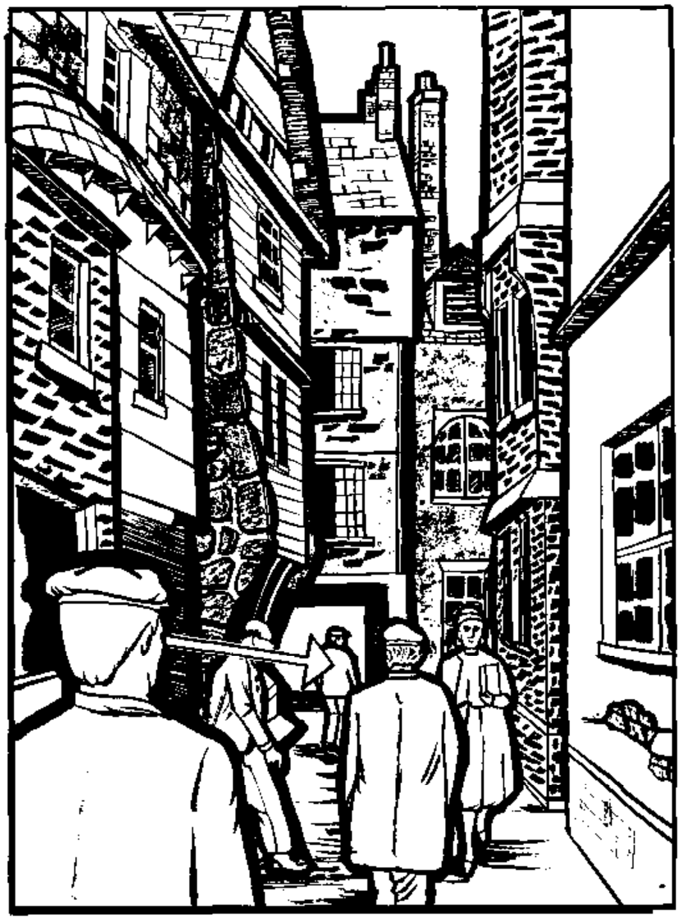

Figure 28-73. Travel with Guide.

have no food reserve, they should try to build up a small stock in case they are forced to abandon the line.

(13) Evaders Conduct:

(a) Be polite by local standards.

(b) Be patient and diplomatic.

(c) Avoid causing jealousy. Disregard the sex of the assisters.

(d) Avoid discussions of a religious or political nature.

(e) Eat and drink if asked, but don't over indulge or become intoxicated.

(f) Don't take sides in arguments between assisters.

(g) Don't become inquisitive or question any instruction.

(h) Help with menial tasks as directed.

(i) Write or say nothing about the other people or places in the net.

(j) Don't be a burden; care for self as much as possible.

(k) Follow all instructions quickly and accurately.

28-19. Combat Recovery:

a. Air recovery is one of several means of transportation for downed crewmembers in their quest for their final goal—returning to their own lines and units. How recovery will be effected will depend on a number of factors, among them are the following:

(1) Terrain.

(2) Capability of the rescue vehicle.

(3) Condition of the survivor (evader).

(4) Enemy activity in the area.

(5) Weather conditions.

(6) High or priority mission.

b. Even though a maximum effort will be made to recover downed aircrew members, survivors can jeopardize the whole rescue operation and the lives of rescue personnel in a combat zone by not taking a responsible role in recovery operations. The responsibilities are many and varied but essential to a successful rescue mission. Evaders should recall that even though they may have little experience in participating in actual rescues, they must nonetheless be very proficient in their actions. Evaders should always remember other people are endangering their lives in an attempt to retrieve them.

c. There are two basic types of air-recovery vehicles—rotary wing and fixed-wing aircraft. The rotary wing type can extract evaders from remote areas. With air-to-air refueling, the range of these aircraft has increased; they are limited only by the endurance of the crewmembers who operate them.

d. Survivors must be proficient in the operation of all the survival equipment at their disposal. For example, they must be able to switch off the beeper on the radio to receive verbal instructions from the rescue commander.

e. The initial contact with rescue or other combat aircraft in the area must be done as directed by authorities in that theater of operations; for example, in Southeast Asia, a contact method of transmitting 15-seconds beeper, 15-seconds voice (call sign), and 15-seconds monitoring was used. The method an evader should use to establish contact with rescue will be briefed before the mission.

f. One important aspect of the rescue process is evader authentication by rescue personnel. Survivors must be able to authenticate their identity through the use of questions and answers, responding as directed before the missions. Authentication in a combat area changes rapidly to reduce the chances of compromising the rescue efforts. Survivors (evaders) must keep up with these changes so their rescue can be made without undue danger to rescue personnel.

g. The purpose of filling out and using a personal authenticator card is positive verification of an evader's identity which is essential before risking search and rescue aircraft or the lives of assisters. The purpose of the photographs, descriptions, fingerprints, etc., on the card is to ensure all possible means are used for the proper identification of personnel. Intelligence personnel are responsible for ensuring that personnel fill out the card completely and that they are aware of the purpose and use of the information. When filled in, the personal authenticator card becomes classified Confidential, and is reviewed at least semiannually by both the crewmember and intelligence personnel. (NOTE: See AFR 64-3, Wartime Search and Rescue (SAR) Procedures.)

h. When selecting a site for possible evasion recovery, there is much for evaders to keep in mind. The area they choose could well decide the success or failure of the mission.

(1) The potential recovery site should be observed for 24 hours, if possible, for signs of:

(a) Enemy or civilian activity.

(b) Roads or trails.

(c) Farming signs.

(d) Orchards.

(e) Tree plantations.

(f) Domesticated animals or droppings.

(g) Buildings or encampment areas.

(2) The rescue site should be observable by aircraft but unobservable from surrounding terrain if possible. There should also be good hiding places around the area. The site should include several escape routes so evaders can avoid being trapped by the enemy if discovered. There should be a small open area for both signaling and recovery. It would be beneficial if the surrounding terrain provided a masking effect for rescue forces in order to avoid enemy observation and fire.

i. The size of evasion ground-to-air signals should be as large as possible but must be concealed from people passing by. Evaders should remember the six-to-one ratio. The contrast these signals make with the surrounding vegetation should be seen from the air only. Any signal displays should be arranged so they can be removed at a moment's notice since enemy aircraft may also fly over the immediate area.

(1) All of the principles of regular (nontactical) signals should be followed by those building evasion signals. Crewmembers may be prebriefed as to the appearance of specially shaped signals (figure 28-74).

(2) Evasion signals, like all others, must be maintained to be effective. At times, evaders may be preinstructed to set out their signals according to a prearranged time schedule.

j. Whatever signal devices are available to evaders, they must be able to use them (mirror, flare, gyro-jet, etc.) with proficiency. These signals, like those used in nontactical situations, are to be used either to gain the attention of friendly aircraft or rescuers, or when directed by rescue. Extreme care must be taken to minimize or eliminate chances of enemy elements spotting the signals. For example, the strobe light and mirror can be directed and aimed instead of being used in an indiscriminate manner.

(1) If evaders are in or on the water, they should use lights, flares, dye, whistles, etc., with extreme care as they are readily distinguishable over water at great distances.

(2) In addition to knowing how to use radios correctly to effect their rescue, evaders should also be familiar with the special points of evasion radio procedures; for example, ensuring the radio transmits continuously during a night recovery operation using the electronic locator finder (ELF) system.

(3) It is also the evaders' responsibility to communicate all signs of enemy activity, such as:

(a) Locating anti-aircraft emplacements.

(b) Identifying when they are firing.

(c) Assisting strikes by spotting hits (high, low, or on target).

(d) Determining effectiveness of hits.

(e) Notifying personnel of changes in small arms positions, etc.

(4) Downed evaders can normally expect to be hoisted to a helicopter by one of five methods: basket, Stokes litter, bell, horse collar, or forest penetrator. Other pickup devices which may be used are the McGuire Rig, Swiss Seat, Motley Rig, Stabo Rig, rope ladder, or rope. Another method of recovery which may be employed to extract evaders is the Surface-to-Air-Recovery (STAR) System (figure 28-75). Evaders should remember whatever the pickup device, it should always be grounded before they grasp it.

(5) There is also fixed-wing capability of rescuing downed crewmembers. Evaders will be prebriefed as to which type rescue vehicle and systems to expect in their areas of operation. No matter which type is used, evaders must be capable of mounting and riding the rescue

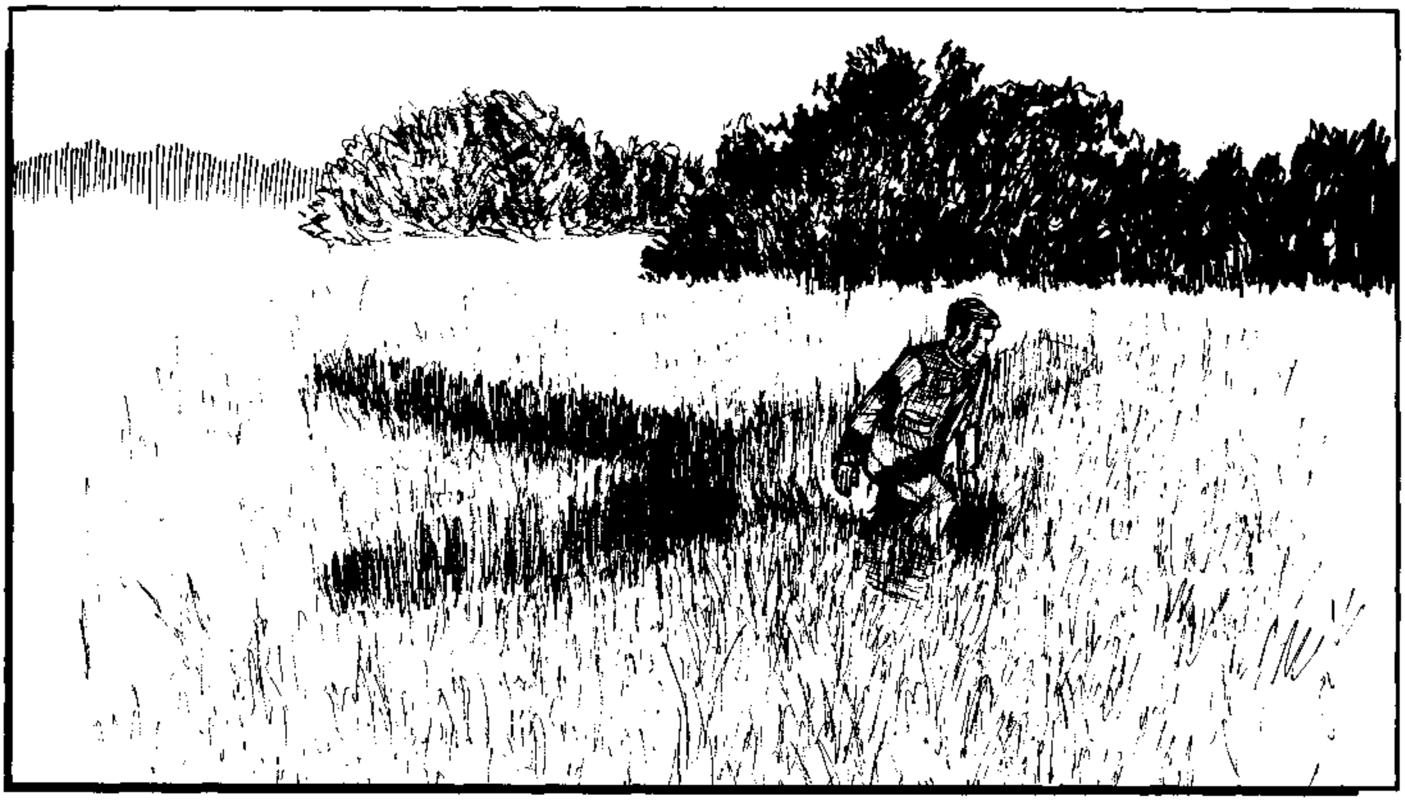

Figure 28-74. Evasion Signals.

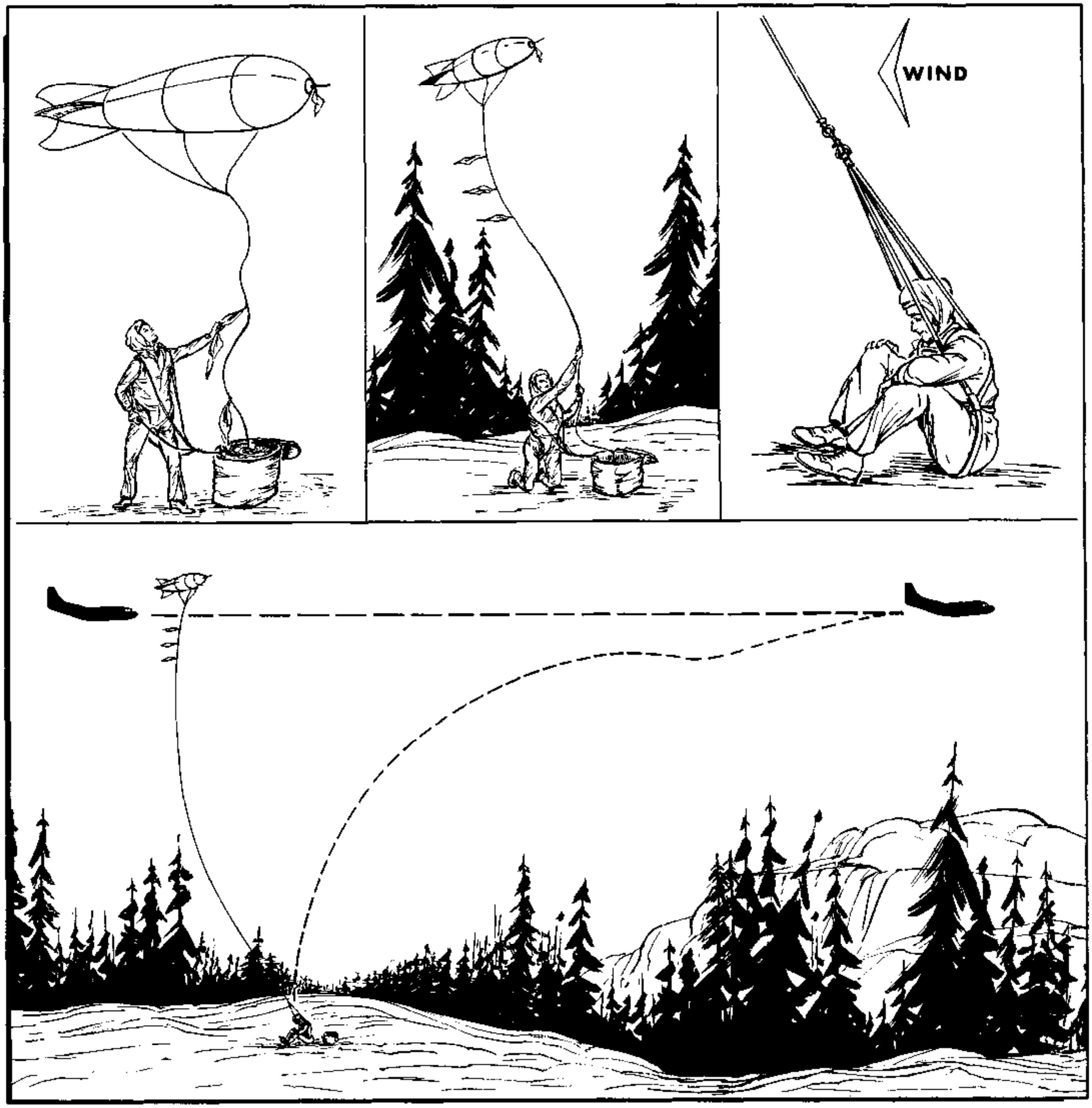

Figure 28-75. Surface-to-Air Rescue.

devices in a minimum amount of time, allowing the rescue craft to effect the rescue and depart quickly.

(6) As downed crewmembers, whether as a result of enemy action or of mechanical failure, it is important they all become familiar with the information that 1 day could prove instrumental in saving their lives. First, they must have prepared themselves to cope with the survival situation before an aircraft emergency. This can be done by including, as part of their preflight planning, a thorough inspection of survival equipment to determine its availability and serviceability. Second, they must realize there are many types of recovery vehicles available in the Air Force inventory, and a knowledge of all recovery techniques is a necessity. They should request detailed briefings from area or local rescue personnel who will explain the operational limita-

tions and recovery potential for each. Third, evaders must be capable of fulfilling their part in the recovery operation. This can be done by knowing when, where, and how to initiate communications and how to cooperate with the rescue aircraft crew.

(7) Other forms of assistance which may be available to evaders are Special Forces, Combat Control, RECON, SEALS, Riverine Operations, and Submarines. While aiding evaders is not the primary mission of these groups, as a secondary mission, they may travel in or near SAFE areas on their return journey to check for the presence of evaders.

Part Ten

INDUCED CONDITIONS — NUCLEAR, BIOLOGICAL, AND CHEMICAL (NBC)

Chapter 29

NUCLEAR CONDITIONS

29-1. Introduction. The possibility of "induced conditions" has served to intensify the difficulties of basic and combat survival because of the serious problems posed by nuclear, biological, and chemical warfare. Though the prescribed survival procedures recommended in other parts of this regulation are still applicable, a number of additional problems are created by the hazards of induced conditions (figure 29-1).

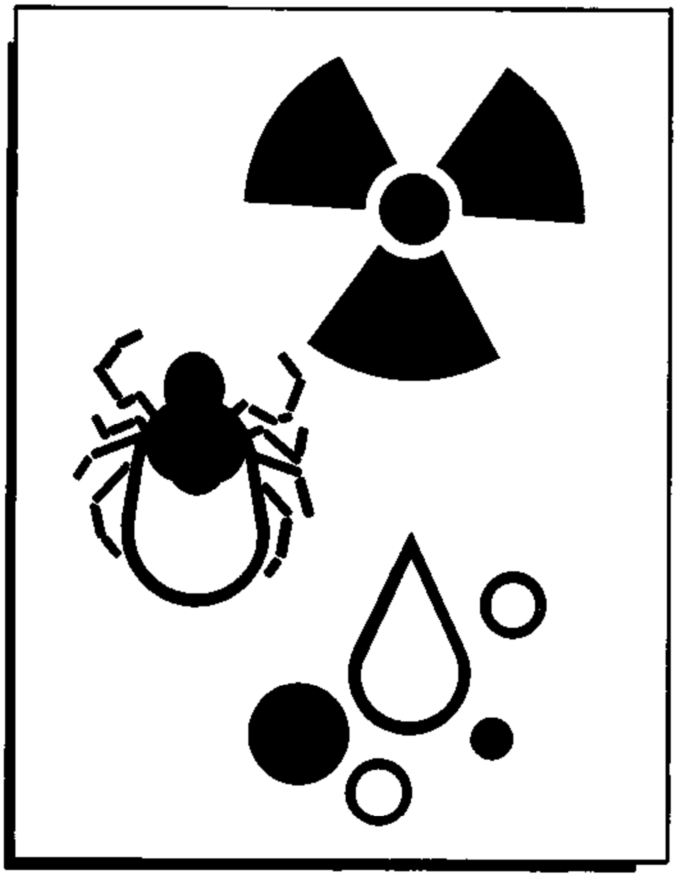

Figure 29-1. Induced Conditions.

29-2. Effects of Nuclear Weapons. Nuclear weapons cause casualties and material damage through the effects of blast, thermal radiation, and nuclear radiation. The degree of hazard from each of these effects depends on the type of weapon, height of the burst, distance from the detonation, hardness of the target, and explosive yield of the weapon.

a. Blast. The blast wave is the cause of most of the destruction accompanying a nuclear blast. After a nuclear detonation, a high-pressure wave develops and moves outward from the fireball. The front of the wave travels rapidly away from the fireball as a moving wall of highly compressed air. An example of the speed of the blast wave is: At 1 minute after the burst, when the fireball is no longer visible, the blast wave has traveled about 40 miles and is still moving slightly faster than the speed of sound. There are strong winds associated with the passage of the blast wave. These winds may have a peak velocity of several hundred miles an hour near ground zero. Ground zero is the point on the ground directly above or below the point of detonation. The overpressure, which is the pressure in excess of the normal atmospheric pressure, and the winds are major contributors to the casualty and damage-producing effects of the nuclear detonation. The overpressure can cause immediate death or injury to personnel and damage to material by its crushing effect. The high-speed winds propel objects, such as tree limbs or debris, at great speed and turn them into potentially lethal missiles. These winds can also physically throw personnel who are not protected, resulting in casualties. People both inside and outside of a structure may be injured as a result of blast damage to that structure; those inside by the collapse of the structure and by fire; and those outside by the flying objects carried by the winds (figure 29-2).

b. Thermal Radiation:

(1) Heat. Within less than a millionth of a second of the detonation of a nuclear weapon, the extremely hot weapon residues radiate great amounts of energy. This leads to the formation of a hot and highly luminous, spherical mass of air and gaseous residue which is the fireball. The heat radiated from the fireball contributes to the overall damage caused by a nuclear burst by igniting combustibles and thus starting fires in buildings and forests. These fires may spread rapidly among the debris produced by the blast. In addition, this intense heat can burn exposed personnel at great distances from ground zero where the effects of blast and initial nuclear radiation become insignificant. The degree of injury from thermal radiation becomes more marked with the increasing size of the weapon. The degree of injury from thermal radiation is also affected by weather and terrain. During periods of limited visibility, the heat effect will be reduced significantly. Additionally, since ther-

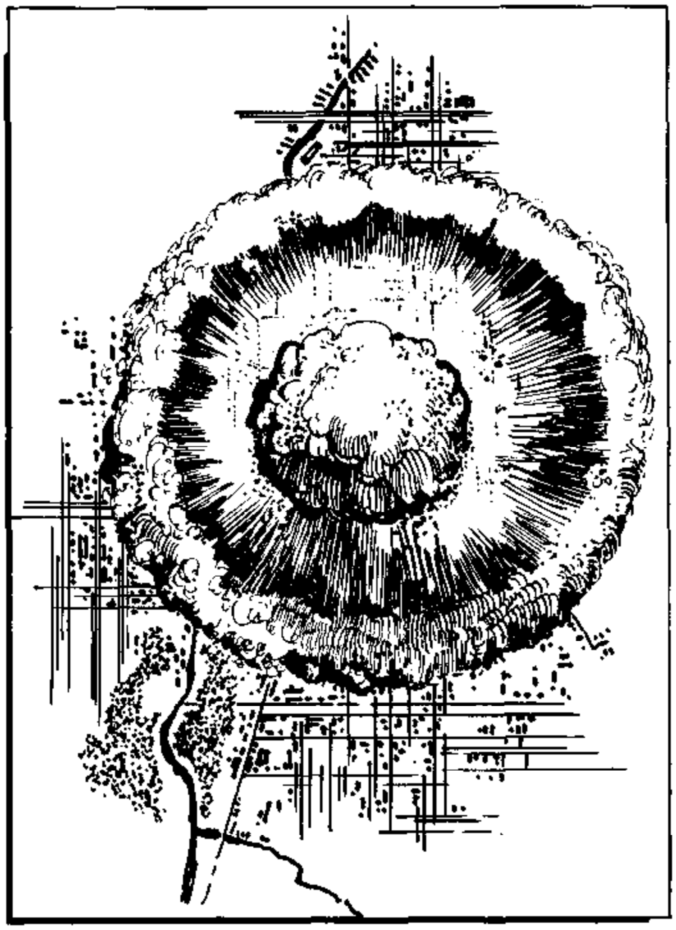

Figure 29-2. Blast.

mal radiation is primarily a line-of-sight phenomena, terrain masking can help reduce its effects (figure 29-3).

(2) Light. The fireball formed at the instant of a nuclear detonation is a source of extremely bright light. To an observer, 135 miles away from the explosion, the fireball of a 1-megaton weapon would appear to be many times more brilliant than the Sun at noon. The surface temperatures of the fireball, upon which the brightness depends, do not vary greatly with the size of the weapon. Consequently, the brightness of the fireball is roughly the same, regardless of the weapon yield. This light can cause injuries to personnel in the form of temporary or permanent blindness. Temporary blindness from a burst during daylight should be of very short duration and is not an important consideration for anyone other than aircrew members. At night, this loss of vision will last for longer periods because the eyes have been adapted to the dark. However, recovery should be complete within 15 minutes. The light flash can cause permanent injury to the eyes due to burns within the eye, but this is only likely to occur in personnel who happen to be looking directly at the fireball at the instant of explosion (figure 29-4).

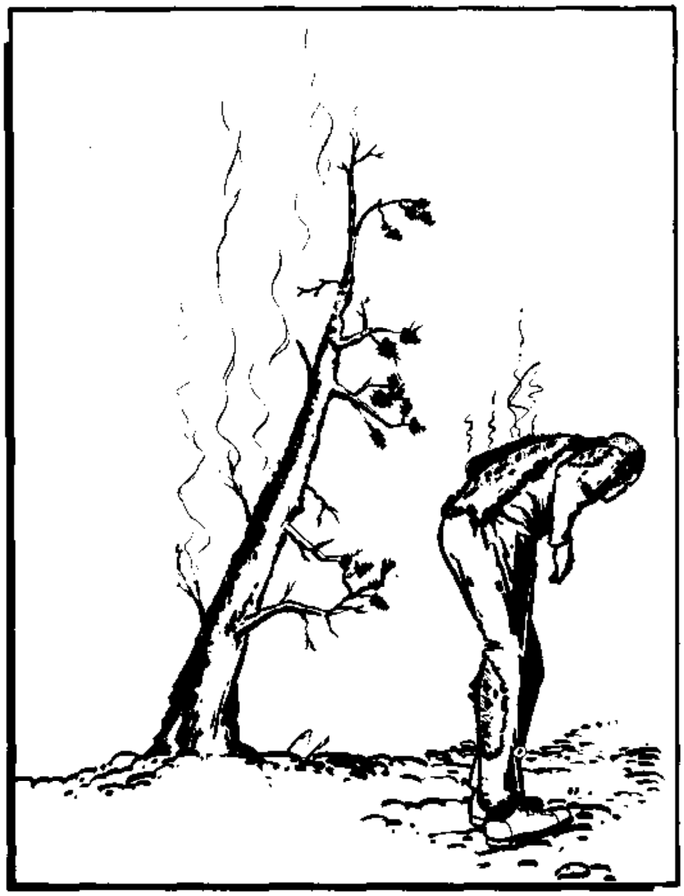

Figure 29-3. Thermal Radiation.

c. Nuclear Radiation:

(1) Initial nuclear radiation is the radiation emitted in the first minute after detonation. For practical pur-

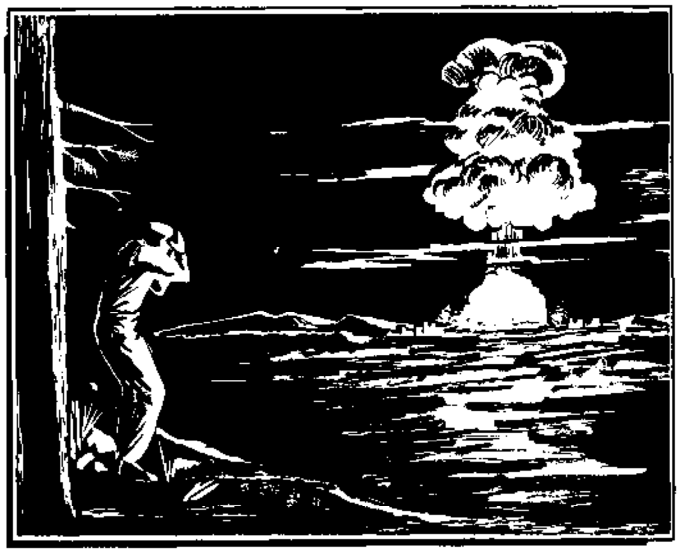

Figure 29-4. Light.

poses, it consists primarily of neutrons and gamma rays. Both of these types of radiation, although different in character, can travel considerable distances through the air and can produce harmful effects in humans. Gamma rays are invisible rays similar to X rays. These penetrating rays interact with the human body and cause damage to tissues and the blood-forming cells. The effects of neutrons on the body resemble those of gamma rays. They are highly penetrating and are easily absorbed by human tissue. Neutron radiation can penetrate several inches of tissue. The neutron radiation produces extensive tissue damage within the body. The major problem in protecting against the effects of initial radiation is that a person may have received a lethal or incapacitating dose of radiation before taking any protective action (figure 29-5).

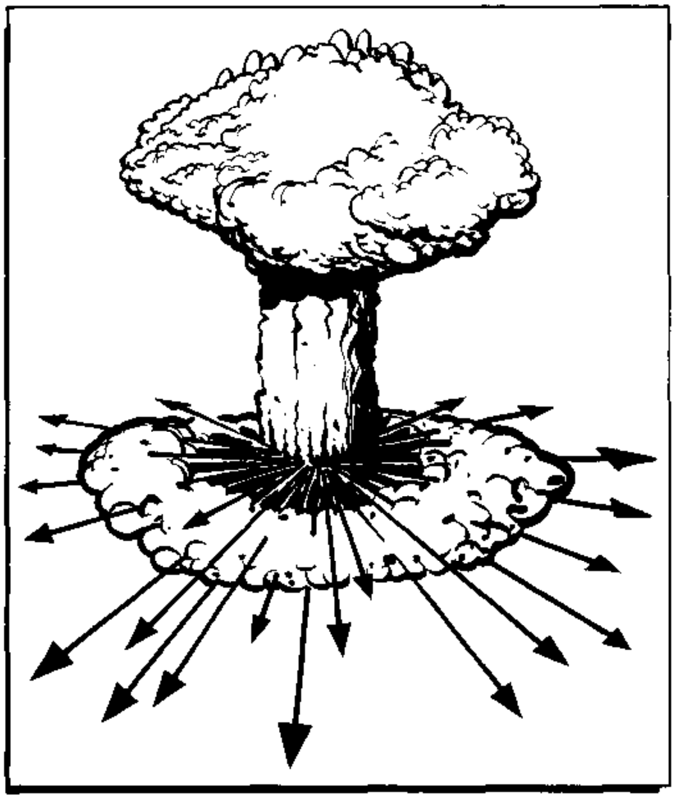

Figure 29-5. Nuclear Radiation.

(2) Residual nuclear radiation is that which lasts after the first minute and consists primarily of fallout and neutron-induced radiation.

(a) The primary hazard of residual radiation results from the creation of fallout. Fallout is produced when material from the Earth is drawn into the fireball, vaporized, combined with radioactive material, and condensed to particles which then fall back to the Earth. The larger particles fall back immediately in the vicinity of ground zero. The smaller particles are carried by the winds until they gradually settle on the Earth's surface. The contaminated areas created by fallout may be very small or may extend over many thousands of square miles. The dose rate may vary from an insignificant level to an extremely dangerous one for all personnel not taking protective measures.

(b) A secondary hazard which may arise is the neutron-induced radioactivity on the Earth's surface in the immediate vicinity of ground zero. The intensity and extent of the induced radiation field depend on the type of soil in the area around ground zero, the height of the burst, and the type and yield of the weapon. The only significant source of residual radiation from an airburst weapon is induced activity in the soil of a limited circular pattern directly beneath the point of burst (figure 29-6).

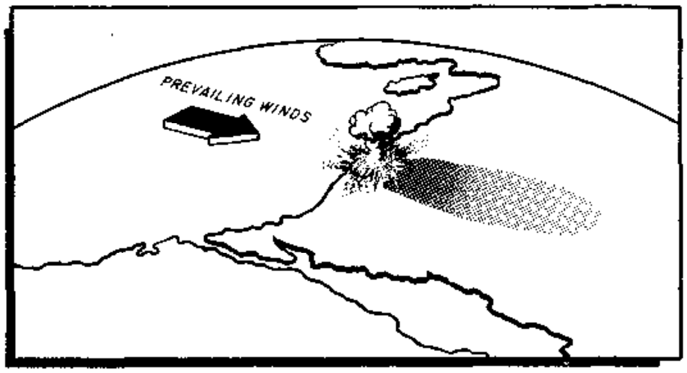

Figure 29-6. Residual Radiation.

29-3. Types of Nuclear Bursts. Nuclear bursts may be classified into three types according to the height of burst—airbursts, surface bursts, and subsurface bursts (figure 29-7).

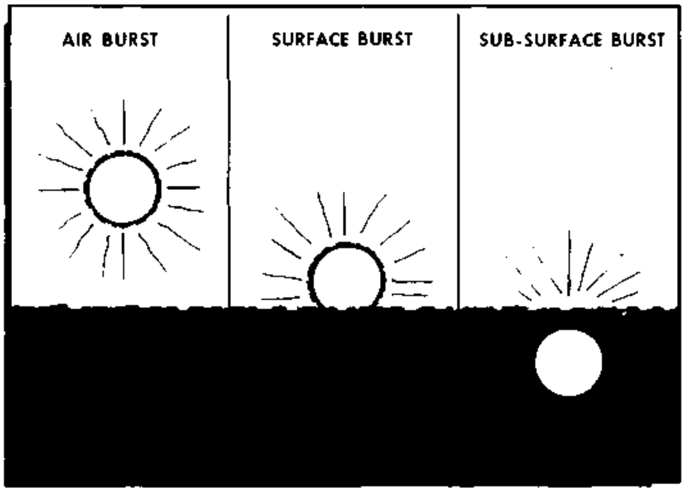

Figure 29-7. Types of Blasts.

a. Airburst. The detonation of a nuclear weapon at such a height that the fireball does not touch the surface of the Earth is called an airburst. Blast, thermal radiation, and initial radiation effects are increased in a low airburst. Fallout of radioactive material from an airburst is not of survival significance unless rain or snow falls through the radioactive cloud and brings the material to Earth. Neutrons from the detonation will cause induced radiation in the soil around ground zero. Except for very high airbursts, neutron-induced radiation in the area of ground zero will be of concern to survivors who are required to go into or across the area. Radiological monitoring will be required as units pass through such an area so that hazardous levels of radiation can be detected and avoided, if possible.

b. Surface Burst. The detonation of a nuclear weapon at such a height that the fireball touches the surface of the Earth or water is called a surface burst. Blast, thermal radiation, and initial nuclear radiation are not as widespread as from an airburst. Induced radiation is present but will be masked by residual radiation from fallout. The fallout produced by a surface burst is by far its most dangerous effect because the burst picks up a great deal more debris and radioactivates this debris; and, depending on the prevailing winds, the fallout covers thousands of square miles with high levels of radioactivity.

c. Subsurface Burst. The detonation of a nuclear weapon so that the center of the fireball is beneath the surface of the Earth or water is called a subsurface burst. If a fireball of this type breaks through the surface, fallout will be produced. Thermal radiation will not be a significant hazard since it will be almost completely absorbed by the soil. Blast effects will also be significantly reduced. Shock waves passing through the ground or water will extend for a limited distance. The range of the initial nuclear radiation will be considerably less than from either of the other two types of bursts because this will also be absorbed to a great extent by the soil. However, extremely hazardous residual radiation will occur in and around any crater. If the fireball does not break the surface, shock waves will pass through the ground and craters may result due to settling.

29-4. Injuries. The explosion of a nuclear bomb can cause three types of injures—blast, thermal radiation, and nuclear radiation. Many survivors receive a combination of two or all three of the above injuries. For example, an unprotected person could be killed by a piece of debris, could be burned to death, or could be killed by initial nuclear radiation if the person is within a few thousand yards from the center of the blast (figure 29-8).

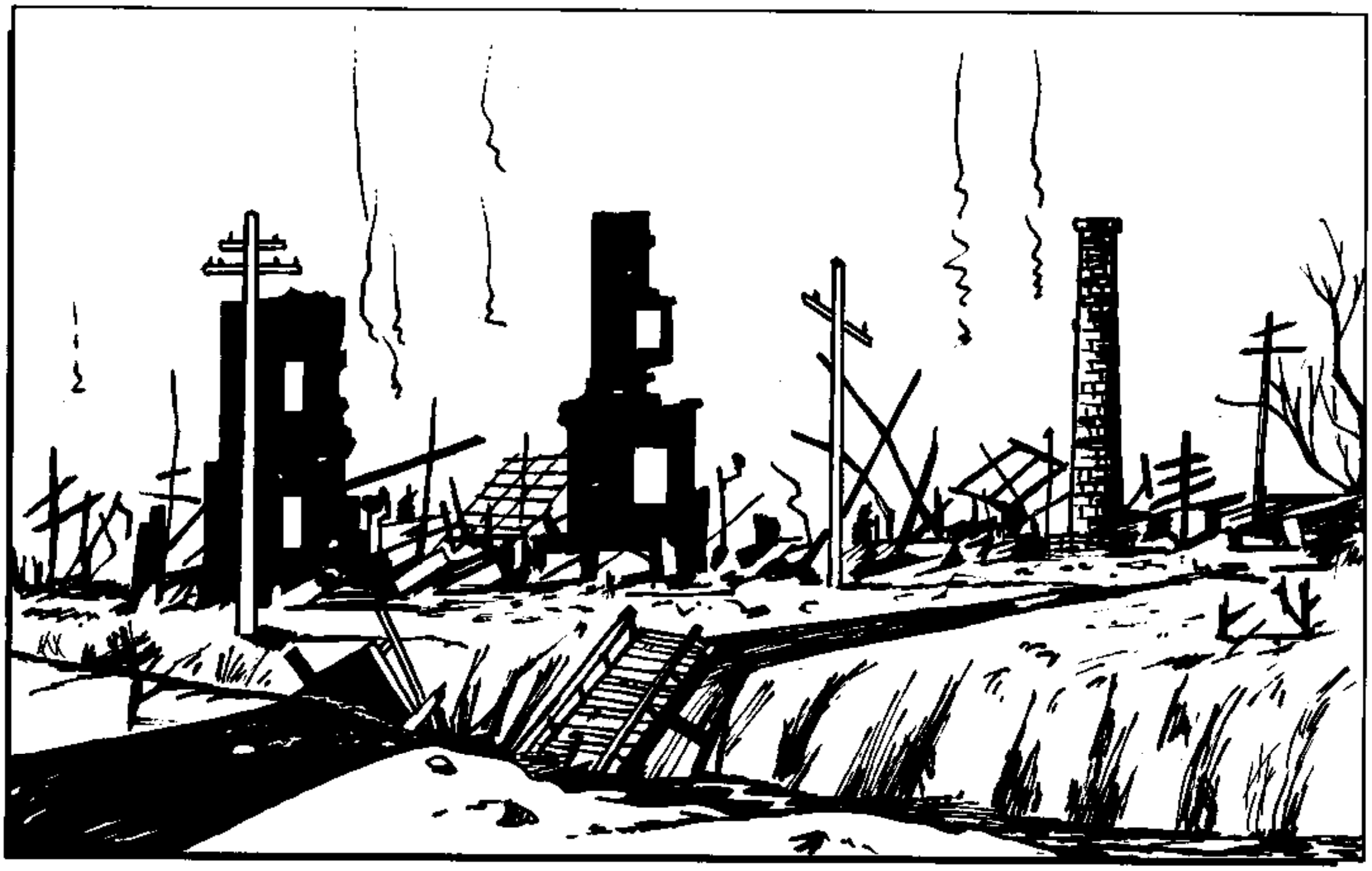

Figure 29-8. Injuries.

a. Blast Injuries. Direct blast can cause damage to lungs, stomach, intestines, and eardrums, or can cause internal hemorrhaging. However, the direct blast is not considered a primary cause of injury because those close enough to suffer serious injury from the direct blast will probably die as a result of initial thermal radiation, or they will be crushed to death. The greatest number of blast injuries are received as an indirect result of the blast from falling buildings, flying objects, and shattered glass.

b. Thermal Radiation Injuries. Burns are classified in degrees according to the depth to which the tissues are injured. In first-degree burns, the skin is reddened as in sunburn. In second-degree burns, the skin is blistered as from contact with boiling water or hot metal. In third-degree burns, the skin is destroyed or charred and the injury extends through the outer skin to deeper tissues. The degree of burn received from thermal radiation depends upon weather conditions, distance from the explosion, and available protection. Many thermal casualties are compounded by nuclear radiation and indirect blast injury. This makes it difficult to attribute casualties to thermal radiation alone.

c. Nuclear Radiation Injuries. The injurious effects of nuclear radiation from a nuclear explosion represent a new threat which is completely absent in conventional explosions. This does not infer that this source of injury is the most important in a nuclear explosion. Rays from radioactive material are not as great a hazard as people fear. The amount of danger from fallout depends upon where and how the nuclear bomb explodes and how well the person is protected. The greatest danger from residual radiation (fallout) comes from exposure for a long period of time to radioactive particles which are nearby, or from dust settling on the body or clothing. Since fallout (like X rays) can destroy living tissue, particularly in the blood-forming system, the exposure of persons working in a radioactive or "hot" area must be controlled so as not to exceed a safe limit. Although a person can become seriously ill and even die from breathing radioactive dust, there is less danger from this than when the whole body is exposed to fallout. Remember, all types of radiation are dangerous (nuclear, thermal, X ray, or even that from an infrared lamp).

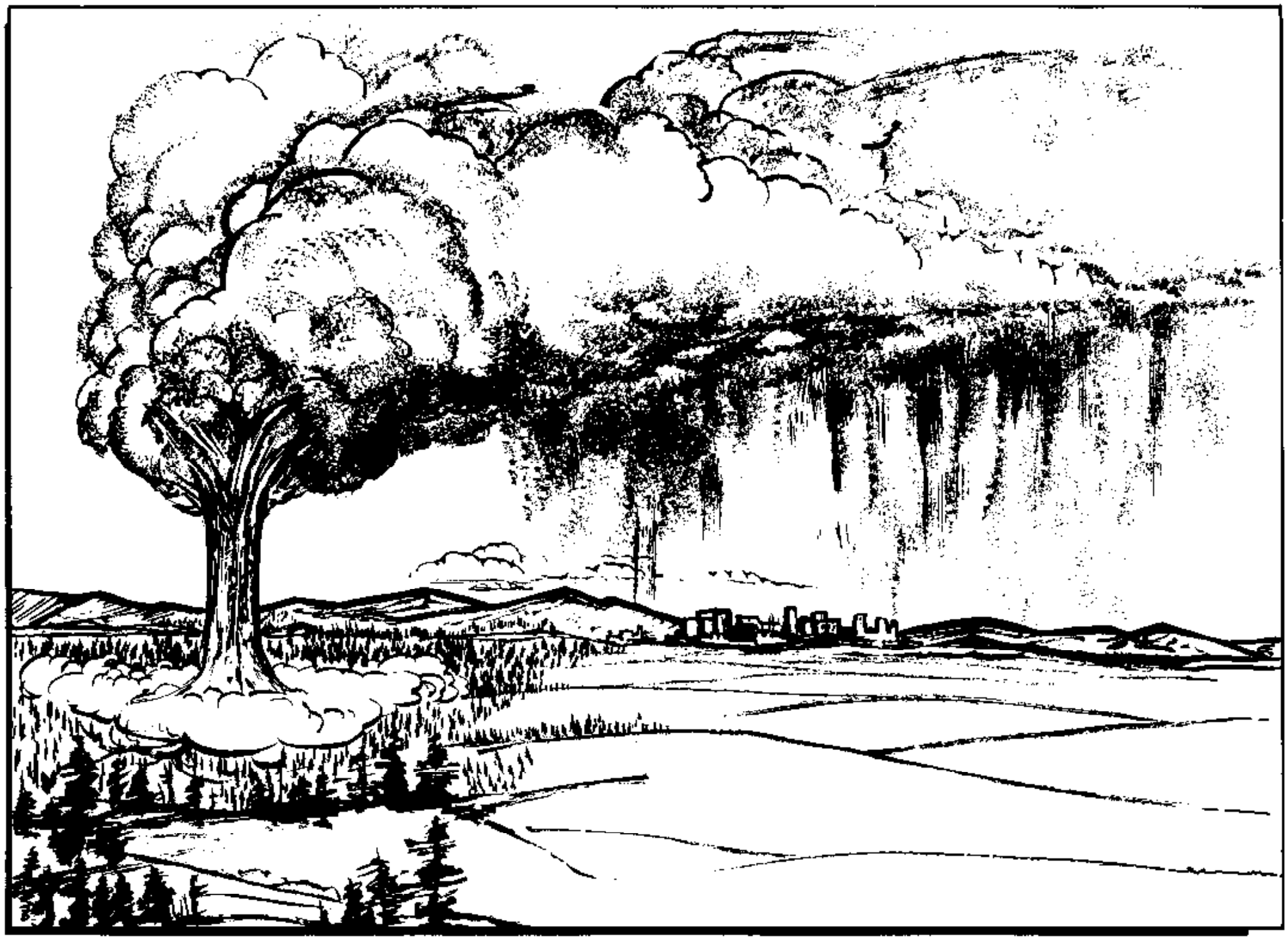

Figure 29-9. Fallout.

29-5. Types of Residual Radiation. The radioactive debris deposited on the surface as fallout contains three types of nuclear radiation—alpha particle emission, beta particle emission, and gamma radiation (figure 29-9).

a. Alpha Emitters. Alpha particles have low-pentrating abilities; therefore, survivors can easily shield themselves against these particles. Although alpha particles emitting (being given off) from radioactive elements will not penetrate the skin, alpha emitters present a serious hazard if ingested, inhaled, or allowed to enter the body by any other means. From a survival standpoint, alpha particles present the least danger of the three radiation hazards.

b. Beta Emitters. Basically, beta particles (radiation) are high-speed electrons which are only slightly penetrating; therefore, survivors can easily shield themselves against beta radiation by wearing materials of moderate thickness such as heavy shoes and clothing. Because serious skin burns may result from the direct contact of beta-emitting materials with the skin, survivors should take special care to brush themselves off, wash any previously unprotected areas of the body, or cover any exposed parts of the body. Since beta radiation is rapidly absorbed by the air, distance will provide a good form of protection; in fact, 6 to 7 feet of atmosphere will stop most of the beta radiation resulting from fallout. In addition to presenting an external hazard, beta radiation will also cause serious internal effects. By using care in decontaminating foods and water and practicing good hygiene, survivors can greatly decrease the seriousness of this hazard.

c. Gamma Radiation. In contrast to either alpha or beta emitters, gamma radiation is highly penetrating. Gamma rays are similar to light rays and X rays, but are composed of shorter wavelengths and contain greater amounts of energy. Because of their penetrating abilities, they are the most hazardous of all types of external radiation.

(1) Fortunately, for the survivor faced with the important problem of obtaining immediate protection against gamma radiation, the amount of shielding is not as great as that required for the initial gamma radiation emitted during the fireball stage of a nuclear detonation. In addition to shielding, other methods of protection against penetrating gamma radiation, later addressed in more detail, include using the factors of time and distance. Because of the low absorption coefficient of gamma rays, their relative internal hazard is greatly reduced and, in this respect, they are far less dangerous than alpha and beta radiation. However, caution should be taken to ensure no radiation is absorbed—internally or externally.

(2) Since the fallout dust or debris contains particles emitting gamma rays, survivors must be especially careful to decontaminate themselves and their shelter area. Though radioactive fallout sometimes has the appearance of white ash or dust, it usually cannot be detected by the human senses. Survivors must assume any suspicious film of dust on water or plant life is radioactive, and they should apply decontamination procedures.

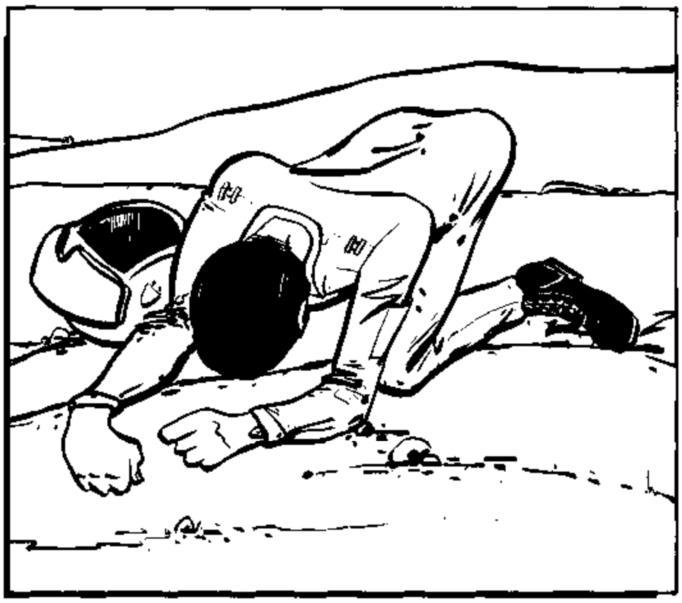

Figure 29-10. Effects.

29-6. Effects of Fallout on the Individual. The most harmful effects of fallout result from the changing of the blood cells. Because of this change in cells, some of the tissues which are essential to normal functioning of the human body are damaged or destroyed. The cells are unable to rebuild, so normal cell replacement in the organism is stopped. In addition, the formed products act as poisons to the remaining cells. The extent of damage to the body cells depends upon the dose received. Therefore, if the body receives a large dose of fallout radiation (gamma rays and possibly some beta particles), so many cells can be affected that survival is unlikely due to infection resulting from the loss of white blood cells (figure 29-10).

a. An overdose of radiation from fallout could be received if the survivor stays in the open and doesn't seek shelter. It is also possible to receive an overexposure from the fallout which settles on the clothing or body. Clothing does not stop gamma rays from penetrating and seriously injuring body tissues. Overexposure can also occur from remaining in a highly radioactive area too long. A survivor should not leave a shelter area unless absolutely necessary.

b. The first indication of an overdose of fallout radiation probably would not show up for several hours or possibly days. The survivor would then most likely become ill and begin to vomit. The time elapsing before the illness would depend on how large a dose was re-

EXPOSURE TYPE	GENERAL SYMPTOMS	ONSET OF SYMPTOMS	ABLE TO WORK	MEDICAL CARE	DEATHS	REMARKS
50-200r	Nausea, weakness, fatigue, possible vomiting, possible radiation burns	Within 24 hours	Yes	No	5%	
200-450r	Nausea, weakness, fatigue, vomiting, diarrhea, loss of hair, radiation burns, easy to bleed	3-6 hours	No	Yes	Less than 50%	Sick a few days, 1-2 weeks temporary recovery, then serious symptoms
450-600r	Same as 200-450r, plus hemorrhaging and infection	Within 3 hours	No	Yes	More than 50%	1 week temporary recovery, die without medical aid
600r+	Same as 450-600r, plus severe bloody diarrhea	Within 1 hour	No	Yes	100%	Death within 14 days
2000r+	Complete incapacitation	Within minutes	No	Yes	100%	Death within hours
Internal Radiation	Varies in proportion to internal dose	Varies	Varies	Varies	Varies	
Mild skin Injury	Itching and/or redness of skin, possible hair loss	1-2 days	Yes	No	None	
Severe Injury	Same as mild, plus radiation burns, & hair loss	1-2 days	Yes	Yes	Varies	Death depends on burn area

Figure 29-11. Radiation Sickness Symptoms.

ceived. When vomiting starts, it does not necessarily mean death will follow. For a few days, a survivor might feel far below par, but with proper medical care, complete recovery is possible (figure 29-11).

29-7. Radiation Detection. Since radiation is invisible and cannot be detected by the physical senses, detection instruments are used. This equipment includes devices for measuring the amount and intensity of area contamination and devices for determining the radiation dosage a survivor has received while in a contaminated area. A single radiation measurement usually has limited operational significance, except to personnel in the immediate area, since it gives information at the point of the reading only. However, a number of individual measurements considered together can give a picture of the radiation pattern over an entire area. A number of readings made at the same point over a period of time are required to determine the rate of decay of the fallout. Several different points for taking readings may be required in varying types of terrain.

a. Measurement of Radiation. Instruments developed for the detection and measurement of radiation are called radiac instruments. Radiac instruments measure the absorption of radiation, either in terms of dose-rate or dosage (figure 29-12).

(1) Dosage is the term applied to the total or accumulated amount of ionizing radiation (beta or gamma) received regardless of the time involved. Dosage is measured in terms of roentgens (r) or milliroentgens (mr). One milliroentgen is .001 of a roentgen. The total quantity of ionization received during a single radiation experience is called a dose. The radiation dose is also referred to as an exposure dose.

(2) When it is desirable to know how fast a certain dosage is being received, the term dose-rate is used to

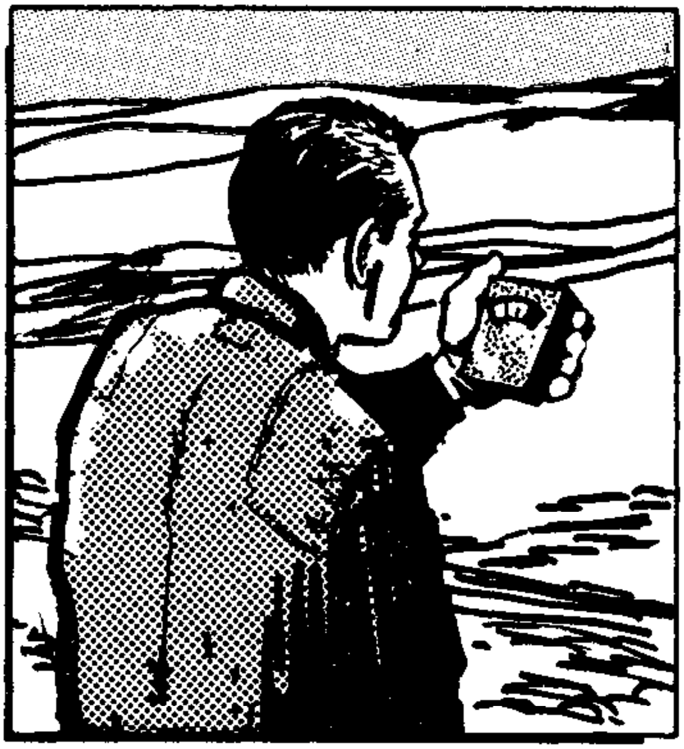

Figure 29-12. Detection.

indicate this rate. Dose-rate is the amount of radiation being received per unit of time. Generally, the dose-rate is expressed in r/hr or mr/hr. The dose-rate is also used to indicate the level of radioactivity in a contaminated area (figure 29-13).

b. Radioactive Decay. The concept of radioactive decay is of vital importance in obtaining protection against nuclear fallout and in determining survival procedures. The debris from a nuclear explosion is made up of a mixture of radioactive materials of many kinds: unfissioned particles, fission products, and numerous other radioisotopes created by the neutron activation of inert material that takes place during the explosion. Fortunately, of the nearly 200 isotopes emerging, 70 percent are short-lived materials with a half-life of less than 1 day. Some elements, however, take a much longer time to decay; these are known as long half-life elements.

(1) Half-life is the time required for a radioactive substance to lose 50 percent of its activity through decay. Half-life for a mixture of isotopes (a term used to define precise species of elements) is not as simple as that described for a single radioactive isotope. The activity of the mixture diminishes very quickly after detonation, but as time passes, the longer-lived species become responsible for the major part of the radiation remaining, so that total radioactivity diminishes much more slowly.

(2) As a rule of thumb, radioactivity may be said to decrease in intensity by a factor of 10 for every sevenfold increase in time following the peak radiation level. Figure 29-13 illustrates the rapidity of the decay of radiation from fallout during the first 2 days after the nuclear explosion which produced it. Notice that it takes about seven times as long for the dose-rate to decay from 1000 roentgens per hour (1000 r/hr) to 10 r/hr (48 hours) as to decay from 1000 r/hr to 100 r/hr (7 hours). Only in high fallout areas would the dose-rate 1 hour after the explosion be as high as 1000 roentgens per hour.

(3) If the dose-rate 1 hour after an explosion is 1000 r/hr, it would take about 2 weeks for the dose-rate to be reduced to 1 r/hr solely as a result of radioactive decay. Weathering effects will reduce the dose-rate further; for example, rain can wash fallout particles from plants and buildings to lower positions on or closer to ground. Surrounding objects would reduce the radiation dose from these low-lying particles. Figure 29-13 also illustrates the fact that at a typical location where a given amount of fallout from an explosion is deposited later than 1 hour after the explosion, the highest dose-rate and the total dose received at that location are less than at a location where the same amount of fallout is deposited 1 hour after the explosion. The longer fallout particles have been airborne before reaching the ground, the less dangerous is their radiation.

(4) Two weeks after the last burst, the occupants of most shelters could begin working outside the shelters, increasing the number of hours each day. Exceptions would be in thermal damaged areas or in areas of extremely heavy fallout such as might occur downwind from important targets attacked with many weapons, especially missile sites and very large cities. To know when to come out safely, occupants would either need a reliable fallout meter to measure the changing radiation dangers, or they would need to receive information based on measurements made nearby with a reliable instrument, using the information in figure 29-14.

29-8. Body Reactions to Radiation. The effects of ionizing radiation upon the human body may be divided into two broad categories—chronic effects and acute effects.

a. Chronic Effects. Chronic effects are defined as those occurring some years or generations after exposure to radiation. Included in this category are the cancer-producing and genetic effects. While of concern because of their possible damage to future generations, these effects are of minor significance insofar as they may affect immediate survival.

b. Acute Effects. Acute effects are of primary significance in survival. Some acute effects appear within a few hours after exposure to radiation. These effects are the result of direct physical damage to tissue caused by

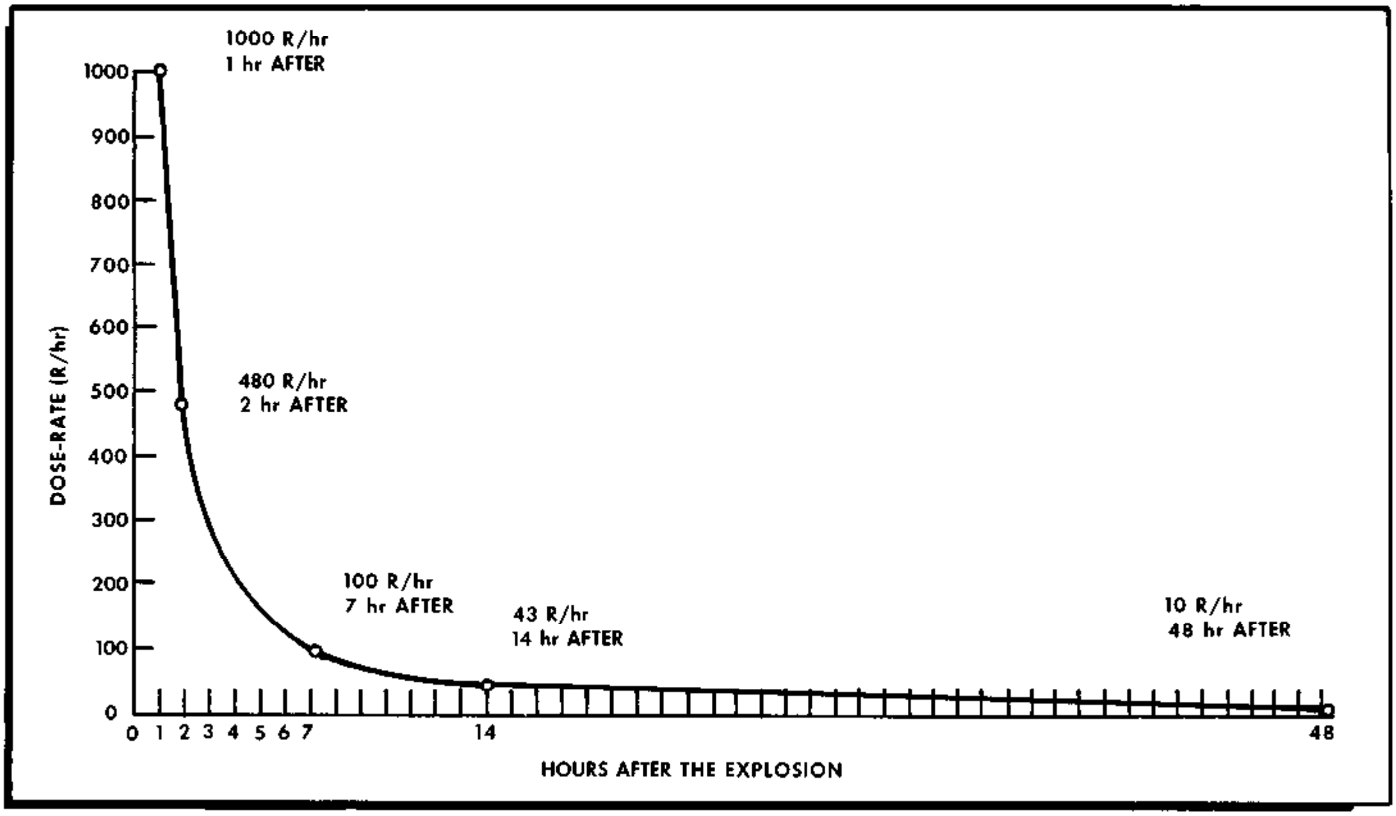

Figure 29-13. Decay Ratio.

radiation exposure; beta burns are examples of acute effects.

c. Body Damage. The extent of body damage depends to a large degree on the part and extent of the body exposed and the ability of the body to recover from radiation damage. Certain parts of the body, such as the brain and the kidneys, have little ability to recover from injury, while other portions of the body, such as the skin and bone marrow, can repair damage. The extent of body exposure to radiation is of great importance in determining the chances of subsequent recovery. If a dose of 350 roentgens was applied to just a small portion of the body, such as the hands or face, there would probably be little effect on overall health. Serious damage would, of course, be created in these exposed parts.

d. Hazards from Residual Radiation. There are two main hazards resulting from residue hazard, resulting from highly penetrating gamma radiation and less penetrating beta radiation which causes burns; and an internal hazard, resulting from the entry of alpha- and beta-emitting particles into the body.

(1) The external hazards result in overall irradiation and serious beta burns, while the internal hazards result in irradiation of the critical organs such as the gastrointestinal tract, thyroid gland, and bone.

(2) A very small amount of radioactive material can cause extensive damage in these and other parts of the body if allowed to enter the body by consumption of contaminated food or water or by absorption by the bloodstream through cuts or abrasions in the skin. By comparison, the material gaining entry by breathing presents a lesser hazard. Fortunately, most of the fallout particles are filtered out by the upper respiratory tract as seen in the observations of victims on the Marshall Islands who were accidentally exposed to fallout during the 1954 nuclear test. By using good hygiene and careful decontamination of food and water, a survivor can further reduce the internal hazard.

e. Degree of Radiation Damage. Because of the sensitivity of the gastrointestinal tract to radiation, the severity and the speed with which vomiting and diarrhea appear after exposure are good indicators of the degree of radiation damage. Almost everyone confined to immediate action shelters after a nuclear attack would be under stress and without clean surroundings. Many would also lack adequate water and food. Under these conditions, perhaps half the survivors who receive a whole-body dose of 200-450 roentgens would die within a few days. The human body can repair most radiation damage if the daily radiation doses are not too large.

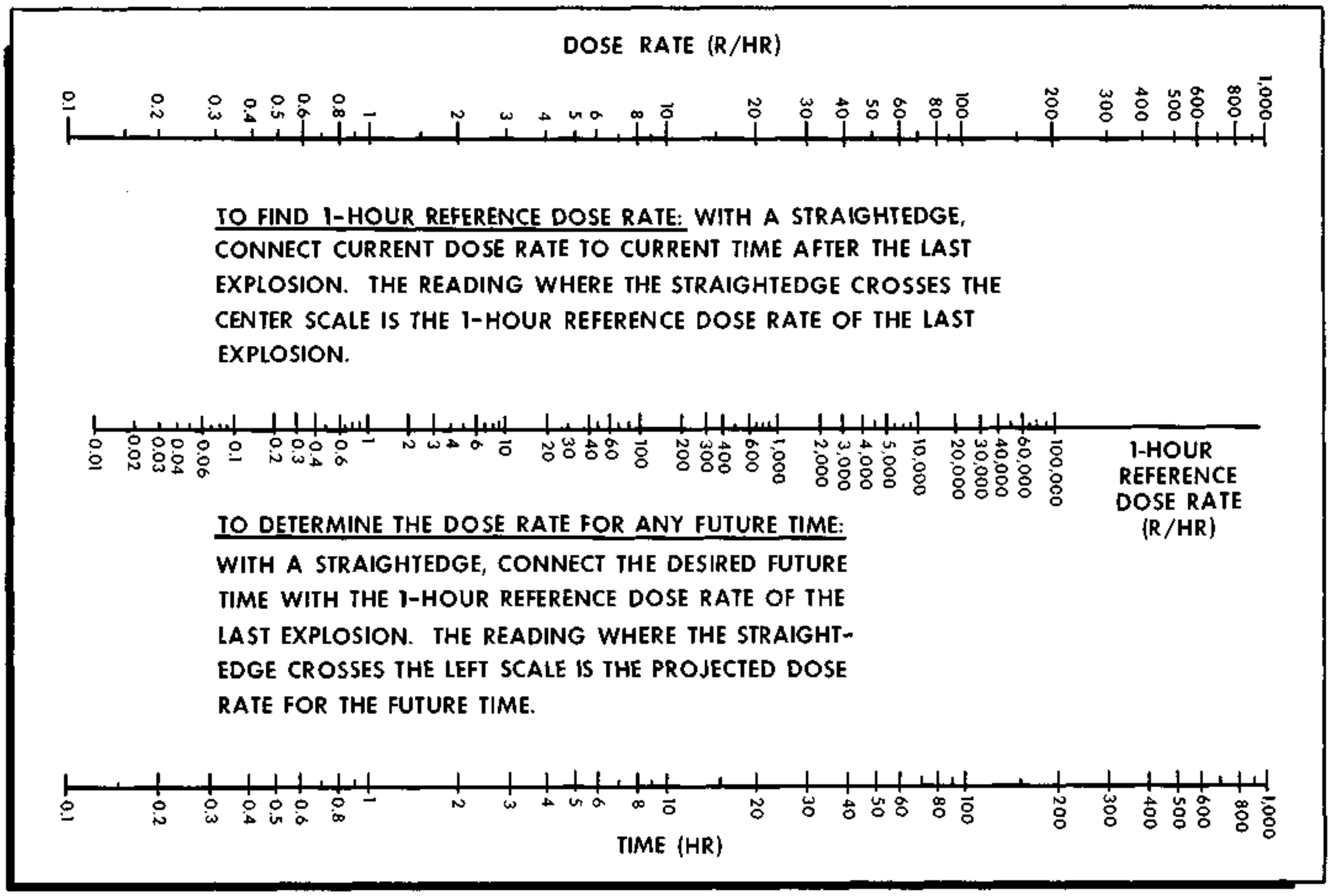
DOSE RATE (R/HR)
0.1 0.2 0.3 0.4 0.5 0.6 0.8 1 2 3 4 5 6 8 10 20 30 40 50 60 80 100 200 300 400 500 600 800 1,000
0.01 0.02 0.03 0.04 0.06 0.1 0.2 0.3 0.4 0.6 1 2 3 4 6 10 20 30 40 60 100 200 300 400 600 1,000 2,000 3,000 4,000 5,000 10,000 20,000 30,000 40,000 60,000 100,000
1-HOUR REFERENCE DOSE RATE (R/HR)
0.1 0.2 0.3 0.4 0.5 0.6 0.8 1 2 3 4 5 6 8 10 20 30 40 50 60 80 100 200 300 400 500 600 800 1,000
TIME (HR)

TO FIND 1-HOUR REFERENCE DOSE RATE: WITH A STRAIGHTEDGE, CONNECT CURRENT DOSE RATE TO CURRENT TIME AFTER THE LAST EXPLOSION. THE READING WHERE THE STRAIGHTEDGE CROSSES THE CENTER SCALE IS THE 1-HOUR REFERENCE DOSE RATE OF THE LAST EXPLOSION.

TO DETERMINE THE DOSE RATE FOR ANY FUTURE TIME: WITH A STRAIGHTEDGE, CONNECT THE DESIRED FUTURE TIME WITH THE 1-HOUR REFERENCE DOSE RATE OF THE LAST EXPLOSION. THE READING WHERE THE STRAIGHTEDGE CROSSES THE LEFT SCALE IS THE PROJECTED DOSE RATE FOR THE FUTURE TIME.

Figure 29-14. Dose Rate.

29-9. Countermeasures Against Penetrating External Radiation:

a. The means of protection from external radiation are threefold—time, distance, and shielding. By controlling the length of exposure time, by controlling the distance between the individual and the source of radiation, and most important of all, by placing some absorbing or shielding material between the survivor and the source of radiation, the survivor's level of radiation can be significantly reduced and thereby increase the chances of survival.

(1) Time. The effect of time on radiation exposure is easy to understand. Take a simple example such as the following: Assume a person is in an area where the radiation level from penetrating external radiation is 100 roentgens per hour and the dose-rate is constant; then in 1 hour, that person would receive 100 roentgens (or more properly, roentgen absorbed dose (rad)); in 2 hours, 200 rad; and in 8 hours, 800 rad. The important implication for a survivor is that, since radiation dosages received should be considered to be essentially cumulative, as little time as possible should be spent in an unprotected area, whether constructing a shelter or searching for water. Time is also of vital importance from another standpoint, that of radioactive decay. Knowledge of this characteristic can serve as one of the primary means of protection against radiation from fallout. The importance of time as a protective factor may be seen clearly by the following example. If survivors were to enter a high-intensity fallout area of 1000 roentgens per hour in which the radiation intensity had just peaked and remain there for 1 hour in the open, they would accumulate a biologically damaging dose of approximately 650 rad. This is a dose strong enough to kill 9 out of 10 people exposed to it. If, however, they entered this same area some 48 hours later and remained there for 1 hour, they would receive a dose of approximately 10 rad. Even if they were then required to remain in the open for 24 hours, their dose would only be 170 rad. Two weeks after the completed fallout, they would receive only one rad per hour in this same area. They could then remain in the area for several months, accumulating a dose of 180 rad, and not develop severe radiation sickness because of the rate at which the dosage was received (figure 29-15).

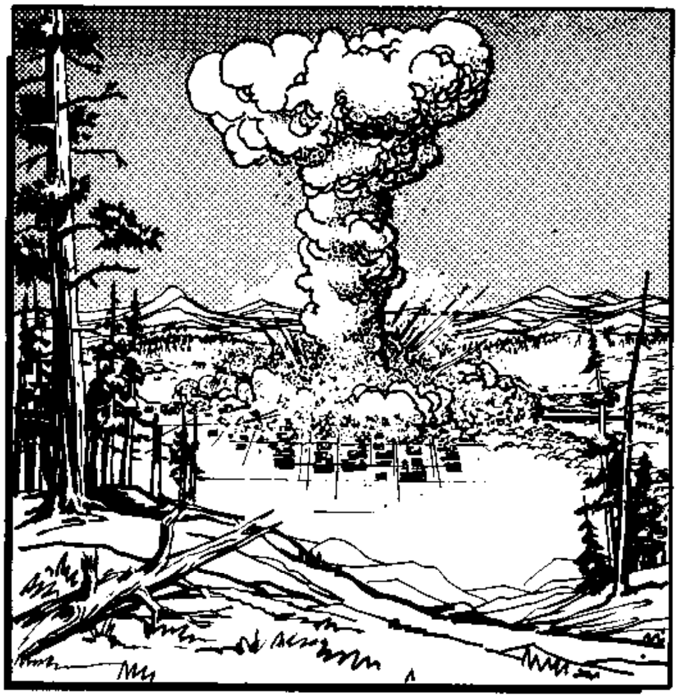

Figure 29-15. Countermeasures.

(2) Distance. Distance is effective as a means of protection against penetrating gamma radiation as shown by the inverse square law. This law states the intensity of radiation decreases by the square of the distance from the source. In other words, survivors exposed to 1000 units of penetrating external radiation at 1 foot from the source would receive only 250 units at 2 feet; when they double the distance, the radiation level is reduced by $(\frac{1}{2})^2$ or one-fourth the amount. When the distance is tripled, the dose is reduced to $(\frac{5}{8})^2$ or one-ninth of the original amount (or 111 units). At 10 feet, this is further reduced to $({}^{1}10)^2$ or one-hundreth of the radiation exposure at 1 foot. This relationship of distance to intensity of radiation exposure is dependent upon the distribution pattern of the radiation source. As just seen, if the radiation is concentrated into a very small area (referred to as a point source), the intensity is decreased to about one-fourth the original amount each time the distance is doubled. A complicated relationship, however, is obtained when the radiation source is not concentrated in a point, but is spread around in random patterns as in the case of surface contamination from fallout.

(3) Shielding. The third and most important method of protection against penetrating radiation is that of shielding. Since the damaging effects of penetrating radiation arise from the fact that the rays interact with electrons in the body, survivors must place dense material between themselves and the source of radiation. The more electrons there are in the makeup of the shielding material, the more gamma radiation will be stopped from entering the body. Lead, iron, concrete, and water are all examples of good shielding materials. Of the three described countermeasures against penetrating external radiation, shielding gives the greatest protection, is the easiest to use under survival conditions, and is, therefore, the most desirable. If shielding is not possible, however, the other precautions should be rigorously followed. The degree of protection they afford is significant and could provide the necessary margin of safety for survival (figure 29-16).

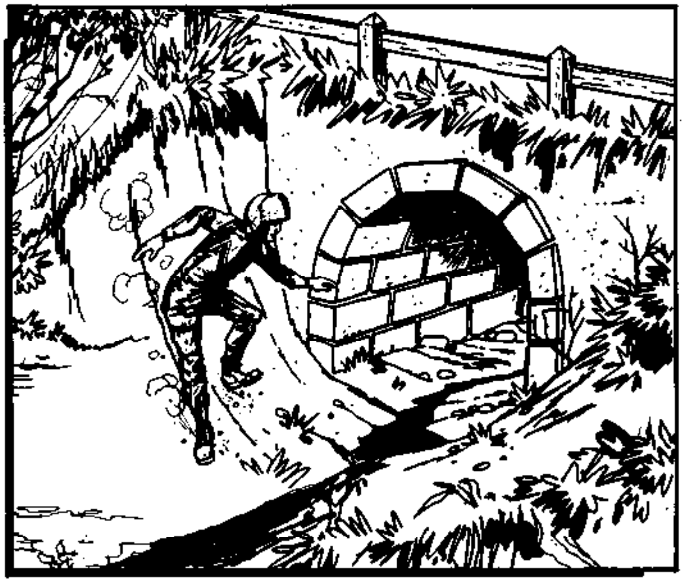

Figure 29-16. Shielding.

b. It is a common misconception that radioactive fallout can impart its radioactivity to the object with which it comes in contact, such as fruits and vegetables. The rays from radioactive fallout particles cannot make something radioactive merely by being in contact with it. For this reason, canned foods and smooth-skinned fruits and vegetables may be eaten once they have been decontaminated.

(1) Everyone entering a contaminated area should wear protective clothing (CW ensemble) to prevent exposure to radioactive dust. The main reason for this precaution is to shield the skin from beta particles. Beta particles won't penetrate far into the body, but they will harm the skin. The effect, a reddening and blistering of the skin, is called a "beta burn." Such damage may not appear until sometime after exposure since beta particles have a delayed action. Survivors may not know they have received skin burns from beta particles until it is too late. Therefore, protective clothing should be worn as instructed. Sleeves should be down and buttoned, wrist and ankle openings taped, and gloves and boots worn, if possible.

(2) Alpha particles are lowest in ability to penetrate, but they are most hazardous when taken into the body through contaminated food and drink. The best defense against alpha particles is to ensure all food is properly decontaminated before consumption.

(3) If at all possible, a survivor should always avoid "hot" areas. Sometimes, a survivor cannot avoid contact with this environment and, when entering a "hot" area, should keep radioactive dust off the body and out of the body's system. If survivors inhale radioactive dust, they can suffer serious internal injury. If there is radioactive dust in "hot" areas, a protective filter should be used over the mouth and nose.

c. The presence of fallout material in the area will require some minor modification of medical procedures. Wounds should be covered to prevent entry of active particles. Burns from beta activity are treated like any other burn, except the burned surface should be washed. Measures to prevent infection should be emphasized since the body will be especially sensitive to infections because of changes in the blood. Extra attention should be given to the prevention of colds and respiratory infections. The eyes should be covered to prevent entry of particles. Improvised goggles may be used for this purpose.

29-10. Shelter. A sufficient thickness of shielding material will reduce the level of radiation to negligible proportions. The thickness required to reduce gamma radiation from fallout is much less than that necessary against initial gamma radiation since fallout radiation has much less energy than the initial radiation from a nuclear explosion. Thus, a comparatively small amount of shielding material can provide good protection against residual gamma radiation.

a. The following table illustrates the thicknesses of various materials required to reduce the penetration from residual fallout gamma sources by 50 percent:

Iron and steel	.7 inches
Concrete	2.2 inches
Brick	2.0 inches
Dirt	3.3 inches
Ice	6.8 inches
Wood (soft)	8.8 inches
Snow	20.3 inches

b. The principle of half-value layer thickness is useful in understanding the absorption of gamma radiation by various materials. According to this principle, if 2 inches of brick reduce the level of gamma radiation by one-half, adding another 2 inches of brick will reduce the intensity by another half, namely to one-fourth the original amount; 6 inches will reduce the level of fallout gamma radiation to one-eighth its original amount; 8 inches to one-sixteenth; and so on. Thus in a shelter protected by 3 feet of earth, a radiation intensity of 1000 roentgens per hour on the outside would be reduced to about one-half roentgen per hour inside the shelter.

c. Areas where terrain provides natural shielding and easy shelter construction, such as the sides of ditches, ravines, rocky outcroppings, hills, and riverbanks, are ideal locations for an emergency shelter (figure 29-17). In level areas lacking natural protection, dig a foxhole or slit trench.

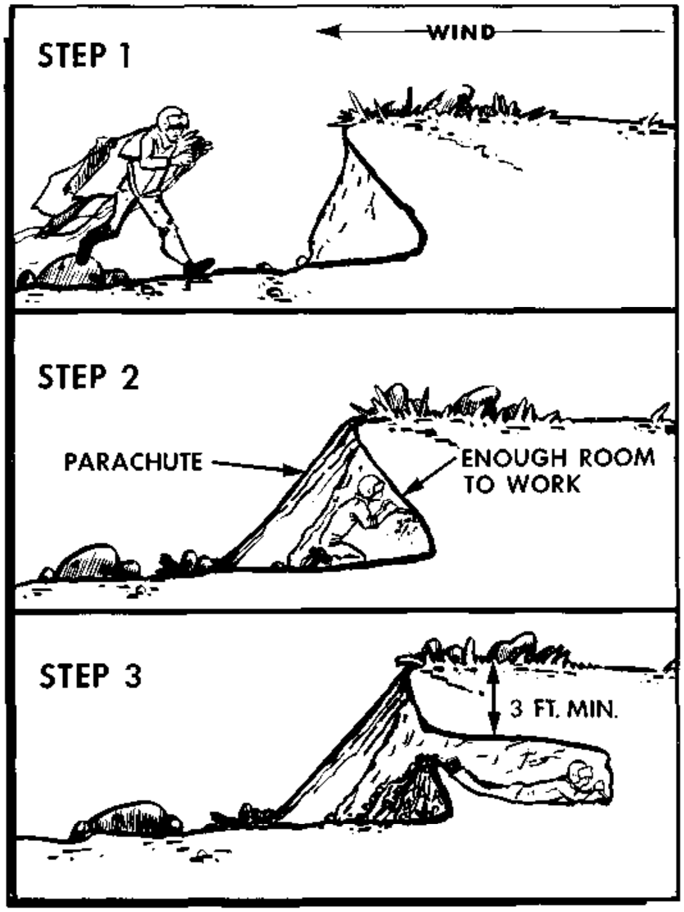

Figure 29-17. Immediate Shelter.

(1) In digging a trench, the survivor should work from the inside of the hole as soon as it is large enough to cover part of the body to prevent exposure of the entire body to radiation. In open country, an attempt should be made to dig the trench from a prone position, stacking the dirt carefully and evenly around the trench. On level ground, the dirt should be around the body for additional shielding. Depending upon soil conditions, the time to construct a shelter will vary from a few minutes to a few hours. Rapid shelter construction will limit the dosage received. Building a shelter under a tree is not recommended because cutting or digging out the numerous roots will be very difficult. Another disadvantage in making a shelter under trees is that more of the gamma rays from fallout particles on the leaves and branches would reach and penetrate the shelter than if

these same particles were on the ground. Many gamma rays from fallout particles on the ground would be scattered or absorbed by striking rocks, clods of earth, tree trunks, or buildings before reaching a below-ground shelter (figure 29-18).

(2) While an underground shelter covered by 3 or more feet of earth would provide the best protection against fallout radiation, the following additional unoccupied structures (in the order listed) offer the next best protection:

(a) Caves and tunnels covered by more than 3 feet of earth.

(b) Storm or storage cellars.

(c) Culverts.

(d) Basements or cellars of abandoned buildings.

(e) Abandoned buildings made of stone or mud.

d. Building a roof on the shelter should not be required. A roof should only be added if the materials are readily available to the survivor and will require only a brief exposure to the outside contamination. If the construction of a roof would require extended exposure to penetrating radiation, it would be wiser to leave the shelter roofless. The function of a roof is to lend distance from the source of fallout to the body, and unless dense roofing is used, a roof provides only scant shielding. A simple roof can be constructed out of a piece of parachute anchored down by dirt, rocks, or other refuse from the shelter. Large particles of dirt and other debris may be removed from the top of this parachute canopy by beating it at frequent intervals. Such a parachute cover will not offer shielding from the radiation emitted by active fission products deposited on the outer surface, but it will increase the distance from the fallout source and keep the shelter area covered from further contamination.

e. The primary criterion for locating and establishing a shelter site should be to obtain protection as rapidly as possible against the high intensity radiation levels of early gamma fallout. Five minutes to locate the shelter site is an excellent guide. Speed in obtaining shelter is essential. Without shelter, the dosage received in the first few hours will exceed that received during the rest of the week in a contaminated area; the dosage received in this first week will exceed that accumulated during the rest of a lifetime spent in the same contaminated area.

(1) Several initial actions should be kept in mind while seeking a shelter location. The survivor should:

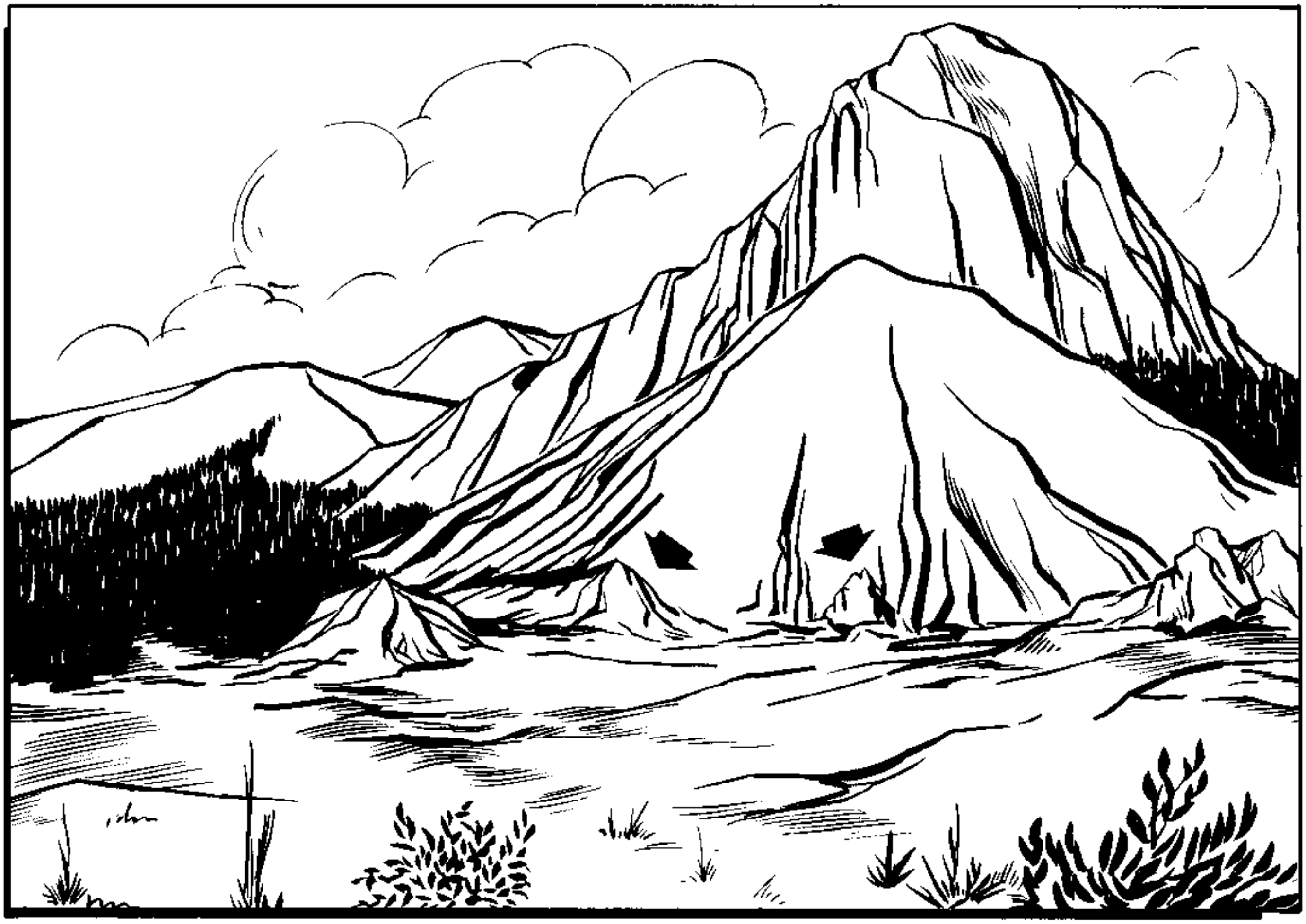

Figure 29-18. Terrain That Provides Natural Shielding.

(a) Where possible, attempt to remain with the aircraft until it is possible to eject or land in an area at least 20 miles upwind or crosswind from any known target. This action will ensure the landing area will have a minimum amount of fallout (figure 29-19).

(b) If it can be controlled, parachute opening should be delayed until a comparatively low altitude is reached in order to reduce the time of exposure to fallout during the descent.

-1. During the descent, select areas on the ground likely to provide good shelter.

-2. Immediately upon landing, take the parachute and survival kit and find cover and protective shelter.

-3. Take precautions to avoid detection and capture, but not at the expense of additional exposure to radiation which will lessen the chance of survival.

(2) In selecting and establishing the shelter, the survivor should keep the following additional factors in mind in order to reduce the time of exposure and the dosage received:

(a) Where possible, seek a crude existing shelter which can be improved. If none is available, dig a trench.

(b) Dig the shelter deep enough to obtain good protection, then enlarge it as required for comfort.

(c) Cover the top of the foxhole or trench with any readily available material and a thick layer of earth, if possible, without leaving the shelter.

(d) During the descent and while constructing a shelter, keep all clothing on, as well as a cap, scarf, and gloves to obtain protection against burns from beta radiation.

(3) The shelter location should be brushed or scraped clean of any surface deposit with a branch or some other object to be certain that contaminated materials are removed from the area to be occupied. The swept area should extend at least 5 feet beyond the area where the shelter is being dug. Any material brought into the shelter should be decontaminated. This includes grass or foliage that is being used as insulation or bedding material, outer clothing (especially footgear), and the parachute if it is to be used. If weather permits and the parachute and outer clothing are heavily contaminated, the survivor may want to remove them and bury them under a foot of earth at the end of the shelter. These may later (after decay factor) be retrieved when leaving the shelter. If these materials are dry, decontamination may be done by beating or shaking them outside the entrance to the shelter to remove the radioactive dust (figure 29-20). Any body of water, even though it may contain contaminated particles, may be used to rid materials of excess fallout particles by simply dipping

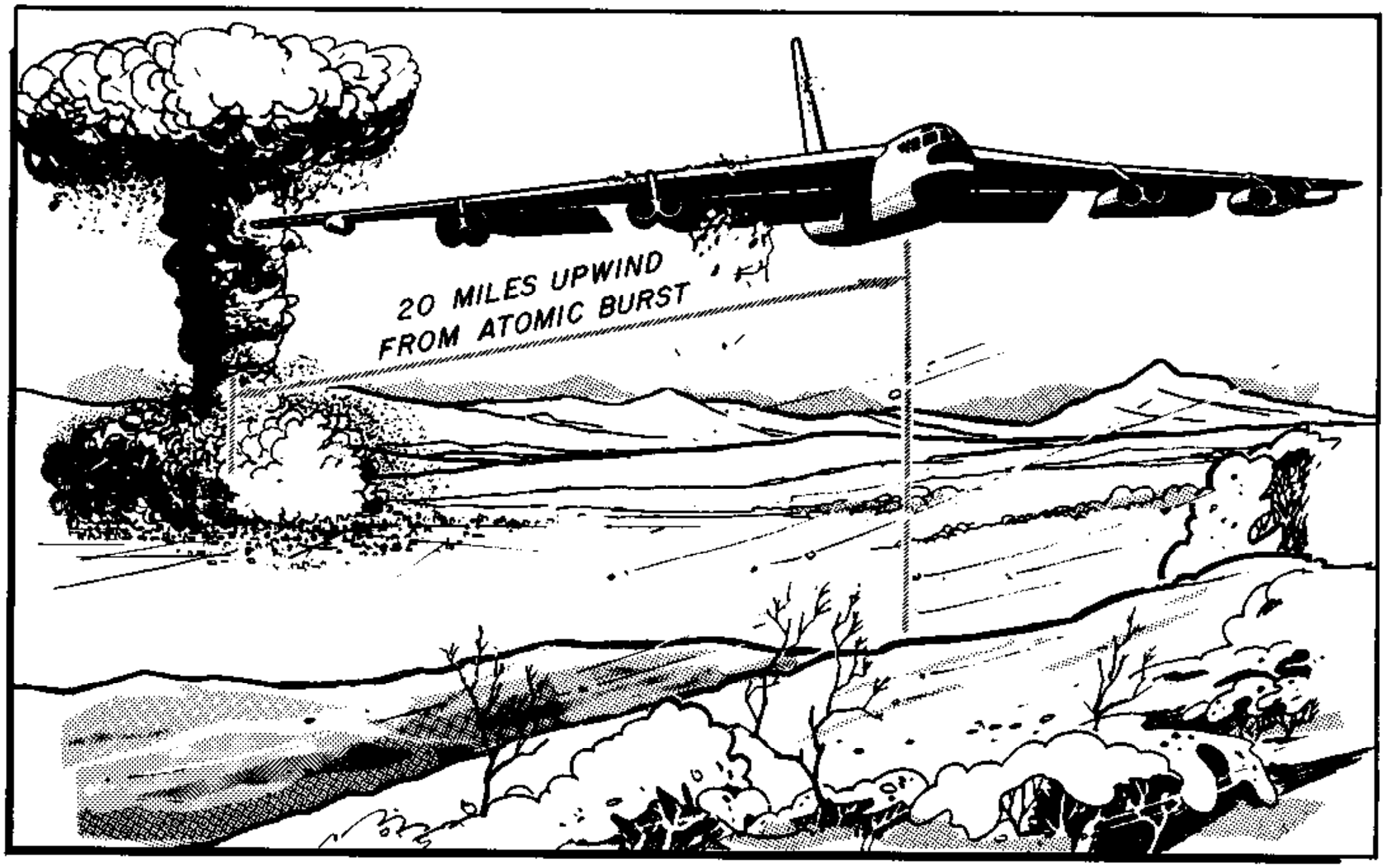
20 MILES UPWIND FROM ATOMIC BURST

Figure 29-19. Route of Travel.

the materials into the water and shaking them to remove excess water. Do not wring out materials since this will trap the contaminated particles. If at all possible, wash the body thoroughly with soap and water without leaving the shelter. This washing will remove most of the harmful radioactive particles which are likely to cause beta burns or other damage. If water is not available, the face and any other exposed skin surfaces should be wiped to remove contaminated dust and dirt. This may be done with a clean piece of cloth or a handful of uncontaminated dirt obtained by scraping off the top few inches of soil and using the "clean" dirt.

f. Upon completion of the shelter, the survivor should lie down, keep warm, sleep, and rest as much as possible during the time spent in the shelter. There is no need to panic if nausea and symptoms of radiation sickness are experienced. Even small doses of radiation can cause these symptoms which may disappear in a short time. The following provides the time necessary to avoid serious dosage and still allow the opportunity to cope with survival problems:

(1) Complete isolation should be maintained from 4 to 6 days following delivery of the last weapon. A very brief exposure for procurement of water on the third day is permissible, but exposure should not exceed 30 minutes.

(2) On day 7, one exposure of not more than 30 minutes.

(3) On day 8, one exposure of not more than 1 hour.

(4) From day 9 through day 12, exposure of 2 to 4 hours per day.

(5) From day 13 on, normal operation, followed by rest in a protected shelter.

29-11. Water. In a fallout contaminated area, available water sources may be contaminated with radioactive particles. If possible, the survivor should wait at least 48 hours before drinking any water to allow radioactive decay to take place. Selecting the safest possible source of water will greatly diminish the danger of ingesting harmful amounts of radioactivity.

a. Although many factors, such as direction of wind, rainfall, and amount of particulate matter (clay, for example) in the water, will influence the choice in selecting water, the following priorities of water sources are

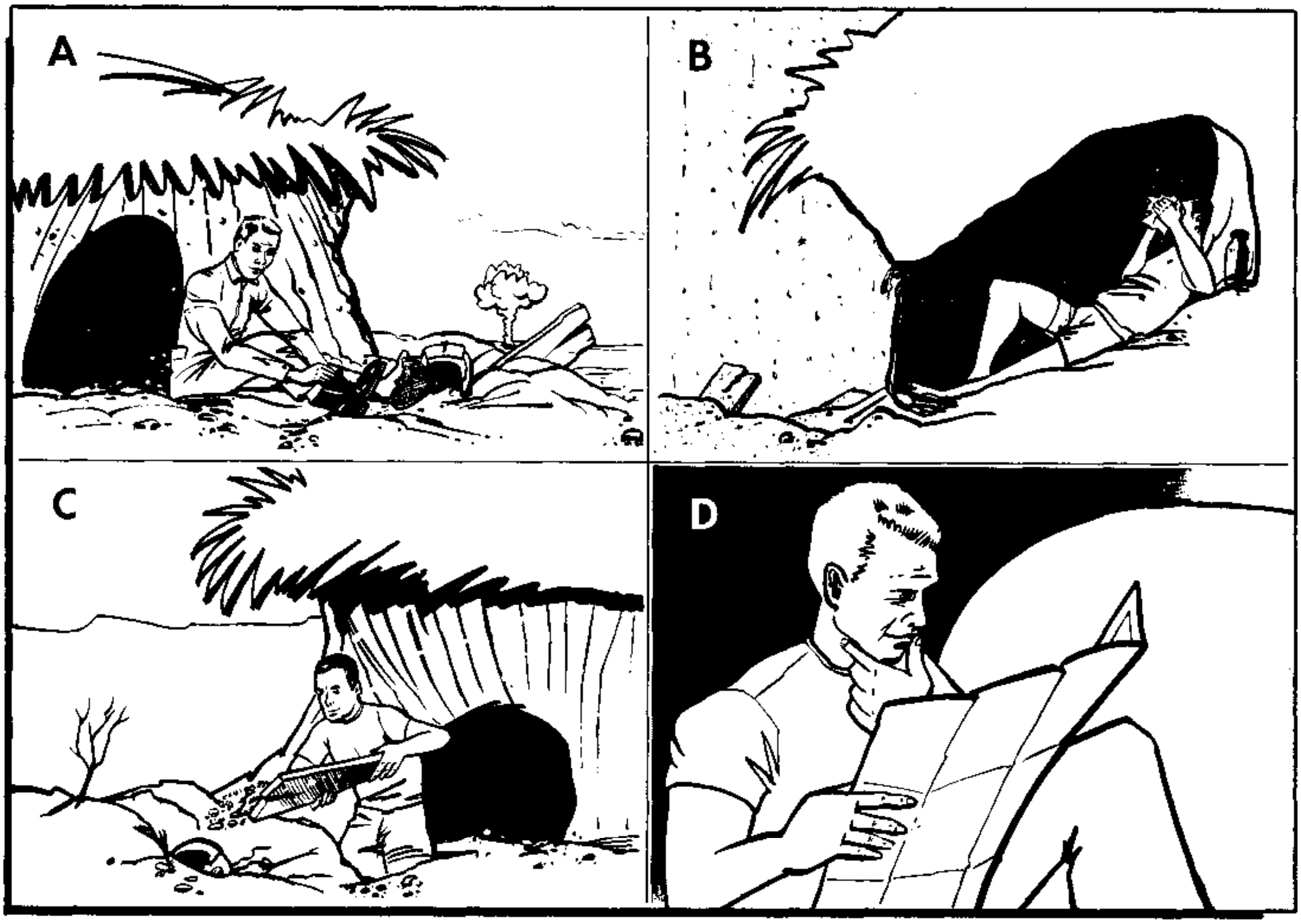

Figure 29-20. Decontamination.

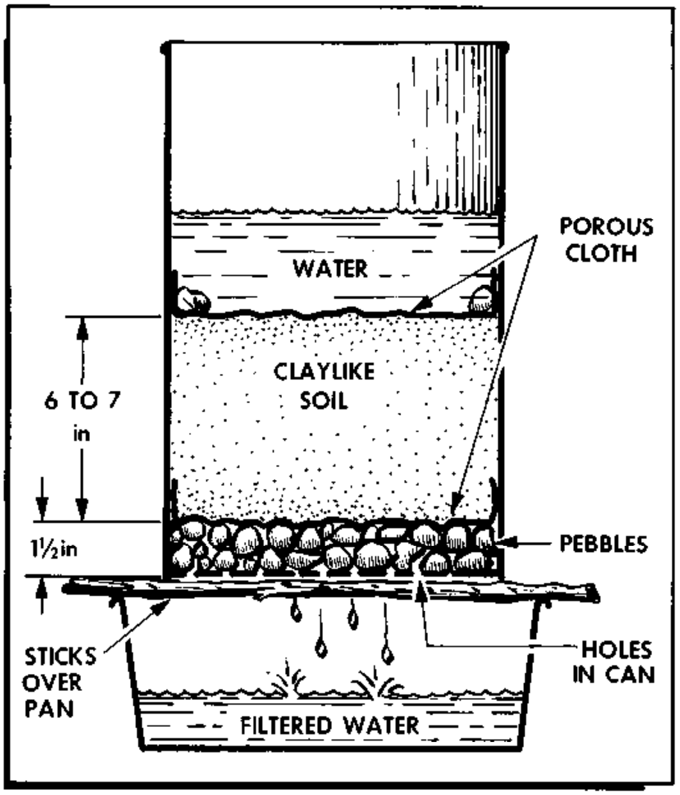

Figure 29-21. Filtering Water.

recommended: (As an additional precaution against disease, all water sources should be treated with water purification tablets from the survival kit or boiled for at least 10 minutes.)

(1) Water from springs, wells, or other underground sources having natural filtration will be the safest sources of water.

(2) Any water in the pipes or containers of abandoned houses or stores will also be free from radioactive particles and, therefore, safe to drink, although precautions will have to be taken against bacteria in the water.

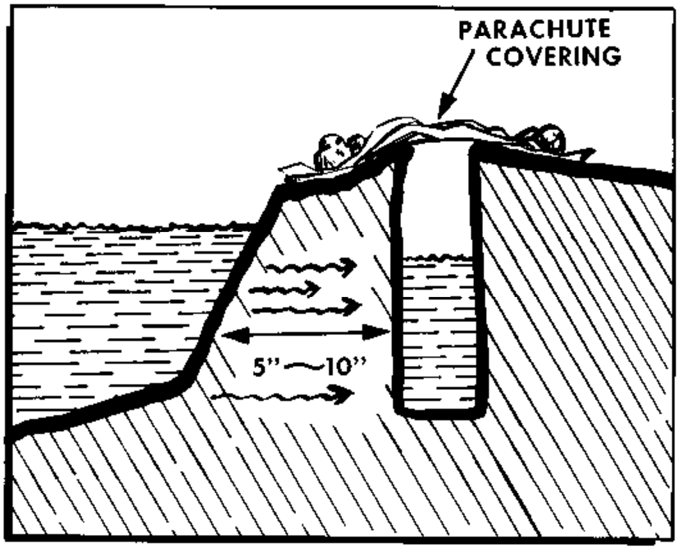

Figure 29-22. Settling Water.

(3) Snow taken from a level which was 6 or more inches below the surface during the fallout should be a safe source of water.

(4) Water from streams and rivers will be comparatively free from fallout within several days after the last nuclear explosion because of the dilution factor. If at all possible, such water should be filtered before drinking.

(5) Water from lakes, pools, ponds, and other standing sources is likely to be heavily contaminated, though most of the heavier, insoluble isotopes will settle to the bottom.

b. The degree of solubility of various isotopes varies. Some fission products are extremely water soluble, but most are relatively insoluble. Certain isotopes, in fact, have been found to be as much as 90 percent insoluble. The significance of this fact is that 99 percent of the radioactivity in water could be removed by filtering it through ordinary earth (figure 29-21). The best method of filtration is to dig sediment holes or seepage basins along the side of a water source. The water will seep laterally into the hole through the intervening soil which will act as a filtering agent and remove the contaminated fallout particles which have settled on the original body of water. It is important the hole be covered in some way (example, with a parachute) to prevent further contamination (figure 29-22).

c. Settling is one of the easiest methods used to remove most fallout particles from water. Furthermore, if the water to be used is muddy or murky, settling it before filtering will extend the life of the filter. The procedure is as follows:

(1) Fill a bucket or other deep container three-fourths full of the contaminated water.

(2) Dig dirt from a depth of 4 or more inches below the ground surface and stir it into the water. Use about a 1-inch depth of dirt for every 4 inches of water.

(3) The water is stirred until practically all of the dirt particles are suspended in the water.

(4) This mixture should be allowed to settle for at least 6 hours. The settling dirt particles will carry most of the suspended fallout particles to the bottom and cover them. The clear water can then be dipped or siphoned out and purified (figure 29-23).

29-12. Food. Obtaining edible food in a radiation contaminated area is possible. Survivors need to follow only a few special procedures in selecting and preparing rations and native foods for use. Since survival rations are protected by secure packaging, they will be perfectly acceptable for use after the ration containers are decontaminated, but survivors should supplement them with any food they can find on their trips away from the shelter. Any processed foods which may be found in abandoned buildings are acceptable for use once they are decontaminated. These include canned and pack-

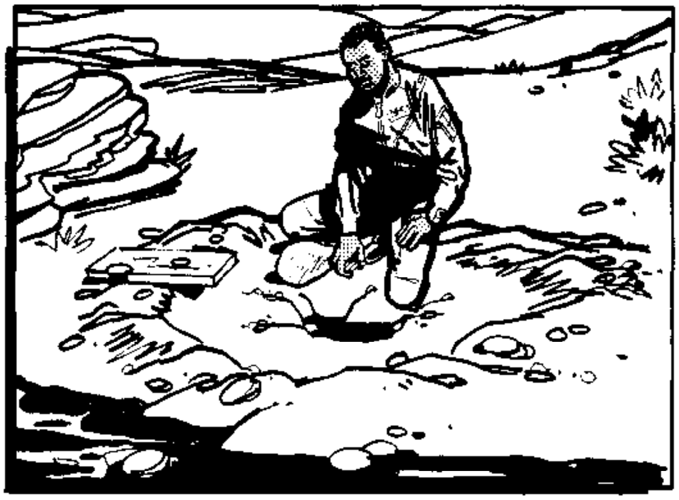

Figure 29-23. Settling Pool.

aged foods after the containers or wrappers are discarded or washed free of fallout particles, food stored in any closed container, and food stored in protected areas (such as cellars). The containers should be washed before opening. For purposes of discussion, native food sources may be divided into two categories: plant food and animal food.

a. Animal Food. In the category of native animal food, survivors must assume all animals, regardless of their habitat or living conditions, will be exposed to radiation. Since the effects of radiation on animals are similar to those in humans, most of the wild animals living in a fallout area are likely to become sick or die from radiation during the first month following the nuclear explosion. Even though animals may not be completely free from radioactive materials, it may be necessary to use them as a food source. With careful preparation and adherence to several important principles, animals can be safe sources of food.

(1) If an animal appears to be sick, it should not be eaten. The animal may have developed a bacterial infection as a result of a radiation dose. Contaminated meat could cause severe illness or death if eaten, even if thoroughly cooked.

(2) All animals should be carefully skinned to prevent any radioactive particles adhering to the outside of the skin from gaining entry into the body.

(3) Meat around the bones and joints should not be eaten. A large percentage of radioactivity in the body of animals is found in the skeleton. The remaining muscle tissue of the animal, however, will be safe to eat. Before cooking the animal, survivors should cut the meat away from the bone, leaving approximately one-eighth of a inch of meat on the bone. All internal organs such as the heart, liver, and kidneys, normally used as survival food, should be discarded (figure 29-24) since they tend to concentrate beta and gamma radioactivity. All meat should be cooked until it is very well done. To be sure the meat is well done, it should be cut into pieces less

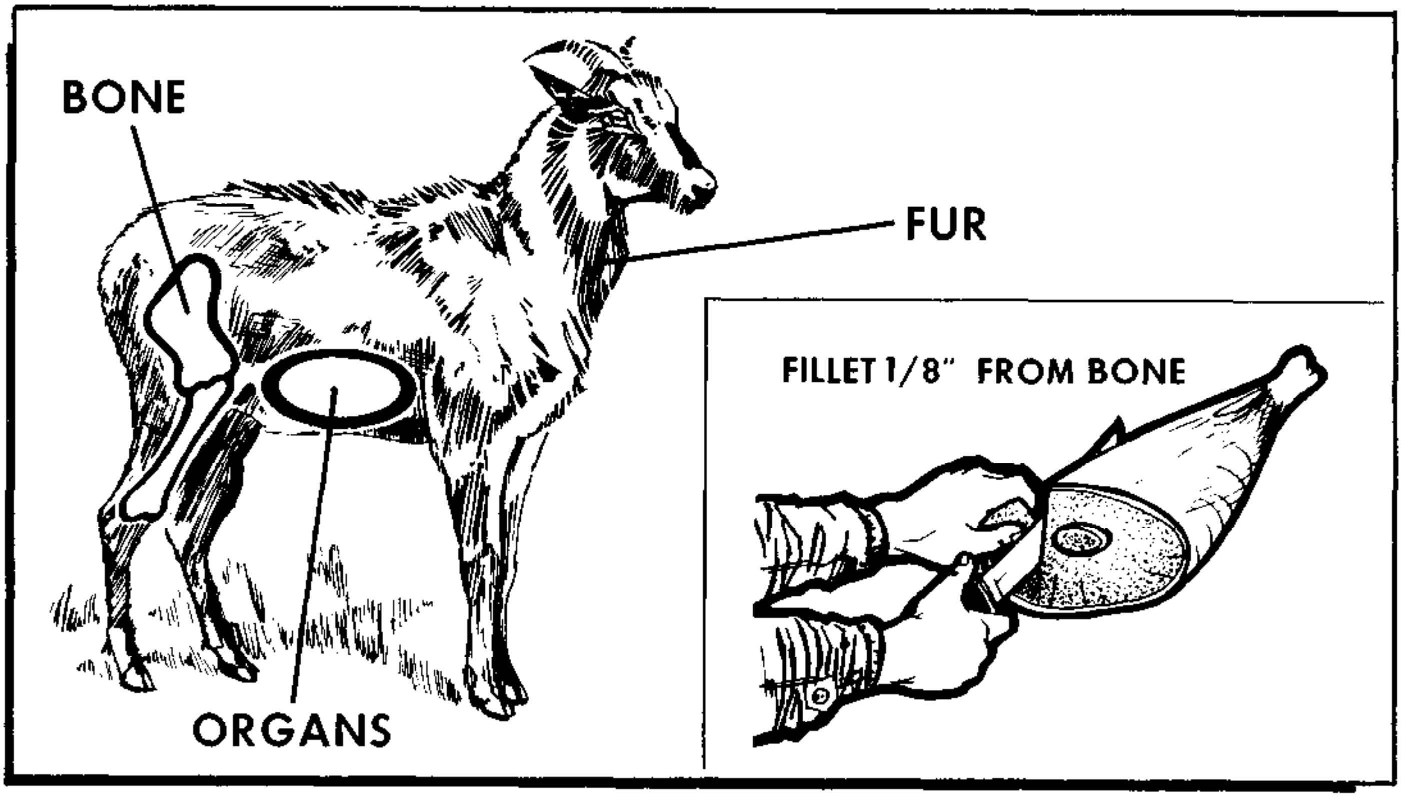

Figure 29-24. Meat Procurement.

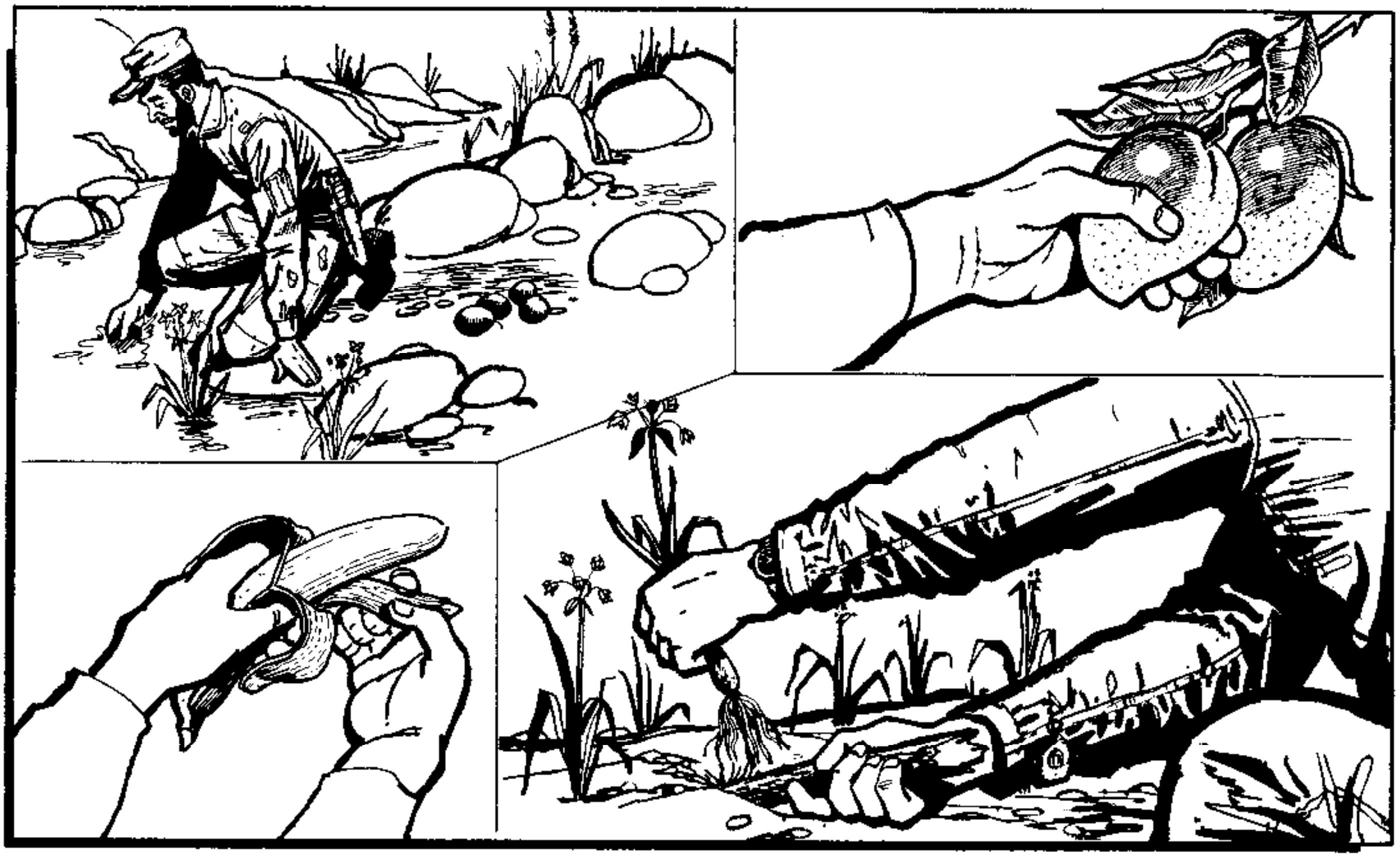

Figure 29-25. Collecting Food.

than one-half inch thick before cooking. This precaution also reduces cooking time and saves fuel.

(4) The extent of contamination of fish and clams following nuclear tests in the Pacific was found to be much greater than that of the land animals. On the basis of these findings and those of other tests showing the high concentration of radioactivity in aquatic plants and animals, especially near coastal areas, it is recommended that aquatic food sources be used only in an extreme emergency.

(5) All eggs (excluding the shell), even if laid during the period of fallout, will be safe to eat. Because animals absorb large amounts of radioactive strontium from the plants upon which they graze, milk from any animals in a fallout area should be avoided.

b. Plant Food. Plant foods are contaminated by the accumulation of fallout on their outer surfaces or by means of absorption through the roots.

(1) The survivor's first choice of plant food should be vegetables, such as potatoes, turnips, carrots, and other plants, whose edible portions grow underground. These are the safest to eat once they are scrubbed and the skins removed. Second in order of preference are those above-ground portions of the plant with edible parts which can be decontaminated by washing and peeling their outer surfaces; examples are bananas, apples, and other such fruits and vegetables. Other smooth-skinned vegetables, fruits, or above-ground plants, which cannot be easily peeled or effectively decontaminated by washing the radioactive particles off their surfaces, will be the third choice of emergency food (figure 29-25).

(2) The effectiveness of decontamination by scrubbing is generally inversely proportional to the roughness of the surface of the fruit. After the Marshall Islands incident, smooth-surfaced fruits were found to lose 90 percent of their radioactivity after washing, while washing rough-surfaced plants removed only 50 percent of the contamination. Plant foods, such as lettuce, having a very rough outer surface which cannot easily or effectively be decontaminated by peeling or washing, should be eaten only as a last resort. Other difficult foods to decontaminate by washing with water are dried fruits, such as figs, prunes, peaches, apricots, pears, and soybeans.

(3) Generally speaking, any plant food ready for harvest can be used for food if it can be effectively decontaminated. Growing plants, however, can absorb some radioactive materials through their leaves as well as from soil, especially if rains have occurred during or after the fallout period. With some elements, such as strontium, which is extremely soluble in water, data has shown greater amounts were taken up by plants through

EQUIVALENT RESIDUAL DOSE (ERD): The body can repair 90% of a dose of radiation damage. The unrepaired damage for any given day is the radiation dose multiplied by the ERD factor of the number of days after the exposure. NOTE: Each subsequent dose ERD Factor is based on the number of days after each subsequent dose.

DAY	FACTOR	DAY	FACTOR	DAY	FACTOR	DAY	FACTOR	DAY	FACTOR
1-4	1.000	10	.873	16	.764	22	.670	28	.590
5	.978	11	.854	17	.748	23	.656	29	.578
6	.956	12	.835	18	.731	24	.642	30	.566
7	.934	13	.817	19	.716	25	.629	31	.554
8	.913	14	.799	20	.700	26	.616	32	.543
9	.893	15	.781	21	.685	27	.602	33	.532

$$\begin{pmatrix}\text{1st}\\ \text{DOSE}\end{pmatrix}\begin{pmatrix}\text{ERD}\\ \text{FACTOR}\end{pmatrix} + \begin{pmatrix}\text{2d}\\ \text{DOSE}\end{pmatrix}\begin{pmatrix}\text{2d ERD}\\ \text{FACTOR}\end{pmatrix} + \ldots + \begin{pmatrix}\text{X}\\ \text{DOSE}\end{pmatrix}\begin{pmatrix}\text{X ERD}\\ \text{FACTOR}\end{pmatrix} + \begin{pmatrix}\text{SHELTER}\\ \text{DOSE}\end{pmatrix}$$

WATER SOURCES: Uncontaminated water--solar still or snow 6 inches below contamination level. Delay as long as possible from drinking from contaminated sources. PRIORITY: (1) underground, (2) running, (3) stationary. Filter sources 2 & 3 through 12 inches of earth and add purification tablets before use. BOILING WILL NOT REMOVE RADIOACTIVITY.

ANIMAL FOOD: DO NOT BUTCHER SICK ANIMALS. Discard internal organs and meat next to bone. Hides/feathers may be heavily contaminated. Eggs OK to eat.

SEA FOOD: Ocean sources OK, others on a risk basis. Fish from stationary bodies of water probably contaminated.

PLANT FOOD: Plants with edible portions below ground first choice. Smooth plant food easy to wash. Rough surfaced plants difficult to wash. Wash, pare, then wash again.

FIRST AID: NO FIRST AID FOR RADIATION SICKNESS. Infection, main danger. Personal hygiene important. Rest, avoid fatigue. Drink liquids.

BURNS: Normal first aid. Cool and cover burn. Do not use grease, etc. Treat for possible shock. Keep warm, lie with feet above head.

FRACTURES: Normal first aid. Immobilize and splint. Possible shock.

BLEEDING: Normal first aid. Apply pressure at break. Tight tourniquet can cause loss of limb--use only as last resort.

Figure 29-26. Equivalent Residual Dose Rate.

NUCLEAR EXPLOSIONS: FALL FLAT. Cover exposed body parts. Present minimal profile to direction of blast. DO NOT LOOK AT FIREBALL. Remain prone until blast effects over.

SHELTER: Pick ASAP, 5 minutes unsheltered max. PRIORITY: (1) cave or tunnel covered with 3 or more feet of earth, (2) storm/storage cellars, (3) culverts, (4) basements, (5) abandoned stone/mud buildings, (6) foxhole 4 ft deep--remove topsoil within 2 ft radius of foxhole lip.

RADIATION SHIELDING EFFICIENCIES: One thickness reduces received radiation dose by one-half. Additional thickness added to any amount of thickness reduces received radiation dose by one-half.

Iron/Steel	.7 in	Earth	3.3 in	Wood (Soft)	8.8 in
Brick	2.0 in	Cinder Block	5.3 in	Snow	20.3 in
Concrete	2.2 in	Ice	6.8 in		

SHELTER SURVIVAL: KEEP CONTAMINATED MATERIALS OUT OF SHELTER. Good weather; bury contaminated clothing outside of shelter--recover later. Bad weather; shake strongly or beat with branches. Rinse and (or) shake wet clothing--DO NOT WRING OUT.

PERSONAL HYGIENE: Wash entire body with soap and ANY water; give close attention to fingernails and hairy parts. No water: Wipe all exposed skin surfaces with clean cloth or uncontaminated soil. Fallout/dusty conditions, keep entire body covered. Keep handkerchief/cloth over mouth and nose. Improvise goggles. DO NOT SMOKE.

EXIT PLANNING: Explosion time and a dose-rate known, use rate decay nomograph.

Explosion time unknown: 1 Take dose-rate readings every hour. 2 When any reading is 1/2 of any previous reading, multiple time difference between two readings by 7/4. 3 Subtract the resultant from the time of first reading. 4 This new time is approximate time of explosion. Use new time with dose rate nomograph to determine safe exit time.

No rate meter, complete isolation first 4-6 days after last explosion.

Day 3/7: Brief exposure, 30 minutes MAX.

Day 8: Brief exposure, 1 hour MAX.

Days 9-12: Exposure of 2-4 hours per day.

Day 13 on: Normal movement.

MAXIMUM SURVIVAL DOSE/ERD: 200 roentgens

Figure 29-27. Nuclear Explosions.

their leaves than through their roots. In the tests conducted in the Marshall Islands, high levels of radioactivity were found in the water and on the external surfaces of the plants early after the detonation, but only small amounts of beta activity and no alpha activity were detected in the edible portions of the fruits and vegetables. High levels of contamination, similar in activity to that of the contaminated water nearby, were found in the sap of the coconut tree. From these and other tests, we know that small amounts of fission products may be immediately absorbed by plants growing on soil contaminated by fallout material.

(4) If the countermeasures recommended for obtaining protection against penetrating external gamma radiation and beta radiation are taken immediately, and the rules for constructing a shelter and selecting food and water are applied, a survivor's chances of surviving a nuclear attack are excellent. Figures 29-26 and 29-27 contain some summarized information applicable to a radioactive environment.

Chapter 30

BIOLOGICAL CONDITIONS

30-1. Introduction. Biological agents are viruses and micro-organisms, or their products, which are used to cause disease, injury, or death to people, animals, or plants (and, to a lesser extent, deterioration of material). Their use is an attempt to produce disease on a large scale. During war, these agents will probably be used primarily in a strategic role to attack rear area bases.

a. Most micro-organisms are harmless to humans, animals, or plants, and a few are helpful. Yeast is used in bread, beer, and cheese production. Some micro-oganisms, however, do produce disease. Biological agents could consist of:

(1) Fungi - mold, mildew, athlete's foot; histoplasmosis; and other pathogenic fungi.

(2) Bacteria - plague, tularemia, anthrax.

(3) Rickettsiae - Rocky Mountain spotted fever, typhus.

(4) Viruses - yellow fever, smallpox.

(5) Biotoxins - mushroom, algae, and bacterial poisons.

b. Many biological agents are living, require moisture, food, and certain temperature limits for life and growth. They are killed by simple acts such as boiling water, adding purification tablets to water, cooking food, exposing them to sunlight, and (or) using germicides. Biological agents enter the body through the nose, mouth, or skin; however, most will not penetrate intact skin. By preventing their entry into the body, a survivor is safe from biological agents.

30-2. Detection. There is no simple method of detecting biological agents. A person cannot see, feel, or taste these agents in a biological attack whether spread through conventional means or sabotage. Additionally, a person cannot taste the toxins in food. The basic methods of disseminating the agents are through the generation of an aerosol, the use of disease-carrying vectors, and food and water.

a. Aerosols are particles composed of many organisms, or a single organism, which are dispersed into the air and transported by air currents. Effective transmission as an aerosal requires that biological agents reach the target area with an effective percentage remaining alive and capable of causing disease. The appearance of certain clues may warn a survivor of an aerosol biological attack. They are:

(1) Aircraft dropping objects or spraying. Both enemy aircraft or friendly aircraft could be engaged in neutralizing or destroying opposing forces.

(2) Breakable containers or new and unusual types of shells and bombs, particularly those which burst with little or no blast.

(3) Smokes and mists of unknown origin.

(4) Unusual substances on the ground or on vegetation. Sick, dead, or dying animals.

b. Vectors, such as mosquitoes and ticks which carry disease, can be delivered to the target area in containers by aircraft or missiles. The containers rupture on impact and release the vectors. Some vectors need a "host" or carrier that can transmit the disease organism through the skin; others may be inanimate objects, such as contaminated food and water. Because disease organisms can infect personnel, the use of a "protective" mask may not help protect against viruses. The vectors can produce disease throughout their entire life spans, regardless of how far or where they travel. Furthermore, they may pass the disease to successive generations. Therefore, survivors must be extremely cautious when skinning wild game by wearing gloves and other protective clothing as the game may host fleas which carry many diseases. (NOTE: Contact with animals should be avoided unless they are to be used as food (figure 30-1).)

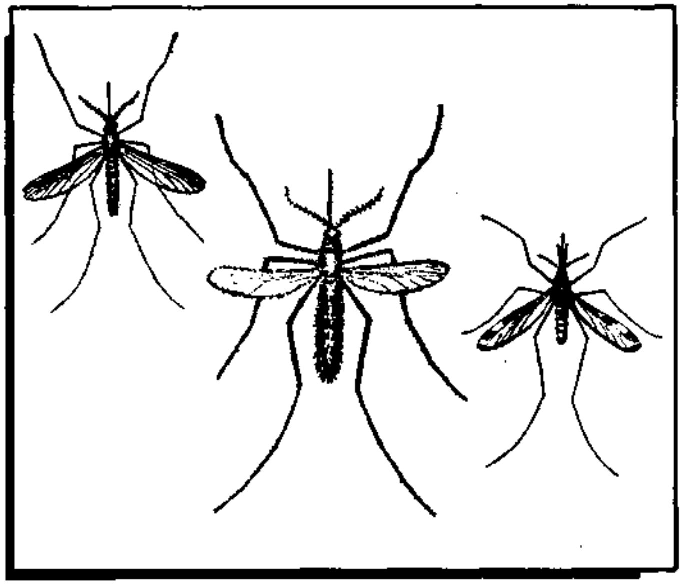

Figure 30-1. Vectors.

30-3. Climate. The various characteristics of aerosols and vectors will affect their utility in varying climates. While a survivor may encounter any form of biological agent once their use in combat occurs, the following factors may help a survivor assess the relative risk of various types of biological agents.

a. Aerosols are generally much more controllable in the area of application than are vector-borne agents. However, most aerosol agents deteriorate to some degree when exposed to direct sunlight. They are more suitable for use against small area targets.

b. Strong winds are necessary to ensure maximum area coverage of both aerosols and vectors. However, the density of coverage may be decreased by extremely high winds.

c. Vectors are less controllable than aerosols in the area of application following release. They are more suited for use against broad area targets. Although vectors tend to last longer in humid climates, many potential vectors (flies, fleas, mosquitos, lice, etc.) will thrive in virtually any environmental area.

d. Generally, nighttime (1 hour before sunset to 1 hour after sunrise) is the best period to dispense vectors. Vector movement and activity is usually greater during the cooler hours of darkness.

30-4. Terrain. Biological agent aerosols tend to follow the contour of rolling terrain and valleys very much like airborne particles, such as fog. Vegetation can slow the downwind travel of agents by removing some particles from the air. Due to a lack of sunlight, densely vegetated areas, such as a jungle (warm humid), allow some agents to thrive for extended periods of time. Because most of the biological agents are more hazardous when inhaled then when directly exposed to the skin, contamination of the ground following an attack is less dangerous to the survivor than exposure to an aerosol attack.

30-5. Protection. Defense against biological warfare is neither simple nor easy. The best defense against these agents is the natural resistance of the survivor's body, a high standard of personal cleanliness, careful attention to sanitation, good nutrition, up-to-date immunization status, proper use of drugs, and immediate self-aid to any break in the skin or a punctured wound. Germs must actually get inside the body to cause disease.

a. A protective mask, properly fitted and in good condition, will greatly reduce the danger of inhaling germs. If a mask is not available, a handkerchief or parachute material over the mouth and nose will suffice to provide protection. Since survivors cannot detect the presence of biological agents, they should wear the mask or some other protective equipment over the mouth and nose until they are rescued, if possible.

b. Cuts and sores, in addition to the nose and mouth, are open doors to germs trying to enter the body. Wounds must be kept clean and protected with a bandage. Any type of clothing will give some protection. Fasten the shirt or jacket collar, roll the sleeves down and button the cuffs, wrap the trouser legs or tuck them inside the boots, and tie down all other clothes to stop the entry of germs which may be in the air or on the ground. If the survivor has a uniform used for protection against chemical agents, it should be worn because it gives a greater degree of protection against germs than ordinary clothing.

c. Survivors should always be careful about eating and drinking during and after a biological attack. One of the easiest ways for germs to enter the body is through food and water.

30-6. Tips for a Survivor:

a. Keep the body and living quarters clean.

b. Don't neglect preventive medicine. Keep the shot record up to date.

c. Keep alert for signs of biological attack.

d. Keep the nose, mouth, and skin covered.

e. Protect food and water. Bottled or canned foods are safe if sealed. If in doubt, boil the food and water for 10 minutes.

f. Build a shelter, if possible. Shelter should be located and constructed to minimize vector and aerosol access to the survivor; for example, shelter enclosed—entrance 90 degrees to the prevailing wind.

g. If traveling, travel crosswind or upwind.

Chapter 31

CHEMICAL CONDITIONS

31-1. Introduction. Chemical agents may be solid particles, liquids, or gases which are toxic (poisonous) chemicals. These agents produce poisonous gas, fire, and smoke. The poisonous agents may produce casualties or irritating effects and render material or areas unusable. The body is attacked by chemical agents which produce specific damage depending on the type and concentration exposure of the agent used. Survivors must have a thorough knowledge of how each of these agents affects the body.

31-2. Chemical Groups. Chemical agents are divided into seven groups—nerve, blood, blister, choking, vomiting, incapacitating, and riot control. These agents can be dispersed by artillery shells, mortar shells, rockets, aircraft spraying, and bombs (figure 31-1).

a. Nerve Agents. The nerve agents are among the deadliest of all chemical agents. They directly affect the nervous system and are highly toxic in both liquid and vapor forms. Examples of G-agents are tabun (GA), sarin (GB), and soman (GD). These nerve agents may be absorbed through any body surface. When dispersed as a vapor, they are absorbed through the respiratory tract or the eyes, but as liquid nerve agents, they can be absorbed through the skin. They are usually quick-acting casualty agents. Symptoms accompanying very small doses are headaches, dizziness, dimmed vision, and nausea. Large doses of nerve agents can interfere with breathing and may cause a tightness in the chest or convulsions, paralysis, and death. Symptoms of large doses of nerve agents are unpredictable, and circulatory collapse can occur without warning. The first effect of eye exposure to agents will probably be a dimming of vision caused by contraction of the eye pupils to pinpoint size. The pinpoint pupils will more noticeably affect vision in dim light. If the nerve agent contaminates the skin only, the pupils may remain normal or be only slightly reduced in size with other symptoms being first to appear. The injuries caused by nerve agents may range from mild disability to death depending on the dose received and the adequacy and speed of self-aid treatment. Nerve agents are odorless, unlike most chemical agents. A survivor must rely on observation of living things and detection devices to identify their presence.

b. Blood Agents (Cyanides). Blood agents produce their effects by interfering with some vital process within the body. The usual route of entry is inhalation. They prevent the body cells from using oxygen. Hydrocyanic acid (AC) and cyanogen Chloride (CK) are the important agents in the group. CK also acts as a choking agent (figure 31-2).

Figure 31-1. Chemical Conditions.

Figure 31-2. Symptoms.

(1) Symptoms associated with blood agent contamination vary. One type of blood agent causes a marked increase in the breathing rate; whereas, another type causes a slow breathing rate, a choking effect, and a strong irritating effect. A slight exposure to still another type of blood agent causes headaches and uneasiness.

(2) Blood agents cause the skin to have a cherry-red color similiar to that seen in carbon monoxide poisoning. This symptom, by itself, may help identify the blood agents' poisoning. The symptoms produced by blood agents also depend upon the concentration of the

agent and the duration of exposure. If irreversible damage has not occurred, removal from the exposure to the agent may enhance recovery. Blood agents are used as quick-acting casualty agents. The speed in donning a protective mask is critical to survival in a blood agent attack.

c. Blister Agents (Vesicants). Blister agents were developed during World War I to circumvent the protective mask that had made chlorine gas obsolete. These agents are primarily designed to attack the body through the skin and eyes. They can also attack through the respiratory or digestive tracts and cause inflammation, blisters, and general destruction of tissue. Some examples of blister agents are mustard (HD), nitrogen mustards (HN), lewisite (L) and other arsenicals, mixtures of mustards and arsenicals, and phosgene oxime (CX). They are effective even in small quantities and produce delayed casualties.

(1) A drop of mustard-type agent the size of a pinhead can produce a blister the size of a quarter. Blister agents are more effective in hot weather than in cold weather. Vapors first affect the moist parts of the body (joints of arms and knees, armpits, and crotch). People who are sweating are especially sensitive to the agents. Blister agents are quickly absorbed through the skin. Reddening of the affected area may appear any time up to about 12 hours after exposure depending on the degree of contamination and the weather conditions. Blisters may appear in a day or less following the reddening. Healing time varies from about 6 days to as much as 8 weeks in severe cases. Since the damage is done during the first few minutes of exposure, speed in administering self-aid is essential.

(2) Damage to the eyes may be worse than the effects on the skin. Even as a liquid, the agent may only mildly irritate the eyes at first or there may be no pain at all. In a few hours, however, the eyes hurt, become inflamed, and are sensitive to light. Tears and great pain follow, and permanent injury can result. Some blister agents cause immediate pain in the eyes.

(3) When inhaled, blister agents inflame the throat and windpipe and cause a harsh cough. Cases of serious exposure may result in pneumonia and death. Immediate detection of blister agents and prompt protection against entry into the eyes and lungs and on the skin is vital.

(4) Blister agents may be absorbed by any material (wood, concrete, clothing, metal, plastics, or rubber). Direct skin contact with these objects can cause blistering. Liquid blister agents will eventually penetrate gloves and other garments. Immediate decontamination after exposure is essential to prevent delayed absorption.

d. Choking Agents (Lung Irritants). These agents cause irritation and inflammation of bronchial tubes and lungs but do not harm the skin or digestive system. They are usually disseminated as gases and inhaled into the body. The best known of these agents is phosgene. During and immediately after exposure, symptoms include coughing, choking, and a feeling of tightness in the chest, nausea, and occasionally vomiting, headache, and crying. If large amounts enter the lungs, they will fill with liquid causing death from lack of oxygen. A properly operating and well-fitted mask protects against all choking agents.

e. Vomiting Agents. These agents produce strong pepper-like irritation in the upper respiratory tract. Other symptoms include irritation of the eyes and uncontrollable tearing. Symptoms of these agents include a very stuffy nose, severe headache, intense burning in the throat, and tightness and pain in the chest. These are followed by uncontrollable sneezing, coughing, nausea, vomiting, and a general feeling of bodily discomfort. These agents are dispersed as aerosols and produce their effects by inhalation or by direct action on the eyes. If survivors inhale a vomiting agent before donning their masks, they may become ill after the respirator is on. As long as a vomiting agent is present, however, mask-wear is essential. The mask should be pulled away from the chin during actual vomiting, but not removed. If the survivor has vomited in the mask, caution should be taken to avoid inhaling or ingesting the vomit. Vomiting agents are not considered a major threat because of the comparative ease of protection against them and their lower toxicity unless used with other agents. Vomiting agents alone seldom produce death (figure 31-3).

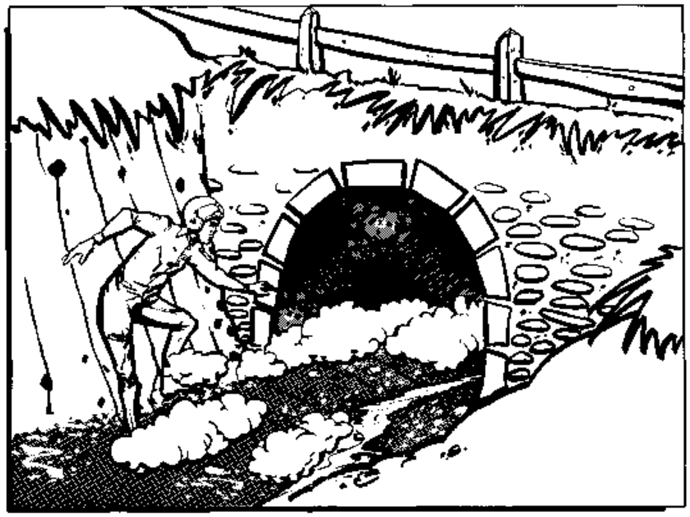

Figure 31-3. Avoiding Agents.

f. Incapacitating Agents. An incapacitating agent is any chemical which produces a temporary disabling condition which persists for hours to days after exposure to the agent has ceased (unlike that produced by riot control agents) and for which medical treatment, while not required, facilitates a more rapid recovery.

(NOTE: Symptoms of riot control agents may not be distinguished from other lethal agents; therefore, the survivor must be prepared to provide treatment for lethal agents.) In actual usage, the term "incapacitating agent" has come to refer primarily to those agents which:

(1) Produce their effects mainly by altering or disrupting the higher regulatory activity of the central nervous system (figure 31-4).

(2) Last for hours or days rather than the very short duration of riot control agents.

(3) Do not seriously endanger life (except when large doses are received) and produce no permanent injury.

Figure 31-4. Effects.

g. Riot Control (RC) Agents (Irritant Agents). RC agents are the least poisonous of the seven groups of chemical agents. They act primarily on the eyes, causing intense pain and tearing. Higher concentrations irritate the upper respiratory tract and the skin and sometimes cause nausea and vomiting. These agents may be dispersed as smoke or in solutions as droplet aerosols. Although they are used primarily in training and in riot control, some agents may be used in combat. When an unmasked person comes in contact with riot control agents, the effects are felt almost immediately. The effects begin in 20 to 60 seconds, depending upon agent concentration. Duration of effects lasts 5 to 10 minutes after removal to fresh air. There is usually no permanent damage to the eyes. For a short time (minutes), a person may be unable to see. If the mask is used before RC agents enter the eyes, increased protection is afforded.

31-3. Detection:

a. General Indications. Detection of a chemical agent requires the recognition of evidence gathered by direct or indirect means. Therefore, every survivor must be alert and able to detect any clues indicating chemical warfare is being used. General symptoms of chemical agents are tears, difficult breathing, choking, itching, coughing, and dizziness. In the presence of agents that are very hard to detect without the use of detection devices, survivors must watch their fellow aircrew members constantly for symptoms. Additionally, a survivor's surroundings may provide valuable clues to the presence of chemical agents; for example, dead animals, sick people, or people displaying abnormal behavior.

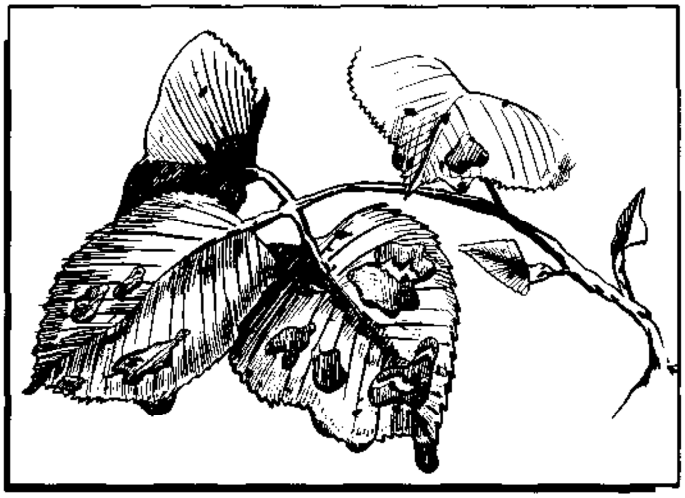

Figure 31-5. Detection.

b. Smell. Survivors cannot rely on the nose as a foolproof means of detecting chemical agents. Although some agents do have a characteristic odor, many others have little or no odor at all. An agent may smell quite differently to different individuals. A mixture of agents will have a different smell than any one agent by itself.

c. Sight. Since chemical agents are in one of three physical states—solid, liquid, or vapors—the sense of sight may help detect their presence. Most chemical agents in the solid or liquid state have some color. In the vapor state, some chemical agents can be seen as mist or thin fog immediately after bomb or shell bursts. Nerve agents are either a colorless liquid or a colorless vapor. Although survivors can't see nerve agents, their eyes may help by detecting the methods used to dispense the agents. Mustard gas, unless purified, is dark brown in its liquid state. As a liquid, it is easy to detect and would appear as oily, dark patches on leaves and buildings. However, liquid mustard changes slowly to a colorless gas. As a gas, it is still very toxic, but now the eyes will not be an effective aid to detection (figure 31-5).

d. Hearing. If survivors know the methods being used by the enemy to spread chemical agents, they can detect the sounds of the enemy's chemical munitions. For example, a bomb filled with an agent would probably cause only a muffled explosion; however, aircrew members untrained in ordnance may have difficulty in making this distinction.

e. Feel and Taste. Irritation in the nose or eyes or on the skin is an urgent warning to protect the body from chemical agents. Additionally, a strange taste in food, water, or cigarettes may serve as a warning they have been contaminated.

31-4. Protection:

a. Protective Actions. Survivors should use the following steps, listed in the order of importance, to protect themselves from chemical attack:

(1) Use protective equipment.

(2) Give quick and correct self-aid when contaminated.

(3) Avoid the areas where chemical agents exist.

(4) Decontaminate equipment and the body as soon as possible.

b. Equipment. Survivors' masks are as vital to them as lifejackets are to sailors or as parachutes are to fliers. If properly adjusted, they protect the face, eyes, and lungs from chemical agents. Survivors are responsible for proper care of the mask and should inspect the masks frequently to ensure they are free from damage and in good condition. Aircrew members located in areas of potential contamination are issued protective clothing.

c. Self-Aid. Survivors must apply self-aid skillfully and promptly after exposure. Not only is it important for them to know what to do, but they must also know what not to do. It is evident from previous information in this chapter that each type of chemical agent produces certain conditions which require special treatment. However, there are certain essentials of self-aid which, if applied soon enough, give some relief and may prevent serious injury.

(1) Since there are definite time limits after which self-aid becomes useless, immediate self-aid or personal decontamination is all-important if survivors are exposed to liquid nerve or blister agents. Since they may not know whether the contamination is by liquid nerve agent or liquid blister agent, the following procedures are recommended:

(a) Don the mask and clear it.

(b) Contact with thickened (persistent) nerve or mustard agents requires the use of decontamination kit, or if not available, tear away the contaminated area of clothing and rinse immediately with water.

(c) Rinse contaminated areas with water (removes nerve agents).

(d) If effects of nerve agents become apparent, then and only then, use an antidote, realizing the antidote provides protection only from nerve agent (GA, GB, GD). It is also incapacitating and is not effective against V agents.

(2) Use the self-aid procedure given in the following paragraphs for specific agents if the agent has been identified.

(a) Nerve Agents. The protective mask and hood, if available, must be donned immediately at the first sign of a nerve agent in the air. Stop breathing until the mask is on and the facepiece cleared and checked. The mask should be worn constantly until the absence of the nerve agent in the air is indicated or the individual moves into a clean area (where there are live animals, etc.) If, after masking, the survivor has an unexplained runny nose, tightness of the chest, dimness of vision, or breathing difficulty, the use of the antidote should be considered. (Use of the antidote is moderately to severely incapacitating. The survivor should consider the severity of symptoms, availability of the buddy care system and requirements for rescue before injecting the antidote.) Exposure to high concentrations of a nerve agent may bring on incoordination, mental confusion, and collapse so rapidly that the survivor cannot perform self-aid. If this happens, a fellow aircrew member must administer first aid. Severe nerve agent exposure may rapidly cause unconsciousness, muscular paralysis, and breathing stoppage. Any remaining survivors should keep their masks on and move out of the area as soon as possible. The following precautions should be used when applying self-aid for nerve agents:

-1. Antidote should not be used until certain it is needed. Pinpointing of the eye pupils or blurred vision, tightness in the chest, and difficulty in breathing are signs it is needed. If certain nerve agents are inhaled, the antidote counteracts them and makes the survivor feel better.

-2. If survivors have inhaled a very large dose of nerve agent vapor, they may need more than one injection of the antidote to relieve their symptoms. If the symptoms are steadily becoming worse and the first injection does not relieve them, or if their mouth does not become dry, it may be necessary to use an extra dose. Inject the second dosage in a different muscle. (NOTE: Do not use your own combo pen to inject a victim; use theirs. Additionally, if you find a deceased aircrew member, remove the combo pens from the deceased and take them with you.)

-3. If the difficulty in breathing is not relieved by the second injection, one more dose may be administered. Dryness of the mouth is a good sign. It means they have had enough antidote to overcome the dangerous effects of the nerve agent.

-4. If a drop or splash of liquid nerve agent gets into the eye, instant action is necessary to avoid serious effects. The eye should be irrigated immediately with water by tilting the head back, looking straight upward. Slowly pour water into the contaminated eye. Hold the eye open with the finger if necessary. Pour the water slowly so that irrigation will last at least 30 seconds. Survivors must irrigate in spite of the danger of breathing nerve agent vapor. Don the mask quickly after completing the irrigation. The pupil of the contaminated eye should be observed during the next minute, in a mirror if one is available or by someone nearby. If the pupil rapidly gets smaller, inject antidote intramuscularly at once. If the pupil does not get smaller, antidote is not needed.

-5. If liquid nerve agent gets on the skin or clothing, it should be removed. The liquid should be blotted off the skin with a handkerchief, a piece of cloth torn from the outer clothing, or personal decontaminating kits. Pinch-blotting the liquid won't spread the contamination. Contaminated clothing must be quickly removed and the skin washed with soap and water. In an emergency, the contaminated portion of the clothing can be cut away and the contaminated skin area flushed with water. The muscles under the contaminated area should be observed for any signs of twitching. If none develops in the next half hour and the survivor has no tightness in the chest, the decontamination was successful. If twitching of the muscles under the area of contaminated skin does develop, the antidote should be administered at once.

-6. Food and water which may be contaminated with nerve agents must be avoided. If a survivor has swallowed contaminated food or water and all of these symptoms occur—nausea, pains in the stomach, increased flow of saliva, and a tightness in the chest—the antidote must be administered.

(b) Blood Agents. If, during any chemical attack, a sudden stimulation of breathing or an odor like bitter almonds is noticed, the survivor must don the mask as quickly as possible. Speed is absolutely essential; this agent acts so rapidly that within a few seconds its effects will make it impossible for survivors to don the mask by themselves. The breath should be held until the mask is on the face, if at all possible. This may be very difficult since the agent strongly stimulates respiration.

(c) Blister Agents. The protective mask, hood, and clothing must be worn when liquid or vaporized blister agents are known to be present. There are two groups of blister agents, one called mustards and the other called arsenicals. Self-aid against mustards and arsenicals is the same. A liquid mustard in the eye will not hurt immediately. A liquid arsenical in the eye will sting and hurt severely.

-1. To remove a liquid blister agent from the eye, the eye is irrigated using the same procedure as for removing nerve agents. Speed in decontaminating the eye is absolutely essential. The self-aid procedure is very effective for mustard within the first few seconds after exposure but is of little value after 2 minutes.

-2. Generally, for any liquid blister agent on the skin, the survivor should:

-a. Pinch-blot to prevent the liquid from spreading, using cloth or any other absorbent material at hand. The used cloth or absorbent material should then be discarded.

-b. Scrub the skin with soap and water and rinse thoroughly with clean water. When scrubbing, special attention must be paid to areas not covered by clothing (neck and ears).

-3. Survivors must decontaminate or remove any clothing which is contaminated with liquid blister agent. Small areas on the clothing can be decontaminated by using soap and water. The contaminated parts of the clothing can also be cut out, thus making the clothing safe to wear.

-4. When mustard vapor is detected, a survivor must put on the mask and leave it on until clear of the area. There is no decontamination procedure of any value when a mustard vapor agent is used. The damage is done as soon as the mustard vapor strikes the eyes, although the full extent of the injury may not appear for several hours.

(d) Choking Agents. The protective mask should be donned immediately upon detection of any choking agents in the air by odor (like cut green corn or grass), irritation of the eyes, or change in the taste of a cigarette (smoking may become tasteless or offensive in taste). Survivors should hold their breath while masking. If an agent has been inhaled, normal survival duties should be continued unless there is difficulty in breathing, nausea and vomiting, or more than the usual shortness of breath from exertion. If any of the above symptoms occur, survivors should rest.

(e) Incapacitating Agents. The mask should be donned immediately. Complete cleaning of the skin with soap and water should be done at the earliest opportunity. The eyes should be flushed with clear water only. Survivors should shake or brush clothing, and when conditions permit, change clothing and thoroughly wash the contaminated clothing.

(f) Vomiting Agents. The protective mask must be worn in spite of coughing, sneezing, salivation, and nausea. The mask can be lifted from the face briefly if necessary to permit vomiting or to drain saliva from the facepiece. Carrying on duties as vigorously as possible will help to lessen and to shorten the symptoms. Survival duties can usually be performed despite the effects of vomiting agents.

(g) Riot Control Agents. After the protective mask has been donned and cleared, the eyes should be kept open as much as possible. When vision clears, activities can continue. The eyes should not be rubbed. If drops or particles have entered the eyes, the eyes can be flushed with water. Chest discomfort can usually be relieved merely by talking.

31-5. Avoiding Chemical Agents. If survivors are hit by a chemical attack, they may have to remain in a contaminated area. After an attack, they should not expose themselves to other enemy weapons and must seek areas which are less contaminated. If the attack is on a very small scale, they might seek an upwind area. Depending on the area and weather conditions, crosswind movement may be best. Chemical agent attacks may cover too large an area to permit area avoidance. Selecting routes on high ground may be advisable because gas is usually heavier than air and tends to settle in low places. Cellars, trenches, gullies, and valleys are examples of places to avoid when possible. Woods, tall grass, and bushes tend to hold chemcial agent vapors. (NOTE: Survivors have a better chance to avoid chemicals if they are familiar with the attack areas and if they know their personal location.)

31-6. Decontamination of Chemical Agents. Decontamination is removing, neutralizing, or destroying the agents. The purpose of personal decontamination is to remove agents from the body or personal equipment before serious injury occurs. An example of decontamination by removal is pinch-blotting the chemical agent from the skin. Neutralization makes the agent harmless. The contaminated cloth could also be buried. Common sense and quick thinking play a big role in personal decontamination. Survivors may have to rely on whatever they have at hand to remove chemical agents from the skin, eyes, or equipment. If liquid nerve or blister agents touch any part of the body, they must be removed as quickly as possible. If survivors are caught without soap and water, then anything which can dilute or remove the agents will have to be used; it may be mud, dirt, or urine. A crude remover may remove only two-thirds of the agent, but it is far better than leaving the agent in full concentration. It must be remembered that nerve and blister agents penetrate very rapidly. (NOTE: Use a scraping action to avoid pressing the agent into the skin.)

a. Soap is excellent for removing chemical agents. Cold water, while not as good as warm water, does dilute or weaken chemical agents. Hot, soapy water removes agents quickly. If the operational situation permits, a bath or a shower should be taken. The mask should be left on until after survivors have washed their hair and thoroughly scrubbed themselves while avoiding wetting the canister or cheek pads. Exposed areas and hairy regions of the body should be given extra attention.

b. When clothes have been exposed only to chemical agent vapors, airing usually decontaminates them (with the exception of mustard vapors which will absorb into the garment and requires washing for removal). If chemical agent droplets or liquid splashes are present, survivors will need detergent or soap and water. Wool clothes are best washed in soapy, lukewarm water. Cotton clothes can be boiled.

c. Boots or shoes can be scrubbed with soap and water and rinsed at least twice. If the survivors' choice is wearing contaminated clothes and shoes or nothing, decontamination of the material must be done the best way possible. Almost any effort will help the survival situation.

31-7. Tips for a Survivor:

a. Keep the body and living area clean.

b. Keep the nose, mouth, and skin covered.

c. If a protective mask is needed, but unavailable, improvise. The charcoal cloth from the CD suit or undergarment makes a moderately effective mask for short-term agent exposure. The use of the aircrew helmet, visor, and charcoal mask may provide a higher level of protection for the eyes and respiratory system.

d. Build a shelter or rest area in a clearing away from vegetation. Decontaminate the ground by removing the top soil. Keep the entrance closed and 90 degrees to the prevailing wind.

e. Do not use wood or vegetation from a contaminated area for a fire.

f. Look at the area around a water source and check for foreign odors (garlic, mustard, geranium, bitter almond), oily spots on the surface or nearby, and the presence of dead fish and animals. If they are present, do not use the water.

g. Keep food and water protected. Bottled or canned foods and water are safe if sealed, and the cans are decontaminated before opening.

h. If possible, obtain water from a closed source, precipitation (if there is no evidence of agent vapor in the air), and from a slow-moving stream after it has been filtered.

i. Do not use plants for food or water in a contaminated area.

j. Do not use sick animals as a food source. When skinning animals in a contaminated area, use protective clothing (gloves).

k. If traveling, travel crosswind or upwind.

MANUAL COMPLETO DE
SUPERVIVENCIA

Hugh McManners

EDITORIAL LA ISLA

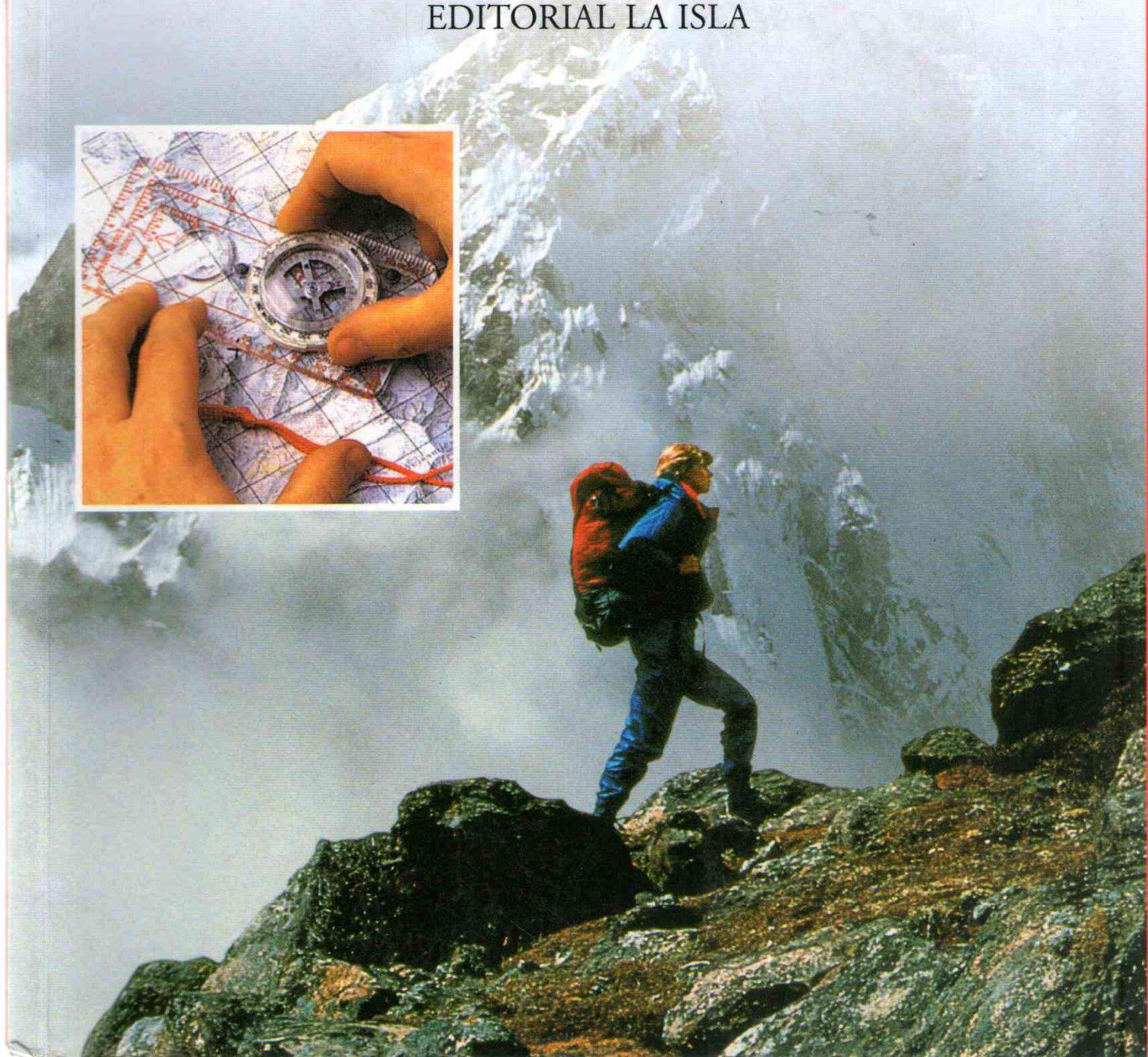

LA ISLA

A DORLING KINDERSLEY BOOK

Título original:
The Commando Survival Manual

Traducción:
Maite Rodríguez Fischer

Revisión científica y técnica de la edición en lengua española:
Ignacio Añibarro; Koldo Sansinenea;
Javier Busselo; Pablo Sánchez; José M. Peña;
Román Muñoz; Iñaki García;
Cándido García; Asier Beltza; Giorgio Studer
Espeleólogos.
Profesionales expertos en vida en exteriores
y acampada en diversos entornos.

Coordinación de la edición en lengua española:
Cristina Rodríguez Fischer

Primera edición argentina, versión en rústica, 1999
Primera edición argentina 1994

© 1994, 1999 La Isla, S.R.L., Buenos Aires
© 1994, 1998 Dorling Kindersley Limited, Londres
© 1994, 1998 del texto, Hugh McManners

I.S.B.N.: 950-637-105-9
Hecho el depósito que marca la ley 11.723
Impreso en Singapur

Todos los derechos reservados.
Queda prohibida la reproducción total o parcial de esta obra, sea por medios mecánicos o electrónicos, sin la debida autorización por escrito del editor.

Venta exclusiva en Argentina,
Chile, Paraguay y Uruguay

NOTA IMPORTANTE

Algunas de las técnicas descritas en este libro únicamente deben utilizarse en situaciones de emergencia grave, cuando la supervivencia de los individuos dependa de ellas. En consecuencia, los editores no pueden hacerse responsables de cualquier daño, pérdida o acción legal que pueda resultar del uso o abuso de la información contenida en este libro. No practique estas técnicas en propiedades privadas sin el permiso del propietario, y obedezca la legislación en lo que concierne a la protección de la tierra, la propiedad, las plantas y los animales.

CONTENIDO

Introducción 6

CAPÍTULO 1 ¿ES USTED UN SUPERVIVIENTE?

Condición física 12
Necesidades nutricionales 14
La adaptación al medio 16
Estar en forma 18

CAPÍTULO 2 EQUIPAMIENTO BÁSICO

La ropa adecuada 22
Calzado 24
Equipo personal de campamento 26
Equipo básico de supervivencia 28
Camas y sacos de dormir 30
Herramientas 32
Cuerdas y nudos 34
El manejo de la cuerda 36

CAPÍTULO 3 LA VIDA EN LA NATURALEZA

La elección del campamento 40
La elección de una tienda 42
Montar una tienda de campaña 44
Desmontar una tienda de campaña 46
La construcción de un refugio 48
Refugios en la nieve 50
Refugios tropicales 52
Refugios permanentes 54
Un campamento seguro 56

El fuego en el campamento 58
Cómo preparar un fuego 60
Cómo encender un fuego 62
Levantar el campamento 64

CAPÍTULO 4 EN BUSCA DE AGUA

La importancia del agua 68
Recoger agua 70
Fuentes naturales de agua 72
La depuración del agua 74
Transporte y almacenamiento del agua 76

CAPÍTULO 5 ENCONTRAR Y PREPARAR ALIMENTOS

Alimentos para el viaje 80
Plantas comestibles 82
Plantas venenosas 84
Raíces, tubérculos y bulbos 86
Nueces y frutos 88
Hongos y líquenes 90
Las plantas como alimento 92
Animales invertebrados 94
Los animales invertebrados como alimento 96
Peces de agua dulce 98
Peces de agua salada 100
Equipo de pesca 102
Técnicas de pesca 104
Redes y trampas de pesca 106
El empleo de redes y trampas de pesca 108
Preparar y cocinar el pescado 110
Utensilios de cocina 112
Cocinas y hornos 114
Métodos de cocción 116
La conservación de los alimentos 118

CAPÍTULO 6 EN MARCHA

La planificación del viaje 122
La orientación 124
El empleo de mapas 126
Encontrar su posición 128
Principios de navegación 130
La evaluación del tiempo 132
Prepararse para el viaje 134
La marcha a pie 136
Caminar sobre hielo y nieve 138
Viajar sobre nieve 140
Cruzar el desierto 142
Viajar a través de la selva 144
Una escalada segura 146
Escalada de emergencia 148
Cruzar las aguas 150
Aguas peligrosas 152
Construir una balsa 154
El empleo de balsas y otras embarcaciones 156
El uso de vehículos 158
Otros medios de transporte 160

APÉNDICE A

Abandonar el barco 162
La supervivencia en el mar 164
Animales peligrosos 166
Peligros naturales 168
Meteorología extrema 170
Códigos y señales para comunicarse 172

APÉNDICE B

Primeros auxilios 174

Glosario 184
Lista de comprobación para supervivencia 187
Índice 188
Agradecimientos 192

INTRODUCCIÓN

Es posible que la sentencia de los grandes filósofos «Cualquier tonto puede sentirse incómodo» la formulara después de un día de campo pasado por agua. Tanto si trata de secar su ropa en un campamento inundado de lluvia como si intenta sobrevivir a un desastre inesperado, el conocimiento de los principios básicos de la vida al aire libre le resultarán de una ayuda inmensurable. Estos principios incluyen la construcción de tiendas y refugios, la elección o confección de sacos de dormir o camas, encender un fuego, así como la prevención o cura de ampollas y de algunas enfermedades. El aprendizaje de algunos principios más especializados de supervivencia, como la navegación en cualquier tipo de terreno o la depuración de agua, es tan interesante y útil para los usuarios con fines recreativos, como vital para quienes desafortunadamente se convierten en «supervivientes».

A LA ESPERA DE LO INESPERADO

En la práctica, nuestro lema diario era «Espera siempre lo inesperado». Los largos años de entrenamiento en lugares lluviosos y potencialmente peligrosos valieron la pena; nuestro equipo de operaciones estuvo destacado en todo el mundo, a menudo trabajando en medios hostiles. La gente que trabaja en las fuerzas especiales es bastante realista sobre su trabajo, así que una vez se había comentado todo en el entrenamiento, no había mayores problemas al llevarlo a cabo en la realidad. Sin embargo, estábamos enfrentándonos a «lo inesperado».

Por el contrario, para los civiles que no se enfrentan a nada más peligroso que la posible avería del automóvil o al mal tiempo, el prepararse para alguna situación de emergencia puede parecerles ridículo. Sin embargo, la más inofensiva caminata familiar, particularmente en un terreno montañoso, puede convertirse rápidamente en una experiencia desagradable si los participantes ignoran los principios básicos de la predicción del tiempo, primeros auxilios, navegación, ropa y calzado. Únicamente conociendo la vida en la naturaleza evitaremos muchos desastres comunes, o por lo menos los identificaremos antes de que se conviertan en una amenaza seria para nuestra vida.

Prepararse para lo inesperado es un proceso constante de intentar entender cómo funcionan las cosas, y establecer líneas paralelas entre lo que forma parte de nuestra propia experiencia y entendemos, y las cosas extrañas y diferentes sobre las que no tenemos experiencia. Incluso cuando decidimos visitar un área desierta o un país subdesarrollado, permanecemos como espectadores, manteniéndonos alejados de la realidad de la vida diaria de aquellos lugares por la naturaleza misma de nuestra vida urbana, llena de tecnologías modernas. Pero debemos cambiar nuestras actitudes de comodidad si queremos entender cómo funciona la vida realmente. Si tuviéramos que establecernos en una remota isla tropical después de un accidente aéreo, por ejemplo, es muy poco probable que dispusiéramos de agua embotellada, de comidas precocinadas o preparadas, e incluso de gafas de sol, y los accesorios urbanos como las tarjetas de crédito serían absolutamente inútiles.

Así pues, es en nuestro propio beneficio que debemos hacer un esfuerzo serio para abandonar nuestras vidas cómodas y protegidas. Si podemos entender cómo encajamos en el mundo, reconociendo honestamente nuestras fuerzas y debilidades físicas y mentales, podemos dar el siguiente paso, convirtiéndonos en actores del mundo real, sin importar en qué parte del mismo actuamos.

CONOCIMIENTOS QUE PUEDEN SALVARLE LA VIDA

Es imposible predecir los desastres que pueden ocurrir, así que no podrá adquirir las habilidades específicas necesarias para cubrir todas las opciones. Entender los principios básicos de la supervivencia es mucho más importante que aprender técnicas de memoria, ya que la mayoría de las técnicas requieren alguna modificación cada vez que se utilizan. Nada es fácil, no importa lo bien entrenado que pueda estar, pero si aprende los principios básicos, será capaz de adaptarse a una gran variedad de situaciones diferentes.

La tecnología moderna enmascara a menudo los principios básicos que rigen todos los trabajos físicos. Creamos máquinas para cualquier trabajo, liberándonos de faenas que ocuparían la mayor parte de nuestro tiempo. Olvidamos los principios de la física que rigen desde el clima hasta las herramientas que improvisamos.

Es muy posible que hayamos aprendido y entendido muchos de estos principios en nuestras vidas, pero la mayoría de ellos están guardados en nuestras mentes, y no siempre somos capaces de aplicarlos a las situaciones de la vida real. La vida urbana tiende generalmente a hacernos inflexibles en la aplicación práctica de nuestros conocimientos, simplemente porque casi nunca nos encontramos en el tipo de situaciones en las que no hay otra opción que la improvisación para sobrevivir.

LOS PRINCIPIOS DE LA SUPERVIVENCIA

Cualquier tipo de conocimiento es importante. Incluso la física o la química casi olvidada del colegio, o una lección de biología, pueden ahorrarnos horas de experimentación. Cuando somos niños, construimos herramientas y otros artefactos continuamente, aprendiendo a aplicar los principios básicos, un proceso que se detiene cuando crecemos y empezamos a comprar artículos fabricados. En la naturaleza debemos volver a nuestra infancia en muchos aspectos, y volver a aprender a improvisar e inventar. Sin embargo, debemos trabajar como adultos, considerando los principios de ingeniería como el esfuerzo, la resistencia y la tensión, que deben combinarse teniendo una idea bien clara del propósito que nos guía. Este proceso está gobernado por la necesidad: si no lo logramos nos mojaremos, nos quedaremos hambrientos o padeceremos frío.

Al combinar paso a paso las instrucciones con fotografías e ilustraciones, este libro contiene información sobre una gran cantidad de técnicas prácticas, que muestran una transición fácil entre la vida urbana y la naturaleza en la que nuestros antepasados primitivos tuvieron que sobrevivir. Los conocimientos que contiene este libro podrían llegar a salvar su vida, si de repente se ve en la necesidad de tener que sobrevivir —incluso temporalmente—, como lo hicieron nuestros ancestros, por ejemplo después de un accidente, desastre o avería de su vehículo, o en condiciones climáticas adversas, cuando los servicios básicos se ven interrumpidos.

Por otra parte, puede que sea un aficionado a las caminatas o un amante de las excursiones en la naturaleza, para alejarse de las presiones y complicaciones de la vida urbana. Cualquiera que sea su situación, este libro le será de ayuda inestimable, ya que le proporcionará gran cantidad de información necesaria, al tocar una amplia gama de temas especializados.

CONDICIÓN FÍSICA Y MENTAL

El capítulo uno trata sobre la condición física y mental del hombre urbano, comparándolo con otros animales y con pueblos primitivos que habitan tierras inhóspitas, a las que han tenido que adaptarse para sobrevivir. En situaciones de supervivencia, nos son impuestas las rígidas leyes naturales, y la realidad irrumpe en nuestra vida moderna de color de rosa. De repente, una buena condición física resulta imprescindible. Al estar en forma, no sólo aumentamos nuestras posibilidades de sobrevivir a un desastre, sino que también sabemos exactamente lo que nuestros cuerpos son capaces de aguantar. En este capítulo se dan algunos consejos sencillos pero efectivos para mantenerse físicamente en forma, incluyendo también información práctica sobre las necesidades nutricionales para mantener una buena salud a lo largo del año.

EL EQUIPAMIENTO ADECUADO

En el capítulo dos se estudia el equipo necesario en zonas de montaña, que va desde la ropa adecuada, los sacos de dormir y las herramientas, hasta un equipo de supervivencia personalizado que puede reunir por sí mismo. Además de comparar artículos manufacturados, este libro le muestra cómo confeccionarlos usted mismo, si carece del equipo apropiado. Si no se dispone de materiales adecuados, los métodos de construcción deben modificarse, los diseños varían y se llega a soluciones intermedias —prácticamente cualquier cosa puede fabricarse a partir de unas cuantas ramas y trozos de cuerda. Que el equipo improvisado no sea particularmente vistoso y que deba utilizarse con precaución, no reviste la menor importancia, siempre y cuando cumpla una función. A lo largo de este capítulo —y de hecho, en todo el libro— se enfatiza la importancia de la conservación de los recursos naturales. No existe ninguna necesidad de dañar el medio ambiente. De hecho, al adaptarnos al medio natural, tendremos muchas más posibilidades de sobrevivir, no sólo como individuos, sino también como especie.

PRINCIPIOS BÁSICOS DE LA VIDA EN UN CAMPAMENTO

Si se encuentra en una situación de supervivencia, la primera acción a efectuar consiste en buscar cobijo, ya sea erigiendo una tienda, o construyendo un refugio a partir de elementos naturales. El capítulo tres considera los distintos tipos de refugio necesarios en cada terreno diferente, ya sea en una jungla tropical o en una región polar. El fuego es el segundo punto a considerar en una situación de supervivencia —con un fuego puede cocinar la comida, mantener su temperatura corporal, depurar agua para que sea potable y alentar su ánimo. Cómo construir y encender distintos tipos de fuego es un punto que se toca en el capítulo tres. El siguiente capítulo trata sobre la importancia del agua en el cuerpo humano, y le muestra cómo obtenerla. Es poco probable que el agua que encuentre sea pura; puede incluso transmitirle enfermedades que amenacen su salud. Así pues, se enfatiza la importancia de la depuración del agua. Después del agua, debe proveerse de alimentos. El capítulo cinco considera los distintos alimentos que puede comprar para llevar en una excursión, y cómo reconocer y recolectar diferentes tipos de comida silvestre, desde plantas y otros invertebrados hasta pescado. Este capítulo también le aconseja sobre cómo preparar la comida silvestre para su consumo, así como su conservación para épocas en las que no se encuentre disponible de forma natural.

VIAJAR POR ÁREAS VÍRGENES

Es probable que no haya tenido la experiencia de cruzar un área totalmente virgen, pero mientras conozca los principios básicos de supervivencia en cada zona, sus oportunidades aumentarán enormemente. El capítulo seis trata sobre la forma de viajar en áreas tan diversas como la jungla y el desierto, así como las regiones polares y las Montañas Rocosas. También se indican algunas técnicas para cruzar ríos y para caminar correctamente a fin de minimizar los accidentes, así como algunos métodos elementales de orientación y navegación que podrá utilizar en todo el mundo, y técnicas básicas de escalada en rocas, útiles si surge alguna emergencia. Además de los viajes a pie, este capítulo también considera el empleo de vehículos motorizados, trineos de perros, balsas, canoas y otras embarcaciones; e incluye consejos sobre la manera de confeccionar, cargar y utilizar este tipo de transportes.

PERMANEZCA TRANQUILO

Su supervivencia dependerá de que tome las decisiones adecuadas. Pero sólo entendiendo lo que puede pasar en una situación de supervivencia será posible tener la suficiente previsión para poder predecir algún desastre potencial, y así tomar la acción correctiva adecuada para prevenirlo. En las situaciones de supervivencia nos encontramos arrojados de repente hacia algo que no podemos evitar y en lo que muy pocas de nuestras habilidades normales son relevantes. Nuestro primer problema consiste en aceptar lo que ha sucedido. El pánico y la depresión son los mayores enemigos de la mente, e impiden que aflore en las personas el deseo innato de sobrevivir. Un reconocimiento lógico y franco de la situación, de los recursos y de las opciones, determinará los primeros pasos a tomar.

En el apéndice A se analizan los desastres y las situaciones peligrosas; así como sus causas, cómo predecirlas, prepararse para afrontarlas y sobrevivir a sus consecuencias. A continuación, se mencionan algunos ejemplos de situaciones peligrosas en las que podría llegar a encontrarse, como quedar a la deriva en una balsa salvavidas en mar abierto, enfrentarse a un animal peligroso, a los desastres naturales como los terremotos o a condiciones climáticas extremas como los huracanes. Este apéndice indica cuándo debe considerarse que tales desastres son inminentes y cómo superarlos si se enfrenta a ellos.

El apéndice B repasa las técnicas básicas de primeros auxilios, con un énfasis especial en las situaciones de supervivencia, en las que no puede simplemente llamar a una ambulancia por teléfono. Asimismo, le proporciona información sobre las enfermedades infecciosas en el mundo y cómo evitarlas.

DEPENDE DE USTED

Por encima de cualquier otra cosa, la supervivencia exige una gran fuerza mental y claridad de pensamiento. La evaluación lógica de todos sus conocimientos y un reconocimiento a conciencia de sus habilidades y de las de su equipo, pueden llevarle a una conclusión desagradable o a una decisión difícil. Sin embargo, a diferencia de lo que ocurre en la vida diaria, en una situación de supervivencia, usted determina su propio futuro, y de sus decisiones dependen su propia vida o muerte.

Aunque este libro le enseña cómo podrían hacerse las cosas, en una situación de supervivencia real debe hacerlas por sí mismo. Si las cosas no salen como se describe en estas páginas, tendrá que arreglárselas solo. Sin embargo, si sigue los detalles específicos de este libro, y adapta los principios básicos, aumentarán notablemente sus posibilidades de supervivencia, incluso en las condiciones más adversas.

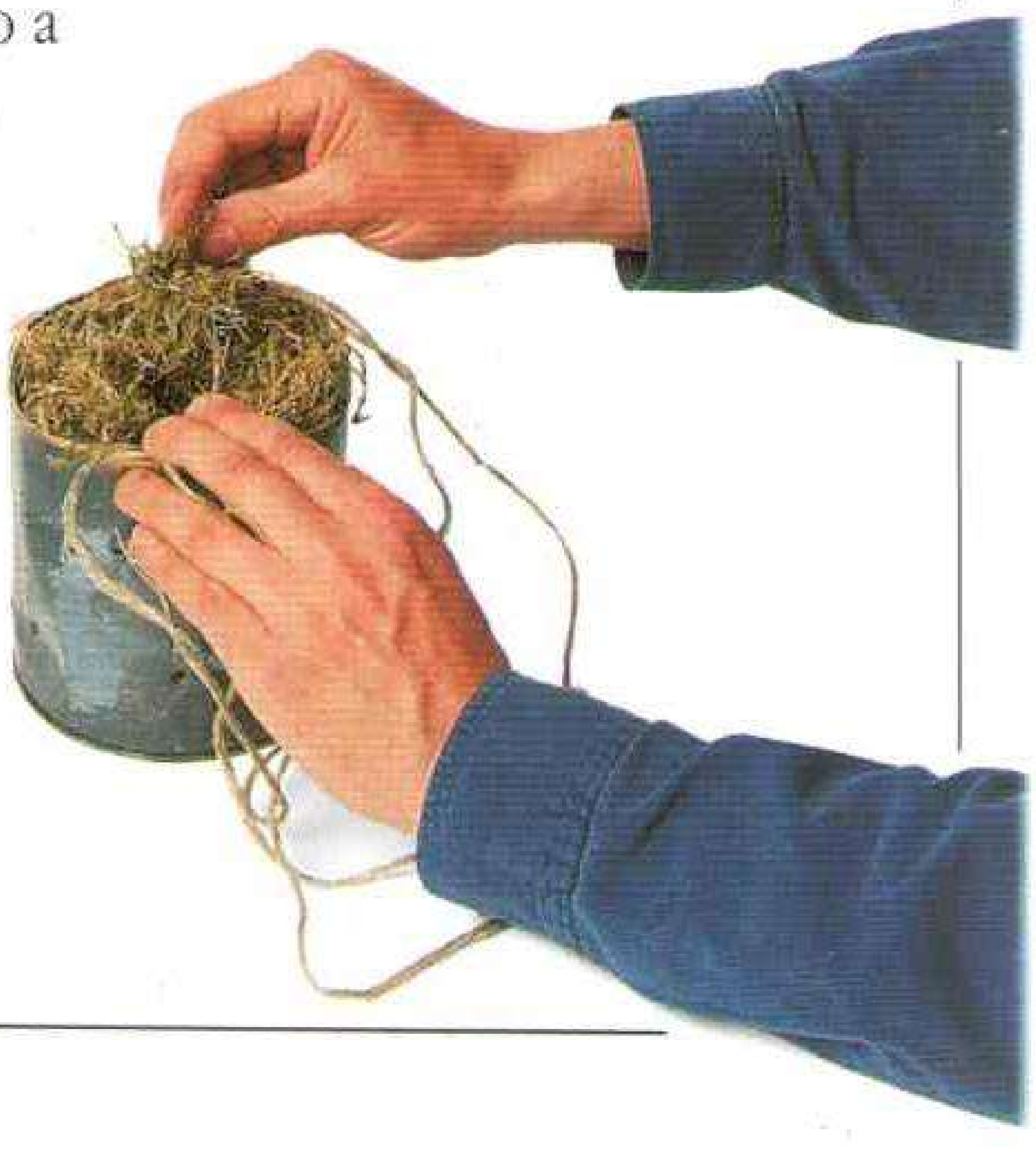

CAPÍTULO UNO

¿ES USTED UN SUPERVIVIENTE?

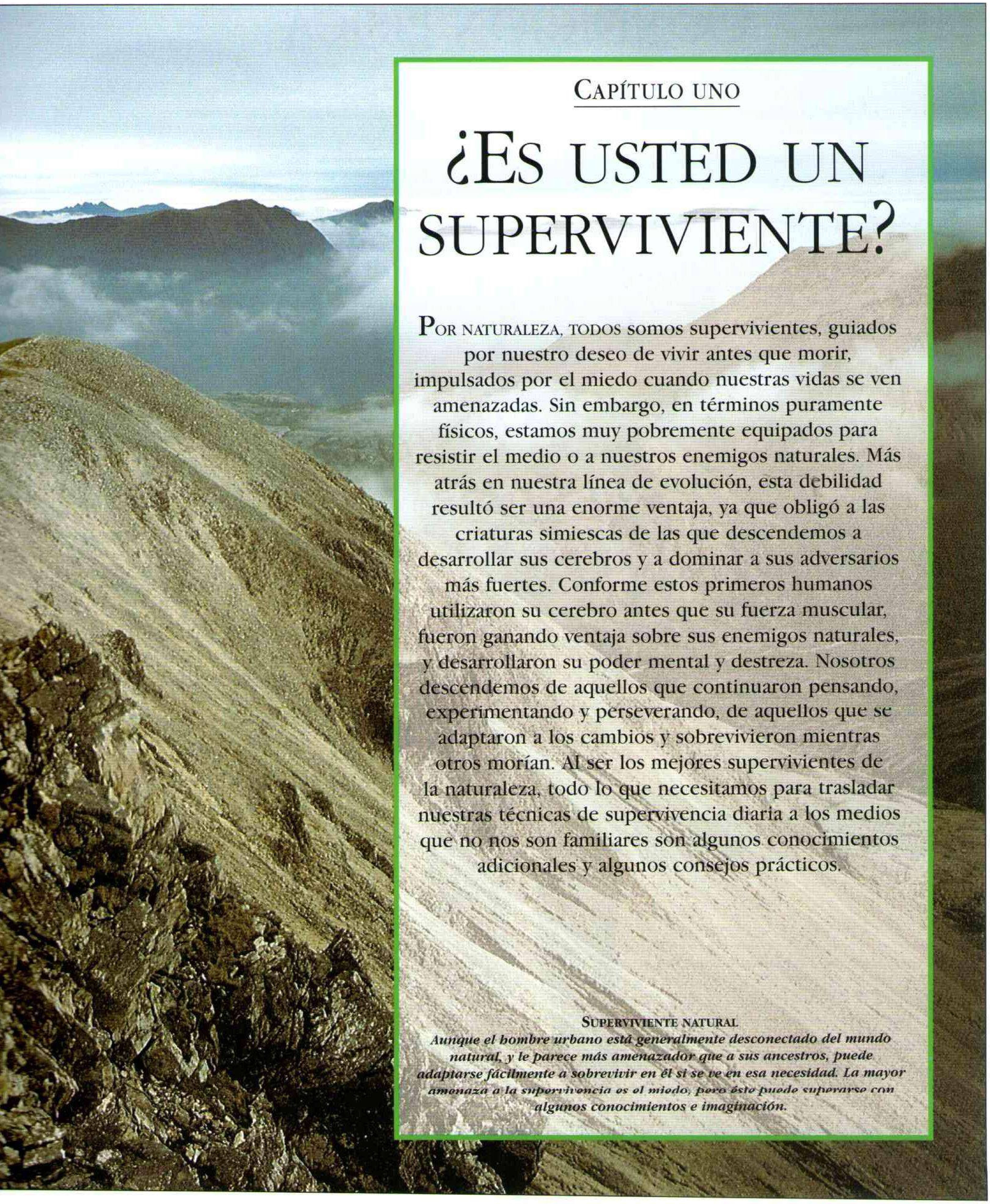

POR NATURALEZA, TODOS somos supervivientes, guiados por nuestro deseo de vivir antes que morir, impulsados por el miedo cuando nuestras vidas se ven amenazadas. Sin embargo, en términos puramente físicos, estamos muy pobremente equipados para resistir el medio o a nuestros enemigos naturales. Más atrás en nuestra línea de evolución, esta debilidad resultó ser una enorme ventaja, ya que obligó a las criaturas simiescas de las que descendemos a desarrollar sus cerebros y a dominar a sus adversarios más fuertes. Conforme estos primeros humanos utilizaron su cerebro antes que su fuerza muscular, fueron ganando ventaja sobre sus enemigos naturales, y desarrollaron su poder mental y destreza. Nosotros descendemos de aquellos que continuaron pensando, experimentando y perseverando, de aquellos que se adaptaron a los cambios y sobrevivieron mientras otros morían. Al ser los mejores supervivientes de la naturaleza, todo lo que necesitamos para trasladar nuestras técnicas de supervivencia diaria a los medios que no nos son familiares son algunos conocimientos adicionales y algunos consejos prácticos.

SUPERVIVIENTE NATURAL
Aunque el hombre urbano está generalmente desconectado del mundo natural, y le parece más amenazador que a sus ancestros, puede adaptarse fácilmente a sobrevivir en él si se ve en esa necesidad. La mayor amenaza a la supervivencia es el miedo, pero éste puede superarse con algunos conocimientos e imaginación.

CONDICIÓN FÍSICA

EL HOMBRE ES físicamente inferior a muchos otros animales, si no a la mayoría. Sin embargo, su gran ventaja es la falta de especialización. Tiene la inteligencia y la capacidad de desarrollar nuevas habilidades y de adaptarse a nuevas situaciones. En un medio urbano, el hombre no utiliza su cuerpo como lo haría en la naturaleza. El hombre primitivo es más dueño de su propio destino que su equivalente urbano.

EL CUERPO HUMANO

Olfato. *El hombre tiene un buen sentido del olfato, aunque no está tan desarrollado como el de otros animales, como el del perro. En la naturaleza, el olfato es útil para la identificación de alimentos, para advertir la presencia de otras criaturas y para prevenir peligros como los fuegos forestales.*

Oído. *Los animales salvajes están acostumbrados a identificar los peligros por su sonido. Sin embargo, el hombre urbano vive en un medio ruidoso que reduce la sensibilidad de sus oídos y su deseo de escuchar con atención.*

Gusto. *Las papilas gustativas en la lengua permiten al hombre la percepción de sabores agrios, dulces, salados y amargos. El cerebro interpreta la información recibida a través de las papilas y del olor del alimento para crear un sabor particular.*

Ritmo respiratorio. *El ritmo respiratorio del hombre depende de la cantidad de oxígeno que el cerebro cree que necesita el cuerpo. Este ritmo aumenta al realizar esfuerzos o al experimentar temor, y en las grandes altitudes, donde el aire contiene una menor cantidad de oxígeno.*

Columna vertebral. *El hombre aún está diseñado para caminar sobre cuatro patas, por lo que el dolor de espalda es el precio de permanecer erguido sobre dos piernas.*

Pulgares opuestos. *El hombre tiene pulgares y dedos que puede mover independientemente para coger los objetos. Esto le permite confeccionar y utilizar herramientas complejas, una habilidad que contribuye a su gran adaptabilidad.*

Articulaciones. *Algunas articulaciones, como las rodillas, son débiles y propensas a los accidentes. En la vida urbana, los humanos no las ejercitan lo suficiente. Éste es uno de los puntos más débiles de su diseño.*

Pies. *Como todo el peso del cuerpo descansa sobre los pies, las lesiones y otros problemas de pies son muy comunes, en especial en terrenos accidentados.*

Cerebro. *El cerebro grande y complejo del hombre compensa con creces la falta de adecuación de sus otros sentidos. Gracias a él, puede pensar con lógica, tener el sentido de la historia, una concepción del futuro y la capacidad de reflexionar sobre su lugar en el universo.*

Vista. *El hombre tiene un campo visual de unos 210 grados, de los cuales unos 120 grados se sobreponen, creando la visión estereoscópica. Esto le permite ver los objetos en tres dimensiones y apreciar las distancias. También puede distinguir los colores a la luz del día.*

Dientes. *Los dientes del hombre sirven tanto para cortar carne como para masticar vegetales. Así pues, los dientes están adaptados a una dieta omnívora.*

Corazón. *El corazón mantiene a los demás órganos funcionando al bombearles sangre rica en oxígeno. Sin un ejercicio regular, el corazón pierde su condición y deteriora el funcionamiento del resto del cuerpo.*

Piel. *La piel no sólo protege los órganos internos, sino que también controla la temperatura corporal gracias al sudor. Al haber perdido a lo largo de los siglos la mayor parte de su vello protector, la piel humana debe protegerse del sol, de la lluvia y del frío.*

El animal humano. *El hombre no puede equipararse a ningún otro animal. En la naturaleza, puede convertirse en presa de cualquier depredador, ya que es incapaz de defenderse con zarpas, garras, dientes o picos, o de protegerse con un caparazón, o de esconderse. Carece de la vista y del olfato de otros cazadores, así como de la velocidad suficiente para alcanzar a sus presas. En algunos ambientes, no puede mantener su temperatura corporal sin ropas especiales. Sin embargo, el hombre es un superviviente nato que, debido a sus debilidades físicas, se ha visto forzado a desarrollar su inteligencia y destreza —aptitudes fundamentales para sobrevivir.*

LA ADAPTACIÓN DE ALGUNOS ANIMALES

Lirón. *Pequeño y nervioso, con un metabolismo acorde, extremadamente rápido, el lirón hiberna durante el invierno cuando hay muy poco o ningún alimento disponible. Durante la hibernación, sus ritmos respiratorio y cardíaco disminuyen, y su cuerpo se alimenta de los depósitos de grasa que ha acumulado antes de su largo sueño.*

Ballena. *Aunque la ballena es un mamífero que necesita respirar aire, está bien adaptada a su medio marino: puede retener la respiración hasta 20 minutos para sumergirse en las profundidades en busca de alimento. Sin embargo, el estar tan bien adaptada es también una limitación, ya que es incapaz de escapar de ese medio.*

Musaraña. *Gracias a su pequeño tamaño y a su rapidez, la musaraña tiene una elevada actividad metabólica. Sin embargo, debe comer casi continuamente para permanecer viva.*

Zorro del Sáhara. *Las grandes orejas del zorro del Sáhara pueden girar para localizar con exactitud el menor de los sonidos, permitiéndole cazar de día o de noche. Su gran área superficial le permite disipar el excesivo calor corporal en su medio ambiente natural, el desierto.*

EL EFECTO DEL MEDIO AMBIENTE

La naturaleza no es un jardín del Edén. Incluso las enfermedades simples que un hombre urbano considera normales, como la caries, la falta de higiene o las infecciones ordinarias, son potencialmente mortales. La vida urbana tampoco es mucho mejor, ya que a menudo crea tantos problemas de salud como los que resuelve. Éstos oscilan desde los dolores de cabeza causados por un ruido excesivo hasta el cáncer producido por un envenenamiento por radiación.

- Un bajo contenido de oxígeno en las grandes altitudes puede dificultar la respiración, y agravar los trastornos respiratorios como el enfisema.
- El humo del tabaco y la contaminación del aire pueden provocar cáncer de pulmón y otros trastornos respiratorios como el asma.
- El agua residual y las aguas contaminadas con organismos causantes de enfermedades pueden ser responsables de problemas digestivos.
- En los casos extremos, las sustancias químicas en los alimentos, así como los pesticidas en los productos agrícolas, pueden causar alergias, cáncer, daño a los ovarios, al hígado, a los testículos e hiperactividad.
- Los rayos ultravioleta del sol pueden causar cáncer y envejecimiento de la piel.
- El polvo y el polen en el aire son responsables de alergias como la fiebre del heno y de otros problemas respiratorios como el asma, la fibrosis y la neumonía.
- Los niveles extremos de ruido pueden causar sordera y zumbido, así como dolores de cabeza, estrés e irritabilidad.
- Los solventes industriales en contacto con la piel, o la inhalación de sus vapores, pueden causar reacciones alérgicas o dañar el hígado y los riñones.
- La radiación proveniente de fugas de máquinas o plantas nucleares puede causar daños en los espermatozoides y aumentar el riesgo de nacimiento de niños con leucemia y otros defectos congénitos.
- El estrés, resultado de una acelerada vida urbana, puede causar una serie de problemas físicos y mentales como la colitis y la úlcera gastroduodenal.

CERCANO AL HOMBRE

Además de a los monos, el hombre es fisiológicamente cercano al cerdo. A semejanza del hombre, el cerdo puede alimentarse a base de carne y vegetales, y sus órganos y fisiología general son comparables a los del cuerpo humano, aunque tiene mayor grasa corporal y su piel es más gruesa que la del hombre. En el pasado, los caníbales de las islas del Pacífico hablaban de comer «cerdo largo», en referencia, quizá, a la similitud entre la carne humana y la de cerdo.

Cerdo doméstico. *El cerdo tiene una fisiología muy similar a la del hombre.*

LA ACTITUD MENTAL

Sea cual sea la seriedad de una situación y la capacidad física de sobreponerse a ella de cada individuo, la actitud mental es la clave de la supervivencia.

El hombre es el gran superviviente, y utiliza su cerebro para mantenerse por delante de sus competidores animales. La vida urbana sólo le obliga a utilizar una pequeña parte de sus atributos físicos y mentales. En una situación de supervivencia en la naturaleza, los hombres tienen que aprender de nuevo muchas habilidades manuales.

El sentido común es una de las características humanas más valiosas, pero el deseo de sobrevivir en cualquier situación es mucho más importante.

NECESIDADES NUTRICIONALES

LAS CARACTERÍSTICAS FÍSICAS de los animales están determinadas por los alimentos que comen. Por ejemplo, los herbívoros tienen estómagos múltiples para digerir las resistentes hierbas, así como molares para pastar; los carnívoros tienen dientes para desgarrar carne, y cuerpos adaptados para moverse sigilosamente o desarrollar grandes velocidades. Sin embargo, el hombre es omnívoro, lo que significa que puede comer carne o prescindir de ella, y tiene dientes tanto para desgarrar carne como para triturar vegetales. Si sigue una dieta vegetariana, debe asegurarse de que le proporciona toda la gama de vitaminas y minerales que obtendría de cualquier otro animal.

LOS CIMIENTOS DE LA VIDA

AVELLANAS

NUECES DEL BRASIL

CACAHUETES

PACANAS

NUECES

Minerales. *Los minerales son necesarios para una gran variedad de funciones, que van desde el control del metabolismo hasta la osificación y la contribución al cierre de las heridas. Una deficiencia de minerales puede ocasionar calambres y convulsiones. Si no come carne, es vital que siga una dieta variada que incluya vegetales verdes, nueces y fruta para obtener todos los minerales, principalmente el hierro. Las setas también contienen minerales, así como las algas.*

SETAS

Grasas. *Las grasas son una valiosa fuente de energía y son necesarias para generar y reparar las células del cuerpo. También contribuyen a la sensación de saciedad después de comer. Recuerde que, a fin de ser digeridas, las grasas requieren una gran cantidad de agua, y sólo deben consumirse cuando hay abundante agua disponible. Las nueces son ricas en grasas y en proteínas.*

Fibra. *La fibra facilita el funcionamiento correcto del intestino y la evacuación de residuos. Se obtiene de las frutas, las verduras y el pan.*

GROSELLAS

KIWI

DÁTILES

CAQUI O PALO SANTO

PAPAYA

PANAL

Hidratos de carbono. *Los hidratos de carbono desempeñan un papel vital en la nutrición al aportar energía al cuerpo. Se encuentran en forma de azúcares o almidones que se degradan y almacenan en el hígado hasta que son necesarios para liberar energía en forma de glucosa. El exceso de glucosa que no es requerido para aportar se convierte en grasa. La miel es excelente fuente de azúcar.*

Vitaminas. *Las vitaminas controlan el crecimiento y el mantenimiento de los tejidos del cuerpo, y estimulan la producción de energía. Las frutas frescas son la mejor fuente de vitaminas. Durante las situaciones de supervivencia en regiones templadas, la fruta debe recolectarse y almacenarse para el invierno.*

PESCADO

Proteínas. *Las proteínas son imprescindibles en la formación de tejidos y para mantener el crecimiento. Algunas proteínas deben obtenerse de los alimentos, ya que proporcionan aminoácidos esenciales que nuestro cuerpo no puede sintetizar. La carne y el pescado proporcionan todos los aminoácidos necesarios y, adicionalmente, vitaminas y minerales. Algunas proteínas pueden obtenerse de productos vegetales.*

CALORÍAS NECESARIAS AL DÍA

La energía obtenida de los alimentos se mide en calorías. El hombre necesita un determinado número de calorías para subsistir. Incluso si permaneciera en la cama todo el día, quemaría cerca de 2.000 calorías. El número de calorías necesarias diariamente depende de varios factores, como la edad. Los supervivientes deben intentar equilibrar la entrada de calorías con el consumo de energía.

Leñador. *Una actividad física pesada, como la tala, requiere un mayor aporte diario de calorías que un estilo de vida sedentario, ya que se necesita más energía para mantener el cuerpo en movimiento y los músculos en uso. En las explotaciones forestales, estas calorías se obtienen gracias a comidas ricas y abundantes.*

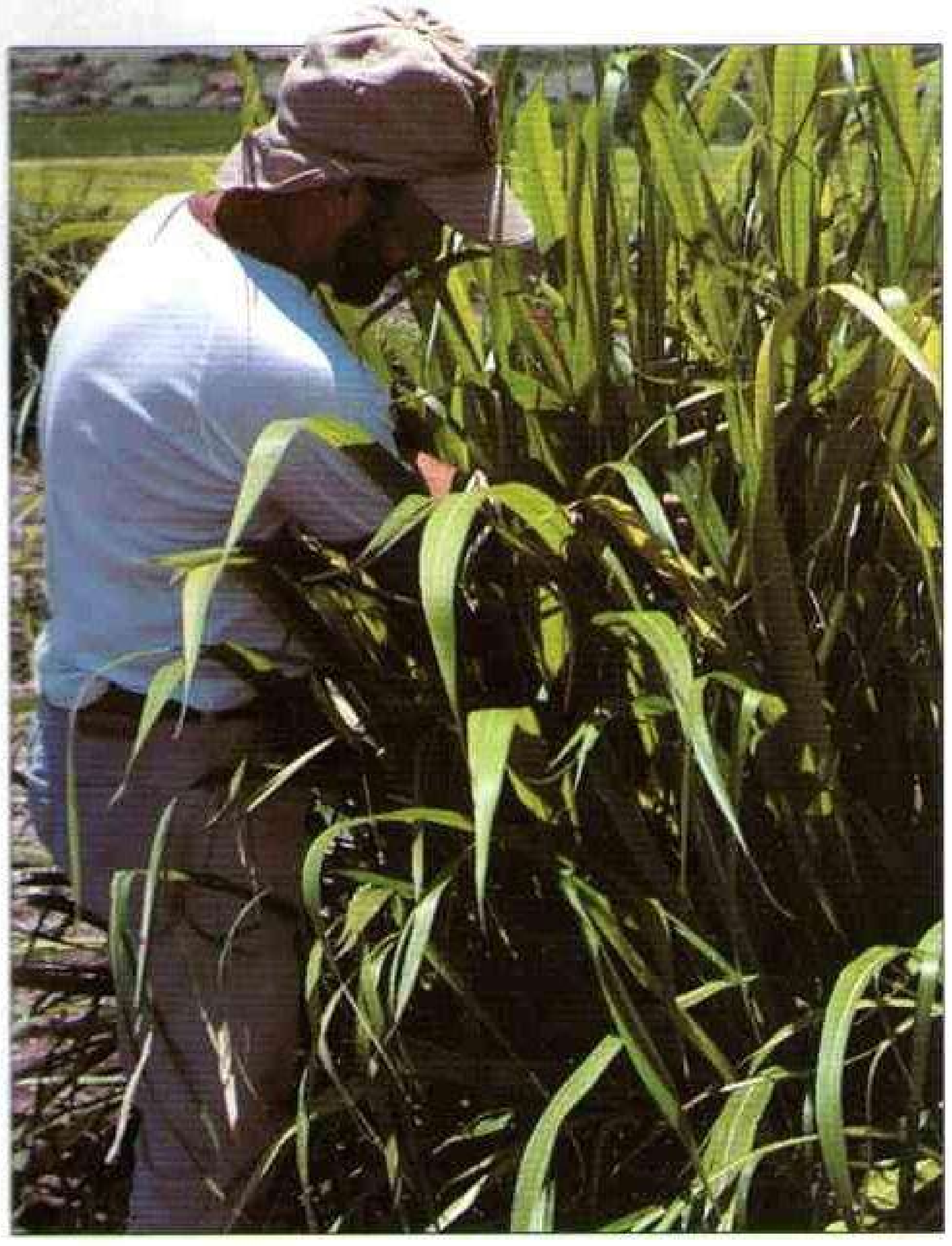

Científico tropical. *Las personas que viven en los trópicos requieren menos calorías que aquellas que viven en climas fríos. Las altas temperaturas provocan un sudor continuo, que produce una pérdida de minerales y sales, por lo que la dieta debe estar equilibrada.*

Alpinista. *Las constantes temperaturas de congelación y la necesidad de estar alerta continuamente, convierten al montañismo en una actividad de gran consumo calórico. Son necesarias más de 5.000 calorías por día para que el cuerpo mantenga su temperatura interna. Pueden requerirse calorías adicionales para otras actividades como tirar de un trineo.*

CONSUMO DE ENERGÍA

Cada actividad requiere un distinto número de calorías, según el clima, la condición física y la cantidad de esfuerzo necesario para cada tarea en particular. Las siguientes cifras son una guía general para un hombre y una mujer «de actividad media». Debe recordar que sus necesidades calóricas particulares, y la cantidad de energía que requiere para cada actividad variarán según el clima y el tipo de terreno en el que trabaje.

Actividad	**Calorías por hora**
Descansar	70
Estar sentado	90
Encender un fuego	135
Caminar	180
Ir en bicicleta	240
Cortar madera	360
Esprintar	360
Correr	400
Nadar	500

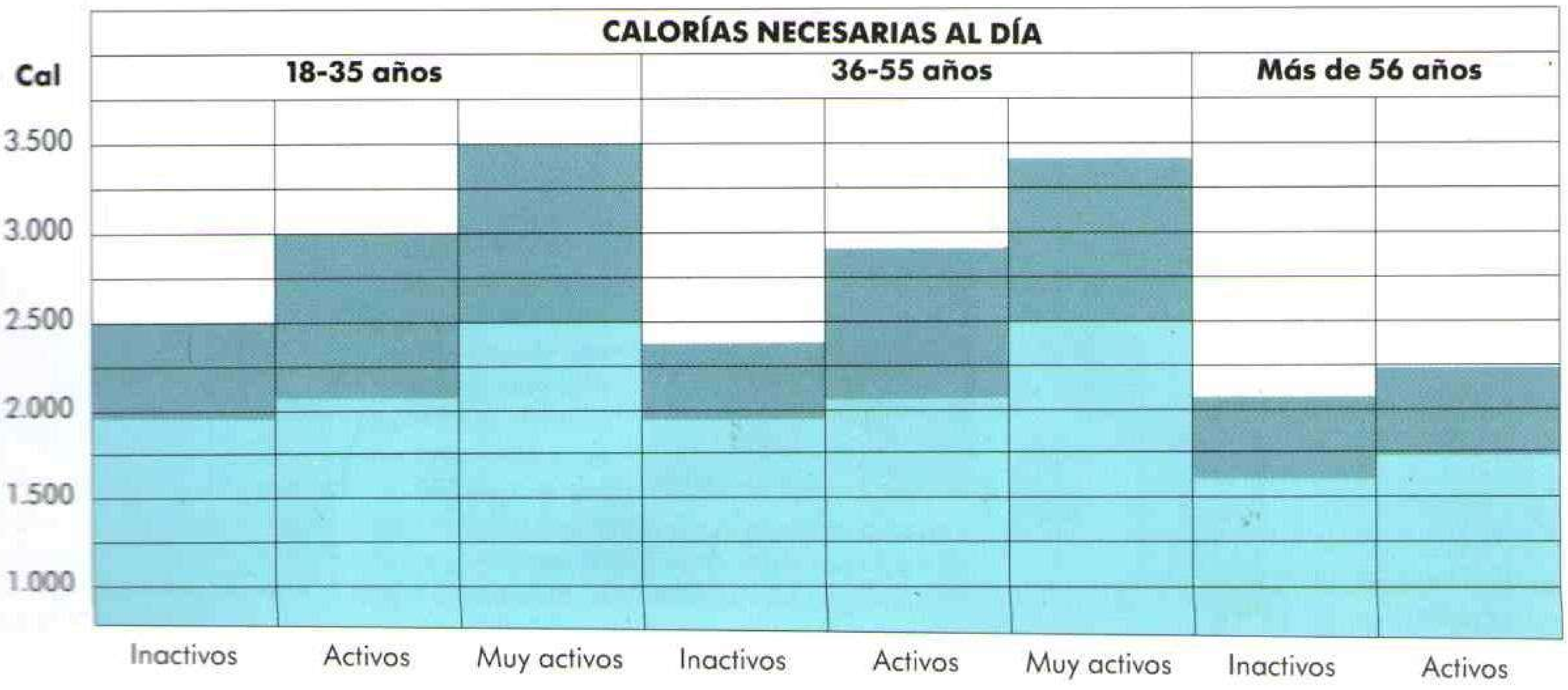

Necesidades calóricas. *El metabolismo de los hombres es menos eficiente que el de las mujeres; los hombres tienen menos aislantes naturales, por lo cual queman más calorías. Las personas jóvenes queman más calorías que las mayores, en parte porque son más activas, pero también porque la madurez tiende a disminuir el ritmo metabólico.*

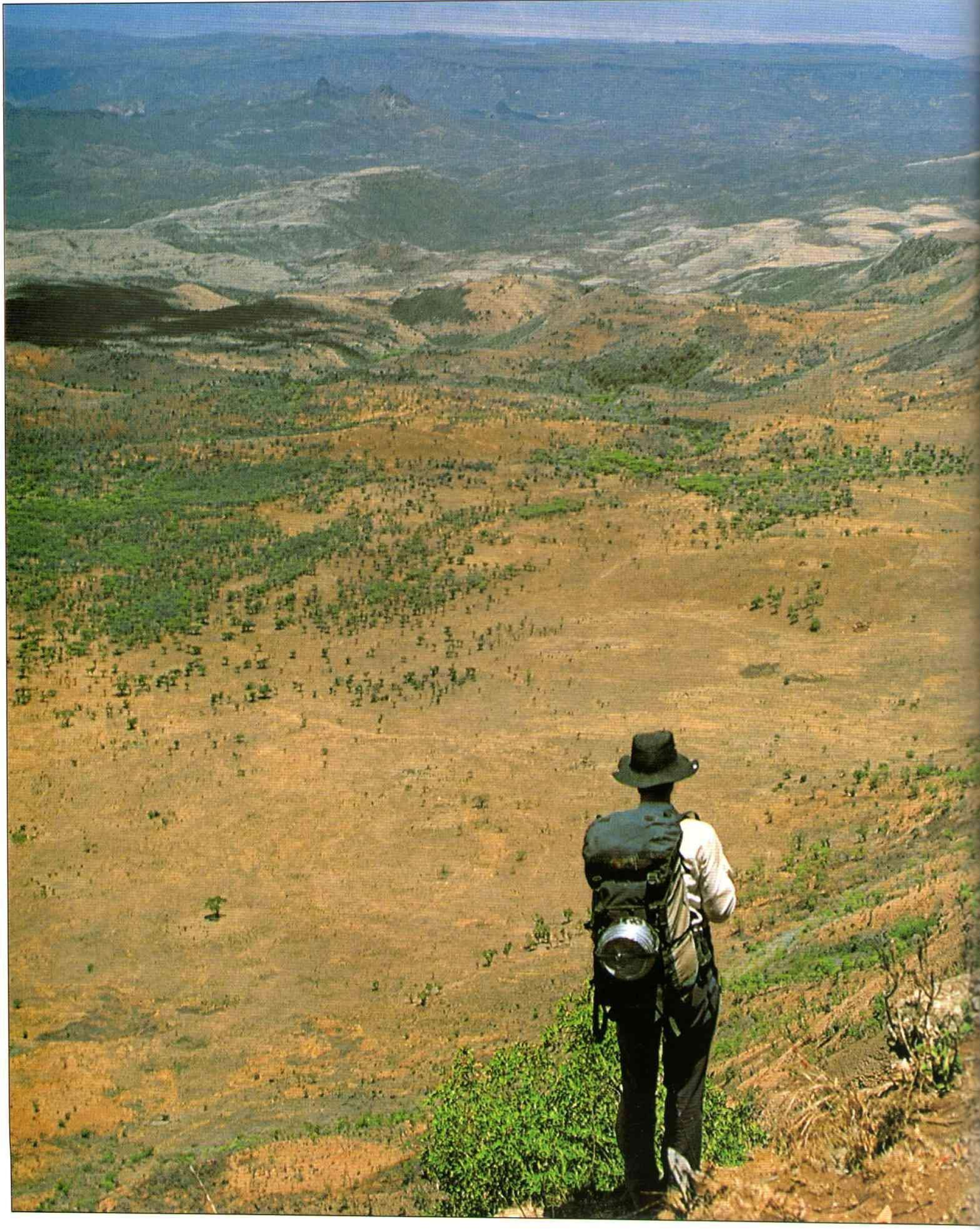

CAPÍTULO DOS

EQUIPAMIENTO BÁSICO

CON EL EQUIPAMIENTO correcto para cada actividad y medio particular reducirá drásticamente la posibilidad de sufrir un accidente y el tener que enfrentarse a una seria situación de supervivencia.

Los proveedores de equipamiento deben proporcionarle el material adecuado para cualquier tipo de clima y terreno. Puede gastar gran cantidad de dinero comprando artículos muy vistosos, pero es más importante dedicarse a encontrar el equipaje que realmente necesita para sobrevivir. Sin una experiencia propia, debería averiguar lo que otras personas, con más experiencia, han utilizado en situaciones similares, y no fiarse de la brillante charlatanería de los vendedores. Algunos artículos del equipo, como, por ejemplo, el cuchillo y la brújula, son esenciales, por lo que debería comprar los mejores que encuentre. Otros artículos, como la almohada hinchable o una colchoneta, sólo mejorarán su comodidad, por consiguiente, puede prescindir de ellos. Una vez haya seleccionado su material, debe prepararse para modificarlo según sus necesidades precisas. Piense en las piezas que podrían desgastarse o romperse, y en cómo asegurar las herramientas y la ropa –en especial los guantes y sombreros– para no perderlos.

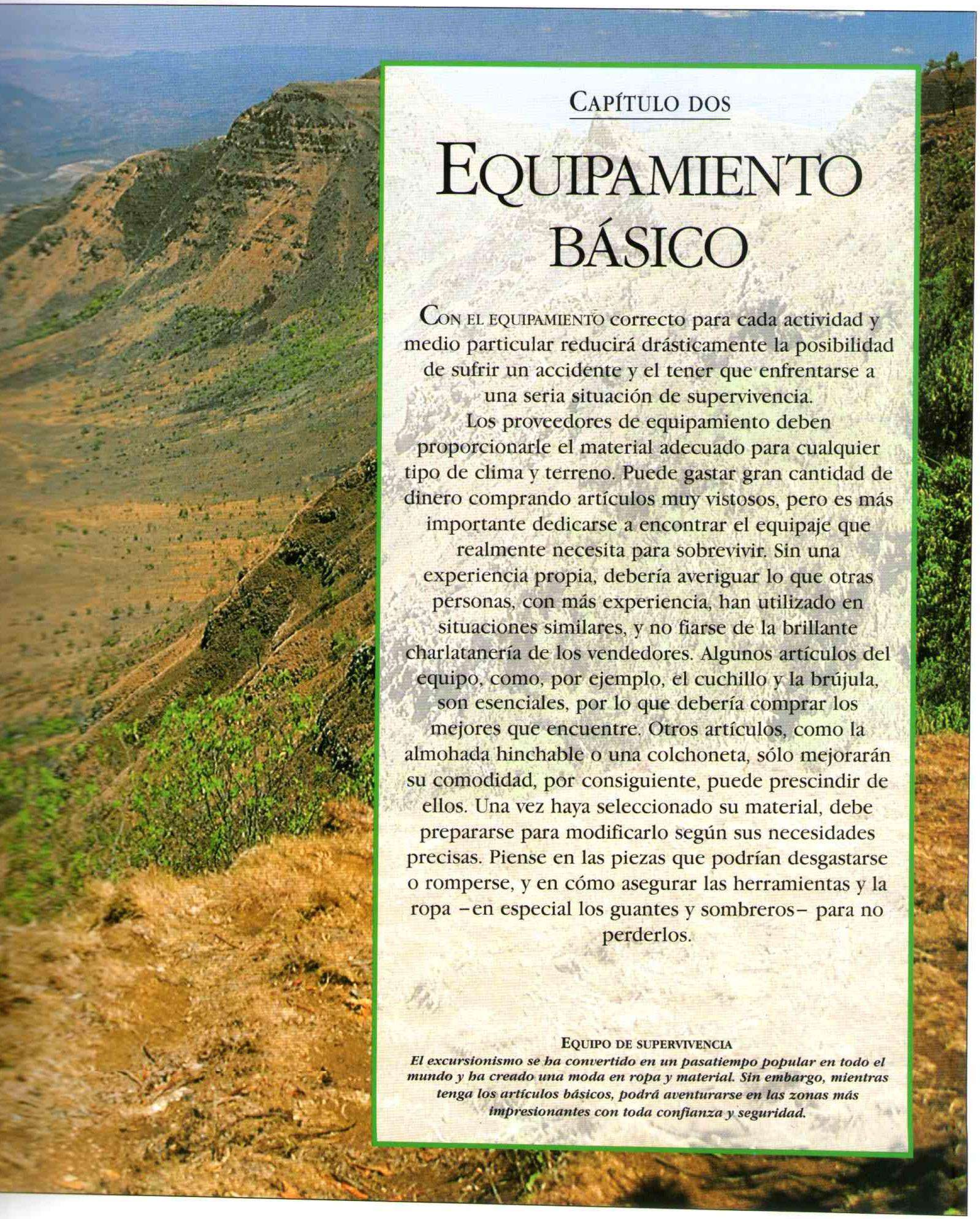

EQUIPO DE SUPERVIVENCIA

El excursionismo se ha convertido en un pasatiempo popular en todo el mundo y ha creado una moda en ropa y material. Sin embargo, mientras tenga los artículos básicos, podrá aventurarse en las zonas más impresionantes con toda confianza y seguridad.

LA ROPA ADECUADA

RARA VEZ LLEVAMOS la ropa apropiada a las condiciones naturales en las que vivimos. La vida urbana nos acostumbra a llevar ropa para estar cómodos y a la moda, y no para mantener la temperatura corporal, y además nos aísla del medio con la calefacción artificial y el aire acondicionado. Incluso si compramos ropa diseñada para la vida exterior, ésta es más moderna que práctica. Puede gastar mucho dinero en conjuntos «de diseño» o comprar los artículos más esenciales, que cumplirán bien su función, aunque no sean del color de última moda.

EL PRINCIPIO DEL SISTEMA DE CAPAS

Un cierto número de capas de tejido mantendrán mejor su temperatura corporal que una sola prenda gruesa. Si tiene demasiado calor, puede controlar su temperatura corporal eliminando capas o «aireándolas» (abriendo cremalleras o botones para dejar escapar el aire caliente y permitir la entrada de aire fresco). Este principio se aplica tanto en los clímas cálidos como en los fríos.

La capa interior. *La primera capa, en contacto con la piel, debe consistir en una camiseta de algodón o una camiseta térmica de manga larga, que se ajuste bien al cuerpo, pero sin que lo apriete. Debe estar hecha de un material que absorba la transpiración y que no se pegue a la piel (llevándola fuera del material). Esta capa debe mantenerse lo más limpia posible para evitar una acumulación de suciedad que llegaría a tapar sus poros.*

La segunda capa. *La segunda capa debe ir más suelta, pero con la posibilidad de proteger y mantener el calor en el cuello y las muñecas. Puede consistir en una camiseta con cuello tipo polo y cremallera, o en una camisa con cuello, mangas que puedan remangarse y puños con botones. En un clima cálido, esta capa puede ser la exterior, si bien pueda requerir una protección para el viento* (véase la página siguiente).

Calzoncillos. *Únicamente requerirá ropa interior larga y térmica en zonas con temperaturas bajo cero. En las regiones árticas, un refuerzo impermeable en la ingle protegerá del viento esa parte del cuerpo, y le será útil si tiene que esquiar. Un pantalón mojado se secará más deprisa si no lleva ropa interior larga, aunque puede combatir este problema llevando pantalones impermeables. En climas templados, use unos calzoncillos cortos de algodón.*

La tercera capa. *La tercera capa debería consistir en un jersey de lana o una chaqueta ligera de fibra polar. Si está en movimiento, incluso en el ártico, ésta es la capa que debe quitarse para evitar un acaloramiento excesivo. Puede airear su plumífero* (véase página siguiente) *si aún tiene calor. Cuando se detenga a descansar, vuelva a ponerse la capa intermedia antes de empezar a sentir frío. Esta capa puede sustituir a la exterior en regiones templadas cuando el clima es suave.*

La capa exterior. *La capa exterior debe ser una chaqueta impermeable o que le proteja del viento, o que cumpla ambas funciones, dependiendo del clima en el que se desenvuelva. En las regiones árticas requerirá un anorak acolchado que le proteja de los fuertes vientos. Debe tener la posibilidad de airear la chaqueta para evitar el exceso de calor. En áreas templadas el principal problema es la lluvia; puede llevar entonces una prenda impermeable sobre la chaqueta.*

Pantalones. *Los pantalones deben permitir la libertad de movimientos y deben estar hechos de un material que se seque rápidamente cuando se ha mojado. En condiciones muy húmedas, puede ponerse pantalones impermeables para proteger sus piernas de una lluvia intensa* (véase página siguiente), *aunque pueden causarle un calor excesivo. En condiciones de frío intenso deben utilizarse pantalones impermeables acolchados por encima de los pantalones y las botas, como protección adicional.*

CARACTERÍSTICAS DE LOS MATERIALES TEXTILES

Material	Ventajas	Inconvenientes
Lana	Esta fibra natural tiene propiedades aislantes incluso cuando está húmeda, manteniéndole cómodo a menos que esté totalmente mojada. Arde sin llama cuando se expone a una llama.	La lana es pesada cuando está mojada y tarda en secarse. Puede dar sensación de picor en contacto directo con la piel. Puede encogerse al lavar.
Algodón	El algodón es resistente, permite la «respiración» y absorbe la humedad. Es un material especialmente indicado para la ropa interior y para todo tipo de ropa que deba llevarse en contacto con la piel.	Es pesado cuando está mojado y puede encogerse si se seca a altas temperaturas. Puede rasgarse con facilidad. No corta el viento y arde fácilmente.
Fibra polar	Se utiliza como tercera capa para alejar la humedad del cuerpo al mismo tiempo que lo mantiene caliente. Pesa poco y es resistente. No absorbe la humedad.	Este tejido sintético no corta el viento. No se comprime fácilmente y puede producir «bolitas» en la parte exterior después de un uso prolongado.
Mat. sintéticos y transpirables	Permiten que el sudor se evapore al mismo tiempo que aíslan de la lluvia. Generalmente protegen del viento, y se utilizan para las prendas exteriores.	Las costuras en estos tejidos permiten el paso del agua. En condiciones muy húmedas, los poros del tejido pueden taparse, y en climas muy fríos la humedad puede condensarse en el interior. La evaporación del sudor puede dar como resultado una pérdida de calor.

La elección del material. *La ropa puede fabricarse con una gran variedad de materiales, desde la lana o el algodón hasta los tejidos sintéticos, algunos de los cuales permiten la transpiración, mientras que aíslan de la lluvia. Los materiales sintéticos no siempre son tan resistentes como los naturales. Las fibras naturales como el algodón y la lana pueden encogerse si se secan a temperaturas elevadas después de lavadas.*

ROPA PARA CLIMAS EXTREMOS

Si al elegir su ropa sigue el sistema de capas *(véase página anterior)* debe ser capaz de equiparse con las prendas adecuadas para cualquier tipo de clima. En una situación de supervivencia, puede que no disponga de las prendas fabricadas para un clima específico, pero puede utilizar el sistema de capas en combinación con refugios y fuegos para protegerse de los elementos adversos. En los climas extremos y en terreno hostil, la ropa debe protegerle de aquellas circunstancias que puedan dañarle o causarle la muerte —por ejemplo las que se encuentran en los desiertos *(véase página 142)*. El viento, el frío, el calor, el sol y la lluvia constituyen la principal amenaza. Algunas prendas que son específicas para un clima determinado —por ejemplo las botas para la selva— no suelen ser adecuadas para áreas menos extremas, pero, en general, la mayor parte de las prendas pueden adaptarse a sus necesidades.

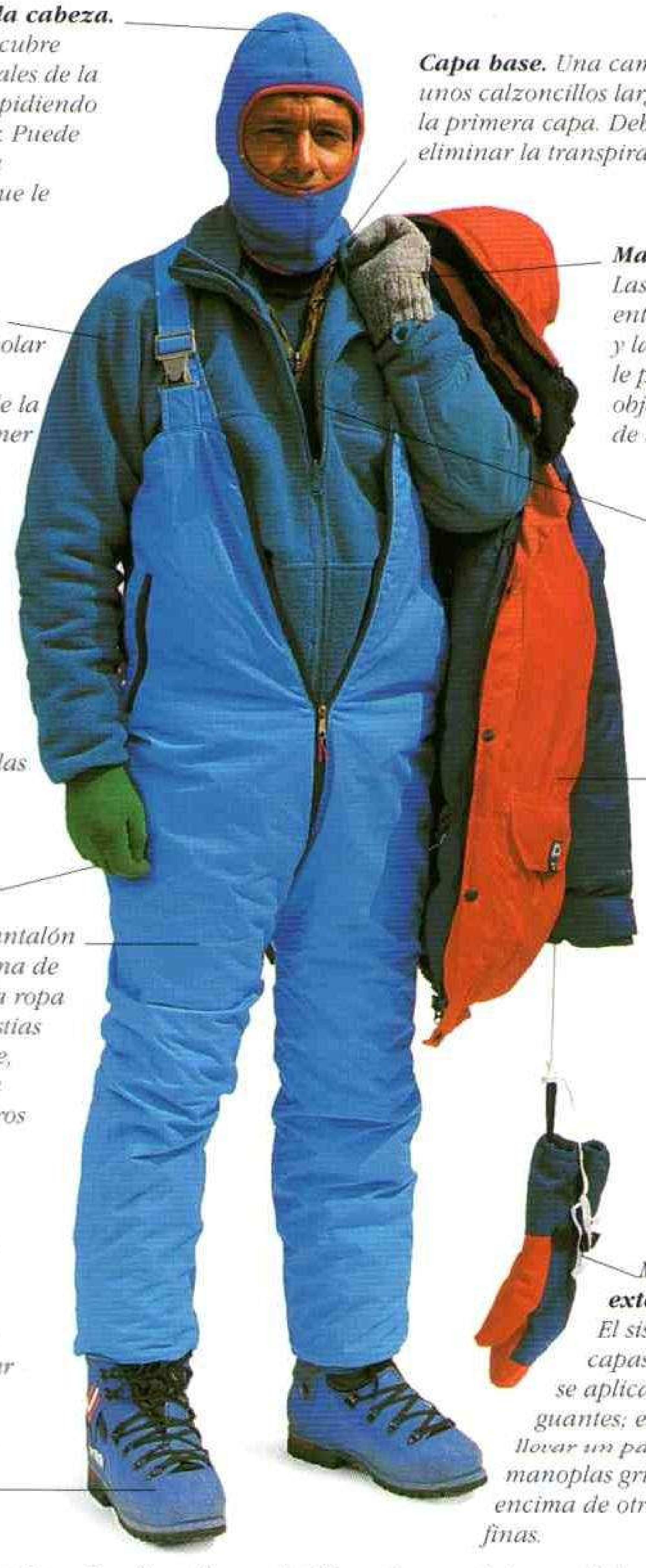

Protección para la cabeza. *Un pasamontañas cubre la cabeza, los laterales de la cara y el cuello, impidiendo la pérdida de calor. Puede llevar también una capucha aislante que le proporcione calor adicional.*

Capa intermedia. *Un jersey de fibra polar o lana debería absorber el sudor de la capa base y mantener una capa de aire caliente alrededor de su cuerpo.*

Guantes interiores. *Evitan que la piel se adhiera a los objetos congelados, y permiten el uso de las manos durante períodos cortos sin necesidad de usar manoplas gruesas.*

Pantalones. *Un pantalón de peto cubre la zona de la cintura, donde la ropa puede causar molestias al caminar; permite, al mismo tiempo, la aireación en bombros y pecho.*

Botas. *Las botas de nieve tienen una coraza plástica y forros térmicos que actúan como un par de botas internas. Deben llevarse dos pares de calcetines debajo de ellas.*

Capa base. *Una camiseta térmica y unos calzoncillos largos constituyen la primera capa. Deben absorber y eliminar la transpiración.*

Manopla intermedia. *Las manoplas de lana entre los guantes interiores y las manoplas exteriores le permiten coger los objetos sin perder el calor de las manos.*

Segunda capa. *Una camisa con cuello tipo polo debe cubrir el pasamontañas a la altura del cuello, y los guantes a la altura de los puños.*

Capa exterior. *Un anorak acolchado y con capucha es esencial. Debe tener un cierre frontal solapado y una capa impermeable en tejido transpirable.*

Manoplas externas. *El sistema de capas también se aplica a los guantes; es aconsejable llevar un par de manoplas gruesas encima de otras más finas.*

Mantenerse en calor. *En situaciones de frío extremo, deben cubrirse todas las partes del cuerpo expuestas; resulta vital el seguimiento del sistema de capas. Prepárese a airear o a quitarse alguna prenda si siente calor excesivo al moverse o trabajar. El sudor atrapado reduce las propiedades aislantes de la ropa y puede causar bipotermia* (véase página 163).

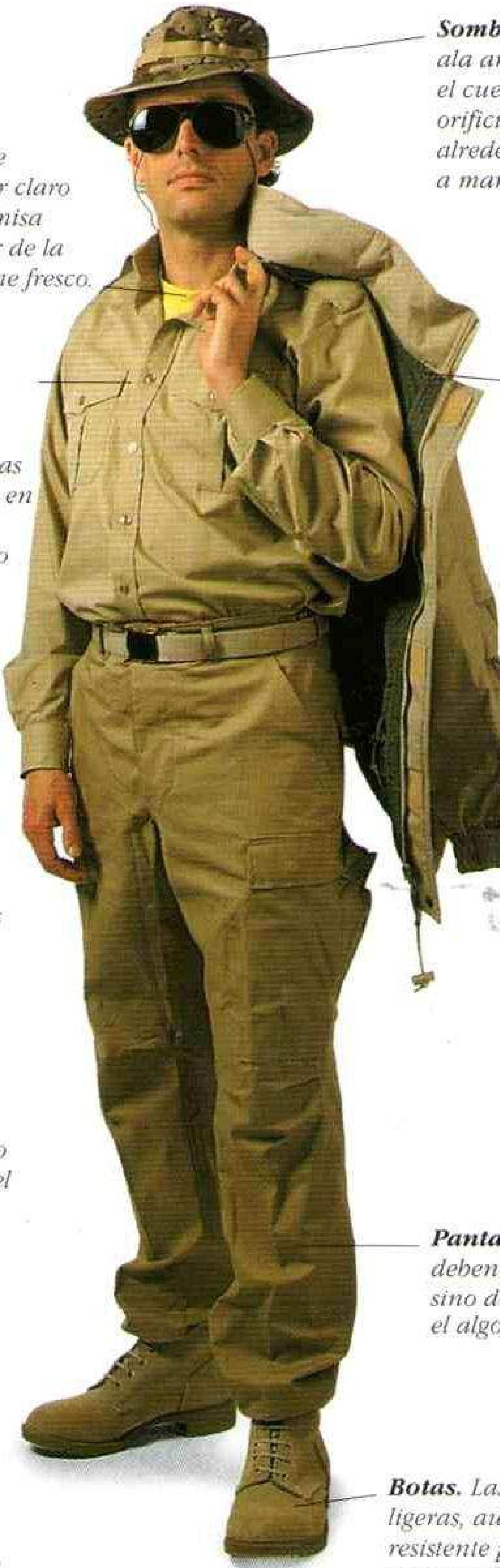

Capa base. *Una camiseta de algodón de color claro debajo de la camisa absorbe el sudor de la piel y le mantiene fresco.*

Segunda capa. *Una camisa ligera actúa como capa principal. Las mangas deben llevarse bajas en situaciones de calor extremo, por ejemplo en el desierto* (véase página 142).

Sombrero. *Un sombrero de ala ancha protege la cabeza y el cuello del sol. Los pequeños orificios de ventilación alrededor del ala contribuyen a mantener la cabeza fresca.*

Capa exterior. *Una chaqueta ligera y aislante sobre la camisa actuará como capa final, protegiéndole del viento y del frío nocturno.*

Mantenerse fresco. *La ropa para el clima cálido debe ser ligera y suelta. Los colores claros reflejan el calor y contribuyen a mantenerle fresco. El principio de las capas también se aplica aquí, de la misma manera que en el clima frío.*

Pantalones. *Los pantalones no deben ser ajustados ni pesados, sino de un material ligero, como el algodón.*

Botas. *Las botas deben ser ligeras, aunque con una suela resistente para caminar* (véase página 24).

MANTENERSE SECO

Estar mojado constituye un grave problema de supervivencia. Debe ponerse el traje impermeable en el momento que empiece a llover, y quitárselo tan pronto cese la lluvia. El equipo de lluvia a base de materiales impermeables no es adecuado para los excursionistas ya que produce sudor —elija únicamente materiales transpirables. Utilice pantalones impermeables sólo cuando llueva y haga frío a la vez. Unas polainas mantendrán secos sus pies y pantorrillas, incluso al cruzar ríos.

Equipo para tiempo lluvioso. *El equipo impermeable debe cubrir todo el cuerpo, desde la cabeza hasta los pies.*

CALZADO

EL PESO DEL cuerpo humano está calculado para ser soportado por cuatro pies en lugar de dos, así que los pies están sometidos a una presión constante. Con el peso añadido que suponen las mochilas cargadas, cruzar un territorio escarpado o húmedo hace que la protección de los pies se convierta en la parte del equipo más importante a considerar. El calzado debe haber cedido antes de ser usado. Para ello, humedezca las botas de cuero y llévelas puestas hasta que se sequen. Una vez en camino, mantenga los pies limpios, lávelos cada día y aplíqueles polvos de talco. Use siempre calcetines limpios y secos, y cubra cualquier rozadura en los pies antes de que se conviertan en ampollas.

CALZADO LIGERO

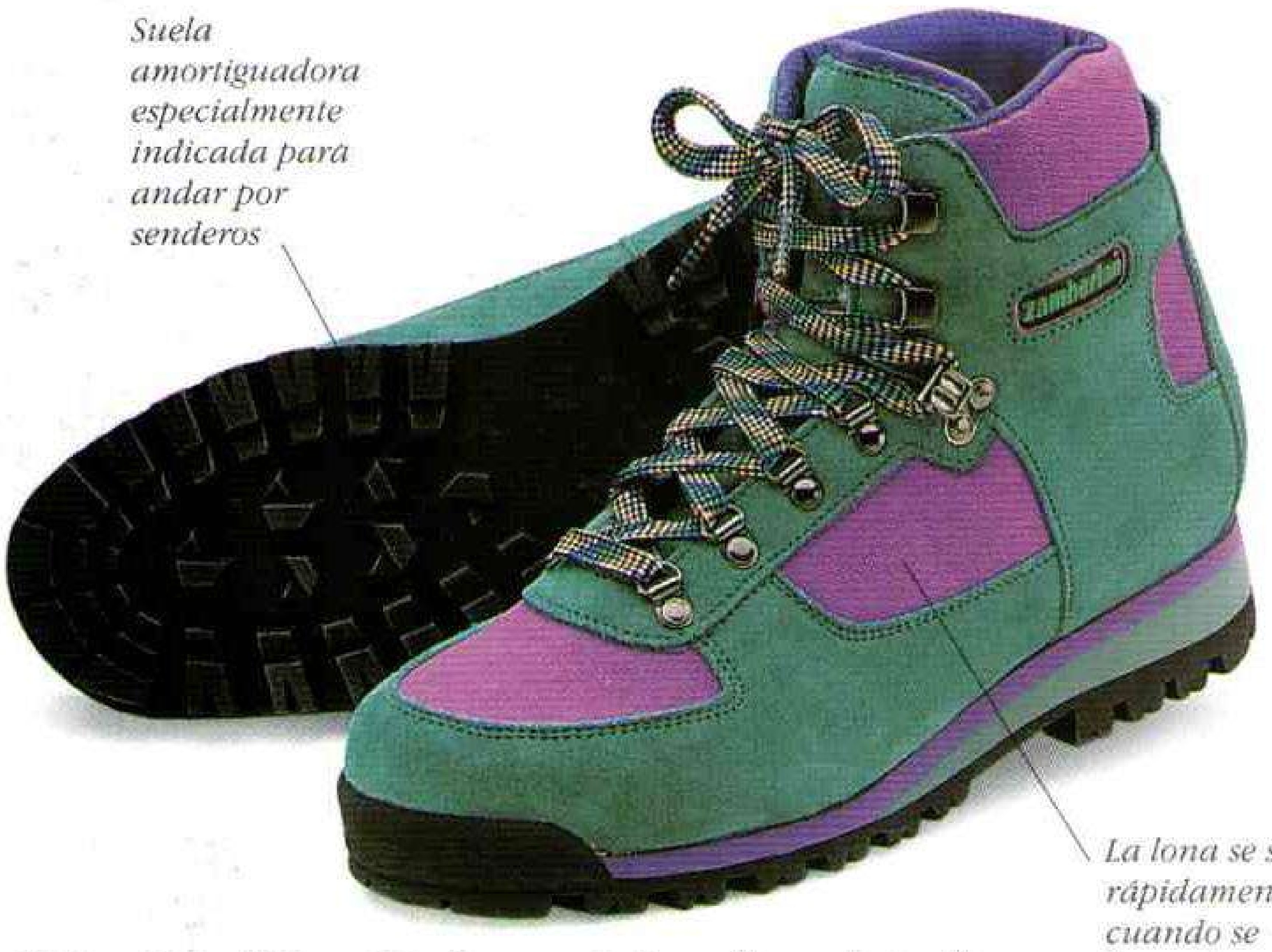

Suela amortiguadora especialmente indicada para andar por senderos

La lona se seca rápidamente cuando se humedece

***Botas de* trekking.** *Muchos montañeros llevan botas ligeras de marcha tanto en el campamento como durante la ascensión a una montaña elevada, y reservan sus pesadas botas de escalada para la nieve y el hielo. Las botas ligeras también pueden utilizarse para las caminatas sencillas y en distancias cortas, siempre y cuando el terreno no sea muy escarpado. Aunque ofrecen menos protección y apoyo que las botas de piel, las botas de marcha se secan rápidamente y dan a los pies una sensación de descanso después de llevar un calzado caluroso y pesado.*

Botas para el desierto. *Con suelas resistentes y fabricadas en cuero ligero, las botas para el desierto permiten que los pies respiren al mismo tiempo que evitan la entrada de arena caliente. Los laterales altos protegen los tobillos de la vegetación espinosa y les proporcionan apoyo. En condiciones húmedas, el cuero se vuelve todavía más poroso por lo que necesitan un tiempo prolongado para secarse.*

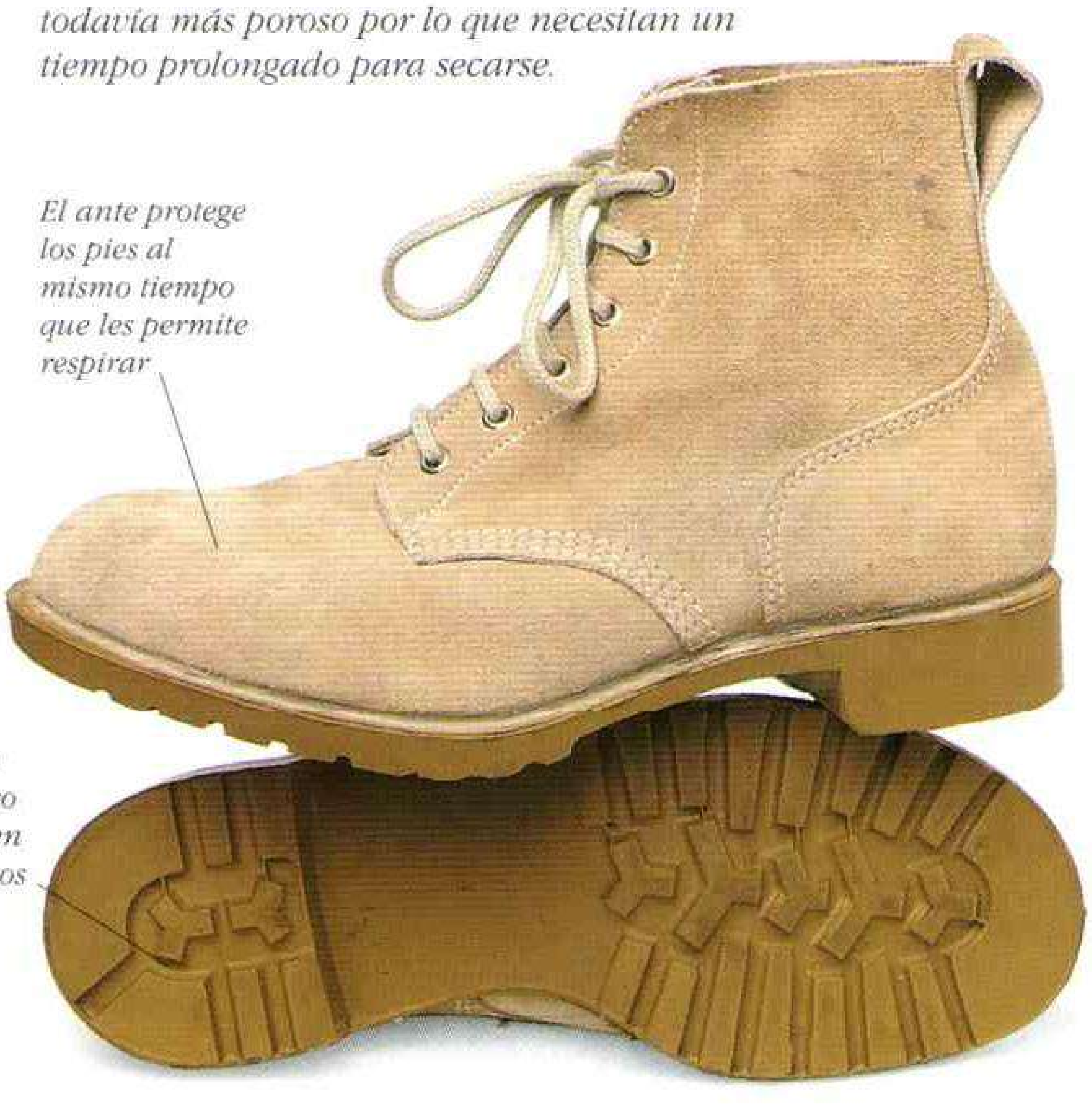

El ante protege los pies al mismo tiempo que les permite respirar

Las suelas resistentes son un buen apoyo para los pies en terrenos rocosos

Suela de espuma moldeada que sirve de apoyo y de amortiguador

Zapatillas. *Una vez en el campamento, las zapatillas alivian los pies doloridos, y son excelentes para llevar en barcos o vehículos. Sin embargo, al ser cómodas y ligeras, no ofrecen tanta protección como las resistentes botas de lona. El uso constante de zapatillas puede restar resistencia a los pies y hacerlos más susceptibles a las heridas. Al caminar sobre un terreno húmedo pueden llegar a quedar absorbidas y perderlas.*

CALCETINES

Los calcetines gruesos, ya sean de lana o algodón, son indispensables. Los calcetines protegen el pie y evitan que las botas rocen la piel, al mismo tiempo que mantienen los pies cálidos y secos eliminando la humedad (trasladan la humedad hacia la capa exterior de las fibras, alejándola de la piel). En climas muy fríos es aconsejable llevar dos pares de calcetines.

CALCETINES INTERIORES

CALCETINES EXTERIORES

CALZADO RESISTENTE

Botas de montaña. *Esta bota es un término medio entre peso, durabilidad y protección; indicada para todo tipo de actividades gracias a su suela gruesa y grabada, material impermeable y protección para los tobillos.*

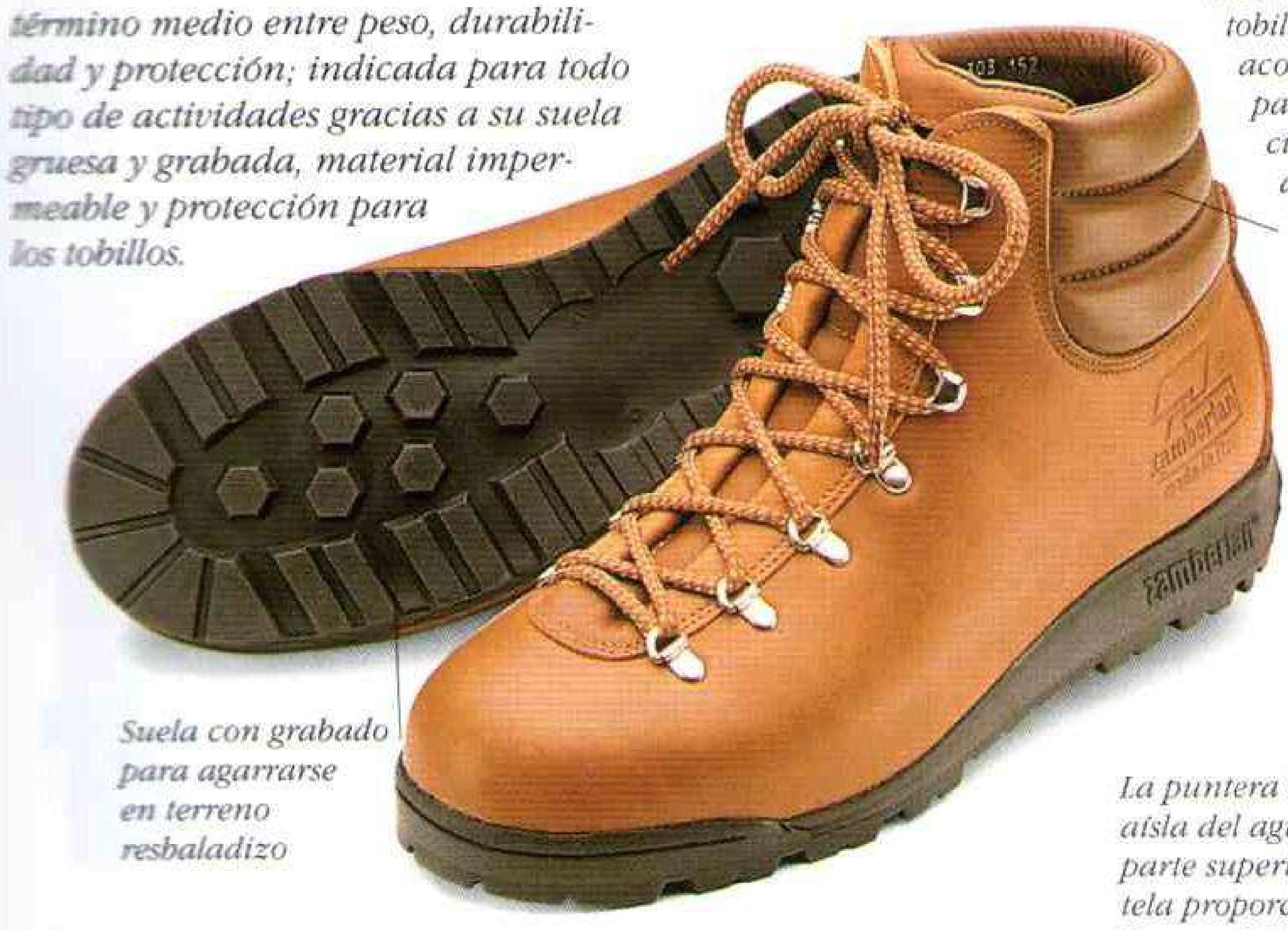

La zona de los tobillos está acolchada para proporcionar un apoyo cómodo

Suela con grabado para agarrarse en terreno resbaladizo

Botas de jungla. *En la jungla se hace necesario proteger el tobillo de ramas y rocas. Con suela de goma, estas botas evitan la absorción de agua, al tiempo que permiten que los pies respiren a través de la lona de la que están hechas. Después de vadear un río, el agua es expulsada a través de ojetes especiales en el empeine. El dibujo de la suela, ancho y profundo, proporciona un buen agarre al suelo húmedo.*

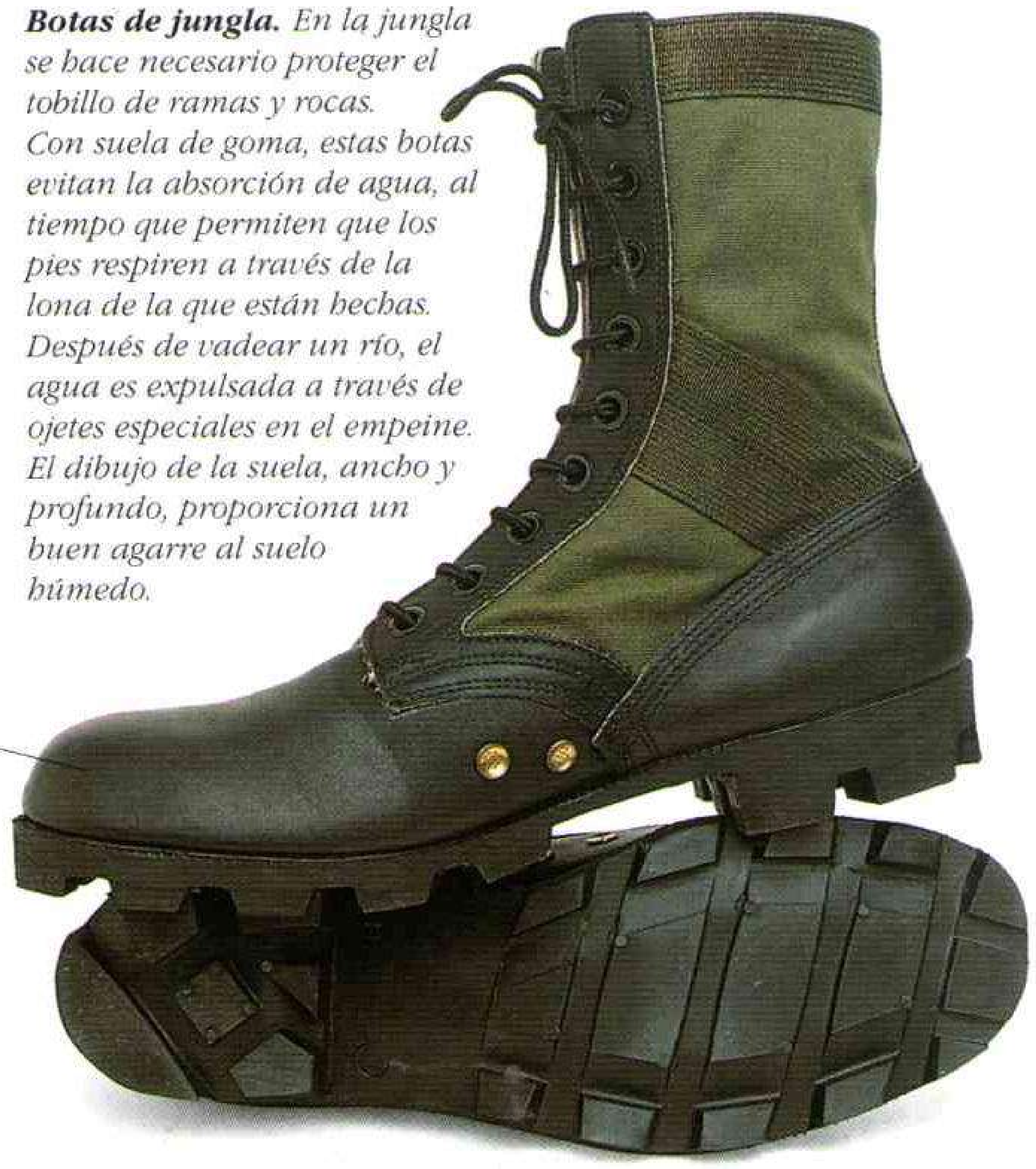

La puntera de goma aísla del agua, y la parte superior de tela proporciona frescor al pie

Botas de plástico para nieve. *Diseñadas para mantener el pie rígido al utilizar garfios para un mejor agarre al hielo, estas botas son especialmente aislantes para ser utilizadas en los climas más fríos. Sin embargo, al igual que sucede con las botas para esquiar, caminar con ellas resulta incómodo. Tienen un forro interno térmico con la forma de la bota, que puede llevarse independientemente, por ejemplo dentro de una tienda de campaña.*

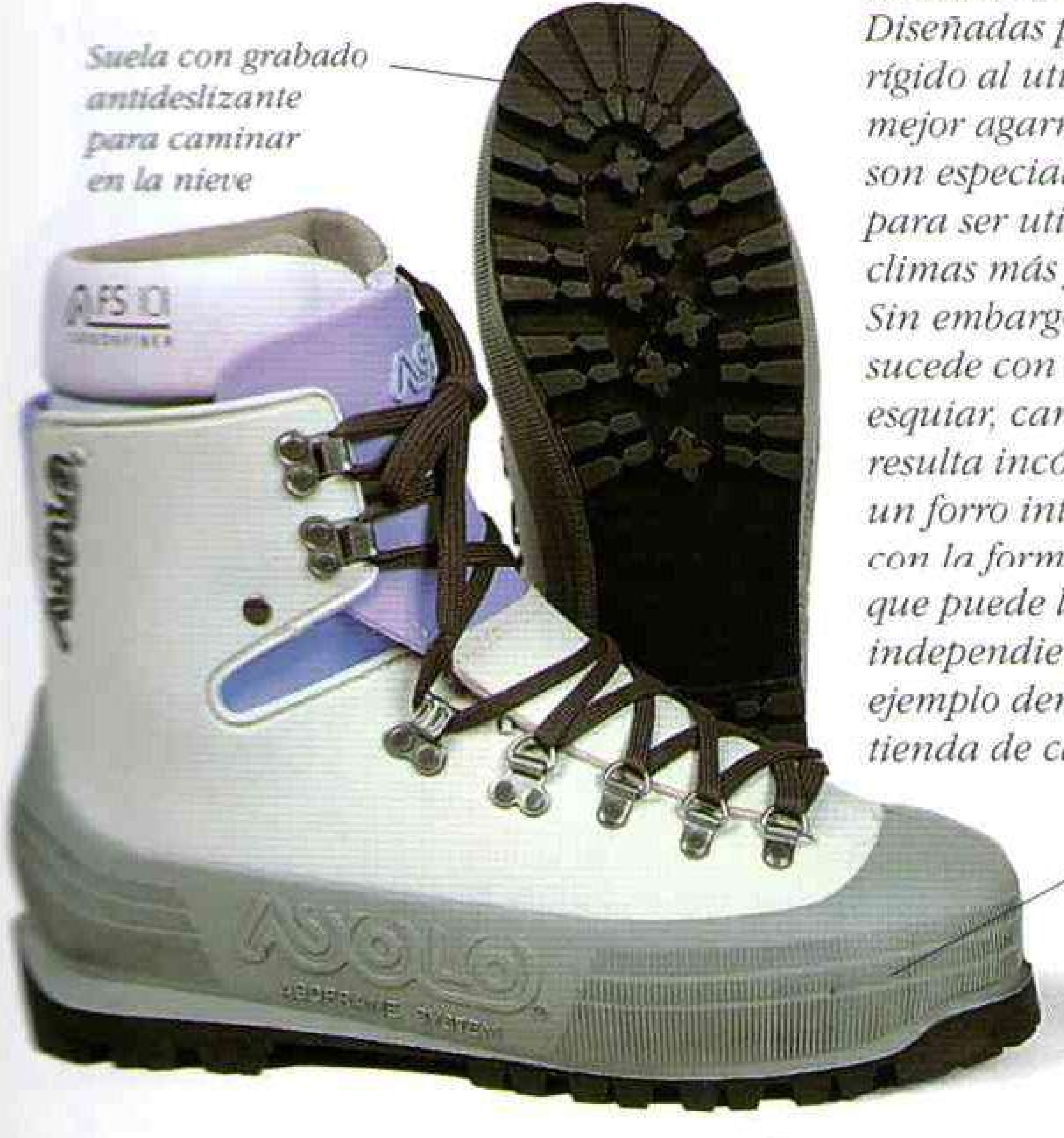

Suela con grabado antideslizante para caminar en la nieve

Los laterales rígidos proporcionan un apoyo excelente en terrenos rocosos

POLAINAS

Las polainas resultan indispensables, tanto en las zonas templadas para mantener secos los pies y los bajos del pantalón como en las regiones polares para evitar el contacto con la nieve. Al cruzar ríos poco profundos, las polainas impedirán que el agua entre en las botas. Son mejores que las botas de goma, ya que no provocan calor ni una transpiración excesiva del pie.

POLAINAS PARA BARRO

POLAINAS PARA NIEVE

CUIDE SUS BOTAS

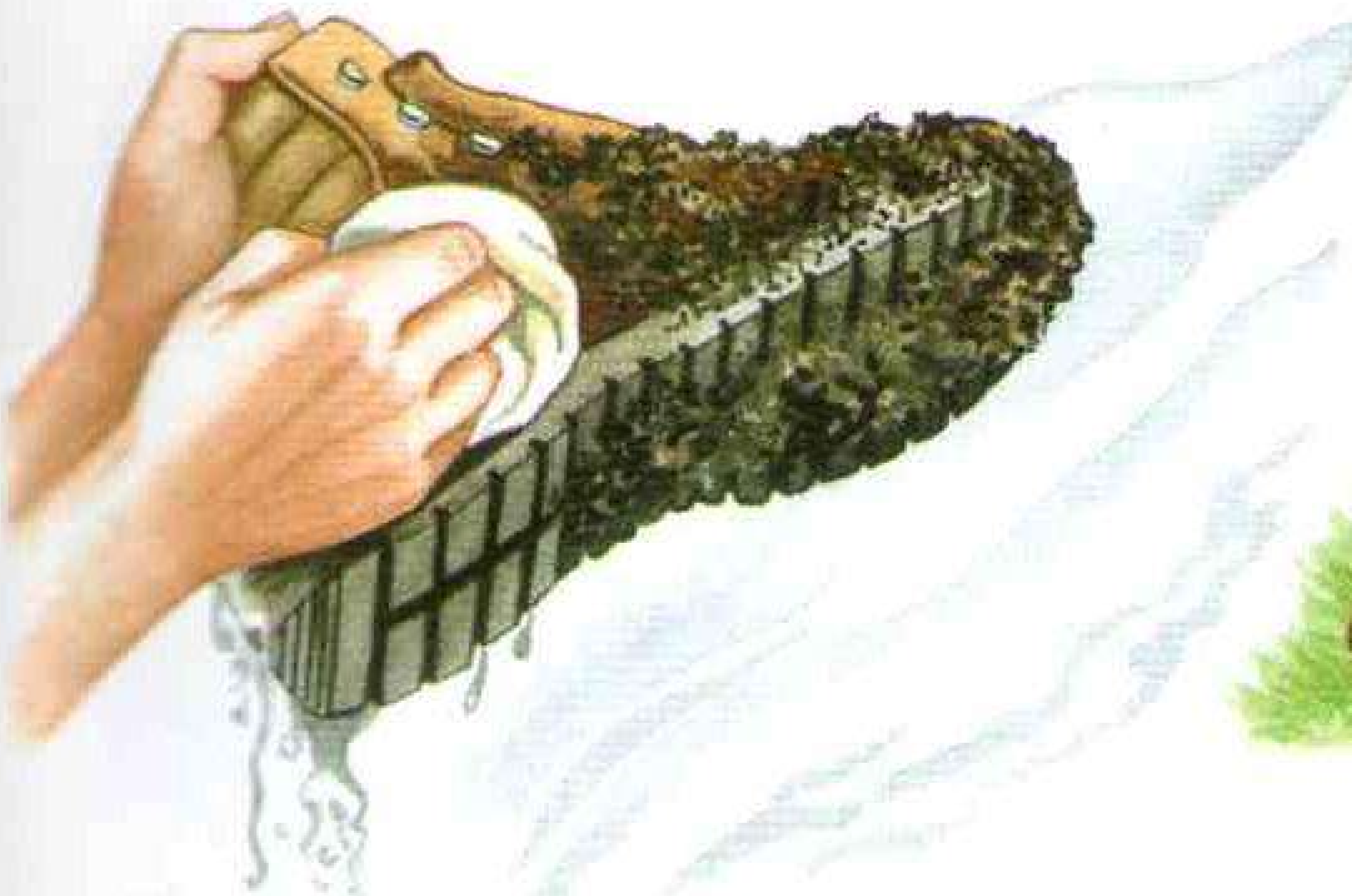

1 Retire los cordones y plantillas interiores y lave las botas para eliminar el barro. Es muy importante quitar toda la turba, cuyo contenido ácido puede dañar la piel de las botas.

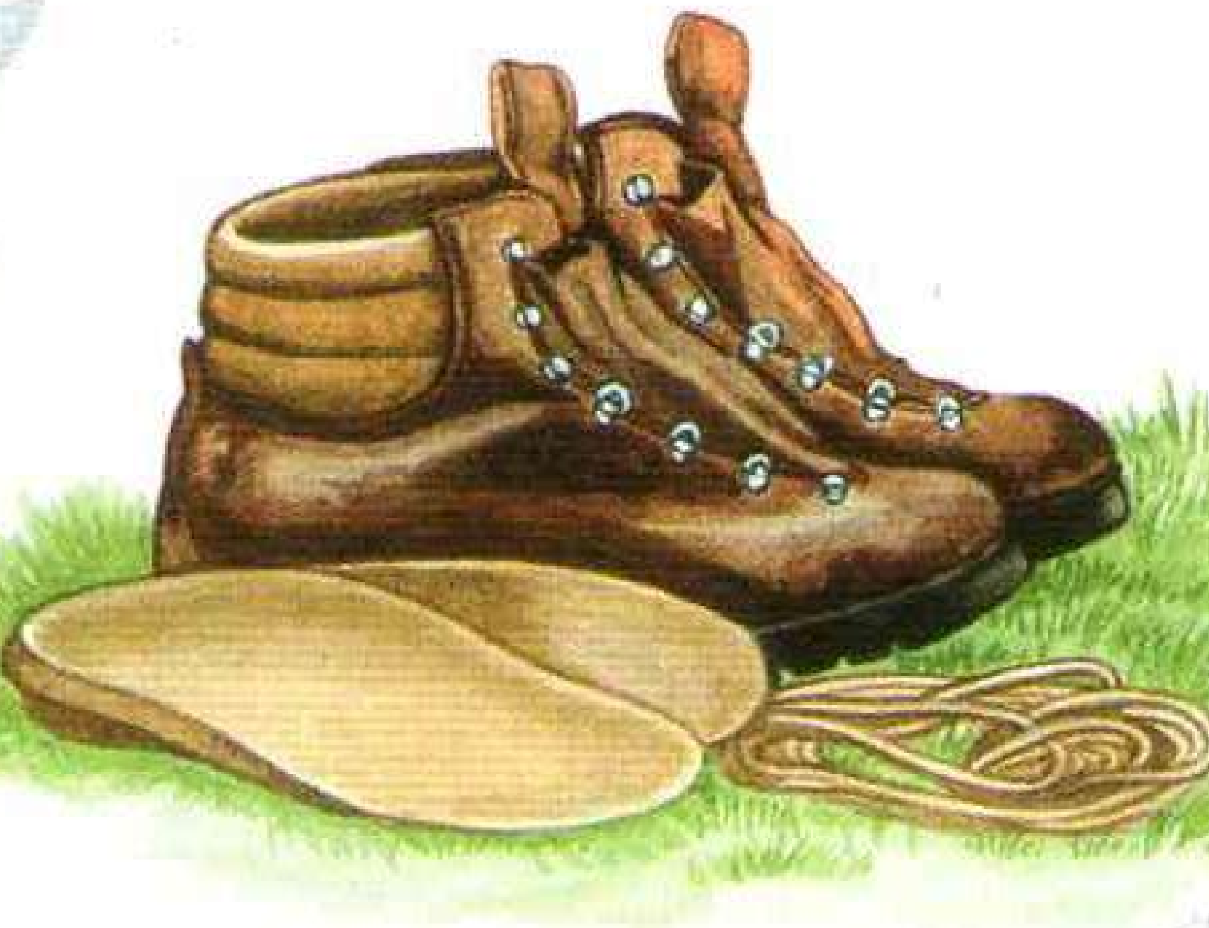

2 Deje secar las botas completamente. Si es posible, manténgalas alejadas de calor directo del fuego o del sol, ya que pueden agrietar la piel.

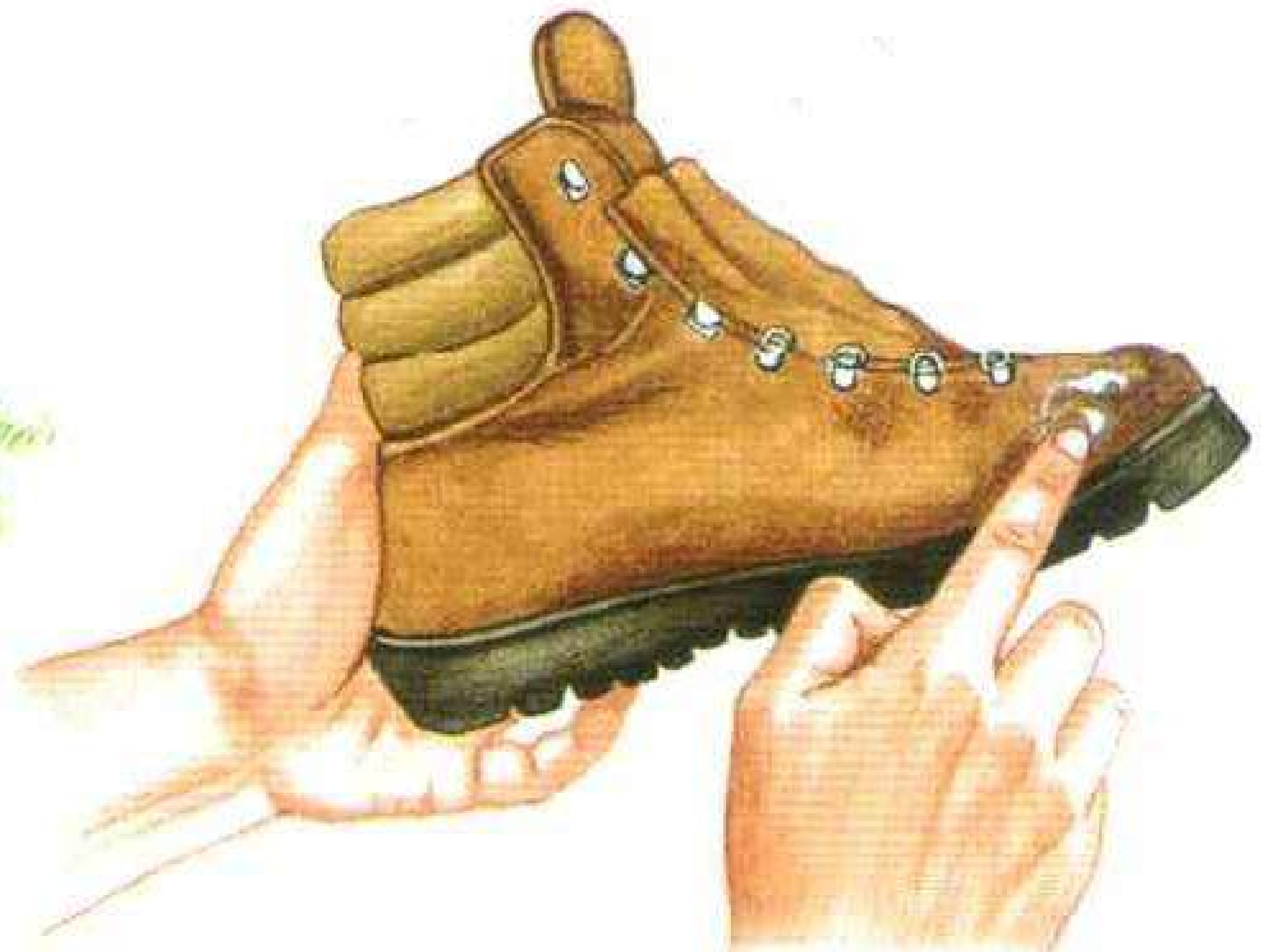

3 Una vez las botas estén secas, frótelas con el dedo con un compuesto impermeable. Si no tiene previsto ponerse las botas enseguida, guárdelas en un lugar fresco y seco.

EQUIPO PERSONAL DE CAMPAMENTO

ÚNICAMENTE REUNIRÁ el equipo perfecto de campamento a base de intentos y equivocaciones. Antes de embarcarse en un viaje, considere la posibilidad de omitir cualquiera de los elementos de su equipo que no utilizó en su viaje anterior. Las modificaciones en el equipo vendrán por sí solas. Algunos artículos serán inamovibles, y pronto mostrarán las señales de su uso continuo. Llegará un punto en el que se encontrará tan identificado con su equipo que no querrá ni considerar la posibilidad de perder o romper cualquiera de sus elementos.

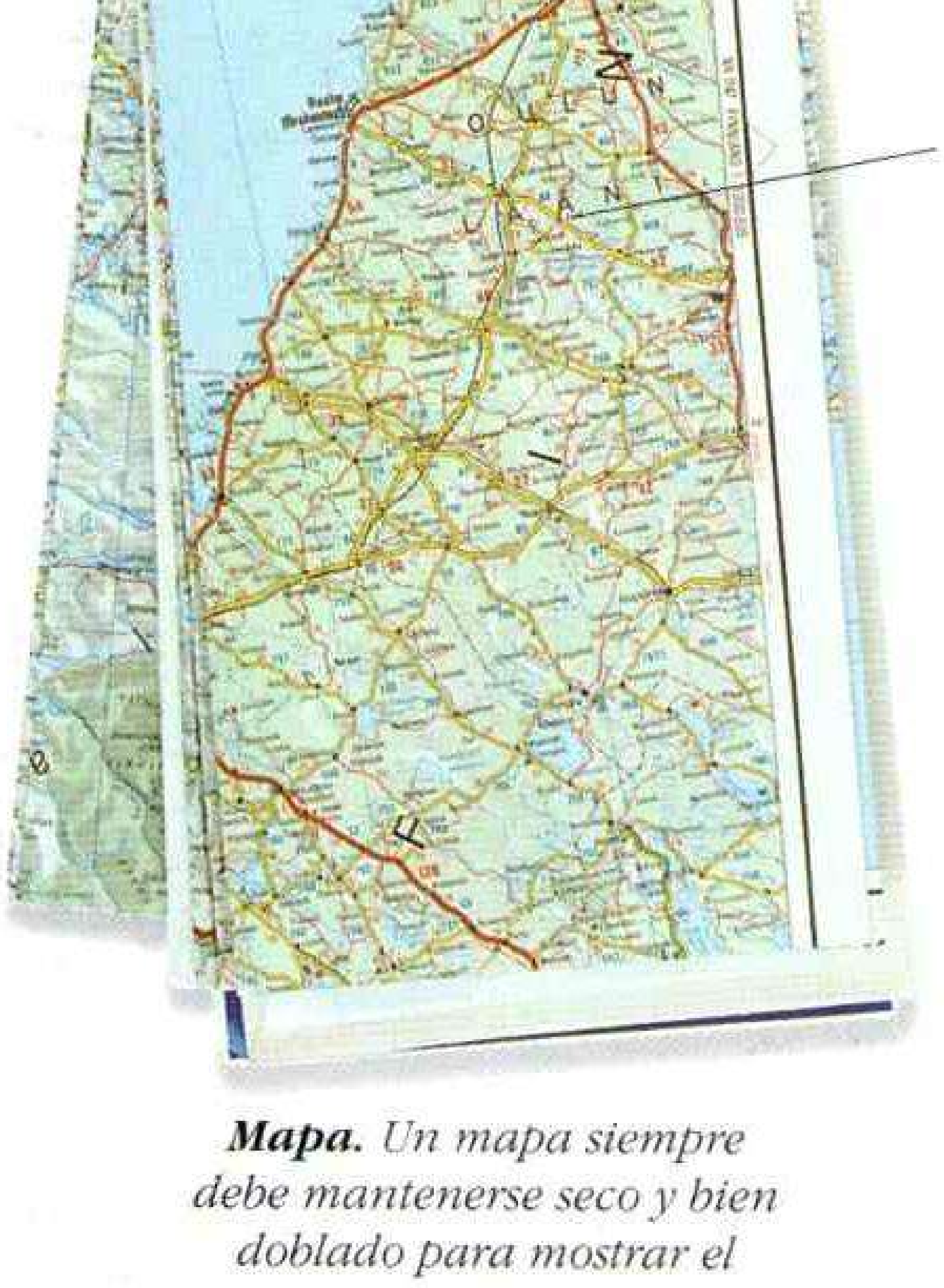

El mapa se dobla abierto mostrando el área en uso

Mapa. *Un mapa siempre debe mantenerse seco y bien doblado para mostrar el área en uso.*

Brújula. *Una brújula es una pieza básica del equipo, por lo que se debe guardar en un sitio seguro.*

Gire la parte superior para encender la linterna

Linterna. *Una linterna debe ser lo más pequeña posible y resistente al agua.*

Prismáticos. *Deben ser lo más pequeños y resistentes posible. Incluso si son estancos, guárdelos en una bolsa de plástico.*

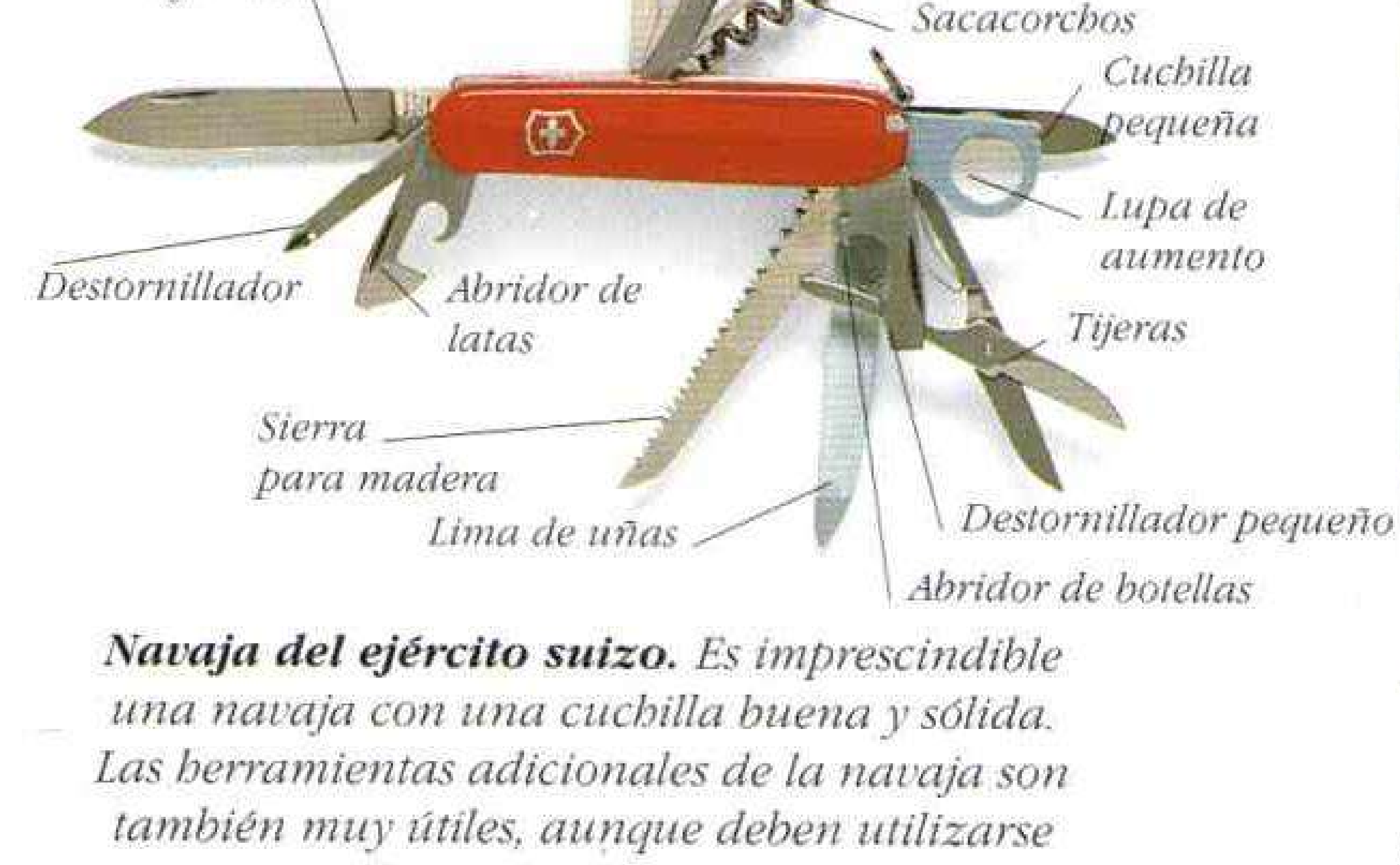
Destornillador fino
Escariador
Cincel
Gancho
Cuchilla grande
Sacacorchos
Cuchilla pequeña
Lupa de aumento
Destornillador
Abridor de latas
Tijeras
Sierra para madera
Lima de uñas
Destornillador pequeño
Abridor de botellas

Navaja del ejército suizo. *Es imprescindible una navaja con una cuchilla buena y sólida. Las herramientas adicionales de la navaja son también muy útiles, aunque deben utilizarse con cuidado.*

Escudilla. *Una buena escudilla es muy útil para cocinar en el campamento* (véase página 112). *Coloque cinta aislante alrededor de las agarraderas metálicas para evitar quemaduras.*

Botella de agua. *Una botella de metal puede agrietarse si el agua que contiene se congela. Una de plástico puede fundirse si se coloca cerca del fuego.*

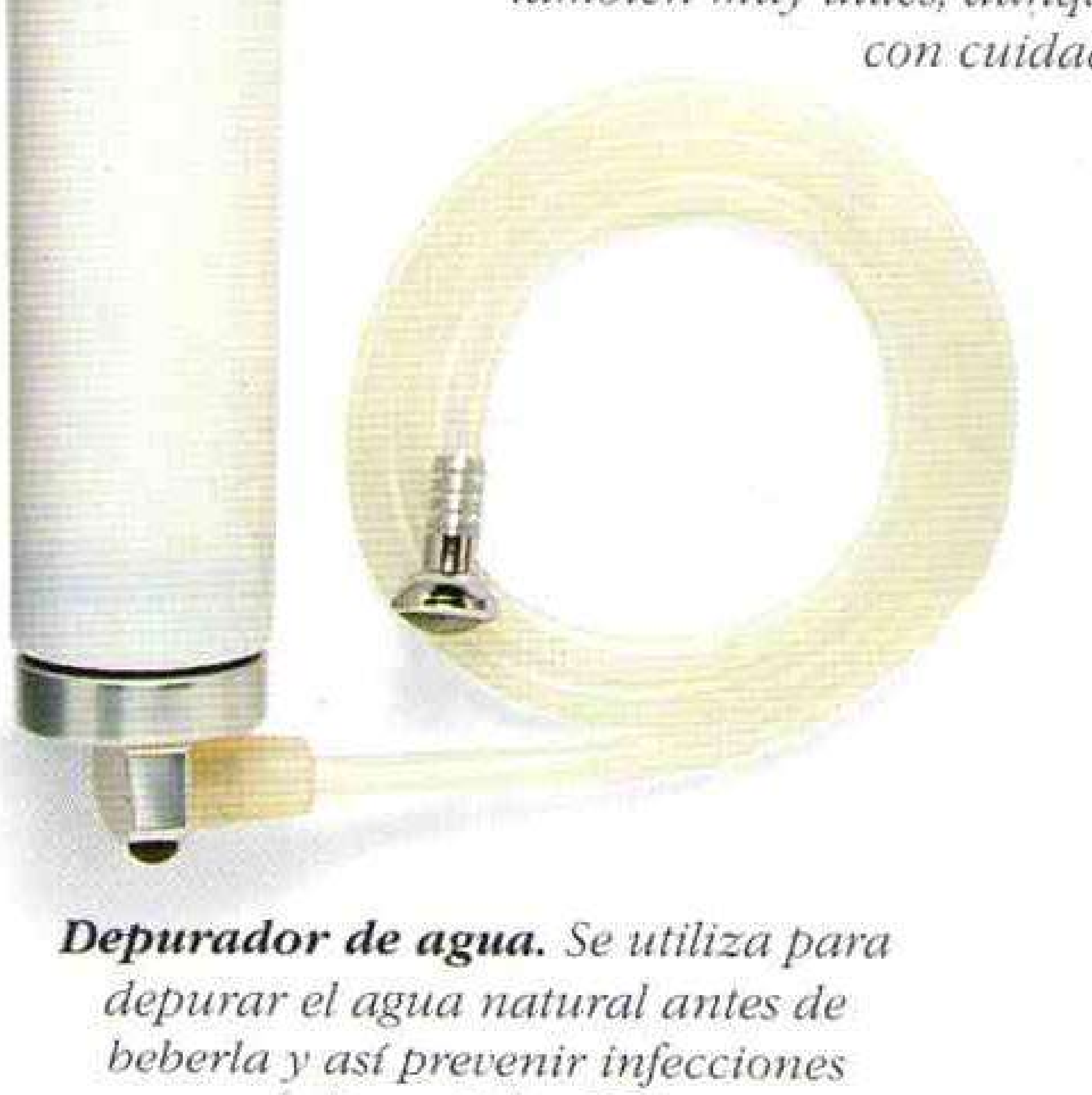

Depurador de agua. *Se utiliza para depurar el agua natural antes de beberla y así prevenir infecciones* (véase página 75).

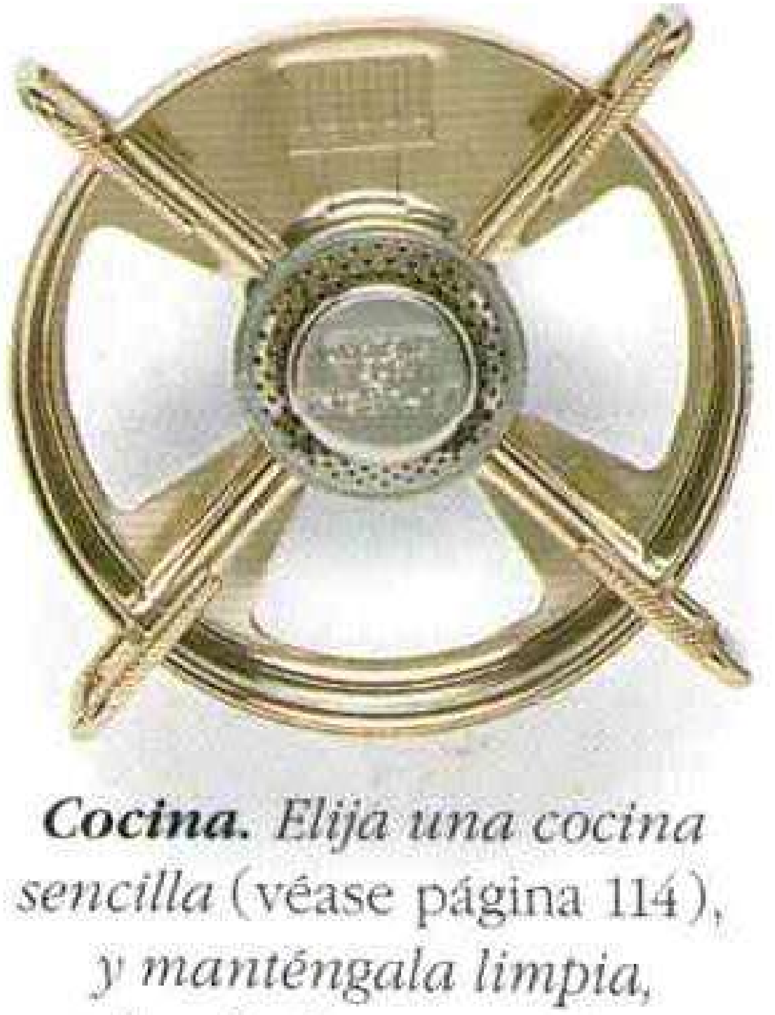

Cocina. *Elija una cocina sencilla* (véase página 114), *y manténgala limpia, además de protegerla de los golpes.*

Combustible para la cocina. *Asegúrese de que no tenga fugas, y de que esté limpio, filtrándolo si es necesario.*

Plato llano
Plato hondo

Platos y recipientes. *Utilice platos y recipientes de plástico o madera para evitar las quemaduras en la boca, que son dolorosas y pueden infectarse* (véase página 112). *Nunca coma de un cazo caliente.*

Taza. *Una taza de plástico es útil para contener bebidas frías o calientes* (véase página 112).

Cuchillo
Tenedor
Cuchara
Funda para cubiertos

Cubiertos. *Mantenga limpios todos los utensilios para evitar intoxicaciones. Si pierde o rompe sus cubiertos, puede tallar unos nuevos de madera* (véase página 113).

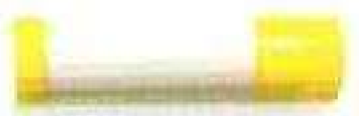

Protector labial. *El protector labial es fundamental en todo tipo de climas, ya que evita que los labios se corten. Utilícelo en cualquier momento del día, cuando los labios se le resequen.*

Protector solar. *Una crema protectora solar es indispensable en la nieve y en los climas soleados.*

Kit de supervivencia. *Contiene artículos básicos que no se encuentran en la naturaleza* (véase página 28).

Toalla. *Una toalla grande puede cortarse en pedazos más pequeños para secarse el sudor o humedecerse la cara. También puede utilizarse como pañuelo* (véase página 142).

Botiquín de primeros auxilios. *Es fundamental para el tratamiento de las heridas* (véase página 174).

Cerillas impermeables. *Tienen cabezas enceradas que las mantienen secas.*

Equipo de costura. *Utilícelo para remendar la ropa y las tiendas.*

La cremallera evita la entrada de suciedad o insectos

Útiles de aseo. *Guarde todos los artículos de aseo personal en una bolsa especial. Cuelgue la bolsa de un árbol mientras se lave para evitar que algún artículo se pierda* (véase página 57).

Caliente y seco

La tienda y el saco de dormir serán sus únicas comodidades cuando se encuentre en plena naturaleza, aunque puede sobrevivir sin ellas. Siempre deben guardarse convenientemente, y los sacos de dormir no deben mojarse nunca.

Tienda de campaña

Saco de dormir

EQUIPO BÁSICO DE SUPERVIVENCIA

EN SITUACIONES DE supervivencia, algunos artículos pueden llegar a constituir la diferencia entre la vida y la muerte. Lleve siempre sobre su persona los artículos de mayor importancia, preferiblemente dentro de una lata pequeña que pueda meter en un bolsillo. Siempre que se aventure en la naturaleza, asegúrese de llevar otros utensilios que pudieran ser vitales para su supervivencia individual y afianzarlos de tal manera que la posibilidad de perderlos o dañarlos sea mínima. Algunos de estos útiles pueden colgarse de cuerdas resistentes alrededor del cuello; métalos debajo de la ropa para que no puedan desprenderse al engancharse con la vegetación. Coloque cada uno de los artículos en una capa diferente de la ropa para que no choquen entre sí y se rompan. Debe reunir un equipo de supervivencia acorde con sus necesidades personales.

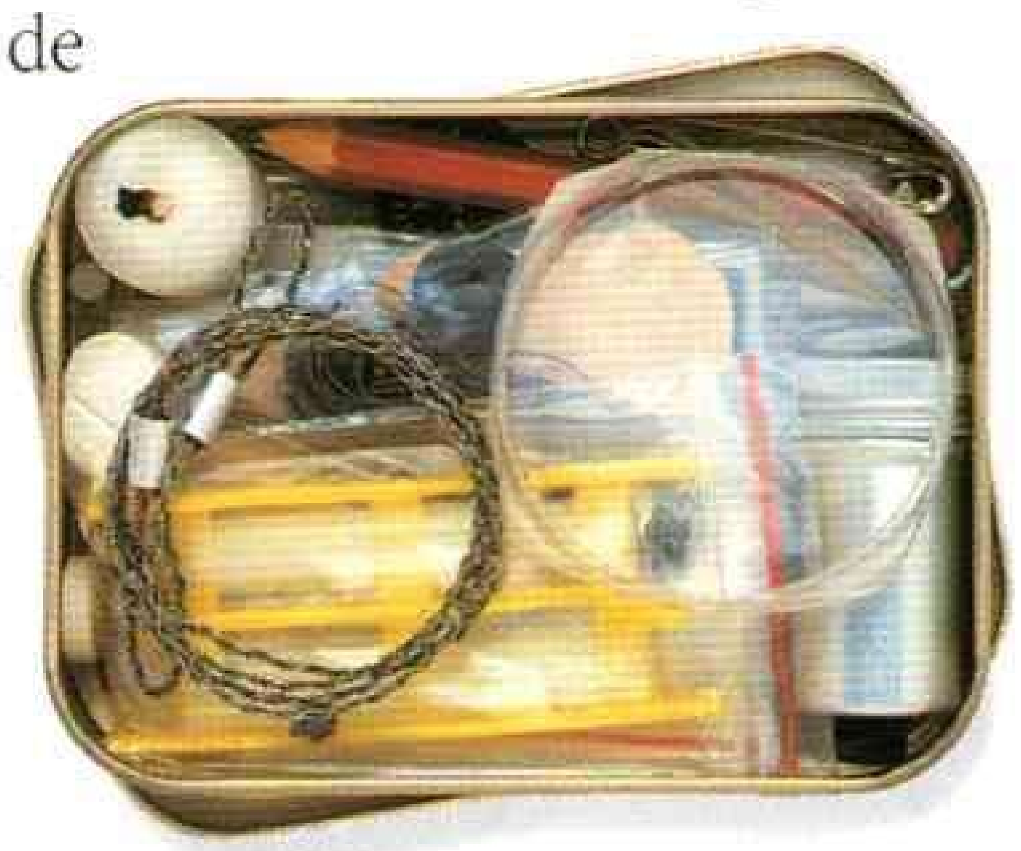

Equipo de supervivencia. *Los artículos más importantes de su equipo de supervivencia deben guardarse en una lata pequeña provista de una tapa que ajuste bien.*

Imperdibles. *Utilícelos para asegurar su ropa, o para arreglar su tienda o saco de dormir.*

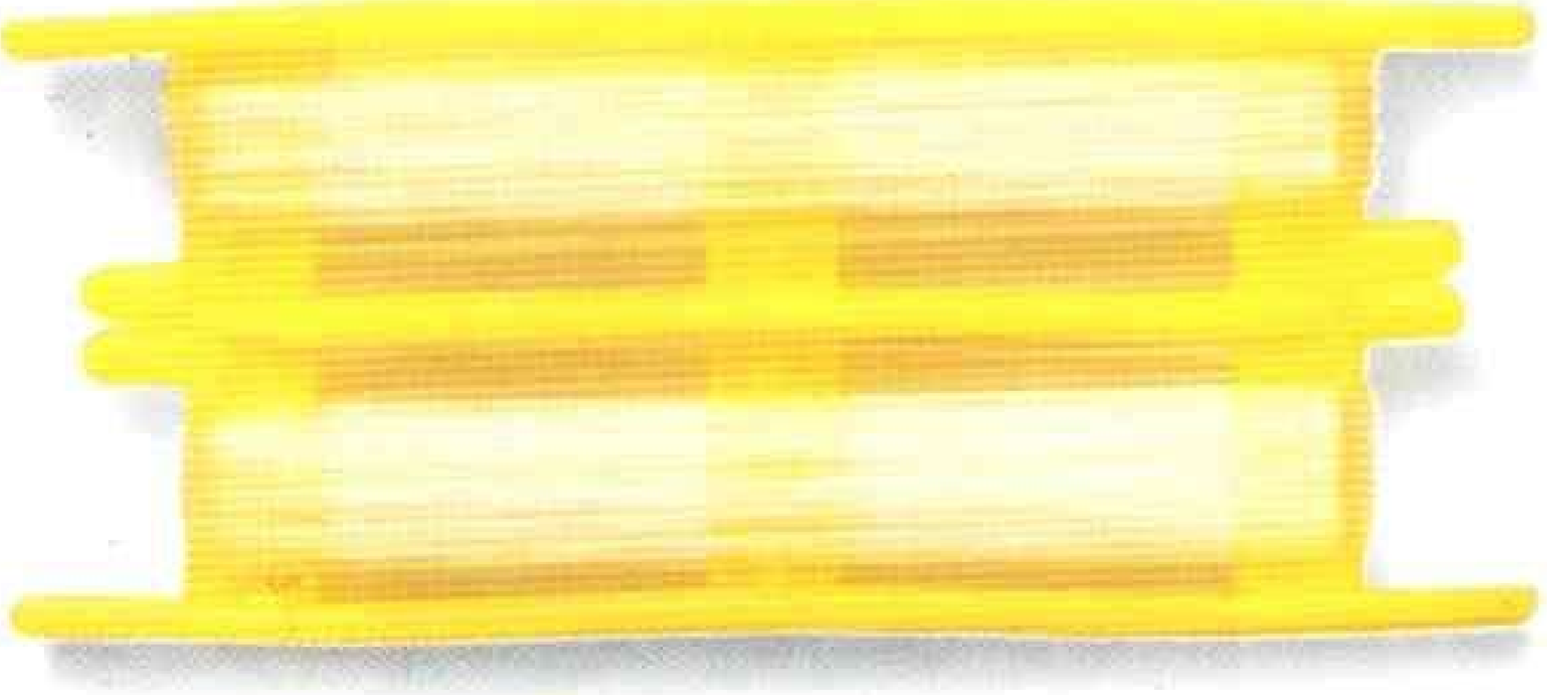

Hilo de pescar. *Lleve siempre consigo un hilo resistente para pescar o para atar herramientas.*

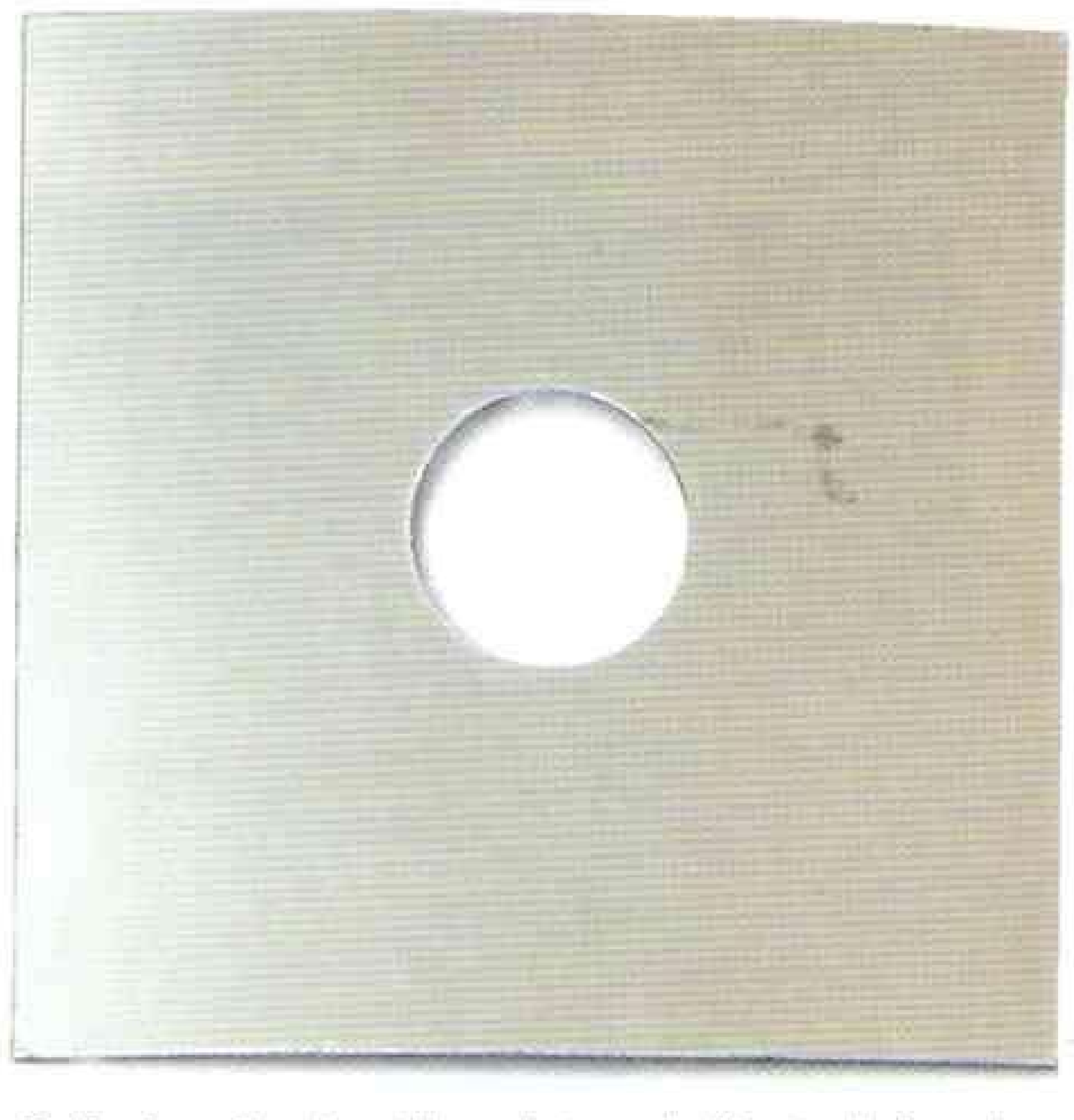

Reflector. *Puede utilizar la tapa brillante de la caja de su equipo de supervivencia, o bien un heliógrafo, para llamar la atención en una situación de supervivencia. Mirando a través del agujero en el centro del heliógrafo puede dirigir la luz reflejada hacia un avión.*

Alambre fino. *El alambre tiene muchas aplicaciones y puede reutilizarse varias veces. Debe ser lo suficientemente fino como para doblarse fácilmente, aunque no tanto como para que se rompa.*

Anzuelos y plomos. *Incluya una buena cantidad de anzuelos, corchos y plomos. Los anzuelos pequeños pueden utilizarse tanto para pescar peces grandes como pequeños.*

Lata de tabaco. *Debe guardar su equipo de supervivencia en una pequeña lata de tabaco provista de una tapa con buen ajuste. Selle la lata con cinta para mantener seco su contenido.*

Brújula. *Es esencial. La brújula ideal debería ser luminosa.*

Vela. *Utilícela para encender un fuego y no como lámpara. No utilice velas de sebo porque pueden pudrirse en climas cálidos.*

Cerillas. *Compre cerillas impermeables, o impermeabilícelas dejando caer una pequeña capa de cera de vela sobre sus cabezas y parte del cuerpo. Elimine la cera de cada cerilla antes de encenderla.*

Agujas, hilo y botones. *Incluya varias agujas en su equipo, todas ellas de ojo grande, así como una buena cantidad de hilo grueso, resistente y preferiblemente impermeable. Envuelva el hilo alrededor de las agujas y guárdelas en una bolsa de plástico. Los botones grandes son útiles para sujetar las pestañas de la tienda.*

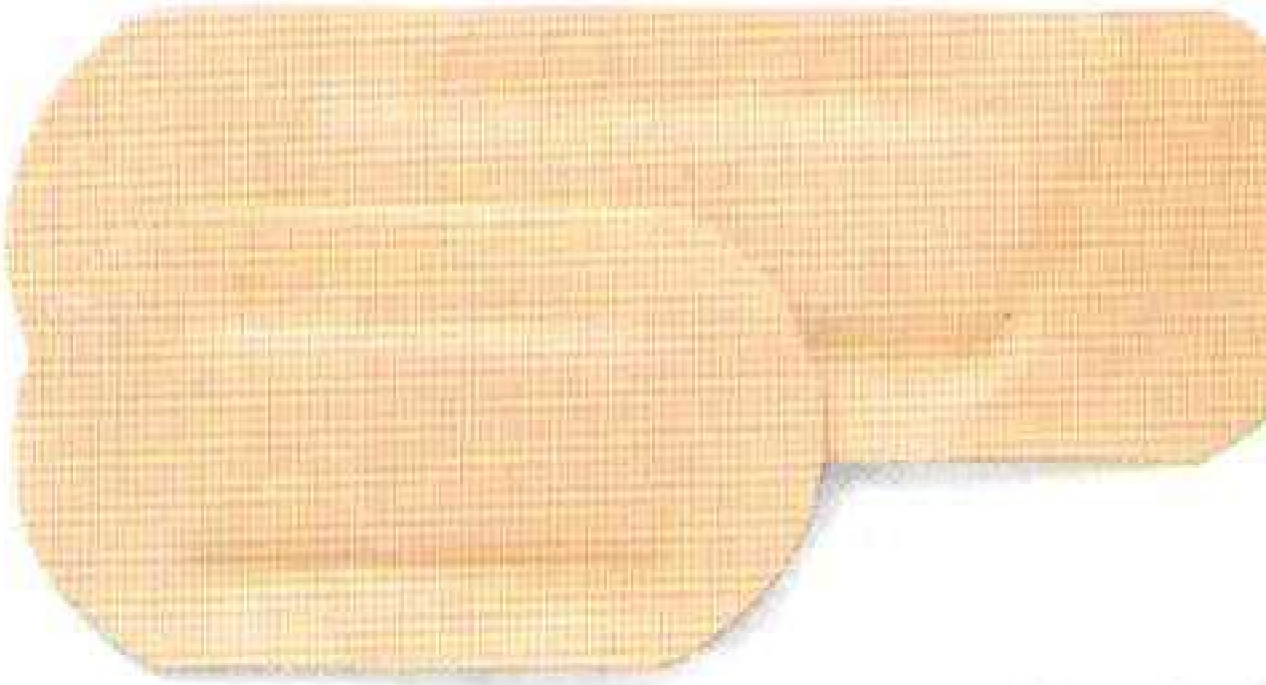

Tiritas adhesivas. *Debe incluir tiritas adhesivas de varios tamaños para evitar que alguna rozadura llegue a infectarse, o para proteger las ampollas del pie hasta que sanen.*

Antibióticos. *No deben utilizarse hasta que sean verdaderamente necesarios.*

Pastillas esterilizadoras. *Guárdelas para una emergencia, cuando le sea imposible hervir agua que le parezca insalubre.*

Cable sierra. *Un cable sierra es de empleo fácil, aunque lento; con él podrá cortar la mayoría de los materiales. Guárdelo cubierto con una película de grasa dentro de una bolsa de plástico, para que no se oxide ni se rompa.*

ENGLAND
STITCH CUTTER

Bisturí. *Puede emplearse para muchos propósitos. Guárdela con la cuchilla engrasada en su envase original.*

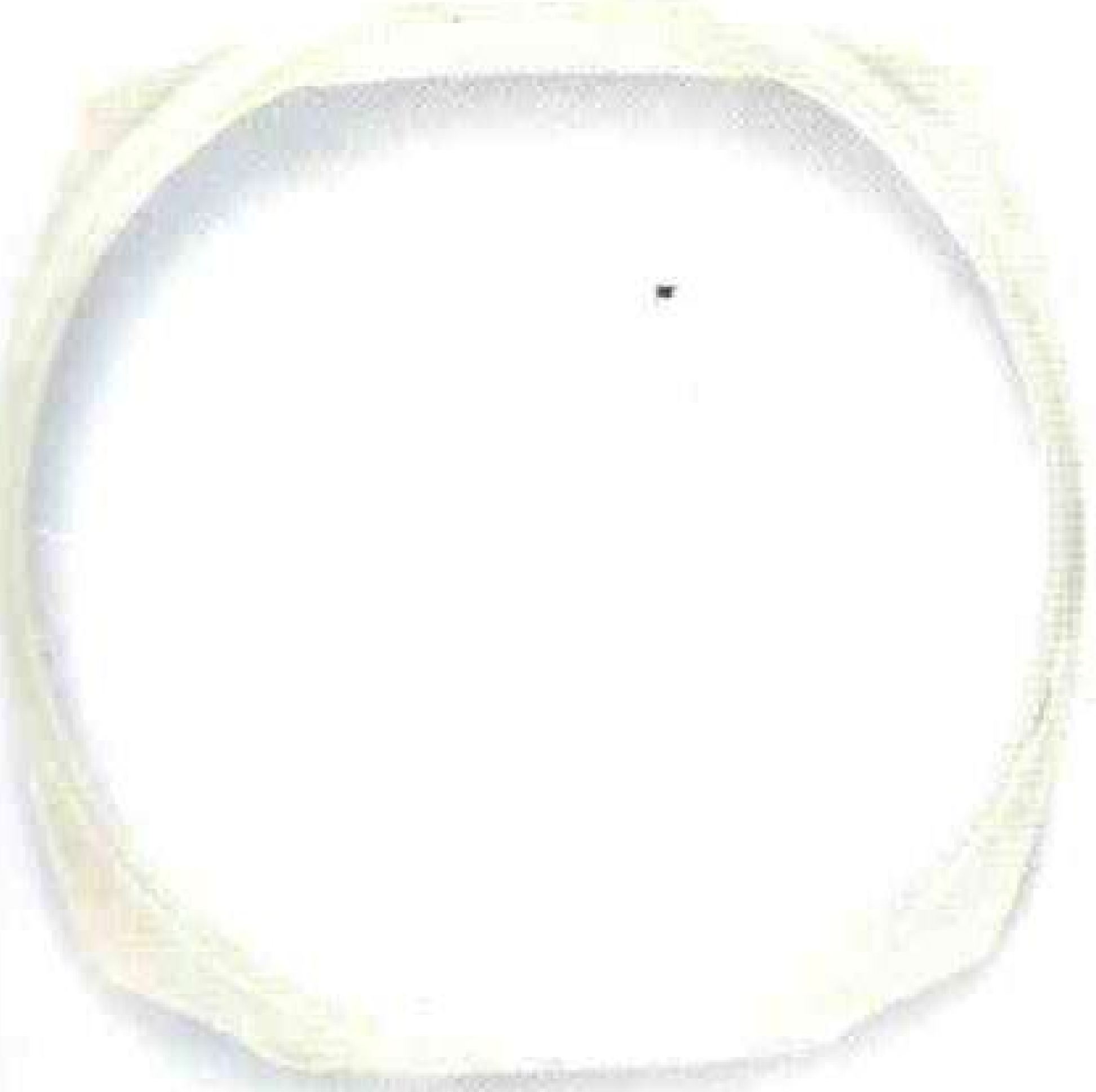

Lupa. *Una lupa puede hacer converger los rayos del sol en la madera seca, para hacerla arder y, de este modo, encender un fuego.*

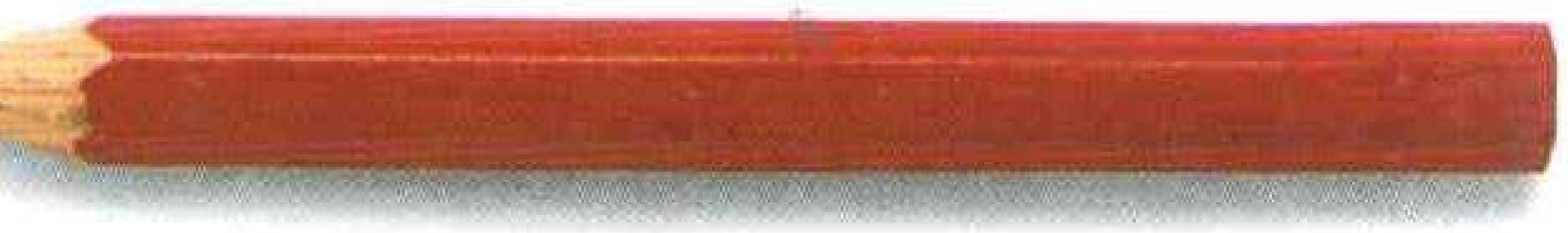

Lápiz. *Un lápiz es de gran utilidad para anotar los sitios en los que ha estado, lo que es comestible, y para dibujar mapas.*

Sal. *Es esencial para sobrevivir. Si no repone la sal perdida en el sudor y la orina, puede llegar a enfermar gravemente.*

Bolsa de plástico. *Una bolsa de plástico resistente tiene múltiples utilidades, como llevar agua desde un río, u obtenerla de la vegetación en una emergencia.*

Permanganato potásico. *Estos cristales son útiles para purificar el agua. Mezclados con agua obtienen un color rojo intenso; pueden utilizarse para el tratamiento de algunas infecciones por hongos, como el pie de atleta. Siga las instrucciones del envase.*

EL EQUIPO BÁSICO ALREDEDOR DE SU CUELLO

No debe prescindir nunca de las piezas más valiosas de su equipamiento, en particular si se adentra en tierra virgen. Debe tomar las precauciones necesarias para no perderlas o romperlas, ya que, alejado de la civilización, no podrán ser arregladas o reemplazadas. Ate cada uno de los artículos a un cordón diferente y cuélgueselos del cuello, de esta manera los tendrá siempre a mano cuando los necesite. Guárdelos entre su ropa cuando no los utilice para evitar que se rompan.

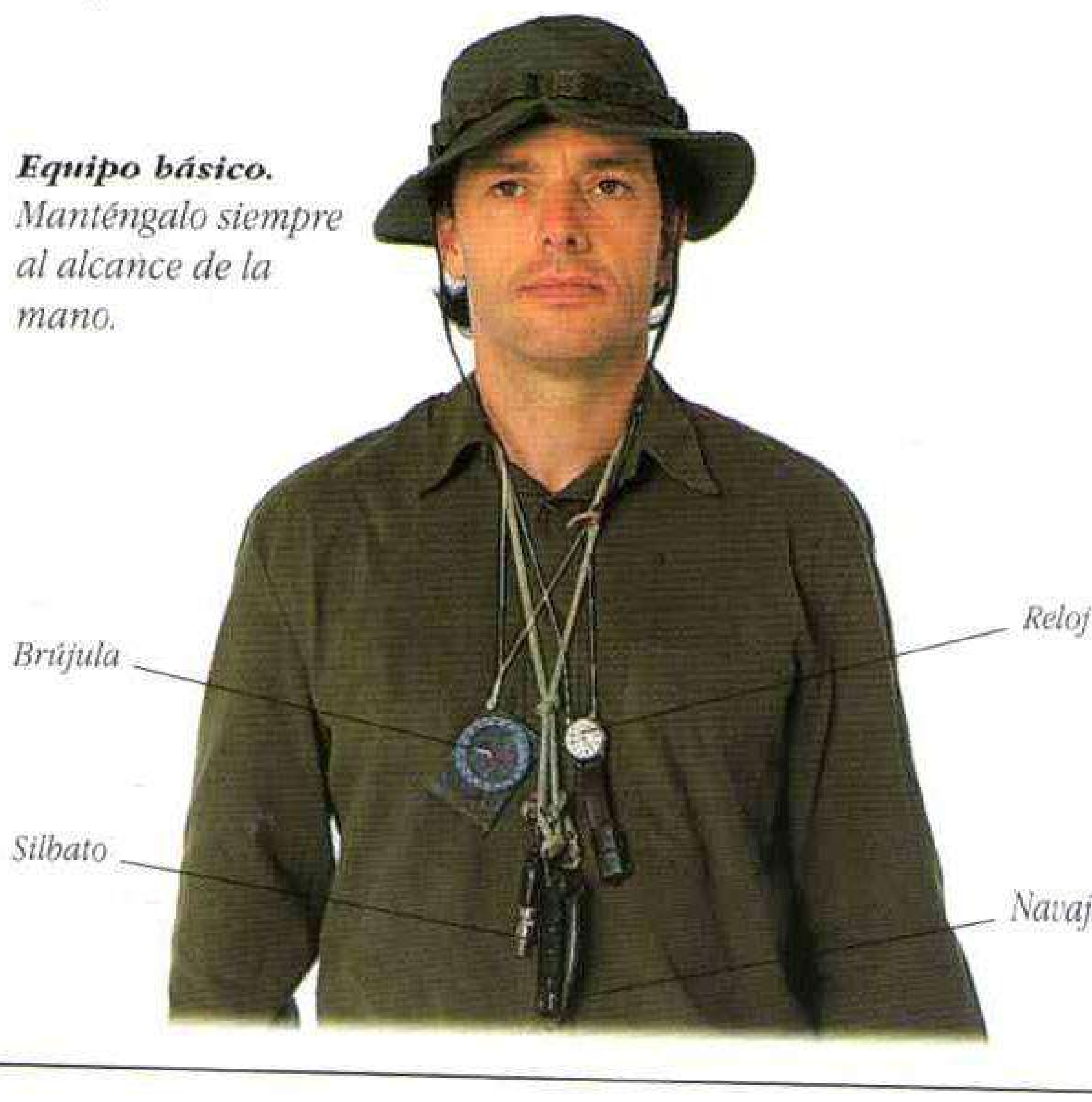

Equipo básico. *Manténgalo siempre al alcance de la mano.*

CAMAS Y SACOS DE DORMIR

EL DESCANSO ES tan importante para la supervivencia como el alimento. Un buen e ininterrumpido descanso nocturno puede aliviar todas las preocupaciones y tensiones y convertir la melancolía pesimista en un optimismo positivo. El estar abrigado y seco le proporciona una mayor sensación de seguridad. Incluso en el calor húmedo de la selva son necesarias la protección y el aislamiento de la humedad del suelo.

CARACTERÍSTICAS DE LOS SACOS DE DORMIR

Aunque hay muchas formas y estilos de sacos de dormir, un saco siempre debe estar lo suficientemente bien acolchado como para mantenerle abrigado incluso si carece de tienda. No debe permitir nunca que un saco de dormir se moje. Guárdelo siempre en una funda impermeable de un material poroso *(véase página 43)*.

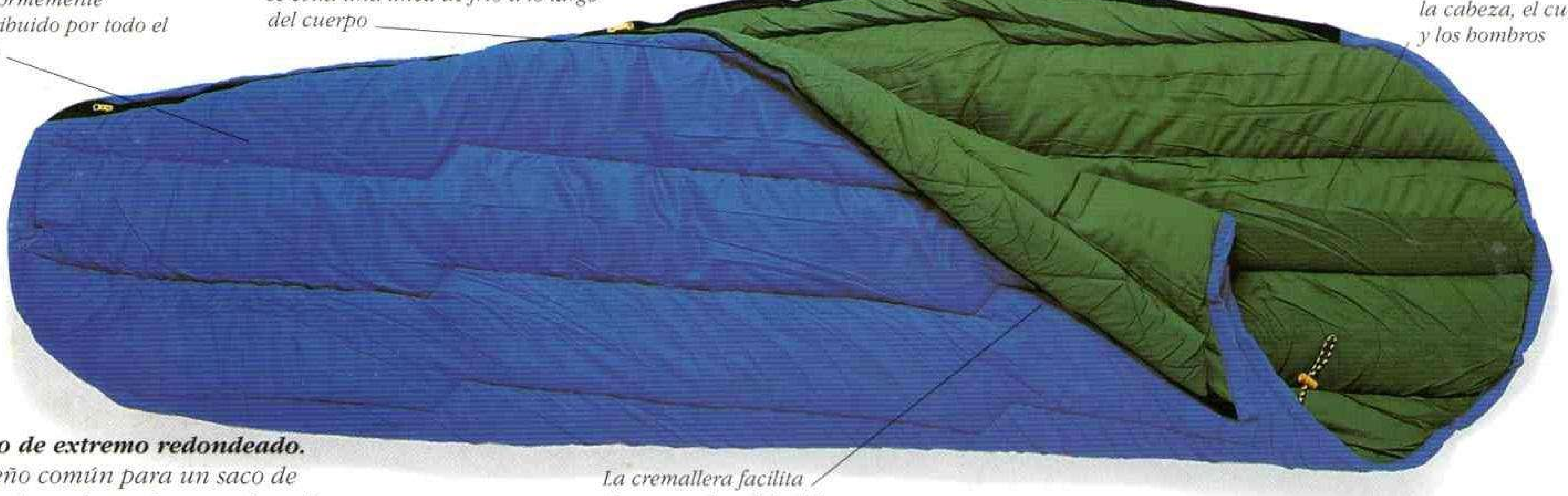
El relleno debe estar uniformemente distribuido por todo el saco
La cremallera debe tener una solapa que la cubra en toda su extensión, de esta manera se evita una línea de frío a lo largo del cuerpo
El fondo curvo ayuda a retener el calor corporal alrededor de la cabeza, el cuello y los hombros
La cremallera facilita el acceso y la salida del saco

Saco de extremo redondeado. *Diseño común para un saco de dormir moderno. La capucha evita que se escape el calor de la cabeza, el pecho y los hombros.*

El estrechamiento en los pies reduce el tamaño y el peso del saco

Saco sin cremallera. *La abertura de este saco puede cerrarse con un cordel para mantener la cabeza y el cuello bien protegidos en climas fríos. Esto es muy importante, ya que puede perder la mitad de su calor corporal a través de la cabeza. Al carecer de cremallera, puede resultar difícil salir del saco rápidamente.*

La almohada complementaria puede doblarse hacia dentro

Saco de forma rectangular. *Este saco es ideal para acampar si se dispone de un vehículo, y no tanto para dormir a la intemperie. Es muy ligero y, por tanto, no adecuado para climas muy fríos como las regiones polares.*

CONFORMACIÓN DE UN SACO DE DORMIR

Un saco de dormir relleno de plumón es ligero y proporciona abrigo. Sin embargo, el plumón puede perder su capacidad aislante al compactarse en bolas cuando se humedece o se almacena muy apretado. Los rellenos sintéticos son más abultados, pesados y carecen de comodidad. Sin embargo, le mantendrán abrigado aun si se mojan, y se secan rápidamente.

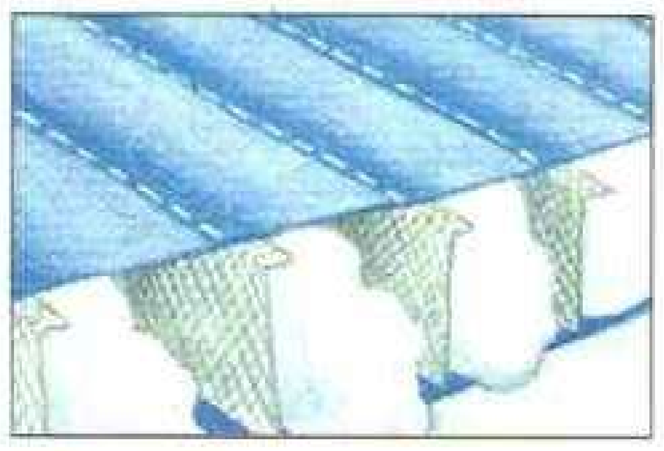

Tabiques. *En los sacos, el plumón se distribuye en «tabiques» de tela para reducir el apelotonamiento del relleno.*

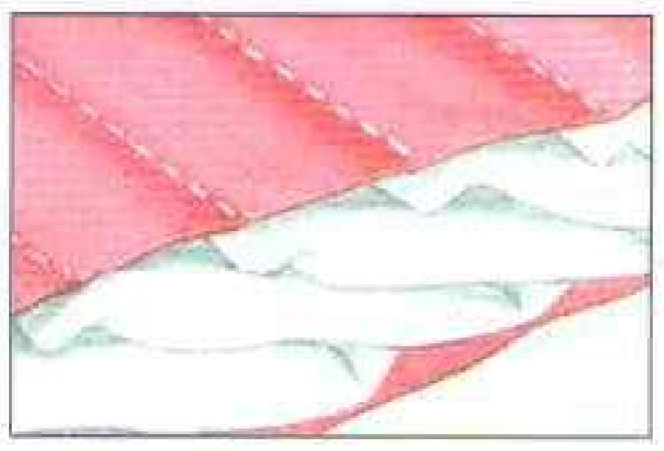

Tablillas. *Tienen capas inclinadas de fibras solapadas que se llenan con aire para proporcionar aislamiento térmico.*

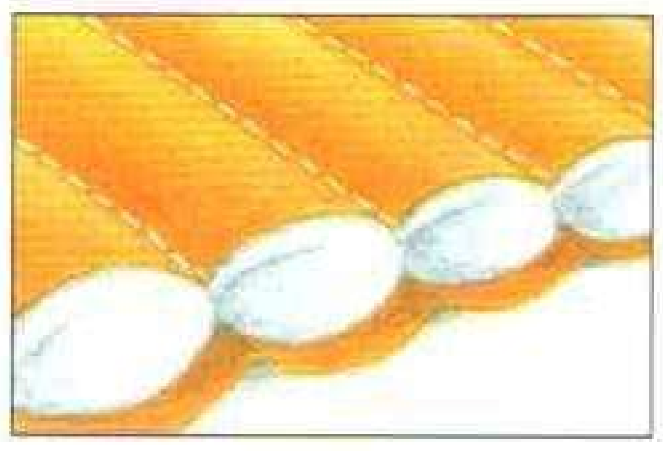

Alvéolos. *El relleno se distribuye en alvéolos ovalados separados. Sin embargo, se pierde calor a través de las costuras.*

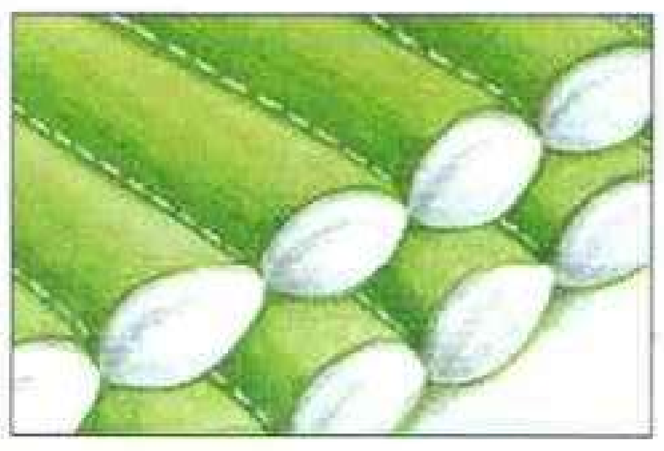

Doble barrera de alvéolos al tresbolillo. *Evita el apelotonamiento del relleno y la pérdida de calor a través de las costuras.*

ACCESORIOS PARA DORMIR

Funda. *Una funda de algodón forma una capa de aire aislante entre su cuerpo y el saco de dormir. Se lava con más facilidad que el saco.*

Aislante. *Es esencial para aislarse de la humedad y el frío del suelo. Debe colocarla debajo del saco de dormir.*

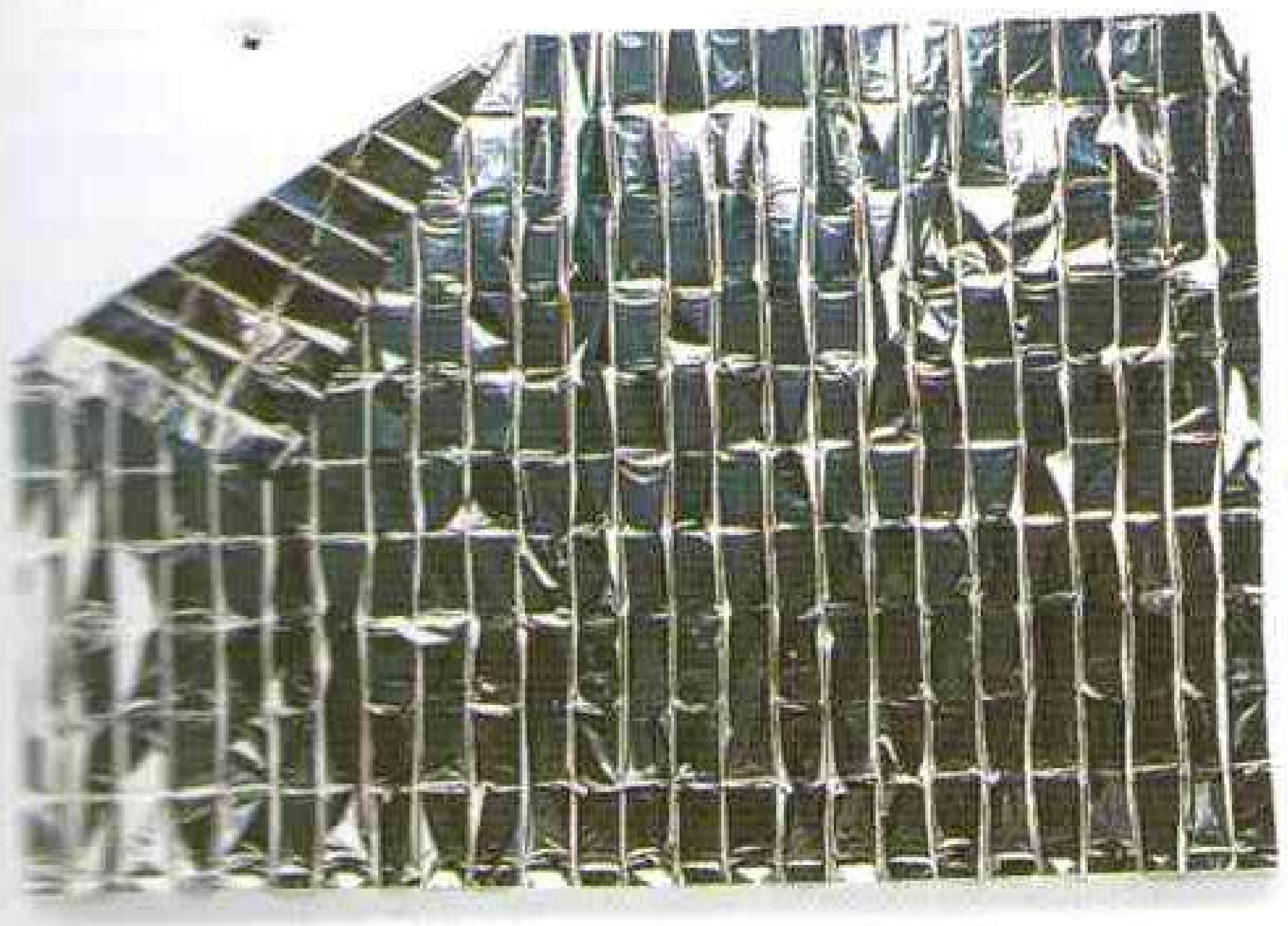

Manta de aluminio. *Se trata de una lámina de aluminio muy ligera para emergencias. El aluminio evita la pérdida de calor del cuerpo. También puede usar la manta como toldo para protegerse del sol, por ejemplo en un desierto.*

Almohada hinchable. *Aunque no es estrictamente necesaria, le proporcionará comodidad adicional en los viajes largos.*

CUIDADOS Y REMIENDOS

Sacuda su saco de dormir antes de utilizarlo para distribuir uniformemente el relleno, y airéelo después de su uso. Si se rasga, remiéndelo inmediatamente con aguja e hilo para evitar que el daño se extienda. Una cinta adhesiva ancha resulta útil para sellar desgarrones de forma temporal. Puede reparar su saco rasgado con un pedazo de tela de la bolsa que utiliza para guardarlo *(véase página 27)*, a menudo confeccionada con el mismo material.

Equipo de costura. Un equipo de costura resulta esencial para remendar los desgarrones en sacos de dormir, tiendas y en la ropa.

LA CONSTRUCCIÓN DE UNA CAMA

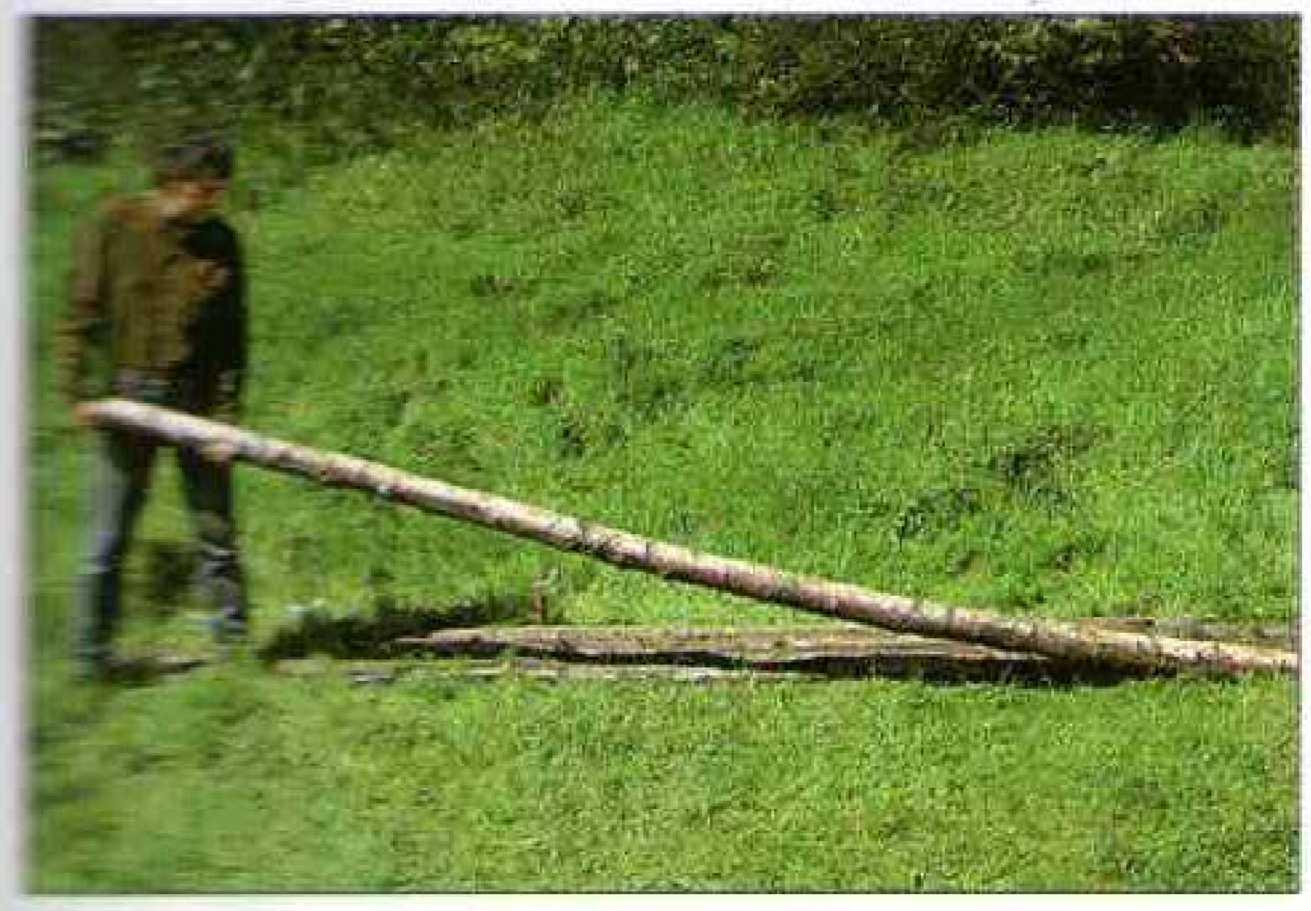

1 Si carece de un saco de dormir, debe construirse una cama levantada del suelo para asegurar un aislamiento adecuado y evitar los insectos y animales. Coloque varios troncos de longitud mayor que su estatura, uno junto a otro, en el suelo para formar la base de la cama.

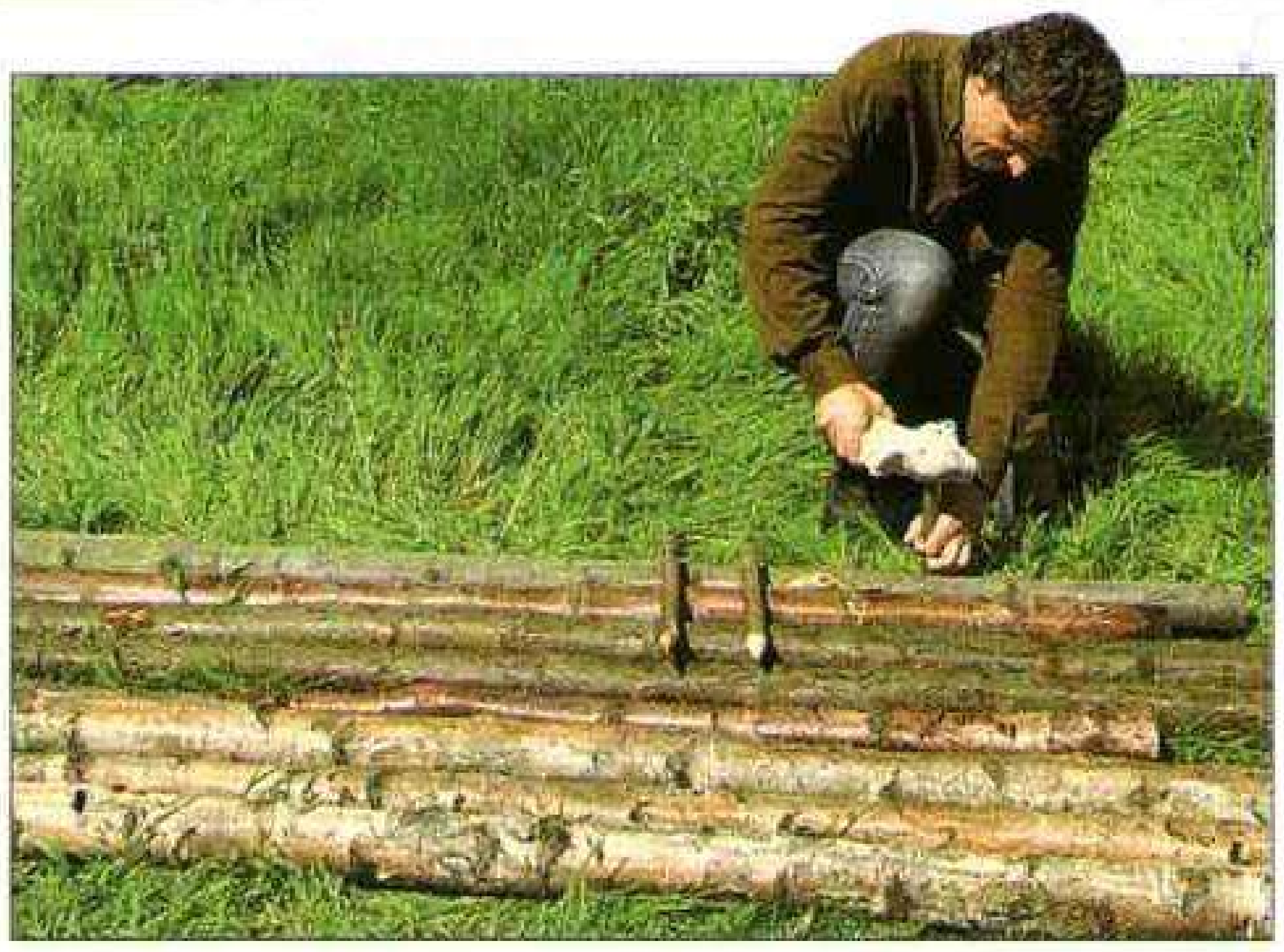

2 Mantenga juntos los troncos y clave en el suelo unas estacas que habrá confeccionado a partir de ramas a las que previamente habrá afilado un extremo.

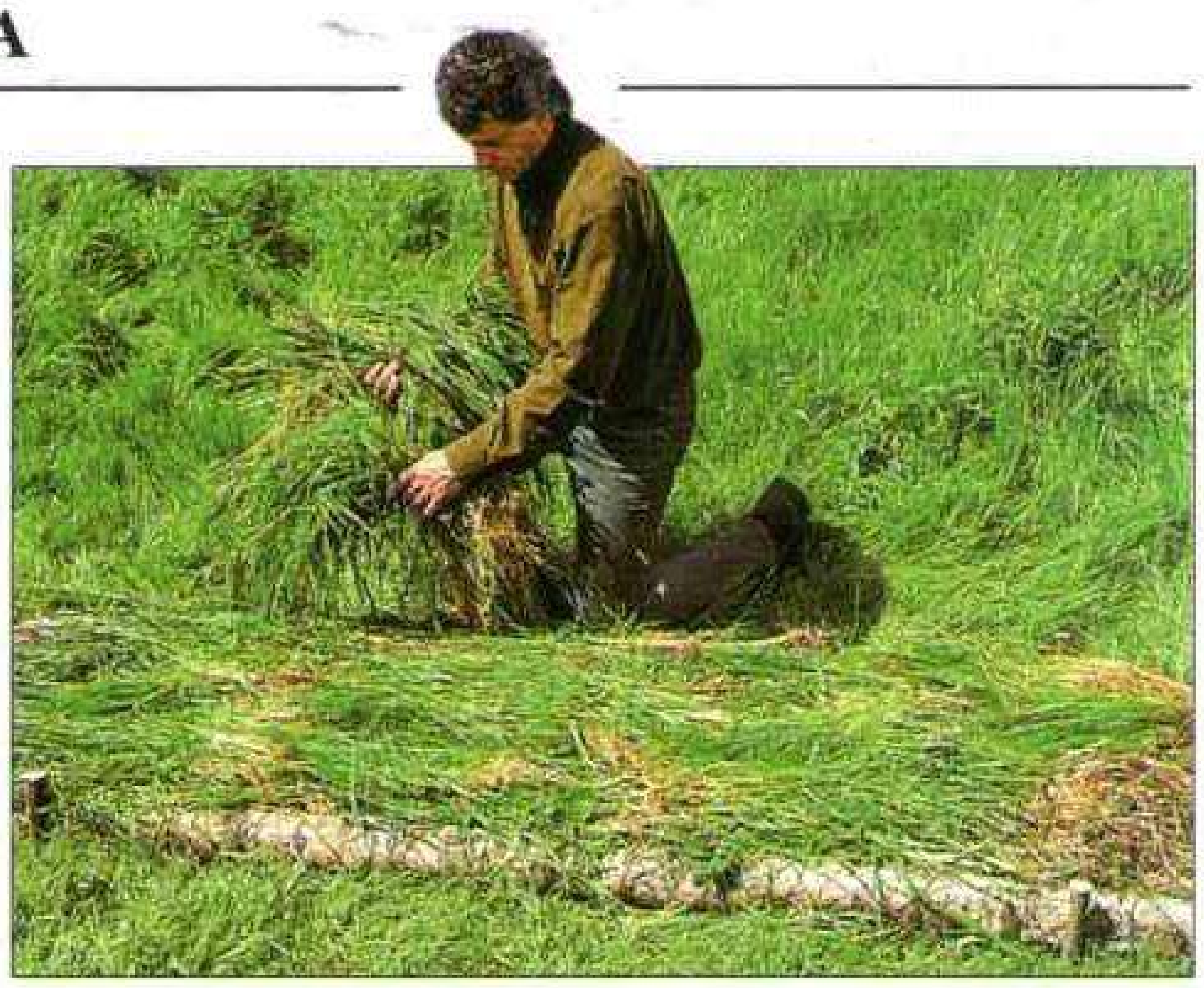

3 Cubra los troncos con una gruesa capa de hierba para esconder las irregularidades de los mismos debidas a la corteza y los nudos.

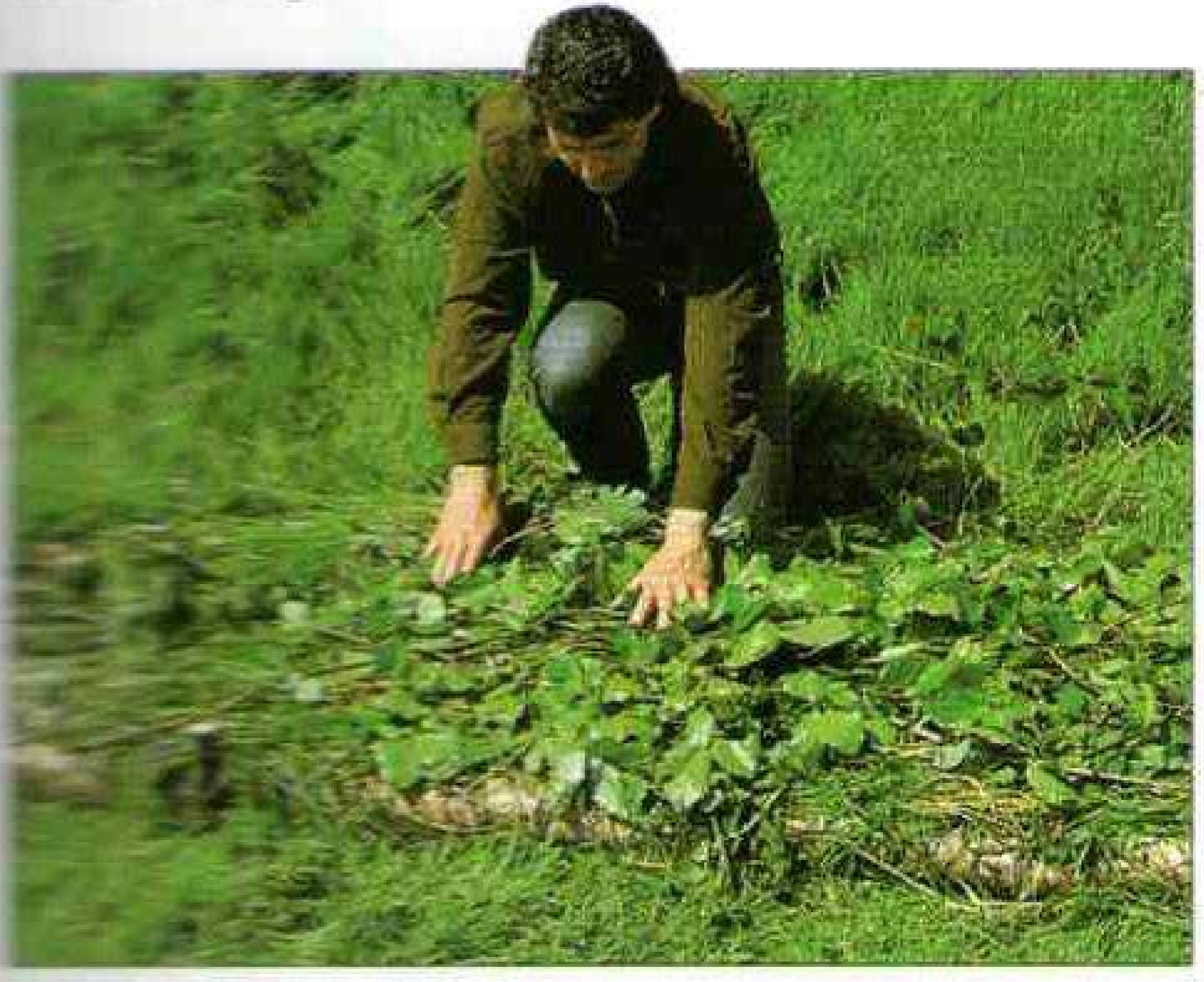

4 Añada más capas de ramas, helechos y follaje denso mezclado con hierbas y vegetación más fina para crear un espacio aireado y dar un efecto de profundidad. Debe construir un emparrado en el cual pueda recostarse cómodamente.

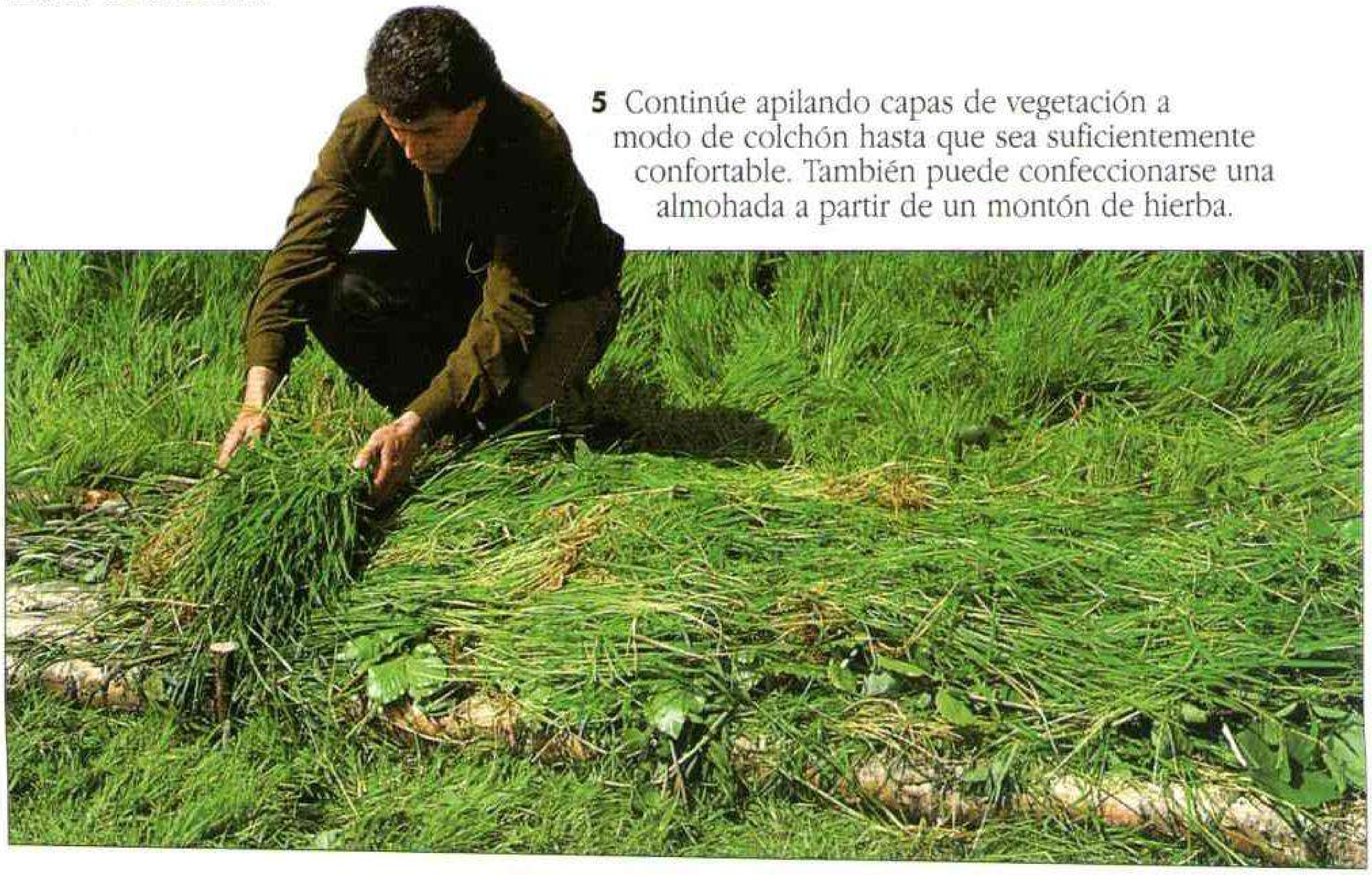

5 Continúe apilando capas de vegetación a modo de colchón hasta que sea suficientemente confortable. También puede confeccionarse una almohada a partir de un montón de hierba.

HERRAMIENTAS

Aunque esté acampado y tenga los últimos artilugios, o sobreviva a un desastre con un equipo mínimo, necesitará algunas herramientas con las que construir un refugio y cortar leña. En áreas vírgenes, no podrá reemplazar las herramientas rotas o perdidas hasta que vuelva a la civilización, por lo que debe cuidarlas. No haga mal uso de ellas de forma que pueda dañarlas, o lo que es más importante, que pueda dañarse a sí mismo. Puede improvisar algunas herramientas, aunque no se arriesgue a perder o estropear otra en el proceso. Trabaje siempre considerando las limitaciones de sus herramientas, así como su propia capacidad.

HERRAMIENTAS BÁSICAS PARA CORTAR MADERA U OTROS MATERIALES

La herramienta básica y esencial en un campamento y en una situación de supervivencia es un cuchillo. Aunque casi cualquier otra herramienta puede improvisarse, si no tiene un cuchillo le será muy difícil fabricar cualquier otro artículo.

Un hacha es muy valiosa para realizar trabajos pesados, pero debido a su peso es poco probable que lleve una.

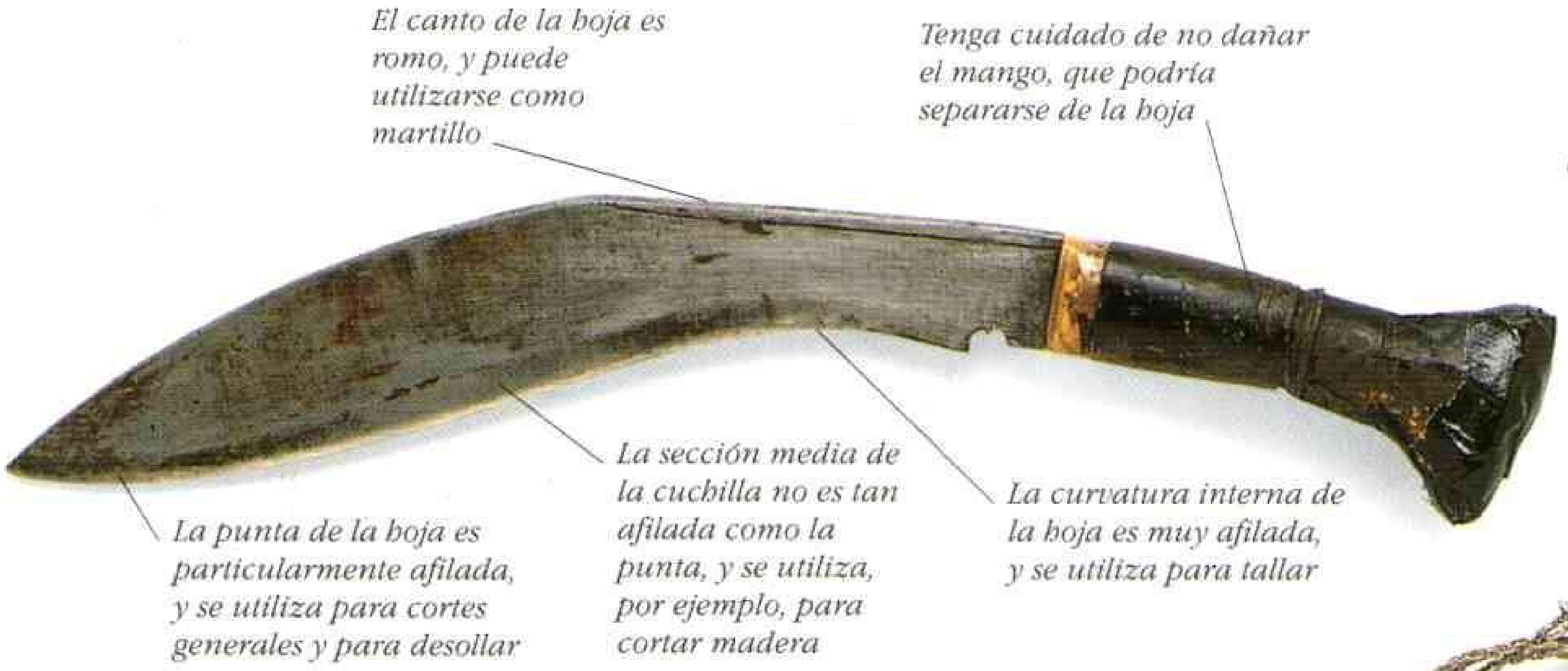
El canto de la hoja es romo, y puede utilizarse como martillo
Tenga cuidado de no dañar el mango, que podría separarse de la hoja
La punta de la hoja es particularmente afilada, y se utiliza para cortes generales y para desollar
La sección media de la cuchilla no es tan afilada como la punta, y se utiliza, por ejemplo, para cortar madera
La curvatura interna de la hoja es muy afilada, y se utiliza para tallar

Kukri. *Este cuchillo es empleado tradicionalmente por los gurkhas del Nepal. Está fabricado a partir de una ballesta trasera de un camión, de acero de alto grado de gran resistencia. En todo el mundo pueden encontrarse cuchillos similares. Si es posible, póngase unos guantes siempre que lo utilice para evitar heridas y rozaduras. Guárdelo en su funda de piel cuando no lo esté utilizando, para evitar herirse a sí mismo, o dañar el cuchillo.*

Navaja. *Una navaja resistente es la segunda herramienta en importancia después de sus dientes. Puede utilizarse en aquellos trabajos para los cuales su cuchillo es demasiado grande* (véase página siguiente)*. Manténgala sujeta a su cuerpo con un cordel fuerte alrededor del cuello o de la cintura en todo momento. Manténgala siempre afilada.*

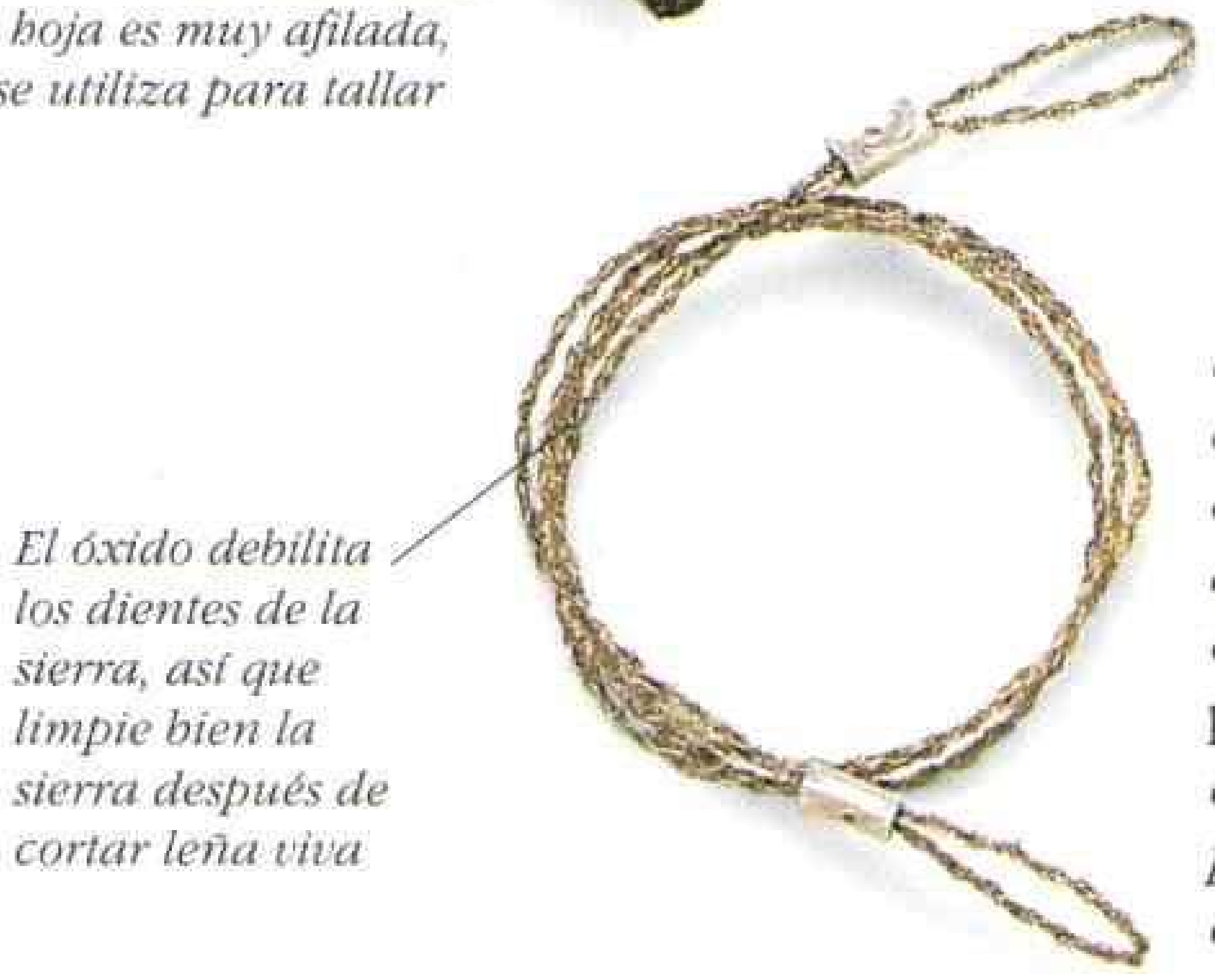
El óxido debilita los dientes de la sierra, así que limpie bien la sierra después de cortar leña viva

Cable sierra. *Un cable sierra consiste en varios hilos trenzados de alambre, con un par de lazadas como agarraderas. Puede incluirlo fácilmente en su equipo de supervivencia* (véase página 28)*. Es una alternativa muy útil a un hacha pesada para cortar ramas, pero es delicado y debe utilizarse con cuidado.*

AFILAR UN CUCHILLO

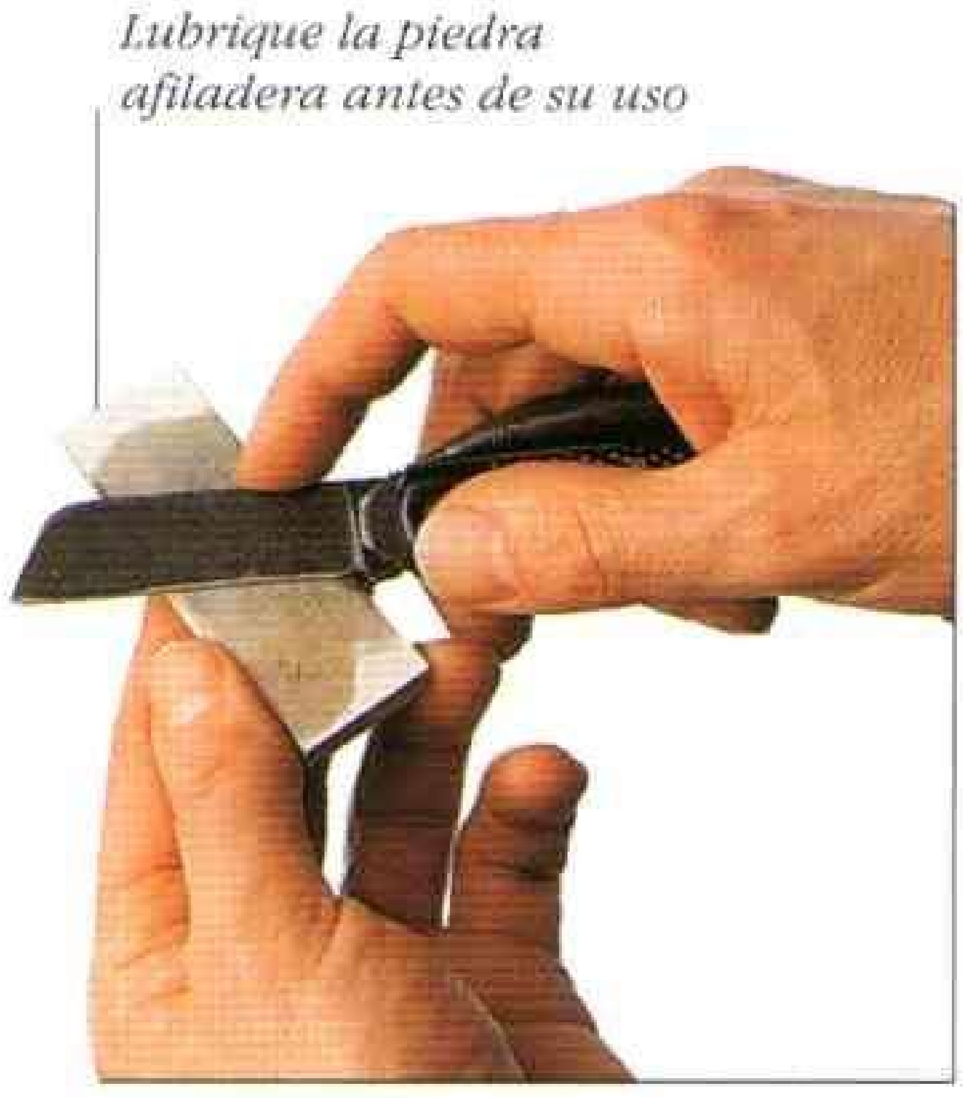
Lubrique la piedra afiladera antes de su uso

1 Humedezca la piedra afiladera con agua. Deslice el cuchillo sobre la piedra, en dirección opuesta al filo de la hoja.

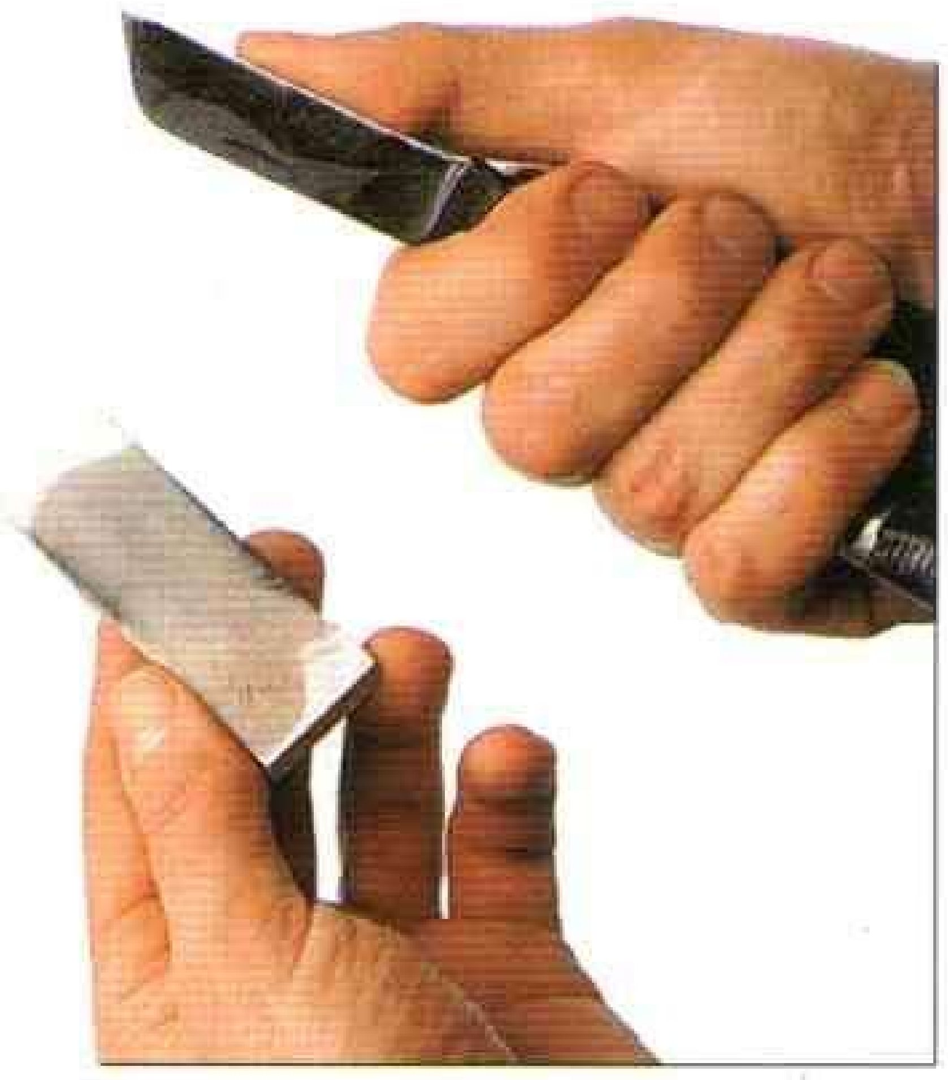

2 Después de afilar el cuchillo por una cara, toque la otra cara para hacer saltar las rebabas de metal provocadas por la abrasión.

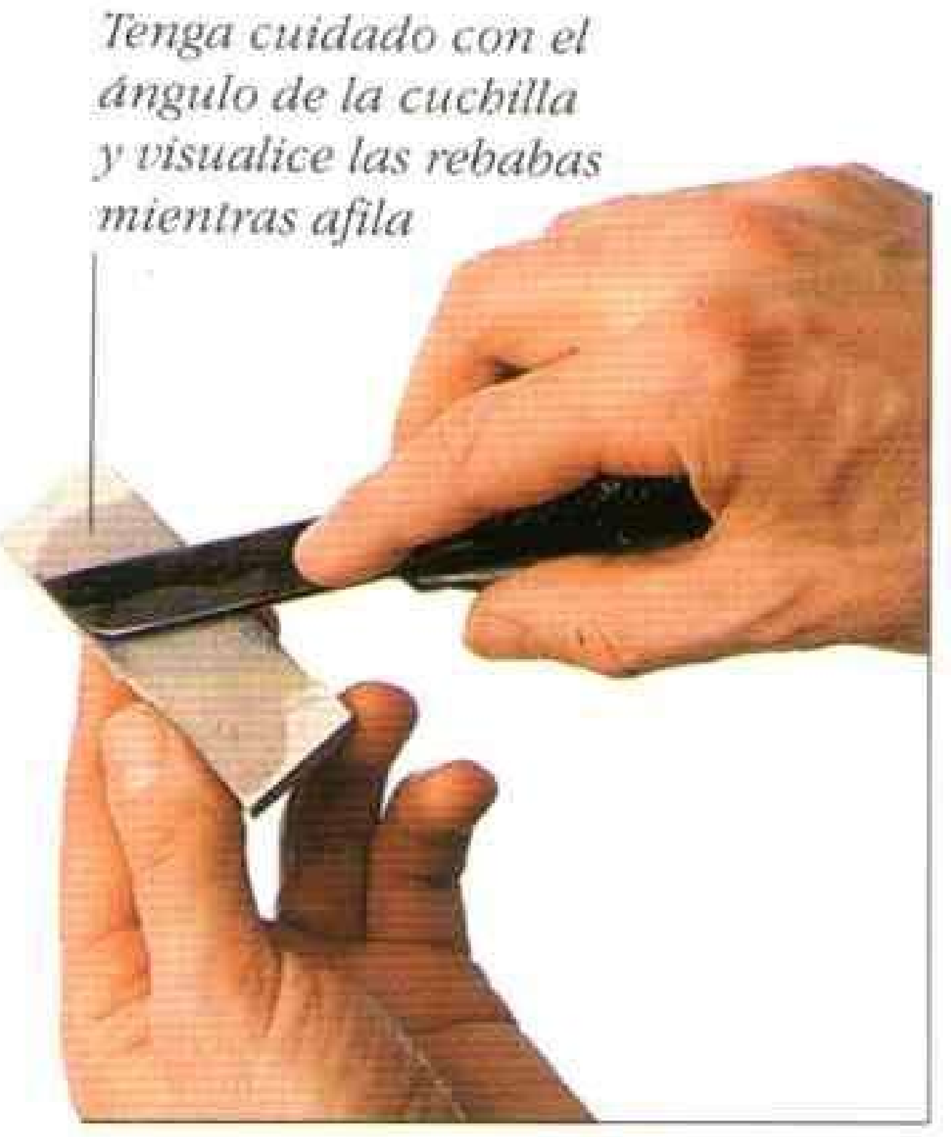
Tenga cuidado con el ángulo de la cuchilla y visualice las rebabas mientras afila

3 Afile la otra cara de la hoja, realineando la rebaba hacia el centro. Puede necesitar más agua.

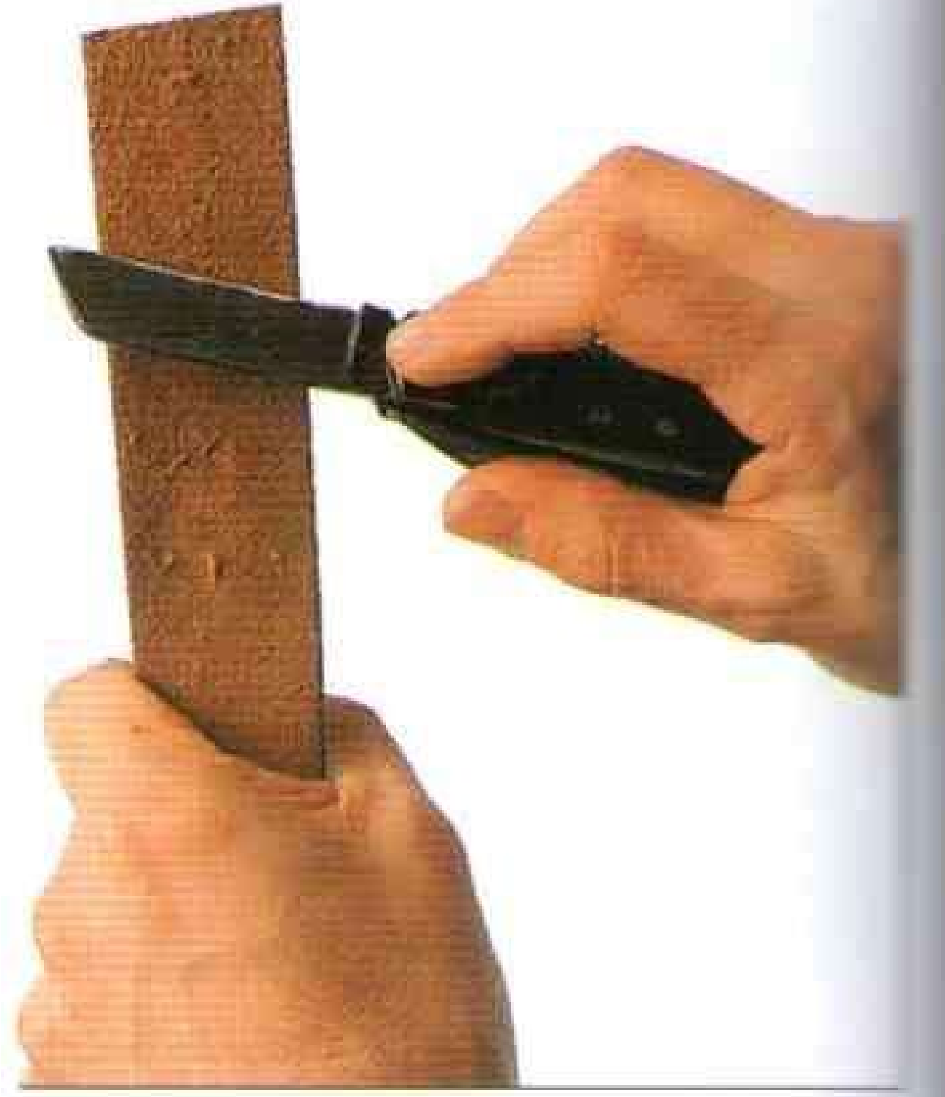

4 Suavice el cuchillo (deslizándolo arriba y abajo) sobre un cinturón de piel. Esto contribuirá a dar mayor uniformidad y a reforzar el filo.

EMPLEO DEL CUCHILLO

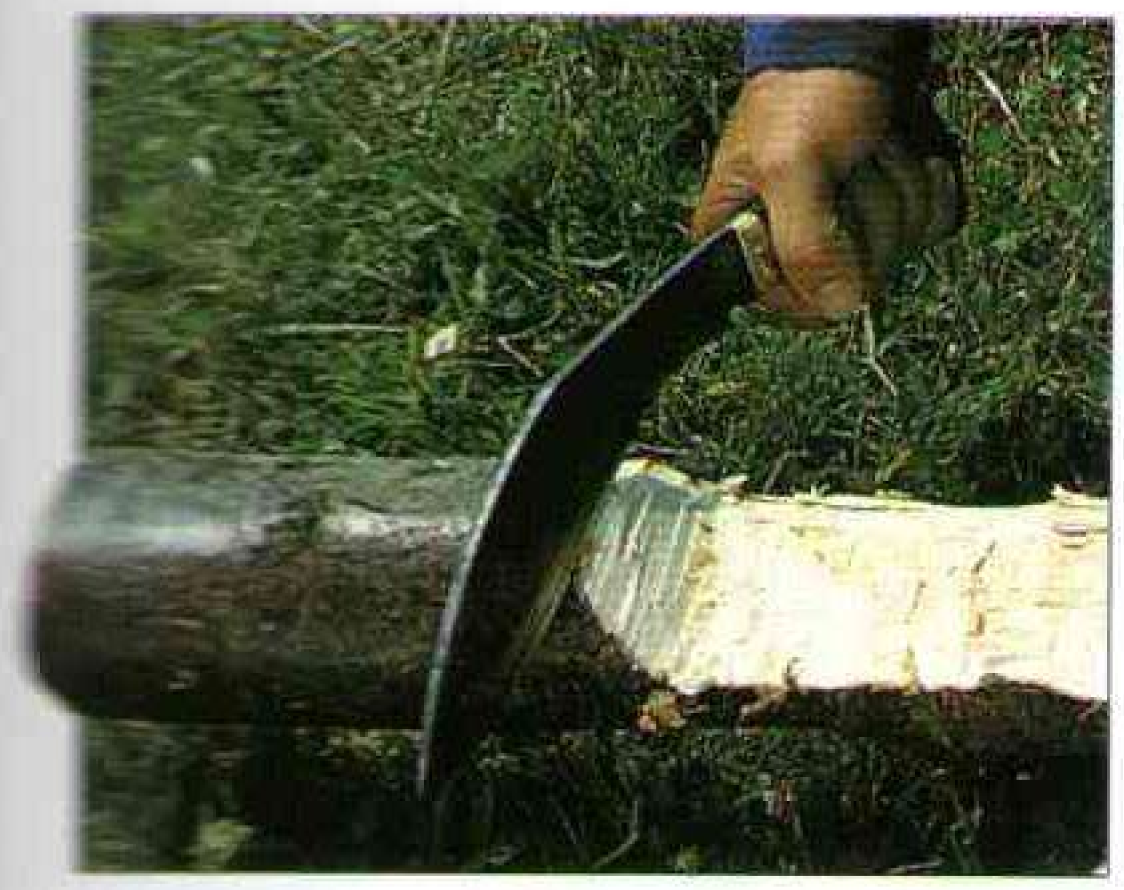

Cortar. *Corte con la sección principal de la hoja y mantenga sus brazos y piernas lo más alejados posible. Corte en dirección perpendicular a la fibra de la madera, con un ángulo de 45 grados para evitar que la hoja rebote y le hiera.*

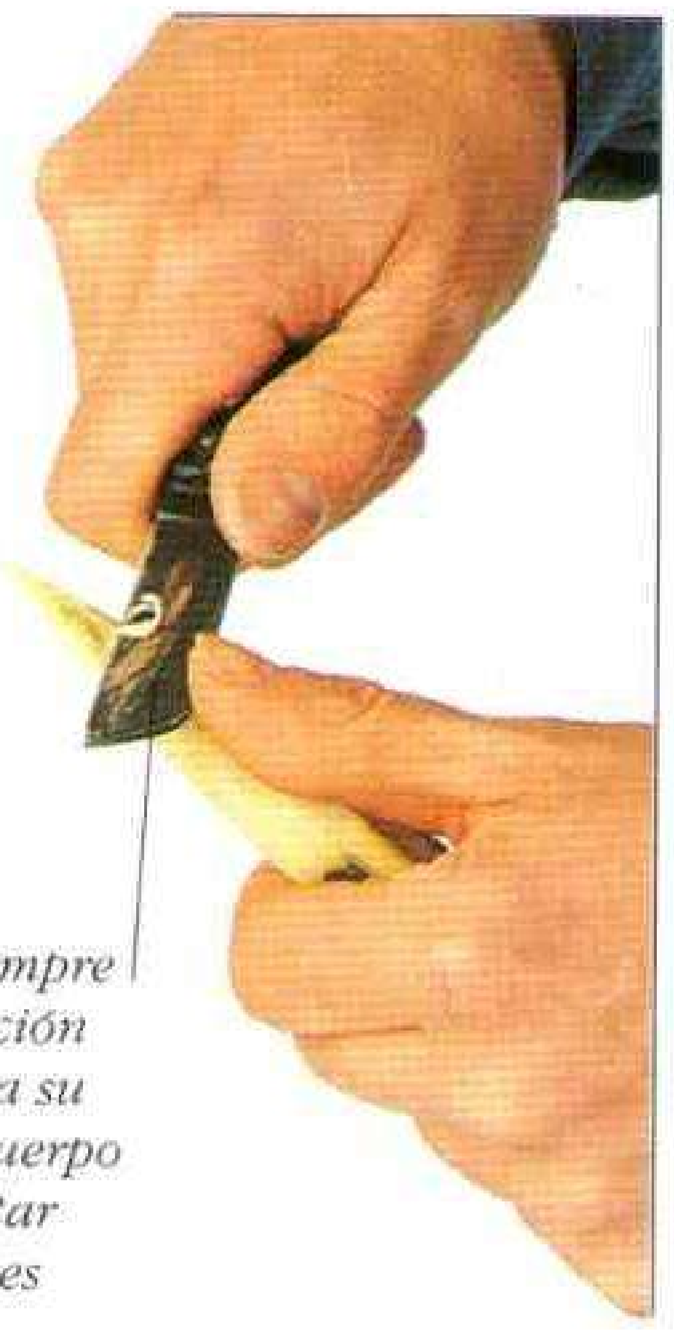

Corte siempre en dirección opuesta a su propio cuerpo para evitar accidentes

Tallar. *La talla es el arte de esculpir en madera pequeños objetos delicados, por ejemplo la punta afilada de un arpón para pescar* (véase página 104). *Para este tipo de trabajo, empuje la hoja cuidadosamente en sentido opuesto a su propio cuerpo con la yema del pulgar. Probablemente advertirá que a este efecto su pequeña navaja es mejor que el cuchillo.*

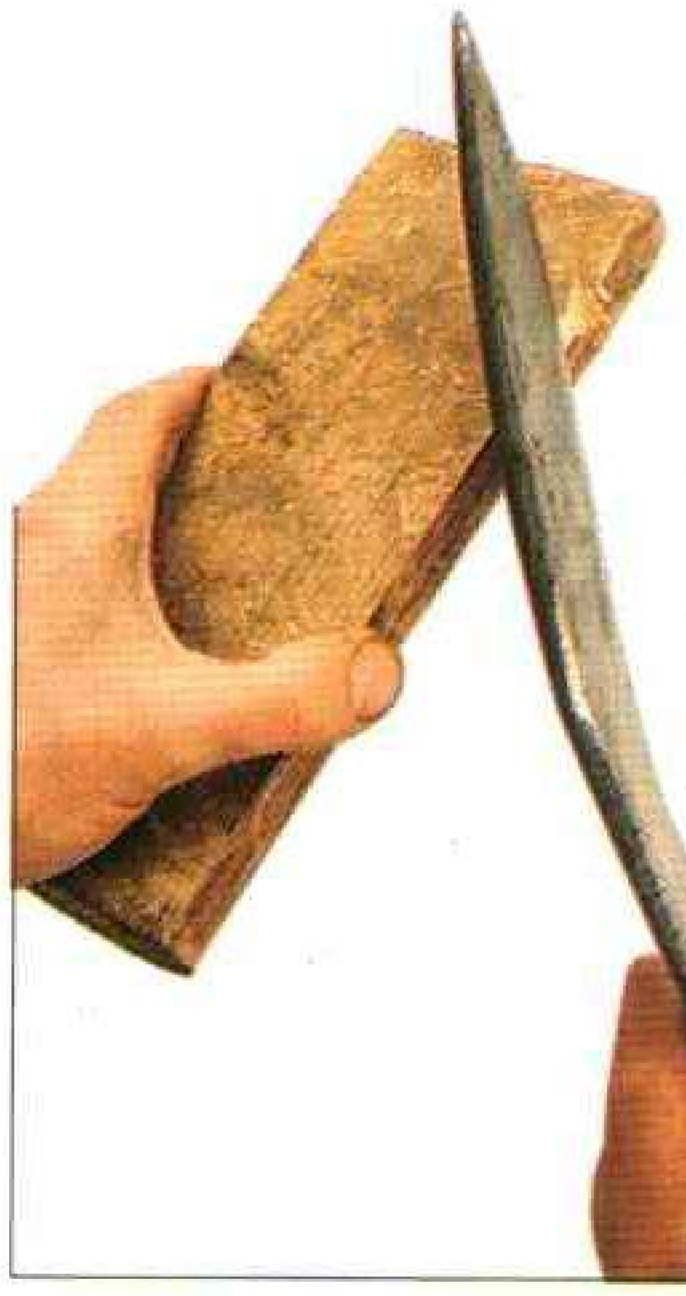

Cincelar. *El cincelado también debe realizarse en dirección opuesta a su cuerpo; tenga cuidado con sus dedos y miembros. Corte en el sentido de la fibra de la madera, dando pequeños golpes.*

EMPLEO DEL CABLE SIERRA

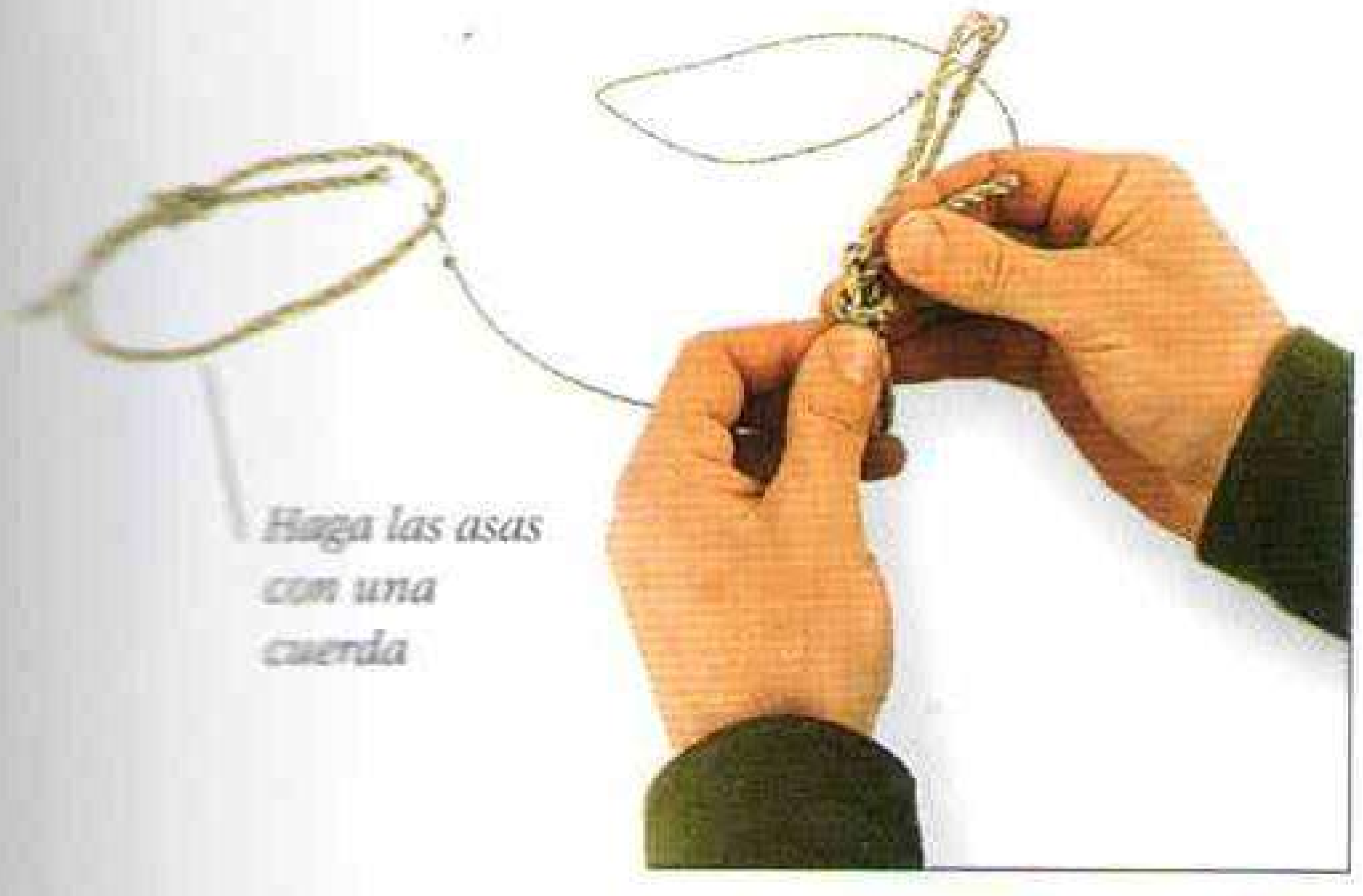

Haga las asas con una cuerda

1 Los lazos de metal pueden herir sus manos, por lo que debe añadir unas asas de cuerda o de tela a la sierra antes de utilizarla.

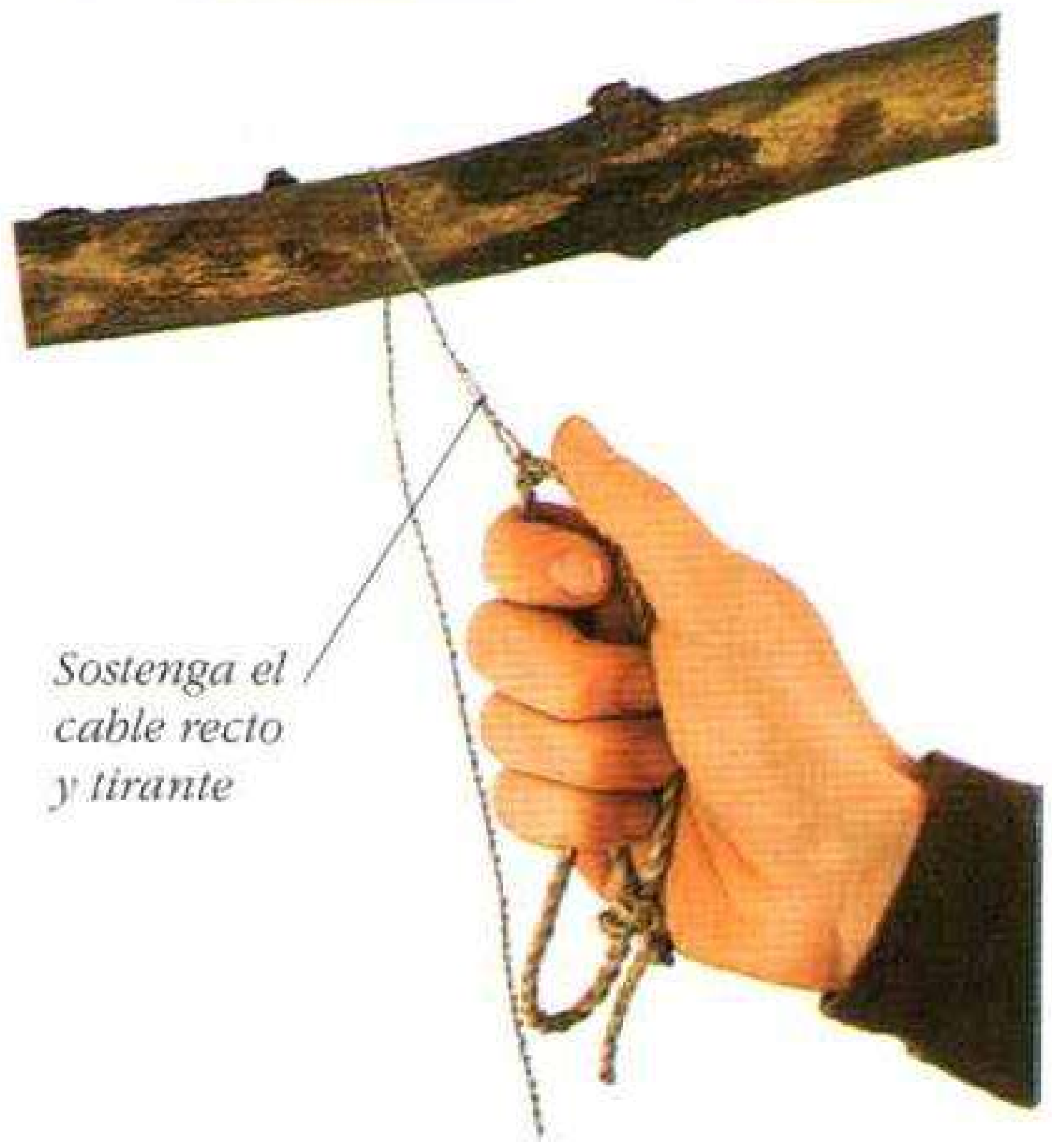

Sostenga el cable recto y tirante

2 Use ambas manos para mover la sierra y cortar la madera. Al cortar, mantenga el cable lo más recto posible, aunque cortará igual si lo curva alrededor de la madera. Mantenga un movimiento regular, pero ligero.

PRECAUCIÓN

Mantenga sus herramientas siempre afiladas, y tenga mucho cuidado al utilizarlas. Cuando saque un cuchillo de su funda, sosténgala por la parte correspondiente al lomo de la hoja. Un cuchillo afilado puede cortar la funda y seccionar los tendones de la mano. Un cable sierra también puede herirle si no lo utiliza adecuadamente.

HERRAMIENTAS IMPROVISADAS

1 Puede obtener una cuchilla natural dejando caer un trozo de pedernal sobre una roca. Al caer, se partirá, dejando los contornos afilados. Tenga cuidado con las astillas.

Un trozo de pedernal puede ser utilizado como cuchillo o como martillo

2 El borde afilado puede ser utilizado como una rudimentaria herramienta cortante. El otro extremo puede usarse como martillo. Cuando el borde pierda su filo, rompa otro de pedernal.

OTROS MATERIALES

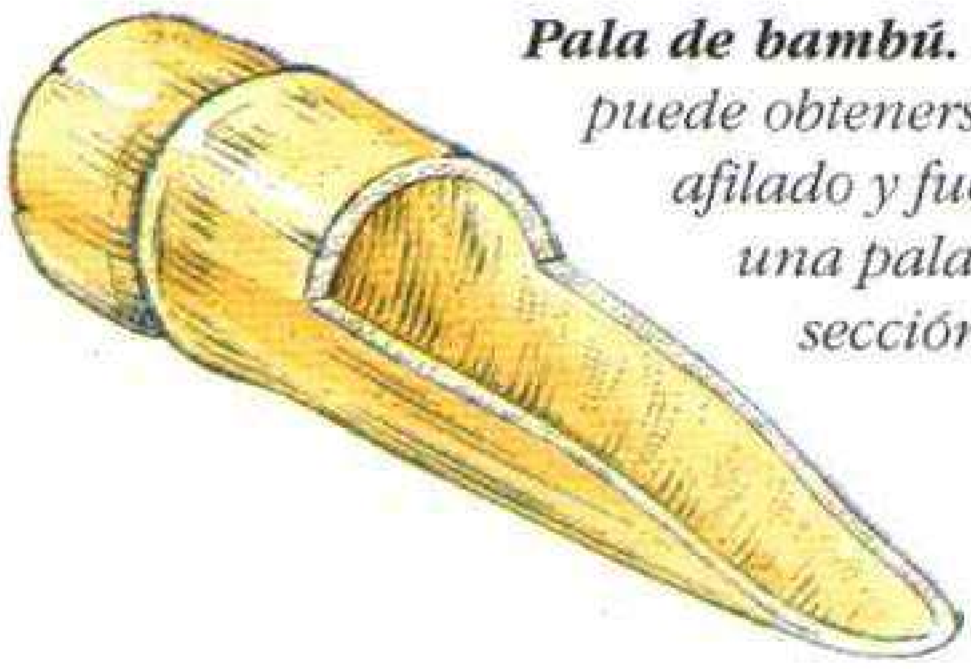

Pala de bambú. *Del bambú puede obtenerse un canto muy afilado y fuerte. Construya una pala cortando una sección del tallo.*

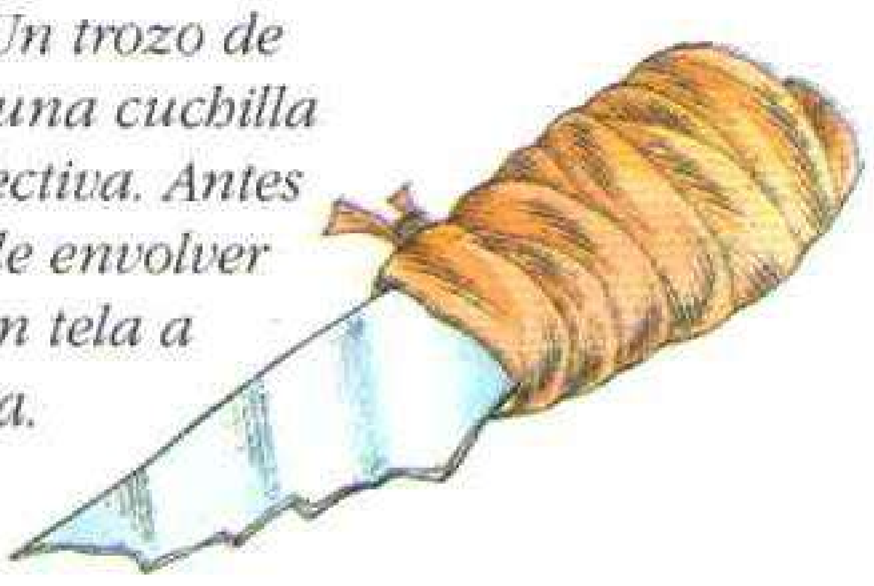

Cuchillo de vidrio. *Un trozo de vidrio roto puede ser una cuchilla rudimentaria pero efectiva. Antes de usarlo, asegúrese de envolver uno de los extremos en tela a modo de empuñadura.*

CUERDAS Y NUDOS

LAS SOGAS Y cuerdas son imprescindibles para improvisar cualquier equipo, a menos que fabrique pinzas de madera o utilice alguna cola para unir las piezas de distintos materiales *(véase página 77)*. Si carece de cuerda o soga, puede fabricarse una a partir de materiales naturales. Utilizando los nudos adecuados, acordes a sus necesidades, es esencial –por su propia seguridad– que lo que haya hecho se mantenga entero, tanto si está usando una cuerda para cruzar un río como para asegurar el éxito de cualquier equipo o instalación que improvise. Si sólo es capaz de recordar un nudo, aprenda el nudo plano *(véase página siguiente)*.

LA CONFECCIÓN DE CUERDAS

Hay dos maneras principales de confeccionar una cuerda. La cuerda de tres cabos consta de tres hilos trenzados, mientras que el tipo *Kernmantle* tiene un núcleo central rodeado por una envuelta exterior tejida. Ambos tipos pueden fabricarse a partir de fibras naturales o sintéticas. El tipo de cuerda que elija dependerá del uso que pretenda darle. Es menos probable que se retuerza una cuerda tejida que una trenzada, mientras que una cuerda fabricada con materiales naturales puede pudrirse cuando se humedece. Las cuerdas de fibras naturales son más propensas a romperse que las sintéticas de nailon o poliéster.

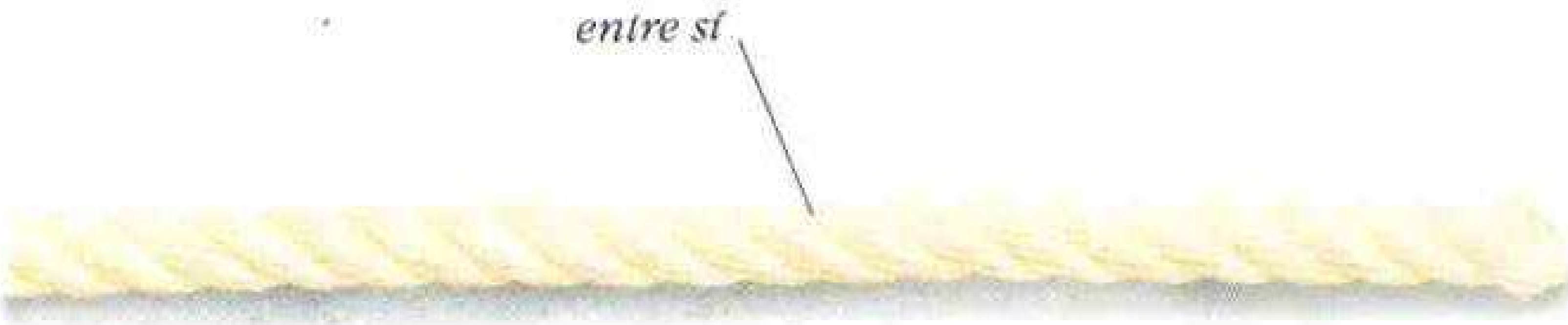
Tres diferentes ramales de hilos se entrelazan entre sí

Cuerda de tres cabos. *Una cuerda de tres cabos es más resistente que una cuerda* Kernmantle, *particularmente cuando se desgasta. Consta de tres hilos de fibra trenzados. Puede fabricarse con materiales naturales como el henequén, el cáñamo o la cáscara de coco, o con materiales sintéticos como el nailon. Las cuerdas de tres cabos se utilizan a menudo para amarrar o anclar barcos, debido a su resistencia.*

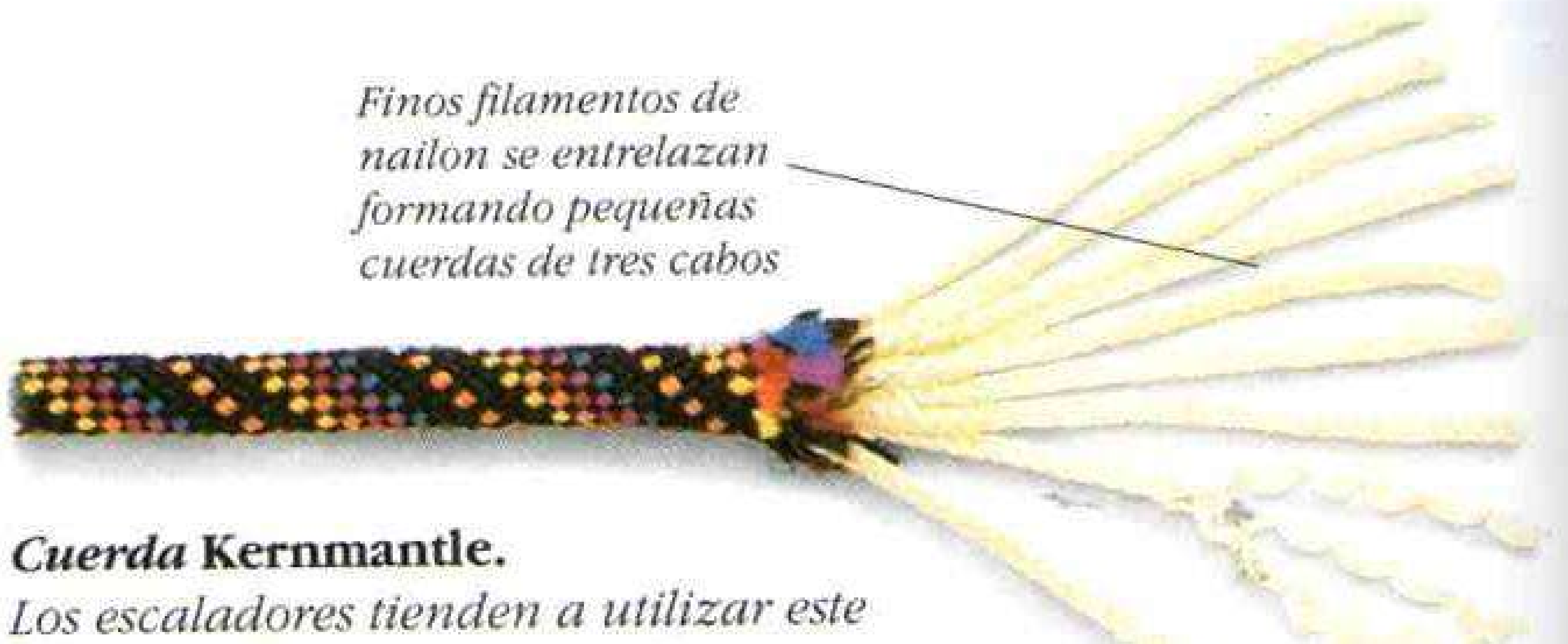
Finos filamentos de nailon se entrelazan formando pequeñas cuerdas de tres cabos

***Cuerda* Kernmantle.** *Los escaladores tienden a utilizar este tipo de cuerda, confeccionada a partir de un núcleo central de cuerdas finas de tres cabos entretejidas (el alma o* kern *en inglés), cubierto por una resistente envoltura exterior (la funda o* mantle *en inglés). Este tipo de cuerda es de fácil manejo, pero difícil de asir si está mojada o helada.*

CORDAJE IMPROVISADO

Para fabricar una cuerda a partir de tallos de ortiga, debe primero remojarlos para que adquieran mayor flexibilidad. Golpéelos después contra una piedra para separar las fibras de la médula, y deje secar las fibras. Podrá entonces entrelazarlos para formar cuerdas, con las que a su vez podrá formar un cordaje. También puede confeccionar cuerdas a partir de la corteza del sauce o del tilo.

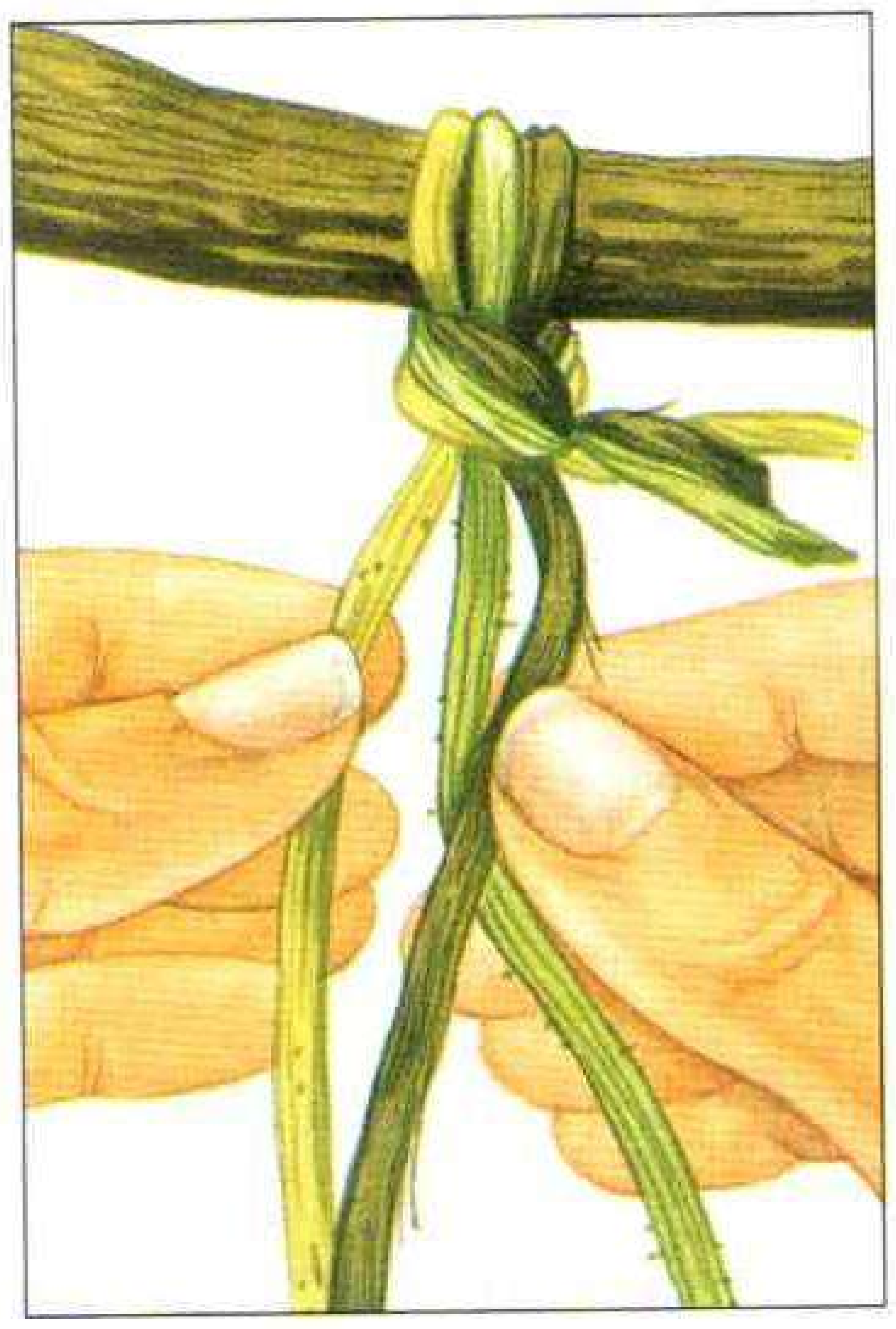

1 Para trenzar las fibras, ate primero los cabos de tres hebras a una rama. Pase la hebra de la derecha por encima de la central.

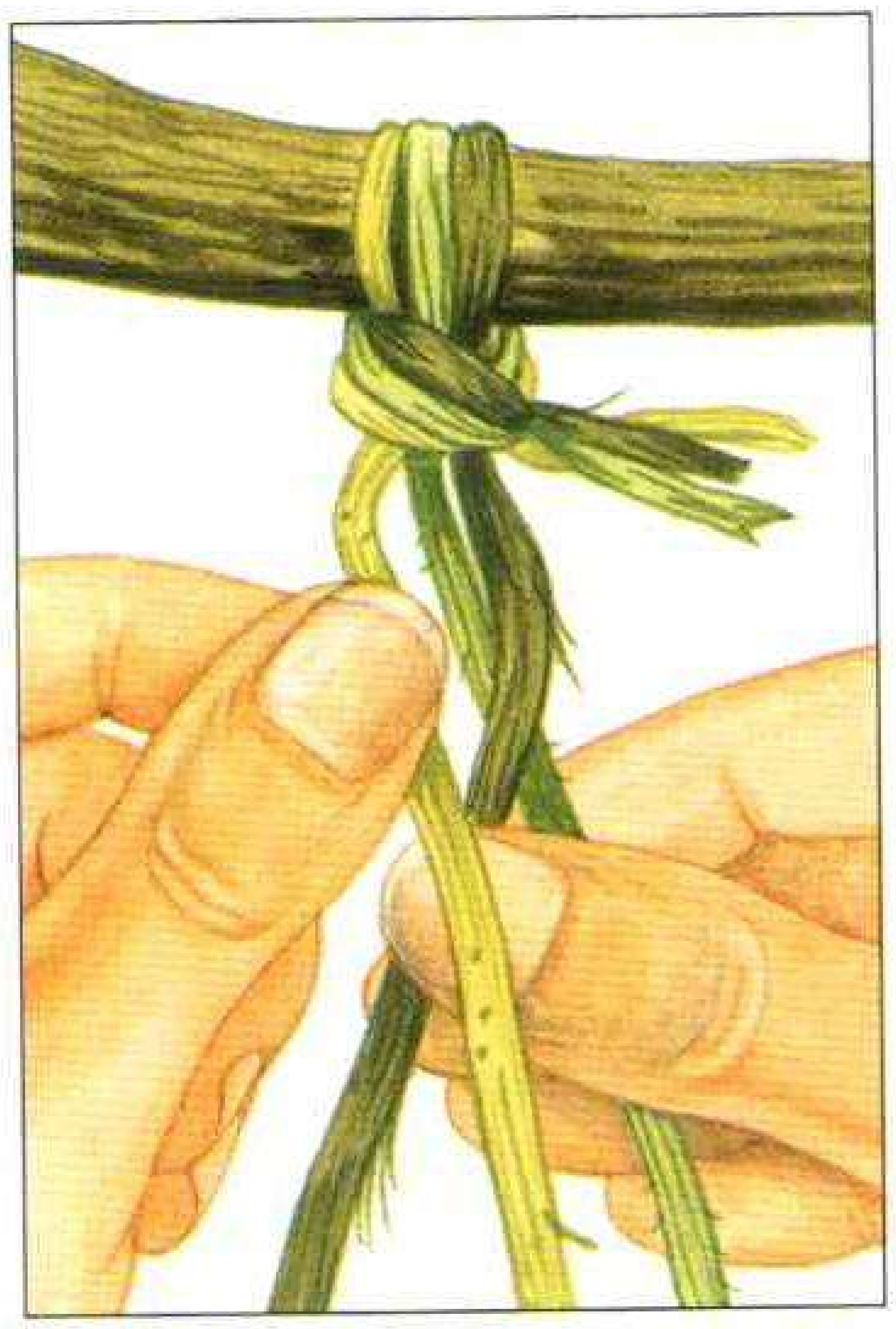

2 A continuación, pase la hebra izquierda por encima de la nueva hebra central, sin dejar de mantener tensas todas las hebras.

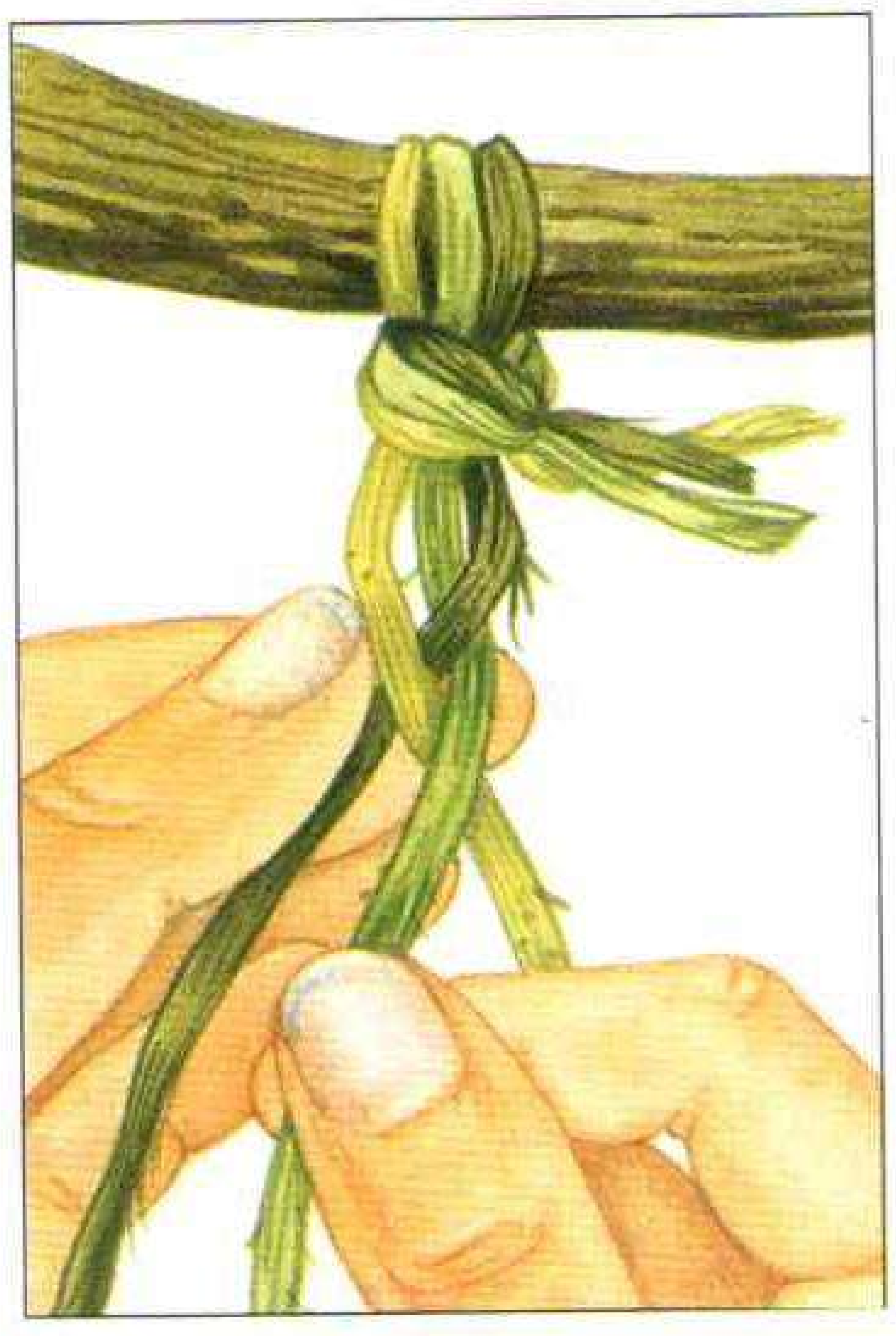

3 Pase la nueva hebra derecha por encima de la nueva hebra central. Con este paso terminará la primera sección de la trenza.

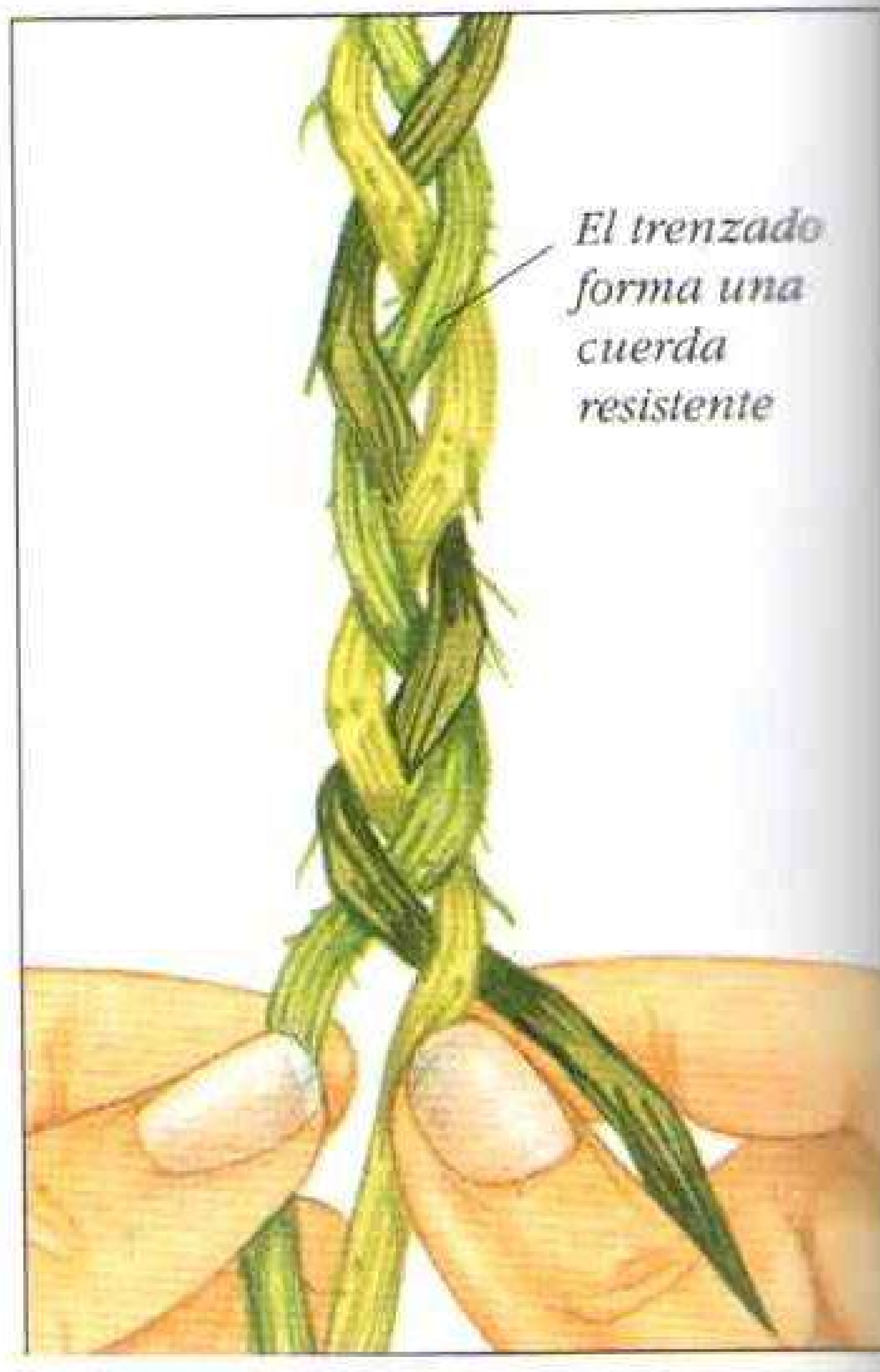
El trenzado forma una cuerda resistente

4 Repita los pasos hasta que la trenza tenga la longitud necesaria. Ate los cabos con fuerza para que no se deshaga la trenza.

NUDO PLANO

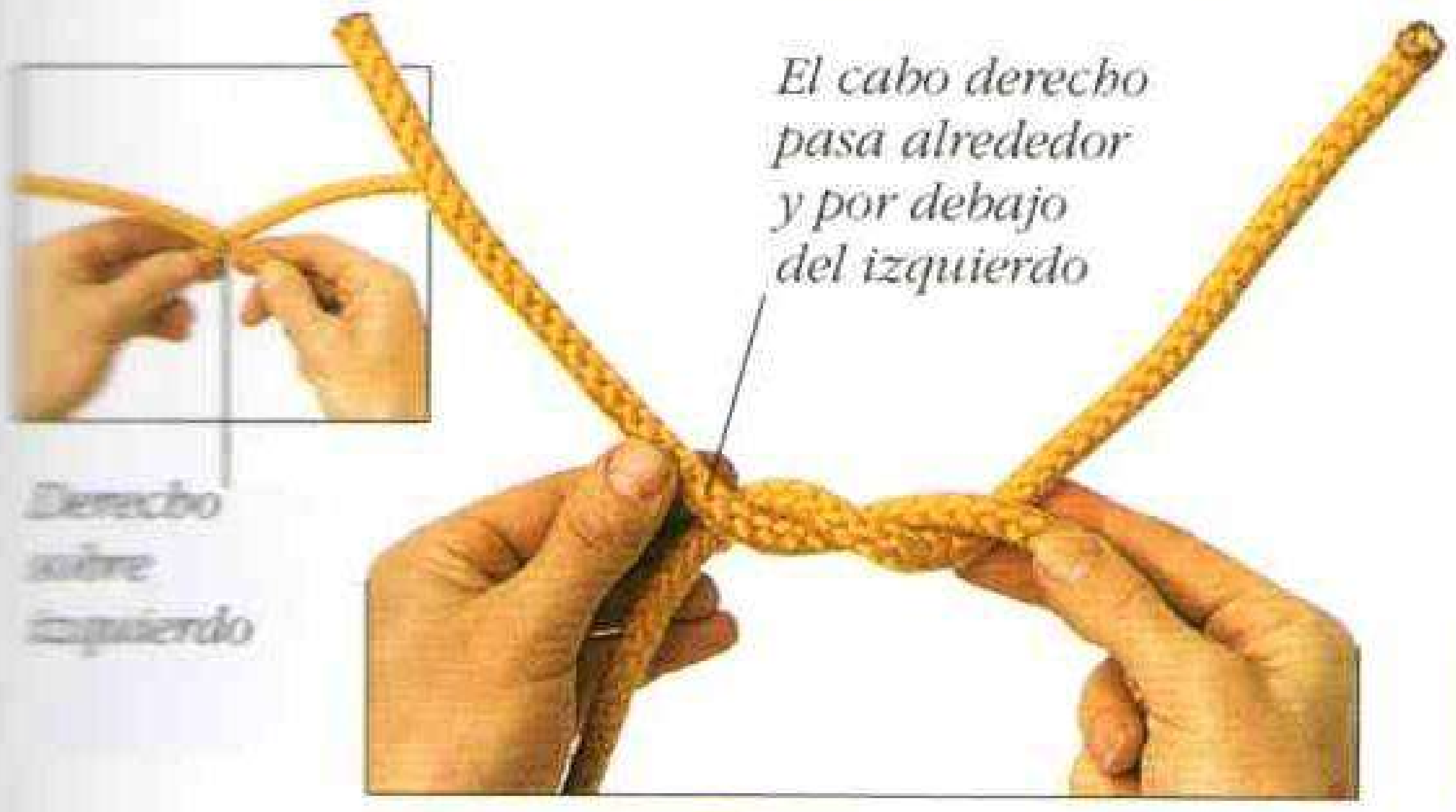
El cabo derecho pasa alrededor y por debajo del izquierdo
Derecho sobre izquierdo

1 Para comenzar el nudo plano, pase el cabo derecho por encima y por debajo del izquierdo.

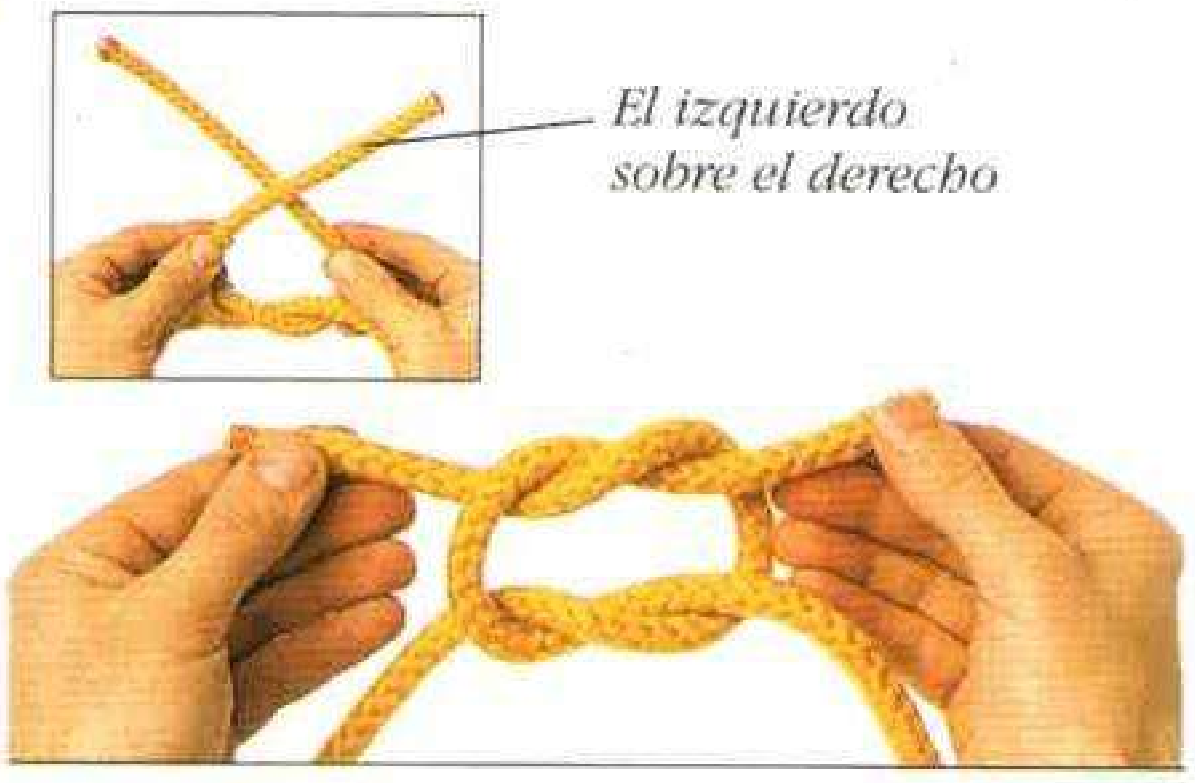
El izquierdo sobre el derecho

2 Pase el cabo izquierdo por encima y por debajo del cabo derecho, sujetando con fuerza ambos cabos.

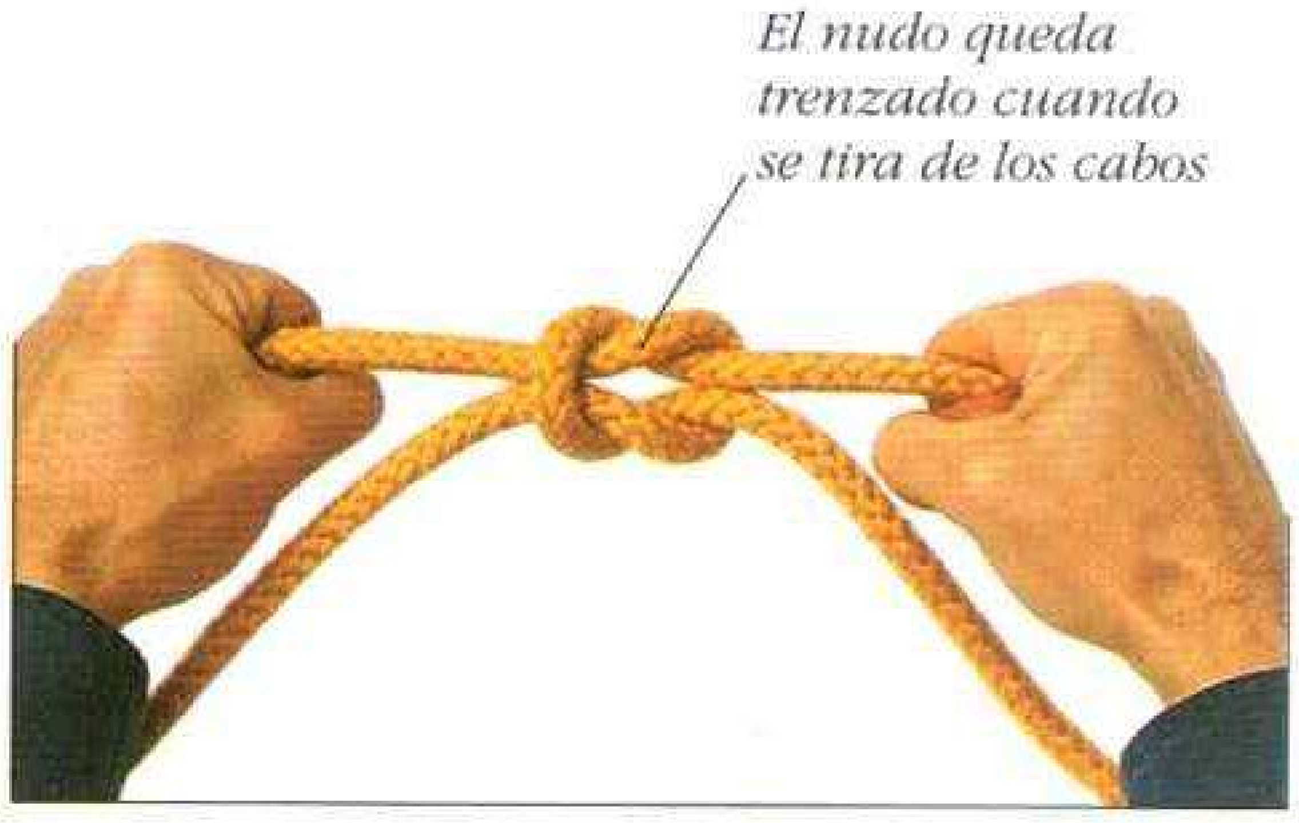
El nudo queda trenzado cuando se tira de los cabos

3 Tire de ambos cabos para tensar el nudo. El nudo puede hacerse en el orden inverso (izquierdo sobre derecho, derecho sobre izquierdo), y puede deshacerse con facilidad.

BOULIN

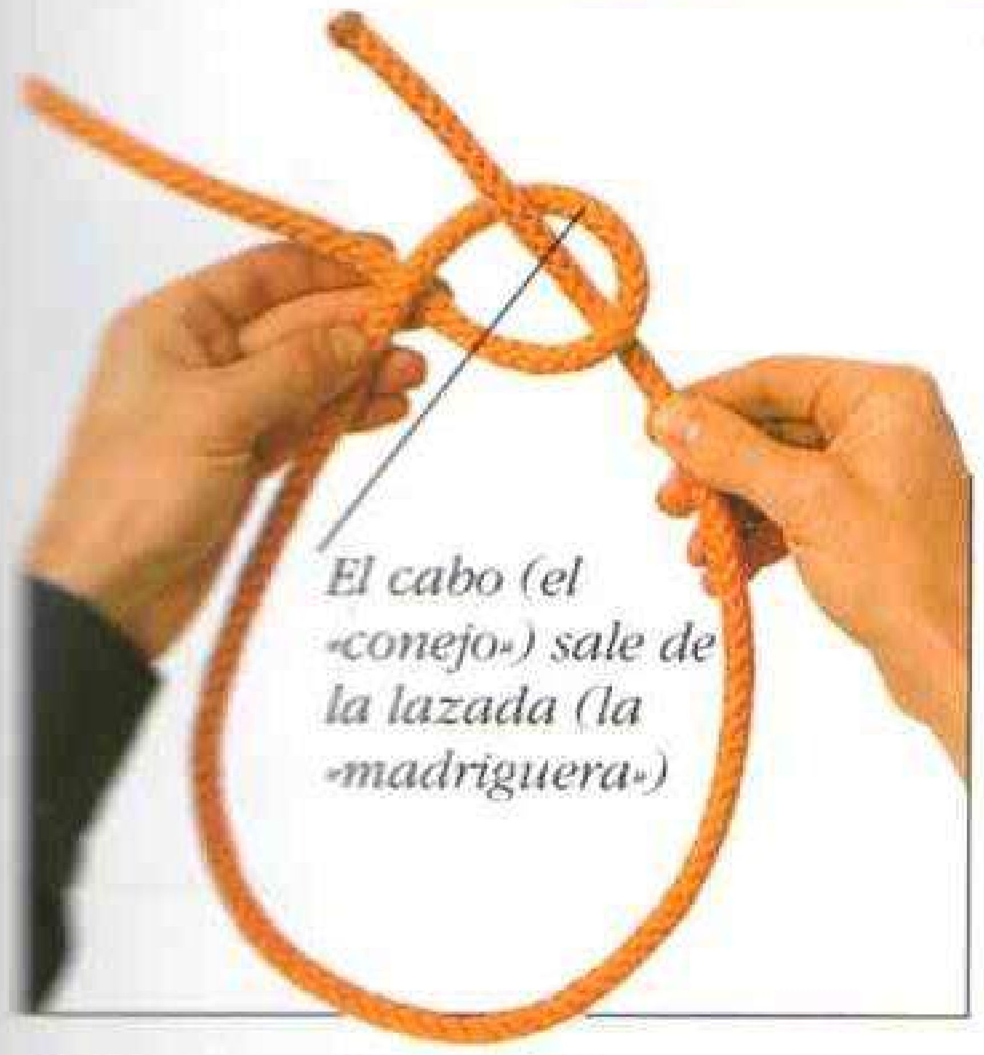
El cabo (el «conejo») sale de la lazada (la «madriguera»)

1 Si se emplea para hacer un lazo, un boulin nunca se tensará o se aflojará. Haga una pequeña lazada y pase el cabo a través de ella por detrás.

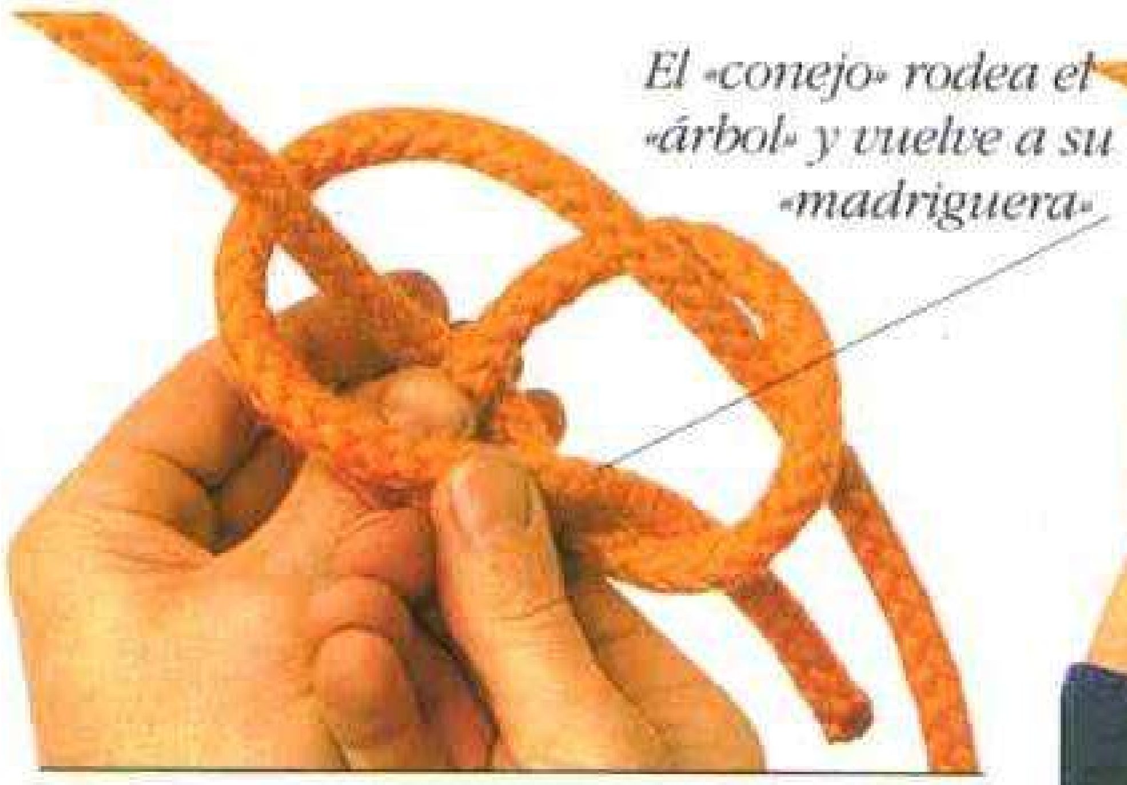
El «conejo» rodea el «árbol» y vuelve a su «madriguera»

2 Pase el cabo alrededor de la cuerda principal, y a continuación nuevamente a través de la lazada (la secuencia le resultará mucho más sencilla si recuerda: «el conejo sale de la madriguera, rodea el árbol y vuelve a entrar en la madriguera»).

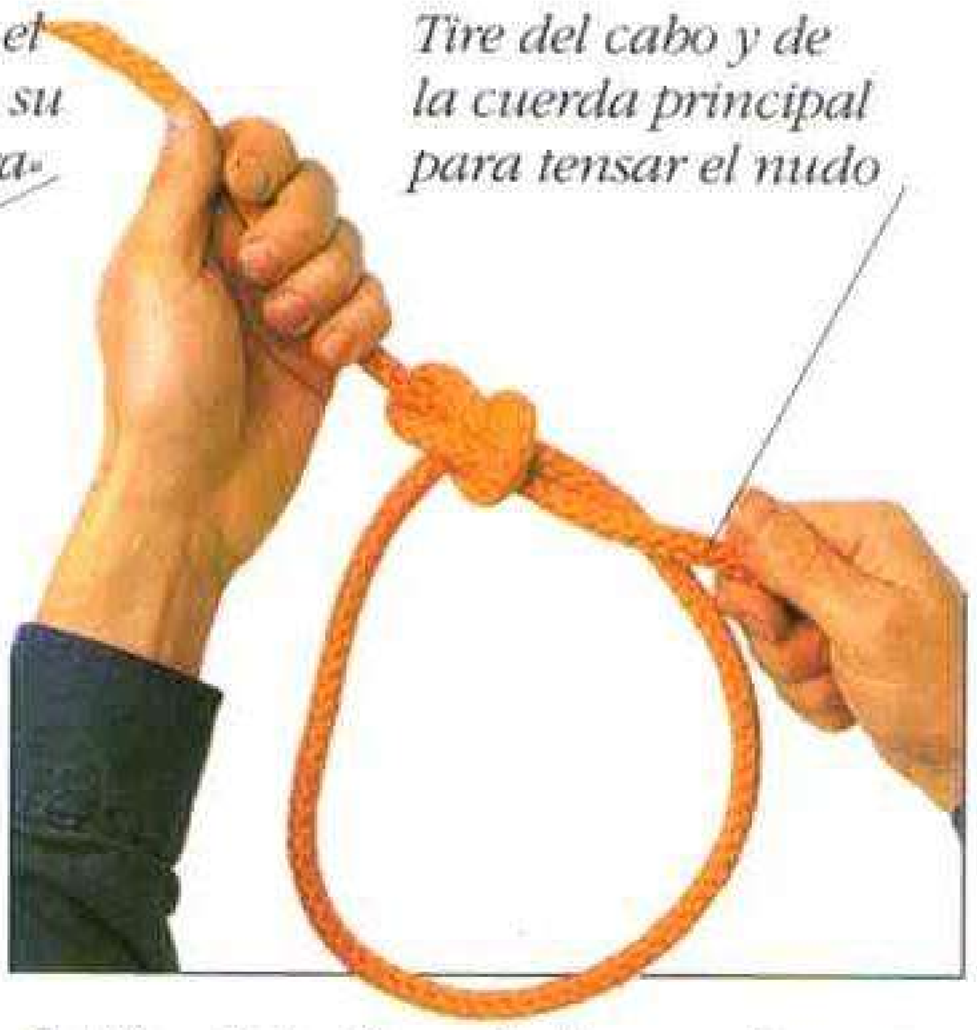
Tire del cabo y de la cuerda principal para tensar el nudo

3 Tire del cabo y de la cuerda principal para tensar el nudo. El boulin resulta útil, pero muchos escaladores prefieren el nudo en ocho para salvamento.

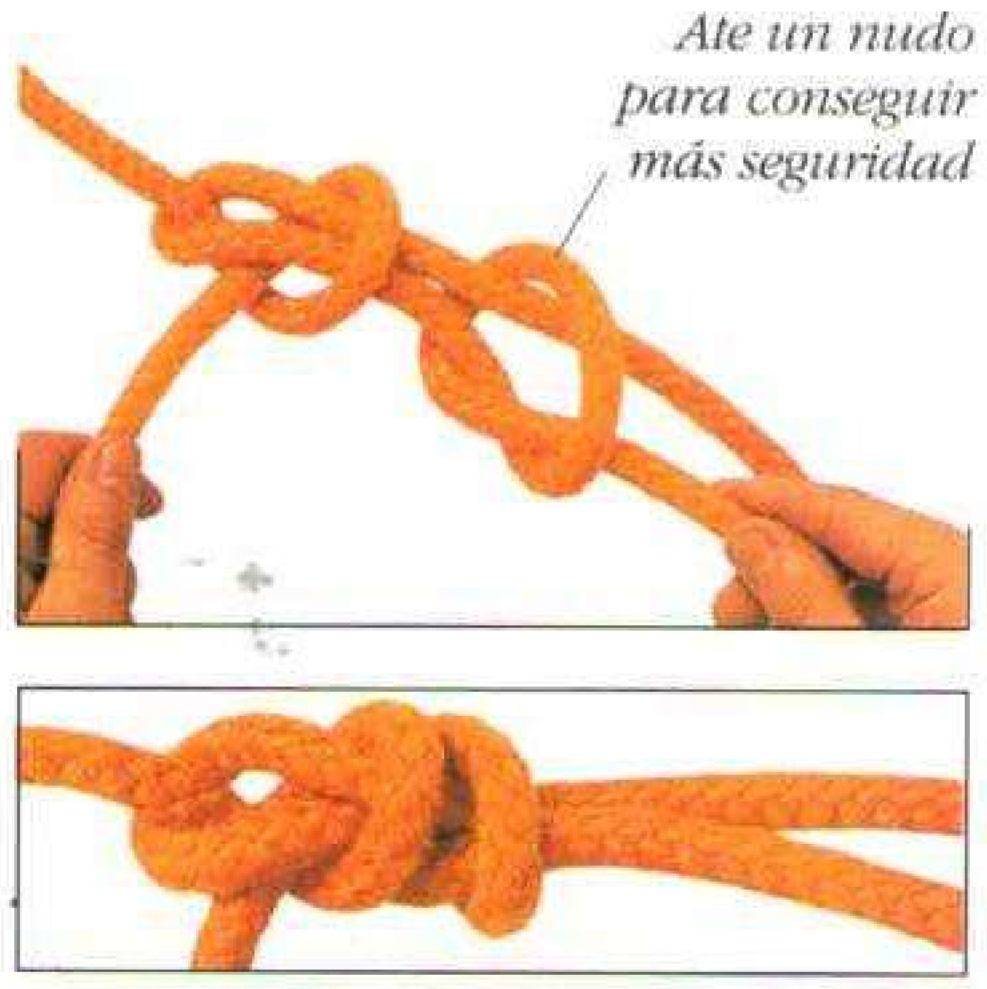
Ate un nudo para conseguir más seguridad

4 Para más seguridad, lleve el extremo sobre la parte derecha de la lazada y a través de ella, y a continuación nuevamente a través de la nueva lazada pequeña desde la parte superior.

NUDO EN OCHO

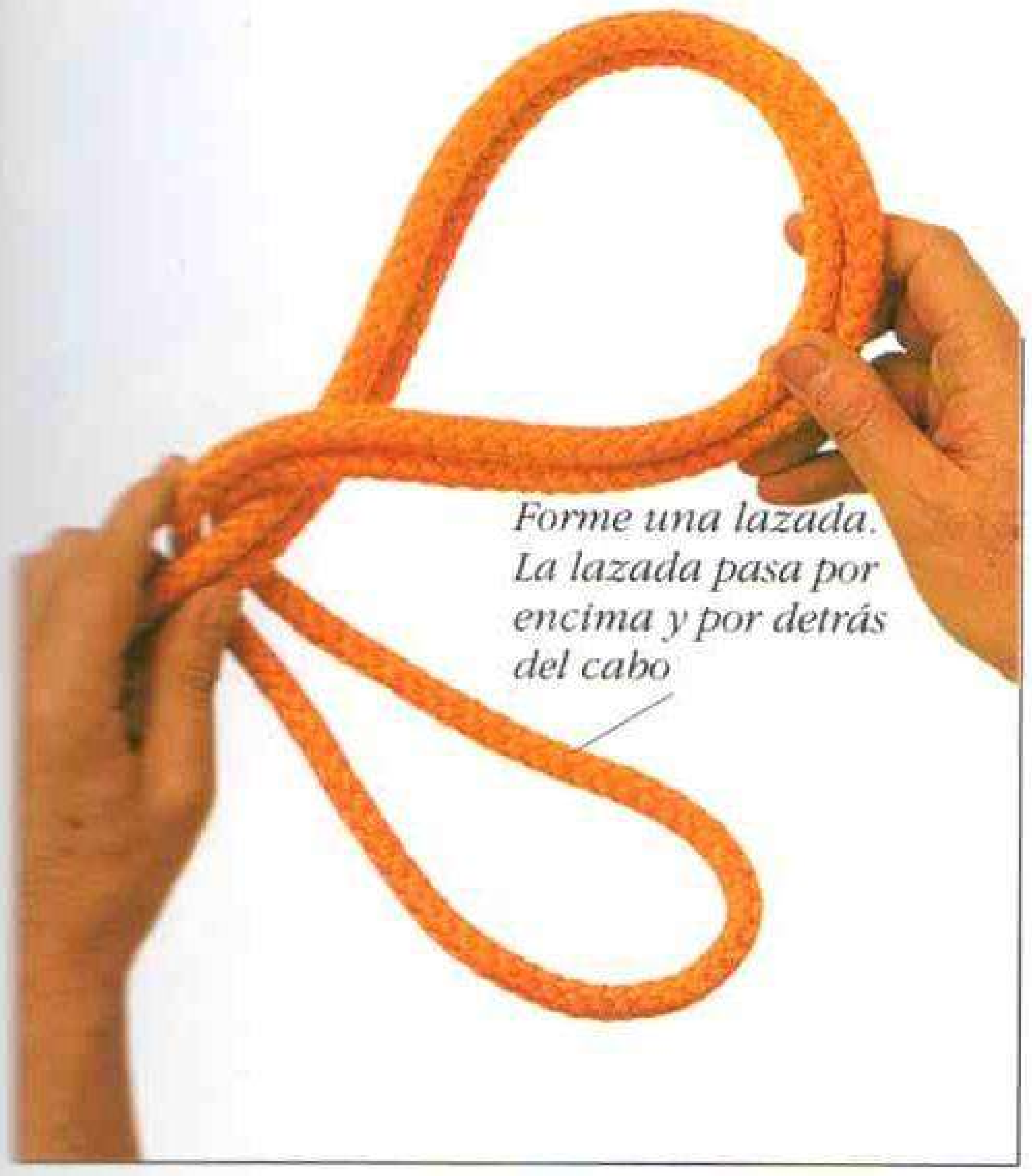
Forme una lazada. La lazada pasa por encima y por detrás del cabo

1 Este nudo es muy útil para hacer un lazo alrededor de un objeto. Forme una lazada doble de 1 m de longitud. Lleve el cabo libre por encima de la cuerda, y páselo a través de la lazada, formando la figura de un ocho.

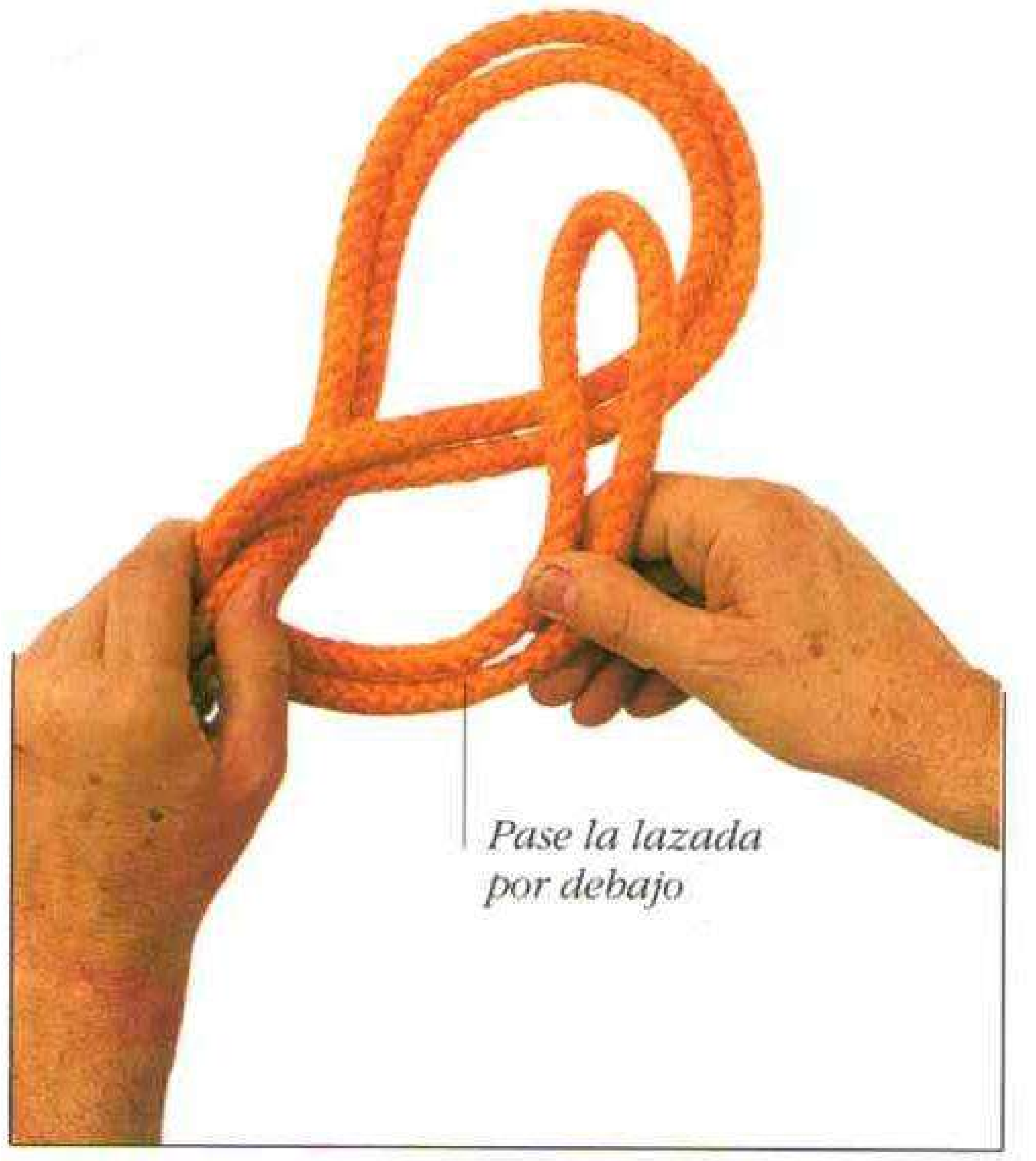
Pase la lazada por debajo

2 Tire de ambos cabos para tensar el nudo. Para formar el lazo, pase el cabo libre de la cuerda alrededor del objeto que quiera asegurar, y vuélvalo a introducir a continuación dentro del nudo en ocho.

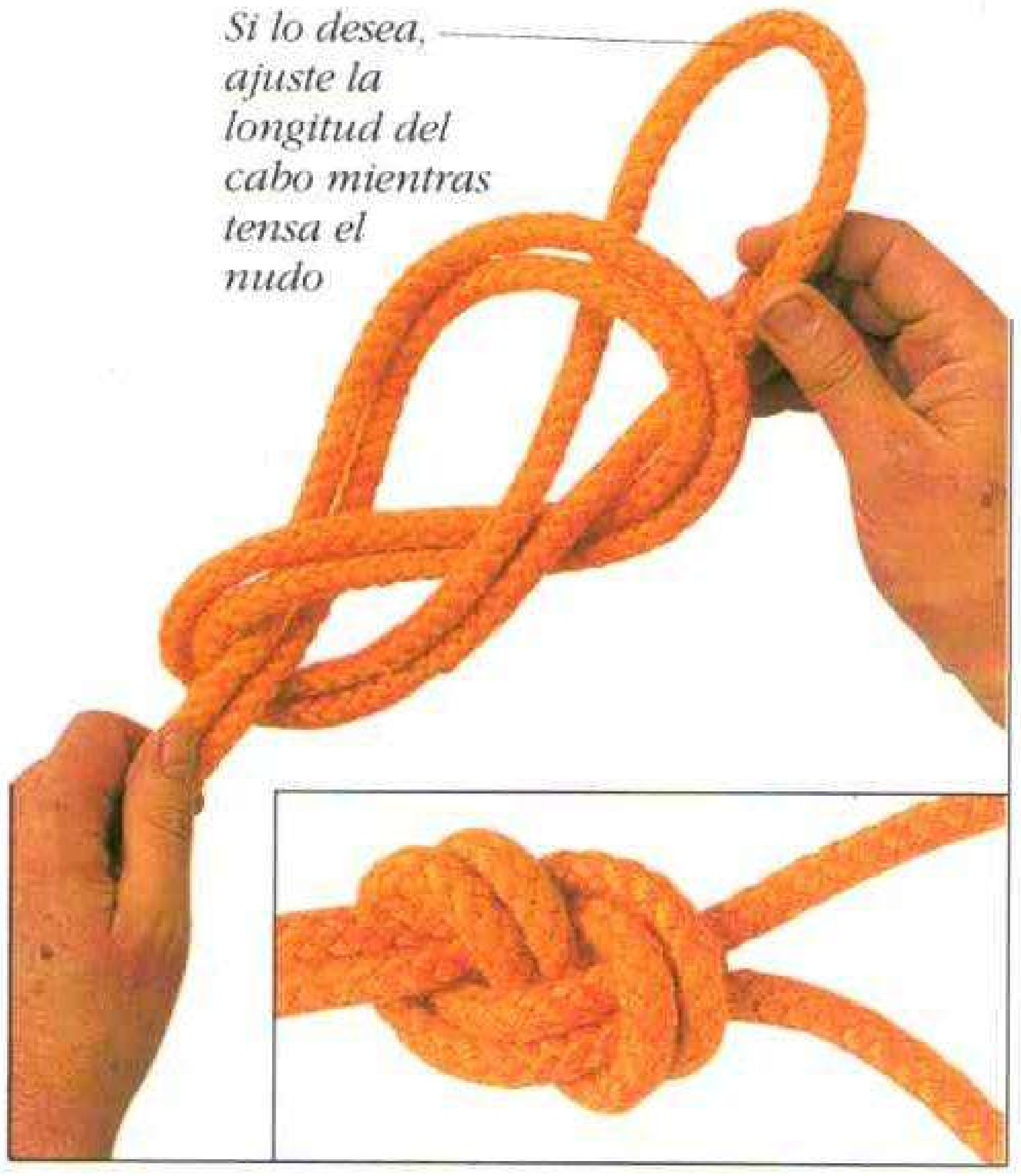
Si lo desea, ajuste la longitud del cabo mientras tensa el nudo

3 Tire con fuerza del lazo terminado y del cabo largo de la cuerda. Este nudo se emplea a menudo en escalada *(véase página 149)*. No se aflojará, aunque puede deshacerse con suma facilidad.

EL MANEJO DE LA CUERDA

UNA CUERDA PUEDE tener muchas aplicaciones, pero debe tratarla con cuidado: Su cuerda de escalar puede ser lo suficientemente resistente como para remolcar un vehículo averiado, pero perderá su capacidad de estirarse, y se ensuciará y estropeará en el proceso. Así pues, será menos probable que pueda salvarle la vida en un futuro, cuando la emplee para su propósito original. Utilice su cuerda uniformemente en sus dos cabos para no desgastarla de forma irregular. Nunca aplique una carga a una cuerda retorcida porque le causaría un daño importante. Mantenga limpia su cuerda lavándola con agua fresca y enrollándola cuando no la utiliza.

EL CUIDADO DE LA CUERDA

La humedad, la luz solar intensa y los insectos fitófagos pueden dañar las cuerdas de fibras naturales. Las fibras sintéticas pueden fundirse con el calor intenso, romperse al rozar con un canto bajo tensión, y son difíciles de asir cuando están húmedas. Las partículas de suciedad en las fibras de una cuerda pueden llegar a estropearla, así que manténgala siempre limpia y seca. Aclare las cuerdas húmedas o sucias en agua limpia y fresca. Cuelgue las cuerdas para que se sequen, en un sitio resguardado del sol y del fuego.

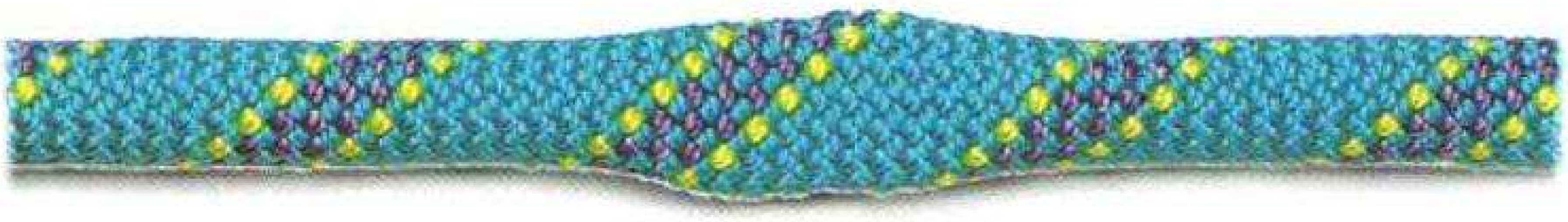

Daño en el alma. *Aunque la funda no está desgastada, esta cuerda no es segura, por lo que no debería utilizarse.*

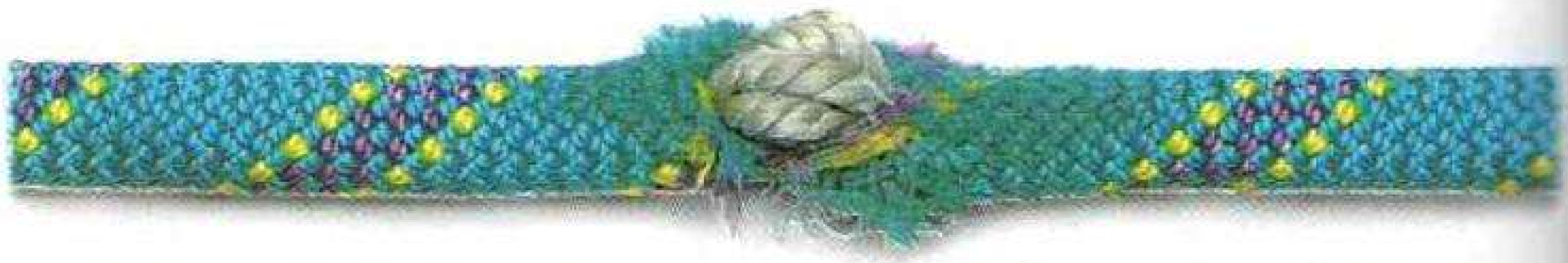

Daño en la funda. *Corte las secciones dañadas, y utilice el resto para hacer ligaduras, ataduras, o empléelo en otros propósitos no tan vitales.*

CÓMO ENROLLAR UNA CUERDA

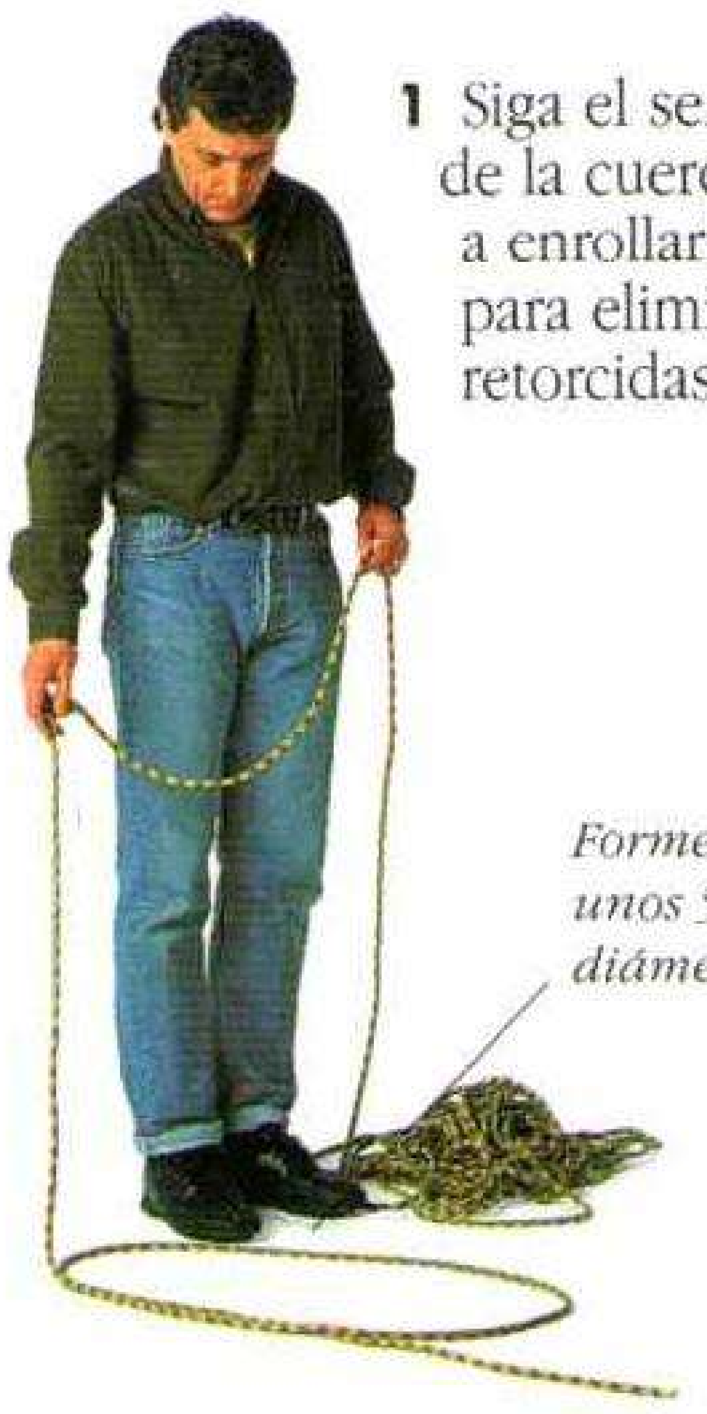

1 Siga el sentido natural de la cuerda al comenzar a enrollarla, y sacúdala para eliminar las zonas retorcidas o dobladas.

Forme círculos de unos 50 cm de diámetro

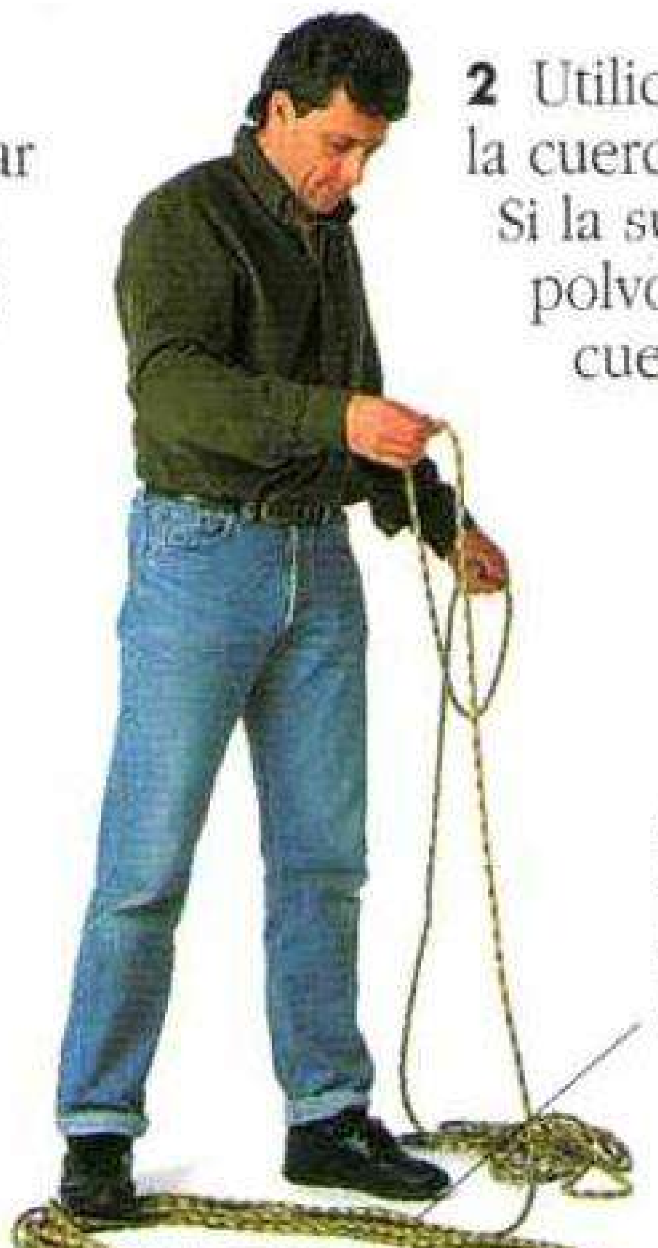

2 Utilice su pie para sujetar la cuerda mientras la enrolla. Si la superficie está sucia o polvorienta, enrolle la cuerda en la mano.

Asegúrese de mantener cada círculo junto al siguiente, sin enredarlos

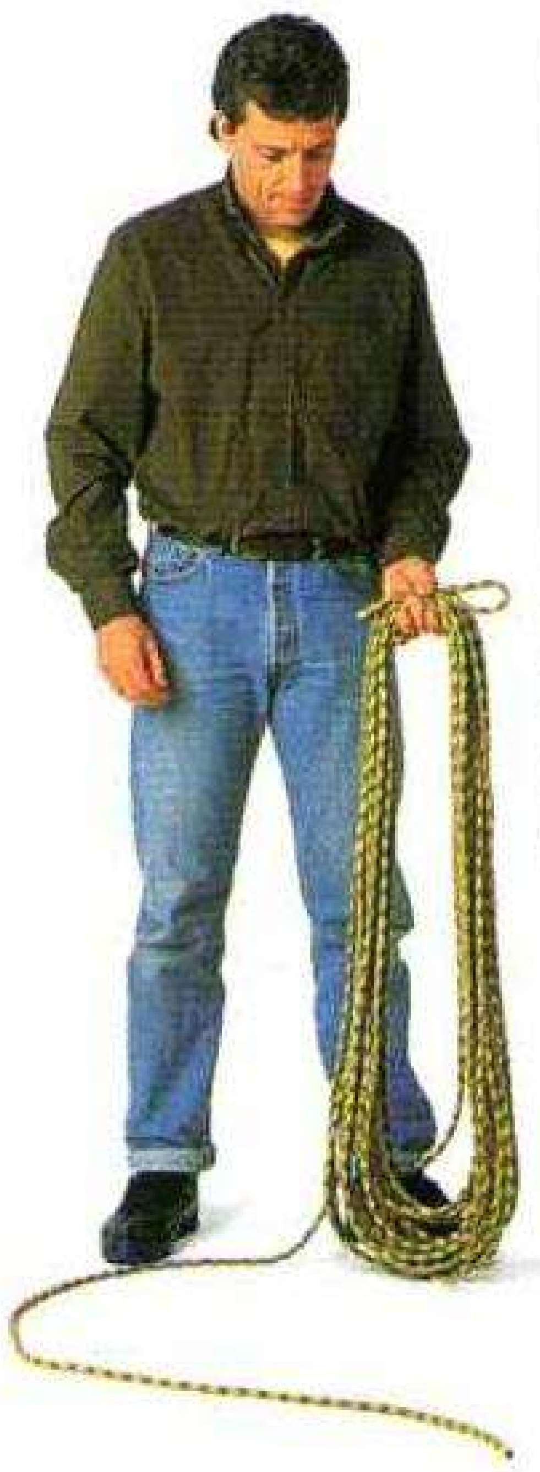

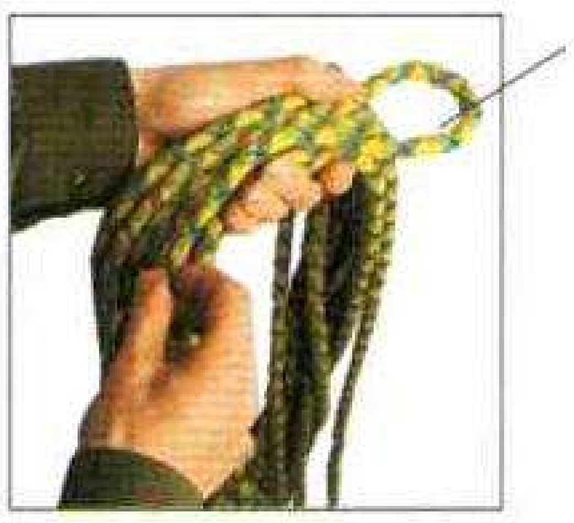

Doble el cabo interior de la cuerda y forme una lazada de unos 30 35 cm de largo

3 Continúe formando círculos hasta que le quede cuerda para un círculo más. Para amarrar la cuerda, haga una lazada doblando el cabo del círculo interior.

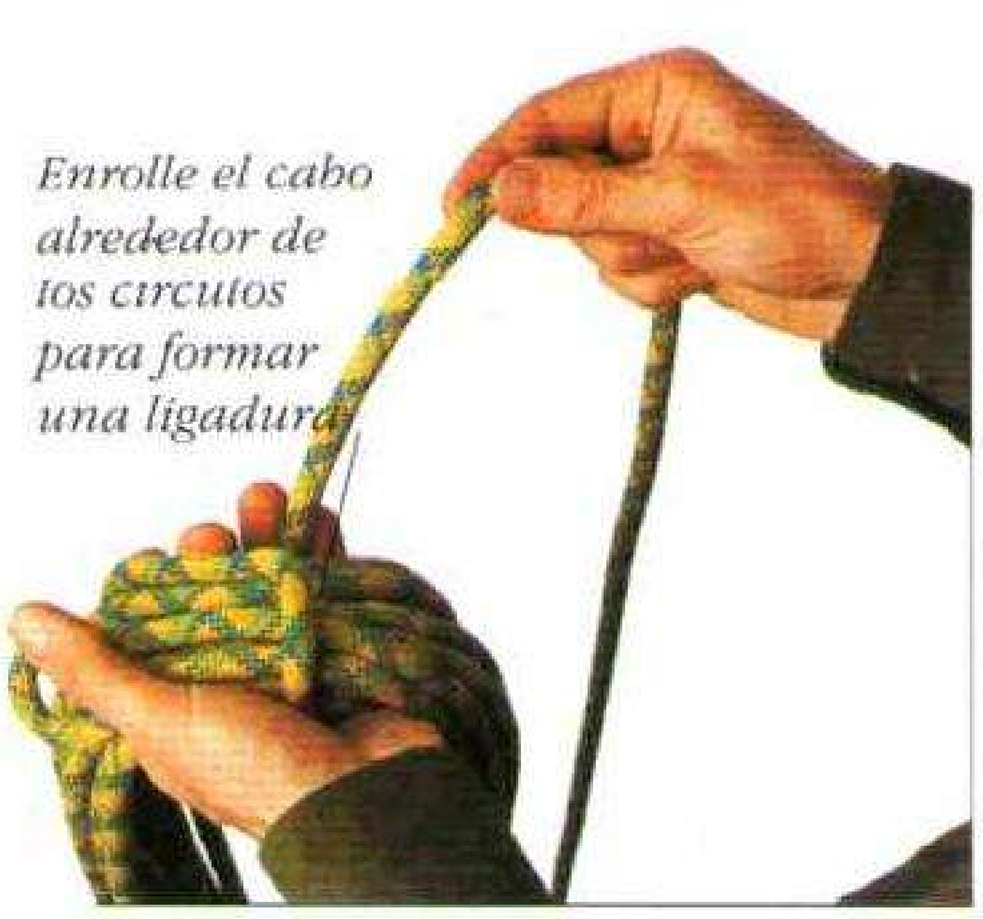

Enrolle el cabo alrededor de los círculos para formar una ligadura

4 Enrolle la cuerda sobrante alrededor de los círculos. Pásela por encima de la lazada, pero deje libre su extremo.

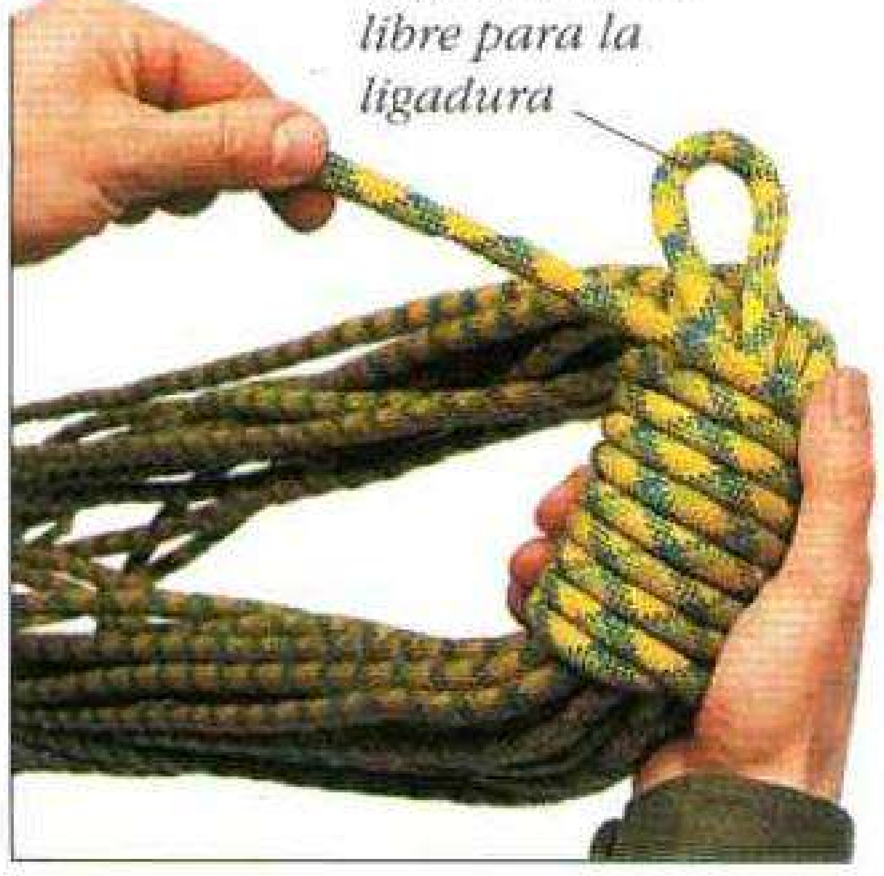

Deje la lazada libre para la ligadura

5 Termine después de haber envuelto con la cuerda sobrante unos 10 cm de rollo, dejando sobresalir 3 cm de lazada.

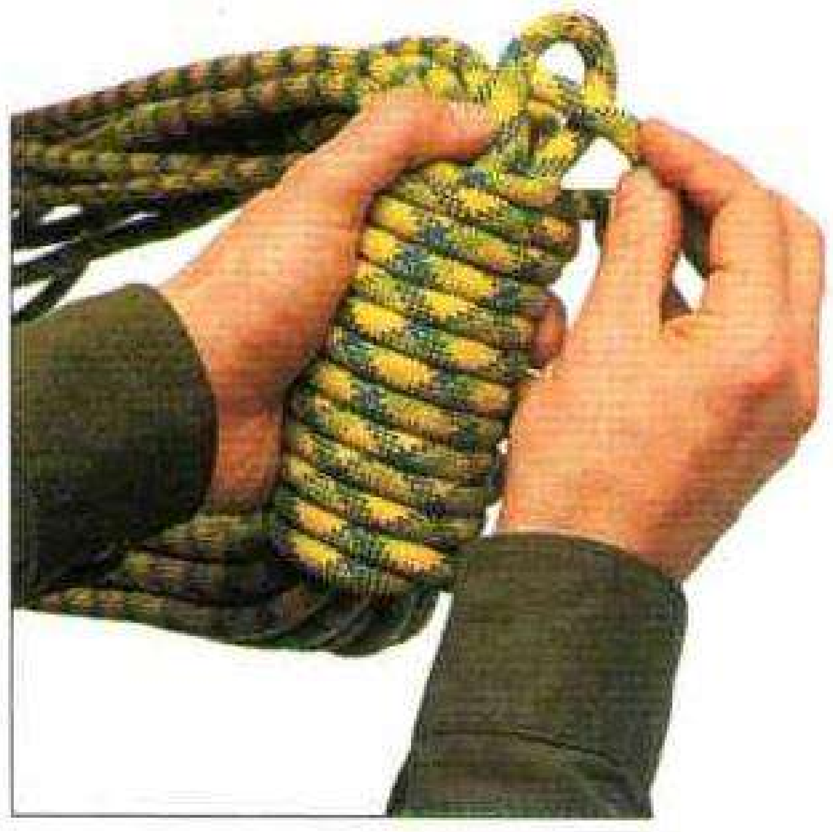

6 Pase el extremo sobrante de la cuerda a través de la lazada y ténselo.

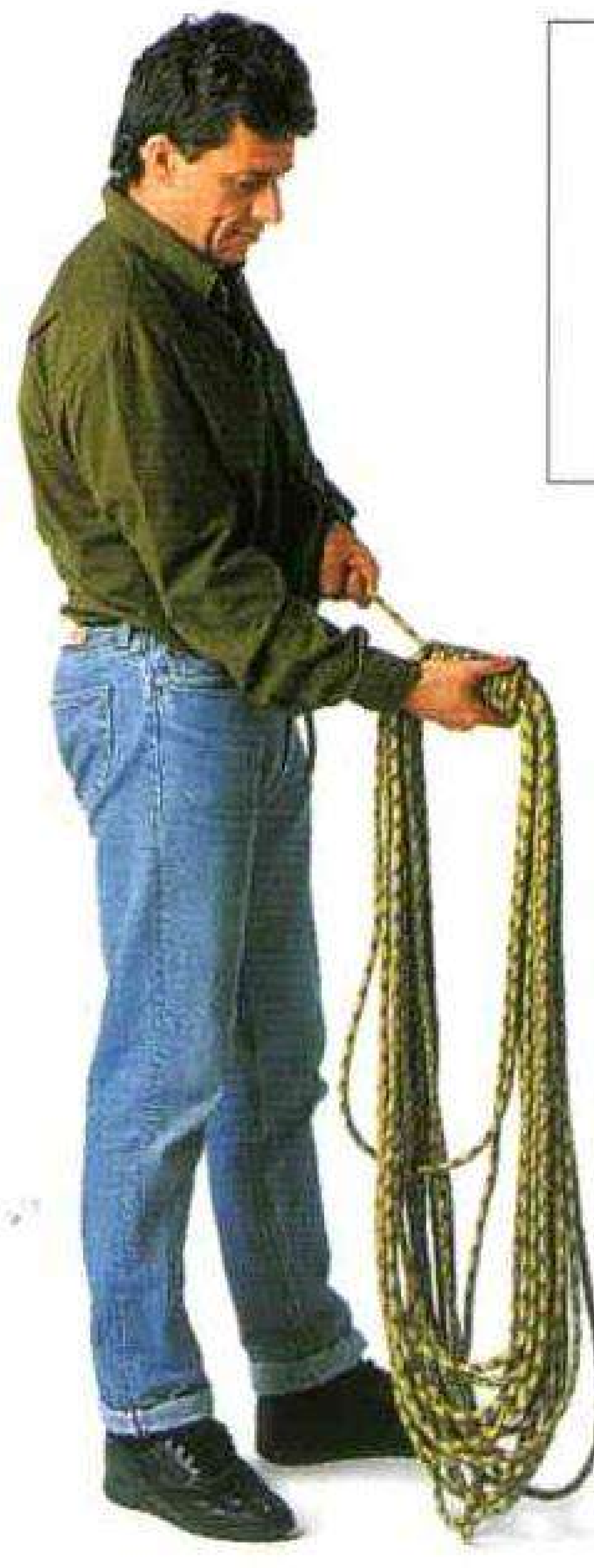

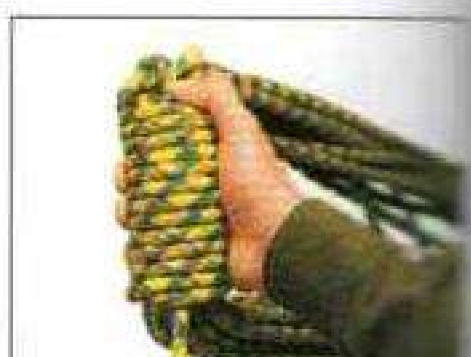

Tire de la lazada, por debajo de la ligadura

7 Tire del cabo de la cuerda que forma la lazada. De esta manera, tirará de la lazada por debajo de la ligadura y la tensará.

CÓMO UNIR DOS CUERDAS

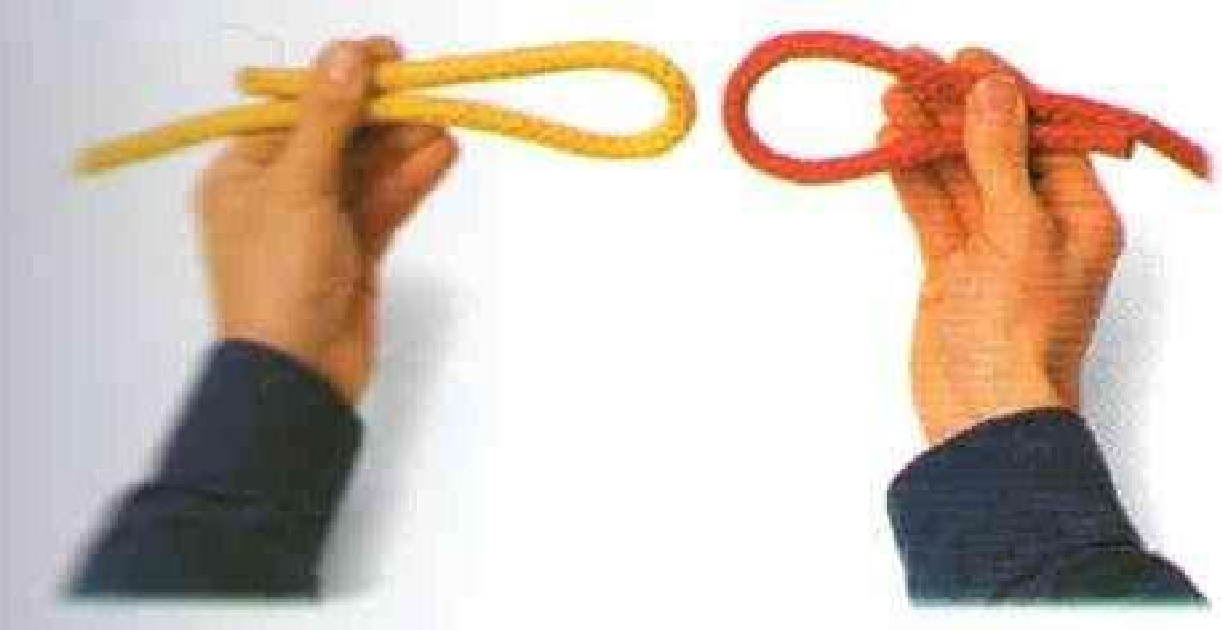

1 Si necesita unir dos cuerdas, átelas convenientemente para evitar que se separen. Comience por hacer una lazada con el cabo de cada cuerda.

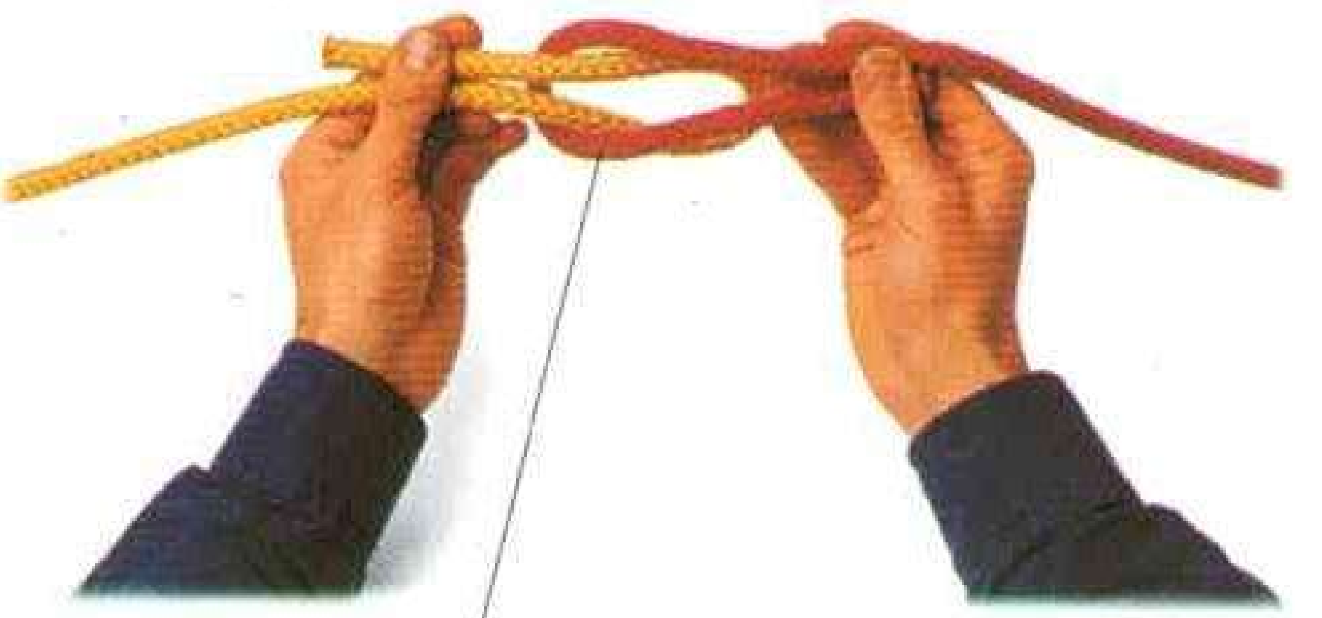

2 Acerque las lazadas hasta que una quede dentro de la otra. Podrá unir cuerdas iguales, o de materiales y grosores diferentes.

Una lazada entra en la otra

Sujete firmemente la primera lazada

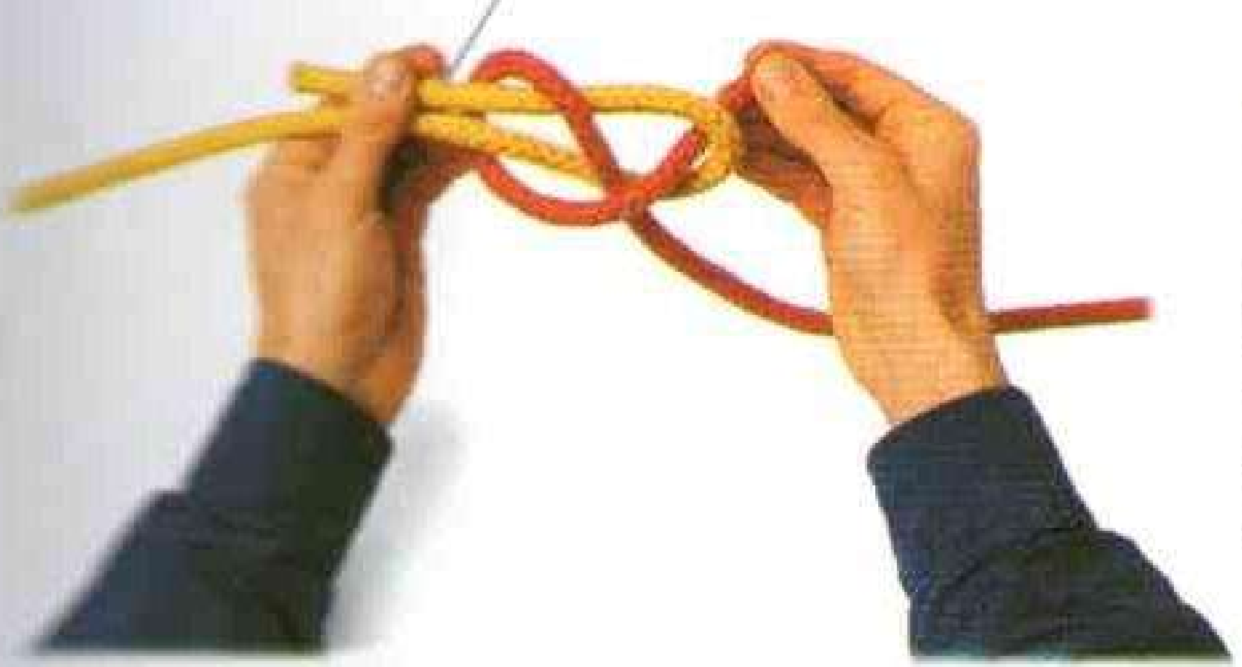

3 Tome el cabo suelto de la lazada exterior y páselo sobre sí mismo y después a través de la lazada interior. No suelte la primera lazada.

Deje que los cabos sean largos para evitar que se deshaga el nudo

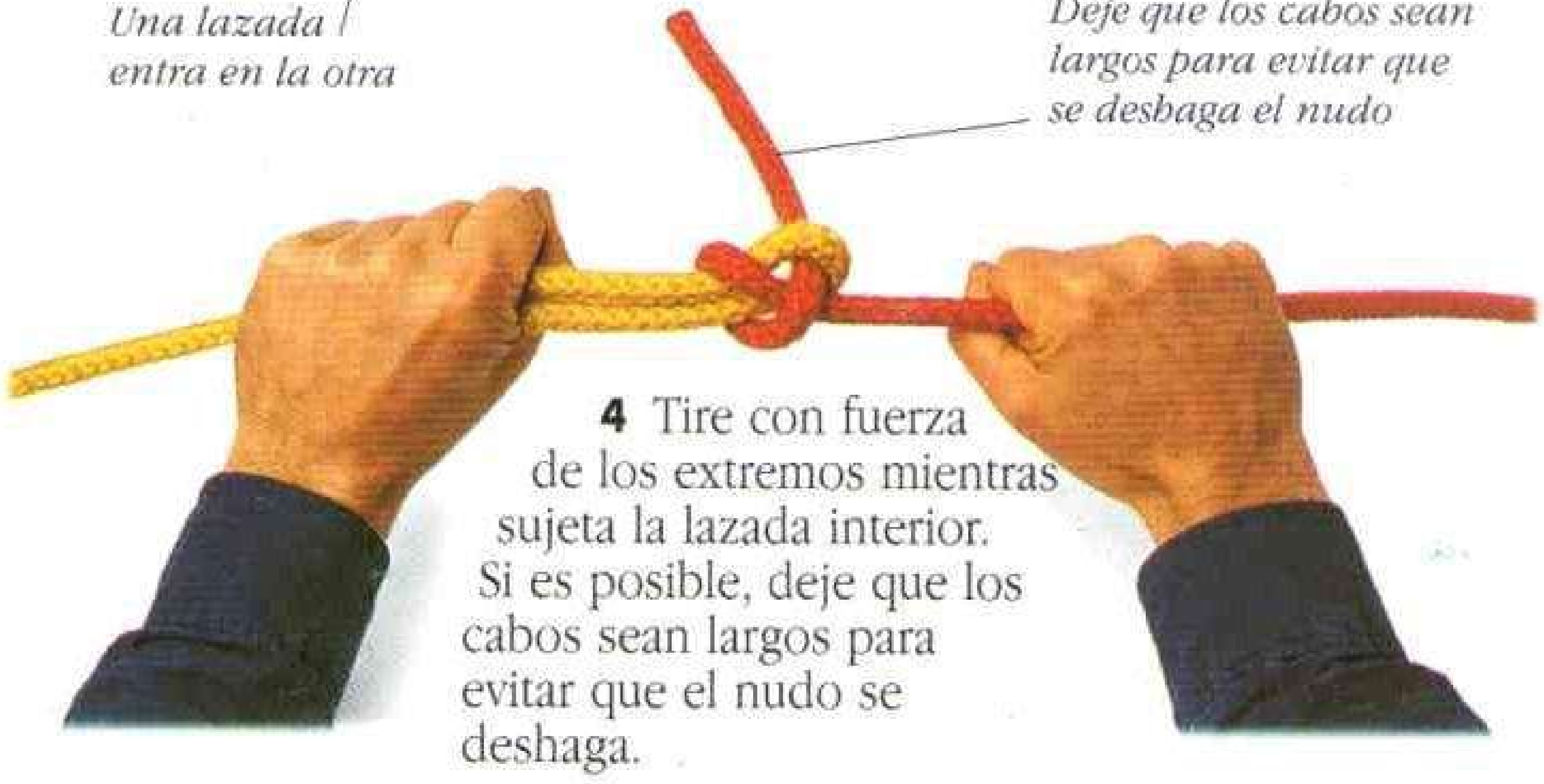

4 Tire con fuerza de los extremos mientras sujeta la lazada interior. Si es posible, deje que los cabos sean largos para evitar que el nudo se deshaga.

ATADURAS

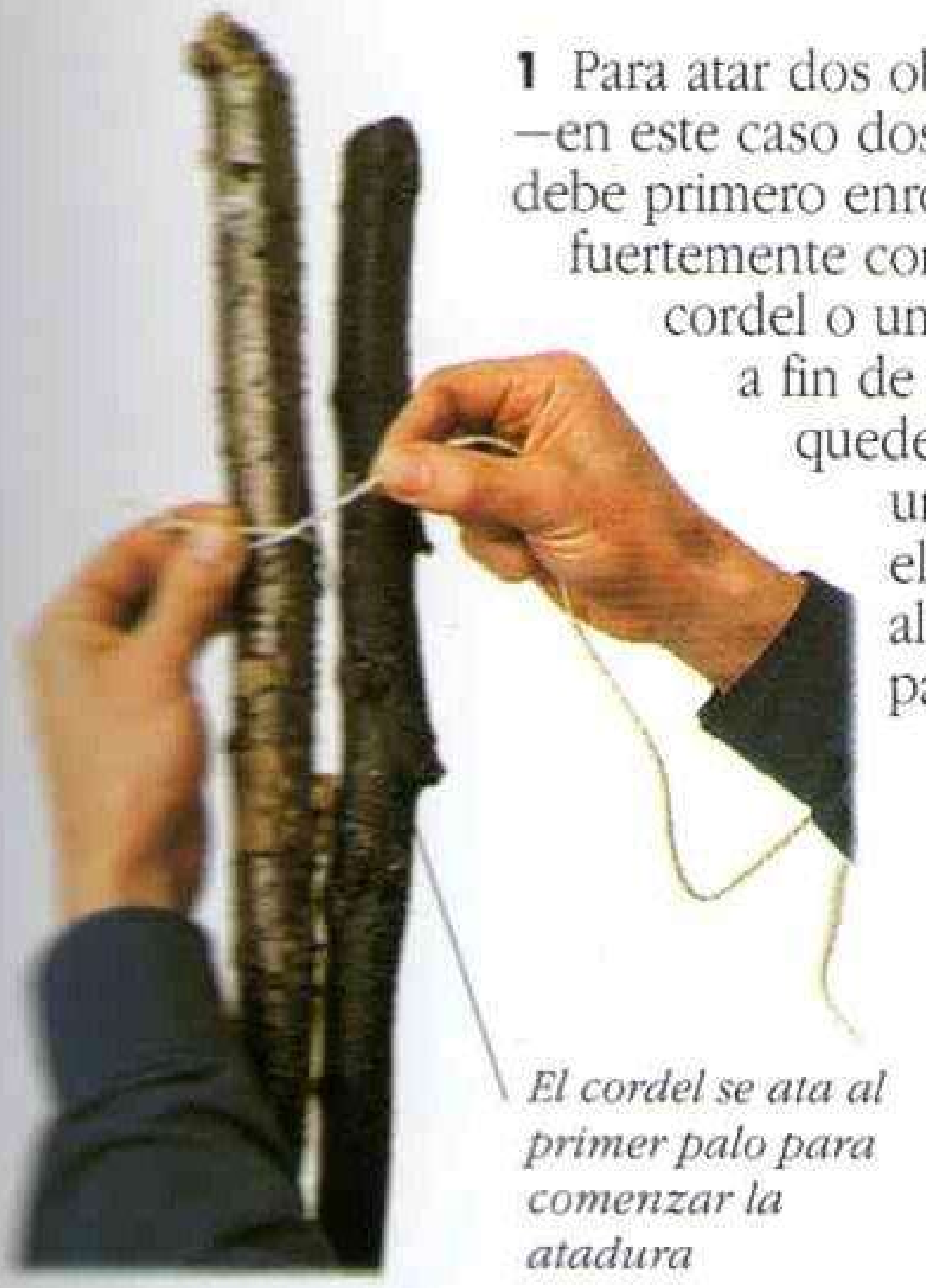

1 Para atar dos objetos —en este caso dos palos—, debe primero enrollarlos fuertemente con un cordel o una cuerda a fin de que queden bien unidos. Ate el cordel al primer palo.

El cordel se ata al primer palo para comenzar la atadura

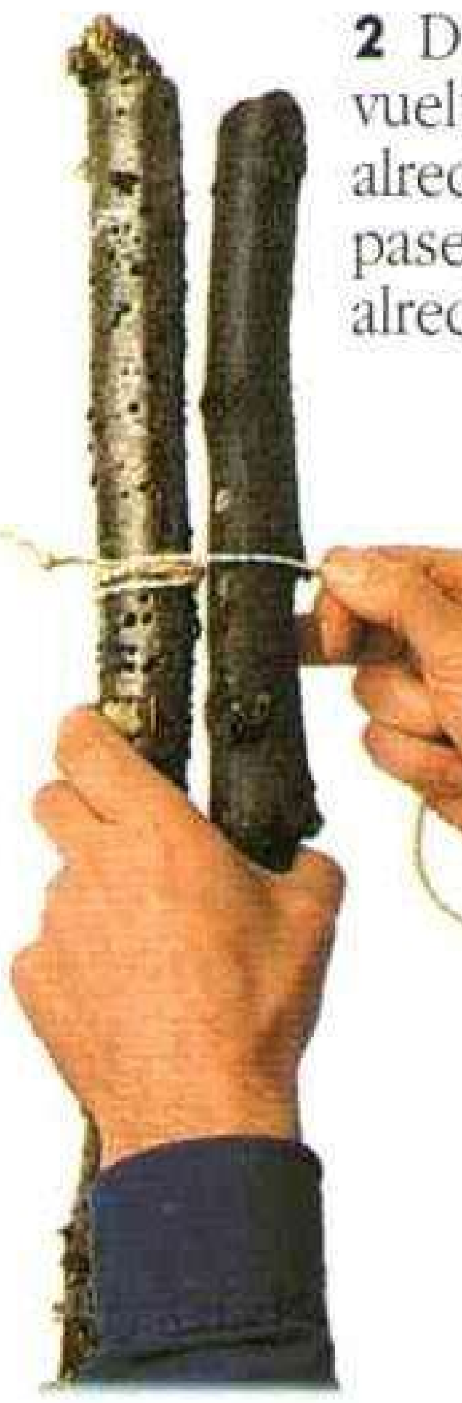

2 Después de dar varias vueltas de anclaje alrededor del primer palo, pase el cordel firmemente alrededor del segundo.

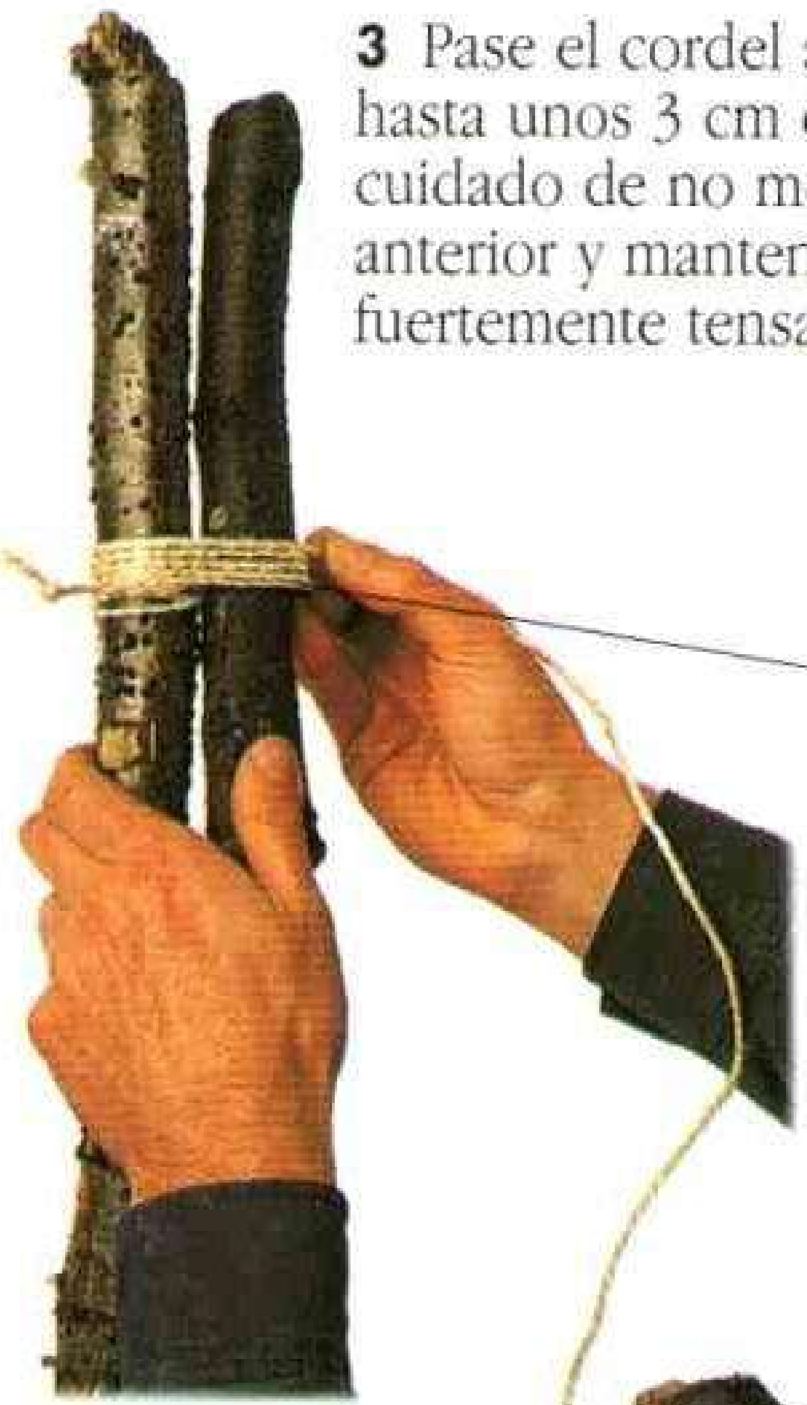

3 Pase el cordel alrededor de ambos palos hasta unos 3 cm de altura, teniendo cuidado de no montar una vuelta sobre la anterior y manteniendo cada nueva vuelta fuertemente tensada.

La atadura debe ser tirante y esmerada

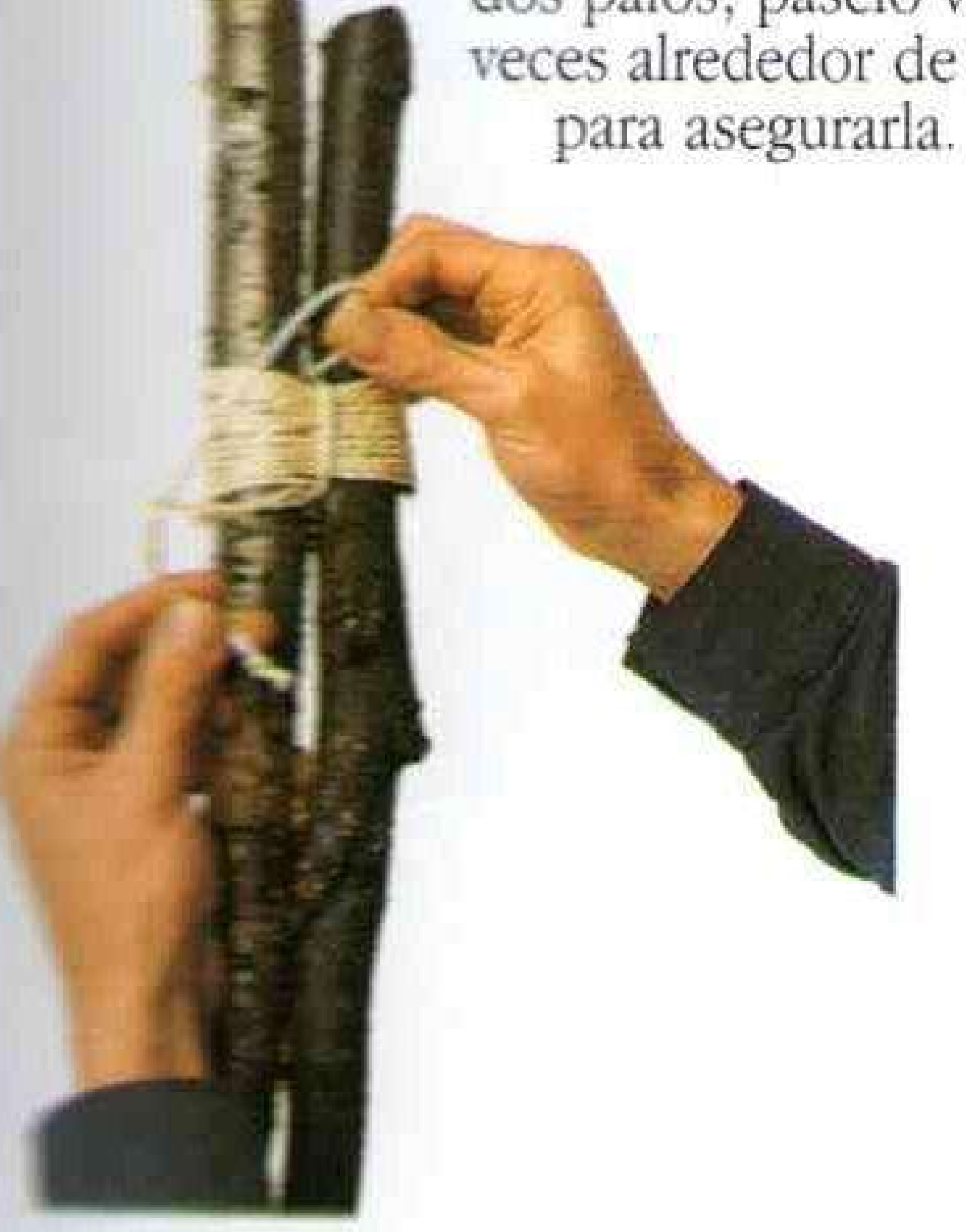

4 Pase el cordel entre los dos palos; páselo varias veces alrededor de la atadura para asegurarla.

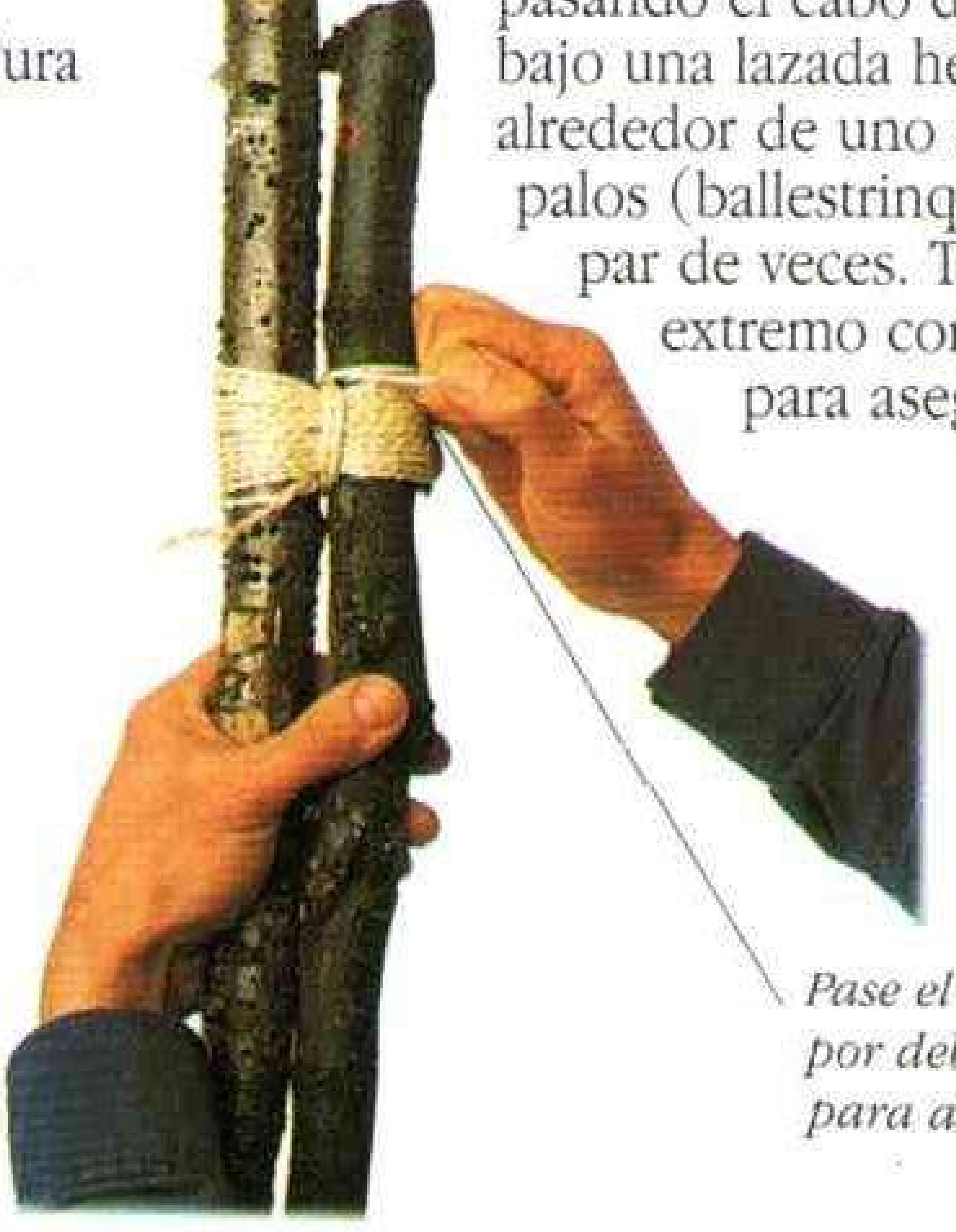

5 Termine la atadura pasando el cabo del cordel bajo una lazada hecha alrededor de uno de los palos (ballestrinque), un par de veces. Tire del extremo con fuerza para asegurarla.

Pase el extremo del cordel por debajo de sí mismo para asegurar la atadura

La atadura debe ser resistente, segura y flexible

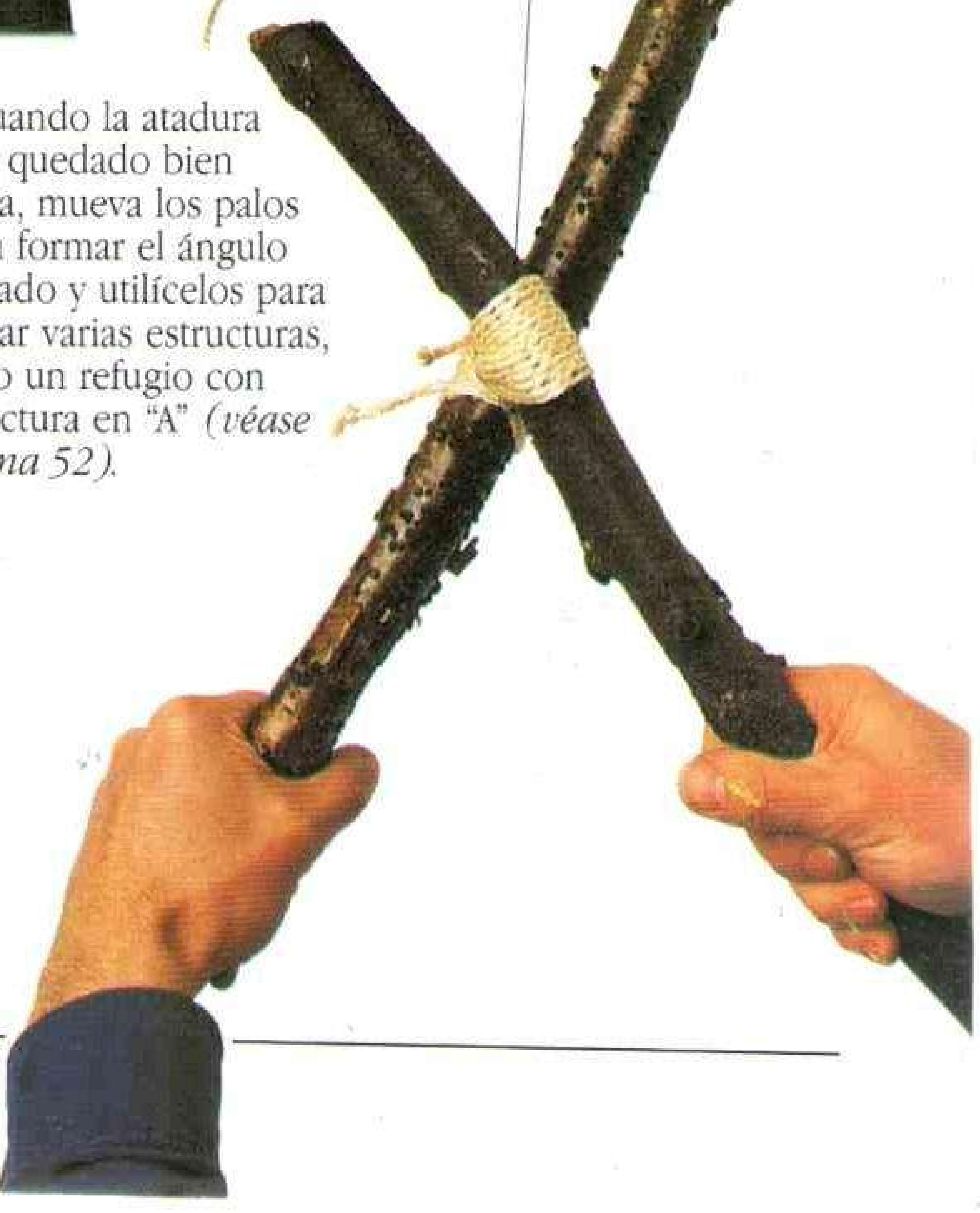

6 Cuando la atadura haya quedado bien sujeta, mueva los palos hasta formar el ángulo deseado y utilícelos para formar varias estructuras, como un refugio con estructura en "A" *(véase página 52).*

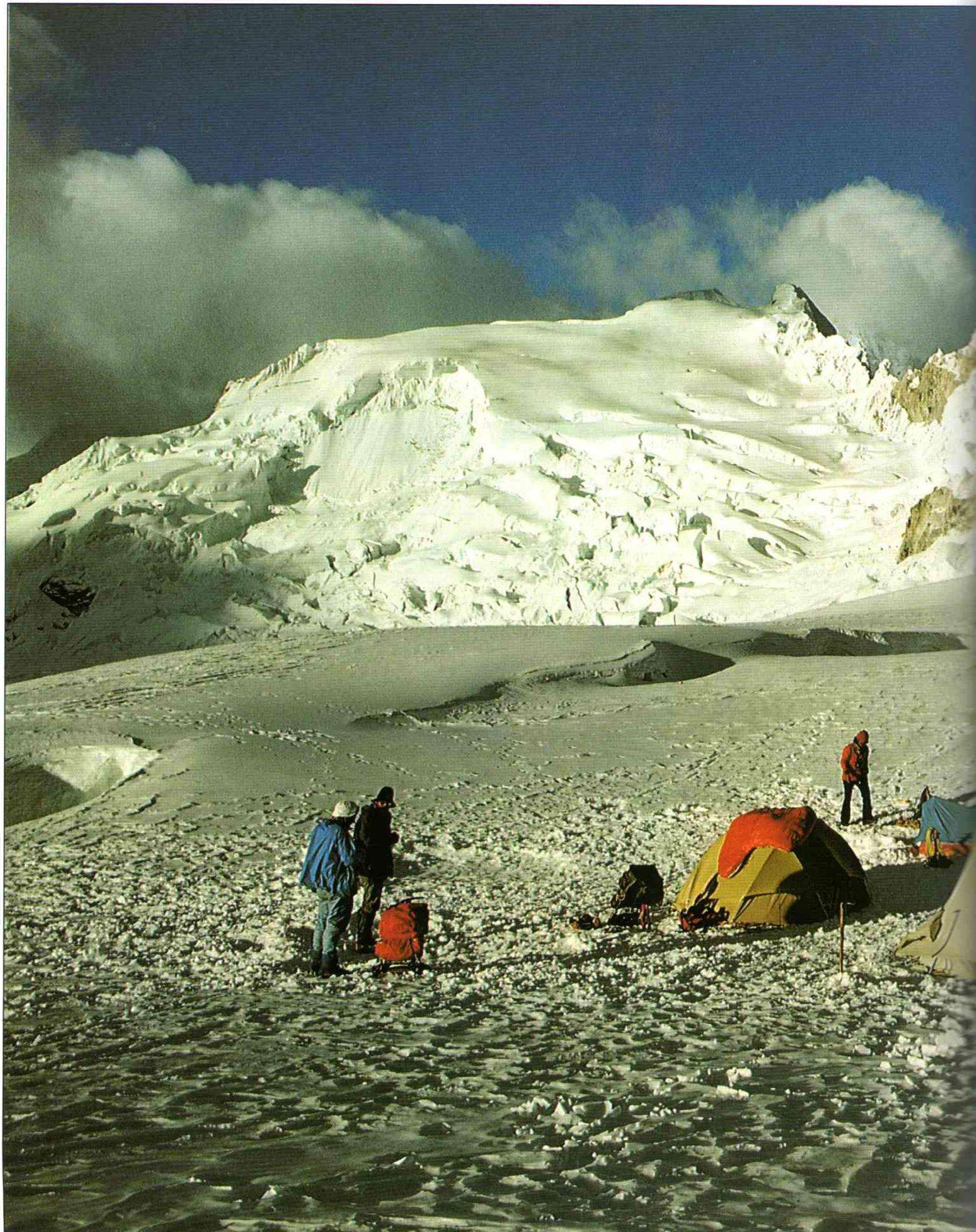

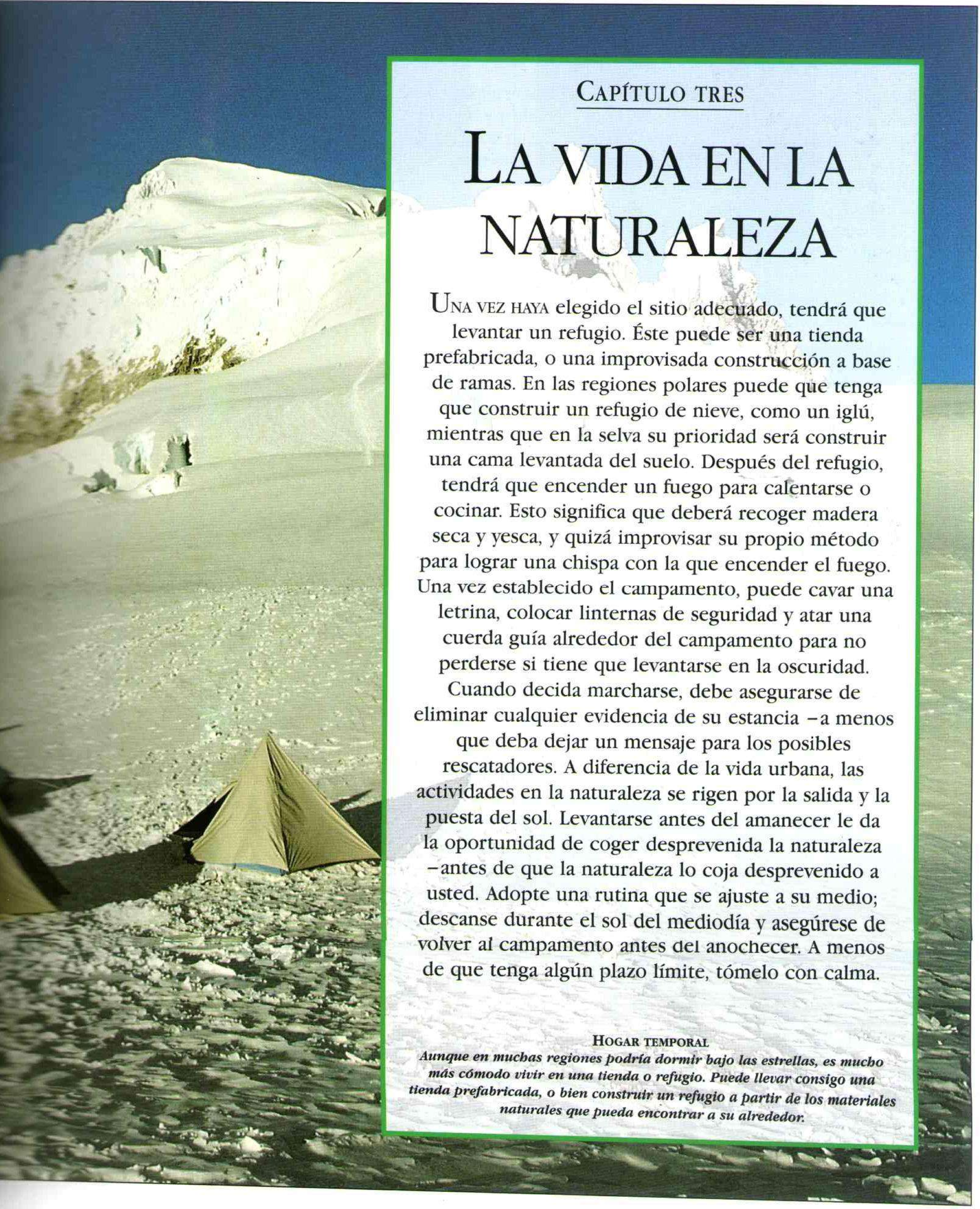

CAPÍTULO TRES

LA VIDA EN LA NATURALEZA

UNA VEZ HAYA elegido el sitio adecuado, tendrá que levantar un refugio. Éste puede ser una tienda prefabricada, o una improvisada construcción a base de ramas. En las regiones polares puede que tenga que construir un refugio de nieve, como un iglú, mientras que en la selva su prioridad será construir una cama levantada del suelo. Después del refugio, tendrá que encender un fuego para calentarse o cocinar. Esto significa que deberá recoger madera seca y yesca, y quizá improvisar su propio método para lograr una chispa con la que encender el fuego. Una vez establecido el campamento, puede cavar una letrina, colocar linternas de seguridad y atar una cuerda guía alrededor del campamento para no perderse si tiene que levantarse en la oscuridad. Cuando decida marcharse, debe asegurarse de eliminar cualquier evidencia de su estancia –a menos que deba dejar un mensaje para los posibles rescatadores. A diferencia de la vida urbana, las actividades en la naturaleza se rigen por la salida y la puesta del sol. Levantarse antes del amanecer le da la oportunidad de coger desprevenida la naturaleza –antes de que la naturaleza lo coja desprevenido a usted. Adopte una rutina que se ajuste a su medio; descanse durante el sol del mediodía y asegúrese de volver al campamento antes del anochecer. A menos de que tenga algún plazo límite, tómelo con calma.

HOGAR TEMPORAL

Aunque en muchas regiones podría dormir bajo las estrellas, es mucho más cómodo vivir en una tienda o refugio. Puede llevar consigo una tienda prefabricada, o bien construir un refugio a partir de los materiales naturales que pueda encontrar a su alrededor.

LA ELECCIÓN DEL CAMPAMENTO

SON MUCHOS LOS factores que influyen en la elección de un campamento, de los cuales la seguridad debe ser el más importante. Es posible que no sea capaz de identificar todos los riesgos potenciales o las desventajas de un sitio particular, pero para tener suerte, debe permitirse todo el tiempo necesario antes de la puesta del sol a fin de encontrar el sitio ideal donde acampar. El tiempo invertido en el reconocimiento nunca será tiempo perdido. Planifique sus actividades considerando la hora en que se pone el sol. Para entonces debe tener la tienda o el refugio montado y la comida casi a punto. Calcule una hora para establecerse, y al menos otra hora más para el reconocimiento. Así pues, si anochece a las 18.00 h, a las 15.00 h debe empezar a pensar en el campamento, y para las 16.00 h debe haberse detenido ya a buscar activamente el lugar adecuado.

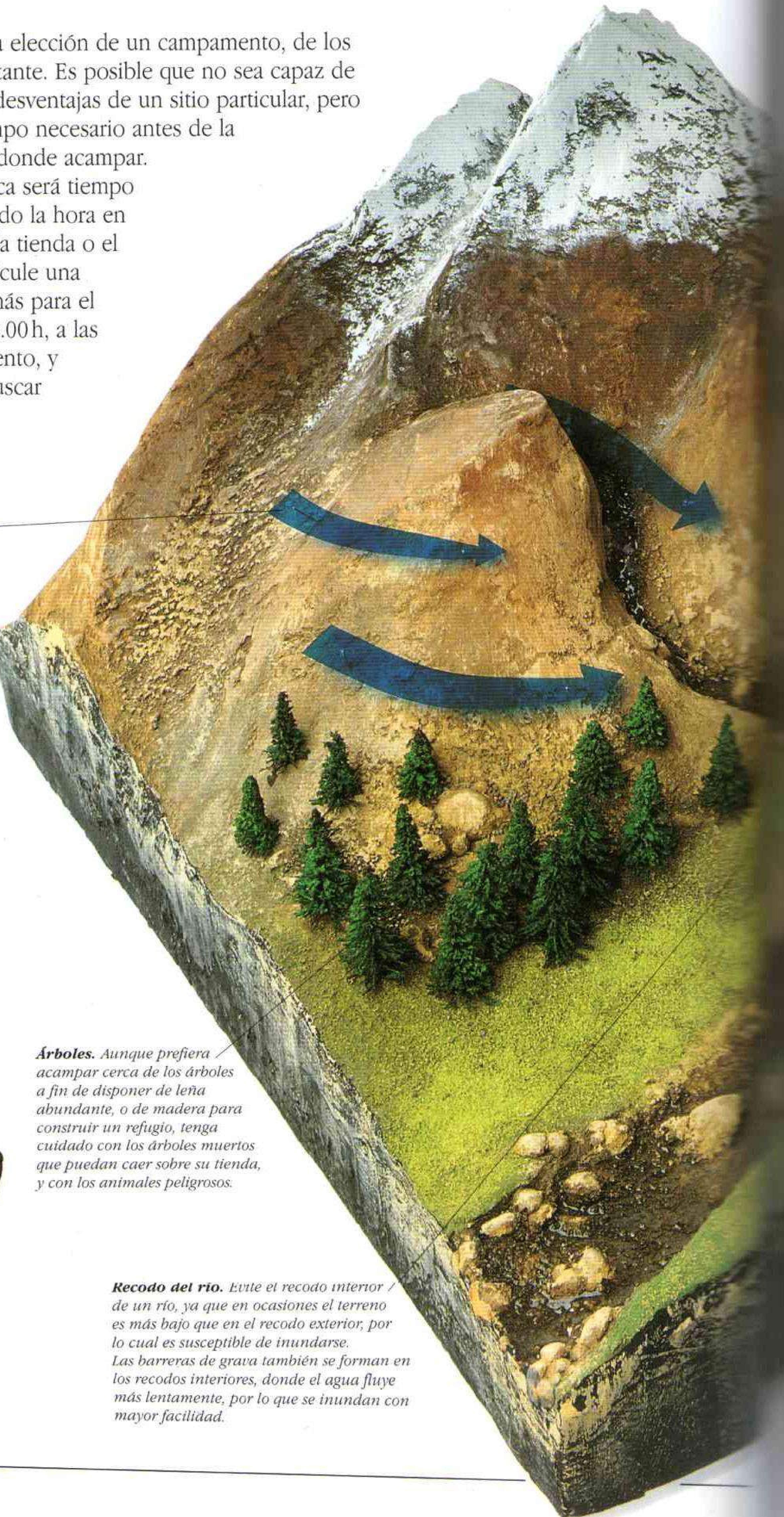

Viento predominante. *Debe intentar identificar la dirección en la que sopla el viento predominante, y levantar su tienda de manera que la entrada se encuentre en la dirección opuesta. Debe cavar la letrina en la dirección del viento, más abajo de su campamento. Encienda el fuego de manera que el humo no entre continuamente en su tienda.*

Árboles. *Aunque prefiera acampar cerca de los árboles a fin de disponer de leña abundante, o de madera para construir un refugio, tenga cuidado con los árboles muertos que puedan caer sobre su tienda, y con los animales peligrosos.*

Recodo del río. *Evite el recodo interior de un río, ya que en ocasiones el terreno es más bajo que en el recodo exterior, por lo cual es susceptible de inundarse. Las barreras de grava también se forman en los recodos interiores, donde el agua fluye más lentamente, por lo que se inundan con mayor facilidad.*

CONSEJOS PARA LA VIDA EN UN CAMPAMENTO

- En tiempo de lluvia, cave un canal alrededor de la base de su tienda, que desemboque colina abajo para evitar las inundaciones. Si el viento es intenso, sujete las cuerdas de la tienda con piedras grandes.
- Deje todo guardado hasta el momento de utilizarlo. Así sabrá dónde están las cosas, será capaz de moverse rápidamente en caso de emergencia y disminuirá las posibilidades de perder algún equipo irremplazable.
- Nunca guarde comestibles dentro de una tienda. Cuelgue su mochila unos tres metros por encima del suelo y a un metro de los troncos de los árboles, de manera que los osos y monos no puedan cogerla.
- Deshaga su equipaje y remiende y airee su ropa y saco de dormir al sol antes de volverlos a meter en la mochila.

Secar las botas. *Deje secar sus botas en palos clavados en el suelo para evitar que entren animales en ellas.*

Peligro de avalancha. *Si acampa cerca de las montañas, no monte su tienda en el camino de una posible avalancha de rocas o nieve. Es más, el agua de los deshielos primaverales puede bajar por una ladera y causar graves inundaciones.*

OTROS SITIOS

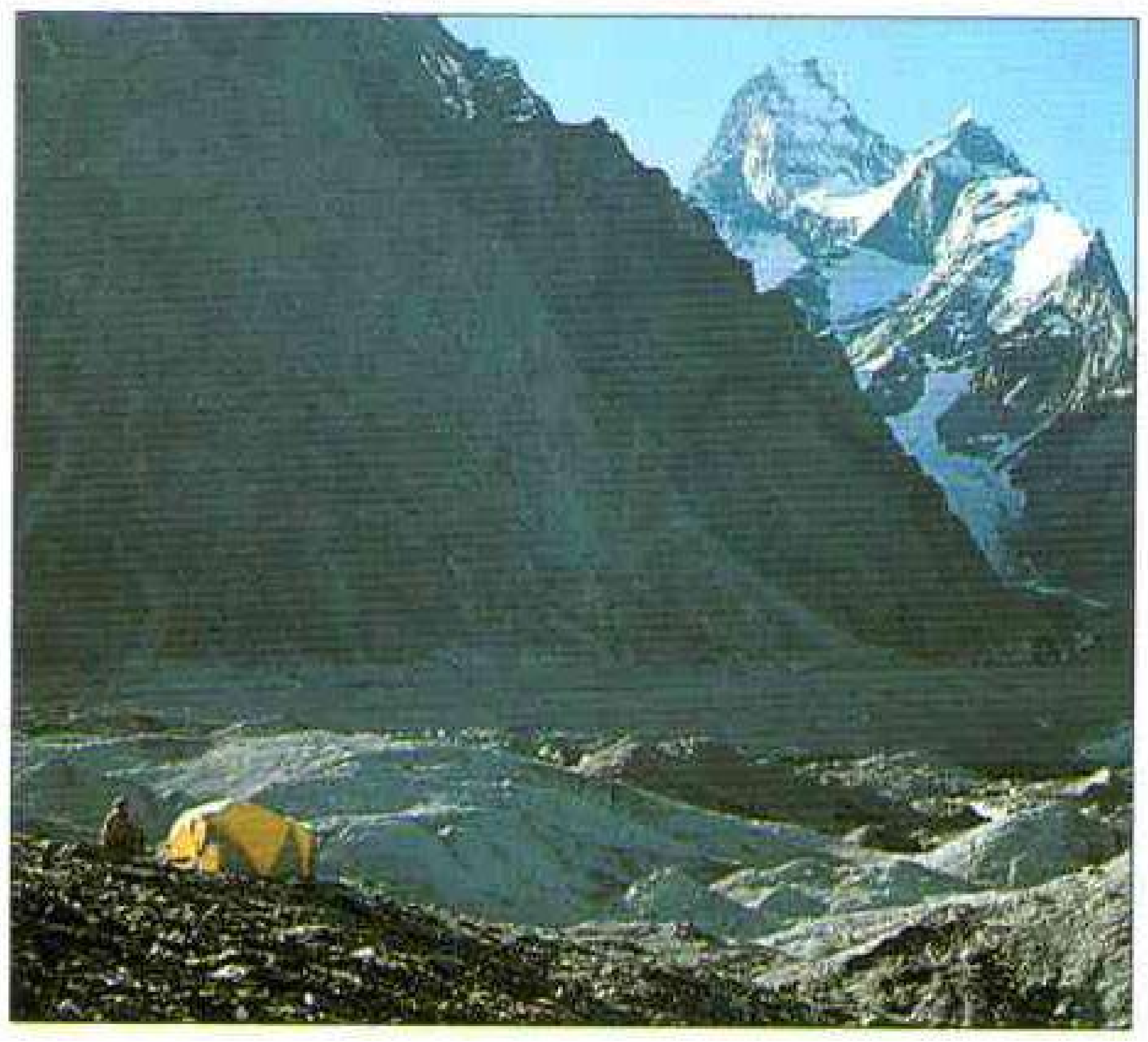

Al elegir un sitio, evite los frentes de aire frío y el suelo húmedo. Imagínese cómo fluirían el agua de una inundación y el aire frío, y elija un sitio por encima de estas líneas de flujo. Si le preocupan las picaduras de insectos, evite los suelos húmedos y busque zonas altas y barridas por el viento.

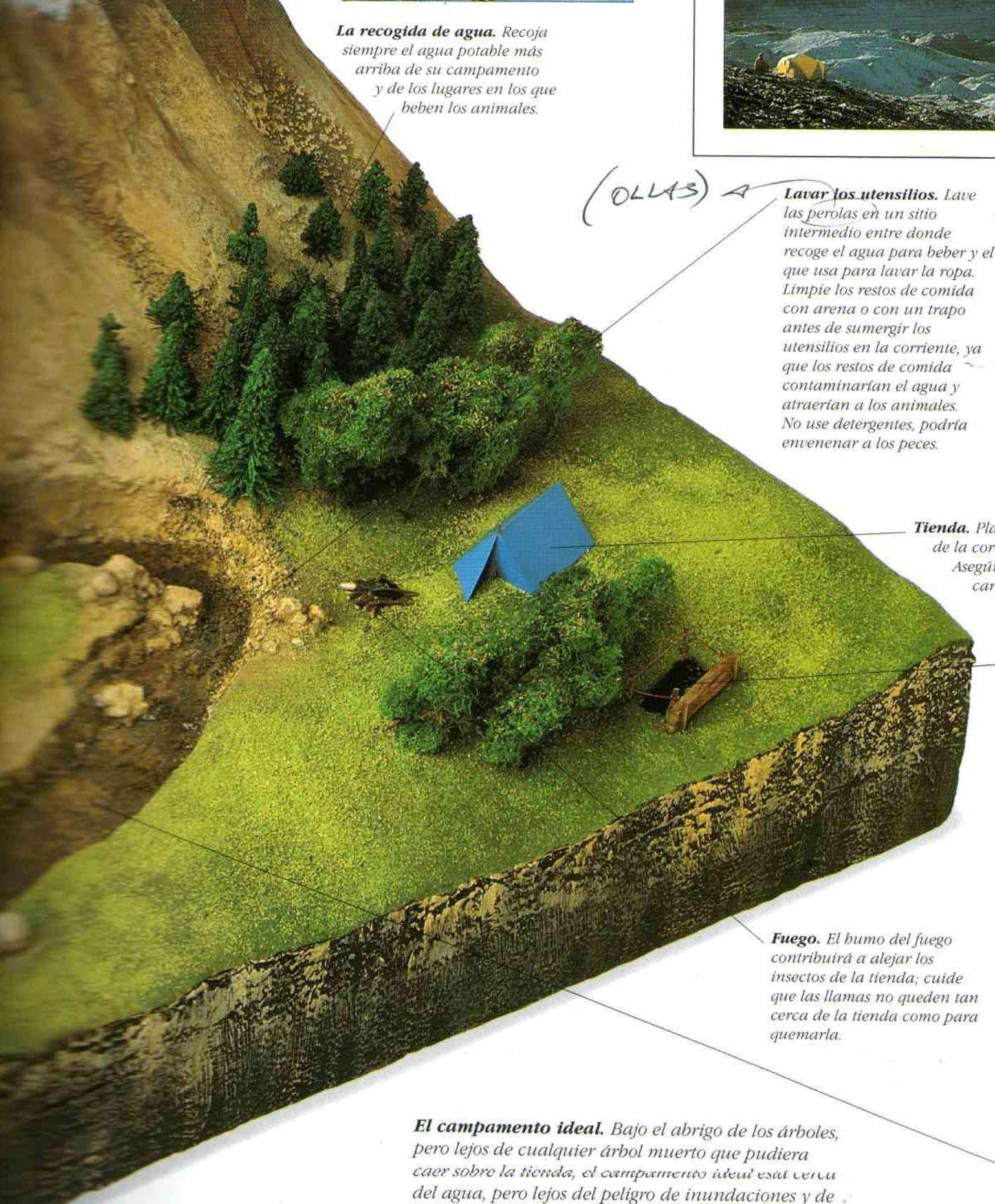

La recogida de agua. *Recoja siempre el agua potable más arriba de su campamento y de los lugares en los que beben los animales.*

(OLLAS)

Lavar los utensilios. *Lave las perolas en un sitio intermedio entre donde recoge el agua para beber y el que usa para lavar la ropa. Limpie los restos de comida con arena o con un trapo antes de sumergir los utensilios en la corriente, ya que los restos de comida contaminarían el agua y atraerían a los animales. No use detergentes, podría envenenar a los peces.*

Tienda. *Plante la tienda lejos del río, donde el sonido de la corriente no encubra el ruido de los animales. Asegúrese de que la entrada no está situada de cara al viento ni cara a la letrina.*

Letrina. *Sitúe la letrina río abajo de la tienda. Asegúrese de que esté lo suficientemente lejos del río para no contaminar el agua.*

Fuego. *El humo del fuego contribuirá a alejar los insectos de la tienda; cuide que las llamas no queden tan cerca de la tienda como para quemarla.*

Lavar la ropa. *Lave su ropa en el río, más abajo de su tienda y de donde recoge agua para beber. Humedézcala primero en el agua y enjabónela en terreno firme. Aclárela en agua recogida en un cubo y no directamente en el río.*

El campamento ideal. *Bajo el abrigo de los árboles, pero lejos de cualquier árbol muerto que pudiera caer sobre la tienda, el campamento ideal está cerca del agua, pero lejos del peligro de inundaciones y de cualquier indicio de tipo animal. En la ladera de un valle, pero no a su pie, en un terreno nivelado y bien drenado, y protegido de los vientos predominantes en la zona.*

LA ELECCIÓN DE UNA TIENDA

LAS TIENDAS LE protegen del viento y de la lluvia y le proporcionan abrigo. La tienda exterior o doble techo debe ser resistente, impermeable y tensa, capaz de desviar incluso la lluvia más intensa y los vientos más fuertes. La tienda interior no debe estar tan tensada, y es de un material permeable para reducir al máximo el vaho al tiempo que retiene el calor, y posibilita la ventilación. Ambas partes de la tienda (interior y exterior) no deben tocarse; si lo hacen, dan lugar a la pérdida de calor y a la condensación de humedad en la parte inferior del doble techo, que puede filtrarse a la tienda interior. El suelo de la tienda interior es de una capa impermeable que la aísla de la humedad.

CARACTERÍSTICAS DE UNA TIENDA

Existen varios tipos de tienda, pero la mayoría siguen la línea básica de una tienda de dos pendientes. La tienda se sostiene gracias a palos hincados en tierra, y se sujeta al suelo mediante estacas que se introducen en los ojales del contorno inferior. Se mantiene tensa con los cables o vientos fijados al suelo con clavijas. Puede llevar una cubierta que la proteja de la humedad o del viento.

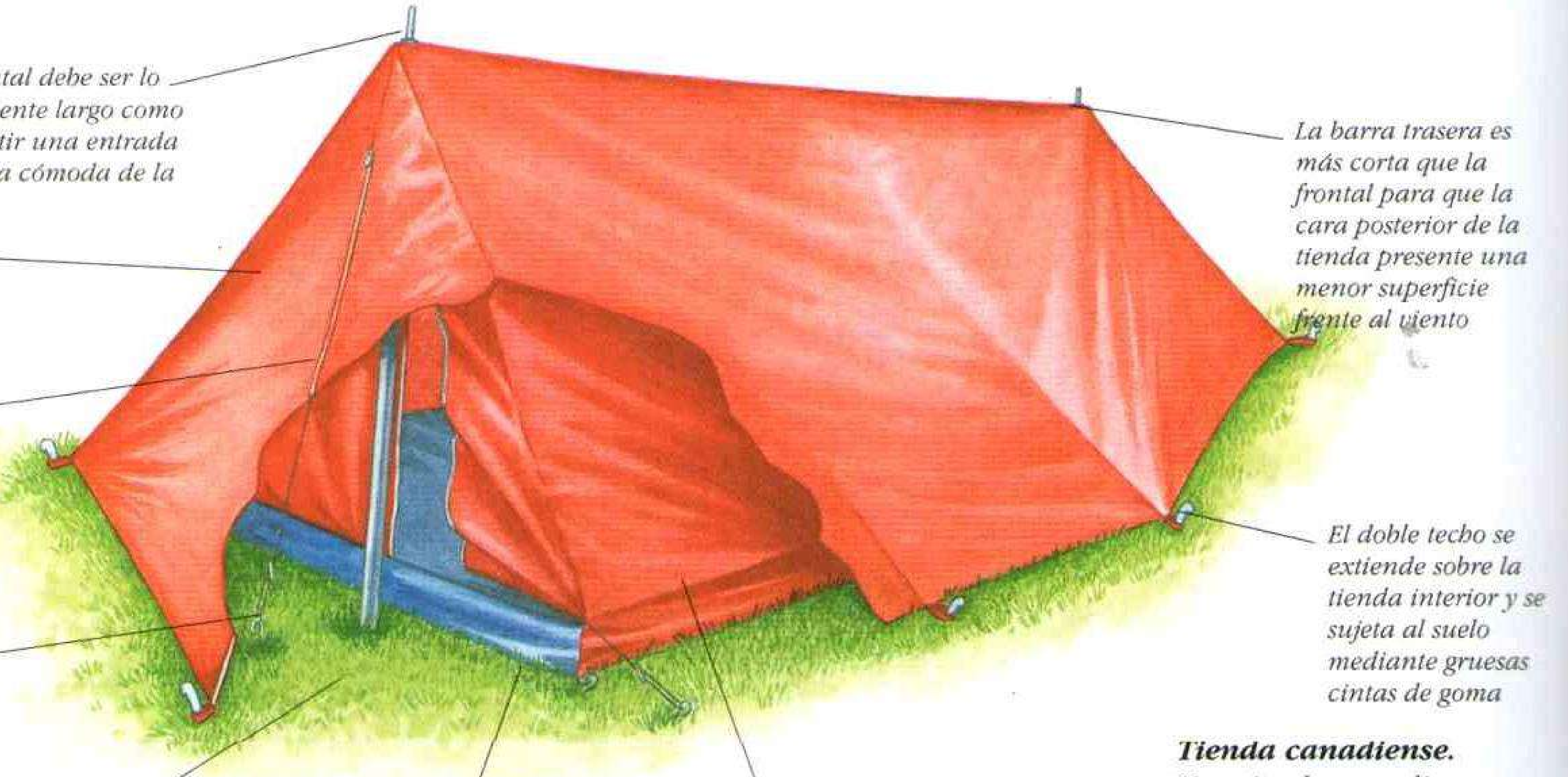
El palo frontal debe ser lo suficientemente largo como para permitir una entrada y una salida cómoda de la tienda
El doble techo impermeable se extiende tensado sobre los palos y la tienda interior
La cremallera le permite cerrar la puerta frontal de noche, o abrirla cuando cocina
Los cables ajustables o vientos tensan y sujetan la tienda interior y los palos
Este pequeño porche (avance) sirve como área para cocinar y espacio para almacenar las mochilas
La capa inferior resistente e impermeable aísla la tienda de la humedad superficial del suelo
La tienda interior no es normalmente necesaria en las regiones cálidas, excepto en invierno
La barra trasera es más corta que la frontal para que la cara posterior de la tienda presente una menor superficie frente al viento
El doble techo se extiende sobre la tienda interior y se sujeta al suelo mediante gruesas cintas de goma

Tienda canadiense. *Una tienda canadiense es adecuada para acampar en cualquier sitio, desde una montaña hasta un jardín doméstico.*

SELECCIÓN DE LA TIENDA ADECUADA

Existen varios tipos diferentes de tiendas, pero todos ellos responden a un patrón parecido. Su aplicación precisa, así como el clima y el terreno de destino determinarán el tipo de tienda a comprar. También debería considerar el tamaño y el peso de la tienda ya que tendrá que cargarla si no utiliza un vehículo. Las tiendas individuales pueden resultar demasiado frías, por lo que son preferibles las tiendas para dos personas; resulta sorprendente la diferencia de calor que generan dos cuerpos.

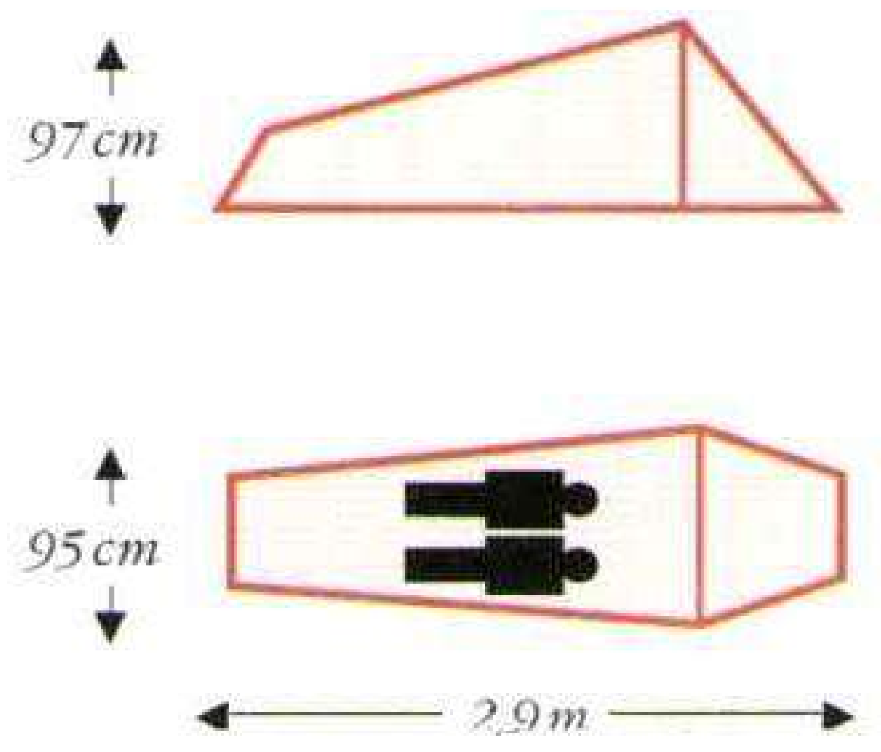
97 cm
95 cm
2,9 m

Tienda para dos personas. *Una tienda canadiense para dos personas es ligera; dispone de sitio para almacenar y cocinar debajo del doble techo y tiene el espacio justo para que dos personas duerman en su interior, una al lado de la otra.*

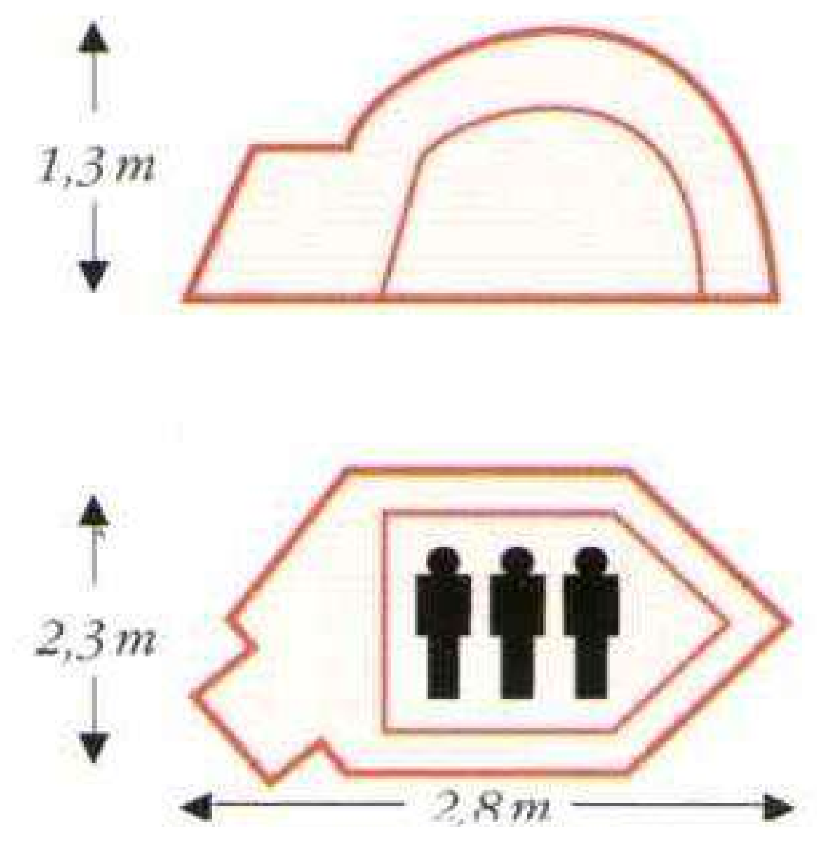

Tienda para tres personas. *Dentro de esta tienda en forma de cúpula hay espacio suficiente tanto para el almacenaje del equipo como para sentarse cómodamente. Tres palos flexibles proporcionan a la estructura una gran resistencia contra los fuertes vientos.*

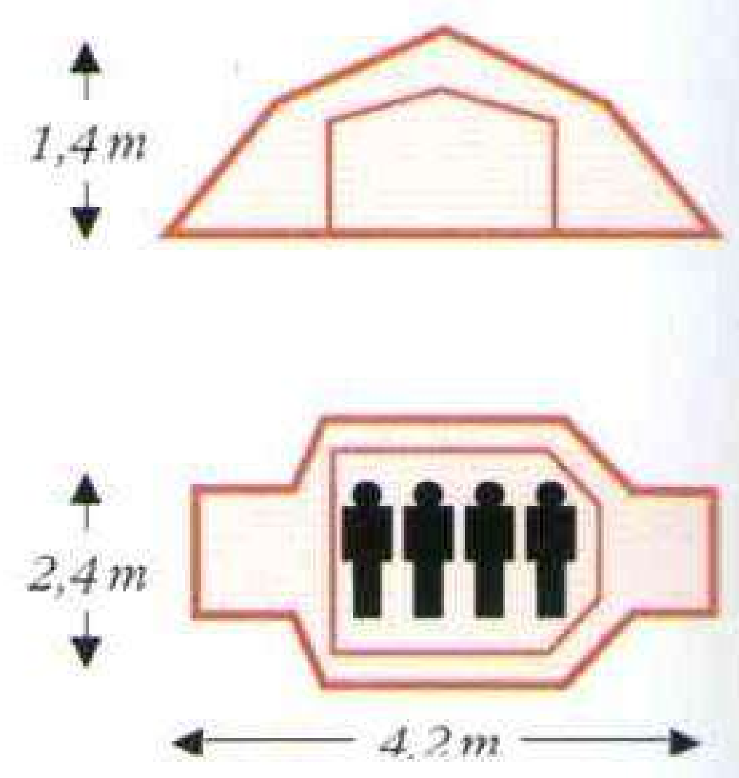

Tienda para cuatro personas. *Esta tienda, de ambiente agradable, tiene capacidad para cuatro personas. En una tienda compartida debe asegurarse de que la ventilación sea buena, y de que cada uno mantenga su equipo bien recogido para evitar un excesivo desorden.*

TIENDAS PARA CUALQUIER EQUIPO

Algunos tipos de tienda son más adecuados para un terreno particular que otros. Una tienda canadiense clásica puede utilizarse prácticamente en cualquier sitio, pero algunos modelos modernos, como las tiendas de cúpula geodésica, se adaptan mejor a los terrenos inhóspitos y a los climas extremos —por ejemplo, a los vientos intensos y a las densas nevadas de las regiones montañosas. Las tiendas con palos exteriores son más fáciles de montar en presencia de vientos fuertes que aquéllas con palos interiores. Las tiendas de dos aros son inestables en climas extremos.

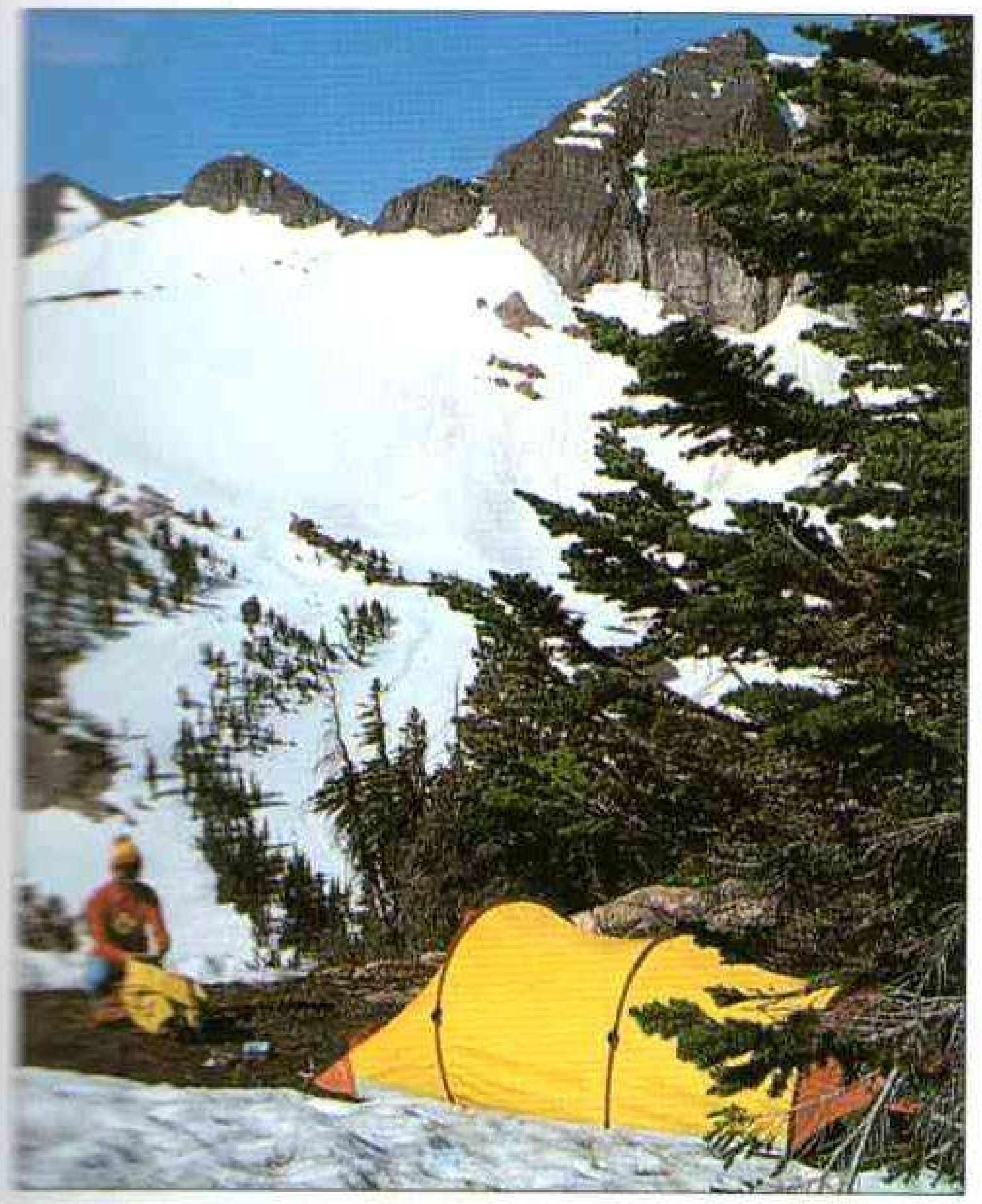

Tienda de túnel de tres aros. *Las tiendas de túnel pueden tener una estructura de hasta tres aros, dos pequeños en los extremos y uno mayor en el centro. Los túneles de tres aros pueden albergar dos o tres personas. Son relativamente estables y gracias a su forma no almacenan el agua de la lluvia. A menudo tienen dos entradas.*

Tienda de cúpula. *Es ideal para las condiciones extremas, como las que se encuentran en las regiones montañosas o polares, ya que puede permanecer estable si se entierra en la nieve. Sin embargo, debe limpiarse antes de que se acumule demasiada nieve. Enterrarla en un banco de nieve es la forma ideal de protegerla del viento.*

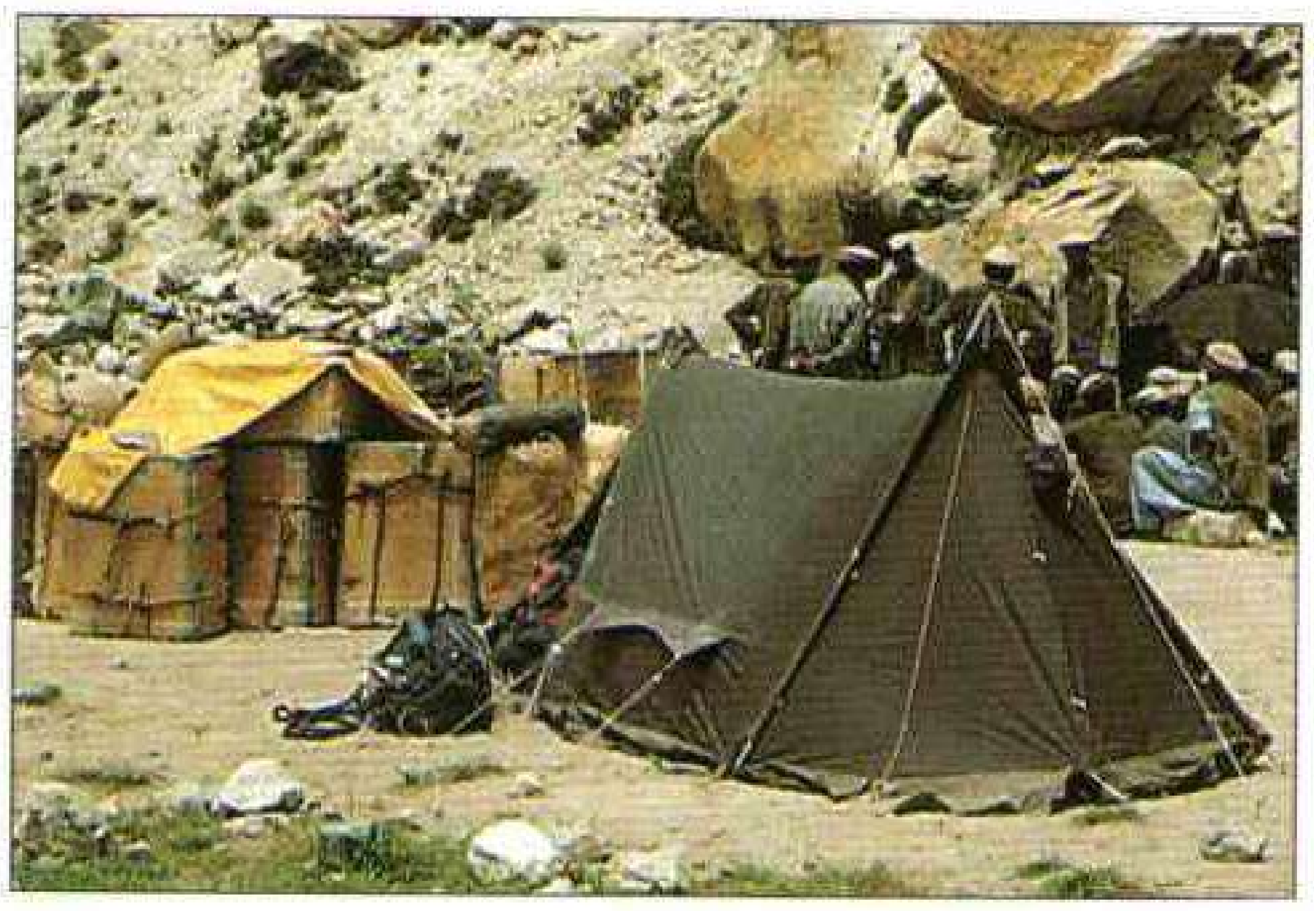

Tienda canadiense. *La tradicional tienda canadiense puede plantarse casi en cualquier sitio. Tiene una barra central horizontal, y dos barras de la misma altura en cada extremo. Este modelo de tienda se fabrica en cualquier tamaño, individual o con capacidad para varias personas.*

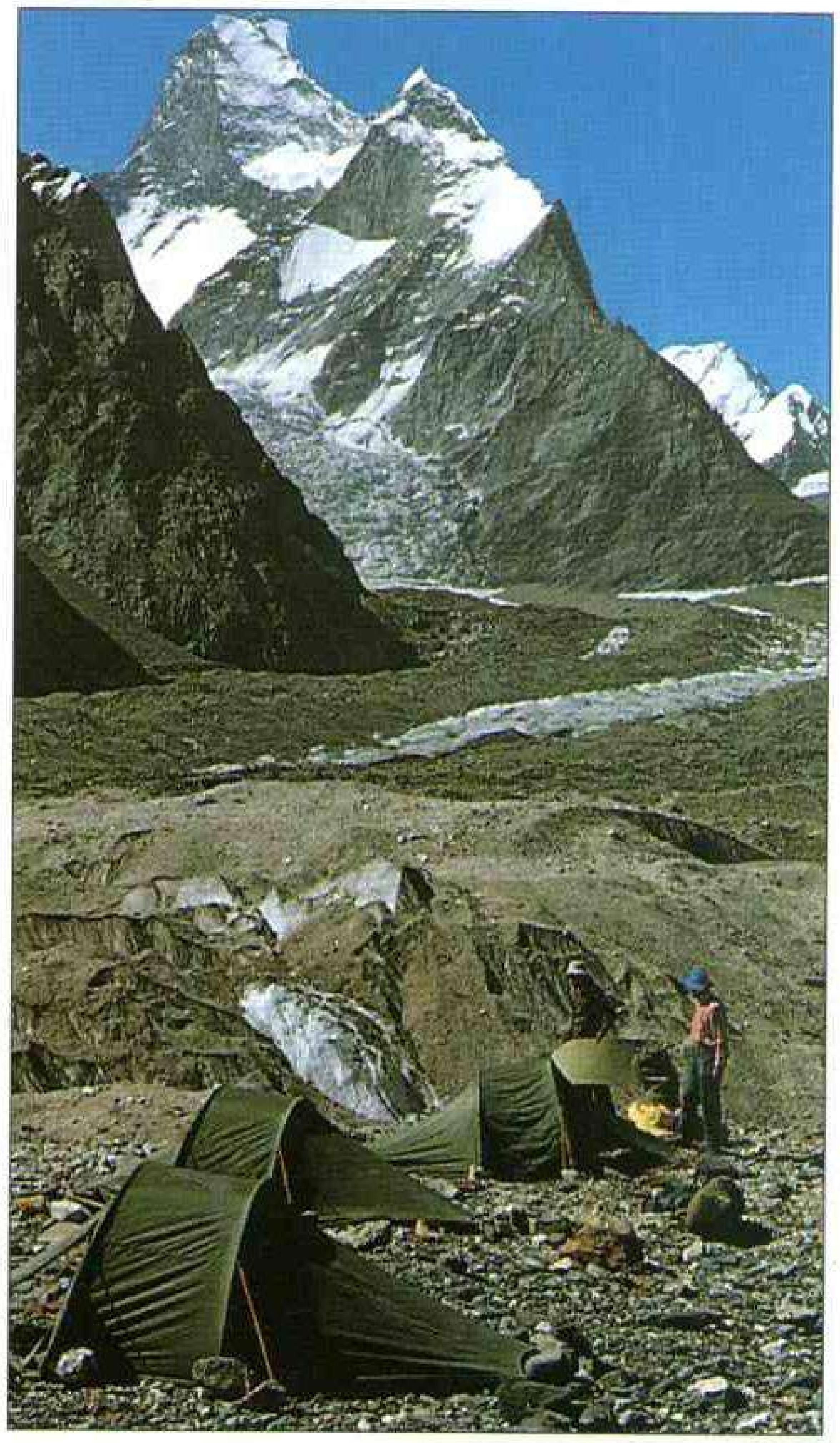

Tienda de túnel. *Una tienda de túnel es un híbrido entre una tienda de cúpula y una tienda canadiense. Las versiones individuales de este tipo de tienda son ideales para los excursionistas, ya que son ligeras y fáciles de levantar. Pueden utilizarse sobre hierba o en valles rocosos, y pueden desviar los vientos fuertes.*

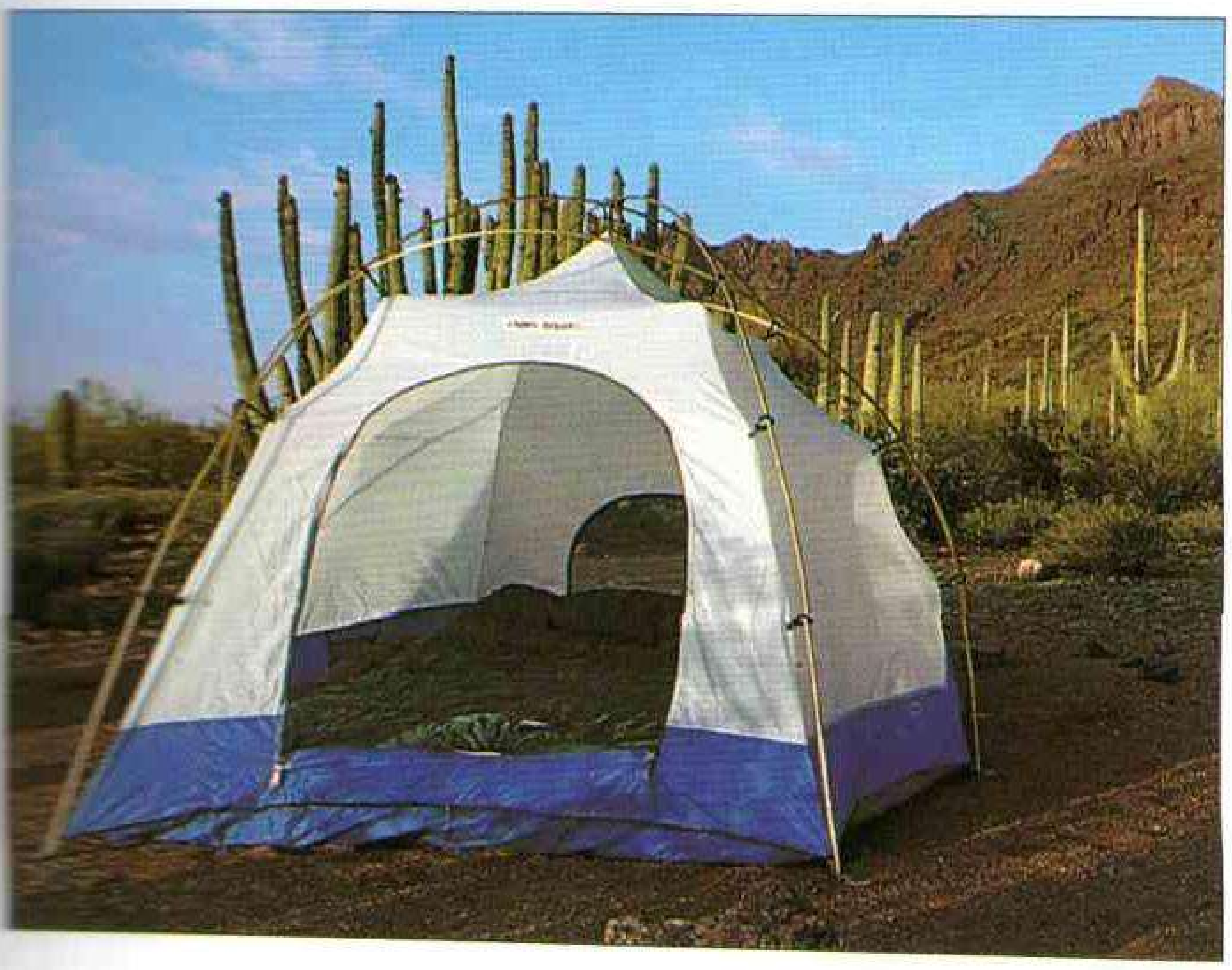

Tienda de cúpula geodésica. *Las tiendas de cúpula geodésica son resistentes, ligeras, fáciles de levantar, e igualmente fáciles de desmontar. En presencia de vientos fuertes deben sujetarse firmemente, pero la disposición de sus palos curvados y entrelazados las hace resistentes. Las tiendas de cúpula geodésica se sujetan gracias a unas varillas flexibles, y no a cables o vientos, que se cruzan a intervalos diferentes para mantenerla tensa. Gracias a su forma, proporcionan un espacio vertical considerable.*

FUNDA DE VIVAC

Una funda de vivac puede ser una simple cubierta impermeable para un saco de dormir; o si dispone de pequeños palos cortos, puede formar un porche y convertir el saco en una pequeña tienda. Debe fabricarse en un material transpirable para reducir al máximo el vaho, y puede utilizarse incluso bajo una lluvia intensa al aire libre —aunque resulta bastante ruidoso. Una funda de vivac puede utilizarse también como saco de supervivencia impermeable.

Cubierta para un saco de dormir. *Una funda de vivac constituye una excelente cubierta impermeable para un saco de dormir; puede utilizarse en lugar de una tienda.*

MONTAR UNA TIENDA DE CAMPAÑA

La elección del campamento entraña la selección de la zona adecuada para su tienda con respecto a los aspectos que ha de incluir su campamento y considerando su seguridad y refugio. Debe decidir hacia dónde encarará la entrada de la tienda y cómo asegurará los vientos, así como tomar en cuenta cualquier peligro potencial *(véase página 40)*. Debe limpiar y nivelar la zona, eliminando o allanando cualquier protuberancia del terreno.

Al igual que cualquier trabajo en un campamento, intente establecer una rutina, realizando las mismas acciones en el mismo orden cada vez, de manera que se convierta en una costumbre. Esto le permitirá plantar su tienda en la oscuridad, en una ventisca o bajo una lluvia intensa. Asegúrese de que sabrá encontrar cada una de las piezas de la tienda cuando llegue el momento de montarla. También debe desmontarla sistemáticamente *(véase página 46)*.

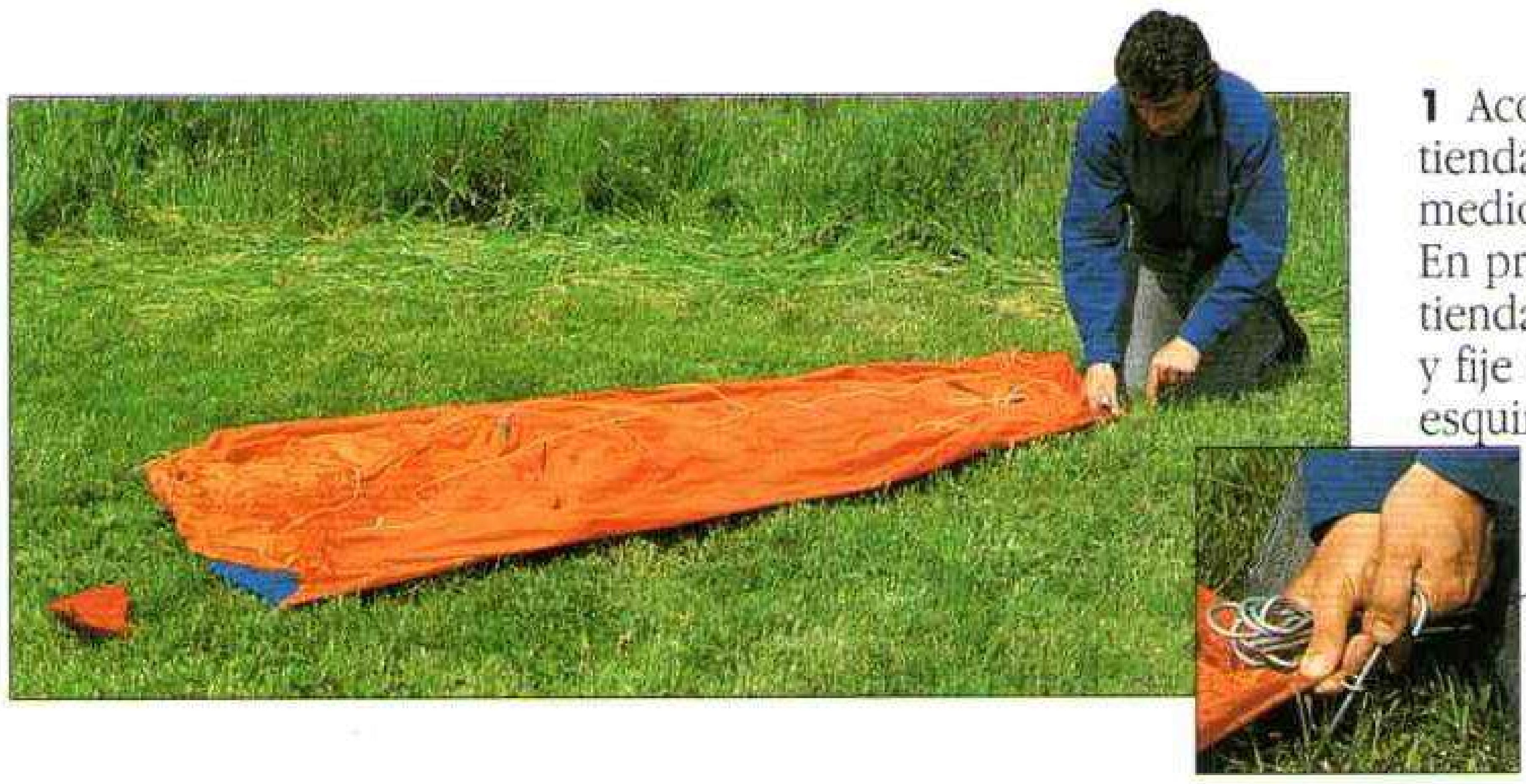

1 Acostúmbrese a montar su tienda como si lo hiciera en medio de una fuerte ventisca. En primer lugar coloque la tienda interior en el suelo y fije inmediatamente las esquinas.

Sujete el ojal de la esquina con la clavija y déjelo recto y tenso

2 Monte los palos de la tienda. Tenga cuidado de unir las piezas correctas, o tendrá dificultades para separarlas posteriormente. Sujete el enganche posterior de la tienda en el palo más corto.

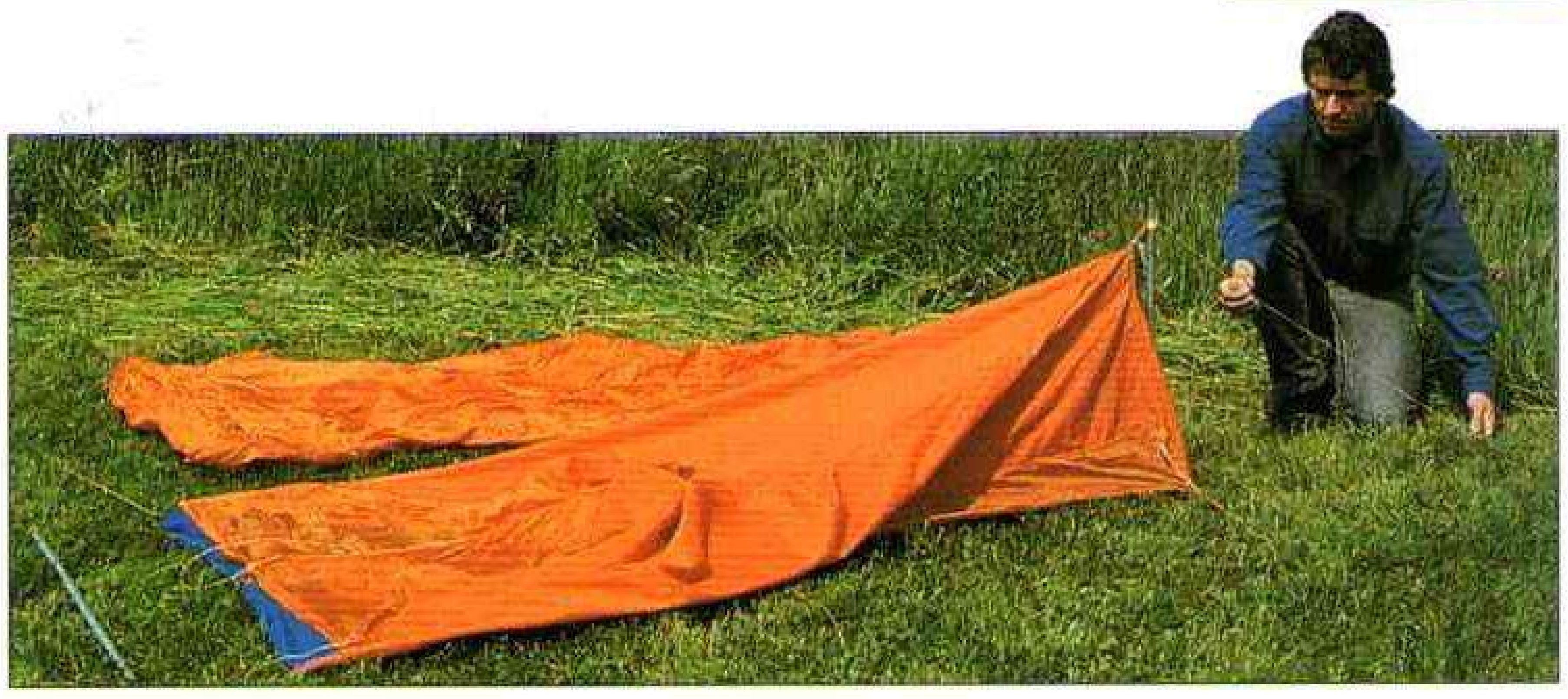

3 Levante el palo corto y fije su viento al suelo, tensándolo hasta que el palo permanezca vertical por sí mismo. Asegúrese de que la pared posterior de la tienda encara al viento.

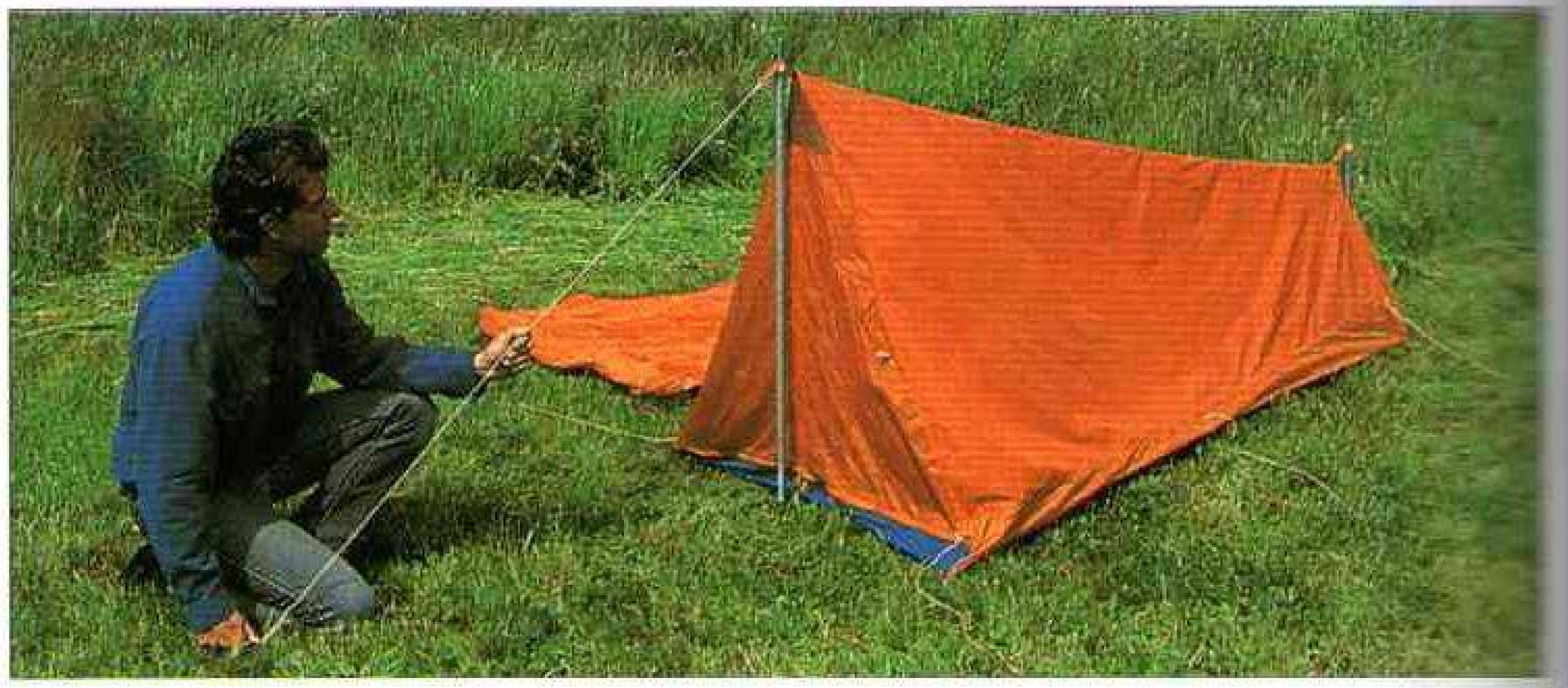

4 Sujete el otro palo al ojal de la parte frontal superior de la tienda. Levántelo y tense el viento. Trabaje deprisa –si hace viento la tienda será poco estable, y si llueve se le mojará.

5 Coloque el doble techo sobre el palo posterior y el extremo más bajo de la tienda, y fíjelo al suelo.

6 Deslice el doble techo sobre el techo de la tienda enganchando el ojal en el mástil principal de la parte alta de la tienda. Sujete la parte frontal del doble techo al suelo con el viento. Es posible que tenga que mover el viento de la tienda interior hacia el palo para que el doble techo se ajuste bien.

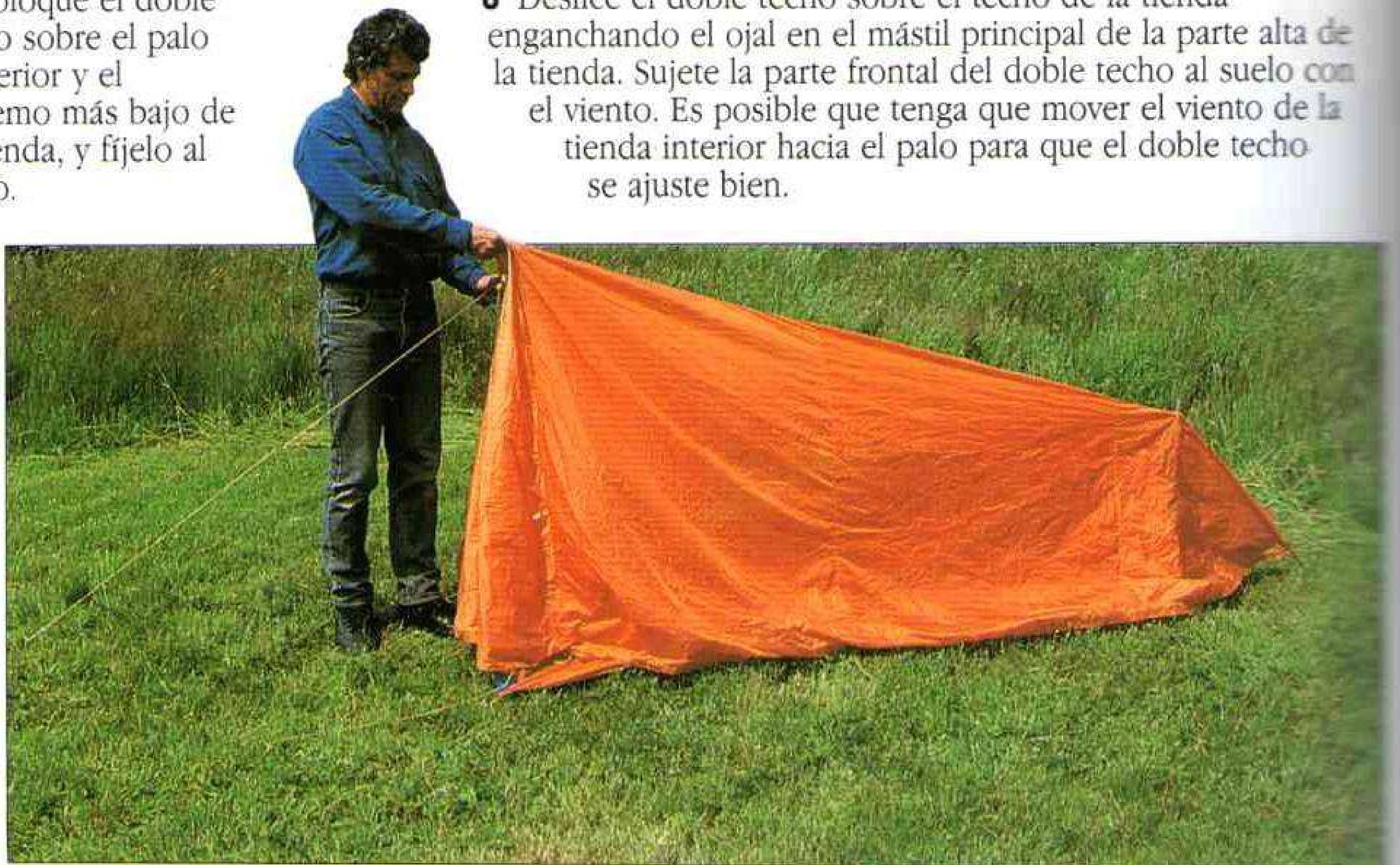

PRECAUCIÓN

No utilice las cintas de goma de la tienda para sacar clavijas clavadas en un lugar equivocado porque las rompería. Para sacar las clavijas de un terreno duro, utilice otra clavija o un gancho especial para este fin.

DESMONTAR UNA TIENDA DE CAMPAÑA

CUANDO DESMONTE UNA tienda, guarde cada pieza de forma ordenada, para que pueda encontrarla rápidamente cuando sea necesario. Si está lloviendo, guarde todas las piezas de la tienda lo más rápidamente posible, en especial la tienda interior. Recuerde que usted es vulnerable si no tiene un refugio donde cobijarse, y que debe desplazarse hasta su siguiente campamento lo más rápidamente posible y levantar su tienda antes de que anochezca. Si el clima es agradable y seco, antes de guardar la tienda extienda todas sus partes al aire libre durante un rato para permitir que el viento y el sol sequen la humedad de la tienda interior y del doble techo. Esto evitará la aparición de moho durante su almacenamiento.

1 En primer lugar, saque todas las clavijas que sujetan la tienda interior y el doble techo. Si el viento es intenso, retenga montada parte de la estructura para evitar que el viento se la lleve. Tenga especial cuidado con las clavijas, nunca utilice los vientos o las cintas de goma para arrancarlas del suelo, y cuéntelas siempre.

Saque las clavijas del suelo teniendo cuidado de no doblarlas

2 Abra la cremallera de la puerta del doble techo y saque el palo frontal. Limpie la tierra o la humedad del palo, desmóntelo cuidadosamente y coloque las distintas partes directamente en su bolsa.

3 Saque el otro palo y desmóntelo. Límpielo y séquelo antes de guardarlo en su bolsa. Extienda la tienda sobre el suelo para poder recoger fácilmente los vértices del tejado.

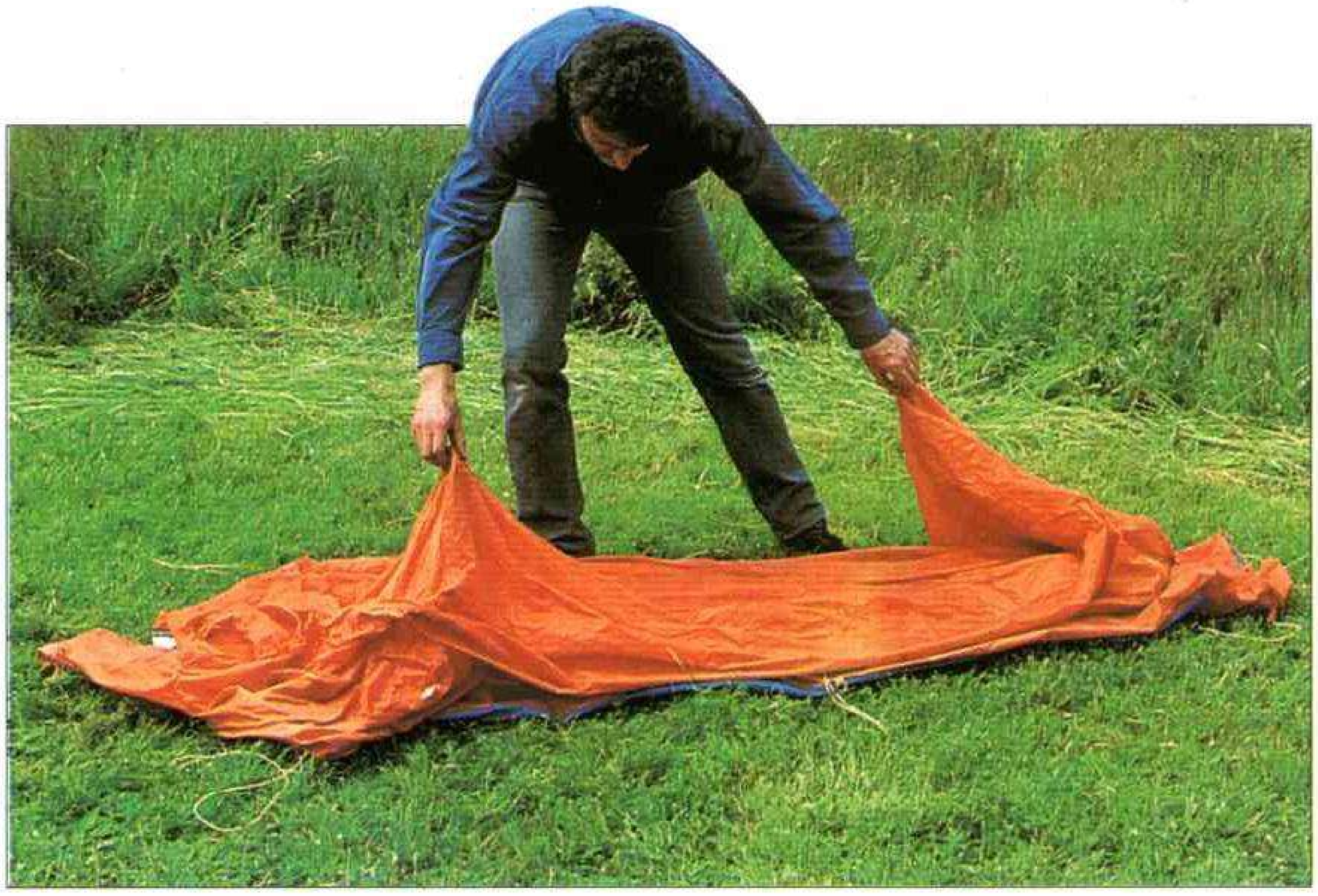

4 La tienda interior y el doble techo deben enrollarse y guardarse separadamente para poder distinguirlos rápidamente al montar de nuevo la tienda. Para guardar el doble techo, levántelo de la tienda interior sujetándolo de los ojales en los vértices del techo. Tenga cuidado de no levantar también la tienda interior.

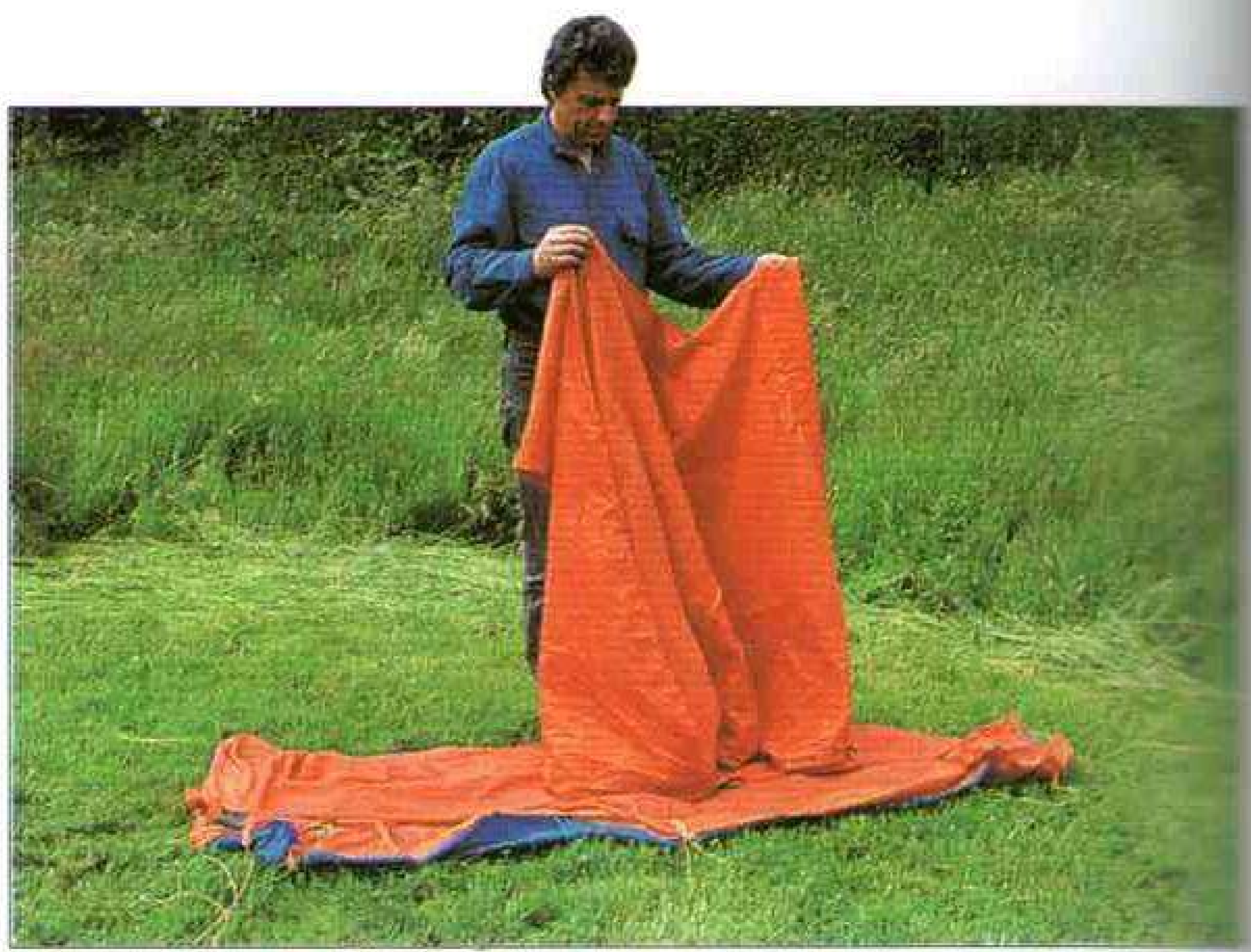

5 Sin dejar de sujetar el doble techo por los ojales, levántelo para que cuelgue según su forma natural. Después dóblela por la mitad para formar una gran «V». Asegúrese de que no esté mojada antes de doblarla.

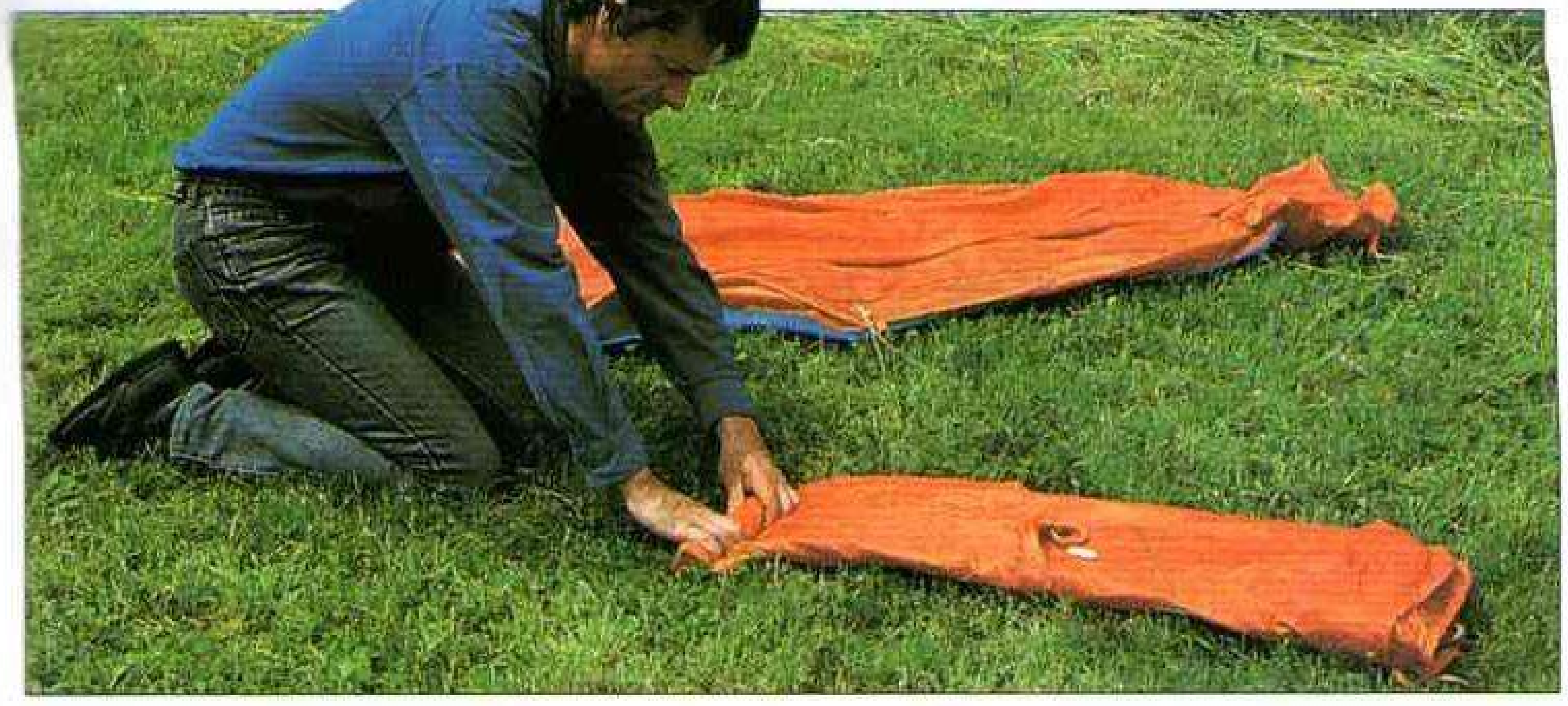

6 Coloque el doble techo sobre un trozo seco de terreno y comience a enrollarlo lo más apretado posible. Doble los laterales conforme lo va enrollando, teniendo cuidado de no enredar los vientos.

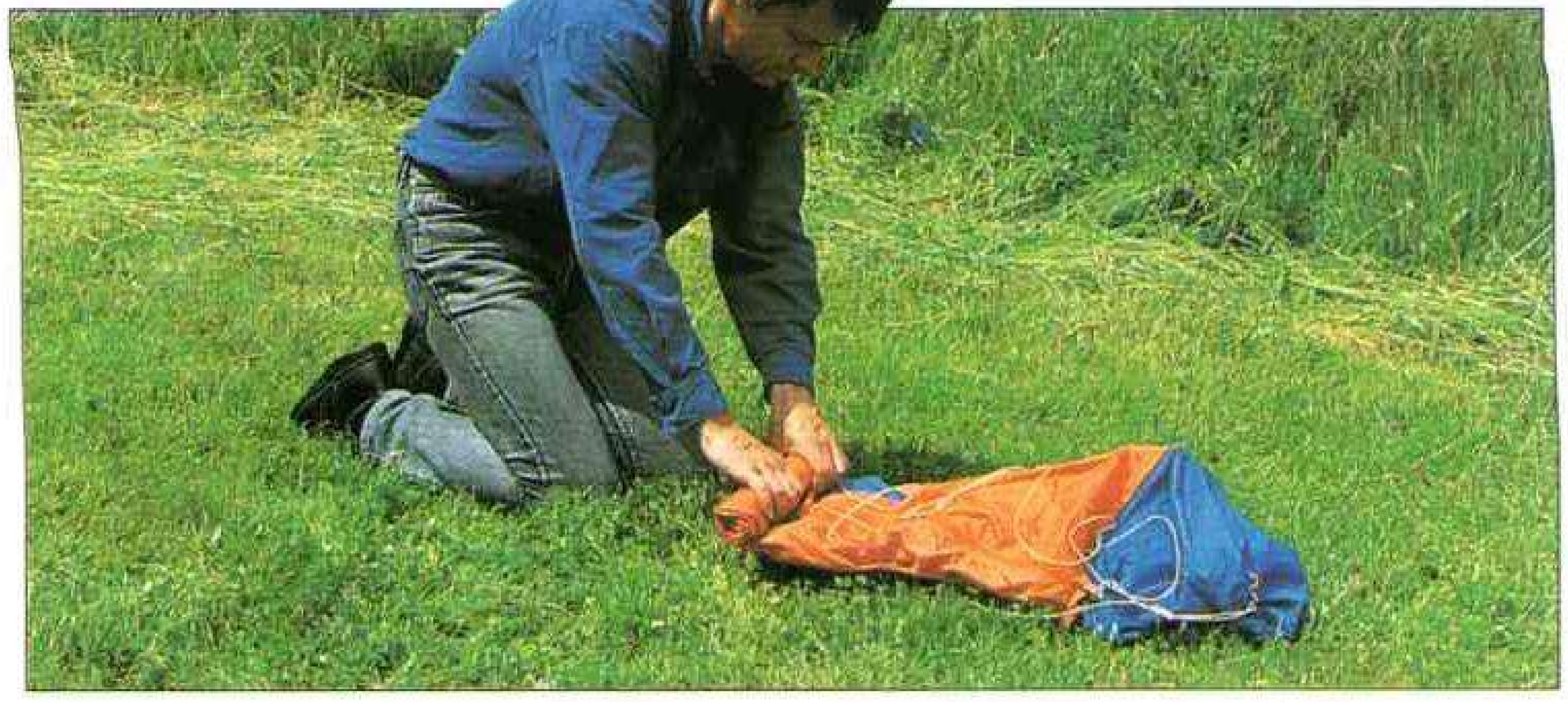

7 Sacuda la tienda interior de manera que los vientos cuelguen libremente, colóquela sobre el suelo y enróllela lo más apretada posible. Doble los laterales conforme la enrolla.

8 Coloque los vientos cuidadosamente en el rollo antes de terminarlo, asegurándose de que no están enredados ni tienen nudos. Si éste es el caso, es el momento de deshacerlos.

9 Guarde las clavijas en una bolsa resistente para que no dañen la tienda. Guarde la bolsa con las clavijas dentro de la bolsa principal de la tienda. El barro o la humedad que pudiera quedar en las clavijas puede corroerlas o dañarlas.

10 Limpie los palos y métalos en una bolsa resistente, y después dentro de la bolsa de la tienda. Si los mantiene limpios, ajustarán fácilmente cuando los utilice.

11 Meta la tienda interior y el doble techo en la bolsa de la tienda. Es una excelente idea colocar antes cada artículo en una bolsa de plástico individual para dar una protección adicional.

PRECAUCIÓN

Asegúrese de no perder ninguna clavija, viento o palo al desmontar su tienda. También debe tener cuidado de no dañar ninguna de sus partes; si esto ocurriera, debe repararla antes de utilizarla de nuevo.

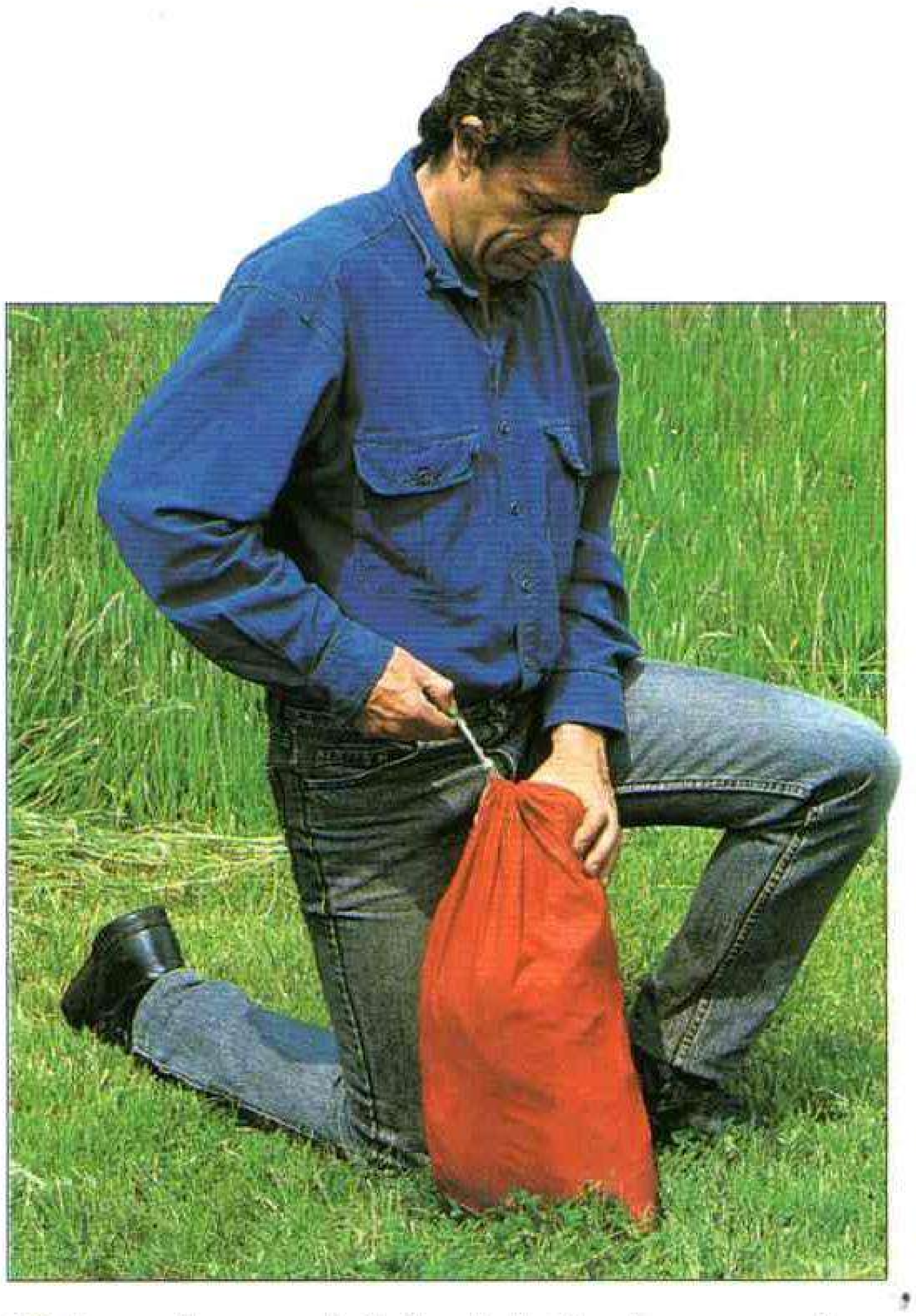

12 Antes de cerrar la bolsa de la tienda, compruebe que todo esté guardado, y que no ha olvidado ninguna clavija o palo. Si la hierba es alta es fácil extraviar alguno de estos artículos.

CUIDADO Y REPARACIÓN DE UNA TIENDA

- Su tienda debe estar totalmente seca antes de guardarla; además de ser más ligera, el tejido se deteriorará menos que si se almacena húmedo.
- Evite montar la tienda o desmontarla cuando hace mucho calor. El aislante impermeable puede derretirse con un calor excesivo.
- Compruebe todas las partes de su tienda. Si es necesario, repárelas inmediatamente.
- Antes de almacenar una tienda, lávela con agua fría. Cuélguela para que se seque antes de guardarla.

Parches. *Puede comprar parches especiales de vinilo con los que remendar temporalmente su tienda.*

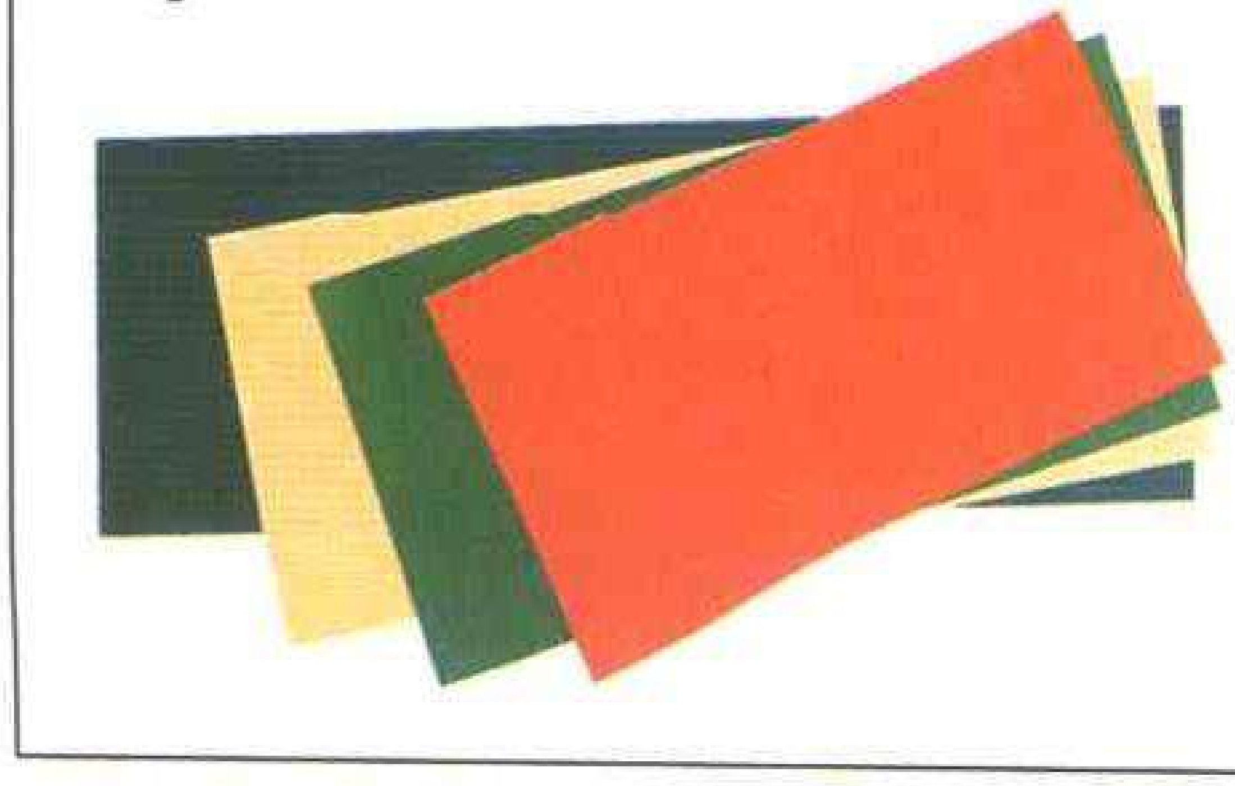

LA CONSTRUCCIÓN DE UN REFUGIO

PROTEGERSE DEL VIENTO, de la lluvia y del sol para poder descansar y dormir resulta vital para sobrevivir. Cuanto mejor sea el refugio construido, más cómodo estará y mejor descansará. El estar descansado es esencial para su salud física y también para su estado psicológico, que determina la sensibilidad y la lógica de su pensamiento, y su decisión de sobrevivir. Un cobertizo es probablemente el tipo de refugio más sencillo y rápido de construir, y es, además, adecuado para la mayor parte de terrenos. Debería construirse en un lugar protegido, seguro y llano *(véase página 40)*, y de tal manera que el techo estuviera de cara al viento. La construcción de su refugio dependerá enteramente de los materiales de que disponga. Si dispone de una lona o de una cubierta de plástico, debe emplearla para construir el techo del refugio; en caso contrario, tendrá que improvisar con los materiales que pueda encontrar.

El travesaño debe ser más alto que el alcance de su brazo

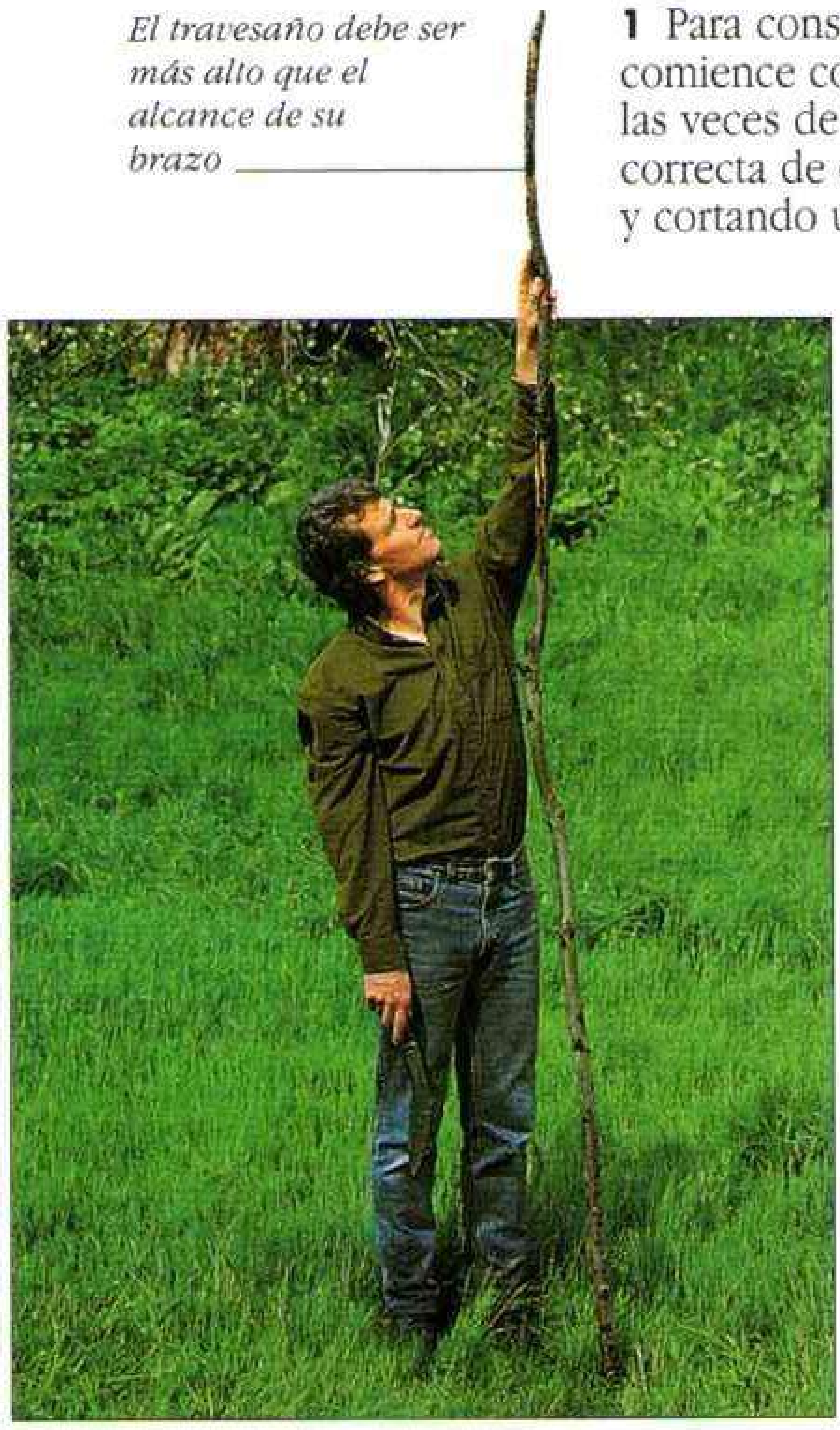

1 Para construir un refugio a un agua, comience cortando una rama larga que hará las veces de travesaño. Mida la longitud correcta de esta rama extendiendo su brazo y cortando unos 60 cm por encima de este punto. Intente que la rama sea lo más recta posible, y recorte todas las ramas que broten de ella. No utilice una rama muerta ya que un viento fuerte podría romperla.

2 Corte dos soportes en forma de «Y» cuya altura llegue hasta su pecho, y córtelos con una longitud adicional de 30 cm, que se clavarán en el suelo. Afile los extremos inferiores.

Afile los extremos de los soportes para que puedan clavarse en el suelo

3 Con la ayuda de una piedra, clave los soportes en el suelo hasta que estén bien seguros y no se [illegible]. La distancia entre ellos [illegible] ser de unos 60 cm menos que la longitud del travesaño.

4 Apoye el travesaño sobre los dos soportes en «Y», dejando la misma distancia en ambos extremos. El travesaño debe ser ligero, aunque lo suficientemente resistente para soportar el peso del techo terminado, así como la acción de la lluvia y el viento, sin derrumbarse.

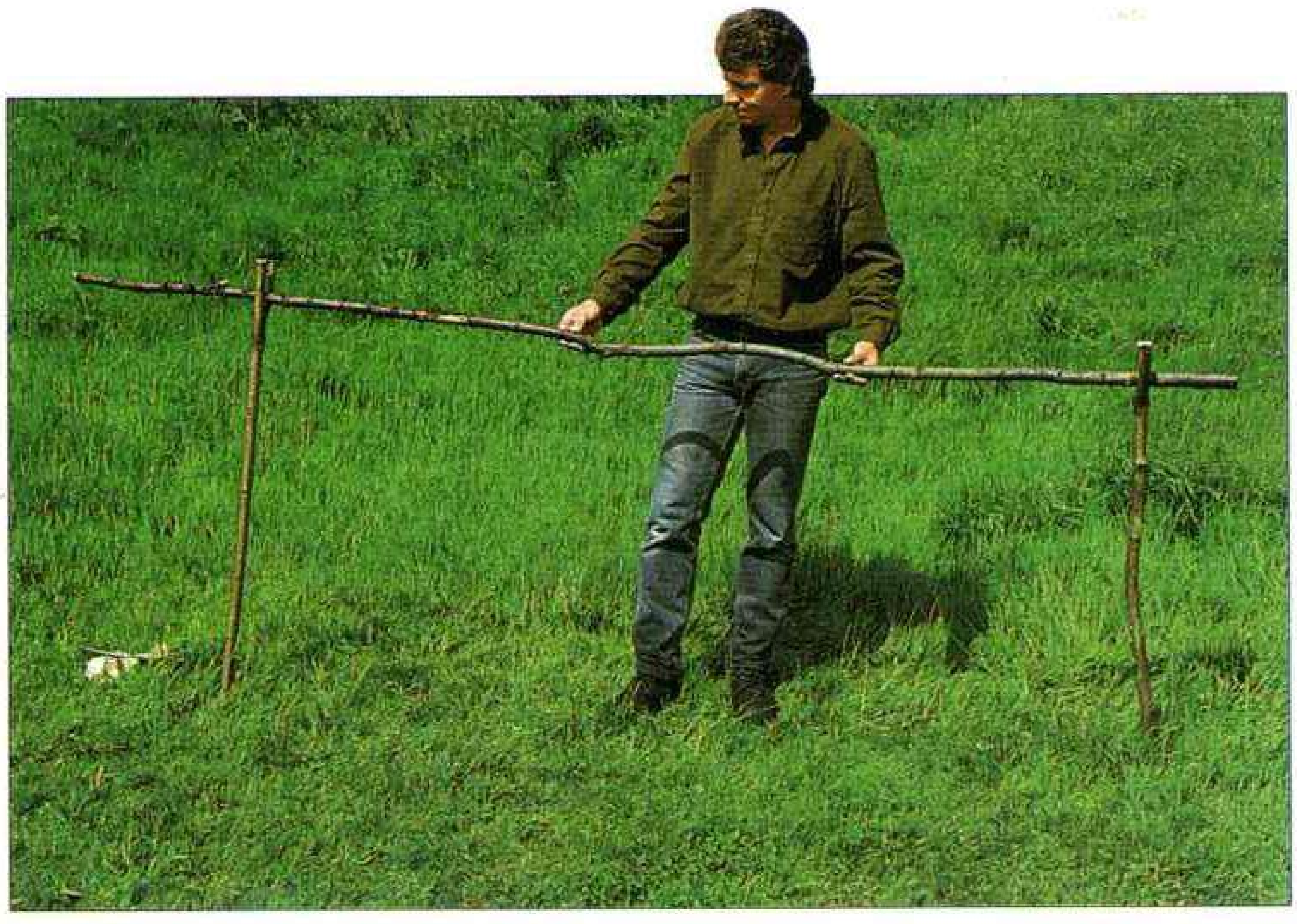

5 Corte varias ramas resistentes lo suficientemente largas para inclinarlas contra el travesaño en un ángulo de 45 grados con el suelo. El techo inclinado permitirá el drenaje del agua de la lluvia y evitará su estancamiento. Colóquelas a unos 20 cm de distancia entre sí. Deben sobrepasar el travesaño unos 10 cm.

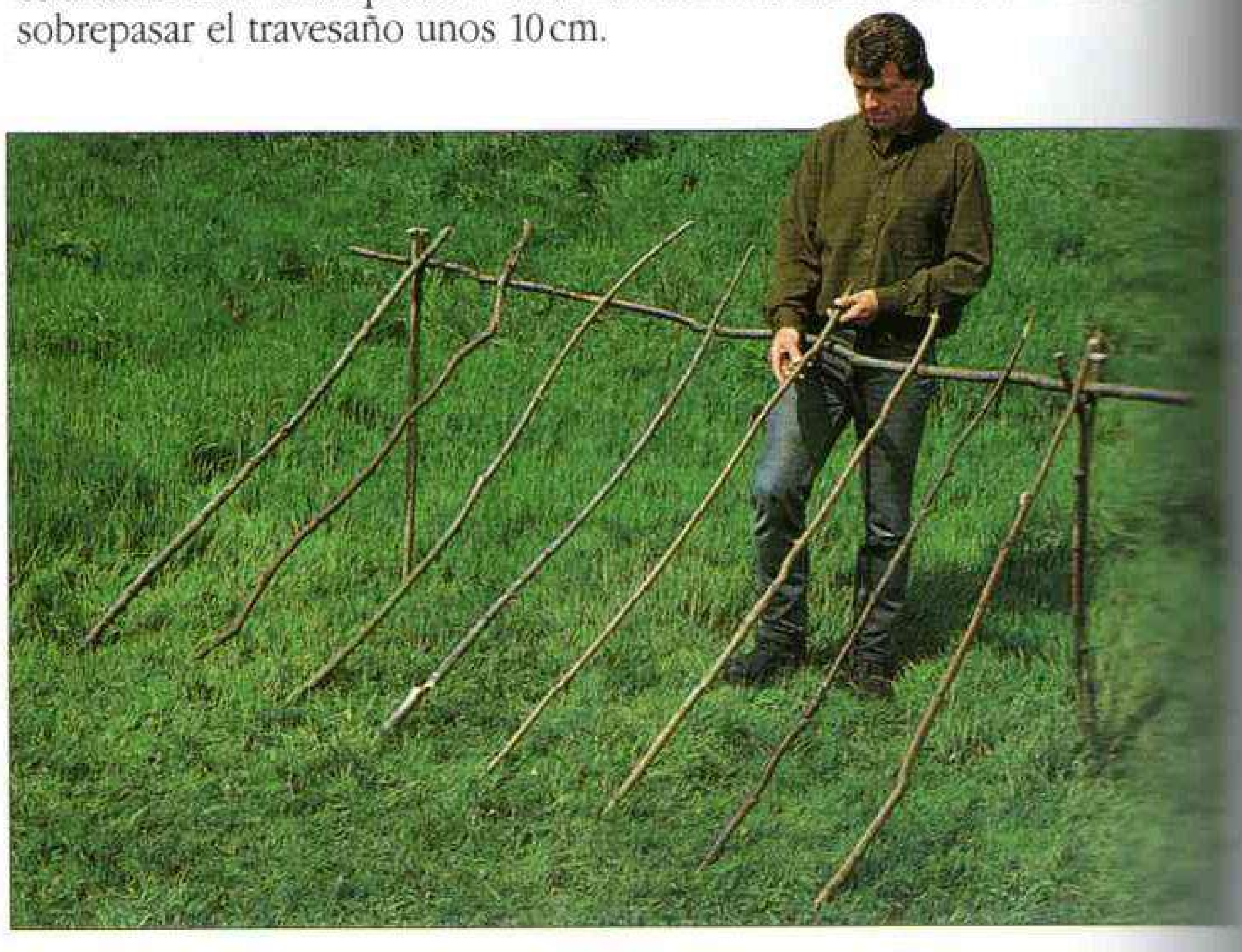

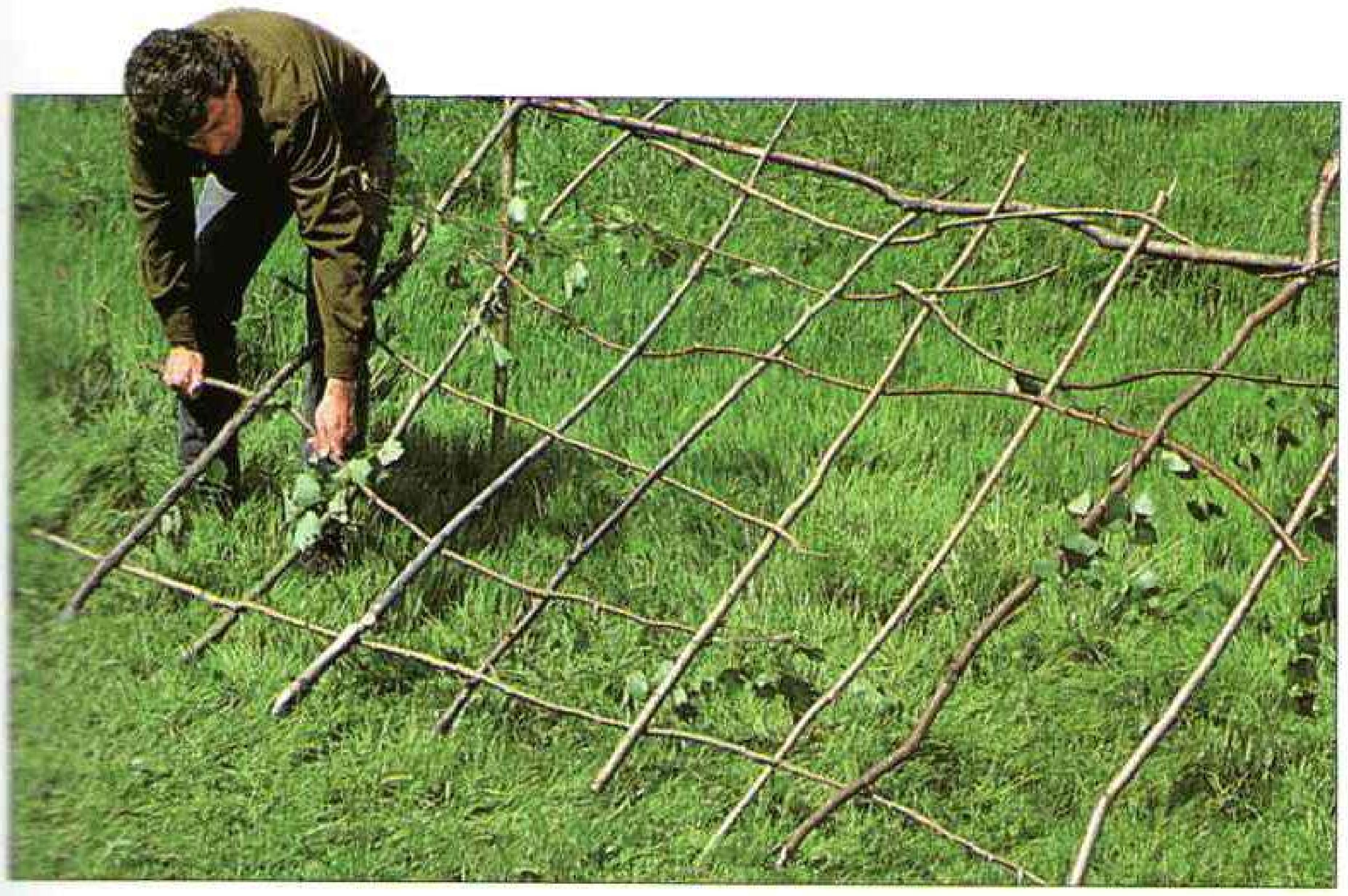

6 Corte varias ramas verdes rectas, más bien fuertes, dejando algunas hojas en ellas. Entretéjalas alternativamente por encima y por debajo de las ramas inclinadas del techo, hasta que todas ellas formen una red relativamente firme. Entreteja también una fila de ramas por encima y por debajo del travesaño para asegurarlo mejor.

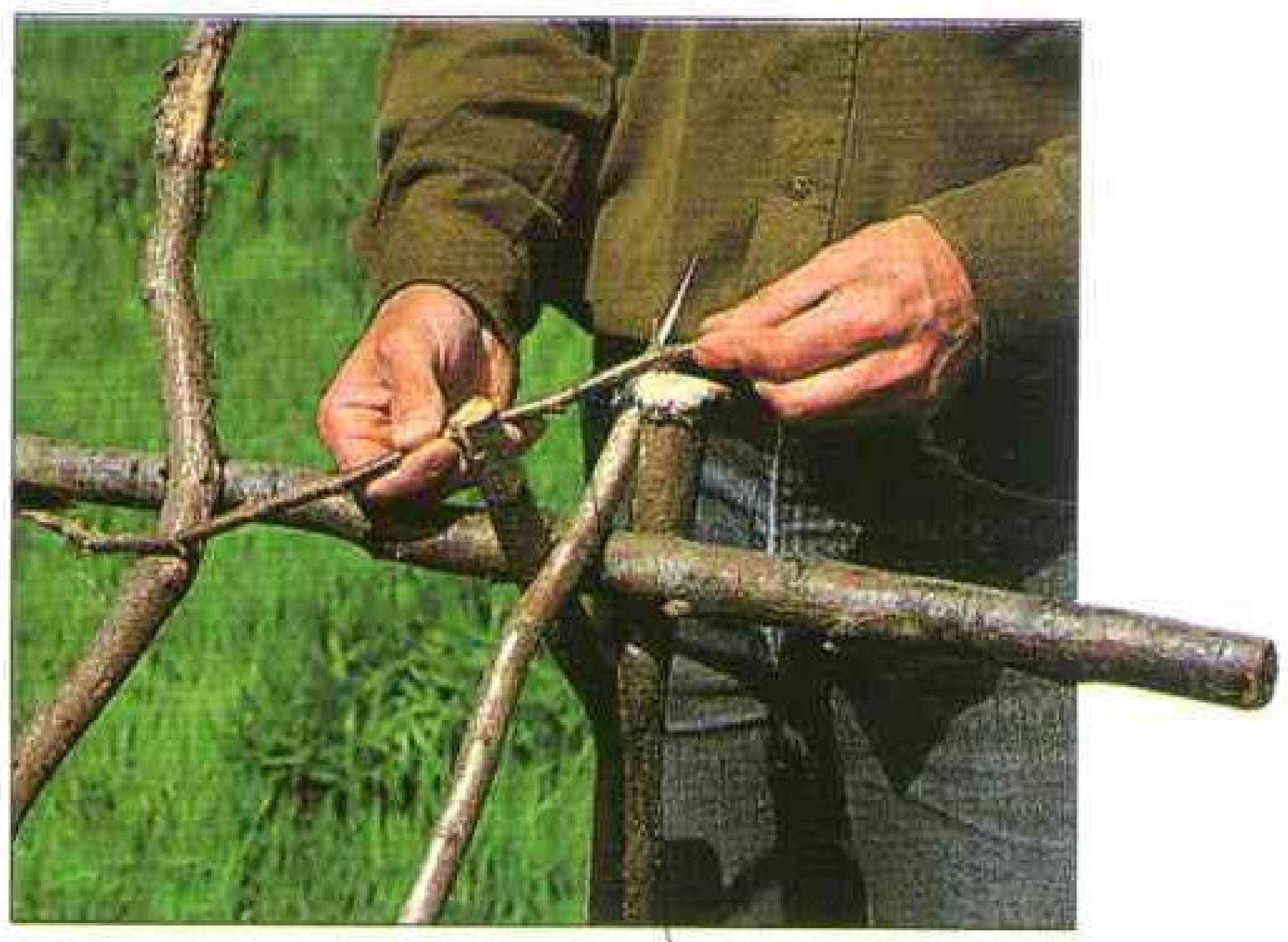

Entreteja las ramas inclinadas con el travesaño para dar mayor estabilidad a la estructura y evitar que las ramas se muevan

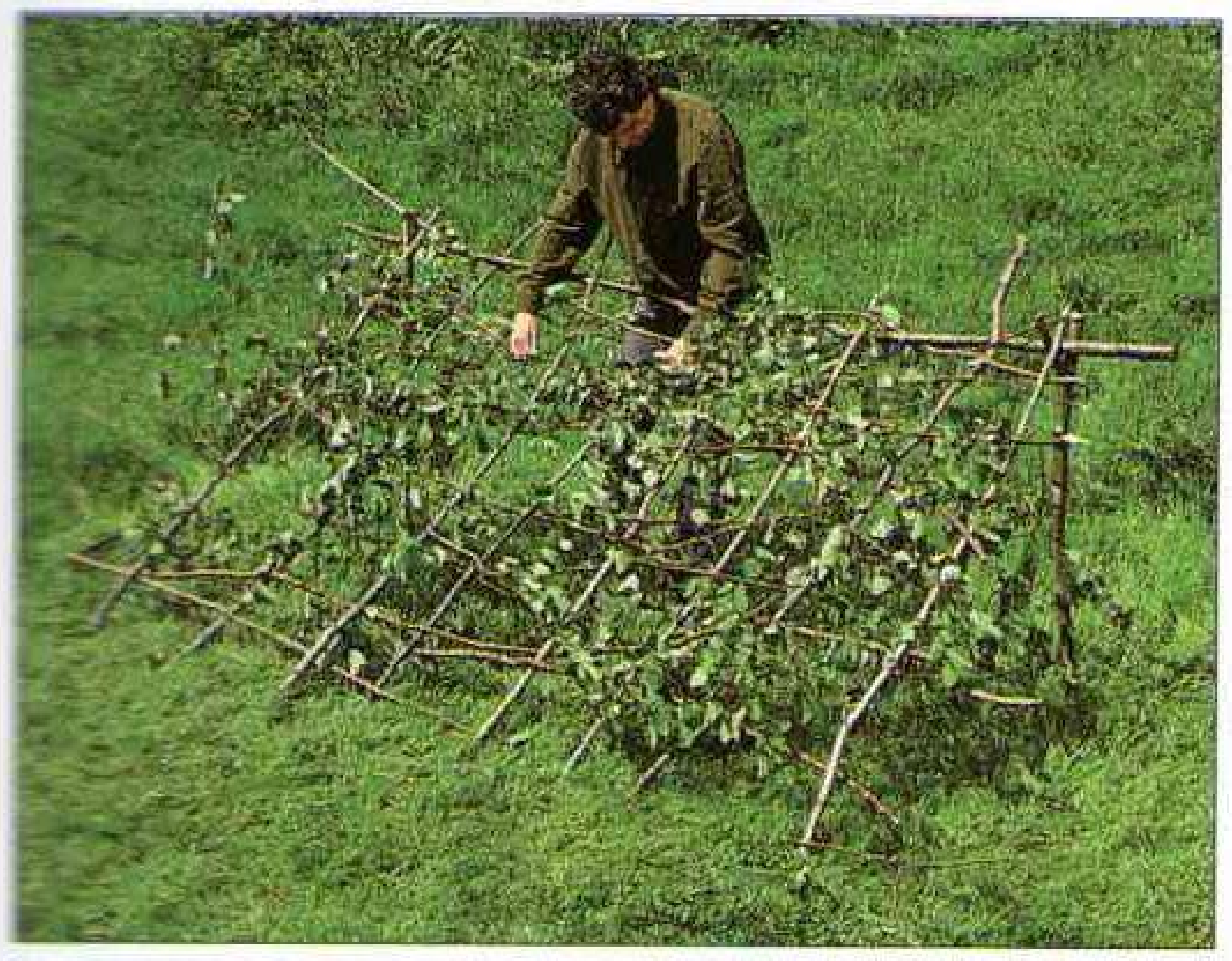

7 Una vez formada la red soporte, entreteja ramas completas, incluyendo sus hojas, para evitar la entrada del viento y de la lluvia. Al emplear las ramas enteras, dará mayor fuerza al refugio, y el follaje permanecerá vivo un mayor tiempo, en especial si es caducifolio.

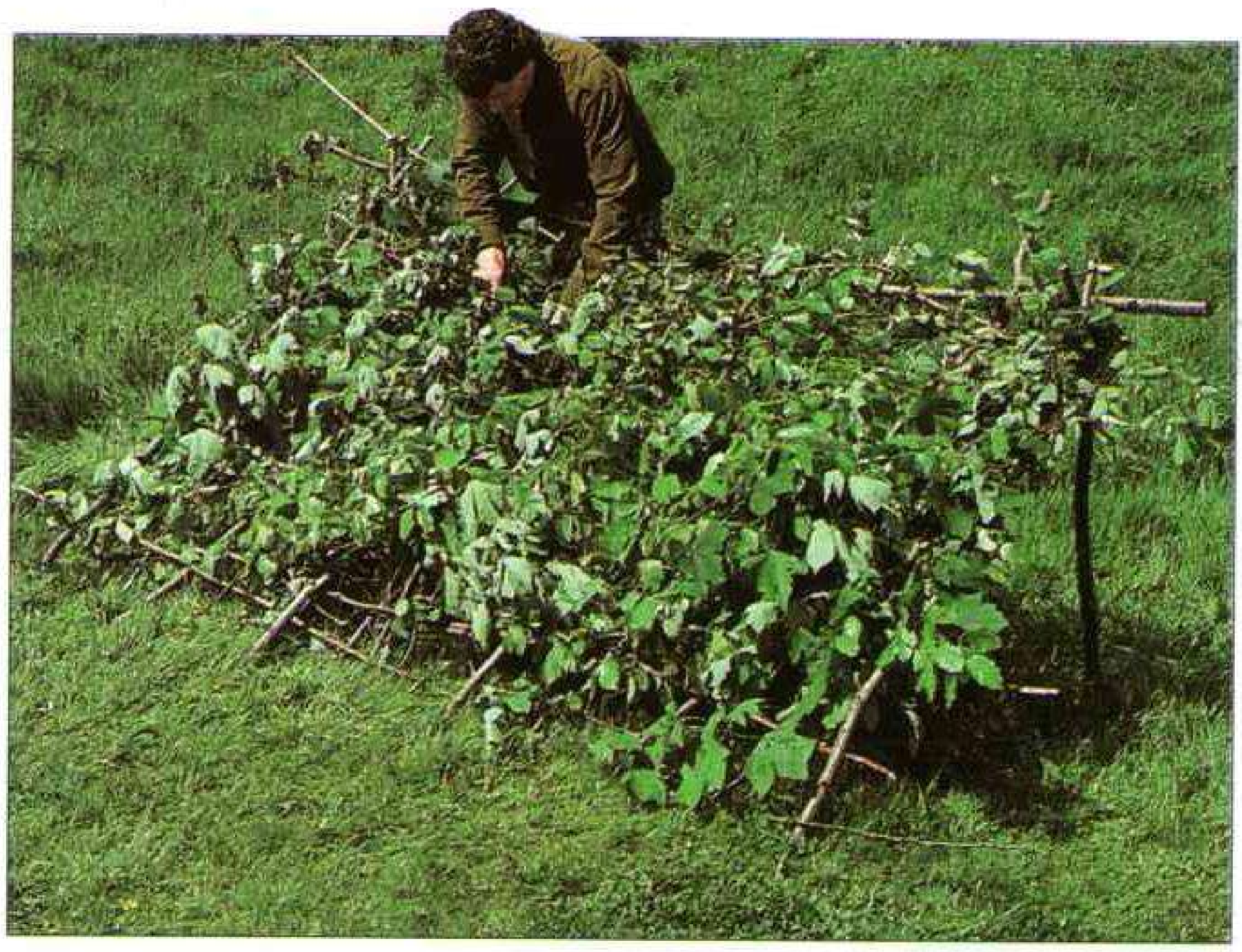

8 Continúe entretejiendo todo tipo de follaje hasta lograr la densidad deseada para su techo. Algunos tipos de hoja se secan más rápidamente que otras, y deberán ser reemplazadas transcurridos algunos días. Resulta mejor entretejer las ramas nuevas por encima de las viejas para ir construyendo un techo en varias capas. Son preferibles las ramas de árboles con hojas pequeñas, ya que las hojas grandes se rasgan, mueren más rápidamente, y son más difíciles de tejer para formar una superficie plana.

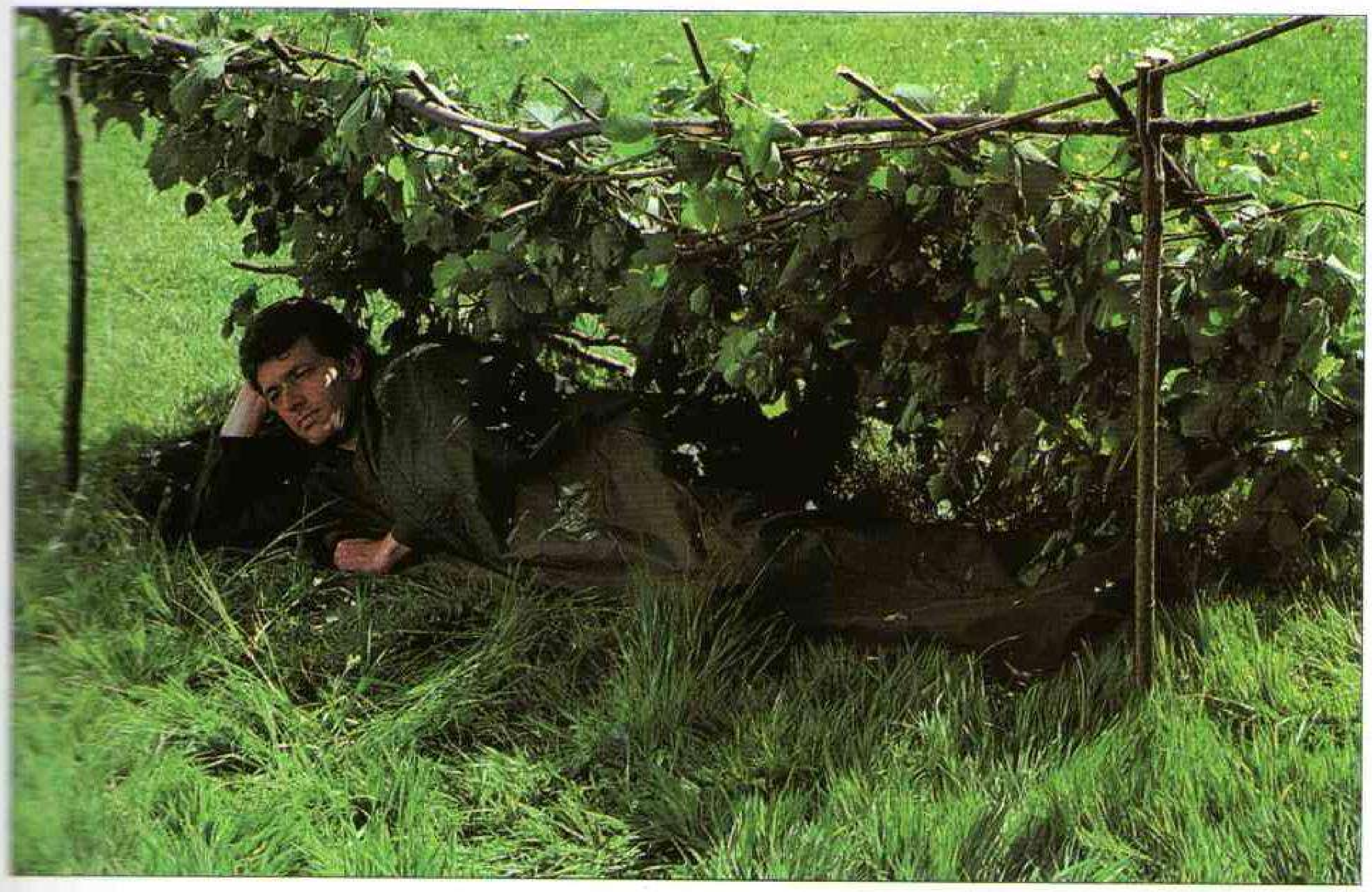

9 El refugio terminado puede requerir algunas mejoras: puede colocar rocas grandes a lo largo de la base para sujetarla en caso de vientos fuertes, o añadir paredes laterales. Encienda el fuego delante del refugio, pero mantenga las llamas alejadas del techo. Renueve las ramas cuando se marchiten.

EL EMPLEO DE MATERIALES ARTIFICIALES

El empleo de materiales artificiales facilita la construcción de un refugio. Las láminas de polietileno, las bolsas de plástico y los cordeles para embalar son comunes entre la basura encontrada en el campo. Las hojas de hierro galvanizado y las cajas de madera pueden resultar útiles.

Otros materiales. *Los refugios pueden construirse a partir de material de desecho.*

REFUGIOS EN LA NIEVE

SI LAS TEMPERATURAS permanecen bajo cero, la construcción de refugios en la nieve es relativamente fácil. Protegerlos del viento es prioritario, ya que el viento puede disminuir drásticamente la temperatura *(véase página 141)*. Las temperaturas inferiores a −10 °C resultan desagradables, por lo que la construcción de refugios que puedan retener muy bien el calor se convierte en una necesidad. Estos pueden variar desde una simple trinchera en la nieve hasta un iglú, en cuya construcción bien puede emplear unas cuantas horas. En un refugio permanente, como un iglú, el aire frío y más denso se disipará del interior si cava un canal que desvíe el aire hacia la parte inferior del iglú y de allí hacia afuera. Una ventilación adecuada es imprescindible en cualquier refugio en la nieve para evitar la asfixia.

LA CONSTRUCCIÓN DE UN IGLÚ

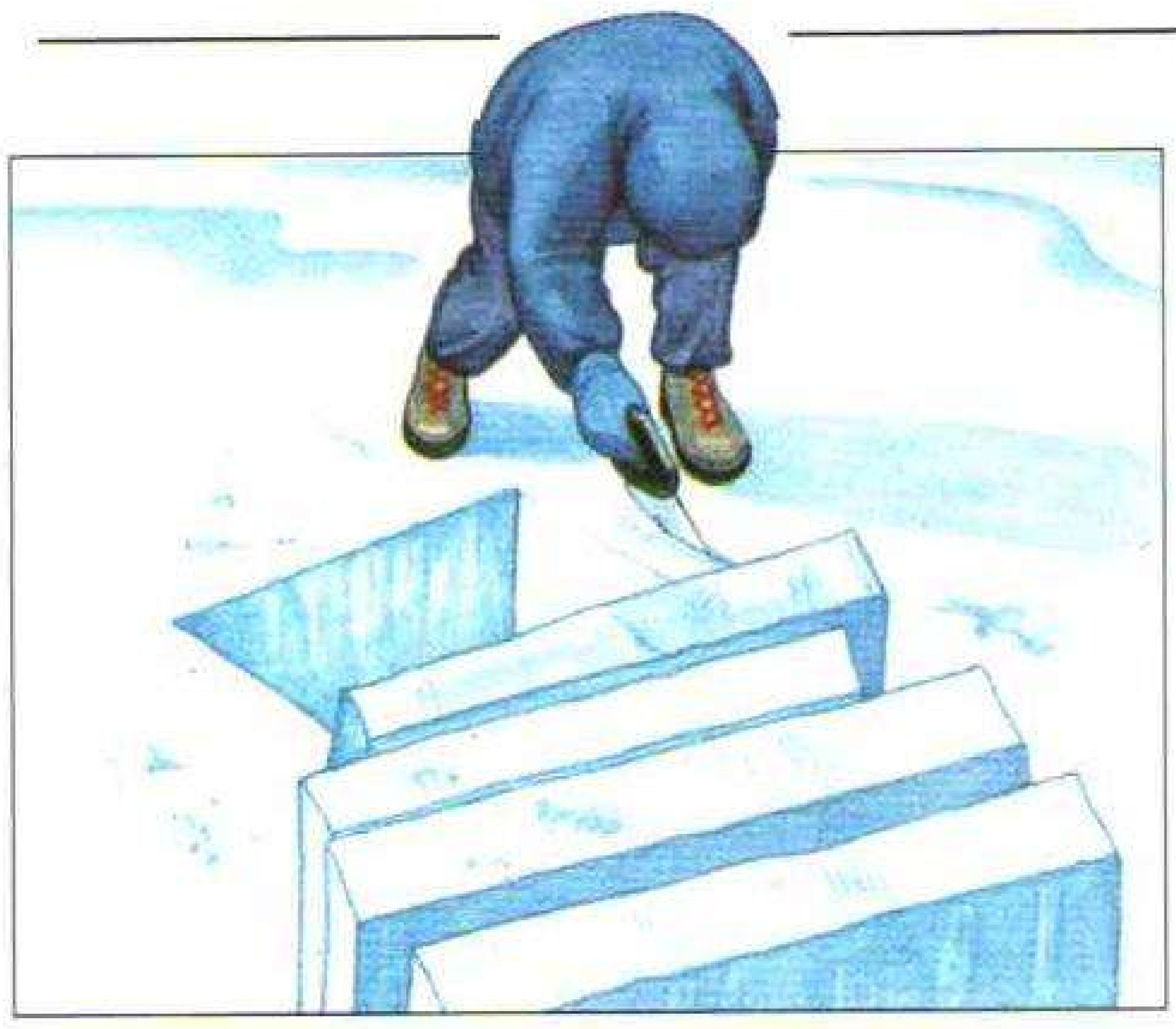

1 Corte bloques de nieve dura empleando una sierra para nieve o un cuchillo grande. Cada bloque debe tener aproximadamente 1 m de largo, 40 cm de alto y 20 cm de ancho.

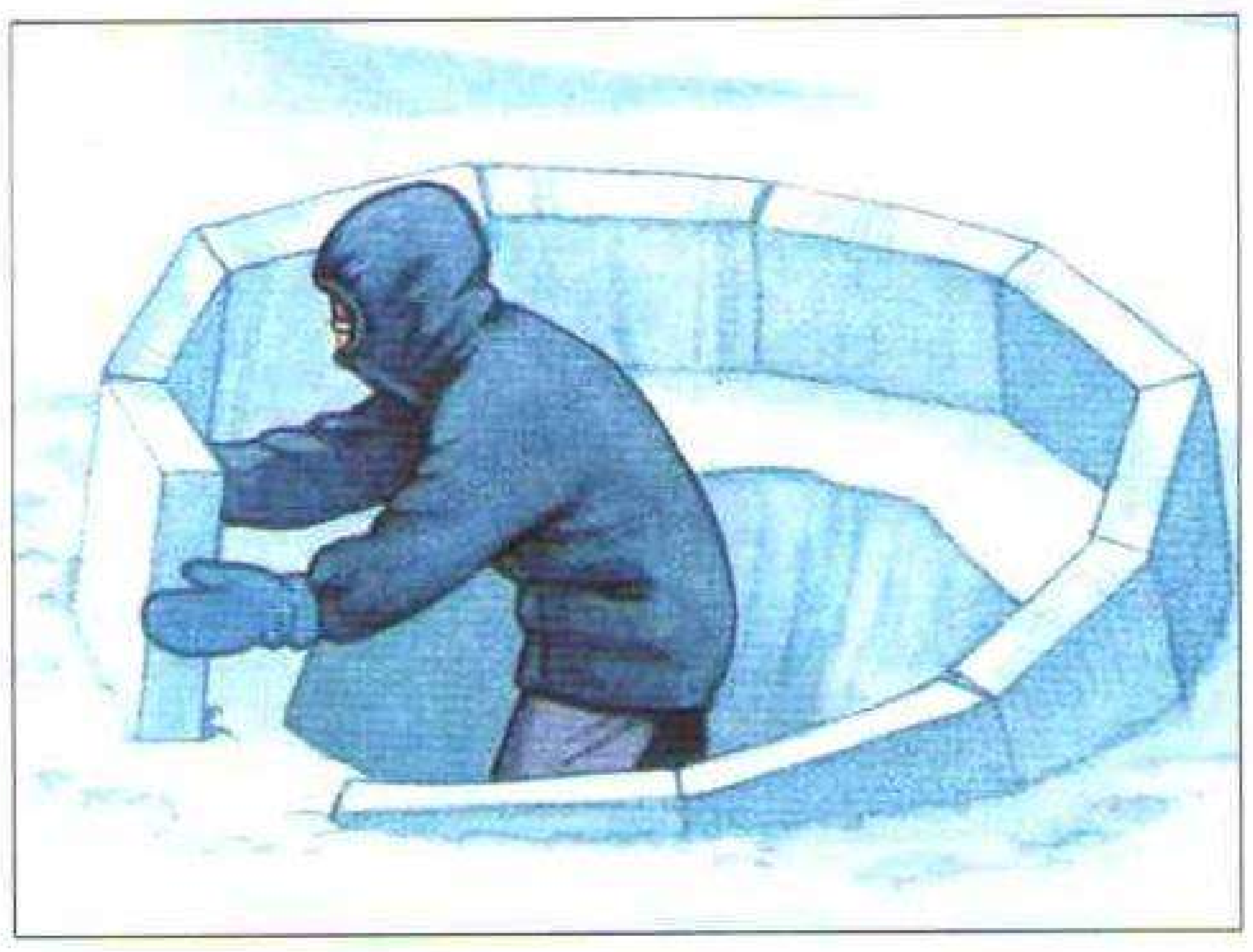

2 Forme un círculo con los bloques alrededor del agujero que se ha formado en el suelo. Corte el círculo en espiral desde la parte superior del último bloque hasta el suelo. Así resultará más fácil construir la cúpula.

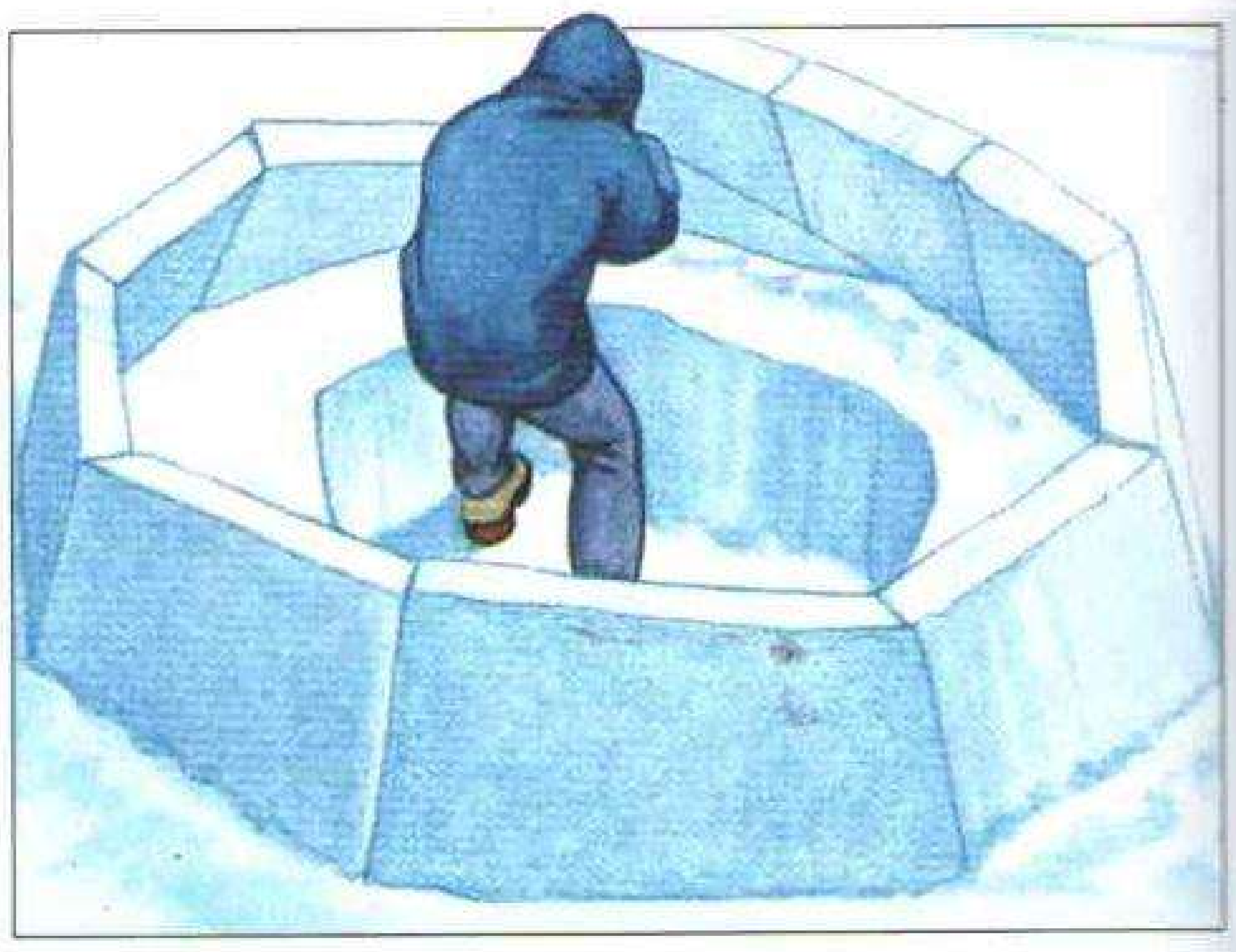

3 Construya las paredes solapando los bloques y colocándolos inclinados hacia el interior. Cave un agujero bajo la pared para la entrada y para disipar el aire frío. Coloque bloques en una pared como plataforma para dormir.

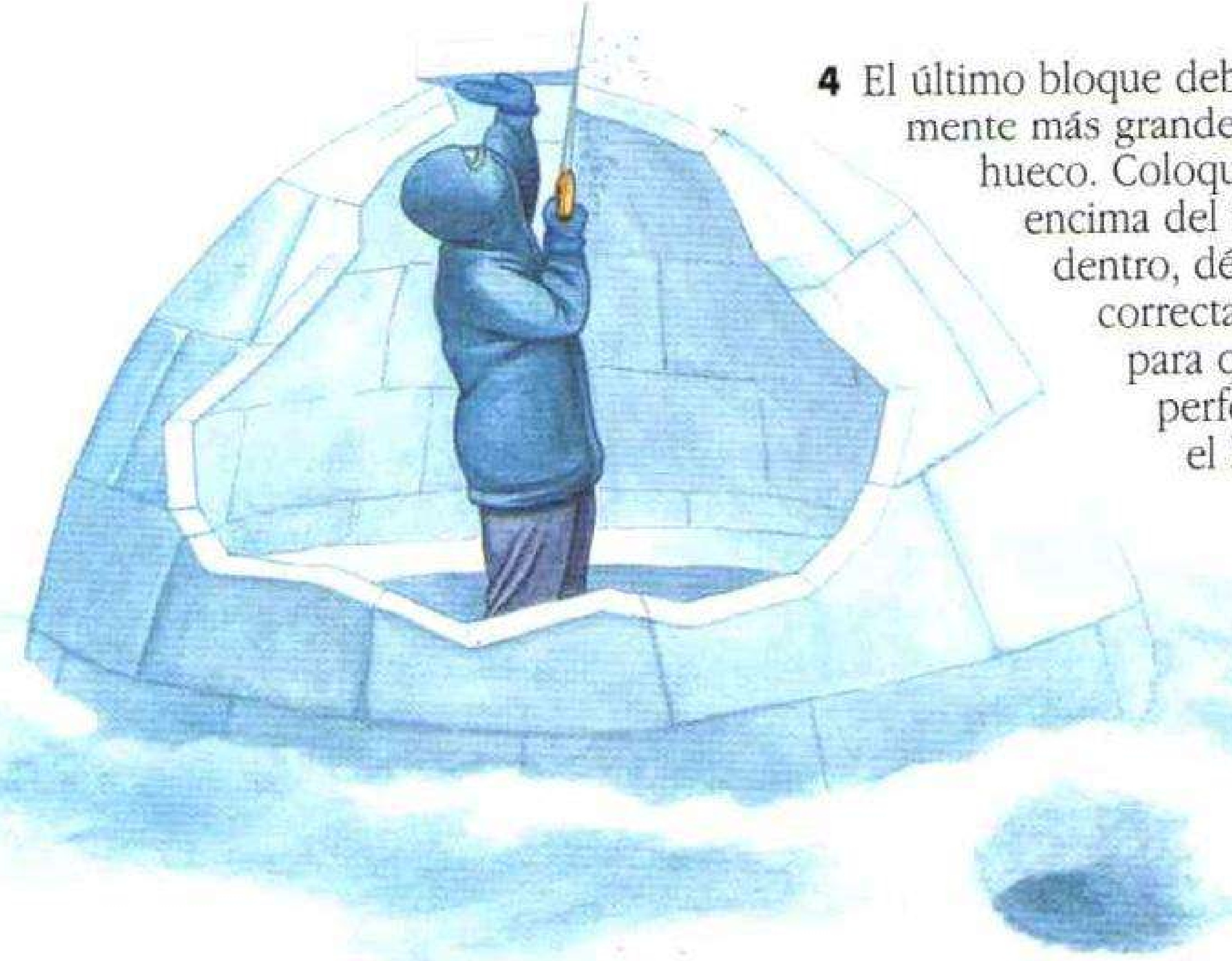

4 El último bloque debe ser inicialmente más grande que el hueco. Coloque el bloque encima del iglú y, desde dentro, déle la forma correcta y deslícelo para que encaje perfectamente en el hueco.

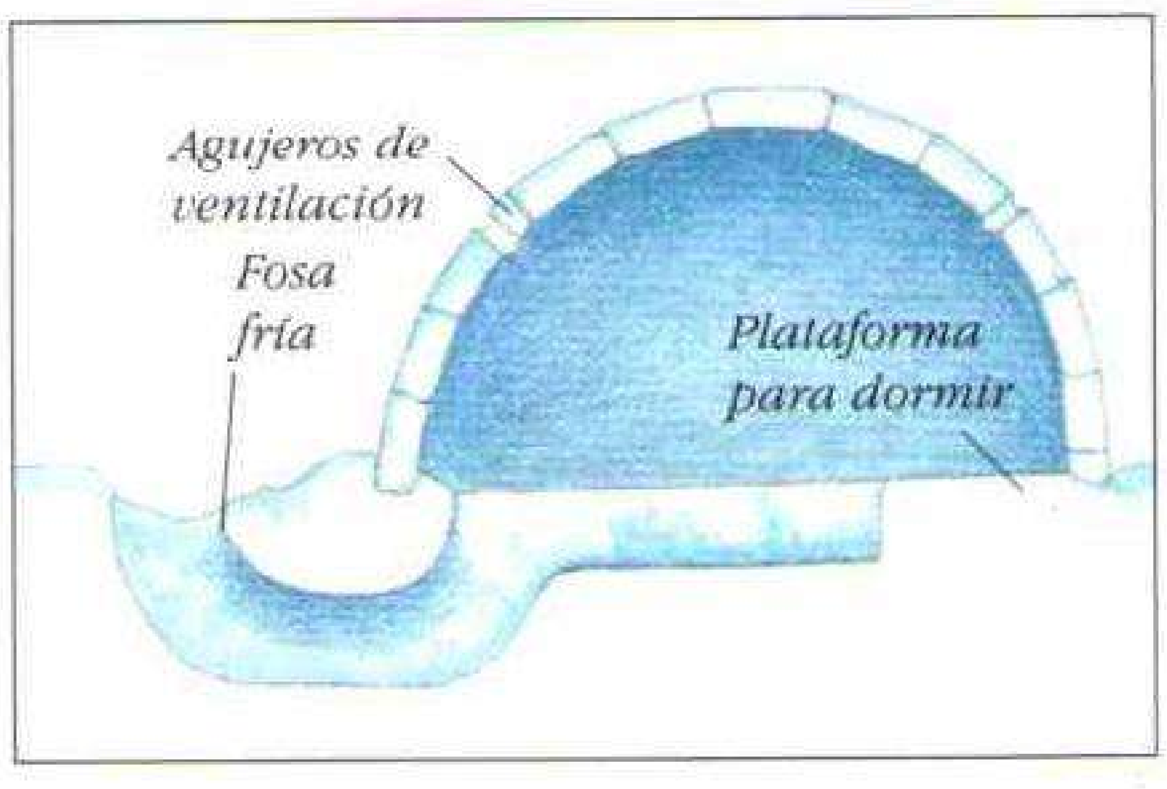

5 El aire caliente proveniente de su cuerpo y de la cocina sube, y queda atrapado por la cúpula. El aire frío baja al túnel y se disipa hacia el exterior. Los agujeros para la ventilación son indispensables; deben realizarse con un hacha especial para hielo.

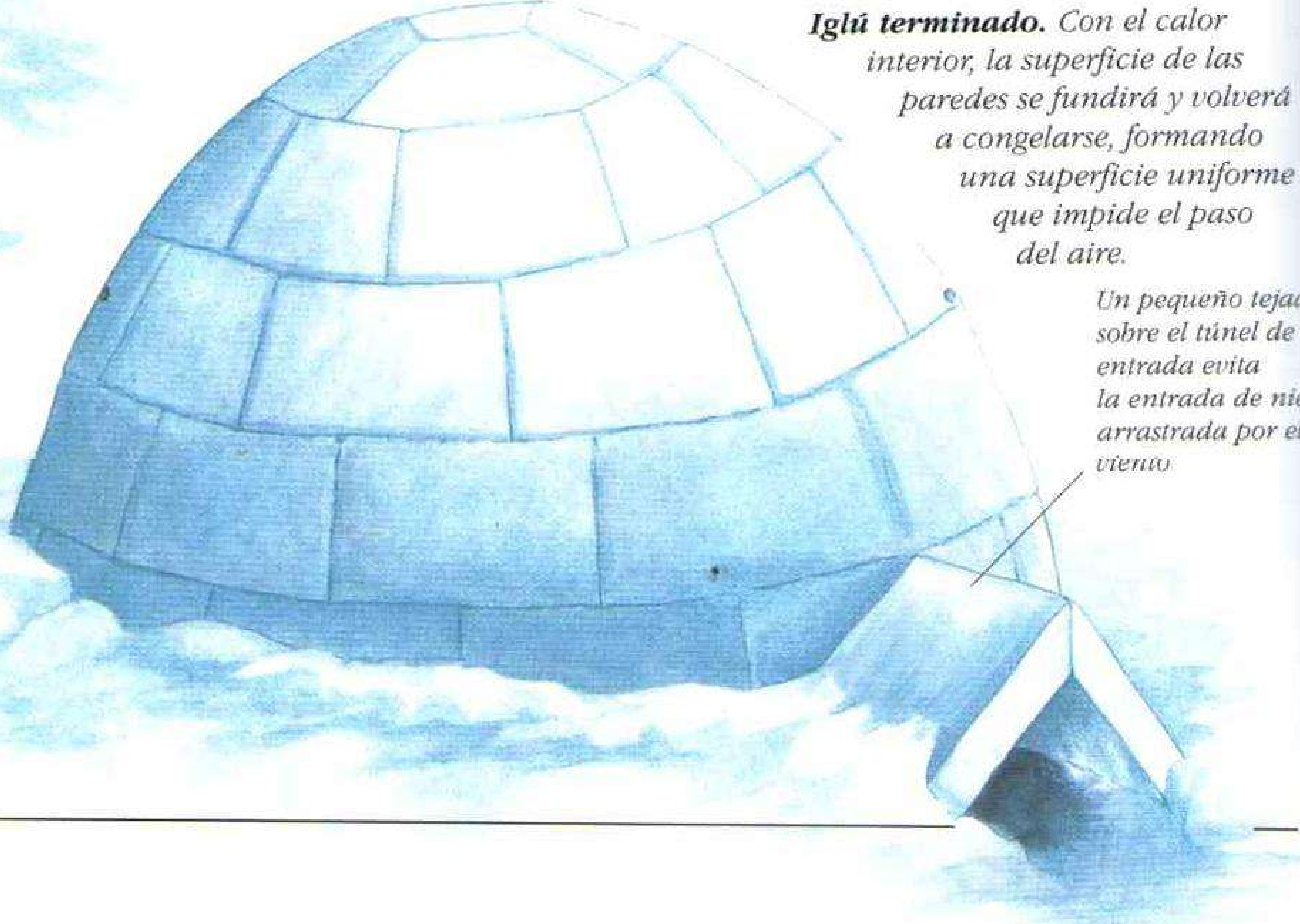

Iglú terminado. *Con el calor interior, la superficie de las paredes se fundirá y volverá a congelarse, formando una superficie uniforme que impide el paso del aire.*

Un pequeño tejado sobre el túnel de entrada evita la entrada de nie arrastrada por el viento

ADVERTENCIA

Es fundamental perforar por lo menos un agujero de ventilación en el techo para evitar la asfixia. El interior del iglú se calentará con el calor de su cuerpo. Sin una ventilación adecuada puede acumularse monóxido de carbono, que es letal. Para construir un iglú se precisa de dos personas.

LA CONSTRUCCIÓN DE UN *QUINZE*

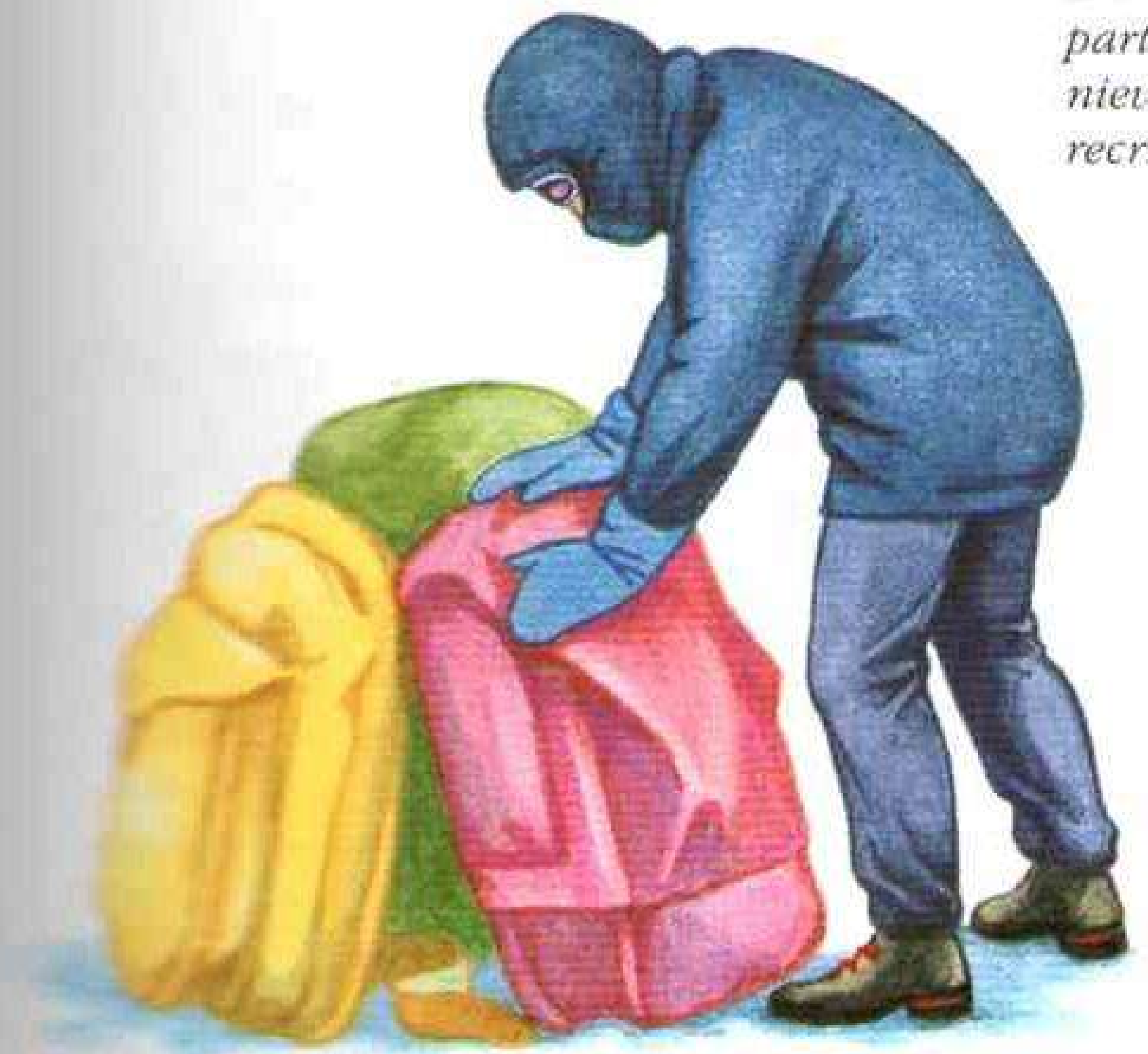

1 Coloque las mochilas y el resto del equipo formando un cono compacto. El equipo determinará la forma interior del refugio y reducirá la cantidad de nieve necesaria para construir el *quinze* (pronunciado «cuinsi»).

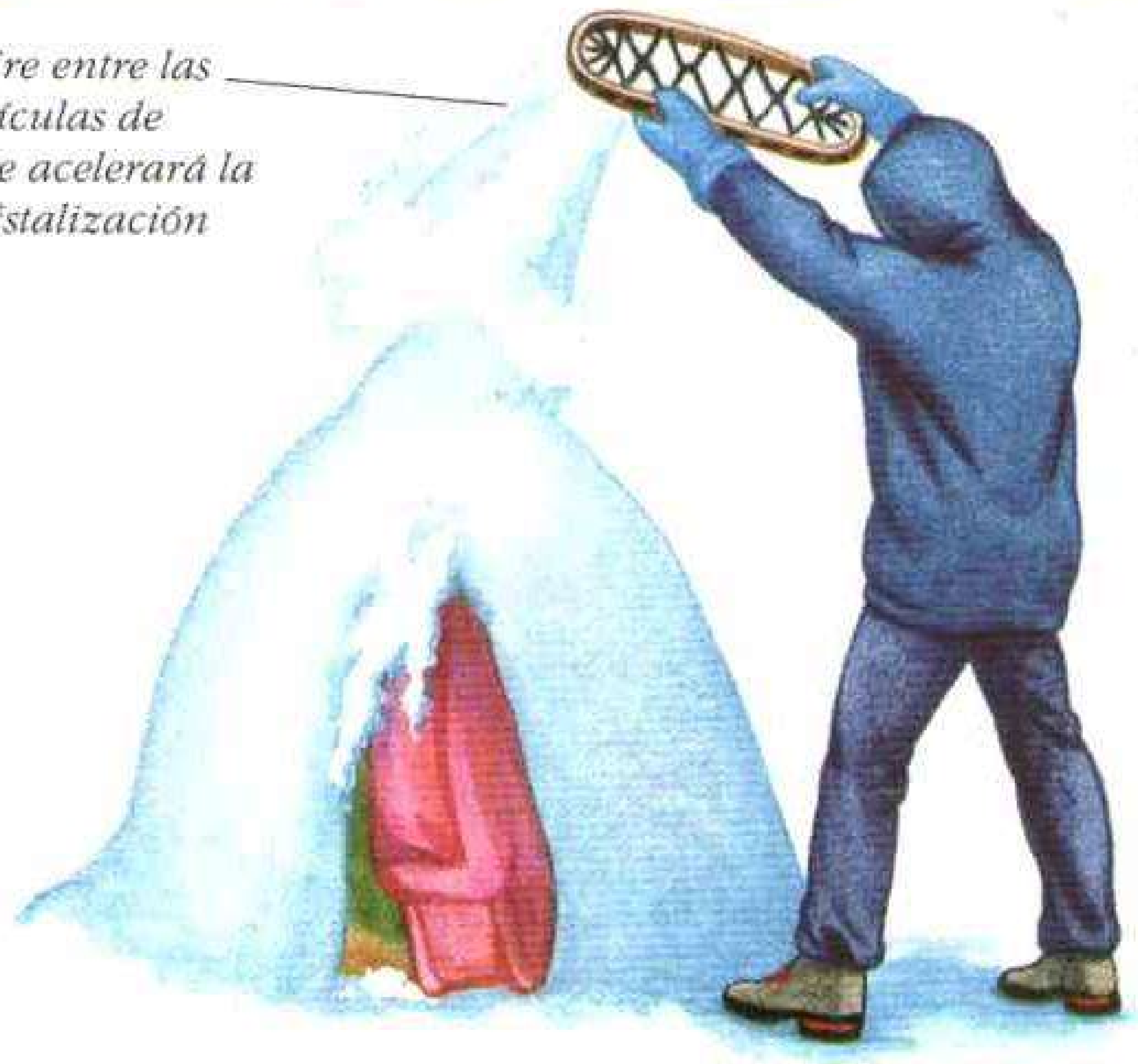

El aire entre las partículas de nieve acelerará la recristalización

2 Con una raqueta para la nieve o una pala, apile nieve sobre las mochilas, y compáctela. Espere al menos 30 minutos a que la nieve se congele antes de añadir más nieve y lograr el grosor deseado para la cúpula.

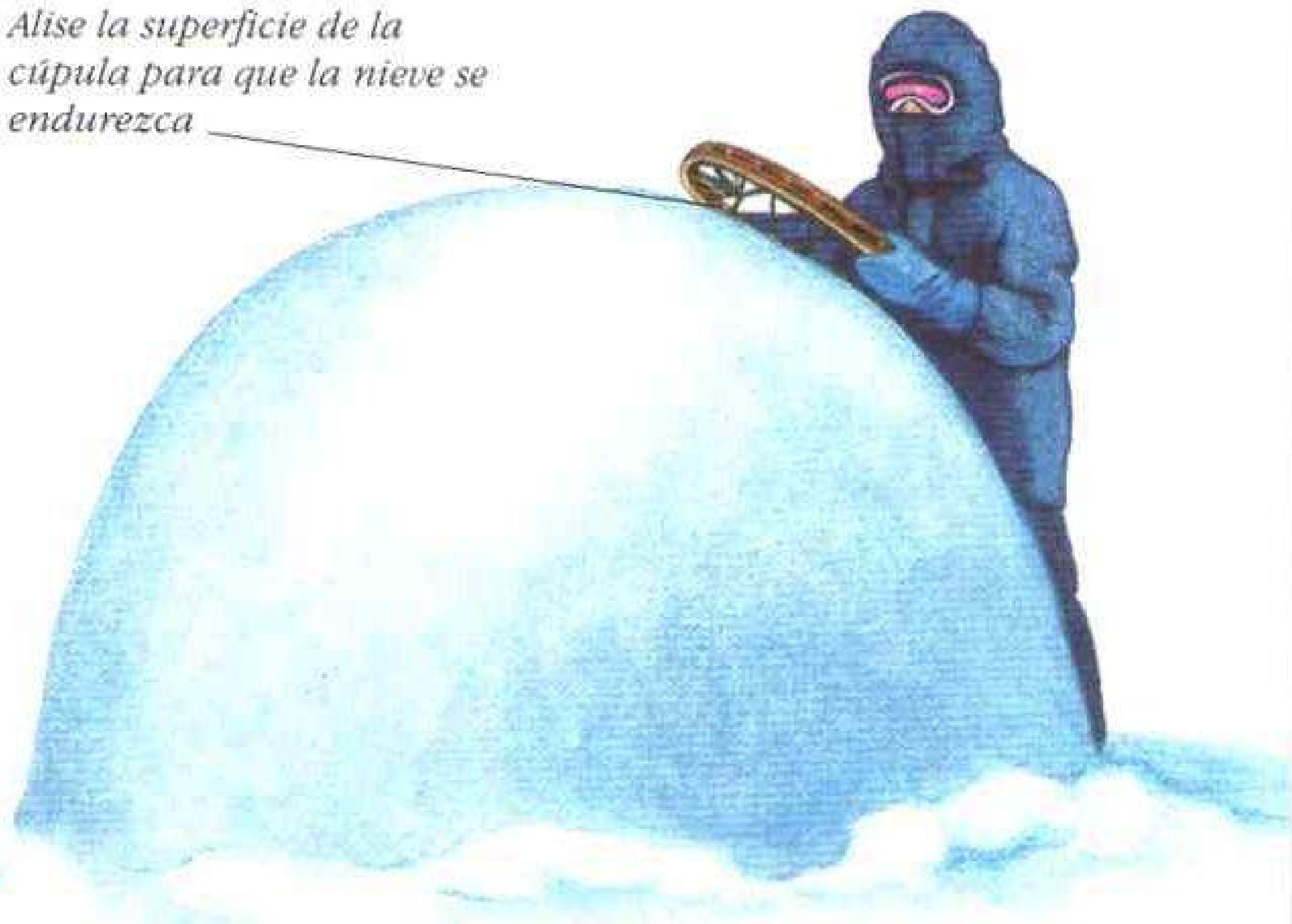

Alise la superficie de la cúpula para que la nieve se endurezca

3 Cuando la nieve en la pila tenga cerca de 1 m de grosor, alise la superficie y espere cerca de una hora para que se endurezca. Este período de espera es muy importante, ya que durante el mismo la nieve se recristaliza.

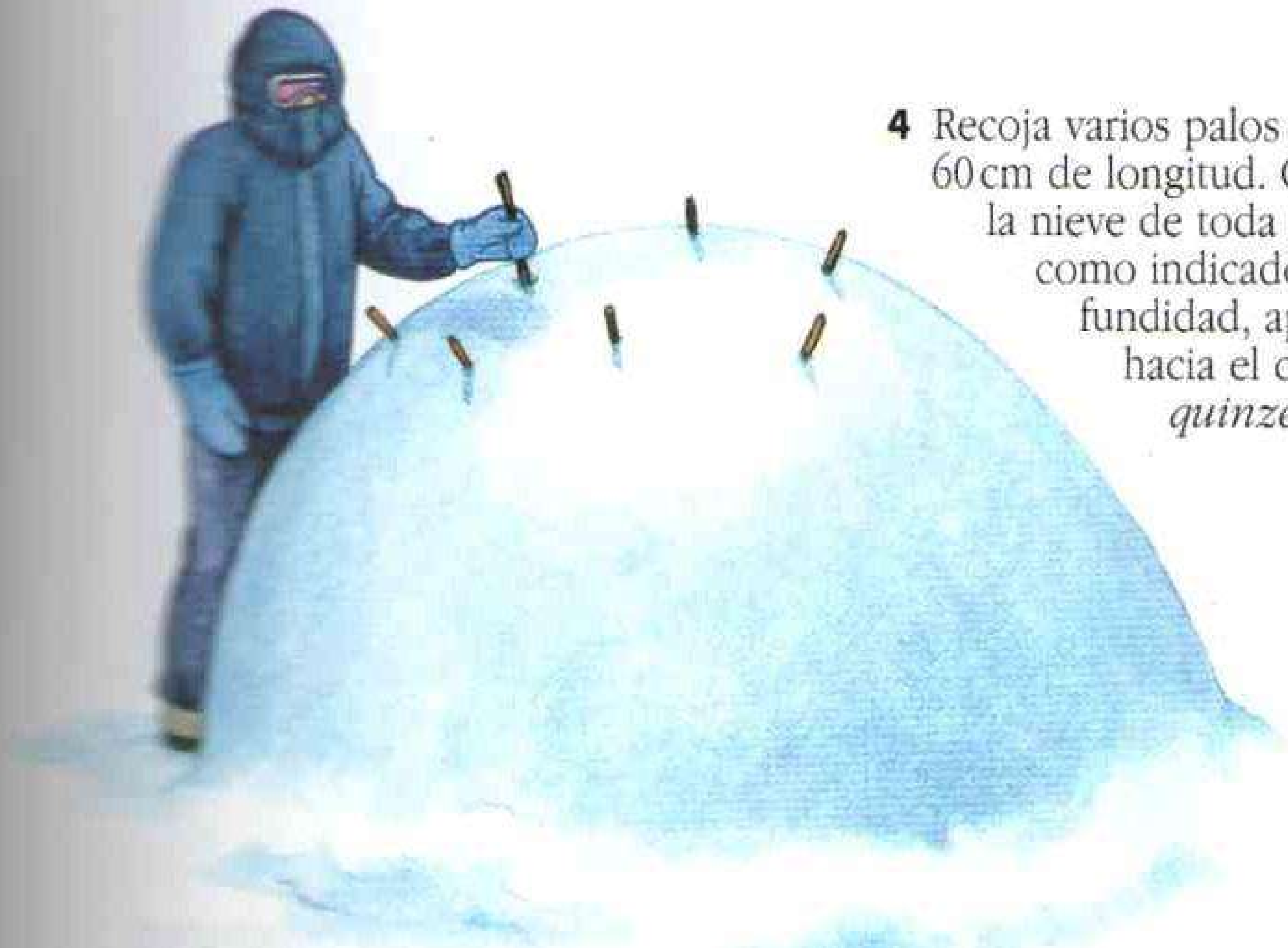

4 Recoja varios palos de unos 60 cm de longitud. Clávelos en la nieve de toda la superficie como indicadores de profundidad, apuntando hacia el centro del *quinze.*

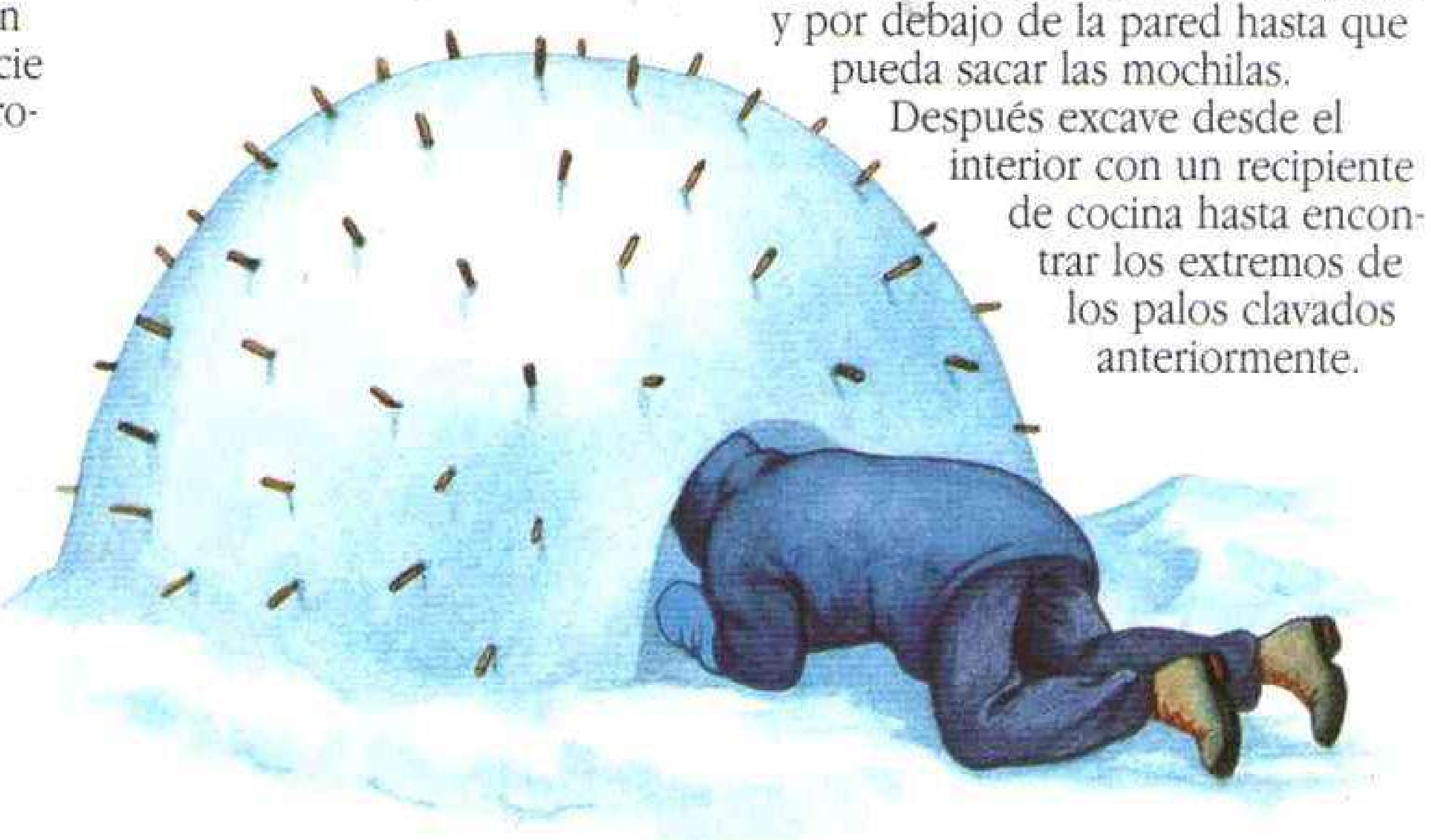

5 Cave desde el exterior del *quinze* y por debajo de la pared hasta que pueda sacar las mochilas. Después excave desde el interior con un recipiente de cocina hasta encontrar los extremos de los palos clavados anteriormente.

OTROS REFUGIOS EN LA NIEVE

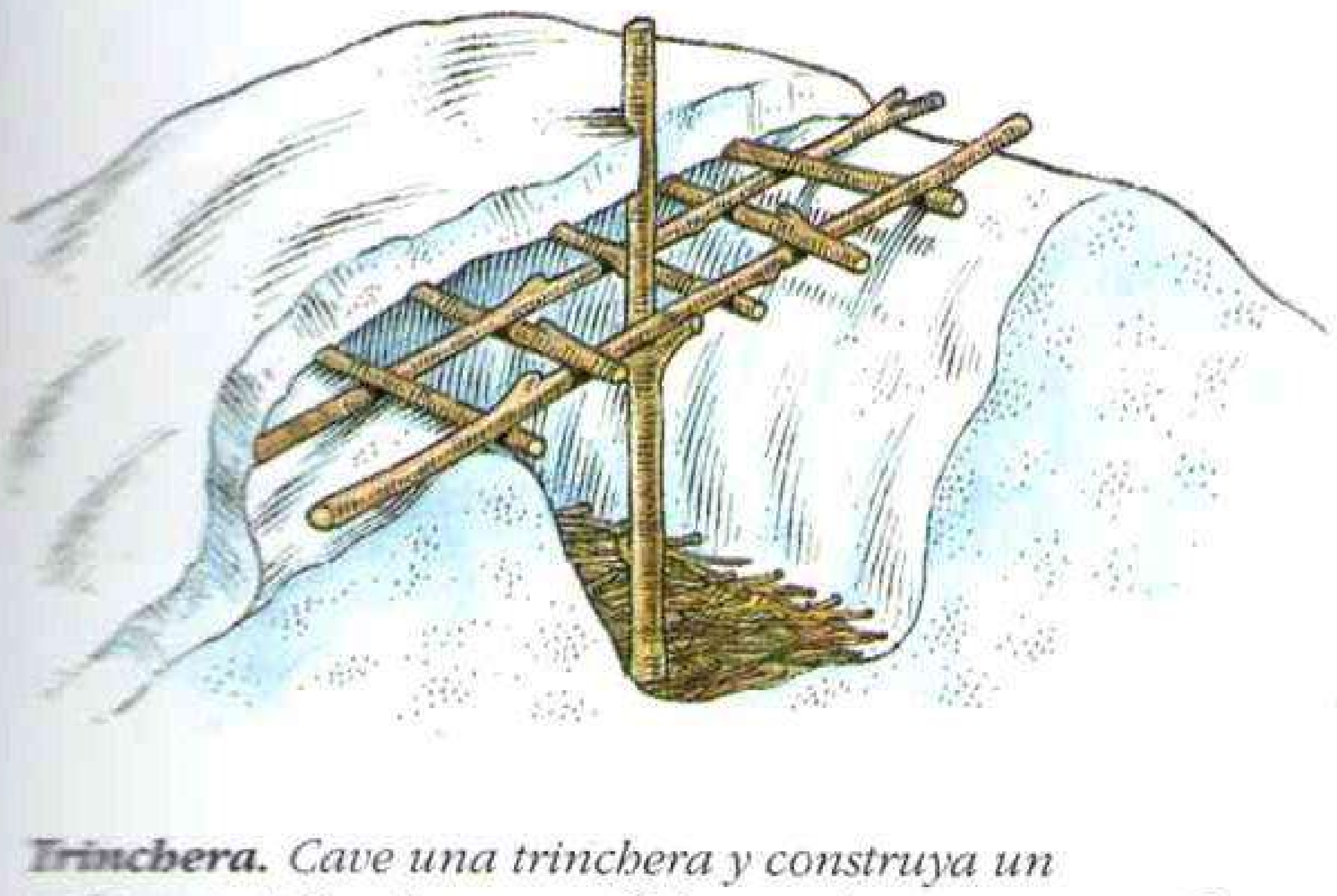

Trinchera. *Cave una trinchera y construya un techo entretejiendo ramas. Coloque nieve compacta sobre el armazón. No olvide colocar una rama vertical a través del techo para tener una ventilación adecuada en la trinchera.*

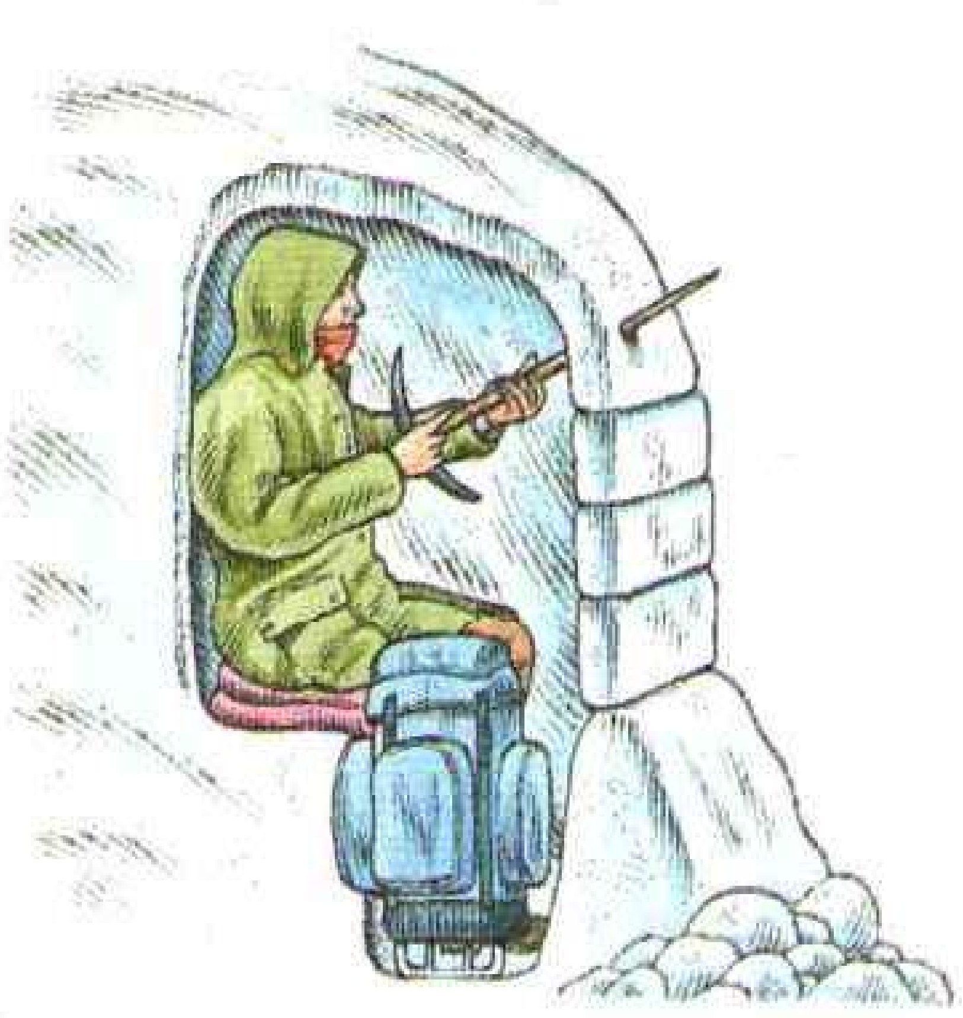

Cueva en la nieve. *En una emergencia, vacíe un banco de nieve y bloquee la entrada a su espalda con bloques de nieve. Haga un agujero en la pared para tener ventilación.*

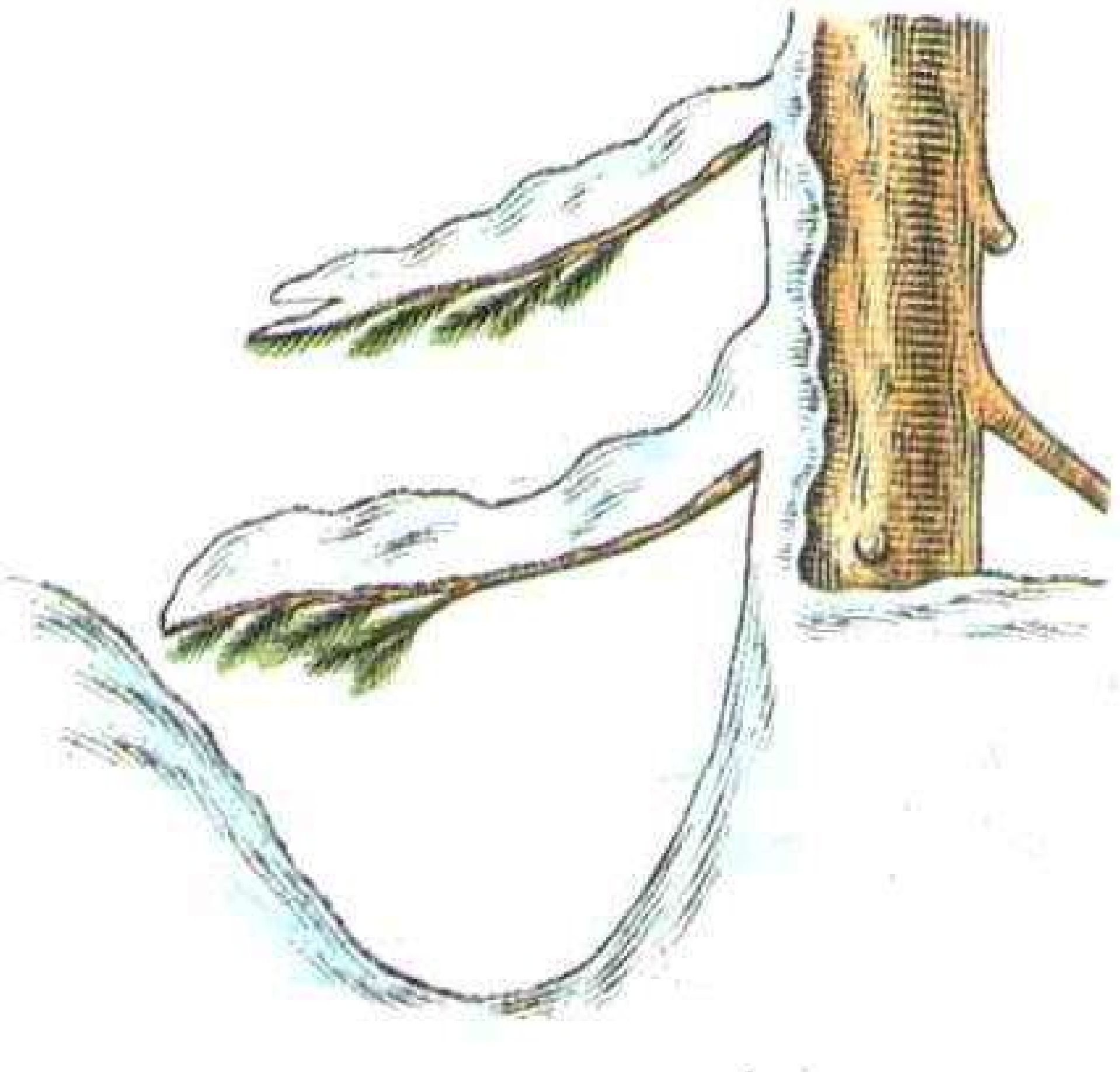

Resguardo natural. *Puede encontrar un refugio cómodo y casi listo para usar debajo de una conífera con ramas anchas y bajas. Tenga cuidado de no sacudir la nieve de las ramas.*

REFUGIOS TROPICALES

LA REGIÓN ENTRE el trópico de Cáncer y el trópico de Capricornio, unos 23 grados al norte y al sur del ecuador, se conoce como *los trópicos*. Esta región comprende selvas pluviales, pantanos y sabanas, que se caracterizan por su vegetación exuberante, las lluvias torrenciales y una abundante vida animal. Refugiarse de la lluvia, del viento y del sol resulta muy importante en todas las áreas tropicales. En las selvas pluviales montañosas, la noche puede ser bastante fría. En las selvas y pantanos, debe dormir a un nivel más elevado que el suelo, a ser posible lo suficientemente elevado para permitir el paso de animales pequeños por debajo, y para asegurarse de que la lluvia torrencial no le salpique al caer al suelo. Un techo impermeable y una mosquitera son también fundamentales. En las sabanas tropicales, dormir a un nivel más elevado que el suelo no es tan importante como en las selvas.

REFUGIO EN «A»

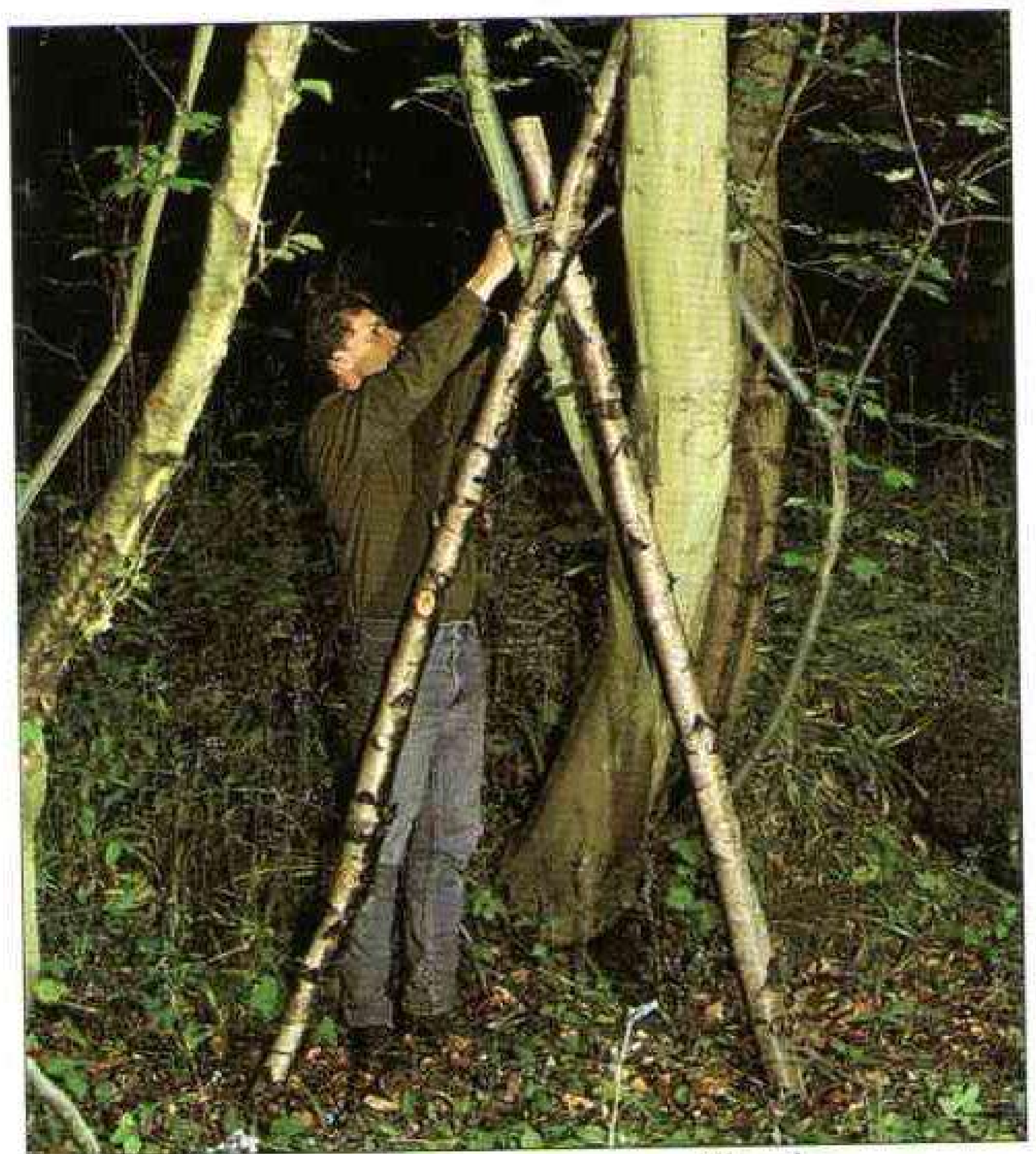

1 El refugio en «A» es muy sencillo de construir. Corte siete ramas largas y ate dos de ellas a un árbol.

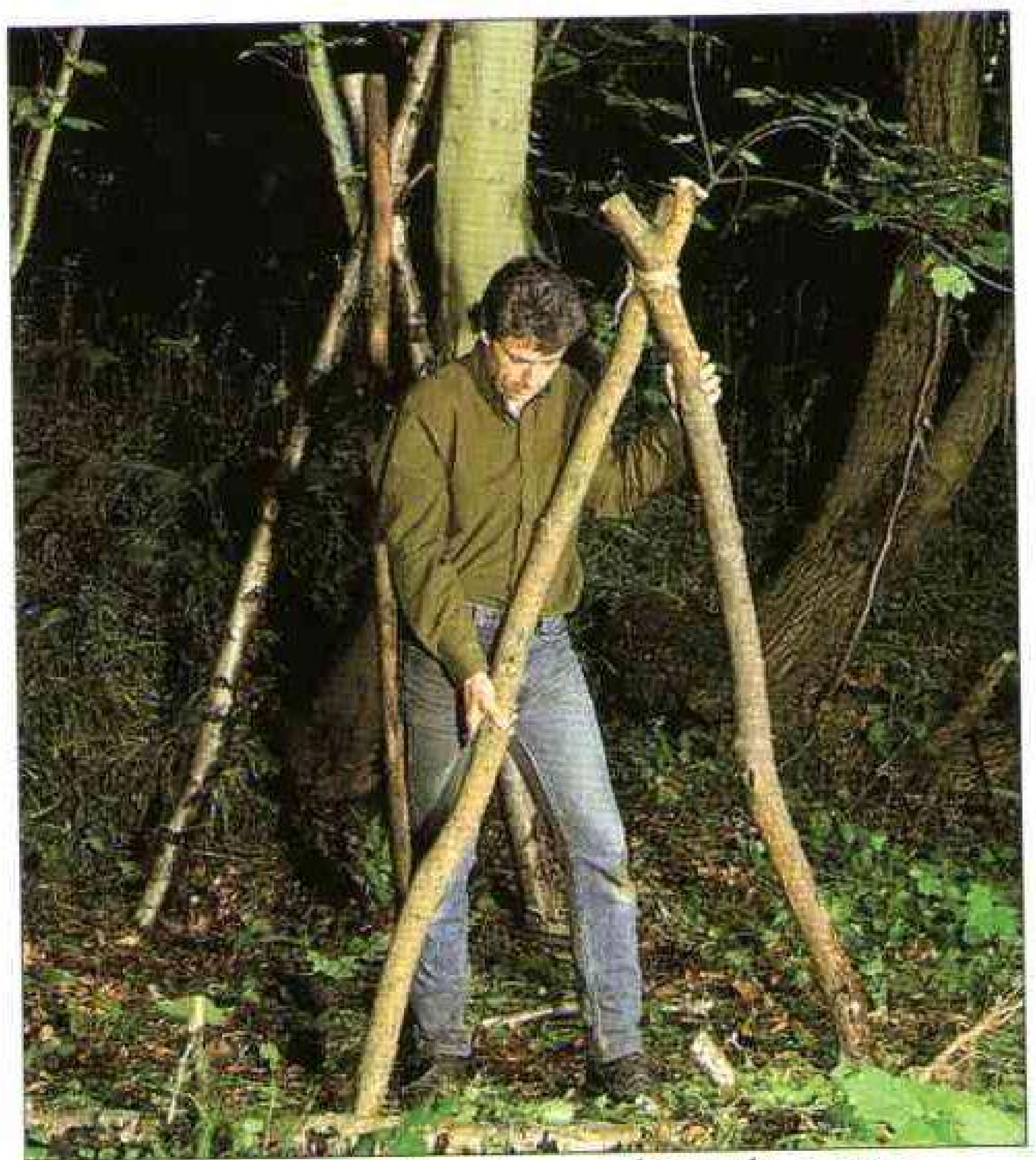

2 Haga una segunda «A» y colóquela a una distancia igual a la de su propia estatura más 60 cm, de la primera de ellas.

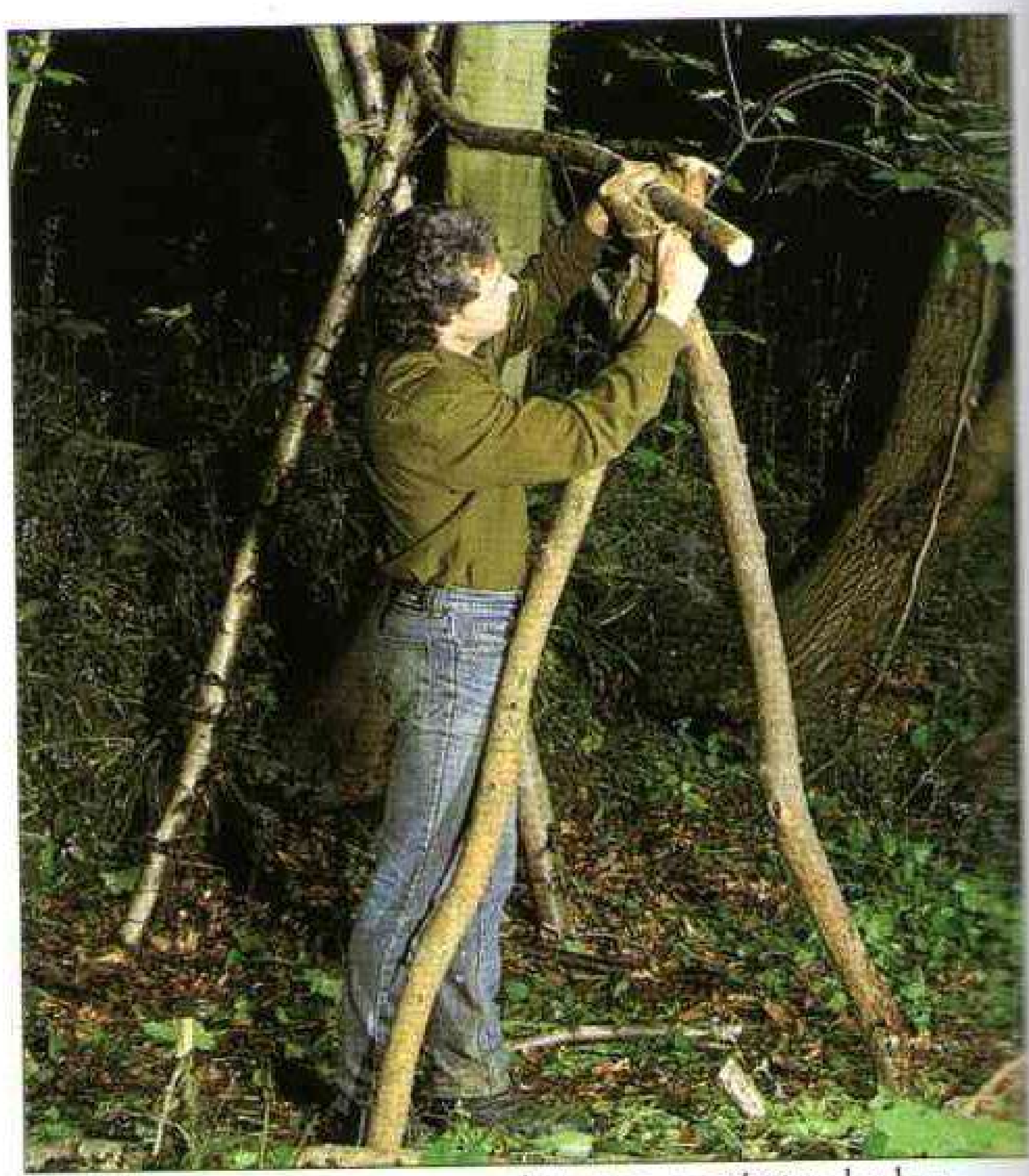

3 Coloque una rama resistente encima de las dos «V» superiores de cada estructura en «A». Esta rama actuará como travesaño.

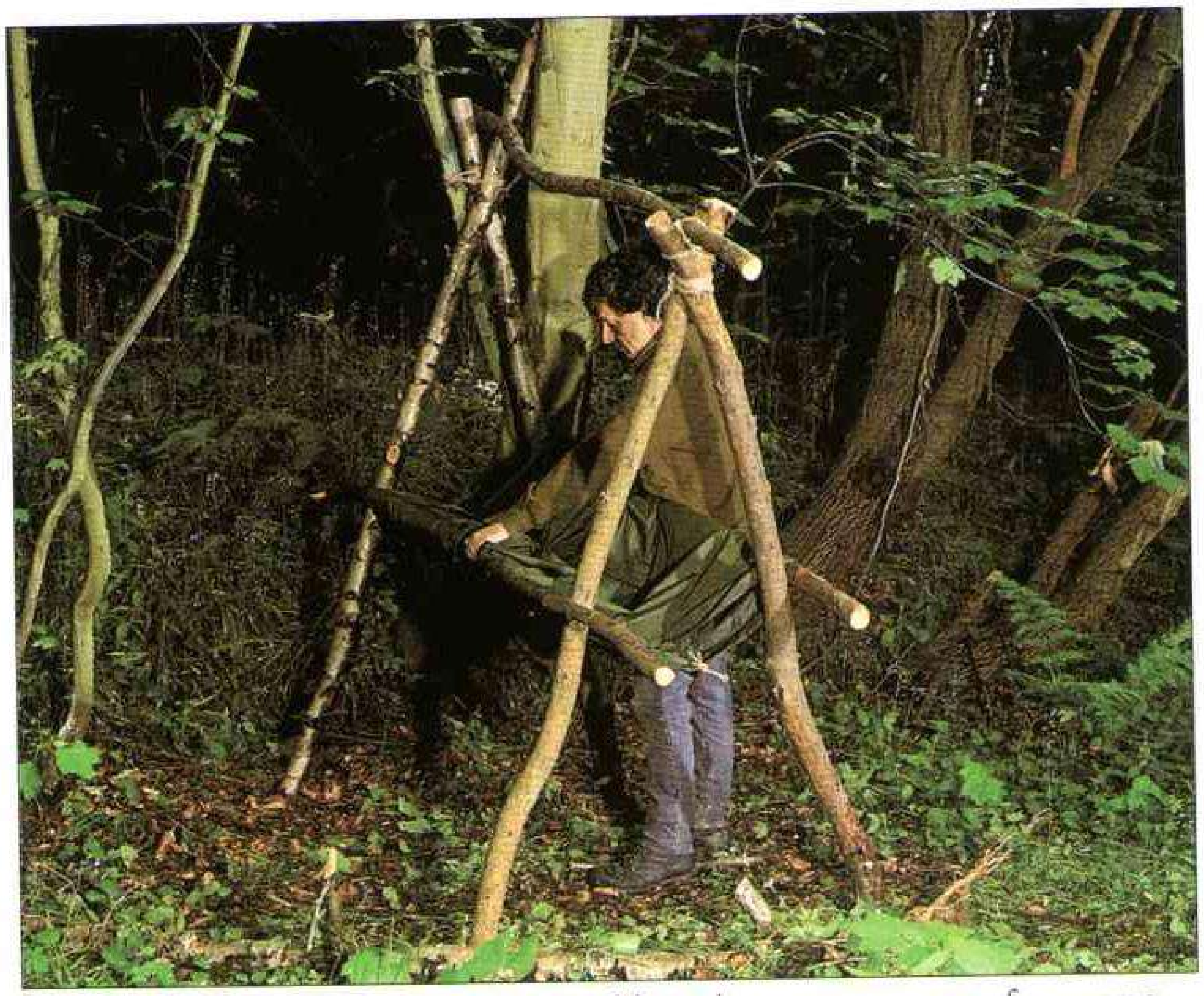

4 Ate los lados de una tela impermeable o de una manta para formar un tubo. Inserte dos ramas largas y sepárelas hasta formar una «camilla». Encaje la camilla entre ambas estructuras en «A» con las ramas en el exterior.

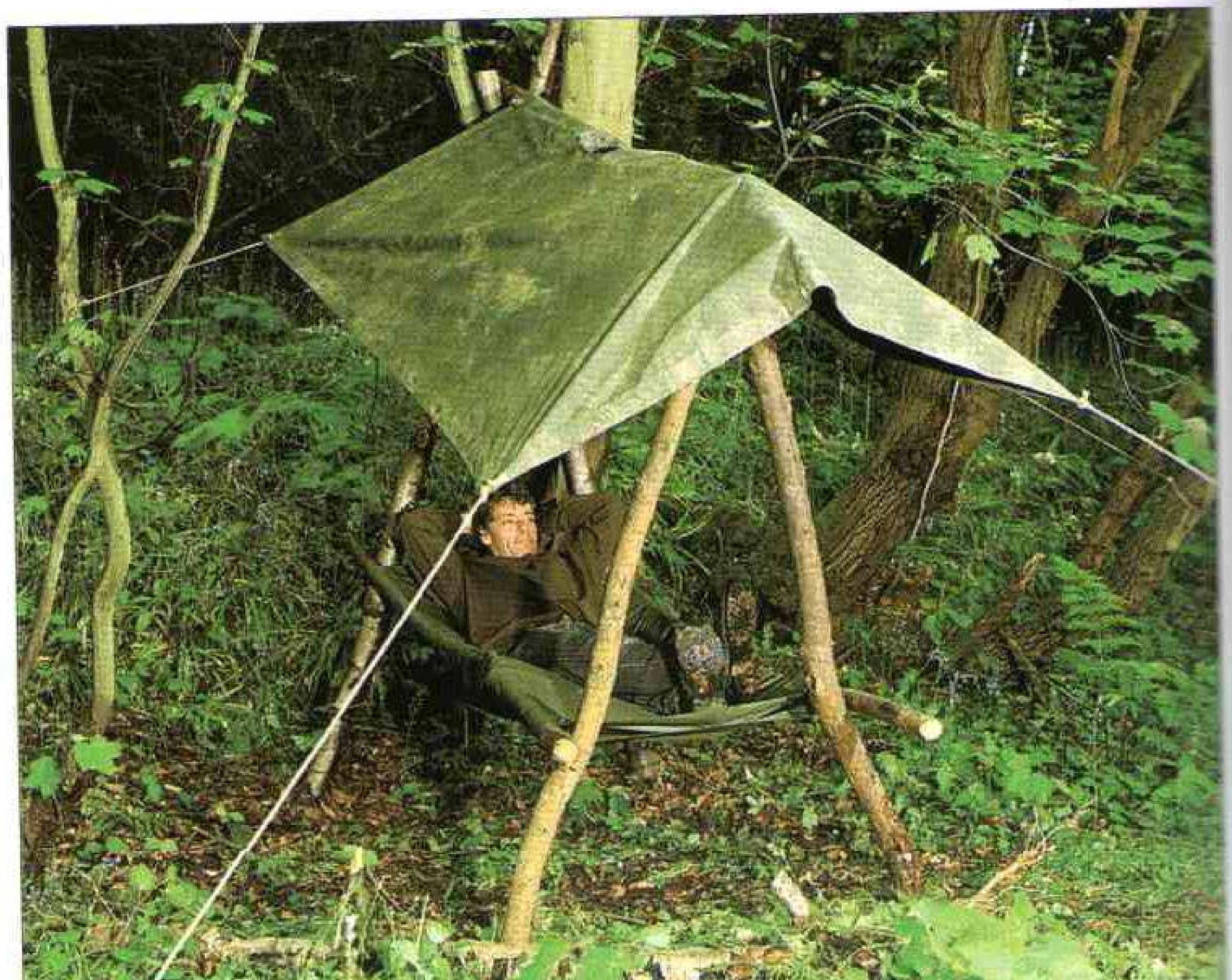

5 Extienda una lona impermeable sobre el travesaño para construir el tejado. Ténselo a ambos lados y átelo a los árboles. El tejado debería protegerle de la lluvia y, al mismo tiempo, permitir la circulación del aire alrededor de su cabeza.

REFUGIO DE BAMBÚ

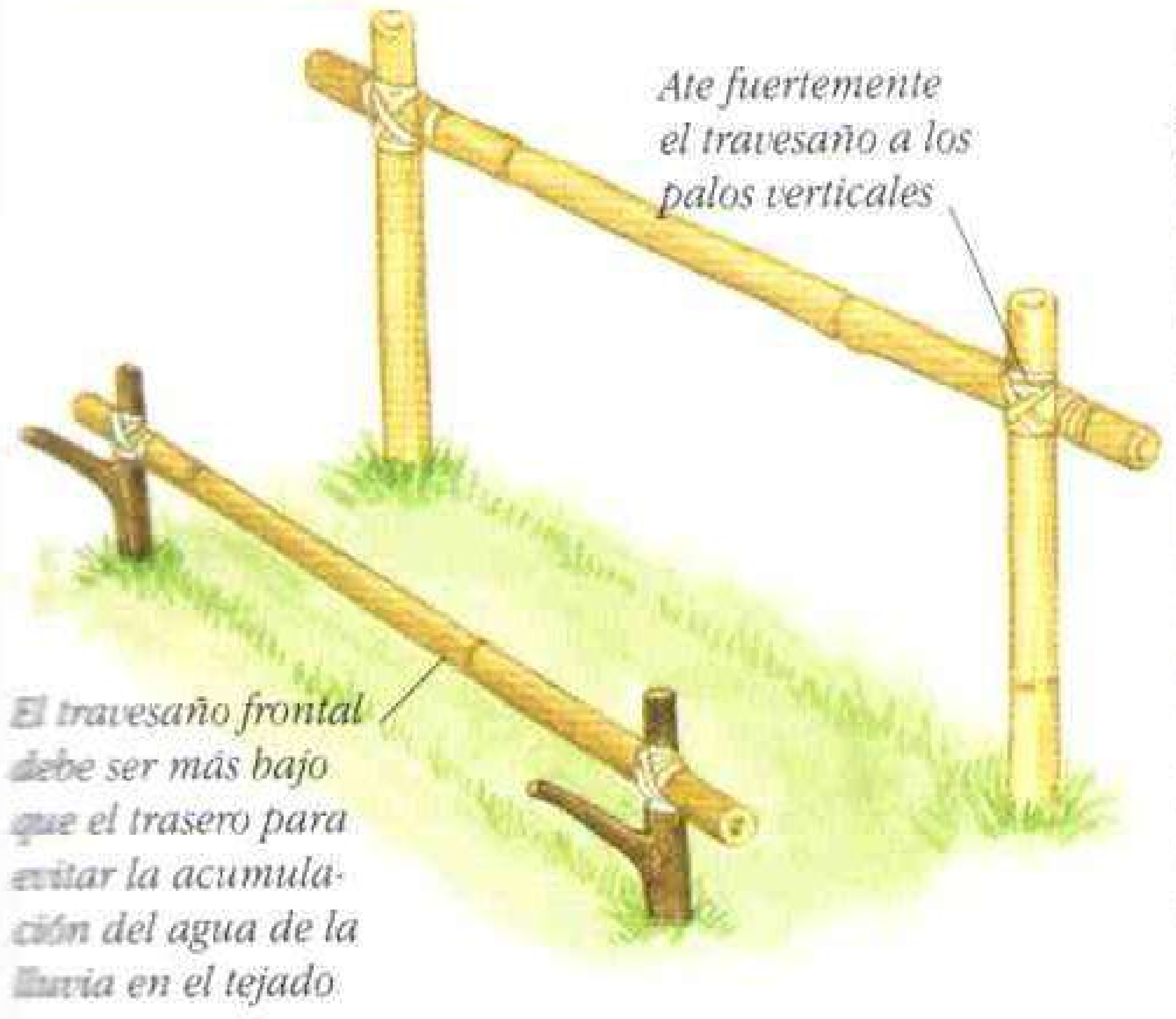

Ate fuertemente el travesaño a los palos verticales

El travesaño frontal debe ser más bajo que el trasero para evitar la acumulación del agua de la lluvia en el tejado

1 Este refugio puede construirse dondequiera que crezca el bambú. Coloque cuatro postes verticales gruesos en el suelo separados por una distancia igual a la de su estatura más unos 60 cm. Los postes frontales deben ser unos 30 cm más cortos que los traseros. Coloque los travesaños entre los postes verticales como se ilustra.

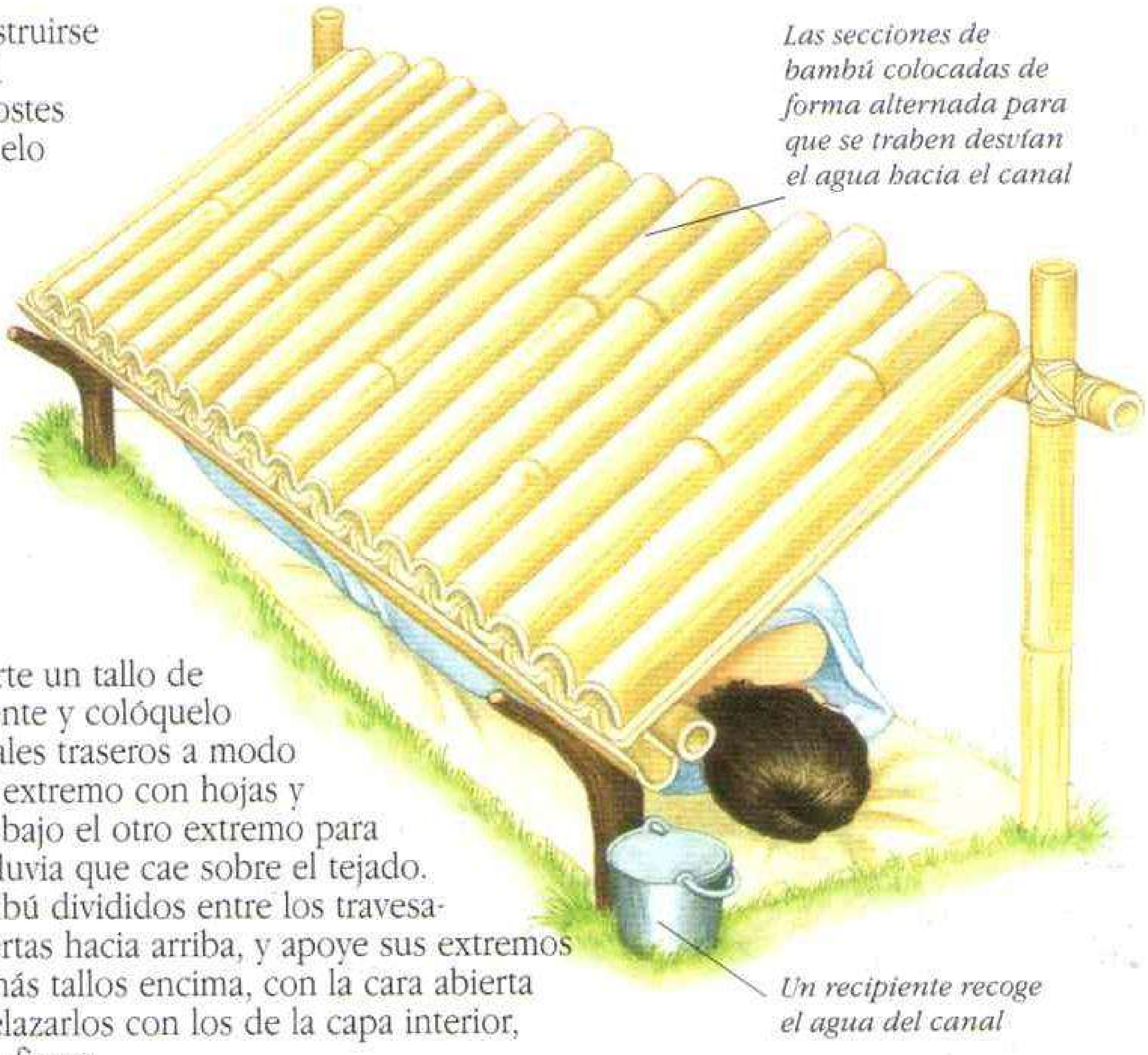

Las secciones de bambú colocadas de forma alternada para que se traben desvían el agua hacia el canal

Un recipiente recoge el agua del canal

2 Con un cuchillo, corte un tallo de bambú longitudinalmente y colóquelo entre los postes verticales traseros a modo de canal. Obstruya un extremo con hojas y coloque un recipiente bajo el otro extremo para recoger el agua de la lluvia que cae sobre el tejado. Coloque tallos de bambú divididos entre los travesaños, con las caras abiertas hacia arriba, y apoye sus extremos en el canal. Coloque más tallos encima, con la cara abierta hacia abajo, para entrelazarlos con los de la capa interior, como se muestra en la figura.

PRECAUCIÓN

Los tallos de bambú pueden partirse fácilmente por el centro con un cuchillo, pero tenga cuidado al cortarlos para evitar la proyección de astillas, afiladas como cuchillas. Tenga cuidado al cortar el bambú, ya que algunos tallos pueden torcerse bajo presión, soltarse y cortarle.

REFUGIO EN LA SELVA

1 Para construir el armazón del refugio, clave cuatro troncos sólidos en el suelo, que harán de postes en las cuatro esquinas, separados por una distancia superior a la de su propia estatura. Coloque cuatro ramas horizontales, atándolas a la parte de los troncos verticales, a los que habrá hecho unas muescas para encajarlos mejor, y constituir así el armazón de la cama *(véase página 48)*.

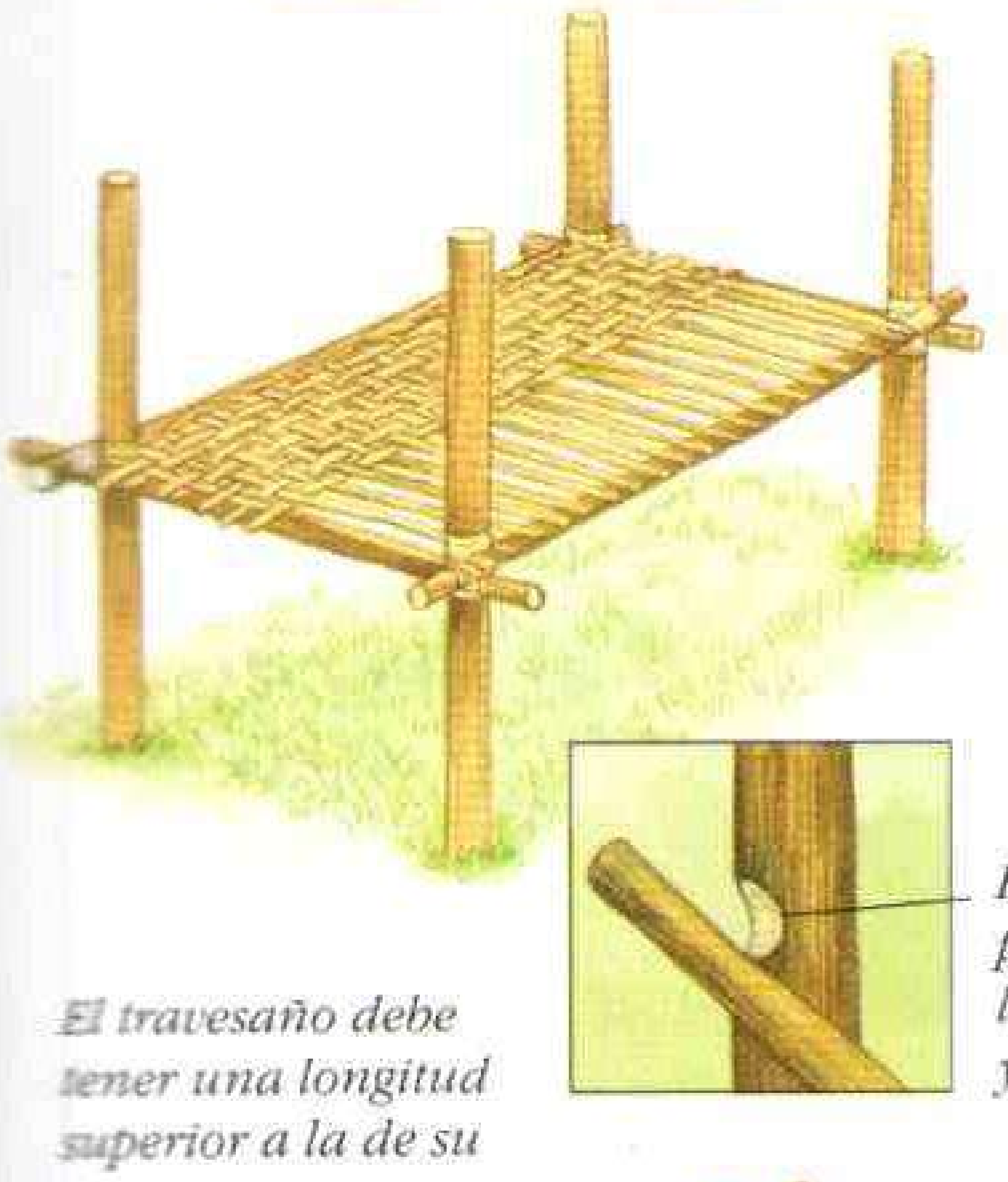

Haga muescas en los postes verticales para lograr un mejor ajuste y fijación

El travesaño debe tener una longitud superior a la de su estatura

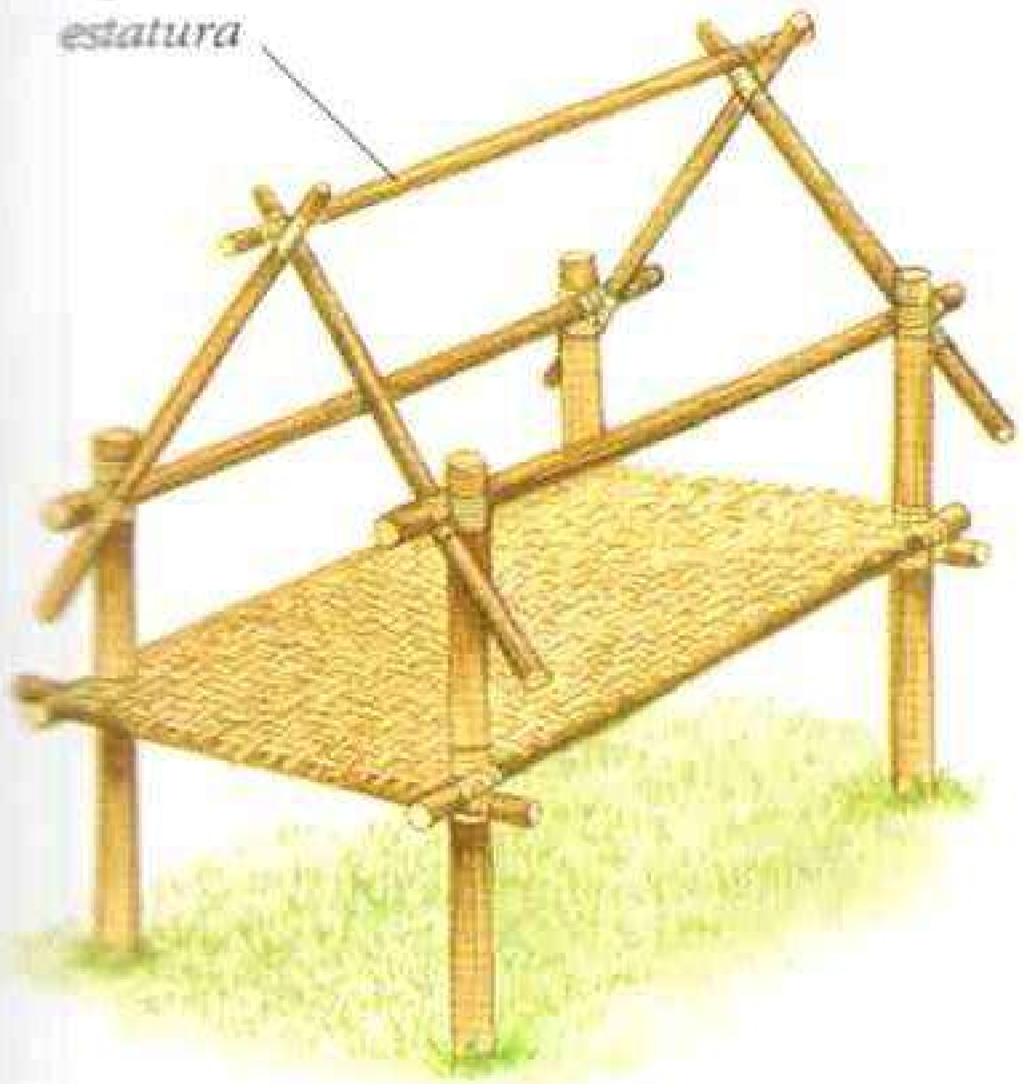

2 Ate siete ramas entre sí para construir la estructura del techo. El techo debe ser inclinado para no acumular el agua de la lluvia, y lo suficientemente resistente para soportar el peso de la cubierta. Átelo a los troncos verticales de las esquinas.

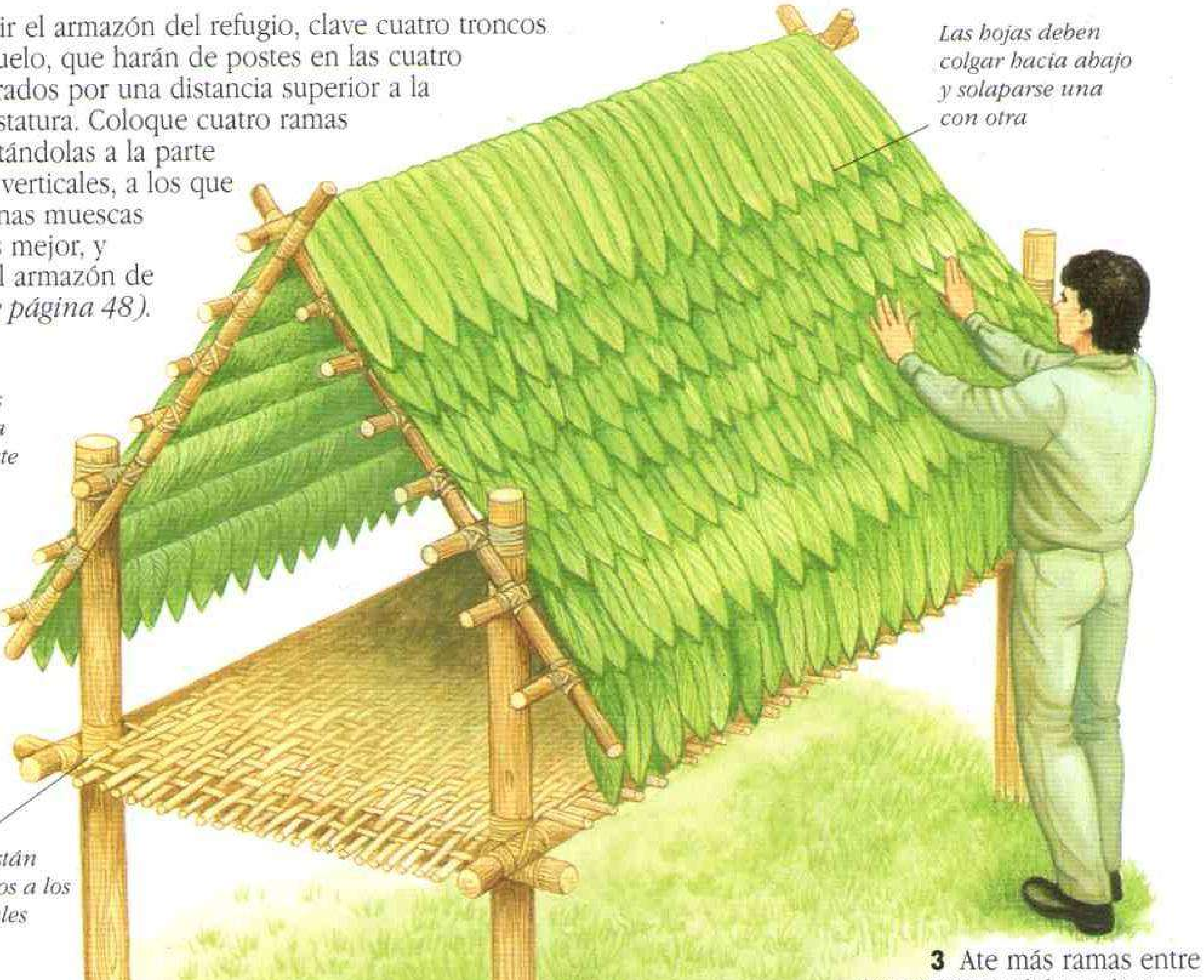

Las hojas deben colgar hacia abajo y solaparse una con otra

Los soportes horizontales están bien asegurados a los troncos verticales

3 Ate más ramas entre los soportes del tejado y cúbralas con hojas grandes. Sobreponga las hojas contiguas y las de una fila con otra. Empezando por abajo, el pedúnculo de cada hoja debe apuntar hacia la parte superior, con el haz (cara brillante) hacia arriba para desviar el agua de la lluvia.

REFUGIOS PERMANENTES

EN UNA SITUACIÓN de supervivencia, desconoce el tiempo en el que permanecerá abandonado, o cuánto tiempo permanecerá en un mismo sitio. Por lo tanto, debe intentar construir el mejor refugio posible. Si se ve forzado a pasar todo un invierno, por ejemplo, necesitará un refugio más sólido, como una cabaña de troncos. Antes de dedicarse durante varios días, o incluso semanas, a construir un refugio permanente, debe asegurarse de que dispone del sitio más adecuado. Sus alimentos, combustibles y agua fresca deben encontrarse en las cercanías; el suelo, sólido, debe estar bien drenado; y el terreno protegido de cualquier peligro natural *(véase página 40)*.

CABAÑA DE TRONCOS

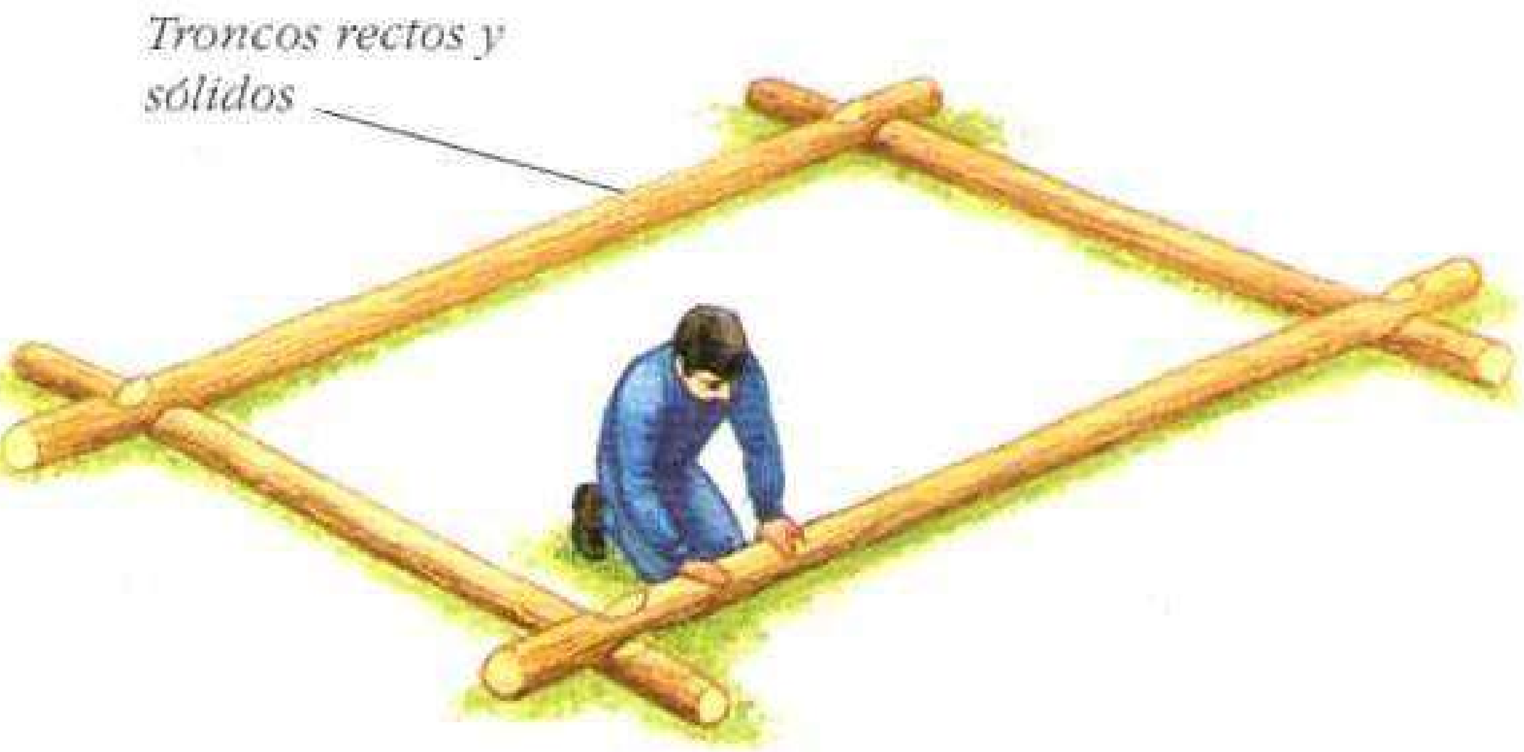
Troncos rectos y sólidos

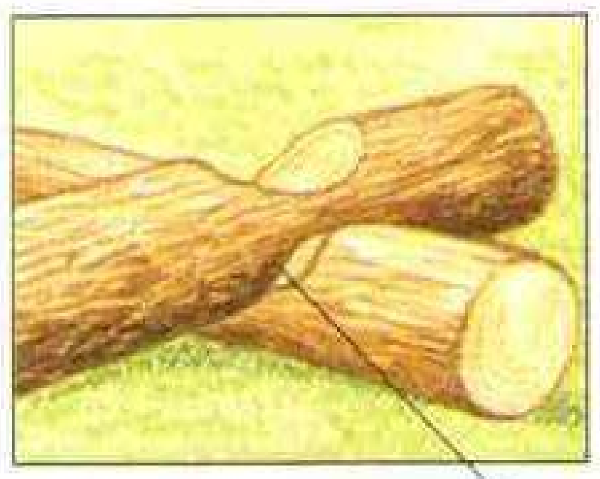
Para asegurar un buen encaje, debe hacer muescas a los troncos, en las puntas y a ambos lados

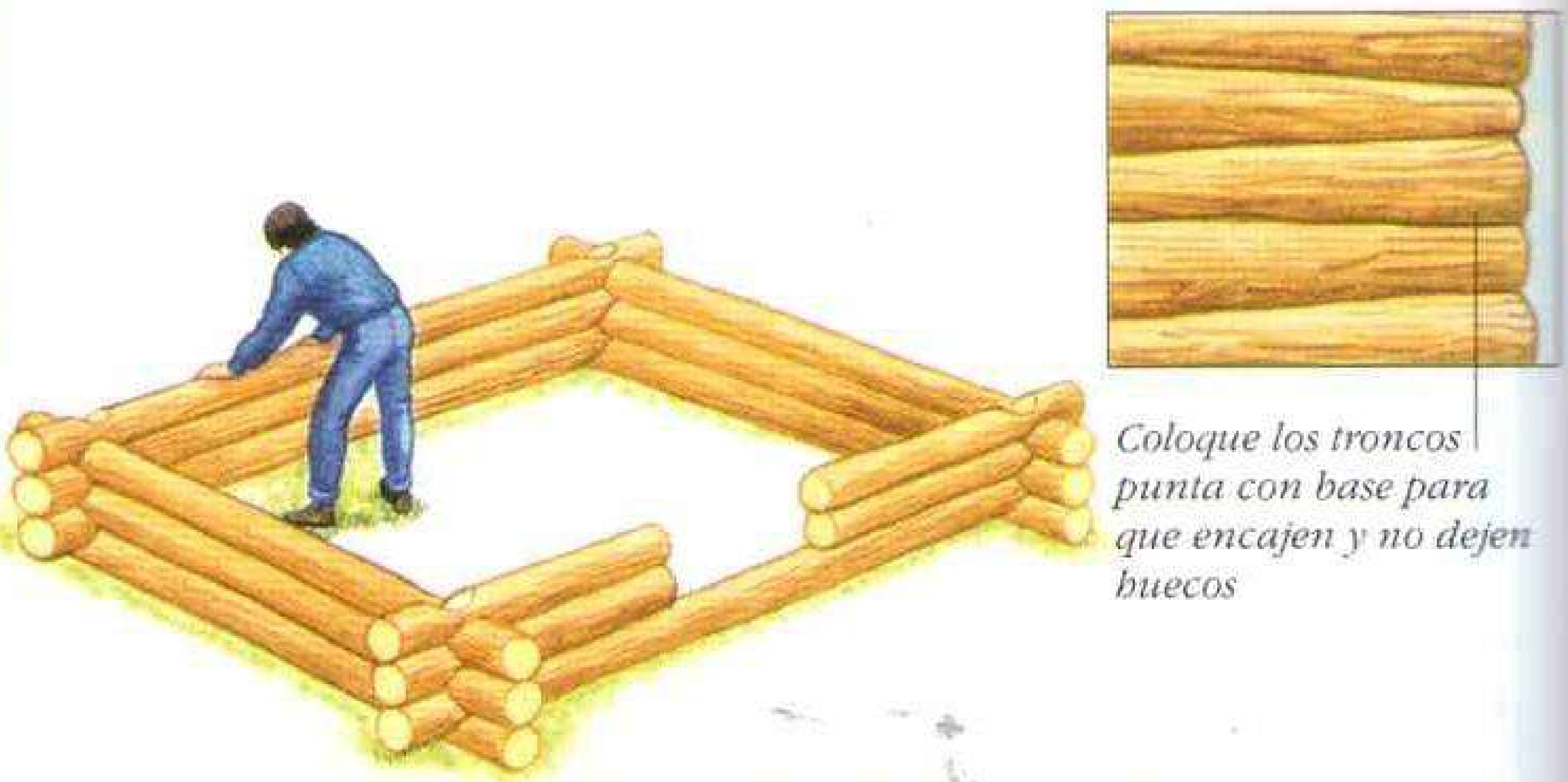
Coloque los troncos punta con base para que encajen y no dejen huecos

1 Limpie el suelo por lo menos hasta 1 m más allá de las paredes de la cabaña, y coloque los primeros troncos formando el contorno de la cabaña.

2 La clave de una buena construcción está en emplear troncos semejantes y regulares, y en hacer las muescas en los sitios correspondientes. Asegúrese de que con las muescas cada tronco queda bien colocado encima de los restantes, para que las paredes queden lo más perpendiculares posible, en ángulo recto con las esquinas. Construya las paredes gradualmente.

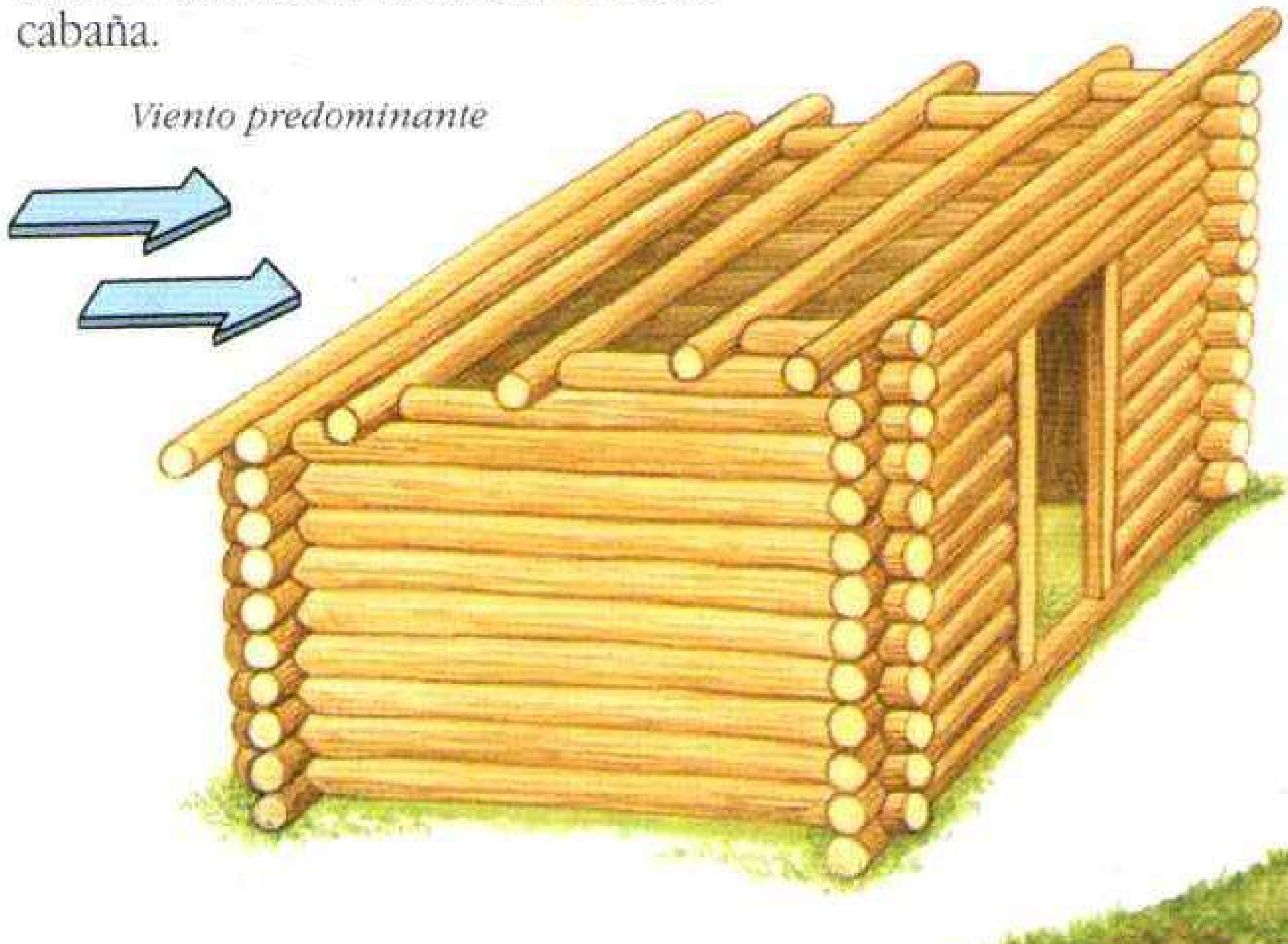
Viento predominante

3 Las paredes traseras y la pendiente del tejado deben estar encaradas al viento y a las inclemencias del tiempo, mientras que la parte frontal con la entrada y cualquier ventana deben encararse hacia el sol. Incline el tejado de tal manera que la parte frontal sea más alta que la trasera, para que no se acumulen el agua de la lluvia o la nieve.

UNA CHIMENEA DE PIEDRA

Debido al peligro de incendio, el hogar y la chimenea deben ser independientes y deben construirse de forma sólida, preferiblemente de piedra o ladrillo. La chimenea puede construirse fuera, con un agujero en la pared a manera de hogar, o bien en el interior, haciendo que la chimenea sobresalga del techo. Para construir la estructura, utilice rocas o ladrillos de barro.

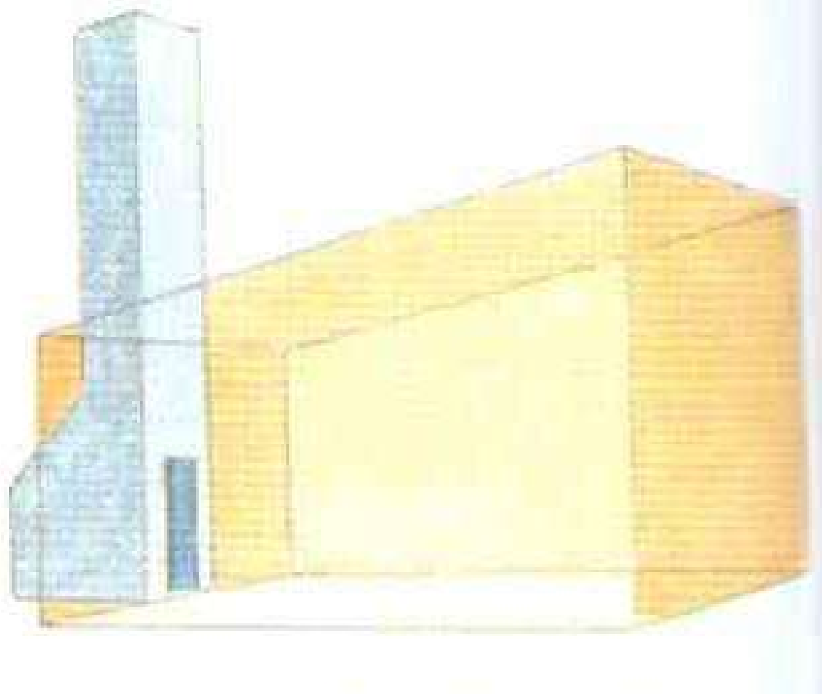

Chimenea de piedra. *Construya una chimenea de piedra fuera de las paredes de la cabaña para que el fuego no queme su refugio.*

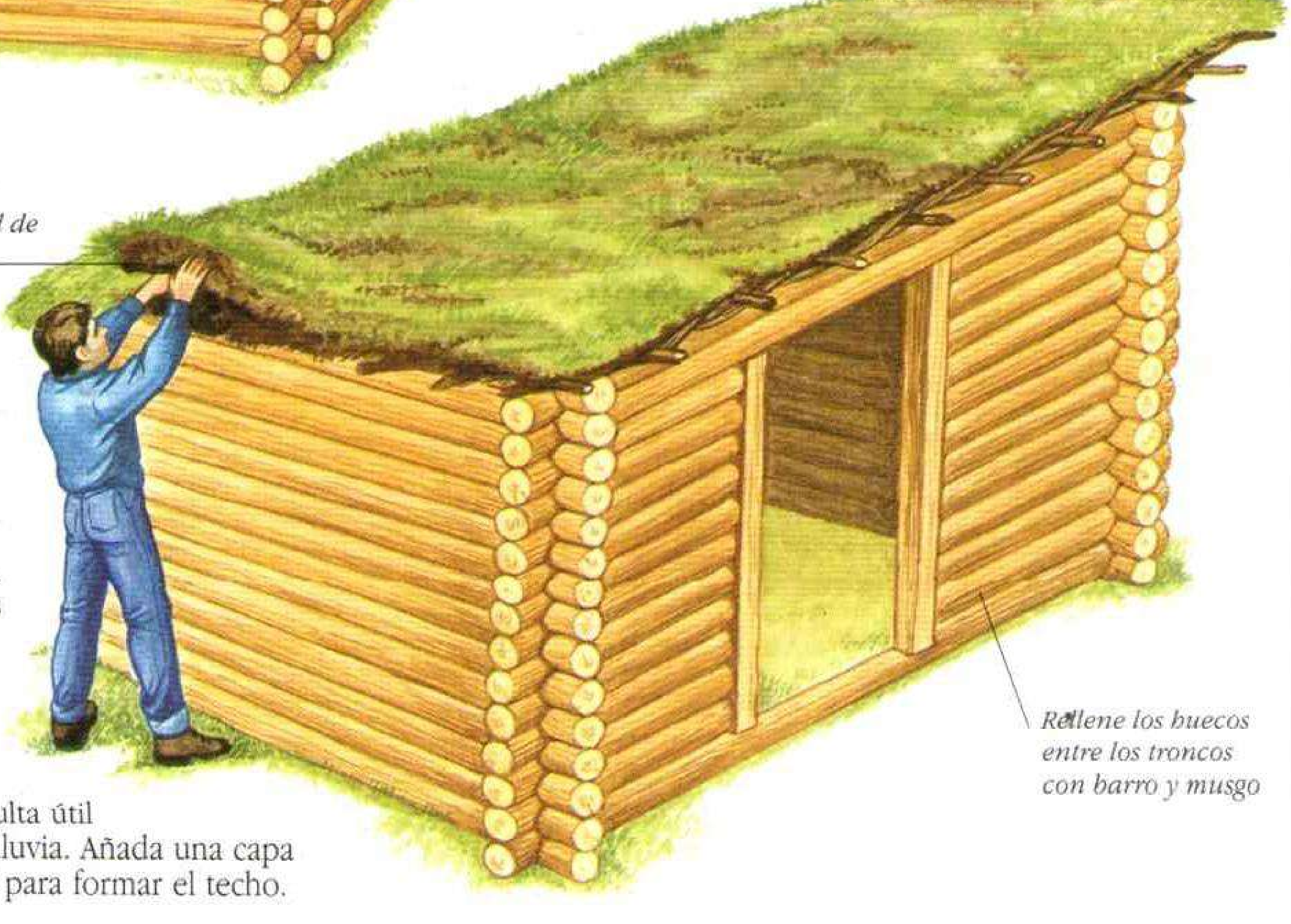
Corte gruesos cuadros de césped con gran cantidad de tierra
Rellene los huecos entre los troncos con barro y musgo

4 La cubierta del tejado puede ser de cualquier material. Coloque un aislante como helechos, hierbas, ramas de abeto, sobre un entramado de troncos. Si dispone de ellas, las bolsas de papel que contienen fertilizante son un aislante particularmente bueno; cualquier lámina de plástico resulta útil para protegerse de la lluvia. Añada una capa de bloques de césped para formar el techo.

CASA DE BLOQUES (TEPES)*

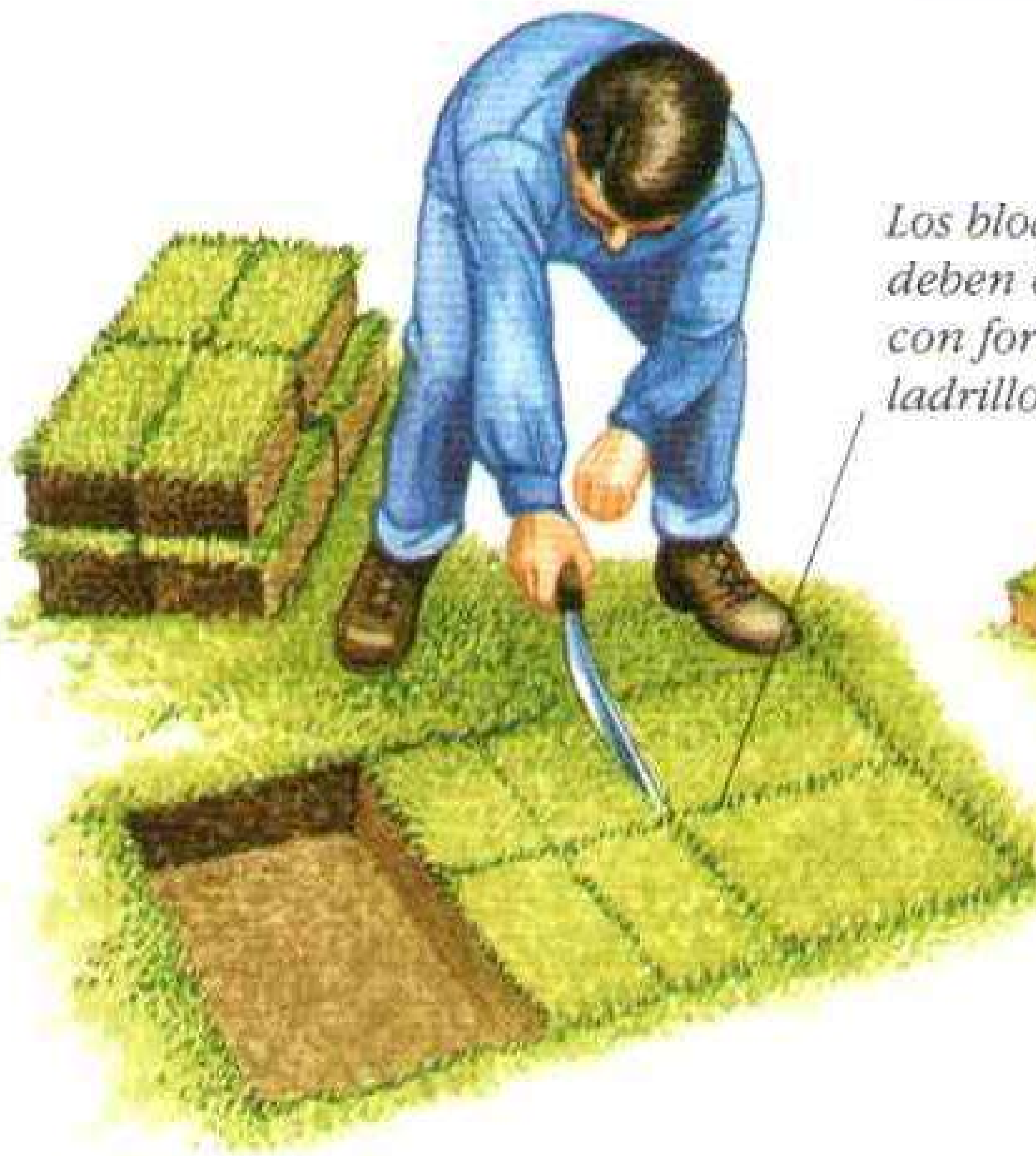

Los bloques deben cortarse con forma de ladrillo

1 Los bloques de césped pueden utilizarse para construir como si se tratase de ladrillos sin secar, cortados y colocados sin mortero. Comience cortando una selección de bloques del suelo.

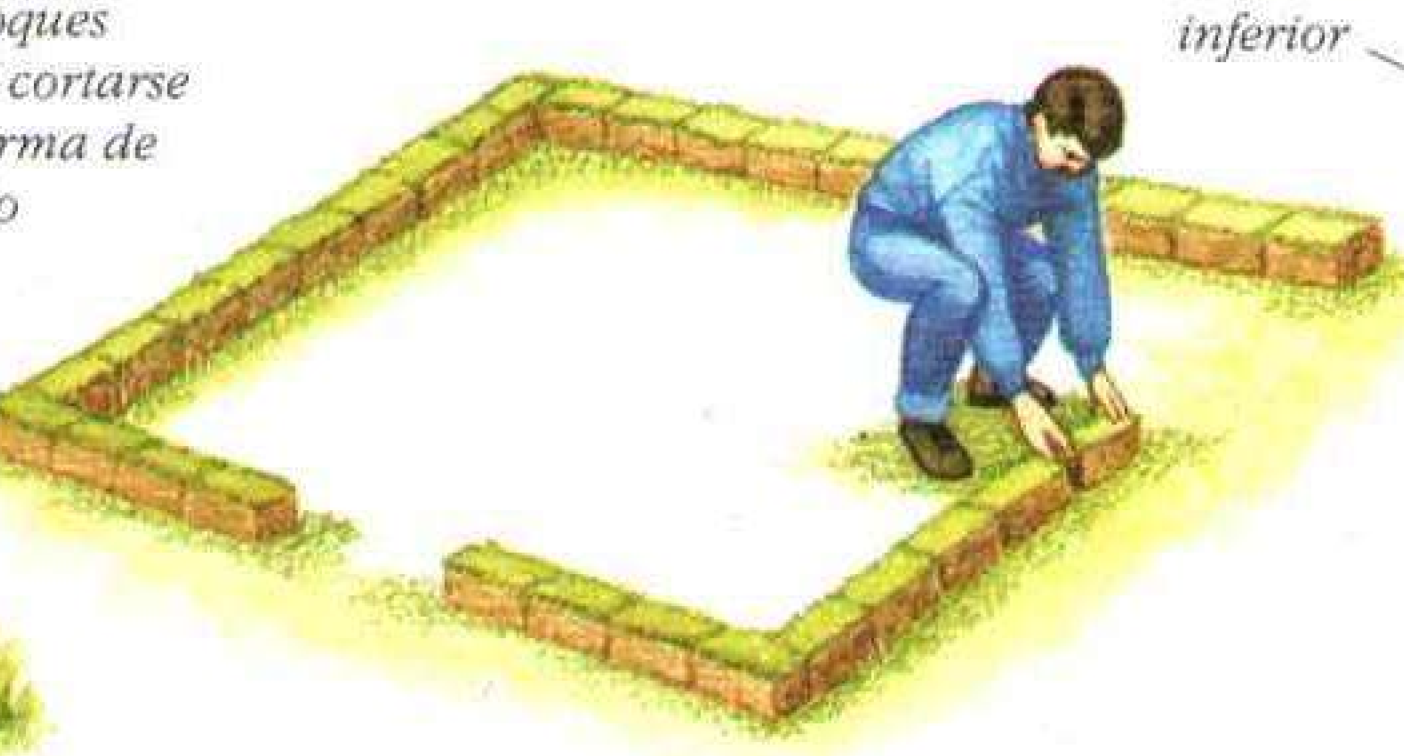

2 Coloque el contorno de la base de la cabaña sobre una porción limpia de terreno, dejando un espacio para la puerta. El suelo debe ser llano y sólido, y capaz de soportar el considerable peso de las paredes y del techo sin ceder. Comience a levantar las paredes.

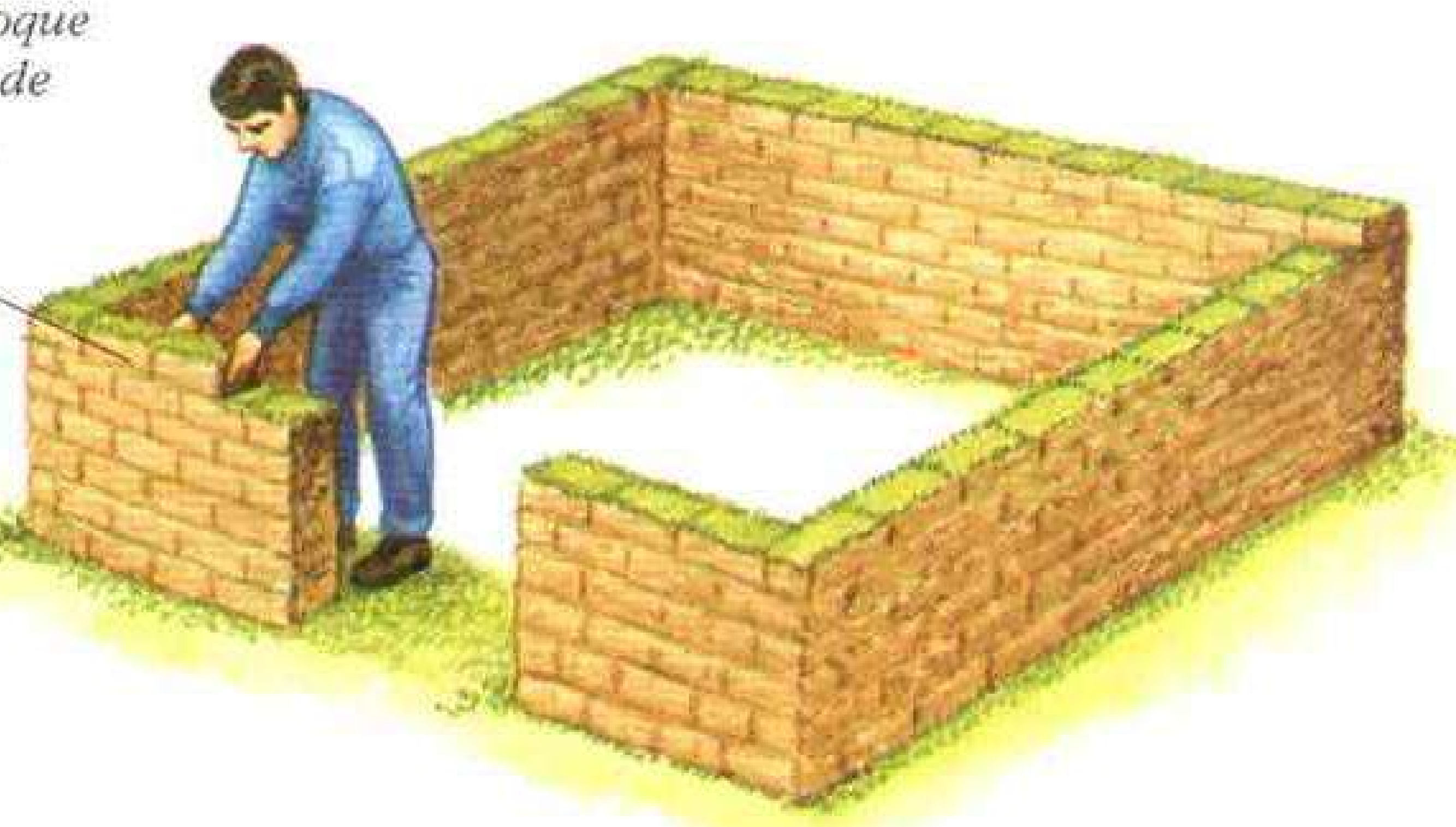

Coloque cada bloque sobre la juntura de los dos de la fila inferior

3 Construya gradualmente las paredes de la cabaña, solapando los bloques en las esquinas. Asegúrese de que cada nueva fila solapa la anterior para que las paredes sean firmes y no se derrumben. Intente hacer paredes rectas, a 90 grados una de otra, y perpendiculares.

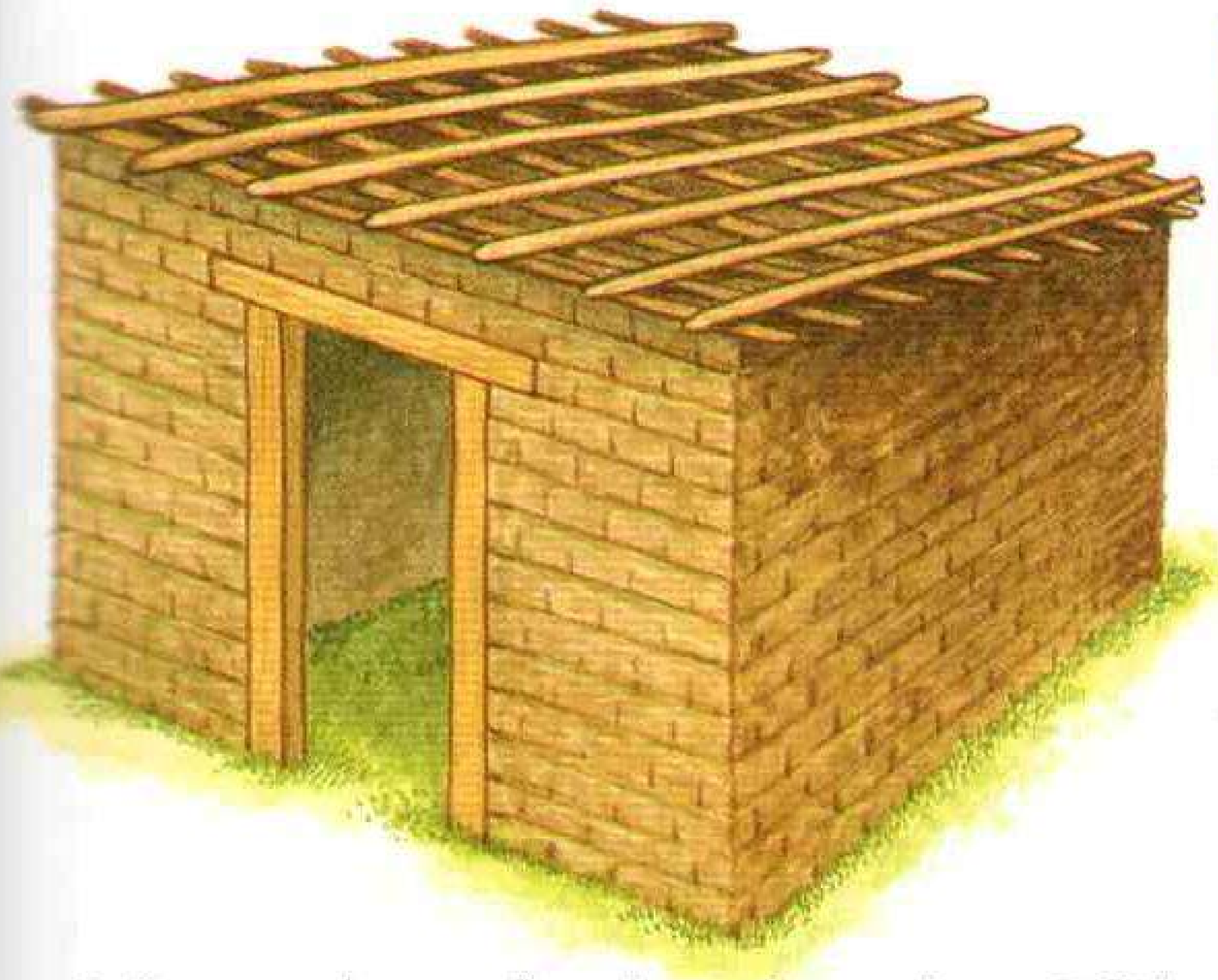

4 Construya la parte frontal más alta que la trasera, inclinando el tejado hacia donde sople el viento para evitar que se acumule el agua de la lluvia. Si es posible, construya una estructura de ramas para el techo.

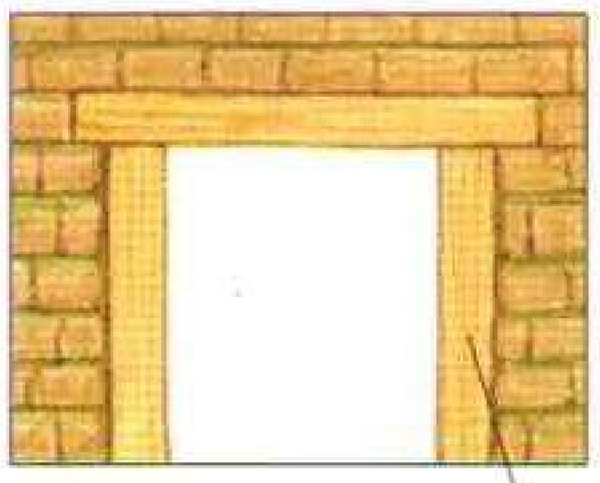

Añada un marco para la puerta hecho de madera, cuyo dintel debe ser lo suficientemente resistente como para soportar el peso de varias filas de bloques

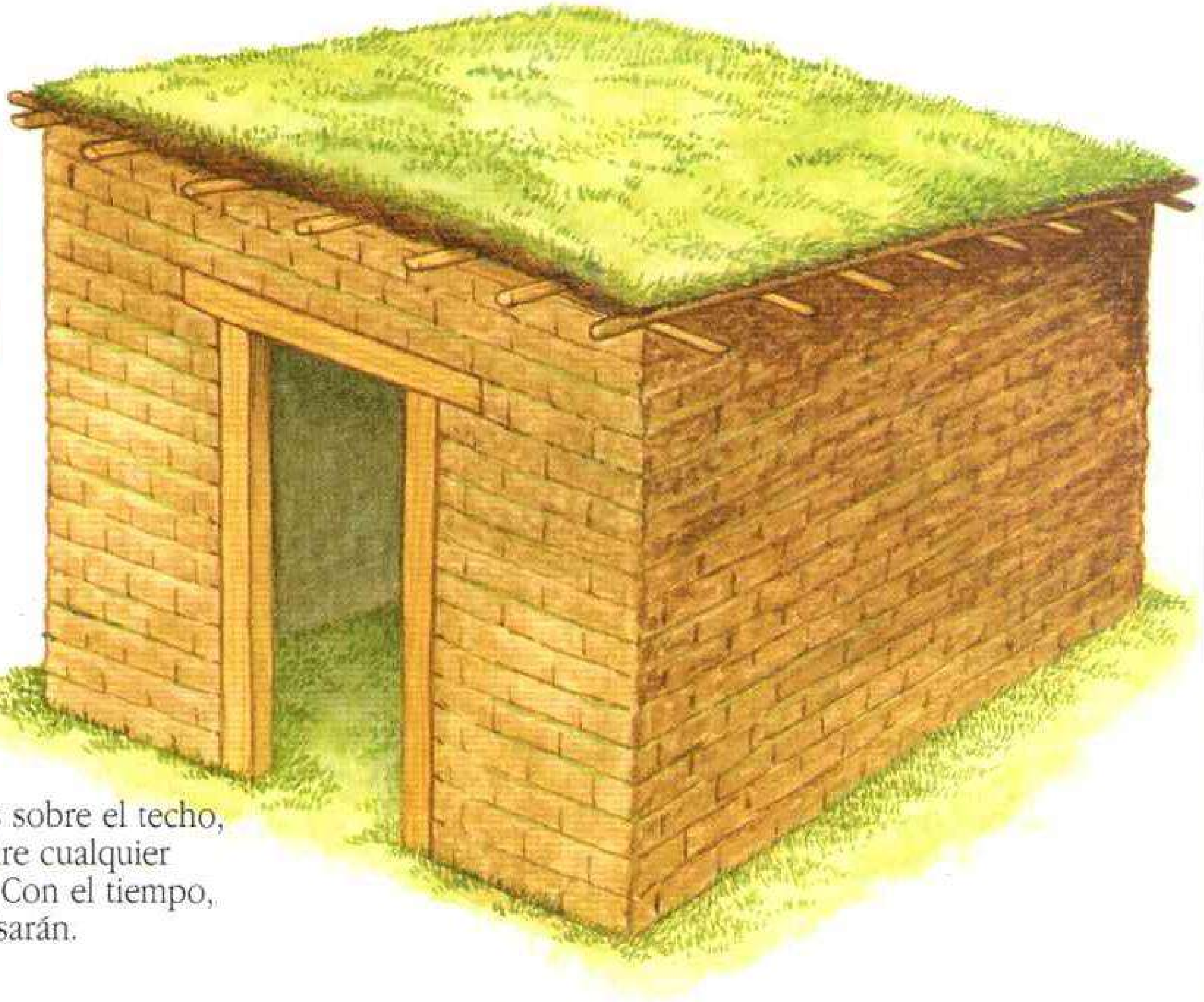

5 Coloque una capa de bloques sobre el techo, con el césped hacia arriba. Repare cualquier grieta en las paredes con barro. Con el tiempo, las paredes se endurecerán y alisarán.

OTROS REFUGIOS PERMANENTES

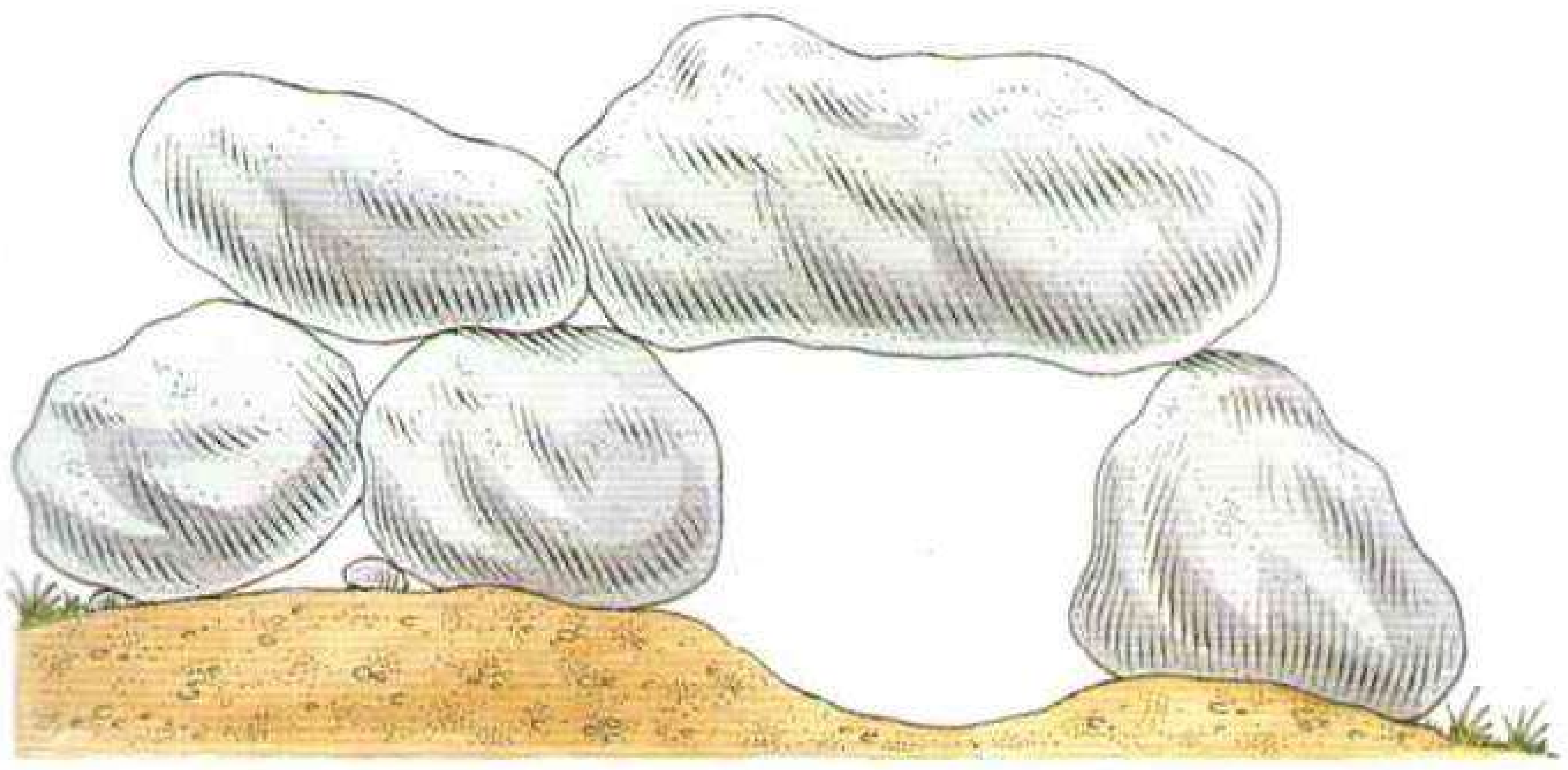

Huecos naturales. *En ocasiones pueden encontrarse refugios naturales bajo las grandes formaciones rocosas, pero debe tener cuidado y asegurarse de que la estructura es sólida e inamovible. Si excava el suelo por debajo, no toque los fundamentos de la estructura.*

Cuevas. *Cuando utilice una cueva como refugio, comience por explorarla en su totalidad. La entrada puede estar obstruida entre dos paredes. El fuego debe encenderse en el fondo de la cueva, de manera que el humo escape y no se acumule en el interior. Puede construir una pared de rocas o de troncos en el lado opuesto al fuego para reflejar nuevamente el calor hacia el interior de la cueva.*

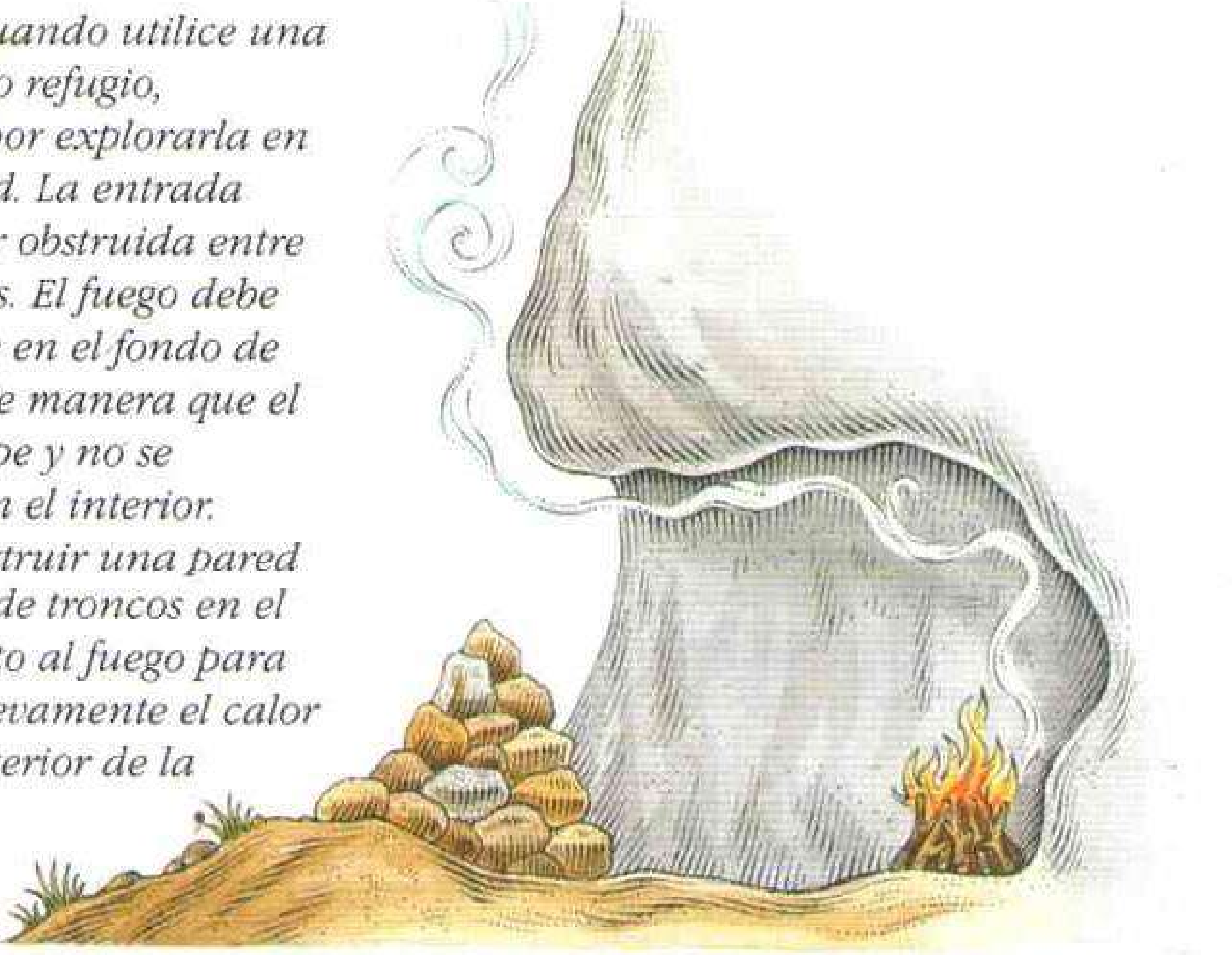

---

* ***Advertencia:*** Este método destruye gran cantidad de cubierta vegetal, la cual tarda años en regenerarse. Por tanto, su construcción sólo debe emprenderse en caso de absoluta necesidad.

UN CAMPAMENTO SEGURO

UNA VEZ ELEGIDO el campamento, se deberán cumplir unas normas básicas de seguridad. En cada situación puede existir un peligro potencial, desde los fuegos y las cocinas sin ventilación hasta la falta de higiene básica. Nunca cocine dentro de una tienda, ya que puede incendiarse rápidamente y arder con fuerza. Cualquier tipo de llama dentro de una tienda reduce la cantidad de oxígeno disponible y puede generar gases venenosos, como el monóxido de carbono. Las tiendas y cualquier otro tipo de refugio deben estar bien ventilados *(véase página 114)*. Sin neveras y aseos, aumenta la posibilidad de sufrir una intoxicación por alimentos u otras enfermedades.

LA SEGURIDAD EN EL CAMPAMENTO

Si acampa por más de una noche, debe vigilar que su campamento sea lo más seguro posible. Los accidentes ocurren a menudo por la noche, por ejemplo, cuando las personas salen de su tienda para ir a la letrina y tropiezan con el fuego, caen al río o se pierden en la oscuridad sin ser capaces de volver a encontrar su tienda.

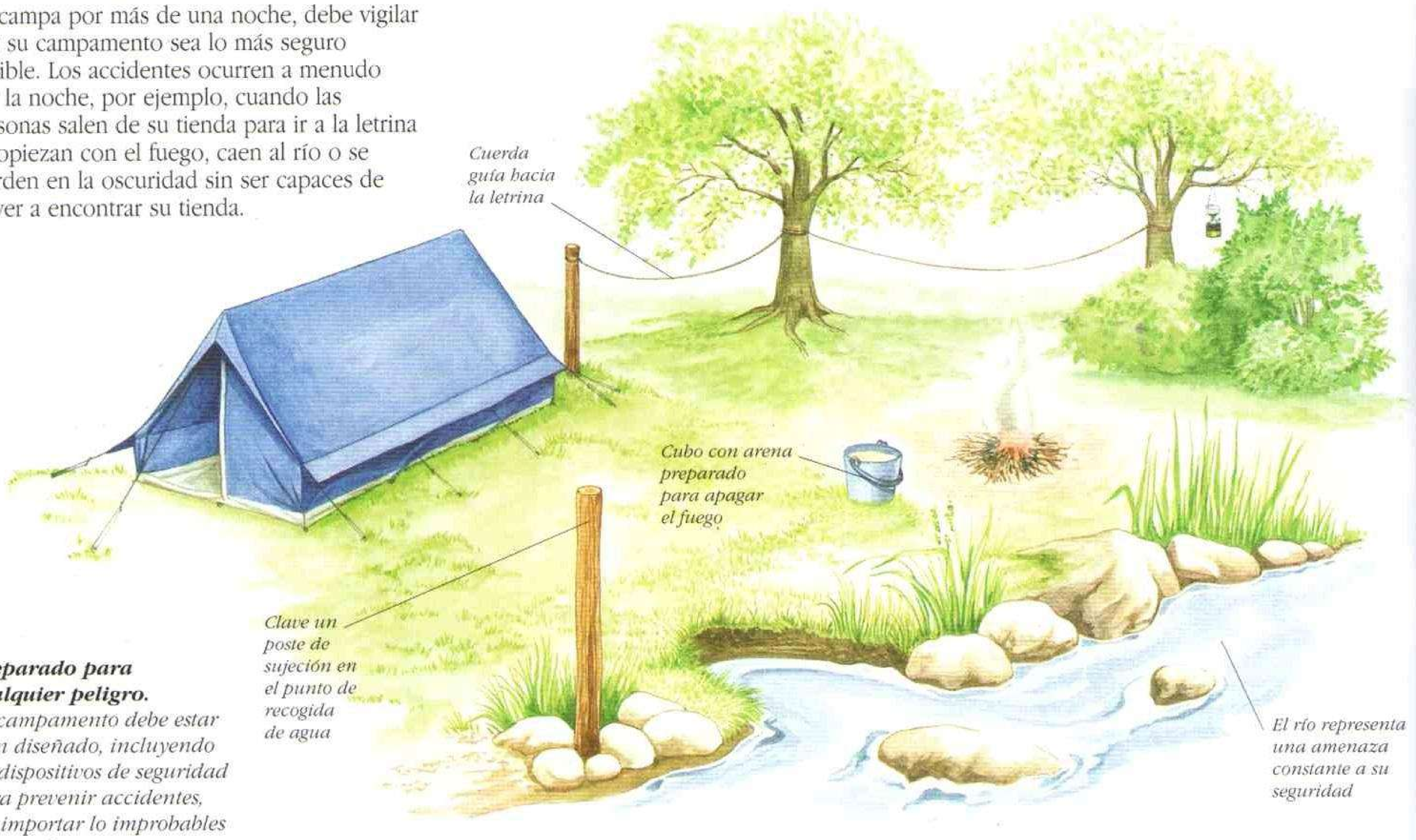

Cuerda guía hacia la letrina

Cubo con arena preparado para apagar el fuego

Clave un poste de sujeción en el punto de recogida de agua

El río representa una amenaza constante a su seguridad

Preparado para cualquier peligro. *Su campamento debe estar bien diseñado, incluyendo los dispositivos de seguridad para prevenir accidentes, sin importar lo improbables que puedan parecer.*

LA CONSTRUCCIÓN DE UNA LETRINA

La letrina debe colocarse a favor del viento y más abajo de su tienda. Espolvoree un poco de tierra cada vez que la use. Nunca vierta sustancias químicas a una letrina, ya que evitan la descomposición natural de la materia fecal y la hacen oler mucho peor de lo que lo haría naturalmente. Construya un punto diferente para orinar, marcado por una estaca, a un lado. No se olvide de rellenar la letrina cuando marche *(véase página 64)*.

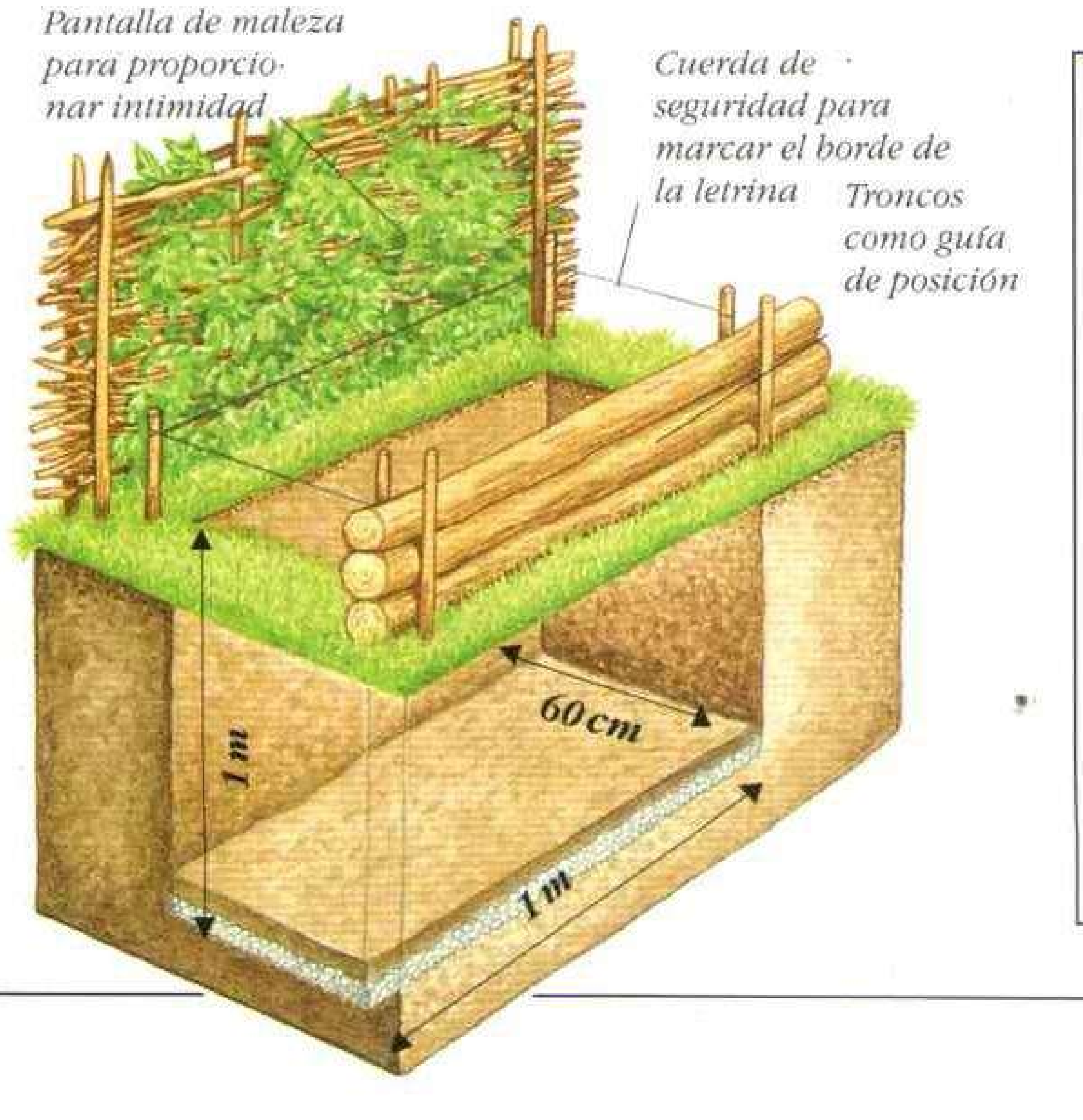

Pantalla de maleza para proporcionar intimidad

Cuerda de seguridad para marcar el borde de la letrina

Troncos como guía de posición

Inodoro de campamento. *Construya una letrina si debe permanecer en el campamento varios días, o si se trata de un grupo numeroso. Utilícelo únicamente para los desechos sólidos.*

LATA COMO INODORO

Allí donde sea ilegal o resulte poco práctico enterrar desechos humanos, construya un inodoro portátil colocando una bolsa de plástico dentro de una lata de café con tapa. Eche un poco de lejía en la bolsa para reducir el olor. Utilice la lata para los desechos sólidos.

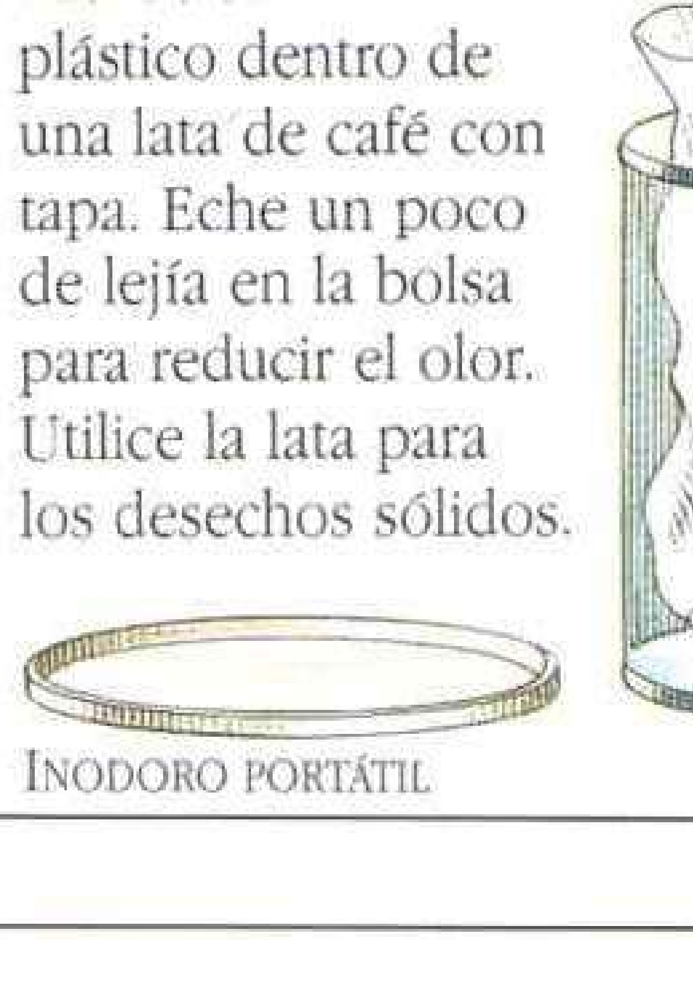

INODORO PORTÁTIL

EL LAVADO DE LA ROPA

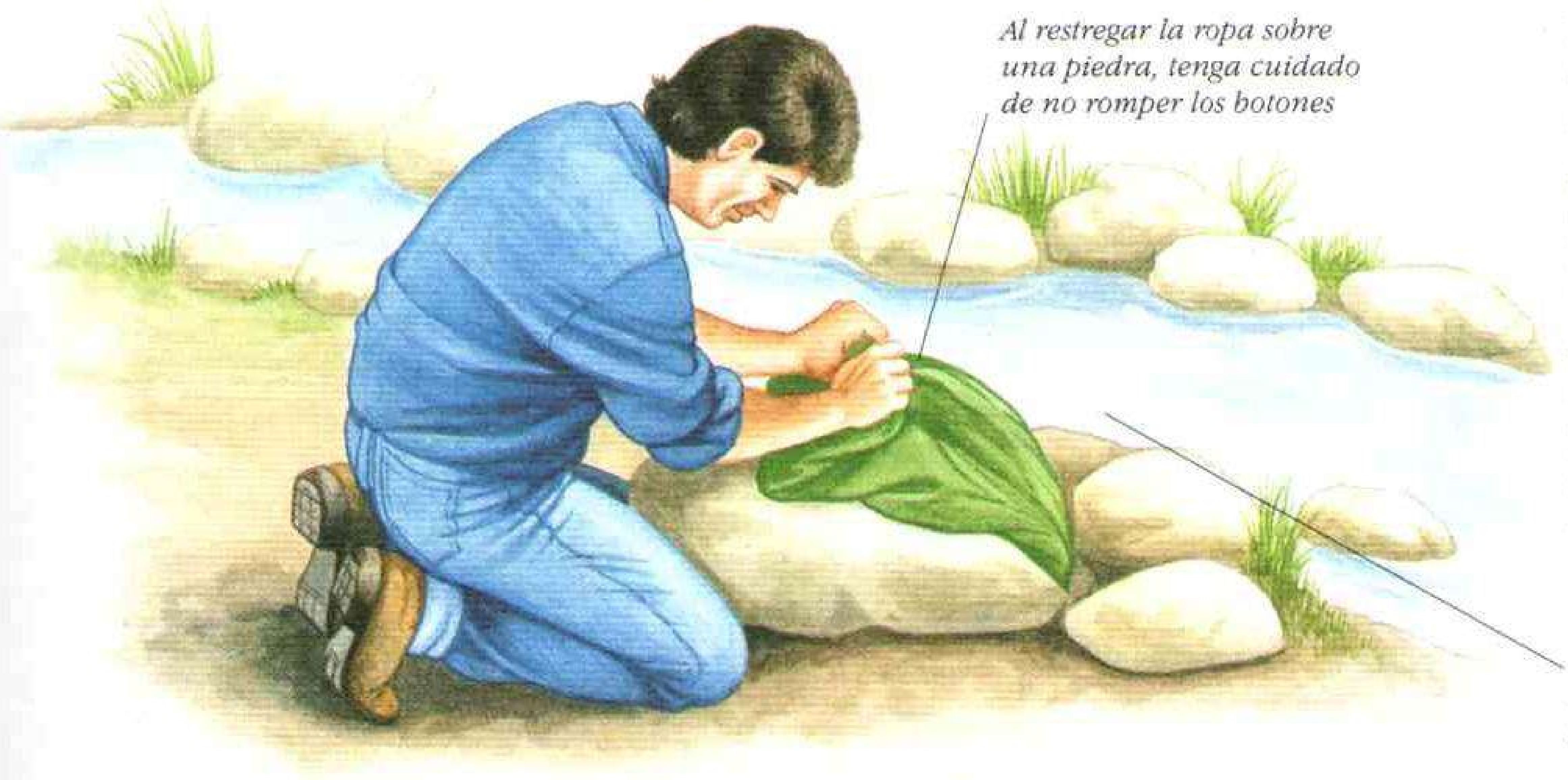

Al restregar la ropa sobre una piedra, tenga cuidado de no romper los botones

No deje que la espuma del detergente contamine su fuente de agua

Si su ropa está limpia, retendrá su capacidad aislante, y también contribuirá a su higiene personal. En las regiones tropicales, la ropa debe lavarse cada día, guardando una muda para llevar por la noche. En los climas fríos, únicamente debe lavar regularmente los calcetines y la ropa interior. La ropa debe remendarse tan pronto se rompa *(véase página 31)*, para que el roto no se haga mayor y la ropa resulte inservible. La ropa limpia y remendada ayuda a mantener la moral en las situaciones difíciles.

HIGIENE PERSONAL

Guarde sus artículos de aseo personal en un neceser impermeable para evitar que se pierdan. Si utiliza una bolsa también los mantendrá limpios. Mientras utilice sus artículos de aseo, cuelgue el neceser de un árbol para acceder rápidamente a cada artículo. Una toalla grande puede tener muchas aplicaciones. También puede cortarse en trozos menores para utilizarla de varias maneras *(véase página 27)*.

NECESER

TOALLA

LIMPIADORES IMPROVISADOS

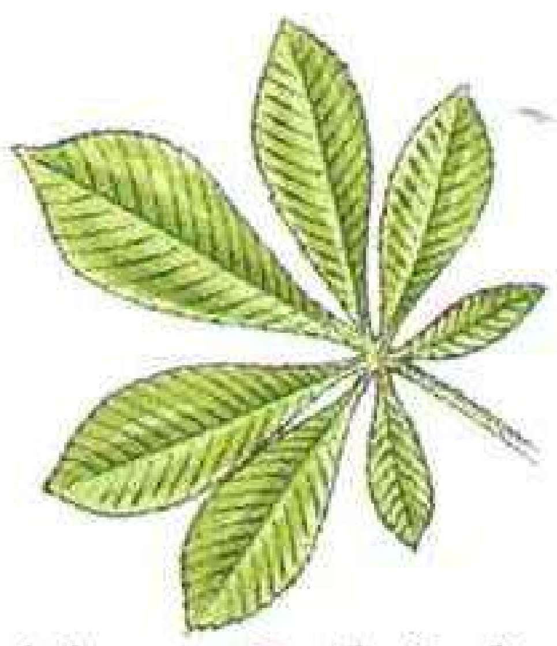

Hoja del castaño de Indias. *Las hojas del castaño de Indias trituradas en agua caliente pueden utilizarse como jabón. También tienen un suave efecto antiséptico.*

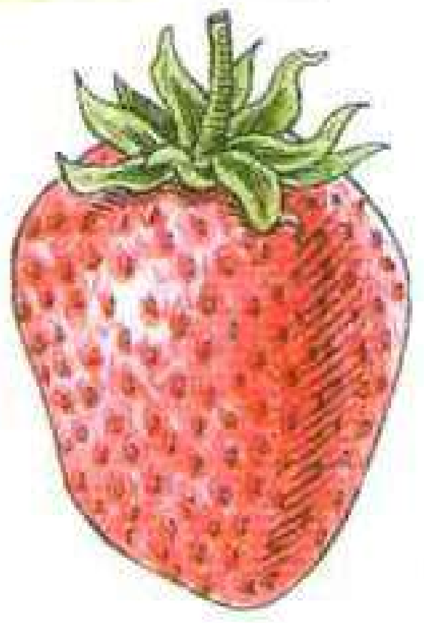

Fresa. *Puede eliminar manchas de los dientes frotando fresas sobre ellos. La ceniza de madera también puede emplearse como pasta de dientes, o como jabón.*

LA PROTECCIÓN DEL MEDIO AMBIENTE

- Evite que la espuma del jabón, los restos de comida o cualquier otro contaminante lleguen a su fuente de agua potable. Recuerde que, aunque lave sus platos y se asee corriente abajo, puede haber otros campamentos río abajo.
- Evite dejar una tienda levantada en el mismo lugar durante un tiempo prolongado para permitir el crecimiento de la hierba.
- No corte los árboles y las plantas a menos que deba construir un refugio.
- Utilice los hoyos para fuego existentes cuando sea posible, antes de hacer uno nuevo, y tenga cuidado de no provocar un incendio.
- Queme toda la basura y llévese el residuo cuando se marche.
- Al abandonar el campamento, recuerde rellenar la letrina y el hoyo del fuego, y eliminar cualquier rastro de su estancia *(véase página 64)*.

MANTENGA SU COMIDA A SALVO

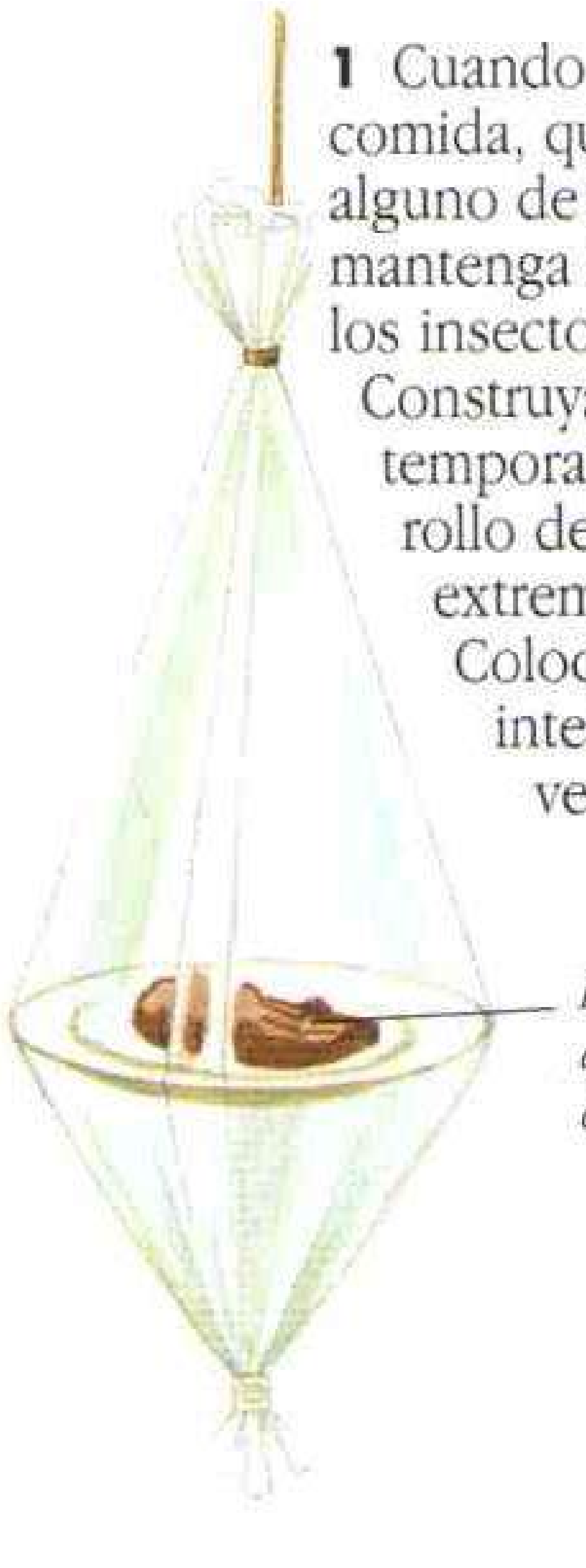

1 Cuando prepare la comida, quizá quiera que alguno de los ingredientes se mantenga fresco y alejado de los insectos y animales. Construya un contenedor temporal a partir de un rollo de tela, atando ambos extremos para cerrarlo. Coloque un plato en el interior, que hará las veces de soporte.

Plato con alimentos a salvo de los insectos

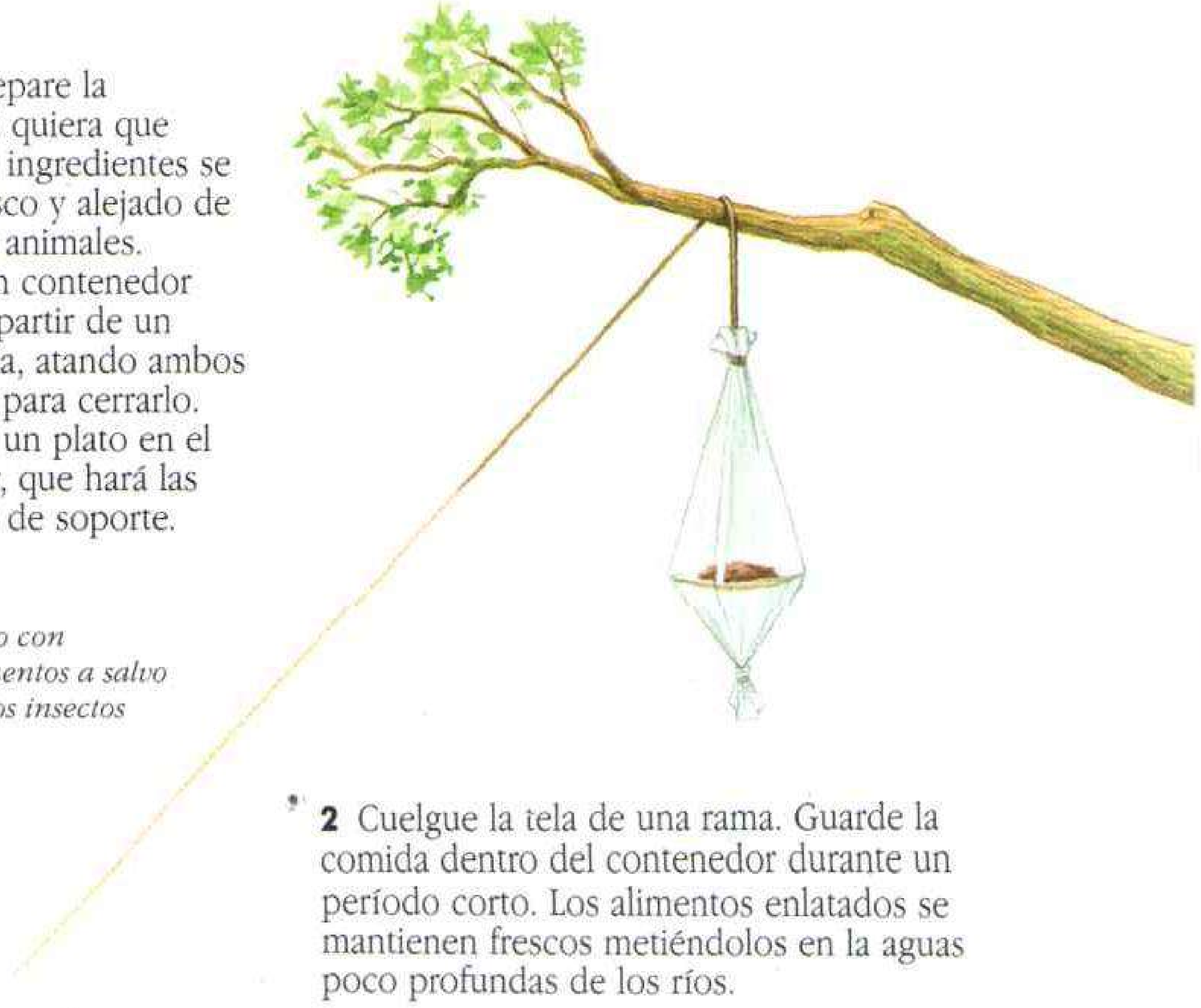

2 Cuelgue la tela de una rama. Guarde la comida dentro del contenedor durante un período corto. Los alimentos enlatados se mantienen frescos metiéndolos en la aguas poco profundas de los ríos.

EL FUEGO EN EL CAMPAMENTO

EL HECHO DE tener o no un fuego en el campamento puede determinar su supervivencia, ya que el fuego le proporcionará calor, un medio para cocinar sus alimentos, y le servirá también de estímulo. El fuego aleja a los animales salvajes y a los insectos; da calor, con lo que evita el consumo de calorías; convierte los alimentos incomibles en una deliciosa comida, y seca la ropa y las botas. Necesitará tres ingredientes para encender un fuego: yesca, astillas y leña. Un fuego también necesita oxígeno para quemar. La mejor fuente de combustible seco es la madera muerta que todavía no ha caído. Si debe usar leña caída, no recoja la que está en contacto con el suelo, sino las ramas que lo cubren.

TIPOS DE YESCA

Musgo. *El musgo verde y húmedo no sirve como yesca. El musgo muerto y seco es fino y denso, ideal para comenzar un fuego. Puede encontrar musgo junto a los troncos de los árboles o en terrenos pantanosos. Quizá deba secarlo antes de su uso.*

Hongos. *Algunos hongos que crecen en los árboles tienen una piel exterior impermeable y una carne suave, seca y esponjosa en el interior, que constituye una yesca excelente. Probablemente deba eliminar la piel exterior con un cuchillo para acceder al suave tejido interior.*

Corteza. *Incluso en climas húmedos, la corteza interior de los troncos caídos puede estar seca. Elimine los trozos húmedos o busque allí donde los insectos se han introducido produciendo en la madera un fino serrín, que es igualmente una buena yesca. La madera podrida también puede emplearse como yesca.*

Hojas muertas. *Siempre hay hojas secas y muertas en el suelo, incluso en los bosques más húmedos. Meta las hojas directamente en una bolsa impermeable según las vaya recogiendo. Puede utilizarlas enteras o estrujarlas para formar trozos pequeños antes de utilizarlas como yesca.*

Hierba seca. *Es muy fácil de recoger. Un manojo de hierba seca se encenderá rápidamente produciendo llamas* (véase inferior).

PREPARAR LA YESCA

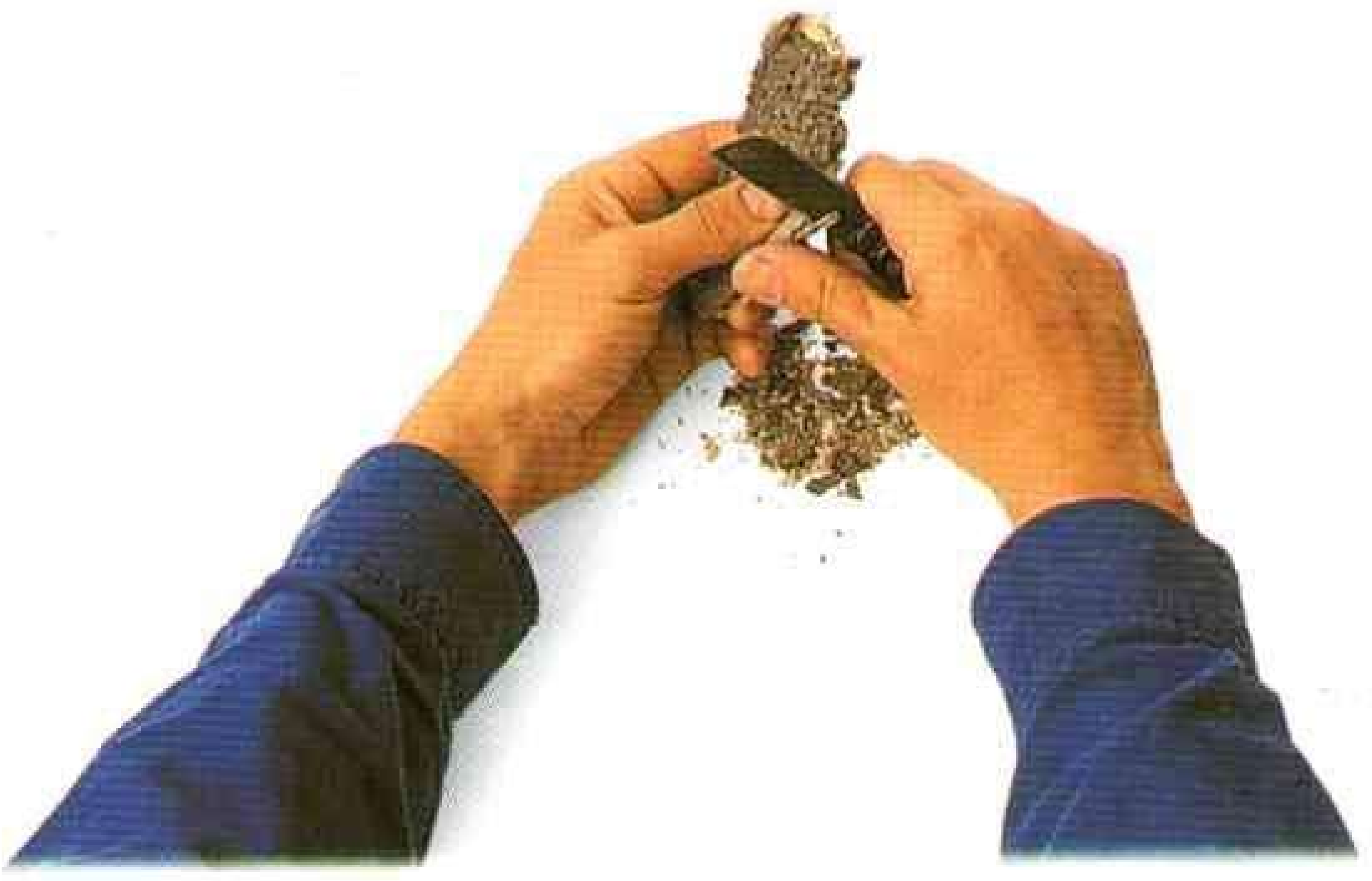

Pulverizar. *Puede convertir los palos secos y los trozos de corteza en yesca en polvo cortándolos en trozos muy pequeños con un cuchillo. Corte los trozos lo más pequeños posible, y guárdelos en una bolsa.*

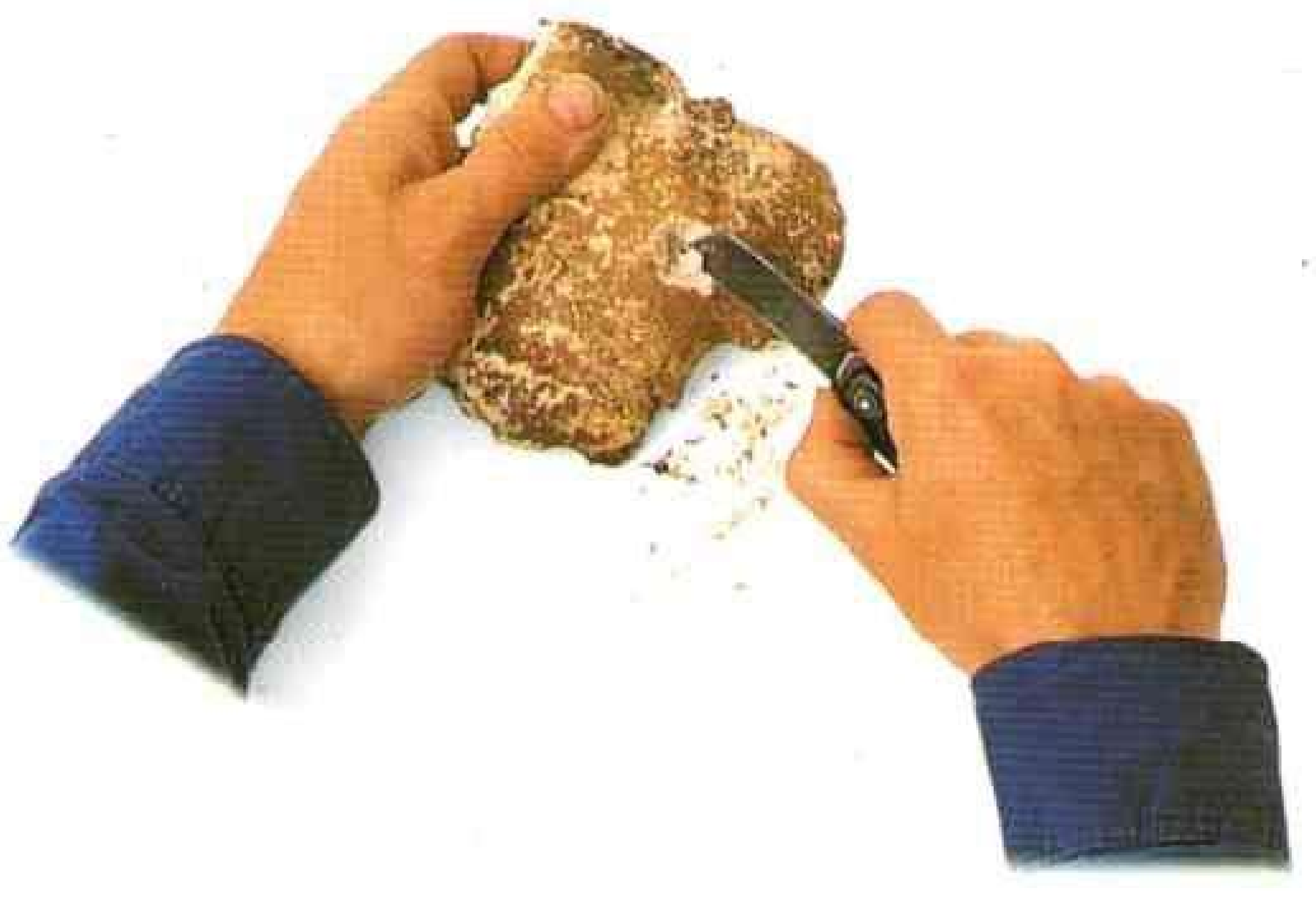

Cortar. *Haga una pequeña depresión en la superficie de la yesca, por ejemplo, en un trozo de hongo de árbol. Si deja caer un rescoldo incandescente en este hueco, éste encontrará una gran área superficial para calentar.*

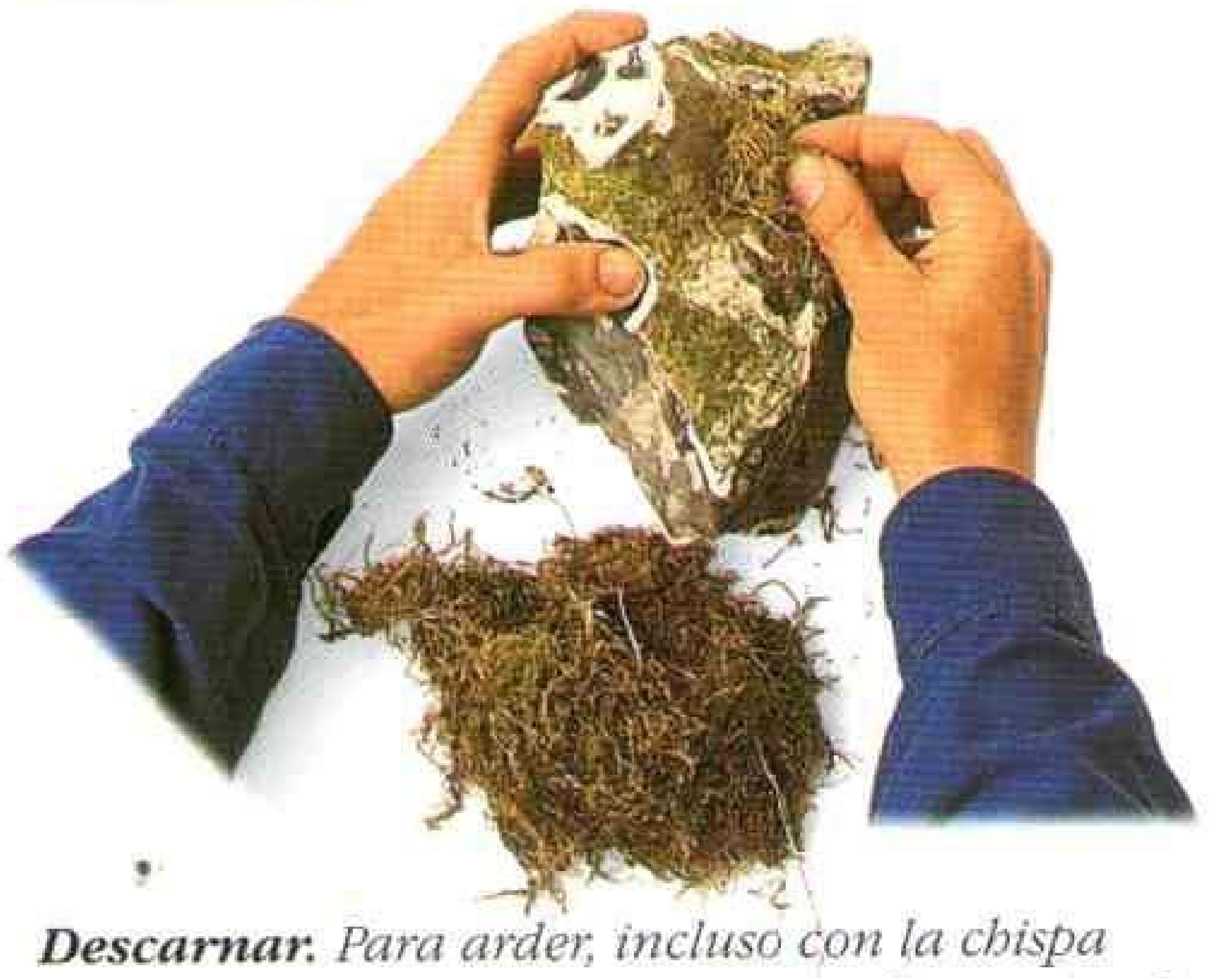

Descarnar. *Para arder, incluso con la chispa más pequeña, la yesca debe ser tan fina como el algodón. Frote y retuerza el material entre sus dedos, o usando una piedra, para separar las fibras entre sí.*

LEÑA

Leña para calentar. *Las maderas blandas como la del fresno, el abeto, el manzano, el avellano y el acebo, arden rápidamente y con mucha luz, desprendiendo mucho calor. Sin embargo, también producen chispas. Las maderas que producen una mayor cantidad de chispas al calentarse son las del cedro y la cicuta. Al arder tan rápidamente, las maderas blandas son ideales para comenzar un fuego, y pueden resultar útiles para cocinar rápidamente o, por ejemplo, para calentar agua u otros líquidos. Sin embargo, se consumen muy rápido y dejan mucha ceniza en lugar de rescoldos o brasas que podrían emplearse para hornear o para otras formas de cocina lenta.*

ACEBO

ABETO

Leña para cocinar. *Las maderas densas y duras como las de la haya, el roble, el abedul, el alerce, el nogal y el sicómoro, arden lenta y uniformemente, desprendiendo mucho calor y produciendo brasas ideales para cocinar lentamente* (véase página 117). *El sabor de la comida cocinada sobre un fuego puede quedar afectado por el olor del humo. Cada tipo de madera produce un olor diferente al quemarse. Por ejemplo, el pino tiende a dar a los alimentos un sabor a resina, mientras que el manzano y el sicómoro pueden enriquecer su sabor.*

ROBLE

HAYA

Combustible de emergencia. *En una situación de supervivencia puede verse en la necesidad de tener que emplear cualquier cosa que encuentre para quemar. Los combustibles de emergencia pueden incluir excrementos animales, líquenes secos, musgo, brezo e incluso bloques de turba; todos deben secarse al sol antes de su uso. En la playa, puede servirse de algas secas. A veces, puede encontrar restos de carbón para quemar, mientras que el petróleo se filtra ocasionalmente formando charcos en la superficie del terreno. En ocasiones, la arena contiene petróleo, y puede quemarse. También puede utilizar la grasa animal como combustible.*

EXCREMENTOS SECOS

ALGAS

MADERAS QUE DEBE EVITAR

Algunas maderas resinosas estallan con fuerza al fuego, y deben evitarse dentro de lo posible. Éstas incluyen el pino y el endrino. Otras maderas como el aliso, el castaño, el álamo y el sauce no queman bien y se limitan a arder sin llama. El bambú estallará sin calentarlo, a menos que lo abra primero.

ÁLAMO

SAUCE

INGREDIENTES PARA UN FUEGO

Debe hacer el fuego gradualmente, comenzando por astillas de madera y progresando hacia las ramas mayores y los troncos. Clasifique su madera en yesca, un puñado de ramitas muy finas y una gran cantidad de ramas pequeñas, ramas grandes y troncos. Deje que cada etapa arda bien antes de añadir ramas más grandes. No utilice ramas verdes, ya que éstas tienden a doblarse.

Yesca. *Necesitará un montón de yesca, al menos del tamaño de un pomelo, cortada muy fina* (véase página anterior). *La yesca es la parte más importante de un fuego, ya que no puede encender un fuego comenzando por los troncos gruesos, a menos de que disponga de un encendedor que sustituya a la yesca* (véase página 62).

Astillas. *Una vez lograda la llama en la yesca, añada leña menuda seca en forma de pequeñas ramas y hojas. Asegúrese de que la leña está totalmente seca. Las ramas deben tener el grosor de un lápiz.*

Leña. *Cuando las astillas ardan correctamente, añada ramas del grosor de un dedo. Las ramas harán que el tamaño de las llamas aumente, desde su fase inicial hasta el tamaño normal para un fuego.*

Leña gruesa. *Las ramas grandes actúan como combustible principal. Deben ser más gruesas que un dedo, pero fáciles de romper. El tamaño de este combustible es el más habitual en los fuegos de campamento.*

Troncos. *Los troncos gruesos son ideales para mantener encendido un fuego estable durante toda la noche. Asegúrese de que están totalmente quemados antes de apagar el fuego.*

CÓMO PREPARAR UN FUEGO

EL SECRETO DE un buen fuego está en disponer de combustible seco y encenderlo poco a poco a partir de una llama pequeña. Debe escoger y construir su fuego con gran cuidado, eligiendo el más adecuado para sus propósitos, al tiempo que se adapta a las condiciones locales *(véase página siguiente)*. Los fuegos son potencialmente peligrosos —puede perder el control del mismo, incendiando la vegetación seca, la tienda o su ropa. Asimismo, pueden gastar todo el oxígeno disponible en un refugio cerrado asfixiando a sus habitantes. Un fuego también puede marcar el paisaje si no se apaga y limpia convenientemente *(véase página 65)*.

FUEGO SENCILLO EN ARMAZÓN DE TIPI

1 Levante un cuadrado de hierba y déjelo a un lado. Coloque una plataforma de ramas verdes en el hueco.

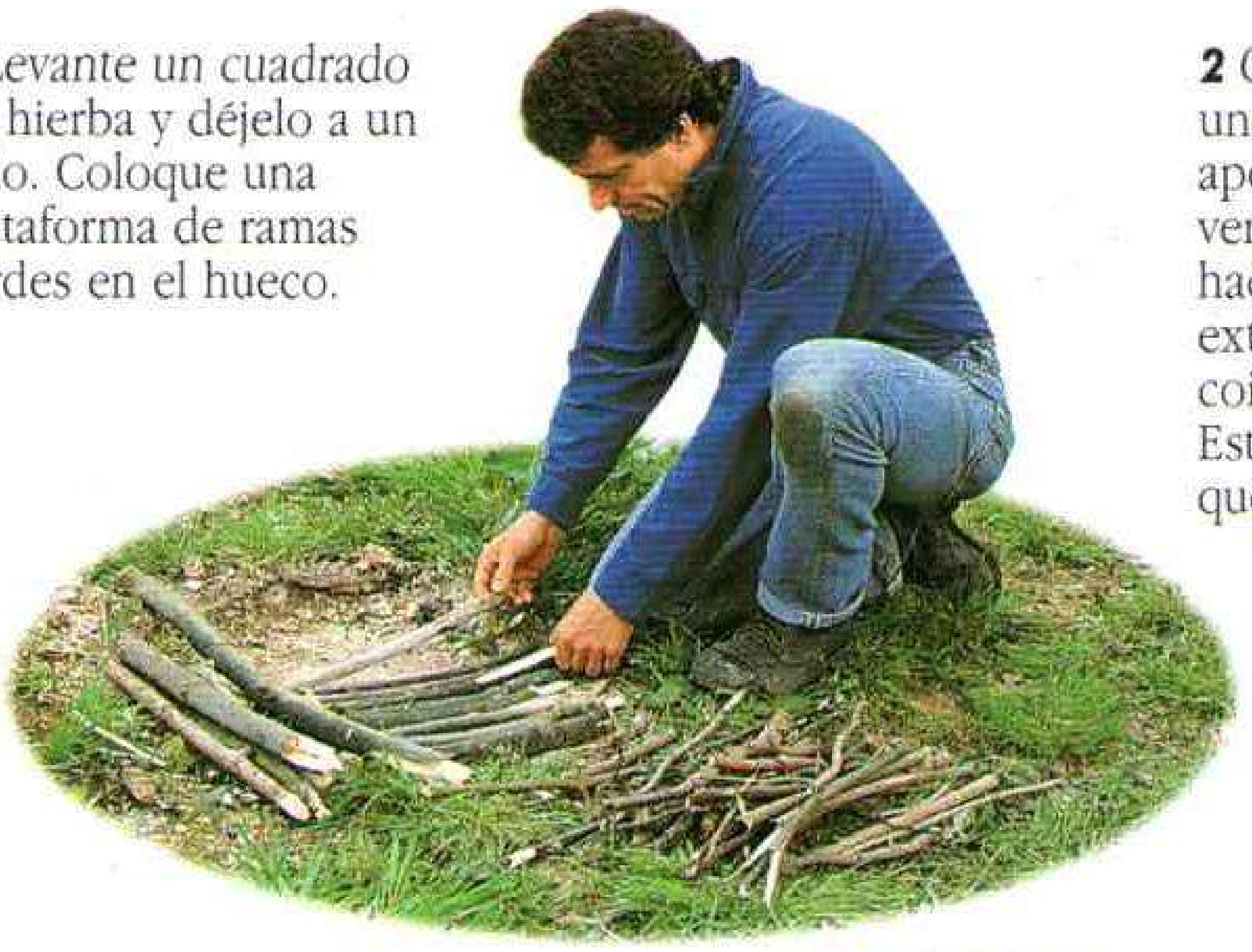

2 Comience a construir un armazón de tipi apoyando cuatro ramas verticales entre sí, haciendo que sus extremos superiores coincidan en un punto. Este armazón no tiene que ser muy grande.

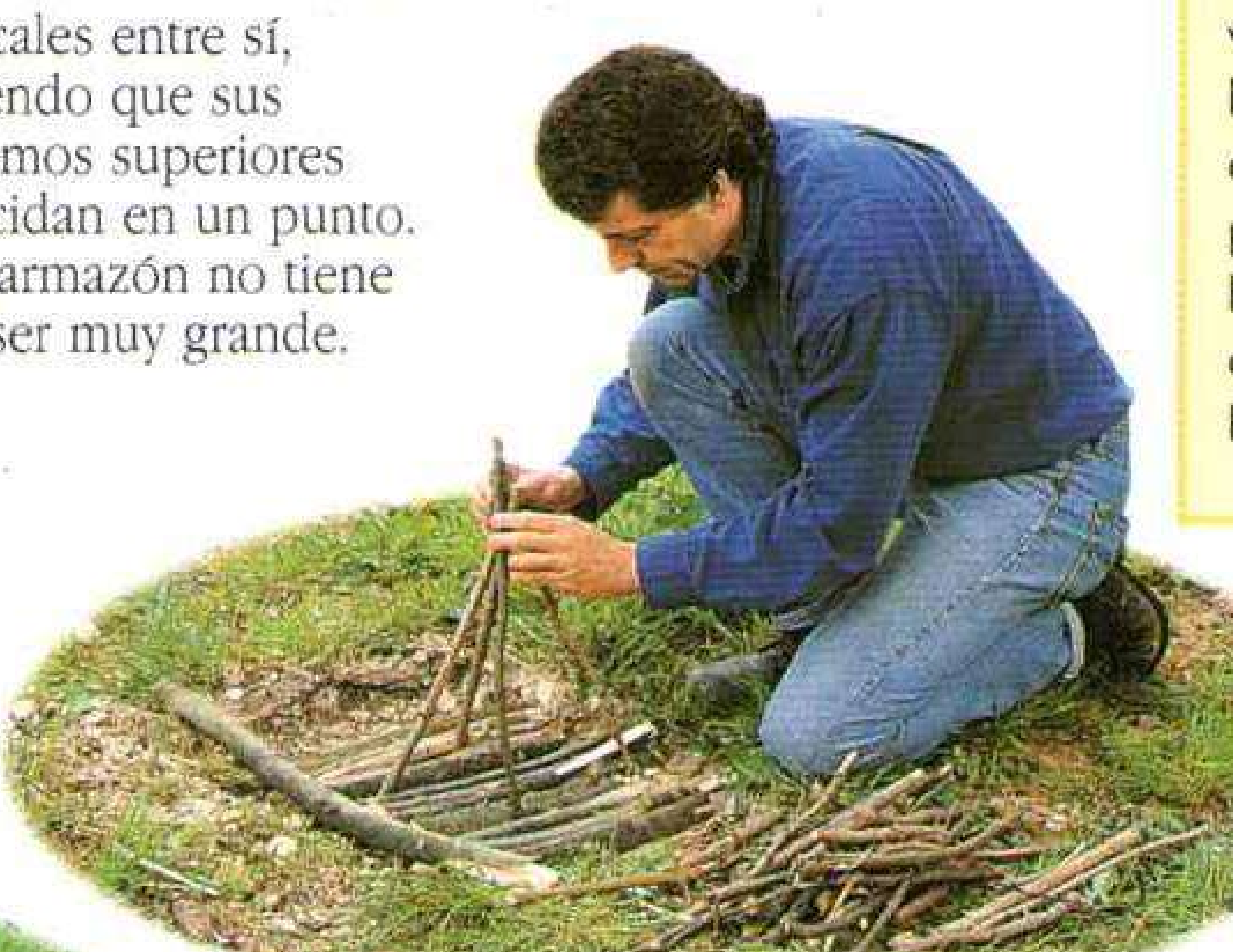

PRECAUCIÓN

No construya un fuego donde la vegetación esté seca debido a la sequía. No construya un fuego de dimensiones mayores de las que pueda controlar, especialmente si hace viento. Tenga cuidado de no construir un fuego debajo de ramas bajas que podrían incendiarse.

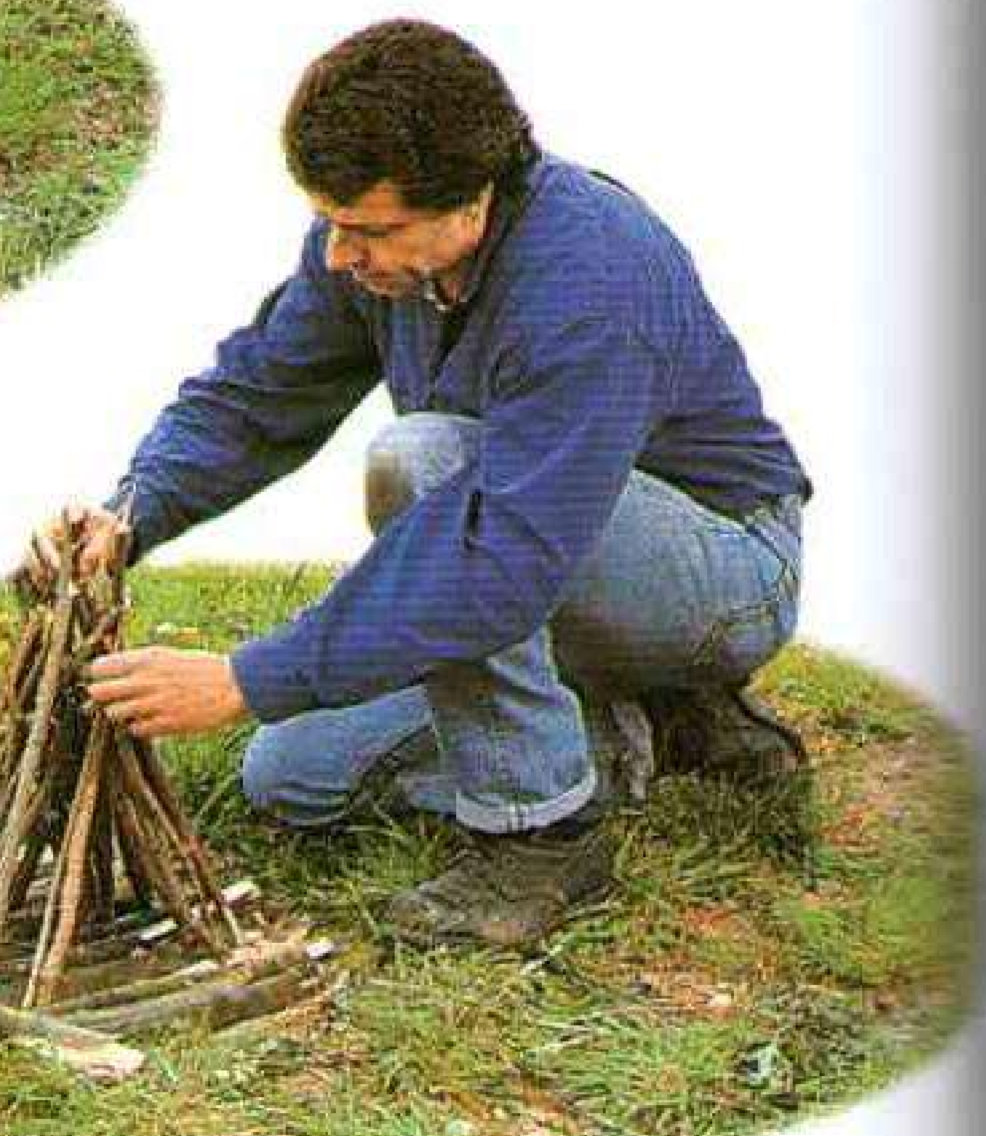

3 Construya un tipi gradualmente, intentando hacerlo lo más robusto posible. Asegúrese de dejar el suficiente espacio debajo del tipi para colocar la yesca, y no olvide dejar una abertura en uno de los lados.

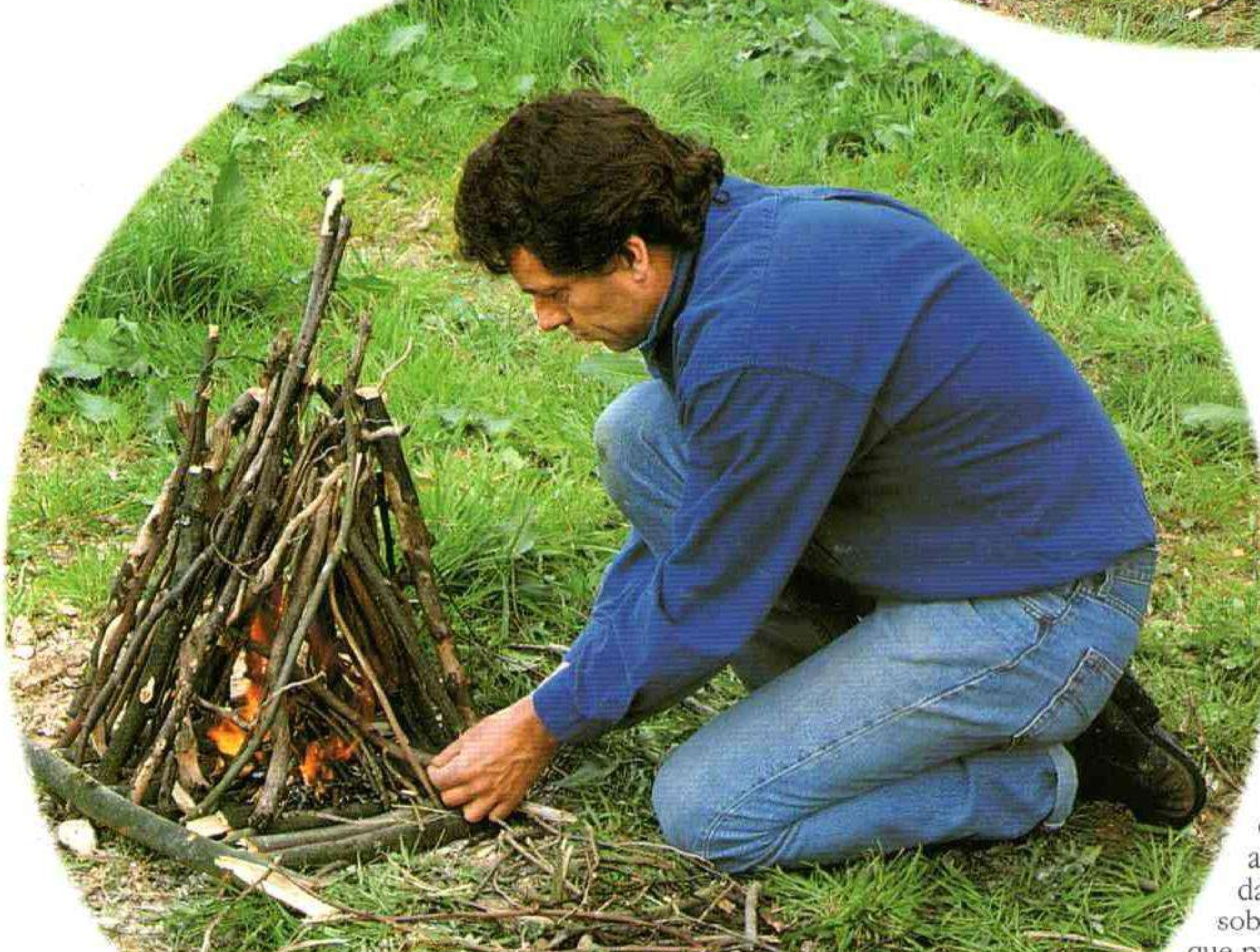

4 Coloque la yesca dentro del tipi terminado. Encienda la yesca *(véase página 58)*, y después añada hojas secas y pequeñas ramas para aumentar el tamaño de las llamas dentro del tipi. Conforme se acumula el calor, el tipi también comienza a quemarse, dando lugar a llamas grandes. El tipi se derrumbará sobre sí mismo, y obtendrá así un lecho de brasas al que puede añadir más leña, o utilizarlo para cocinar. La plataforma de madera también llegará a quemarse, originando más brasas.

Tenga ramitas y hojas a mano para alimentar el fuego

TIPOS DE FUEGO

Aunque puede utilizar el fuego sencillo en armazón de tipi en casi cualquier área, puede encontrarse con que deba adaptarlo a ciertas condiciones especiales. Por ejemplo, si hace mucho viento, puede construir el fuego en una trinchera para que esté protegido y arda correctamente. En la nieve, tendrá que escarbar hasta llegar al suelo sobre el que colocará la plataforma para construir el fuego.

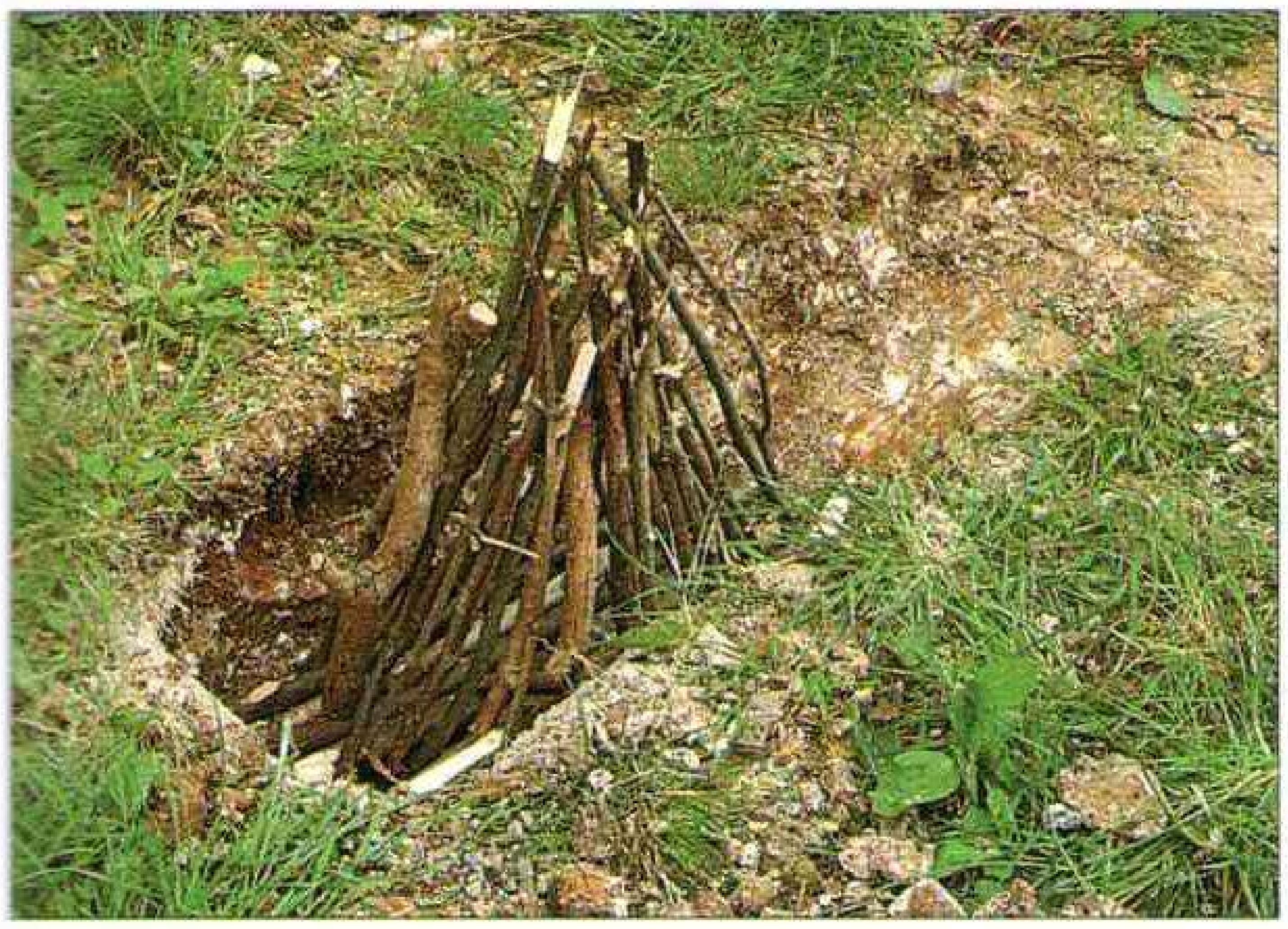

Fuego de trinchera. *En condiciones de viento fuerte, cave una trinchera de unos 30 cm de profundidad. Los fuegos de trinchera son excelentes para cocinar, y tienen un bajo consumo de leña. No llegan a producir grandes llamaradas pero, a pesar de su construcción, disponen del aire suficiente para arder correctamente.*

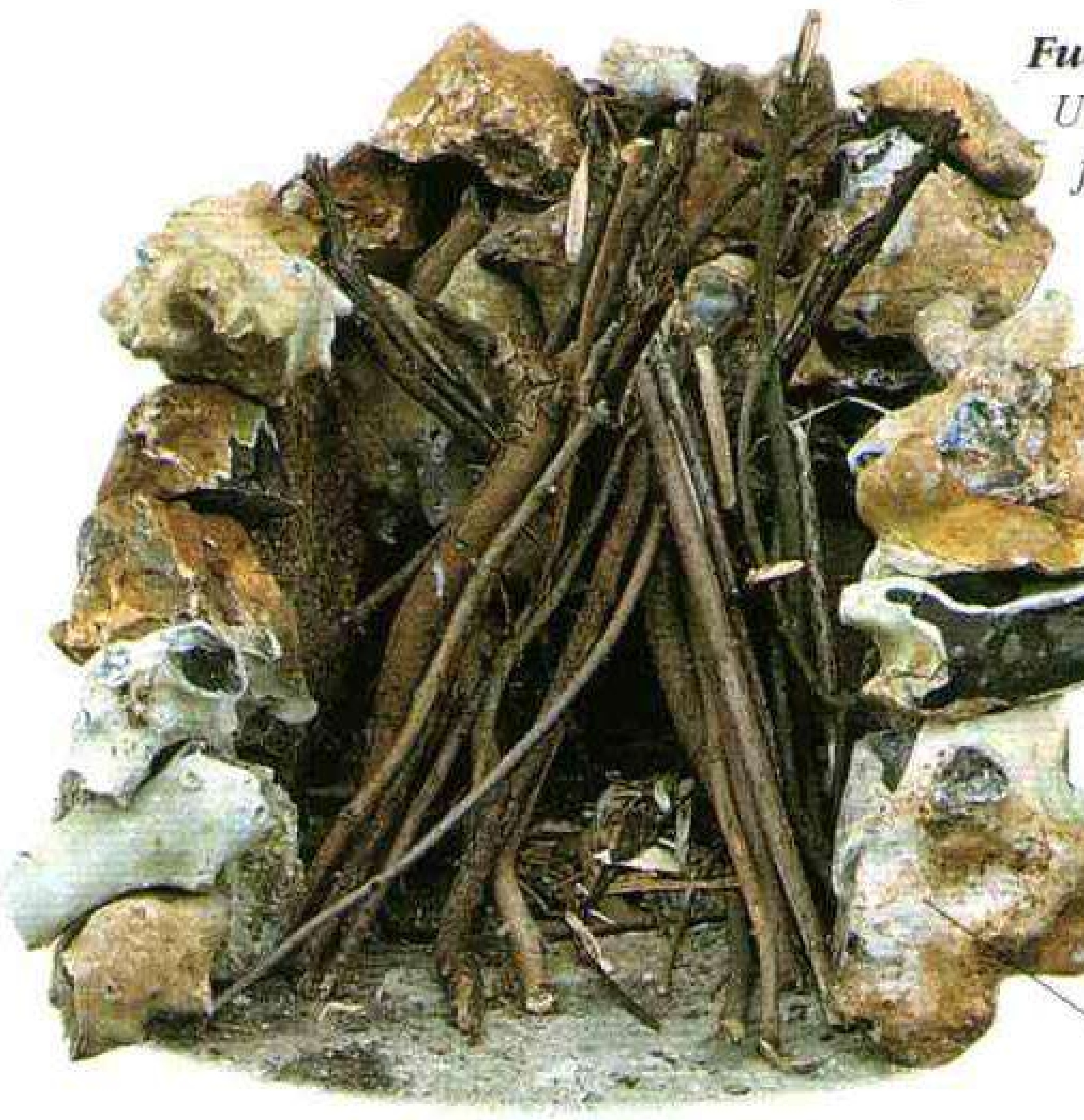

Fuego en una pared rocosa. *Una pared rocosa protegerá el fuego del viento, impedirá las llamaradas peligrosas y limitará el consumo de leña. Elija rocas secas, no porosas, y evite la pizarra, que puede explotar al calentarla. Utilice las rocas más estables y de mayor tamaño que pueda encontrar. Puede rellenar los huecos con barro o arcilla para formar un área funcional para cocinar.*

La pared rocosa protege el fuego del viento

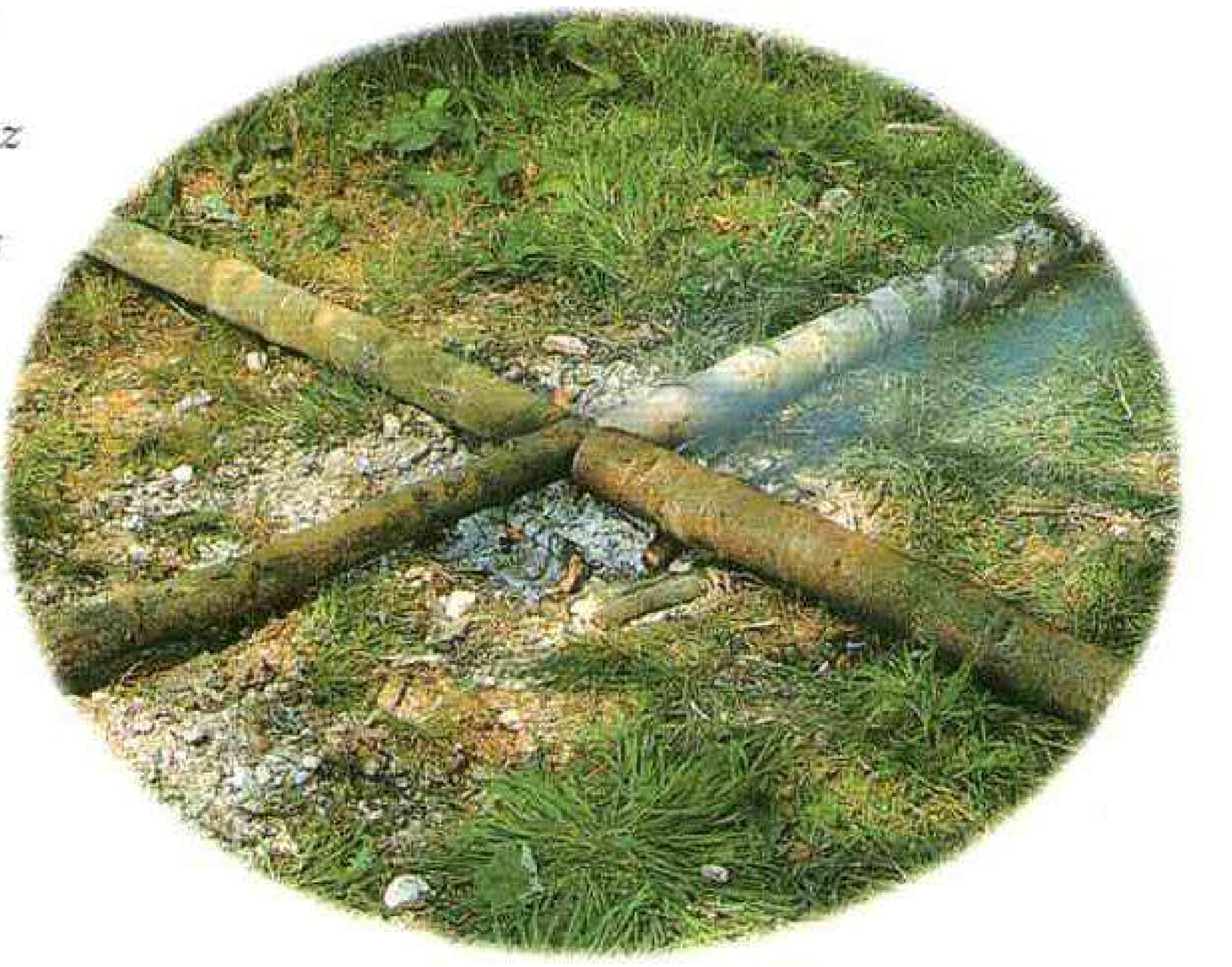

Fuego de estrella. *Una vez que su fuego dispone de un buen lecho de rescoldos o brasas, coloque cuatro troncos grandes en el centro. Conforme se van quemando, debe irlos empujando hacia el centro. Utilice este tipo de fuego únicamente en un campamento a largo plazo, para asegurarse de quemar los troncos en su totalidad antes de apagar el fuego* (véase página 64).

FUEGO DE LARGA DURACIÓN

Para que un fuego dure toda una noche, o todo un día, coloque tres troncos gruesos muy juntos sobre un lecho profundo de brasas. Si quiere calor adicional, puede construir una pared de madera u otro material detrás del fuego para reflejar el calor hacia usted.

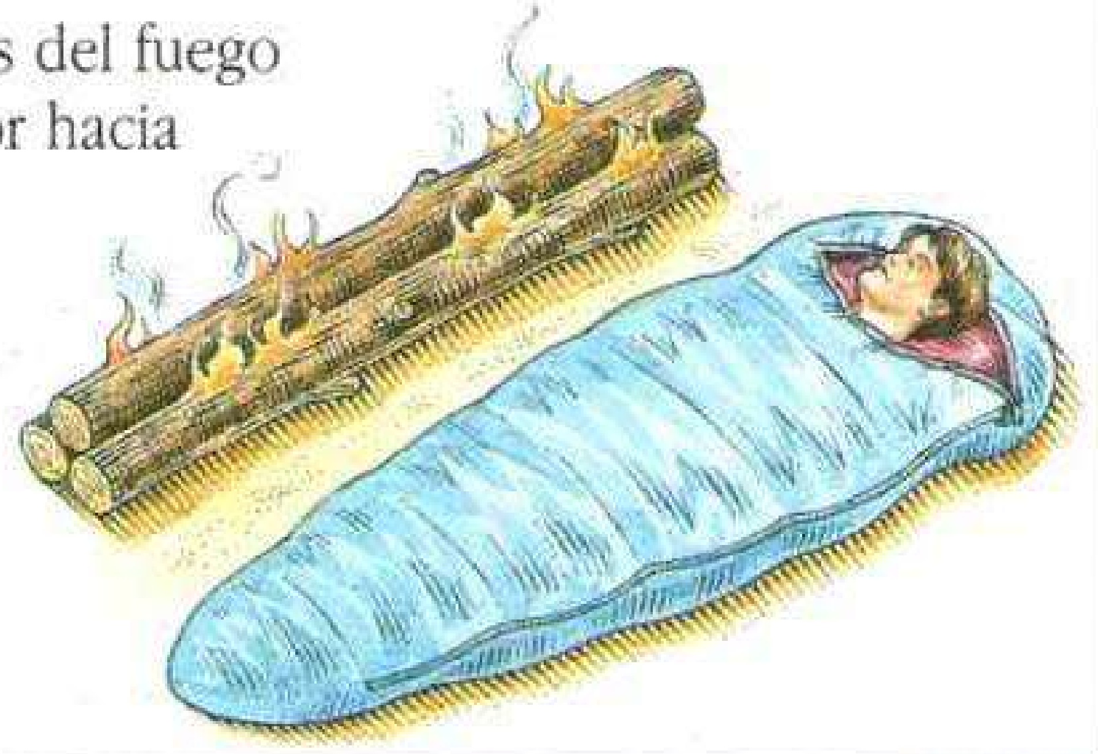

Fuego de larga duración. *Un tronco largo puede arder toda la noche.*

CÓMO TRANSPORTAR EL FUEGO

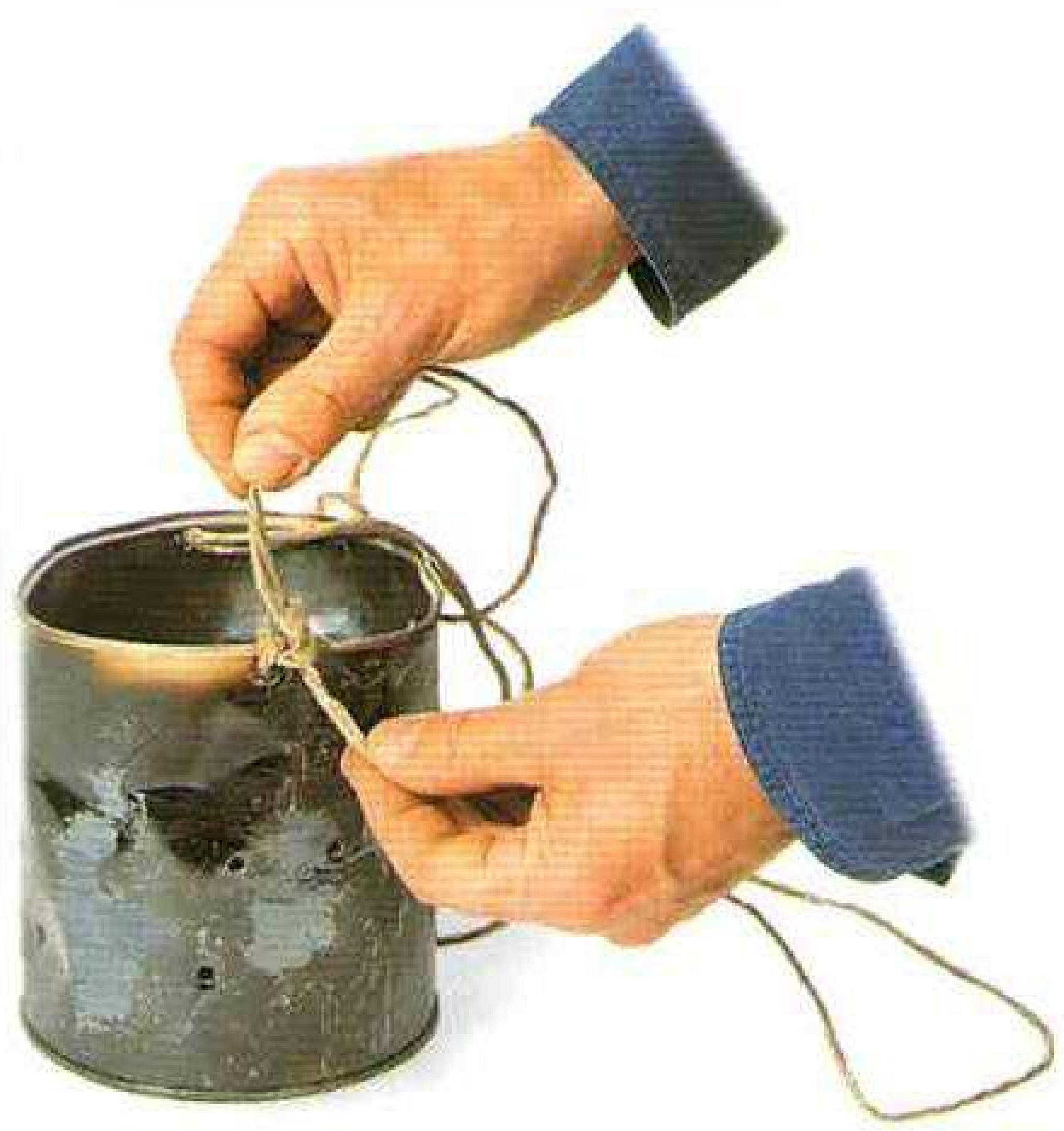

1 Puede mantener vivos los rescoldos incandescentes varios días para comenzar un nuevo fuego dondequiera que esté. Anude un asa de cuerda o alambre a la lata.

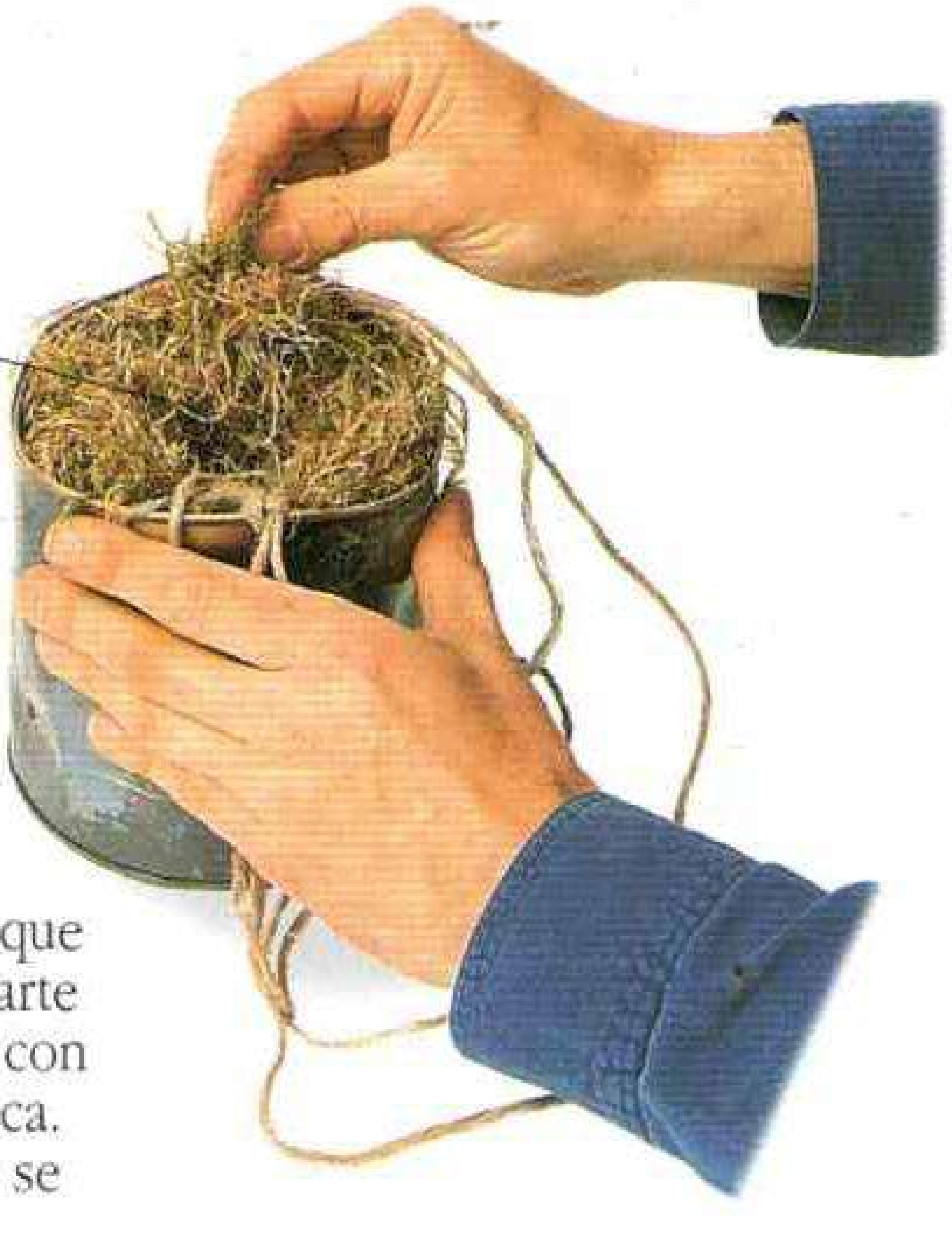

Coloque el musgo seco en la parte superior

2 Llene la lata con musgo seco. Si únicamente puede encontrar musgo húmedo, puede utilizarlo, siempre que lo coloque en la parte inferior y lo cubra con musgo o hierba seca. El musgo húmedo se secará pronto.

3 Coloque unos rescoldos incandescentes en el nido de musgo y cúbralos con más musgo seco. Sople suavemente sobre los rescoldos cuando le parezca que vayan a apagarse.

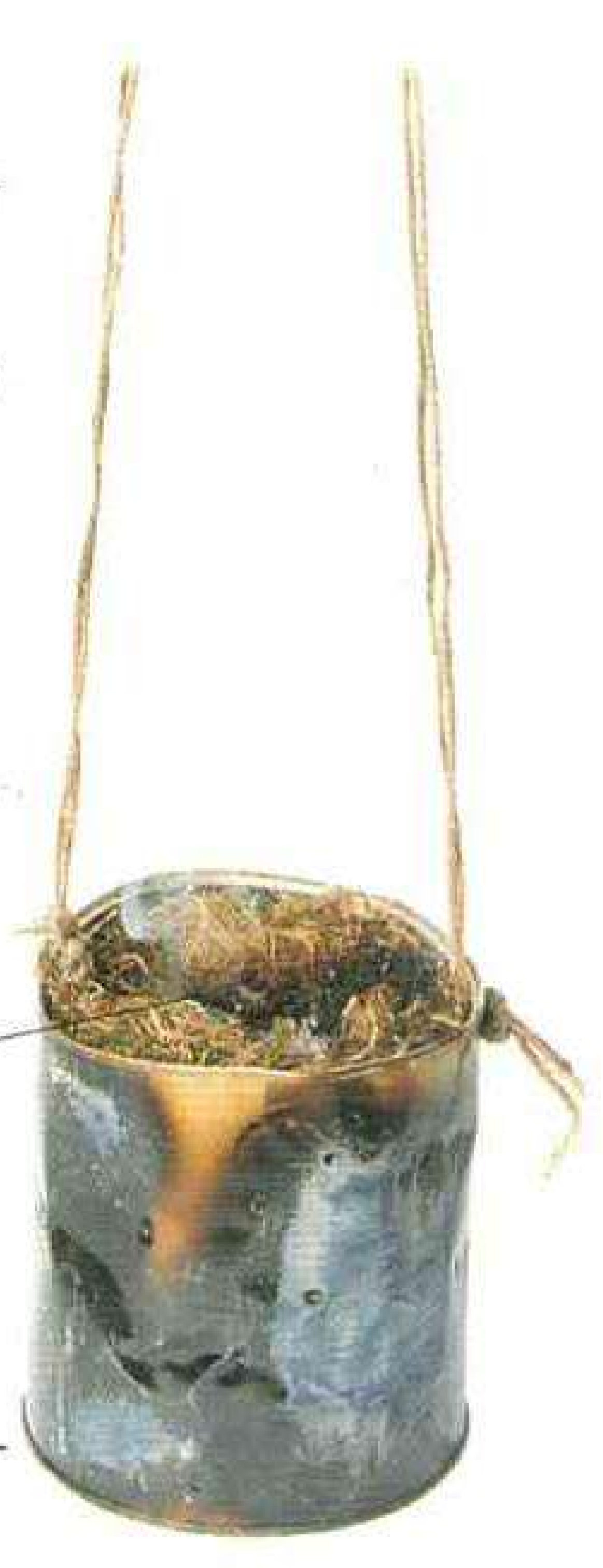

Las brasas vivas hacen que el musgo se encienda

CÓMO ENCENDER UN FUEGO

JUNTO CON LOS primeros auxilios, saber encender un fuego es la habilidad de supervivencia más necesaria. Cuando la madera u otro combustible se calienta, despide gases que se encienden originando llamas. Los fuegos deben encenderse en pequeña escala y reforzarse gradualmente. Es muy importante disponer de leña seca, ya que la madera húmeda apagará prácticamente cualquier fuego. Los fuegos no sólo necesitan calor y combustible seco, sino también una gran cantidad de oxígeno, especialmente al encenderlos. Puede apagar un fuego cubriéndolo con tierra para impedir que las llamas tengan el suficiente oxígeno para continuar ardiendo.

MÉTODOS PARA ENCENDER UN FUEGO

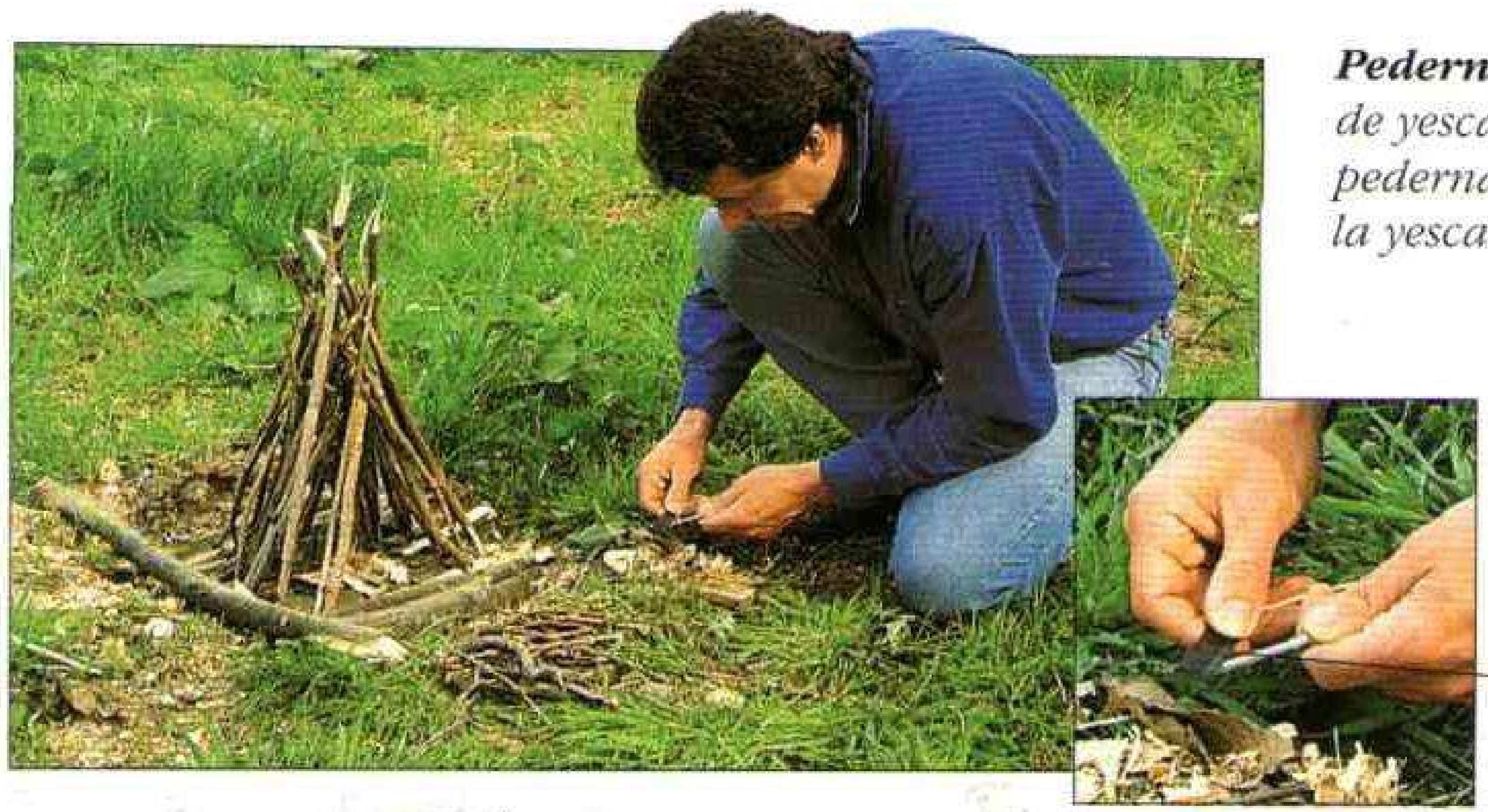

Pedernal y acero. *Haga un nido de yesca junto al tipi. Frote el pedernal con el acero, encima de la yesca. Sople sobre las chispas para crear una llama. Una vez ha encendido la yesca, empújela dentro del tipi.*

Frote el pedernal con la sierra de acero dirigiendo las chispas hacia la yesca

CONSEJOS

- Asegúrese de que el fuego está protegido del viento *(véase página 60)*; recuerde que un fuego necesita oxígeno para arder.
- Mantenga una reserva de yesca seca.
- Reúna todos los materiales necesarios antes de encender un fuego.
- Impermeabilice las cerillas dejando caer la cera de una vela sobre sus cabezas. Elimine la cera antes de encenderlas.

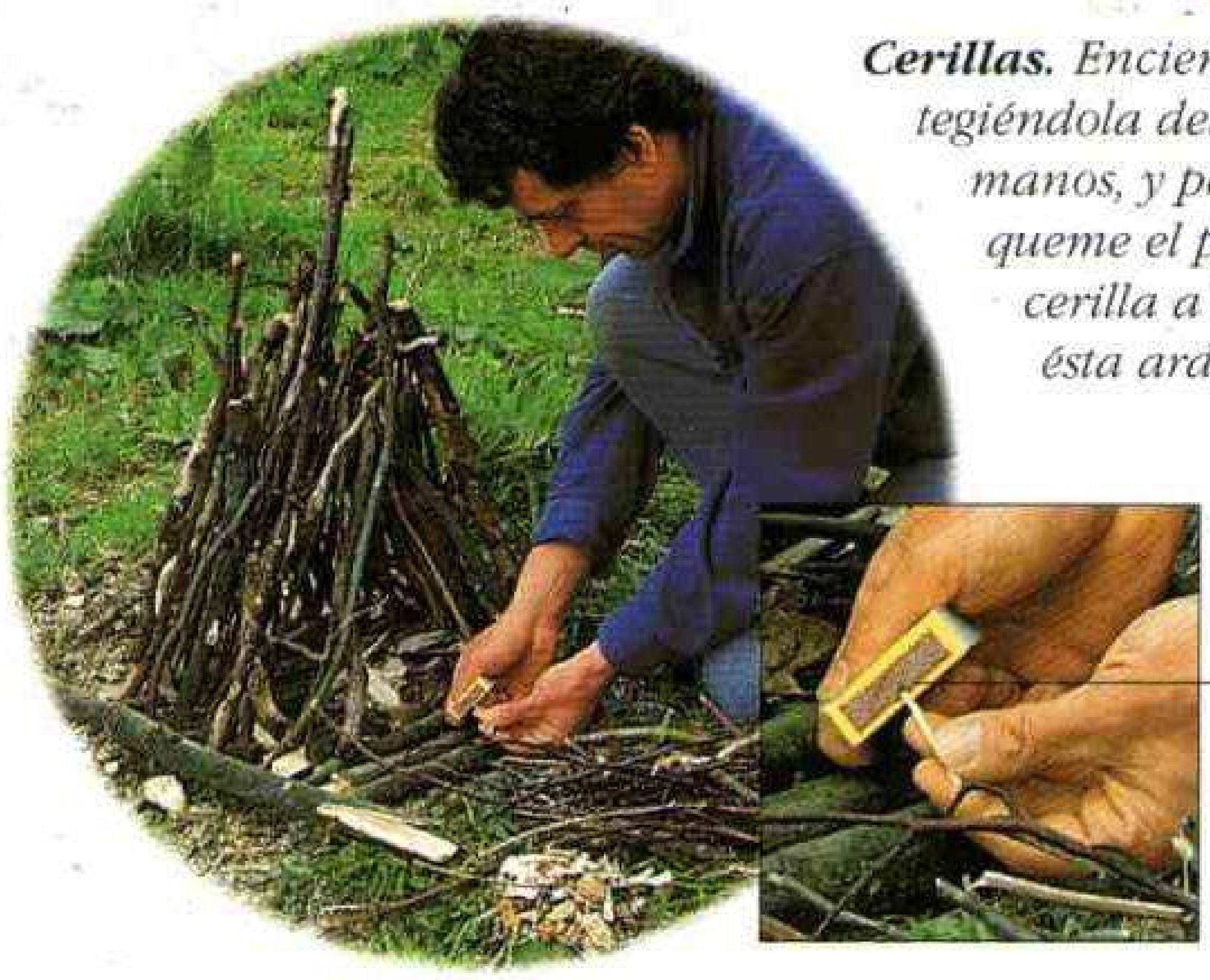

Cerillas. *Encienda una cerilla protegiéndola del viento con las manos, y permita que la llama queme el palo. Acerque la cerilla a la yesca hasta que ésta arda. Deje caer la cerilla, añada más yesca y después leña.*

Proteja la cerilla del viento con las manos

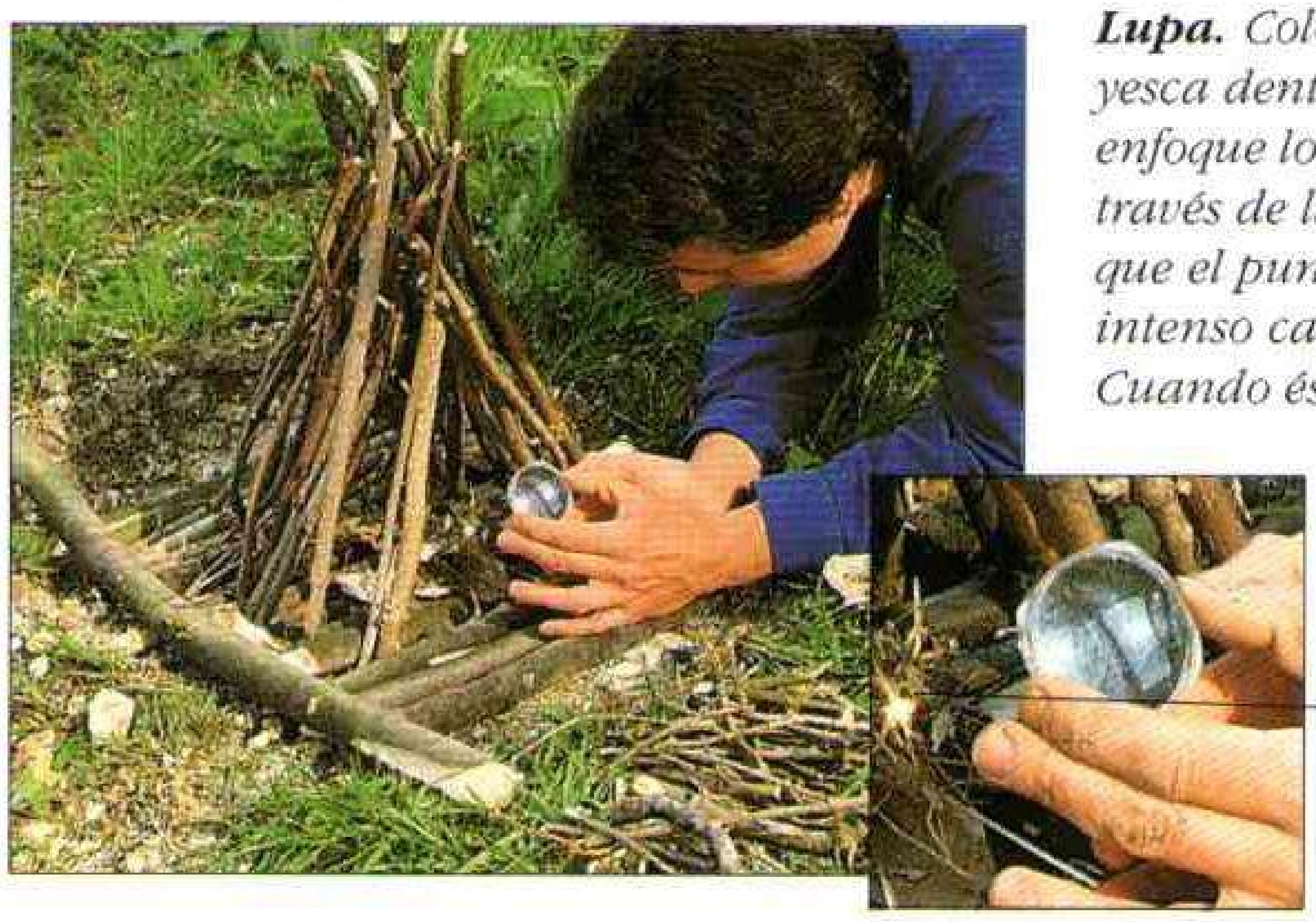

Lupa. *Coloque un nido de yesca dentro del tipi, y enfoque los rayos del sol a través de la lupa de manera que el punto de luz más intenso caiga sobre la yesca. Cuando ésta comience a arder, sople ligeramente hasta que haga llama.*

Enfoque los rayos de sol sobre la yesca a través de la lente de aumento

ÚTILES PARA ENCENDER FUEGO

Hay muchas maneras diferentes de encender un fuego empleando utensilios diversos. Éstos varían desde el tradicional bloque de pedernal y la sierra de acero hasta las modernas pastillas de parafina y los palitos de astillas de madera tratadas con varios materiales químicos. Escoja el método que le parezca más sencillo de usar.

Cerillas impermeables. *Quite la cera de la cabeza de las cerillas antes de utilizarlas.*

Pedernal y acero. *Esta sierra de acero y el «pedernal» de aleación de magnesio producen chispas mayores que las del pedernal natural.*

Lupa. *Cuanto mayor sea la lupa, mejores resultados obtendrá, particularmente con el sol débil de las latitudes boreales.*

Algodón. *El algodón impregnado de parafina es una alternativa de peso ligero a la yesca natural.*

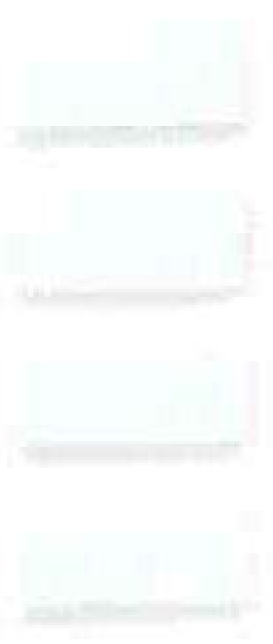

Pastillas de parafina. *Pueden utilizarse para encender ramas grandes sin yesca.*

Palos de fuego. *Los palos de virutas de madera tratada con materiales químicos pueden encenderse con una cerilla para prender fuego fácilmente, sin necesidad de yesca.*

FUEGO DE ARCO

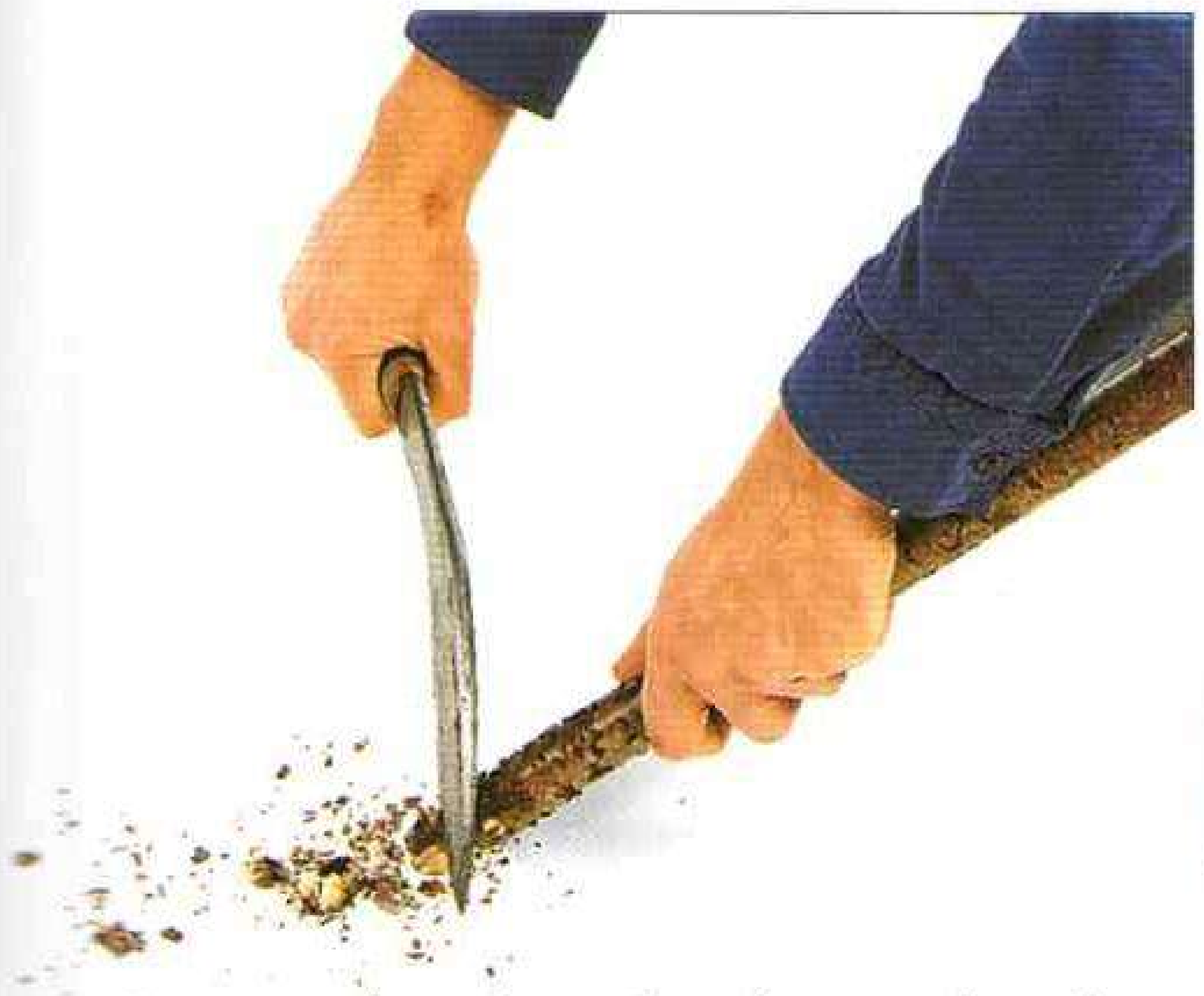

1 Corte una rama de madera dura que haga las veces de eje *(véase página 59)*. Debe tener unos 40 cm de longitud y unos 4 cm de grosor. Redondee uno de los extremos y talle una punta con el otro.

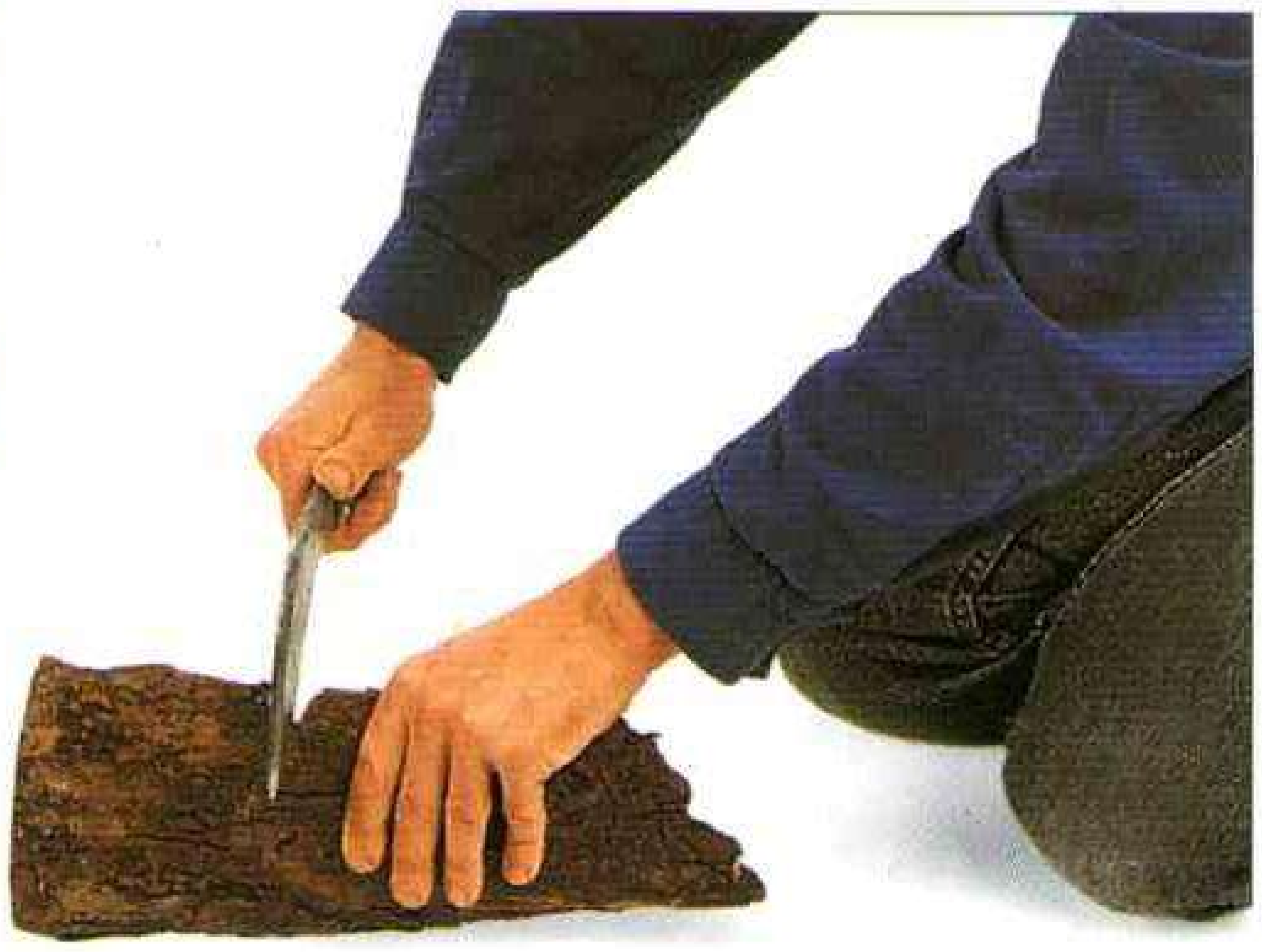

2 Busque un trozo de madera suave, como de pino o de balsa, para utilizar como base, y corte una muesca en forma de «V» en uno de los lados. La fricción entre el eje y la base producirá cenizas.

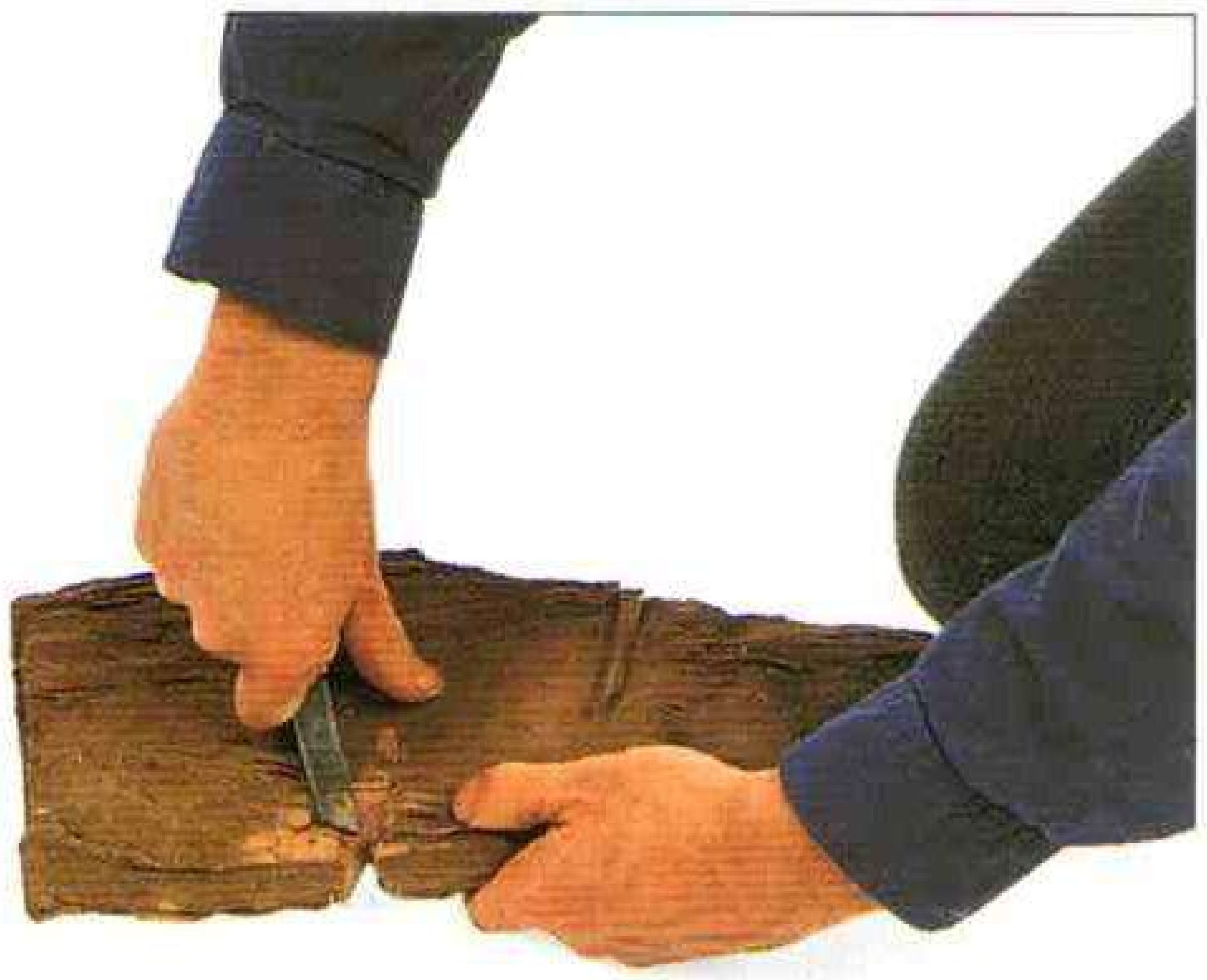

3 Talle un agujero pequeño al lado de la muesca, en el que quepa la punta del eje. La ceniza caliente y negra creada por la fricción del eje girando en el agujero caerá en la yesca a través de la muesca.

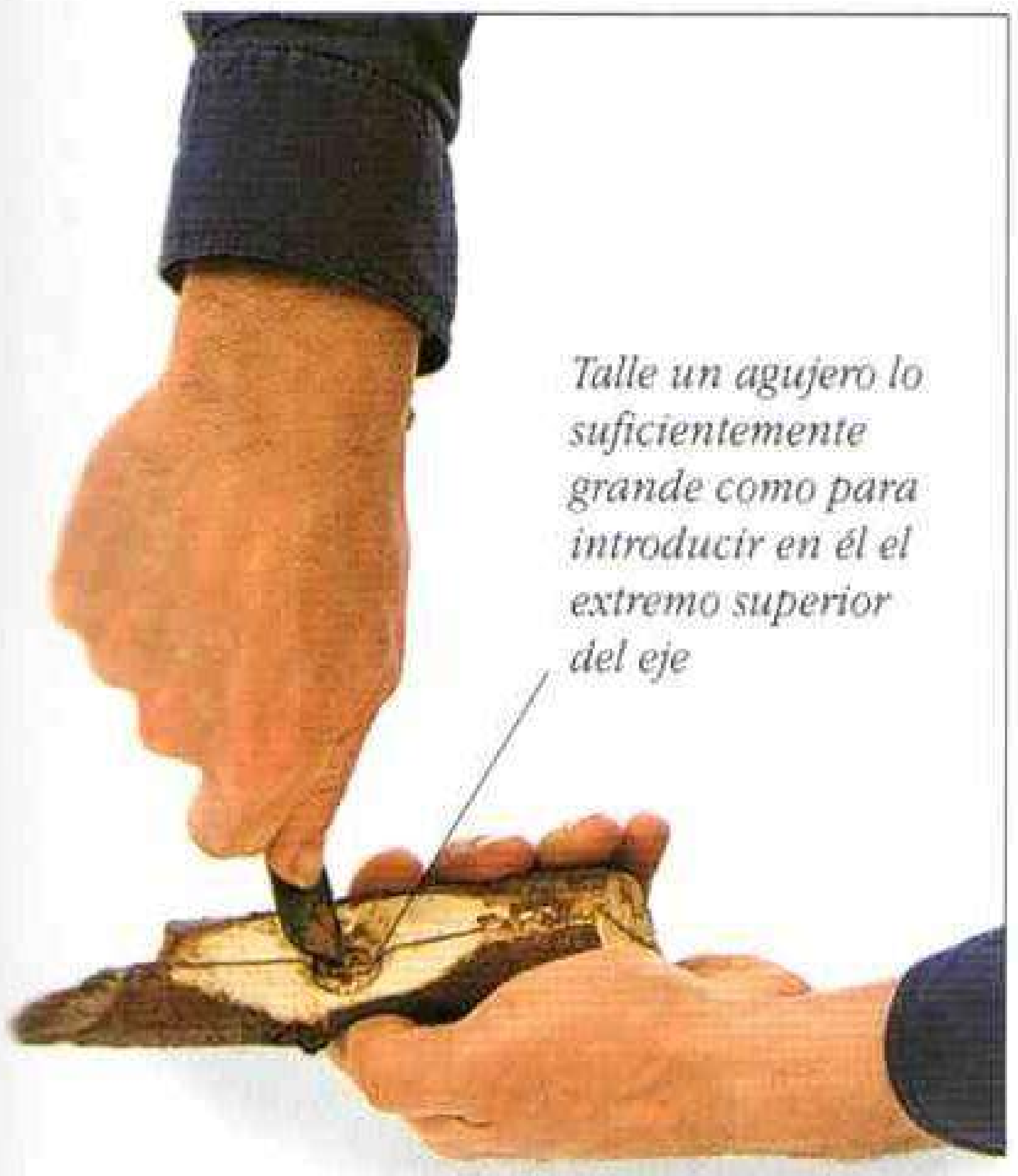

Talle un agujero lo suficientemente grande como para introducir en él el extremo superior del eje

4 Utilice un trozo de madera pesada como mango para sujetar y presionar el eje. Talle un agujero en su centro, para la parte superior del eje.

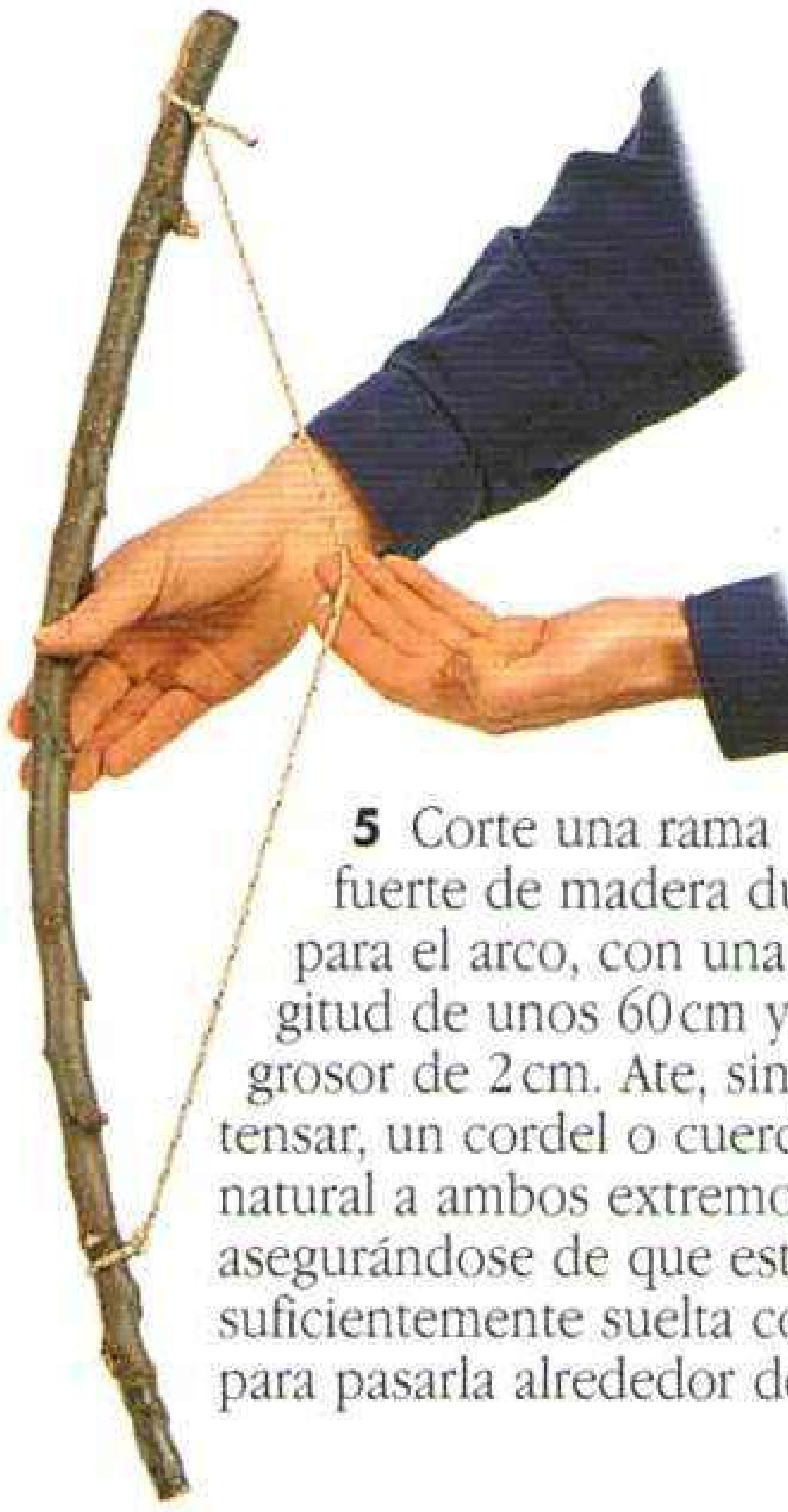

5 Corte una rama fuerte de madera dura para el arco, con una longitud de unos 60 cm y un grosor de 2 cm. Ate, sin tensar, un cordel o cuerda natural a ambos extremos, asegurándose de que está lo suficientemente suelta como para pasarla alrededor del eje.

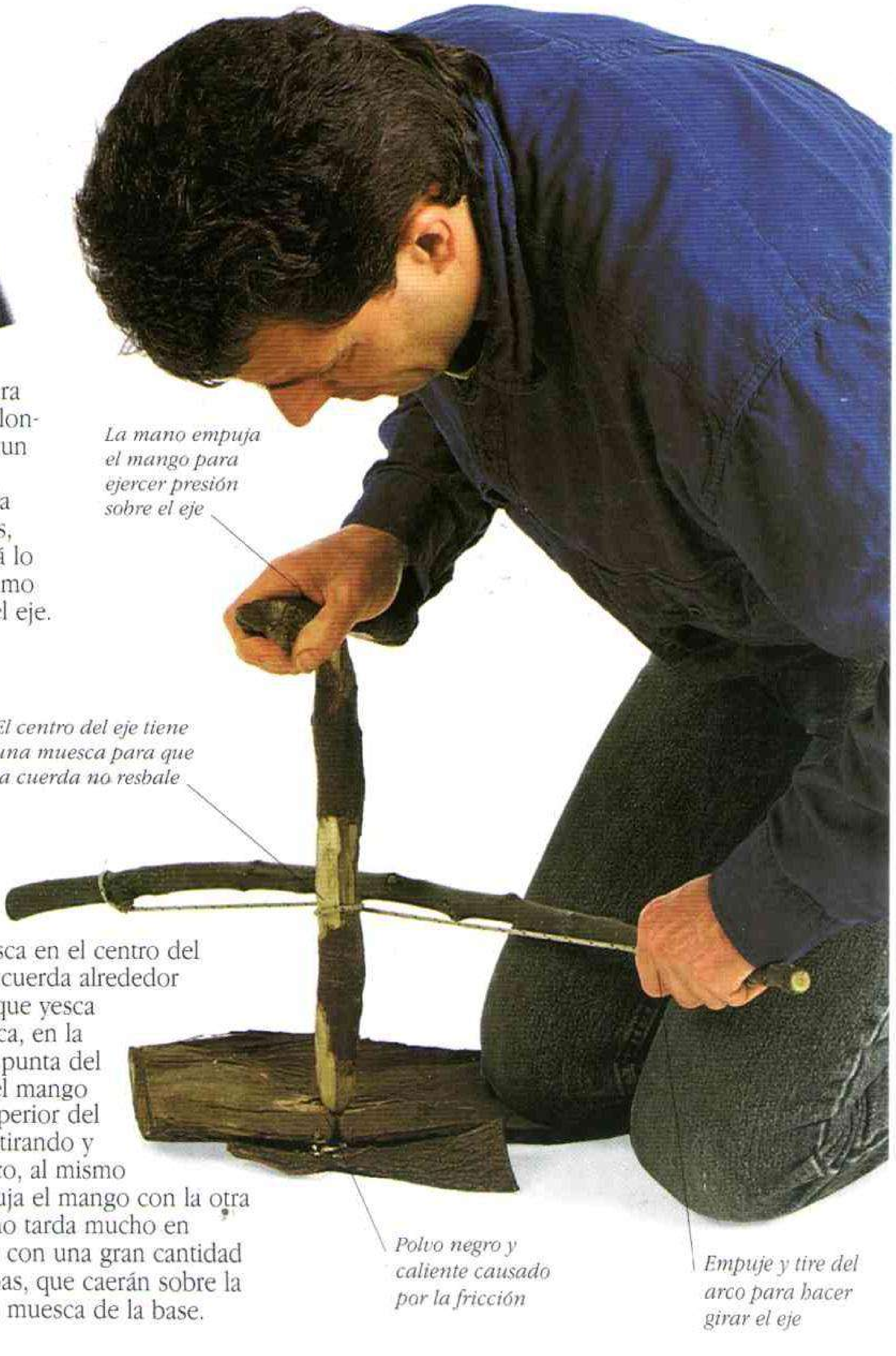

La mano empuja el mango para ejercer presión sobre el eje

El centro del eje tiene una muesca para que la cuerda no resbale

Polvo negro y caliente causado por la fricción

Empuje y tire del arco para hacer girar el eje

6 Talle una muesca en el centro del eje y envuelva la cuerda alrededor del mismo. Coloque yesca cerca de la muesca, en la base. Coloque la punta del eje en la base y el mango en el extremo superior del eje. Hágalo girar tirando y empujando el arco, al mismo tiempo que empuja el mango con la otra mano. El humo no tarda mucho en producirse, junto con una gran cantidad de ceniza y chispas, que caerán sobre la yesca gracias a la muesca de la base.

VARIANTE MANUAL

Un método más sencillo que el fuego de arco es la rotación con las manos de un palo terminado en punta sobre una base. Sin embargo, ya que las manos crean menor fricción que el arco, este método únicamente funciona eficientemente en los climas cálidos, donde la madera esté muy seca. Empuje el eje hacia la base al girarlo, y hágalo rotar en impulsos cortos.

Técnica manual. *El empleo del eje manual es más sencillo, aunque no tan efectivo como el arco.*

LEVANTAR EL CAMPAMENTO

LEVANTAR EL CAMPAMENTO debe convertirse en una rutina en la que todos trabajan para lograr ponerse en marcha a una hora previamente acordada. Si las condiciones del tiempo son extremas, esperar a una persona puede ser muy molesto, e incluso peligroso. El segundo jefe —quien debe cerrar la marcha del grupo *(véase página 136)*— debe comprobar el campamento antes de ponerse en marcha. Es importante limpiar el lugar de manera que parezca que nunca haya estado allí —a menos que esté en una situación de supervivencia y deba dejar alguna señal para los rescatadores. Desmonte su tienda o refugio en último lugar, para poder aprovecharla durante el mayor tiempo posible.

LIMPIAR EL CAMPAMENTO

Debe dejar el campamento como si nunca hubiera estado en él: llévese toda la basura, reemplace cualquier bloque de césped que hubiera levantado y obedezca las normas locales. De esta manera, no sólo se asegurará de que la naturaleza permanezca inalterada, sino que también contribuirá al bienestar de los futuros visitantes a la zona. En una situación de supervivencia, si necesita dejar algún mensaje para los posibles rescatadores, puede hacerlo sin alterar el medio ambiente *(véase página siguiente)*.

Tienda. *Las tiendas deben desmontarse en último lugar, para que, en caso de mal tiempo, pueda refugiarse hasta el último momento. Debe acordarse un tiempo para el cual todo debe estar recogido y listo para marchar. Una vez guardadas las tiendas, abandone el lugar inmediatamente.*

Letrina. *La letrina debe rellenarse y volverse a cubrir de hierba; debe marcar la fecha de partida para información de los futuros campistas. En las zonas en las cuales el medio era demasiado sensible para cavar una letrina, debe llevarse todos sus desechos en bolsas de plástico.*

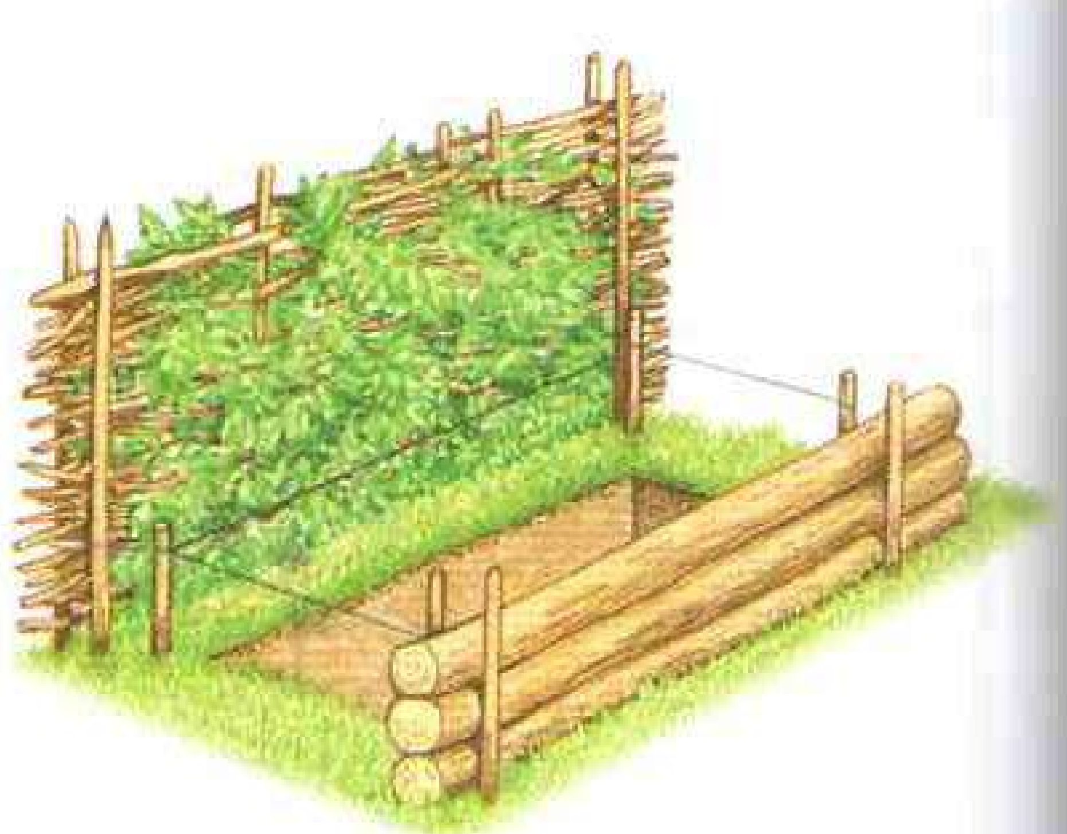

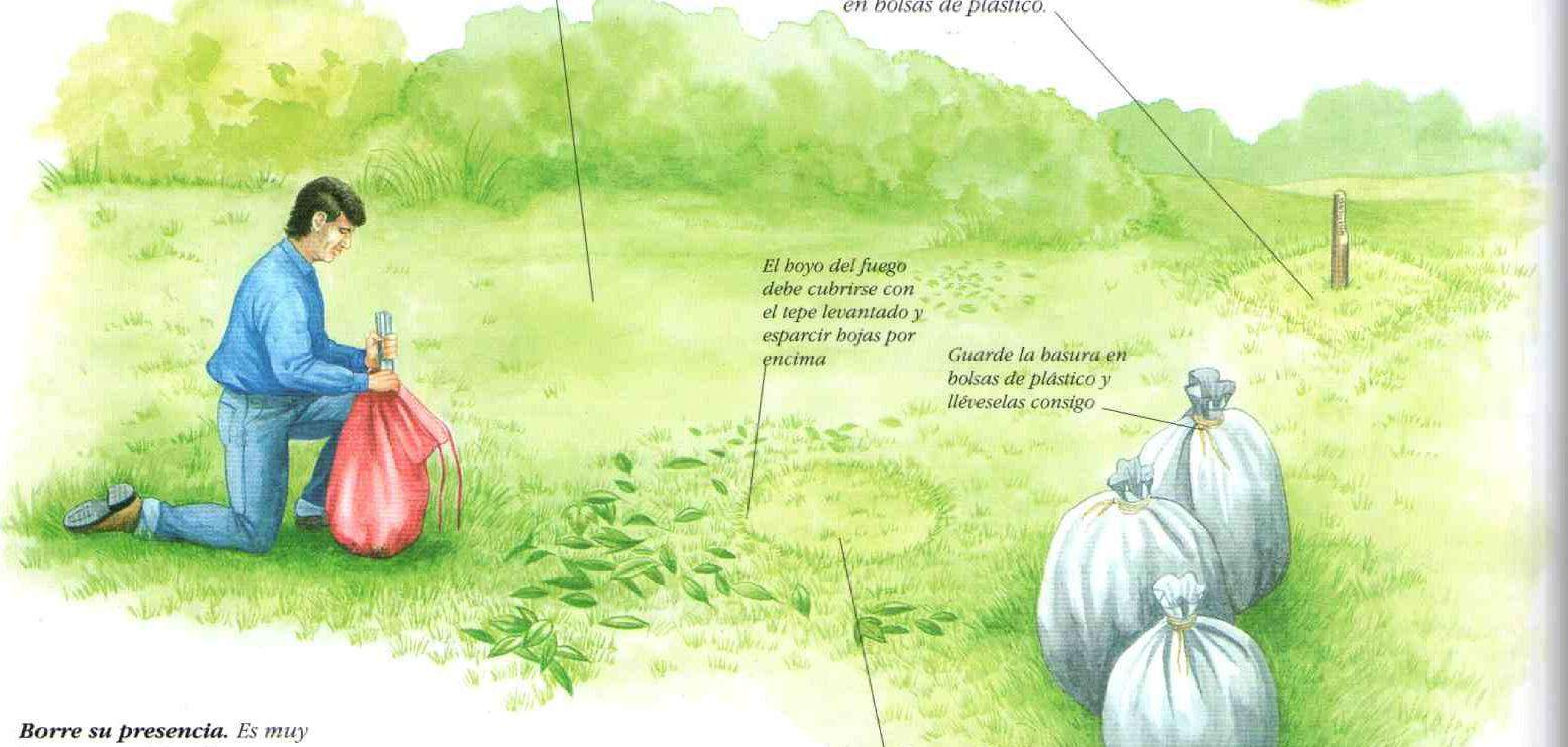

El hoyo del fuego debe cubrirse con el tepe levantado y esparcir hojas por encima

Guarde la basura en bolsas de plástico y lléveselas consigo

Borre su presencia. *Es muy importante dejar el campamento como si nunca hubiera estado allí. Desmonte el fuego y disfrace el hoyo del fuego, llévese todos sus desechos, basura y residuos de comida y rellene la letrina. Si siempre actúa de esta manera al abandonar el campamento, no alterará la naturaleza en beneficio de la vida silvestre y de los futuros campistas.*

Fuego. *El fuego debe quedar completamente apagado. Las cenizas deben esparcirse y enterrarse en el suelo. Cualquier residuo sin quemar debe recogerse y guardarse para eliminarlo posteriormente.*

LIMPIAR EL FUEGO

Debe asegurarse de que el fuego esté totalmente apagado en el momento de abandonar el campamento. Incluso si ha cubierto el hoyo del fuego, éste puede contener rescoldos ardientes que podrían originar un incendio forestal. Utilice la menor cantidad de leña posible para preparar el desayuno el día de su marcha, y asegúrese de que toda la leña queda consumida.

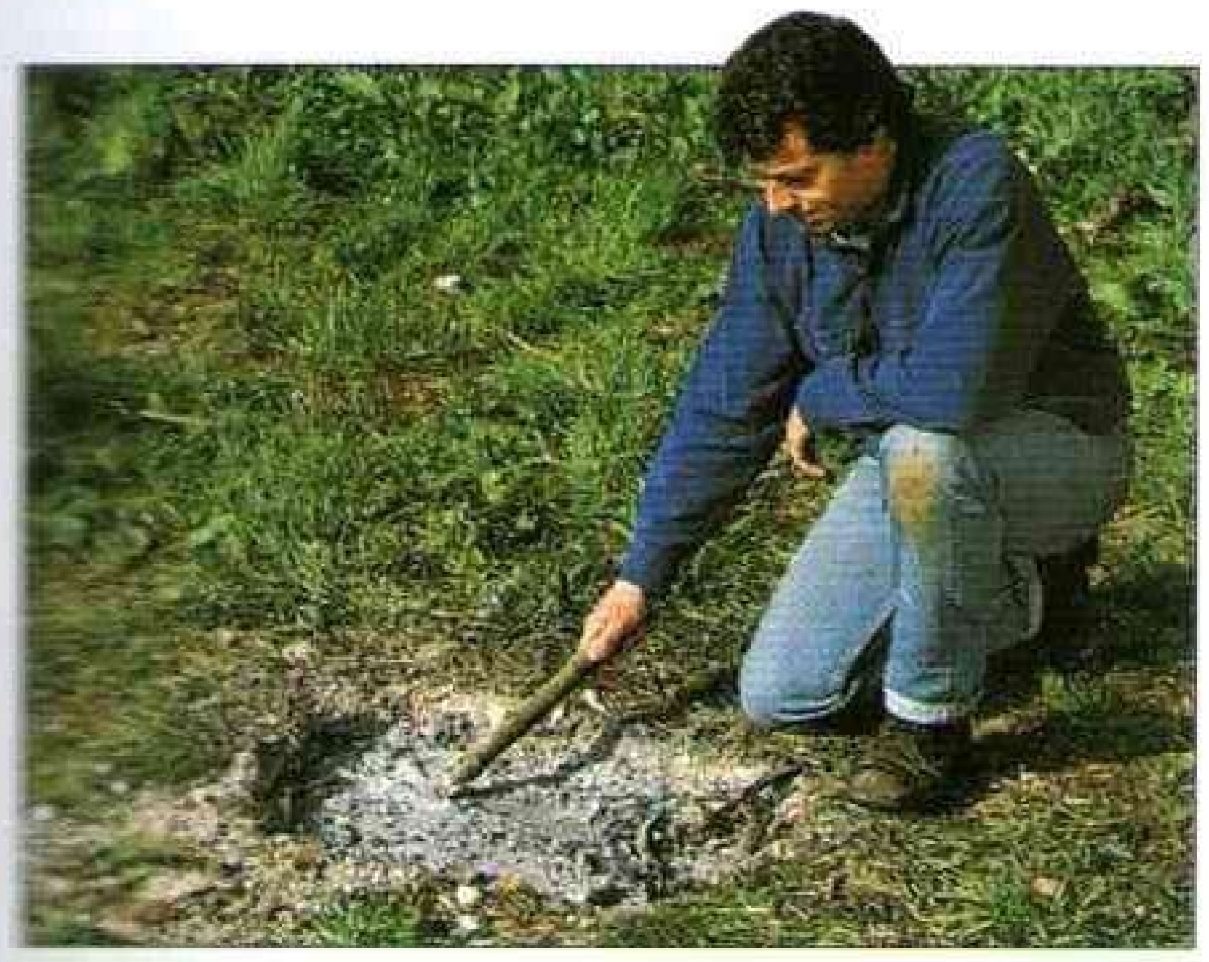

1 Una vez el fuego deja de arder, remueva los restos hacia el centro hasta que queden reducidos a cenizas. Cuando las cenizas estén frías, espárzalas y entiérrelas en el suelo para dispersarlas.

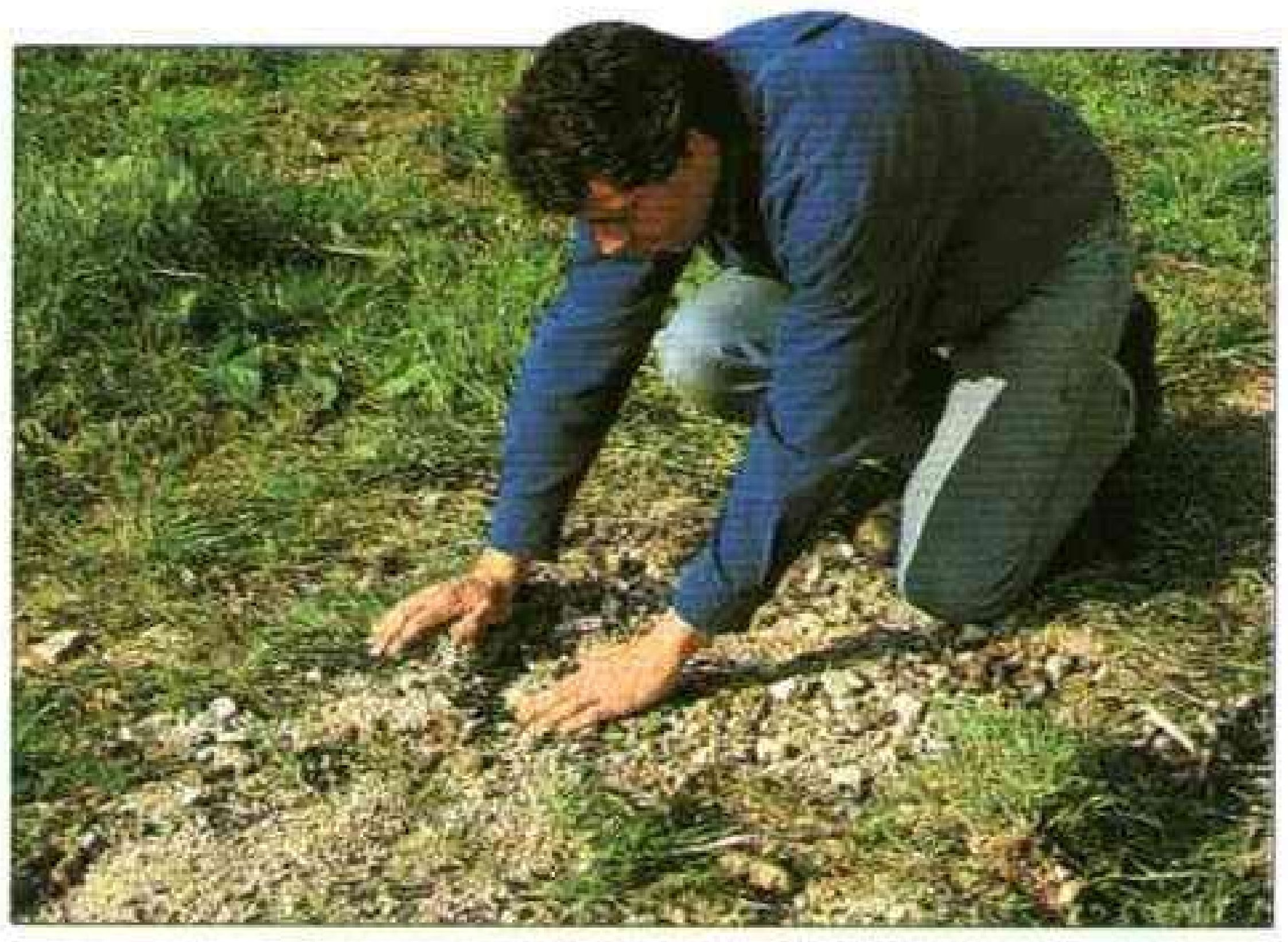

2 Asegurándose de que no quedan cenizas en la superficie que puedan matar la hierba, tape el hoyo del fuego con tierra. Alísela con las manos. Después coloque el bloque de césped que había sacado previamente al construir el fuego. Si empleó rocas, dispérselas y camúflelas con mantillo.

3 Rellene los bordes del bloque de césped con tierra y hierbas para que no queden marcas de la unión. Esparza hojas y hierba encima del lugar para que se asemeje al área circundante. Guarde cualquier desperdicio en una bolsa de plástico y llévela consigo hasta que pueda quemarla o deshacerse de ella.

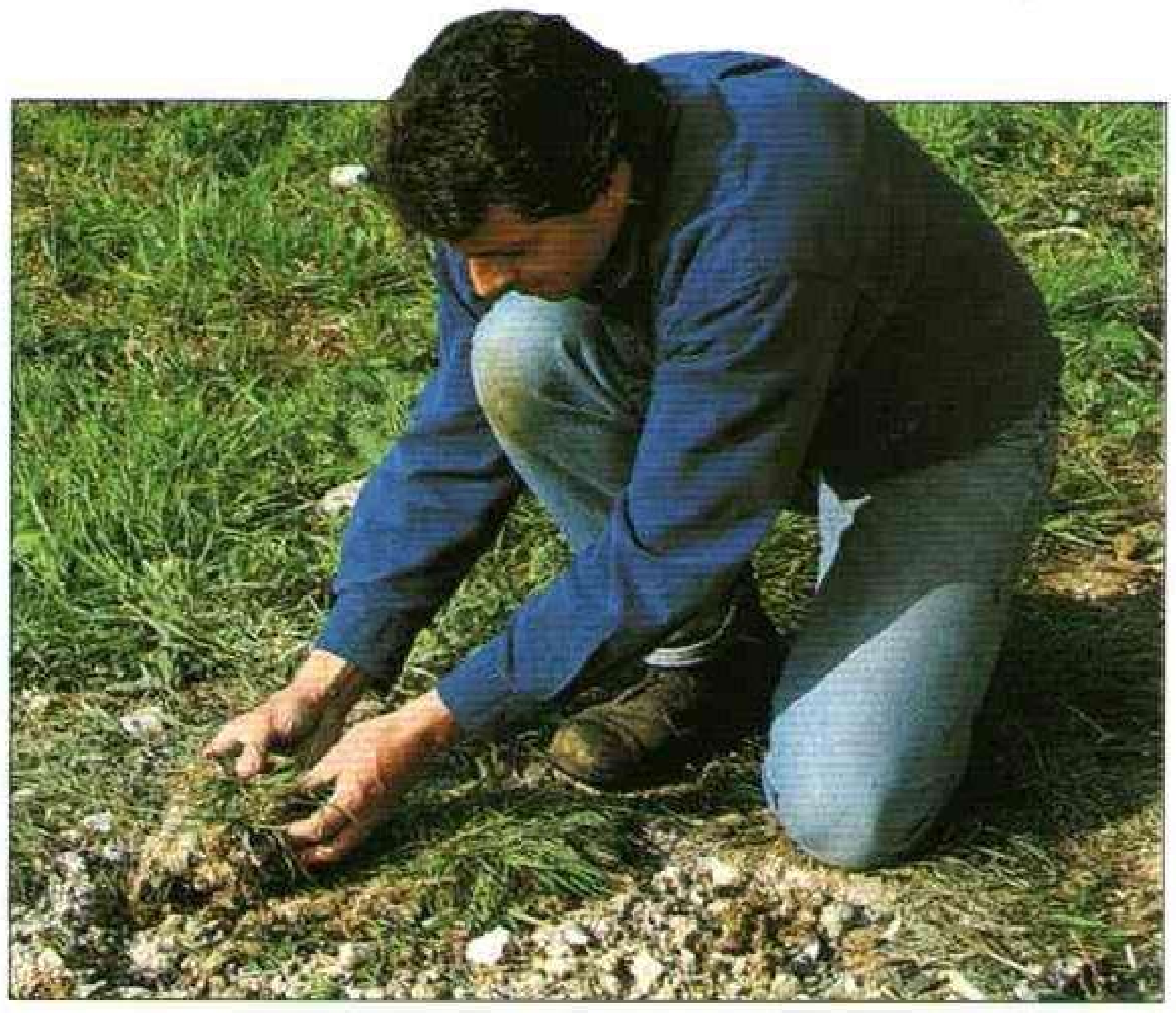

NO DEJE HUELLAS

Muchas áreas naturales quedan dañadas cuando los visitantes dejan basuras que estropean el paisaje y que pueden llegar a matar a los animales; alteran la vegetación y los árboles; causan incendios forestales, y contaminan los ríos. Debe dejar todas las áreas naturales como si nunca hubiera estado en ellas, poniendo especial cuidado en su campamento. En algunas áreas naturales vigiladas, está prohibido construir fuegos o cavar letrinas; respete estas reglas en beneficio del medio ambiente y de los futuros campistas. Si deja residuos de comida, los animales pueden convertirse en una molestia para los campistas, llegando incluso a poner en peligro no sólo la vida de los animales, sino también la de las personas.

DEJAR MENSAJES A LOS RESCATADORES

Cualquier signo fuera de lo común será avistado por aquéllos que le sigan. Indique la dirección de su marcha, así como el tiempo y la fecha, en cualquier señal que deje a los rescatadores. Estos signos permitirán trazar su ruta. Coloque los mensajes lo más alto posible, a salvo de los animales. También puede levantar un montón de piedras de modo que sea visible y colocar en él una nota.

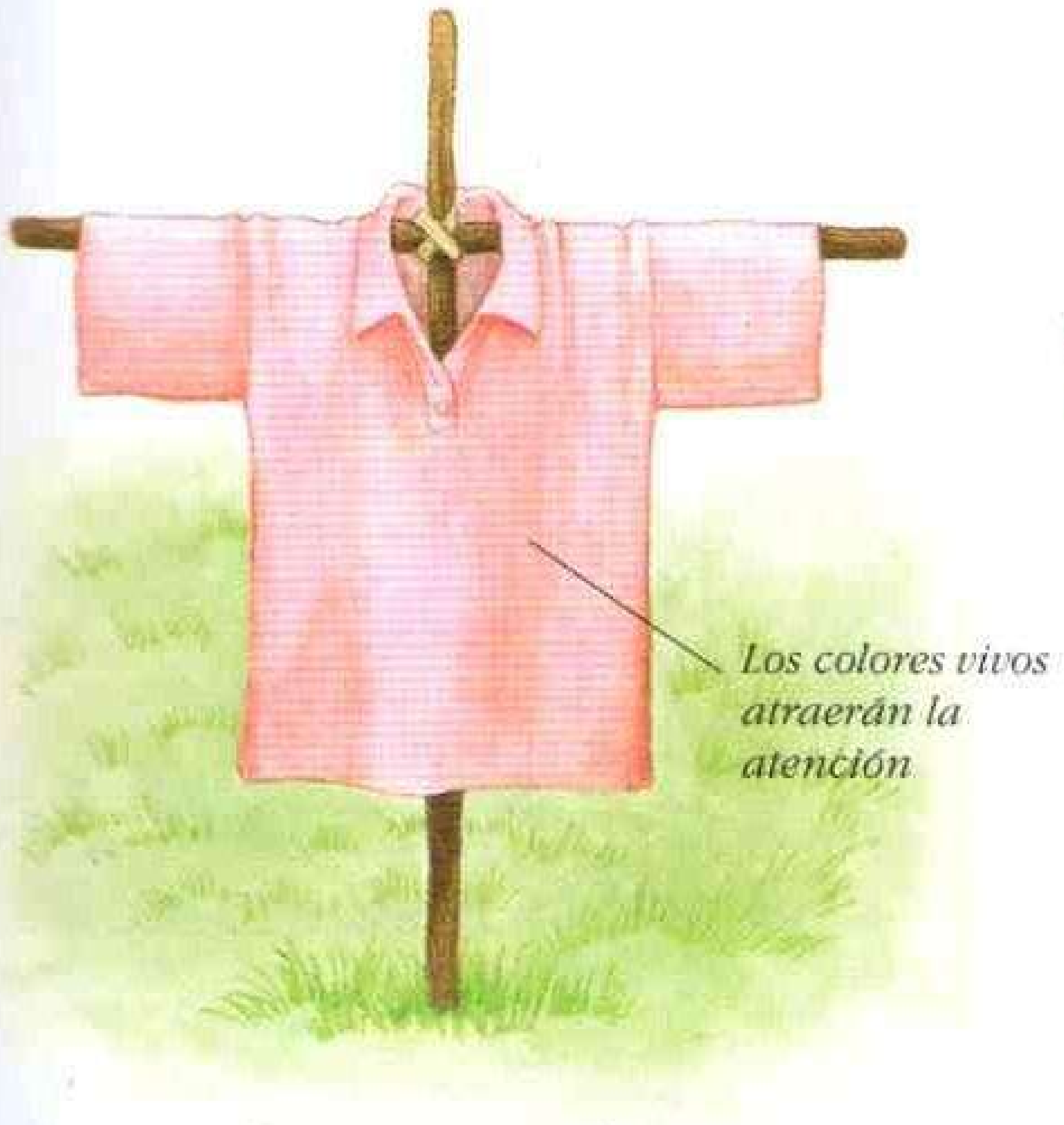
Los colores vivos atraerán la atención

Bandera. *Ate dos ramas para formar una cruz y cuelgue de ella una camiseta de un color vivo. Coloque la cruz en un claro. Si utiliza dos o más de estas señales, dispóngalas de tal manera que apunten hacia la dirección que haya tomado.*

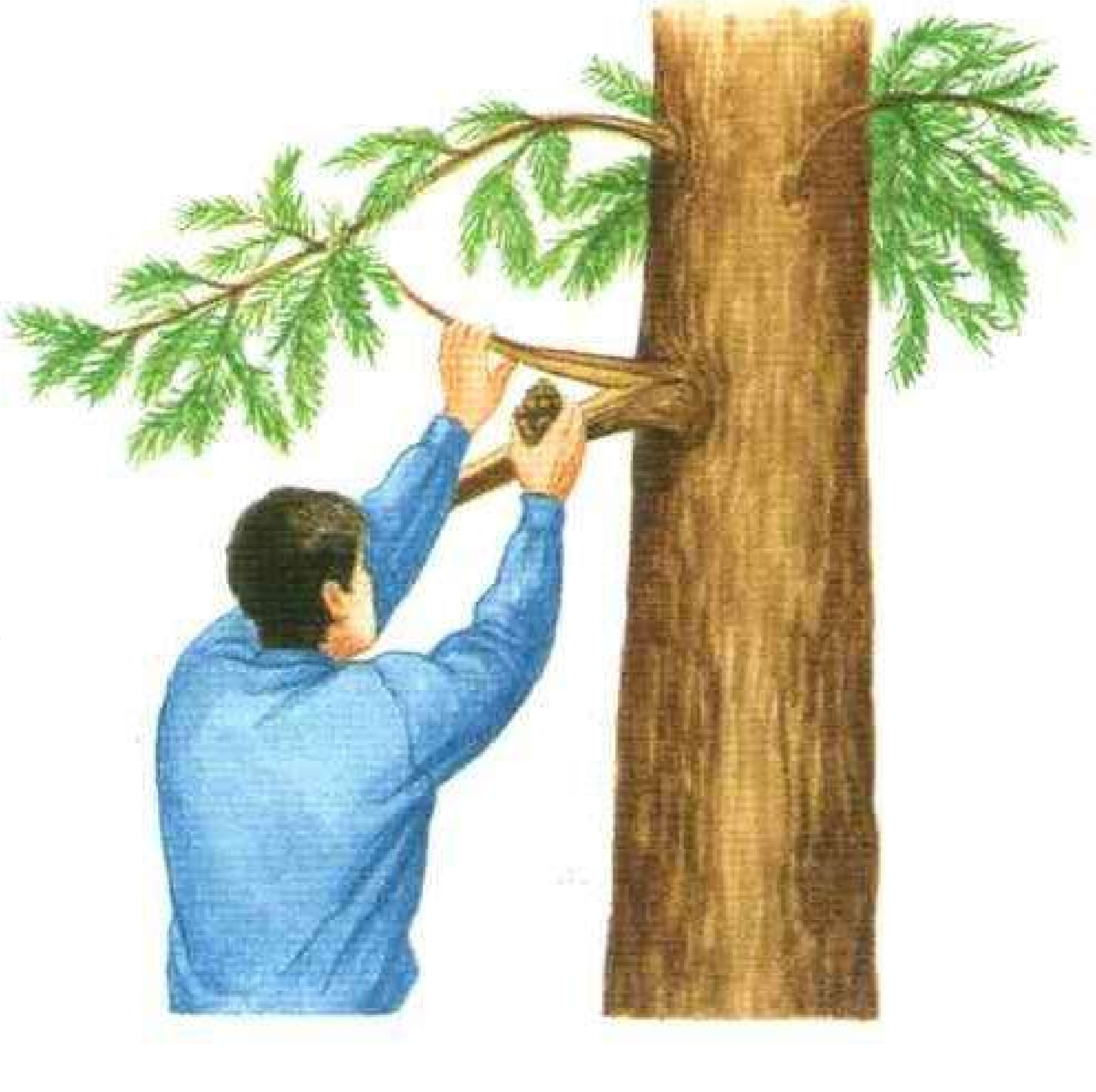

Piña. *Coloque una piña en una rama dividida en el lado del camino que haya tomado. Cualquier suceso que no ocurra naturalmente será avistado por cualquiera que le busque. Tenga cuidado de no emplear algo que los animales puedan comerse.*

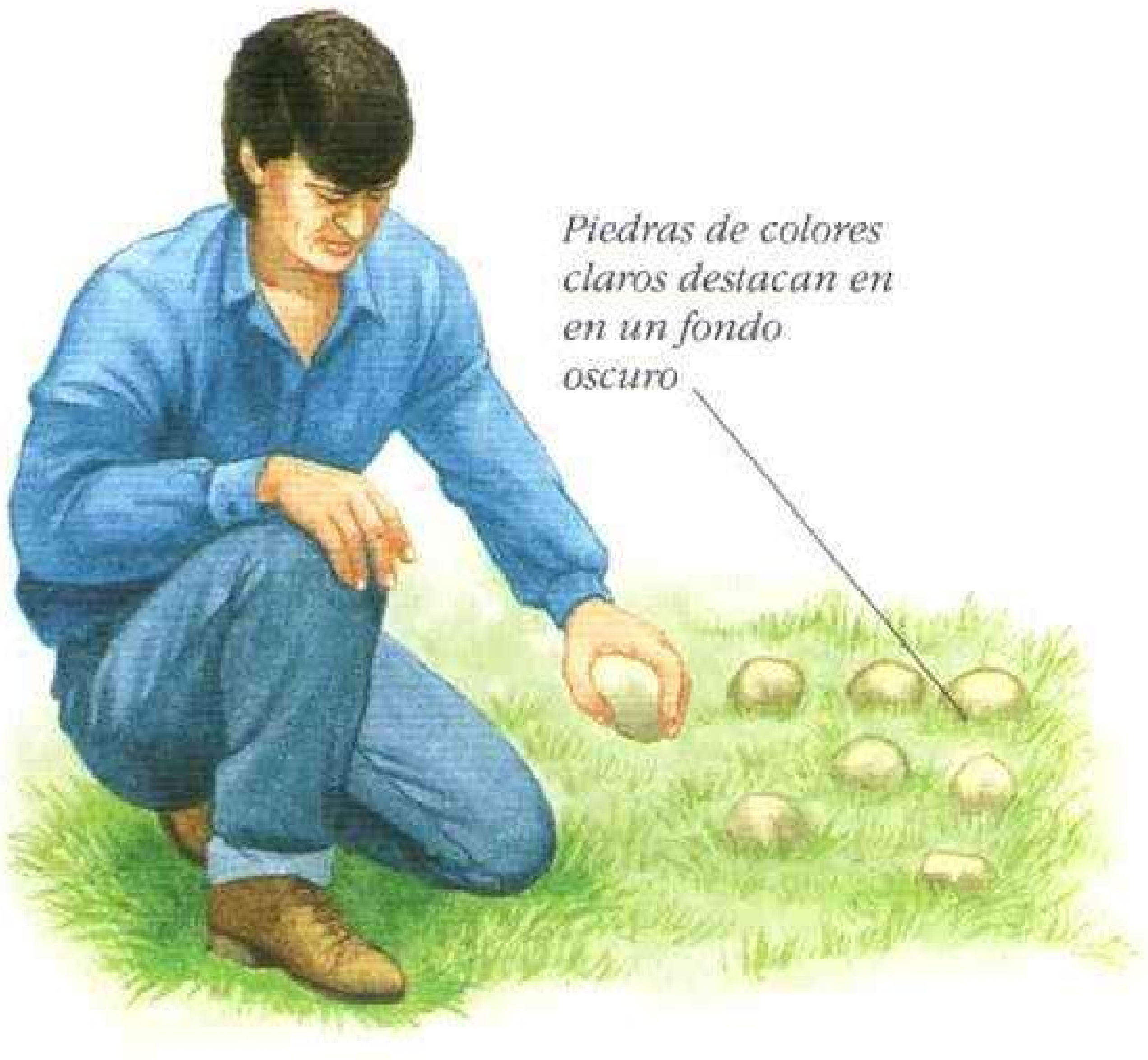
Piedras de colores claros destacan en en un fondo oscuro

Flechas. *Colocadas en un claro, las piedras en forma de flecha indicarán la dirección de su viaje. La flecha también será visible desde el aire. Haga flechas nuevas a intervalos, y siempre en los puntos en los que el camino se divida.*

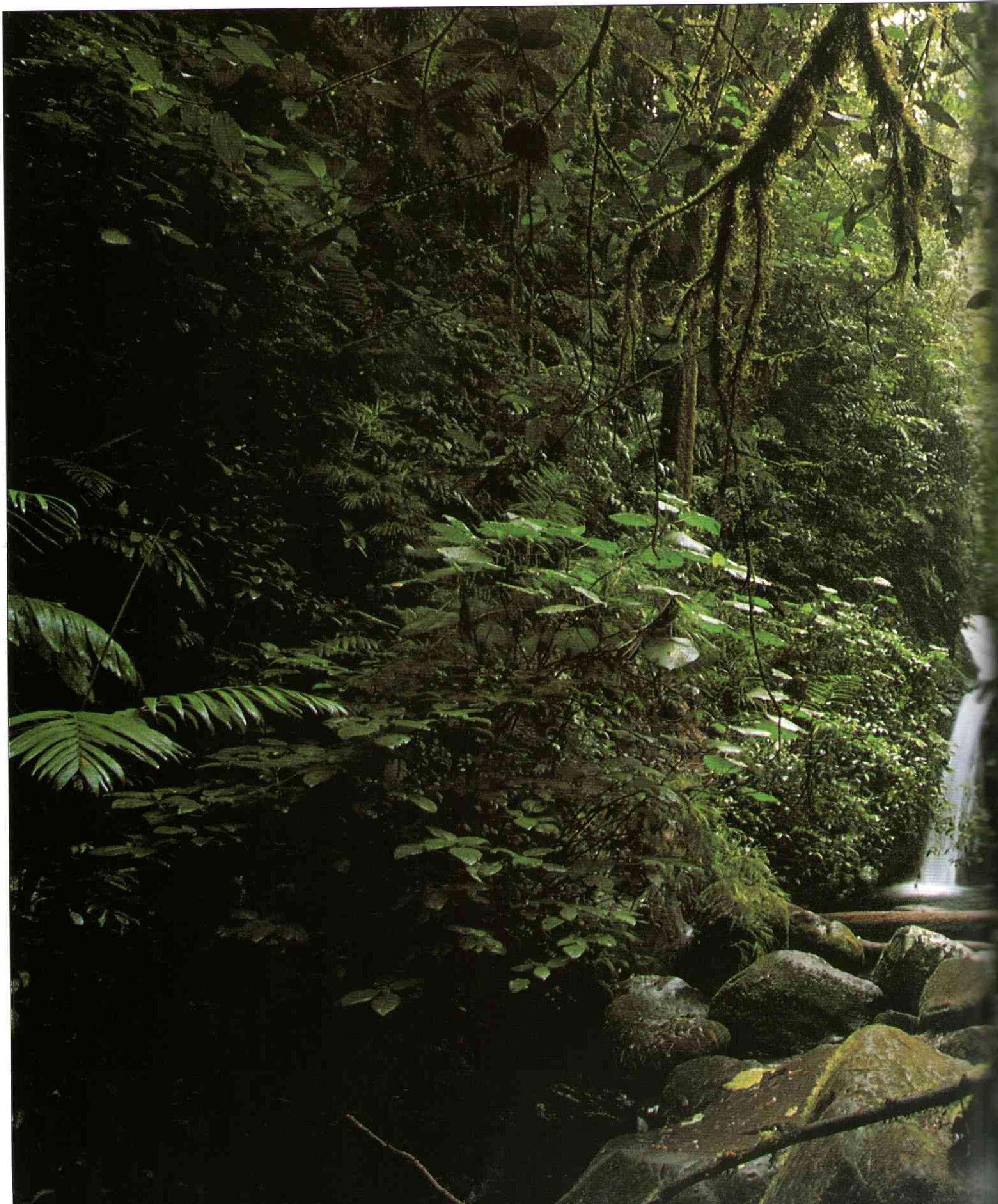

CAPÍTULO CUATRO

EN BUSCA DE AGUA

TODOS LOS SERES vivos están compuestos en su mayor parte de agua, sin la cual mueren. En una situación de supervivencia, después de encontrar refugio, su prioridad principal debe ser la localización de una fuente adecuada de agua limpia. Ni siquiera la comida es tan importante. El agua es la base de cualquier asentamiento humano y organización social. Gracias a los grandes ríos, las fértiles tierras permitieron la producción de excedentes de alimentos en las antiguas civilizaciones. Este hecho permitió el desarrollo del comercio y de la cultura, dando lugar a que la gente se apartara de la línea de supervivencia cosecha a cosecha. El agua también ha traído la muerte a las civilizaciones humanas, extendiendo las epidemias letales de cólera y tifus, o destruyendo ciudades enteras cuando un río se secaba, o bien cambiaba de curso y se desbordaba. Hoy en día, el hombre urbano no valora el agua limpia y fresca. Sin embargo, ésta es una de las grandes falacias de la vida occidental, que pronto se esfumará cuando se agoten nuestras fuentes de agua subterránea y se sequen nuestros ríos. Para sobrevivir debe aprender a valorar la vida y el agua.

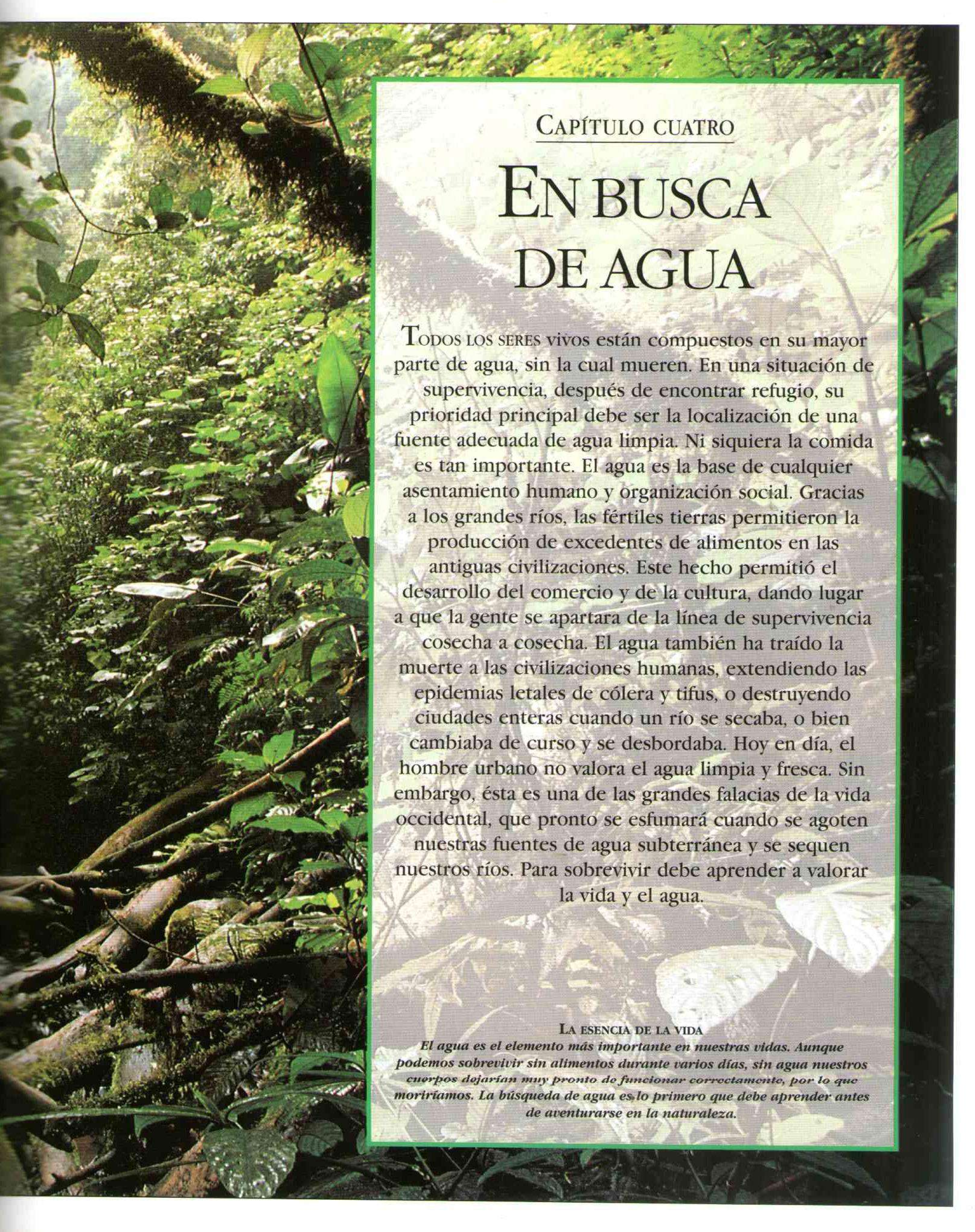

LA ESENCIA DE LA VIDA
El agua es el elemento más importante en nuestras vidas. Aunque podemos sobrevivir sin alimentos durante varios días, sin agua nuestros cuerpos dejarían muy pronto de funcionar correctamente, por lo que moriríamos. La búsqueda de agua es lo primero que debe aprender antes de aventurarse en la naturaleza.

LA IMPORTANCIA DEL AGUA

EL HOMBRE DEBE beber un mínimo de tres litros de agua cada día. Así pues, una fuente constante de agua resulta vital para la vida. Viajar está, por lo general, determinado por la posibilidad de agua, ya que incluso los vehículos no pueden transportar más que la necesaria para unos cuantos días. Si viaja a pie, debe llevar suficiente agua al menos para un día, así como una reserva de emergencia. Resulta vital llegar a una fuente de agua al menos una vez al día. Saber dónde encontrar el agua y cómo obtenerla es, pues, esencial.

DÓNDE BUSCAR AGUA

Cuando la lluvia cae en las montañas, se recoge en pequeños riachuelos que se juntan formando torrentes mayores. En cuanto el agua alcanza las tierras bajas, estos torrentes se han convertido en grandes ríos, más lentos, que finalmente desembocan en el mar. El agua potable puede encontrarse en cualquier punto de este recorrido. Cualquier agua recogida en la naturaleza debe purificarse antes de beberse, para evitar posibles infecciones *(véase página 75).*

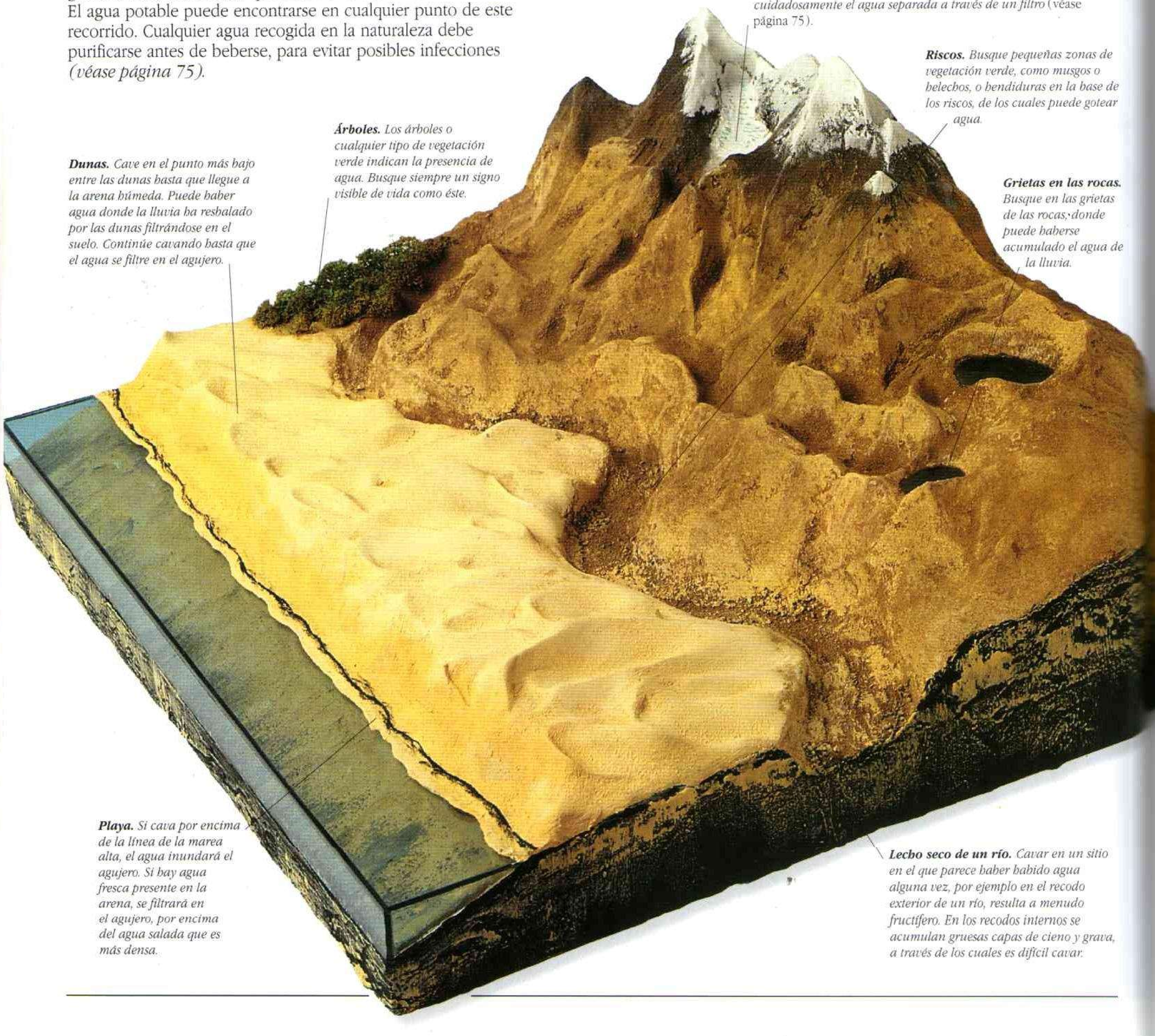

Glaciar. *Los glaciares a menudo generan vigorosos torrentes de agua de deshielos. Sin embargo, el agua de los glaciares contiene una gran cantidad de polvo de roca abrasivo, que puede causar diarrea. Antes de beber el agua, debe dejarla reposar una noche para que las partículas de mineral se asienten; vierta entonces cuidadosamente el agua separada a través de un filtro* (véase página 75).

Riscos. *Busque pequeñas zonas de vegetación verde, como musgos o helechos, o hendiduras en la base de los riscos, de los cuales puede gotear agua.*

Árboles. *Los árboles o cualquier tipo de vegetación verde indican la presencia de agua. Busque siempre un signo visible de vida como éste.*

Dunas. *Cave en el punto más bajo entre las dunas hasta que llegue a la arena húmeda. Puede haber agua donde la lluvia ha resbalado por las dunas filtrándose en el suelo. Continúe cavando hasta que el agua se filtre en el agujero.*

Grietas en las rocas. *Busque en las grietas de las rocas, donde puede haberse acumulado el agua de la lluvia.*

Playa. *Si cava por encima de la línea de la marea alta, el agua inundará el agujero. Si hay agua fresca presente en la arena, se filtrará en el agujero, por encima del agua salada que es más densa.*

Lecho seco de un río. *Cavar en un sitio en el que parece haber habido agua alguna vez, por ejemplo en el recodo exterior de un río, resulta a menudo fructífero. En los recodos internos se acumulan gruesas capas de cieno y grava, a través de los cuales es difícil cavar.*

INDICIOS DE AGUA

La presencia de agua está a menudo indicada por indicios de vida como vegetación verde, huellas animales o viviendas humanas. Sin embargo, hay otros indicios naturales que muestran la existencia de agua oculta. En un terreno aparentemente sin vida, busque signos de vida, aunque tenga en cuenta que algunas aves, como el buitre, pueden viajar a grandes distancias de sus fuentes de agua.

Árboles y otro tipo de vegetación. *Todos los árboles necesitan agua para vivir, aunque algunos tienen raíces muy largas que se extienden bajo el suelo en busca de agua. Las palmeras crecen, por lo general, donde hay algún tipo de agua cerca de la superficie, y algunas variedades almacenan agua en sus troncos o raíces* (véase página 72).

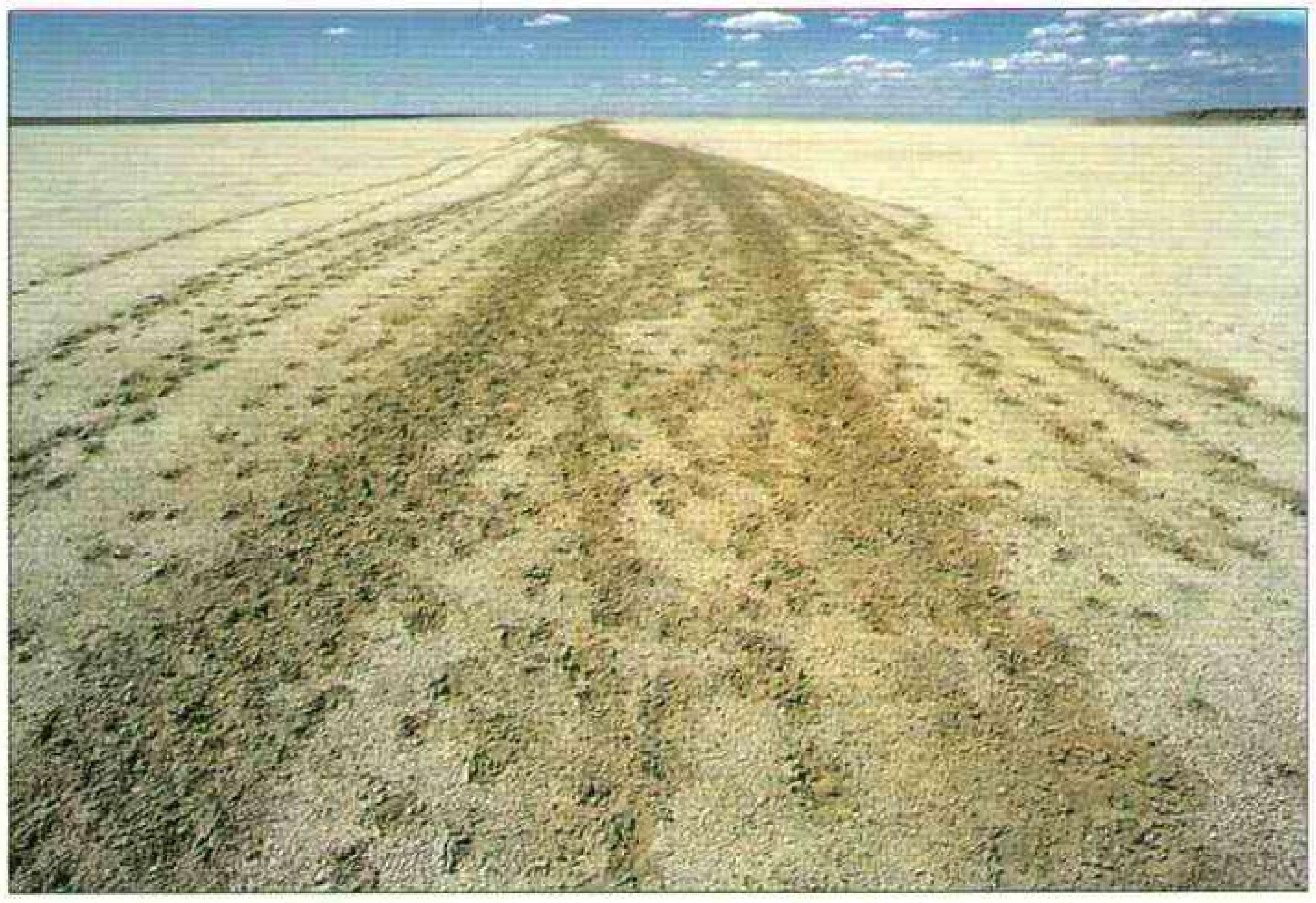

Huellas animales. *Los rumiantes necesitan beber por lo menos dos veces al día, por lo que se dirigen desde o hacia el agua. Busque el punto en el que convergen las huellas de los animales: puede ser un indicativo de una fuente de agua. Tenga cuidado si se acerca a una charca: cerca puede haber depredadores.*

Abejas y moscas. *Las abejas no se alejan, generalmente, más de 5 km de sus nidos y necesitan una reserva constante de agua. Observe la dirección en la que vuelan cuando abandonan el nido. Las moscas aún permanecen más cerca del agua —unos 100 m aproximadamente. Sin embargo, en el desierto parecen venir de ningún sitio, independientemente de donde esté la fuente más cercana de agua.*

Hormigas en movimiento. *Las hormigas dependen de una reserva constante de agua. Si observa algunas de ellas subiendo a un árbol, es muy probable que se dirijan hacia una reserva de agua de lluvia.*

LA DESHIDRATACIÓN Y LAS NECESIDADES DE AGUA DEL CUERPO

El agua es un constituyente primordial del cuerpo humano. No disponemos de ningún medio para almacenar agua en nuestros cuerpos, como hacemos con las grasas. Necesitamos una provisión constante de agua para controlar la temperatura corporal (transpiración), la eliminación de productos de desecho a través de los riñones y el funcionamiento del sistema nervioso. Sin alimentos podemos sobrevivir con nuestras reservas corporales durante unas tres semanas, pero si el agua perdida no se repone en cinco días, la muerte es inminente.

EFECTOS DE LA PÉRDIDA DE AGUA

Pérdida entre 1-5 %	Pérdida entre 6-10 %	Pérdida entre 11-12 %
Sed	Dolor de cabeza	Delirio
Malestar	Mareo	Lengua hinchada
Letargo	Sequedad en la boca	Contracciones
Impaciencia	Dificultad para respirar	Sordera
Falta de apetito	Hormigueo en las extremidades	Visión borrosa
Enrojecimiento de la piel	Tono azulado de la piel	Falta de sensibilidad en la piel
Aumento del pulso	Dificultad en el habla	La piel comienza a arrugarse
Náuseas	Incapacidad para caminar	Incapacidad de tragar
Debilidad	Visión borrosa	Muerte

Necesidades de agua. *Cerca de un 75 % del cuerpo humano es agua, imprescindible para la vida. Cuando el cuerpo la pierde a través de la transpiración y de la orina, debe ser reemplazada. Si no es así, se desencadenan diversos problemas de salud que se agravan hasta terminar en la muerte.*

PREVENIR LA PÉRDIDA DE AGUA

Incluso mientras descansa en la sombra, una persona normal pierde más de un litro de agua cada día debido a la respiración y a la orina. El ejercicio aumenta la pérdida de agua a través de la transpiración, especialmente en climas templados, así que, si su reserva de agua es limitada, debe descansar de día y trabajar únicamente de noche. Intente mantenerse fresco y a la sombra. Respire a través de la nariz para reducir la pérdida de agua, y no fume. No se recueste sobre el suelo caliente. Limítese a comer lo indispensable para mantenerse vivo. Omita los alimentos grasos en su dieta, y no beba alcohol —ambos requieren una gran cantidad de agua para ser digeridos. No espere a haber agotado su reserva de agua antes de comenzar a buscar más.

RECOGER AGUA

EN UNA SITUACIÓN de supervivencia, después de encontrar refugio, la búsqueda de agua debe convertirse en su siguiente prioridad, ya que sin agua no sobreviviría más que unos cuantos días. Es posible que no siempre tenga la fortuna de acampar al lado de una fuente de agua de confianza *(véase página 40)*. Sin embargo, puede sobrevivir en una situación de emergencia recogiendo el agua de lluvia o del rocío. El agua recogida de la atmósfera, de las plantas o del suelo, tiene una gran ventaja sobre las otras fuentes de agua naturales: su pureza *(véase página 74)*. El amanecer es la mejor hora para recoger agua.

RECOGER AGUA DE LA LLUVIA

Mientras permanezca en el campamento, ponga varios recipientes para recoger el agua de la lluvia que pudiera caer. También debe recoger el agua que fluye por el tejado de su refugio, improvisando un canal para encauzar el agua hacia un recipiente *(véase página 53)* o hacia un estanque *(véase página 77)*. Incluso si tiene un río o arroyo cerca, recoja el agua de la lluvia, que será pura y le permitirá ahorrar en medios de depuración químicos o le evitarán tener que hervirla *(véase página 75)*. Para recoger el agua de lluvia, extienda una tela impermeable tensa sobre un área extensa, preferiblemente en una pendiente. Sujete las esquinas al suelo con palos y recoja el agua en un recipiente.

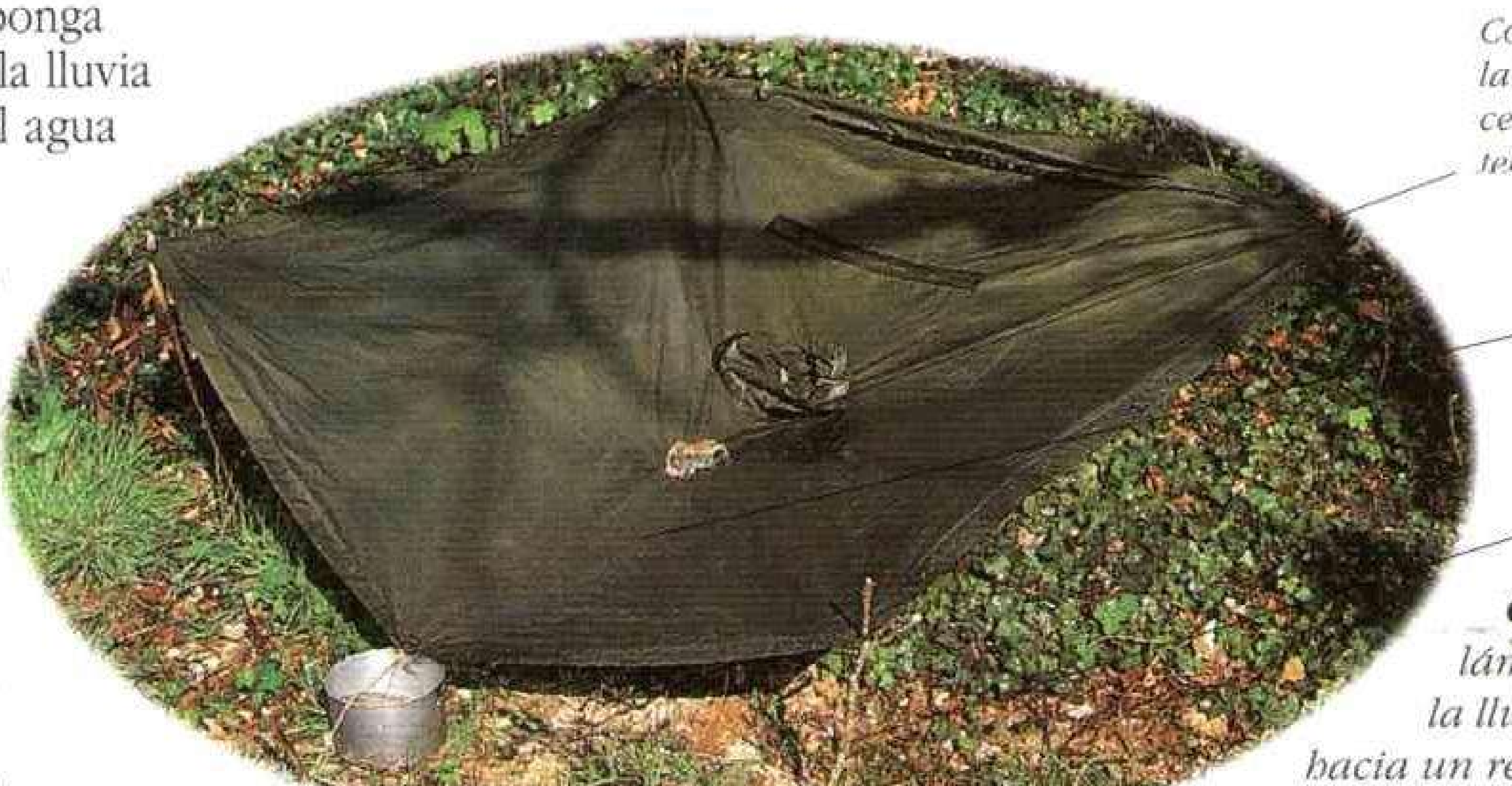

Con el peso de la piedra, el centro de la tela se hunde

El agua fluye por los lados hacia el recipiente

Los palos sujetan tensa la tela

Colector de lluvia. *Cualquier lámina impermeable recogerá la lluvia y la canalizará hacia un recipiente. Si la lluvia es intensa, vigile continuamente para reemplazar rápidamente el recipiente cuando se llene.*

RECOGER AGUA DEL ROCÍO

Cuando el aire se enfría por la noche, el vapor de agua que contiene se condensa en rocío sobre el terreno bajo, la vegetación y los vehículos. El agua se evapora rápidamente al salir el sol. Muchas plantas, insectos y animales dependen del rocío para sobrevivir; los hombres también pueden utilizar esta fuente natural de agua.

1 Puede recoger el rocío empapando un paño en la hierba larga y húmeda. La mejor hora para recoger el rocío es el amanecer, ya que después de la salida del sol se evapora rápidamente. Muchos pueblos indígenas de las tierras áridas utilizan este método regularmente para obtener agua pura.

2 Cuando el trapo esté empapado, exprímalo en un recipiente. Repita la operación hasta que tenga suficiente agua.

CAVAR EN BUSCA DE AGUA

A menudo el agua se filtra en un agujero excavado en un terreno pantanoso. Cave un agujero de unos 30 cm de profundidad. Recoja el agua que se acumula en el agujero. El agua estará ligeramente cenagosa las primeras veces que se llene el agujero, pero posteriormente subirá agua limpia que podrá depurar y beber.

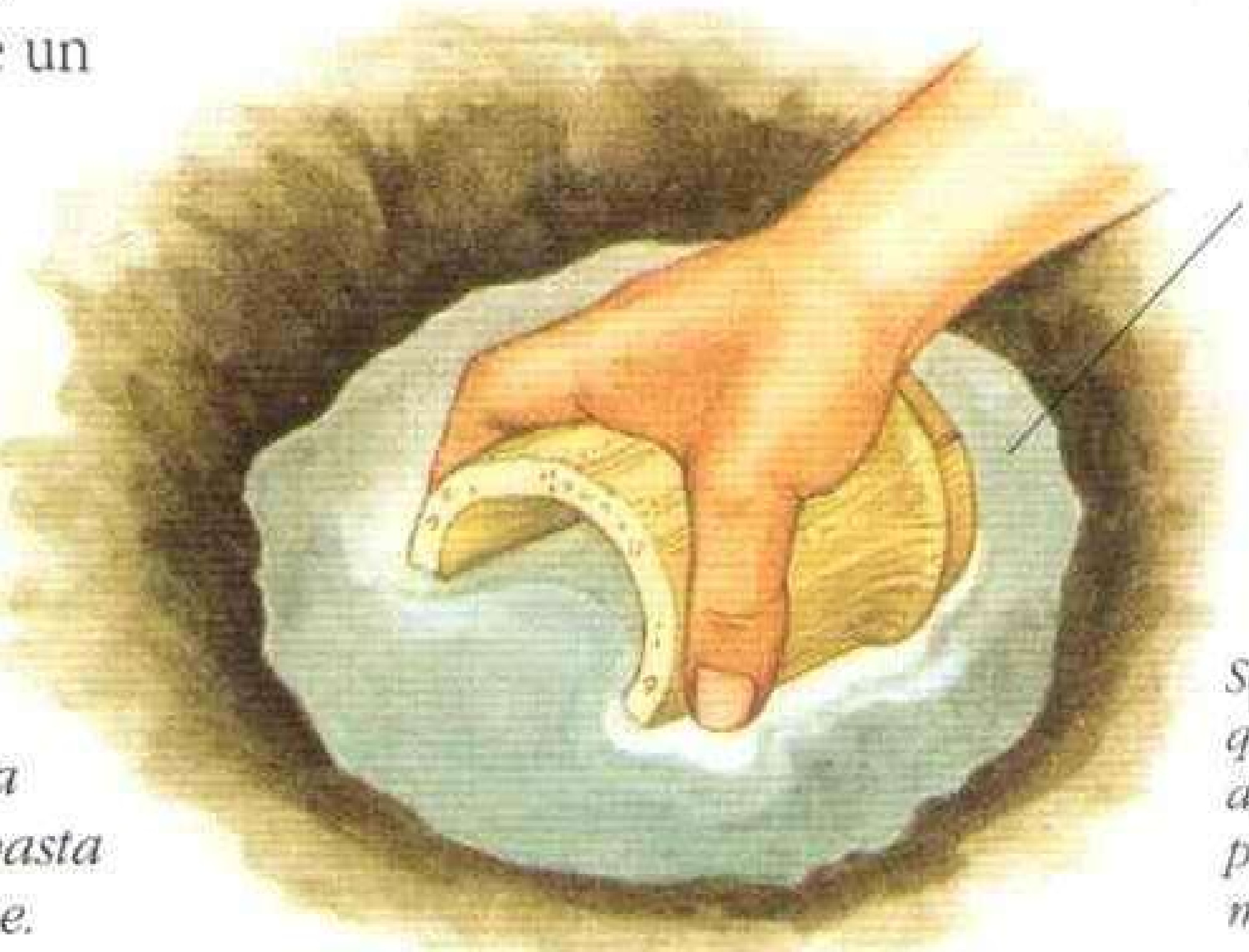

El agua que se filtra en el agujero proviene del suelo circundante

Agua cenagosa. *Saque el agua lodosa de un agujero en terreno pantanoso hasta que el agua fresca aflore a la superficie.*

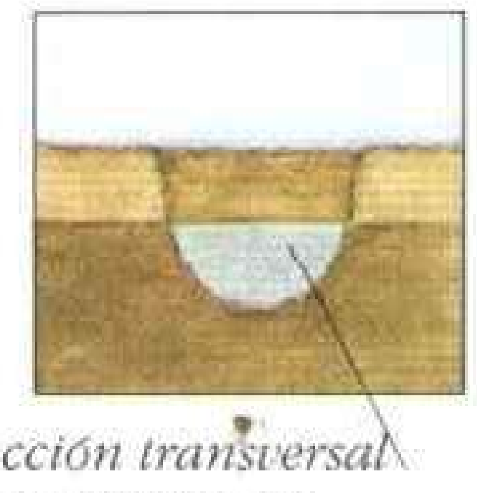

Sección transversal que muestra un agujero con una profundidad mayor que el nivel en el cual el agua inunda la tierra

ADVERTENCIA

Antes de cavar un agujero de este tipo estudie el terreno. Nunca cave donde el suelo tenga un olor muy potente, o tenga limo verde en la superficie —cualquier agua en este lugar estará probablemente contaminada. No recoja agua donde encuentre animales muertos. Depure siempre el agua antes de beberla *(véase página 74)*.

CÓMO OBTENER AGUA DE LA VEGETACIÓN

Las hojas de las plantas desprenden vapor de agua. Ésta puede recogerse fácilmente envolviendo el follaje en una bolsa de plástico de manera que el vapor se condense en la bolsa en forma de pequeñas gotas.

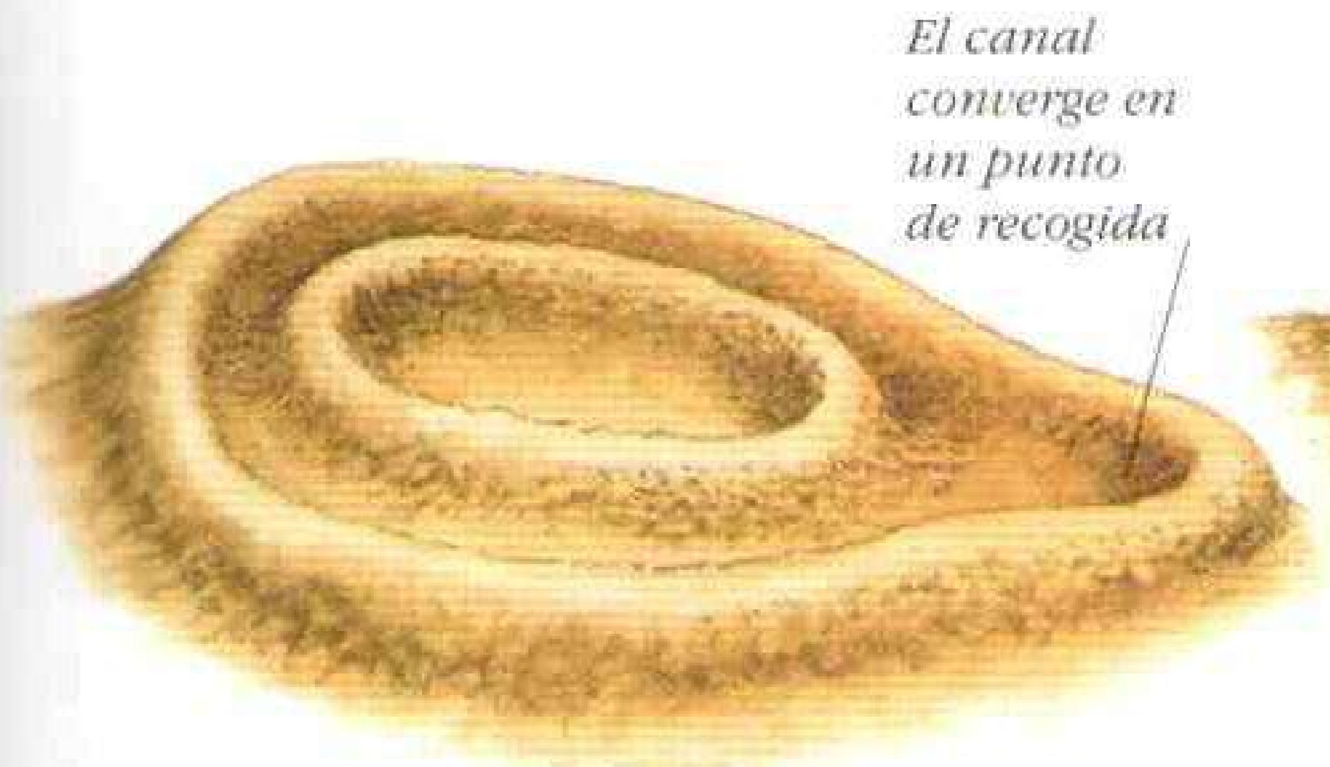
El canal converge en un punto de recogida

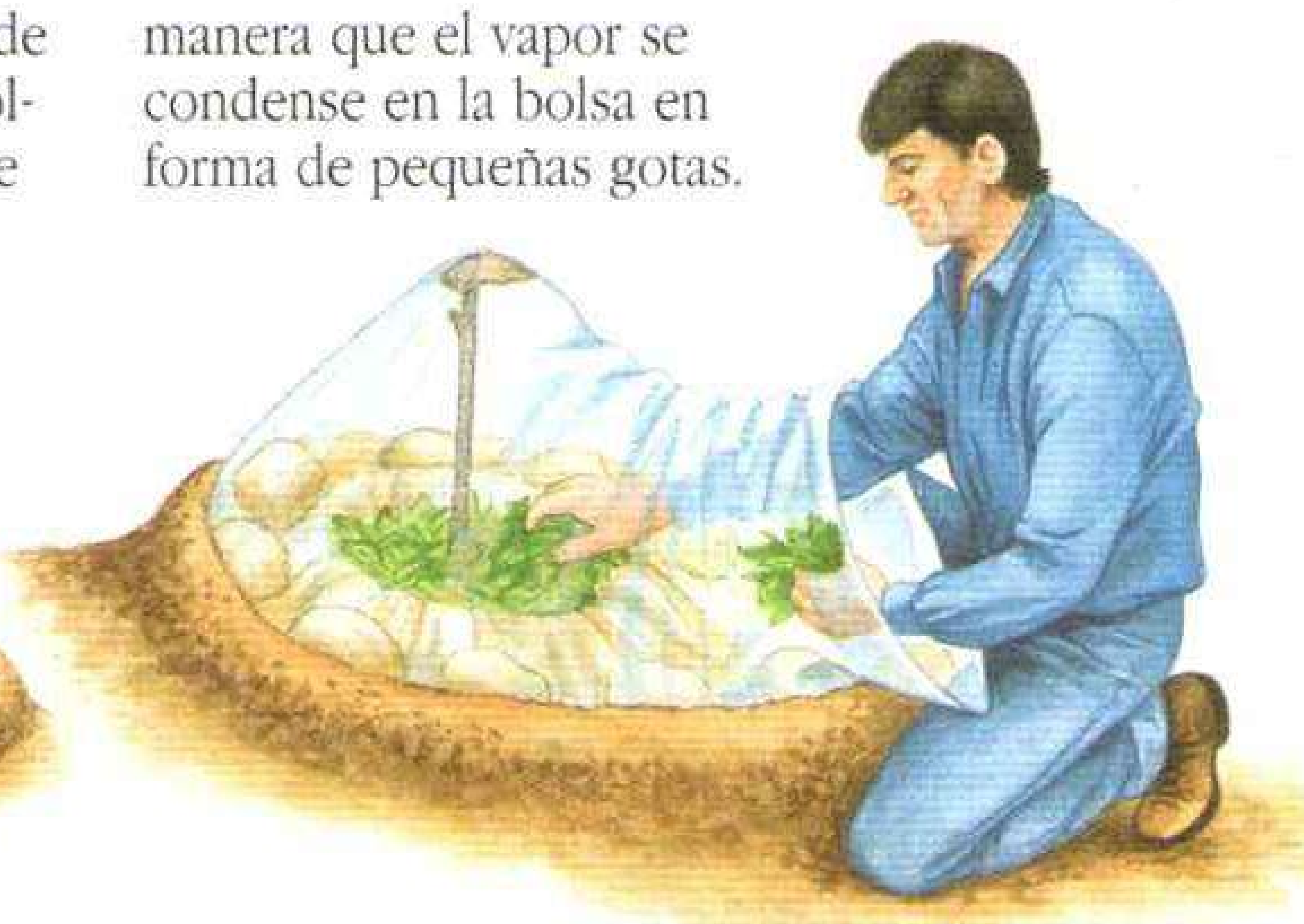

1 Cave un cráter poco profundo en una pendiente, rodeado por un cráter mayor canalizado hacia un punto de recogida. Los cráteres deben quedar separados por un caballón elevado de tierra.

2 Coloque una bolsa grande de plástico sobre ambos cráteres, alzada en el centro con una pequeña estaca y sujeta a los bordes interiores con piedras. Coloque hojas verdes y hierbas dentro del cráter central.

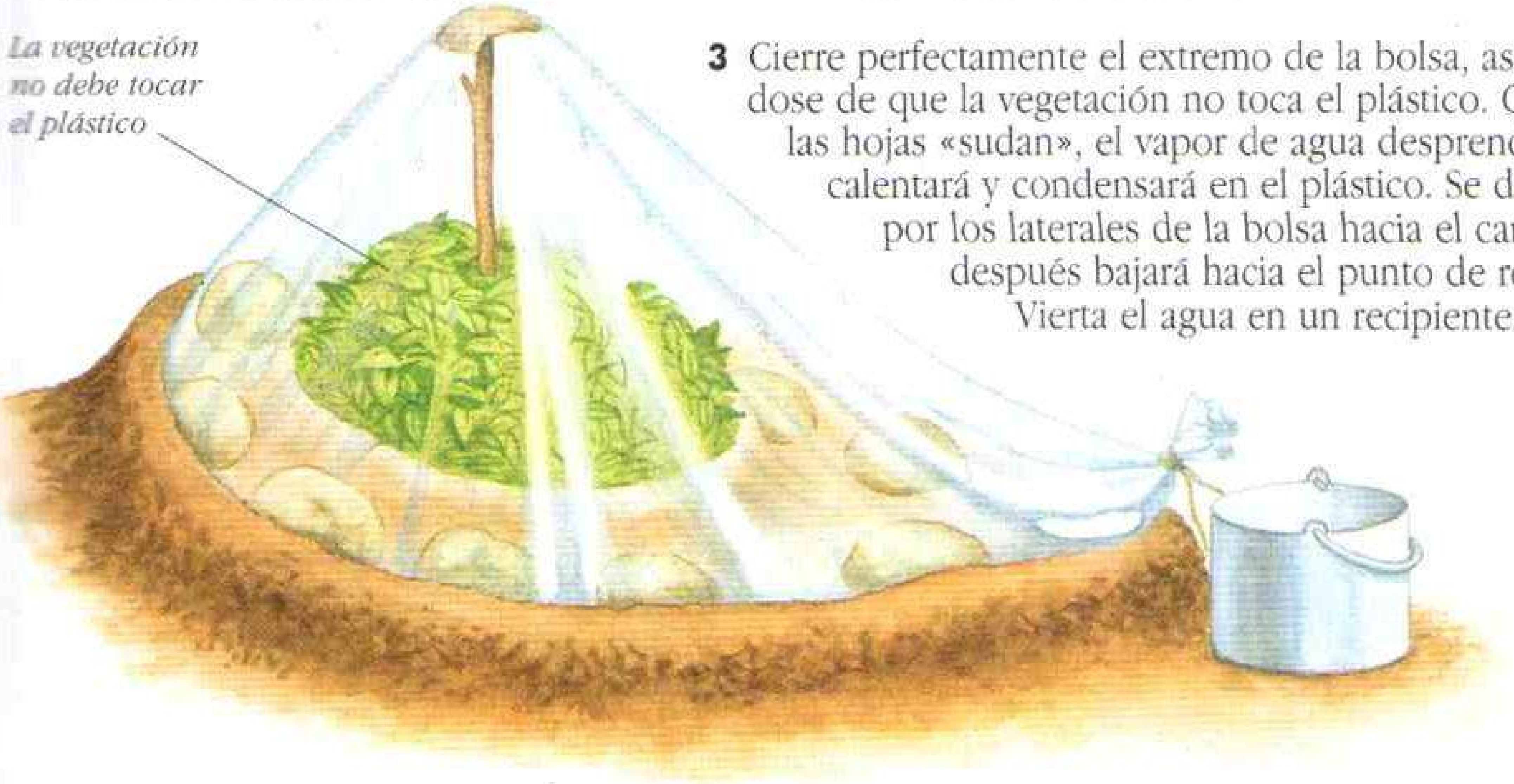
La vegetación no debe tocar el plástico

3 Cierre perfectamente el extremo de la bolsa, asegurándose de que la vegetación no toca el plástico. Conforme las hojas «sudan», el vapor de agua desprendido se calentará y condensará en el plástico. Se deslizará por los laterales de la bolsa hacia el canal, y después bajará hacia el punto de recogida. Vierta el agua en un recipiente.

OTROS MÉTODOS

En ocasiones puede recoger pequeñas gotas de agua de la neblina o de una niebla muy densa colocando un trapo en el suelo, o colgándolo entre los árboles. En un caso de emergencia, coloque una lámina de plástico sobre la hierba durante la noche. Conforme el aire se enfría, el aire caliente del suelo se condensará en pequeñas gotas de agua en la cara inferior de la lámina. Aunque con este método no recogerá mucha agua, puede ser suficiente para mantenerle vivo hasta que encuentre una fuente mejor.

Empleo de una rama. *Ate una bolsa de plástico alrededor de una rama o una planta entera. El vapor de agua que desprende el follaje se calentará dentro del plástico y se condensará en agua en la superficie interna de la bolsa.*

El vabo se forma en la cara interna de la bolsa de plástico

CÓMO EXTRAER AGUA DEL SUELO

También puede extraer agua del suelo empleando un destilador solar. Mientras exista una diferencia de temperatura entre dos superficies, el aire entre ambas se calentará y saturará, condensando el agua en forma de pequeñas gotas sobre la superficie más fría.

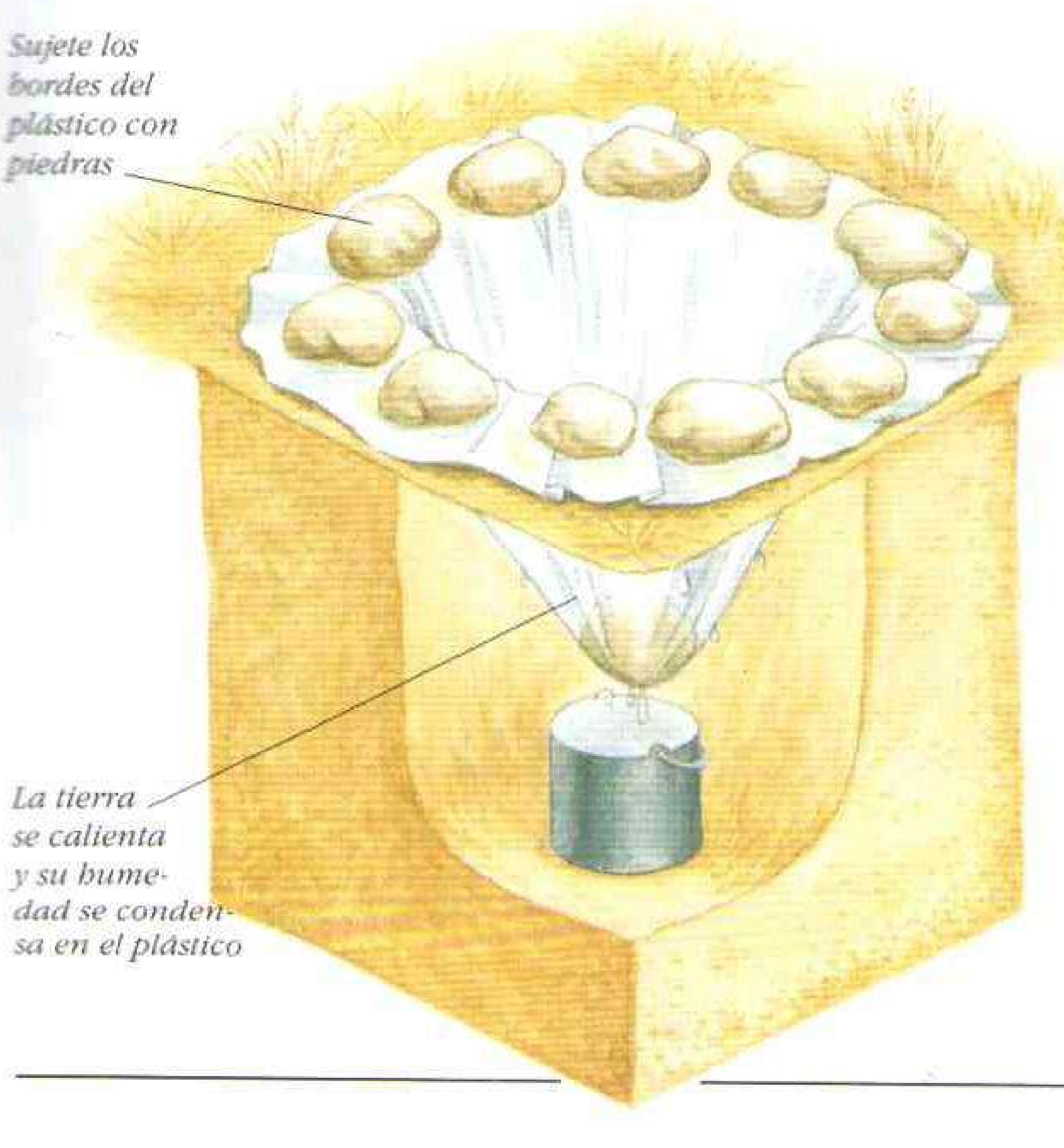
Sujete los bordes del plástico con piedras
La tierra se calienta y su humedad se condensa en el plástico

Construcción de un destilador solar. *Cave un agujero de 1 m de ancho y unos 60 cm de profundidad. Coloque un recipiente en el fondo. Extienda una bolsa de plástico sobre el agujero y sujétela con piedras. Sobrecargue el centro de la lámina sobre el recipiente con una piedra del tamaño de un puño. Conforme el aire en el agujero se caliente, el vapor de agua se condensará en la parte inferior de la lámina fría, y se deslizará hacia el recipiente. Este tipo de destilador funciona muy bien en el desierto. Por la noche, el aire exterior enfría la lámina, y como el aire dentro del agujero es más caliente, el agua se condensa en la lámina. Cave otro agujero cuando se haya agotado la humedad.*

CONDENSACIÓN DEL AGUA DE MAR

El agua de mar o la orina pueden condensarse en agua potable empleando el mismo principio del destilador solar. Coloque el agua salada en un recipiente, con una taza en el centro. Coloque una lámina de plástico sobre el recipiente y sujétela bien. Ponga una piedra en el centro para formar un cono sobre la taza. El aire bajo el plástico se calienta, y conforme se calienta el agua salada se condensa en agua dulce en la parte inferior de la lámina.

DESTILADOR DE AGUA DE MAR

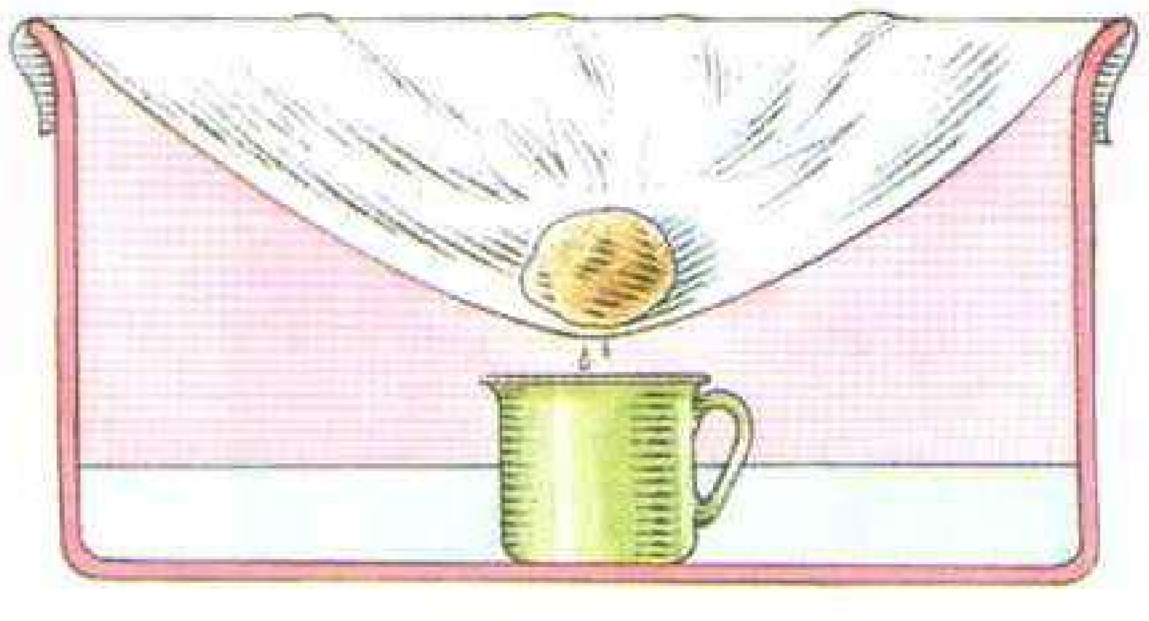

FUENTES NATURALES DE AGUA

EXISTEN MUCHAS FUENTES naturales de agua. Por ejemplo, puede obtener agua de las plantas, de la sangre y de los ojos de los animales, y de algunas ranas que retienen la humedad en su piel mientras hibernan enterradas en el barro de las charcas secas. Los nativos saben dónde encontrar agua, que es su recurso más preciado. Al utilizar estas fuentes naturales de agua, puede estar destruyendo un ecosistema cuidadosamente mantenido, y puede levantar las objeciones de las gentes del lugar. Debe tener cuidado de no dañar las charcas o fuentes naturales; asegúrese de que puede recoger este agua, y no desperdicie ni una gota.

EJEMPLOS DE PLANTAS QUE ALMACENAN AGUA

Muchas variedades diferentes de plantas almacenan agua, bien en sus hojas o en sus raíces. Algunas recogen el agua de la lluvia para atrapar insectos o comida, mientras que otras segregan líquidos especiales para su propio uso, que pueden beberse en caso de emergencia. Abajo se muestran algunos ejemplos de plantas de todo el mundo que almacenan agua.

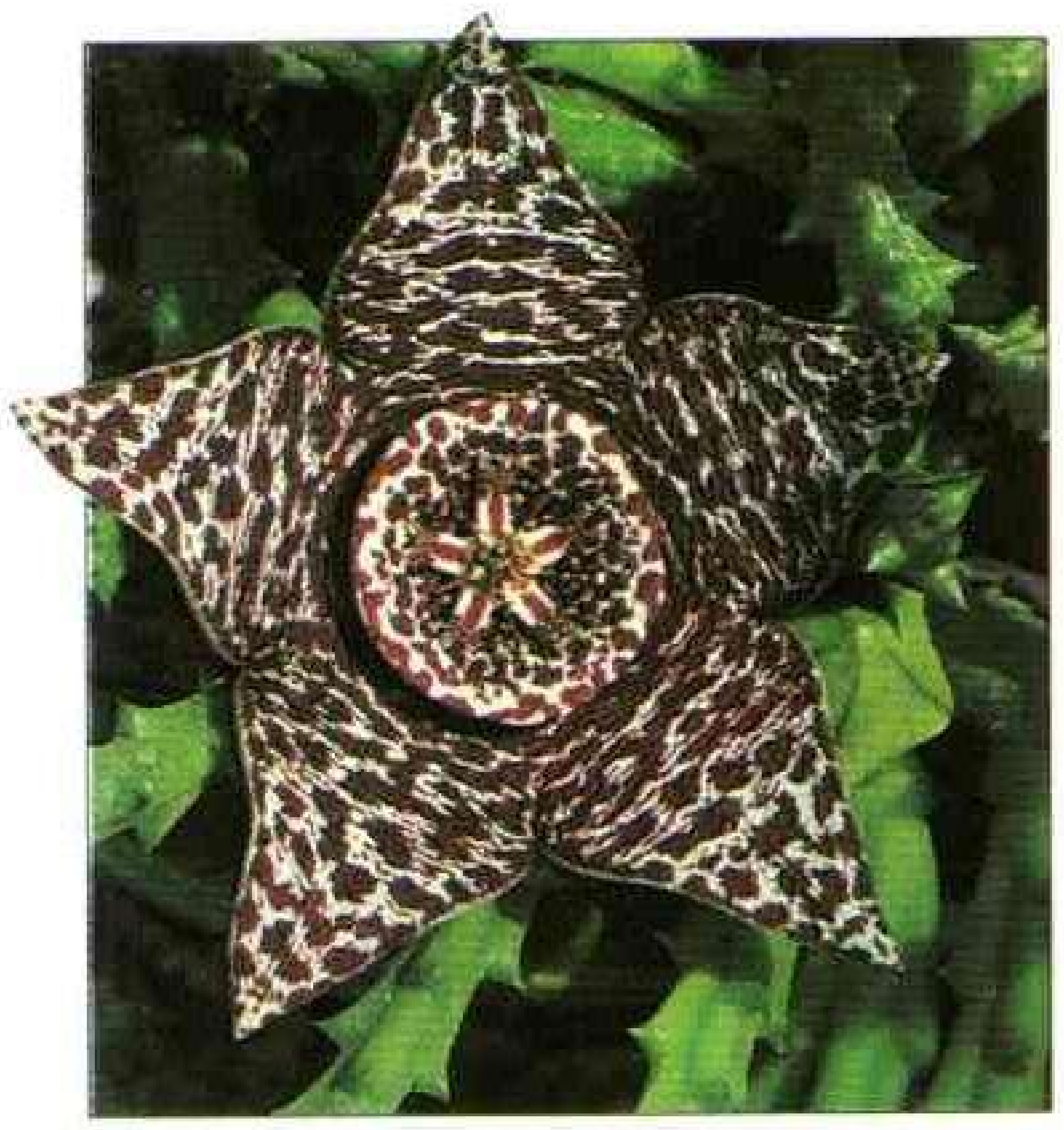

Estrella africana (Stapelia). *Al igual que otras plantas carnosas, la estrella africana almacena agua en el tallo. También puede masticar sus hojas carnosas para obtener humedad.*

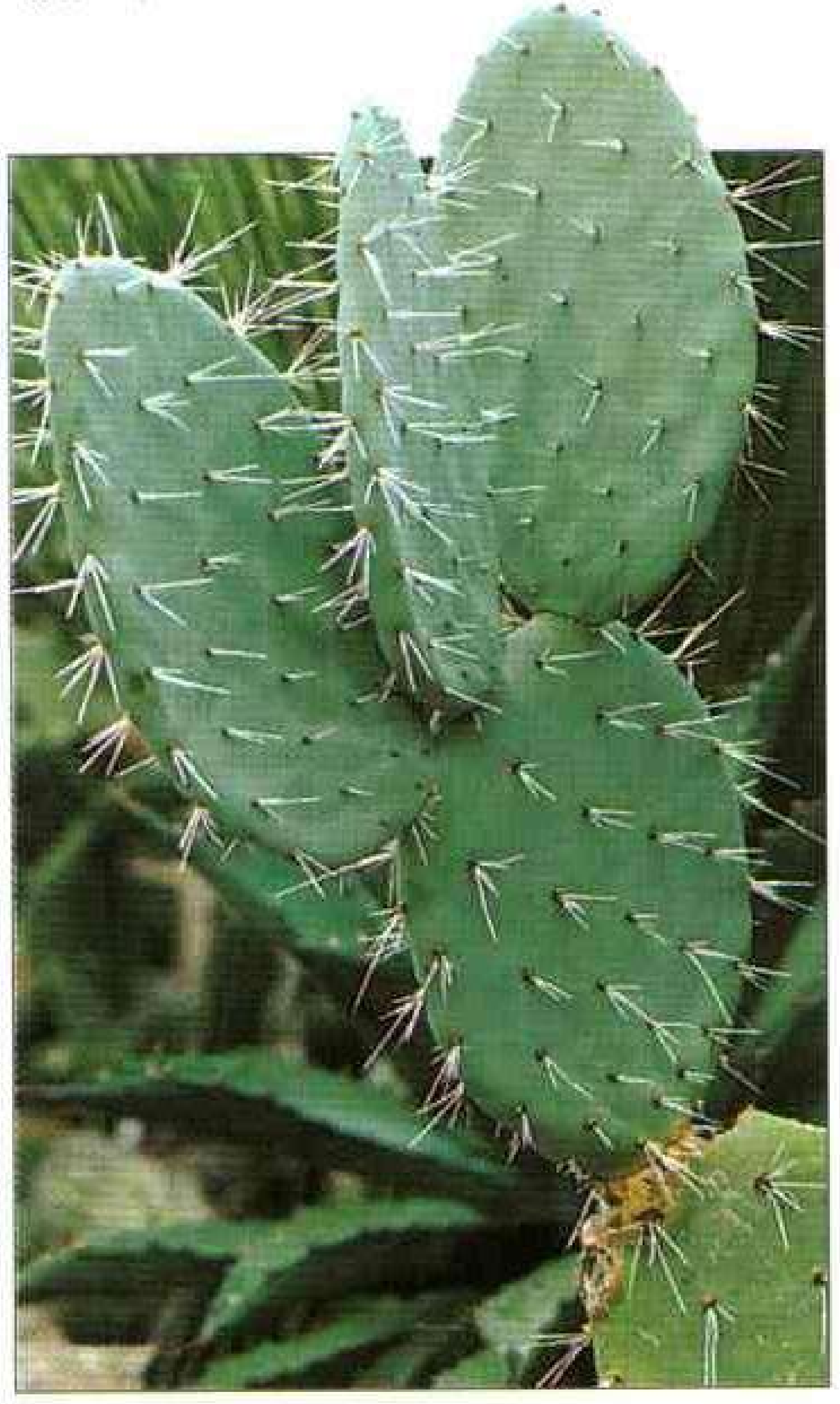

Chumbera (Opuntia). *Algunos cactus, como la chumbera, tienen hojas carnosas que desprenden humedad al masticarlas.*

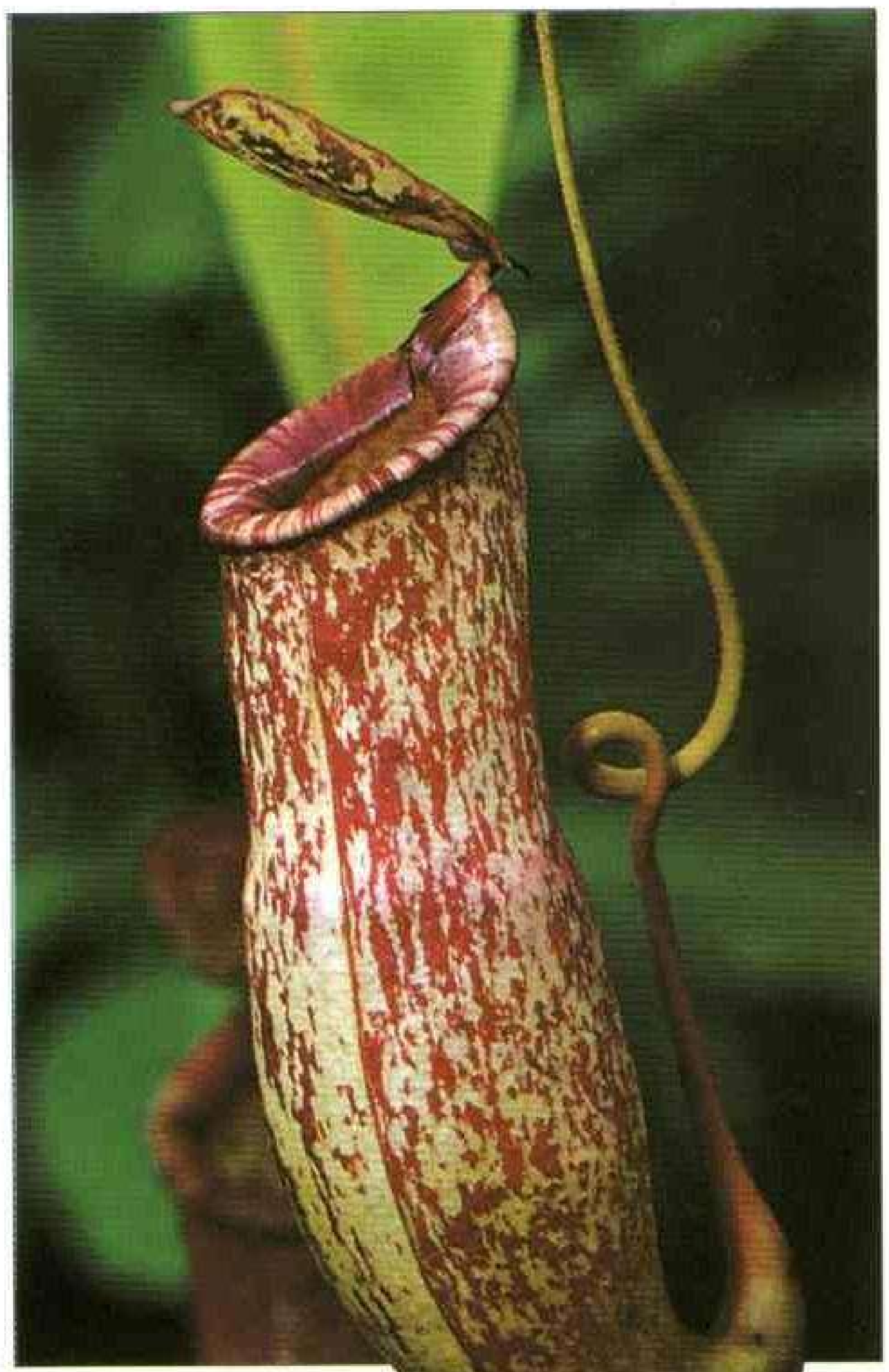

Nepente o sarracenia (Nepenthes). *Esta planta atrapa insectos en un líquido acuoso en su «jarra». Puede extraer agua, pero debe filtrarla antes para separar los insectos (que puede comer también).*

OTRAS FUENTES DE AGUA

Muchas palmeras contienen un líquido azucarado que puede extraerse de sus tallos floridos. Las raíces de las plantas del desierto son almacenes de agua, pero son difíciles de arrancar. Algunas ranas africanas y australianas almacenan agua en sus cuerpos para utilizarla durante su época de hibernación. En una emergencia extrema puede estrujarlas para extraer el agua.

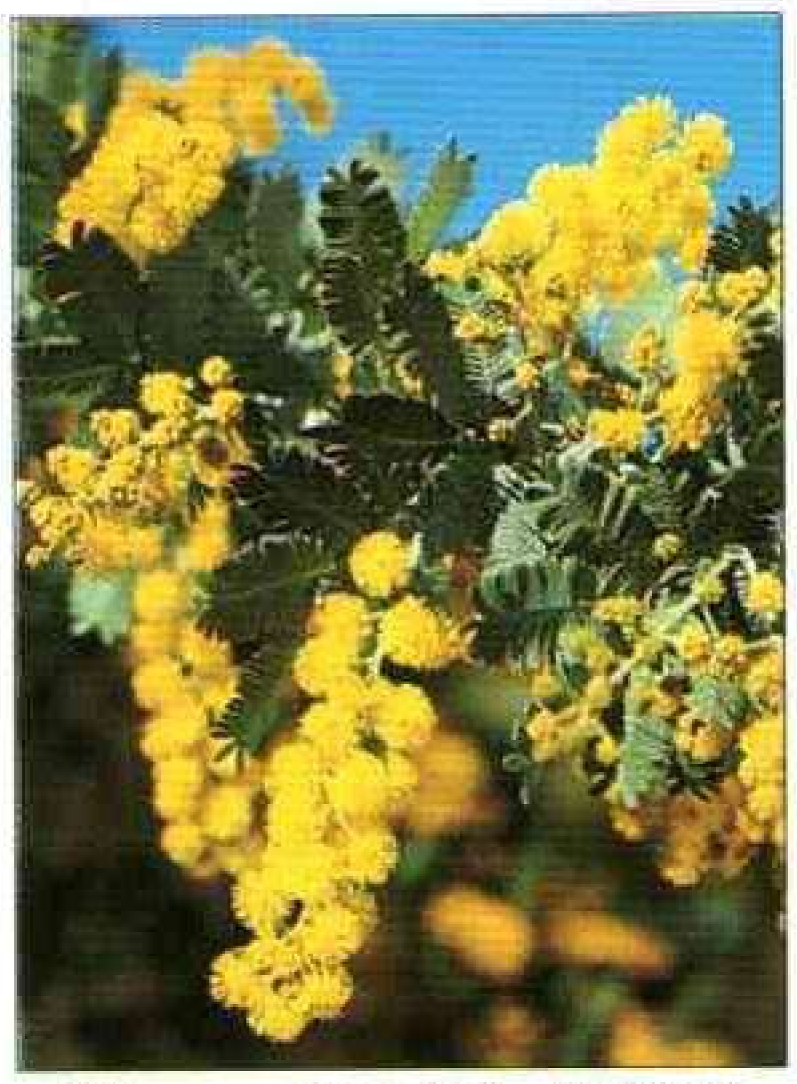

Mimosa (Acacia). *Al igual que otros árboles del desierto, la mimosa almacena agua en sus raíces, justo por debajo de la superficie del suelo.*

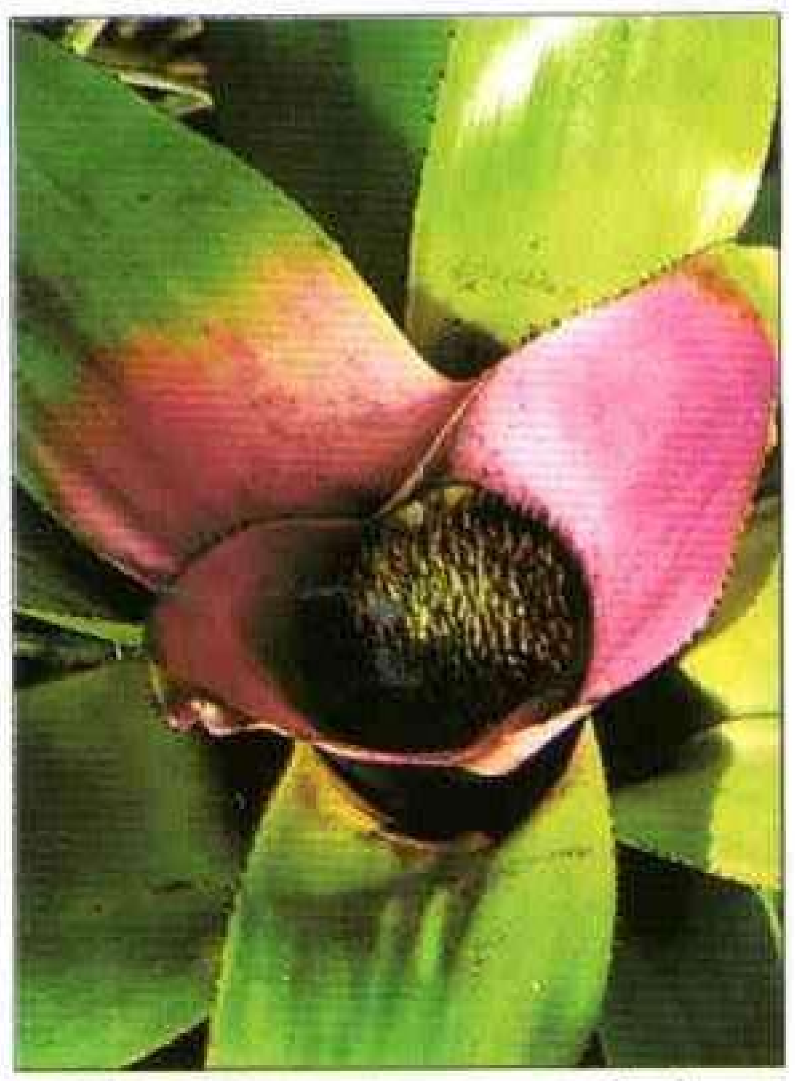

Neoregalia. *A menudo las bromeliáceas recogen agua de lluvia en sus centros, y tienen hojas comestibles que pueden masticarse para extraer agua.*

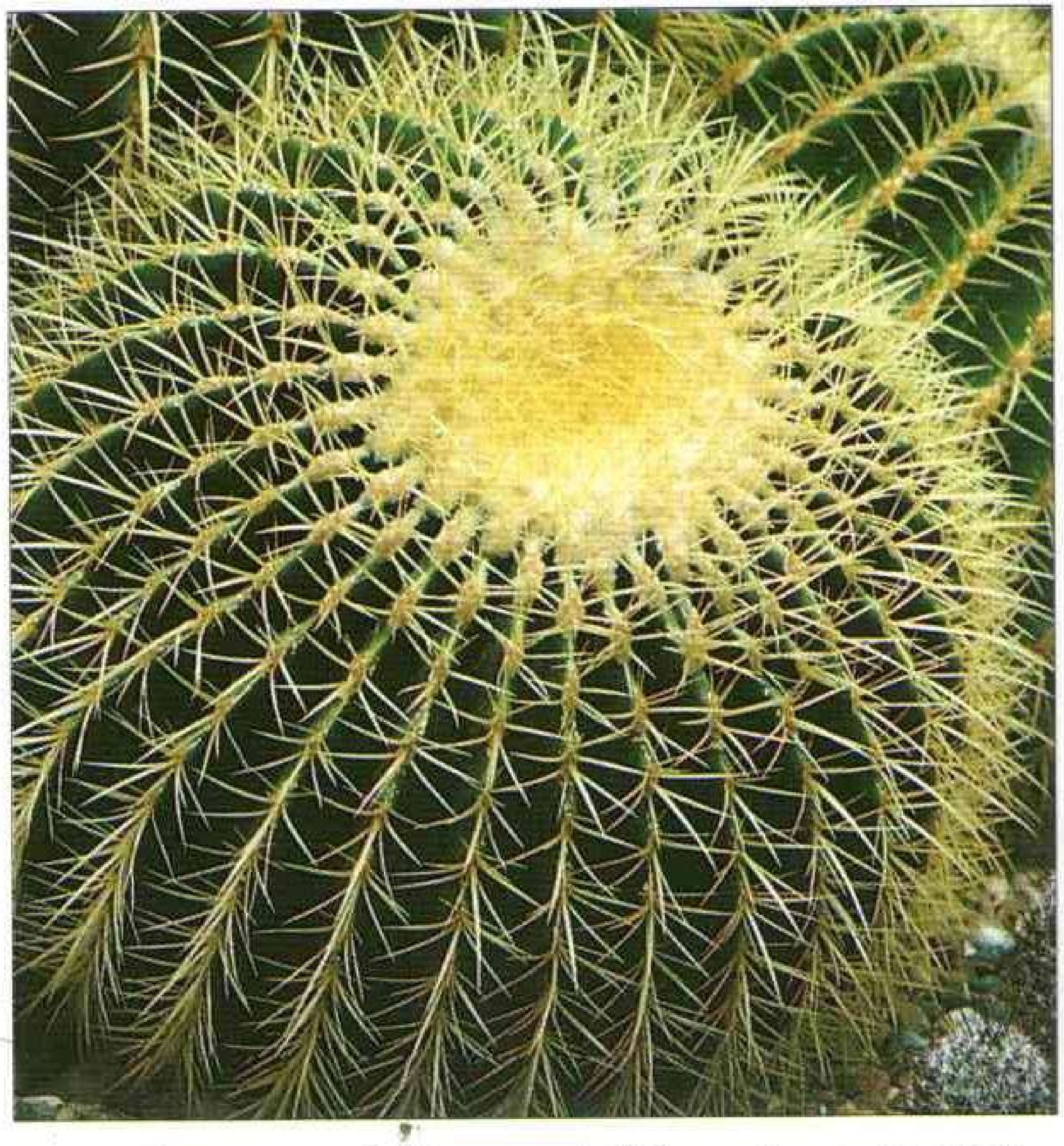

Cactus (Ferocactus, Echinocactus). *Estos cactus en forma de tonel son la única excepción a la regla de no beber ningún líquido lechoso proveniente de una planta* (véase página siguiente). *Hay varias especies de cactus de este tipo que pueden crecer hasta 1 m.*

CÓMO EXTRAER AGUA DE UN CACTUS

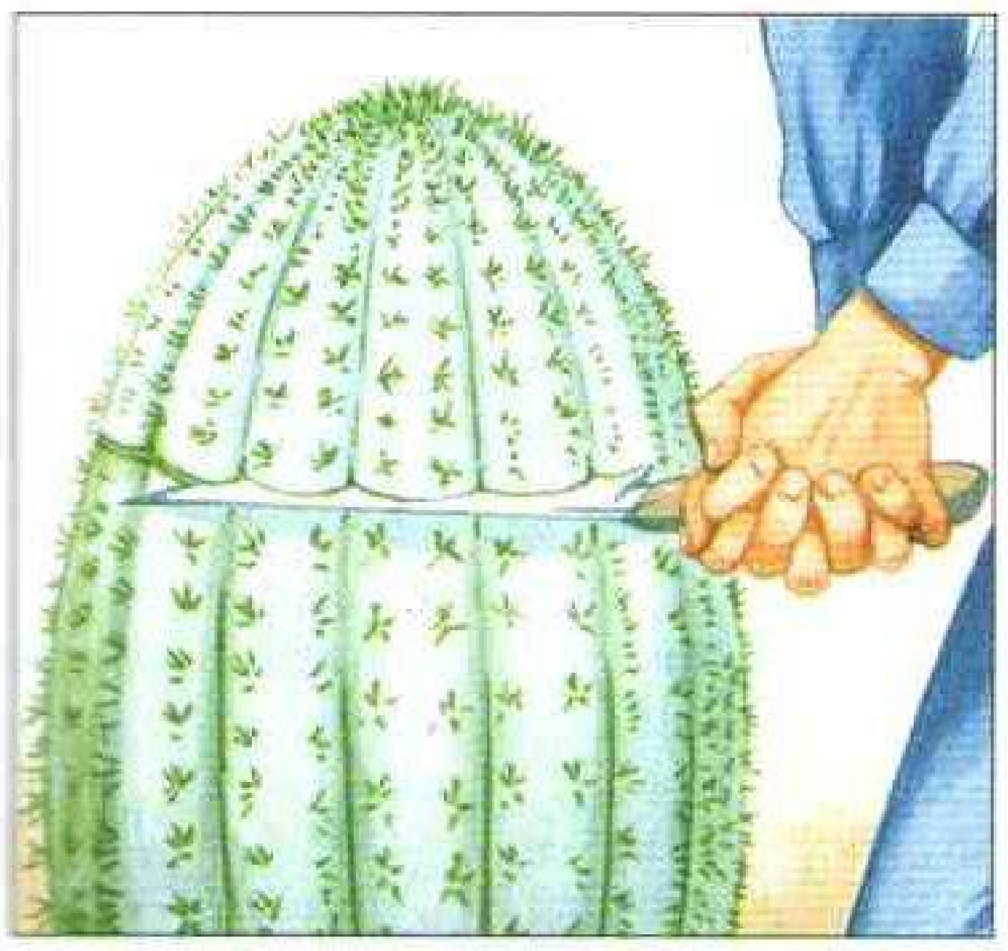

1 Para extraer agua de un cactus, corte cuidadosamente la parte superior con un cuchillo muy afilado o un machete, evitando las espinas.

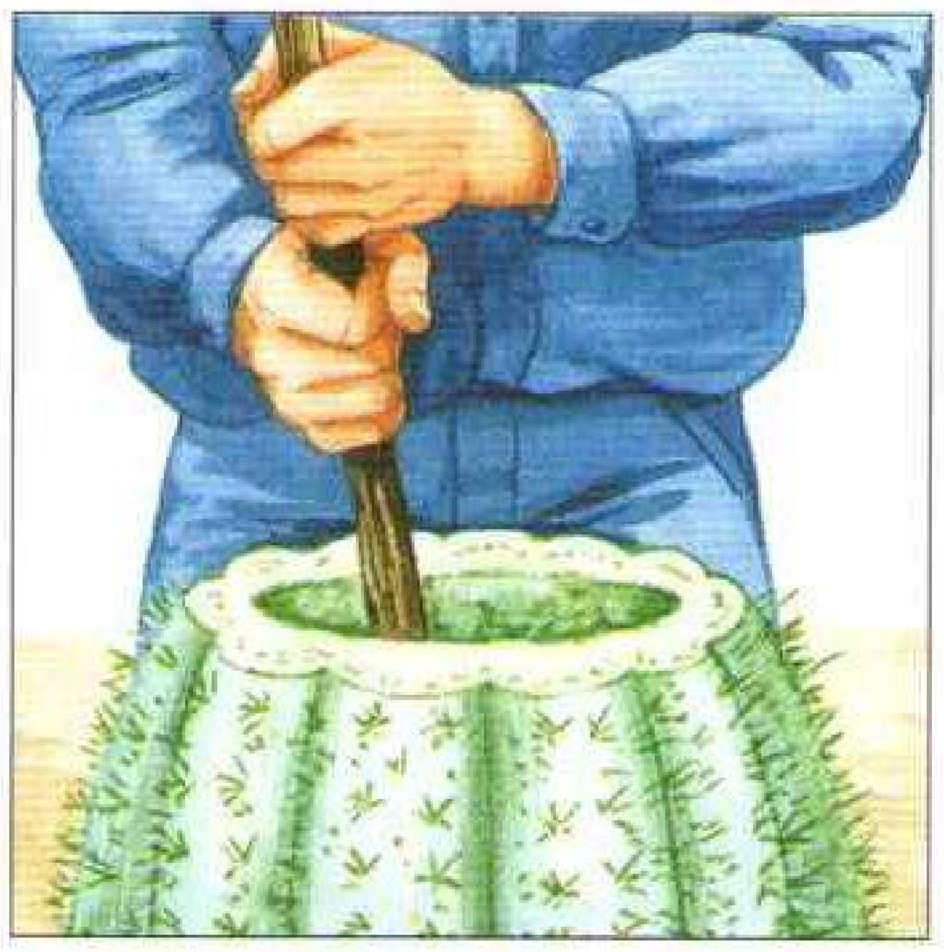

2 Machaque la carne del interior del cactus con un palo hasta formar una pulpa. De esta manera separará el agua que contiene.

3 Chupe el jugo de la pulpa a través del tallo hueco de una hierba o caña. Únicamente debe utilizar un cactus para extraer agua en caso de emergencia, ya que su crecimiento es muy lento, y algunas especies podrían extinguirse en ciertas áreas si se utilizan en exceso.

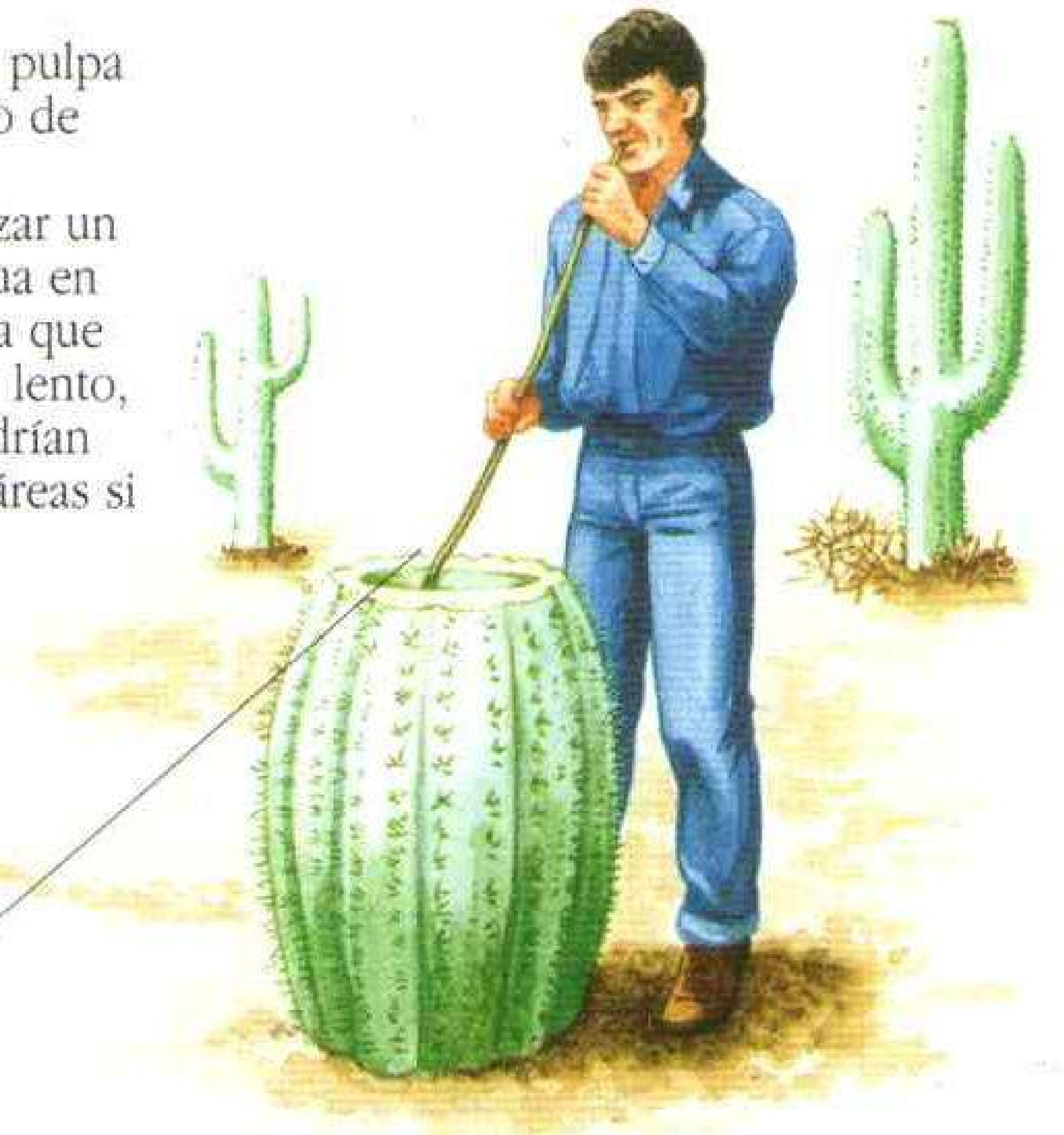

Chupe el jugo de la pulpa utilizando una caña hueca

CÓMO EXTRAER AGUA DE UNA LIANA

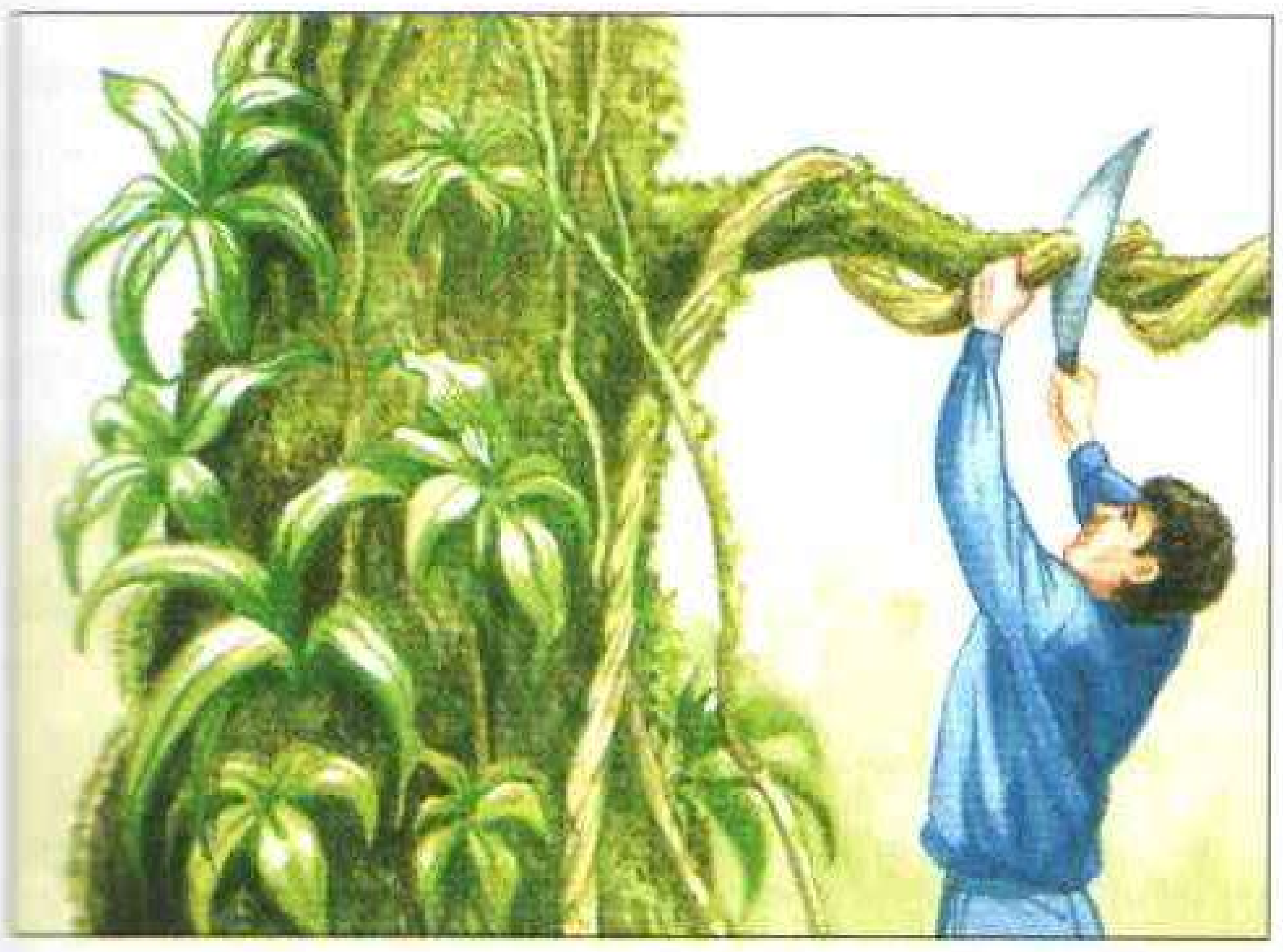

1 Corte la liana en el punto más alto que pueda alcanzar, con un cuchillo largo o un machete. No corte primero la parte inferior, ya que el líquido fluiría hacia arriba por la acción capilar.

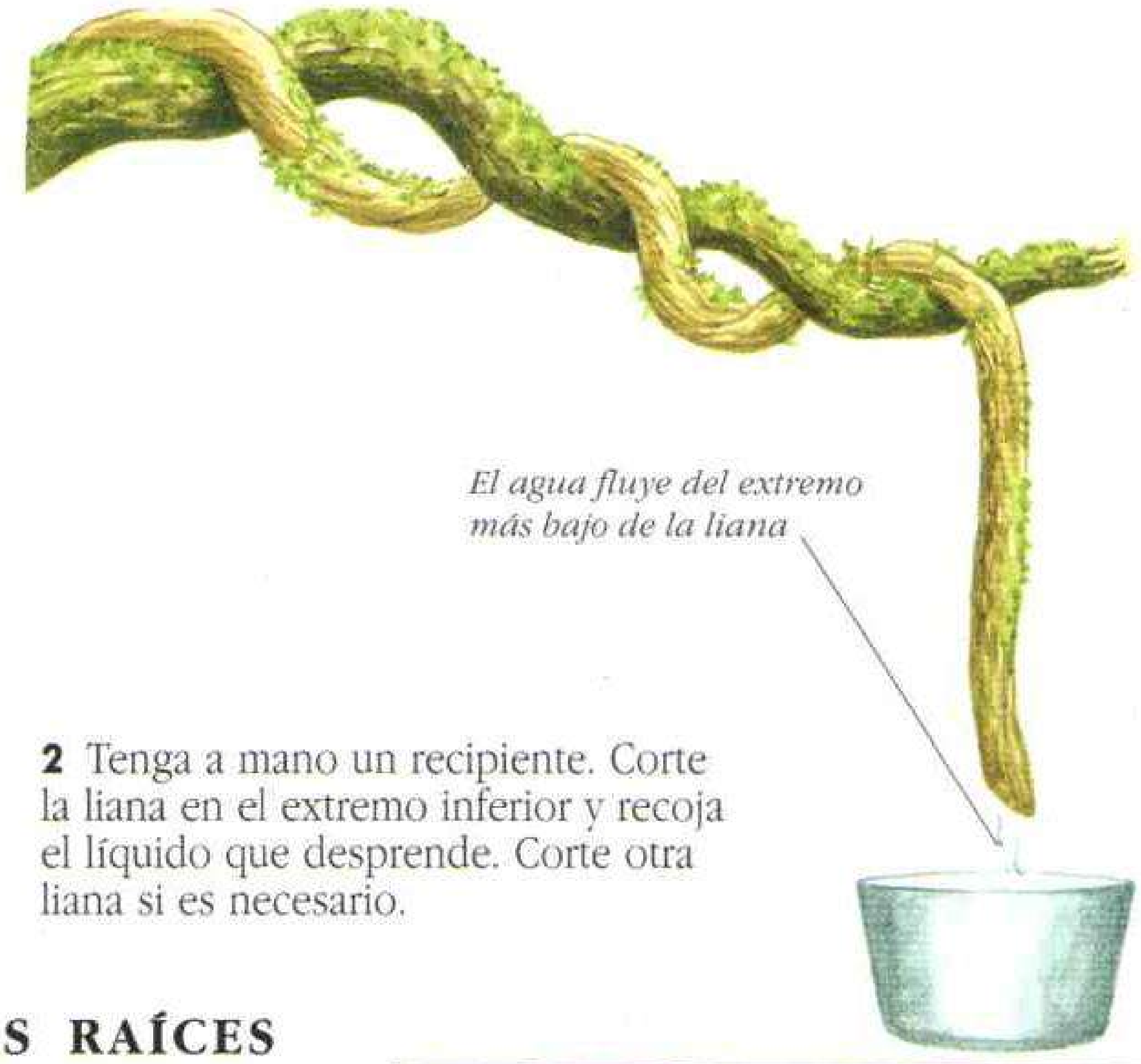

El agua fluye del extremo más bajo de la liana

2 Tenga a mano un recipiente. Corte la liana en el extremo inferior y recoja el líquido que desprende. Corte otra liana si es necesario.

ADVERTENCIA

La savia lechosa de las plantas es generalmente venenosa, al igual que el líquido del cactus saguaro gigante *(Carnegiea)* del sudoeste americano. La leche de coco y los líquidos animales contienen proteínas, que requieren agua adicional para ser digeridos por el cuerpo. Además, la espesa y rica leche de coco maduro es un fuerte laxante, que podría causarle diarrea, cansancio y deshidratación.

CÓMO EXTRAER AGUA DE LAS RAÍCES DE LOS ÁRBOLES

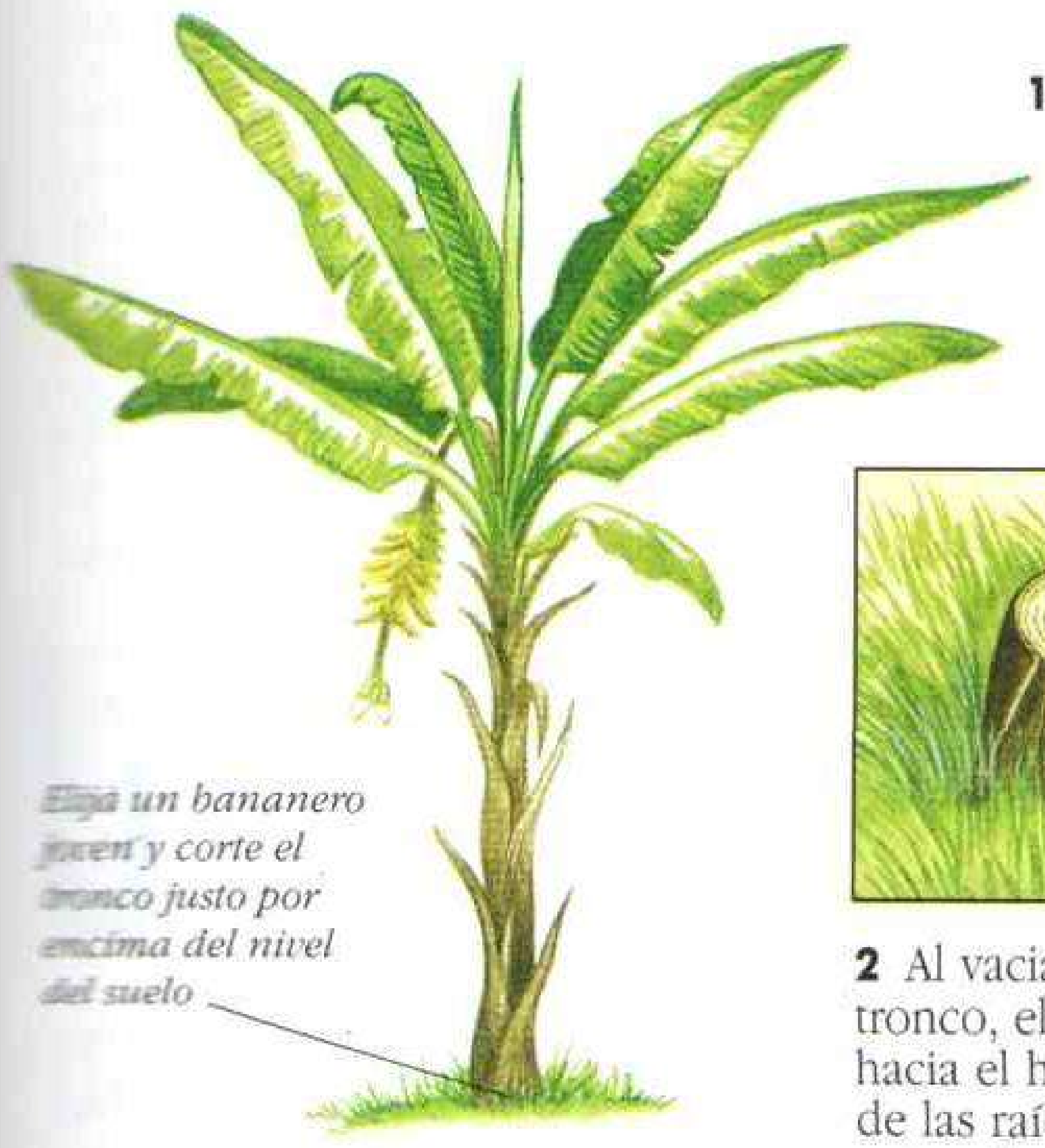

Elija un bananero joven y corte el tronco justo por encima del nivel del suelo

1 Corte un banano joven a una altura de 8 cm por encima del suelo.

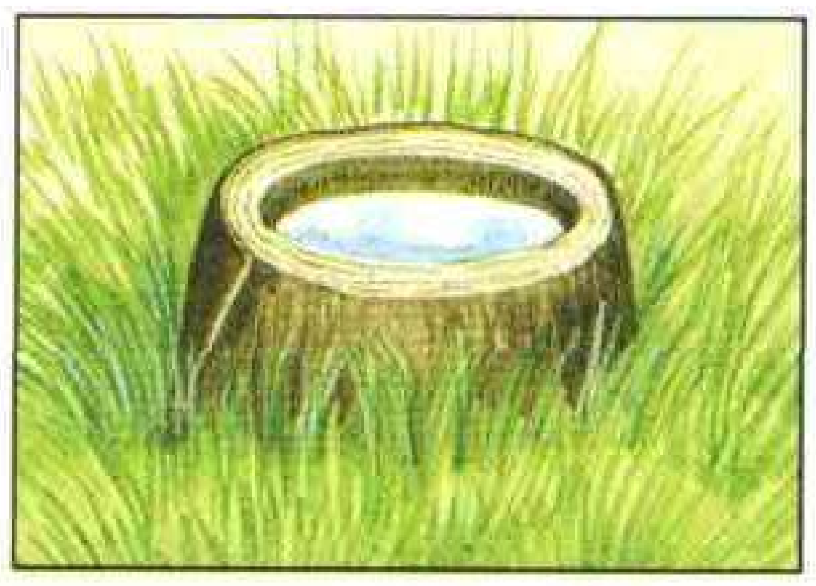

2 Al vaciar el centro del tronco, el agua se filtrará hacia el hueco proveniente de las raíces, y podrá extraerla.

AGUA DEL HIELO Y DE LA NIEVE

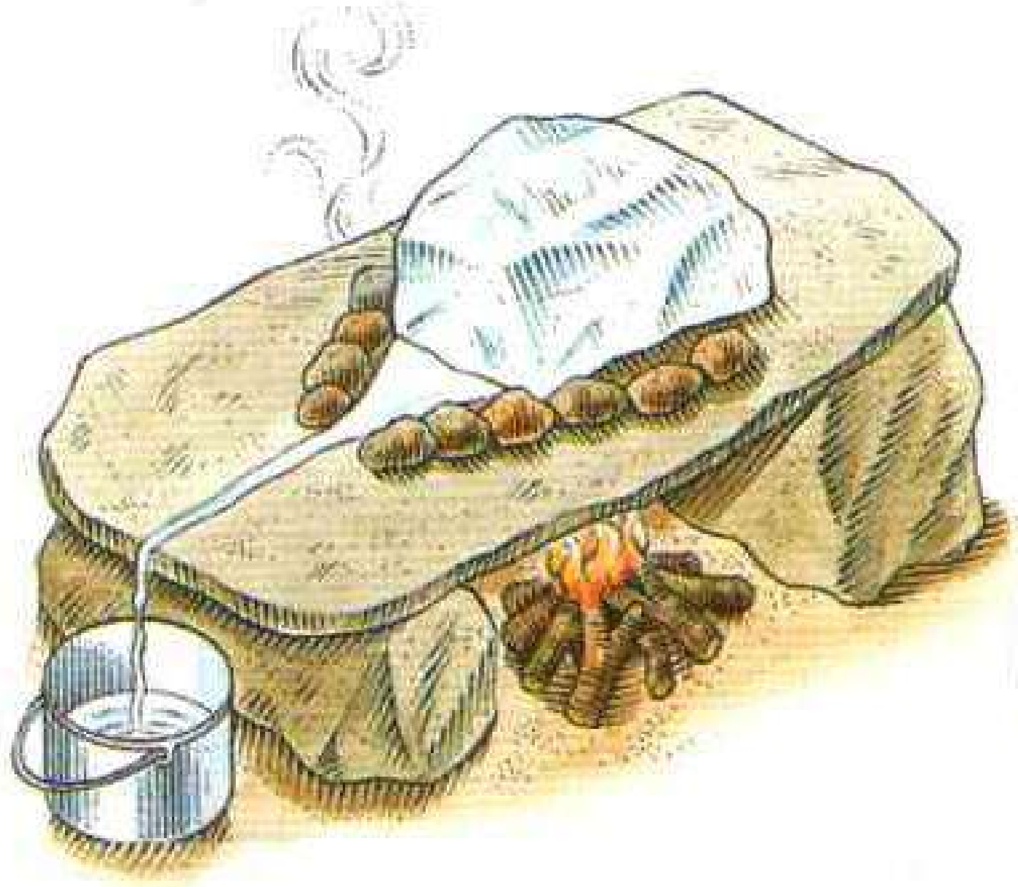

Fundir hielo. *Funda el hielo sobre una roca inclinada sobre un fuego. No utilice hielo marino recién congelado, ya que contiene sal. Utilice hielo viejo, de color azulado, que contiene mucha menos sal.*

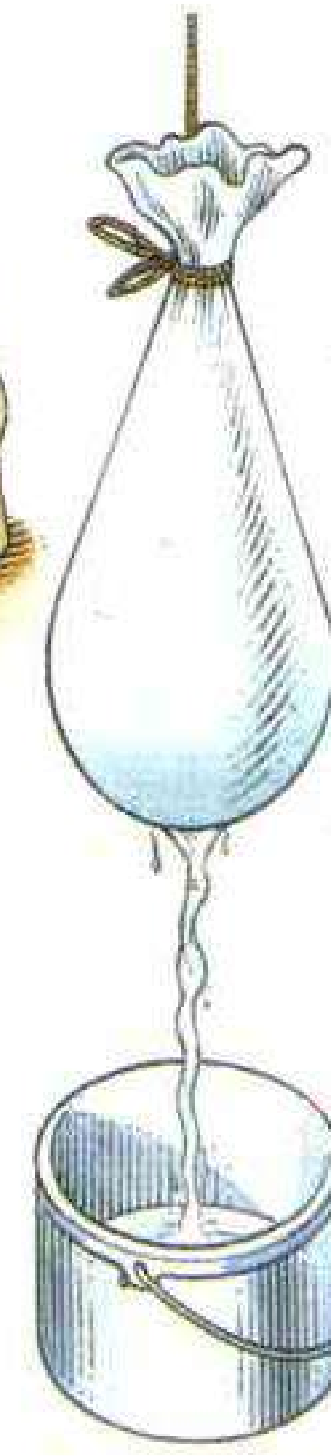

Fundir nieve. *Utilice la nieve densa, extraída de la parte más profunda posible. Cuélguela en una tela sobre un recipiente cercano al fuego, y se derretirá gradualmente. No funda nieve en un recipiente sobre el fuego, ya que el agua producida será absorbida rápidamente por la nieve remanente y el recipiente se quemaría. Unos 40 cm de nieve producen 3 cm de agua potable.*

LA DEPURACIÓN DEL AGUA

En la naturaleza el agua rara vez es pura, y siempre debe depurarse antes de beberla, ya que puede contener microorganismos patógenos que pueden transmitir enfermedades. También deben separarse las partículas de limo y otros contaminantes. En los países occidentales, la gente supone que el agua siempre es potable, especialmente si sale de un grifo. Sin embargo, en muchas zonas, ni siquiera el agua del grifo es pura. Algunas personas toleran bien el agua ligeramente impura, pero también se acostumbran a sufrir toda una serie de problemas gastrointestinales.

Agua potable. *El agua natural debe filtrarse, por lo menos a través de un calcetín o un pañuelo. Asimismo, debe purificarse siempre para eliminar los organismos dañinos.*

MÉTODOS PARA DEPURAR EL AGUA

A excepción de la lluvia, incluso el agua con el aspecto más puro no es únicamente H_2O, sino que contiene sales, minerales y microorganismos. El agua de la naturaleza debe, en primer lugar, filtrarse para eliminar las partículas y, a continuación, debe depurarse para destruir los organismos patógenos hirviéndola por lo menos durante cinco minutos o tratándola con agentes químicos. Nunca debe beber orina o agua de mar en su estado natural, pero en caso de una emergencia pueden destilarse *(véase página siguiente)*. Existen muchos tipos de depuradores, la mayoría de los cuales filtran el agua y después la purifican por la acción de agentes químicos. Como forma alternativa, puede utilizar las pastillas depuradoras o hervir el agua después de filtrarla *(véase página siguiente)*.

Asegúrese de que el cuentagotas esté bien cerrado

Yodo. *El yodo puede ensuciarle al usarlo, y tiñe el agua de rosa. También da al agua un sabor particular.*

Permanganato potásico. *Ésta es una sustancia química de uso general para esterilizar el agua. Utilice el suficiente para teñir el agua de un ligero color rosado.*

Pastillas. *Estas pastillas contienen una base de cloro que da al agua un sabor que recuerda al de las piscinas, pero son muy seguras. Utilice una pastilla por cada medio litro de agua.*

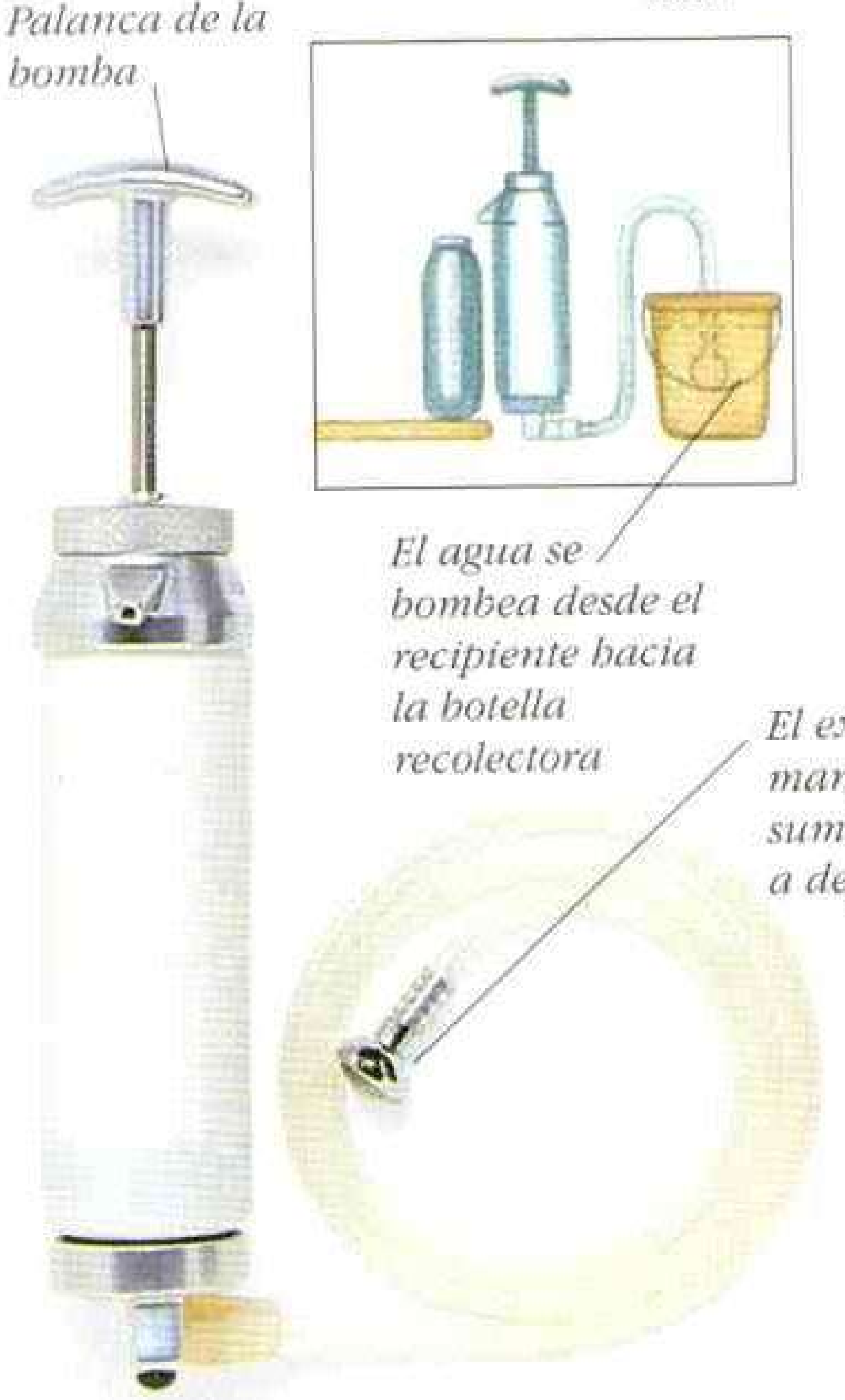
Palanca de la bomba
El agua se bombea desde el recipiente hacia la botella recolectora
El extremo de la manguera se sumerge en el agua a depurar

Mini filtro portátil. *El extremo de la manguera se sumerge en un recipiente con agua impura. El agua pura sale del pico de la bomba conforme el agua se bombea al subir y bajar la palanca, y se recoge finalmente en un recipiente.*

La tapa del filtro sirve de taza

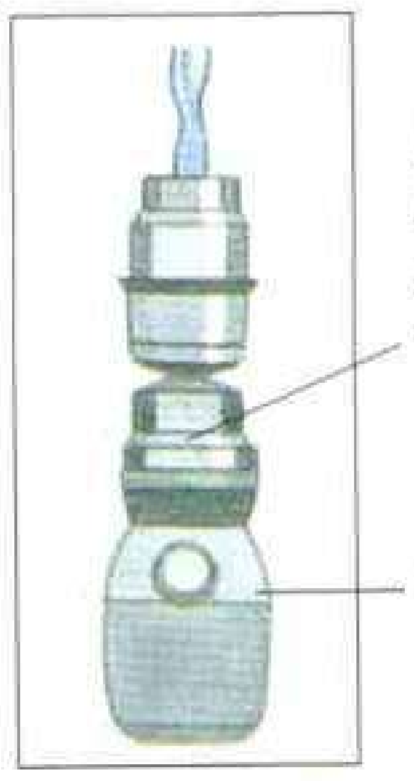
El filtro se ajusta a la cantimplora para filtrar pequeñas cantidades de agua
Cantimplora

Las piezas del filtro encajan en una botella de agua
Lleva incorporado un filtro con sustancias químicas

Depurador con taza. *Éste puede guardarse en un bolsillo cuando se desmonta. El agua impura se vierte en el filtro, que se ajusta sobre la parte superior de una botella de agua. El agua pura caerá gota a gota en la cantimplora.*

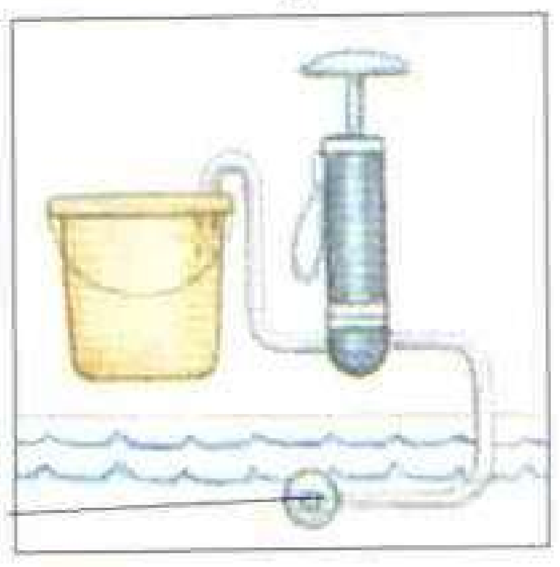
El agua a purificar es aspirada por el bulbo, y pasa a través de las sustancias químicas antes de salir purificada por el otro extremo de la manguera

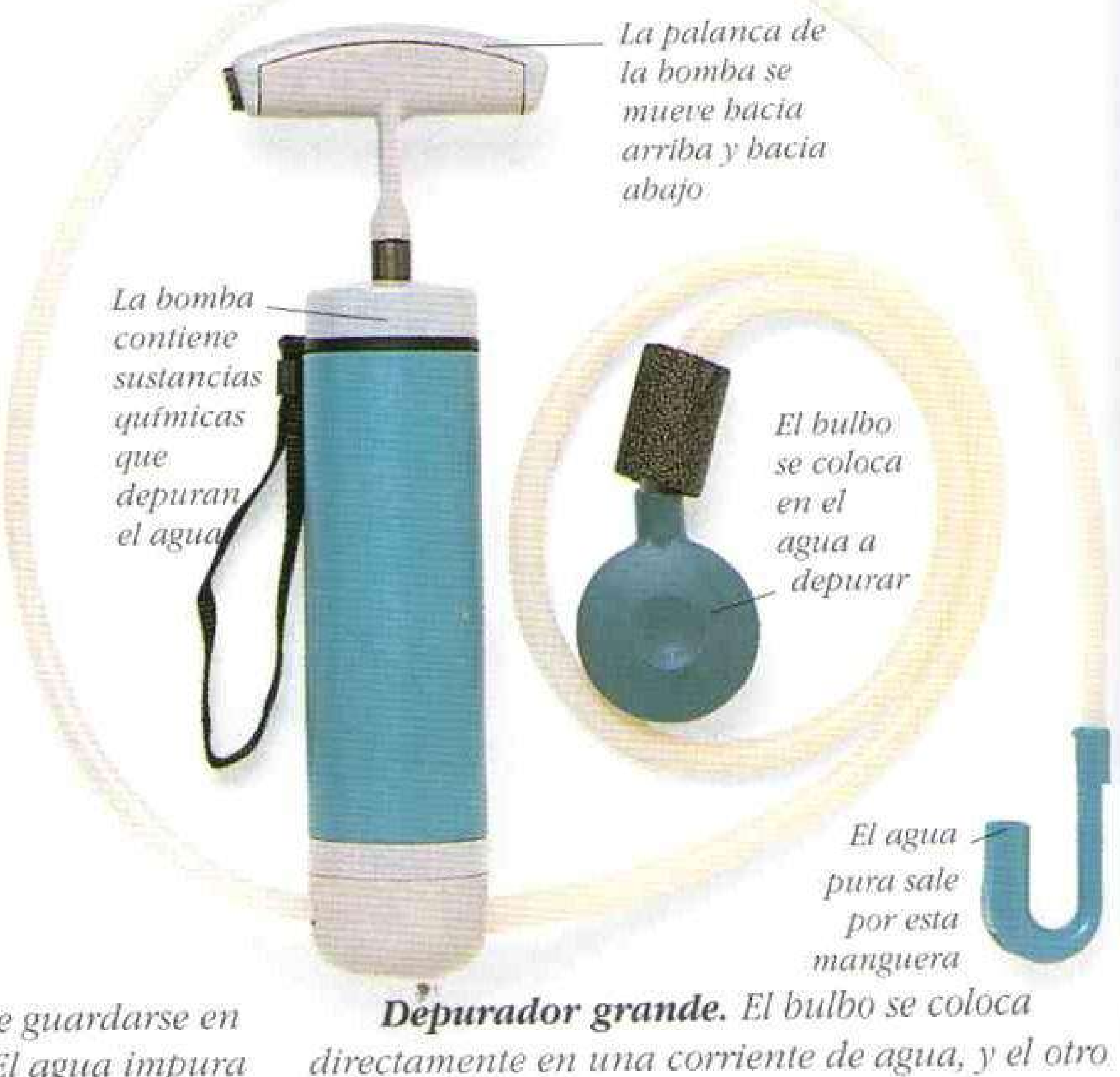
La palanca de la bomba se mueve hacia arriba y hacia abajo
La bomba contiene sustancias químicas que depuran el agua
El bulbo se coloca en el agua a depurar
El agua pura sale por esta manguera

Depurador grande. *El bulbo se coloca directamente en una corriente de agua, y el otro extremo de la manguera dentro de un cubo. El agua de la corriente es aspirada por el depurador al subir y bajar la palanca.*

TRATAMIENTO IMPROVISADO DE AGUA

Filtro en un trípode. *Si no dispone de un depurador prefabricado, puede utilizar un calcetín para filtrar el agua antes de hervirla para depurarla* (véase página anterior). *Si utiliza un calcetín, rellénelo con otro calcetín mucho más fino, arena o un pañuelo, y después de cada sesión de filtrado déle la vuelta y aclárelo bien. Construya un trípode para sujetar su filtro* (véase página 119).

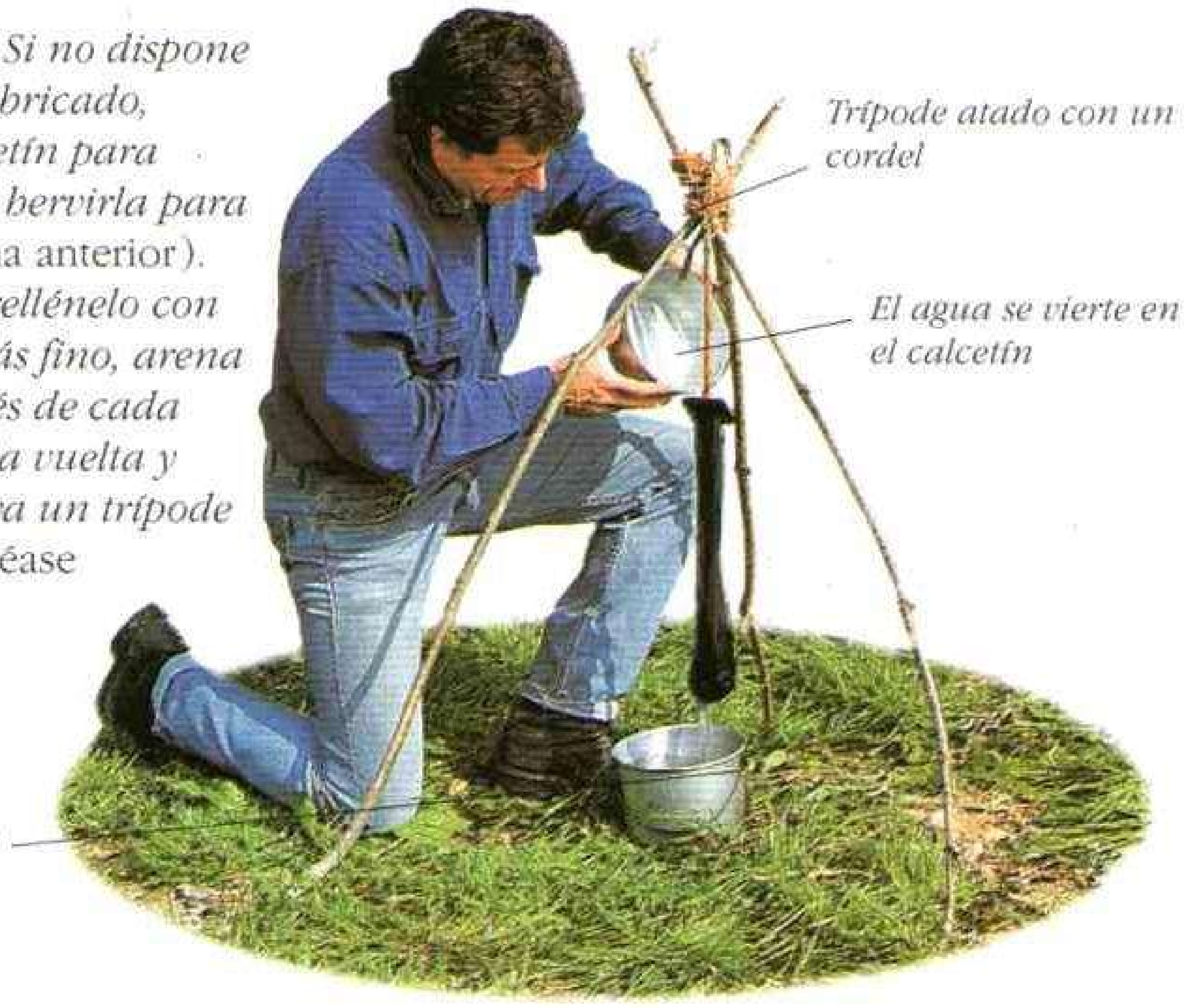
Trípode atado con un cordel
El agua se vierte en el calcetín
El agua filtrada cae en el recipiente

DESTILACIÓN DE AGUA SALADA

En un caso de emergencia puede obtener agua pura y potable a partir de agua del mar, por destilación. Hierva el agua y recoja el vapor en un trapo colocado sobre el recipiente. Utilice varios trapos a la vez, uno después de otro. Levántelos del recipiente con una rama para evitar quemarse. Una vez se han enfriado, exprima los trapos y recoja el agua.

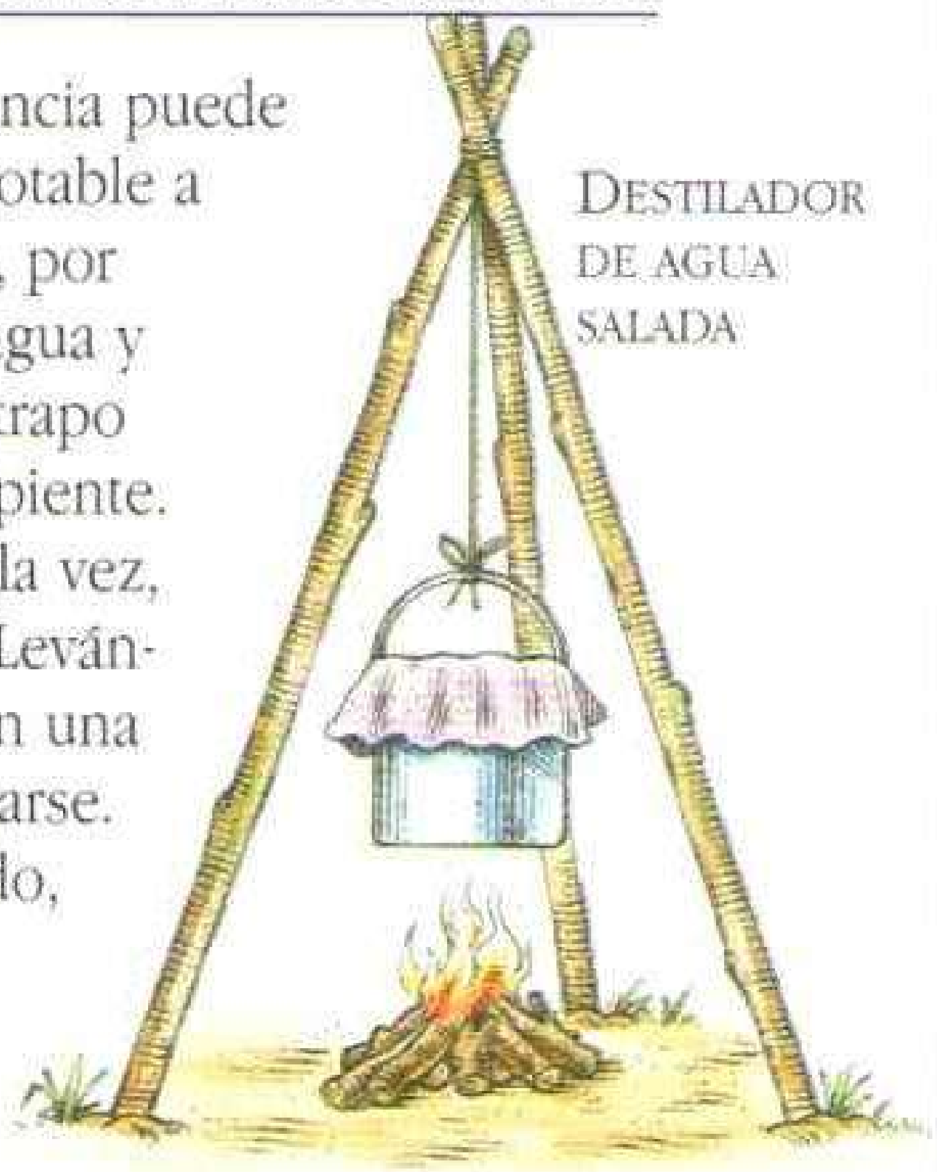

LOS PELIGROS DEL AGUA EN LA NATURALEZA

Cuando los animales beben, transmiten al agua una variedad de microorganismos y parásitos intestinales muy perjudiciales para los seres humanos. Algunos de ellos necesitan un período de incubación, y otros tienen un efecto inmediato. Como el hombre urbano no ha desarrollado una inmunidad o tolerancia a estos contaminantes, se ve muy afectado por ellos. Aunque los arroyos de las montañas deberían ser puros, la gran variedad de sustancias químicas industriales utilizadas en la agricultura, así como la posible filtración en el suelo desde los asentamientos urbanos, hacen necesario depurar el agua recogida antes de beberla.

Peligro al acecho. *La recogida de agua en estado natural puede resultar peligrosa. Si utiliza charcas en las que beben los animales salvajes tenga cuidado y evite los depredadores.*

INDICADORES DE AGUA EN MAL ESTADO

Las calaveras y huesos de animales pueden no siempre estar presentes para indicarnos que el agua no está en buenas condiciones. Si el agua está contaminada por sustancias químicas, busque sedimentos polvorientos alrededor de las orillas. La ausencia de vegetación o numerosas algas verdes en la superficie indican que el agua no es potable.

Agua estancada. *Las espadañas y los juncos indican que el agua está estancada: debe evitarse.*

ENFERMEDADES PROPAGADAS POR EL AGUA CONTAMINADA

Enfermedad	Origen	Síntomas
Leptospirosis (en su forma severa = enfermedad de Weil)	Transmitida al ser humano a partir de animales infectados con la bacteria *Leptospirum*. Contraída por contacto con orina o líquidos fetales infectados. Penetra en el organismo a través de las heridas en la piel o las mucosas de la boca, nariz, garganta u ojos.	Provoca síntomas semejantes a los de la gripe (fiebre, escalofríos, dolor de cabeza, dolor muscular). En su forma más severa causa meningitis, ictericia, trastorno renal, hemorragias y problemas cardíacos.
Bilharziosis o Bilharzia (Esquistosomiasis)	Causada por un platelminto parásito de las corrientes de agua dulce y de curso lento. Penetra en la piel directamente, alojándose en el intestino. También se transmite por los parásitos de los caracoles de agua dulce.	Causa irritación de la piel, urticaria, ataques asmáticos, dilatación del hígado e irritación del tracto urinario.
Disentería amebiana	Contraída por beber agua contaminada con aguas residuales infectadas.	Causa diarrea con sangre y/o pus, e infección del colon. Una complicación de la infección conlleva hepatitis, abscesos en el hígado y los pulmones, y úlcera intestinal.
Anquilostomiasis	Las larvas del parásito penetran en el cuerpo humano al beber agua o directamente a través de la piel.	Los gusanos adultos se alojan en el intestino, provocando anemia y letargia. Las larvas en el torrente sanguíneo pueden provocar neumonía.
Giardiosis	Provocada por el parásito *Giardia* en agua contaminada con orina o heces infectadas.	Provoca diarrea y dolores abdominales. Cada vez más abundante en Norteamérica, África y Asia.

Cómo prevenir las enfermedades transmitidas por el agua contaminada. *La única manera de evitar contraer una enfermedad transmitida por el agua es purificándola. Si un miembro de su grupo enferma por beber agua contaminada, debe aislarlo junto con alguien que le asista. Debe tener un cuidado extremo con los excrementos y ropa de la víctima, ya que contendrán grandes cantidades del organismo patógeno. Tenga cuidado con la higiene del campamento para que nadie se contagie* (véase página 56).

TRANSPORTE Y ALMACENAMIENTO DEL AGUA

Incluso si ha acampado al lado de un río, necesitará algún tipo de recipiente para contener el agua. Por ejemplo, querrá tener agua a mano para cocinar. También necesitará un recipiente para recoger agua del río. Como acampar junto a una fuente de agua no es siempre la opción más adecuada, el ir a buscar agua será una actividad que tendrá que desarrollar al menos una o dos veces al día. Además, al salir en busca de alimentos, en exploración o para trabajar, puede no encontrar agua cuando la necesite y tener que llevarla con usted.

Conservar el agua. *El agua es un bien muy preciado; nunca debe desperdiciarla. Puede que no le sea tan fácil renovar su reserva.*

ENVASES PARA AGUA

Hay muchos tipos de envases para transportar y conservar líquidos fríos o calientes. Varían desde las botellas y termos de plástico o acero hasta las bolsas en varios tipos de tejido que pueden doblarse cuando están vacías. Asegúrese de que puede diferenciar las botellas de agua de aquéllas en las que transporta el combustible para la cocina.

Termo. *Un termo metálico es prácticamente irrompible y puede conservar líquidos fríos o calientes.*

Botella de plástico. *Una resistente es ideal, pero se fundirá si se pone muy cerca del fuego.*

Botella de acero. *Es una botella resistente, pero pesada si está llena de agua.*

Bolsa para la cintura. *Una bolsa para la cintura le permite transportar agua mientras se desplaza, al tiempo que le deja las manos libres.*

Cantimplora con taza. *Ésta tiene una taza de plástico como tapa, igual que un termo.*

Pequeña bolsa para agua. *Puede doblarse cuando no se utiliza. No la deje en el suelo, ya que se rompe fácilmente.*

Cantimplora plegable. *Éste es un tipo de recipiente tradicional y muy popular. Tenga cuidado de no romper el asa que une el tapón con la botella, ya que podría perder el tapón.*

Bolsa grande de agua. *Muy útil para un campamento grande, esta bolsa de agua puede colgarse de un árbol desprotegido del viento para mantener frío su contenido.*

CÓMO TRANSPORTAR AGUA

Si ir a buscar agua es una actividad diaria, y si el campamento está a cierta distancia de la fuente diaria –como se aconseja para un campamento seguro–, llene grandes recipientes de agua para después transportarlos hasta el campamento. Lo mejor es llevar el envase a la espalda, dejando las manos libres para sortear los obstáculos. Si es posible, utilice una botella pequeña para llenar el recipiente. Esto le permite aprovechar mejor el agua.

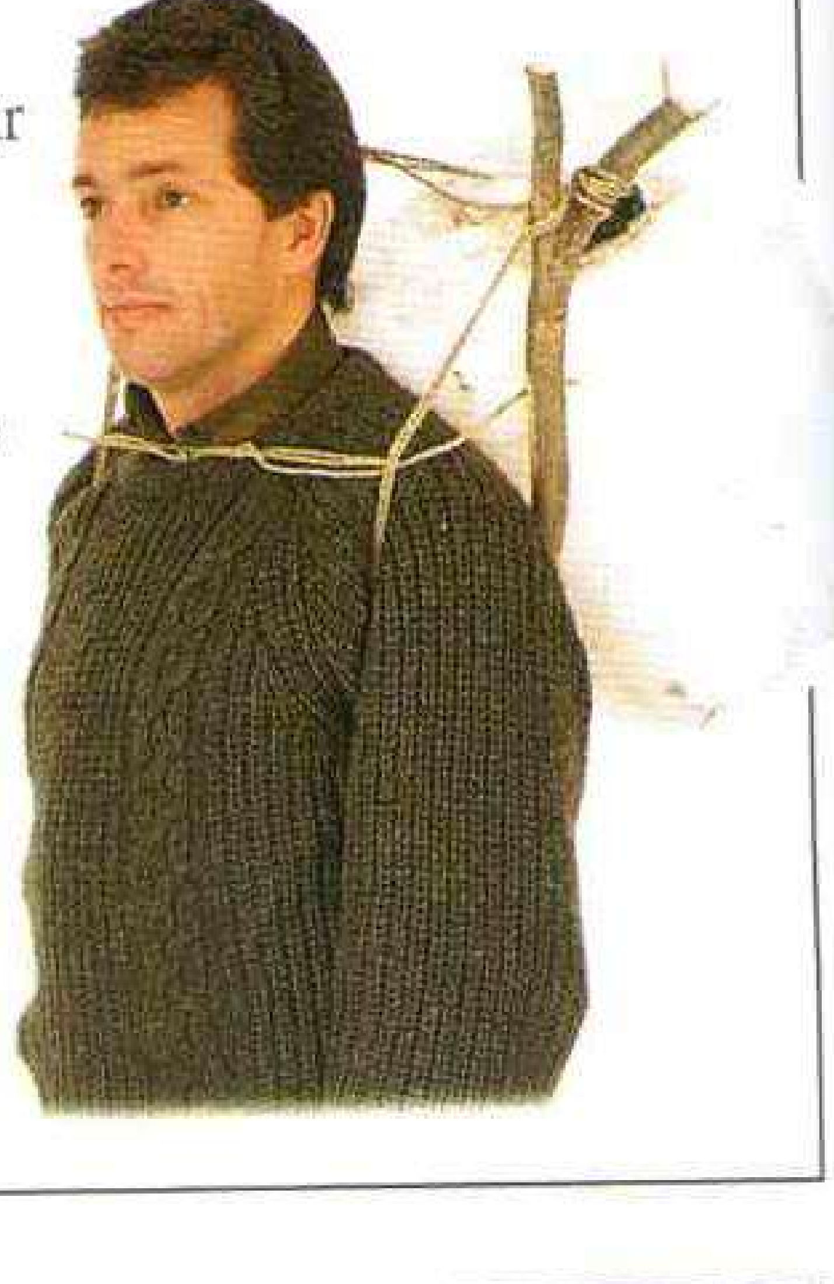

Transporte de agua. *Construya un armazón* (véase página 135) *para cargar un recipiente grande.*

CONSTRUCCIÓN DE UN RECIPIENTE DE CORTEZA

1 Se puede improvisar un recipiente para agua a partir de corteza verde y flexible, como la del abedul. Corte un trozo en forma de rectángulo.

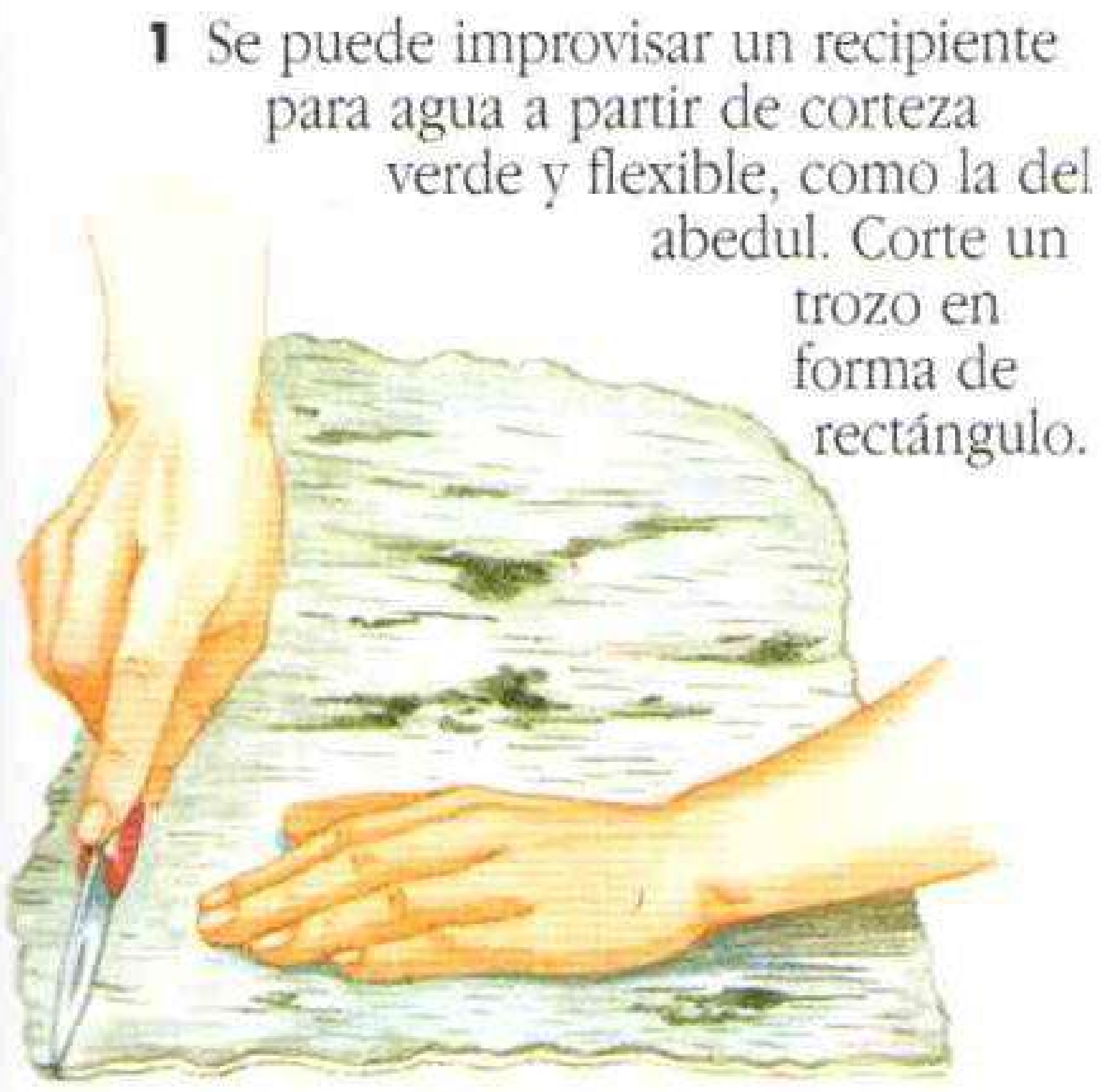

2 Sumerja las esquinas en agua para ablandar la corteza y doble los lados con suavidad.

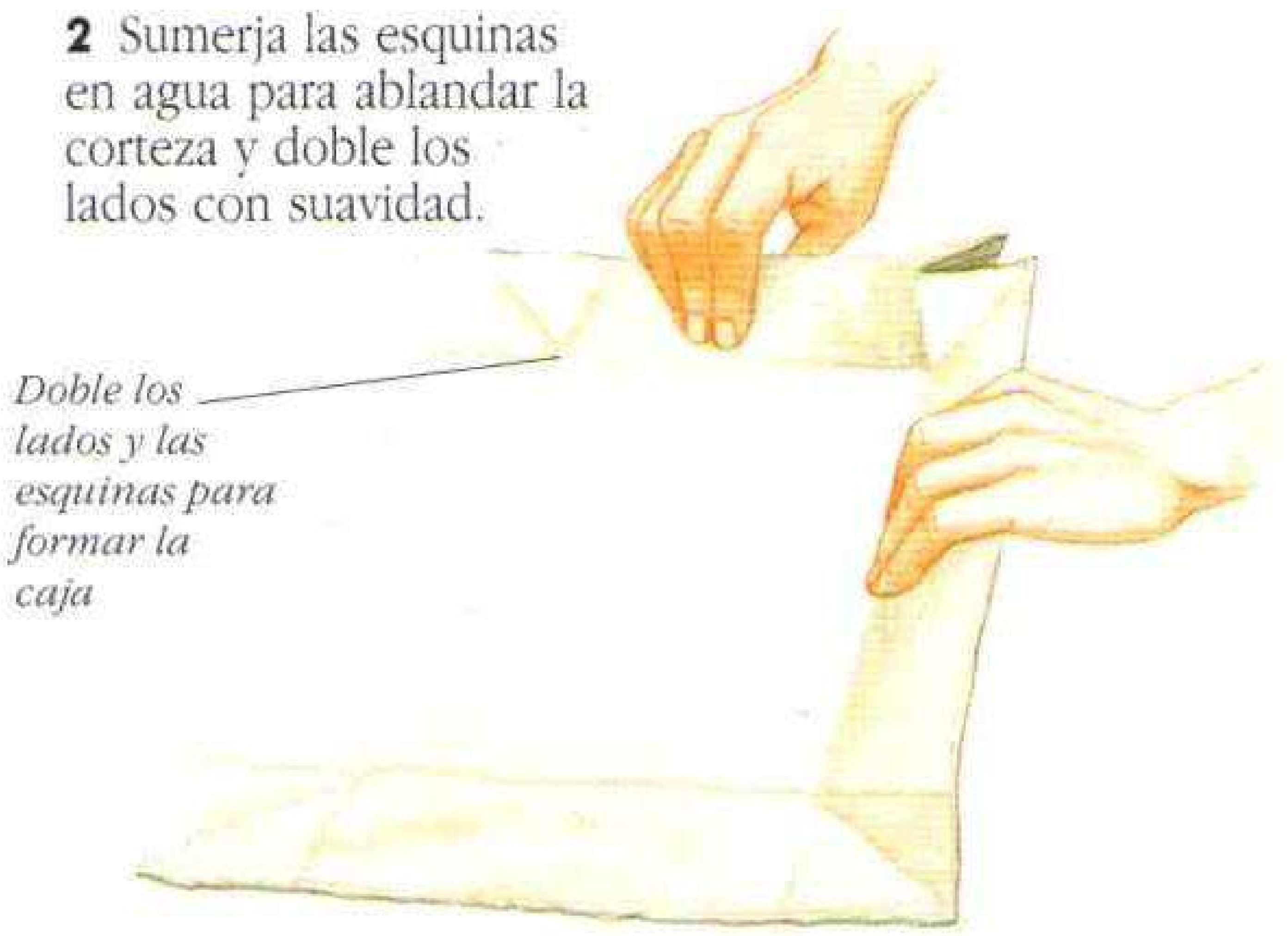
Doble los lados y las esquinas para formar la caja

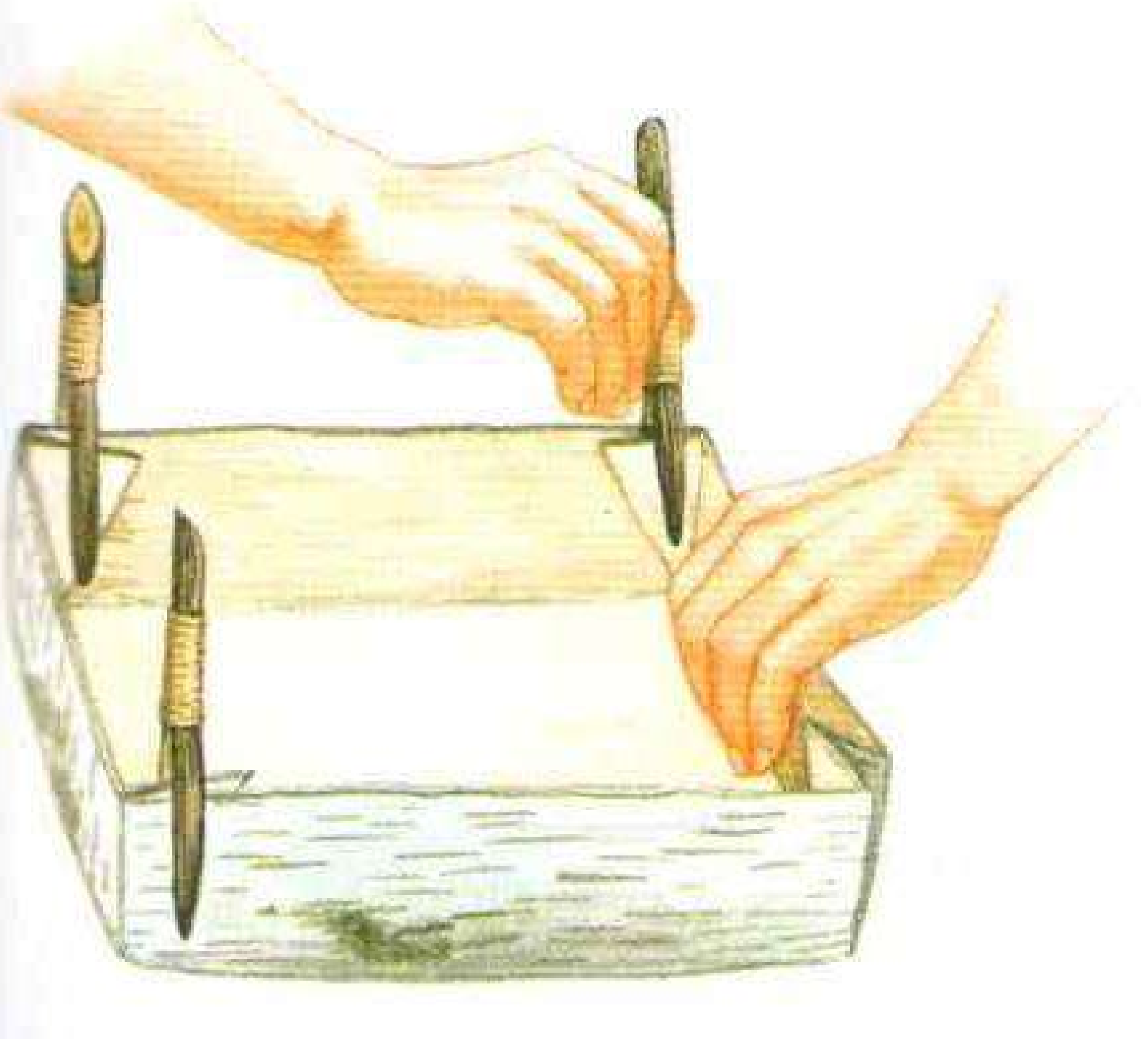

3 Construya unas pinzas a partir de ramitas partidas longitudinalmente y atadas con hierbas. Encole los lados de la caja con resina de pino, que puede desprender de la corteza de estos árboles. Deje las pinzas hasta que se seque.

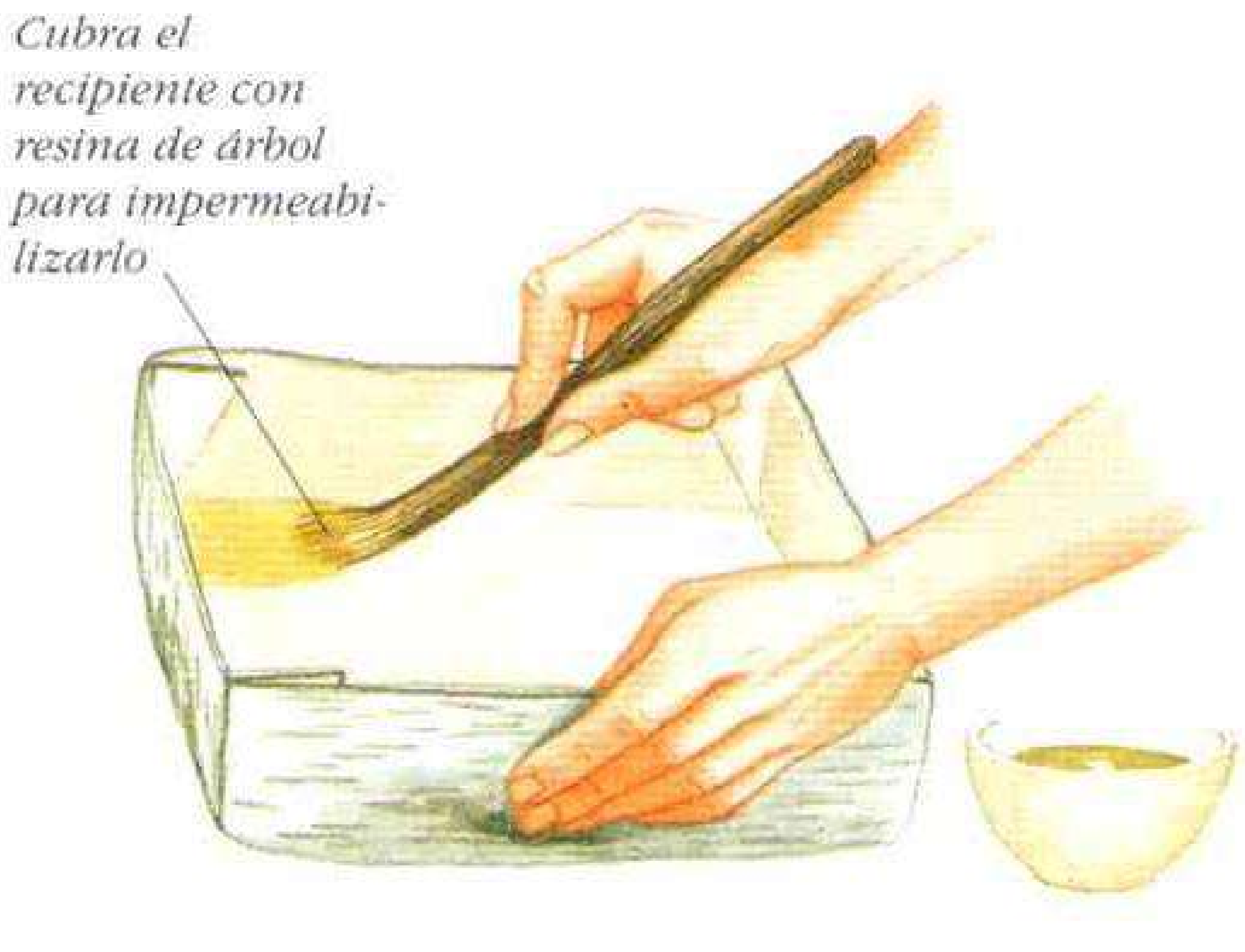
Cubra el recipiente con resina de árbol para impermeabilizarlo

4 Cuando el recipiente esté seco y retenga la forma, elimine las pinzas e impermeabilícelo con resina de pino empleando una rama astillada como pincel. Deje que la caja se seque antes de utilizarla.

RECIPIENTES IMPROVISADOS

Segmento de bambú. *Los tallos de bambú están formados por segmentos huecos divididos por «paredes» y constituyen excelentes recipientes para agua. Corte un tallo grueso de bambú unos 2 cm por debajo de una unión entre segmentos, y en el mismo lugar del segmento siguiente. Así obtendrá un recipiente hueco con un extremo abierto.*

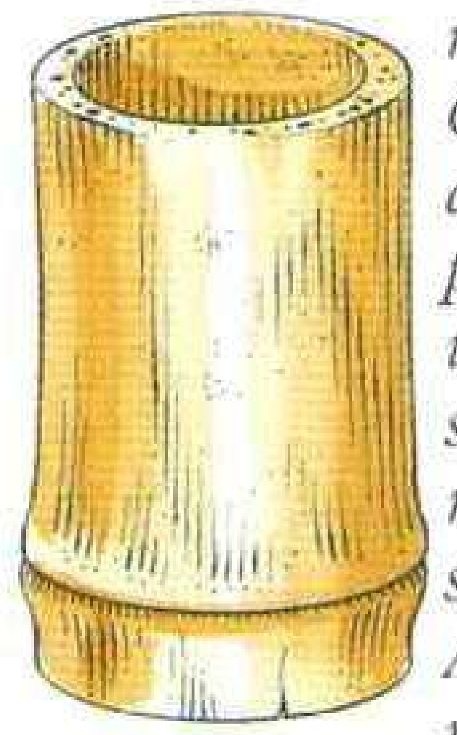

Calabazas. *Las calabazas pueden vaciarse para emplearlas como recipientes para agua. Corte la parte superior, saque toda pulpa y deje las paredes limpias; utilice un cuchillo o una rama de punta afilada.*

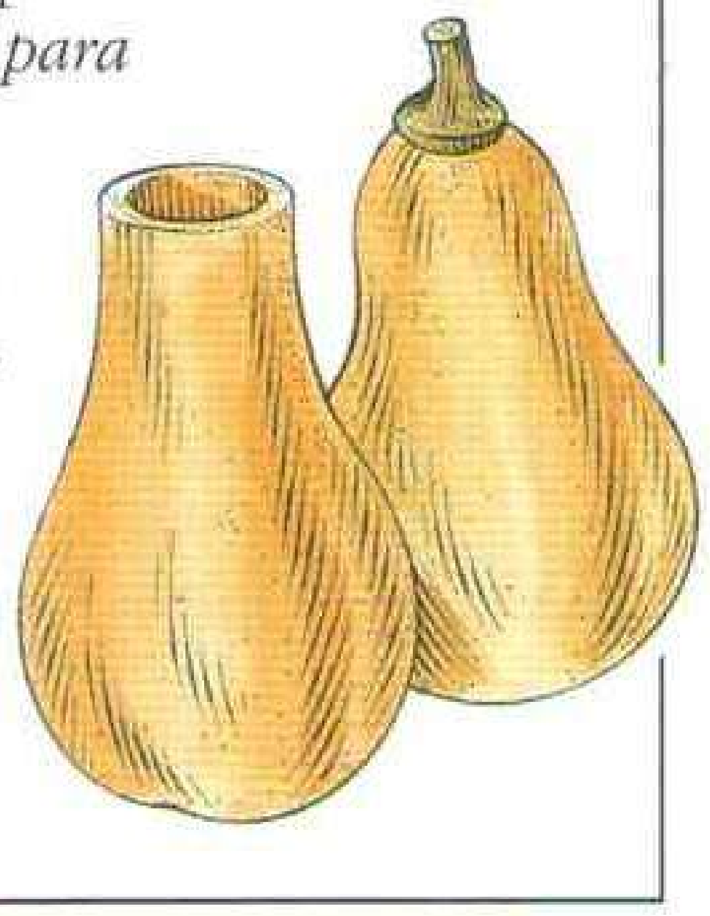

LAS REGLAS DEL AGUA

- Filtre y purifique siempre el agua recogida *(véase página 74)*.
- Intente recoger la cantidad suficiente de agua para cubrir sus necesidades y para guardar una reserva de emergencia para cuando no pueda encontrarla.
- Cuando llueva, disponga la mayor cantidad de recipientes posible para recoger el agua de la lluvia –nunca se sabe cuándo puede necesitarse.
- No la desperdicie nunca, ni cuando la tenga en exceso, ya que algún día le puede hacer falta.
- Nunca beba orina o agua de mar sin tratar. En caso de emergencia puede destilarlas *(véase página 75)*.
- Beba agua en abundancia en condiciones de frío intenso como lo haría en climas más cálidos, ya que caminar en la nieve o esquiar pueden hacerle sudar, y la falta de agua podría causarle deshidratación.
- No coma si no tiene una reserva adecuada de agua, porque necesitará agua para la digestión.

CÓMO ALMACENAR AGUA

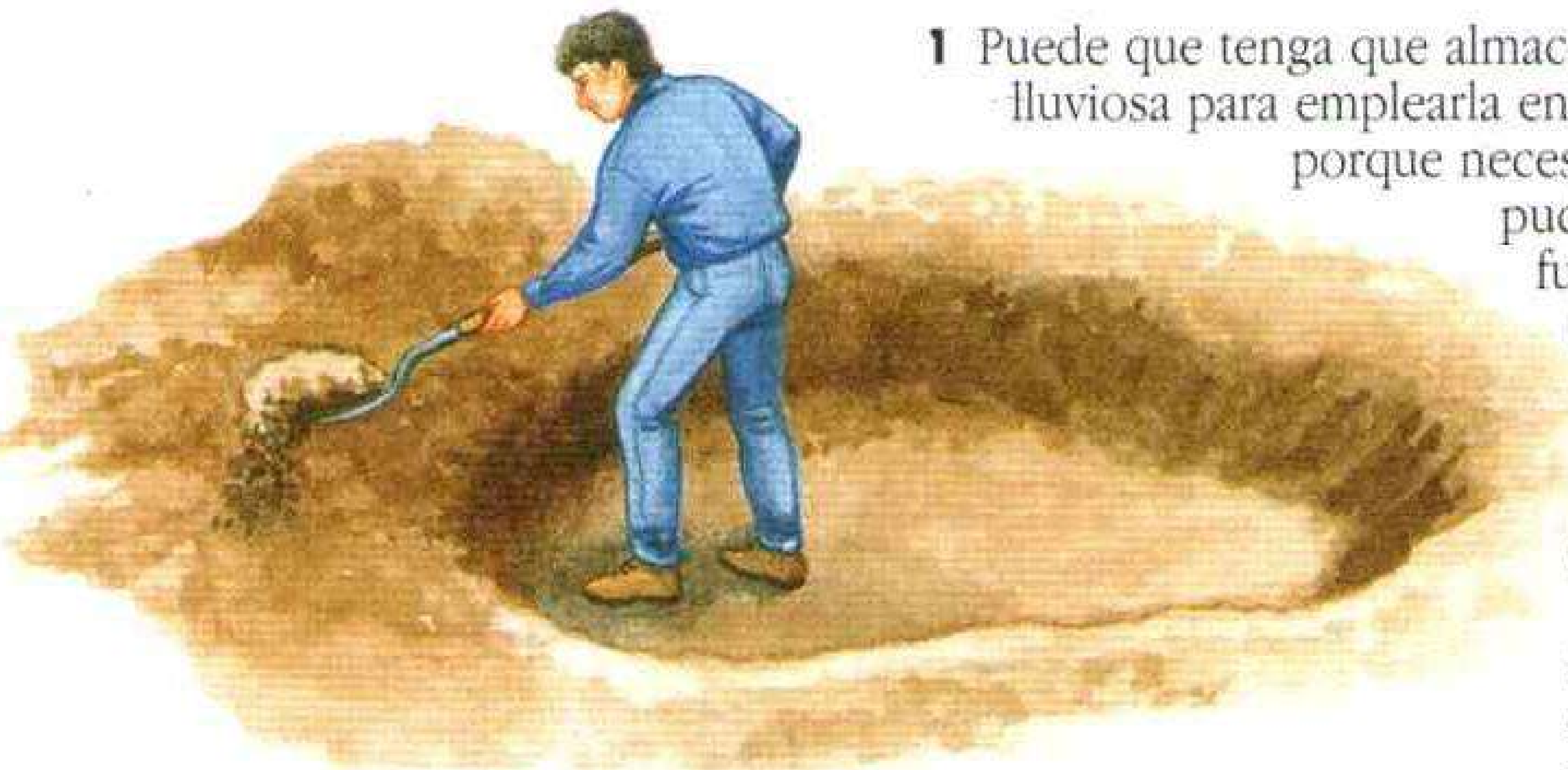

1 Puede que tenga que almacenar agua durante la época lluviosa para emplearla en los meses de sequía, o porque necesite más agua que la que pudiera transportar desde la fuente. Escoja un lugar hacia el cual fluiría el agua naturalmente, por ejemplo una pequeña hondonada, y cave hasta encontrar arcilla o piedra.

2 Cubra el hoyo con una tela impermeable o recúbralo con arcilla húmeda, alisándola hasta que se convierta en una superficie hermética. Cubra el hoyo durante el día con una cubierta de vinilo o teja una estera de vegetación *(véase página 49)* para reducir al mínimo la evaporación. Retire la cubierta de noche y cuando llueva para que la lluvia y el rocío caigan en el estanque. Recoja el agua con regularidad, antes de que se evapore.

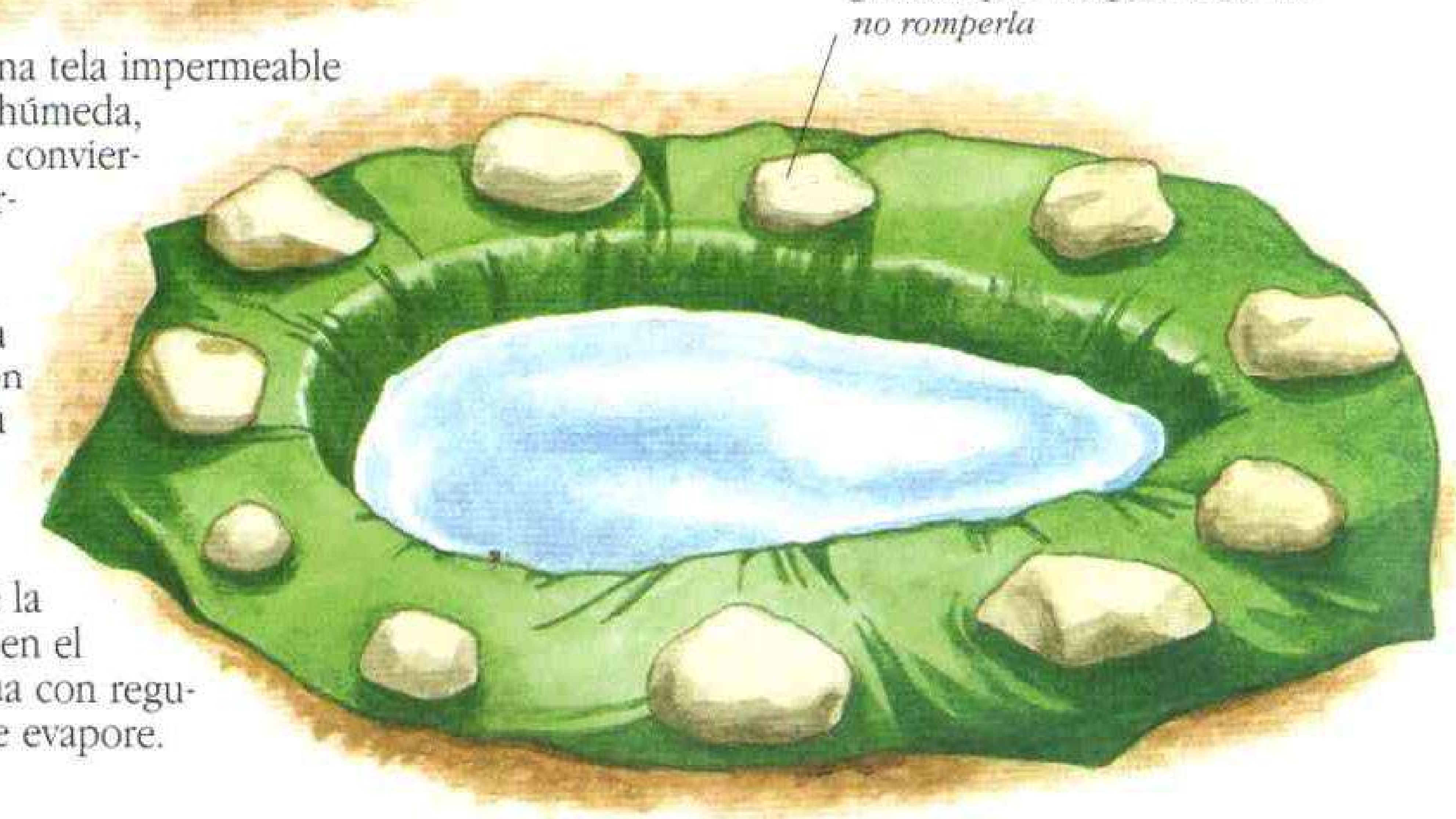
Sujete la cubierta con piedras grandes, pero tenga cuidado de no romperla

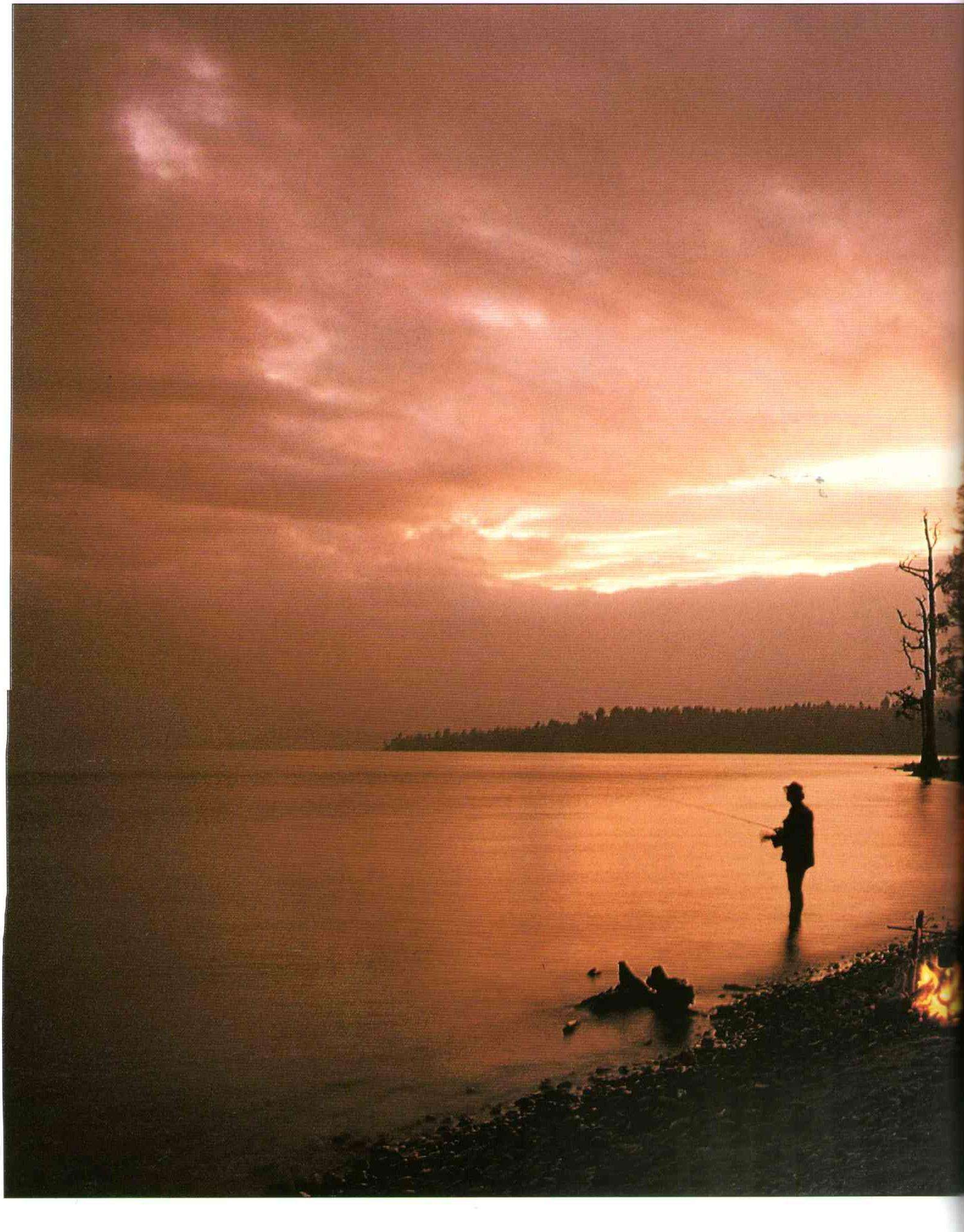

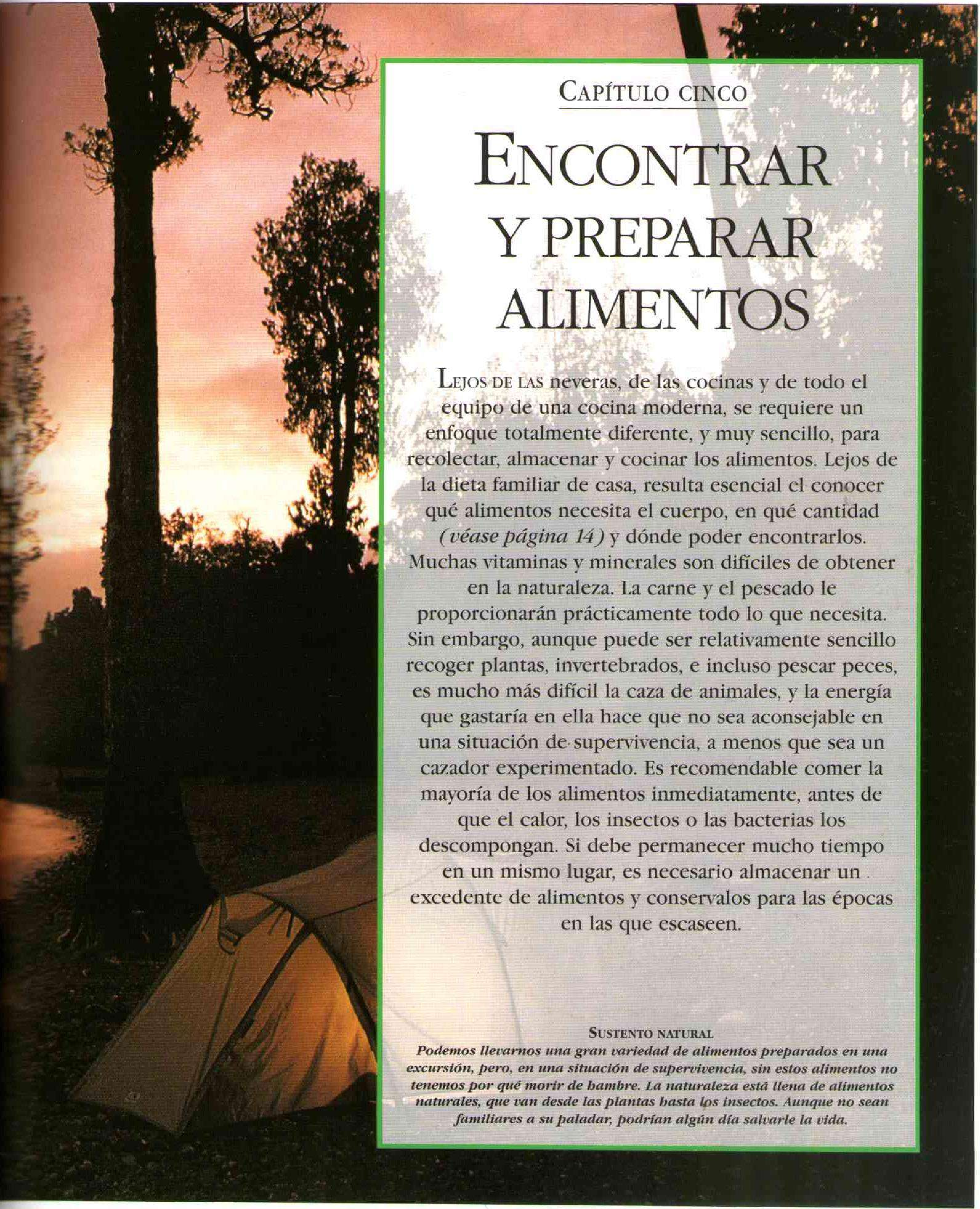

CAPÍTULO CINCO

ENCONTRAR Y PREPARAR ALIMENTOS

LEJOS DE LAS neveras, de las cocinas y de todo el equipo de una cocina moderna, se requiere un enfoque totalmente diferente, y muy sencillo, para recolectar, almacenar y cocinar los alimentos. Lejos de la dieta familiar de casa, resulta esencial el conocer qué alimentos necesita el cuerpo, en qué cantidad *(véase página 14)* y dónde poder encontrarlos. Muchas vitaminas y minerales son difíciles de obtener en la naturaleza. La carne y el pescado le proporcionarán prácticamente todo lo que necesita. Sin embargo, aunque puede ser relativamente sencillo recoger plantas, invertebrados, e incluso pescar peces, es mucho más difícil la caza de animales, y la energía que gastaría en ella hace que no sea aconsejable en una situación de supervivencia, a menos que sea un cazador experimentado. Es recomendable comer la mayoría de los alimentos inmediatamente, antes de que el calor, los insectos o las bacterias los descompongan. Si debe permanecer mucho tiempo en un mismo lugar, es necesario almacenar un excedente de alimentos y conservalos para las épocas en las que escaseen.

SUSTENTO NATURAL
Podemos llevarnos una gran variedad de alimentos preparados en una excursión, pero, en una situación de supervivencia, sin estos alimentos no tenemos por qué morir de hambre. La naturaleza está llena de alimentos naturales, que van desde las plantas hasta los insectos. Aunque no sean familiares a su paladar, podrían algún día salvarle la vida.

ALIMENTOS PARA EL VIAJE

EN UNA EXCURSIÓN puede llevarse comida enlatada, deshidratada en bolsas de aluminio, o en su estado natural en algún tipo de recipiente. Los alimentos enlatados son fáciles de cocinar y comer (fríos o calientes), pero son pesados. Los alimentos deshidratados son mucho más ligeros, pero necesitan una gran cantidad de agua para rehidratarlos, y en algunas regiones se vería obligado a transportar agua con este único propósito. Debe resistirse a la tentación de comer un alimento antes de rehidratarlo totalmente, ya que absorbería el agua de su cuerpo causando obstrucciones intestinales molestas o incluso peligrosas.

Alimentos altos en calorías. *Los dulces tienen un bajo contenido nutricional, pero son siempre bienvenidos. Los montañeros y exploradores polares obtienen a menudo la gran cantidad de calorías necesarias para la supervivencia de los dulces* (véase página 15).

CHOCOLATE EN POLVO

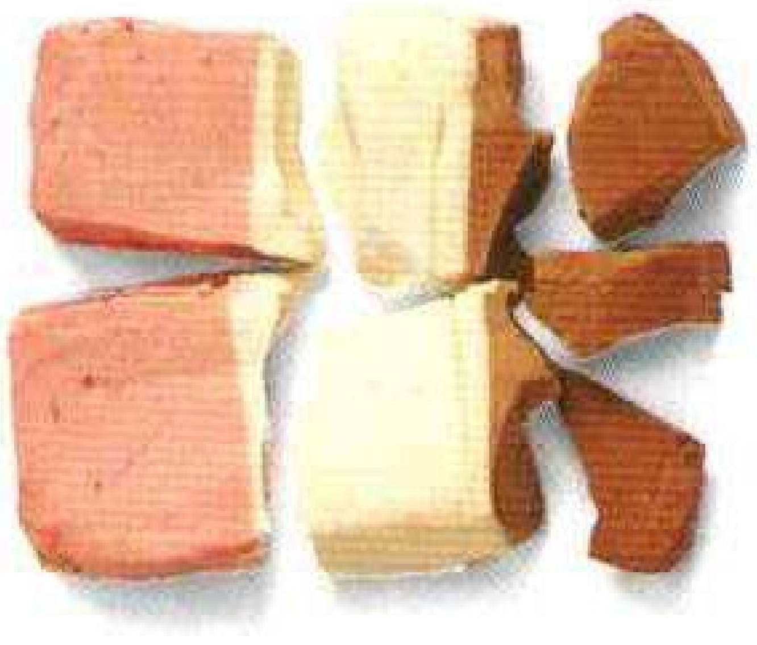

HELADO DESHIDRATADO

PREPARADO PARA FLAN DE FRESA

CARAMELOS CON SABOR A FRUTAS

Alimentos para el desayuno. *Los alimentos para el desayuno son fundamentales para obtener energía al comienzo del día. Los cereales como la avena y el* muesli *también aportan fibra que evita las obstrucciones del tracto digestivo.*

NUECES Y SIMILARES

MUESLI

FRUTOS SECOS

AVENA

Granos y legumbres. *El arroz proporciona los hidratos de carbono necesarios, pero puede ser difícil de preparar en una cocina de campamento. Las judías y las legumbres proporcionan fibra y proteínas. Deben estar bien remojadas antes de cocinarlas, en especial las judías. Siga las instrucciones del fabricante.*

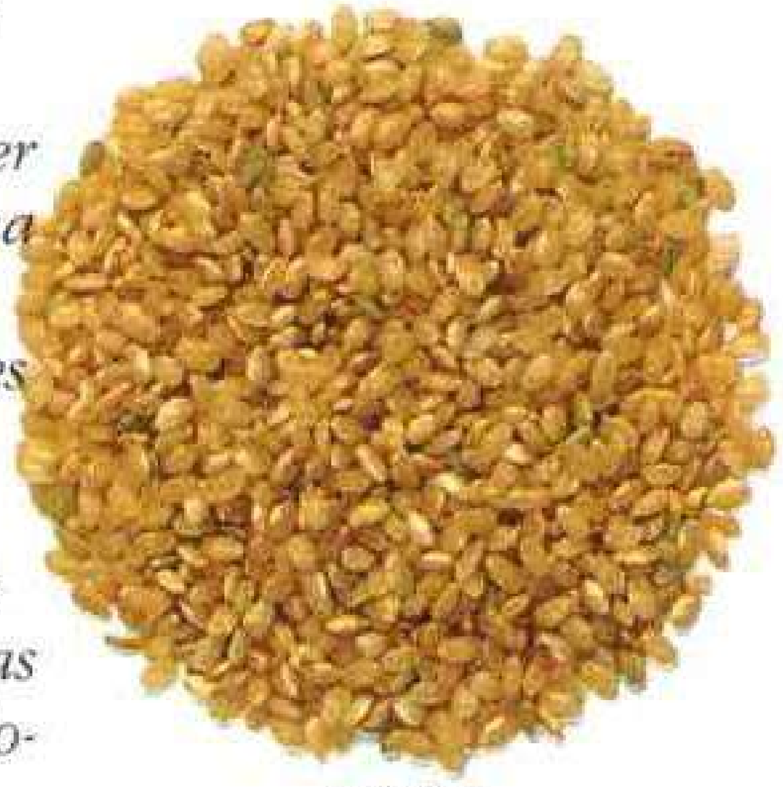

ARROZ

LENTEJAS

JUDÍAS

GUISANTES SECOS

Tentempiés. *Para mantener la energía, lo mejor es ir picando continuamente durante el día para evitar las punzadas de hambre, y hacer únicamente una comida importante al día, que pueda digerir totalmente durante la noche. Estos tentempiés incluyen caramelos, chocolate y galletas. Ellos aportan energía e hidratos de carbono.*

GALLETA DE MANTEQUILLA

GALLETA DE CHOCOLATE

GALLETAS DIGESTIVAS

BARRA DE GRANOLA

TABLETA DE CHOCOLATE

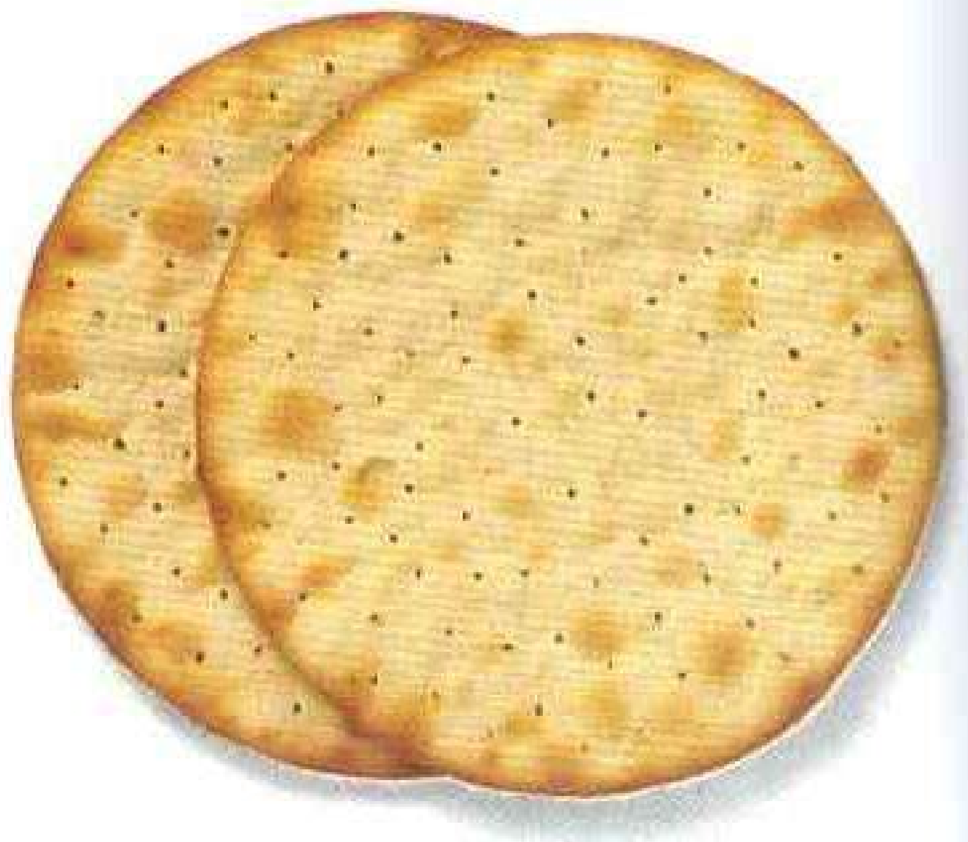

GALLETAS SALADAS

Bebidas. *Las bebidas son un lujo que aporta poco contenido nutricional. Sin embargo, proporcionan una sensación de calor y comodidad. La leche en polvo puede ser una fuente de calcio y da mejor sabor a las bebidas. El chocolate a la taza es una bebida deliciosa en una tarde fría.*

Café

Té

Chocolate a la taza

Leche en polvo

Comidas principales. *La liofilización es un método de deshidratación que conserva intacta la estructura y la textura de los alimentos, en especial la de la fruta. Los alimentos deshidratados son ligeros y fáciles de transportar, pero deben rehidratarse en agua antes de cocinarlos.*

Harina de soja

Comida liofilizada

Comida deshidratada

Comida vegetariana deshidratada

Condimentos. *La harina y la sal son artículos de primera necesidad; a partir de ellos pueden prepararse una gran variedad de platos* (véase página 117). *El sebo puede ser de gran valor si no puede obtener las grasas necesarias a partir de su comida. El azúcar hará más gustosa la comida silvestre.*

Harina

Sebo

Azúcar moreno

Sal

Condimentos. *Cualquier condimento que realce el sabor de una comida de campamento vale la pena de llevar* (véase página 93). *El curry en polvo da un sabor distinto a los alimentos, al igual que el ajo, las cebollas, las especias, la salsa de tomate y demás salsas.*

Cubos de caldo de carne

Salsa de cebolla

Sopa

Salsa de tomate

CARNE Y PESCADO

Aunque la carne roja y el pescado proporcionan las proteínas necesarias, cualquier tipo de carne es difícil de conservar fresca. Lo mejor es llevarla enlatada o en conserva. Una vez las latas se han abierto, su contenido debe consumirse inmediatamente, por lo que debe comprar latas que se puedan consumir en una sola comida. Los alimentos enlatados pueden ser, sin embargo, pesados para llevar en una mochila. Las carnes curadas, como el salami, son más ligeras, pero su variedad es limitada. Puede conservar su propia carne o pescado de muchas maneras *(véase página 118)*. Una alternativa a los alimentos enlatados o curados es una selección de alimentos completos preparados a base de proteína de soja o *tofu*, así como la pasta y los vegetales.

Sardinas

Pasta

Salchichas y judías

Salami

PLANTAS COMESTIBLES

A MENOS QUE se encuentre en medio de un desierto árido, siempre habrá plantas a su alrededor, y muchas de ellas serán comestibles y nutritivas. El problema radica en conocer cuáles de estas plantas son nutritivas y en evitar las venenosas o las que podrían causarle alguna enfermedad. Algunas plantas nutritivas tienen finas vellosidades o aristas que producen irritaciones en la boca y en la garganta, o bien sus semillas son venenosas. Más aún, algunas plantas no son comestibles —o incluso son venenosas— en determinadas épocas del año. Algunas de ellas son comestibles pero no muy nutritivas; su recolección e ingestión consumen más calorías de las que realmente aportan.

PLANTAS DE CLIMA TEMPLADO

Evite las plantas y brotes viejos, que son fibrosos y duros, y pueden tener un sabor desagradable. Los brotes jóvenes son mucho mejores que los viejos, e incluso pueden ser comestibles sin necesidad de cocinarlos. También debe elegir las hojas jóvenes en lugar de las viejas, y debe cortarlas enteras. Aquí encontrará algunos ejemplos de plantas comestibles que crecen en climas templados.

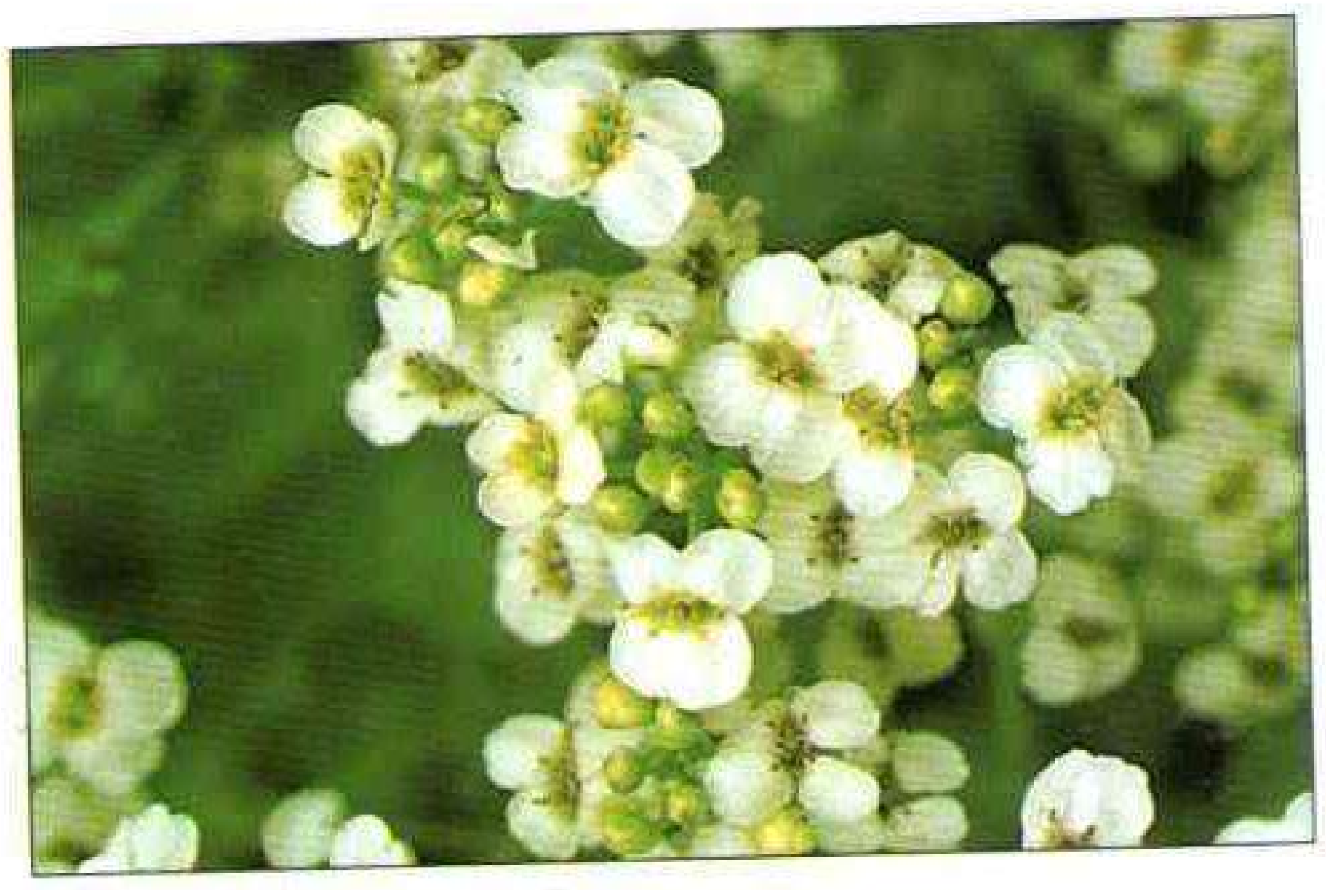

Col marina (Crambe). *La col marina crece en las costas templadas. Sus gruesas hojas y tallos subterráneos pueden comerse hervidos; su sabor recuerda al de la col.*

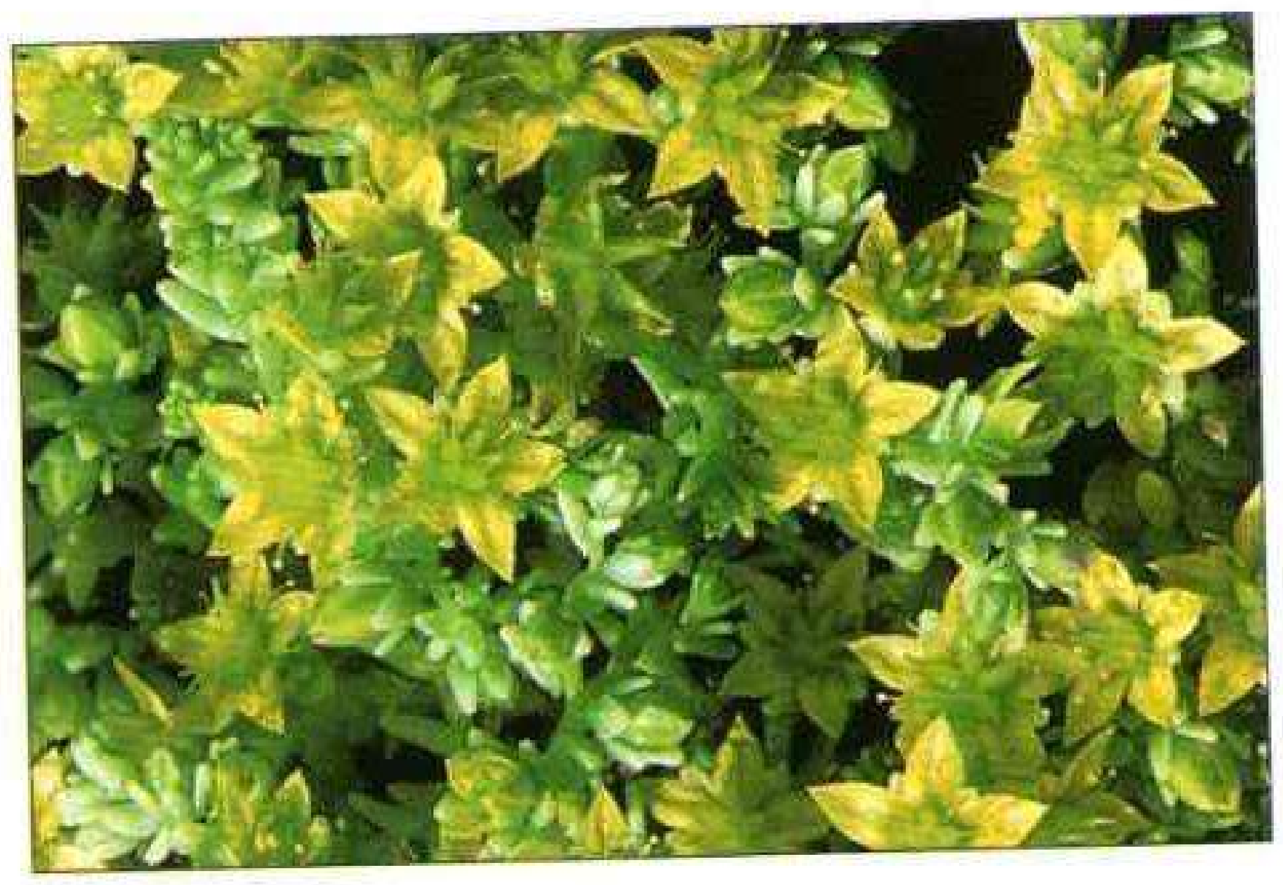

Pimienta de muros (Sedum). *La pimienta de muros crece en rocas, paredes y guijarros en las regiones templadas. Tiene un fuerte sabor a pimienta. Sus hojas pueden comerse crudas como ensalada, o puede cocerlas para utilizarlas para dar sabor a una sopa.*

OTRAS PLANTAS DE CLIMA TEMPLADO

Algunas algas como las ovas *(Porphyria)* pueden ser hervidas, horneadas o secadas. El musgo de Islandia *(Chondrus crispus)* produce una especie de gelatina que puede utilizar para preparar un postre *(véase página 93)*. Las hojas del diente de león *(Taraxacum)* y de la acedera *(Rumex)* pueden cocinarse como espinacas, mientras que el ajo silvestre *(Allium)* puede utilizarse como condimento.

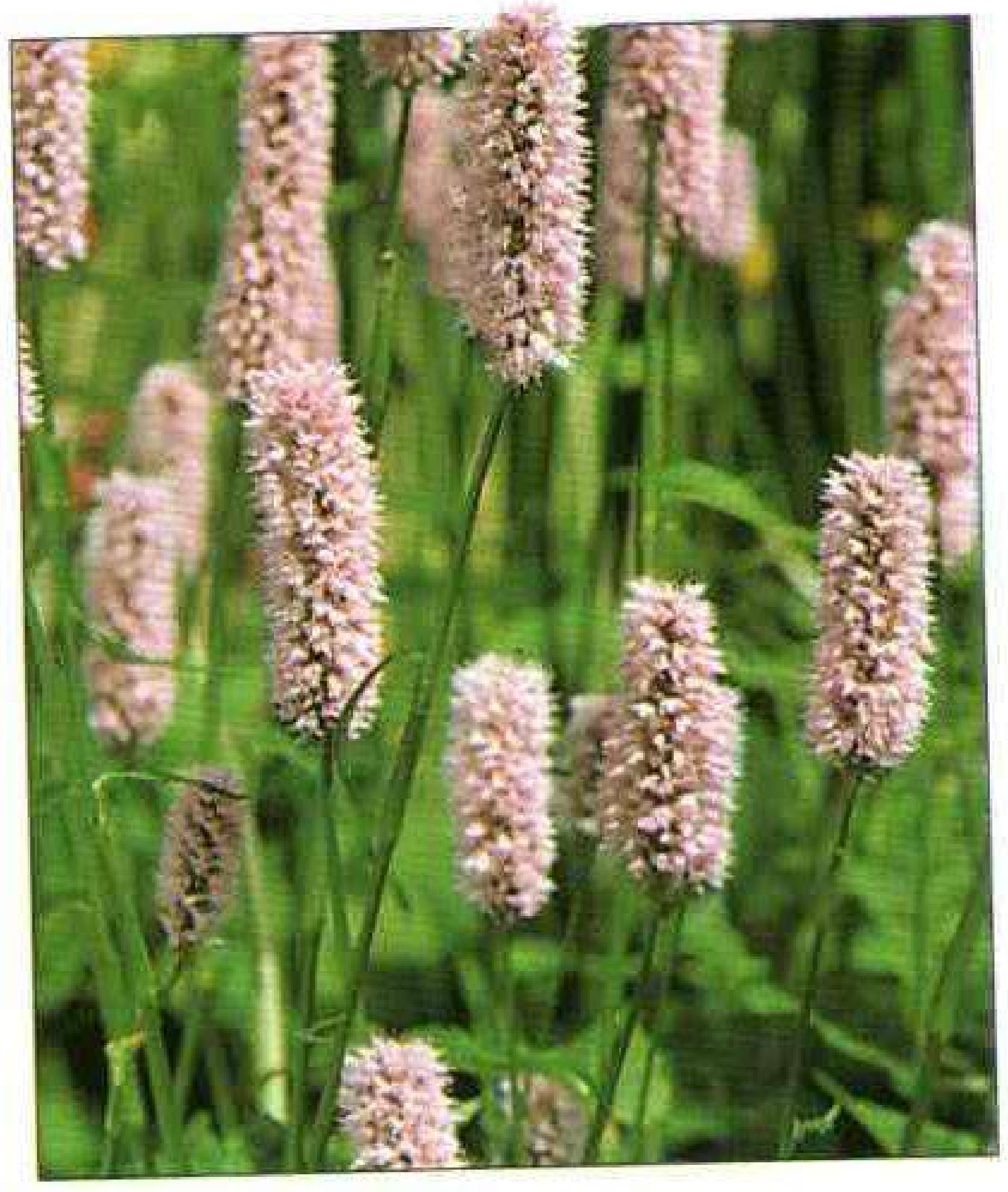

Bistorta (Polygonum). *Esta planta se encuentra en toda Europa, y algunas especies afines en Norteamérica. Crece hasta una altura de 60 cm en áreas herbosas o en bosques de zonas templadas, y es fácilmente reconocible por su larga cabeza floral de color rosa. Las hojas y brotes pueden comerse hervidos. Las raíces pueden comerse asadas, pero primero deben remojarse en agua.*

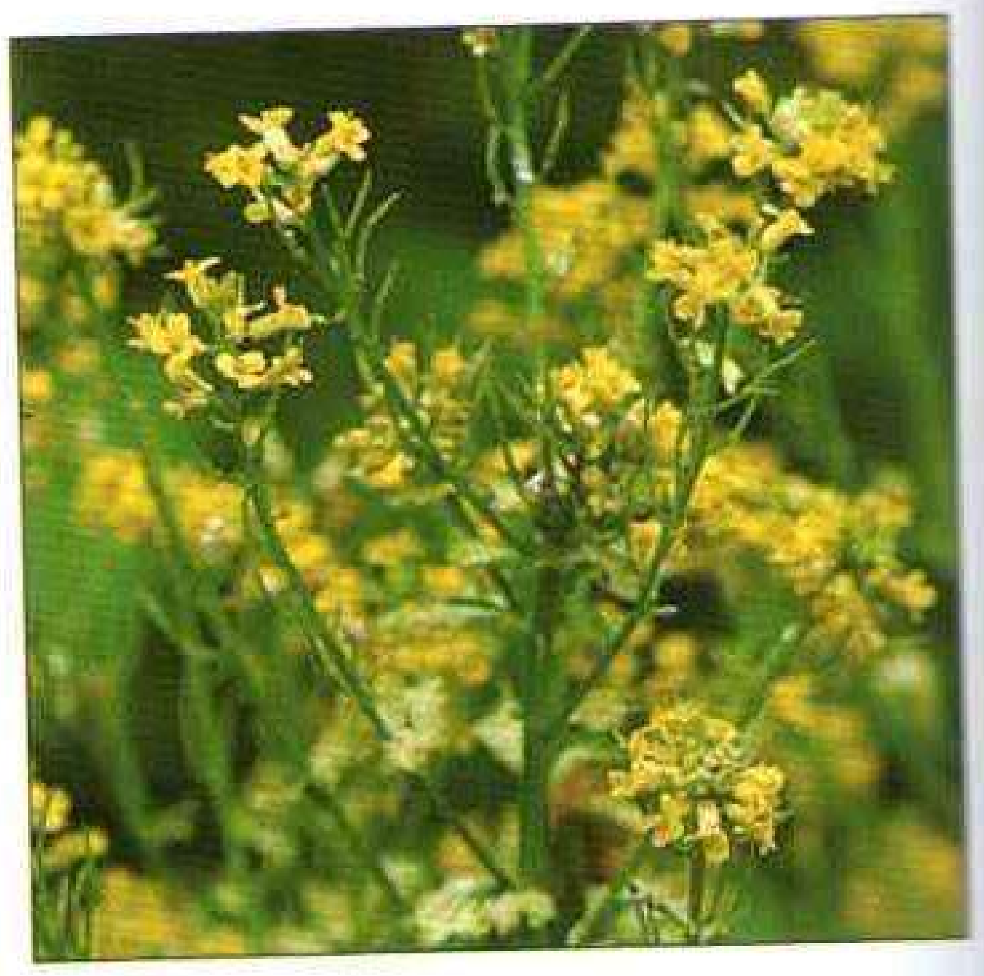

Berro terrestre (Barbarea). *Esta planta crece en Europa, Norteamérica y Nueva Zelanda. Las hojas pueden comerse crudas o hervidas.*

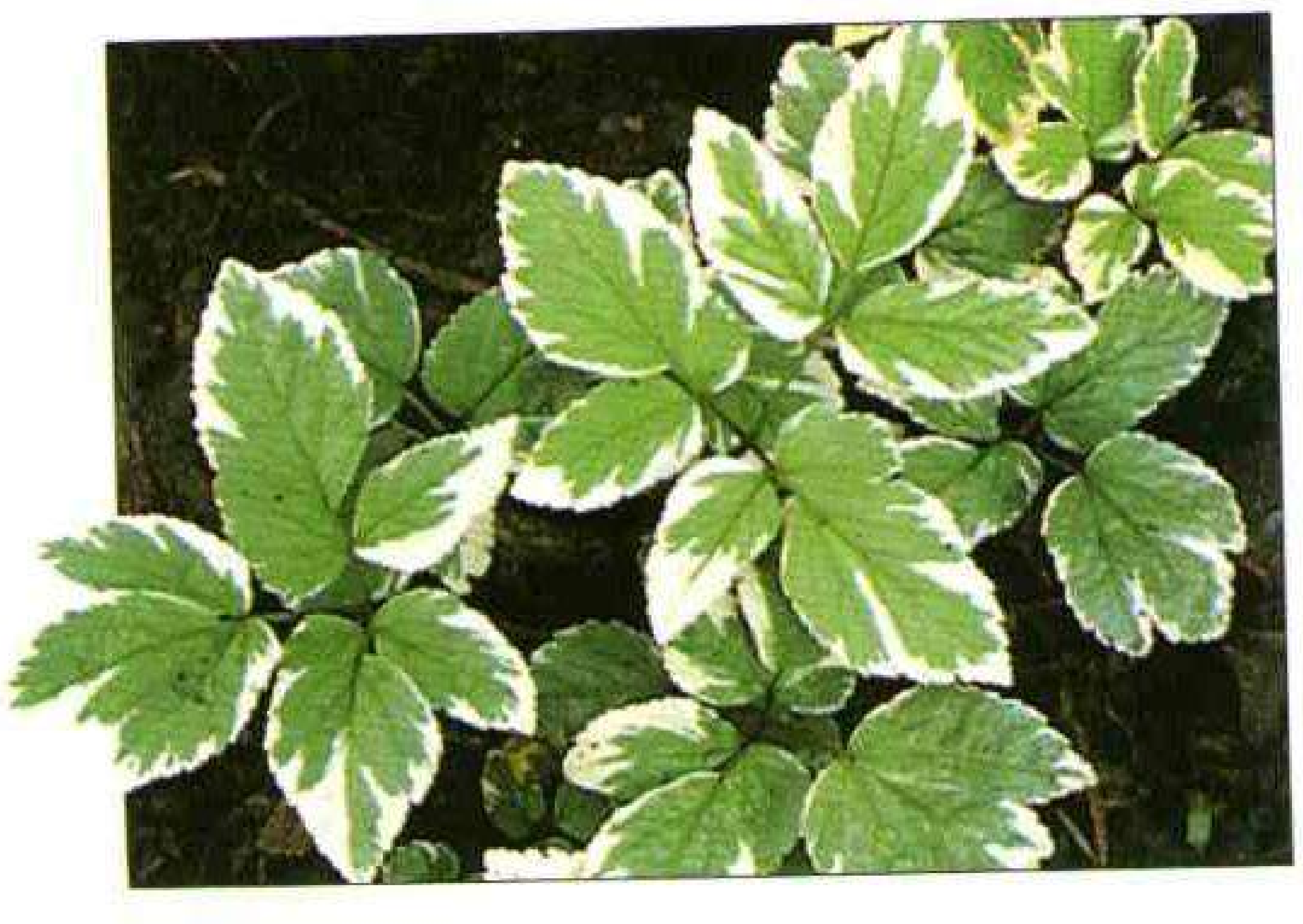

Saúco trepador (Aegopodium). *Esta hierba se encuentra a lo largo y ancho de Europa y Asia. Las hojas jóvenes y los tallos frondosos adquieren su mejor sabor cuando los brotes tienen unos 15 cm de longitud.*

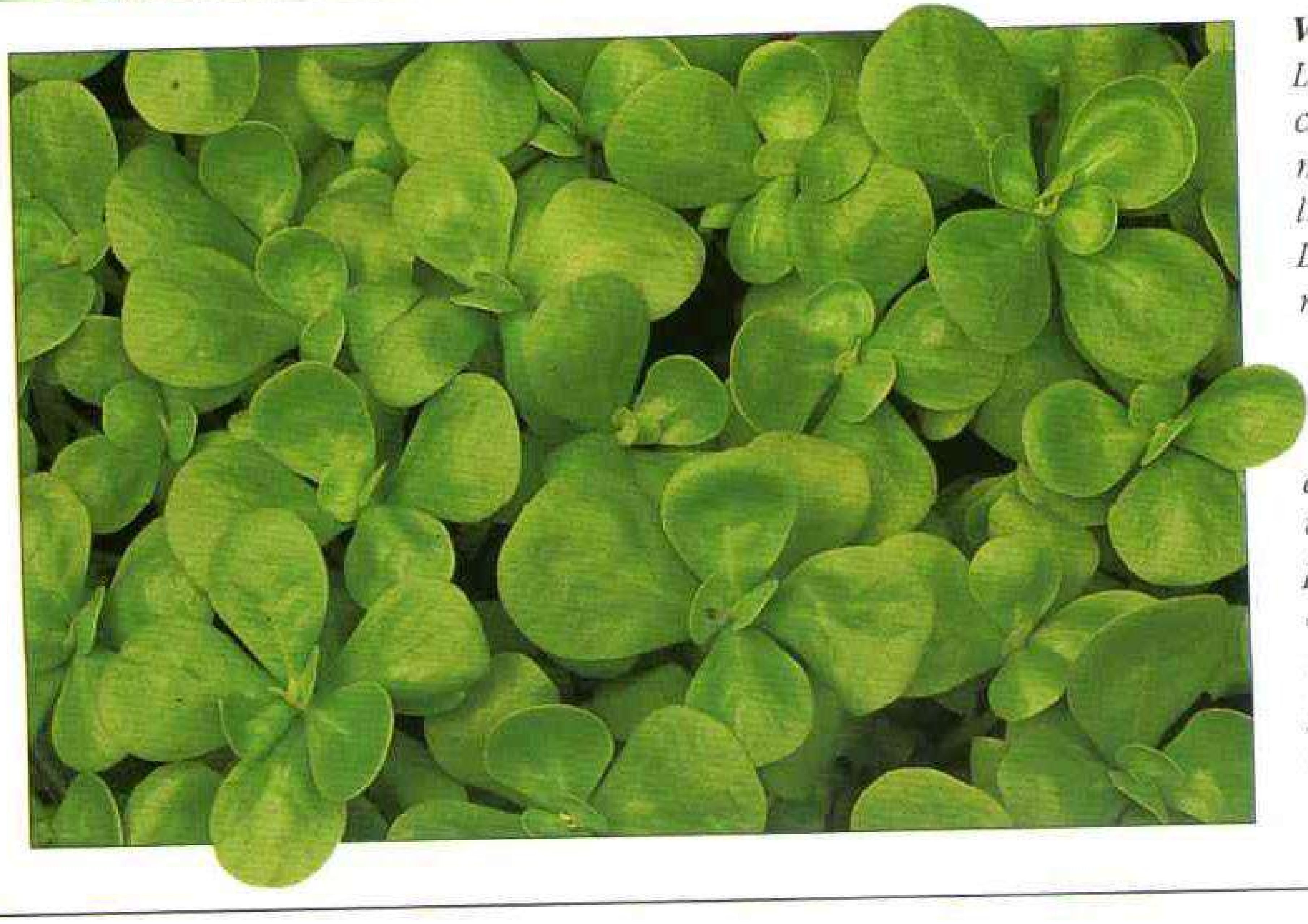

Verdolaga (Portulaca). *La verdolaga crece en climas templados, en las marismas debajo de la línea de la marea alta. Las hojas pueden recolectarse en cualquier época del año. Hiérvalas en agu y sazónelas con zumo de limón. Coma la verdolaga con otras plantas o alimentos, ya que si únicamente come un tipo de alimento, puede caer en un estado de desnutrición.*

PLANTAS TROPICALES

Con el calor y la humedad de los trópicos se desarrollan todo tipo de plantas. Con la ausencia de estaciones, crecen durante todo el año. Existen muchas variedades diferentes, siendo particularmente nutritivos muchos frutos y hortalizas. Pruebe las plantas que no le sean familiares antes de comerlas. Aquí se describen algunas especies comunes encontradas en zonas tropicales.

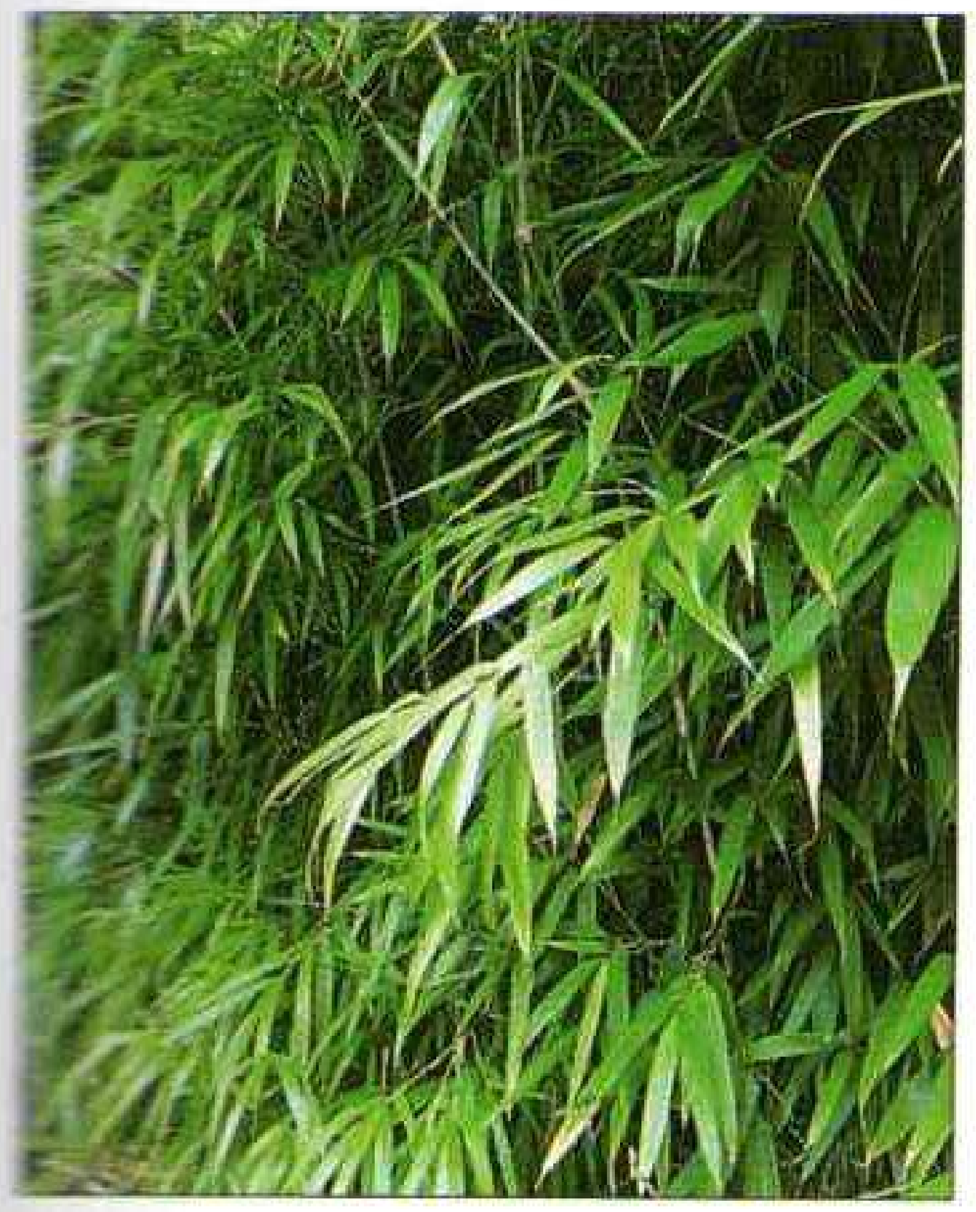

Bambú. (Pseudosasa). *Corte los tallos jóvenes de bambú desde la base de la planta y abra el revestimiento exterior con un cuchillo. La tierna pulpa interior puede hervirse en agua como si fueran espárragos. También puede comer las semillas. Existen muchos tipos diferentes de bambú, distribuidos en el mundo entero.*

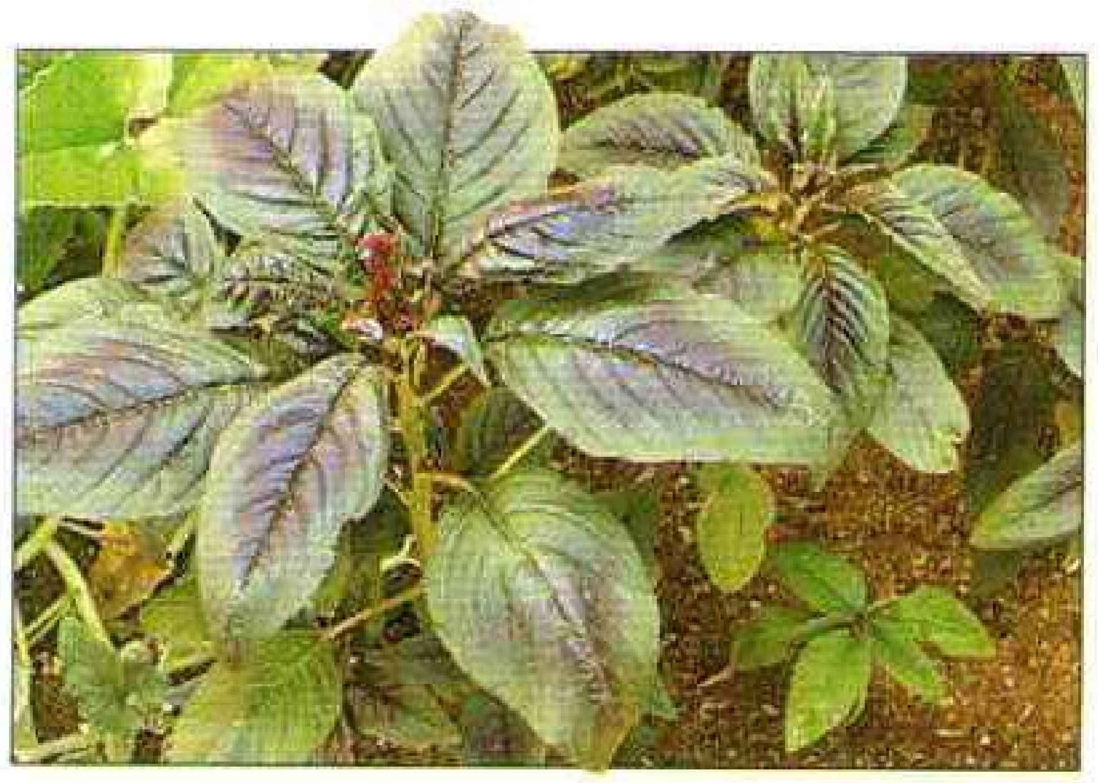

Amaranto (Amaranthus). *Esta planta puede alcanzar hasta 1 m de altura. Corte las hojas y los tallos en pequeños trozos y hiérvalos en agua con sal como si fueran espinacas. Los brotes jóvenes pueden comerse crudos en ensalada.*

OTRAS PLANTAS TROPICALES

Las especies de plantas varían de región a región, y cientos de ellas son comestibles. Muchas tienen frutos deliciosos, aunque será incapaz de alcanzarlos si crecen en la parte más alta de una selva tropical. Las plantas que pueden alcanzarse incluyen la espinaca de Ceilán *(Basella)*, cuyos tallos y hojas son ricos en vitaminas, y la caña de azúcar *(Saccharum)*, cuyos tallos pueden masticarse crudos.

PRUEBA DE SABOR

En primer lugar triture una hoja. Si huele mal o a almendras, deséchela. Frote la savia en la parte interior de su brazo. Si no le produce irritación, coloque un trozo pequeño en sus labios, después en la comisura de la boca, a continuación en la punta de la lengua, y finalmente bajo la lengua; cada vez durante cinco segundos. Si no es urticante, trague una pequeña cantidad y espere cinco horas. Si no se presenta ninguna reacción desagradable, puede comerla.

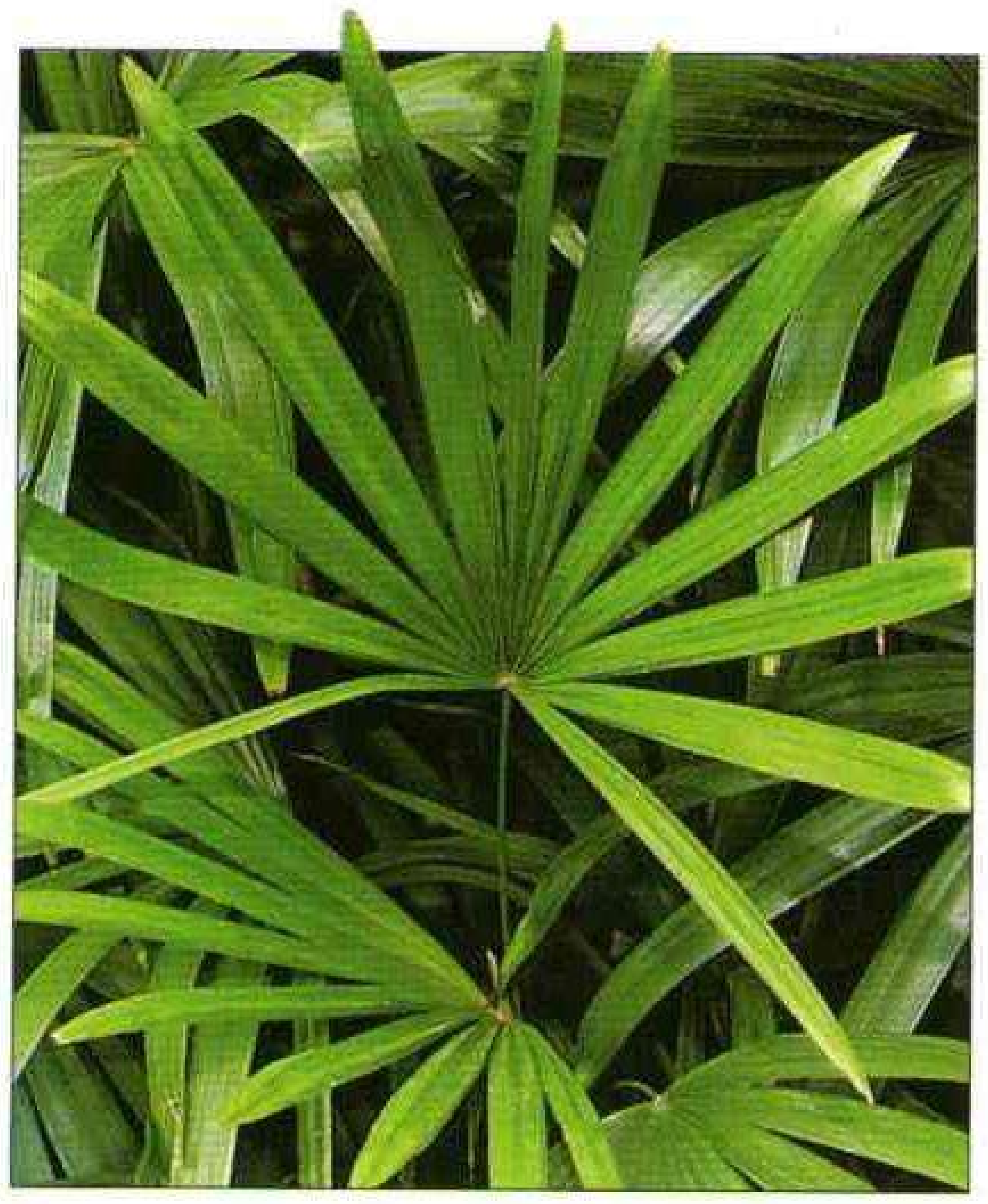

Palmas. *Los brotes jóvenes de algunas palmas, como la del cocotero* (Cocos), *la palma de azúcar* (Arenga) *y la palma sagú* (Metroxylon) *son comestibles. La palma sagú también tiene una médula feculenta en el interior del tronco que puede hervirse para preparar un pudín de sagú, semejante al de arroz. Sin embargo, no todas las palmas son comestibles, debe aprender a reconocerlas.*

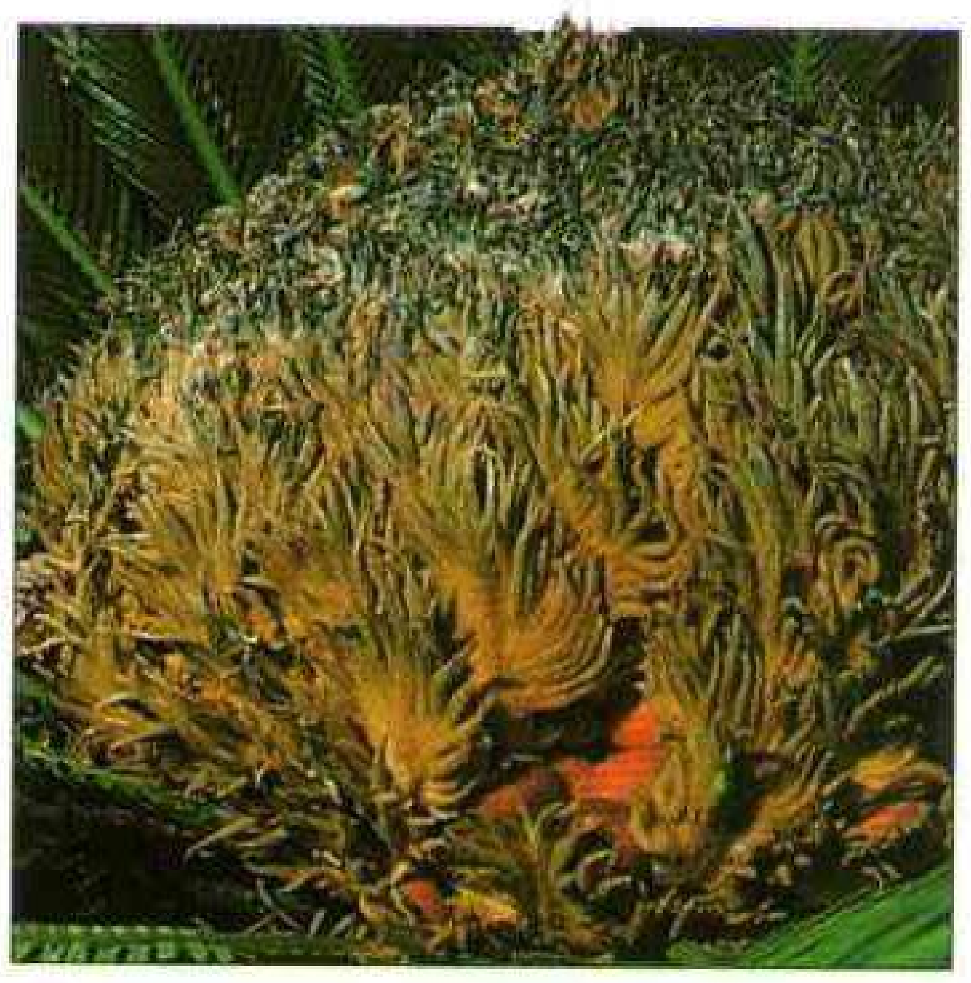

Corazones de palma. *La punta o «corazón» de algunas palmas es comestible, tanto cruda como cocida. No coma frutos de palma a menos que pueda identificarlos.*

ÁRBOLES

La mayoría de los árboles producen frutos, bayas o nueces de algún tipo, que pueden ser muy nutritivos *(véase página 88)*. Los brotes, yemas y corteza interior de muchos árboles también son comestibles, y la savia de algunos árboles puede beberse. Las agujas de algunos árboles perennes son ricas en vitamina C, y pueden remojarse en agua caliente para hacer té *(véase página 93)*.

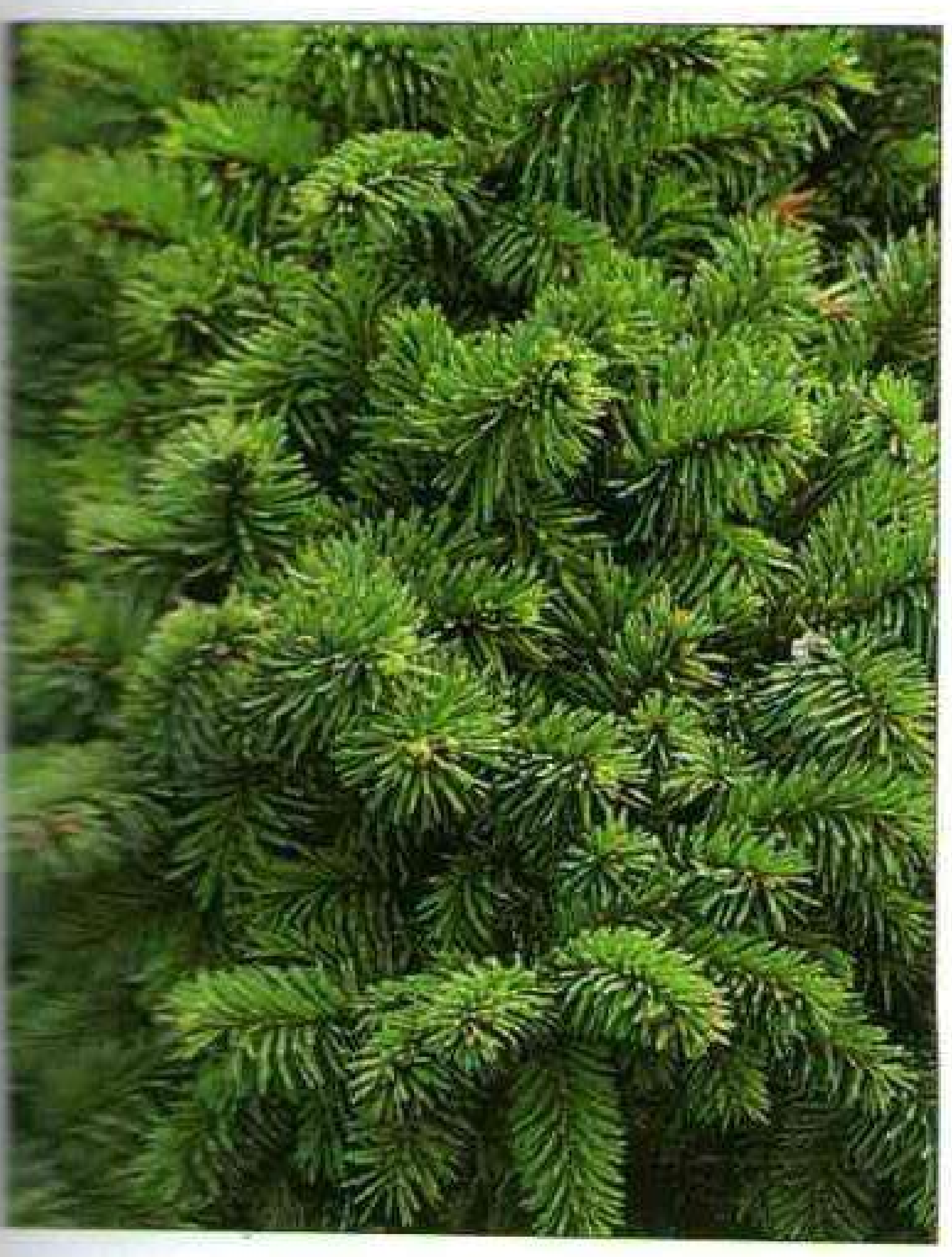

Pícea (Picea). *La corteza interior de la pícea es rica en vitamina C, especialmente en la parte inferior del tronco, cerca de las raíces. Pele la corteza exterior y hierva la corteza interior hasta que esté tierna, antes de comerla. Utilice las agujas para preparar una bebida refrescante* (véase página 93).

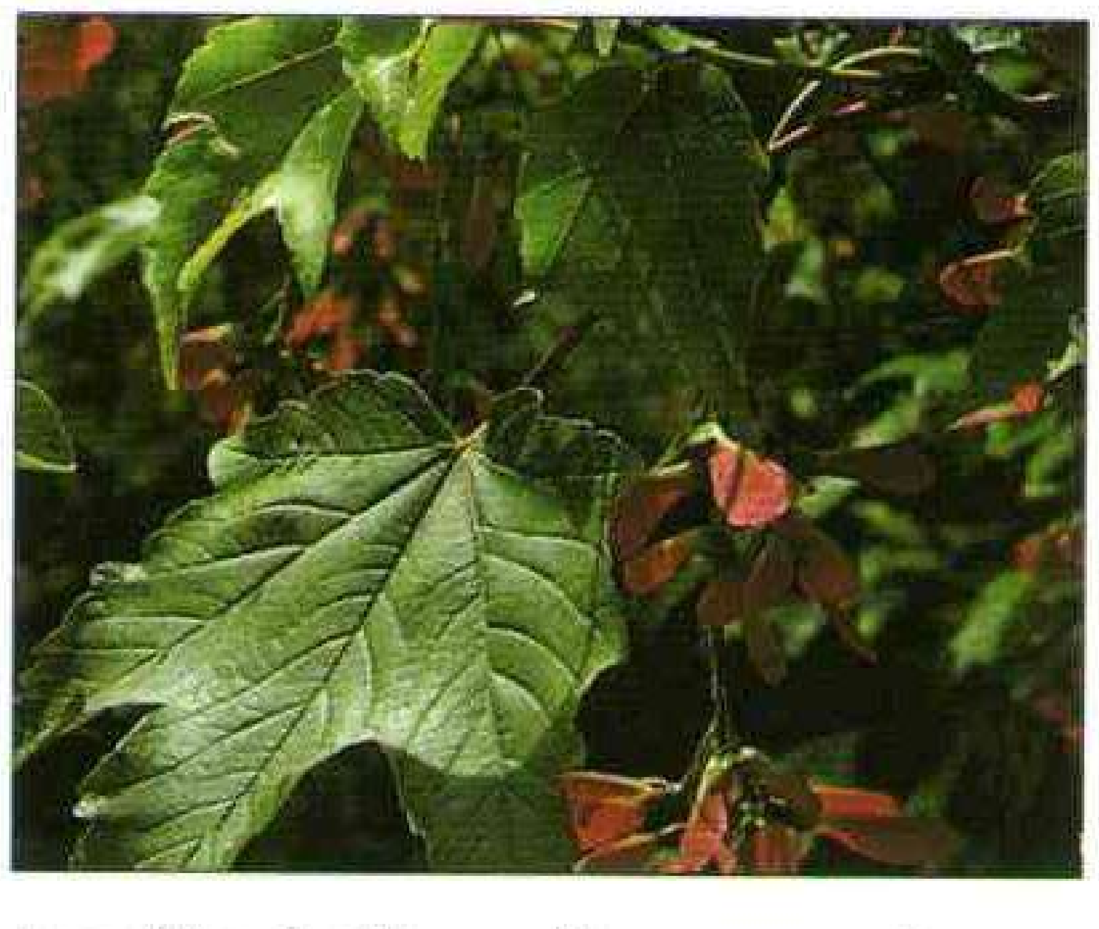

Arce (Acer). *El arce tiene una savia azucarada que puede recogerse de las aberturas naturales, o realizando un corte en la corteza. Déjela hervir hasta que se espese para formar un jarabe rico en energía.*

OTROS ÁRBOLES

Las agujas jóvenes del pino *(Pinus)* tienen un buen sabor y pueden utilizarse para hacer té *(véase página 93)*. La corteza interior del álamo temblón *(Populus)* es muy nutritiva, mientras que la savia del abedul *(Betula)* es deliciosa. Recoja la savia cortando una «V» en la corteza, pero no la haga demasiado ancha. Otros árboles comestibles son el algarrobo *(Ceratonia)* y el tamarindo *(Tamarindus)*.

PLANTAS VENENOSAS

EL NÚMERO DE plantas venenosas es mucho menor que el de las comestibles, particularmente en las regiones templadas. Algunas son venenosas por contacto, y causan urticaria o graves irritaciones *(véase página 181)*. Otras son tóxicas por ingestión y pueden provocar vómitos, diarrea, e incluso la muerte. Algunas plantas venenosas se asemejan a las comestibles, así que la única manera de evitar un desastre consiste en aprender a diferenciar las plantas venenosas. Tenga siempre un cuidado extremo al probar las plantas *(véase página 83)*, particularmente en una situación de supervivencia, en la cual puede que no disponga de asistencia médica.

PLANTAS DE CLIMA TEMPLADO

Las hojas y tallos de algunas plantas son venenosas, así como algunas flores, raíces y savia. Incluso el roce con otras plantas puede inducirlas a liberar el veneno almacenado en vellos o poros. Existen muchas más plantas comestibles que venenosas en las regiones templadas, pero las plantas tóxicas pueden muchas veces ser mortales.

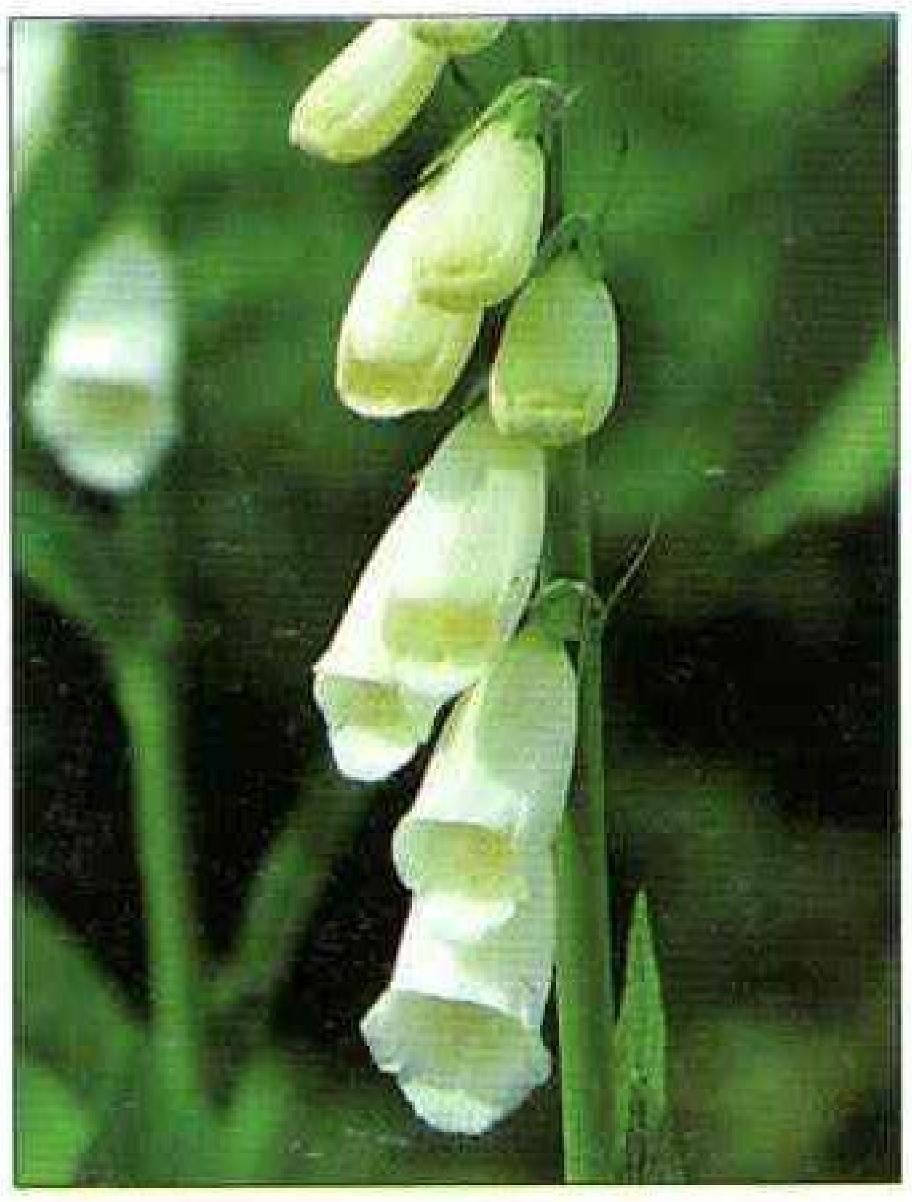

Dedalera (Digitalis). *Todas las partes de la dedalera contienen digitalina, un estimulante cardíaco altamente tóxico. La planta puede alcanzar una altura de 1,5 m, y tiene flores muy reconocibles, de forma acampanada y colores violeta, rosa, amarillo o blanco.*

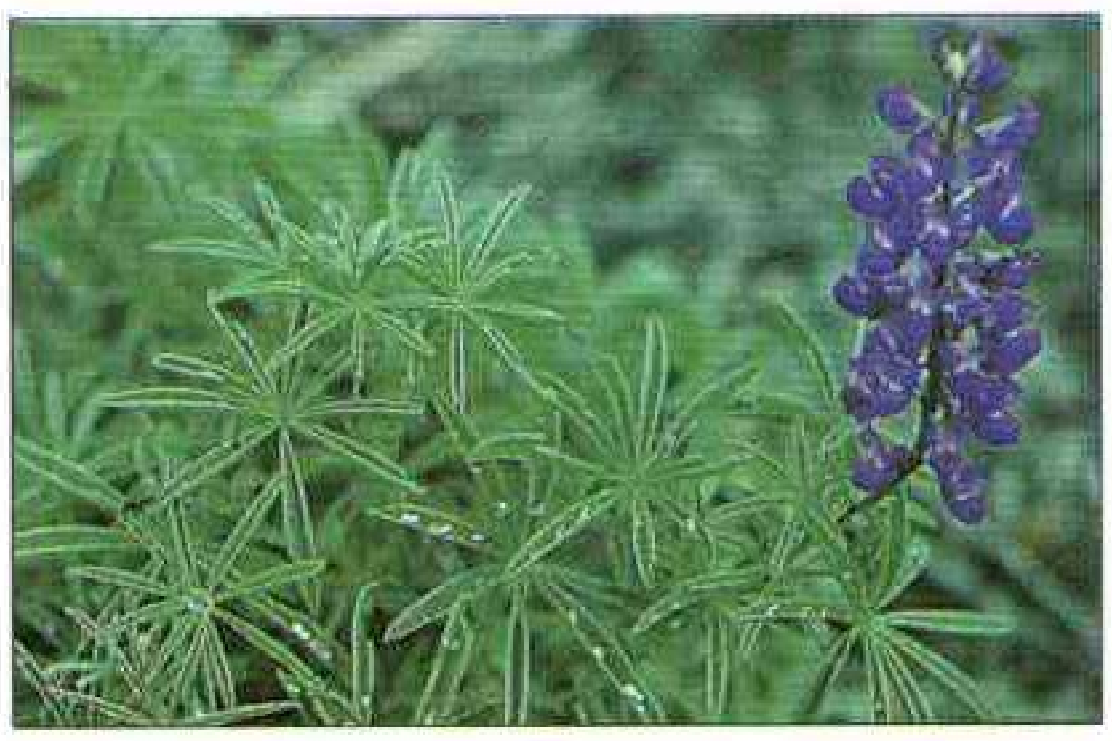

Altramuz (Lupinus). *Esta planta crece en Norteamérica, Europa y Asia. Todas sus partes causan inflamaciones mortales del estómago y los intestinos. La planta tiene flores púrpuras, rosas, blancas o amarillas.*

OTRAS PLANTAS DE CLIMA TEMPLADO

Una de las plantas más tóxicas de las regiones templadas es el acónito *(Aconitum)*, con flores de color púrpura. La mortal cicuta *(Conium)* y la cicuta acuática *(Cicuta)* se reconocen por sus flores blancas, dispuestas en forma de paraguas. Evite cualquier planta umbelífera: la mayoría son peligrosas. El zumaque venenoso y el árbol de las pulgas *(Toxicodendron)* causan irritación por contacto.

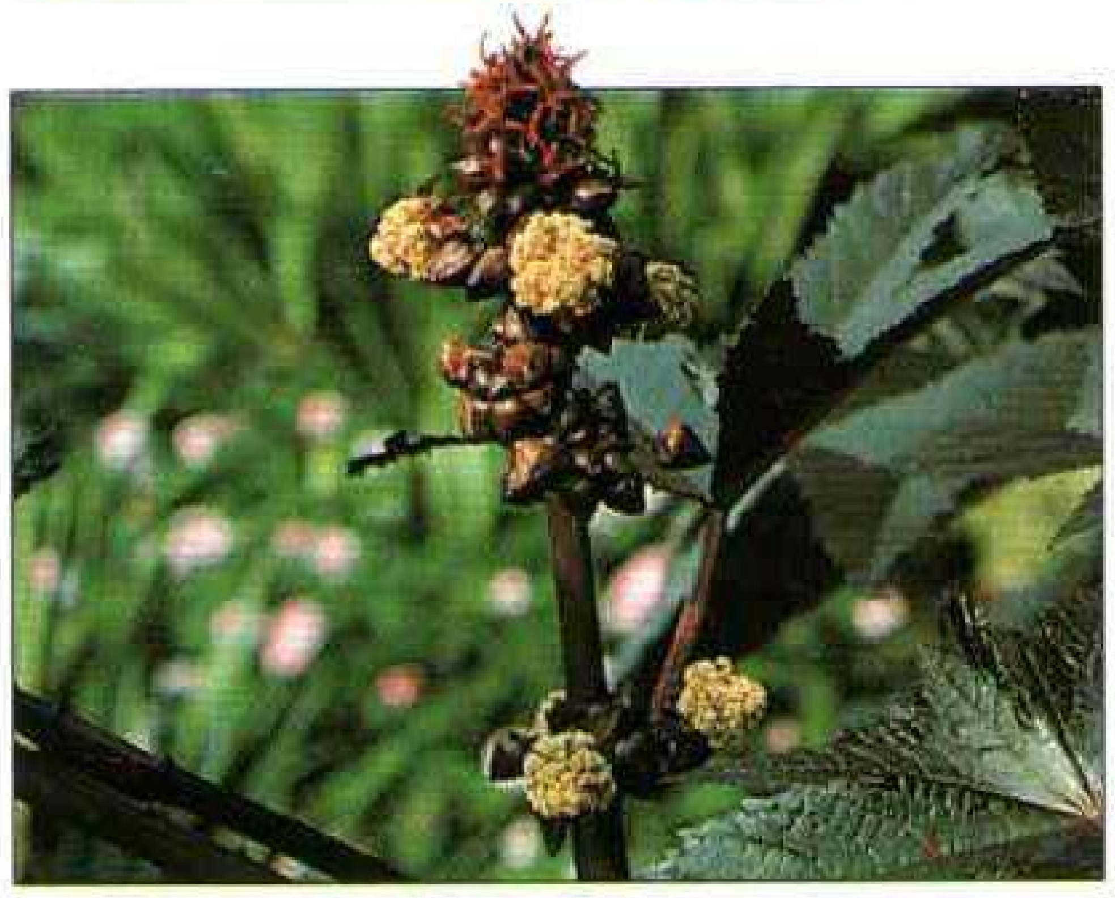

Ricino (Ricinus). *Aunque esta planta es originaria de las regiones tropicales, ha sido introducida en las áreas templadas. Se cultiva a menudo por el aceite con efecto purgante obtenido de las semillas. Sin embargo, si estas semillas se ingieren en su estado silvestre, pueden ser mortales.*

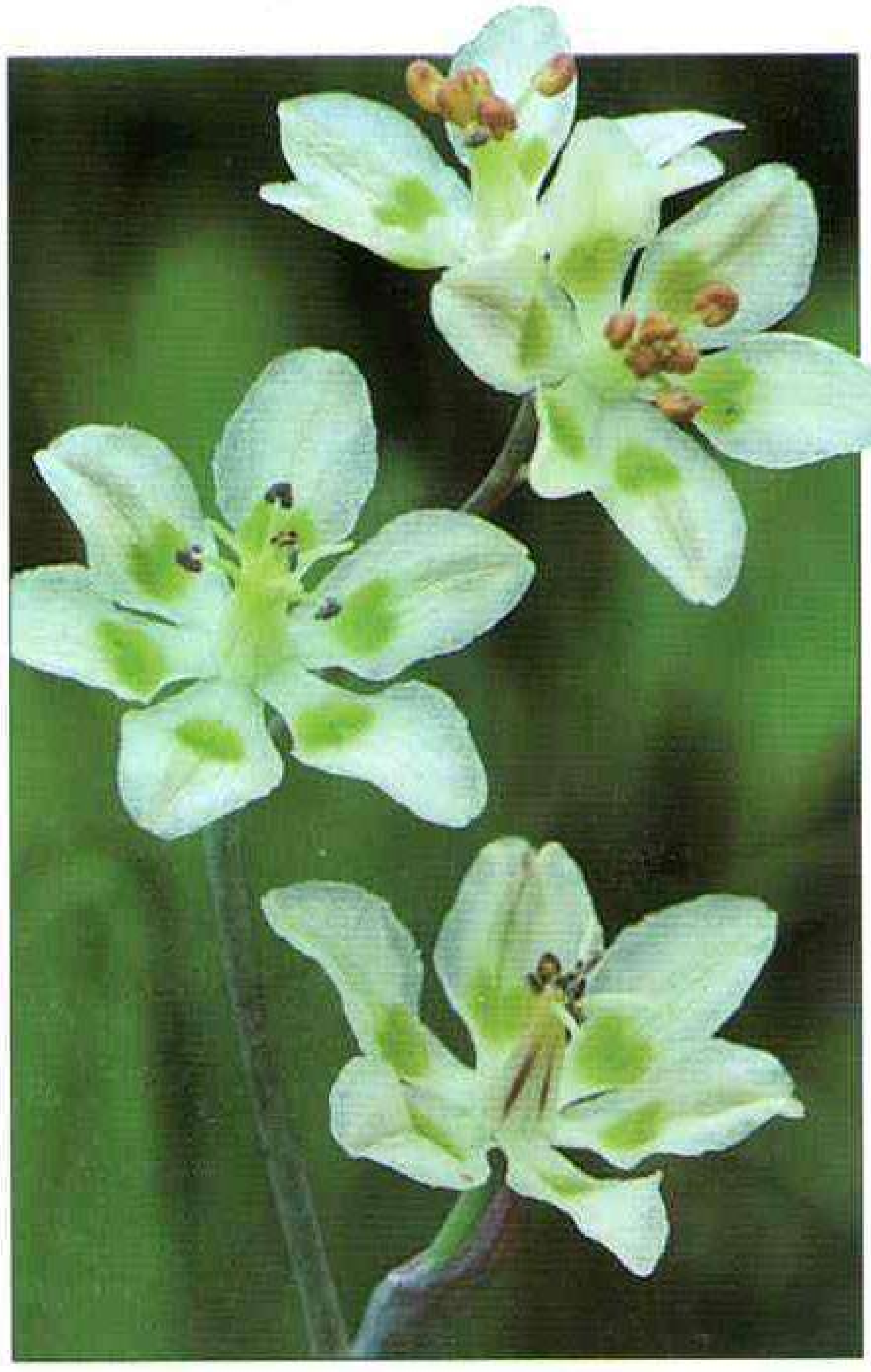

Zigadenus. *El bulbo letal de esta planta puede confundirse con el de una cebolla silvestre* (véase página 86). *Tiene largas hojas y flores blancas, y crece hasta unos 60 cm de altura. Se encuentra comúnmente en áreas boscosas y en pastos de Norteamérica.*

Hiedra venenosa (Toxicodendron). *La hiedra venenosa se encuentra en los bosques de Norteamérica. Puede ser reptante o de crecimiento vertical, como una hiedra común, pero se distingue por sus flores verdosas y sus bayas blancas. Produce urticaria al contacto con la piel.*

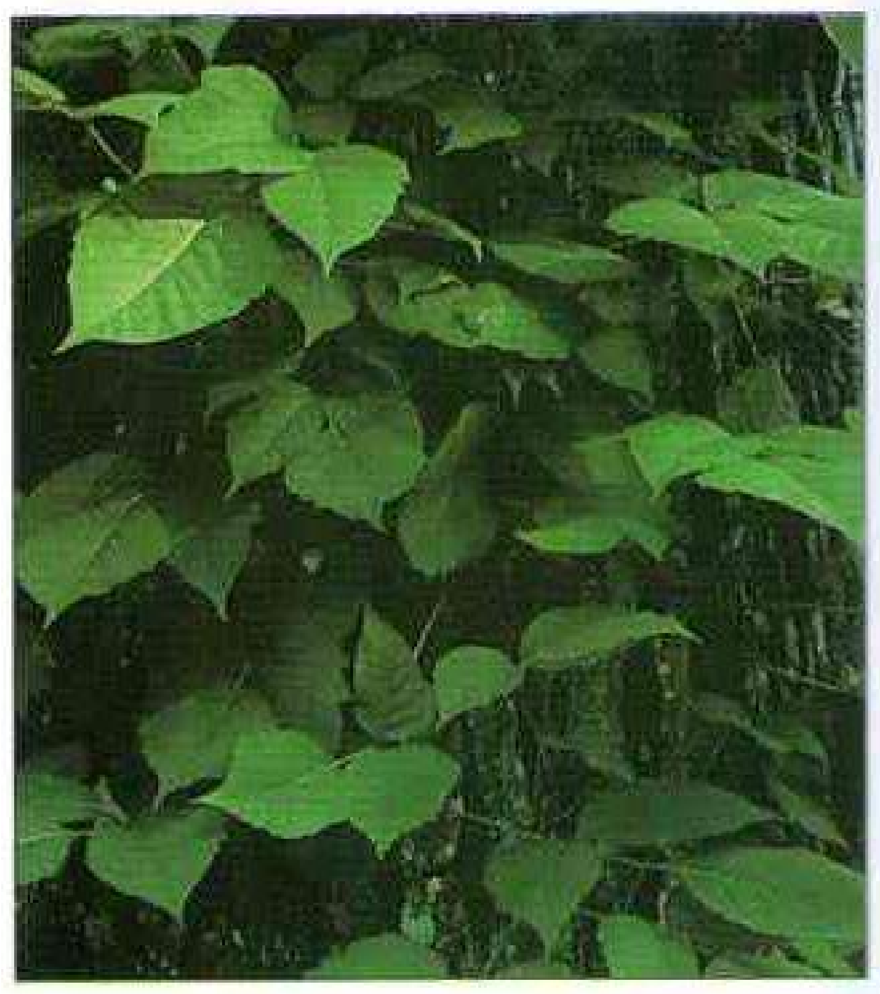

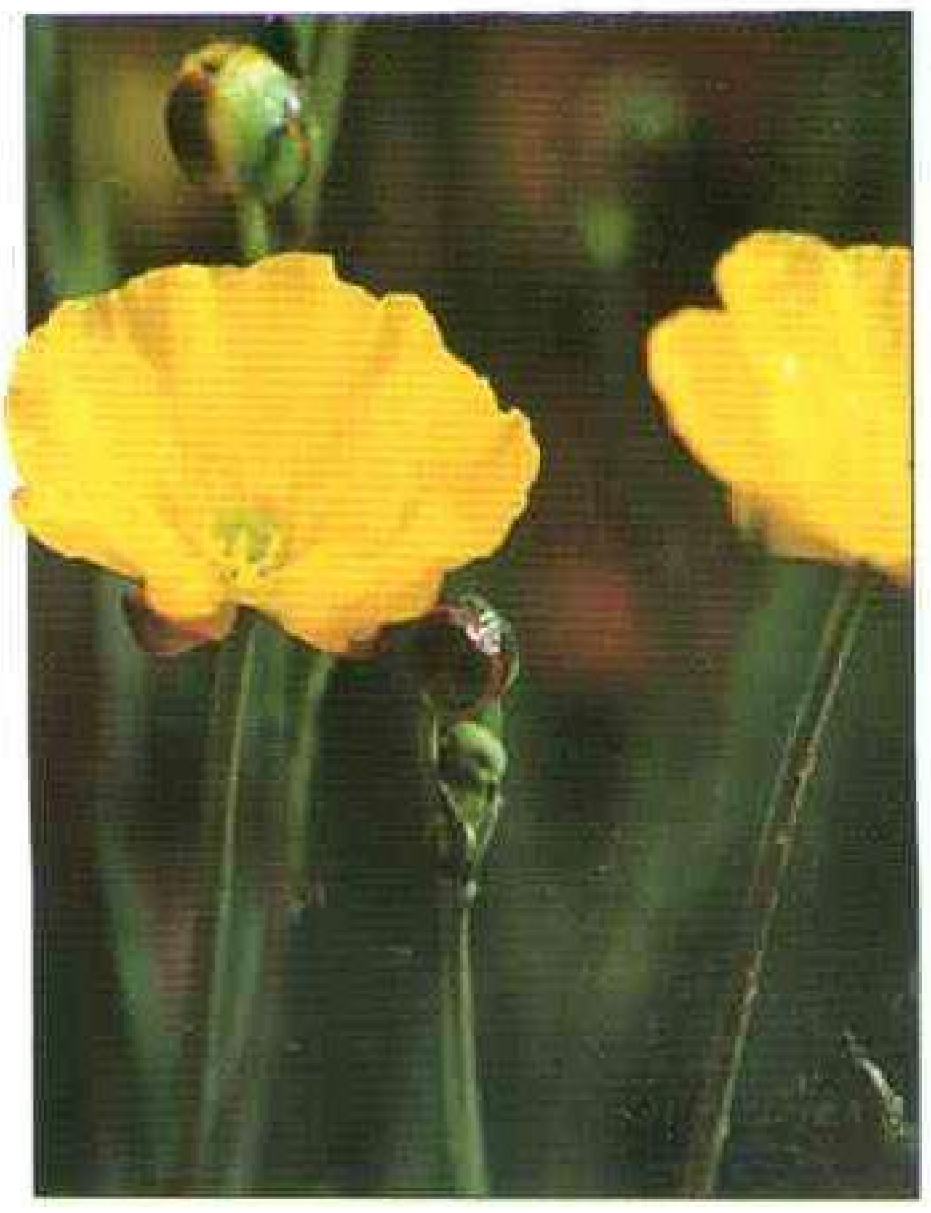

Botón de oro (Ranunculus). *Todas las especies de este género causan graves y dolorosas inflamaciones de los intestinos, en caso de ingestión. Se encuentran en todo el mundo, y todas las especies tienen al menos cinco pétalos amarillos, brillantes y superpuestos.*

ADVERTENCIA

Muchas plantas venenosas comunes de las regiones templadas se asemejan a alguna planta comestible, así que, a menos de que esté totalmente seguro de identificarla correctamente, evite comerla. Incluso un bocado de alguna de ellas, como la cicuta acuática *(véase superior)* puede causar la parálisis y la muerte en pocos minutos.

PLANTAS TROPICALES

Las plantas tropicales son tan variadas y abundantes que, por su propia seguridad, sólo debe comer aquéllas que pueda identificar. Si debe considerar la opción de comer una especie que no le sea familiar, aplique en primer lugar la prueba de sabor *(véase página 83)*, y coma únicamente una cantidad pequeña. A continuación se describen algunos ejemplos de plantas venenosas.

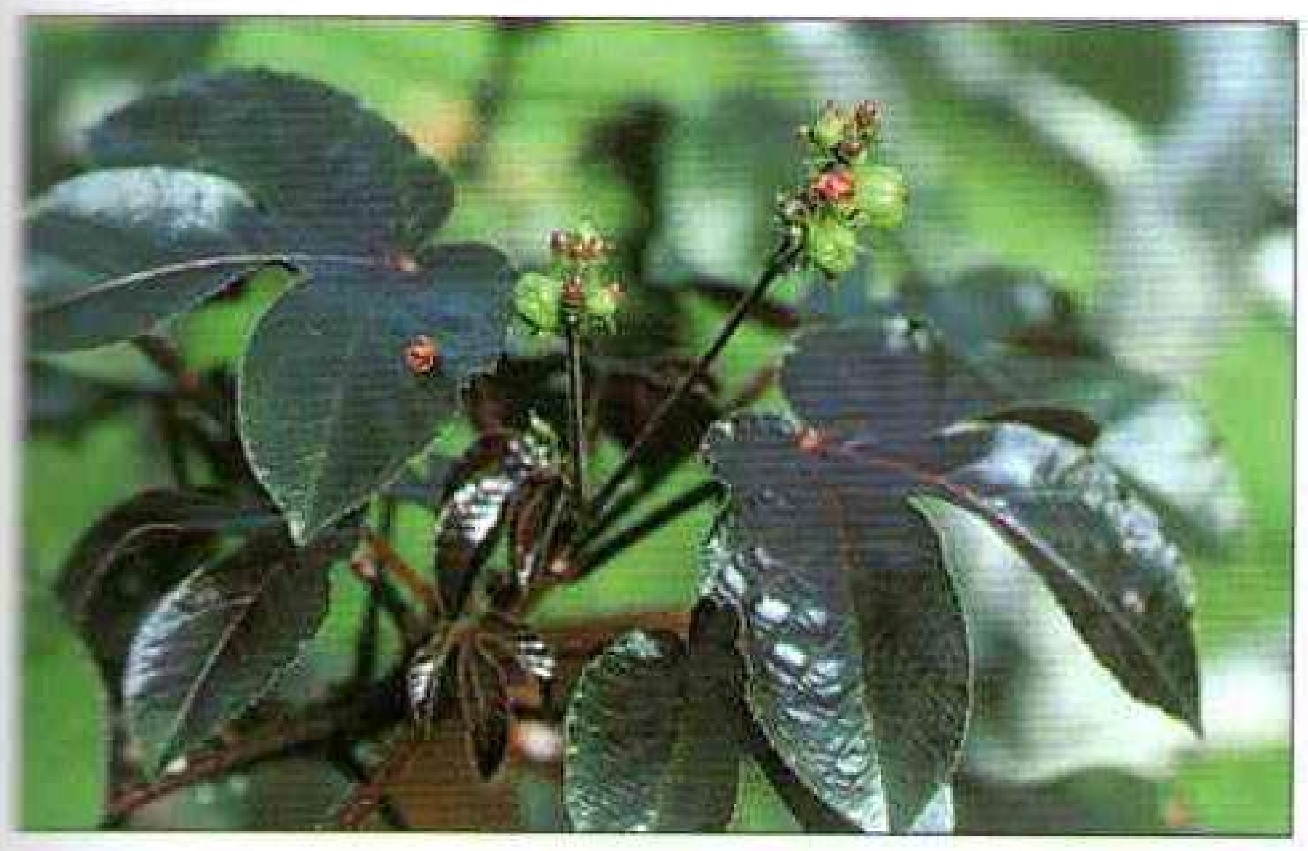

Jatropha glandulosa. *Varias especies de plantas tropicales del género* Jatropha *tienen hojas, semillas, savia o frutos venenosos. Las semillas tienen un fuerte efecto purgante.*

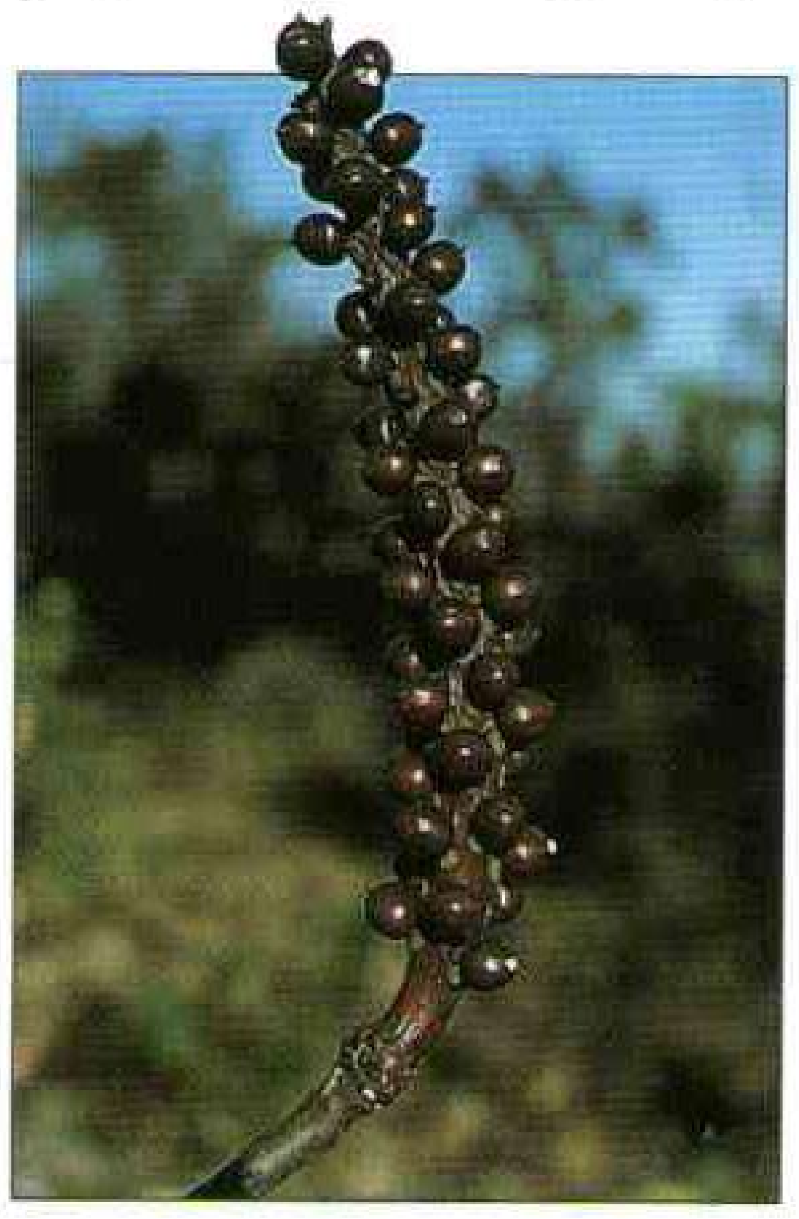

Sapium insigne. *La savia de esta planta es lechosa y altamente tóxica. No la ponga en contacto con su piel ya que le produciría ampollas. Evite todas las plantas con savia lechosa, ya que es muy probable que sean venenosas.*

OTRAS PLANTAS TROPICALES

La ortiga *(Laportea)* crece cerca del agua en las regiones tropicales. Causa urticaria por contacto, y sus semillas son venenosas. La manzana de la playa *(Hippomane)* tiene frutos y savia venenosos. *Mucuna* crece en los bosques y matorrales, y sus vainas y flores irritan la piel al contacto. En contacto con los ojos, estas partes de la planta pueden causar ceguera.

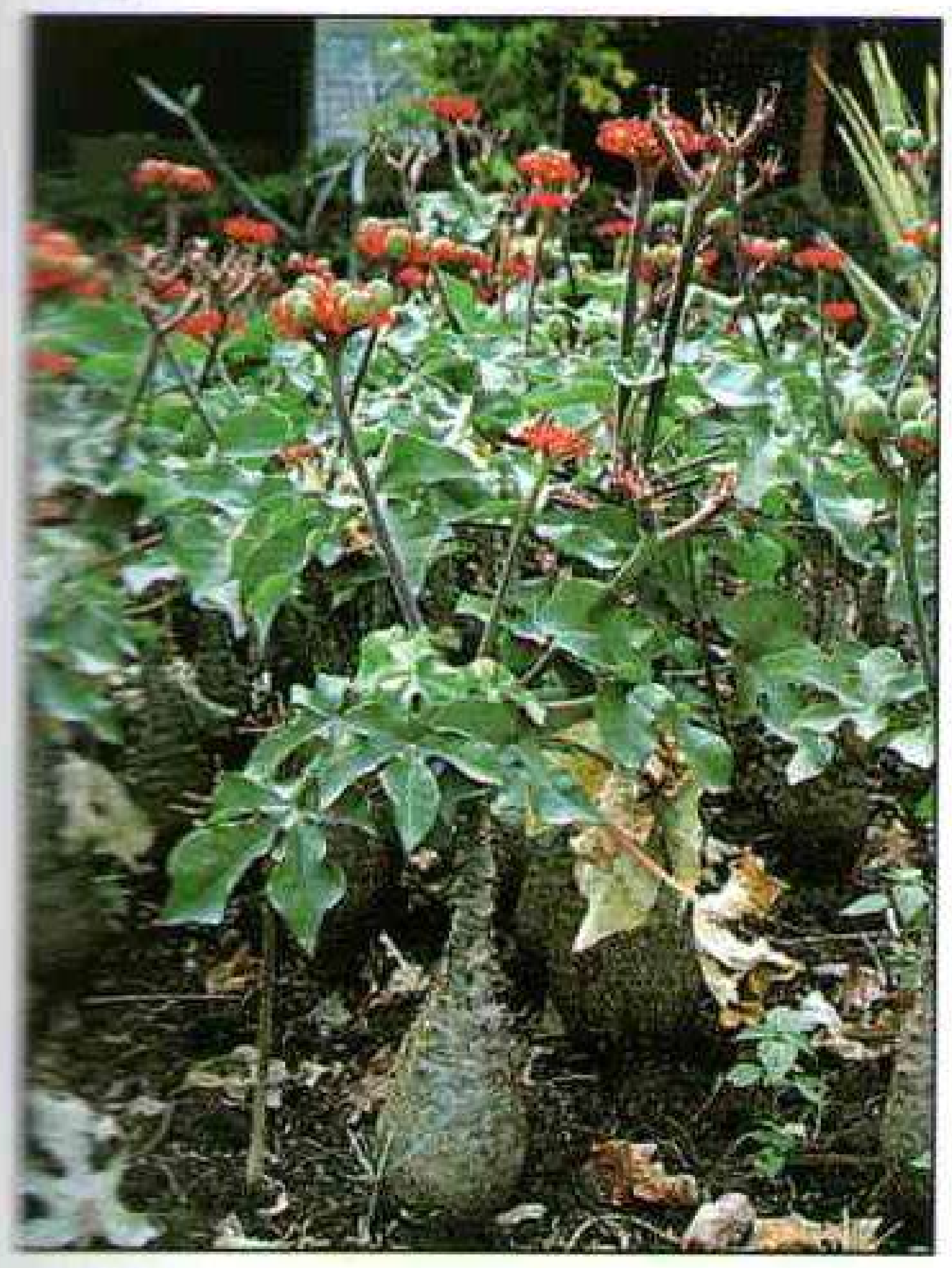

Jatropha podagrica. *Esta planta tiene la base semejante a un nabo, pero todas sus partes son muy venenosas. La única manera de identificar las plantas desconocidas es reconociéndolas a partir de una fuente fidedigna, como una guía local, o a través de la gente que habita en la zona.*

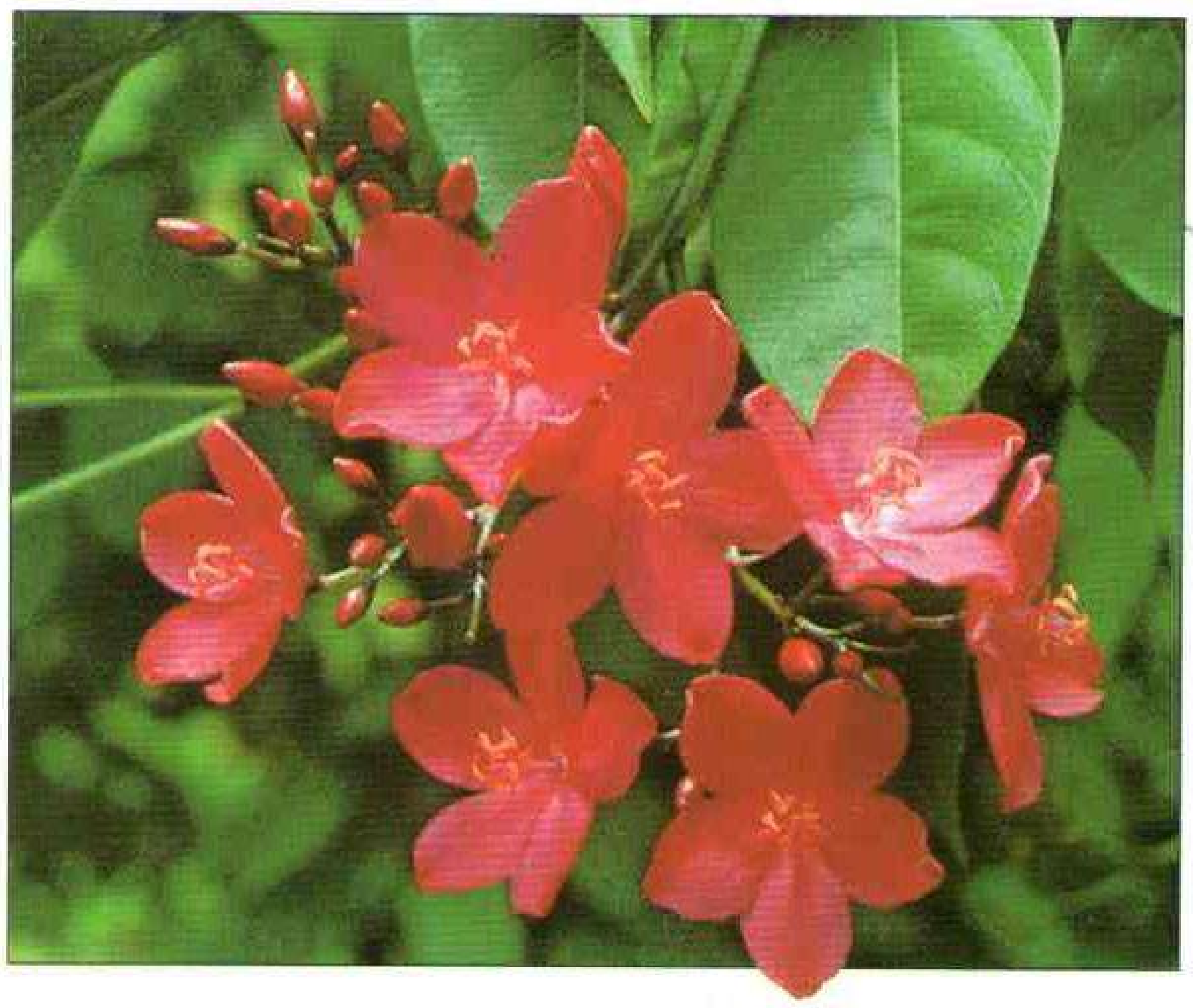

Jatropha integerrima. *Esta planta se encuentra en las regiones boscosas de los trópicos. Todas sus partes pueden causar inflamación de los intestinos, además de diarrea y vómitos. Estas condiciones pueden provocar deshidratación* (véase página 181), *que podría amenazar su vida en una situación de supervivencia, si no es capaz de encontrar ayuda médica inmediata.*

ÁRBOLES

No existe regla alguna para distinguir los árboles venenosos de los que no lo son. Algunas partes de algunos árboles pueden comerse, mientras que otras son venenosas. Debe tener especial cuidado con los árboles tropicales, ya que las semillas y los frutos de algunos de ellos pueden ser mortales, mientras que la savia de otros puede causar ampollas o ceguera.

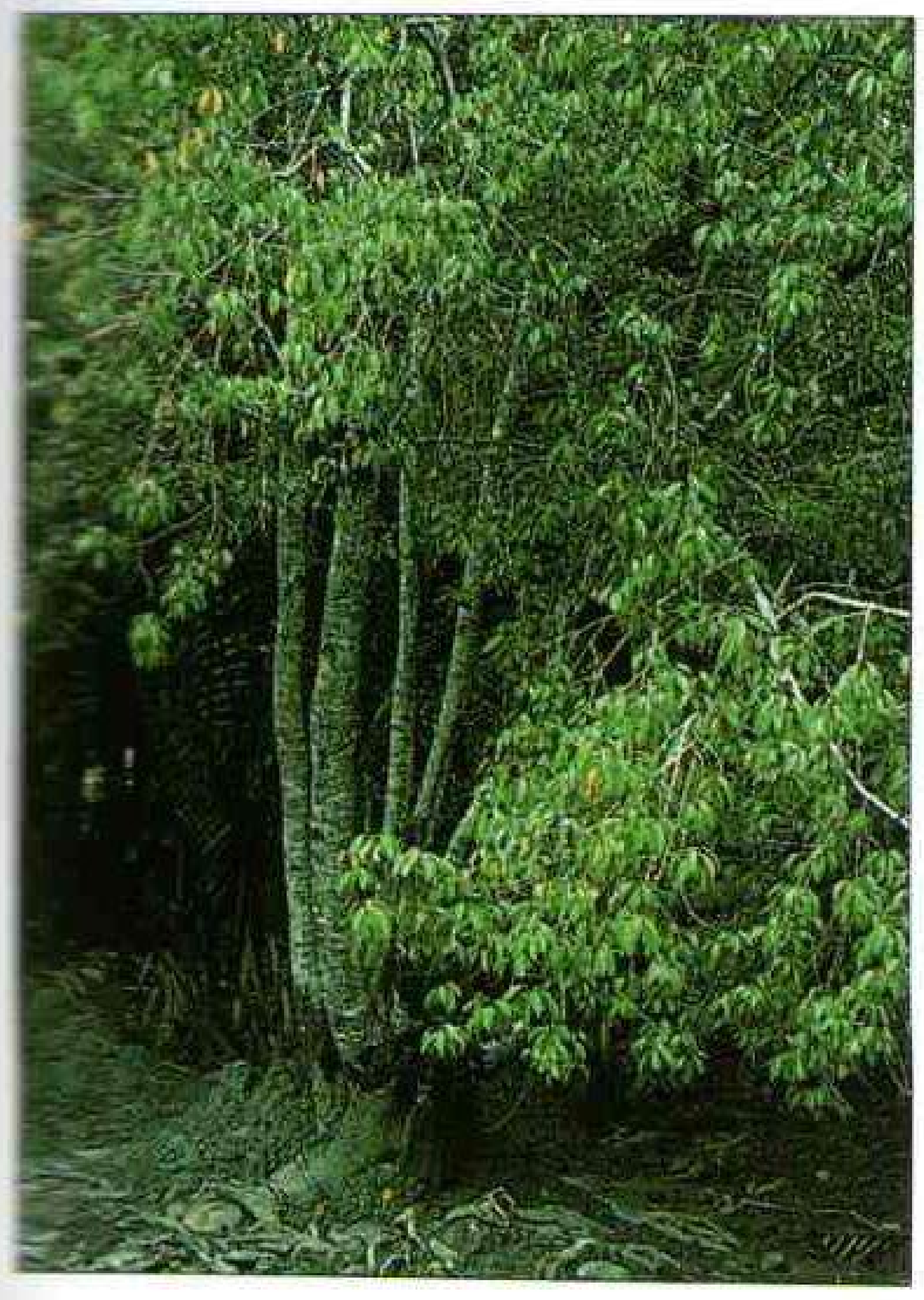

Mangle cegador (Excoecaria). *Este pequeño árbol se encuentra en los manglares y estuarios de Australasia, sudeste de Asia y África tropical. Debe evitarlo siempre que sea posible, ya que su savia produce ampollas en la piel, y ceguera al contacto con los ojos.*

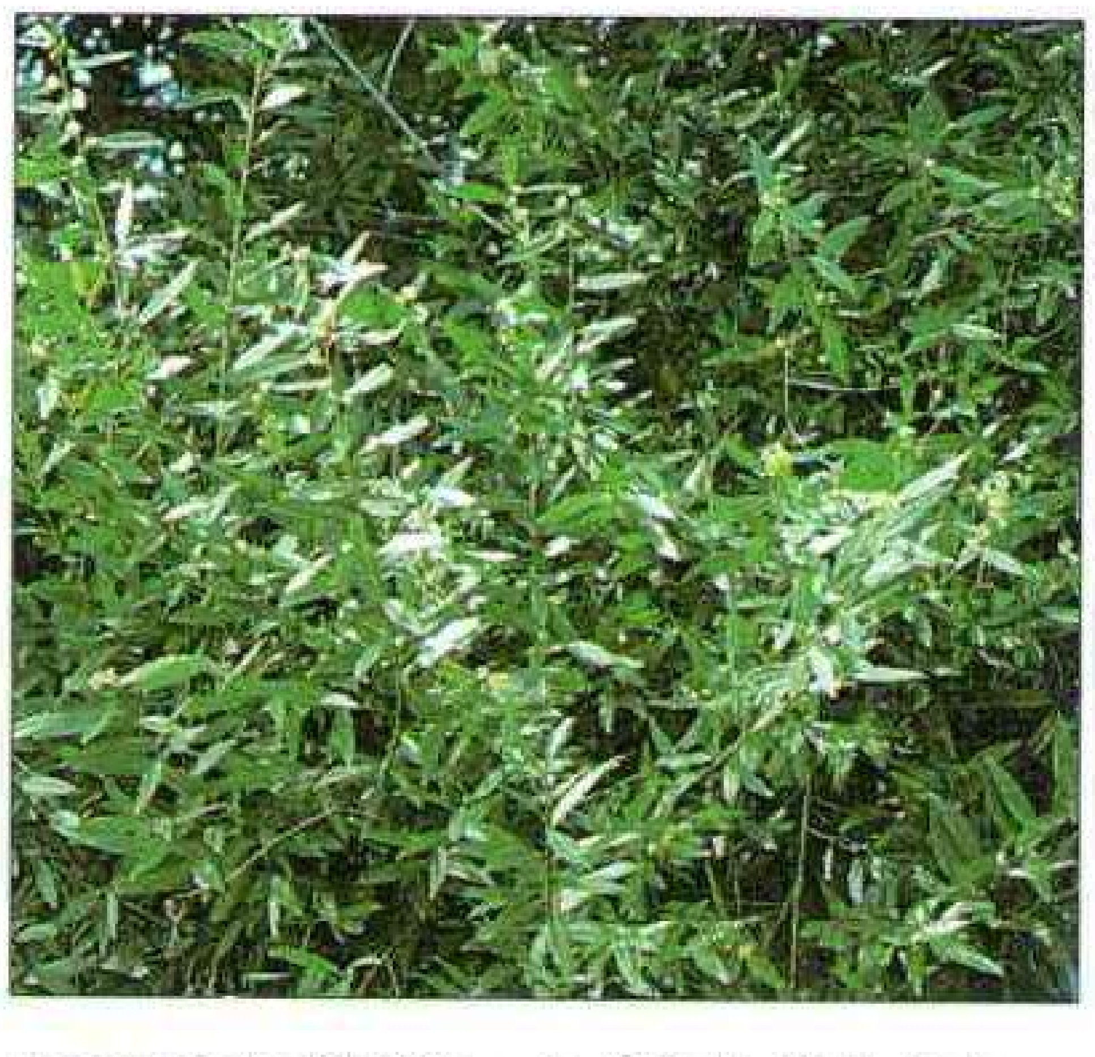

Laurel de California (Umbellularia). *El laurel de California crece en Norteamérica. Sus resistentes hojas ovaladas y perennes son fuertemente aromáticas y venenosas. Tiene flores amarillas y bayas verdes o púrpuras.*

OTROS ÁRBOLES

Las vainas del codeso *(Laburnum)* son mortales, como lo son las bayas del tejo *(Taxus)*. El algarrobo negro *(Robinia)* de Norteamérica contiene sustancias tóxicas en la corteza, flores y vainas. Los árboles rhengas *(Gluta)* del sudeste asiático tienen una savia irritante, mientras que la savia del árbol arenero *(Hua)* de Sudamérica puede producir ceguera. Evite comer cualquier parte de un árbol a menos que pueda identificarlo con seguridad.

RAÍCES, TUBÉRCULOS Y BULBOS

LAS RAÍCES, TUBÉRCULOS y bulbos de algunas plantas son ricos en vitaminas, tienen un contenido especialmente elevado de fécula y pueden comerse en situaciones de supervivencia. Algunas raíces son tóxicas si se comen crudas, por lo que se recomienda cocinarlas bien antes de comerlas. Límpielas bien, hiérvalas después y termine asándolas al fuego *(véase página 116)*. La piel de muchas raíces es rica en vitaminas y minerales, así que evite pelar las raíces si las pieles están en buen estado. Sin embargo, algunas raíces deben pelarse para eliminar sustancias peligrosas. Aprenda a reconocer algunas raíces comestibles y deseche las demás.

RAÍCES COMESTIBLES

La única manera de determinar si una raíz o bulbo es comestible es aprendiendo a reconocer la planta que genera. Las plantas de diferentes regiones pueden parecerse, o tener bulbos similares, así que no suponga que porque un bulbo le parezca familiar es comestible; así, por ejemplo, el bulbo de la mortal *Zigadenus* parece el de una inofensiva cebolla *(véase página 84)*. Casi todos los bulbos y raíces deben hervirse antes de comerlos.

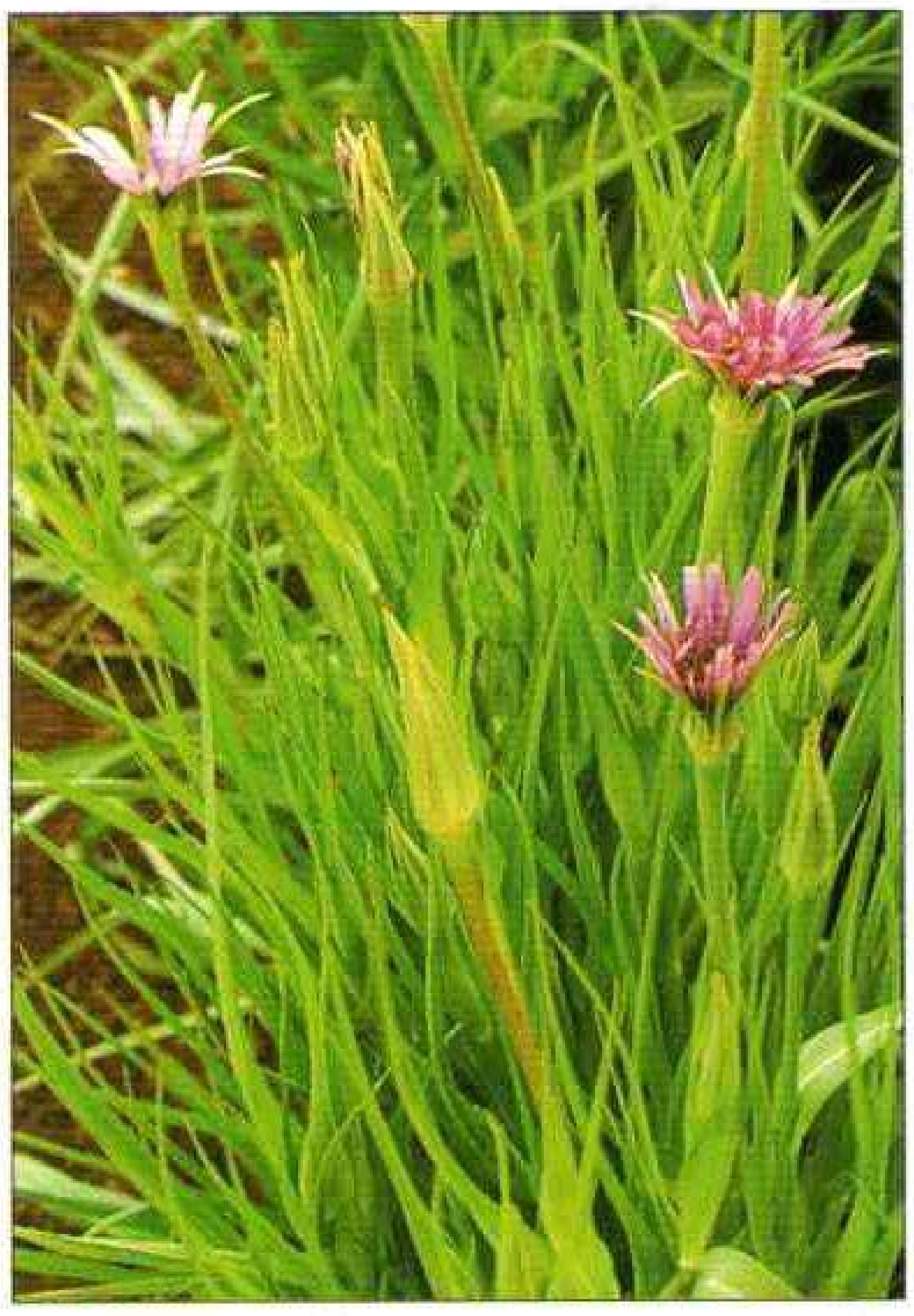

Salsifí (Tragopogon). *El salsifí tiene entre 60 y 90 cm de altura y crece en áreas secas. Tiene flores parecidas a las del diente de león. Sus hojas largas y el tubérculo parecido a la chirivía son comestibles, y llega a cultivarse.*

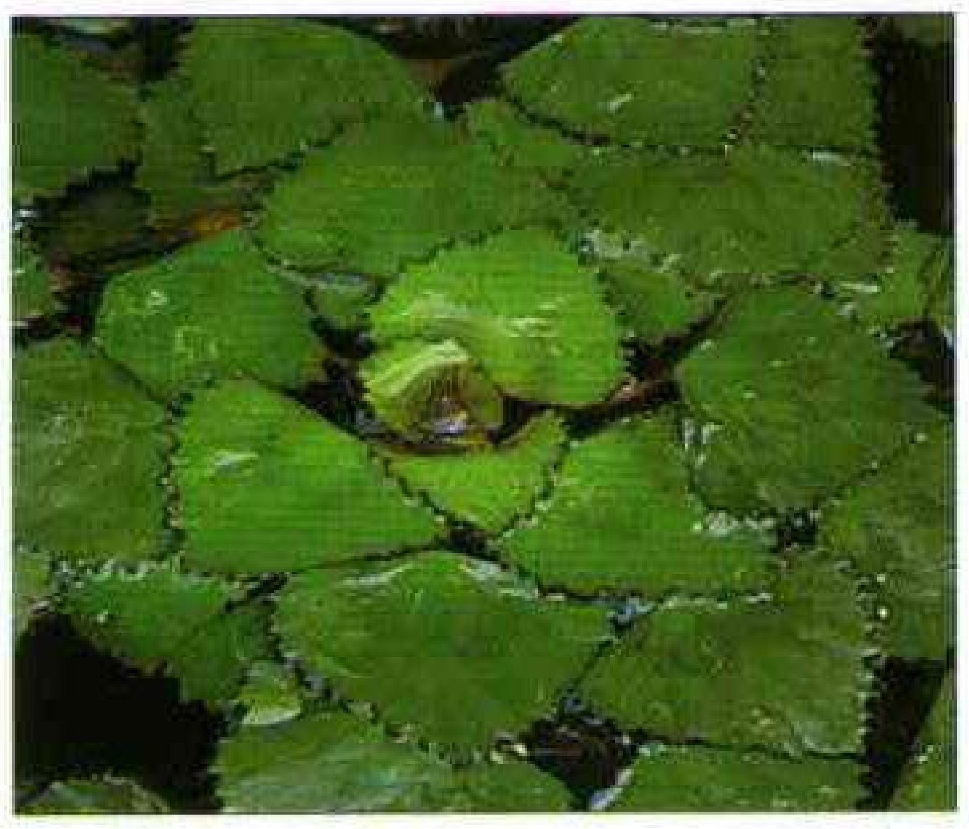

Castaño acuático (Trapa). *El castaño acuático crece en aguas dulces de Europa y Asia. Sus duras semillas son comestibles, tanto crudas como tostadas.*

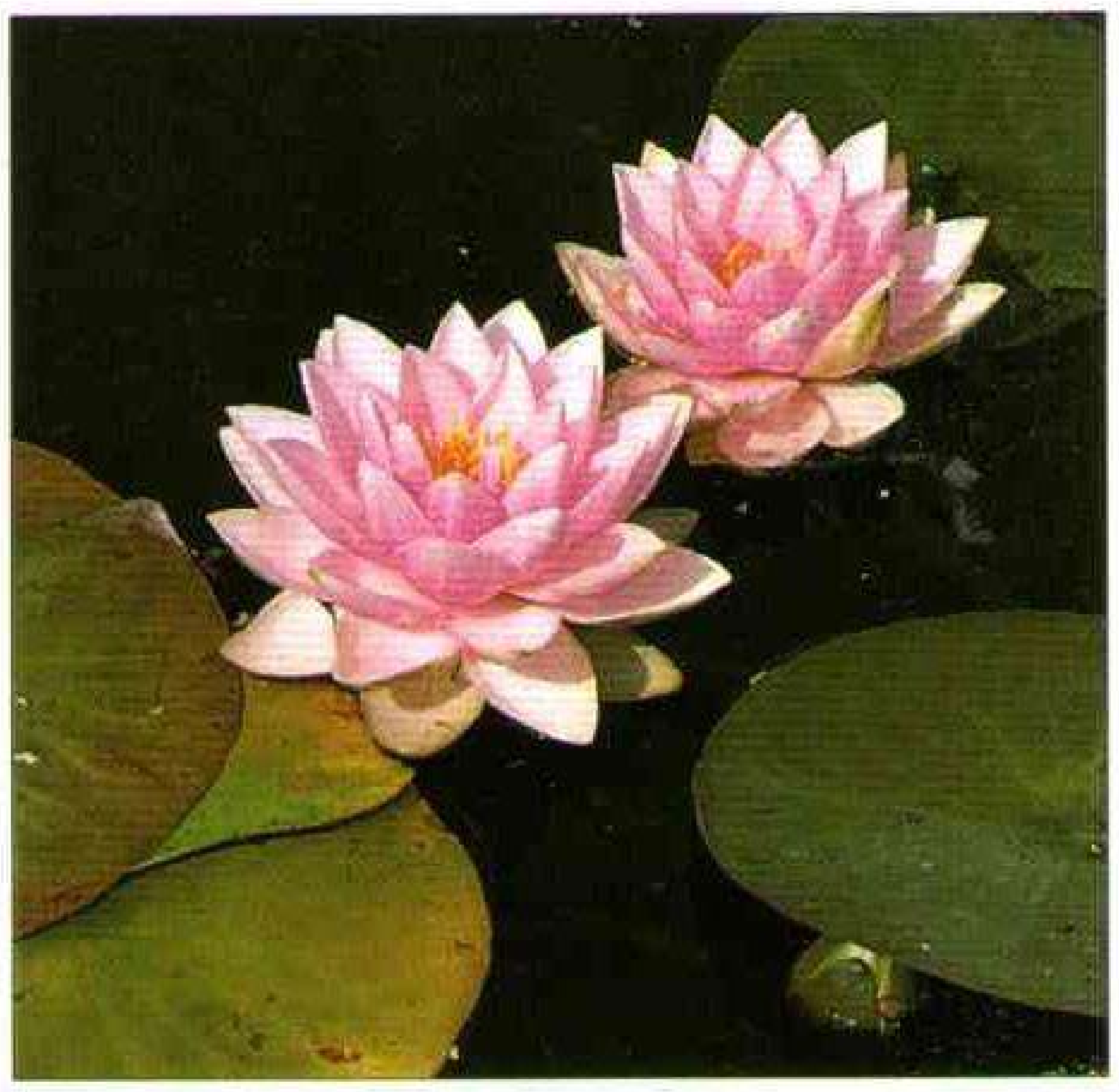

Nenúfar (Nymphaea). *El nenúfar tiene hojas flotantes en forma de corazón, y crece en aguas dulces, tropicales y templadas, casi en todo el mundo. Sus raíces, tubérculos y tallos son comestibles, aunque las semillas son ligeramente amargas.*

OTRAS RAÍCES COMESTIBLES

Las raíces del diente de león *(Taraxacum)*, de la juncia *(Cyperus)*, y de algunas otras plantas pueden tostarse para obtener un sustituto del café. Las raíces de plantas como el aro *(Calla)* pueden molerse hasta obtener un polvo fino que puede utilizarse como harina *(véase página 92)*. Las raíces del palo dulce *(Astralagus)* pueden comerse crudas o cocidas, al igual que los tubérculos de la alcachofa de Jerusalén *(Helianthus)*, que crece silvestre en Norteamérica.

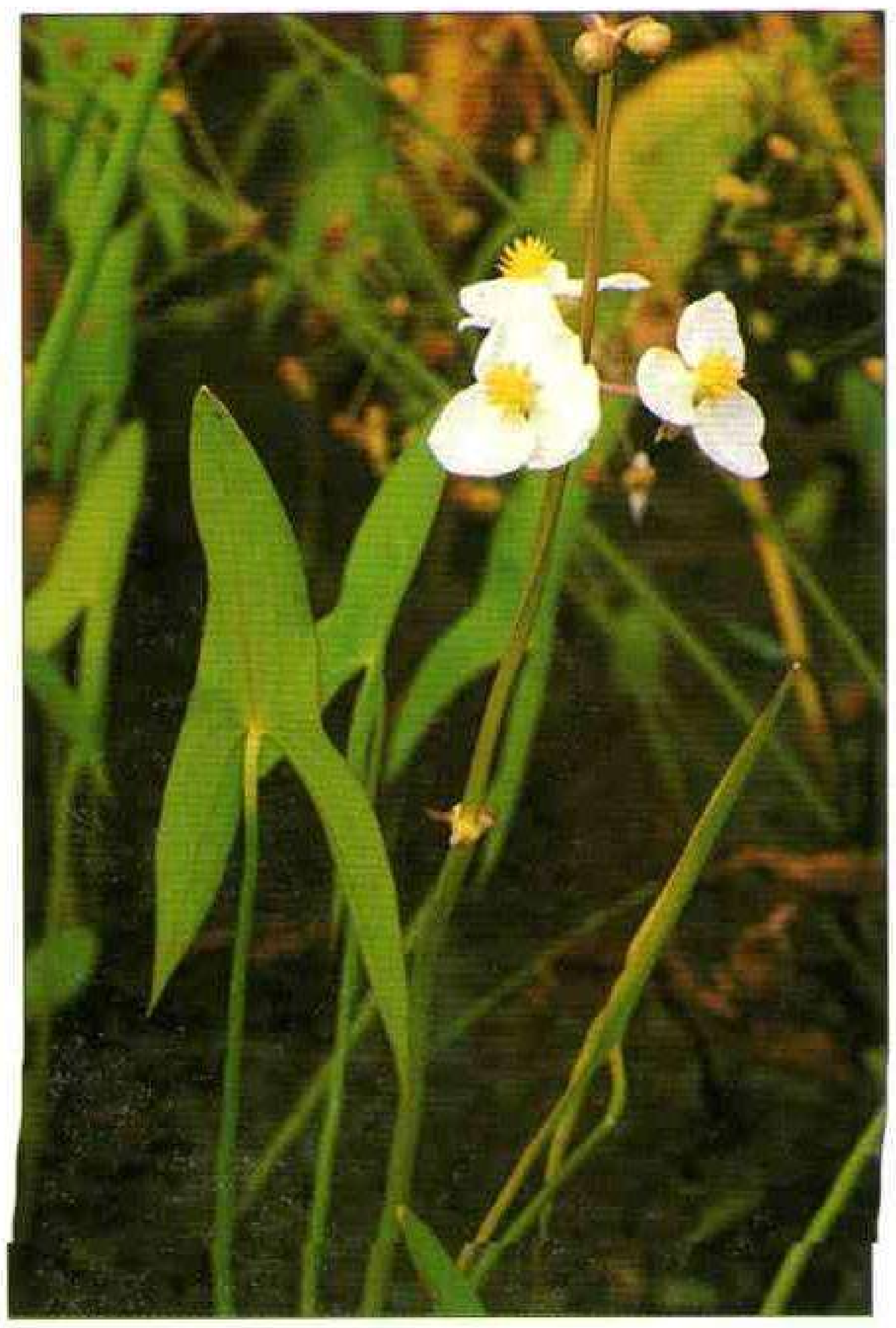

Saetas de agua (Sagittaria). *Pueden encontrarse cerca del agua dulce. Crece hasta 1 m de altura y tiene hojas erectas en forma de saeta, y pequeñas flores de tres pétalos. Sus tubérculos son comestibles en crudo, aunque son mejores cocidos.*

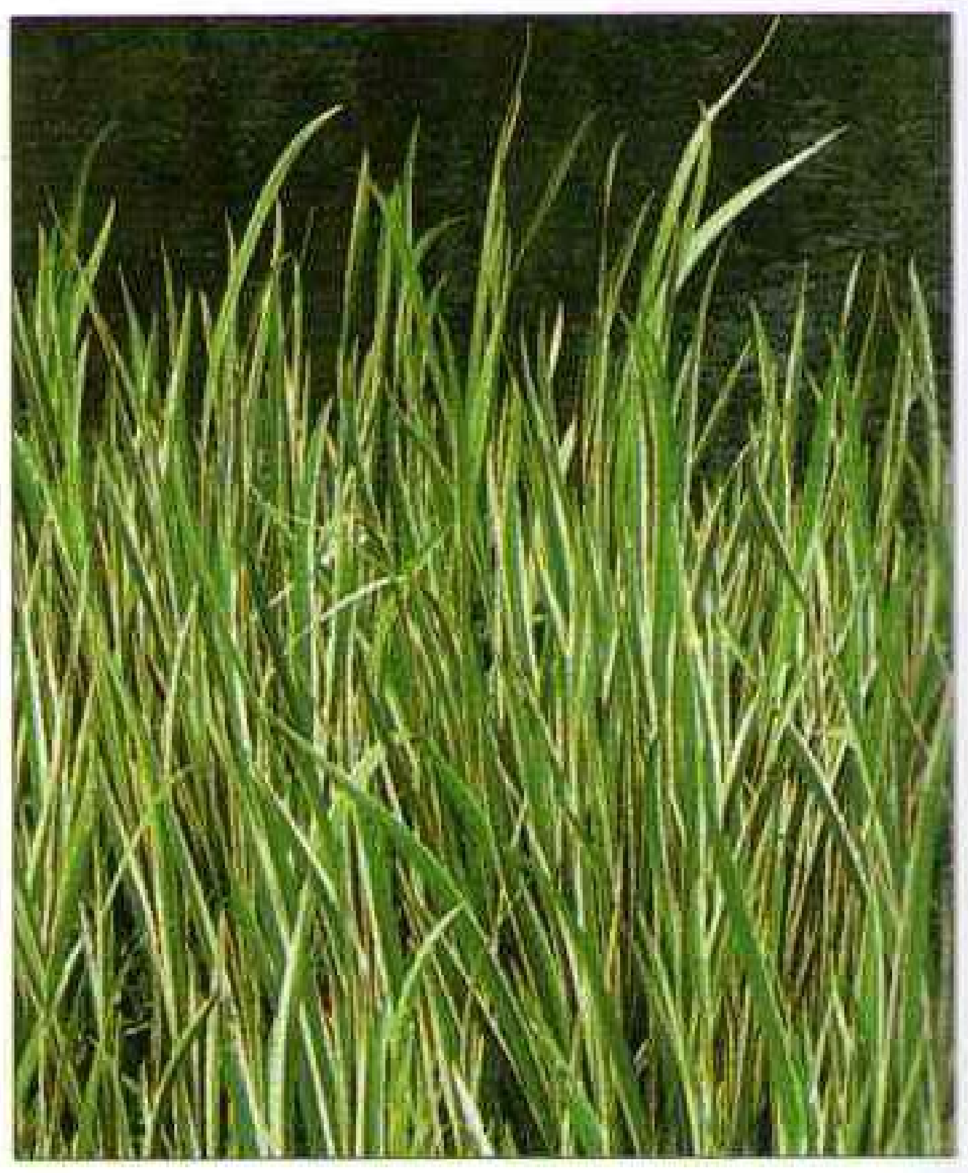

Ácoro o lirio dulce (Acorus). *El reconocible ácoro puede crecer hasta una altura de 1,3 m, y siempre crece junto a agua dulce. Sus raíces son fuertemente aromáticas y comestibles. Deben cortarse y hervirse hasta formar un jarabe, antes de comerlas.*

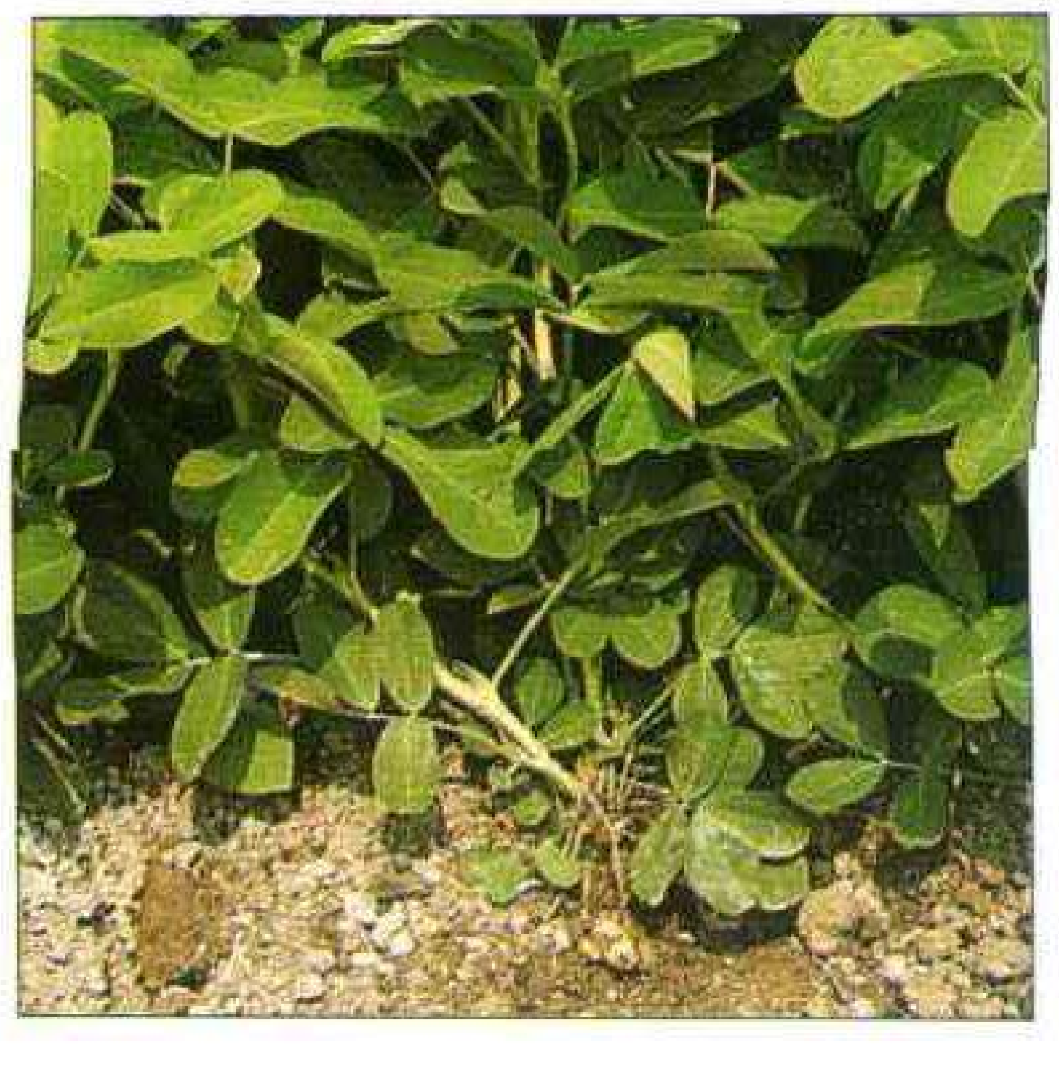

Cacahuete (Arachis). *Los cacahuetes se obtienen de las vainas que crecen bajo tierra, unidas a los tallos. La planta del cacahuete es pequeña y arbustiva, con hojas ovaladas y flores amarillas y planas.*

RAÍCES VENENOSAS

Hablando en términos generales, si las raíces de una planta son venenosas, sus hojas, tallo, flores y savia también lo serán. La celidonia menor es un buen ejemplo —sólo con tocar la planta aparecen ampollas en la piel. Sin embargo, hay excepciones a esta regla. Algunas raíces únicamente son comestibles una vez hervidas, y pueden ser terriblemente venenosas si se comen crudas. Tenga siempre mucho cuidado. A menos que esté seguro de haber identificado una planta correctamente, no la coma.

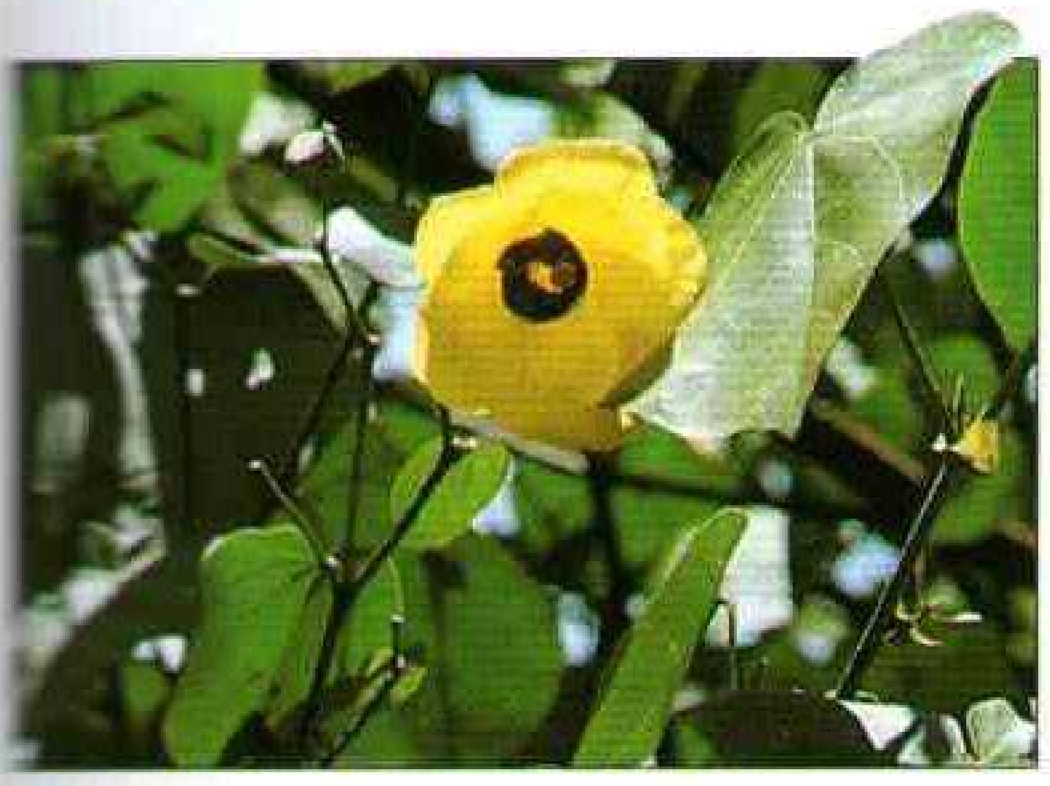

Mandioca (Manihot). *Los tubérculos de la mandioca o yuca son letales si se comen crudos. Deben remojarse durante 48 horas y hervirse a conciencia antes de comerlos.*

ADVERTENCIA

Ya que resulta tan difícil distinguir las raíces venenosas de las comestibles —a menos que la planta sea fácilmente reconocible— es mejor evitar comer raíces. Mientras pueda identificar con seguridad unas cuantas especies, aténgase a éstas e ignore las demás.

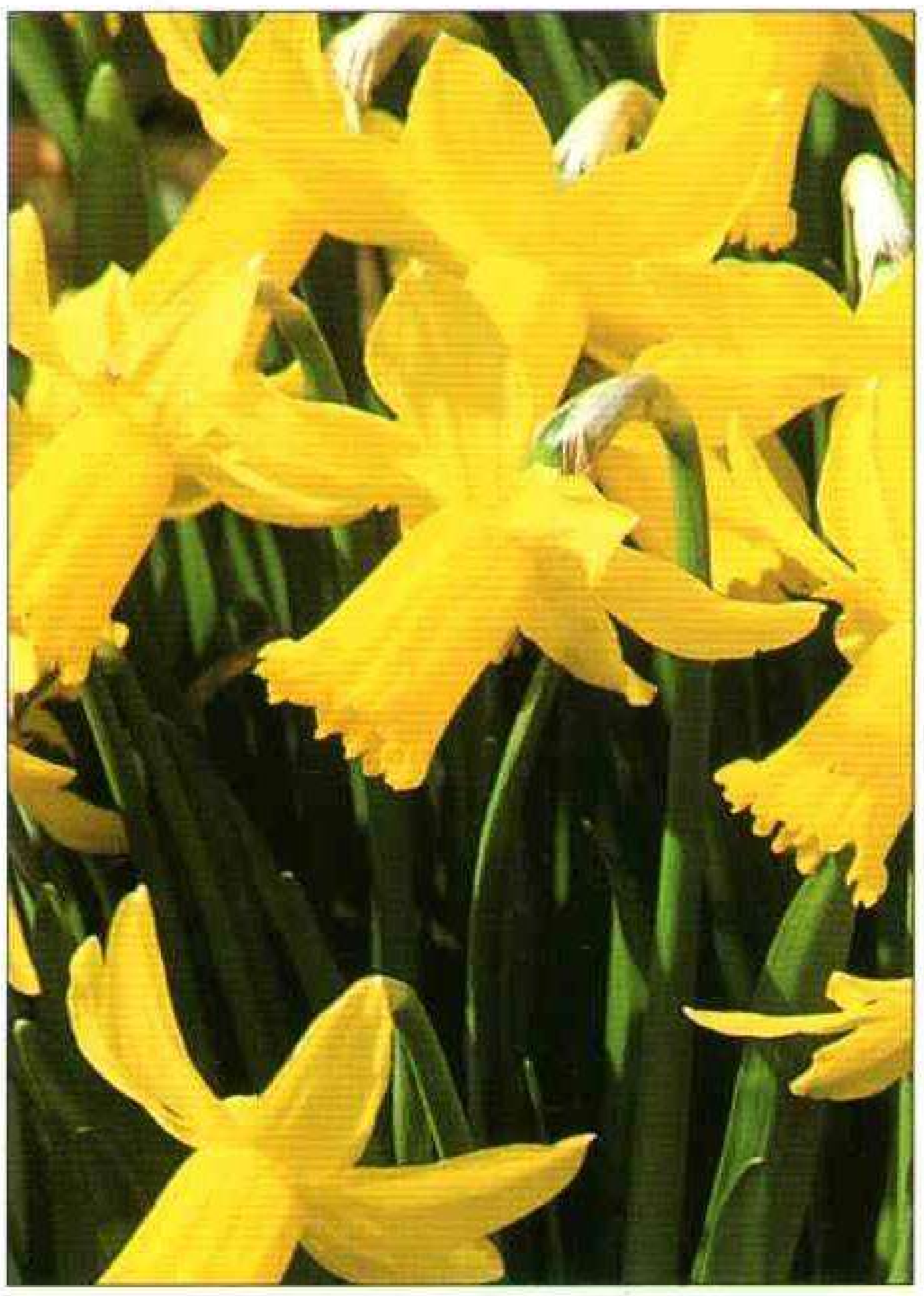

Narciso silvestre (Narcissus). *Es fácil de reconocer por su conocida flor, semejante a las variedades domésticas. Si no tiene flores, no toque la planta.*

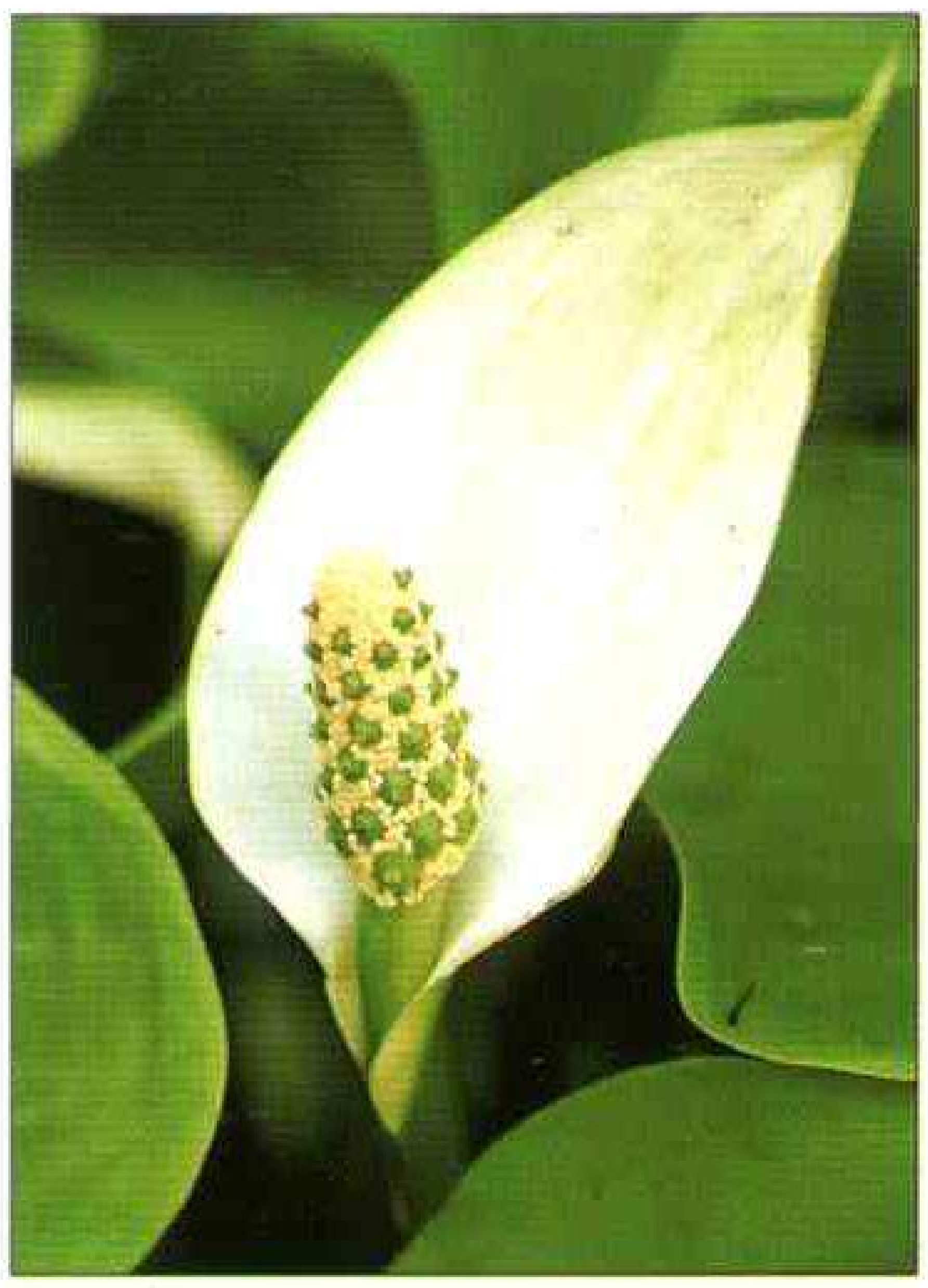

Aro de los pantanos (Calla). *También conocida como cala silvestre, el aro de los pantanos tiene una espiga floral distintiva dentro de un capuchón. Crece cerca de agua dulce. Las raíces pueden comerse o ser molidas para obtener harina* (véase página 92), *pero deben hervirse antes. Todas las demás partes de la planta son venenosas, y deben evitarse.*

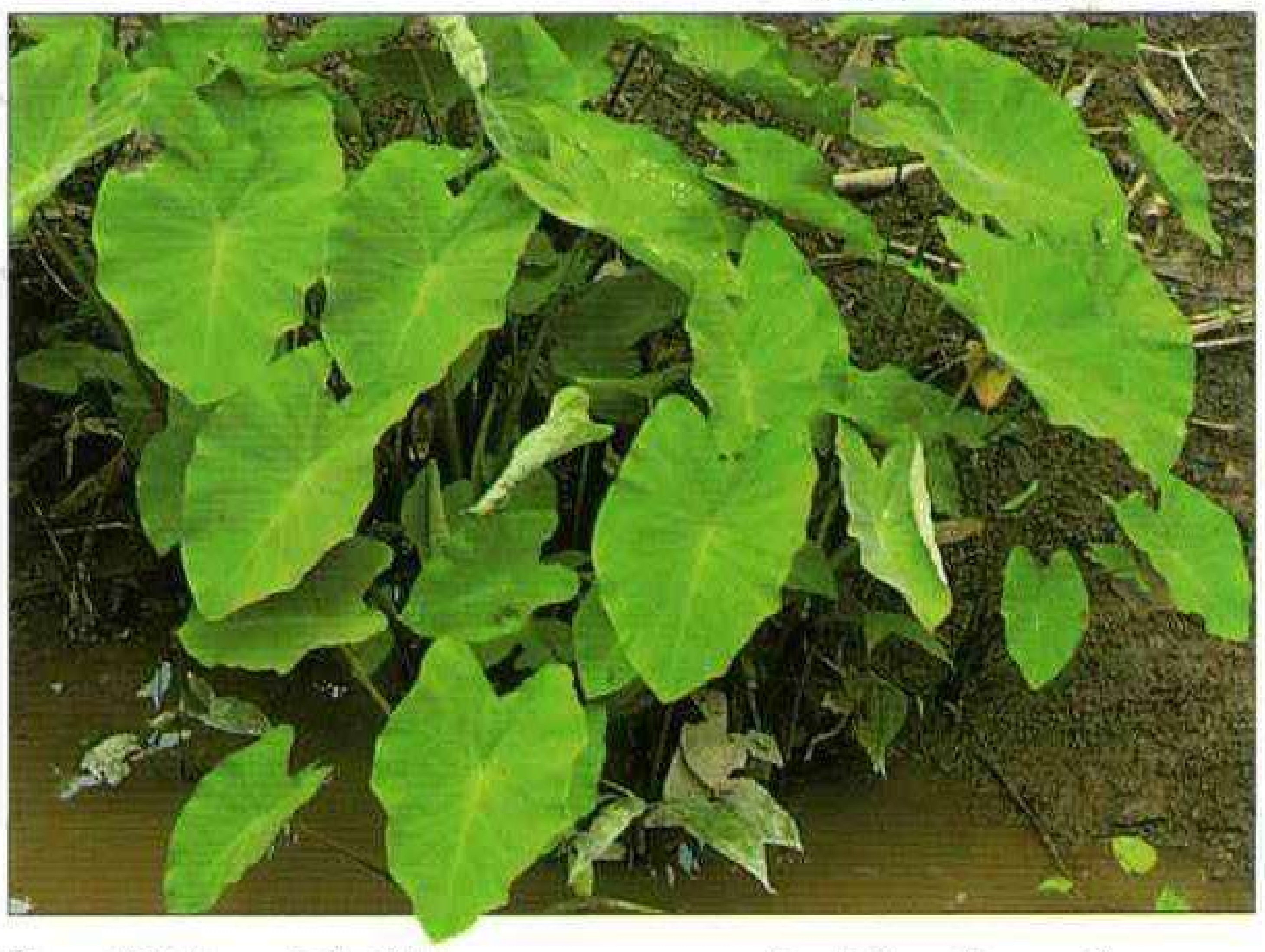

Taro (Colocasia). *El taro crece en suelos húmedos en las regiones tropicales. La planta alcanza una altura de hasta 1,5 m y tiene grandes hojas verdes en forma de cuña, y una flor de color amarillo naranja. Sus tubérculos son venenosos si se comen crudos, pero pueden consumirse una vez cocidos. Tienen un sabor semejante al de la patata.*

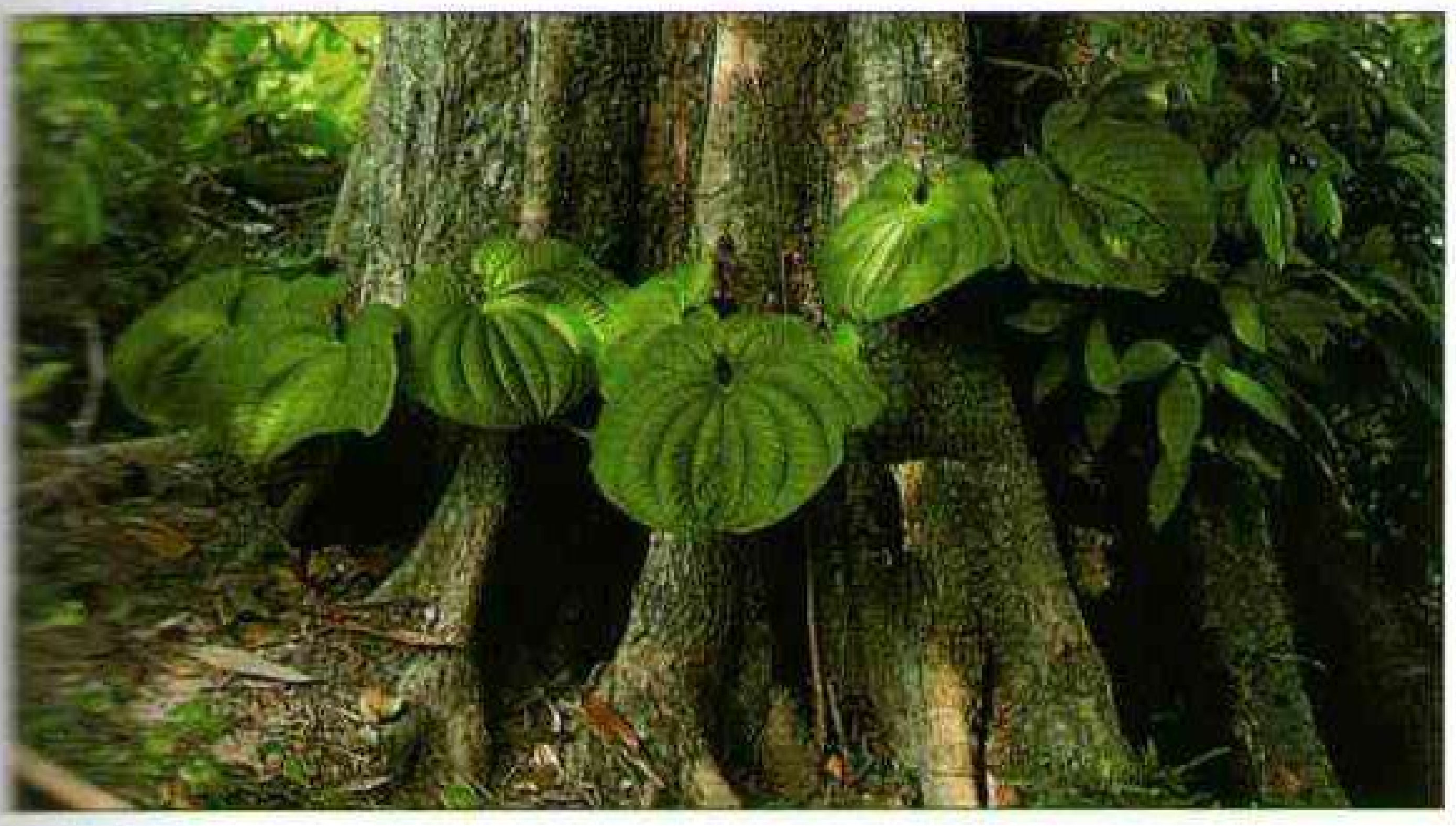

Ñame silvestre (Dioscorea). *Aunque algunas especies de ñame silvestre se cultivan en zonas tropicales, algunas de ellas son venenosas a menos que se pelen y se hiervan. Si no puede identificar las variedades, lo ideal es pelar y hervir los ñames silvestres. Tienen hojas grandes y crecen a menudo alrededor de los troncos de los árboles.*

OTRAS RAÍCES VENENOSAS

Los tubérculos de la patata silvestre *(Solanum)* son comestibles, pero deben hervirse antes. Sus frutos, que asemejan a un tomate, son venenosos. El tomate silvestre *(Lycopersicon)*, que es comestible, tiene un aspecto muy similar a la planta de la patata: es aconsejable evitar ambos. *Zigadenus* tiene un bulbo parecido a una cebolla, pero es letal *(véase página 84)*. Los tubérculos y las raíces de la celidonia menor o botón de oro *(Ranunculus)* y de la cicuta acuática *(Cicuta)* son mortales.

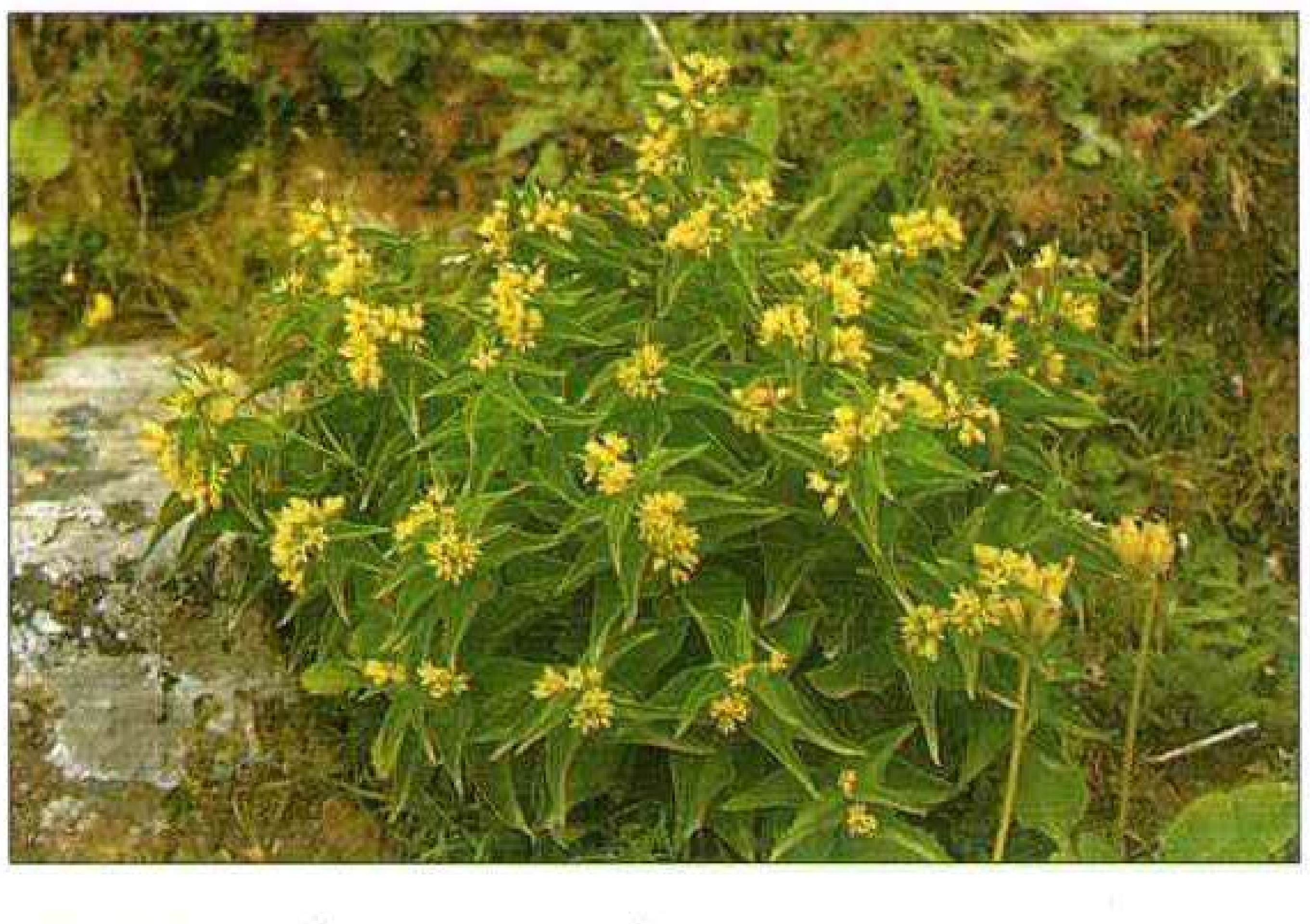

Vencetósigo (Vincetoxicum). *Esta planta tiene hojas apuntadas en forma de corazón y pequeñas flores de color blanco amarillento. Si se ingiere, sus raíces y vainas producen vómitos y pérdida de agua; su savia lechosa también es tóxica. Es común en regiones templadas.*

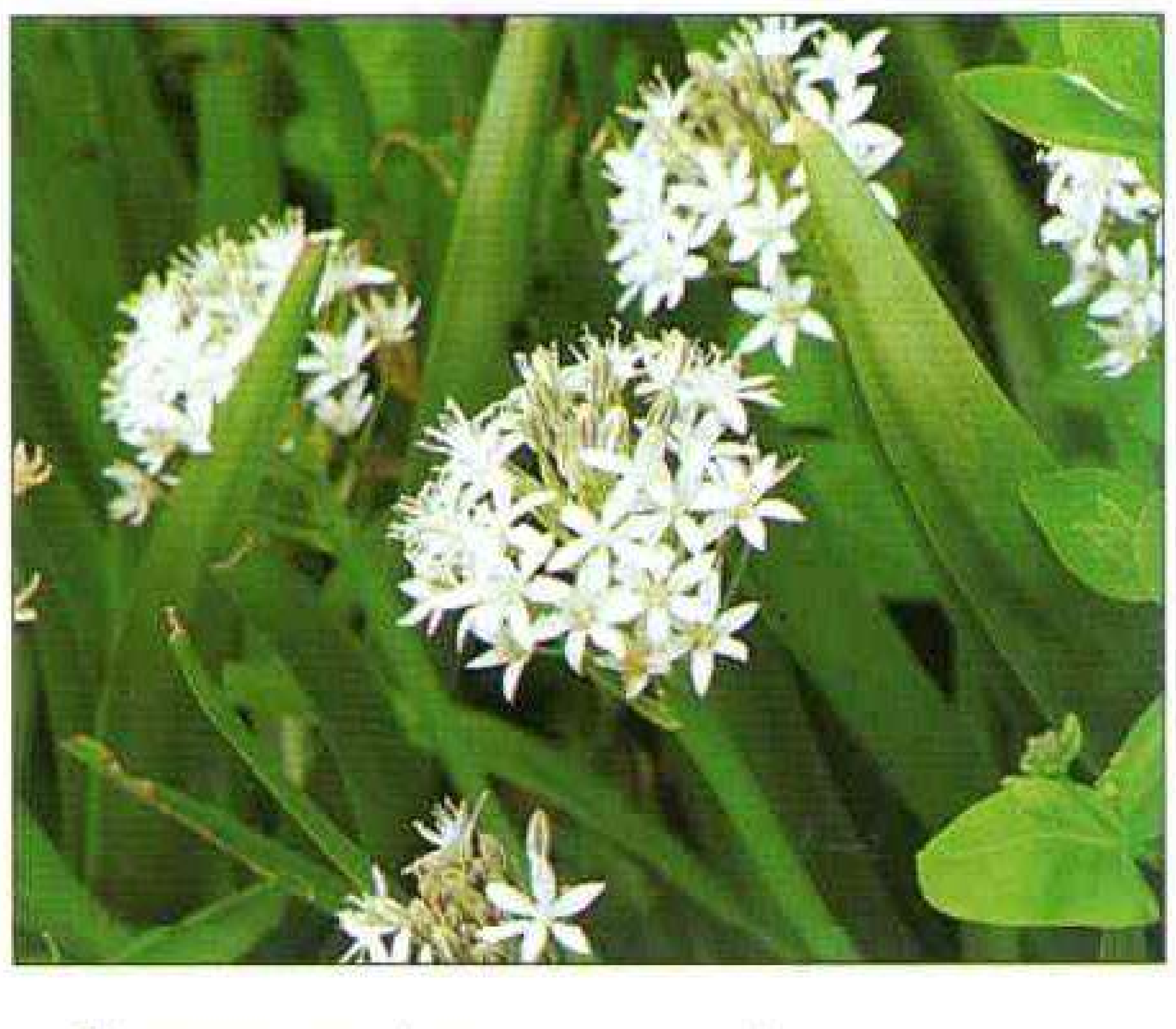

Matacandiles (Ornithogalum). *Esta planta propia de las regiones templadas crece hasta unos 30 cm de altura. Sus bulbos son comestibles siempre y cuando se hiervan, pero evite comer el resto de la planta.*

NUECES Y FRUTOS

TANTO LAS NUECES como los frutos constituyen la fuente más importante de alimento en una situación de supervivencia. En las regiones templadas, los frutos silvestres aparecen a partir de mediados de verano, mientras que las nueces aparecen más tarde, en otoño. La mayoría de los frutos silvestres son difíciles de almacenar a menos que haga una mermelada o seque las especies que lo permiten *(véase página 118)*. Recuerde que, al recoger frutos y nueces, está compitiendo con aves y mamíferos, particularmente cuando se trata de nueces, que algunos animales recogen como fuente principal de alimento para el invierno. No recoja frutos o nueces demasiado maduros o mohosos: podría enfermar.

NUECES COMESTIBLES

Las nueces proporcionan tanto proteínas como grasas. Con las nueces oleosas como las de la haya, puede separar su aceite y almacenarlo para cocinar. Hierva los hayucos lentamente en agua y separe el aceite de la superficie. También pueden ser molidas para obtener harina *(véase página 92)*. Las nueces son el alimento silvestre más fácil de almacenar, y se recomienda comerlas durante un traslado.

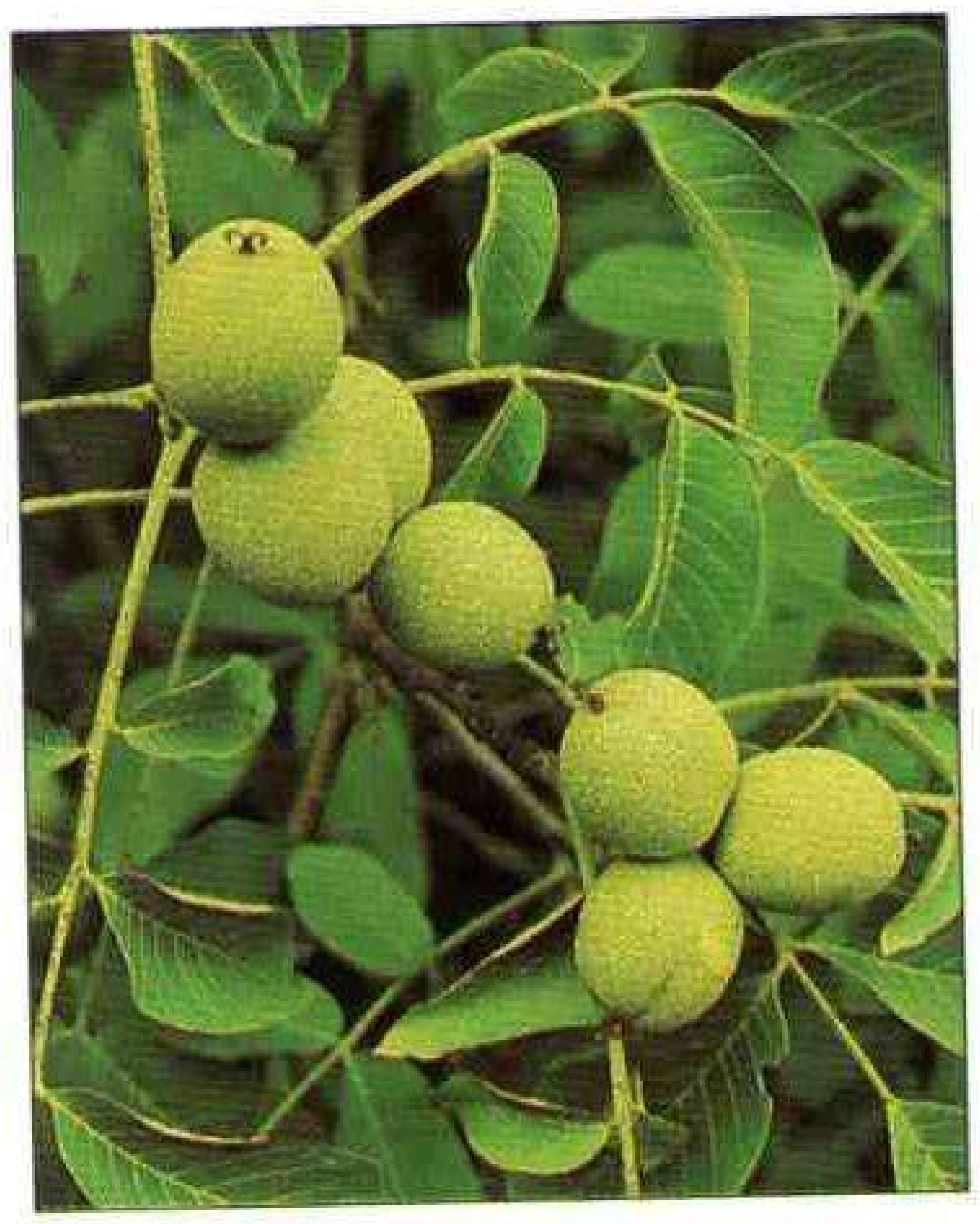

Nogal (Juglans). *Los nogales crecen en áreas templadas. Tienen una corteza especialmente nudosa. Las nueces tienen una cáscara verde y con pulpa, que se pudre después de caer del árbol. Las nueces tienen un alto contenido calórico.*

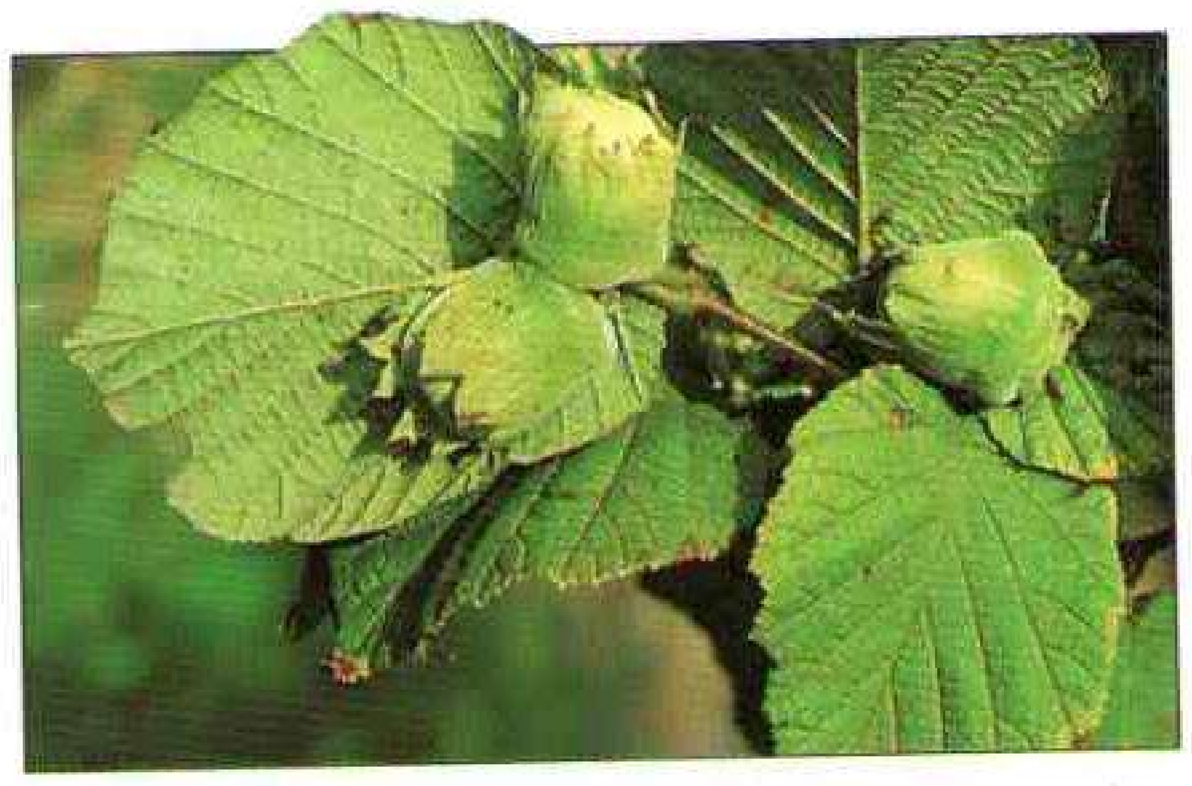

Avellanos (Corylus). *Las avellanas crecen en grandes arbustos y árboles de áreas templadas, de hojas serradas y acorazonadas, y amentos de color amarillo marrón. Las nueces crecen en cáscaras pilosas.*

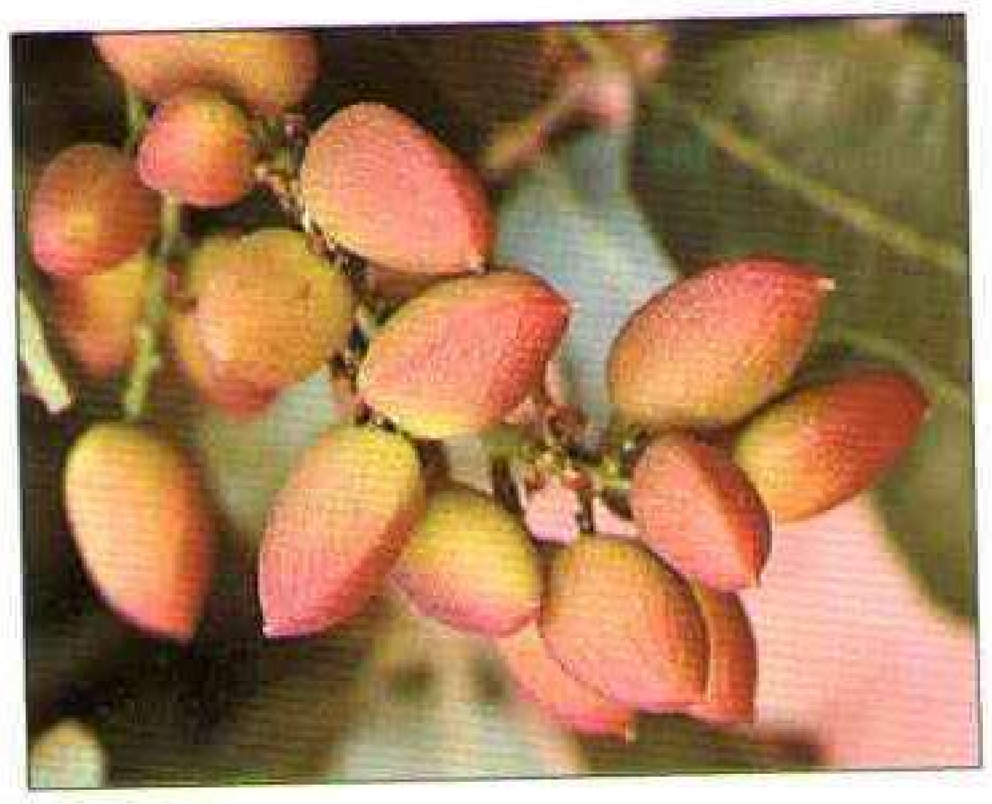

Pistacho (Pistacia). *Los pistachos crecen en Europa, Asia y Australia. Sus nueces rojas se comen crudas o tostadas.*

OTRAS NUECES COMESTIBLES

La pacana *(Carya)* es relativamente común en las áreas húmedas de Norteamérica. Contiene más grasa que cualquier otra nuez, vegetal o fruto. Las nueces de macadán *(Macadamia)* de Australia también tienen un sabor agradable. Los piñones *(Pinus)* son comestibles.

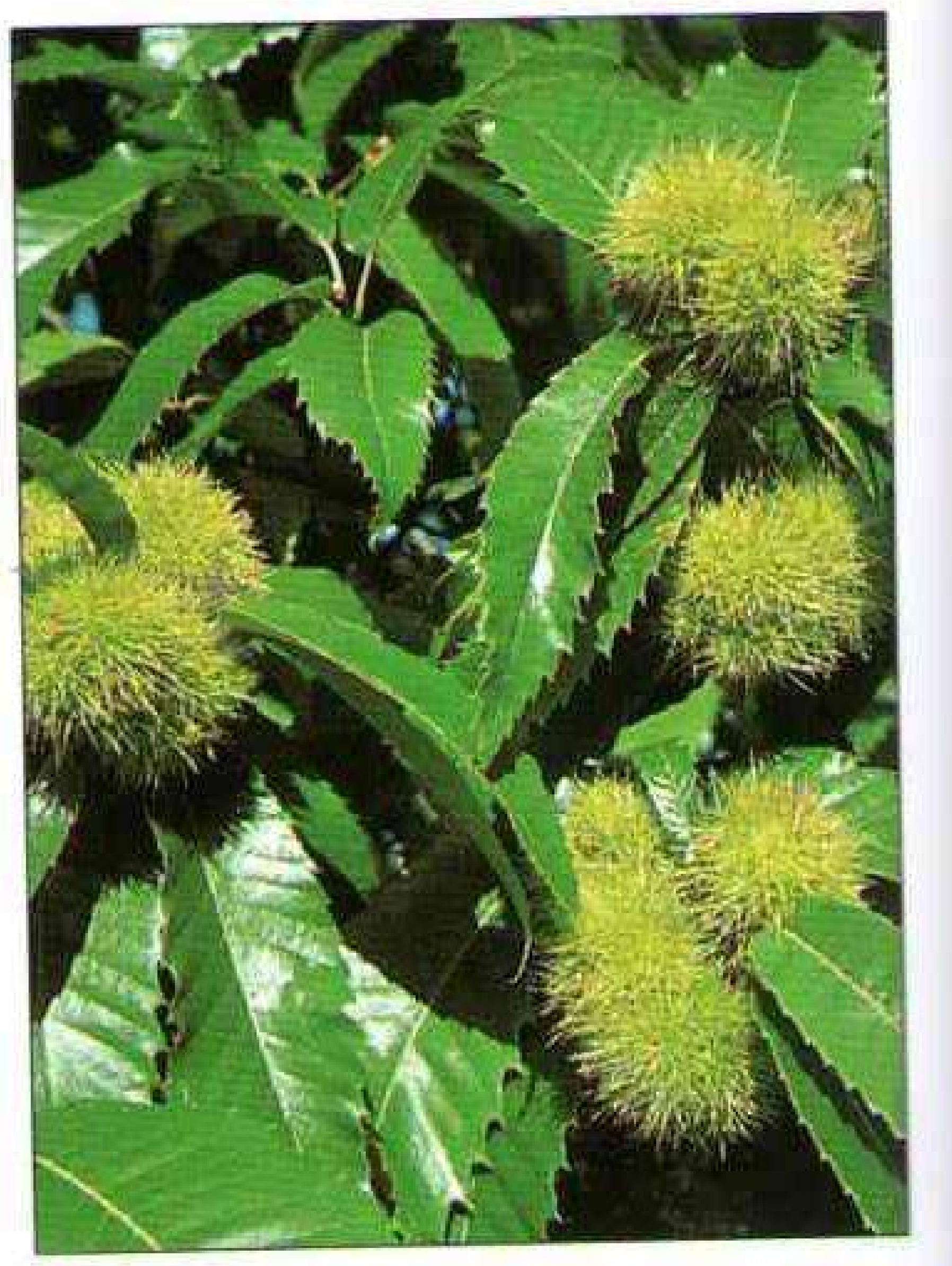

Castaño (Castanea). *Los castaños tienen largas hojas serradas y cáscaras espinosas, dentro de las cuales se desarrollan las castañas. Para comer las castañas debe pelarlas, hervirlas y reducirlas a una pasta, o tostarlas al fuego.*

NUECES VENENOSAS

Pruebe las nueces con cuidado, y utilice la prueba del sabor *(véase página 83)*. Aunque hay pocas nueces venenosas, algunas pueden ser peligrosas a menos que las hierva, mientras que otras le causarán malestar si están mohosas. Si una nuez tiene un sabor fuera de lo común, dulce o semejante al de las almendras, deséchela porque puede contener ácido prúsico, que causa irritación.

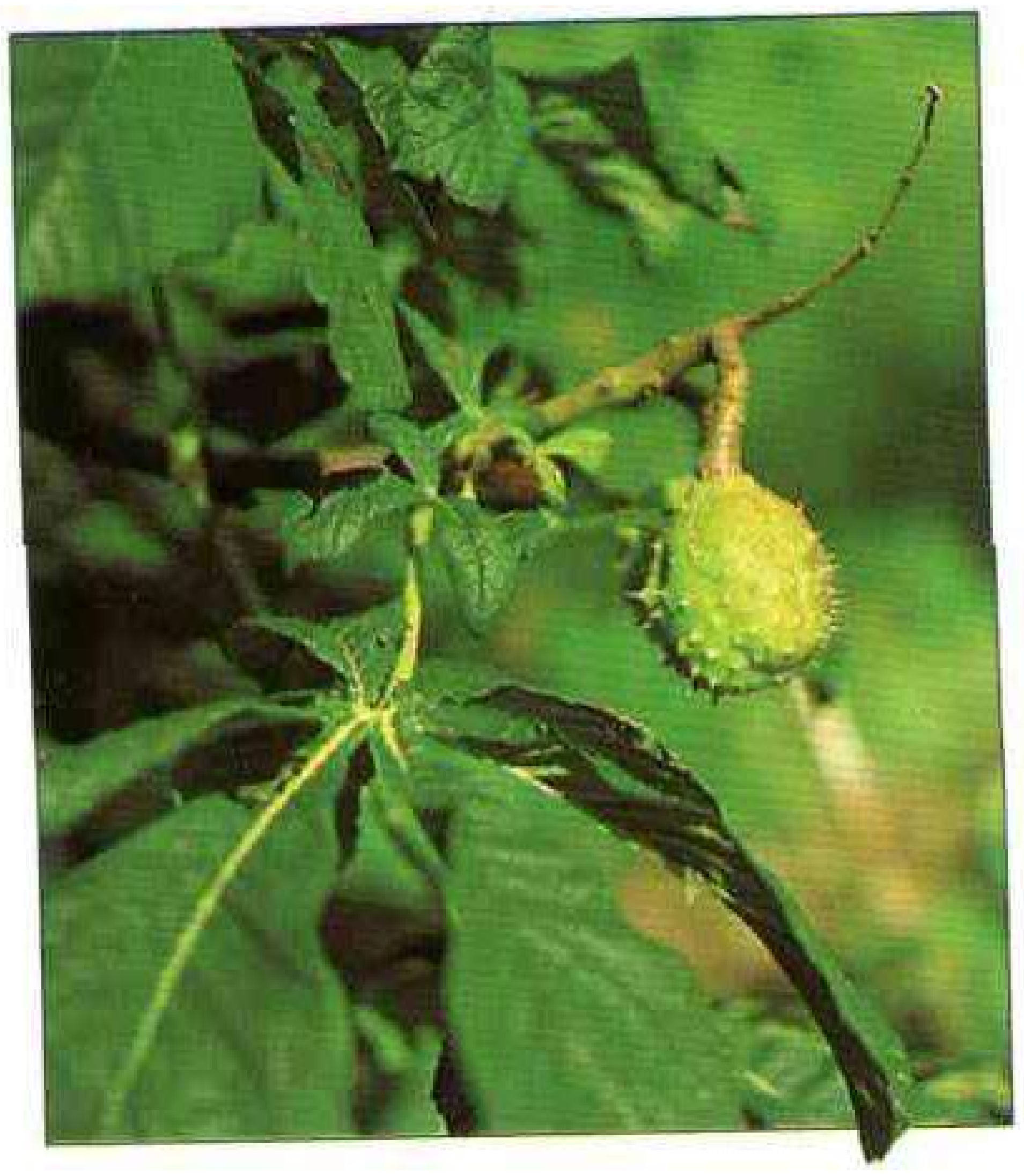

Castaño de Indias (Aesculus). *El castaño de Indias, que crece en las regiones templadas, tiene grandes hojas palmeadas, flores pálidas y yemas pegajosas. Puede alcanzar los 30 m de altura en algunas áreas. Sus hojas venenosas tienen cáscaras menos espinosas que el castaño común, y las nueces dentro de las cáscaras son mucho mayores que las castañas comestibles.*

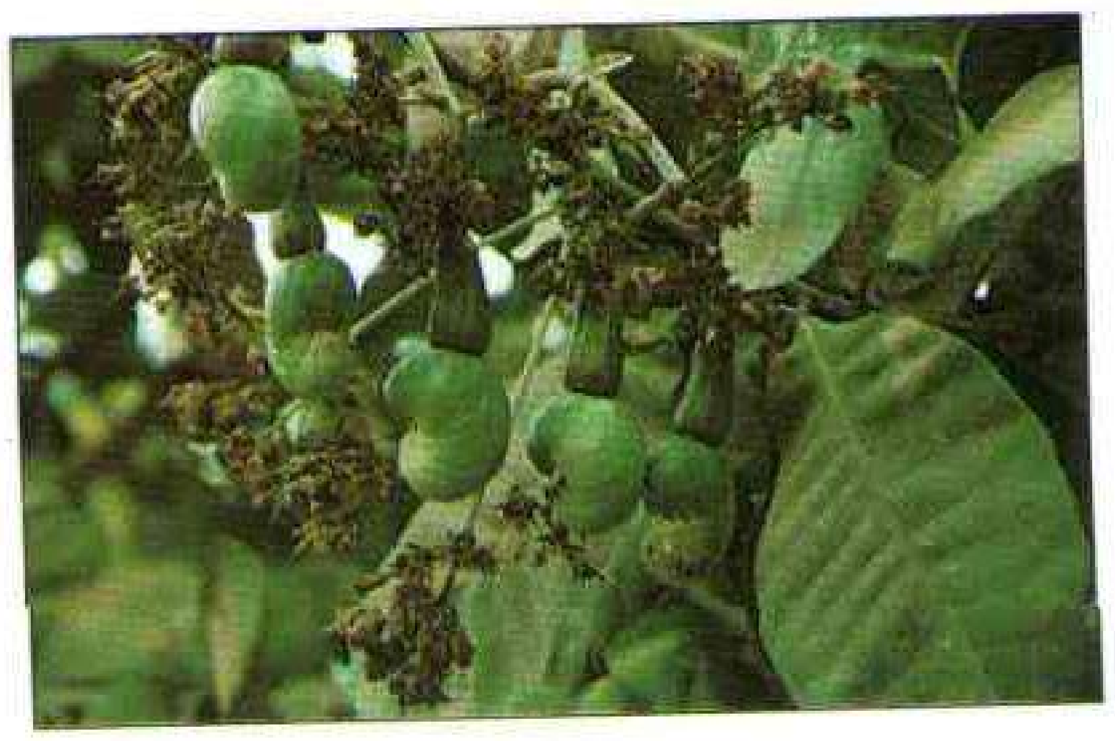

Anacardo (Anacardium). *Las nueces del anacardo son venenosas a menos que las pele y hierva. Tenga mucho cuidado mientras hierven, ya que los vapores pueden producir ceguera. El anacardo crece en áreas tropicales.*

OTRAS NUECES VENENOSAS

Aunque son comestibles, las almendras *(Prunus)* pueden contener ácido prúsico; tienen como resultado un sabor amargo. Las bellotas de los robles *(Quercus)* también son comestibles, pero deben hervirse o remojarse en agua fría durante varias horas, para después tostarlas a fin de eliminar su sabor amargo.

FRUTOS COMESTIBLES

Los frutos comestibles contienen muchos nutrientes esenciales, que incluyen azúcares y vitaminas A, B_2 y C. Las aves y los mamíferos también lo saben, por lo que tendrá que competir con ellos. Los insectos también atacan la fruta, pero puede beneficiarse de ello comiéndose cualquier gusano nutritivo que llegue a contener. No coma cantidades excesivas de cualquier tipo de fruta.

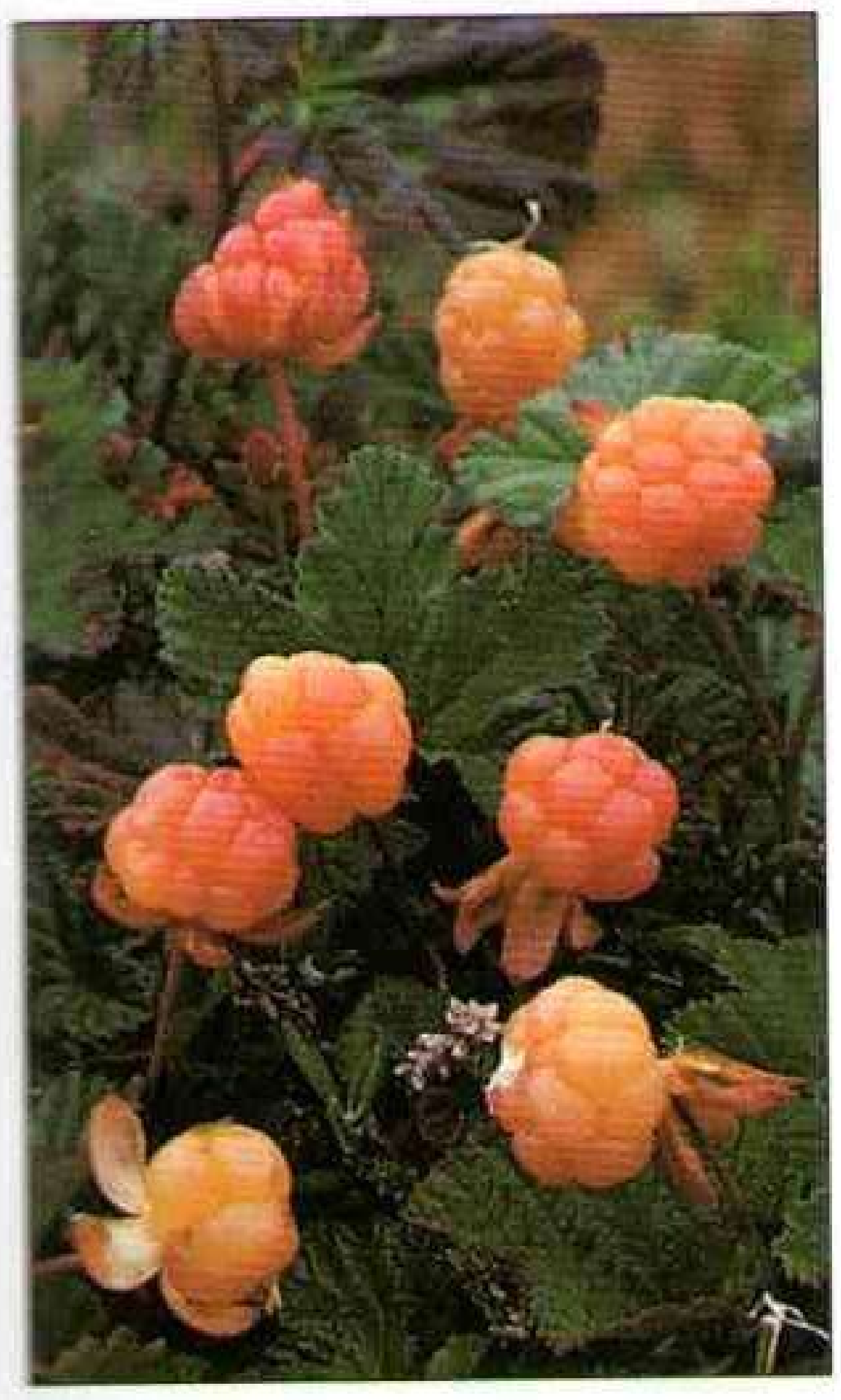

Camemoro (Rubus). *El camemoro crece en el ártico y en las regiones templadas del norte. Alcanza una altura de unos 30 cm, tiene flores blancas y frutos que se asemejan a las frambuesas.*

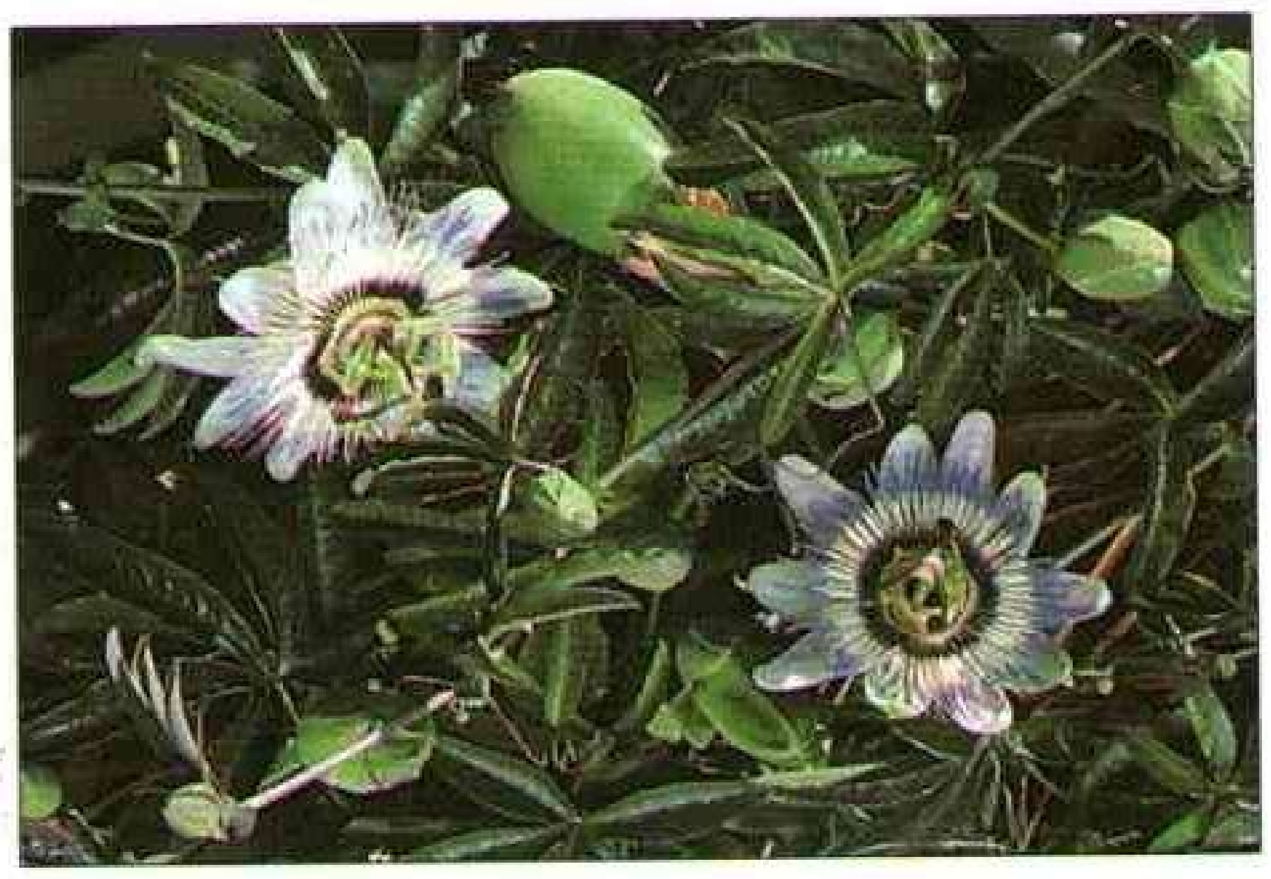

Fruta de la pasión (Passiflora). *La fruta de la pasión tiene largos tallos reptantes, hojas trilobuladas y flores de color blanco y violeta que crecen sobre zarcillos. Crece en áreas tropicales, sobre todo en Sudamérica. Los frutos comestibles son de color violeta, ovalados y jugosos.*

OTROS FRUTOS COMESTIBLES

Las uvas silvestres *(Vitis)* son comunes en las zonas cálidas. Sus hojas pueden comerse hervidas. Varias especies de fresa silvestre *(Fragaria)* se encuentran en los bosques y en zonas de hierbas secas. Son ricas en vitamina C, muy dulces y deliciosas.

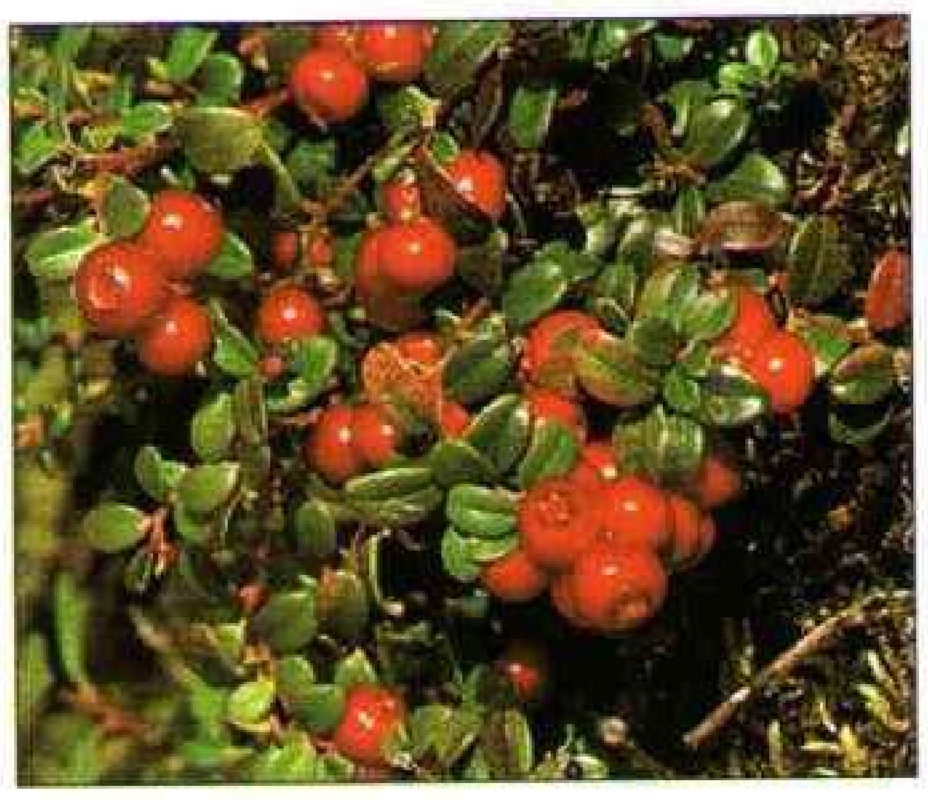

Arándano (Vaccinium). *Crece en los brezales templados, en la tundra y en los bosques.*

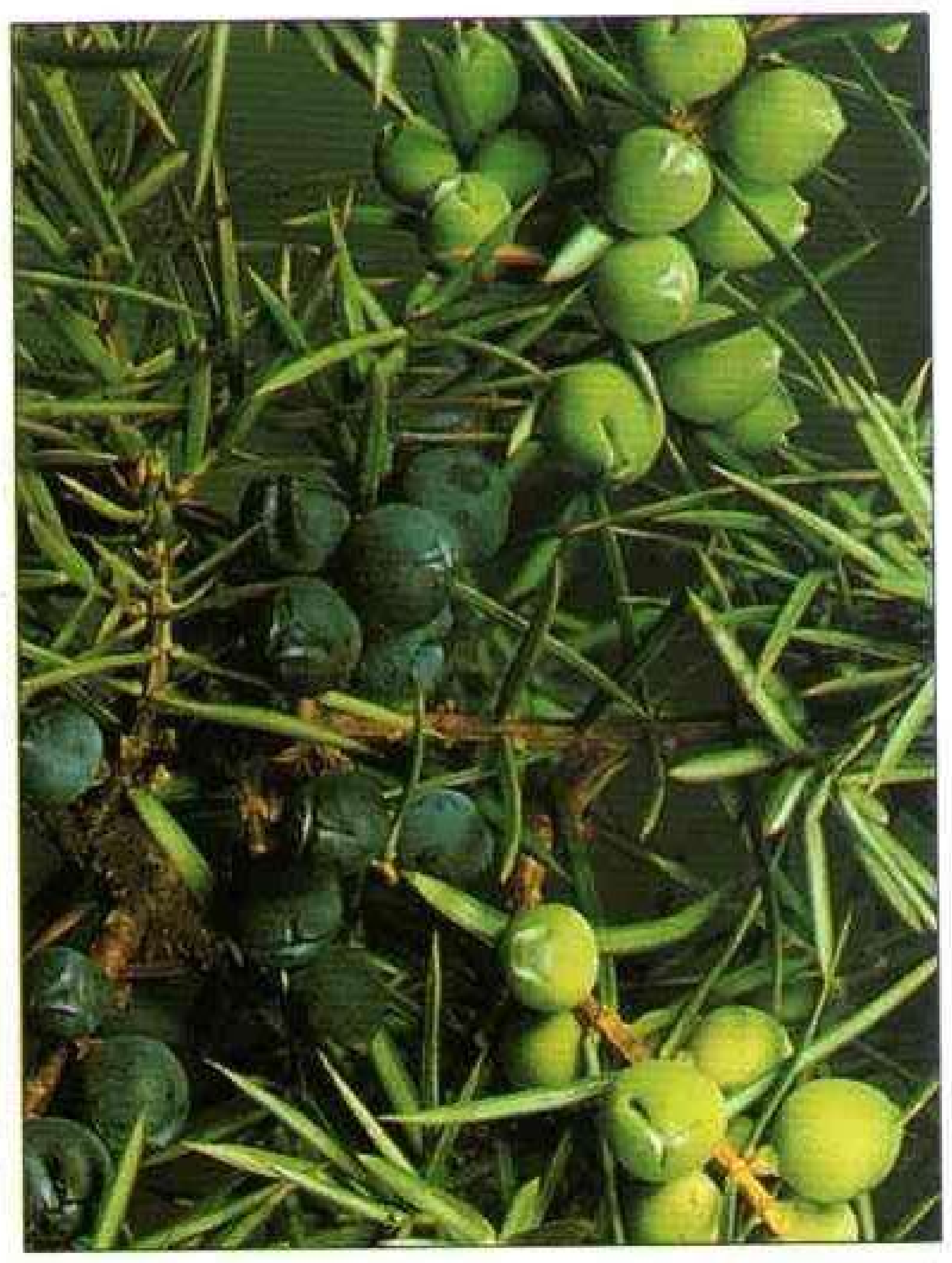

Enebro (Juniperus). *El enebro crece en las montañas templadas del norte, en forma de arbusto grande o pequeño, o en densos matorrales bajos. En algunas partes de Norteamérica se le conoce como* uvas de Oregón. *Las bayas maduras de color azul oscuro son comestibles, aunque un poco amargas.*

FRUTOS VENENOSOS

Ya que los frutos están diseñados para atraer a los animales y a los pájaros (los cuales, una vez han comido el fruto, diseminan las semillas en sus heces), muchas bayas venenosas se asemejan a las comestibles. Esto puede ser una amenaza para quienes intentan sobrevivir a base de alimentos naturales. Al igual que las demás plantas y setas, debe aprender a reconocer algunos frutos comestibles.

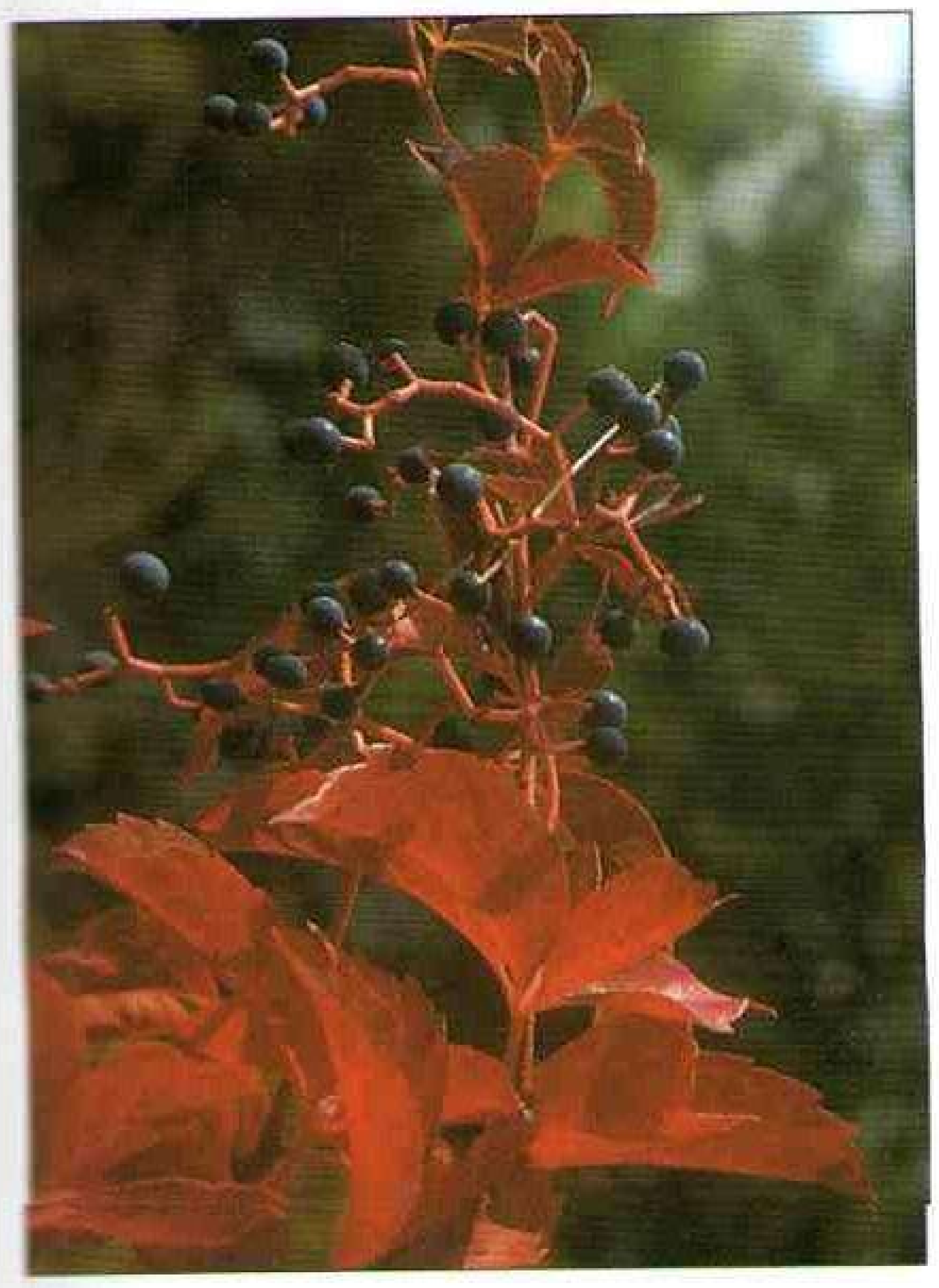

Enredadera de Virginia (Parthenocissus). *Las bayas de profundo color azul o púrpura de esta planta de clima templado se asemejan a las uvas. Ninguna baya azul o púrpura de una planta trepadora con zarcillos es comestible.*

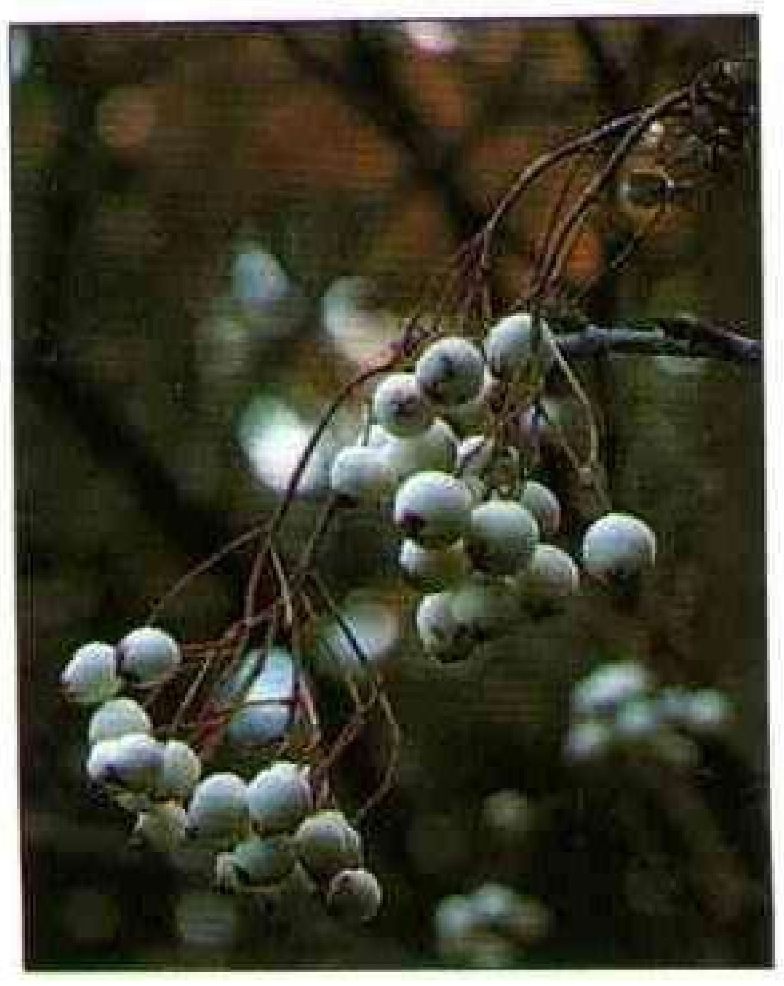

Serbal. *Aunque los frutos de algunas especies de* Sorbus *son comestibles, en general las bayas blancas son siempre venenosas. No se arriesgue a comer ninguna baya de este color. Otras plantas con frutos blancos venenosos son la hierba de San Cristóbal y algunos tipos de belladona.*

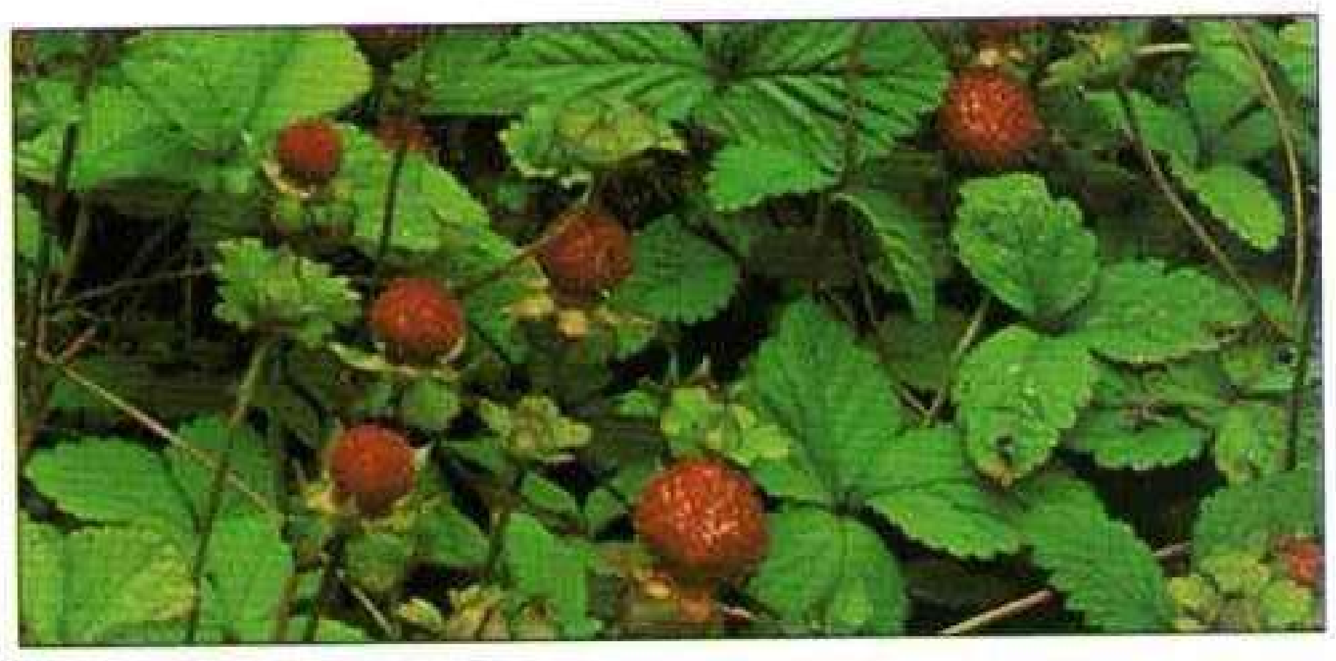

Duchesnea. *Los frutos de* Duchesnea, *también conocida como* falsa fresa, *pueden ser mortales. La planta crece en las áreas tropicales de Asia y Norteamérica.*

OTROS FRUTOS VENENOSOS

Existen muchos frutos venenosos, en particular en las regiones tropicales. Algunos de ellos son sólo medianamente venenosos y pueden provocar diarreas y náuseas al ingerirse en grandes cantidades. Otros son altamente tóxicos y una pequeña cantidad basta para causar la muerte. En una situación de supervivencia, los riesgos se incrementan, especialmente si se encuentra débil. Algunos frutos venenosos comunes incluyen las bayas azules de la mortal belladona *(Atropha)*, que son letales si se ingieren. *Arum* tiene bayas rojas que pueden causar un grave malestar digestivo si se ingieren. La hierba de San Cristóbal *(Actaea)* tiene bayas blancas o negras que pueden producir vómitos e irritación interna. Los frutos de la manzana silvestre *(Malus)* pueden causar diarrea si se comen en grandes cantidades. El árbol tropical de la estricnina *(Strychnos)* tiene frutos que parecen naranjas y sus semillas son mortales.

HONGOS Y LÍQUENES

PARA MUCHAS PERSONAS en todo el mundo, la recogida de hongos y líquenes se ha convertido en una actividad habitual. Además de ser absolutamente deliciosos, los hongos frescos contienen minerales y algunas proteínas, y constituyen el elemento básico de muchos platos nacionales. Sin embargo, debe tener una gran precaución al recogerlos y prepararlos. Algunos hongos son letales, y muchos se asemejan a las variedades comestibles. Incluso los recolectores experimentados pueden equivocarse: en ocasiones, lo que no representa ningún peligro en un país se asemeja mucho a una variedad venenosa en otro. Consuma hongos únicamente si está seguro de poder identificarlos correctamente.

HONGOS COMESTIBLES

No existe una manera lógica de identificar los hongos comestibles. Recójalos únicamente si va acompañado de un experto y aprenda a identificar una o dos variedades comestibles, con referencia especial a sus hábitats. Aprenda a reconocer las variedades venenosas, en particular las que son mortales. Rechace cualquier hongo que no esté en perfectas condiciones. Aquí se muestran algunos comestibles.

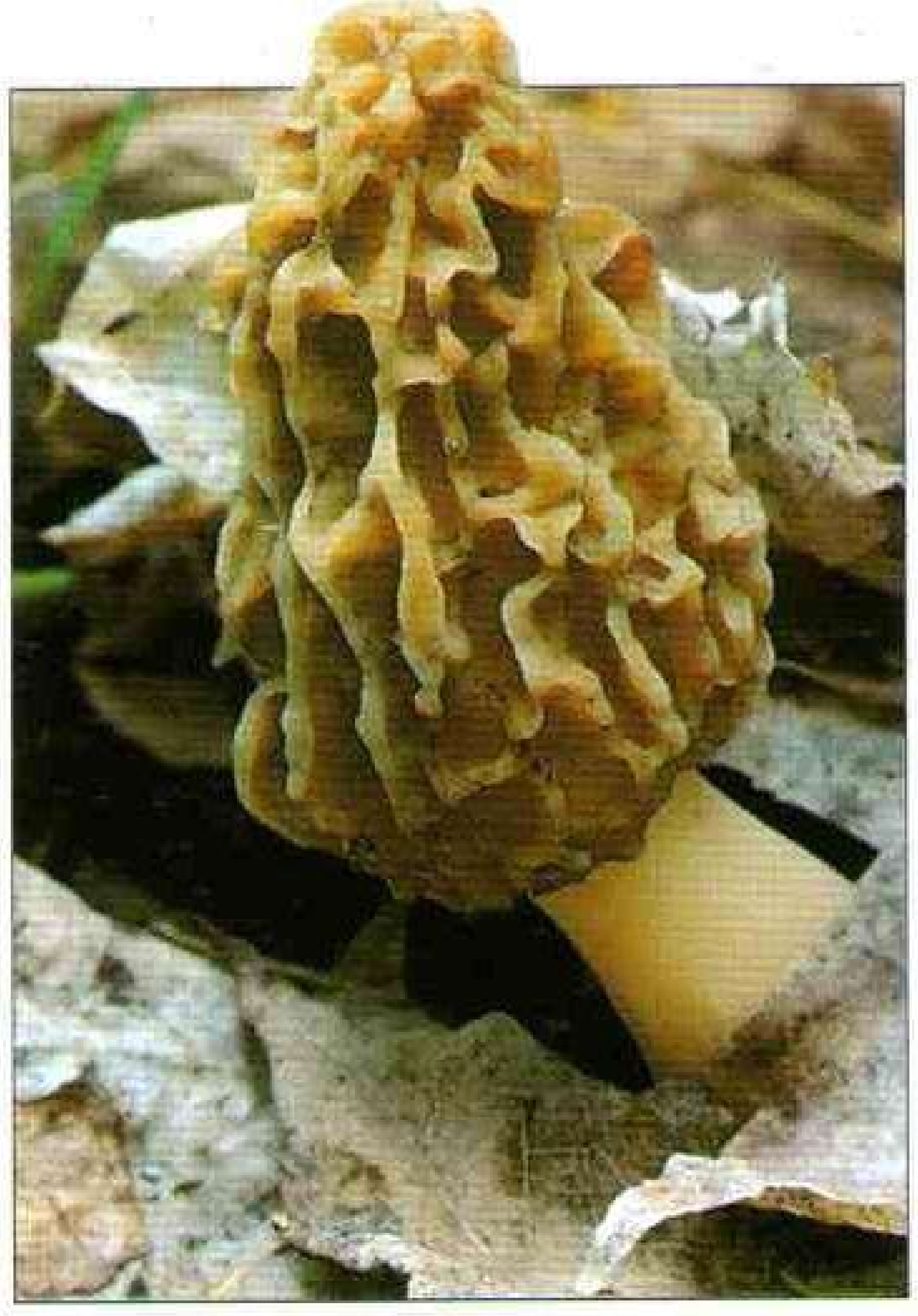

Colmenilla redonda, morilla (Morchella esculenta). *Este hongo blanco crece en primavera en suelos arenosos o arenoso-arcillosos, bajo los árboles o al descubierto.*

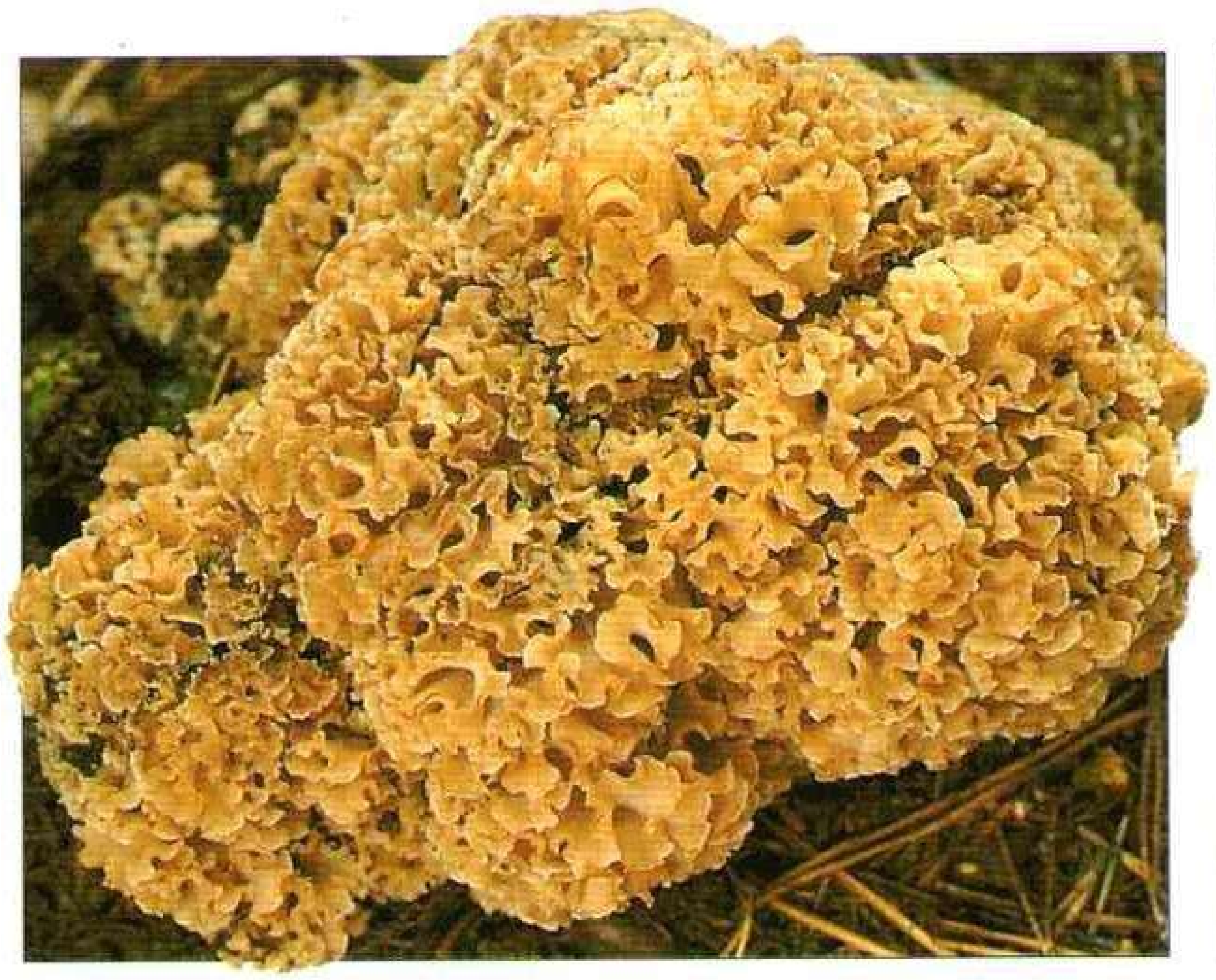

Seta coliflor (Sparassis crispa). *Este hongo se encuentra en otoño en los bosques de coníferas; crece al pie de los árboles y sobre sus raíces. Su carne despide un característico olor a anís, y tiene un sabor similar al de una nuez.*

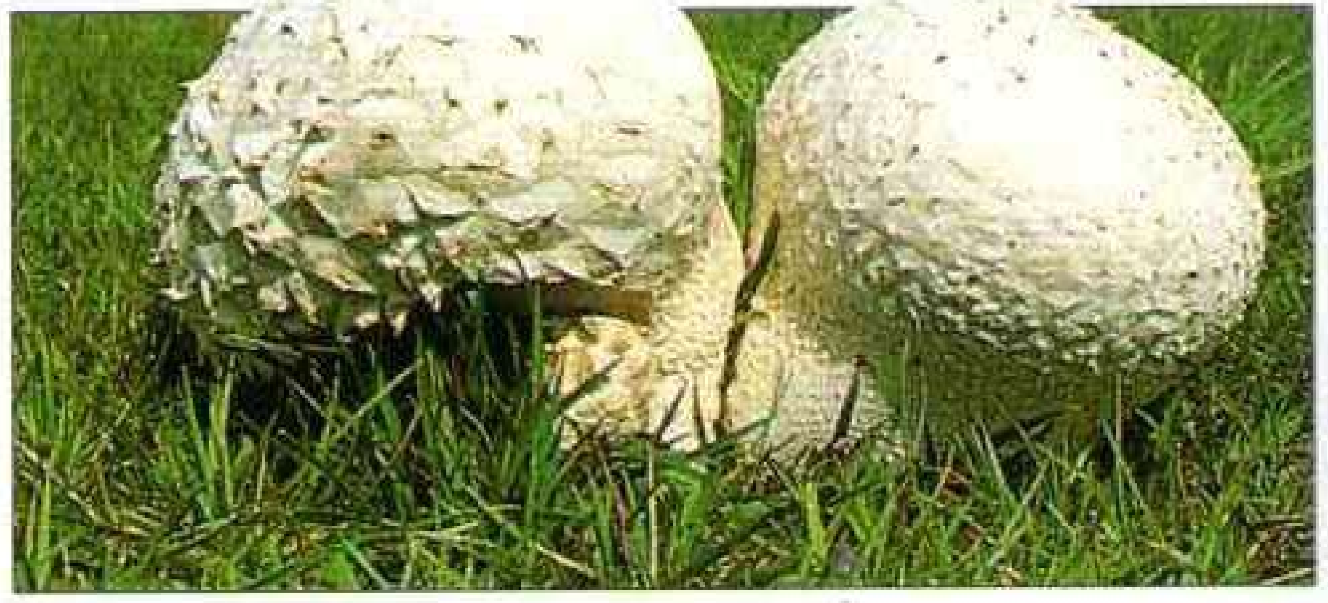

Bejín areolado (Calvatia utriformis). *Se encuentra en verano y en otoño en los campos y prados. Es preferible consumirlo joven, cuando la carne todavía se mantiene blanca.*

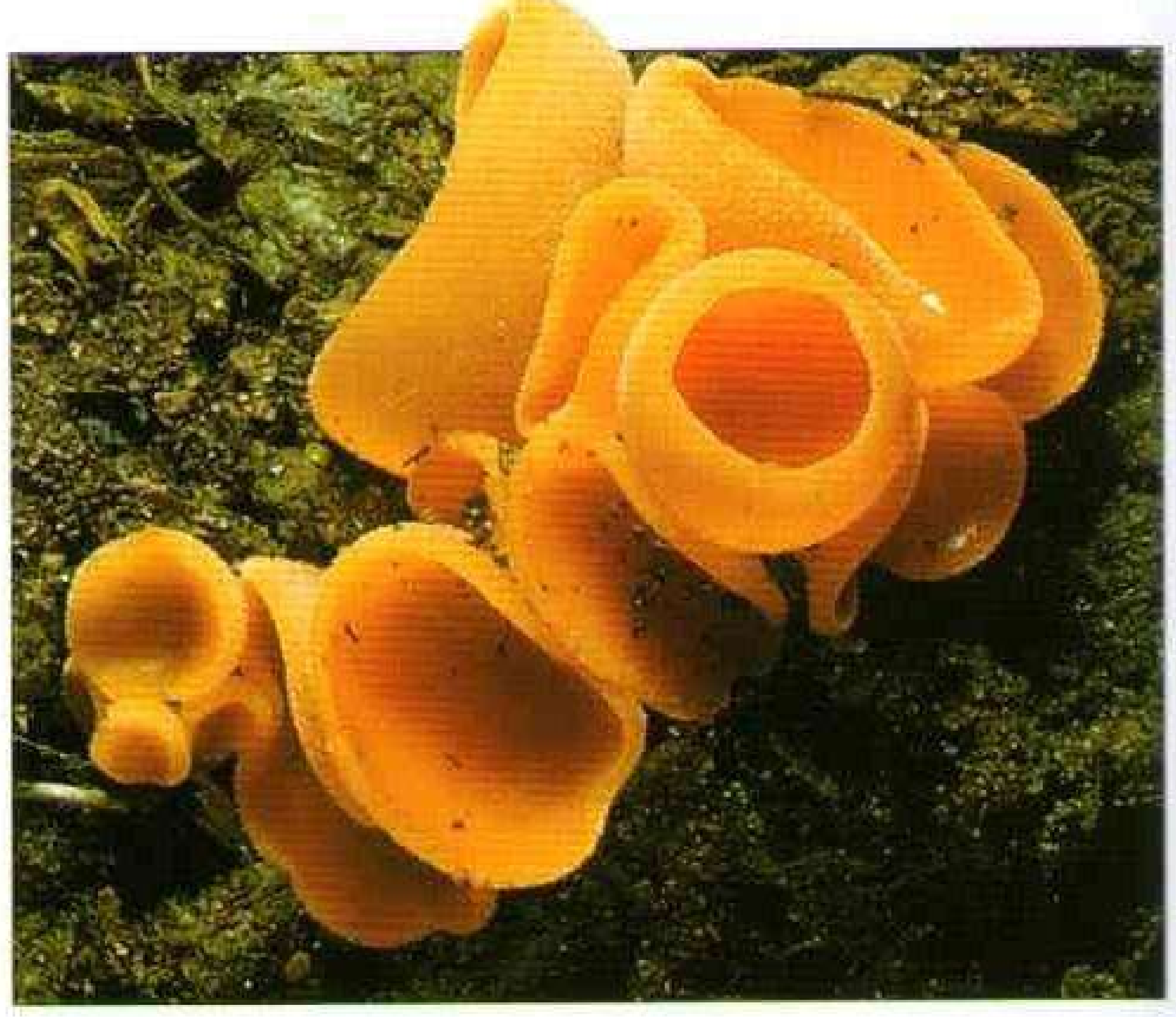

Peziza anaranjada (Aleuria aurantia). *Este hongo necesita luz solar para crecer, y se encuentra en los claros de las regiones boscosas y en áreas de pasto en otoño. A pesar de ser comestible, su carne no tiene un olor o sabor destacable.*

OTROS HONGOS COMESTIBLES

Las variedades *Boletus* tienen la forma tradicional de una seta, pero tienen poros en la parte inferior del sombrero, en lugar de láminas. Generalmente son deliciosos. Algunas especies son medianamente venenosas, pero pueden distinguirse de las comestibles ya que su sabor en crudo es desagradable, y tienen los poros y el pie de color rojizo o naranja. Elija únicamente las setas con poros amarillos o de color crema. Los hongos que crecen sobre los troncos de los árboles son generalmente comestibles, pero algunas variedades son excesivamente duras y con poco sabor. Todos los hongos silvestres deben hervirse antes de consumirse, para así destruir los compuestos ligeramente tóxicos. Recuerde, sin embargo, que no con sólo cocinarlos convertirá en comestible una especie venenosa. Los hongos difieren entre sí según la región, por lo que debe aprender a reconocer los de su entorno, o bien los de la región que pretenda visitar.

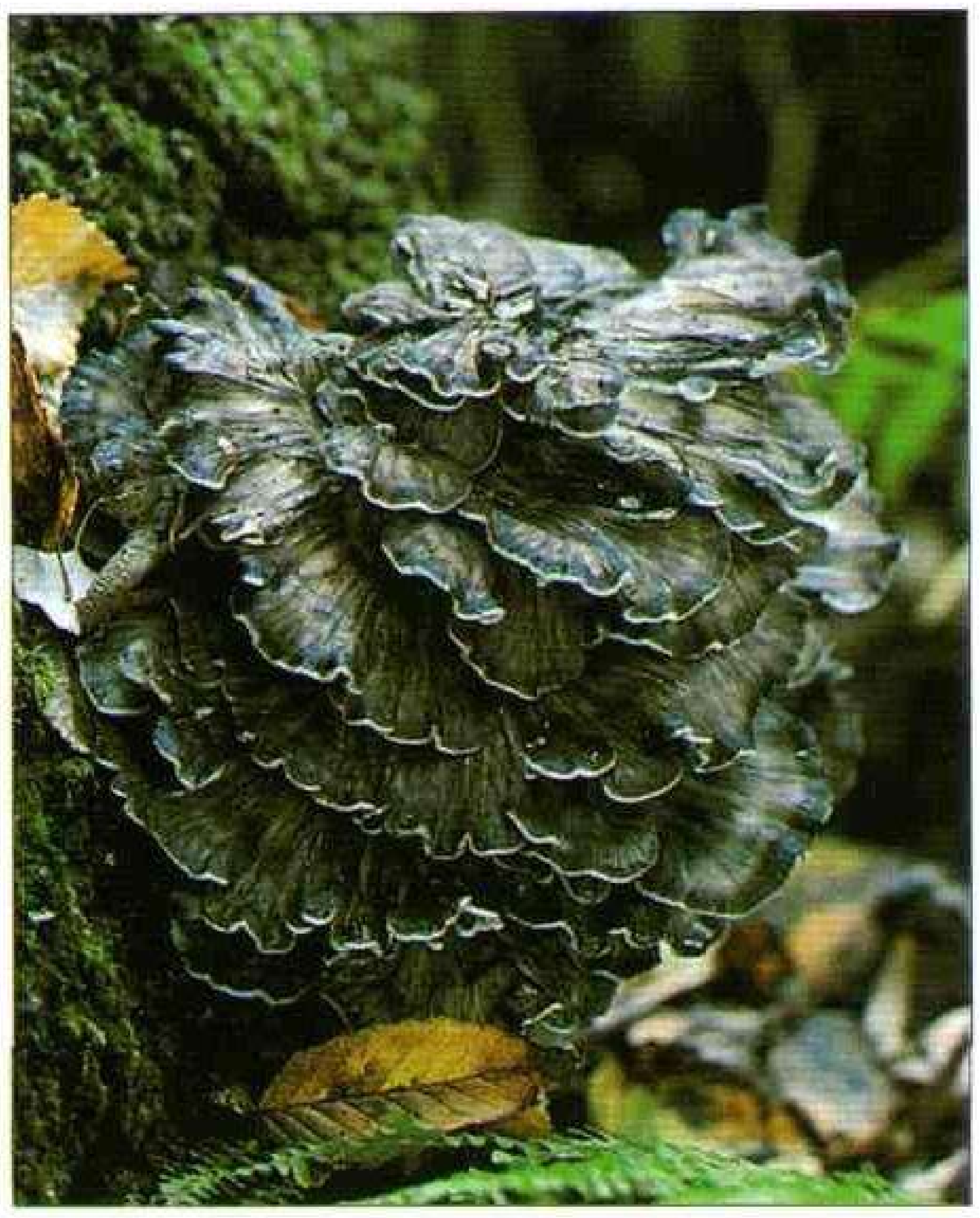

Políporo frondoso (Grifola frondosa). *Este hongo de árbol, que puede encontrarse desde la primavera hasta el otoño, crece en árboles de hoja caduca. Despide un característico olor a seta, y tiene un sabor dulce.*

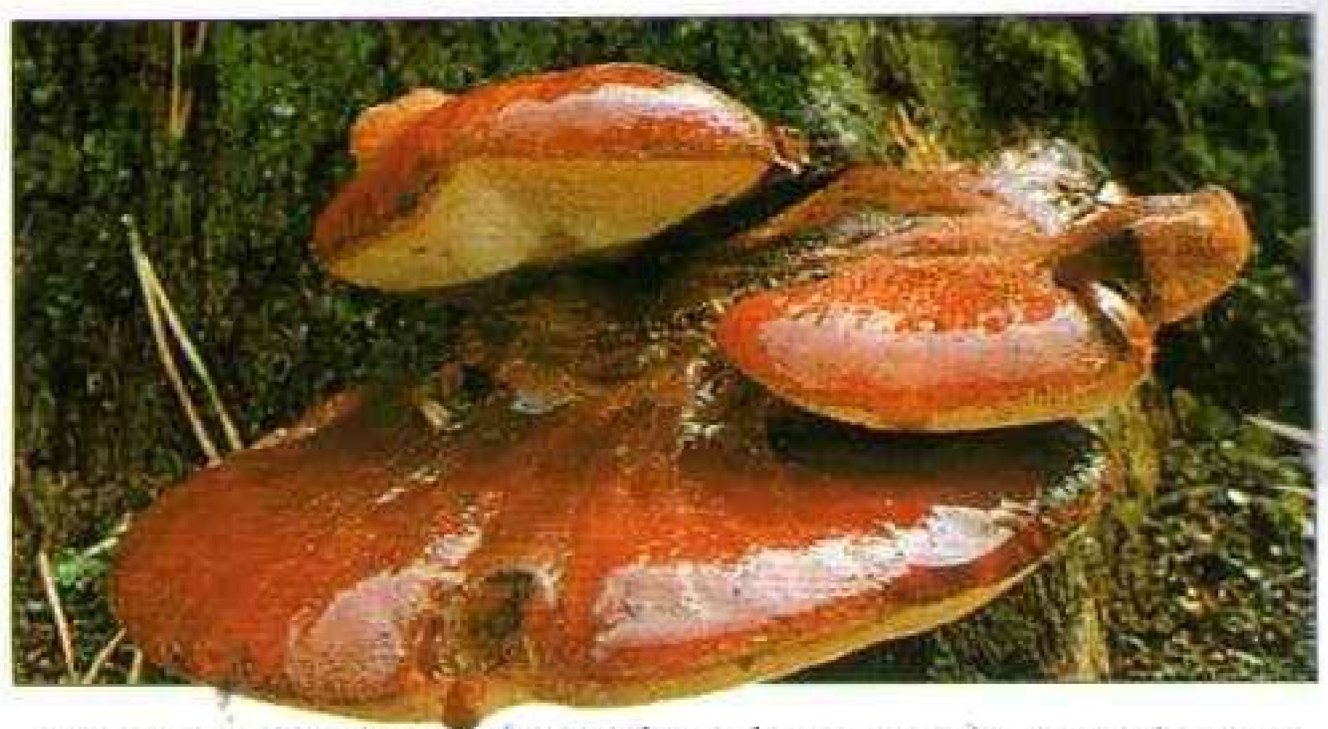

Lengua de buey (Fistulina hepatica). *Este hongo es fácil de encontrar durante el otoño, sobre los robles. Su carne roja tiene un sabor amargo en crudo, por lo que debe remojarlo y cocerlo antes de consumirlo.*

HONGOS VENENOSOS

No hay características específicas que permitan distinguir los hongos venenosos. Las amanitas se encuentran entre los más peligrosos, y son las más fáciles de identificar, aunque algunas especies tienen un aspecto totalmente diferente. Peor aún, algunos hongos venenosos se asemejan a los hongos comestibles de otros países. Los síntomas de envenenamiento por hongos son muy variados. A menos que pueda identificar positivamente un hongo como variedad comestible, no lo toque. A continuación se muestran algunas de las especies más peligrosas.

ADVERTENCIA

Aunque los hongos del género *Amanita* incluyen algunos de los más mortales, no todos los hongos venenosos se les parecen. Siempre debe realizar una identificación positiva de una variedad comestible, rechazando las demás a menos que esté bien seguro de que son comestibles.

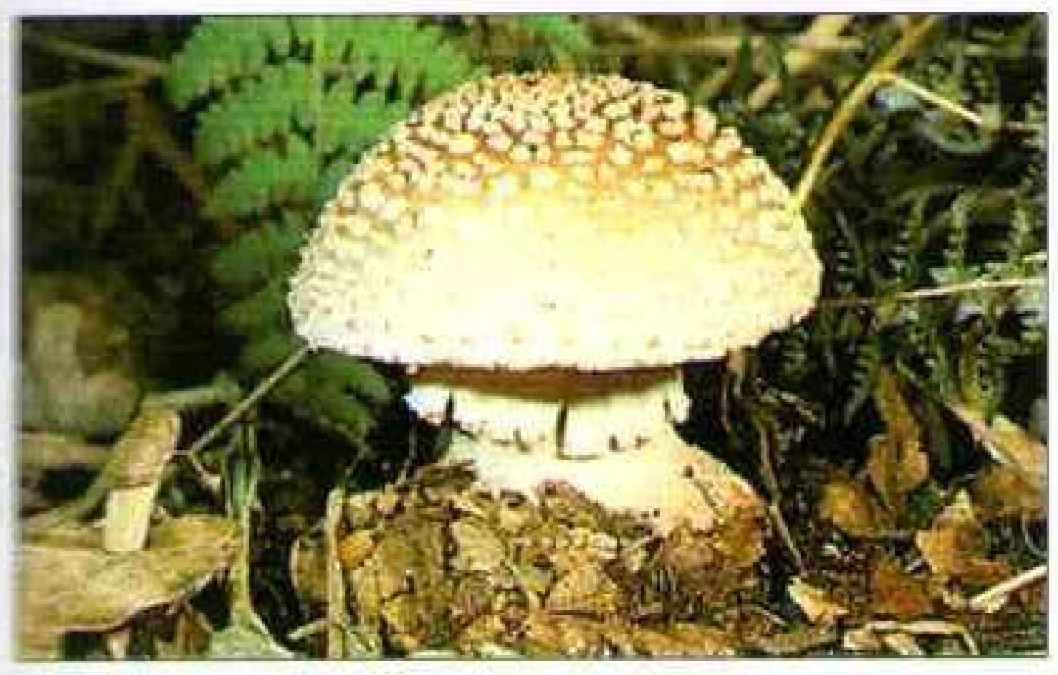

Amanita pantera (Amanita pantherina). *Suele ser mortal. Se encuentra en los bosques de hayas; tiene un sombrero marrón moteado de blanco, láminas blancas y anillos alrededor del pie.*

CÓMO RECONOCER UN HONGO VENENOSO

Los hongos venenosos del género *Amanita* son relativamente fáciles de identificar. Todos tienen una volva en la base, láminas blancas y anillos alrededor del pie.

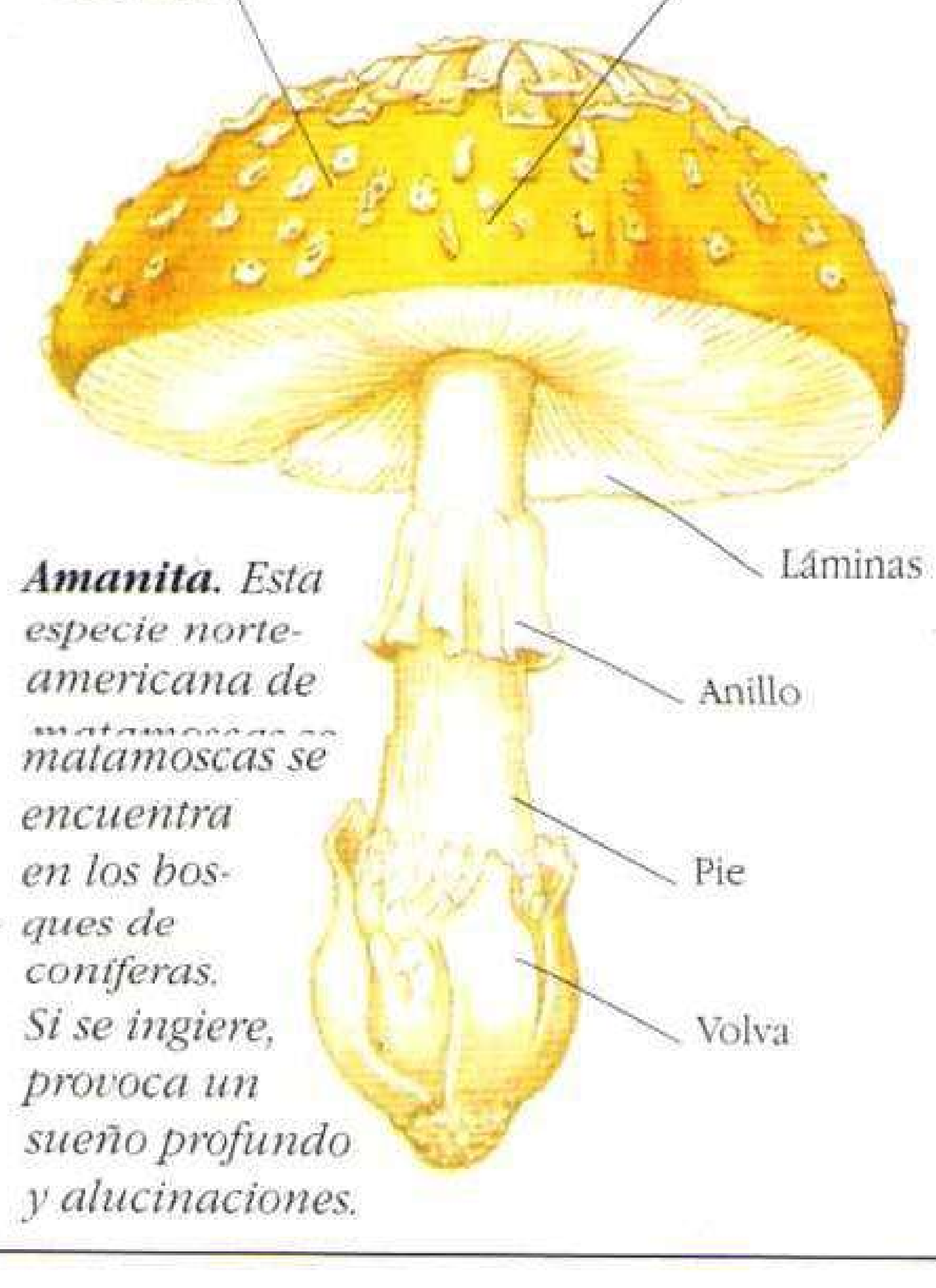

Amanita. *Esta especie norteamericana de matamoscas se encuentra en los bosques de coníferas. Si se ingiere, provoca un sueño profundo y alucinaciones.*

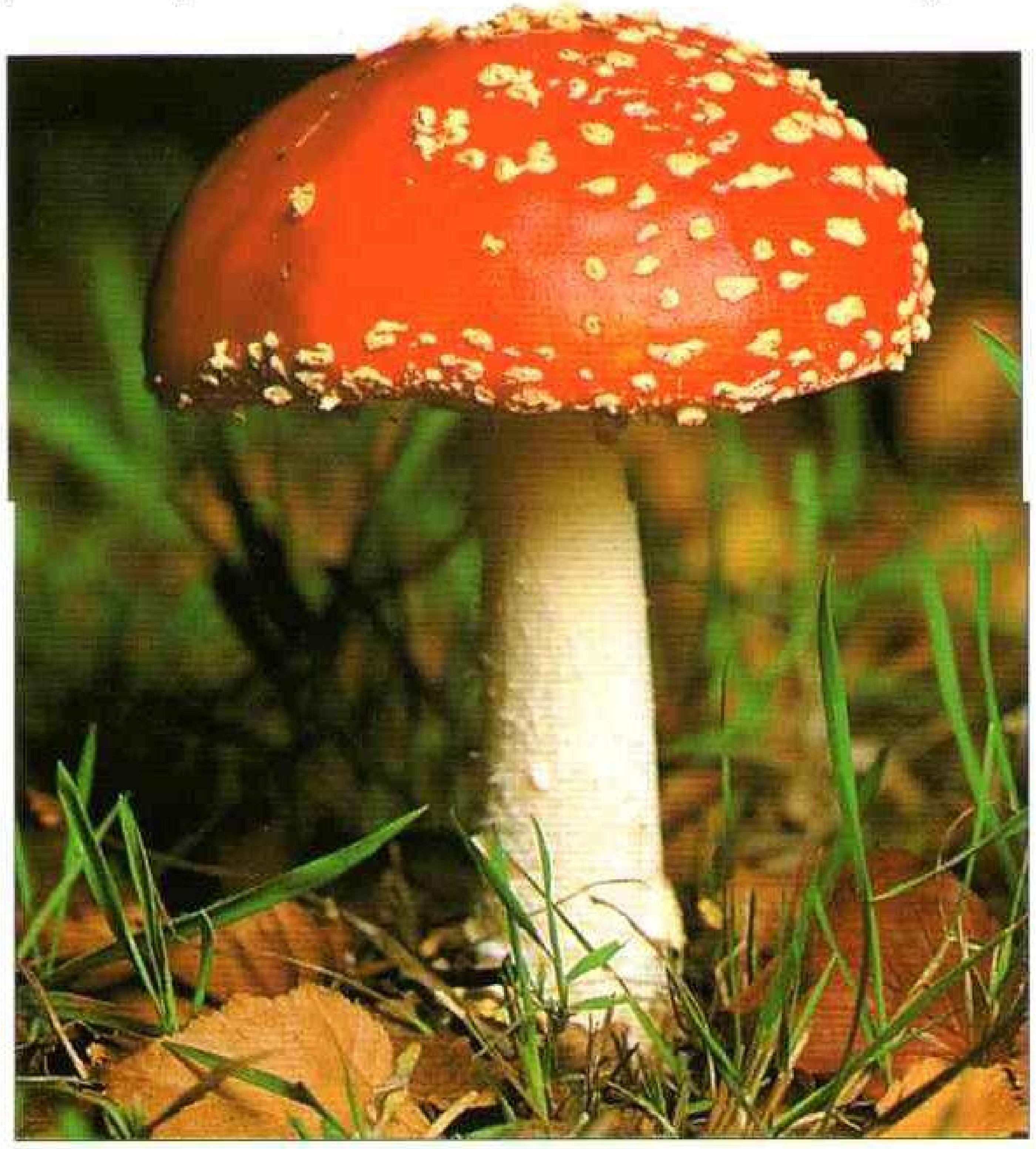

Matamoscas (Amanita muscaria). *Este hongo fácilmente reconocible crece durante el otoño, particularmente en los bosques de hayas y de pinos. Tiene un sombrero de color rojo brillante, moteado de blanco.*

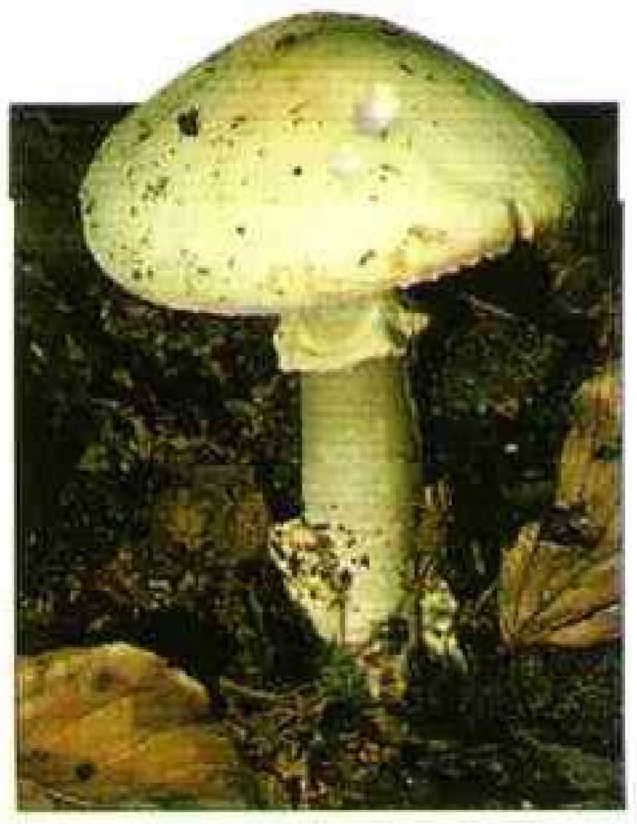

Amanita maloliente (Amanita virosa). *Es mortalmente venenoso, y se encuentra en verano y otoño en los bosques. Su carne blanca tiene un olor dulce y nauseabundo. Los especímenes jóvenes pueden parecerse a los* Agaricus *comestibles.*

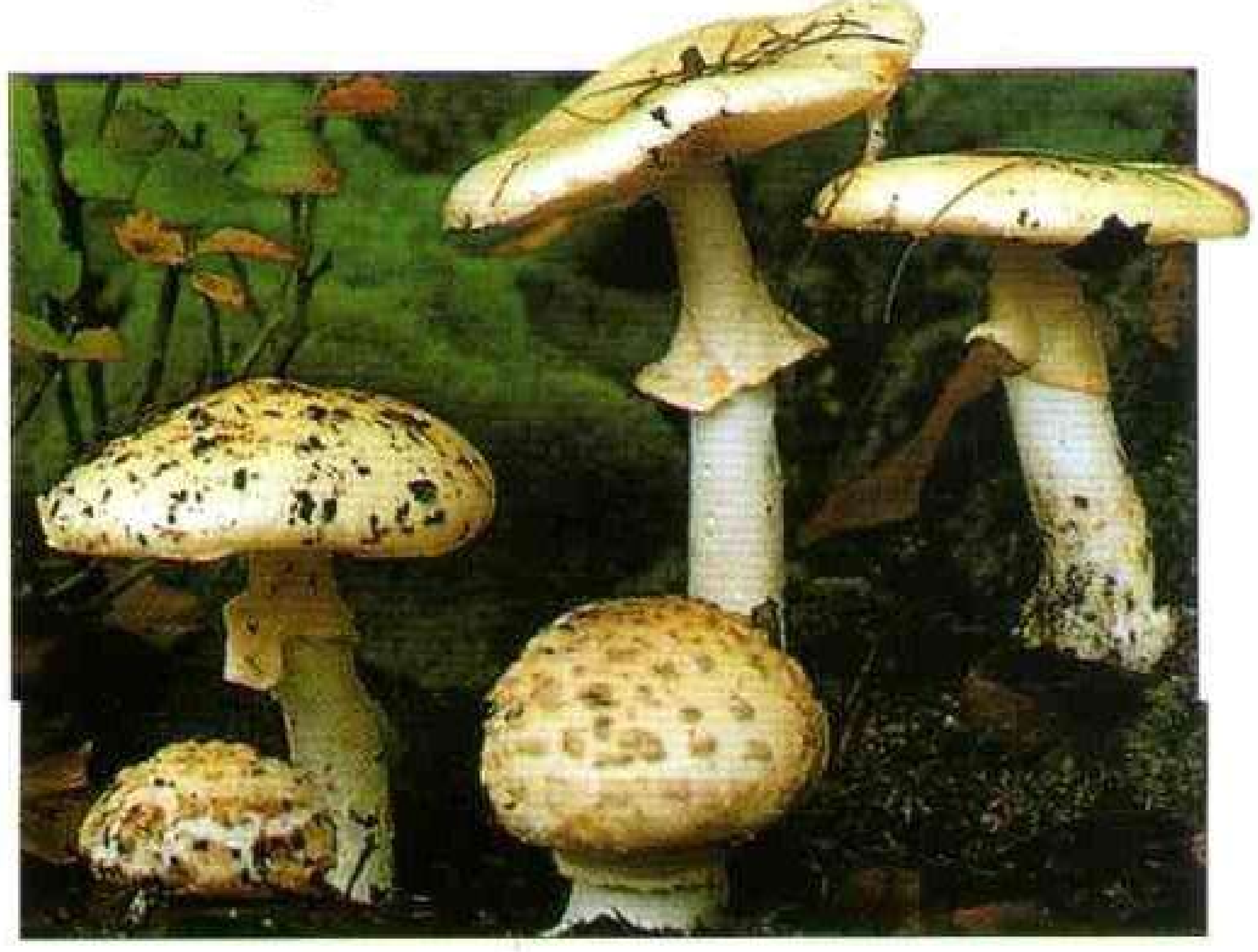

Cicuta verde (Amanita phalloides). *La especie más venenosa de todos los hongos se encuentra en las regiones boscosas, bajo hayas y robles. Tiene el sombrero de color verde oliva, el pie blanco, láminas y carne blancas, y una gran volva.*

OTROS HONGOS VENENOSOS

Algunas variedades venenosas parecen idénticas a las especies comestibles de otras regiones. Por ejemplo, mientras que muchas especies del género *Agaricus* son comestibles, el agárico amarilleante *(Agaricus xanthodermus)* es muy venenoso. Puede reconocerlo por su intenso olor a yodo y la coloración amarilla que aparece al tocarlo. Otras variedades venenosas incluyen la cicuta blanca *(Amanita verna)*, de color blanco, y que se encuentra en suelos arenosos bajo árboles de hoja ancha, así como varias especies de *Cortinarius*, algunas de las cuales son mortales.

LÍQUENES COMESTIBLES

Los líquenes crecen sobre las rocas, a menudo en las regiones boreales o árticas. No hay variedades venenosas, pero todos deben remojarse una noche en agua y hervirse después a fin de eliminar cualquier ácido irritante, antes de ingerirlos.

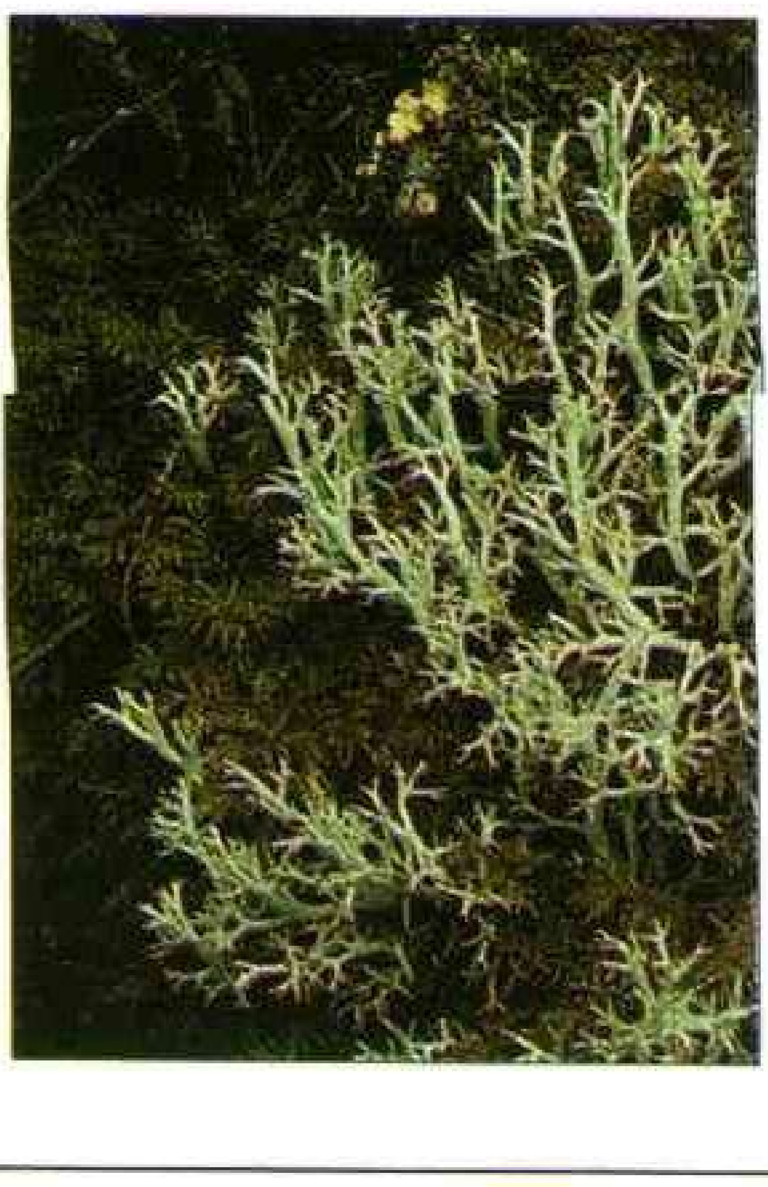

Liquen de los renos (Cladonia rangiferina). *Esta planta resistente crece en las regiones árticas y su aspecto recuerda los cuernos de los renos, de ahí su nombre. Antes de comerlo debe remojarlo en agua durante varias horas y hervirlo bien para que resulte digerible.*

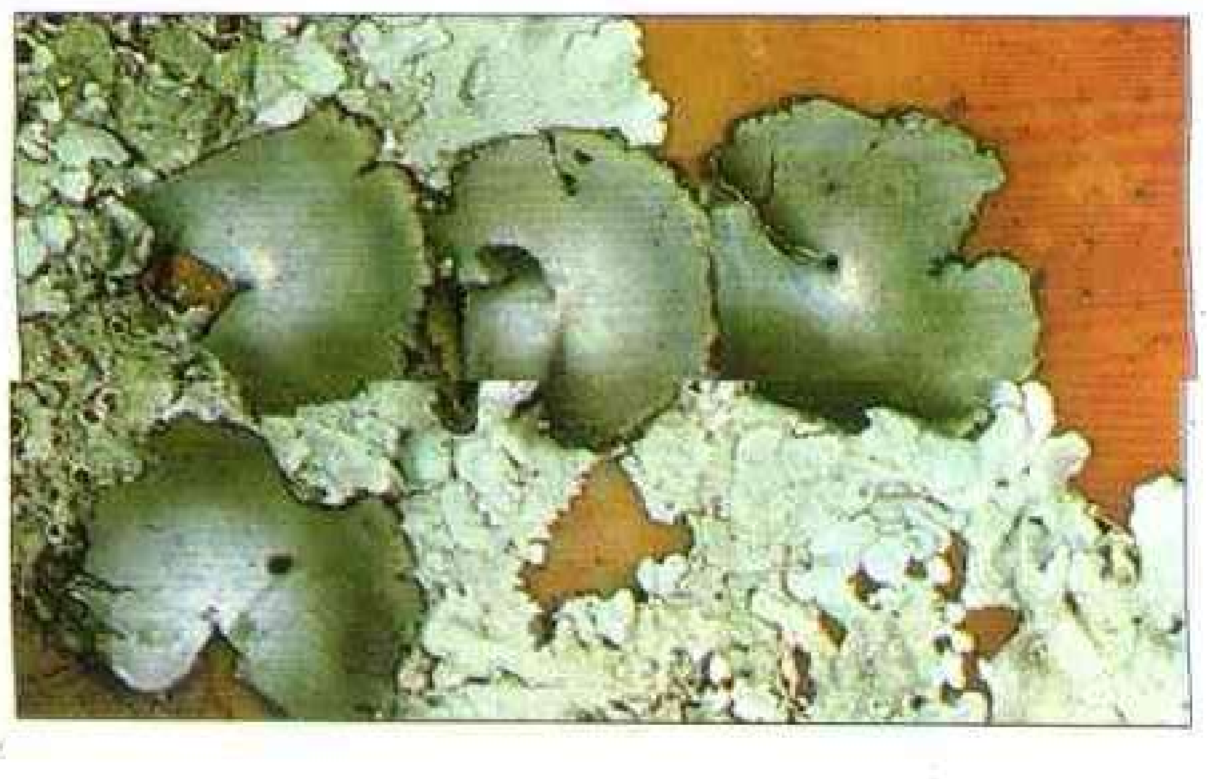

Callos de roca. *Estos líquenes nutritivos crecen en las regiones templadas del norte y en el ártico. Deben remojarse y hervirse antes de consumirse.*

LAS PLANTAS COMO ALIMENTO

LA RECOLECCIÓN NO consiste en recoger cualquier planta que parezca comestible. Debe reunir la cantidad suficiente de una o más especies para preparar una comida. Si se limita a recolectar unas cuantas especies, reducirá las posibilidades de incluir alguna especie no comestible o venenosa *(véase página 84)*. Al recolectar plantas, asegúrese de dejar una cantidad suficiente de cada especie para permitir su regeneración para, de este modo, poder abastecerse en el futuro. Las plantas cortadas se estropean muy rápidamente, por lo que es mejor guardarlas en una caja o bolsa.

RAÍCES Y TUBÉRCULOS

Algunas plantas almacenan almidón en sus raíces para utilizarlo durante el invierno. En primavera se convierte en azúcar para alimentar los nuevos brotes. De esta manera, las raíces son más nutritivas en otoño, y más dulces en primavera. Tenga cuidado al recoger las raíces en primavera, ya que algunas plantas tienen entonces algunos brotes y pueden ser difíciles de identificar.

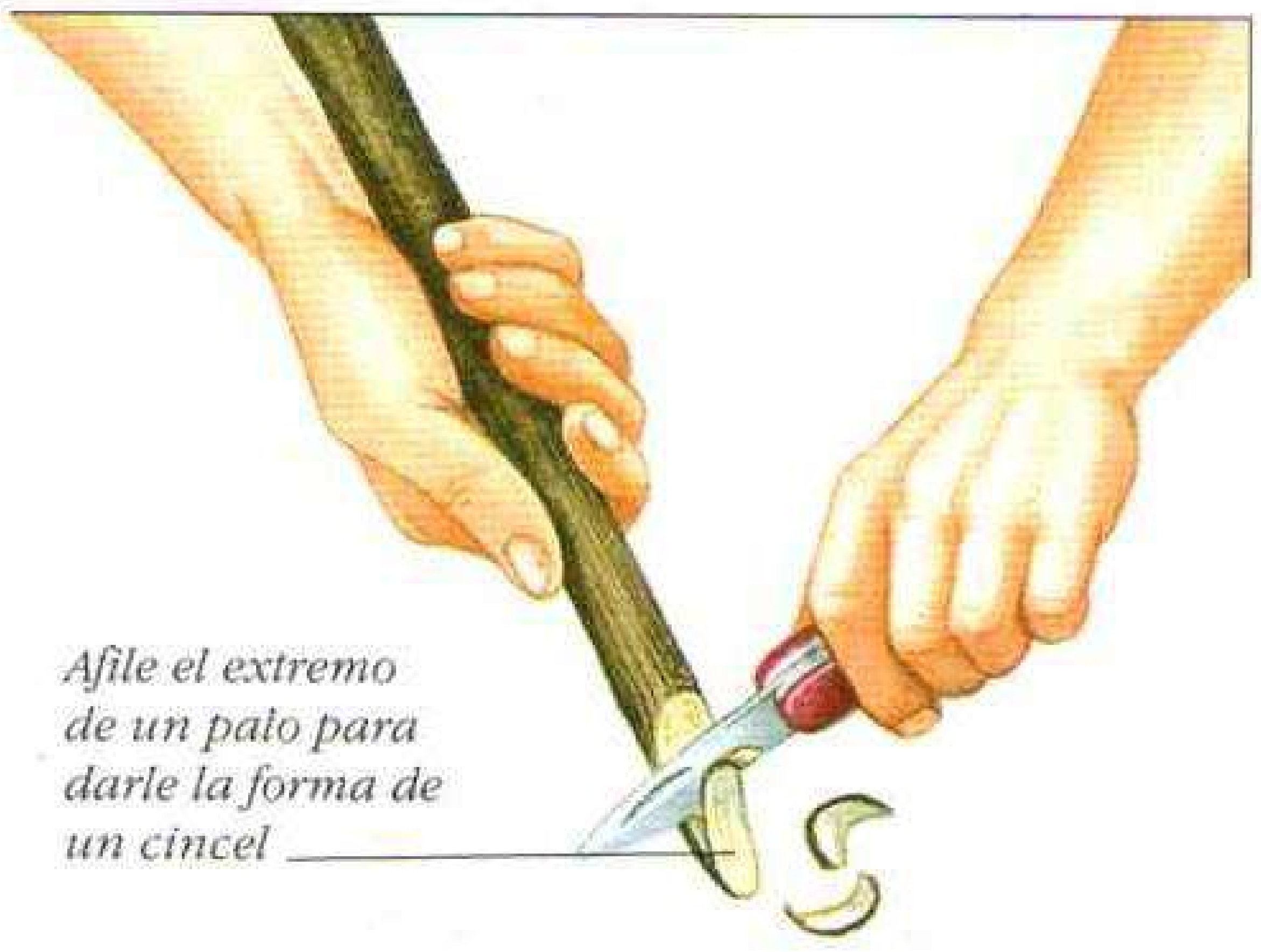
Afile el extremo de un palo para darle la forma de un cincel

1 Corte una rama gruesa de madera dura como el nogal, y afile uno de los extremos hasta darle la forma de un cincel. La punta puede endurecerse al fuego.

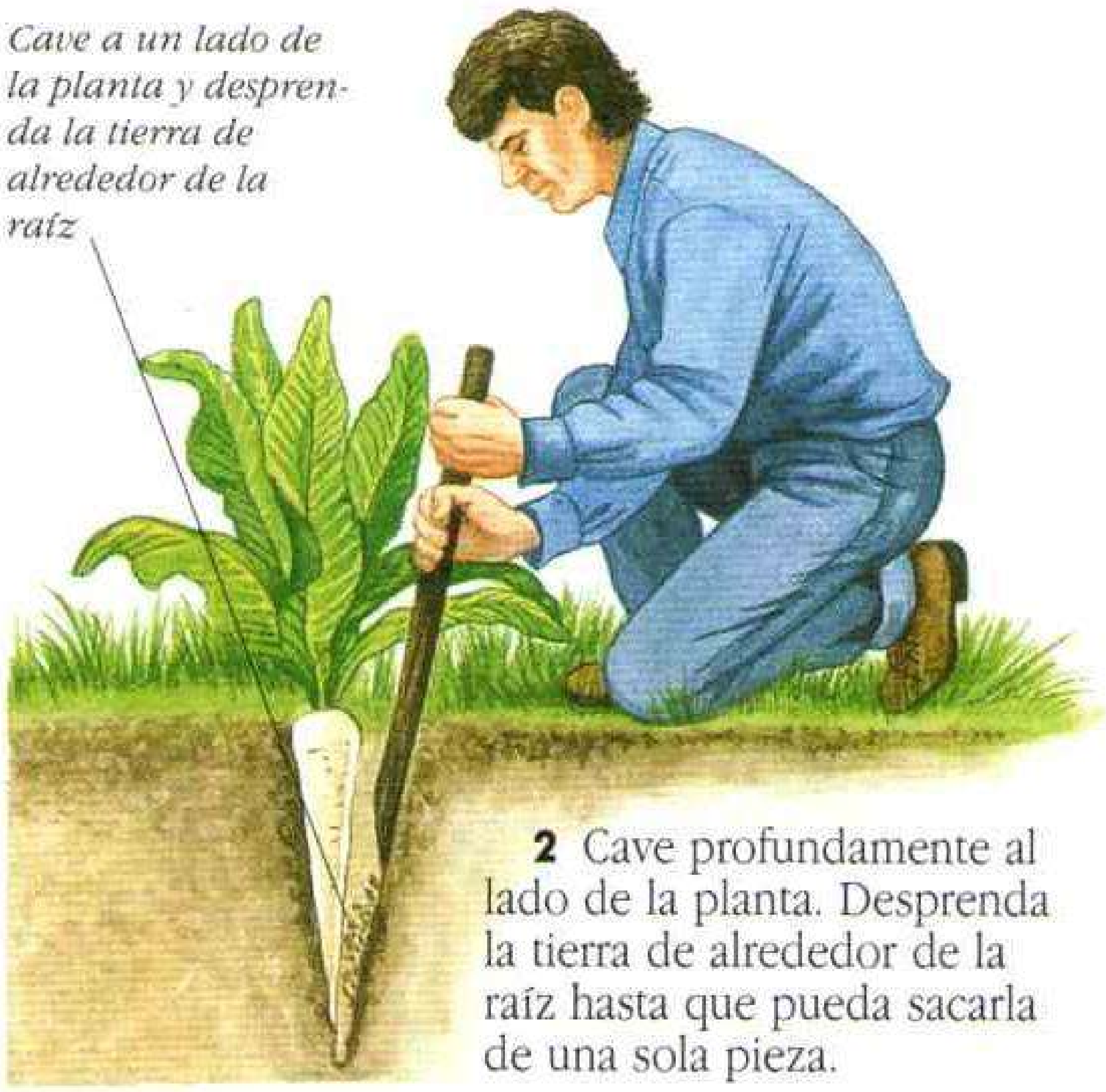
Cave a un lado de la planta y desprenda la tierra de alrededor de la raíz

2 Cave profundamente al lado de la planta. Desprenda la tierra de alrededor de la raíz hasta que pueda sacarla de una sola pieza.

CÓMO PREPARAR LAS RAÍCES

Debe preparar y cocinar las raíces tan pronto las haya recogido para evitar que se estropeen. Las raíces deben limpiarse cuidadosamente, aunque no debe pelarlas ya que la piel contiene a menudo muchas vitaminas. Corte las raíces en trozos grandes y hiérvalos en agua. Así eliminará las sustancias químicas amargas que pudieran contener. Los trozos deben hervirse hasta que estén lo suficientemente blandos como para poder pincharlos con un tenedor.

CÓMO OBTENER HARINA

La mayoría de las semillas no pueden digerirse bien a menos que se muelan para obtener harina. Puede entonces añadirla a un estofado o mezclarla con agua para hacer una masa para preparar una pasta o pan húmedo (*damper bread* es su nombre original) *(véase página 117)*. También puede obtener harina de las nueces *(véase página 88)*, así como de raíces o cabezuelas florales.

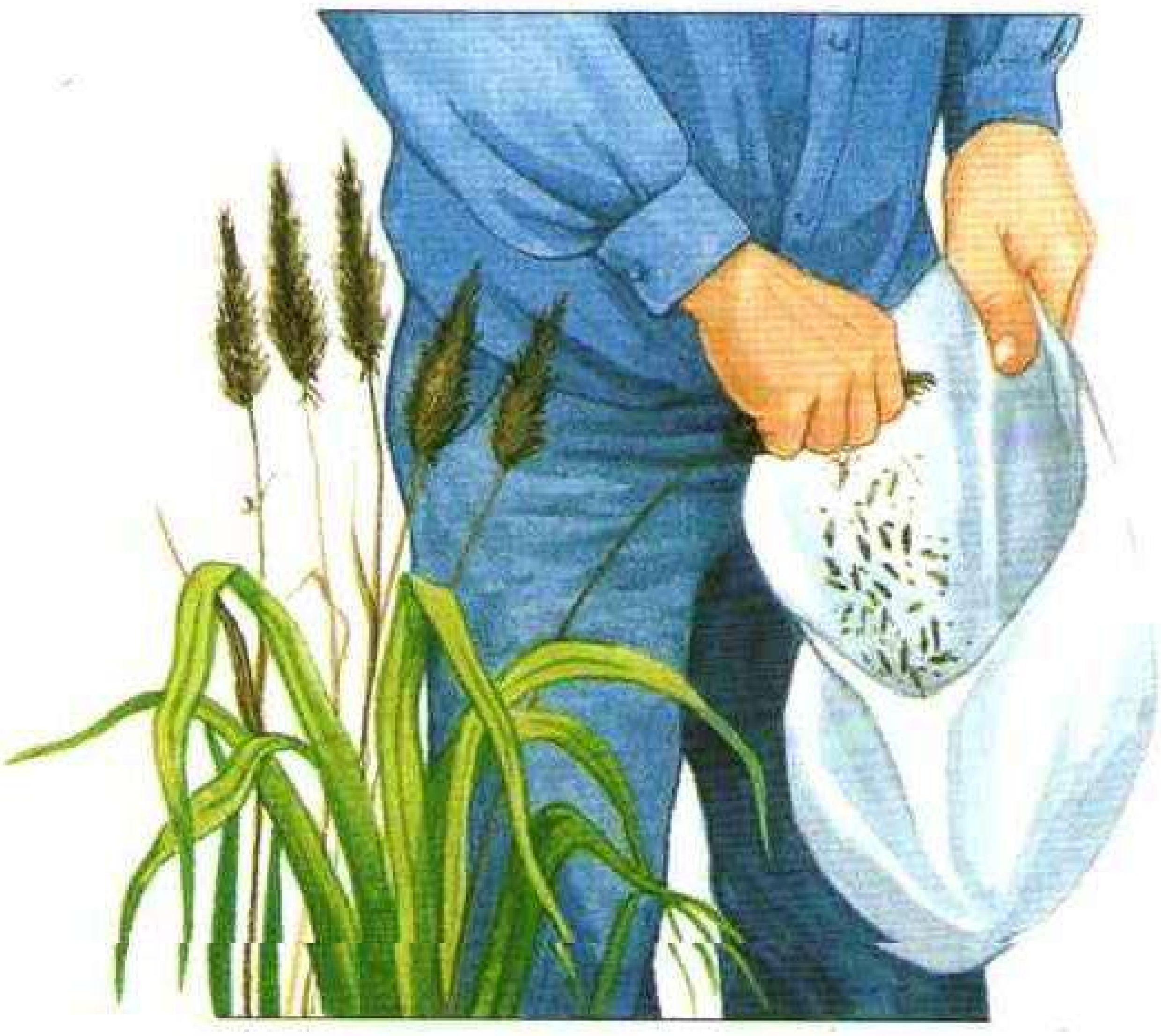

1 Para obtener harina a partir de semillas, comience por doblar la espiga de la planta que las contiene hacia una bolsa. Sacuda la espiga para desprender las semillas, o arránquelas con la mano. Deje secar las semillas al sol.

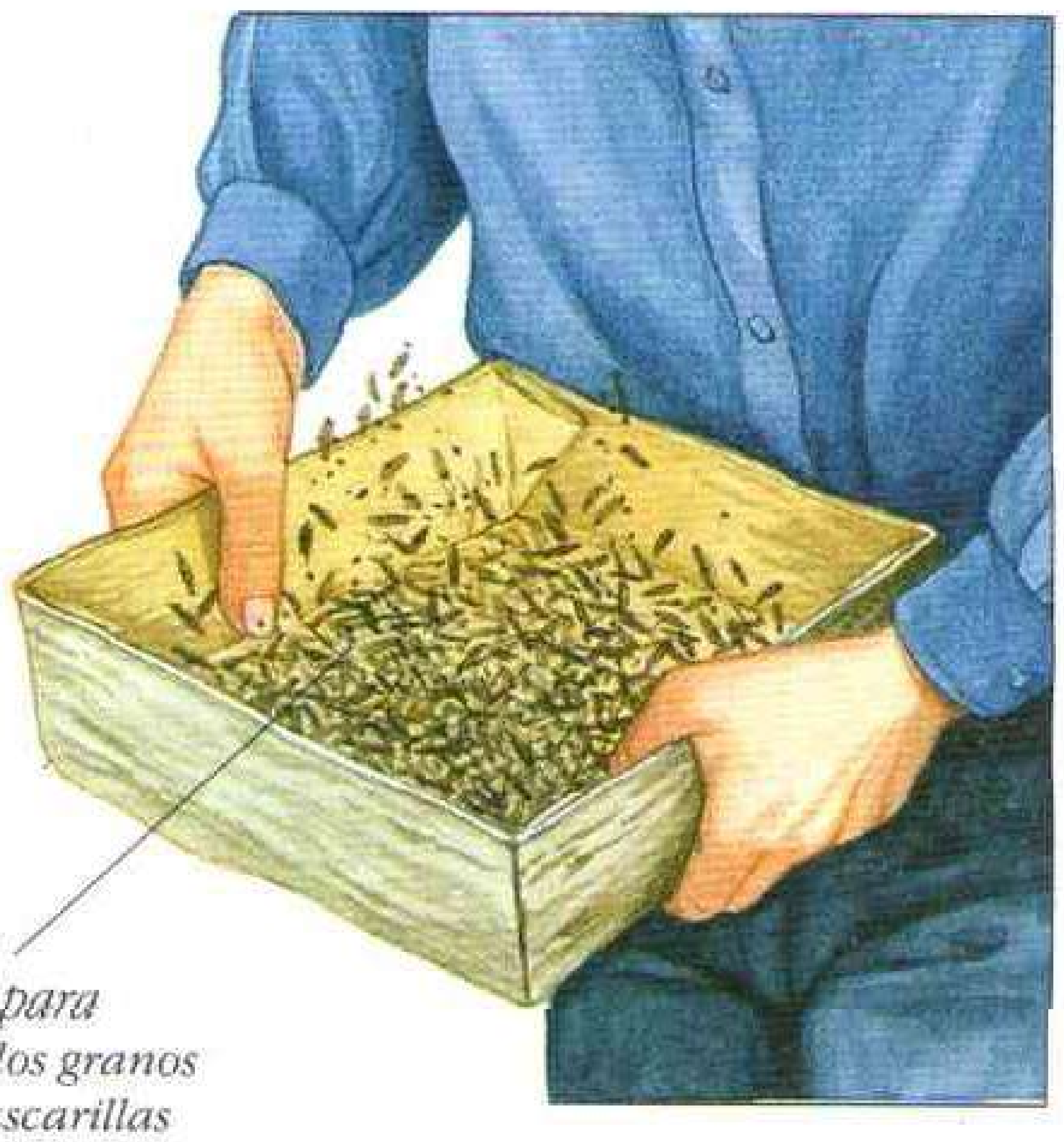
Agite las semillas para separar los granos de las cascarillas

2 Trille las semillas frotándolas entre sí para desprender las cascarillas. A continuación, aviente las semillas agitándolas en un recipiente. Como las cascarillas son más ligeras, podrá separarlas de las semillas soplando.

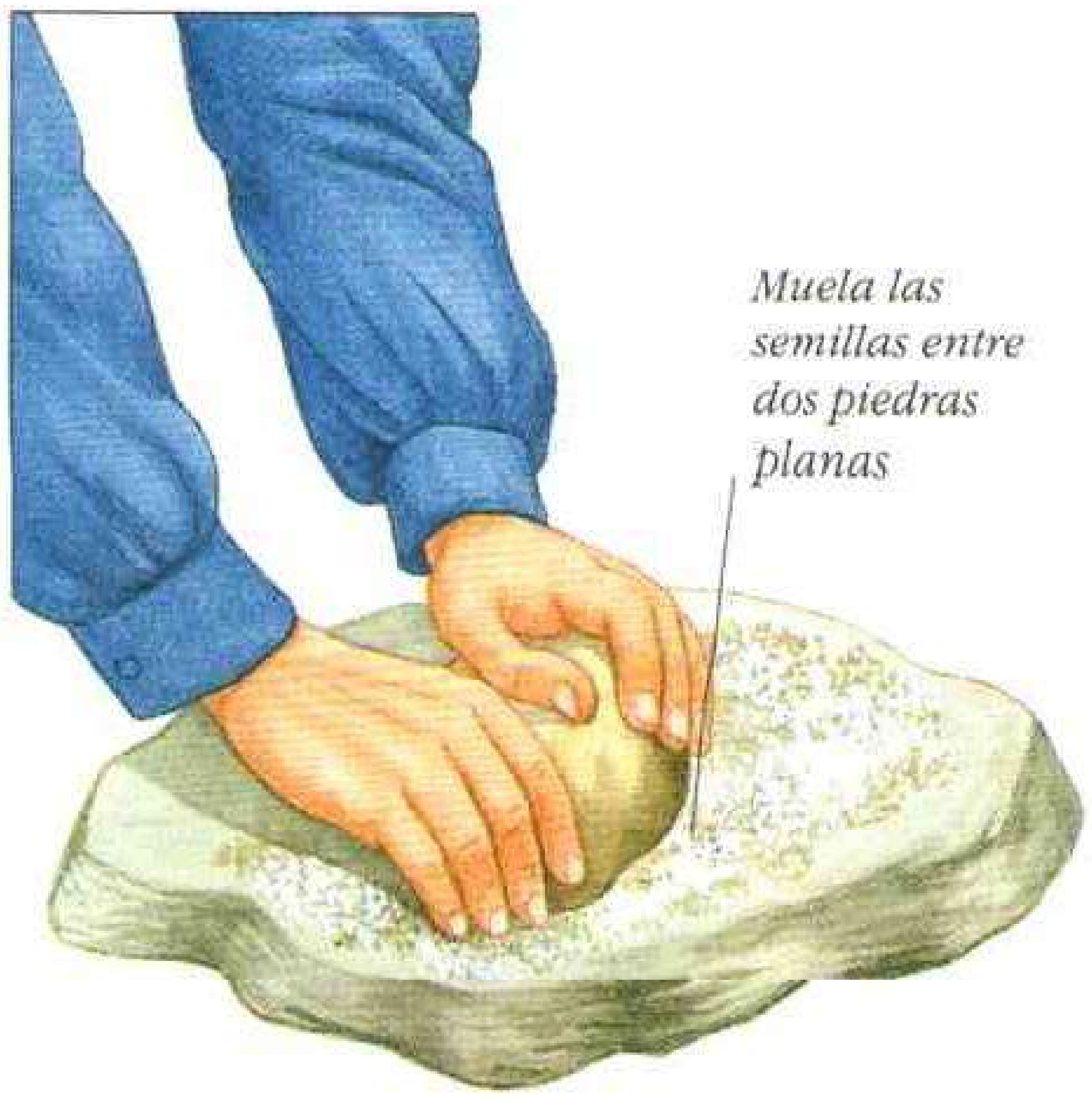
Muela las semillas entre dos piedras planas

3 Muela las semillas entre una piedra grande con una depresión en el centro y otra menor plana. Mantenga un movimiento continuo y circular. Tenga cuidado que el viento no se lleve la harina obtenida.

CÓMO PREPARAR LAS ALGAS

El musgo de Islandia *(Chondrus)* puede comerse fresco, seco o cocinarlo para obtener una especie de pudín de gelatina. Es rico en yodo, minerales y vitaminas A y B. Se encuentra en las rocas bajas de la playa, en primavera y a principios del verano. También puede hervir y comer las hojas de las ovas *(Porphyria)*, de la lechuga marina *(Ulva)* y de los varecs *(Alaria)*.

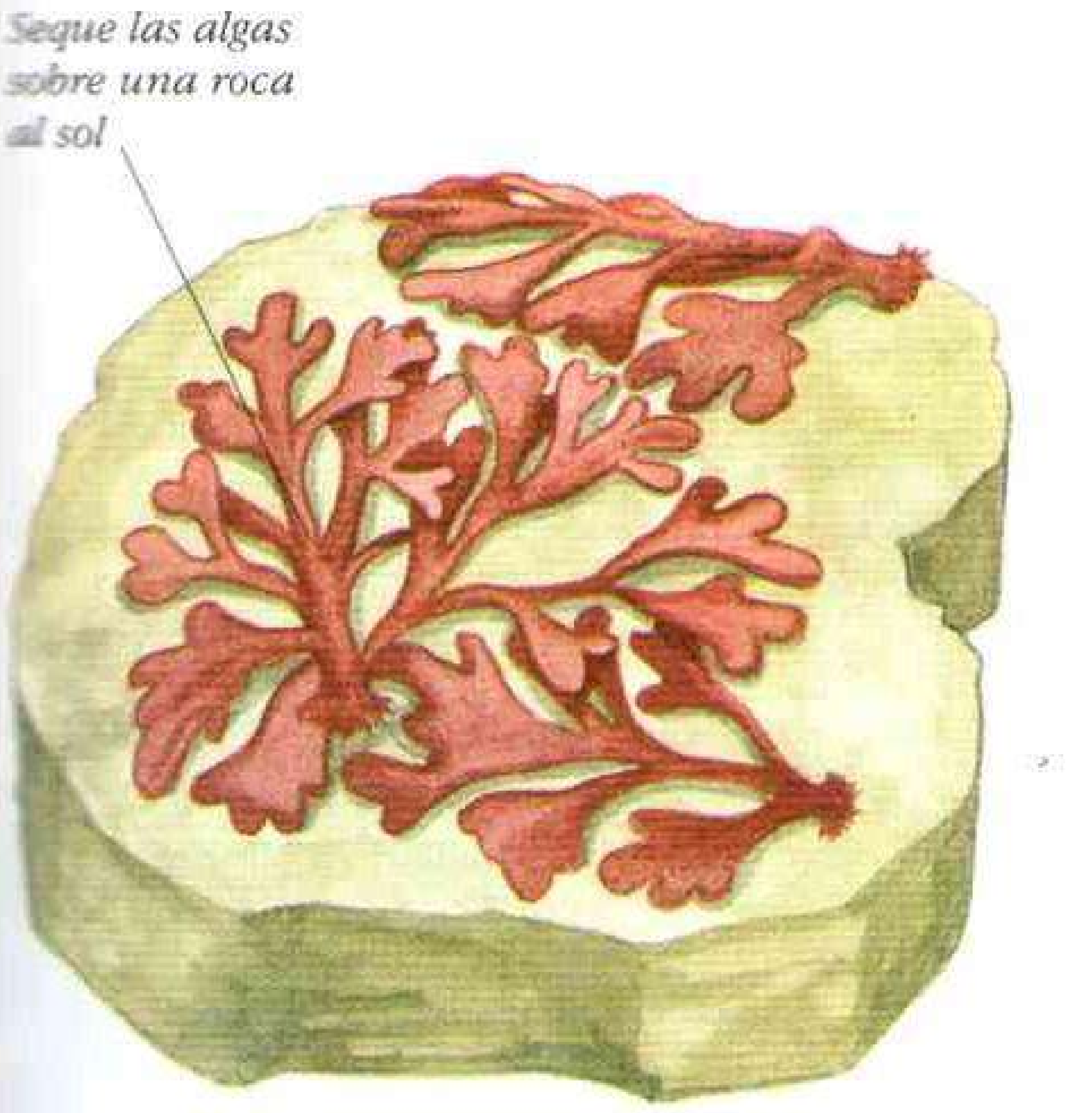
Seque las algas sobre una roca al sol

1 Recoja las algas desprendiéndolas de las rocas con su cuchillo. Extiéndalas al sol para que se sequen. Al secarse, las fibras palidecerán y adquirirán una textura correosa.

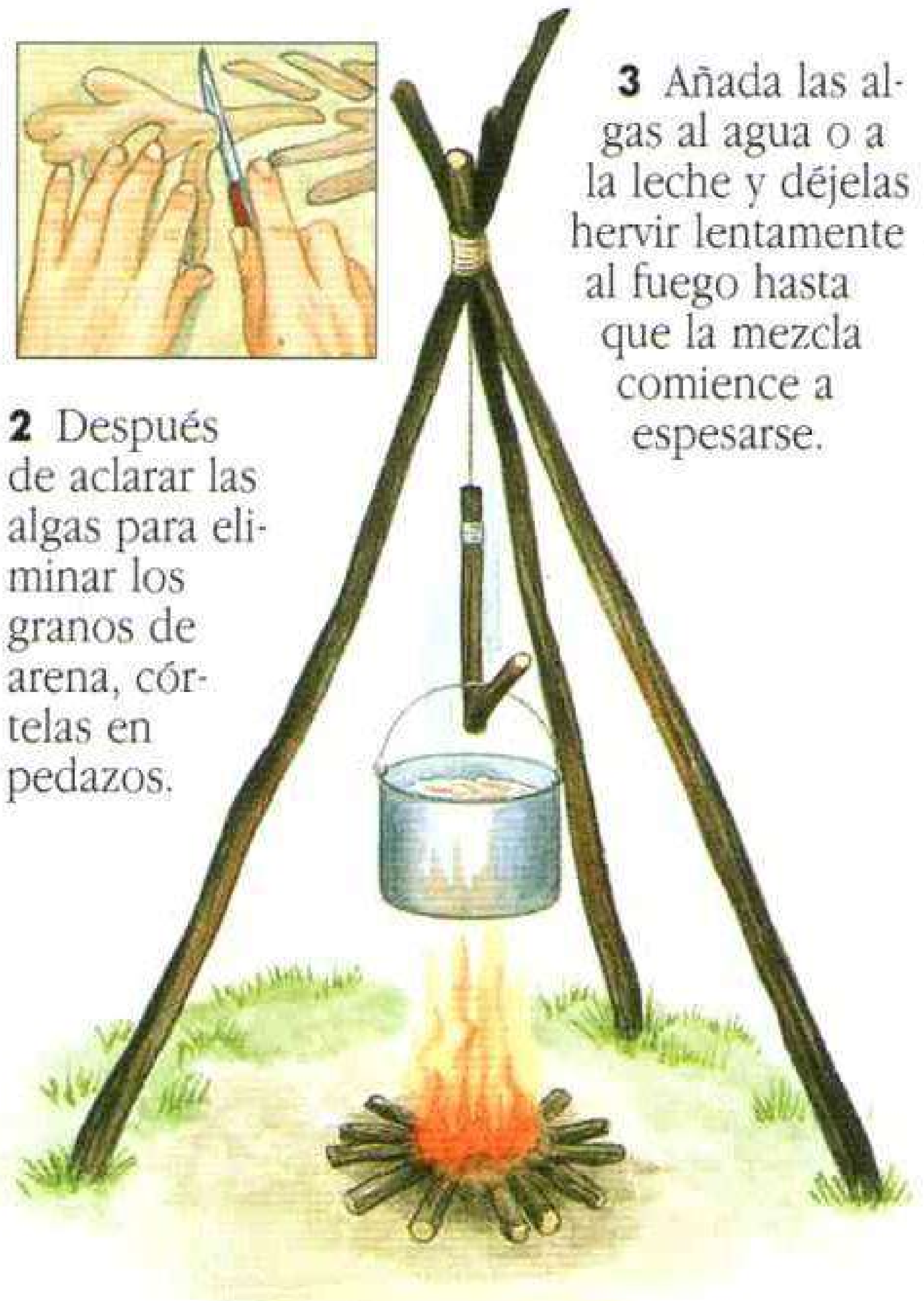

2 Después de aclarar las algas para eliminar los granos de arena, córtelas en pedazos.

3 Añada las algas al agua o a la leche y déjelas hervir lentamente al fuego hasta que la mezcla comience a espesarse.

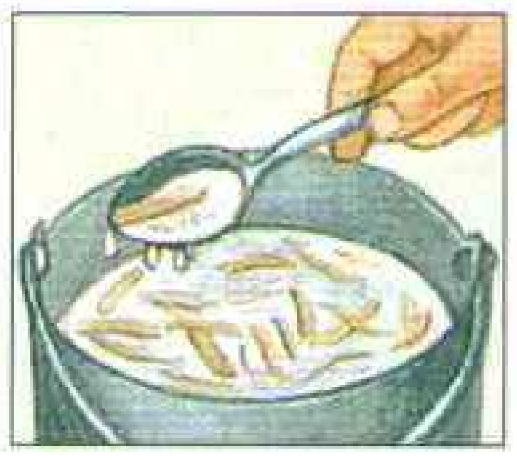

4 Saque las algas y endulce la mezcla con bayas o azúcar.

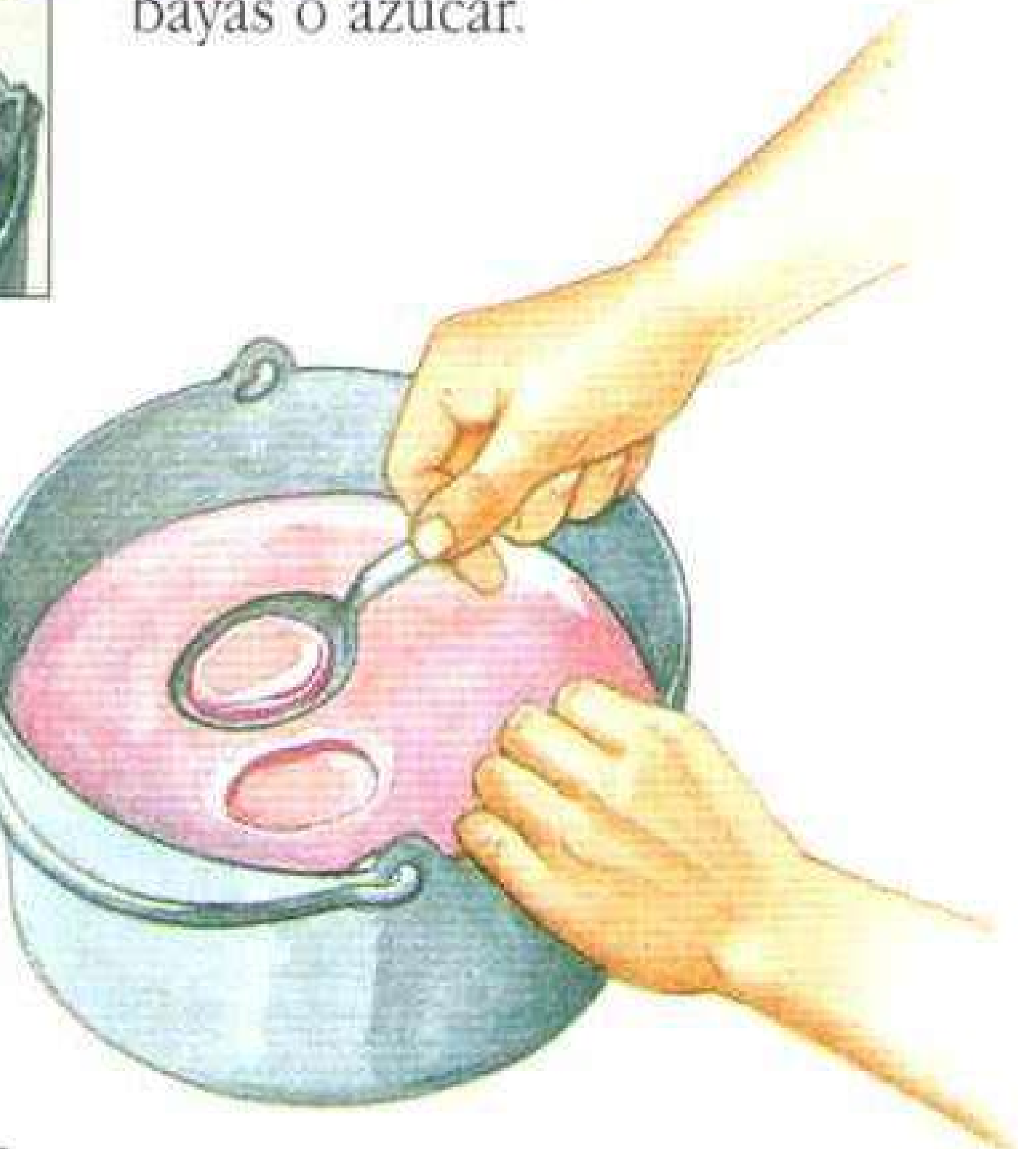

5 La mezcla cuaja conforme se va enfriando; sáquela del recipiente con una cuchara. Tenga cuidado de no comerla hasta que esté completamente fría, o se quemaría la boca.

CÓMO PREPARAR BEBIDAS

Las agujas de los árboles perennes como el pino *(Pinus)* y las píceas *(Picea)* son ricas en vitamina A y C, y pueden utilizarse para preparar té. Recoja únicamente agujas frescas y verdes no decoloradas. Puede preparar un sucedáneo del café a partir de hojas de achicoria *(Chicorium)* o de bellotas *(Quercus)*.

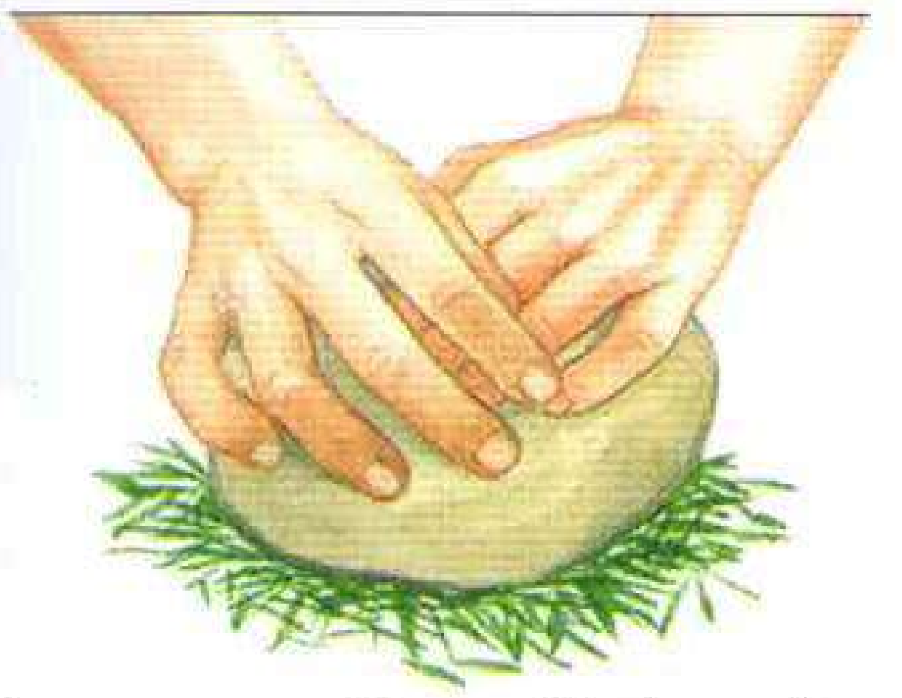

1 Para preparar té a partir de agujas perennes, utilice dos cucharadas de té de agujas por taza de agua. Májelas ligeramente con una piedra grande.

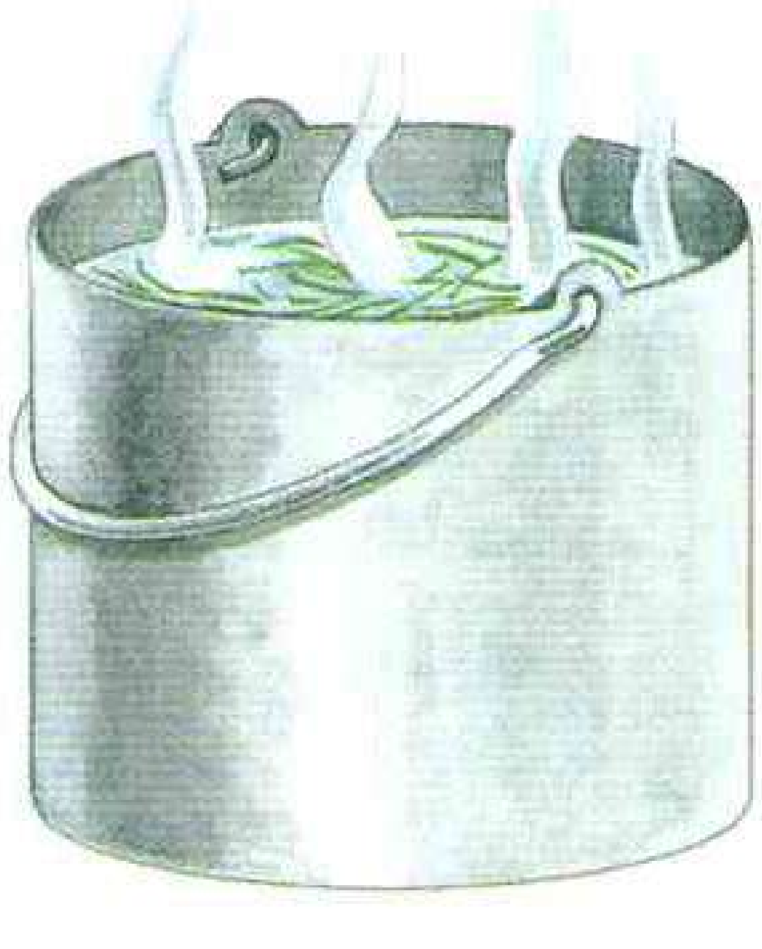

2 Sumerja las agujas en agua hirviendo. Mantenga el recipiente caliente sin permitir que el agua hierva, y deje remojar las agujas entre 5 y 10 minutos, removiendo de vez en cuando.

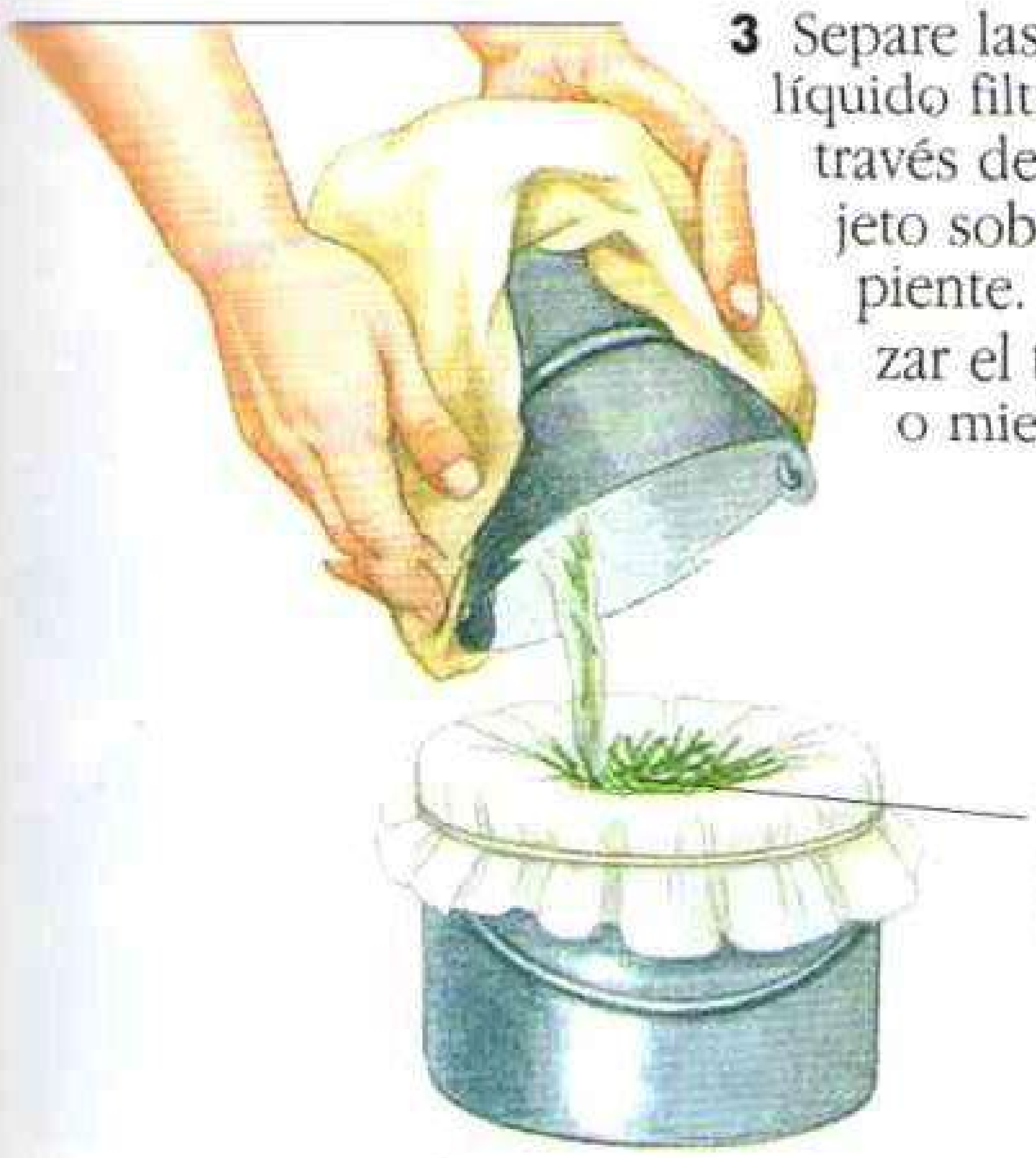
Filtre el té a través de un paño

3 Separe las agujas del líquido filtrando el té a través de un paño sujeto sobre un recipiente. Puede endulzar el té con azúcar o miel.

OTRAS BEBIDAS

En Australia, las hojas del eucalipto *(Eucalyptus)* dan un peculiar sabor mentolado refrescante al té habitual. Esta infusión se conoce como *billy tea*. Las flores secas del trébol *(Trifolium)*, y las hojas de la ortiga *(Urtica)* y de la hiedra rastrera *(Glechoma)*, también pueden utilizarse para preparar té. La savia de los arces *(Acer)* puede beberse al mezclarla con agua.

CONDIMENTOS NATURALES

Las plantas silvestres tienen a menudo un sabor muy suave; quizá quiera realzarlo un poco con especias u otros condimentos. Puede llevar especias preparadas como curry en polvo, canela o clavo en su mochila para añadirlos a cualquier plato cuando sea necesario. Como alternativa, puede recoger especias en su estado silvestre y secarlas para utilizarlas más adelante. Un ejemplo de especia que puede encontrar en estado silvestre es la nuez moscada, la semilla de un fruto de una planta nativa de Indonesia Oriental *(Myristica)*. Una vez seca, la semilla puede rallarse o trocearse para dar sabor a los alimentos horneados, a cualquier pudín *(véase superior)*, o a otros platos preparados a partir de vegetales como la espinaca o la calabaza. También las hierbas como la albahaca o el tomillo pueden conferir sabores novedosos.

CANELA

CURRY EN POLVO

CHILE EN POLVO

PIMIENTA

NUEZ MOSCADA

AJO

ANIMALES INVERTEBRADOS

LA IDEA DE comer una babosa, un caracol o un saltamontes no resulta demasiado apetecible, y nos hace pensar que preferiríamos pasar hambre. Sin embargo, en una situación de supervivencia es muy poco probable que una comida apetitosa sea fácil de obtener, y mientras espera se debilitará cada vez más hasta llegar al punto de no ser capaz de buscar fuentes alternativas de alimento. La experimentación es fundamental, pero para sobrevivir no debe rechazar ninguna fuente potencial de alimento, ni reparar en los escrúpulos que pueda sentir. Algunos invertebrados contienen incluso más proteínas que los vegetales, por lo que vale la pena no prescindir de ellos.

ANIMALES INVERTEBRADOS COMESTIBLES

Aunque muchos animales invertebrados son comestibles, no es tan sencillo meterse una babosa o un saltamontes en la boca, masticarlos y tragarlos. Únicamente algunas partes de estas criaturas son comestibles, y todas ellas deben cocinarse bien antes de consumirse. Algunas incluso deben prepararse de una manera especial *(véase página 96)*. A continuación se muestran algunos ejemplos.

Saltamontes. *Todos los insectos que saltan tienen grandes músculos en las patas que pueden comerse. Al asarlos destruirá los parásitos que puedan contener. Dan a la carne, según dicen los entendidos, un excelente sabor.*

Abejas. *Las abejas, sus crisálidas y larvas, son comestibles. Además, las abejas producen miel, una fuente instantánea de energía. Sin embargo, las picaduras de abeja pueden ser mortales, así que tenga cuidado al manipular estos insectos.*

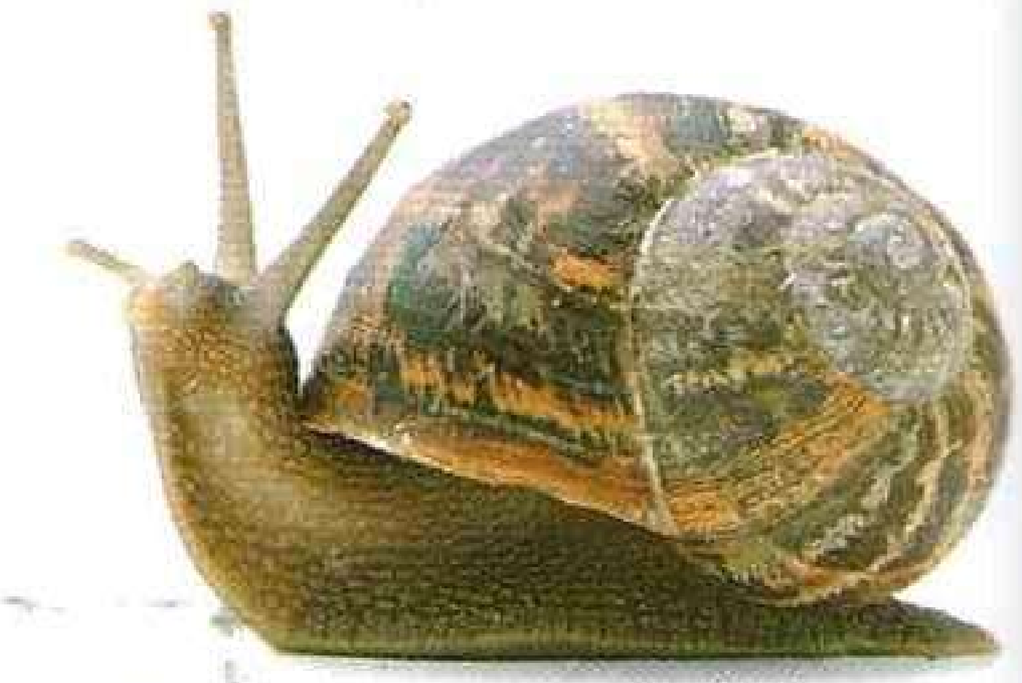

Caracoles. *Los gusanos, babosas y caracoles son una comida excelente, pero evite todos los caracoles marinos y cualquier caracol de tierra de colores brillantes, ya que podrían ser venenosos.*

Hormigas. *Las hormigas pueden picar o morder, pero son comestibles si se cogen con cuidado* (véase página 96). *Algunas hormigas colmeneras almacenan néctar y agua en sus abdómenes.*

Mariposas y polillas. *Las mariposas y polillas son comestibles, pero no aportan muchos nutrientes a menos que logre reunirlas en gran número. Las orugas son más fáciles de atrapar y constituyen una mejor comida.*

Estrellas de mar. *Pueden encontrarse adheridas a la madera a la deriva. Los cangrejos y las gambas también son comestibles y pueden encontrarse en las zonas rocosas cuando baja la marea.*

Erizos de mar. *Algunos equinodermos, como el erizo de mar* (Echinocyamus), *pueden desenterrarse de la arena de la playa para comerlos. Tenga cuidado al cogerlos, ya que algunos tienen largas púas.*

OTROS ANIMALES INVERTEBRADOS COMESTIBLES

Los insectos tienen un valor dietético superior al de los vegetales y son abundantes en cualquier región, aunque deben recolectarse en gran cantidad para obtener un aporte dietético adecuado. Las larvas de los insectos son particularmente jugosas. Para algunos pueblos indígenas constituyen un elemento primordial de su dieta. No olvide los insectos acuáticos, como los girinos *(Gyrinidae)*, aunque pueden resultar difíciles de atrapar a menos que se use una fina red de malla. Entre los invertebrados comestibles se incluyen también las termitas, la carcoma y el marisco. No es recomendable comer arañas, ya que muchas de ellas tienen vello o toxinas irritantes en la boca para matar a sus presas que, a su vez, pueden resultar dañinas para los humanos.

ANIMALES INVERTEBRADOS VENENOSOS

Debe aplicar el sentido común al recolectar animales invertebrados. Evite los insectos de colores brillantes, aquellos que comen carroña, los que parezcan enfermos o muertos, o los que tengan un olor fuerte, ya que podrían ser venenosos. Los animales invertebrados que comen otros animales pueden ser nocivos para los humanos. Tenga cuidado al recolectar animales invertebrados *(véase página 96).*

Escarabajos rojos. *El color rojo o brillante de algunos insectos, como los escarabajos y gorgojos, advierte a sus depredadores que tienen un sabor desagradable, por lo que no debe cogerlos. Su toxina puede no ser mortal, pero sí puede provocarle una grave indisposición.*

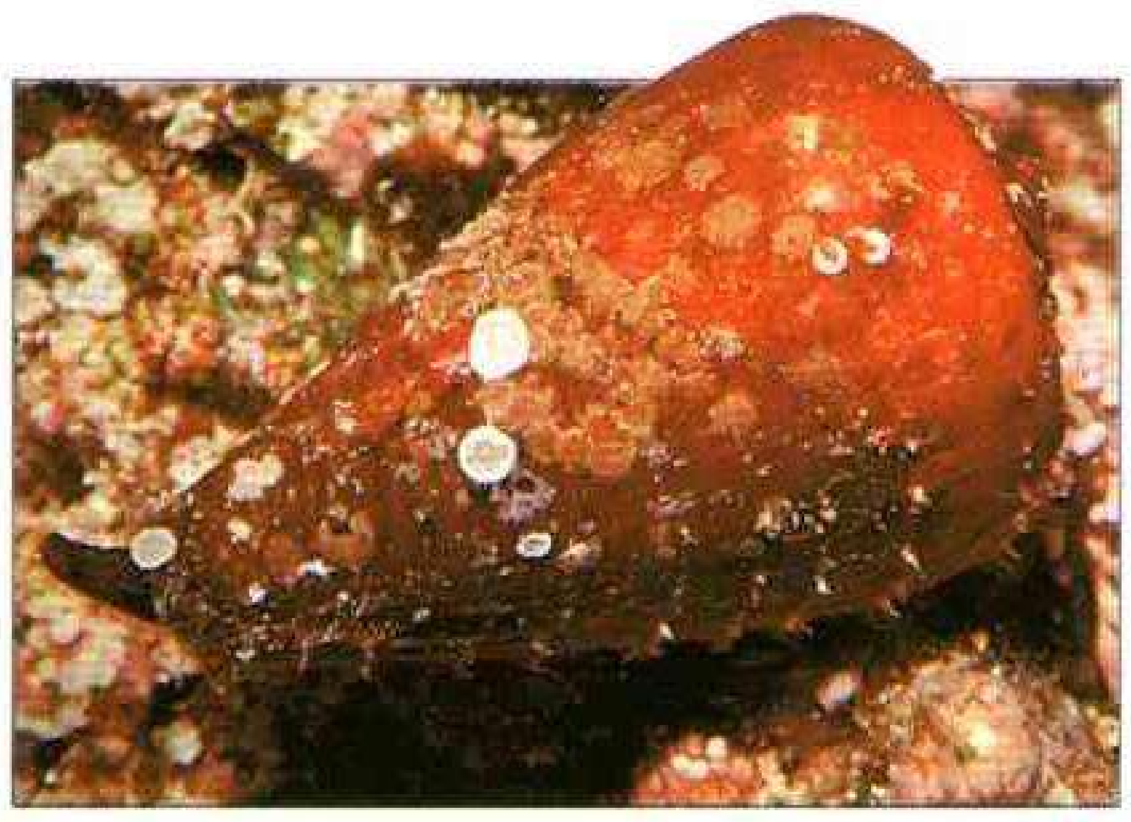

Bonidos. *Hay cientos de tipos de caracoles marinos que disparan un aguijón que inyecta veneno. En algunas variedades resulta mortal. Estos caracoles se entierran en la arena, por lo que puede pisar uno accidentalmente.*

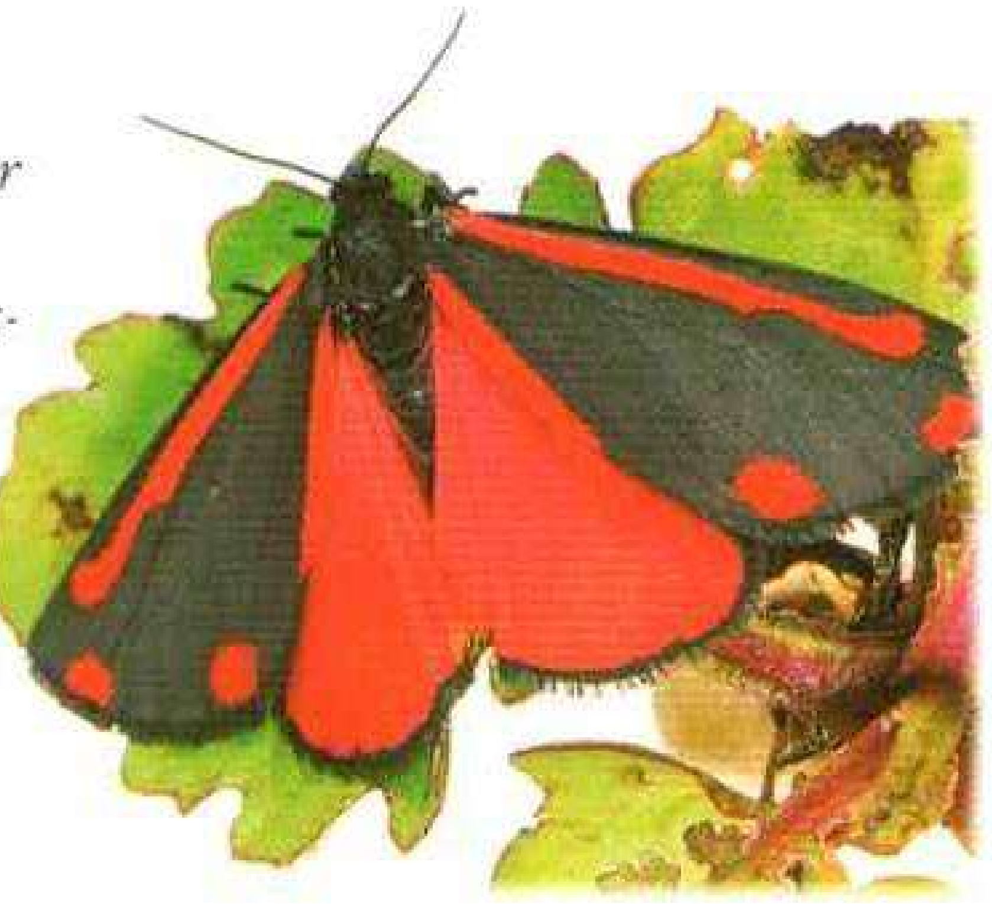

Polillas rojas. *La coloración roja de la polilla escarlata* (Callimorpha) *indica que tiene un sabor desagradable. Aunque algunas especies de insectos no tóxicos imitan los colores de las variedades venenosas, lo mejor es evitar aquellos que tengan colores brillantes, a menos que esté completamente seguro de que son inofensivos.*

ADVERTENCIA

Al igual que ocurre con la comida silvestre, es preferible aprender a identificar unos cuantos invertebrados comestibles. Si se encuentra en una situación de supervivencia, puede estar débil y ser más susceptible de lo habitual a las cantidades pequeñas de veneno.

OTROS ANIMALES INVERTEBRADOS VENENOSOS

Las orugas y los gusanos espinosos tienen sustancias irritantes en las púas que pueden ocasionar una erupción al tocarlos. Los ciempiés, sobre todo las especies grandes, tropicales, tienen a menudo sustancias venenosas en la piel. Algunas especies de mariquita amarilla *(Coccinellidae)* también son venenosas. Aunque la mayoría de las especies de pulpo son comestibles, el pulpo de anillo azul de Australia *(Hapalochlaena)* tiene una picadura mortal, por lo que es preferible evitar cualquier especie a menos que pueda identificarla positivamente. También debe evitar las crisálidas enterradas, así como algunas especies de hormiga *(Formica)* que lanzan ácido fórmico a sus atacantes. Si no logra identificar algún invertebrado, no lo toque.

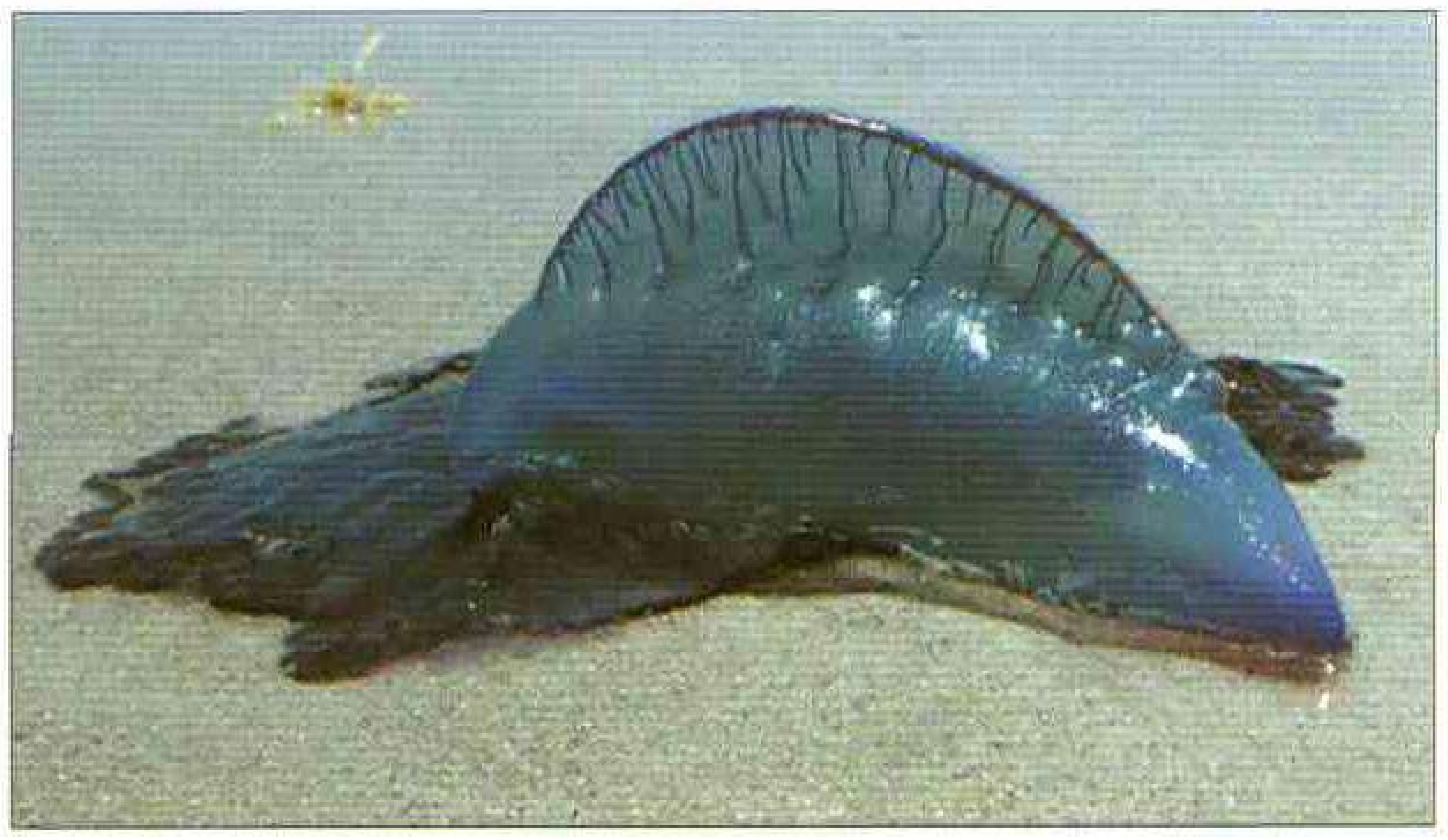

Hidrozoos marinos. *Algunos hidrozoos urticantes, como la hidra* (Hydra) *se encuentran en las corrientes cálidas. La fisalia* (Physalia) *puede tener tentáculos urticantes de hasta 12 m de longitud. Evite estos animales dentro de lo posible. También debe evitar las medusas, algunas de las cuales tienen una picadura venenosa* (véase página 180).

Avispas. *Son muy agresivas, y sus aguijones trabajan incluso después de muertas. Nunca intente atrapar avispas para comerlas. Si le pican en la cara, la hinchazón le impedirá respirar, lo que puede causarle la muerte.*

Crustáceos y moluscos. Las distintas especies de crustáceos y moluscos son en general comestibles una vez se han cocido, pero también son potencialmente peligrosos. Para alimentarse filtran grandes cantidades de agua y pueden retener abundantes sustancias tóxicas en sus cuerpos. Los moluscos deben cogerse vivos y comerse lo más pronto posible.

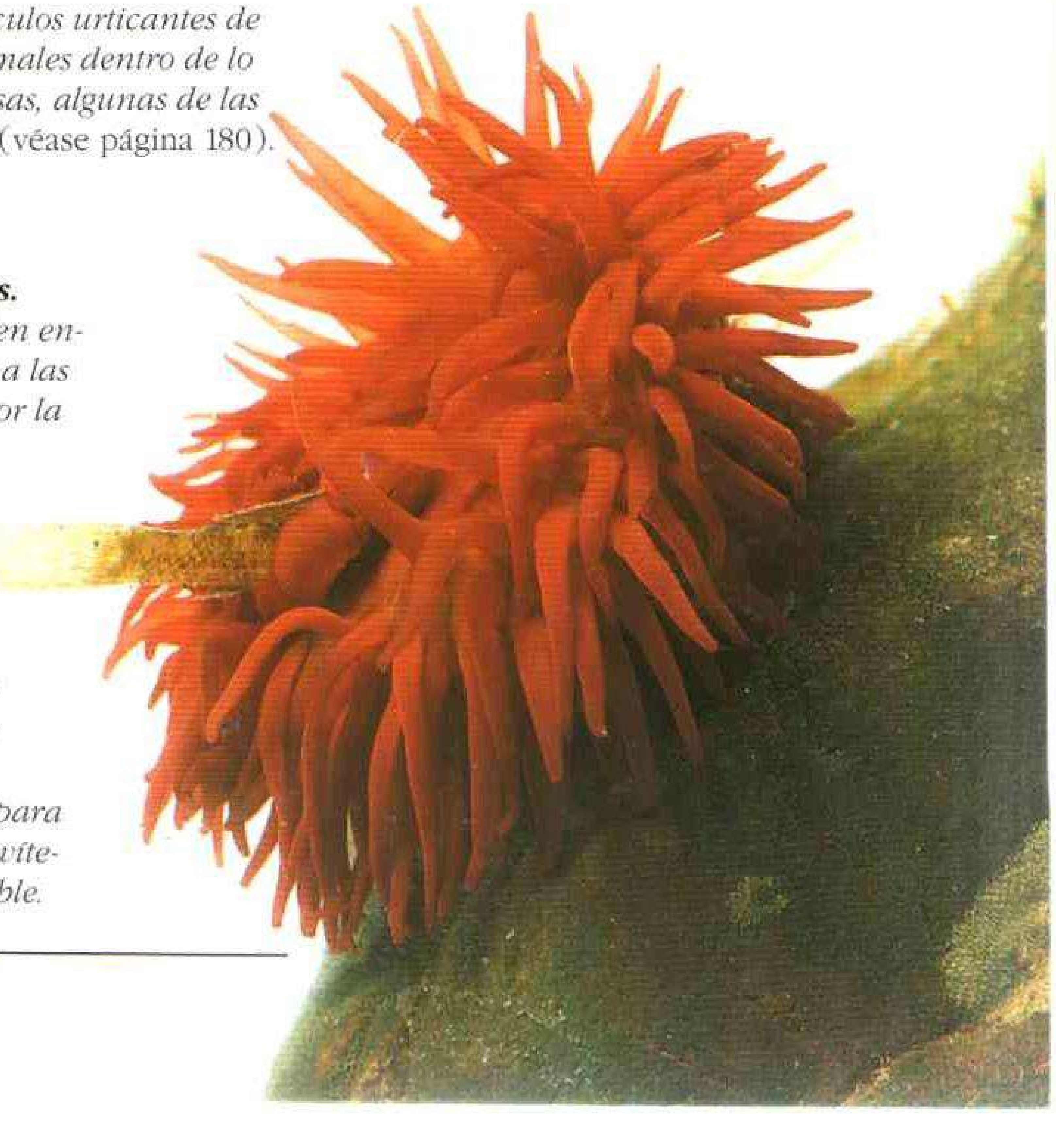

Anémonas marinas. *Las anémonas pueden encontrarse adheridas a las rocas descubiertas por la marea. La mayoría tienen tentáculos urticantes, por lo que debe evitarlas. La mayor parte de criaturas que matan a sus presas por medio de picaduras también pueden ser nocivas para los seres humanos: evítelas dentro de lo posible.*

LOS ANIMALES INVERTEBRADOS COMO ALIMENTO

LOS ANIMALES INVERTEBRADOS son, con diferencia, los más abundantes del reino animal, y a menudo son las únicas criaturas que habitan en los medios y condiciones más extremos. Es más fácil atrapar animales invertebrados como los insectos que pescar, y en esa actividad se gasta menos energía. Sin embargo, los animales invertebrados parecen menos apetitosos que otras criaturas. Para superar este inconveniente, puede picarlos o molerlos, y después añadirlos a otros platos más sabrosos, antes que comerlos enteros.

INSECTOS VOLADORES

La mayoría de los insectos voladores son comestibles, pero las abejas productoras de miel proporcionan el mejor alimento. Tanto las abejas adultas como sus larvas son comestibles. Sin embargo, ya que las abejas pueden ser peligrosas, atrápelas únicamente si no le queda otra opción.

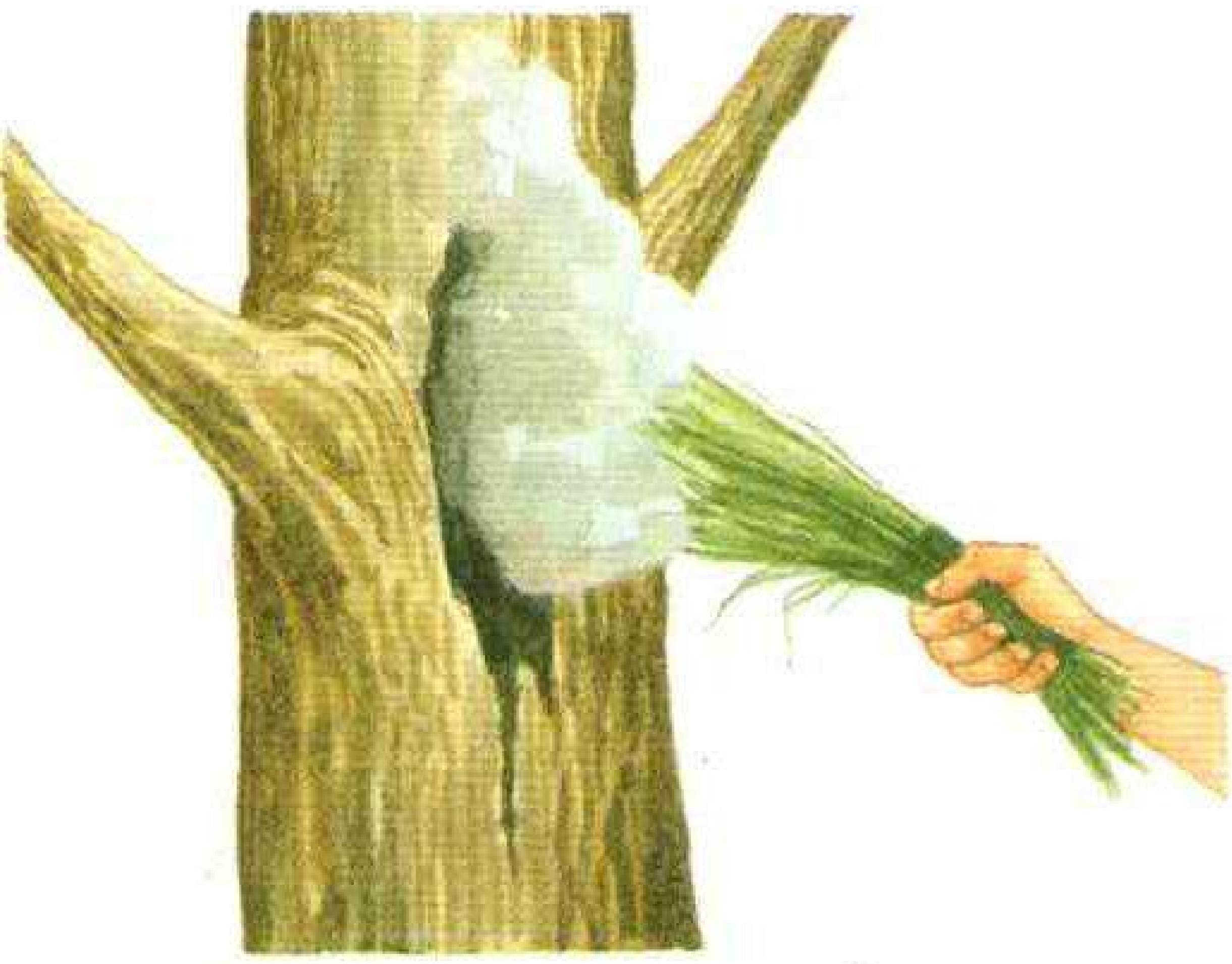

Abejas y miel. *Para atrapar abejas y recoger miel, acerque hierbas humeantes a la entrada de la colmena hasta que ésta se llene de humo. Selle el agujero y espere una noche. A la mañana siguiente, destápelo cuidadosamente con una rama y saque las abejas muertas y el panal.*

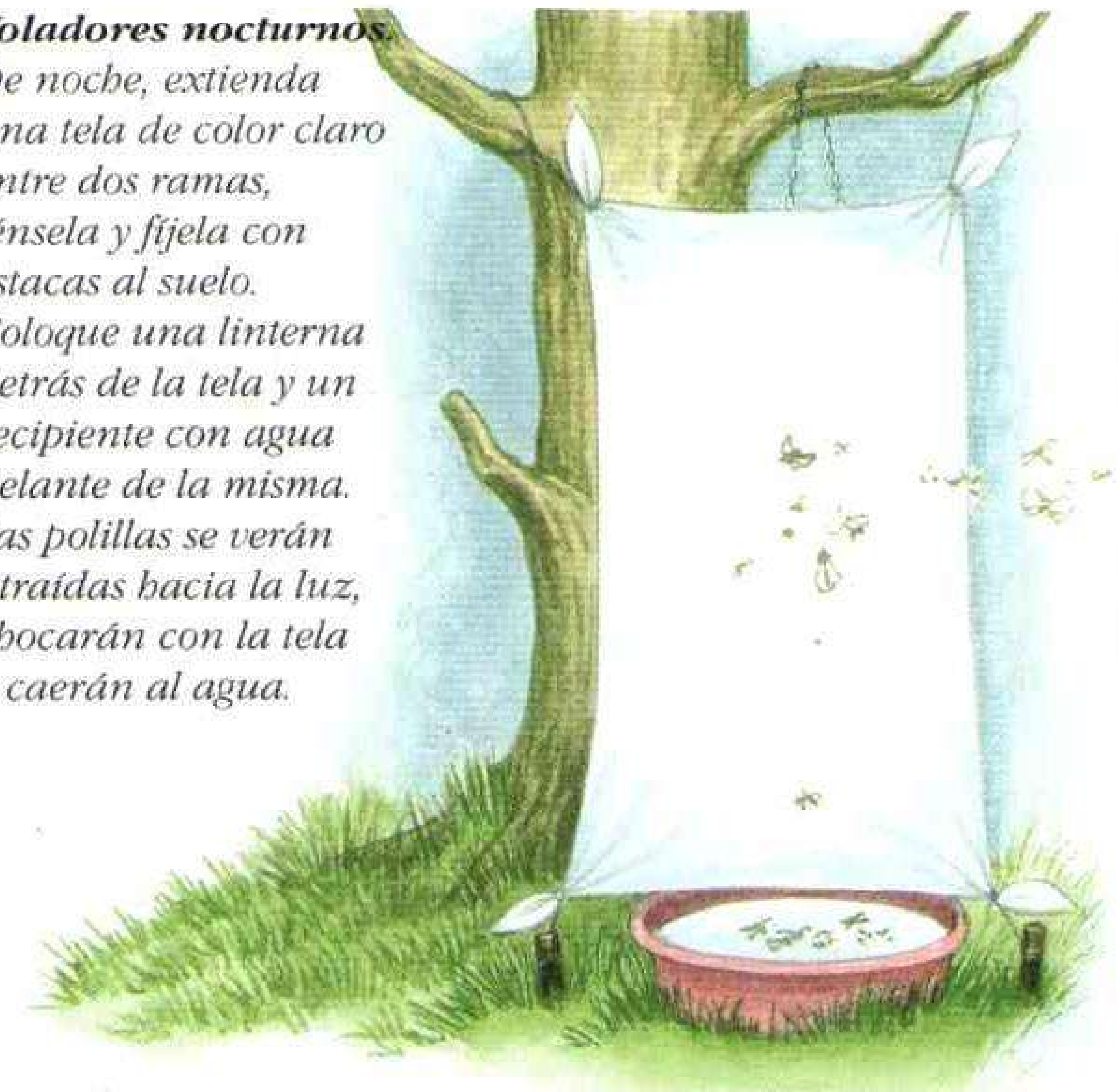

Voladores nocturnos. *De noche, extienda una tela de color claro entre dos ramas, ténsela y fíjela con estacas al suelo. Coloque una linterna detrás de la tela y un recipiente con agua delante de la misma. Las polillas se verán atraídas hacia la luz, chocarán con la tela y caerán al agua.*

CÓMO PREPARAR LOS INSECTOS

Elimine los aguijones de las abejas. Separe las alas y las patas de todos los insectos. Desde el punto de vista nutricional, los insectos deben comerse crudos, pero resultan más apetecibles hervidos o asados.

Molienda. *Los insectos pueden molerse entre dos piedras y añadirse a un estofado.*

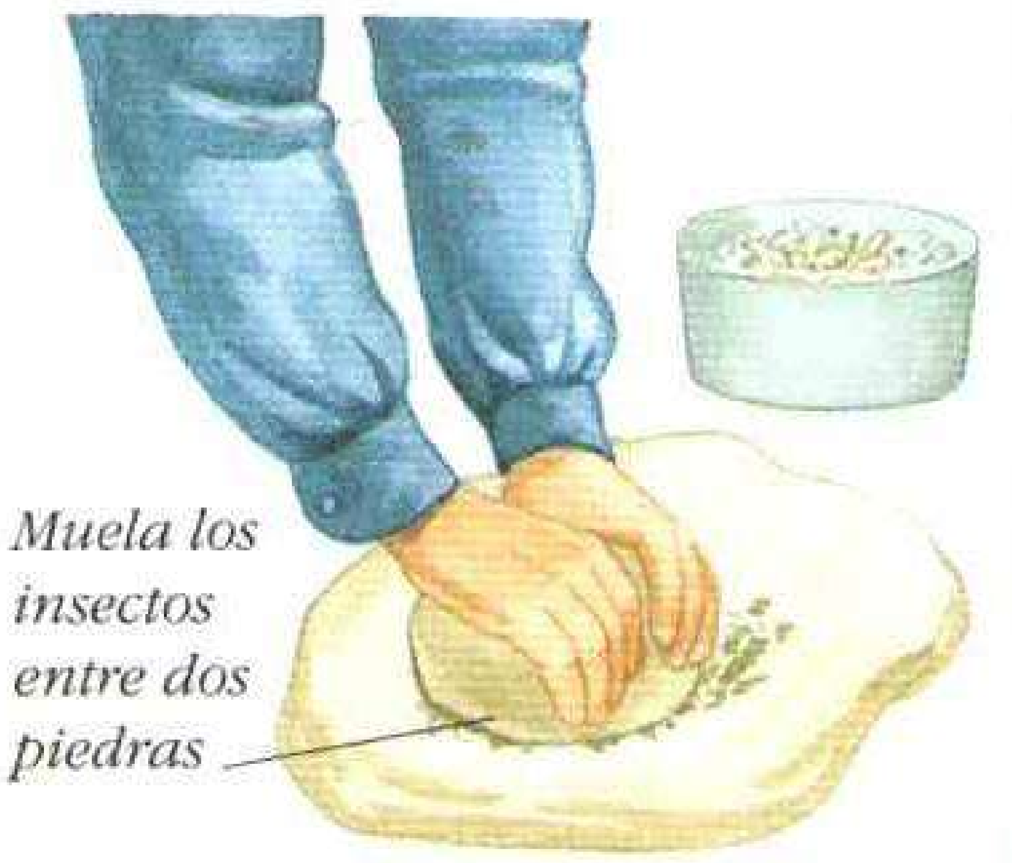
Muela los insectos entre dos piedras

INSECTOS REPTANTES

Las larvas de los escarabajos son muy nutritivas. Examine la madera podrida con ayuda de una rama para comprobar si hay alguna larva bajo la superficie. Las orugas se encuentran a menudo sobre las plantas que les sirven de alimento, y frecuentemente las encontrará en grandes colonias, pero recuerde que muchas tienen vellos urticantes en su piel, y pueden producir irritaciones al tocarlas.

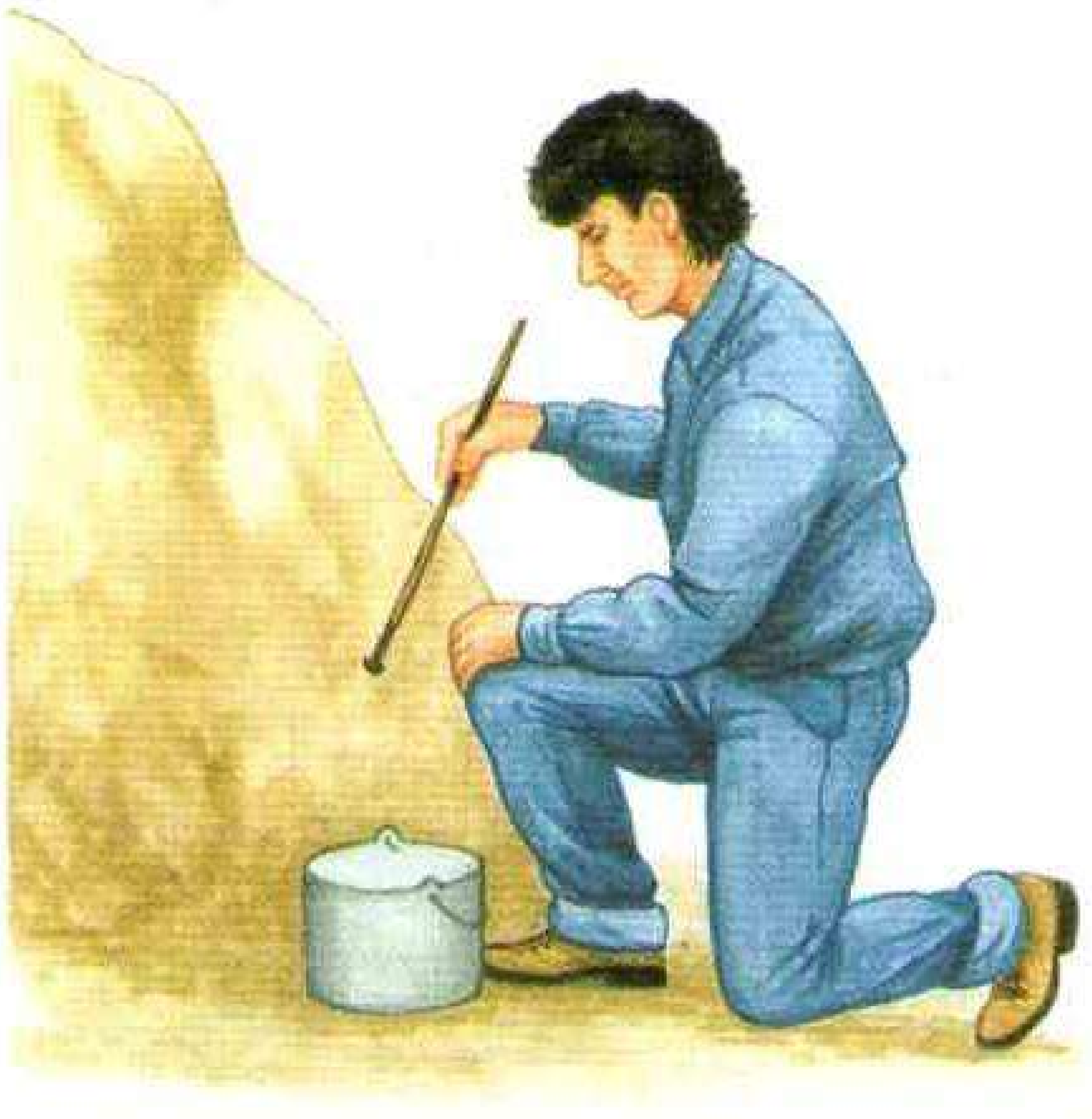

1 Para atrapar termitas u hormigas, corte una rama larga y recta y elimine su corteza hasta dejarla lisa. Después introdúzcala lentamente en el nido de hormigas o termitas. Tenga a mano un recipiente para recoger los insectos sin que se le caigan al suelo.

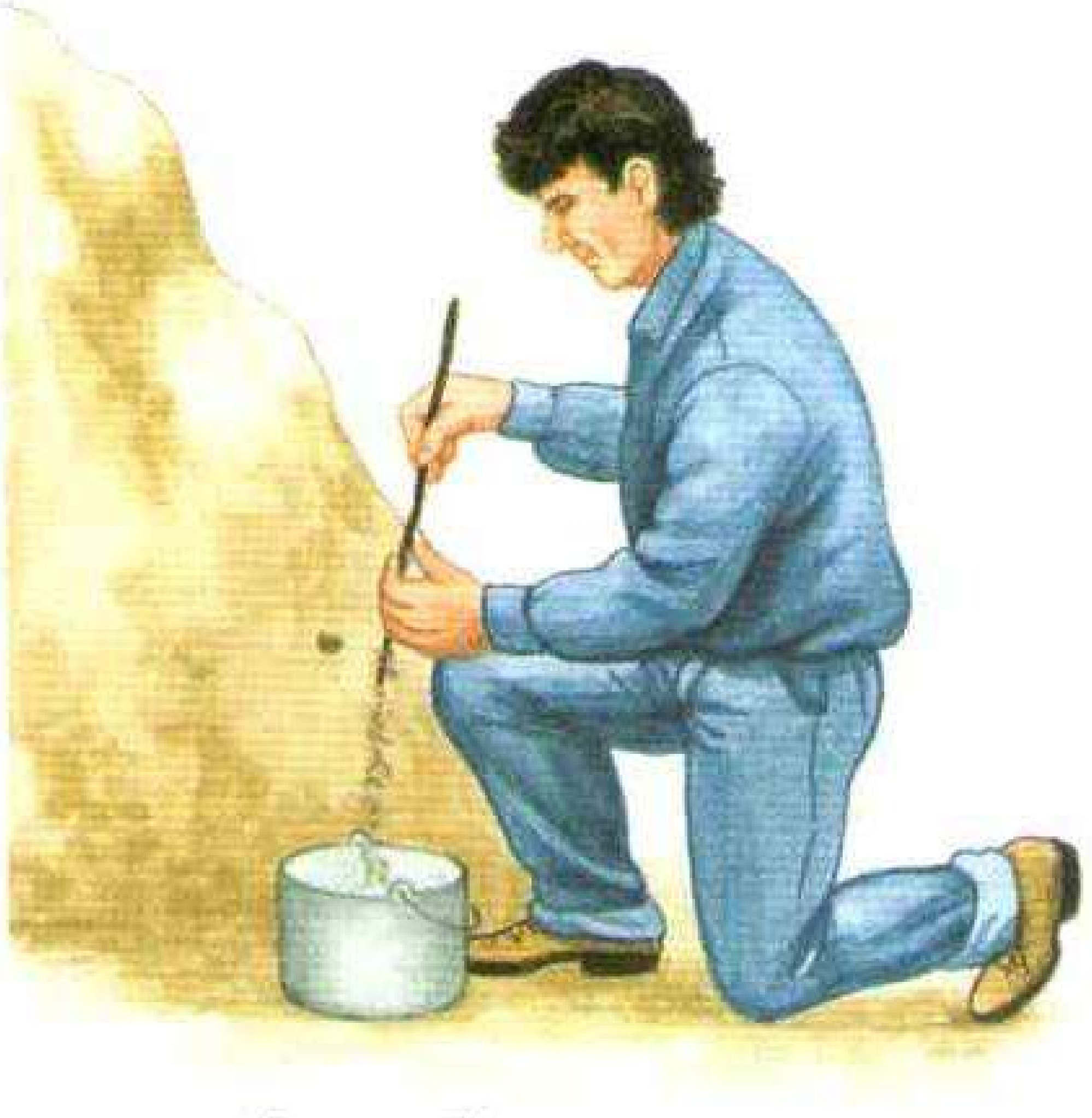

2 Extraiga la rama lenta y suavemente. Las termitas habrán atacado la madera y aún estarán sujetas a la misma. Raspe la rama y hágalas caer al recipiente. Puede entonces prepararlas como se ha indicado anteriormente.

CÓMO ATRAPAR INSECTOS

Puede preparar una trampa para insectos. Cave un agujero con cabida suficiente para un recipiente, y coloque éste en su interior. Coloque una tabla de madera encima del agujero, sobre dos piedras, para dejar un pequeño espacio entre la madera y el suelo alrededor del agujero. Los insectos serán atraídos por la sombra que les proporciona la madera, y caerán en la trampa.

TRAMPA PARA INSECTOS

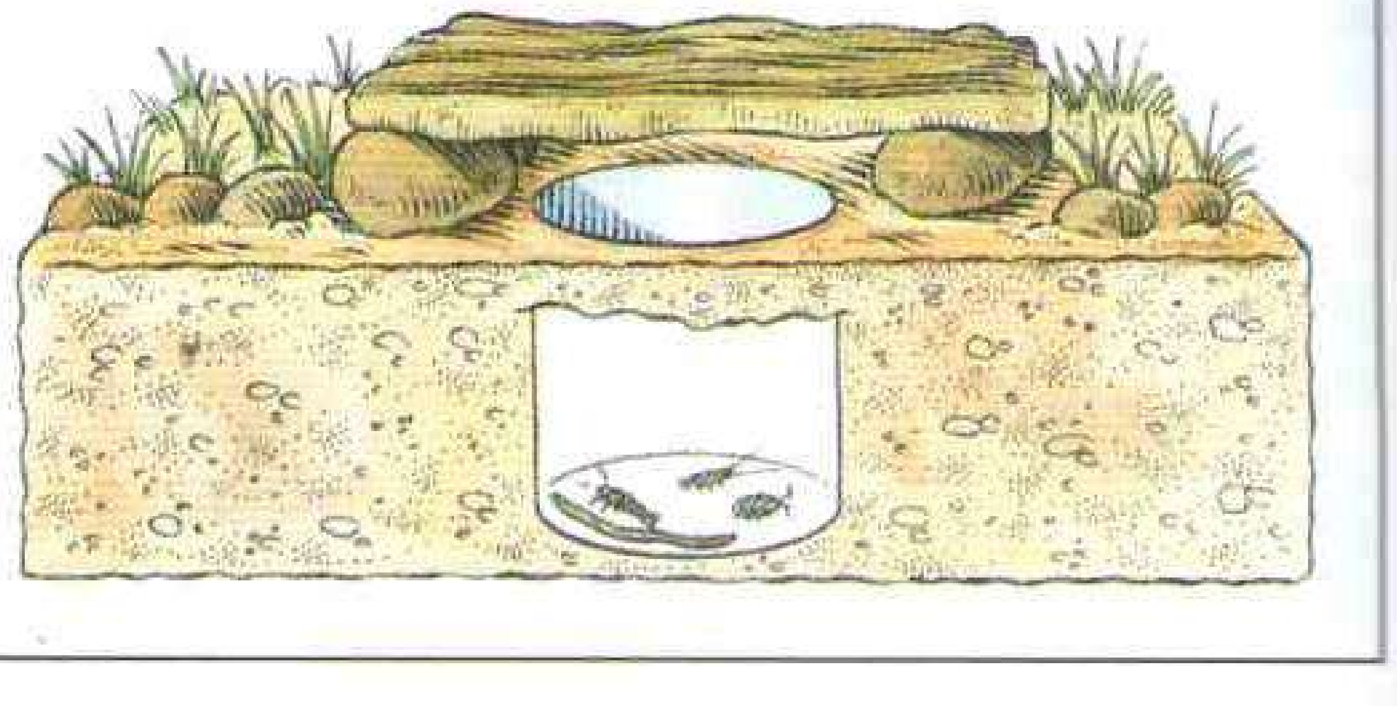

CRUSTÁCEOS Y MOLUSCOS

Los crustáceos y moluscos son fáciles de pescar en la mayoría de las playas, y pueden ser muy nutritivos. Sin embargo, debe asegurarse de que aún están vivos cuando los pesque, ya que su carne se deteriora rápidamente después de morir. Cocínelos siempre y consúmalos inmediatamente después de cogerlos. No los pesque cerca de cualquier posible fuente de contaminación.

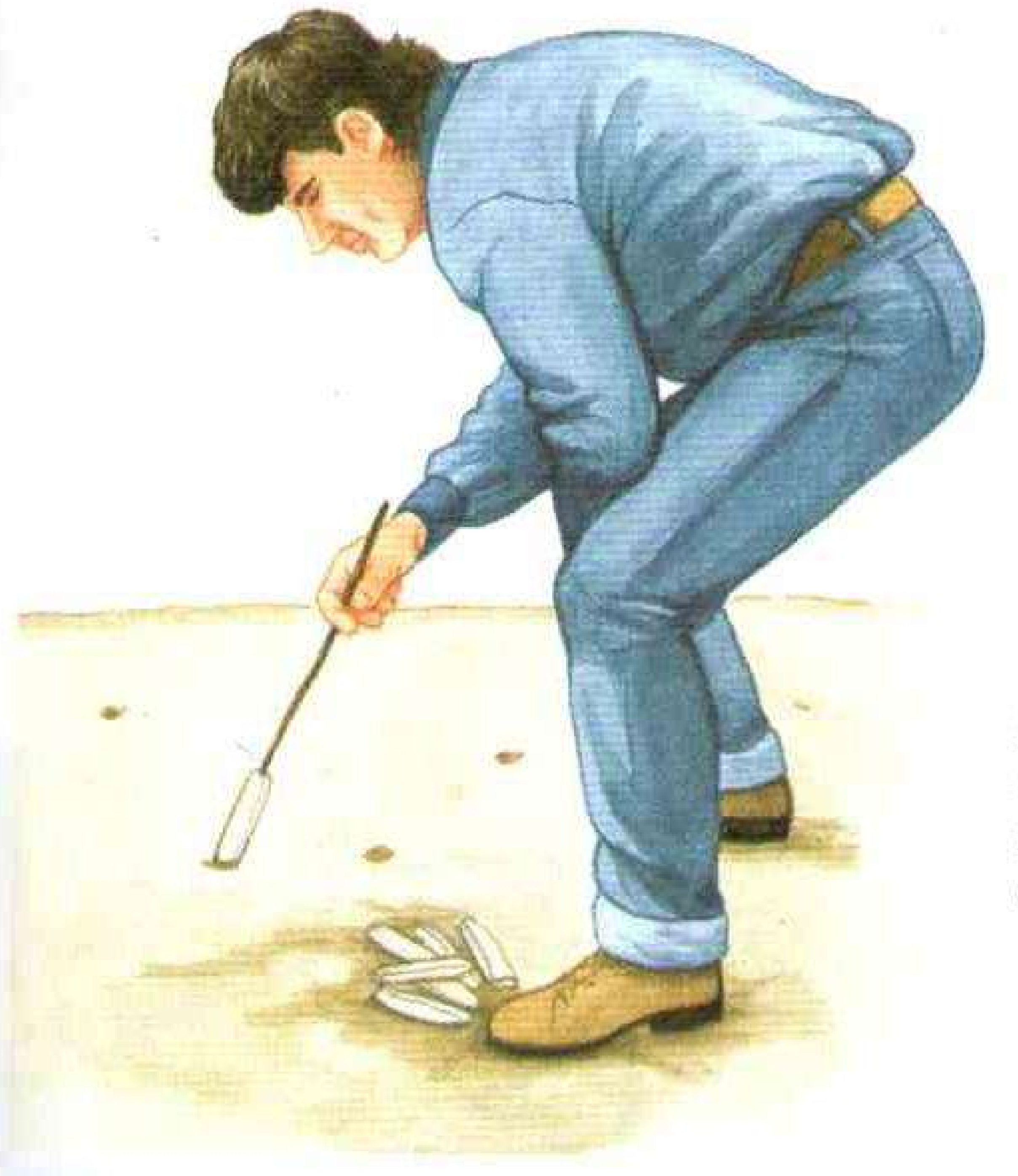

En busca de navajas. *Las navajas se entierran verticalmente en la arena dejando agujeros para respirar en la parte superior de sus túneles. Puede atraparlas haciendo pasar un trozo de alambre o un palo a través de estos agujeros. Asimismo, puede espolvorear sal alrededor de la entrada del túnel y atraparlas cuando salen a la superficie.*

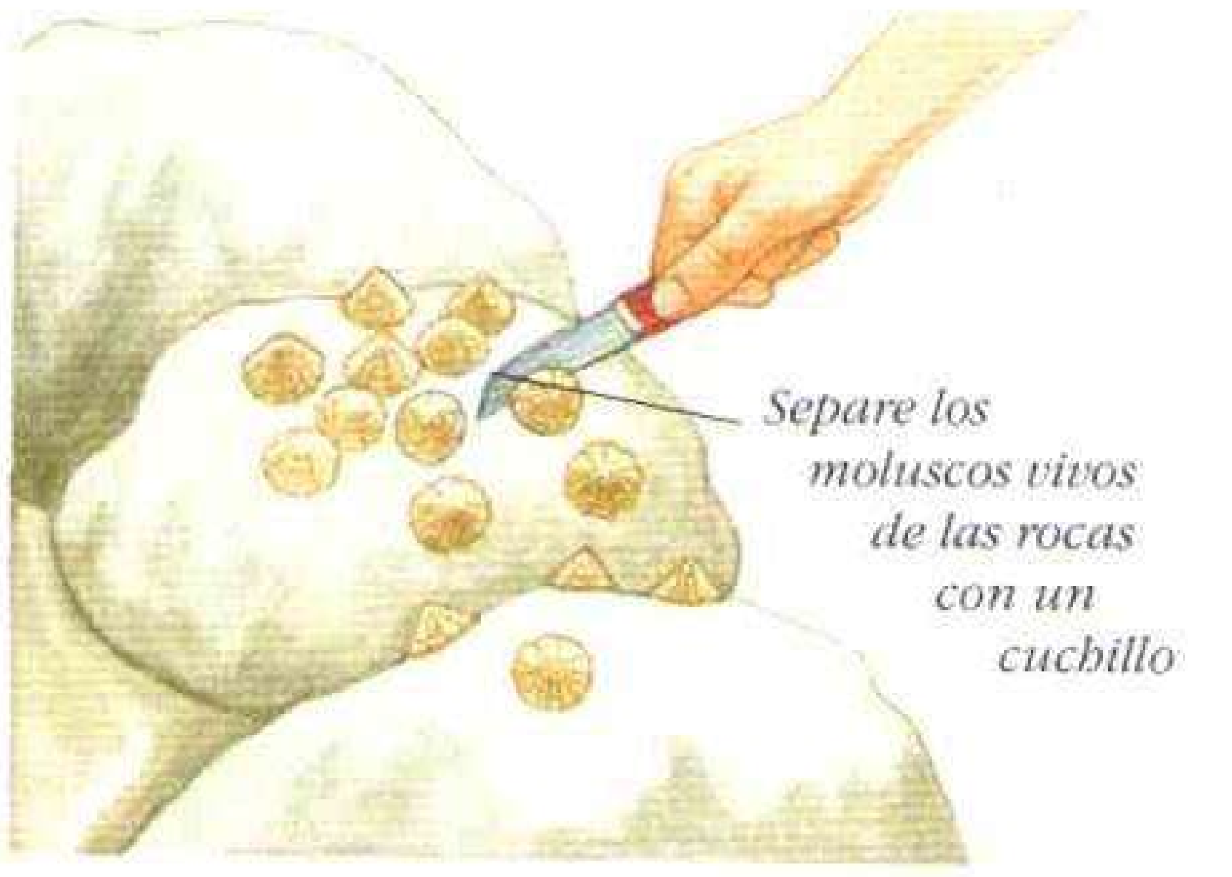

Separe los moluscos vivos de las rocas con un cuchillo

Bígaros y lapas. *Coja los moluscos pequeños, como las lapas adheridas a las rocas, cuando baje la marea. Tome únicamente aquellas que estén vivas —que estarán fuertemente adheridas.*

ADVERTENCIA

Nunca está de más advertir del cuidado que debe tener al manipular los crustáceos y moluscos. No debe comerlos nunca crudos, y deben hervirse o cocinarse al vapor por lo menos durante cinco minutos, y siempre inmediatamente después de pescarlos.

CÓMO COCINAR CRUSTÁCEOS Y MOLUSCOS

Hierva los crustáceos y moluscos en sus conchas por lo menos durante cinco minutos antes de comerlos. Deseche aquellos que no se abran en agua hirviendo, lo que significa que están en mal estado. Después de hervirlos, extraiga los bígaros y lapas de sus conchas con una ramita o un tenedor.

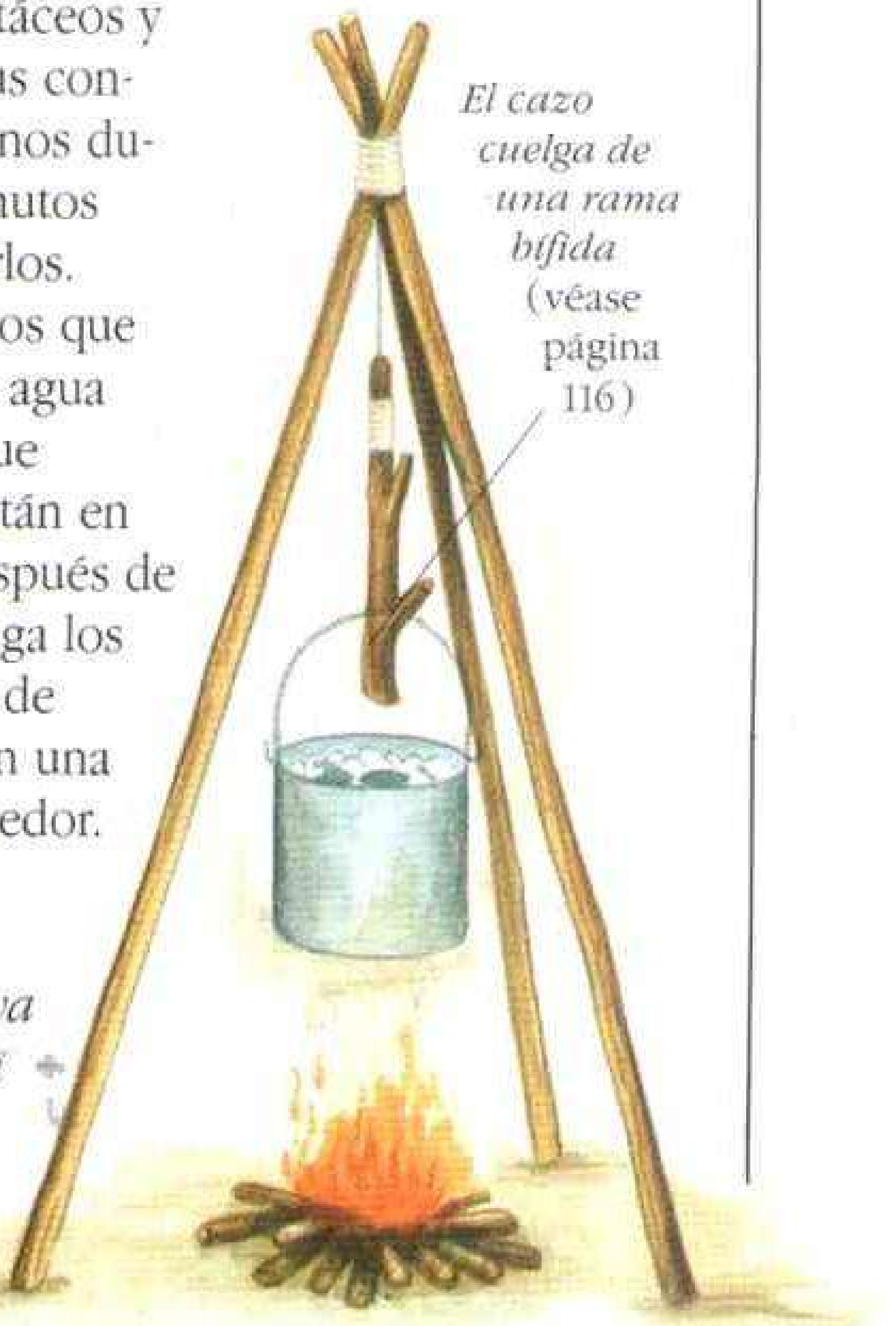

El cazo cuelga de una rama bífida (véase página 116)

Estofado de marisco. *Hierva el marisco a fin de eliminar las sustancias nocivas.*

OTRAS DELICIAS INVERTEBRADAS

Los caracoles, las babosas y las lombrices son comestibles y pueden cogerse al amanecer. Mantenga los caracoles y las babosas vivos, con una dieta a base de hojas verdes durante 24 horas, antes de matarlos colocándolos en agua hirviendo. Puede hervir los caracoles, pero las babosas saben mejor asadas. Remoje las lombrices en agua salada durante 24 horas, y después extraiga sus partes blandas antes de añadirlas a un estofado o de secarlas.

Hormigas colmeneras. *Las hormigas colmeneras se alimentan de néctar y agua, y cada una de ellas se convierte en un auténtico almacén viviente que cuelga del nido.*

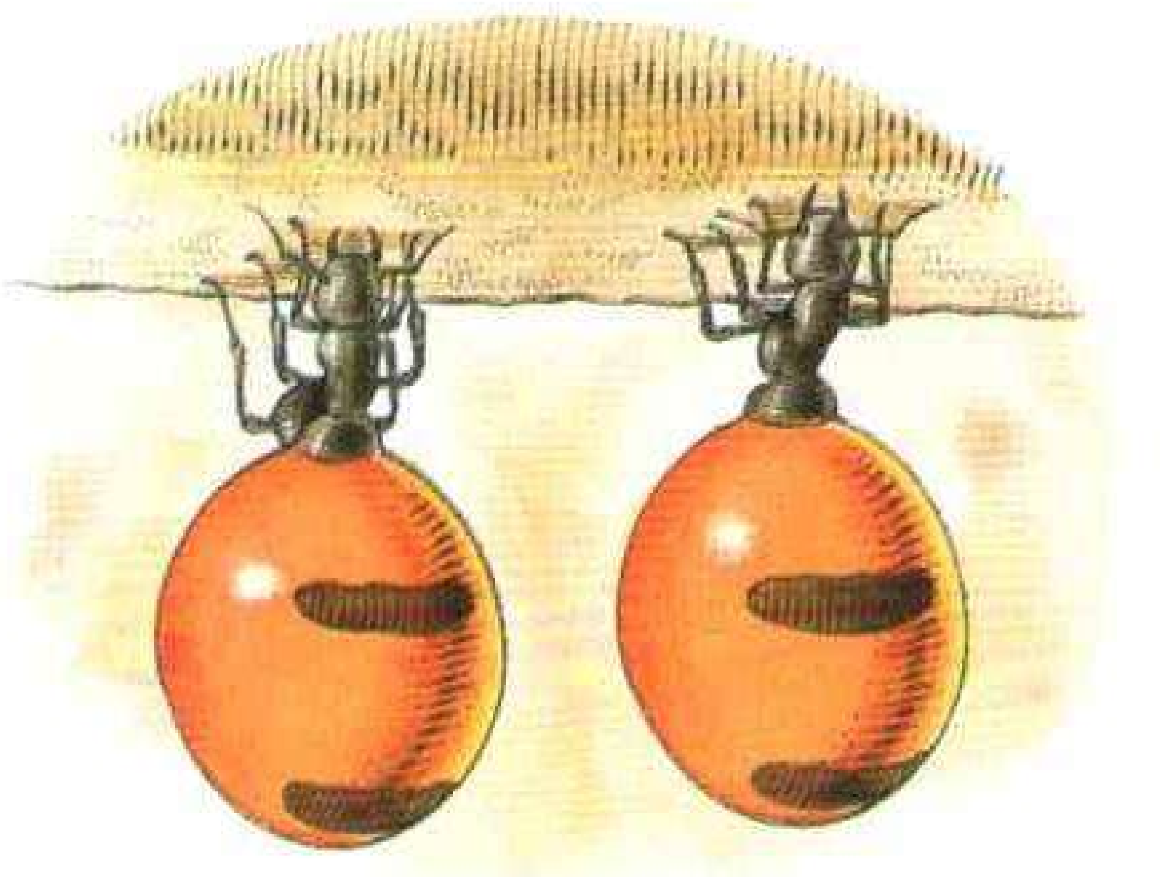

CÓMO PREPARAR UN CANGREJO

Cuando coja cangrejos, tenga cuidado con sus pinzas. Quizá se haga necesario atar las pinzas de las especies más grandes para evitar que se peleen y se hieran entre sí. Mátelos sumergiéndolos en agua hirviendo.

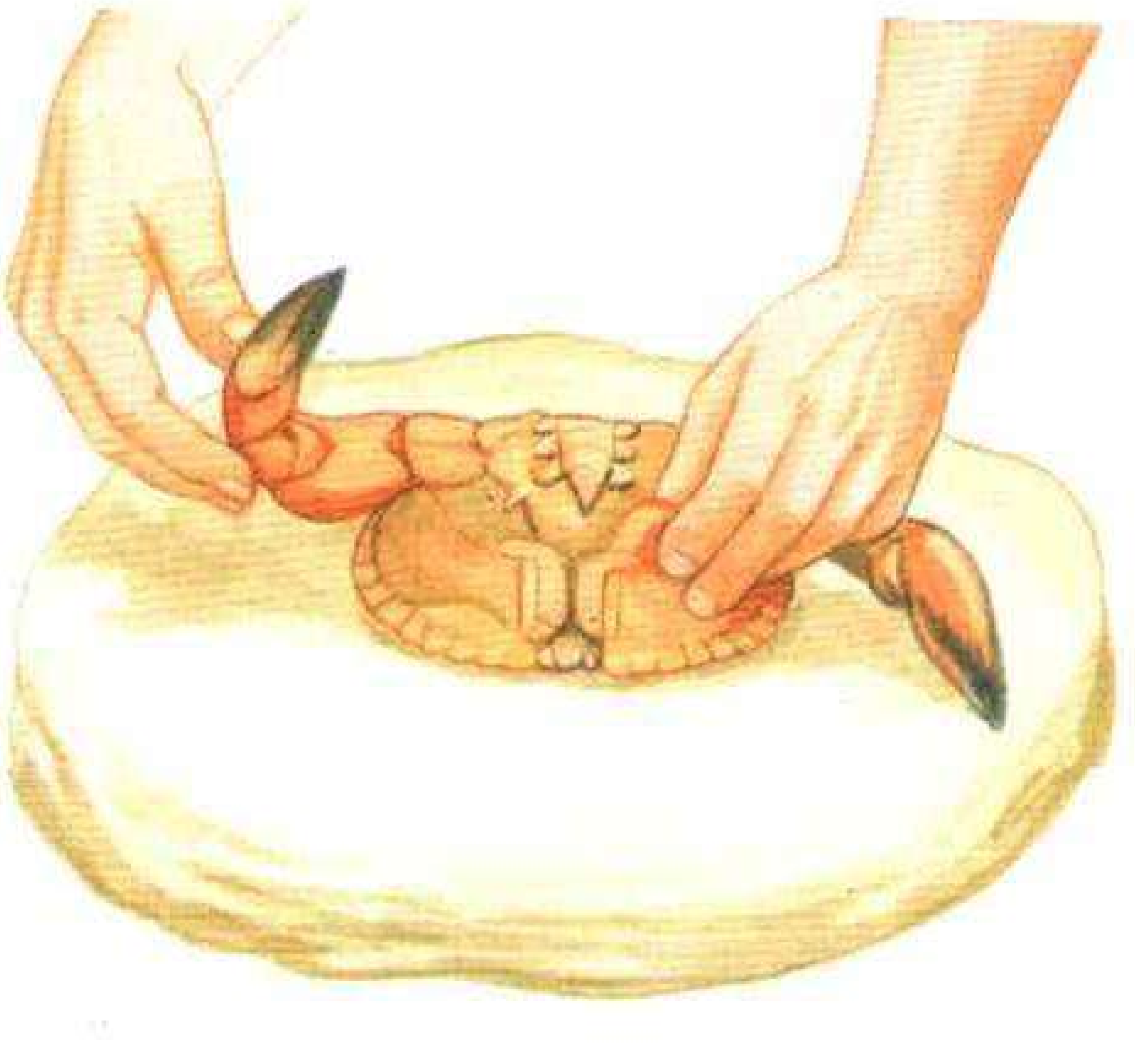

1 Un cangrejo debe prepararse con cuidado a fin de que todos sus órganos venenosos sean eliminados. Para preparar un cangrejo para comer, arranque primero las patas y las pinzas grandes. Todas ellas contienen carne comestible.

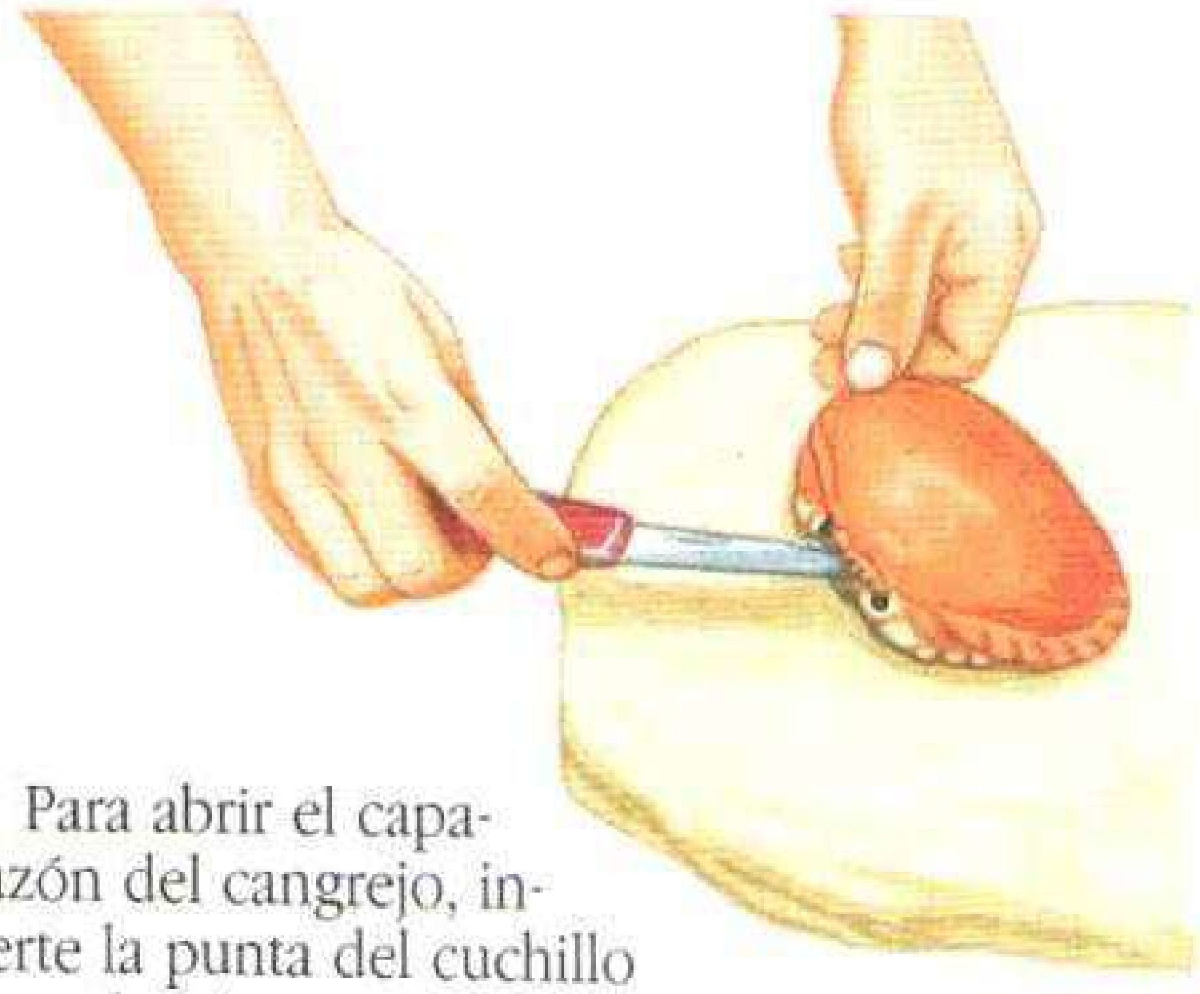

2 Para abrir el caparazón del cangrejo, inserte la punta del cuchillo entre las dos mitades y hágalo girar.

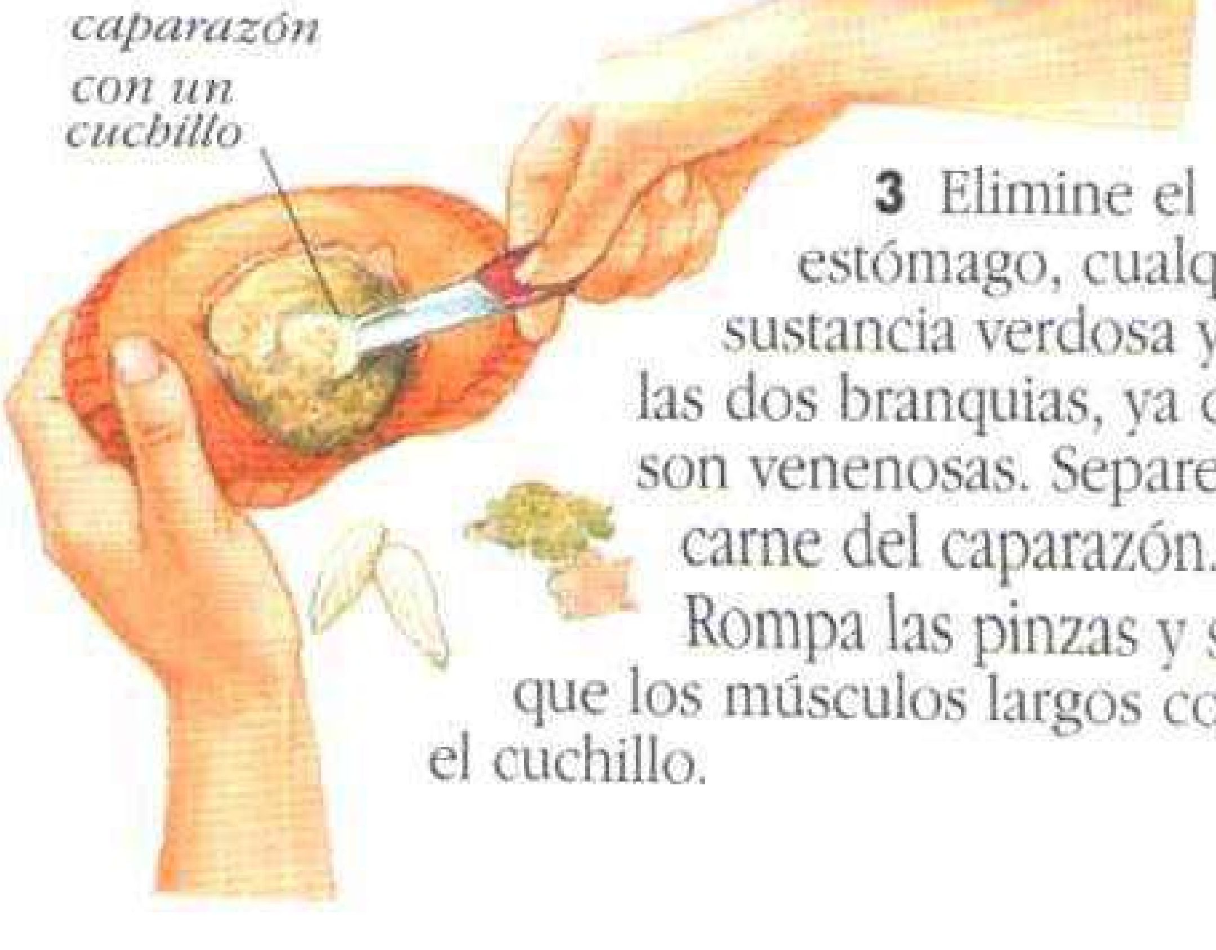

Separe la carne del caparazón con un cuchillo

3 Elimine el estómago, cualquier sustancia verdosa y las dos branquias, ya que son venenosas. Separe la carne del caparazón. Rompa las pinzas y saque los músculos largos con el cuchillo.

PECES DE AGUA DULCE

LOS PECES DE agua dulce viven en ríos, corrientes y lagos, así como en los estuarios de los ríos más arriba de donde el agua dulce se sala y penetra en el mar. Algunas especies recorren los ríos en toda su extensión, mientras que algunos peces marinos, principalmente el salmón, migran hacia el interior para desovar en la relativa seguridad de las corrientes tierra adentro. Las condiciones y el tamaño de los peces de agua dulce varían de acuerdo a la cantidad y variedad de su alimento, y con las estaciones. Aunque los lagos y algunos ríos pueden tener tramos profundos lejos de la orilla, la mayoría de los peces de agua dulce viven en las aguas relativamente poco profundas de las orillas.

PECES COMESTIBLES

La mayoría de los peces que se encuentran en agua dulce pueden comerse, aunque algunas especies tienen un sabor más agradable, o son más fáciles de preparar que otras. Algunos peces tienen tal cantidad de pequeñas espinas que dificultan su preparación; estas espinas pueden ser peligrosas si se atraganta con ellas. A continuación se muestran algunos ejemplos de peces de agua dulce comestibles.

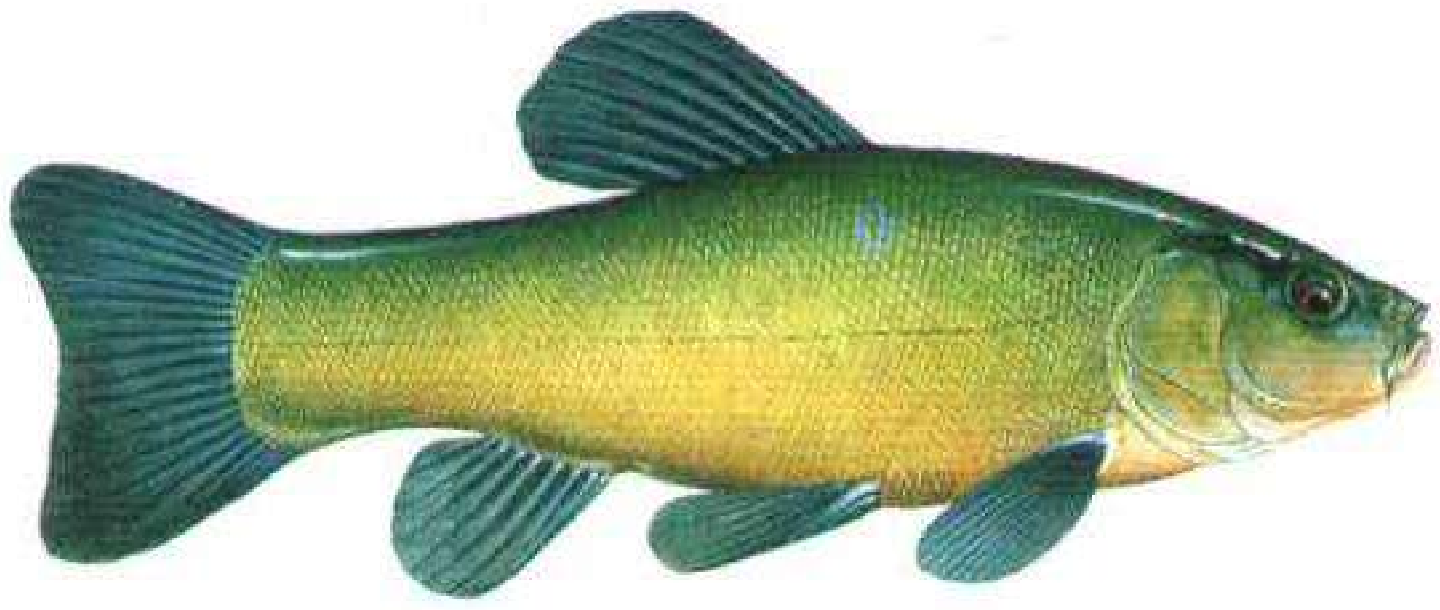

Tenca (Tinca). *La tenca vive en las aguas tranquilas de los tramos bajos de los ríos, antes de desembocar en el mar. Se encuentra en Europa, Norteamérica y Australia.*

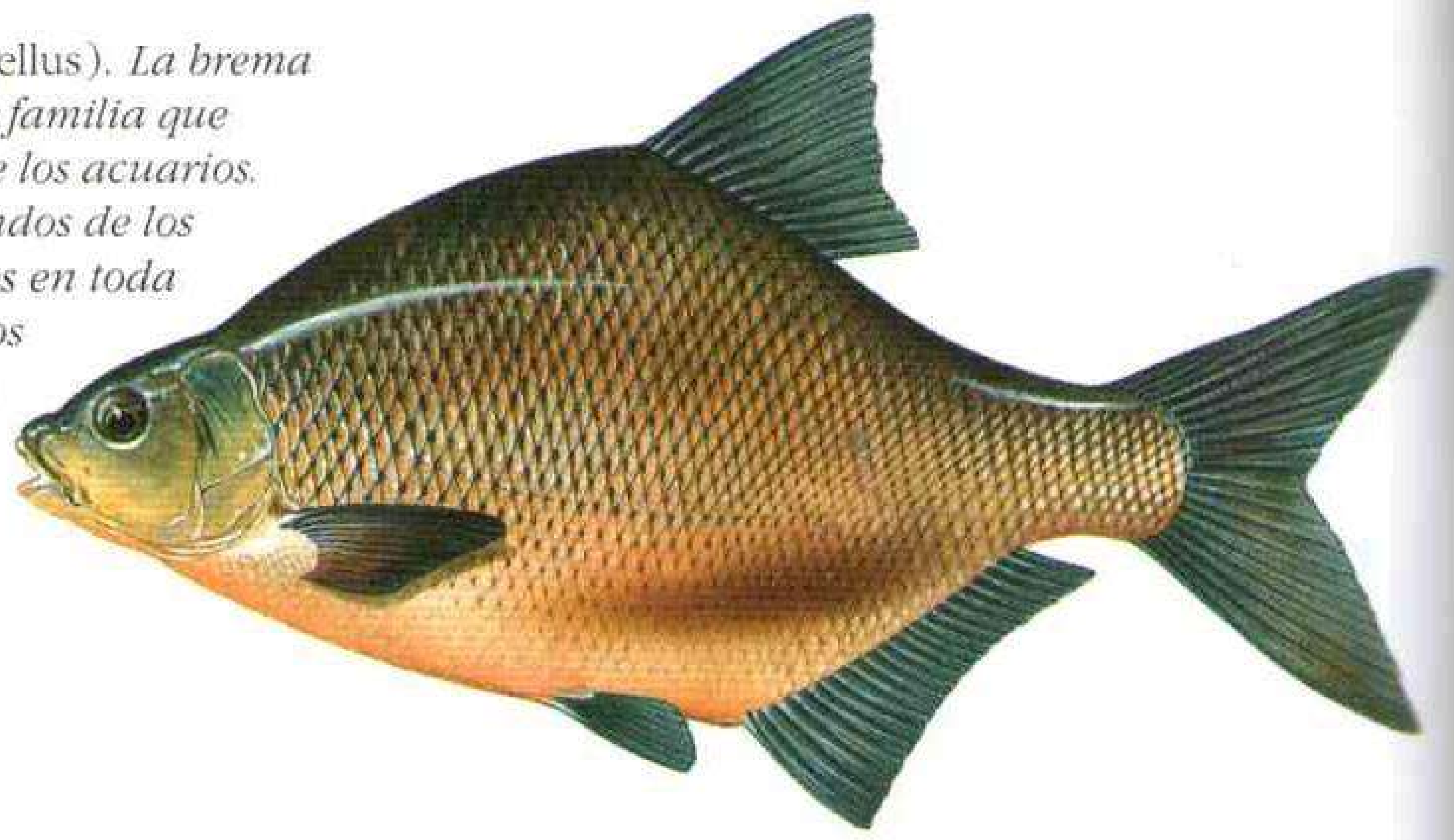

Brema (Abramis, Pagellus). *La brema pertenece a la misma familia que los peces de colores de los acuarios. Se alimenta en los fondos de los ríos profundos y lentos en toda Europa y Asia. Algunos individuos alcanzan un tamaño considerable.*

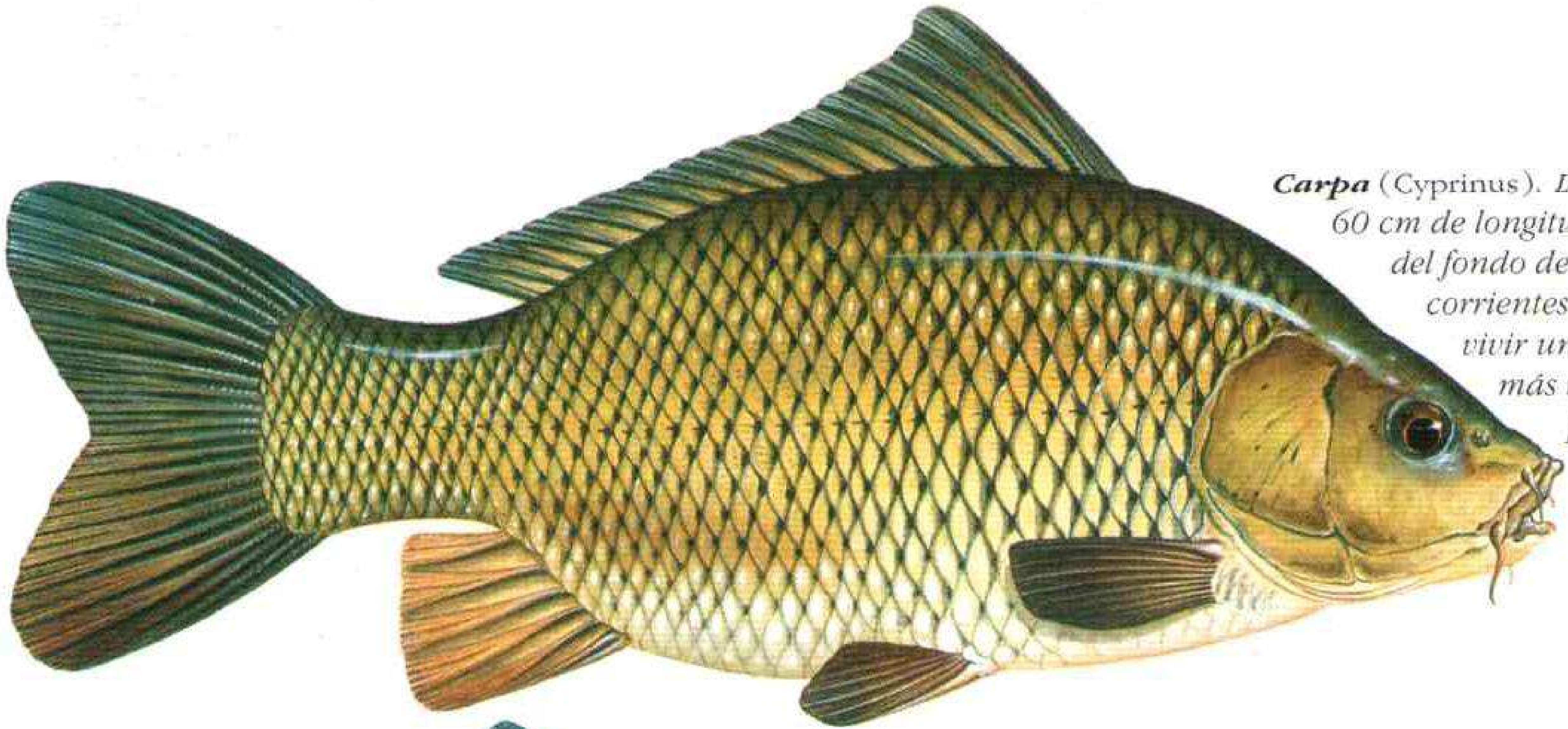

Carpa (Cyprinus). *La carpa común crece hasta unos 60 cm de longitud y se alimenta en las aguas oscuras del fondo de los ríos, o en la superficie de las corrientes y ríos. Aunque la leyenda dice que puede vivir unos cuantos siglos, 40 años es una cifra más verosímil, así que no debe sentirse culpable si pesca un ejemplar grande.*

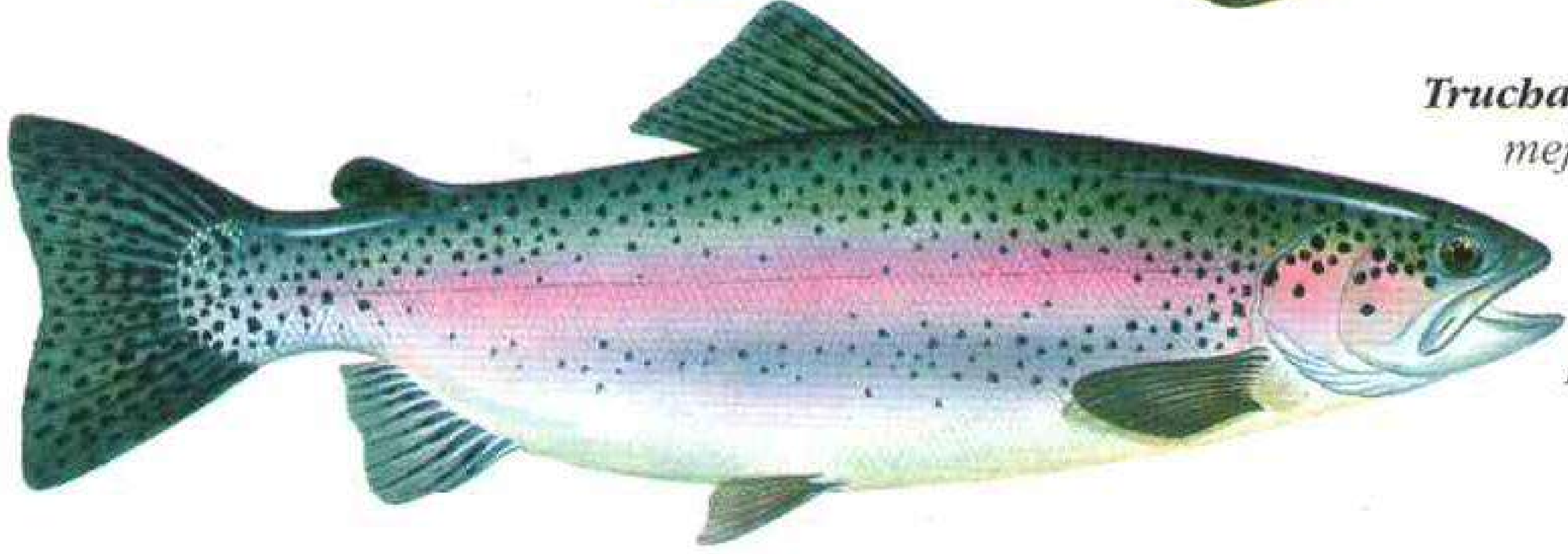

Trucha (Oncorhynchus). *La trucha se asemeja mucho al salmón* (véase página 100). *La trucha arco iris vive en las corrientes frescas y limpias, y en los lagos de Escandinavia, Norteamérica, Australasia, África e India.*

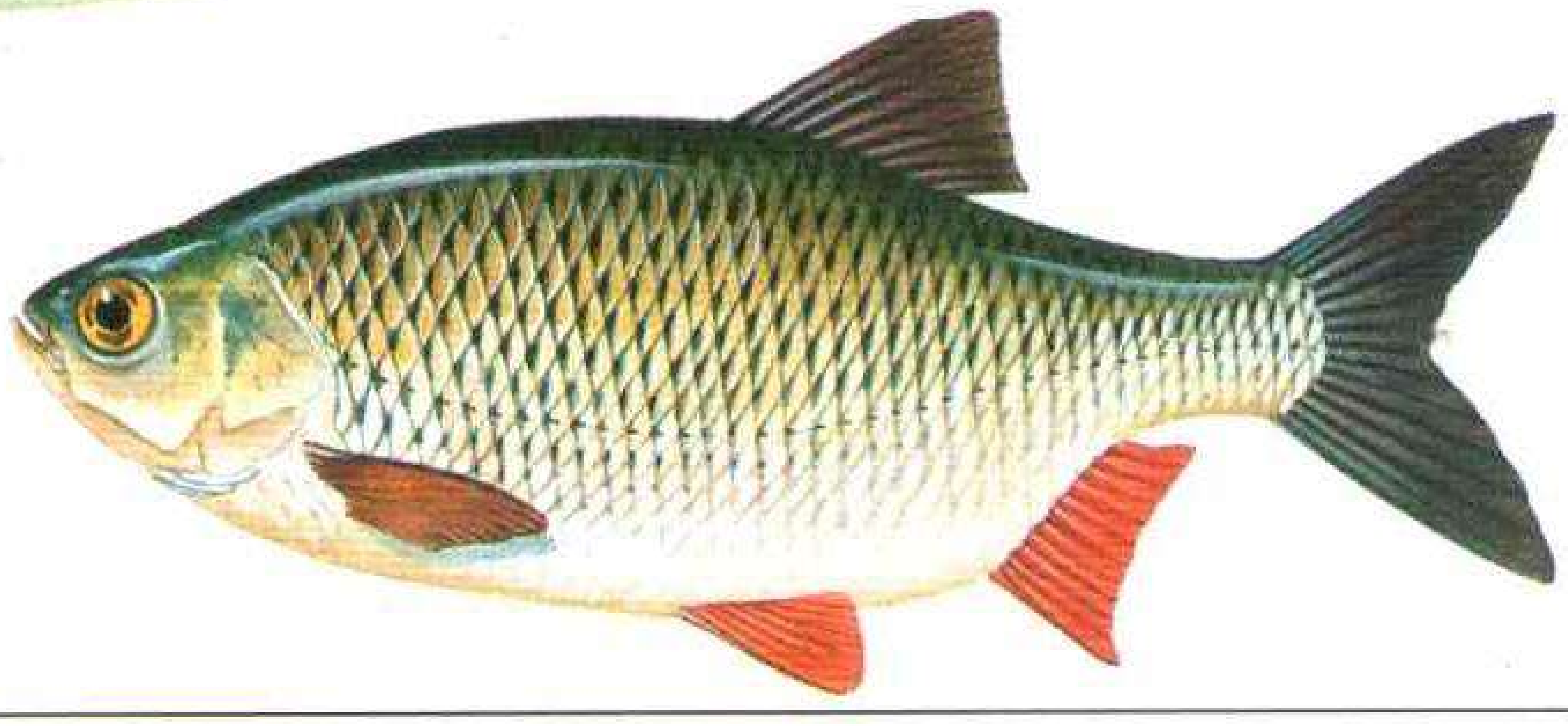

Escardinio (Scardinus). *El escardinio vive en aguas tranquilas con vegetación abundante, entre la que se oculta. Se encuentra en Norteamérica, Europa y Escandinavia. Se alimenta en la superficie, aunque algunos especímenes viejos y grandes a menudo descansan en el fondo.*

OTROS PECES COMESTIBLES

Existen muchas otras especies de agua dulce comestibles, en especial en los ríos y lagos tropicales. Las especies más comunes en todo el mundo incluyen el albur y el cacho *(Leuciscus)*, y el gobio *(Rutilus)*. La observación determinará qué especies se alimentan en el fondo o en la superficie, y la experiencia le dirá cuáles son las mejores para comer. Sin embargo, tenga cuidado con las especies desconocidas. Ante la duda, áselos a la parrilla, elimine la piel y las espinas, y después hierva la carne.

PECES PELIGROSOS

Los peces depredadores son más agresivos que sus presas, e incluso llegan a comerse a los miembros menores de su propia especie. Muchos de estos peces tienen aguijones afilados, que pueden estar conectados a glándulas venenosas. La mucosidad de su piel puede ser tóxica y causar erupciones en la piel. Incluso puede hacer que el pescado sea venenoso por ingestión a menos que se elimine la piel.

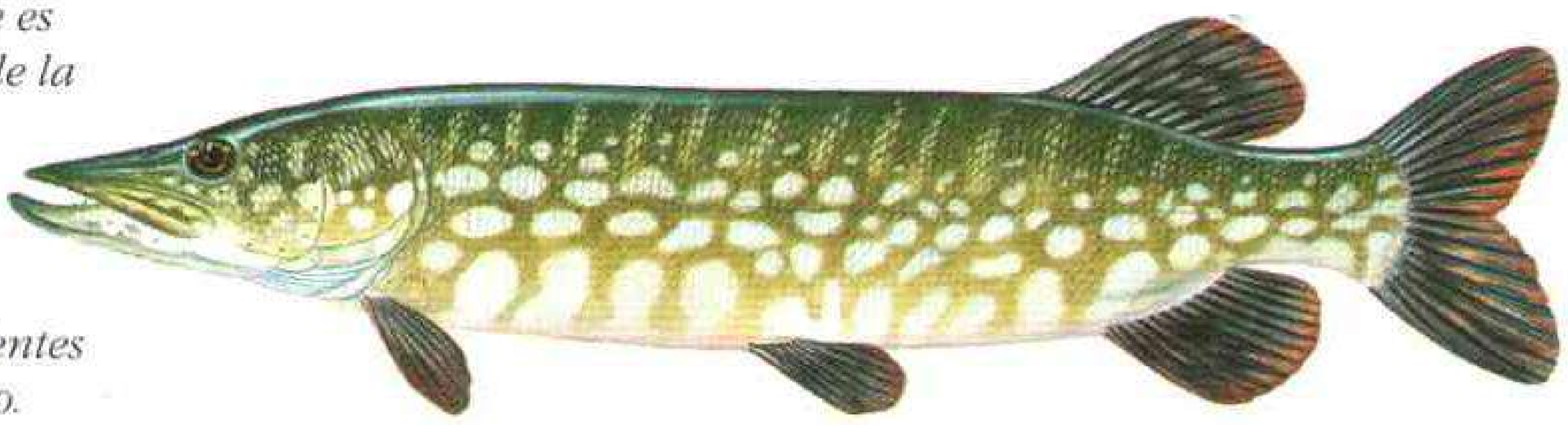

Lucio americano (Esox). *Éste es un miembro norteamericano de la familia de los lucios, y puede llegar a medir hasta 1 m de longitud. Prefiere las aguas tranquilas y claras con vegetación, y habita los ríos y corrientes que desembocan en el Atlántico. Tiene dientes afilados.*

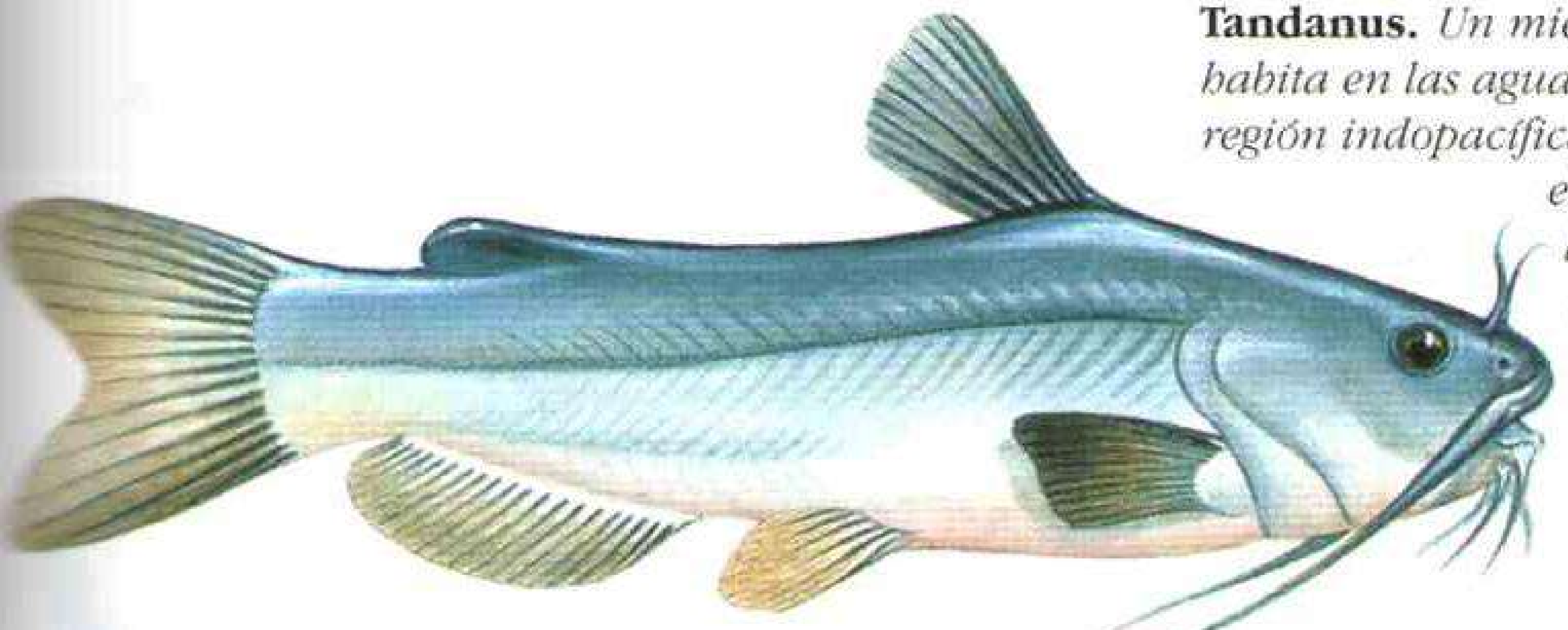

Tandanus. *Un miembro de la familia de los siluros, habita en las aguas tranquilas de Australasia y de la región indopacífica. Otros siluros se encuentran en todo el mundo. La aleta dorsal o pectoral de los siluros puede tener una espina dura, en ocasiones en contacto con una glándula de veneno. Los siluros también están cubiertos por una mucosidad venenosa.*

OTROS PECES PELIGROSOS

Muchos peces comestibles de agua dulce pueden causar heridas con sus espinas o dientes afilados, o bien tener la piel revestida de una mucosidad venenosa. La conocida piraña *(Serrasalmus)* de la Sudamérica tropical es un ejemplo de pez peligroso y voraz, mientras que la anguila eléctrica *(Electrophorus)*, también originaria de Sudamérica, puede herir a una persona con una potente descarga eléctrica al nadar o vadear las aguas.

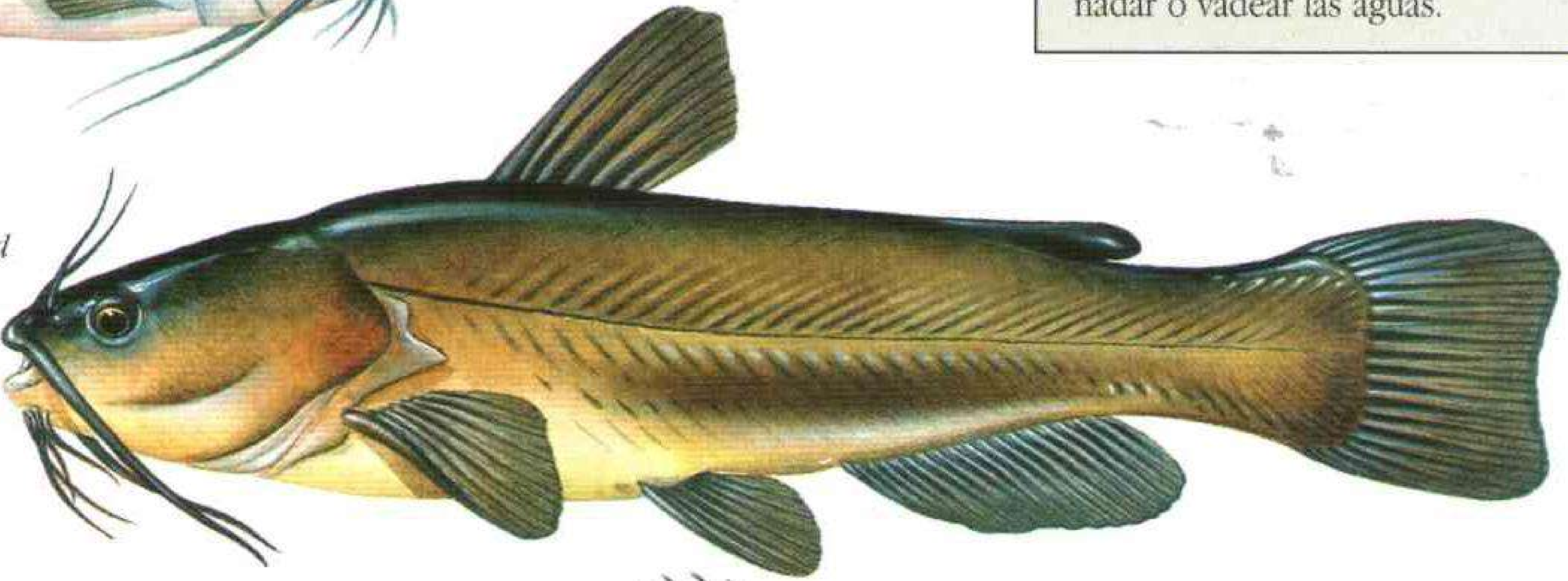

Siluro (Ictalurus). *El siluro tiene espinas venenosas y está cubierto por una mucosidad tóxica. Debe manipularse con cuidado y eliminar la piel antes de comerlo* (véase página 110). *El siluro come plantas e insectos, y se alimenta mayoritariamente en los fondos de las corrientes lentas y tranquilas, particularmente en Norteamérica y Europa. Es una especie de bagre, y tiene los bigotes característicos de esta familia.*

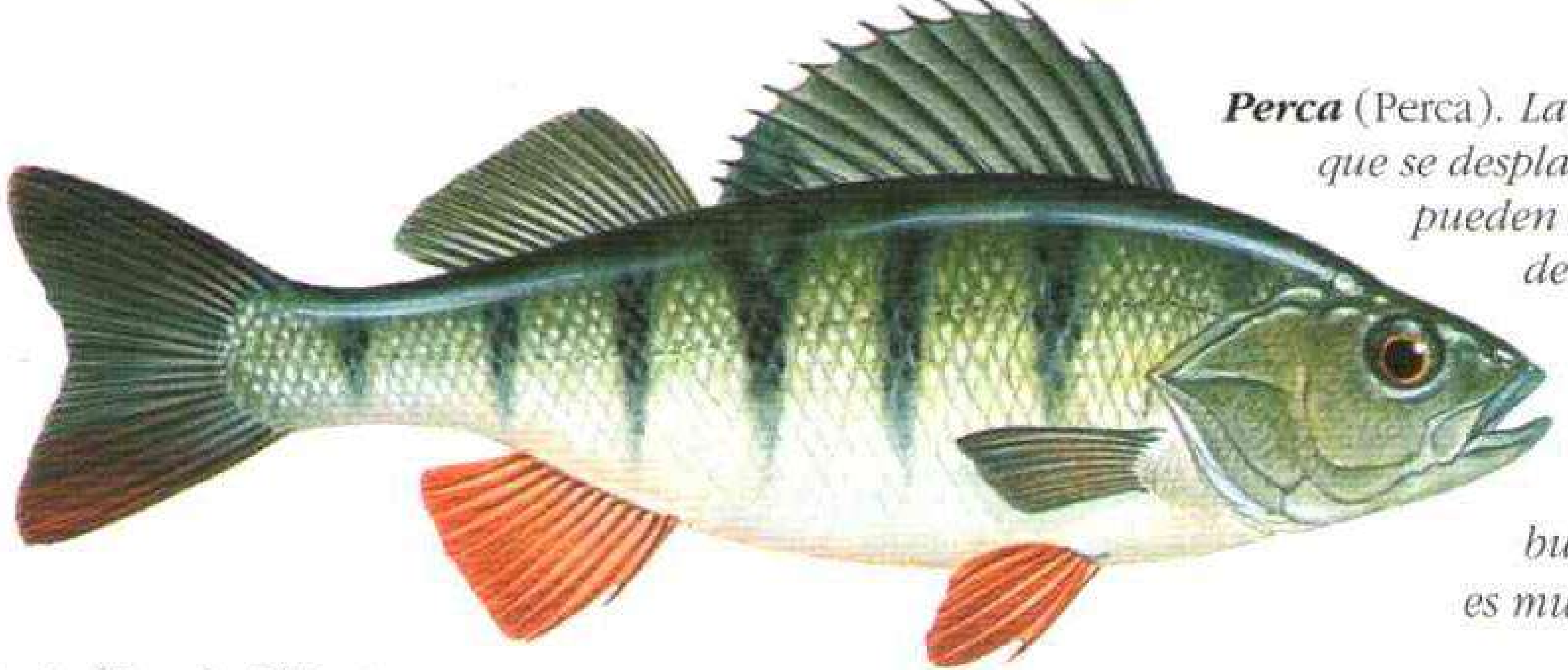

Perca (Perca). *La perca es un devorador voraz que se desplaza en grandes cardúmenes que pueden llegar a mermar la población de otros peces de menor tamaño. Se encuentra en los ríos y corrientes tranquilas y lentas de Europa, Asia, Norteamérica y Australia. Tiene un buen sabor, pero su aleta dorsal es muy afilada.*

ADVERTENCIA

Una vez muerto, a menos que se conserve seco, ahumado o en escabeche (*véase página 118*), todo pescado se convierte en una fuente potencial de intoxicación debido a la proliferación de bacterias. A menos que le haya dado muerte usted mismo, es recomendable no comer ningún pescado muerto, independientemente de lo fresco que parezca.

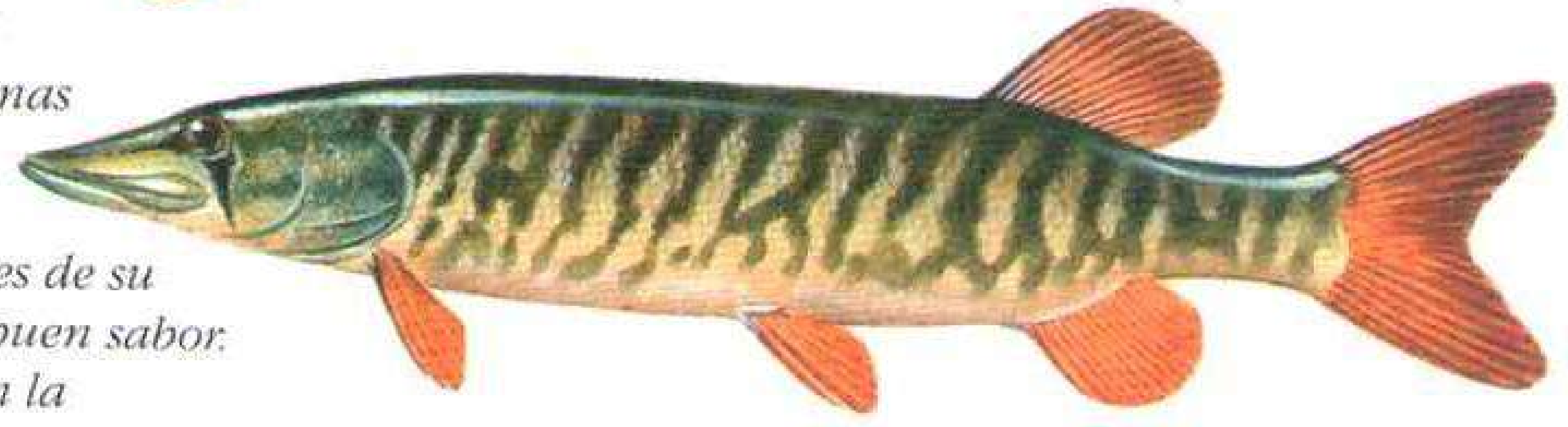

Lucio (Esox). *El lucio es un pez solitario y agresivo, con dientes muy afilados. En algunas áreas se le conoce como* lobo de río *debido a su apetito y poder de destrucción, ya que llega a pelear y comerse incluso peces de su mismo tamaño. A pesar de ello, tiene muy buen sabor. Se encuentra en Europa y Norteamérica, en la mayor parte de tipos de agua.*

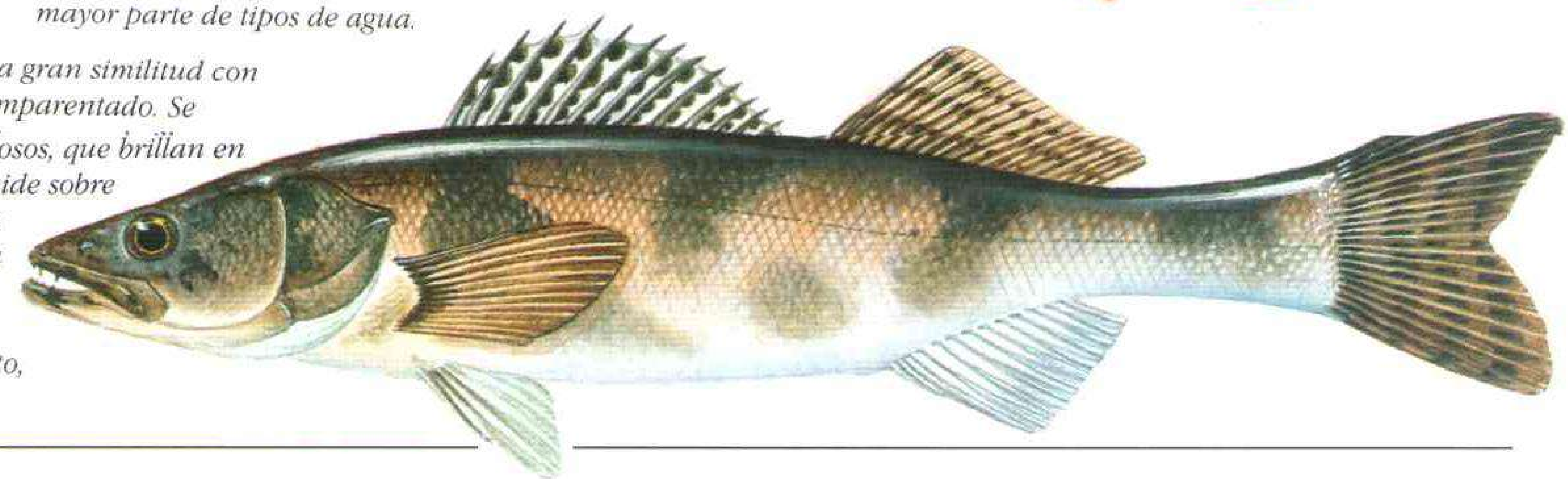

Lucioperca (Stizostedion). *Tiene una gran similitud con la perca y el lucio, con quienes está emparentado. Se caracteriza por sus grandes ojos vidriosos, que brillan en la oscuridad reflejando la luz que incide sobre ellos. Se encuentra principalmente en Norteamérica y pertenece a la familia de las percas, por lo que tiene un sabor agradable. Sin embargo, tiene una peligrosa aleta dorsal; por lo tanto, tenga cuidado al manipularlo.*

PECES DE AGUA SALADA

COMO REGLA GENERAL, los grandes peces de agua salada viven en los mares profundos, mientras que los grandes cardúmenes de peces de menor tamaño viven en las aguas poco profundas cercanas a la costa. Los peces pueden llegar a migrar a grandes distancias en búsqueda de alimento, y procrear, asediados por multitud de predadores. Los peces grandes pueden llegar a acercarse lo suficiente a la costa para ser atrapados con una red o una caña, pero es más fácil pescarlos desde un barco mar adentro. Si carece de embarcación, puede vadear las zonas poco profundas de los mares o estuarios para pescar peces de agua salada, o bien adentrarse en los canales más profundos cuando baja la marea.

PECES COMESTIBLES

En realidad, existen muy pocos peces que no puedan comerse. La forma más adecuada y saludable de comer pescado es crudo, para conservar todos sus aceites naturales y su frágil valor nutricional. Para cocinarlo, prepárelo al vapor, a la brasa u horneado *(véase página 117)*. Sin embargo, deseche la piel porque puede estar revestida de una mucosidad tóxica, o tener espinas peligrosas.

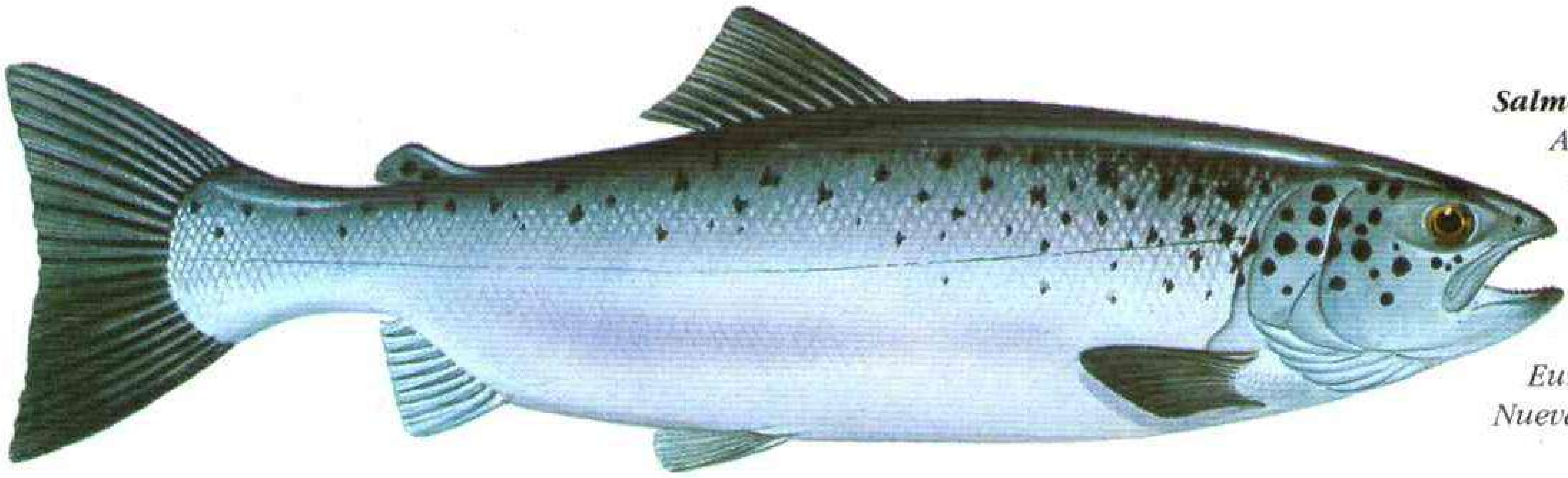

Salmón (Salmo). *El salmón del Atlántico, el «rey de los peces», recorre grandes distancias río arriba para desovar; es un pescado delicado y su carne es muy apreciada como alimento. Se encuentra en Europa, Norteamérica, Australia, Nueva Zelanda, Sudamérica y Asia.*

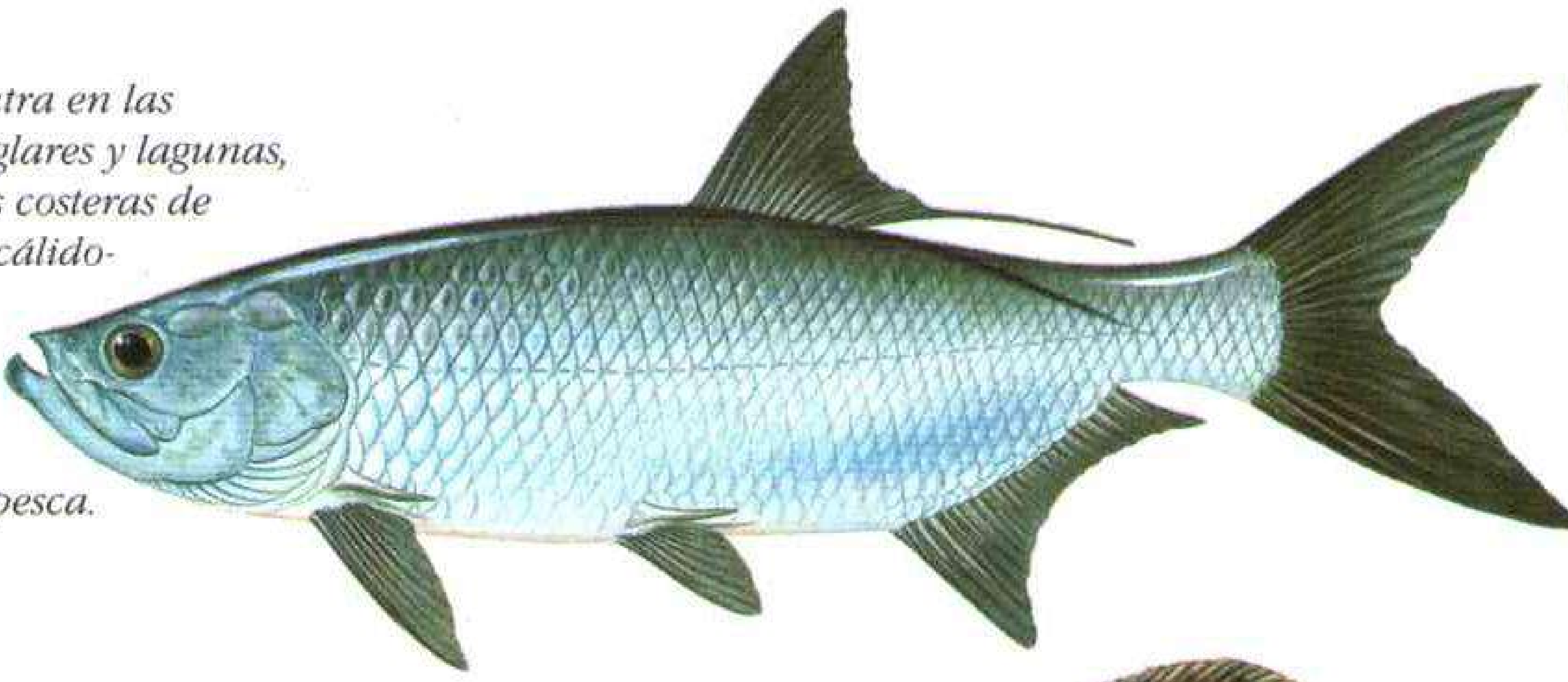

Megalops. *Se encuentra en las pequeñas calas, manglares y lagunas, así como en las aguas costeras de las áreas tropicales y cálido-templadas. Tiene escamas grandes y recias, y puede llegar a ofrecer una gran resistencia si se pesca.*

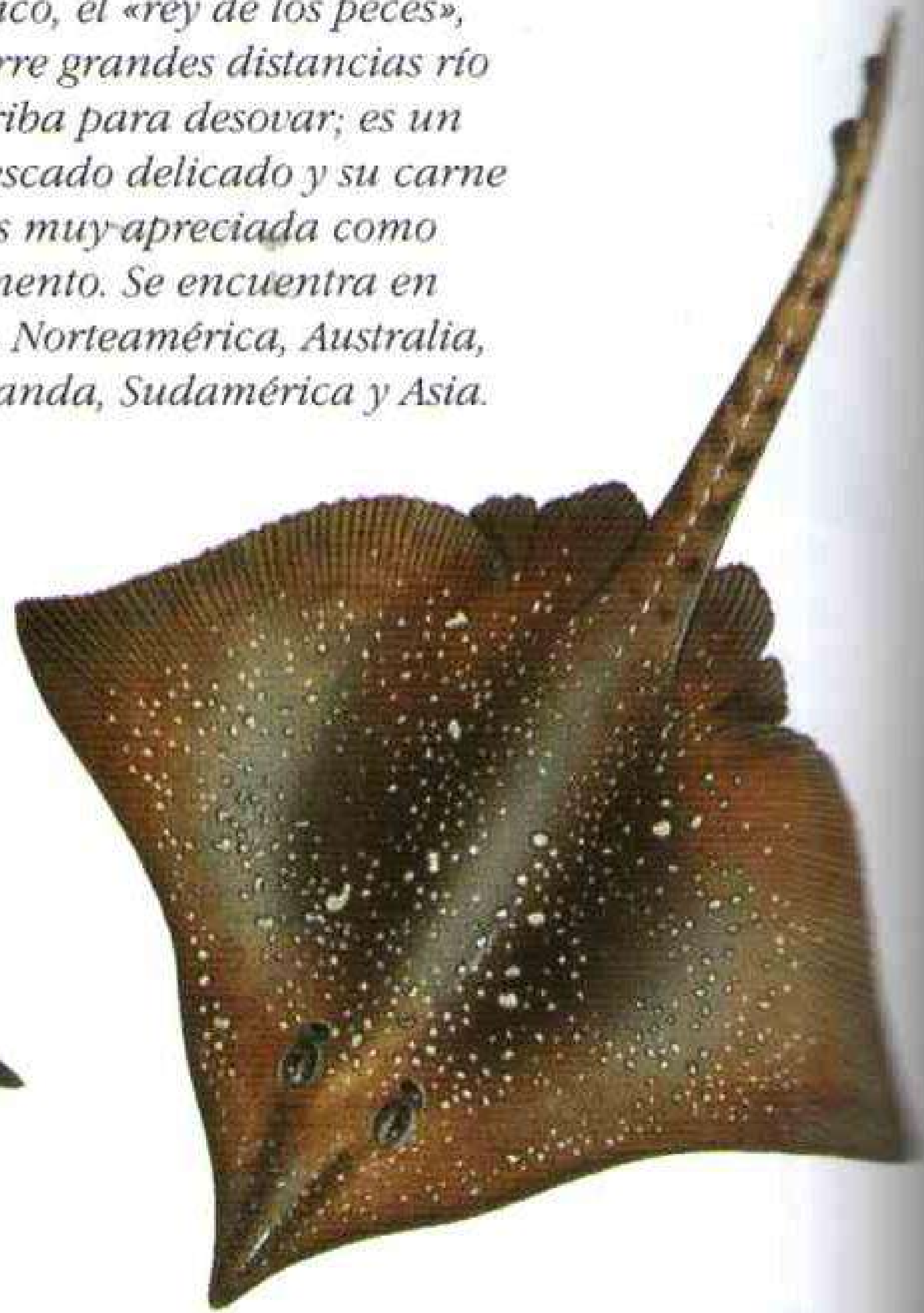

Raya (Raja). *La raya pasa la mayor parte del día enterrada en la arena o en el fango del fondo de las aguas del Pacífico o del Atlántico, emergiendo de noche para comer moluscos u otros peces. Los individuos de algunas especies pueden alcanzar hasta 2,4 m de longitud.*

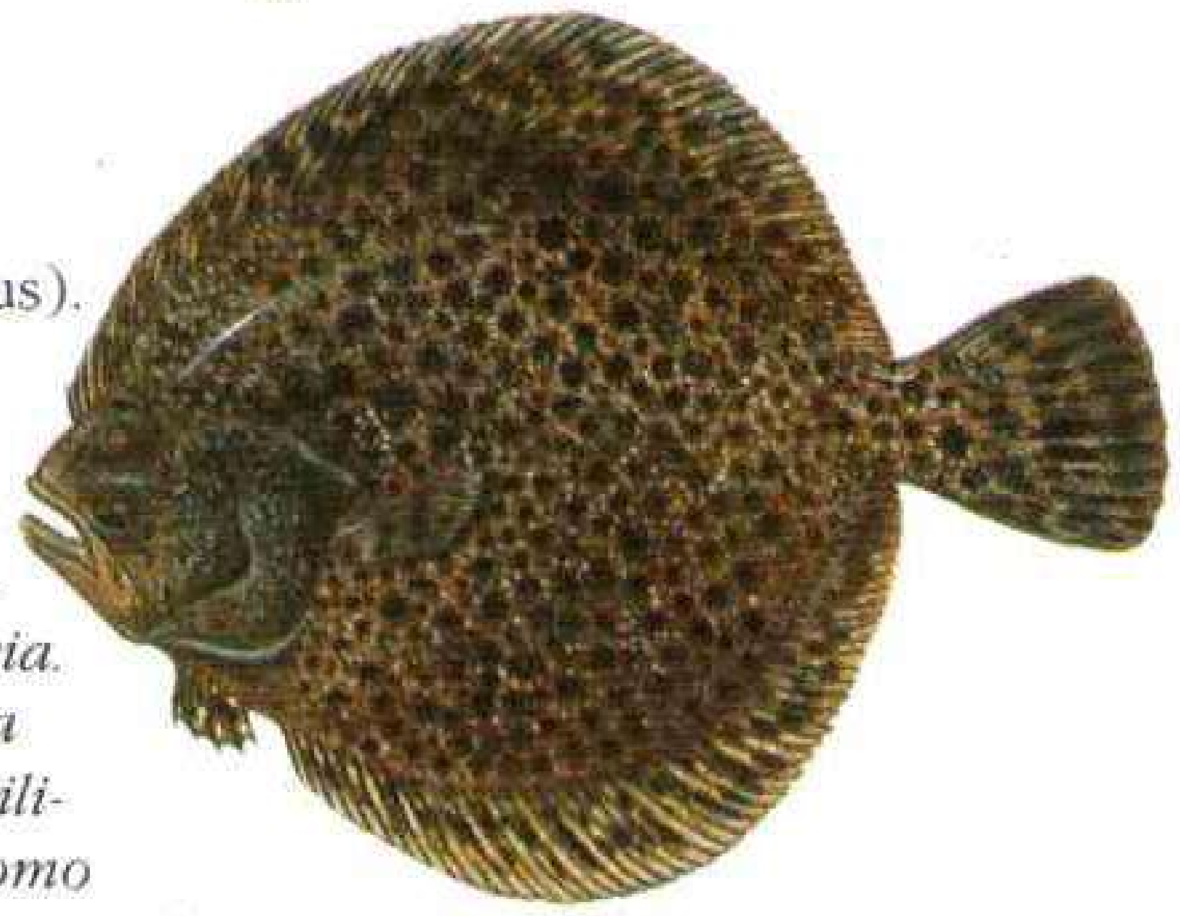

Rodaballo (Scophthalmus). *Un pez popular para la mesa, el rodaballo, se encuentra cerca de las costas en Europa, principalmente en Escandinavia. Se puede pescar de forma relativamente sencilla utilizando pequeños peces como carnada.*

OTROS PECES COMESTIBLES

El arenque *(Clupea)* se ha convertido en una rareza en las aguas templadas cercanas a la costa, pero aún constituye una captura habitual para los pescadores de todo el mundo. Se mueve en grandes cardúmenes que pueden llegar a pasar cerca de la costa, donde algunos pueden ser pescados con red o anzuelo. La lubina *(Morone)* es un pez delicioso, de tamaño medio, que puede pescarse con anzuelo de noche cuando baja la marea, ya que se acerca a la costa para alimentarse. Otros peces comunes comestibles, de agua salada, son las variadas especies de mújol *(Mullidae, Muglidae)*, tímalos *(Thymallus)* y meros *(Serranidae)*.

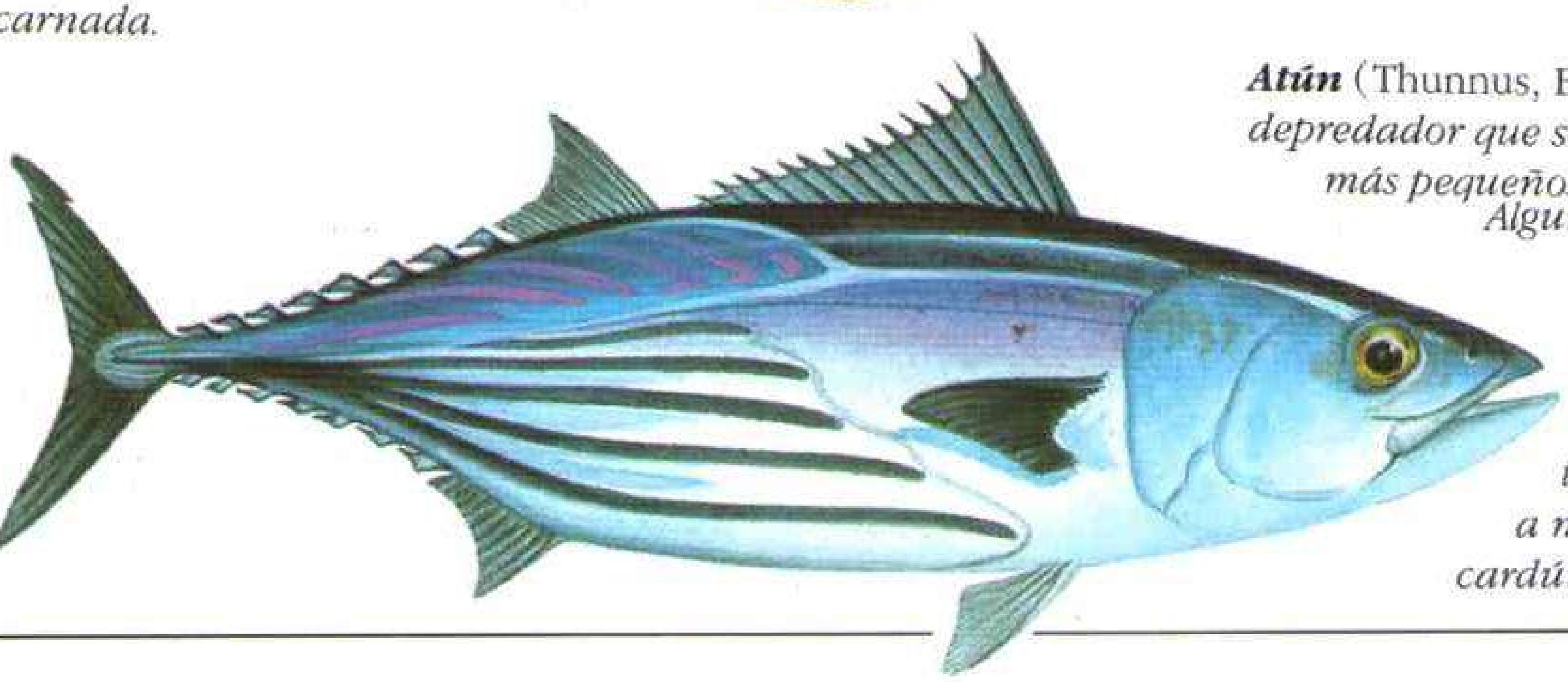

Atún (Thunnus, Euthynnus). *Pez depredador que se alimenta de peces más pequeños, vivos o muertos. Algunas especies pueden llegar a alcanzar un tamaño considerable. El atún se encuentra en aguas templadas o tropicales, a menudo en grandes cardúmenes cerca de la costa.*

PECES PELIGROSOS

A pesar de que los peces de mayor tamaño tienden a habitar las aguas más profundas, las especies grandes y peligrosas también patrullan las aguas poco profundas en busca de comida fácil. Es bien sabido que los tiburones hambrientos llegan hasta las playas y atacan tanto a animales como a bañistas. Pueden detectar la sangre y el sonido de los peces o animales heridos incluso a grandes distancias. Algunos peces tienen espinas o púas que pueden causar un dolor muy intenso. Otros tienen dientes muy afilados.

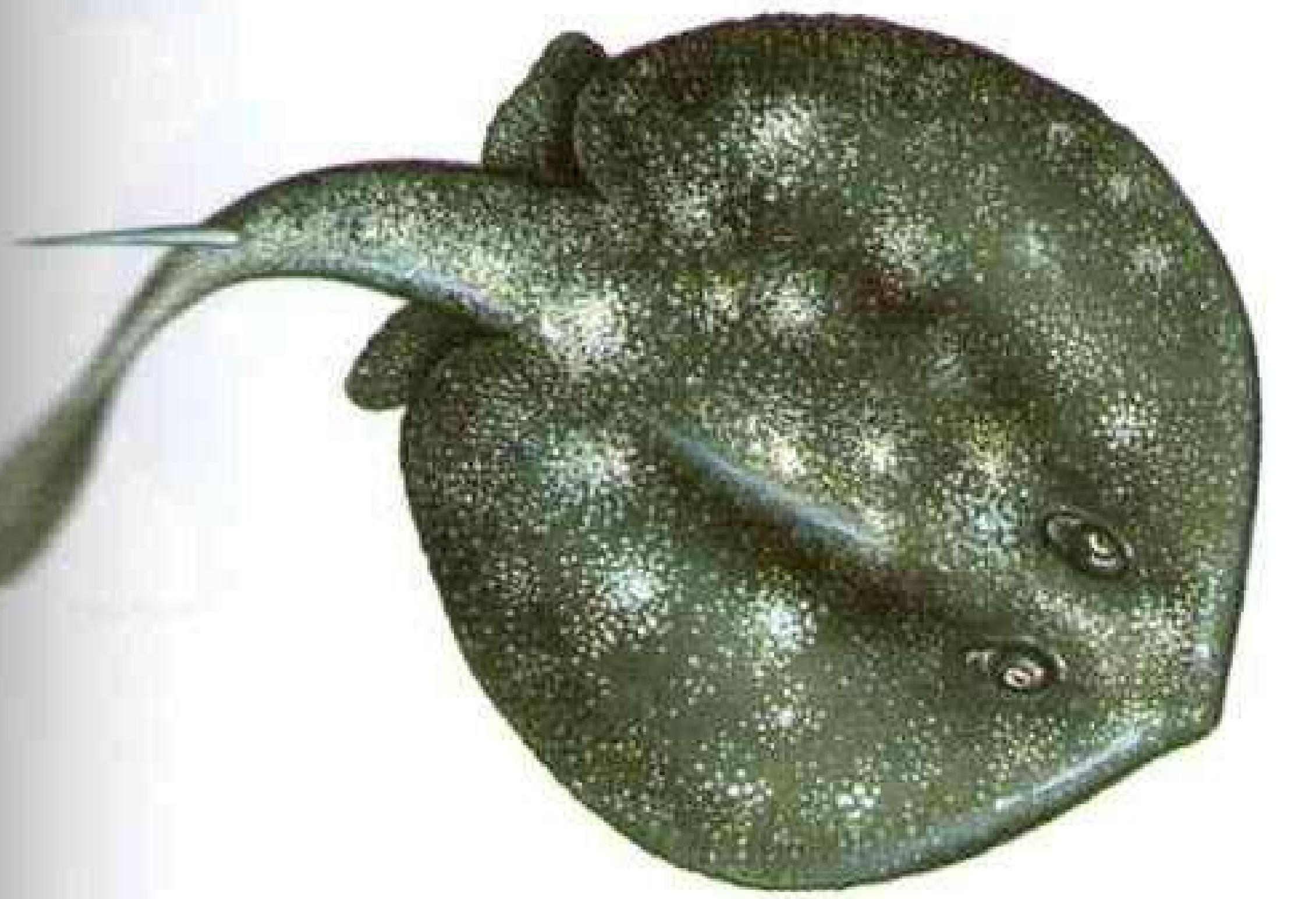

Pastinaca (Urolophus, Dasyatis). *La pastinaca tiene aguijones venenosos y una cola que asemeja un látigo. Su picadura puede ser fatal si no se trata adecuadamente. Vive en el Pacífico Norte y en aguas tropicales.*

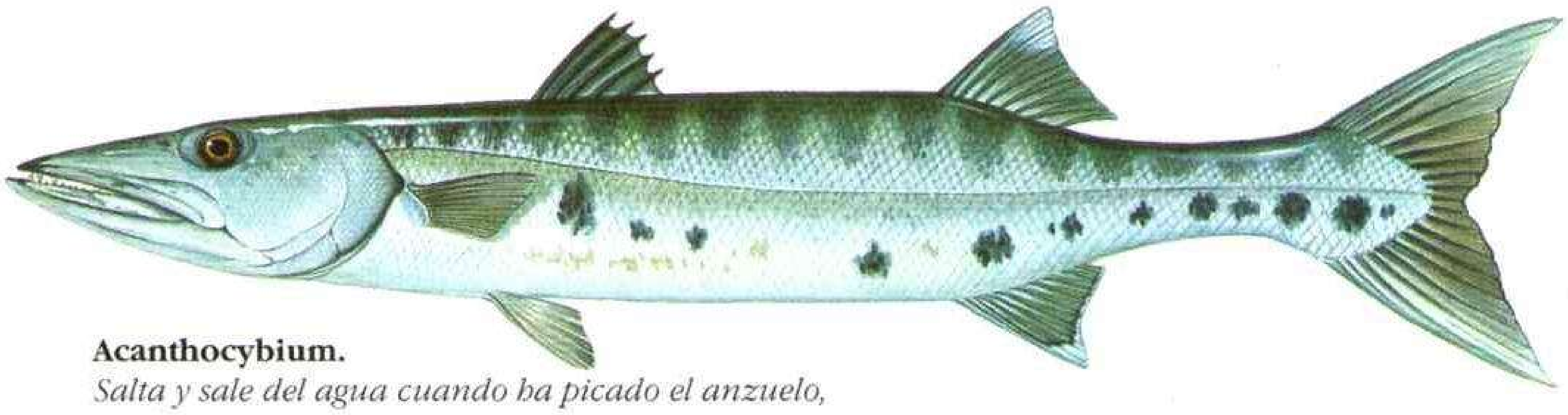

Acanthocybium.
Salta y sale del agua cuando ha picado el anzuelo, por lo que se hace difícil atraparlo. También tiene dientes muy afilados: puede ser peligroso incluso en tierra. Se encuentra en todo el mundo en las aguas cálidas y templadas.

OTROS PECES PELIGROSOS

El pez ballesta *(Balistidae)*, que vive en los mares poco profundos y tropicales, tiene un único aguijón en el abdomen, a semejanza de un gatillo. Su carne es venenosa por ingestión. Tanto el pez piedra *(Synanceia)*, como el pejesapo *(Batrachoididae)* y la escórpora o pez cebra *(Scorpaenidae)*, tienen aguijones venenosos. El peje araña *(Trachinidae)* de las costas europeas y del occidente de África se entierra en la arena, y sus aguijones pueden causar un dolor muy intenso.

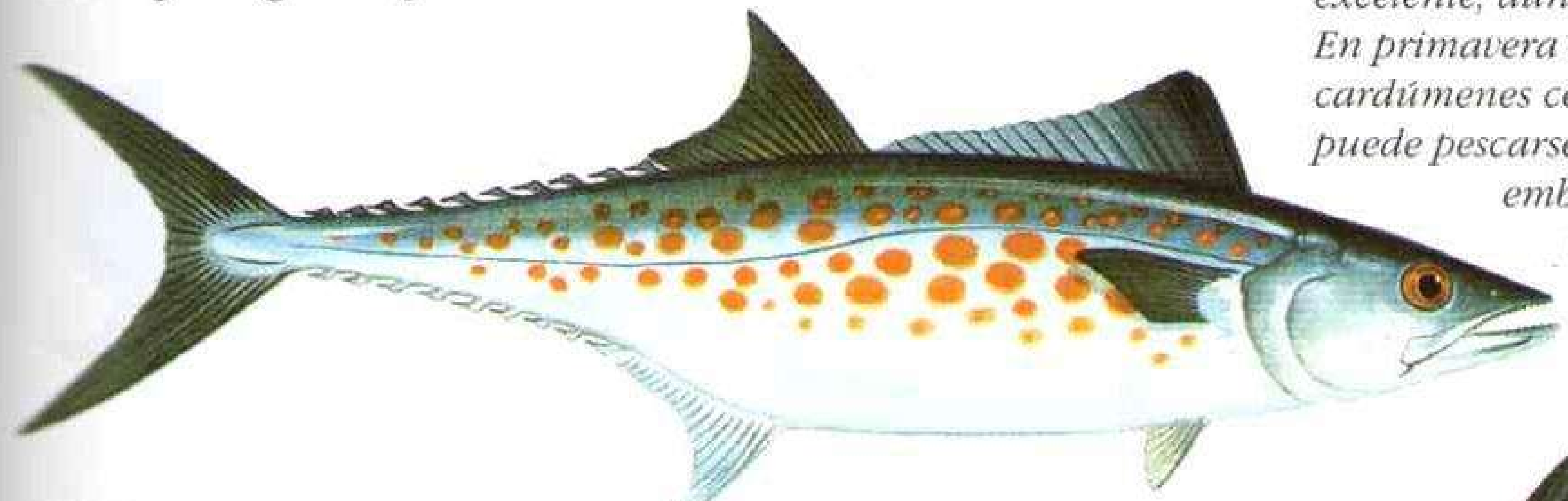

Caballa (Scomberomorus). *La caballa constituye un plato excelente, aunque un poco grasoso. En primavera se desplaza en grandes cardúmenes cerca de la costa, donde puede pescarse con redes. Sin embargo, es un nadador rápido, con dientes afilados, que ofrece gran resistencia s ser capturado. Se encuentra distribuida mundialmente.*

ADVERTENCIA

Cuando pesque con arpón, salga del agua tan pronto atrape su presa. El olor de la sangre y las vibraciones causadas por sus movimientos atraerán a los depredadores. Evite remover mucho el agua, ya que los depredadores interpretarán estas vibraciones como las de un pez herido.

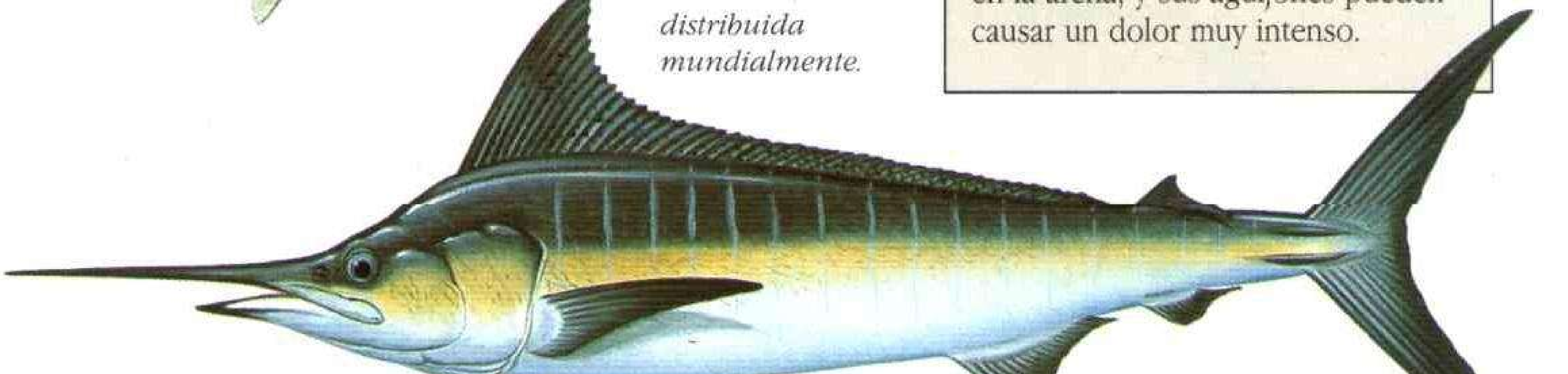

Pez vela (Makaira, Tetrapturus). *El pez vela se encuentra en las aguas cálido-templadas y tropicales. Aunque de sabor agradable, es difícil de pescar ya que es un nadador rápido y ofrece una gran resistencia una vez pescado. También tiene la mandíbula superior fuerte y afilada. La forma más adecuada de pescarlo es con curricán, en la superficie o en aguas medias.*

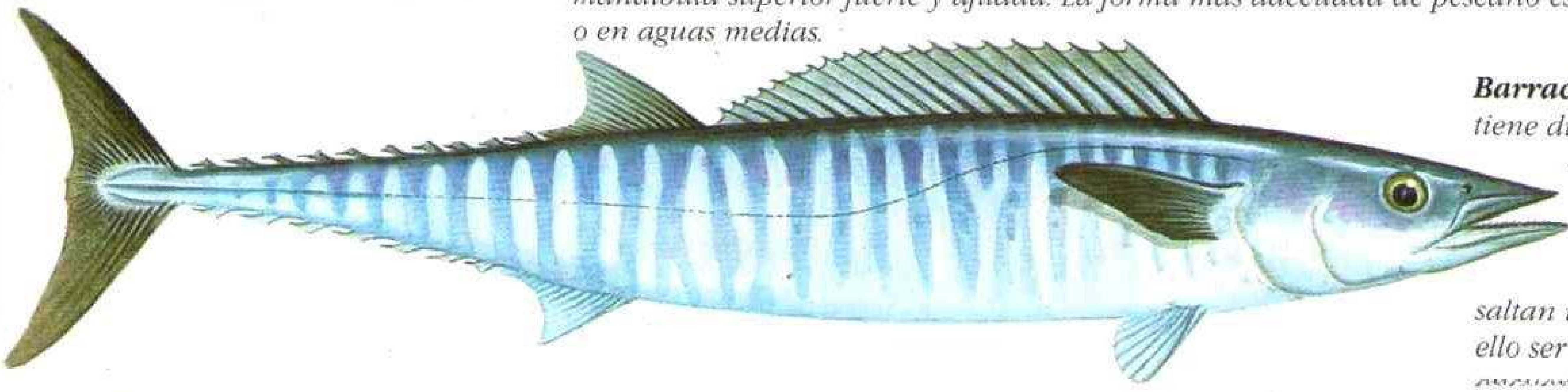

Barracuda (Sphyraena). *La barracuda tiene dientes muy afilados y puede llegar a atacar a los bañistas. Las especies pequeñas, aunque muerden, son comestibles, y pueden encontrarse en aguas poco profundas. Asimismo, saltan una vez han picado, y pueden por ello ser difíciles de atrapar. La barracuda se [illegible] tropicales de todo el mundo.*

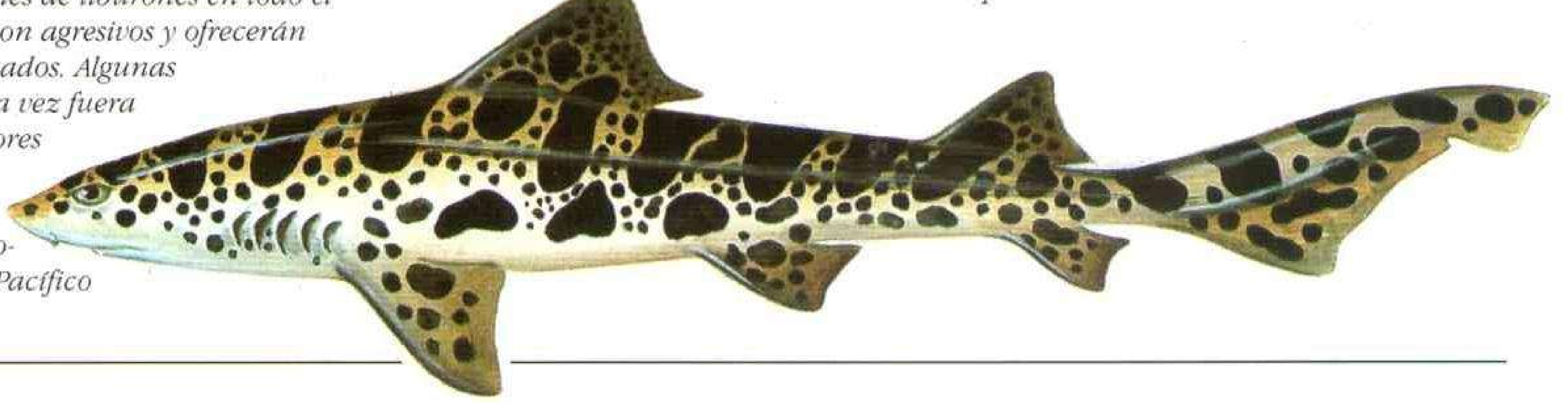

Tiburón. *Hay cientos de especies de tiburones en todo el mundo, muchos de los cuales son agresivos y ofrecerán una gran resistencia a ser pescados. Algunas especies «hacen el muerto» una vez fuera del agua, y cogen a los pescadores desprevenidos. El tiburón leopardo* (Triakis) *vive en aguas poco profundas e interiores a lo largo de las costas del Pacífico de Norteamérica.*

EQUIPO DE PESCA

LA PESCA ES, en gran medida, cuestión de tiempo y paciencia. Sin un equipo prefabricado y resistente, los peces pueden picar y escaparse. Los peces pueden pescarse con anzuelo, trampas, redes, arpones, o incluso puede capturarlos con las manos *(véase página 104)*. Los anzuelos, cebos y corchos pueden construirse a partir de cualquier material tanto natural como artificial. Fabrique un sedal a partir de hierbas naturales *(véase página 34)*, o bien utilice el hilo de pescar de su equipo de supervivencia *(véase página 28)*. Puede improvisar una caña a partir de una rama. Tenga cuidado de no romper su equipo improvisado cuando los peces piquen, y asegúrese de tener preparado algún método alternativo de pesca (redes) para cerciorarse de que no se le escapa ninguno.

ANZUELO DE ESPINAS

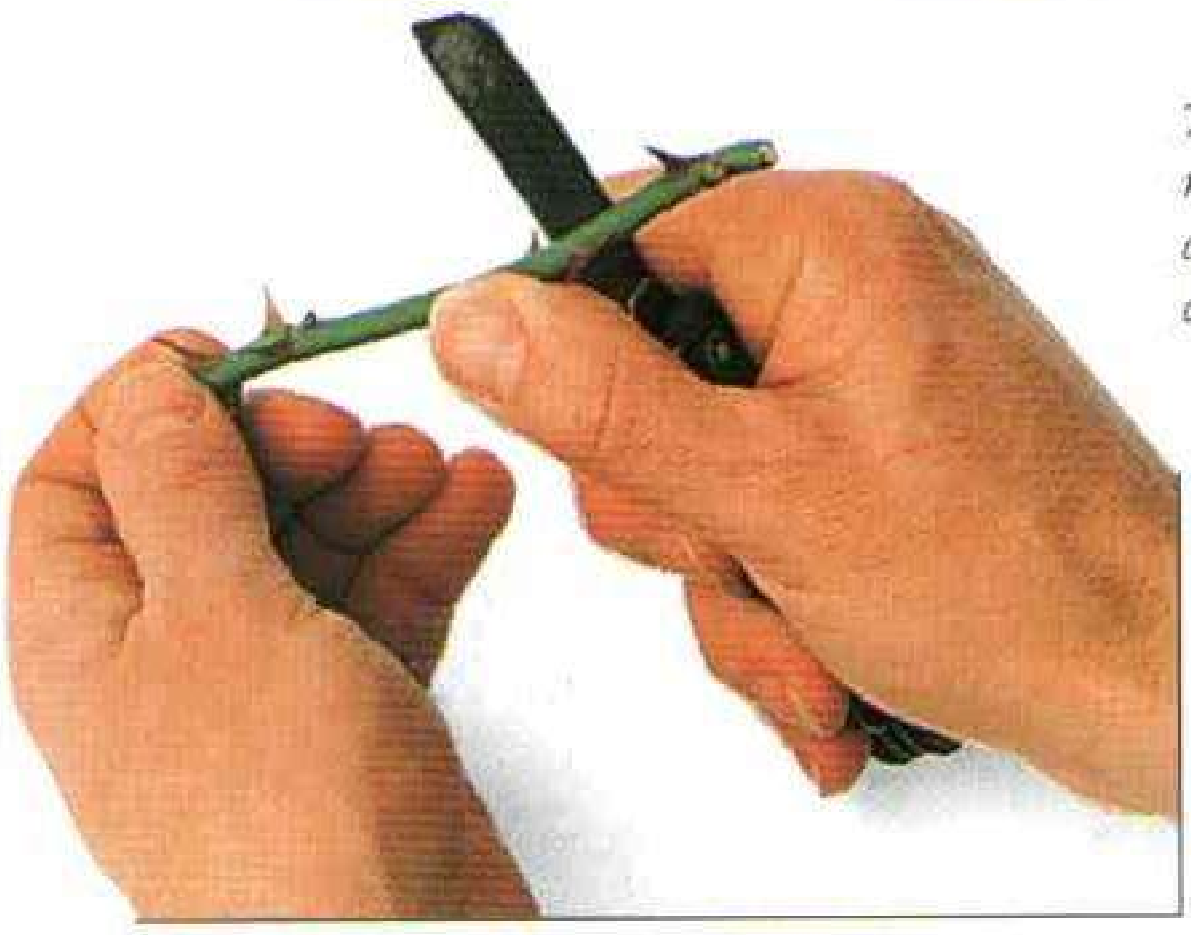

1 Corte una sección de un tallo de zarza, de unos 2,5 cm de longitud, que tenga una espina grande y fuerte.

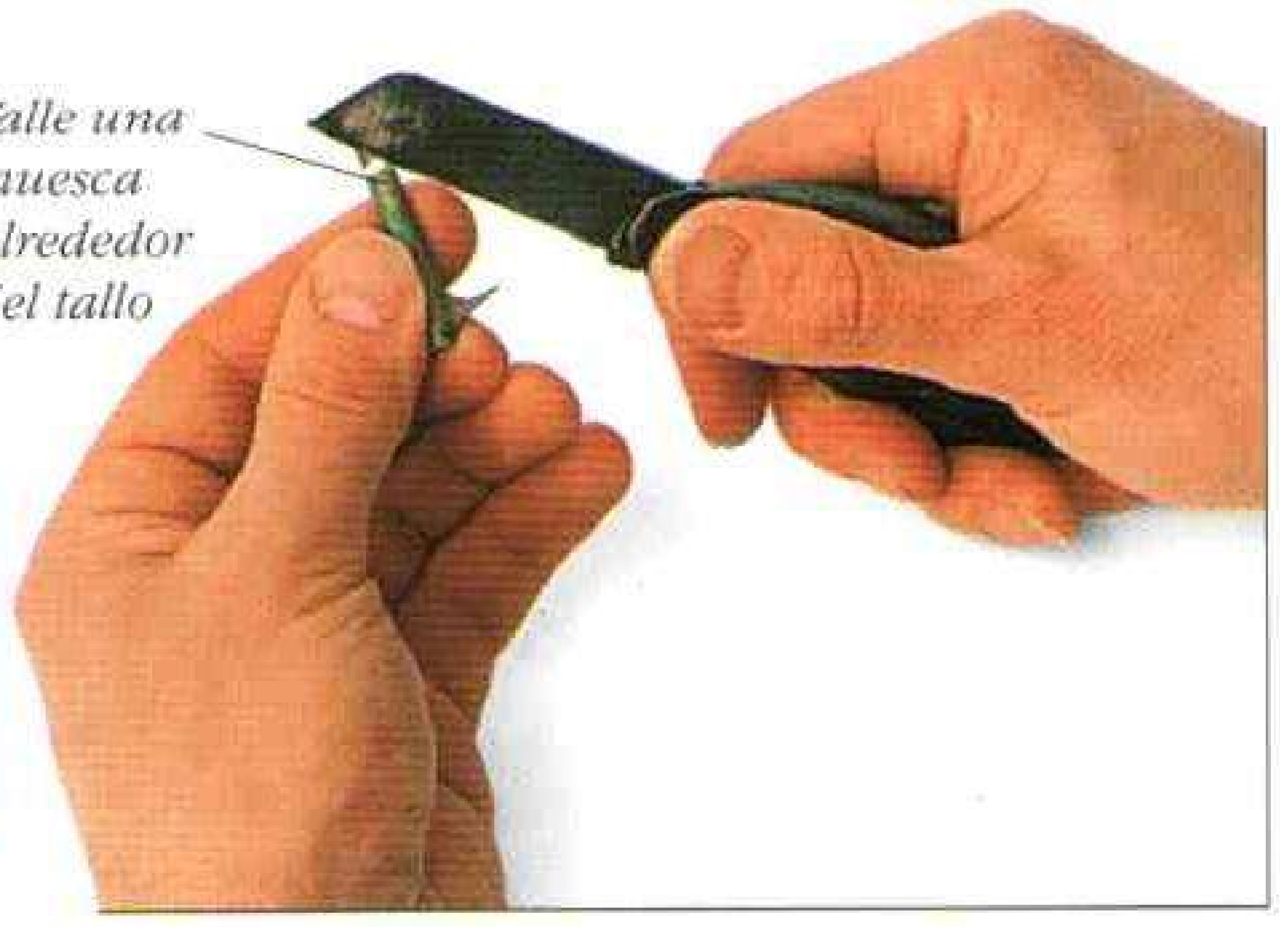

Talle una muesca alrededor del tallo

2 Con una navaja, talle una muesca en el extremo del tallo para evitar que el sedal resbale cuando lo ate alrededor del mismo.

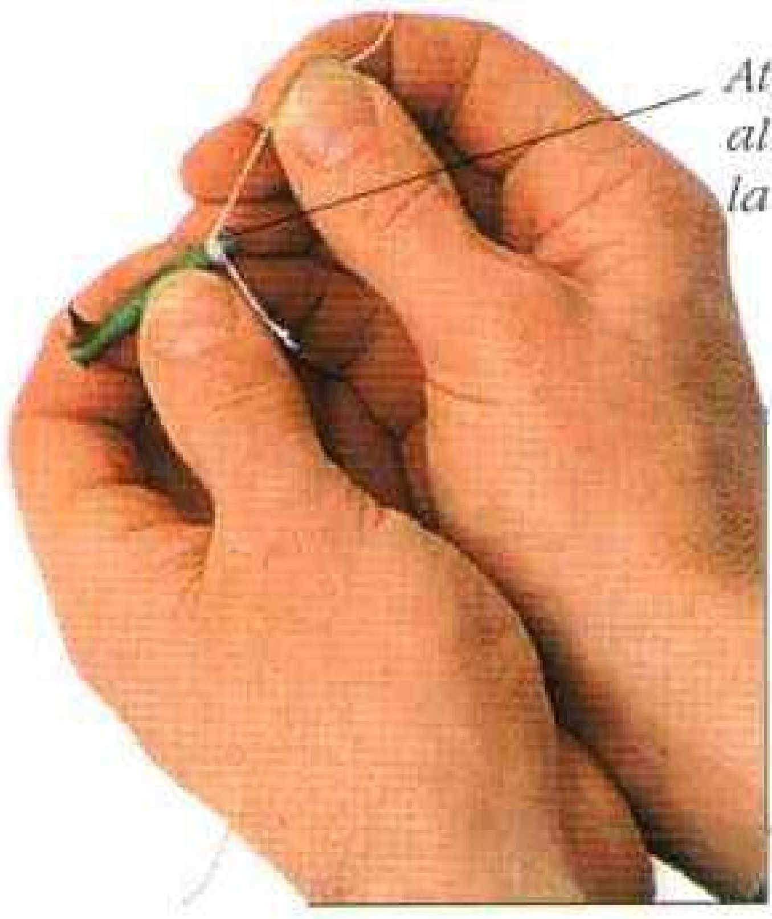

Ate el sedal alrededor de la muesca

3 Ate el sedal alrededor de la muesca en el tallo dando varias vueltas.

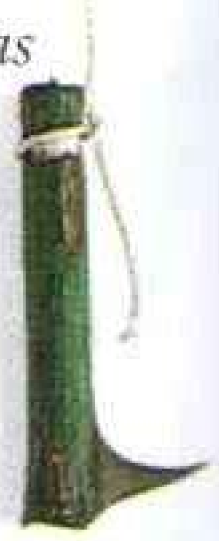

Anzuelo terminado. *Utilícelo para pescar peces que se traguen la carnada entera, como las anguilas y los bagres. El anzuelo también puede engancharse en las agallas o bocas de otros peces.*

UN CLAVO COMO ANZUELO Y UNA PLUMA COMO CEBO

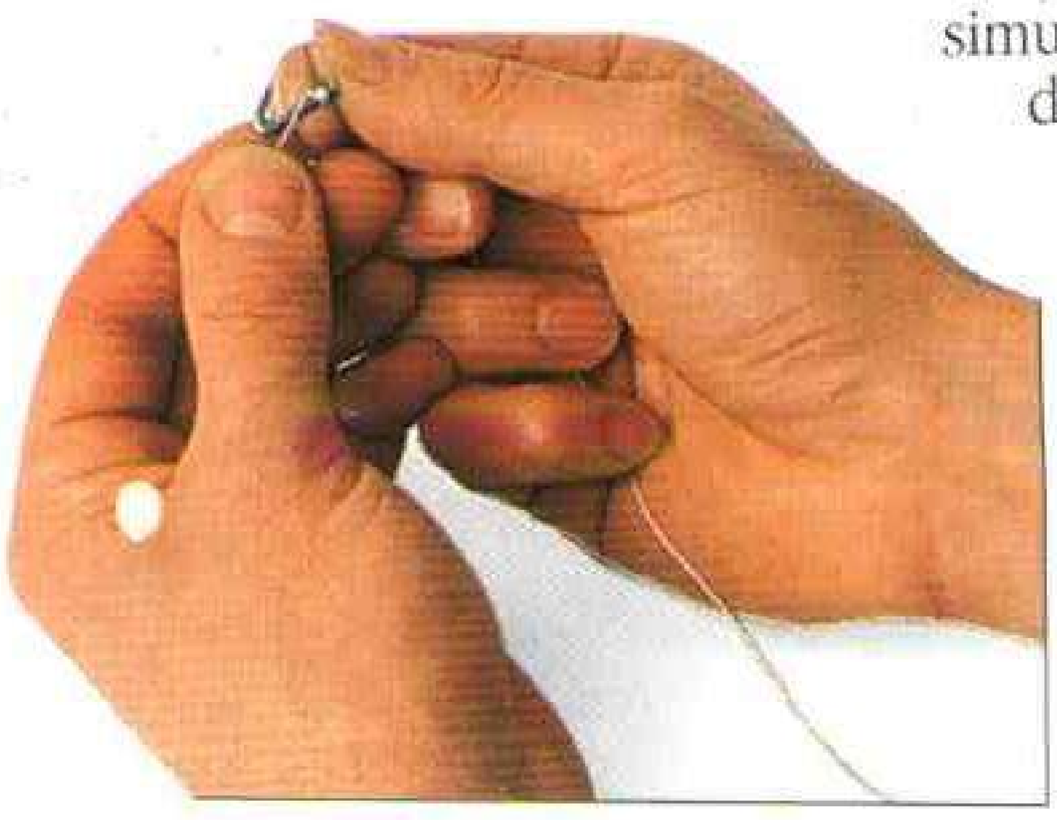

1 El objetivo de un cebo es simular la carnada, a fin de atraer al pez y para que muerda y se trague el anzuelo. Estudie lo que comen los peces en su hábitat y construya un cebo adecuado. Para construir un anzuelo simple, doble un clavo y átele el sedal.

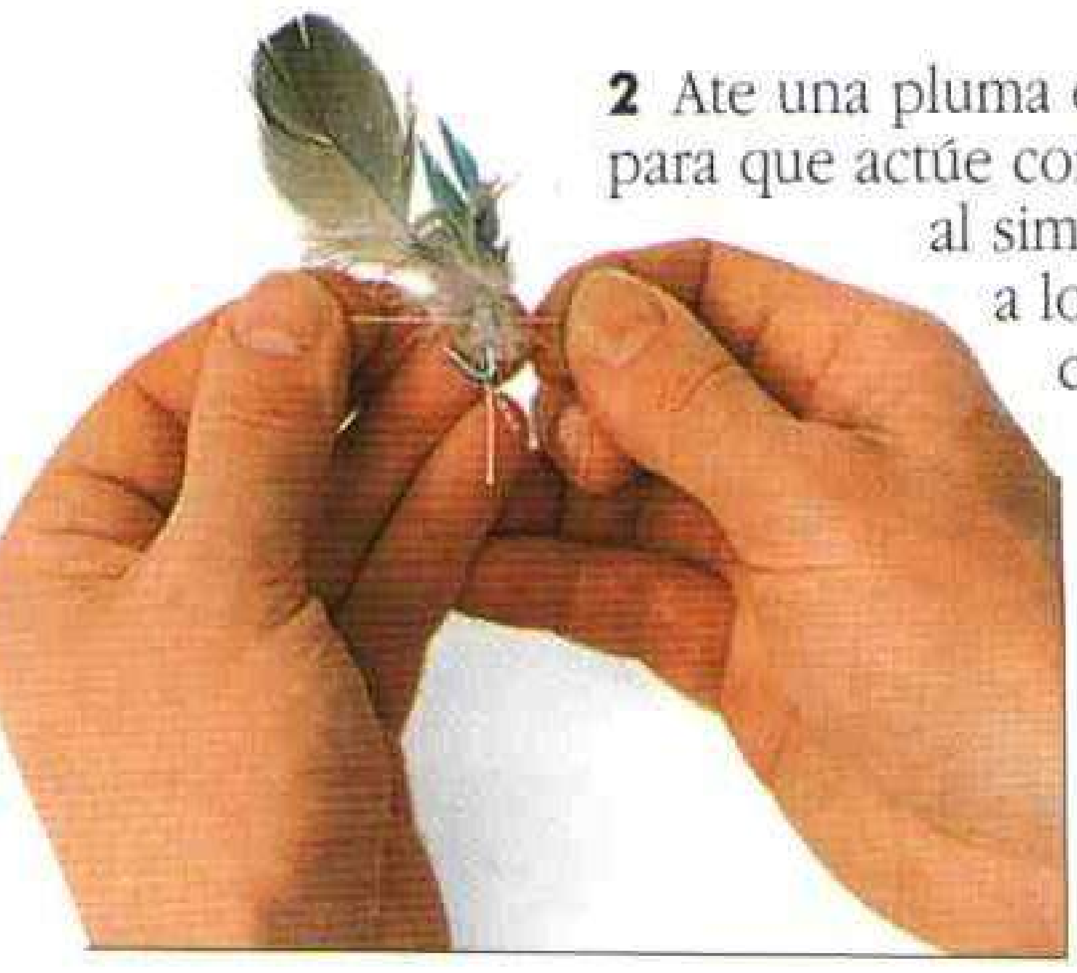

2 Ate una pluma encima del anzuelo para que actúe como señuelo. La pluma, al simular un insecto, atraerá a los peces. Roce la superficie del río o corriente con el cebo y después sumerja repetidamente el anzuelo en el agua a fin de atraer a los peces para que muerdan el anzuelo.

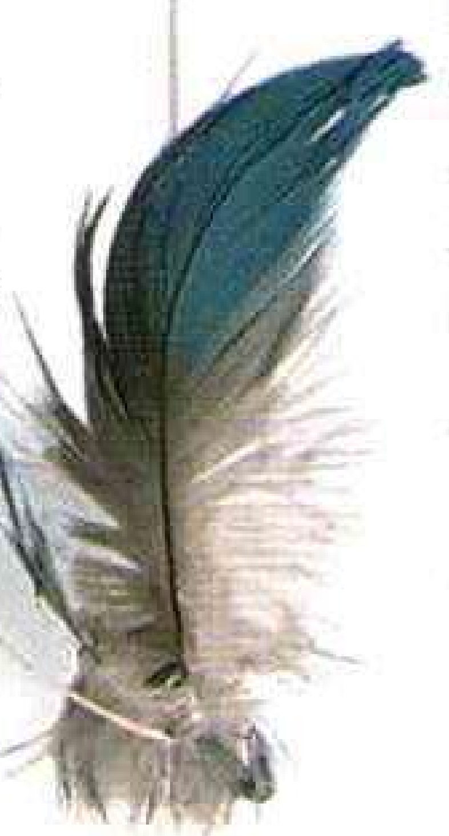

Cebo terminado. *Asegúrese de que el anzuelo está bien sujeto al sedal para no perderlo cuando pique el pez.*

UN ESCARAMUJO COMO CORCHO Y UN IMPERDIBLE COMO ANZUELO

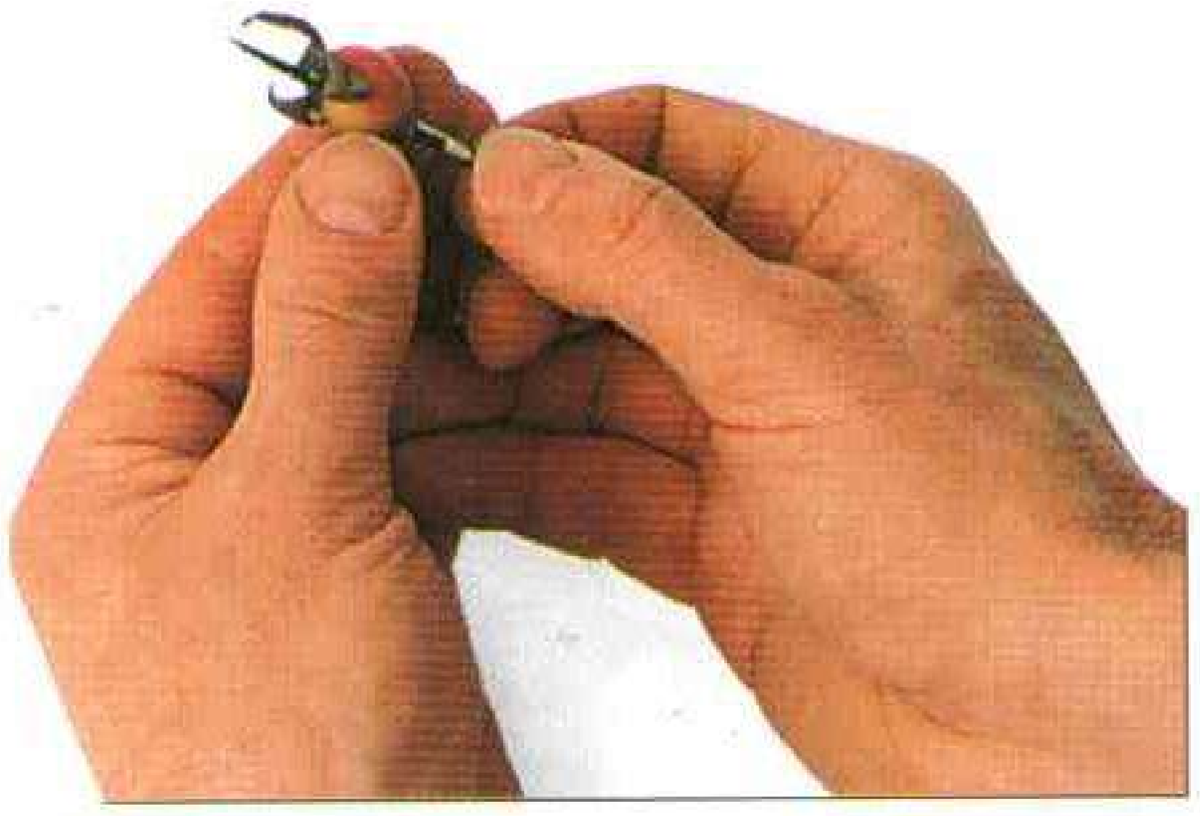

1 Puede emplear materiales naturales para construir un corcho. Si utiliza un escaramujo, agujeréelo con un cuchillo, un palo afilado, o una espina.

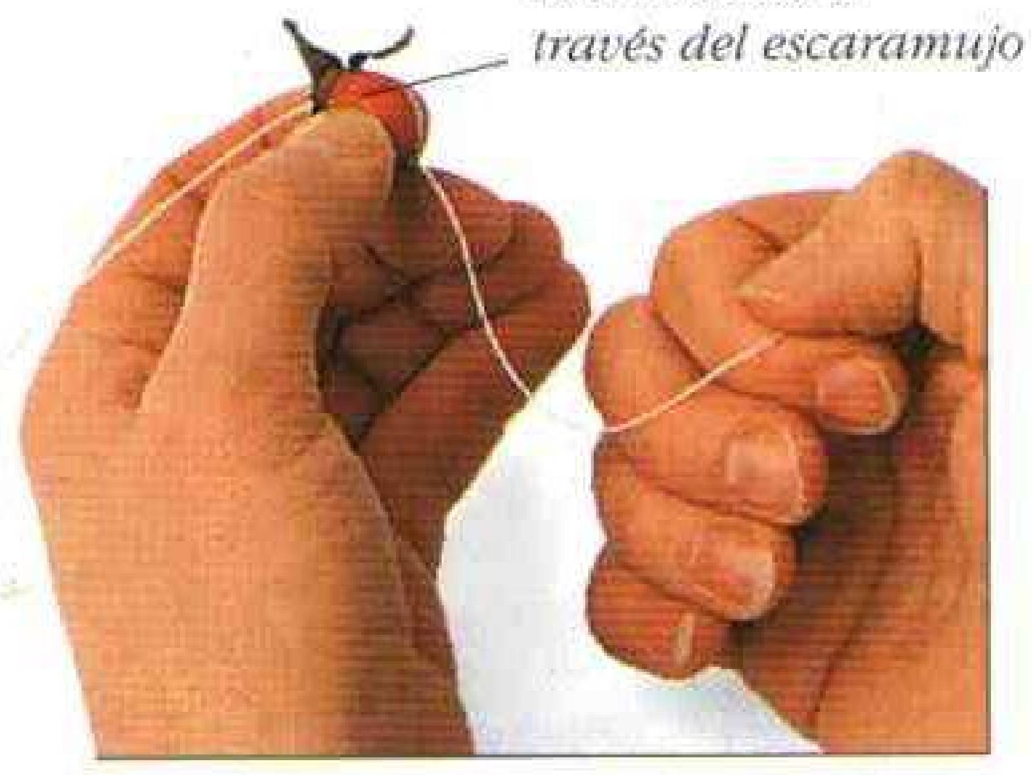

Pase el sedal a través del escaramujo

2 Pase el hilo de pescar o un cordel a través del escaramujo. La función del corcho consiste en mantener el anzuelo a la profundidad adecuada para el pez que se pretenda pescar.

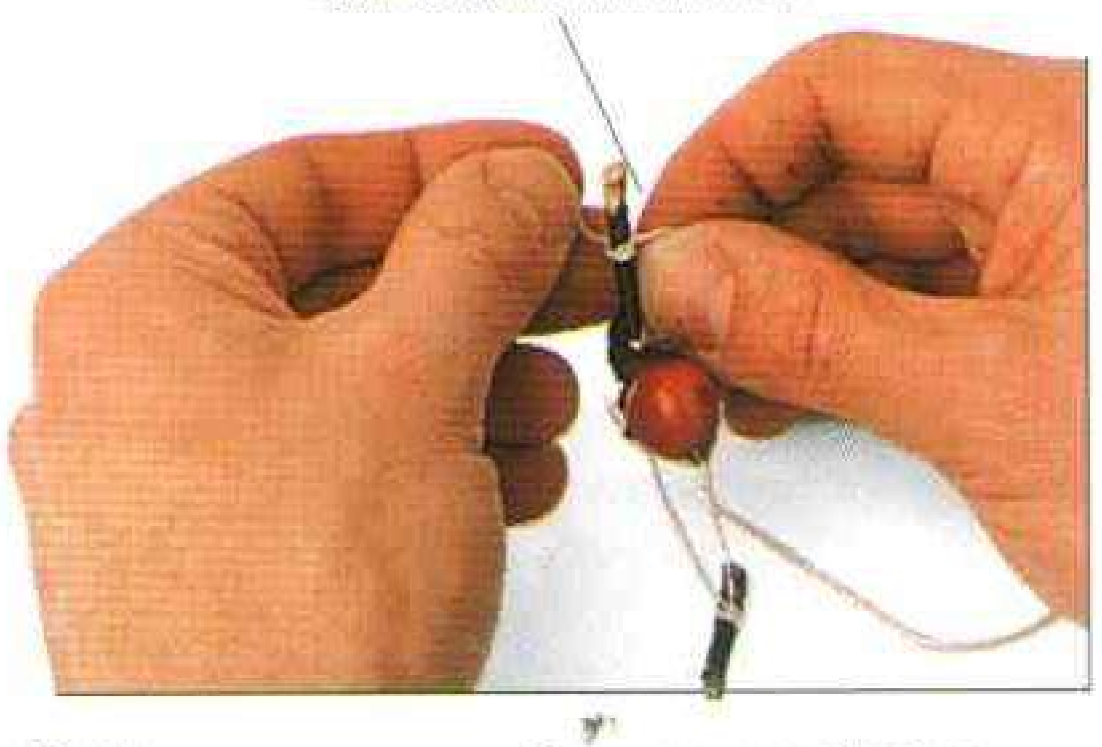

Las ramitas funcionan como estabilizadores

3 Ate ramas pequeñas por encima y por debajo del corcho para que actúen como estabilizadores y para evitar que el corcho se deslice a lo largo del sedal.

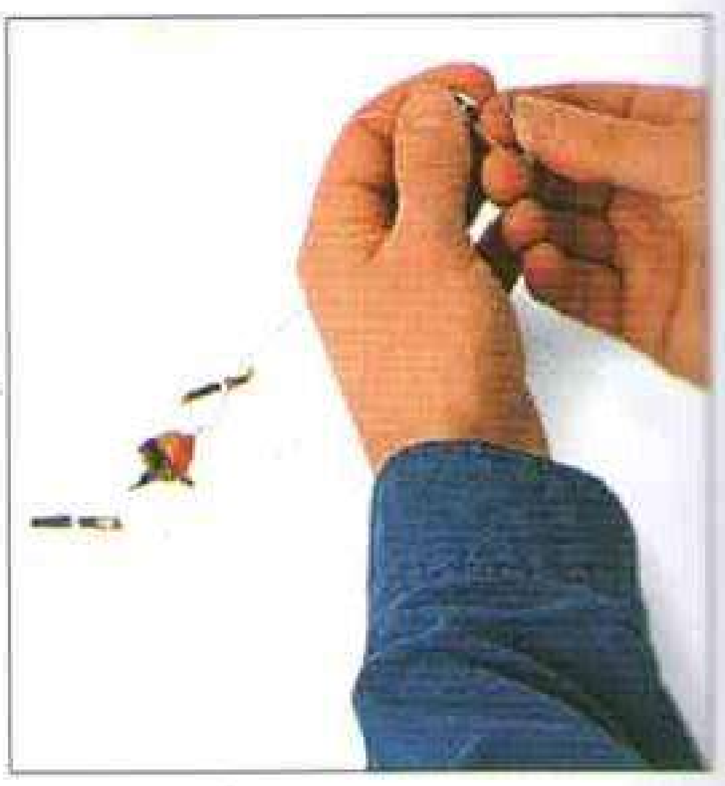

4 Ate el anzuelo (en este caso un imperdible) por debajo del corcho a la profundidad deseada.

CEBO

El cebo puede extenderse en la superficie del agua, sujetarse a un anzuelo o colocarse dentro de una trampa. Los insectos, los frutos pequeños, el pan, la carne cruda y los trozos de pescado son buenos cebos. La carnada viva atrae a los peces con sus movimientos, ya que indican que está herida y que es una comida fácil. El secreto de una carnada adecuada está en tentar al pez con su dieta habitual.

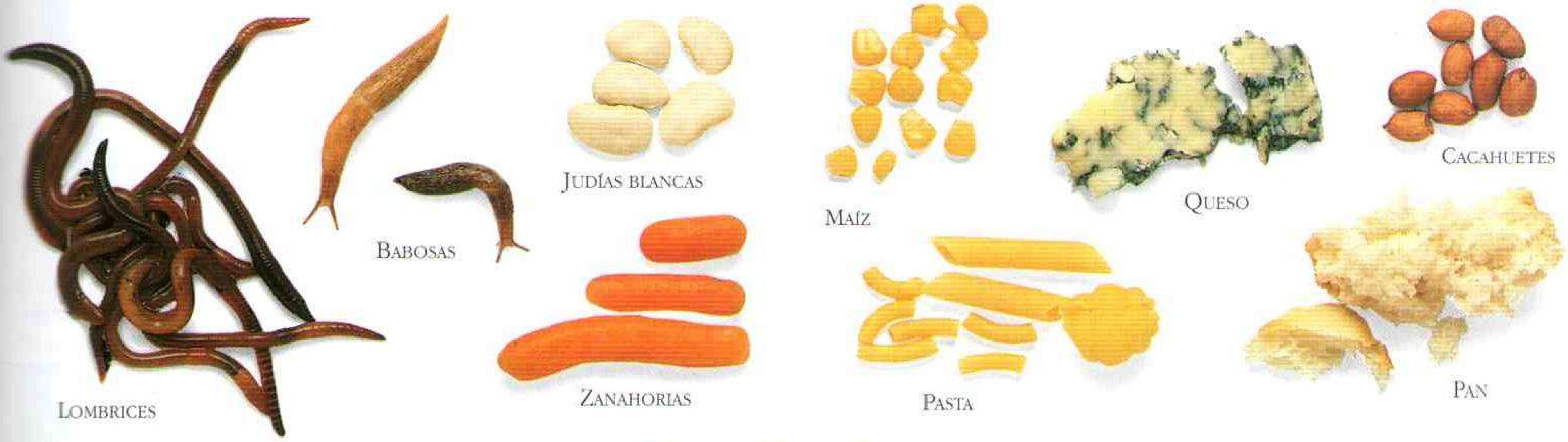

ARPÓN CON TRAMPA

Corte los vástagos de la rama para construir el arpón

1 Este tipo de arpón se emplea principalmente en aguas profundas *(véase página 104)*. En primer lugar, corte los vástagos laterales para obtener una rama recta y esbelta.

Afile el arpón hasta conseguir una punta larga y afilada

2 Afile el extremo más grueso de la rama hasta lograr una punta afilada y larga. Puede endurecer la punta pasándola cuidadosamente entre las llamas de un fuego. Talle los laterales de la punta para formar una superficie lisa a la que atará la trampa utilizando una cuerda.

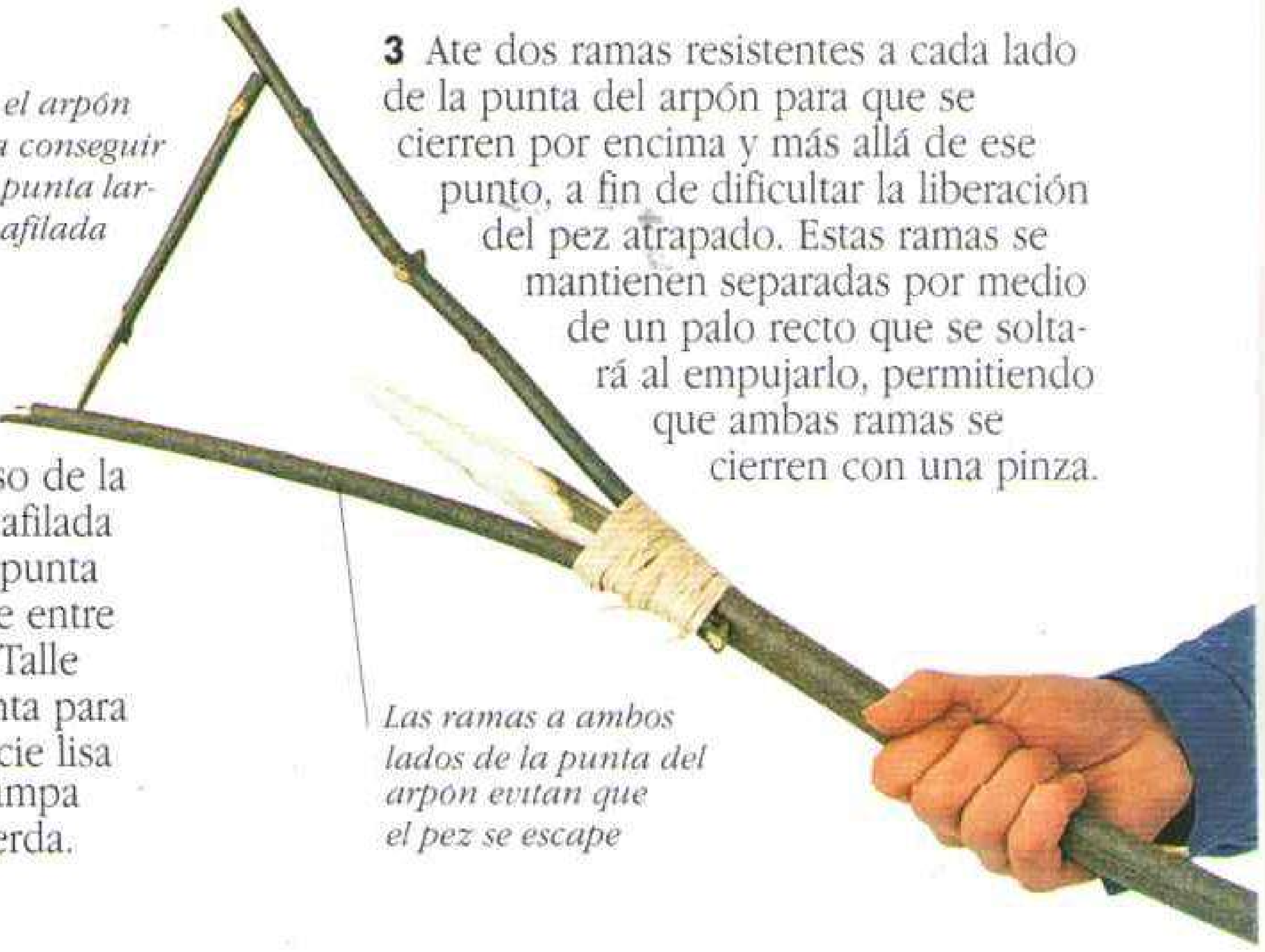

3 Ate dos ramas resistentes a cada lado de la punta del arpón para que se cierren por encima y más allá de ese punto, a fin de dificultar la liberación del pez atrapado. Estas ramas se mantienen separadas por medio de un palo recto que se soltará al empujarlo, permitiendo que ambas ramas se cierren con una pinza.

Las ramas a ambos lados de la punta del arpón evitan que el pez se escape

ARPÓN CLAVETEADO

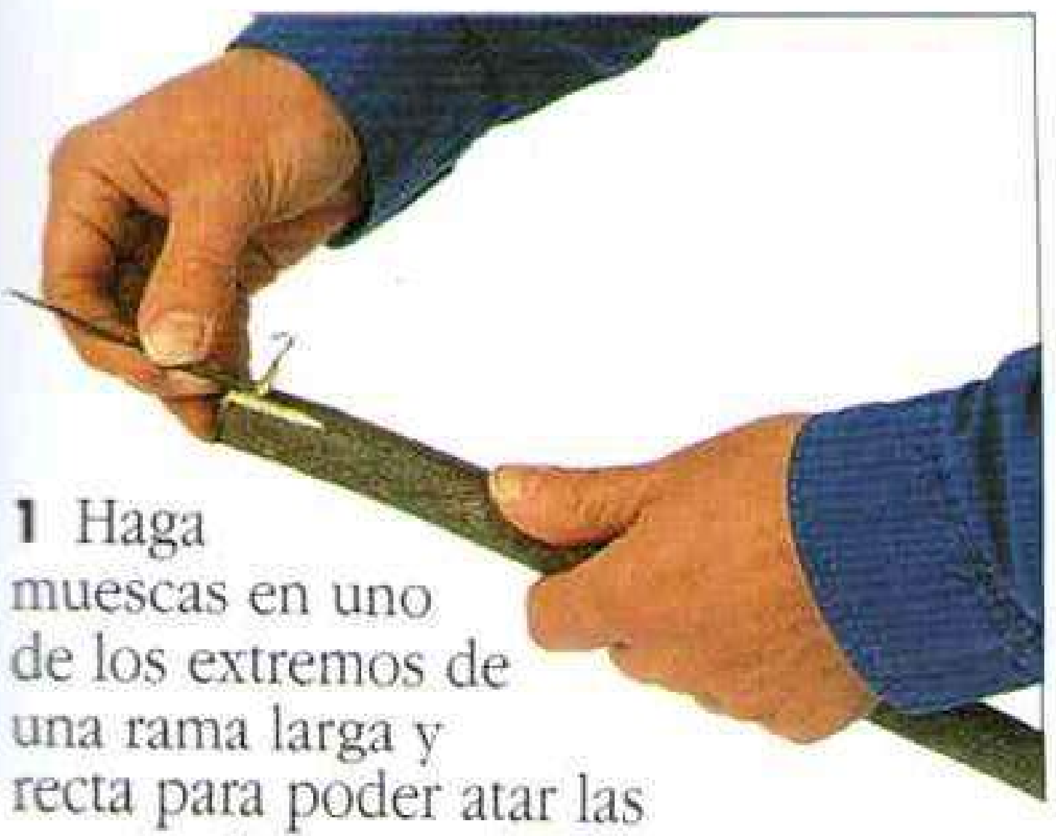

1 Haga muescas en uno de los extremos de una rama larga y recta para poder atar las puntas al arpón.

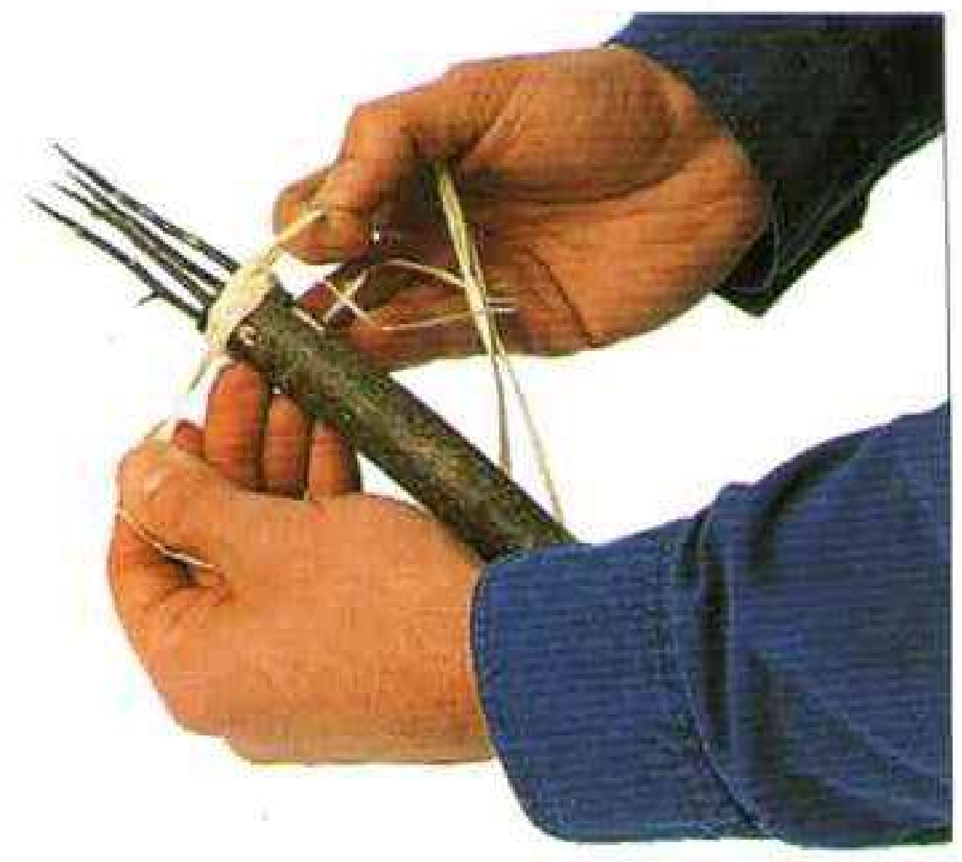

2 Ate firmemente pequeñas ramas espinosas alrededor del astil. Recuerde que tendrán que soportar parte del peso del pescado.

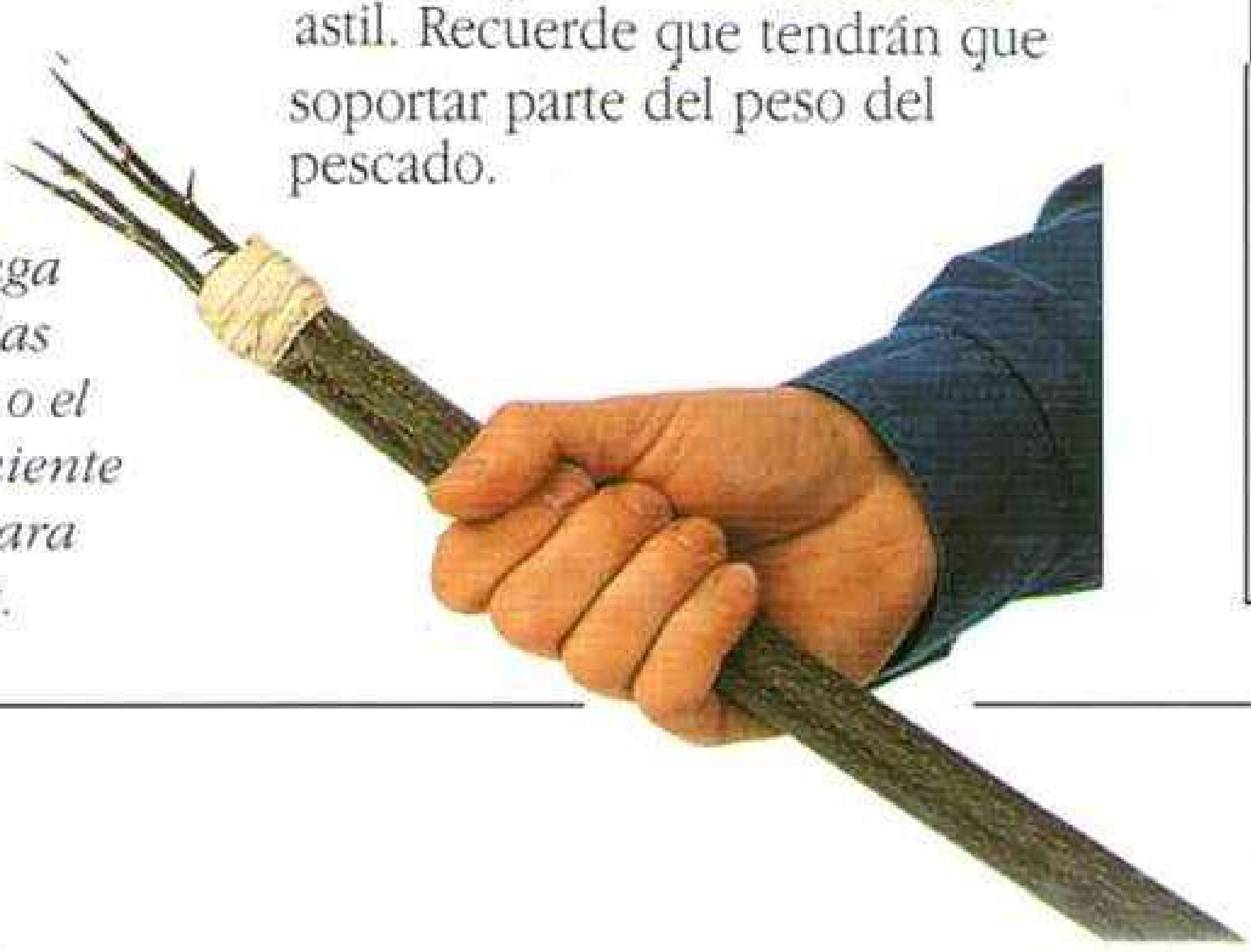

Arpón terminado. *Tenga cuidado de no romper las puntas contra las rocas o el fondo del río. Es conveniente utilizar un salabardo para sacar los peces del agua.*

TOME PRECAUCIONES AL PESCAR

Tenga mucho cuidado al utilizar equipos de pesca improvisados, ya que pueden romperse con facilidad y hacerle perder la presa, o el equilibrio. Si cae al agua, además del riesgo de ahogarse, puede contraer neumonía o alguna infección transmitida por el agua *(véase página 75)*. No entre en un río turbio sin una rama como soporte que a su vez le sirva para tantear el terreno. Recuerde que otros animales —entre ellos los depredadores— pueden encontrarse pescando o bebiendo en las inmediaciones

A la caza del pescador. *Busque las huellas de los animales que se dirigen a beber, y asegúrese de no convertirse en el cebo para un depredador.*

TÉCNICAS DE PESCA

Para un pescador deportivo moderno, estas técnicas tradicionales pueden parecer muy rudimentarias. Sin embargo, el éxito del pescador y el suyo propio dependen, no del equipo, sino del conocimiento de la presa. Tómese el tiempo necesario para observar dónde van los peces durante las distintas horas del día, cuándo y dónde se alimentan, y qué comen. A continuación, diseñe la estrategia para pescarlos. Cuando utilice un equipo de pesca improvisado, tenga cuidado de no romperlo o perderlo, y piense siempre primero en su propia seguridad. Las técnicas descritas aquí pueden utilizarse en ríos y en corrientes de agua dulce o salada. Algunas de estas técnicas son ilegales en muchos lugares del mundo, y únicamente debe emplearlas si se encuentra en una situación de vida o muerte.

EL EMPLEO DE ANZUELOS Y SEDALES

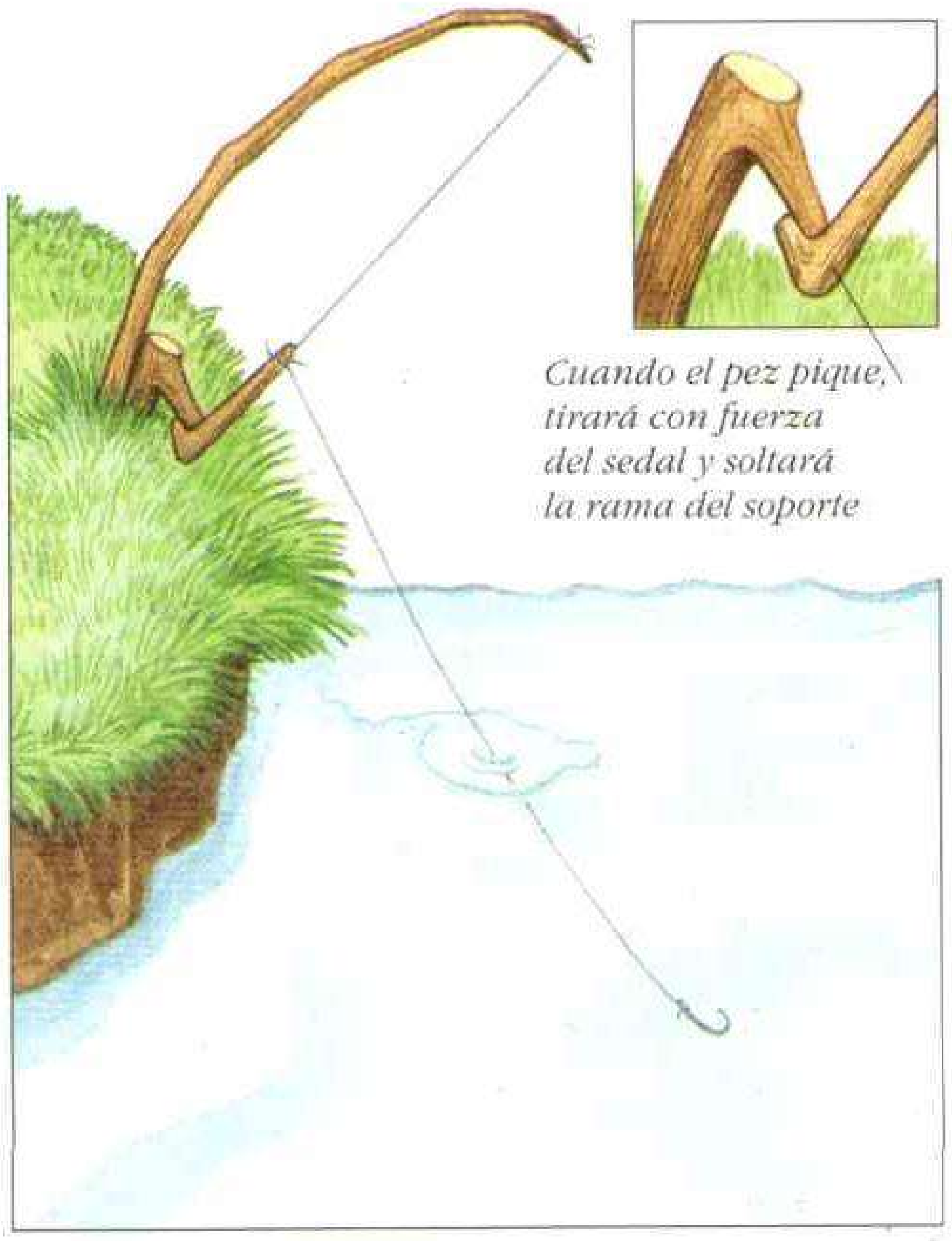

Cuando el pez pique, tirará con fuerza del sedal y soltará la rama del soporte

Sedal autónomo. *Un sedal de este tipo funcionará solo. Cuando el pez pique, el sedal se verá desplazado del soporte que lo sujeta y la caña volverá a extenderse; el anzuelo se clavará entonces en la boca del pez.*

Sedal múltiple. *Cuantos más anzuelos tenga en el agua, mayores serán sus posibilidades de pescar. Con este método, los peces picarán más rápidamente que con un solo anzuelo. Utilice este método únicamente en una situación de supervivencia, ya sea en agua dulce o en una ensenada de agua salada.*

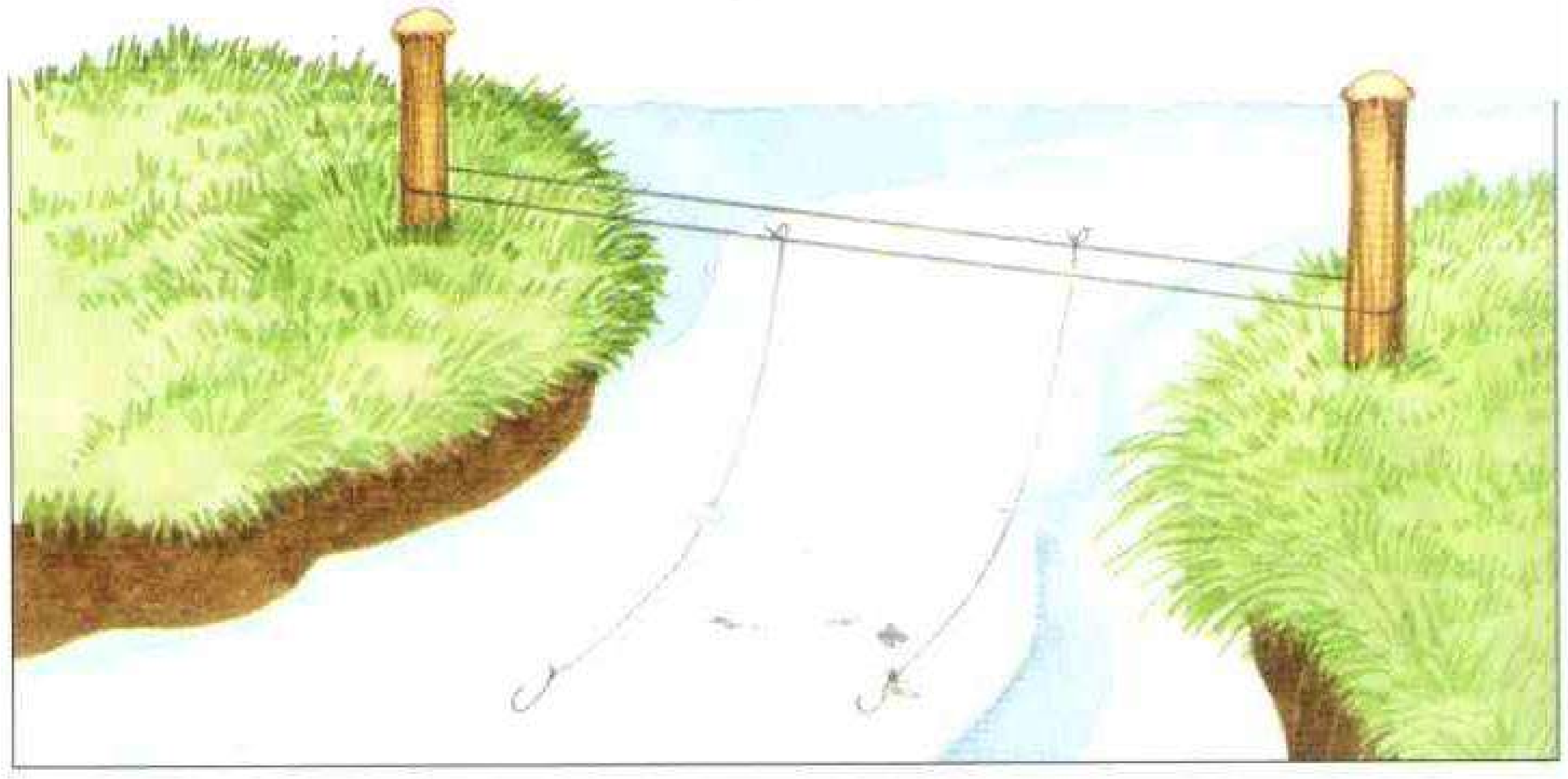

PESCA EN EL HIELO

Con una sierra o cuchillo para hielo, corte un pequeño agujero en el hielo, cerca de la orilla, a fin de comprobar que tenga como mínimo 5 cm de espesor y para poder soportar su peso. Corte un agujero de unos 30 cm de diámetro. Procure no romper el hielo porque podría ceder y provocar su inmediata caída al agua. Ate el sedal a dos palos en forma de cruz a los que previamente habrá atado una bandera.

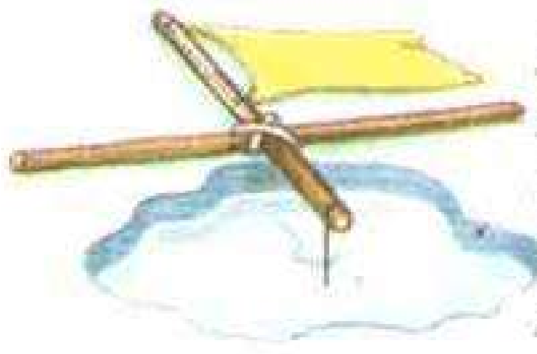

Anzuelo con cebo. *Ate el sedal a una bandera para evitar tener que sacarlo para comprobar si algún pez ha mordido el cebo.*

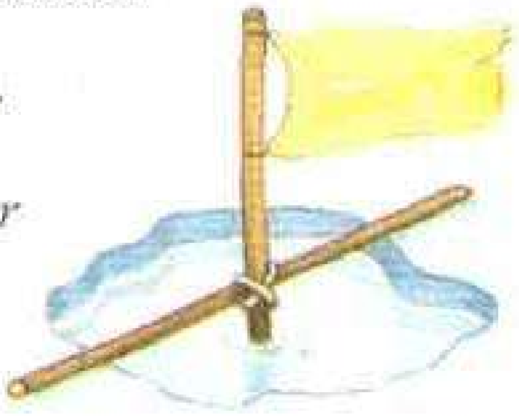

El pez ha picado. *Si la bandera está en posición vertical significa que algún pez ha picado.*

PESCA MANUAL

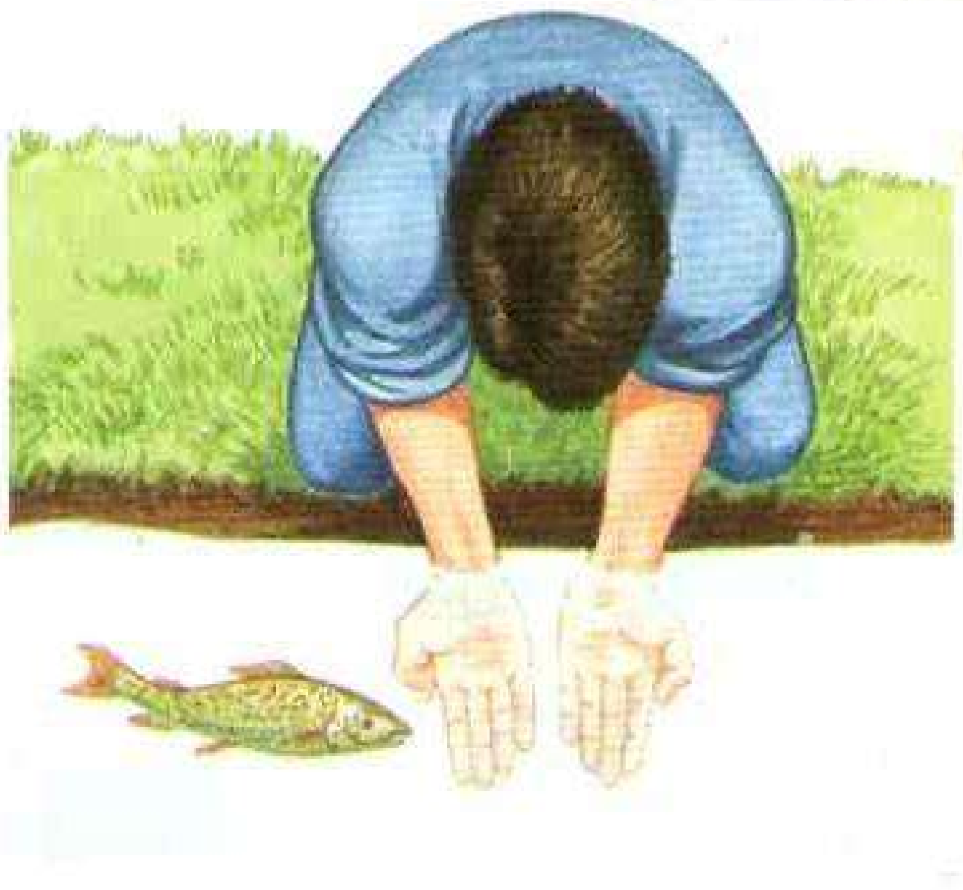

1 Este método es ilegal en la mayoría de los países. Es más adecuado para las aguas resguardadas. Tantee a lo largo de la orilla con las palmas de las manos extendidas hacia arriba.

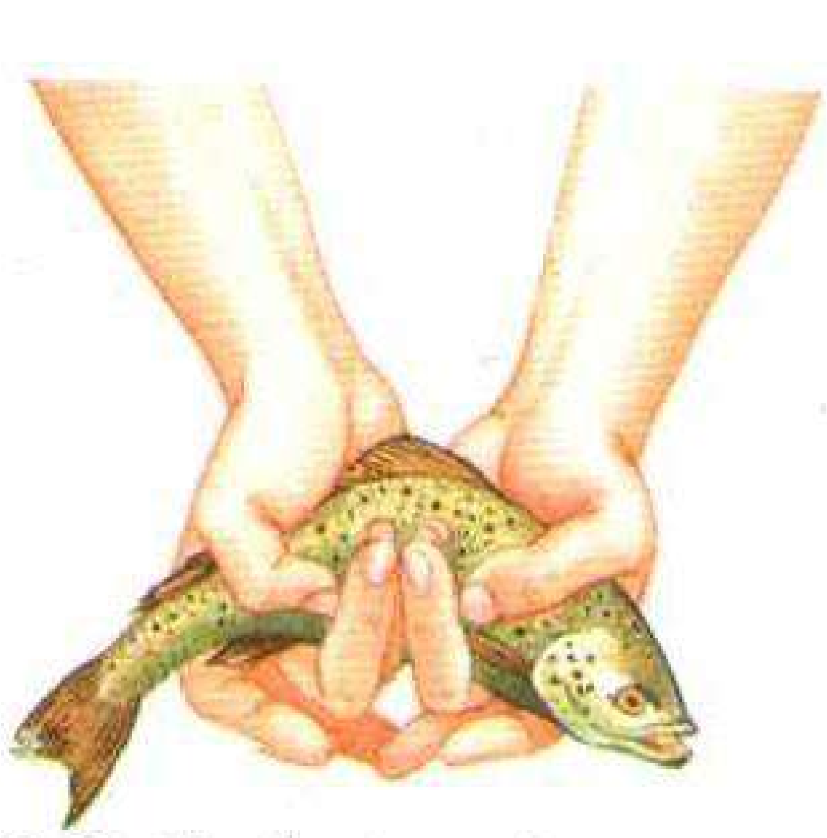

2 Deslice lentamente sus manos por debajo de un pez que permanezca quieto. Coja el pez rápidamente y láncelo a tierra.

PESCA CON ARPÓN

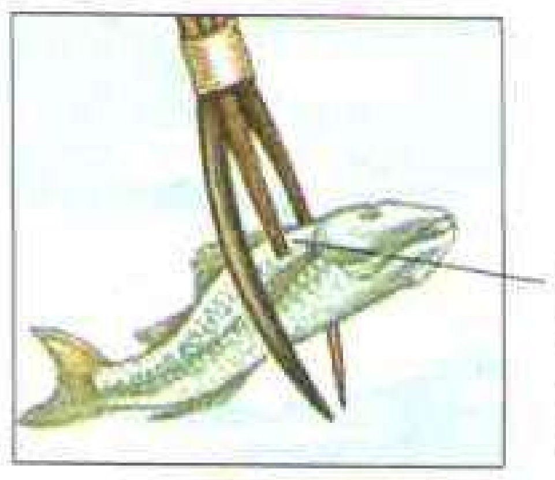

Arpón con trampa. *Este tipo de arpón atrapa al pez en su interior, evitando que se escape.*

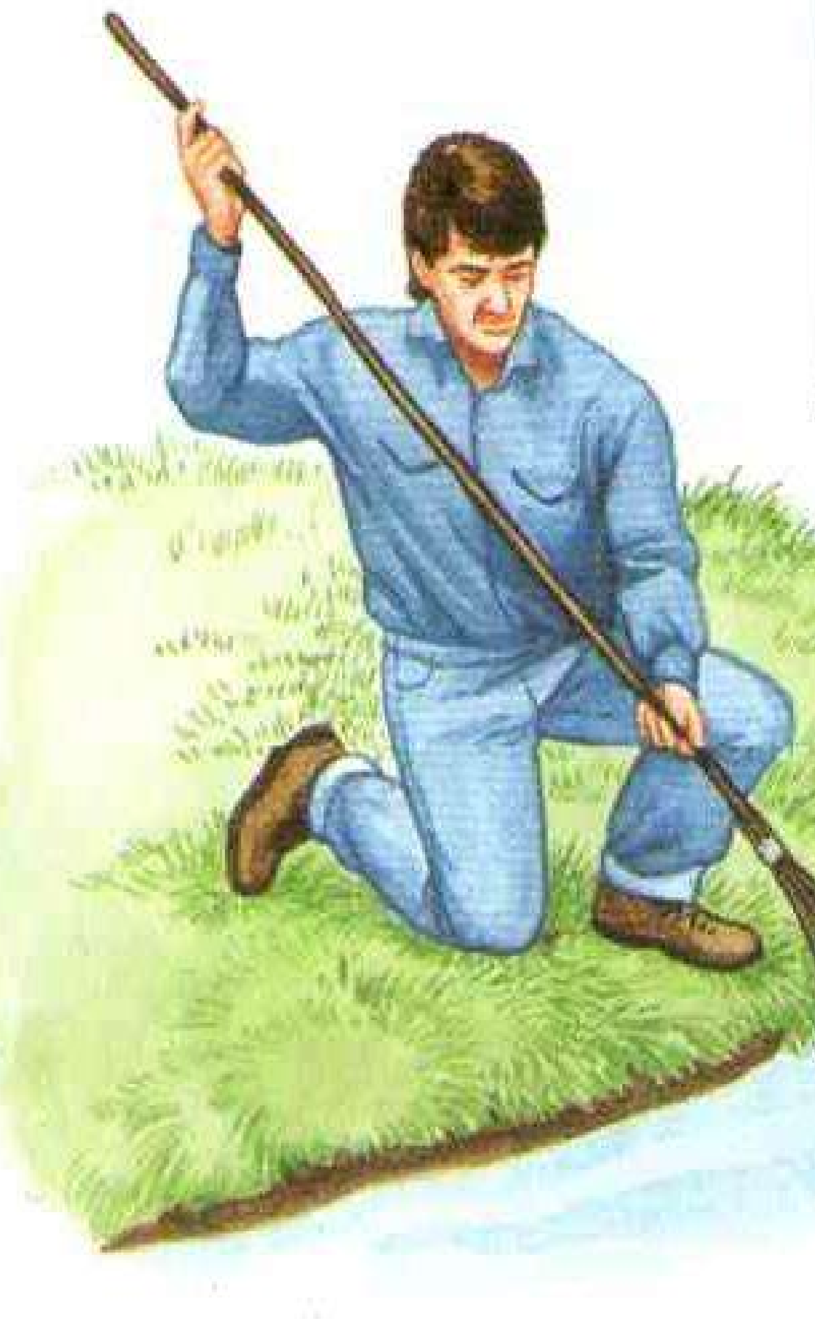

El empleo del arpón. *La pesca con arpón requiere tiempo y mucha paciencia. Tenga cuidado de no romper sus arpones contra las rocas o el fondo del río. Si mantiene la punta del arpón en el agua, evitará las salpicaduras. Aseste el golpe rápidamente una vez haya divisado la posible presa.*

CÓMO SACAR LOS PECES DEL AGUA

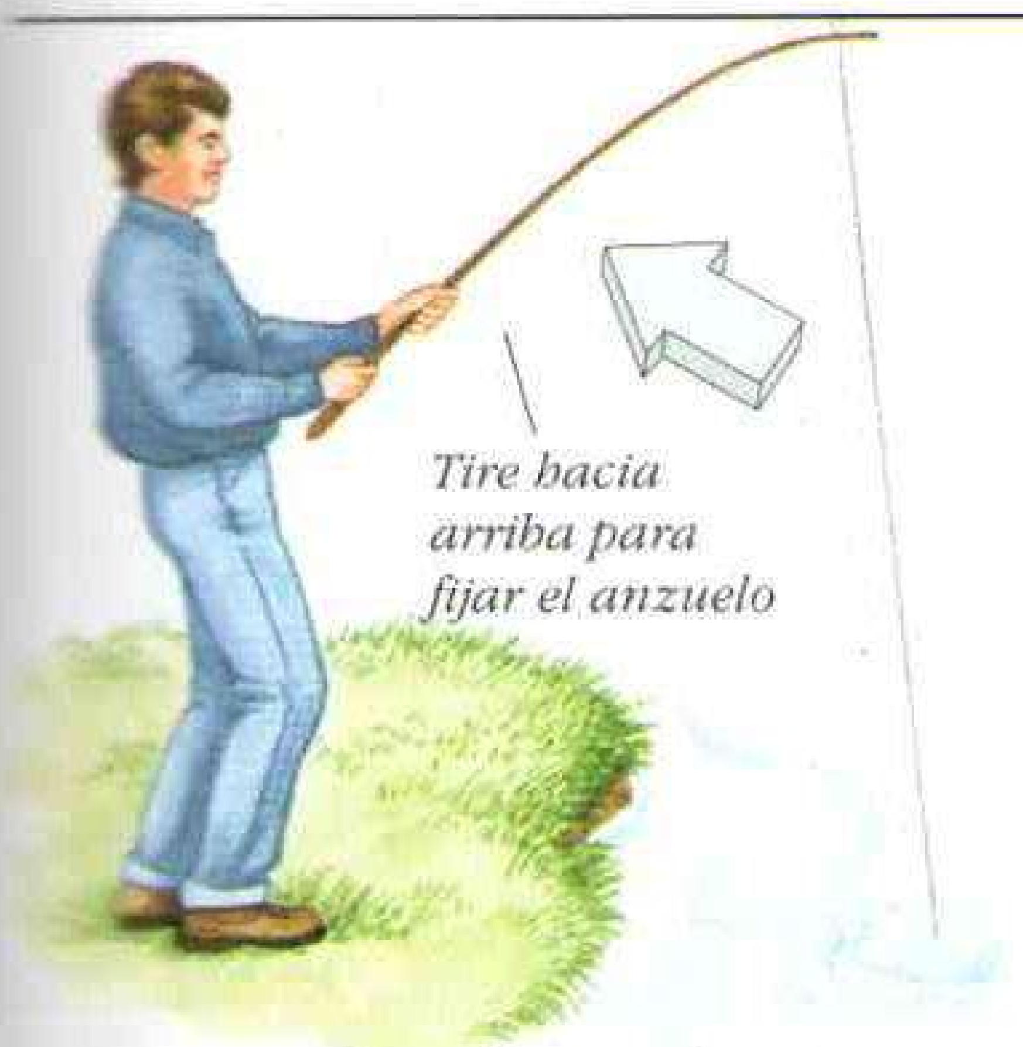

Tire hacia arriba para fijar el anzuelo

1 Una vez el pez ha picado, tire con fuerza de la caña hacia arriba. Tenga cuidado de no romper su equipo, en especial al acercar el pez a la orilla.

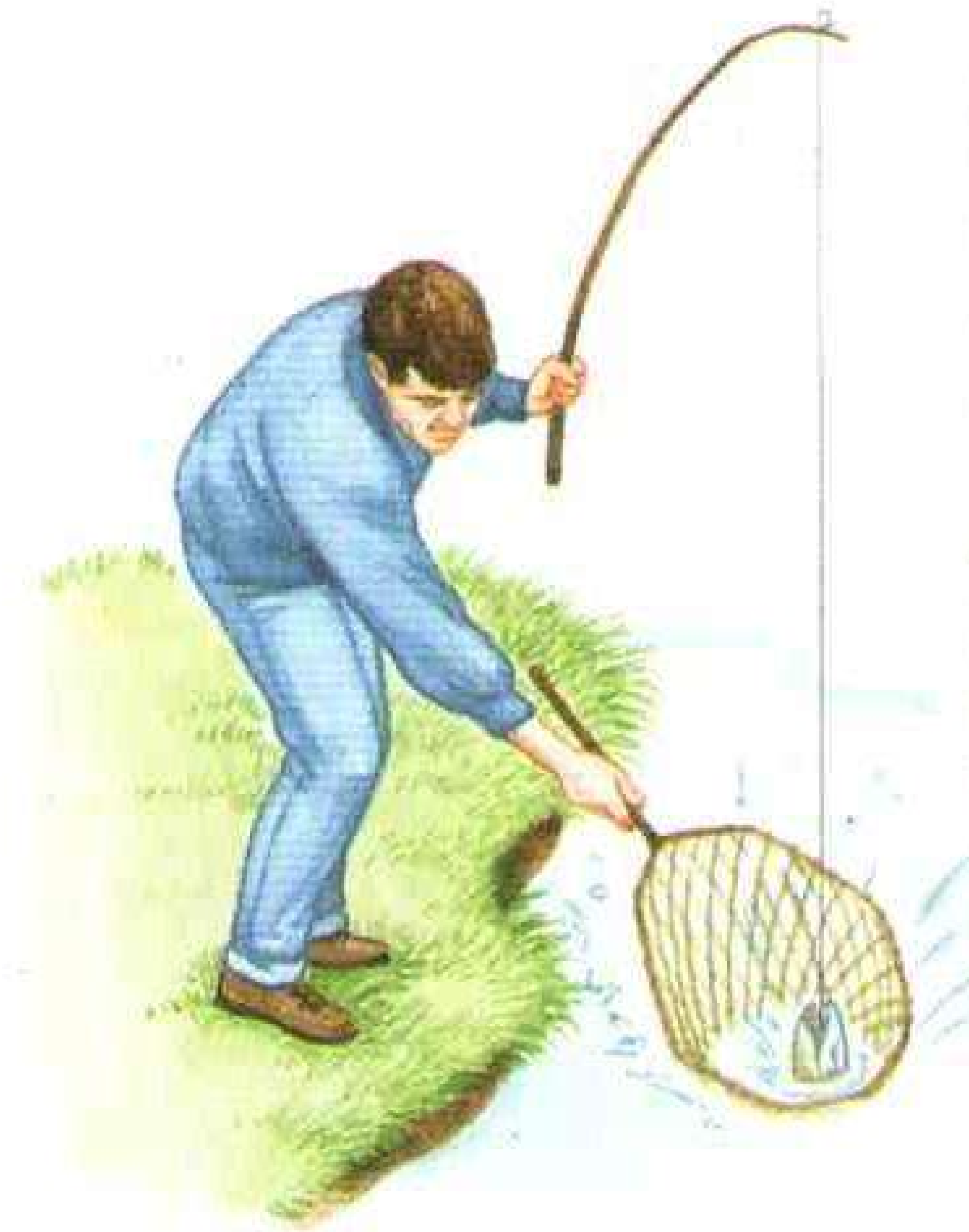

2 Proceda lentamente a sacar el pez del agua. Utilice la red tan pronto le sea posible para liberar el peso de la caña; entre en el agua si es necesario. No se deje llevar ni por su entusiasmo ni por la corriente. Su seguridad debe prevalecer por encima de todo.

EL EMPLEO DE UN GARFIO

Garfio. *Como alternativa al empleo de una red, saque el pescado con un garfio. No levante la pieza en el aire.*

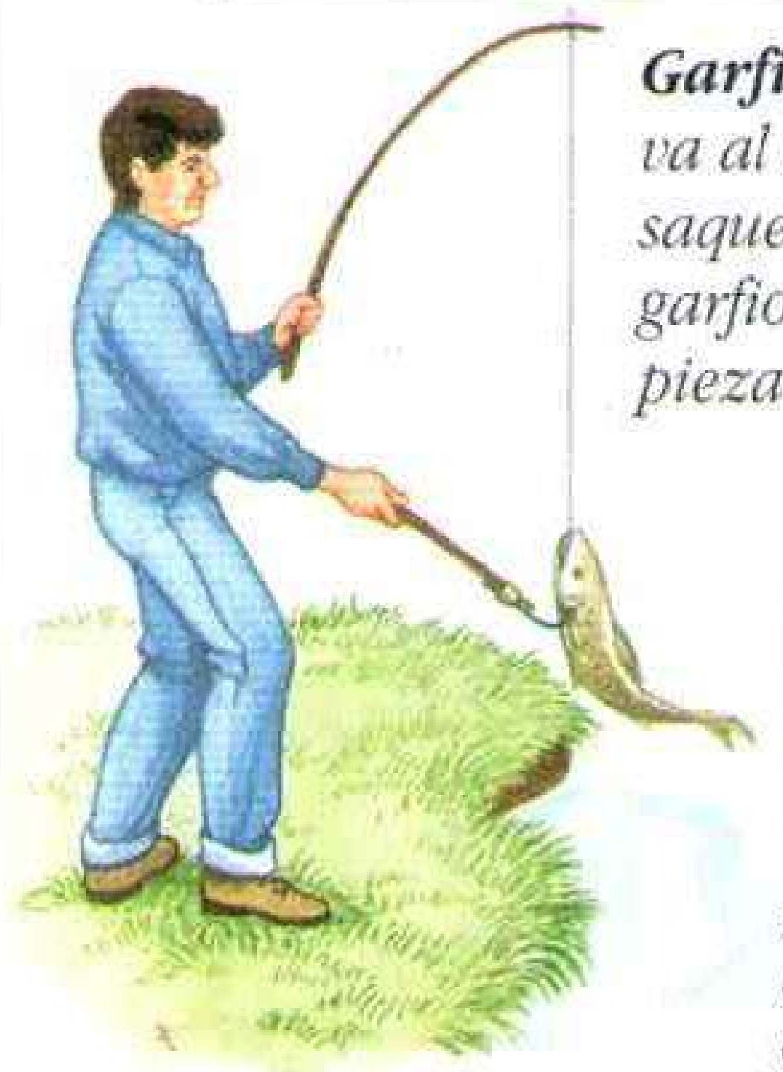

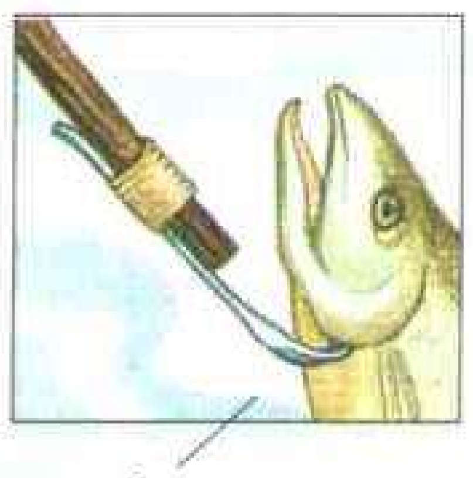

El garfio construido con un tenedor sujeta el pescado

DÓNDE BUSCAR LOS PECES

Un regalo escondido. *Encontrará peces en cualquier lugar donde puedan alimentarse sin sentirse inseguros; éste varía según la especie, el medio y la hora del día. Las aguas muertas (cuando un río se inunda), las orillas salientes, los troncos grandes y las rocas ofrecen buenos refugios, y en días cálidos y soleados, los peces buscarán el agua bajo la sombra o los lugares más profundos. Si el clima es frío, busque los peces en aguas poco profundas, por ejemplo en las orillas de un lago, en una zona calentada por el sol.*

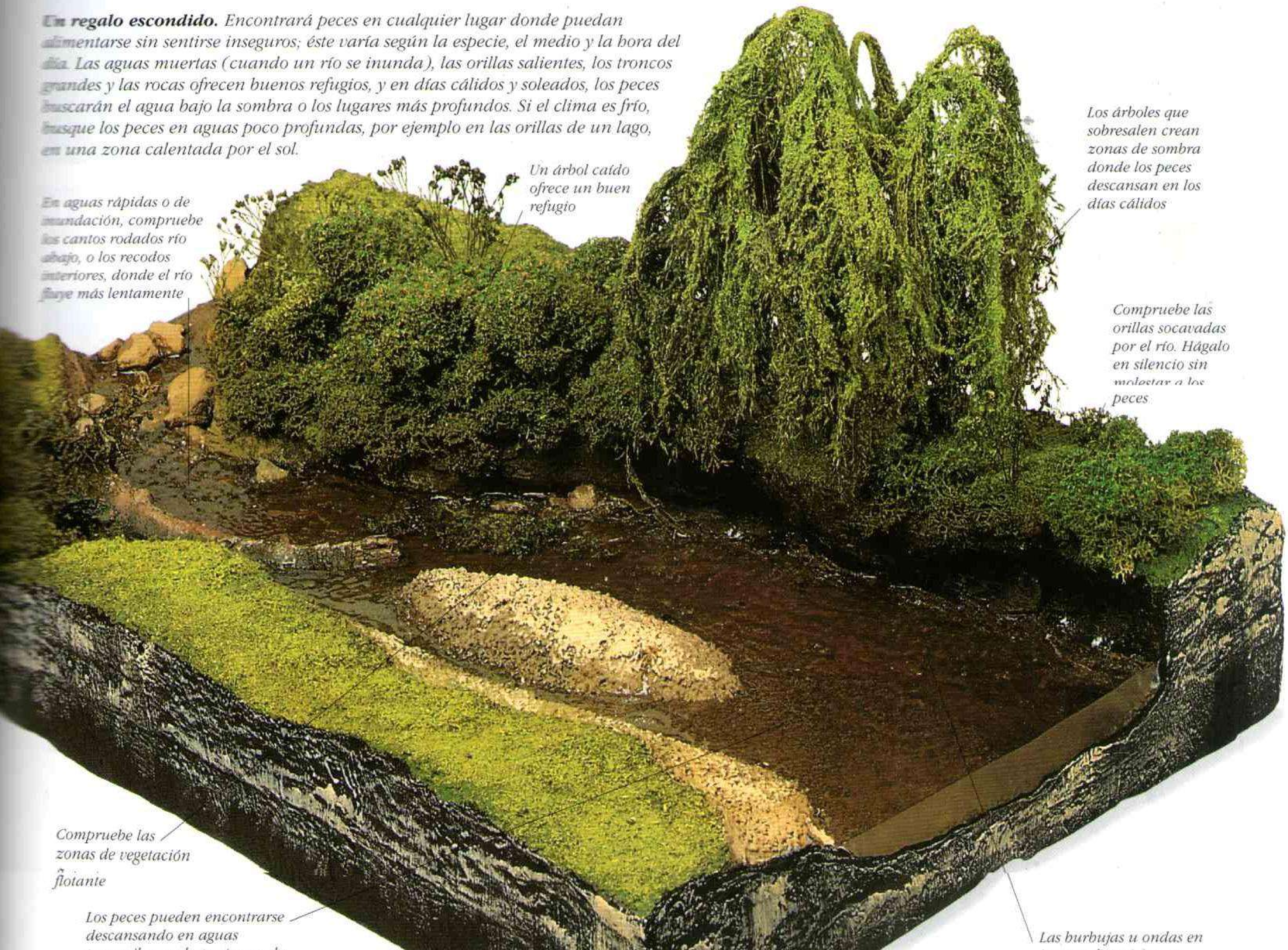

Un árbol caído ofrece un buen refugio

Los árboles que sobresalen crean zonas de sombra donde los peces descansan en los días cálidos

En aguas rápidas o de inundación, compruebe los cantos rodados río abajo, o los recodos interiores, donde el río fluye más lentamente

Compruebe las orillas socavadas por el río. Hágalo en silencio sin molestar a los peces

Compruebe las zonas de vegetación flotante

Los peces pueden encontrarse descansando en aguas tranquilas, en la parte que da río abajo de las graveras, donde el agua rápida se arremolina para formar un pozo

Las burbujas u ondas en la superficie del agua indican que hay peces comiendo

REDES Y TRAMPAS DE PESCA

En una situación de supervivencia, las redes y trampas de pesca son más efectivas que los anzuelos y los sedales sencillamente porque no requieren una atención continua. Las redes pueden fabricarse a partir de cordeles, cuerdas o fibras naturales *(véase página 34)*, utilizando ramas de árboles jóvenes como asas. También puede fabricar una red con ropa que no necesite. Las trampas pueden construirse a partir de ramas y cuerdas, pueden tejerse con hierbas resistentes, e incluso improvisarlas a partir de materiales fabricados por el hombre, como las botellas. El equipo improvisado únicamente debe usarse en caso de emergencia *(véase página 108)*.

CONSTRUCCIÓN DE UN SALABARDO

CON UNA CAMISETA

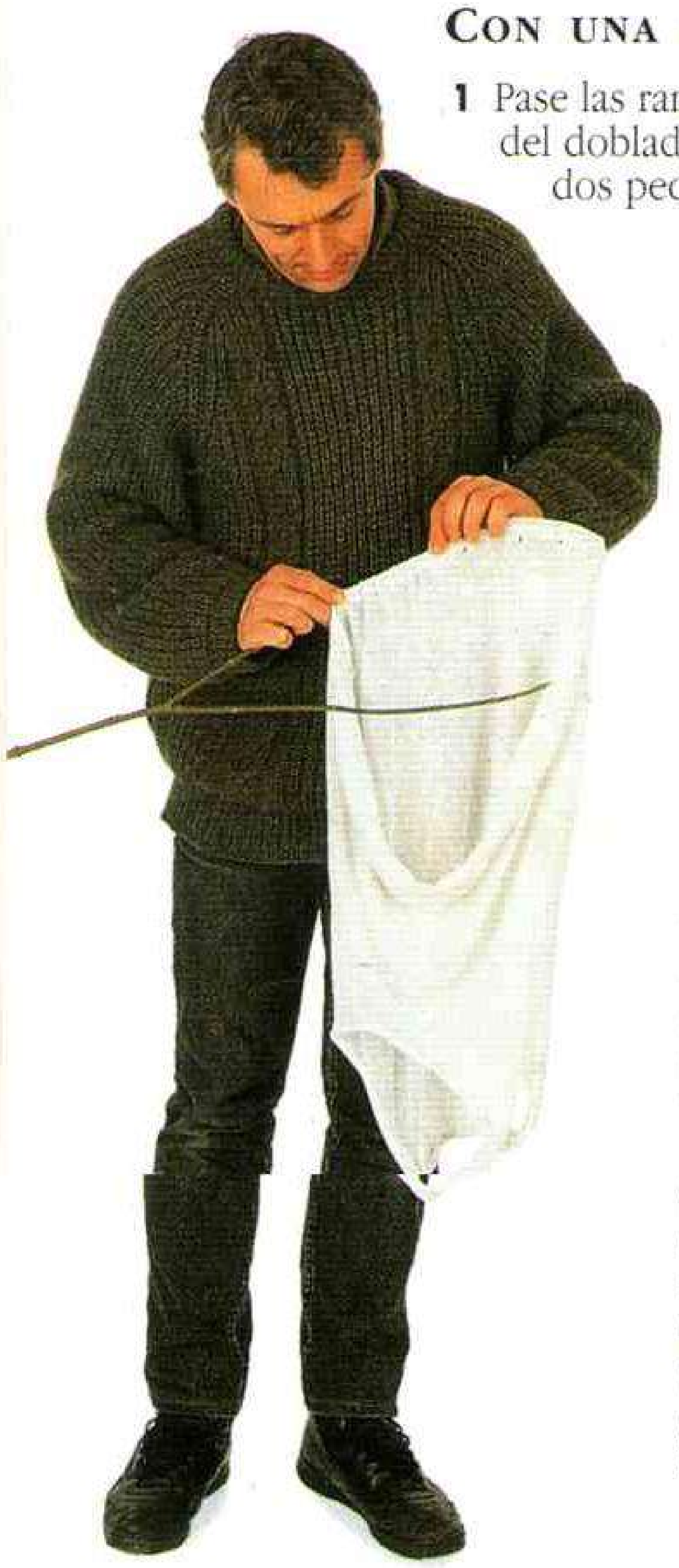

1 Pase las ramas de un árbol joven a través del dobladillo de una camiseta, haciendo dos pequeñas aberturas en la costura.

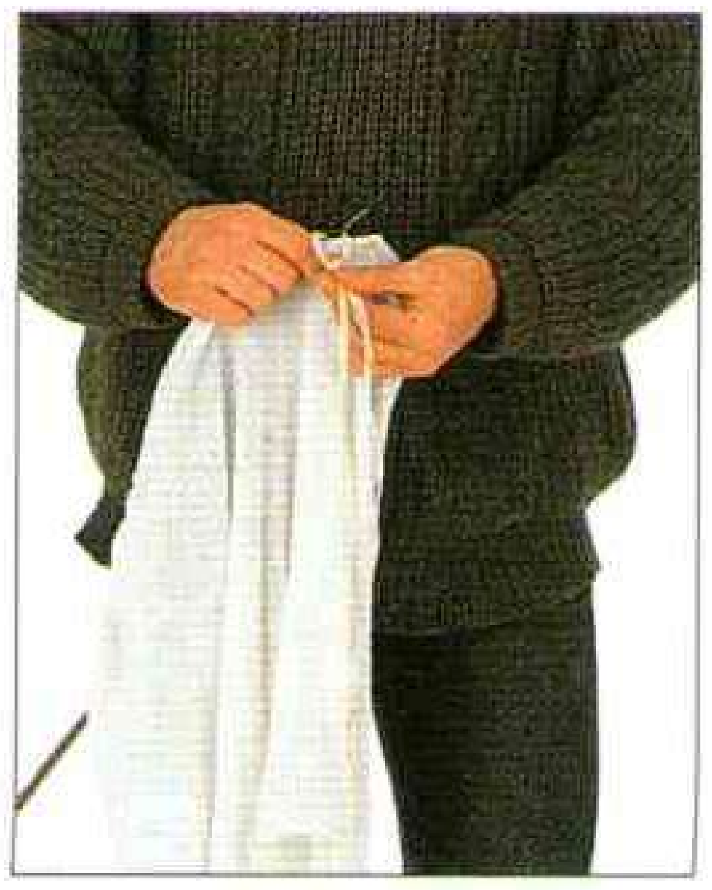

2 Cuando las ramas se encuentren, haga una nueva abertura en la camiseta en este punto y pase ambas ramas a través de ella. Ate ambos extremos entre sí utilizando una cuerda, y vuélvalos a introducir en el dobladillo.

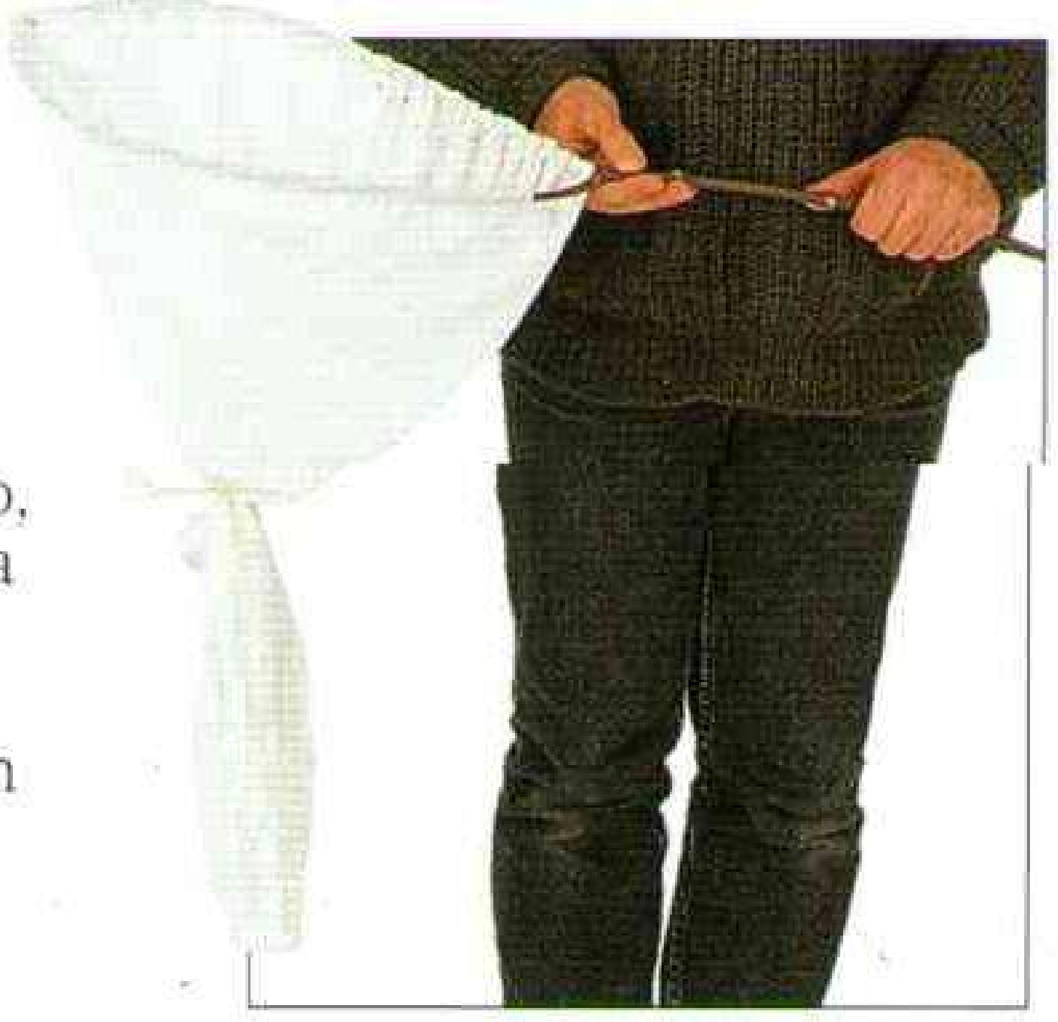

3 Ate la camiseta por encima de las mangas y el cuello. Puede cortar la tela debajo del nudo, o bien invertir la red para que el exceso de tela no se arrastre en el agua cuando la utilice.

CON CUERDAS

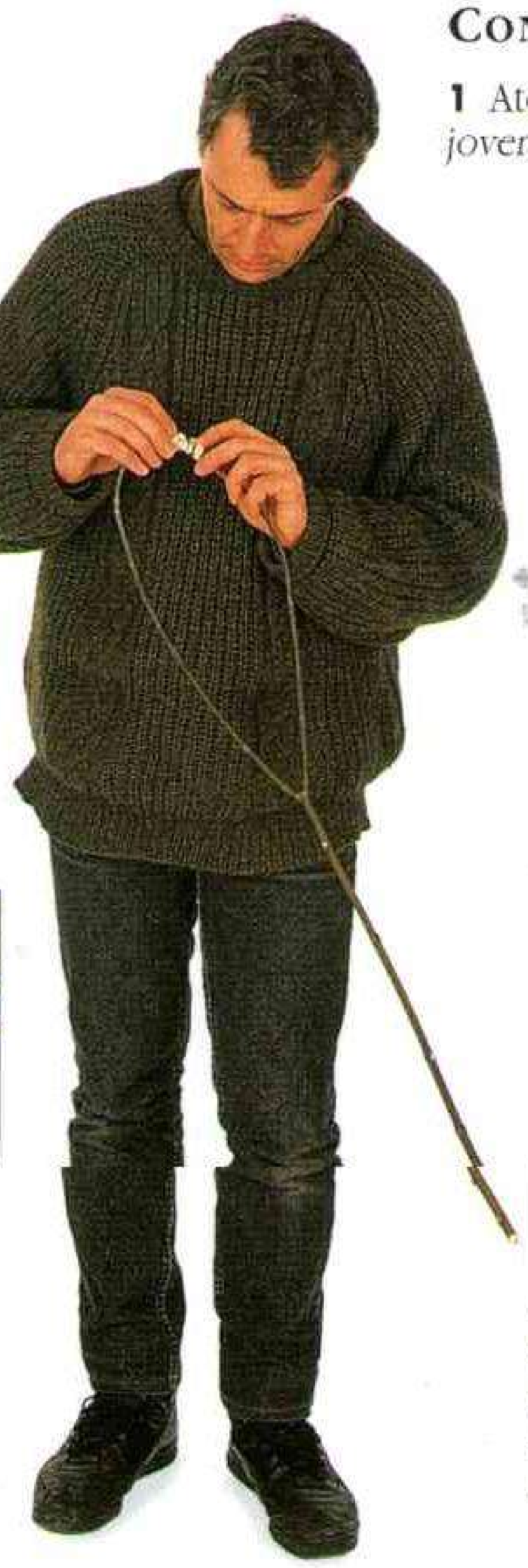

1 Ate los extremos de las ramas de un árbol joven para formar un círculo.

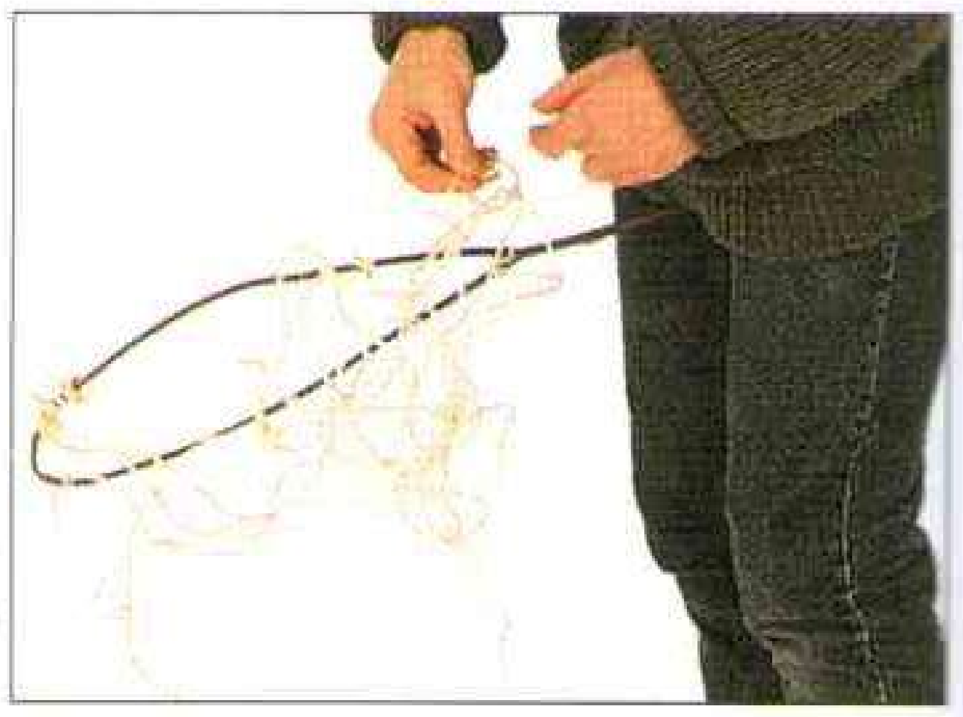

2 Ate uno de los cabos de la cuerda en el círculo. Ate el otro cabo en el lado opuesto del círculo formando aros.

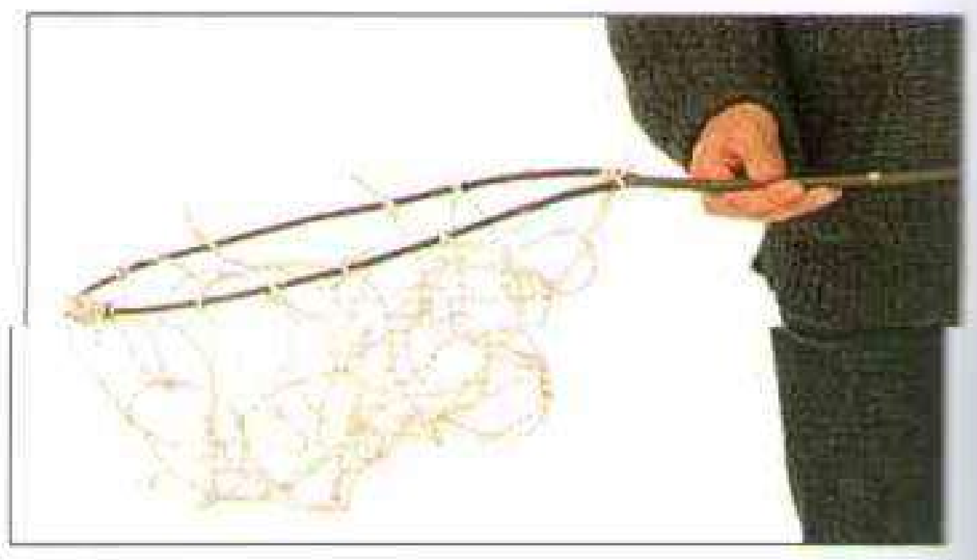

3 Entreteja más cuerdas entre los aros que cuelgan del círculo formado por las ramas, hasta formar una red.

CONFECCIÓN DE UNA RED VERTICAL (PARA PESCA POR ENMALLE)

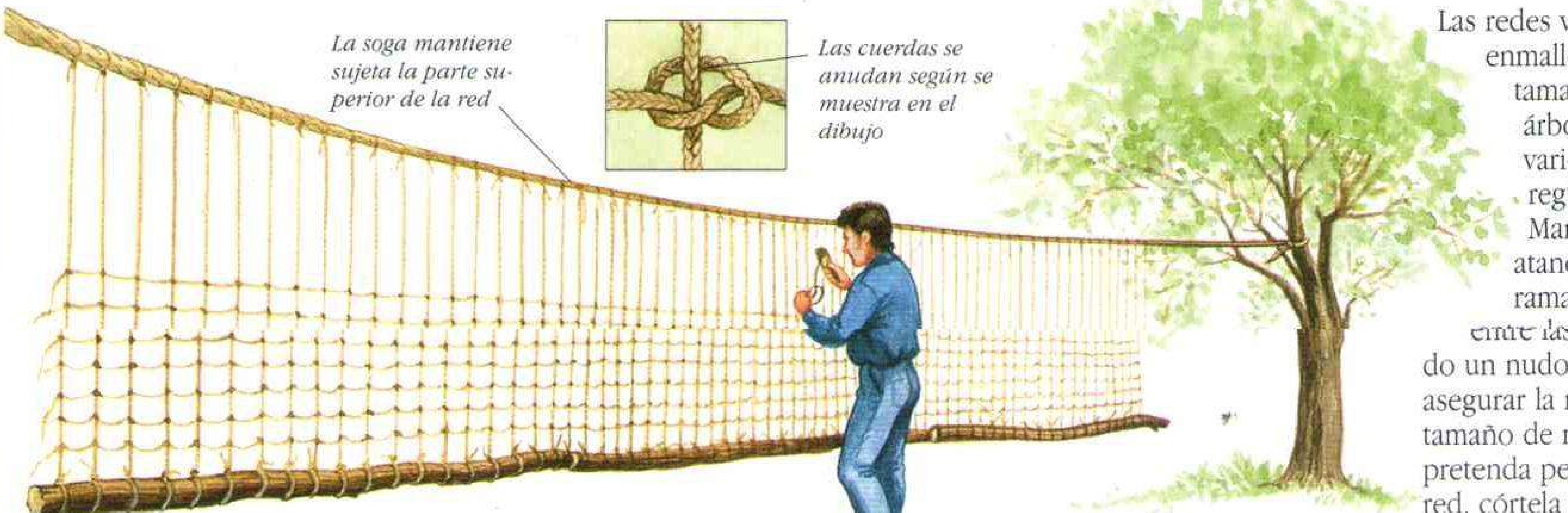
La soga mantiene sujeta la parte superior de la red
Las cuerdas se anudan según se muestra en el dibujo

Las redes verticales (para pesca por enmalle) pueden ser de cualquier tamaño. Ate una soga entre dos árboles y anude los cabos de varios cordeles a intervalos regulares a lo largo de la misma. Mantenga estos cordeles tensos atando sus cabos inferiores a una rama. Entreteja más bramantes entre las cuerdas colgantes, haciendo un nudo alrededor de cada una para asegurar la malla. Puede ajustar el tamaño de malla al tipo de pez que pretenda pescar. Una vez terminada la red, córtela para liberarla de la soga y de la rama.

CONSTRUCCIÓN DE UNA NASA

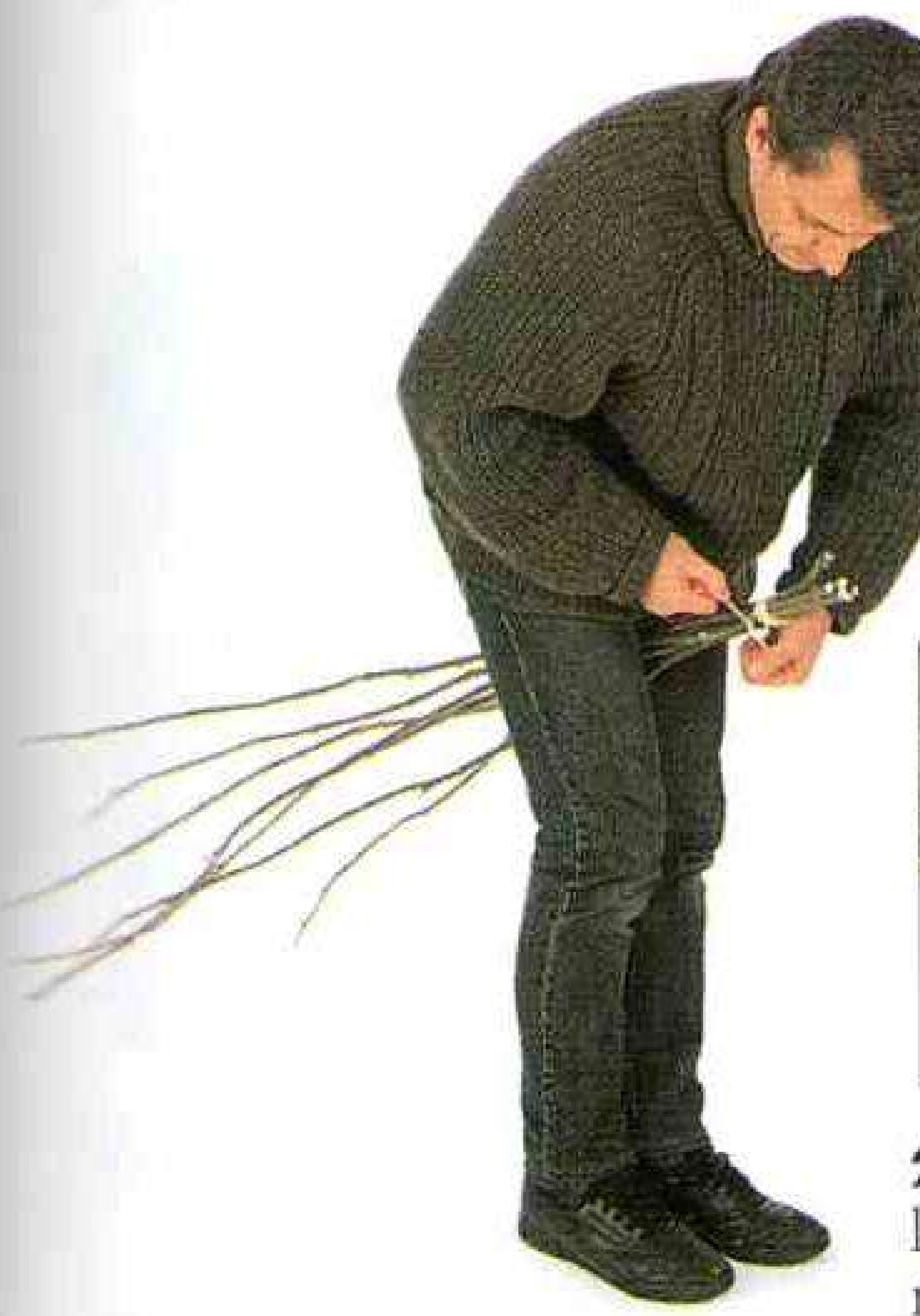

1 La nasa puede utilizarse en un río o en una cala sujeta al efecto de la marea. Ate una gavilla de ramas o árboles jóvenes por sus extremos más gruesos.

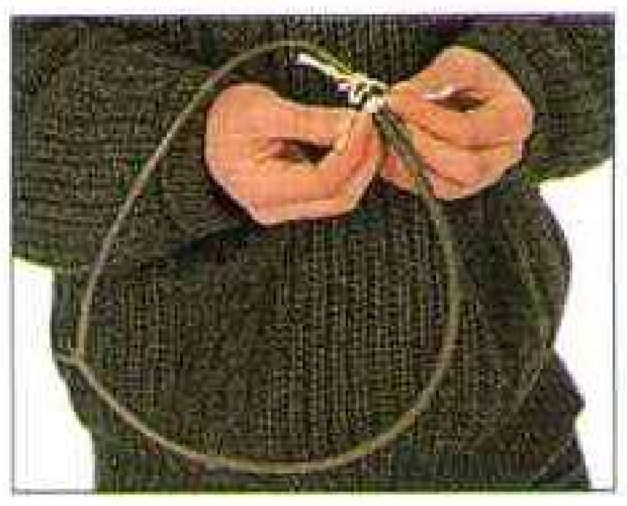

2 Forme un aro atando los extremos de una rama joven.

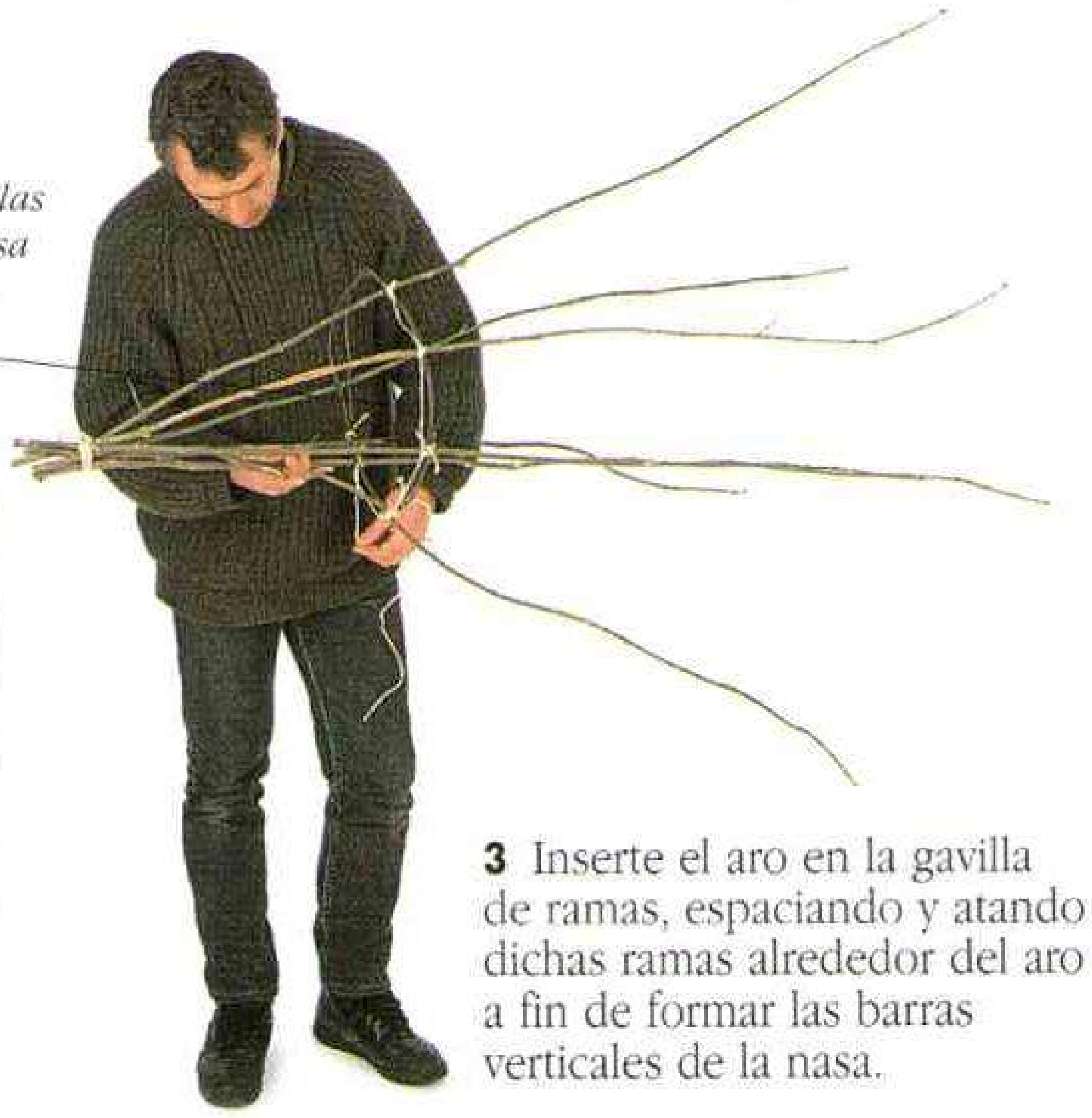

El aro refuerza las barras de la nasa y las mantiene separadas

3 Inserte el aro en la gavilla de ramas, espaciando y atando dichas ramas alrededor del aro a fin de formar las barras verticales de la nasa.

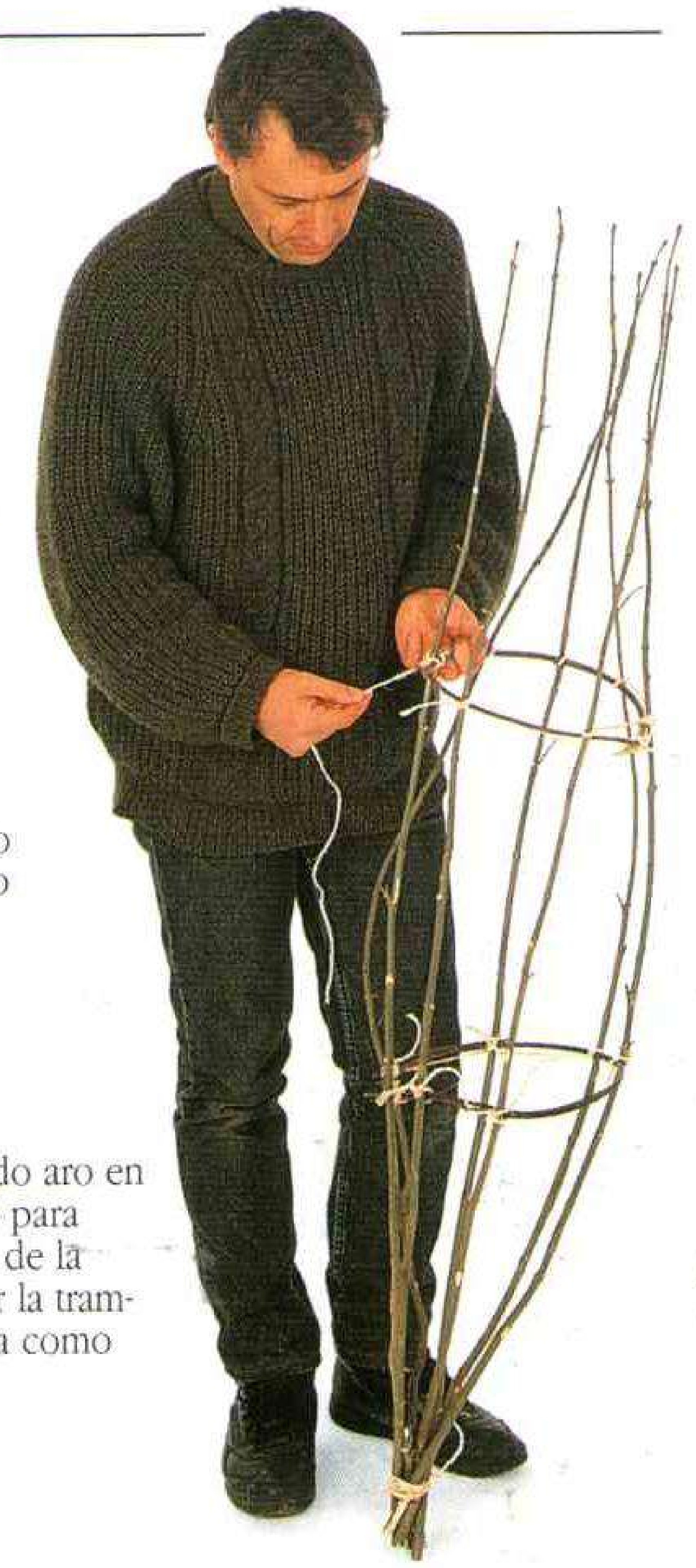

4 Forme un segundo aro en el extremo opuesto para construir la cámara de la trampa. Puede hacer la trampa tan larga y ancha como desee.

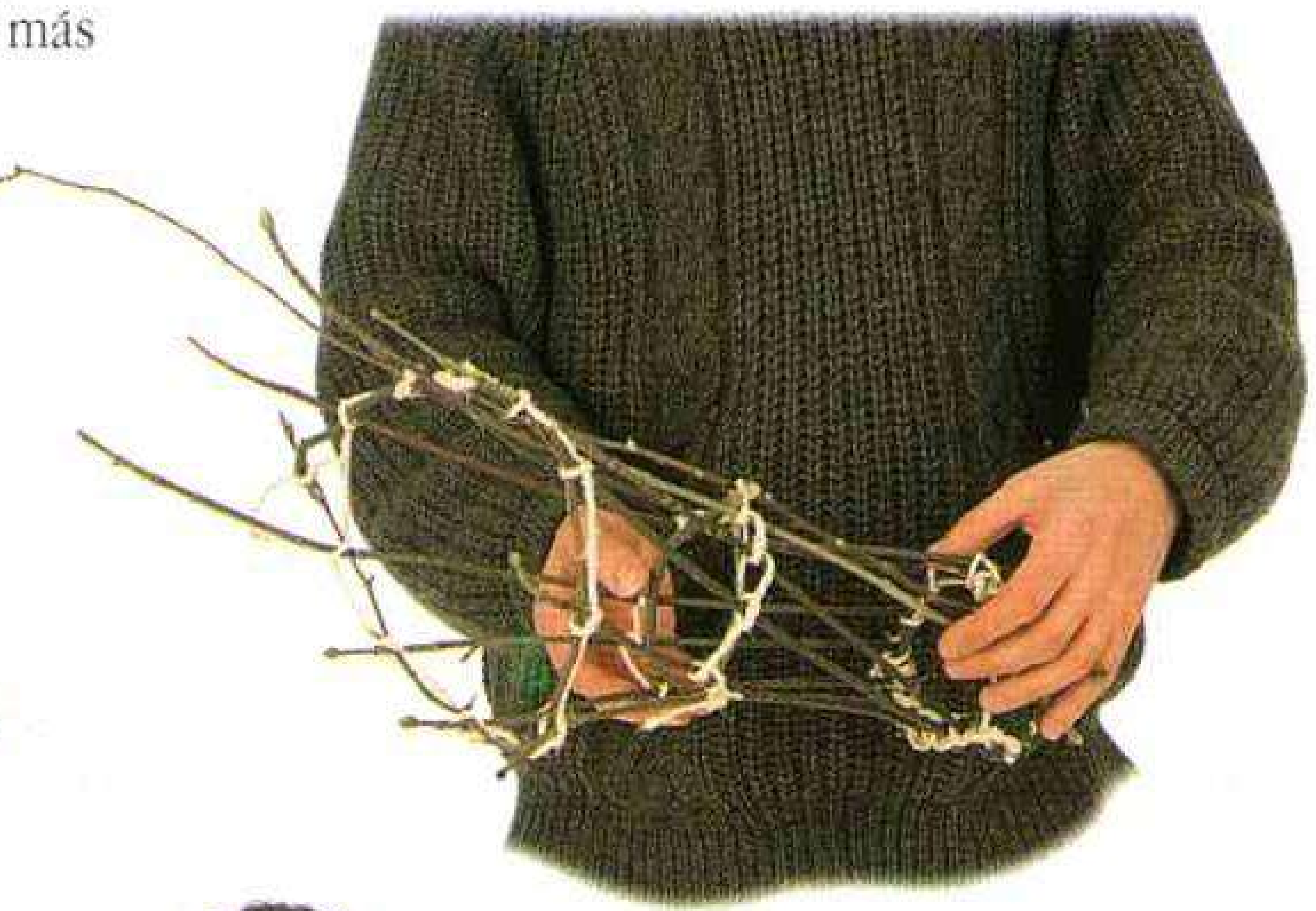

5 Para construir la entrada de la trampa encaje dos aros, uno grande y uno pequeño, sobre un manojo de ramas en forma de embudo, con una salida ancha.

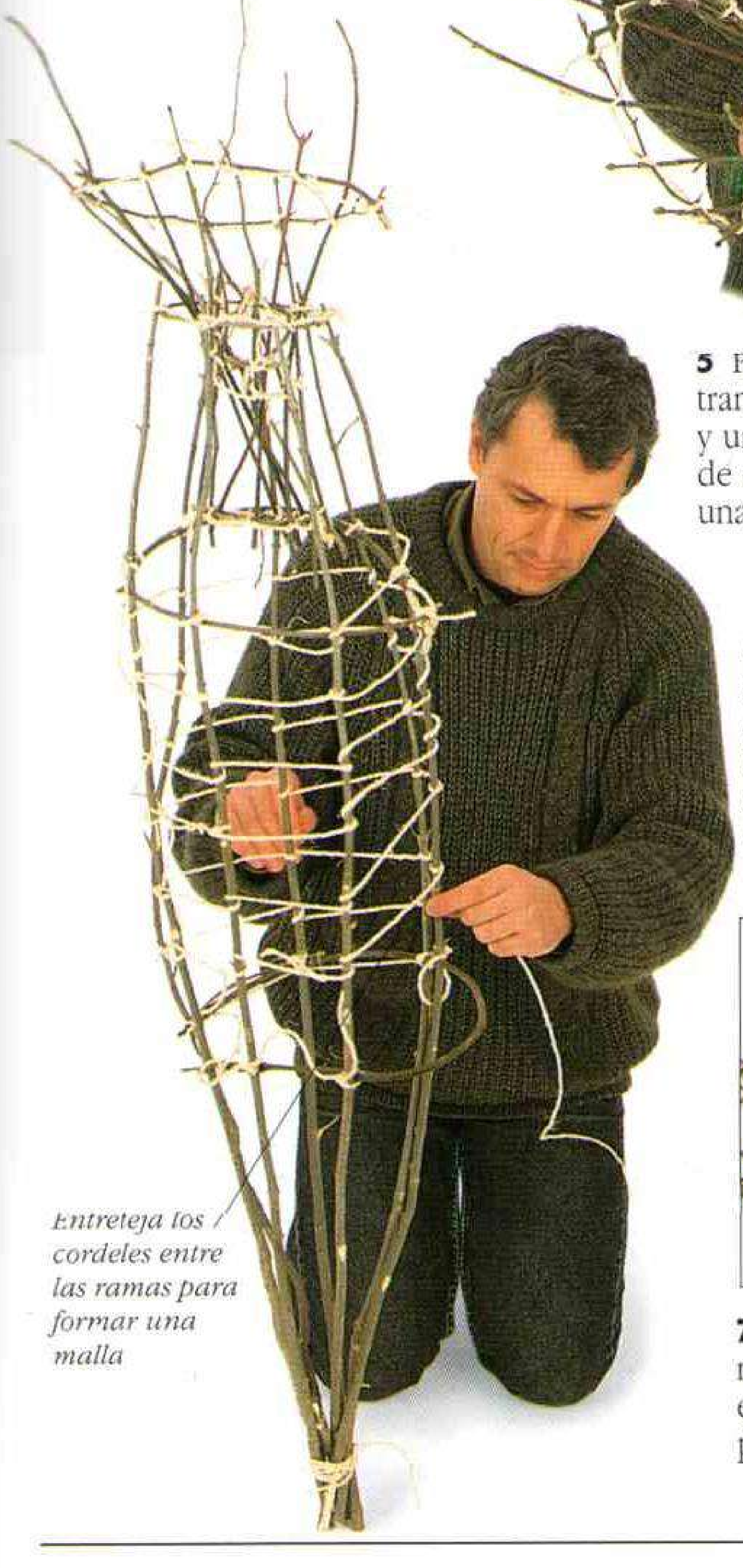

Entreteja los cordeles entre las ramas para formar una malla

6 Invierta el embudo en el extremo abierto de la nasa y átelo en esta posición. Comience a entretejer cuerdas alrededor de las barras para construir los laterales de la nasa.

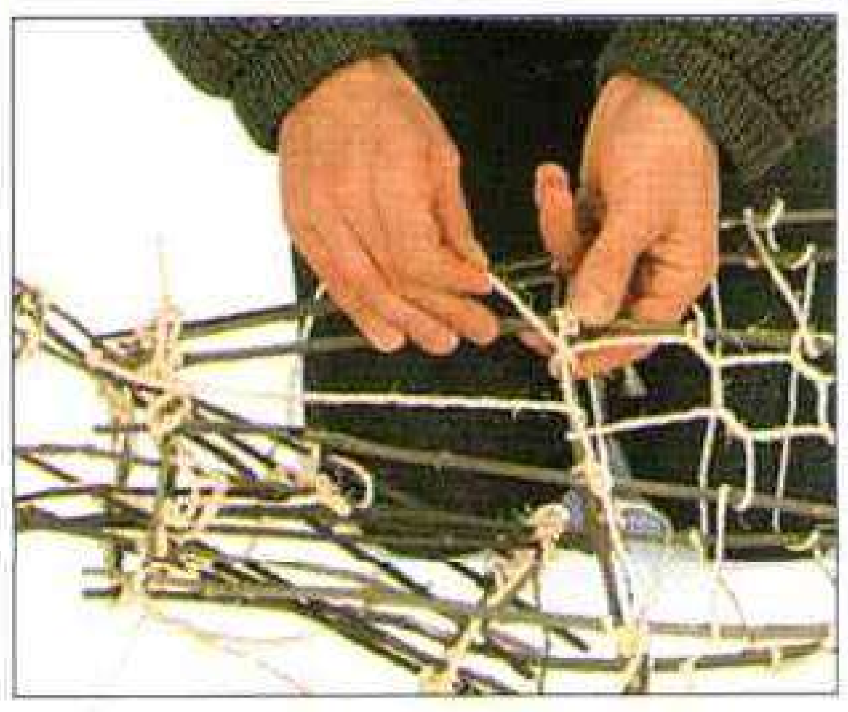

7 Rellene los laterales de la nasa con cordeles. Los peces entrarán en la trampa pero no podrán salir *(véase página 109)*.

OTRAS TRAMPAS

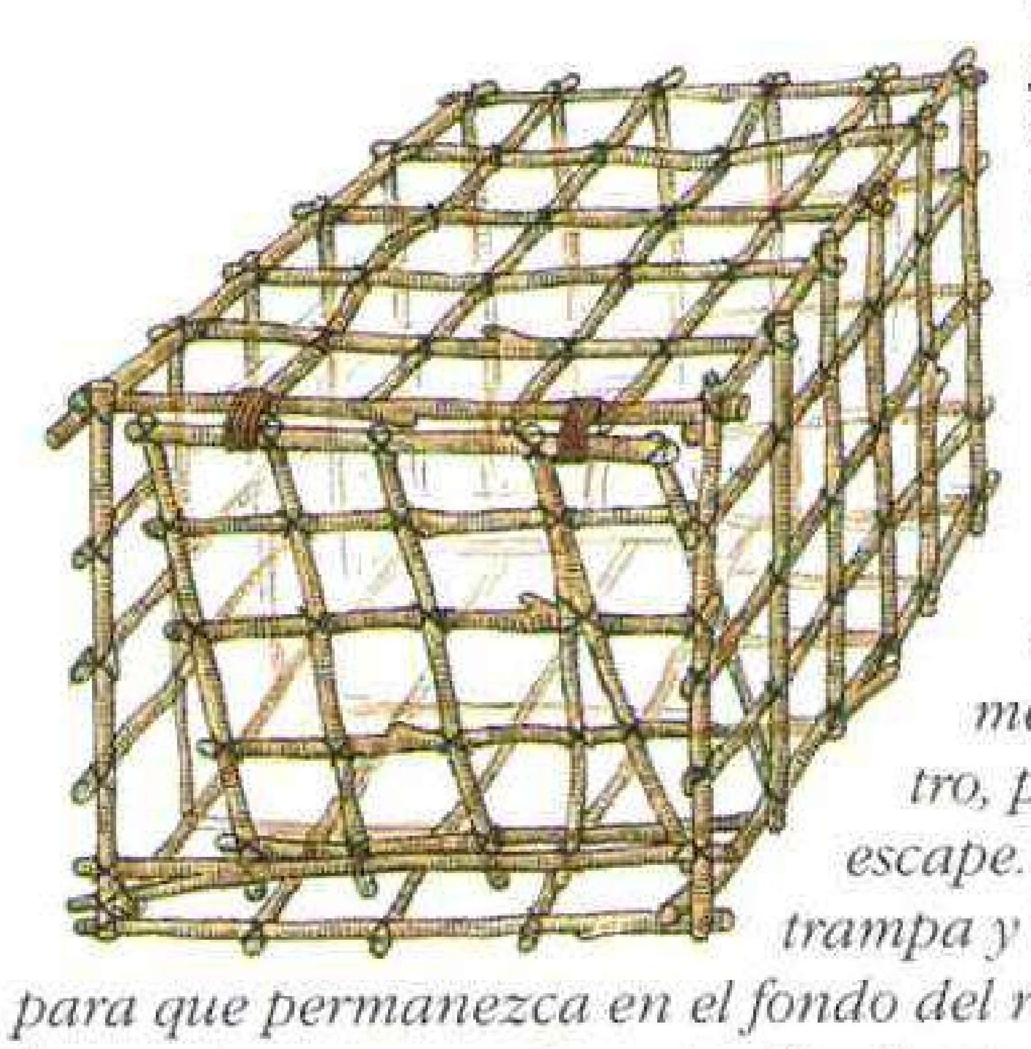

Trampa para cangrejos. *Esta trampa es muy sencilla de construir. Ate varias ramas entre sí para formar una caja. Ajuste el tamaño de la caja a la especie que pretenda capturar. La puerta de la trampa únicamente se abre hacia adentro, para evitar que la presa escape. Coloque un cebo en la trampa y lástrela con una roca para que permanezca en el fondo del río o del mar. Coloque la trampa en mares poco profundos para pescar cangrejos y langostas, o en aguas dulces para pescar cangrejos de río.*

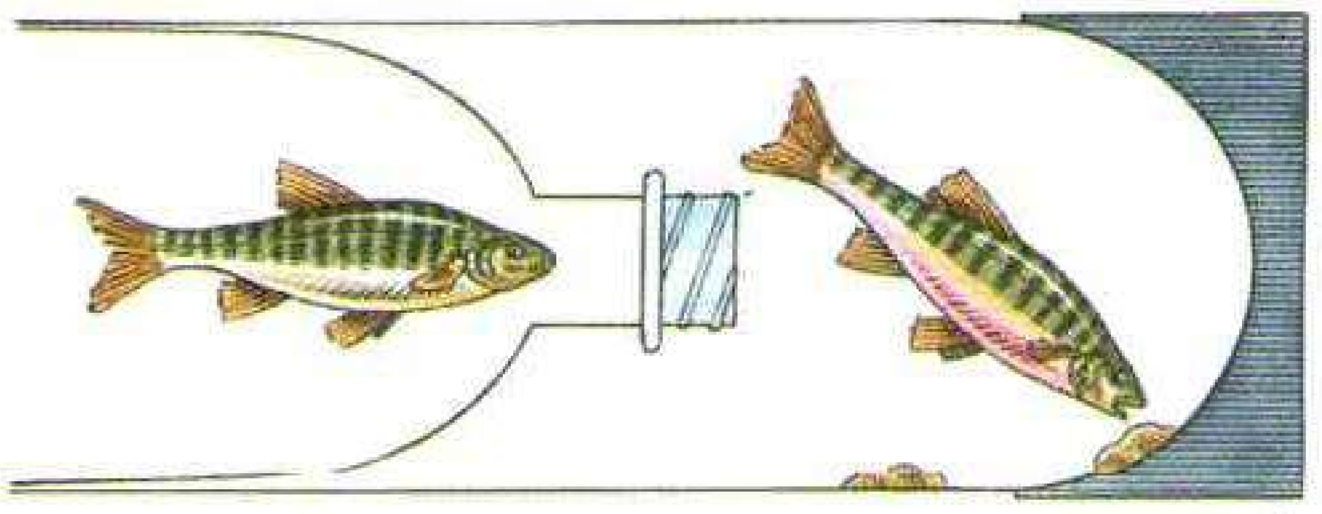

Trampa con botella. *Corte la parte superior de una botella de plástico, inviértala e insértela en la base. Los peces entrarán en la botella para comer el cebo colocado en el fondo, pero no podrán encontrar la salida. Coloque la trampa en la corriente y compruébela a intervalos regulares.*

EL EMPLEO DE REDES Y TRAMPAS DE PESCA

PUEDE DEJAR UN sedal con anzuelo durante toda una noche, pero a menos que lo compruebe al rayar el alba del día siguiente, cualquier pez que haya picado habrá sido engullido por un pez mayor, o se habrá soltado. A diferencia de los anzuelos, las trampas pueden dejarse puestas durante varios días. Las redes no deben utilizarse de la misma manera: deben colocarse durante un período corto de tiempo, o bien utilizarse como redes de arrastre. Los peces de cualquier tamaño pueden quedarse atrapados en una red vertical, y si la deja extendida a lo ancho de un río, puede matar o herir a cualquier especie. Por esta razón, las redes verticales son ilegales en muchas partes del mundo. Estas redes únicamente deben emplearse en una situación de supervivencia en caso de emergencia.

EL EMPLEO DE REDES VERTICALES

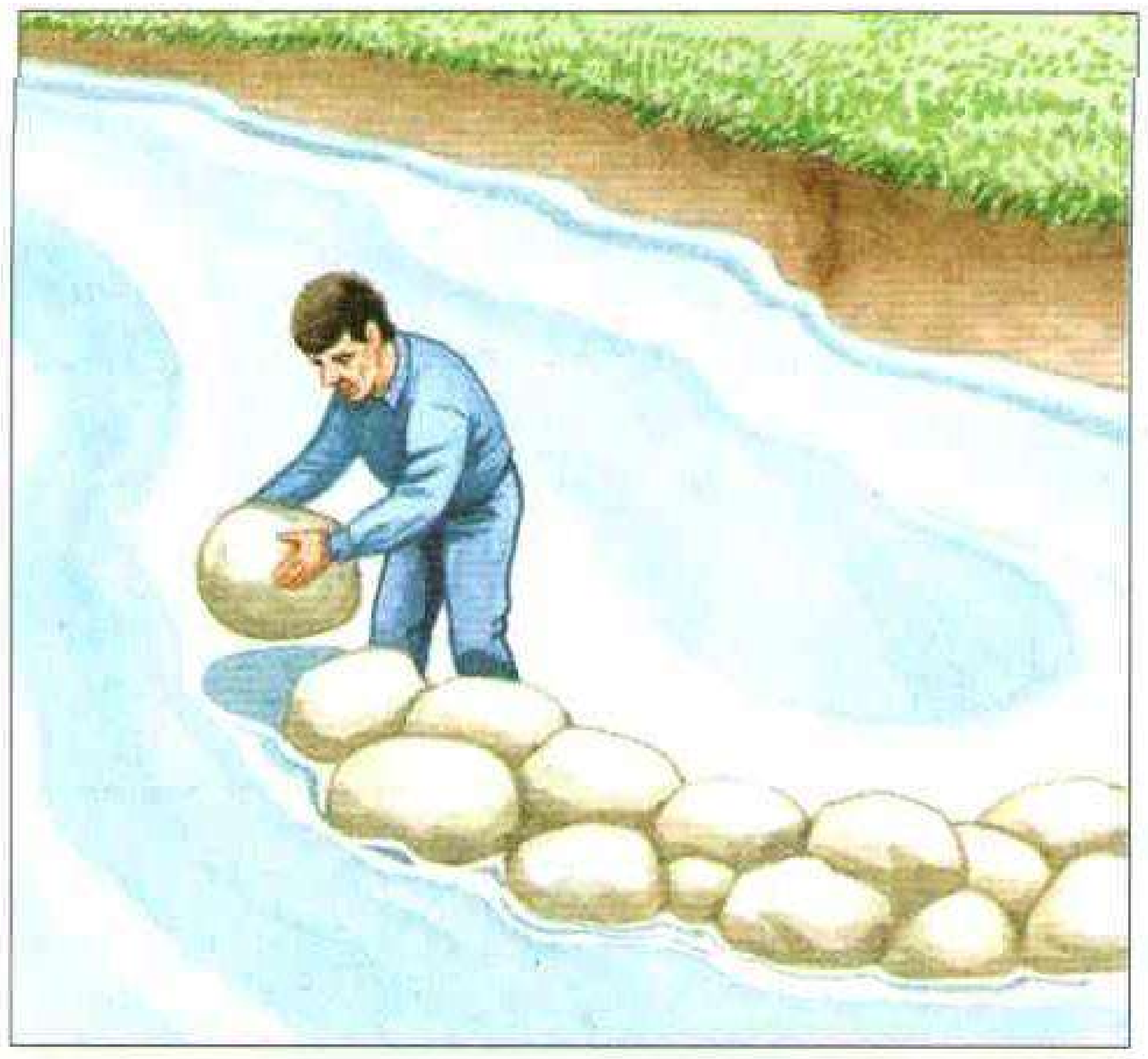

1 En una situación de supervivencia puede que tenga que construir un dique en un río a fin de pescar la suficiente cantidad de peces como para preparar conservas para el invierno. Construya una pared con rocas en el centro de una corriente, escogiendo un recodo, donde el agua fluye más lentamente.

Construya una pared que haga ángulo con la corriente

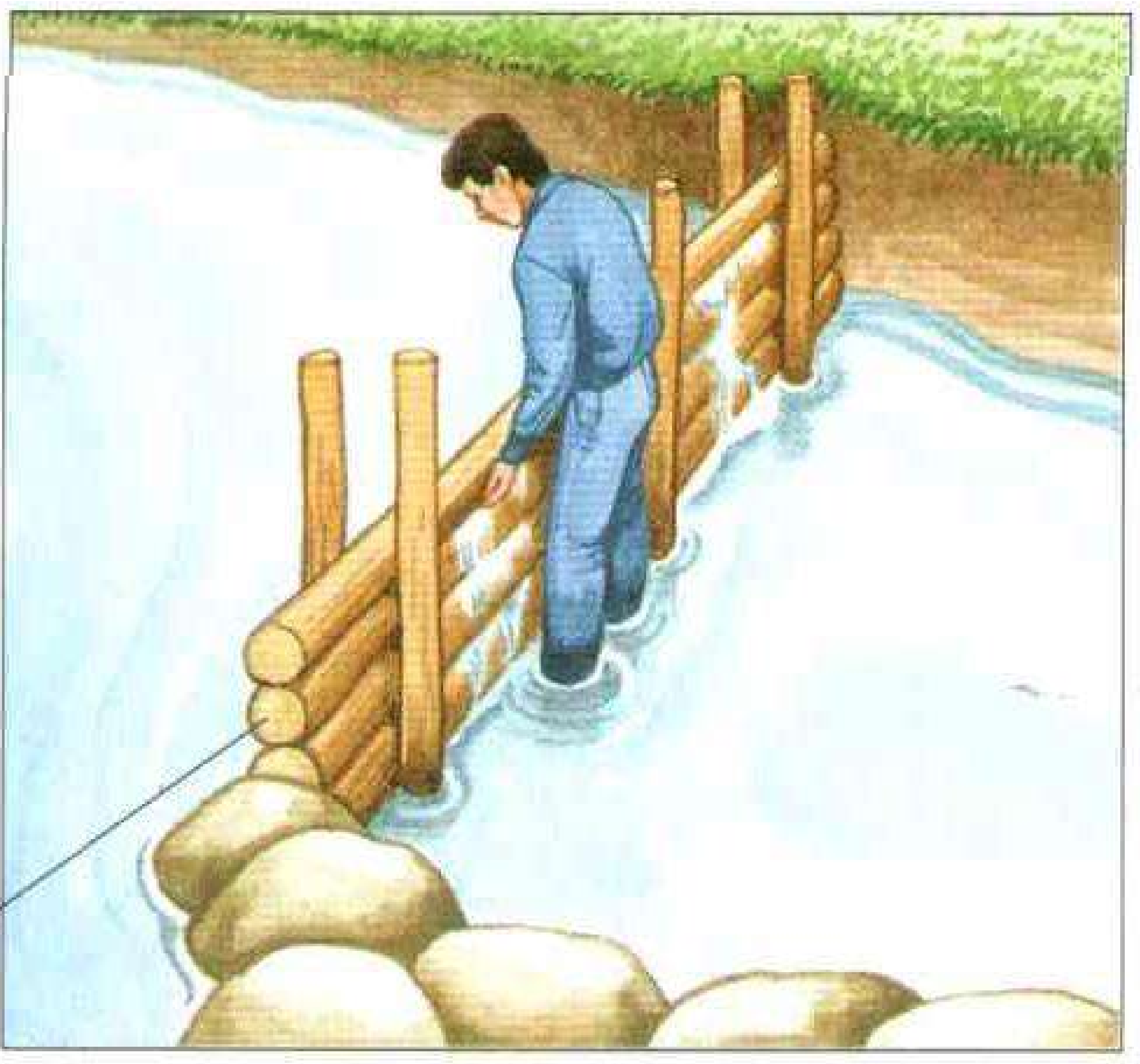

2 Clave tres postes sólidos en el lecho del río separados unos 30 cm entre sí. Colóquelos en diagonal para que el agua no derribe la pared de troncos una vez terminada. Coloque un par de troncos contra los postes en el lado superior de la corriente. Clave otros tres postes como retención. Después añada más troncos para construir la pared.

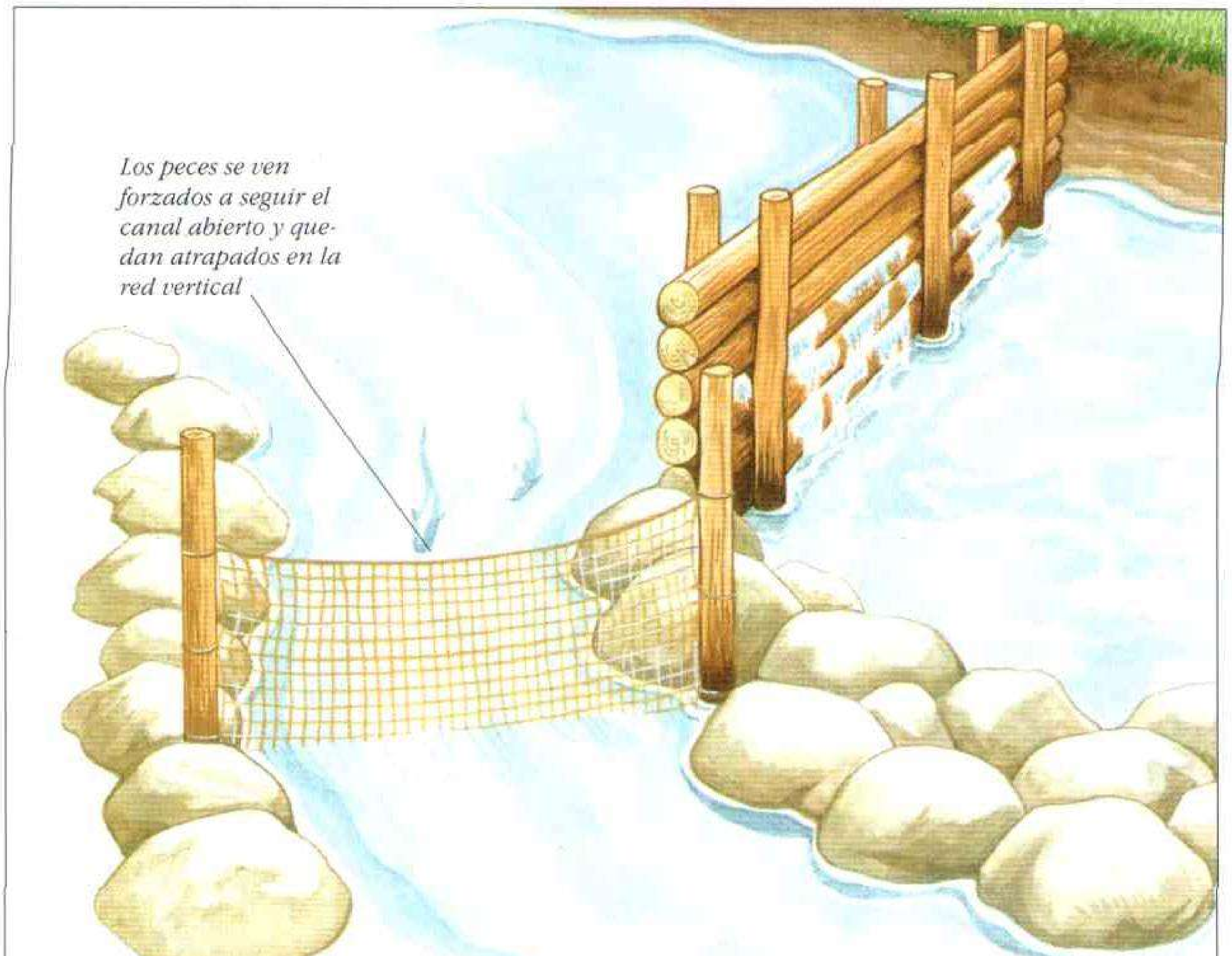

Los peces se ven forzados a seguir el canal abierto y quedan atrapados en la red vertical

3 Clave dos postes a cada lado del canal cerca de la pared y ate a ellos su red vertical *(véase página 106)*. La presa forzará a los peces a seguir por el canal y quedarán atrapados en la red. Cuando dé por finalizada la pesca, retire la red y los postes horizontales de la pared para no atrapar o herir peces sin necesidad.

MANGA DE PESCAR

Una persona toma el extremo de la red, lo coloca sobre sus hombros y se adentra en el agua, mientras la otra permanece en la orilla. La primera persona camina hasta estirar la red y después describe una curva volviendo hacia la playa. Al llegar a la orilla, cualquier pez dentro de los límites de la red quedará atrapado. La red se lleva entonces a la playa y se recogen los peces.

Pesca en la playa. *Dos personas pueden pescar cardúmenes enteros de peces en las aguas poco profundas utilizando una red grande.*

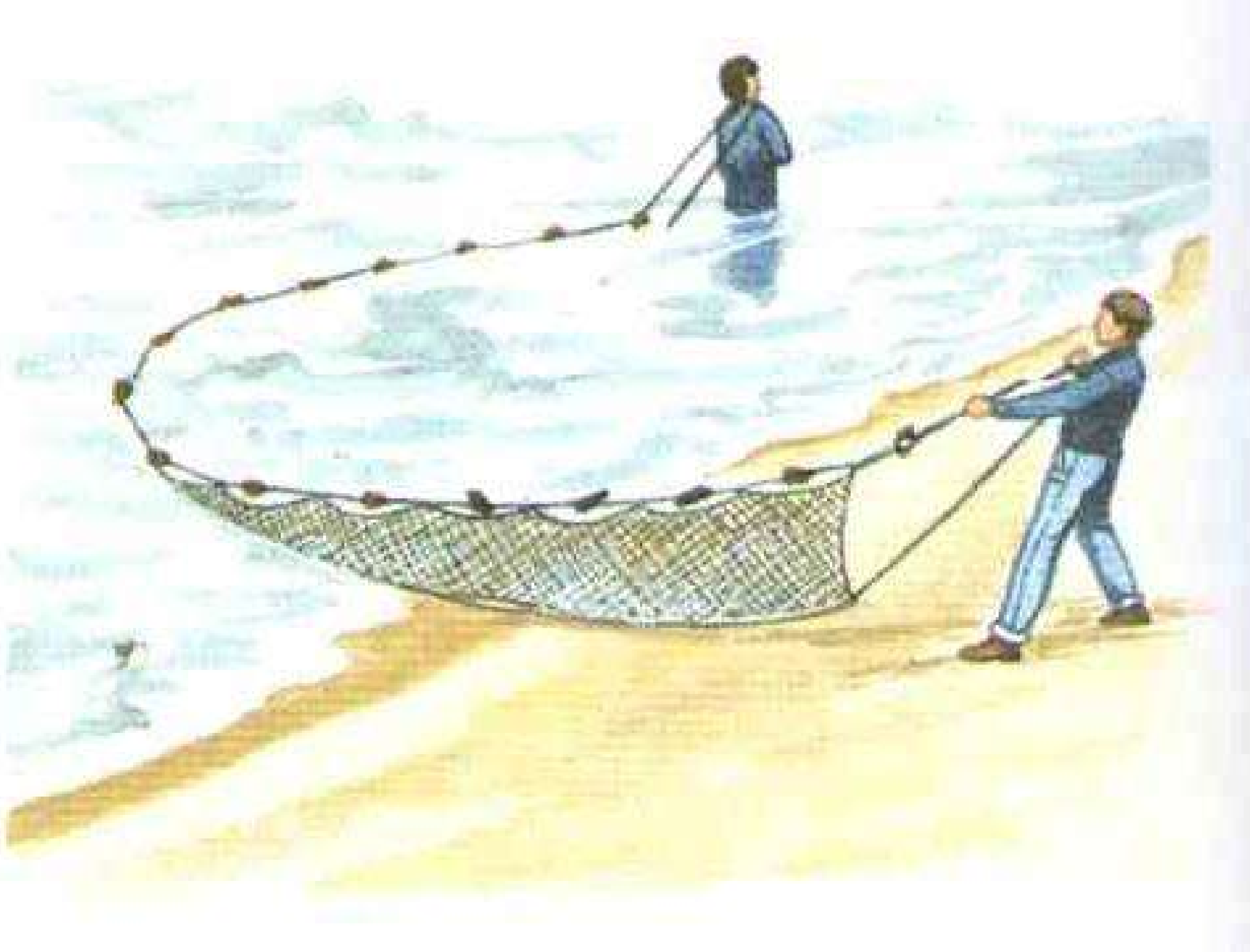

TRAMPAS CON PAREDES DE ROCAS

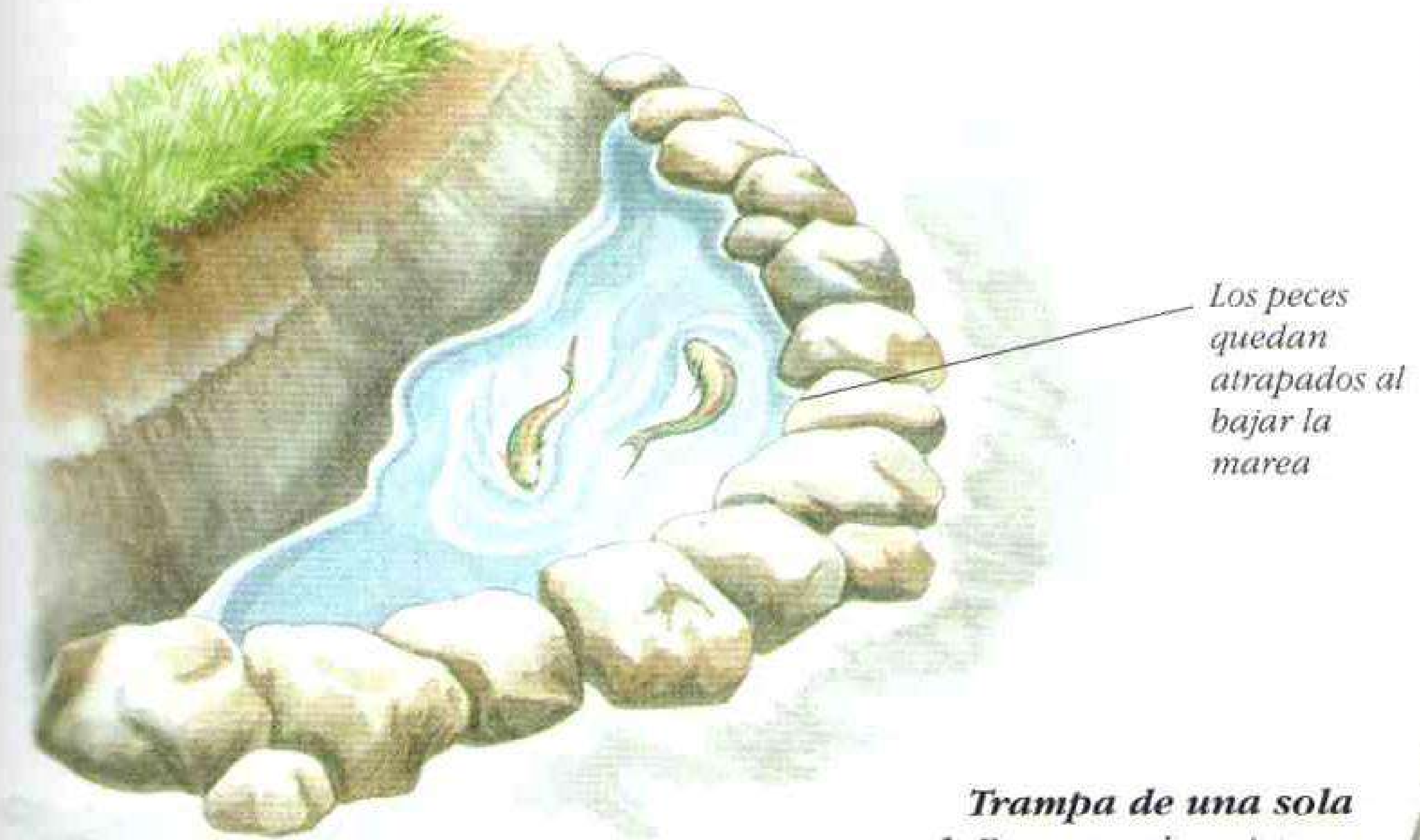
Los peces quedan atrapados al bajar la marea

Trampa de una sola pared. *En una cala sujeta a mareas, puede construir una pared de rocas grandes cerca de la orilla, dándole una forma curva. Cuando la marea baja, los peces quedan atrapados en la pequeña charca que se crea entre la pared y la orilla. En una playa, si no hay una orilla o un arrecife, puede seguir el mismo principio, pero debe construir un círculo completo. Cualquier hueco entre las rocas grandes puede taparse con piedras pequeñas para evitar que los peces se escapen.*

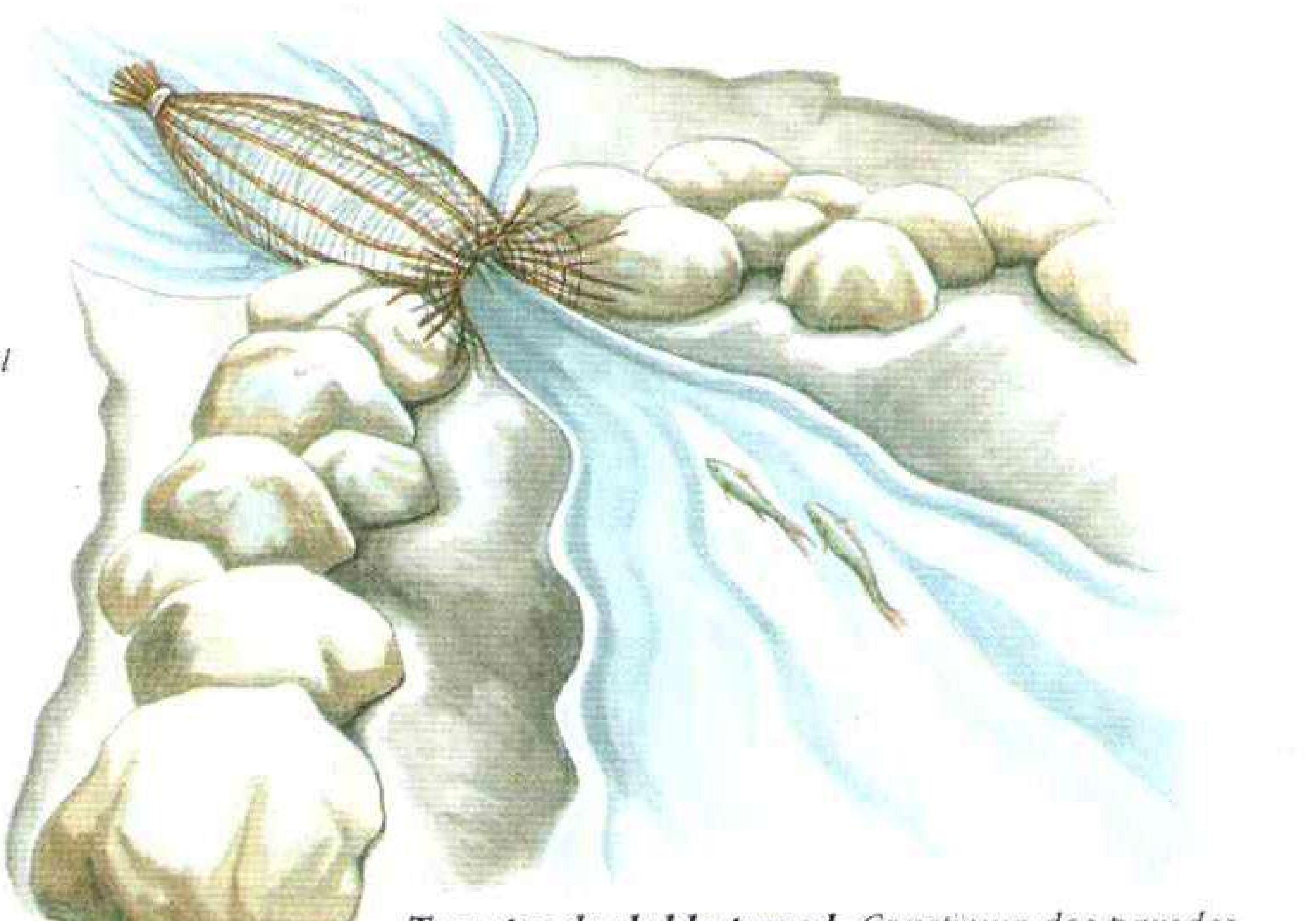

Trampa de doble pared. *Construya dos paredes con rocas en forma de embudo, con la salida corriente abajo. Los peces que naden corriente abajo se dejarán guiar por las paredes hacia el embudo. Coloque su trampa o red en la angostura que forman las dos paredes* (véase página 106).

EL EMPLEO DEL SALABARDO

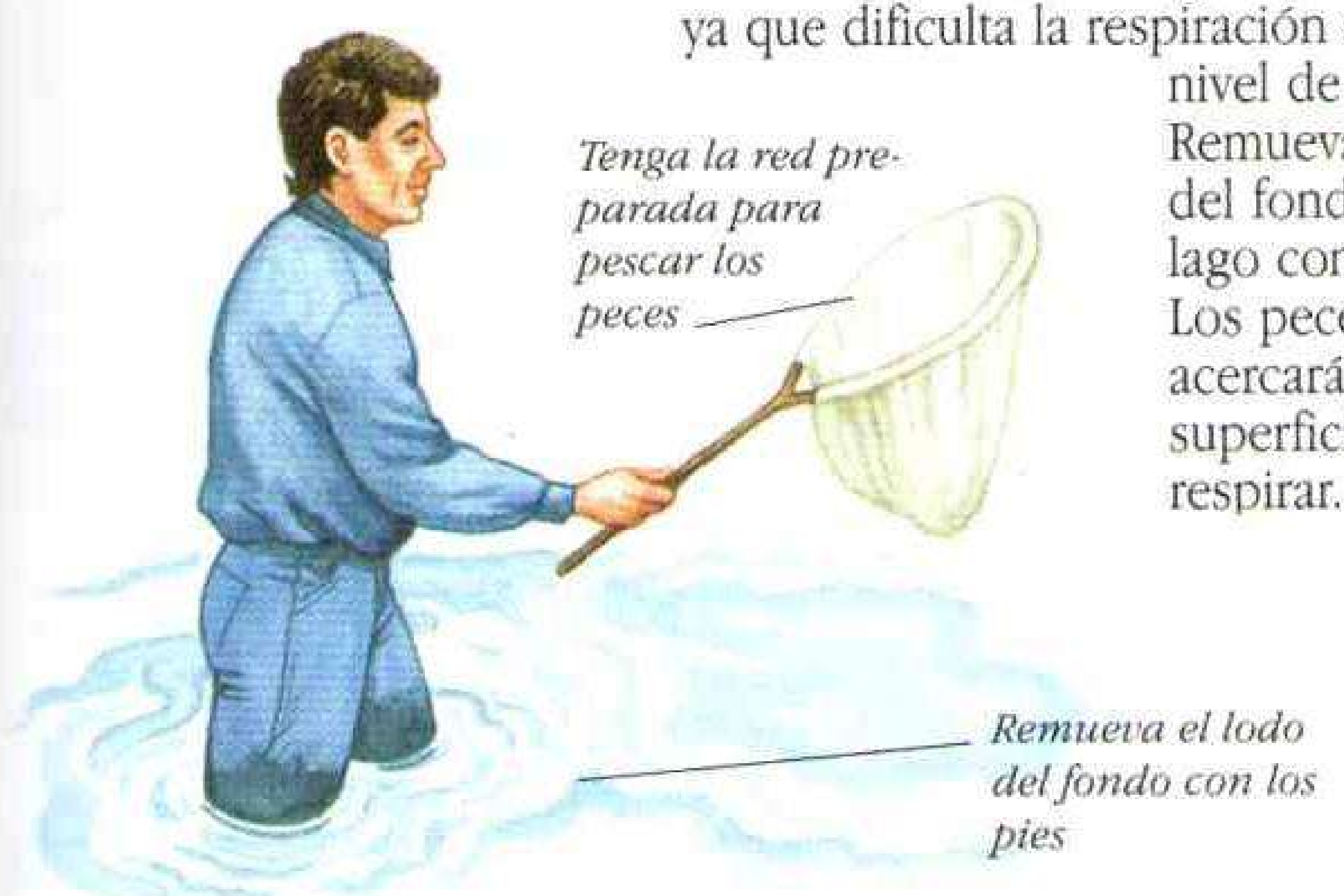
Tenga la red preparada para pescar los peces
Remueva el lodo del fondo con los pies

1 A los peces les desagrada el agua lodosa, ya que dificulta la respiración al disminuir el nivel de oxígeno. Remueva el lodo del fondo de un lago con sus pies. Los peces se acercarán a la superficie para respirar.

2 Utilice la red lenta y cuidadosamente, pero una vez que el pez esté dentro, levántela rápidamente del agua, y cierre la parte superior de la red, como si de una bolsa se tratara, para evitar que el pez salte fuera.

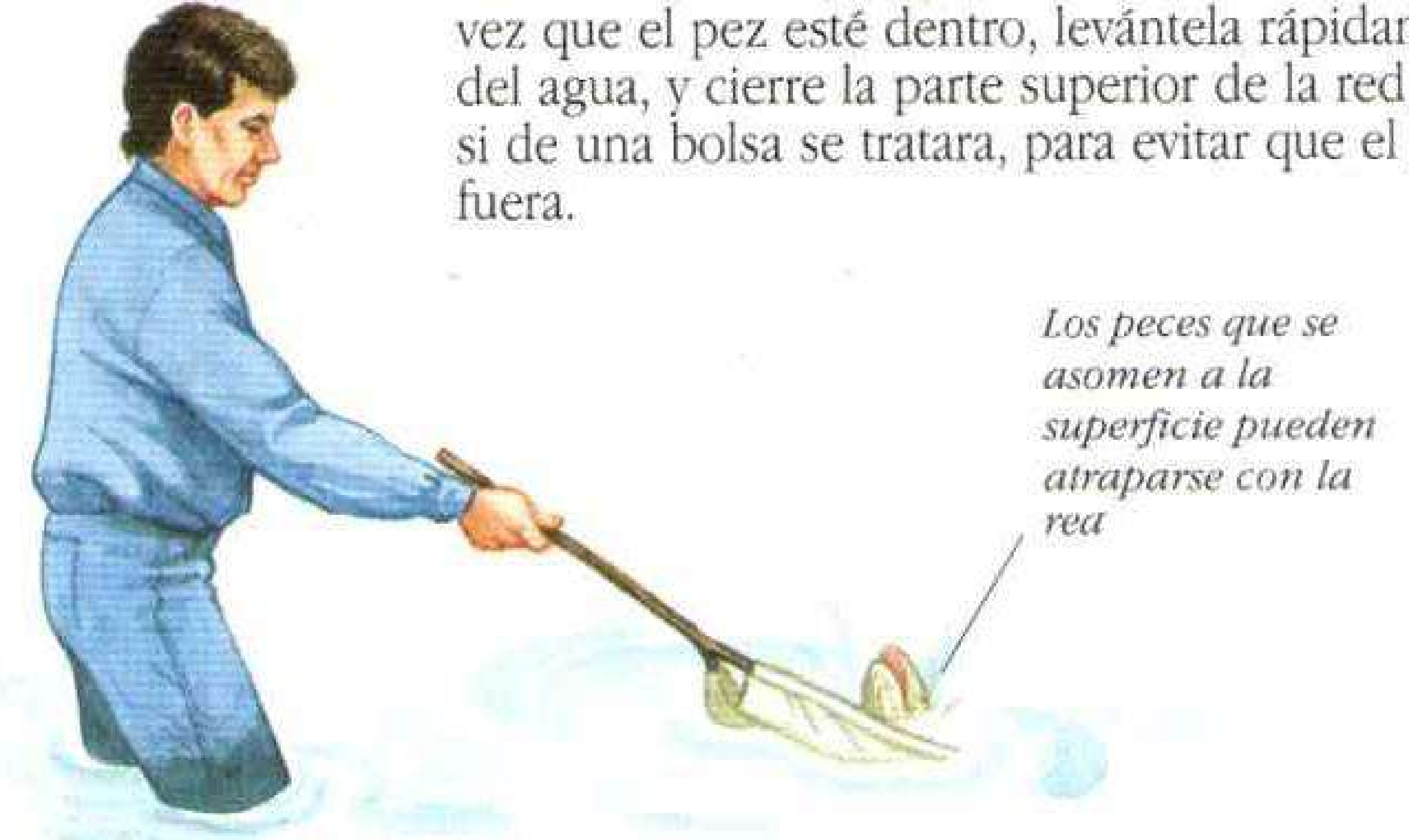
Los peces que se asomen a la superficie pueden atraparse con la rea

EL EMPLEO DE REDES CON CEBO

Algunos peces pueden pescarse con una red con cebo. Utilice una red fabricada con una camiseta *(véase página 106)*, o confeccione una con un calcetín. Pase una rama bífida a través del dobladillo del calcetín, que le servirá de asa. Coloque excrementos o despojos de pescado como cebo, y muévala en el agua.

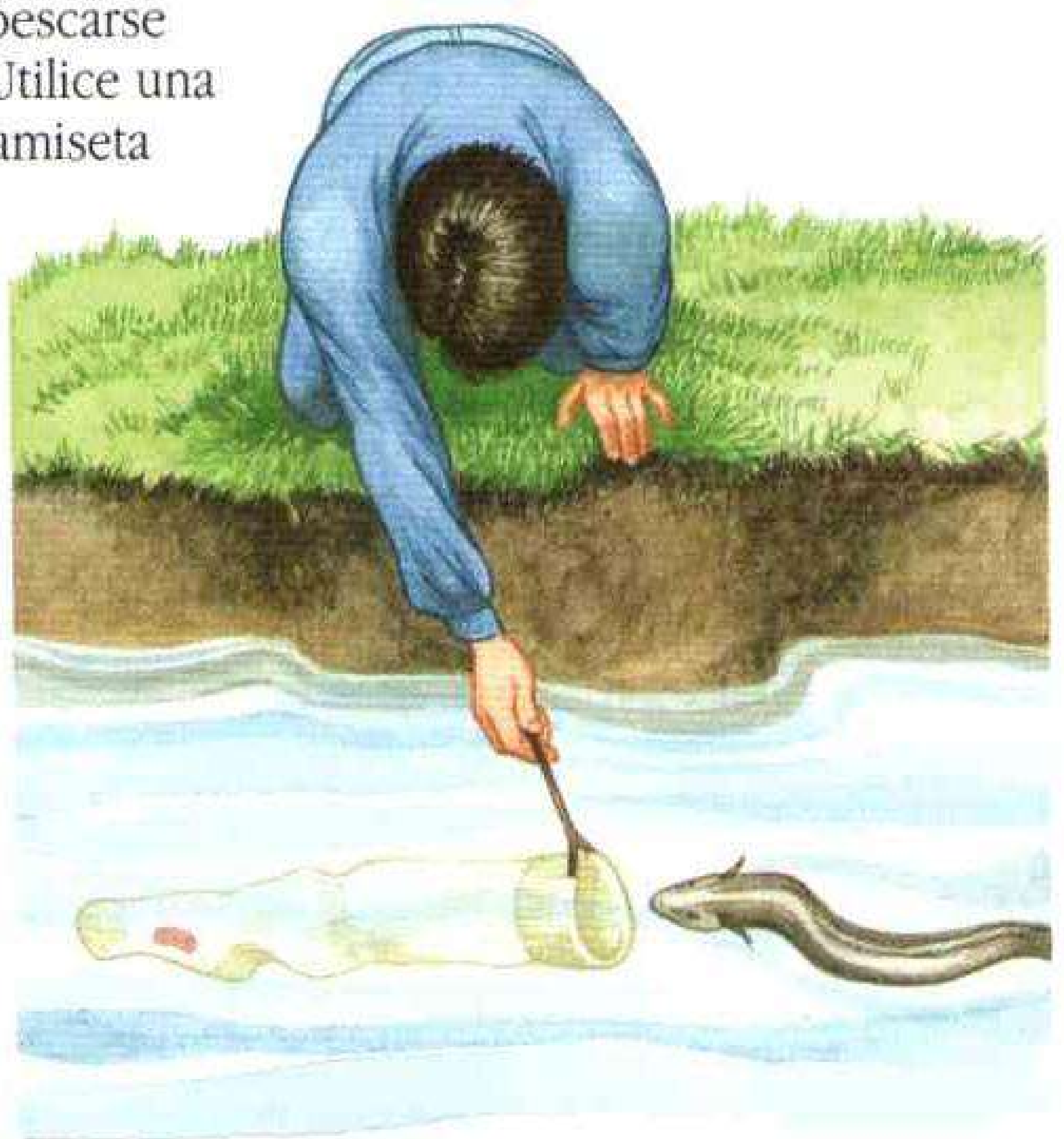

Cebo para anguilas. *Las anguilas se sienten atraídas hacia una red con despojos o excrementos animales.*

PUNTOS QUE DEBE TENER EN CUENTA AL PESCAR

- El agua siempre es peligrosa, fluya de forma rápida o lenta, sea poco o muy profunda. Incluso una caída sobre las rocas húmedas puede ser fatal.
- Considere concienzudamente todos los aspectos sobre su técnica de pesca, intentando imaginar los peligros antes de que ocurra un accidente; tome siempre las precauciones necesarias.
- Los peces tienen tendencia a volver al agua. Piense cómo llevará los peces a tierra una vez los ha pescado.
- Los peligros submarinos varían desde las algas en las que puede enredarse, o las criaturas peligrosas como los cocodrilos, hasta los pozos profundos y las fuertes corrientes. No entre en un río o corriente si puede evitarlo, aun si puede ver el fondo.
- Nunca salte al agua, y menos de cabeza. El reconocimiento de la zona resulta vital antes de tomar un baño con fines recreativos.
- Tenga mucho cuidado al utilizar equipos improvisados de pesca como el arpón, ya que podría romperse y hacerle caer al agua.

PREPARAR Y COCINAR EL PESCADO

INMEDIATAMENTE DESPUÉS DE llevar un pez a tierra, mátelo asestándole un golpe seco en la parte posterior de la cabeza, utilizando un palo o una piedra. Debe limpiar, cocinar y comer el pescado lo más pronto posible, o bien conservarlo para su posterior consumo *(véase página 118)*. En los climas cálidos, basta una corta exposición al sol para que el pescado se estropee, ya que su piel húmeda es un excelente foco de cultivo para las bacterias y las moscas. Una vez descamado, desollado, cortado y limpio es menos probable que se descomponga. En los climas fríos, puede esperar entre seis y doce horas antes de limpiar y cortar el pescado, lo cual permitirá que todas las terminales nerviosas mueran, y que el proceso sea mucho más sencillo que con una pesca reciente. Las espinas, piel y cabeza del pescado pueden ser hervidas en agua para hacer un caldo –siempre y cuando lo haga inmediatamente. Debe darse al caldo el mismo tratamiento que al pescado sin conservar: manténgalo fresco y consúmalo tan pronto como sea posible.

CÓMO LIMPIAR EL PESCADO

La limpieza elimina las partes del pescado con mayor tendencia a la descomposición, dejando la mayor parte de carne posible. Además, el pescado puede alojar en sus órganos internos sustancias contaminantes. Por ejemplo, los metales pesados como el mercurio, presentes en el mar en pequeñas cantidades, se concentran en el hígado y pueden resultar mortales para los hombres si se consumen en grandes cantidades. Las sustancias peligrosas no tienden a acumularse, por lo general, en los músculos del pescado.

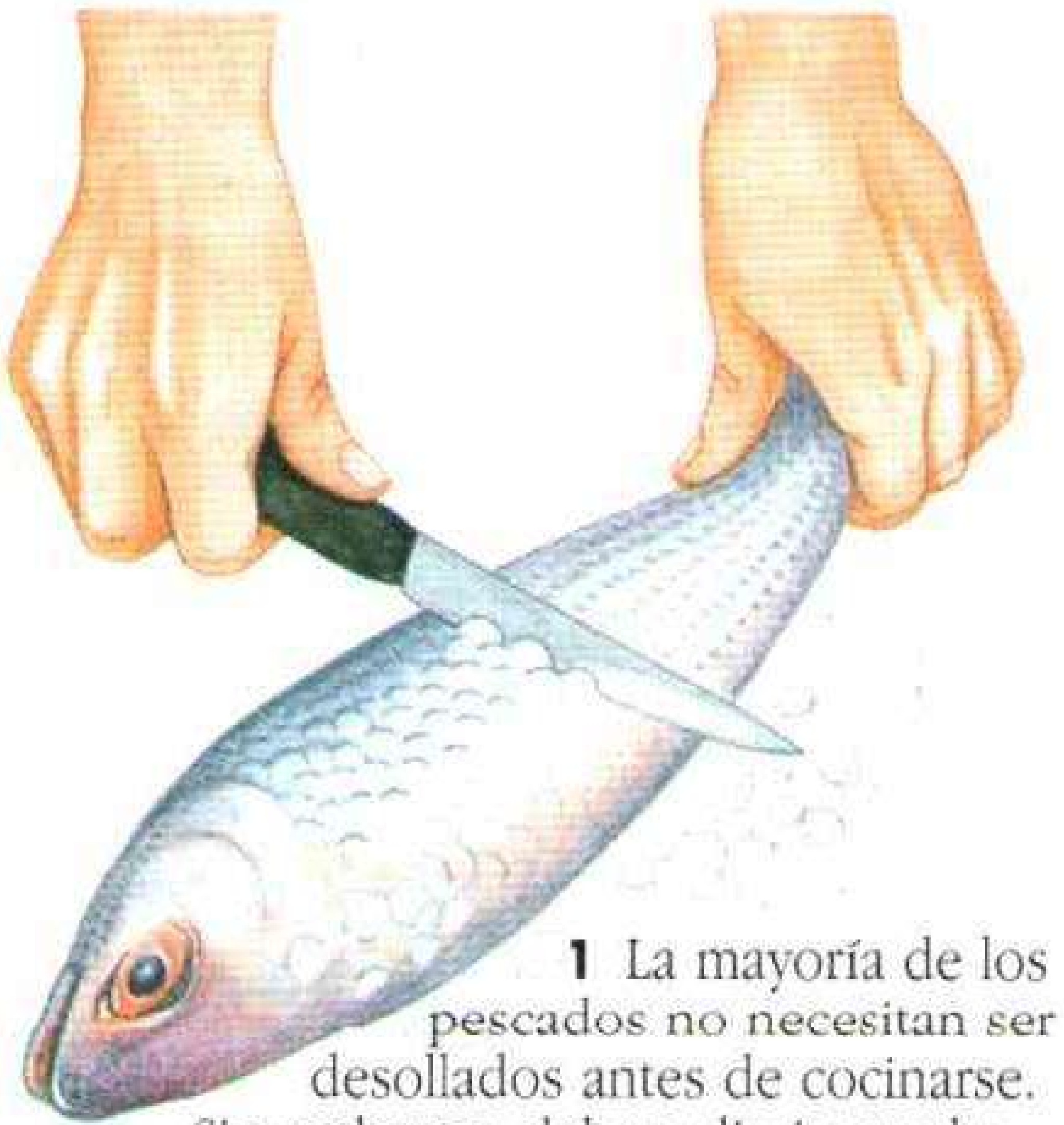

1 La mayoría de los pescados no necesitan ser desollados antes de cocinarse. Sin embargo, deben eliminarse las escamas. Raspe las escamas desde la cola hacia la cabeza con el cuchillo, deslizando el cuchillo en dirección opuesta a su propio cuerpo.

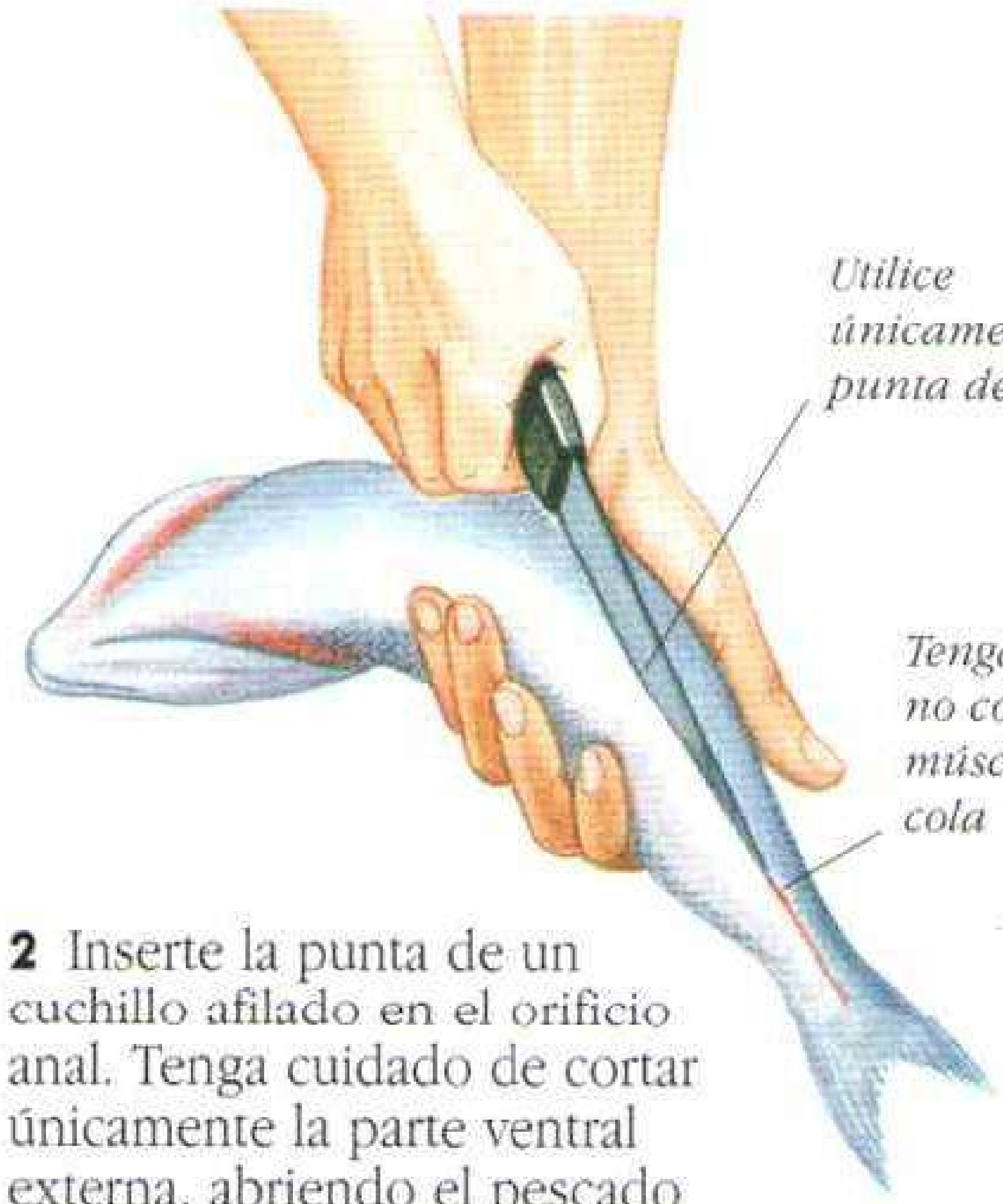

Utilice únicamente la punta del cuchillo

Tenga cuidado de no cortar los músculos de la cola

2 Inserte la punta de un cuchillo afilado en el orificio anal. Tenga cuidado de cortar únicamente la parte ventral externa, abriendo el pescado justo hasta las agallas. No corte los músculos de la cola ni perfore los órganos internos.

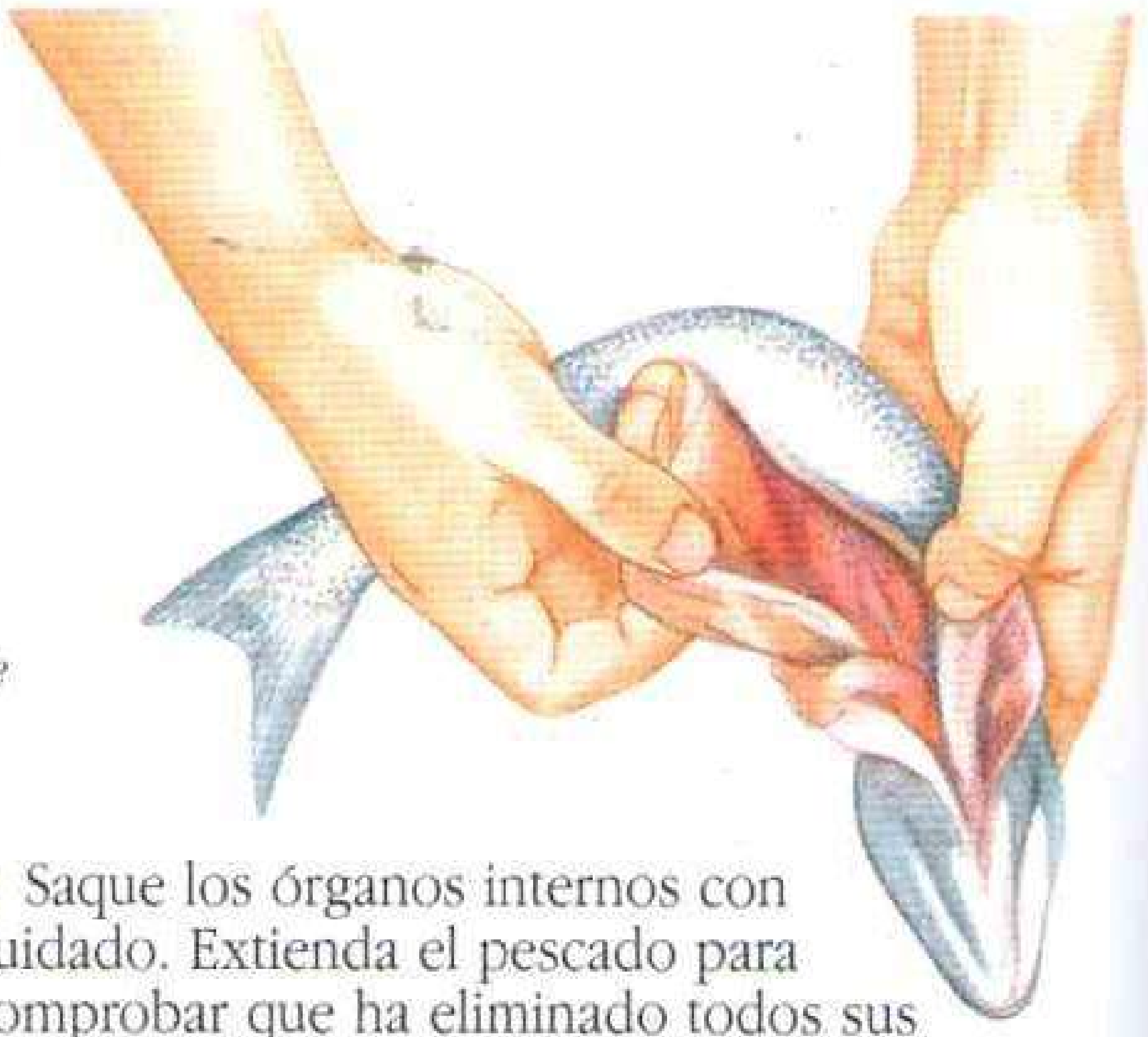

3 Saque los órganos internos con cuidado. Extienda el pescado para comprobar que ha eliminado todos sus órganos. Guarde las huevas (huevos en las hembras, o lechaza en los machos) para comer, y después lave el pescado por dentro y por fuera.

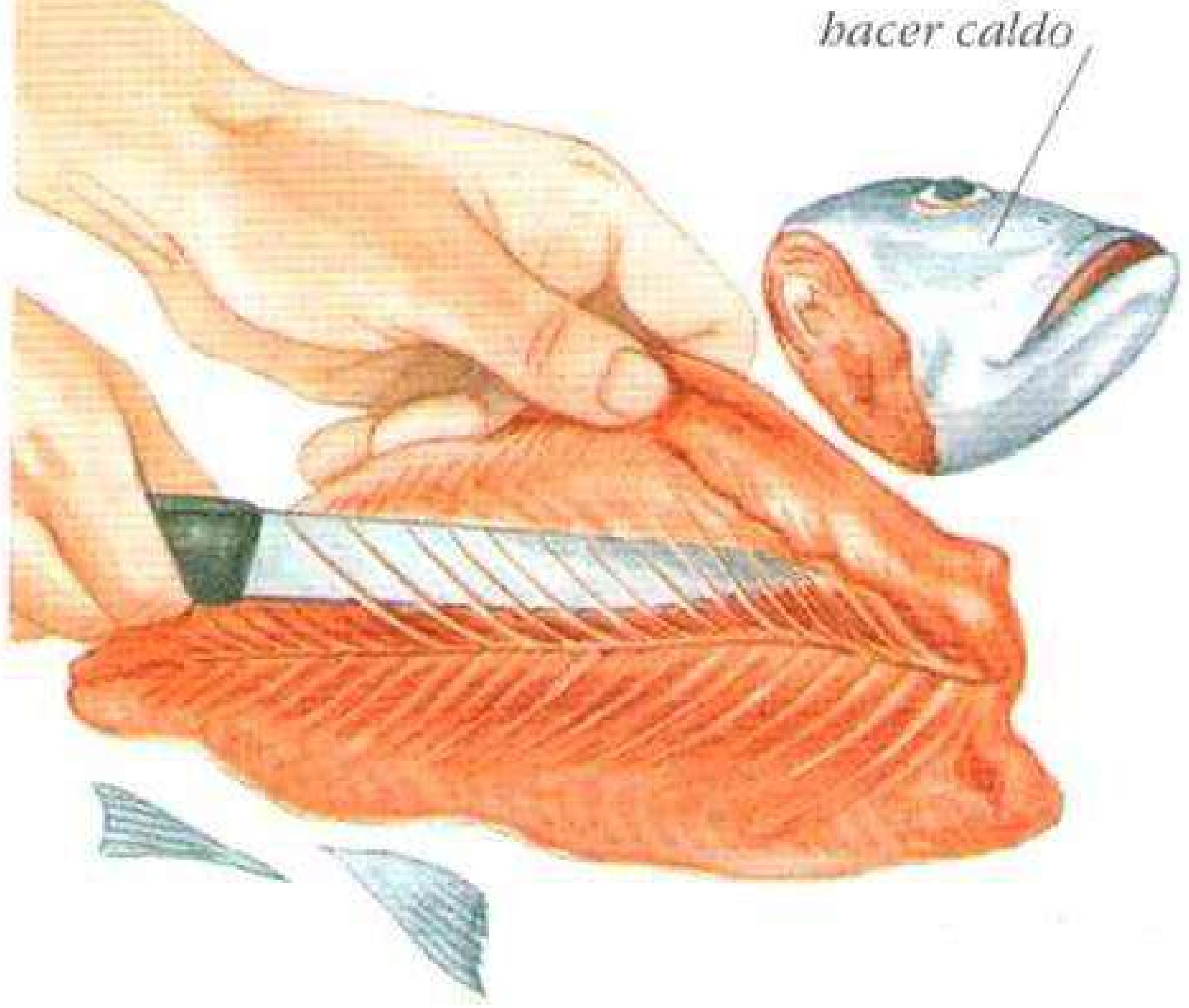

Corte la cabeza y utilícela para hacer caldo

4 Corte la cabeza, la cola y las aletas. Abra el pescado y separe las espinas de la carne con la punta del cuchillo. Puede ser más fácil eliminar las espinas una vez el pescado esté cocido.

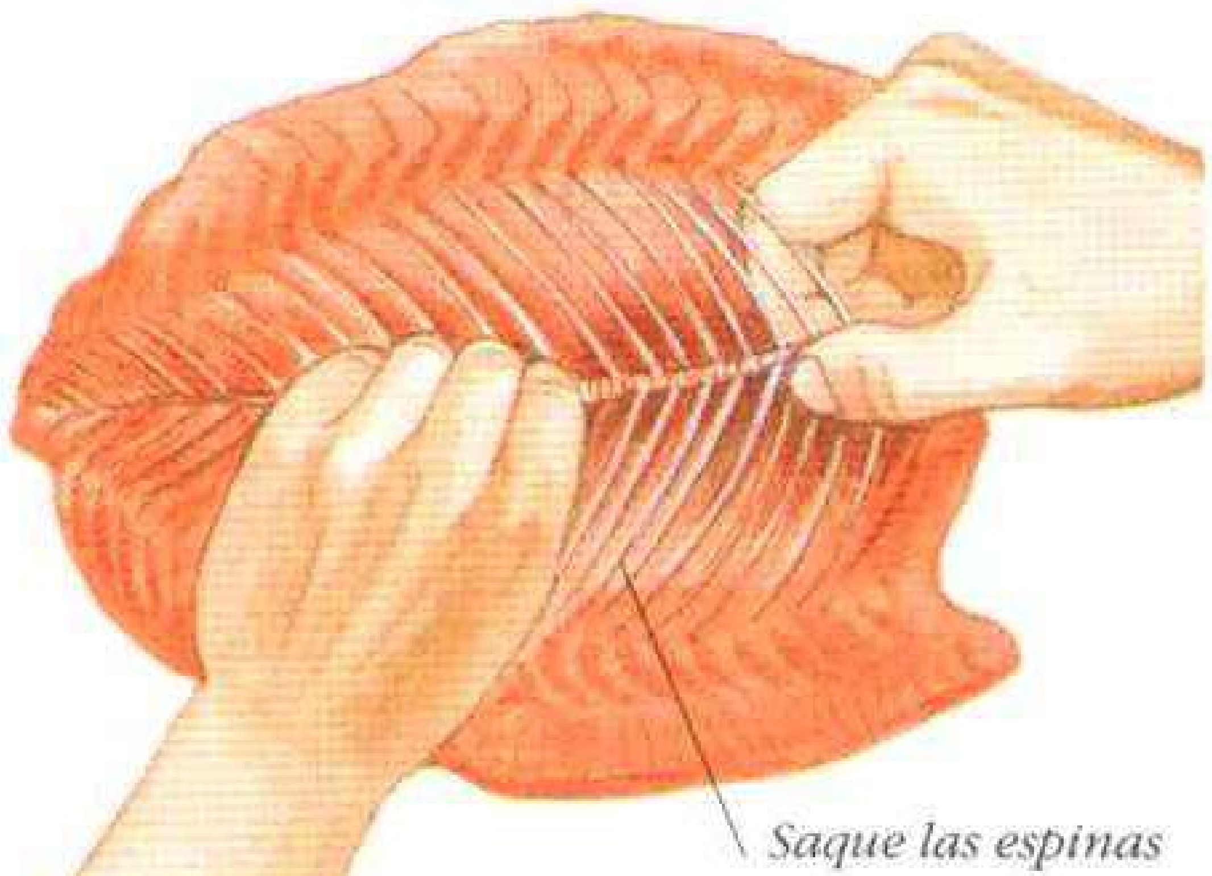

Saque las espinas

5 Separe la cabeza de la espina de la carne con la punta de su cuchillo. Después, despacio pero con firmeza, separe las espinas de la carne en una sola pieza. Si las espinas no salen bien limpias, tendrá que limpiar la carne de ellas con el cuchillo.

CÓMO DESOLLAR EL PESCADO

No todas las especies de pescado necesitan ser desolladas antes de comerse, en particular las más pequeñas. Sin embargo, algunos peces segregan en su piel mucosidades urticantes o venenosas. Otros, como los bagres o los tiburones, tienen pieles muy duras. Es pues recomendable eliminar y desechar la piel antes de comer el pescado.

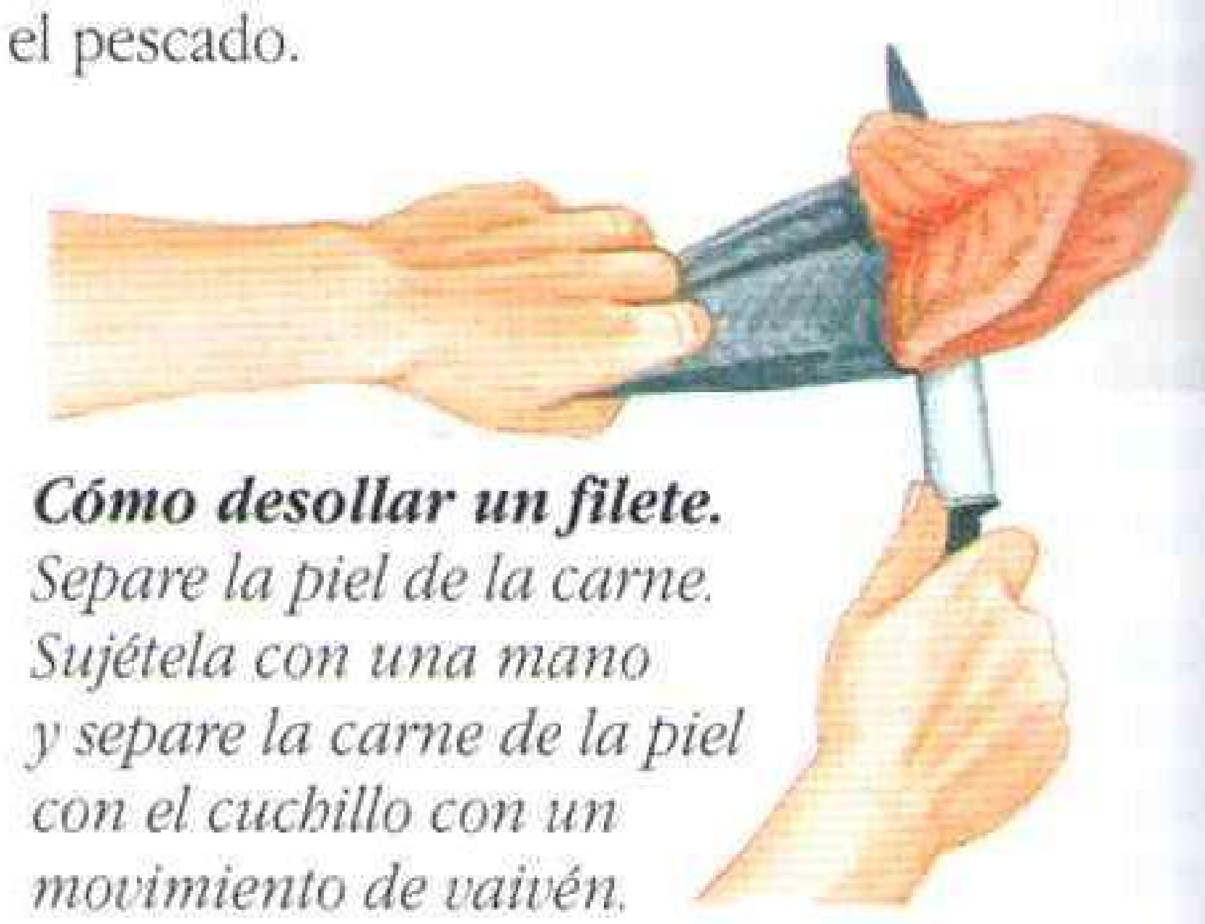

Cómo desollar un filete. *Separe la piel de la carne. Sujétela con una mano y separe la carne de la piel con el cuchillo con un movimiento de vaivén.*

CÓMO ESPETAR EL PESCADO PARA COCINARLO

Es recomendable comer el pescado inmediatamente después de ser pescado. Los métodos más sencillos y rápidos de cocción son los que proporcionan el pescado con mejor sabor. El pescado no requiere mucha cocción. Áselo sobre un fuego de ascuas o empálelo en un espetón *(véase página 116)*. Los peces grandes pueden freírse en una sartén, o cocerse al vapor o en un horno de barro *(véase página 117)*.

1 Abra el pescado limpio, y ensarte dos palos delgados como soporte a lo largo de ambos extremos. La madera verde no se quemará con tanta facilidad. No utilice madera con gran contenido en savia porque podría chisporrotear.

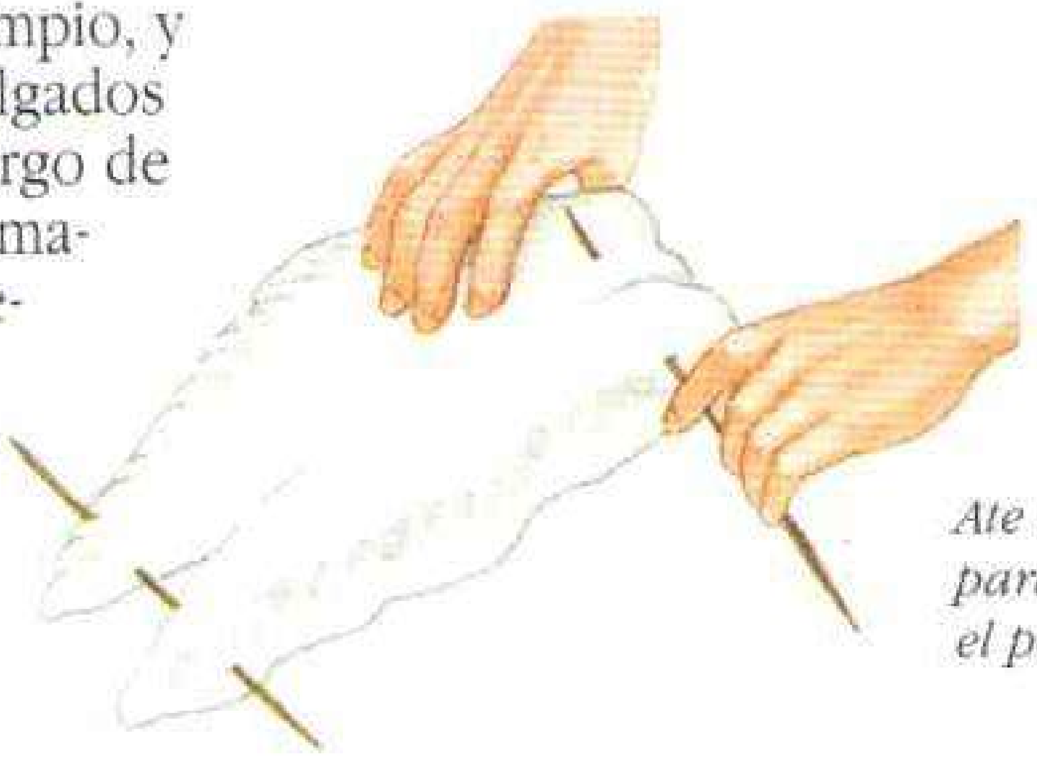

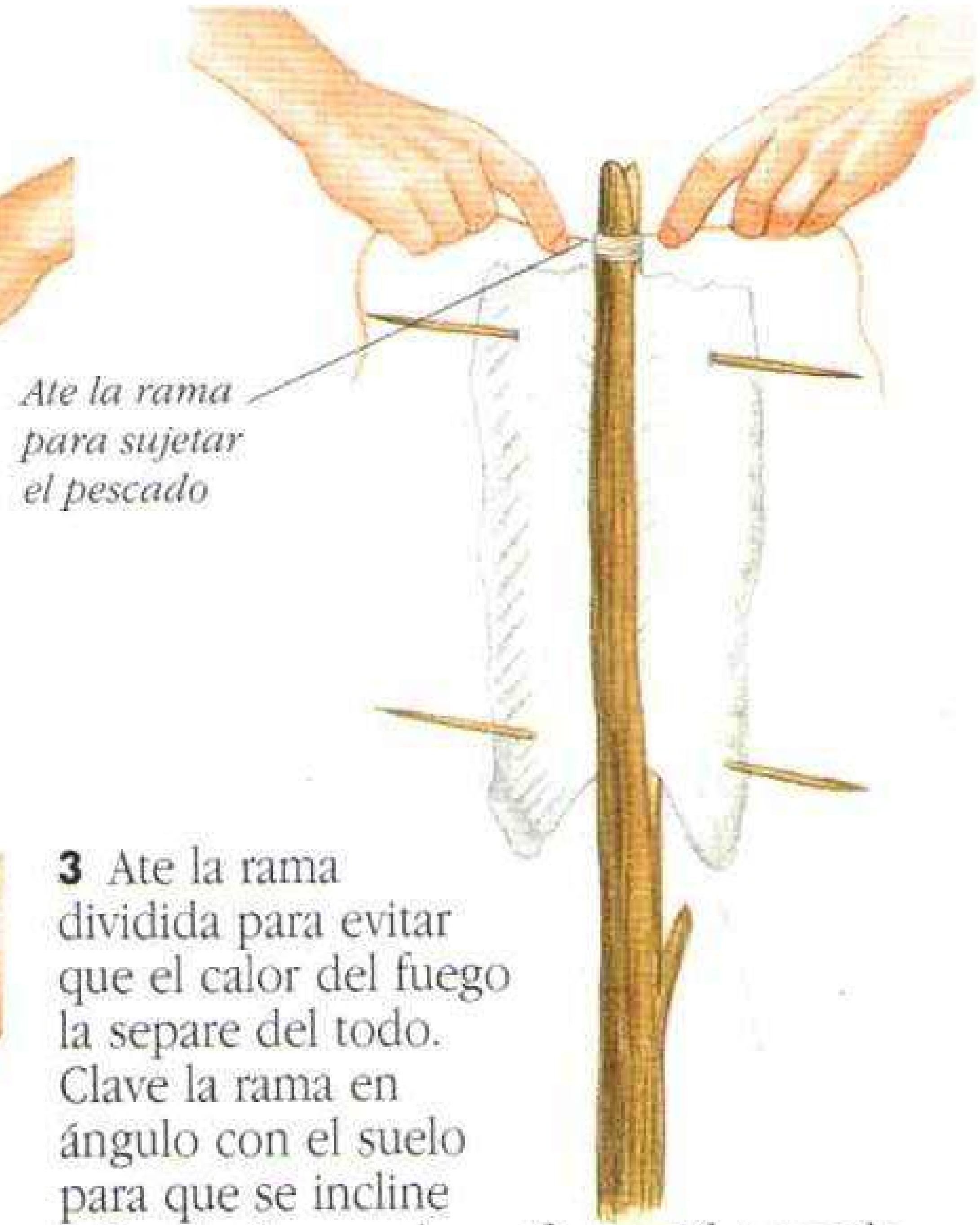
Ate la rama para sujetar el pescado

2 Parta una rama joven longitudinalmente hasta la mitad. Inserte el pescado limpio en la hendidura y extiéndalo, separando ligeramente la parte central del filete para que se cueza uniformemente.

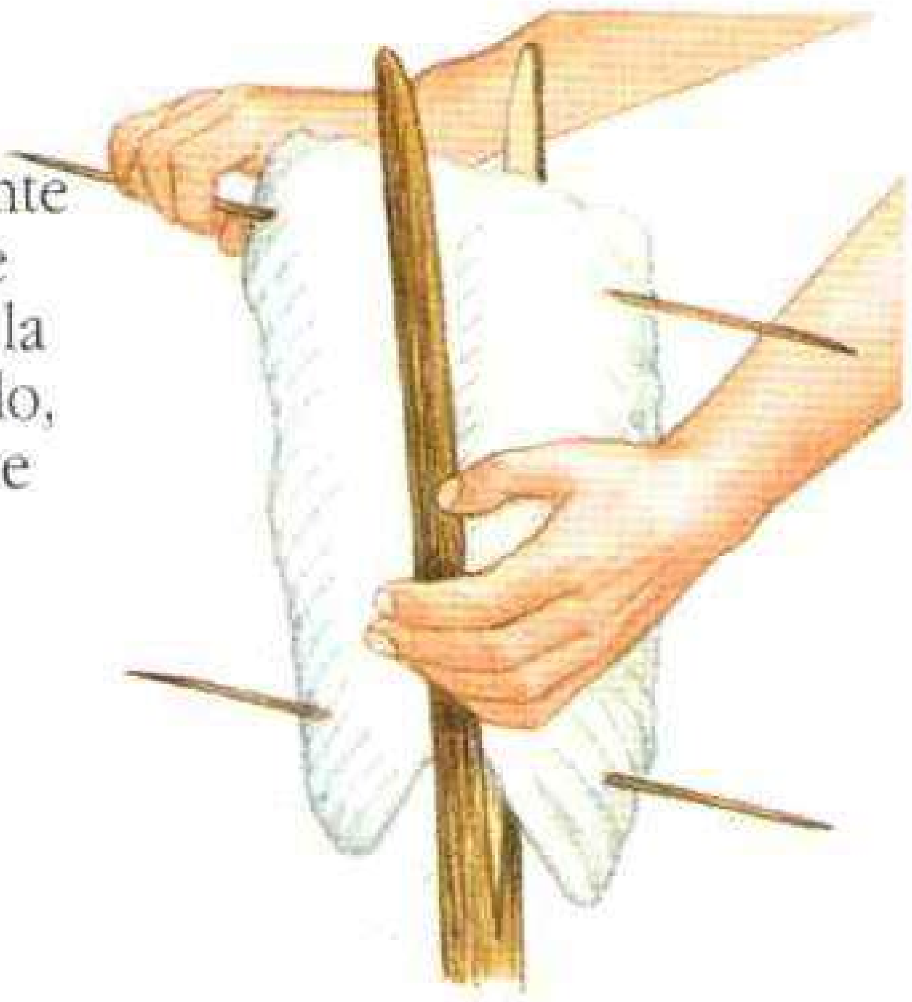

3 Ate la rama dividida para evitar que el calor del fuego la separe del todo. Clave la rama en ángulo con el suelo para que se incline sobre las brasas de un fuego. El pescado estará cocido cuando la carne parezca separarse.

CÓMO COCINAR PESCADOS PEQUEÑOS

Fría o ase los pescados pequeños enteros sin limpiarlos —no es necesario limpiar cualquier pieza de menos de 15 cm de longitud. Una vez limpios, clávelos en un palo verde y áselos sobre las ascuas o en una sartén. No elimine las cabezas y colas, ya que ayudan a conservar el pescado entero.

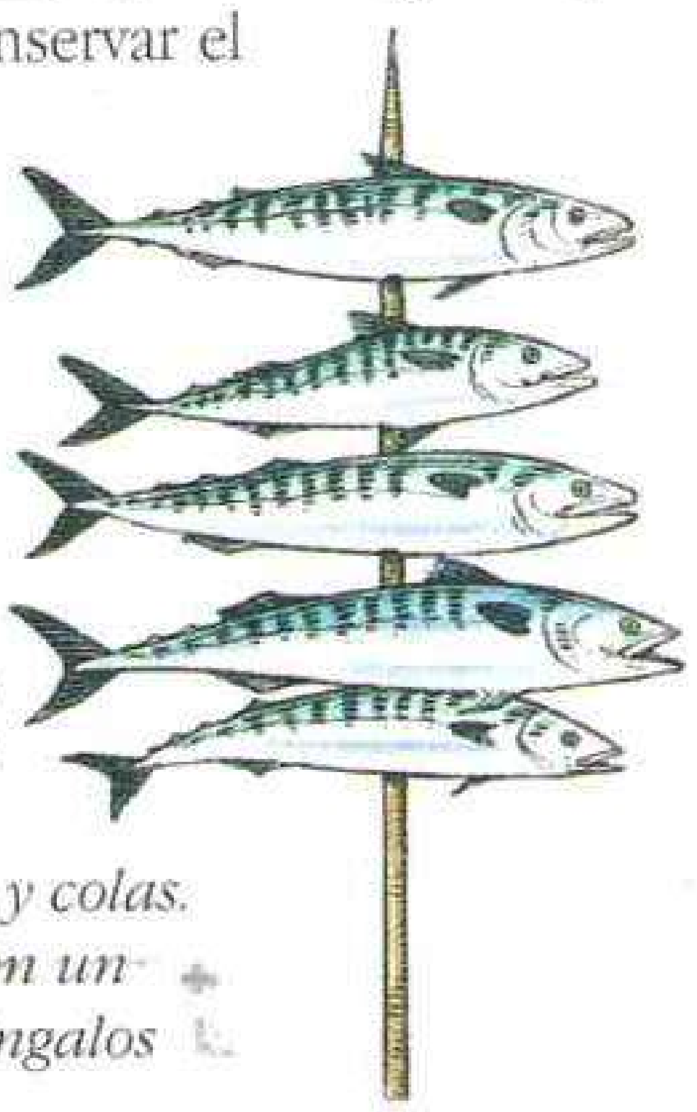

Pescado espetado. *Limpie los pescados pequeños, pero no elimine las cabezas y colas. Clave los pescados en un palo. Después, sosténgalos sobre un fuego.*

CÓMO DESOLLAR UNA ANGUILA

Las anguilas constituyen un plato excelente, pero deben destriparse, desollarse y limpiarse antes de cocinarlas. Su longitud y su piel viscosa y escurridiza hacen que el desollarlas sea muy difícil, a menos que efectúe esta operación inmediatamente después de pescarlas. Limpie una anguila utilizando el método mostrado en la página anterior y desóllela como se muestra a continuación.

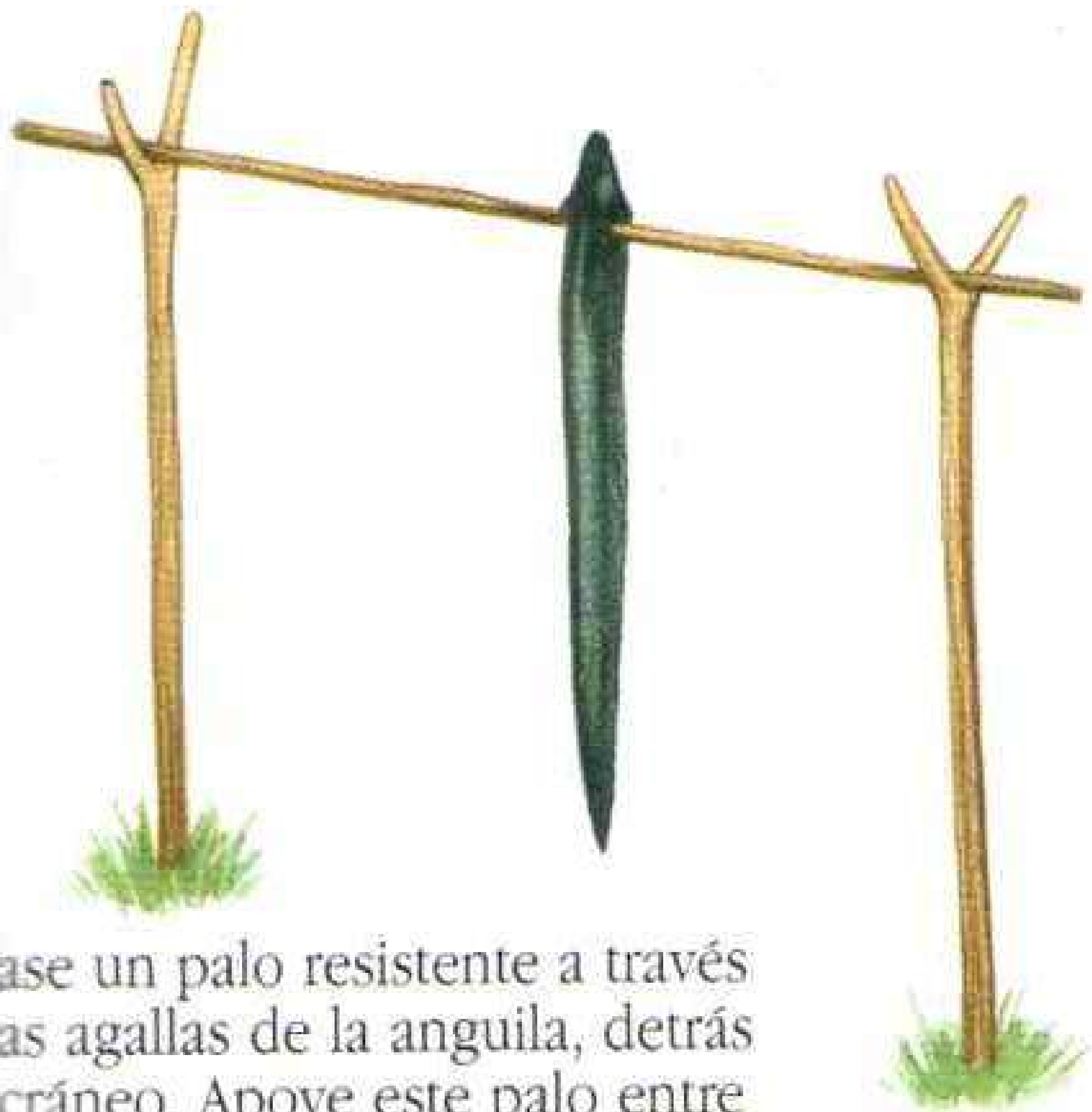

1 Pase un palo resistente a través de las agallas de la anguila, detrás del cráneo. Apoye este palo entre otros dos, verticales y bífidos.

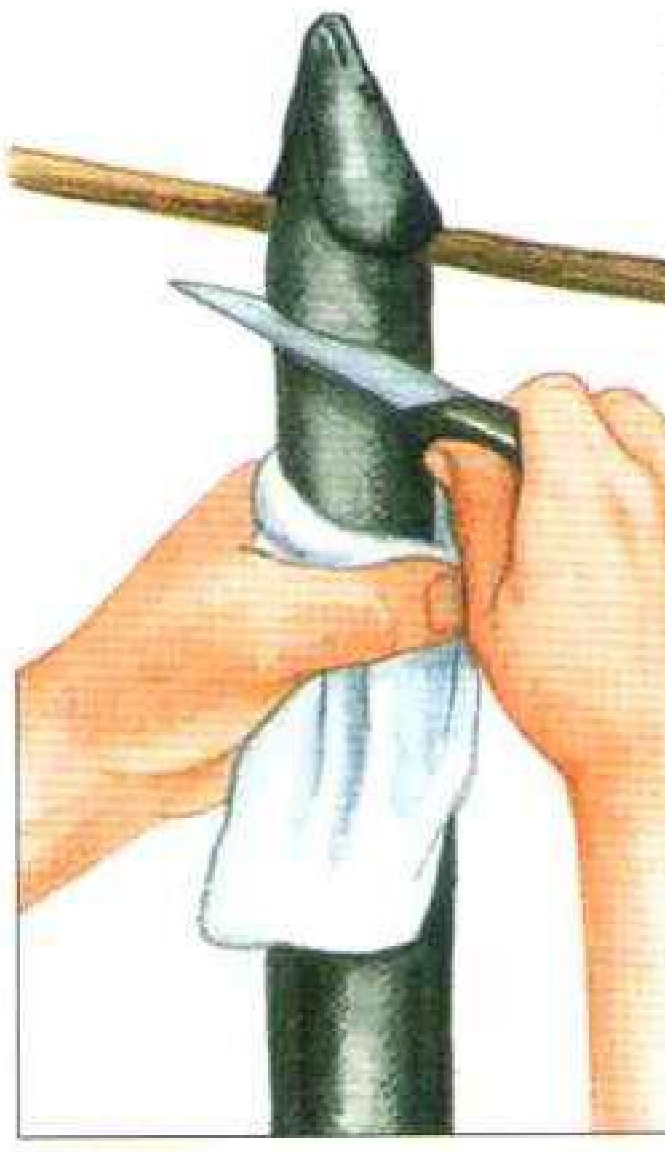

2 Sujete la anguila con un trapo y corte la piel alrededor de la cabeza. (Aunque la anguila esté muerta, aún puede llegar a moverse.)

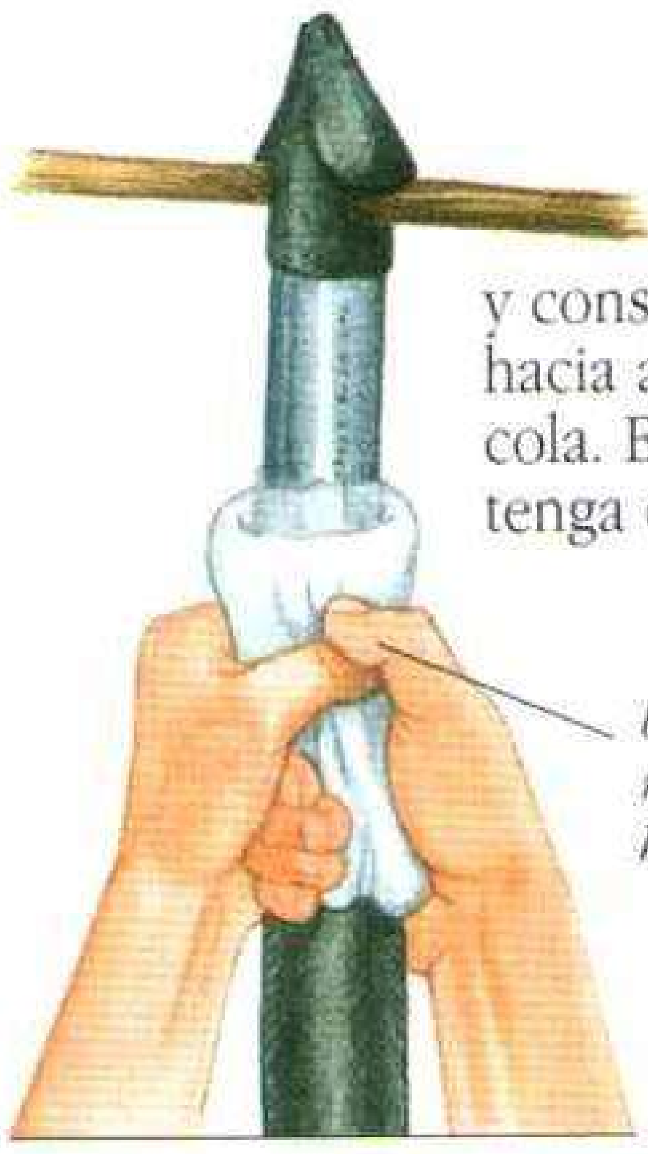
Utilizando ambas manos, estire la piel hacia abajo

3 Separe la parte superior de la piel de la carne. Después, con un movimiento firme y constante, deslice la piel hacia abajo, en dirección a la cola. Es muy probable que tenga que usar ambas manos.

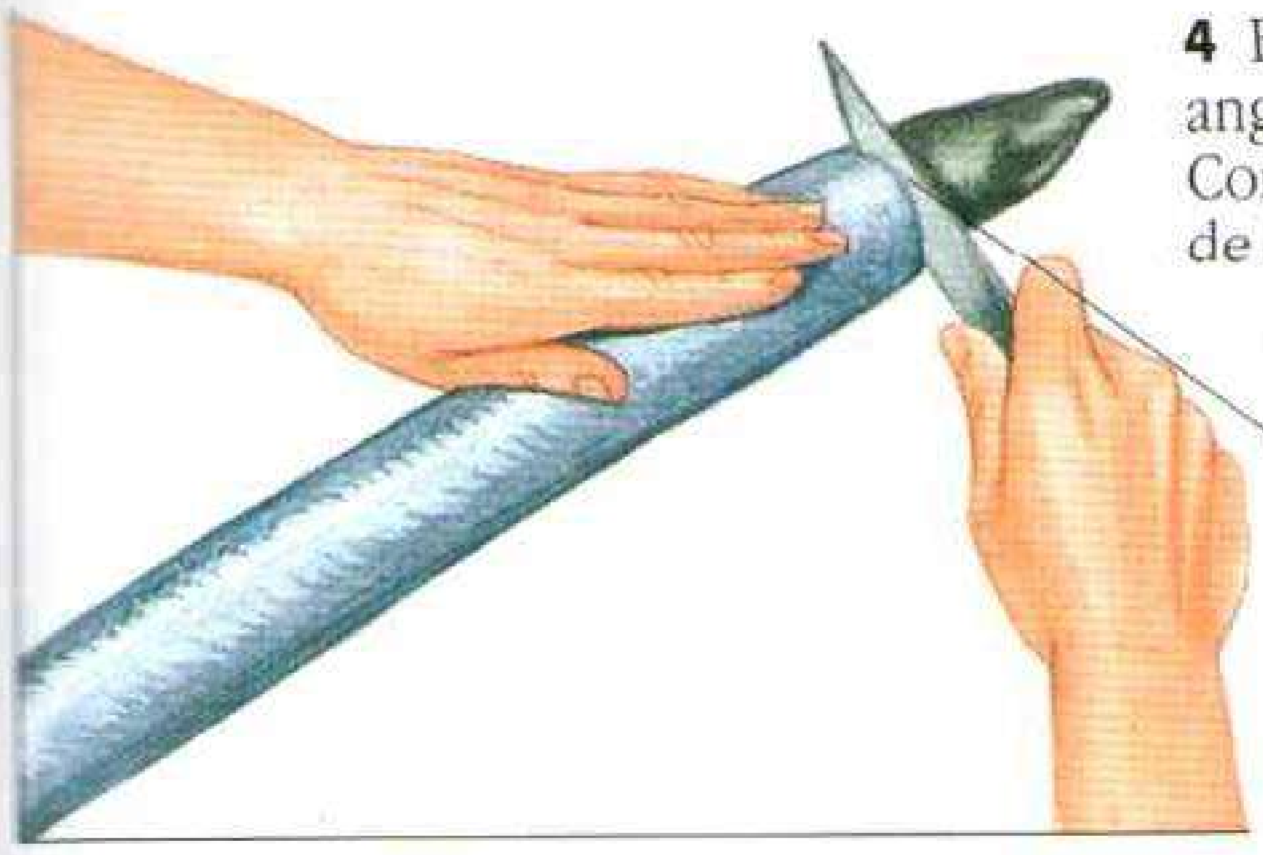
Corte la cabeza de la anguila

4 Elimine la piel y quite la anguila del palo que la sujetaba. Corte la cabeza y la parte final de la cola.

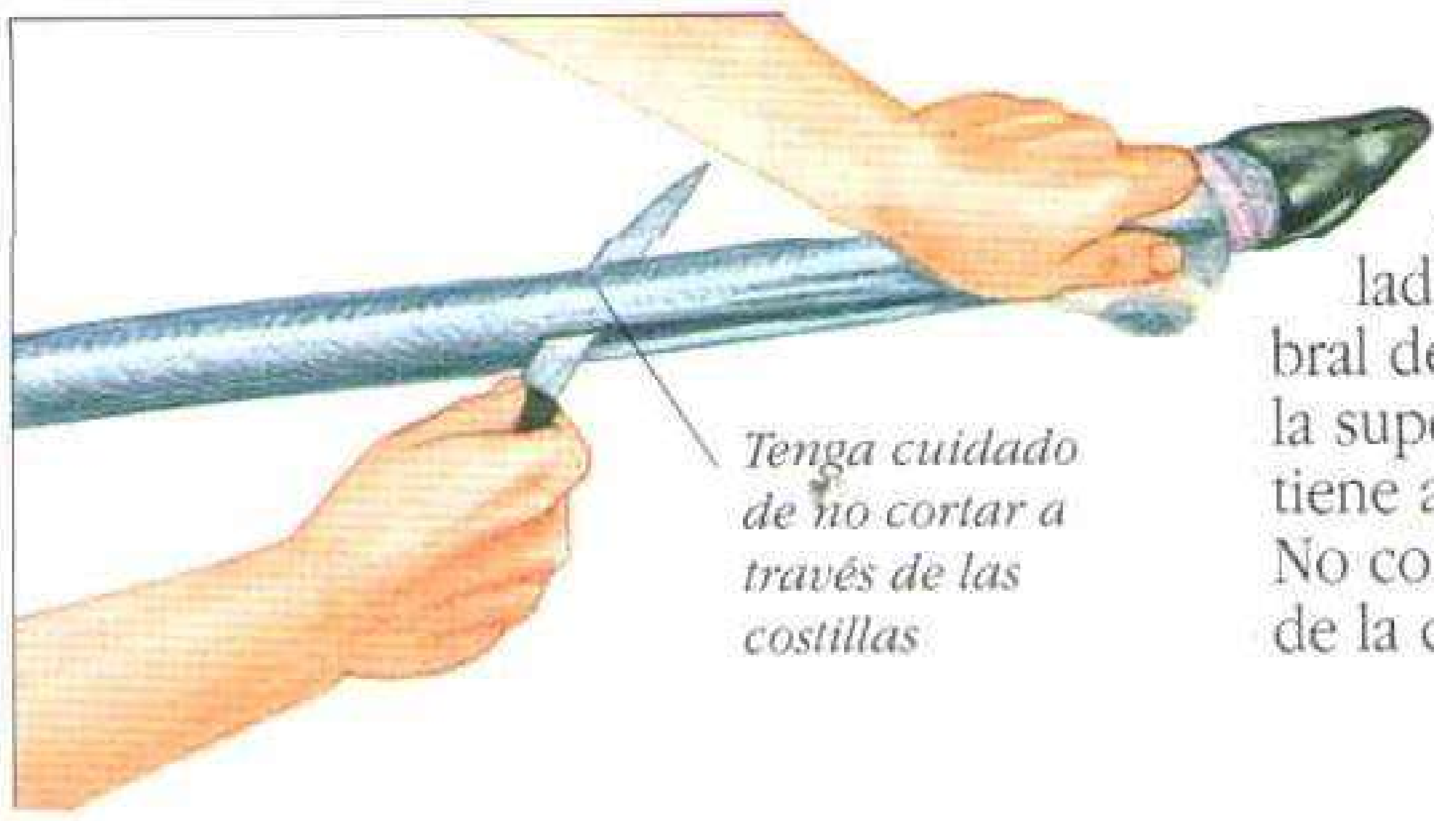
Tenga cuidado de no cortar a través de las costillas

5 Sujetando la anguila con un trapo, recorra con un cuchillo afilado la columna vertebral de forma paralela a la superficie sobre la cual tiene apoyada la anguila. No corte demasiado cerca de la columna.

UTENSILIOS DE COCINA

LOS UTENSILIOS BÁSICOS de cocina en un campamento son un cazo, una cocina o fuego donde guisar los alimentos, y una cuchara de madera. Al cocinar, tanto el cazo como la cocina o el fuego deben protegerse del viento, y es recomendable que el cazo disponga de una tapa. De esta manera economizará en combustible. La higiene y el orden al cocinar resultan fundamentales. Lave sus manos como primer paso, y después saque de su mochila únicamente lo indispensable para preparar la comida. La mayoría de los alimentos para mochileros es preferible cocinarla y comerla en un solo guiso, preparada con mucha agua y en un mismo cazo. Los mochileros experimentados tienden a mezclar la fruta, el postre y los platos principales. De esta manera ahorran combustible y lavan menos platos.

Cocina al aire libre. *Mantenga la cocina resguardada del viento y asegúrese de que no pueda volcar. Tenga cuidado de no tocar un cazo caliente con las manos desnudas.*

CAZOS Y SARTENES

Si come directamente de un cazo, en especial si cocina en un fuego al aire libre, el riesgo de quemarse los dedos, los labios, la boca o la lengua es mayor. El empleo de vasos y platos de plástico evita este riesgo. Resulta innecesario cargar con una gran cantidad de recipientes para cocinar, a menos que viaje con un grupo numeroso o en un vehículo. Elija el tipo de cazo que utilizará con mayor frecuencia, y cuide de él.

Tazón. *Un tazón de plástico resistente es de gran utilidad. Debe tener asimismo un asa resistente.*

Plato. *Un plato permite separar una porción de comida mientras el resto se mantiene caliente dentro del cazo.*

Escudilla. *Una escudilla puede utilizarse para comer los cereales del desayuno, unas gachas o un estofado.*

CUBIERTOS

Recuerde que los cubiertos deben ser de uso particular y que debe mantenerlos escrupulosamente limpios. Los cubiertos sucios pueden provocar úlceras bucales y otras infecciones. Para limpiarlos, sumérjalos en agua hirviendo siempre que sea posible. Si pierde o rompe su cuchillo, tenedor o cuchara, puede tallar uno nuevo en madera *(véase página siguiente)*. Puede improvisar los cubiertos y la vajilla a partir de ramas, corteza, hongos de árbol y huesos.

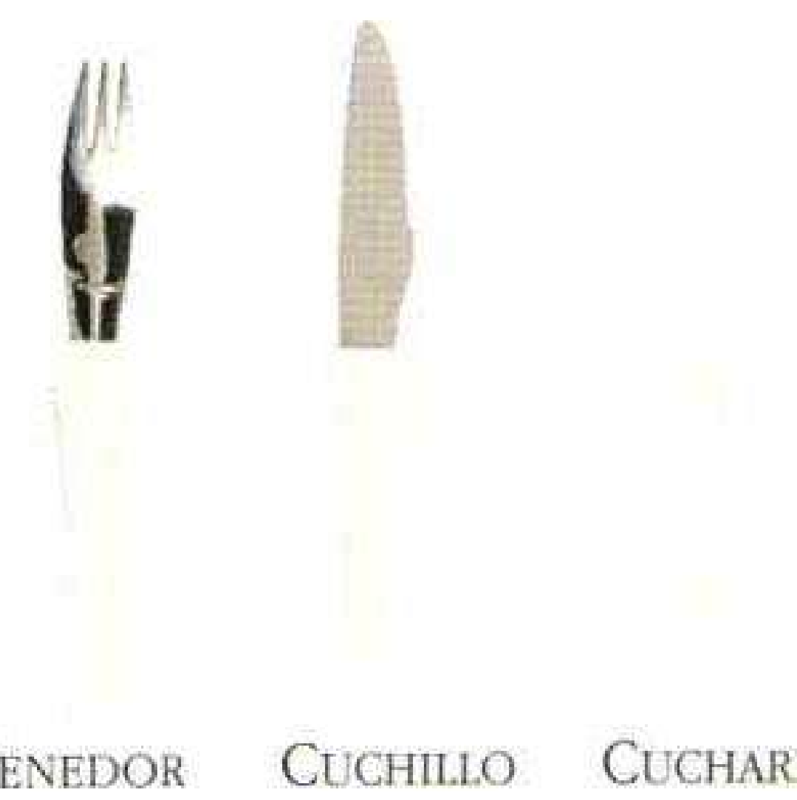

TENEDOR CUCHILLO CUCHARILLA

ESTUCHE PARA CUBIERTOS

Sartén antiadherente plana. *Una sartén antiadherente es de gran ayuda en un campamento, pero tenga cuidado de no dañar la capa que la recubre, ya que entonces los alimentos se pegarían.*

Sartén antiadherente honda. *Una sartén honda es de utilidad al cocinar para más de una o dos personas.*

Hervidor. *Un hervidor de poco peso le permite utilizar el resto de sus recipientes para cocinar, y no para calentar el agua para las bebidas calientes.*

Juego de cazos. *Un juego de cazos de buena calidad tiene poco peso y contiene un conjunto de cazos, y en ocasiones incluso un hervidor.*

Sartén. *Se pueden encontrar sartenes ligeras especiales para campamentos. Tenga cuidado de no quemarlas.*

Cazo de aluminio. *Los ca de aluminio calientan l comida rápidamente, pe pueden quemar los alime no se vigilan continuame*

CÓMO HACER UN RECIPIENTE DE BARRO

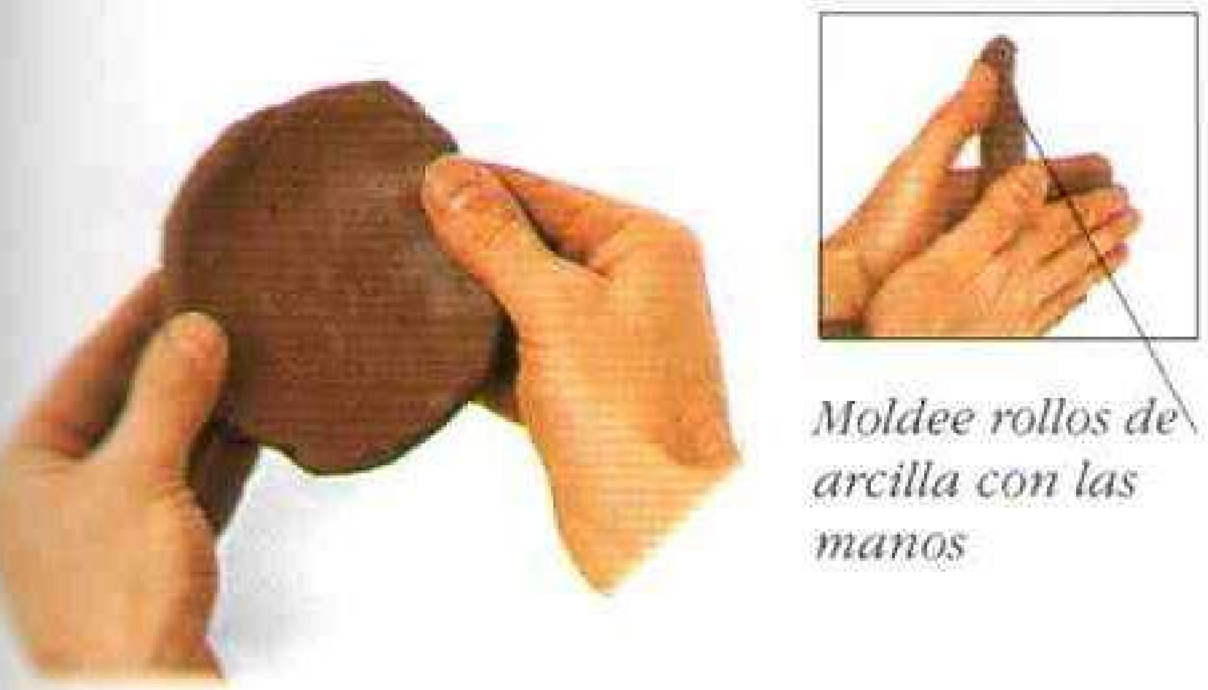

Moldee rollos de arcilla con las manos

1 Moldee una base circular que deberá ser más gruesa que las paredes para dar resistencia y estabilidad al recipiente.

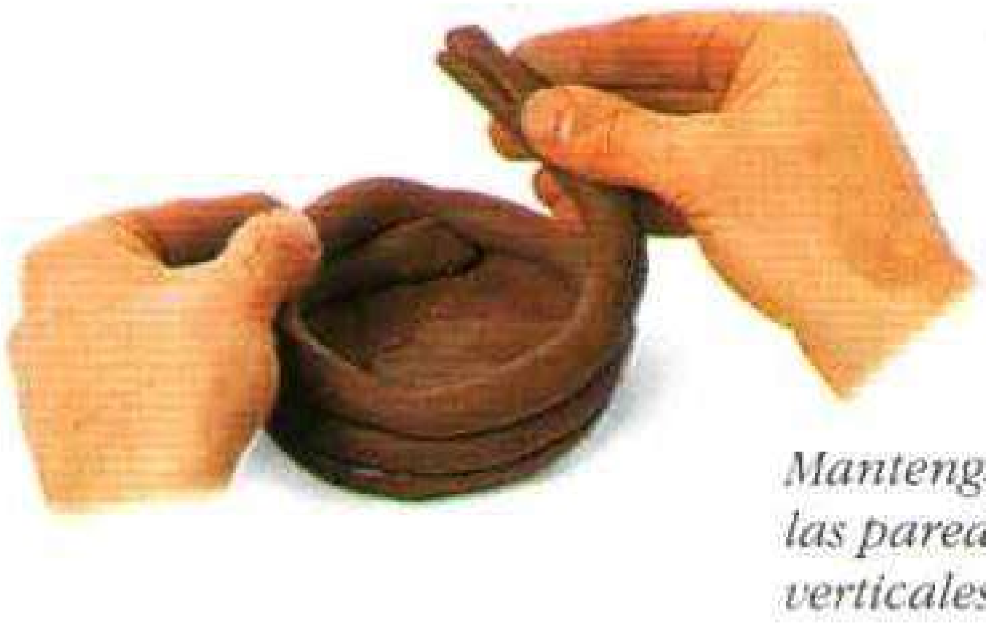

2 Construya las paredes del recipiente con rollos de arcilla previamente moldeados.

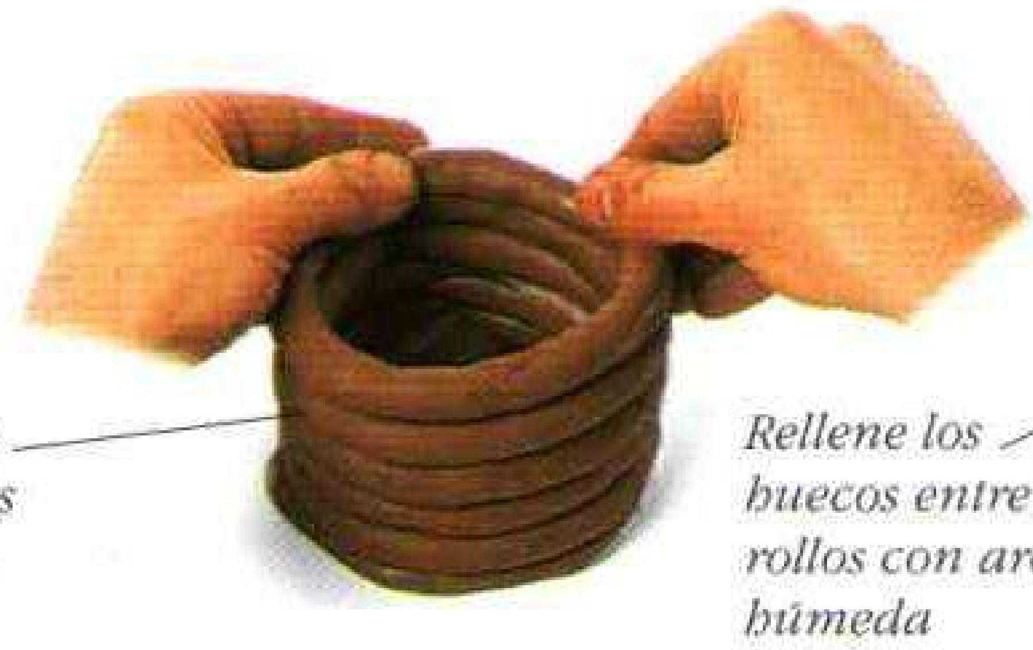

Mantenga las paredes verticales

3 Suavice la boca del recipiente con el pulgar una vez obtenga la altura deseada.

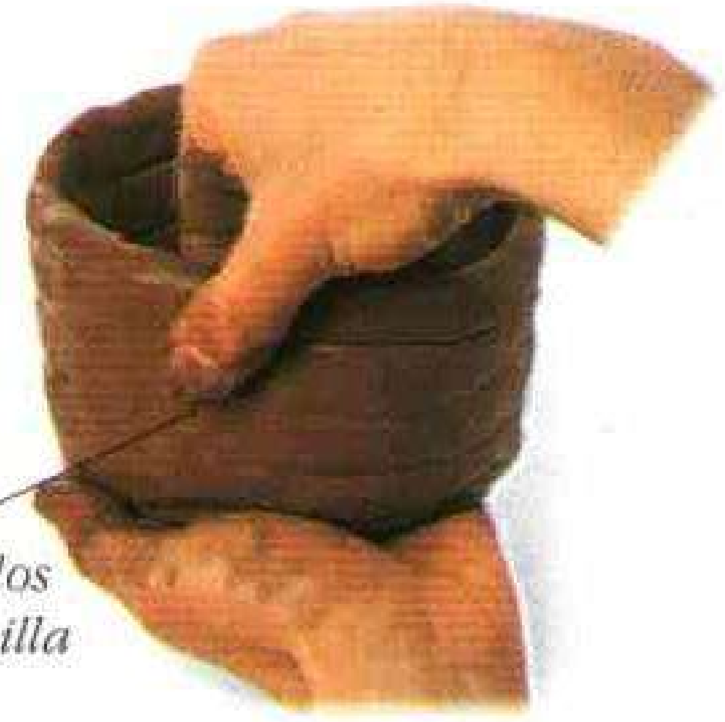

Rellene los huecos entre los rollos con arcilla húmeda

4 Alise el recipiente con agua. Una vez esté seco, póngalo al fuego para que se endurezca.

CÓMO TALLAR UNA CUCHARA DE MADERA

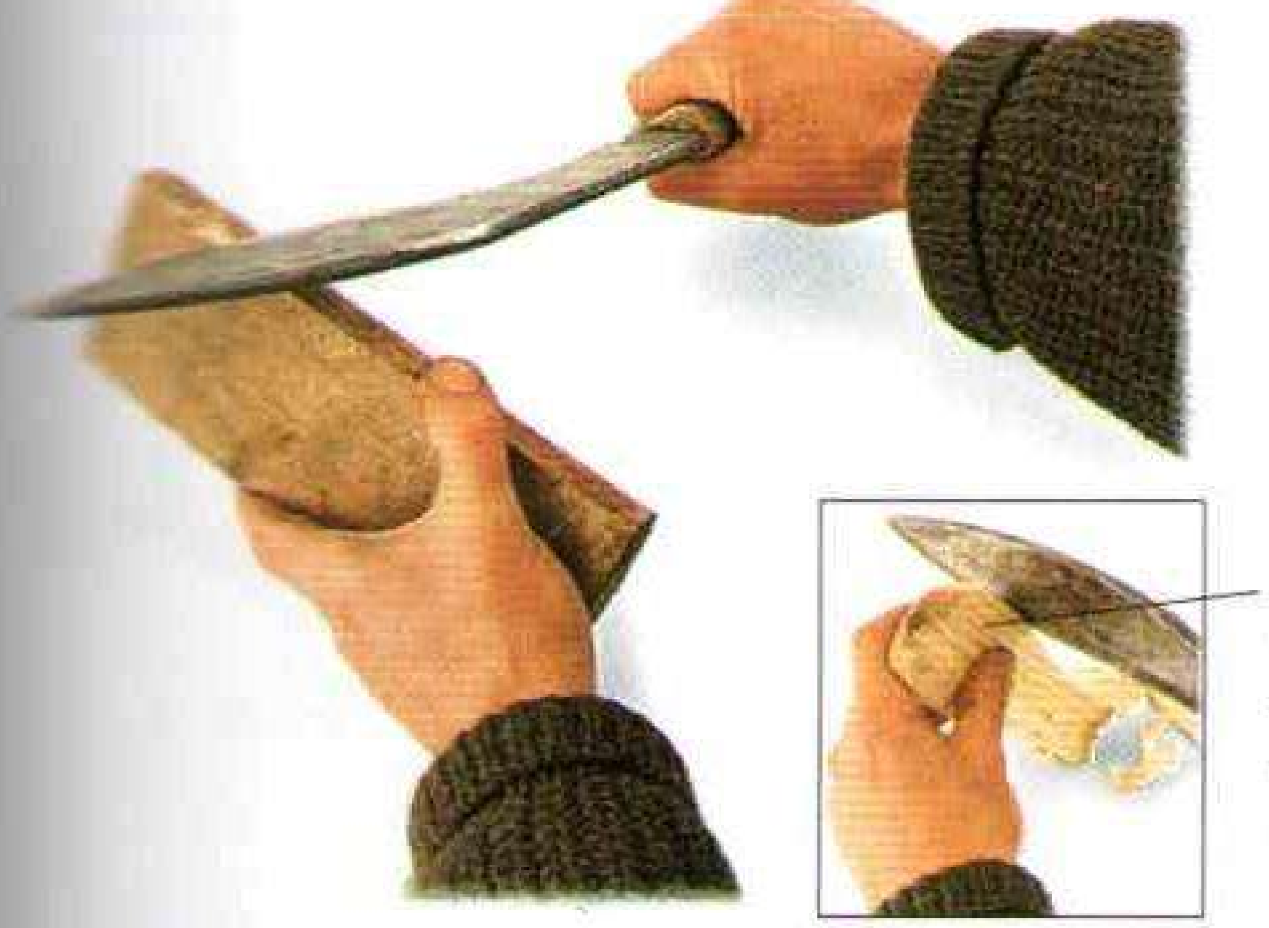

Comience por tallar la forma de la cuchara en un extremo de la madera

1 Corte un trozo de madera de grano fino y comience por tallarla en el sentido longitudinal de la fibra. Evite las maderas resinosas, ya que se rompen con facilidad.

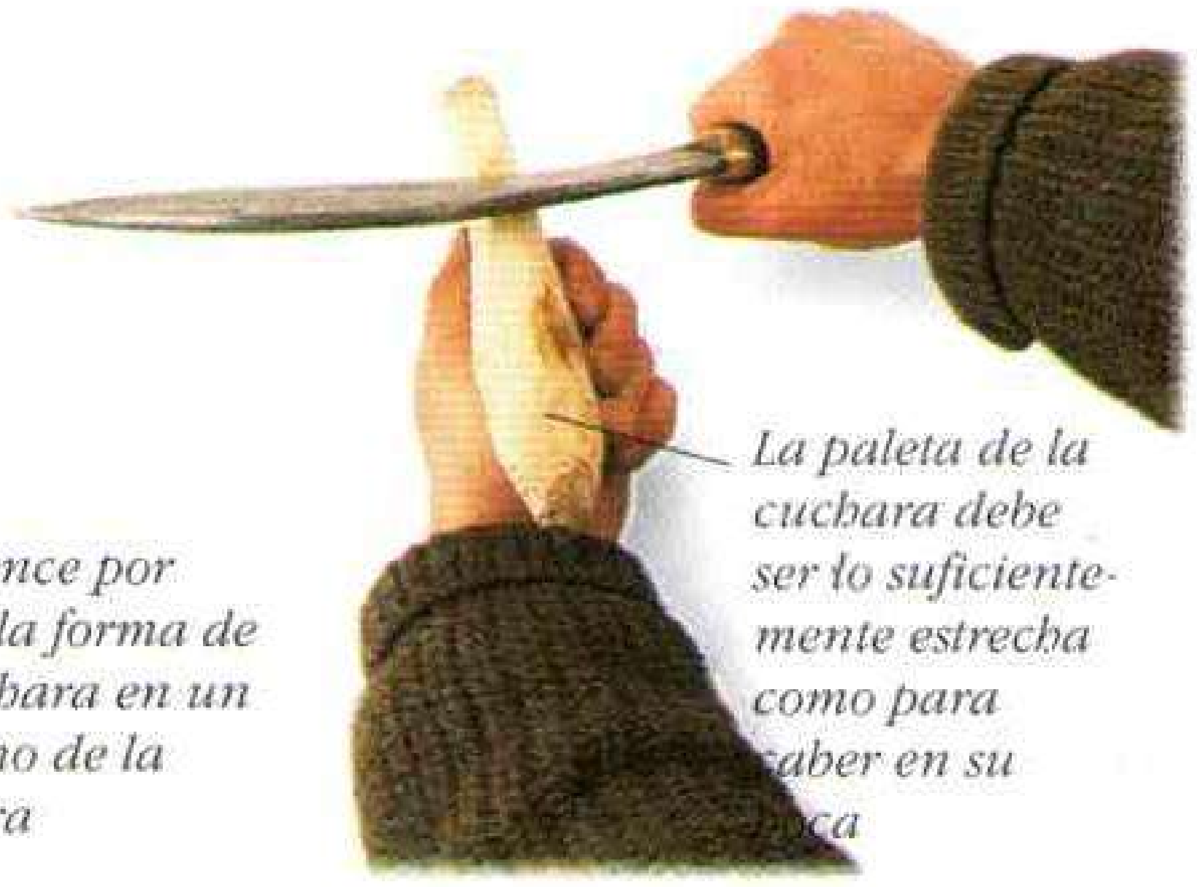

La paleta de la cuchara debe ser lo suficientemente estrecha como para aber en su ca

2 Corte siempre en dirección opuesta a su cuerpo para evitar accidentes. La cuchara debe ser lo suficientemente pequeña como para caber en su boca; en cambio, un mango estrecho no es tan importante.

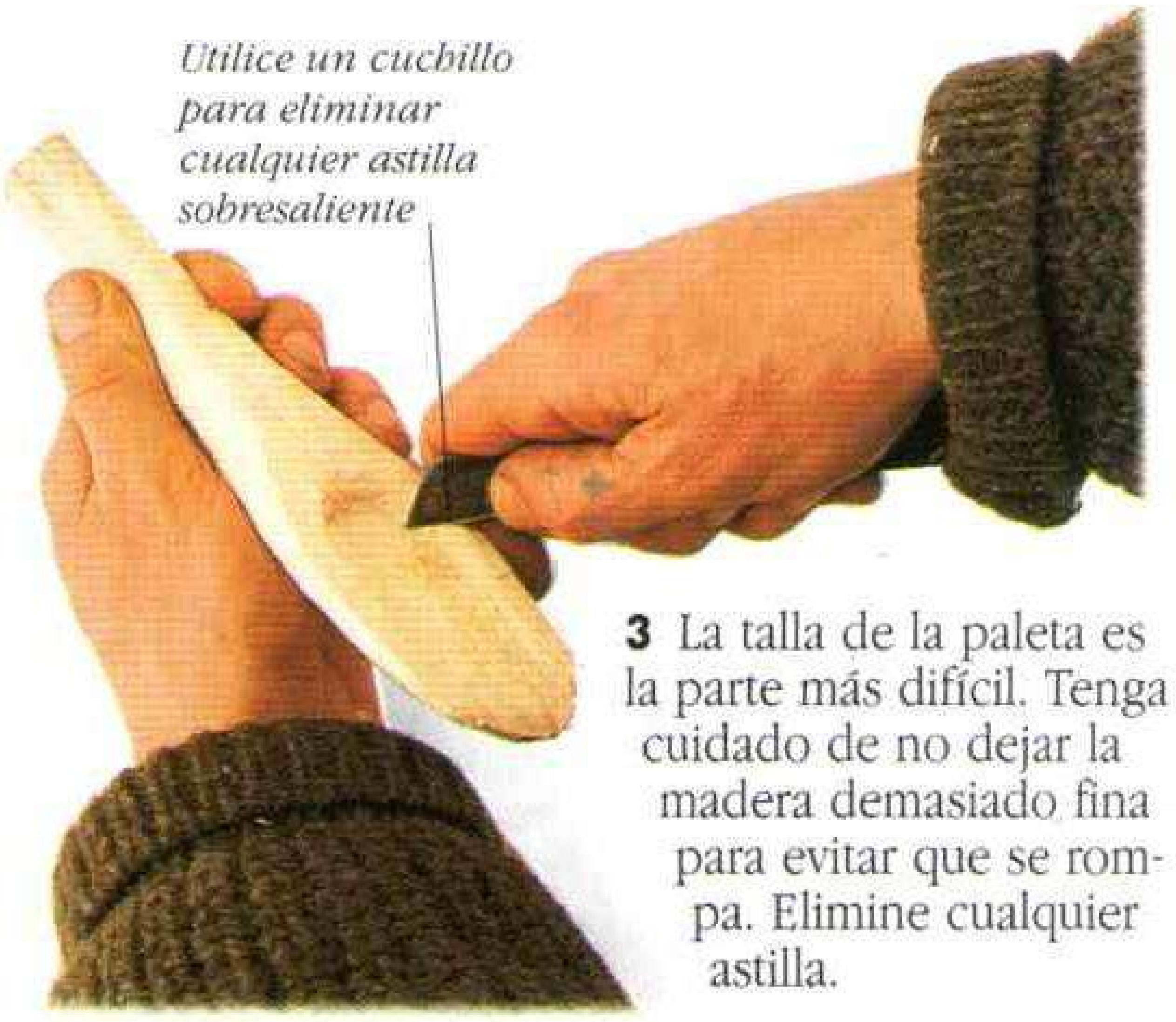

Utilice un cuchillo para eliminar cualquier astilla sobresaliente

3 La talla de la paleta es la parte más difícil. Tenga cuidado de no dejar la madera demasiado fina para evitar que se rompa. Elimine cualquier astilla.

CÓMO CONSTRUIR UN CUENCO DE MADERA

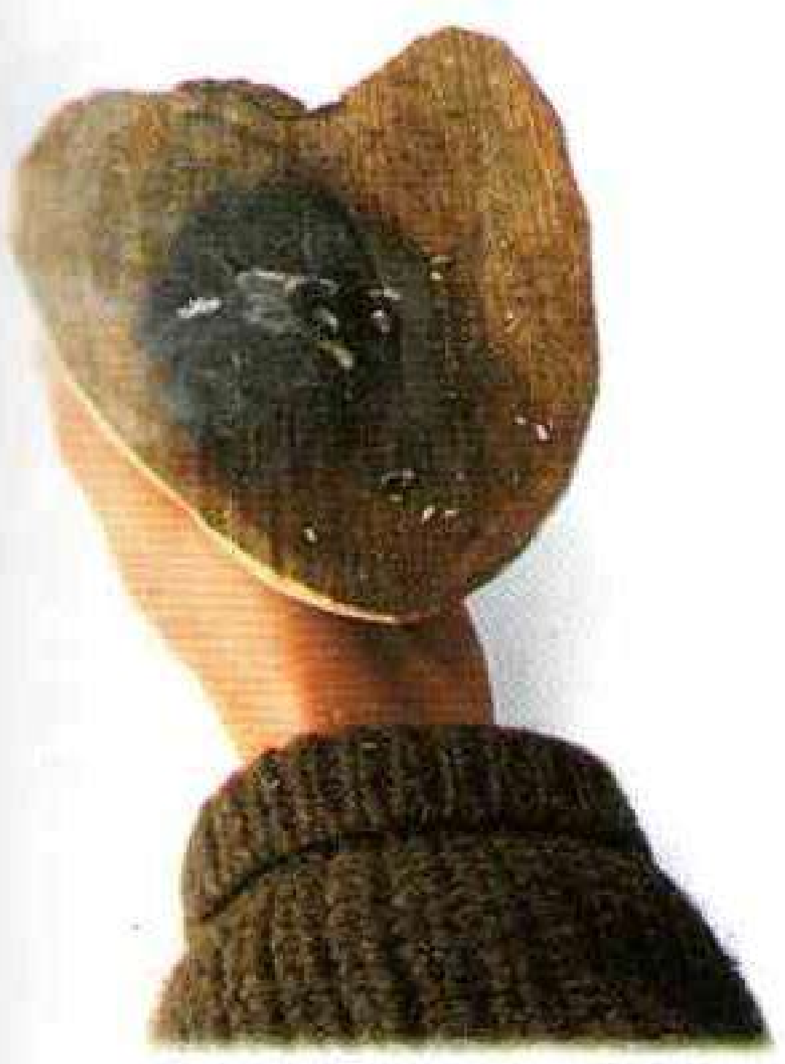

1 Talle un cuenco a partir de un bloque o un nudo de madera. Los nudos de madera pueden ser extremadamente duros, por lo que requerirá largo tiempo para tallarlos. El empleo cuidadoso de las ascuas de un fuego puede ayudarle a acelerar el proceso.

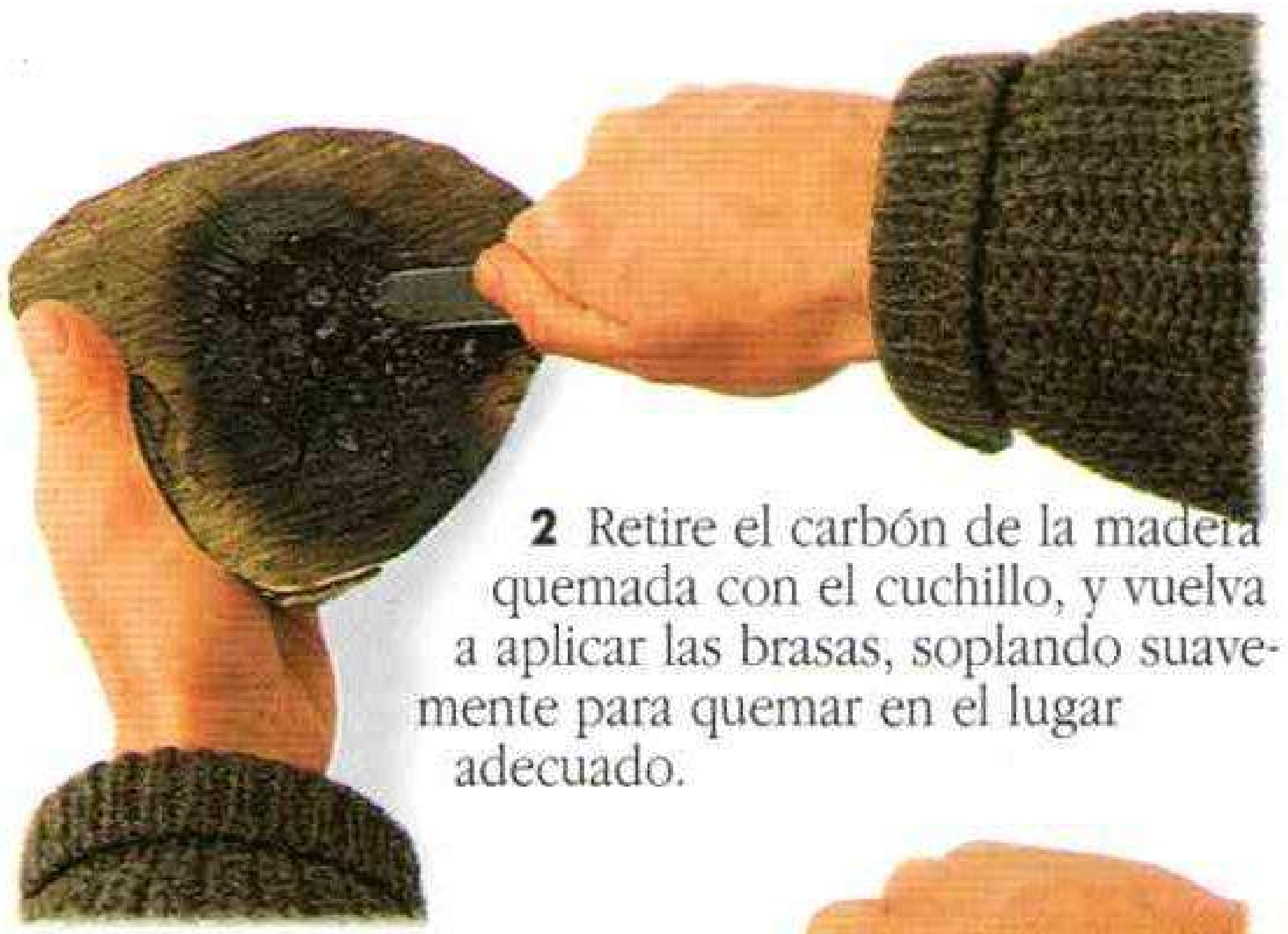

2 Retire el carbón de la madera quemada con el cuchillo, y vuelva a aplicar las brasas, soplando suavemente para quemar en el lugar adecuado.

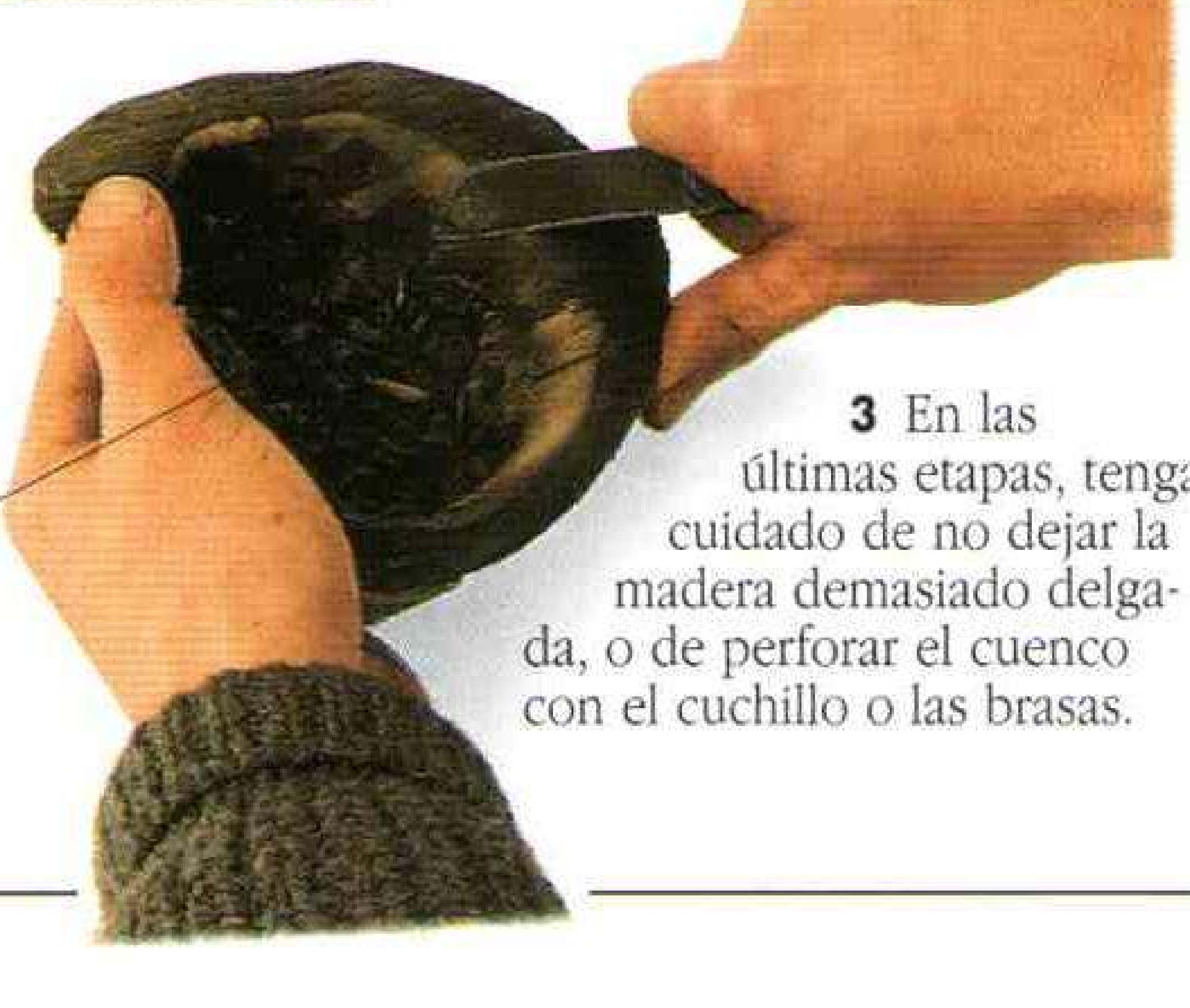

Continúe quemando y tallando para ahuecar el cuenco

3 En las últimas etapas, tenga cuidado de no dejar la madera demasiado delgada, o de perforar el cuenco con el cuchillo o las brasas.

CAZOS Y SARTENES IMPROVISADOS

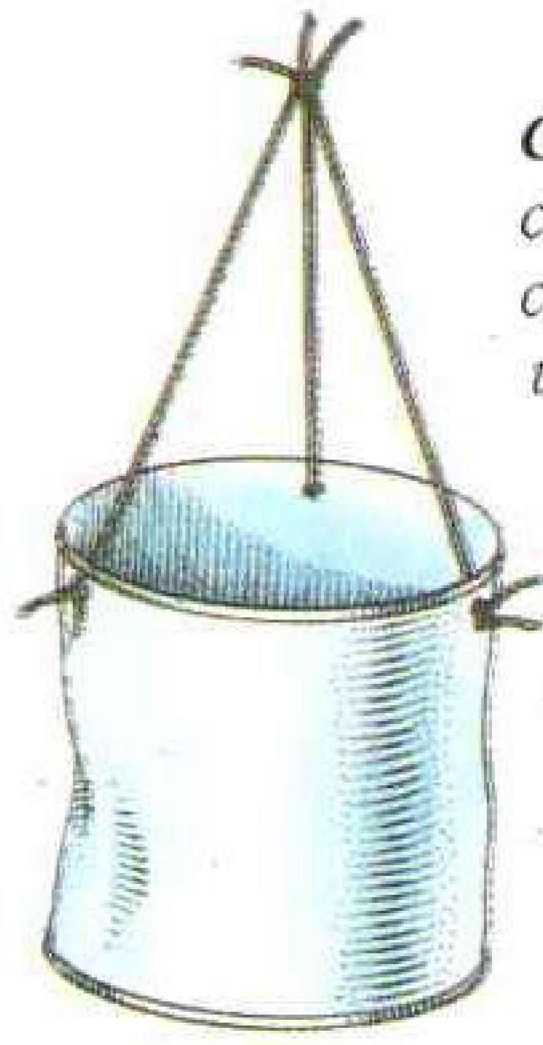

Cazo de lata. *Construya un cazo atando un asa de cuerda o alambre a una lata vacía. Tenga cuidado de mantener el cazo por encima de las llamas del fuego, para evitar que el asa de cuerda se queme.*

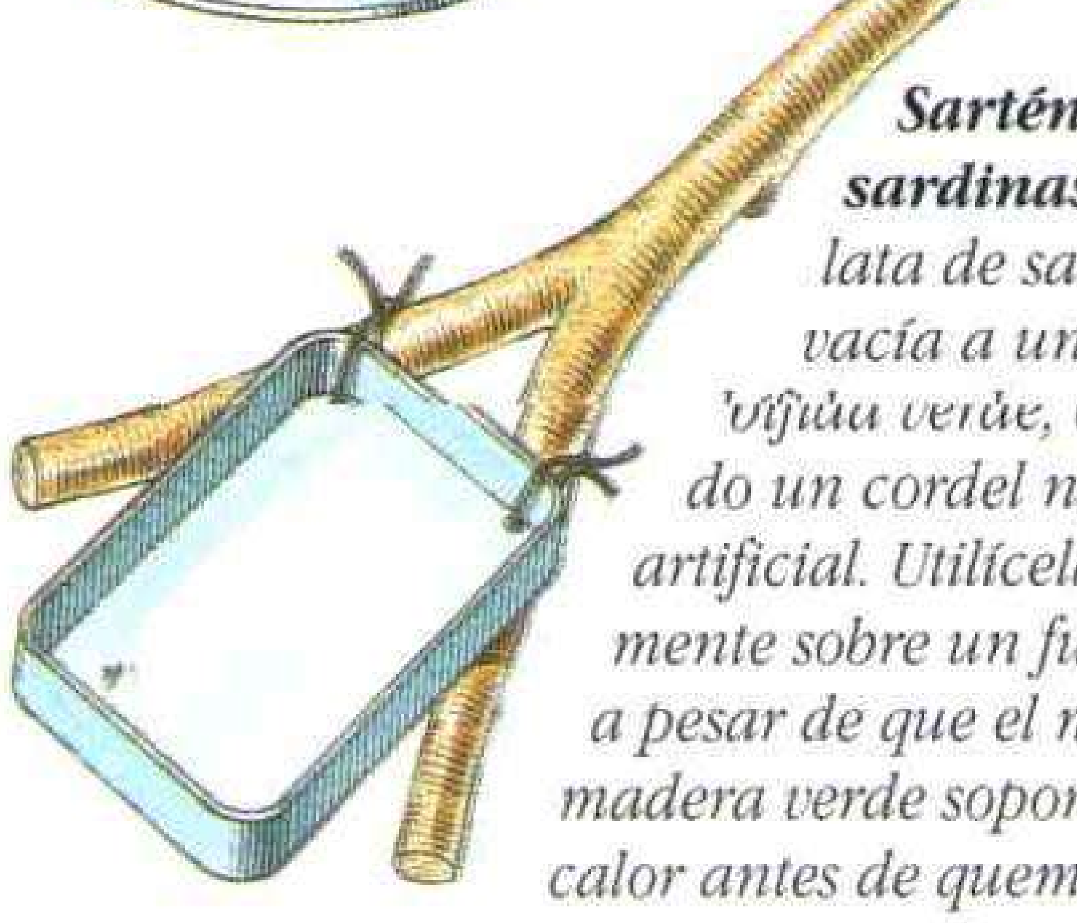

Sartén de lata de sardinas. *Ate una lata de sardinas vacía a una rama bífida verde, empleando un cordel natural o artificial. Utilícela únicamente sobre un fuego lento, a pesar de que el mango de madera verde soportará mucho calor antes de quemarse.*

COCINAS Y HORNOS

La forma más sencilla de cocinar es en un fuego al aire libre, o con bloques de combustible sólido que arden con una llama regular durante el tiempo suficiente para preparar una bebida caliente. Si no puede construir el fuego con facilidad, o que sea seguro, necesitará una cocina con una llama controlable, que se encienda fácilmente y que queme el combustible con buen rendimiento. Para una cocina más elaborada, puede construir un horno o un foso de vapor. Guarde la cocina, el combustible sólido, las cerillas y el cazo con su tapa en un bolsillo lateral de su mochila, para poder encontrarlos fácilmente cuando los necesite.

Recargar una cocina. *Elija la cocina que más se ajuste a sus necesidades culinarias, y en función de la posibilidad de encontrar gas o combustible para recargarla. Al recargar una cocina, asegúrese de que la llama esté realmente apagada.*

COCINAS DE CAMPAMENTO

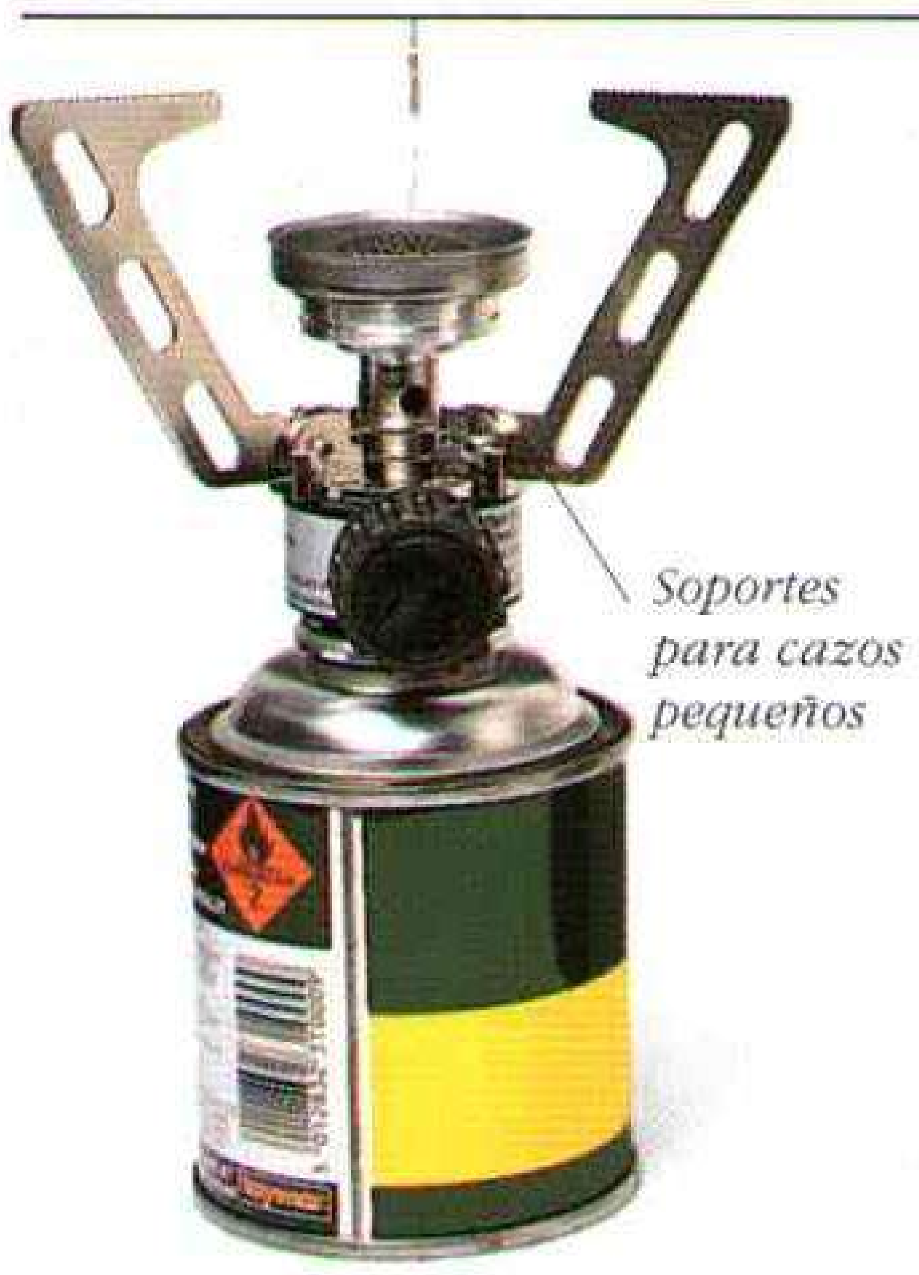

Mini-cocina. *Esta cocina ultraligera requiere una mezcla de butano y propano. No es adecuada para soportar cazos grandes.*

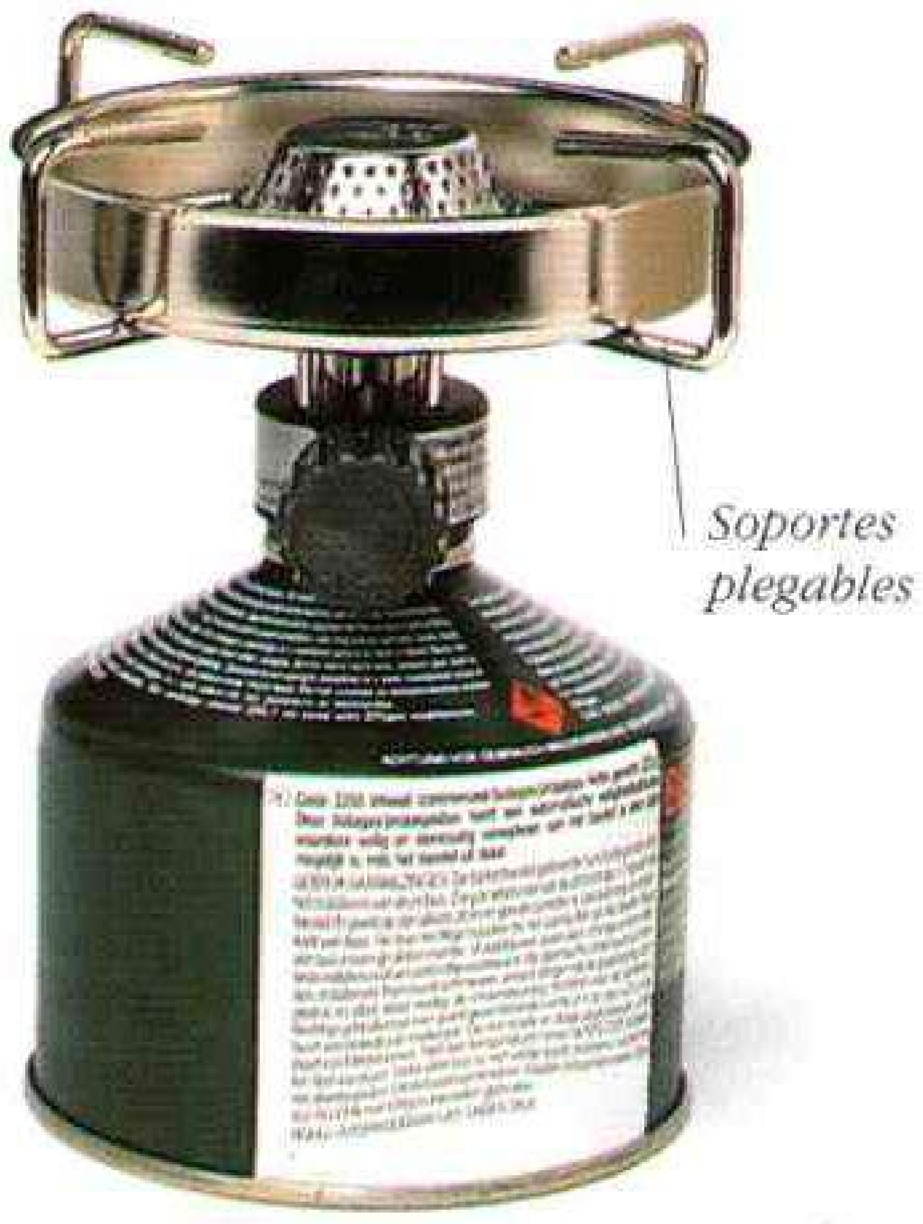

Cocina para uso general. *Esta cocina rápida es de uso general, y tiene soportes plegables para cocinar con cazos grandes.*

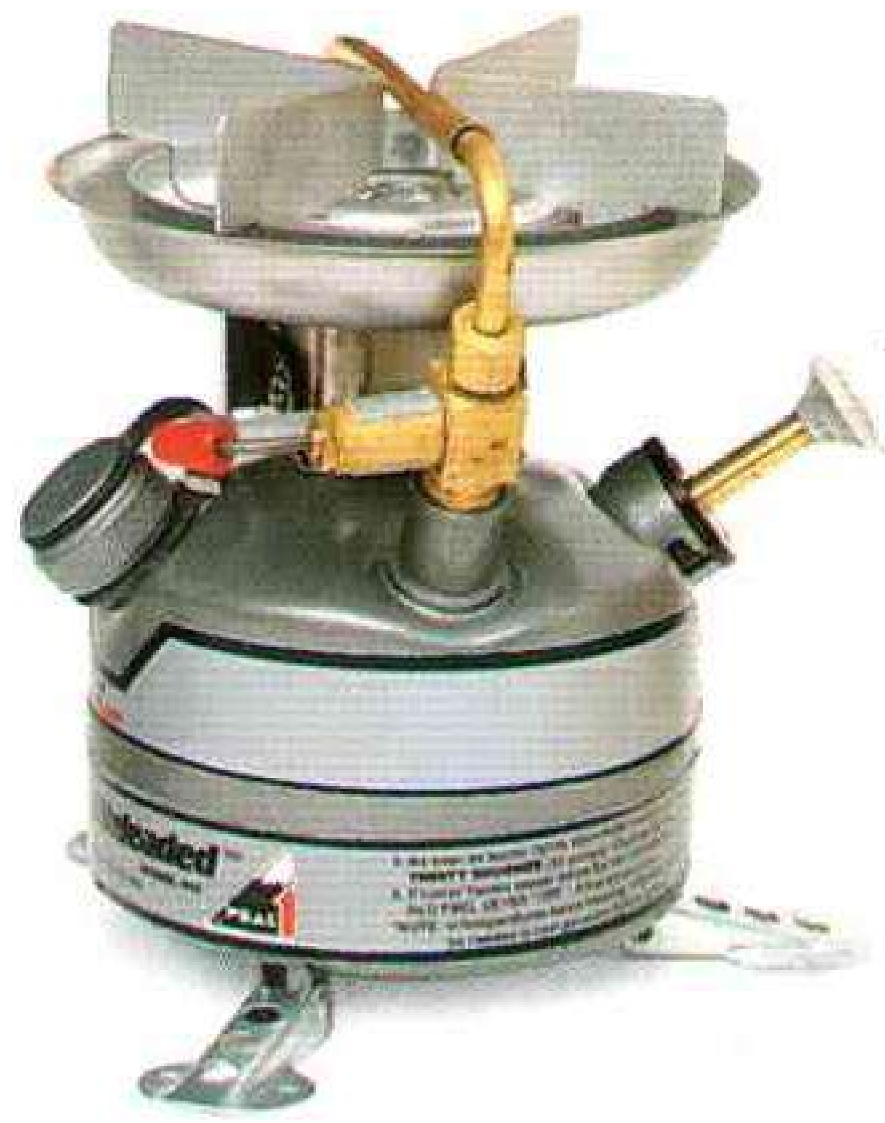

Cocina de gasolina sin plomo. *Esta cocina requiere gasolina sin plomo y tiene un accionamiento de purga para limpiar el generador.*

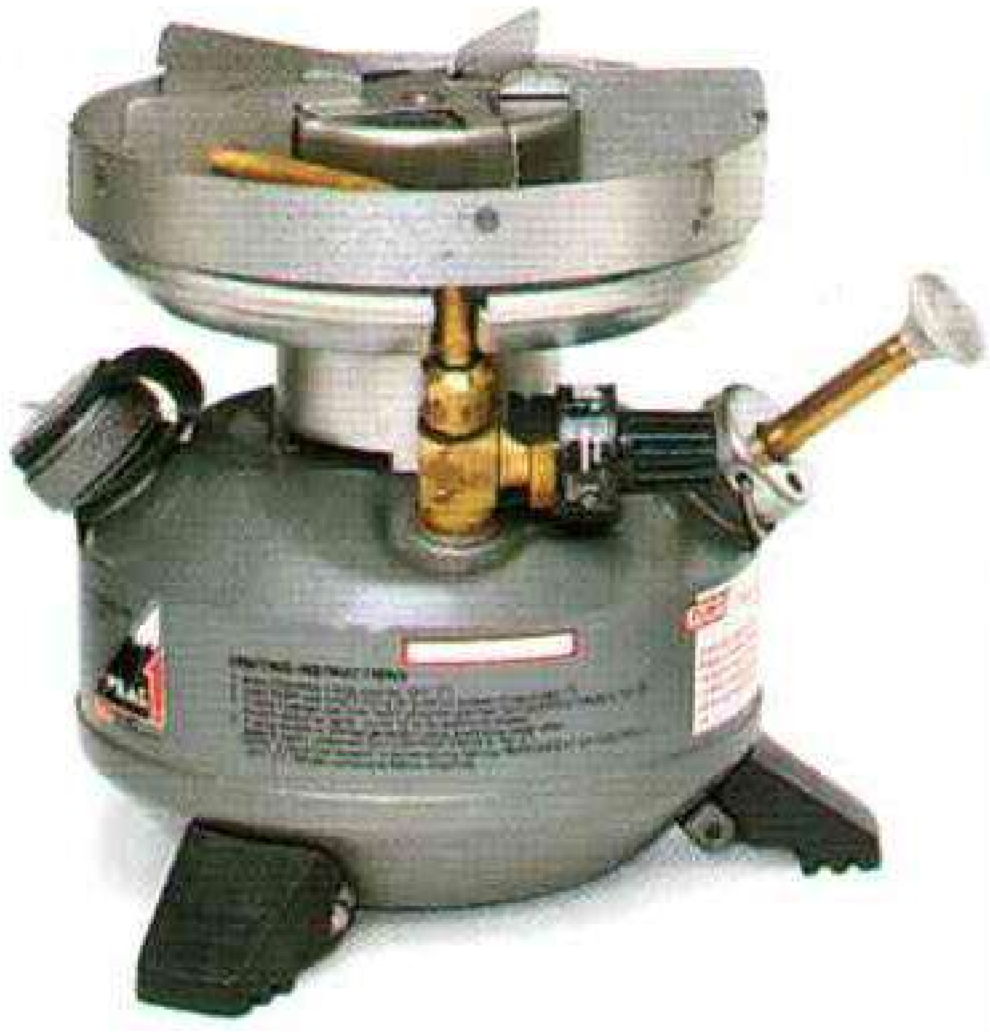

Cocina de combustible múltiple. *Es una de las cocinas más populares del mundo; funciona con gasolina, parafina o queroseno.*

ADVERTENCIA

Cualquier cocina debe tratarse con sumo cuidado. Recuerde que siempre existe el riesgo de explosión. Las áreas de cocina deben estar bien ventiladas, ya que las cocinas consumen oxígeno y despiden monóxido de carbono tóxico. Cuando se apaga la llama de una cocina con gas a presión, el vapor se almacena y cualquier llama o chispa dará lugar a una gran explosión.

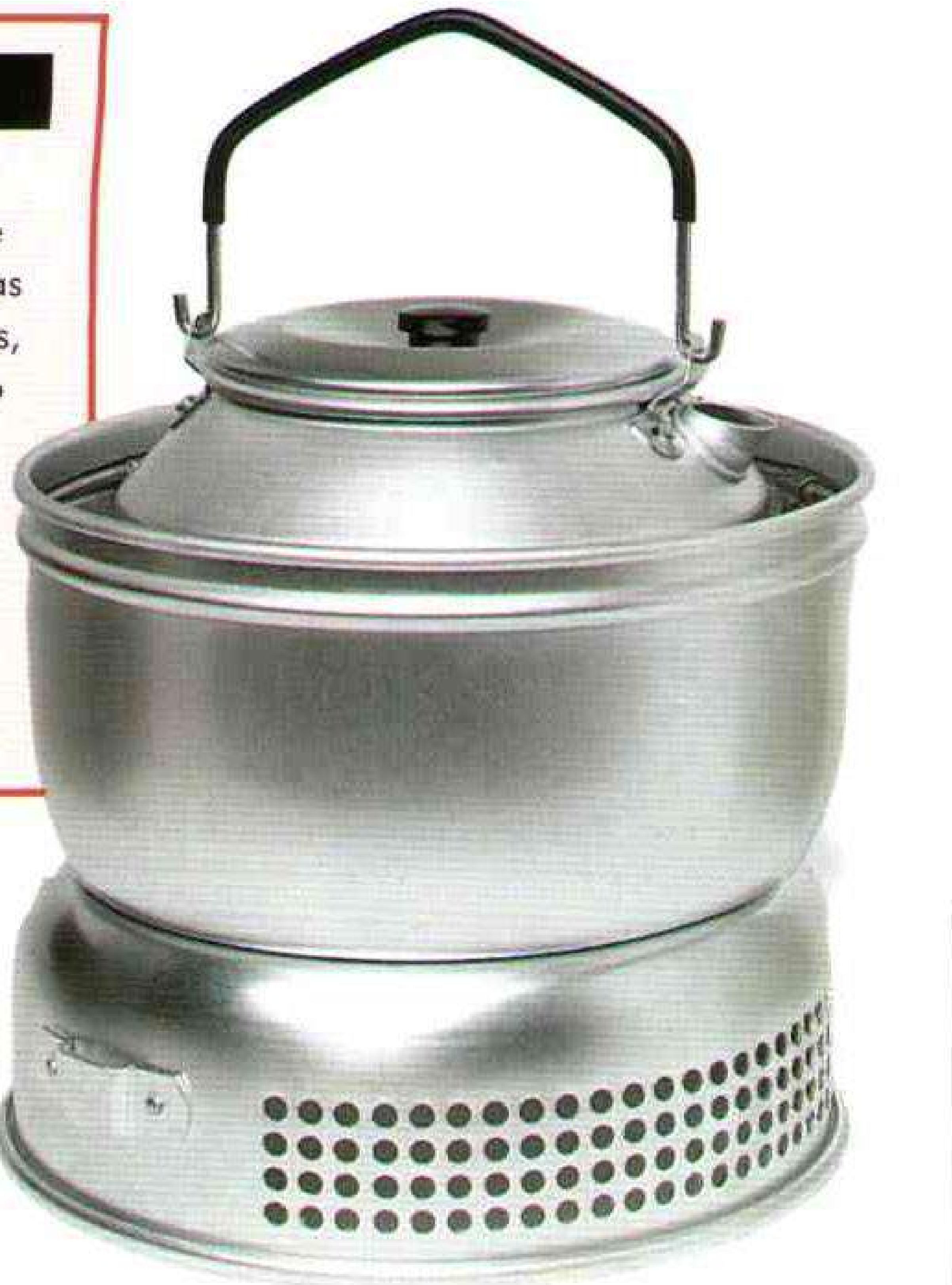

Cocina no presurizada. *Esta cocina de uso popular funciona con alcohol de quemar, y lleva incorporado un paraviento, así como una batería de cocina. Debido a su forma, el riesgo de rotura de algún componente es mínimo; es bastante estable, pero no calienta con demasiada rapidez.*

COMBUSTIBLE

Las botellas de combustible deben distinguirse fácilmente de aquellas que contienen agua, para evitar confundirlas. No deben presentar fuga alguna que pudiera contaminar los alimentos y deteriorar las ropas y el equipo. La disponibilidad de combustible debe preverse antes de viajar, así como la eliminación de los recipientes usados.

COMBUSTIBLE SÓLIDO A BASE DE ALCOHOL

BOTELLA DE GASOLINA

BOTELLA DE PARAFINA

BOMBONA DE BUTANO/PROPANO

HORNO DE BARRO

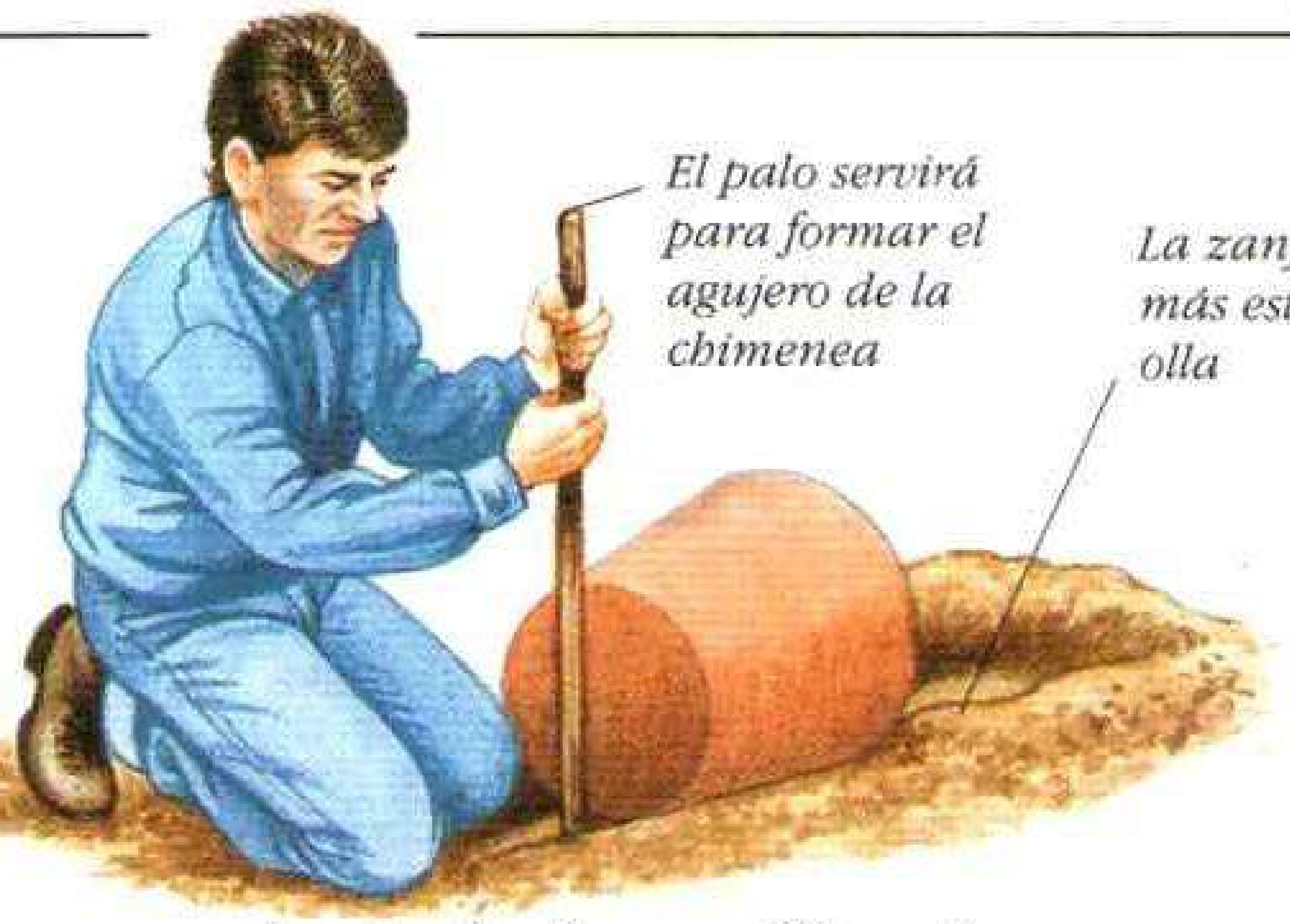

El palo servirá para formar el agujero de la chimenea

La zanja debe ser más estrecha que la olla

1 Cave una zanja estrecha de unos 20 cm de profundidad. Coloque una olla con tapa de unos 23 cm de profundidad sobre uno de sus costados. Clave un palo largo en el suelo justo en la base de la olla.

2 Cubra la olla completamente con tierra o arcilla húmedas, moviendo el palo para dejar libre el agujero de la chimenea. Compacte la tierra para que la cúpula sea lisa y estable.

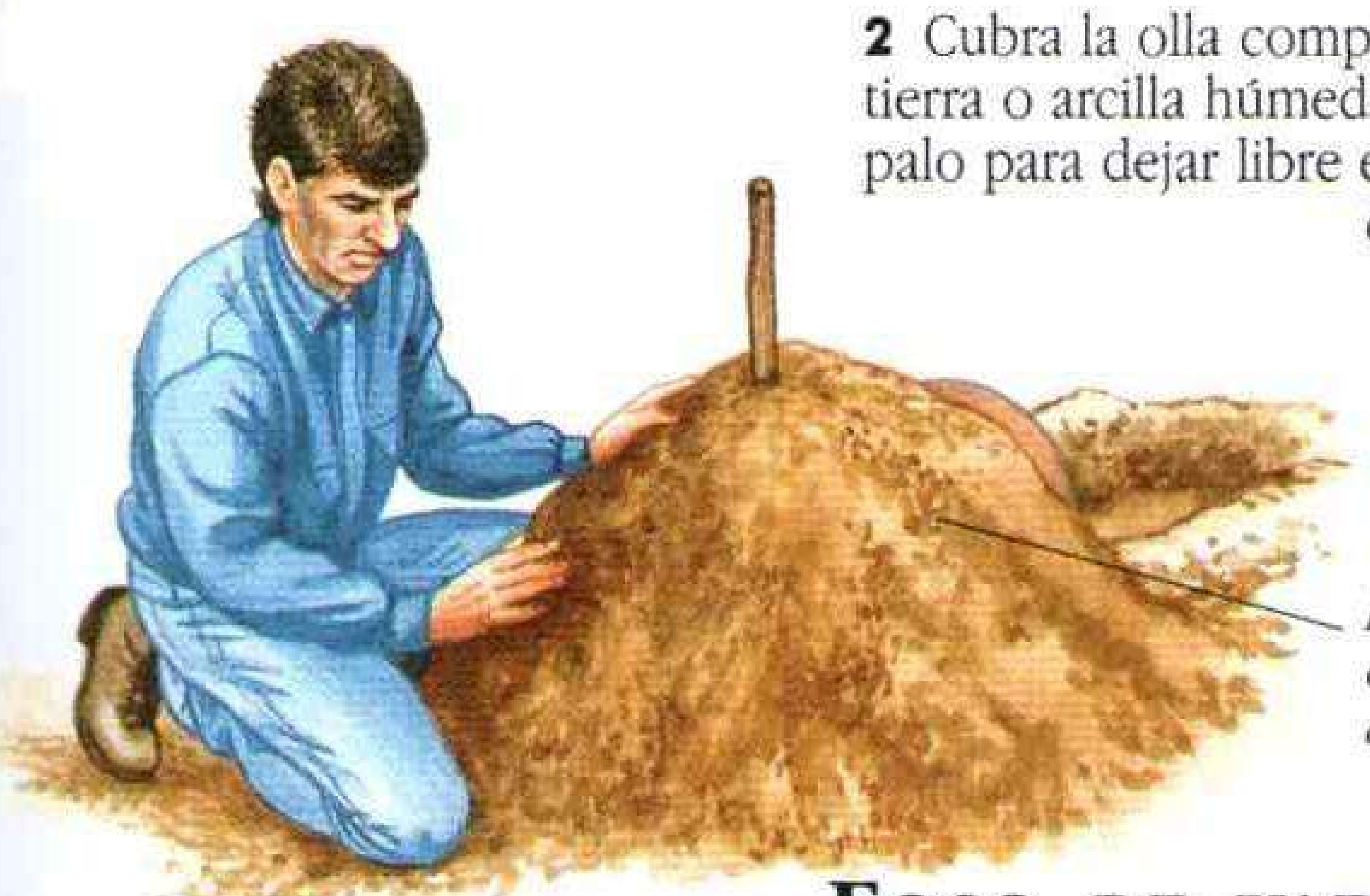

La olla se cubre con arcilla

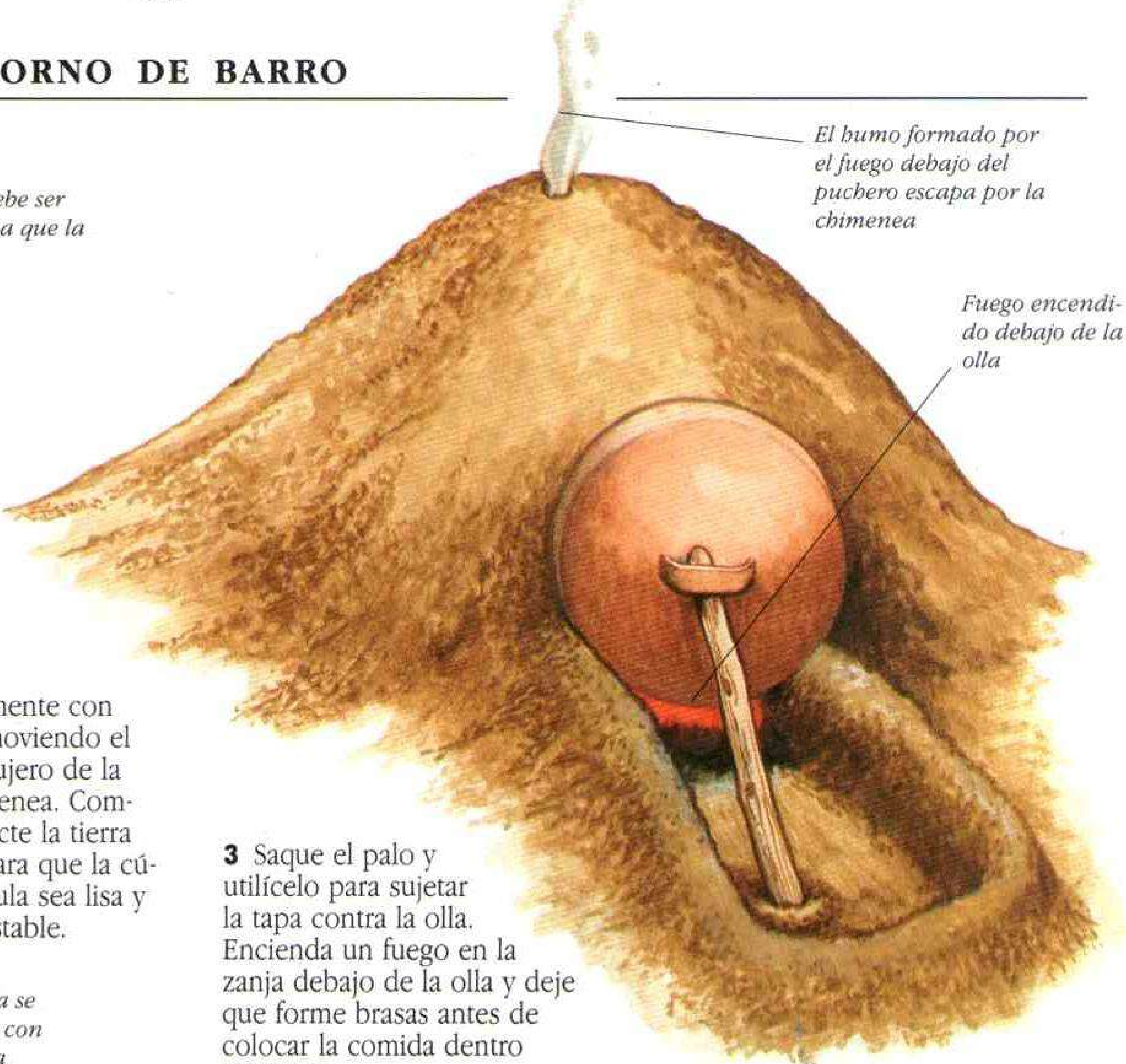

El humo formado por el fuego debajo del puchero escapa por la chimenea

Fuego encendido debajo de la olla

3 Saque el palo y utilícelo para sujetar la tapa contra la olla. Encienda un fuego en la zanja debajo de la olla y deje que forme brasas antes de colocar la comida dentro del horno.

FOSO DE VAPOR

1 Encienda un fuego sobre una plataforma de ramas encima de una zanja *(véase superior)*. Coloque piedras pequeñas sobre la plataforma.

2 Una vez la plataforma se ha quemado y las piedras han caído a la zanja, clave un palo vertical en la zanja. Lleve las ascuas a un lado. Envuelva el alimento en hierbas y colóquelo en la zanja sobre una capa gruesa de hierba.

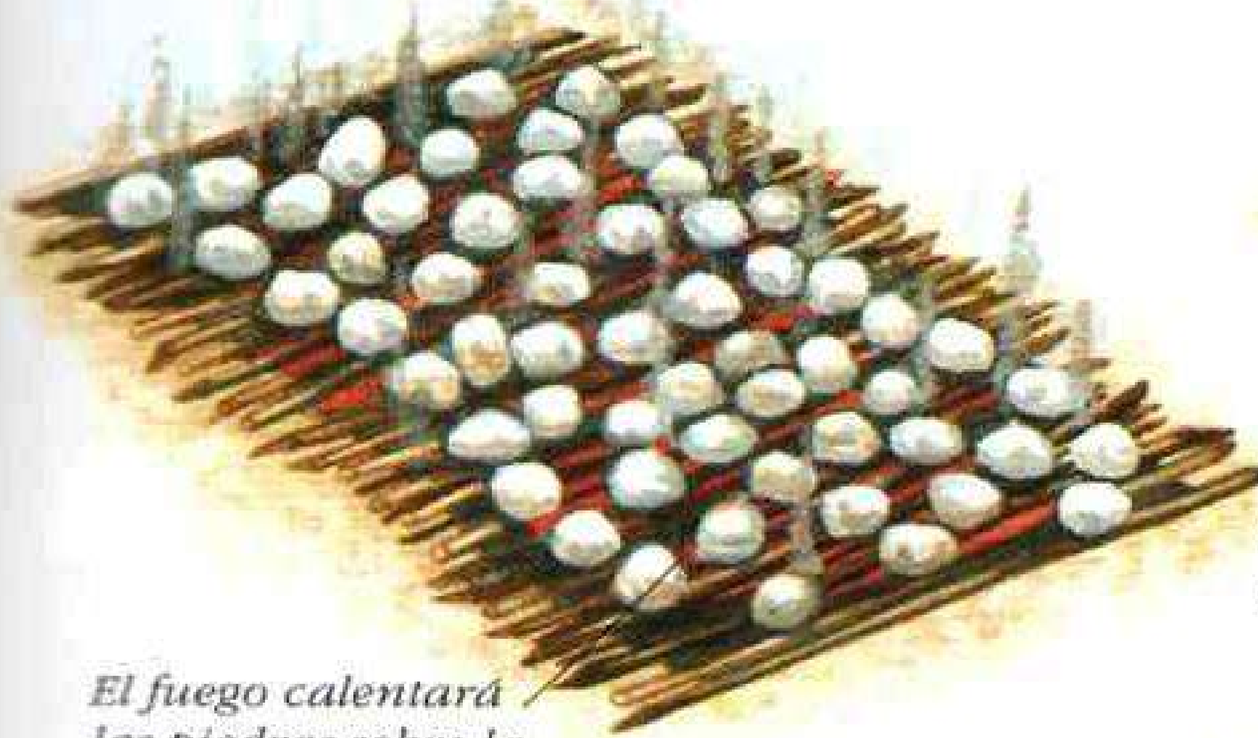

El fuego calentará las piedras sobre la plataforma de madera

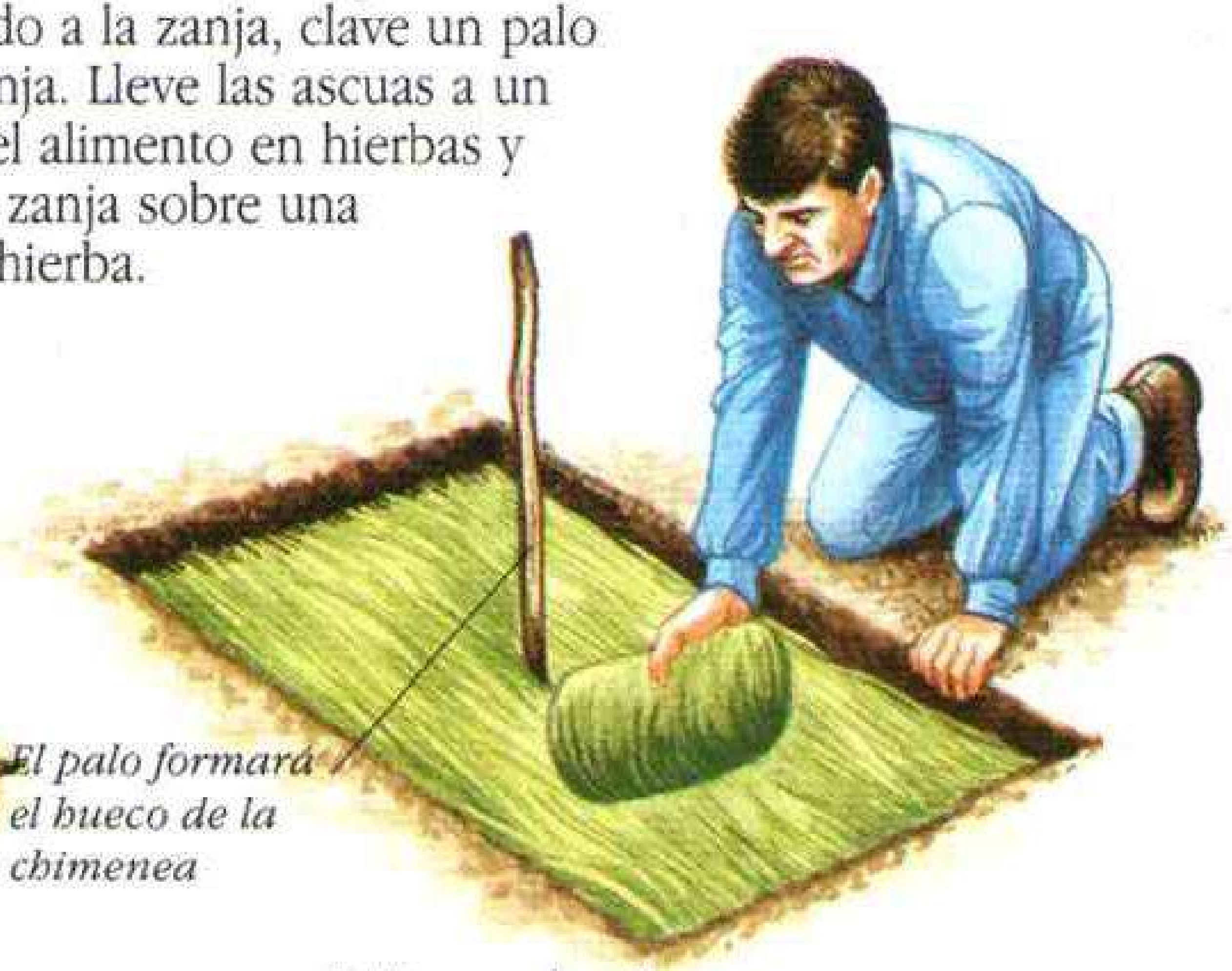

El palo formará el hueco de la chimenea

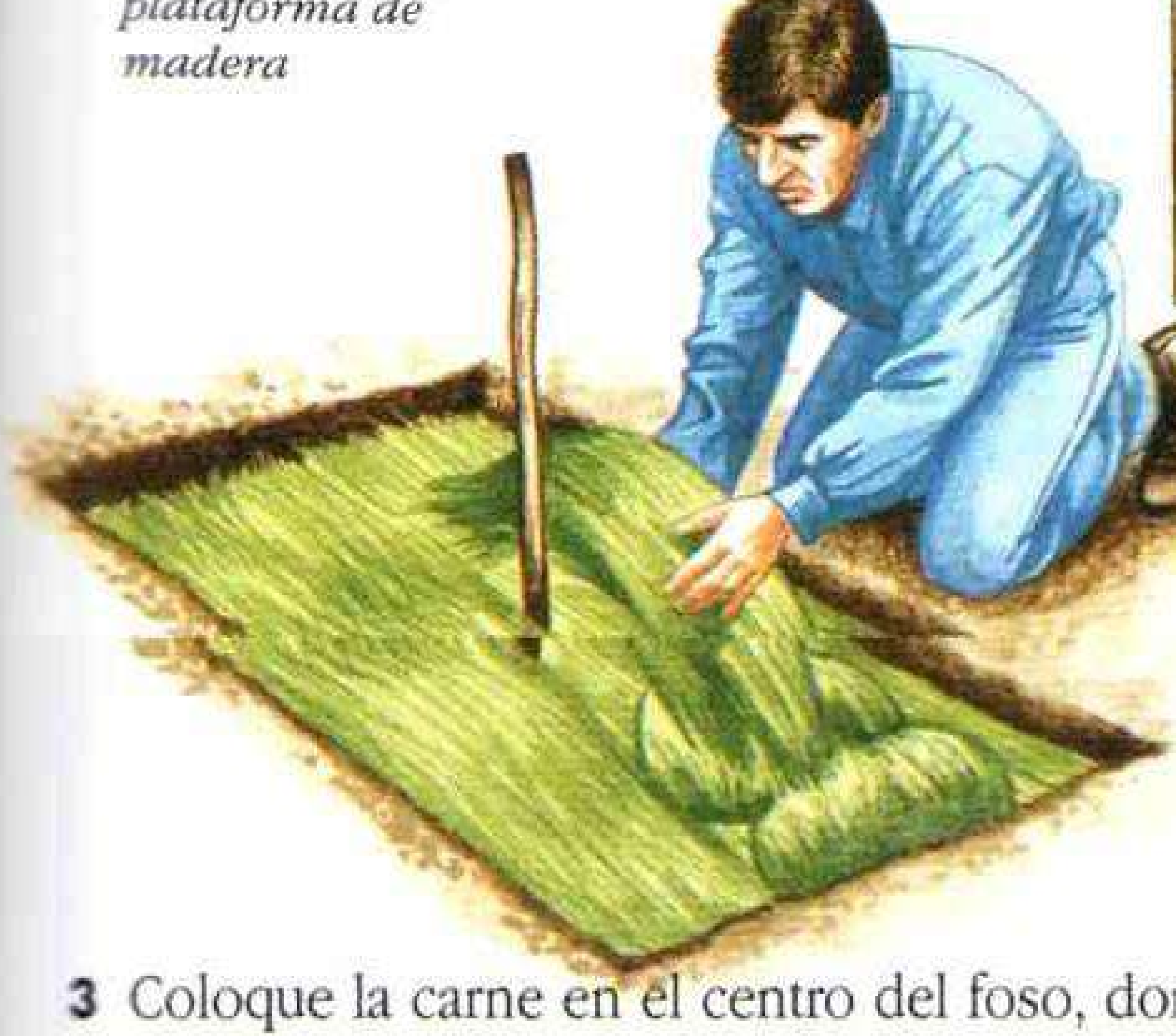

3 Coloque la carne en el centro del foso, donde se acumula la mayor parte del calor. Extienda una capa gruesa de hierba sobre el alimento. Cubra el foso con tierra.

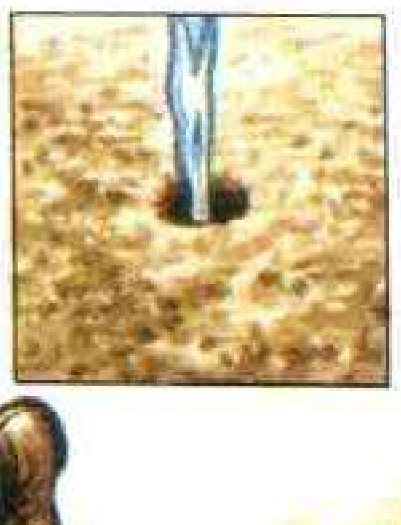

4 Saque el palo y vierta un poco de agua en el hueco para generar vapor.

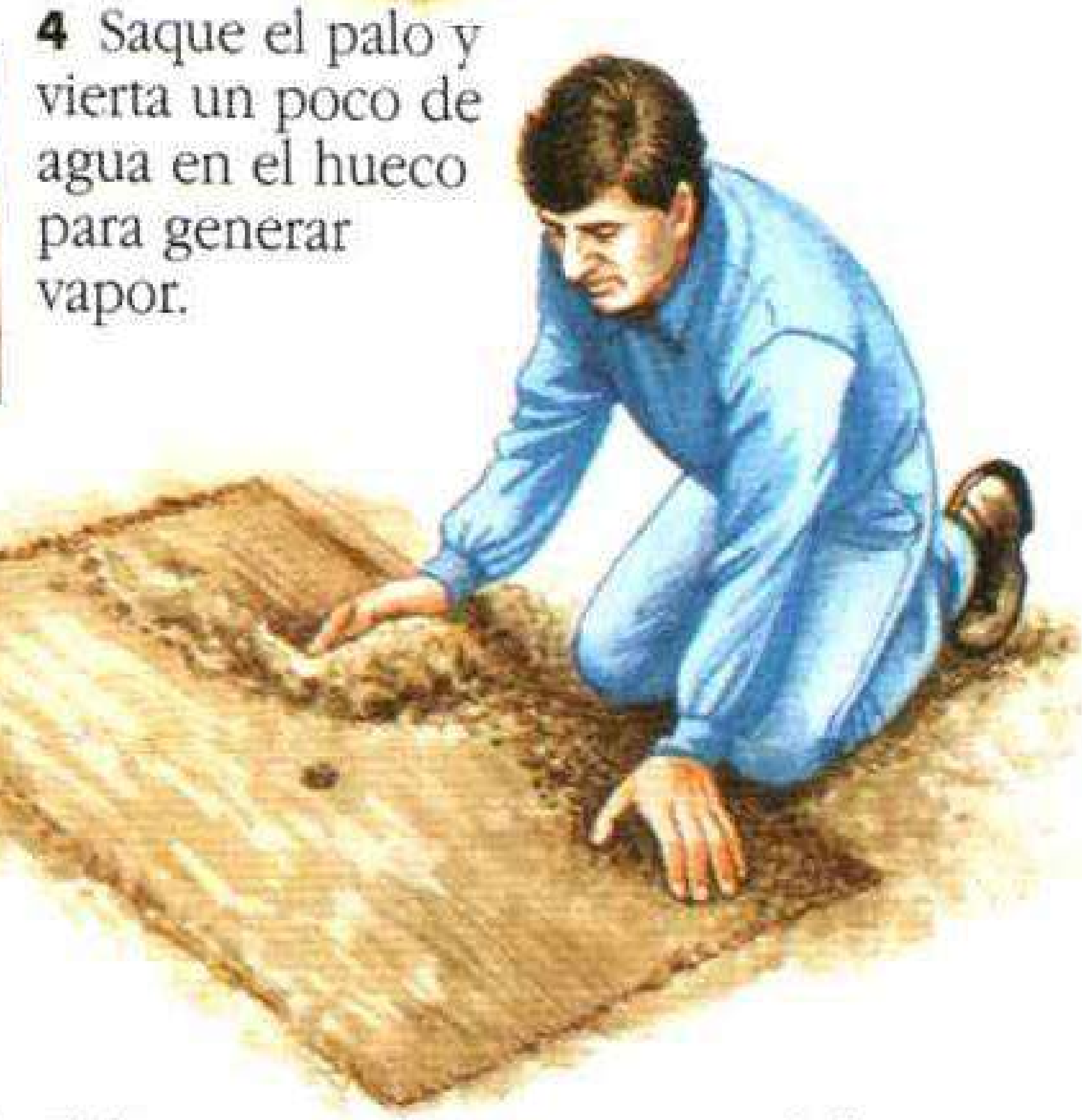

5 Selle el foso con tierra, compactándola hasta dejarla lisa. El alimento estará cocido transcurridas unas cuatro horas.

HORNO DE YUKÓN

Se trata de una chimenea encima de un fuego, que quemará incluso leña húmeda. Su parte superior constituye una cocina muy efectiva. Construya primero un fuego de tipi *(véase página 60)*, pero no lo encienda. Levante una pared de piedras alrededor del tipi, utilizando arcilla como mortero, y selle el exterior con arcilla. La forma cónica de la estructura incrementa el flujo de aire alrededor del fuego, creando más calor. Encienda el extremo de un palo largo y empújelo hacia el fuego desde arriba. Utilice madera seca para encender el fuego. Una vez esté bien encendido puede secar la madera húmeda sobre el horno.

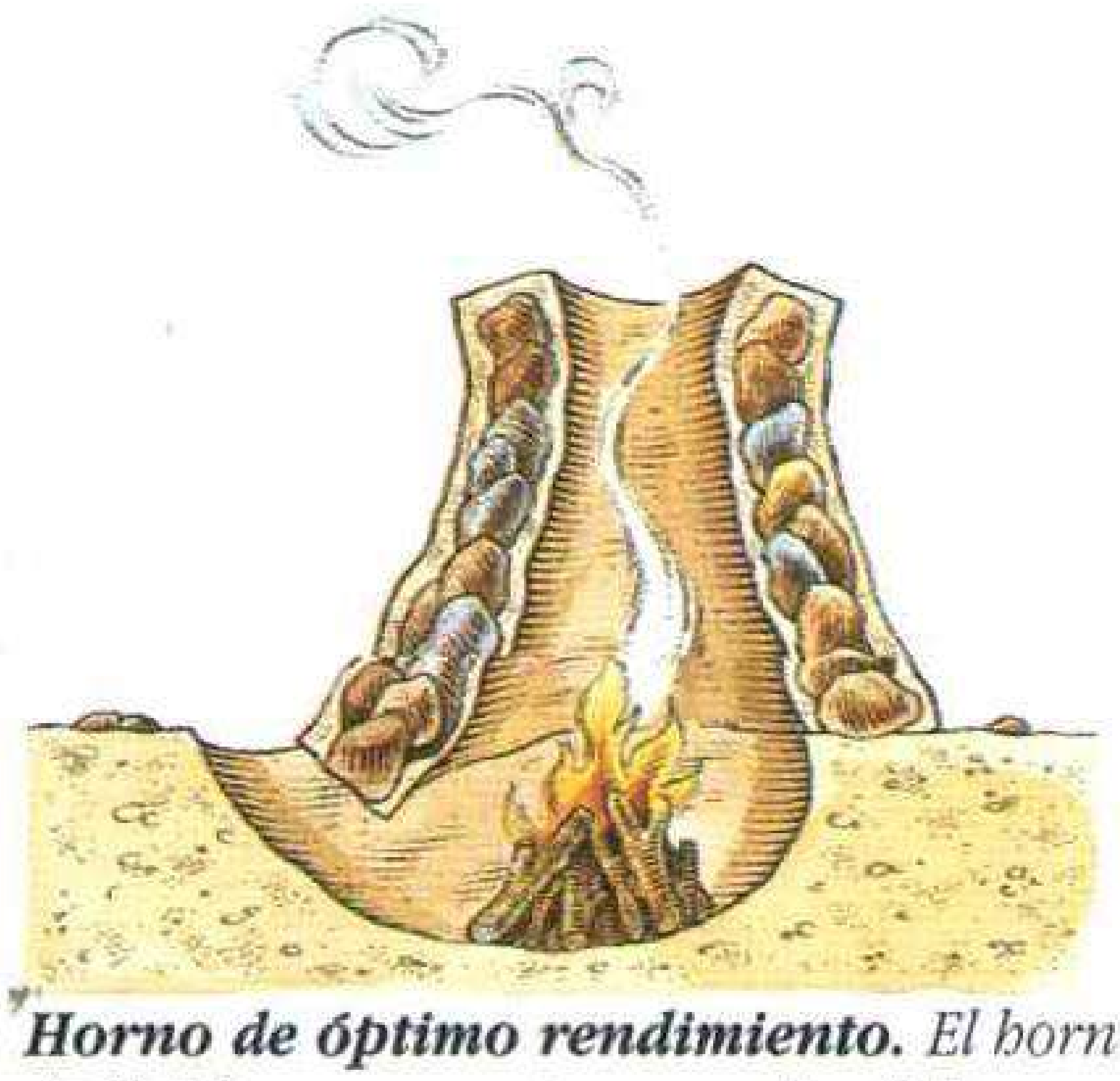

Horno de óptimo rendimiento. *El horno de Yukón consume poco combustible y desprende gran cantidad de calor.*

MÉTODOS DE COCCIÓN

LA FORMA MÁS básica de cocción consiste en asar los alimentos empleando, por lo general, un espetón sobre las llamas o las brasas de un fuego. En una situación de supervivencia es importante cocinar todos los alimentos que den lugar a dudas, a fin de matar cualquier bacteria peligrosa, posibles parásitos o sustancias químicas. Cocinados los alimentos son más apetecibles y digeribles que en su estado natural. Además, una comida caliente levanta el ánimo a cualquiera.

Cocinar sobre un fuego al aire libre. *Puede preparar un estofado en un cazo suspendido sobre el fuego, o puede esperar hasta que las llamas se conviertan en brasas para utilizarlas en otras modalidades de cocina* (véase página siguiente). *Si dispone de grandes cantidades de agua, tenga siempre a mano agua hervida para preparar bebidas calientes.*

SOPORTES PARA CAZOS

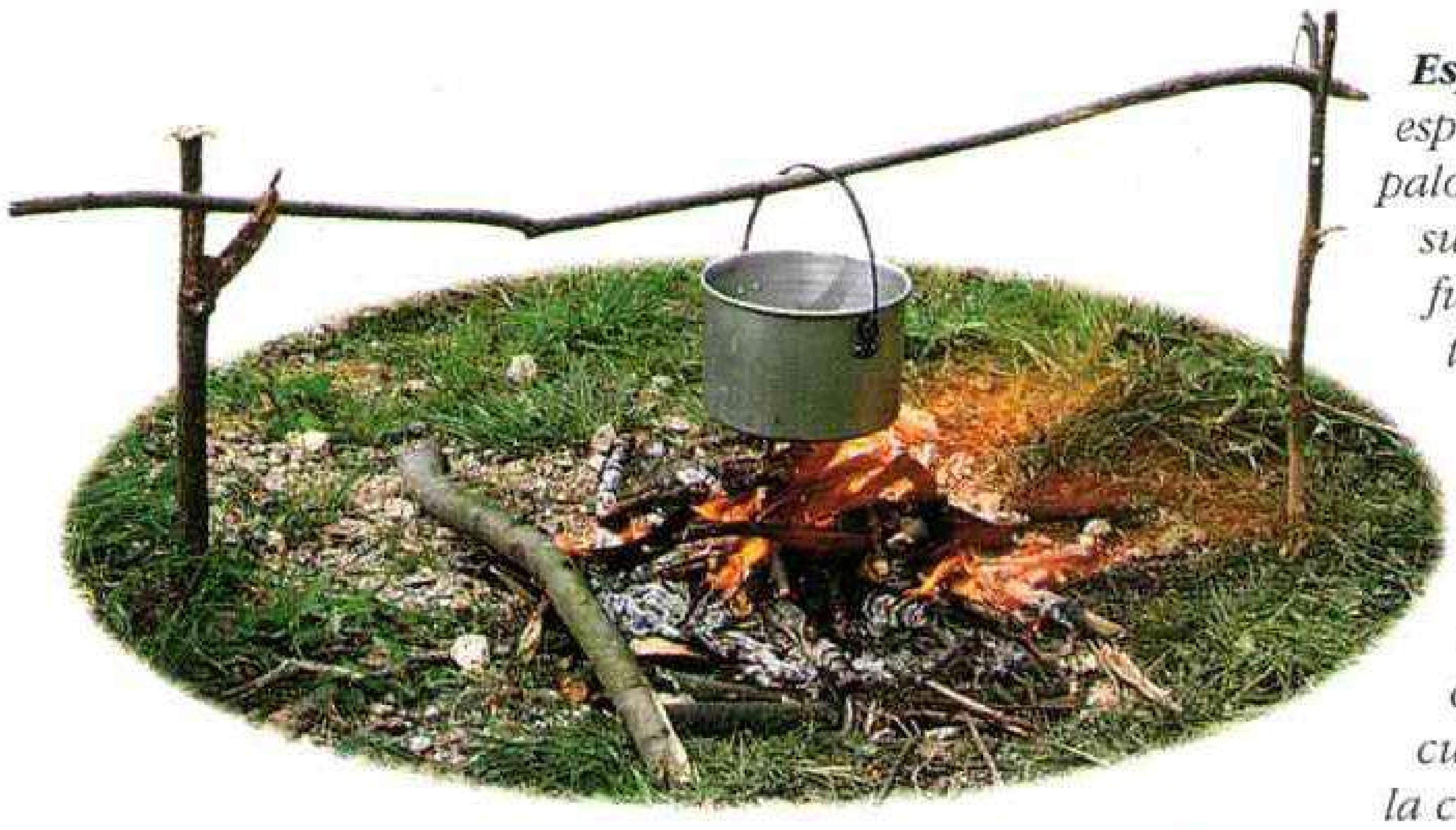

Espetón. *Prepare un espetón clavando dos palos ahorquillados en el suelo a ambos lados del fuego y apoyando un tercer palo entre ellos. Evite que el fuego arda demasiado y envuelva el cazo. Utilice un palo para empujar el cazo hacia el extremo del espetón cuando quiera servirse la comida.*

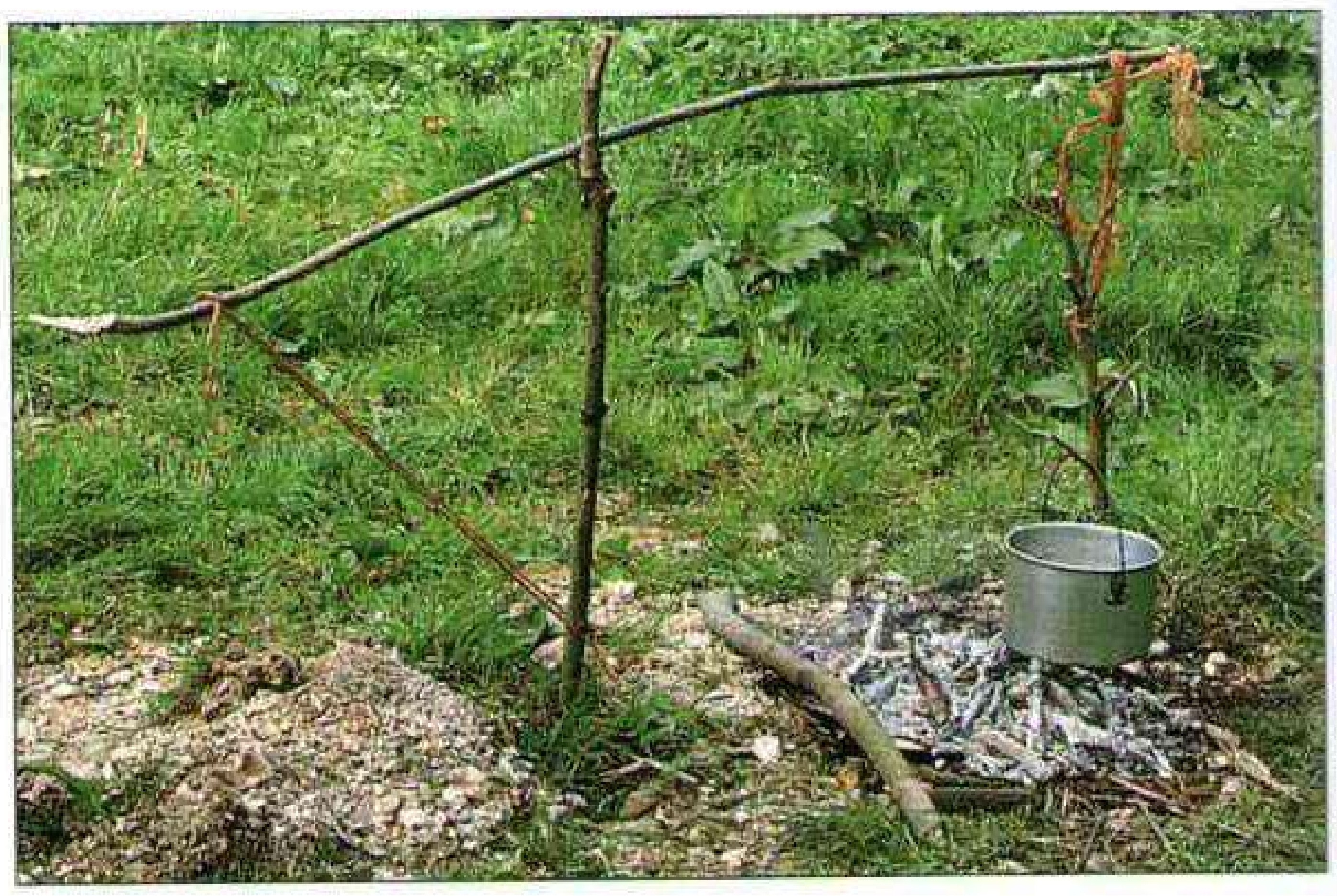

Soporte oscilante. *Resulta muy útil al cocinar estofados que deban apartarse del fuego para ajustar su temperatura. Clave un palo ahorquillado en el suelo y equilibre una rama larga sobre la horquilla para que de un extremo cuelgue el cazo sobre el fuego. Ate el otro extremo al palo vertical. El cazo puede apartarse del fuego con un palo.*

Trípode. *Este método resulta muy estable y muy seguro. Construya un trípode* (véase página 75), *y cuelgue una rama en forma de «U» de su centro, atada con un cordel. Cuelgue el cazo de la rama. Esto le permitirá levantar o bajar el cazo sobre el fuego y controlar así la temperatura de la comida. Permita que el fuego baje antes de colocar el trípode sobre los rescoldos.*

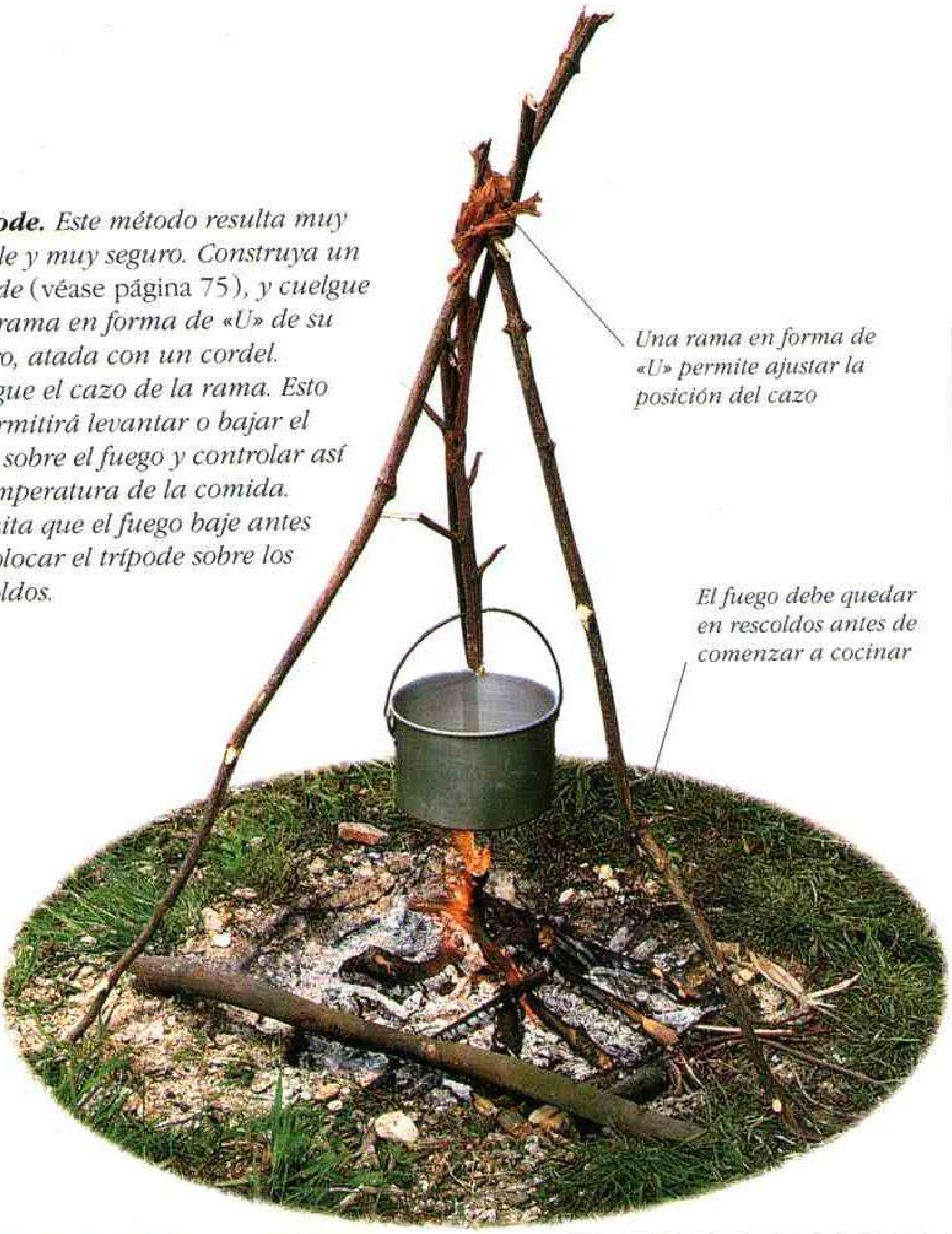

Una rama en forma de «U» permite ajustar la posición del cazo

El fuego debe quedar en rescoldos antes de comenzar a cocinar

EL EMPLEO DEL BAMBÚ

Al ser resistente y hueco, el bambú constituye un excelente recipiente para cocinar. Con una rama terminada en punta, haga un agujero en cada una de las paredes que separan los compartimentos huecos, dejando intacta la de un extremo. Vierta agua en el tallo hasta llegar al primer anillo. Apoye el tallo sobre un fuego sujetando el extremo superior con una rama en forma de «U». Caliente el tallo para hervir el agua y cocinar al vapor los alimentos colocados en la parte superior.

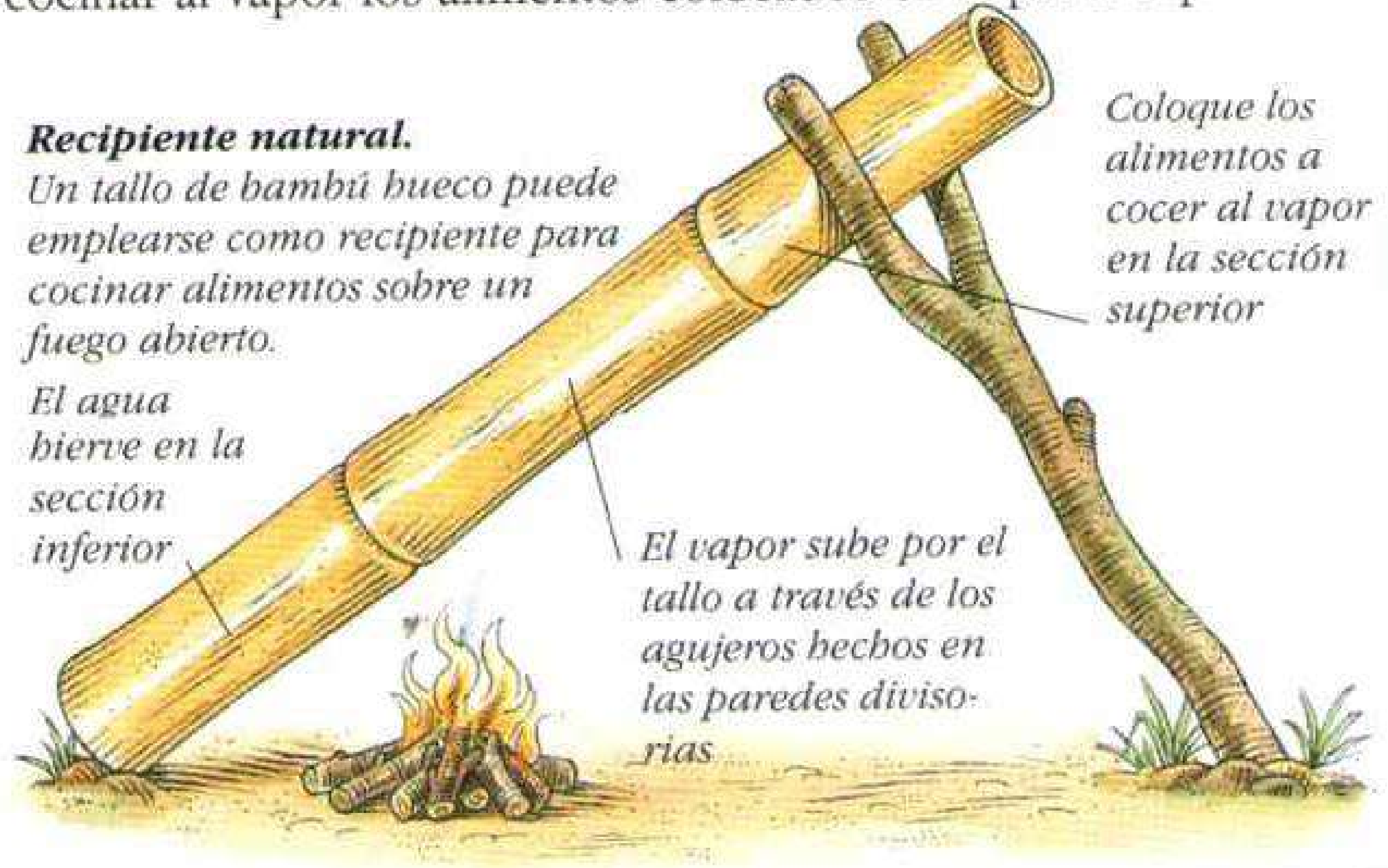

Recipiente natural. *Un tallo de bambú hueco puede emplearse como recipiente para cocinar alimentos sobre un fuego abierto.*

El agua hierve en la sección inferior
El vapor sube por el tallo a través de los agujeros hechos en las paredes divisorias
Coloque los alimentos a cocer al vapor en la sección superior

COCINAR SOBRE PIEDRAS CALIENTES

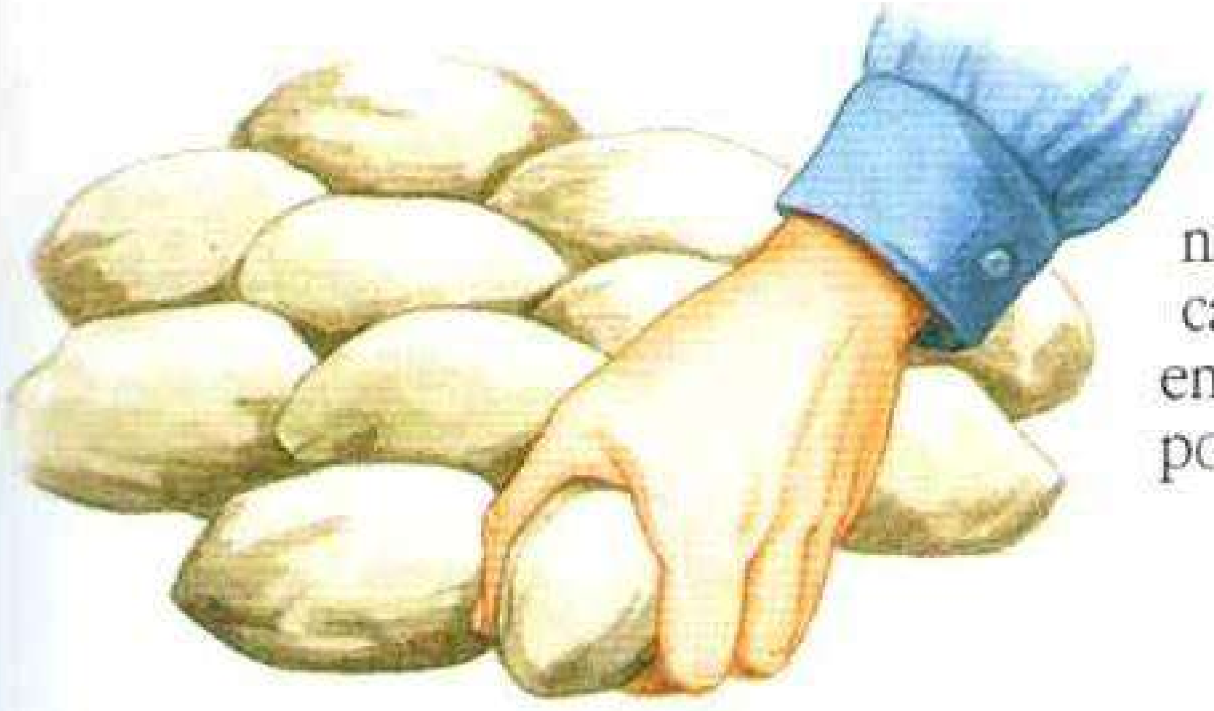

1 Construya un lecho denso de piedras grandes y planas. No utilice pizarra ni otro tipo de roca estratificada porque podría estallar en pedazos al ser calentada por la acción del fuego.

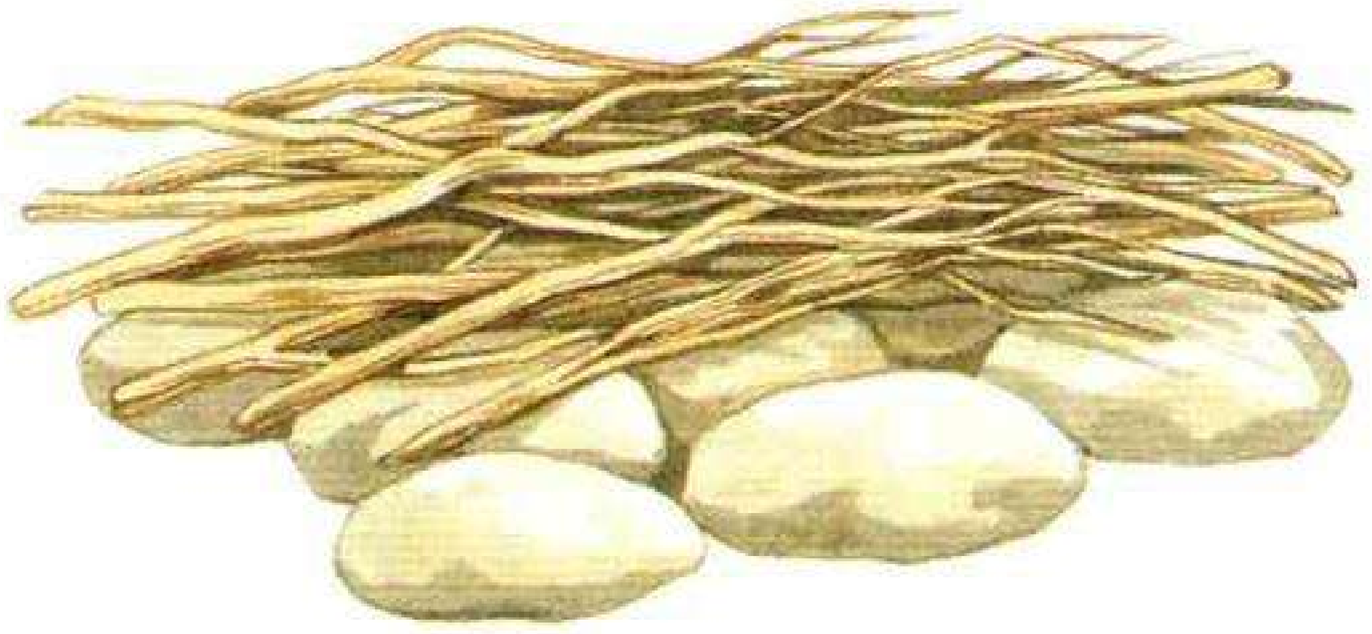

2 Coloque yesca sobre las piedras y cúbrala con ramas secas Después encienda la yesca. Puede dejar que el fuego se reduzca a cenizas mientras prepara la comida *(véase página 110)*.

3 Con un manojo de ramas verdes, limpie las cenizas y los rescoldos de la superficie de las piedras. Tenga cuidado de no tocarlas con las manos porque estarán extremadamente calientes.

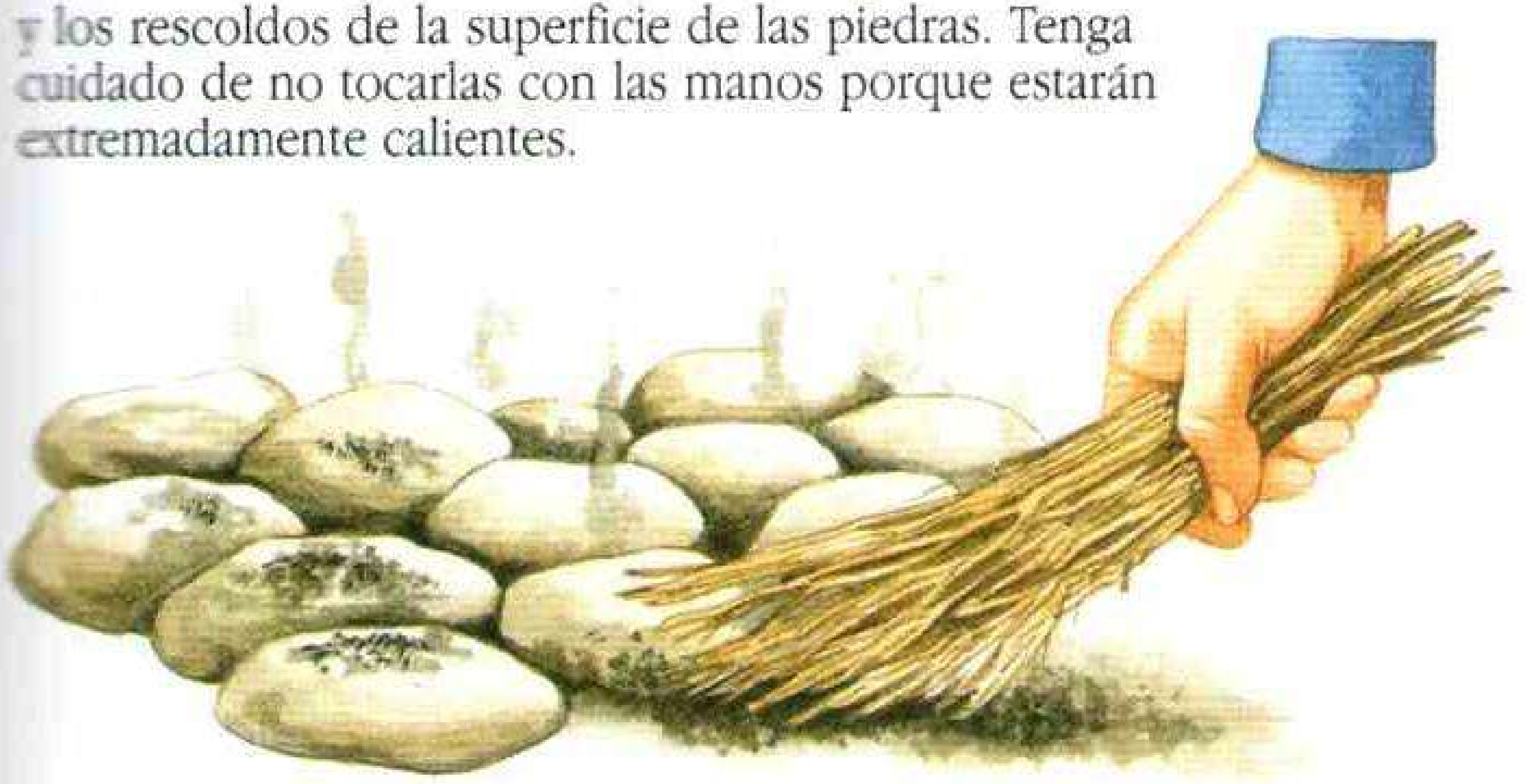

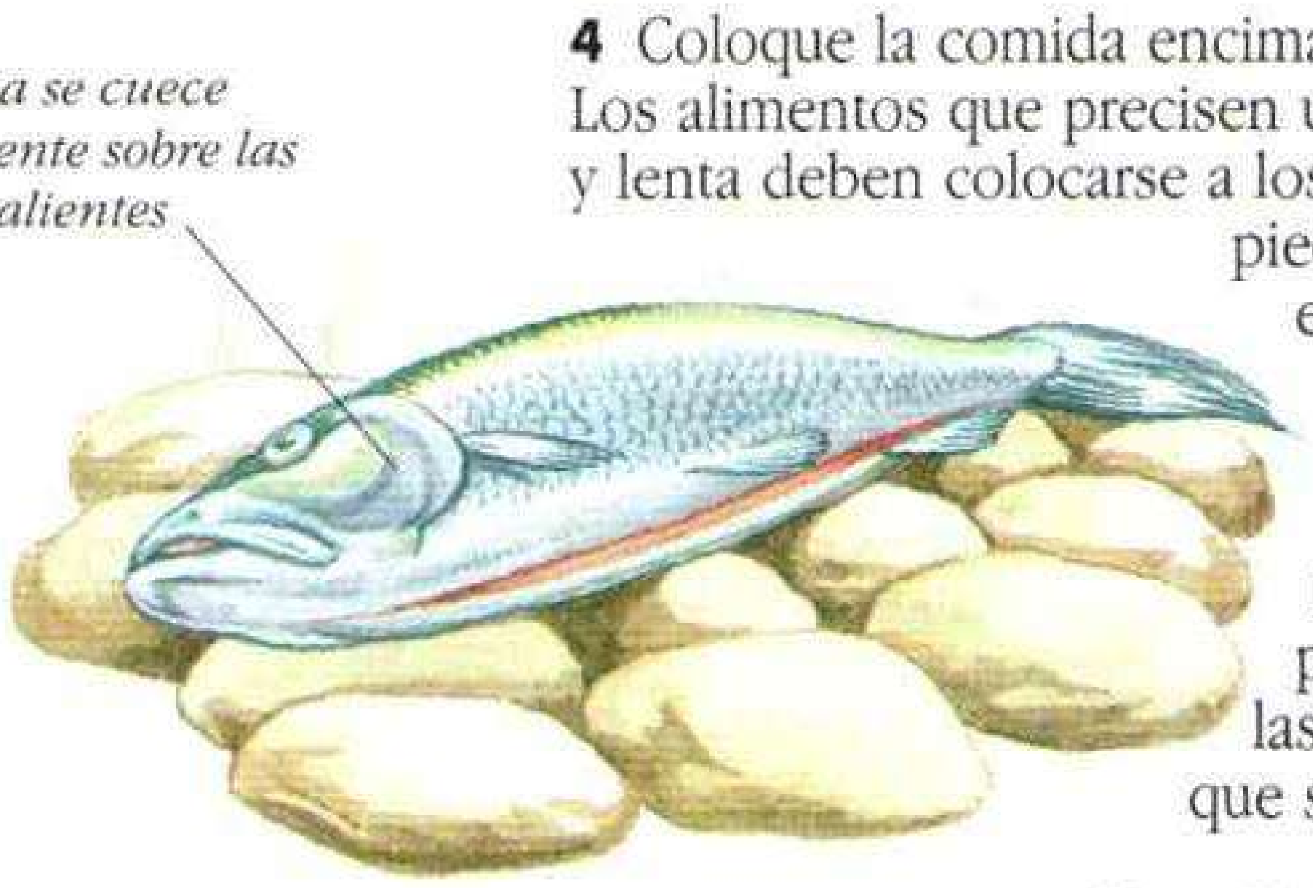
La comida se cuece directamente sobre las piedras calientes

4 Coloque la comida encima de las piedras. Los alimentos que precisen una cocción prolongada y lenta deben colocarse a los lados, donde las piedras están más frías que en el centro. Continúe colocando comida, hasta que las piedras se hayan enfriado. Las patatas sin pelar pueden enterrarse bajo las piedras calientes y dejar que se horneen.

HORNEAR EN BARRO

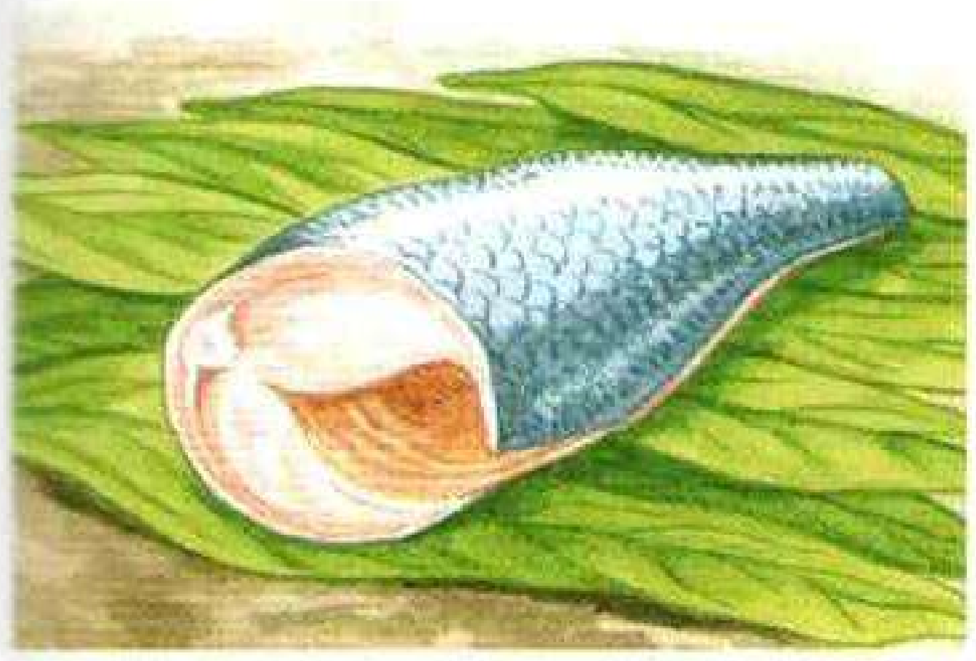

1 Puede hornear alimentos como el pescado envueltos en barro al fuego. Recoja un manojo grande de hierbas y largas hojas.

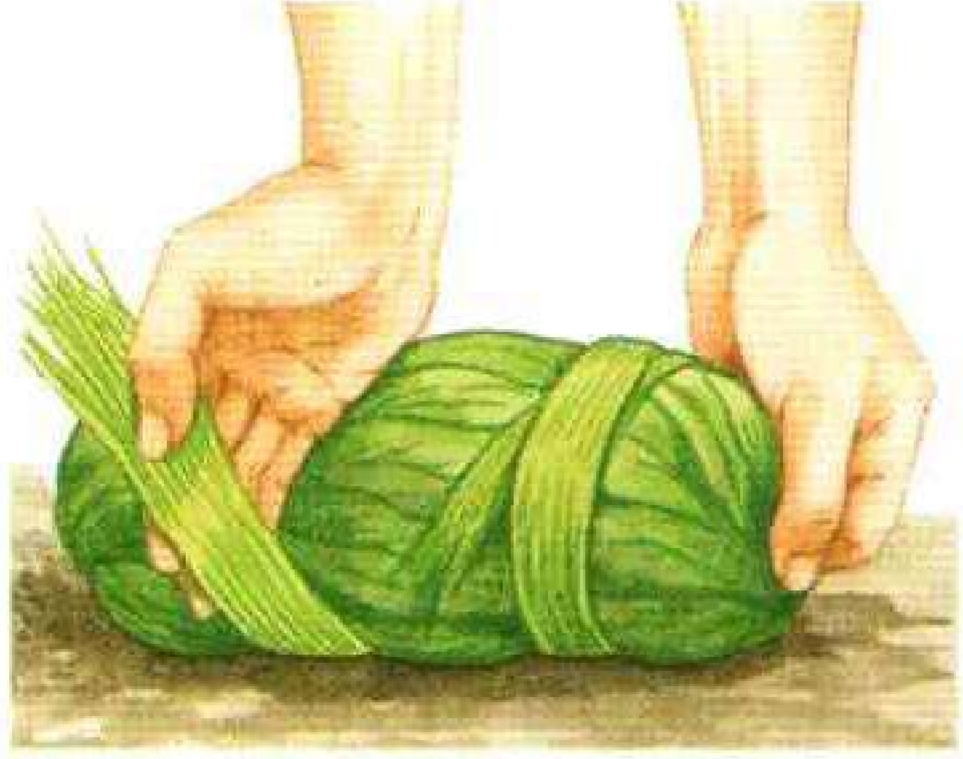

2 Envuelva el alimento en las hojas y ate el envoltorio con hebras trenzadas de hierbas.

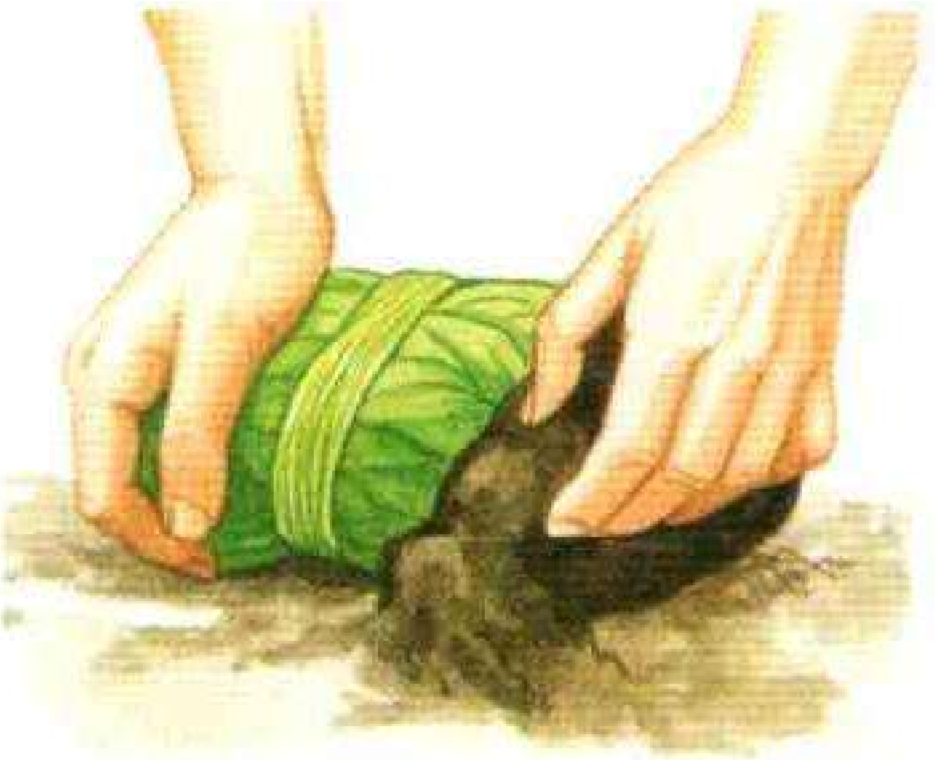

3 Cubra la totalidad del paquete con barro. Asegúrese de que queda completa y uniformemente cubierto, y de que esté bien cerrado.

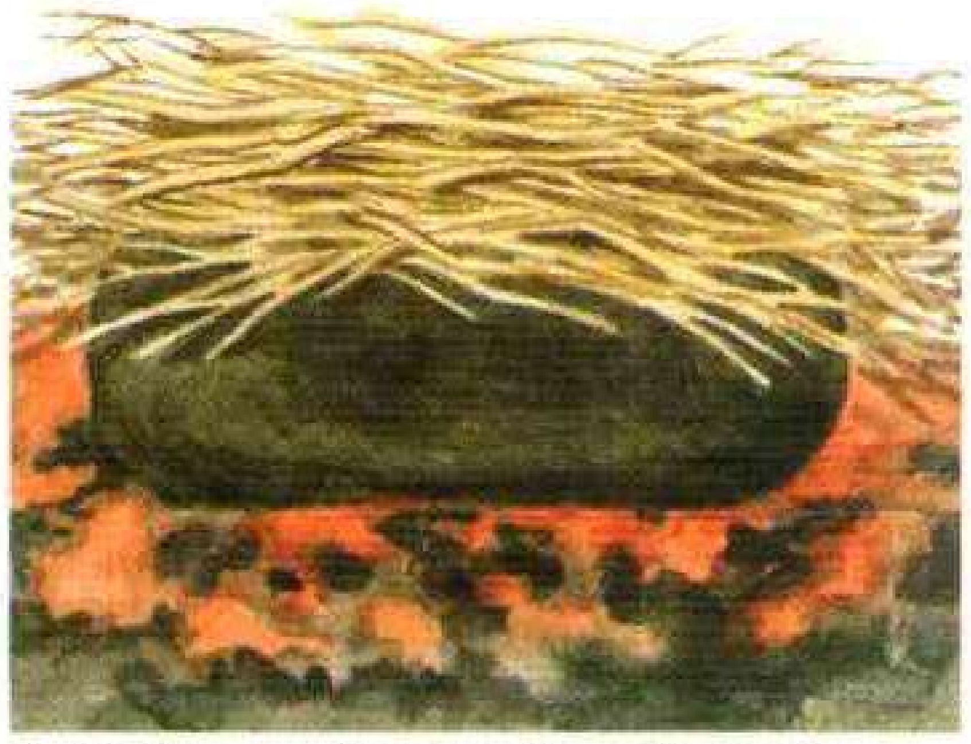

4 Coloque el paquete sobre un lecho de rescoldos calientes y construya un fuego en la parte superior. El pescado puede quedar cocido en una hora.

CONSEJOS CULINARIOS

- Las hierbas y especias son de gran valor al preparar la comida ya que la hacen más apetecible. También son útiles el ajo fresco, las cebollas y los cubos de caldo.
- La forma más sencilla de cocinar es hirviendo agua y utilizándola para hacer un estofado «todo en uno». La mezcla de una conserva de pescado con los bizcochos y los frutos secos parece desagradable, pero con la adición de curry en polvo y una gran cantidad de aire fresco, le parecerá una comida deliciosa.
- Si hierve sus alimentos en agua, es importante beberla también, para aprovechar al máximo el valor nutricional del alimento (a menos que lo hierva para eliminar las sustancias tóxicas).

PAN HÚMEDO *(DAMPER BREAD)*

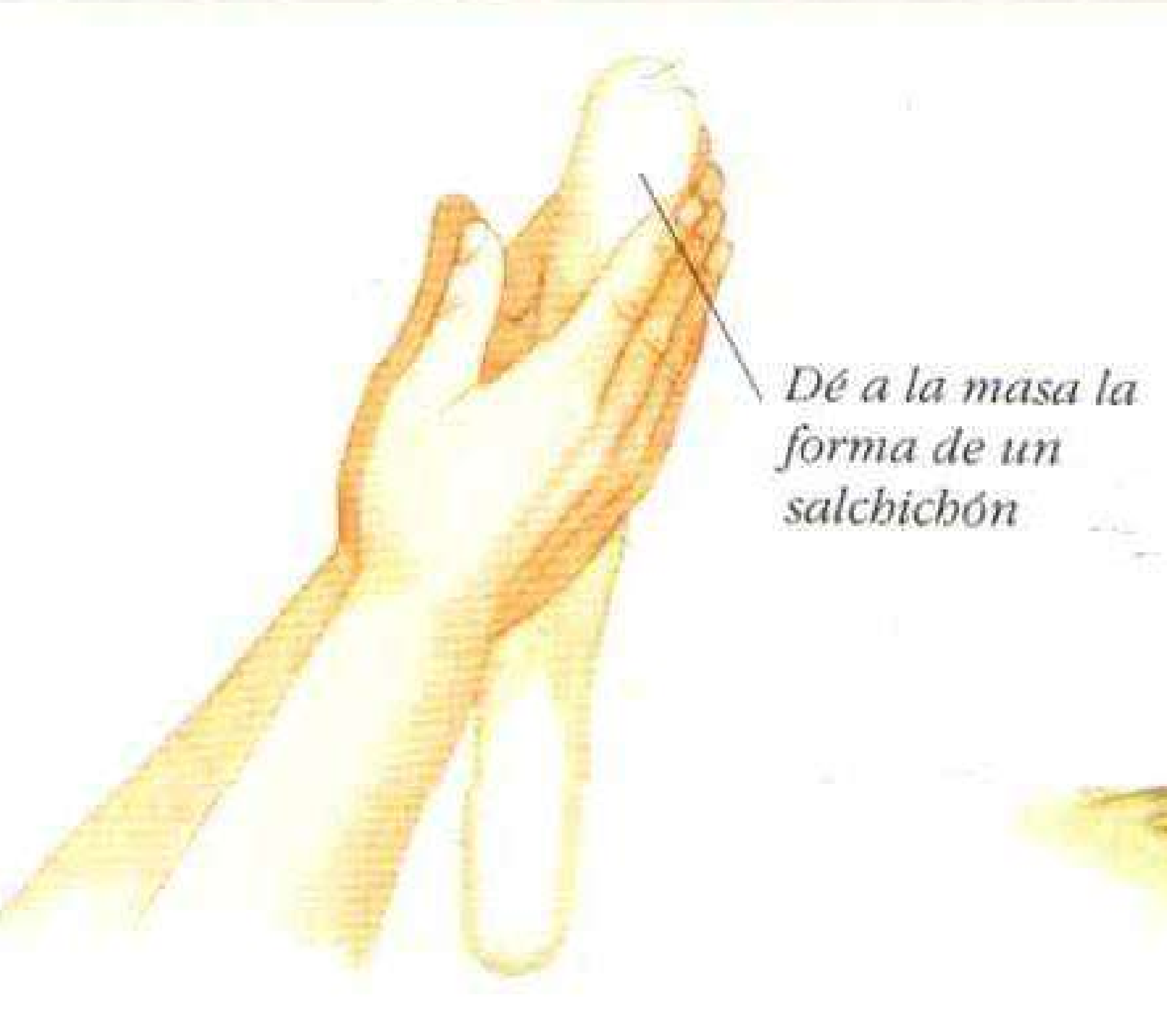
Dé a la masa la forma de un salchichón

1 Este manjar típico de un campamento australiano es fácil y rápido de preparar. En primer lugar, prepare una masa elástica a partir de harina y agua, y déle la forma de un salchichón largo y grueso.

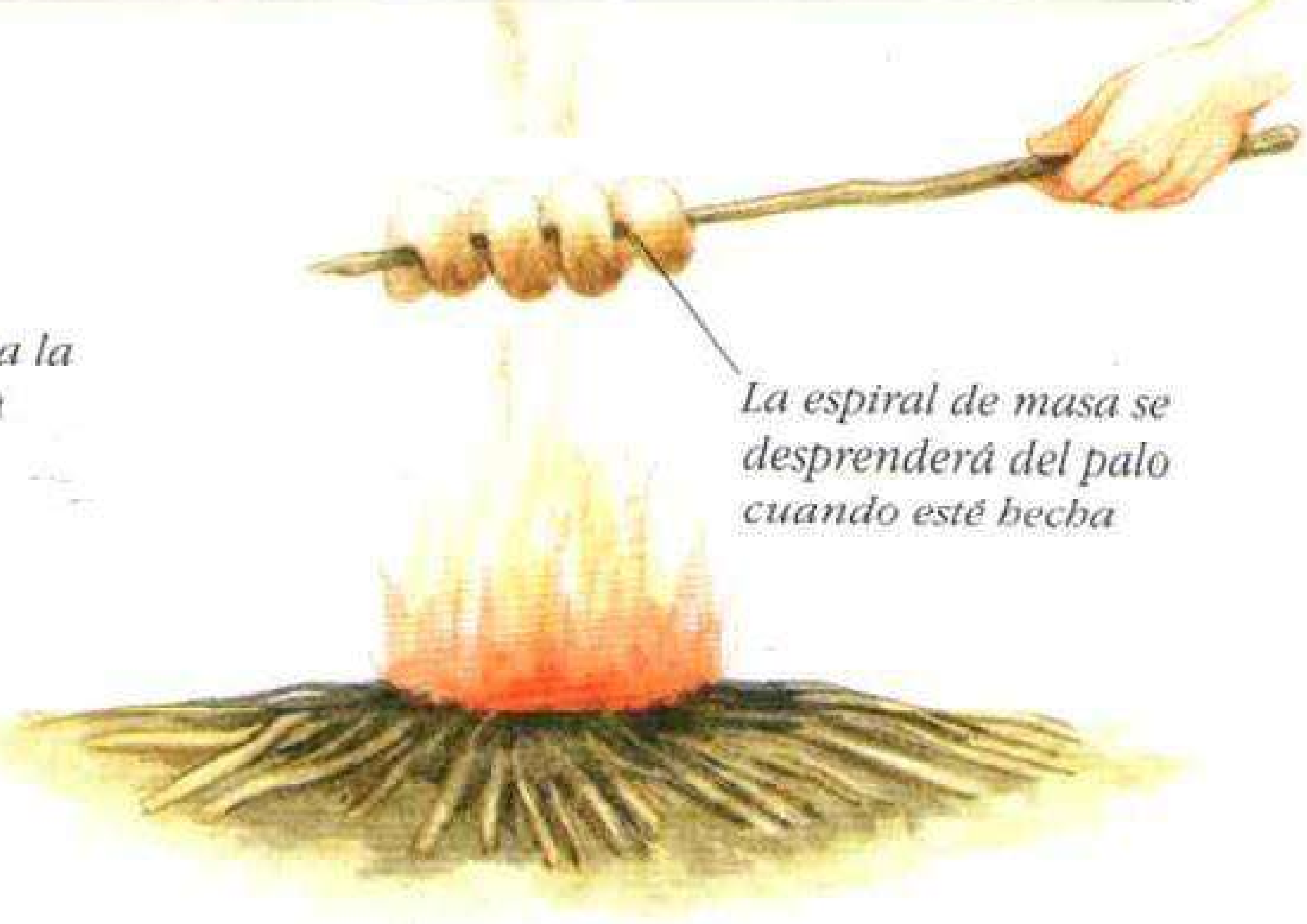
La espiral de masa se desprenderá del palo cuando esté hecha

2 Enróllelo en forma de espiral alrededor de un palo de madera y sujételo sobre los rescoldos de un fuego mientras se cuece. El pan se desprenderá del palo cuando esté listo.

LA CONSERVACIÓN DE LOS ALIMENTOS

UNA VEZ RECOLECTADOS, los alimentos comienzan a deteriorarse. Sin refrigeración, la carne fresca y el pescado son de consumo dudoso un día después de ser capturados en cualquier clima, a excepción de las regiones polares, mientras que las frutas y las verduras pierden sus valores nutricionales sin un almacenamiento adecuado. Para una situación de supervivencia a medio o largo plazo, debe recolectarse una gran cantidad de alimento en la estación de mayor abundancia, y conservarlo para las épocas de escasez, cuando resulta difícil seguir una dieta equilibrada. Recorrer los alrededores en busca de alimento para un solo día requiere una gran cantidad de calorías. En cualquier situación de supervivencia, a menos que sea muy corta, debe comenzar a conservar y almacenar alimentos inmediatamente, y a guardar sus provisiones, como las conservas, como raciones de emergencia real.

CÓMO SECAR LOS ALIMENTOS

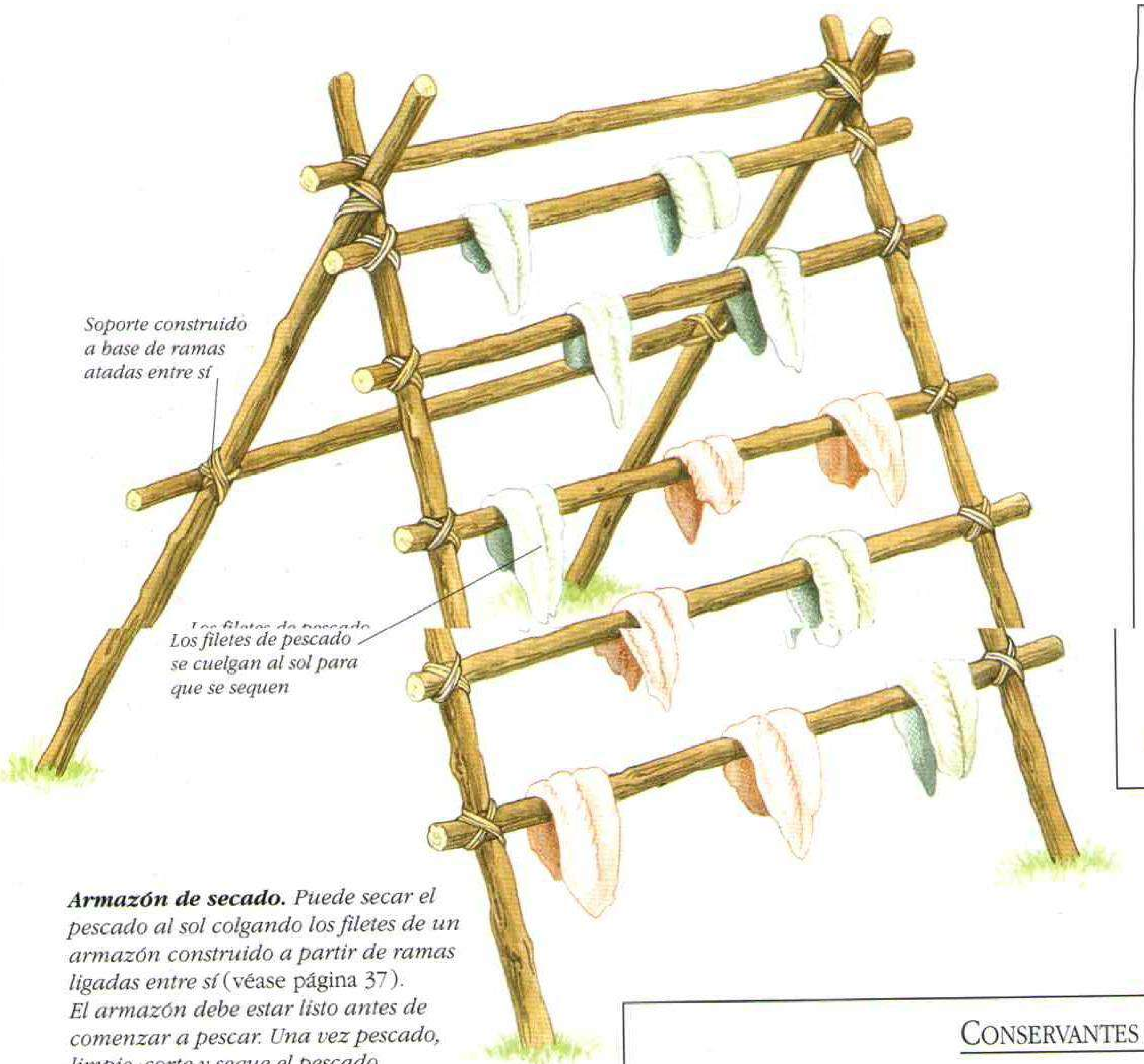

Soporte construido a base de ramas atadas entre sí

Los filetes de pescado se cuelgan al sol para que se sequen

Armazón de secado. *Puede secar el pescado al sol colgando los filetes de un armazón construido a partir de ramas ligadas entre sí* (véase página 37). *El armazón debe estar listo antes de comenzar a pescar. Una vez pescado, limpie, corte y seque el pescado inmediatamente* (véase página 110). *Es necesario un sol intenso para que el pescado se seque antes de que empiece a estropearse. Coloque el armazón en dirección hacia donde sople la brisa para alejar las moscas. Cuelgue los filetes de pescado del armazón dejando mucho espacio entre las piezas para permitir la circulación del aire. Los filetes estarán listos cuando toda el agua que contenían haya sido eliminada y muestren un aspecto quebradizo. Cuando los filetes de pescado estén suficientemente secos, puede almacenarlos para consumirlos más adelante, o puede molerlos entre dos piedras y utilizarlos para espesar sopas o estofados, o para preparar* pemmican (véase página siguiente).

OTROS MÉTODOS DE SECADO

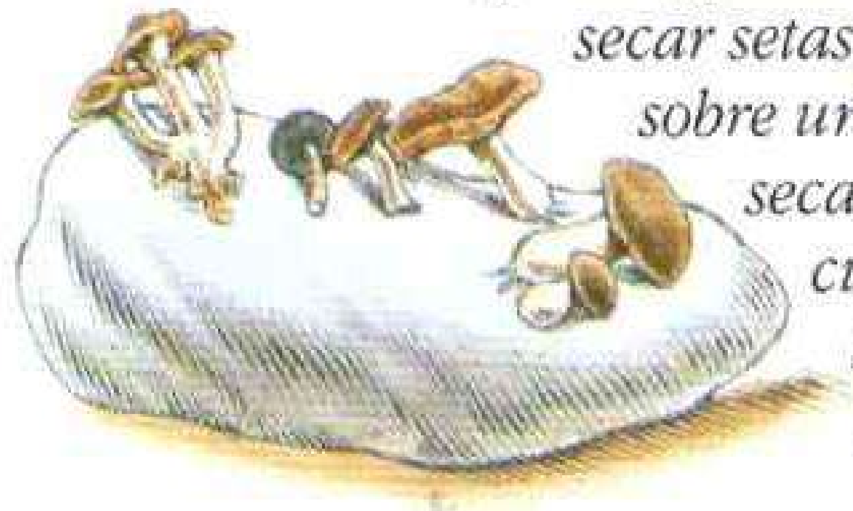

Secado natural. *En los climas cálidos y secos, puede poner a secar setas y plantas sobre una roca. Para secar hierbas, cuélguelas por los tallos, boca abajo.*

Secado sobre el fuego. *Para secar frutas, córtelas en rodajas, y cuélguelas de un espetón sobre un fuego humeante. Tenga cuidado de que el fuego no se avive demasiado y cocine la fruta.*

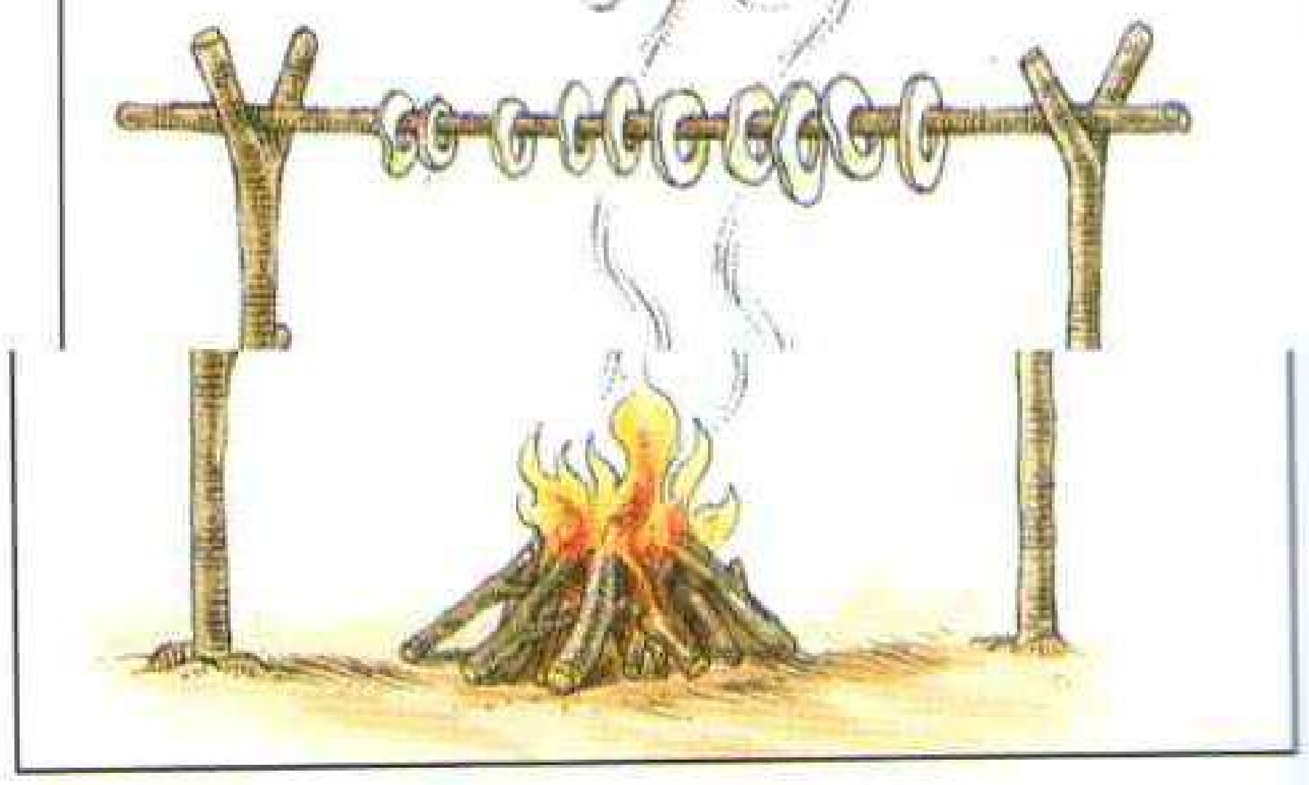

CONSERVANTES NATURALES

Puede conservar los alimentos empleando ingredientes naturales. Sale el pescado y sumerja los vegetales en jugo de lima o de limón, o hiérvalos en agua que contenga especias para encurtir. El pescado salado se conservará bien a temperaturas de hasta 11 °C.

SAL

ESPECIAS PARA ENCURTIR

LIMA

LIMÓN

CONSTRUCCIÓN DE UN TIPI PARA AHUMAR

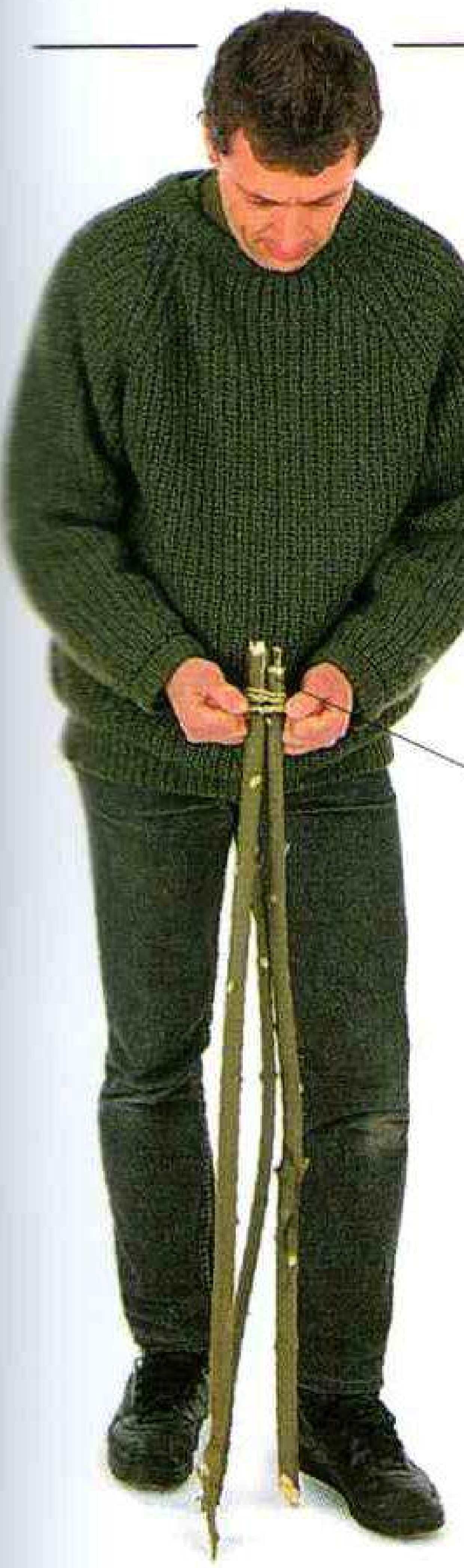
Ate las ramas con un cordel

1 Si no hay suficiente sol para secar el pescado, puede ahumarlo en un tipi especialmente construido para este fin. Ate tres ramas verticales por un extremo con un cordel *(véase página 37)*.

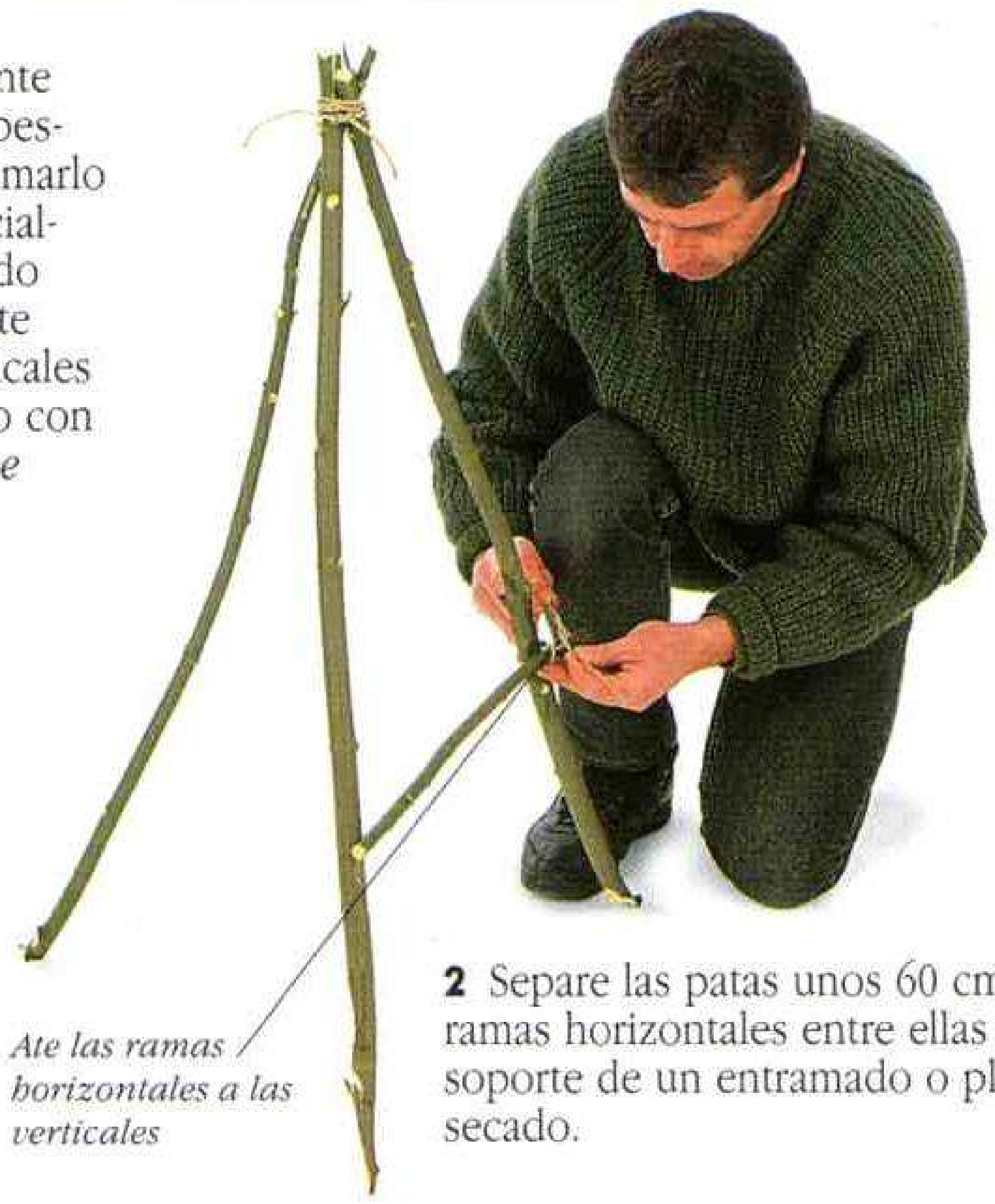
Ate las ramas horizontales a las verticales

2 Separe las patas unos 60 cm y ate unas ramas horizontales entre ellas para formar el soporte de un entramado o plataforma de secado.

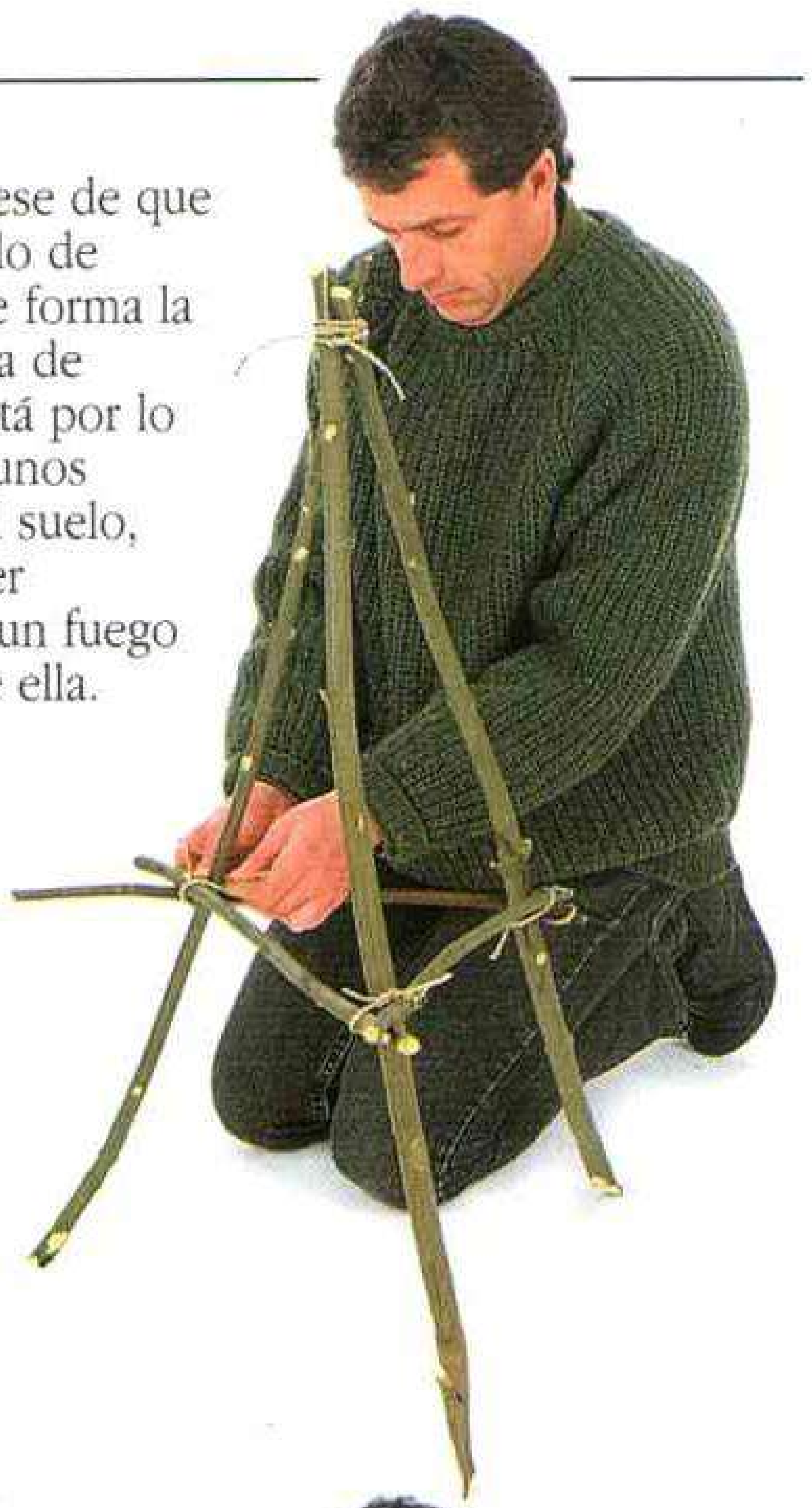

3 Asegúrese de que el triángulo de ramas que forma la plataforma de secado está por lo menos a unos 30 cm del suelo, para poder construir un fuego debajo de ella.

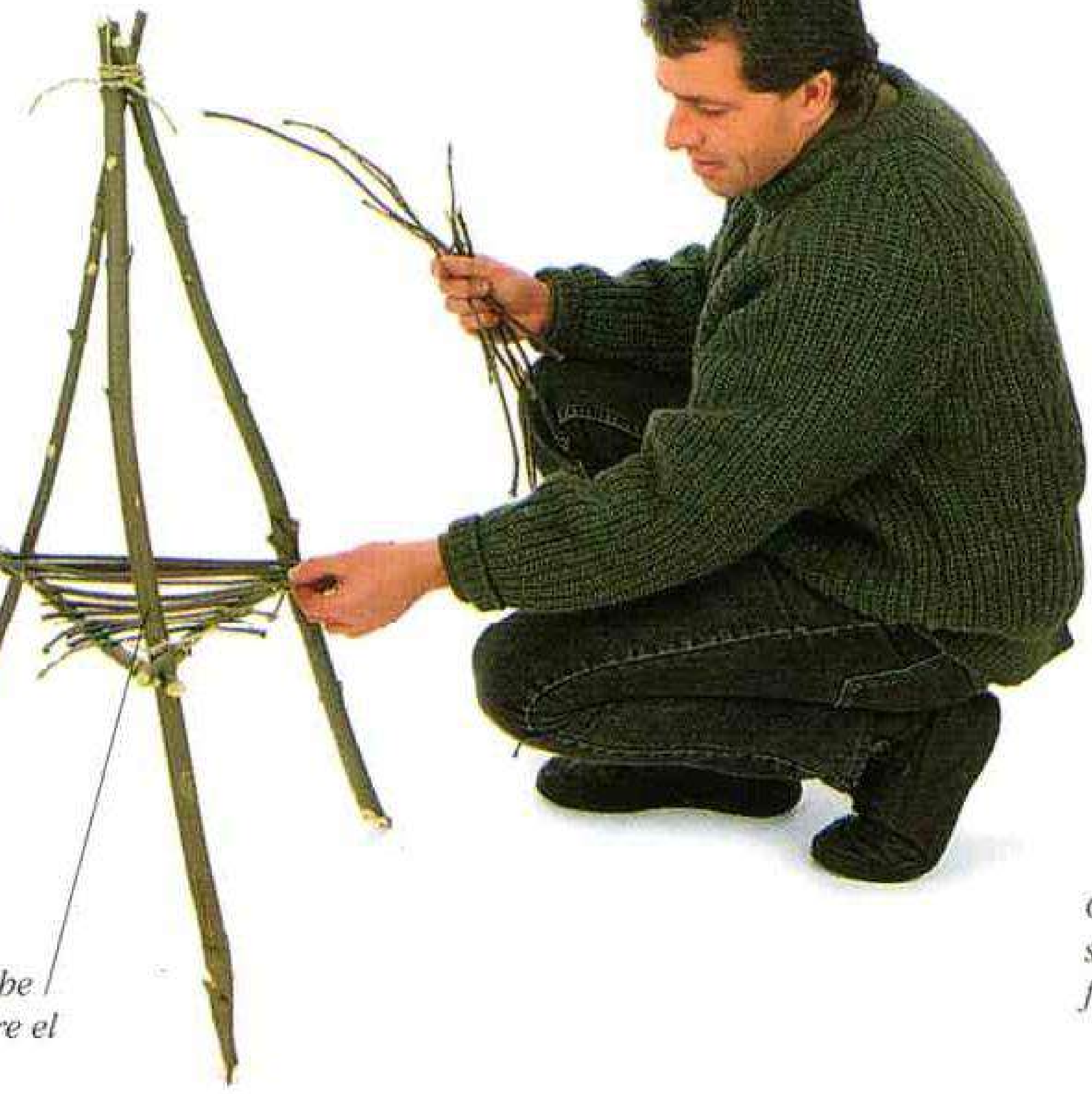
La comida debe colocarse sobre el soporte

4 Coloque ramas más pequeñas a través del triángulo para formar un entramado sobre el cual ahumar los alimentos.

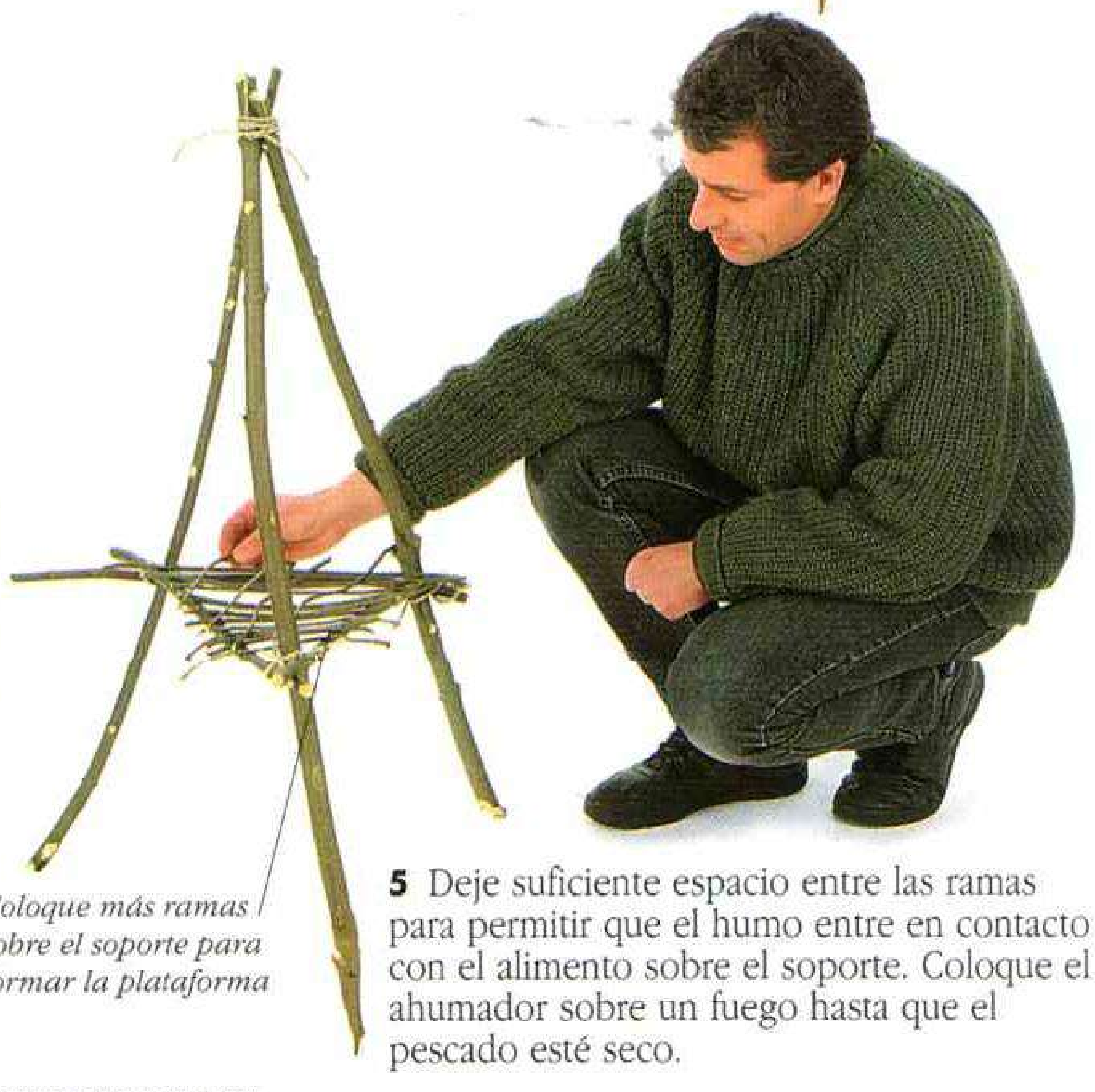
Coloque más ramas sobre el soporte para formar la plataforma

5 Deje suficiente espacio entre las ramas para permitir que el humo entre en contacto con el alimento sobre el soporte. Coloque el ahumador sobre un fuego hasta que el pescado esté seco.

CÓMO PREPARAR *PEMMICAN*

Amase la grasa con las manos

1 Puede conservar el pescado cocido preparando *pemmican*. En primer lugar, trabaje con las manos una porción de grasa de pescado para ablandarla.

2 Corte o ralle el pescado cocinado en trozos muy pequeños y mézclelos con la grasa. Puede añadir condimentos, si así lo desea.

Sazone con bayas y especias

3 Añada bayas para dar mejor sabor, y forme porciones del tamaño de una pelota de golf.

4 Dé a cada porción la forma de una salchicha y colóquela al sol para que se endurezca. El *pemmican* se conserva durante años y es muy nutritivo.

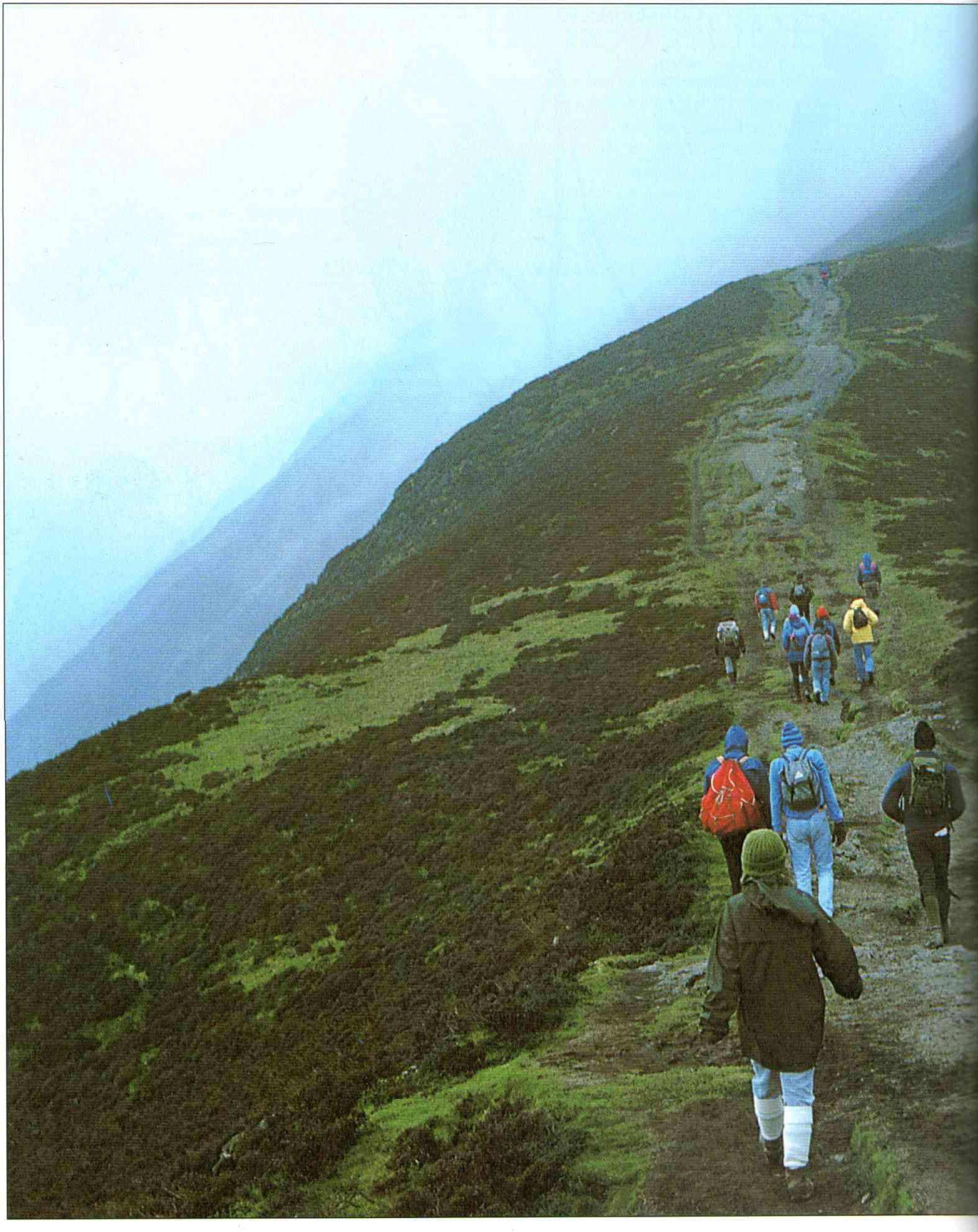

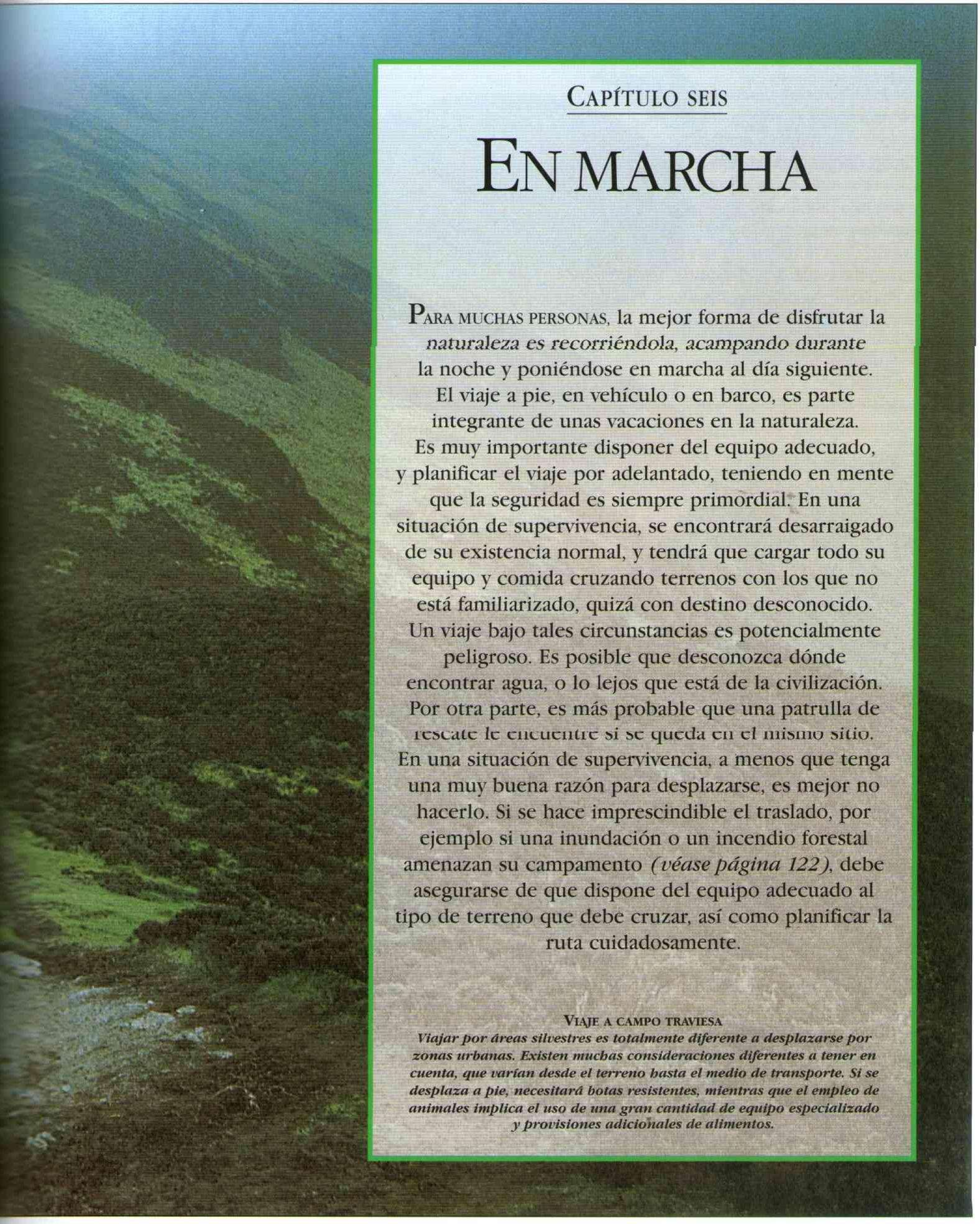

CAPÍTULO SEIS

EN MARCHA

PARA MUCHAS PERSONAS, la mejor forma de disfrutar la *naturaleza es recorriéndola, acampando durante* la noche y poniéndose en marcha al día siguiente. El viaje a pie, en vehículo o en barco, es parte integrante de unas vacaciones en la naturaleza. Es muy importante disponer del equipo adecuado, y planificar el viaje por adelantado, teniendo en mente que la seguridad es siempre primordial. En una situación de supervivencia, se encontrará desarraigado de su existencia normal, y tendrá que cargar todo su equipo y comida cruzando terrenos con los que no está familiarizado, quizá con destino desconocido. Un viaje bajo tales circunstancias es potencialmente peligroso. Es posible que desconozca dónde encontrar agua, o lo lejos que está de la civilización. Por otra parte, es más probable que una patrulla de rescate le encuentre si se queda en el mismo sitio. En una situación de supervivencia, a menos que tenga una muy buena razón para desplazarse, es mejor no hacerlo. Si se hace imprescindible el traslado, por ejemplo si una inundación o un incendio forestal amenazan su campamento *(véase página 122)*, debe asegurarse de que dispone del equipo adecuado al tipo de terreno que debe cruzar, así como planificar la ruta cuidadosamente.

VIAJE A CAMPO TRAVIESA
Viajar por áreas silvestres es totalmente diferente a desplazarse por zonas urbanas. Existen muchas consideraciones diferentes a tener en cuenta, que varían desde el terreno hasta el medio de transporte. Si se desplaza a pie, necesitará botas resistentes, mientras que el empleo de animales implica el uso de una gran cantidad de equipo especializado y provisiones adicionales de alimentos.

LA PLANIFICACIÓN DEL VIAJE

AUN FORMANDO PARTE de una expedición bien planificada, el viaje puede exponerle, junto con su grupo, a muchos riesgos imprevisibles. En una situación de supervivencia, estos riesgos son aún más pronunciados. Es importante estar preparado y tener un plan de acción. Si espera ser rescatado debe quedarse donde está. Sin embargo, puede verse forzado a abandonar su campamento por varias razones. Los animales pueden acabar con sus provisiones, o un desastre natural, como una inundación o un incendio forestal, pueden destruir la vegetación que le rodea. Antes de emprender el viaje, debe reconocer cuidadosamente el terreno que pretende cruzar.

Inundaciones. *Las inundaciones destruyen la vegetación y dejan una profunda capa de lodo que puede dificultar la búsqueda de plantas comestibles o de madera con la que construir refugios o encender fuego.*

EL RECONOCIMIENTO DEL TERRENO

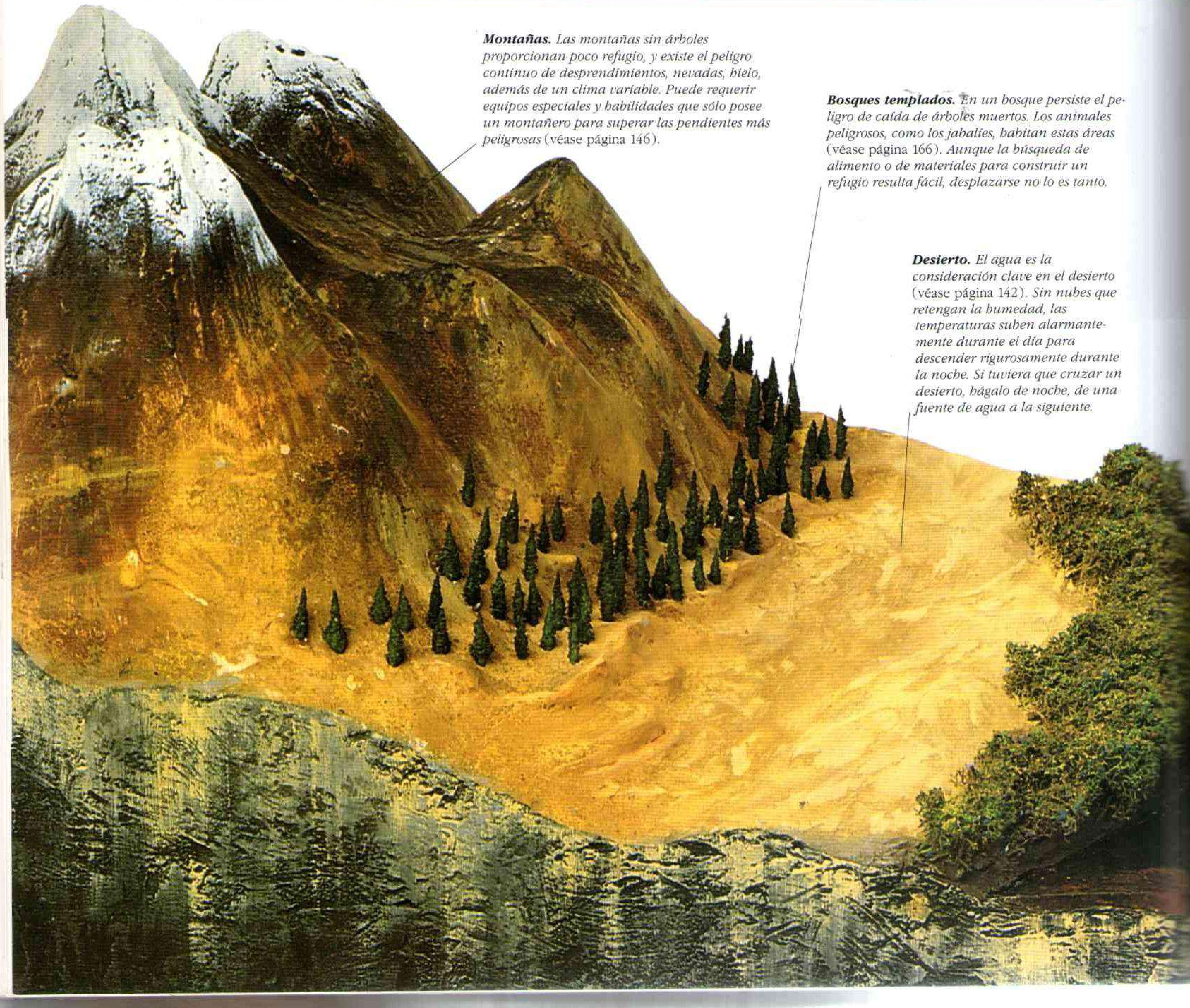

Montañas. *Las montañas sin árboles proporcionan poco refugio, y existe el peligro continuo de desprendimientos, nevadas, hielo, además de un clima variable. Puede requerir equipos especiales y habilidades que sólo posee un montañero para superar las pendientes más peligrosas* (véase página 146).

Bosques templados. *En un bosque persiste el peligro de caída de árboles muertos. Los animales peligrosos, como los jabalíes, habitan estas áreas* (véase página 166). *Aunque la búsqueda de alimento o de materiales para construir un refugio resulta fácil, desplazarse no lo es tanto.*

Desierto. *El agua es la consideración clave en el desierto* (véase página 142). *Sin nubes que retengan la humedad, las temperaturas suben alarmantemente durante el día para descender rigurosamente durante la noche. Si tuviera que cruzar un desierto, hágalo de noche, de una fuente de agua a la siguiente.*

EVALUAR A SUS COMPAÑEROS

Un grupo de supervivientes debe elegir a un líder, capaz de evaluar las capacidades de cada individuo antes de establecer un plan de viaje. El líder debe considerar, con cuidado y en privado, los puntos fuertes y débiles de cada miembro del grupo, y así establecer un plan de acción basándose en la capacidad de cada uno. El liderazgo en esta situación suele ser muy difícil, debido a la presión de una situación de vida o muerte y, especialmente, cuando algún individuo no acepta someterse a la dirección del líder.

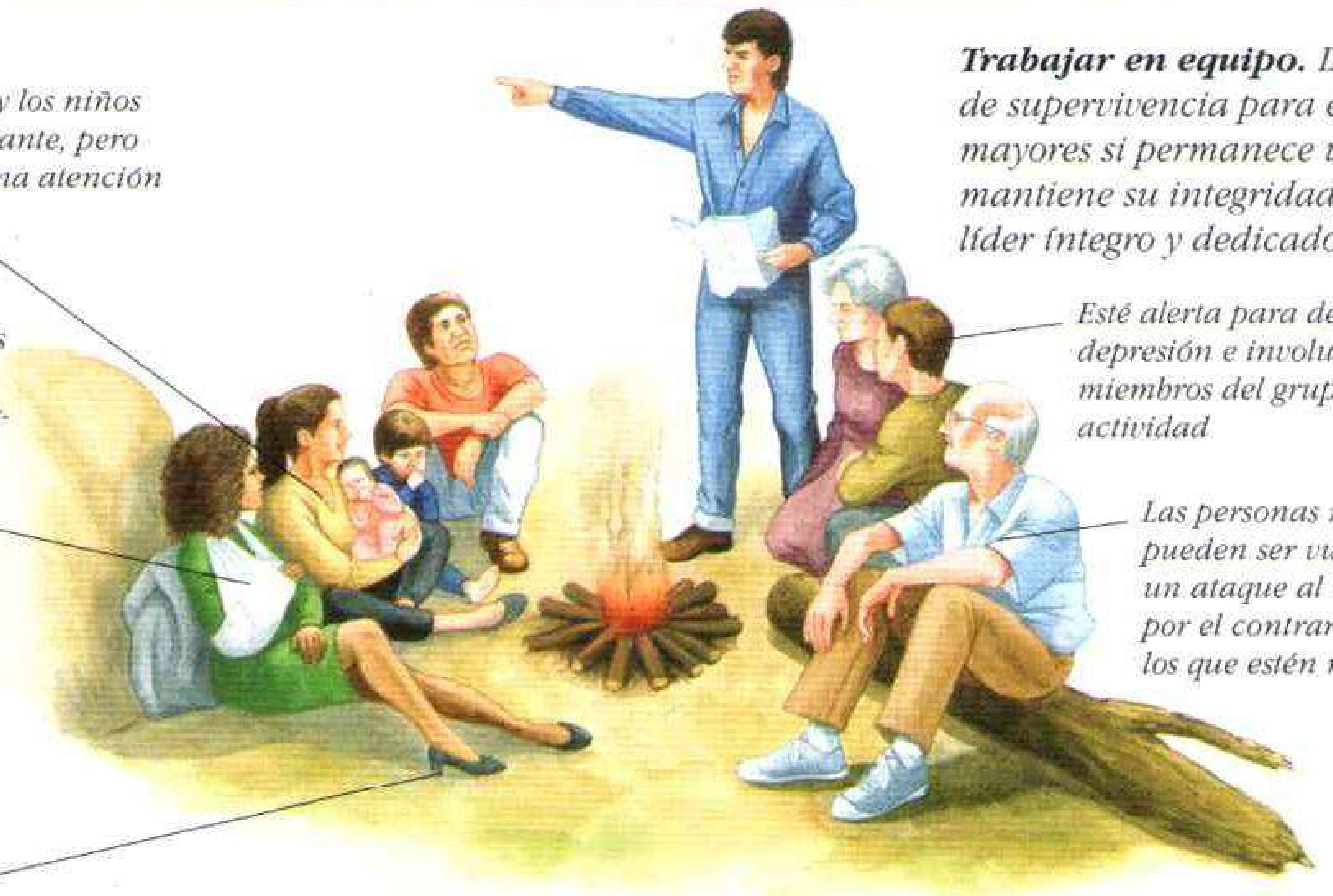

Las madres y los niños resisten bastante, pero necesitan una atención especial

Las personas heridas pueden retrasar a los demás y exponerlos a situaciones peligrosas

Puede que algún miembro del grupo no esté adecuadamente vestido o calzado para caminar

***Trabajar en equipo.** Las posibilidades de supervivencia para el grupo serán mayores si permanece unido y mantiene su integridad gracias a un líder íntegro y dedicado.*

Esté alerta para detectar síntomas de depresión e involucre a todos los miembros del grupo en alguna actividad

Las personas mayores pueden ser vulnerables a un ataque al corazón o, por el contrario, que sean los que estén más en forma

***Estar preparados.** Antes de desplazarse, debe hacerse con la mayor cantidad de información posible acerca de la región que pretende cruzar. No suponga siempre que los mapas son exactos* (véase página 126)*, en particular cuando planifique buscar agua. Vale la pena enviar una serie de expediciones cortas, con equipo ligero, para reconocer la zona antes de emprender cualquier desplazamiento.*

EL EMPLEO DE PRISMÁTICOS

Unos prismáticos le ahorrarán muchos kilómetros. Ajuste la imagen hasta que sea nítida y observe el terreno. Explore la zona desde distintos puntos de observación, e intente descubrir por dónde fluyen las corrientes o los ríos, o la situación de los montes.

***Escrutar.** Recorra el paisaje con la mirada, observando cuidadosamente, y no simplemente mirando los alrededores.*

***Ríos.** Los ríos son peligrosos, y debe acercarse a ellos con cuidado. Para cruzar un río, puede que tenga que buscar troncos caídos para emplearlos como puente, o buscar zonas más angostas y poco profundas para vadearlos* (véase página 150)*. También pueden estar habitados por animales peligrosos, como los cocodrilos* (véase página 75).

***Selvas.** El calor y la humedad de la selva pueden llegar a ser muy molestos, además de poco saludables para los humanos, pero permiten el desarrollo de los insectos y las bacterias. Las heridas más pequeñas se infectan, la comida se descompone y la ropa está continuamente empapada debido a la lluvia o al sudor, o a ambos factores. En la selva, los desplazamientos se hacen particularmente difíciles.*

***Sabana.** El agua es un factor importante a considerar en este ecosistema. Puede tener problemas para encontrar refugio* (véase página 53)*. Recuerde que también estará expuesto al peligro que suponen los depredadores que cazan los grandes rebaños de herbívoros que pastan habitualmente en estas enormes áreas.*

***Mar y playa.** No subestime nunca el mar. Siempre es peligroso. En las zonas de mareas es mucho más fácil caminar por la arena durante la marea baja que superar las dunas o los acantilados. Sin embargo, debe vigilar continuamente para asegurarse de que sabe en qué dirección viene la marea. Puede encontrarse con criaturas peligrosas en la playa o en las aguas poco profundas* (véase página 109).

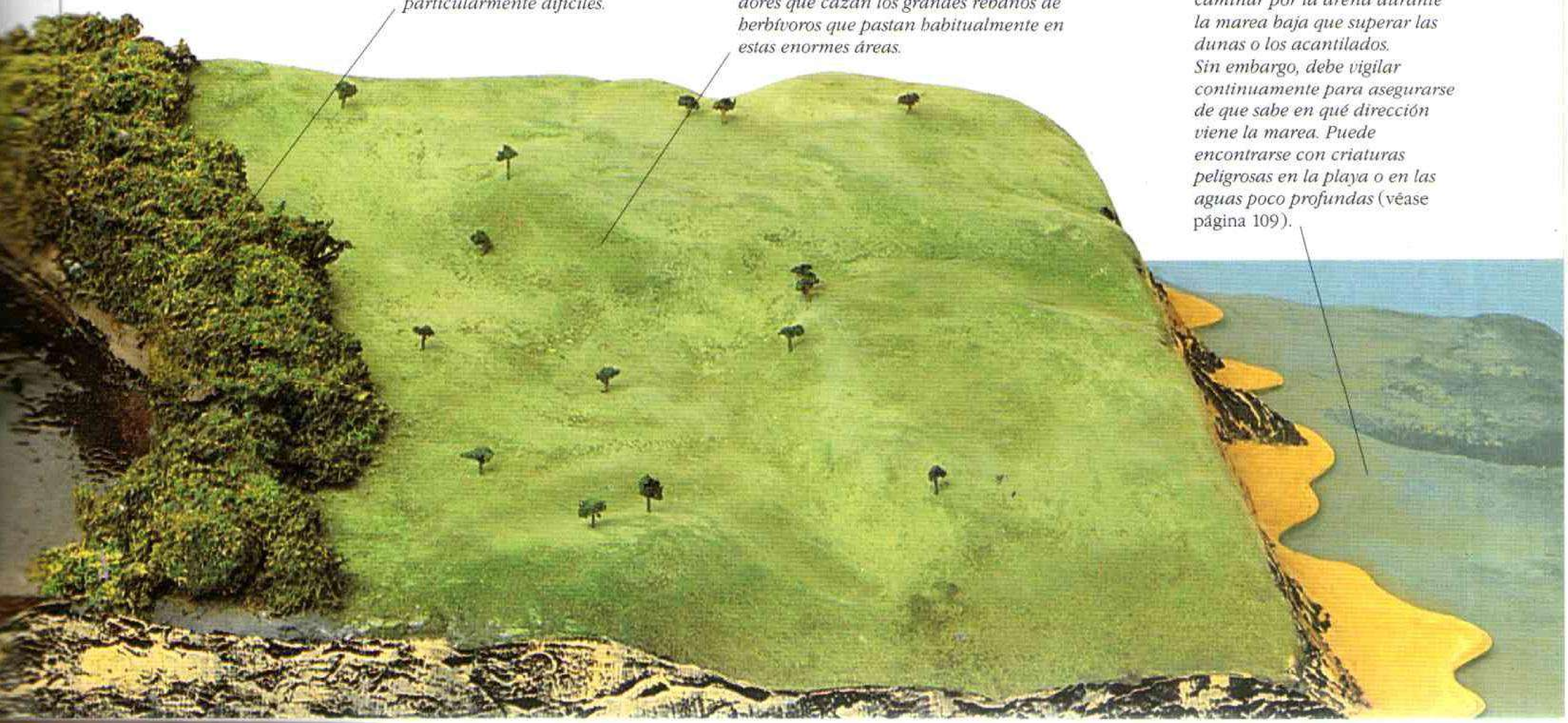

LA ORIENTACIÓN

LA ORIENTACIÓN ES el aspecto más importante de la navegación, así que una brújula fiable se convertirá en su herramienta más importante. Un reloj o cronómetro le será de gran utilidad, ya que puede indicar tanto la dirección como el tiempo recorrido, con lo que podrá estimar la distancia recorrida *(véase página siguiente).* Existen muchos tipos diferentes de brújula que funcionan atraídos por los polos del campo magnético de la Tierra. Si no dispone de una brújula, podrá orientarse observando la posición del sol y de las estrellas, los animales y las plantas.

BRÚJULAS

Las agujas de una brújula señalan el norte magnético y el sur magnético, los polos del campo magnético natural de la Tierra. Imagínese que la Tierra contiene un gran imán vertical, y que los polos se encuentran en la parte superior e inferior del imán, de los cuales emana la carga magnética. Hay muchos tipos diferentes de brújula. Si únicamente pretende dar un paseo, basta con una brújula sencilla para un caso de emergencia *(véase página 28)*, pero utilice una brújula como transportador con fines de navegación y orientación básica *(véase inferior).* Las brújulas prismáticas incluyen un prisma a través del cual se toman los rumbos, y son resistentes y precisas. Algunas de estas brújulas incluyen un espejo además de, o en lugar del prisma. Sin embargo, son bastante caras. Una brújula Silva es lo único que necesita realmente para una navegación precisa, y le evitará complicaciones que pueden provocar errores.

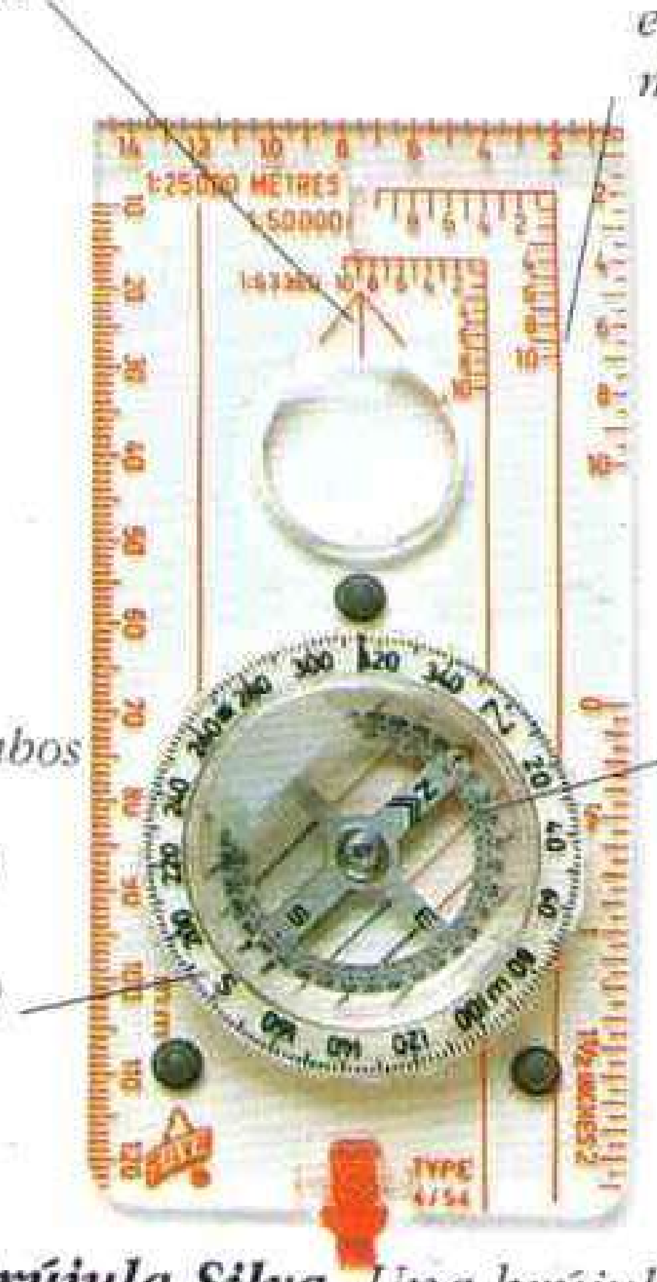

La flecha indica la dirección a seguir una vez la brújula ha sido regulada

Medidas de distancia para las diferentes escalas de los mapas

Los rumbos se leen sobre el dial (limbo)

Las líneas paralelas sirven para orientar la brújula sobre las cuadrículas norte-sur del mapa

Brújula Silva: *Una brújula con transportador le permite medir rumbos sobre un mapa sin tener que mover la brújula* (véase página 130). *De esta manera evita el tener que usar un transportador aparte y tener que hacer cálculos aritméticos mentales para determinar la ruta.*

CÓMO FABRICAR UNA BRÚJULA

Si se suspende libremente una pieza de metal ferroso (que contiene hierro) imantado, girará hasta orientarse en un eje norte-sur. Esto se debe a que se ve atraída por el campo magnético de la Tierra. Aprovéchese de este fenómeno para construir una brújula, imantando un trozo de metal.

En lugar de utilizar un paño de seda para imantar una aguja, puede utilizar un imán, que hará la misma función pero más deprisa

1 Magnetice una aguja frotándola repetidamente en una dirección sobre un paño de seda, como su pañuelo *(véase página 142).* Esto genera una carga de electricidad estática.

Permita que la aguja se mueva libremente haciéndola flotar en un cuenco lleno de agua

2 Haga flotar una brizna de hierba en un cuenco lleno de agua, y coloque la aguja sobre ella. La aguja se orientará por sí misma apuntando en dirección norte-sur.

3 Una vez la aguja esté fija, compruebe la posición del norte utilizando otros indicadores *(véase página siguiente)* y marque el extremo norte de la aguja. Ahora puede utilizarla como brújula.

UTILICE UNA CUCHILLA DE AFEITAR COMO BRÚJULA

En un caso de emergencia, puede construir una brújula con una cuchilla de afeitar en lugar de una aguja, si dispone de ella. Sin embargo, tenga la precaución de desgastar la hoja sobre una piedra, y tenga mucho cuidado al imantarla.

Cuchilla de afeitar como brújula. *Imante una cuchilla de afeitar sin filo, y suspéndala de un hilo para que cuelgue libremente y apunte hacia el norte.*

CÓMO ORIENTARSE A TRAVÉS DEL SOL

Como el sol siempre sale por el este y se oculta por el oeste, puede ser utilizado como indicador de dirección. Mientras pueda verlo y sepa la hora, podrá orientarse en relación al sol. También puede dar con la dirección aproximada construyendo un reloj de sol, o empleando un reloj como transportador.

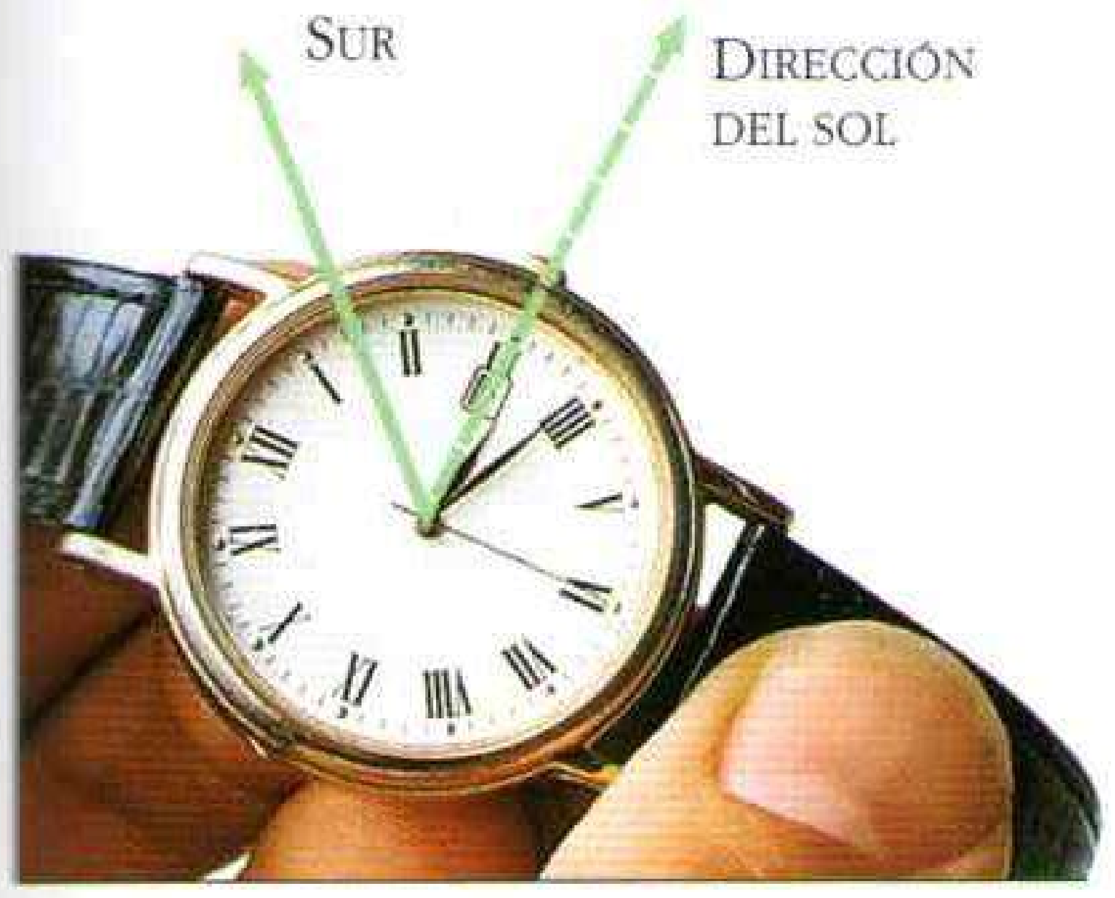

Hemisferio norte. *Apunte hacia el sol la manecilla horaria. Imagine una línea que divida en dos el ángulo entre la manecilla horaria y las 12.00h. Esta línea apuntará hacia el sur.*

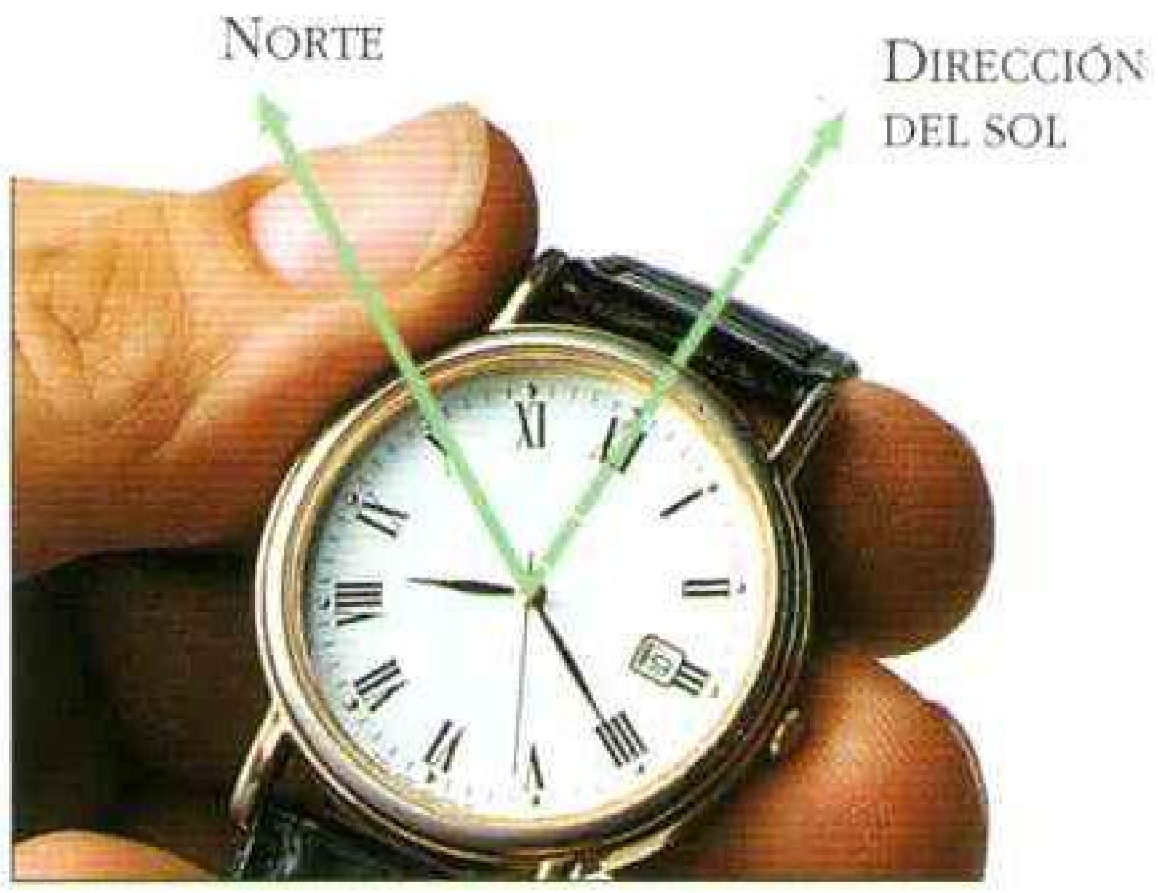

Hemisferio sur. *Si se encuentra en el hemisferio sur, apunte las 12.00h del reloj hacia el sol. El norte se encuentra en el punto medio entre las 12.00h y la manecilla horaria.*

RELOJ DE SOL

Clave una rama vertical en el suelo. Por la mañana, observe dónde cae la sombra y marque el punto exacto en el suelo (será el oeste en el hemisferio norte). Ate una cuerda de la misma longitud de la sombra a un palo pequeño y dibuje un arco alrededor de la rama. Marque el punto en el que la sombra toca el arco por la tarde (el este).

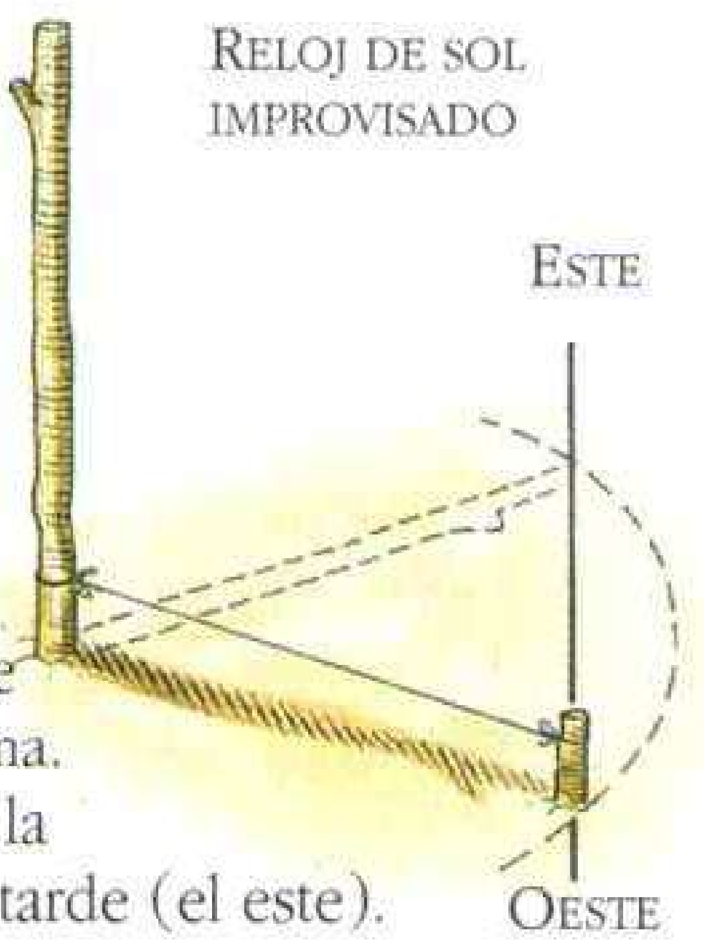

CÓMO ORIENTARSE A TRAVÉS DE LAS ESTRELLAS

Las estrellas permanecen siempre en la misma posición. Sólo parecen moverse en el cielo debido al movimiento de la Tierra. Únicamente hay una estrella que aparenta no moverse: la Estrella Polar. Esta estrella puede utilizarse en el hemisferio norte para encontrar el norte. En el hemisferio sur, debe localizar la Cruz del Sur para encontrar el sur.

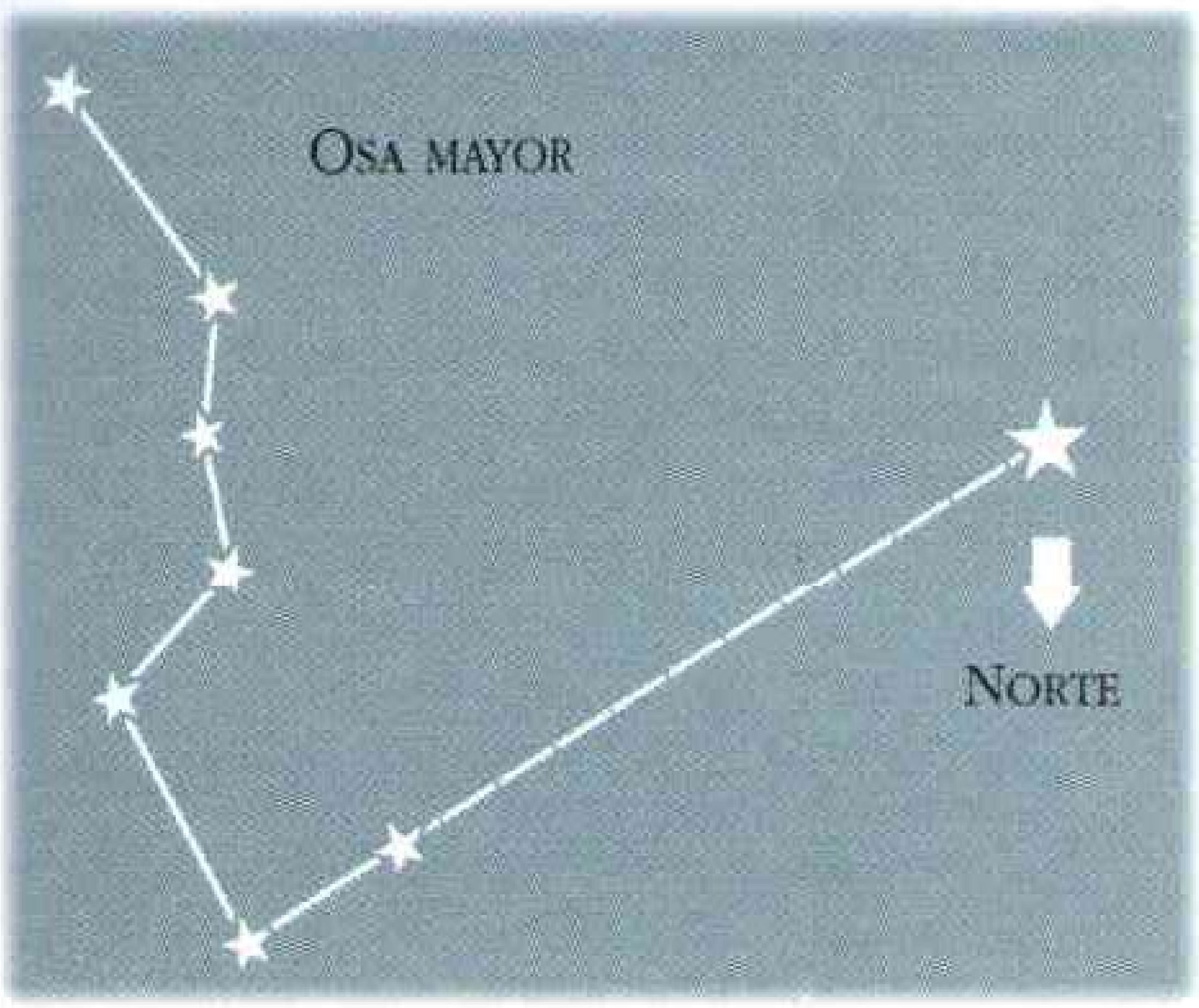

Hemisferio norte. *Mirando hacia la Osa Mayor (el Carro), trace una línea imaginaria entre las dos estrellas que forman la parte frontal de la Osa y prolónguela multiplicada por cuatro hasta encontrar la Estrella Polar. Esta brillante estrella está situada al norte en el horizonte.*

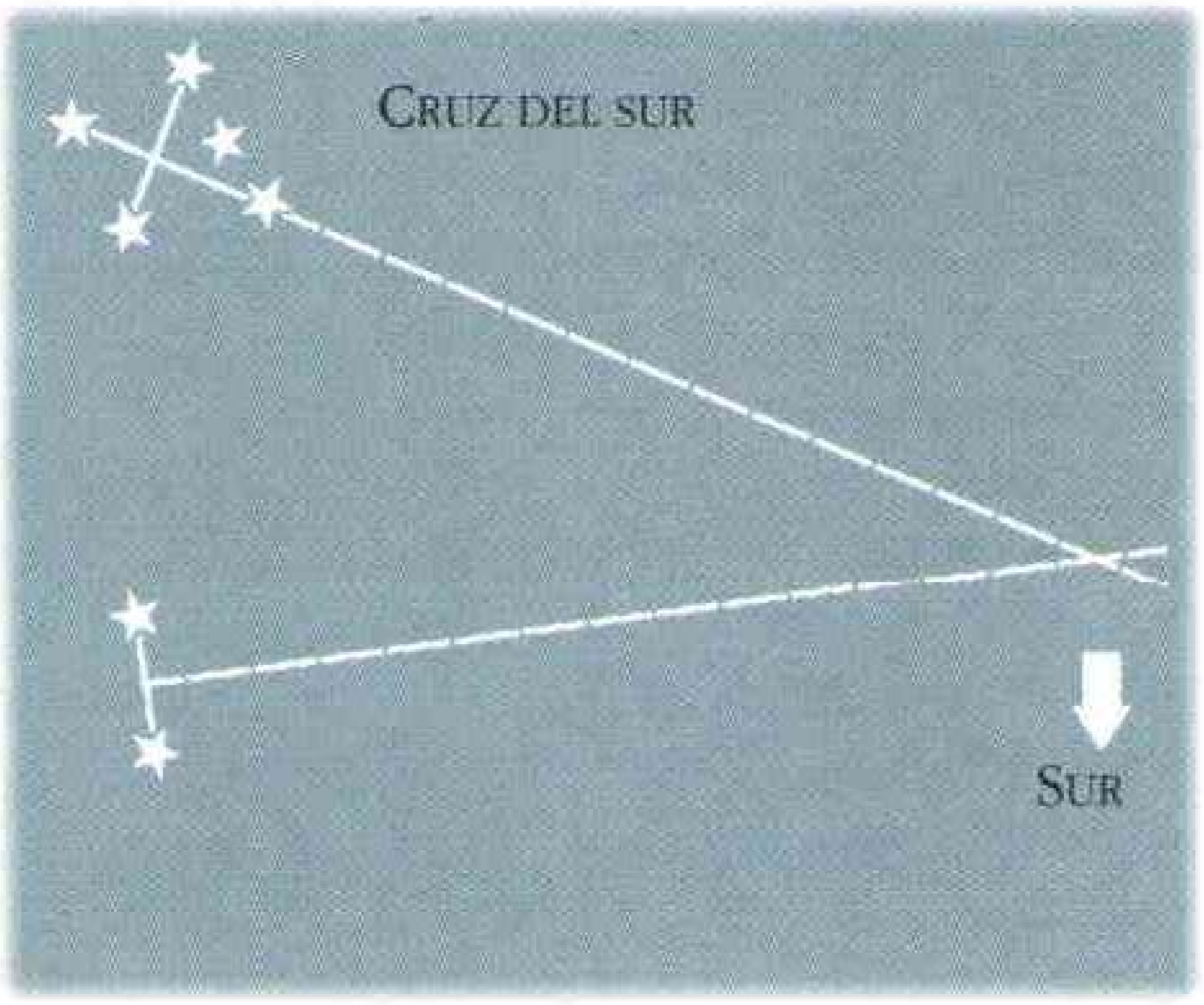

Hemisferio sur. *Utilice la Cruz del Sur para saber de forma aproximada dónde está el sur. Trace una línea imaginaria desde el travesaño, aproximadamente unas 4 veces y media su longitud. El sur debe estar en el horizonte debajo de este punto. Dos estrellas brillantes debajo de la Cruz le ayudarán a encontrar el punto correcto.*

INDICADORES DE DIRECCIÓN NATURALES

La naturaleza está mejor orientada y adaptada al campo magnético de la Tierra que los humanos. Las plantas siempre crecen en dirección al sol —hacia el sur en el hemisferio norte y hacia el norte en el hemisferio sur. Los girasoles siguen el movimiento del sol cruzando el cielo. La planta denominada Polo Norte en Sudáfrica se inclina hacia el norte. Los anillos de los árboles son a menudo mucho más anchos en el lado en el que da el sol.

Nidos de termitas. *Algunos nidos de termitas están construidos siguiendo un eje norte-sur. De esta manera reciben el máximo calor durante la mañana y la tarde, y están protegidos del sol al mediodía.*

Pájaros tejedores. *Algunos pájaros pueden utilizarse como indicadores de dirección básicos. Por ejemplo, algunas especies de pájaros tejedores construyen sus nidos únicamente en la cara oeste de los árboles. Si se encuentra en una zona donde viven estos pájaros —por ejemplo, en Sudáfrica— le servirán como indicativo de dirección bastante fiable.*

EL EMPLEO DE MAPAS

EXISTEN MUCHOS TIPOS diferentes de mapas, cada uno diseñado para un propósito diferente. La elección del mapa adecuado resulta fundamental si se adentra en un territorio desconocido. Necesitará mapas a gran escala de la totalidad del área que pretenda visitar, así como mapas detallados de aquellas zonas que recorrerá a pie. Los mapas a escala 1:50.000 son ideales para traslados a pie (esta escala significa que, por cada unidad de medida en el mapa, la distancia real es 50.000 veces mayor). Por lo general, tienen una cuadrícula de 2 cm, cada una de las cuales representa 1 km, y están marcados con dos coordenadas para cada línea, que va de 0 a 99. Por muchas razones, algunos mapas pueden ser bastante inexactos. Modifíquelos sobre la marcha y aprenda a sortear esta falta de exactitud.

CÓMO DESCIFRAR UN MAPA

El agua determina el paisaje, ya que lo cruza formando corrientes y ríos, creando a su paso colinas, montañas y valles. Las construcciones humanas, como las carreteras, pueden cambiar rápidamente. Para entender las características de un terreno utilizando un mapa, identifique en primer lugar los ríos, luego los valles y las colinas, que apenas varían durante la vida de un mapa.

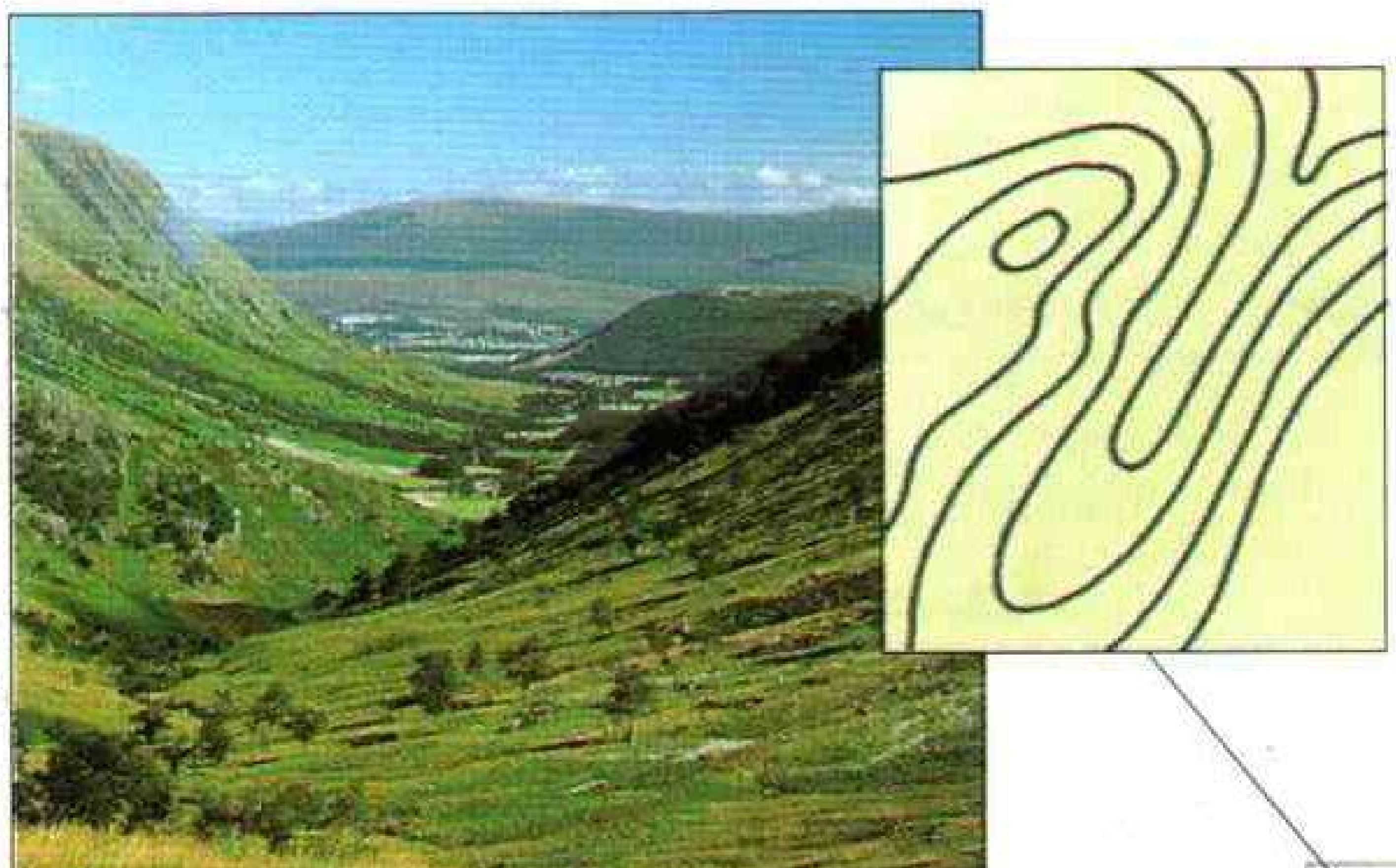

Valle. *Un valle, y el río que lo ha formado, aparecen en un mapa como un remolino complicado de líneas de igual altura, mostrando su forma exacta, en particular las zonas que no son visibles para el observador que se acerca a ellas desde tierra. La altura sobre el nivel del mar se marca con números a lo largo de las líneas.*

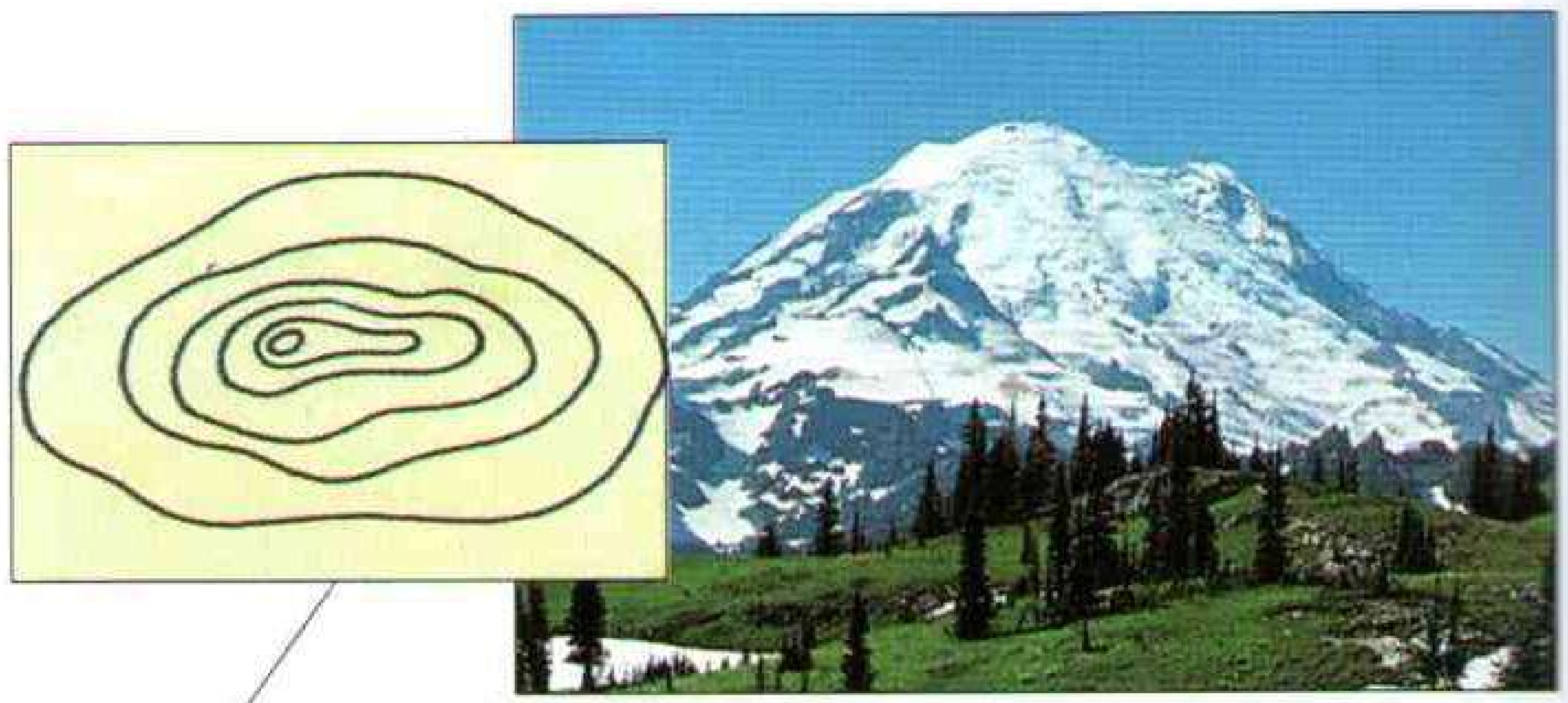

Montaña. *Un mapa muestra un plano completo de la montaña, la mayor parte de la cual no es visible desde cualquier posición en tierra. Debe imaginarse la parte que no ve desde el punto en que esté situado. La altura se indica en el mapa como si estuviera escrita sobre la ladera.*

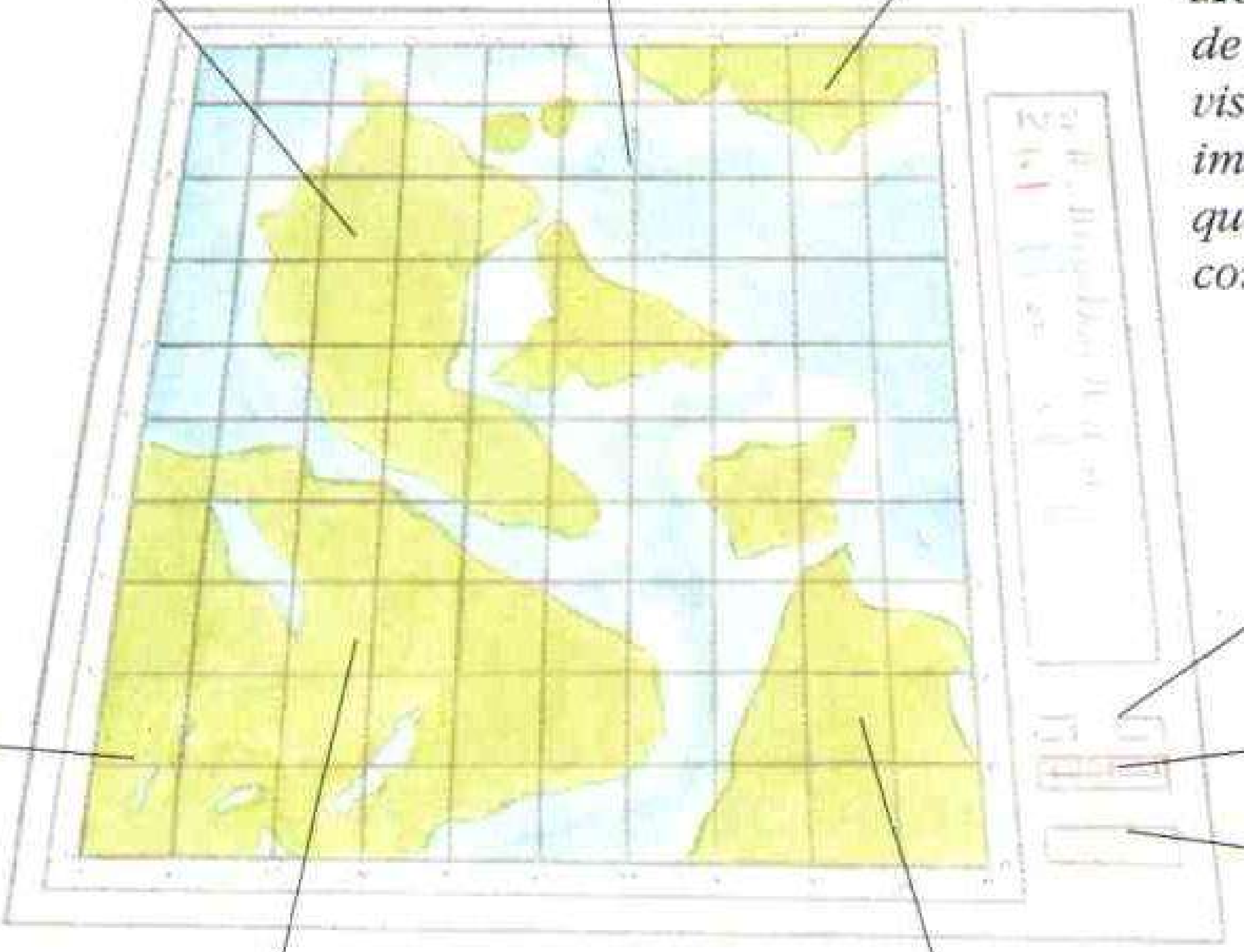

Las líneas verticales se llaman meridianos

Las líneas horizontales se llaman paralelos

Clave para la información adicional al mapa, como las ampliaciones

Detalles sobre la variación magnética

Escala

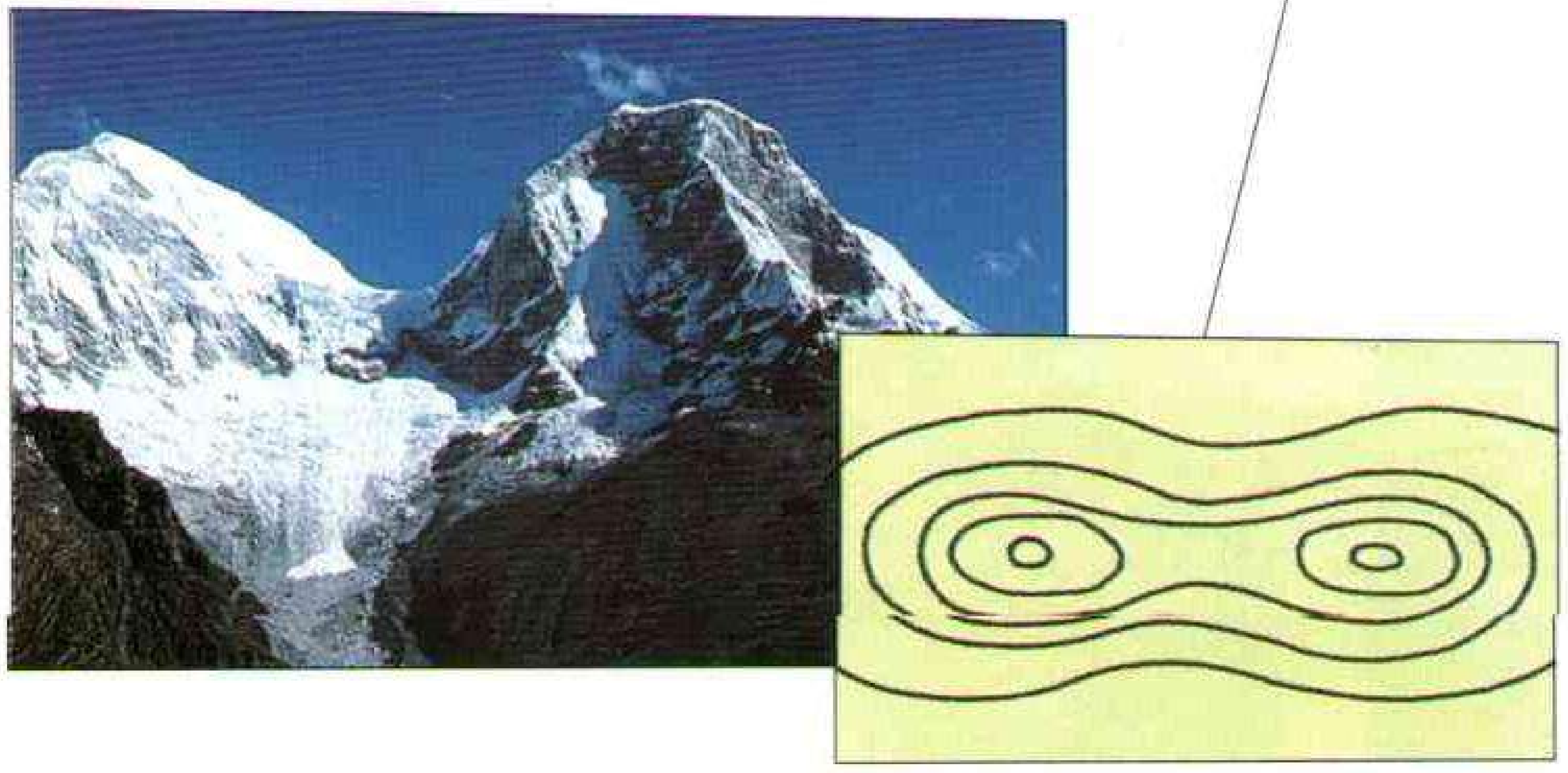

Collados. *Un collado es una depresión en una cordillera entre dos montañas o picos. A nivel del suelo, si mira desde un extremo, únicamente verá una de las montañas. En un mapa, un paso aparece representado por dos círculos enlazados por curvas de nivel.*

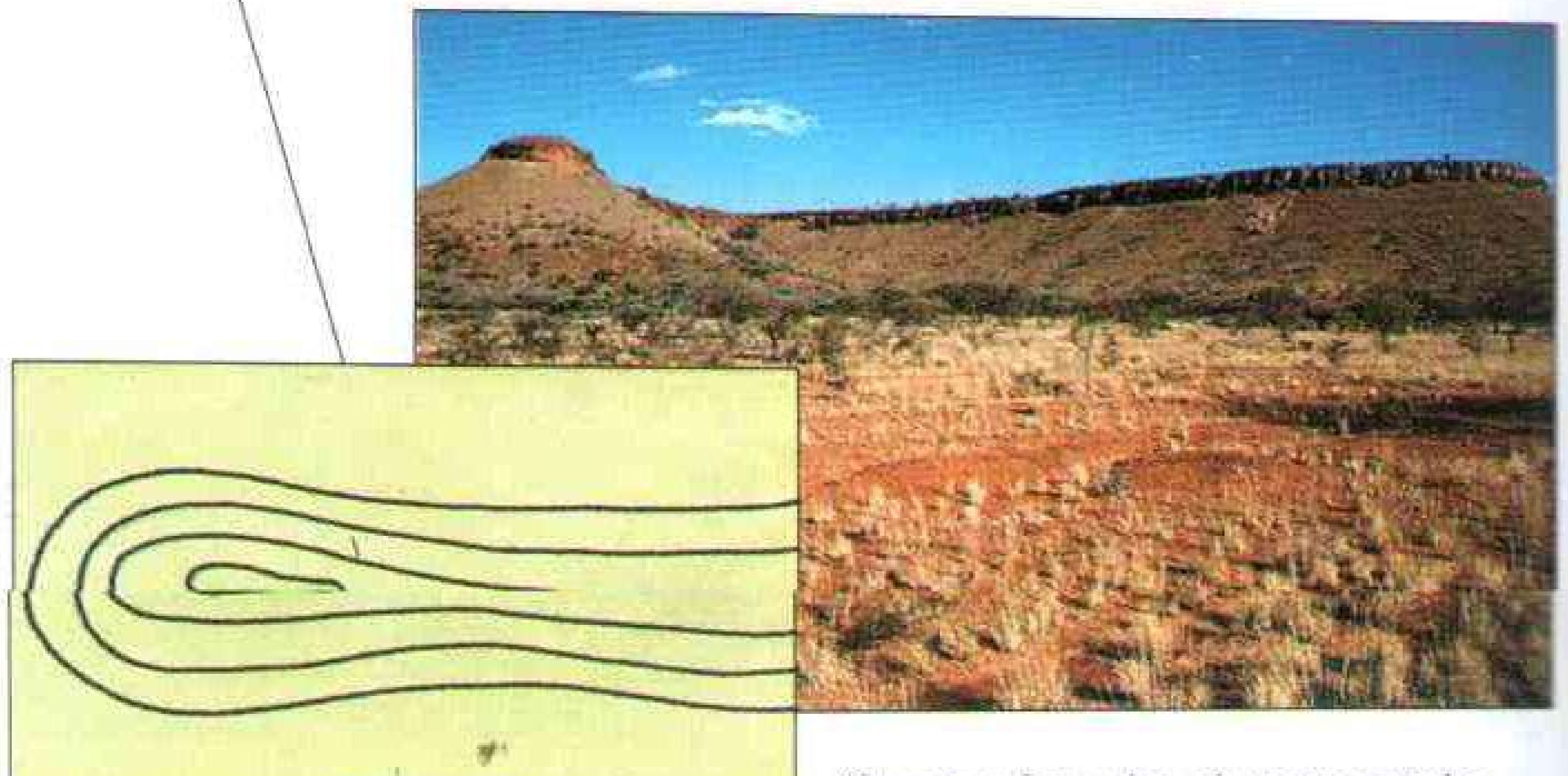

Cresta. *Los ríos siempre están separados por crestas. Éstas se dibujan en un mapa como los dedos de una mano, sobresaliendo de alguno de los lados de una montaña. A menudo es más fácil caminar a lo largo de la parte superior de la cresta que por el valle, en particular al cruzar una selva densa.*

GRADIENTES

Las curvas de nivel en un mapa unen puntos de igual altitud. Las alturas de algunas de las curvas de nivel están indicadas en los mapas. Por lo general, se escriben encarando la parte inferior de la pendiente, como si estuvieran escritas sobre la ladera de la montaña.

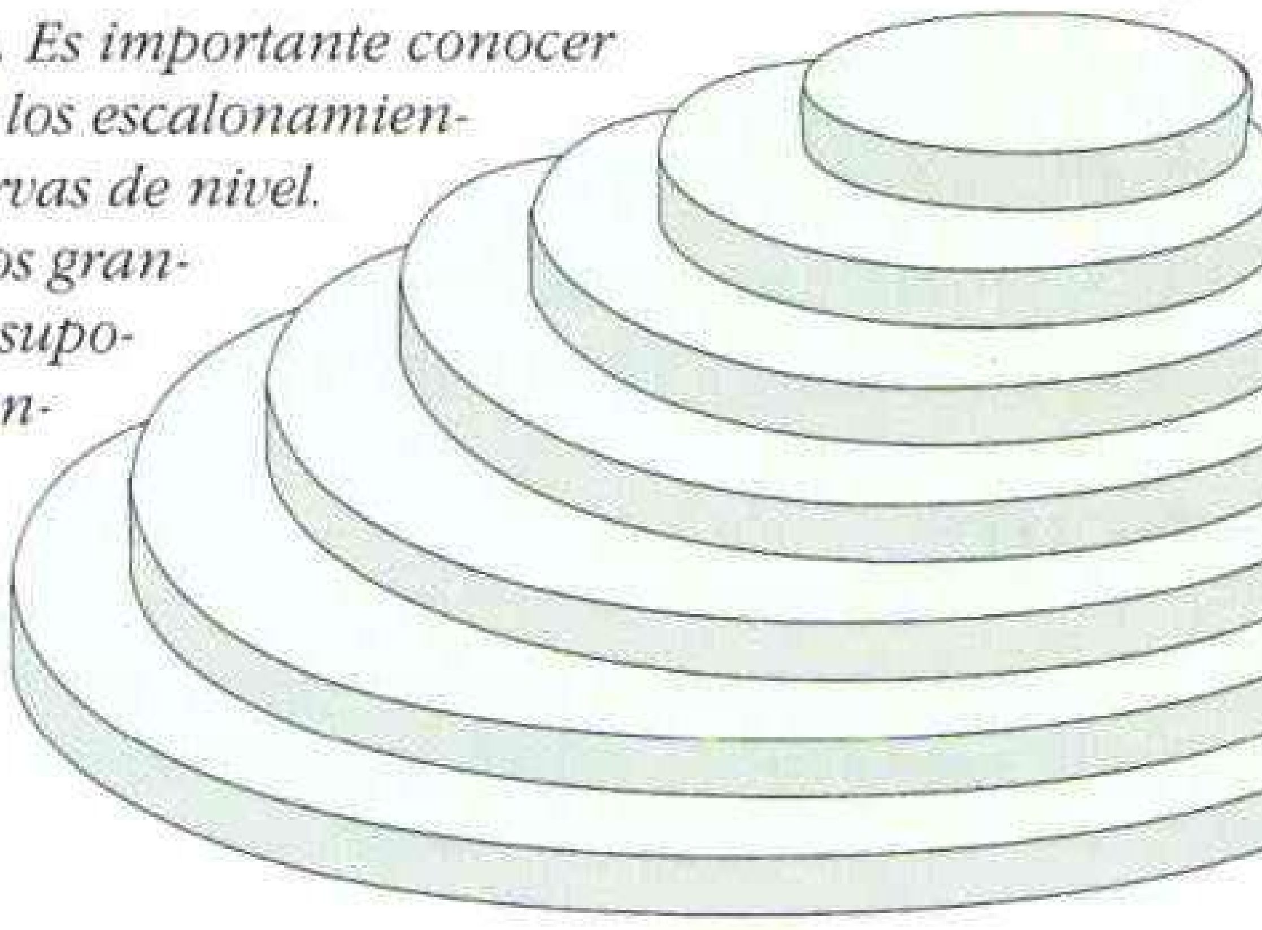

Acantilado. *Es importante conocer la altura de los escalonamientos entre curvas de nivel. Los intervalos grandes pueden suponer la existencia de un acantilado.*

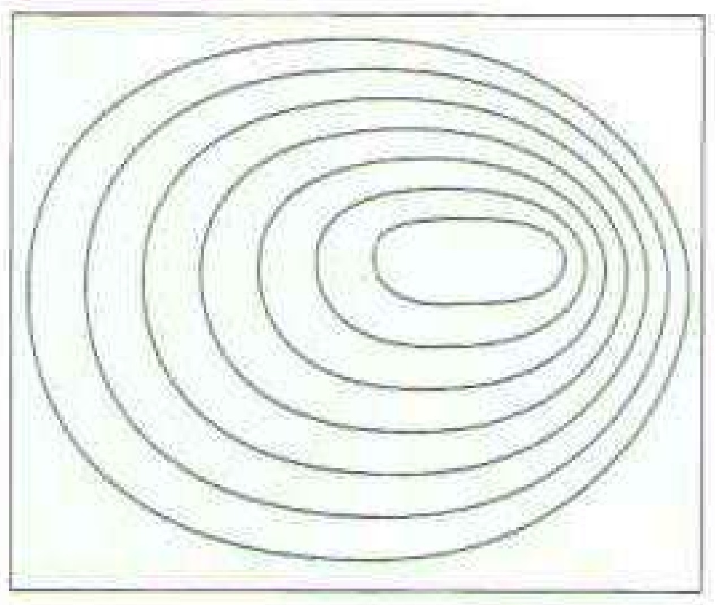

Pendiente pronunciada. *Generalmente existirá una pendiente gradual entre curvas de nivel. Unas líneas muy contiguas indican que la pendiente es muy pronunciada.*

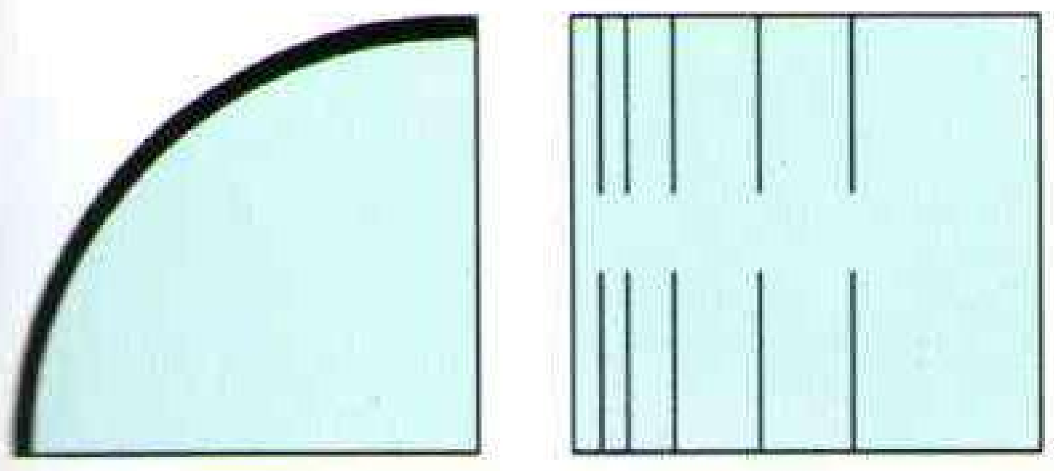

Pendiente convexa. *No es posible vislumbrar la cima de una pendiente convexa si se encuentra a su pie. Las curvas de nivel se acercan mucho en la base de la pendiente y se alejan en la cima.*

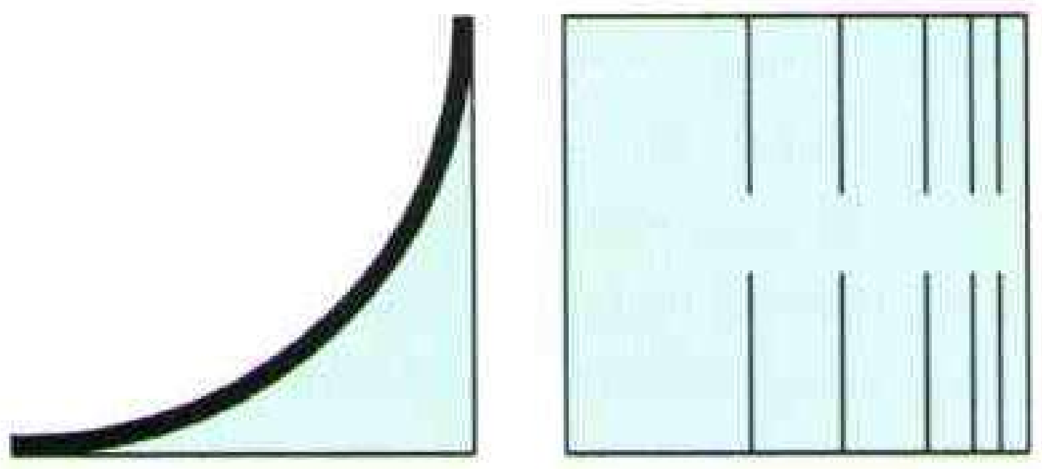

Pendiente cóncava. *Puede ver la cima de una pendiente cóncava desde su base, ya que presenta un declive suave en la parte inferior. En un mapa, sus curvas de nivel se juntan en la cima.*

REFERENCIAS EN CUADRÍCULA

Una cuadrícula le permite encontrar con precisión una posición en el mapa. Las líneas verticales se llaman *meridianos*, y las horizontales *paralelos*. Al dar una posición siempre se hace referencia primero a los meridianos. También pueden darse letras como referencia para una sección del mapa. Al determinar una referencia en la cuadrícula, siga la línea vertical más cercana a la posición en el pie del mapa para encontrar su coordenada —por ejemplo, 04. Estime el número de décimas desde la línea de cuadrícula hasta la posición —en este caso, 5. Repita la operación con la línea horizontal más cercana a la posición (410). Añada las letras de referencia, si es necesario.

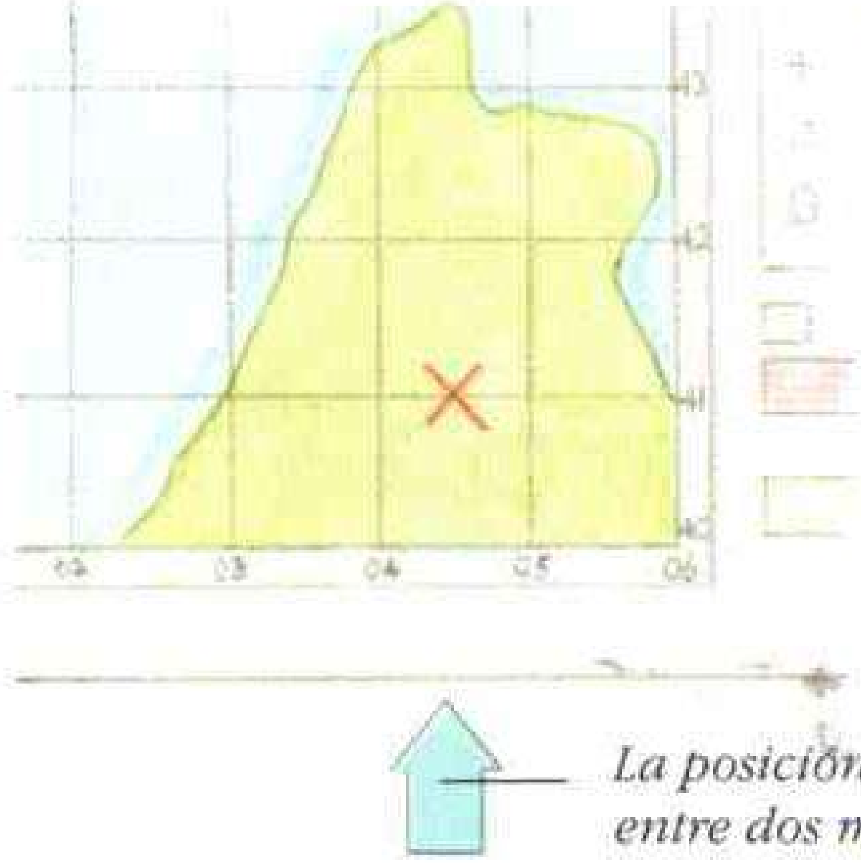

La posición se encuentra exactamente sobre un paralelo

La posición se encuentra entre dos meridianos

La posición exacta. *El método antes descrito proporciona una referencia de seis cifras sobre una cuadrícula (045410). Si su mapa tiene letras puede insertar las que sean relevantes para su posición anteponiéndolas a los números de referencia.*

PROYECCIONES

Un mapa intenta mostrar, con un grado elevado de exactitud, los detalles superficiales de la corteza terrestre, que es una esfera irregular. La proyección de la misma sobre una superficie cilíndrica o cónica crea muchas imprecisiones y distorsiones, que deben ajustarse al imprimir un mapa. Debido a este fenómeno, la mayoría de los mapas no son del todo precisos, particularmente en las latitudes mayores. La mejor proyección conocida es la de Mercator, empleada universalmente en el mar.

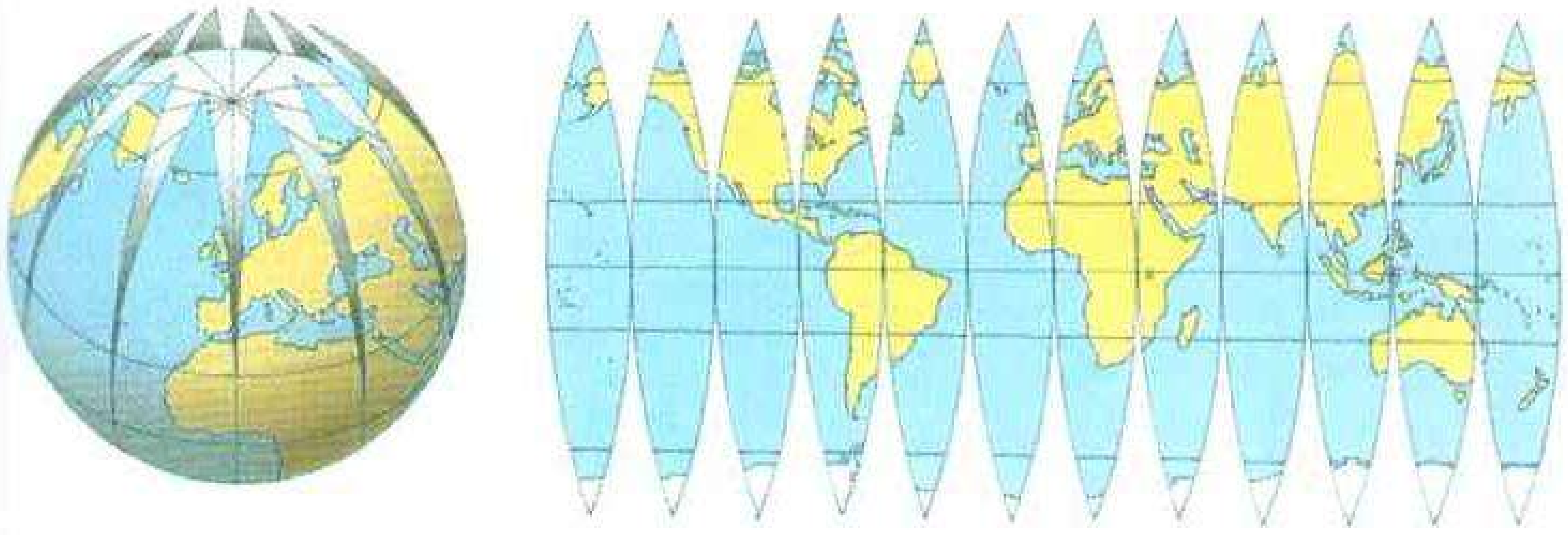

«Pelar» la Tierra. *Si «pelamos» la Tierra como una naranja, su piel quedará plana, pero habrá grandes huecos entre los segmentos en las latitudes más altas. Si se rellenan para dibujar un mapa plano, se introducen grandes distorsiones.*

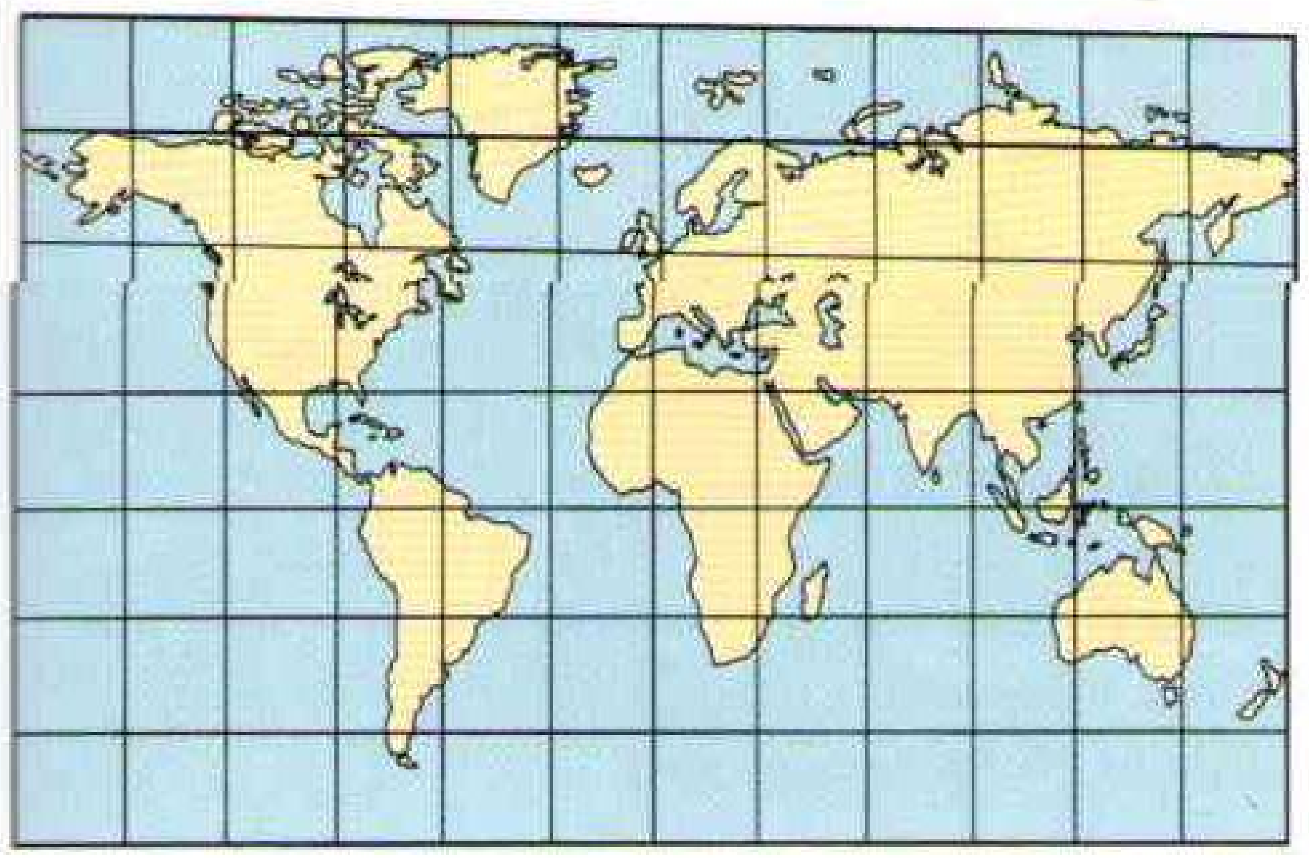

Proyección cilíndrica. *La proyección de Mercator muestra el mundo entero como si se hubiera proyectado sobre un cilindro de papel. Al desenrollar el cilindro, el mundo puede verse sobre una superficie plana. Aunque es rectangular por convención, y los meridianos y paralelos se cruzan en ángulos rectos, al igual que lo hacen en el globo, la Proyección de Mercator hace que los países de las latitudes polares aparezcan mucho mayores* (véase superior).

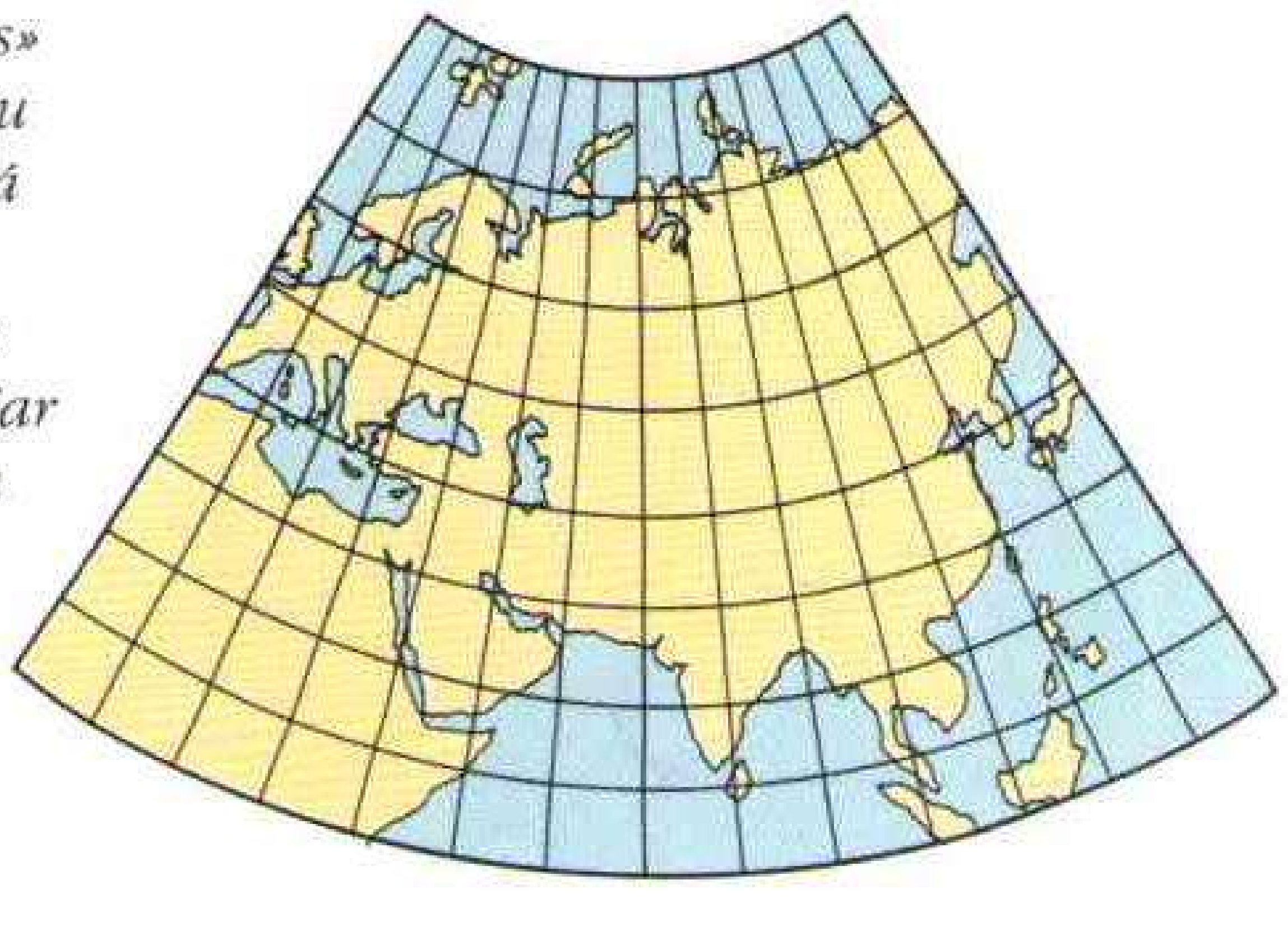

Proyección cónica. *Existen muchos tipos de proyecciones, cada una con propiedades diferentes. El mapa de la ilustración se forma proyectando el mundo sobre un cono y desenrollándolo. La proyección cónica se utiliza a menudo en los atlas de mapas regionales. Aunque existen todavía distorsiones, particularmente en la base del cono, éstas no son tan importantes como en otras proyecciones.*

ENCONTRAR SU POSICIÓN

ES IMPOSIBLE ESTAR totalmente perdido. Incluso las víctimas de un accidente aéreo conocen el país o la zona en que han caído. Averiguar el lugar exacto en que se halla es un proceso que consiste en ir descartando opciones hasta poder determinar un punto en un mapa. El ser capaz de dar una referencia precisa sobre un mapa *(véase página 127)* por radio, le asegura un rescate inmediato. Si conoce su posición sobre un mapa particular, es difícil que esté perdido. Determinando las características del terreno y encontrando alguna referencia geográfica prominente y relacionándola con el mapa, el proceso de descartar opciones no le tomará mucho tiempo. Es posible que tenga que dibujar su propio mapa por varias razones: puede que no disponga de un mapa, o que el que tenga no sea preciso. Si dibuja su propio mapa, se asegurará de poder volver a encontrar su campamento, el agua y las fuentes de comida.

DETERMINACIÓN DE LA POSICIÓN

Para determinar su situación utilizando un mapa y una brújula, debe relacionar su posición con las características del terreno que pueda ver e identificar en el mapa. Un rumbo le dará la dirección desde alguna referencia hasta su punto de observación.

Si toma estos rumbos desde una o más referencias, podrá delimitar su posición de forma bastante exacta.

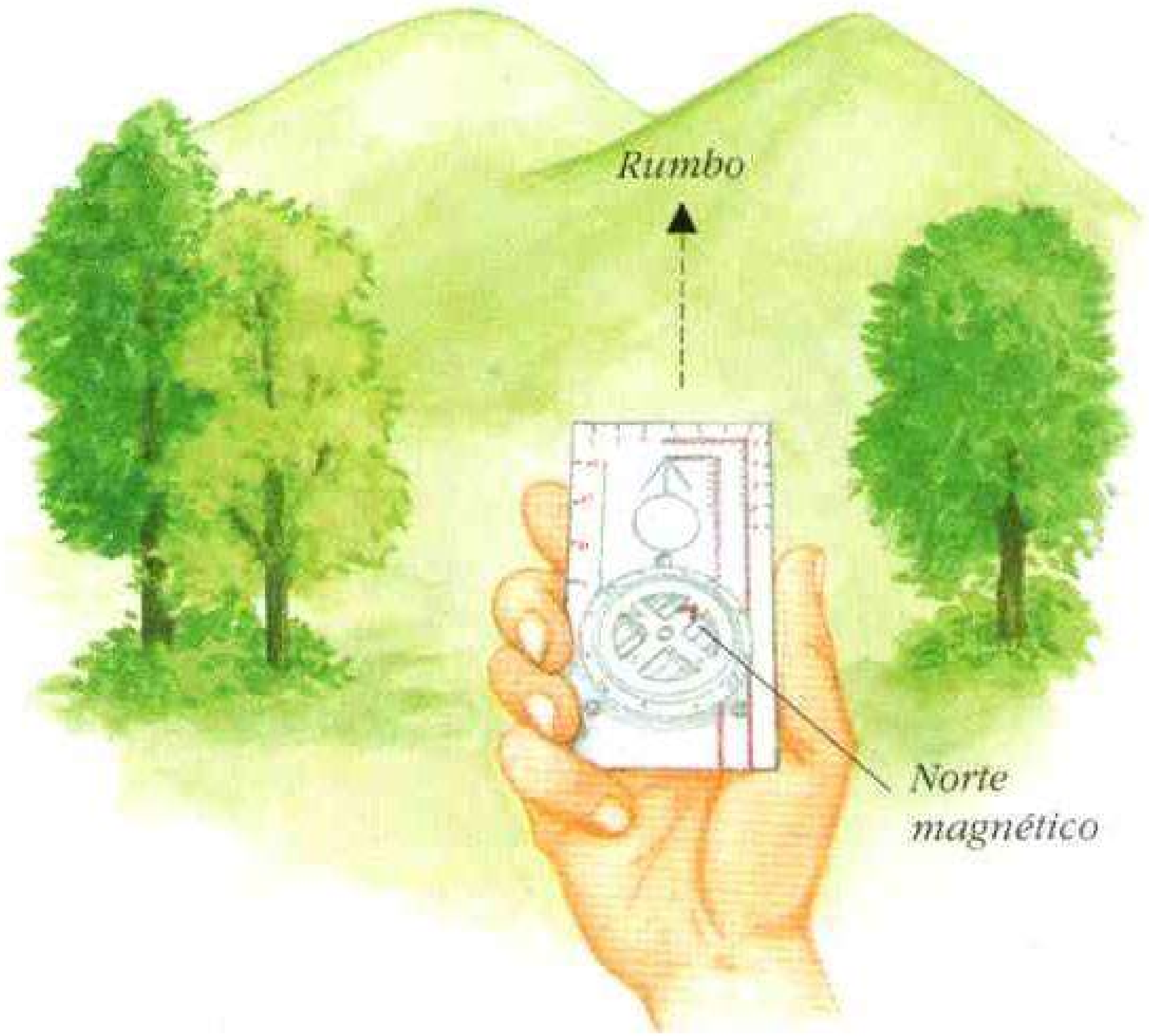

1 Apunte la brújula hacia una referencia destacable a su lado derecho, por ejemplo un paso entre dos montañas. Gire el dial (limbo) de su brújula hasta que la flecha norte del dial apunte en la misma dirección que la aguja magnética.

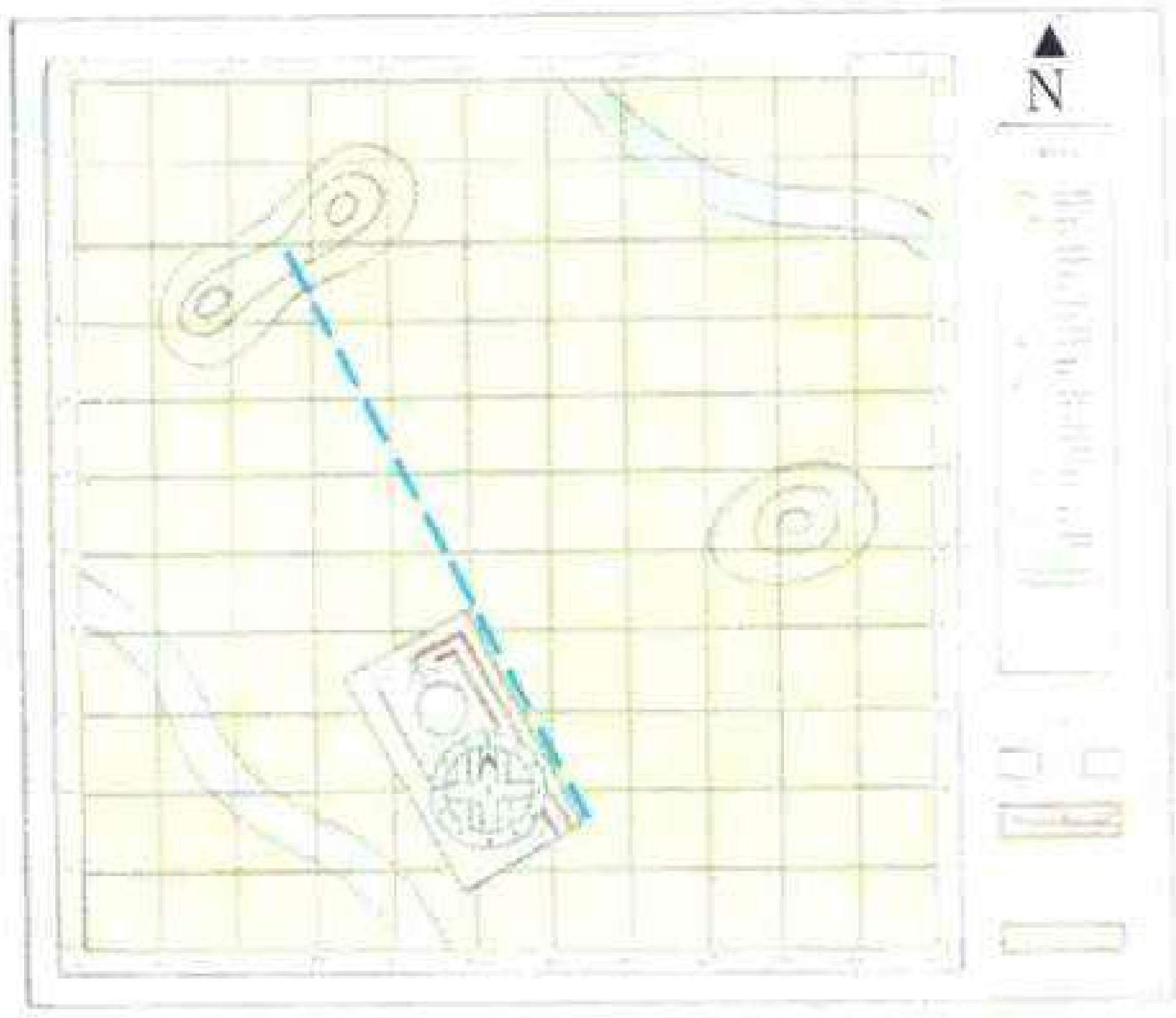

2 Coloque la brújula sobre el símbolo del collado dibujado en el mapa. Gire la brújula hasta que la flecha del dial apunte hacia el norte en el mapa. Dibuje una línea desde el collado a lo largo del rumbo (rumbo inverso).

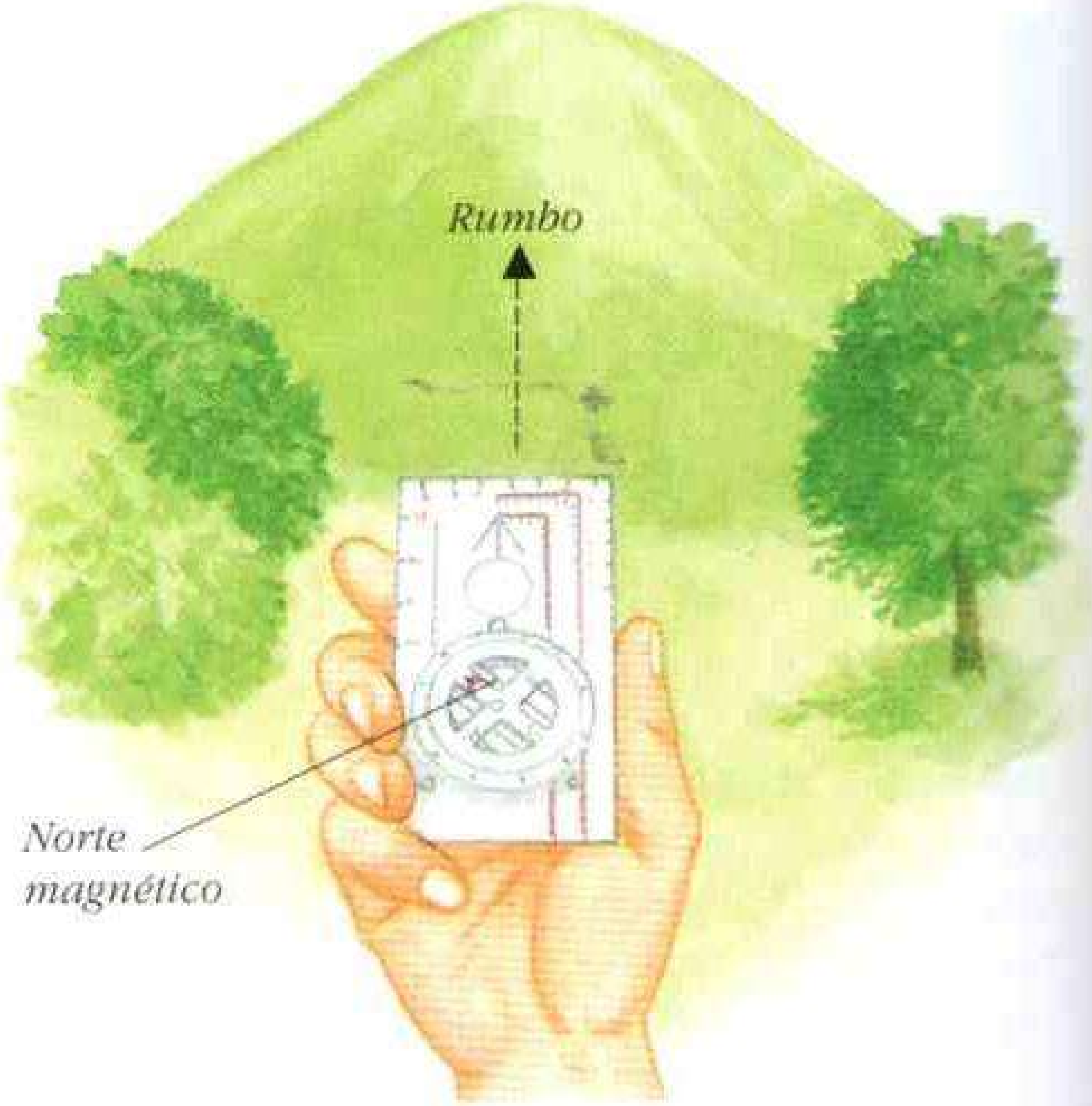

3 Apunte la brújula hacia otra referencia que pueda ver con claridad. Esta segunda referencia debería estar situada en un ángulo de unos 90 grados de la primera, y por lo menos a 1 km de su posición actual. Lo ideal es la cima de una montaña bien definida.

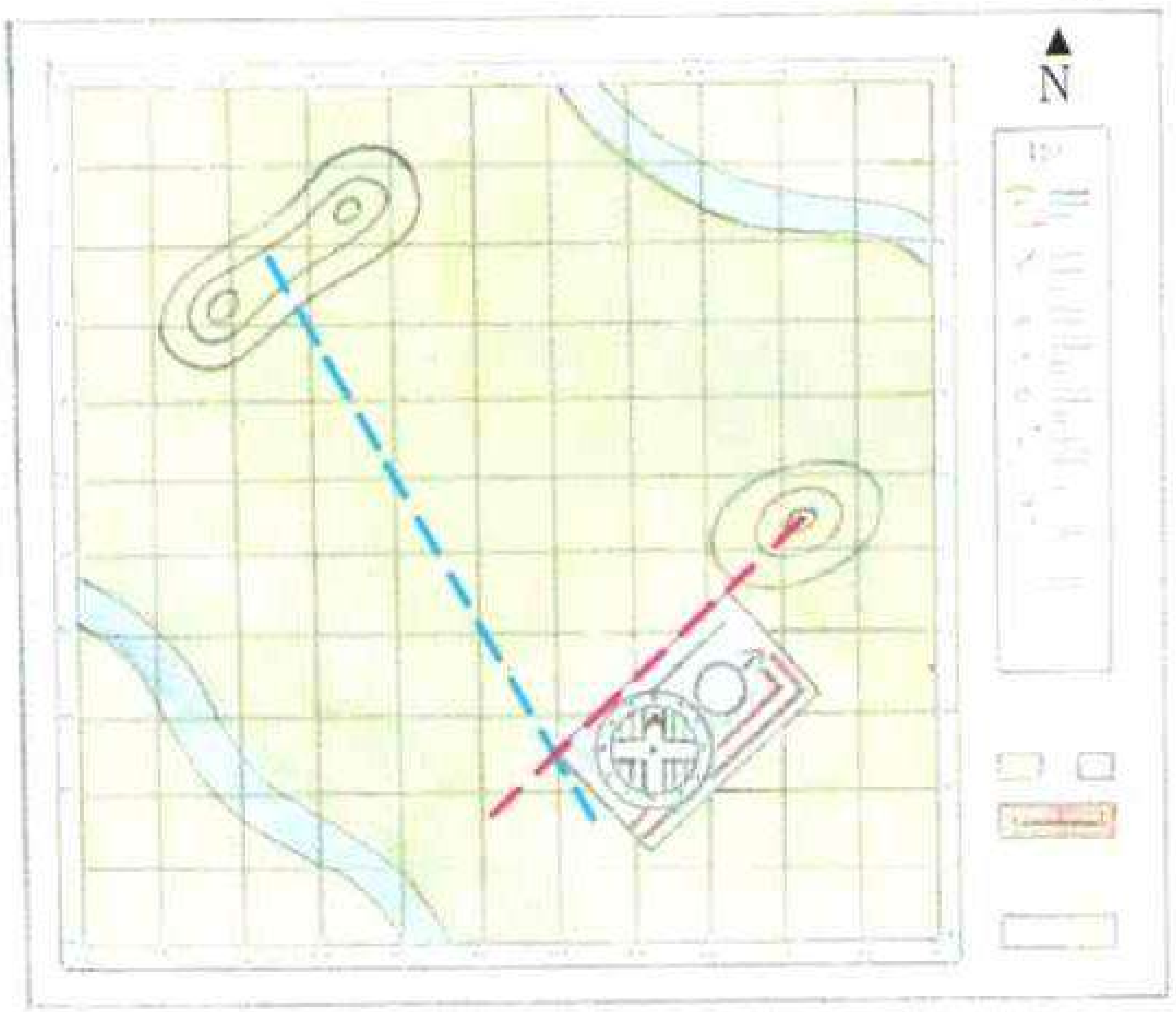

4 Trace un rumbo inverso desde la segunda referencia sobre el mapa, siguiendo la metodología antes indicada. Su posición se encuentra en la intersección de las líneas. Si así lo desea, puede tomar un tercer rumbo para obtener mayor precisión.

EL EMPLEO DE UN GPS

El GPS (siglas en inglés correspondientes a *Global Positioning System* = Sistema de Situación Global) ha revolucionado la navegación. Este sistema utiliza una constelación de 24 satélites, que transmiten señales de radio hacia la Tierra. Empleando un receptor, puede interceptar estas señales para determinar su posición de forma relativamente exacta. Los transmisores fijos calculan la posición de los satélites en órbita. La navegación en un desierto sin apenas referencias geográficas, o en el mar, se hacía antes a base de determinar rumbos con la brújula y una confirmación de la posición con un sextante y un cronómetro. Los métodos tradicionales son un refuerzo al GPS.

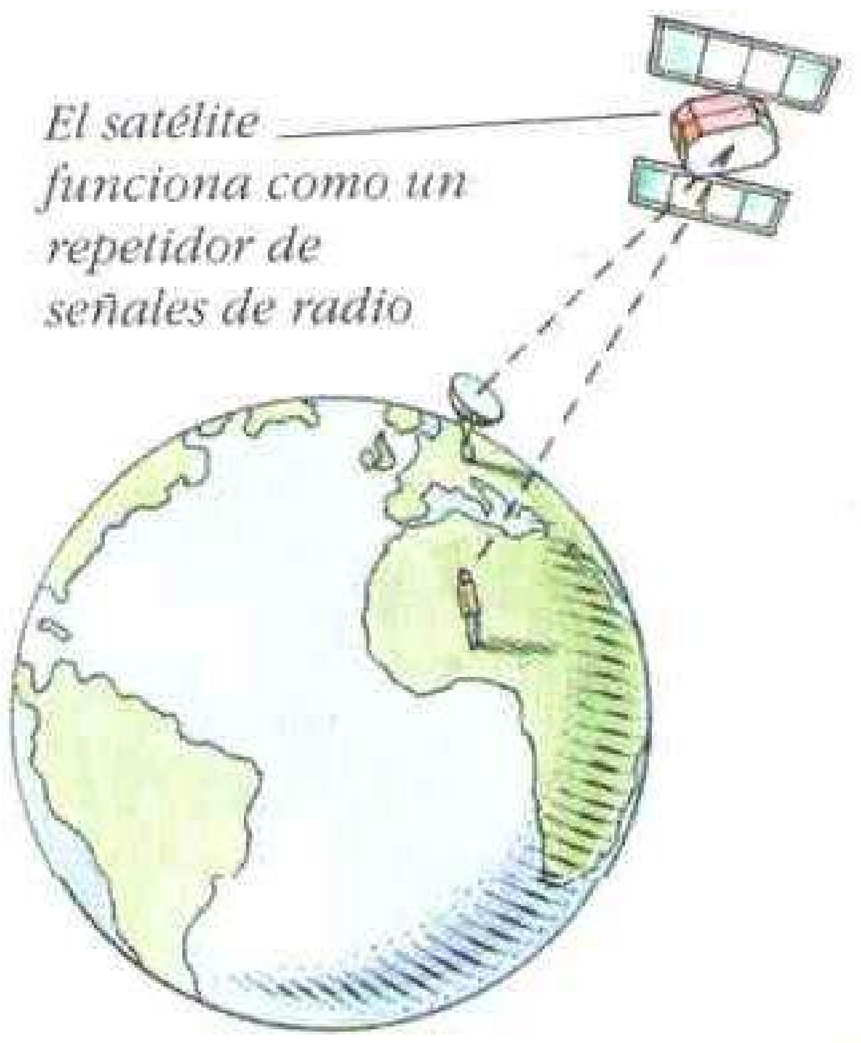
El satélite funciona como un repetidor de señales de radio

Enlace por satélite. *Las terminales en tierra emiten y reciben señales de los satélites en órbita. Estas señales pueden ser recogidas por unos receptores especiales que son de gran ayuda en la navegación.*

CÓMO DIBUJAR SU PROPIO MAPA

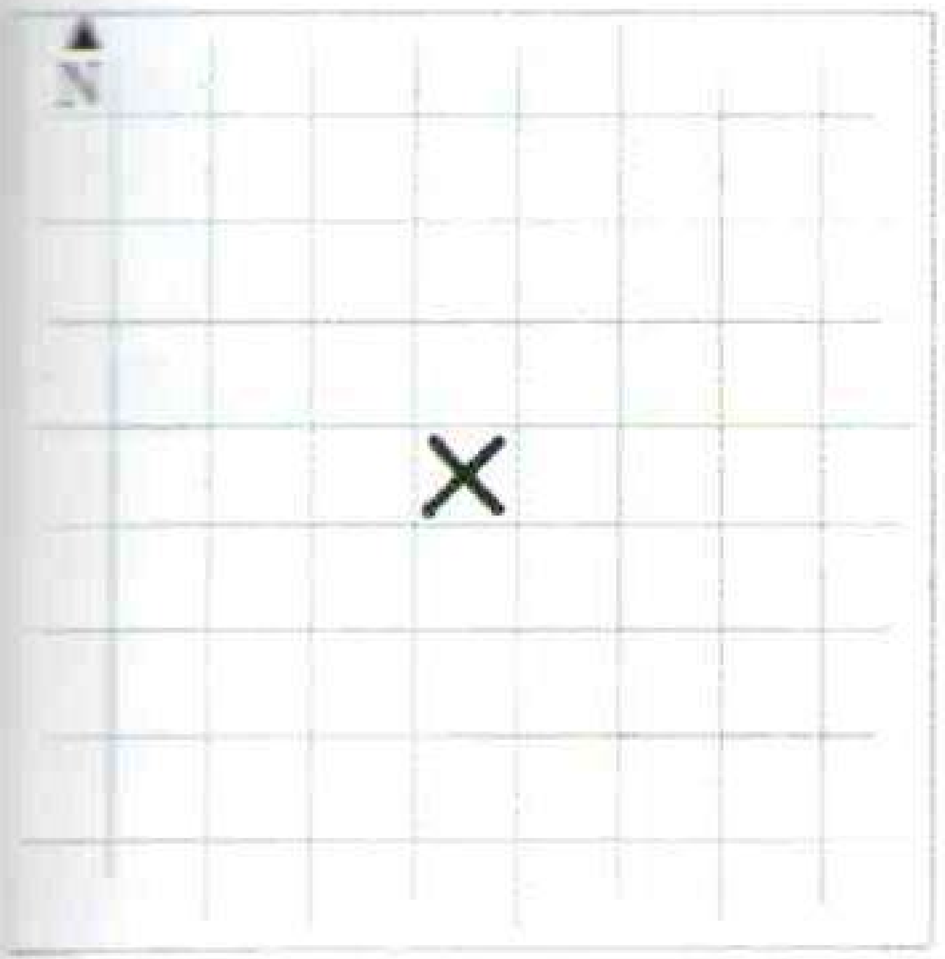

1 Dibuje una cuadrícula, cada cuadrado representará 1 km². Trace una X en el centro que representará su posición actual.

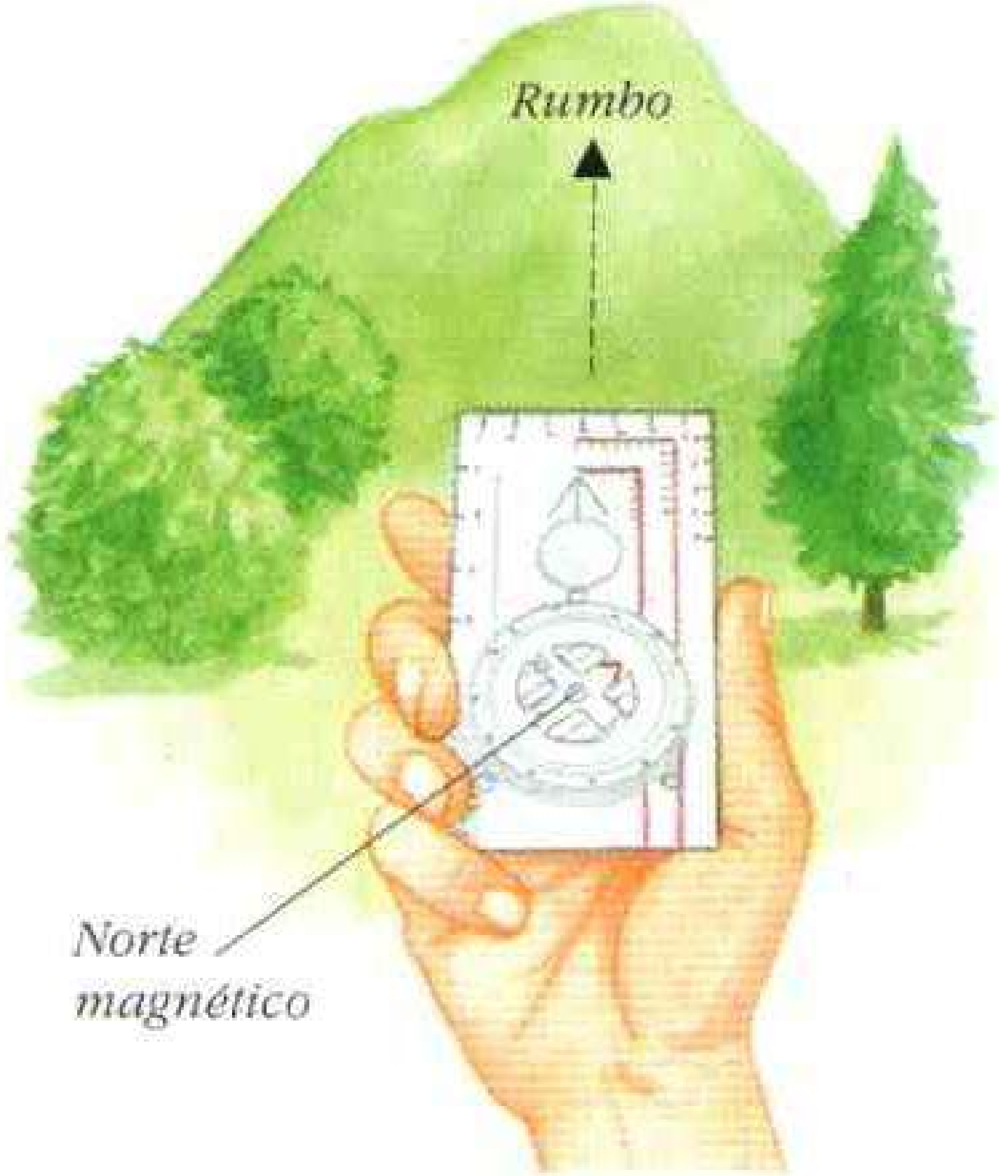

2 Desplácese hasta una posición ventajosa. Tome un rumbo hacia una referencia del terreno y estime la distancia a la misma *(véase inferior).*

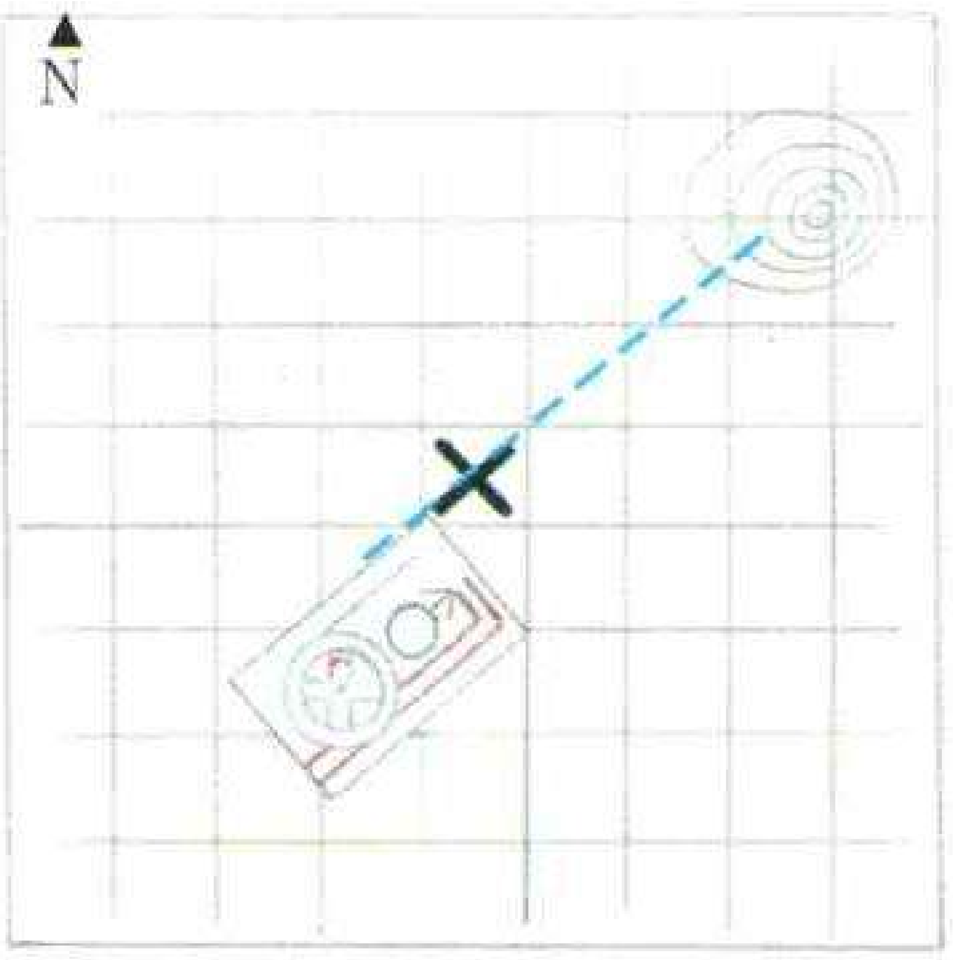

3 Señale el rumbo en su mapa, alineando la cuadrícula en dirección norte-sur. Dibuje la montaña a la distancia estimada.

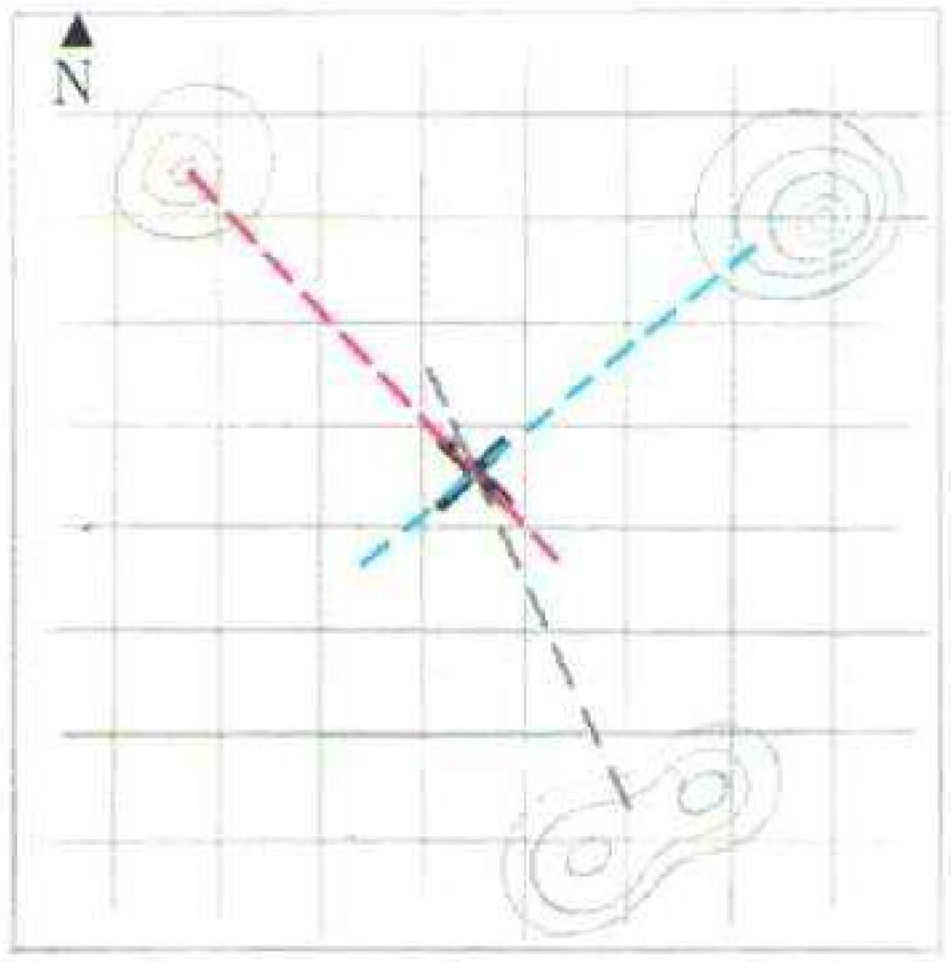

4 Tome rumbos a otras dos referencias del terreno y dibújelas en su mapa. El punto en el cual se cruzan es su posición actual.

LA MEDICIÓN DE LA DISTANCIA

Para dibujar una referencia del terreno sobre su mapa, debe, en primer lugar, determinar la distancia a la que se encuentra. Puede emplear rumbos inversos para hacerlo, pero para ello debe moverse de «X» a otra posición, que puede marcarse con toda precisión en el mapa. El rumbo de su nueva posición a la referencia le dará una intersección con el primer rumbo —que podrá marcarse entonces en el mapa. Su distancia puede calcularse con bastante precisión.

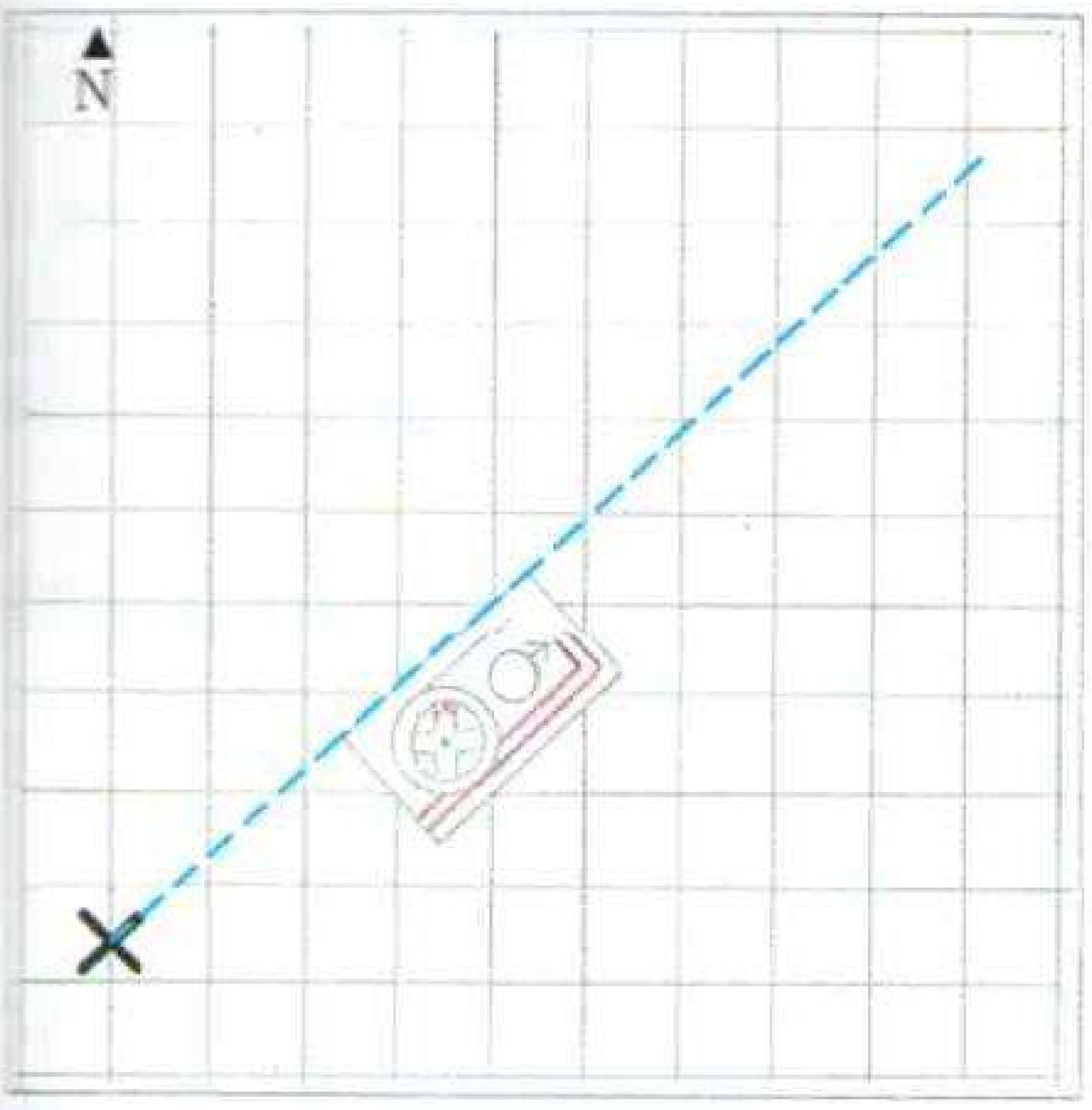

1 Puede medir la distancia de su posición a una referencia, tanto sobre un mapa impreso como en uno confeccionado por usted. Si ha dibujado una cuadrícula, los cuadrados deben ser exactos. En primer lugar, tome un rumbo a una referencia, por ejemplo una montaña. Dibuje el rumbo como una línea en la cuadrícula desde la cruz (su posición actual) en dirección a la referencia.

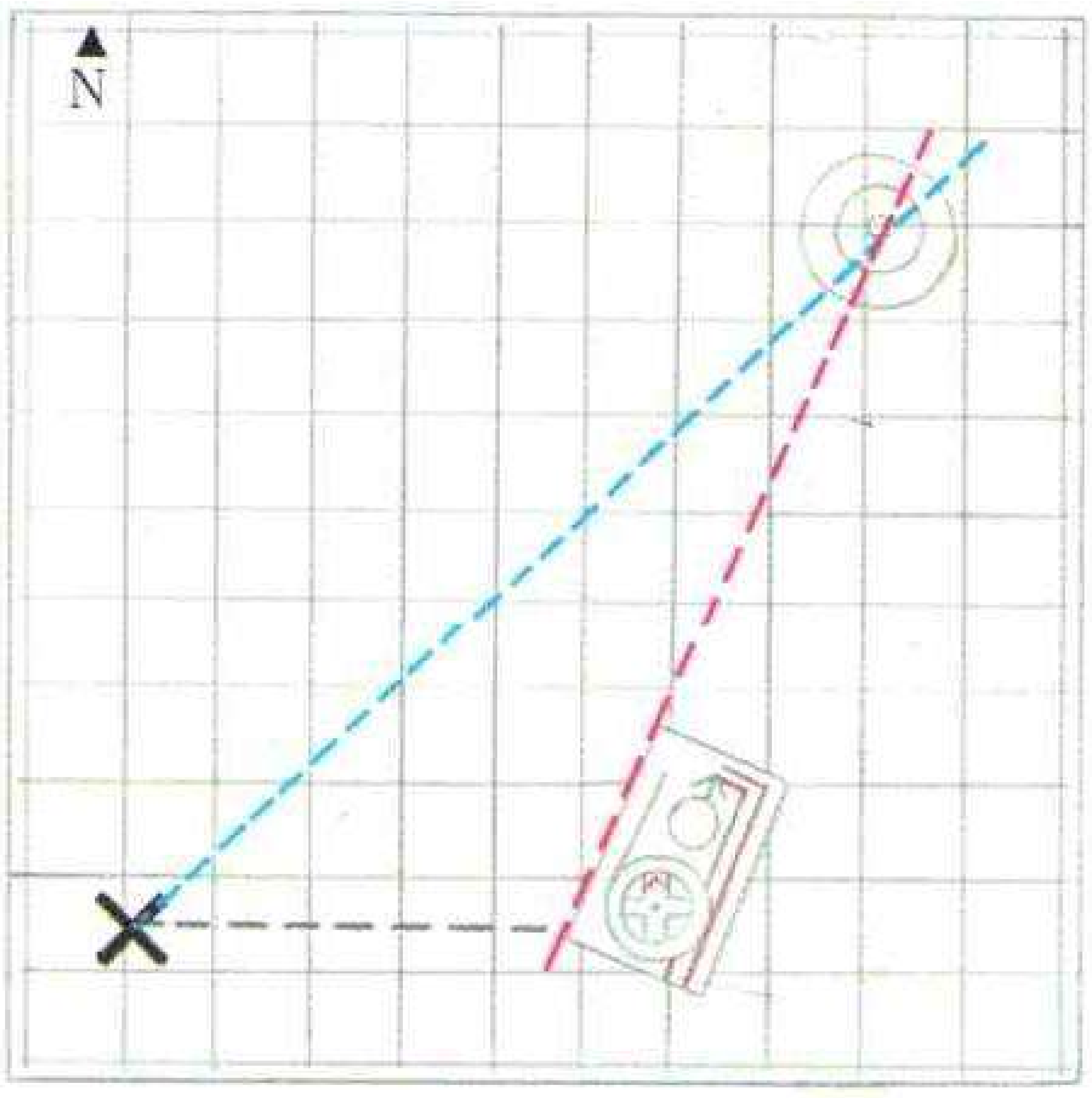

2 Camine con un rumbo de 90 grados con respecto al norte durante una distancia conocida *(véase página 131)* hasta que su siguiente rumbo a la montaña difiera en, al menos, 30 grados con respecto al primero. Señale en el gráfico su nueva posición, así como el nuevo rumbo. El punto en el que ambos se intersectan es la situación de la montaña. Puede determinar la distancia hasta la montaña contando las cuadrículas.

VARIACIÓN MAGNÉTICA

La orientación puede resultar confusa al existir tres nortes ligeramente diferentes. En la mayor parte del globo terráqueo, esta variación magnética es lo suficientemente pequeña como para que la mayoría de los caminantes no tengan que preocuparse por ella. Sin embargo, en otras zonas, como en las altas latitudes, la variación magnética se incrementa de forma significativa, llegando incluso a resultar inútil el uso de una brújula. Es, pues, vital conocer la variación magnética del área en que se navega. Ésta debe añadirse o sustraerse del rumbo, dependiendo de su localización. La variación magnética del área en cuestión se indica al pie de la mayoría de los mapas, utilizando tres flechas.

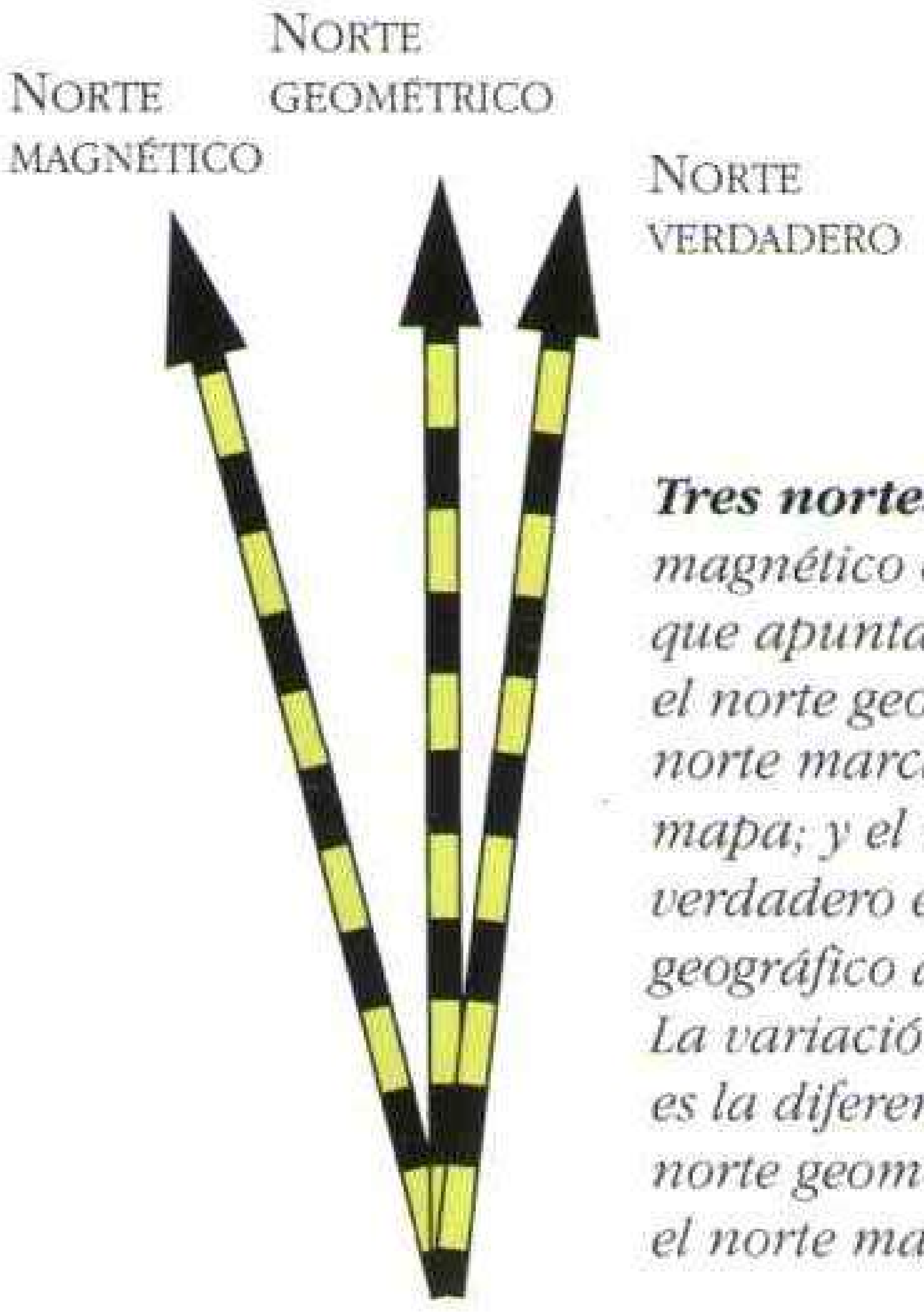

Tres nortes. *El norte magnético es aquél al que apunta la brújula; el norte geométrico es el norte marcado en el mapa; y el norte verdadero es el polo geográfico astral real. La variación magnética es la diferencia entre el norte geométrico y el norte magnético.*

ALGUNOS CONSEJOS PARA NO PERDERSE

- La regla de oro es: «Confíe en su brújula». Aun así, muchas personas se pierden al confiar antes en su propio sentido de la orientación que en su brújula.
- Mida siempre el rumbo y las distancias.
- Asegúrese de comprender cómo se indican los distintos tipos de terreno en los mapas, localizando los valles, las montañas y las crestas.
- Contraste el mapa utilizando la brújula *(véase página 130)*, busque con la vista las referencias geográficas más relevantes, e identifique su posición sobre el terreno antes de tomar un rumbo.

PRINCIPIOS DE NAVEGACIÓN

LA NAVEGACIÓN EN tierra consiste en una combinación de trabajos sobre el mapa y con la brújula, con el propósito de desplazarse por un terreno determinado de la forma más segura y sencilla. De ninguna manera es tan precisa como la navegación en el mar, donde las únicas variaciones son las mareas y el viento. La velocidad de desplazamiento se ve sobremanera afectada por el tipo de terreno, las rocas y el suelo, el clima, el estado del tiempo y la vegetación. La navegación en tierra requiere un seguimiento constante de la ruta, no sólo para evitar los peligros, sino también para no perderse.

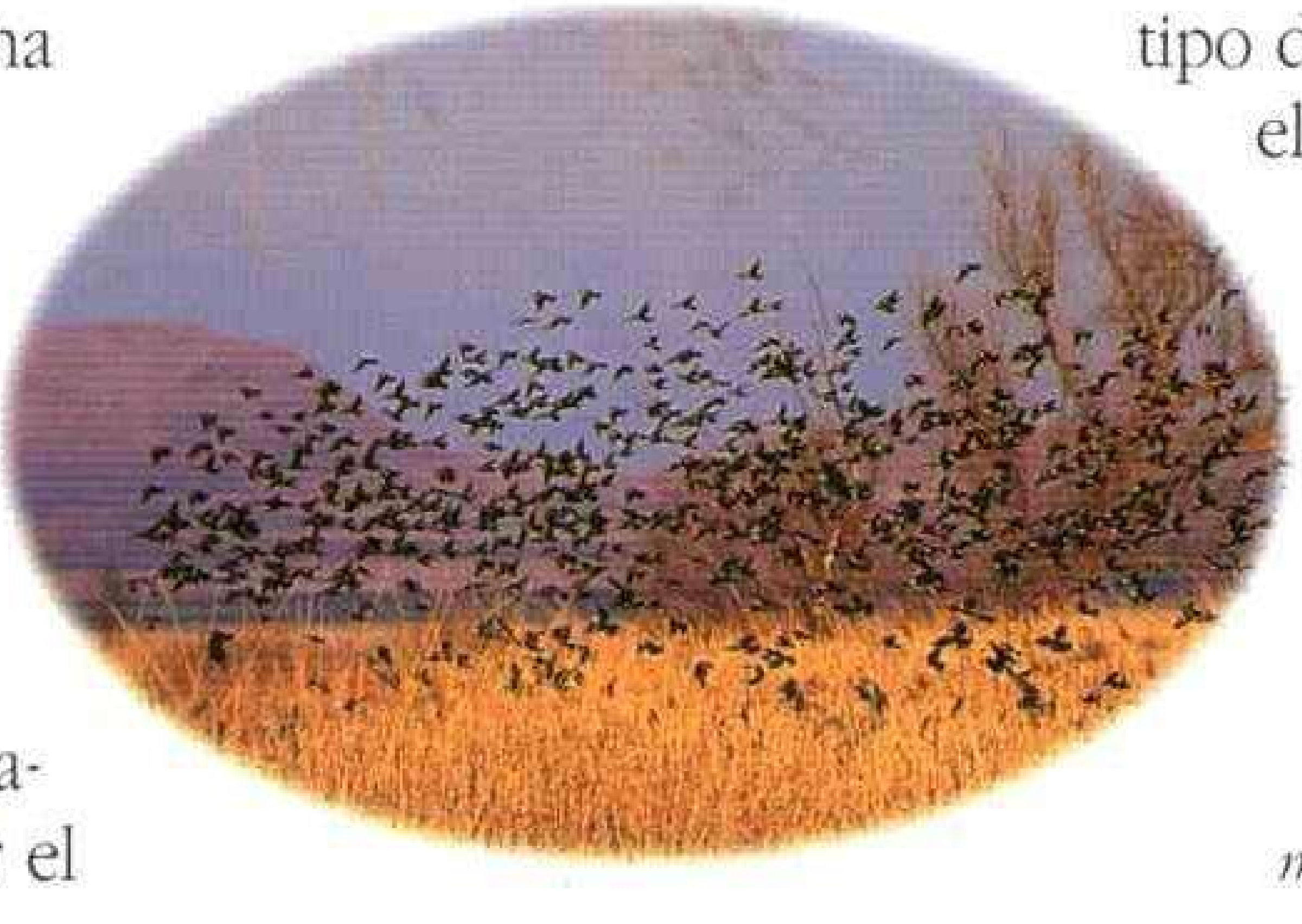

La migración. *Muchas aves han desarrollado una capacidad de navegación tal, que les permite regresar desde el otro lado del Globo a los mismos lugares de cría, un año tras otro.*

CONTRASTAR SU MAPA

Antes de emprender un recorrido, debe contrastar su mapa con la brújula, determinando dónde se encuentra su posición actual y su destino en el mapa. Además de trazar el rumbo, debe asimismo saber a qué distancia se encuentra del lugar de destino. Puede determinar dicha distancia utilizando la escala del lateral de su brújula, o bien la escala de la parte inferior de su mapa. Mientras mide esta distancia, compruebe en el mapa el tipo de terreno que espera cruzar.

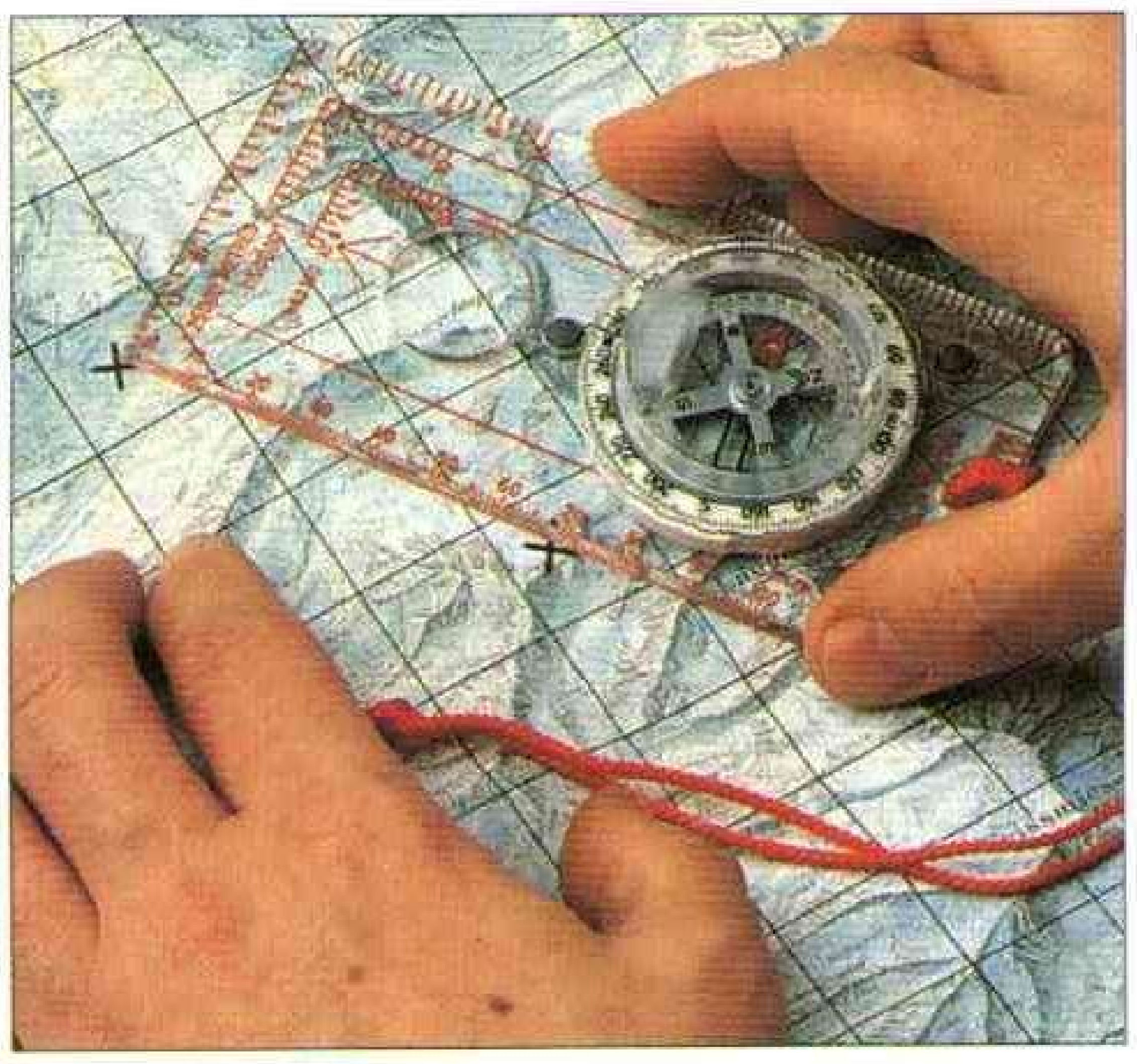

1 Para determinar el rumbo del punto A al punto B, coloque la brújula sobre el mapa entre A y B, y apunte la flecha de dirección del extremo de la brújula hacia la dirección deseada. Lea la distancia entre A y B en la escala del lateral de la brújula y compárela con la escala del mapa.

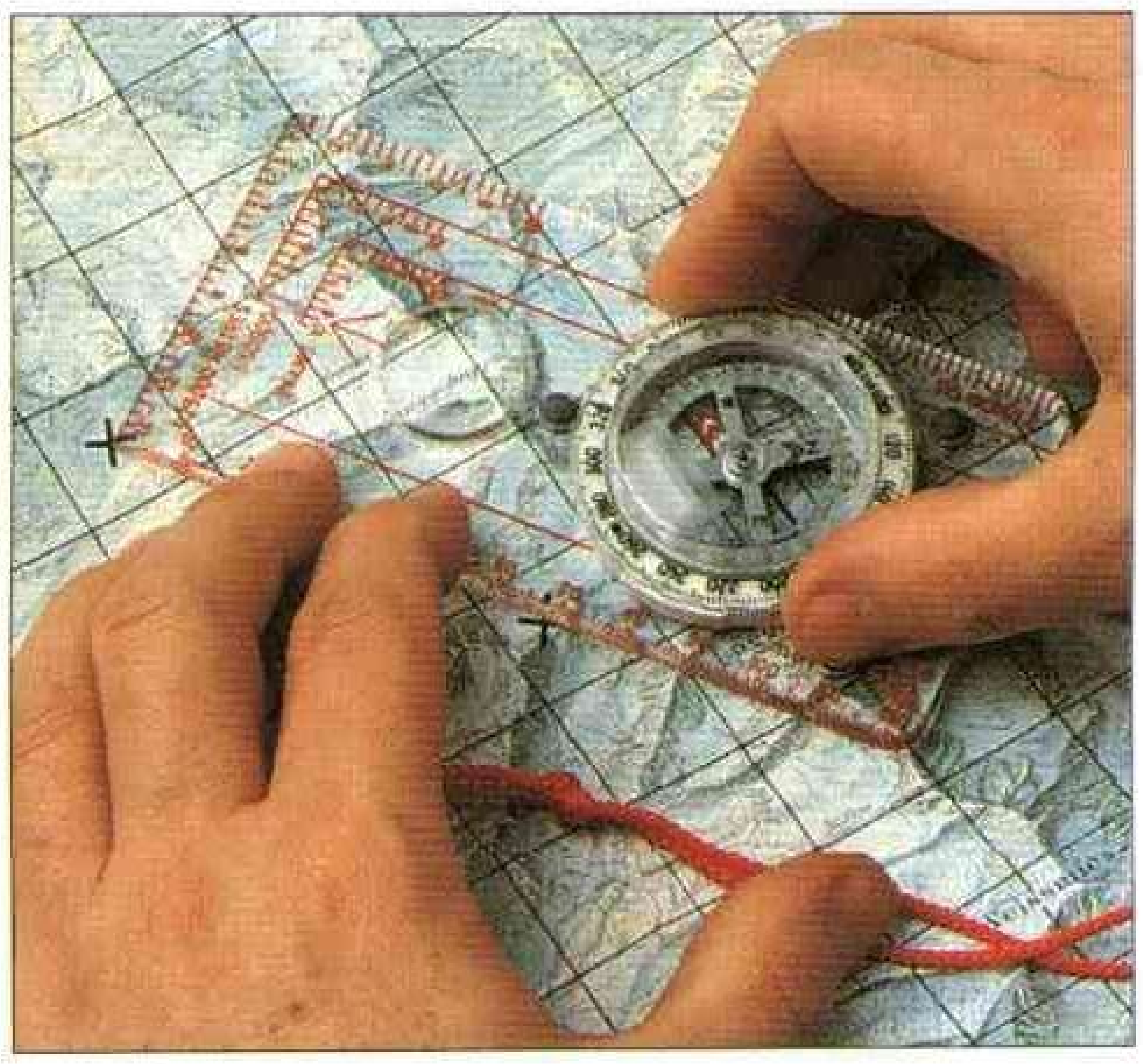

2 Sin mover la brújula, gire el dial (limbo) central hasta que las líneas paralelas norte-sur estén alineadas con las líneas de la cuadrícula del mapa. La flecha norte (roja) del dial (limbo) debería apuntar hacia el norte geométrico *(véase página 129)*. De esta manera, establecerá el rumbo (el ángulo entre la línea A-B y el norte magnético) con la brújula.

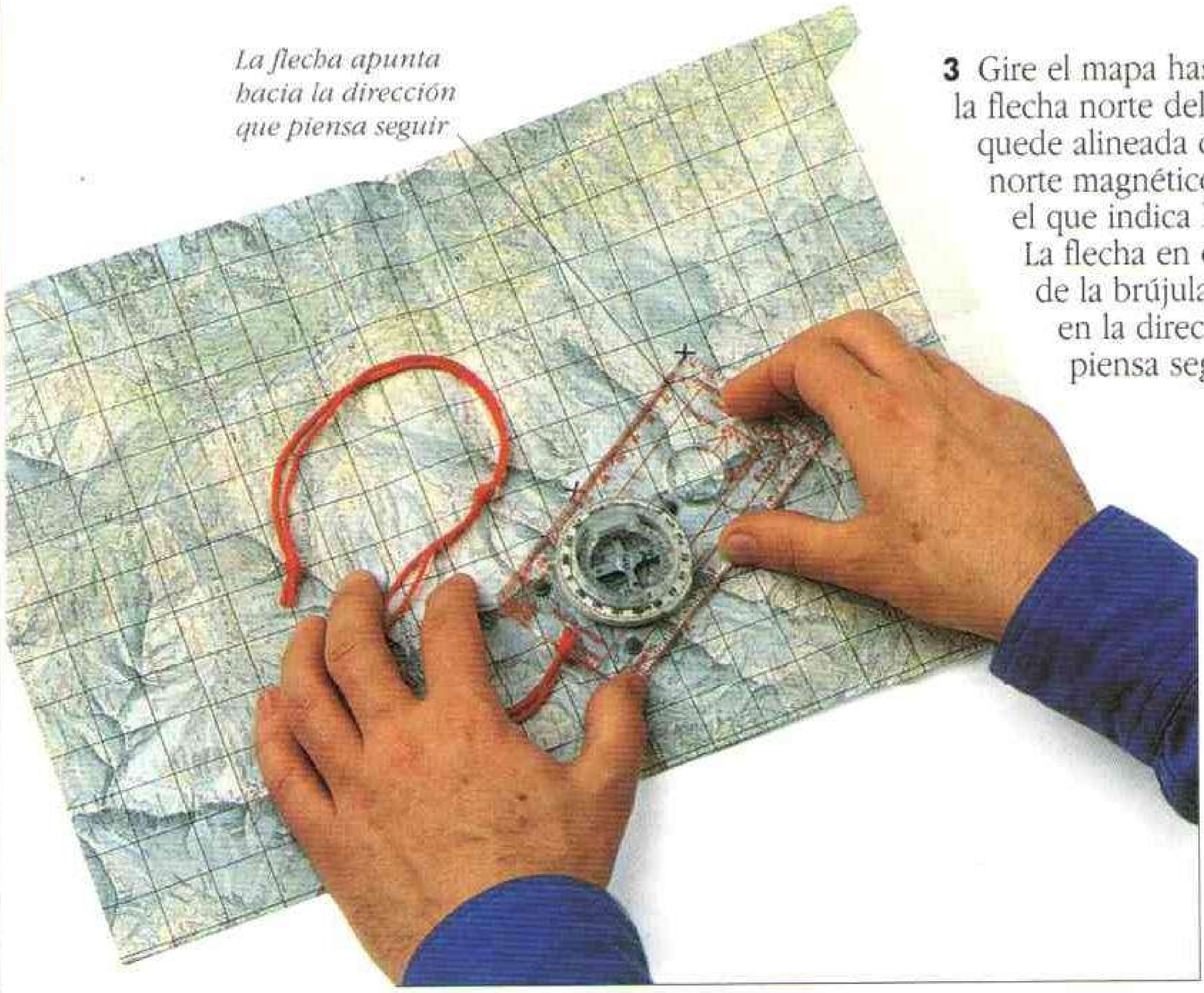

La flecha apunta hacia la dirección que piensa seguir

3 Gire el mapa hasta que la flecha norte del dial quede alineada con el norte magnético, que es el que indica la aguja. La flecha en el extremo de la brújula apuntará en la dirección que piensa seguir.

4 Ahora puede sujetar la brújula con la mano y seguir la flecha de dirección. Asegúrese de mantener alineadas la flecha norte del dial y la aguja magnética. Al seguir un rumbo, mantenga la brújula en posición horizontal.

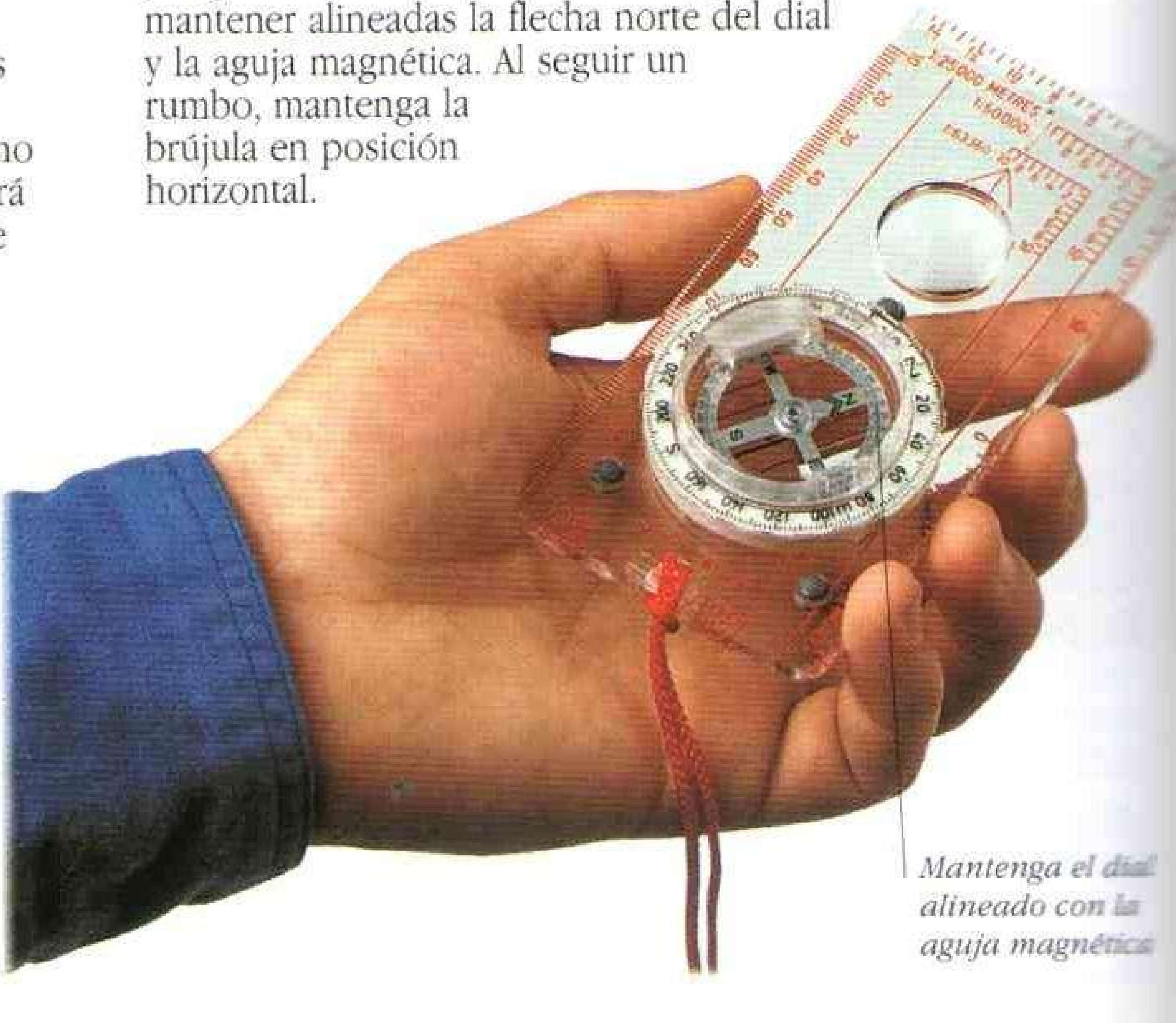

Mantenga el dial alineado con la aguja magnética

TÉCNICAS DE NAVEGACIÓN

En tierra no siempre es posible o recomendable caminar en línea recta siguiendo un rumbo. Los siguientes «trucos del oficio» le mostrarán cómo determinar el rumbo a seguir sorteando los terrenos difíciles y siguiendo la ruta más sencilla sin perder tiempo para comprobar la situación, y sin perderse. Es importante recordar que siempre debe haber establecido un rumbo en su brújula *(véase página anterior)*, aunque sea únicamente para emplearlo como referencia.

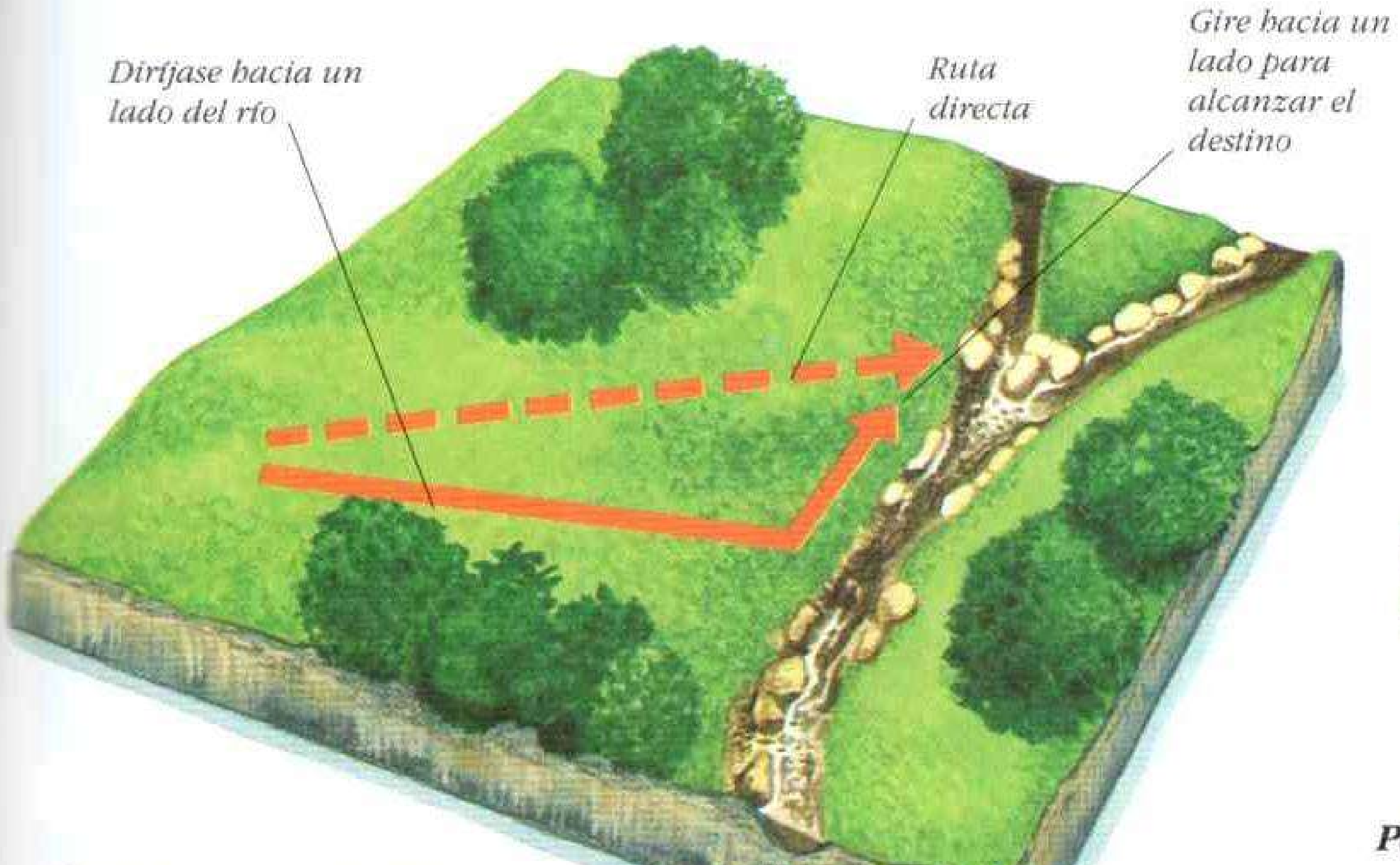
Diríjase hacia un lado del río
Ruta directa
Gire hacia un lado para alcanzar el destino

Desviación intencional. *Su brújula le llevará a unos 20 o 30 grados de su objetivo. A una distancia de 1 km podría perder esta bifurcación en el río y aparecer a 300 m de ella a cualquiera de sus lados. Más aún, no sabría en qué dirección buscarla. Para evitarlo, desvíese intencionalmente hacia uno de los lados. Cuando alcance el río, sabrá si girar hacia la derecha o a la izquierda de la bifurcación.*

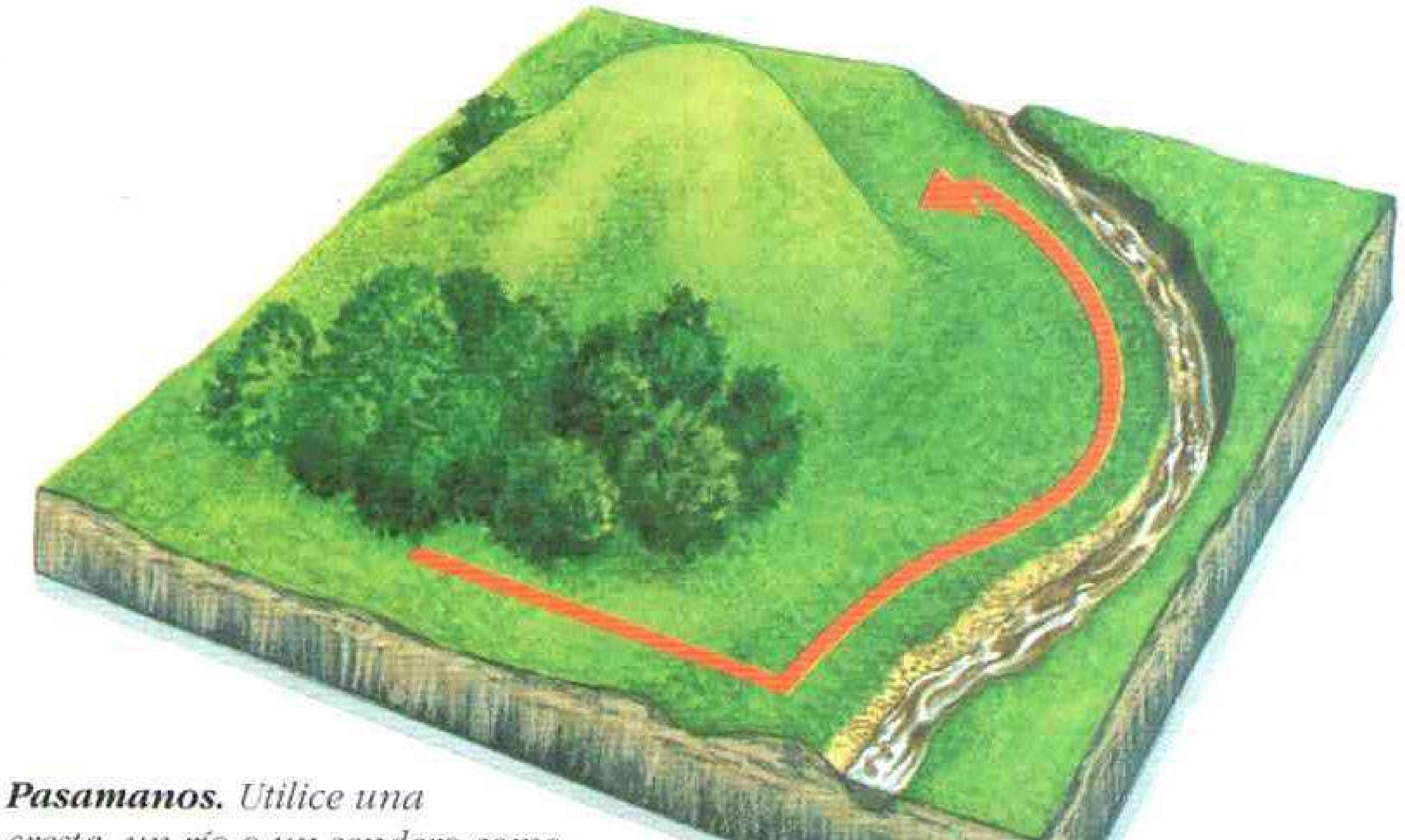

Pasamanos. *Utilice una cresta, un río o un sendero como guía o «pasamanos» a seguir en caso de no poder ver su destino. En primer lugar, siga un rumbo hasta el pasamanos (en este caso un río) y camine hasta él. En este ejemplo, la montaña se encuentra a unos 250 grados del punto más cercano al río. Siga el río hasta que aparezca la montaña (a 250 grados) y gire entonces a la izquierda para seguir ese rumbo.*

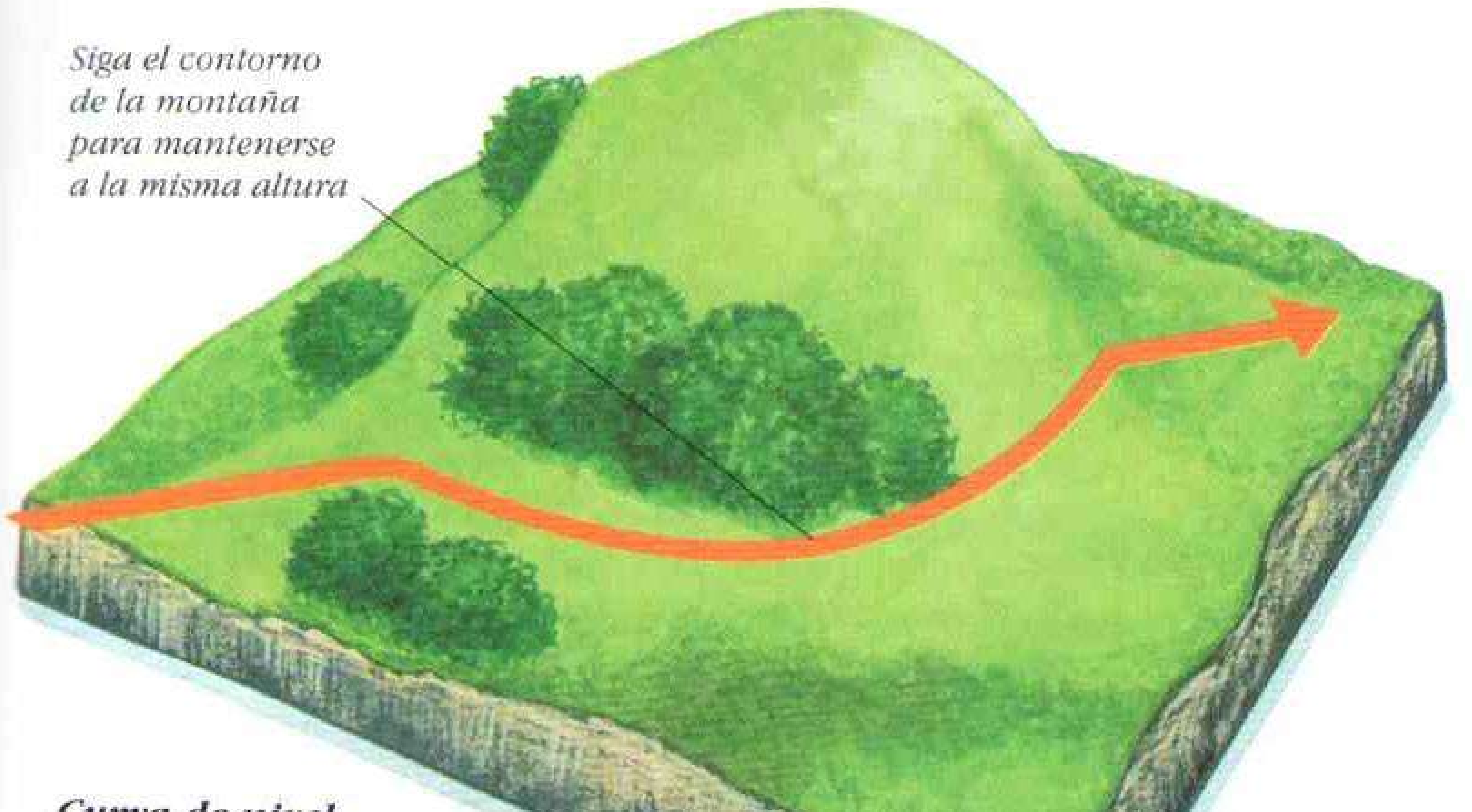
Siga el contorno de la montaña para mantenerse a la misma altura

Curva de nivel. *Al mantenerse a la misma altura, estará siguiendo una curva de nivel en el mapa, un método muy conveniente de orientación, en particular en la selva, donde puede resultar mucho más exacto que seguir un rumbo. Seguir una curva de nivel también le permitirá ahorrar energía al no perder ni ganar altura.*

La estimación de la distancia recorrida

Al moverse a pie, puede contar sus pasos (los que da con el pie izquierdo) o utilizar el reloj. La marcha varía de acuerdo con el terreno y el ritmo de viaje. El tiempo es una forma más adecuada de medir la distancia recorrida. Deténgase después de los primeros 10 minutos para determinar en el mapa la distancia que ha recorrido, y entonces, una vez que ha determinado su velocidad, puede detenerse cada hora para comprobar el mapa y el equipo. Aprenderá a determinar la velocidad de su marcha con una exactitud sorprendente. En un terreno escabroso, y con un equipo pesado, una velocidad de 4 km/h es muy buena.

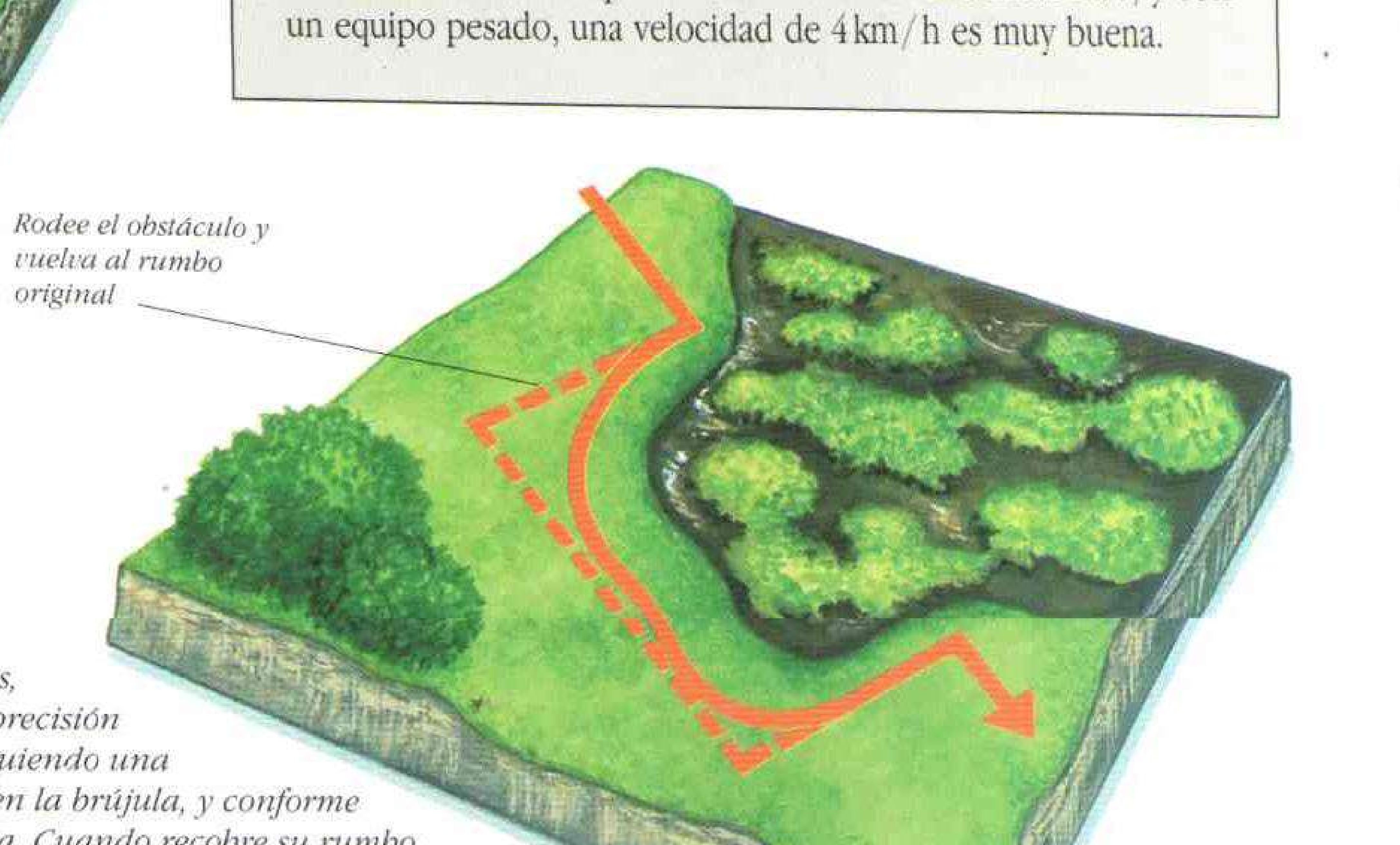
Rodee el obstáculo y vuelva al rumbo original

Rodeo / Desviación. *Los obstáculos importantes, como una ciénaga, no siempre se marcan con precisión en los mapas. Al rodearlos, al igual que hace siguiendo una curva de nivel, siga la línea recta de su rumbo en la brújula, y conforme va rodeando el obstáculo, determine la distancia. Cuando recobre su rumbo, sabrá la distancia que tendrá que recorrer para volver a la línea original.*

LA EVALUACIÓN DEL TIEMPO

EL TIEMPO ES el día-a-día del clima. Los cambios locales son causados por las diferencias en la presión atmosférica y la temperatura. Al pasar por encima de los diferentes tipos de terreno, los movimientos de aire localizados producen vientos y lluvias. Las áreas costeras tienen, por lo general, vientos que soplan hacia el interior durante el día y hacia la costa durante la noche. En algunos lugares, como en el interior de los continentes, el tiempo es relativamente estable, mientras que en otros el clima es muy variable. Los climas realmente ecuatoriales no tienen distinción entre estaciones, sino una rutina diaria de lluvia y sol.

El tiempo en el mundo. *Un tipo de tiempo o clima puede localizarse en un área específica, o bien, puede desplazarse alrededor del mundo, afectando a zonas muy alejadas entre sí —como es el caso de los huracanes.*

LEER LAS NUBES

Las nubes son grandes masas de vapor de agua en condensación. Su presencia, tipo y tamaño indican la temperatura y la presión de las masas de aire, permitiendo la predicción del tiempo. Cuanto más altas sean las nubes, mejor y más estable será el tiempo. Las nubes tormentosas son generalmente oscuras, bajas y agrupadas en grandes masas, mientras que las nubes que presagian buen tiempo son altas y blancas. Las nubes muy bajas pueden envolver las tierras altas en la niebla.

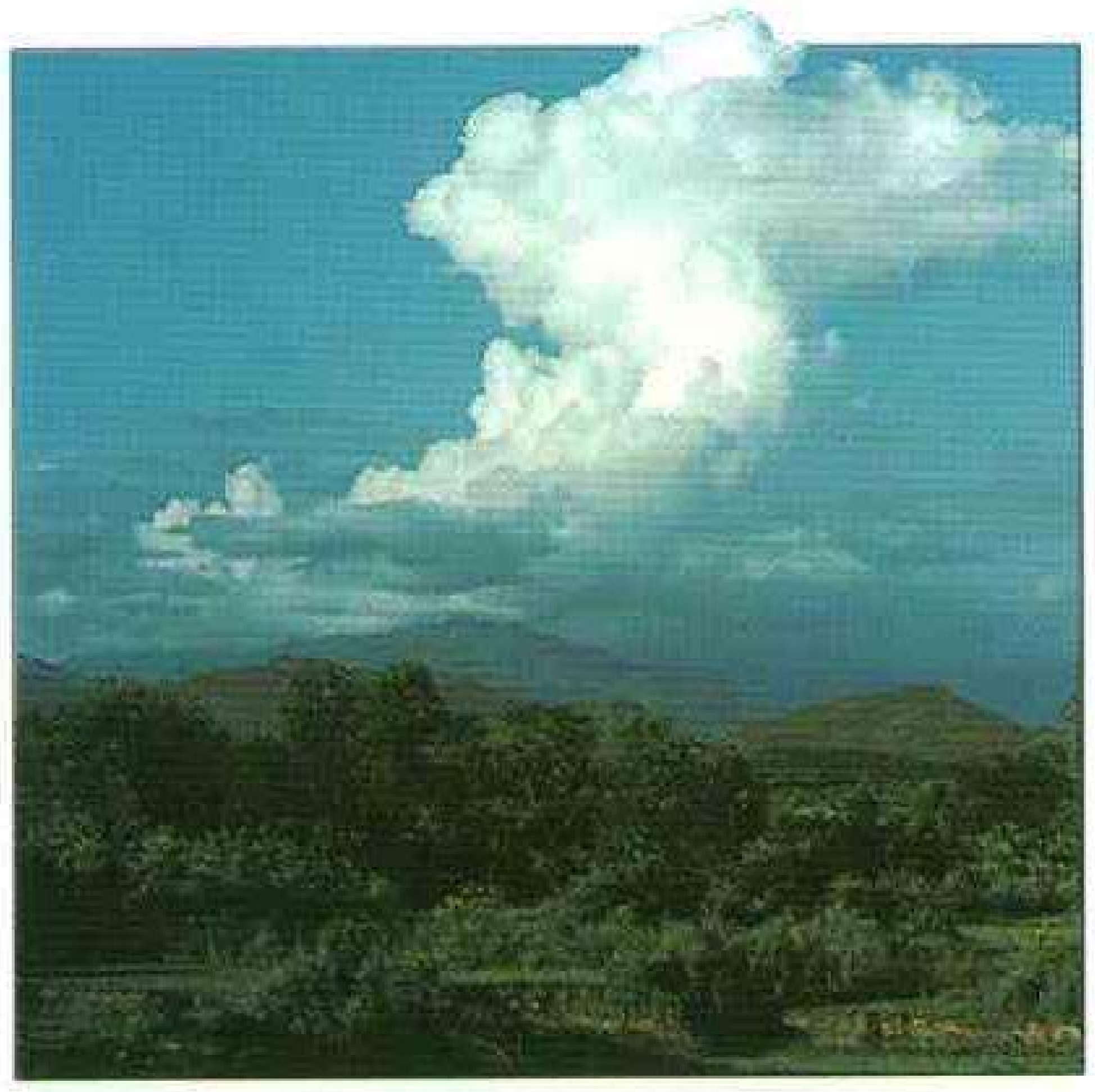

Cumulonimbos. *Los cumulonimbos traen consigo granizo, vientos y tormentas. Las formaciones de estas nubes oscuras y densas pueden alcanzar un gran tamaño y un desarrollo vertical considerable, tomando formas extrañas. En ocasiones, anuncian tormenta eléctrica* (véase página 171).

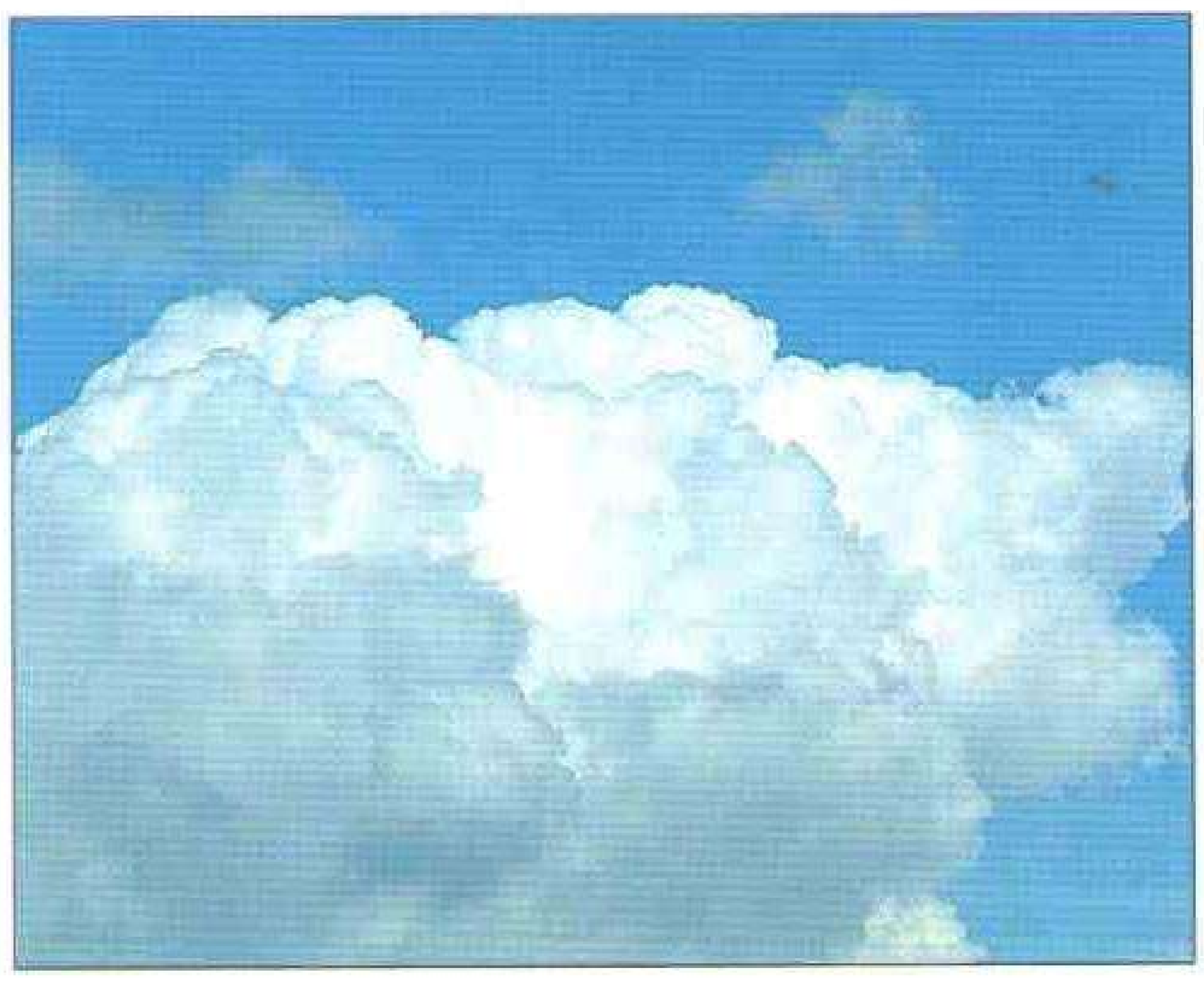

Cúmulos. *Los cúmulos de aspecto algodonoso indican buen tiempo, pero si las nubes están muy juntas y se tornan de un color oscuro, indican lluvia. Sobre mar abierto, pueden indicar la proximidad de tierra* (véase página 165).

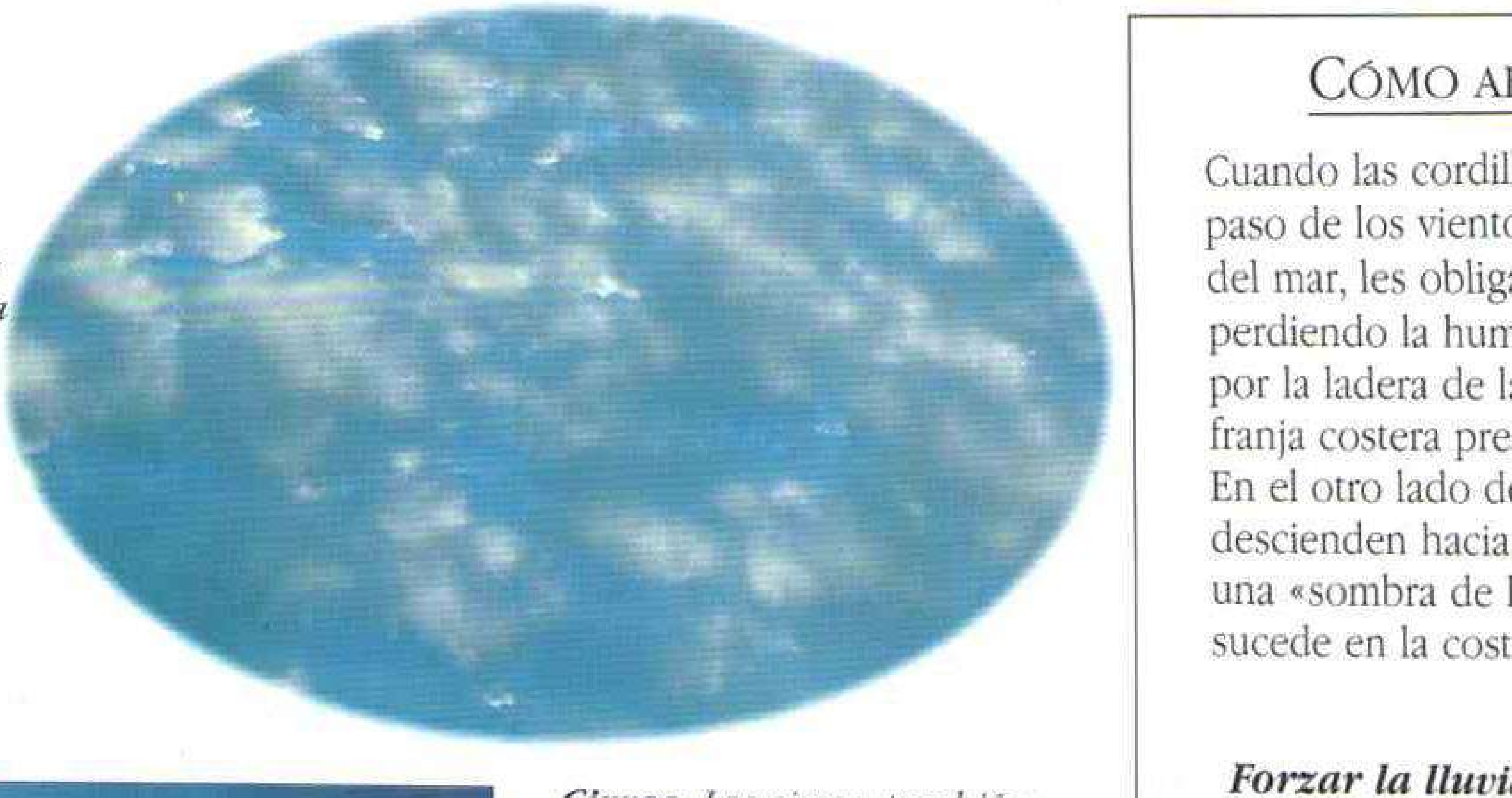

Estratocúmulos. *Los estratocúmulos se forman encima de los cúmulos, extendiéndose para formar una gruesa capa. Los estratocúmulos pueden producir lluvias ligeras que se dispersan al atardecer.*

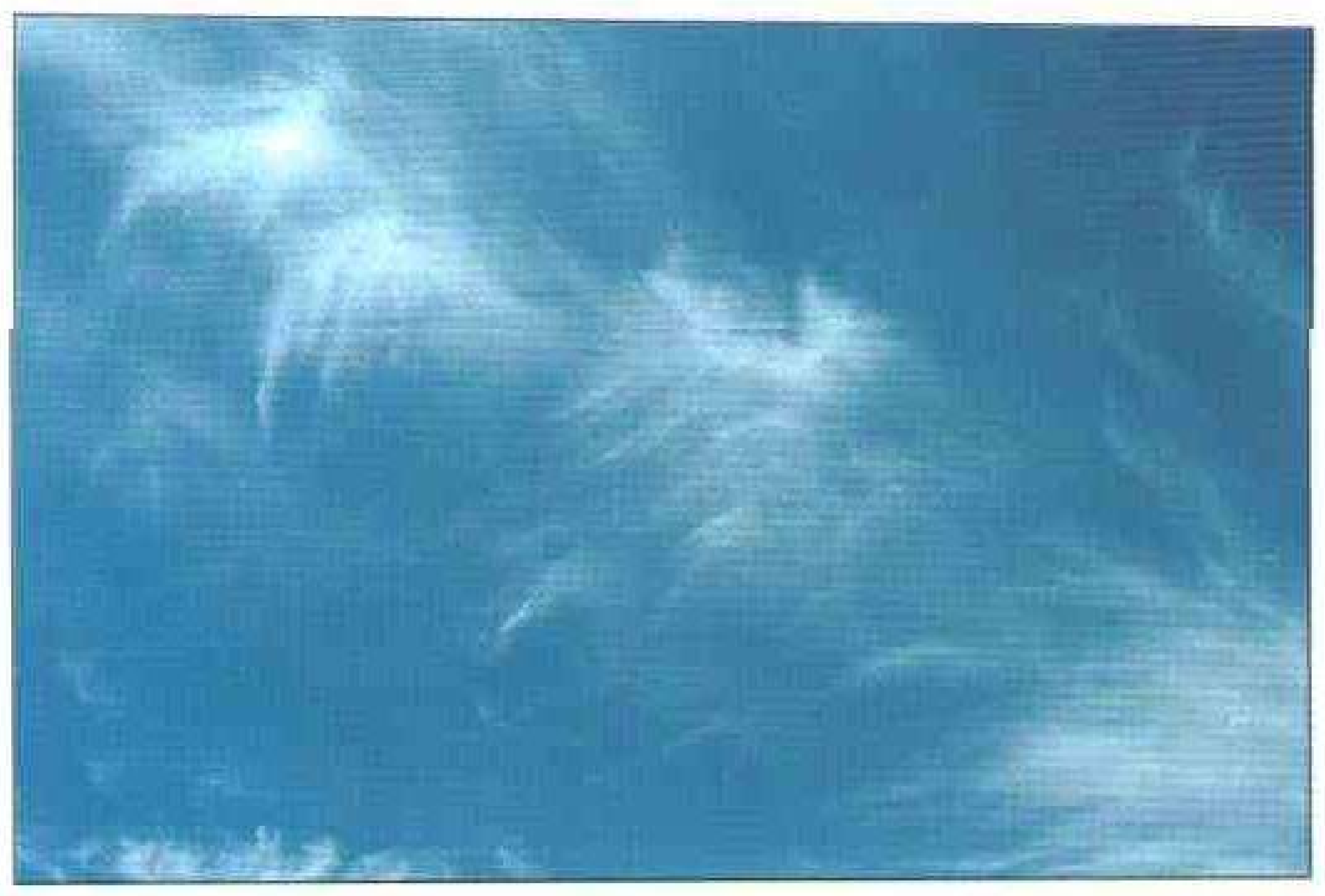

Cirros. *Los cirros, también conocidos como penachos, son nubes muy blancas en forma de filamentos. Tienden a formarse en las grandes alturas si el tiempo es bueno. Como la atmósfera es tan fría a esas alturas, estas pequeñas nubes están formadas totalmente por cristales de hielo.*

CÓMO AFECTA EL PAISAJE AL TIEMPO

Cuando las cordilleras o las montañas obstaculizan el paso de los vientos cargados de humedad provenientes del mar, les obligan a elevarse, formando nubes y perdiendo la humedad en forma de lluvia. Esta lluvia cae por la ladera de la montaña hasta el mar, haciendo que la franja costera presente una intensa precipitación pluvial. En el otro lado de la montaña, los vientos, ahora secos, descienden hacia el interior. Al no contener lluvia, se crea una «sombra de lluvia» que deja el interior seco (como sucede en la costa oriental de Australia).

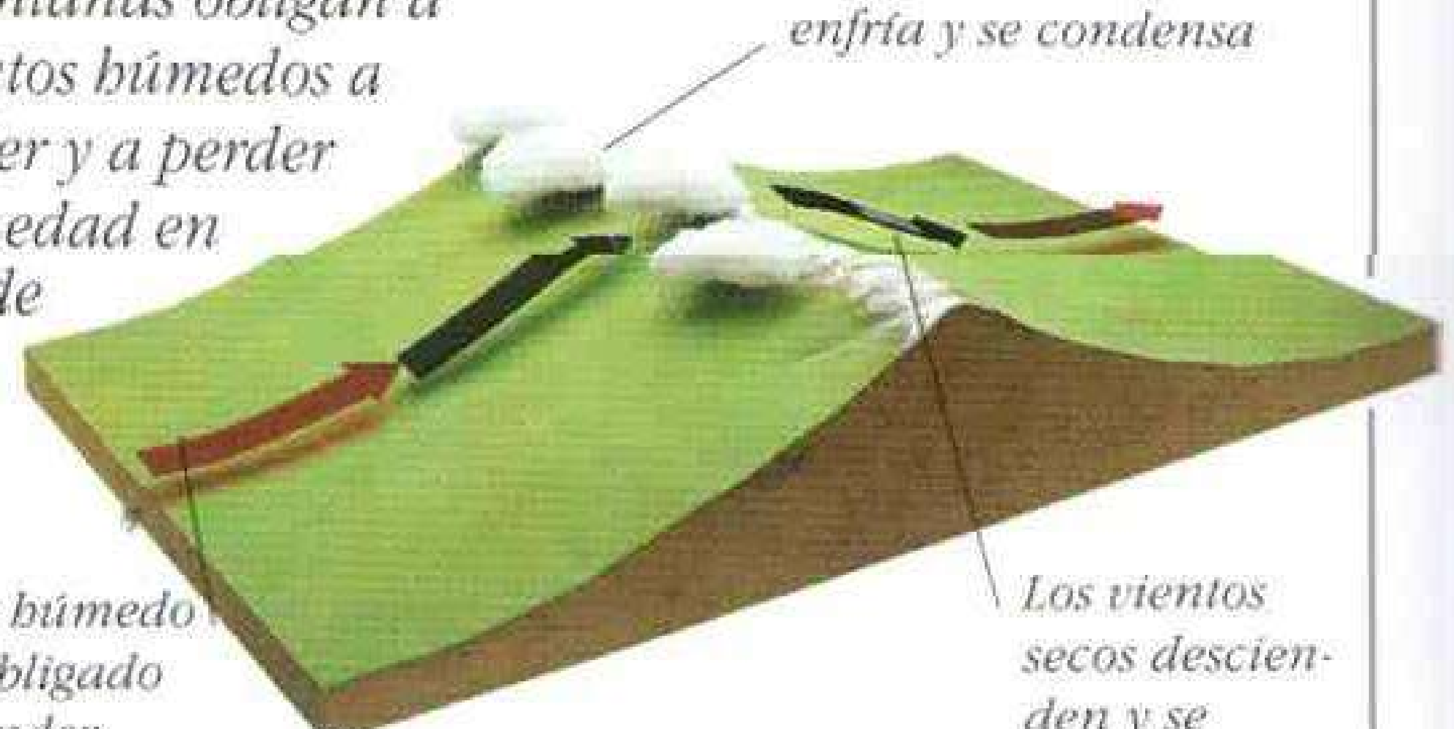

Forzar la lluvia. *Las montañas obligan a los vientos húmedos a ascender y a perder su humedad en forma de lluvia.*

Las nubes se forman cuando el aire se enfría y se condensa

El aire húmedo se ve obligado a ascender

Los vientos secos descienden y se calientan

LA PREDICCIÓN NATURAL DEL TIEMPO

Muchos de los llamados *cuentos de viejas* están basados en la observación. Por ejemplo, la antiquísima rima inglesa «Cielo rojo al atardecer —delicia del pastor; cielo rojo al amanecer —alarma del pastor» puede resultar ser una predicción precisa. El comportamiento animal también puede indicar cambios en el tiempo, al igual que algunos fenómenos naturales (arco iris).

Arco iris. *Un arco iris se origina cuando la luz del sol se refracta y refleja en las pequeñas gotas de vapor de agua en el aire después de la lluvia. Las gotas actúan como prisma, dividiendo la luz del sol en los diferentes colores del espectro. La presencia de un arco iris indica que el tiempo será bueno, particularmente si aparece durante la tarde.*

Animales en migración. *Los animales detectan los cambios de la presión atmosférica que preceden cualquier cambio en el tiempo. Cuando se aproxima una tormenta o nevada, muchos animales gregarios que viven en las grandes alturas se reúnen y bajan a las tierras más bajas y seguras.*

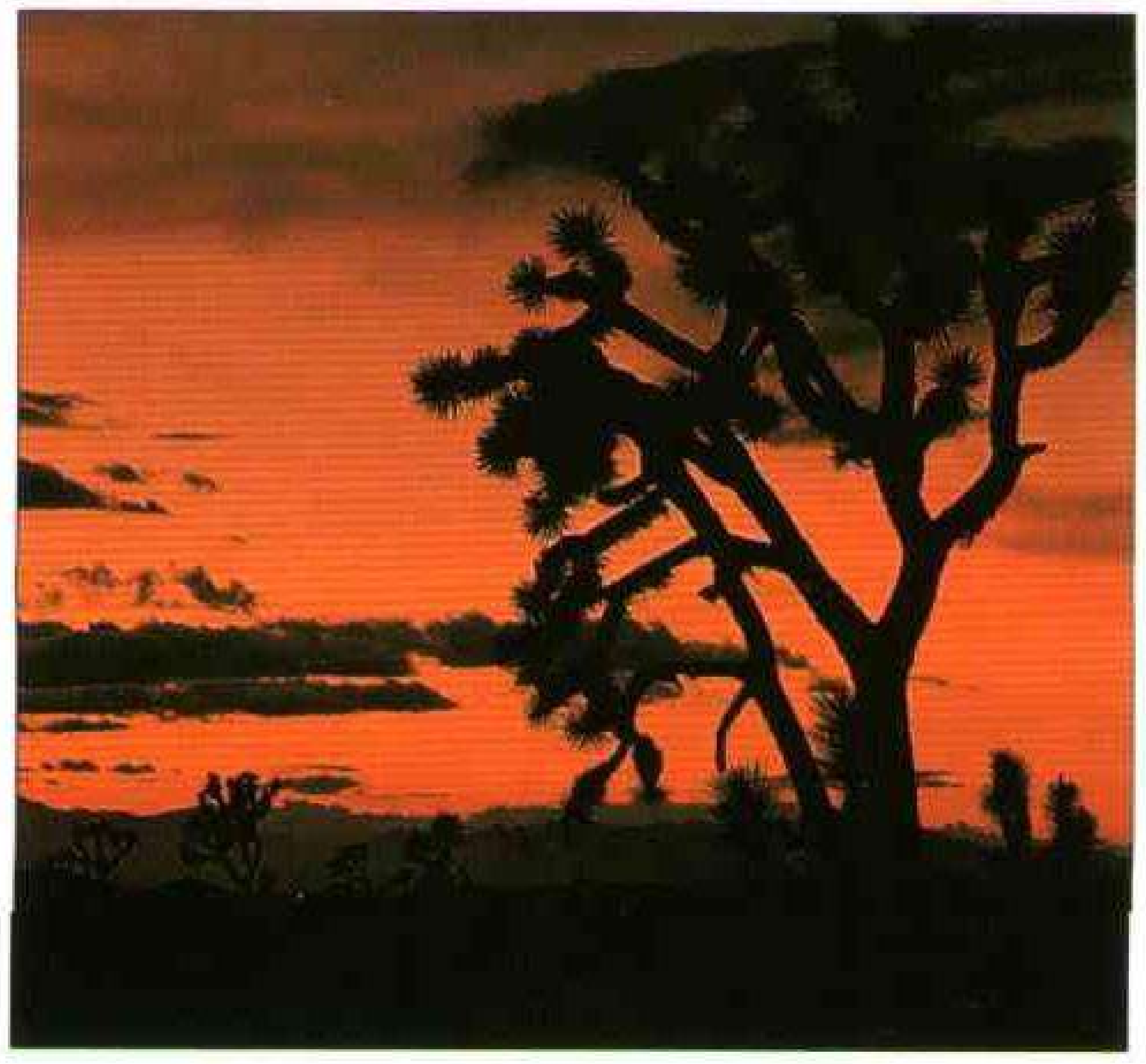

Amanecer rojizo. *Un cielo rojo al amanecer a menudo indica que hay una gran cantidad de humedad en el aire, ya que el sol se refleja en las nubes. Esto significa, generalmente, que se aproxima una tormenta, como bien dice una antigua rima* (véase superior).

LOS DISTINTOS SIGNOS DE LA NATURALEZA

Un cambio en el tiempo puede ser indicado por un cambio en la dirección o en la fuerza del viento, o por cambios en la formación de las nubes. Los vientos predominantes traen, por lo general, un tiempo particular en cada ocasión. Un viento seco y continuado que cambia de dirección o amaina suele preceder a la lluvia. El rocío matinal o la niebla indican un tiempo estable, pero el viento —especialmente junto con niebla en las montañas bajas— puede traer lluvia. Un cielo claro al anochecer indica que la noche será fría y que probablemente helará, ya que no hay nubes que retengan el calor. Antes de llover, el aumento de la presión atmosférica incrementa los dolores reumáticos, hace que las plantas abran sus poros y produce la dilatación de los objetos de madera. Asimismo, los sonidos parecen viajar más lejos.

Piñas de los pinos. *Las piñas de los pinos también reaccionan ante la humedad. Si el aire es seco, las glumas de la piña se encogen y ésta se abre. Poco antes de la lluvia, las glumas absorben la humedad y la piña recobra su forma.*

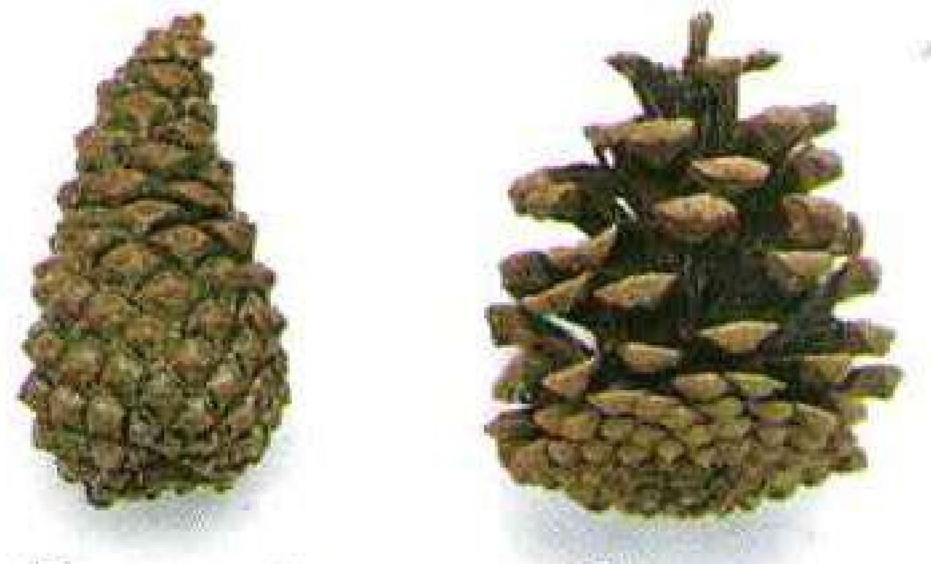

TIEMPO HÚMEDO TIEMPO SECO

ESCALA DE BEAUFORT PARA MEDIR LA VELOCIDAD DEL VIENTO

Fuerza	Descripción	Velocidad (km / h; m / s)	Efectos
0	Calma	0-1; menos de 0,3	El humo sube verticalmente
1	Ventolina	1-6; 0,3-1,5	El humo, aunque no el viento, indican la dirección
2	Viento flojito o brisa	6-12; 1,6-3,3	El viento se nota en la cara; las hojas revolotean
3	Viento flojo	12-20; 3,4-5,4	Hojas en continuo movimiento; el viento hace ondear las banderas
4	Viento bonancible	20-29; 5,5-7,9	Se levantan el polvo y los papeles sueltos; se mueven las ramas pequeñas
5	Viento fresquito	29-39; 8,0-10,7	Se inclinan los árboles pequeños con hojas
6	Viento fresco	39-50; 10,8-13,8	Las ramas grandes se mueven; los paraguas se usan con dificultad
7	Viento frescachón	50-62; 13,9-17,1	Los árboles se mueven; dificultad para caminar
8	Viento duro	62-75; 17,2-20,7	Se rompen las ramas de los árboles; la circulación es prácticamente imposible
9	Viento muy duro	75-92; 20,8-25,4	Ligeros daños en las construcciones; se levantan las tejas de las chimeneas
10	Temporal	92-102; 25,5-28,4	Árboles desarraigados; daños estructurales considerables
11	Borrasca	102-117; 28,5-32,6	Daños generalizados
12	Huracán	Más de 117; más de 32,6	Zona devastada

Escala de Beaufort para medir la velocidad del viento. *Los cambios de tiempo están anunciados por un cambio en la velocidad del viento. Utilizando la escala de Beaufort, puede evaluarse la velocidad del mismo. En esta escala, la velocidad del viento se divide en 12 fuerzas, desde la fuerza 0 o calma hasta la fuerza 12 o huracán.*

PREPARARSE PARA EL VIAJE

UNA VEZ SE ha decidido a viajar, es fundamental prepararlo todo con antelación. Lave, repare y guarde su equipo. Evalúe el terreno que planea cruzar *(véase página 122)*, y planifique su ruta con la mayor exactitud posible. Reúna reservas de agua y de alimento, y lleve consigo todo aquello que pudiera serle útil en el camino o en su próximo campamento, como los materiales para el refugio. En una situación de supervivencia, si se ve obligado a cambiar de campamento, asegúrese de que tanto usted como sus compañeros tienen la condición física necesaria para realizar el viaje. Puede ser necesaria la construcción de algún equipo antes de viajar *(véase página siguiente)*.

EL USO DE LA MOCHILA

Su mochila debe estar preparada de tal manera que pueda acceder rápidamente a la comida y a la ropa durante la marcha, además de llevar el resto de su equipo de campamento con toda comodidad. No ate ningún artículo a la parte exterior; podría dañarse o perderse. Coloque todo aquello a lo que deba acceder rápidamente en los bolsillos exteriores. En el interior, dentro de una bolsa grande, resistente e impermeable, debe guardar los artículos más ligeros y voluminosos en el fondo, y los más pesados encima, para mantener el centro de gravedad de la mochila lo más alto posible. Dentro de su mochila debe guardar la ropa, la comida, los artículos de limpieza, un botiquín de primeros auxilios y otros artículos, en bolsas separadas e impermeables. Mientras permanezca en el campamento, mantenga todo lo que no necesite en la mochila para reducir al mínimo el desorden, y para que, en caso de necesidad, pueda abandonar el campamento rápidamente.

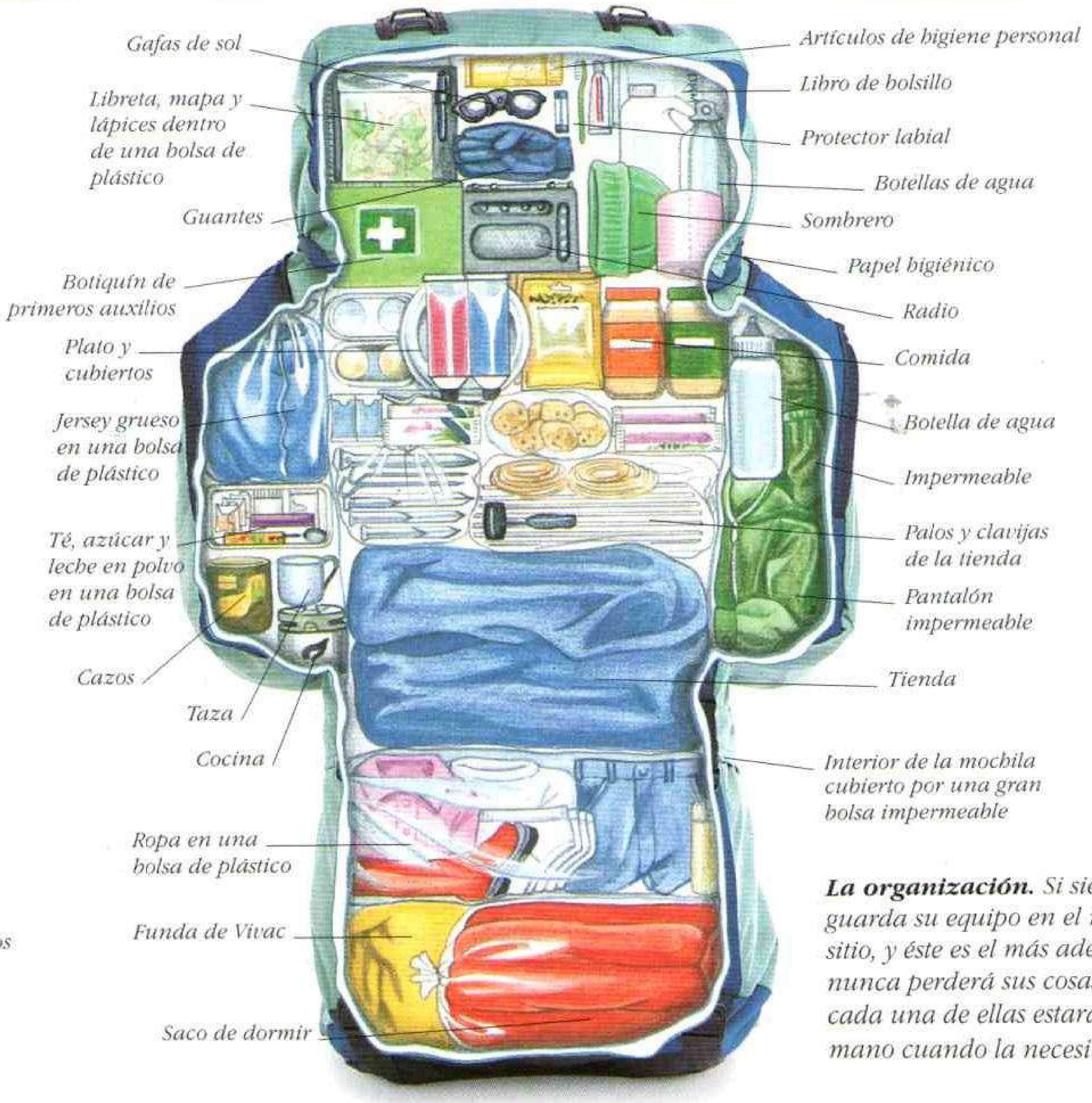
Gafas de sol
Libreta, mapa y lápices dentro de una bolsa de plástico
Guantes
Botiquín de primeros auxilios
Plato y cubiertos
Jersey grueso en una bolsa de plástico
Té, azúcar y leche en polvo en una bolsa de plástico
Cazos
Taza
Cocina
Ropa en una bolsa de plástico
Funda de Vivac
Saco de dormir
Artículos de higiene personal
Libro de bolsillo
Protector labial
Botellas de agua
Sombrero
Papel higiénico
Radio
Comida
Botella de agua
Impermeable
Palos y clavijas de la tienda
Pantalón impermeable
Tienda
Interior de la mochila cubierto por una gran bolsa impermeable

La organización. *Si siempre guarda su equipo en el mismo sitio, y éste es el más adecuado, nunca perderá sus cosas, y cada una de ellas estará a mano cuando la necesite.*

El peso debe llevarse sobre los hombros

Llevar una mochila. *Mantenga el centro de gravedad de la mochila lo más alto posible, para evitar que tire de usted hacia atrás. Sus piernas deben realizar el trabajo de cargar la mochila. Un cinturón acolchado para la cadera permite que una parte del peso sea transferida de los hombros a la pelvis, liberando una parte del peso de la columna y de los hombros. Las correas reparten una parte del peso al pecho. Al ser ajustables, permiten que las mochilas puedan llevarse en la parte superior de la espalda.*

BOLSAS Y MOCHILAS

Para un viaje corto, es mucho más cómodo llevar una bolsa pequeña que contenga alimento, ropa y provisiones, que toda la mochila. Los artículos de higiene personal pueden llevarse en un pequeño neceser, mientras que los documentos de viaje pueden guardarse en portadocumentos impermeables.

MOCHILA

NECESER

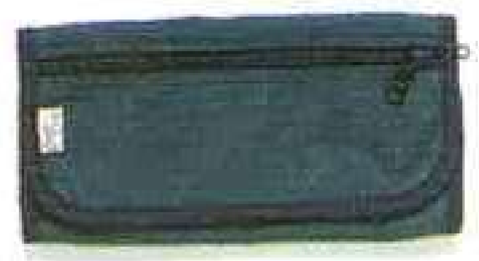
PORTADOCUMENTOS

CONSTRUCCIÓN DE UN ARMAZÓN PARA LA MOCHILA

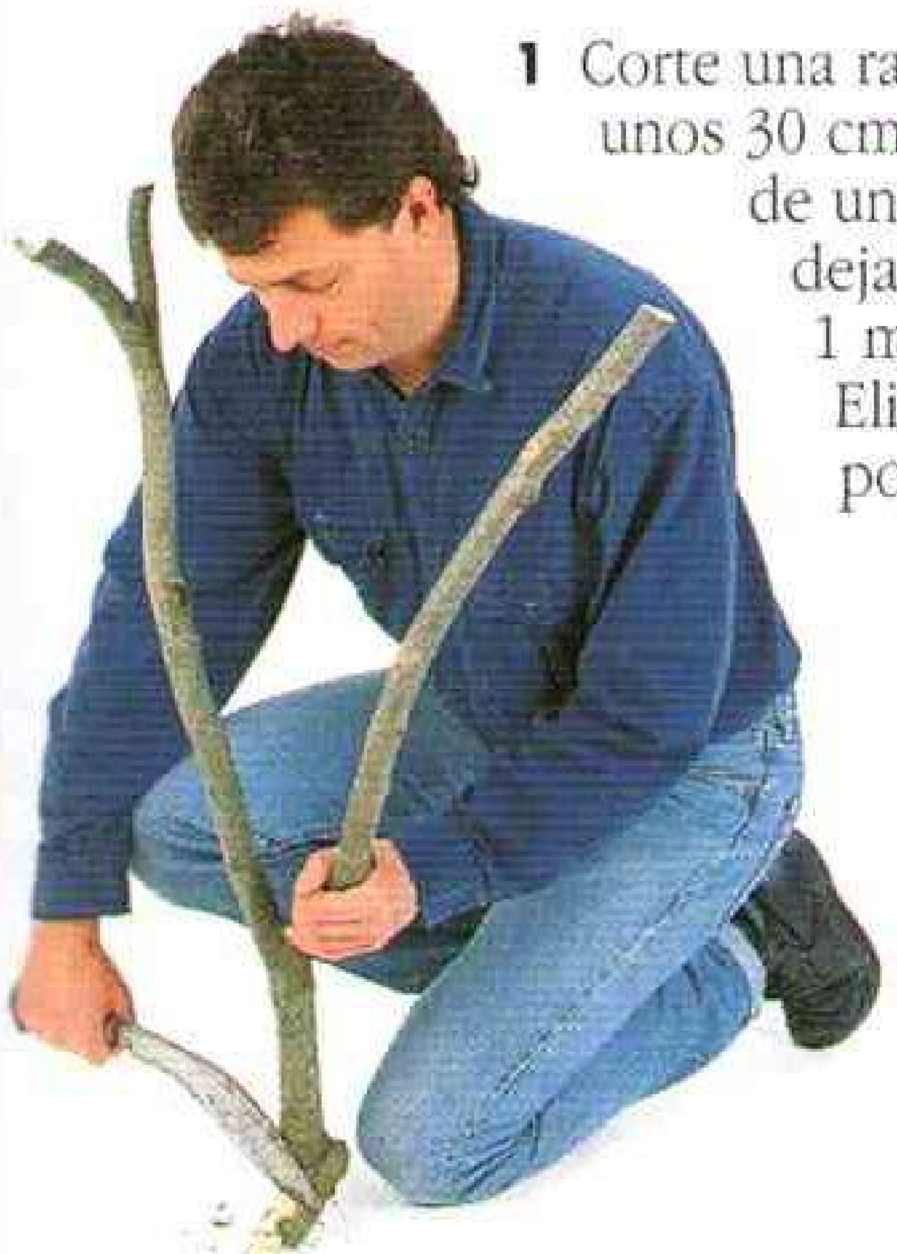

1 Corte una rama ligera unos 30 cm por debajo de una bifurcación, dejando cerca de 1 m de horquilla. Elimine cualquier posible nudo.

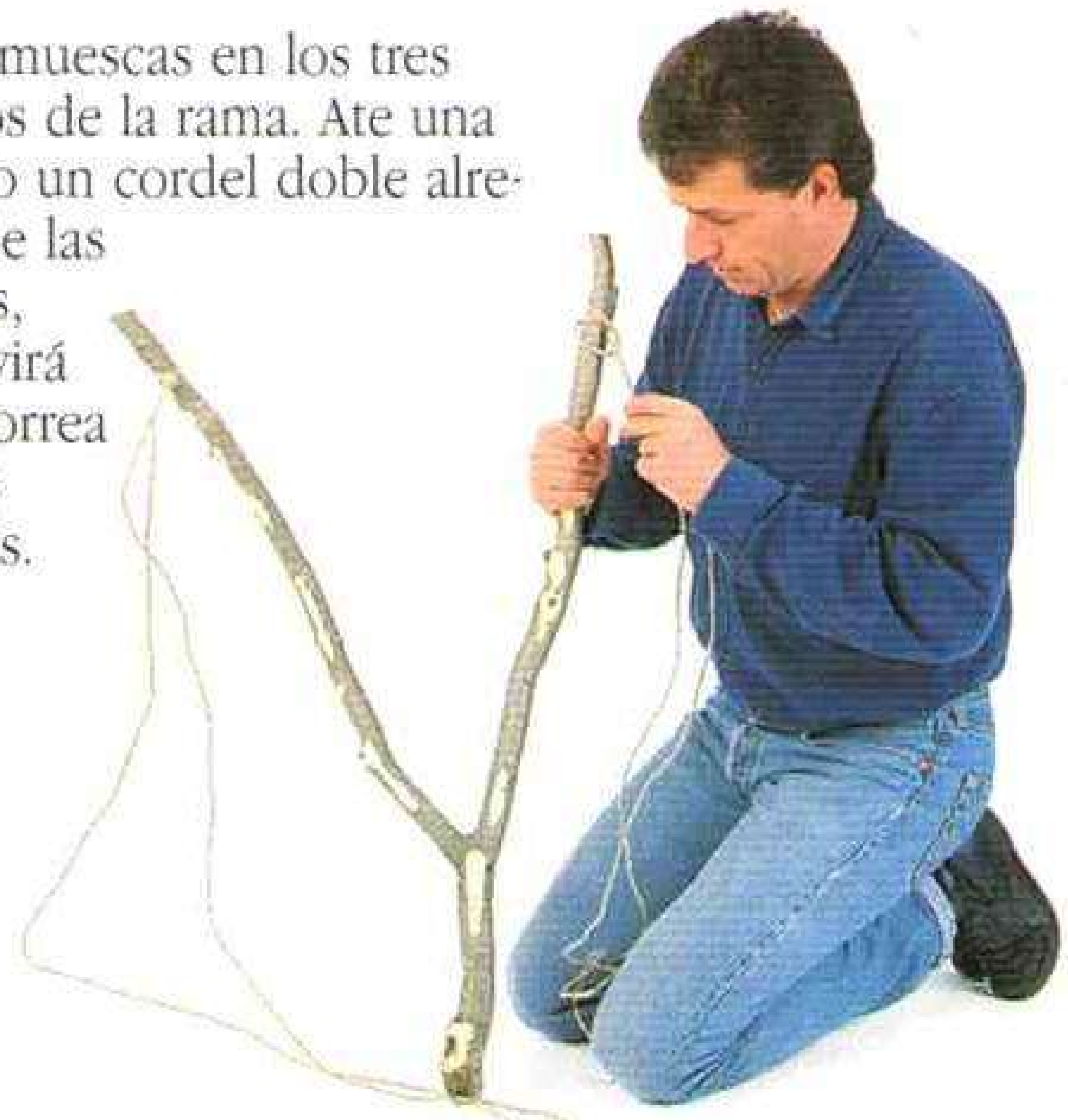

2 Talle muescas en los tres extremos de la rama. Ate una cuerda o un cordel doble alrededor de las muescas, que servirá como correa para los hombros.

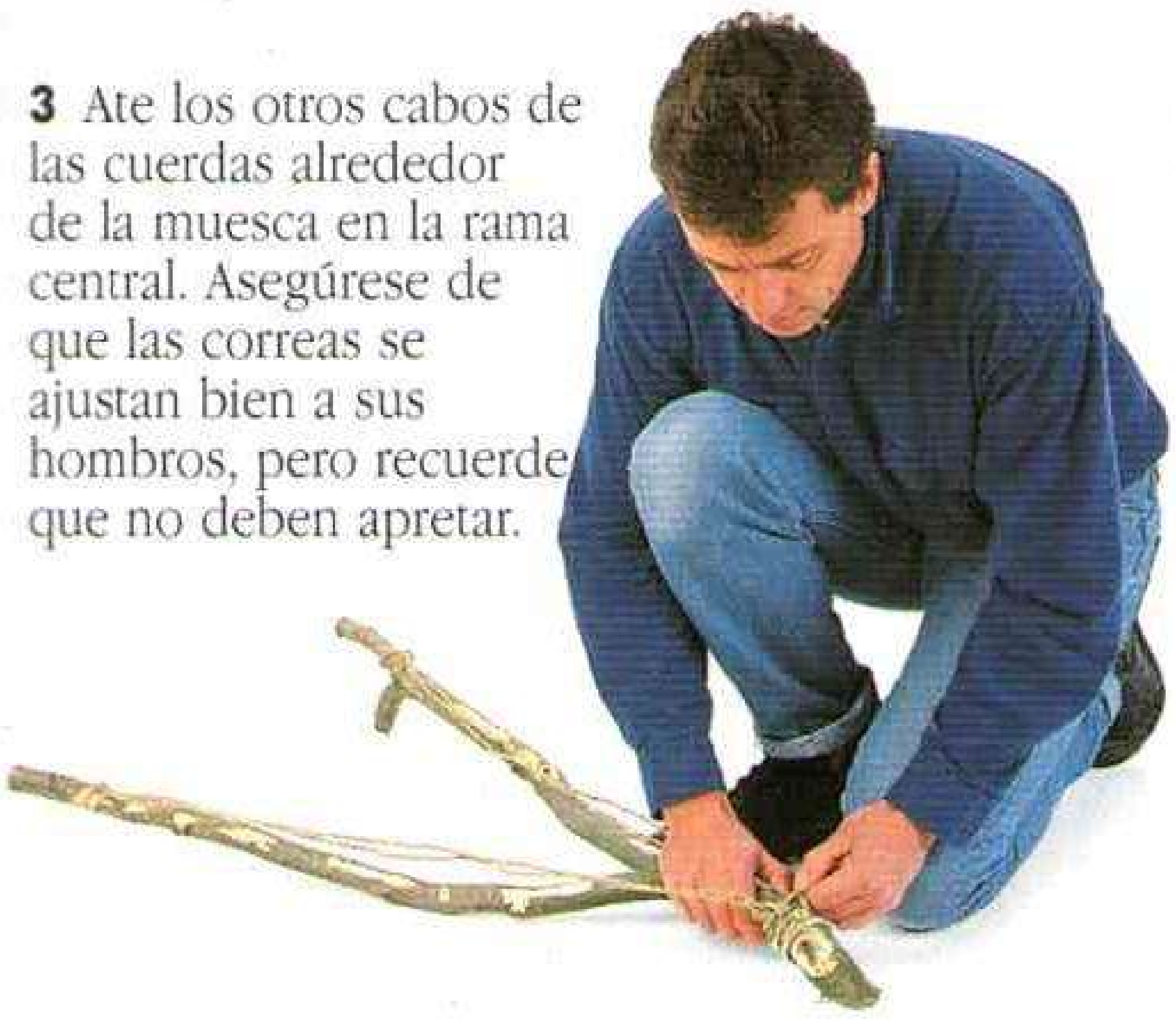

3 Ate los otros cabos de las cuerdas alrededor de la muesca en la rama central. Asegúrese de que las correas se ajustan bien a sus hombros, pero recuerde que no deben apretar.

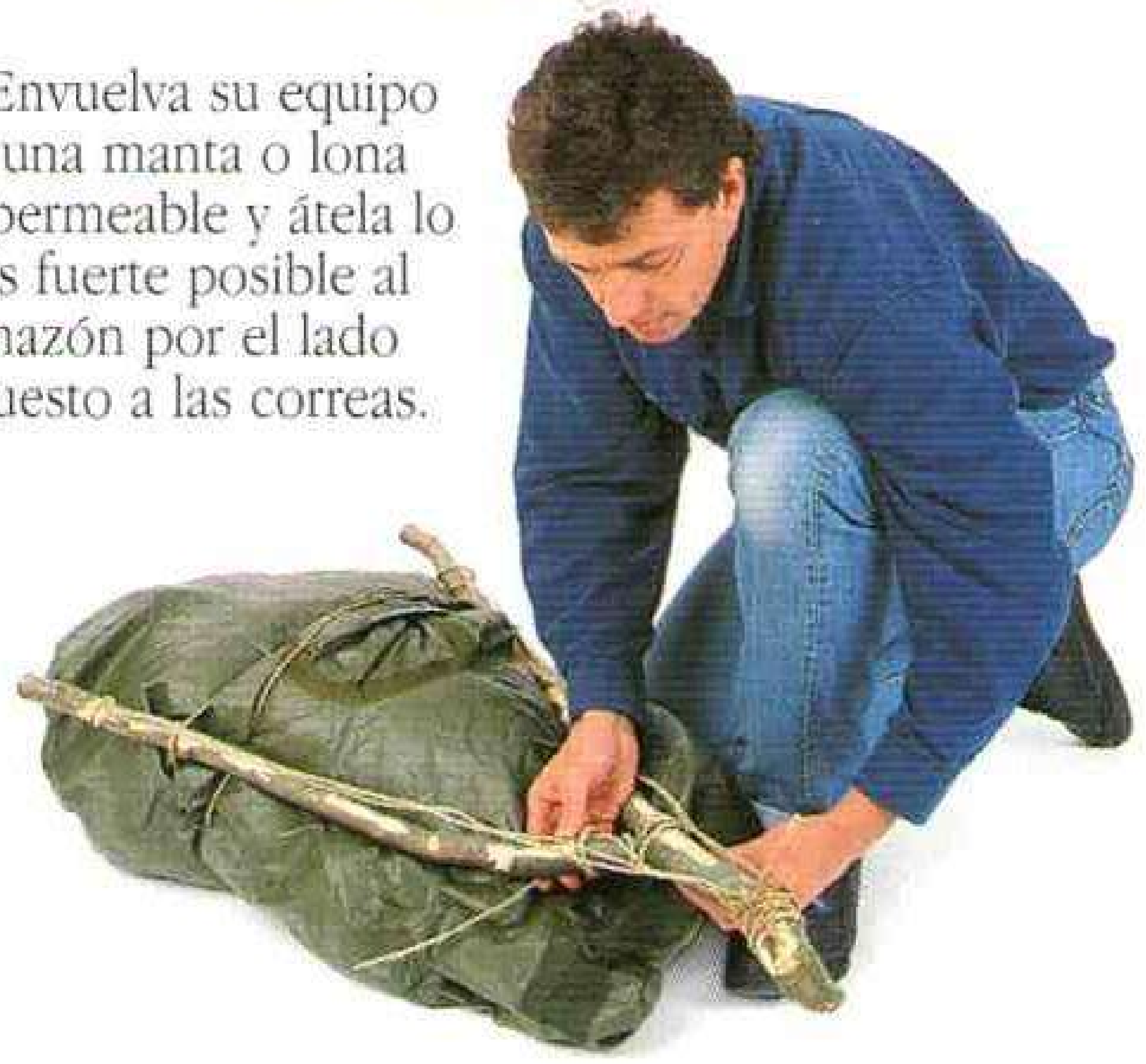

4 Envuelva su equipo en una manta o lona impermeable y átela lo más fuerte posible al armazón por el lado opuesto a las correas.

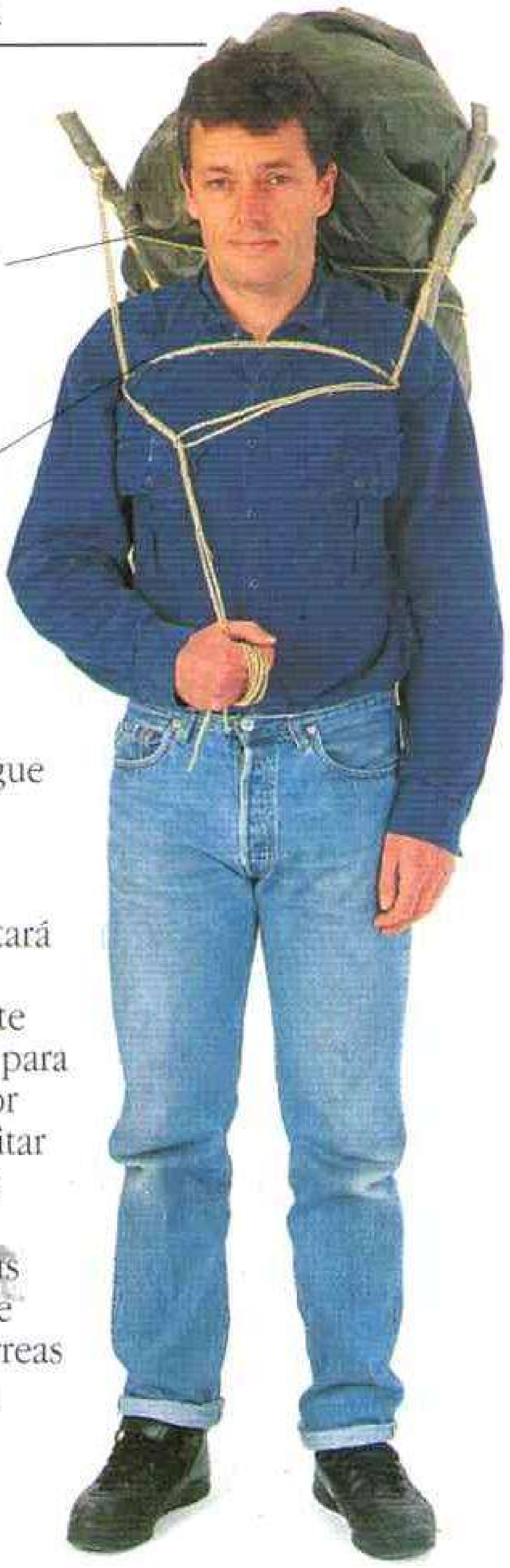

El equipo debe estar firmemente atado al armazón

Debe mantener las correas alejadas de los hombros para evitar irritaciones o rozaduras

5 No sobrecargue su mochila improvisada —aunque el armazón soportará bien un peso considerable. Ate las dos correas para los hombros por delante para evitar que las cuerdas finas causen rozaduras en sus hombros. Puede acolchar las correas con vegetación.

CONSTRUCCIÓN DE UN REMOLQUE O *TRAVOIS*

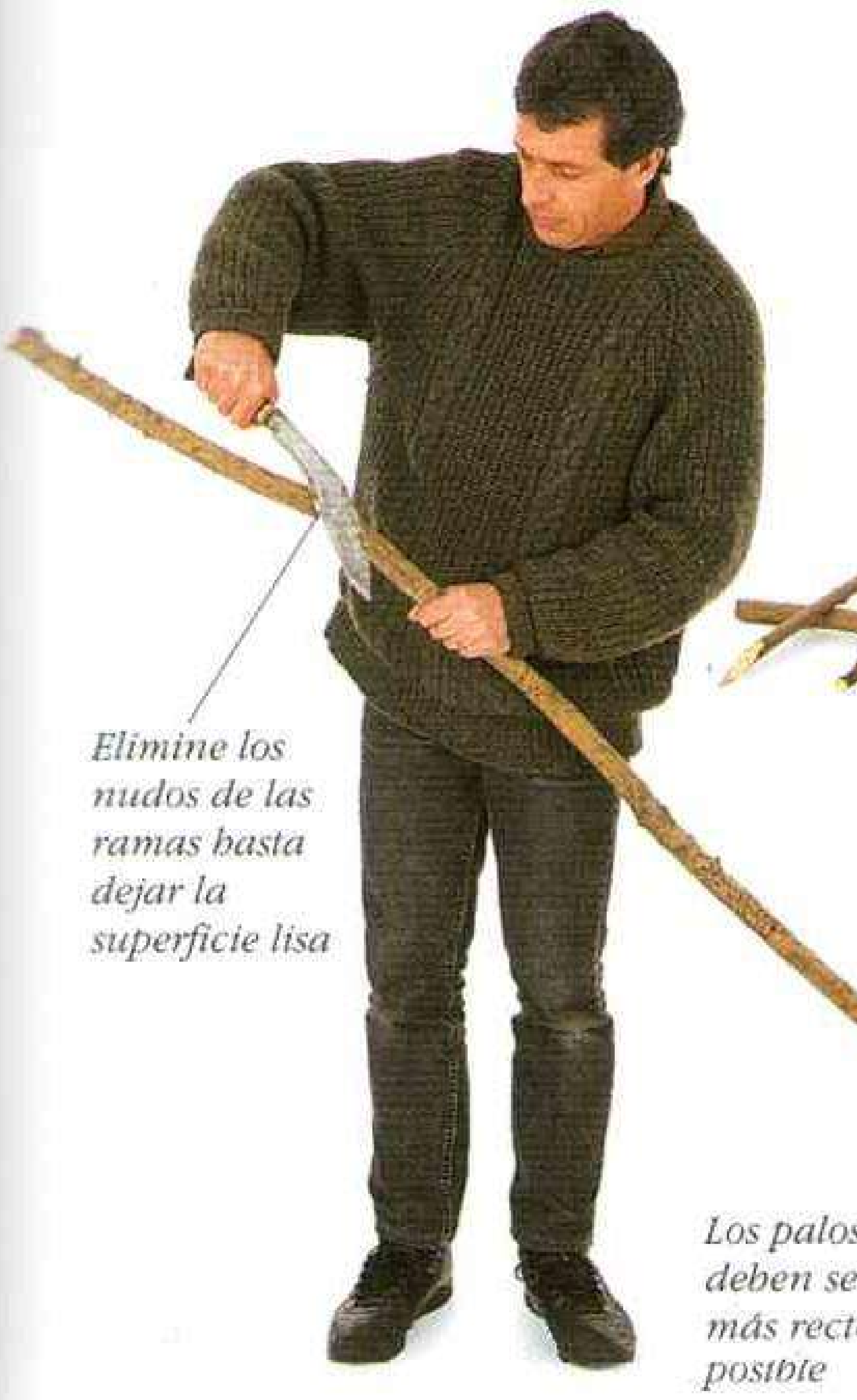

Elimine los nudos de las ramas hasta dejar la superficie lisa

Los palos deben ser lo más rectos posible

1 Corte dos palos largos y verticales de al menos 2 m de longitud. Desbaste los palos con un cuchillo para que las riostras, que deberán soportar la carga, se ajusten firmemente sobre los mismos.

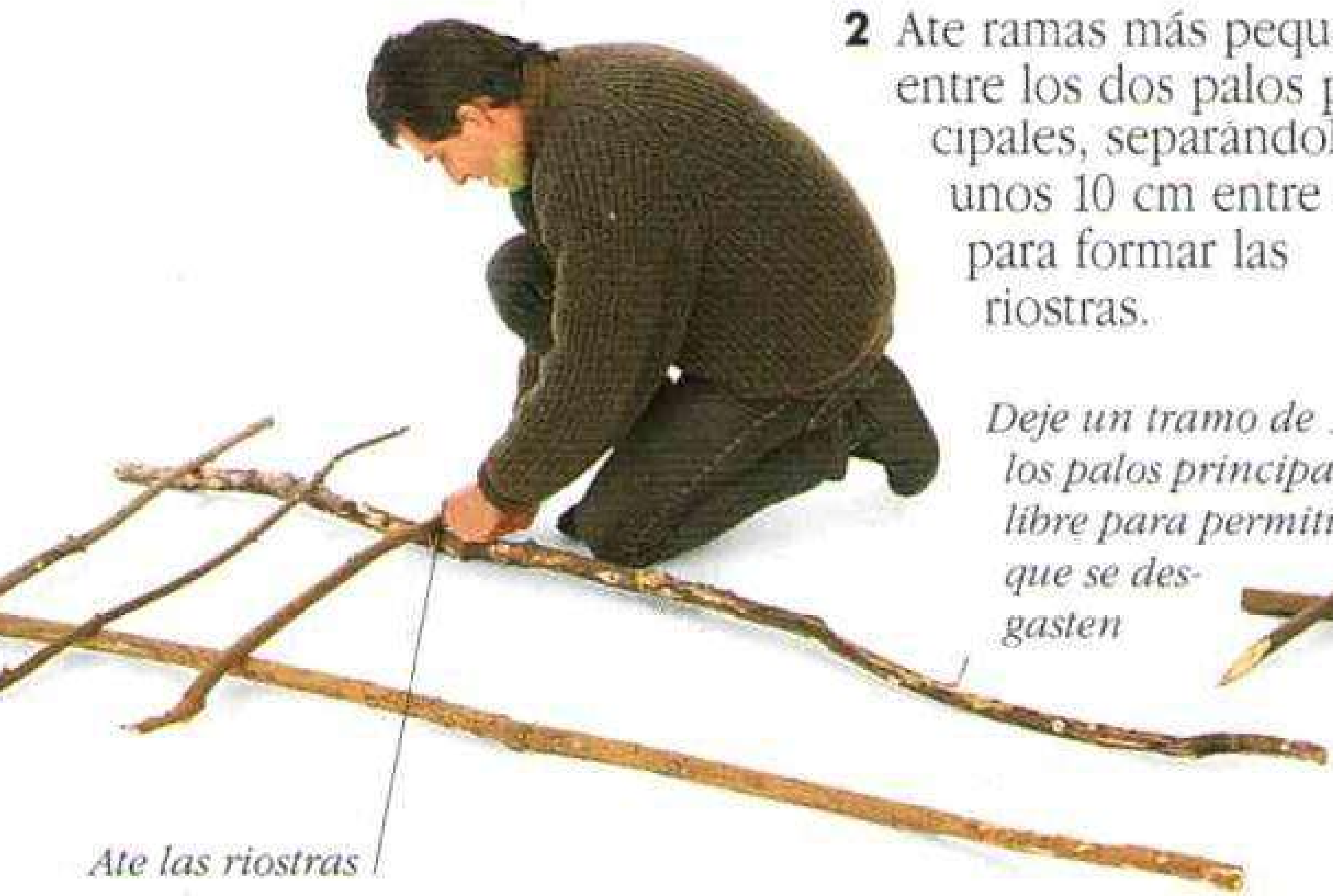

Ate las riostras con fuerza a los palos principales

Deje un tramo de los palos principales libre para permitir que se desgasten

2 Ate ramas más pequeñas entre los dos palos principales, separándolas unos 10 cm entre sí para formar las riostras.

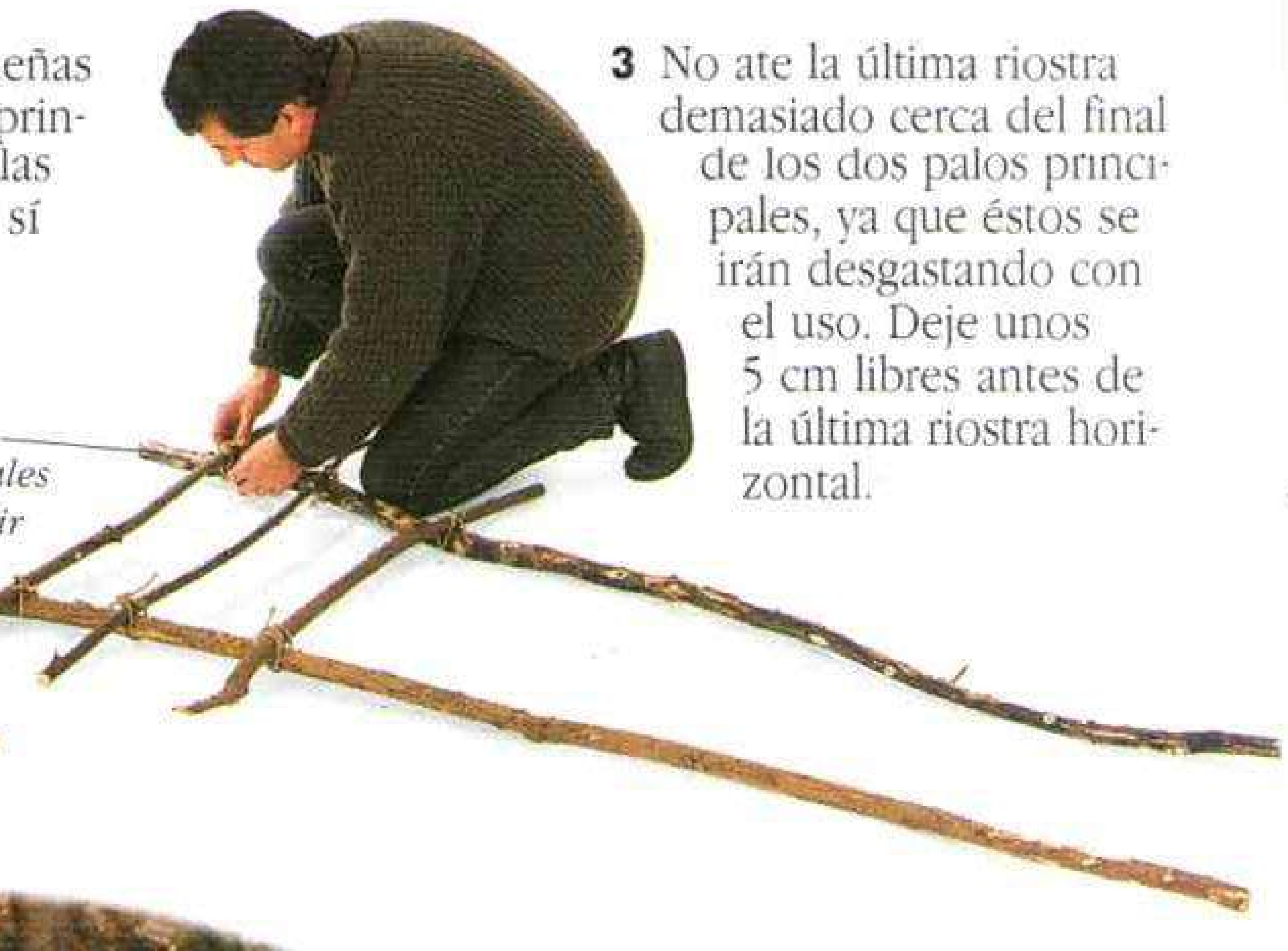

3 No ate la última riostra demasiado cerca del final de los dos palos principales, ya que éstos se irán desgastando con el uso. Deje unos 5 cm libres antes de la última riostra horizontal.

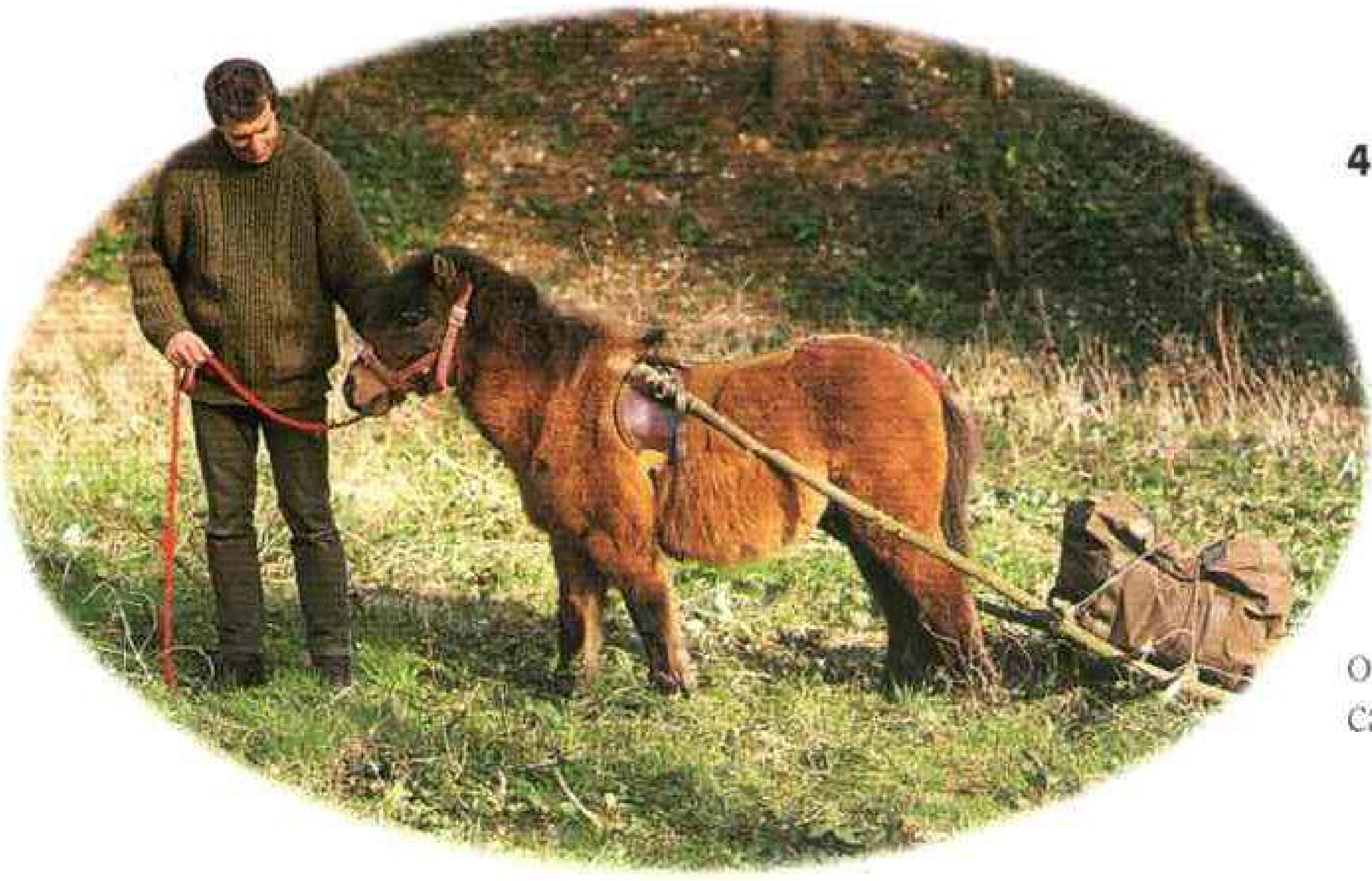

4 Asegúrese de atar todo el equipo sobre el remolque. Puede arrastrarlo usted mismo o engancharlo a un animal, como un caballo o un perro. Compruebe que la carga no toque sus talones o los del animal, o que no se suelte por el camino.

LA MARCHA A PIE

EL IR A pie llevando un equipo pesado nada tiene que ver con nuestros paseos diarios urbanos. Al caminar en grupo hay que seguir el ritmo del miembro más lento, y se requiere una planificación, organización y previsión muy detalladas. Una vez comenzado el viaje, hay que detenerse después de 10 minutos para ajustar los calcetines, botas, vestimenta y equipo, comprobando que se mantiene la dirección correcta y que todos se encuentran bien. A partir de ese momento, hay que seguir un ritmo, aminorando la marcha al ir cuesta arriba o cuesta abajo.

TÉCNICAS DE MARCHA

Camine lenta y regularmente, balanceando los brazos para mantener el impulso y el equilibrio, y no fuerce los pasos. Intente mantener un paso uniforme y continuado —aquél que pueda seguir durante un período prolongado sin perder el aliento o lastimarse. Intente detenerse a intervalos regulares para descansar y comprobar su equipo.

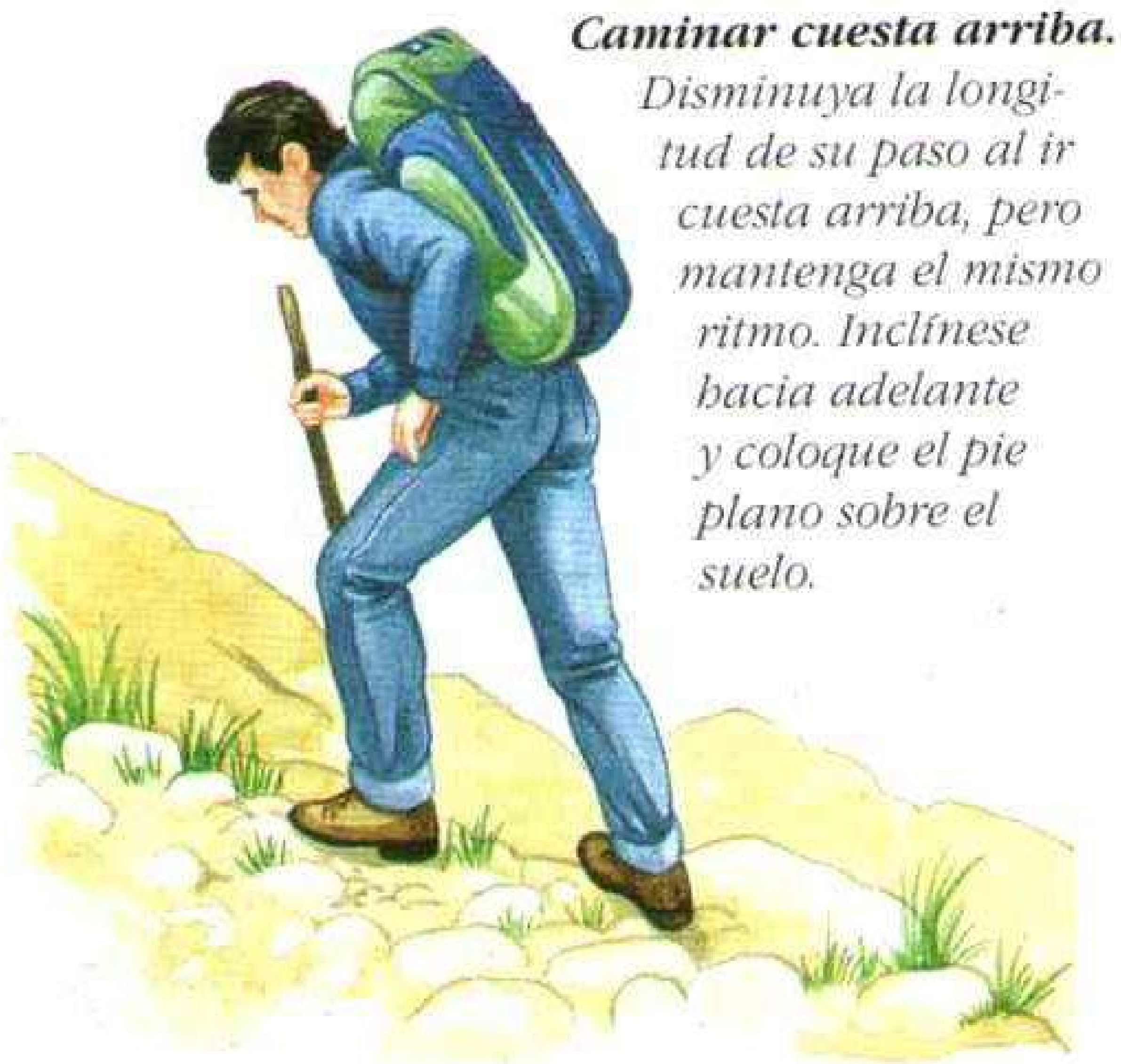

Caminar cuesta arriba. *Disminuya la longitud de su paso al ir cuesta arriba, pero mantenga el mismo ritmo. Inclínese hacia adelante y coloque el pie plano sobre el suelo.*

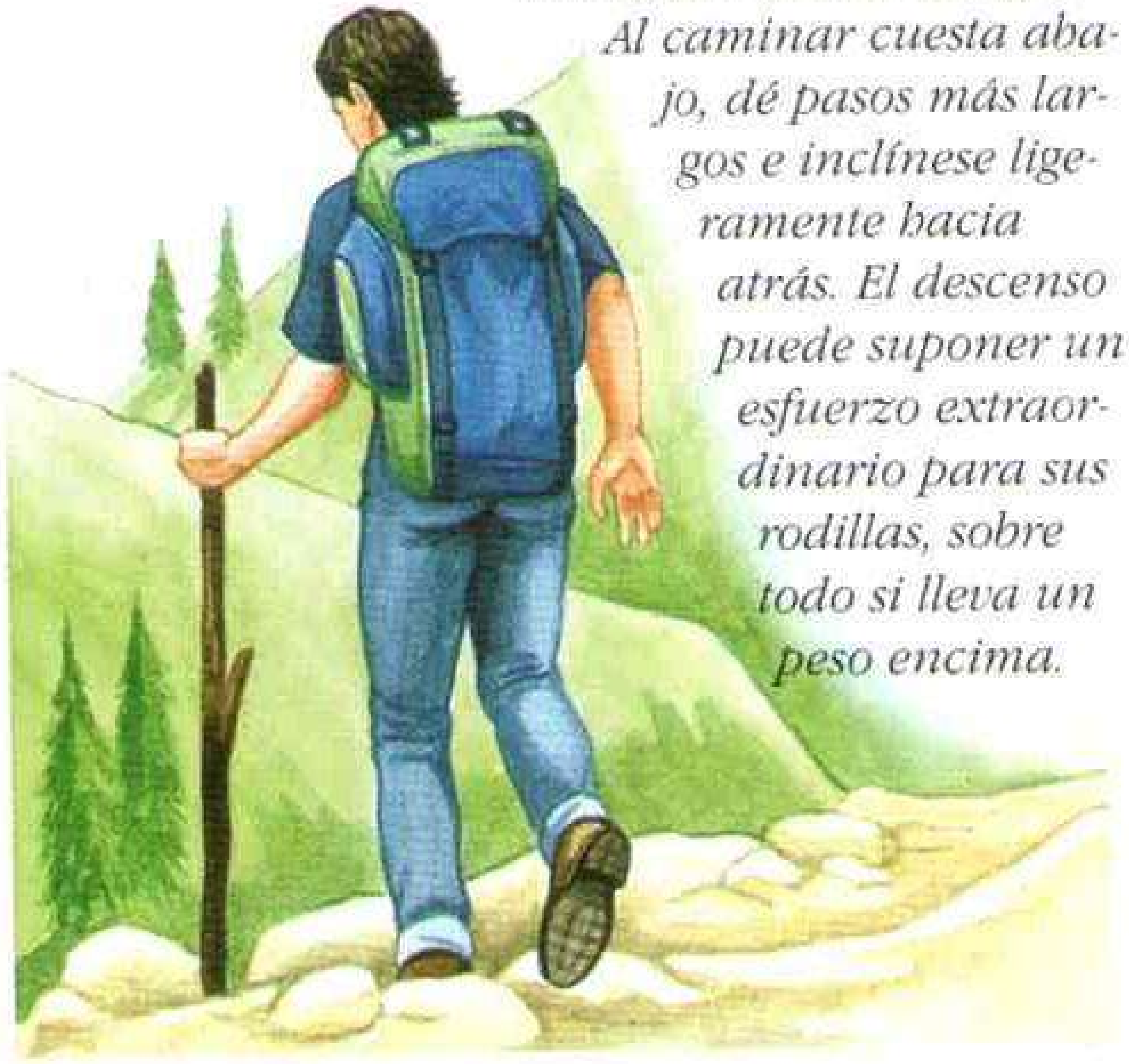

Caminar cuesta abajo. *Al caminar cuesta abajo, dé pasos más largos e inclínese ligeramente hacia atrás. El descenso puede suponer un esfuerzo extraordinario para sus rodillas, sobre todo si lleva un peso encima.*

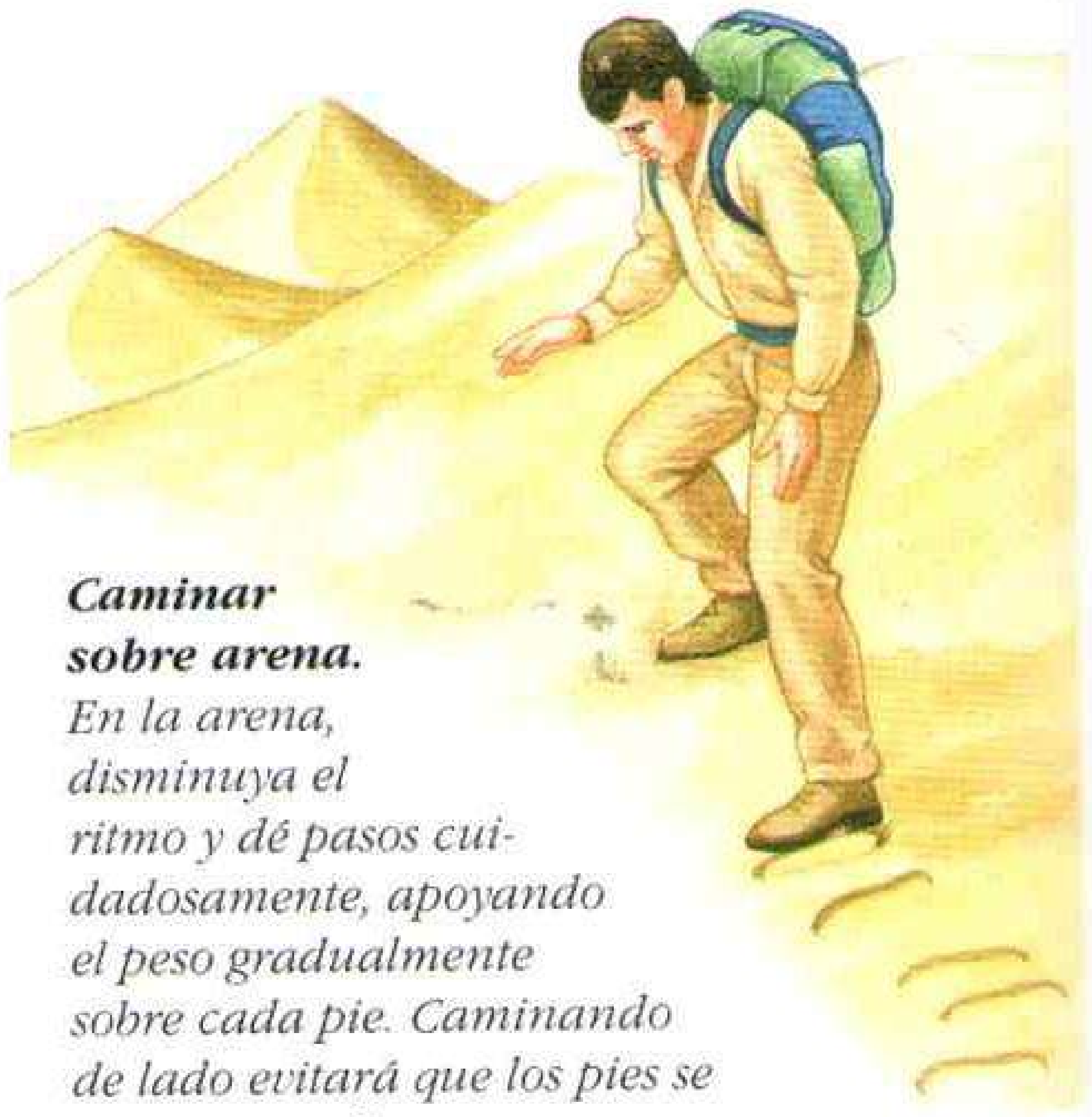

Caminar sobre arena. *En la arena, disminuya el ritmo y dé pasos cuidadosamente, apoyando el peso gradualmente sobre cada pie. Caminando de lado evitará que los pies se entierren en la arena.*

CAMINAR EN GRUPO

Al caminar en grupo, envíe una avanzadilla de personas en buena forma física para comprobar la ruta, encontrar un sendero y buscar la manera de sortear los obstáculos. Al final de la jornada, estos mismos miembros del grupo pueden encontrar también un sitio donde acampar, levantar las tiendas o refugios, recoger agua y encender el fuego. Sin embargo, deben permanecer con el grueso del grupo la cantidad suficiente de personas en forma para ayudar a aquellos que no puedan caminar tan deprisa como los demás, y para ayudar a cualquier persona enferma o herida.

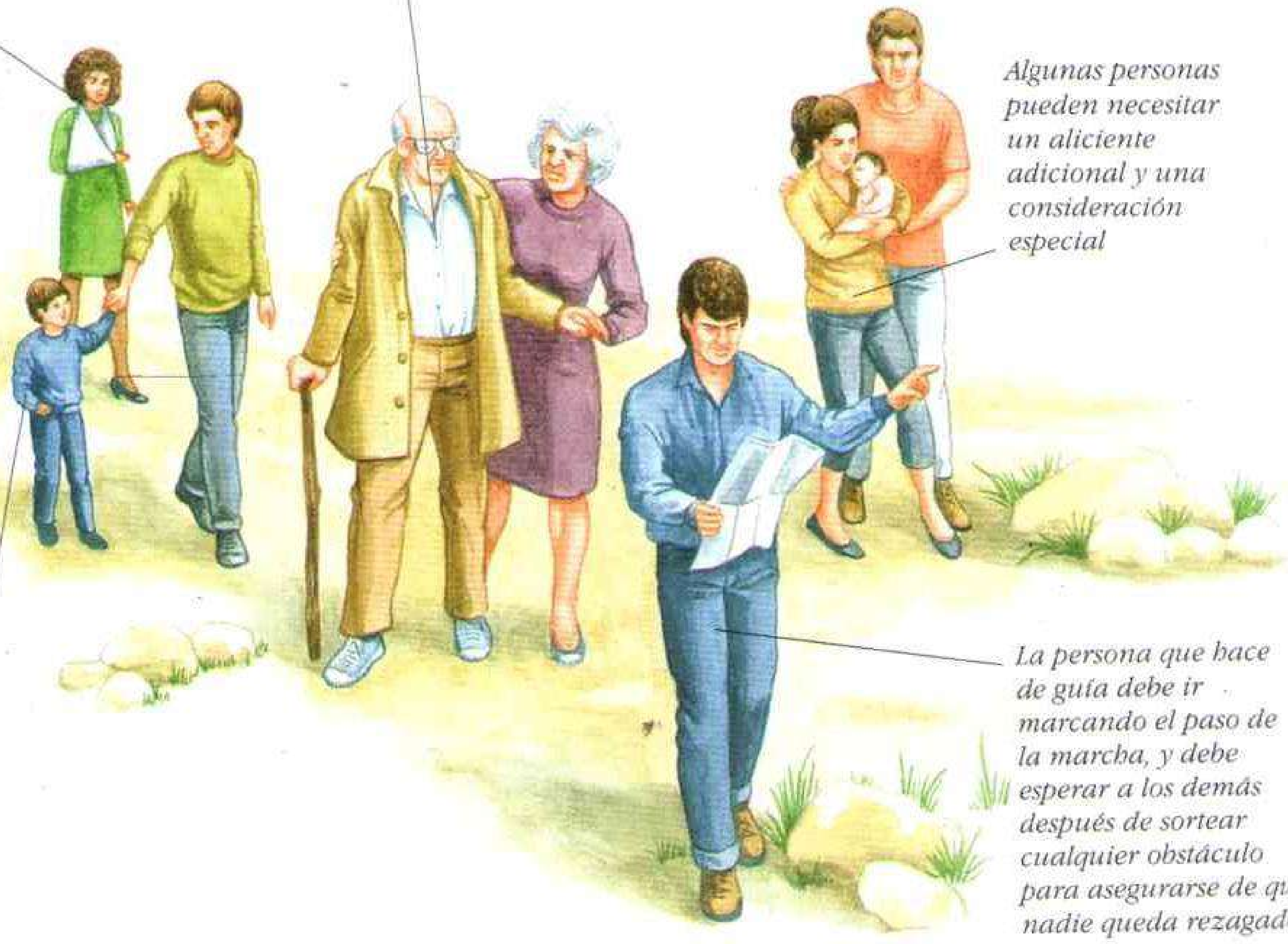

La velocidad del grupo debe ajustarse a la de sus miembros más lentos

Toda persona disminuida o enferma debe tener un acompañante, y ambas deben caminar en medio del grupo para no quedar atrás

Algunas personas pueden necesitar un aliciente adicional y una consideración especial

Vigile que los niños no se adelanten, jugueteen o se retrasen

La persona que hace de guía debe ir marcando el paso de la marcha, y debe esperar a los demás después de sortear cualquier obstáculo para asegurarse de que nadie queda rezagado

Mantenerse juntos. *Es importante que un grupo se mantenga unido —a excepción del grupo de avanzadilla—, en especial en una situación de supervivencia. No deje que los niños o las personas heridas se retrasen. Mantenga un paso que todos puedan seguir.*

PLANIFICAR LA RUTA

La mejor forma de planificar su ruta es recorriéndola primero, aunque esto es generalmente poco factible, en especial en una situación de supervivencia. Antes de ponerse en marcha, busque un punto elevado, como una loma o un árbol al que pueda trepar fácilmente, y utilice los prismáticos y un mapa, así como cualquier otro elemento disponible, para determinar la ruta más fácil *(véase página 122)*. En la selva, las crestas siempre tienen menos vegetación que el fondo de los valles, y son más fáciles de seguir. Sin embargo, puede que existan senderos marcados a lo largo de los cursos de agua, creados por los habitantes del lugar. Intente seguir siempre la ruta más fácil, particularmente si se encuentra en una situación de supervivencia.

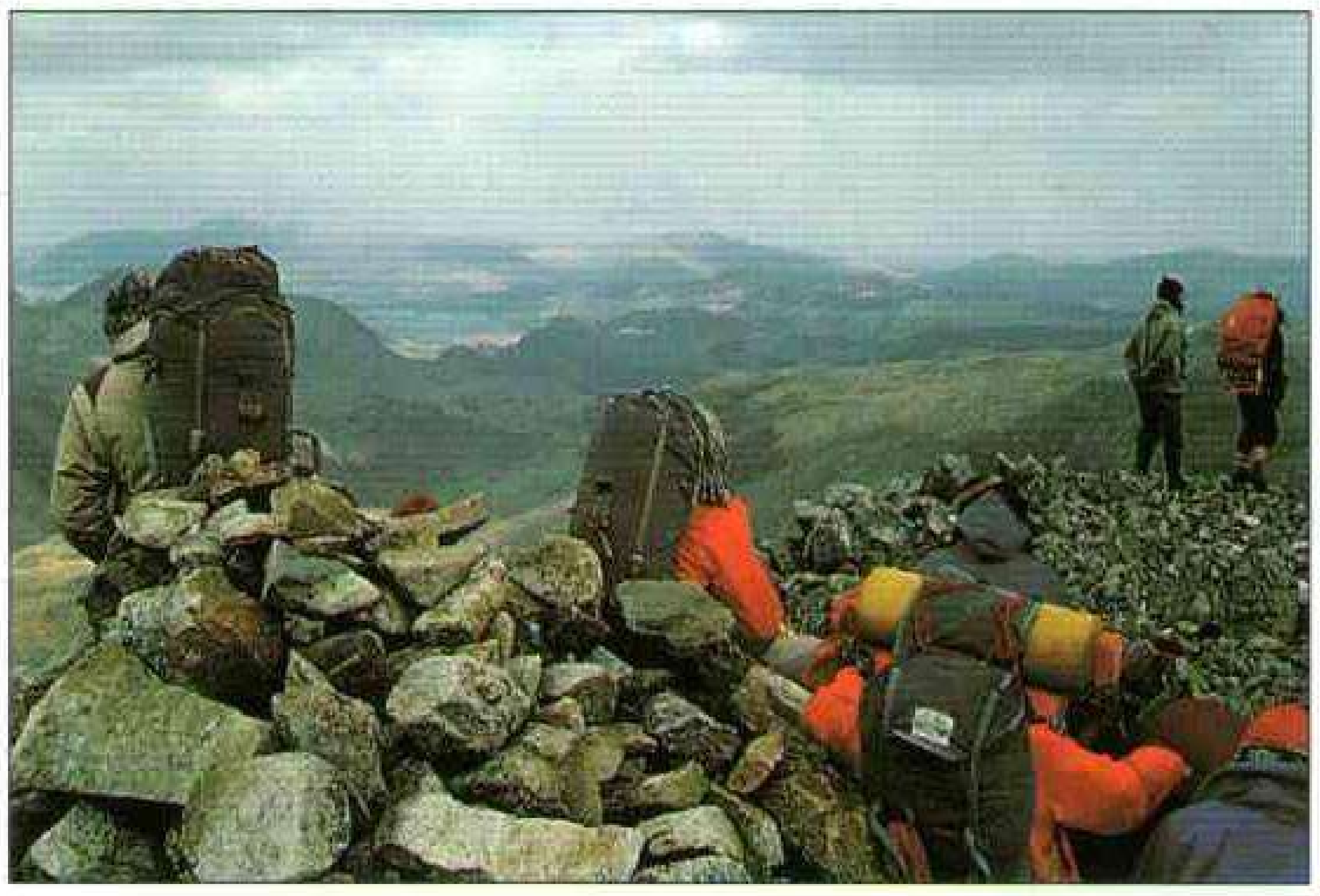

Parada de descanso. *Las paradas para descansar son esenciales, en particular si camina en grupo. Intente no prolongarlas más de 10 minutos en cada ocasión, pero no empiece a contabilizar los minutos hasta que todos los miembros hayan llegado y estén sentados. Mientras el grupo descansa, la avanzadilla puede ir comprobando la ruta.*

LOS PELIGROS DE LA MARCHA

Existen muchos peligros al caminar en la naturaleza, que varían desde las caídas hasta las mordeduras de insectos y serpientes. Por ejemplo, hay varias especies de mosca negra urticante en los bosques y tierras húmedas de Norteamérica. Las garrapatas también constituyen otro peligro, y pueden pasar accidentalmente de los arbustos a su ropa *(véase página 180)*. Tendrá que adaptar su estilo de marcha al tipo de terreno que atraviese para evitar lesionarse *(véase página 145)*. Otro peligro es el agotamiento por calor. Asegúrese de que su ropa permite la transpiración para refrescarse sin pasar frío *(véase página 23)*.

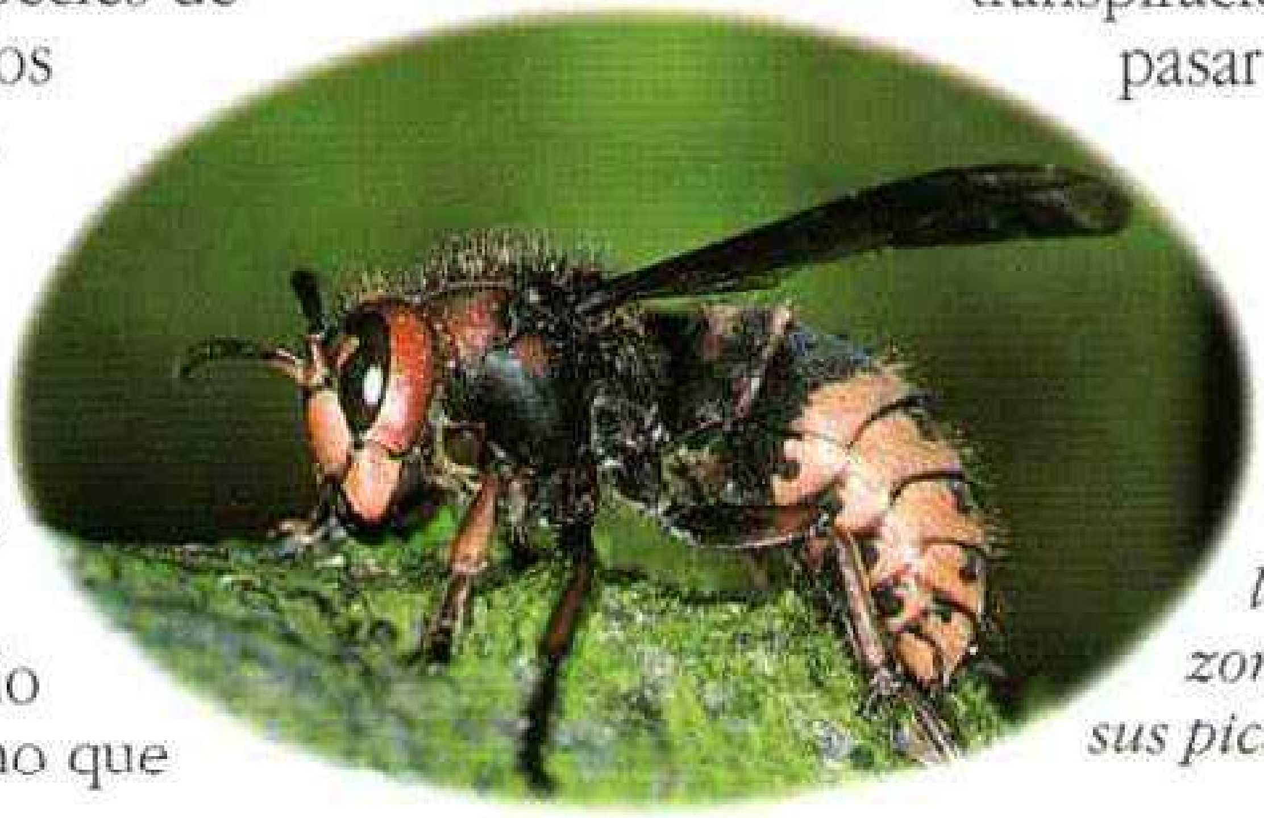

Avispón. *Varias especies de avispones* (Vespa) *pueden constituir una amenaza para los caminantes en las zonas templadas, debido a sus picaduras muy dolorosas.*

CAMINAR DE NOCHE

Camine lentamente durante la noche, comprobando a cada paso el estado del terreno, antes de apoyar todo el peso sobre el pie. Aproveche la luz de la luna y las estrellas, y camine por terrenos abiertos, alejados de los árboles. Agáchese y mire hacia arriba para comprobar el contorno del camino contra el cielo. Mire hacia los lados de los objetos y no directamente a los mismos. Deténgase a intervalos regulares, permaneciendo completamente en silencio, y escuche con atención.

Conserve la visión nocturna. *Nuestros ojos necesitan unos 30 minutos o más para adaptarse a la oscuridad. Para mirar el mapa y la brújula con la linterna, cierre un ojo. De esta manera, recobrará más rápidamente la visión.*

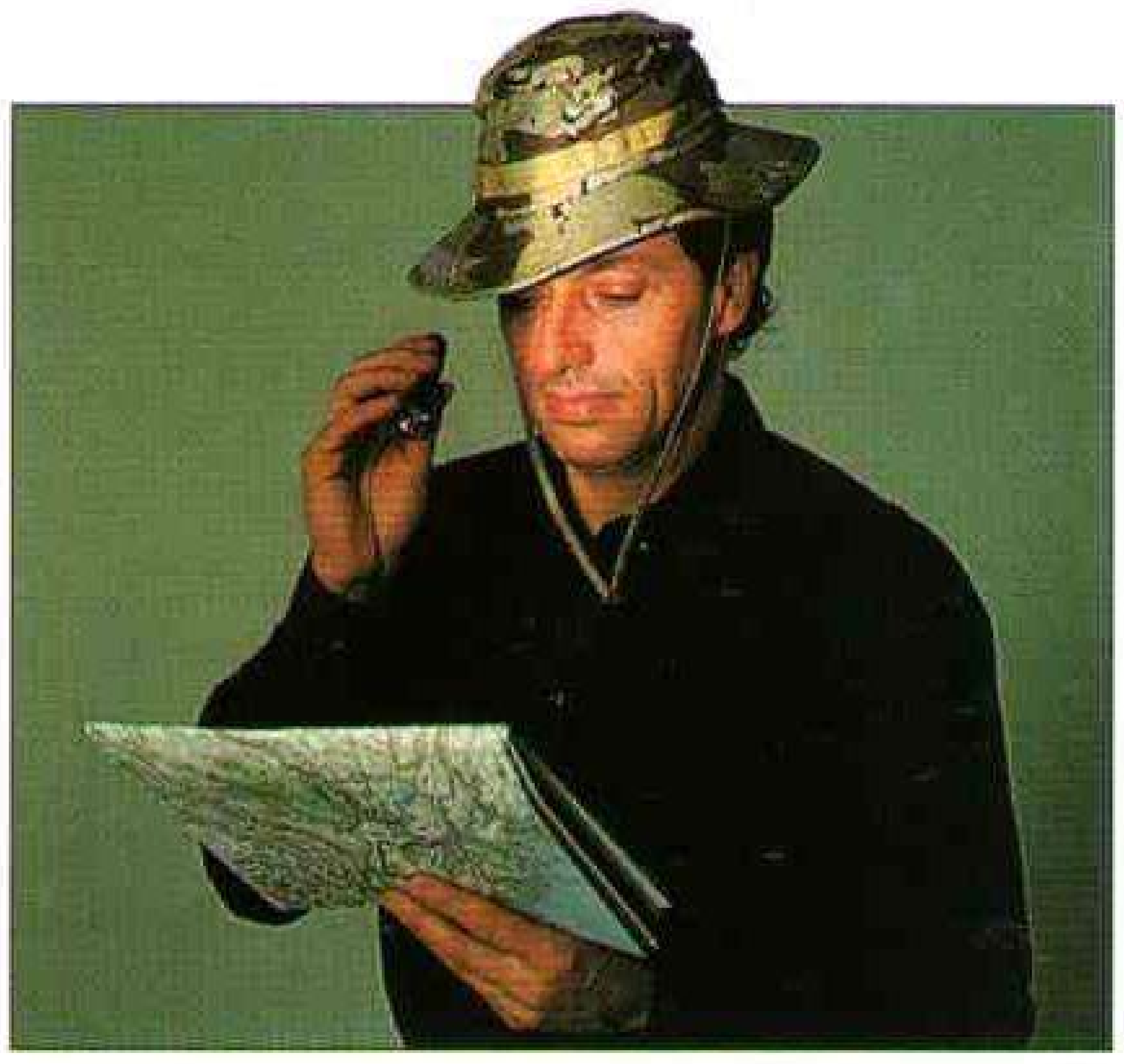

MOCHILA «HUDSON BAY»

1 Si no dispone de una mochila, puede llevar su equipo en esta mochila improvisada. En primer lugar, envuelva una piedra en cada una de las esquinas de una lona o manta impermeable para proporcionar un buen punto de sujeción.

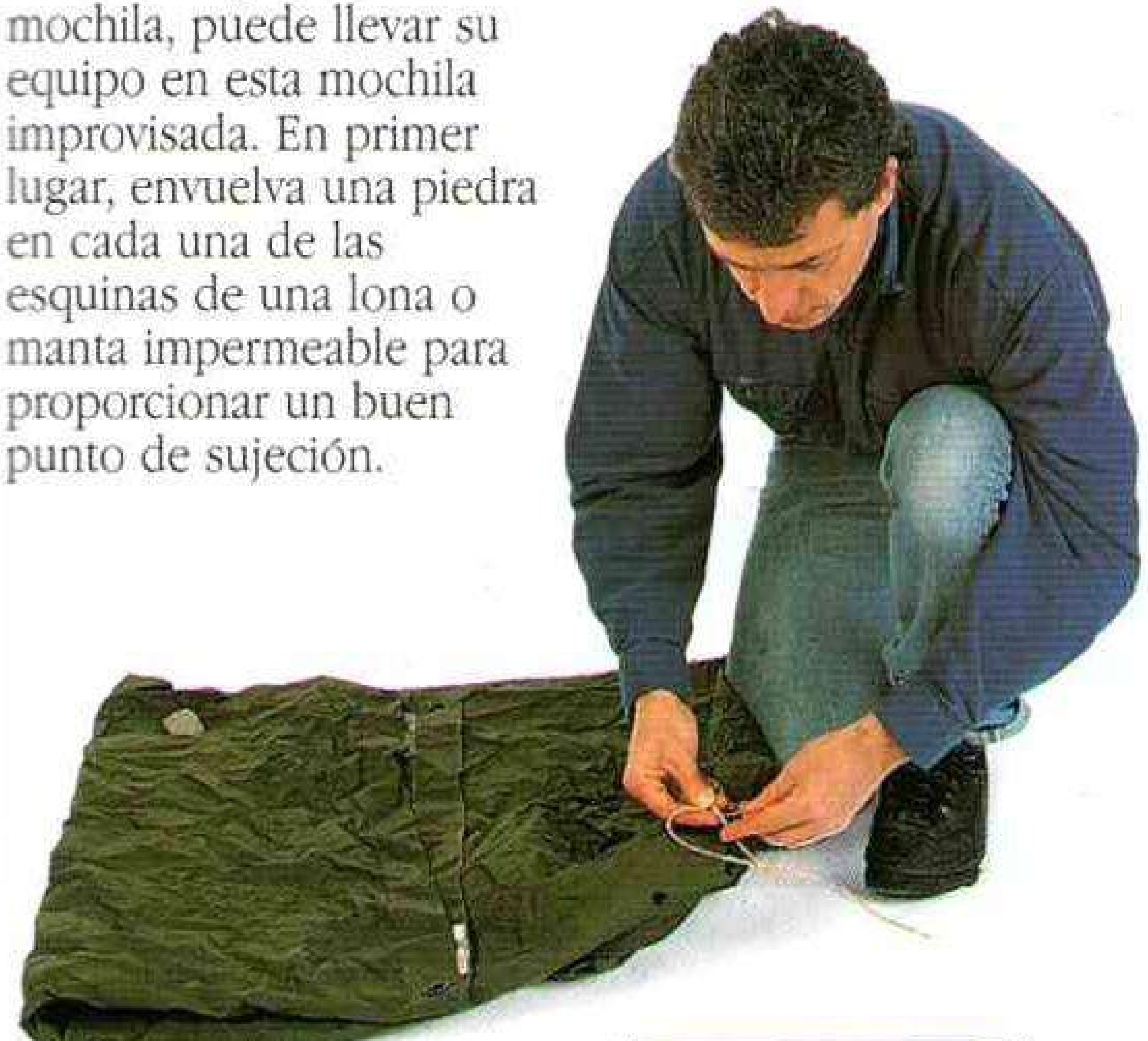

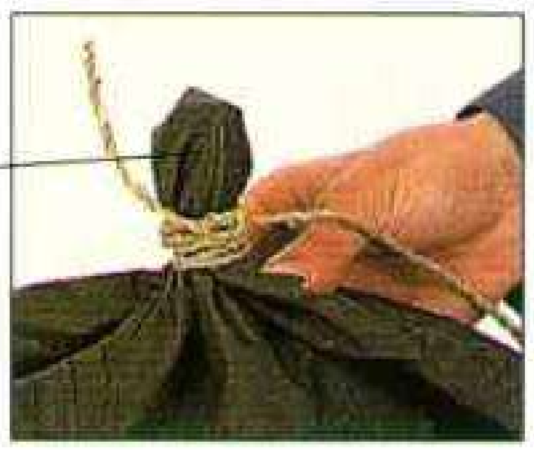

Ate una esquina de la lona alrededor de una piedra

2 Extienda su equipo sobre la lona y enróllela con la ropa dentro haciendo un hatillo en forma de salchicha.

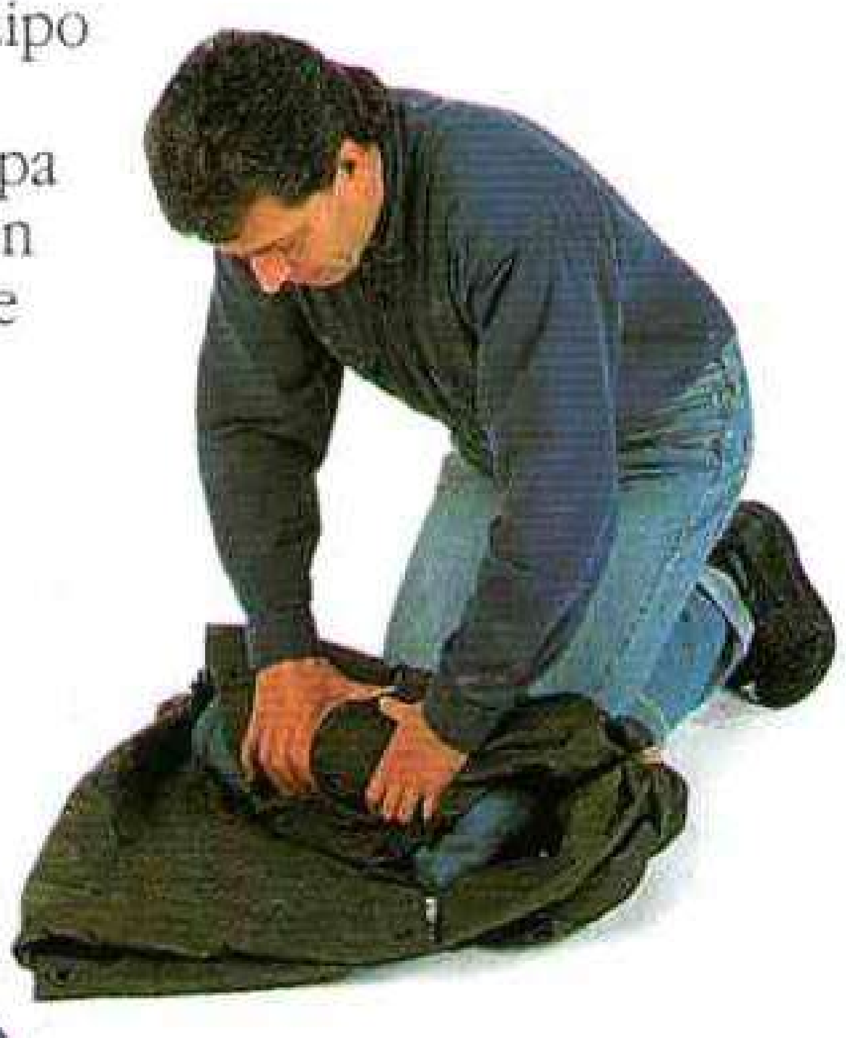

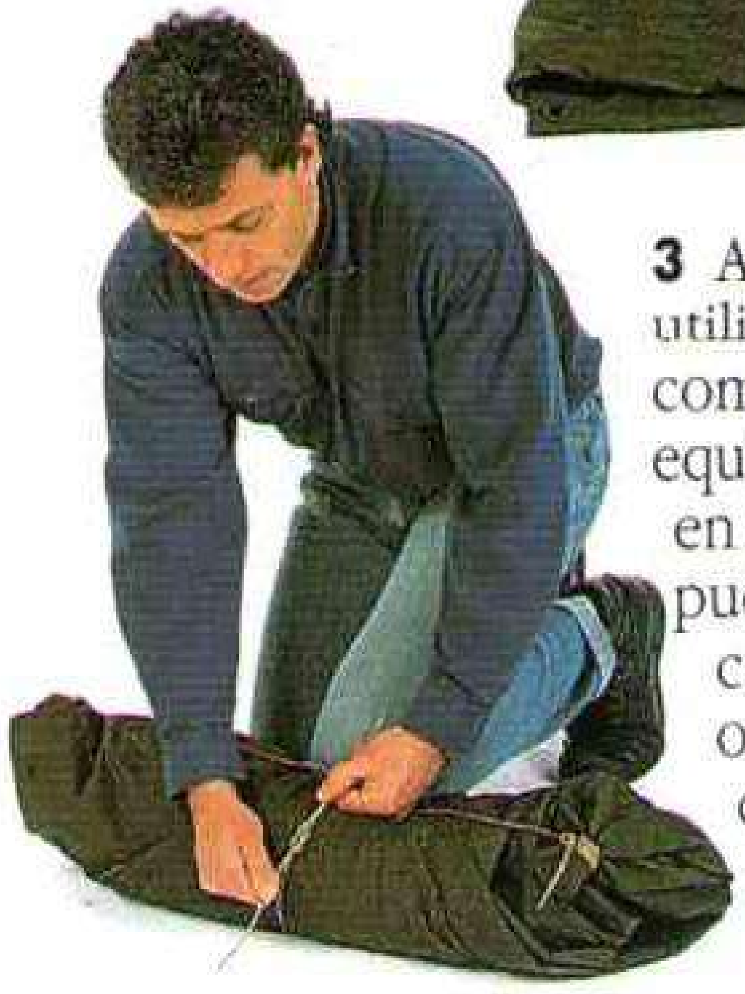

3 Ate el hatillo con fuerza utilizando una cuerda; compruebe que todo su equipo esté bien enrollado en el interior y que no pueda perderlo. Ate una cuerda de un extremo al otro, para que haga de correa.

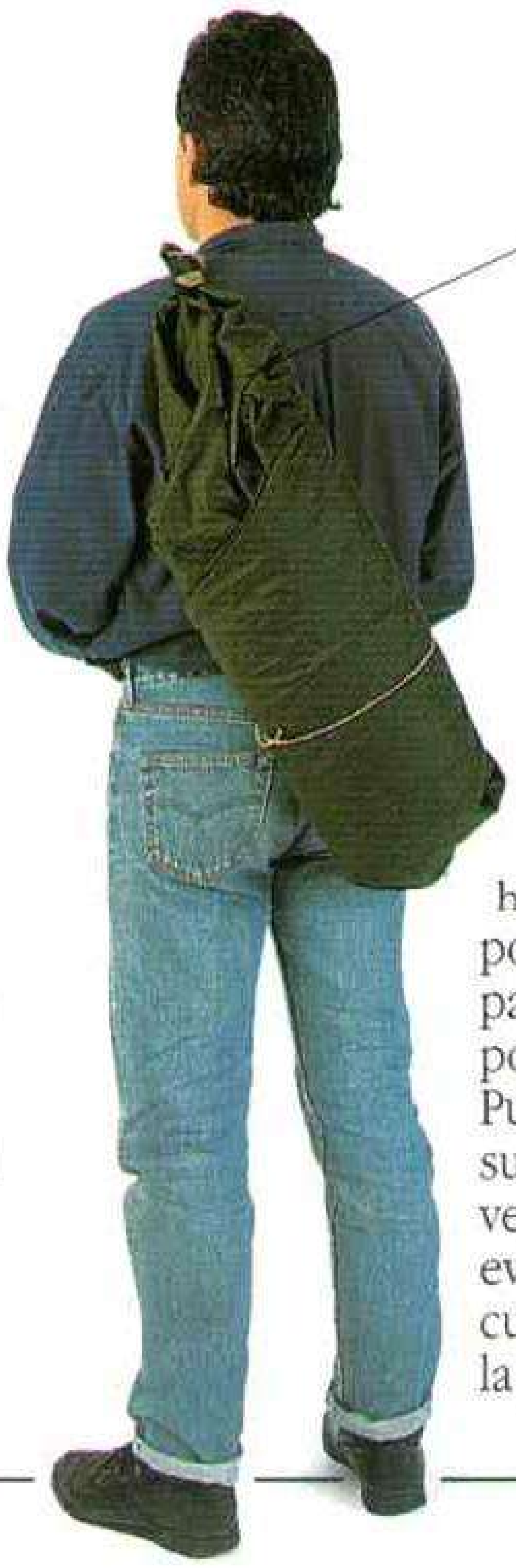

El hatillo se lleva cruzado sobre la espalda, como una mochila

4 Cuelgue el hatillo cruzándolo por su espalda, pasando la cuerda por el pecho. Puede acolchar su hombro con vegetación para evitar que la cuerda le roce la piel.

CAMINAR SOBRE HIELO Y NIEVE

LOS PELIGROS DE caminar sobre el hielo y la nieve radican tanto en la naturaleza del terreno como en el efecto del frío sobre el cuerpo humano. En las regiones polares, mantener el calor no es tanto un problema como el exceso de calor seguido de sudor al trabajar o viajar. Debe tener una excelente condición física para recorrer un territorio polar, además de disponer del equipo adecuado, sin el cual no podría sobrevivir durante mucho tiempo. Caminar sobre el hielo y la nieve puede ser muy peligroso en las áreas templadas, y no debe intentarlo sin una planificación apropiada.

INDUMENTARIA ADECUADA PARA LAS REGIONES POLARES

El intenso frío de las regiones polares requiere una indumentaria especial. Para caminar en estas zonas necesita ayudas especiales como las que le proporcionan las raquetas de nieve, los crampones o los esquís, mientras que para cruzar un área de hielo, como los glaciares, requerirá un piolet. Para evitar un sobrecalentamiento, debe ser capaz de ventilar su vestimenta, sea abriendo cremalleras o eliminando capas. Es fundamental seguir el sistema de capas *(véase página 22)* y llevar prendas resistentes al viento como capa exterior. Las temperaturas muy bajas y los vientos helados exigen cubrir incluso las mínimas áreas de piel expuestas al aire para evitar la congelación *(véase página 141)*. También son esenciales unas gafas para impedir la ceguera producida por el reflejo de la nieve.

EQUIPO INDISPENSABLE

En las regiones polares, unas gafas pueden evitar la ceguera producida por el reflejo de la nieve. Los crampones evitan que resbale sobre el hielo, mientras que las raquetas le permiten caminar sobre la nieve profunda. Un piolet para apoyarse en las pendientes heladas y unos palos de esquí le ayudan a mantener el equilibrio.

GAFAS

RAQUETA

PIOLET

PALO DE ESQUÍ

CRAMPÓN

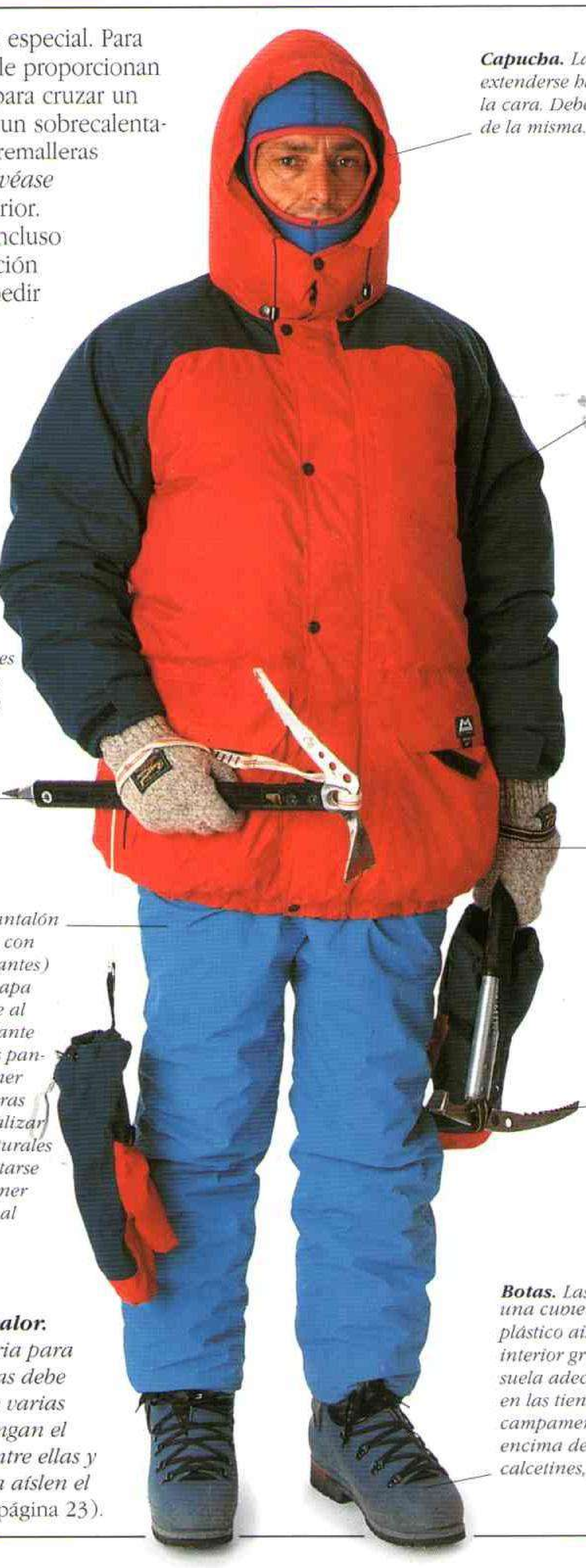

Capucha. *La capucha debe extenderse hacia adelante para cubrir la cara. Debe llevar un gorro debajo de la misma.*

Anorak. *Un anorak debe ser resistente al viento y fácil de ventilar, de la talla adecuada para permitir llevar muchas capas de ropa por debajo. Asimismo, debe ser aislante y de peso ligero.*

Piolet. *Un piolet es útil para construir escalones en las pendientes heladas. También puede utilizarlo como freno en caso de resbalar y caer.*

Guantes. *Son necesarios tres guantes: unos finos para evitar que la piel se adhiera al metal, unos intermedios de lana y unas manoplas exteriores resistentes al viento e impermeables. Los guantes deben sujetarse al cuerpo y entre sí por una cinta que suba por los brazos y pase por los hombros; así evitará perderlos.*

Pantalón. *Un pantalón o peto (pantalón con cintura alta y tirantes) debe tener una capa exterior resistente al viento y una aislante en el interior. Los pantalones deben tener cremalleras traseras para permitir realizar las funciones naturales sin tener que quitarse el anorak o exponer parte del cuerpo al aire frío.*

Martillo. *Se utiliza para clavar clavijas y asideros en el hielo al escalar* (véase página 146).

Mantener el calor. *La indumentaria para las regiones frías debe componerse de varias capas que retengan el aire caliente entre ellas y de esta manera aíslen el cuerpo* (véase página 23).

Botas. *Las botas constan de una cubierta exterior de plástico aislante y una bota interior gruesa, con una suela adecuada para el uso en las tiendas y en el campamento. Deben llevarse encima de varias capas de calcetines, finos y gruesos.*

RAQUETAS IMPROVISADAS

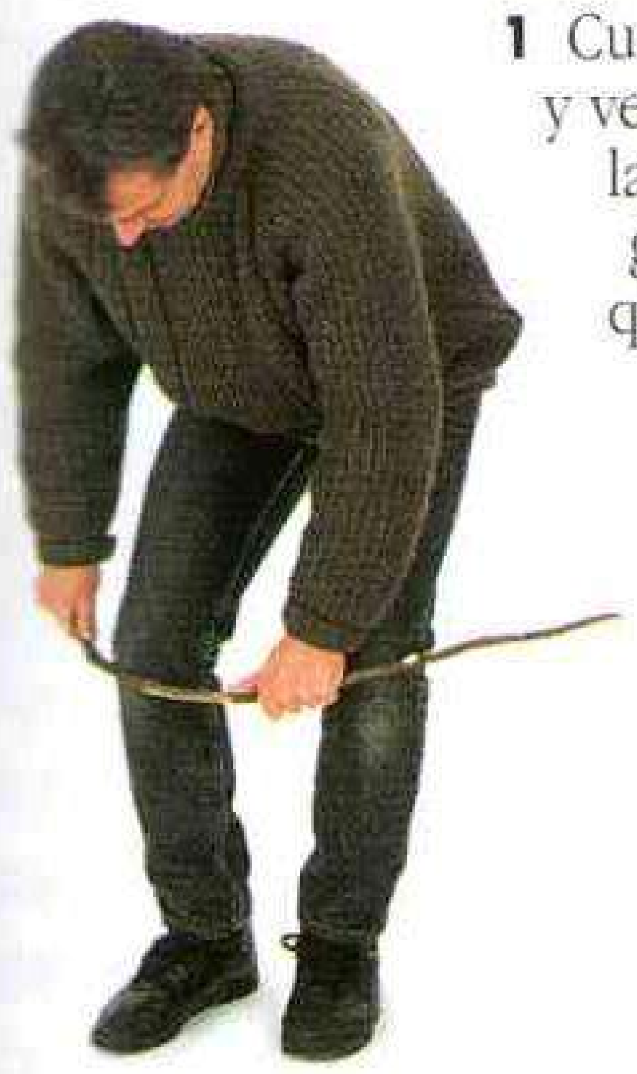

1 Curve una rama joven y verde ayudándose con la rodilla y dóblela gradualmente hasta que sea flexible.

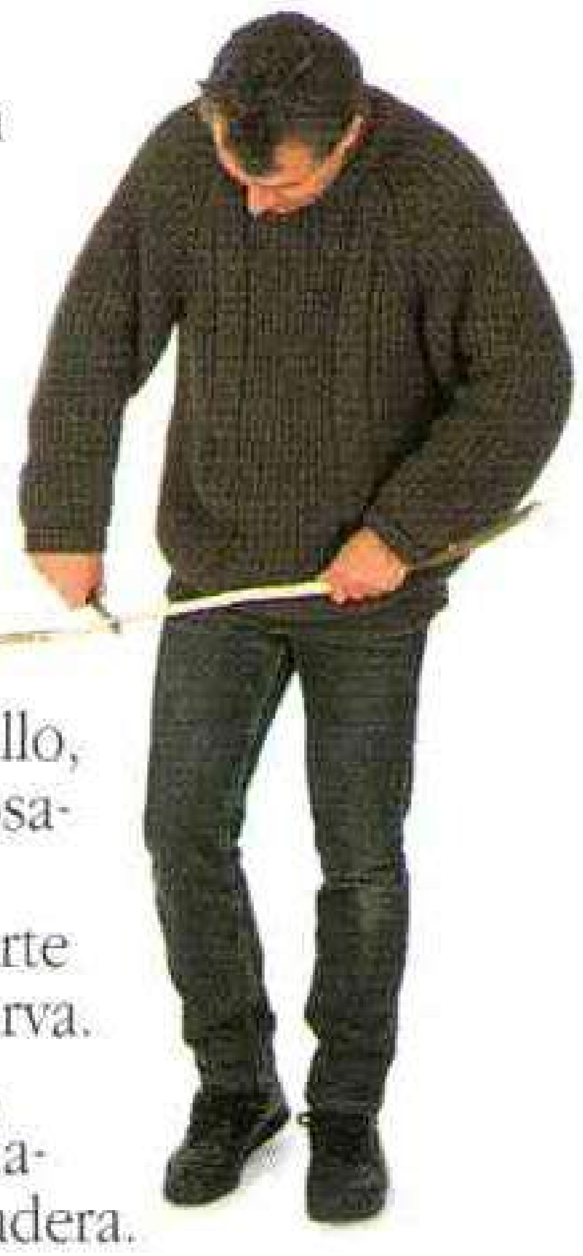

2 Con un cuchillo, elimine cuidadosamente toda la corteza en la parte interior de la curva. De esta manera, facilitará el doblamiento de la madera.

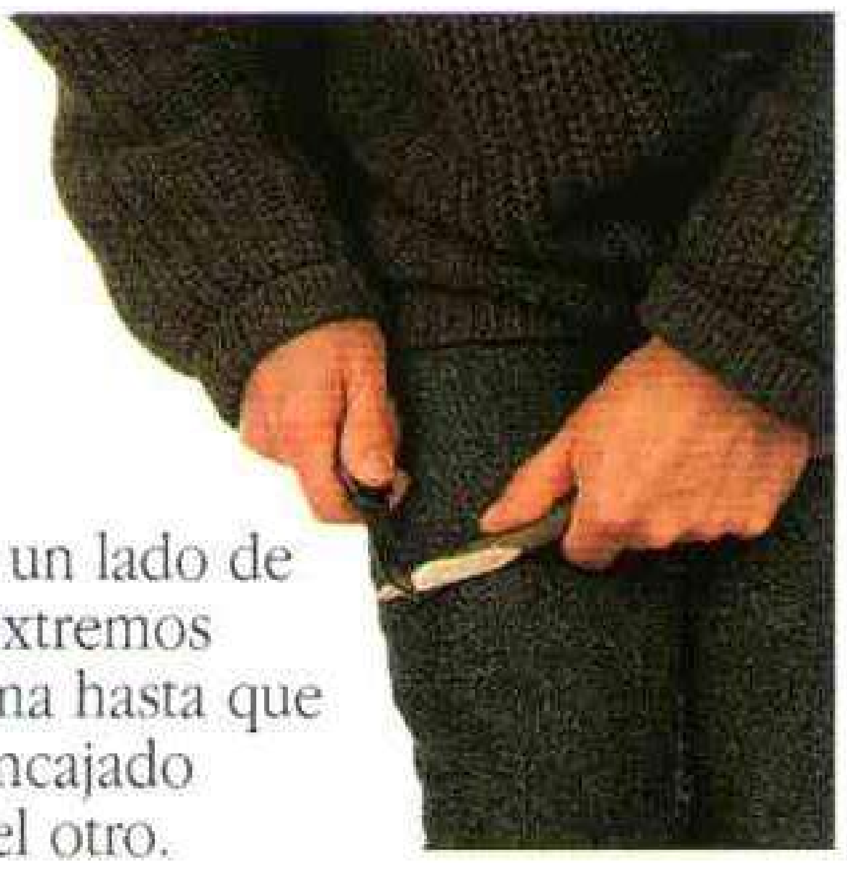

3 Corte un lado de ambos extremos de la rama hasta que quede encajado uno en el otro.

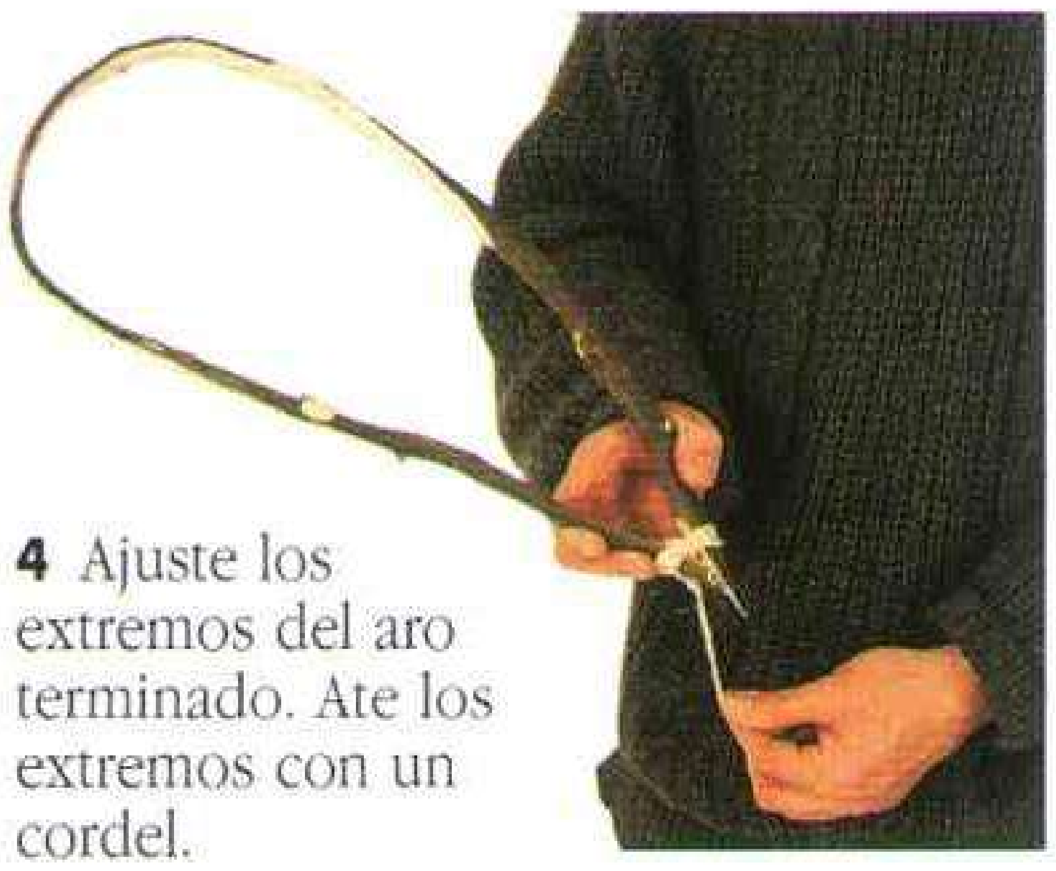

4 Ajuste los extremos del aro terminado. Ate los extremos con un cordel.

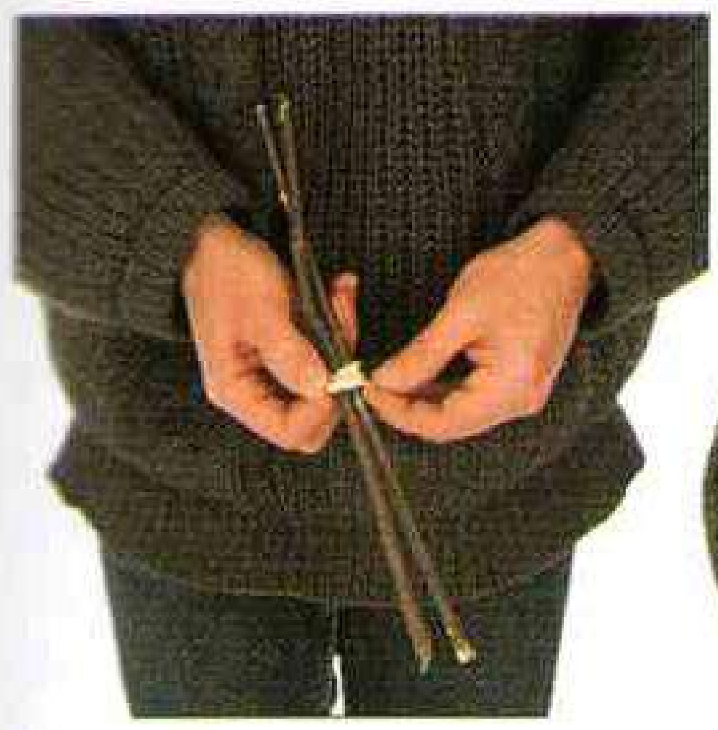

5 Ate dos palos cortos entre sí por el centro, con una cuerda. Repita la operación con dos pares más de palos.

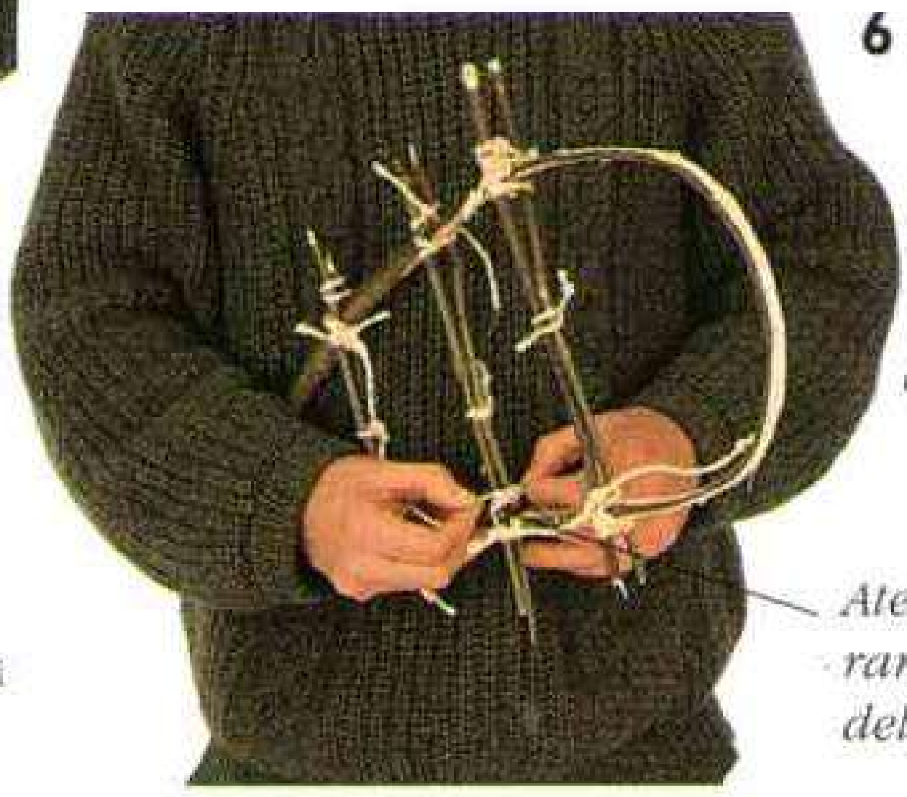

Ate cada par de ramas a lo ancho del aro

6 Ate tres pares de ramas pequeñas de un lado a otro del aro. Éstas servirán de soporte al pie.

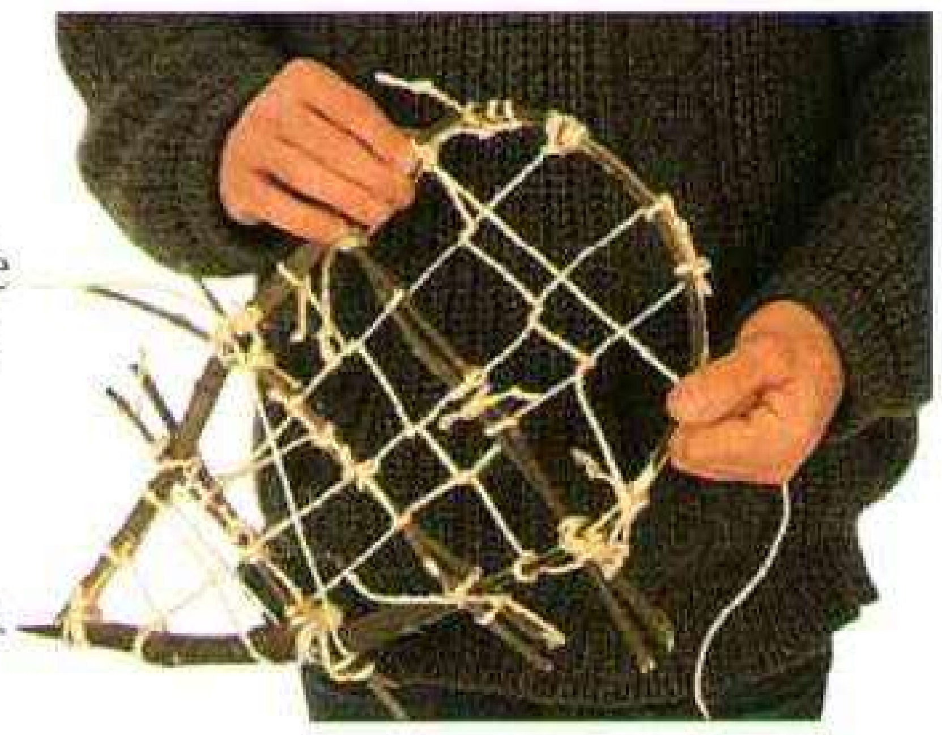

7 Entreteja un cordel entre las ramas y alrededor de las mismas para formar la base de la raqueta.

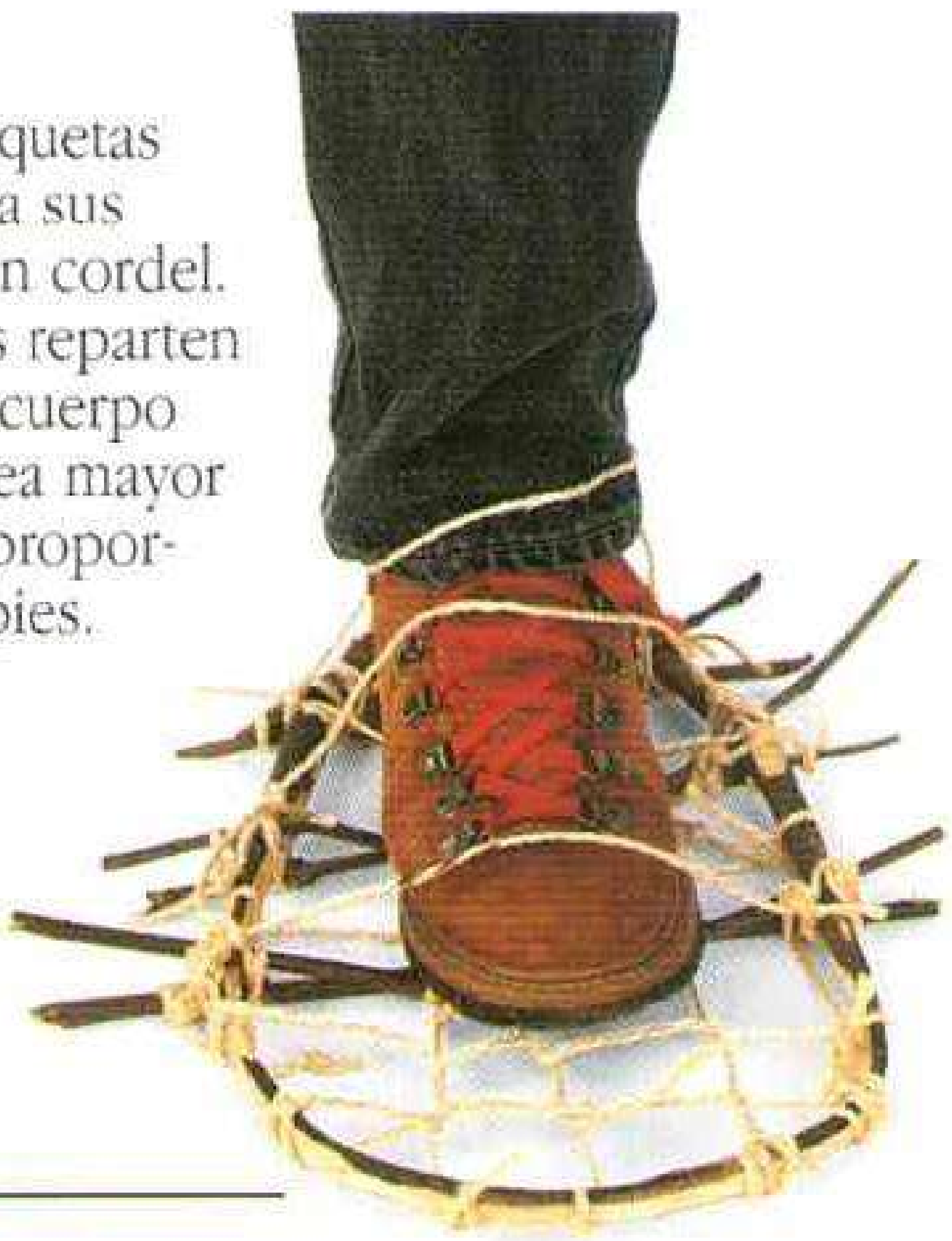

8 Ate las raquetas terminadas a sus botas con un cordel. Las raquetas reparten el peso del cuerpo sobre un área mayor que la que proporcionan los pies.

TÉCNICAS PARA CRUZAR HIELO Y NIEVE

Nieve profunda. *Para cruzar una gran extensión de terreno nevado en grupo, lo ideal es caminar en fila india, dejando poca distancia entre cada miembro del grupo. Esto evita que alguno se pierda en una ventisca, y permite que, en el caso de que alguien cayera, los demás pudieran socorrerle rápidamente. Una capa de hielo sobre la nieve profunda puede soportar su peso, pero tenga mucho cuidado. Utilice el piolet como soporte para caminar.*

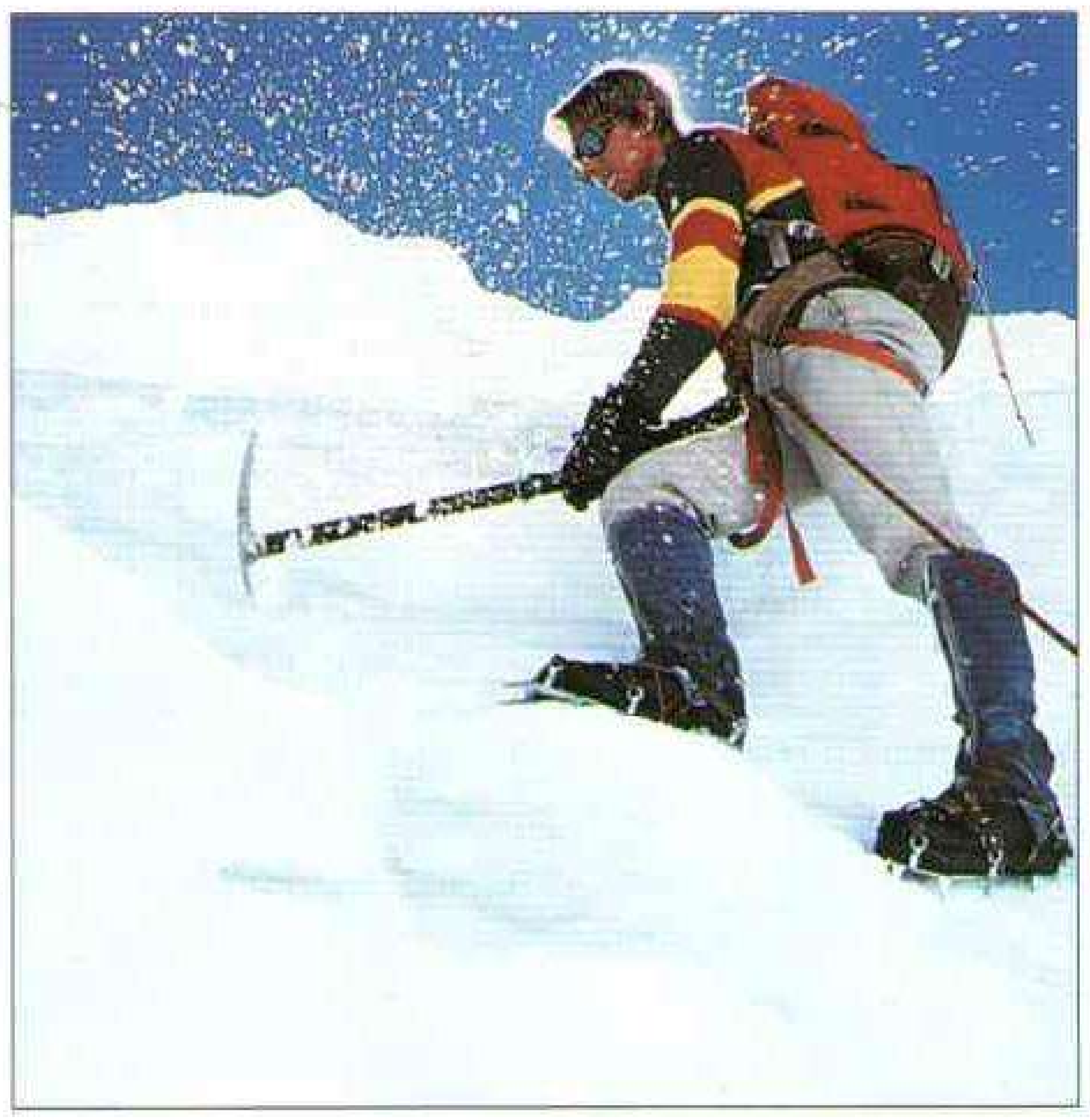

Hacer frente a las pendientes. *Cruzar pendientes heladas o lagos congelados requiere el mayor de los cuidados. Debe atarse a sus compañeros y emplear su piolet para tantear el camino por si encontrara grietas ocultas, o para sujetarse en caso de resbalar. Si el hielo se cuartea, reparta su peso en la mayor superficie posible, y utilice el piolet para empujarse fuera. No camine nunca encima de un embalse congelado. Si éste se ha vaciado, podría existir un hueco de varios metros debajo del hielo, y si el hielo se rompiera no podría volver a la superficie. Si resbala sobre una pendiente helada, clave la punta de su piolet en el terreno y apoye todo su peso sobre ella. Puede actuar como freno y detenerle en la caída por la pendiente.*

ESQUÍ DE FONDO

Los esquís empleados en el esquí de fondo son largos y estrechos. Están encerados para que se deslicen y tengan un buen agarre a la vez, permitiendo un movimiento frontal y evitando los deslizamientos hacia atrás. El impulso proviene del empuje de la pierna que queda atrás, mientras que la otra se desliza hacia delante. Los esquís de fondo son muy recomendables para desplazarse sobre la nieve en polvo muy profunda, donde resulta imposible caminar.

Deslizamiento silencioso. *Los esquís de fondo permiten desplazarse de forma eficaz y casi silenciosa en zonas silvestres.*

VIAJAR SOBRE NIEVE

CON PATINES Y esquís especiales, podrá desplazarse rápida y eficientemente sobre el hielo y la nieve profunda. Sin embargo, cruzar una extensión helada, o incluso con una pequeña cantidad de nieve, puede suponer un enorme esfuerzo. Únicamente las personas experimentadas y con un buen entrenamiento, y dotadas del equipo adecuado, deberían emprender viajes a las regiones polares, donde las bajas temperaturas embotan la mente, las noches invernales casi interminables deprimen a la persona más alegre, e incluso los vientos moderados aumentan drásticamente el riesgo de congelación.

Vehículos para la nieve. *Con fines deportivos o prácticos, las motonieves han reemplazado a los trineos de perros en la vida cotidiana de las regiones árticas.*

TRINEOS DE TRACCIÓN HUMANA

Los trineos de tracción humana son muy eficientes para su empleo sobre hielo o nieve helada, en especial si lleva esquís. Esquiar sobre nieve blanda requiere un mayor esfuerzo, y resulta más adecuado usar raquetas, aunque el esfuerzo sigue siendo considerable. Debe tener mucho cuidado de no acalorarse y sudar, porque el sudor mojaría su ropa. Asegúrese de que puede ventilar su ropa para liberar el exceso de calor.

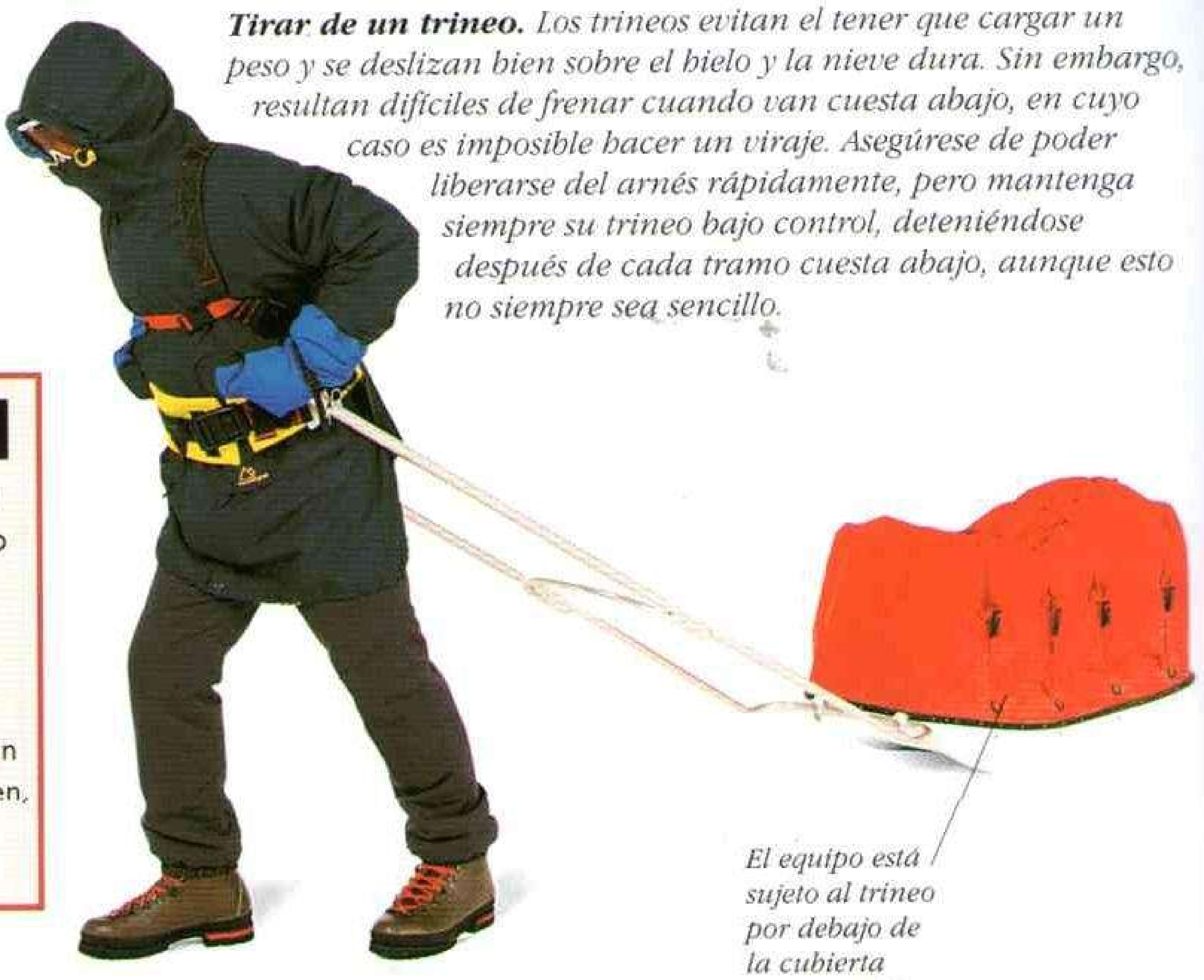

Tirar de un trineo. *Los trineos evitan el tener que cargar un peso y se deslizan bien sobre el hielo y la nieve dura. Sin embargo, resultan difíciles de frenar cuando van cuesta abajo, en cuyo caso es imposible hacer un viraje. Asegúrese de poder liberarse del arnés rápidamente, pero mantenga siempre su trineo bajo control, deteniéndose después de cada tramo cuesta abajo, aunque esto no siempre sea sencillo.*

El equipo está sujeto al trineo por debajo de la cubierta impermeable

ADVERTENCIA

Los dedos, la nariz, los pies, las orejas y la cara son las partes del cuerpo que primero se congelan. Cuando la piel comienza a congelarse da una sensación de hormigueo. Después aparecen unas zonas entumecidas y de aspecto ceroso. Si no vuelven a calentarse, estas zonas adquieren un aspecto granuloso, después se enrojecen, forman ampollas, y mueren antes de desprenderse totalmente.

TRINEO DE PERROS

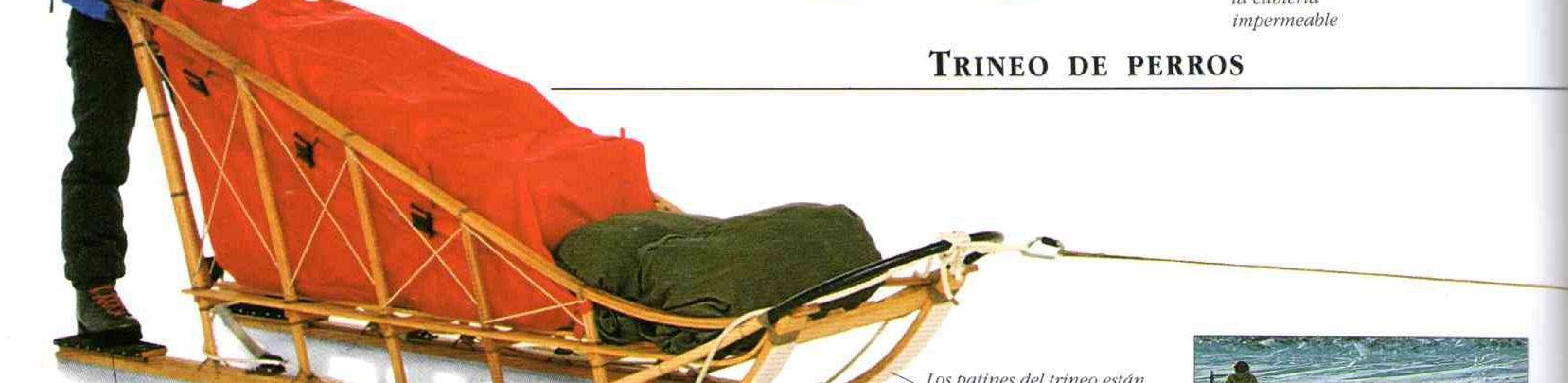

Los patines del trineo están revestidos de un plástico rígido que aumenta la fricción

El conductor maniobra el trineo desplazando su peso de un lado a otro

El trineo se sujeta con cuerdas para darle flexibilidad y resistencia

Transporte tradicional. *Los trineos de perros fueron utilizados tradicionalmente para el transporte polar, y algunos exploradores aún lo consideran el mejor método para el transporte de provisiones a través de los terrenos con hielo o nieve. Increíblemente resistentes, capaces de dormir en medio de una ventisca y de desplazarse rápidamente en casi todo tipo de terreno, los perros son más fáciles de mantener y son una compañía mucho más agradable que las motonieves.*

Trabajo en equipo. *El equipo trabaja en conjunto como si se tratara de una manada, siendo el perro guía su jefe. Éste considera al conductor el guía de la manada.*

EL EFECTO REFRIGERANTE DEL VIENTO

Velocidad	Temperatura (°C)														
Calma	-4	-7	-9	-12	-15	-18	-20	-23	-26	-29	-31	-34	-37	-40	-43
	Temperatura equivalente de congelación (°C)														
8 km/h	-7	-9	-12	-15	-18	-20	-23	-26	-29	-32	-34	-37	-40	-43	-45
16 km/h	-12	-15	-18	-23	-26	-29	-32	-37	-40	-43	-45	-51	-54	-57	-59
24 km/h	-18	-20	-23	-29	-32	-34	-40	-43	-45	-51	-54	-57	-62	-65	-68
32 km/h	-18	-23	-26	-32	-34	-37	-43	-45	-51	-54	-59	-62	-65	-70	-73
40 km/h	-20	-26	-29	-34	-37	-43	-45	-51	-54	-59	-62	-68	-70	-76	-79
48 km/h	-23	-29	-32	-34	-40	-45	-48	-54	-57	-62	-65	-70	-73	-79	-82
56 km/h	-23	-29	-34	-37	-40	-45	-51	-54	-59	-62	-68	-73	-76	-82	-84
64 km/h	-26	-29	-34	-37	-43	-48	-51	-57	-59	-65	-70	-73	-79	-82	-87

🟧 Las condiciones son relativamente agradables si se toman las precauciones normales

🟨 Las condiciones son muy desagradables; se requiere una vestimenta exterior térmica

⬜ La piel comienza a congelarse cuando se expone al aire libre durante un período prolongado

🟦 Un desplazamiento al aire libre puede ser peligroso; la piel expuesta puede congelarse en un minuto

⬜ Condiciones extremadamente peligrosas; la piel expuesta puede congelarse en 30 segundos

Vientos fríos. *El viento puede llegar a matar en las regiones del norte, ya que puede enfriar el aire por debajo de la temperatura ambiental. Conforme aumenta la velocidad del viento, su efecto sobre la temperatura se multiplica.*

CÓMO CONSTRUIR UN TRINEO

1 Puede construir un trineo a partir de una rama de árbol ahorquillada. Ate los extremos de la rama a la curva que forma la horquilla para tensarla.

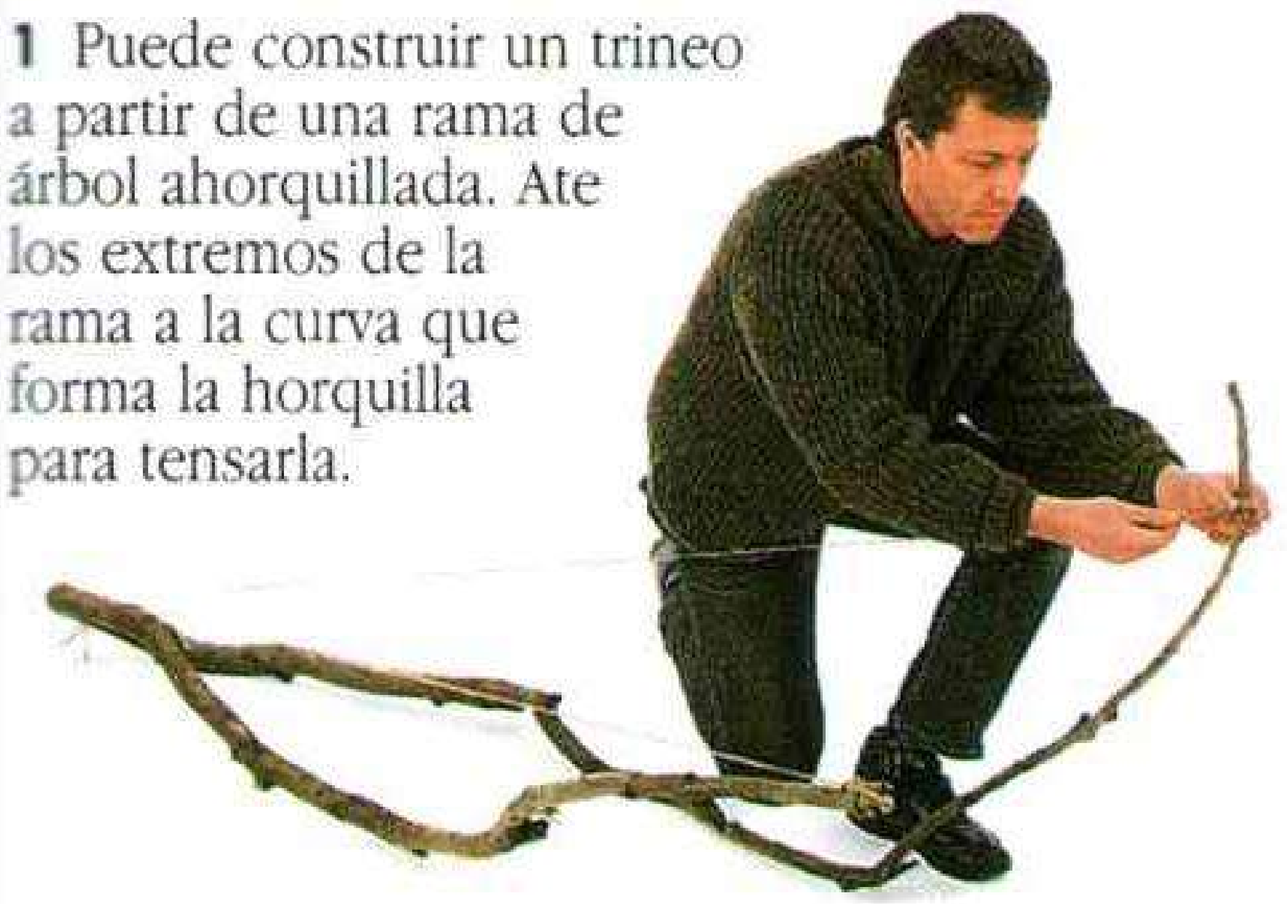

2 Cualquier cuerda que toque el suelo se desgastará rápidamente, por lo que debe atar ramas de refuerzo a los patines para construir un soporte para la plataforma.

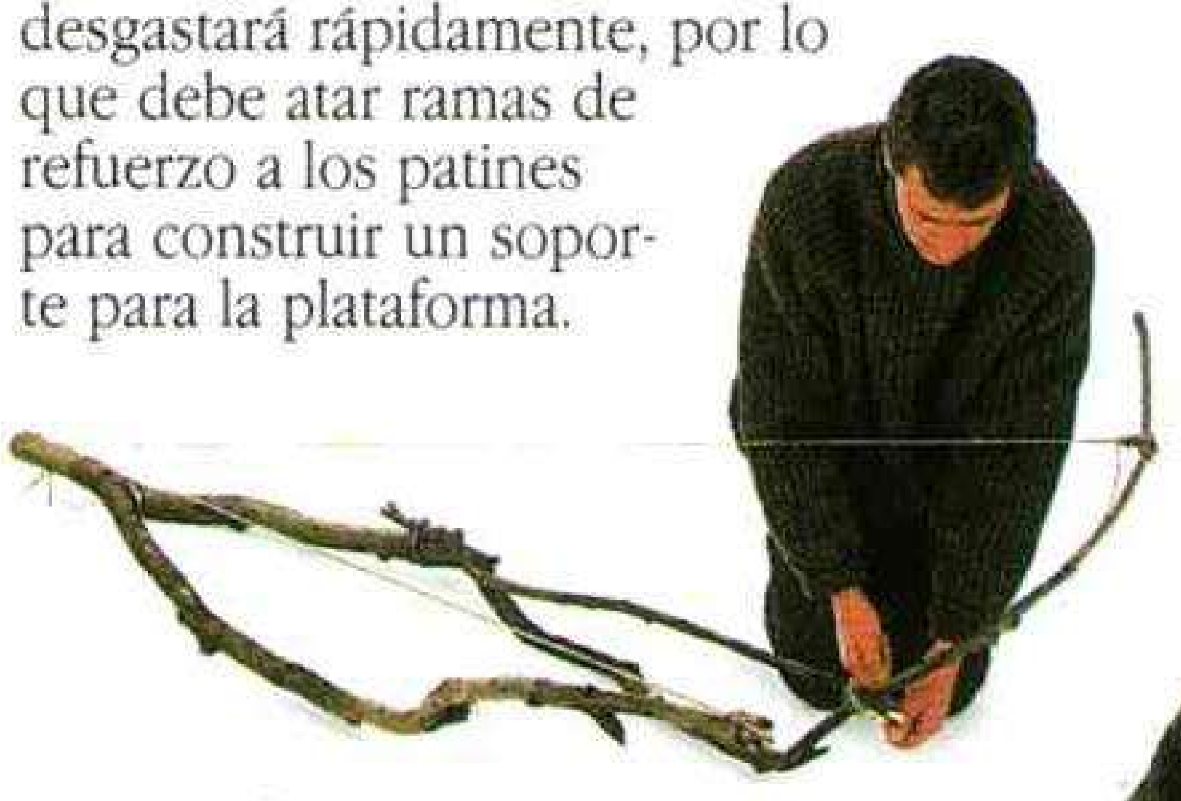

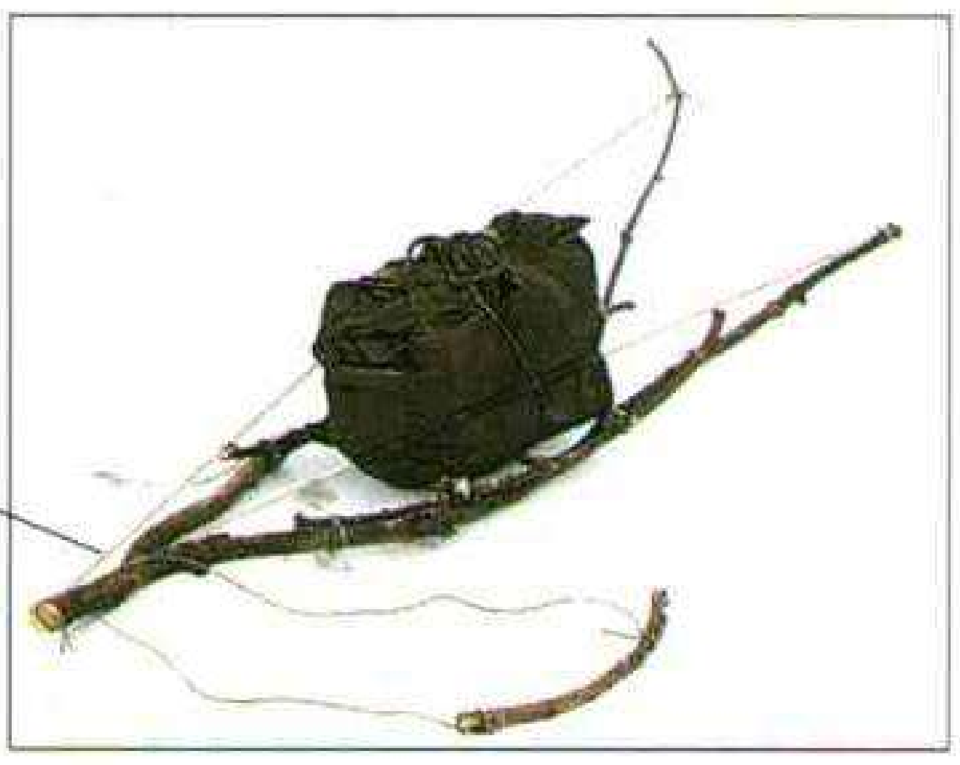

Construya un asa con una rama y una cuerda

Trineo cargado. *Envuelva su equipo en una manta o tela impermeable y átelo con fuerza al trineo.*

3 Ate algunas ramas sobre las piezas de refuerzo para formar la plataforma principal del trineo.

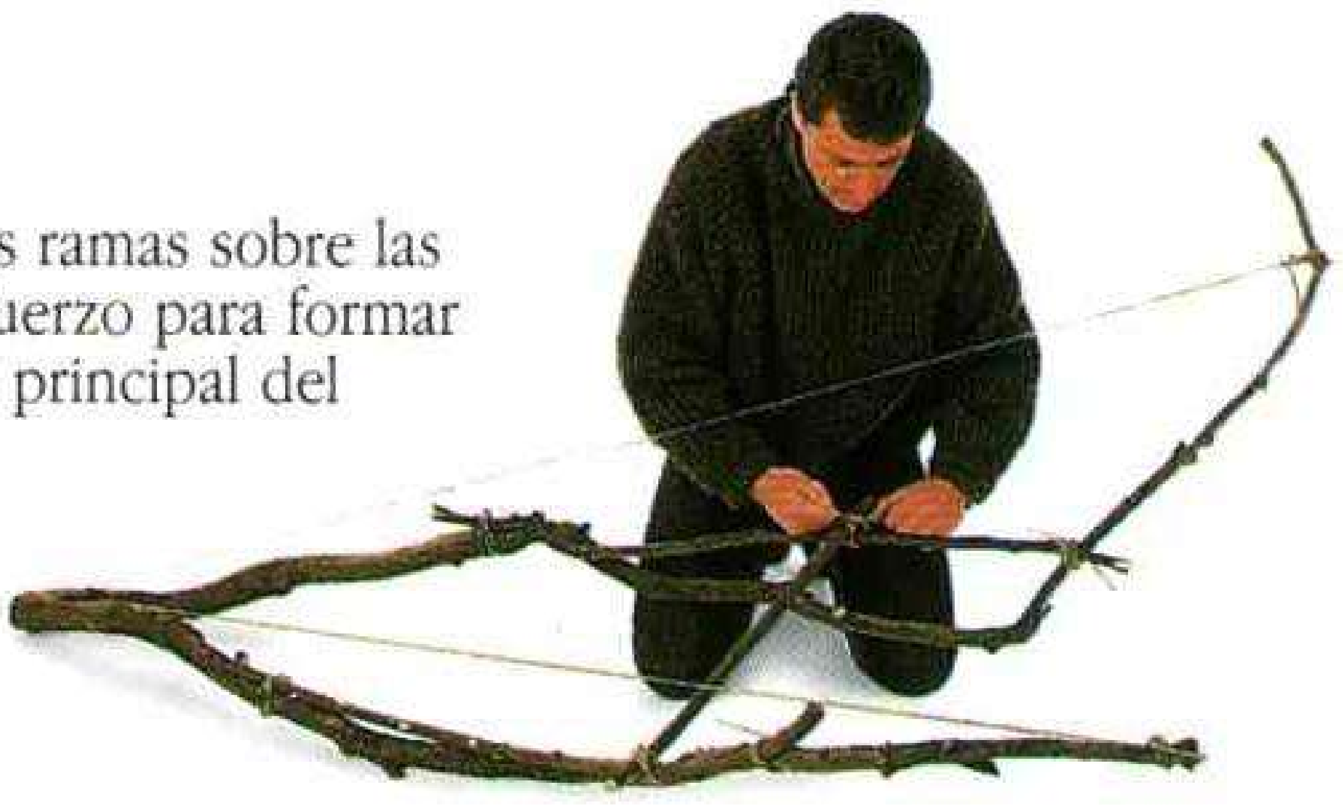

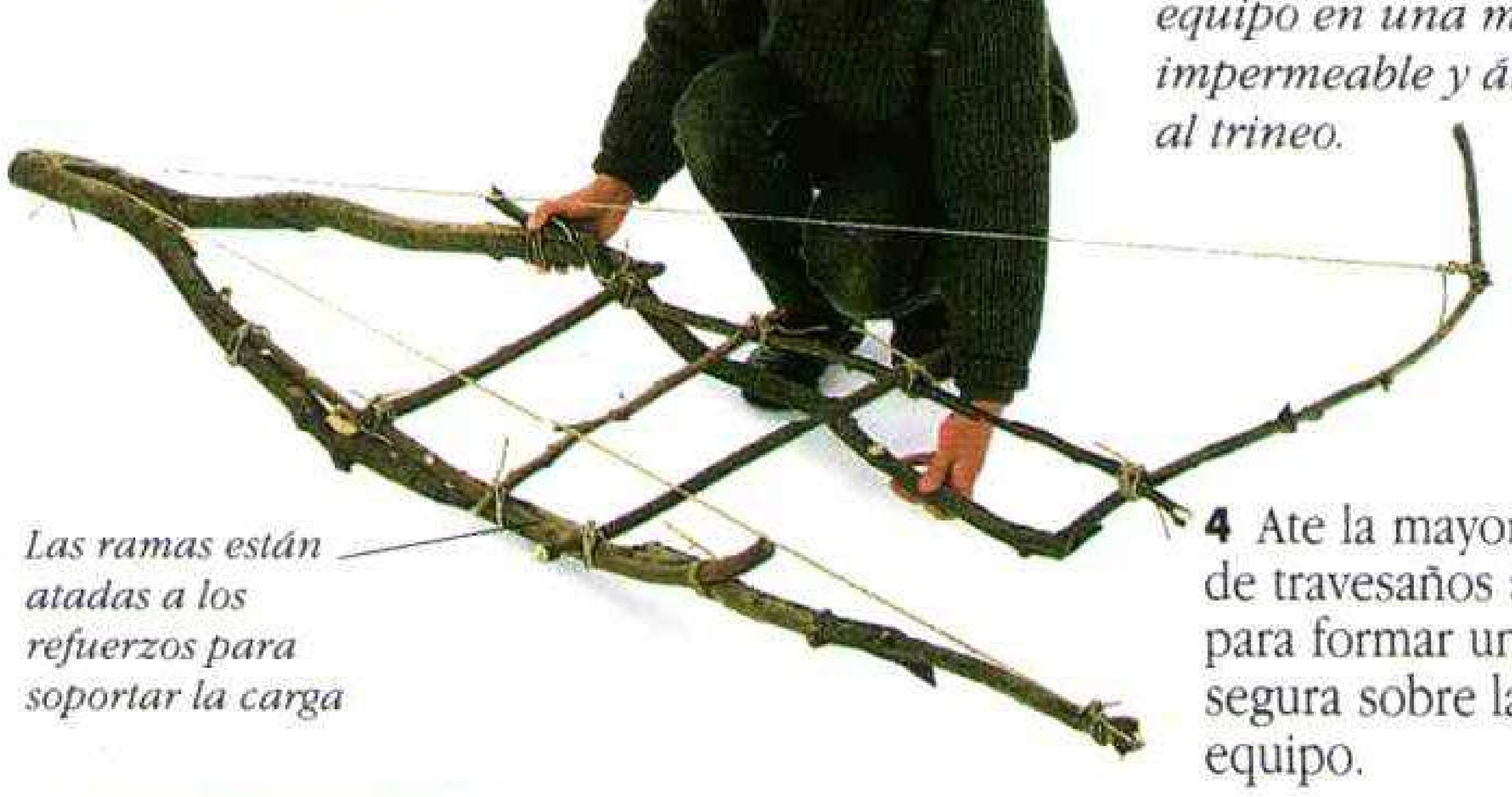

Las ramas están atadas a los refuerzos para soportar la carga

4 Ate la mayor cantidad posible de travesaños a los refuerzos para formar una plataforma segura sobre la cual sujetar su equipo.

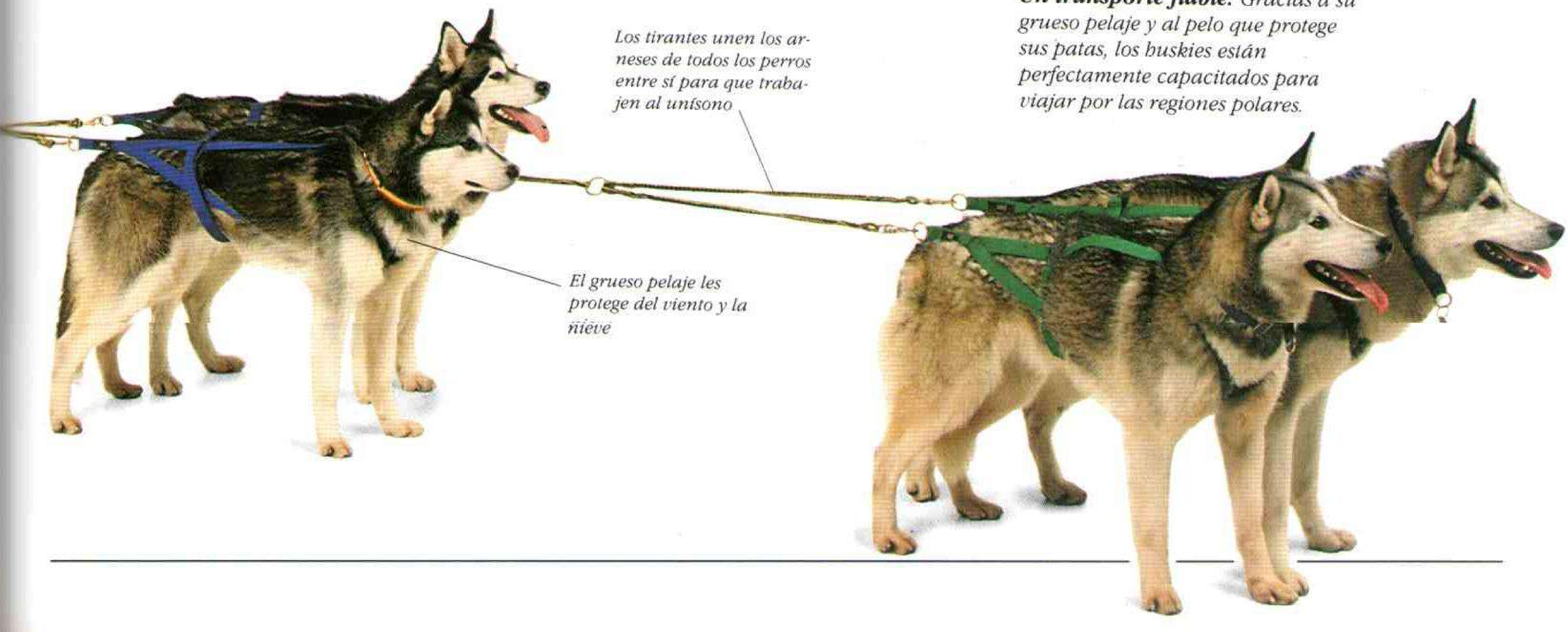

Los tirantes unen los arneses de todos los perros entre sí para que trabajen al unísono

El grueso pelaje les protege del viento y la nieve

Un transporte fiable. *Gracias a su grueso pelaje y al pelo que protege sus patas, los huskies están perfectamente capacitados para viajar por las regiones polares.*

CRUZAR EL DESIERTO

LOS DESIERTOS ESTÁN generalmente deshabitados porque no existe la suficiente cantidad de agua ni de vegetación para el sustento de una comunidad permanente. Son lugares con climas muy extremos: a menudo hace mucho calor durante el día y el frío es extremo por la noche. Las condiciones climáticas también pueden variar rápidamente debido a las violentas tormentas de arena y a las lluvias torrenciales. El desierto es un medio inhóspito, en el que el agua es una necesidad esencial y en el que únicamente sobreviven las criaturas más resistentes.

INDUMENTARIA ADECUADA PARA EL DESIERTO

El objetivo principal de la indumentaria para el desierto es mantener constante la temperatura corporal. La ropa debe proteger al cuerpo del calor del sol durante el día, y retener el calor corporal durante la noche, cuando las temperaturas bajan drásticamente. Todas las partes del cuerpo deben estar cubiertas y protegidas del sol y de la incomodidad que supone una tormenta de arena. En los desiertos no hay prácticamente humedad; esto hace que el calor sea más soportable que en otras regiones cálidas. Como la ropa para el desierto no debe protegerle de la lluvia, puede ser cómodamente porosa, permitiendo la ventilación entre las capas, y proporcionándole así un buen aislante durante la noche.

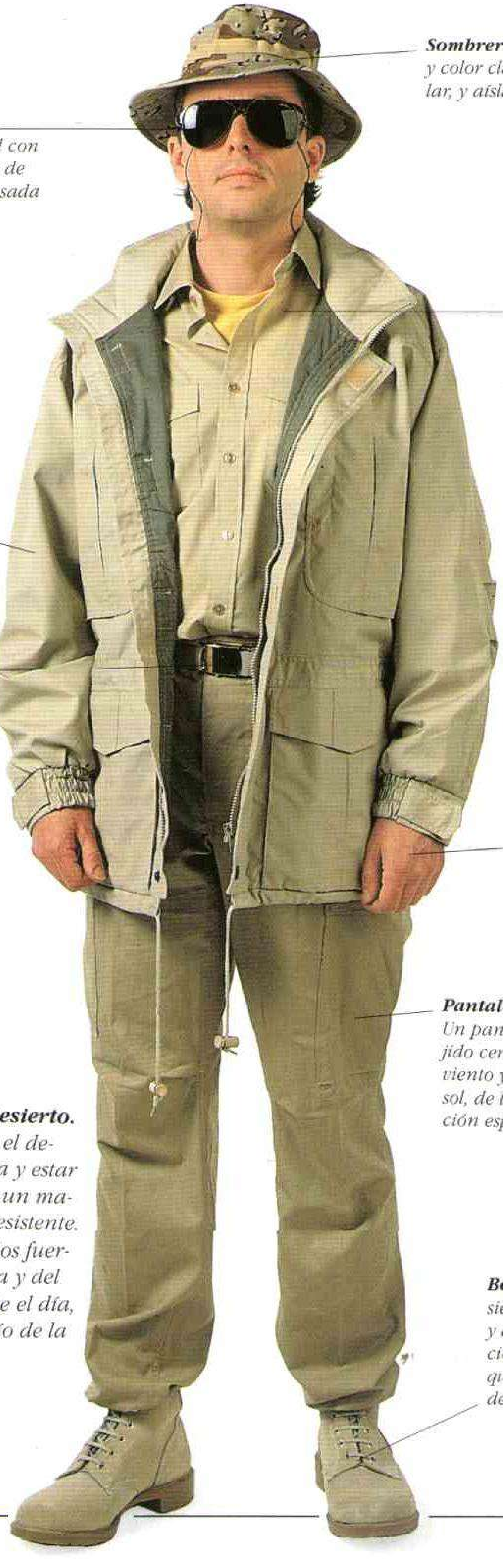

Sombrero. *Un sombrero de ala ancha y color claro refleja la caliente luz solar, y aísla la cabeza del frío nocturno.*

Gafas de sol. *Las gafas de sol con filtros protegen de la ceguera causada por el sol.*

Camisa de algodón. *Una camisa ligera de algodón debe ser de manga larga para protegerle los brazos del sol. Debajo de la camisa debe llevar una camiseta que controlará la evaporación del sudor y ayudará a mantenerle fresco.*

Chaqueta. *Una chaqueta resistente al viento protege de la vegetación espinosa, del sol, del viento y del frío nocturno.*

Cinturón. *Un cinturón tejido o trenzado no sólo evitará que la arena entre en sus ropas, sino que también podrá utilizarlo como cabestrillo en caso de emergencia* (véase página 179).

Manos. *Proteja sus manos de las quemaduras solares con cremas protectoras con filtro solar.*

Pantalón de algodón. *Un pantalón de algodón de tejido cerrado es resistente al viento y protege las piernas del sol, de la arena y de la vegetación espinosa.*

La ropa para el desierto. *La vestimenta para el desierto debe ser ligera y estar confeccionada con un material poroso pero resistente. Debe protegerle de los fuertes rayos ultravioleta y del calor del sol durante el día, así como del aire frío de la noche.*

Botas. *Las botas para el desierto tienen suelas gruesas y aislantes, y están confeccionadas con ante poroso que permite la respiración de los pies.*

IMPROVISAR EL EQUIPO

Como el desierto puede ser un lugar muy peligroso, es importante contar con el equipo adecuado, en especial con ropa apropiada que le proteja del calor y de los rayos ultravioleta del sol. Puede improvisar unas gafas de sol a partir de película fotográfica, un pañuelo a partir de una toalla para proteger su cuello, o hacerse unas botas con tela.

Gafas. *Debe proteger sus ojos del sol en todo momento. Improvise unas gafas cortando ranuras muy finas en una tira de película de fotografía, que atará alrededor de su cabeza con la ayuda de un cordel.*

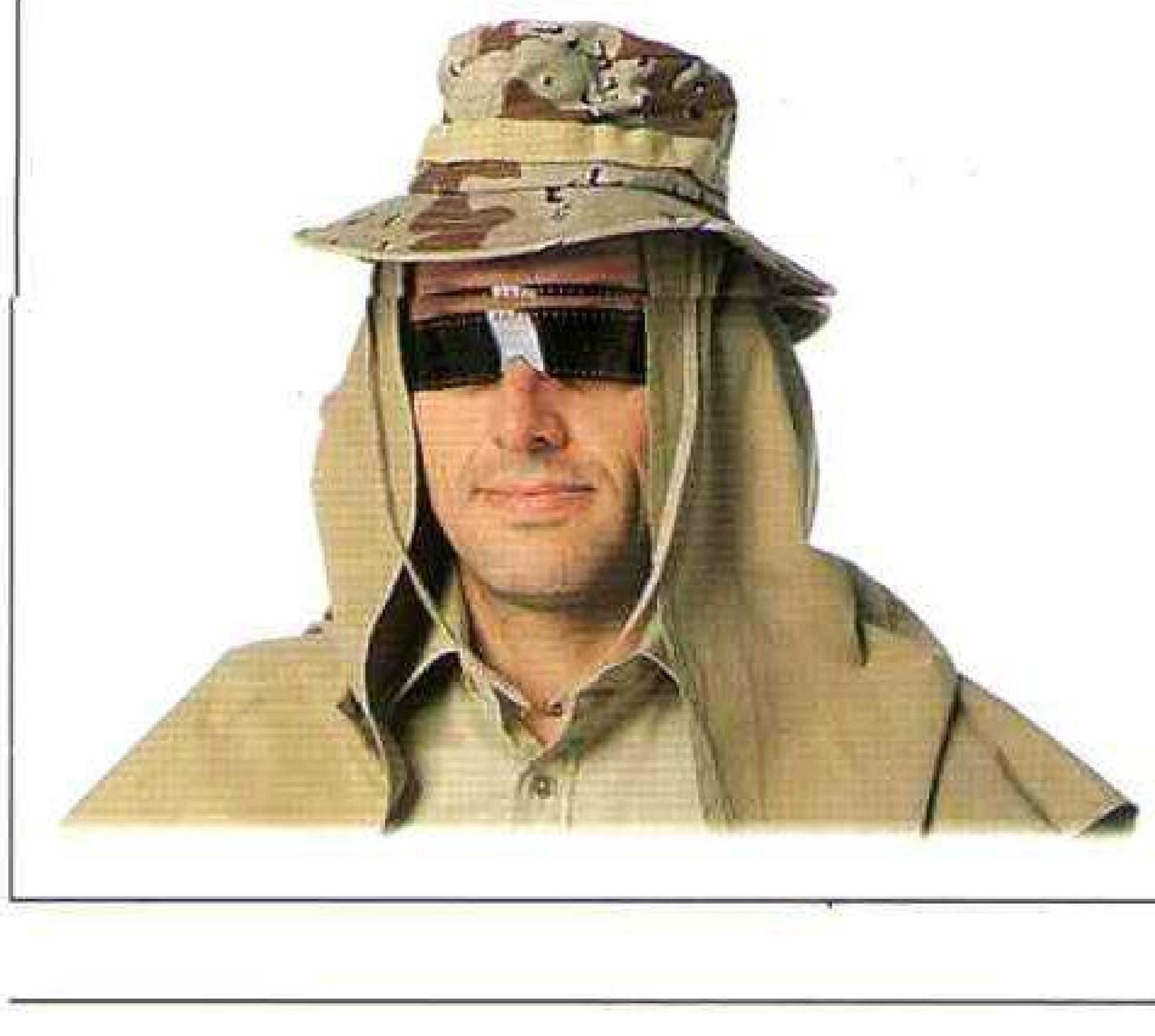

TIPOS DE DESIERTO

Una zona se clasifica como desierto si recibe menos de 25 cm de lluvia al año. Los desiertos pueden ser muy diferentes, según su localización y clima *(véase página 16)*. Pueden ser áreas rocosas, enormes extensiones de dunas de arena, montañas rocosas con matorrales o simplemente una tierra desolada que no puede soportar una vida animal o vegetal especializada.

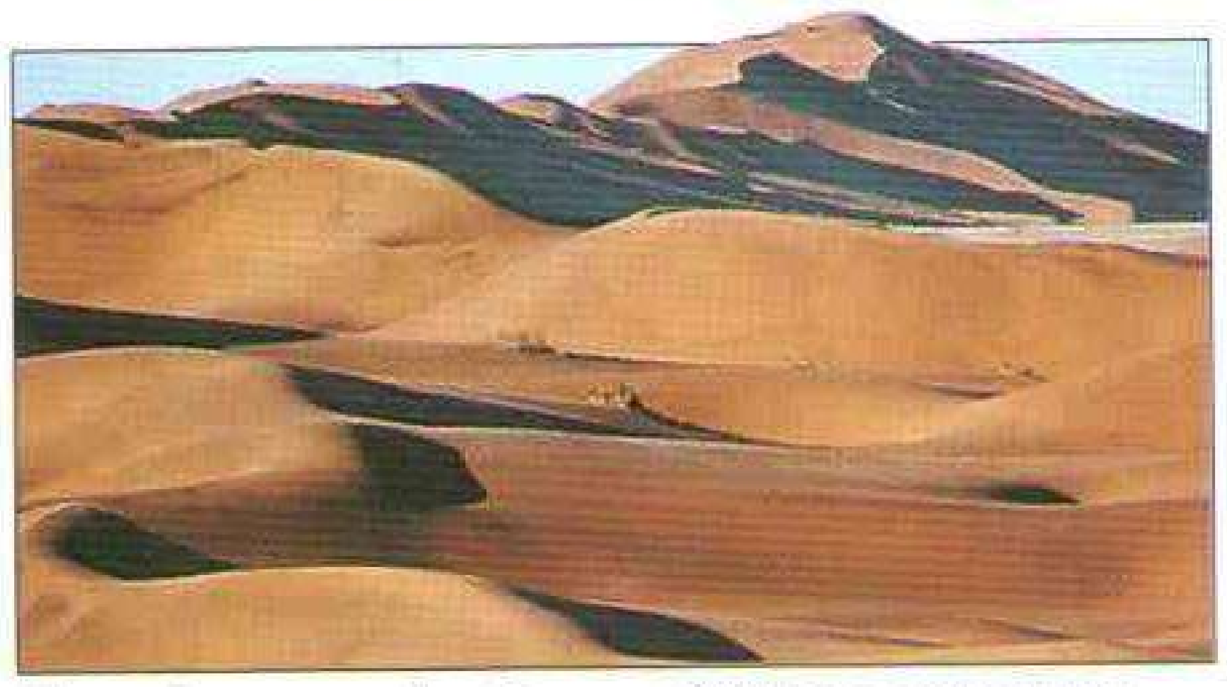

Erg. *Las grandes áreas cálidas y arenosas con dunas se conocen como ergs. Se encuentran en Australia, en el norte de África y en Asia.*

Desierto rocoso. *Las áreas llanas y rocosas fueron en su día los lechos de los antiguos mares. Estos desiertos, que se encuentran en todo el mundo, son muy inhóspitos debido a la falta de refugios y a los vientos constantes.*

Desierto de cactus. *Algunos desiertos pueden soportar algún tipo de vida especializada, como los cactus, que retienen agua en sus tejidos. Estos desiertos se encuentran en Norteamérica, Sudamérica, Australia y Asia.*

Semidesierto templado. *A pesar del crecimiento de algunos arbustos leñosos, es demasiado seco para soportar el crecimiento de hierbas. Los semidesiertos templados se encuentran en Norteamérica, Sudamérica, Asia central y Australia.*

CÓMO PROTEGERSE EN UNA TORMENTA DE ARENA

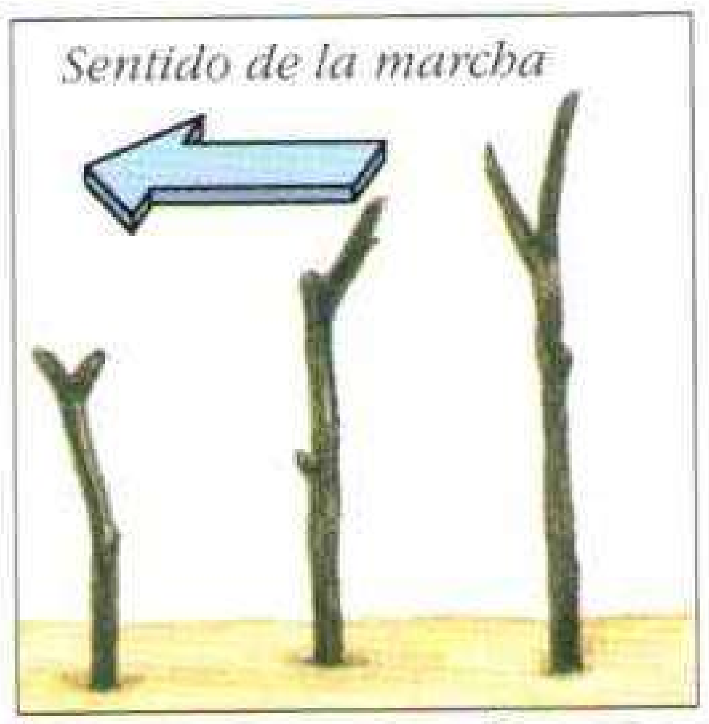

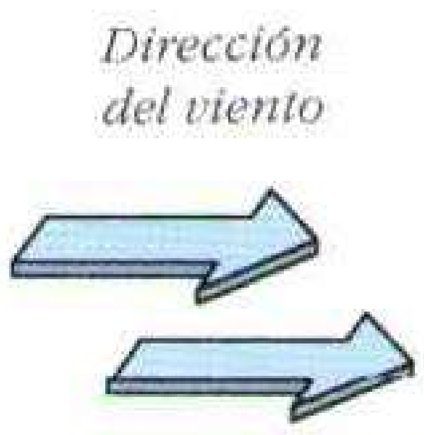

1 Las tormentas de arena son muy desagradables y pueden llegar a alterar el aspecto de un desierto. Debe marcar el sentido de su marcha con piedras o ramas.

2 Refúgiese donde pueda, con el viento a su espalda. Cubra su cuerpo totalmente, protegiendo su cara y cuello de la arena. Espere hasta que la tormenta de arena haya cesado por completo antes de reemprender la marcha.

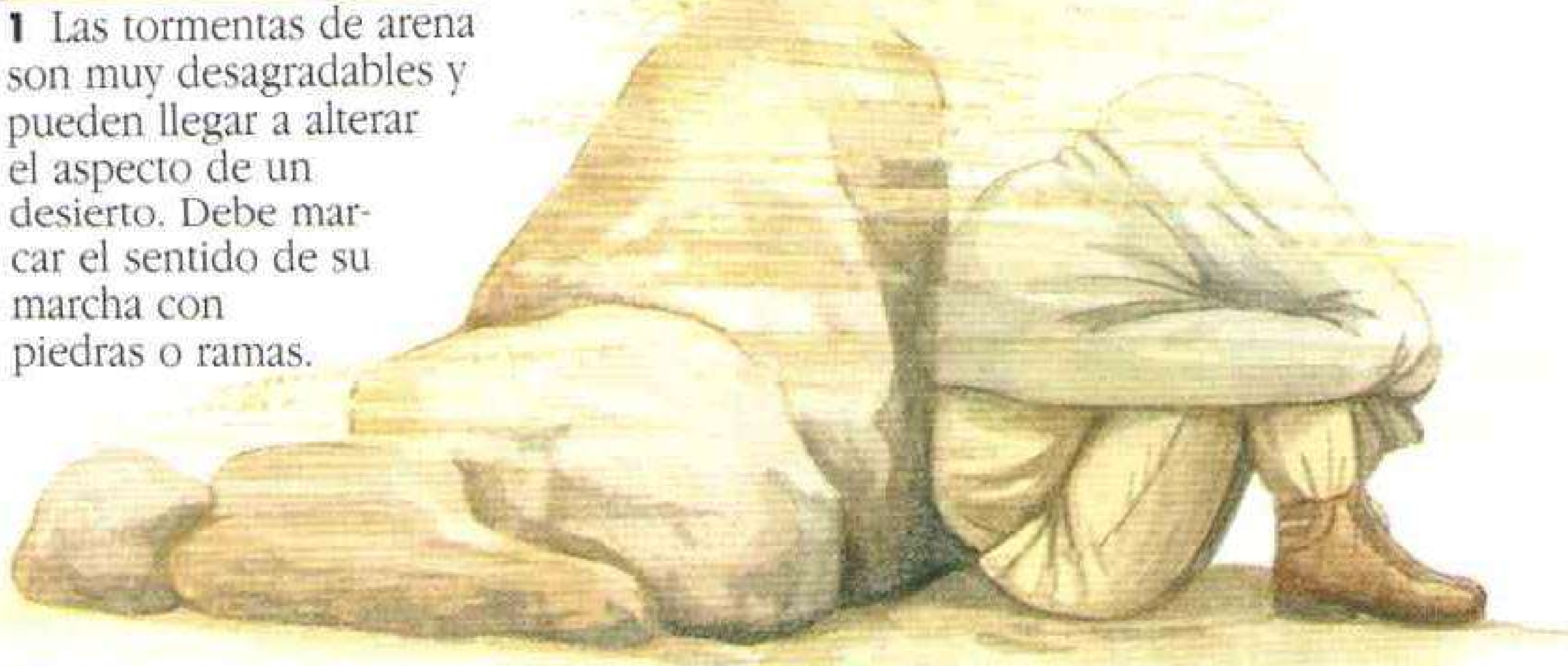

EL VIAJE POR EL DESIERTO

En una situación de supervivencia en el desierto, no debe moverse a menos que sea imprescindible. Si debe desplazarse, por ejemplo para buscar agua, o porque no existe la esperanza de que pueda ser rescatado desde donde se encuentra, hágalo únicamente de noche, cuando refresca, y si es posible bajo la luz de la luna. Como es muy difícil determinar dónde se encuentra en un desierto sin referencia alguna, la búsqueda de un oasis (suponiendo que sabe dónde se encuentran) está lejos de ser fácil.

Sombra improvisada. *Puede preparar un espacio sombreado colgando una manta sobre una cuerda atada entre dos estacas. Cave una depresión en la arena para tener espacio suficiente para moverse bajo el tejado.*

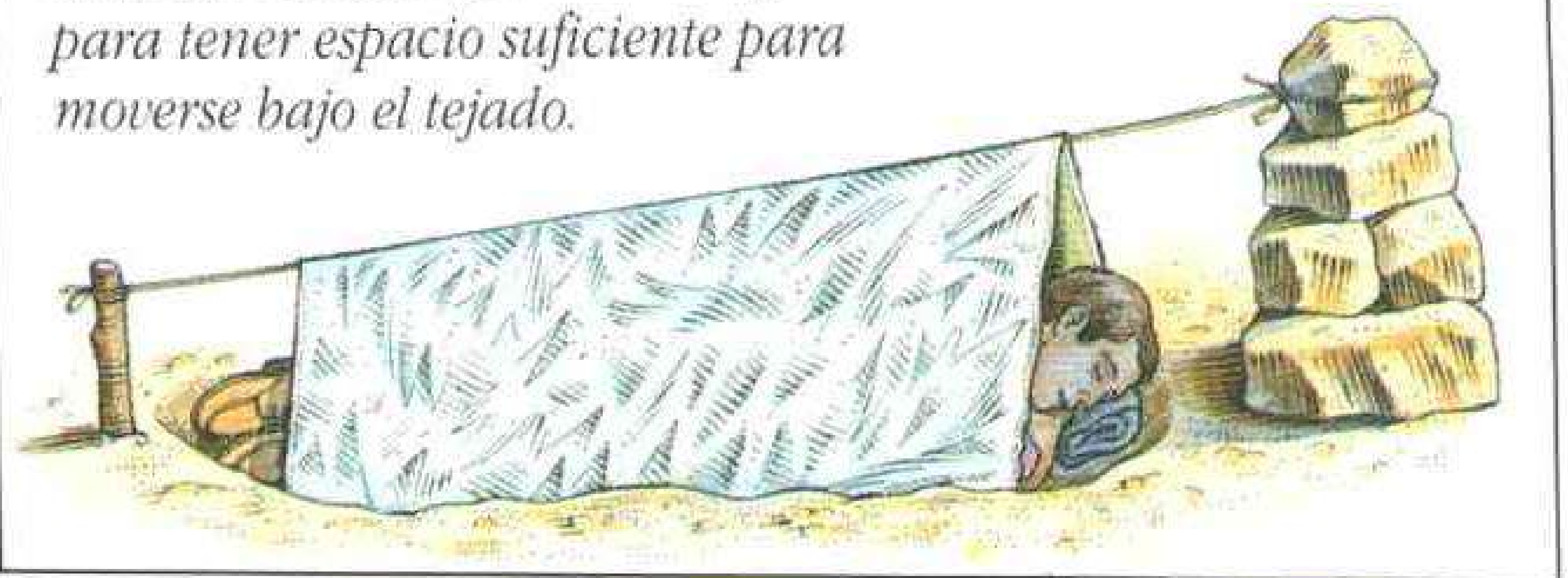

TIEMPO DE SUPERVIVENCIA EN EL DESIERTO CON RESERVA LIMITADA DE AGUA				
		Descansando en la sombra		
°C	SIN AGUA	2 LITROS	10 LITROS	°F
50	2-5 días	2-5 días	3-5 días	122
30	7 días	5-8 días	14 días	86
20	12 días	14 días	23-25 días	68
		Caminando de noche		
50	1 día	2 días	3-5 días	122
30	4 días	5 días	5-7 días	86
20	9 días	10-15 días	5-15 días	68

Supervivencia sin agua. *Los seres humanos necesitan un mínimo de 4-5 litros de agua por día para ser capaces de viajar a pie con un mínimo grado de seguridad. En el desierto, la conservación del agua de su cuerpo es primordial, por lo que únicamente debe trabajar de noche, hacer los mínimos esfuerzos, e intentar evitar la transpiración. A temperaturas de 50 °C, sin agua para beber, únicamente sobrevivirá cinco días, si se limita a descansar en la sombra. Este tiempo se reducirá a la mitad si camina, incluso de noche. En una situación de supervivencia, debe cuestionarse seriamente la necesidad de cualquier tipo de movimiento a menos de que disponga de una gran reserva de agua.*

VIAJAR A TRAVÉS DE LA SELVA

A PESAR DE SER cálida, húmeda y estar llena de insectos, la selva es, en muchos aspectos, un jardín del Edén. Los alimentos son abundantes, no hay estaciones de carencia, y el agua fresca cae del cielo en grandes cantidades. Sin embargo, para aquellos que no estén familiarizados, la selva es húmeda, oscura y ruidosa: un lugar incómodo y amenazador. La humedad y el intenso esfuerzo que supone viajar a través de la selva hacen que la insolación sea un peligro constante, por lo que debe llevar consigo una reserva importante de agua para evitarla.

INDUMENTARIA ADECUADA PARA LA SELVA

Debe cubrirse completamente cuando viaje a través de la selva, y llevar ropa ligera y resistente que se seque rápidamente. Es posible que su ropa esté húmeda la mayor parte del día; deberá lavarla en agua fresca cada día para evitar que se pudra. No tiene sentido llevar ropa impermeable porque únicamente le haría sudar. Un jersey ligero resulta útil en las noches frescas. Tenga siempre a punto una muda de ropa limpia y seca para dormir. Guarde todo su equipo a menos que lo esté utilizando. Los artículos que no estén bien sujetos pueden engancharse en la vegetación, así que sujete sus objetos valiosos con cuerdas resistentes alrededor del cuello.

BOTIQUÍN PARA LA SELVA

No permita que las pequeñas heridas lleguen a infectarse. Al tener que abrirse camino entre una vegetación espesa y espinosa, las heridas en las manos son inevitables. Los antibióticos en polvo son ideales para prevenir infecciones. Cuando se abra paso entre la densa vegetación, tenga cuidado y evite herirse con el machete, porque podría dejarle imposibilitado o causarle una herida que difícilmente no se le infectaría. Necesitará una gran cantidad de tiritas impermeables para cubrir las pequeñas heridas y cortes.

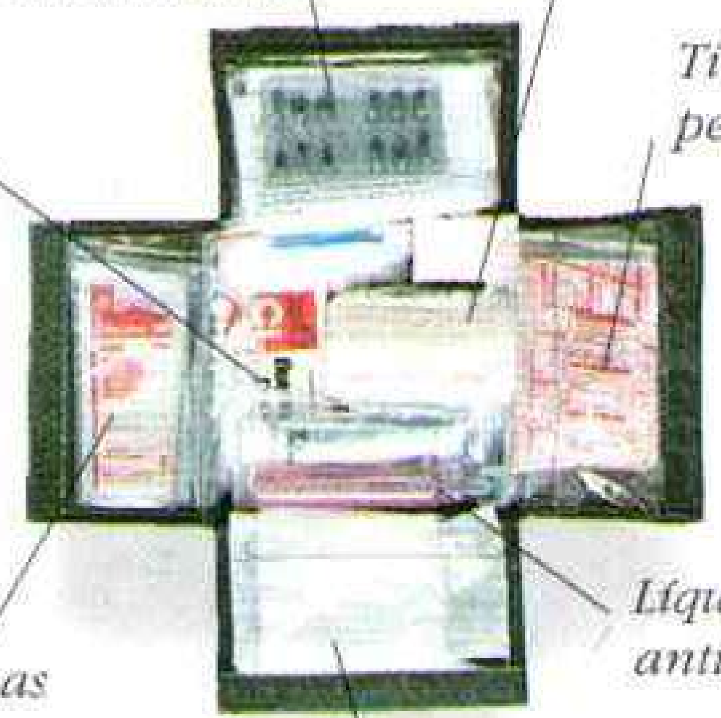

Medicación tropical. *Existen botiquines especiales para las regiones tropicales. Contienen artículos comunes para primeros auxilios y medicamentos para contrarrestar las enfermedades tropicales.*

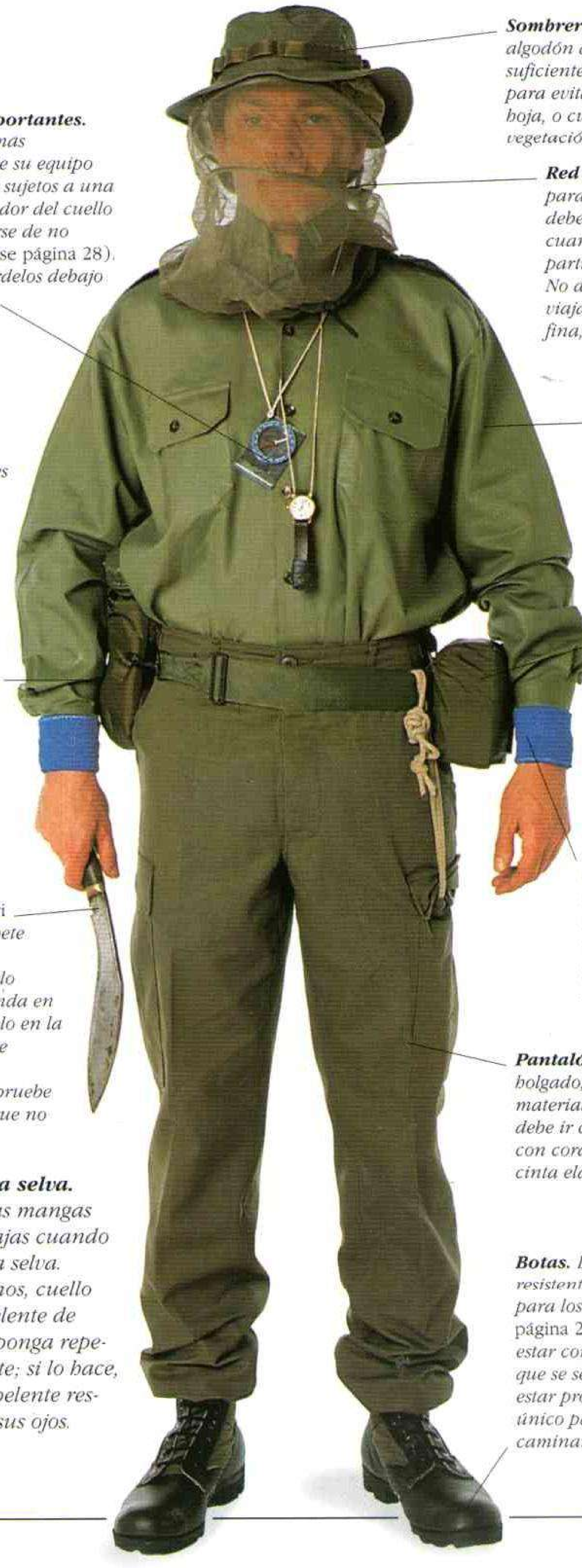

Artículos importantes. *Los artículos más importantes de su equipo deben quedar sujetos a una cuerda alrededor del cuello para asegurarse de no perderlos* (véase página 28). *Al viajar, guárdelos debajo de la camisa.*

Cantimplora. *Una cantimplora es absolutamente necesaria. Llévela sujeta al cinturón para no perderla. Recuerde filtrar y depurar toda el agua que recoja antes de beberla* (véase página 74).

Cuchillo. *Su* kukri *(cuchillo) o machete debe estar bien protegido; guárdelo dentro de una funda en su cinturón. Llévelo en la mano únicamente mientras lo esté empleando. Compruebe constantemente que no lo ha perdido.*

Equipo para la selva. *Lleve siempre las mangas de la camisa bajas cuando se adentre en la selva. Proteja sus manos, cuello y cara con repelente de insectos. No se ponga repelente en la frente; si lo hace, el sudor y el repelente resbalarán hacia sus ojos.*

Sombrero. *Un sombrero de algodón debe tener un ala lo suficientemente ancha como para evitar que una rama, una hoja, o cualquier tipo de vegetación lastime sus ojos.*

Red para mosquitos. *La red para mosquitos únicamente debe utilizarse al atardecer, cuando estos insectos son particularmente molestos. No debe utilizarla mientras viaja, ya que la malla, aunque fina, obstruye la visibilidad.*

Camisa. *La camisa debe ser de algodón y de manga larga. Asimismo, debe tener el cuello que pueda levantarse y abotonarse para protegerle la garganta.*

Botiquín de primeros auxilios. *Lleve siempre consigo un botiquín de primeros auxilios, sea en el cinturón o en un bolsillo.*

Tiras repelentes de insectos. *Estas tiras que se llevan en las muñecas y tobillos están impregnadas con repelente de insectos para alejar a los mosquitos y otros parásitos.*

Pantalón. *El pantalón debe ser holgado, confeccionado con material resistente pero ligero, y debe ir ceñido a los tobillos, sea con cordones incorporados o con cinta elástica.*

Botas. *Las botas deben tener suelas resistentes, con protección especial para los dedos y el empeine* (véase página 25). *La parte superior debe estar confeccionada en tela, para que se seque rápidamente, y deben estar provistas de válvulas de paso único para evacuar el agua al caminar.*

CARACTERÍSTICAS DE LA SELVA

Bóveda verde. *En una selva primaria, los árboles pueden crecer hasta una altura de más de 60 m, creando una cubierta densa que oculta el sol. Esta bóveda va siguiendo la forma del suelo que hay debajo, pero desde el aire, lo único que puede verse es el denso techo de vegetación. La mayoría de los animales que habitan este medio lo hacen en la copa de los árboles.*

Vegetación exuberante. *La selva secundaria crece allí donde los hombres quemaron árboles para obtener tierras que cultivar. Después de la cosecha, abandonaron estos claros permitiendo que el sol llegara al nivel del suelo y estimulara el enorme crecimiento. Este tipo de selva es difícil de cruzar. Debe abrirse un sendero a través de la vegetación, y éste es un trabajo extenuante.*

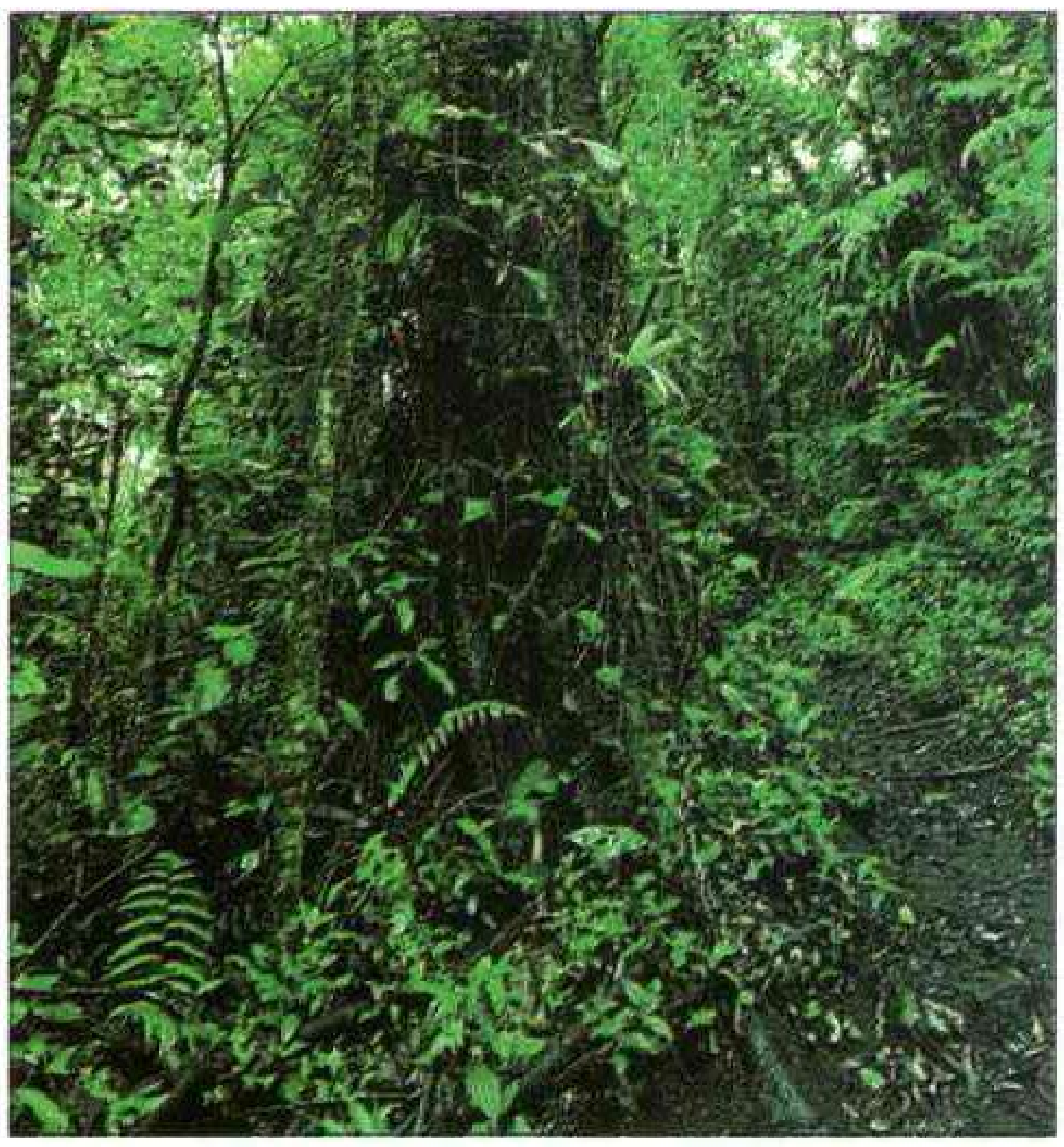

Selva umbría. *Debajo de la bóveda, el suelo de la jungla permanece relativamente libre de vegetación. Las lianas gigantes crecen hacia abajo, y una gran variedad de flora y fauna habita los distintos niveles de este medio húmedo y sombreado.*

DESPLAZARSE A TRAVÉS DE LA SELVA

Aunque es fácil desplazarse en la selva primaria, la selva secundaria es densa, llena de obstáculos y desagradable. Cortar la vegetación es un trabajo árduo que hace muy lento el avance. Abundan los insectos y las hierbas urticantes, muchas de las cuales tienen espinas peligrosas. Si queda atrapado por una de ellas, la mejor manera de liberarse por sí mismo es moviéndose hacia atrás, invirtiendo el proceso por el que quedó atrapado.

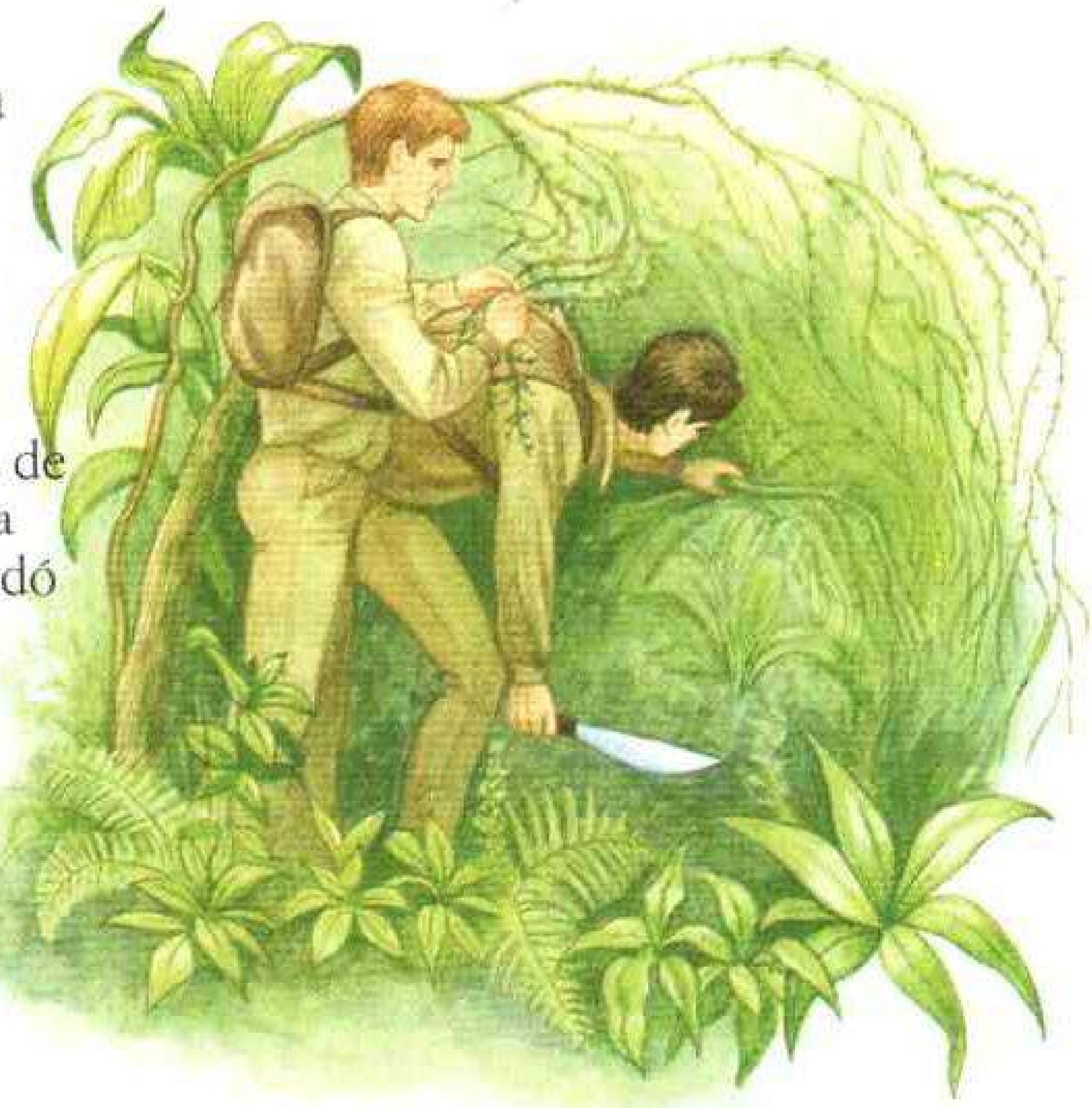

Aguarda un momento. *Algunas especies de* rattan, *conocidas como «aguarda un momento» pueden convertir su ropa en jirones. Si queda atrapado, espere pacientemente a que un compañero le libere.*

LA RUTINA EN LA SELVA ECUATORIAL

Conforme se acerca al ecuador, los días se hacen más regulares. En las áreas ecuatoriales, el amanecer y el atardecer tienen lugar a la misma hora cada día, y a menudo llueve a la misma hora en punto —por lo general, justo al atardecer. No hay una puesta de sol gradual, como ocurre en las latitudes más altas, sino una oscuridad repentina y total. La rutina diaria consiste en levantarse al amanecer, trabajar hasta el mediodía y descansar durante las horas de mucho calor. Si se está desplazando, comience a buscar un lugar para acampar no más tarde de las 15.00 horas.

Eliminar las sanguijuelas. *No se limite a arrancar las sanguijuelas, ya que las partes bucales podrían quedar dentro de la herida y causar infecciones. Tóquelas con una rama ardiendo o empápelas en repelente de insectos para que suelten su presa, y quítelas entonces.*

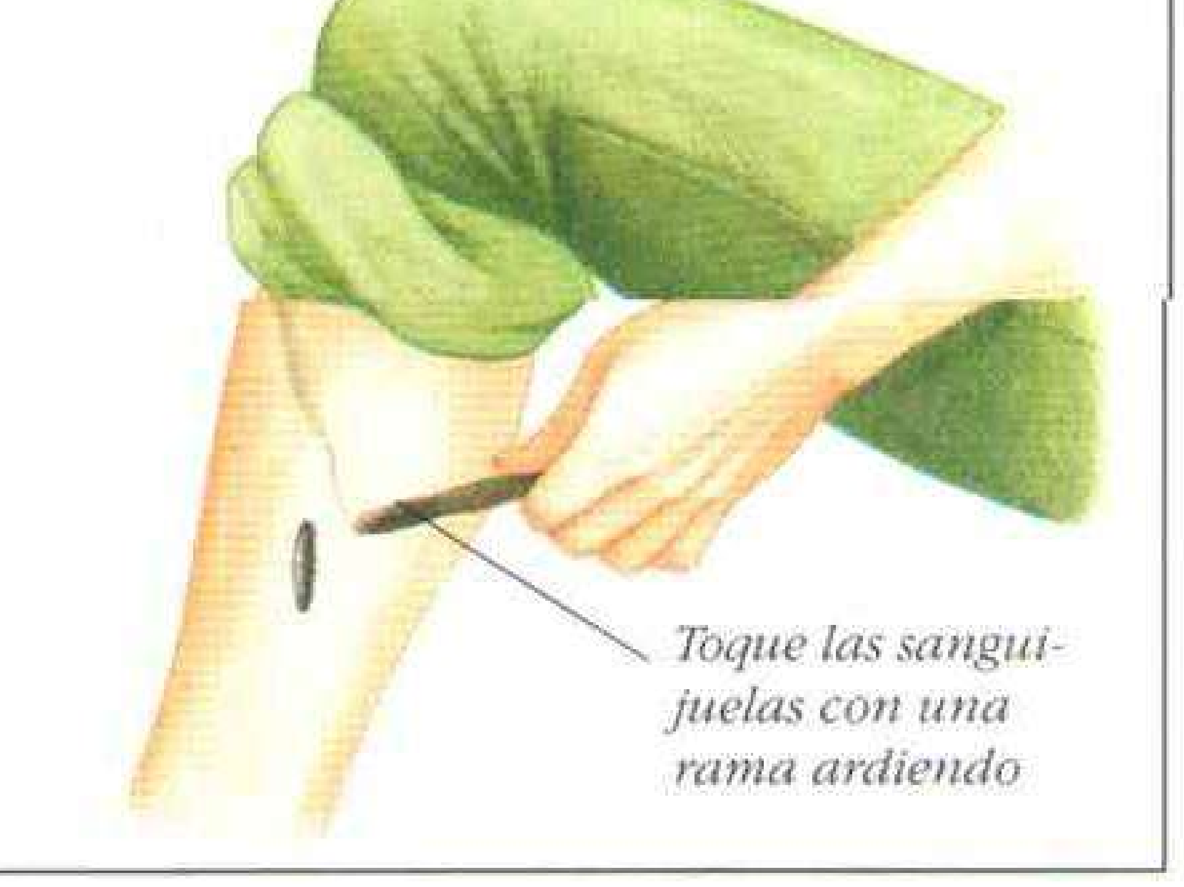

Toque las sanguijuelas con una rama ardiendo

LOS PELIGROS DE LA SELVA

La selva tiene muchos peligros, que van desde los animales salvajes hasta enfermedades como la malaria. La constante humedad puede causar infecciones fúngicas en los pies, por lo que es recomendable aplicar polvos de talco regularmente *(véase página 24)*. Sacuda la ropa y las botas antes de ponérselas, en caso de que se hayan convertido en refugio de arañas o serpientes venenosas.

Asegúrese de vacunarse contra la malaria, y tome la adecuada medicación antimalaria antes de adentrarse en la selva. Utilice los repelentes de insectos y cuelgue una mosquitera durante la noche para evitar sus picaduras.

Enemigo mortal. *La hembra del mosquito* Anopheles *puede transmitir el mortal virus de la malaria.*

UNA ESCALADA SEGURA

ÚNICAMENTE DEBE EMPRENDER una escalada si no hay otra manera de sortear un obstáculo, y nunca si está solo. Si cae o se queda atrapado, nadie podrá ir en busca de ayuda. La regla de oro de todo escalador es mantener tres puntos de contacto con la roca, emulando la estabilidad de un trípode. Piense cuidadosamente cada movimiento, intentando mantener tres partes de su cuerpo (los dedos de los pies, los dedos de las manos, los pies o las manos —nunca las rodillas—) en contacto con la roca en todo momento. Tenga cuidado al evaluar el estado de la roca y la estabilidad de las presas para manos y pies. Utilice siempre una cuerda, si dispone de ella.

PRESAS PARA MANOS Y PIES

Las presas para manos y pies son la base de la escalada en roca. Antes de poner el pie o la mano en una presa, póngala a prueba con una patada o con un golpe de mano. No intente asir una muy lejana, o ponerse de puntillas para alcanzar una muy alta. Utilice sus piernas, más que las manos, para soportar su peso y empujarse hacia arriba. No se levante con los brazos.

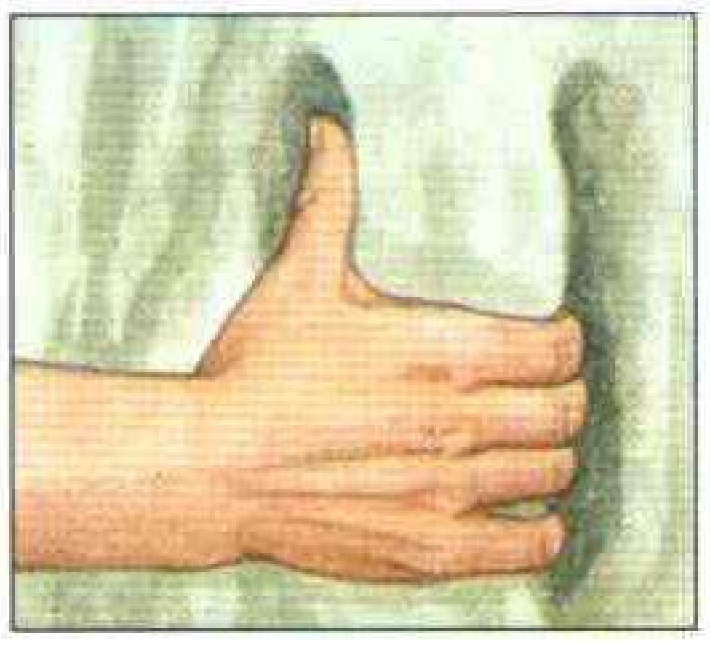

Presa de oposición. *Puede servirse de presas laterales para equilibrarse. Las presas de oposición son ideales si puede utilizar el pulgar para presionar hacia el lado contrario.*

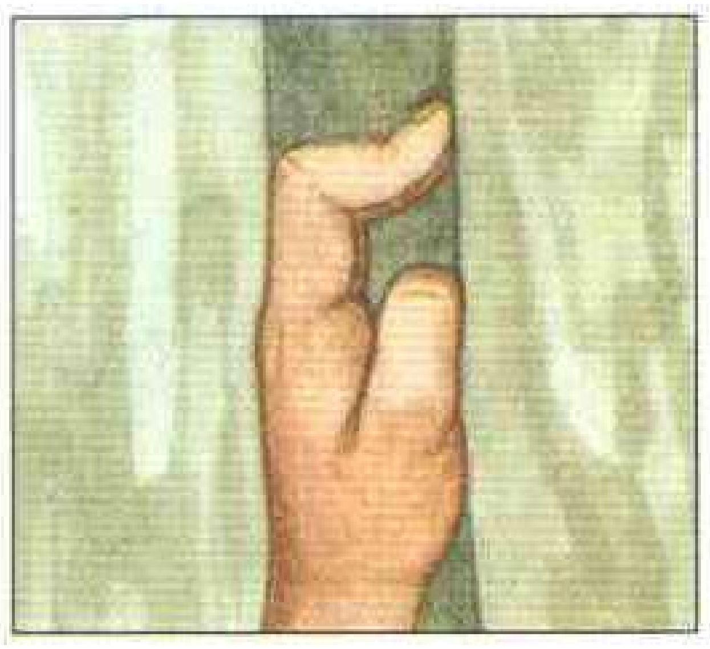

Empotramiento de mano. *Puede encajar su mano o puño en una pequeña grieta, e impulsarse hacia arriba cuando no exista una presa de otro tipo.*

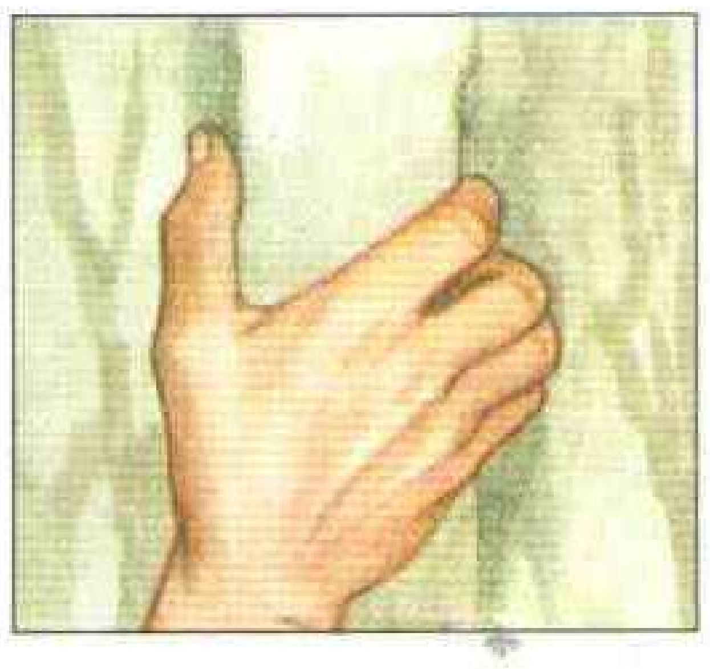

Presa de pellizco. *Cuando no pueda aferrarse bien a una presa, agárrela como si se tratase de un ladrillo. Intente pellizcar la roca.*

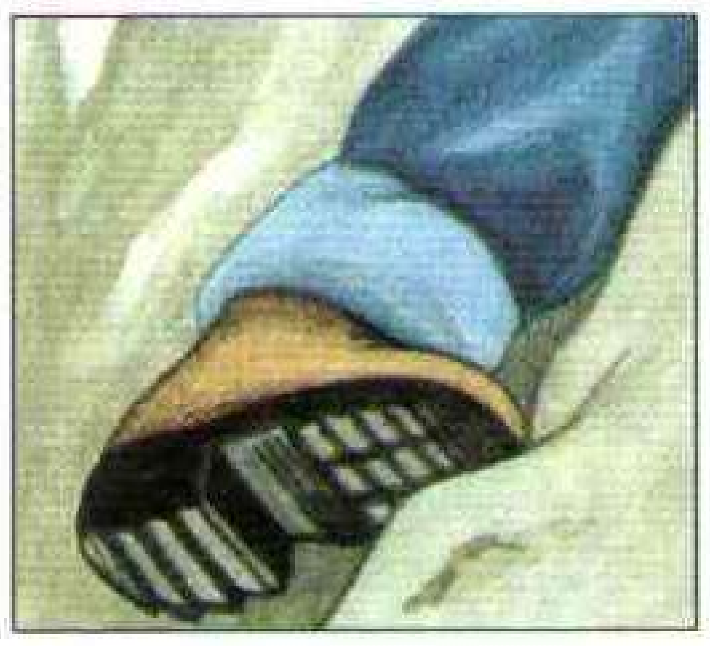

Empotramiento de pie. *Utilice una hendidura de la roca para apoyar la parte interior del pie. Deslice la suela a lo largo de la hendidura, en un ángulo de 90 grados con la pared.*

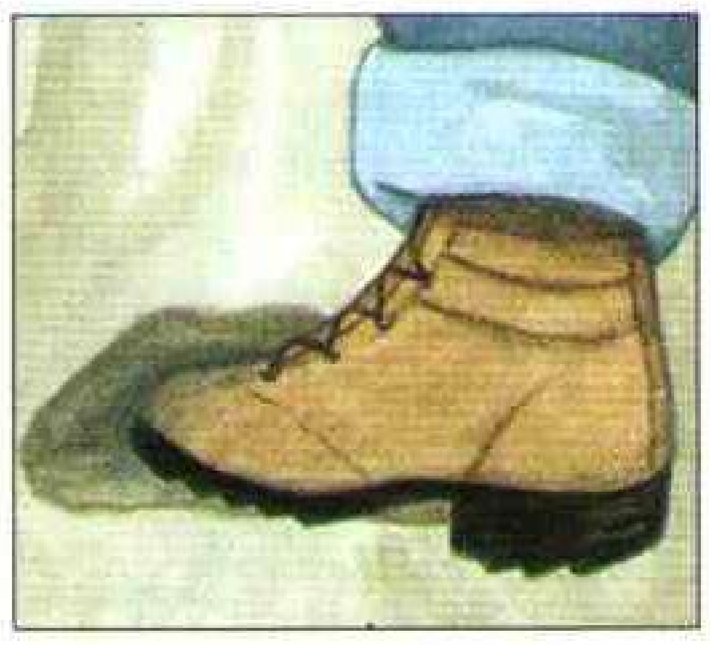

Presa grande. *Los agujeros en la roca se conocen como* presas grandes *y constituyen apoyos de pie ideales. Equilíbrese para que sus brazos y piernas soporten su peso por igual.*

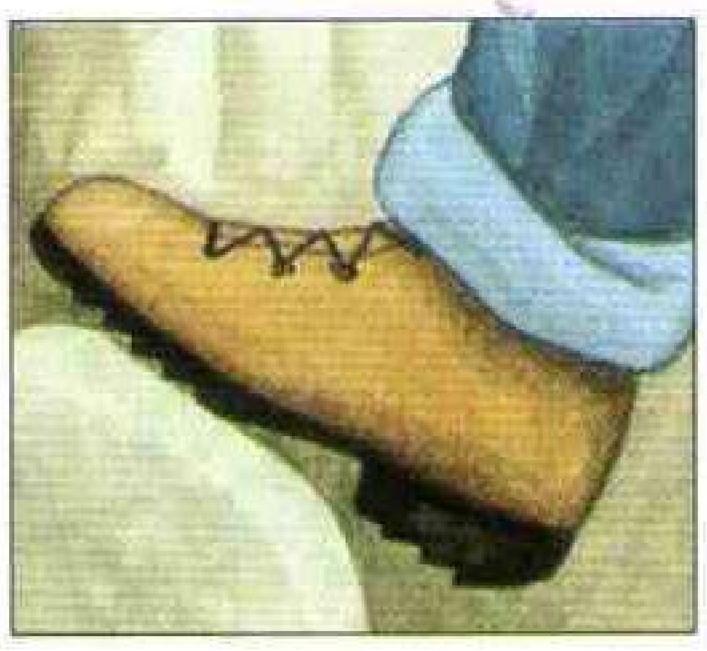

Presa de pie en pendiente. *Intente colocar la totalidad de la suela sobre la presa. Si únicamente puede apoyar la punta del pie, doble el tobillo para mejorar la adherencia.*

ESCALADA BÁSICA

Si es usted un novato, no debería escalar sin estar sujeto a una cuerda asegurada por un escalador experimentado. Sin embargo, en una situación de supervivencia, puede no tener otra opción. Antes de escalar, busque con la vista el recorrido más fácil y seguro, y continúe mirando mientras va escalando. Si tiene cualquier duda sobre la ruta, retroceda y empiece de nuevo.

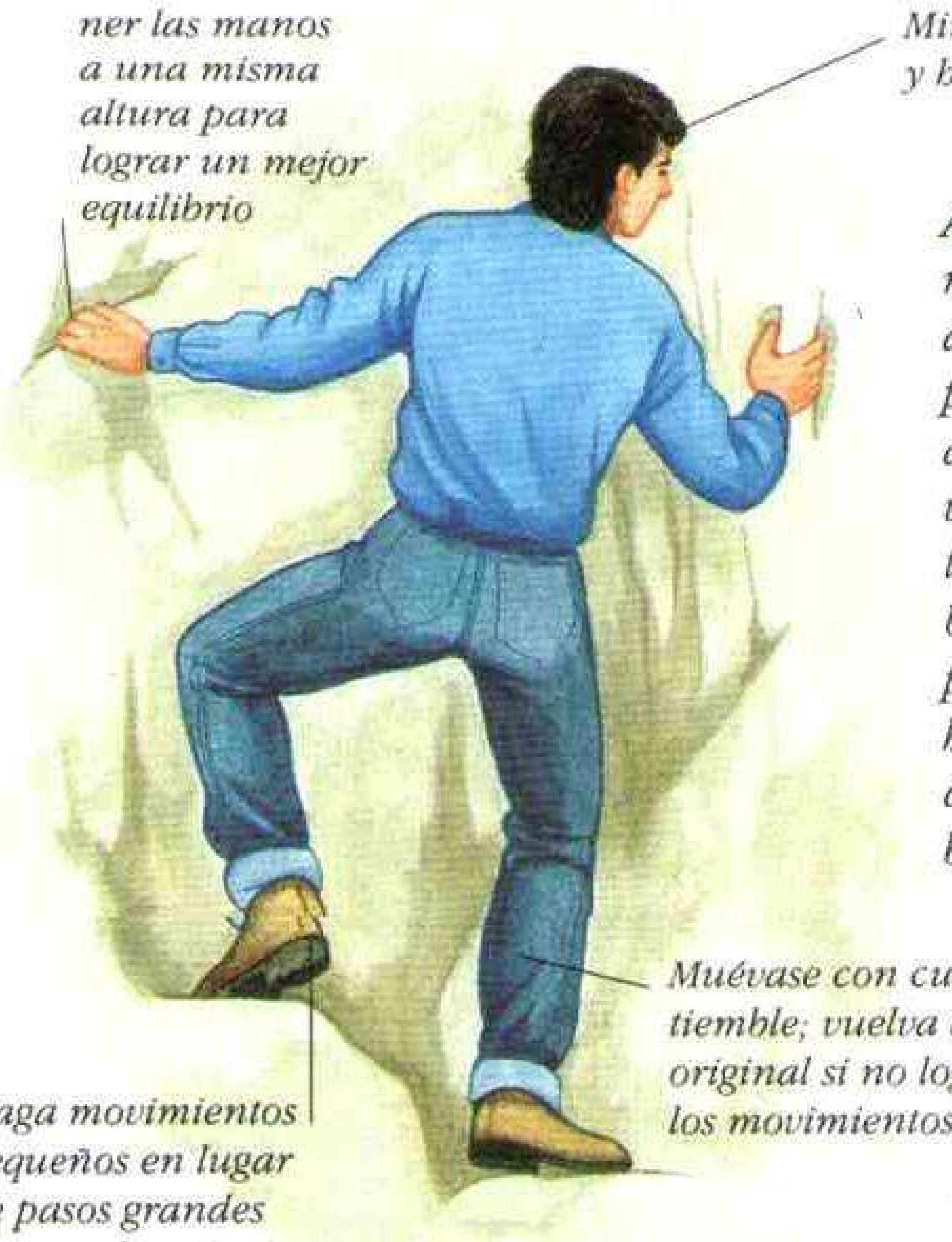

Intente mantener las manos a una misma altura para lograr un mejor equilibrio
Mire hacia arriba y hacia los lados
Muévase con cuidado y no tiemble; vuelva a su posición original si no logra realizar los movimientos
Haga movimientos pequeños en lugar de pasos grandes que puedan significar un riesgo excesivo

Ascenso. *Planee varios movimientos con antelación, y esté preparado para desplazarse, no sólo verticalmente, sino también lateralmente. Utilice los músculos de las piernas para impulsarse hacia arriba, y no dependa sólo de los brazos.*

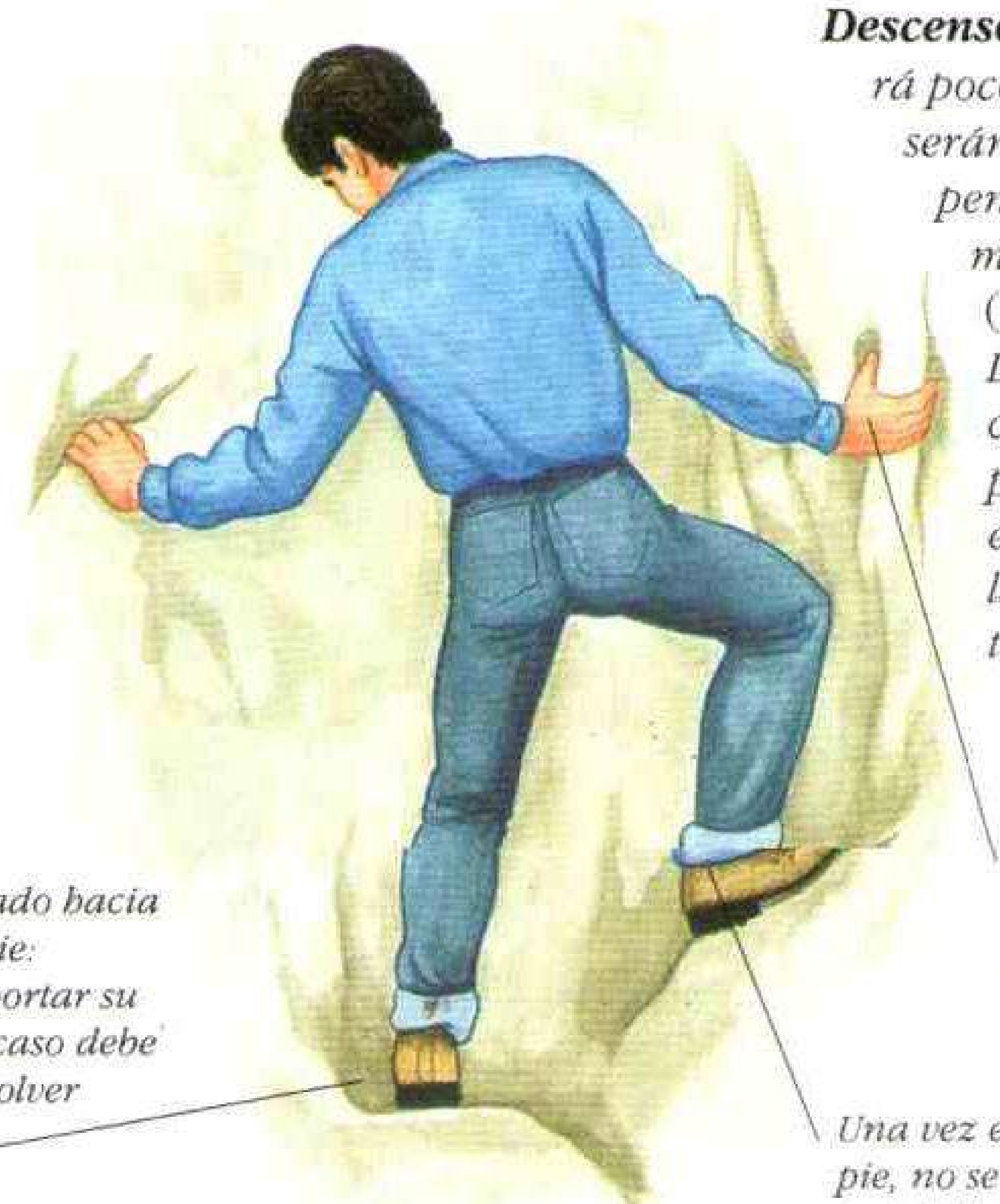

Descenso. *El descenso le parecerá poco natural y las presas serán difíciles de ver. Si la pendiente es suave, baje mirando hacia afuera* (véase página 136). *Descienda de lado conforme aumenta la pendiente, hasta llegar a encarar la roca. Busque las presas con los ojos, no tantee con los pies.*

Baje con cuidado hacia las presas de pie: pueden no soportar su peso, en cuyo caso debe ser capaz de volver hacia arriba
Busque siempre presas de mano y pie más arriba o a los lados, en caso de no poder seguir el recorrido previsto
Una vez encontrada una presa de pie, no se apoye de puntillas, y doble los tobillos para mejorar la adherencia

ASEGURADO

El seguro es el sistema por el cual un escalador se sujeta a otro que está escalando con una cuerda para parar una posible caída. El asegurador debe sujetarse a la pared rocosa («anclaje») antes de comenzar. Para sujetar a alguien que desciende, o si su compañero necesita más cuerda para ascender, vaya soltando cuerda en la forma inversa a como la recogería.

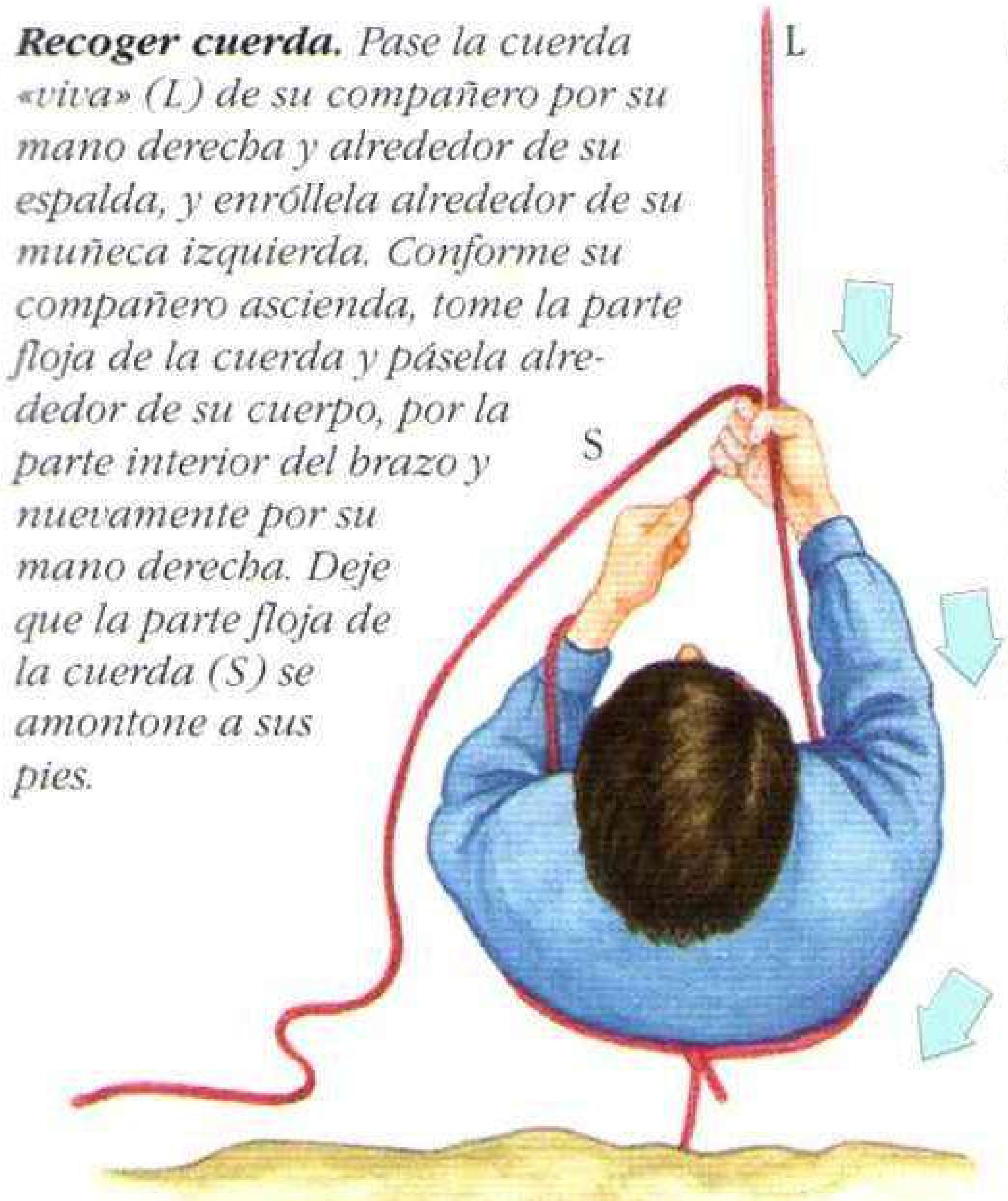

Recoger cuerda. *Pase la cuerda «viva» (L) de su compañero por su mano derecha y alrededor de su espalda, y enróllela alrededor de su muñeca izquierda. Conforme su compañero ascienda, tome la parte floja de la cuerda y pásela alrededor de su cuerpo, por la parte interior del brazo y nuevamente por su mano derecha. Deje que la parte floja de la cuerda (S) se amontone a sus pies.*

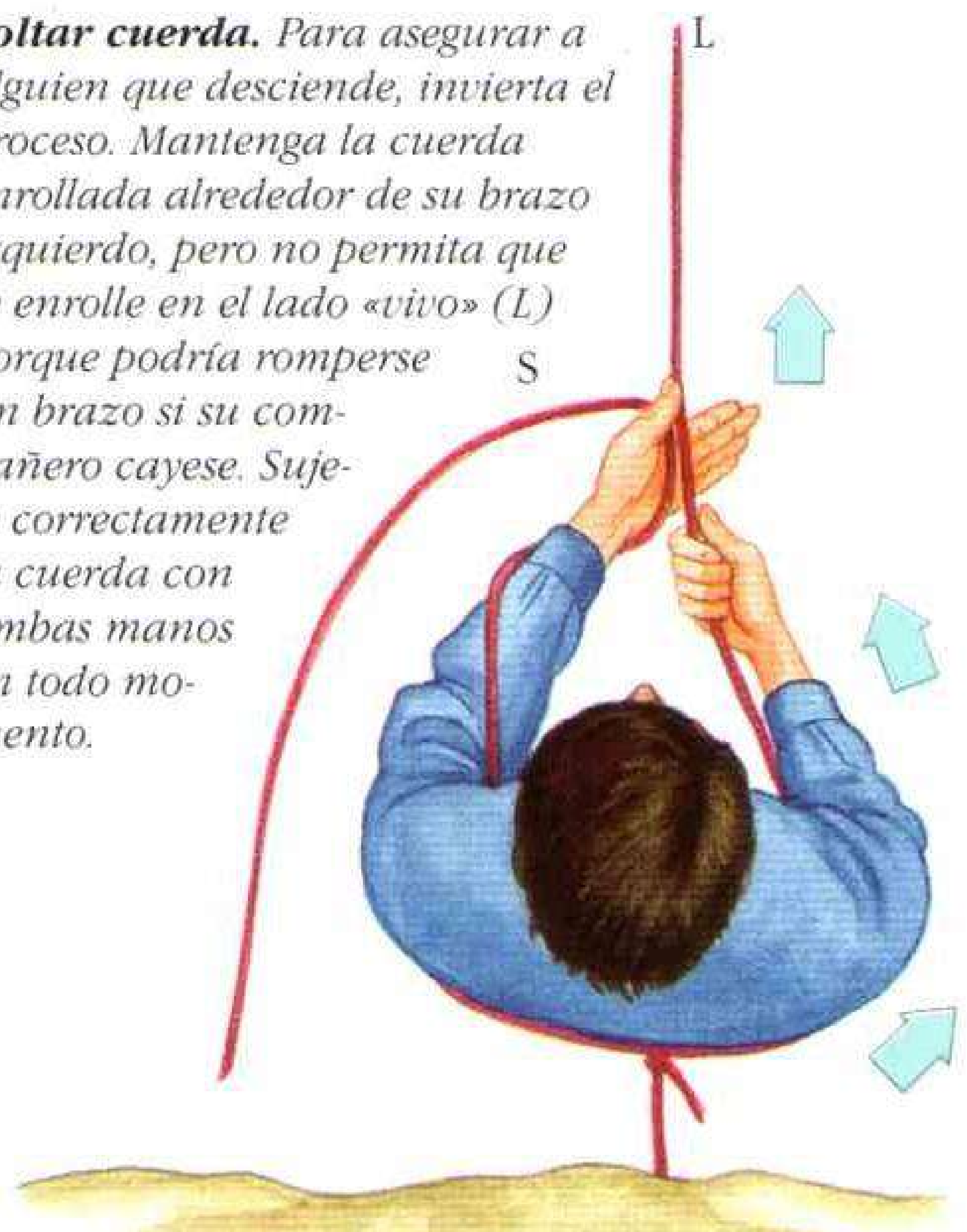

Soltar cuerda. *Para asegurar a alguien que desciende, invierta el proceso. Mantenga la cuerda enrollada alrededor de su brazo izquierdo, pero no permita que se enrolle en el lado «vivo» (L) porque podría romperse un brazo si su compañero cayese. Sujete correctamente la cuerda con ambas manos en todo momento.*

AGUANTAR UNA CAÍDA

Esté siempre preparado para aguantar a su compañero por si se cae. Para parar una caída, cruce su brazo izquierdo sobre la parte delantera de su cuerpo sujetando con fuerza su antebrazo contra el pecho. Reclínese en el punto de anclaje y haga fuerza.

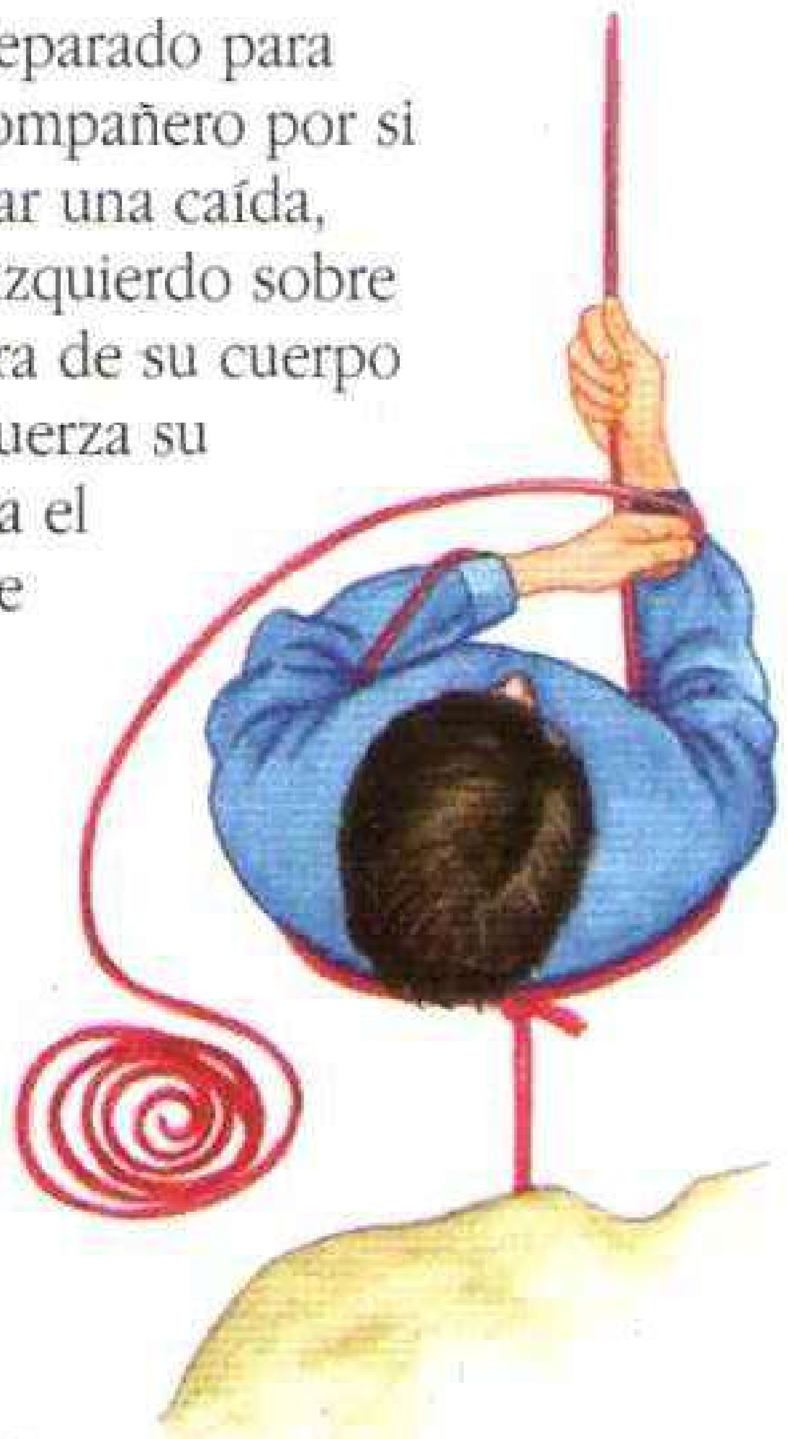

Bloqueo. *Esté preparado para sujetar a su compañero si cayese.*

RAPPEL ESPAÑOL O *DÜLFER*

El *rappel* español o *dülfer* es una técnica rápida y sencilla de descenso con cuerda. Pase la cuerda alrededor de un punto de anclaje seguro y suéltela. Colóquese frente al anclaje, ponga la cuerda entre las dos piernas, pásela alrededor de la cadera de modo que cruce el pecho hasta el hombro contrario; pásela por el hombro y déjela caer por la espalda. Agárrela con la mano que está del mismo lado que la cadera rozada por la cuerda. La otra mano aguanta la cuerda por arriba para mantener la posición vertical. Descienda caminando de espaldas, dejando que la cuerda se deslice por sus manos y alrededor de su cuerpo.

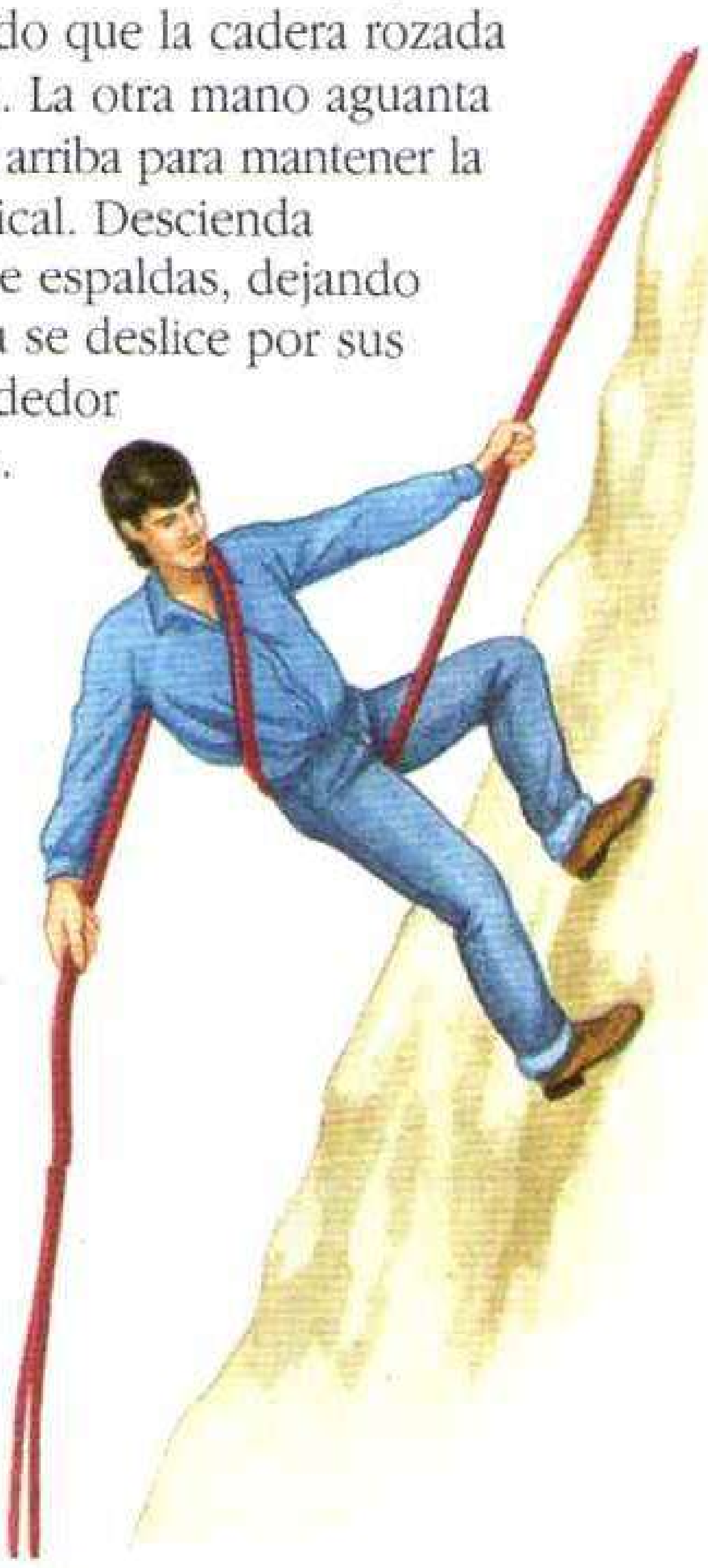

Dülfer. *Controle su velocidad con la mano derecha.*

ASCENSOS EN VARIOS LARGOS

Si ha de ascender por paredes de longitud superior a la de su cuerda, hágalo en varios largos, deteniéndose en puntos de reunión para asegurar a los demás miembros del grupo antes de continuar el ascenso.

1 El escalador que ascienda primero deberá buscar el recorrido más sencillo y seguro. Si la pared es muy empinada y peligrosa, puede ir insertando empotradores o clavijas conforme asciende, a fin de sujetar anillas de seguridad que aguanten la cuerda. El segundo escalador va soltando cuerda mientras permanece convenientemente anclado a la pared.

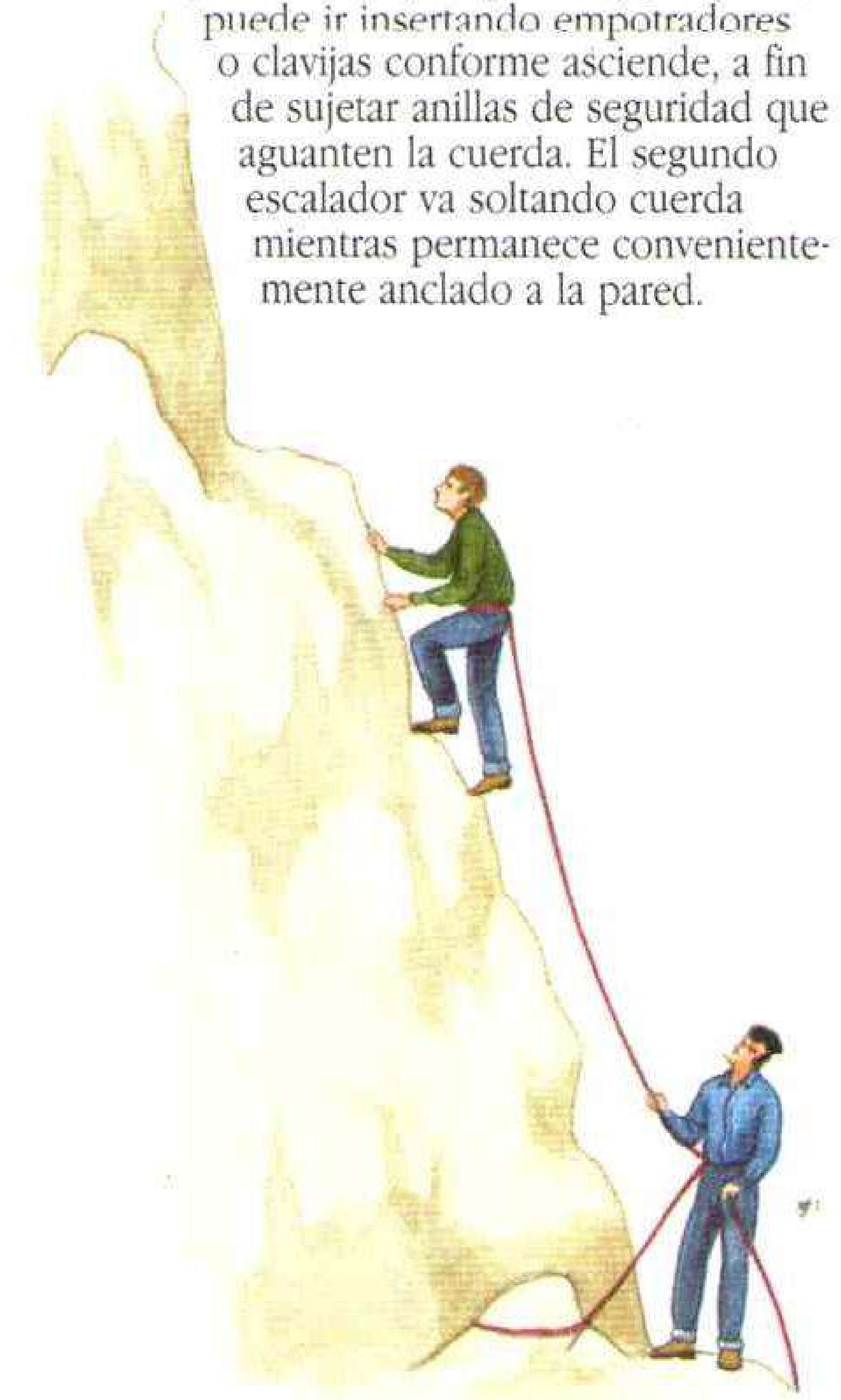

2 Una vez alcanzado el punto de reunión, se ancla a la pared y asegura al segundo escalador en su ascenso. Si el primer escalador ha empleado empotradores, el segundo deberá retirarlos conforme ascienda.

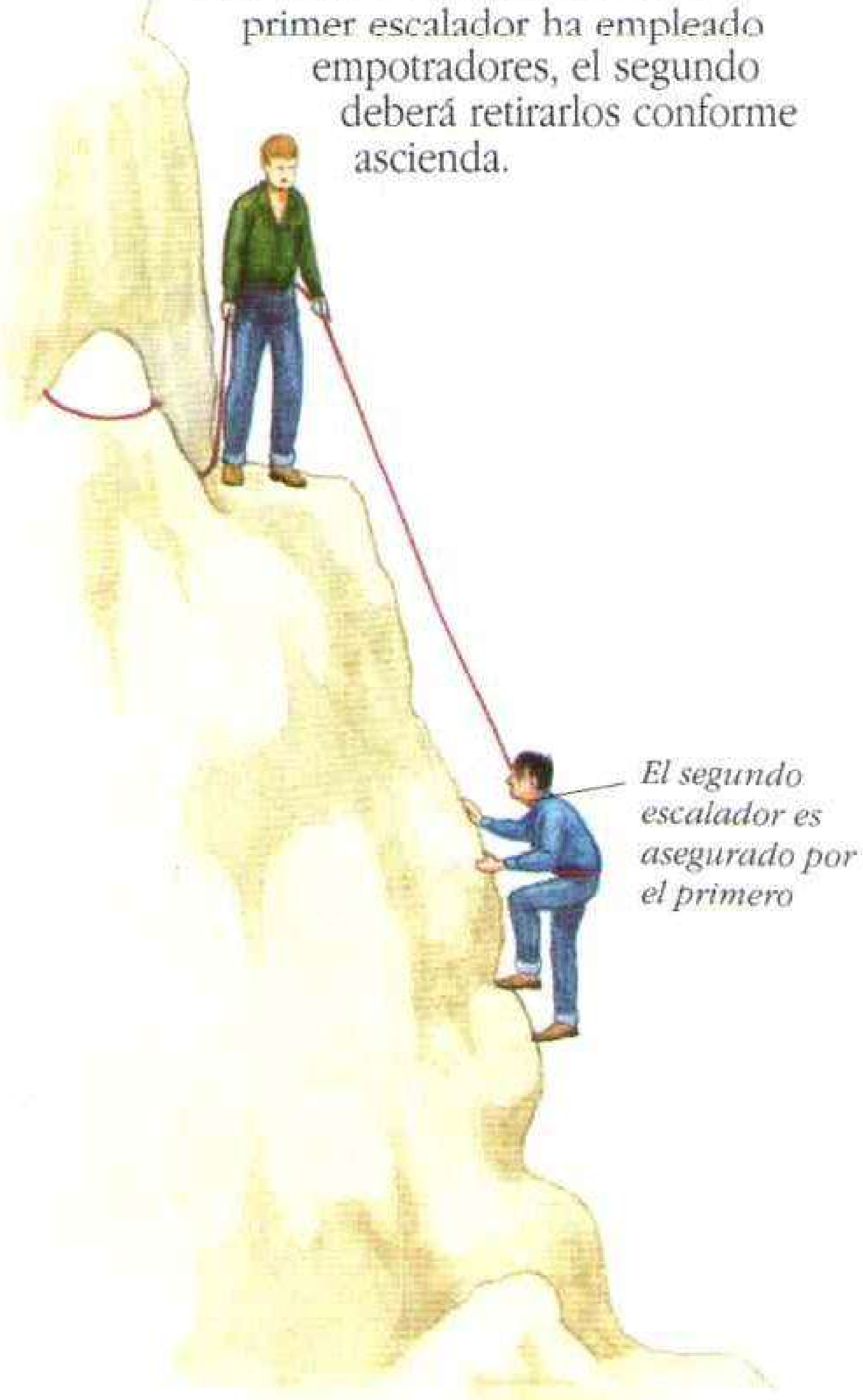

El segundo escalador es asegurado por el primero

ESCALADA DE EMERGENCIA

AUN LA ESCALADA más sencilla puede complicarse. Incluso escaladores experimentados no pueden seguir, caen o tienen que ser rescatados. Los expertos utilizan avanzadas técnicas para librarse de algún problema. Son peligrosas y no tan fáciles como parecen en las ilustraciones. En una situación de supervivencia, los problemas más graves deben solucionarse utilizando los recursos disponibles. Atienda a estas técnicas por si tuviera que emplearlas. Si intenta probarlas antes, consulte un manual de escalada y pida el consejo de un escalador con experiencia.

OPOSICIONES

El avance en oposición es una técnica de escalada que consiste en escalar por el interior de grietas lo suficientemente anchas para el cuerpo humano. Debe prever la forma en que saldrá una vez llegue a la parte superior. Es posible que entonces no sepa cómo salir.

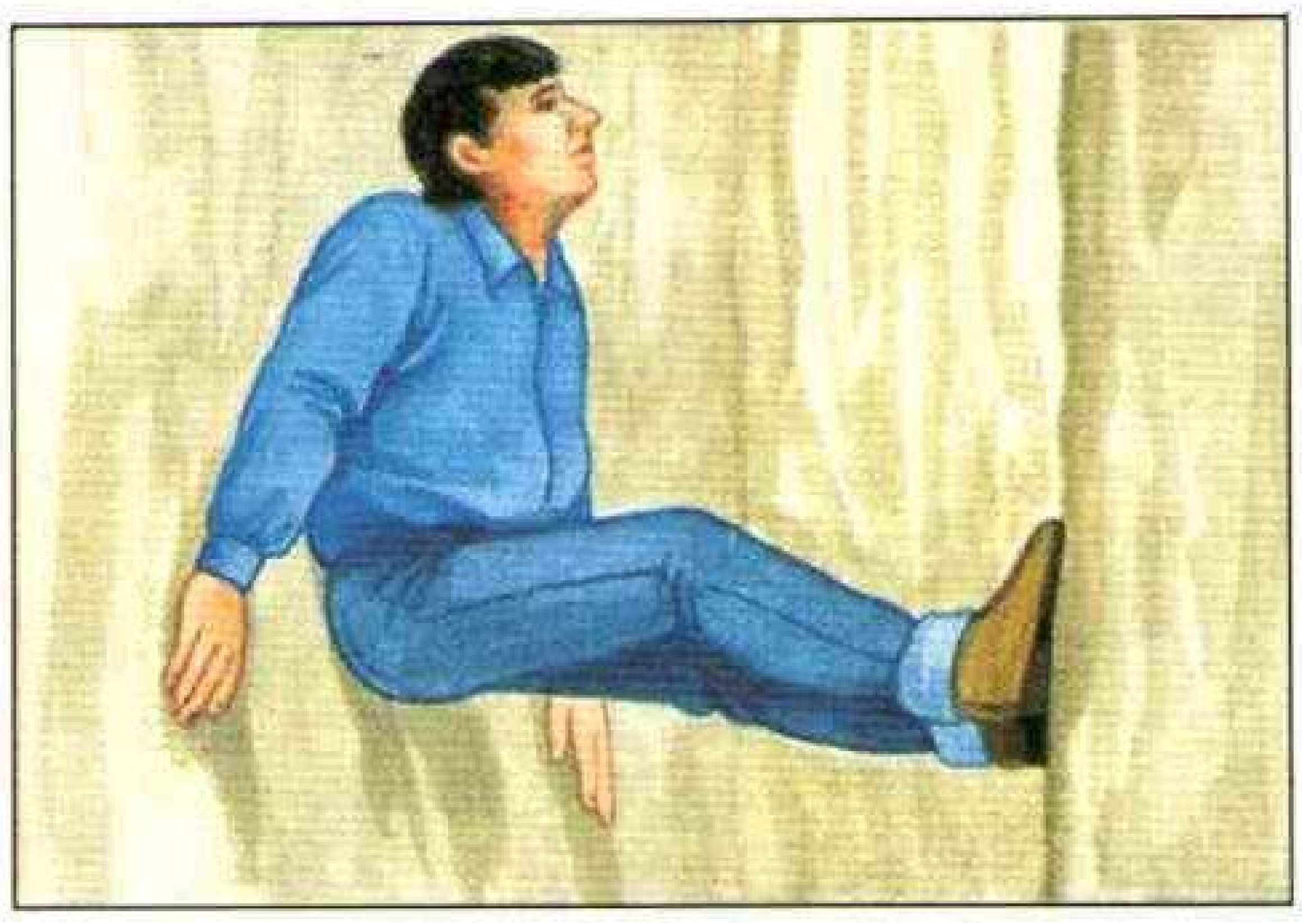

1 Compruebe si hay un recorrido alternativo antes de ascender por una grieta. Manténgase cerca del exterior de la grieta en una roca seca. Sujétese con ambas piernas y manos empujando contra las paredes.

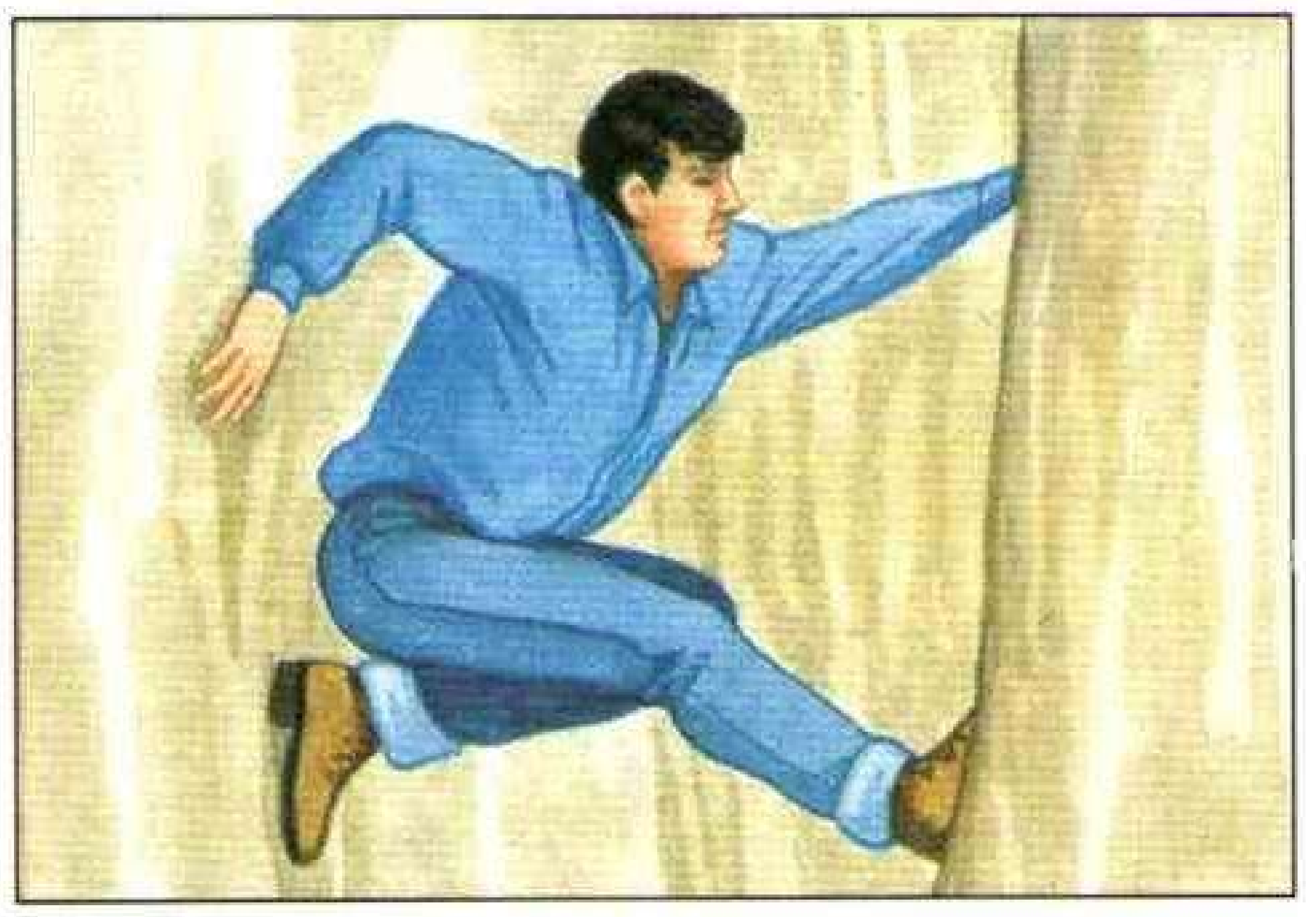

2 Presione un pie contra la pared trasera y la mano de ese mismo lado contra la pared delantera. Deslice la espalda hacia arriba, empujando con las piernas.

TOMARSE UN DESCANSO

El avance en oposición supone un esfuerzo muscular constante. Descanse a intervalos regulares apoyando la espalda contra la pared y manteniendo los brazos y piernas rectos.

DESCANSANDO

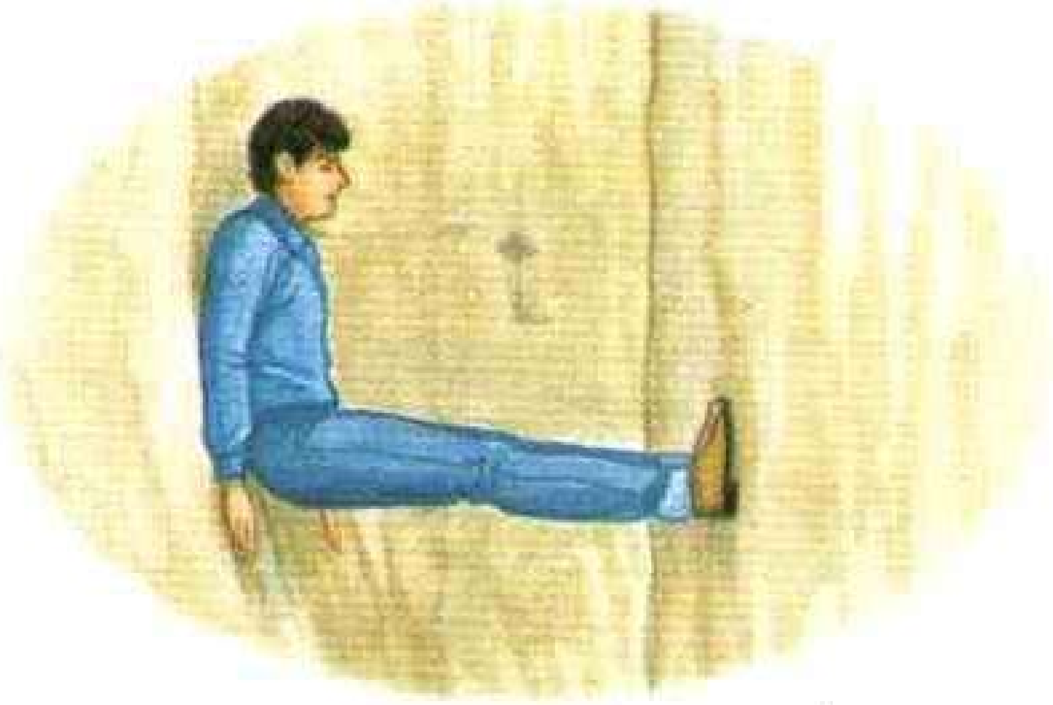

TÉCNICA DE CUBIERTA

Desde abajo, los resaltes en la pared de roca pueden parecer imponentes, en particular cuando no puede ver más allá de las presas. La técnica de cubierta se utiliza para superar estos resaltes. Únicamente debe intentarlo si no existe otra manera de superar un obstáculo. Es una técnica muy difícil, que requiere una gran condición física y habilidad.

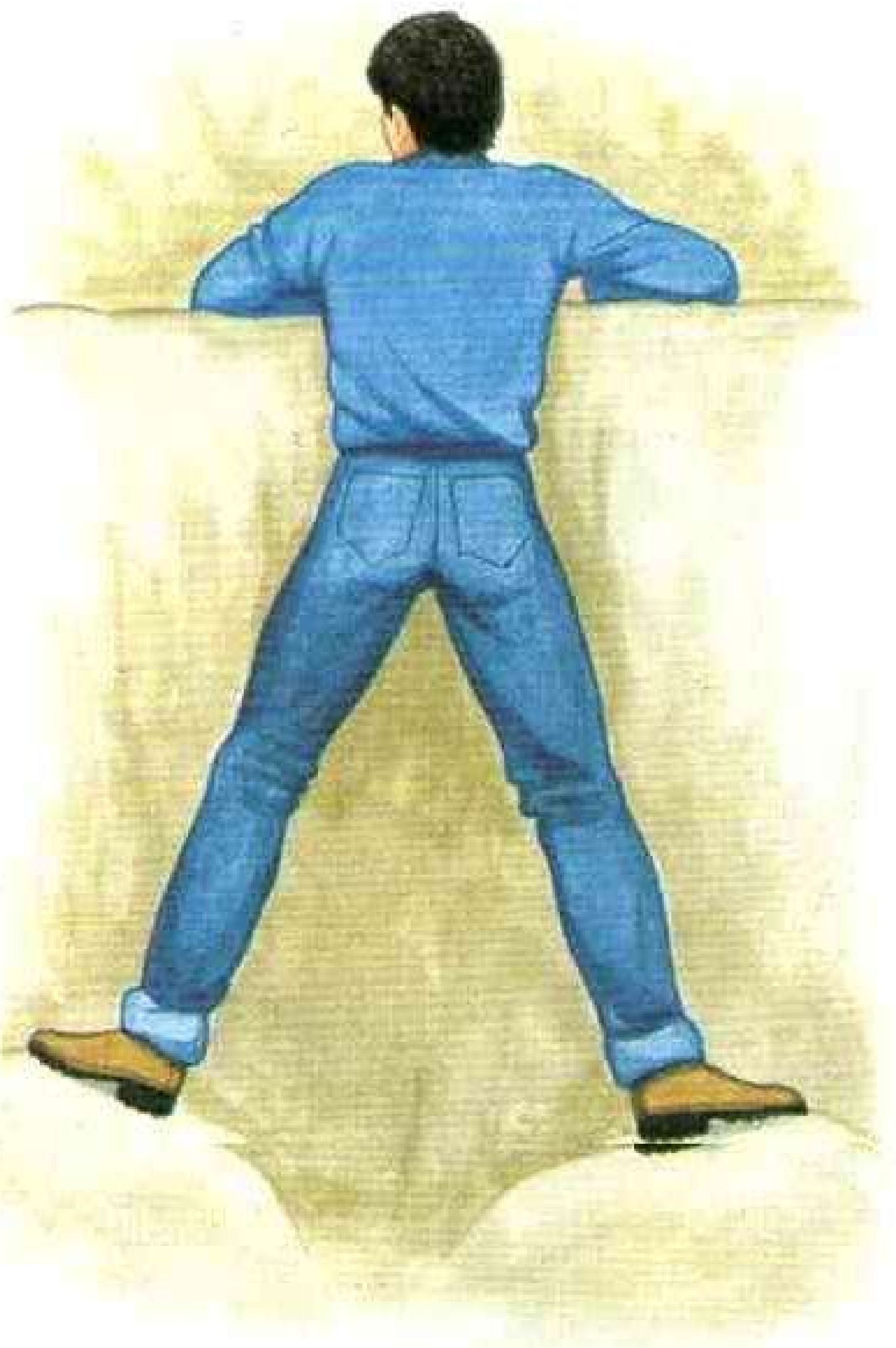

1 Escale hasta que sea capaz de colocar ambos codos sobre el resalte, soportando su peso en los pies.

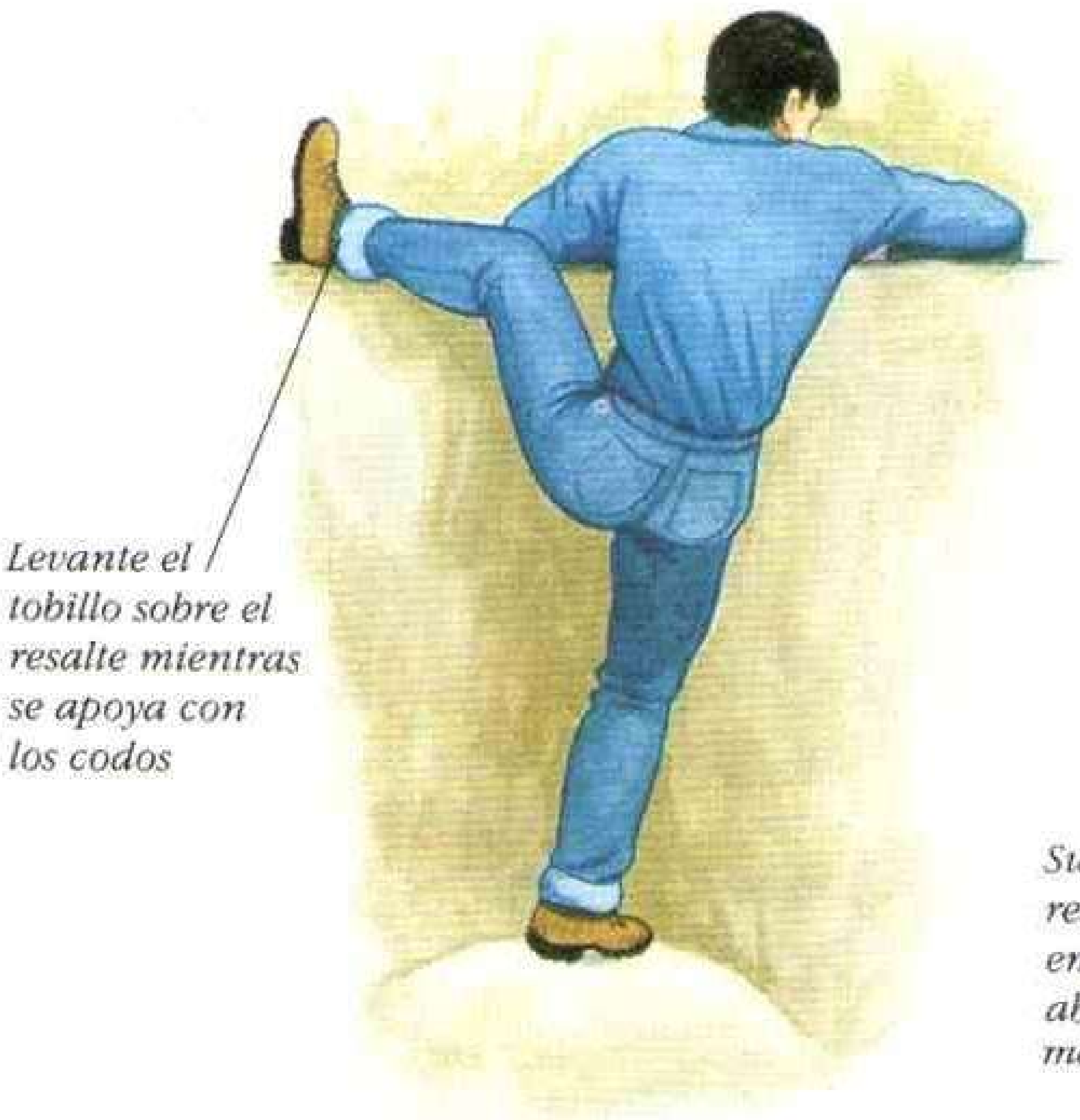

Levante el tobillo sobre el resalte mientras se apoya con los codos

2 Apóyese sobre un codo, manteniendo el pecho y la cabeza lo más cerca posible de la roca. Apoye un tobillo sobre el resalte.

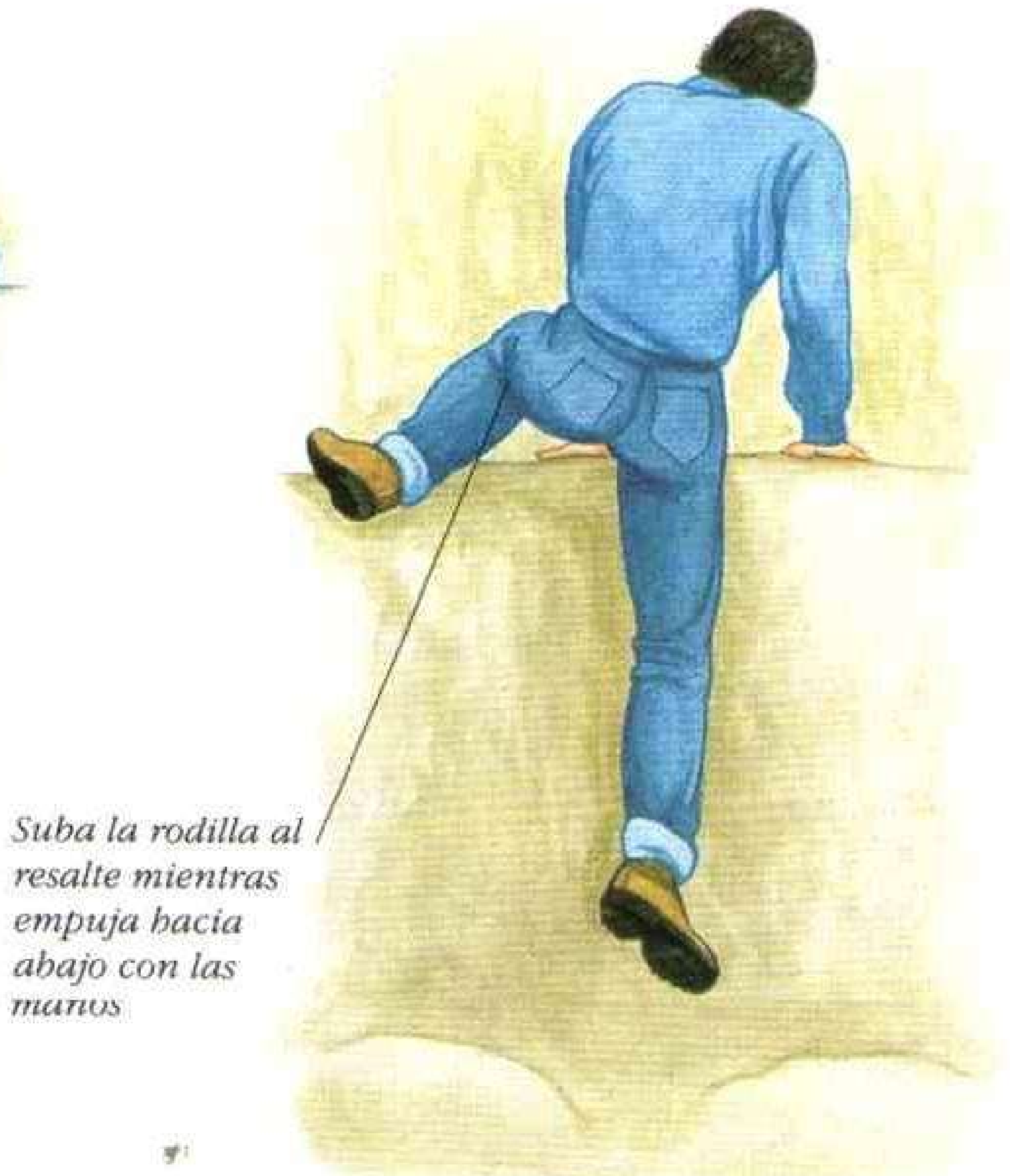

Suba la rodilla al resalte mientras empuja hacia abajo con las manos

3 Empujándose con ambas manos, mueva la rodilla hacia el resalte. Levántese con cuidado y suba al resalte.

NUDO PRUSSIK*

Un nudo prussik es un lazo de tipo corredizo que se aprieta bajo un peso, pero que puede desatarse fácilmente. Los nudos prussik constituyen una forma segura de atarse a una cuerda, y pueden utilizarse para trepar por una cuerda, quizá después de una caída, o para subir un saliente.

1 Haga un aro en su mano con una cuerda (aquí coloreada de amarillo) y dóblela sobre otra cuerda (coloreada en rojo). Pase ambos cabos de la cuerda amarilla a través del aro amarillo.

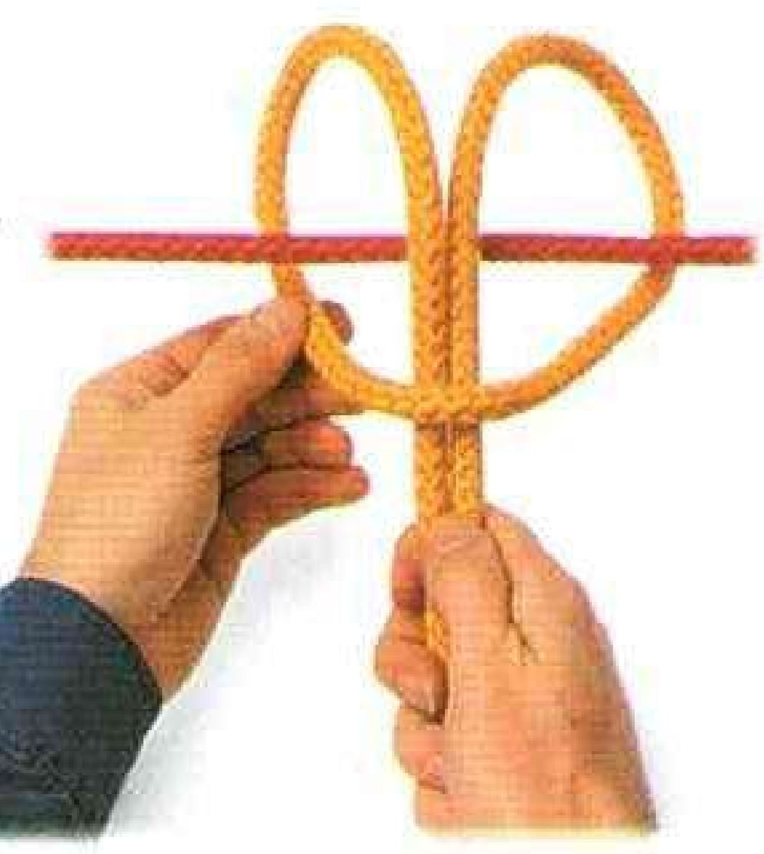

2 A continuación, enrolle la cuerda amarilla alrededor de la cuerda roja, igual que hizo en el paso 1, para obtener cuatro vueltas.

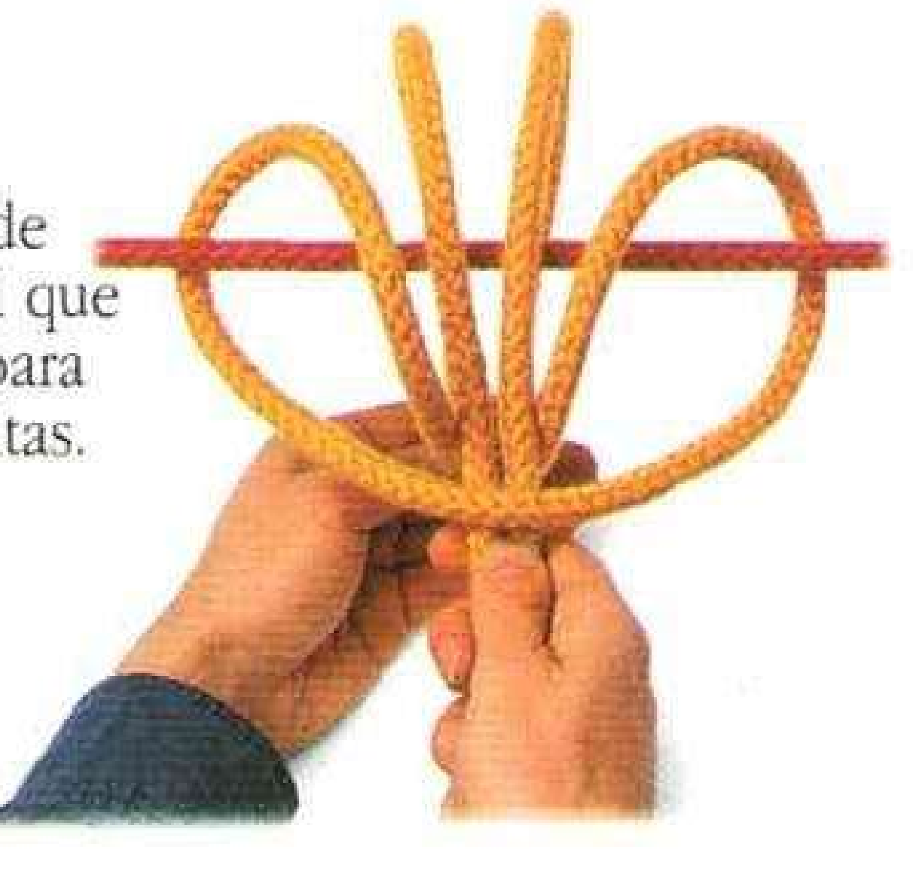

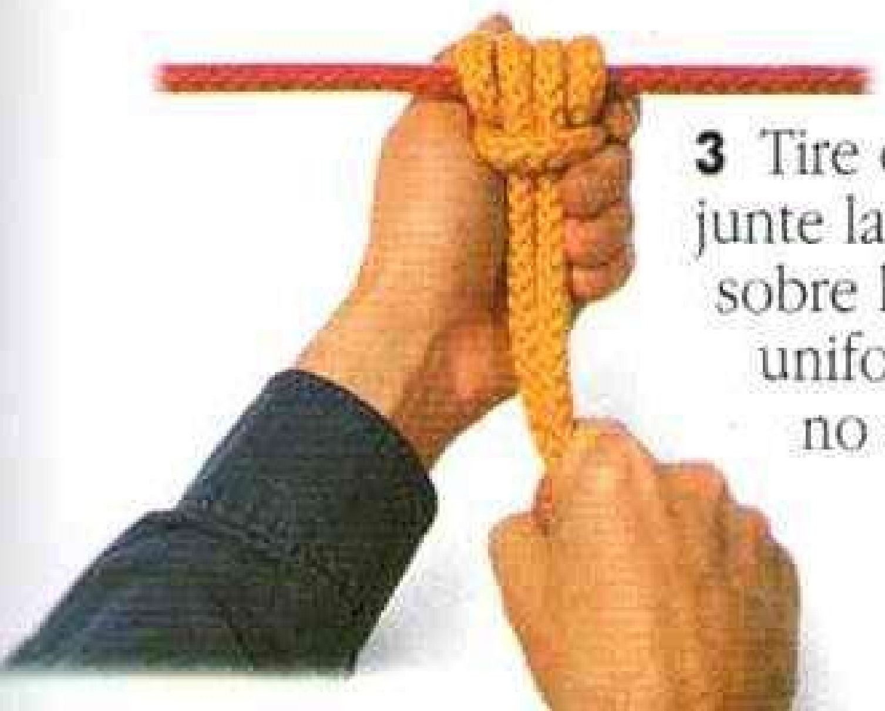

3 Tire de la cuerda amarilla y junte las vueltas para tensarlas sobre la cuerda roja y hacerlas uniformes. Tenga cuidado de no superponerlas.

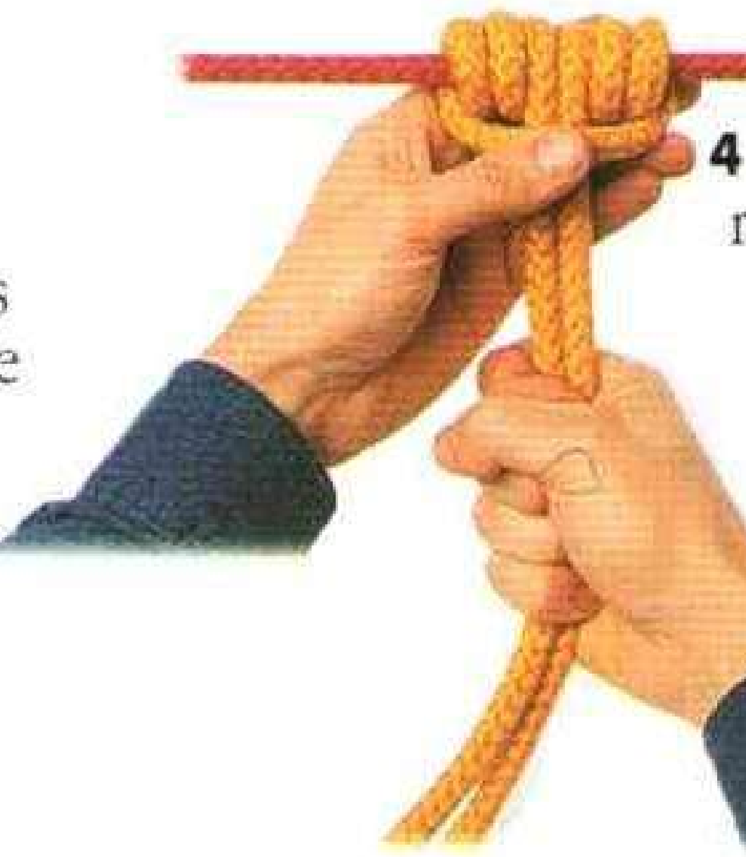

4 Para una mayor tirantez, repita los pasos 2 y 3 hasta obtener cuatro vueltas ajustadas alrededor de la cuerda roja. Retoque el nudo hasta que las vueltas sean uniformes y tensas.

MOSQUETÓN

Un mosquetón es un sistema de enganche útil para unir cuerdas o equipos *(véase página 152)*. Los mosquetones con abertura de muelle son más rápidos y fáciles de utilizar que aquéllos con cierre de seguridad (con tuerca), pero estos últimos son los más recomendables para el asegurado, el *rappel* y otras situaciones peligrosas *(véase página 147)*.

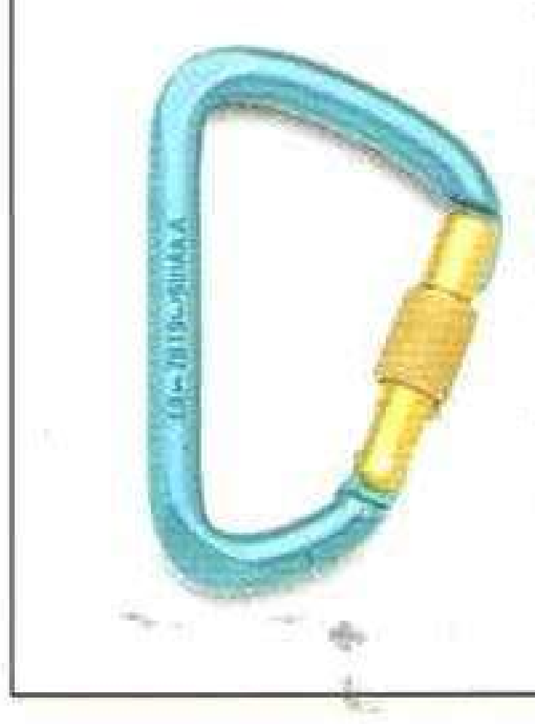

Mosquetón con cierre de seguridad. *Utilícelos para las técnicas que requieran una conexión muy segura, como el asegurado y el* rappel.

ESCALADA CON NUDOS PRUSSIK

Con dos cuerdas cortas y un arnés, puede ascender por una cuerda fija y salvar una situación difícil. En una emergencia, si no se dispone de un arnés apropiado, puede improvisar uno con cuerda. Haga un lazo en el que se pueda sentar y asegúrelo a una cuerda alrededor de su cintura con un nudo doble de tejedor *(véase página 184)*.

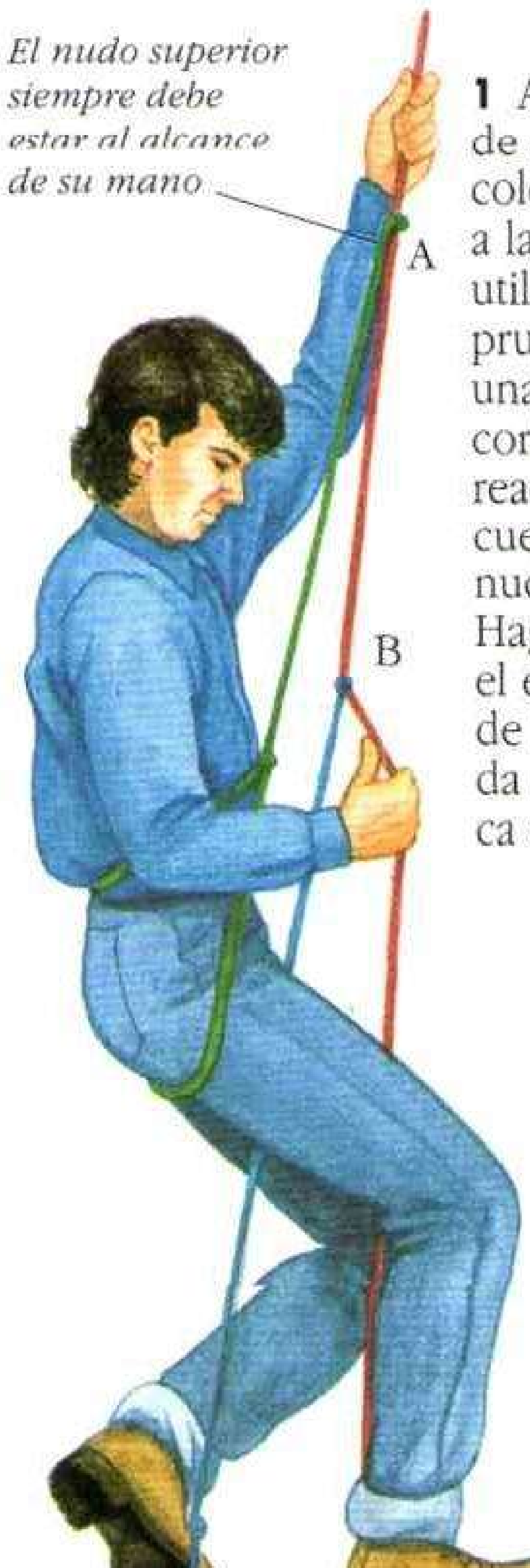

El nudo superior siempre debe estar al alcance de su mano

1 Asegure la cuerda de su arnés (aquí, coloreada de verde) a la cuerda fija, utilizando un nudo prussik (A). Sujete una segunda cuerda corta (aquí, coloreada de azul) a la cuerda fija con un nudo prussik (B). Haga una lazada en el extremo inferior de la segunda cuerda corta e introduzca en ella el pie.

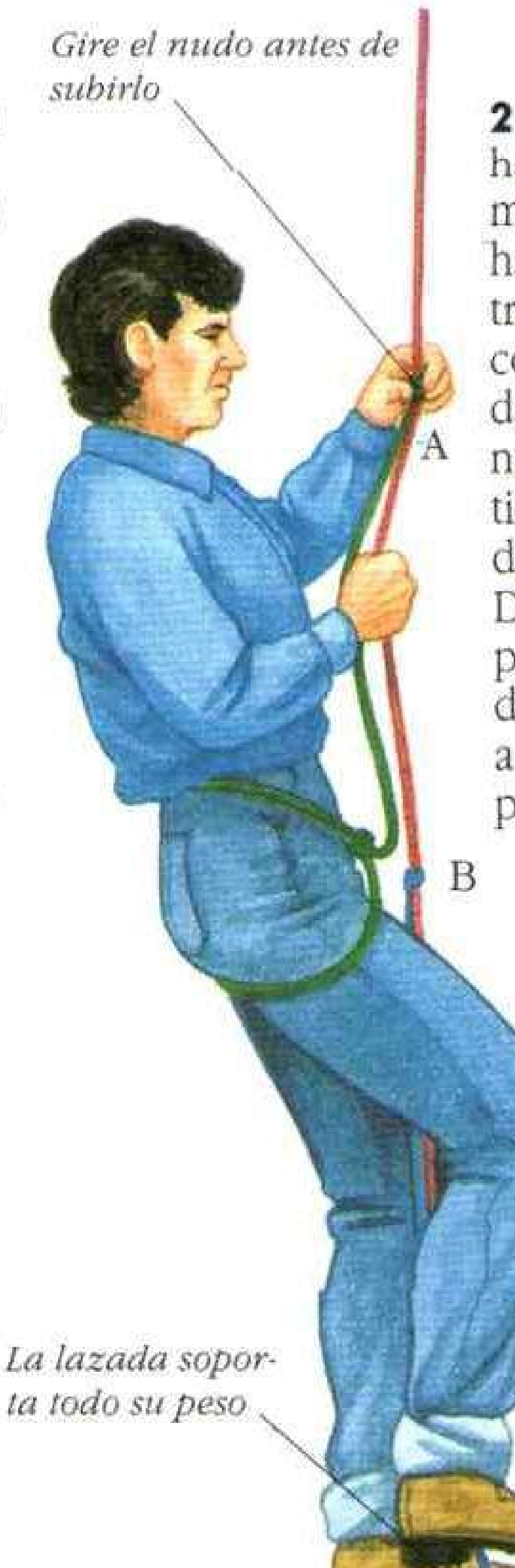

Gire el nudo antes de subirlo

La lazada soporta todo su peso

2 Para subir, empuje hacia arriba con los músculos de la pierna hasta que se encuentre en posición vertical con la pierna extendida. Para mover el nudo superior (A), tire de la cuerda fija debajo del nudo. Después gire el nudo para aflojarlo y deslícelo hacia arriba a la mayor distancia posible.

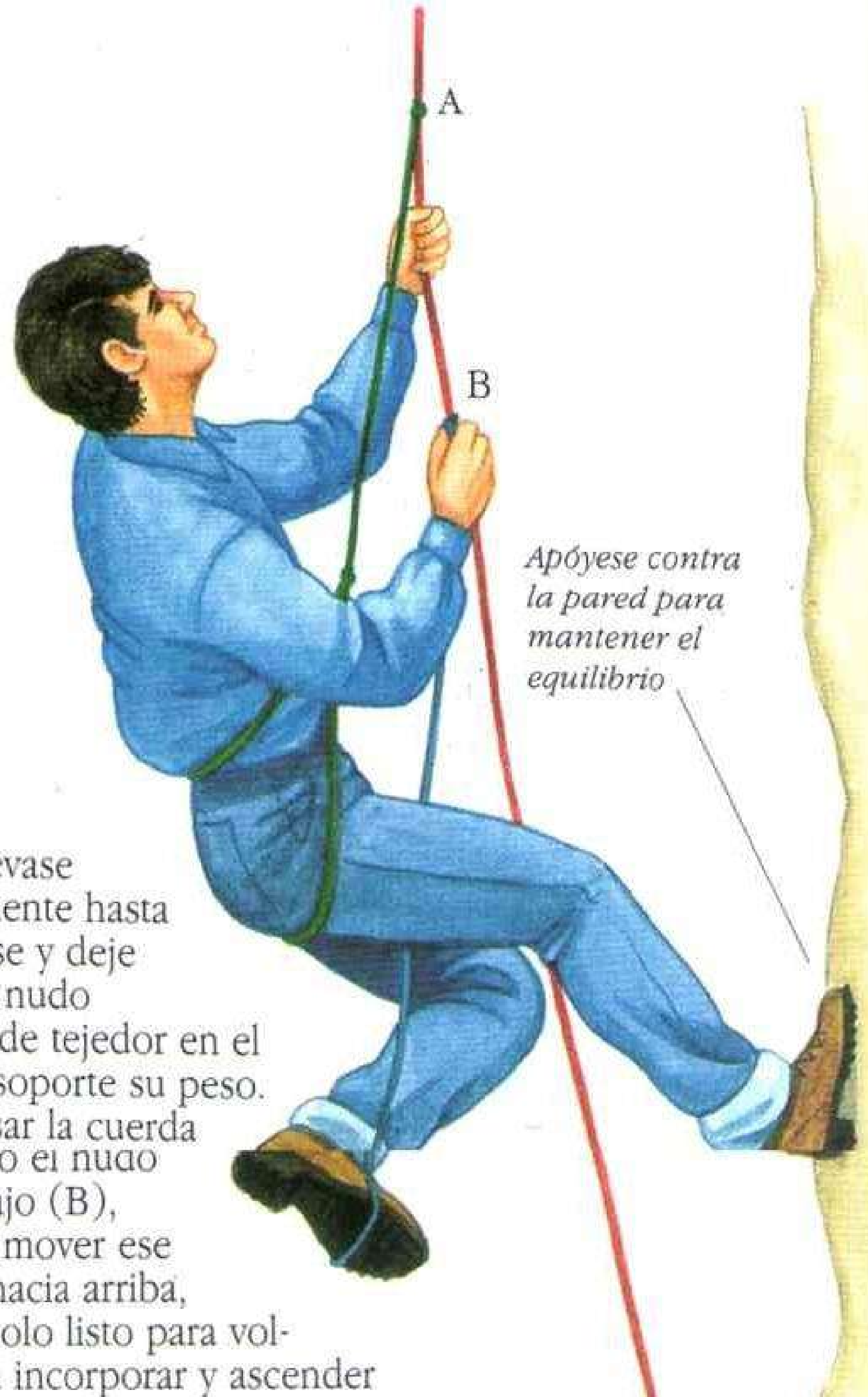

Apóyese contra la pared para mantener el equilibrio

3 Muévase lentamente hasta sentarse y deje que el nudo doble de tejedor en el arnés soporte su peso. Al tensar la cuerda fija bajo el nudo más bajo (B), puede mover ese nudo hacia arriba, dejándolo listo para volverse a incorporar y ascender un nuevo tramo.

---

* El prussik no es un nudo autobloqueante. De la misma manera que es necesario aflojarlo para subir, puede ser necesario apretarlo para que soporte el peso sin deslizarse hacia abajo. El diámetro de la cuerda que forme el prussik debería ser inferior al de la cuerda fija.

CRUZAR LAS AGUAS

Los ríos y las corrientes son siempre peligrosos, desde sus cabeceras, o fuentes, donde son rápidos, angostos y poco profundos, hasta las profundidades tranquilas de sus cuencas más bajas. Aun si el agua aparenta ser tranquila y de movimiento lento, poco profunda y segura, suponga siempre que hay algún peligro oculto. Ni en las aguas más claras podrá ver todo lo que hay debajo de la superficie. Las aguas de cabecera son más fáciles de cruzar que las aguas profundas, pero tenga cuidado al atravesar aguas rápidas. Las aguas tranquilas son generalmente más profundas que las rápidas, y pueden contener algas traicioneras, bancos de fango y otros obstáculos ocultos muy peligrosos. No vadee ni nade si existe alguna otra opción más segura, y pregúntese siempre si es realmente necesario cruzar.

DÓNDE CRUZAR

El lugar más indicado para cruzar cualquier corriente de agua es un puente, un pontón o un transbordador. Así pues, antes de mojarse, compruebe las inmediaciones, tanto río arriba como río abajo. Es posible que encuentre un puente, o un tramo ancho y regular donde el lecho del río sea firme, y por donde pueda cruzar con toda seguridad. Compruebe que la orilla opuesta no sea demasiado escarpada.

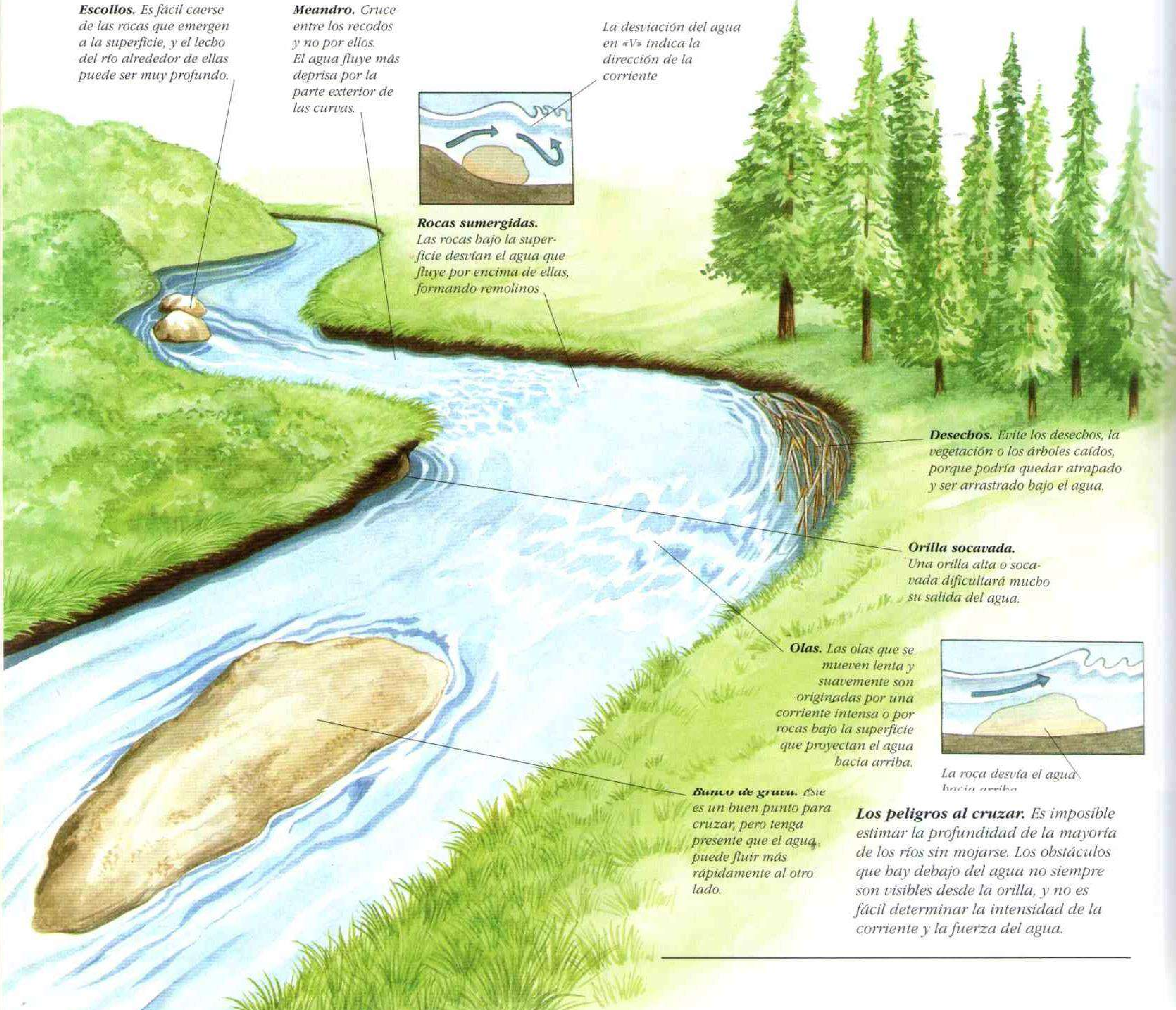

Escollos. *Es fácil caerse de las rocas que emergen a la superficie, y el lecho del río alrededor de ellas puede ser muy profundo.*

Meandro. *Cruce entre los recodos y no por ellos. El agua fluye más deprisa por la parte exterior de las curvas.*

La desviación del agua en «V» indica la dirección de la corriente

Rocas sumergidas. *Las rocas bajo la superficie desvían el agua que fluye por encima de ellas, formando remolinos*

Desechos. *Evite los desechos, la vegetación o los árboles caídos, porque podría quedar atrapado y ser arrastrado bajo el agua.*

Orilla socavada. *Una orilla alta o socavada dificultará mucho su salida del agua.*

Olas. *Las olas que se mueven lenta y suavemente son originadas por una corriente intensa o por rocas bajo la superficie que proyectan el agua hacia arriba.*

La roca desvía el agua hacia arriba

Banco de grava. *Éste es un buen punto para cruzar, pero tenga presente que el agua puede fluir más rápidamente al otro lado.*

Los peligros al cruzar. *Es imposible estimar la profundidad de la mayoría de los ríos sin mojarse. Los obstáculos que hay debajo del agua no siempre son visibles desde la orilla, y no es fácil determinar la intensidad de la corriente y la fuerza del agua.*

VADEAR

Lleve algún tipo de calzado para proteger sus pies y tener un mejor punto de apoyo. Esté preparado para afrontar un fondo lodoso, con vegetación, o un cambio repentino en la profundidad del río. Observe el comportamiento de las olas y si hay algún signo de obstrucción bajo el agua *(véase página anterior)*. Cruce siempre muy despacio.

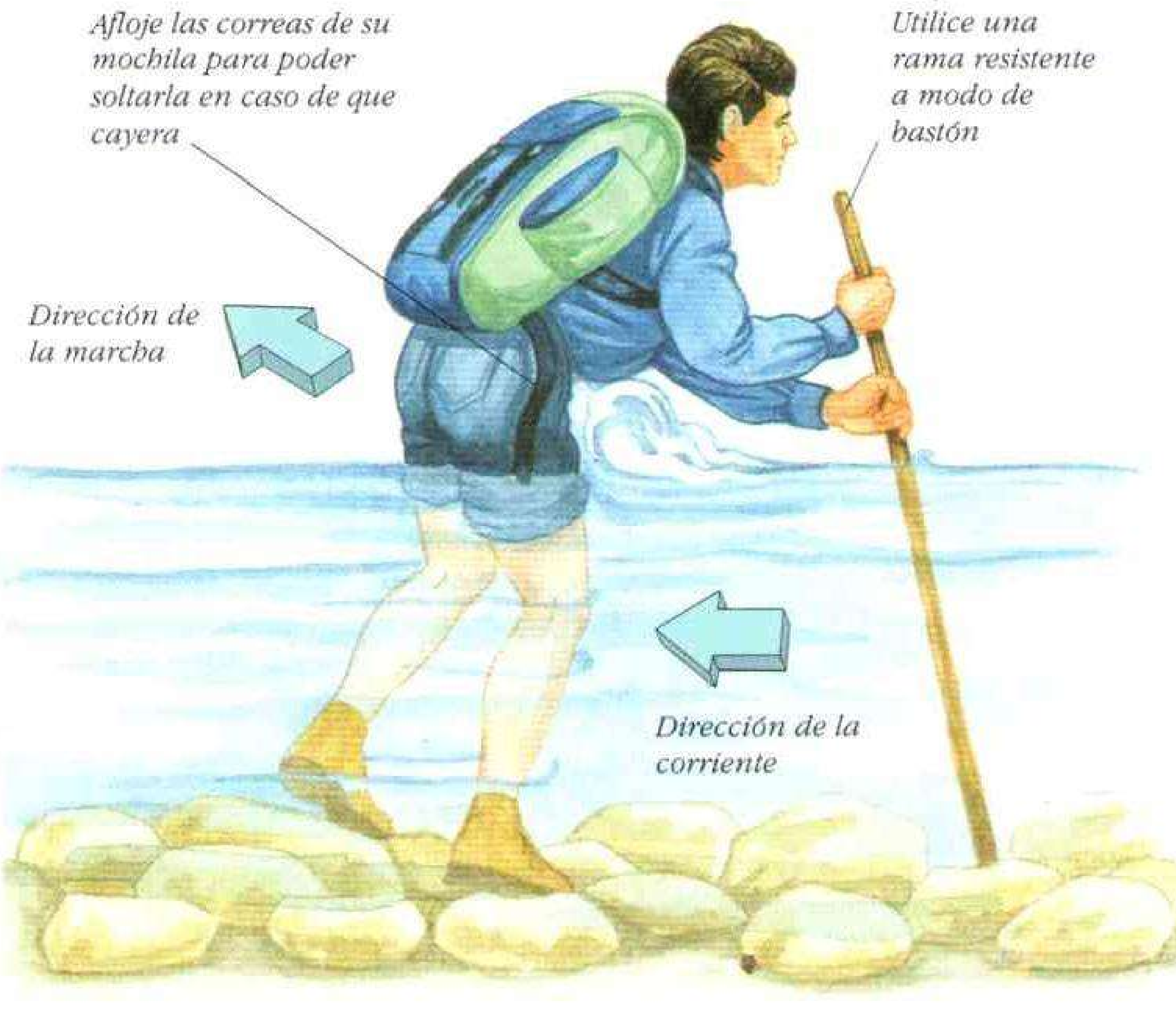
Afloje las correas de su mochila para poder soltarla en caso de que cayera
Utilice una rama resistente a modo de bastón
Dirección de la marcha
Dirección de la corriente

Cruzar solo. *Utilice un palo a modo de sonda y, una vez en el agua, como si se tratase de una tercera pierna, para mantener una forma de «T» extendida, como un trípode. Coloque el bastón corriente arriba y apóyese en él al dar un paso lateral con respecto a la corriente, volviendo a pisar firmemente el lecho del río. Dé pasos cortos, arrastrando los pies, para asegurar que la corriente no empuje su pierna hacia atrás y le haga caer.*

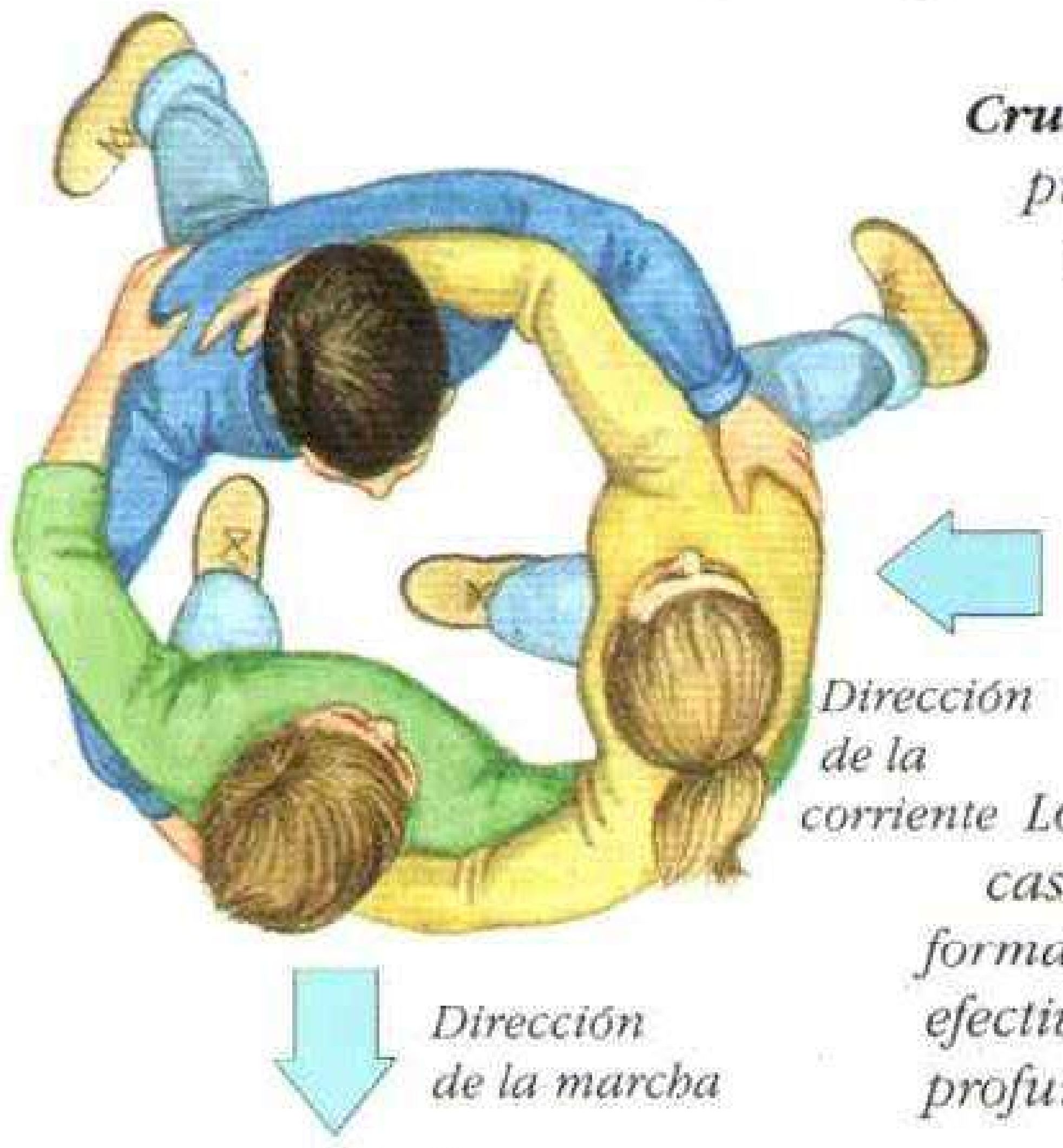

Cruzar en grupo. *Tres personas pueden formar una especie de trípode para cruzar un río. Deben entrelazar los brazos con fuerza e inclinarse hacia el centro, doblando la cintura. La persona más fuerte debe situarse corriente arriba y debe dar siempre el primer paso. Los demás deben sujetarle en caso de que caiga. Ésta es una formación muy estable y muy efectiva en aguas rápidas y poco profundas.*

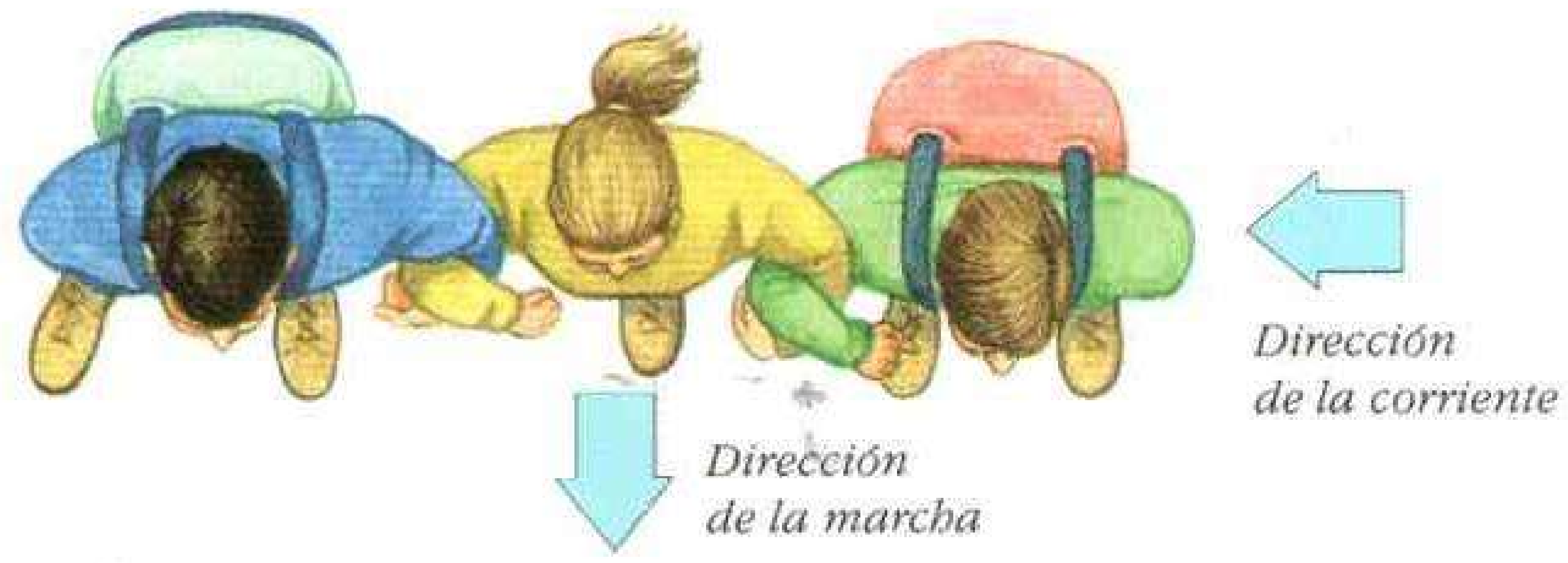

Cruzar en línea. *Un grupo de personas puede cruzar la corriente en línea. La persona más fuerte debe colocarse corriente arriba, y las restantes, dando estabilidad y sujetando a cualquiera que pudiera caer. El líder debe decidir por dónde cruzar y dar los primeros pasos. Los demás deben entrelazar sus brazos con los de él; la persona más débil y menos pesada debe colocarse en el centro. Hay que cruzar despacio y con cuidado. Debe mantener el equilibrio y moverse paso a paso.*

NADAR

Si el agua es muy profunda y no puede vadearla, será necesario cruzar a nado. Fabrique un flotador que le ayude. Antes de entrar en el agua, busque un sitio adecuado para desembarcar en la orilla opuesta.

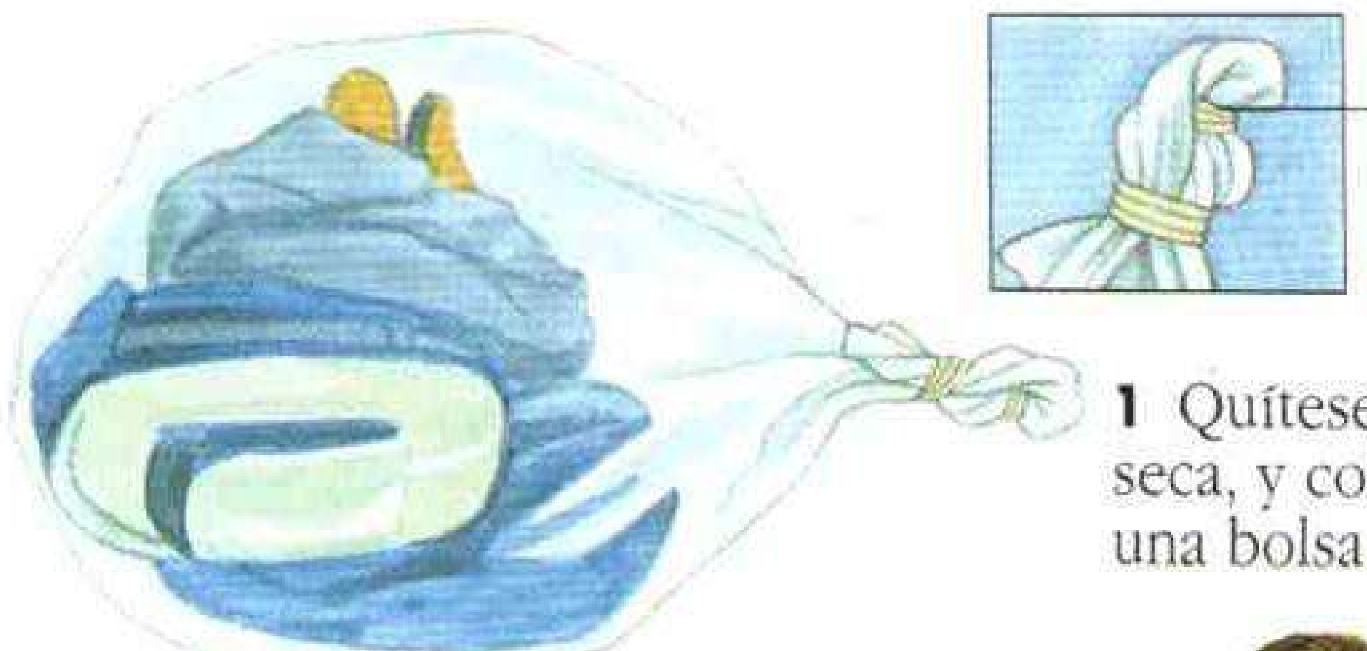

Retuerza la parte superior de la bolsa dejándola bien apretada, dóblela sobre sí misma y átela con fuerza

1 Quítese la ropa para que se mantenga seca, y colóquela junto con su equipo en una bolsa impermeable.

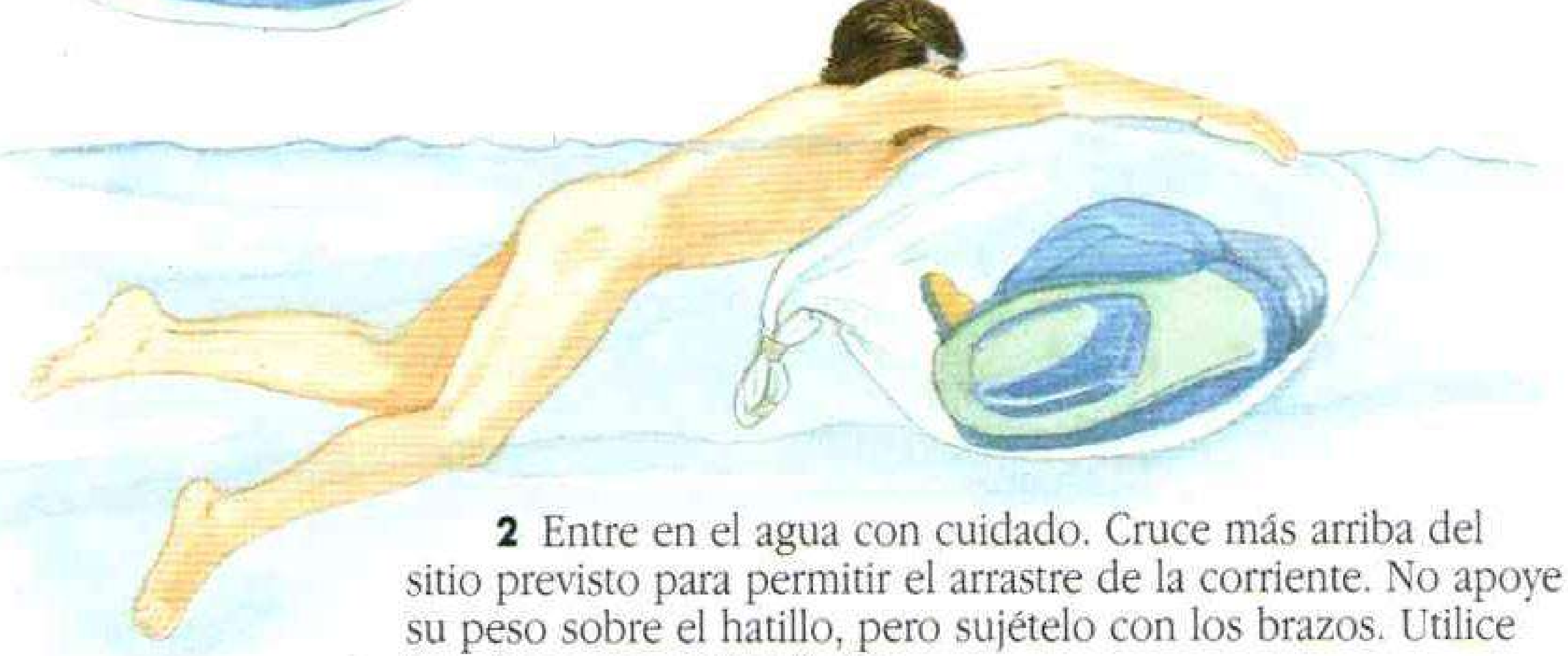

2 Entre en el agua con cuidado. Cruce más arriba del sitio previsto para permitir el arrastre de la corriente. No apoye su peso sobre el hatillo, pero sujételo con los brazos. Utilice las piernas para impulsarse.

HIPOTERMIA

Desnúdese (salvo el calzado) antes de entrar en el agua. Evalúe la conveniencia de llevar alguna prenda impermeable o resistente al viento para darle algo de protección. El frío le producirá letargo. Inmediatamente después de cruzar, séquese y vístase con ropa de abrigo seca y, si es posible, prepárese una bebida caliente y azucarada. Auxilie a quien esté temblando. Trabaje en parejas. El líder debe comprobar si en el grupo hay síntomas de hipotermia, y debe estar preparado para dar el tratamiento adecuado. Si quien la sufre no recupera el calor inmediatamente, puede llegar a morir *(véase página 163)*. La mejor manera de proporcionar calor a una persona es metiéndola en un saco de dormir con otra persona. Puede darse a la víctima una bebida caliente y ropa de abrigo, además de palabras tranquilizadoras, para que recupere el calor corporal. Los síntomas de hipotermia aparecen habitualmente en el siguiente orden:

- Temblores, carne de gallina, piel pálida y entumecida.
- Apatía, confusión, comportamiento irracional, amnesia, incoherencia y agresividad.
- Letargo seguido de arranques de energía frenética.
- Lapsos de consciencia con respiración lenta y poco profunda, arritmia cardíaca.
- Pulso lento y débil. Puede llegar a producir un paro cardíaco *(véase página 176)*.

AGUAS PELIGROSAS

CRUZAR CUALQUIER TIPO de corriente de agua debe considerarse una acción peligrosa. Las fluctuaciones en las precipitaciones pueden cambiar la seguridad de cualquier punto de cruce. La única manera de evaluar la profundidad y la intensidad de la corriente es enviando a una persona a cruzarla primero, sujetándola con una cuerda de salvamento atada a la orilla y controlada por un miembro del grupo. Si se ve obligado a nadar, utilice siempre un flotador, y en los rápidos, zonas rocosas o de aguas muy turbulentas, lleve un casco protector de color brillante.

EL EMPLEO DE CUERDAS DE SEGURO

Si está solo y tiene que cruzar un río peligroso, únicamente puede emplear una cuerda de seguro pasándola alrededor de una roca o árbol y sujetándola con nudos fáciles de aflojar. Con dos o más personas pueden utilizarse las cuerdas de seguro para disminuir los posibles peligros. Utilice siempre una cuerda de seguro, y asegúrese de que las personas que estén en el agua no se enreden en la parte floja de la cuerda.

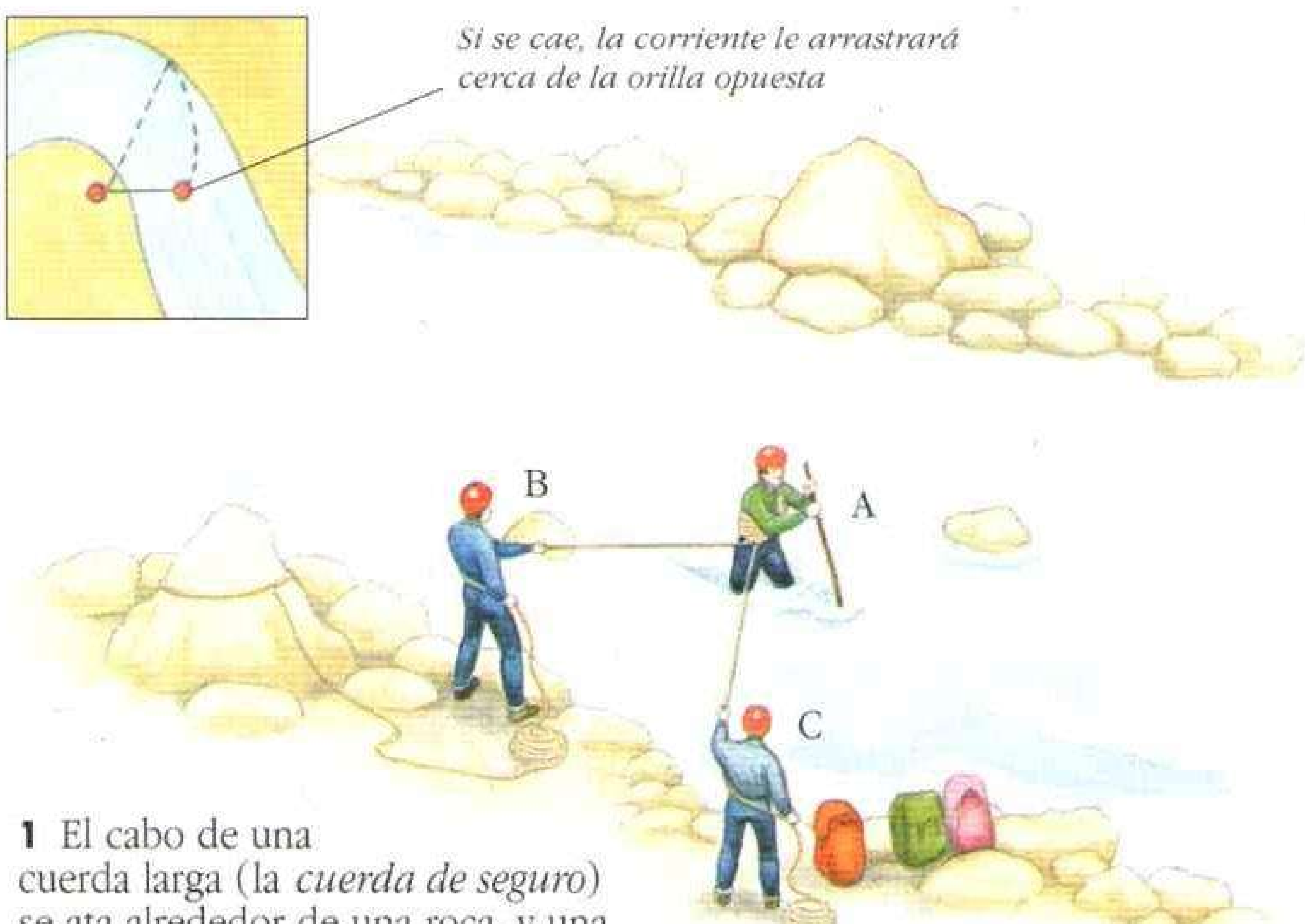
Si se cae, la corriente le arrastrará cerca de la orilla opuesta
B
A
C

1 El cabo de una cuerda larga (la *cuerda de seguro*) se ata alrededor de una roca, y una persona (B) la va soltando gradualmente. La persona más fuerte (A) toma el otro cabo de la cuerda y cruza el río. Asimismo, lleva consigo el cabo de una segunda cuerda (la *cuerda de cruce*), cuyo otro extremo está sujeto por una tercera persona (C). Esta cuerda lleva un mosquetón en su centro *(véase página 147)*. La persona A lleva un casco y botas, y utiliza un palo para sondear el camino.

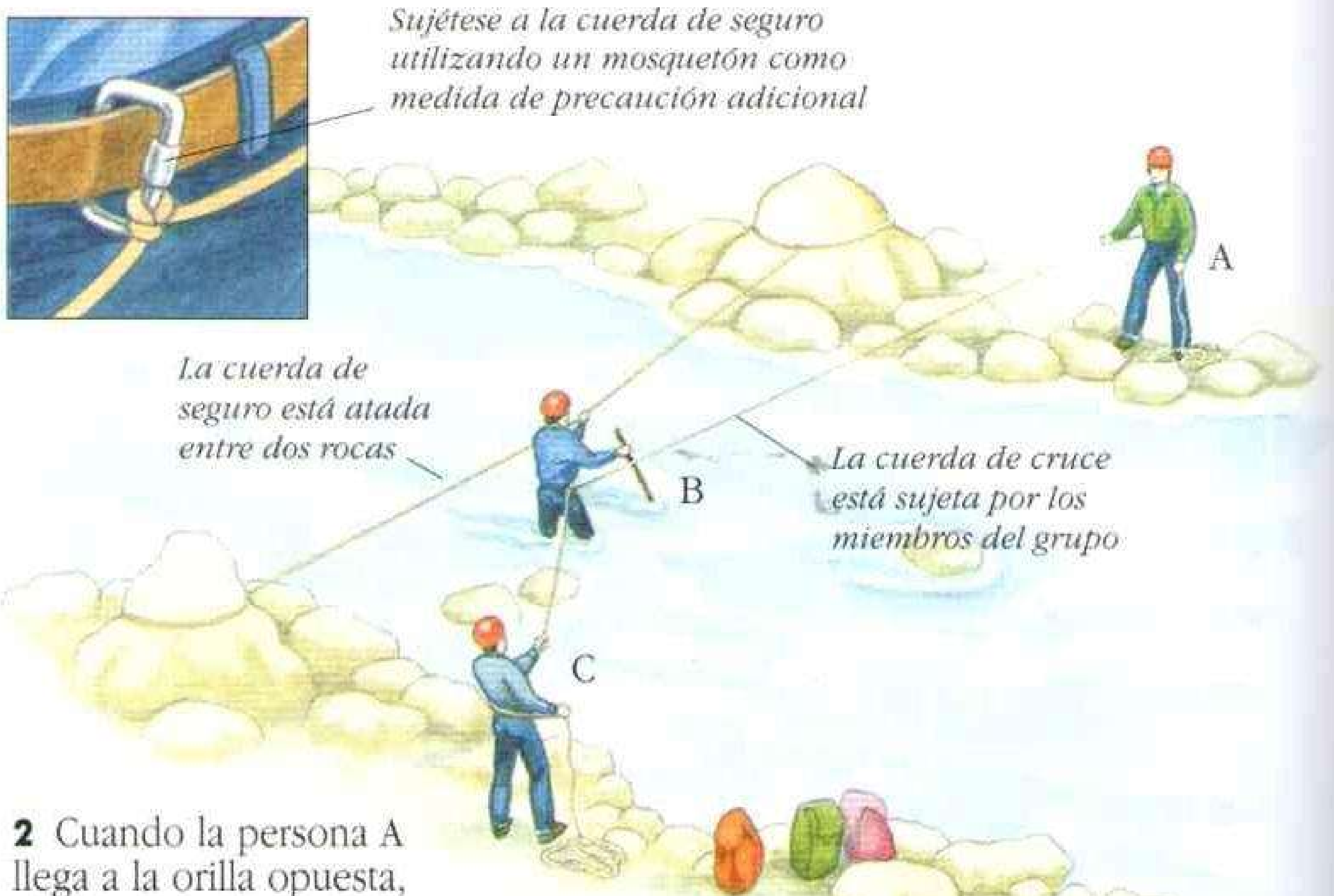
Sujétese a la cuerda de seguro utilizando un mosquetón como medida de precaución adicional
A
La cuerda de seguro está atada entre dos rocas
B
La cuerda de cruce está sujeta por los miembros del grupo
C

2 Cuando la persona A llega a la orilla opuesta, ata el cabo de la cuerda de seguro a una roca. Sujeta el mosquetón a la cuerda de seguro y ata el cabo de la cuerda de cruce alrededor de sí mismo para sujetar a la siguiente persona en cruzar. El mosquetón se lleva a la primera orilla, y la persona B lo sujeta a su cinturón. Agarrándose a la cuerda de seguro, cruza lentamente en el lado que da río arriba, utilizando el palo como sonda. La persona C va soltando la cuerda de cruce, mientras que la persona A va tirando del otro lado.

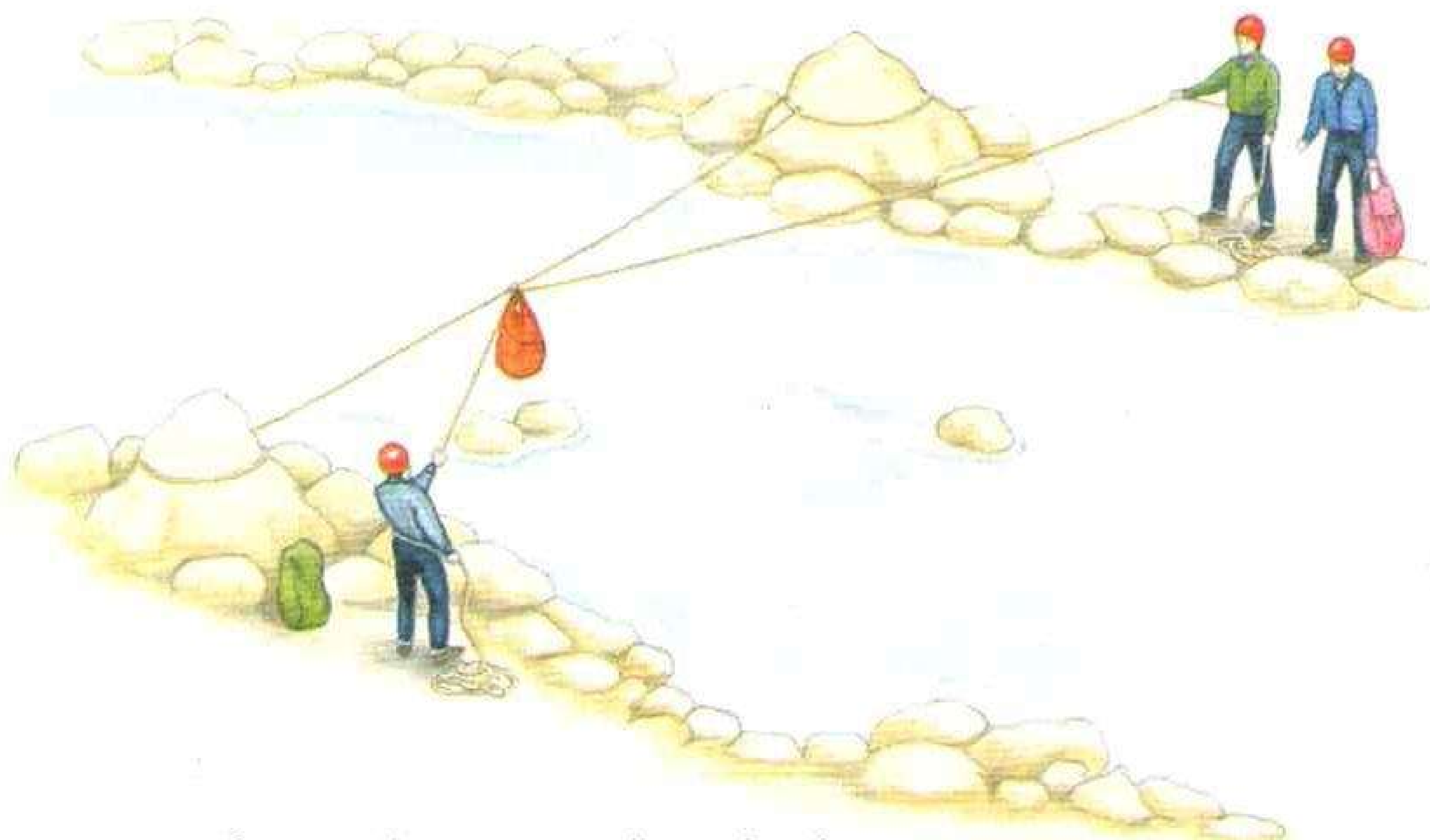

3 Conforme cada persona va llegando a la orilla opuesta, desprende el mosquetón de su cinturón y lo sujeta a la cuerda de seguro. A continuación se recoge la cuerda de cruce hacia la orilla de partida. Las mochilas se llevan a la orilla opuesta de la misma manera, sujetando las correas al mosquetón. Es posible que tenga que tensar la cuerda de seguro para evitar que las mochilas se mojen.

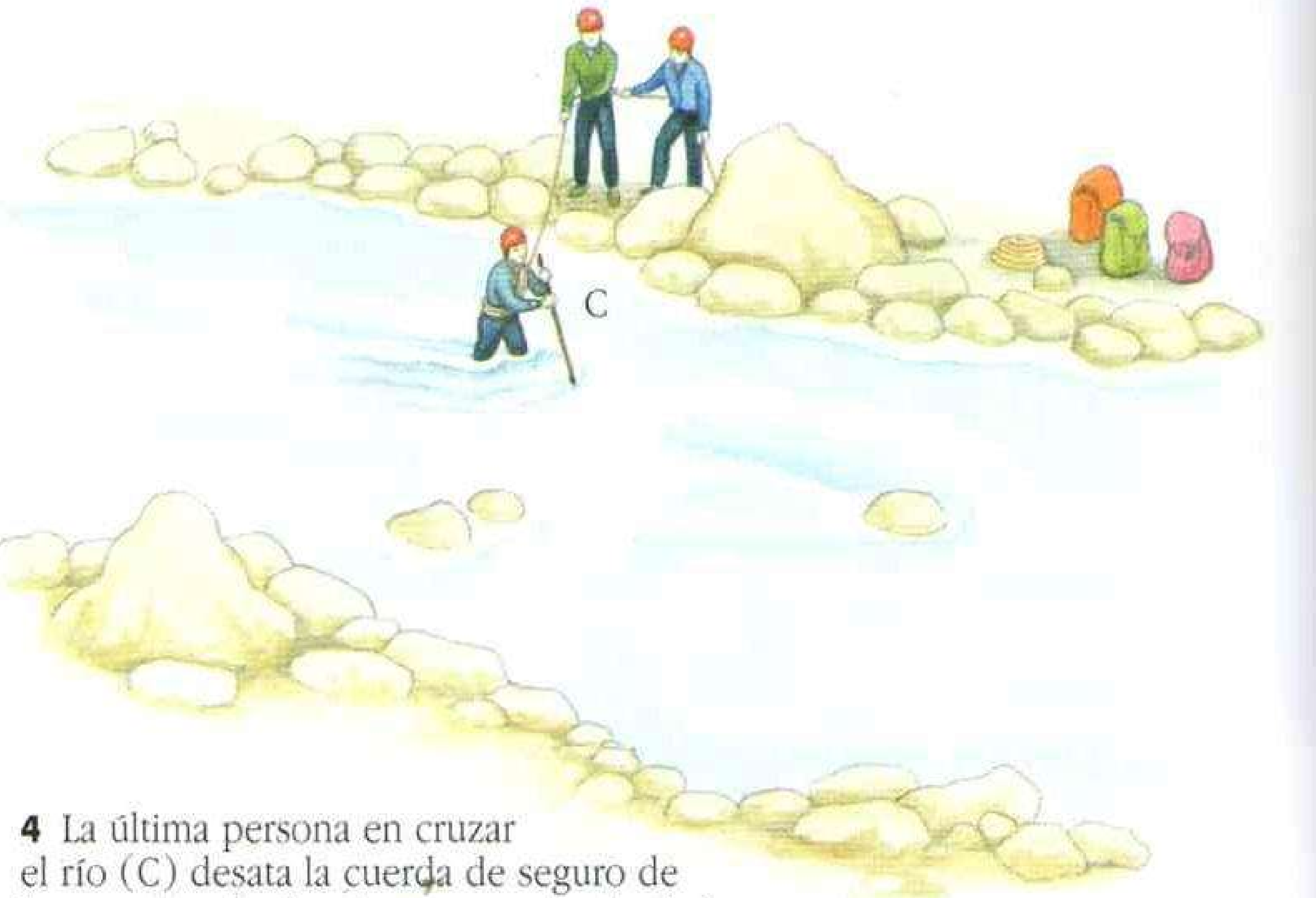

4 La última persona en cruzar el río (C) desata la cuerda de seguro de la roca. Ata el cabo de esta cuerda alrededor de sí mismo y cruza el río, tanteando el terreno con un palo. Los miembros restantes del grupo sujetan la cuerda de cruce, preparados para una eventual caída.

CÓMO CONSTRUIR UN PUENTE

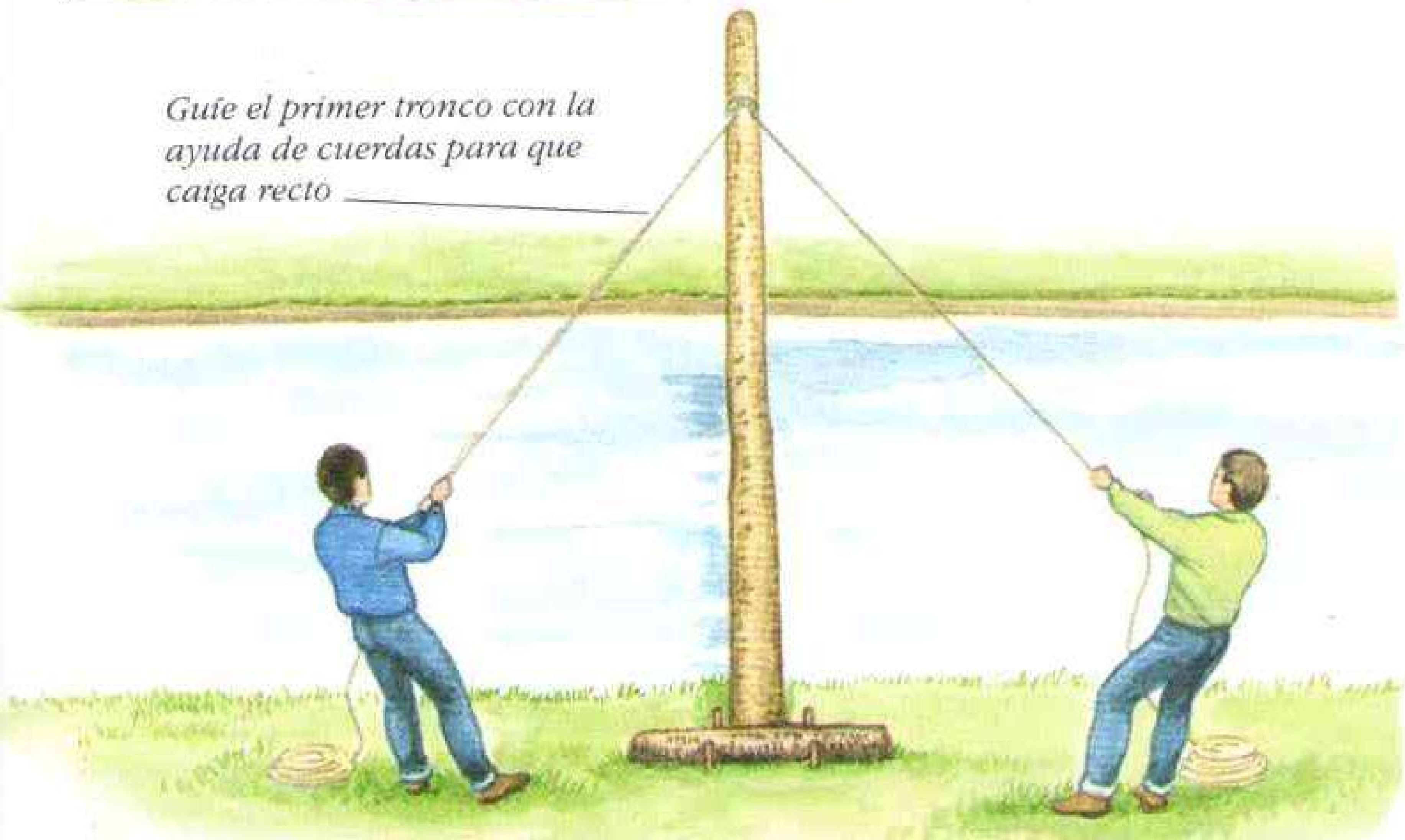
Guíe el primer tronco con la ayuda de cuerdas para que caiga recto

1 Para construir un puente, coloque primero un tronco largo entre las dos orillas con la ayuda de dos cuerdas. El extremo del tronco que permanece en la primera orilla debe asegurarse a otro tronco más corto, que se mantiene en su sitio gracias a unas estacas.

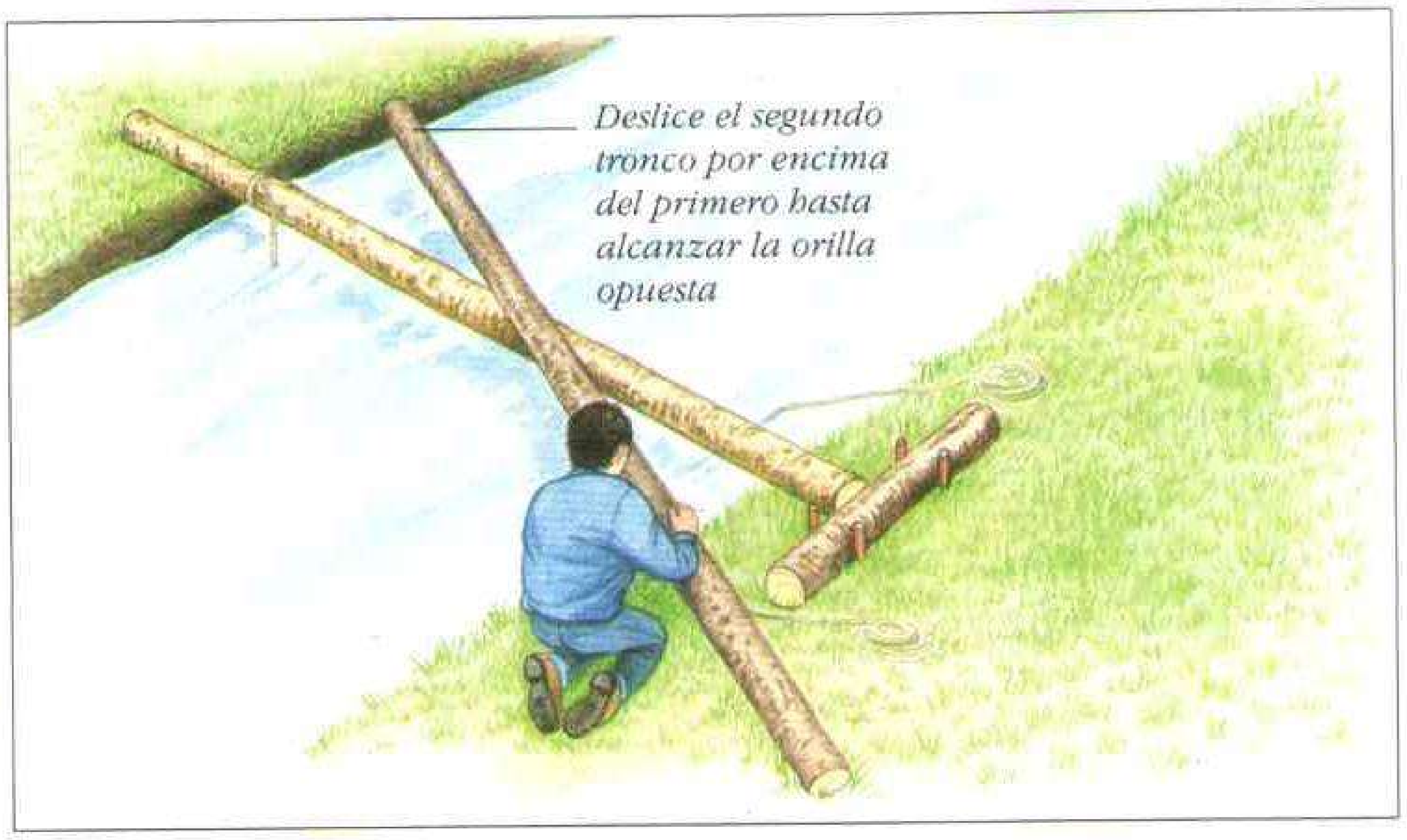
Deslice el segundo tronco por encima del primero hasta alcanzar la orilla opuesta

2 Una vez ha colocado el primer tronco, deslice un segundo tronco por encima del primero hasta que el extremo alcance la orilla opuesta. Haga caer el segundo tronco junto al primero y manténgalos juntos.

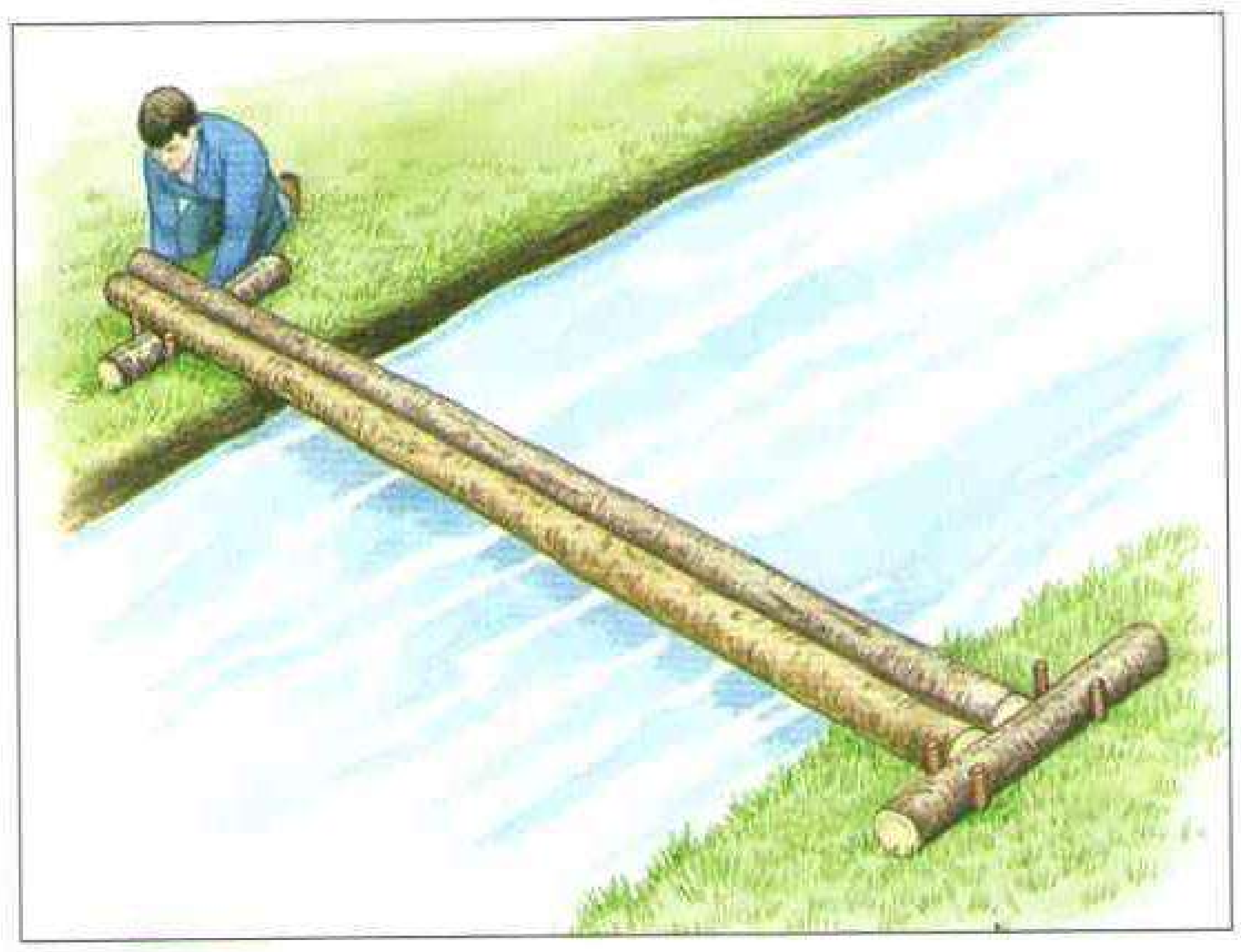

3 Cruce al otro lado sobre los dos troncos y coloque un tronco más corto bajo sus extremos para evitar que se entierren en el suelo. Fije los troncos con estacas para que no se separen.

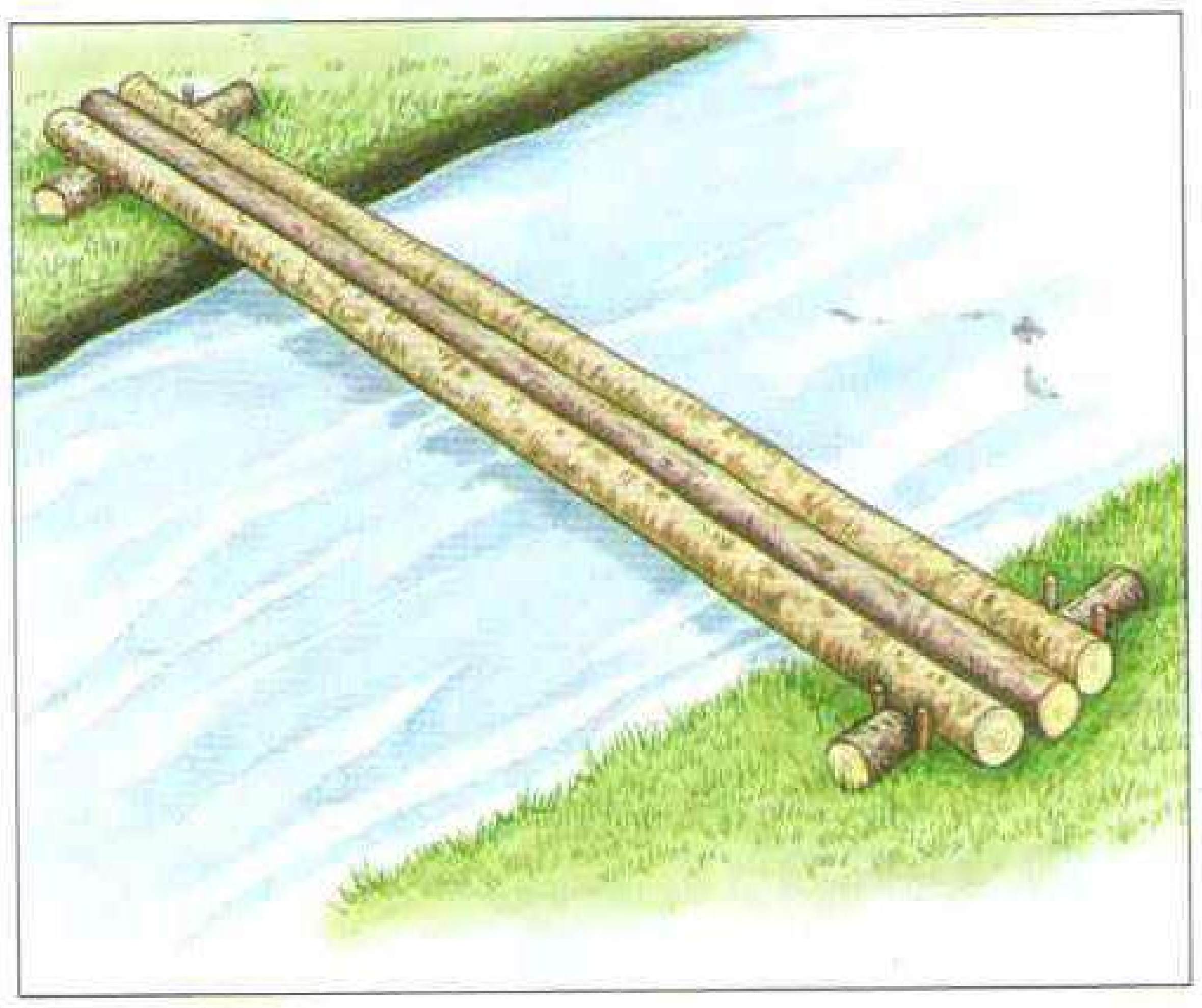

4 Deslice un tercer tronco junto a los otros dos. Coloque el tronco corto bajo sus extremos en la primera orilla y fíjelos con estacas como hizo en el paso 3. Puede hacer un puente tan ancho como desee, aunque con tres troncos suele ser suficiente.

CÓMO CRUZAR UN TERRENO PANTANOSO

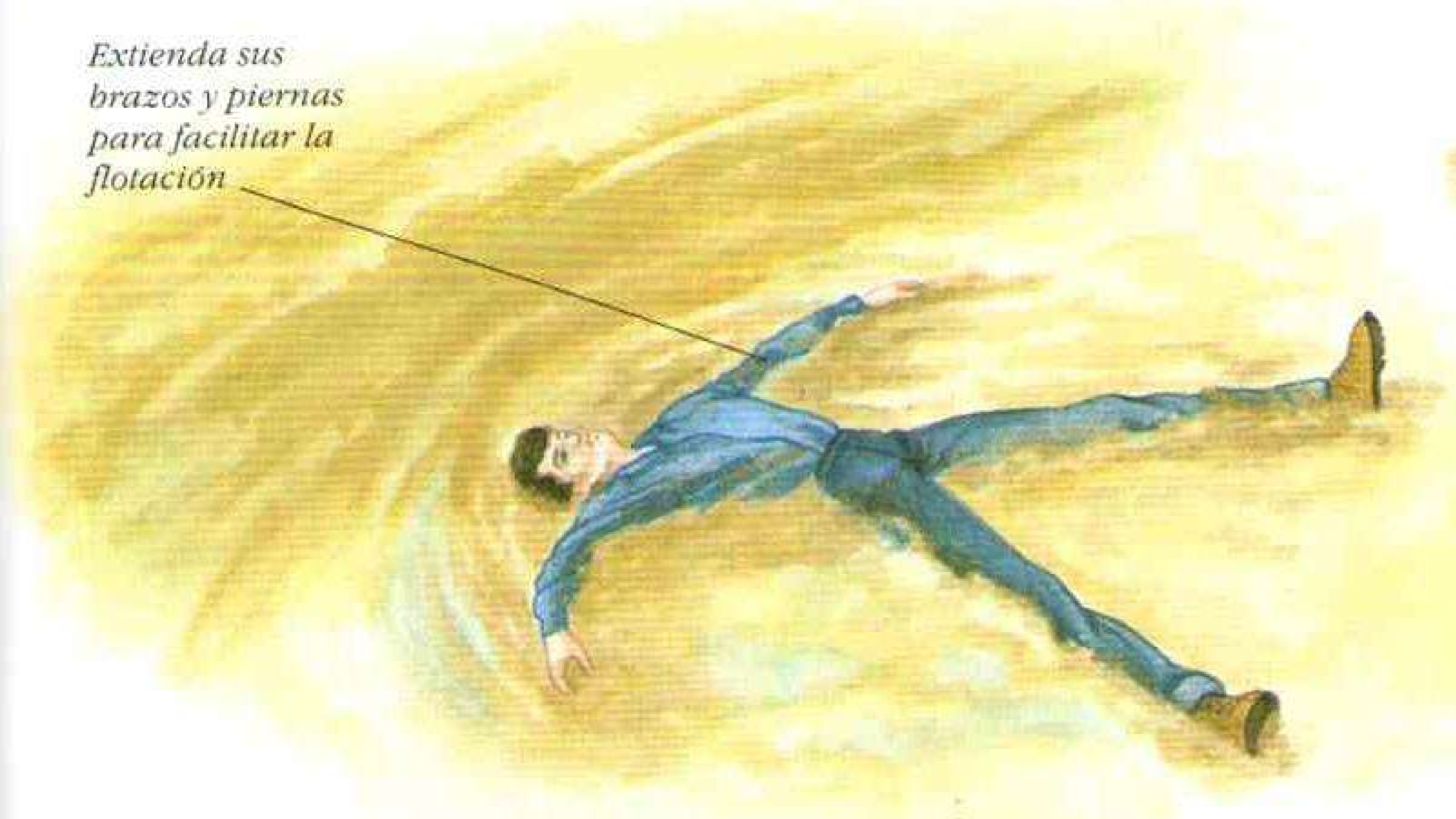
Extienda sus brazos y piernas para facilitar la flotación

Si está solo. *Las arenas movedizas y los pantanos no son más que suelos con agua estancada, por lo que si cae en ellos debe ser capaz de flotar o cruzarlos a nado, con tal de que logre repartir su peso sobre la mayor superficie posible y así evitar hundirse. Trate primero de cruzarlos corriendo, pisoteando las matas de hierba. Si comienza a hundirse, suelte la mochila y flote sobre su espalda. Extienda los brazos y las piernas y nade lentamente hacia la orilla* (véase página 163). *No se asuste ni forcejee.*

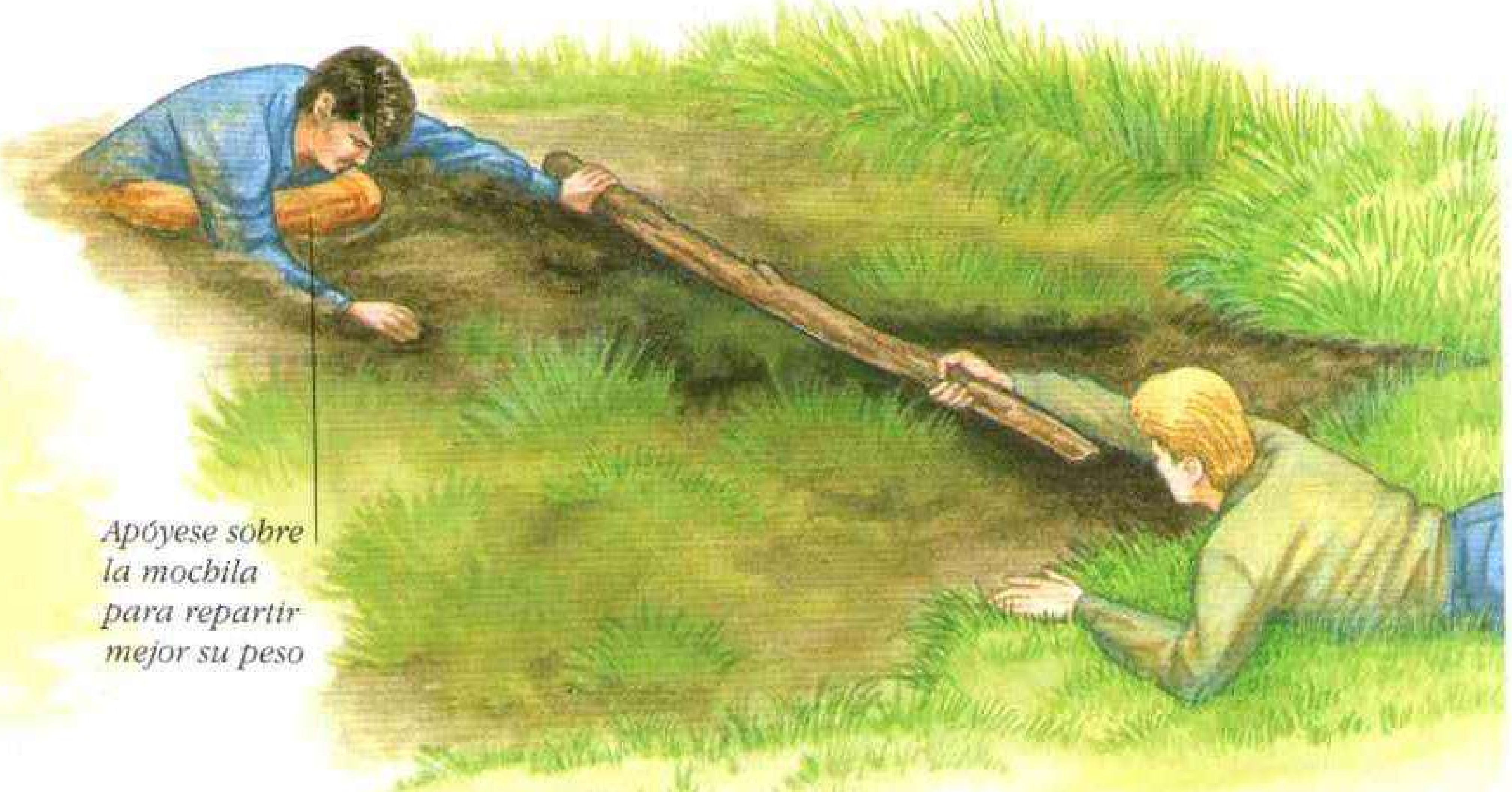
Apóyese sobre la mochila para repartir mejor su peso

Con ayuda. *Si no está solo cuando comienza a hundirse en las arenas movedizas o en un pantano, otra persona puede recostarse a orillas de la arena, pasarle el extremo de una rama larga y ayudarle a salir tirando de ella. Nuevamente, es importante que no forcejee ni se asuste, ya que los movimientos rápidos agitarán la superficie y le harán hundirse más rápidamente que si se limita a flotar o a moverse lentamente. Si se apoya sobre su mochila, repartirá su peso sobre una mayor superficie y evitará hundirse más.*

CONSTRUIR UNA BALSA

Si dispone de provisiones y equipo abundante y tiene que cruzar un río, o si quiere navegar río abajo y no dispone de un barco, una balsa es la embarcación más sencilla que puede construir. Los troncos son los materiales tradicionales empleados en la construcción de una balsa, aunque puede utilizar también cañas de bambú o bidones de aceite. El principio básico de la construcción es hacer la menor cantidad posible de cortes en la madera de la embarcación para que permanezca a flote, y para que sea lo más estable posible. Una balsa nunca debe utilizarse en un río con rápidos, porque podría romperse y usted resultar herido. Si elige construir la balsa con maderos, es recomendable cortar árboles inclinados, ya que puede estimar con certeza dónde caerán. Si es posible, pruebe la balsa en aguas poco profundas antes de botarla en un río profundo.

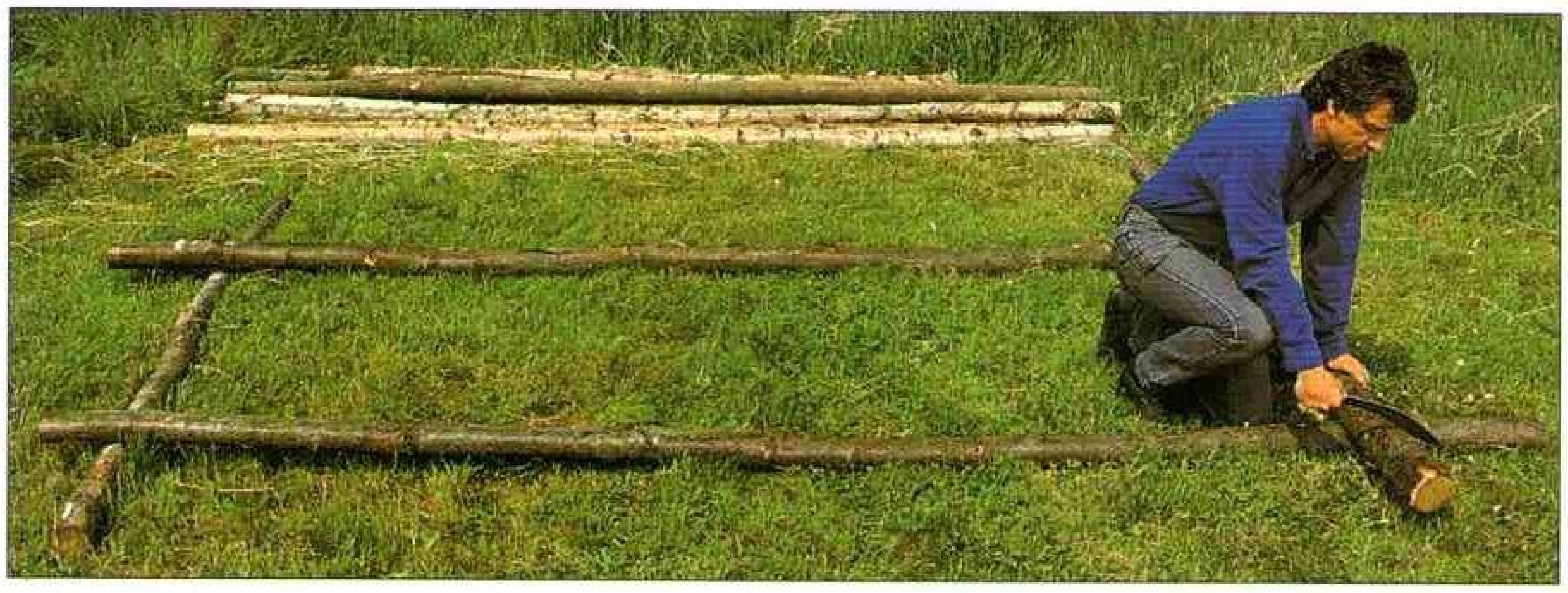

1 Para formar la cubierta, corte entre 12 y 14 troncos resistentes de la misma longitud. Corte seis troncos más, cada uno unos 30 cm más largo que la anchura prevista para la balsa. Éstos actuarán como retenes, manteniendo unida la balsa. Antes de unir los maderos de la cubierta, coloque dos troncos largos debajo de dos de los maderos más cortos para actuar como dispositivo de botadura *(véase página 156)*.

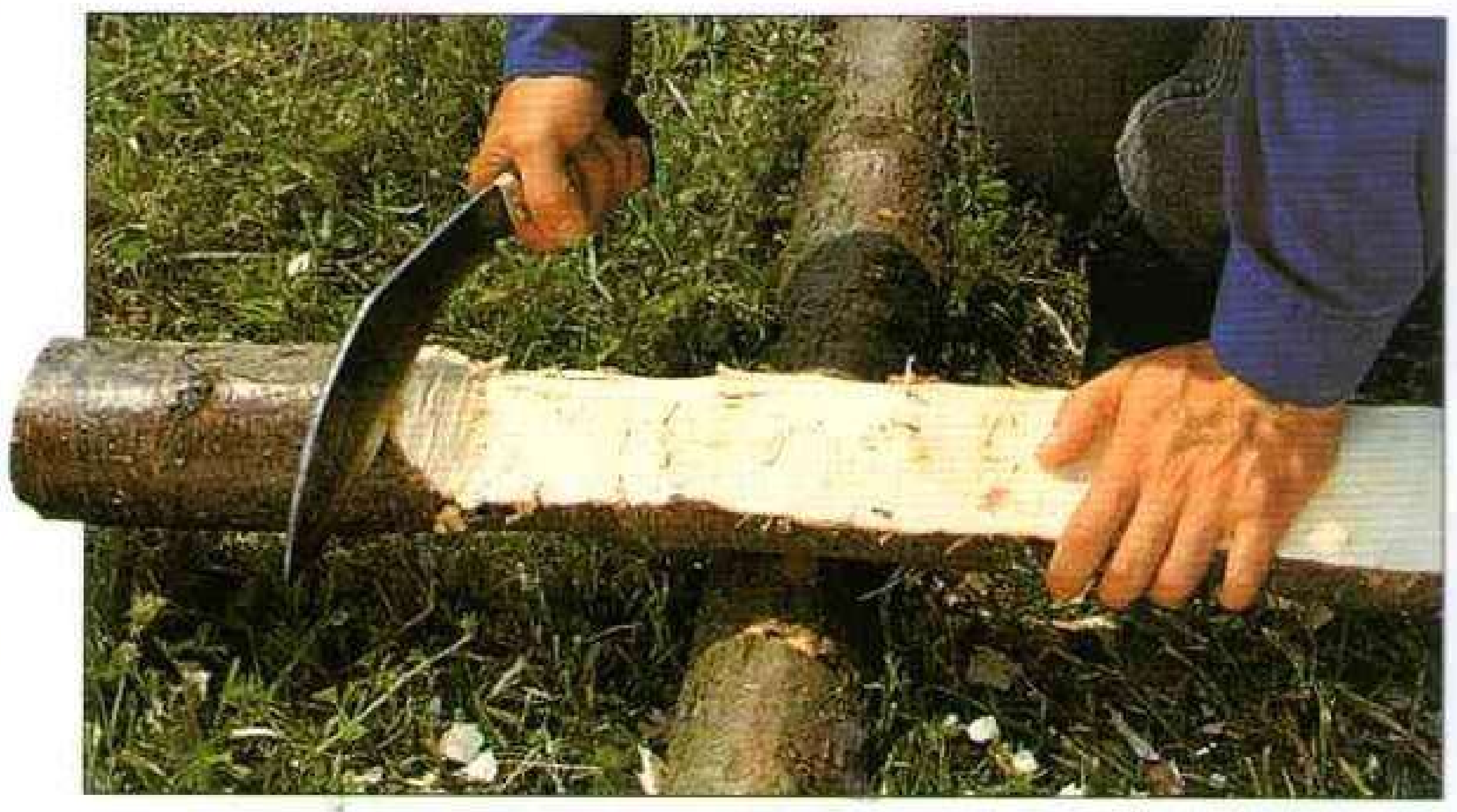

2 Con un cuchillo grande, corte una muesca a todo lo largo de cada uno de los troncos de retén, dejando unos 30 cm en cada extremo. Las muescas deben tener una profundidad de casi medio tronco. Tenga cuidado de no cortar con demasiada profundidad porque el tronco de retén podría ceder.

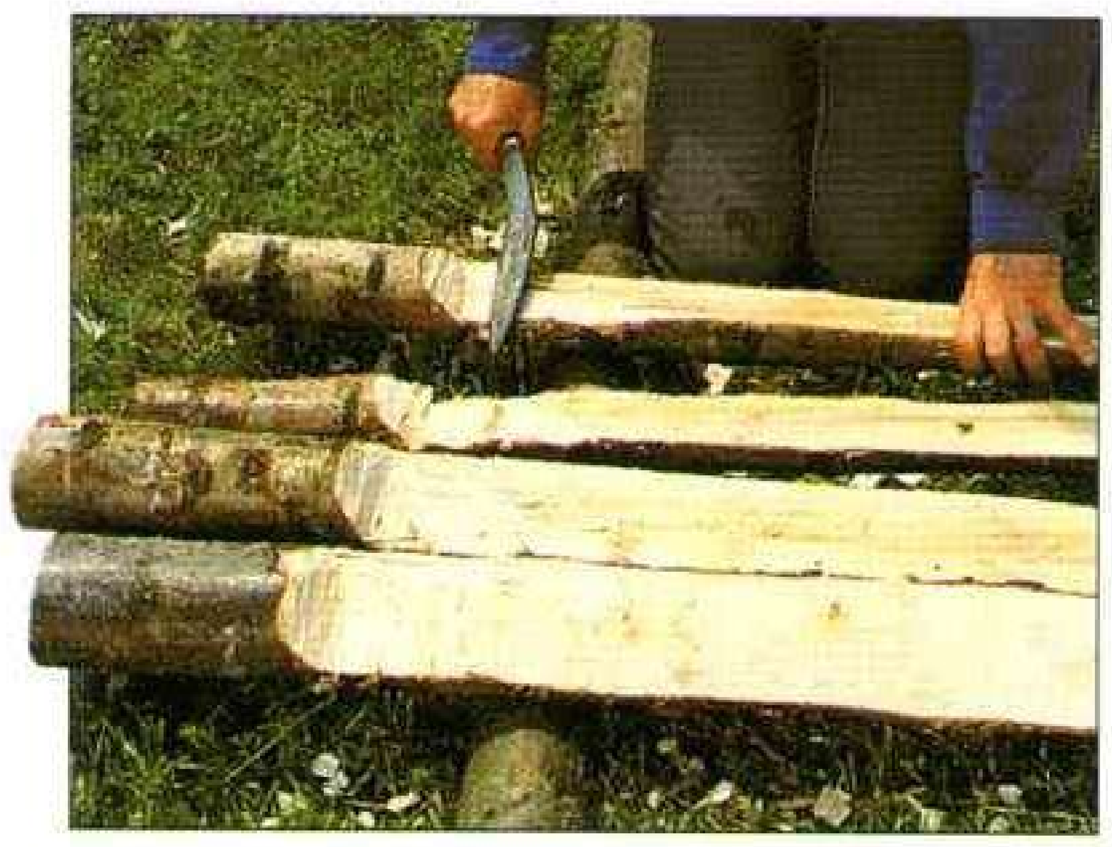

3 Asegúrese de que todas las muescas tengan la misma longitud y sean uniformes. Los maderos largos que formarán la cubierta se apoyarán sobre estos troncos de retén; la cubierta debe ser lo más plana y uniforme posible.

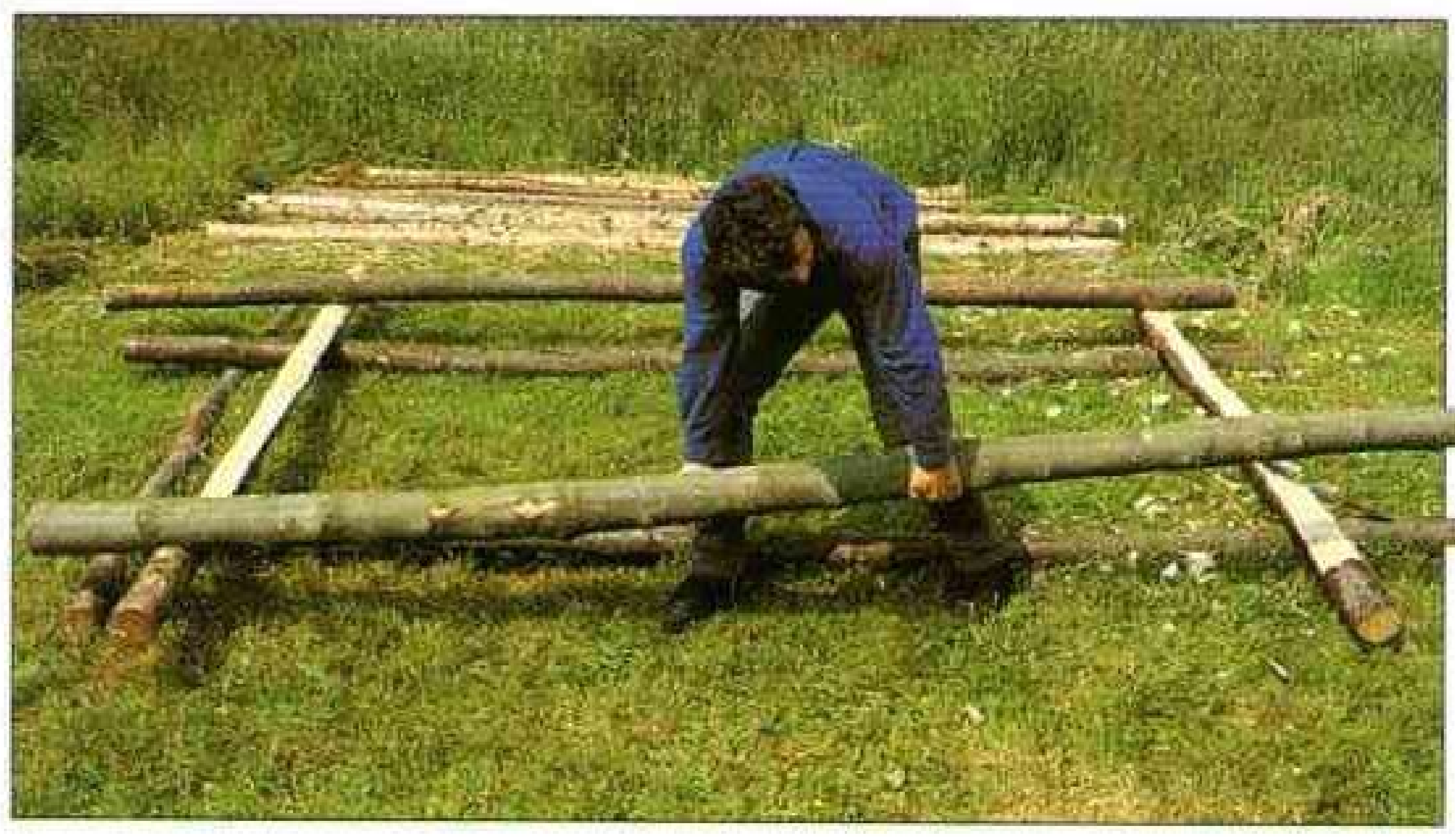

4 Coloque dos troncos con muesca, con la cara tallada hacia arriba, en extremos opuestos al dispositivo de botadura. Éstos actuarán como soporte base. Comience a colocar los maderos largos entre estos dos soportes para formar la cubierta de la balsa.

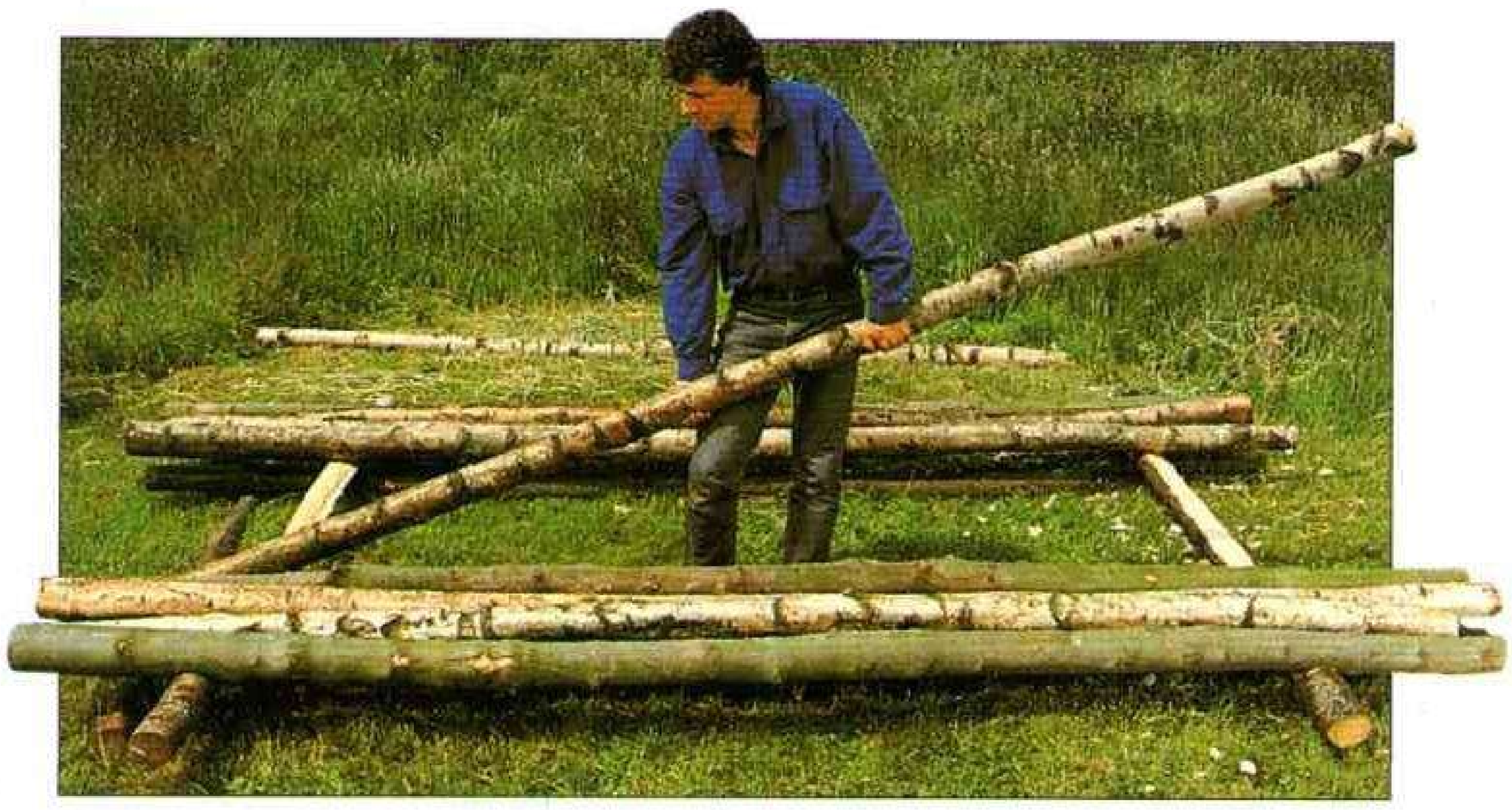

5 Coloque el extremo de cada uno de los maderos en las muescas de los troncos de soporte, trabajando desde el exterior hacia el centro de la balsa. Alterne el grosor de los maderos para que su peso esté bien repartido, alineándolos de manera que se ajusten bien entre sí.

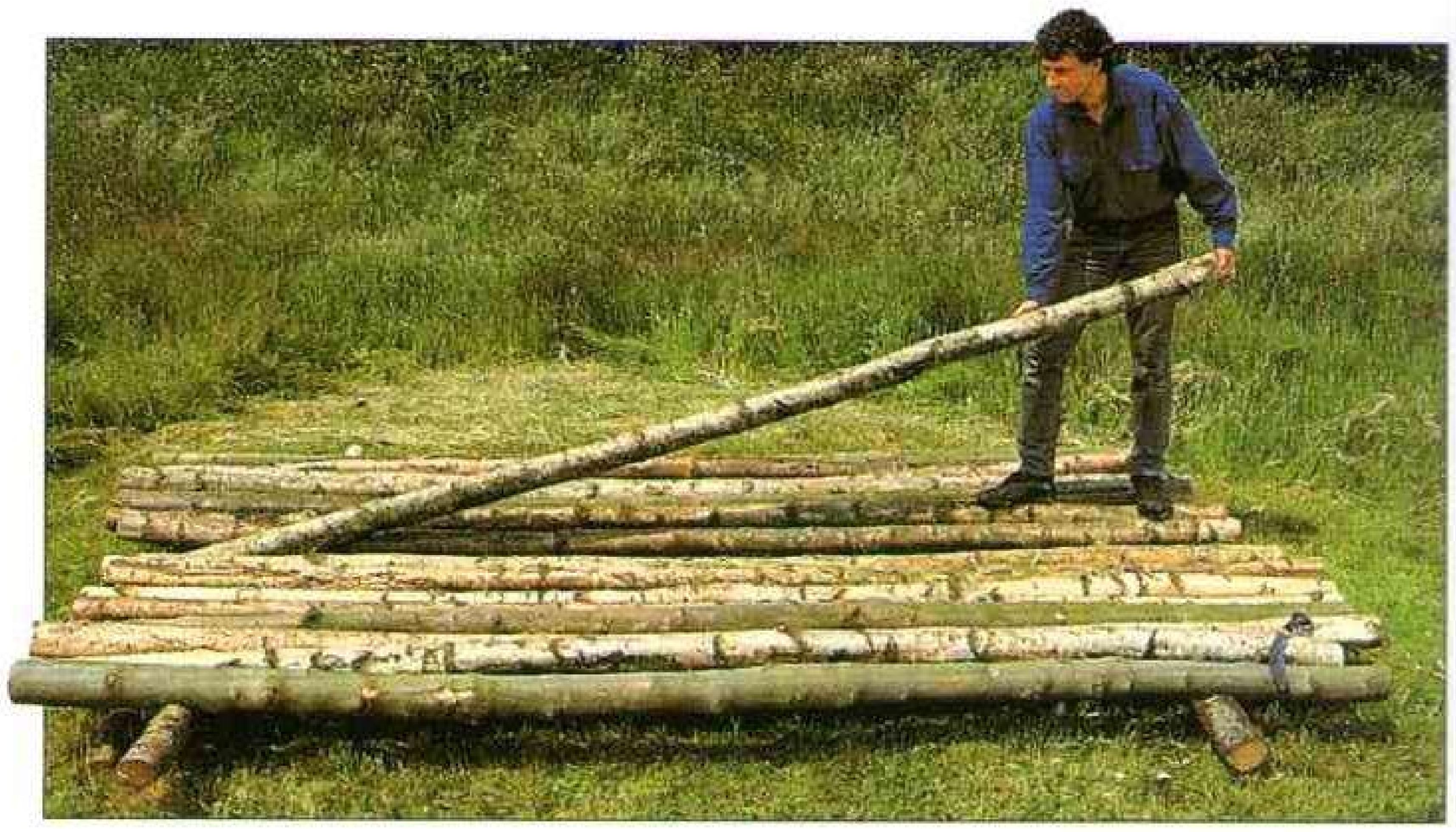

6 El último tronco debe entrar bien ajustado, apretando a los otros troncos entre sí para que la balsa se mantenga unida. Una vez en el agua, tendrá que pisar los troncos para mantenerlos en posición, añadiendo ligaduras adicionales si es necesario. Es normal que algo de agua se filtre entre los troncos.

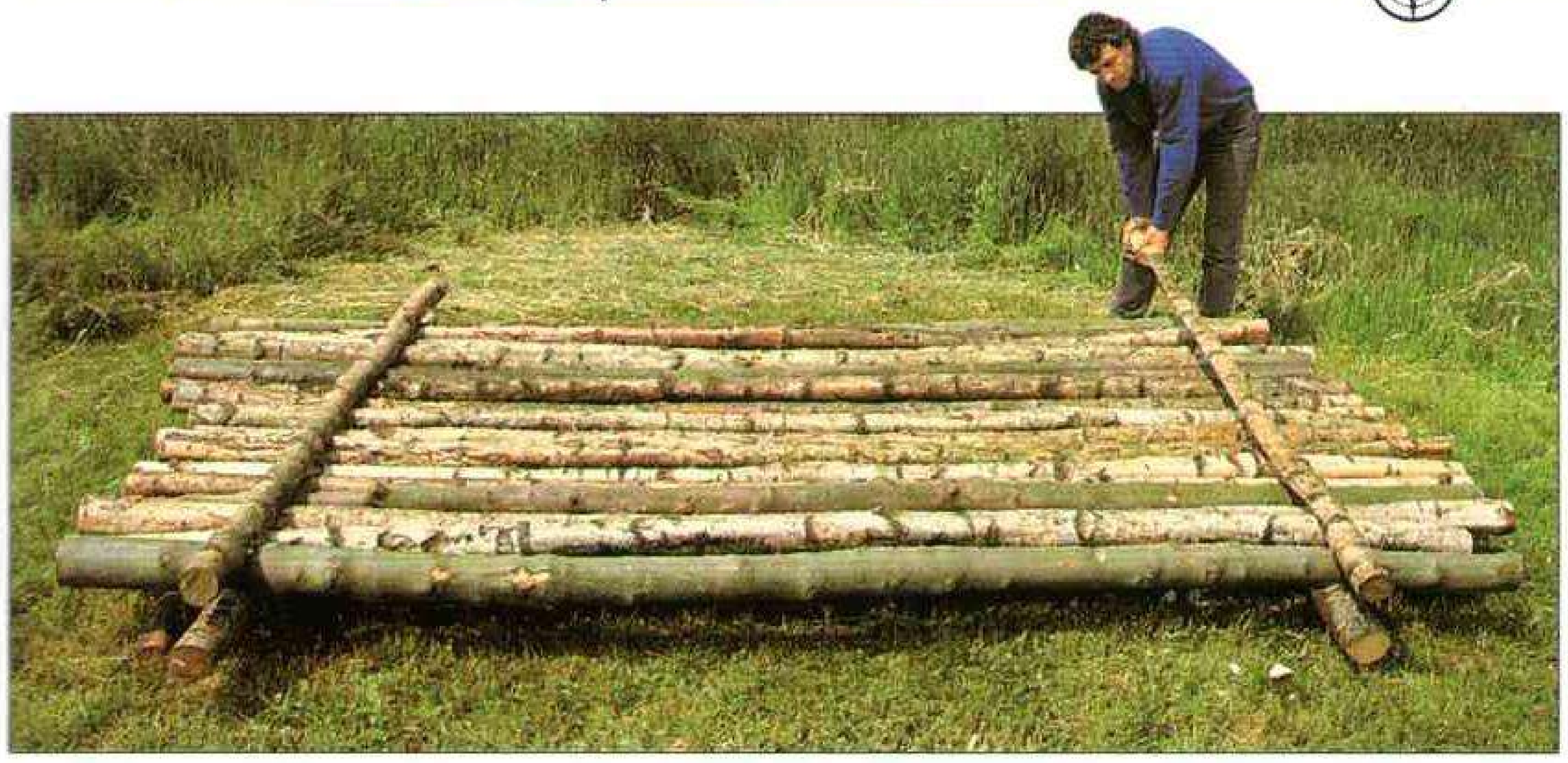

7 Coloque los dos últimos troncos con muesca, con la cara tallada hacia abajo, encima de los maderos de la cubierta. Éstos actuarán como retenes. Recuerde que cuando la balsa esté en el agua, los troncos flotarán hasta estos retenes, que servirán para mantener unidas todas las partes de la balsa.

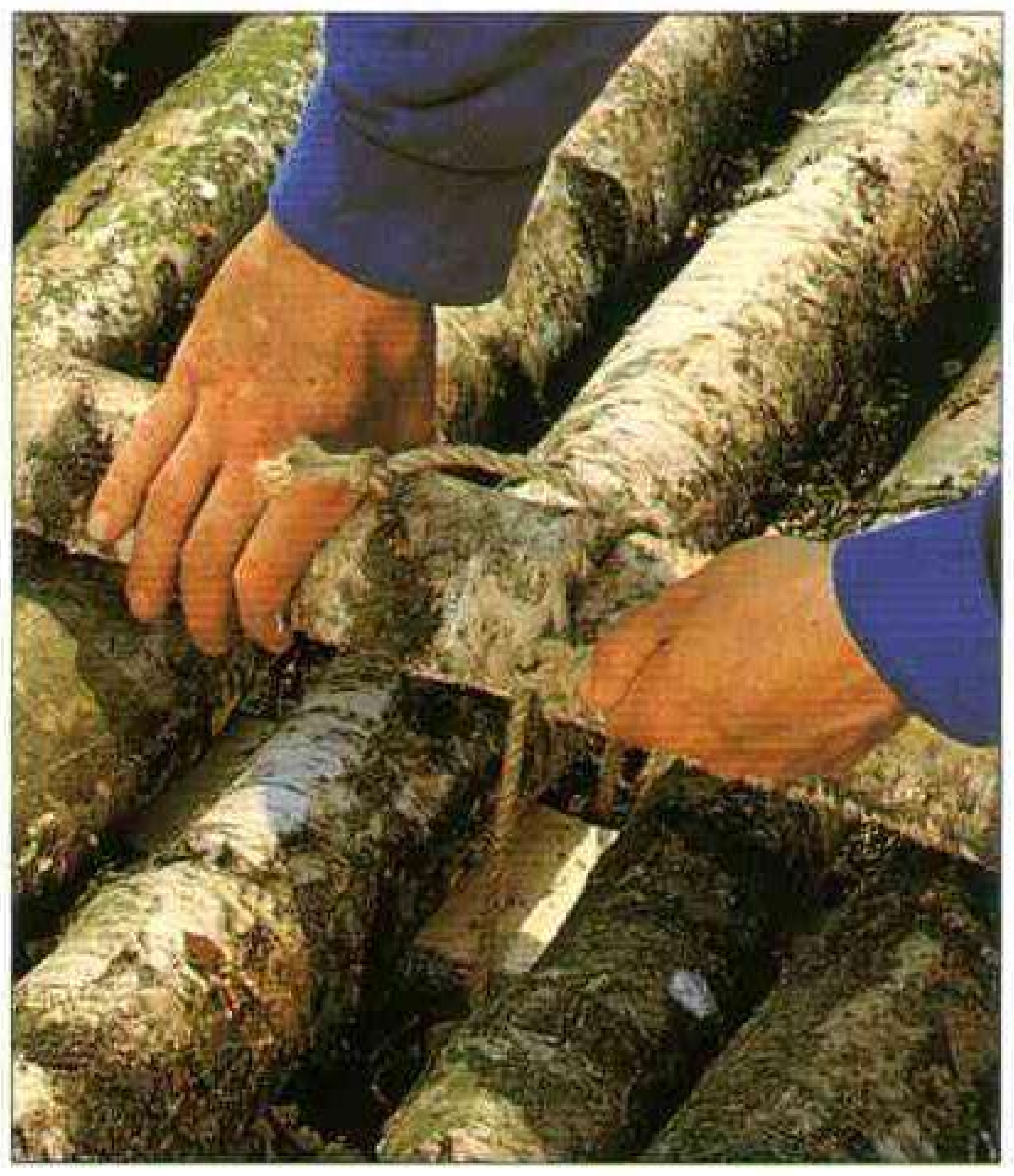

8 Los dos troncos de retén en cada extremo, por encima y por debajo de la cubierta, deben atarse firmemente entre sí, tensándolos tanto como le sea posible. Mantenga los nudos a la vista para que, una vez la balsa esté en el agua, pueda comprobarlos y ajustarlos si es necesario.

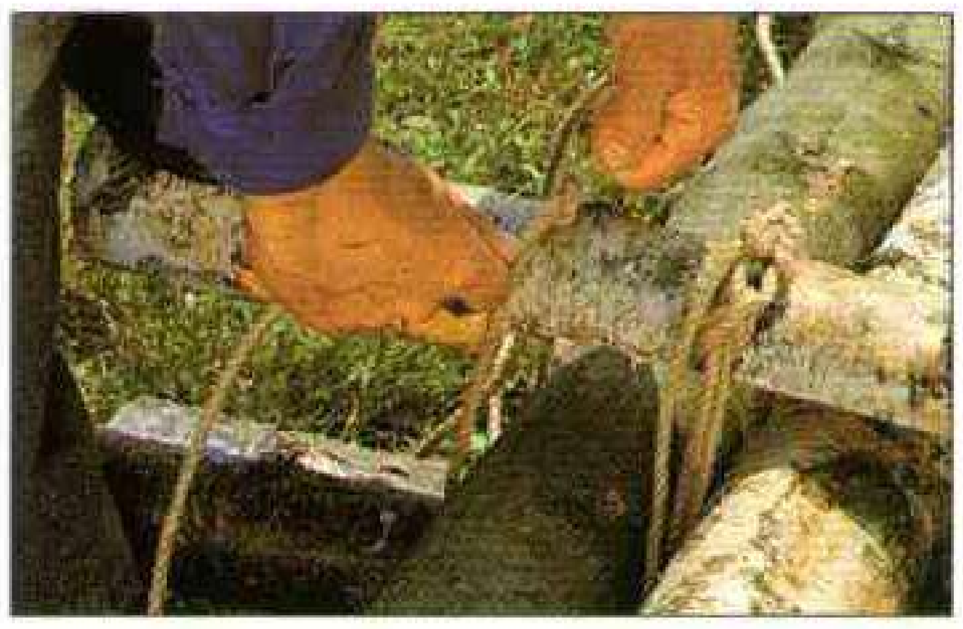

9 Ate los troncos de retén a ambos lados de los troncos de la cubierta utilizando nudos planos *(véase página 35).*

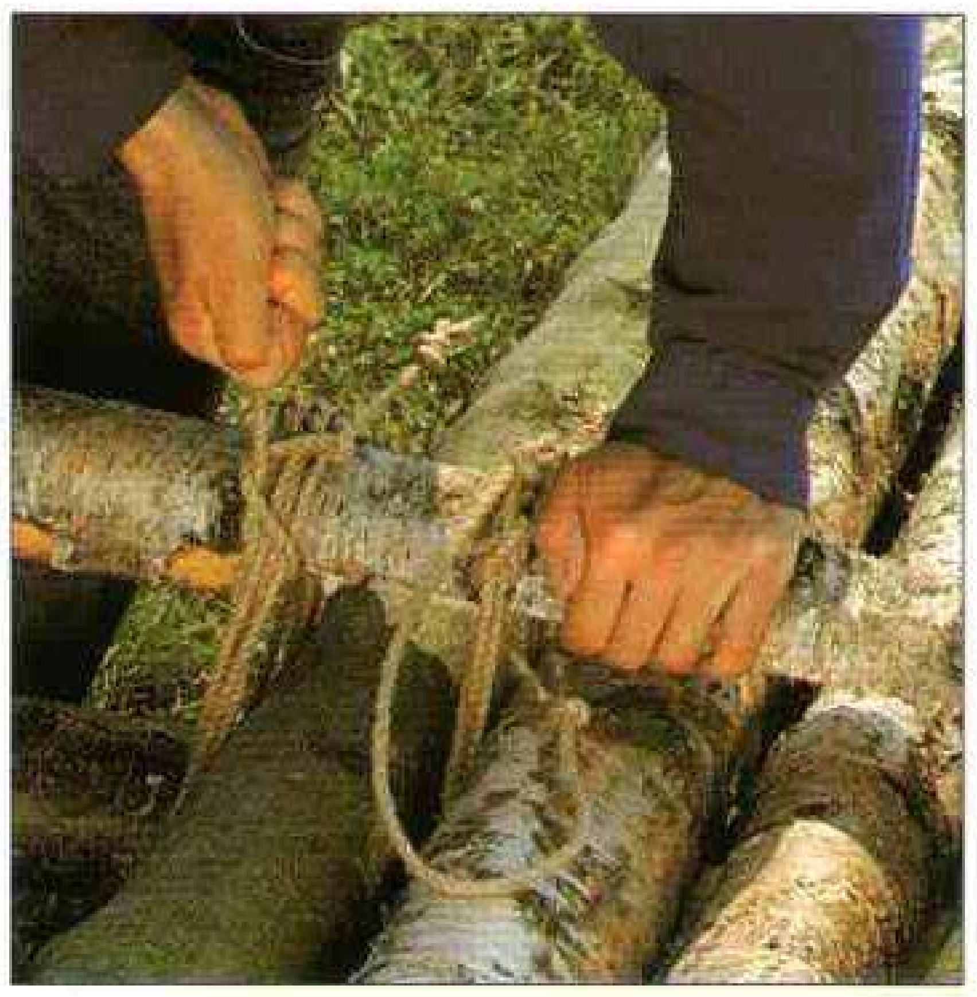

10 Cerciórese de que las ligaduras estén bien sujetas y correctamente anudadas. Si dispone de tiempo, puede hacer más ligaduras entre cada uno de los troncos.

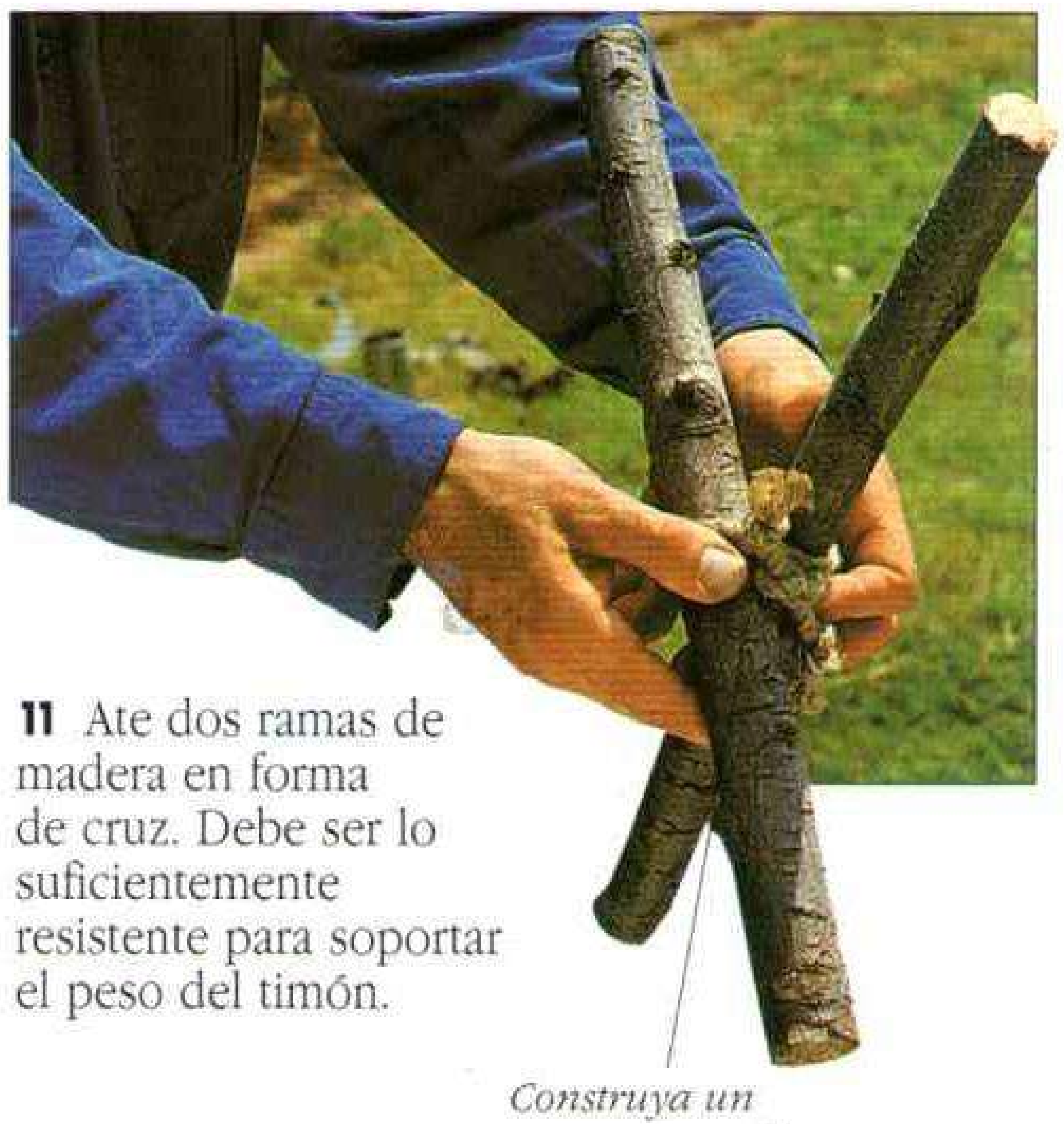

Construya un soporte para el timón

11 Ate dos ramas de madera en forma de cruz. Debe ser lo suficientemente resistente para soportar el peso del timón.

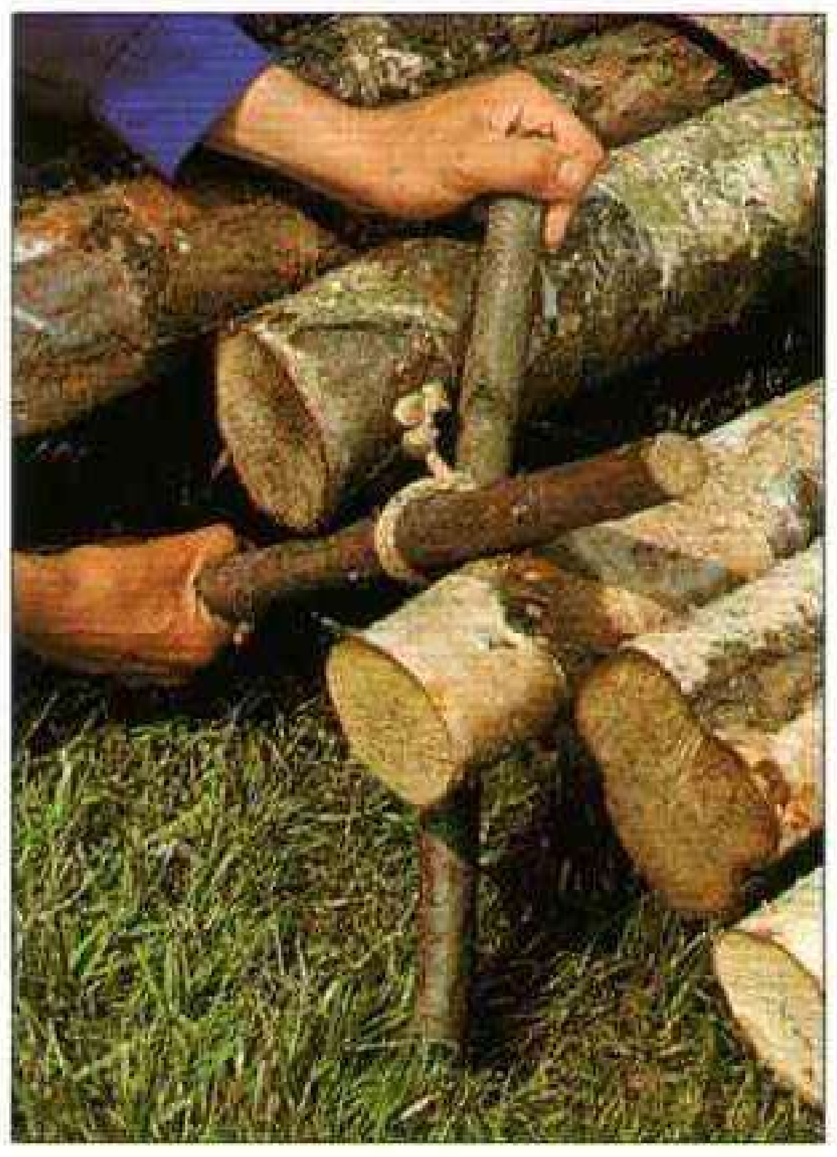

12 Encaje la cruz en la parte posterior de la balsa, entre dos de los troncos de la cubierta, y asegúrese de que queda sujeta. Ajuste las ligaduras si es necesario. No la ate aún en esta etapa.

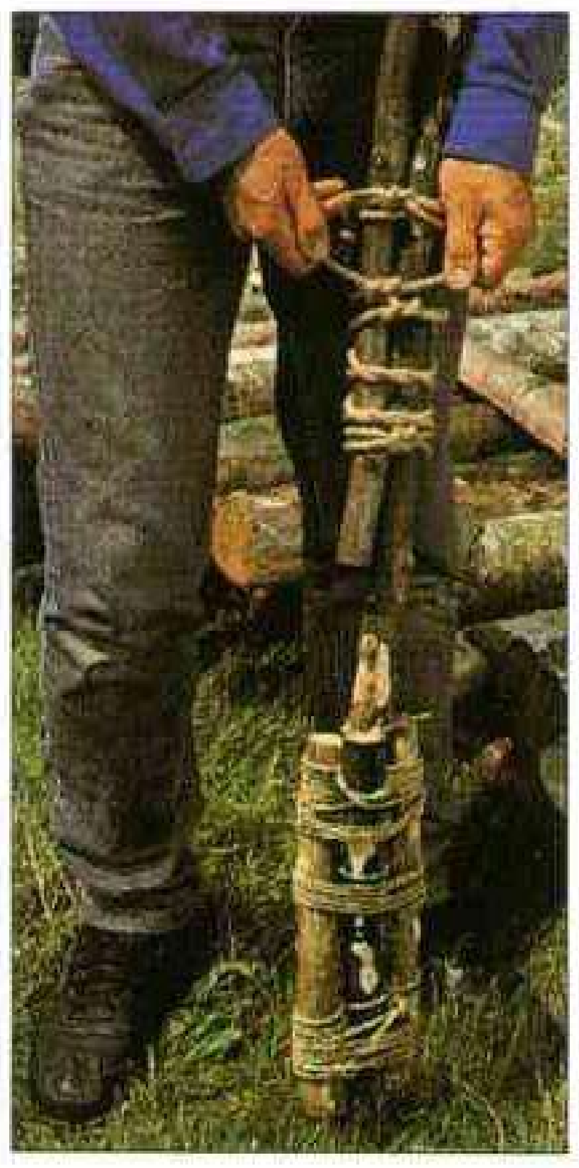

13 Construya un remo *(véase página 157).* Éste actuará como timón y le permitirá guiar la balsa. Si quiere mantenerse de pie mientras gobierna, la balsa tendrá que alargar la caña del «timón» (el mango de la pala) añadiéndole una rama larga.

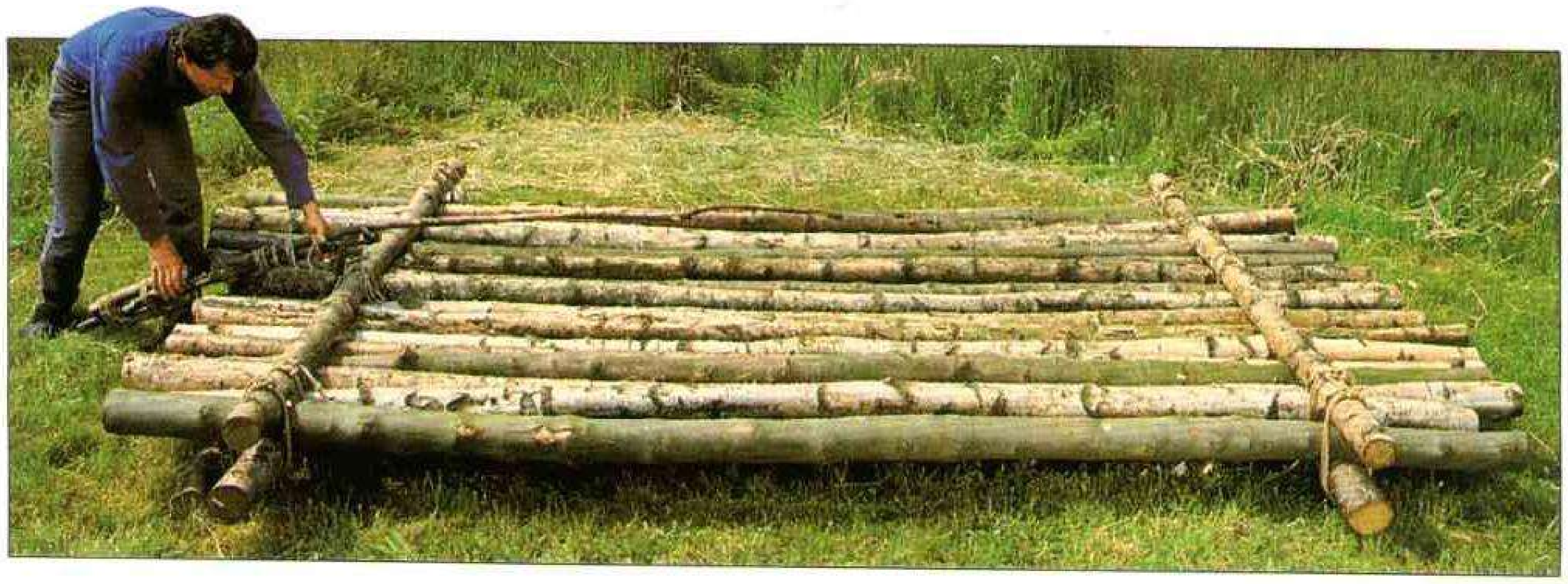

14 Compruebe que el timón se ajuste al soporte y átelo firmemente. Para evitar que el timón y el soporte se rompan al botar la balsa, retírelos de la misma. Sujételos en su sitio cuando la balsa esté en el agua. La balsa debe ser lo suficientemente grande como para llevarle a usted y su equipo al otro lado del río, o río abajo. Compruebe si la balsa flota antes de cargarla.

LAS BALSAS Y SU FORMA DE NAVEGAR

- En lugar de madera, puede utilizar gruesos tallos de bambú para construir la cubierta de la balsa. Deberá unir dos capas de bambú porque una sola no soportaría su peso. En lugar de atar retenes por encima y debajo de la balsa, puede hacer agujeros a ambos lados de los tallos y pasar unas ramas finas o cuerda a través de ellos para unir los tallos entre sí y formar la cubierta. Pase una cuerda alrededor de cada uno de los tallos como medida de seguridad adicional.
- No sabrá cómo flota su balsa hasta que la bote. Con algunos tipos de madera, una balsa cargada flotará ligeramente por debajo de la superficie del agua, aunque esto no constituye un problema importante.
- Una balsa virará de forma relativamente lenta. Téngalo en cuenta al maniobrar.
- Evite las colisiones con la orilla o los objetos grandes, ya que pueden romper las ataduras y desmontar la balsa.

EL EMPLEO DE BALSAS Y OTRAS EMBARCACIONES

EN MUCHAS ZONAS silvestres, la vegetación densa y el terreno escarpado hacen recomendable el viaje por río, lago o a lo largo de la costa. Sin embargo, recuerde que aventurarse en el agua puede resultar peligroso, en especial si lo hace en una barca improvisada. Debe asegurarse, por lo tanto, de que su embarcación sea la adecuada para las aguas que pretende cruzar, sea improvisada o no. El estado del agua puede variar rápidamente, convirtiendo una travesía segura en peligrosa. Por esta razón, no debe aventurarse a navegar a menos que tenga experiencia o no le quede otra opción. Lleve siempre algún tipo de chaleco salvavidas.

CÓMO BOTAR UNA BALSA

Una balsa sólo deberá utilizarse en aguas tranquilas. Constrúyala cerca del agua, ya que podría resultar ser muy pesada para arrastrarla hasta la orilla. Antes de meter la balsa en el agua, átela a un objeto inamovible, dejando suficiente cuerda libre para permitir botarla.

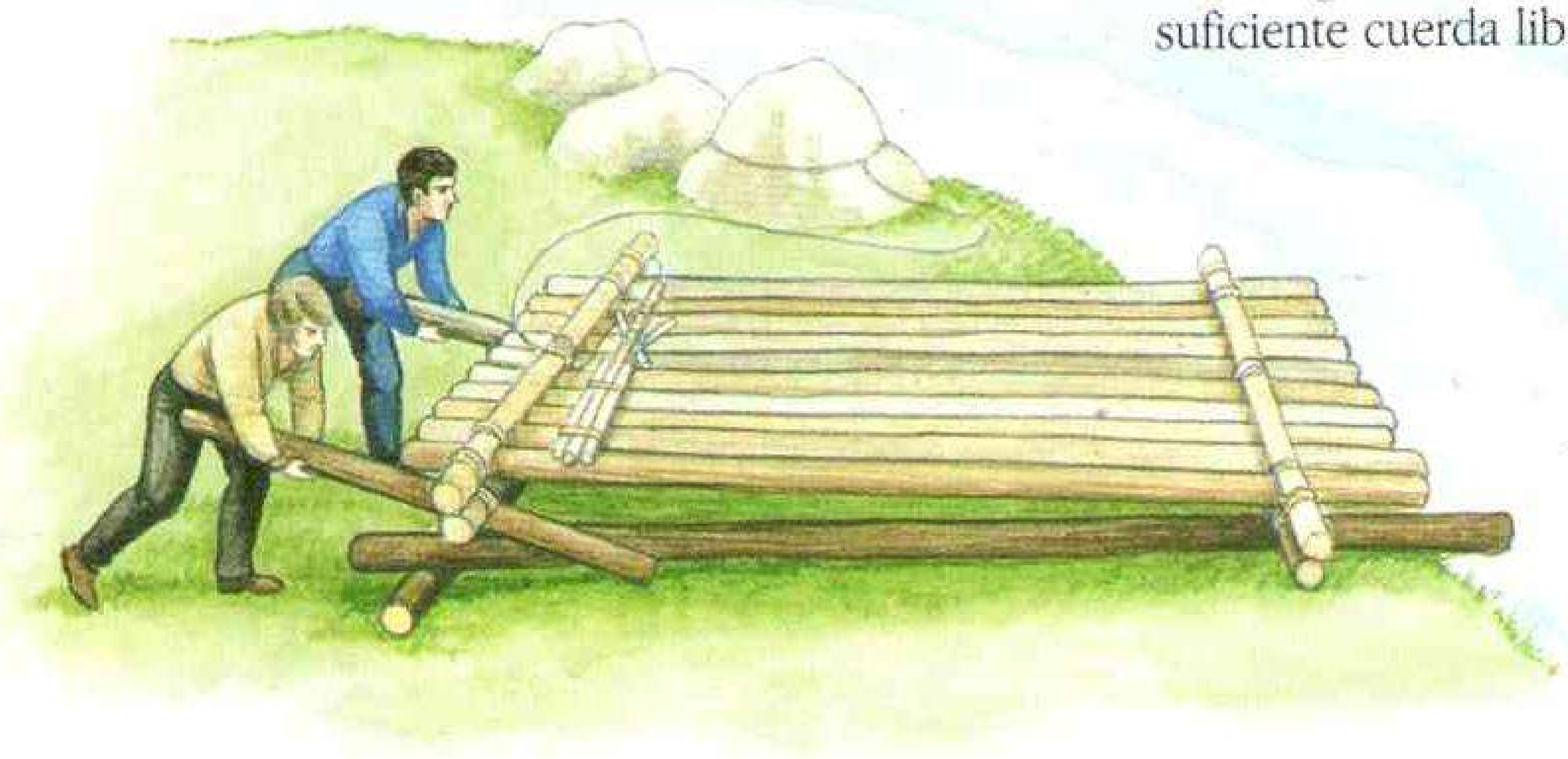

Haciendo palanca. *Bote la balsa haciendo palanca con unos troncos en uno de los lados hasta lograr deslizarla al agua.*

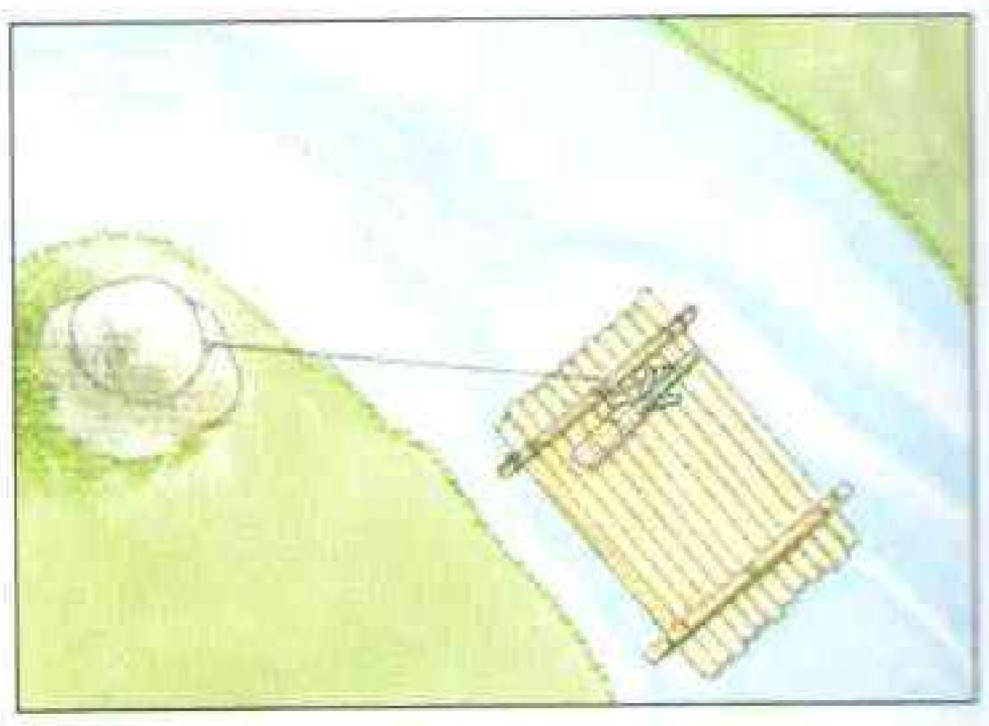

A flote. *Una vez ha comprobado que la balsa no se hunde en el agua, puede sujetar el timón y comenzar a cargar su equipo en la parte central.*

APAREJAR UNA BALSA

Para travesías prolongadas a lo largo de un río recto, levante un mástil y una vela para aprovechar el viento al máximo —siempre y cuando el viento sople en la dirección adecuada. Sin una quilla, el timón únicamente controlará la dirección si el viento es muy suave, pero le permitirá mantener la balsa recta. Si la proa se hunde en el agua, reduzca la superficie de la vela. Únicamente podrá moverse río abajo siguiendo la corriente.

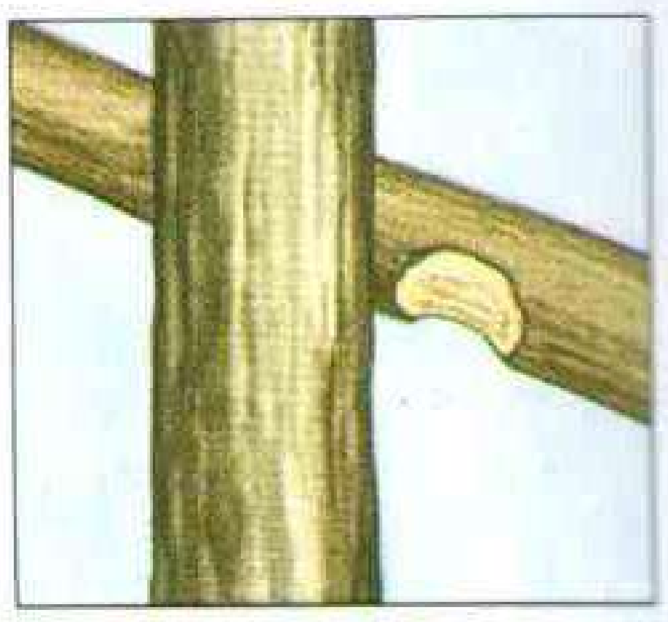

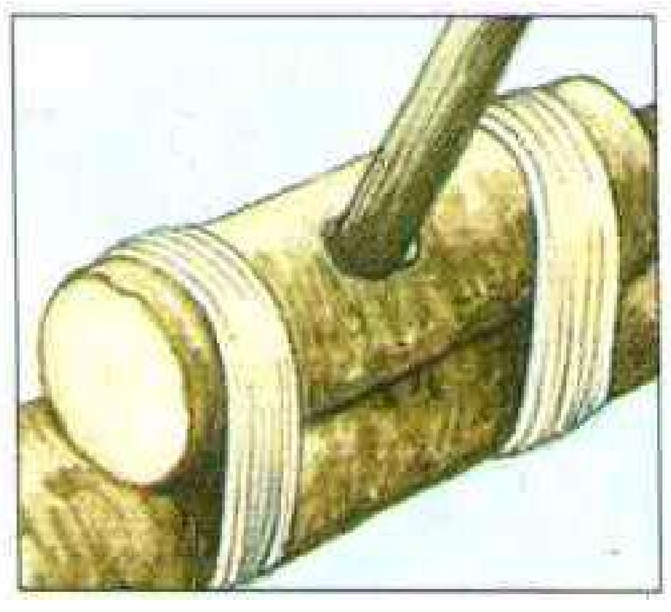

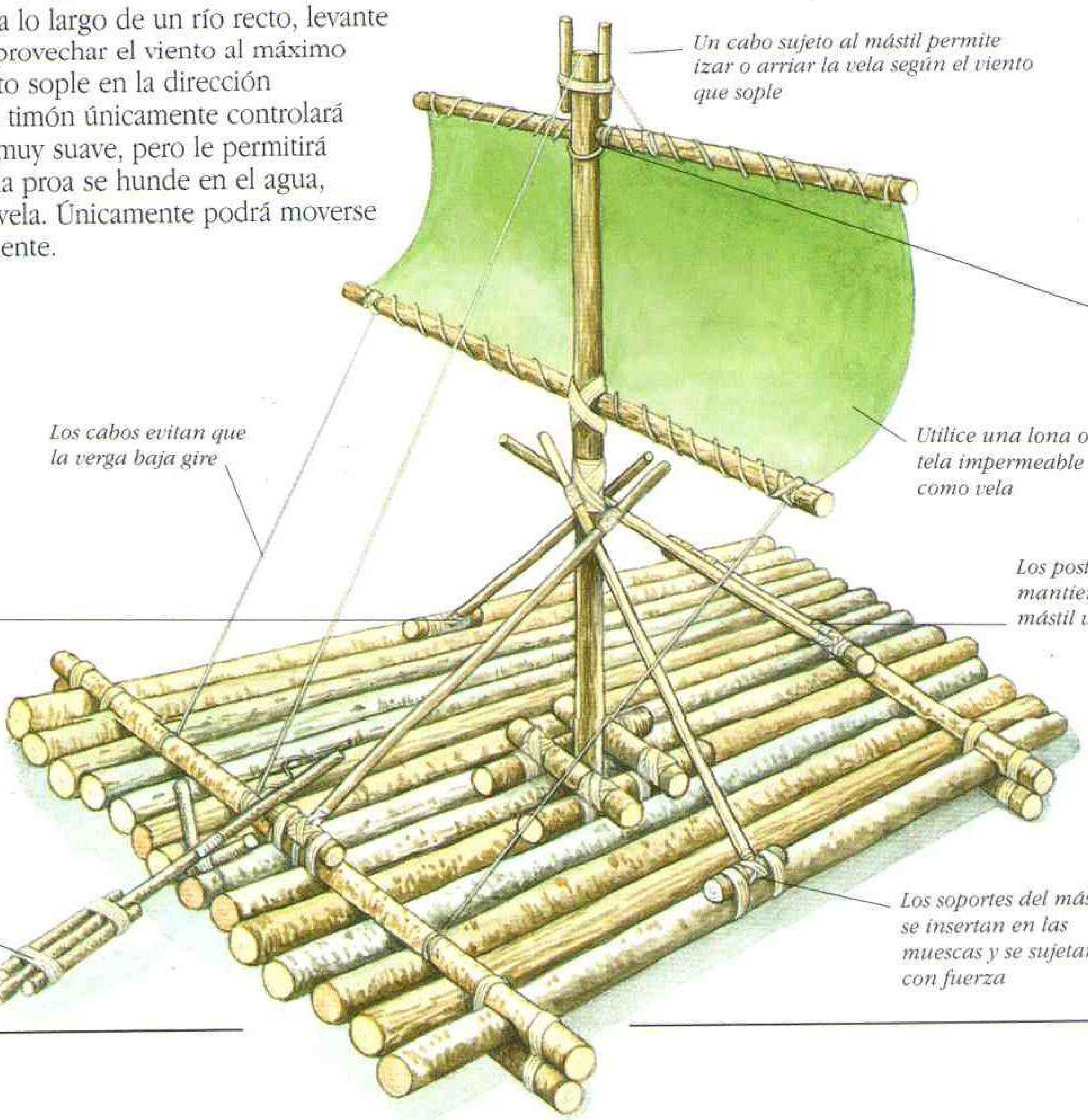

Un cabo sujeto al mástil permite izar o arriar la vela según el viento que sople

La verga alta tiene una muesca que permite subirla o bajarla empleando el cabo que pasa por encima del mástil, y así reducir la vela

Los cabos evitan que la verga baja gire

Utilice una lona o tela impermeable como vela

Pequeños troncos de madera en los cuatro costados de la balsa permiten la talla de muescas para insertar palos de refuerzo

Los postes mantienen el mástil vertical

Un remo sujeto a la popa hace las veces de timón

Los soportes del mástil se insertan en las muescas y se sujetan con fuerza

Levantar un mástil con vela. *Sujete un mástil largo entre cuatro troncos y asegúrelo con fuerza. Refuerce con cuatro palos, cada uno ligado a la cubierta. Utilice una manta o una lona impermeable como vela.*

LA ELECCIÓN DE LA EMBARCACIÓN

Cualquier tipo de embarcación menor tiene sus ventajas y sus inconvenientes, pero nunca debe utilizarse para un propósito inadecuado, en condiciones que vayan más allá de sus posibilidades. Algunas embarcaciones, como el coracora, son aptas para las aguas tranquilas, mientras que otras, como las neumáticas modernas, pueden utilizarse en condiciones inadecuadas para otro tipo de embarcaciones.

Neumática. *Una neumática puede utilizarse en condiciones relativamente adversas, y es la embarcación preferida por muchos científicos y exploradores. Resulta manejable incluso en mares agitados. Sin embargo, el hielo superficial y otros obstáculos pueden perforar los compartimientos de flotación. Cuando utilice una neumática deberá mantener el peso en la proa para evitar que la barca se levante hacia atrás.*

Motor fuera borda. *Una embarcación de fondo plano con un motor fuera borda es ideal para los ríos anchos de flujo uniforme. Es capaz de cubrir grandes distancias. Un motor fuera borda sobrecarga la popa de una embarcación pequeña, en particular durante el avance, por lo que debe equilibrarse con un peso en la proa. La longitud del eje de cola reducirá el calado del barco hasta cierto punto.*

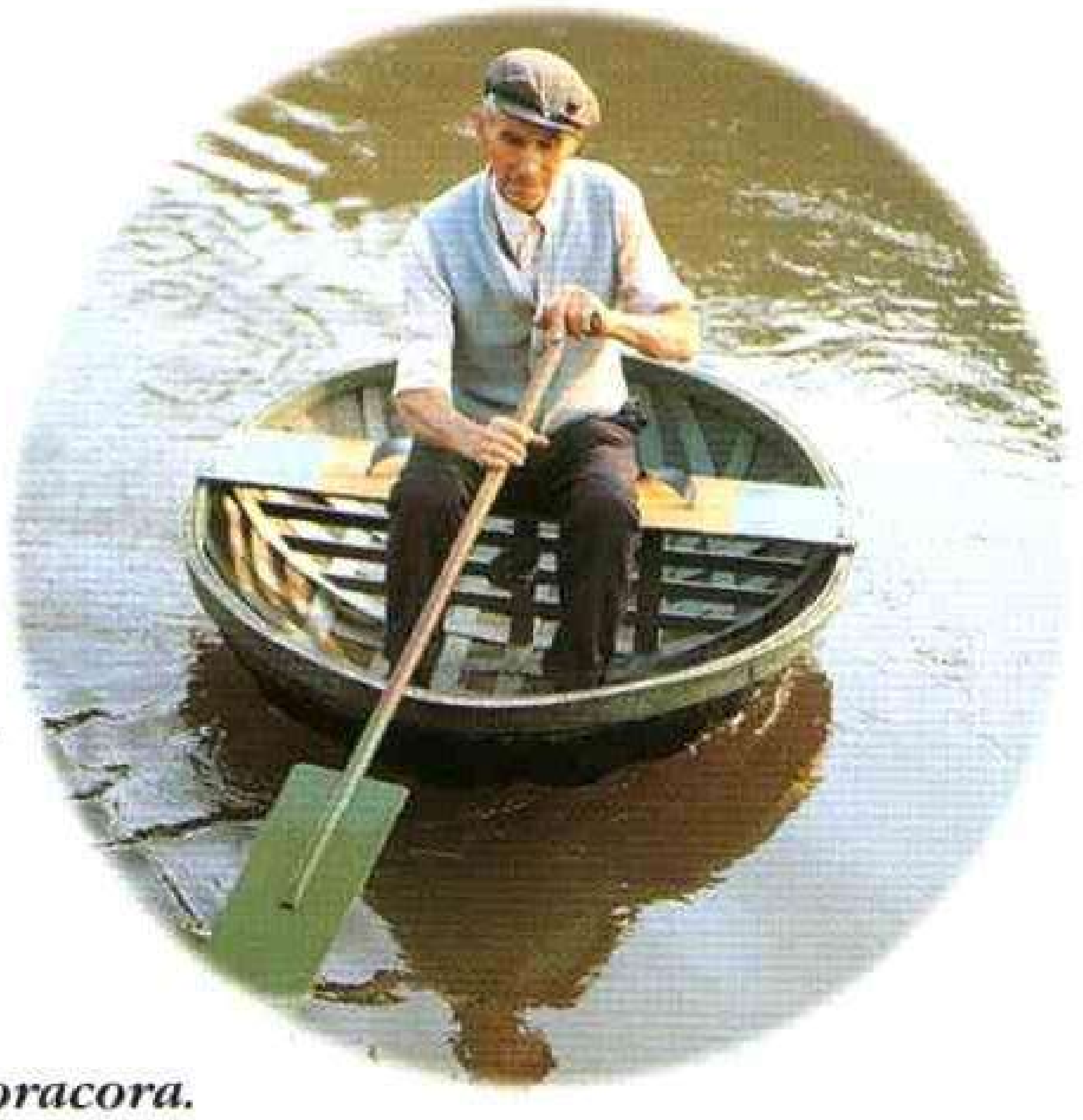

Coracora.
Un coracora requiere una gran habilidad para su construcción y manejo. Consiste en un armazón de maderas flexibles en forma de media esfera, cubierto con piel o lona impermeable. Un coracora únicamente es recomendable para las aguas muy protegidas y tranquilas.

Kayak. *El tradicional kayak esquimal ha servido como modelo para la construcción de las resistentes canoas modernas capaces de sortear las aguas más rápidas. Las embarcaciones de piel o tela no son tan robustas como las versiones en fibra de vidrio, pero son más ligeras y pueden gobernarse fácilmente por los rápidos demasiado difíciles de pasar en barca. Se necesita una gran habilidad para controlar un kayak, en particular cuando va cargado con el equipo, o cuando se utiliza en aguas bravas. Al carecer de quilla, vuelcan fácilmente.*

CONSEJOS PARA IR EN BARCA

- Lleve siempre un chaleco salvavidas, o improvise uno llevando botellas de agua vacías en sus bolsillos.
- Ate cualquier artículo de su equipo a la embarcación con una cuerda.
- No sobrecargue la embarcación.
- Explore el río que pretende recorrer antes de botar su embarcación.
- Si tiene un motor fuera borda, acelere hasta que la barca adquiera una posición horizontal, y reduzca después para mantener la marcha.

CÓMO CONSTRUIR UN REMO

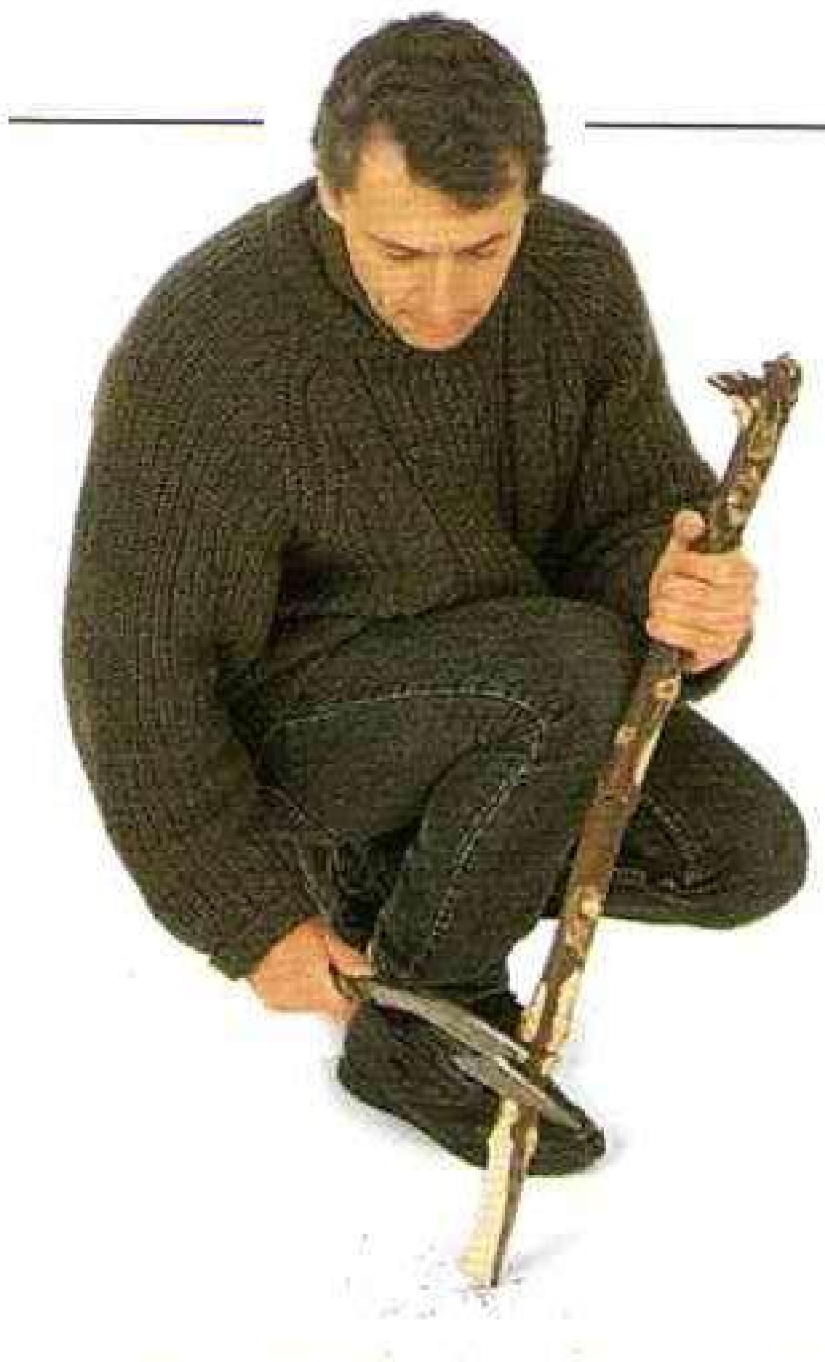

1 Para construir un remo, elimine todos los ramos de una rama fuerte y verde con un cuchillo. Talle uno de los extremos en forma de cuña.

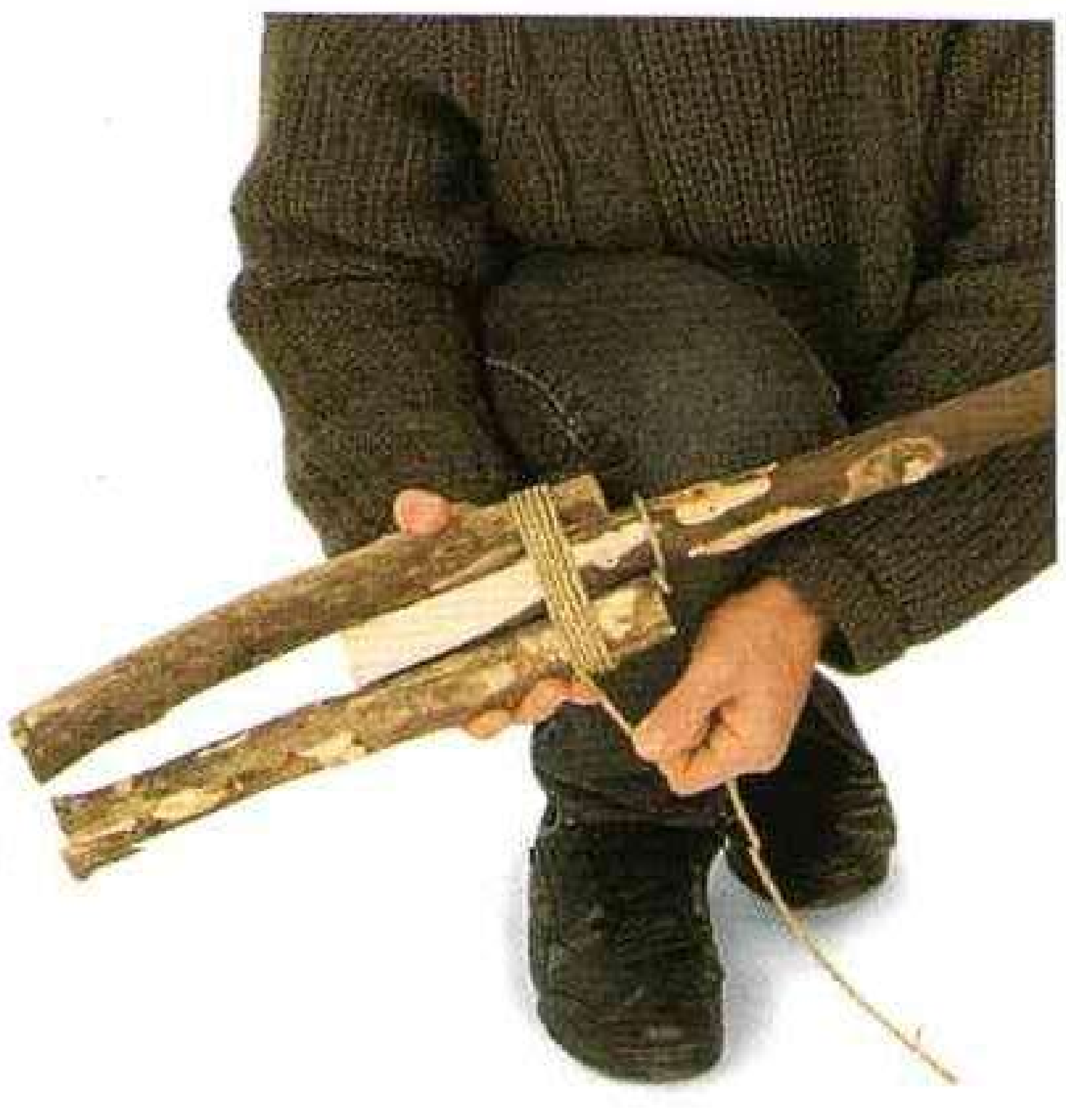

2 Ate dos ramas más cortas y rectas a ambos lados de la cuña, que actuarán como pala del remo.

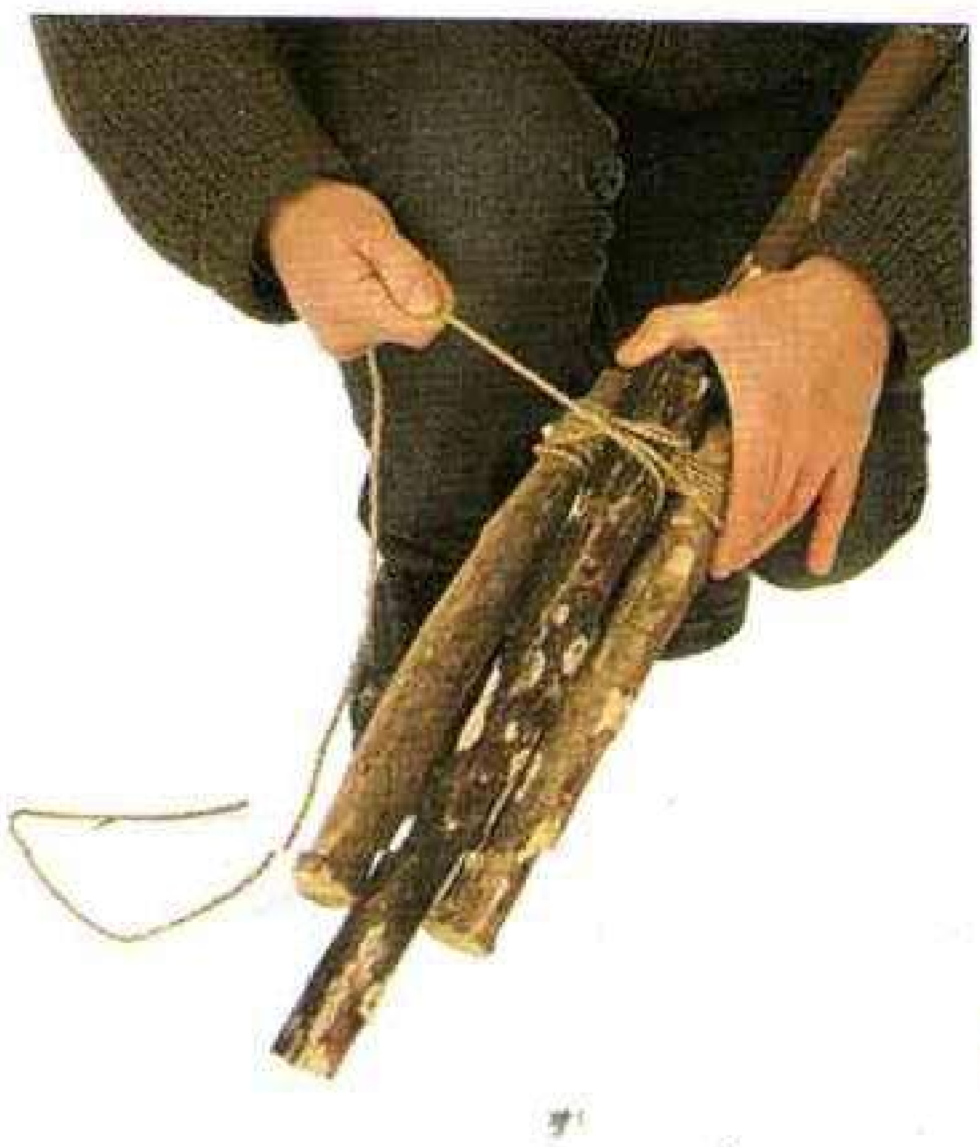

3 Ate una tercera rama en la parte central, forzándola en el hueco formado por las otras dos, y apoyándola en la cuña al final del mango.

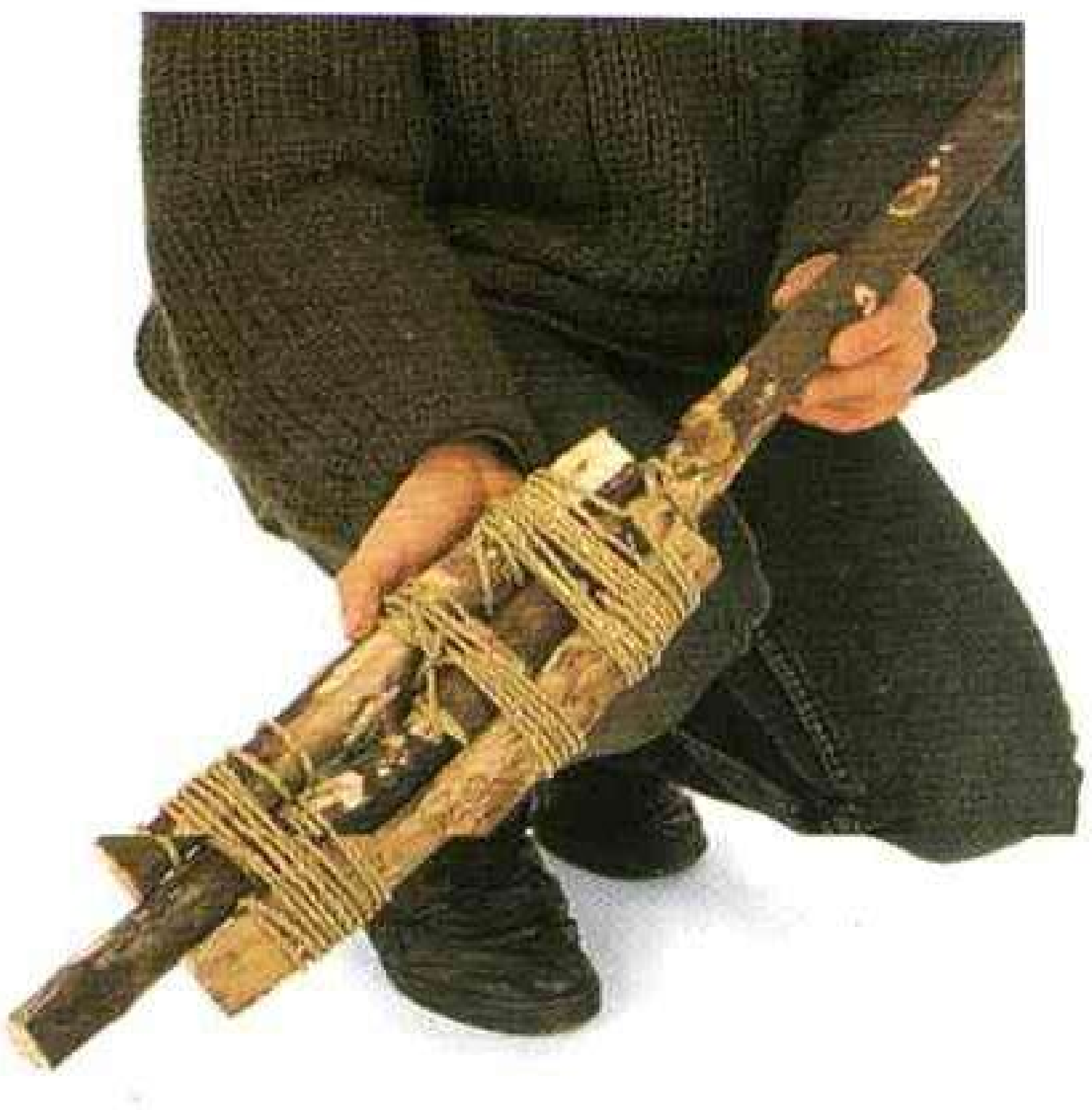

Remo terminado. *Es posible que tenga que volver a atar el remo cuando se humedezca, ya que la cuerda puede aflojarse.*

A GUIDE TO
MEDICINAL PLANTS

An Illustrated, Scientific and Medicinal Approach

Koh Hwee Ling
Chua Tung Kian
Tan Chay Hoon

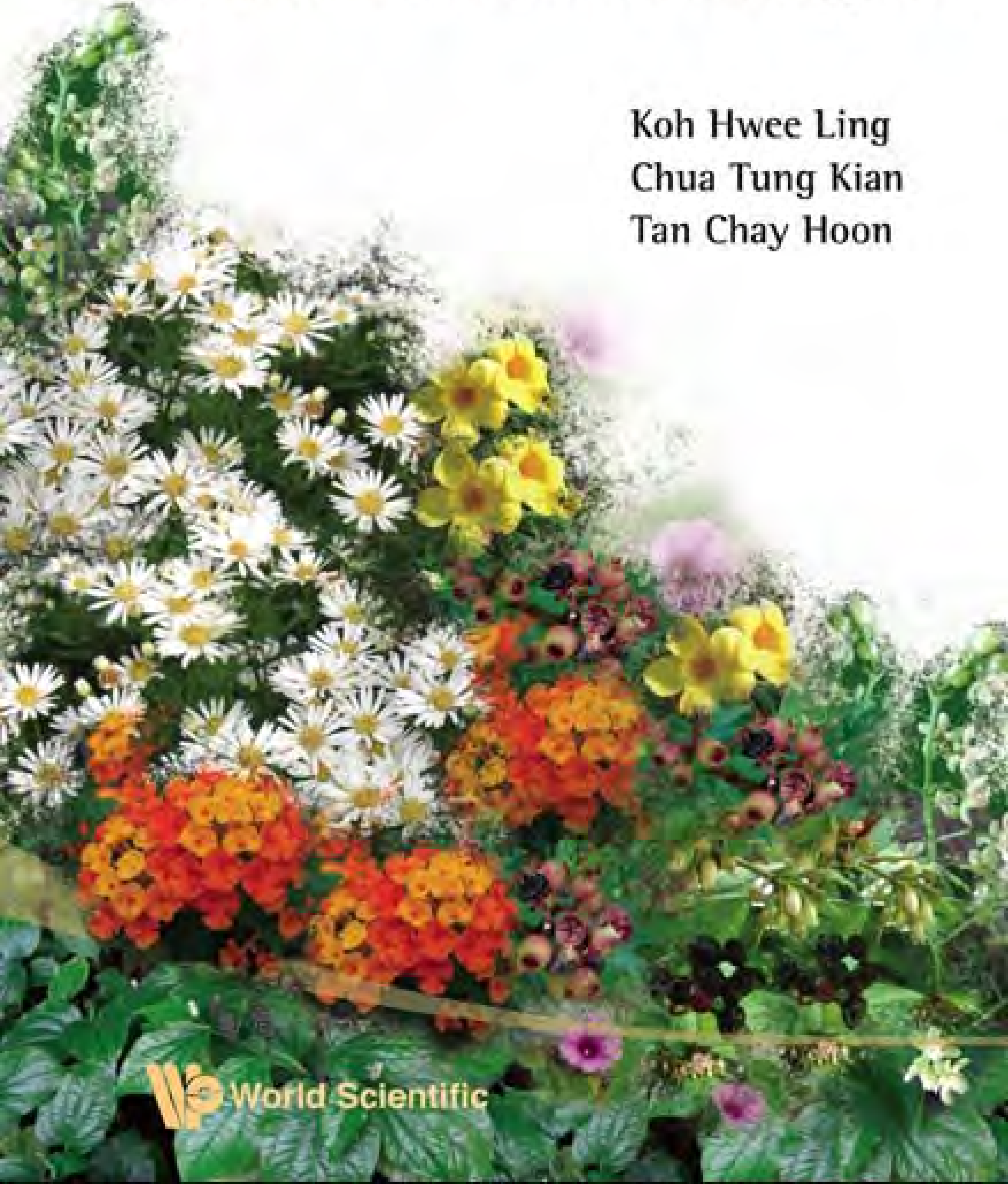

World Scientific

1. *Abrus precatorius* L. (Leguminosae)

Rosary Pea, Indian Licorice, Precatory Bean

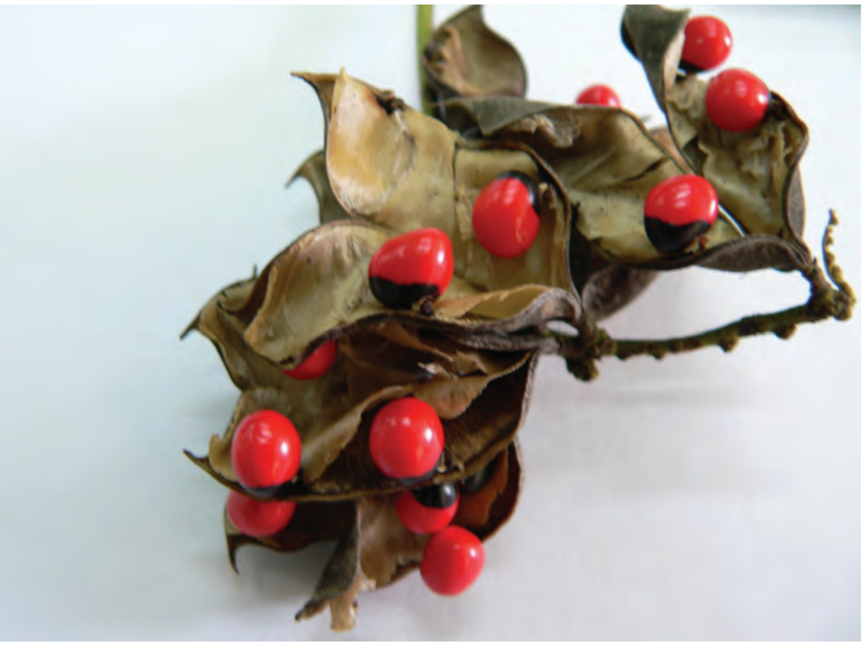

Seeds of *Abrus precatorius*

Description: *Abrus precatorius* L. is a perennial climber with a slender stem. Leaves are pinnate and 5–8 cm long. Leaflets are rhomboid, numbering 20–24 or more, opposite and are 1.2–1.8 cm long. Leaf margin is entire. It bears pink flowers arranged in dense axillary racemes. Pods are oblong, cylindrical, inflated, 5–6 cm by 1 cm and contains 3–6 round, glossy, black and red seeds.[1–3]

Origin: Native to Pakistan, India, Ceylon and tropical Africa; and introduced widely in the New and Old World.[4]

Phytoconstituents: Abrectorin, abricin, abridin, abrins A–D, (+)-abrine, abruslactone A, abrusgenic acid, abrusogenin, abrusoside A–D, precatorine, abruquinones, abraline, abrusic acid, abruquinone G and others.[2,5–18]

Traditional Medicinal Uses: A decoction of the leaves has been prescribed for scurvy, cough, bronchitis, sprue and hepatitis and as a refrigerant. They are also applied on painful swellings, eye inflammation, cancer, syphilis and on leucodermic spots.[19] The leaves are also effective in the treatment of coryza, cough, fever, and jaundice resulting from viral hepatitis and intoxications.[6] The seeds have been used to treat fever, malaria, headache, dropsy and to expel worms.[3] A decoction of the seeds is applied for abdominal complaints, conjunctivitis, trachoma and malarial fever.[2] Central Africans use powdered seed as an oral contraceptive.[5] It is also used to lower high blood pressure and relieve severe headache.[5] The seeds are very toxic and can be applied externally to treat bacterial infection and accelerate the bursting of boils and to cure mastitis and galactophoritis.[6] The seed has purgative properties and is used as an emetic, tonic, aphrodisiac, and for nervous disorders. The poultice can be used as suppository, abortifacient, or tonic for pregnant women and children and to treat severe headaches.[19] Water from the boiled roots is used to cure cough, bronchitis, sore throat and also applied as an emetic agent.[2,5]

Pharmacological Activities: Antibacterial,[20] Anthelmintic,[21] Antiviral,[17] Anti-inflammatory,[10,22,23] Anticancer/Antitumour,[11,24–30] Antiplatelet,[23] Antiprotozoal,[31] Immunomodulatory,[27,32–35] Antioxidant,[23] Antiplasmodial,[17] Antitubercular[17] and Molluscicidal.[36]

Dosage: In Central Africa, 200 mg powdered seeds are used as an oral contraceptive which can last 13 menstrual cycles.[5] Approximately 14 g of powdered seeds are used as tonic for pregnant women and children.[19] About 5–7 of seed grains are prescribed for pertussis.[37]

Adverse Reactions: Ingestion of *Abrus* seeds resulted in pulmonary oedema and hypertension.[38] Abrine can cause coma, confusion, convulsions, dehydration, gastroenterosis and hypotension.[37]

Toxicity: The LD_{50} of abrin in mice is 0.02 mg/kg body weight.[2] Ingestion of seeds causes severe stomach cramping accompanied by nausea, severe diarrhoea, cold sweat, fast pulse, organ failure, coma and circulatory collapse.[2,5,39] Oral administration of a 50% ethanol extract of *A. precatorius* seeds (250 mg/kg) in albino rats for 30 and 60 days induced infertility in males which was reversible.[40] Dose-dependent degenerative changes in the testicular weights, sperm count, spermatogenesis and Leydig cells were observed in testes of rats treated with steroidal fraction of the seeds.[41] The methanol extract of the seeds caused a concentration-related impairment of percentage human sperm motility with an EC_{50} of 2.29 mg/ml.[42]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

2. *Adiantum capillus-veneris* L. (Polypodiaceae)

Black Maidenhair Fern, Southern Maidenhair Fern,
Venus Hair Fern

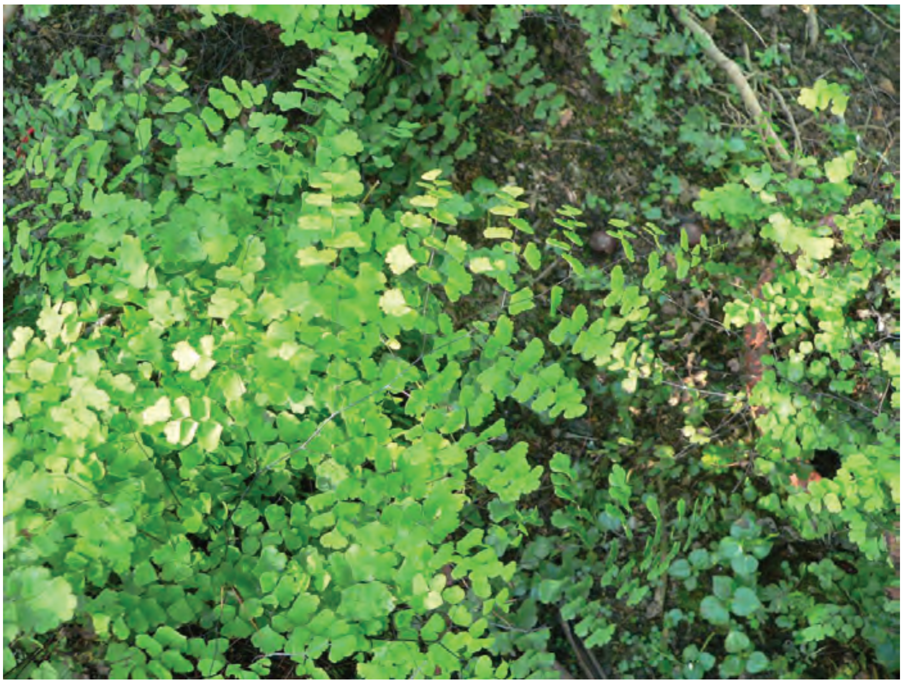

Adiantum capillus-veneris fern

Description: *Adiantum capillus-veneris* L. is a perennial fern with short creeping stems. Leaf blades are lanceolate, pinnate, 10–45 cm by 4–15 cm and glabrous. Ultimate segments are various but generally cuneate or fan-shaped to irregularly rhombic, about as long as well as broad with its base broadly to narrowly cuneate. Plant is delicate, brittle and has dark stalks.[1–6]

Origin: Native to America, Mexico, West Indies and South America. It can also be found in temperate regions of Eurasia and Africa.[6]

Phytoconstituents: Adiantoxide, adiantone, isoadiantone, isoadiantol, hydroxyadiantone, capesterol and others.[1,4,7–11]

Traditional Medicinal Uses: The fern is considered an astringent, demulcent, depurative, diaphoretic, diuretic, emmenagogue, emollient, expectorant, laxative, stimulant, sudorific and a tonic.[1] It is also used for alopecia, asthma, bladder ailments, catarrh, chest ailments, chills, cold, dropsy, dysmenorrhoeal, fever, head ailments, hepatosis, labour, lung ailments, respiratory problems, rheumatism, sclerosis, snake bite, sores, sore throat, splenosis, stones and other urinary calcification.[1,4] The plant expels worm, induces vomitting and relieves fever. Used externally, it is poulticed on snakebites and as a treatment for impetigo.[1,3] The Russians use the herb for rhinitis while the French, with orange flowers and honey, uses it for pulmonary catarrh.[1] A handful of leaves is made into a tea and drunk as an expectorant, astringent, tonic for coughs, throat afflictions and bronchitis.[1] The plant is also used as a hair wash for dandruff and to promote hair growth in Latin America.[1,3] The ashes mixed with vinegar and olive oil is rubbed into the scalp to cure alopecia.[5] In Traditional Chinese Medicine, the leaves are used for bronchial diseases and as an expectorant;[3] and in Africa, the leaves are smoked to prevent head and chest colds.[1]

Pharmacological Activities: Antibacterial.[12]

Dosage: Taken internally as a tea prepared from powdered dried fronds. The standard single dose is 1.5 g of drug to 1 cup of fluid per dose.[13]

Adverse Reactions: No known side effects with therapeutic dosages.[13]

Toxicity: No information as yet.

Contraindications: Should not be used during pregnancy.[13]

Drug-Herb Interactions: No information as yet.

3. *Allamanda cathartica* L. (Apocynaceae)

Allamanda, Common Allamanda, Golden Trumpet

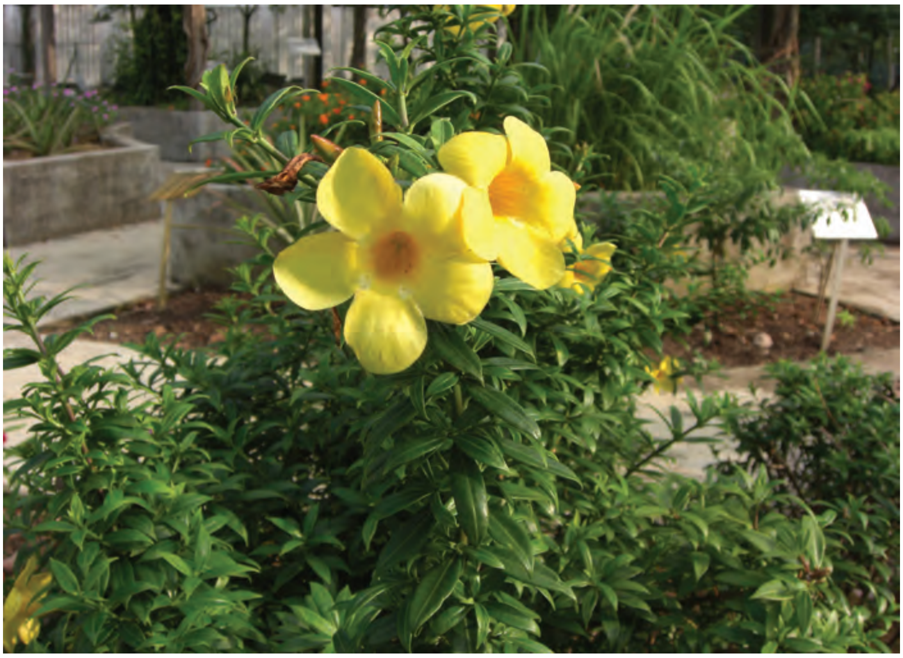

Allamanda cathartica flowers and leaves

Description: *Allamanda cathartica* L. is a woody shrub that can grow up to 4 m tall. The stems exude milky white sap when incised. Leaves are simple, exstipulate, glossy, leathery and glabrous. Leaf blade is oblong-lanceolate, 8–15 cm by 4–5 cm and arranged in opposites of 3–5 sessiles. Flowers are large, tubular, bright yellow and 4–5 cm long.[1–3]

Origin: Native to South America; cultivated in China for medicine.[4]

Phytoconstituents: Allamandin, plumericin, plumieride, ursolic acid and others.[3,5–10]

Traditional Medicinal Uses: The plant has been used as a purgative to induce vomiting at low dosage.[3] Its leaves are cathartic and the bark is used as a hydragogue for ascites.[11] In Surinam's traditional medicine, its roots are used against jaundice, for complications with malaria and enlarged spleen.[11]

Pharmacological Activities: Anthelmintic,[12] Antifungal,[7,13–15] Antineoplastic,[16] Antivenom[17] and Wound healing.[18]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: Every part of the plant was reported to be poisonous.[19] The sap of the plant was reported to cause mild and occasional oral irritation and slight nausea when sucking cut stems. Rash or dermatitis were also reported when sap was in contact with sensitive skin.[19]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

4. *Aloe vera* Mill. (Aloaceae)

Aloe, Lidah Buaya

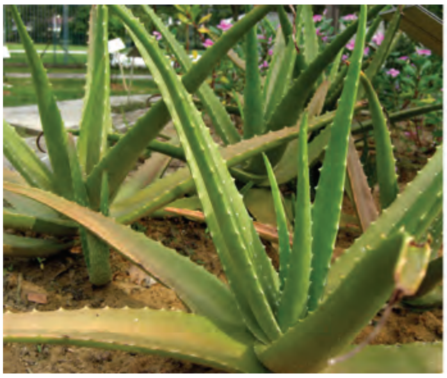

Aloe vera

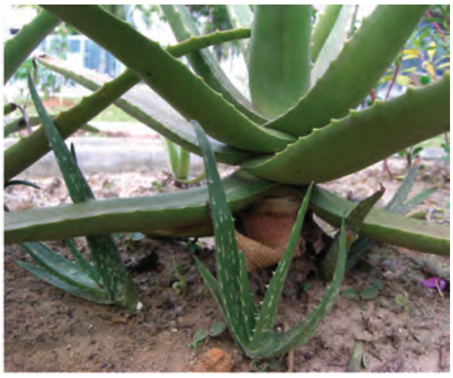

Young and old plants

Description: *Aloe vera* Mill. is a short-stemmed, up to 50 cm, succulent herb with thick green leaves that have a sharp and pointed apex, 15–50 cm by 4–7 cm, arranged in a rosette around the short stem. Blade is green to variegate with small white or glaucous dots, irregular bands, lanceolate, tapering from base to apex, glabrous with green and spiny-toothed margins. The leaves contain a thick colourless juice. Flowers are yellow, orange or red, crowded into a rosette and in panicles.[1–3]

Origin: Native to North Africa; cultivated in China for medicine and widely used as indoor ornamental plants.[3,4]

Phytoconstituents: Aloin (barbaloin), arabinose, aloe-emodin, aloetinic acid, emodin, aloeresin A–C, aloesone, aloeride and others.[2,5–12]

Traditional Medicinal Uses: The plant has been used in cosmetic preparations for the treatment of pimples, acne and mouth ulcers.[2] It has also been used to control bleeding, itching of piles, and relief from arthritic pains.[2] The Chinese uses the plant juice as a mild laxative, wash for piles, abscesses and scabies. In the Philippines, it is used to treat dysentery and pain in the kidneys.[2] The plant has been found to treat bacterial infection, as a cathartic, emmenagogue, purgative and vermifuge. It can be used in the treatment of burns, oedema, pain, swellings and wounds; treatment of leukemia, lung cancer; treatment of constipation, eczema, piles and pertussis.[6,13–15] The whole plant has also been used for the treatment of rectal fissures and piles while the root is used to treat colic.[16] The juice from the leaves is used to increase menstrual flow.[13] The jelly is used as aperient, for wounds, applied on the abdomen in fever, after confinement, on swelling and is especially useful in correcting constipation due to intake of iron medication.[17,18] Fresh juice of the leaves is cathartic and cooling and used for various eye diseases.[16] The dried juice is applied with lime juice for reducing swellings and promoting granulation in ulcers. In Malaysia, it is used for treating wounds, fever, swellings and put on the abdomen of women after confinement. The mixture of sugar with sap obtained from heated leaves is taken for asthma. The mucilaginous flesh and the sap are used for burns. The watery extract is used as a hair tonic. It is also used in cosmetics for decreasing wrinkles and other skin problems. It is mixed with milk and given for dysentery and pains in the kidney.[15] It is used in Ayurveda to alleviate pain and is also mentioned in folk medicine of Arabian Peninsula for the management of diabetes.[19]

Pharmacological Activities: Angiogenic,[20] Antifungal,[21–23] Antidiabetic,[24,25] Anti-inflammatory,[12,26–30] Anticancer,[31–38] Antimicrobial,[39] Antioxidant,[11,40–45] Antiproliferative,[12,38] Chemopreventive,[35,46] Gastric mucosal protection,[47,48] Hepatoprotective,[19,49,50] Neuroprotective,[51] Hypolipidaemic,[52] Immunomodulatory,[11,33,53–57] Immunostimulatory,[58] Antimutagenic,[31] Alloantigenic,[59] Antileishmanial,[60,61] Prevention of kidney stones,[62,63] Radioprotective[64,65] and Wound healing.[66–69]

Dosage: Single dose of powdered Aloe, 50–200 mg at bedtime; tincture BPC 1949 (1:40, 45% ethanol), 2–8 ml. Aloes should only be taken for short periods, maximum 8–10 days.[70] Doses of 10–30 mg act as a bitter stomachic; 60–200 mg as a laxative and 300–1000 mg as a purgative.[18,71] A dose of 1 teaspoon after meals, or otherwise advised by manufacturers and practitioners has been reported.[72] To prevent kidney stones, a dose of 2 to 3 tablespoon daily is reported. As a laxative, the recommended dose is 500 to 1000 mg daily. For burns or wound healing, fresh gel from plant may be applied topically and liberally. For haemorrhoids, as a stool softener, 0.05 to 2 g of dry aloe extract is administered. In the treatment of HIV, 800 to 1600 mg of acemannan daily (equivalent to 0.5 to 1 L of *Aloe vera* juice) is administered. To relieve constipation, 20 to 30 mg hydroxyanthracene derivatives daily, calculated as anhydrous aloin is prescribed.[73]

Adverse Reactions: Barbaloin was shown to have a laxative effect.[70] Ingestion of *A. vera* is associated with diarrhoea, electrolyte imbalance, kidney dysfunction, and conventional drug interactions; episodes of contact dermatitis, erythema, and phototoxicity have been reported from topical applications.[74] *A. vera* could also induce acute liver damage.[75]

Toxicity: Severe gastrointestinal cramping can occur if the latex (which is just below the leaf surface) is taken internally.[72] Toxic doses cause severe haemorrhagic diarrhoea and kidney damage, and sometimes death. The lethal dose of the dried plant extract is stated to be 1 g/day taken for a period of several days.[76]

Contraindications: Contraindicated in intestinal obstruction, acute inflammatory intestinal diseases (e.g., Crohn's disease, ulcerative colitis), appendicitis and idiopathic abdominal pain. Should not be used during pregnancy. Should not be given to children below 12 years of age.[77] Should avoid application of *Aloe* topically on deep, vertical wounds. Contraindicated in menstruation and if the person has kidney complaints.[72]

Drug-Herb Interactions: Increase the actions of cardiac glycosides and antiarrhythmic drugs (chronic use of aloe causes potassium loss), thiazide diuretics, loop diuretics, licorice and corticosteroids.[77] Aloe gel, when taken orally, can reduce the absorption of many medications. Thus, it should be taken two hours apart from all medications.[72] A study reported that *Aloe vera* preparations improved the absorption of both vitamins C and E.[78]

5. *Andrographis paniculata* (Barm.f.) Nees (Acanthaceae)

Hempedu Bumi, Sambiloto, Chuan Xin Lian

Andrographis paniculata plants

Andrographis paniculata flowers

Description: *Andrographis paniculata* (Barm.f.) Nees is an annual herb that grows up to 1 m in height. Stems are glabrous and articulated. Leaves are simple, opposite and exstipulate. Blade is dark green, bitter, glossy, simple, lanceolate, opposite and 4–8 cm by 1.3–2.5 cm. Its small and white flowers grow in terminal or axillary panicles. Both the bracts and the 5-lobed calyx are small. Fruits are upright, fusiform, capsular and contains 2–4 seeds.[1]

Origin: Native to Indian subcontinent and cultivated elsewhere.[2]

Phytoconstituents: Andrographolide, andropanolide, andrographic acid and andrographidine A, andrographatoside, andropaniculosin A and andropaniculoside A and others.[3–14]

Traditional Medicinal Uses: The plant is used orally to prevent and treat common cold, influenza, pharyngotonsilitis, allergies and sinusitis. Traditionally, it is used for many conditions including anorexia, atherosclerosis, insect and snake bites, bronchitis, prevention of cardiovascular disease, diabetes, hypertension, cholera and as a tonic.[1,15,16]

Pharmacological Activities: Antiapoptotic,[17] Antibacterial,[18–20] Antifungal,[18] Anticancer,[21–30] Antidiabetic/Hypoglycaemic,[31–36] Antifertility,[37] Anti-inflammatory,[38–48] Antioxidant,[45,49–51] Antiplatelet,[52,53] Antiprotozoal,[54–56] Antiviral,[57,58] Cardioprotective,[59] Chemopreventive,[60] Hepatoprotective,[61–63] Hypotensive,[64] Immunomodulatory,[35,42,65,66] Psychopharmacological activities,[67] Vasorelaxant[68] and Cytotoxic.[12]

Dosage: For decreasing symptoms of common cold, doses of 400 mg of standardised andrographolide are required three times daily; for preventing common cold, a dose of 200 mg daily for 5 days in a week; for relieving fever and sore throat in pharyngotonsilitis, doses of 3 g and 6 g daily were used.[16] Use 6–9 g for influenza with fever, sore throat, ulcers in the mouth, acute or chronic cough, colitis, dysentery, urinary tract infection, carbuncles, sores and venomous snake bite.[69]

Adverse Reactions: Orally, large doses of Andrographis may cause gastrointestinal distress, anorexia, emesis and urticaria. Androgapholide taken orally at 5 mg/kg three times a day may cause headache, fatigue, rash, abnormal taste, diarrhoea, itching, lymphadenopathy, anaphylactic reactions, etc.[16]

Contraindications: Contraindicated in pregnancy, likely to be unsafe due to abortifacient effect.[16]

Toxicity: No toxic effect was observed after administration of a decoction of *Andrographis paniculata* leaves to rabbits.[70] LD_{50} of androgapholide in mice through oral route is > 40 g/kg body weight, which indicates low toxicity.[71]

Drug-Herb Interactions: Simultaneous application of *A. paniculata* and warfarin did not produce significant effects on the pharmacokinetics of warfarin, and practically no effect on its pharmacodynamics.[72]

6. *Ardisia elliptica* Thunb. (Myrsinaceae)

Mata Pelanduk/Ayam, Sea-Shore Ardisia, Shoebutton Ardisia

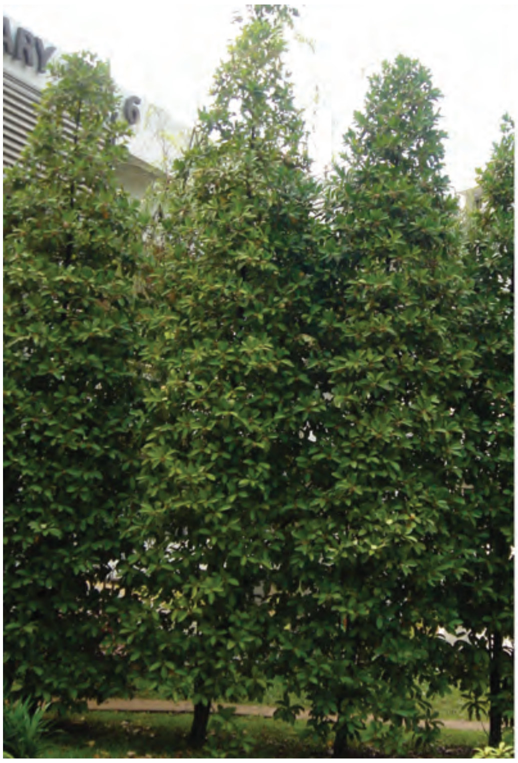

Ardisia elliptica trees

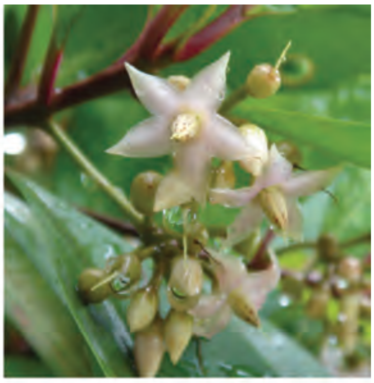

Flowers of *Ardisia elliptica*

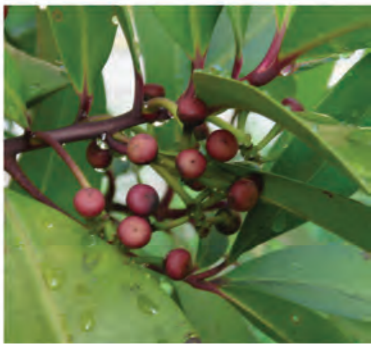

Ardisia elliptica fruits

Description: *Ardisia elliptica* Thunb. is a small shrub that can grow up to 10 m tall. The leaves are obovate, 6–9 cm long with smooth margins. They have an acute apex and a cuneate leaf base. The leaves have a leathery texture. The plant bears whitish pink axillary inflorescences and the drupes are globular, 1–1.2 cm in diameter and grow in clumps, pale red when immature, and turning dark purplish upon maturity.[1–3]

Origin: Native to tropical and temperate Asia.[4]

Phytoconstituents: Rapanone,[5] bauerenol, α- and β-amyrin,[6,7] syringic acid, isorhamnetin and quercetin,[8] bergenin,[9] 5-(*Z*-Heptadec-4′-enyl)resorcinol and 5-pentadecylresorcinol.[10]

Traditional Medicinal Uses: The decoction of the leaves is used by the Malays to treat pain in the region of the heart.[10] The *Kadazan Dusun* tribes in Malaysia used paste made from the leaves of *Ardisia elliptica* to treat herpes and measles.[1,11] The fruits are used in Thai Traditional Medicine to cure diarrhoea with fever.[8]

Pharmacological Activities: Antiplatelet[10,11] and Antibacterial.[8]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

7. *Areca catechu* L. (Palmae)

Betel Nut Palm, Areca Nut, Pinang

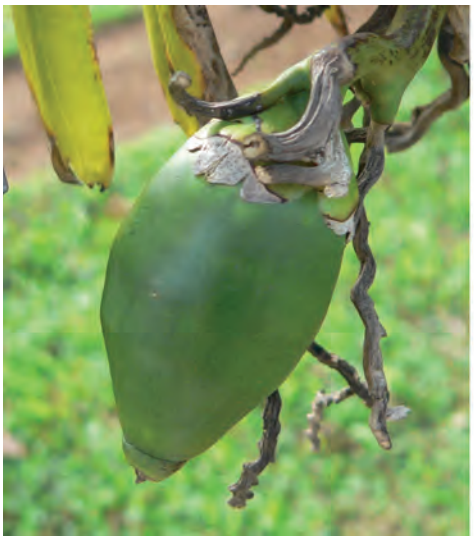

Fruit of *Areca catechu*

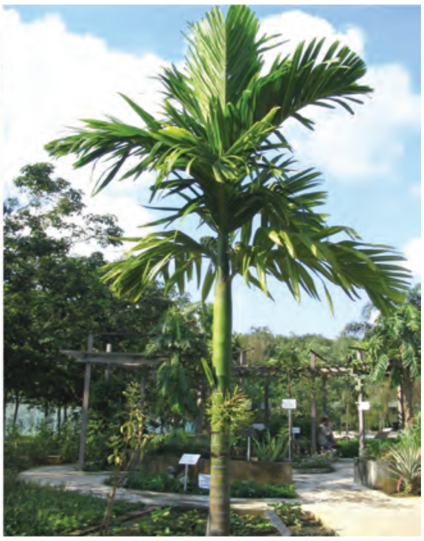

Areca catechu tree

Description: *Areca catechu* L. is a tall, slender palm that can grow up to 10 m. Leaves are dark green, pinnate and up to 1.2–2 m long. Inflorescence is branched and male flowers grow in one row surrounding the female flower at the base of branch. Fruit is a one seeded ovoid berry about 5 cm long.[1-4]

Origin: Originate from the Philippines.[5]

Phytoconstituents: Arecoline, arecaidin, arecaine, catechin, glucides, guvacine, guvacoline, arecolidine, isoguvacine, nicotine and others.[1,6–8]

Traditional Medicinal Uses: In Irian Jaya, parts of this tree are used on wounds, swellings and other skin afflictions. The pericarp is effective in the treatment of flatulence, oedema, dysuria and hyperemesis of pregnancy. On the Finschhafen coast, Papua New Guinea, the inner seed is chopped, heated over fire and pressed on sores caused by sea urchins. Chewing the betel nut with lime and the leaves of catkins or *Piper betel* gives a stimulant effect as well as an attributed sedative effect.[9] This may also be used to soothe a mad person. The red mixture is applied to ulcers in New Britain and to treat sores caused by venereal disease in Northern Province; whereas in New Ireland, the scraped bark is mixed with sea water with a leaf of *Indocarpus fagiferus* and drunk to treat asthma.[9] The Malays use a decoction of the leaves to treat diarrhoea in children.[1] The dried ripe fruits have been used by the Chinese to expel tapeworms and roundworms, treat diarrhoea, indigestion, lumbago, urinary problems and increase menstrual flow.[1,10] The kernel of the fruit is chewed as a narcotic, fresh or cured with slacked lime and betel leaves.[2] The fruits are also used for beriberi, dysentery, dyspepsia, dysuria, oedema and malaria.[6] The fruit can also be applied on venereal sores (ground fresh nut with betel leaf, *Nigella sativa* and roots of *Gymnema hirsutus*, cooked in mustard oil or butter and applied).[11]

Pharmacological Activities: Analgesic,[12] Anthelmintic,[6,13] Antibacterial,[14] Anticancer/Antineoplastic,[15,16] Anticonvulsant,[17] Antidepressant,[18,19] Antihypertensive,[20,21] Antimitotic,[22] Antioxidant,[23] Apoptotic,[24] Hypocholesterolaemic,[25] Immunomodulatory,[26,27] Immunostimulatory,[28–31] Antihyaluronidase,[32] Antivenom,[33] Cell growth inhibitor,[34] Molluscicidal[35] and causes periodontitis.[36]

Dosage: For the treatment of diarrhoea, 30 g areca powder in 200 cm^3 water, simmered for 1 hour is taken before breakfast. If expulsion does not take place within 9 hours, 50 cm^3 of 50% magnesium sulphate solution may be taken.[11] A decoction of the pericarp has been prescribed in a daily dose of 6 to 12 g to treat flatulence, oedema, dysuria and hyperamesis during pregnancy. To treat diarrhoea and dysentery, a daily dose of 0.5 to 4 g of the kernel has been used.[7] For sore throat, the pressed juice is used as a gargle. 2 g of fresh nut can be chewed for 15 min or more before spitting it out. Another reported usage is rolling the leaves and placing them between teeth and gums/lips.[37]

Adverse Reactions: Excessive chewing can cause dizziness, nausea, vomiting, diarrhoea, and seizures.[38]

Toxicity: A dose of 8–10 g is toxic to humans.[6,39] Heavy consumption may cause the development of cancer in the upper and middle third of the oesophagus respectively[40] and chronic kidney disease.[41] Betel quid (a mixture of areca nut and flavouring ingredients with or without processed tobacco leaves) chewing resulted in a statistically significant increase in the risk of total and cerebrovascular deaths in the elderly population.[42] At higher doses, bradycardia, reflex excitability, tremor, spasms and eventual paralysis may occur. Long term effect as stimulant causes malignant tumours of oral cavity through formation of nitrosamines.[39] Betel quid chewing during pregnancy has a substantial effect on a number of birth outcomes, including sex ratio at birth, lower birth weight and reduced birth length.[43] It is toxic during pregnancy[44,45] as it also possesses cytotoxic and genotoxic activities.[46,47]

Contraindications: Should not be used during pregnancy and lactation. Should not be given to children. Patients with oral or oesophageal cancers, ulcers, oesophagitis, or renal disease should avoid its use.[37]

Drug-Herb Interactions: Decrease action of antiglaucoma agents. Increase action of beta-blockers, calcium channel blockers, cardiac glycosides (digoxin, digitoxin). For neuroleptics, extrapyramidal symptoms can occur.[37] Avoid taking alcohol and atropine.[38]

8. *Asplenium nidus* L. (Aspleniaceae)

Bird's Nest Fern

Asplenium nidus growing on the ground

Asplenium nidus on a tree

Description: *Asplenium nidus* L. is a common epiphytic fern found growing on trees. Fronds are long, simple, green and grow from a central rhizome attached to the tree branch in the shape of a nest. Parallel lines of spores are found on the undersides of the fronds and radiate away from the midrib towards the leaf margin.[1]

Origin: Native to tropical Africa, temperate and tropical Asia and Australasia.[2]

Phytoconstituents: Kaempferol-3-*O*-gentiobiosie-7,4′-bisglucoside, kaempferol-3-*O*-diglucoside, kaempferol-3,7-diglycoside and kaempferol-3-*O*-vicianoside.[3]

Traditional Medicinal Uses: *A. nidus* is regarded as depurative. Infusion of the fronds is used to ease labour pains by Malaysia native tribes. The Malays pound the leaves in water and apply the resulting lotion to feverish head.[4] Two young fronds are eaten when they are still coiled, just after menstruation, in the morning as a contraceptive. Tea made from the fronds is recommended for general weakness.[5,6]

Pharmacological Activities: Oxytocic activity.[6]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

9. *Aster tataricus* L.f. (Compositae)

Tatarian Aster, Tatarian Daisy

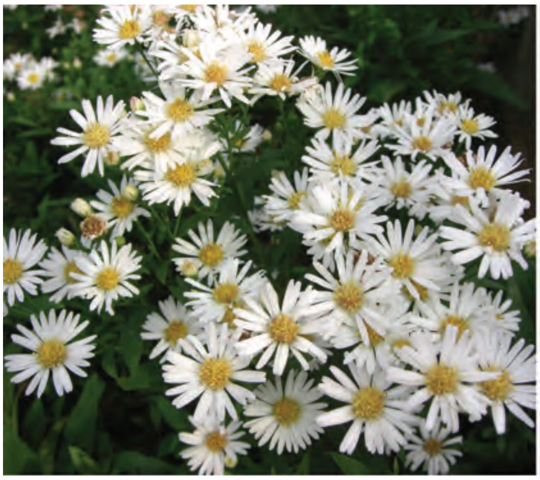

Aster tataricus flowers

Aster tataricus shrub

Description: *Aster tataricus* L.f. is a small shrub with abundant fibrous roots. Leaf blades are oblanceolate to lanceolate, margins serrate or entire, 4–18 by 1–5 cm and acute. Flower heads are in large bunches with white petals and yellow centre.[1–3]

Origin: Native to Siberia.[3]

Phytoconstituents: Shinone, friedelin, epifriedelinol, shinoside A–C, asterinin A–F, astins A–E, astertarone A&B and others.[4–11]

Traditional Medicinal Uses: The underground rootstock is used as a purgative, treats colds, coughs with excessive sputum or with blood and painful menstruation.[2] It is also used as a bechic-expectorant.[12] Used as an aromatic tonic in chronic gastroenteritis.[13]

Pharmacological Activities: Anticancer[14,15] and Antioxidant.[4]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

10. *Azadirachta indica* A. Juss. (Meliaceae)

Neem

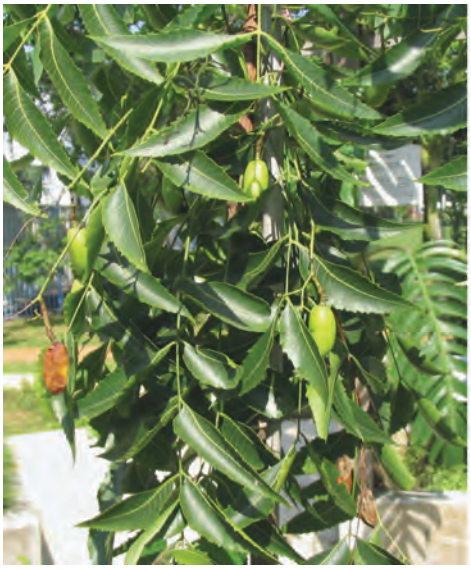

Fruits of *Azadirachta indica*

Azadirachta indica tree

Description: *Azadirachta indica* A. Juss. is a tall evergreen tree, growing up to 30 m in height. Leaves are pinnate with opposite or alternate, lanceolate, serrated and glabrous 8–16 leaflets, 20–32 cm long. Flowers are yellowish white. Fruits are small, ellipsoid, about 5 cm long and green.[1–3]

Origin: A native of India and China, cultivated and naturalised throughout India, Malay Peninsula, Indonesia and Pakistan.[4]

Phytoconstituents: Azadirachtin O–Q, nimbin, deacylnimbin, salanin, nimbidin, nimbinin, nimbidol, azadirone, melianol, meliacinol, nimbothalin, nimonol, azharone and others.[3,5–28]

Traditional Medicinal Uses: It is used for the treatment of a variety of human and veterinary ailments including head lice, mange, fleas, fever, convulsions, leprosy, scrofula, rheumatism, asthma, worm infestations, treat bacterial infection, insecticide, local application for indolent ulcer and consumed as tonic after childbirth.[8] It is used for boils, heart disease, fever, tuberculosis, diarrhoea, jaundice, dysentery, to promote healing, measles, smallpox, sores, inflamed gums, syphilis, leprosy, piles, urinary diseases, to expel worms, purgative, emollient, local stimulant, treat fever (crushed leaves added to lemon), disinfectant (oil from nuts), astringent, contraceptive and tonic.[29,30] It is also used for dermatological problems in Nigeria.[31] Neem has also been used to protect crops, stored grains and library books from insects. Neem leaves buried in grain bins are used to keep the stored crops insect-free. Crushed seeds soaked in water produce a potent pesticide that does not harm mammals, birds, earthworms and bees.[3]

Pharmacological Activities: General review.[32] Antibacterial,[33–37] Anticancer,[38–46] Anticarcinogenic,[47–55] Anticonvulsant,[56] Antifertility,[57,58] Antifungal,[35,59–61] Anti-inflammatory,[62] Antimalarial,[63–68] Antimicrobial,[69] Antioxidant,[70–74] Antiproliferative,[75,76] Antipyretic,[77] Antiviral,[78,79] Gastroprotective,[80–82] Hepatoprotective,[83–85] Hypoglycaemic,[86,87] Hypotensive,[88–90] Immunostimulatory,[38,91–93] Neuroprotective,[94] Anthelmintic,[95] Antihaemorrhagic,[37] Antileishmanial,[96] Antimutagenic,[97–99] Molluscicidal,[100] Insecticidal and Insect repellent.[101–110]

Dosage: Approximately 100 g of bark is soaked in 1 L of water and approximately 3 ml of this infusion is consumed daily for one month as a male contraceptive.[30]

Adverse Reactions: Nausea, vomiting, anorexia, hypersensitivity, Reye's syndrome (infants).[111]

Toxicity: Mice injected with the tetranortriterpenoid fractions (>86 mg/feeding) into the tail vein died within 24 hours.[8] Various acute, subacute and chronic toxicity tests of extracts of neem have been reported, especially on human and animal fertility.[112] Non-aqueous extracts appeared to be the most toxic neem-based pesticide products.[112]

Contraindications: Should not be used during pregnancy and lactation, and in children. Should not be used in persons with hypersensitivity to neem.[111]

Drug-Herb Interactions: No information as yet.

[**Authors' Note:** An extensive review[112] on the safety of neem derived pesticides concluded that the use of neem derived pesticides as an insecticide should not be discouraged.]

11. *Barringtonia asiatica* L. (Lecythidaceae)

Beach Barringtonia, Fish-killer tree, Putat Laut

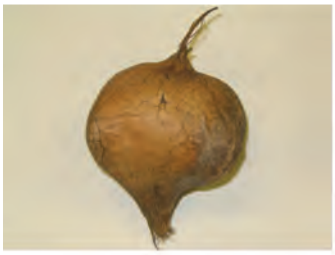

Barringtonia asiatica fruit

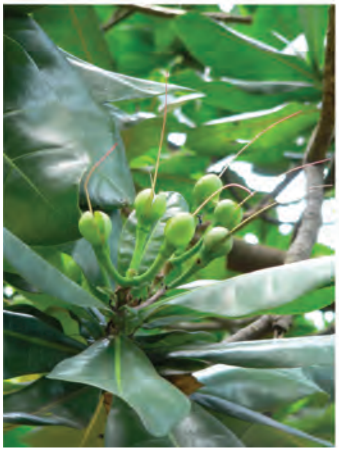

Developing fruits of *B. asiatica*

Barringtonia asiatica tree

Description: *Barringtonia asiatica* L. is a large tree bearing large simple leaves, 20–30 cm long which taper to the leaf base. Flowers are large and white with several white stamens. The flowers are actinomorphic and have four petals. The fruit is oblong, green, large, 8–10 cm across and contains one seed.[1–3]

Origin: Native to Africa, temperate and tropical Asia and Australasia.[4]

Phytoconstituents: A_1-barrinin, ranuncoside VIII, A_1-barrigenin and others.[4,5–7]

Traditional Medicinal Uses: The plant is used to treat fungal infections,[8] burns and wounds.[9] The leaves are heated and used to treat stomachache and rheumatism in the Philippines. Its fruits are used as a fish poison and the fruit juice for controlling scabies while the seeds are used for the expulsion of intestinal worms and also as a fish poison.[10] They are also used to treat sores, cough, influenza, sore throat, diarrhoea, swollen spleen after malaria.[11] In other provinces of Vietnam, the fresh nut is scraped and applied to sores; dried nut is ground into a powder, mixed with water and drunk to cure coughs, influenza, sore throat, bronchitis, diarrhoea and swollen spleen.[12] The bark is used in the treatment of tuberculosis.[11] In Yambio (Sudan), the inner bark is crushed and mixed with water and drunk to ease the aching associated with malaria. It is also used in combination with other plants as a medicine to treat tuberculosis in New Ireland and the Solomon Islands.[12]

Pharmacological Activities: Insect repellent,[5] Antibacterial and Antifungal.[13]

Dosage: To relieve the aching associated with malaria, inner bark is crushed and mixed with water and drunk, 2 cups per day for 2 days.[12]

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

12. *Barringtonia racemosa* (L.) K. Spreng
(Lecythidaceae)

Putat Kampong, Samundrapandu

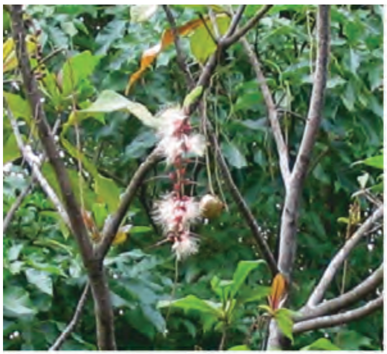

Barringtonia racemosa flowers

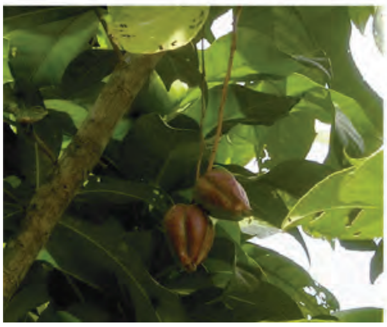

Barringtonia racemosa fruits

Barringtonia racemosa tree

Description: *Barringtonia racemosa* (L.) K. Spreng is a small tree with large, simple leaves about 20 cm long. Flowers are large with numerous pinkish stamens. The fruit is ellipsoid, green, 8 cm long, and turns red upon maturity.[1,2]

Origin: Native to Africa, temperate and tropical Asia, Australasia and the Pacific.[3]

Phytoconstituents: Nasimalun A and B, barringtonin, R_1-barrigenol, R_2-barrigenol, barringtogenol, barringtogenic acid and others.[4–9]

Traditional Medicinal Uses: The plant is widely used in the form of a decoction in Sri Lankan traditional medicine.[10] The leaves and bark are used for rat and snake bites, rat poisoning and on boils.[10,11] The fruits are used for cough, asthma and diarrhoea.[11,12] Kernels of the drupe are mixed with milk to treat bilious diseases and jaundice whereas the seed has been used as an insecticide and tonic.[11] The seeds along with other ingredients are employed in preparations for the treatment of itch, piles and typhoid fever, while the bark is also used for gastric ulcers.[10] Its roots act as a coolant and deobstruent.[11]

Pharmacological Activities: Antinociceptive,[10] Antibacterial,[12] Glucosidase and Amylase Inhibition,[13] Anticancer [14] and Cytotoxic.[15]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: Aqueous extracts (500, 750, 1000 or 1500 mg/kg) of *B. racemosa* bark in male rats did not produce any unwanted side effects or toxicity or alter fertility, gestational length, peri- and neonatal development and appeared to be non-teratogenic.[10] *B. racemosa* seed extract was found to be devoid of acute and short-term toxicity to mice, when administered daily, intraperitoneally for a fortnight up to a dose of 12 mg/kg. The treated mice showed conspicuous toxic symptoms only at 24 mg/kg. The LD_{50} of male mice for a single i.p. dose is 36 mg/kg.[14]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

13. *Bauhinia purpurea* L. (Leguminosae)

Butterfly Tree

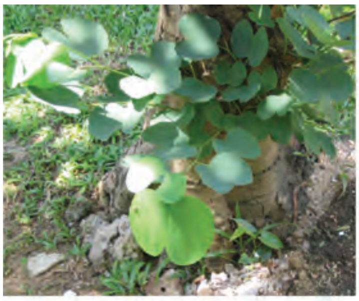

Leaves of *Bauhinia purpurea*

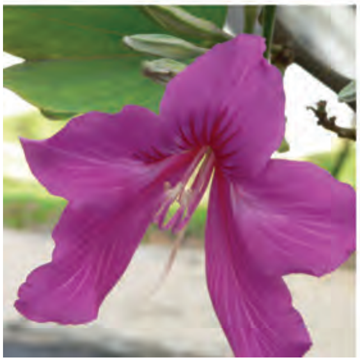

Flower of *Bauhinia purpurea*

Bauhinia purpurea tree

Description: *Bauhinia purpurea* L. is a deciduous tree. Leaves are simple and stipulate. Blade is butterfly shaped, coriaceous with 9–11 pairs of secondary veins. Flowers are showy, pink and arranged in axillary. The fruits are darkish pods, 1.5 by 15 cm and woody.[1,2]

Origin: Native to India and grown in the Asia-Pacific as ornamental plants.[2]

Phytoconstituents: Bauhiniastatins 1–4, bauhinoxepin C–J, bauhibenzofurin A, bauhispirorin A, bauhinol E and others.[3–9]

Traditional Medicinal Uses: The root is grated with water and is drunk to treat common fever. The flowers are used as laxative and leaves applied to sores and boils, and for cough treatment.[10]

Pharmacological Activities: Antibacterial,[9,11] Antifungal, Antimalarial, Cytotoxic,[9] Anticancer,[8] Anti-inflammatory[9,12] Antinociceptive, Antipyretic[12] and Thyroid hormone regulating.[13]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

14. *Bixa orellana* L. (Bixaceae)

Annatto, Lipstick Tree

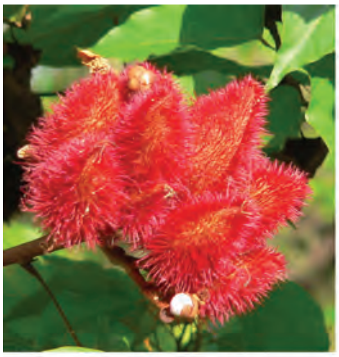

Fruits of *Bixa orellana*

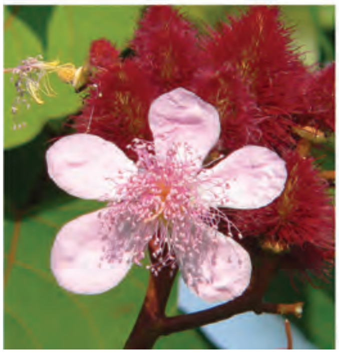

Bixa orellana flower

Bixa orellana tree

Description: *Bixa orellana* L. is a small tree with simple and spiral leaves, 10–20 cm by 6.3–12.5 cm, dark green, ovate, acuminate, truncate at the base and glabrous. Flowers are large, 5 cm in diameter, pinkish or white, arranged in terminal panicles. Fruits are dehiscent, ovoid capsules containing 15–20 trigonous seeds in bright red pulp.[1–3]

Origin: Originate from tropical America. Cultivated Pantropically.[4]

Phytoconstituents: Bixin, valencene, β-elemene, β-selinene, copaene, δ-cadinene, spathulenol, γ-cadinene, δ-elemene, ledol, α-muurolene, α-cadinol and others.[5–10]

Traditional Medicinal Uses: The leaves have been used to treat snakebites and jaundice and the seed is considered a good cure for gonorrhoea.[11] The bark of the root is used to treat fever and as an aperient.[11] In Cambodia, the leaves are a popular febrifuge while in Indonesia, water in which the leaves are rubbed is poured over the head of children with fever. In Malaysia, the leaves are used in a postpartum medicine and in the Philippines the leaves are pounded in coconut oil and heated, then applied to the abdomen to relieve tympanites. Pastes of the fresh leaves are rubefacient and used in dysentery. In Vietnam, lotions or baths of leaves are used during fever. Its unripe fruits are emollient in leprosy.[3] Alcoholic extracts of seed coat are taenifuge and laxative.[3] Decoctions of barks are used for catarrh. Infusions of seeds are used to treat asthma and excessive nasopharynx mucus production.[3] Traditionally, it is also used as a gargle for sore throats and oral hygiene.[12] In Trinidad and Tobago, the leaves and roots are used for hypertension, diabetes and jaundice.[13] Leaves and seed pods are used as a female aphrodisiac.[14]

Pharmacological Activities: Antibacterial,[12,15–18] Anticancer,[19] Anticonvulsant,[16] Antidiarrhoeal,[16] Anti-inflammatory,[19,20] Antioxidant,[3,16] Antiplatelet,[21] Hypoglycaemic,[3,22] Immunostimulatory,[3] Sedative,[16] Antigenotoxic and Antimutagenic,[23] Antifungal, Antileishmanial[24] and Radioprotective.[3]

Dosage: Approximately 9 seed pods are boiled in 3 cups of water for 10 mins and drunk as a diuretic before each meal. 10 g powdered seed/40 ml oil for topical pastes.[25] As a female aphrodisiac, 3 leaves in 0.5 L of water and red paste of seed pods.[14]

Adverse Reactions: Urticaria and angiooedema are possible adverse reactions with annatto dye. A patient developed these symptoms and hypotension within 20 min of ingestion of annatto containing fibres.[3]

Toxicity: Toxic to dogs dosed with 60 mg/kg trans-bixin.[26] However, annatto containing 5% bixin was non-genotoxic and non-carcinogenic to rat livers even at the highest concentration tested at 1000 ppm (4.23 bixin/kg body weight/day).[27] Annatto given through gavage to Wistar rats on days 6–15 of pregnancy showed no adverse effect on the mothers and foetus.[28]

Contraindications: Trans-bixin is hyperglycaemic and should not be ingested by patients with diabetes mellitus.[25]

Drug-Herb Interactions: No information as yet.

15. *Calophyllum inophyllum* L. (Guttiferae)

Indian Laurel, Penaga Laut, Borneo Mahogany

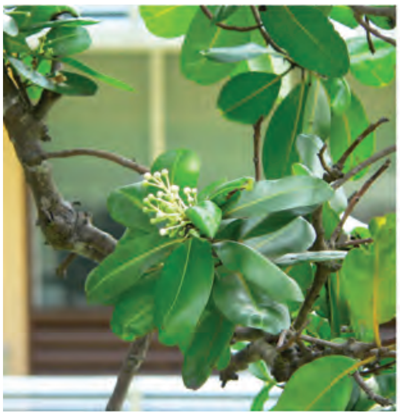

Flowers of *Calophyllum inophyllum*

Calophyllum inophyllum tree

Description: *Calophyllum inophyllum* L. is a large tree with broad, glossy, leathery, elliptic-oblong leaf blades, 8–16 cm by 4–8 cm, and with numerous parallel side veins. The tree bears sweetly scented white flowers in erect racemes. Fruits are globose, 2 cm across and are green in colour.[1,2]

Origin: It is found in Africa, tropical and temperate Asia, Australasia and the Pacific.[3]

Phytoconstituents: Inophynone, canophyllol, canophyllic acid, calophyllolide, inophyllolide, inophyllum B, C, P, and E, jacareubin, (+)-calanolide A, inocalophyllins A and B, calophinone, calophyllumin C, inophyllin A and others.[4–28]

Traditional Medicinal Uses: The whole plant is used as a crude drug for curing rheumatism and skin affections in South India.[14] Its juice is a purgative and the seed oil is specific for rheumatism and various skin diseases (i.e., scabies, ringworm and dermatosis). The bark is used for internal haemorrhage and as an astringent.[29] In Buso, Papua New Guinea, the milky latex from the leaves is diluted with water and the solution is applied to irritated eyes.[30] The gum is emetic and purgative.[29]

Pharmacological Activities: Antibacterial,[31,32] Anticancer/Antineoplastic,[19] Anti-inflammatory,[33] Antiplatelet,[34,35] Antipsychotic,[33] Antiviral,[15,18,36–38] Photoprotective,[39] Molluscicidal[40] and Piscicidal.[5]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: The unrefined oil is toxic.[41] There was a significant difference in the plasma cholesterol levels of the rats fed with *C. inophyllum* oil when compared with the control. Mild, focal to severe and widespread lesions were found in the kidneys, hearts and livers of rats fed with *C. inophyllum* seed oil.[42]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

16. *Cananga odorata* (Lam.) Hook. f. & Th.

(Annonaceae)

Kenanga, Ylang-Ylang

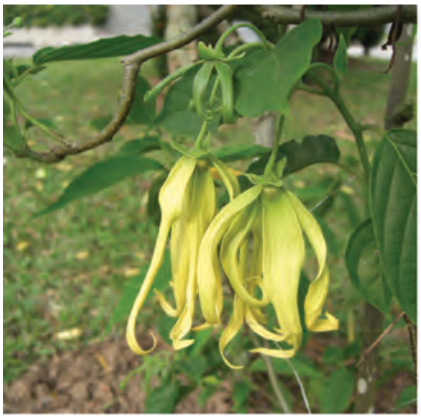

Flowers of *Cananga odorata*

Cananga odorata shrub

Description: *Cananga odorata* (Lam.) Hook. f. & Th. is a shrub which grows to a height of 4 m. Its bark is grey and smooth. Leaves are simple, alternate, exstipulate, oblong to broadly elliptic, large, 3–6.5 cm across, 8.5–29 cm long, with distinct venation pattern. Flowers are fragrant, in clusters on older branches. Fruit turns black on maturity and has many seeds in two rows.[1–3]

Origin: Originates from Indochina, Malesia and tropical Australia.[4]

Phytoconstituents: Acetogenin, aporphine, liriodenine, canangone, α-humulene, β-cubebene, germacrene D, cananodine, γ-eudesmol and others.[3,5–7]

Traditional Medicinal Uses: It is used for asthma, malaria, fever, cholera, typhoid, scabies, dermatitis, ulcer and wounds.[8] The seeds are used for stomach complaints with fever and in Indonesia, the bark is used for scabies.[3] In Malaysia, a paste of fresh flowers is applied to the chest for asthma and to treat malaria. In Solomon islands, a paste of fresh flowers is applied to boils while in India, the essential oil from the flowers makes an external remedy for cephalgia, ophthalmia and gout.[3,9]

Pharmacological Activities: Antibacterial,[10,11] Antifungal,[11,12] Antihypertensive,[3,13] Antioxidant,[10] Antineoplastic[7,12,14,15] and Antiprotozoal.[12,16]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: A 50% ethanolic root bark extract administered orally to male albino rats at the dose of 1 g/kg body weight/day for 60 days resulted in decreased epididymal sperm motility and sperm count, and morphological abnormalities in the sperms.[17] However, it is non-toxic at the current level of intake as a food ingredient (0.0001 mg/kg/day). Although sometimes Ylang-Ylang oil has been reported to cause dermal sensitisation reactions in animals and humans, it is unclear what constituent(s) within the essential oil comprise the offending agent.[18]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

17. *Capsicum annuum* L. (Solanaceae)

Chilli, Red Pepper

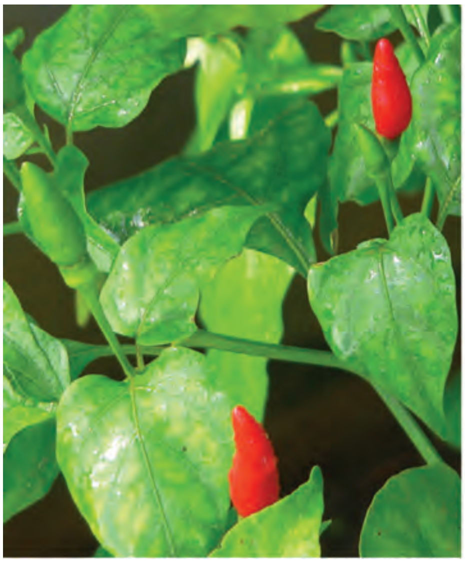

Capsicum annuum fruits

Capsicum annuum plant

Description: *Capsicum annuum* L. is a small herb that can grow up to 1 m tall. Leaves are oblong-ovate, ovate, or ovate-lanceolate, 4–13 cm by 1.5–4 cm with entire margin. Flowers are small, white or tinged purple. Fruits are mostly red, but can be green, orange, yellow and can grow up to 15 cm. Seeds are pale yellow, discoid or reniform and 3–5 mm.[1–5]

Origin: Native to Mexico and South America, widely cultivated throughout the world.[5]

Phytoconstituents: Capsaicin, capsicosides E–G, capsianosides 1–4, capsianosides VIII, IX, X, XIII, XV and XVI, solanidine, solanine, solasdine, scopoletin and others.[6–15]

Traditional Medicinal Uses: The leaves are used to treat toothache. The fruits are used to stimulate gastric activities and increase blood circulation.[4] It is also a stimulant, carminative, and used locally for neuralgia and for rheumatism.[15] Uterine pain associated with childbirth is treated with soup containing the fruit.[16] The Commision E approved *Capsicum annuum* for painful muscle spasms in areas of shoulder, arm and spines. Preparations are used to treat arthritis, neuralgia, lumbago and chilblains.[17]

Pharmacological Activities: Antibacterial,[18,19] Antifungal,[20,21] Anticancer,[22,23] Antioxidant,[10,24–26] Antiprotozoal,[27] Hypocholesterolaemic/Hypolipidemic,[19,28] Immunomodulatory,[29] Antimutagenic[30,31] and Pesticidal.[32]

Dosage: Liquid extract is prepared by percolating 100 gm of the plant extract with 60 mg of ethanol, to be used as an antirheumatic. External daily dose of semi solid preparations containing maximum of 50 mg of capsaicin in 100 gm neutral base is also used as an antirheumatic and applied to the affected area not more than 3 or 4 times daily.[33]

Adverse Reactions: Internally, it may cause gastrointestinal cramping, pain, and diarrhoea. Topically, it may cause painful irritation of mucous membrane.[34]

Toxicity: Oral LD_{50} values for capsaicin are 161.2 mg/kg (rats) and 118.8 mg/kg (mice), with haemorrhage of the gastric fundus observed in some of the animals that died. However, capsaicin is considered to be safe and effective as an external analgesic counterirritant.[35] Rabbits fed with *C. annuum* powder at 5 mg/kg per day in the diet daily for 12 months showed damaged liver and spleen. A rabbit skin irritation test of *C. annuum* fruit extract at 0.1% to 1.0% produced no irritation but caused neoplastic changes in the liver and intestinal tumours were observed in rats fed red chili powder at 80 mg/kg per day for 30 days.[35] High doses administered over extended period of time can cause chronic gastritis, kidney damage, liver damage and neurotoxic effects.[33]

Contraindications: Should not be used during pregnancy and lactation, in people with hypersensitivity and in children. Should not be used on open wounds or abrasions, or near the eyes.[34]

Drug-Herb Interactions: Reported with concomitant administration with aspirin and salicylic compounds.[33,36] Decrease the actions of α-adrenergic blockers, clonidine (anti-hypertensive) and methyldopa (antihypertensive). Hypertensive crisis with monoamine oxidase inhibitors.[34]

18. *Cassia fistula* L. (Leguminosae)

Golden Shower Tree, Indian Laburnum, Purging Cassia

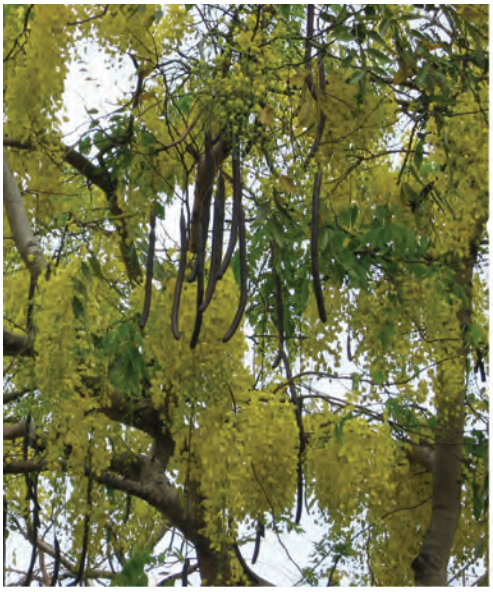

Cassia fistula fruits and flowers

Cassia fistula tree

Description: *Cassia fistula* L. is a large tree, which grows to 10 m tall. Leaves are alternate, pinnate, 3–8 pairs of leaflets, broadly ovate and pointed. Flowers are dense, bright yellow and about 4–5 cm across. Fruit pods are long, 30–60 cm, cylindrical, brown in colour and contains many seed.[1–3]

Origin: Native to India and Sri Lanka.[2,4]

Phytoconstituents: Fistucacidin, chrysophanic acid, chrysophanol, clitorin, sennosides A and B, chrysophanein and others.[2,4–9]

Traditional Medicinal Uses: The whole plant is used for anthrax, burns, cancer, constipation, convulsions, delirium, diarrhoea, dysentery, dysuria, epilepsy, fever, influenza, gravel, haematuria, pimples, syphilis, tumours and worms.[4,10] The leaf is used for skin diseases (juice), healing ulcers, for ringworm and irritation of skin (juice of young leaves), facial paralysis and rheumatism (paste).[10] The raw black pulp found between the seeds is a popular remedy for constipation.[3,4] It is also used as a cathartic, for rheumatism and snakebite (pulp), treats bacterial infections (pulp mixed with leaves of *Cassia angustifolia*), liver complaints, heart disease, reduce fever, as abortifacient, demulcent and is useful in liver, throat, eye diseases, convulsions and sores. The seed is an emetic, carminative, appetiser, and is used for constipation, jaundice, cancer on face and syphilis.[10] The roots act as a purgative while the rootbark is used for cleansing wounds.[4] The root is also used as an astringent, tonic, febrifuge, for skin diseases, leprosy, tuberculous glands, syphilis and epilepsy.[10]

Pharmacological Activities: Antimicrobial,[11–14] Anticholinergic,[15] Antifertility,[16] Anti-inflammatory,[17] Antineoplastic,[18] Antioxidant,[19–22] Depressant,[23] Hepatoprotective,[24–26] Hypocholesterolaemic,[27] Antileishmanial,[28] Larvicidal[29] and Wound healing.[30,31]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: In cases of overdose or prolonged administration, loss of electrolytes, especially potassium ions, aldosteronism, albuminuria, haematuria, inhibition of intestinal motility and muscle weakness may occur. Rarely, cardiac arrhythmia, nephropathy, oedema, and accelerated osteoclasis may arise.[32]

Contraindications: Contraindicated with acute inflammatory diseases of intestine and appendicitis. Should not be used during pregnancy and while nursing. Should not be used in children under 12 years of age.[32]

Drug-Herb Interactions: Interaction with anthranoid laxatives.[33] Enhancement of effects of cardioactive steroids may occur. Effects of antiarrhythmics may also be affected.[32]

19. *Catharanthus roseus* (L.) G. Don

(Apocynaceae)

Madagascar Periwinkle, Rose Periwinkle

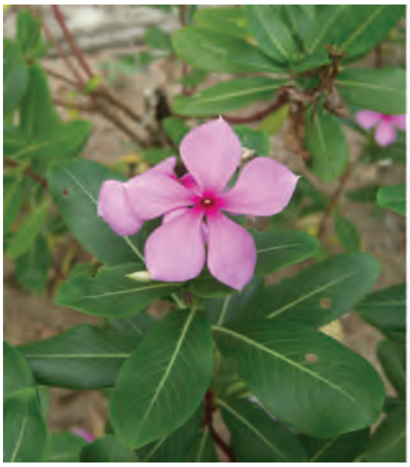
Catharanthus roseus flower

Catharanthus roseus shrub

Description: *Catharanthus roseus* (L.) G. Don is a herb, up to 80 cm tall. Stem is woody, slightly branched and all parts contain white milky latex. Leaves are simple, dark green, glossy, obovate-elliptic, 4–5 cm by 2–3 cm with prominent lateral veins on the abaxial surface. Flowers are bisexual, white, purple, pink, red or white with a red or pink centre. Fruits consist of pairs of greenish succulent follicles, 2–3 cm long and contain small oblong seeds.[1,2]

Origin: Native to Madagascar. Cultivated or naturalised in all tropical countries.[2,3]

Phytoconstituents: Vinblastine, vincristine, leurosine, akuammicine, carosine, catharanthine, catharicine, catharine, catharosine, cathovaline, catharanthiole, vindoline, vindolinine, vincaleucoblastine, secologanin, mauritianin, rosicine and others.[4–11]

Traditional Medicinal Uses: The plant is used as a remedy for diarrhoea, malaria, diabetes, astringent, diaphoretic, bechic, emmenagogue, menstrual pain, hypertension, insomnia and depurative after parturition in Indochina, the Philippines, Jamaica, West Indies, South Africa, Southeast Asia, India and Queensland.[2,5–7,12,13] The plant is also used for cold, cough, fever and bronchitis.[14] In Malaysia, the crushed leaves are applied to scalds, burns, sores, mumps, swollen neck, tonsillitis and insect bites.[13] In Puerto Rico and Cuba, the flowers are decocted and used as an eyewash.[12]

Pharmacological Activities: Anticancer/Antineoplastic,[5,12,15–19] Antioxidant,[20] Antiangiogenic,[21] Chemopreventive,[22] Hypoglycaemic[23–28] and Wound healing.[29,30]

Dosage: Leaves are useful in treating oliguria, haematuria, diabetes mellitus, and menstrual disorders in a daily dose of 4 to 8 g as a decoction or liquid extract.[7] A decoction of 30 g of the plant is taken for diabetes, dysentery, enteritis, menstrual pains, hypertension, insomnia and cancer in Malaysia.[13] For the treatment of cold and sore throat, tea is made from 9 pink flowers in 1 pint of water and sipped throughout the day.[31] Dilute infusions of roots are used to treat diabetes.[32]

Adverse Reactions: *Catharanthus roseus* pollen can trigger IgE-mediated respiratory allergy in the people living in close proximity.[33]

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

[**Authors' Note:** Clinically, vinblastine and vincristine are administered by intravenous injection as solutions of their sulphate salts.[34]]

20. *Celosia argentea* L. (Amaranthaceae)

Feather Cockscomb, Red Spinach

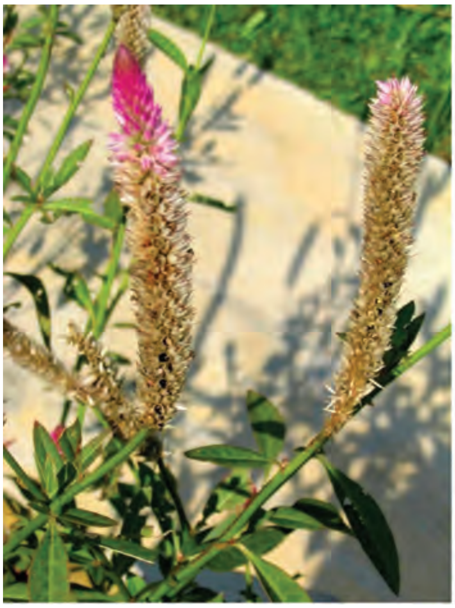

Flowers of *Celosia argentea*

Celosia argentea plant

Description: *Celosia argentea* L. is an annual tropical herb, up to 1 m tall. Stems are cylindrical and the aerial part is branched. Leaves are simple, small, spirally arranged, about 5–8 cm by 1–3 cm, alternate and exstipulate. The blade is lanceolate and ovate. The apex is acuminate. It bears several pinkish or white flowers which are minute. The fruits are globose and seeds are black.[1–5]

Origin: Native to India.[6]

Phytoconstituents: Celosian, nicotinic acid, celogenamide A, celogentin A–D, H, J and K, moroidin and others.[3,5,7–11]

Traditional Medicinal Uses: *C. argentea* is used internally for haematological and gynaecologic disorders and externally to treat inflammation and as a disinfectant. The whole plant is used to treat dysentery and dysuria, and used externally as poultices for broken bones.[5] The plant is used for eye and liver ailments in Yunnan, China[12] and also for the treatment of mouth sores and blood diseases and used as an aphrodisiac.[13] The petioles are used to treat sores, wounds, boils and swellings.[4,12] The seeds are used for the treatment of conjunctivitis and hypertension. In China, the seeds are used for haemorrhage, menorrhagia and opthalmia.[5,12,14] In Indonesia, the flowering tops are used for bleeding lungs whereas in Malaysia, the red flowering tops are prepared as decoctions which are given in cases of white discharges, excessive menstruations, haematuria, dysentery, proteinuria, bleeding piles and bleeding nose.[5]

Pharmacological Activities: Antibacterial,[15] Antimitotic,[16] Antineoplastic,[17] Diuretic,[5] Hypoglycaemic,[18] Hepatoprotective,[11,19] Immunomodulatory,[5,17,20] Cytoprotective[21] and Wound healing.[22]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: Leaves should not be eaten by menstruating women.[23]

Drug-Herb Interactions: No information as yet.

21. *Centella asiatica* (L.) Urban (Umbelliferae)

Indian Pennywort, Asiatic Pennywort

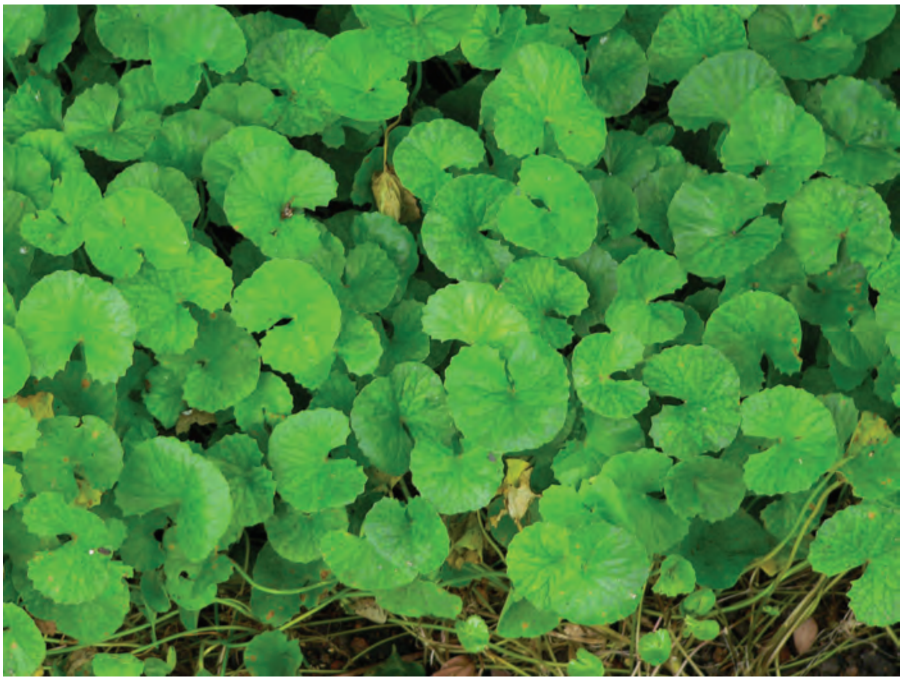

Top view of *Centella asiatica* herb

Description: *Centella asiatica* (L.) Urban is a creeping, perennial herb with long slender horizontal stolons, characterised by long internodes. Leaves are green, fan-shaped or round-reniform, 1–4 cm by 1–7 cm with a crenate or dentate margin. Flowers are umbels with white or light purple-to-pink petals and bear small oval fruit.[1,2]

Origin: Native to India, found in tropical America, Africa, West Pakistan, China, Japan and the Pacific.[3]

Phytoconstituents: Asiaticoside, asiatic acid, brahmic acid, brahmoside, centellic acid, centellose, indocentelloside, madecassic acid, madecassoside, thankuniside, vellarin, bayogenin, centellin, asiaticin, and centellicin.[4–15]

Traditional Medicinal Uses: The plant is used in cooling drinks when boiled, for diarrhoea, diuretic, gravel, leprosy treatment, stones, wound healing and as a tonic.[2,4] It is also part of a mixture to treat colic in Indonesia.[16] The Chinese uses it to improve appetite, aid digestion, treat sores and ulcers. In India, it is used to treat skin, nervous system and blood diseases.[2,7,17] The plant is used to treat fever and rheumatism, as a detoxicant, sedative and peripheral vasodilator.[7,18] The plant is reported as treatment for cancer, circulatory stimulant, hypotensive, stimulant, tonic, cicatrizant (leaf), used for the treatment of haematuria, gonorrhoea, peptic ulcer and sore throat.[5] It is also used for tuberculosis, headache (decoction), dysentery (decoction with other ingredients), boils and tumours (paste applied), leucoderma, anaemia, urinary discharges, bronchitis, insanity, leprosy (decoction), mental deficiency, dysentery (juice or paste on empty stomach for 2–3 days), cough (decoction with ginger and black pepper), cooling (paste with pepper and salt), tonic (juice with palm jaggery given to women after childbirth), elephantiasis, enlarged spleen (ointment from leaves), cures stuttering or stammering, diuretic, small pox and as a local stimulant.[19] External application in the form of poultices is prescribed for contusions, closed fractures, sprain and furunculosis.[18]

Pharmacological Activities: Antibacterial,[20] Antidepressant,[21] Antiemetic,[22] Antineoplastic,[8,23–26] Antioxidant,[27–32] Antithrombotic,[33] Anxiolytic,[34,35] Gastroprotective,[9,36] Immunomodulatory,[37,38] Antigenotoxic,[39] Nerve-regenerative,[40] Radioprotective[41,42] and Wound healing.[43,44]

Dosage: Approximately 0.6 g dry weight of whole plant taken three times a day (condition not indicated).[45] Larger amounts of fresh leaves are sometimes eaten as vegetable and dietary supplement.[45] 60 mg of *C. asiatica* extract given daily for varicose veins.[46] 30 or 60 mg of Total Triterpenic fraction of *Centella asiatica* (TTFCA) three times a day for improving venous hypertension.[47] Titrated extract of *C. asiatica* given 60 or 120 mg daily for chronic venous insufficiency.[48] For the treatment of fever, measles, haematemesis, epistaxis, diarrhoea, dysentery, constipation, leucorrhoea, jaundice, dysuria, furunculosis, dysmenorrhoea, varices, daily dose of 30 to 40 g of fresh plant in the form of extracted juice or decoction is taken.[18]

Adverse Reactions: Allergic contact dermatitis.[49] Sedation, increased blood glucose and cholesterol levels.[50]

Toxicity: Hepatotoxic.[51]

Contraindications: Should not be used during pregnancy and lactation and not to be given to children.[50]

Drug-Herb Interactions: Reported with ephedrine, theophylline, atropine and codeine.[52] Decrease effectiveness of antidiabetic and antilipidemic drugs.[50]

[**Authors' Note:** Topical creams containing the active component asiaticoside is available commercially for wound healing.]

22. *Cerbera odollam* Gaertn. (Apocynaceae)

Pong Pong Tree, Indian Suicide Tree, Sea Mango

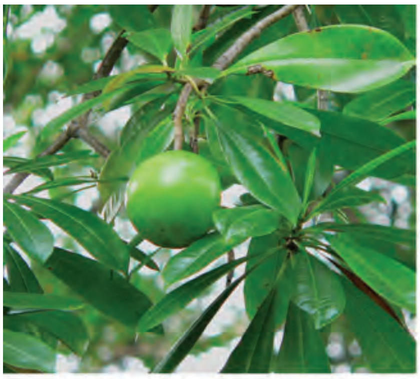

Fruit of *Cerbera odollam*

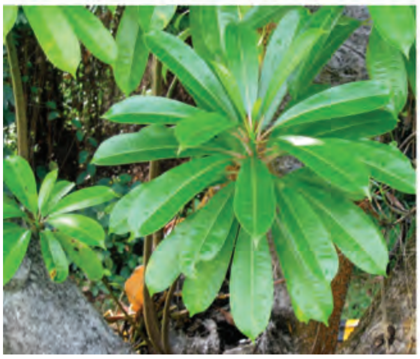

Cerbera odollam leaves

Cerbera odollam tree

Description: *Cerbera odollam* Gaertn. is a medium-sized tree with smooth and grey bark. Leaves are simple, few, without stipules, 12–16 cm by 3–5 cm, arranged in a spiral. The blade is succulent, dark green, glossy and lanceolate. Flowers are white, large and bisexual with a yellow eye in the throat of the corolla tube and arranged in terminal. Fruits are round, waxy surfaced, large, 5–10 cm across, with fibrous husk covering the single seed. The fruit turns from green to reddish brown upon maturity.[1–4]

Origin: Native to Indian subcontinent, Indochina, Malesia and the Pacific.[5]

Phytoconstituents: Cerberin, cerleaside A, 17α-neriifolin, 17β-neriifolin, thevetin B, acetyl-thevetin B, diacetylneriifolin, cerberoside, odollin and others.[3,6–12]

Traditional Medicinal Uses: In Malaysia, rheumatism is treated with embrocations of the fruits.[11] The seeds are poisonous and have been used to poison rats and dogs.[3,11] The seeds are also narcotic. In Indonesia, oil obtained from the seeds is rubbed on the body as remedy for colds, scabies, and rheumatism. In the Philippines, oil of the seeds is also used to treat rheumatism.[11] It is used in Burma as an insecticide or insect repellent when mixed with other oils.[13] The bark, latex and roots are used as purgatives and emetics in India.[3,11,13]

Pharmacological Activities: Antineoplastic[8,10] and Antipsychotic.[14]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: Seeds are poisonous and have been used to poison rats and dogs.[3,11] The seeds have a long history as an ordeal poison in Madagascar, due to the highly toxic cardiac glycosides they contain. The kernel contains cerberin, cerberoside and odollin which are toxins.[15] Humans poisoned by oral consumption of half to one seed kernel would result in sinus bradycardia, wandering pacemaker and second-degree sino-atrial block and nodal rhythm.[15,16] Other symptoms include nausea, retching and vomiting.[16] The poisoning may result in hyperkalemia, which can cause death.[17]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

23. *Cissus quadrangularis* L. (Vitaceae)

Grape Leaf, Veld Grape

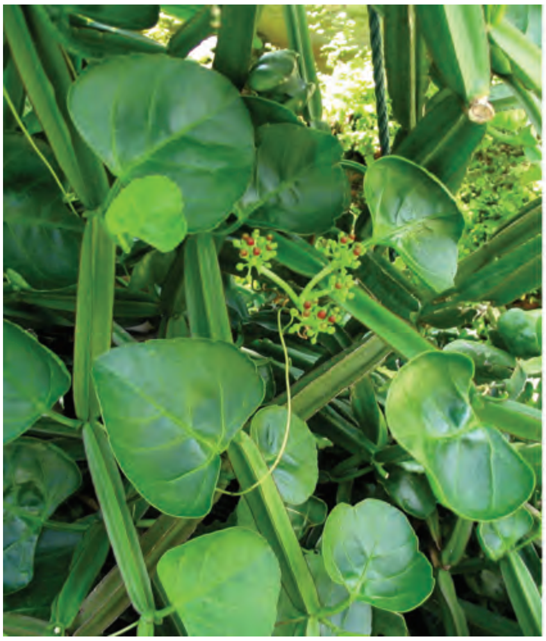

Flowers of *Cissus quadrangularis*

Cissus quadrangularis plant

Description: *Cissus quadrangularis* L. is a herbaceous plant with a thick, quadrangular, succulent stem and is constricted at regular intervals. It has long, slender and simple tendrils. Leaves are ovate, entire, crenate-serrate, 3–5 cm by 5–13 cm and glabrous on both sides. Flowers are pink and white, 2 mm long, while the berries contain one or two seeds.[1,2]

Origin: India, Pakistan, Indonesia, Malaysia, East Africa, Sri Lanka and Arabia.[2]

Phytoconstituents: Quadrangularins A–C, δ-amyrin, δ-amyrone, resveratrol, piceatannol, pallidol, parthenocissine A and others.[3–7]

Traditional Medicinal Uses: Whole plant is used for urinary *schistosomiasis* in Mali.[8]

Pharmacological Activities: Antibacterial and Antioxidant,[9] Analgesic and Anti-inflammatory,[10] Antimalarial,[11] Gastroprotective[12–16] and Antiosteoporotic.[17]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

24. *Cocos nucifera* L. (Palmae)

Coconut Palm, Kelapa

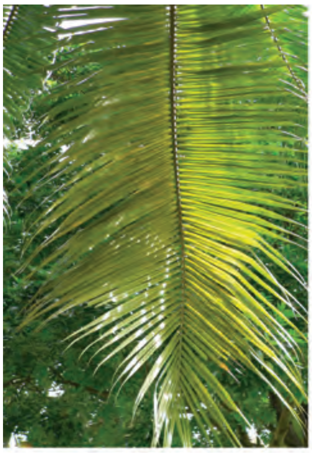

Cocos nucifera leaves

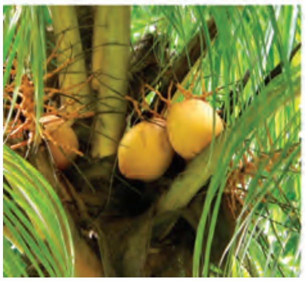

Fruits of *Cocos nucifera*

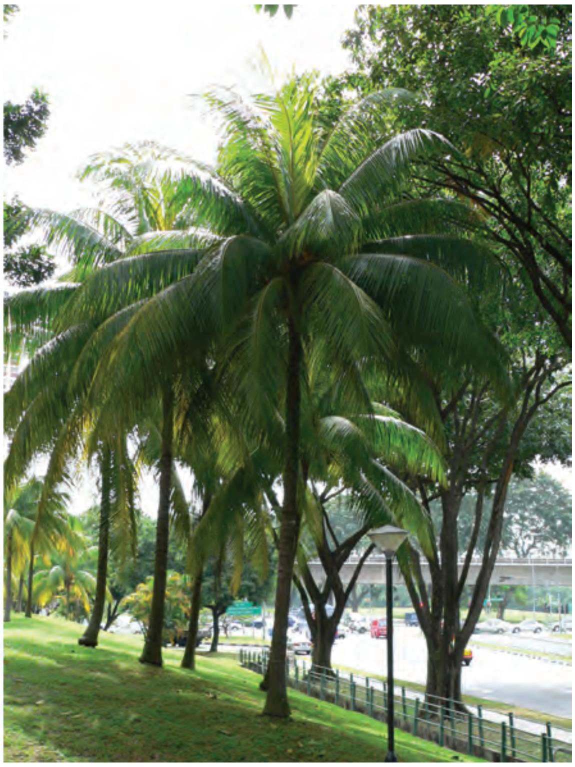

Cocos nucifera trees

Description: *Cocos nucifera* L. is a tall palm with a ringed stem that can grow up to 30 m tall. Leaves are pinnate, 2–6 m long with numerous pairs of narrow leaflets. Flowers are arranged in large panicles among the leaves with female flowers near the base of the inflorescence. The fruit is symmetrical, ovoid and about 20–30 cm across. Thick fibrous husk encloses the hard shell (endocarp) in which the fleshy pericarp adheres.[1,2]

Origin: Native to the Pacific.[3]

Phytoconstituents: Trans-zeatin, dihydrozeatin, dihydrozeatin-*O*-glucoside, *meta*-topolin riboside, N^6-isopentenyladenine, N^6-benzylaminopurine and others.[4,5]

Traditional Medicinal Uses: Its fruit juice is used to treat poisoning, cholera and is a diuretic. Fresh coconut juice with rice flour is poulticed onto carbuncles, gangrenous sores and indolent ulcers.[6] Coconut water is also used for fever, urinary complaints and to stop vomiting.[7] Juice of the green, unripe coconut is boiled and drunk to relieve diarrhoea. In Somoa, coconut oil is used as a laxative and to relieve stomach ailments.[8] The decoction of husk fibre has been used in northeastern Brazil traditional medicine for the treatment of diarrhoea and arthritis.[9] Coconut water has also been used as short-term intravenous hydration and resuscitation fluid in emergencies.[10] The rootbark is used as an astringent, styptic and in haemorrhages.[6] It is also used for uterine diseases, bronchitis, liver complaints and dysentery.[7]

Pharmacological Activities: Analgesic,[11] Antibacterial,[9,12] Antifungal,[13] Antineoplastic,[14,15] Antioxidant,[11,16] Antiprotozoal,[17] Antiviral,[9] Hypoglycaemic,[18] Hypolipidaemic,[19,20] Hypotensive,[21] Immunomodulatory[22] and Antitrichomonal.[23]

Dosage: No information as yet.

Adverse Reactions: A case report of occupational allergic conjunctivitis due to coconut fibre dust has been reported.[24]

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

25. *Coix lacryma-jobi* L. (Gramineae)

Job's Tears, Adlay

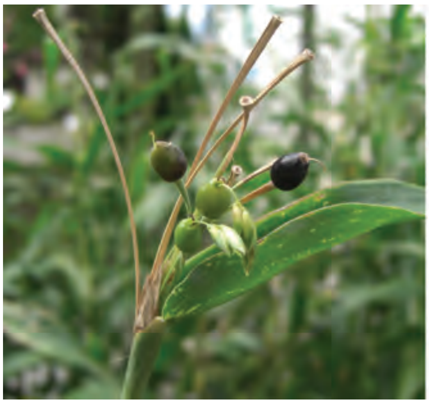

Fruits of *Coix lacryma-jobi*

Coix lacryma-jobi plants

Description: *Coix lacryma-jobi* L. is a small herbaceous plant which grows up to 2.5 m. Leaf sheaths are glabrous and the leaf blades are narrowly lanceolate, 20–50 cm by 1.5–4 cm. The midrib is prominent. Fruit is tear-shaped, 8 mm by 1.1 cm, with glossy berries which turn black upon maturity.[1–3]

Origin: Native to tropical and temperate Asia.[4]

Phytoconstituents: Coixol, coixenolide, α-coixins, γ-coixins, syringaresinol, mayuenolide, coixan A–C, coixic acid and others.[5–9]

Traditional Medicinal Uses: A decoction is believed to benefit the blood and breath and used to wash newborns to prevent diseases.[10] The kernels are used to treat lung and chest complaints, rheumatism, dropsy and gonorrhoea.[3] Fruit is used for intestinal or lung cancers and warts. The fruit is also used as a vermifuge and for hypertension. As the seed is diuretic and refrigerant, a decoction is used for appendicitis, arthritis, beriberi, bronchitis, cancer, diarrhoea, dryskin, dysuria, oedema, hydrothorax, inflammation, pleurisy, pneumonia, pulmonary abscesses, rheumatism and tuberculosis.[10,11] Seeds are also used for the treatment of enteritis, persistent diarrhoea in children, urinary lithiasis, rheumatism and acrodynia.[12] An infusion of the seeds is prescribed for bronchitis, pulmonary abscess, pleurisy and hydrothorax.[11] The root is given along with roots of long pepper and other herbs for fever with drying of saliva and intense thirst, for dysentery, diarrhoea and puerperal fever.[13]

Pharmacological Activities: Anti-inflammatory,[14,15] Anticancer/Antineoplastic,[16–21] Antioxidant,[8,18] Hypoglycaemic,[9,22] Hypolipidaemic[22–24] and Hypotensive.[25]

Dosage: For the treatment of enteritis, persistent diarrhoea in children, oedema, urinary lithiasis, rheumatism and acrodynia, doses of 10 to 30 g daily in the form of powder or decoction are given.[12]

Adverse Reactions: No information as yet.

Toxicity: Embryotoxicity in pregnant rats was observed. Oral administration of 1 g/kg body weight of water extracts caused an increase in foetal resorptions and postimplantation mortality.[26]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

26. *Crinum asiaticum* L. (Amaryllidaceae)

Crinum Lily, Spider Lily, Bawang Tanah

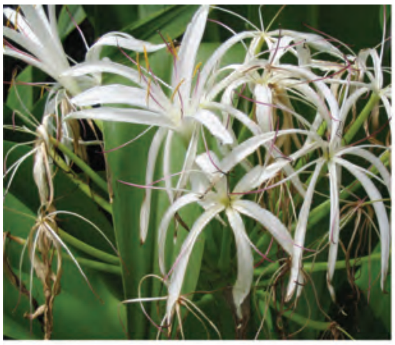

Crinum asiaticum flowers

Leaf blade of *Crinum asiaticum*

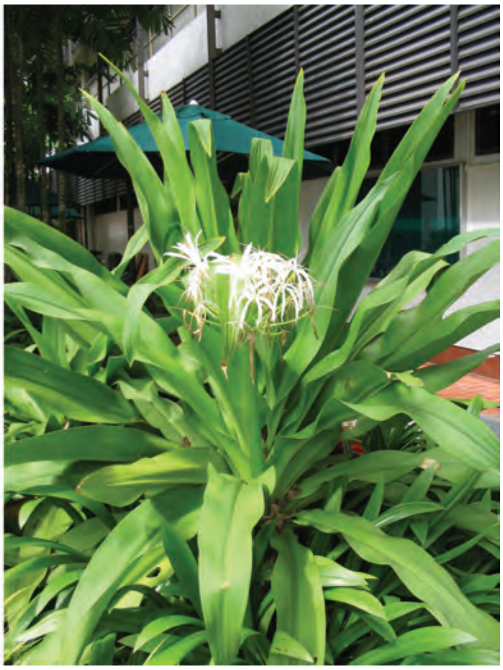

Crinum asiaticum plant

Description: *Crinum asiaticum* L. is a bulbous herb with a flowering stalk in the centre of the plant. The leaves are narrowly lanceolate, acuminate, 0.5–1.5 m long, greenish and have a hairy texture. The flowering stem is about 1–1.2 m. Flowers are white, 6–12 in an umbel. Filaments are reddish. The fruit is irregularly globose, 4–5 cm across, with one or a few large green seeds.[1–4]

Origin: Native to Tropical Asia, introduced in Northern America.[5]

Phytoconstituents: Crinamine, lycoricidine, hamayne, isocraugsodine, palmilycorine, lycoriside, ambelline, crinasiatin, hippadine, bakonine, pratorimine, crinine, powelline, ungeremine, criasbetaine, crinasiatine, phenanthridone II and others.[6–17]

Traditional Medicinal Uses: In Southeast Asian countries, *C. asiaticum* has a considerable medicinal reputation as a potent folk medicine in the treatment of injury and inflamed joints.[18] The plant has been used for carbuncles and cancer.[7] In Indonesia, the oiled and heated leaves are useful to treat wounds by poisoned arrows, bites and stings. In Malaysia, poultices of the leaves are applied to swellings, swollen joints, lumbago, pains and in cases of headache and fever. The leaves are also an emollient. In Northwest Solomon Islands, the leaves make a topical treatment for inflammation.[19] In Malaysia, the leaves are used as a rheumatic remedy and to relieve local pain.[18] On Karkar Island and in Simbu, Papua New Guinea, the latex from the leaves is applied to cuts. In India, the leaves are applied to skin diseases and inflammation.[20] The crushed leaves are used to wash piles or mixed with honey and applied to wounds and abscesses.[3] Its seeds are considered purgative and emmenagogic.[7] In the Trobriands, Papua New Guinea, the stem fibres are used to stop bleeding and in New Ireland, the milky sap from the stem is used for stonefish wounds.[20] The bulb is an emetic and counter-irritant. In Papua New Guinea, juice obtained from bulb is ingested regularly for 2 months to treat gonorrhoea. In the Philippines, the bulbs are crushed and applied as an ointment.[19] Juice from the fresh bulbs, taken several times per month induces vomiting. It is also instilled in the ear to treat otitis.[21] The root is also an emetic, diaphoretic and nauseant when fresh.[22] In a Finschhafen area village, Papua New Guinea, the cut root is cooked in a banana leaf, then cooled and placed on an aching tooth. Roots are used in New Caledonia, Indonesia, and Malaysia in a poultice for wounds, ulcers and swellings.[20]

Pharmacological Activities: Analgesic and Antibacterial,[19] Anti-inflammatory,[18] Antiviral,[18,23] Anticancer,[13,24] Antitumour,[15] Mast cells degranulation,[10] Antimitotic and Membrane stabilising.[8]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

27. *Cymbopogon citratus* (DC.) Stapf. (Gramineae)

Lemon grass

Cymbopogon citratus plants

Description: *Cymbopogon citratus* (DC.) Stapf. is a grass composed of dense leafy clumps that grows up to 1.8 m tall. Leaf sheath is tubular and acts as a pseudostem, 12–25 cm long, thickening towards the base and 1–2 cm in diameter. Leaf blades are sessile, simple, green, linear, glabrous, 60–90 cm by 1–2 cm, and possess parallel venation. Inflorescence is a raceme.[1–3]

Origin: Native to South Asia, Southeast Asia and Australia.[3]

Phytoconstituents: Citral, citronellal, cymbogonol, α-terpineol, citronellic acid, α-camphorene, geranial, isoorientin, isoscoparin and others.[1,4–8]

Traditional Medicinal Uses: The plant is used to treat digestive problems and relieve cramping pains.[4] A decoction of the plant is used by the Chinese to treat coughs, colds and blood in sputum.[2,9] The entire plant is used to treat bacterial infection and possesses fever-reducing and stomachic properties.[5] In many Asian countries, leaves in water provide a bath to reduce swelling, to remove body odour, improve blood circulation, treat cuts, wounds, bladder problems and leprosy.[9] Its oil is used as an insect repellent, for aerosols, deodorants, floor polishes and household detergents.[4,5] The oil is carminative for cholera, and is prescribed for dyspepsia, vomiting, fever, and headache,[10] and used externally to treat eczema.[5] Its roots are taken to induce sweating, increase flow of urine,[9] treat coryza and influenza.[5]

Pharmacological Activities: Analgesic,[11,12] Anthelmintic,[13] Antibacterial,[14–18] Antifungal,[16,19–25] Anticancer/Antineoplastic,[26,27] Antimalarial,[28] Antioxidant,[8,29,30] Antiplatelet,[31] Hepatoprotective,[32] Hypoglycaemic,[33] Sedative,[34] Vasorelaxant,[35] Antimutagenic,[36] Insecticidal[37] and Radioprotective.[38]

Dosage: A dose of 10 to 20 g of roots is used for treating coryza, influenza and fever. It is also prescribed for dyspepsia and vomiting and as a carminative, by using 3 to 4 drops of the essential oil diluted in water.[5]

Adverse Reactions: Application of thick ointments with the volatile oil on the skin has led to rare incidence of allergy.[39]

Toxicity: Alveolitis occurred as a result of inhalation of the volatile oil.[39] It is also reported to be cytotoxic and genotoxic.[40]

Contraindications: Should not be used during pregnancy.[41]

Drug-Herb interactions: No information as yet.

28. *Dolichos lablab* L. (Leguminosae)

Lablab Bean, Hyacinth Bean

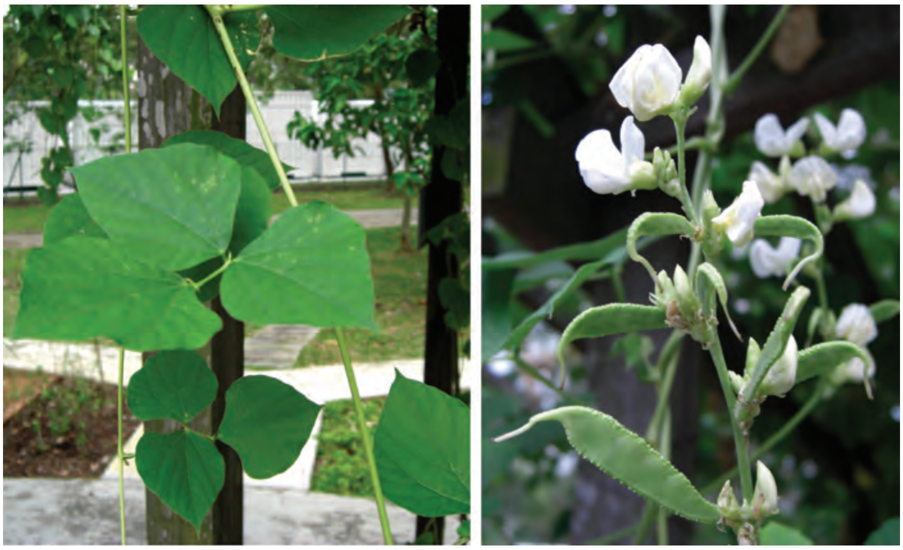
Dolichos lablab leaves *Dolichos lablab* flowers and beans

Description: *Dolichos lablab* L. is a woody climbing herb which can reach a length of 5 m. Leaves are pinnate and generally 3-foliolate. Leaflets are acute, entire, 6–12 cm by 5–9 cm. Flowers are white or purplish pink. Fruits are green pods, 6 cm long by 2 cm wide, flattened, contain 4–5 seeds and turn light brown when mature.[1–4]

Origin: Native to Africa.[5]

Phytoconstituents: Dolichin, arabinogalactan 1 & 2, lablabosides A–F, phytin, pantothenic acid, saponin I, putrescine, spermidine, spermine and others.[3,6–14]

Traditional Medicinal Uses: The plant is decocted for alcoholic intoxication, cholera, diarrhoea, globefish poisoning, gonorrhoea, leucorrhoea and nausea.[15] Its seeds are used to stimulate gastric activities, as antidote against poisoning, to treat colic, cholera, diarrhoea, rheumatism and sunstroke.[4,16] The juice from the fruit pods are used for inflamed ears and throats. The fruit is also astringent, digestive, stomachic and used to expel worms. The seeds are reportedly alexiteric, aphrodisiac, febrifuge, stomachic and used for menopause and spasms.[15] The flowers are used to treat dysentery, inflammation of uterus and to increase menstrual flow.[4,15] They are also used for leucorrhoea, menorrhagia, and summer heat disorders, as they have alexiteric and carminative properties.[15]

Pharmacological Activities: Antifungal,[6,17] Antiviral[6] and Haemagglutinating activities.[18–20]

Dosage: The reported dose of the seeds is 8 to 16 g daily in the form of powder or a decoction for the treatment of nausea, vomiting, diarrhoea, enteritis, abdominal pains and alcoholism.[21]

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

29. *Elephantopus scaber* L. (Compositae)

Elephant's Foot, Tutup Bumi, Tapak Sulaiman

Elephantopus scaber flowers

Elephantopus scaber herbs

Description: *Elephantopus scaber* L. is a tropical herb that can grow up to 30–60 cm tall. Leaves are simple, without stipules and when fully developed, form a rosette on the ground. Blades are obovate or oblong obtuse, hairy, large, 5–10 cm by 1.2–3 cm. The petioles are short, hairy, often crowded at the base of stem. Flowers are small and whitish pink.[1–3]

Origin: Native to Africa, tropical and temperate Asia and Australasia.[4]

Phytoconstituents: Molephantin, crepiside E, deoxyelephantopin, stigmasterol, stigmasteryl, scabertopin, lupeol and others.[5–9]

Traditional Medicinal Uses: In Vietnam, the plant is considered diuretic and administered at parturition.[10] A decoction of whole plant is bechic, used to treat pulmonary disease and scabies. In India, it is used as a tonic, laxative, analgesic, aphrodisiac and to treat inflammation. In the Philippines, the plant is febrifuge, diuretic and emollient.[10] In Malaysia, it is used as a preventive medicine after childbirth, to expel intestinal worms, for coughs and venereal diseases. In Chinese medicine, it is used to cure "dumpheat", which includes indigestion, swollen legs and loss of appetite.[1] In Indonesia, the plant is prescribed when there is a yellowish discharge from the vagina.[1] The plant is also considered as diuretic, emollient, used for relief of anuria and blennorrhoea, remedy for leucorrhoea and anaemia, to treat fever, inflammation and as a disinfectant.[6] The whole plant is decocted for abscesses, cold, dysentery, oedema, gastroenteritis, gonorrhoea, influenza, pharyngitis and snakebite.[11] In Burma, a decoction of the stem and leaves is used in cases of menstrual disorders.[10] The roots are used for fever in children, on pimples, wounds of cattle, as an abortifacient, for urinary complaints, amoebic dysentery and other digestive problems.[12]

Pharmacological Activities: Antibacterial,[13] Anticancer/Antineoplastic,[14,15] Anti-inflammatory,[16] Antiviral[17] and Hepatoprotective.[18,19]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: Molephantin isolated from plant demonstrated cytotoxic activity.[5]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

30. *Euphorbia hirta* L. (Euphorbiaceae)

Asthma Weed

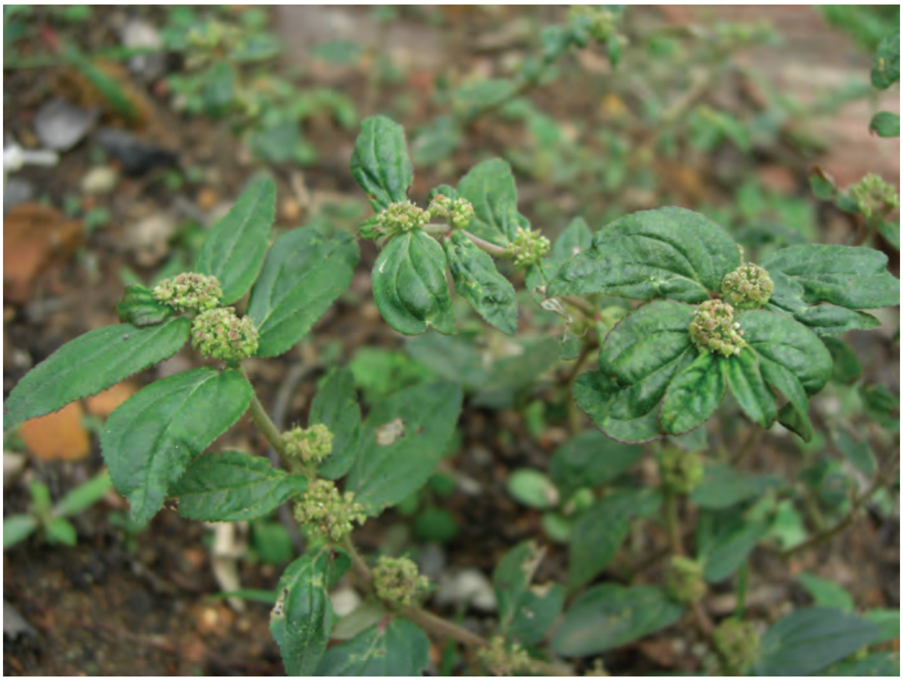

Euphorbia hirta herb

Description: *Euphorbia hirta* L. is an annual herb, which can grow up to 30 cm tall, and branched near the base. Leaves are elliptic-oblong, 2–3 cm by 0.8–1.5 cm. Inflorescence is small, in axillary dense clusters, each with one female flower and 4–5 male flowers inside.[1]

Origin: Native to the Pantropic.[2]

Phytoconstituents: Euphorbon, euphosterol, camphol, leucocyanidol, xanthorhamnin, taraxerol, taraxerone, myricitrin, euphorbianin and others.[3–6]

Traditional Medicinal Uses: The whole plant is decocted for athlete's foot, dysentery, enteritis, fever, gas, itch, and skin conditions.[3] It is also regarded as anodyne, depurative, diuretic, lactogogue, purgative, and vermifuge. The plant is used for asthma, bronchitis, calculus, colic, cough, dyspnoea, eruptions, excrescences, influenza, fractures, gonorrhoea, headache, hypertension, measles, nausea, ophthalmia, sores, splinters, stomachache, tumours, urogenital ailments, warts and wounds.[3,7] In Central Province of Papua New Guinea, the plant is boiled and the solution is taken by patients who pass blood in the urine. The Chinese use the plant to treat fever, dysentery and skin conditions. In the Philippines and Indonesia, the plant is used to treat bowel problems.[8] The latex is used on warts and abscesses.[7]

Pharmacological Activities: Analgesic,[9] Antibacterial,[10–14] Antidiarrhoeal,[6,15] Anti-inflammatory,[9,16,17] Antiplatelet,[18] Antiprotozoal,[19,20] Antipyretic,[9] Anxiolytic,[21] Diuretic,[22] Sedative,[23] Antianaphylactic[24] and Molluscicidal.[25,26]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: Toxic to brine shrimp.[11]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

31. *Eurycoma longifolia* Jack (Simaroubaceae)

Tongkat Ali, Ali's Umbrella, Pasak Bumi

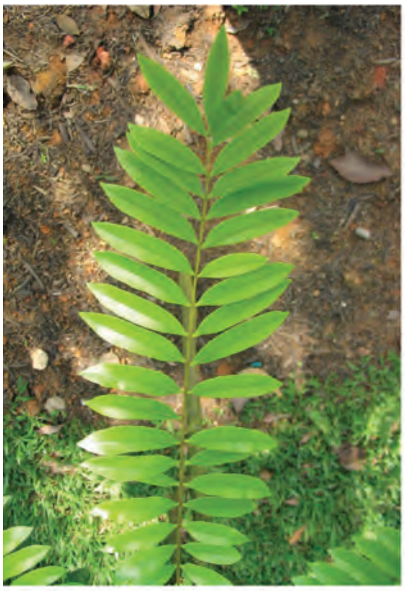

Eurycoma longifolia leaves

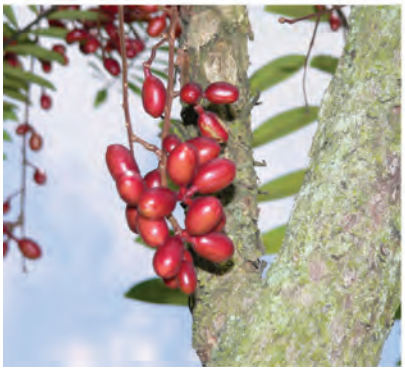

Eurycoma longifolia fruits

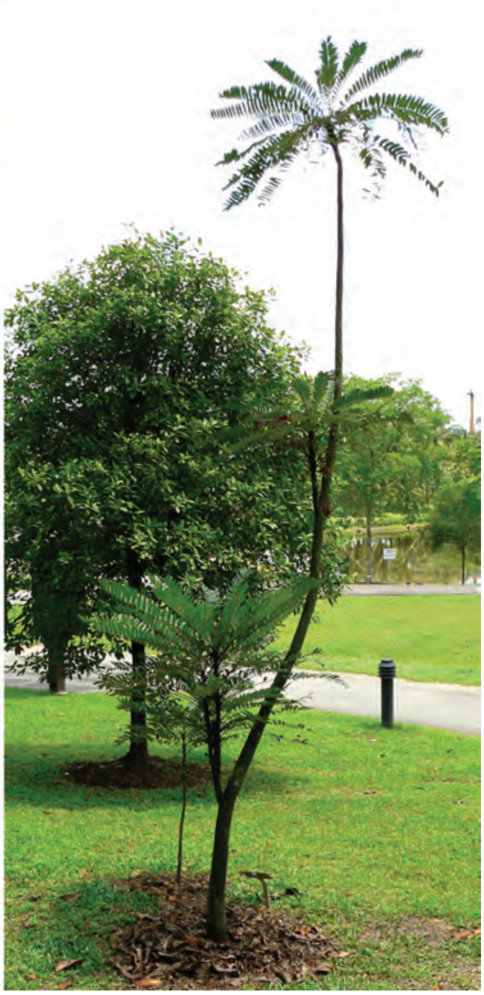

Eurycoma longifolia tree

Description: *Eurycoma longifolia* Jack is a small tree with compound leaves on branches that can grow up to 1 m long. The numerous leaflets are opposite or subopposite, lanceolate to ovate-lanceolate, 5–20 cm by 1.5–6 cm, with smooth margins. Flowers are tiny, reddish, unisexual and are densely arranged. The drupes are ovoid with a distinct ridge, 1–2 cm by 0.5–1.2 cm and they turn dark reddish brown when ripe.[1–3]

Origin: Native to Malesia and Indochina.[4]

Phytoconstituents: Eurycomalactone, eurycomanol, eurycomanone, eurylactone, eurylene, laurycolactone A and B, longilactone, pasakbumins A to D, eurycomalide A and B, piscidinol A and others.[3,5–19]

Traditional Medicinal Uses: The plant is used to cure indigestion and lumbago. It is used as a tonic after childbirth, to relieve pains in the bone and for treatment of jaundice, dropsy, cachexia and fever.[3,20] Tongkat Ali is one of the most well known folk medicines for intermittent fever (malaria) in Southeast Asia.[16] Decoction of the leaves is used for washing itches, while the fruits are used in curing dysentery.[3] Its bark is used as a vermifuge.[3] The taproots are used to lower high blood pressure, while the root bark is used for the treatment of fever and diarrhoea.[18] The roots of this plant are used as folk medicine for the treatment of sexual insufficiency, aches, persistent fever, malaria, dysentery, glandular swelling and also as health supplements.[18]

Pharmacological Activities: Antianxiety,[21] Antibacterial,[22] Anticancer,[6,8,17,18,24–26] Antitumour,[23,27] Antimalarial/Antiplasmodial,[5,13,18,23,28–32] Antischistosomal,[23] Antiulcer,[33] Aphrodisiac[15,18,34–38] and Plant growth inhibitor.[39]

Dosage: 1 g daily is recommended to be the maximum dose for supplemental use.[39]

Adverse Reactions: No information as yet.

Toxicity: One animal study found that the LD_{50} in mice was 1500–2000 mg/kg of the alcohol extract and 3000 mg/kg of the water extract. A subacute toxicity study with the alcohol extract indicated that 600 mg/kg daily was associated with signs of toxicity while 200 mg/kg daily was not, and another study found no toxic effects at 270–350 mg/kg daily but toxic effects were observed at 430 mg/kg daily.[39] Eurycomanone was identified as the most toxic component from its butanol extract.[40]

Contradindications: Should be used with caution and preferably not for extended periods without taking periodic breaks from use when it is used as a supplement. Tongkat Ali should not be taken by methods other than oral administration.[39]

Drug-Herb Interactions: No information as yet.

32. *Hibiscus mutabilis* L. (Malvaceae)

Cotton Rose, Chinese Rose

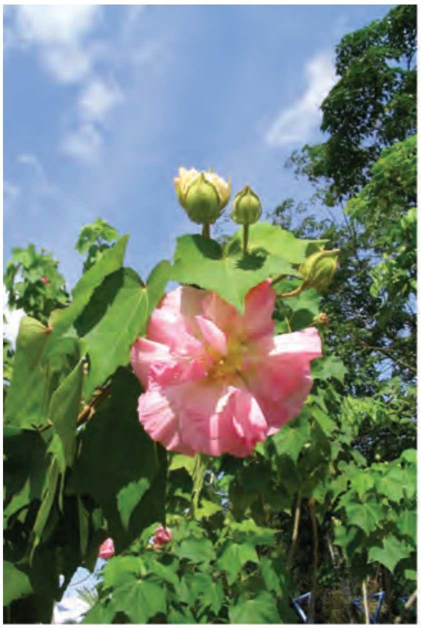

Hibiscus mutabilis flower

Hibiscus mutabilis tree

Description: *Hibiscus mutabilis* L. is a small tree that can grow up to 5 m tall. Leaf blades are heart-shaped, broadly ovate to round-ovate or cordate, 5–7-lobed, 10–15 cm in diameter, and papery. Abaxially, they are densely stellate and minutely tomentose; adaxially they are, sparsely stellate and minutely hairytoothed, 8–15 cm wide. Flowers are solitary and with multi-petals, white colour in the morning, changing to pink in the afternoon.[1–3]

Origin: Native to China.[4]

Phytoconstituents: Isoquercitrin, hyperoside, rutin, quercetin, naringenin, tetracosanoic acid, daucosterol, salicylic acid, quercimeritrin, meratrin and others.[5–10]

Traditional Medicinal Uses: The plant is used for fistulae, pustules and tumours.[11] The leaves and flowers are used as an analgesic, to expel phlegm, treat excessive bleeding during menstruation, painful urination, inflammation and snake bites.[3] The leaves and flowers are also used as demulcent, diuretic and treat bacterial infection. They are used to treat boils, particularly on the chin, in the form of a poultice made of powdered dried leaves and flowers mixed with concentrated tea infusion which makes the boils burst earlier and less painfully. They are also used for treating impetigo, prurigo, metritis, leucorrhoea, mastitis, nephritis, cystitis, dysuria and infections.[12] The leaf is applied to swellings, crushed and compressed and applied onto abscesses, burns, and ulcers. It is also used as anodyne, alexipharmic, demulcent, expectorant, and refrigerant.[11] The flowers are used for lung ailments, with leaves for burns, inflammation, and snake bite. They are also prescribed for cough, dysuria and menorrhagia.[11]

Pharmacological Activities: Anti-inflammatory.[13]

Dosage: For the treatment of impetigo, prurigo, metritis, leucorrhoea, mastitis, nephritis, cystitis, dysuria and infections, a dose of 5 to 20 g of leaves and flowers is taken daily in the form of a decoction.[12]

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

33. *Hibiscus rosa-sinensis* L. (Malvaceae)

Hawaiian Hibiscus, China Rose, Bunga Raya

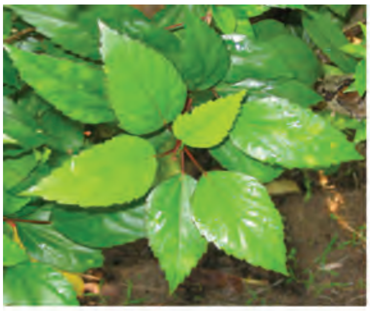

Leaves of *Hibiscus rosa-sinensis*

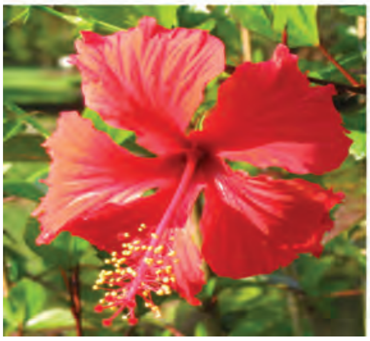

Hibiscus rosa-sinensis flower

Hibiscus rosa-sinensis shrub

Description: *Hibiscus rosa-sinensis* L. is a small evergreen perennial tree that can grow up to 3.6 m tall. It produces flower all year round. The large glossy leaves are ovate, alternate and vary in colour from pale green to dark green, with serrated edges. The flowers have five petals, and their colour varies but are mostly red.[1,2]

Origin: Native to South Eastern Asia and cultivated throughout the world as decorative plants.[3]

Phytoconstituents: Gossypetin, anthocyanin, myristic acid, palmitic acid, ambrettolide, campesterol, methyl sterculate, malvalate and others.[4–6]

Traditional Medicinal Uses: In Suquang, Papua New Guinea, the whole plant juice is applied directly to sores, to treat headaches and irregular periods.[7] In Central Province, the plant is used to treat eye sores whereas in Northern Province and North Solomons Province, it is used to induce labour. In Indonesia, it is used as a purgative, an abortifacient and to regulate menstruation.[7] The juice also provides a soothing effect on mucous membrane that line the respiratory and digestive tracts.[5] The flowers and leaves are used to treat skin diseases, mumps, to relieve fever,[2,4] as well as to be used as emollient, anodyne and laxative.[8] The flower is also used as an astringent,[5] for excessive menstruation, fever and skin diseases. Its roots are used to treat gonorrhoea.[8] In Finschhafen, the roots and leaves are crushed and the juice is drunk to treat diarrhoea.[7]

Pharmacological Activities: Antianxiety,[9] Anticancer,[10] Anticonvulsant,[9] Antifertility,[11–14] Antioxidant,[15] Hepatoprotective,[16] Hypoglycaemic,[17–19] Hypolipidaemic[18,20] and Wound healing.[21]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: Not to be taken by small children, and during pregnancy and lactation.[22]

Drug-Herb Interactions: No information as yet.

34. ***Hibiscus tiliaceus*** L. (Malvaceae)

Linden Hibiscus, Sea Hibiscus, Mahoe

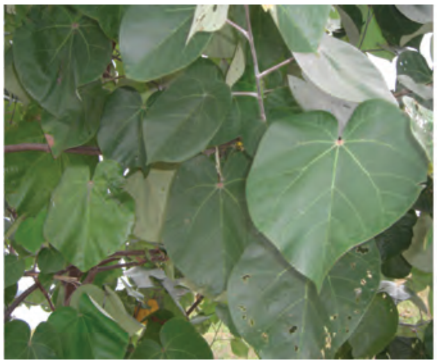

Leaves of *Hibiscus tiliaceus*

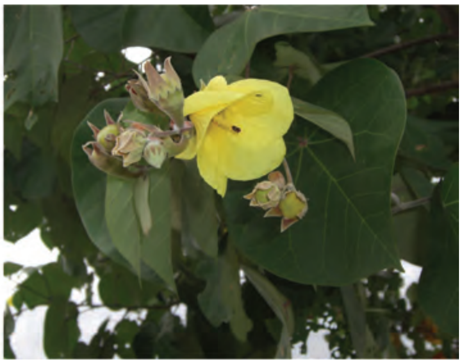

Hibiscus tiliaceus flowers

Description: *Hibiscus tiliaceus* L. is an evergreen coastal tree that can grow up to 10 m tall. Leaves are heart-shaped, 10–15 cm long and wide, leathery and green. The large showy flowers are yellow with maroon centre. Fruit is capsular, subglobose to ovoid, about 2–3 cm long and wide, surrounded by the calyx. Seeds are reniform, smooth and glabrous.[1–4]

Origin: Native to the Pantropic.[4]

Phytoconstituents: Hibiscones A–D, hibiscusin, hibiscusamide, friedelin, epifriedelanol, pachysandiol A and others.[2,5–7]

Traditional Medicinal Uses: In Malaysia and Indonesia, the leaves are considered cooling and are used to treat fever. They are also used as a soothing agent and to remove phlegm from the respiratory passages.[2] The flowers are widely used as a means of birth control in traditional medicine in Asian and African countries.[8] In the Philippines, the mucilaginous water obtained by soaking the fresh bark in water is prescribed for dysentery[2] while the root is an ingredient of embrocations and is used as a febrifuge.[8]

Pharmacological Activities: Tyrosinase inhibitory,[9] Antioxidant[10,11] and Cytotoxic.[5]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

35. *Impatiens balsamina* L. (Balsaminaceae)

Balsam Plant

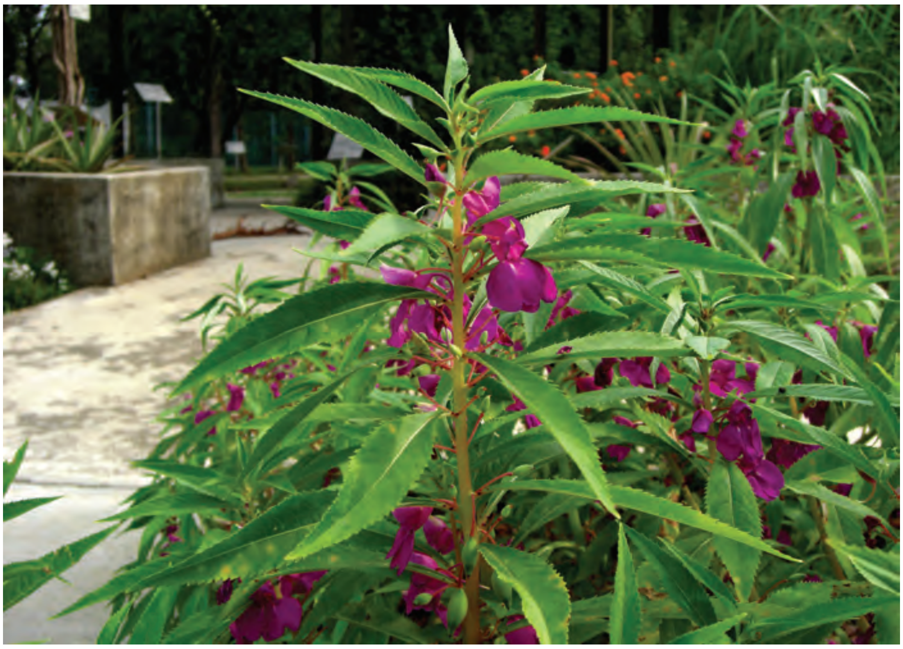

Impatiens balsamina plants

Description: *Impatiens balsamina* L. is an annual, erect, succulent plant, which is about 30–90 cm tall. Leaves are simple, alternate, 15 cm long, serrated and arranged spirally. Flowers are white, pink, purple or variegated and usually in leaf axils. Fruits are capsular, tomentose and contain several seeds.[1–5]

Origin: Native to India.[5]

Phytoconstituents: Impatienol, pelargonidine, delphinidine, cyanidine, balsaminones A and B, hosenkosides F–O and others.[4,6–10]

Traditional Medicinal Uses: A lotion of fresh leaves is used to treat eczema, itches and insect bites.[4] In Vietnam, decoctions of leaves are used to stimulate growth and to wash hair.[4] The juice is also used for warts, cancer treatment and expectorant.[11] A decoction of flowers is taken for infections, vomiting, urine retention and as a tonic. In India, flowers are regarded as cooling, tonic and useful when applied to burns and scalds.[4] The flowers are also used for lumbago and intercostal neuralgia, snakebite, improves circulation and relieves stasis.[11] In Japan, juice squeezed from the white flower petals are applied on the skin to alleviate dermatitis.[6] In China, the seeds are prescribed for difficult labour, puerperal pains, difficult menstruation, cough, hiccups and poisonings. The seeds are mixed with arsenious acid for removing teeth.[4] In Malaysia, the seeds are taken for gastrointestinal tract cancer, and to dislodge fish or chicken bones in throat.[4]

Pharmacological Activities: Antibacterial,[12–15] Antihypotensive,[16,17] Antifungal,[18] Anti-inflammatory[4–8] and Antipruritic.[6,19]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

36. *Imperata cylindrica* (L.) P. Beauv. (Gramineae)

Lalang, Alang-alang, Speargrass

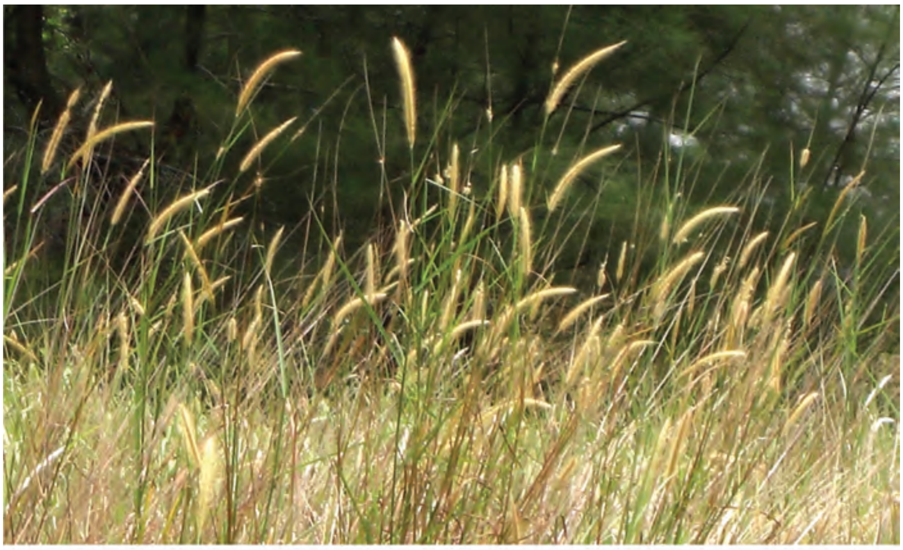

Imperata cylindrica

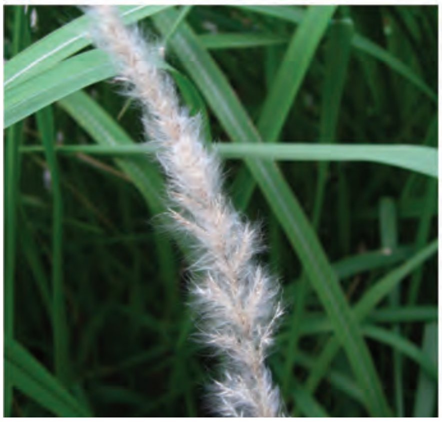

Inflorescences of *Imperata cylindrica*

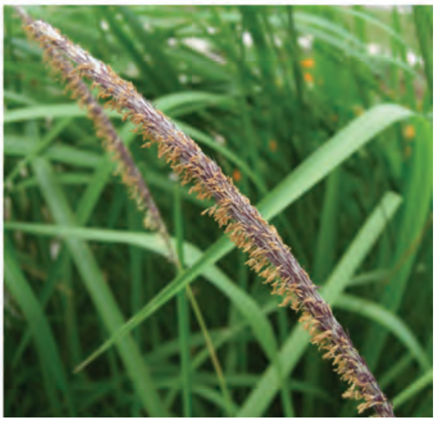

Imperata cylindrica fruits

Description: *Imperata cylindrica* (L.) P. Beauv. is a perennial erect grass about 0.6–1.5 m tall. Leaf blades are linear to lanceolate, long, 0.5–1 cm by 15–30 cm, narrow, with sharp margins and prominent nerves. Inflorescences are white and fluffy. Rhizome is hard, coriaceous and deep within the soil.[1]

Origin: Native to China and Africa but this widespread weed can be found throughout the world.[2]

Phytoconstituents: Arborinione, arundoin, anemonin, isoorientin, imperanene, cylindol A and B, graminones A and B, cylindrene, isoarborinol, impecyloside and others.[3–10]

Traditional Medicinal Uses: Rotted grass from thatch boiled with wine is used to treat bug bite, haemoptysis, severe constipation and vaginismus. It is also used for drug withdrawal symptoms.[3] Besides being used to quench thirst, its flowers and rhizome are also used to treat blood in the sputum, nose bleeds, lung and kidney diseases, jaundice,[11] haemorrhage, wounds, haemoptysis, epistaxis, haematemesis, haematuria, nephritic oedema, high fever, and urinary tract infections.[3] The roots are used to treat fever, cough with phlegm,[11] asthma, cancer, dropsy, epistaxis, haematuria, jaundice, nephritis, diarrhoea, gonorrhoea[3] and dysuria.[12]

Pharmacological Activities: Antidiuretic,[12] Anti-inflammatory,[6] Neuroprotective[9] and Antibacterial.[13]

Dosage: Rhizome is used for the treatment of urodynia, pollakiuria, haematuria and fever. The recommended usual daily dose is 10 to 40 g in the form of a decoction, to be administered orally.[1]

Adverse Reactions: Allergic contact dermatitis[14] and sensitivity to grass pollen (Type I allergy).[15]

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

37. *Ipomoea pes-caprae* (L.) Sweet (Convolvulaceae)

Beach Morning Glory, Goat's Foot Creeper, Bayhops

Ipomoea pes-caprae plants

Ipomoea pes-caprae flower

Description: *Ipomoea pes-caprae* (L.) Sweet is a seaside trailing herb. The stems are long and the roots are found at the nodes. Leaves are simple, ovate, quadrangular or rounded, 2.5 cm by 10 cm, with slender petioles. Cymes are 1–3 flowered; the 5 sepals are ovate and blunt. The flower funnel is about 3–5 cm long. The plant bears dehiscent capsules (1.2 cm) containing hairy seeds that break up easily.[1–4]

Origin: Native to the Pantropic.[5]

Phytoconstituents: Pescaproside A & B, pescapreins I–IX, stoloniferin III, E-phytol, β-damascenone and others.[4,6–12]

Traditional Medicinal Uses: The juice squeezed from the plant is used by the Malays to treat fish stings.[3] In Burma, infusions of the plant with rusted iron are used to cure gynaecologic haemorrhages. [4] In India, the plant is known to be cooling, astringent and laxative. It is used internally as an emollient in Vietnam.[4,13] The plant has also been administered for headache.[14] In Thailand, it is used to treat dermatitis caused by jellyfish (e.g., Portuguese man-of-war).[15] The leaves are used in Indonesia to hasten the bursting of boils and used externally for injured feet. Sap from the young leaves are boiled in coconut oil and used to treat sores and ulcers. In the Philippines, the boiled leaves are used to treat rheumatism and fungal infection.[3,4,13] In Papua New Guinea, decoctions of the leaves are applied to sores. On the Solomon Islands, preparations of the leaves are ingested for digestive troubles.[4] The seeds are chewed with areca nut to sooth abdominal pains and cramps or consumed alone as a purgative.[3,13] Decoctions of the tubers are also ingested to treat bladder problems.[4]

Pharmacological Activities: Analgesic,[16] Antidiabetic,[17] Anti-inflammatory,[1,16,18] Antiplatelet aggregation,[14] Antiviral,[4] Antivenom,[4,19] Collagenase inhibitory activity,[12] Muscle relaxant[4,15] and Prostaglandin synthesis inhibition.[9]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

38. *Ixora chinensis* Lam. (Rubiaceae)

Chinese Ixora

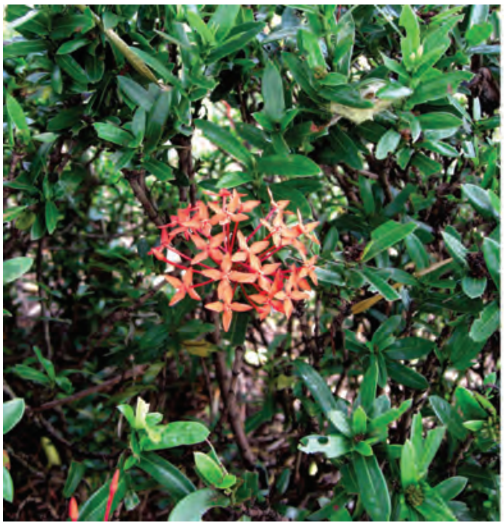

Ixora chinensis flowers

Ixora chinensis shrub

Description: *Ixora chinensis* Lam. is a small, dense shrub that can grow up to 2 m tall. Leaves are short-stalked, obovate-oblong, waxy, 6–10 cm long. Flowers are densely arranged, with 4 petals in bright red.[1]

Origin: This common garden flower is native to China and Thailand.[1]

Phytoconstituents: Ixoric acid, ixoroside, ixoside, geniposidic acid and others.[2–4]

Traditional Medicinal Uses: It is used for hypertension.[5] The whole plant is used to treat rheumatism, abscesses, bruises, and wounds. It is also used as an anodyne, and resolvent. It is beneficial to bone marrow and pregnant uterus as well.[6,7]

Pharmacological Activities: No information as yet.

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

39. *Jatropha curcas* L. (Euphorbiaceae)

Barbados Nut, Physic-nut

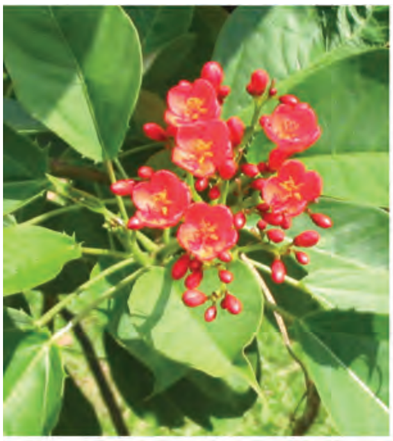

Flowers of *Jatropha curcas*

Jatropha curcas leaf

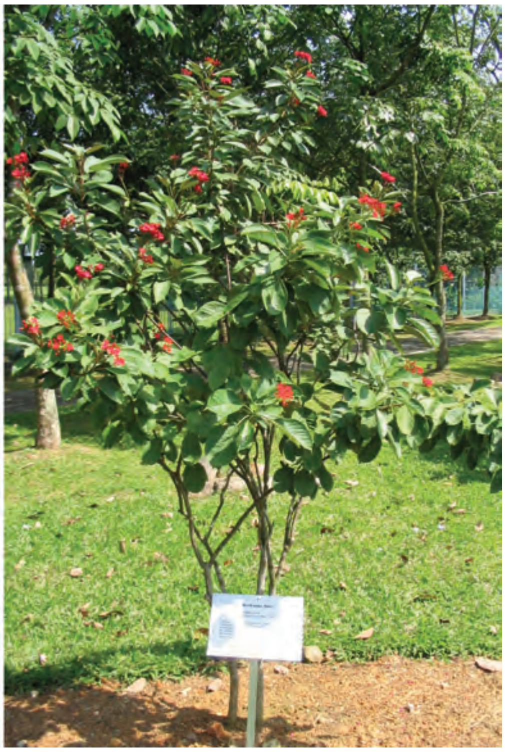

Jatropha curcas shrub

Description: *Jatropha curcas* L. is a woody shrub, 2–5 m tall. Stems are smooth and leaves are ovate, 5–15 cm long, green and glabrous. The blade is thin and 5-lobed-cordate. The flowers are 6 mm by 7 mm, hermaphrodite, and with fragrance. Fruit is 3-lobed fleshy capsule of about 2.5 cm in length, containing black seeds.[1,2]

Origin: Native to Mexico,[2] widely cultivated and naturalised in New and Old World Tropics.[3]

Phytoconstituents: Curcain, curcasin, curcacycline A, jatropholone A and B, heudelotinone, nobiletin, jatrocurin, curcin, curcusone B and others.[4–14]

Traditional Medicinal Uses: The leaves are used to treat scabies, parasites and as a rubefacient for paralysis and rheumatism.[15] The fruit is used for dropsy and anasarca.[16] The seed oil is emetic, laxative, purgative and for skin ailments.[4] The latex is applied to bee and wasp stings, to dress ulcers, sores and inflamed tongues. It is also used as a haemostatic agent,[15] and to treat whitlow, carbuncle and sores in mouth.[16] The roots are used in the form of a decoction as mouthwash for bleeding gums and toothache.[15]

Pharmacological Activities: Antidiabetic,[17] Anti-inflammatory,[18] Anticancer,[13,14,19] Antiprotozoal,[20] Antiviral,[21,22] Coagulant,[23] Abortifacient,[24] Haemolytic,[25] Lipolytic,[26] Insecticidal,[27–29] Molluscicidal[30] and Wound healing.[31]

Dosage: Two seeds have been used as a purgative.[32]

Adverse Reactions: No information as yet.

Toxicity: Human poisoning may lead to amnesia, convulsions, delirium, diarrhoea, nausea, vertigo, visual disturbances, acute abdominal pain, nausea, depression and collapse may also occur.[32] Intake of 4–5 seeds may cause death. It is toxic to rats, chicks and human.[32–40]

Contraindications: No information as yet.

Drug-Herb Interactions: May have interactions with other drugs that are metabolised by cytochrome P450 enzymes.[20]

[**Authors' Note:** *Jatropha curcas* is also cultivated as a source of biofuel.]

40. *Juniperus chinensis* L. (Cupressaceae)

Chinese Juniper

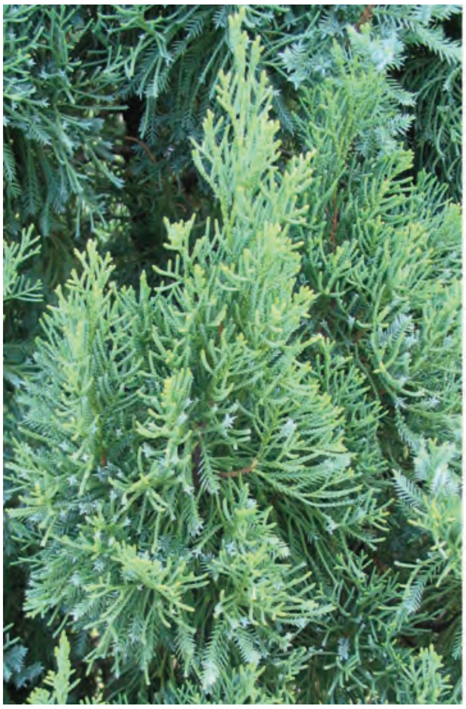
Leaves of *Juniperus chinensis*

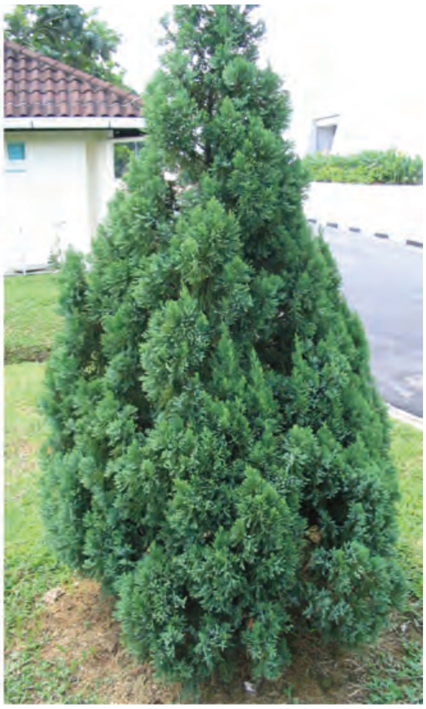
Juniperus chinensis tree

Description: *Juniperus chinensis* L. is conical shaped tree that can grow up to 25 m tall. Leaves are both scale-like and needle-like. Needle-like leaves are present on both young and adult plants, decussate or in whorls of three, loosely arranged, ascending, nearly lanceolate whereby scale-like leaves are present only in the adult. Seed cones are brown when ripe, usually glaucous, subglobose and 4–9 mm in diameter.[1]

Origin: Native to Myanmar and temperate Asia.[2]

Phytoconstituents: Chinensiol, yatein, podophyllotoxin, thujopsenal, widdrol, sandaracopimaric acid, hinokiic acid and others.[3–11]

Traditional Medicinal Uses: The leaves are used as a tonic to treat bleeding resulting from coughs.[12] It is also used for cold and haemorrhage. Liquor brewed from fresh leaves is used as a tonic and to treat hemoptysis. Others include treatment for convulsions, excessive sweating and hepatitis.[3] Its roots are used on burns, scalds and to promote hair growth on scars.[3]

Pharmacological Activities: Antibacterial,[13] Antifungal,[13,14] Antineoplastic,[15] Antioxidant,[6] Hypoglycemic[16] and Hypolipidaemic.[8,16]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

41. *Kaempferia galanga* L. (Zingiberaceae)

Galangal, Sand Ginger, Kencur

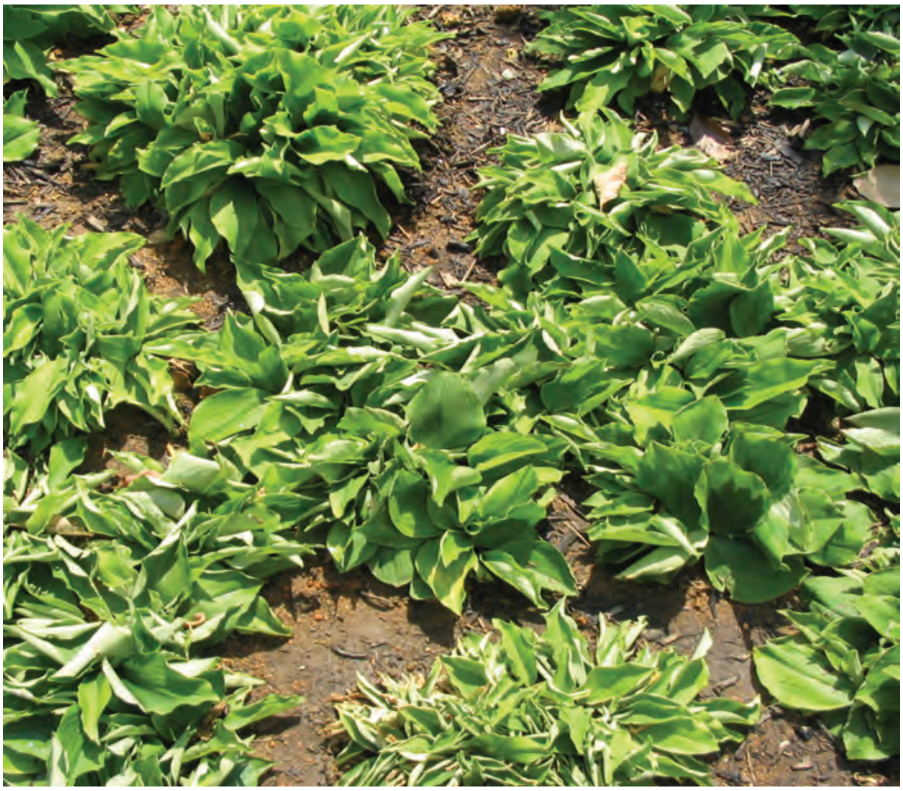

Kaempferia galanga herbs

Description: *Kaempferia galanga* L. is a small herb with short underground stems. Leaves are usually in pairs, oval, glabrous, pointed, 6–15 cm long, and spread out above ground with prominent veins. Flowers are in short stalked spikes. The corolla is white or pinkish, with violet spotted lip.[1]

Origin: Native to tropical Asia.[2]

Phytoconstituents: Ethyl cinnamate, 1,8-cineole, δ-3-carene, α-pinene, camphene, borneol, cyene, α-terpineol, α-gurjunene, germacrenes, cadinenes, caryophyllenes and others.[3–8]

Traditional Medicinal Uses: The whole plant is used as a postpartum protective medicine, treatment for stomachache, diarrhoea, dysentery, treatment for rheumatism, swellings, fever, coughs, asthma and as a tonic/lotion.[8] In Malaya, the leaves and rhizomes are chewed to stop cough. In Indonesia, it is used for abdominal pain, for swelling and muscular rheumatism. In the Philippines, the rhizome is used for boils, chills, dyspepsia, headache and malaria. The Indians also use the rhizomes as lotions, poultices for fever, rheumatism, sore eyes, sore throat and swellings.[3] The rhizomes are stimulant, used to treat toothache, chest pains and constipation.[9] They are also used as carminative, prophylactic, stomachic, for dandruff and scabs. A decoction of the rhizome is used for cholera, contusion, dyspepsia, headache, lameness, lumbago, and malaria. It is also roasted and applied to rheumatism and tumours.[3] To facilitate delivery during birth, it is mixed with the juice of *Curcuma montana, C. aromatica* and ginger rhizomes and consumed.[10]

Pharmacological Activities: Antibacterial,[11,12] Antifungal,[13] Antihypertensive,[14] Anti-inflammatory,[8] Antineoplastic,[15–18] Antioxidant,[19] Antiprotozoal,[20] Depressant,[21] Immunomodulatory,[21] Vasorelaxant,[22] Antiallergy,[23] Insect repellent,[24,25] Insecticidal[25–28] and Wound healing.[29]

Dosage: Oral doses range from 3 to 6 g of the rhizome per day, administered in the form of decoction, powder or pill for the treatment of pectoral and abdominal pains, headache, toothache and cold.[30]

Adverse Reactions: No information as yet.

Toxicity: The ethanolic extract injected intraperitoneally in increasing doses of 25, 100, 250 and 800 mg/kg body weight resulted in a decrease in motor activity, respiratory rate, loss of screen grip and analgesia in rats. A dose of 2000 mg/kg body weight was observed to be lethal. Acute and subacute oral toxicity test of *Kaempferia galanga* produced neither mortality nor significant differences in the body and organ weights between controls and treated rats.[21]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

42. *Lantana camara* L. (Verbenaceae)

Common Lantana, Bunga Tahi Ayam, Wild Sage

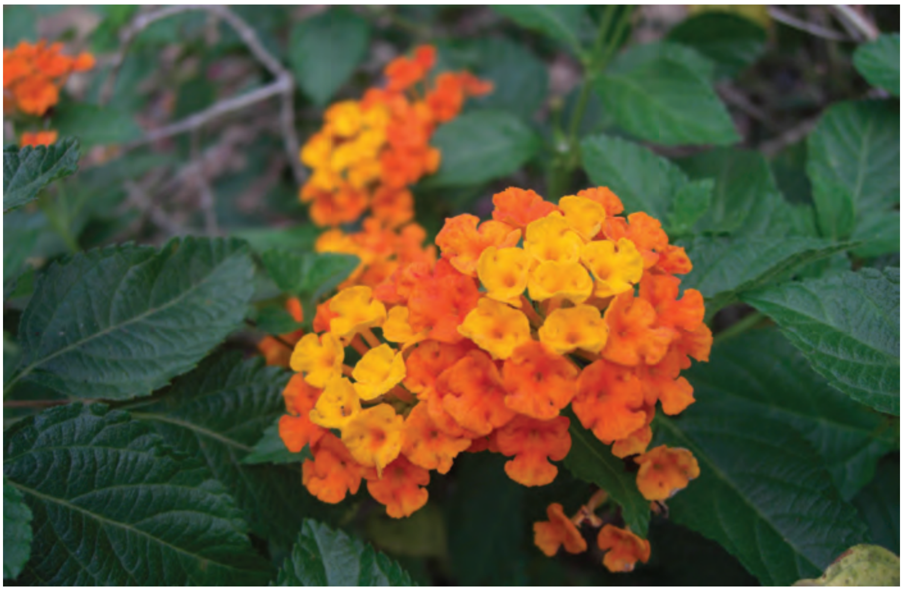

Lantana camara flowers

Description: *Lantana camara* L. is a small shrub with long, weak branches. Leaves are opposite, ovate, 5–12 cm long with a pungent scent. Flowers are in dense spikes, with salver-shaped corolla, 1–1.2 cm long, orange, red, pink or variegated. Fruit is bluish, globose and 4 mm in diameter.[1]

Origin: Native to Mexico and Southern America.[2]

Phytoconstituents: Lantadene A and B, lantanic acid, lantanilic acid, icterogenin, lantanose A & B, lamiridoside, geniposide, δ-guaiene, camarinic acid and others.[3–14]

Traditional Medicinal Uses: The whole plant is used as a bath for scabies and leprosy.[3,15] The entire plant is also a carminative, diaphoretic, vulnerary, used for fistulae, pustules, tumours, treating malaria, rheumatism, tetanus and spasms.[3] In Budiope county, Uganda, the leaves are used for treating malaria.[16] The root and leaves are decocted to treat fever, mumps, neurodermatitis and traumatic injuries.[3,16]

Pharmacological Activities: Anthelmintic,[7,17] Antimutagenic,[13] Antibacterial,[13,18–21] Anticancer/Antineoplastic,[22,23] Antifertility,[24] Anti-inflammatory,[25] Antimicrobial,[26–28] Antimotility,[29] Antioxidant,[30] Antithrombotic,[31] Antiviral,[32] Antifilarial,[33] Insect repellent,[34,35] Insecticidal,[36–39] Larvicidal,[40] Molluscicidal[10,41] and Spermatotoxic.[42]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: Ingestion of green berries had been reported to cause human fatalities.[43] *L. camara* is harmful to animals[44–58] and causes teratogenicity in rats.[59]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

43. *Lonicera japonica* Thunb. (Caprifoliaceae)

Japanese Honeysuckle, Jin Yin Hua

Lonicera japonica shrub

Lonicera japonica leaves

Description: *Lonicera japonica* Thunb. is a climbing shrub having tomentose young leaves and stems. Leaves are simple, opposite and exstipulate. Blade is elliptic, 3–8 cm by 2–3 cm, truncate at base, obtuse and chartaceous. Flowers are axillary, white, and turns yellow upon maturity. Fruits are globose and black.[1]

Origin: A native of East Asia, widely cultivated and naturalised throughout the world.[2]

Phytoconstituents: Linalool, luteolin, geraniol, aromadendrene, eugenol, loniceroside A, B, C, L-phenylalaninosecologanin, (Z)-aldosecologanin, (E)-aldosecologanin and others.[3–8]

Traditional Medicinal Uses: In China, the flowers are used for influenza, boils and carbuncle.[3] In Malaysia, decoctions of dried flowers are used for cooling, flu, fever, headache, and boils. Distilled flowers are used to produce a medicine for treating postprandial stomachaches.[3] Flower tea is prescribed to treat fever, sore throat, mouth sores, headache, conjunctivitis, keratitis, corneal ulcers, breast infections, muscle and joint pain, stomach problems, diarrhoea, and painful urination.[9] They are used in the treatment of arthritis and inflammation.[10] Flower buds are used in infusions for cutaneous infections, scabies, as diuretic and treat bacterial infection. Decoction is used for bacterial dysentery, cold, enteritis, infected boils, laryngitis, lymphadentitis, rheumatism and sores. The flowers and stems are regarded as cooling and are used to treat aching bones and boils. Other uses include intestinal inflammation, stomach ulcers, painful haemorrhoids, sore throats and intoxication.[3]

Pharmacological Activities: Antibacterial,[11,12] Anticancer/Antineoplastic,[12–15] Antifungal,[16] Antihypertensive,[17] Anti-inflammatory,[9,18–25] Antioxidant,[7] Antiplatelet,[26] Antiviral,[27] Hepatoprotective[28] and Antiatherogenic.[29]

Dosage: In China, 10–60 g of dried floral buds are used for decoction. A combination of 10 g of honeysuckle, 10 g of forsythia, a little mint and bamboo leaf is a prescription for a bad cold from a drugstore in China.[30] About 9–15 g dried flowers has been used in decoction, pills, powder or poultice of the powder.[31] The recommended daily dose is 4 to 8 g of flowers or 10 to 20 g of stems and leaves in the form of a decoction, infusion, extract or alcoholic maceration for the treatment of boils, impetigo, urticaria, allergic rhinitis, fever, malaria, erythema, measles, diarrhoea, dysentery, syphilis, rheumatism and lichen tropicus.[32] For carbuncles, boils, erysipelas, acute dysentery, pharyngitis, upper respiratory infection and epidemic febrile diseases, 6–15 g of dried flowers are used.[33]

Adverse Reactions: Linear, itchy, raised blisters on the skin may occur on contact.[34]

Toxicity: No sign of acute and subacute toxicity was observed.[35]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

44. *Mangifera indica* L. (Anacardiaceae)

Mango, Mangga

Mangifera indica tree

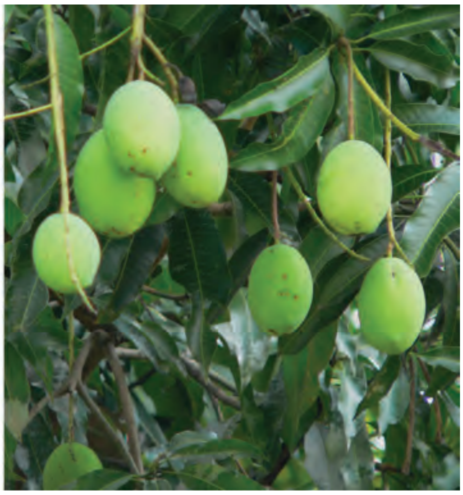
Fruits of *Mangifera indica*

Description: *Mangifera indica* L. is a fruit tree which grows up to 8 m high. Bark is grey and fissured. Leaves are simple, 12–30 cm by 4–9 cm, narrowly elliptic, pointed, and slightly leathery with wavy edges. Flowers are very small, greenish-yellow or white, fragrant and arranged in panicles. The fruit is kidney-shaped and the yellow to orange flesh is edible. Seed is elongated, fibrous and flattened.[1]

Origin: Native to India and Indochina.[2]

Phytoconstituents: Mangiferin, ambolic acid, ambonic acid, arabinan, mangiferonic acid, quercitin, violaxanthin and others.[3–7]

Traditional Medicinal Uses: The leaves are used in the form of ashes for burns and scalds; chewed to strengthen the gums, while the burning leaf smoke is inhaled for hiccups and other throat ailments. They are also used for skin ailments, asthma and cough.[4] Paste of leaves is applied to warts and used as styptic ointments.[3] The leaves are also used in the treatment of malaria in Budiope county, Uganda.[8] The fruits are used to treat pain in abdomen, diarrhoea, and to quench thirst (with Aegle and salt).[9] Pulp of the fruit is used in China to promote blood circulation, the fruit rind as tonic in Burma and dried slices of young fruits are used for septicaemia in Palau.[3] Its seeds are used for stubborn colds, coughs, diarrhoea, vermifuge and menorrhagia.[4] They are also used for asthma.[9] In the Philippines, raw seeds are used to expel worms and the roasted ones are for diarrhoea.[3] The bark is used for the treatment of fever or sunstroke, cholera, rheumatism, sty in eye, ulcerated tongue (with roots of *Ichnocarpus* and bark of *Zizyphus rugosa*), haemiplegia, diarrhoea (with bark of *Streblus asper*, roots of *Oroxylum indicum* and *Helianthus annuus*), dysentery (with bark of *Streblus* and *Spondias pinnata*), poisoning, uterine haemorrhage and jaundice.[9]

Pharmacological Activities: Analgesic,[10] Anthelmintic,[11] Antibacterial,[12–14] Anticonvulsant,[15] Antidiarrhoeal,[13] Antifungal,[16] Anti-inflammatory,[10,11,17–19] Anticancer,[20-23] Antioxidant,[24–38] Antiviral,[39,40] Gastroprotective,[41,42] Hepatoprotective,[43,44] Antiprotozoal,[15] Hypoglycaemic,[45–47] Hypolipidaemic,[46,48] Immunomodulatory,[49] Larvicidal[50] and Radioprotective.[51,52]

Dosage: As an anthelmintic, it is recommended to take 20–30 g of powdered seed.[53]

Adverse Reactions: Anaphylactic reaction following ingestion of fruit.[54]

Toxicity: Intraperitoneal LD_{50} (50% ethanolic extract) is > 1000 mg/kg in mice.[55]

Contraindications: No information as yet.

Drug-Herb Interactions: Modulation of P450 isozymes.[56,57]

45. *Manihot esculenta* Crantz (Euphorbiaceae)

Tapioca, Cassava

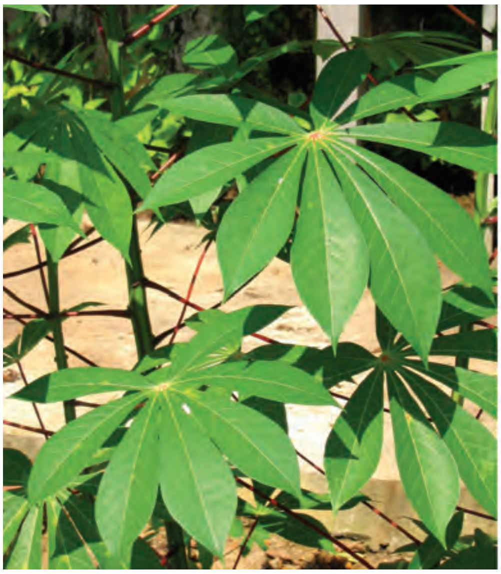

Leaves of *Manihot esculenta*

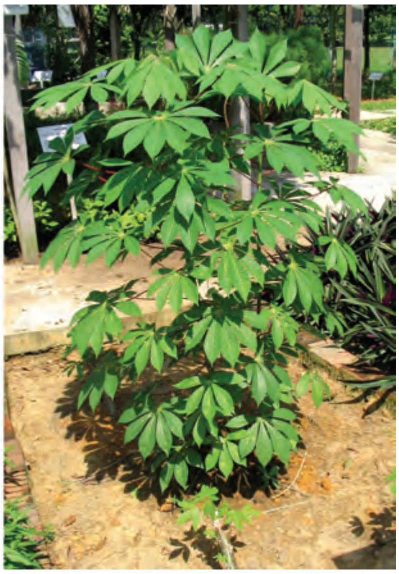

Manihot esculenta shrub

Description: *Manihot esculenta* Crantz is an erect woody shrub that can grow up to 5 m tall. Leaf blades are palmately 3–9-lobed, 5–20 cm. Lobes are oblanceolate to narrowly elliptic, entire, 8–18 cm by 1.5–4 cm. Root tubers beneath the ground yield tapioca, which is fleshy, starchy and edible.[1,2]

Origin: Native to Brazil, cultivated throughout the tropics.[2]

Phytoconstituents: Linamarin, esculentoic acid A and B, esculentin, esculin, scopoletin, scopolin, oxalic acid, saponins and others.[3–8]

Traditional Medicinal Uses: Rhizome is made into a poultice and applied to sores.[5]

Pharmacological Activities: Antifungal,[6] Antineoplastic,[7] Hepatoprotective,[9] Hypercholesterolaemic,[8] Antithyroidal,[10] Neurotoxic[11] and Superoxide dismutase inhibition.[12]

Dosage: No information as yet.

Adverse Reactions: Several minor acute intoxications were seen, with complaints of nausea, vomiting, abdominal pain and headache following a meal of cassava.[12]

Toxicity: Hydrocyanic acid — oral, human LD_{50} 570 μg/kg; Oxalic acid — oral, human LD_{50} 700 mg/kg; Saponin — oral, mouse LD_{50} 3000 mg/kg; Tryptophane — oral, rat TD_{50} 1100 mg/kg. Bitter cassava juice can cause death due to cyanide poisoning.[3] Haemorrhage, necrosis, fibrosis and atrophy of the acinar tissue and fibrosis of the islets of Langerhans of the pancreas occurred in dogs fed on cassava.[13] Epidemic spastic paraparesis (konzo) is known to be due to long-term intake of cassava (*M. esculenta*).[10] Tropical ataxic neuropathy, a polyneuropathy with sensorineural hearing loss and optic atrophy can also result from intake of cassava in humans.[14]

Contraindications: No information as yet.

Drug-Herb Interactions: Co-administration of cassava rich diet and alcohol is found to reduce the alcohol induced toxicity in rats.[7]

46. *Melaleuca cajuputi* Roxb. (Myrtaceae)

Gelam, Paper-bark Tree, Kayu Puteh

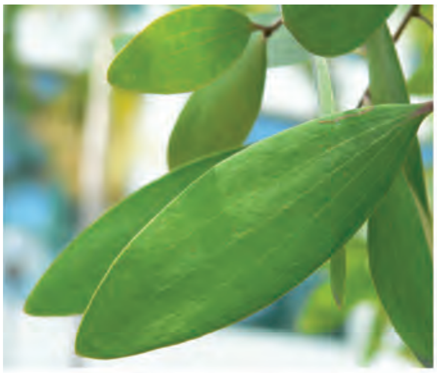

Leaves of *Melaleuca cajuputi*

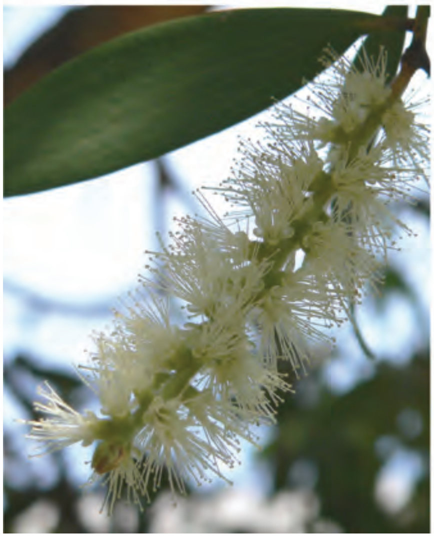

Melaleuca cajuputi flowers

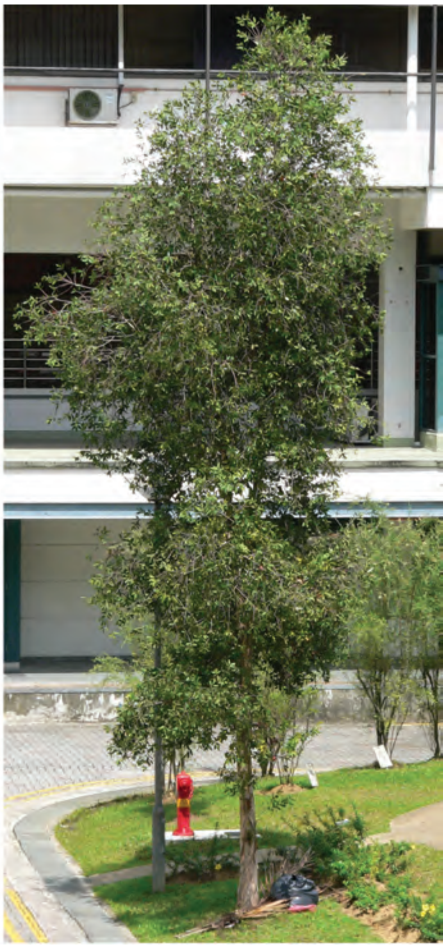

Melaleuca cajuputi tree

Description: *Melaleuca cajuputi* Roxb. is a medium sized tree with an often twisted trunk when the tree is very old. Bark is light brown, flaky and peeling. Leaves are simple, slightly curved, with 5–7 longitudinal veins. Flowers are white and borne in dense spikes. Seeds are borne in a capsule.[1–3]

Origin: Native to tropical Asia and Australasia.[4]

Phytoconstituents: Cajeputol, cineole, β-pinene, eugenol, phellandrene, α-terpineol, eugenetin, isoeugenetin and others.[2,5,6]

Traditional Medicinal Uses: In Malaysia, it is used for the treatment of colic and cholera. It is also used externally for thrush, vaginal infection, acne, athlete's foot, verruca, warts, insect bites, cold sore and nits.[5] Cajuput oil is distilled from the leaves and used by the Burmese to treat gout. The Indochinese uses cajeput oil for rheumatism and pain in the joints and as an analgesic.[2] The oil is used externally in Indonesia for burns, colic, cramps, earache, headache, skin diseases, toothache and wounds. When administered internally, it can induce sweating and act as a stimulant and antispasmodic. In the Philippines, the leaves are used to treat asthma.[2,6]

Pharmacological Activities: Antibacterial.[6]

Dosage: For treatment of coryza, influenza, cough, asthma, dyspepsia, earache, toothache, rheumatism, osteodynia, neuralgia, wounds, burns, post partum haematometra. The reported dose for the above ailments is 20 to 40 g of fresh leaves or 5 to 10 g of dried leaves in the form of a decoction or infusion.[7]

Adverse Reactions: Contact dermatitis may occur. Glottal spasms or bronchial spasms or asthma-like attacks may occur if the oil is applied to the facial areas of infants and small children.[8]

Toxicity: Due to the high cineole content, life-threatening poisonings can occur with overdoses of cajuput oil (more than 10 g). Symptoms of poisoning are reduction in blood pressure, circulatory disorders, collapse and respiratory failure.[8]

Contraindications: Do not consume orally in severe liver diseases and presence of inflammatory condition of gastrointestinal tract or of the biliary ducts. Cajuput oil preparations are not to be applied to the faces of infants or small children.[8]

Drug-Herb Interactions: No information as yet.

47. *Melastoma malabathricum* L. (Melastomaceae)

Sendudok, Singapore Rhododendron

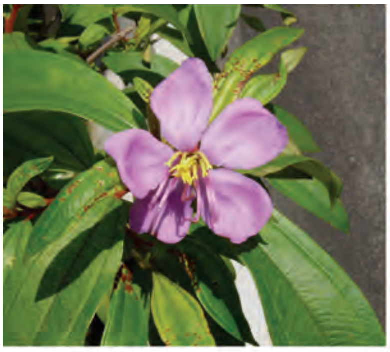

Melastoma malabathricum flower

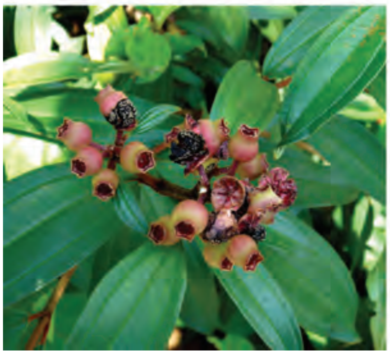

Fruits of *Melastoma malabathricum*

Melastoma malabathricum shrub

Description: *Melastoma malabathricum* L. is a small shrub that can grow up to 1 m tall, densely clothed with tiny scales. Leaves are simple, ovate, elliptic, 4–14 cm by 1.7–3.5 cm, and stiffly papery. Flowers are purple with five petals and fruits are urceolate-globular, 6–15 mm by 6–12 mm and succulent.[1,2]

Origin: Native to tropical and temperate Asia and the Pacific Islands.[2]

Phytoconstituents: Malabathrins A–F, nobotanins B, G, H and J, kaempferol and others.[3–9]

Traditional Medicinal Uses: In Indonesia and Malaysia, the leaves and roots are used as a remedy for diarrhoea and dysentery. The leaves are also used to wash ulcers, to treat piles and for small pox pustules to prevent development of scars.[10] Its bark is used to treat skin diseases.[11]

Pharmacological Activities: Analgesic,[12] Anti-inflammatory,[13] Antineoplastic,[14] Antinociceptive,[13] Antiviral,[14,15] Antioxidant[15] and Antiplatelet.[16]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: *M. malabathricum* was reported to be an aluminium (Al) accumulator (with up to 10 g Al/kg of dry weight in old leaves and up to 7 g Al/kg dry weight in new leaves) and could lead to accumulation of aluminium in the bone and brain, leading to neurotoxicity, when consumed in large amount.[17]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

48. *Mimosa pudica* L. (Leguminosae)

Touch-me-not, Sensitive Plant, Rumput Simalu

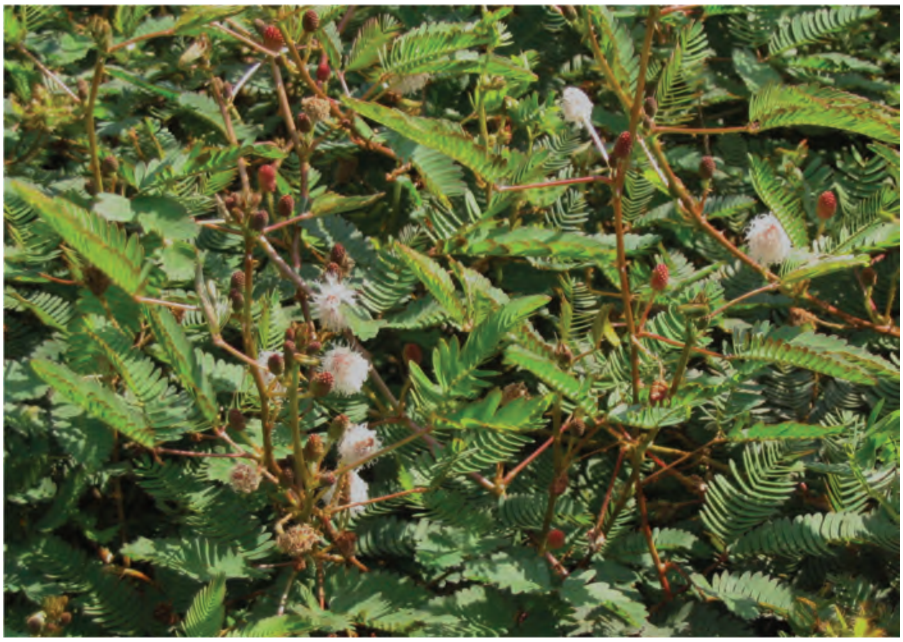

Mimosa pudica leaves and flowers

Description: *Mimosa pudica* L. is a prickly, herbaceous weed. Leaves bipinnate, very sensitive, fold together when touched, in rain or at night. Leaflets are 15–20 pairs, small oblong, nearly sessile. Flowers are pink and fruits are flat pods covered with bristles.[1–3]

Origin: Native to Brazil.[3]

Phytoconstituents: Mimosine, 2-Hydroxymethyl-chroman-4-one and others.[4–6]

Traditional Medicinal Uses: The plant is used on cuts and wounds.[4,7] It is also used for childbirth and infertility in Trinidad and Tobago.[8] Bath with plant decoction relieves insomnia.[4] The leaves are used for hydrocele, dressing for sinus, sores, piles and swelling of feet. In Mexico, aqueous extracts from dried leaves are employed to alleviate depression.[9] In the Philippines, the leaves soaked in coconut oil is used for ulcers.[4]

Pharmacological Activities: Anthelmintic.[10] Antibacterial,[11] Anticonvulsant,[12] Antidepressant,[9] Antifertility,[13] Antifungal,[5] Hyperglycaemic,[14] Antioestrogenic[15] and Antivenom.[16,17]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

49. *Mirabilis jalapa* L. (Nyctaginaceae)

Four O'Clock Flower, Bunga Pukul Empat

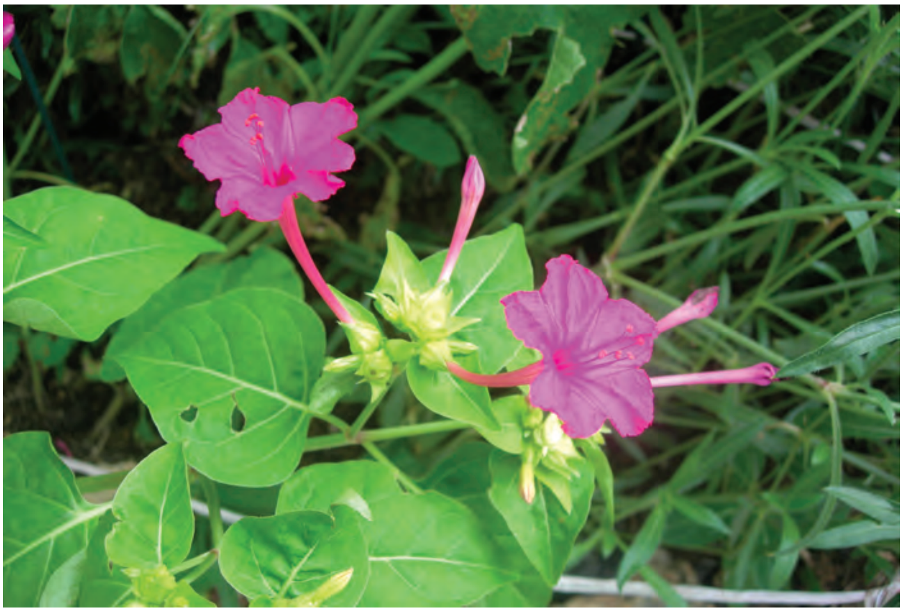

Flowers of *Mirabilis jalapa*

Description: *Mirabilis jalapa* L. is an erect herb that can grow up to 1 m tall. Leaves are simple, heart-shaped, 3–12 cm long, opposite, tapering to a pointed end. Flowers are bisexual, red, pink, yellow or white, with perianth distinctly constricted above, and they bloom late in the afternoon. Fruits are black and globose, 5–8 mm in diameter.[1,2]

Origin: Native to tropical America, introduced in China and in many tropical areas.[2]

Phytoconstituents: Trigonellin, 2′-*O*-methylabronisoflavone, 6-methoxy-boeravinone C and betaxanthins.[3,4]

Traditional Medicinal Uses: Its leaves are used as a decoction for abscesses, juice for wounds and cooked with pork as tonic.[5] The leaves are also placed on boils, blisters, and to relieve urticaria.[6] In Indochina, the seeds are used as a purgative.[3] The flowers release a strong odour at night which will stupefy or drive away mosquitoes.[7] The roots are used as a laxative in a decoction, with or without pork, for colds, inflammation and leucorrhoea.[5] In Malagassy, Madagascar, they are used to treat intestinal pains. In South Africa, the roots are used as purgative agents.[7]

Pharmacological Activities: Antibacterial,[8,9] Antifungal,[4,10] Antineoplastic and Abortifacient.[11]

Dosage: Approximately 8–10 g of roots are taken as a purgative.[12]

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

50. *Morinda citrifolia* L. (Rubiaceae)

Mengkudu, Indian Mulberry, Noni

Flowers and fruits of *Morinda citrifolia*

Morinda citrifolia tree

Description: *Morinda citrifolia* L. is a tree that grows up to 9 m. Leaves are simple, large, 10–15 cm by 20–30 cm, decussate and stipulate. Blade is broadly elliptic to obovate, glossy, soft and succulent. Flowers are small, sessile, white, 1.5–2 cm across, terminal and axillary. The fruit is succulent, oval, 5–7 cm across, light greyish green and turns yellow upon maturity.[1–5]

Origin: Native to tropical and temperate Asia and Australasia.[6]

Phytoconstituents: Morintrifolins A and B, morindin, morindone, rubiadin, morindadiol, morindicone, morinthone, morindicinone, morindicininone, noniosides E–H, morinaphthalenone and others.[1,4,7–28]

Traditional Medicinal Uses: The whole plant is used to treat aching bones and arthritis. In Malaysia, the heated leaves are applied to the chest and abdomen to treat coughs, nausea, colic and enlarged spleen. In Japanese and Chinese medicine, *M. citrifolia* is used to treat fever and as a tonic whereas in Indochina, the fruit is prescribed for lumbago, asthma and dysentery.[1] A decoction of the leaves taken orally is effective for the treatment of fever, dysentery and diarrhoea. Poultice of fresh leaves can be used to cure furunculosis. The fruit when consumed together with a little salt is stomachic, aperient, and active on dysentery, uterine haemorrhage, cough, coryza, oedema and neuralgia. The root bark has beneficial effects in hypertension, osteodynia and lumbago.[7] The fruit is also used for throat and gum complaints, dysentery and leucorrhoea while the root is used as a cathartic and febrifuge.[29] The processed fruit juice is also in great demand for various ailments such as diabetes, high blood pressure, headaches, heart disease, AIDS, cancers, gastric ulcers, mental depression, senility, poor digestion, atherosclerosis and drug addiction.[30]

Pharmacological Activities: Analgesic,[30] Anthelmintic,[30,31] Antiangiogenic,[32] Antibacterial,[11,30,33] Anticancer/Antineoplastic,[34–37] Antitumour,[30] Antihypertensive,[30] Antiviral,[30] Anti-inflammatory,[24,25,38] Antioxidant,[14,16,39–44] Antiprotozoal,[45] Antidiabetic,[26] Sedative/Anxiolytic,[25] Chemopreventive,[24] Insecticidal,[9] Wound healing[46] and Hepatoprotective.[47]

Dosage: For treatment of hypertension, osteodynia and lumbago, 10 to 20 g of rootbark is prescribed daily as a decoction or alcoholic maceration of torrefied material.[7]

Adverse Reactions: Sedation, nausea, vomiting, anorexia, hypersensitivity, hyperkalemia may occur.[48, 49]

Toxicity: The lyophilised aqueous extract of *M. citrifolia* roots did not exhibit any toxic effects but did show a significant, dose-related, central analgesic activity in the writhing and hotplate tests.[50] Chemical analysis and genotoxicity tests revealed that noni juice does not have a genotoxic potential and that genotoxic anthraquinones do not exist in noni juice.[51] However, according to some, anthraquinones are the most likely hepatotoxic components found in *M. citrifolia*.[52] Hepatitis induced by Noni juice had been reported.[53,54]

Contraindications: Should not be used during pregnancy and lactation. Contraindicated in people with hyperkalemia or with hypersensitivity to *M. citrifolia*.[48]

Drug-Herb Interactions: No information as yet.

51. *Nelumbo nucifera* Gaertn. (Nymphaeaceae)

Sacred Lotus, East Indian Lotus, Oriental Lotus

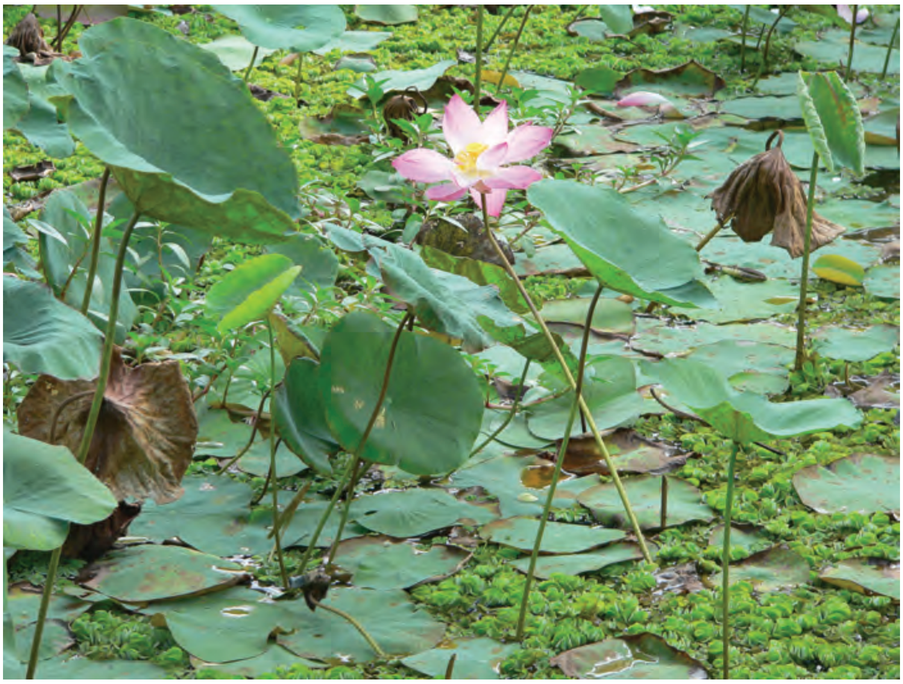

Nelumbo nucifera plants with flowers

Description: *Nelumbo nucifera* Gaertn. is an aquatic plant that grows in shallow waters. Leaves are green, round, 30–60 cm across and with long petiole. Flowers are pink, white or red, 10–30 cm and solitary. Fruits are non-edible and non-fleshy.[1]

Origin: Native to tropical and temperate Asia, Australia and Eastern Europe.[2]

Phytoconstituents: Nuciferin, nornuciferin, nelumboroside A & B, nelumstemine, dotriacontane, ricinoleic, roemerin, liensinine, neferine, lotusine, liriodenine, asimilobin, pronuciferine and others.[3–8]

Traditional Medicinal Uses: The leaves are used to treat sunstroke, diarrhoea, dysentery, fever, dizziness and vomiting of blood.[9] The plant is used as an antidote for mushroom poisoning[9] and for smallpox.[10] In Ayurveda, the plant is used to treat cholera, diarrhoea, worm infestation, vomiting, exhaustion and intermittent fever.[3] The fruits are used in decoction for agitation, fever, heart and haematemesis while the stamens are used to "purify the heart, permeate the kidneys, strengthen virility, to blacken the hair, for haemoptysis and spermatorrhoea".[10] They are also used to treat premature ejaculation,[9] as astringent for bleeding,[3] excessive bleeding from the uterus,[9] abdominal cramps, bloody discharges, metrorrhagia, non-expulsion of the amniotic sac,[10] and as cooling agent during cholera.[11] The seeds are believed to promote virility, for leucorrhoea and gonorrhoea.[10] Powdered beans are used in treating digestive disorders, particularly diarrhoea.[3] They are also used as a tonic,[9] for enteritis, insomnia, metrorrhagia, neurasthenia, nightmare, spermatorrhoea, splenitis and seminal emissions.[10] The roots are for the treatment of diarrhoea, dysentery, dyspepsia, ringworm and other skin ailments and as a tonic as well.[10,11]

Pharmacological Activities: Antianxiety,[12] Antiarrhythmic,[13] Antibacterial,[4] Anticonvulsant,[14] Antidiarrhoeal,[15] Anti-inflammatory,[16] Hepatoprotective,[18] Antioxidant,[8,17–20] Antiplatelet,[21] Antiproliferative,[22] Antipyretic,[23,24] Antiviral,[25,26] Hypoglycaemic,[27,28] Hypolipidaemic,[29] Immunomodulatory[30] and Insecticidal.[4]

Dosage: A daily dose of 10 to 30 g of the ripe seeds as a decoction or powder is used for the treatment of neurasthenia, spermatorrhoea and metrorrhoea. The pericarps should be removed before using the seeds. Decoctions of 15 to 20 g and 2 to 4 g of dried leaves and seed cores respectively have been used for treating insomnia, haemorrhage and haematemesis. 6 to 12 g of plumules, 5 to 10 g of filaments, or 15 to 30 g of the receptacles in the form of a decoction, are used in the treatment of bloody stools, haematuria, uterine haemorrhage and haematemesis.[31] A daily dose of 5 to 8 g of rhizome or seeds taken by mouth is recommended.[32]

Adverse Reactions: Stamen, receptacle, rhizome node, leaf, and embryo can be safely consumed with proper usage.[33] No known side effects with appropriate therapeutic dosages.[3]

Toxicity: No information as yet.

Contraindications: Seed is contraindicated in constipation and stomach distention.[33]

Drug-Herb Interactions: No information as yet.

52. *Nephelium lappaceum* L. (Sapindaceae)

Rambutan, Hairy Lychee

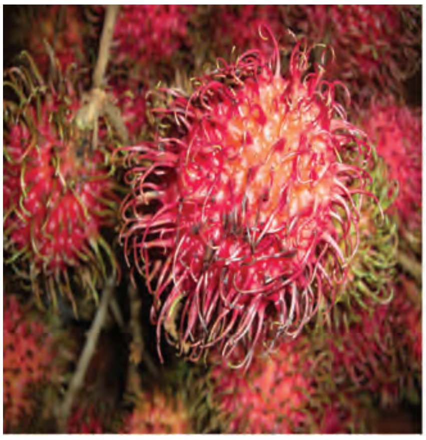

Nephelium lappaceum fruits

Nephelium lappaceum tree

Description: *Nephelium lappaceum* L. is an evergreen tree that can grow up to 10 m tall. Leaves are pinnate, 25–40 cm long, compound with 2–4 pairs of leaflets. Blades are elliptic or obovate, 6–18 cm by 4–7.5 cm, thinly leathery, glabrous, with 7–9 pairs of lateral veins. Flowers are greenish white, fragrant and with no petals. The fruit is hairy, turning from green to red when mature and contains one hard seed.[1–3]

Origin: Native to Indonesia, Malaysia, Philippines and Thailand.[3]

Phytoconstituents: Type II and III cyanolipids, β-caryophyllene, monoterpene lactones 1 & 2, paullinic acid and others.[4–7]

Traditional Medicinal Uses: The fruits are used to treat diarrhoea and fever,[2] and also for the treatment of dysentery and dyspepsia.[8] Its seeds are narcotic[8] and the bark is used to treat disease of the tongue while the roots are used for fever.[2]

Pharmacological Activities: Antiviral,[9] Antifungal,[6] Antibacterial,[10] Antioxidant[10–13] and Cytotoxic.[12,13]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

53. *Nerium oleander* L. (Apocynaceae)

Oleander

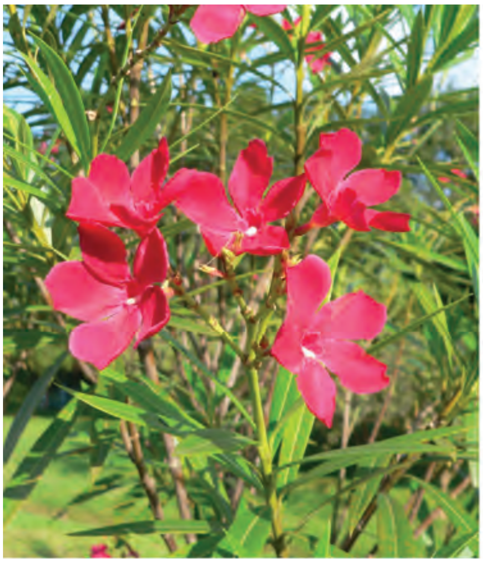

Flowers of *Nerium oleander*

Nerium oleander shrub

Description: *Nerium oleander* L. is a small shrub up to 2 m high. Leaves are very narrowly elliptic, 5–21 by 1–3.5 cm, dark green, without stipules, leathery and arranged in whorls of three. Flowers are showy and fragrant. Sepals are narrowly triangular to narrowly ovate, 3–10 mm. Corolla is purplish red, pink, white, salmon, or yellow. Fruits consist of cylindrical follicles, 12–23 cm. Seeds are oblong, coma, about 0.9–1.2 cm.[1,2]

Origin: Native to southern Europe,[2] and widely cultivated and naturalised in Asia, Europe and North America.[1]

Phytoconstituents: Oleandroside, kaneroside, neriaside, nerigoside, neriumoside, neridiginoside, nerizoside, neritaloside, proceragenin, neridienone A, cardenolides N-1 to N-4 and others.[3–11]

Traditional Medicinal Uses: The plant is used in Ayurveda to treat scabies, eye disease and haemorrhoids.[3] It is used to treat parasitic infection in Calabria (Southern Italy).[12] Leaf decoction is used to treat diabetes in southeastern Morocco.[13] Bark, leaf, flower are used medicinally as a cardiotonic and diuretic.[14]

Pharmacological Activities: Analgesic,[15] Anti-inflammatory,[8,9,15] Antibacterial,[16] Anticancer/Antineoplastic,[8,9,17–22] Antifungal,[23] Depressant,[24,25] Antimitotic,[26] Insecticidal,[27] Larvicidal,[28] Muscle stimulatory[29–30] and inhibits Nuclear factor-kappa B (NF-κB) activation.[31]

Dosage: No information as yet.

Adverse Reactions: Depression, dizziness, stupor, headache, nausea, vomiting, anorexia, abdominal cramps, spontaneous abortion, hypersensitivity, contact dermatitis, hyperkalemia and tachypnea.[32]

Toxicity: Toxic to humans[33–39] and animals.[40–48] The plant contains numerous toxic compounds, many of which can be deadly to people.[33–39] Ingestion can cause both gastrointestinal and cardiac effects and also affect the central nervous system.[33–39]

Contraindications: Should not be used during pregnancy and lactation, in children and in persons with hypersensitivity to oleander.[32] Should not be taken internally.[49]

Drug-Herb Interactions: Fatal digitalis toxicity can occur with concurrent usage of cardiac glycosides such as digoxin and digitoxin.[32] Concurrent use of quinidine, calcium salts, saluretics, laxatives or glucocorticids increases efficacy as well as side effects.[3]

54. *Ophiopogon japonicus* Ker-Gawl. (Liliaceae)

Dwarf Lilyturf, Mondo Grass, Mai Men Dong

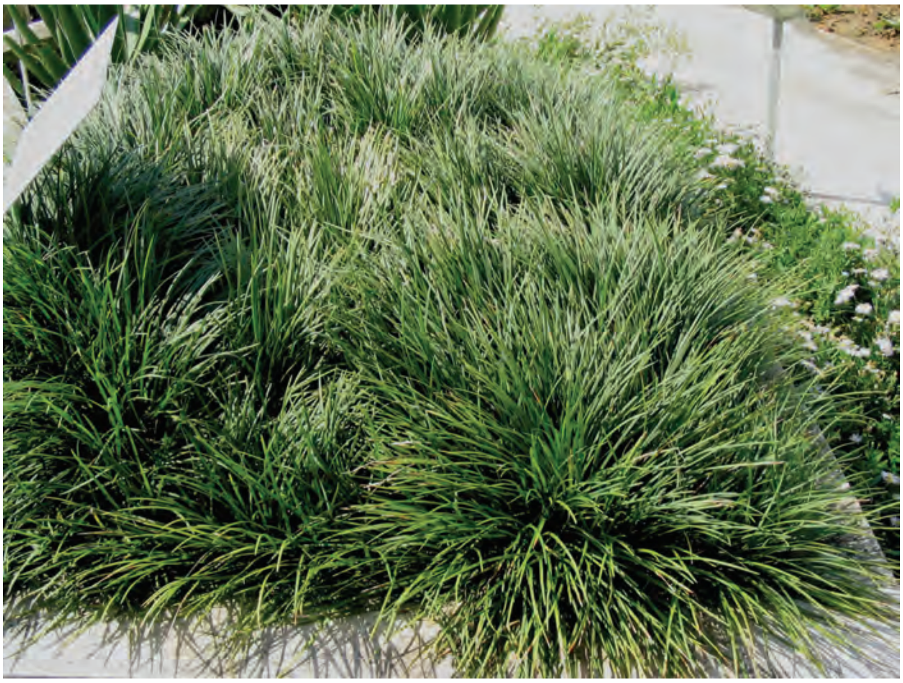

Ophiopogon japonicus herb

Description: *Ophiopogon japonicus* Ker-Gawl. is an evergreen, stemless, rhizomatous herb. Leaves are sessile, long, 10–50 cm by 2–4 mm, linear, grass-like, 3–7 veined, and have pointed tips. Flowers are either solitary or paired.[1,2]

Origin: Native to Japan and Korea.[2]

Phytoconstituents: Ophiopogonin D and E, ophiopogonin C′ and D′, bornanol, ophiopogonanone A, C, E and F, ophiopojaponin D and others.[3–9]

Traditional Medicinal Uses: The roots are used to cool the body system, as a tonic, purgative, thirst quencher, treatment for sore throat, cough and fever.[10] In China, the roots are also used to treat bronchitis, cold, dysuria, haemoptysis, laryngitis, restlessness, thirst, tuberculosis, stress, as an aphrodisiac, promoting fertility and memory, also as a sialogogue, to treat cancer and frequently included in polyherbal prescriptions for diabetes mellitus. The Indochinese uses the rhizomes to treat fever and inflammation, as a febrifuge, galactagogue, and also for intestinal, kidney and liver ailments.[3]

Pharmacological Activities: Antiarrhythmic,[11] Anti-inflammatory,[12,13] Antithrombotic,[14,15] Immunomodulatory,[16] Immunostimulatory,[17,18] Cardioprotective[19,20] and Chemoprotective.[21]

Dosage: The daily dose of tuberous roots of 6 to 20 g in the form of a decoction, pills or syrup consumed as an expectorant or antitussive.[22]

Adverse Reactions: No information as yet.

Toxicity: LD_{50} value of *O. japonicus* in mice was more than 2 g/kg intraperitoneally.[12]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

55. *Peltophorum pterocarpum* Backer ex K. Heyne (Leguminosae)

Jemerlang Laut, Yellow Flame, Yellow Flamboyant

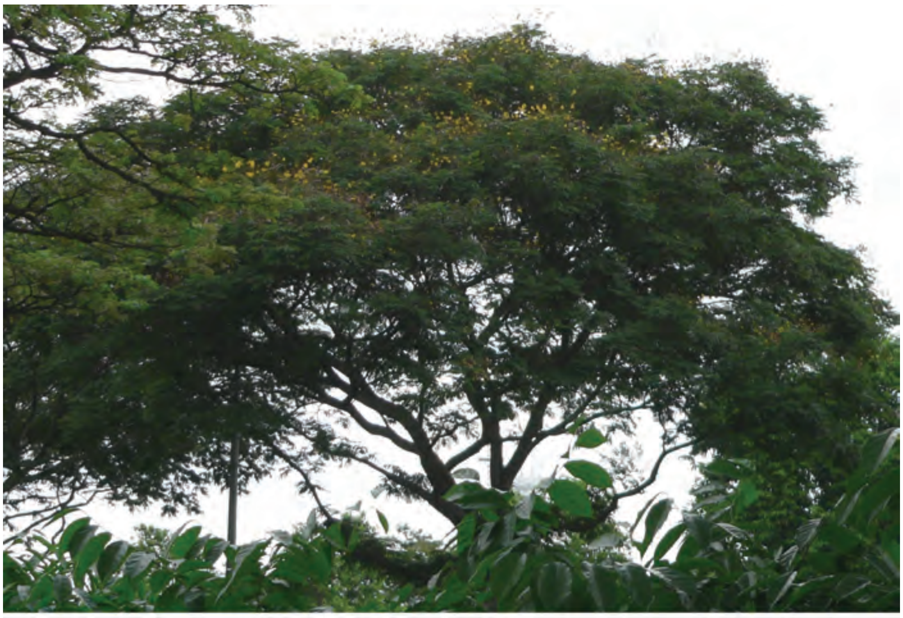

Peltophorum pterocarpum tree

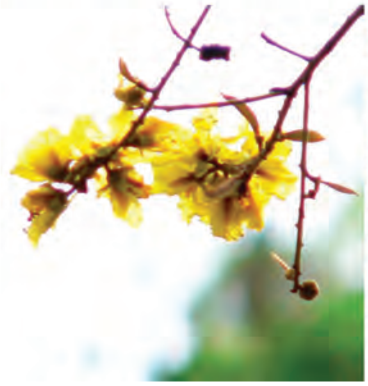

Peltophorum pterocarpum flowers

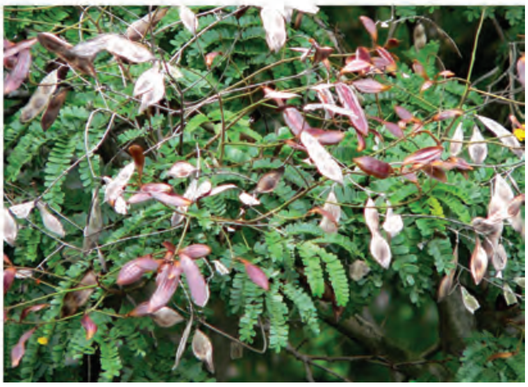

Fruits of *Peltophorum pterocarpum*

Description: *Peltophorum pterocarpum* Backer ex K. Heyne is a large tree with dome-shaped crown that can grow up to 24 m tall. The main rachis are 15–30 cm long and the pinnae are 6–20 paired with each pinna having 20–30 pairs of oblong leaflets. Flowers are yellow and in large bunches. The tree bears pods which are oblong, 5–10 cm long, flat and thin.[1–3]

Origin: Native to Malaysia, Ceylon, the Andamans and North Australia.[3]

Phytoconstituents: Rhamnetin, hirusitidin, bergenin and others.[4]

Traditional Medicinal Uses: The bark extract is used internally to cure dysentery and externally as a lotion to treat sprains, muscular aches, ulcers, and as an eye lotion, gargle and tooth powder.[2]

Pharmacological Activities: Antibacterial[5–7] and Antifungal.[7]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

56. *Persicaria hydropiper* L. (Polygonaceae)

Water Pepper, Laksa Plant

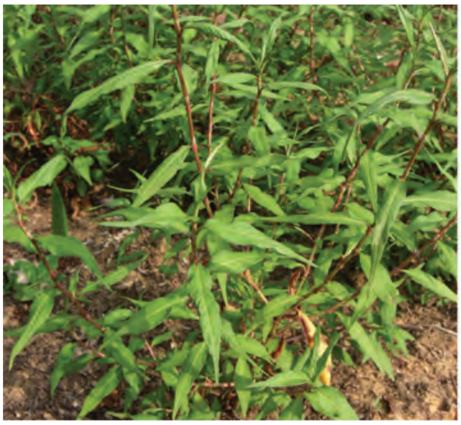

Persicaria hydropiper herbs

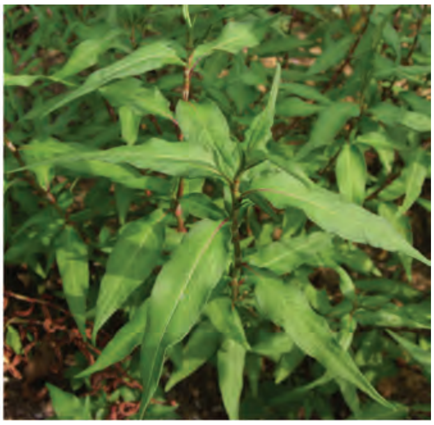

Persicaria hydropiper top view

Description: *Persicaria hydropiper* L. is an erect herb that can grow up to 30–50 cm tall. Leaves are oblong-lanceolate, short petioled, tapering to a pointed end, ciliate on the under surface. Flowers are white, in sparse, thin and with hanging false ears. Fruit is black, nut-like and has a flat and domed side.[1]

Origin: Native to Northern Africa, tropical and temperate Asia, Australia and Europe.[2]

Phytoconstituents: Rhamnazin, hydropiperoside, polygoidal, warburganal, isopolygodial, isodrimeninol, drimenol, confertifolin and others.[1,3–5]

Traditional Medicinal Uses: The liquid extract of the plant is used as a contraceptive and a haemostatic.[6] The plant is also used alone or with other herbs decocted for diarrhoea, dyspepsia, dysentery, enteritis, diuretic, expelling worms, heat stroke, itching skin, haemorrhage, jaundice and cancer as well.[7] In folk medicine, it is used internally for uterine bleeding, menstrual bleeding, bleeding of haemorrhoids, gastrointestinal bleeding, rheumatic pain, as a diuretic, for bladder and kidney disease, and gout. It is used externally for poorly healing wounds, sprains and contusions.[1] The leaves are pounded and applied to skin diseases, for uterine disorders,[7] while its seeds are used as carminative, diuretic and stimulant.[7]

Pharmacological Activities: Analgesic,[8] Anthelmintic,[9] Antifertility,[10,11] Antifungal,[12–14] Antineoplastic,[15] Antioxidant,[12] Antimutagenic[16] and Insect repellent.[17,18]

Dosage: A tea prepared by pouring 0.25 L of hot water over 1 heaped teaspoon of the dried plant extract and strained after 10 min is to be drunk 3 times a day.[1] Homeopathically, it is used to treat varicose veins, 5 drops, 1 tablet or 10 globules are to be taken every 30 to 60 min in acute cases, or 1 to 3 times daily for chronical cases.[1]

Adverse Reactions: No known side effects with therapeutic dosages. Larger amounts can cause gastroenteritis. Skin irritation may occur if applied externally.[1]

Toxicity: Toxic to animals[19–21] and the LD_{50} for its chloroform leaf extract was 758.58 mg/kg in male albino mice.[22]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

57. *Phyllanthus amarus* Schum. & Thonn.

(Euphorbiaceae)

Pick-a-back, Carry Me Seed, Ye Xia Zhu

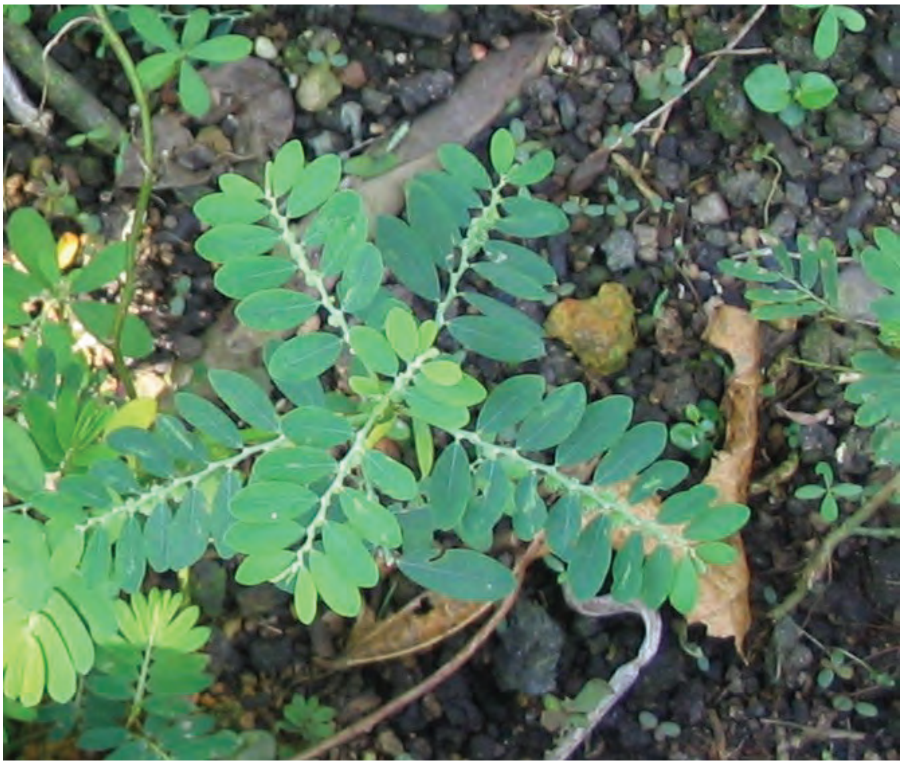

Phyllanthus amarus herbs

Description: *Phyllanthus amarus* Schum. & Thonn. is a small herbaceous annual plant that can grow up to 60 cm tall. Leaves are simple, alternate, green and stipulate. The blade is 3–8 mm by 2–4.5 mm and oblong-elliptic. The fruits are green, depressed globose in shape, 3-lobed and smooth. Both the flowers and fruits are borne under the branches.[1,2]

Origin: Native to Mexico and South America.[3]

Phytoconstituents: Phyllanthusin D, geraniin, corilagin, elaeocarpusin, amariin, amariinic acid, amarosterol-A and B, phyllantin, hypophyllantin and others.[4–11]

Traditional Medicinal Uses: The aerial part of the plant is used for various conditions. In Chinese medicine, the plant is made into a tea to cure kidney problems, venereal diseases, stones in the kidneys and bladder. The Malays use it to increase menstrual flow, reduce fever and cure colic. It is used by the Indians as a fish poison.[5] Indians also use the plant as liver tonic to treat liver ailments, ascites, jaundice, diarrhoea, dysentery, intermittent fever, conditions of the urogenital tract, eye disease, scabies, ulcers and wounds.[12] In Vietnam, it is used to induce sweating, and increase menstrual flow. It is also prescribed for toothache, muscle spasms and gonorrhoea. It is considered a diuretic, colic remedy and abortifacient in Southeast Asia.[5] It is also commonly used in Benin, Africa, as folk medicine against malaria.[13]

Pharmacological Activities: Analgesic,[14] Antibacterial,[15,16] Antidiarrhoeal,[17] Antifertility,[18] Antifungal,[19] Anti-inflammatory,[20–24] Antineoplastic,[25–29] Antioxidant,[30–32] Antiplasmodial,[13,33] Antiviral,[34–39] Diuretics,[40] Hepatoprotective,[41–43] Hypoglycaemic,[11,30,44] Inhibition of gastric lesion,[22] Antimutagenic,[27,45] Insecticidal[46] and Radioprotective.[47]

Dosage: A decoction may be prepared with 10 plants in 1 L of water.[12]

Adverse Reactions: No known side effects with therapeutic dosages.[12]

Toxicity: Non-toxic to mice at a dose of 100 mg/kg body weight[18] but at doses of 400 mg/kg, 800 mg/kg and 1000 mg/kg body weight in rats.[48]

Contraindications: No information as yet.

Drug-Herb Interactions: An alcoholic extract of *P. amarus* was found to inhibit cytochrome P450 enzymes both *in vivo* as well as *in vitro*.[48]

58. *Piper nigrum* L. (Piperaceae)

Pepper, Lada

Leaves and fruits of *Piper nigrum*

Piper nigrum herbs

Description: *Piper nigrum* L. is a perennial woody vine with aerial roots at the stem nodes. Leaves are shiny, broad, alternate and heart-shaped. Flowers are small and white. The fruits are globular berries, about 5 mm in diameter, green when unripe but turn bright red upon maturity and each bears a single seed.[1]

Origin: Native to Southeast Asia.[2]

Phytoconstituents: Piperine, sabinene, nigramides A–S, pipertipine, pipercitine; pellitorine, guineensine, piperettine, pipericine, dipiperamides A–C, pipnoohine, pipyahyine and others.[3–10]

Traditional Medicinal Uses: The plant is used in many Asian countries as a stimulant, for the treatment of colic, rheumatism, headache, diarrhoea, dysentery, cholera, menstrual pains, removing excessive gas and increasing the flow of urine.[11] It is also used in folk medicine for stomach disorders and digestive problems, neuralgia and scabies. In Ayurveda, it is used for arthritis, asthma, fever, cough, catarrh, dysentery, dyspepsia, flatulence, haemorrhoids, urethral discharge and skin damage. In Chinese medicine, it is used for vomiting, diarrhoea and gastric symptoms. Homeopathically, it is used for irritation of mucous membrane and galactorrhoea.[3] A heavy dose of pepper with wild bamboo shoots is said to cause abortion.[4] In Assam, a method of birth control include *Cissampelos pareira* in combination with *Piper nigrum*, root of *Mimosa pudica* and *Hibiscus rosa-sinensis*.[12] The fruits are used to remove excessive gas in system, increase flow of urine, treat colic, rheumatism, headache, diarrhoea, dysentery, cholera, and menstrual pains.[13] White pepper is used for cholera, malaria, stomachache while black pepper is used for abdominal fullness, adenitis, cancer, cholera, cold, colic, diarrhoea, dysentery, dysmenorrhoea, dysuria, furuncles, headache, gravel, nausea, poisoning due to fish, mushrooms or shellfish.[4]

Pharmacological Activities: Antibacterial,[14–16] Antidiabetic,[17] Antifungal,[18] Anti-inflammatory,[19] Antineoplastic,[20] Antioxidant,[21–26] Gastroprotective,[27] Hepatoprotective,[28,29] Hypolipidaemic,[30,31] Antimutagenic,[32] Antithyroidal[33] and Insecticidal.[5,6,9,34–38]

Dosage: A single dose ranges from 300–600 mg of the berries is taken internally for stomach disorder.[3] Homeopathic recommendations for irritation of mucous membranes and galactorrhoea are 5–10 drops, 1 tablet or 5–10 globules 1–3 times daily.[3] For the treatment of haemorrhoids, 5–15 whole peppercorns are recommended to be taken.[39] For congestion, cold, head cold, chicken soup with black pepper can be taken.[40] The average daily dose of the berries is stated to be 1 to 3 g as a decoction, powder or pills, for the treatment of dyspepsia, vomiting, diarrhoea and colic resulting from cold.[41]

Adverse Reactions: No known side effects with appropriate therapeutic dosages.[3] Hypersensitivity may occur in certain people. The fruit can be safely consumed with proper usage.[42] However, when taken in large amounts in children, apnoea can occur.[43]

Toxicity: *P. nigrum* did not cause any adverse effects when tested with rats at doses 5 to 20 times normal human intake.[44]

Contraindications: Not to be taken during pregnancy and lactation or by children and people who are allergic to it.[43]

Drug-Herb Interactions: *P. nigrum* extracts inhibits the activity of cytochrome P450 enzymes.[8,45–49] Piperine enhanced the bioavailability of beta lactam antibiotics, amoxycillin trihydrate and cefotaxime sodium significantly in rats.[50] Piperine potentiated pentobarbitone sleeping time in a dose dependant manner, with peak effect at 30 min.[19] Piperine enhanced the bioavailability of phenytoin significantly, possibly by increasing the absorption.[51] Piperine from black peppers increased the AUC of phenytoin, propranolol and theophylline in healthy volunteers and plasma concentrations of rifampicin in patients with pulmonary tuberculosis.[52] Avoid concurrent use with drugs metabolised by cytochrome P450 enzymes.[43] Avoid concurrent use with anticoagulant agents as well.[53]

59. *Piper sarmentosum* Roxb. (Piperaceae)

Wild Pepper, Kadok, Sirih Tanah

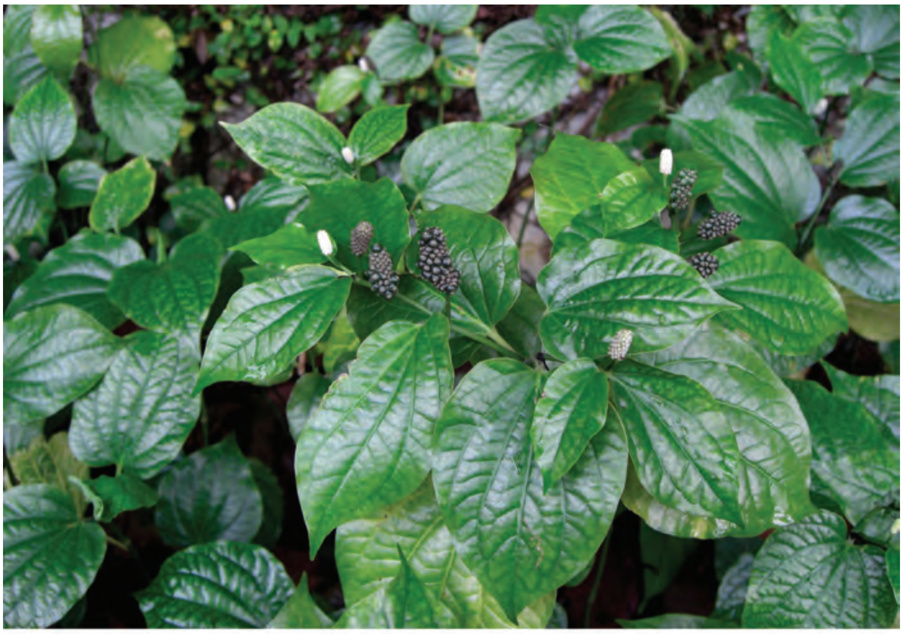

Piper sarmentosum herb

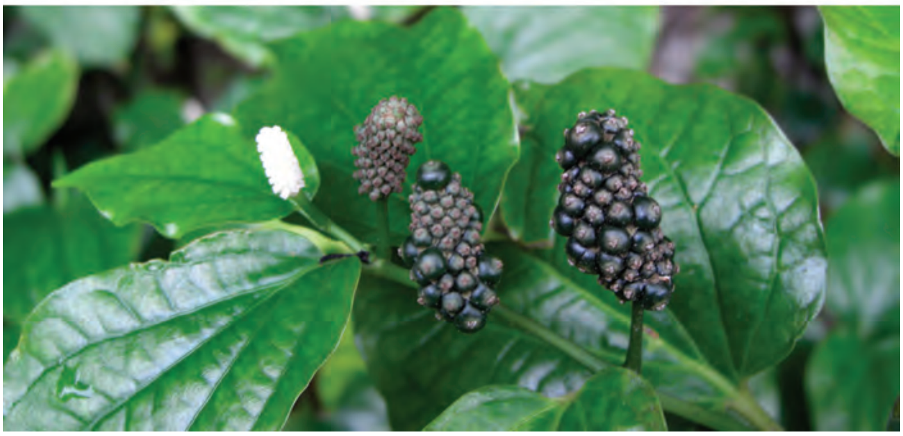

Close up of *Piper sarmentosum* flowers and fruits

Description: *Piper sarmentosum* Roxb. is a climbing herb that can grow up to 10 m long with long runners that can develop into plantlets. Leaves are alternate, simple, 7–14 cm by 6–13 cm, heart-shaped and young leaves have a waxy surface. Flowers are bisexual or unisexual, in terminal or leaf opposite spikes. Fruit is small, dry, with several rounded bulges. Plant has a characteristic pungent odour.[1–5]

Origin: Native to Cambodia, India, Indonesia, Laos, Malaysia, Philippines and Vietnam.[5]

Phytoconstituents: Sarmentosine, sarmentine, (+)-sesamin, horsfieldin, brachystamide B, sarmentamide A, B, and C, (+)-asarinin, methyl piperate and others.[6–11]

Traditional Medicinal Uses: The whole plant can be used to treat fever and aids digestion.[3] The fruit is used as an expectorant[6] while the roots are used to treat toothache, fungal dermatitis on the feet, coughing, asthma and pleurisy.[6,12]

Pharmacological Activities: Antibacterial,[8,10,13] Antineoplastic,[6] Antiprotozoal,[10,14,15] Antipsychotic,[16] Hypoglycaemic,[17] Antifungal,[10] Insecticidal,[18,19] Antituberculosis,[11] Antimalarial[14] and Antioxidant.[20]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

60. *Plantago major* L. (Plantaginaceae)

Common Plantain, Whiteman's Foot, Daun Sejumbok

Plantago major herbs

Description: *Plantago major* L. is a small perennial herb. Leaves are nearly all basal, exstipulate, lanceolate to ovate, 5–20 cm long and rosette. Flowers are small, white, in dense spike-like inflorescence. Sepals are broadly elliptic, oblong to rounded obtuse or subacute and corolla are greenish or yellowish, with four lobed and imbricate. Seeds are dull black and endospermous.[1,2]

Origin: It is found in Europe, Northern and Central Asia, and introduced all over the world.[2]

Phytoconstituents: Aucubin, catalpol, scutellarein, nepetin, chlorogenic acid, neochlorogenic acid, hispidulin, homoplantaginin, nepitrin, ursolic acid and others.[3,4]

Traditional Medicinal Uses: The Greeks and Romans used it as an astringent, to heal wounds, asthma, fever and eye disorders.[5] In Brazil, it has been used to treat skin ulceration (cutaneous leishmaniasis) caused by *Leishmania braziliensis*.[6] *P. major* has been used in Turkey in the treatment of ulcers by taking the powdered dried leaves together with honey daily before breakfast.[7] Infusion of the leaf has been taken for diarrhoea, ulcers, bloody urine, digestive disorders, and excess mucous discharge. The American Indian groups make use of a poultice of the leaves for pain, swelling, cuts, wounds, sores, infections, blisters, insect bites, snakebites and haemorrhoids.[5] Its seeds are used to induce sweating, increase flow of urine, treat diarrhoea, dysentery, rheumatism, malaria, asthma, kidney problems, bladder diseases, gonorrhoea and piles.[3] Its roots are used to treat fever, respiratory infections and constipation.[5] The Commision E approved the internal use of plantain for catarrhs of the respiratory tract and inflammatory alterations of the oral and pharyngeal mucosa while its external application is approved for inflammatory reactions of the skin.[8]

Pharmacological Activities: Analgesic,[9] Antibacterial,[10,11] Antidiarrhoeal,[12] Anti-inflammatory,[4,13–15] Anticancer,[10,16–18] Antioxidant,[19] Antiprotozoal,[20,21] Antiviral,[22] Immunomodulatory,[23,24] Immunostimulatory,[25,26] Proliferative,[10] Antiulcerogenic,[6] Antimutagenic,[27,28] Uterotonic[29] and Wound healing.[30]

Dosage: A daily dose of 8 to 16 g of the whole plant or seeds in the form of a decoction or extract is used to treat oedema, dysuria, haematuria, persistent cough, bronchitis and ophthalmia.[31] Approximately 2 to 4 ml of the fluid extract taken orally three times daily serves for general well being.[32] As a rinse or gargle, 1.4 g of cut herb is immersed in 150 ml of cold water for 1 to 2 hours.[8] For internal use, 1.4 g of herb is immersed in 150 ml of boiled water for 10–15 min, drunk as infusion, for 3 to 4 times daily.[8]

Adverse Reactions: Nausea, vomiting, diarrhoea, anorexia, bloating, hypersensitivity and dermatitis may arise. Life threatening anaphylaxis may occur in more serious cases.[32]

Toxicity: The 70% ethanol extract was found to be toxic to shrimps[33] but *P. major* possesses a low toxicity in rats after oral and intraperitoneal administration.[34]

Contraindications: Should not be used during pregnancy and lactation. Should not be used in persons with intestinal obstruction or those who developed hypersensitivity to plantain.[32]

Drug-Herb Interactions: Decreases the effects of carbamazepine and enhances the effects of cardiac glycosides, *β*-blockers and calcium channel blockers.[32]

61. *Punica granatum* L. (Punicaceae)

Pomegranate, Delima

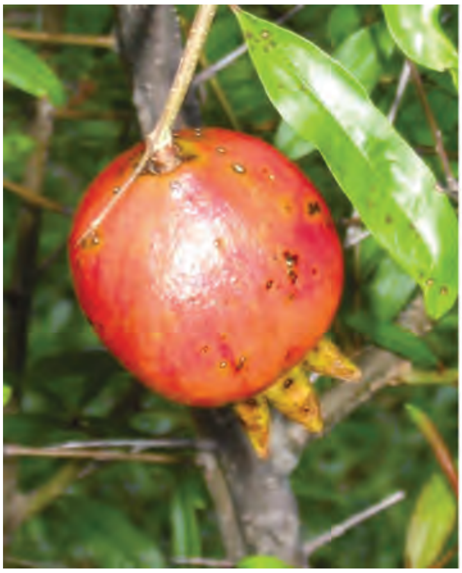

Punica granatum fruit

Punica granatum tree

Description: *Punica granatum* L. is a small woody shrub that can grow up to 3–4.5 m tall. Leaves are opposite, oblong-lanceolate, 3–8 cm and branches are spiny. Flowers are large, trumpet-shaped, and bright orange-red in colour. Fruit is globose, with waxy surface, tough leathery skin, turns deep pink or red upon maturity, and contains numerous seeds with fleshy covering.[1]

Origin: Native to temperate and tropical Asia.[2]

Phytoconstituents: Punicafolin, punicalagin, friedelin, betulic acids, estrone, estradiol, piperidine, pomegranatate, pseudopelletierine and others.[3–7]

Traditional Medicinal Uses: The plant is used in folk medicine to expel worms, for diarrhoea, dysentery, as an abortifacient, astringent, for haemorrhoids (when used externally) and as a gargle for sore throat. In Ayurveda it is used for diarrhoea, dysentery, vomiting, and eye pain. Used in Chinese medicine for chronic diarrhoea and dysentery, it is also used to treat blood in stools, worm infestation, and anal prolapse. Homeopathically, it is used for gastrointestinal disturbances.[5] The leaves are used to relieve itch. The pericarp is used to treat diarrhoea, rectocele and to expel pinworms.[8] The dried pericarp is decocted with other herbs for colic, colitis, diarrhoea, dysentery, leucorrhoea, menorrhagia, oxyuriasis, paralysis and rectocele,[3] to treat piles, yellowish discharge from the vagina, sore throat, bad breath and nose bleed.[8] The stem bark is used as an emmenagogue and for dysentery.[3]

Pharmacological Activities: Anthelmintic,[3,9,10] Antibacterial,[11–21] Anticancer/Antineoplastic,[22–36] Antidiabetic/Hypoglycaemic,[37–41] Antidiarrhoeal,[42] Antifertility,[43] Antifungal,[44,45] Anti-inflammatory,[46,47] Antimalarial,[48] Antioxidant,[34,48–62] Antiviral,[63–65] Gastroprotective,[66,67] Hepatoprotective,[68] Hypolipidaemic,[41,69] Immunomodulatory,[70,71] Neuroprotective,[72–75] Antiatherogenic,[76–78] Wound healing,[79,80] Larvicidal[81] and Molluscicidal.[82,83]

Dosage: Daily dose for the treatment of tapeworm is one part pericarp, root, or stem bark boiled with 5 parts of water. Bark juice extract is recommended for tapeworm at a single dose of 20 g. Otherwise, 250 parts powdered bark are boiled in 1500 parts water for 30 minutes.[5] Doses of 4–5 g of powdered dried flower are used in haematuria, haemorrhoids, dysentery, chronic diarrhoea, and bronchitis. Either 1.5–3 g of root and bark powder, 100–200 ml of bark decoction (for children, 28–56 ml) is used for anthelmintic purposes.[84] 7 g flower in 300 ml water is prescribed for inflamed mouth and throat; 5–12 g root or stem bark boiled in 240 cm^3 water until 1/3 has evaporated and taken on a 3 hourly basis on empty stomach 2 hours after taking 40 ml castor oil.[85] A daily dose of 20 to 50 g of dried root bark or stem bark as a decoction is used in the treatment of taeniasis (tapeworm infection). Daily dose of 15 to 20 g fruit rind in the form of a decoction is taken to treat dysentery and diarrhoea.[86]

Adverse Reactions: No known side effects with appropriate therapeutic dosages. In some cases, gastrointestinal disturbances may occur due to high tannin content.[5] Pomegranate fruits have rarely been reported to cause immediate hypersensitivity.[87]

Toxicity: The bark[5,9,88,89] and peel[90,91] are toxic.

Contraindications: Contraindicated in the elderly, children and pregnant women.[92] Pericarp is contraindicated in diarrhoea. Should not be taken with fats or oils when it is used for killing parasites.[88] Should not be used by patients with hepatic diseases or asthma, and who are hypersensitive to pomegranate.[91]

Drug-Herb Interactions: Inhibition of cytochrome P450 enzymes,[93,94] interaction with carbamazepine[95] and interaction with tolbutamide.[96]

62. *Rhodomyrtus tomentosa* (Ait.) Hassk
(Myrtaceae)

Rose Myrtle

Rhodomyrtus tomentosa plants

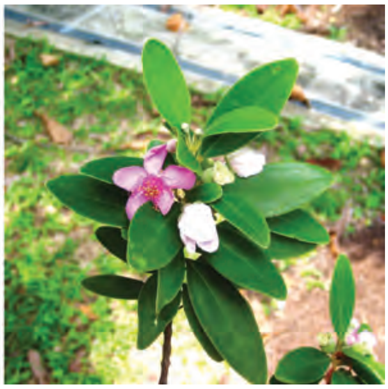

Flowers of *Rhodomyrtus tomentosa*

Description: *Rhodomyrtus tomentosa* (Ait.) Hassk is a bushy shrub with woolly young parts. Leaves are simple, oblong, 4–10 cm, opposites with three main longitudinal veins. Flowers are large, showy and pink, 3–4 cm wide, solitary and auxillary. Berries are oblong, 1–1.25 cm long, dark purple, fleshy and aromatic.[1–3]

Origin: Native to tropical and temperate Asia.[4]

Phytoconstituents: Rhodomyrtone, casuariin, castalagin, friedelin and others.[5–11]

Traditional Medicinal Uses: The leaves are used by the Chinese as a painkiller. They are also used in Indonesia to heal wounds.[1] In Malaysia, the plant, including the roots, are decocted to treat diarrhoea and heartburn.[1,5] The buds and young leaves are used for treatment of colic, diarrhoea, dysentery, abscesses, furunculosis and haemorrhage. Concentrated decoction of leaves is used as a disinfectant for wounds, impetigo and abscesses.[12] The seeds are used as a digestive tonic and to treat snake bites.[1]

Pharmacological Activities: Antibacterial.[6]

Dosage: Daily dose of 10 to 30 g of fresh buds or young leaves in the form of an extracted juice or dried for use as a powder or in a decoction is used for the treatment of colic, diarrhoea, dysentery, abscesses, furunculosis and haemorrhage.[12]

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

63. *Rhoeo spathacea* (Sw.) Stearn (Commelinaceae)

Purple-leaved Spider Wort, Moses-in-the-Cradle, Oyster Plant

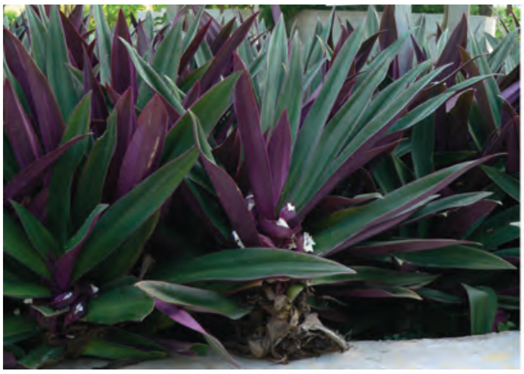

Rhoeo spathacea plants

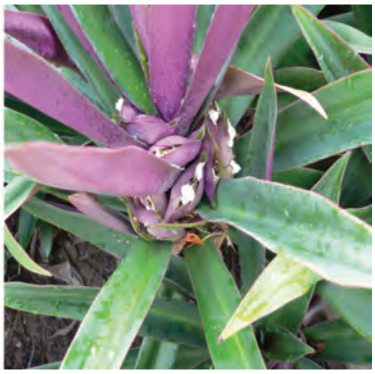

Flowers of *Rhoeo spathacea*

Description: *Rhoeo spathacea* (Sw.) Stearn is a short-stemmed herb with underground rhizome that can grow up to 90 cm tall. Leaves pointed upwards, fleshy, hard, long, narrow, 15–25 cm by 3–5 cm, pointed, dark green on the adaxial surface and purple on abaxial surface. Flowers are small, white, and concealed within a boat-shaped envelope of two bracts.[1,2]

Origin: Native to Central America and West Indies, widely cultivated as ornamental plants in most tropical countries.[2]

Phytoconstituents: No information as yet.

Traditional Medicinal Uses: They are used to treat intestinal bleeding, blood in the sputum and dysentery.[3]

Pharmacological Activities: Anti-inflammatory,[4] Antiadrenergic[5] and Uterostimulatory.[6]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

64. *Ricinus communis* L. (Euphorbiaceae)

Castor Oil Plant, Castor Bean

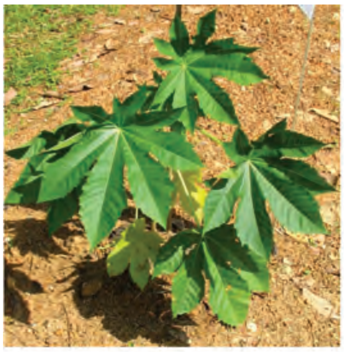

Ricinus communis leaf

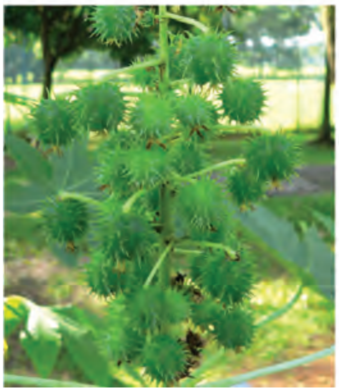

Fruits of *Ricinus communis*

Ricinus communis plant

Description: *Ricinus communis* L. is an erect herb, growing up to 3.6 m high, having pinkish succulent stem and large alternate palmate leaves that are green or reddish brown. Leaves are lobed, consisting of 6–8 radiating leaflets with serrated edges and prominent central veins. Flowers are green, pink or red and inconspicuous, with no petals. The fruits are capsular, with three lobes, prickly and green, containing three seeds.[1]

Origin: Native to Africa, naturalised throughout tropics and subtropics.[2]

Phytoconstituents: Ricin, ricinoleic acid, ricinine, p-coumaric acid, ferulic acid, o-coumaric acids, syringic acid, cinnamic acids, stigmasterol, fucosterol and others.[3–8]

Traditional Medicinal Uses: Its leaf poultice is applied to boils and sores in India; to treat headaches and fever in Hawaii.[9] The leaves and roots are used in a decoction for anal prolapse, arthritis, constipation, facial palsy, lymphadenopathy, strabismus, uteral prolapse, cough, and also as a discutient and expectorant. The heated leaves are applied to gout and swellings as well.[5] The leaves and oil are used for dermatological purposes in Nigeria.[10] Its seeds are used to treat abscesses and skin eruptions, deafness, headache, skin problems, bleeding, constipation, boils, piles and to promote labour.[4] They are rubbed on the temple for headache, powdered for abscesses, boils, and carbuncles. The plant is also used for dogbite, scrofula and several skin infections. The Chinese rub the oil on the body for skin ailments. The seeds are crushed and made into a pulp and rubbed into the palms for palsy, introduced into the urethra in stricture and rubbed on the soles of feet of parturient woman to hasten birth or expulsion of the placenta. The seeds are also used to treat colic, diarrhoea, dysentery, enteritis, acute constipation, for itching, ringworms, warts, dandruff, hair loss and haemorrhoids. It is also used as a laxative before X-ray examination of bowels.[3,5] Midwives sometimes use castor oil to induce labour.[11]

Pharmacological Activities: Antifertility,[12–18] Antioxidant,[19] Antipsychotic,[20] Antiviral,[21] Anti-inflammatory,[19] Convulsant,[20] Hepatoprotective,[22] Filaricidal,[23] Haemaglutination[6] and Insecticidal.[24]

Dosage: For internal use in the treatment of constipation or to expel worms, at least five 2 g capsules or ten 1 g capsules are recommended.[3] Externally, a paste made of ground seeds may be applied to affected skin areas 2 times daily, up to 15 days.[3] Homeopathically, 5 drops, 1 tablet or 10 globules are recommended every 30–60 min for acute treatment or 1–3 times daily for chronic cases.[3] Standard adult dose is reported to be 5.0–20.0 ml/dose, not exceeding 60.0 ml per day. Standard child dose is reported to be 4.0–12.0 ml/dose.[11] The seed oil is recommended for use as a laxative in a dose of 2 to 5 ml and as a purgative in a dose of 20 to 30 ml. 15 seeds may be crushed and applied to the plantar surface of the feet for the induction of labour at term and placenta delivery in cases of placenta retention.[25] The woman's feet are advised to be thoroughly washed after delivery of foetus and placenta.[25] For constipation, it is reported that 15 to 60 ml of castor oil may be taken orally a day.[26]

Adverse Reactions: Gastric irritation, nausea, vomiting, colic, and severe diarrhoea may occur. Long term use may result in loss of electrolytes, especially potassium, and causes intestinal motility inhibition.[3]

Toxicity: Seeds are highly poisonous due to the toxic lectins, principally the albumin and ricin.[27] The seeds are toxic to both animals and humans.[3,12,26–35]

Contraindications: Not to be taken by children under 12 years of age, patients with acute inflammatory intestinal diseases, inflammatory bowel disease and appendicitis.[3] contraindicated in pregnancy, intestinal obstruction and abdominal pain.[3,11] Not to be given with potentially toxic oil-soluble anthelmintics.[36]

Drug-Herb Interactions: Effect of cardioactive steroids may be increased.[3] To prevent decreased absorption of castor, it is advised not to take within one hour of antacids, other drugs and milk.[37]

65. *Ruta graveolens* L. (Rutaceae)

Herb of Grace, Common Rue

Ruta graveolens herb

Ruta graveolens leaves

Description: *Ruta graveolens* L. is a glabrous herb with stem that can grow up to 14–45 cm. Lower leaves are more or less long-petiolate with ultimate segments 2–9 mm wide, lanceolate to narrowly oblong. Inflorescence is rather lax; pedicels are as long as or longer than the capsule; bracts are lanceolate, leaf-like. Sepals are lanceolate and acute. Petals are oblong-ovate, denticulate and undulate. Capsule is glabrous; segments somewhat narrowed above to an obtuse apex.[1]

Origin: Native to Europe.[2]

Phytoconstituents: Rutoside, rutaverine, arborinine, rutin, elemol, pregeijerene, geijerene, furocoumarins, bergapten, xanthotoxin, fagarine, graveolinine and others.[3–5]

Traditional Medicinal Uses: It is frequently used to treat worm and parasitic infection.[6] It has been commonly used for the treatment of psoriasis and vitiligo due to the psoralens and methoxypsoralens present.[7] It is also used to relieve muscle spasms, as carminative, emmenagogue, haemostat, uteronic, vermifuge, to treat hepatitis, dyspepsia, diarrhoea, bug bite, cancer, cold, fever, snakebite, earache, toothache and as an antidote especially in malarial poisoning.[3,4,8] It is also used as an abortifacient to terminate pregnancy.[9] The plant has been used for pain relief in Mexico.[10]

Pharmacological Activities: Analgesic,[11] Antibacterial,[12] Anticancer,[13–17] Antifertility,[18–20] Anti-inflammatory,[5,21] Antimicrobial,[22–24] Antiprotozoal[25] and Cytotoxic.[24]

Dosage: For the treatment of earache, the oil is poured on cotton and inserted into the affected ear. For oral administration, 1 capsule or 0.5 to 1 teaspoon of extract is taken three times daily with meals. The topical cream may be applied to the affected area when necessary.[26] For delayed menstruation, 2 cups of infusion per day is used.[8]

Adverse Reactions: Hypotension, hypersensitivity, rash, erythema and blisters may occur when applied topically.[26] Therapeutic dosages could bring about melancholic moods, sleep disorders, tiredness, dizziness and spasms.[27] Misuse as an abortive during pregnancy can lead to vomiting, epigastric pain, kidney damage, depression, sleep disorders, feelings of vertigo, delirium, fainting, tremor, spasm and sometimes may end up with fatal outcome.[8]

Toxicity: Toxic to humans[27–33] and animals.[34]

Contraindications: Should not be used during pregnancy.[9] Should not be used during lactation, in children and in person with hypersensitivity to rue. Patients with heart diseases should use it with caution.[26] Contraindicated in patients with poor renal function.[9]

Drug-Herb Interactions: Concurrent use of cardiac glycosides (e.g. digoxin, digitoxin) with rue may cause increased inotropic effects. Concurrent use of antihypertensives with rue may cause increased vasodilation.[26]

66. *Saccharum officinarum* L. (Gramineae)

Sugarcane, Tebu

Saccharum officinarum plants

Saccharum officinarum stems

Description: *Saccharum officinarum* L. is a perennial plant that forms tall clumps that can grow up to 6 m tall. Stems are greenish, yellowish or dark purplish and juicy. Leaf blades are broadly linear, glabrous and 80–150 cm by 4–6 cm. Inflorescence is a large silky panicle.[1,2]

Origin: Native to Southeast Asia and Pacific Islands, widely cultivated elsewhere.[2]

Phytoconstituents: Octacosanol, policosanol, orientin, tricin-7-*O*-glycoside, palmitic acid, oleic acid and linolenic acid and others.[3–9]

Traditional Medicinal Uses: Consumption of beets (*Beta vulgaris*) combined with molasses from *S. officinarum*, is used by Dominican healers to shrink fibroids or to "strengthen and fortify the uterus after the fibroid had been drained from the body".[10] The cane juice promotes expulsion of phlegm from the respiratory passages, stimulates gastric activities, treats wounds, ulcers and boils.[11]

Pharmacological Activities: Analgesic,[3] Anticancer,[6] Anti-inflammatory,[3,9,12] Antiosteoporotic,[5] Antioxidant,[6,13,14] Antiplatelet,[15–17] Antithrombotic,[18] Hypocholesterolaemic,[19–24] Immunomodulatory,[25,26] Immunostimulatory,[27] Antiatherogenic[28] and Myocardial protective.[4,17]

Dosage: Short-term studies have shown the efficacy and tolerability of policosanol at 10 mg/day on hypercholesterolaemia in obese patients with Type 2 diabetes.[19,20]

Adverse Reactions: Pollen extract of *S. officinarum* showed strong sensitising potential which induces allergy.[29]

Toxicity: Studies on the toxicity of higher aliphatic primary acids (D003) and higher aliphatic primary alcohols (policosanol) isolated from *S. officinarum* did not show any toxic effects.[30–33]

Contraindications: No information as yet.

Drug-Herb Interactions: Pretreatment with high doses of policosanol significantly increased propranolol-induced hypotensive effects, while the effects of nifedipine remained unchanged. Policosanol does not antagonise the hypotensive effect of β-blockers but it can increase the hypotensive effect of beta-blockers without modifying cardiac frequency.[34] Pretreatment with single doses of policosanol significantly increased the nitroprusside-induced hypotensive effect.[35] Warfarin alone and the combination of policosanol and warfarin induced a moderate, but significant prolongation of the bleeding time.[36]

[**Authors' Note:** Freshly squeezed sugarcane juice is commonly consumed and sugar is commercially obtained from sugarcane.]

67. *Sauropus androgynus* (L.) Merr.

(Euphorbiaceae)

Sweet Leaf Bush, Cekup Manis, Daun Katuk

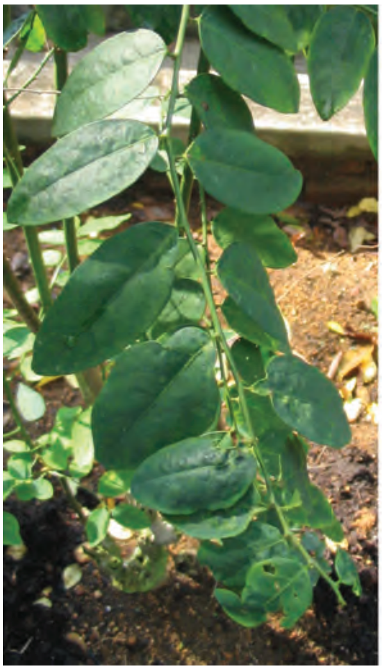

Leaves of *Sauropus androgynus*

Sauropus androgynus plant

Description: *Sauropus androgynus* (L.) Merr. is a small shrub that can grow up to 3 m tall. Leaf blade is ovate-lanceolate, oblong-lanceolate, or lanceolate, 3–10 cm by 1.5–3.5 cm and thinly papery. Flowers are small and borne in axillary clusters. Mature fruit is about 5 cm wide. Young branches and leaves are used as vegetable.[1,2]

Origin: Native to China and tropical Asia.[3]

Phytoconstituents: Sauroposide, (–)-isolariciresinol, corchoionoside C and others.[4–10]

Traditional Medicinal Uses: The fresh leaves or roots possess uterotonic activity and are used for the treatment of retained placenta. A mouth wash made from the juice of the fresh leaves and honey, and applied to the tongue and gums cures thrush of the tongue in infants. The leaves are used for erythema, measles and dysuria. Its roots serve as a diuretic and relieve congestion.[4] The root decoction is used for fever and urinary bladder complaints.[11]

Pharmacological Activities: Anthelmintic.[12]

Dosage: For the treatment of retained placenta, a dose of 40 g, in the form of an extracted juice is administered in 2 subdoses at 10 minutes interval.[4]

Adverse Reactions: Excessive consumption of the leaf reportedly caused dizziness, drowsiness and constipation.[13]

Toxicity: Reported to cause bronchiolitis obliterans,[14–16] ischaemic bronchi necrosis,[17] irreversible obstructive ventilatory defect,[18,19] breathing difficulties[20,21] and is cytotoxic.[22]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

68. *Sesbania grandiflora* Pers. (Leguminosae)

Scarlet Wisteria Tree, Red Wisteria, Daun Turi

Sesbania grandiflora leaves

Fruit of *Sesbania grandiflora*

Sesbania grandiflora tree

Description: *Sesbania grandiflora* Pers. is a tree that can grow to 8–10 m in height. The compound leaves are about 30 cm long with 12 to 20 pairs of rounded, narrow, oblong leaflets, 3–4 cm by 1 cm. Flowers are 5–10 cm by 3 cm, in pale pink, red, purple or white. The pods are 25–50 cm, slender, and cylindrical with many light brown to red brown seeds.[1]

Origin: Native to Malesia and cultivated in the tropics.[2]

Phytoconstituents: Grandiflorol, (+)-leucocyanidin, oleanolic acid, lutein, beta-carotene, violaxanthin, neoxanthin, zeaxanthin and others.[3–9]

Traditional Medicinal Uses: In the Philippines, the plant is used for its hypotensive properties.[10] It is used in Indian folk medicine for the treatment of liver disorders.[11] The juice of the leaves and flowers are popularly used for nasal catarrh and headache when taken as snuff. Various leaf preparations are used to treat epileptic fits. Applied externally for treatment of leprous eruptions. A poultice of the leaves is used for bruises. The leaf juice is mixed with honey for congenital bronchitis or cold in babies.[12]

Pharmacological Activities: Antibacterial,[13] Anticonvulsant,[14] Anti-inflammatory,[15] Anxiolytic,[14] Depressant,[15] Diuretic,[15] Hepatoprotective,[11] Hypoglycaemic,[16] Hypotensive[15] and Haemolytic.[17,18]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

69. *Solanum nigrum* L. (Solanaceae)

Black Nightshade, Terong Meranti, Poison Berry

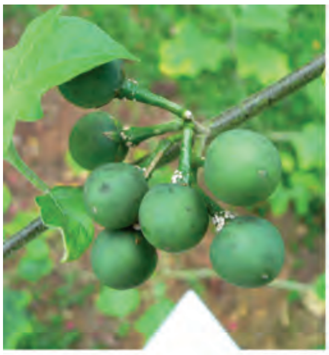

Fruits of *Solanum nigrum*

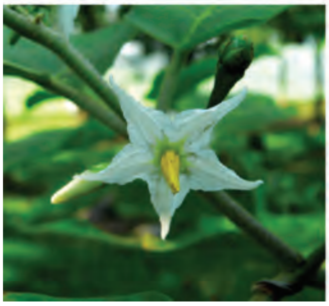

Solanum nigrum flower

Solanum nigrum shrub

Description: *Solanum nigrum* L. is a small herb, up to 1.5 m tall. Leaves are ovate, ovate-oblong, glabrous, hairy, 1–16 cm by 0.25–12 cm. Inflorescence of 2–10 in an extra-axillary cluster, with white or purple corolla and yellow central protrusion. Fruit is globose, black in colour but is green when immature, 0.5 cm in diameter, with many seeds.[1]

Origin: Native to Southwest Asia, Europe, India and Japan.[2]

Phytoconstituents: Solanidine, α-,β-,γ-chaconine, desgalactotigonin, α-,β-solamargine, diosgenin, solanadiol, α-,β-,γ-solanines, soladulcidine, solanocapsine, α-,β-solansodamine, solasodine, α-solasonine, tigogenin, tomatidenol, uttronins A and B, uttrosides A and B, solanigroside A–H and others.[3–9]

Traditional Medicinal Uses: The stem, leaves and roots are used as a decoction for wounds, tumours and cancerous growths, sores and as an astringent.[3,5] They are also used as a condiment, stimulant, tonic, for treatment of piles, dysentery, abdominal pain, inflammation of bladder, relief of asthma, bronchitis, coughs, eye ailments, itch, psoriasis, skin diseases, eczema, ulcer, relief of cramps, rheumatism, neuralgia and expulsion of excess fluids. The roots are used as an expectorant.[5] The plant has yielded medicines for sore throats, coughs and digestive problems. It has also been used as an agricultural insecticide.[10] Europeans in Africa used the plant to treat convulsions. It is used by the Africans for treating headache, ulcers and as a sedative. The whole plant is used for the treatment of dermatitis, inflammation, heavy female discharge, diarrhoea and dysentery.[11] It is also used as a diuretic and febrifuge. Whole plant is decocted for abscesses, cancer of the cervix, inflammation, leucorrhoea and open sores. Young shoots are consumed as virility tonic for men and to treat dysmenorrhoea in females.[3] In Indochina, the leaves are used as purgative and high blood pressure lowering agents while the fruits are used as laxatives.[12]

Pharmacological Activities: Antibacterial,[13] Anticancer/antineoplastic,[8,14–33] Antiulcerogenic,[36,37] Antinociceptive,[15] Anti-inflammatory,[15] Antioxidant,[34,35] Antiviral,[38] Depressant,[39] Hepatoprotective,[40–43] Hypolipidaemic,[44] Antimutagenic,[45] Enzyme modulation,[46] Larvicidal,[47] Molluscicidal[47–49] and Parasiticidal.[50]

Dosage: 10 drops of extract is taken internally 2 to 3 times a day or 5 to 10 g of tincture may be taken daily for gastric irritation, cramps and whooping cough. For external use as a rinse of moist compress, it is boiled in 1 L water for 10 minutes before usage for psoriasis, haemorrhoids, abscesses, eczema and bruising.[4]

Adverse Reactions: No known side effects with appropriate therapeutic dosage.[4]

Toxicity: Harmful to rats.[51] It is toxic to cattle as it can cause acute nitrate toxicity which leads to death in cattle. In chronic cases, decreased milk yield, abortion, impaired vitamin A and iodine nutrition can occur. The proposed LD_{50} for nitrate toxicity is 160–224 mg/NO_3/kg for cattle.[52] Overdoses can lead to headache, queasiness and vomiting, due to high alkaloid content. Mydriasis may also occur, although rare.[4,52] Solanine, in doses of 200–400 mg, may cause gastroenterosis, tachycardia, dyspnea, vertigo, sleepiness, lethargy, twitching of the extremities and cramps. It is also teratogenic.[52]

Contraindications: No information as yet.

Drug-herb Interactions: No information as yet.

70. *Swietenia macrophylla* King (Meliaceae)

Honduras Mahogany, Broad-leaved Mahogany

A fruiting *Swietenia macrophylla* tree

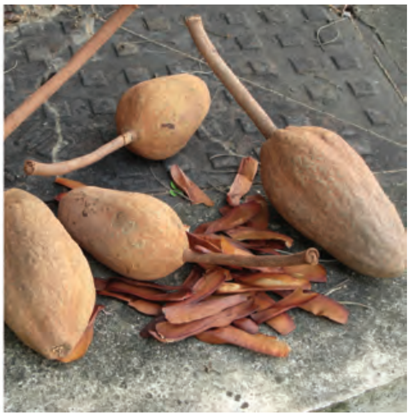

Seeds and fruits of *Swietenia macrophylla*

Description: *Swietenia macrophylla* King is an evergreen tree, up to 30–35 m tall. Bark is grey and smooth when young, turning dark brown, ridged and flaky when old. Leaves are up to 35–50 cm long, alternate, glabrous with 4–6 pairs of leaflets. Each leaflet is 9–18 cm long. Flowers are small and white; and the fruit is dehiscent, usually 5-lobed capsule, erect, 12–15 cm long, grayish brown, smooth or minutely verrucose. The seed is woody, glossy and possesses wing-like structure at the base that aids its dispersion by wind.[1]

Origin: Native to South America, cultivated in the Asia-Pacific and the Pacific for its quality wood.[1]

Phytoconstituents: Swietenine, swietenolide, andirobin, khayasin T, swietemahonins E–G, swietenine acetate, swietenolide tiglate and others.[2–7]

Traditional Medicinal Uses: The seeds of *Swietenia macrophylla* are traditionally used in several indigenous systems of medicine for the treatment of various ailments such as hypertension, diabetes and malaria.[9] The local folks of Malaysia believe that the seeds are capable of “curing” hypertension and diabetes. The seeds are usually consumed raw by chewing.[8] A decoction of seeds of *Swietenia macrophylla* is reported to treat malaria in Indonesia.[9] Among the Amazonian Bolivian ethnic groups, the seeds are traditionally used to induce abortion by drinking a decoction of the seeds and to heal wounds and various ailments of the skin via external application of the mashed seeds.[10]

Pharmacological Activities: Antimalarial,[11] Antihypertensive[8] and Antidiarrhoeal.[12]

Dosage: No information as yet.

Adverse Reactions: No information as yet.

Toxicity: Uterine haemorrhage.[13]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

71. *Terminalia catappa* L. (Combretaceae)

Indian Almond, Katapang

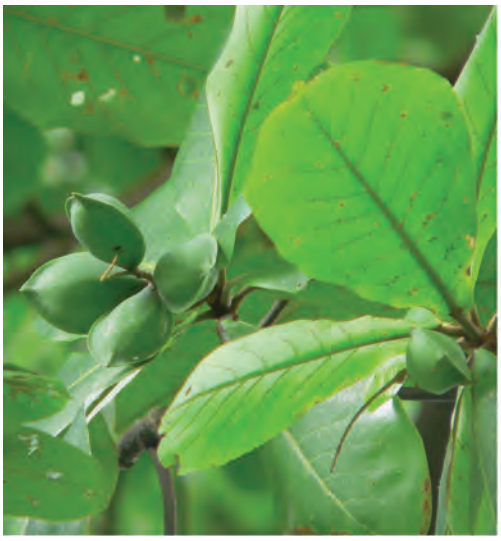

Fruits of *Terminalia catappa*

Terminalia catappa tree

Description: *Terminalia catappa* L. is a tall tree, up to 25 m tall. Branches are horizontally whorled, giving it a pagoda shape. Leaves are shiny, obovate, 10–25 cm long, tapering to a short thick petiole. Leaves are yellow that turn red before shedding. Flowers are small and white. Fruits have smooth outer coat, 3–6 cm long, flattened edges, with a pointed end. Pericarp is fibrous and fleshy.[1–3]

Origin: Native to tropical and temperate Asia, Australasia, the Pacific and Madagascar.[4]

Phytoconstituents: Catappanin A, chebulagic acid, 1-desgalloylleugeniin, geraniin, granatin B, punicalagin, punicalin, tercatain, terflavins A & B, tergallagin, euginic acid and others.[2,5–13]

Traditional Medicinal Uses: *Terminalia catappa* has been used to treat dysentery in a number of Southeast Asian countries. In Indonesia, the leaves are used as a dressing for swollen rheumatic joints while in the Philippines, they are used to expel worms.[2] In Karkar Island, New Guinea, juice from the squeezed leaves is applied to sores and the sap from the white stem pith is squeezed and drunk to relieve cough. In Nasingalatu, Papua New Guinea, the flower is crushed, mixed with water and drunk to induce sterility. In New Britain, the old yellow leaves are crushed in water and drunk to sooth sore throat. In Bougainville, the leaves are heated and placed on pimples and the bark is applied to sores. In Tonga, the juice from pounded leaves and bark is applied to mouth sore. In Irian Jaya, the leaves are applied to wounds and burns while in Somoa, it is used to cure cough and sore throat.[14] The fruits are used after childbirth to strengthen the back. An enema made from the crushed fruit mixed with *Trigonella foenum-graecum*, animal fat and warm water is administered to the new mother after childbirth.[15] The leaves are used for the treatment of scabies and skin diseases while the juice is used to treat headache and colic.[13,14] The bark is used as a diuretic, cardiotonic and for dysentery.[16] The leaves of this plant have also been used as a folk medicine for treating hepatitis in India and Philippines.[5]

Pharmacological Activities: Antimicrobial,[17–19] Anticancer/Antineoplastic,[20,21] Anti-inflammatory,[5,22] Hypoglycaemic,[8] Antioxidant,[6,10,11,23,24] Hepatoprotective,[7,11,25–29] Antiviral,[30] Chemopreventive,[31] Aphrodisiac[32] and Antimutagenic.[33]

Dosage: 2 tablespoons of a decoction (few leaves in 200 ml water) is given every 2 hours to stop diarrhoea.[34]

Adverse Reactions: No information as yet.

Toxicity: Preliminary oral LD_{50} doses of petroleum ether, methanol and aqueous extracts of *T. catappa* in mice were found to be 343, 195, and 210 mg/kg respectively.[8] Rats fed on *T. catappa* diet maintained their body weight but suffered from stomach, small intestine and pancreas hypertrophies as well as spleen atrophy.[35] Larger doses enhanced liver damage.[36]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

72. *Thevetia peruviana* (Pers.) K. Schum.

(Apocynaceae)

Yellow Oleander, Trumpet Flower

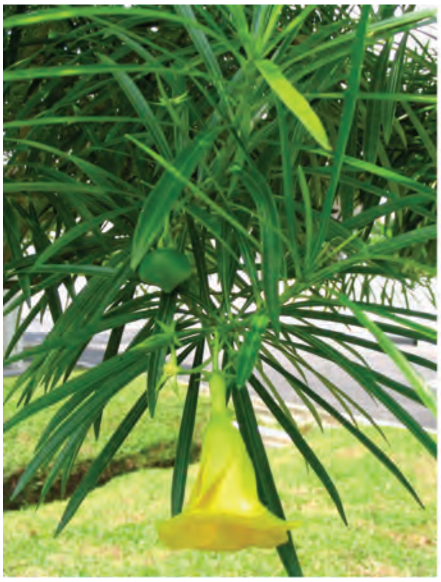

Thevetia peruviana flower and fruit

Thevetia peruviana shrub

Description: *Thevetia peruviana* (Pers.) K. Schum. is a shrub, up to 6 m tall. All parts contain highly poisonous milky latex. Leaves are simple, few, exstipulate and spirally arranged. Blade is linear, 7–13 cm by 0.5–1 cm and glossy. Flowers are large, yellow, 5 cm across, gathered in few flowered terminal cymes. Fruits are green, shiny, globose, 4–5.5 cm across with 4 or less poisonous seeds.[1]

Origin: Native to Central and South America.[2]

Phytoconstituents: Thevetins A and B, thevetosides, acetylperuvoside, epiperuviol, perusitin, theveneriin, thevebioside, thevefolin, pervianoside I–III and others.[3–6]

Traditional Medicinal Uses: Used as an abortifacient, to treat congestive heart failure, malaria, leprosy, indigestion, ringworm, venereal disease and even as a suicide instrument.[7] Used in India as an astringent to the bowel, useful in urethral discharge, worms, skin diseases, wounds, piles, eye problems and itch. Used in continental Europe and is considered particularly useful in mild myocardial insufficiency and digitalis intolerance.[5] Its bark is used as an emetic, febrifuge, insecticidal, poison and for reviving patients with heart failure.[4]

Pharmacological Activities: Antiarrhythmic,[5] Antifungal,[8] Hepatotoxicity and Nephrotoxicity,[5] Larvicidal, [9] Molluscicidal,[10] and Cardiotonic.[5]

Dosage: No information as yet.

Adverse Reactions: Vomiting, dizziness, and cardiac dysrhythmias such as conduction block affecting the sinus and AV nodes may occur.[11]

Toxicity: Toxic to humans,[3,11–16] fishes[17–19] and animals.[5,20] Ingestion of half to fifteen seeds may cause a varying degree of vomiting, palpitation, arrhythmia including sinus bradycardia and could also lead to death.[21] The kernels of about 10 fruits may be fatal to adults while the kernel of one fruit may be fatal to children.[22]

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

[**Authors' Note:** The purified glucoside thevetin, extracted from the seeds, is prescribed as a cardiotonic drug in a 0.1% solution orally, in a dose of 1 to 2 ml daily, or in 1 mg/2 ml ampoules parenterally. 1 to 2 ampoules have been given daily for its cardiotonic property.[23] A dose of 2.4 mg followed by average maintenance dose of 600 μg has been shown to be effective in 22 patients with congestive heart failure.[5] The herb has been administered as a thevetin tablet of 0.25 mg/tablet. A maintenance dose consists of 0.25 mg/day. For injection administration, an injection ampoule is prepared from 0.25 mg/ml diluted in a 5% glucose solution.[24] However, note that thevetin is not conventionally used in clinical practice and its dosage is not found in pharmacopoeial monographs.]

73. *Tinospora crispa* (L.) Diels (Menispermaceae)

Akar Putarwali, Batang Wali

Tinospora crispa shrub

Flowers of *Tinospora crispa*

Description: *Tinospora crispa* (L.) Diels is a woody climber with numerous protrusions on the stem. Leaves are oblong-ovate, cordate, 8–9 cm by 7–8 cm and tapering to a pointed end. Flowers are small, with 6 petals, 2 mm in length and 8–27 cm racemes. Male flowers have yellow sepals whereas female flowers have green sepals. Drupelets are red, juicy and 7–8 mm long.[1,2]

Origin: Native to Malesia, Indochina, Indian subcontinent and China.[3]

Phytoconstituents: Boropetol B, borapetoside B, C & F, jatrorhizine, magnoflorine, palmatine, protoberberine, tembolarine, diosmetin, cycloeucalenol, cycloeucalenone and others.[4–8]

Traditional Medicinal Uses: It is used for hypertension, diabetes mellitus,[9] to treat malaria, remedy for diarrhoea and as vermifuge.[6] In Malaysia, *T. crispa* extract is taken orally by Type 2 (non-insulin-dependent) diabetic patients to treat hyperglycaemia.[10]

Pharmacological Activities: Anti-inflammatory,[11] Antioxidant,[7] Antimalarial,[12,13] Antiprotozoal[14] and Hypoglycaemic.[10,15,16]

Dosage: No information as yet.

Adverse Reactions: The plant may result in an increased risk of hepatic dysfunction due to marked elevation of liver enzymes but is reversible upon discontinuation of *T. crispa*.[17]

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: *T. crispa* extract showed 61.3% increase of cytochrome P450 3A4 (CYP3A4) enzyme inhibition after 20 min of preincubation with human liver microsomes.[18]

[**Authors' Note:** In a randomised double blind placebo controlled trial[17] to determine the efficacy of *T. crispa* as an additional treatment in patients with Type 2 diabetes mellitus who did not respond to oral hypoglycaemic drugs and refused insulin injection, 20 patients received *T. crispa* powder in capsule form at a dose of 1 gram three times daily for six months. Twenty patients received a placebo. The results showed no evidence to support the use of *T. crispa* for additional therapy in such patients. Two patients had elevated levels of liver enzymes, which was reversible on discontinuing *T. crispa*. Patients receiving *T. crispa* had significant weight reduction and cholesterol elevation. Note that the study only included patients that did not respond to oral hypoglycaemic therapy.]

74. *Vitex rotundifolia* L. f. (Verbenaceae)

Round Leaf Chastetree, Beach Vitex

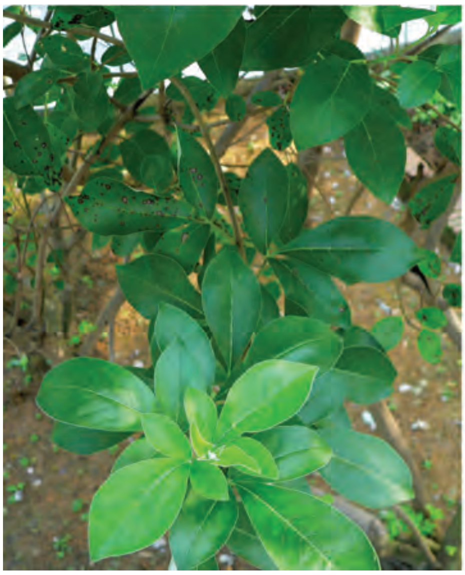

Vitex rotundifolia leaves

Vitex rotundifolia tree

Description: *Vitex rotundifolia* L. f. is an evergreen woody tree, densely covered with short hairs. Leaves are opposite, simple, ovate, broadly oblong-elliptic, 2–5 cm long by 1.5–3 cm wide, rounded or abruptly acute at the base. Inflorescence panicles are at the terminal, densely flowered, 4–7 cm long with purple corolla. Fruits are globose, 5–7 mm.[1]

Origin: Native to Temperate and Tropical Asia, Australasia and Pacific.[2]

Phytoconstituents: Rotundifuran, prerotundifuran, vitexilactone, previtexilactone, vitexicarpin, vitricine, vitetrifolins D–G, vitexifolins A–E, isoambreinolide and others.[1,3–9]

Traditional Medicinal Uses: In Malaysia, various parts of the plants are considered panacea for illnesses ranging from headache to tuberculosis.[10] In China, the plant has been used for the treatment of cancer.[11] A poultice of the leaves is used to treat rheumatism, contusions, swollen testicles and as a discutient in sprains. In Indonesia, leaves have been used in medicinal baths, as a tincture or for intestinal complaints.[10] In Papua New Guinea, sap from crushed heated leaves is diluted with water and drunk to relieve headaches. The fruits are used to expel worms and in Vietnam, a decoction of dried fruits has been used to treat colds, headache, watery eyes and mastitis.[10] In Thailand, fruits have been used for asthmatic cough and haemorrhoids.[10] Infusion of the boiled roots is regarded as diaphoretic and diuretic, drunk for fever, after child-birth and for liver diseases.[10]

Pharmacological Activities: Analgesic,[12] Antibacterial,[13,14] Antifungal,[15] Anti-inflammatory,[16] Antineoplastic,[11,15,17–20] Antinociceptive,[8] Antioxidant,[6] Antiprotozoal,[21] Hypotensive,[22] Immunomodulatory,[17] Antimutagenic[23] and Insect repellant.[15]

Dosage: No information as yet.

Adverse Reactions: *V. rotundifolia* may trigger various allergic reactions such as sneezing, respiratory problems, dizziness, headache and nausea.[10]

Toxicity: No information as yet.

Contraindications: No information as yet.

Drug-Herb Interactions: No information as yet.

75. *Zingiber officinale* Roscoe (Zingiberaceae)

Common Ginger

Zingiber officinale rhizome

Zingiber officinale plant

Description: *Zingiber officinale* Roscoe is an herbaceous plant that grows up to 1.2 m high and with an underground rhizome. The stem grows above ground and leaves are narrow, long, lanceolate, with distinct venation pattern and pointed apex. Flowers are white or yellowish-green, streaked with purple and fragrant.[1]

Origin: Originate from tropical Asia, widely cultivated in the tropics.[2]

Phytoconstituents: Gingerol, zingiberene, farnesene, camphene, neral, nerol, 1,8-cineole, geranial, geraniol, geranyl acetate and others.[3–7]

Traditional Medicinal Uses: Ginger is the folk remedy for anaemia, nephritis, tuberculosis, and antidote to *Arisaema* and *Pinellia***.**[8] Sialogogue when chewed, causes sneezing when inhaled and rubefacient when applied externally. Antidotal to mushroom poisoning, ginger peel is used for opacity of the cornea. The juice is used as a digestive stimulant and local application in ecchymoses.[8] Underground stem is used to treat stomach upset, nausea, vomiting, nose bleeds, rheumatism, coughs, blood in stools, to improve digestion, expel intestinal gas, and stimulate appetite.[9] The rhizomes are used to treat bleeding, chest congestion, cholera, cold, diarrhoea, dropsy, dysmenorrhoea, nausea, stomachache, and also for baldness, cancer, rheumatism, snakebite and toothache.[8] It is also used as postpartum protective medicine, treatment for dysentery, treatment for congestion of the liver, complaints with the urino-genital system/female reproduction system and sinus.[10] Besides that, it is used to alleviate nausea, as a carminative, circulatory stimulant and to treat inflammation and bacterial infection.[3] The Commision E approved the internal use of ginger for dyspepsia and prevention of motion sickness.[11] The British Herbal Compendium indicates ginger for atonic dyspepsia, colic, vomiting of pregnancy, anorexia, bronchitis and rheumatic complaints.[12] European Scientific Cooperative on Phytotherapy (ESCOP) indicates its use for prophylaxis of the nausea and vomiting of motion sickness and to alleviate nausea after minor surgical procedures.[13]

Pharmacological Activities: Analgesic,[14–16] Anthelmintic,[17] Antiarthritic,[18] Anticancer,[19–33] Antidiabetic,[34,35] Antidiarrhoeal,[36] Antiemetic,[37–53] Antihyperlipidaemic,[54] Antihypertensive,[55] Anti-inflammatory,[14,15,56–66] Antimicrobial,[67–78] Antioxidant,[6,7,79–82] Antiplatelet,[83–87] Antispasmodic,[88] Antiulcer,[89,90] Antiviral,[91–93] Anxiolytic,[94] Hepatoprotective,[93–97] Hypocholesterolaemic,[57,98,99] Hypoglycaemic,[14,100,101] Hypolipidaemic,[102] Hypotensive,[103] Immunomodulatory,[104–106] Neuroprotective,[107] Insect repellent[69,108] and Radioprotective.[109–111]

Dosage: A tea is prepared by pouring boiling water over 0.5 to 1 g of the coarsely powdered ginger for 5 min and passing through a tea strainer, taken to prevent vomitting.[112]

Adverse Reactions: Fresh rhizome can be safely consumed with proper usage.[113] Contact dermatitis of the fingertips has been reported in sensitive patients.[114]

Toxicity: It is nontoxic when tested in rats[115] but overdose may cause cardiac arrhythmia and CNS depression.[116]

Contraindications: Consult physician before using ginger preparations in patients with blood coagulation disorders, taking anticoagulant drugs or with gallstones. Avoid dried rhizomes during pregnancy.[113]

Drug-Herb Interactions: Interacts with anticoagulants such as heparin, warfarin, drugs used in chemotherapy and ticlopidine. Ginger taken prior to 8-MOP (treatment for patients undergoing photopheresis) may substantially reduce nausea caused by 8-MOP.[117] Ginger appears to increase the risk of bleeding in patients taking warfarin.[118] However, ginger at recommended doses does not significantly affect clotting status, the pharmacokinetics or pharmacodynamics of warfarin in healthy subjects.[119] Ginger also significantly decreased the oral bioavailability of cyclosporine.[120]

[**Authors' Note:** Ginger is widely eaten as a food ingredient and used in many different cultures as traditional medicine. Ongoing scientific research has shown diverse pharmacological activities.]

This page intentionally left blank

Appendix

List of Native and Non-native Plants

Native	Non-native	
Ardisia elliptica	*Abrus precatorius*	*Juniperus chinensis*
Asplenium nidus	*Adiantum capillus-veneris*	*Kaempferia galanga*
Barringtonia asiatica	*Allamanda cathartica*	*Lantana camara*
Barringtonia racemosa	*Aloe vera*	*Lonicera japonica*
Calophyllum inophyllum	*Andrographis paniculata*	*Mangifera indica*
Cerbera odollam	*Areca catechu*	*Manihot esculenta*
Crinum asiaticum	*Aster tataricus*	*Mimosa pudica*
Eurycoma longifolia	*Azadirachta indica*	*Mirabilis jalapa*
Hibiscus tiliaceus	*Bauhinia purpurea*	*Morinda citrifolia*
Ipomoea pes-caprae	*Bixa orellana*	*Nelumbo nucifera*
Melaleuca cajuputi	*Cananga odorata*	*Nerium oleander*
Melastoma malabathricum	*Capsicum annuum*	*Ophiopogon japonicus*
Nephelium lappaceum	*Cassia fistula*	*Persicaria hydropiper*
Peltophorum pterocarpum	*Catharanthus roseus*	*Phyllanthus amarus*
Piper sarmentosum	*Celosia argentea*	*Piper nigrum*
Rhodomyrtus tomentosa	*Centella asiatica*	*Plantago major*
Terminalia catappa	*Cissus quadrangularis*	*Punica granatum*
	Cocos nucifera	*Rhoeo spathacea*
	Coix lacryma-jobi	*Ricinus communis*
	Cymbopogon citratus	*Ruta graveolens*
	Dolichos lablab	*Saccharum officinarum*
	Elephantopus scaber	*Sauropus androgynus*
	Euphorbia hirta	*Sesbania grandiflora*
	Hibiscus mutabilis	*Solanum nigrum*
	Hibiscus rosa-sinensis	*Swietenia macrophylla*
	Impatiens balsamina	*Thevetia peruviana*
	Imperata cylindrica	*Tinospora crispa*
	Ixora chinensis	*Vitex rotundifolia*
	Jatropha curcas	*Zingiber officinale*

Note: Native plants listed above refer to those that are indigenous to Singapore while the non-native plants are those that originate elsewhere and have been subsequently introduced to Singapore. Many of the non-native plants are now growing well in Singapore.

List of Plants by Scientific Name

Scientific Name	Common Name
Abrus precatorius	Rosary Pea, Indian Licorice, Precatory Bean
Adiantum capillus-veneris	Black Maidenhair Fern, Southern Maidenhair Fern, Venus Hair Fern
Allamanda cathartica	Allamanda, Common Allamanda, Golden Trumpet
Aloe vera	Aloe, Lidah Buaya
Andrographis paniculata	Hempedu Bumi, Sambiloto, Chuan Xin Lian
Ardisia elliptica	Mata Pelanduk/Ayam, Sea-Shore Ardisia, Shoebutton Ardisia
Areca catechu	Betel Nut Palm, Areca Nut, Pinang
Asplenium nidus	Bird's Nest Fern
Aster tataricus	Tatarian Aster, Tatarian Daisy
Azadirachta indica	Neem
Barringtonia asiatica	Beach Barringtonia, Fish-Killer Tree, Putat Laut
Barringtonia racemosa	Putat Kampong, Samundrapandu
Bauhinia purpurea	Butterfly Tree
Bixa orellana	Annatto, Lipstick Tree
Calophyllum inophyllum	Indian Laurel, Penaga Laut, Borneo Mahogany
Cananga odorata	Kenanga, Ylang-Ylang
Capsicum annuum	Chilli, Red Pepper
Cassia fistula	Golden Shower Tree, Indian Laburnum, Purging Cassia
Catharanthus roseus	Madagascar Periwinkle, Rose Periwinkle
Celosia argentea	Feather Cockscomb, Red Spinach
Centella asiatica	Indian Pennywort, Asiatic Pennywort
Cerbera odollam	Pong Pong Tree, Indian Suicide Tree, Sea Mango
Cissus quadrangularis	Grape Leaf, Veld Grape
Cocos nucifera	Coconut Palm, Kelapa
Coix lacryma-jobi	Job's Tears, Adlay
Crinum asiaticum	Crinum Lily, Spider Lily, Bawang Tanah
Cymbopogon citratus	Lemon Grass
Dolichos lablab	Lablab Bean, Hyacinth Bean
Elephantopus scaber	Elephant's Foot, Tutup Bumi, Tapak Sulaiman
Euphorbia hirta	Asthma Weed
Eurycoma longifolia	Tongkat Ali, Ali's Umbrella, Pasak Bumi

Scientific Name	Common Name
Hibiscus mutabilis	Cotton Rose, Chinese Rose
Hibiscus rosa-sinensis	Hawaiian Hibiscus, China Rose, Bunga Raya
Hibiscus tiliaceus	Linden Hibiscus, Sea Hibiscus, Mahoe
Impatiens balsamina	Balsam Plant
Imperata cylindrica	Lalang, Alang-Alang, Speargrass
Ipomoea pes-caprae	Beach Morning Glory, Goat's Foot Creeper, Bayhops
Ixora chinensis	Chinese Ixora
Jatropha curcas	Barbados Nut, Physic-Nut
Juniperus chinensis	Chinese Juniper
Kaempferia galanga	Galangal, Sand Ginger, Kencur
Lantana camara	Common Lantana, Bunga Tahi Ayam, Wild Sage
Lonicera japonica	Japanese Honeysuckle, Jin Yin Hua
Mangifera indica	Mango, Mangga
Manihot esculenta	Tapioca, Cassava
Melaleuca cajuputi	Gelam, Paper-Bark Tree, Kayu Puteh
Melastoma malabathricum	Sendudok, Singapore Rhododendron
Mimosa pudica	Touch-Me-Not, Sensitive Plant, Rumput Simalu
Mirabilis jalapa	Four O'clock Flower, Bunga Pukul Empat
Morinda citrifolia	Mengkudu, Indian Mulberry, Noni
Nelumbo nucifera	Sacred Lotus, East Indian Lotus, Oriental Lotus
Nephelium lappaceum	Rambutan, Hairy Lychee
Nerium oleander	Oleander
Ophiopogon japonicus	Dwarf Lilyturf, Mondo Grass, Mai Men Dong
Peltophorum pterocarpum	Jemerlang Laut, Yellow Flame, Yellow Flamboyant
Persicaria hydropiper	Water Pepper, Laksa Plant
Phyllanthus amarus	Pick-A-Back, Carry Me Seed, Ye Xia Zhu
Piper nigrum	Pepper, Lada
Piper sarmentosum	Wild Pepper, Kadok, Sirih Tanah
Plantago major	Common Plantain, Whiteman's Foot, Daun Sejumbok
Punica granatum	Pomegranate, Delima
Rhodomyrtus tomentosa	Rose Myrtle
Rhoeo spathacea	Purple-Leaved Spider Wort, Moses-In-The-Cradle, Oyster Plant
Ricinus communis	Castor Oil Plant, Castor Bean
Ruta graveolens	Herb of Grace, Common Rue
Saccharum officinarum	Sugarcane, Tebu

Scientific Name	Common Name
Sauropus androgynus	Sweet Leaf Bush, Cekup Manis, Daun Katuk
Sesbania grandiflora	Scarlet Wisteria Tree, Red Wisteria, Daun Turi
Solanum nigrum	Black Nightshade, Terong Meranti, Poison Berry
Swietenia macrophylla	Honduras Mahogany, Broad-Leaved Mahogany
Terminalia catappa	Indian Almond, Katapang
Thevetia peruviana	Yellow Oleander, Trumpet Flower
Tinospora crispa	Akar Putarwali, Batang Wali
Vitex rotundifolia	Round Leaf Chastetree, Beach Vitex
Zingiber officinale	Common Ginger

List of Plants by Common Name

Common Name	Scientific Name
Akar Putarwali	*Tinospora crispa*
Alang-Alang	*Imperata cylindrica*
Ali's Umbrella	*Eurycoma longifolia*
Allamanda	*Allamanda cathartica*
Aloe	*Aloe vera*
Annatto	*Bixa orellana*
Areca Nut	*Areca catechu*
Asiatic Pennywort	*Centella asiatica*
Asthma Weed	*Euphorbia hirta*
Balsam Plant	*Impatiens balsamina*
Barbados Nut	*Jatropha curcas*
Batang Wali	*Tinospora crispa*
Bawang Tanah	*Crinum asiaticum*
Bayhops	*Ipomoea pes-caprae*
Beach Barringtonia	*Barringtonia asiatica*
Beach Morning Glory	*Ipomoea pes-caprae*
Beach Vitex	*Vitex rotundifolia*
Betel Nut Palm	*Areca catechu*
Bird's Nest Fern	*Asplenium nidus*
Black Maidenhair Fern	*Adiantum capillus-veneris*
Black Nightshade	*Solanum nigrum*
Borneo Mahogany	*Calophyllum inophyllum*
Broad-Leaved Mahogany	*Swietenia macrophylla*
Bunga Pukul Empat	*Mirabilis jalapa*
Bunga Raya	*Hibiscus rosa-sinensis*
Bunga Tahi Ayam	*Lantana camara*
Butterfly Tree	*Bauhinia purpurea*
Carry Me Seed	*Phyllanthus amarus*
Cassava	*Manihot esculenta*
Castor Bean	*Ricinus communis*
Castor Oil Plant	*Ricinus communis*
Cekup Manis	*Sauropus androgynus*
Chilli	*Capsicum annuum*
China Rose	*Hibiscus rosa-sinensis*

Common Name	Scientific Name
Chinese Ixora	*Ixora chinensis*
Chinese Juniper	*Juniperus chinensis*
Chinese Rose	*Hibiscus mutabilis*
Chuan Xin Lian	*Andrographis paniculata*
Coconut Palm	*Cocos nucifera*
Common Allamanda	*Allamanda cathartica*
Common Ginger	*Zingiber officinale*
Common Lantana	*Lantana camara*
Common Plantain	*Plantago major*
Common Rue	*Ruta graveolens*
Cotton Rose	*Hibiscus mutabilis*
Crinum Lily	*Crinum asiaticum*
Daun Katuk	*Sauropus androgynus*
Daun Sejumbok	*Plantago major*
Daun Turi	*Sesbania grandiflora*
Delima	*Punica granatum*
Dwarf Lilyturf	*Ophiopogon japonicus*
East Indian Lotus	*Nelumbo nucifera*
Elephant's Foot	*Elephantopus scaber*
Feather Cockscomb	*Celosia argentea*
Fish-Killer Tree	*Barringtonia asiatica*
Four O'Clock Flower	*Mirabilis jalapa*
Galangal	*Kaempferia galanga*
Gelam	*Melaleuca cajuputi*
Goat's Foot Creeper	*Ipomoea pes-caprae*
Golden Shower Tree	*Cassia fistula*
Golden Trumpet	*Allamanda cathartica*
Grape Leaf	*Cissus quadrangularis*
Hairy Lychee	*Nephelium lappaceum*
Hawaiian Hibiscus	*Hibiscus rosa-sinensis*
Hempedu Bumi	*Andrographis paniculata*
Herb of Grace	*Ruta graveolens*
Honduras Mahogany	*Swietenia macrophylla*
Hyacinth Bean	*Dolichos lablab*
Indian Almond	*Terminalia catappa*
Indian Laburnum	*Cassia fistula*
Indian Laurel	*Calophyllum inophyllum*

Common Name	Scientific Name
Indian Licorice	*Abrus precatorius*
Indian Mulberry	*Morinda citrifolia*
Indian Pennywort	*Centella asiatica*
Indian Suicide Tree	*Cerbera odollam*
Japanese Honeysuckle	*Lonicera japonica*
Jemerlang Laut	*Peltophorum pterocarpum*
Jin Yin Hua	*Lonicera japonica*
Job's Tears	*Coix lacryma-jobi*
Kadok	*Piper sarmentosum*
Katapang	*Terminalia catappa*
Kayu Puteh	*Melaleuca cajuputi*
Kelapa	*Cocos nucifera*
Kenanga	*Cananga odorata*
Kencur	*Kaempferia galanga*
Lablab Bean	*Dolichos lablab*
Lada	*Piper nigrum*
Laksa Plant	*Persicaria hydropiper*
Lalang	*Imperata cylindrica*
Lemon Grass	*Cymbopogon citratus*
Lidah Buaya	*Aloe vera*
Linden Hibiscus	*Hibiscus tiliaceus*
Lipstick Tree	*Bixa orellana*
Madagascar Periwinkle	*Catharanthus roseus*
Mahoe	*Hibiscus tiliaceus*
Mangga	*Mangifera indica*
Mango	*Mangifera indica*
Mata Pelanduk/Ayam	*Ardisia elliptica*
Mengkudu	*Morinda citrifolia*
Mondo Grass	*Ophiopogon japonicus*
Moses-In-The-Cradle	*Rhoeo spathacea*
Neem	*Azadirachta indica*
Noni	*Morinda citrifolia*
Oleander	*Nerium oleander*
Oriental Lotus	*Nelumbo nucifera*
Oyster Plant	*Rhoeo spathacea*
Paper-Bark Tree	*Melaleuca cajuputi*
Pasak Bumi	*Eurycoma longifolia*

Common Name	Scientific Name
Penaga Laut	*Calophyllum inophyllum*
Pepper	*Piper nigrum*
Physic-Nut	*Jatropha curcas*
Pick-A-Back	*Phyllanthus amarus*
Pinang	*Areca catechu*
Poison Berry	*Solanum nigrum*
Pomegranate	*Punica granatum*
Pong Pong Tree	*Cerbera odollam*
Precatory Bean	*Abrus precatorius*
Purging Cassia	*Cassia fistula*
Purple-Leaved Spider Wort	*Rhoeo spathacea*
Putat Kampong	*Barringtonia racemosa*
Putat Laut	*Barringtonia asiatica*
Rambutan	*Nephelium lappaceum*
Red Pepper	*Capsicum annuum*
Red Spinach	*Celosia argentea*
Red Wisteria	*Sesbania grandiflora*
Rosary Pea	*Abrus precatorius*
Rose Myrtle	*Rhodomyrtus tomentosa*
Rose Periwinkle	*Catharanthus roseus*
Round Leaf Chastetree	*Vitex rotundifolia*
Rumput Simalu	*Mimosa pudica*
Sacred Lotus	*Nelumbo nucifera*
Sambiloto	*Andrographis paniculata*
Samundrapandu	*Barringtonia racemosa*
Sand Ginger	*Kaempferia galanga*
Scarlet Wisteria Tree	*Sesbania grandiflora*
Sea Hibiscus	*Hibiscus tiliaceus*
Sea Mango	*Cerbera odollam*
Sea-Shore Ardisia	*Ardisia elliptica*
Sendudok	*Melastoma malabathricum*
Sensitive Plant	*Mimosa pudica*
Shoebutton Ardisia	*Ardisia elliptica*
Singapore Rhododendron	*Melastoma malabathricum*
Sirih Tanah	*Piper sarmentosum*
Southern Maidenhair Fern	*Adiantum capillus-veneris*
Speargrass	*Imperata cylindrica*

Common Name	Scientific Name
Spider Lily	*Crinum asiaticum*
Sugarcane	*Saccharum officinarum*
Sweet Leaf Bush	*Sauropus androgynus*
Tapak Sulaiman	*Elephantopus scaber*
Tapioca	*Manihot esculenta*
Tatarian Aster	*Aster tataricus*
Tatarian Daisy	*Aster tataricus*
Tebu	*Saccharum officinarum*
Terong Meranti	*Solanum nigrum*
Tongkat Ali	*Eurycoma longifolia*
Touch-Me-Not	*Mimosa pudica*
Trumpet Flower	*Thevetia peruviana*
Tutup Bumi	*Elephantopus scaber*
Veld Grape	*Cissus quadrangularis*
Venus Hair Fern	*Adiantum capillus-veneris*
Water Pepper	*Persicaria hydropiper*
Whiteman's Foot	*Plantago major*
Wild Pepper	*Piper sarmentosum*
Wild Sage	*Lantana camara*
Ye Xia Zhu	*Phyllanthus amarus*
Yellow Flamboyant	*Peltophorum pterocarpum*
Yellow Flame	*Peltophorum pterocarpum*
Yellow Oleander	*Thevetia peruviana*
Ylang-Ylang	*Cananga odorata*

List of Abbreviations and Symbols

α	alpha
Al	aluminium
AUC	area under the curve
β	beta
cm^3	cubic centimetre
δ	delta
EC_{50}	effective concentration, i.e. minimum concentration required for 50% of the test subjects to exhibit the desired response
γ	gamma
g	gramme
IgE	immunoglobulin E
kg	kilogramme
L	litre
LD_{50}	lethal dose, i.e. minimum concentration required to kill 50% of the test subjects
mg	milligramme
ml	millimetre
mg/kg	milligrammes per kilogram of body weight
mg/ml	milligrammes per millimetre of fluid
mg/m^2	milligrammes per square metre of body surface area
μg	microgramme
NO_3	nitrate
ppm	parts per million
%	percent
TD_{50}	toxic dose, i.e. minimum concentration required to induce toxicity in 50% of the test subjects

This page intentionally left blank

References

Abrus precatorius L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Nguyen V. D. 1993. Medicinal Plants of Vietnam, Cambodia and Laos. Nguyen Van Duong: Vietnam.
3. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
4. eFloras.org 2008 http://www.efloras.org/florataxon.aspx?flora_id=5&taxon_id=200011844 Date Accessed: 30th March 2008
5. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
6. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
7. Saxena, V. K. and Sharma, D. N. 1999. *Fitoterapia* 70: 328–329.
8. Kennelly, E. J., Cai, L. N., Kim, N. C. and Kinghorn, A. D. 1996. *Phytochemistry* 41: 1381–1383.
9. Tahirov, T. H., Lu, T. H., Liaw, Y. C., Chen, Y. L. and Lin, J. Y. 1995. *Journal of Molecular Biology* 250: 354–367.
10. Wang, J. P., Hsu, M. F., Chang, L. C., Kuo, J. S. and Kuo, S. C. 1995. *European Journal of Pharmacology* 273: 73–81.
11. Lin, J. Y., Lee, T. C. and Tung, T. C. 1978. *International Journal of Peptide and Protein Research* 12: 311–317.
12. Lin, J. Y., Lee, T. C., Hu, S. T. and Tung, T. C. 1981. *Toxicon* 19: 41–51.
13. Bhardwaj, D. K., Bisht, M. S. and Mehta, C. K. 1980. *Phytochemistry* 19: 2040–2041.
14. Ghosal, S. and Dutta, S. K. 1971. *Alkaloids of Abrus Precatorius. Phytochemistry* 10: 195–198.
15. Kim, N. C., Kim, D. S. and Kinghorn, A. D. 2002. *Natural Product Letters* 16: 261–266.
16. Yadava, R. N. and Reddy, V. M. 2002. *Journal of Asian Natural Product Research* 4: 103–107.
17. Limmatvapirat, C., Sirisopanaporn, S. and Kittakoop, P. 2004. *Planta Medica* 70: 276–278.
18. Singh, R. B. and Shelley. 2007. *Journal of Environmental Biology* 28: 461–464.
19. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
20. Zore, G. B., Awad, V., Thakre, A. D., Halde, U. K., Meshram, N. S., Surwase, B. S. and Karuppayil, S. M. 2007. *Natural Product Research* 21: 933–940.
21. Mølgaard, P., Nielsen, S. B., Rasmussen, D. E., Drummond, R. B., Makaza, N. and Andreassen, J. 2001. *Journal of Ethnopharmacology* 74: 257–264.
22. Anam, E. M. 2001. *Phytomedicine* 8: 24–27.
23. Kuo, S. C., Chen, S. C., Chen, L. H., Wu, J. B., Wang, J. P. and Teng, C. M. 1995. *Planta Medica* 61: 307–312.
24. Moriwaki, S. O. H., Bakalova, R., Yasuda, S. and Yamasaki, N. 2004. *Toxicology and Applied Pharmacology* 195: 182–193.
25. Ramnath, V., Kuttan, G. and Kuttan, R. 2002. *Indian Journal of Physiology and Pharmacology* 46: 69–77.
26. Ohba, H., Moriwaki, S., Bakalova, R., Yasuda, S. and Yamasaki, N. 2004. *Toxicology and Applied Pharmacology* 195: 182–193.
27. Ghosh, D. and Maiti, T. K. 2007. *Immunobiology* 212: 589–599.
28. Ghosh, D. and Maiti, T. K. 2007. *Immunobiology* 212: 667–673.
29. Bhutia, S. K., Mallick, S. K., Stevens, S. M., Prokai, L., Vishwanatha, J. K. and Maiti, T. K. 2008. *Toxicology in Vitro* 22: 344–351.
30. Bhutia, S. K., Mallick, S. K., Maiti, S. and Maiti, T. K. 2008. *Chemico-Biological Interactions*, 174: 11–18.

31. Schottelius, J., Marinkelle, C. J. and Gomez-Leiva, M. A. 1986. *Tropical Medicine and Parasitology* 37: 54–58.
32. Ramnath, V., Kuttan, G. and Kuttan, R. 2002. *Indian Journal of Experimental Biology* 40: 910–913.
33. Tripathi, S. and Maiti, T. K. 2003. *International Immunopharmacology* 3: 375–381.
34. Ramnath, V., Kuttan, G. and Kuttan, R. 2006. *Immunopharmacology and Immunotoxicology* 28: 259–268.
35. Tripathi, S. and Maiti, T. K. 2005. *International Journal of Biochemistry and Cell Biology* 37: 451–462.
36. Singh, S. and Singh, D. K. 1999. *Phytotherapy Research* 13: 210–213.
37. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of medicinal herbs. CRC Press: Boca Raton, FL.
38. Fernando, C. 2001. *Anaesthesia.* 56: 1178–1180.
39. Bradberry, S. 2007. *Medicine* 35: 576–577.
40. Sinha, R. 1990. *Journal Ethnopharmacology* 128: 173–181.
41. Sinha, S. and Mathur, R. S. 1990. *Indian Journal of Experimental Biology* 28: 752–756.
42. Ratnasooriya, W. D., Amarasekera, A. S., Perera, N. S. D. and Premakumara, G. A. S. 1991. *Journal of Ethnopharmacology* 33: 85–90.

Adiantum capillus-veneris L.

1. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
2. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
3. Foster, S. and Duke, J. A. 1990. A Field Guide to Medicinal Plants, Eastern and Central North America. Houghton Mifflin Company: New York.
4. Schauenberg, P. and Paris, F. 1990. Guide to Medicinal Plants (translated from French by Maurice Pugh Jones). Keats Publishing Inc.: New Canaan, Connecticut.
5. Bianchini, F. and Corbetta, F. 1977. Health Plants of the World, Atlas of Medicinal Plants. Newsweek Books: New York.
6. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=200003518
   Date Accessed: 30th March 2008
7. Imperato, F. 1982. *Phytochemistry* 21: 2158–2159.
8. Imperato, F. 1982. *Phytochemistry* 21: 2717–2718.
9. Berti, G., Bottari, F. and Marsili, A. 1969. *Tetrahedron* 25: 2939–2947.
10. Nakane, T., Maeda, Y., Ebihara, H., Arai, Y., Masuda, K., Takano, A., Ageta, H., Shiojima, K., Cai, S. Q. and Abdel-Halim, O. B. 2002. *Chemical and Pharmaceutical Bulletin (Tokyo)* 50: 1273–1275.
11. Marino, A., Elberti, M. G. and Cataldo, A. 1989. *Bollettino della Società italiana di biologia sperimentale* 65: 461–463.
12. Singh, M., Singh, N., Khare, P. B. and Rawat, A. K. 2008. *Journal of Ethnopharmacology* 115: 327–329.
13. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.

Allamanda cathartica L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=210000031
   Date Accessed: 30th March 2008
5. Carballeira, N. M. and Cruz, C. 1998. *Phytochemistry* 49: 1253–1256.
6. Coppen, J. J. W. 1983. *Phytochemistry* 22: 179–182.

7. Tiwari, T. N., Pandey, V. B. and Dubey, N. K. 2002. *Phytotherapy Research* 16: 393–394.
8. Agrawal, P. K. and Sharma, M. P. 1983. *Indian Journal of Pharmaceutical Sciences* 45: 246–248.
9. De Melo, S. J. and De Mello, J. F. 1997. *Fitoterapia* 68: 478.
10. Abdel-Kader, M. S., Wisse, J., Evans, R., van der Werff, H. and Kingston, D. G. 1997. *Journal of Natural Products* 60: 1294–1297.
11. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
12. Alen, Y., Nakajima, S., Nitoda, T., Baba, N., Kanzaki, H. and Kawazu, K. 2000. *Zeitschrift für Naturforschung [C]* 55: 295–299.
13. Dixit, S. N., Tripathi, S. C. and Ojaha, T. N. 1982. *Bokin Bobai* 10: 197–199.
14. Connér, D. E. and Beuchat, L. R. 1984. *Journal of Food Science* 49: 429–434.
15. Tripathi, S. C. and Dixit, S. N. 1977. *Experientia* 33: 207–208.
16. Tiwari, T. N., Dubey, N. K. and Pandey, V. B. 1998. *Indian Drugs* 35: 415–416.
17. Otero, R., Núñez, V., Barona, J., Fonnegra, R., Jiménez, S. L., Osorio, R. G., Saldarriaga, M. and Díaz, A. 2000. *Journal of Ethnopharmacology* 73: 233–241.
18. Nayak, S., Nalabothu, P., Sandiford, S., Bhogadi, V. and Adogwa, A. 2006. *BMC Complementary and Alternative Medicine* 6: 12.
19. Apollo, M., Dash, S. K. and Padhy, S. 2006. *Journal of Human Ecology* 19: 239–248.

Aloe vera Mill.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
3. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=200027555
   Date Accessed: 30th March 2008
4. van Wyk, B. E. and Wink, M. 2004. Medicinal plants of the world: an illustrated scientific guide to important medicinal plants and their uses. Timber Press: Portland.
5. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
6. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
7. Djeraba, A. and Quere, P. 2000. *International Journal of Immunopharmacology* 22: 365–372.
8. Okamura, N., Hine, N., Tateyama, Y., Nakazawa, M., Fujioka, T., Mihashi, K. and Yagi, A. 1998. *Phytochemistry* 49: 219–223.
9. Pugh, N., Ross, S. A., ElSohly, M. A. and Pasco, D. S. 2001. *Journal of Agriculture and Food Chemistry* 49: 1030–1034.
10. Wang, H. M., Shi, W., Xu, Y. K., Liu, Y., Lu, M. J. and Pan, J. Q. 2003. Spectral study of a new dihydroisocoumarin. *Magnetic Resonance in Chemistry* 41: 718–720.
11. Zhang, X. F., Wang, H. M., Song, Y. L., Nie, L. H., Wang, L. F., Liu, B., Shen, P. P. and Liu, Y. 2006. *Bioorganic and Medicinal Chemistry Letters* 16: 949–953.
12. Esua, M. F. and Rauwald, J. W. 2006. *Carbohydrate Research* 341: 355–364.
13. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
14. Goh, S. H. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 7–26. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
15. Jaganath, I. B. and Ng, L. T. 2000. Herbs: The green pharmacy of Malaysia. Vinpress Sdn. Bhd: Kuala Lumpur, in collaboration with the Malaysian Agricultural Research and Development (MARDI).
16. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
17. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
18. Keys, J. D. 1976. Chinese herbs: their botany, chemistry, and pharmacodynamics.

Charles E. Tuttle Co. of Rutland: Vermont & Tokyo, Japan.

19. Grover, J. K., Yadav, S. and Vats, V. 2002. *Journal of Ethnopharmacology* 81: 81–100.
20. Moon, E. J., Lee, Y. M., Lee, O. H., Lee, M. J., Lee, S. K., Chung, M. H., Park, Y. I., Sung, C. K., Choi, J. S. and Kim, K. W. 2000. *Angiogenesis* 3: 117–123.
21. Jasso de Rodríguez, D., Hernández-Castillo, D., Rodríguez-García, R. and Angulo-Sánchez, J. L. 2005. *Industrial Crops and Product* 21: 81–87.
22. Stuart, R. W., Lefkowitz, D. L., Lincoln, J. A., Howard, K., Gelderman, M. P. and Lefkowitz, S. S. 1997. *International Journal of Immunopharmacology* 19: 75–82.
23. Rosca-Casian, O., Parvu, M., Vlase, L. and Tamas, M. 2007. *Fitoterapia* 78: 219–222.
24. Rajasekaran, S., Ravi, K., Sivagnanam, K. and Subramanian, S. 2006. *Clinical and Experimental Pharmacology and Physiology* 33: 232–237.
25. Tanaka, M., Misawa, E., Ito, Y., Habara, N., Nomaguchi, K., Yamada, M., Toida, T., Hayasawa, H., Takase, M., Inagaki, M. and Higuchi, R. 2006. *Biological and Pharmaceutical Bulletin* 29: 1418–1422.
26. Duansak, D., Somboonwong, J. and Patumraj, S. 2003. *Clinical Hemorheology and Microcirculation* 29: 239–246.
27. Langmead, L., Makins, R. J. and Rampton, D. S. 2004. *Alimentary Pharmacology and Therapeutics* 19: 521–527.
28. Lee, C. K., Han, S. S., Shin, Y. K., Chung, M. H., Park, Y. I., Lee, S. K. and Kim, Y. S. 1999. *International Journal of Immunopharmacology* 21: 303–310.
29. Prabjone, R., Thong-Ngam, D., Wisedopas, N., Chatsuwan, T. and Patumraj, S. 2006. *Clinical Hemorheology and Microcirculation* 35: 359–366.
30. Habeeb, F., Stables, G., Bradbury, F., Nong, S., Cameron, P., Plevin, R. and Ferro, V. A. 2007. *Methods* 42: 388–393.
31. Lee, K. H., Kim, J. H., Lim, D. S. and Kim, C. H. 2000. *Journal of Pharmacy and Pharmacology* 52: 593–598.
32. Pecere, T., Gazzola, M. V., Mucignat, C., Parolin, C., Dalla, V. F., Cavaggioni, A., Basso, G., Diaspro, A., Salvato, B., Carli, M. and Palu, G. 2000. *Cancer Research* 60: 2800–2804.
33. Acevedo-Duncan, M., Russell, C., Patel, S. and Patel, R. 2004. *International Immunopharmacology* 4: 1775–1784.
34. Akev, N., Turkay, G., Can, A., Gurel, A., Yildiz, F., Yardibi, H., Ekiz, E. E. and Uzun, H. 2007. *European Journal of Cancer Prevention* 16: 151–157.
35. Akev, N., Turkay, G., Can, A., Gurel, A., Yildiz, F., Yardibi, H., Ekiz, E. E. and Uzun, H. 2007. *Phytotherapy Research* 21: 1070–1075.
36. Chen, S. H., Lin, K. Y., Chang, C. C., Fang, C. L. and Lin, C. P. 2007. *Food and Chemical Toxicology* 45: 2296–2303.
37. Esmat, A. Y., El-Gerzawy, S. M. and Rafaat, A. 2005. *Cancer Biology and Therapy* 4: 108–112.
38. Nićiforović, A., Adzić, M., Spasić, S. D. and Radojcić, M. B. 2007. *Cancer Biology and Therapy* 6: 1200–1205.
39. Habeeb, F., Shakir, E., Bradbury, F., Cameron, P., Taravati, M. R., Drummond, A. J., Gray, A. I. and Ferro, V. A. 2007. *Methods* 42: 315–320.
40. Yagi, A., Kabash, A., Mizuno, K., Moustafa, S. M., Khalifa, T. I. and Tsuji, H. 2003. *Planta Medica* 69: 269–271.
41. Saada, H. N., Ussama, Z. S. and Mahdy, A. M. 2003. *Pharmazie* 58: 929–931.
42. Rajasekaran, S., Sivagnanam, K. and Subramanian, S. 2005. *Pharmacological Reports* 57: 90–96.
43. Gupta, R. and Flora, S. J. 2005. *Phytotherapy Research* 19: 23–28.
44. Wu, J. H., Xu, C., Shan, C. Y. and Tan, R. X. 2006. *Life Sciences* 78: 622–630.
45. Kardosová, A. and Machová, E. 2006. *Fitoterapia* 77: 367–373.
46. Yoo, E. J. and Lee, B. M. 2005. *Journal of Toxicology and Environmental Health, Part A* 68: 1841–1860.
47. Maze, G., Terpolilli, R. N. and Lee, M. K. 1997. *Medical Science Research* 25: 765–766.
48. Eamlamnam, K., Patumraj, S., Visedopas, N. and Thong-Ngam, D. 2006. *World Journal of Gastroenterology* 12: 2034–2039.
49. Etim, O. E., Farombi, E. O., Usoh, I. F. and Akpan, E. J. 2006. *Pakistan Journal of Pharmaceutical Sciences* 19: 337–340.

50. Chandan, B. K., Saxena, A. K., Shukla, S., Sharma, N., Gupta, D. K., Suri, K. A., Suri, J., Bhadauria, M. and Singh, B. 2007. *Journal of Ethnopharmacology* 111: 560–566.
51. Lin, H. J., Lai, C. C., Lee Chao, P. D., Fan, S. S., Tsai, Y., Huang, S. Y., Wan, L. and Tsai, F. J. 2007. *Pharmacology and Therapeutics* 23: 152–171.
52 Geremias, R., Pedrosa, R. C., Locatelli, C., de Fávere, V. T., Coury-Pedrosa, R. and Laranjeira, M. C. 2006. *Phytotherapy Research* 20: 288–293.
53. Hart, L. A. T., Van Enckevort, P. H., Van Dijk, H., Zaat, R., De Silva, K. T. D. and Labadie, R. P. 1988. *Journal of Ethnopharmacology* 23: 61–71.
54. Hart, L. A. T., Van den Berg, A. J. J., Kuis, L., Van Dijk, H. and Labadie, R. P. 1989. *Planta Medica* 55: 509–512.
55. Chen, X. D., Huang, L. Y., Wu, B. Y., Jiang, Q., Wang, Z. C. and Lin, X. H. 2005. *Zhongguo Wei Zhong Bing Ji Jiu Yi Xue* 17: 296–298.
56. Chen, X. D., Wu, B. Y., Jiang, Q., Huang, L. Y. and Wang, Z. C. 2005. *Zhongguo Zhong Yao Za Zhi* 30: 1944–1946.
57. Liu, C., Leung, M. Y., Koon, J. C., Zhu, L. F., Hui, Y. Z., Yu, B. and Fung, K. P. 2006. *International Immunopharmacology* 6: 1634–1641.
58. Djeraba, A. and Quere, P. 2000. *International Journal of Immunopharmacology* 22: 365–372.
59. Womble, D. and Helderman, J. H. 1988. *International Journal of Immunopharmacology* 10: 967–974.
60. Dutta, A., Mandal, G., Mandal, C. and Chatterjee, M. 2007. *Glycoconjugate Journal* 24: 81–86.
61. Dutta, A., Bandyopadhyay, S., Mandal, C. and Chatterjee, M. 2007. *Journal of Medical Microbiology* 56: 629–636.
62. Kirdpon, S., Kirdpon, W., Airarat, W., Trevanich, A. and Nanakorn, S. 2006. *Journal of the Medical Association of Thailand* 89 (Suppl 2): S9–14.
63. Kirdpon, S., Kirdpon, W., Airarat, W., Thepsuthammarat, K. and Nanakorn, S. 2006. *Journal of the Medical Association of Thailand* 89: 1199–1205.
64. Wang, Z. W., Zhou, J. M., Huang, Z. S., Yang, A. P., Liu, Z. C., Xia, Y. F., Zeng, Y. X. and Zhu, X. F. 2004. *Journal of Radiation Research (Tokyo)* 45: 447–454.
65 Wang, Z. W., Huang, Z, S., Yang, A. P., Li, C. Y., Huang, H., Lin, X., Liu, Z. C. and Zhu, X. F. 2005. *Ai Zheng* 24: 438–442.
66. Abdullah, K. M., Abdullah, A., Johnson Mary, L., Bilski, J. J., Petry, K., Redmer, D. A., Reynolds, L. P. and Grazul-Bilska, A. T. 2003. *Journal of Alternative and Complementary Medicine* 9: 711–718.
67. Chithra, P., Sajithlal, G. B. and Chandrakasan, G. 1998. *Journal of Ethnopharmacology* 59: 179–186.
68. Bezáková, L., Oblozinský, M., Sýkorová, M., Paulíková, I. and Kostálová, D. 2005. *Ceská a Slovenská Farmacie* 54: 43–46.
69. Maenthaisong, R., Chaiyakunapruk, N., Niruntraporn, S. and Kongkaew, C. 2007. *Burns* 33: 713–718.
70. Bradley, P. R. (Ed.). 1992. British herbal compendium. Volume 1. A handbook of scientific information on widely used plant drugs. Companion to Volume 1 of the British Herbal Pharmacopoeia. British Herbal Medicine Association: United Kingdom.
71. Fetrow, C. W. and Avila, J. R. 2000. The complete guide to herbal medicines. Springhouse Corp: Springhouse, PA.
72. Kuhn, M. A. and Winston, D. 2000. Herbal therapy and supplements: a scientific and traditional approach. Lippincott: Philadelphia.
73. Quick access: Professional guide to conditions, herbs & supplements. 2000. Newton, MA: Integrative Medicine Communications.
74. Boudreau, M. D. and Beland, F. A. 2006. *Journal of Environmental Science and Health. Part C, Environmental Carcinogenesis & Ecotoxicology Reviews* 24: 103–154.
75. Bottenberg, M. M., Wall, G. C., Harvey, R. L. and Habib, S. 2007. *The Annals of Pharmacotherapy* 41: 1740–1743.
76. Bisset, N. G. 1994. Herbal drugs and phytopharmaceuticals. CRC Press: Boca Raton, FL.
77. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
78. Vinson, J. A., Al Kharrat, H. and Andreoli, L. 2005. *Phytomedicine* 12: 760–765.

Andrographis paniculata (Barm.f.) Nees

1. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?104887 Date Accessed: 30th June 2008
3. Jain, D. C., Gupta, M. M., Saxena, S. and Kumar, S. 2000. *Journal of Pharmaceutical and Biomedical Analysis* 22: 705–709.
4. Cheung, H. Y., Cheung, C. S. and Kong, C. K. 2001. *Journal of Chromatography A* 930: 171–176.
5. Reddy, M. K., Reddy, M. V., Gunasekar, D., Murthy, M. M., Caux, C. and Bodo, B. 2003. *Phytochemistry*, 62: 1271–1275.
6. Koteswara, R. Y., Vimalamma, G., Rao, C. V. and Tzeng, Y. M. 2004. *Phytochemistry* 65: 2317–2321.
7. Chen, L. X., Qu, G. X. and Qiu, F. 2006. *Zhongguo Zhong Yao Za Zhi* 31: 391–395.
8. Chen, L. X., Qu, G. X. and Qiu, F. 2006. *Zhongguo Zhong Yao Za Zhi* 19: 1594–1597.
9. Reddy, V. L., Reddy, S. M., Ravikanth, V., Krishnaiah, P., Goud, T. V., Rao, T. P., Ram, T. S., Gonnade, R. G., Bhadbhade, M. and Venkateswarlu, Y. 2005. *Natural Product Research* 19: 223–230.
10. Pramanick, S., Banerjee, S., Achari, B., Das, B., Sen, A. K. Sr., Mukhopadhyay, S., Neuman, A. and Prangé, T. 2006. *Journal of Natural Products* 69: 403–405.
11. Shen, Y. H., Li, R. T., Xiao, W. L., Xu, G., Lin, Z. W., Zhao, Q. S. and Sun, H. D. 2006. *Journal of Natural Products* 69: 319–322.
12. Li, W., Xu, X., Zhang, H., Ma, C., Fong, H., van Breemen, R. and Fitzloff, J. 2007. *Chemical and Pharmaceutical Bulletin* 55: 455–458.
13. Wu, T. S., Chern, H. J., Damu, A. G., Kuo, P. C., Su, C. R., Lee, E. J. and Teng, C. M. 2008. *Journal of Asian Natural Product Research* 10: 17–24.
14. Chan, W. R., Taylor, D. R., Willis, C. R. and Fehlhaber, H. W. 1968. *Tetrahedron Letters* 9: 4803–4806.
15. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.
16. Natural Medicines Comprehensive Database 2008 http://www.naturaldatabase.com/S(fpd5gq45bydk4145u4wt5dij))/nd/Search.aspx?cs=&s=ND&pt=9&Product=andrographis+paniculata&btnSearch.x=21&btnSearch.y=9 Date Accessed: 23rd June2008
17. Chen, J. H., Hsiao, G., Lee, A. R., Wu, C. C. and Yen, M. H. 2004. *Biochemical Pharmacology* 67: 1337–1345.
18. Singha, P. K., Roy, S. and Dey, S. 2003. *Fitoterapia* 74: 692–694.
19. Limsong, J., Benjavongkulchai, E. and Kuvatanasuchati, J. 2004. *Journal of Ethnopharmacology* 92: 281–289.
20. Zaidan, M. R., Noor Rain, A., Badrul, A. R., Adlin, A., Norazah, A. and Zakiah, I. 2005. *Tropical Biomedicine* 22: 165–170.
21. Kumar, R. A., Sridevi, K., Kumar, N. V., Nanduri, S. and Rajagopal, S. 2004. *Journal of Ethnopharmacology* 92: 291–295.
22. Cheung, H. Y., Cheung, S. H., Li, J., Cheung, C. S., Lai, W. P., Fong, W. F. and Leung, F. M. 2005. *Planta Medica* 71: 1106–1111.
23. Jada, S. R., Hamzah, A. S., Lajis, N. H., Saad, M. S., Stevens, M. F. and Stanslas, J. 2006. *Journal of Enzyme Inhibition and Medicinal Chemistry* 21: 145–155.
24. Zhou, J., Zhang, S., Ong, C. N. and Shen, H. M. 2006. *Biochemical Pharmacology* 72: 132–144.
25. Jada, S. R., Subur, G. S., Matthews, C., Hamzah, A. S., Lajis, N. H., Saad, M. S., Stevens, M. F. G. and Stanslas, J. 2007. *Phytochemistry* 68: 904–912.
26. Sheeja, K. and Kuttan, G. 2007. *Immunopharmacology and Immunotoxicology* 29: 81–93.
27. Sheeja, K. and Kuttan, G. 2007. *Integrative Cancer Therapies* 6: 66–73.
28. Sheeja, K., Guruvayoorappan, C. and Kuttan, G. 2007. *International Immunopharmacology* 7: 211–221.
29. Ji, L., Liu, T., Liu, J., Chen, Y. and Wang, Z. 2007. *Planta Medica* 73: 1397–1401.
30. Jiang, C. G., Li, J. B., Liu, F. R., Wu, T., Yu, M. and Xu, H. M. 2007. *Anticancer Research* 27: 2439–2447.
31. Husen, R., Pihie, A. H. and Nallappan, M. 2004. *Journal of Ethnopharmacology* 95: 205–208.

32. Yu, B. C., Hung, C. R., Chen, W. C. and Cheng, J. T. 2003. *Planta Medica* 69: 1075–1079.
33. Zhang, X. F. and Tan, B. K. 2000. *Clinical and Experimental Pharmacology and Physiology* 27: 358–363.
34. Reyes, B. A., Bautista, N. D., Tanquilut, N. C., Anunciado, R. V., Leung, A. B., Sanchez, G. C., Magtoto, R. L., Castronuevo, P., Tsukamura, H. and Maeda, K. I. 2006. *Journal of Ethnopharmacology* 105: 196–200.
35. Xu, Y., Chen, A., Fry, S., Barrow, R. A., Marshall, R. L. and Mukkur, T. K. 2007. *International Immunopharmacology* 7: 515–523.
36. Umamaheswari, S. and Mainzen Prince, P. S. 2007. *Acta Poloniae Pharmaceutica* 64: 53–61.
37. Akbarsha, M. A. and Murugaian, P. 2000. *Phytotherapy Research* 14: 432–435.
38. Habtemariam, S. 1998. *Phytotherapy Research* 12: 37–40.
39. Hidalgo, M. A., Romero, A., Figueroa, J., Cortes, P., Concha, I. I., Hancke, J. L. and Burgos, R. A. 2005. *British Journal of Pharmacology* 144: 680–686.
40. Shen, Y. C., Chen, C. F. and Chiou, W. F. 2002. *British Journal of Pharmacology* 135: 399–406.
41. Madav, S., Tripathi, H. C., Tandan, S. K., Kumar, D. and Lal, J. 1998. *Indian Journal of Pharmaceutical Sciences* 60: 176–178.
42. Iruretagoyena, M. I., Tobar, J. A., Gonzalez, P. A., Sepulveda, S. E., Figueroa, C. A., Burgos, R. A., Hancke, J. L. and Kalergis, A. M. 2005. *The Journal of Pharmacology and Experimental Therapeutics* 312: 366–372.
43. Ji, L. L., Wang, Z., Dong, F., Zhang, W. B. and Wang, Z. T. 2005. *Journal of Cellular Biochemistry* 95: 970–978.
44. Ko, H. C., Wei, B. L. and Chiou, W. F. 2006. *Journal of Ethnopharmacology* 107: 205–210.
45. Sheeja, K., Shihab, P. K. and Kuttan, G. 2006. *Immunopharmacology and Immunotoxicology* 28: 129–140.
46. Liu, J., Wang, Z. T. and Ji, L. L. 2007. *The American Journal of Chinese Medicine* 35: 317–328.
47. Liu, J., Wang, Z. T., Ji, L. L. and Ge, B. X. 2007. *Molecular and Cellular Biochemistry* 298: 49–57.
48. Liu, J., Wang, Z. T. and Ge, B. X. 2008. *International Immunopharmacology* 8: 951–958.
49. Kamdem, R. E., Sang, S. and Ho, C. T. 2002. *Journal of Agricultural and Food Chemistry* 50: 4662–4665.
50. Tripathi, R. and Kamat, J. P. 2007. *Indian Journal of Experimental Biology* 45: 959–967.
51. Tripathi, R., Mohan, H. and Kamat, J. P. 2007. *Food Chemistry* 100: 81–90.
52. Burgos, R. A., Imilan, M., Sanchez, N. S. and Hancke, J. L. 2000. *Journal of Ethnopharmacology* 71: 115–121.
53. Thisoda, P., Rangkadilok, N., Pholphana, N., Worasuttayangkurn, L., Ruchirawat, S. and Satayavivad, J. 2006. *European Journal of Pharmacology* 553: 39–45.
54. Rahman, N. N. N. A., Furuta, T., Kojima, S., Takane, K. and Mustafa-Ali, M. 1999. *Journal of Ethnopharmacology* 64: 249–254.
55. Zaridah, M. Z., Idid, S. Z., Omar, A. W. and Khozirah, S. 2001. *Journal of Ethnopharmacology* 78: 79–84.
56. Dua, V. K., Ojha, V. P., Roy, R., Joshi, B. C., Valecha, N., Devi, C. U., Bhatnagar, M. C., Sharma, V. P. and Subbarao, S. K. 2004. *Journal of Ethnopharmacology* 95: 247–251.
57. Calabrese, C., Berman, S. H., Babish, J. G., Ma, X., Shinto, L., Dorr, M., Wells, K., Wenner, C. A. and Standish, L. J. 2000. *Phytotherapy Research: PTR* 14: 333–338.
58. Wiart, C., Kumar, K., Yusof, M. Y., Hamimah, H., Fauzi, Z. M. and Sulaiman, M. 2005. *Phytotherapy Research* 19: 1069–1070.
59. Woo, A. Y., Waye, M. M., Tsui, S. K., Yeung, S. T. and Cheng, C. H. 2008. *The Journal of Pharmacology and Experimental Therapeutics* 325: 226–235.
60. Chang, K. T., Lii, C. K., Tsai, C. W., Yang, A. J. and Chen, H. W. 2008. *Food and Chemical Toxicology* 46: 1079–1088.
61. Singh, R. P., Banerjee, S. and Rao, A. R. 2001. *Phytotherapy Research: PTR* 15: 382–390.
62. Visen, P. K., Shukla, B., Patnaik, G. K. and Dhawan, B. N. 1993. *Journal of Ethnopharmacology* 40: 131–136.
63. Trivedi, N. P., Rawal, U. M. and Patel, B. P. 2007. *Integrative Cancer Therapies* 6: 271–280.

64. Zhang, C. Y. and Tan, B. K. 1996. *Clinical and Experimental Pharmacology and Physiology* 23: 675–678.
65. Panossian, A., Davtyan, T., Gukassyan, N., Gukasova, G., Mamikonyan, G., Gabrielian, E. and Wikman, G. 2002. *Phytomedicine* 9: 598–605.
66. Burgos, R. A., Seguel, K., Perez, M., Meneses, A., Ortega, M., Guarda, M. I., Loaiza, A. and Hancke, J. L. 2005. *Planta Medica* 71: 429–434.
67. Mandal, S. C., Dhara, A. K. and Maiti, B. C. 2001. *Phytotherapy Research: PTR* 15: 253–256.
68. Yoopan, N., Thisoda, P., Rangkadilok, N., Sahasitiwat, S., Pholphana, N., Ruchirawat, S. and Satayavivad, J. 2007. *Planta Medica* 73: 503–511.
69. Pharmacopoeia of the People's Republic of China, volume 1. English Edition, 2000. Compiled by the State Pharmacopoeia Commission of People's Republic of China. Chemical Industry Press: Beijing.
70. Mills, S. and Bone, K. 2000. Principles and Practice of Phytotherapy. Churchill Livingstone: New York.
71. Chang, H. and But, P. 1987. Pharmacology and Application of China Materia Medica, volume two. World Scientific Publication: Singapore.
72. Hovhannisyan, A. S., Abrahamyan, H., Gabrielyan, E. S. and Panossian, A. G. 2006. *Phytomedicine* 13: 318–323.

Ardisia elliptica Thunb.

1. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
2. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
3. Wee, Y. C. and Corlett, R. 1986. The City and the Forest: Plant Life in Urban Singapore. Singapore University Press: National University of Singapore.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?403463
   Date Accessed: 30th June 2008
5. Chow, P. W., Sim, K. Y., Lim, P. L. and Chung, V. C. 1991. *Bulletin of the Singapore National Institute of Chemistry* 19: 87–93.
6. Ahmad, S. A., Catalano, S., Marsili, A., Morelli, I. and Scartoni, V. 1977. *Planta Medica* 32: 162–164.
7. Kobayashi, H. and de Mejia, E. 2005. *Journal of Ethnopharmacology* 96: 347–354.
8. Phadungkit, M. and Luanratana, O. 2006. *Natural Product Research* 20: 693–697.
9. Liu, N., Li, Y., Gua, J. X. and Qian, D. G. 1993. *Acta Academiae Medicinae Shanghai* 20: 49–54.
10. Jalil, J., Jantan, I., Shaari, K. and Rafi, I. A. A. 2004. *Pharmaceutical Biology* 42: 457–461.
11. Ibrahim, B. J., Kang, Y. H., Suh, D. Y. and Han, B. H. 1996. *Natural Product Sciences* 2: 86–89.

Areca catechu L.

1. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
2. Palaniswamy, U. R. 2003. A guide to Medicinal Plants of Asian origin and culture. CPL Press: United States of America.
3. Hsuan, K., Chin, S. C. and Tan, H. T. W. 1998. The Concise Flora of Singapore volume II: Monocotyledons. Singapore University Press: Singapore.
4. Wagner, W. L., Herbst, D. R. and Sohmer, S. H. 1999. Manual of the flowering plants of Hawaii (Revised edition). University of Hawaii Press: Honolulu.
5. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?104887
   Date Accessed: 30th June 2008
6. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
7. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
8. Holdsworth, D. K., Jones, R. A. and Self, R. 1998. *Phytochemistry* 48: 581–582.

9. Woodley, E. (Ed.). 1991. Medicinal plants of Papua New Guinea, part I. Veerlag Josef Margraf Scientific Books: Weikersheim.
10. Keys, J. D. 1976. Chinese herbs: their botany, chemistry, and pharmacodynamics. Charles E. Tuttle Co. of Rutland: Vermont & Tokyo, Japan.
11. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
12. Ghelardini, C., Galeotti, N., Lelli, C. and Bartolini, A. 2001. *Il Farmaco* 56: 383–385.
13. Muhammad, M. M., Abdul, M. A., Mohd, A. A., Rozita, M. N., Nashriyah, B. M., Abdul, R. R. and Kazuyoshi, K. 1997. *Pesticide Science* 51: 165–170.
14. Wang, Y. C. and Huang, T. L. 2005. Screening of anti-*FEMS Immunology and Medical Microbiology* 43: 295–300.
15. Chang, Y. C., Tai, K. W., Li, C. K., Chou, L. S. and Chou, M. Y. 1999. *Clinical Oral Investigations* 3: 25–29.
16. Sundqvist, K., Liu, Y., Erhardt, P., Nair, J., Bartsch, H. and Grafstrom, R. C. 1991. *IARC Scientific Publications* 105: 281–285.
17. Lodge, D., Johnston, G. A. R., Curtis, D. R. and Brand, S. J. 1977. *Brain Research* 136: 513–522.
18. Ahsana, D. and Shagufta, K. 1997. *Phytotherapy Research* 11: 174–176.
19. Ahsana, D. and Shagufta, K. 2000. *Pharmacology Biochemistry and Behavior* 65: 1–6.
20. Inokuchi, J., Okabe, H., Yamauchi, T., Nagamatsu, A., Nonaka, G. and Nishioka, I. 1986. *Life Sciences* 38: 1375–1382.
21. Chung, F. M., Shieh, T. Y., Yang, Y. H., Chang, D. M., Shin, S. J., Tsai, J. C., Chen, T. H., Tai, T. Y. and Lee, Y. J. 2007. *Translational Research* 150: 58–65.
22. Chang, M. C., Ho, Y. S., Lee, P. H., Chan, C. P., Lee, J. J., Hahn, L. J., Wang, Y. J. and Jeng, J. H. 2001. *Carcinogenesis* 22: 1527–1535.
23. Owen, P. L., Matainaho, T., Sirois, M. and Johns, T. 2007. *Journal of Biochemical and Molecular Toxicology* 21: 231–242.
24. Jeng, J. H., Ho, Y. S., Chan, C. P., Wang, Y. J., Hahn, L. J., Lei, D., Hsu, C. C. and Chang, M. C. 2000. *Carcinogenesis* 21: 1365–1370.
25. Park, Y. B., Jeon, S. M., Byun, S. J., Kim, H. S. and Choi, M. S. 2002. *Life Sciences* 70: 1849–1859.
26. Lai, Y. L., Lin, J. C., Yang, S. F., Liu, T. Y. and Hung, S. L. 2007. *Journal of Periodontal Research* 42: 69–76.
27. Hung, S. L., Cheng, Y. Y., Peng, J. L., Chang, L. Y., Liu, T. Y. and Chen, Y. T. 2005. *Journal of Periodontal Research* 76: 373–379.
28. Chiang, C. P., Hsieh, R. P., Chen, T. H., Chang, Y. F., Liu, B. Y., Wang, J. T., Sun, A. and Kuo, M. Y. 2002. *Journal of Oral Pathology and Medicine* 31: 402–409.
29. Hung, S. L., Lin, Y. J., Chien, E. J., Liu, W. G., Chang, H. W., Liu, T. Y. and Chen, Y. T. 2007. *Journal of Periodontal Research* 42: 393–401.
30. Chang, L. Y., Wan, H. C., Lai, Y. L., Liu, T. Y. and Hung, S. L. 2006. *Journal of Periodontal Research* 77: 1969–1977.
31. Wang, C. C., Liu, T. Y., Wey, S. P., Wang, F. I. and Jan, T. R. 2007. *Food and Chemical Toxicology* 45: 1410–1418.
32. Lee, K. K., Cho, J. J., Park, E. J. and Choi, J. D. 2001. *International Journal of Cosmetic Science* 23: 341–346.
33. Pithayanukul, P., Ruenraroengsak, P., Bavovada, R., Pakmanee, N., Suttisri, R. and Saen-oon, S. 2005. *Journal of Ethnopharmacology* 97: 527–533.
34. Chang, M. C., Kuo, M. Y., Hahn, L. J., Hsieh, C. C., Lin, S. K. and Jeng, J. H. 1998. *Journal of Periodontal Research* 69: 1092–1097.
35. Jaiswal, P. and Singh, D. K. 2008. *Veterinary Parasitology* 152: 264–270.
36. Chatrchaiwiwatana, S. 2006. *Journal of the Medical Association of Thailand* 89: 1004–1011.
37. Skidmore-Roth, L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
38. Fetrow, C. W. and Avila, J. R. 2000. The complete guide to herbal medicines. Springhouse Corp: Springhouse, PA.
39. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
40. Wu, M. T., Wu, D. C., Hsu, H. K., Kao, E. L. and Lee, J. M. 2003. *British Journal of Cancer* 89: 1202–1204.

41. Kang, I. M., Chou, C. Y., Tseng, Y. H., Huang, C. C., Ho, W. Y., Shih, C. M. and Chen, W. 2007. *Journal of Occupational and Environmental Medicine* 49: 776–769.
42. Lan, T. Y., Chang, W. C., Tsai, Y. J., Chuang, Y. L., Lin, H. S. and Tai, T. Y. 2007. *American Journal of Epidemiology* 165: 677–683.
43. Yang, M. S., Lee, C. H., Chang, S. J., Chung, T. C., Tsai, E. M., Ko, A. M. and Ko, Y. C. 2008. *Drug and Alcohol Dependence* 95: 134–139.
44. López-Vilchez, M. A., Seidel, V., Farré, M., García-Algar, O., Pichini, S. and Mur, A. 2006. *Pediatrics* 117: e129–131.
45. García-Algar, O., Vall, O., Alameda, F., Puig, C., Pellegrini, M., Pacifici, R. and Pichini, S. 2005. *Archives of Diseases in Childhood: Fetal and neonatal edition* 90: F276–277.
46. Chou, W. W., Guh, J. Y., Tsai, J. F., Hwang, C. C., Chen, H. C., Huang, J. S., Yang, Y. L., Hung, W. C. and Chuang, L. Y. 2008. *Toxicology* 243: 1–10.
47. Lai, K. C. and Lee, T. C. 2006. *Mutation Research* 599: 66–75.

Asplenium nidus L.

1. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?5590
   Date Accessed: 30th June 2008
3. Imperato, F. 1987. *Chemistry & Industry* 14: 487.
4. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.
5. Bourdy, G. and Walter, A. 1992. *Journal of Ethnopharmacology* 37: 179–196.
6. Bourdy, G., Francois, C., Andary, C. and Boucard, M. 1996. *Journal of Ethnopharmacology* 52: 139–143.

Aster tataricus L.f.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
3. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=200023501
   Date Accessed: 30th March 2008
4. Ng, T. B., Liu, F., Lu, Y. H., Cheng, C. H. K. and Wang, Z. T. 2003. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* 136: 109–115.
5. Schumacher, K. K., Hauze, D. B., Jiang, J. J., Szewczyk, J., Reddy, R. E., Davis, F. A. and Joullié, M. M. 1999. *Tetrahedron Letters* 40: 455–458.
6. Cheng, D. L. and Shao, Y. 1993. *Phytochemistry* 35: 173–176.
7. Cheng, D. L., Shao, Y., Hartmann, R., Roeder, E. and Zhao, K. 1996. *Phytochemistry* 41: 225–227.
8. Morita, H., Nagashima, S., Shirota, O., Takeya, K. and Itokawa, H. 1993. *Chemistry Letters* 11: 1877–1880.
9. Morita, H., Nagashima, S., Takeya, K., Itokawa, H. and Iitaka, Y. 1995. *Tetrahedron* 51: 1121–1132.
10. Akihisa, T., Kimura, Y., Koike, K., Tai, T., Yasukawa, K., Arai, K., Suzuki, Y. and Nikaido, T. 1998. *Chemical & Pharmaceutical Bulletin* 46: 1824–1826.
11. Akihisa, T., Kimura, Y., Tai, T. and Arai, K. 1999. *Chemical & Pharmaceutical Bulletin* 47: 1161–1163.
12. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.
13. Keys, J. D. 1976. Chinese herbs: their botany, chemistry, and pharmacodynamics. Charles E. Tuttle Co. of Rutland: Vermont & Tokyo, Japan.
14. Rossi, F., Zanotti, G., Saviano, M., Iacovino, R., Palladino, P., Saviano, G., Amodeo, P., Tancredi, T., Laccetti, P., Corbier, C. and Benedetti, E. 2004. *Journal of Peptide Science* 10: 92–102.
15. Cozzolino, R., Palladino, P., Rossi, F., Calì, G., Benedetti, E. and Laccetti, P. 2005. *Carcinogenesis* 26: 733–739.

Azadirachta indica A. Juss.

1. Palaniswamy, U. R. 2003. A guide to Medicinal Plants of Asian origin and culture. CPL Press: United States of America.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Sumner J. 2000. The Natural History of Medicinal Plants. Timber Press, Portland.
4. eFloras.org 2008 http://www.efloras.org/florataxon.aspx?flora_id=5&taxon_id=220001427 Date Accessed: 30th March 2008
5. Siddiqui, B. S., Afshan, F. and Faizi, S. 2001. *Tetrahedron* 57: 10281–10286.
6. Siddiqui, B. S., Afshan, F., Ghiasuddin, Faizi, S., Naqvi, S. N. H. and Tariq, R. M. 2000. *Phytochemistry* 53: 371–376.
7. Siddiqui, S., Faizi, S., Mahmood, T. and Siddiqui, B. S. 1986. *Tetrahedron* 42: 4849–4856.
8. Sinniah, D., Goh, S. H. and Baskaran, G. 1988. Proceedings: Malaysian Traditional Medicines, Kuala Lumpur, pp. 65–69. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
9. Siddiqui, B. S., Afshan, F., Sham-Sul-Arfeen and Gulzar, T. 2006. *Natural Product Research* 20: 1036–1040.
10. Siddiqui, B. S., Tariq Ali, S. and Kashif Ali, S. 2006. *Natural Product Research* 20: 241–245.
11. Siddiqui, B. S., Rasheed, M., Ilyas, F., Gulzar, T., Tariq, R. M. and Naqvi, S. N. 2004. *Zeitschrift für Naturforschung. C, Journal of Biosciences* 59: 104–112.
12. Siddiqui, B. S., Afshan, F., Gulzar, T. and Hanif, M. 2004. *Phytochemistry* 65: 2363–2367.
13. Siddiqui, B. S., Afshan, F., Gulzar, T., Sultana, R., Naqvi, S. N. and Tariq, R. M. 2003. *Chemical and Pharmaceutical Bulletin (Tokyo)* 51: 415–417.
14. Siddiqui, B. S., Afshan, F., Faizi, S., Naqvi, S. N. and Tariq, R. M. 2002. *Journal of Natural Products* 65: 1216–1218.
15. Garg, H. S. and Bhakuni, D. S. 1985. *Phytochemistry* 24: 866–867.
16. Garg, H. S. and Bhakuni, D. S. 1984. *Phytochemistry* 23: 2115–2118.

17 Garg, H. S. and Bhakuni, D. S. *Phytochemistry* 23: 2383–2385.

18. Govindachari, T. R. and Gopalakrishnan, G. 1997. *Phytochemistry* 45: 397–399.
19. Lee, S. M., Olsen, J. I., Schweizer, M. P. and Klocke, J. A. 1988. *Phytochemistry* 27: 2773–2775.
20. Luo, X. D., Wu, S. H., Ma, Y. B. and Wu, D. G. 2000. *Fitoterapia* 71: 668–672.
21. Mangalam, S. N., Sindhu, G. and Deepa, I. 1997. *Phytochemistry* 46: 1177–1178.
22. Nanduri, S., Thunuguntla, S. S. R., Nyavanandi, V. K., Kasu, S., Kumar, P. M., Ram, P. S., Rajagopal, S., Kumar, R. A., Deevi, D. S., Rajagopalan, R. and Venkateswarlu, A. 2003. *Bioorganic & Medicinal Chemistry Letters* 22: 4111–4115.
23. Ramji, N., Venkatakrishnan, K. and Madyastha, K. M. 1996. *Phytochemistry* 42: 561–562.
24. Rojatkar, S. R. and Nagasampagi, B. A. 1992. *Phytochemistry* 32: 213–214.
25. Suresh, G., Narasimhan, N. S. and Palani, N. 1997. *Phytochemistry* 45: 807–810.
26. Singh, U. P., Maurya, S. and Singh, D. P. 2005. *Journal of Herbal Pharmacotherapy* 5: 35–43.
27. Kanokmedhakul, S., Kanokmedhakul, K., Prajuabsuk, T., Panichajakul, S., Panyamee, P., Prabpai, S. and Kongsaeree, P. 2005. *Journal of Natural Products* 68: 1047–1050.
28. Sharma, V., Bali, A. and Singh, M. 1998. *Phytochemistry* 49: 2121–2123.
29. Ghazanfar, S. A. 2004. Handbook of Arabian Medicinal Plants. CRC Press: Boca Raton, FL.
30. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
31. Ajose, F. O. 2007. *International Journal of Dermatology* 46(Suppl 1): 48–55.
32. Subapriya, R. and Nagini, S. 1995. *Current Medicinal Chemistry, Anticancer Agents* 5: 149–146.
33. Baswa, M., Rath, C. C., Dash, S. K. and Mishra, R. K. 2001. *Microbios* 105: 183–189.
34. Pai, M. R., Acharya, L. D. and Udupa, N. 2004. *Journal of Ethnopharmacology* 90: 99–103.

35. SaiRam, M., Ilavazhagan, G., Sharma, S. K., Dhanraj, S. A., Suresh, B., Parida, M. M., Jana, A. M., Devendra, K. and Selvamurthy, W. 2000. *Journal of Ethnopharmacology* 71: 377–382.
36. Mahfuzul Hoque, M. D., Bari, M. L., Inatsu, Y., Juneja, V. K. and Kawamoto, S. 2007. *Foodborne Pathogens and Disease* 4: 481–488.
37. Thakurta, P., Bhowmik, P., Mukherjee, S., Hajra, T. K., Patra, A. and Bag, P. K. 2007. *Journal of Ethnopharmacology* 111: 607–612.
38. Baral, R. and Chattopadhyay, U. 2004. *International Immunopharmacology* 4: 355–366.
39. Bose, A. and Baral, R. 2007. *Human Immunology* 68: 823–831.
40. Baral, R., Mandal, I. and Chattopadhyay, U. 2005. *International Immunopharmacology* 5: 1343–1352.
41. Roy, M. K., Kobori, M., Takenaka, M., Nakahara, K., Shinmoto, H., Isobe, S. and Tsushida, T. 2007. *Phytotherapy Research* 21: 245–250.
42. Roy, M. K., Kobori, M., Takenaka, M., Nakahara, K., Shinmoto, H. and Tsushida, T. 2006. *Planta Medica* 72: 917–923.
43. Kumar, S., Suresh, P. K., Vijayababu, M. R., Arunkumar, A. and Arunakaran, J. 2006. *Journal of Ethnopharmacology* 105: 246–250.
44. Koul, A., Mukherjee, N. and Gangar, S. C. 2006. *Molecular and Cellular Biochemistry* 283: 47–55.
45. Sastry, B. S., Suresh Babu, K., Hari Babu, T., Chandrasekhar, S., Srinivas, P. V., Saxena, A. K. and Madhusudana Rao, J. 2006. *Bioorganic and Medicinal Chemistry Letters* 16: 4391–4394.
46. Subapriya, R., Kumaraguruparan, R. and Nagini, S. 2006. *Clinical Biochemistry* 39: 1080–1087.
47. Gangar, S. C. and Koul, A. 2007. *Indian Journal of Biochemistry and Biophysics* 44: 209–215.
48. Subapriya, R., Bhuvaneswari, V. and Nagini, S. 2005. *Asian Pacific Journal of Cancer Prevention* 6: 515–520.
49. Subapriya, R., Bhuvaneswari, V., Ramesh, V. and Nagini, S. 2005. *Cell Biochemistry and Function* 23: 229–238.
50. Sarkar, K., Bose, A., Laskar, S., Choudhuri, S. K., Dey, S., Roychowdhury, P. K. and Baral, R. 2007. *International Immunopharmacology* 7: 306–312.
51. Niture, S. K., Rao, U. S. and Srivenugopal, K. S. 2006. *International Journal of Oncology* 29: 1269–1278.
52. Gangar, S. C., Sandhir, R., Rai, D. V. and Koul, A. 2006. *World Journal of Gastroenterology* 12: 2749–2755.
53. Sritanaudomchai, H., Kusamran, T., Kuakulkiat, W., Bunyapraphatsara, N., Hiransalee, A., Tepsuwan, A. and Kusamran, W. R. 2005. *Asian Pacific Journal of Cancer Prevention* 6: 263–269.
54. Arakaki, J., Suzui, M., Morioka, T., Kinjo, T., Kaneshiro, T., Inamine, M., Sunagawa, N., Nishimaki, T. and Yoshimi, N. 2006. *Asian Pacific Journal of Cancer Prevention* 7: 467–471.
55. Dasgupta, T., Banerjee, S., Yadava, P. K. and Rao, A. R. 2004. *Journal of Ethnopharmacology* 92: 23–36.
56. Parshad, O., Young, L. E. and Young, R. E. 1997. *Phytotherapy Research* 11: 398–400.
57. Parshad, O., Gardner, M. T., The, T. L., Williams, L. A. D. and Fletcher, C. K. 1997. *Phytotherapy Research* 11: 168–170.
58. Khillare, B. and Shrivastav, T. G. 2003. *Contraception* 68: 225–229.
59. Amadioha, A. C. 2000. *Crop Protection* 19: 287–290.
60. Govindachari, T. R., Suresh, G. and Masilamani, S. 1999. *Fitoterapia* 70: 417–420.
61. Polaquini, S. R., Svidzinski, T. I., Kemmelmeier, C. and Gasparetto, A. 2006. *Archieves of Oral Biology* 51: 482–490.
62. Kaur, G., Alam, M. S. and Athar, M. 2004. *Phytotherapy Research* 18: 419–424.
63. El Tahir, A., Satti, G. M. and Khalid, S. A. 1999. *Journal of Ethnopharmacology* 64: 227–233.
64. RaviDhar, Zhang, K. Y., Talwar, G. P., Garg, S. and Kumar, N. 1998. *Journal of Ethnopharmacology* 61: 31–39.
65. Nathan, S. S., Kalaivani, K. and Murugan, K. 2005. *Acta Tropica* 96: 47–55.
66. Udeinya, I. J., Brown, N., Shu, E. N., Udeinya, F. I. and Quakeyie, I. 2006. *Annals of Tropical Medicine and Parasitology* 100: 17–22.

67. Kirira, P. G., Rukunga, G. M., Wanyonyi, A. W., Muregi, F. M., Gathirwa, J. W., Muthaura, C. N., Omar, S. A., Tolo, F., Mungai, G. M. and Ndiege, I. O. 2006. *Journal of Ethnopharmacology* 106: 403–407.
68. Soh, P. N. and Benoit-Vical, F. 2007. *Journal of Ethnopharmacology* 114: 130–140.
69. Gopal, M., Gupta, A., Arunachalam, V. and Magu, S. P. 2007. *Bioresource Technology* 98: 3154–3158.
70. Gupta, S., Kataria, M., Gupta, P. K., Murganandan, S. and Yashroy, R. C. 2004. *Journal of Ethnopharmacology* 90: 185–189.
71. Subapriya, R., Kumaraguruparan, R., Abraham, S. K. and Nagini, S. 2005. *Journal of Herbal Pharmacotherapy* 5: 39–50.
72. Sithisarn, P., Supabphol, R. and Gritsanapan, W. 2006. *Medical Principles and Practice* 15: 219–222.
73. Sithisarn, P., Supabphol, R. and Gritsanapan, W. 2005. *Journal of Ethnopharmacology* 99: 109–112.
74. Prakash, D., Suri, S., Upadhyay, G. and Singh, B. N. 2007. *International Journal of Food Science and Nutrition* 58: 18–28.
75. Di Ilio, V., Pasquariello, N., van der Esch, A. S., Cristofaro, M., Scarsella, G. and Risuleo, G. 2006. *Molecular and Cellular Biochemistry* 287: 69–77.
76. Chaube, S. K., Prasad, P. V., Khillare, B. and Shrivastav, T. G. 2006. *Fertility and Sterility* 85 Suppl 1: 1223–1231.
77. Ashorobi, R. B. 1998. *Phytotherapy Research* 12: 41–43.
78. Parida, M. M., Upadhyay, C., Pandya, G. and Jana, A. M. 2002. *Journal of Ethnopharmacology* 79: 273–278.
79. Joshi, S. N., Katti, U., Godbole, S., Bharucha, K., Kishore Kumar, B., Kulkarni, S., Risbud, A. and Mehendale, S. 2005. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 99: 769–774.
80. Yinusa, R., Isiaka, A. O., Caleb, A. O. and Godwin, J. 2004. *Journal of Ethnopharmacology* 90: 167–170.
81. Bandyopadhyay, U., Biswas, K., Chatterjee, R., Bandyopadhyay, D., Chattopadhyay, I., Ganguly, C. K., Chakraborty, T., Bhattacharya, K. and Banerjee, R. K. 2002. *Life Sciences* 71: 2845–2865.
82. Dorababu, M., Joshi, M. C., Bhawani, G., Kumar, M. M., Chaturvedi, A. and Goel, R. K. 2006. *Indian Journal of Physiology and Pharmacology* 50: 241–249.
83. Chattopadhyay, R. R. 2003. *Journal of Ethnopharmacology* 89: 217–219.
84. Koul, A., Ghara, A. R. and Gangar, S. C. 2006. *Phytotherapy Research* 20: 169–177.
85. Koul, A., Binepal, G. and Gangar, S. C. 2007. *Indian Journal of Experimental Biology* 45: 359–366.
86. Chattopadhyay, R. R. 1999. *Journal of Ethnopharmacology* 67: 367–372.
87. Waheed, A., Miana, G. A. and Ahmad, S. I. 2006. *Pakistan Journal of Pharmaceutical Sciences* 19: 322–325.
88. Oiefuna, I. and Young, R. 2004. Novel Compounds from Natural Products in the New Millennium. Potential and Challenges. World Scientific Publishing: Singapore.
89. Chattopadhyay, R. R. 1997. *General Pharmacology: The Vascular System* 28: 449–451.
90. Obiefuna, I. and Young, R. 2005. *Phytotherapy Research* 19: 792–795.
91. Beuth, J., Schneider, H. and Ko, H. L. 2006. *In Vivo* 20: 247–251.
92. Haque, E. and Baral, R. 2006. *Immunobiology* 211: 721–731.
93. Haque, E., Mandal, I., Pal, S. and Baral, R. 2006. *Immunopharmacology and Immunotoxicology* 28: 33–50.
94. Yanpallewar, S., Rai, S., Kumar, M., Chauhan, S. and Acharya, S. B. 2005. *Life Sciences* 76: 1325–1338.
95. Hördegen, P., Cabaret, J., Hertzberg, H., Langhans, W. and Laurer, V. 2006. *Journal of Ethnopharmacology* 108: 85–89.
96. Fatima, F., Khalid, A., Nazar, N., Abdalla, M., Mohomed, H., Toum, A. M., Magzoub, M. and Alı, M. S. 2005. *Turkish Society for Parasitology* 29: 3–6.
97. Nakahara, K., Roy, M. K., Ono, H., Maeda, I., Ohnishi-Kameyama, M., Yoshida, M. and Trakoontivakorn, G. 2003. *Journal of Agriculture and Food Chemistry* 51: 6456–6460.
98. Nakahara, K., Trakoontivakorn, G., Alzoreky, N. S., Ono, H., Onishi-Kameyama, M. and Yoshida, M. 2002. *Journal of Agriculture and Food Chemistry* 50: 4796–4802.

99. Farah, M. A., Ateeq, B. and Ahmad, W. 2006. *The Science of the Total Environment* 364: 200–214.
100. Singh, K., Singh, A. and Singh, D. K. 1996. *Journal of Ethnopharmacology* 52: 35–40.
101. Abdel-Shafy, S. and Zayed, A. A. 2002. *Veterinary Parasitology* 106: 89–96.
102. Al-Rajhy, D. H., Alahmed, A. M., Hussein, H. I. and Kheir, S. M. 2003. *Pest Management Science* 59: 1250–1254.
103. Paul, P. K. and Sharma, P. D. 2002. *Physiological and Molecular Plant Pathology* 61: 3–13.
104. Singh, U. P. and Prithiviraj, B. 1997. *Physiological and Molecular Plant Pathology* 51: 181–194.
105. Thacker, J. R. M., Bryan, W. J., McGinley, C., Heritage, S. and Strang, R. H. C. 2003. *Crop Protection* 22: 753–760.
106. Nathan, S. S., Savitha, G., George, D. K., Narmadha, A., Suganya, L. and Chung, P. G. 2006. *Bioresource Technology* 97: 1316–1323.
107. Nathan, S. S., Kalaivani, K. and Murugan, K. 2006. *Ecotoxicology and Environmental Safety* 65: 102–107.
108. Nathan, S. S., Chung, P. G. and Murugan, K. 2006. *Ecotoxicology and Environmental Safety* 64: 382–389.
109. Anuradha, A., Annadurai, R. S. and Shashidhara, L. S. 2007. *Insect Biochemistry and Molecular Biology* 37: 627–634.
110. Okumu, F. O., Knols, B. G. and Fillinger, U. 2007. *Malaria Journal* 6: 63.
111. Skidmore-Roth L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
112. Boeke, S. J., Boersma, M. G., Alink, G. M., van Loon, J. J., van Huis, A., Dicke, M. and Rietjens, I. M. 2004. *Journal of Ethnopharmacology* 94: 25–41.

Barringtonia asiatica L.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?6512
   Date Accessed: 30th June 2008
5. Herlt, A. J., Mander, L. N., Pongoh, E., Rumampuk, R. J. and Tarigan, P. 2002. *Journal of Natural Products* 65: 115–120.
6. Nozoe, T. 1934. *Nippon Kagaku Kaishi* 55: 1106–1114.
7. Burton, R. A., Wood, S. G. and Owen, N. L. 2003. *Arkivoc* Part 13: 137–146.
8. Locher, C. P., Burch, M. T., Mower, H. F., Berestecky, J., Davis, H., Van Poel, B., Lasure, A., Vanden Berghe, D. A. and Vlietinck, A. J. 1995. *Journal of Ethnopharmacology* 49: 23–32.
9. Whistler, A. W. 1992. Polynesian Herbal Medicine. Everbest Publishing Co.: Hong Kong.
10. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.
11. Steiner, R. P. 1986. Folk medicine: the art and the science. American Chemical Society: Washington D. C.
12. Woodley, E. (Ed.). 1991. Medicinal plants of Papua New Guinea, part I. Veerlag Josef Margraf Scientific Books: Weikersheim.
13. Khan, M. R. and Omoloso, A. D. 2002. *Fitoterapia* 73: 255–260.

Barringtonia racemosa (L.) K. Spreng

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?6514
   Date Accessed: 30th June 2008
4. Anantaraman, R., Pillai, K. S. and Madhavan. 1956. *Journal of the Chemical Society, Abstracts*: 4369–4373.

5. Lin, Y. T., Lo, T. B. and Su, S. C. 1957. *Journal of the Chinese Chemical Society, Taipei, Taiwan* 4 (Ser. II): 77–81.
6. Pillai, K. S. and Madhavan. 1959. *Bull. Cent. Research Inst. University of Kerala, Trivandrum, Ser. A* 6: 15–20.
7. Hasan, C. M., Khan, S., Jabbar, A. and Rashid, M. A. 2000. *Journal of Natural Products* 63: 410–411.
8. Sun, H. Y., Long, L. J. and Wu, J. 2006. *Zhong Yao Cai* 29: 671–672.
9. Yang, Y., Deng, Z., Proksch, P. and Lin, W. 2006. *Pharmazie* 61: 365–366.
10. Deraniyagala, S. A., Ratnasooriya, W. D. and Goonasekara, C. L. 2003. *Journal of Ethnopharmacology* 86: 21–26.
11. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
12. Khan, S., Jabbar, A., Hasan, C. M. and Rashid, M. A. 2001. *Fitoterapia* 72: 162–164.
13. Gowri. P. M., Tiwari, A. K., Ali, A. Z. and Rao, J. M. 2007. *Phytotherapy Research* 2: 796–799.
14. Thomas, T. J., Panikkar, B., Subramoniam, A., Nair, M. K. and Panikkar, K. R. 2002. *Journal of Ethnopharmacology* 82: 223–227.
15. Mackeen, M. M., Ali, A. M., El-Sharkawy, S. H., Manap, M. Y., Salleh, K. M., Lajis, N. H. and Kawazu, K. 1997. *International Journal of Pharmacognosy* 35: 174–178.

Bauhinia purpurea L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
3. Wu, A. M., Wu, J. H., Liu, J. H. and Singh, T. 2004. *Life Sciences* 74: 1763–1779.
4. Yadava, R. N. and Tripathi, P. 2000. *Fitoterapia* 71: 88–90.
5. Kuo, Y. H., Yeh, M. H. and Huang, S. L. 1998. *Phytochemistry* 49: 2529–2530.
6. Bhartiya, H. P. and Gupta, P. C. 1981. *Phytochemistry* 20: 2051.
7. Bhartiya, H. P., Dubey, P., Katiyar, S. B. and Gupta, P. C. 1979. *Phytochemistry* 18: 689.
8. Pettit, G. R., Numata, A., Iwamoto, C., Usami, Y., Yamada, T., Ohishi, H., Cragg, G. M. 2006. *Journal of Natural Products* 69: 323–327.
9. Boonphong, S., Puangsombat, P., Baramee, A., Mahidol, C., Ruchirawat, S. and Kittakoop, P. 2007. *Journal of Natural Products* 70: 795–801.
10. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.
11. Brennan, M. J., Cisar, J. O., Vatter, A. E. and Sandberg, A. L. 1984. *Infection and Immunity* 46: 459–464.
12. Zakaria, Z. A., Wen, L. Y., Abdul Rahman, N. I., Abdul Ayub, A. H., Sulaiman, M. R. and Gopalan, H. K. 2007. *Medical Principles and Practice* 16: 443–449.
13. Panda, S. and Kar, A. 1999. *Journal of Ethnopharmacology* 67: 233–239.

Bixa orellana L.

1. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
4. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200014319
Date Accessed: 30th March 2008
5. Houssiau, L., Felicissimo, M., Bittencourt, C. and Pireaux, J. J. 2004. *Applied Surface Science* 231–232: 416–419.
6. Mercadante, A. Z., Steck, A. and Pfander, H. 1997. *Phytochemistry* 46: 1379–1383.
7. Mercadante, A. Z., Steck, A. and Pfander, H. 1999. *Phytochemistry* 52: 135–139.
8. Mercadante, A. Z., Steck, A., Rodriguez-Amaya, D., Pfander, H. and Britton, G. 1996. *Phytochemistry* 41: 1201–1203.

9. Lawrence, B. M. and Hogg, J. W. 1973. *Phytochemistry* 12: 2995.
10. Schneider, W. P., Caron, E. L. and Hinman, J. W. 1965. *Journal of Organic Chemistry* 30: 2856–2857.
11. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
12. Fleischer, T. C., Ameade, E. P. K., Mensah, M. L. K. and Sawer, I. K. 2003. *Fitoterapia* 74: 136–138.
13. Lans, C. A. 2006. *Journal of Ethnobiology and Ethnomedicine* 2: 45.
14. Morton, J. F. 1981. Atlas of Medicinal Plants of Middle America: Bahamas to Yucatan. C.C. Thomas Publishers Co.: Springfield, IL.
15. Cáceres, A., Menéndez, H., Méndez, E., Cohobón, E., Samayoa, B. E., Jauregui, E., Peralta, E. and Carrillo, G. 1995. *Journal of Ethnopharmacology* 48: 85–88.
16. Shilpi, J. A., Taufiq-Ur-Rahman, M., Uddin, S. J., Alam, M. S., Sadhu, S. K. and Seidel, V. 2006. *Journal of Ethnopharmacology* 108: 264–271.
17. Rojas, J. J., Ochoa, V. J., Ocampo, S. A. and Muñoz, J. F. 2006. *BMC Complementary and Alternative Medicine* 6: 2.
18. Galindo-Cuspinera, V. and Rankin, S. A. 2005. *Journal of Agriculture and Food Chemistry* 53: 2524–2529.
19. Reddy, M. K., Alexander-Lindo, R. L. and Nair, M. G. 2005. *Journal of Agriculture and Food Chemistry* 53: 9268–9273.
20. Nunez, V., Otero, R., Barona, J., Saldarriaga, M., Osorio, R. G., Fonnegra, R., Jimenez, S. L., Diaz, A. and Quintanam J. C. 2004. *Brazilian Journal of Medical and Biological Research* 37: 969–977.
21. Villar, R., Calleja, J. M., Morales, C. and Cáceres, A. 1997. *Phytotherapy Research* 11: 441–445.
22. Russell, K. R., Morrison, E. Y. and Ragoobirsingh, D. 2005. *Phytotherapy Research* 19: 433–436.
23. Júnior, A. C., Asad, L. M., Oliveira, E. B., Kovary, K., Asad, N. R. and Felzenszwalb, I. 2005. *Genetics and Molecular Research* 4: 94–99.
24. Braga, F. G., Bouzada, M. L., Fabri, R. L., de O Matos, M., Moreira, F. O., Scio, E. and Coimbra, E. S. 2007. *Journal of Ethnopharmacology* 111: 396–402.
25. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of medicinal herbs. CRC Press: Boca Raton, FL.
26. Duke, J. A. and duCellier, J. L. 1993. Handbook of Alternative Cash Crops. CRC Press: Boca Raton, FL.
27. Agner, A. R., Barbisan, L. F., Scolastici, C. and Salvadori, D. M. F. 2004. *Food and Chemical Toxicology* 42: 1687–1693.
28. Paumgartten, F. J. R., De-Carvalho, R. R., Araujo, I. B., Pinto, F. M., Borges, O. O., Souza, C. A. M. and Kuriyama, S. N. 2002. *Food and Chemical Toxicology* 40: 1595–1601.

Calophyllum inophyllum L.

1. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?8631
   Date Accessed: 30th June 2008
4. Govindachari, T. R. and Viswanathan, B. R. P. N., Ramadas, R. U. and Srinivasan, M. 1967. *Tetrahedron* 23: 1901–1910.
5. Kawazu, K., Ohigashi, H. and Mitsui, T. 1968. *Tetrahedron Letters* 9: 2383–2385.
6. Jackson, B., Locksley, H. D. and Scheinmann, F. 1969. *Phytochemistry* 8: 927–929.
7. Subramanian, S. S. and Nair, A. G. R. 1971. *Phytochemistry* 10: 1679–1680.
8. Al-Jeboury, F. S. and Locksley, H. D. 1971. *Phytochemistry* 10: 603–606.
9. Bhushan, B., Rangaswami, S. and Seshadri, T. R. 1975. *Indian Journal of Chemistry* 13: 746–747.
10. Kumar, V., Ramachandran, S. and Sultanbawa, M. U. S. 1976. *Phytochemistry* 15: 2016–2017.
11. Goh, S. H. and Jantan, I. 1991. *Phytochemistry* 30: 366–367.
12. Iinuma, M., Tosa, H., Tanaka, T. and Yonemori, S. 1994. *Phytochemistry* 35: 527–532.

13. Iinuma, M., Tosa, H., Tanaka, T. and Yonemori, S. 1995. *Phytochemistry* 38: 725–728.
14. Khan, N. U. D., Parveen, N., Singh, M. P., Singh, R., Achari, B., Dastidar, P. P. G. and Dutta, P. K. 1996. *Phytochemistry* 42: 1181–1183.
15. Spino, C., Dodier, M. and Sotheeswaran, S. 1998. *Bioorganic & Medicinal Chemistry Letters* 8: 3475–3478.
16. Muhammad, S. A., Shaukat, M., Shaista, P., Viqar, U. A. and Ghazala, H. R. 1999. *Phytochemistry* 50: 1385–1389.
17. Yao, C. F. and Zeng, H. P. 2000. *Huanan Shifan Daxue Xuebao, Ziran Kexueban* 3: 62–64.
18. Joshi, S. P., Deodhar, V. B. and Phalgune, U. D. 2000. *Indian Journal of Chemistry, Section B: Organic Chemistry Including Medicinal Chemistry* 39B: 560–561.
19. Itoigawa, M., Ito, C., Tan, H. T. W., Kuchide, M., Tokuda, H., Nishino, H. and Furukawa, H. 2001. *Cancer Letters* 169: 15–19.
20. Shen, Y. C., Hung, M. C., Wang, L. T. and Chen, C. Y. 2003. *Chemical and Pharmaceutical Bulletin Tokyo* 51: 802–806.
21. Wu, Y., Zhang, P. C., Chen, R. Y., Yu, D. Q. and Liang, X. T. 2003. *Huaxue Xuebao* 61: 1047–1051.
22. Cheng, H. C., Wang, L. T., Khalil, A. T., Chang, Y. T., Lin, Y. C. and Shen, Y. C. 2004. *Journal of the Chinese Chemical Society, Taipei, Taiwan* 51: 431–435.
23. Sekino, E., Kumamoto, T., Tanaka, T., Ikeda, T. and Ishikawa, T. 2004. *Journal of Organic Chemistry* 69: 2760–2767.
24. Yimdjo, M. C., Azebaze, A. G., Nkengfack, A. E., Meyer, A. M., Bodo, B. and Fomum, Z. T. 2004. *Phytochemistry* 65: 2789–2795.
25. Laure, F., Herbette, G., Faure, R., Bianchini, J. P., Raharivelomanana, P. and Fogliani, B. 2005. *Magnetic Resonance in Chemistry: MRC* 43: 65–68.
26. Crane, S., Aurore, G., Joseph, H., Mouloungui, Z. and Bourgeois, P. 2005. *Phytochemistry* 66: 1825–1831.
27. Ee, G. C., Kua, A. S., Lim, C. K., Jong, V. and Lee, H. L. 2006. *Natural Product Research* 20: 485–491.
28. Li, Y. Z., Li, Z. L., Hua, H. M., Li, Z. G. and Liu, M. S. 2007. *Zhongguo Zhong Yao Za Zhi* 32: 692–694.
29. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
30. Woodley, E. (Ed.). 1991. Medicinal plants of Papua New Guinea, part I. Veerlag Josef Margraf Scientific Books: Weikersheim.
31. Sundaram, B. M., Gopalkrishnan, C. and Subramanian, S. 1986. *Arogya, Manipal, India* 12: 48–49.
32. Ali, M. S., Mahmud, S., Perveen, S., Rizwani, G. H. and Ahmad, V. U. 1999. *Journal of the Chemical Society of Pakistan* 21: 174–178.
33. Gopalakrishnan, C., Shankaranarayanan, D., Nazimudeen, S. K., Viswanathan, S. and Kameswaran, L. 1980. *Indian Journal of Pharmacology* 12: 181–191.
34. Jantan, I. B., Jalil, J. and Abd, W. N. M. 2001. *Pharmaceutical Biology* 39: 243–246.
35. Oku, H., Ueda, Y., Iinuma, M. and Ishiguro, K. 2005. *Planta Medica* 71: 90–92.
36. Patil, A. D., Freyer, A. J., Eggleston, D. S., Haltiwanger, R. C., Bean, M. F., Taylor, P. B., Caranfa, M. J., Breen, A. L., Bartus, H. R., Johnson, R. K., Hertzberg, R. P. and Westley, J. W. 1993. *Journal of Medicinal Chemistry* 36: 4131–4138.
37. Clercq, E. D. 2000. *Medicinal Research Reviews* 20: 323–349.
38. Ishikawa, T. 2000. *Heterocycles* 53: 453–474.
39. Said, T., Dutot, M., Martin, C., Beaudeux, J. L., Boucher, C., Enee, E., Baudouin, C., Warnet, J. M. and Rat, P. 2007. *European Journal of Pharmaceutical Sciences* 30: 203–210.
40. Ravelonjato, B., Libot, F., Ramiandrasoa, F., Kunesch, N., Gayral, P. and Poisson, J. 1992. *Planta Medica* 58: 51–55.
41. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.
42. Ajayi, I. A., Oderinde, R. A., Talwo, V. O. and Agbedana, E. O. 2008. *Food Chemistry* 106: 458–465.

Cananga odorata (Lam.) Hook. f. & Th.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.

2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?8805 Date Accessed: 30th June 2008
5. Caloprisco, E., Fourneron, J. D., Faure, R. and Demarne, F. E. 2002. *Journal of Agriculture and Food Chemistry* 50: 78–80.
6. Phan, T. S., Phan, M. G. and Nguyen, D. H. 2001. *Tap Chi Duoc Hoc* 7: 9–11.
7. Hsieh, T. J., Chang, F. R., Chia, Y. C., Chen, C. Y., Chiu, H. F. and Wu, Y. C. 2001. *Journal of Natural Products* 64: 616–619.
8. Kamarudin, M. S. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 80–87. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
9. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
10. Baratta, M. T., Dorman, H. J. D., Deans, S. G., Figueiredo, A. C., Barroso, J. G. and Ruberto, G. 1998. *Flavour and Fragrance Journal* 13: 235–244.
11. Rahman, M. M., Lopa, S. S., Sadik, G., Harun-Or-Rashid, Islam, R., Khondkar, P., Alam, A. H. and Rashid, M. A. 2005. *Fitoterapia* 76: 758–761.
12. Kluza, J., Clark, A. M. and Bailly, C. 2003. *Annals of the New York Academy of Sciences* 1010: 331–334.
13. Hongratanaworakit, T. and Buchbauer, G. 2004. *Planta Medica* 70: 632–636.
14. Woo, S. H., Reynolds, M. C., Sun, N. J., Cassady, J. M. and Snapka, R. M. 1997. Inhibition of topoisomerase II by liriodenine *Biochemical Pharmacology* 54: 467–473.
15. Kluza, J., Mazinghien, R., Degardin, K., Lansiaux, A. and Bailly, C. 2005. *European Journal of Pharmacology* 525: 32–40.
16. Chu, D. M., Miles, H., Toney, D., Ngyuen, C. and Marciano-Cabral, F. 1998. *Parasitology Research* 84: 746–752.
17. Anitha, P. and Indira, M. 2006. *Indian Journal of Experimental Biology* 44: 976–980.
18. Burdock, G. A. and Carabin, I. G. 2008. *Food and Chemical Toxicology* 46: 433–445.

Capsicum annuum L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
5. eFloras.org 2008 http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200020513 Date Accessed: 30th March 2008
6. Uquiche, E., del Valle, J. M. and Ortiz, J. 2004. *Journal of Food Engineering* 65: 55–66.
7. Mazida, M. M., Salleh, M. M. and Osman, H. 2005. *Journal of Food Composition and Analysis* 18: 427–437.
8. Materska, M., Piacente, S., Stochmal, A., Pizza, C., Oleszek, W. and Perucka, I. 2003. *Phytochemistry* 63: 893–898.
9. Yahara, S., Ura, T., Sakamoto, C. and Nohara, T. 1994. *Phytochemistry* 37: 831–835.
10. Ochi, T., Takaishi, Y., Kogure, K. and Yamauti, I. 2003. *Journal of Natural Products* 66: 1094–1096.
11. Iorizzi, M., Lanzotti, V., Ranalli, G., De Marino, S. and Zollo, F. 2002. *Journal of Agricultural and Food Chemistry* 50: 4310–4316.
12. Lee, J. H., Kiyota, N., Ikeda, T. and Nohara, T. 2007. *Chemical and Pharmaceutical Bulletin* 55: 1151–1156.
13. Lee, J. H., Kiyota, N., Ikeda, T. and Nohara, T. 2006. *Chemical and Pharmaceutical Bulletin (Tokyo)* 54: 1365–1369.
14. Nagy, V., Agócs, A., Turcsi, E., Molnár, P., Szabó, Z. and Deli, J. 2007. *Tetrahedron Letters* 48: 9012–9014.

15. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
16. Sumner J. 2000. The Natural History of Medicinal Plants. Timber Press, Portland.
17. Blumenthal, M., Goldberg, A. and Gruenwald, J. (Eds.). 2000. Herbal Medicine. Expanded Commission E Monographs. Integrative Medicine Communications: Austin, Texas.
18. Careaga, M., Fernández, E., Dorantes, L., Mota, L., Jaramillo, M. E. and Hernandez-Sanchez, H. 2003. *International Journal of Food Microbiology* 83: 331–335.
19. Kuda, T., Iwai, A. and Yano, T. 2004. *Food and Chemical Toxicology* 42: 1695–1700.
20. Ribeiro, S. F. F., Carvalho, A. O., Cunha M. D., Rodrigues, R., Cruz, L. P., Melo, V. M. M., Vasconcelos, I. M., Melo, E. J. T. and Gomes, V. M. 2007. *Toxicon* 50: 600–611.
21. Diz, M. S. S., Carvalho, A. O., Rodrigues, R., Neves-Ferreira, A. G. C., Cunha, M. D., Alves, E. W., Okorokova-Façanha, A. L., Oliveira, M. A., Perales, J., Machado, O. L. T. and Gomes, V. M. 2006. *Biochimica et Biophysica Acta* 1760: 1323–1332.
22. Han, S. S., Keum, Y. S., Chun, K. S. and Surh, Y. J. 2002. *Archives of Pharmacal Research* 25: 475–479.
23. Maoka, T., Mochida, K., Kozuka, M., Ito, Y., Fujiwara, Y., Hashimoto, K., Enjo, F., Ogata, M., Nobukuni, Y., Tokuda, H. and Nishino, H. 2001. *Cancer Letters* 172: 103–109.
24. Maoka, T., Goto, Y., Isobe, K., Fujiwara, Y., Hashimoto, K. and Mochida, K. 2001. *Journal of Oleo Science* 50: 663–665.
25. Sun, T., Xu, Z., Wu, C.-T., Janes, M., Prinyawiwatkul, W. and No, H. K. 2007. *Journal of Food Science* 72: 98–102.
26. Materska, M. and Perucka, I. 2005. *Journal of Agriculture and Food Chemistry* 53: 1750–1756.
27. Frischkorn, C. G., Frischkorn, H. E. and Carrazzoni, E. 1978. *Naturwissenschaften* 65: 480–483.
28. Kwon, M. J., Song, Y. S., Choi, M. S. and Song, Y. O. 2003. *Clinica Chimica Acta* 332: 37–44.
29. Takano, F., Yamaguchi, M., Takada, S., Shoda, S., Yahagi, N., Takahashi, T. and Ohta T. 2007. *Life Sciences* 80: 1553–1563.
30. El Hamss, R., Idaomar, M., Alonso-Moraga, A. and Munoz Serrano, A. 2003. *Food and Chemical Toxicology* 41: 41–47.
31. Laohavechvanich, P., Kangsadalampai, K., Tirawanchai, N. and Ketterman, A. J. 2006. *Food and Chemical Toxicology* 44: 1348–1354.
32. Antonious, G. F., Meyer, J. E. and Snyder, J. C. 2006. *Journal of Environmental Science and Health B* 41: 1383–1391.
33. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
34. Skidmore-Roth L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
35. Johnson, W. Jr. 2007. *International Journal of Toxicology* 26: 3–106.
36. Lininger, S. W., Gaby, A. R., Austin, S., Batz, F., Yarnell, E., Brown, D. J. and Constantine, G. 1999. A–Z guide to drug-herb-vitamin interactions: how to improve your health and avoid problems when using common medications and natural supplements together. Prima Health: Rocklin, CA.

Cassia fistula L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Palaniswamy, U. R. 2003. A guide to Medicinal Plants of Asian origin and culture. CPL Press: United States of America.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
5. Sircar, P. K., Dey, B., Sanyal, T., Ganguly, S. N. and Sircar, S. M. 1970. *Phytochemistry* 9: 735–736.
6. Murty, V. K., Rao, T. V. P. and Venkateswarlu, V. 1967. *Tetrahedron* 23: 515–518.
7. Kuo, Y. H., Lee, P. H. and Wein, Y. S. 2002. *Journal of Natural Products, Technical Note* 65: 1165–1167.

8. Yadava, R. N. and Verma, V. 2003. *Journal of Asian Natural Product Research* 5: 57–61.
9. Chowdhury, S. A. C., Mustafa, A. K. M., Alam, M. N., Gafur, M. A., Ray, B. K., Ahmed, K. and Faruq, O. 1996. *Bangladesh Journal of Scientific and Industrial Research* 31: 91–97.
10. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
11. Kumar, V. P., Chauhan, N. S., Padh, H. and Rajani, M. 2006. *Journal of Ethnopharmacology* 107: 182–188.
12. Perumal Samy, R., Ignacimuthu, S. and Sen, A. 1998. *Journal of Ethnopharmacology* 62: 173–182.
13. Duraipandiyan, V. and Ignacimuthu, S. 2007. *Journal of Ethnopharmacology* 112: 590–594.
14. Ali, N. H., Kazmi, S. U. and Faizi, S. 2007. *Pakistan Journal of Pharmaceutical Sciences* 20: 140–145.
15. Ingkaninan, K., Temkitthawon, P., Chuenchom, K., Yuyaem, T. and Thongnoi, W. 2003. *Journal of Ethnopharmacology* 89: 261–264.
16. Yadav, R. and Jain, G. C. 1999. *Advances in Contraception* 15: 293–301.
17. Sunil Kumar, K. C. and Klaus, M. 1998. *Phytotherapy Research* 12: 526–528.
18. Gupta, M., Mazumder, U. K., Rath, N. and Mukhopadhyay, D. K. 2000. *Journal of Ethnopharmacology* 72: 151–156.
19. Munasinghe, T. C. J., Seneviratne, C. K., Thabrew, M. I. and Abeysekera, A. M. 2001. *Phytotherapy Research* 15: 519–523.
20. Siddhuraju, P., Mohan, P. S. and Becker, K. 2002. *Food Chemistry* 79: 61–67.
21. Prakash, D., Suri, S., Upadhyay, G. and Singh, B. N. 2007. *International Journal of Food Sciences and Nutrition* 58: 18–28.
22. Manonmani, G., Bhavapriya, V., Kalpana, S., Govindasamy, S. and Apparanantham, T. 2005. Antioxidant activity of *Cassia fistula* (Linn.) flowers in alloxan induced diabetic rats. *Journal of Ethnopharmacology* 97: 39–42.
23. Mazumder, U. K., Gupta, M. and Rath, N. 1998. *Phytotherapy Research* 12: 520–522.
24. Bhakta, T., Mukherjee, P. K., Mukherjee, K., Banerjee, S., Mandal, S. C., Maity, T. K., Pal, M. and Saha, B. P. 1999. *Journal of Ethnopharmacology* 66: 277–282.
25. Pradeep, K., Mohan, C. V., Gobianand, K. and Karthikeyan, S. 2006. *Chemico-Biological Interactions* 167: 12–18.
26. Pradeep, K., Mohan, C. V., Anand, K. G. and Karthikeyan, S. 2005. *Indian Journal of Experimental Biology* 43: 526–530.
27. el-Saadany, S. S., el-Massry, R. A., Labib, S. M. and Sitohy, M. Z. 1991. *Die Nahrung* 35: 807–815.
28. Sartorelli, P., Andrade, S. P., Melhem, M. S., Prado, F. O. and Tempone, A. G. 2007. *Phytotherapy Research* 21: 644–647.
29. Govindarajan, M., Jebanesan, A. and Pushpanathan, T. 2008. *Parasitology Research* 102: 289–292.
30. Bhakta, T., Mukherjee, P. K., Mukherjee, K., Pal, M. and Saha, B. P. 1998. *Natural Product Sciences* 4: 84–87.
31. Senthil Kumar, M., Sripriya, R., Vijaya Raghavan, H. and Sehgal, P. K. 2006. *Journal of Surgical Research* 131: 283–289.
32. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
33. de Smet, P. A. G. M., Keller, K., Hasen, R. and Chandler, R. F. 1993. Adverse Effects of Herbal Drugs, Volume 2. Springer-Verlag: Berlin.

Catharanthus roseus (L.) G. Don

1. Hsuan, K., Chin, S. C. and Tan, H. T. W. 1998. The Concise Flora of Singapore volume II: Monocotyledons. Singapore University Press: Singapore.
2. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
3. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200018366
   Date Accessed: 30th March 2008
4. Chan, G. L., Chang, S., Chow, K. K. G., Goh, K. L. H. and Hoo, K. 1998. A Guide to Toxic Plants of Singapore. Singapore Science Centre Publication: Singapore.

5. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
6. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
7. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
8. Contin, A., Heijden, R. V. D. and Verpoorte, R. 1999. *Plant Science* 147: 177–183.
9. Gilles, B., Marie-Geneviève, D., Bruno, D. and Anne-Marie, M. 1999. *Phytochemistry* 50: 167–169.
10. Atta-ur-Rahman and Fatimak, J. A. 1984. *Tetrahedron Letters* 25: 6051–6054.
11. Atta-ur-Rahman, Bashir, M., Kaleem, S. and Fatima, T. 1983. *Phytochemistry* 22: 1021–1023.
12. Sumner, J. 2000. The Natural History of Medicinal Plants. Timber Press, Portland.
13. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
14. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
15. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
16. Goh, S. H. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 7–26. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
17. Koh, Y. K., Chang, P. and Geh, S. L. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 27–31. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
18. Ueda, J. Y., Yasuhiro, T., Hari, B. A., Quan, L. T., Kim, T. Q., Yuko, H., Ikuo, S. and Shigetoshi, K. 2002. *Biological & Pharmaceutical Bulletin* 25: 753–760.
19. Haque, N., Chowdhury, S. A. R., Nutan, M. T. H., Rahman, G. M. S., Rahman, K. M. and Rashid, M. A. 2000. *Fitoterapia* 71: 547–552.
20. Zheng, W. and Wang, S. Y. 2001. *Journal of Agricultural and Food Chemistry* 49: 5165–5170.
21. Wang, S. S., Zheng, Z. G., Weng, Y. Q., Yu, Y. J., Zhang, D. F., Fan, W. H., Dai, R. H. and Hu, Z. B. 2004. *Life Sciences* 74: 2467–2478.
22. Mohanta, B., Sudarshan, M., Boruah, M. and Chakraborty, A. 2007. *Biological Trace Element Research* 117: 139–151.
23. Singh, S. N., Vats, P., Suri, S., Shyam, R., Kumria, M. M. L., Ranganathan, S. and Sridharan, K. 2001. *Journal of Ethnopharmacology* 76: 269–277.
24. Chattopadhyay, R. R. 1999. *Journal of Ethnopharmacology* 67: 367–372.
25. Nammi, S., Boini, M. K., Lodagala, S. D. and Behara, R. B. 2003. *BMC Complementary and Alternative Medicine* 3: 4.
26. Elgorashi, E. E., Taylor, J. L. S., Maes, A., Staden, J. V., Kimpe, N. D. and Verschaeve, L. 2003. *Toxicology Letters* 143: 195–207.
27. Ahmed, A. U., Ferdous, A. H., Saha, S. K., Nahar, S., Awal, M. A. and Parvin, F. 2007. *Mymensingh Medical Journal* 16: 143–148.
28. Rau, O., Wurglics, M., Dingermann, T., Abdel-Tawab, M. and Schubert-Zsilavecz, M. 2006. *Pharmazie* 61: 952–956.
29. Nayak, B. S., Anderson, M. and Pinto Pereira, L. M. 2007. *Fitoterapia* 78: 540–544.
30. Nayak, B. S. and Pinto Pereira, L. M. 2006. *BMC Complementary and Alternative Medicine* 6: 41.
31. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of medicinal herbs. CRC Press: Boca Raton, FL.
32. van Wyk, B. E. and Wink, M. 2004. Medicinal plants of the world: an illustrated scientific guide to important medicinal plants and their uses. Timber Press: Portland.
33. Ghosh, D., Roy, I., Chanda, S. and Gupta-Bhattacharya, S. 2007. *Annals of Agricultural and Environmental Medicine* 14: 39–43.
34. Sweetman, S. C. (Ed). 2007. Martindale: the complete drug reference, 35th edition, volume one. Pharmaceutical Press: London.

Celosia argentea L.

1. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
5. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
6. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=200006992
   Date Accessed: 30th March 2008
7. Suzuki, H., Morita, H., Iwasaki, S. and Kobayashi, J. 2003. *Tetrahedron* 59: 5307–5315.
8. Suzuki, H., Morita, H., Shiro, M. and Kobayashi, J. 2004. *Tetrahedron* 60: 2489–2495.
9. Morita, H., Suzuki, H. and Kobayashi, J. 2004. *Journal of Natural Products* 67: 1628–1630.
10. Morita, H., Shimbo, K., Shigemori, H. and Kobayashi, J. 2000. *Bioorganic and Medicinal Chemistry Letters* 10: 469–471.
11. Hase, K., Kadota, S., Basnet, P., Namba, T. and Takahashi, T. 1996. *Phytotherapy Research* 10: 387–392.
12. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
13. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
14. Keys, J. D. 1976. Chinese herbs: their botany, chemistry, and pharmacodynamics. Charles E. Tuttle Co. of Rutland: Vermont & Tokyo, Japan.
15. Gnanamani, A., Shanmuga Priya, K., Radhakrishnan, N. and Babu, M. 2003. *Journal of Ethnopharmacology* 86: 59–61.
16. Kobayashi, J., Suzuki, H., Shimbo, K., Takeya, K. and Morita, H. 2001. *Journal of Organic Chemistry* 66: 6626–6633.
17. Hayakawa, Y., Fujii, H., Hase, K., Ohnishi, Y., Sakukawa, R., Kadota, S., Namba, T. and Saiki, I. 1998. *Biological and Pharmaceutical Bulletin* 21: 1154–1159.
18. Vetrichelvan, T., Jegadeesan, M. and Devi, B. A. 2002. *Biological and Pharmaceutical Bulletin* 25: 526–528.
19. Hase, K., Kadota, S., Basnet, P., Takahashi, T. and Namba, T. 1996. Protective effect of celosian, an acidic polysaccharide, on chemically and immunologically induced liver injuries. *Biological and Pharmaceutical Bulletin* 19: 567–572.
20. Hase, K., Basnet, P., Kadota, S. and Namba, T. 1997. *Planta Medica* 63: 216–219.
21. Iwalewa, E. O., Adewunmi, C. O., Omisore, N. O., Adebanji, O. A., Azike, C. K., Adigun, A. O., Adesina, O. A. and Olowoyo, O. G. 2005. *Journal of Medical Food* 8: 539–544.
22. Priya, K. S., Arumugam, G., Rathinam, B., Wells, A. and Babu, M. 2004. *Wound Repair and Regeneration* 12: 618–625.
23. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.

Centella asiatica (L.) Urban

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
3. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=5&taxon_id=200015478
   Date Accessed: 30th March 2008
4. Goh, S. H. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur,

pp. 7–26. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
5. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
6. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
7. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
8. Park, B. C., Bosire, K. O., Lee, E. S., Lee, Y. S. and Kim, J. A. 2005. *Cancer Letters* 218: 81–90.
9. Cheng, C. L., Guo, J. S., Luk, J. and Koo, M. W. L. 2004. *Life Sciences* 74: 2237–2249.
10. Inamdar, P. K., Yeole, R. D., Ghogare, A. B. and de Souza, N. J. 1996. *Journal of Chromatography A* 742: 127–130.
11. Yu, Q. L., Duan, H. Q., Takaishi, Y. and Gao, W. Y. 2006. *Molecules* 11: 661–665.
12. Yu, Q. L., Gao, W. Y., Zhang, Y. W., Teng, J. and Duan, H. Q. 2007. *Zhongguo Zhong Yao Za Zhi* 32: 1182–1184.
13. Govindan, G., Sambandan, T. G., Govindan, M., Sinskey, A., Vanessendelft, J., Adenan, I. and Rha, C. K. 2007. *Planta Medica* 73: 597–599.
14. Siddiqui, B. S., Aslam, H., Ali, S. T., Khan, S. and Begum, S. 2007. *Journal of Asian Natural Product Research* 9: 407–414.
15. Zheng, C. J. and Qin, L. P. 2007. *Zhong Xi Yi Jie He Xue Bao* 5: 348–351.
16. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
17. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
18. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
19. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
20. Zaidan, M. R., Noor Rain, A., Badrul, A. R., Adlin, A., Norazah, A. and Zakiah, I. 2005. *Tropical Biomedicine* 22: 165–170.
21. Chen, Y., Han, T., Rui, Y., Yin, M., Qin, L. and Zheng, H. 2005. *Zhong Yao Cai* 28: 492–496.
22. Shobi, V. and Goel, H. C. 2001. *Physiology & Behavior* 73: 19–23.
23. Babu, T. D., Kuttan, G. and Padikkala, J. 1995. *Journal of Ethnopharmacology* 48: 53–57.
24. Bunpo, P., Kataoka, K., Arimochi, H., Nakayama, H., Kuwahara, T., Bando, Y., Izumi, K., Vinitketkumnuen, U. and Ohnishi, Y. 2004. *Food and Chemical Toxicology* 42: 1987–1997.
25. Govindan, G., Sambandan, T. G., Govindan, M., Sinskey, A., Vanessendelft, J., Adenan, I. and Rha, C. K. 2007. *Planta Medica* 73: 597–599.
26. Yoshida, M., Fuchigami, M., Nagao, T., Okabe, H., Matsunaga, K., Takata, J., Karube, Y., Tsuchihashi, R., Kinjo, J., Mihashi, K. and Fujioka, T. 2005. *Biological and Pharmaceutical Bulletin* 28: 173–175.
27. Jayashree, G., Muraleedhara, G. K., Sudarslal, S. and Jacob, V. B. 2003. *Fitoterapia* 74: 431–434.
28. Zainol, M. K., Abd-Hamid, A., Yusof, S. and Muse, R. 2003. *Food Chemistry* 81: 575–581.
29. Ramanathan, M., Sivakumar, S., Anandvijayakumar, P. R., Saravanababu, C. and Pandian, P. R. 2007. *Indian Journal of Experimental Biology* 45: 425–431.
30. Gupta, R. and Flora, S. J. 2006. *Journal of Applied Toxicology* 26: 213–222.
31. Subathra, M., Shila, S., Devi, M. A. and Panneerselvam, C. 2005. *Experimental Gerontology* 40: 707–715.
32. Bajpai, M., Pande, A., Tewari, S. K. and Prakash, D. 2005. *International Journal of Food Science and Nutrition* 56: 287–291.
33. Satake, T., Kamiya, K., An, Y., Oishi Nee Taka, T. and Yamamoto, J. 2007. *Biological and Pharmaceutical Bulletin* 30: 935–940.
34. Awad, R., Levac, D., Cybulska, P., Merali, Z., Trudeau, V. L. and Arnason, J. T. 2007.

Canadian Journal of Physiology and Pharmacology 85: 933–942.
35. Wijeweera, P., Arnason, J. T., Koszycki, D. and Merali, Z. 2006. *Phytomedicine* 13: 668–676.
36. Cheng, C. L. and Koo, M. W. L. 2000. *Life Sciences* 67: 2647–2653.
37. Punturee, K., Wild, C. P. and Vinitketkumneun, U. 2004. *Journal of Ethnopharmacology* 95: 183–189.
38. Punturee, K., Wild, C. P., Kasinrerk, W. and Vinitketkumnuen, U. 2005. *Asian Pacific Journal of Cancer Prevention* 6: 396–400.
39. Siddique, Y. H., Ara, G., Beg, T., Faisal, M., Ahmad, M. and Afzal, M. 2007. *Toxicology in vitro* 22: 10–17.
40. Soumyanath, A., Zhong, Y. P., Gold, S. A., Yu, X., Koop, D. R., Bourdette, D. and Gold, B. G. 2005. *Journal of Pharmacy and Pharmacology* 57: 1221–1229.
41. Sharma, J. and Sharma, R. 2002. *Phytotherapy Research* 16: 785–786.
42. Sharma, R. and Sharma, J. 2005. *Phytotherapy Research* 19: 605–611.
43. Shukla, A., Rasik, A. M., Jain, G. K., Shankar, R., Kulshrestha, D. K. and Dhawan, B. N. 1999. *Journal of Ethnopharmacology* 65: 1–11.
44. Shetty, B. S., Udupa, S. L., Udupa, A. L. and Somayaji, S. N. 2006. *International Journal of Lower Extremity Wounds* 5: 137–143.
45. van Wyk, B. E. and Wink, M. 2004. Medicinal plants of the world: an illustrated scientific guide to important medicinal plants and their uses. Timber Press: Portland.
46. Arpaia, M. R., Ferrone, R., Amitrano, M., Nappo, C., Leonardo, G. and del Guercio, R. 1990. *Journal of Clinical Pharmacology Research* 10: 229–233.
47. Belcaro, G. V., Grimaldi, R. and Guidi, G. 1990. *Angiology* 41: 12–18.
48. Pointel, J. P., Boccalon, H., Cloarec, M. Ledevehat, C. and Joubert, M. 1987. *Angiology* 38(1 Pt 1): 46–50.
49. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
50. Skidmore-Roth L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
51. Jorge, O. A. and Jorge, A. D. 2005. *Revista española de enfermedades digestivas* 97: 115–124.
52. Lininger, S. W., Gaby, A. R., Austin, S., Batz, F., Yarnell, E., Brown, D. J. and Constantine, G. 1999. A-Z guide to drug-herb-vitamin interactions: how to improve your health and avoid problems when using common medications and natural supplements together. Prima Health: Rocklin, CA.

Cerbera odollam Gaertn.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
5. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?9932
Date Accessed: 30th June 2008
6. Abe, F., Yamauchi, T. and Wan, A. S. C. 1988. *Phytochemistry* 27: 3627–3631.
7. Chantrapromma, S., Usman, A., Fun, H. K., Laphookhieo, S., Karalai, C., Rat-a-pa, Y. and Chantrapromma, K. 2003. *Acta Crystallographica C* 59: 68–70.
8. Laphookhieo, S., Cheenpracha, S., Karalai, C., Chantrapromma, S., Rat-a-pa, Y., Ponglimanont, C. and Chantrapromma, K. 2004. *Phytochemistry* 65: 507–510.
9. Rao, E. V. and Rao, M. A. 1976. *Phytochemistry* 15: 848.
10. Shen, L. R., Jin, S. M., Yin, B. W., Du, X. F., Wang, Y. L. and Huo, C. F. 2007. *Chemistry & Biodiversity* 4: 1438–1449.
11. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
12. Yamauchi, T., Abe, F. and Wan, A. S. C. 1987. *Chemical & Pharmaceutical Bulletin* 35: 4813–4818.

13. Chopra, R. N., Nayar, S. L. and Chopra, I. C. 1956. Glossary of Indian Medicinal Plants. Council of Scientific and Industrial Research: New Delhi.
14. Hiên, T. T. M., Navarro-Delmasure, C. and Vy, T. 1991. *Journal of Ethnopharmacology* 34: 201–206.
15. Gaillard, Y., Krishnamoorthy, A. and Bevalot, F. 2004. *Journal of Ethnopharmacology* 95: 123–126.
16. Iyer, G. V. and Narendranath, M. 1975. *Indian Journal of Medical Research* 63: 321–324.
17. Narendranathan, M., Krishna Das, K. V. and Vijayaraghavan, G. 1975. *Indian Heart Journal* 27: 283–286.

Cissus quadrangularis L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=5&taxon_id=242444305
   Date Accessed: 30th March 2008
3. Adesanya, S. A., Nia, R., Marie-Therese, M., Boukamcha, N., Montagnac, A. and Paies, M. 1999. *Journal of Natural Products* 62: 1694–1695.
4. Bhutani, K. K., Kapoor, R. and Atal, C. K. 1984. *Phytochemistry* 23: 407–410.
5. Gupta, M. M. and Verma, R. K. 1991. *Phytochemistry* 30: 875–878.
6. Gupta, M. M. and Verma, R. K. 1990. *Phytochemistry* 29: 336–337.
7. Singh, G., Rawat, P. and Maurya, R. 2007. *Natural Product Research* 21: 522–528.
8. Bah, S., Diallo, D., Dembélé, S. and Paulsen, B. S. 2006. *Journal of Ethnopharmacology* 105: 387–399.
9. Chidambara Murthy, K. N., Vanitha, A., Mahadeva Swamy, M. and Ravishankar, G. A. 2003. *Journal of Medicinal Food* 6: 99–105.
10. Panthong, A., Supraditaporn, W., Kanjanapothi, D., Taesotikul, T. and Reutrakul, V. 2007. *Journal of Ethnopharmacology* 110: 264–270.
11. Bah, S., Jäger, A. K., Adsersen, A., Diallo, D. and Paulsen, B. S. 2007. *Journal of Ethnopharmacology* 110: 451–457.
12. Jainu, M. and Devi, C. S. 2004. *Journal of Medicinal Food* 7: 372–376.
13. Jainu, M., Mohan, K. V. and Devi, C. S. 2006. *Journal of Ethnopharmacology* 104: 302–305.
14. Jainu, M. and Shyamala Devi, C. S. 2005. *Journal of Herbal Pharmacotherapy* 5: 33–42.
15. Jainu, M., Vijai Mohan, K. and Shyamala Devi, C. S. 2006. *Indian Journal of Medical Research* 123: 799–806.
16. Jainu, M. and Devi, C. S. 2006. *Chemico-Biological Interactions* 161: 262–270.
17. Shirwaikar, A., Khan, S. and Malini, S. 2003. *Journal of Ethnopharmacology* 89: 245–250.

Cocos nucifera L.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K., Chin, S. C. and Tan, H. T. W. 1998. The Concise Flora of Singapore volume II: Monocotyledons. Singapore University Press: Singapore.
3. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=200027077
   Date Accessed: 30th March 2008
4. Ge, L. Y., Yong, J. W. H., Tan, S. N., Yang, X. H. and Ong, E. S. 2004. *Journal of Chromatography A* 48: 119–126.
5. Kooiman, P. 1971. *Carbohydrate Research* 20: 329–337.
6. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
7. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
8. Woodley, E. (Ed.). 1991. Medicinal plants of Papua New Guinea, part I. Veerlag Josef Margraf Scientific Books: Weikersheim.

9. Esquenazi, D., Wigg, M. D., Miranda, M. M. F. S., Rodrigues, H. M., Tostes, J. B. F., Rozental, S., da Silva, A. J. R. and Alviano, C. S. 2002. *Research in Microbiology* 153: 647–652.
10. Pummer, S., Heil, P., Maleck, W. and Petroianu, G. 2001. *American Journal of Emergency Medicine* 19: 287–289.
11. Alviano, D. S., Rodrigues, K. F., Leitão, S. G., Rodrigues, M. L., Matheus, M. E., Fernandes, P. D., Antoniolli, A. R. and Alviano, C. S. 2004. *Journal of Ethnopharmacology* 92: 269–273.
12. Alanís, A. D., Calzada, F., Cervantes, J. A., Torres, J. and Ceballos, G. M. 2005. *Journal of Ethnopharmacology* 100: 153–157.
13. Wang, H. X. and Ng, T. B. 2005. *Peptides* 26: 2392–2396.
14. Namasivayam, N., Vaiyapuri, M. and Venugopal, P. M. 2004. *Clinica Chimica Acta* 342: 203–210.
15. Koschek, P. R., Alviano, D. S., Alviano, C. S. and Gattass, C. R. 2007. *Brazilian Journal of Medical and Biological Research* 40: 1339–1343.
16. Mantena, S. K., Jagadish, Badduri, S. R., Siripurapu, K. B. and Unnikrishnan, M. K. 2003. *Nahrung* 47: 126–131.
17. Mendonça-Filho, R. R., Rodrigues, I. A., Alviano, D. S., Santos, A. L. S., Soares, R. M. A., Alviano, C. S., Lopes, A. H. C. S. and Rosa, M. D. S. S. 2004. *Research in Microbiology* 155: 136–143.
18. Sindurani, J. A. and Rajamohan, T. 2000. *Indian Journal of Physiology and Pharmacology* 44: 97–100.
19. Salil, G. and Rajamohan, T. 2001. *Journal of Experimental Biology* 39: 1028–1034.
20. Sandhya, V. G. and Rajamohan, T. 2006. *Journal of Medicinal Food* 9: 400–407.
21. Alleyne, T., Roache, S., Thomas, C. and Shirley, A. 2005. *West Indian Medical Journal* 54: 3–8.
22. Kirszberg, C., Esquenazi, D., Alviano, C. S. and Rumjanek, V. M. 2003. *Phytotherapy Research* 17: 1054–1058.
22. Calzada, F., Yépez-Mulia, L. and Tapia-Contreras, A. 2007. *Journal of Ethnopharmacology* 113: 248–251.
23. Wittczak, T., Pas-Wyroslak, A. and Palczynski, C. 2005. *Allergy* 60: 970–971.

Coix lacryma-jobi L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
3. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?11129
   Date Accessed: 30th June 2008
5. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
6. Targon, M. L. N., Ottoboni, L. M. M., Leite, A., Ludevid, D., Puigdomenech, P. and Arruda, P. 1992. *Plant Science* 83: 169–180.
7. Yamada, H., Yanahira, S., Kiyohara, H., Cyong, J. C. and Otsuka, Y. 1987. *Phytochemistry* 26: 3269–3275.
8. Kuo, C. C., Chiang, W., Liu, G. P., Chien, Y. L., Chang, J. Y., Lee, C. K., Lo, J. M., Huang, S. L., Shih, M. C. and Kuo, Y. H. 2002. *Journal of Agriculture and Food Chemistry* 50: 5850–5855.
9. Takahashi, M., Konno, C. and Hikino, H. 1986. *Planta Medica* 52: 64–65.
10. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
11. Keys, J. D. 1976. Chinese herbs: their botany, chemistry, and pharmacodynamics. Charles E. Tuttle Co. of Rutland: Vermont & Tokyo, Japan.
12. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
13. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.

14. Hsu, H. Y., Lin, B. F., Lin, J. Y., Kuo, C. C. and Chiang, W. 2003. *Journal of Agriculture and Food Chemistry* 51: 3763–3769.
15. Hung. W. C. and Chang, H. C. 2003. *Journal of Agriculture and Food Chemistry* 51: 7333–7337.
16. Hsu, H. Y., Lin, B. F., Lin, J. Y., Kuo, C. C. and Chiang, W. 2003. *Journal of Agriculture and Food Chemistry* 51: 3656–3660.
17. Kubota, T., Haramaki, Y., Oki, Y., Saito, F., Kuroshima, A. and Numata, M. 1988. *Nippon Nogei Kagaku Kaishi* 62: 23–28.
18. Kuo, C. C., Shih, M. C., Kuo, Y. H. and Chiang, W. 2001. *Journal of Agriculture and Food Chemistry* 49: 1564–1570.
19. Huang, T., Wu, W., Li, Y., Zhang, J and Huang, L. 2002. *Zhonghua Gan Zang Bing Za Zhi* 10: 452–454.
20. Shih, C. K., Chiang, W. C. and Kuo, M. L. 2004. *Food and Chemical Toxicology* 42: 1339–1347.
21. Bao, Y., Yuan, Y., Xia, L., Jiang, H., Wu, W. and Zhang, X. 2005. *Journal of Gastroenterology and Hepatology* 20: 1046–1053.
22. Yeh, P. H., Chiang, W. and Chiang, M. T. 2006. *International Journal for Vitamin and Nutrition Research* 76: 299–305.
23. Kim, S. O., Yun, S. J., Jung, B., Lee, E. H., Hahm, D. H., Shim, I. and Lee, H. J. 2004. *Life Sciences* 75: 1391–1404.
24. Huang, B. W., Chiang, M. T., Yao, H. T. and Chiang, W. 2005. *Phytomedicine* 12: 433–439.
25. Ko, B. J. and Lee, S. U. 1974. *Chonyonmul Hwahak Yonguso Yongu Pogo (Yongnam Taehakkyo)* 2: 19–21.
26. Tzeng, H. P., Chiang, W., Ueng, T. H. and Liu, S. H. 2005. *Journal of Toxicology and Environmental Health A* 68: 1557–1565.

Crinum asiaticum L.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K., Chin, S. C. and Tan, H. T. W. 1998. The Concise Flora of Singapore volume II: Monocotyledons. Singapore University Press: Singapore.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Wee, Y. C. and Corlett, R. 1986. The City and the Forest: Plant Life in Urban Singapore. Singapore University Press: National University of Singapore.
5. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=200028036
Date Accessed: 30th March 2008
6. Ochi, M., Otsuki, H. and Nagao, K. 1976. *Bulletin of the Chemical Society of Japan* 49: 3363–3364.
7. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
8. Ghosal, S., Saini, K. S., Razdan, S. and Kumar, Y., 1985a. *Journal of Chemical Research, Synopses* 3: 100–101.
9. Ghosal, S., Shanthy, A., Kumar, A. and Kumar, Y. 1985b. *Phytochemistry* 24: 2703–2706.
10. Ghosal, S., Shanthy, A., Mukhopadhyay, M., Sarkar, M. K. and Das, P. K. 1986. *Pharmaceutical Research* 3: 240–243.
11. Ghosal, S., Shanthy, A. and Singh, S. K. 1988. *Phytochemistry* 27: 1849–1852.
12. Kobayshi, S., Ishikawa, H., Kihara, M., Shing, T. and Uyeo, S. 1976. *Chemical and Pharmaceutical Bulletin* 24: 2553–2555.
13. Yui, S., Mikami, M., Kitahara, M. and Yamazaki, M. 1998. *Immunopharmacology* 40: 151–162.
14. Tang, R. J., Bi, N. J. and Guang, E. 1994. *Chinese Chemical Letters* 5: 855–858.
15. Min, B. S., Jiang, J., Nakamura, N., Kim, Y. H. and Hattori, M. 2001. *Chemical & Pharmaceutical Bulletin* 49: 1217–1219.
16. Phan, T. S., Tran, B. D., Phan, M. G., Nguyen, T. M., Nguyen, Q. V. and Taylor, W. C. 2002. *Tap Chi Hoa Hoc* 40: 53–58.
17. Sun, Q., Zhang, W. D., Shen, Y. H., Zhang, C. and Li H. L. 2008. *Chinese Chemical Letters* 19: 447–449.
18. Samud, A. M., Asmawi, M. Z., Sharma, J. N. and Yusof, A. P. M. 1999. *Immunopharmacology* 43: 311–316.
19. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.

20. Woodley, E. (Ed.). 1991. Medicinal plants of Papua New Guinea, part I. Veerlag Josef Margraf Scientific Books: Weikersheim.
21. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
22. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
23. Min, B. S., Kim, Y. H., Tomiyama, M., Nakamura, N., Miyashiro, H., Otake, T. and Hattori, M. 2001. *Phytotherapy Research* 15: 481–486.
24. Kim, Y. H., Park, E. J., Park, M. H., Badarch, U., Woldemichael, G. M. and Beutler, J. A. 2006. *Biological & Pharmaceutical Bulletin* 29: 2140–2142.

Cymbopogon citratus (DC.) Stapf.

1. Hsuan, K., Chin, S. C. and Tan, H. T. W. 1998. The Concise Flora of Singapore volume II: Monocotyledons. Singapore University Press: Singapore.
2. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
3. Palaniswamy, U. R. 2003. A guide to Medicinal Plants of Asian origin and culture. CPL Press: United States of America.
4. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
5. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
6. Hanson, S. W., Crawford, M., Koker, M. E. S. and Menezes, F. A. 1976. *Phytochemistry* 15: 1074–1075.
7. Chisowa, E. H., Hall, D. R. and Farman, D. I. 1998. *Flavour and Fragrance Journal* 13: 29–30.
8. Cheel, J., Theoduloz, C., Rodríguez, J. and Schmeda-Hirschmann, G. 2005. *Journal of Agriculture and Food Chemistr*y 53: 2511–2517.
9. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
10. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
11. Viana, G. S. B., Vale, T. G., Pinho, R. S. N. and Matos, F. J. A. 2000. *Journal of Ethnopharmacology* 70: 323–327.
12. Lorenzetti, B. B., Souza, G. E., Sarti, S. J., Santos Filho, D. and Ferreira, S. H. 1991. *Journal of Ethnopharmacology* 34: 43–48.
13. Muhammad, M. M., Abdul, M. A., Mohd, A. A., Rozita, M. N., Nashriyah, B. M., Abdul, R. R. and Kazuyoshi, K. 1997. *Pesticide Science* 51: 165–170.
14. Ohno, T., Kita, M., Yamaoka, Y., Imamura, S., Yamamoto, T., Mitsufuji, S., Kodama, T., Kashima, K. and Imanishi, J. 2003. *Helicobacter* 8: 207–215.
15. Ogunlana, E. O., Hoglund, S., Onawunmi, G. and Skold, O. 1987. *Microbios* 50: 43–59.
16. Adegoke, G. O. and Odesola, B. A. 1996. *International Biodeterioration & Biodegradation* 37: 81–84.
17. Betoni, J. E., Mantovani, R. P., Barbosa, L. N., Di Stasi, L. C. and Fernandes Junior, A. 2006. *Memórias do Instituto Oswaldo Cruz* 101: 387–390.
18. Wannissorn, B., Jarikasem, S., Siriwangchai, T. and Thubthimthed, S. 2005. *Fitoterapia* 76: 233–236.
19. Abe, S., Sato, Y., Inoue, S., Ishibashi, H., Maruyama, N., Takizawa, T., Oshima, H. and Yamaguchi, H. 2003. *Nippon Ishinkin Gakkai Zasshi* 44: 285–291.
20. Paranagama, P. A., Abeysekera, K. H., Abeywickrama, K. and Nugaliyadde, L. 2003. *Letters in Applied Microbiology* 37: 86–90.
21. Helal, G. A., Sarhan, M. M., Abu Shahla, A. N. and Abou El-Khair, E. K. 2007. *Journal of Basic Microbiology* 47: 5–15.
22. Helal, G. A., Sarhan, M. M., Abu Shahla, A. N. and Abou El-Khair, E. K. 2006. *Journal of Basic Microbiology* 46: 456–469.
23. Helal, G. A., Sarhan, M. M., Abu Shahla, A. N. and Abou El-Khair, E. K. 2006. *Journal of Basic Microbiology* 46: 375–386.

24. Dutta, B. K., Karmakar, S., Naglot, A., Aich, J. C. and Begam, M. 2007. *Mycoses* 50: 121–124.
25. Pawar, V. C. and Thaker, V. S. 2006. In vitro efficacy of 75 essential oils against *Aspergillus niger*. *Mycoses* 49: 316–323.
26. Kinouchi, T., Suaeyun, R., Chewonarin, T., Intiyot, Y., Arimochi, H., Kataoka, K., Akimoto, S., Vinitketkumnuen, U. and Ohnishi, Y. 1997. *Mutation Research/ Fundamental and Molecular Mechanisms of Mutagenesis* 379: S181.
27. Dudai, N., Weinstein, Y., Krup, M., Rabinski, T. and Ofir, R. 2005. *Planta Medica* 71: 484–488.
28. Tchoumbougnang, F., Zollo, P. H., Dagne, E. and Mekonnen, Y. 2005. *Planta Medica* 71: 20–23.
29. Melo, C. D. F., Soares, S. F., da Costa, R. F., da Silva, C. R. and de Oliveira, *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 496: 33–38.
30. Tapia, A., Cheel, J., Theoduloz, C., Rodríguez, J., Schmeda-Hirschmann, G., Gerth, A., Wilken, D., Jordan, M., Jiménez-González, E., Gomez-Kosky, R. and Mendoza, E. Q. 2007. *Zeitschrift für Naturforschung. C, Journal of Biosciences* 62: 447–457.
31. Tognolini, M., Barocelli, E., Ballabeni, V., Bruni, R., Bianchi, A., Chiavarini, M. and Impicciatore, M. 2006. *Life Sciences* 78: 1419–1432.
32. Puatanachokchai, R., Kishida, H., Denda, A., Murata, N., Konishi, Y., Vinitketkumnuen, U. and Nakae, D. 2002. *Cancer Letters* 183: 9–15.
33. Adeneye, A. A. and Agbaje, E. O. 2007. *Journal of Ethnopharmacology* 112: 440–444.
34. Blanco, M. M., Costa, C. A., Freire, A. O., Santos, J. G. Jr. and Costa, M. 2007. Article in Press, Available online 11th June 2007.
35. Runnie, I., Salleh, M. N., Mohamed, S., Head, R. J. and Abeywardena, M. Y. 2004. *Journal of Ethnopharmacology* 92: 311–316.
36. Vinitketkumnuen, U., Puatanachokchai, R., Kongtawelert, P., Lertprasertsuke, N. and Matsushima, T. 1994. *Mutation Research/ Genetic Toxicology* 341: 71–75.
37. Chungsamarnyart, N. and Jiwajinda, S. 1992. *Kasetsart Journal: Natural Sciences* 26: 46–51.
38. Fuentes, J. L., Alonso, A., Cuétara, E., Vernhe, M., Alvarez, N., Sánchez-Lamar, A. and Llagostera, M. 2006. *International Journal of Radiation Biology* 82: 323–329.
39. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
40. Akinboro, A. and Bakare, A. A. 2007. *Journal of Ethnopharmacology* 112: 470–475.
41. Skidmore-Roth, L. 2001. Mosby's Handbook of Herbs and Natural Supplements. Mosby, Inc.: St. Louis.

Dolichos lablab L.

1. Nguyen, V. D. 1993. Medicinal Plants of Vietnam, Cambodia and Laos. Nguyen Van Duong: Vietnam.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
5. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?104887
   Date Accessed: 30th June 2008
6. Ye, X. T., Wang, H. X. and Ng, T. B. 2000. *Biochemical and Biophysical Research Communications* 269: 155–159.
7. Yoshiki, Y., Kim, J. H., Okubo, K., Nagoya, I., Sakabe, T. and Tamura, N. 1995. *Phytochemistry* 38: 229–231.
8. Yoshikawa, M., Murakami, T., Komatsu, H. and Matsuda, H. 1998. *Chemical and Pharmaceutical Bulletin (Tokyo)* 46: 812–816.
9. Yokota, T., Kobayashi, S., Yamane, H. and Takahashi, N. 1978. *Agricultural and Biological Chemistry* 42: 1811–1812.
10. Yokota, T., Ueda, J. and Takahashi, N. 1981. *Phytochemistry* 20: 683–686.
11. Salimath, P. V. and Tharanathan, R. N. 1982. *Carbohydrate Research* 104: 341–347.

12. Salimath, P. V. and Tharanathan, R. N. 1982. *Carbohydrate Research* 106: 251–257.
13. Hamana, K., Niitsu, M., Samejima, K. and Matsuzaki, S. 1992. *Phytochemistry* 31: 893–894.
14. Ghosh, R. and Das, A. 1984. *Carbohydrate Research* 126: 287–296.
15. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
16. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
17. Kim, Y. H., Woloshuk, C. P., Cho, E. H., Bae, J. M., Song, Y. S. and Huh, G. H. 2007. *Plant Cell Reports* 26: 395–405.
18. Favero, J., Miquel, F., Dornand, J. and Mani, J. C. 1988. *Cellular Immunology* 112: 302–314.
19. Silva-Lima, M., Pusztai, A., Nunes, D. C. and Farias, M. E. 1988. *Brazilian Journal of Medical and Biological Research* 21: 219–222.
20. Guran, A., Ticha, M., Filka, K. and Kocourek, J. 1983. *Biochemical Journal* 209: 653–657.
21. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.

Elephantopus scaber L.

1. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. and Corlett, R. 1986. The City and the Forest: Plant Life in Urban Singapore. Singapore University Press: National University of Singapore.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?401632
   Date Accessed: 30th June 2008
5. Goh, S. H. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 7–26. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
6. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
7. But, P. P. H., Hon, P. M., Cao, H., Chan, T. W. D., Wu, B. M., Mak, T. C. W. and Che, C. T. 1997. *Phytochemistry* 44: 113–116.
8. DeSilva, L. B., Herath, W. H. M. W., Jennings, R. C., Mahendran, M. and Wannigama, G. E. 1982. *Phytochemistry* 21: 1173–1175.
9. Sim, K. Y. and Lee, H. T. 1969. *Phytochemistry* 8: 933–934.
10. Wiart, C. 2000. Medicinal Plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
11. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
12. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
13. Chen, C. P., Lin, C. C. and Namba, T. 1989. *Journal of Ethnopharmacology* 27: 285–295.
14. Ichikawa, H., Nair, M. S., Takada, Y., Sheeja, D. B., Kumar, M. A., Oommen, O. V. and Aggarwal, B. B. 2006. *Clinical Cancer Research* 12: 5910–5918.
15. Xu, G., Liang, Q., Gong, Z., Yu, W., He, S. and Xi, L. 2006. *Experimental Oncology* 28: 106–109.
16. Tsai, C. C. and Lin, C. C. 1998. *Journal of Ethnopharmacology* 64: 85–89.
17. Li, Y., Ooi, L. S., Wang, H., But, P. P. and Ooi, V. E. 2004. *Phytotherapy Research* 18: 718–722.
18. Lin, C. C., Tsai, C. C. and Yen, M. H. 1995. *Journal of Ethnopharmacology* 45: 113–123.
19. Rajesh, M. G. and Latha, M. S. 2001. Hepatoprotection by *Elephantopus scaber* Linn. in CCl4. *Indian Journal of Physiology and Pharmacology* 45: 481–486.

Euphorbia hirta L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?400049
   Date Accessed: 30th June 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Chen, L. 1991. *Zhongguo Zhong Yao Za Zhi* 16: 38–39.
5. Gupta, D. R. and Garg, S. K. 1966. *Bulletin of the Chemical Society of Japan* 39: 2532–2534.
6. Galvez, J., Zarzuelo, A., Crespo, M. E., Lorente, M. D., Ocete, M. A. and Jiménez, J. 1993. *Planta Medica* 59: 333–336.
7. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
8. Woodley, E. (Ed.). 1991. Medicinal plants of Papua New Guinea, part I. Veerlag Josef Margraf Scientific Books: Weikersheim.
9. Lanhers, M. C., Fleurentin, J., Dorfman, P., Mortier, F. and Pelt, J. M. 1991. *Planta Medica* 57: 225–231.
10. Vijaya, K., Ananthan, S. and Nalini, R. 1995. *Journal of Ethnopharmacology* 49: 115–118.
11. Oyewale, A. O., Mika, A. and Peters, F. A. 2002. *Global Journal of Pure and Applied Sciences* 8: 49–55.
12. Wang, Y. C. and Huang, T. L. 2005. *FEMS Immunology and Medical Microbiology* 43: 295–300.
13. Ndip, R. N., Malange Tarkang, A. E., Mbullah, S. M., Luma, H. N., Malongue, A., Ndip, L. M., Nyongbela, K., Wirmum, C. and Efange, S. M. 2007. *Journal of Ethnopharmacology* 114: 452–457.
14. Sudhakar, M., Rao, Ch. V., Rao, P. M., Raju, D. B. and Venkateswarlu, Y. 2006. *Fitoterapia* 77: 378–380.
15. Hore, S. K., Ahuja, V., Mehta, G., Kumar, P., Pandey, S. K. and Ahmad, A. H. 2006. *Fitoterapia*. 77: 35–38.
16. Martinez-Vazquez, M., Apan, T. O. R., Lazcano, M. E. and Bye, R. 1999. *Revista de la Sociedad Quimica de Mexico* 43: 103–105.
17. Singh, G. D., Kaiser, P., Youssouf, M. S., Singh, S., Khajuria, A., Koul, A., Bani, S., Kapahi, B. K., Satti, N. K., Suri, K. A. and Johri, R. K. 2006. *Phytotherapy Research* 20: 316–321.
18. Pretorius, E., Ekpo, O. E. and Smit, E. 2007. *Experimental and Toxicologic Pathology* 59: 105–114.
19. Tona, L., Cimanga, R. K., Mesia, K., Musuamba, C. T., De Bruyne, T., Apers, S., Hernans, N., Van Miert, S., Pieters, L., Totté, J. and Vlietinck, A. J. 2004. *Journal of Ethnopharmacology* 93: 27–32.
20. Tona, L., Ngimbi, N. P., Tsakala, M., Mesia, K., Cimanga, K., Apers, S., De Bruyne, T., Pieters, L., Totté, J. and Vlietinck, A. J. 1999. *Journal of Ethnopharmacology* 68: 193–203.
21. Anuradha, H., Srikumar, B. N., Shankaranarayana Rao, B. S. and Lakshmana, M. 2008. *Journal of Neural Transmission* 115: 35–42.
22. Johnson, P. B., Abdurahman, E. M., Tiam, E. A., Abdu-Aguye, I. and Hussaini, I. M. 1999. *Journal of Ethnopharmacology* 65: 63–69.
23. Lanhers, M. C., Fleurentin, J., Cabalion, P., Rolland, A., Dorfman, P., Misslin, R. and Pelt, J. M. 1990. *Journal of Ethnopharmacology* 29: 189–198.
24. Youssouf, M. S., Kaiser, P., Tahir, M., Singh, G. D., Singh, S., Sharma, V. K., Satti, N. K., Haque, S. E. and Johri, R. K. 2007. *Fitoterapia* 78: 535–539.
25. Singh, S. K., Yadav, R. P., Tiwari, S. and Singh, A. 2005. *Chemosphere* 59: 263–270.
26. Singh, S. K., Yadav, R. P., Singh, D. and Singh, A. 2004. *Environmental Toxicology and Pharmacology* 15: 87–93.

Eurycoma longifolia Jack.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons.

Singapore University Press: National University of Singapore.

3. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?417514
Date Accessed: 30th June 2008
5. Chan, K. L., O'neill, M. J., Phillipson, J. D. and Warhurst, D. C. 1986. *Planta Medica* 52: 105–107.
6. Kardono, L. B., Angerhofer, C. K., Tsauri, S., Padmawinata, K., Pezzuto, J. M. and Kinghorn, A. D. 1991. *Journal of Natural Products* 54: 1360–1367.
7. Chan, K. L., Lee, S. P., Sam, T. W., Tan, S. C., Noguchi, H. and Sankawa, U. 1991. *Phytochemistry* 30: 3138–3141.
8. Itokawa, H., Kishi, E., Morita, H. and Takeya, K. 1992. *Chemical and Pharmaceutical Bulletin* 40: 1053–1055.
9. Morita, H., Kishi, E., Takeya, K. and Itokawa, H. 1992. *Phytochemistry* 31: 3993–3995.
10. Morita, H., Kishi, E., Takeya, K., Itokawa, H. and Iitaka, Y. 1993a. *Phytochemistry* 33: 691–696.
11. Morita, H., Kishi, E., Takeya, K., Itokawa, H. and Iitaka, Y. 1993b. *Phytochemistry* 34: 765–771.
12. Chan, K. L., Iitaka, Y., Noguchi, H., Sugiyama, H., Saito, I. and Sankawa, U. 1992. *Phytochemistry* 31: 4295–4298.
13. Ang, H. H., Chan, K. L. and Mak, J. W. 1995. *Planta Medica* 61: 177–178.
14. Mitsunaga, K., Koike, K., Tanaka, T., Ohkawa, Y., Kobayashi, Y., Sawaguchi, T. and Ohmoto, T. 1994. *Phytochemistry* 35: 799–802.
15. Ang, H. H., Cheang, H. S. and Yusof, A. P. 2000. *Experimental Animals* 49: 35–38.
16. Jiwajinda, S., Santisopasri, V., Murakami, A., Hirai, N. and Ohigashi, H. 2001. *Phytochemistry* 58: 959–962.
17. Kuo, P. C., Shi, L. S., Damu, A. G., Su, C. R., Huang, C. H., Ke, C. H., Wu, J. B., Lin, A. J., Bastow, K. F., Lee, K. H. and Wu, T. S. 2003. *Journal of Natural Products* 66: 1324–1327.
18. Kuo, P. C., Damu, A. G., Lee, K. H. and Wu, T. S. 2004. *Bioorganic and Medicinal Chemistry* 12: 537–544.
19. Bedir, E., Abou-Gazar, H., Ngwendson, J. N. and Khan, I. A. 2004. *Chemical and Pharmaceutical Bulletin* 51: 1301–1303.
20. Latiff, A. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 1–6. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
21. Ang, H.H. and Cheang, H.S. 1999. *Japanese Journal of Pharmacology* 79: 497–500.
22. Farouk A. and Benafri A. 2007. *Saudi Medical Journal* 28: 477–479.
23. Jiwajinda, S., Santisopasri, V., Murakami, A., Kawanaka, M., Kawanaka, H., Gasquet, M., Eilas, R., Balansard, G. and Ohigashi, H. 2002. *Journal of Ethnopharmacology* 82: 55–58.
24. Mohd-Fuat, A. R., Aidoo, K. E. and Candlish, A. A. G. 2007. *Tropical Biomedicine* 24: 49–59.
25. Tee, T. T. and Azimahtol, H. L. 2005. *Anticancer Research* 25: 2205–2213.
26. Nurhanan, M. Y., Azimahtol, Hawariah, L. P., Mohd Ilham, A. and Mohd Shukri, M. A. 2005. *Phytotherapy Research.* 19: 994–996.
27. Li, Y., Liang, F., Jiang, W., Yu, F., Cao, R., Ma, Q., Dai, X., Jiang, J., Wang, Y. and Si, S. 2007. *Cancer Biology and Therapy* 6: 1193–1199.
28. Chan, K. L., Choo, C. Y., Abdullah, N. R. and Ismail, Z. 2004. *Journal of Ethnopharmacology* 92: 223–227.
29. Chan, K. L., Choo, C. Y. and Abdullah, N. R. 2005. *Planta Medica* 71: 967–969.
30. Mohd Ridzuan, M. A., Rain, A. N., Zhari, I. and Zakiah, I. 2005. *Tropical Biomedicine* 22: 155–163.
31. Hout, S., Chea, A., Bun, S. S., Elias, R., Gasquet, M., Timon-David, P., Balansard, G. and Azas, N. 2006. *Journal of Ethnopharmacology* 107: 12–18.
32. Mohd Ridzuan, M. A., Sow, A., Noor Rain, A., Mohd Ilham, A. and Zakiah, I. 2007. *Tropical Biomedicine* 24: 111–118.
33. Tada, H., Yasuda, F., Otani, K., Doteuchi, M., Ishara, Y. and Shiro, M. 1991. *European Journal of Medical Chemistry* 26: 345–349.
34. Widjajakusama, L. E. 1987. Androgenic Activity in Root and Stem of Pasak Bumi (*Eurycoma longifolia*) as Determined by a Modified Chick-comb Growth Method of Dorfman. XVI Pacific Science Congress, Seoul, Korea.

35. Ang, H. H. and Sim, M. K. 1998. *Archives of Pharmacal Research* 21: 779–781.
36. Ang, H. H. and Ngai, T. H. 2001a. *Fundamental and Clinical Pharmacology* 15: 265–268.
37. Ang, H. H., Ikeda, S. and Gan, E. K. 2001b. *Phytotherapy Research* 15: 435–436.
38. Ang, H. H. and Cheang, H. S. 2001c. *Archieves of Pharmacal Research* 24: 437–440.
39. Satayavivad, J., Soonthornchareonnon, N., Somanabandhu, A. and Thebtaranonth, Y. 1998. *Thai Journal of Phytopharmacy* 52: 14–27.
40. Chan, K. L. and Choo, C. Y. 2002. *Planta Medica* 68: 662–664.

Hibiscus mutabilis L.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
4. eFloras.org 2008 http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200013711 Date Accessed: 30th March 2008
5. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
6. Chauhan, J. S., Vidyapati, T. J. and Gupta, A. K. A. 1979. *Phytochemistry* 18: 1766–1767.
7. Chen, R. T., Chen, L., Wei, K. X. and Fang, S. D. 1993. *Zhongcaoyao* 24: 227–229.
8. Lowry, J. B. 1971. *Phytochemistry* 10: 673–674.
9. Vidyapati, T. J., Gupta, A. K. and Chauhan, J. S. 1979. *Indian Journal of Chemistry, Section B: Organic Chemistry Including Medicinal Chemistry* 17B: 536.
10. Yao, L. Y., Lu, Y. and Chen, Z. N. 2003. *Zhongcaoyao* 34: 201–203.
11. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
12. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
13. Fu, S. C., Zhang, F. H., Shi, W. B., Du, N., Yao, L. Y. and Lu, Y. 2001. *Shanghai Dier Yike Daxue Xuebao* 21: 14–16.

Hibiscus rosa-sinensis L.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Palaniswamy, U. R. 2003. A guide to Medicinal Plants of Asian origin and culture. CPL Press: United States of America.
3. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?19075 Date Accessed: 30th June 2008
4. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
5. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
6. Nakatani, M., Matsuoka, K., Uchio, Y. and Hase, T. 1994. *Phytochemistry* 35: 1245–1247.
7. Woodley, E. (Ed.). 1991. Medicinal plants of Papua New Guinea, part I. Veerlag Josef Margraf Scientific Books: Weikersheim.
8. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
9. Kasture, V. S., Chopde, C. T. and Deshmukh, V. K. 2000. *Journal of Ethnopharmacology* 71: 65–75.
10. Sharma, S., Khan, N. and Sultana, S. 2004. *European Journal of Cancer Prevention* 13: 53–63.
11. Kholkute, S. D. 1977. *Planta Medica* 31: 127–135.
12. Reddy, C. M., Murthy, D. R. and Patil, S. B. 1997. *Indian Journal of Experimental Biology* 35: 1170–1174.

13. Murthy, D. R., Reddy, C. M. and Patil, S. B. 1997. *Biological and Pharmaceutical Bulletin* 20: 756–758.
14. Nivsarkar, M., Patel, M., Padh, H., Bapu, C. and Shrivastava, N. 2005. *Contraception* 71: 227–230.
15. Gauthaman, K. K., Saleem, M. T., Thanislas, P. T., Prabhu, V. V., Krishnamoorthy, K. K., Devaraj, N. S. and Somasundaram, J. S. 2006. *BMC Complementary and Alternative Medicine* 6: 32.
16. Obi, F. O., Ichide, I. I., Okororo, A. A., Esiri, O. and Oghre, S. 2002. *Global Journal of Pure and Applied Sciences* 8: 259–264.
17. Sachdewa, A. and Khemani, L. D. 1999. *Biomedical and Environmental Science* 12: 222–226.
18. Sachdewa, A. and Khemani, L. D. 2003. *Journal of Ethnopharmacology* 89: 61–66.
19. Prakash, A. O. 1979. *Experientia* 35: 1122–1123.
20. Singwi, M. S. and Lall, S. B. 1981. *Indian Journal of Experimental Biology* 19: 359–362.
21. Shivananda Nayak, B., Sivachandra Raju, S., Orette, F. A. and Chalapathi Rao, A. V. 2007. *International Journal of Lower Extremity Wounds* 6: 76–81.
22. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of medicinal herbs. CRC Press: Boca Raton, FL.

Hibiscus tiliaceus L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
3. Wee, Y. C. and Corlett, R. 1986. The City and the Forest: Plant Life in Urban Singapore. Singapore University Press: National University of Singapore.
4. eFloras.org 2008 http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200013733 Date Accessed: 30th March 2008
5. Melecchi, M. I. S., Martinez, M. M., Abad, F. C., Zini, P. P., Filho, I. D. N. and Caramão, E. B. 2002. *Journal of Separation Science* 25: 86–90.
6. Chen, J. J., Huang, S. Y., Duh, C. Y., Chen, I. S., Wang, T. C. and Fang, H. Y. 2006. *Planta Medica* 72: 935–938.
7. Li, L., Huang, X., Sattler, I., Fu, H., Grabley, S. and Lin, W. 2006. *Magnetic Resonance Chemistry* 44: 624–628.
8. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
9. Masuda, T., Yamashita, D., Takeda, Y. and Yonemori, S. 2005. *Bioscience, Biotechnology, and Biochemistry* 69: 197–201.
10. Rosa, R. M., Melecchi, M. I., da Costa Halmenschlager, R., Abad, F. C., Simoni, C. R., Caramão, E. B., Henriques, J. A., Saffi, J. and de Paula Ramos, A. L. 2006. *Journal of Agricultural and Food Chemistry* 54: 7324–7330.
11. Rosa, R. M., Moura, D. J., Melecchi, M. I., dos Santos, R. S., Richter, M. F., Camarão, E. B., Henriques, J. A., de Paula Ramos, A. L. and Saffi, J. 2007. *Toxicology in vitro* 21: 1442–1452.

Impatiens balsamina L.

1. Hsuan, K. 1978. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
5. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
6. Ishiguro, K., Ohira, Y. and Oku, H. 1998. *Journal of Natural Products* 61: 1126–1129.
7. Ishiguro, K., Oku, H. and Kato, T. 2000. *Phytotherapy Research* 14: 54–56.

8. Oku, H. and Ishiguro, K. 2002. *Biological and Pharmaceutical Bulletin* 25: 658–660.
9. Shoji, N., Umeyama, A., Saitou, N., Yoshikawa, K., Nagai, M. and Arihara, S. 1994. *Chemical and Pharmaceutical Bulletin* 42: 1422–1426.
10. Shoji, N., Umeyama, A., Yoshikawa, K., Nagai, M. and Arihara, S. 1994. *Phytochemistry* 37: 1437–1441.
11. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
12. Yang, X., Summerhurst, D. K., Koval, S. F., Ficker, C., Smith, M. L. and Bernards, M. A. 2001. *Phytotherapy Research* 15: 676–680.
13. Choi, S. C. and Jung, J. S. 1999. *Journal of the Korean Fiber Society* 36: 338–343.
14. Lim, Y. H., Kim, I. H. and Seo, J. J. 2007. *Journal of Microbiology* 45: 473–477.
15. Tailor, R. H., Acland, D. P., Attenborough, S., Cammue, B. P., Evans, I. J., Osborn, R. W., Ray, J. A., Rees, S. B. and Broekaert, W. F. 1997. *Journal of Biological Chemistry* 272: 24480–24487.
16. Ishiguro, K., Ohira, Y. and Oku, H. 2002. *Biological and Pharmaceutical Bulletin* 25: 505–508.
17. Oku, H. and Ishiguro, K. 1999. *Phytotherapy Research* 13: 521–525.
18. Thevissen, K., François, I. E., Sijtsma, L., van Amerongen, A., Schaaper, W. M., Meloen, R., Posthuma-Trumpie, T., Broekaert, W. F. and Cammue, B. P. 2005. *Peptides* 26: 1113–1119.
19. Oku, H. and Ishiguro, K. 2001. *Phytotherapy Research* 15: 506–510.

Imperata cylindrica (L.) P. Beauv.

1. Nguyen, V. D. 1993. Medicinal Plants of Vietnam, Cambodia and Laos. Nguyen Van Duong: Vietnam.
2. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200025558
Date Accessed: 30th March 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
5. Matsunaga, K., Shibuya, M. and Ohizumi, Y. 1994. *Journal of Natural Products* 57: 1183–1184.
6. Matsunaga, K., Shibuya, M. and Ohizumi, Y. 1995. *Journal of Natural Products* 58: 138–139.
7. Matsunaga, K., Ikeda, M., Shibuya, M. and Ohizumi, Y. 1994. *Journal of Natural Products* 57: 1290–1293.
8. Matsunaga, K., Shibuya, M. and Ohizumi, Y. 1994. *Journal of Natural Products* 57: 1734–1736.
9. Yoon, J. S., Lee, M. K., Sung, S. H. and Kim, Y. C. 2006. *Journal of Natural Products* 69: 290–291.
10. Lee, D. Y., Han, K. M., Song, M. C., Lee, D. G., Rho, Y. D. and Baek, N. I. 2008. *Journal of Asian Natural Products Research* 10: 337–341.
11. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
12. Sripanidkulchai, B., Wongpanich, V., Laupattarakasem, P., Suwansaksri, J. and Jirakulsomchok, D. 2001. *Journal of Ethnopharmacology* 75: 185–190.
13. De Leon, R. M., Panlilio, B. G., Bamba, E. M., Bognot, F. L. and Dizon, D. S. 2001. Antibacterial properties of local medicinal plants against selected test organisms. Abstracts of Papers, 22nd ACS National Meeting, Chicago, IL, United States.
14. Koh, D., Goh, C. L., Tan, H. T., Ng, S. K. and Wong, W. K. 1997. *Contact Dermatitis* 37: 32–34.
15. Kumar, L., Sridhara, S., Singh, B. P., Gangal, S. V. 1998. *International Archieves of Allergy and Immunology* 117: 174–179.

Ipomoea pes-caprae (L.) Sweet

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons.

Singapore University Press: National University of Singapore.

3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
5. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200018852
Date Accessed: 16th April 2008
6. Amor-Prats, D. and Harborne, J. B. 1993. *Biochemical Systematics and Ecology* 21: 455–461.
7. Escobedo-Martinez, C. and Pereda-Miranda, Rogelio. 2007. *Journal of Natural Products* 70: 974–978.
8. Pereda-Miranda, R., Escalante-Sanchez, E. and Escobedo-Martinez, C. 2005. *Journal of Natural Products* 68: 226–230.
9. Pongprayoon, U., Baeckstrom, P., Jacobsson, U., Lindstrom, M. and Bohlin, L. 1991. *Planta Medica* 57: 515–518.
10. Pongprayoon, U., Baeckström, P., Jacobsson, U., Lindström, M. and Bohlin, L. 1992. *Phytotherapy Research* 6: 104–107.
11. Pongprayoon, U., Bohlin, L., Baeckström, P., Jacobsson, U. and Lindström, M. 1992. *Planta Medica* 58: 19–21.
12. Teramachi, F., Koyano, T., Kowithayakorn, T., Hayashi, M., Komiyama, K. and Ishibashi, M. 2005. *Journal of Natural Products* 68: 794–796.
13. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
14. Rogers, K. L., Grice, I. D. and Griffiths, L. R. 2000. *European Journal of Pharmaceutical Sciences* 9: 355–363.
15. Pongprayoon, U., Bohlin, L., Sandberg, F. and Wasuwat, S. 1989. *Acta Pharmaceutica Nordica* 1: 41–44.
16. Maria de Souza, M., Madeira, A., Berti, C., Krogh, R., Yunes, R. A. and Cechinel-Filho, V. 2000. *Journal of Ethnopharmacology* 69: 85–90.
17. Khan, M. M., Ahmad, F., Rastogi, A. K., Kidwai, J. R., Lakshmi, V. and Bhakuni, D. S. 1994. *Fitoterapia* 65: 231–234.
18. Pongprayoon, U., Bohlin, L., Soonthornsaratune, P. and Wasuwat, S. 1991. *Phytotherapy Research* 5: 63–66.
19. Pongprayoon, U., Bohlin, L. and Wasuwat, S. 1991. *Journal of Ethnopharmacology* 35: 65–69.

Ixora chinensis Lam.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Takeda, Y., Nishimura, H. and Inouye, H. 1975. *Phytochemistry* 14: 2647–2650.
3. Huang, M. 1990. *Phytochemistry* 29: 1317–1319.
4. Hui, W. H. and Ho, C. T. 1968. *Australian Journal of Chemistry* 21: 547–549.
5. Goh, S. H., Liew, S. L. and Soepadmo, E. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 104–110, Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
6. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
7. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.

Jatropha curcas L.

1. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200012585
Date Accessed: 30th June 2008
2. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
3. USDA, ARS GRIN Taxonomy for Plants 2008
http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?20692
Date Accessed: 30th June 2008
4. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.

5. Ravindranath, N., Ramesh, C. and Das, B. 2003. *Biochemical Systematics and Ecology* 31: 431–432.
6. Ravindranath, N., Reddy, M. R., Mahender, G., Ramu, R., Kumar, K. R. and Das, B. 2004. *Phytochemistry* 65: 2387–2390.
7. Naengchomnong, W., Thebtaranonth, Y., Wiriyachitra, P., Okamoto, K. T. and Clardy, J. 1986. *Tetrahedron Letters* 27: 2439–2442.
8. Van den Berg, A. J., Horsten, S. F, Kettenes-van den Bosch, J. J., Kroes, B. H., Beukelman, C. J., Leeflang, B. R. and Labadie, R. P. 1995. *FEBS Letters* 358: 215–218.
9. Staubmann, R., Schubert-Zsilavecz, M., Hiermann, A. and Kartnig, T. 1998. *Phytochemistry* 50: 337–338.
10. Kong, L., Min, Z. and Shi, J. 1996. *Zhiwu Xuebao* 38: 161–166.
11. Talapatra, S. K., Mandal, K. and Talapatra, B. 1993. *Journal of the Indian Chemical Society* 70: 543–548.
12. Abigor, R. D. and Uadia, P. O. 2001. *Rivista Italiana delle Sostanze Grasse* 78: 163–165.
13. Lin, J., Yan, F., Tang, L. and Chen, F. 2003. *Acta Pharmacologica Sinica* 24: 241–246.
14. Muangman, S., Thippornwong, M. and Tohtong, R. 2005. *In vivo* 19: 265–268.
15. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.
16. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
17. Rau, O., Wurglics, M., Dingermann, T., Abdel-Tawab, M. and Schubert-Zsilavecz, M. 2006. *Pharmazie* 61: 952–956.
18. Mujumdar, A. M. and Misar, A. V. 2004. *Journal of Ethnopharmacology* 90: 11–15.
19. Luo, M. J., Yang, X. Y., Liu, W. X., Xu, Y., Huang, P., Yan, F. and Chen, F. 2006. *Acta Biochimica et Biophysica Sinica (Shanghai)* 38: 663–668.
20. Ankrah, N. A., Nyarko, A. K., Addo, P. G., Ofosuhene, M., Dzokoto, C., Marley, E., Addae, M. M. and Ekuban, F. A. 2003. *Phytotherapy Research* 17: 697–701.
21. Matsuse, I. T., Lim, Y. A., Hattori, M., Correa, M. and Gupta, M. P. 1999. *Journal of Ethnopharmacology* 64: 15–22.
22. Muanza, D. N., Euler, K. L., Williams, L. and Newman, D. J. 1995. *International Journal of Pharmacognosy* 33: 98–106.
23. Osoniyi, O. and Onajobi, F. 2003. *Journal of Ethnopharmacology* 89: 101–105.
24. Goonasekera, M. M., Gunawardana, V. K., Jayasena, K., Mohammed, S. G. and Balasubramaniam, S. 1995. *Journal of Ethnopharmacology* 47: 117–123.
25. Gandhi, V. M., Cherian, K. M. and Mulky, M. J. 1995. *Food and Chemical Toxicology* 33: 39–42.
26. Staubmann, R., Ncube, I., Gubitz, G. M., Steiner, W. and Read, J. S. 1999. *Journal of Biotechnology* 75: 117–126.
27. Karmegam, N., Sakthivadivel, M., Anuradha, V. and Daniel, T. 1997. *Bioresource Technology* 59: 137–140.
28. Meshram, P. B., Kulkarni, N. and Joshi, K. C. 1996. *Journal of Environmental Biology* 17: 295–298.
29. Georges, K., Jayaprakasam, B., Dalavoy, S. S. and Nair, M. G. 2008. *Bioresource Technology* 99: 2037–2045.
30. Liu, S. Y., Sporer, F., Wink, M., Jourdane, J., Henning, R., Li, Y. L. and Ruppel, A. 1997. *Tropical Medicine ad International Health* 2: 179–188.
31. Shetty, S., Udupa, S. L., Udupa, A. L. and Vollala, V. R. 2006. *Saudi Medical Journal* 27: 1473–1476.
32. Duke, J. A. 1985. CRC Handbook of Medicinal Herbs. CRC Press: Boca Raton, FL.
33. Felke, J. 1914. *Landw. Vers. Stat.* 82: 427–463.
34. Liberalino, A. A. A., Bambirra, E. A., Moraes-Santos, T. and Vieira, E. C. 1988. *Arquivos de Biologia e Tecnologia* 31: 539–550.
35. Abdu-Aguye, I., Sannusi, A., Alafiya-Tayo, R. A. and Bhusnurmath, S. R. 1986. *Human Toxicology* 5: 269–274.
36. el Badwi, S. M. and Adam, S. E. 1992. *Veterinary and Human Toxicology* 34: 112–115.
37. Joubert, P. H., Brown, J. M., Hay, I. T. and Sebata, P. D. 1984. *South African Medical Journal* 65: 729–730.
38. Ahmed, O. M. and Adam, S. E. 1979. *Research in Veterinary Science* 27: 89–96.
39. Gandhi, V. M., Cherian, K. M. and Mulky, M. J. 1995. *Food and Chemical Toxicology* 33: 39–42.

40. Kulkarni, M. L., Sreekar, H., Keshavamurthy, K. S. and Shenoy, N. 2005. *Indian Journal of Pediatrics* 72: 75–76.

Juniperus chinensis L.

1. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=210000896
   Date Accessed: 30th March 2008
2. USDA, ARS GRIN Taxonomy for Plants 2008
   http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?300290
   Date Accessed: 30th June 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Lee, C. K. and Cheng, Y. S. 2001. *Journal of the Chinese Chemical Society* 48: 1077–1080.
5. Lee, C. K., Fang, J. M. and Cheng, Y. S. 1995. *Phytochemistry* 39: 391–394.
6. Lim, J. P., Song, Y. C., Kim, J. W., Ku, C. H., Eun, J. S., Leem, K. H. and Kim, D. K. 2002. *Archieves of Pharmacal Research* 25: 449–452.
7. Miyata, M., Itoh, K. and Tachibana, S. 1998. Extractives of *Juniperus chinensis* L. *Journal of Wood Science* 44: 397–400.
8. Kim, Y. K. and Kim, W. K. 2001. *Han'guk Sikp'um Yongyang Kwahak Hoechi* 30: 1204–1214.
9. Kuo, Y. H. and Chen, W. C. 1994. *Chemical & Pharmaceutical Bulletin* 42: 2187–2189.
10. Kuo, Y. H. and Shiu, L. L. 1996. *Chemical & Pharmaceutical Bulletin* 44: 1758–1760.
11. Shiu, L. L., Chen, W. C. and Kuo, Y. H. 1999. *Chemical & Pharmaceutical Bulletin* 47: 557–560.
12. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
13. Bagci, E. and Digrak, M. 1996. *Turkish Journal of Biology* 20: 191–198.
14. Mohamed, S., Saka, S., El-Sharkawy, S. H., Ali, A. M. and Muid, S. 1996. *Pesticide Science* 47: 259–264.
15. Ali, A. M., Mackeen, M. M., Intan-Safinar, I., Hamid, M., Lajis, N. H., el-Sharkawy, S. H. and Murakoshi, M. 1996. *Journal of Ethnopharmacology* 53: 165–169.
16. Ju, J. B., Kim, J. S., Choi, C. W., Lee, H. K., Oh, T. K. and Kim, S. C. 2008. *Journal of Ethnopharmacology* 115: 110–115.

Kaempferia galanga L.

1. Hsuan, K., Chin, S. C. and Tan, H. T. W. 1998. The Concise Flora of Singapore volume II: Monocotyledons. Singapore University Press: Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008
   http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?20916
   Date Accessed: 30th June 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Duke, J. A. 1985. CRC Handbook of Medicinal Herbs. CRC Press: Boca Raton, FL.
5. Jenie, U. A. and Sudibyo, R. S. 2000. *Majalah Farmasi Indonesia* 11: 168–173.
6. Jirovetz, L., Buchbauer, G., Shafi, P. M. and Abraham, G. T. 2001. *Acta Pharmaceutica Turcica* 43: 107–110.
7. Kiuchi, F., Nakamura, N. and Tsuda, Y. 1987. *Phytochemistry* 26: 3350–3351.
8. bin Din, L., Zakaria, Z., Hj. Abd Malek, S. N. and Samsudin. M. W. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 118–124. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
9. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
10. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
11. Arambewela, L., Perera, A., Wijesundera, R. T. R. L. C. and Gunatileke, J. 2000. *Journal of the National Science Foundation of Sri Lanka* 28: 225–230.
12. Bhamarapravati, S., Pendland, S. L., Mahady, G. B. 2003. *In Vivo* 17: 541–544.
13. bin Jantan, I., Yassin, M. S. M., Chin, C. B., Chen, L. L. and Sim, N. L. 2003. *Pharmaceutical Biology (Lisse, Netherlands)* 41: 392–397.

14. Othman, R., Ibrahim, H., Mohd, M. A., Mustafa, M. R. and Awang, K. 2006. *Phytomedicine* 13: 61–66.
15. Xue, Y. and Chen, H. 2002. *Wei Sheng Yan Jiu* 31: 247–248.
16. Xue, Y., Murakami, A., Koizumi, K. and Chen, H. 2002. *Zhongguo Zhongyao Zazhi* 27: 522–524.
17. Vimala, S., Norhanom, A. W. and Yadav, M. 1999. *British Journal of Cancer* 80: 110–116.
18. Kosuge, T., Yokota, M., Sugiyama, K., Saito, M., Iwata, Y., Nakura, M. and Yamamoto, T. 1985. *Chemical & Pharmaceutical Bulletin* 33: 5565–5567.
19. Chirangini, P., Sharma, G. J. and Sinha, S. K. 2004. *Journal of Environmental Pathology, Toxicology and Oncology* 23: 227–236.
20. Chu, D. M., Miles, H., Toney, D., Ngyuen, C. and Marciano-Cabral, F. 1998. *Parasitology Research* 84: 746–752.
21. Kanjanapothi, D., Panthong, A., Lertprasertsuke, N., Taesotikul, T., Rujjanawate, C., Kaewpinit, D., Sudthayakorn, R., Choochote, W., Chaithong, U., Jitpakdi, A. and Pitasawat, B. 2004. *Journal of Ethnopharmacology* 90: 359–365.
22. Othman, R., Ibrahim, H., Mohd, M. A., Awang, K., Gilani, A. U. and Mustafa, M. R. 2002. *Planta Medica* 68: 655–657.
23. Tewtrakul, S. and Subhadhirasakul, S. 2007. *Journal of Ethnopharmacology* 109: 535–538.
24. Choochote, W., Chaithong, U., Kamsuk, K., Jitpakdi, A., Tippawangkosol, P., Tuetun, B., Champakaew, D. and Pitasawat, B. 2007. *Fitoterapia* 78: 359–364.
25. Choochote, W., Kanjanapothi, D., Panthong, A., Taesotikul, T., Jitpakdi, A., Chaithong, U. and Pitasawat, B. 1999. *Southeast Asian Journal of Tropical Medicine and Public Health* 30: 470–476.
26. Pandji, C., Grimm, C., Wray, V., Witte, L. and Proksch, P. 1993. *Phytochemistry* 34: 415–419.
27. Kiuchi, F., Nakamura, N., Tsuda, Y., Kondo, K. and Yoshimura, H. 1988. *Chemical & Pharmaceutical Bulletin* 36: 412–415.
28. Suzuki, J., Yasuda, I., Murata, I., Murata, R. and Nishitani, K. 2003. *Tokyo-toritsu Eisei Kenkyusho Kenkyu Nenpo* 53: 35–39.
29. Tara Shanbhag, V., Chandrakala, S., Sachidananda, A., Kurady, B. L., Smita, S. and Ganesh, S. 2006. *Indian Journal of Physiology and Pharmacology* 50: 384–390.
30. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.

Lantana camara L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?310628
   Date Accessed: 30th June 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Pan, W. D., Li, Y. J., Mai, L. T., Ohtani, K., Kasai, R. and Tanaka, O. 1992. *Yao Xue Xue Bao* 27: 515–521.
5. Pan, W. D., Li, Y. J., Mai, L. T., Ohtani, K. H., Kasai, R. T., Tanaka, O. and Yu, D. Q. 1993. *Yao Xue Xue Bao* 28: 40–44.
6. Pan, W. D., Mai, L. T., Li, Y. J., Xu, X. L. and Yu, D. Q. 1993. *Yao Xue Xue Bao* 28: 35–39.
7. Begum, S., Wahab, A., Siddiqui, B. S. and Qamar, F. 2000. *Journal of Natural Products* 63: 765–767.
8. Begum, S., Wahab, A. and Siddiqui, B. S. 2003. *Chemical and Pharmaceutical Bulletin (Tokyo)* 51: 134–137.
9. Begum, S., Wahab, A and Siddiqui, B. S. 2002. *Natural Product Letters* 16: 235–238.
10. Refahy, L. A. 2003. *Egyptian Journal of Chemistry* 46: 537–543.
11. Yadav, S. B. and Tripathi, V. 2003. *Fitoterapia* 74: 320–351.
12. Misra, L. and Laatsch, H. 2000. *Phytochemistry* 54: 969–974.
13. Barre, J. T., Bowden, B. F., Coll, J. C., DeJesus, J., De La Fuente, V. E., Janairo, G. C. and Ragasa, C. Y. 1997. *Phytochemistry* 45: 321–324.
14. Alitonou, G., Avlessi, F., Bokossa, I., Ahoussi, E., Dangou, J. and Sohounhloue, D. C. K.

2004. *Comptes Rendus Chimie* 7: 1101–1105.

15. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
16. Tabuti, J. R. 2007. *Journal of Ethnopharmacology* 116: 33–42.
17. Avadhoot, Y., Dixit, V. K. and Varma, K. C. 1980. *Indian Drugs & Pharmaceuticals Industry* 15: 19–20.
18. Kasali, A, A., Ekundayo, O., Oyedeji, A. O., Adeniyi, B. A. and Adeolu, E. O. 2002. *Journal of Essential Oil-Bearing Plants* 5: 108–110.
19. Sukul, S. and Chaudhuri, S. 1999. *Journal of Phytological Research* 12: 119–121.
20. Hernandez, T., Canales, M., Avila, J. G., Duran, A., Caballero, J., Romo de Vivar, A. and Lira, R. 2003. *Journal of Ethnopharmacology* 88: 181–188.
21. Basu, S., Ghosh, A. and Hazra, B. 2005. *Phytotherapy Research* 19: 888–894.
22. Inada, A., Nakanishi, T., Tokuda, H., Nishino, H. and Sharma, O. P. 1997. *Planta Medica* 63: 272–274.
23. Sharma, M., Sharma, P. D., Bansal, M. P. and Singh J. 2007. *Chemistry and Biodiversity* 4: 932–939.
24. Fatope, M. O., Salihu, L., Asante, S. K. and Takeda, Y. 2002. *Pharmaceutical Biology (Lisse, Netherlands)* 40: 564–567.
25. Basu, S. and Hazra, B. 2006. *Phytotherapy Research* 20: 896–900.
26. Kothari, M. and Chaudhary, B. L. 2001. *Journal of Experimental Biology* 39: 1194–1198.
27. Kumar, V. P., Chauhan, N. S., Padh, H. and Rajani, M. 2006. *Journal of Ethnopharmacology* 107: 182–188.
28. Begum, S., Wahab, A. and Siddiqui, B. S. 2008. *Natural Product Research* 22: 467–470.
29. Sagar, L., Sehgal, R. and Ojha, S. 2005. *BMC Complementary and Alternative Medicine* 5: 18.
30. Saini, N., Singh, J., Sehgal, R., Ojha, S. 2007. *Cellular and Molecular Biology (Noisy-le-grand)* 53: 79–83.
31. O'Neill, M. J., Lewis, J. A., Noble, H. M., Holland, S., Mansat, C., Farthing, J. E., Foster, G., Noble, D., Lane, S. J., Sidebottom, P. J., Lynn, S. M., Hayes, M. V. and Dix, C. J. 1998. *Journal of Natural Products* 61: 1328–1331.
32. Inada, A., Nakanishi, T., Tokuda, H., Nishino, H., Iwashima A. and Sharma, O. P. 1995. *Planta Medica* 61: 558–559.
33. Misra, N., Sharma, M., Raj, K., Dangi, A., Srivastava, S. and Misra-Bhattacharya, S. 2007. *Parasitology Research* 100: 439–448.
34. Dua, V. K., Gupta, N. C., Pandey, A. C. and Sharma, V. P. 1996. *Journal of the American Mosquito Control Association* 12: 406–408.
35. Seyoum, A., Killeen, G. F., Kabiru, E. W., Knols, B. G. and Hassanali, A. 2003. *Tropical Medicine and International Health* 8: 1005–1011.
36. Bouda, H., Tapondjou, L. A., Fontem, D. A. and Gumedzoe, M. Y. 2001. *Journal of Stored Products Research* 37: 103–109.
37. Morallo-Rejesus, B. and Tantengco, G. B. 1986. *NSTA Technology Journal* 11: 37–46.
38. Pandey, U. K., Verma, G. S., Lekha, C. and Singh, A. K. 1982. *Indian Journal of Agricultural Sciences* 52: 205–206.
39. Verma, R. K. and Verma, S. K. 2006. *Fitoterapia* 77: 466–468.
40. Abdel-Hady, N. M., Abdei-Halim, A. S. and Al-Ghadban, A. M. 2005. *Journal of the Egyptian Society of Parasitology* 35: 687–698.
41. Garg, S. K., Shah, M. A., Garg, K. M., Farooqui, M. M. and Sabir, M. 1997. *Indian Journal of Experimental Biology* 35: 1315–1318.
42. de Mello, F. B., Jacobus, D., de Carvalho, K. C. and de Mello, J. R. 2003. *Veterinary and Human Toxicology* 45: 20–23.
43. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of medicinal herbs. CRC Press: Boca Raton, FL.
44. Johnson, J. H. and Jensen, J. M. 1998. *Journal of Zoo and Wildlife Medicine* 29: 203–207.
45. Ide, A. and Tutt, C. L. 1998. *Journal of the South African Veterinary Association* 69: 30–32.
46. Sharma, O. P., Dawra, R. K. and Makkar, H. P. 1989. *Veterinary and Human Toxicology* 31: 10–13.

47. Sharma, O. P., Dawra, R. K. and Makkar, H. P. S. 1982. *Research Communications in Chemical Pathology and Pharmacology* 38: 153–156.
48. Sharma, O. P., Vaid, J., Pattabhi, V. and Bhutani, K. K. 1992. *Journal of Biochemistry and Toxicology* 7: 73–79.
49. Sharma, O. P. and Dawra, R. K. 1984. *Chemico-Biological Interactions* 49: 369–374.
50. Sharma, O. P., Dawra, R. K., Krishna, L. and Makkar, H. P. S. 1988. *Veterinary and Human Toxicology* 30: 214–218.
51. Pass, M. A. and Goosem, M. W. 1983. *Toxicon* 3: 337–340.
52. Pass, M. A. and Heath, T. 1977. *Journal of Comparative Pathology* 87: 301–306.
53. Pass, M. A. 1986. *Australian Veterinary Journal* 63: 169–171.
54. Pass, M. A., Seawright, A. A., Lamberton, J. A. and Heath, T. J. 1979. *Pathology* 11: 89–94.
55. Ganai, G. N. and Jha, G. J. 1991. *Indian Journal of Experimental Biology* 29: 762–766.
56. Akhter, M. H., Mathur, M. and Bhide, N. K. 1990. *Indian Journal of Physiology and Pharmacology* 34: 13–16.
57. Uppal, R. P. and Paul, B. S. 1982. *Indian Veterinary Journal* 59: 18–24.
58. Sharma, O. P., Sharma, S., Pattabhi, V., Mahato, S. B. and Sharma, P. D. 2007. *Critical Reviews in Toxicology* 37: 313–352.
59. Mello, F. B., Jacobus, D., Carvalho, K. and Mello, J. R. 2005. *Toxicon* 45: 459–466.

Lonicera japonica Thunb.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. eFloras.org 2008 http://www.efloras.org/florataxon.aspx?flora_id=5&taxon_id=200022324 Date Accessed: 30th March 2008
3. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
4. Machida, K., Sasaki, H., Iijima, T. and Kikuchi, M. 2002. *Chemical and Pharmaceutical Bulletin (Tokyo)* 50: 1041–1044.
5. Kwak, W. J., Han, C. K., Chang, H. W., Kim, H. P., Kang, S. S. and Son, K. H. 2003. *Chemical and Pharmaceutical Bulletin (Tokyo)* 51: 333–335.
6. Son, K. H., Jung, K. Y., Chang, H. W., Kim, H. P. and Kang, S. S. 1994. *Phytochemistry* 35: 1005–1058.
7. Choi, C. W., Jung, H. A., Kang, S. S. and Choi, J. S. 2007. *Archives of Pharmacal Research* 30: 1–7.
8. Kumar, N., Singh, B., Bhandari, P., Gupta, A. P., Uniyal, S. K. and Kaul, V. K. 2005. *Phytochemistry* 66: 2740–2744.
9. Foster, S. and Hobbs, C. 2002. A Field Guide to Western Medicinal Plants and Herbs. Houghton Mifflin Company: New York.
10. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
11. Wong-Leung, Y. L. 1981. *Xianggang Qinhui Xueyuan Xue Bao* 8: 115–124.
12. Phan, M. G., Nguyen, T. M. and Phan, T. S. 2002. *Tap Chi Hoa Hoc* 40: 103–107.
13. Yip, E. C., Chan, A. S., Pang, H., Tam, Y. K. and Wong, Y. H. 2006. *Cell Biology and Toxicology* 22: 293–302. Erratum in: *Cell Biology and Toxicology* 2007, 23: 139.
14. Leung, H. W., Kuo, C. L., Yang, W. H., Lin, C. H. and Lee, H. Z. 2006. *European Journal of Pharmacology* 534: 12–18.
15. Leung, H. W., Wu, C. H., Lin, C. H. and Lee, H. Z. 2005. *European Journal of Pharmacology* 508: 77–83.
16. Lee, H. K. and Chung, Y. S. 1963. *Kisul Yon'guso Pogo* 2: 76–78.
17. Cheng, J. T., Lee, Y. Y., Hsu, F. L., Chang, W. and Niu, C. S. 1994. *Chinese Pharmaceutical Journal (Taipei, Taiwan)* 46: 575–582.
18. Tae, J., Han, S. W., Yoo, J. Y., Kim, J. A., Kang, O. H., Baek, O. S., Lim, J. P., Kim, D. K., Kim, Y. H., Bae, K. H. and Lee, Y. M. 2003. *Clinica Chimica Acta* 330: 165–171.
19. Kwak, W. J., Han, C. K., Chang, H. W., Kim, H. P., Kang, S. S. and Son, K. H. 2003. *Chemical and Pharmaceutical Bulletin (Tokyo)* 51: 333–335.
20. Lee, S. J., Son, K. H., Chang, H. W., Kang, S. S. and Kim, H. P. 1998. *Phytotherapy Research* 12: 445–447.

21. Xu, Y., Oliverson, B. G. and Simmons, D. L. 2007. *Journal of Ethnopharmacology* 111: 667–670.
22. Son, M. J., Moon, T. C., Lee, E. K., Son, K. H., Kim, H. P., Kang, S. S., Son, J. K., Lee, S. H. and Chang, H. W. 2006. *Archives of Pharmacal Research* 29: 282–286.
23. Suh, S. J., Chung, T. W., Son, M. J., Kim, S. H., Moon, T. C., Son, K. H., Kim, H. P., Chang, H. W. and Kim, C. H. 2006. *Archives of Biochemistry and Biophysics* 447: 136–146.
24. Park, E., Kum, S., Wang, C., Park, S. Y., Kim, B. S. and Schuller-Levis, G. 2005. *American Journal of Chinese Medicine* 33: 415–424.
25. Kim, J. A., Kim, D. K., Kang, O. H., Choi, Y. A., Park, H. J., Choi, S. C., Kim, T. H., Yun, K. J., Nah, Y. H. and Lee, Y. M. 2005. *International Immunopharmacology* 5: 209–217.
26. Chang, W. C. and Hsu, F. L. 1992. *Prostaglandins, Leukotrienes and Essential Fatty Acids* 45: 307–312.
27. Chang, C. W., Lin, M. T., Lee, S. S., Liu, K. C., Hsu, F. L. and Lin, J. Y. 1995. *Antiviral Research* 27: 367–374.
28. Jeong, C. S., Suh, I. O., Hyun, J. E. and Lee, E. B. 2003. *Natural Product Sciences* 9: 87–90.
29. Suh, S. J., Jin, U. H., Kim, S. H., Chang, H. W., Son, J. K., Lee, S. H., Son, K. H. and Kim, C. H. 2006. *Journal of Cellular Biochemistry* 99: 1298–1307.
30. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
31. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of medicinal herbs. CRC Press: Boca Raton, FL.
32. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
33. Pharmacopoeia of the People's Republic of China, volume 1. English Edition, 2000. Compiled by the State Pharmacopoeia Commission of People's Republic of China. Chemical Industry Press: Beijing.
34. Webster, R. M. 1993. *Cutis* 51: 424.
35. Thanabhorn, S., Jaijoy, K., Thamaree, S., Ingkaninan, K. and Panthong, A. 2006. *Journal of Ethnopharmacology* 107: 370–373.

Mangifera indica L.

1. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?23351
   Date Accessed: 30th June 2008
3. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
4. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
5. Nunez Selles, A. J., Velez Castro, H. T., Aguero-Aguero, J., Gonzalez-Gonzalez, J., Naddeo, F., De Simone, F. and Rastrelli, L. 2002. *Journal of Agriculture and Food Chemistry* 50: 762–766.
6. Rocha Ribeiro, S. M., Queiroz, J. H., Lopes Ribeiro de Queiroz, M. E., Campos, F. M. and Pinheiro Sant'ana, H. M. 2007. *Plant Foods for Human Nutrition* 62: 13–17.
7. Berardini, N., Fezer, R., Conrad, J., Beifuss, U., Carle, R. and Schieber, A. 2005. *Journal of Agriculture and Food Chemistry* 53: 1563–1570.
8. Tabuti, J. R. 2008. *Journal of Ethnopharmacology* 116: 33–42.
9. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
10. Ojewole, J. A. 2005. *Methods and Findings in Experimental and Clinical Pharmacology* 27: 547–554.
11. Garcia, D., Escalante, M., Delgado, R., Ubeira, F. M. and Leiro, J. 2003. *Phytotherapy Research* 17: 1203–1208.
12. Nkuo-Akenji, T., Ndip, R., McThomas, A. and Fru, E. C. 2001. *Central African Journal of Medicine* 47: 155–158.
13. Sairam, K., Hemalatha, S., Kumar, A., Srinivasan, T., Ganesh, J., Shankar, M. and Venkataraman, S. 2003. *Journal of Ethnopharmacology* 84: 11–15.
14. Aqil, F. and Ahmad, I. 2007. *Methods and Findings in Experimental and Clinical Pharmacology* 29: 79–92.

15. Tona, L., Kambu, K., Ngimbi, N., Mesia, K., Penge, O., Lusakibanza, M., Cimanga, K., De Bruyne, T., Apers, S., Totte, J., Pieters, L. and Vlietinck, A. J. 2000. *Phytomedicine* 7: 31–38.
16. Cojocaru, M., Droby, S., Glotter, E., Goldman, A., Gottlieb, H. E., Jacoby, B. and Prusky, D. 1986. *Phytochemistry* 25: 1093–1095.
17. Garrido, G., Gonzalez, D., Lemus, Y., Garcia, D., Lodeiro, L., Quintero, G., Delporte, C., Nunez-Selles, A. J. and Delgado, R. 2004. *Pharmacology Research* 50: 143–149.
18. Sá-Nunes, A., Rogerio, A. P., Medeiros, A. I., Fabris, V. E., Andreu, G. P., Rivera, D. G., Delgado, R. and Faccioli, L. H. 2006. *International Immunopharmacology* 6: 1515–1523.
19. Garrido, G., González, D., Lemus, Y., Delporte, C. and Delgado, R. 2006. *Phytomedicine* 13: 412–418.
20. Muanza, D. N., Euler, K. L., Williams, L. and Newman, D. J. 1995. *International Journal of Pharmacognosy* 33: 98–106.
21. Ali, A. M., Mooi, L. Y., Yih, K. Y., Norhanom, A. W., Saleh, K. M., Lajis, N. H., Yazid, A. M., Ahmad, F. B. H. and Prasad, U. 2000. *Natural Product Sciences* 6: 147–150.
22. Mooi, L. Y., Ali, A. M., Norhanom, A. B., Salleh, K. M., Murakami, A. and Koshimizu, K. 1999. *Natural Product Sciences* 5: 33–38.
23. Percival, S. S., Talcott, S. T., Chin, S. T., Mallak, A. C., Lounds-Singleton, A. and Pettit-Moore, J. 2006. *Journal of Nutrition* 136: 1300–1304.
24. Martinez, G., Delgado, R., Perez, G., Garrido, G., Nunez Selles, A. J. and Leon, O. S. 2000. *Phytotherapy Research* 14: 424–427.
25. Martinez Sanchez, G., Candelario-Jalil, E., Giuliani, A., Leon, O. S., Sam, S., Delgado, R. and Nunez Selles, A. J. 2001. *Free Radical Research* 35: 465–473.
26. Sanchez, G. M., Re, L., Giuliani, A., Nunez-Selles, A. J., Davison, G. P. and Leon-Fernandez, O. S. 2000. *Pharmacological Research* 42: 565–573.
27. Pardo-Andreu, G. L., Sánchez-Baldoquín, C., Avila-González, R., Yamamoto, E. T., Revilla, A., Uyemura, S. A., Naal, Z., Delgado, R. and Curti, C. 2006. *Pharmacology Research* 54: 389–395.
28. Pardo-Andreu, G. L., Delgado, R., Núñez-Sellés, A. J., Vercesi, A. E. 2006. *Phytotherapy Research* 20: 120–124.
29. Pardo-Andreu, G. L., Philip, S. J., Riaño, A., Sánchez, C., Viada, C., Núñez-Sellés, A. J. and Delgado, R. 2006. *Archives of Medical Research* 37: 158–164.
30. Pardo-Andreu, G., Delgado, R., Velho, J., Inada, N. M., Curti, C. and Vercesi, A. E. 2005. *Pharmacology Research* 51: 427–435.
31. Pardo-Andreu, G. L., Paim, B. A., Castilho, R. F., Velho, J. A., Delgado, R., Vercesi, A. E. and Oliveira, H. C. 2008. *Pharmacology Research* 57: 332–338.
32. Pardo-Andreu, G. L., Barrios, M. F., Curti, C., Hernández, I., Merino, N., Lemus, Y., Martínez, I., Riaño, A. and Delgado, R. 2008. *Pharmacology Research* 57: 79–86.
33. Larrauri, J. A., Ruperez, P. and Saura-Calixto, F. 1997. *Zeitschrift fuer Lebensmittel-Untersuchung und — Forschung A: Food Research and Technology* 205: 39–42.
34. Ajila, C. M. and Prasada Rao, U. J. 2007. *Food and Chemical Toxicology* 46: 303–309.
35. Rodríguez, J., Di Pierro, D., Gioia, M., Monaco, S., Delgado, R., Coletta, M. and Marini, S. 2006. *Biochimica et Biophysica Acta* 1760: 1333–1342.
36. Prabhu, S., Jainu, M., Sabitha, K. E. and Devi, C. S. 2006. *Journal of Ethnopharmacology* 107: 126–133.
37. Remirez, D., Tafazoli, S., Delgado, R., Harandi, A. A. and O'Brien, P. J. 2005. *Drug Metabolism and Drug Interactions* 21: 19–29.
38. Bafna, P. A. and Balaraman, R. 2005. *Phytotherapy Research* 19: 216–221.
39. Yoosook, C., Bunyapraphatsara, N., Boonyakiat, Y. and Kantasuk, C. 2000. *Phytomedicine* 6: 411–419.
40. Zhu, X. M., Song, J. X., Huang, Z. Z., Wu, Y. M. and Yu, M. J. 1993. *Zhongguo Yao Li Xue Bao* 14: 452–454.
41. Carvalho, A. C., Guedes, M. M., de Souza, A. L., Trevisan, M. T., Lima, A. F., Santos, F. A. and Rao, V. S. 2007. *Planta Medica* 73: 1372–1376; Erratum in: *Planta Medica* 73: 1522.
42. Lima, Z. P., Severi, J. A., Pellizzon, C. H., Brito, A. R., Solis, P. N., Cáceres, A., Girón, L. M., Vilegas, W. and Hiruma-Lima, C. A.

2006. *Journal of Ethnopharmacology* 106: 29–37.

43. Prasad, S., Kalra, N. and Shukla, Y. 2007. *Molecular Nutrition and Food Research* 51: 352–359.
44. Sanchez, G. M., Rodriguez, H., MA, Giuliani, A., Nunez Selles, A. J., Rodriguez, N. P., Leon Fernandez, O. S. and Re, L. 2003. *Phytotherapy Research* 17: 197–201.
45. Prashanth, D., Amit, A., Samiulla, D. S., Asha, M. K. and Padmaja, R. 2001. *Fitoterapia* 72: 686–688.
46. Muruganandan, S., Srinivasan, K., Gupta, S., Gupta, P. K. and Lal, J. 2005. *Journal of Ethnopharmacology* 97: 497–501.
47. Aderibigbe, A. O., Emudianughe, T. S. and Lawal, B. A. 2001. *Phytotherapy Research* 15: 456–458.
48. Anila, L. and Vijayalakshmi, N. R. 2002. *Journal of Ethnopharmacology* 79: 81–87.
49. Garcia, D., Delgado, R., Ubeira, F. M. and Leiro. J. 2002. *International Immunopharmacology* 2: 797–806.
50. Mohammed, A. and Chadee, D. D. 2007. *Journal of the American Mosquito Control Association* 23: 172–176.
51. Jagetia, G. C. and Baliga, M. S. 2005. *Phytomedicine* 12: 209–215.
52. Jagetia, G. C. and Venkatesha, V. A. 2005. *Environmental and Molecular Mutagenesis* 46: 12–21.
53. Watt, G. 1972. Dictionary of the Economic Products of India, volume 1–6. Cosmo Publications: New Delhi.
54. Hegde, V. L. and Venkatesh, Y. P. 2007. *Journal of Investigational Allergology and Clinical Immunology* 17: 341–344.
55. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of Medicinal Herbs. CRC Press: Boca Raton, FL.
56. Rodeiro, I., Donato, M. T., Lahoz, A., González-Lavaut, J. A., Laguna, A., Castell, J. V., Delgado, R. and Gómez-Lechón, M. J. 2008. *Chemico-Biological Interactions* 172: 1–10.
57. Rodeiro, I., Donato, M. T., Jiménez, N., Garrido, G., Delgado, R. and Gómez-Lechón, M. J. 2007. *Food and Chemical Toxicology* 45: 2506–2512.

Manihot esculenta **Crantz.**

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200012596
   Date Accessed: 30th March 2008
3. Duke, J. A. 1985. CRC Handbook of Medicinal Herbs. CRC Press: Boca Raton, FL.
4. Sumner J. 2000. The Natural History of Medicinal Plants. Timber Press, Portland.
5. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
6. Gomez-Vasquez, R., Day, R., Buschmann, H., Randles, S., Beeching, J. R. and Cooper, R. M. 2004. *Annals of Botany (London)* 94: 87–97.
7. Chaturvedula, V. S., Schilling, J. K., Malone, S., Wisse, J. H., Werkhoven, M. C. and Kingston, D. G. 2003. *Planta Medica* 69: 271–274.
8. Padmaja, G. and Panikkar, K. R. 1989. *Indian Journal of Physiology and Pharmacology* 27: 635–639.
9. Boby, R. G. and Indira, M. 2004. *Indian Journal of Physiology and Pharmacology* 48: 41–50.
10. Gaitan, E., Cooksey, R. C., Legan, J., Lindsay, R. H., Ingbar, S. H. and Medeiros-Neto, G. 1994. *European Journal of Endocrinology* 131: 138–144.
11. Sreeja, V. G., Nagahara, N., Li, Q. and Minami, M. 2003. *British Journal of Nutrition* 90: 467–472.
12. Izokun-Etiobhio, B. O. and Ugochukwu, E. N. 1995. *Nigerian Journal of Physiological Sciences* 11: 46–49.
13. Akintonwa, A., Tunwashe, O. and Onifade, A. 1994. *Acta Horticulturae* 375: 285–288.
14. Kamalu, B. P. 1991. *British Journal of Nutrition* 65: 365–372.
15. Llorens, J. 2004. *Endocrinologia y Nutricion* 51: 418–425.

Melaleuca cajuputi **Roxb.**

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons.

Singapore University Press: National University of Singapore.
2. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
3. Wee, Y. C. and Corlett, R. 1986. The City and the Forest: Plant Life in Urban Singapore. Singapore University Press: National University of Singapore.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?23778
Date Accessed: 30th June 2008
5. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
6. bin Din, L., Zakaria, Z., Hj. Abd Malek, S. N. and Samsudin, M. W. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 118–124. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
7. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
8. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.

Melastoma malabathricum L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=242413697
Date Accessed: 30th March 2008
3. Das, K. K. and Kotoky, J. 1988. *Journal of the Indian Chemical Society* 65: 385–386.
4. Dinda, B. and Saha, M. K. 1988. *Journal of the Indian Chemical Society* 65: 209–211.
5. Yoshida, T., Nakata, F., Hosotani, K., Nitta, A. and Okuda, T. 1992a. *Chemical and Pharmaceutical Bulletin* 40: 1727–1732.
6. Yoshida, T., Nakata, F., Hosotani, K., Nitta, A. and Okuda, T. 1992b. *Phytochemistry* 31: 2829–2833.
7. Yoshida, T., Ito, H. and Hipolito, I. J., 2005. *Phytochemistry* 66: 1972–1983.
8. Mohandoss, S. and Ravindran, P. 1993. *Fitoterapia* 64: 277–278.
9. Malek, S. N. H. A., Baek, S. H. and Asari, A. 2003. *ACGC Chemical Research Communications* 16: 28–33.
10. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
11. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
12. Sulaiman, M. R., Somchit, M. N., Israf, D. A., Ahmad, Z. and Moin, S. 2004. *Fitoterapia* 75: 667–672.
13. Zakaria, Z. A., Nor, R. N. S. R. M., Kumar, G. H., Ghani, Z. D. F. A., Sulaiman, M. R., Devi, G. R., Jais, A. M. M., Somchit, M. N. and Fatimah, C. A. 2006. *Canadian Journal of Physiology and Pharmacology* 84: 1291–1299.
14. Lohezic-Le Devehat, F., Bakhtiar, A., Bezivin, C., Amoros, M. and Boustie, J. 2002. *Fitoterapia* 73: 400–405.
15. Susanti, D., Sirat, H. M., Ahmad, F., Mat Ali, R., Aimi, N. and Kitajima, M. 2007. *Food Chemistry* 103: 710–716.
16. Mazura, M. P., Susanti, D. and Rasadah, M. A. 2007. *Pharmaceutical Biology* 45: 372–375.
17. Rengel, Z. 2004. *Biometals* 17: 669–689.

Mimosa pudica L.

1. Wee, Y. C. and Corlett, R. 1986. The City and the Forest: Plant Life in Urban Singapore. Singapore University Press: National University of Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Nguyen, V. D. 1993. Medicinal Plants of Vietnam, Cambodia and Laos. Nguyen Van Duong: Vietnam.
4. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.

5. Kang, J. G., Shin, S. Y., Kim, M. J., Bajpai, V., Maheshwari, D. K. and Kang, S. C. 2004. *Journal of Antibiotics* 57: 726–731.
6. Yuan, K., Lü, J. L. and Yin, M. W. 2006. *Yao Xue Xue Bao* 41: 435–438.
7. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
8. Lans, C. 2007. *Journal of Ethnobiology and Ethnomedicine* 3: 13.
9. Molina, M., Contreras, C. M. and Tellez-Alcantara, P. 1999. *Phytomedicine* 6: 319–323.
10. Robinson, R. D., Williams, L. A., Lindo, J. F., Terry, S. I. and Mansingh, A. 1990. *West Indian Medical Journal* 39: 213–217.
11. Akinsinde, K. A. and Olukoya, D. K. 1995. *Journal of Diarrhoeal Diseases Research* 13: 127–129.
12. Ngo Bum, E., Dawack, D. L., Schmutz, M., Rakotonirina, A., Rakotonirina, S. V., Portet, C., Jeker, A., Olpe, H. R. and Herrling, P. 2004. *Fitoterapia* 75: 309–314.
13. Ganguly, M., Devi, N., Mahanta, R. and Borthakur, M. K. 2007. *Contraception* 76: 482–485.
14. Amalraj, T. and Ignacimuthu, S. 2002. *Fitoterapia* 73: 351–352.
15. Alsala, S. V. 2000. *Journal of Ecotoxicology & Environmental Monitoring* 10: 25–29.
16. Girish, K. S., Mohanakumari, H. P., Nagaraju, S., Vishwanath, B. S. and Kemparaju, K. 2004. *Fitoterapia* 75: 378–380.
17. Mahanta, M. and Mukherjee, A. K. 2001. *Journal of Ethnopharmacology* 75: 55–60.

Mirabilis jalapa L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200007009
   Date Accessed: 30th March 2008
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Yang, S. W., Ubillas, R., McAlpine, J., Stafford, A., Ecker, D. M., Talbot, M. K. and Rogers, B. 2001. *Journal of Natural Products* 64: 313–317.
5. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
6. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
7. Watt, J. M. and Breyer-Brandwijk, M. G. 1962. The Medicinal and Poisonous Plants of South Africa. E. and S. Livingstone: London.
8. Kusamba, C., Byamana, K. and Mbuyi, W. M. 1991. *Journal of Ethnopharmacology* 35: 197–199.
9. Michalet, S., Cartier, G., David, B., Mariotte, A. M., Dijoux-franca, M. G., Kaatz, G. W., Stavri, M. and Gibbons, S. 2007. *Bioorganic and Medicinal Chemistry Letters* 17: 1755–1758.
10. De Bolle, M. F. C., Terras, F. R. G., Cammue, B. P. A., Rees, S. B. and Broekaert, W. F. 1993. *Developments in Plant Pathology* 2: 433–436.
11. Wong, R. N., Ng, T. B., Chan, S. H., Dong, T. X. and Yeung, H. W. 1992. *Biochemistry International* 28: 585–593.
12. Morton, J. F. 1981. Atlas of Medicinal Plants of Middle America: Bahamas to Yucatan. C.C. Thomas Publishers Co.: Springfield, IL.

Morinda citrifolia L.

1. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Goh, S. H. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 7–26. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
4. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.

5. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
6. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?318237
Date Accessed: 30th June 2008
7. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
8. Sang, S. M., Liu, G. M., He, K., Zhu, N. Q., Dong, Z. G., Zheng, Q. Y., Rosen, R. T. and Ho, C. T. 2003. *Bioorganic & Medicinal Chemistry* 11: 2499–2502.
9. Farine, J. P., Legal, L., Moreteau, B. and Le Quere, J. 1996. *Phytochemistry* 41: 433–438.
10. Dyas, L., Threlfall, D. R. and Goad, L. J. 1994. *Phytochemistry* 35: 655–660.
11. Saludes, J. P., Garson, M. J., Franzblau, S. G. and Aguinaldo, A. M. 2002. *Phytotherapy Research* 16: 683–685.
12. Inoue, K., Nayeshiro, H., Inouyet, H. and Zenk, M. 1981. *Phytochemistry* 20: 1693–1700.
13. Liu, G., Bode, A., Ma, W. Y., Sang, S., Ho, C. T. and Dong, Z. 2001. *Cancer Research* 61: 5749–5756.
14. Su, B. N., Pawlus, A. D., Jung, H. A., Keller, W. J., McLaughlin, J. L. and Kinghorn, A. D. 2005. *Journal of Natural Products* 68: 592–595.
15. Kamiya, K., Tanaka, Y., Endang, H., Umar, M. and Satake, T. 2005. *Chemical & Pharmaceutical Bulletin* 53: 1597–1599.
16. Pawlus, A. D., Su, B. N., Keller, W. J. and Kinghorn, A. D. 2005. *Journal of Natural Products* 68: 1720–1722.
17. Siddiqui, B. S., Sattar, F. A., Begum, S., Gulzar, T. and Ahmad F. 2006. *Natural Product Research* 20: 1136–1144.
18. Dalsgaard, P. W., Potterat, O., Dieterle, F., Paululat, T., Kühn, T. and Hamburger, M. 2006. *Planta Medica* 72: 1322–1327.
19. Bui, A. K. T., Bacic, A. and Pettolino, F. 2006. *Phytochemistry* 67: 1271–1275.
20. Samoylenko, V., Zhao, J., Dunbar, D. C., Khan, I, A., Rushing, J. W. and Muhammad, I. 2006. *Journal of Agriculture and Food Chemistry* 54: 6398–6402.
21. Takashima, J., Ikeda, A. Y., Komiyama, A. K., Hayashi, M., Kishida, A., Ohsaki, A. 2007. *Chemical & Pharmaceutical Bulletin* 55: 343–345.
22. Siddiqui, B. S., Sattar, F. A., Ahmad, F. and Begum, S. 2007. *Archieves of Pharmacal Research* 30: 919–923.
23. Siddiqui, B. S., Sattar, F. A., Begum, S., Gulzar, T. and Ahmad, F. 2007. *Archieves of Pharmacal Research* 30: 793–798.
24. Akihisa, T., Matsumoto, K., Tokuda, H. and Yasukawa, K., Seino, K., Nakamoto, K., Kuninaga, H., Suzuki, T. and Kimura, Y. 2007. *Journal of Natural Products* 70: 754–757.
25. Deng, S., West, B. J., Palu, A.K., Zhou, B. N. and Jensen, C. J. 2007. *Phytomedicine* 14: 517–522.
26. Kamiya, K., Hamabe, W., Harada, S., Murakami, R., Tokuyama, S. and Satake, T. 2008. *Biological and Pharmaceutical Bulletin* 31: 935–938.
27. Deng, Y., Chin, Y. W., Chai, H., Keller, W. J. and Kinghorn, A. D. 2007. *Journal of Natural Products* 70: 2049–2052.
28. Lin, C. F., Ni, C. L., Huang, Y. L., Sheu, S. J. and Chen, C. C. 2007. *Natural Product Research* 21: 1199–1204.
29. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
30. Wang, M.Y., West, B. J., Jensen, J. C, Nowicki, D., Su, C., Palu, A. K. and Anderson, G. 2002. *Acta Pharmacologica Sinica* 23: 1127–1141.
31. Muhammad, M. M., Abdul, M. A., Mohd, A. A., Rozita, M. N., Nashriyah, B. M., Abdul, R. R. and Kazuyoshi, K. 1997. *Pesticide Science* 51: 165–170.
32. Hornick, C. A., Myers, A., Sadowska-Krowicka H., Anthony, C. T. and Woltering, E. A. 2003. *Angiogenesis* 6: 143–149.
33. Zaidan, M. R. S., Noor Rain, A., Badrul, A. R., Adlin, A., Norazah, A. and Zakiah, I. 2005. *Tropical Biomedicine* 22: 165–170.
34. Hirazumi, A., Furusawa, E., Chou, S. C. and Hokama, Y. 1994. *Proceedings of the Western Pharmacology Society* 37: 145–146.
35. Furusawa, E., Hirazumi, A., Story, S. and Jensen, J. 2003. *Phytotherapy Research* 17: 1158–1164.

36. Arpornsuwan, T. and Punjanon, T. 2006. *Phytotherapy Research* 20: 515–517.
37. Hirazumi, A. and Furusawa, E. 1999. *Phytotherapy Research* 13: 380–387.
38. Li, R. W., Myers, S. P., Leach, D. N., Lin, G. D. and Leach, G. 2003. *Journal of Ethnopharmacology* 85: 25–32.
39. Mohd Zin, Z., Abdul Hamid, A. and Osman, A. 2002. *Food Chemistry* 78: 227–231.
40. Mohd Zin, Z., Abdul Hamid, A., Osman, A. and Saari, N. 2006. *Food Chemistry* 94: 169–178.
41. Calzuola, I., Gianfranceschi, G. L. and Marsili, V. 2006. *International Journal of Food Sciences and Nutrition* 57: 168–177.
42. Wang, M. Y. and Su, C. 2001. *Annals of the New York Academy of Sciences* 952: 161–168.
43. Yang, J., Paulino, R., Janke-Stedronsky, S. and Abawi, F. 2007. *Food Chemistry* 102: 302–308.
44. Owen, P. L., Matainaho, T., Sirois, M. and Johns, T. 2007. *Journal of Biochemical and Molecular Toxicology* 21: 231–242.
45. Ancolio, C., Azas, N., Mahiou, V., Ollivier, E., Di Giorgio, C., Keita, A., Timon-David, P. and Balansard, G. 2002. *Phytotherapy Research* 16: 646–649.
46. Nayak, B. S., Isitor, G. N., Maxwell, A., Bhogadi, V. and Ramdath, D. D. 2007. *Journal of Wound Care* 16: 83–86.
47. Wang, M. Y., Nowicki, D., Anderson, G., Jensen, J. and West, B. 2008. *Plant Foods for Human Nutrition* 63: 59–63.
48. Skidmore-Roth, L. 2001. Mosby's Handbook of Herbs and Natural Supplements. Mosby, Inc.: St. Louis.
49. Mueller, B. A., Scott, M. K., Sowinski, K. M. and Prag, K. A. 2000. *American Journal of Kidney Diseases* 35: 310–312.
50. Younos, C., Rolland, A., Fleurentin, J., Lanhers, M. C., Misslin, R. and Mortier, F. 1990. *Planta Medica* 56: 430–434.
51. Westendorf, J., Effenberger, K., Iznaguen, H. and Basar, S. 2007. *Journal of Agriculture and Food Chemistry* 55: 529–537.
52. Stadlbauer, V., Fickert, P., Lackner, C., Schmerlaib, J., Krisper, P., Trauner, M. and Stauber, R. E. 2005. *World Journal of Gastroenterology* 11: 4758–4760.
53. Millonig, G., Stadlmann, S. and Vogel, W. 2005. *European Journal of Gastroenterology and Hepatology* 17: 445–447.
54. Yuce, B., Gulberg, V., Diebold, J. and Gerbes, A. L. 2006. *Digestion* 73: 167–170.

Nelumbo nucifera Gaertn.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?25110
   Date Accessed: 30th June 2008
3. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
4. Saeed, A., Omer, E. and Hashem, A. 1993. *Bulletin of the Faculty of Pharmacy (Cairo University)* 31: 347–351.
5. Ohkoshi, E., Miyazaki, H., Shindo, K., Watanabe, H., Yoshida, A. and Yajima, H. 2007. *Planta Medica* 73: 1255–1259.
6. Xu, H. D. and Jian, Q. J. 2008. *Chinese Chemical Letters* 19: 308–310.
7. Xiao, H., Chen, A. H., Ji, A. M., Li, Z. L. and Song, X. D. 2007. *Zhong Yao Cai* 30: 289–291.
8. Hyun, S. K., Jung, Y. J., Chung, H. Y., Jung, H. A. and Choi, J. S. 2006. *Archives of Pharmacal Research* 29: 287–292.
9. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
10. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
11. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
12. Mukherjee, P. K., Saha, K., Balasubramanian, R., Pal, M. and Saha, B. P. 1996. *Journal of Ethnopharmacology* 54: 63–67.
13. Li, G. R., Qian, J. Q. and Lu, F. H. 1990. *Zhongguo Yao Li Xue Bao* 11: 158–161.
14. Shoji, N., Umeyama, A., Saito, N., Iuchi, A., Takemoto, T., Kajiwara, A. and Ohizumi, Y. 1987. *Journal of Natural Products* 50: 773–774.

15. Talukder, M. J. and Nessa, J. 1998. *Bangladesh Medical Research Council Bulletin* 24: 6–9.
16. Mukherjee, P. K., Saha, K., Das, J., Pal, M. and Saha, B. P. 1997. *Planta Medica* 63: 367–369.
17. Wu, M. J., Wang, L., Weng, C. Y. and Yen, J. H. 2003. *American Journal of Chinese Medicine* 31: 687–698.
18. Sohn, D. H., Kim, Y. C., Oh, S. H., Park, E. J., Li, X. and Lee, B. H. 2003. *Phytomedicine* 10: 165–169.
19. Rai, S., Wahile, A., Mukherjee, K., Saha, B. P. and Mukherjee, P. K. 2006. *Journal of Ethnopharmacology* 104: 322–327.
20. Ling, Z. Q., Xie, B. J. and Yang, E. L. 2005. *Journal of Agriculture and Food Chemistry* 53: 2441–2445.
21. Yu, J. and Hu, W. S. 1997. *Yao Xue Xue Bao* 32: 1–4.
22. Xiao, J. H., Zhang, J. H., Chen, H. L., Feng, X. L. and Wang, J. L. 2005. *Planta Medica* 71: 225–230.
23. Mukherjee, P. K., Das, J., Saha, K., Giri, S. N., Pal, M. and Saha, B. P. 1996. *Indian Journal of Experimental Biology* 34: 275–276.
24. Sinha, S., Mukherjee, P. K., Mukherjee, K., Pal, M., Mandal, S. C. and Saha, B. P. 2000. *Phytotherapy Research* 14: 272–274.
25 Kashiwada, Y., Aoshima, A., Ikeshiro, Y., Chen, Y. P., Furukawa, H., Itoigawa, M., Fujioka, T., Mihashi, K., Cosentino, L. M., Morris-Natschke, S. L. and Lee, K. H. 2005. *Bioorganic & Medicinal Chemistry* 13: 443–448.
26. Kuo, Y. C., Lin, Y. L., Liu, C. P. and Tsai, W. J. 2005. *Journal of Biomedical Sciences* 12: 1021–1034.
27. Mukherjee, P. K., Saha, K., Pal, M. and Saha, B. P. 1997. *Journal of Ethnopharmacology* 58: 207–213.
28. Lee, M. W., Kim, J. S., Cho, S. M., Kim, J. H. and Lee, J. S. 2001. *Natural Product Sciences* 7: 107–109.
29. la Cour, B., Molgaard, P. and Yi, Z. 1995. *Journal of Ethnopharmacology* 46: 125–129.
30. Liu, C. P., Tsai, W. J., Shen, C. C., Lin, Y. L., Liao, J. F., Chen, C. F. and Kuo, Y. C. 2006. *European Journal of Pharmacology* 531: 270–279.
31. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
32. van Wyk, B. E. and Wink, M. 2004. Medicinal plants of the world: an illustrated scientific guide to important medicinal plants and their uses. Timber Press: Portland.
33. McGuffin, M., Hobbs, C., Upton, R. and Goldberg, A. (Eds). 1997. *American Herbal Products Association's Botanical Safety Handbook.* CRC Press: Boca Raton, FL.

Nephelium lappaceum L.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
3. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200013210
Date Accessed: 30th March 2008
4. Nishizawa, M., Adachi, K., Sastrapradja, S. and Hayashi, Y. 1983. *Phytochemistry* 22: 2853–2855.
5. Wong, K. C., Wong, S. N., Loi, H. K. and Lim, C. L. 1996. *Flavour and Fragrance Journal* 11: 223–229.
6. Ragasa, C. Y., de Luna, R. D., Cruz, W. C. Jr. and Rideout, J. A. 2005. *Journal of Natural Products* 68: 1394–1396.
7. Avato, P., Rosito, I., Papadia, P. and Fanizzi, F. P. 2006. *Natural Product Communications* 1: 751–755.
8. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
9. Nawawi, A., Nakamura, N., Hattori, M., Kurokawa, M. and Shiraki, K. 1999. *Phytotherapy Research* 13: 37–41.
10. Thitilertdecha, N., Teerawutgulrag, A. and Rakariyatham N. 2008. *LWT-Food Science and Technology* 41: 2029–2035.
11. Wall, M. M. 2006. *Journal of Food Composition and Analysis* 19: 655–663.
12. Palanisamy, U., Cheng, H. M., Masilamani, T., Subramaniam, T., Ling, L. T. and

Radhakrishnan, A. K. 2008. *Food Chemistry* 109: 54–63.

13. Okonogi, S., Duangrat, C., Anuchpreeda, S., Tachakittirungrod, S. and Chowwanapoonpohn, S. 2007. *Food Chemistry* 103: 839–846.

Nerium oleander L.

1. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200018424
   Date Accessed: 30th March 2008
2. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
3. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
4. Siddiqui, B. S., Begum, S., Siddiqui, S. and Lichter, W. 1995. *Phytochemistry* 39: 171–174.
5. Begum, S., Siddiqui, B. S., Sultana, R., Zia, A. and Suria, A. 1999. *Phytochemistry* 50: 435–438.
6. Huq, M. M., Jabbar, A., Rashid, M. A., Hasan, C. M., Ito, C. and Furukawa, H. 1999. *Journal of Natural Products* 62: 1065–1067.
7. Mueller, B. M., Rosskopf, F., Paper, D. H., Kraus, J. and Franz, G. 1991. *Pharmazie* 46: 657–663.
8. Zhao, M., Bai, L., Wang, L., Toki, A., Hasegawa, T., Kikuchi, M., Abe, M., Sakai, J., Hasegawa, R., Bai, Y., Mitsui, T., Ogura, H., Kataoka, T., Oka, S., Tsushima, H., Kiuchi, M., Hirose, K., Tomida, A., Tsuruo, T. and Ando, M. 2007. *Journal of Natural Products* 70: 1098–1103.
9. Bai, L., Wang, L., Zhao, M., Toki, A., Hasegawa, T., Ogura, H., Kataoka, T., Hirose, K., Sakai, J., Bai, J. and Ando, M. 2007. *Journal of Natural Products* 70: 14–18.
10. Zhao, M., Zhang, S., Fu, L., Li, N., Bai, J., Sakai, J., Wang, L., Tang, W., Hasegawa, T., Ogura, H., Kataoka, T., Oka, S., Kiuch, M., Hirose, K. and Ando, M. 2006. *Journal of Natural Products* 69: 1164–1167.
11. Fu, L., Zhang, S., Li, N., Wang, J., Zhao, M., Sakai, J., Hasegawa, T., Mitsui, T., Kataoka, T., Oka, S., Kiuchi, M., Hirose, K. and Ando. M. 2005. *Journal of Natural Products* 68: 198–206.
12. Passalacqua, N. G., De Fine, G. and Guarrera, P. M. 2006. *Journal of Ethnobiology and Ethnomedicine* 11: 2:52.
13. Tahraoui, A., El-Hilaly, J., Israili, Z. H. and Lyoussi, B. 2007. *Journal of Ethnopharmacology* 110: 105–117.
14. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
15. Erdemoglu, N., Kupeli, E. and Yesilada, E. 2003. *Journal of Ethnopharmacology* 89: 123–129.
16. Khafagi, I. K. 2001. *Egyptian Journal of Microbiology* 34: 613–627.
17. Afaq, F., Saleem, M., Aziz, M. H. and Mukhtar, H. 2004. *Toxicology and Applied Pharmacology* 195: 361–369.
18. Smith, J. A., Madden, T., Vijjeswarapu, M. and Newman, R. A. 2001. *Biochemical Pharmacology* 62: 469–472.
19. Pathak, S., Multani, A. S., Narayan, S., Kumar, V. and Newman, R. A. 2000. *Anticancer Drugs* 11: 455–463.
20. Newman, R. A., Kondo, Y., Yokoyama, T., Dixon, S., Cartwright, C., Chan, D., Johansen, M. and Yang, P. 2007. *Integrative Cancer Therapies* 6: 354–364.
21. Turan, N., Akgün-Dar, K., Kuruca, S. E., Kiliçaslan-Ayna, T., Seyhan, V. G., Atasever, B., Meriçli, F. and Carin, M. 2006. *Journal of Experimental Therapeutics and Oncology* 6: 31–38.
22. Mekhail, T., Kaur, H., Ganapathi, R., Budd, G. T., Elson, P. and Bukowski, R. M. 2006. *Investigational New Drugs* 24: 423–427.
23. Male, O. and Holubar, K. 1967. *Proceedings of the 5th International Congress of Chemotherapy* 2: 73–78.
24. Siddiqui, B. S., Sultana, R., Begum, S., Zia, A. and Suria, A. 1997. *Journal of Natural Product* 60: 540–544.
25. Zia, A., Siddiqui, B. S., Begum, S., Siddiqui, S. and Suria, A. 1995. *Journal of Ethnopharmacology* 49: 33–39.
26. Tarkowska, J. A. 1971. *Hereditas (Lund, Sweden)* 67: 205–211.
27. el-Shazly, M. M., Nassar, M. I. and el-Sherief, H. 1996. *Journal of the Egyptian Society of Parasitology* 26: 461–473.

28. Pushpalatha, E. and Muthukrishnan, J. 1995. *Indian Journal of Malariology* 32: 14–23.
29. Adome, R. O., Gachihi, J. W., Onegi, B., Tamale, J. and Apio, S. O. 2003. *African Health Sciences* 3: b77–86.
30. Murad, T. M. C., Gazzinelli, N. M., Sarsur Neto, J. M. and Murad, J. E. 1980. *Ciencia e Cultura (Sao Paulo)* 32: 172–179.
31. Sreenivasan, Y., Sarkar, A. and Manna, S. K. 2003. *Biochemical Pharmacology* 66: 2223–2239.
32. Skidmore-Roth L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
33. Gupta, A., Joshi, P., Jortani, S. A., Valdes, R. Jr., Thorkelsson, T., Verjee, Z. and Shemie, S. 1997. *Therapeutic Drug Monitoring* 19: 711–714.
34. Nishioka, S. A. and Resende, E. S. 1995. *Revista da Associação Médica Brasileira* 41: 60–62.
35. Haynes, B. E., Bessen, H. A. and Wightman, W. D. 1985. *Annals of Emergency Medicines* 14: 350–353.
36. Arriola Martínez, P., Montero Aparicio, E., Martínez Odriozola, P. and Miguel de la Villa, F. 2006. *Medicina Clinica* 127: 759.
37. Bourgeois, B., Incagnoli, P., Hanna, J. and Tirard, V. 2005. *Annales françaises d'anesthèsie et de reanimation* 24: 640–642.
38. Pietsch, J., Oertel, R., Trautmann, S., Schulz, K., Kopp, B. and Dressler, J. A. 2005. *International Journal of Legal Medicine* 119: 236–240.
39. Wojtyna, W. and Enseleit, F. 2004. *Pacing and Clinical Electrophysiology* 27: 1686–1688.
40. Kirsch, M. 1997. *Tierarztl Prax* 25: 398–400.
41. Mahin, L., Marzou, A. and Huart, A. 1984. *Veterinary and Human Toxicology* 26: 303–304.
42. De Pinto, F., Palermo, D., Milillo, M. A. and Iaffaldano, D. 1981. *Clinica Veterinaria* 104: 15–18.
43. Haiba, M. H., Mohamed, A. I., Mehdi, A. W. R. and Nair, G. A. 2003. *Pollution Research* 22: 157–161.
44. Aslani, M. R., Movassaghi, A. R., Mohri, M., Abbasian, A. and Zarehpour, M. 2004. *Veterinary Research Communications* 28: 609–616.
45. Adam, S. E., Al-Yahya, M. A. and Al-Farhan, A. H. 2001. *American Journal of Chinese Medicine* 29: 525–532.
46. Al-Yahya, M. A., Al-Farhan, A. H. and Adam, S. E. 2000. *Fitoterapia* 71: 385–391.
47. Al-Yahya, M. A., Al-Farhan, A. H. and Adam, S. E. 2002. *American Journal of Chinese Medicine* 30: 579–587.
48. Barbosa, R. R., Fontenele-Neto, J. D. and Soto-Blanco, B. 2008. *Research in Veterinary Science* 85: 279–281.
49. Fetrow, C. W. and Avila, J. R. 2000. The Complete Guide to Herbal Medicines. Springhouse Corp: Springhouse, PA.

Ophiopogon japonicus Ker-Gawl.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200027794
   Date Accessed: 30th March 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Chang, J. M., Shen, C. C., Huang, Y. L., Chien, M. Y., Ou, J. C., Shieh, B. J. and Chen, C. C. 2002. *Journal of Natural Products* 65: 1731–1733.
5. Nguyen, T. H. A., Tran, V. S. and Wessjohann, L. 2003. *Tap Chi Hoa Hoc* 41: 117–121.
6. Nguyen, T. H. A., Tran, V. S. and Wessjohann, L. 2003. *Tap Chi Hoa Hoc* 41: 136–142.
7. Tada, A., Saitoh, T. and Shoji, J. 1980. *Chemical & Pharmaceutical Bulletin* 28: 2487–2493.
8. Cheng, Z. H., Wu, T. and Yu, B. Y. 2006. *Journal of Asian Natural Product Research* 8: 555–559.
9. Dai, H. F. and Mei, W. L. 2005. *Archieves of Pharmacal Research* 28: 1236–1238.
10. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
11. Chen, M., Yang, Z. W., Zhu, J. T., Xiao, Z. Y. and Xiao, R. 1990. *Zhongguo Yao Li Xue Bao* 11: 161–165.
12. Shibata, M., Noguchi, R., Suzuki, M., Iwase, H., Soeda, K., Niwayama, K., Kataoka, E.,

Kusuda, K. and Hamano, M. 1971. *Hoshi Yakka Daigaku Kiyo* 13: 66–76.

13. Kou, J., Sun, Y., Lin, Y., Cheng, Z., Zheng, W., Yu, B. and Xu, Q. 2005. *Biological and Pharmaceutical Bulletin* 28: 1234–1238.
14. Kou, J., Tian, Y., Tang, Y., Yan, J. and Yu, B. 2006. *Biological and Pharmaceutical Bulletin* 29: 1267–1270.
15. Kou, J., Yu, B. and Xu, Q. 2005. *Vascular Pharmacology* 43: 157–163.
16. Wang, Y., Yan, T., Shen, J., Guo, H. and Xiang, X. 2007. *Journal of Ethnopharmacology* 114: 246–253.
17. Yu, B., Yin, X., Zhang, C. and Xu, G. 1991. *Zhongguo Yaoke Daxue Xuebao* 22: 286–288.
18. Wu, X., Dai, H., Huang, L., Gao, X., Tsim, K. W. and Tu, P. 2006. *Journal of Natural Products* 69: 1257–1260.
19. Tao, J., Wang, H., Chen, J., Xu, H. and Li, S. 2005. *American Journal of Chinese Medicine* 33: 797–806.
20. Tao, J., Wang, H., Zhou, H. and Li, S. 2005. *Life Sciences* 77: 3021–3030.
21. Liu, J. Q. and Wu, D. W. 1993. *Zhongguo Zhong Xi Yi Jie He Za Zhi* 13: 150–152, 132.
22. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.

Peltophorum pterocarpum Backer ex K. Heyne.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Tan, H. T. W. and Morgany, T. 2001. A Guide to Growing the Native Plants of Singapore. Singapore Science Centre Publication: Singapore.
3. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=5&taxon_id=200012271
   Date Accessed: 30th March 2008
4. Menon, P. S., Gangabai, G., Swarnalakshmi, T., Sulochana, N. and Amala, B. 1982. *Indian Drugs* 19: 345–347.
5. Voravuthikunchai, S., Lortheeranuwat, A., Jeeju, W., Sririrak, T., Phongpaichit, S. and Supawita, T. 2004. *Journal of Ethnopharmacology* 94: 49–54.
6. Voravuthikunchai, S. P. and Limsuwan, S. 2006. *Journal of Food Protection* 69: 2336–2341.
7. Duraipandiyan, V., Ayyanar, M. and Ignacimuthu, S. 2006. *BMC Complementary and Alternative Medicine* 6: 35.

Persicaria hydropiper L.

1. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
2. USDA, ARS GRIN Taxonomy for Plants 2008
   http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?400981
   Date Accessed: 30th June 2008
3. Fukuyama, Y., Sato, T., Asakawa, Y. and Takemoto, T. 1980. *Phytochemistry* 21: 2895–2898.
4. Fukuyama, Y., Sato, T., Miura, I., Asakawa, Y. and Takemoto, T. 1983. *Phytochemistry* 22: 549–552.
5. Miyazawa, M. and Tamura, N. 2007. *Biological and Pharmaceutical Bulletin* 30: 595–597.
6. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
7. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
8. Rahman, E., Goni, S. A., Rahman, M. T. and Ahmed, M. 2002. *Fitoterapia* 73: 704–706.
9. Chaudhuri, P. S., Chaudhuri, D., Nanda, D. K., Achari, B., Bhattacharya, D. and Saha, C. 1996. *Indian Journal of Experimental Biology* 34: 277–278.
10. Kapoor, M., Garg, S. K. and Mathur, V. S. 1974. *Indian Journal of Medical Research* 62: 1225–1227.
11. Vohora, S. B., Garg, S. K. and Chaudhury, R. R. 1969. *Indian Journal of Medical Research* 57: 893–899.
12. Haraguchi, H., Hashimoto, K. and Yagi, A. 1992. *Journal of Agricultural and Food Chemistry* 40: 1349–1351.

13. Fujita, K., Fujita, T. and Kubo, I. 2000. Molecular design of anti-*Saccharomyces cerevisiae* agents and their modes of action. Book of Abstracts, 219th ACS National Meeting, San Francisco, CA, March 26–30.
14. Fujita, K. and Kubo, I. 2005. *Journal of Agricultural and Food Chemistry* 53: 5187–5191.
15. Matsumoto, T. and Tokuda, H. 1990. *Basic Life Sciences* 52: 423–427.
16. Fujimoto, T., Ose, Y., Sato, T., Matsuda, H., Nagase, H. and Kito, H. 1987. *Mutation Research* 178: 211–216.
17. Zhang, Z. and Fang, Y. 2001. *Kunchong Zhishi* 38: 207–210.
18. Zhang, Z., Liu, X., Lou, Z., Li, H., Zhu, S. and Zou, F. 1993. *Kunchong Xuebao* 36: 172–176.
19. Cribb, J. W. and Cribb, A. B. 1981. Wild Plants in Australia. Collins: Sydney.
20. Kuroiwa, K., Shibutani, M., Inoue, K., Lee, K. Y., Woo, G. H. and Hirose, M. 2006. *Food and Chemical Toxicology* 44: 1236–1244.
21. Rahman, I., Gogoi, I., Dolui, A. K. and Handique, R. 2005. *Journal of Environmental Biology* 26: 239–241.

Phyllanthus amarus Schum. & Thonn.

1. eFloras.org 2008 http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=242337368 Date Accessed: 30th March 2008
2. Palaniswamy, U. R. 2003. A guide to Medicinal Plants of Asian origin and culture. CPL Press: United States of America.
3. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?401814 Date Accessed: 30th June 2008
4. Ahmad, B. and Alam, T. 2003. *Indian Journal of Chemistry, Section B: Organic Chemistry Including Medicinal Chemistry* 42B: 1786–1790.
5. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
6. Foo, L. Y. 1993. *Phytochemistry* 33: 487–491.
7. Foo, L. Y. 1995. *Phytochemistry* 39: 217–224.
8. Foo, L. Y. and Wong, H. 1992. *Phytochemistry* 31: 711–713.
9. Houghton, P. J., Woldemariam, T. Z., O'Shea, S. and Thyagarajan, S. P. 1996. *Phytochemistry* 43: 715–717.
10. Rajakanan, V., Sripathi, M. S., Selvanayagam, S., Velmurugan, D., Murthy, U. M., Vishwas, M., Thyagarajan, S. P., Raj, S. S. S. and Fun, H. K. 2003. *Acta Crystallographica* 59: 203–205.
11. Ali, H., Houghton, P. J. and Soumyanath, A. 2006. *Journal of Ethnopharmacology* 107: 449–455.
12. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
13. Adjobimey, T., Edaye, I., Lagnika, L., Gbenou, J., Moudachirou, M. and Sanni, A. 2004. *Comptes Rendus Chimie* 7: 1023–1027.
14. Santos, A. R., De Campos, R. O., Miguel, O. G., Filho, V. C., Siani, A. C., Yunes, R. A. and Calixto, J. B. 2000. *Journal of Ethnopharmacology* 72: 229–238.
15. Okigbo, R. N. and Igwe, D. I. 2007. *Acta microbiologica et immunologica Hungarica* 54: 353–366.
16. Mazumder, A., Mahato, A. and Mazumder, R. 2006. *Natural Product Research* 20: 323–326.
17. Odetola, A. A. and Akojenu, S. M. 2000. *African Journal of Medicine and Medical Sciences* 29: 119–122.
18. Rao, M. V. and Alice, K. M. 2001. *Phytotherapy Research* 15: 265–267.
19. Agrawal, A., Srivastava, S., Srivastava, J. N. and Srivasa, M. M. 2004. *Biomedical and Environmental Sciences* 17: 359–365.
20. Kassuya, C. A. L., Silvestre, A. A., Rehder, V. L G. and Calixto, J. B. 2003. *European Journal of Pharmacology* 478: 145–153.
21. Kiemer, A. K., Hartung, T., Huber, C. and Vollmar, A. M. 2003. *Journal of Hepatology* 38: 289–297.
22. Raphael, K. R. and Kuttan, R. 2003. *Journal of Ethnoparmacology* 87: 193–197.
23. Kassuya, C. A., Silvestre, A., Menezes-de-Lima, O. Jr., Marotta, D. M., Rehder, V. L. and Calixto, J. B. 2006. *European Journal of Pharmacology* 546: 182–188.
24. Kassuya, C. A., Leite, D. F., de Melo, L. V., Rehder, V. L. and Calixto, J. B. 2005. *Planta Medica* 71: 721–726.

25. Gowrishanker, B. and Vivekanandan, O. S. 1994. *Mutation Research* 322: 185–192.
26. Rajeshkumar, N. V., Joy, K. L., Kuttan, G., Ramsewak, R. S., Nair, M. G. and Kuttan, R. 2002. *Journal of Ethnoparmacology* 81: 17–22.
27. Sripanidkulchai, B., Tattawasart, U., Laupatarakasem, P., Vinitketkumneun, U., Sripanidkulchai, K., Furihata, C. and Matsushima, T. 2002. *Phytomedicine* 9: 26–32.
28. Joy, K. L. and Kuttan, R. 1998. *Journal of Clinical Biochemistry and Nutrition* 24: 133–139.
29. Leite, D. F., Kassuya, C. A., Mazzuco, T. L., Silvestre, A., de Melo, L. V., Rehder, V. L., Rumjanek, V. M. and Calixto, J. B. 2006. *Planta Medica* 72: 1353–1358.
30. Raphael, K. R., Sabu, M. C. and Kuttan, R. 2002. *Indian Journal of Experimental Biology* 40: 905–909.
31. Kumaran, A. and Karunakaran, R. J. 2005. *LWT — Food Science and Technology* 40: 344–352.
32. Lim, Y. Y. and Murtijaya, J. 2007. *LWT — Food Science and Technology* 40: 1664–1669.
33. Traoré, M., Diallo, A., Nikièma, J. B., Tinto, H., Dakuyo, Z. P., Ouédraogo, J. B., Guissou, I. P. and Guiguemdé, T. R. 2008. *Phytotherapy Research* 22: 550–551.
34. Notka, F., Meier, G. R. and Wagner, R. 2003. *Antiviral Research* 58: 175–186.
35. Notka, F., Meier, G. R. and Wagner, R. 2004. *Antiviral Research* 64: 93–102.
36. Yeh, S. F., Hong, C. Y., Huang, Y. L., Liu, T. Y., Choo, K. B. and Chou, C. K. 1993. *Antiviral Research* 20: 185–192.
37. Dhiman, R. K. and Chawla, Y. K. 2005. *Digestive Diseases and Sciences* 50: 1807–1812.
38. Jayaram, S. and Thyagarajan, S. P. 1996. *Indian Journal of Pathology and Microbiology* 39: 211–215.
39. Huang, R. L., Huang, Y. L., Ou, J. C., Chen, C. C., Hsu, F. L. and Chang, C. 2003. *Phytotherapy Research* 17: 449–453.
40. Wright, C. I., Van-Buren, L., Kroner, C. I. and Koning, M. M. 2007. *Journal of Ethnoparmacology* 114: 1–31.
41. Pramyothin, P., Ngamtin, C., Poungshompoo, S. and Chaichantipyuth, C. 2007. *Journal of Ethnoparmacology* 114: 169–173.
42. Naaz, F., Javed, S. and Abdin, M. Z. 2007. *Journal of Ethnoparmacology* 113: 503–509.
43. Lee, C. Y., Peng, W. H., Cheng, H. Y., Chen, F. N., Lai, M. T., Chiu, T. H. 2006. *American Journal of Chinese Medicine* 34: 471–482.
44. Adeneye, A. A., Amole, O. O. and Adeneye, A. K. 2006. *Fitoterapia* 77: 511–514.
45. Raphael, K. R., Ajith, T. A., Joseph, S. and Kuttan, R. 2002. *Teratogenesis, Carcinogenesis, and Mutagenes*is 22: 285–291.
46. Khanna, S., Srivastava, C. N., Srivasa, M. M. and Srivastava, S. 2003. *Journal of Environmental Biology* 24: 391–394.
47. Harikumar, K. B. and Kuttan, R. 2007. *Journal of Radiation Research (Tokyo)* 48: 469–476.
48. Adedapo, A. A., Adegbayibi, A. Y. and Emikpe, B. O. 2005. *Phytotherapy Research* 19: 971–976.
49. Hari Kumar, K. B. and Kuttan, R. 2006. *Biological and Pharmaceutical Bulletin* 29: 1310–1313.

Piper nigrum L.

1. Palaniswamy, U. R. 2003. A guide to Medicinal Plants of Asian origin and culture. CPL Press: United States of America.
2. eFloras.org 2008 http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200005581 Date Accessed: 30th March 2008
3. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
4. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
5. Siddiqui, B. S., Gulzar, T. and Begum, S. 2002. *Heterocycles* 57: 1653–1658.
6. Siddiqui, B. S., Gulzar, T., Mahmood, A., Begum, S., Khan, B. and Afshan, F. 2004. *Chemical and Pharmaceutical Bulletin* 52: 1349–1352.
7. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
8. Tsukamoto, S., Tomise, K., Miyakawa, K., Cha, B. C., Abe, T., Hamada, T., Hirota, H.

and Ohta, T. 2002. *Bioorganic and Medicinal Chemistry* 10: 2981–2985.

9. Rasheed, M., Afshan, F., Tariq, R. M., Siddiqui, B. S., Gulzar, T., Mahmood, A., Begum, S. and Khan, B. 2005. *Natural Product Research* 19: 703–712.
10. Wei, K., Li, W., Koike, K., Chen, Y. and Nikaido, T. 2005. *Journal of Organic Chemistry* 70: 1164–1176.
11. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
12. Tiwari, K. C., Majumder, R. and Bhattacharjee, S. 1982. *International Journal of Crude Drug Research* 20: 133–137.
13. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
14. Perez, C. and Anesini, C. 1994. *The American Journal of Chinese Medicine* 22: 169–174.
15. Pradhan, K. J., Variyar, P. S. and Bandekar, J. R. 1999. *Lebensmittel-Wissenschaft and — Technologie* 32: 121–123.
16. Chaudhry, N. M. and Tariq, P. 2006. *Pakistan Journal of Pharmaceutical Sciences* 19: 214–218.
17. Kaleem, M., Sheema, Sarmad, H. and Bano, B. 2005. *Indian Journal of Physiology and Pharmacology* 49: 65–71.
18. Hector, R. J., Simon, J. E., Ramboatiana, M. M. R., Behra, O., Garvey, A. S. and Raskin, I. 2004. *Acta Horticulturae* 629: 77–81.
19. Mujumdar, A. M., Dhuley, J. N., Deshmukh, V. K., Raman, P. H. and Naik, S. R. 1990. *Journal of Medical Science and Biology* 43: 95–100.
20. Pradeep, C. R. and Kuttan, G. 2002. *Clinical and Experimental Metastasis* 19: 703–708.
21. Vijayakumar, R. S. and Nalini, N. 2006. *Cell Biochemistry and Function* 24: 491–498.
22. Vijayakumar, R. S., Surya, D. and Nalini, N. 2004. *Redox Report: Communications in Free Radical Research* 9: 105–110.
23. Natarajan, K. S., Narasimhan, M., Shanmugasundaram, K. R. and Shanmugasundaram, E. R. 2006. *Journal of Ethnopharmacology* 105: 76–83.
24. Gulcin, I. 2005. *International Journal of Food Sciences and Nutrition*. 56: 491–499.
25. Topal, U., Sasaki, M., Goto, M. and Otles, S. 2007. *International Journal of Food Science and Nutrition* 18: 1–16.
26. Chatterjee, S., Niaz, Z., Gautam, S., Adhikari, S., Variyar, P. S. and Sharma, A. 2007. *Food Chemistry* 101: 515–523.
27. Nalini, N., Sabitha, K., Viswanathan, P. and Menon, V. P. 1998. *Journal of Ethnopharmacology* 62: 15–24.
28. Koul, I. B. and Kapil, A. 1993. *Planta Medica* 59: 413–417.
29. Singh, A. and Rao, A. R. 1993. *Cancer Letters* 72: 5–9.
30. Vijayakumar, R. S. and Nalini, N. 2006. *Journal of Basic and Clinical Physiology and Pharmacology* 17: 71–86.
31. Vijayakumar, R. S., Surya, D., Senthilkumar, R. and Nalini, N. 2002. *Journal of Clinical Biochemistry and Nutrition* 32: 31–42.
32. El Hamss, R., Idaomar, M., Alonso-Moraga, A. and Munoz Serrano, A. 2003. *Food and Chemical Toxicology* 41: 41–47.
33. Panda, S. and Kar, A. 2003. *Hormone and Metabolic Research* 35: 523–526.
34. Park, I. K., Lee, S. G., Shin, S. C., Park, J. D. and Ahn, Y. J. 2002. *Journal of Agricultural and Food Chemistry* 50: 1866–1870.
35. Simas, N. K., da Costa Lima, E., Kuster, R. M. and Salgueir, C. L. 2007. *Revista da Sociedade Brasileira de Medicina Tropical* 40: 405–407.
36. Siddiqui, B. S., Gulzar, T., Begum, S. and Afshan, F. 2004. *Natural Product Research* 18: 473–477.
37. Siddiqui, B. S., Gulzar, T., Begum, S., Afshan, F. and Sattar, F. A. 2005. *Natural Product Research* 19: 143–150.
38. Scott, I. M., Gagnon, N., Lesage, L., Philogène, B. J. and Arnason, J. T. 2005. *Journal of Economic Entomology* 98: 845–855.
39. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of medicinal herbs. CRC Press: Boca Raton, FL.
40. Rinzler, C. A. 1990. The complete book of herbs, spices and condiments. Facts on File: New York.
41. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western

Pacific, Manila, Institute of Materia Medica, Hanoi.

42. McGuffin, M., Hobbs, C., Upton, R. and Goldberg, A. (Eds). 1997. *American Herbal Products Association's Botanical Safety Handbook.* CRC Press: Boca Raton, FL.
43. Skidmore-Roth L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
44. Bhat, B. G. and Chandrasekhara, N. 1986. *Toxicology* 40: 83–92.
45. Cha, B. C. 2003. *Saengyak Hakhoechi* 34: 86–90.
46. Tsukamoto, S., Cha, B. C. and Ohta, T. 2002. *Tetrahedron* 58: 1667–1671.
47. Subehan, Usia, T., Iwata, H., Kadota, S. and Tezuka, Y. 2006. *Journal of Ethnopharmacology* 105: 449–455.
48. Usia, T., Iwata, H., Hiratsuka, A., Watabe, T., Kadota, S. and Tezuka, Y. 2006. *Phytomedicine* 13: 67–73.
49. Subehan, Usia, T., Kadota, S. and Tezuka, Y. 2006. *Planta Medica* 72: 527–532.
50. Hiwale, A. R., Dhuley, J. N. and Naik, S. R. 2002. *Indian Journal of Experimental Biology* 40: 277–281.
51. Pattanaik, S., Hota, D., Prabhakar, S., Kharbanda, P. and Pandhi, P. 2006. *Phytotherapy Research* 20: 683–686.
52. Hu, Z., Yang, X., Ho, P. C., Chan, S. Y., Heng, P. W., Chan, E., Duan, W., Koh, H. L. and Zhou, S. 2005. *Drugs* 65: 1239–1282.
53. Fetrow, C. W. and Avila, J. R. 2000. The complete guide to herbal medicines. Springhouse Corp: Springhouse, PA.

Piper sarmentosum Roxb.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
4. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
5. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200005594
Date Accessed: 30th March 2008
6. Toong, Y. V. and Wong, B. L. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 280. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
7. Strunz, G. M. and Finlay, H. 1995. *Phytochemistry* 39: 731–733.
8. Masuda, T., Inazumi, A., Yamada, Y., Padolina, W. G., Kikuzaki, H. and Nakatani, N. 1991. *Phytochemistry* 30: 3227–3228.
9. Stohr, J. R., Xiao, P. G. and Bauer, R. 1999. *Planta Medica* 65: 175–177.
10. Tuntiwachwuttikul, P., Phansa, P., Pootaeng-On, Y. and Taylor, W. C. 2006. *Chemical and Pharmaceutical Bulletin* 54: 149–151.
11. Rukachaisirikul, T., Siriwattanakit, P., Sukcharoenphol, K., Wongvein, C., Ruttanaweang, P., Wongwattanavuch, P. and Suksamrarn, A. 2004. *Journal of Ethnopharmacology* 93: 173–176.
12. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
13. Zaidan, M. R. S., Noor Rain, A., Badrul, A. R., Adlin, A., Norazah, A. and Zakiah, I. 2005. *Tropical Biomedicine* 22: 165–170.
14. Rahman, N. N. N. A., Furuta, T., Kojima, S., Takane, K. and Ali Mohd, M. 1999. *Journal of Ethnopharmacology* 64: 249–254.
15. Sawangjaroen, N., Sawangjaroen, K. and Poonpanang, P. 2004. *Journal of Ethnopharmacology* 91: 357–360.
16. Ridtitid, W., Rattanaprom, W., Thaina, P., Chittrakarn, S. and Sunbhanich, M. 1998. *Journal of Ethnopharmacology* 61: 135–142.
17. Peungvicha, P., Thirawarapan, S. S., Temsiririrkkul, R., Watanabe, H., Kumar Prasain, J. and Kadota, S. 1998. *Journal of Ethnopharmacology* 60: 27–32.
18. Chaithong, U., Choochote, W., Kamsuk, K., Jitpakdi, A., Tippawangkosol, P., Chaiyasit, D., Champakaew, D., Tuetun, B. and Pitasawat B. 2006. *Journal of Vector Ecology: Journal of the Society for Vector Ecology* 31: 138–144.

19. Choochote, W., Chaithong, U., Kamsuk, K., Rattanachanpichai, E., Jitpakdi, A., Tippawangkosol, P., Chaiyasit, D., Champakaew, D., Tuetun, B. and Pitasawat, B. 2006. *Revista do Instituto de Medicina Tropical de São Paulo* 48: 33–37.
20. Chanwitheesuk, A., Teerawutgulrag, A. and Rakariyatham, N. 2005. *Food Chemistry* 92: 491–497.

Plantago major L.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=5&taxon_id=200022050
   Date Accessed: 30th March 2008
3. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
4. Ringbom, T., Segura, L., Noreen, Y., Perera, P. and Bohlin, L. 1998. *Journal of Natural Products* 61: 1212–1215.
5. Foster, S. and Hobbs, C. 2002. A Field Guide to Western Medicinal Plants and Herbs. Houghton Mifflin Company: New York.
6. Franca, F., Lago, E. L. and Marsden, P. D. 1996. *Revista da Sociedade Brasileira de Medicina Tropical* 29: 229–232.
7. Yesilada, E., Sezik, E., Fujita, T., Tanaka, S. and Tabata, M. 1993. *Phytotherapy Research* 7: 263–265.
8. Blumenthal, M., Goldberg, A. and Gruenwald, J. (Eds.). 2000. Herbal Medicine. Expanded Commission E Monographs. Integrative Medicine Communications: Austin, Texas.
9. Atta, A. H. and El-Sooud, K. A. 2004. *Journal of Ethnopharmacology* 95: 235–238.
10. Hetland, G., Samuelsen, A. B., Lovik, M., Paulsen, B. S., Aaberge, I. S., Groeng, E. C. and Michaelsen, T. E. 2000. *Scandinavian Journal of Immunology* 52: 348–355.
11. Velasco-Lezama, R., Tapia-Aguilar, R., Román-Ramos, R., Vega-Avila, E., Pérez-Gutiérrez, Ma. S. 2006. *Journal of Ethnopharmacology* 103: 36–42.
12. Atta, A. H. and Mouneir, S. M. 2005. *Phytotherapy Research* 19: 481–485.
13. Murai, M., Tamayama, Y. and Nishibe, S. 1995. *Planta Medica* 61: 479–480.
14. Shipochliev, T., Dimitrov, A. and Aleksandrova, E. 1981. *Veterinarno-meditsinski nauki* 18: 87–94.
15. Ikawati, Z., Wahyuono, S. and Maeyama, K. 2001. *Journal of Ethnopharmacology* 75: 249–256.
16. Galvez, M., Martin-Cordero, C., Lopez-Lazaro, M., Cortes, F. and Ayuso, M. J. 2003. *Journal of Ethnopharmacology* 88: 125–130.
17. Lin, L. T., Liu, L. T., Chiang, L. C. and Lin, C. C. 2002. *Phytotherapy Research* 16: 440–444.
18. Ozaslan, M., Didem Karagöz, I., Kalender, M. E., Kilic, I. H., Sari, I. and Karagöz, A. 2007. *American Journal of Chinese Medicine* 35: 841–851.
19. Campos, A. M. and Lissi, E. A. 1995. *Boletin de la Sociedad Chilena de Quimica* 40: 375–381.
20. Ponce-Macotela, M., Navarro-Alegria, I., Martinez-Gordillo, M. N. and Alvarez-Chacon, R. 1994. *Revista de Investigación Clínica* 46: 343–347.
21. Weenen, H., Nkunya, M. H. H., Bray, D. H., Mwasumbi, L. B., Kinabo, L. S., Kilimali, V. A. E. B. 1990. *Planta Medica* 56: 368–370.
22. Chiang, L. C., Chiang, W., Chang, M. Y., Ng, L. T., Lin, C. C. 2002. *Antiviral Research* 55: 53–62.
23. Basaran, A. A., Ceritoglu, I., Undeger, U. and Basaran, N. 1997. *Phytotherapy Research* 11: 609–611.
24. Dorhoi, A., Dobrean, V., Zăhan, M. and Virag, P. 2006. Modulatory effects of several herbal extracts on avian peripheral blood cell immune responses. *Phytotherapy Research* 20: 352–358.
25. Chiang, L. C., Ng, L. T., Chiang, W., Chang, M. Y. and Lin, C. C. 2003. *Planta Medica* 69: 600–604.
26. Gomez-Flores, R., Calderon, C. L., Scheibel, L. W., Tamez-Guerra, P., Rodriguez-Padilla, C., Tamez-Guerra, R. and Weber, R. J. 2000. *Phytotherapy Research* 14: 617–622.
27. Lim-Sylianco, C. Y. and Shier, W. T. 1985. *Journal of Toxicology-Toxin Reviews* 4: 71–105.

28. Basaran, A. A., Yu, T. W., Plewa, M. J. and Anderson, D. 1996. *Teratogenesis, Carcinogenesis and Mutagenesis* 16: 125–138.
29. Shipochliev, T. 1981. *Veterinarno-meditsinski nauki* 18: 94–98.
30. Krasnov, M. S., Margasiuk, D. V., Iamskov, I. A. and Iamskova, V. P. 2003. *Radiatsionnaia biologiia, radioecologiia* 43: 269–272.
31. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
32. Skidmore-Roth L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
33. Schmeda-Hirschmann, G., Loyola, J. I., Retamal, S. R. and Rodriguez, J. 1992. *Phytotherapy Research* 6: 184–188.
34. Angelov, A., Lambev, I., Markov, M., Yakimova, K., Leseva, M. and Yakimov, A. 1980. *Medical Archives* 18: 47–52.

Punica granatum L.

1. Hsuan, K. 1992. Orders and Families of Malayan Seed Plants. Singapore University Press: Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?30372
   Date Accessed: 30th June 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
5. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
6. Wang, R., Wang, W., Wang, L., Liu, R., Ding, Y. and Du, L. 2006. *Fitoterapia* 77: 534–537.
7. Wang, R. F., Xie, W. D., Zhang, Z., Xing, D. M., Ding, Y., Wang, W., Ma, C. and Du, L. J. 2004. *Journal of Natural Products* 67: 2096–2098.
8. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
9. Wren, R. C. 1988. Potter's new cyclopaedia of botanical drugs and preparations. The C.W. Daniel Company Ltd.: Saffron Walden, Essex.
10. Raj, R. K. 1975. *Indian Journal of Physiology and Pharmacology* 19, Issue 1.
11. Voravuthikunchai, S., Lortheeranuwat, A., Jeeju, W., Sririrak, T., Phongpaichit, S. and Supawita, T. 2004. *Journal of Ethnopharmacology* 94: 49–54.
12. Braga, L. C., Shupp, J. W., Cummings, C., Jett, M., Takahashi, J. A., Carmo, L. S., Chartone-Souza, E. and Nascimento, A. M. 2005. *Journal of Ethnopharmacology* 96: 335–339.
13. Guevara, J. M., Chumpitaz, J. and Valencia, E. 1994. *Revista de gastroenterología del Perú* 14: 27–31.
14. Aqil, F. and Ahmad, I. 2007. *Methods and Findings in Experimental and Clinical Pharmacology* 29: 79–92.
15. Aqil, F., Khan, M. S., Owais, M. and Ahmad, I. 2005. *Journal of Basic Microbiology* 45: 106–114.
16. Braga, L. C., Leite, A. A., Xavier, K. G., Takahashi, J. A., Bemquerer, M. P., Chartone-Souza, E. and Nascimento, A. M. 2005. *Canadian Journal of Microbiology* 51: 541–547.
17. Voravuthikunchai, S. P. and Limsuwan, S. 2006. *Journal of Food Protection* 69: 2336–2341.
18. Voravuthikunchai, S. P. and Kitpipit, L. 2005. *Clinical Microbiology and Infection* 11: 510–512.
19. Naz, S., Siddiqi, R., Ahmad, S., Rasool, S. A. and Sayeed, S. A. 2007. *Journal of Food Science* 72: M341–345.
20. Mathabe, M. C., Nikolova, R. V., Lall, N. and Nyazema, N. Z. 2006. *Journal of Ethnopharmacology* 105: 286–293.
21. Alanís, A. D., Calzada, F., Cervantes, J. A., Torres, J. and Ceballos, G. M. 2005. *Journal of Ethnopharmacology* 100: 153–157.
22. Mehta, R. and Lansky, E. P. 2004. *European Journal of Cancer Prevention* 13: 345–348.
23. Kim, N. D., Mehta, R., Yu, W., Neeman, I., Livney, T., Amichay, A., Poirier, D., Nicholls, P.,

Kirby, A., Jiang, W., Mansel, R., Ramachandran, C., Rabi, T., Kaplan, B. and Lansky, E. 2002. *Breast Cancer Research and Treatment* 71: 203–217.

24. Kawaii, S. and Lansky, E. P. 2004. *Journal of Medicinal Food* 7: 13–18.
25. Kohno, H., Suzuki, R., Yasui, Y., Hosokawa, M., Miyashita, K. and Tanaka, T. 2004. *Cancer Science* 95: 481–486.
26. Lansky, E. P., Jiang, W., Mo, H., Bravo, L., Froom, P., Yu, W., Harris, N. M., Neeman, I. and Campbell, M. J. 2005. *Investigational New Drugs* 23: 11–20.
27. Afaq, F., Saleem, M., Krueger, C. G., Reed, J. D. and Mukhtar, H. 2005. *International Journal of Cancer* 113: 423–433.
28. Lansky, E. P., Harrison, G., Froom, P. and Jiang, W. G. 2005. *Investigational New Drug* 23: 121–122; Erratum in: *Investigational New Drugs* 23: 379.
29. Lansky, E. P. and Newman, R. A. 2006. *Journal of Ethnopharmacology* 109: 177–206.
30. Malik, A. and Mukhtar, H. 2006. *Cell Cycle* 5: 371–373.
31. Malik, A., Afaq, F., Sarfaraz, S., Adhami, V. M., Syed, D. N. and Mukhtar, H. 2005. *Proceedings of the National Academy of Sciences of the USA* 102: 14813–14818.
32. Khan, N., Afaq, F., Kweon, M. H., Kim, K. and Mukhtar, H. 2007. *Cancer Research* 67: 3475–3482.
33. Khan, N., Hadi, N., Afaq, F., Syed, D. N., Kweon, M. H., Mukhtar, H. 2007. *Carcinogenesis* 28: 163–173.
34. Seeram, N. P., Adams, L. S., Henning, S. M., Niu, Y., Zhang, Y., Nair, M. G. and Heber, D. 2005. *Journal of Nutritional Biochemistry* 16: 360–367.
35. Seeram, N. P., Aronson, W. J., Zhang, Y., Henning, S. M., Moro, A., Lee, R. P., Sartippour, M., Harris, D. M., Rettig, M., Suchard, M. A., Pantuck, A. J., Belldegrun, A. and Heber, D. 2007. *Journal of Agriculture and Food Chemistry* 55: 7732–7737.
36. Pantuck, A. J., Leppert, J. T., Zomorodian, N., Aronson, W., Hong, J., Barnard, R. J., Seeram, N., Liker, H., Wang, H., Elashoff, R., Heber, D., Aviram, M., Ignarro, L. and Belldegrun, A. 2006. *Clinical Cancer Research* 12: 4018–4026.
37. Das, A. K., Mandal, S. C., Banerjee, S. K., Sinha, S., Saha, B. P. and Pal, M. 2001. *Phytotherapy Research* 15: 628–629.
38. Jafri, M. A., Aslam, M., Javed, K. and Singh, S. 2000. *Journal of Ethnopharmacology* 70: 309–314.
39. Li, Y., Qi, Y., Huang, T. H., Yamahara, J. and Roufogalis, B. D. 2008. *Diabetes, Obesity and Metabolism* 10: 10–17.
40. Katz, S. R., Newman, R. A. and Lansky, E. P. 2007. *Journal of Medicinal Food* 10: 213–217.
41. Huang, T. H., Peng, G., Kota, B. P., Li, G. Q., Yamahara, J., Roufogalis, B. D. and Li, Y. 2005. *Toxicology and Applied Pharmacology* 207: 160–169.
42. Das, A. K., Mandal, S. C., Banerjee, S. K., Sinha, S., Das, J., Saha, B. P. and Pal, M. 1999. *Journal of Ethnopharmacology* 68: 205–208.
43. Prakash, A. O., Saxena, V., Shukla, S., Tewari, R. K., Mathur, S., Gupta, A., Sharma, S. and Mathur, R. 1985. *Acta Europaea Fertilitatis* 16: 441–448.
44. Vasconcelos, L. C., Sampaio, M. C., Sampaio, F. C. and Higino, J. S. 2003. *Mycoses* 46: 192–196.
45. Holetz, F. B., Pessini, G. L., Sanches, N. R., Cortez, D. A., Nakamura, C. V. and Filho, B. P. 2002. *Memórias do Instituto Oswaldo Cruz* 97: 1027–1031.
46. Jung, K. H., Kim, M. J., Ha, E., Uhm, Y. K., Kim, H. K., Chung, J. H. and Yim, S. V. 2006. *Biological and Pharmaceutical Bulletin* 29: 1258–1261.
47. Adams, L. S., Seeram, N. P., Aggarwal, B. B., Takada, Y., Sand, D. and Heber, D. 2006. *Journal of Agriculture and Food Chemistry* 54: 980–985.
48. Reddy, M. K., Gupta, S. K., Jacob, M. R., Khan, S. I. and Ferreira, D. 2007. *Planta Medica* 73: 461–467.
49. Noda, Y., Kaneyuki, T., Mori, A. and Packer, L. 2002. *Journal of Agriculture and Food Chemistry* 50: 166–171.
50. Singh, R. P., Chidambara Murthy, K. N. and Jayaprakasha, G. K. 2002. *Journal of Agriculture and Food Chemistry* 50: 81–86.
51. Sezer, E. D., Akçay, Y. D., Ilanbey, B., Yildirim, H. K. and Sözmen, E. Y. 2007. *Journal of Medicinal Food* 10: 371–374.

52. Zaid, M. A., Afaq, F., Syed, D. N., Dreher, M. and Mukhtar, H. 2007. *Photochemistry and Photobiology* 83: 882–888.
53. Zhang, Q., Jia, D. and Yao, K. 2007. *Natural Product Research* 21: 211–216.
54. Kulkarni, A. P., Mahal, H. S., Kapoor, S. and Aradhya, S. M. 2007. *Journal of Agriculture and Food Chemistry* 55: 1491–1500.
55. Rout, S. and Banerjee, R. 2007. *Bioresource Technology* 98: 3159–3163.
56. Ricci, D., Giamperi, L., Bucchini, A. and Fraternale, D. 2006. *Fitoterapia* 77: 310–312.
57. Afaq, F. and Mukhtar, H. 2006. *Experimental Dermatology* 15: 678–684.
58. Kaur, G., Jabbar, Z., Athar, M. and Alam, M. S. 2006. *Food and Chemical Toxicology* 44: 984–993.
59. Rosenblat, M., Hayek, T. and Aviram, M. 2006. *Atherosclerosis* 187: 363–371.
60. Sudheesh, S. and Vijayalakshmi, N. R. 2005. *Fitoterapia* 76: 181–186.
61. Faria, A., Monteiro, R., Mateus, N., Azevedo, I. and Calhau, C. 2007. *European Journal of Nutrition* 46: 271–278.
62. Sestili, P., Martinelli, C., Ricci, D., Fraternale, D., Bucchini, A., Giamperi, L., Curcio, R., Piccoli, G. and Stocchi, V. 2007. *Pharmacology Research* 56: 18–26.
63. Neurath, A. R., Strick, N., Li, Y. Y. and Debnath, A. K. 2004. *BMC Infectious Diseases* 4: 41.
64. Zhang, J., Zhan, B., Yao, X., Gao, Y. and Shong, J. 1995. *Zhongguo Zhong Yao Za Zhi* 20: 556–568.
65. Li, Y., Ooi, L. S., Wang, H., But, P. P. and Ooi, V. E. 2004. *Phytotherapy Research* 18: 718–722.
66. Gharzouli, K., Khennouf, S., Amira, S. and Gharzouli, A. 1999. *Phytotherapy Research* 13: 42–45.
67. Ajaikumar, K. B., Asheef, M., Babu, B. H. and Padikkala, J. 2005. *Journal of Ethnopharmacology* 96: 171–176.
68. Toklu, H. Z., Dumlu, M. U., Sehirli, O., Ercan, F., Gedik, N., Gökmen, V. and Sener, G. 2007. *Journal of Pharmacy and Pharmacology* 59: 1287–1295.
69. Esmaillzadeh, A., Tahbaz, F., Gaieni, I., Alavi-Majd, H. and Azadbakht, L. 2006. *International Journal for Vitamin and Nutrition Research* 76: 147–151.
70. Gracious Ross, R., Selvasubramanian, S. and Jayasundar, S. 2001. *Journal of Ethnopharmacology* 78: 85–87.
71. Yamasaki, M., Kitagawa, T., Koyanagi, N., Chujo, H., Maeda, H., Kohno-Murase, J., Imamura, J., Tachibana, H. and Yamada, K. 2006. *Nutrition* 2006 22: 54–59.
72. West, T., Atzeva, M. and Holtzman, D. M. 2007. *Developmental Neuroscience* 29: 363–372.
73. Kwak, H. M., Jeon, S. Y., Sohng, B. H., Kim, J. G., Lee, J. M., Lee, K. B., Jeong, H. H., Hur, J. M., Kang, Y. H. and Song, K. S. 2005. *Archives of Pharmacal Research* 28: 1328–1332.
74. Hartman, R. E., Shah, A., Fagan, A. M., Schwetye, K. E., Parsadanian, M., Schulman, R. N., Finn, M. B. and Holtzman, D. M. 2006. *Neurobiology of Disease* 24: 506–515.
75. Loren, D. J., Seeram, N. P., Schulman, R. N. and Holtzman, D. M. 2005. *Pediatric Research* 57: 858–864; Erratum in: *Pediatric Research* 2007 62: 363.
76. de Nigris, F., Williams-Ignarro, S., Sica, V., Lerman, L. O., D'Armiento, F. P., Byrns, R. E., Casamassimi, A., Carpentiero, D., Schiano, C., Sumi, D., Fiorito, C., Ignarro, L. J. and Napoli, C. 2007. *Cardiovascular Research* 73: 414–423.
77. de Nigris, F., Williams-Ignarro, S., Lerman, L. O., Crimi, E., Botti, C., Mansueto, G., D'Armiento, F. P., De Rosa, G., Sica, V., Ignarro, L. J. and Napoli, C. 2005. *Proceedings of the National Academy of Sciences of the USA* 102: 4896–4901.
78. Fuhrman, B., Volkova, N. and Aviram, M. 2005. *Journal of Nutritional Biochemistry* 16: 570–576.
79. Sastravaha, G., Yotnuengnit, P., Booncong, P. and Sangtherapitikul, P. 2003. *Journal of the International Academy of Periodontology* 5: 106–115.
80. Murthy, K. N., Reddy, V. K., Veigas, J. M. and Murthy, U. D. 2004. *Journal of Medicinal Food* 7: 256–259.
81. Morsy, T. A., Mazyad, S. A. and el-Sharkawy, I. M. 1998. *Journal of Egyptian Society of Parasitology* 28: 699–709.
82. Tripathi, S. M. and Singh, D. K. 2000. *Brazilian Journal of Medical and Biological Research* 33: 1351–1355.

83. Tripathi, S. M., Singh, V. K., Singh, S. and Singh, D. K. 2004. *Phytotherapy Research* 18: 501–506.
84. Kapoor, L. D. 1990. Handbook of Ayurvedic Medicinal Plants. CRC Press: Boca Raton, FL.
85. Morton, J. F. 1981. Atlas of Medicinal Plants of Middle America: Bahamas to Yucatan. C.C. Thomas Publishers Co.: Springfield, IL.
86. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
87. Enrique, E., Utz, M., De Mateo, J. A., Castelló, J. V., Malek, T. and Pineda, F. 2006. *Annals of Allergy, Asthma and Immunology* 96: 122–123.
88. McGuffin, M., Hobbs, C., Upton, R. and Goldberg, A. (Eds). 1997. *American Herbal Products Association's Botanical Safety Handbook.* CRC Press: Boca Raton, FL.
89. Gaig, P., Botey, J., Gutierrez, V., Pena, M., Eseverri, J. L. and Marin, A. 1992. *Journal of Investigational Allergology and Clinical Immunology* 2: 216–218.
90. Fetrow, C. W. and Avila, J. R. 2000. The complete guide to herbal medicines. Springhouse Corp: Springhouse, PA.
91. Skidmore-Roth, L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
92. Duke, J. A., Bogenschutz-Godwin, M. J., duCellier, J. and Duke, P. A. K. 2002. Handbook of medicinal herbs. CRC Press: Boca Raton, FL.
93. Summers, K. M. 2006. Potential drug-food interactions with pomegranate juice. *Annals of Pharmacotherapy* 40: 1472–1473.
94. Usia, T., Iwata, H., Hiratsuka, A., Watabe, T., Kadota, S. and Tezuka, Y. 2006. *Phytomedicine* 13: 67–73.
95. Hidaka, M., Okumura, M., Fujita, K., Ogikubo, T., Yamasaki, K., Iwakiri, T., Setoguchi, N. and Arimori, K. 2005. *Drug Metabolism and Disposition* 33: 644–648.
96. Nagata, M., Hidaka, M., Sekiya, H., Kawano, Y., Yamasaki, K., Okumura, M. and Arimori, K. 2007. *Drug Metabolism and Disposition* 35: 302–305.

Rhodomyrtus tomentosa (Ait.) Hassk.

1. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. Wee, Y. C. and Corlett, R. 1986. The City and the Forest: Plant Life in Urban Singapore. Singapore University Press: National University of Singapore.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?31664
   Date Accessed: 30th June 2008
5. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
6. Dachriyanus, S., Sargent, M. V., Skelton, B. W., Soediro, I., Sutisna, M., White, A. H. and Yulinah, E. 2002. *Australian Journal of Chemistry* 55: 229–232.
7. Dachriyanus, Fahmi, R., Sargent, M. V., Skelton, B. W. and White, A. H. 2004. *Acta Crystallographica Section E-Structure Reports Online* 60: O86–O88.
8. Liu, Y. Z., Hou, A. J., Ji, C. R. and Wu, Y. J. 1997. *Chinese Chemical Letters* 8: 39–40.
9. Hou, A. J., Wu, Y. J. and Liu, Y. Z. 1999. *Zhongcaoyao* 30: 645–648.
10. Hui, W. H., Li, M. M. and Luk, K. 1975. *Phytochemistry* 14: 833–834.
11. Hui, W. H. and Li, M. M. 1976. *Phytochemistry* 15: 1741–1743.
12. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.

Rhoeo spathacea (Sw.) Stearn

1. Hsuan, K., Chin, S. C. and Tan, H. T. W. 1998. The Concise Flora of Singapore volume II: Monocotyledons. Singapore University Press: Singapore.

2. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=5&taxon_id=242356469
   Date Accessed: 30th March 2008
3. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
4. Perez, G. R. M. 1996. *Phytomedicine* 3: 163–167.
5. Garcia, M., Miyares, C., Menendez, E. and Sainz, F. 1971. *Canadian Journal of Physiology and Pharmacology* 49: 1106–1110.
6. Weniger, B., Haag-Berrurier, M. and Anton, R. 1982. *Journal of Ethnopharmacology* 6: 67–84.

Ricinus communis L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008
   http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?31896
   Date Accessed: 30th June 2008
3. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
4. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
5. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
6. Selvaraj, O. S. and Bhalla, V. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 221–225. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
7. Chakravartula, S. V. and Guttarla, N. 2007. *Natural Product Research* 21: 1073–1077.
8. Zhang, X., Han, F., Gao, P., Yu, D. and Liu, S. 2007. *Natural Product Research* 2007 21: 982–989.
9. Foster, S. and Hobbs, C. 2002. A Field Guide to Western Medicinal Plants and Herbs. Houghton Mifflin Company: New York.
10. Ajose, F. O. 2007. *International Journal of Dermatology* 46: 48–55.
11. McGuffin, M., Hobbs, C., Upton, R. and Goldberg, A. (Eds). 1997. American Herbal Products Association's Botanical Safety Handbook. CRC Press: Boca Raton, FL.
12. Isichei, C. O., Das, S. C., Ogunkeye, O. O., Okwuasaba, F. K., Uguru, V. E., Onoruvwe, O., Olayinka, A. O., Dafur, S. J., Ekwere, E. O. and Parry, O. 2000. *Phytotherapy Research* 14: 40–42.
13. Makonnen, E., Zerihun, L., Assefa, G. and Rostom, A. A. 1999. *East African Medical Journal* 76: 335–337.
14. Okwuasaba, F. K., Osunkwo, U. A., Ekwenchi, M. M., Ekpenyong, K. I., Onwukeme, K. E., Olayinka, A. O., Uguru, M. O. and Das, S. C. 1991. *Journal of Ethnopharmacology* 34: 141–145.
15. Sandhyakumary, K., Bobby, R. G. and Indira, M. 2003. *Phytotherapy Research* 17: 508–511.
16. Okwusasaba, F. K., Das, S. C., Isichei, C. O., Ekwenchi, M. M., Onoryvwe, O., Olayinka, A. O., Uguru, V. E., Dafur, S. J., Ekwere, E. O. and Parry, O. 1997. *Phytotherapy Research* 11: 97–100.
17. Salhab, A. S., Shomaf, M. S., Gharaibeh, M. N. and Amer, N. A. 1999. *Contraception* 59: 395–399.
18. Raji, Y., Oloyo, A. K. and Morakinyo, A. O. 2006. *Asian Journal of Andrology* 8: 115–121.
19. Ilavarasan, R., Mallika, M. and Venkataraman, S. 2006. *Journal of Ethnopharmacology* 103: 478–480.
20. Ferraz, A. C., Angelucci, M. E., Da Costa, M. L., Batista, I. R., De Oliveira, B. H. and Da Cunha, C. 1999. *Pharmacology, Biochemistry and Behaviour* 63: 367–375.
21. Wang, H. X. and Ng, T. B. 2001. *Planta Medica* 67: 669–672.
22. Shukla, B., Visen, P. K. S., Patnaik, G. K., Kapoor, N. K. and Dhawan, B. N. 1992. *Drug Development Research* 26: 183–193.
23. Nisha, M., Kalyanasundaram, M., Paily, K. P., Abidha, Vanamail, P. and Balaraman, K. 2007. *Parasitology Research* 100: 575–579.
24. Upasani, S. M., Kotkar, H. M., Mendki, P. S. and Maheshwari, V. L. 2003. *Pest Management Science* 59: 1349–1354.

25. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
26. Fetrow, C. W. and Avila, J. R. 2000. The complete guide to herbal medicines. Springhouse Corp: Springhouse, PA.
27. Duke, J. A. 1985. CRC Handbook of Medicinal Herbs. CRC Press: Boca Raton, FL.
28. Balint, G. A. 1993. *Experimental Toxicology and Pathology* 45: 303–304.
29. Bradberry, S. M., Dickers, K. J., Rice, P., Griffiths, G. D. and Vale, J. A. 2003. *Toxicological Reviews* 22: 65–70.
30. Kumar, O., Sugendran, K. and Vijayaraghavan, R. 2003. *Toxicon* 41: 333–338.
31. Kumar, O., Lakshmana Rao, P. V., Pradhan, S., Jayaraj, R., Bhaskar, A. S., Nashikkar, A. B. and Vijayaraghavan, R. 2007. *Cellular and Molecular Biology* 53: 92–102.
32. Kanerva, L., Estlander, T. and Jolanki, R. 1990. *Journal of the American Academy of Dermatology* 23: 351–355.
33. Griffiths, G. D., Phillips, G. J. and Holley, J. 2007. *Inhalation Toxicology* 19: 873–887.
34. Kaur, C. and Ling, E. A. 1993. *Journal für Hirnforschung* 34: 493–501.
35. Aslani, M. R., Maleki, M., Mohri, M., Sharifi, K., Najjar-Nezhad, V. and Afshari, E. 2007. *Toxicon* 49: 400–406.
36. Wren, R. C. 1988. Potter's new cyclopaedia of botanical drugs and preparations. The C.W. Daniel Company Ltd.: Saffron Walden, Essex.
37. Skidmore-Roth L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.

Ruta graveolens L.

1. Tutin, T. G., Heywood, V. H., Burges, N. A., Moore, D. M., Valentine, D. H., Walters, S. M. and Webb, D. A. (Eds.). 1968. Flora Europaea 2, Rosaceae to Umbelliferae. Cambridge University Press: London.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?32578
   Date Accessed: 30th June 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
5. Raghav, S. K., Gupta, B., Shrivastava, A. and Das, H. R. 2007. *European Journal of Pharmacology* 560: 69–80.
6. Guarrera, P. M. 1999. *Journal of Ethnopharmacology* 68: 183–192.
7. Gravot, A., Larbat, R., Hehn, A., Lievre, K., Gontier, E., Goergen, J. L. and Bourgaud, F. 2004. *Archieves of Biochemistry and Biophysics* 422: 71–80.
8. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition. Medical Economics Company: New Jersey.
9. McGuffin, M., Hobbs, C., Upton, R. and Goldberg, A. (Eds). 1997. American Herbal Products Association's Botanical Safety Handbook. CRC Press: Boca Raton, FL.
10. Waldstein. A. 2006. *Journal of Ethnopharmacology* 108: 299–310.
11. Atta, A. H. and Alkofahi, A. 1998. *Journal of Ethnopharmacology* 60: 117–124.
12. Alzoreky, N. S. and Nakahara, K. 2003. *International Journal of Food Microbiology* 80: 223–230.
13. Trovato, A., Monforte, M. T., Rossitto, A. and Forestieri, A. M. 1996. *Bollettino Chimico Farmaceutico* 135: 263–266.
14. Pathak, S., Multani, A. S., Banerji, P. and Banerji, P. 2003. *International Journal of Oncology* 23: 975–982.
15. Es, S., Kuttan, G., Kc, P. and Kuttan, R. 2007. *Asian Pacific Journal of Cancer Prevention* 8: 390–394.
16. Preethi, K. C., Kuttan, G. and Kuttan, R. 2006. *Asian Pacific Journal of Cancer Prevention* 7: 439–443.
17. Réthy, B., Zupkó, I., Minorics, R., Hohmann, J., Ocsovszki, I. and Falkay, G. 2007. *Planta Medica* 73: 41–48.
18 Kong, Y. C., Lau, C. P., Wat, K. H., Ng, K. H., But, P. P., Cheng, K. F. and Waterman, P. G. 1989. *Planta Medica* 55: 176–178.
19. Prakash, A. O., Saxena, V., Shukla, S., Tewari, R. K., Mathur, S., Gupta, A., Sharma, S. and Mathur, R. 1985. *Acta Europaea Fertilitatis* 16: 441–448.

20. de Freitas, T. G., Augusto, P. M. and Montanari, T. 2005. *Contraception* 71: 74–77.
21. Raghav, S. K., Gupta, B., Agrawal, C., Goswami, K. and Das, H. R. 2006. *Journal of Ethnopharmacology* 104: 234–239.
22. Oliva, A., Meepagala, K. M., Wedge, D. E., Harries, D., Hale, A. L., Aliotta, G. and Duke, S. O. 2003. *Journal of Agriculture and Food Chemistry* 51: 890–896.
23. Meepagala, K. M., Schrader, K. K., Wedge, D. E. and Duke, S. O. 2005. *Phytochemistry* 66: 2689–2695.
24. Ivanova, A., Mikhova, B., Najdenski, H., Tsvetkova, I. and Kostova, I. 2005. *Fitoterapia* 76: 344–347.
25. Al-Heali, F. M. and Rahemo, Z. 2006. *Turkiye Parazitolojii Dergisi* 30: 272–274.
26. Skidmore-Roth L. 2001. Mosby's handbook of herbs and natural supplements. Mosby, Inc.: St. Louis.
27. Blumenthal, M., Busse, W. R., Goldberg, A., Gruenwald, J., Hall, T., Klein, S., Riggins, C. W. and Rister, R. S. (Eds.). 1998. The Complete German Commission E monographs, Therapeutic guide to herbal medicines. American Botanical Council, Boston: Integrative Medicine Communications: Austin, Texas.
28. Schempp, C. M., Schopf, E. and Simon, J. C. 1999. *Hautarzt* 50: 432–434.
29. Heskel, N. S., Amon, R. B., Storrs, F. J. and White, C. R. Jr. 1983. *Contact Dermatitis* 9: 278–280.
30. Sumner J. 2000. The Natural History of Medicinal Plants. Timber Press, Portland.
31. Furniss, D. and Adams, T. 2007. *Journal of Burn Care and Research* 28: 767–769.
32. Eickhorst, K., DeLeo, V. and Csaposs, J. 2007. *Dermatitis* 18: 52–55.
33. Seak, C. J. and Lin, C. C. 2007. *Clinical Toxicology (Philadelphia)* 45: 173–175.
34. El Agraa, S. E., el Badwi, S. M. and Adam, S. E. 2002. *Tropical Animal Health and Production* 34: 271–281.

Saccharum officinarum L.

1. Hsuan, K., Chin, S. C. and Tan, H. T. W. 1998. The Concise Flora of Singapore volume II: Monocotyledons. Singapore University Press: Singapore.
2. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200026232
   Date Accessed: 30th June 2008
3. Ledon, N., Casaco, A., Rodriguez, V., Cruz, J., Gonzalez, R., Tolon, Z., Cano, M. and Rojas, E. 2003. *Planta Medica* 69: 367–369.
4. Noa, M., Herrera, M., Magraner, J. and Mas, R. 1994. *Journal of Pharmacy and Pharmacology* 46: 282–285.
5. Noa, M., Mendoza, S., Mas, R., Mendoza, N. and Leon, F. 2004. *Drugs in R&D* 5: 281–290.
6. Duarte-Almeida, J. M., Negri, G., Salatino, A., de Carvalho, J. E. and Lajolo, F. M. 2007. *Phytochemistry* 68: 1165–1171.
7. Colombo, R., Yariwake, J. H., Queiroz, E. F., Ndjoko, K. and Hostettmann, K. 2006. *Phytochemical Analysis* 17: 337–343.
8. Colombo, R., Yariwake, J. H., Queiroz, E. F., Ndjoko, K. and Hostettmann, K. 2005. *Journal of Chromatography A* 1082: 51–59.
9. Ledón, N., Romay, Ch., Rodríguez, V., Cruz, J., Rodríguez, S., Ancheta, O., González, A., González, R., Tolón, Z., Cano, M., Rojas, E., Capote, A. and Valdes, T. 2005. *Planta Medica* 71: 126–129.
10. Balick, M. J., Kronenberg, F., Ososki, A. L., Reiff, M., Fugh-Berman, A., O'Connor, B., Roble, M., Lohr, P. and Atha, D. 2000. *Economic Botany* 54: 344–357.
11. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
12. Ledón, N., Casacó, A., Remirez, D., González, A., Cruz, J., González, R., Capote, A., Tolón, Z., Rojas, E., Rodríguez, V. J., Merino, N., Rodríguez, S., Ancheta, O. and Cano, M. C. 2007. *Phytomedicine* 14: 690–695.
13. Noa, M., Mendoza, S., Mas, R. and Mendoza, N. 2002. *Drugs under Experimental and Clinical Research* 28: 177–183.
14. Mendoza, S., Noa, M., Mas, R. and Mendoza, N. 2003. *International Journal of Tissue Reactions* 25: 91–98.
15. Gamez, R., Mas, R., Noa, M., Menendez, R., Garcia, H., Gonzalez, J., Perez, Y. and Goicochea, E. 2004. *Drugs under Experimental and Clinical Research* 30: 75–88.

16. Arruzazabala, M. L., Carbajal, D., Mas, R., Garcia, M. and Fraga, V. 1993. *Thrombosis Research* 69: 321–327.
17. Molina, V., Carbajal, D., Arruzazabala, L. and Más, R. 2005. *Journal of Medicinal Food* 8: 232–236.
18. Carbajal, D., Arruzazabala, M. L., Mas, R., Molina, V. and Valdes, S. 1994. *Prostaglandins, Leukotrienes, and Essential Fatty Acids* 50: 249–251.
19. Mas, R., Castano, G., Fernandez, J., Gamez, R., Illnait, J., Fernandez, L., Lopez, E., Mesa, M., Alvarez, E. and Mendoza, S. 2004. *Asia Pacific Journal of Clinical Nutrition* 13: S102.
20. Mas, R., Castano, G., Fernandez, J., Gamez, R. R., Illnait, J., Fernandez, L., Lopez, E., Mesa, M., Alvarez, E. and Mendoza, S. 2004. *Asia Pacific Journal of Clinical Nutrition* 13: S101.
21. Menendez, R., Amor, A. M., Rodeiro, I., Gonzalez, R. M., Gonzalez, P. C., Alfonso, J. L. and Mas, R. 2001. *Archieves of Medical Research* 32: 8–12.
22. Torres, O., Agramonte, A. J., Illnait, J., Mas Ferreiro, R., Fernandez, L. and Fernandez, J. C. 1995. *Diabetes Care* 18: 393–397.
23. Aneiros, E., Más, R., Calderón, B., Illnait, J., Fernández, L., Castaño, G. and Fernández, J. C. 1995. *Current Therapeutic Research* 56: 176–182.
24. Crespo, N., Alvarez, R., Más, R., Illnait, J., Fernández, L. and Fernández, J. C. 1997. *Current Therapeutic Research* 58: 44–51.
25. Hikosaka, K., Koyama, Y., Motobu, M., Yamada, M., Nakamura, K., Koge, K., Shimura, K., Isobe, T., Tsuji, N., Kang, C. B., Hayashidani, H., Wang, P. C., Matsumura, M. and Hirota, Y. 2006. *Bioscience, Biotechnology and Biochemistry* 70: 2853–2858.
26. Hikosaka, K., El-Abasy, M., Koyama, Y., Motobu, M., Koge, K., Isobe, T., Kang, C. B., Hayashidani, H., Onodera, T., Wang, P. C., Matsumura, M. and Hirota, Y. 2007. *Phytotherapy Research* 21: 120–125.
27. Amer, S., Na, K. J., El-Abasy, M., Motobu, M., Koyama, Y., Koge, K. and Hirota, Y. 2004. *International Immunopharmacology* 4: 71–77.
28. Noa, M., Mendoza, S. and Más, R. 2005. *Journal of Medicinal Food* 8: 237–241.
29. Boral, D., Chatterjee, S. and Bhattacharya, K. 2004. *Annals of Agriculture and Environmental Medicine* 11: 45–52.
30. Gamez, R., Aleman, C. L., Mas, R., Noa, M., Rodeiro, I., Garcia, H., Hernandez, C., Menendez, R. and Aguilar, C. 2001. *Journal of Medicinal Food* 4: 57–65.
31. Gamez, R., Mas, R., Noa, M., Menendez, R., Aleman, C., Acosta, P., Garcia, H., Hernandez, C., Amor, A., Perez, J. and Goicochea, 2000. *Toxicology Letters* 118: 31–41.
32. Aleman, C. L., Puig, M. N., Elias, E. C., Ortega, C. H., Guerra, I. R., Ferreiro, R. M. and Brinis, F. 1995. *Food and Chemical Toxicology* 33: 573–578.
33. Rodriguez, M. D. and Garcia, H. 1994. *Teratogen, Carcinogen and Mutagen* 14: 107–113.
34. Molina Cuevas, V., Arruzazabala, M. L., Carbajal Quintana, D., Mas Ferreiro, R. and Valdes Garcia, S. 1998. *Archieves of Medical Research* 29: 21–24.
35. Arruzazabala, M. L., Carbajal, D., Mas, R., Valdes, S. and Molina, V. 2001. *Journal of Medicinal Food* 4: 67–70.
36. Carbajal, D., Arruzazabala, M. L., Valdes, S. and Mas, R. 1998. *Pharmacology Research* 38: 89–91.

Sauropus androgynus (L.) Merr.

1. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200012609
Date Accessed: 30th March 2008
2. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
3. USDA, ARS GRIN Taxonomy for Plants 2008
http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?33189
Date Accessed: 30th June 2008
4. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
5. Kanchanapoom, T., Chumsri, P., Kasai, R., Otsuka, H. and Yamasaki, K. 2003. *Phytochemistry* 63: 985–988.

6. Ching, L. S. and Mohamed, S. 2001. *Journal of Agriculture and Food Chemistry* 49: 3101–3105.
7. Lin, C. Q., Lin, W. B., Pan, W. D. and Li, Y. J. 1999. *Redai Yaredai Zhiwu Xuebao* 7: 255–256.
8. Wang, P. H. and Lee, S. S. 1997. *Journal of the Chinese Chemical Society (Taipei)* 44: 145–149.
9. Liao, X. K. and Li, Y. H. 1996. *Redai Yaredai Zhiwu Xuebao* 4: 70–71.
10. Yu, S. F., Shun, C. T., Chen, T. M. and Chen, Y. H. 2006. *Biological and Pharmaceutical Bulletin* 29: 2510–2513.
11. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
12. Muhammad, M. M., Abdul, M. A., Mohd, A. A., Rozita, M. N., Nashriyah, B. M., Abdul, R. R. and Kazuyoshi, K. 1997. *Pesticide Science* 51: 165–170.
13. Padmavathi, P. and Rao, M. P. 1990. *Plant Foods for Human Nutrition (Dordrecht, Netherlands)* 40: 107–113.
14. Ger, L. P., Chiang, A. A., Lai, R. S., Chen, S. M. and Tseng, C. J. 1997. *American Journal of Epidemiology* 145: 842–849.
15. Oonakahara, K., Matsuyama, W., Higashimoto, I., Machida, K., Kawabata, M., Arimura, K., Osame, M., Hayashi, M., Ogura, T., Imaizumi, K. and Hasegawa, Y. 2005. *Respiration* 72: 221.
16. Hayashi, M., Tagawa, A., Ogura, T., Kozawa, S., Nakamura, M., Watanuki, Y. and Takahashi, H. 2007. *Nihon Kokyuki Gakkai Zasshi* 45: 81–86.
17. Chang, Y. L., Yao, Y. T., Wang, N. S. and Lee, Y. C. 1998. *American Journal of Respiratory and Critical Care Medicine* 157: 594–598.
18. Hsiue, T. R., Guo, Y. L., Chen, K. W., Chen, C. W., Lee, C. H. and Chang, H. Y. 1998. *Chest* 113: 71–76.
19. Wu, C. L., Hsu, W. H. and Chiang, C. D. 1998. *Zhonghua yi xue za zhi (Chinese medical journal, Free China Ed)* 61: 34–38.
20. Lin, T. J., Lu, C. C., Chen, K. W., Yueh, W. C. and Deng, J. F. 1997. *Toxicon* 35: 502.
21. Kao, C. H., Hom Y. J., Wu, C. L. and ChangLai, S. P. 1999. *Respiration* 66: 46–51.
22. Yu, S. F., Chen, T. M. and Chen, Y. H. 2007. *Journal of the Formosan Medical Association* 106: 537–547.

Sesbania grandiflora Pers.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?33770
   Date Accessed: 30th June 2008
3. Tiwari, R. D., Bajpai, R. K. and Khana, S. S. 1964. *Archiv der Pharmazie* 297: 310–312.
4. Tiwari, R. D. and Bajpai, R. K. 1964. *Proceedings of the National Academy of Sciences, India, Section A: Physical Sciences* 34: 239–244.
5. Saxena, V. K. and Mishra, L. N. 1999. *Research Journal of Chemistry and Environment* 3: 69–70.
6. Saxena, V. K. and Mishra, L. N. 1999. *Journal of the Institution of Chemists (India)* 71: 234–236.
7. Das, C. and Tripathi, A. K. 1999. *Oriental Journal of Chemistry* 15: 561–562.
8. Andal, K. R. and Sulochana, N. 1986. *Fitoterapia* 57: 293–294.
9. Lakshminarayana, R., Raju, M., Krishnakantha, T. P. and Baskaran, V. 2005. *Journal of Agriculture and Food Chemistry* 53: 2838–2842.
10. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
11. Pari, L. and Uma, A. 2003. *Therapie* 58: 439–443.
12. Nadkarni, K. M. 1982. Indian Materia Medica, Vol I. Popular Prakashan: Mumbai.
13. Solis, C. S. 1969. *Acta Manilana, Series A: Natural and Applied Sciences* 4: 52–109.
14. Kasture, V. S., Deshmukh, V. K. and Chopde, C. T. 2002. *Phytotherapy Research* 16: 455–460.
15. Fojas, F. R., Barrientos, C. M., Capal, T. V., Cruzada, S. F., Sison, F. M., Co, Y. C.,

Chua, N. G. and Gavina, T. L. 1982. *Philippine Journal of Science* 111: 157–181.

16. Boonmee, A., Reynolds, C. D. and Sangvanich, P. 2007. *Planta Medica* 73: 1197–1201.
17. Kalyanagurunathan, P., Sulochana, N. and Murugesh, N. 1985. *Fitoterapia* 56: 188–189.
18. Kumar, V. R., Murugesh, N., Vembar, S. and Damodaran, C. 1982. *Toxicology Letters* 10: 157–161.

Solanum nigrum L.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. eFloras.org 2008
   http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200020597
   Date Accessed: 17th April 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Fleming, T. (Ed.). 2000. PDR for Herbal Medicines, 2nd edition Medical Economics Company: New Jersey.
5. Goh, S. H. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 7–26. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
6. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
7. Wang, L. Y., Wang, N. L. and Yao, X. S. 2007. *Zhong Yao Cai* 30: 792–794.
8. Zhou, X. L., He, X. J., Wang, G. H., Gao, H., Zhou, G. X., Ye, W. C. and Yao, X. S. 2006. *Journal of Natural Products* 69: 1158–1163.
9. Zhou, X. L., He, X. J., Zhou, G. X., Ye, W. C. and Yao, X. S. 2007. *Journal of Asian Natural Products Research* 9: 517–523.
10. Sumner, J. 2000. The Natural History of Medicinal Plants. Timber Press, Portland.
11. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
12. Perry, L. M. 1980. Medicinal plants of East and Southeast Asia: Attributed properties and uses. The MIT Press: Cambridge, Massachusetts and London.
13. Rani, P. and Khullar, N. 2004. *Phytotherapy Research* 18: 670–673.
14. Prashanth Kumar, V., Shashidhara, S., Kumar, M. M. and Sridhara, B. Y. 2001. *Fitoterapia* 72: 481–486.
15. Zakaria, Z. A., Gopalan, H. K., Zainal, H., Mohd Pojan, N. H., Morsid, N. A, Aris, A. and Sulaiman, M. R. 2006. *Yakugaku Zasshi* 126: 1171–1178.
16. Heo, K. S., Lee, S. J. and Lim, K. T. 2004. *Environmental Toxicology and Pharmacology* 17: 45–54.
17. Lee, S. J., Oh, P. S., Ko, J. H., Lim, K. and Lim, K. T. 2004. *Cancer Chemotherapy and Pharmacology* 54: 562–572.
18. Son, Y. O., Kim, J., Lim, J. C., Chung, Y., Chung, G. H. and Lee, J. C. 2003. *Food and Chemical Toxicology* 41: 1421–1428.
19 Hu, K., Kobayashi, H., Dong, A., Jing, Y., Iwasaki, S. and Yao X. 1999. *Planta Medica* 65: 35–38.
20. Saijo, R., Murakami, K., Nohara, T., Tomimatsu, T., Sato, A. and Matsuoka, K. 1982. *Yakugaku Zasshi* 102: 300–305.
21. Ji, Y. B., Gao, S. Y., Ji, C. F. and Zou, X. 2008. *Journal of Ethnopharmacology* 115: 194–202.
22. Jeong, J. B., Jeong, H. J., Park, J. H., Lee, S. H., Lee, J. R., Lee, H. K., Chung, G. Y., Choi, J. D. and de Lumen, B. O. 2007. *Journal of Agricultural and Food Chemistry* 55: 10707–10713.
23. Li, J., Li, Q., Feng, T., Zhang, T., Li, K., Zhao, R., Han, Z. and Gao, D. 2007. *Phytotherapy Research* 21: 832–840.
24. Lin, H. M., Tseng, H. C., Wang, C. J., Chyau, C. C., Liao, K. K., Peng, P. L. and Chou, F. P. 2007. *Journal of Agricultural and Food Chemistry* 55: 3620–3628.
25. Lee, S. J. and Lim, K. T. 2008. *Journal of Nutritional Biochemistry* 19: 166–174.
26. Oh, P. S. and Lim, K. T. 2007. *Journal of Biomedical Science* 14: 223–232.
27. Lee, S. J. and Lim, K. T. 2006. *Cancer Chemotherapy and Pharmacology* 57: 507–516.

28. Lee, S. J., Ko, J. H. and Lim, K. T. 2005. *Oncology Reports* 14: 789–796.
29. Lim, K. T. 2005. *Journal of Medicinal Food* 8: 215–226.
30. Heo, K. S. and Lim, K. T. 2005. *Journal of Medicinal Food* 8: 69–77.
31. An, L., Tang, J. T., Liu, X. M. and Gao, N. N. 2006. *Zhongguo Zhong Yao Za Zhi* 31: 1225–1226.
32. An, H. J., Kwon, K. B., Cho, H. I., Seo, E. A., Ryu, D. G., Hwang, W. J., Yoo, S. J., Kim, Y. K., Hong, S. H. and Kim, H. M. 2005. *European Journal of Cancer Prevention* 14: 345–350.
33. Wang, W. and Lu, D. P. 2005. *Beijing Da Xue Xue Bao* 37: 240–244.
34. Heo, K. S. and Lim, K. T. 2004. *Journal of Medicinal Food* 7: 349–357.
35. Al-Fatimi, M., Wurster, M., Schröder, G. and Lindequist, U. 2007. *Journal of Ethnopharmacology* 111: 657–166.
36. Akhtar, M. S. and Munir, M. 1989. *Journal of Ethnopharmacology* 27: 163–176.
37. Jainu, M. and Devi, C. S. 2006. *Journal of Ethnopharmacology* 104: 156–163.
38. Ikeda, T., Ando, J., Miyazono, A., Zhu, X. H., Tsumagari, H., Nohara, T., Yokomizo, K. and Uyeda, M. 2000. *Biological & Pharmaceutical Bulletin* 23: 363–364.
39. Perez, R. M., Perez, J. A., Garcia, L. M. and Sossa, H. 1998. *Journal of Ethnopharmacology* 62: 43–48.
40. Raju, K., Anbuganapathi, G., Gokulakrishnan, V., Rajkapoor, B., Jayakar, B. and Manian, S. 2003. *Biological & Pharmaceutical Bulletin* 26: 1618–1619.
41. Sultana, S., Perwaiz, S., Iqbal, M. and Athar, M. 1995. *Journal of Ethnopharmacology* 45: 189–192.
42. Lin, H. M., Tseng, H. C., Wang, C. J., Lin, J. J., Lo, C. W. and Chou, F. P. 2008. *Chemico-biological Interactions* 171: 283–293.
43. Huseini, H. F., Alavian, S. M., Heshmat, R., Heydari, M. R. and Abolmaali, K. 2005. *Phytomedicine* 12: 619–624.
44. Lee, S. J., Ko, J. H., Lim, K. and Lim, K. T. 2005. *Pharmacological Research* 51: 399–408.
45. Yen, G. C., Chen, H. Y. and Peng, H. H. 2001. *Food and Chemical Toxicology* 39: 1045–1053.
46. Moundipa, P. F. and Domngang, F. M. 1991. *British Journal of Nutrition* 65: 81–91.
47. Ahmed, A. H., Kamal, I. H. and Ramzy, R. M. 2001. *Journal of the Egyptian Society of Parasitology* 31: 843–852.
48. Ahmed, A. H. and Rifaat M. M. 2004. *Journal of the Egyptian Society of Parasitology* 34: 1041–1050.
49. El-Ansary, A. K. and Al Daihan, S. K. 2007. *Journal of the Egyptian Society of Parasitology* 37: 39–50.
50. Ahmed, A. H. and Rifaat M. M. 2005. *Journal of the Egyptian Society of Parasitology* 35: 33–40.
51. Dugan, G. M. and Gumbmann, M. R. 1990. *Food and Chemical Toxicology* 28: 101–117.
52. Duke, J. A. 1985. CRC Handbook of Medicinal Herbs. CRC Press: Boca Raton, Florida.

Swietenia macrophylla King

1. Wiart, C. 2006. Medicinal Plants of the Asia-Pacific: Drugs for the future? World Scientific Publishing Co. Pte. Ltd.: Singapore.
2. Connolly, J. D., McCrindle, R., Overton, K. H. and Warnock, W. D. C. 1965. *Tetrahedron Letters* 33: 2937–2940.
3. Solomon, K. A., Malathi, R., Rajan, S. S., Narasimhan, S. and Nethaji, M. 2003. *Acta Crystallographica E* 59: O1519–O1521.
4. Connolly, J. D., Henderson, R., McCrindle, R., Overton, K. H. and Bhacca, N. S. 1964. *Tetrahedron Letters* 5: 2593–2597.
5. Taylor Anne, R. H. and Taylor David, A. H. 1983. *Phytochemistry* 22: 2870–2871.
6. Chan, K. C., Tang, T. S. and Toh, H. T. 1976. *Phytochemistry* 15: 429–430.
7. Mootoo, B. S., Ali, A., Motilal, R., Pingal, R., Ramlal, A., Khan, A., Reynolds, W. F. and McLean, S. 1999. *Journal of Natural Products* 62: 1514–1517.
8. Tang, T. S. 1976. Some chemical studies of oil palm & *Swietenia macrophylla* (Master's Thesis). University of Malaya (Malaysia), Faculty of Science.
9. Kadota, S., Marpaung, L., Kikuchi, T. and Ekimoto, H. 1990. *Chemical and Pharmaceutical Bulletin* 38: 894–901.
10. Munoz, V., Sauvain, M., Bourdy, G., Callapa, J., Rojas, I., Vargas, L., Tae, A. and Deharo, E.

2000. *Journal of Ethnopharmacology* 69: 139–155.

11. Sodiero, I., Padmawinata, K., Wattimena, J. R. and Rekita, S. 1990. *Acta Pharmaceutica Indonesia* 15: 1–13.
12. Maiti, A., Dewanjee, S. and Mandal, S. C. 2007. *Tropical Journal of Pharmaceutical Research* 6: 711–716.
13. Bourdy, G., DeWalt, S. J., Chávez de Michel, L. R., Roca, A., Deharo, E., Muñoz, V., Balderrama, L., Quenevo, C. and Gimenez, A. 2000. *Journal of Ethnopharmacology* 70: 87–109.

Terminalia catappa L.

1. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
2. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
3. Wee, Y. C. and Corlett, R. 1986. The City and the Forest: Plant Life in Urban Singapore. Singapore University Press: National University of Singapore.
4. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?36334
   Date Accessed: 30th June 2008
5. Fan, Y. M., Xu, L. Z., Gao, J., Wang, Y., Tang, X. H., Zhao, X. N. and Zhang, Z. X. 2004. *Fitoterapia* 75: 253–260.
6. Masuda, T., Yonemori, S., Oyama, Y., Takeda, Y., Tanaka, T., Andoh, T., Shinohara, A. and Nakata, M. 1999. *Journal of Agricultural and Food Chemistry* 47: 1749–1754.
7. Lin, C. C., Hsu, Y. F., Lin, T. C., Hsu, F. L. and Hsu, H. Y. 1998. *The Journal of Pharmacy and Pharmacology* 50: 789–794.
8. Nagappa, A. N., Thakurdesai, P. A., Venkat Rao, N. and Singh, J. 2003. *Journal of Ethnopharmacology* 88: 45–50.
9. Mau, J. L., Ko, P. T. and Chyau, C. C. 2003. *Food Research International* 36: 97–104.
10. Ko, T. F., Weng, Y. M. and Chiou, R. Y. 2002. *Journal of Agriculture and Food Chemistry* 50: 5343–5348.
11. Kinoshita, S., Inoue, Y., Nakama, S., Ichiba, T. and Aniya, Y. 2007. *Phytomedicine* 14: 755–762.
12. Lin, Y. L., Kuo, Y. H., Shiao, M. S., Chen, C. C. and Ou, J. C. 2000. *Journal of the Chinese Chemical Society* 47: 253–256.
13. Lin, T. C. and Hsu, F. L. 1999. *Journal of the Chinese Chemical Society* 46: 613–618.
14. Woodley, E. (Ed.). 1991. Medicinal plants of Papua New Guinea, part I. Veerlag Josef Margraf Scientific Books: Weikersheim.
15. Ghazanfar, S. A. 1994. Handbook of Arabian Medicinal Plants. CRC Press: Boca Raton, FL.
16. Jain, S. K. and DeFilipps, R. A. 1991. Medicinal Plants of India, volume one and two. Reference Publications Inc.: United States of America.
17. Pawar, S. P. and Pal, S. C. 2002. *Indian Journal of Medical Sciences* 56: 276–278.
18. Goun, E., Cunningham, G., Chu, D., Nguyen, C. and Miles, D. 2003. *Fitoterapia* 74: 592–596.
19. Kloucek, P., Polesny, Z., Svobodova, B., Vlkova, E. and Kokoska, L. 2005. *Journal of Ethnopharmacology* 99: 309–312.
20. Chu, S. C., Yang, S. F., Liu, S. J., Kuo, W. H., Chang, Y. Z. and Hsieh, Y. S. 2007. *Food and Chemical Toxicology* 45: 1194–1201.
21. Chen, P. S. and Li, J. H. 2006. *Toxicology Letters* 163: 44–53.
22. Lin, C. C., Hsu, Y. F. and Lin, T. C. 1999. *American Journal of Chinese Medicine* 27: 371–376.
23. Chyau, C. C., Ko, P. T. and Mau, J. L. 2006. *LWT — Food Science and Technology* 39: 1099–1108.
24. Chen, P. S., Li, J. H., Liu, T. Y. and Lin, T. C. 2000. *Cancer Letters* 152: 115–122.
25. Tang, X. H., Gao, J., Wang, Y. P., Xu, L. Z., Zhao, X. N. and Xu, Q. 2003. *Zhongguo Zhong Yao Za Zhi* 28: 1170–1174.
26. Tang, X. H., Gao, L., Gao, J., Fan, Y. M., Xu, L. Z., Zhao, X. N. and Xu, Q. 2004. *American Journal of Chinese Medicine* 32: 509–519.
27. Tang, X. H., Gao, J., Wang, Y., Fan, Y. M., Xu, L. Z., Zhao, X., Xu, Q. and Qian Z. M. 2006. *The Journal of Nutritional Biochemistry* 17: 177–182.
28. Gao, J., Tang, X., Dou, H., Fan, Y., Zhao, X. and Xu, Q. 2004. *Journal of Pharmacy and Pharmacology* 56: 1449–1455.

29. Gao, J., Chen, J., Tang, X., Pan, L., Fang, F., Xu, L., Zhao, X. and Xu, Q. 2006. *Journal of Pharmacy and Pharmacology* 58: 227–233.
30. Nonaka, G., Nishioka, I., Nishizawa, M., Yamagishi, T., Kashiwada, Y., Dutschman, G. E., Bodner, A. J., Kilkuskie, R. E., Cheng, Y. C. and Lee, K. H. 1990. *Journal of Natural Products* 53: 587–595.
31. Morioka, T., Suzui, M., Nabandith, V., Inamine, M., Aniya, Y., Nakayama, T., Ichiba, T. and Yoshimi, N. 2005. *European Journal of Cancer Prevention* 14: 101–105.
32. Ratnasooriya, W. D. and Dharmasiri, M. G. 2000. *Asian Journal of Andrology* 2: 213–219.
33. Ko, T. F., Weng, Y. M., Lin, S. B. and Chiou, R. Y. 2003. *Journal of Agriculture and Food Chemistry* 51: 3564–3567.
34. Morton, J. F. 1981. Atlas of Medicinal Plants of Middle America: Bahamas to Yucatan. C.C. Thomas Publishers Co.: Springfield, IL.
35. Oliveira, J. T. A., Vasconcelos, I. M., Bezerra, L. C. N. M., Silveira, S. B., Monteiro, A. C. O. and Moreira, R. A. 2000. *Food Chemistry* 70: 185–191.
36. Lin, C. C., Hsu, Y. F., Lin, T. C. and Hsu, H. Y. 2001. *Phytotherapy Research* 15: 206–212.

Thevetia peruviana (Pers.) K. Schum.

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. eFloras.org 2008
http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200018473
Date Accessed: 30th March 2008
3. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
4. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
5. Wiart, C. 2000. Medicinal plants of Southeast Asia. Pelanduk Publications (M) Sdn Bhd: Malaysia.
6. Tewtrakul, S., Nakamura, N., Hattori, M., Fujiwara, T. and Supavita, T. 2002. *Chemical & Pharmaceutical Bulletin* 50: 630–635.
7. Morton, J. F. 1981. Atlas of Medicinal Plants of Middle America: Bahamas to Yucatan. C.C. Thomas Publishers Co.: Springfield, IL.
8. Gata-Goncalves, L., Nogueira, J. M., Matos, O. and Bruno de Sousa, R. 2003. *Journal of Photochemistry and Photobiology B* 70: 51–54.
9. Lapcharoen, P., Apiwathnasorn, C., Komalamisra, N., Dekumyoy, P., Palakul, K. and Rongsriyam, Y. 2005. *Southeast Asian Journal of Tropical Medicine and Public Health* 36: 167–175.
10. Singh, A. and Singh, S. K. 2005. *Fitoterapia* 76: 747–751.
11. Eddleston, M., Ariaratnam, C. A., Sjostrom, L., Jayalath, S., Rajakanthan, K., Rajapakse, S., Colbert, D., Meyer, W. P., Perera, G., Attapattu, S., Kularatne, S. A., Sheriff, M. R. and Warrell, D. A. 2000. *Heart* 83: 301–306.
12. Eddleston, M., Ariaratnam, C. A., Meyer, W. P., Perera, G., Kularatne, A. M., Attapattu, S., Sheriff, M. H. and Warrell, D. A. 1999. *Tropical Medicine and International Health* 4: 266–273.
13. Duke, J. A. 1985. CRC Handbook of Medicinal Herbs. CRC Press: Boca Raton, FL.
14. Samal, K. K., Sahu, H. K., Kar, M. K., Palit, S. K., Kar, B. C. and Sahu, C. S. 1989. *Journal of the Associations of Physicians of India* 37: 232–233.
15. Maringhini, G., Notaro, L., Barberi, O., Giubilato, A., Butera, R. and Di Pasquale, P. 2002. *Italian Heart Journal* 3: 137–140.
16. Camphausen, C., Haas, N. A. and Mattke, A. C. 2005. *Zeitschrift für Kardiologie* 94: 817–823.
17. Sambasivam, S., Karpagam, G., Chandran, R. and Khan, S. A. 2003. *Journal of Environmental Biology* 24: 201–204.
18. Singh, D. and Singh, A. 2005. *Chemosphere* 60: 135–140.
19. Singh, D. and Singh, A. 2002. *Chemosphere* 49: 45–49.
20. Oji, O. and Okafor, Q. E. 2000. *Phytotherapy Research* 14: 133–135.
21. Bose, T. K., Basu, R. K., Biswas, B., De, J. N., Majumdar, B. C. and Datta, S. 1999. *Journal of the Indian Medical Association* 97: 407–410.

22. Saravanapavananthan, T. 1985. *Jaffna Medical Journal* 20: 17–21.
23. Nguyen, V. D. and Doan, T. N. 1989. Medicinal Plants in Vietnam. World Health Organisation, Regional Office for the Western Pacific, Manila, Institute of Materia Medica, Hanoi.
24. Kee, C. H. 1998. The Pharmacology of Chinese Herbs, 2nd edition. CRC Press: Boca Raton, Fl.

Tinospora crispa (L.) Diels

1. Hsuan, K. 1990. The Concise Flora of Singapore, Gymnosperms and Dicotyledons. Singapore University Press: National University of Singapore.
2. Nguyen, V. D. 1993. Medicinal Plants of Vietnam, Cambodia and Laos. Nguyen Van Duong: Vietnam.
3. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?428217
   Date Accessed: 30th June 2008
4. Kongkathip, N., Dhumma-upakorn, P., Kongkathip, B., Chawananoraset, K., Sangchomkaeo, P. and Hatthakitpanichakul, S. 2002. *Journal of Ethnopharmacology* 83: 95–99.
5. Goh, S. H., Chuah, C. H., Mok, J. S. L. and Soepadmo, E. 1995. Malaysian Medicinal Plants for the Treatment of Cardiovascular Diseases. Pelanduk Publications (M) Sdn Bhd: Malaysia.
6. Kalsom, Y. U. and Noor, H. 1995. *Fitoterapia* 66: 280.
7. Cavin, A., Hostettmann, K., Dyatmyko, W. and Potterat, O. 1998. *Planta Medica* 64: 393–396.
8. Pachaly, P., Adnan, A. Z. and Will, G. 1992. *Planta Medica* 58: 184–187.
9. Marzuki, A. and Iljaz, S. 1988. Proceedings: Malaysian Traditional Medicine, Kuala Lumpur, pp. 183–191. Institute of Advanced Studies, University of Malaya, Kuala Lumpur.
10. Noor, H., Hammonds, P., Sutton, R. and Ashcroft, S. J. 1989. *Diabetologia* 32: 354–359.
11. Higashino, H., Suzuki, A., Tanaka, Y. and Pootakham, K. 1992. *Nippon Yakurigaku Zasshi* 100: 339–344.
12. Rahman, N. N. N. A., Furuta, T., Kojima, S., Takane, K. and Mustafa Ali Mohd. 1999. *Journal of Ethnopharmacology* 64: 249–254.
13. Bertani, S., Bourdy, G., Landau, I., Robinson, J. C., Esterre, P. and Deharo, E. 2005. *Journal of Ethnopharmacology* 98: 45–54.
14. Zaridah, M. Z., Idid, S. Z., Omar, A. W. and Khozirah, S. 2001. *Journal of Ethnopharmacology* 78: 79–84.
15. Noor, H. and Ashcroft, S. J. 1998. *Journal of Ethnopharmacology* 62: 7–13.
16. Noor, H. and Ashcroft, S. J. 1989. *Journal of Ethnopharmacology* 27: 149–161.
17. Sangsuwan, C., Udompanthurak, S., Vannasaeng, S. and Thamlikitkul, V. 2004. *Journal of the Medical Association of Thailand* 87: 543–546.
18. Subehan, Usia, T., Iwata, H., Kadota, S. and Tezuka, Y. 2006. *Journal of Ethnopharmacology* 105: 449–455.

Vitex rotundifolia L. f.

1. Byung, H. H., Youngbae, S. and Hyung-Joon, C. 1998. Medicinal plants in the Republic of Korea. World Health Organization, Regional Office for the Western Pacific, Manila, Natural Products Research Institute, Seoul National University.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?41839
   Date Accessed: 2nd July 2008
3. Chen, Y. L., Wang, Y. S., Kao, B. C. and Chang, Y. A. 1978. *Zhongguo Nongye Huaxue Huizhi* 16: 99–102.
4. Tada, H. and Yasuda, F. 1984. *Heterocycles* 22: 2203–2205.
5. Ono, M., Ito, Y. and Nohara, T. 2001. *Chemical and Pharmaceutical Bulletin (Tokyo)* 49: 1220–1222.
6. Ono, M., Yamamoto, M., Masuoka, C., Ito, Y., Yamashita, M. and Nohara, T. 1999. *Journal of Natural Products* 62: 1532–1537.
7. Ono, M., Yanaka, T., Yamamoto, M., Ito, Y. and Nohara, T. 2002. *Journal of Natural Products* 65: 537–541.
8. Hu, Y., Xin, H. L., Zhang, Q. Y., Zheng, H. C., Rahman, K. and Qin, L. P. 2007. *Phytomedicine* 14: 668–674.

9. Yoshioka, T., Inokuchi, T., Fujioka, S. and Kimura, Y. 2004. *Zeitschrift für Naturforschung. C, Journal of biosciences* 59: 509–514.
10. de Padua, L. S., Bunyapraphatsara, N. and Lemmens, R. H. M. J. (Eds). 1999. Plant resources of South-East Asia, No. 12, Medicinal and Poisonous Plants 1. Backhuys Publishers: Leiden, the Netherlands.
11. Li, W. X., Cui, C. B., Cai, B. and Yao, X. S. 2005. *Journal of Asian Natural Product Research* 7: 95–105.
12. Okuyama, E., Fujimori, S., Yamazaki, M. and Deyama, T. 1998. *Chemical and Pharmaceutical Bulletin (Tokyo)* 46: 655–662.
13. Hossain, M. M., Paul, N., Sohrab, M. H., Rahman, E. and Rashid, M. A. 2001. *Fitoterapia* 72: 695–697.
14. Kawazoe, K., Yutani, A., Tamemoto, K., Yuasa, S., Shibata, H., Higuti, T. and Takaishi, Y. 2001. *Journal of Natural Products* 64: 588–591.
15. Hernandez, M. M., Heraso, C., Villarreal, M. L., Vargas-Arispuro, I. and Aranda, E. 1999. *Journal of Ethnopharmacology* 67: 37–44.
16. Shin, T. Y., Kim, S. H., Lim, J. P., Suh, E. S., Jeong, H. J., Kim, B. D., Park, E. J., Hwang, W. J., Rye, D. G., Baek, S. H., An, N. H. and Kim, H. M. 2000. *Journal of Ethnopharmacology* 72: 443–450.
17. You, K. M., Son, K. H., Chang, H. W., Kang, S. S. and Kim, H. P. 1998. *Planta Medica* 64: 546–550.
18. Ko, W. G., Kang, T. H., Lee, S. J., Kim, N. Y., Kim, Y. C., Sohn, D. H. and Lee, B. H. 2000. *Food and Chemical Toxicology* 38: 861–865.
19. Ko, W. G., Kang, T. H., Lee, S. J., Kim, Y. C. and Lee, B. H. 2001. *Phytotherapy Research* 15: 535–537.
20. Kobayakawa, J., Sato-Nishimori, F., Moriyasu, M. and Matsukawa, Y. 2004. *Cancer Letters* 208: 59–64.
21. Kiuchi, F., Matsuo, K., Ito, M., Qui, T. K. and Honda, G. 2004. *Chemical and Pharmaceutical Bulletin (Tokyo)* 52: 1492–1494.
22. Ko, S. T., Moon, Y. H. and Ko, O. H. 1977. *Saengyak Hakhoechi* 8: 55–60.
23. Miyazawa, M., Shimamura, H., Nakamura, S. and Kameoka, H. 1995. *Journal of Agricultural and Food Chemistry* 43: 3012–3015.

Zingiber officinale Roscoe.

1. Palaniswamy, U. R. 2003. A guide to Medicinal Plants of Asian origin and culture. CPL Press: United States of America.
2. USDA, ARS GRIN Taxonomy for Plants 2008 http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?42254
   Date Accessed: 30th June 2008
3. Li, T. S. C. 2000. Medicinal Plants. Culture, Utilisation and Phytopharmacology. Technomic Publishing Co., Inc.: Lancaster.
4. Ma, J., Jin, X., Yang, L. and Liu, Z. L. 2004. *Phytochemistry* 65: 1137–1143.
5. Jolad, S. D., Lantz, R. C., Chen, G. J., Bates, R. B. and Timmermann, B. N. 2005. *Phytochemistry* 66: 1614–1635.
6. Tao, Q. F., Xu, Y., Lam, R. Y., Schneider, B., Dou, H., Leung, P. S., Shi, S. Y., Zhou, C. X., Yang, L. X., Zhang, R. P., Xiao, Y. C., Wu, X., Stöckigt, J., Zeng, S., Cheng, C. H. and Zhao, Y. 2008. *Journal of Natural Products* 71: 12–17.
7. Sekiwa, Y., Kubota, K. and Kobayashi, A. 2000. *Journal of Agriculture and Food Chemistry* 48: 373–377.
8. Duke, J. A. and Ayensu, E. S. 1985. Medicinal Plants of China, volume one and two. Reference Publications, Inc.: United States of America.
9. Wee, Y. C. and Hsuan, K. 1990. An illustrated Dictionary of Chinese Medicinal Herbs. Times Edition and Eu Yan Seng Holdings Ltd: Singapore.
10. Wee, Y. C. 1992. A Guide to Medicinal Plants. Singapore Science Centre Publication: Singapore.
11. Blumenthal, M., Goldberg, A. and Gruenwald, J. (Eds.). 2000. Herbal Medicine. Expanded Commission E Monographs. Integrative Medicine Communications: Austin, Texas.
12. Bradley, P. R. (Ed.). 1992. British Herbal Compendium, vol. 1. British Herbal Medicine Association: Bournemouth.
13. ESCOP. 1997. “Zingiberis rhizoma.” Monographs on the Medicinal Uses of Plant Drugs. European Scientific Cooperative on Phytotherapy: Exeter, UK.
14. Ojewole, J. A. 2006. *Phytotherapy Research* 20: 764–772.
15. Young, H. Y., Luo, Y. L., Cheng, H. Y., Hsieh, W. C., Liao, J. C. and Peng, W. H. 2005. *Journal of Ethnopharmacology* 96: 207–210.

16. Cady, R. K., Schreiber, C. P., Beach, M. E. and Hart, C. C. 2005. *Medical Science Monitor* 11: PI65–69.
17. Iqbal, Z., Lateef, M., Akhtar, M. S., Ghayur, M. N. and Gilani, A. H. 2006. *Journal of Ethnopharmacology* 106: 285–287.
18. Phan, P. V., Sohrabi, A., Polotsky, A., Hungerford, D. S., Lindmark, L. and Frondoza, C. G. 2005. *Journal of Alternative and Complementary Medicine* 11: 149–154.
19. Shukla, Y., Prasad, S., Tripathi, C., Singh, M., George, J. and Kalra, N. 2007. *Molecular Nutrition and Food Research* 51: 1492–1502.
20. Shukla, Y. and Singh, M. 2007. *Food and Chemical Toxicology* 45: 683–690.
21. Sagar, S. M., Yance, D. and Wong, R. K. 2006. *Current Oncology* 13: 99–107.
22. Sagar, S. M., Yance, D. and Wong, R. K. 2006. *Current Oncology* 13: 14–26.
23. Yance, D. R. Jr. and Sagar, S. M. 2006. *Integrative Cancer Therapy* 5: 9–29.
24. Rhode, J., Fogoros, S., Zick, S., Wahl, H., Griffith, K. A., Huang, J. and Liu, J. R. 2007. *BMC Complementary and Alternative Medicine* 20: 7: 44.
25. Lee, S. H., Cekanova, M. and Baek, S. J. 2008. *Molecular Carcinogenesis* 47: 197–208.
26. Vijaya Padma, V., Arul Diana Christie, S. and Ramkuma, K. M. 2007. *Basic and Clinical Pharmacology and Toxicology* 100: 302–307.
27. Hsu, M. H., Kuo, S. C., Chen, C. J., Chung, J. G., Lai, Y. Y. and Huang, L. J. 2005. *Leukemia Research* 29: 1399–1406.
28. Hofbauer, S., Kainz, V., Golser, L., Klappacher, M., Kiesslich, T., Heidegger, W., Krammer, B., Hermann, A. and Weiger, T. M. 2006. *Forsch Komplementmed* 13: 18–22.
29. Chen, C. Y., Liu, T. Z., Liu, Y. W., Tseng, W. C., Liu, R. H., Lu, F. J., Lin, Y. S., Kuo, S. H. and Chen, C. H. 2007. *Journal of Agriculture and Food Chemistry* 55: 948–954.
30. Kim, E. C., Min, J. K., Kim, T. Y., Lee, S. J., Yang, H. O., Han, S., Kim, Y. M., Kwon, Y. G. 2005. *Biochemical and Biophysical Research Communications* 335: 300–308.
31. Ishiguro, K., Ando, T., Maeda, O., Ohmiya, N., Niwa, Y., Kadomatsu, K. and Goto, H. 2007. *Biochemical and Biophysical Research Communications* 362: 218–223.
32. Vimala, S., Norhanom, A. W. and Yadav, M. 1999. *British Journal of Cancer* 80: 110–116.
33. Surh, Y. J., Park, K. K., Chun, K. S., Lee, L. J., Lee, E. and Lee, S. S. 1999. *Journal of Environmental Pathology, Toxicology and Oncology* 18: 131–139.
34. Al-Amin, Z. M., Thomson, M., Al-Qattan, K. K., Peltonen-Shalaby, R. and Ali, M. 2006. *British Journal of Nutrition* 96: 660–666.
35. Kato, A., Higuchi, Y., Goto, H., Kizu, H., Okamoto, T., Asano, N., Hollinshead, J., Nash, R. J. and Adachi, I. 2006. *Journal of Agriculture and Food Chemistry* 54: 6640–6644.
36. Chen, J. C., Huang, L. J., Wu, S. L., Kuo, S. C., Ho, T. Y. and Hsiang, C.Y. 2007. *Journal of Agriculture and Food Chemistry* 55: 8390–8397.
37. Keating, A. and Chez, R. A. 2002. *Alternative Therapies in Health and Medicine* 8: 89–91.
38. Eberhart, L. H., Mayer, R., Betz, O., Tsolakidis, S., Hilpert, W., Morin, A. M., Geldner, G., Wulf, H. and Seeling, W. 2003. *Anesthesia and Analgesia* 96: 995–998.
39. Sharma, S. S., Kochupillai, V., Gupta, S. K., Seth, S. D. and Gupta, Y. K. 1997. *Journal of Agriculture and Food Chemistry* 57: 93–96.
40. Sharma, S. S. and Gupta, Y. K. 1998. *Journal of Ethnopharmacology* 62: 49–55.
41. Vishwakarma, S. L., Pal, S. C., Kasture, V. S. and Kasture, S. B. 2002. *Phytotherapy Research* 16: 621–626.
42. Abdel-Aziz, H., Windeck, T., Ploch, M. and Verspohl, E. J. 2006. *European Journal of Ethnopharmacology* 530: 136–143.
43. Abdel-Aziz, H., Nahrstedt, A., Petereit, F., Windeck, T., Ploch, M. and Verspohl, E. J. 2005. *Planta Medica* 71: 609–616.
44. Borrelli, F., Capasso, R., Aviello, G., Pittler, M. H. and Izzo, A. A. 2005. *Obstetrics and Gynecology* 105: 849–856.
45. Chittumma, P., Kaewkiattikun, K. and Wiriyasiriwach, B. 2007. *Journal of the Medical Association of Thailand* 90: 15–20.
46. Apariman, S., Ratchanon, S. and Wiriyasirivej, B. 2006. *Journal of the Medical Association of Thailand* 89: 2003–2009.
47. Nanthakomon, T. and Pongrojpaw, D. 2006. *Journal of the Medical Association of Thailand* 89: S130–136.

48. Pongrojpaw, D., Somprasit, C. and Chanthasenanont, A. 2007. *Journal of the Medical Association of Thailand* 90: 1703–1709.
49. Ben-Arye, E., Oren, A. and Ben-Arie, A. 2006. *Harefuah* 145: 738–742.
50. Chaiyakunapruk, N., Kitikannakorn, N., Nathisuwan, S., Leeprakobboon, K. and Leelasettagool, C. 2006. *American Journal of Obstetric and Gynecology* 194: 95–99.
51. Boone, S. A. and Shields, K. M. 2005. *Annals of Pharmacotherapy* 39: 1710–1713.
52. Thompson, H. J. and Potter, P. J. 2006. *Evidence Based Nursing* 9: 80.
53. Bryer, E. 2005. *Journal of Midwifery and Women's Health* 50: e1–3.
54. Kadnur, S. V. and Goyal, R. K. 2005. *Indian Journal of Experimental Biology* 43: 1161–1164.
55. Ghayur, M. N. and Gilani, A. H. 2005. *Journal of Cardiovascular Pharmacology* 45: 74–80.
56. Kim, S. O., Chun, K. S., Kundu, J. K. and Surh, Y. J. 2004. *Biofactors* 21: 27–31.
57. Shen, C. L., Hong, K. J. and Kim, S. W. 2003. *Journal of Medicinal Food* 6: 323–328.
58. Penna, S. C., Medeiros, M. V., Aimbire, F. S., Faria-Neto, H. C., Sertie, J. A. and Lopes-Martins, R. A. 2003. *Phytomedicine* 10: 381–385.
59. Thomson, M., Al-Qattan, K. K., Al-Sawan, S. M., Alnaqeeb, M. A., Khan, I. and Ali, M. 2002. *Prostaglandins, Leukotrienes, and Essential Fatty Acids* 67: 475–478.
60. Altman, R. D. and Marcussen, K. C. 2001. *Arthritis and Rheumatism* 44: 2531–2538.
61. Minghetti, P., Sosa, S., Cilurzo, F., Casiraghi, A., Alberti, E., Tubaro, A., Loggia, R. D. and Montanari, L. 2007. *Planta Medica* 73: 1525–1530.
62. Aimbire, F., Penna, S. C., Rodrigues, M., Rodrigues, K. C., Lopes-Martins, R. A. and Sertié, J. A. 2007. *Prostaglandins, Leukotrienes, and Essential Fatty Acids* 77: 129–138.
63. Levy, A. S., Simon, O., Shelly, J. and Gardener, M. 2006. *BMC Pharmacology* 6: 12.
64. Aktan, F., Henness, S., Tran, V. H., Duke, C. C., Roufogalis, B. D. and Ammit, A. J. 2006. *Planta Medica* 72: 727–734.
65. Lantz, R. C., Chen, G. J., Sarihan, M., Sólyom, A. M., Jolad, S. D. and Timmermann, B. N. 2007. *Phytomedicine* 14: 123–128.
66. Grzanna, R., Lindmark, L. and Frondoza, C. G. 2005. *Journal of Medicinal Food* 8: 125–132.
67. Mahady, G. B., Pendland, S. L., Yun, G. S., Lu, Z. Z. and Stoia, A. 2003. *Anticancer Research* 23: 3699–3702.
68. Akoachere, J. F., Ndip, R. N., Chenwi, E. B., Ndip, L. M., Njock, T. E. and Anong, D. N. 2002. *East African Medical Journal* 79: 588–592.
69. Agarwal, M., Walia, S., Dhingra, S. and Khambay, B. P. 2001. *Pest Management Science* 57: 289–300.
70. Ficker, C. E., Arnason, J. T., Vindas, P. S., Alvarez, L. P., Akpagana, K., Gbeassor, M., De Souza, C. and Smith, M. L. 2003. *Mycoses* 46: 29–37.
71. Betoni, J. E., Mantovani, R. P., Barbosa, L. N., Di Stasi, L. C. and Fernandes Junior, A. 2006. *Memórias do Instituto Oswaldo Cruz* 101: 387–390.
72. Nagoshi, C., Shiota, S., Kuroda, T., Hatano, T., Yoshida, T., Kariyama, R. and Tsuchiya, T. 2006. *Biological and Pharmaceutical Bulletin* 29: 443–447.
73. Gupta, S. and Ravishankar, S. 2005. *Foodborne Pathogens and Disease* 2: 330–340.
74. López. P., Sánchez, C., Batlle, R. and Nerín, C. 2005. *Journal of Agriculture and Food Chemistry* 53: 6939–6946.
75. Norajit, K., Laohakunjit, N. and Kerdchoechuen, O. 2007. *Molecules* 12: 2047–2060.
76. Mahady, G. B., Pendland, S. L., Stoia, A., Hamill, F. A., Fabricant, D., Dietz, B. M. and Chadwick, L. R. 2005. *Phytotherapy Research* 19: 988–991.
77. Wang, H. and Ng, T. B. 2005. *Biochemical and Biophysical Research Communions* 336: 100–104.
78. Samy, R. P. 2005. *Fitoterapia* 76: 697–699.
79. Ansari, M. N., Bhandari, U. and Pillai, K. K. 2006. *Indian Journal of Experimental Biology* 44: 892–897.
80. Natarajan, K. S., Narasimhan, M., Shanmugasundaram, K. R. and Shanmugasundaram, E. R. 2006. *Journal of Ethnopharmacology* 105: 76–83.
81. Ippoushi, K., Ito, H., Horie, H. and Azuma, K. 2005. *Planta Medica* 71: 563–566.

82. Asnani, V. and Verma, R. J. 2007. *Acta Poloniae Pharmaceutica* 64: 35–37.
83. Nurtjahja-Tjendraputra, E., Ammit, A. J., Roufogalis, B. D., Tran, V. H. and Duke, C. C. 2003. *Thrombosis Research* 111: 259–265.
84. Koo, K. L., Ammit, A. J., Tran, V. H., Duke, C. C. and Roufogalis, B. D. 2001. *Thrombosis Research* 103: 387–397.
85. Guh, J. H., Ko, F. N., Jong, T. T. and Teng, C. M. 1995. *Journal of Pharmacy and Pharmacology* 47: 329–332.
86. Young, H. Y., Liao, J. C., Chang, Y. S., Luo, Y. L., Lu, M. C. and Peng, W. H. 2006. *American Journal of Chinese Medicine* 34: 545–551.
87. Jantan, I., Rafi, I. A. and Jalil, J. 2005. *Phytomedicine* 12: 88–92.
88. Riyazi, A., Hensel, A., Bauer, K., Geissler, N., Schaaf, S. and Verspohl, E. J. 2007. *Planta Medica* 73: 355–362.
89. Yoshikawa, M., Hatakeyama, S., Taniguchi, K., Matuda, H. and Yamahara, J. 1992. *Chemical and Pharmaceutical Bulletin Tokyo* 40: 2239–2241.
90. Siddaraju, M. N. and Dharmesh, S. M. 2007. *Molecular Nutrition and Food Research* 51: 324–332.
91. Koch, C., Reichling, J., Schneele, J. and Schnitzler, P. 2008. *Phytomedicine* 15: 71–78.
92. Schnitzler, P., Koch, C. and Reichling, J. 2007. *Antimicrobial Agents and Chemotherapy* 51: 1859–1862.
93. Sookkongwaree, K., Geitmann, M., Roengsumran, S., Petsom, A. and Danielson, U. H. 2006. *Pharmazie* 61: 717–721.
94. Vishwakarma, S. L., Pal, S. C., Kasture, V. S. and Kasture, S. B. 2002. *Phytotherapy Research* 16: 621–626.
95. Ajith, T. A., Hema, U. and Aswathy, M. S. 2007. *Food and Chemical Toxicology* 45: 2267–2272.
96. Yemitan, O. K. and Izegbu, M. C. 2006. *Phytotherapy Research* 20: 997–1002.
97. Nie, H., Meng, L. Z. and Zhang, H. 2006. *Zhongguo Zhong Xi Yi Jie He Za Zhi* 26: 529–532.
98. Bhandari, U., Sharma, J. N. and Zafar, R. 1998. *Journal of Ethnopharmacology* 61: 167–171.
99. Verma, S. K., Singh, M., Jain, P. and Bordia, A. 2004. *Indian Journal of Experimental Biology* 42: 736–738.
100. Akhani, S. P., Vishwakarma, S. L. and Goyal, R. K. 2004. *Journal of Pharmacy and Pharmacology* 56: 101–105.
101. Srinivasan, K. 2005. *International Journal of Food Science and Nutrition* 56: 399–414.
102. Bhandari, U., Kanojia, R. and Pillai, K. K. 2005. *Journal of Ethnopharmacology* 97: 227–230.
103. Ghayur, M. N. and Gilani, A. H. 2005. *Journal of Cardiovascular Pharmacology* 45: 74–80.
104. Chang, C. P., Chang, J. Y., Wang, F. Y. and Chang, J. G. 1995. *Journal of Ethnopharmacology* 48: 13–19.
105. Imanishi, N., Andoh, T., Mantani, N., Sakai, S., Terasawa, K., Shimada, Y., Sato, M., Katada, Y., Ueda, K. and Ochiai, H. 2006. *American Journal of Chinese Medicine* 34: 157–169.
106. Shen, C. L., Hong, K. J. and Kim, S. W. 2005. *Journal of Medicinal Food* 8: 149-153.
107. Kim, D. S., Kim, J. Y. and Han, Y. S. 2007. *Journal of Alternative and Complementary Medicine* 13: 333–340.
108. Prajapati, V., Tripathi, A. K., Aggarwal, K. K. and Khanuja, S. P. 2005. *Bioresource Technology* 96: 1749–1757.
109. Jagetia, G., Baliga, M. and Venkatesh, P. 2004. *Cancer Biotherapy and Radiopharmaceuticals* 19: 422–435.
110. Haksar, A., Sharma, A., Chawla, R., Kumar, R., Arora, R., Singh, S., Prasad, J., Gupta, M., Tripathi, R. P., Arora, M. P., Islam, F. and Sharma, R. K. 2006. *Pharmacology, Biochemistry and Behavior* 84: 179–188.
111. Sharma, A., Haksar, A., Chawla, R., Kumar, R., Arora, R., Singh, S., Prasad, J., Islam, F., Arora, M. P. and Kumar Sharma, R. 2005. *Pharmacology, Biochemistry and Behavior* 81: 864–870.
112. Bisset, N. G. 1994. Herbal drugs and phytopharmaceuticals. CRC Press: Boca Raton, FL.
113. McGuffin, M., Hobbs, C., Upton, R. and Goldberg, A. (Eds.). 1997. *American Herbal Products Association's Botanical Safety Handbook*. CRC Press: Boca Raton, FL.

114. Seetheram, K. A. and Pasricha, J. S. 1987. *Indian Journal of Dermatology, Venereology and Leprology* 53: 325–328.
115. Weidner, M. S. and Sigwart, K. 2001. *Reproductive Toxicology* 15: 75–80.
116. Iwu, M. M. 1993. Handbook of African Medicinal Plants. CRC Press: Boca Raton, FL.
117. Meyer, K., Schwartz, J., Crater, D. and Keyes, B. 1995. *Dermatology Nursing* 7: 242–244.
118. Shalansky, S., Lynd, L., Richardson, K., Ingaszewski, A. and Kerr, C. 2007. *Pharmacotherapy* 27: 1237–1247.
119. Jiang, X., Williams, K. M., Liauw, W. S., Ammit, A. J., Roufogalis, B. D., Duke, C. C., Day, R. O. and McLachlan, A. J. 2005. *British Journal of Clinical Pharmacology* 59: 425–432.
120. Chiang, H. M., Chao, P. D., Hsiu, S. L., Wen, K. C., Tsai, S. Y. and Hou, Y. C. 2006. *American Journal of Chinese Medicine* 34: 845–855.

This page intentionally left blank

Botanical Glossary

abaxial: away from the axis, referring to the surface of an organ that is furthest from the axis in bud

actinomorphic: having radially arranged floral segments which are more or less equal in size and shape

acuminate: tapering gradually to an extended point

acute: terminating in a distinct but not extended point, the converging edges forming an angle of less than 90 degrees

adaxial: towards the axis, referring to the surface of an organ that is closest to the axis in bud

alternate: borne singly and spaced around and along the axis, applied to leaves or other organs on an axis

annual: a plant/tree whose life cycle is only one year

articulated: consisting of segments held together by joints

axillary: situated at the angle between one part of a plant and another part, e.g., a branch and a leaf

basal: arising from or positioned at the base

bipinnate: 2-pinnate; twice pinnately divided

blade: part of the leaf above the sheath or petiole

bract: a leaf-like structure, usually different in form from the normal leaves, associated with the inflorescence

bulbous: shaped like a bulb; having an underground storage organ made up of enlarged and fleshy scales wrapped around each other from which flowers and leaves are produced

calyx: the outermost part of a flower, usually green

capsule: a dry fruit formed from two or more carpels that splits at maturity to release the seeds

carpel: the female reproductive organ of a flower

chartaceous: papery

compound: consisting of two or more anatomically or morphologically equivalent units

cordate: heart-shaped in outline

coriaceous: leathery

corolla: consists of petals or a corolla tube and corolla lobes

crenate: with obtuse or rounded teeth which either point forwards or are perpendicular to the margin

cuneate: obtriangular, i.e., wedge-shaped

cylindrical: tubular- or rod-shaped

cyme: an inflorescence in which each flower, in turn, is formed at the tip of a growing axis, further flowers being formed on branches arising below. *adj*. cymose

decussate: having paired organs with successive pairs at right angles to give four rows

dehiscent: breaking open at maturity to release the contents

dentate: with sharp, spreading, rather coarse teeth standing out from the margin

denticulate: finely dentate

discoid: resembling a disc

drupe: a fleshy or pulpy fruit with the inner portion of the pericarp hard or stony

ellipsoid: elliptic in outline and with a length:breadth ratio between 3:2 and 2:1

elliptic: widest at the middle of the blade, with curved margins and pointed apex and base

elongate: lengthened; stretched out

endocarp: the innermost layer of the fruit wall, derived from the innermost layer of the carpel wall

endospermous: possessing the nutritive tissue of a seed, consisting of carbohydrates, proteins, and lipids

entire: without any incisions or teeth

epiphyte: a plant growing on, but not parasitic on, another plant. *adj*. epiphytic

erect: upright; perpendicular

evergreen: a plant/tree that has leaves all year round

exocarp: the outermost layer of the fruit wall, derived from the outermost layer of the carpel wall. Sometimes called epicarp

exstipulate: leaves without stipules

filament: the stalk of a stamen below the point of attachment to the anther

-foliolate: used with a number prefix to denote the number of leaflets

follicle: a dry fruit, derived from a single carpel and dehiscing along one suture

frond: the leaf of a fern or cycad

fusiform: spindle-shaped, i.e., is circular in cross-section and tapering at both ends

glabrous: without hairs

glaucous: blue-green in colour, with a whitish bloom

globose: spherical or globular; circular in outline

herb: a plant which is non-woody or woody at the base only; some having medicinal properties

herbaceous: herb-like; often applied to bracts, bracteoles or floral parts that are green and soft in texture

hermaphrodite: self-pollinating plants containing both male and female reproductive organs, namely stamens and ovary respectively

imbricate: closely packed and overlapping

indehiscent: not opening or splitting to release the contents at maturity

inflorescence: the arrangement of flowers in relation to the axis and to each other

internode: the part of an axis between two successive nodes, joints or point of attachment of the leaves

introduced: not indigenous; not native to the area in which it now occurs

lanceolate: lance-shaped, much longer than wide, the widest point below the middle

leaflet: one segment of a compound leaf

leathery: possess the feel or texture of leather

linear: long and narrow, with essentially parallel margins

lobe: a usually rounded or pointed projecting part, usually one of two or more, each separated by a gap

margin: the edge of the leaf blade

mesocarp: the middle layer of the fruit wall derived from the middle layer of the carpel wall

mucilaginous: soft, moist, viscous and slimy

native: a plant indigenous to the locality

oblanceolate: reverse lanceolate, widest above the middle of the blade

oblong: rectangular, with nearly parallel margins, about two times as long as wide

obovate: reverse ovate, with the broadest part above

obtuse: blunt or rounded at the apex, the converging edges separated by an angle greater than 90 degrees

opposite: describing leaves or other organs which are at the same level but on opposite sides of the stem

ovate: broader part at the end of the base

ovoid: egg-shaped; ovate in outline

palmate: describing a leaf which is divided into several lobes

panicle: a compound raceme; an indeterminate inflorescence in which the flowers are borne on branches of the main axis or on further branches of these. *adj.* paniculate

paripinnate: pinnate with an even number of leaflets and without a terminal leaflet

pedicel: the stalk of an individual flower

perennial: with a life span extending over more than two growing seasons

perianth: the outer floral whorl or whorls of a monocotyledonous flower

pericarp: the wall of a fruit developed from the ovary wall. Composed of the exocarp, mesocarp and endocarp

petal: free segment of the corolla

petiole: the stalk of a leaf

pinna: a primary segment of the blade of a compound leaf or frond. *pl.* pinnae

pinnate: with the same arrangement as a feather

plumule: the primary bud of an embryo or germinating seed

pod: a dry dehiscent fruit containing many seeds

prostrate: lying flat on the ground

raceme: an unbranched flower spike where the flowers are borne on pedicels

rachis: the axis of a pinna in a bipinnate leaf

receptacle: the region at the end of a pedicel or on an axis which bears one or more flowers

reniform: kidney-shaped in outline

rhizome: a creeping stem, usually below ground, consisting of a series of nodes and internodes with adventitious roots

rhombic: diamond-shaped, widest at the middle and with straight margins

rosette: a tuft of leaves or other organs resembling the arrangement of petals in a rose

serrate: toothed so as to resemble a saw; with regular, asymmetric teeth pointing forward

sessile: without a stalk

shrub: a woody plant usually less than 5 m high and many-branched without a distinct main stem except at ground level

simple: not divided, e.g., applied to a leaf not divided into leaflets

spathe: a large bract ensheathing an inflorescence or its peduncle

spike: an unbranched inflorescence of sessile flowers or spikelets. *adj.* spicate

stamen: one of the male organs of a flower, consisting typically of a stalk (filament) and a pollen-bearing portion (anther)

stellate: star-shaped, usually referring to hairs with radiating branches

stipe: a stalk or support such as the petiole of a frond or the stalk of an ovary or fruit

stipule: one of a pair of leaf-like, scale-like or bristle-like structures inserted at the base or on the petiole of a leaf or phyllode. *adj.* stipulate

stolon: the creeping stem of a rosetted or tufted plant, giving rise to another plant at its tip

strobilus: a cone-like structure formed from sporophylls or sporangiophores. *pl.* strobili

succulent: fleshy, juicy, soft in texture and usually thickened

tendril: a slender organ formed from a modified stem, leaf or leaflet which, by coiling around objects, supports a climbing plant

terminal: at the apex or distal end

tomentose: covered with not very long cottony hairs

trigonous: obtusely 3-angled; triangular in cross-section with plane faces

truncate: with an abruptly transverse end as if cut off

tuber: a stem, usually underground, enlarged as a storage organ and with minute scale-like leaves and buds or “eyes”

umbel: an inflorescence in which the pedicels originate from one point on top of the peduncle and are usually of equal length

undulate: with an edge or edges wavy in a vertical plane

urceolate: urn-shaped

variegate: diverse in colour or marked with irregular patches of different colours

vein: the vascular tissue of the leaf

verrucose: warty

whorl: a ring-like arrangement of similar parts arising from a common point or node

This page intentionally left blank

Medical Glossary

abortifacient: inducing abortion

abscesses: areas with necrotic liquid or pus accumulation

aching: causing physical pain or distress

acne: a disorder of the skin caused by inflammation of the skin glands and hair follicles

acrodynia: a disease of infancy and early childhood marked by pain and swelling in, and pink coloration of, the fingers and toes and by listlessness, irritability, failure to thrive, profuse, perspiration, and sometimes scarlet coloration of the cheeks and tip of the nose

adenitis: inflammation of a gland

AIDS: Acquired Immune Deficiency Syndrome

albuminuria: presence of serum albumin (protein) in the urine

aldosteronism: hyperaldosteronism; an abnormality of electrolyte balance caused by excessive secretion of aldosterone

alexipharmic: medicine that is intended to counter the effects of poison; an antidote to poison or infection

alexiteric: protecting against infection, venom and poison

alopecia: hair loss; baldness; absence of hair from skin areas where it is normally present

amylase: an enzyme that catalyses the hydrolysis of starch into sugars

analgesic: an agent that reduces or eliminates pain

anaphylaxis: hypersensitivity resulting from sensitisation following prior contact with the causative agent

anasarca: generalised massive oedema

angiogenic: an agent that promotes the growth of new blood vessels from pre-existing vessels

anodyne: an agent that relieves pain

anorexia: lack or loss of appetite for food

anthelmintic: an agent that destroys or expels parasitic intestinal worms

anthrax: a serious disease caused by *Bacillus anthracis*, a bacterium that forms spores

antianaphylactic: an agent that prevents anaphylaxis (exaggerated reaction of an organism to which it has previously become sensitised)

antiangiogenic: preventing the growth of new blood vessels

antianxiety: allaying anxiety

antiapoptotic: preventing genetically programmed cell death

antiarrhythmic: preventing or alleviating irregular heart beats

antiatherogenic: an agent that is capable of stopping the formation of fat deposits in blood vessels

antibacterial: an agent that destroys or stops the growth of bacteria

anticancer: reducing the frequency or rate of uncontrolled cell growth

anticarcinogenic: describing an agent that will stop the formation of a cancer

anticholinergic: opposing or blocking the physiological action of a neurotransmitter, acetylcholine, which contract muscles and causing excitatory actions in the central nervous system

anticoagulant: acting to prevent clotting of blood

anticonvulsant: an agent that prevents or relieves seizures or epileptic attack

antidepressant: an agent that stimulates the mood of a patient

antidiabetic: an agent that controls high blood sugar

antidiarrhoeal: substances used to prevent or treat diarrhoea

antidotal: counteracting the effects of a poison

antiemetic: an agent that prevents or alleviates nausea and vomiting

antifertility: agent that is capable of reducing or adversely affecting fertility

antifilarial: an agent that counters nematode (roundworm) of the superfamily Filarioidea

antifungal: an agent that destroys or inhibits the growth of fungi

antigenotoxic: an agent that is capable of preventing damage to DNA molecules in genes, causing mutations, tumours, etc.

antihyaluronidase: directed or effective against an enzyme (hyaluronidase) that breaks down hyaluronic acid, thus facilitating the spread of fluid through tissues

antihypertensive: an agent that controls high blood pressure

anti-inflammatory: an agent that counteracts inflammation

antileishmanial: directed or effective against leishmania, a parasitic protozoan that causes leishmaniasis

antimalarial: an agent that prevents or treat the disease malaria caused by the Plasmodium parasite

antimicrobial: destroying or inhibiting the growth of microorganisms

antimitotic: pertaining to certain substances capable of stopping cell division (mitosis)

antimutagenic: directed or effective against substances that tends to increase the frequency or extent of mutation

antineoplastic: inhibiting or preventing the growth and spread of neoplasms or cancerous cells

antinociceptive: relieving pain

antiosteoporotic: directed or effective against osteoporosis (characterised by decrease in bone mass with increased porosity and brittleness)

antioxidant: inhibiting oxidation or an agent that does so

antioxidative: protect body cells from the damaging effects of oxidation

antiplasmodial: directed or effective against the malarial parasites (Plasmodium species)

antiplatelet: inhibiting or stopping the formation of platelet aggregation

antiproliferative: inhibiting cell growth

antiprotozoal: tending to destroy or inhibit the growth of protozoans

antipruritic: preventing or relieving itching

antipsychotic: any of the powerful tranquilisers used especially to treat psychosis and believed to act by blocking dopamine nervous receptors

antipyretic: reducing fever or an agent that reduces fever

antirheumatic: relieving or preventing rheumatism

antischistosomal: directed or effective against infection caused by trematode worms (schistosomes) which are parasitic in the blood of humans and other mammals

antiseptic: any substance that inhibits the growth and reproduction of microorganisms

antispasmodic: an agent that prevents or relieves muscle spasms

antithyroidal: opposing thyroid function

antitrichomonal: directed or effective against some protozoa (Trichomonas)

antituberculosis: effective in the treatment of tuberculosis, a condition due to *Mycobacterium tuberculosis*

antitumour: inhibiting the growth of tumour cells

antitussive: effective against cough; an agent that suppresses coughing

antiulcer: prevent or cure ulcers or irritation of the gastrointestinal tract

antiulcerogenic: directed or effective against development of ulcers

antivenom: an agent used in the treatment of poisoning by animal toxins

antiviral: acting, effective, or directed against viruses

anuria: lack of urine production

anxiolytic: relieving anxiety or an agent that relieves anxiety

aperient: a mild laxative or gentle purgative

aphrodisiac: an agent that stimulates sexual desire

apnoea: a condition whereby breathing is stopped

apoptotic: a term used to describe the state of programmed cell death

appendicitis: inflammation of the appendix

arrhythmia: abnormal heart beat

arthritis: inflammation of joints

ascites: effusion and accumulation of fluid in the abdominal cavity

asthma: a chronic, inflammatory lung disease characterised by recurrent breathing problems usually triggered by allergens

astringent: having the property of causing contraction of soft organic tissues for the control of bleeding or secretions

atherosclerosis: thickening, hardening and loss of elasticity of the walls of arteries

athlete's foot: fungal infections on the feet, often smelly

atrophy: a wasting away; reduction in the size of a cell, tissue or organ

bacillary: pertaining to the bacteria bacilli or to rod-like structures

bechic: pertaining to cough

beriberi: a disease due to vitamin B1 deficiency, marked by inflammation of nerves, heart problems and oedema

bilious: relating to or containing bile

blennorrhoea: any free discharge of mucus, especially from urethra or vagina

blisters: a vesicle

boils: furuncle; tender, swollen areas of infection that usually form around hair follicles

bronchial: pertaining to or affecting one or both windpipe

bronchiolitis obliterans: a disease of the lungs where the bronchioles are inflamed and plugged with granulation tissue

bronchitis: inflammation of mucous membrane of the windpipe (bronchus)

cachexia: general ill health and malnutrition

calculus: abnormal hard deposit, usually composed of mineral salts, occurring within the body

cancer: a malignant tumour of growth that expands locally by invasion and systemically by metastasis

carbuncles: swollen lumps or masses under the skin due to skin infections that often involve a group of hair follicles

cardiac: pertaining to the heart

cardioprotective: protecting the heart function

cardiotonic: an agent that increases tonicity of the heart muscles

carminative: an agent that relieves and removes gas from the digestive system

cataracts: conditions in which there is a loss or reduction of transparency of the lens of the eye causing progressive loss of clarity and detail of images

catarrh: a condition of the mucous membranes characterised by inflammation and mucous

cathartic: an active purgative, producing bowel movements and evacuation of the bowels

cephalgia: headache

chemopreventive: a chemical agent that can help prevent the development of cancer

chilblains: inflammation or blister of the hands and feet caused by exposure to cold or moisture

cholera: a form of intestinal infection that results in frequent watery stools, cramping abdominal pain and eventual collapse from dehydration

cicatrizant: a medicine or application that promotes the healing of a sore or wound, or the formation of a cicatrix, which is a scar left by the formation of new connective tissue over a healing sore or wound

coagulant: an agent promoting clotting of blood

cold sore: a recurrent, small blister in and around the mouth caused by the herpes simplex virus

colic: acute paraxysmal abdominal pain or pertaining to the colon

coma: a state of profound unconsciousness from which the patient cannot be aroused, even by powerful stimuli

condiment: something used to give a special flavour to food, as mustard, ketchup, salt, or spices

congestive heart failure: the heart fails to pump efficiently, resulting in swelling, shortness of breath, weakness, etc.

conjunctivitis: inflammation of the conjunctiva, generally associated with discharge

constipation: infrequent or difficult evacuation of faeces

contraceptive: an agent capable of preventing the formation of a foetus

contusion: injury to a part without a break in the skin; a bruise

convulsions: spasms, epilepsy

coolant: an agent that produces cooling

coryza: profuse discharge from the mucous membrane of the nose

counter-irritant: an agent producing counter-irritation, so that less pain at a particular site is experienced

cramps: a painful spasmodic muscular contraction

Crohn's disease: a chronic inflammatory disease of the digestive tract, particularly the small intestine and colon

cystitis: inflammation of the urinary bladder

cytotoxic: toxic to cells, preventing their reproduction or growth

dehydration: depletion of body fluids

delirium: mental state characterised by reduced ability to maintain attention to external stimuli and disorganised thinking

demulcent: an agent that soothes and protects the part to which it is applied

deobstruent: removes obstructions by ducts or pores of the body

depressant: an agent that diminishes any functional activity

depurative: purifying the blood or the humours

dermatitis: inflammation of the skin

dermatosis: any skin disease, especially one not characterised by inflammation

diabetes: metabolic disorder resulting in high blood sugar and discharge of large amounts of sugar in the urine.

diaphoretic: an agent that increases perspiration

diarrhoea: abnormally frequent discharge of watery stools

digitalis: the dried leaf of *Digitalis purpurea*; used as a tonic for the heart

dipsia: thirst

discutient: an agent that causes the dispersal or causes something, such as a tumour or any pathologic accumulation, to disappear

diuretic: an agent that increases the production of urine

dropsy: an abnormal accumulation of serous fluid in cellular tissues or in a body cavity

dysentery: infection of the gut caused by a bacterium called shigella, characterised by abdominal pain, diarrhoea with passage of mucus or blood

dysmenorrhoea: difficult and painful menstruation

dyspepsia: indigestion

dyspnoea: difficult or laboured breathing

dysrhythmias: an abnormality in an otherwise normal rhythmic pattern

dysuria: painful or difficult urination

ecchymoses: small haemorrhagic spots in the skin or mucous membrane, forming a non-elevated, rounded or irregular, blue or purplish patch

eczema: an inflammatory condition of the skin characterised by redness, itching, scales, crusts or scabs alone or in combination

elephantiasis: a disease caused by tumours in lymph node or filarial worms, causing excessive swelling in limbs and genitalis

embrocations: alcohol-based treatments rubbed into the skin to relieve pain or which produce reddening of the skin

embryotoxicity: any toxicity that affects an embryo

emetic: an agent that induces vomiting

emmenagogue: a substance that promotes or assists the flow of menstrual fluid

emollient: an agent that will soften, soothe and protect the part when applied locally

enteritis: inflammation of the intestine, especially the small intestine

epigastric: pertaining to the upper and middle region of the abdomen

epilepsy: convulsions

epistaxis: nosebleed

eruptions: lesions on the skin that are usually red, raised, and easily visible

erysipelas: is an acute streptococcus bacterial infection marked by deep red inflammation of the skin and mucous membranes. This disease is also known as Saint Anthony's fire

erythema: redness of the skin due to congestion of the capillaries

excrescences: abnormal outgrowth; a projection of morbid origin

expectorant: an agent that promotes the discharge or expulsion of mucus from the respiratory tract

febrifuge: an agent that reduces fever

fertility: the capacity to conceive or induce conception

fibroids: non-cancerous tumours made of muscle cells and tissues that grow in and around the wall of the uterus or womb

fibrosis: an abnormal formation of fibrous tissue as a reparative or reactive process, as opposed to the formation of fibrous tissue that is a normal constituent of an organ or tissue

filaricidal: an agent that kills filariae, nematodes that as adults are parasites in the blood or tissues of mammals and as larvae usually develop in biting insects

fistulae: abnormal connections between an organ, vessel or intestine and another structure, usually due to injury, surgery or result from infection or inflammation

flatulence: the presence of excessive gas in the intestinal tract, causing discomfort

fluxes: excessive flow or discharge

furuncle: infection of a hair follicle

furunculosis: the persistent sequential occurrence of furuncles over a period of weeks or months or the simultaneous occurrence of a number of furuncles

galactophoritis: inflammation of the milk ducts

galactorrhoea: excessive or spontaneous milk flow; persistent secretion of milk irrespective of nursing

gangrenous: characterised by the decay of body tissues, which become black and smelly

gastritis: inflammation of the stomach

gastroenteritis: inflammation of the lining membrane of the stomach and the intestinal tract

gastroprotective: protect the stomach

gavage: forced feeding, especially through a tube passed into the stomach

genotoxic: a toxic agent that damages DNA molecules in genes, causing mutations, tumours, etc.

glottal: pertaining to the vocal apparatus of the larynx, consisting of the true vocal cords and the opening between them

glucosidase: an enzyme of the hydrolase class that breaks down a glucoside

gonorrhoea: infectious sexual disease due to *Neisseria gonorrhoeae*

gout: a metabolic disease that is a form of acute arthritis, marked by inflammation of the joints and great pain

gravel: calculi occurring in small particles

haemaglutination: agglutination of red blood cells

haematemesis: the vomiting of blood

haematometra: an accumulation of blood in the uterus

haematuria: the finding of blood in the urine

haemiplegia: paralysis of one side of the body

haemolytic: lysis of red blood cells liberating haemoglobin in the plasma

haemoptysis: coughing up of blood from the airways

haemorrhage: profuse bleeding from the blood vessels

haemorrhoids: piles

haemostatic: stop bleeding

heartburn: painful, burning feeling in the chest caused by stomach acid flowing back into the esophagus

hepatitis: inflammation of the liver, usually due to viruses or toxins

hepatoprotective: protecting the liver functions

hepatotoxic: an agent that is toxic to liver cells

herpes: inflammation of the skin or mucuous membrane with clusters of deep seated vesicles; a family of viruses that infect humans: herpes simplex causes lip and genital sores; herpes zoster causes shingles

hiccups: sharp inspiratory sound with spasm of the glottis and diaphragm

HIV: Human Immunodeficiency Virus, the virus that causes AIDS

hydragogue: producing watery discharge, especially from the bowels or a cathartic that causes watery purgation

hydrocele: swelling of and fluid on the testicles

hydrothorax: a collection of serous fluid within the pleural cavity

hypercholesterolaemic: high level of cholesterol in the blood

hyperemesis: excessive vomiting

hyperkalemia: high concentration of potassium in the blood

hypertension: high blood pressure

hypocholesterolaemic: an abnormal deficiency of cholesterol in the blood

hypoglycaemic: producing a decrease in the blood sugar level

hypolipidaemic: producing or resulting from a decrease in the level of lipids in the blood

hypotension: low blood pressure

hypotensive: marked by low blood pressure

immunoglobulin: a protein of animal origin with known antibody activity

immunomodulatory: having the ability to modify the immune response or function of the immune system

immunostimulatory: the ability to stimulate the immune response

immunosuppressive: being able to reduce the immune response

impetigo: a highly contagious skin infection caused by bacteria, usually occurring around the nose and mouth; commonly occurring in children

infertility: diminution or absence of ability to produce offspring

inflammation: localised protective reaction of tissue to irritation, injury, infection, chemicals, electricity, heat, cold or microorganisms. Characterised by pain, redness, swelling and possible loss of function

influenza: the flu; an infectious viral respiratory disease characterised by chills, fever, prostration, headache, muscle aches, sore throat, and a dry cough

inotropic: affecting the force or energy of muscular contractions

insanity: a legal term for mental illness, roughly equivalent to psychosis and implying inability to be responsible for one’s acts

insomnia: inability to sleep

intoxication: poisoning

jaundice: yellowish discoloration of the whites of the eyes, skin and mucous membranes caused by deposition of bile salts in these tissues

keratitis: inflammation of the cornea

lactogogue: a substance which stimulates the flow of milk

lameness: a state of being incapable of normal locomotion

larvicidal: larvae killing (especially of parasites)

laryngitis: inflammation of the lining of the larynx causing hoarseness

laxative: having a tendency to loosen or relax, specifically in relieving constipation or an agent that relieves constipation

leishmaniasis: infection with Leishmania, a parasitic protozoa

leprosy: a chronic mycobacterial disease caused *Mycobacterium leprae*, characterised by skin lesions and necrosis

lethargy: a condition of indifference or drowsiness

leucoderma: partial or total loss of skin pigmentation, often occurring in patches. Also called vitiligo.

leucodermic: an agent that can cause partial or total loss of skin pigmentation, often occurring in patches. Also called vitiligo

leucorrhoea: a gynecologic disorder resulting in abnormal, thick, whitish, non-bloody discharge from the genital tract (uterus of vagina)

leukemia: a progressive, maglinant disease of the blood-forming organs, marked by distorted proliferation and development of leukocytes and their precursors in the blood and bone marrow

lipolytic: an agent that breaks down fats

lumbago: pain in the lumbar or loin region

lymphadenopathy: enlargement of lymph nodes usually associated with inflammation or infection

malaria: a febrile disease caused by infection with Plasmodial parasites, causing periodic attacks of chills, fever and sweating

mange: a skin disease of domestic animals or pets due to mites

mastitis: inflammation of the breast

measles: acute highly infectious viral human disease caused by a virus, specifically a paramyxovirus of the genus Morbillivirus, characterised by cough, fever, small red lesions each with whitish centre, formed in the mouth in early stages of the measles

melancholic: a depressed and unhappy emotional state with abnormal inhibition of mental and bodily activity

menorrhagia: excessive or prolonged menstruation

metritis: inflammation of the uterus

metrorrhagia: uterine bleeding; usually abnormal amount, occurring at completely irregular intervals, the period of flow sometimes being prolonged

molluscicidal: effective for destroying molluscs

mumps: infectious disease caused by a paramyxovirus, marked by fever and inflammation of parotid gland and swelling of the neck and throat

mydriasis: pupil dilation

myocardial: referring to the middle and the thickest layer of the heart wall, composed of cardiac muscles

narcotic: a drug or agent which in moderate doses depresses the central nervous system, relieving pain and producing sleep but in large doses, produces unconsciousness, stupor, coma and possibly death

nausea: an unpleasant sensation vaguely referred to the epigastrium and abdomen, with a tendency to vomit

necrosis: the morphological changes indicative of cell death

neonatal: pertaining to the first four weeks after birth

nephritis: inflammation of the kidneys

nephropathy: disease of the kidneys

nephroprotective: an agent that protects or prevents damage to the kidney cells

nephrotoxic: toxic to kidney cells

neuralgia: pain occurring in the area served by a sensory nerve

neurasthenia: nervous debility dependent upon impairment in the functions of the spinal cord

neurodermatitis: a general term for a dermatosis presumed to be caused by itching due to emotional causes

neuroleptics: effects on cognition and behavior of antipsychotic drugs, which produce a state of apathy, lack of initiative and limited range of emotion

neuroprotective: an agent that protects or prevents damage to the nerves

neurotoxic: toxic to the nerves or nervous tissue

neurotoxicity: the quality of exerting a destructive or poisonous effect on nerves

nits: the eggs or youngs of a parasitic insect, such as a lice

nodal rhythm: a type of heart rhythm disorder; the cardiac rhythm that results when the heart is controlled by the atrioventricular node in which the impulse arises in the atrioventricular node, ascends to the atria, and descends to the ventricles more or less simultaneously

non-teratogenic: referring to substance which will not cause malformations of an embryo or fetus

oedema: the presence of abnormally large amounts of fluid in the intercellular tissue spaces of the body

oesophagitis: inflammation of the esophagus

oliguria: diminished urine secretion in relation to fluid intake

ophthalmia: severe inflammation of the eye

osteoclasis: surgical fracture or refracture of bones

osteodynia: ostealgia; pain in the bones

otitis: inflammation of the ear

oxytocic: an agent that promotes rapid labour by stimulating contractions of the uterus

oxyuriasis: infection with *Enterobius vermicularis* (in humans) or with other worms

pacemaker: a device that regulates the rhythm of the heart beat

palsy: paralysis

panacea: a remedy for all diseases

parasiticidal: destructive to the parasites

parturition: the act or process of giving birth to a child

periodontitis: inflammation of the tissues surrounding and supporting the teeth

pertussis: whooping cough

pharyngitis: inflammation of the pharynx (throat)

photoprotective: protecting against harmful effects of lights (UV-irradiation)

phototoxicity: toxic effect triggered by exposure to light

piles: haemorrhoids

piscicidal: a substance that kills fish

pleurisy: inflammation of the serous membrane investing the lungs and lining the walls of the thoracic cavity

pneumonia: inflammation of the lungs due to a bacterial or viral infection, which causes fever, shortness of breath, and the coughing up of phlegm

pollakiuria: abnormally frequent urination

postpartum: occurring after childbirth, with reference to the mother

prolapse: the falling down, or downward displacement, of a part or viscus

prophylactic: an agent that prevents or protects against a disease or condition

prurigo: any of several itchy skin eruptions in which the characteristic lesion is dome-shaped with a small transient vesicle on top, followed by crusting or lichenification

psoriasis: a chronic skin disease characterised by inflammation of the skin and formation of red patches

puerperal: pertaining to a woman who has just given birth to a child or to the period or state of confinement after childbirth

purgative: an agent that will cause evacuation of the intestinal contents

pustule: a small, circumscribed elevation of the skin containing pus

radioprotective: serving to protect or aiding in protecting against the harmful effects of radiation

rectocele: hernial protrusion of part of the rectum into the vagina

refrigerant: an agent that produces coolness or reduces fever, allays thirst and gives a sensation of coolness to the system.

resolvent: promoting resolution or the dissipation of a pathologic growth

retching: strong involuntary effort to vomit

Reye's syndrome: a rare disorder in children and teenagers while recovering from childhood infections, such as chicken pox, flu, and other viral infections. Reye's syndrome include nausea, severe vomiting, fever, lethargy, stupor, restlessness, and possibly delirium. Also caused by taking aspirin in children less than 16 years old

rheumatism: general term for acute and chronic conditions characterised by inflammation (arthritis, tendonitis and bursitis), soreness and stiffness of muscles, and pain in joints

rhinitis: inflammation of the nasal mucous membrane

ringworm: any of a number of contagious skin diseases caused by certain parasitic fungi and characterised by the formation of ring-shaped eruptive patches

rubefacient: an agent that produces a mild irritation, reddening of the skin, and local vasodilation increasing the blood supply to the area of application.

saluretics: agents that promotes urinary excretion of sodium and chloride ions

scabies: a contagious skin disease characterised by itching, inflammation, hair loss and secondary bacterial infection.

schistosomiasis: infection with Schistosoma (blood flukes) which causes infection in man by penetrating the skin of people coming in contact with infected waters

sclerosis: an induration or hardening, especially from inflammation and in diseases of the interstitial tissues

scrofula: primary tuberculosis of the cervical lymph nodes; the inflamed structures being subject to a cheesy degeneration

scurvy: disease characterised by spongy bleeding gums, loosening of teeth, and mucous membranes that is caused by lack of Vitamin C

sedative: tending to calm, moderate, or transquilise nervousness or excitement or an agent that does so

senility: the physical and mental deterioration associated with old age

septicaemia: blood poisoning

sialogogue: an agent that stimulates the flow of saliva

sinus bradycardia: a normal but slow heart rhythm

sinusitis: inflammation of a sinus

smallpox: an acute infectious disease due to poxvirus, marked by sustained high fever and the appearance of skin eruptions and pustules, leaving small, depressed, depigmented scars

sores: any lesion of the skin or mucous membranes

spasm: a sudden, violent, involuntary muscular contraction

spastic paraparesis: partial paralysis of the lower extremities due to muscular stiffness and spasms

spermatogenesis: formation and development of sperms

spermatorrhoea: involuntary escape of semen, without orgasm

spermatotoxic: an agent that is toxic to spermatozoa

splenitis: inflammation of the spleen

splenosis: Implantation and subsequent growth of splenic tissue within the abdomen as a result of disruption of the spleen

sprue: a chronic form of malabsorption syndrome

stammering: a disorder of speech behavious marked by involuntary pauses in speech

stasis: stoppage of flow, as of blood or other body fluid

stimulant: an agent that excites the functional activity of an organ or system

stomachic: an agent that stimulates the appetite and gastric secretion

strabismus: squint

sty: a circumscribed abscess caused by bacterial infection of the glands on the edge of the eyelid

styptic: astrigent or arresting haemorrhage by means of an astringent quality

sudorific: an agent that promotes sweating; diaphoretic

sunstroke: a condition caused by excessive exposure to the sun, marked by high skin temperature, convulsions and coma

suppository: an easily fusible medicated mass to be introduced into a body orifice

syphilis: a contagious sexual disease caused by *Treponema pallidum,* characterised by local formation of ulcerous skin eruptions and systemic infection

tachycardia: increased heart beat

tachypnea: very rapid respiration

taeniasis: infection with tapeworms

taenifuge: an agent that expels tapeworms

teratogenic: an agent that that causes physical defects in the developing embryo

tetanus: an acute, sometimes fatal, disease of the central nervous system; caused by the toxin of the tetanus bacterium

thrush: candidiasis of the oral mucous membranes, fungal infection by *Candida albicans* with formation of whitish spots

tonic: producing and restoring normal tone or characterised by continuous tension

tonsillitis: inflammation of the tonsils

trachoma: a contagious disease of the conjunctiva and cornea, producing photophobia, pain, and lacrimation, caused by *Chlamydia trachomatis*

tuberculosis: any of the infectious diseases of man and other animals due to species of Mycobacterium and marked by formation of tubercles and caseous necrosis in tissues of any organs, usually the lung

tumour: an abnormal growth of tissue resulting from uncontrolled multiplication of cells and serving no physiological function

tympanites: abnormal distention due to the presence of gas or air in the intestine or the peritoneal cavity

typhoid: acute infectious disease caused by *Salmonella typhii* and characterised by fever, severe physical and mental depression, diarrhoea and headache

tyrosinase: an oxidising enzyme, occurring in plant and animal tissues, that catalyses the aerobic oxidation of tyrosine into melanin and other pigments

ulcerative colitis: a chronic inflammatory disease of the colon or colon and rectum, characterised by abdominal pain and diarrhoea, often mixed with blood and mucus

urinary calcification: the deposit of calcium salts in the tissues of the urinary tract

urinary lithiasis: the formation of urinary stones

urodynia: pain in urination

urticaria: an allergic disorder marked by raised patches of skin causing itchy swellings of the skin

uteral: relating to uterus

uterine haemorrhage: bleeding in the uterus

uterotonic: increasing the tone of uterine muscle

vaginismus: painful spasm of the vagina, hence often resulting in unsuccessful penetration

varices: enlarged tortuous vein, artery or lymphatic vessel. *adj.* varicose

vasodilation: dilation of blood vessels

vasorelaxant: an agent that relaxes the blood vessels

venereal: due to or propagated by sexual intercourse

vermifuge: an agent that causes the expulsion of intestinal worms

verruca: a wart or one of the wart-like elevations on the endocardium in various types of endocarditis

virility: the condition or quality of being virile

vitiligo: a condition in which destruction of melanocytes in small or large circumscribed areas results in patches of depigmentation often having a hyperpigmented border, and often enlarging slowly

vulnerary: a remedy useful in healing wounds

wandering pacemaker: an atrial arrhythmia that occurs when the natural cardiac pacemaker site shifts between the sino-atrial node, the atria and/or the atrioventricular node

warts: abnormal growths on the skin due to viral infection caused by human papillomavirus

whitlow: an infection of the fingers caused by herpes simplex virus

whooping cough: a bacterial infection especially of the children caused by bacterium *Bordetella pertussis* marked by convulsive spasmodic coughs

Navigation Down Under

For
Northern-Hemisphere Tourists
Walkers and Scouts
Boating and Outdoor Orientation

1001 Ways to Extend

Your Navigational Know-How

2010

CONTENTS

Section page
1 BIG-PICTURE TOP-DOWN————————3
Global Perspective
Sun
Stars
Earth
Moon
Wind
Clockwise Navigation
Anticlockwise Systems
Revision, Overview, Explanations
2 THE INSIDE STORY—Mental Manoeuvres—34
Learn
Think Interpret
Look
Integrate
Orientation
Disorientation
Reorientation
Lost?
Miscellaneous: Angle; Night-time
3 REALLY INTERESTING STUFF
—Bush Navigation————————————56
Nature's Norths
Memory & Naming Skills
Lie of the Land
Tracks and Bends
4 DISTANCE————————————————90
Progress
Pace
Range
MAP and/or COMPASS
Map Use
Magnetic Compass Use
Mapmaking
5 Navigational Mnemonic————————110
6 Appendices———————————————116
Tide and Moon Dial
Star Charts; Zodiac
Morse Code & Rhythm Mnemonics
Semaphore Code & Mnemonics
Symbols for Track-Notes & Mud-maps
A List of Star Coordinates
Great Circles; Calendars in the Sky
Tropical Countries
7 Contents/Index—————————————128

Big-Picture Rule One: Locate a Distant Reference Direction First

Until you Orient yourself to the bigger picture, by finding an unchanging direction to latch on to,
- •Small-scale pictures and large-scale mental layouts cannot come together. •And the local features won't fit into the bigger perspective.
- •Your mental photographs will float around in chaos, like so many loose jigsaw pieces, or letters in an alphabet soup.

In choosing a Reference Direction, your first choice should be...
- •ASTRONOMICAL — Sun, Moon, stars, planets, satellites—e.g. Locate the direction of the North Pole Star. There is an astronomical overview p32f
- Then •GLOBAL — Compass or Botanical or Geomorphological North e.g. find Canada's North Magnetic Pole.
- Next •CONTINENTAL — Weather Systems, radio transmitters, powerline networks, highways, sea-coast—e.g. notice the Sydney to Hobart flight path.
- Then •DISTANT landforms — Lie, Lay, Lines, Hills; Lights, etc.—e.g. the direction of the main river valley draining to the sea.

Last of all comes: •What you can see in front of your nose! I.e. don't let your mind latch on to some unreliable or meaningless reference like how the car is parked, or where the wind is, or the initial direction of the track/the drive in/the main street, or uphill, or the way you got out of bed…
- •If you can't see a thing beyond a few metres, and have no compass, you will need to "keep a straight line" (see p. 87b) by lining up markers.

Now that you have a reference direction, you can implement the other "Rule One"s...
- •Match what you see in front of your nose to the reference;
- •Predict your overall trend before you start; •Notice your start-off direction as you start.

Visualisation is your Key Skill

This is most especially relevant to orienting your nose, in your mind's eye, to the Bigger or Invisible Pictures, but is needed at all scales.

Imagine	where the Southern Cross is during the day;	or	where the cars are, now; how you got here.
Visualise	the Earth's invisible spin-axis	and	the weather map you saw, yesterday
Mentally picture	the Galactic spiral! — which clock-sense is it?	or	the turnoff you need — will it be uphill?
Remember	the tropical countries under the Sun	and	the way back
Describe	the path of a shadow-tip in mid-Winter	or	what makes this place recognisable
Picture	sunrise in Antarctica tomorrow morning	or	old blazes, in your mind, so you don't miss any
Interpret numbers, labels and words into reality	e.g. what the shape of the night-sky has to do with your global position		
Express visually	what each clue means	e.g.	what "high noon" means to you right now
Mentally photograph	the map as you study it beforehand	and	how the climb opened out onto the plateau
Mentally match	the invisible landscape trends to the reference direction and to what you see in front of your nose		

At night, and in the rain, navigation gets to be 'all in the mind'.

By recreating the reality you can't see, you can navigate what is in front of your toes, by intelligent imagination.

Big-Picture Orientation is a Mental Layout Acquired Beforehand

Because the invisible picture is so large, you can do your global and continental orientation at home before you start.
Travelling all day to go walking somewhere else, will hardly alter it.

You use: a globe; an atlas; a roadmap of the continent — to learn the lay beforehand. These are layout maps... The relative positions shown are fixed.
But layout views — as seen from above — will hardly ever be seen in your reality, first-hand.
(In contrast, when you come to putting yourself in the middle of the big-picture, you see the landscape first-hand, and you will need "Nose Navigation" — as seen from the ground — which is all about the apparent relative directions.)
The global constants won't change as you walk, e.g. the behaviour of the Sun and stars, because everything is so distant.

You need extra care to visualise the Big Picture, and its meaning, because it is (a) Round (global) (b) Layout (i.e. mental only) (c) Invisible (d) Big!
Do this before you attempt to nose your way through the local scenery.

Smaller-scale clues can then 'fit-in', sensibly, to the larger context.

Compass Navigation Robs You of a Global Perspective

Flat maps must distort the true picture. Only a globe is true. Mercator-projection maps (diagram 1) distort land areas and longitude lines, in order to "preserve the straightness" (!) of a compass course, which never was straight!
If you glue your nose to a compass needle, to follow a compass-course, the course which looks straight on a flat Mercator map will actually take you on a spiral, into either the South or North Pole (as when you keep any landmark at a diagonal).
The sense of the spiral (turning to your right/left) changes as you cross the Equator...
Because you do need a big-picture approach, **you need a global perspective**.
(Here's a good place to set out the basics...

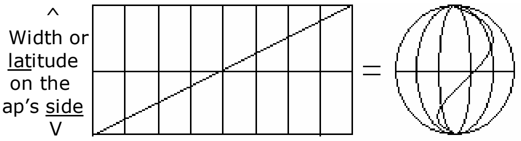

Remember: "Latitude **equates** to the distance from the Equator" — the short dimension.
"Longitude turns upon the time — how **long** has the globe turned since midnight")

Here's a test: Imagine the midsummer sunset from well Down Under, looking even further South. Do you think of a tropical country in that direction?
The following section — about Sun Moon Earth and Stars — aims to restore a true perspective to people who can hardly think except with a flat diagram, and can hardly visualise anything without divorcing it from the real landscape in front of their nose. (There is always a tropical country under the Sun.)
Think of this 'global orientation' section as a global **Re**-Orientation and don't forget, that... **Global means Round!**

The Big Picture is Invisible — Your Job is to Visualise It!

As a matter of principle, we should work top-down
i.e. from a framework down to the detail, from general to specific, large to small, overall to local, unchanging to changeable.
As a matter of fact, we are lazy. The point of doing "useless" grand-scale orientation exercises is to:
- make global orientation second nature — easier next time around;
- to make you forget flat pictures of a round Earth and to imagine it properly;
- to give your struggling memory some comprehensive and relable framework to work within;
- to integrate what you can't see with what you can, e.g. to link different walks and areas;
- to orient to, and to understand, the Sun's movement, and to orient to the Moon and stars and planets as a bonus;
- to make extra clues and techniques available, e.g. to open up astro-navigation techniques;
- to work "top-down" — to avoid jumbling up your mind.

The idea is to read the following double-pages horizontally, as one topic, i.e. Sun-Sun; Stars-Stars; Clockwise-Anticlockwise.
To follow one topic through, flip over the pages and look in the corresponding positions.

Global Re-orientation — Light-Half/Dark-Half

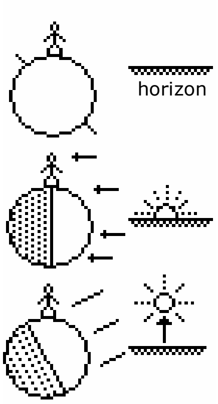

Step 1. **Always Position Yourself on Top of the World.** ..whether in a diagram or mentally or in reality.
This makes your reorientation realistic, doesn’t it?
Look at your feet and imagine you are standing on a huge basketball in the sunlight.
To visualise the light-half/dark-half of the globe, carry a real ball with you, preferably a globe of the world.
Imagine that everything which you see, this side of the horizon, is the tiny top-fraction of a huge ball.

Step 2. Face the Sun, but imagine it low down, setting, with horizontal rays coming to you.
Notice the circle which divides day from night, passing to left and to right underneath your feet.
“He divided light from darkness by a circle drawn on the face of the sea.” Job 26:10

Step 3. ‘Raise’ the Sun, and its sunbeams, to their actual height.
And as you do, let the dark-half fall back behind you by the same amount
i.e. degree for degree around a round globe—one degree of fall-back for one degree of Sun-rising.

After this you can do your best to answer: (Help will follow soon) What countries are seeing the Sun as horizontal?
i.e. Which countries does the ‘circle’ pass through at the moment? Which ones are seeing the Sun rise?
Which ones see it setting? Which parts see a 24-hour sunset, with the Sun skimming the horizon? (Use a globe)

Your Best Reference Direction: Imagine the Sun at High Noon — Always

(This won’t work so well in the tropics)

Night or day, your mental reference direction is North or South by the Sun. So imagine the Sun North of you (Southern Hemisphere) rain or shine.
It is the natural interpretation… of looking at your watch, and
of looking at the Sun, when it is not North of you. (Warning: Don’t look directly at the Sun! It will blind you!)
of looking at the compass. (You don’t want to locate ‘Canada’s magnetic pole’, after all!)
of judging botanical North/South (p.74a);
of using a map in conjunction with the landscape,
and even when using the Moon

Imagine the *high-noon* Sun in a real landscape — this one here!
— at all times of day or night.
— at the correct height, on its circular path,
— which is tilted at its rise-set angle.
Warning:
You will get disoriented
when you cross the Equator!

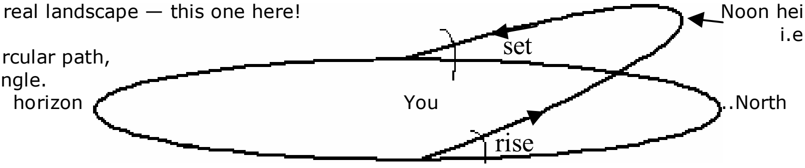

Noon height (Southern Hemisphere)
i.e. your best guide to “North”

The 4 Bright Outer Stars of Orion Point out Where the Pot is Going

=Find N/S E/W & Equator

Hold this chart over your head

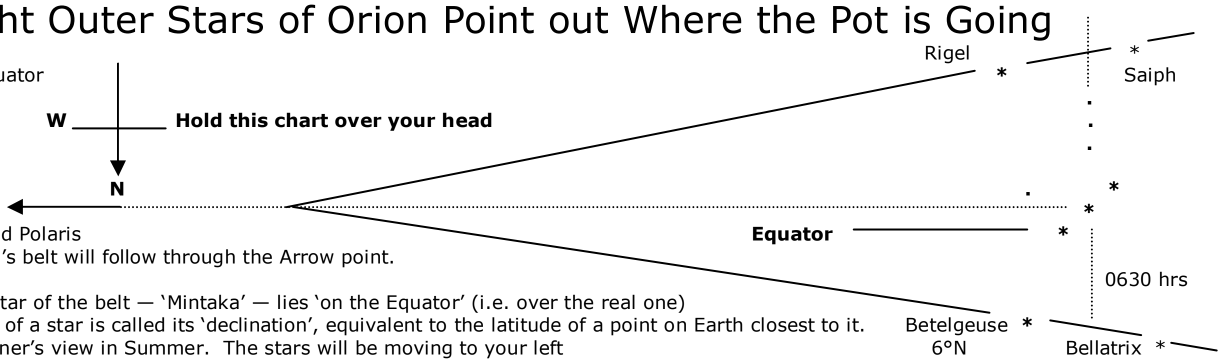

Movement West
Anticlockwise around Polaris
The middle of Orion’s belt will follow through the Arrow point.

The north-west star of the belt — ‘Mintaka’ — lies ‘on the Equator’ (i.e. over the real one)
The N-S position of a star is called its ‘declination’, equivalent to the latitude of a point on Earth closest to it.
This is a Southerner’s view in Summer. The stars will be moving to your left
The Pot handle down through Meissa points to Polaris

Look for the Line Joining Sun, Moon, Planets & Zodiac Stars—the Ecliptic

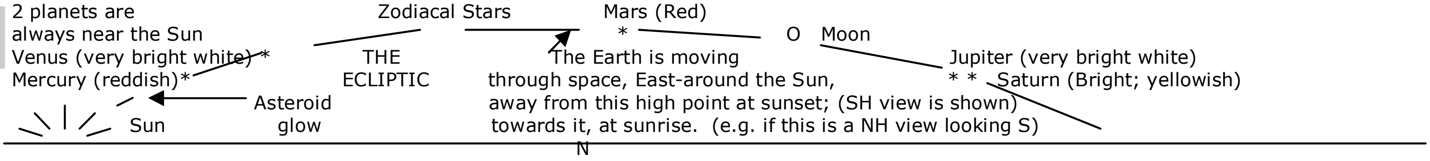

The Earth is moving
through space, East-around the Sun,
away from this high point at sunset; (SH view is shown)
towards it, at sunrise. (e.g. if this is a NH view looking S)

The solar system — Sun and planets — is flat, and the Zodiac marks that plane, by design. It’s **the ‘edge-on’ view, which we see, from within**.
The Moon orbits within 5° of that plane, and so is never far off line. **The line is called the Ecliptic** — a great circle.
(It looks as straight as the horizon, unlike the diagram. See ‘Optical illusion’ in the Moon box, p.8d)
During the day, you can orient, not by the Sun only or the Moon only, but by the line joining them. Sometimes Venus may be visible too.
You have to imagine the rotation of that line (more on this soon). It does not move in the same direction in which it points, but its path is up to 23°½ tilted from that.

Mnemonic: “Clock-wise Northern Hemisphere **SWiNES**”

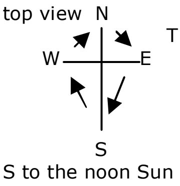

Europeans make chronometers to follow their Sun.
They see, looking South (diagram top right)… rise in the East ; South at noon ; set in the West.
That’s ‘clockwise’ (seen from above; East to South to West) for them
The swines! They force us, down under, to place our clocks on the southern wall and think behind us!
(We see, looking North: rise East, to North at noon, to West i.e. SH; anti-clockwise)……

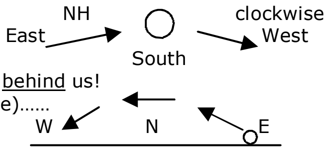

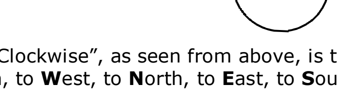

Anyway, changing your direction “Clockwise”, as seen from above, is turning to your right, and, **i**gnoring the **i**, the mnemonic is:
South, to **W**est, to **N**orth, to **E**ast, to **S**outh or “clockwise: **SW**i**NES**”

A mnemonic for global orientation: **“True Polarity”**

Earth

“**True**” courses refer to the Earth’s **Poles** — i.e. the axis of rotation of the Earth — **the single most directional thing on our spherical Earth**.
“True” bearings measure the angle between the course and the longitude line — a North/South line which connects the poles, and which passes through your location.

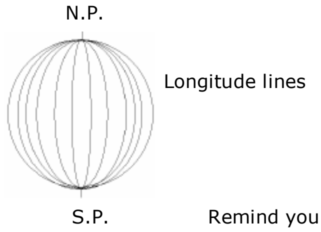
N.P.
Longitude lines
S.P.

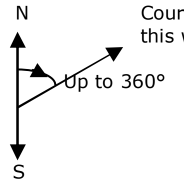
N
Course this way
Up to 360°
S

The angle is measured **clockwise from ‘North’**
e.g. 065°T
“Oh Six Five degrees True (-polarity)”!

Remind yourself, whenever you use “True”, by saying “True Polarity”!

Mnemonic: “**N**oon Sun Passes **N**orth of My **N**ose Down Under”— It Rhymes in ‘ **n** ’ (dow-nun-der)

Sun

(Now, Keeping your **n**ose to **N**orth, say…) (The second parts work for both hemispheres provided you face North, not noon)
“Sun **Ri**ses on My **Ri**ght (to the right of North)” —It Rhymes in ‘ **ri** ’
“Sun S**e**ts on My L**e**ft (in the W**e**st)” —It Rhymes in ‘ **e** ’
“That’s Anticlockwise around the North Pole Star, if you are facing North from the SH; [Clockwise around the South Pole Star”, if facing S from NH]
“The **SH** horizon **shi**fts to the **ri**ght”
“The **NH** ho**ri**zon **N**ever **H**eads **ri**ght, it moves left”

For Southerners, looking at the noon Sun i.e. from South of the tropics

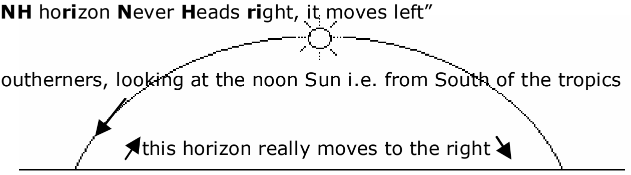
this horizon really moves to the right

West/Set/Left **North**/Nose/**Noon**/**Down Under** East/Rise/Right

Paper Diagrams aren’t Enough.

Imagination isn’t enough.

Go outside and take a look.
(You will need to shield your eyes from the Sun, e.g. with your hand)

Let the reality sink in.

The Two Brightest Stars Point out the **South** Pole Star

Stars

Canopus and Sirius line up almost South/North.
Use the star next to Sirius (‘Mirzam’) and the line-up is perfect.
The ‘Dog Star’ Sirius is the brightest star in the sky.
The Big Dog seems to have a front leg and back leg and a tail...

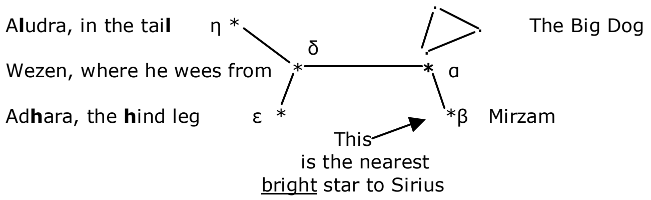
A**l**udra, in the tai**l**
η *
δ
Wezen, where he wees from
α
Ad**h**ara, the **h**ind leg
ε
β
Mirzam
This is the nearest bright star to Sirius
The Big Dog

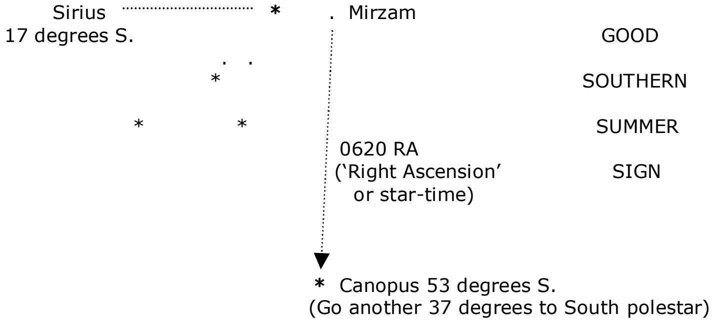
Sirius
17 degrees S.
Mirzam
0620 RA
(‘Right Ascension’ or star-time)
Canopus 53 degrees S.
(Go another 37 degrees to South polestar)

A
GOOD
SOUTHERN
SUMMER
SIGN

Mnemonic: “Wind? Weather Map!”

Wind

For navigating by the wind, don’t just leave it at the level of “The wind is from over there”.

* Look at the skyline “from Bruce Valley”
* Look at the Sunshine “from 30° to the left of the Sun at the moment” (Don’t look at the Sun!)
* Look at the Sun-path “a ‘sunset’ breeze” i.e. from sunset (Simply imagine its rise-set path.)
* Imagine the continent beyond the skyline “from the West Coast”
* Meld that reality to the weather map you remembered.

Don’t stop interpreting the wind until you can fit it into a Big Invisible Picture connected to that skyline over there.

Mnemonic: Down Under, The Sun Moves **ANTI**-BigB**ENWiSE** **i.e. anti-clockwise**

Anticlockwise

Look at the ‘ENWiSE’ in the mnemonic.
Anticlockwise is **E**ast to **N**orth to **W**est to **S**outh to **E**ast.
Ignore the **i**. Look down from above onto the horizontal directions involved…
Anticlockwise is turning to your left. “ Anti-clockwise is ENWISE ”

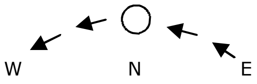

(If you put a clockface on the north wall and look at the Sun’s movement, (Don’t look at the Sun!), you will see that the Southern Hemisphere’s second-hands tick backwards! i.e. against the Sun’s movement. So put it on the south wall facing North, and the clock hands will move ‘with’ the Sun.)

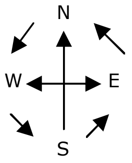
N
W
E
S
top view

In the tropics, the Sun will pass somewhere near overhead. To make it seem “anti-clockwise” anyway, face North.

Don’t get mixed up and think of Big Ben wise as clockwise. Down here our Sun moves anti- , and that is ENWiSE.

Mnemonic: "Time Tells Turning"

What's the use of looking at your watch and spouting numbers! At least, glance at the Sun, or rather at its shadows (don't look directly at the Sun) and then, imagine the Sun's direction at high noon as a reference direction.

That's what time tells you! — the Sun's position relative to its noon direction, North/South, called the "meridian".
A.m. says "ante meridiem" i.e. "before noon", and p.m. says "post meri-diem" i.e. "after mid-day"
Midday = 12 noon (but only in an ideal world, because noon is only approximately at 12)
One pm = one hour past 12md, and so on.
24-hour time tells you how far around the Earth has turned, since the Sun was opposite you i.e. since midnight.
(And then don't forget that the local significance of the time is: progress — distance made good — "watch" your pace)
Mnemonic:

"Watch" the Sun (without looking at it) whenever you watch your watch (and "watch" your steps too, for progress)

Notice these, Yesterday — to Guess the Time Today

- Sunrise time (and direction) ..If you know the time, you can guess at the directions.
- Sunset time (and direction) ..These first 2 are symmetric around North.
- The horizontal angle of separation between sunrise and set
- Midday time (and direction, and height) ..This is halfway between. Memorise how short your shadow is.
- When did it get light enough? To see. To walk.
- When did it get too dark? To cook. To walk.
- How long is the usable daylight? ..The day-length, between dawn and dusk.
- How many hours of dark?
- When did the Sun reach due East and due West? (How high?) ..This won't happen in winter. They are symmetric around midday.
- Times for the Sun at magnetic East, noon, West? ..These don't depend on guessing or calculating the magnetic variation.

Update yourself every week.

The Spinning Earth Makes **Circular Paths** for the Sun, Moon, Etc.

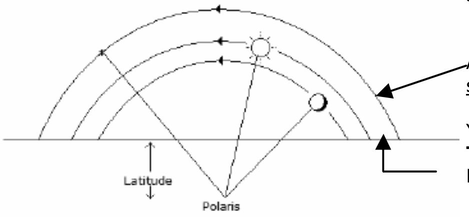

Sun, Moon, stars, planets — they all move only slightly in relation to each other during 1 day/night — but they all zoom around at about 15° per hour from East to West.

Although I can draw this on flat paper as a circular path, the Equatorial path actually looks as straight as the horizon, in reality! (Southern Hemisphere view)

You can now guess at time or direction by your familiarity with the circular 'orbits' e.g...
The highest point of the path will be North-South of you, if you can guesstimate it
Latitude determines the rise/set angle of the Equatorial path.
Rising and setting places are symmetrical around North.

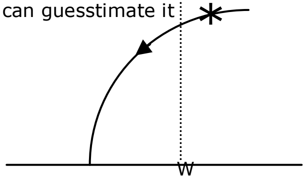

Although individual rising and setting is symmetrical about North, whole constellations that rise vertically in mid-latitude may set horizontally! and vice versa e.g. The Big Dog or Scorpio rise flat but dive vertically into the western horizon.

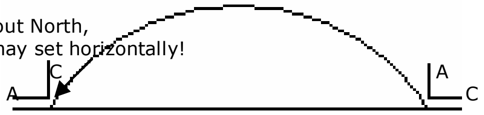

The Moon; Rule 4: Look for It

You have a 1 in 4 chance of finding it during the day. Provided you remember to look!

Rules 1 and 2 (like so many other navigational clues) are...

- PREDICT when and where it should be visible
- GUESS where to look And Rule 3 goes with the guess:
- LEARN from that mistake.

Keep looking for it. It's easy to miss. E.g. through the tree canopy; or rising through the horizon haze. Polaroid glasses help; you need to twist them

You will then have clues to… North, East, and West
Time of night (i.e. later)
Ocean tides
Tomorrow's Moon — Phase; rise, set, direction, height
Upper-level winds, if there are clouds about.

Hint: Measure the Sun-Moon separation with a piece of string (Don't look directly or indirectly at the Sun!) so you know exactly where to look later.

"VEERING" with the Sun is "Clockwise" only in the Northern Hemisphere "BACKING", against the Sun's movement, is "Anticlockwise" — but only in the Northern Hemisphere.

(The terms refer to horizontal directions changing either with or against the Sun's movement)
"Veering and Backing" are Eurocentric images, so they can be very confusing Down Under. Don't use those words in the Southern Hemisphere!

Say "Clockwise" and "Anticlockwise" instead
..for both hemispheres, as a matter of habit. Although these meanings too are Eurocentric, they are not confusing!
(the terms refer to a bird's-eye view looking down from above as if onto a clock-face)
Meteorologists and mariners still use the terms 'veering' and 'backing' and they do mean 'clockwise' and 'anticlockwise'
But the Sun-based image is backwards, south of the tropics.

The sky turns **15° per hour** — Use your **thumb and forefinger**

Earth

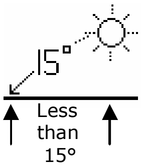

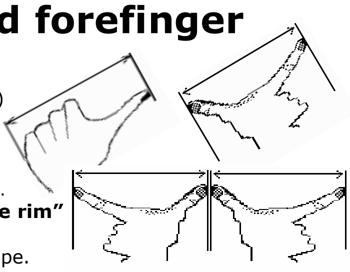

The Earth rotates once a day. Yes? That's 360° in 24 hours, yes?
Well that works out at 15° per hour — the sky spins 15°/hour to the West.
That calls for **recognising 15°** E.g. (diagrams left, and right) "one hour of sunlight left" (SH)

For Accuracy:
(0) Measure each handspan along the Sun's 'circular' path, not horizontally.
(1) Use constant stretch — of arms and fingers — 'just strained' — for consistency.
(2) Use constant chest-arm geometry. It is best to use two hands squarely in front of you…
(3)**Test for 15°, horizontally,** against the full 360° horizon — **"24 spans go around the rim"**
Span off, 24 times, from point-to-point right around the horizon.
Adjust your hand configuration until you find one that fits the "24 Hours" recipe.
(4) Try gripping one hand with the other, or touching one, at a specified point, with the other
, or, (5) Look along your shoulder, sideways, at one hand.

My recipe for accuracy fits my arms:"4 spread fingers, wrist bent back fully (at arm's length in front=60cm or 1cm/degree")
What do your arms require for 24 spans = 360°?

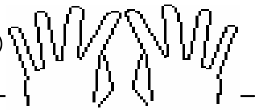

Mnemonic: The Southern Summer Sun Sets Somewhat South

Sun

Surprising perhaps. (If you equate North with 'hot' you get caught on this.) (for Northern Hemisphere substitute 'North' for 'South')
When the Sun comes South for summer its sunrise and sunset places also shift South with it.
Memory hexagon for Mid-Winter and Mid-Summer Down-Under at 37° S…
(top view) N. (the Sun moves anti-clockwise through North)

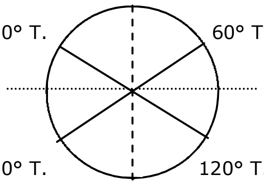

300° T. 60° T. … In Mid-Winter, Down-Under, at 37°S., sunrise is at 60°T. and sunset is at 300°T. … A 9½ hour day.
In Winter the Sun will never show East or West of you.

W At Equinoxes — Mid-Autumn, and Mid-Spring — rise and set are East and West… a 12 hour day.

240° T. 120° T. … In Mid-Summer, Down-Under, at 37°S., sunrise is at 120°T., and sunset is at 240°T. … A 14½ hour day.
The Sun will be East and West at **mid**-morning and **mid**-afternoon respectively, and **mid**-high — about 45°.
180° T. **Your main aim is to get familiar with the Sun's circular path**.

Stars Leave Star-Trails ..if you photograph them with time-exposure...

Stars

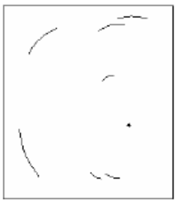

Star-trails are bits of circles, because… ..stars circle around the North and South Pole-Stars...
i.e. the Earth rotates on an N-S axis, which points to the Pole-stars. The pole star stays fixed in position all night (if you do).
Southern Stars Circle Clockwise, once per night-and-day, around the South Pole Star.
Northern Stars Circle Anticlockwise, around the North pole star, Polaris. These clock senses are the same in both hemispheres.
But **Note**: low Southern stars will move anti-clockwise around you, (i.e. to your left around the horizon, anti-clockwise as seen from above you). (Clockwise — to your right in the NH) (More on this later)

It helps you with global orientation to use 'star compasses'.

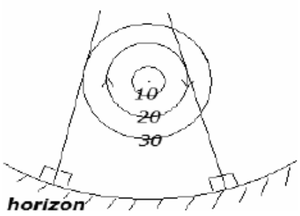

One of the clearest compasses is to blur the star circles into full circles then take vertical tangents to those circles.
In the following diagram, the left-hand direction is "30°" to the left of the Earth's axis (or to 60° maximum latitude).
The right-hand one is "20°" to the right (or to 80° maximum latitude). They are global.
If you follow these star directions, they will take you right around the earth in a 'straight' line — a big circle.
You will pass through places equal in latitude to the star which is overhead (p.11c), so you can already *see* the maximum latitude that such a 'great circle course' would take you to; then it takes you back towards the equator.
HINT: You won't get far in star-navigation, without a detailed star list and star chart. See p. 123 for a list.

Mnemonic: "Wind in Your Face"

Wind

The Southern Hemisphere version of Buys Ballot's Law… [Try to use unobstructed wind i.e. not in a valley nor in a mountain eddy]

- "**Face** the Wind [N.H. rule uses 'Back to the wind, Sun-twist…]
- Sun-twist yourself 15°-30° [i.e. anti-clockwise in S. H.]
  — to counteract landscape friction on the surface winds.
  [When the winds are impeded, they slow down, and so stop circling around the pressure centres, and start to cross the isobars, toward the low pressure (a 'sink' hole) at an angle of 15°-30°, less at sea and over flat land; more over rough terrain.]
- Rhyme: **"Low on My Left** &/or **High on My Righ-t"**

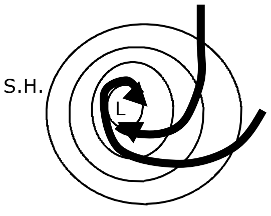

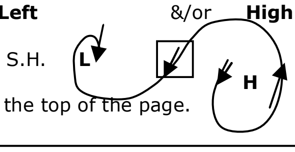

Top view of you, in the square, facing the top of the page.

Southerners Need **Anti**clockwise Systems

Anticlockwise

Our clocks seem to run backwards, compared to the Sun.
Our compass-bearings are measured 'backwards'. (see p.10e)
'Veering' and 'Backing' seem to mean the wrong clock-sense.
When the Sun moves forwards, all the other systems seem to push backwards.
We lose the global perspective.
And this section is all about aiming to put that right.
Consequently, we need a Southerner's viewpoint…

I hope you find that the anticlockwise systems which will follow soon are much more workable than Eurocentric systems.
Some of the other systems given on this right-hand side of the page, are universal — they can be used in either hemisphere.

Global Orientation: Latitude and Polar Distance

Step 1. Stand up on the globe again (in your mind and in reality).
Step 2. Put your left shoulder to the Equator i.e. Face East (sunrise).
(If you are in the northern hemisphere: face west, sunset.)
Step 3. Mnemonic: EQUATOR on My Left equals My LATITUDE is on My LEFT,
RISEN POLE on My RIGHT equals My POLAR DISTANCE is on My RIGHT.

Backviews...
Facing E.

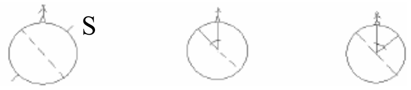

Equator — Latitude on left — P.D. on your right

Since the South Pole is 90° latitude, it is further South from the Equator than you are, and will be on your Right, as you face East.
Step 4. Measure the degrees of distance around the Earth's circumference as an angle at the centre of the Earth (by slicing the globe in two).
The two angles mentioned must add up to 90° i.e. they are "complementary" angles (spelt with an 'e').
If you are, say, 45° south of the Equator, then the closest pole must be 45° further south. The 'distance' to the Pole is 45°.
If you are 30° south, then the South Pole will be 60° away from you. Your '**colatitude'**—or *90° minus latitude*—or "Polar Distance"—is 60°.
The goal is to imagine those two angles — Latitude and Polar Distance — at the Globe's centre beneath your feet, and around the circumference.

Mnemonic: "Sun at High Noon — How High?"

When you navigate by the Sun, think: "Sun": the mnemonic is: "Sun at high noon — how high?".
You look: North, there, then up there, that high.— Southern Hemisphere (cf Noon will be South in NH) (Don't look at the Sun, if it is already there!)
It might be morning : "It is going up there"
It might be noon : "See how high it gets!"
It might be afternoon: "It came from up there"
Visualising the height and path of the Sun also helps you to interpret plants for 'bush noon' (p.74a).
"Walking Stick North" depends on knowing the Sun's maximum height. (See page 30b)

On Overcast Days, Average the Brightest Sky over 1/4 hour

If the clouds are moving fast enough, and not entirely uniform, sit still for long enough to see many clouds pass over. Keep pointing at the brightest patch of sky using some reference like a tree, and watch to see whether you can confidently locate the average position of brightness.

RISE and SET Directions for a (**Local**) STAR COMPASS

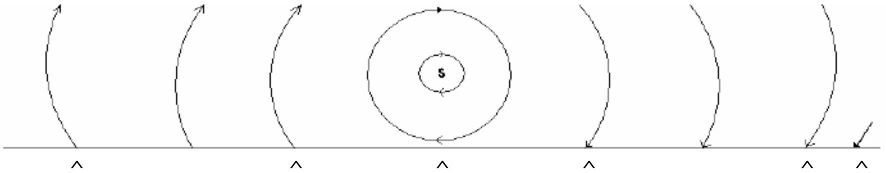

The star directions don't change night by night or with longitude, but they do change with latitude.

The circles in the diagram will rise from the horizon as you travel further south, and so change their intersections with the horizon.

This star rises here every night for the seasons when it is visible

You need to know your stars and visualise their circular paths

Stars which share the same star path indicate the same direction at staggered times during the night

Sun-set and Moon-set (examples only) can be fitted in to the scheme to span the day

Just make sure to mark sunset to a landmark on the skyline horizon and carry it over into night-time.

(SH view)

(Some stars stay up all night every night; these don't move far either, and so are good orientation value too)

From the Moon, Locate the Sun-Moon Direct Line, Carefully.

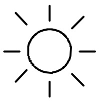

The line is the Approximate Ecliptic, and being a great-circle is a straight line.
OPTICAL ILLUSION!
The Moon seems to be lit from above the 'direct' line to the Sun.
But that may only prove that the Sun is "up" and shining down.

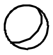

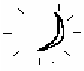

Bust the illusion with a piece of string…
Caution: Don't look directly at the Sun!
horizon

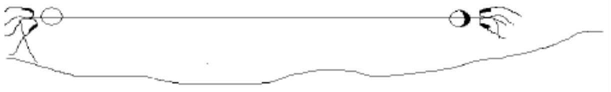

Prove it in 3 dimensions, by using a torch and a ball.
When the ball, in line with the Moon, looks right, the torch direction mimics actual sunlight.

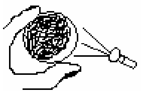

Below horizontal = at night

By locating the Sun, you can guess the time of night, and guesstimate North more reliably, because the path of the Sun is more predictable and familiar than that of the Moon.

Mnemonic: (when you say) "COMPASS" (say) "CANADA"

Compass needles point to Northern Canada. That's where the North Magnetic Pole lives. Lat. 75° N.; long. 100° E.
E.g. from Australia, Canada lies to the East of North…

Check it out on a globe. Locate the Queen Elizabeth Islands.

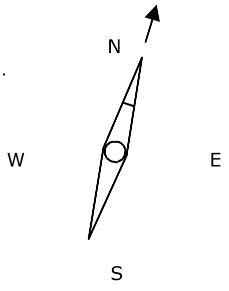

To find the Southern Magnetic Pole, (but this doesn't work so well)
locate the Antarctic coast south of Adelaide, at Lat. 65° S.; long. 140° E.

The Angle of Rising and Setting — is Determined by Your Latitude

To post-dict sunrise or pre-dict sunset direction, you can use your latitude to visualise the tilt at which the Sun **rises** (or the rise-angle of a near-Equatorial Star/Moon/planet) and the deviation from straight down, at which it **sets**.
Remember to use vertical, not horizontal as the expected norm. (To judge vertical correctly, see p19c. To practise, see 30b)

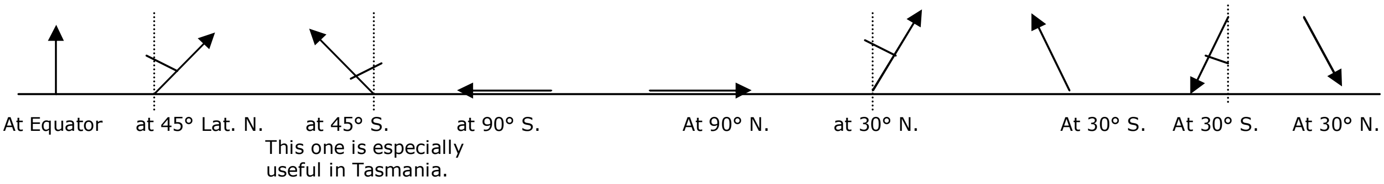

The Sun's path is very nearly a straight line, despite the illusion of being circular—see p.15a.
Hint: Whenever you see the shadow of your head, provided the Sun is not too high, you can make the "anti-Sun" rise or set positions obvious, by following up from that shadow, to the horizon at the correct angle. In midwinter the anti-Sun will be in a midsummer position and vice-versa.

High Noon — How High?

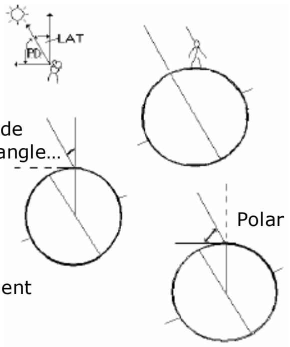
LAT
PD
S.P. and two parallel rays
Latitude angle…
Polar distance angle = elevation of the average Sun

Answer one: the same as last week. Take a look. It changes only slowly.
Answer two: **within 23½ degrees of your co-latitude (or 'polar distance') high**
i.e. 90° – Lat., plus or minus 23½ degrees, depending on the season.
Mid-Summer… co-lat+23.5° (very high in the sky)
Mid-Winter… co-lat–23.5° (very low in the sky)
Equinoxes… co-lat (average height).
Why? At 0° Latitude, i.e. at the Equator, the average Sun is 90° high at noon i.e. overhead. Yes?
At the Poles, 90° Lat., the Sun skims the horizon at equinoxes. Yes?
If you need further proof (that its height is 90° minus Lat.) draw it.

See how the height of the average Sun is 'latitude degrees below vertical'.
That's an angle of elevation of 90° minus lat, isn't it? Now adjust for the Sun's seasonal movement which is 23½ degrees each side of that.

Global Star-Compasses e.g. 'Perpendicular to' rise and set

This is one step harder and three steps better. It allows 'great-circle' course keeping.
A global Star-compass yields 'direct', global, 'straightest path' courses — ones which take you around the globe and back to your starting point.
E.g. You take a rise or set compass-point and head at right angles to it, and the new horizontal direction is global—no longer tied to one latitude.
That's easier to imagine by visualising the light half/dark half of the Earth: All who see the Sun on the horizon could point to it, but each would be pointing in a different "direction". But if they all pointed horizontally at right angles to the Sun, they would all be pointing in truly the same way —around the Great Circle which divides day from night. The same goes for any horizontal star. Imagine the star as a pole star, and use its "equator"
There are many alternative star appearances, besides rise or set & **you don't have to 'go at right angles to' these to set your course...**

E.g. aim at **one star above another**
The midpoint of One star **level with another**
One star **rising or falling vertically**.

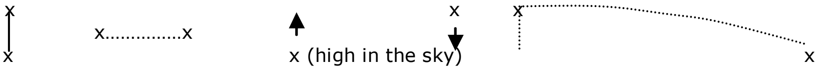

Toward a star which has a 2nd star off to one side **at right angles** i.e. lower. You would pass *under* the point you kept aiming for.
You don't need a clear sea horizon — the stars are high, and a plumb-line will suffice.
These guides occur every night, in the same global direction, for several seasons, then disappear for a while due to daylight, but reappear.
They will occur at the same times next year. **They are always worth learning more of**, adding to your familiarity of favourite 'Star directions'.
These global directions are labelled by saying *how close to the poles* they would take you, and you can see or guess that from the stars, see p. 7c.

'Low' Clouds Come Clockwise Down Under

In the Southern Hemisphere

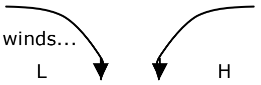

- Low-**pressure-system winds** spiral **clockwise** around. I.e. into your face from your left as you face their winds...
- 'Low **clouds**' (i.e. below freezing level) follow the surface level isobars i.e. the weather-map lines.

The double-meaning for ' Low ' in the mnemonic reminds you of this.
'Lows' cause more clouds than 'Highs', so you focus on 'Lows'.
I.e. when you see the low clouds; you say 'low clouds come clockwise' and automatically imagine………… '**L**ow On My **L**eft' Plus 'H**igh** On My Ri**gh**t'
The Alternative Southern Hemisphere version of Buys Ballot's law. This **new version goes like this** :
Face the low clouds (unaffected by friction—winds at 600m high are considered unimpeded),
then rhyme: L— L ;igh—igh. (Don't Suntwist yourself.)

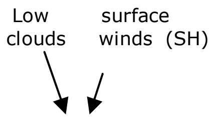

[Cf "Face the wind, (In the northern hemisphere you have to say "back to the wind")
Sun-twist yourself up to about 30° to counteract friction (which makes the surface winds spiral inward to the low centre)]
Surface winds could also be expected to be half the speed of unimpeded winds, less impediment over sea, more over land—about one third.

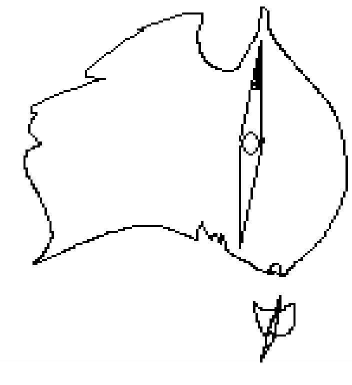

A Compass-Needle Continent—Swinging the Continent

Mentally glue a miniature continent under the compass needle, to swing with it.

The idea is twofold:
1. To reprogram your mind — "The continent is not swinging, so... **I** must be!"
2. To orient yourself meaningfully — to the invisible picture
3. To remind yourself of compass variation

When using a compass you could: just stare at the compass-needle, like a lone night sailor in a fog.
Or: look at what the compass is pointing to — that bump!
Even better: also 'see' the needle lying across the whole continent
Keep going: pointing to Northern Canada
Don't forget: the map you are using — the real one, not this imaginary one — should also line up properly — twist it around until it does.

Global Orientation: Point to the South Pole… ..half its polar distance below horizontal

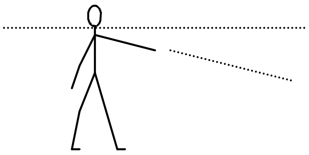

If you had to point "South" you need to point horizontally.
But when you have to point out "the South Pole", you must point below horizontal.
(It is similar when Northerners use the North Pole)
Imagine the true scale… you are on the circumference, not above it...

The question is: "What is the exact angle below horizontal?"
The answer is : "The South Pole is **HALF its polar distance** below South."
Example: at 45° S., you point 22½ degrees below the horizon. (90-45) /2
Example: latitude 30°; colatitude = 90–30 = 60°; half that equals 30°
Example: latitude 60°; polar distance 30°; 15° below horizontal.

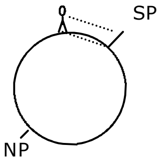

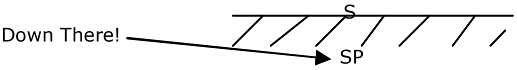

Optical Illusion — Beware! — Use Your Shadow Not the Sun, For Direction

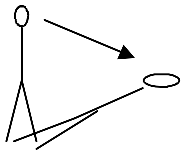

When the Sun is directly behind you, it seems to be behind on the left when you look over your left shoulder, and behind on the right when you look over your right shoulder! And when the Sun is exactly to the side, it nonetheless seems to be ahead of you. (Every sports coach knows that when you turn your head, your body follows it.) Don't rely on where the Sun seems to be… USE A SHADOW INSTEAD! It is also easier on your eyes.

And then don't get fooled by the shadow of your finger, when you throw a shadow onto your map to find the Sun's direction. It may look straight but it doesn't point to the Sun, until you line it up with your finger**tip** while looking from directly above. The shadows will always be parallel, but the Sun's rays will always join object to image, fingertip to fingertip-shadow… The first diagram is illuminated from behind you……
To avoid looking at the Sun, try this:
Point your walking stick from somewhere beneath your eyes, outwards, and rotate yourself until it is in line over its own shadow.

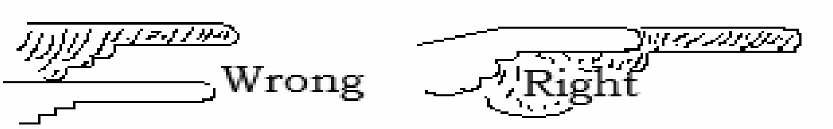

Right Around the Horizon, Stars Move to Your **Left**, Down South

Looking South, all the lower stars move left.

It is only the high Southern stars which move to your right.

Looking East all the stars rise to the left (i.e. we see the bottom half of the circles down at horizon level, looking South)

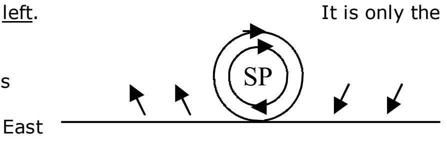

Looking West all the stars set to the left

Looking North, all the stars move left. (we see the top of the circles)

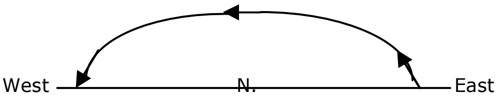

The same goes for Sun, Moon, Planets.

Looking directly overhead, or behind your head, will not work!

(it is all opposite in the northern hemisphere)

Mentally Rotate the Sun-Moon Line Around Polaris

A whole line through the sky gives a better 'feel' for direction than a single point — than either the Sun or the Moon alone —
So join the two and watch them rotate together during the rest of the day.

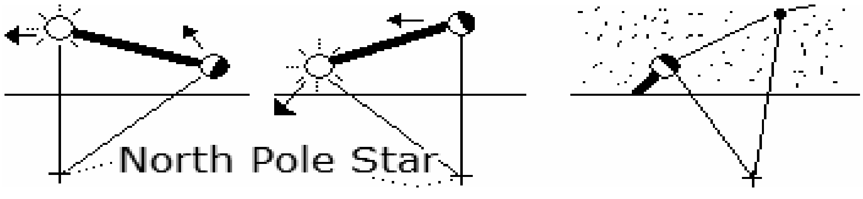

(Southern Hemisphere view)

When you can see both, and you know where North is, you watch them.
When you see only the one, you look for the other, and so find north more exactly than by using either one's path by itself.
Add in any planets to continue the process at night, then from planet to planet or to Zodiac stars, after moonset.

"Bears Circle Clockwise"
Compass Bearings are **C**lockwise, A Full **C**ircle, and Euro**c**entric, **C**ee?

In Europe the Sun moves clockwise… from North (Where the Sun is at midnight—0 hours) = 0° as far as the horizontal angle involved is concerned.

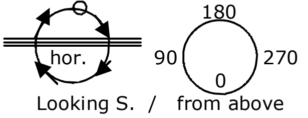

East is at about 06 hours (sunrise) or 90° True — Sun-wise or Clock-wise from North
South at about 12 hours or 180° T (noon)
West at about 18 hours or 270° T (sunset) — so you see, the bearings circle clockwise, from North, 360°.

A **Southerner**, of course, would prefer to go 360° **anti**-clockwise, from **South**, but they didn't get a vote.
Everyone in the world is expected to submit to the (NH) rule: compass bearings should be measured in degrees, up to 360°, from North, clockwise around through East — like the ancient Babylonians; like the Northern temperate Sun.
Not many people consider using 400 grads (4 right-angles of 100 grads make up the circle), or consider going left-around, through West,
Or even think of starting at South. So don't you dare (unless you **don't** call it "a compass bearing").

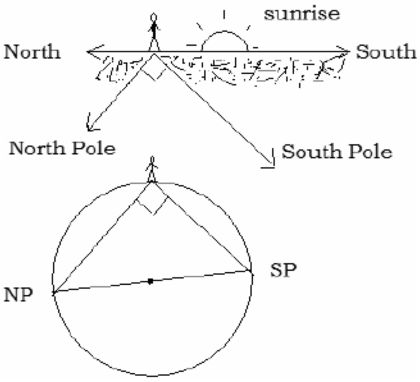

The North Pole is *Perpendicular to the South Pole*

Since the poles are 180° apart — opposite — you see them as 90° apart — half the arc distance.
The numbers can't work out any differently.

Example: 10° S. Lat. = Equator is 10° North of you. South Pole is 80° away — the "Polar Distance".
So S.P. = 40° below S.
North Pole is another 90° past the Equator.
So N.P. = (90 + 10)/2 = 50° below N. And the gap is 90°.

A simple way of looking at it is to say
"Diameters yield right-angles at the circumference"

Now go outside and point to the Poles.

The Dark-Blue Polarisation Band Perpendicular to the Sun

90° from the Sun there is a circle of darker sky. Bees navigate by this clue. The whole sky shows degrees of polarisation.
You can see it naked-eye, but since it is due to polarisation, it is much clearer through Polaroid sunglasses or a camera filter.
Point the top of your polaroid sunglasses toward the Sun for best effect.
The dark blue band is narrowest highest up (in the direction of your shadow, and perpendicular to that sunbeam) and varies with visibility and haze.
The sky is darkest away from the Sun and away from the horizon haze and 90° from the Sun.
You can always guess where the Sun is, from the gradations of blue. After sunset the band is caused by the brightest part of sky.
Example: You are in a forest at dusk. The sky is blue overhead, but the sunset direction is not visible. Find the Sun from the blue!
See the "Sunny Skies System" (p.19e) for how to use the ends of the dark blue band, and its trend.

The Centre of the Rainbow is Down-Sun

The shadow of your head coincides with the centre of the rainbow's circle. (Pilots often see the full circle of a rainbow)
The radius of this circle is two wide hand-spans — spread out your 10 fingers fully, thumbs touching.
Whenever you see a rainbow, treat it as a clue to 'shadow-South' (p.14b) i.e. relate its position to the time of day and to South.
The shadow of your head is on the end of the shadow of you, of course, but when you can't see it due to clouds, you might be able to see a rainbow.

Identify the Stars That Pass Directly Overhead

Each star has its own latitude, called "declination", which is very nearly constant.
So the same series of stars of the same latitude will pass overhead that latitude, each night, each in turn, each at some fixed star-time.
This is true anywhere on the globe.
Use a plumb-line. The top of the plumb-line points 'up' to overhead (when looked at from any side, so) where the 'up' lines cross is 'overhead'.
Look past it from 2 or 3 directions to find overhead, especially looking North/South for longitude-timing, and East/West for latitude-positioning.
Then use a star-chart to identify "exactly overhead", latitude-wise, at least. If you can do this, even without a watch, this will fix your **latitude**.
Plus or minus a few kilometres is the expected maximum accuracy from using a plumb-line.
For **longitude**, you need to determine the star-time at which one of these stars is overhead, or at least passing from East of you to West of you.
Greenwich (absolute) star-time, not Local (relative) star-time, that is. To the nearest 20 seconds is okay. See p30c. A 4 min error is 1° error in long.
Alternatively, set your reference star-time to some other longitude, e.g. home, or start, or destination, or a round number.
This time & position data is life-long, give or take a few kilometres of predictable precessional drift per decade.

You can get back to that place later — Match the appearance and timing. It's easy to tell which way to move in order to "match up" (p22a,c; 23c).
You need to know a star's longitude, so you need an almanac or list. Look down the list for bright stars with your latitude, and go look for them.
You will quickly learn about having to wait for the right 'star-time'!

Weather Fronts Distort The (Circular) Wind Circulation

Buys Ballot's law is only approximate, because of the sudden wind-shift at a cold front.
Cols, Ridges and Troughs are likewise noncircular. E.g. 'centres' will sometimes split into two…

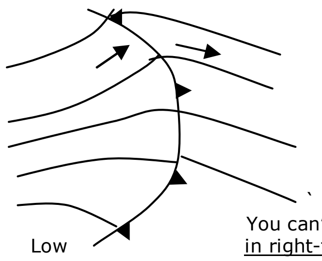

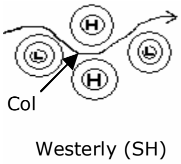

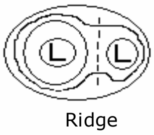

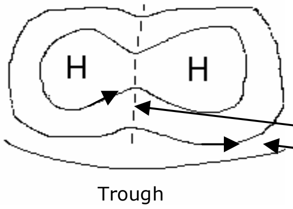

In the trough, crossing the ridge, the winds are turning right.
This airstream is turning left (SH) around the High centre.

You can't accurately say that these winds are 'circulating' around the Low or High! But you can say that in right-turning winds the clouds are forming, under the influence of a Low or a trough (SH), and left-turning winds are under the influence of a High or a ridge—the clouds are dissipating. Swap left and right in NH.

Mnemonic: (Interpret a Compass Needle as) "RED HOT — WHITE FROST"

Southerners are looking North to the "red, hot, tropics, dripping with sweat, in the Sun"…

When southerners look South, it is to the "white, icy, Antarctica, frosty cold, in the shadow"…

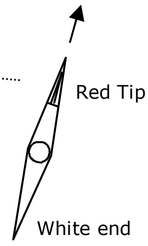

So when you notice the red point of the compass needle you say to yourself "red-hot!"
and try to feel boiled on that side (as if by a bon-fire), and then visualise the invisible tropics North of you.
And when you notice the white point (e.g. during a back-bearing) you say to yourself "white-icy cold-frosty-Antarctica"
And then try to feel frozen on your 'Shadow-South side', and imagine the South Pole, e.g. point (down) to it.

You can also remember that the Canadian flag is red (but it is not hot in Canada).

Global Orientation: the South Pole Star, and Polaris

(In the Southern Hemisphere…)
Polaris is below the northern horizon, by exactly as much as your South latitude.
E.g. at 45°S. lat. Polaris is below North by 45° Around it the northern stars circle anticlockwise……

The equivalent "South Pole Star" (a dim pinprick in the midst of dim pinpricks) is **above the southern horizon by exactly as much as your South latitude**. E.g. 30° S. = 30° above South
Around it the Southern stars circle clockwise, once per day-night.
Locating the South Pole Star, helps you in the following...

- You can use the bright stars — Cross, Pointers, Achernar — as hour-hands on a 24-hour clockface.
- Star-paths Circle around the Pole Stars
- So do the horizon, the Sun, the Moon and the planets — You can predict all their positions.

Finding the Pole Stars is Fundamental to Visualising Rotation.

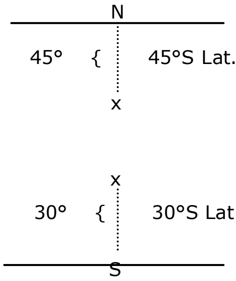

- Your Shadow Rotates with the **same clock-sense** as the Sun — not surprisingly!
- Shadow-Rise and Shadow-Set is **opposite direction** to the Sun — not surprisingly!
- Mnemonic: (Face the opposite direction as for the Sun mnemonic p.5b)

} Together they can surprise you!

Face **Sh**adowy **S**out**h** for **S**outhern **H**emisphere (Face Su**nn**y South for **N**ortherners)

"Shadow **Ri**se is on the **Ri**ght; S**e**t is on the L**e**ft" (Same rhyme as for sunrise and sunset, but facing South, both hemispheres)
Your Shadow Passes **South** of Southerners At Southern Noon (compare: 'The Sun is North of Your Nose at Noon')
(North of Northerners at Noon)
Your Shadow Moves **Anti**clockwise — around to Your **Left** — Down Under
(Clockwise — Around to Your Right — in N. H.)
(The Southern Hemisphere Horizon still shifts to the right, in real terms. The Northern Hemisphere Horizon still Never Heads Right')

In addition, notice: **The shadow *tip* moves due East, at around noon,** and in a much straighter line than you expect, all during the day, as the shadow itself changes, from pointing to opposite sunrise, then to pointing to South of the object, then to lying opposite sunset.

---

15 Ways to Locate the **South** Pole Star Or the N./S. Meridian

1. Halfway between Achernar and Epsilon Centauri (located from the six bright stars near it)
2. Rigil Kentaurus to S., at right angles to Hadar…………
3. Beta Crucis to S., parallel to gamma-to-alpha Crucis, 5 Cross lengths … Use 2. and 3. together
4. Beta Hydri to alpha Crucis, as far as the beta-to-gamma-Hydri length … most accurate
5. Mirzam through Canopus
6. Suhail through the top half of the False Cross through Miaplacidus
7. Make a triangle out of the two Magellanic clouds toward R. K. — very crude but useful
8. Gamma Hydri to beta Triangulum Australe — the "back" star of the Triangle
9. Alpha Hydri to Hadar
10. Arcturus to between the two 'Pointers' (R. K. & Hadar) } These are used in reverse
11. Spica to between the Cross and Pointers } — to locate Arcturus
12. Middle of Corvus through middle of Cross } Spica and Corvus
13. Atria to the star between Canopus and Achernar, alpha Dorado
14. Peacock to Avior
15. 'Your Latitude' degrees above horizontal, due South

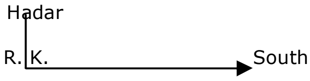

("Crucis" says "of the Crux" i.e. of the Cross)
(gamma looks a bit like a thin alpha twisted 90°)

"Cusp North" — Approximate Direction from the Moon

The Sun-Moon line marks the approximate Ecliptic.
The "centre" of that circle of Sun-planets-asteroids is located in "Draco", quite near "Polaris", the North Pole Star. For a start, ignore the difference…

- Join the Cusps, tip to tip
- Drop the line diagonally
- Keep going down to (near) Polaris (latitude degrees below horizontal)
- Come vertically back up to "Moon North" (in the worst-case +/-60° "accuracy"!)

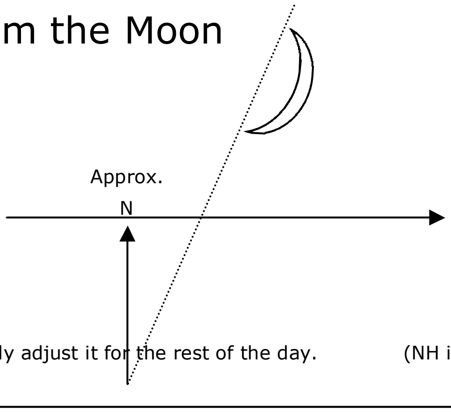

You have to do this only once or twice a day. Compare it with True North and automatically adjust it for the rest of the day. (NH is similar)

Mariner's "Compass Points" work on Successive Halvings

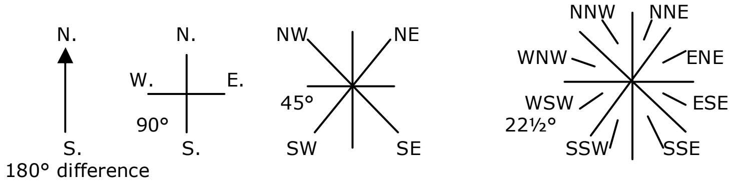

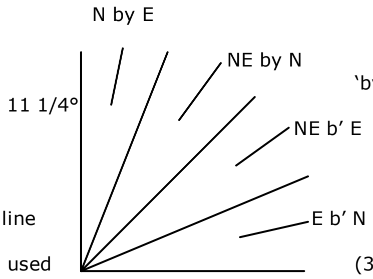

Start with the nearest major point; add 'by'
'by' means 'come 11 1/4° towards' (or 'one point')
There are 32 points altogether on the compass.
One Point is approximately One Palm wide
(32 around the horizon—test it)

Make your own emergency protractor (e.g. as a compass, sextant or plotter) by folding any piece of paper, and folding it again and again, matching line to line so as to exactly halve it with every fold.
One more halving and you get 64 divisions, each of "100 mils" (so named and used because **1mil is approximately 1 in 1000 e.g. 1 m in 1 km**) i.e. 6400 mils around, clockwise, 17.7778 mils per degree.

Pole Star Locations — "(Latitude)° above and below, South and North"

Why? Think: at 0° Lat., they are 0° high. Yes? At 90° Lat., they are 90° above and below horizontal. Yes?
If you need further proof: Draw a Diagram of two parallel rays to South Pole Star...

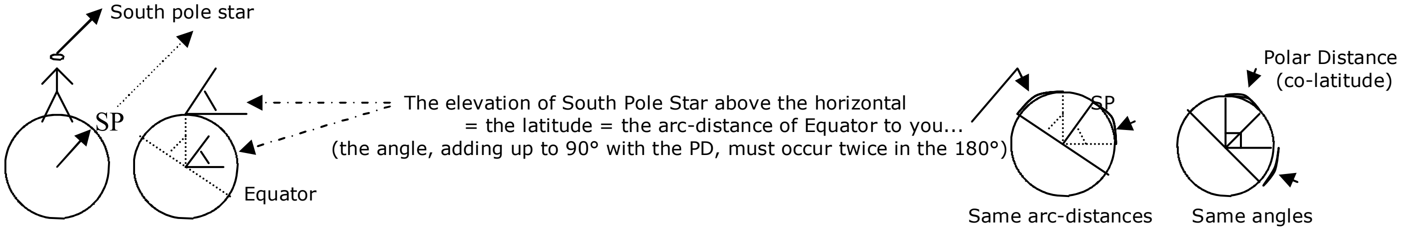
South pole star
SP
Equator
SP
Polar Distance
(co-latitude)
Same arc-distances
Same angles

The elevation of South Pole Star above the horizontal
= the latitude = the arc-distance of Equator to you...
(the angle, adding up to 90° with the PD, must occur twice in the 180°)

The Sun Passes Quickly Past North-South Nearer the Tropics

This pattern shows
the hourly change
in Sun or shadow direction
at 30° latitude, Mid-Summer...

The Sun is East/West
at (noon −/+ 2 hrs 44½ m.)
at 52.9° high.

This shows the
seven hours' sunlight
(one side of North/South).

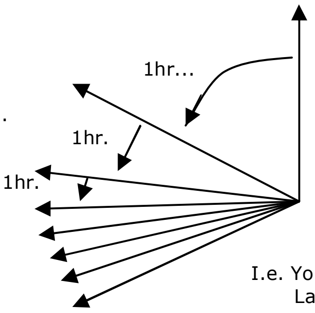

Don't get caught misjudging the Sun's horizontal direction near midday.
When it is high, it moves more than 15° per hour in horizontal direction.

At other latitudes the pattern will be different — more uniform
away from the Equator, and with a different day-length,
depending upon the season.
I.e. You need to calculate each particular pattern from the
Latitude and declination and season (as if calibrating a vertical stick sundial).

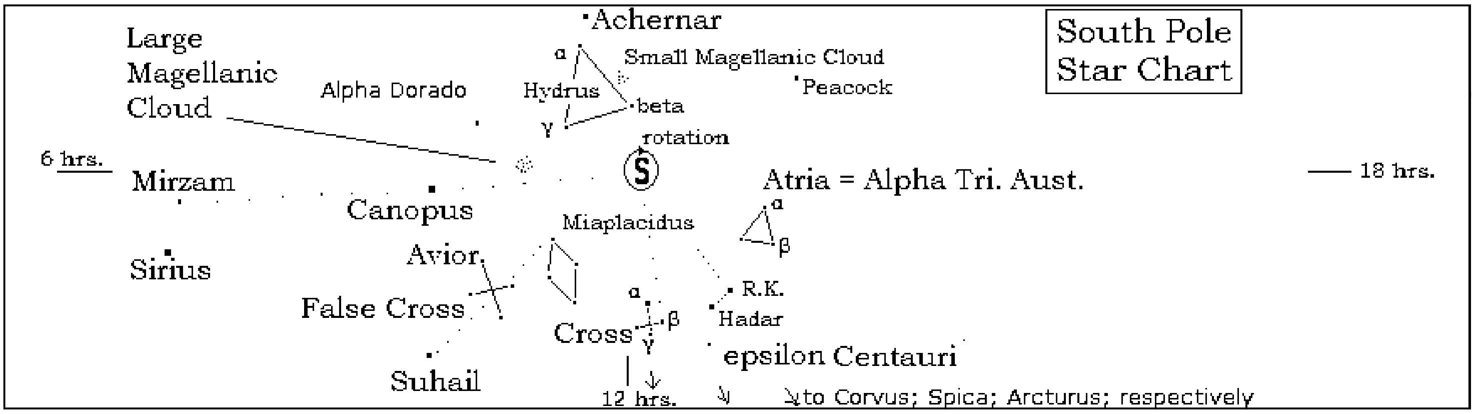
South Pole
Star Chart
Achernar
Large
Magellanic
Cloud
Alpha Dorado
α
Hydrus
γ
Small Magellanic Cloud
beta
Peacock
rotation
S
6 hrs.
Mirzam
Canopus
Atria = Alpha Tri. Aust.
α
β
18 hrs.
Miaplacidus
Sirius
Avior
False Cross
α
β
γ
Cross
R.K.
Hadar
epsilon Centauri
Suhail
12 hrs.
to Corvus; Spica; Arcturus; respectively

Wind-Direction-Change Rule for Buys Ballot's Law

If "the Low is on Your Left Or the High is on Your Right"…**and** the wind is gradually **changing**: Add in this new information as follows...
Imagine how the Low and High are moving across the ocean/continent as you **reenact the wind change**...
Exaggerate the change as you *twist yourself to follow the speeded-up movement* of the wind change in your face.
Point to the two choices (left & right) while you twist, then ask yourself which of your hands is moving to the east.
Usually only one of the two options above — 'Low' or 'High' — will make sense of the general movement of the weather systems to the East.
Temperate systems usually do move to the East from the West. (Both Hemispheres)
..From 0 knots (stationary highs) to 15 (average) up to 25 knots (Lows in winter).
Tropical cyclonic hurricanes move Westerly initially at 25-40 knots, then recurve to the pole, but the "re-enact" rule still works to show up its movement.
In this way you can decide: "These winds are from a Low" or "from a High".
Don't, of course, confuse local winds, e.g. a sea breeze, with the more general 'synoptic' circulation due to the large-scale pressure systems.
To further help you decide where the wind has travelled from —from your right or from your left— check out the temperature and absolute humidity — i.e. the air-mass characteristics. Is the air hot and dry, or cool and moist, or warm and humid — which one is supported by the geography?

Four Reference Directions — Name Them: "**Rise, Noon, Set, Midnight**"

..not E N W S

The trouble with "North" is that it is sometimes behind you... then you have to think backwards.
Use all four and you always have one in front of you... N., S., E., W.
Give them meaningful names as in the title... and you can picture the meaning.
Base it on the Sun's daily movement and the effect is vivid.
To flesh out the meaning even more, **add in a second clue: "Rise/Start; Noon/Hottest; Set/End; Midnight/Coldest"**.

What this does is to take the outside picture inside with you, when you enter a new building, a hospital complex, a shopping centre.
E.g. "I'm now looking toward the mid-morning Sun direction, i.e. 'after rise' or 'before noon', which is 'warming up' ".

It is then up to you to
1) get to know the Sun's daily movement.
2) locate the Sun's path outside, once.
3) *keep track* of your twists and turns when the Sun is not visible e.g. at night.

Ignore the time of day or night—it is the directions which you are imagining. You use the real Sun only as a start-up hint.

Point to 2 Places on the Equator — Halfway between the Poles

You can find the 2 Poles, apparently 90° apart, so that means you can point 45° — halfway — between the Poles, to find the Equator, then 90° to that for the opposite Equator (the "45°" is because the Equator is 90° from each Pole and we halve that arc).
It locates (1) the furthest point on the Equator from you e.g. Brazil then...
(2) the nearest point on the Equator from you e.g. New Guinea—ish.

From central Australia…

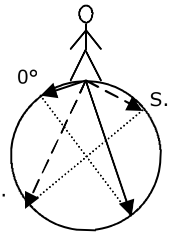
Closest Equator 0°
(PNG)
S.P.
N.P.
Farthest Equator. Notice that it lies to your South! (Brazil)

Shortest Shadow South — Opposite Highest Noon. Always Imagine "Shadow-Noon" on your "Shadow South" Side

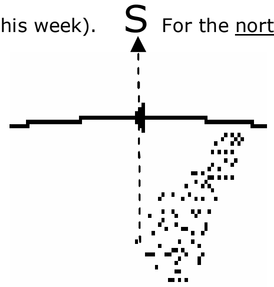

i.e. Your shortest shadow is South of you at noon. Memorise how short it is (this week). For the northern hemisphere use "North" ("shadowy-North")
(Shadow Noon is Opposite Direction to "Noon-North") .."Noon-South"
Imagine your shadow moving left around… ..right-around

In the morning: South must be "ahead" of my shadow ..North
At midday : Look — see how short it is! … **Due South at noon** **..North**
After noon : My shadow has "passed" South ..North

To guess North from your own shadow, see "Walking stick North". (p.30b)
Try to feel the icy South, freezing that side of you! ..North
(As on a cold night facing away from a campfire).
Mentally **visualise**: Antarctica below horizontal; an aurora above that horizon; Southern stars behind that blue sky. ..Arctic; Aurora Borealis
In the tropics, you'll have to be more observant, to know your shadow's expected movements more accurately. ..Northern stars

---

Find the Polar Axis by Star-Trail Bisection

You need to imagine the centre of the star-circle.
A Hypothetical Example: Lost in a stormy night! You want to find North-South.
You catch the occasional glimpses of unknown stars.

Lie down and line up one such star with a tree branch.
Maintain your exact position and wait for another glimpse.
It may only need a couple of minutes to see its direction of movement.
You can also point something straight at the star and fix it in place, to see the star's movement.
Bisect the result. It points to Polaris, or to the South Pole Star.
You now need to know your latitude, roughly, to locate Polaris or South Pole Star.
The higher the star the clearer the result — it will be moving West.
The only ambiguity is a low star moving horizontally (to your left).
The Southern stars will be moving slowly; northern ones, quickly
— so you may be able to guess which is which.
Northern Hemisphere is similar — but Polaris is visible.

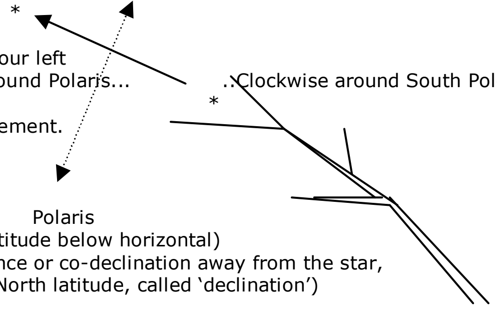
South Pole Star (latitude above horizontal)
(This star is moving to your left
Anti clockwise around Polaris...
..Clockwise around South Pole Star)
Polaris
(latitude below horizontal)
(also polar distance or co-declination away from the star,
if you know its North latitude, called 'declination')

The Rabbit Looks up Ahead at Full Moon Using the Ecliptic

When the 'cusps' disappear at full Moon, the Rabbit ears and face appear,
and let you know which way the Moon is moving.
These views are for the Southern Hemisphere — where the 'Rabbit' is upright.
In the first diagram, the cusps are difficult to distinguish, and so is the Rabbit.
In the second diagram, you don't need the cusps, because the ears show.
The dashed arrows point down to near Polaris.
As an alternative to cusp-North, you can use the Ecliptic perpendicular to the cusps.
The other, solid arrow, falls roughly along the Ecliptic, with the light side of the Moon pointing to the Sun and the dark side pointing to the anti-Sun.
Therefore you can judge approximate sunrise or anti-sunrise, sunset or anti-sunset, just as soon as you see the Moon, either from its cusps, or from the Full Moon.
Find the nearest horizon on the Ecliptic and picture which of those 4 directions it is.
A setting Moon in early daylight will show anti-sunrise, but after dusk will show up sunset.
An eastern Moon before dawn indicates sunrise, but in the p.m. will point to anti-sunset.

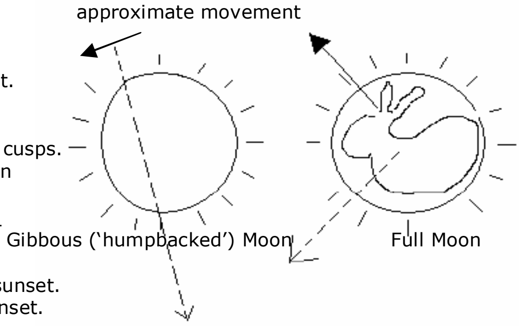

"E. W. N. South" **Of What**!? Rule: **Say** What

- Say 'True' and you mean 'degrees East of True North' 270° T = due W.
- Write "Gee" and you mean: relative to the map **grid**. 270°G ≠ 270°T Both are clockwise around, but they start off from different directions
- S20°E means '(Start at)South(;come)20°(toward the)East' e.g.S20°E=20° to the right of the white end of the compass needle. (old system)
  Clockwise/to the Right/East-around-from-North is standard if not otherwise specified.
- 270° M says 'Magnetic' — what a good compass should read.
- 270° C says 'Compass' — what this compass does read.

Note that these are standard: 'G' does not mean 'Geographical' or 'Great Circle' or 'Grads' but is recognised as 'Grid'.
'M' does not mean 'Map', 'Map grid' etc.
xyz° can not mean anti-clockwise.
The remaining problem is how to shift between different methods… See section 4: "East Least West Best" rule. (p.107b)

Perspective Distortion — The Sun Moves in a Straight Line!

The 'horizon' (at sea) seems "straight", yes? but joins to itself! This is because you see it from the centre of a circle. That's no problem yet.
Two different great-circles circling around you do cut twice, but *it certainly looks strange*, since looking at the gap between them, you see "parallel straight lines"; Yet to the left **and** right of that you will see them both join!
Great Circles Intersect Twice. **Nowhere does either circle seem to 'curve', yet they manage to meet.**
The Sun 'comes up and goes down' but is actually 'moving' in a nearly straight line—as straight as the horizon…

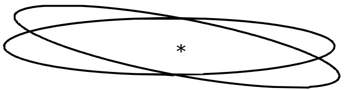

PUT YOUR HEAD INTO A HOOP; GET OUTSIDE; CHECK IT OUT; GET IT INTO YOUR HEAD: GET USED TO IT!
Caution: Don't look directly at the Sun. Watch your shadow and you will see that it moves in a nearly straight line from East to West.
The Sun's average path — the 'Equator' up in the sky — where the 'Saucepan' (Orion; the Pot) moves — is a great circle.
It intersects the horizon due East and due West. At times of year other than at Equinox, the Sun's path is still almost a great circle.
The Milky Way is another maximum sized circle; and the Zodiac (i.e. the Ecliptic) is another such 'Great Circle'.
They hit the horizon twice, in opposite directions (not E/W).
2 other 'great circles' you need to visualise are: Satellite paths; and the Dark Blue Band. (Star-trails are usually 'small' circles, and do look curved.)

Locate Polaris (and Find North) by Imagining the Centre of Circling

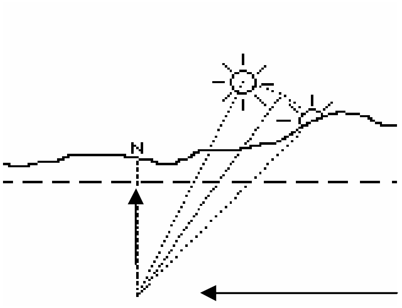

This works for the Sun, Moon, planets or stars.
You *find the centre of their circular paths by bisecting the chord*.
(Northerners locate the South Pole Star, South)

The **perpendicular bisector** of an arc or chord passes through the centre of a circle...

Moonrise varies quickly from day to day
— make sure you check its exact position.

Don't look directly at the Sun to do this!

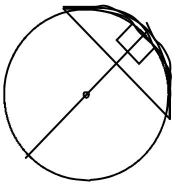

Polaris is, in degrees, at 'your south latitude below horizontal' (not the horizon) — judge it carefully.
and/or, 'polar distance (co-declination) away from the Sun': 90°-N dec, e.g. dec=10°S 90-(-10)=100°

The Eagle Clips the Equator, South of Altair

Find the star 40% along the southwest side of the rhombus…

Finding the Equator helps you

find the Ecliptic (not shown)
and find where the Milky Way crosses the Equator…

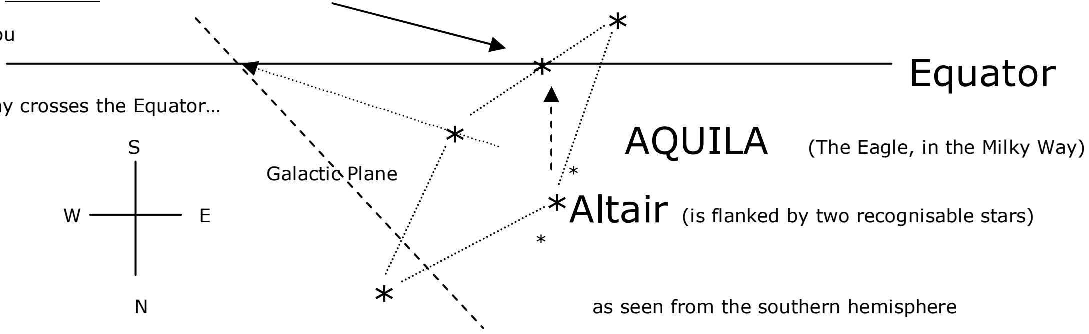
Equator
AQUILA
(The Eagle, in the Milky Way)
Galactic Plane
S
W
E
N
Altair
(is flanked by two recognisable stars)
as seen from the southern hemisphere

Steering Winds

Mid-to-high-level (upper) winds tend to push sea-level pressure-systems along
at **half the speed of the upper winds, and in the same direction**.

This rule is said to work best with small, weak, less well-developed systems; and with hurricanes.

E.g. If the low is on your left And if the upper winds are coming from your left
Then the storm centre might pass straight over you.

Since the upper level winds often oscillate in the 'Rossby Wave', and tend to generate high and low pressure areas beneath them (because the 'wave' causes the upper winds to slow down or speed up) the High-pressure systems will on average be steered slightly toward the Equator, and the Lows will go slightly toward the Poles, as the systems move to the East — a rule of thumb worth remembering.

But this extra steering-wind rule will alert you to non-average situations.

"Rise-Noon-Set-Midnight" carries its meaning across the Equator! N.S.E.W. doesn't.

Name the directions according to the average outside picture…
This is a visual system and addresses the disorientation of turning around.
For back-bearings i.e. when you turn around: replace 'rise' with 'set',
'noon' with 'midnight' (start with end, & hottest with coldest)
e.g. if the old direction was "after setting" the new one will be "after rising".
When you turn right or left adjust by one of the four notches
e.g. 'after midnight' becomes 'after rising' if you turn 'with' the Sun.
Note: 'after the setting' (or 'before midnight') is a direction of the evening Sun
— S.W. — and can be bright sunlight, in midsummer !
And 'after midnight' ('before East') is the direction of the South-East Sun.
At Mid-Summer sunrise it can similarly be very much bright !

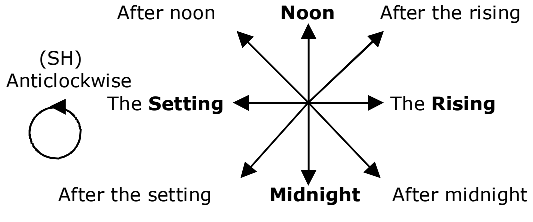

(The Northerner's version needs to go clockwise!)

Example: You enter a hospital with "setting" (i.e. due West) on your right, so you start at "midnight" (SH). You then navigate through look-alike wings, look-alike wards, and look-alike staircases, remaining oriented, and finally emerge, expecting "rise" on your right — having not looked out of a window for hours!

Global Orientation: Half the Arc-Distance below Horizontal

Point out any country or place on Earth, by first finding the (arc-)distance there, and then pointing below the horizon…
The method works **from any place on Earth, in any direction, to any place on Earth**.
Step 1. Measure the angle around the circumference to it. E.g. guess from a globe of the world (This is an angle at the centre of the globe)
Step 2. **Halve** it. That is how far below horizontal it is.
E.g. South Pole from 30° S. Point 30° below since it is 60° away
E.g. Perth to Peking (which is directly North) is 32 S. to 40 N. Therefore the 'arc' to Peking/Beijing is 72°…

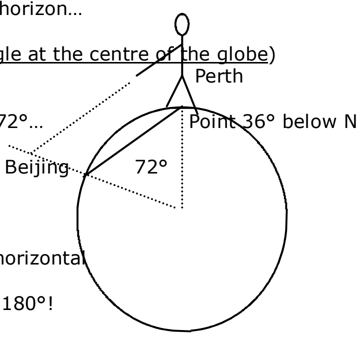

Try this (outside): Go through the motions of pointing out a whole circle of real places on earth, from the horizon ahead of you, to the horizon behind you, using your index finger.
E.g. point North, then down a bit to below horizontal e.g. New Guinea, now onto the North Pole.
Continue down until pointing down underneath you e.g. mid North Atlantic.
At this stage you have pointed through 90°, but have pointed out 180° around the Earth.
Turn around to continue, (Brazil, Antarctica) bringing your finger higher and higher until it is back to horizontal pointing at the hill on the Southern skyline.
Your index finger has pointed out a 360° circle (as measured from the centre of the Earth) by traversing only 180°!
Get Hold of a Globe. Go Outside. Point Out A Few Countries.

" Noon " is Not at 12 O'clock

' High Noon ' is at 'midday' i.e. in the middle, between sunrise & sunset. And that is **not '12 noon' or '12 midday'** but at North — Noon North. (SH)

- In summer-time 'high noon' may be near 1 or 1:15 p.m. or so. If you are on Summer Time, true noon is one hour late.

- Even in winter-time it may be over half an hour away from 1200 hrs. This is because we use a one-hour time zone to cover 15° of longitude.
  i.e. you may be near the next time-zone, longitude-wise, and therefore be up to one half an hour 'out'.
  You can check this up if you have a map with longitude on it. The middle of the time zone is at 7½°, 22½°, etc, in steps of 15°.
  Then it will take 4 minutes to travel an extra degree westwards. Work out how early or late the Sun will be in your longitude.

- Plus; the Sun speeds up in its 'orbit' 'around' the Earth when it is close in the Southern Summer, getting up to 20 minutes 'fast', by the end of summer; likewise it can be up to 20 minutes 'slow' at other times, in addition to the time zone effect. See the Equation of Time prediction opposite.

The best thing to do about this, to predict when, exactly, it will be at 'midday' today, is to look at the Sun yesterday.
(Don't! Use a shadow instead. Notice when it is due South.) Or see 26a.

Equatorial Stars Rise and Set Due East and West

Always

Globally

- The Sun at Equinox — about March 21st, September 22nd
- The Pot in Orion — Mintaka, the north-west star of Orion's belt — the one that sets first — is right on the Equator
- Various dull stars — like the one south of Altair
- A Half Moon at solstice — give or take 5° or so — It must be 90° from the Sun which is at the solstice position, so that the Moon is at the ecliptic equinox (Equatorial) position.
  Other phases can be worked out, e.g. Full Moon at Equinox, but see p. 29d on how to predict an equinoxial Moon every fortnight.
- Planets 90° from the Sun-at-solstice
- The equinoctial points, if you can identify them yet. One is between Spica and Regulus; the other is between Diphda and Markab.

The Moon Cusps Push and Pull the Hour Hand **Backwards**, Down South

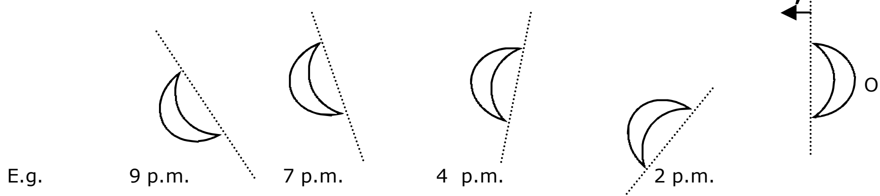

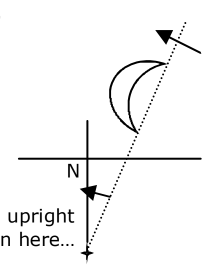

Watch the Moon rotate during the day, and night. Tell the time from it. Update yourself once per day.
The Moon rotates like this about 14½ degrees or 15° per hour. So it is like a 24-hour clock hour-hand running "backwards".
The correct vertical to use as a reference is not at the Moon, but at the Pole Star, if you want to be accurate.

(NH — the Moon seems to turn clockwise)

Back-Bearings Killed Sleepy Pilots

Back-bearings are opposite directions — "going back home".
The Babylonian system of 360° in a circle has a lot to answer for, when you want to turn around!

What is required, is to add or subtract 180° (groan)!

But after an all-night mission, tired navigators can't be relied upon to come up with the right answer.
In World War II, many air-crews perished from following the wrong back-bearing back.
Happy, relaxed, but lost.
It was a problem requiring a solution, and a simple solution was found. What is it? (page 18e)

Half the Arc-Distance below Horizontal — 2 Explanations Why

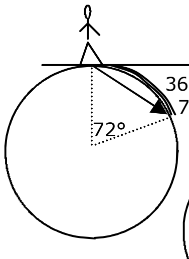

1a.
A horizontal tangent
36° angle below horizontal
72° 'arc' around the circumference — twice the angle from the tangent — but why?

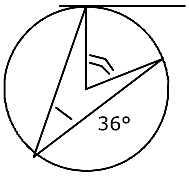

1b.
The angle at the centre
(used for 'arc-distance')
is always double the angle
at the circumference...

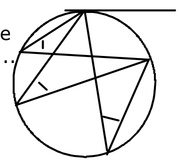

1c.
..It doesn't matter which point
on the circle you use,
as long as it is on the one side
—on the larger arc here—
you get the same angle each time.
The tangent point is one such acceptable point — the last one.

2. (Isosceles triangle angles... (180 – 2α)/2)
= 90 – a

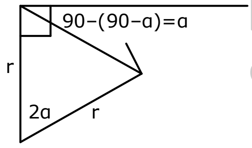

1d.

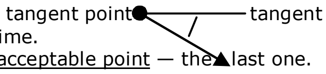

The "Equation of Time" Diagram

The real Sun moves quicker through the sky near Christmas because it is closer to the Earth (i.e. we orbit it quicker, elliptically!).
I.e. The Sun 'gets ahead of' its predicted average position, according to our clocks, in the New Year,
And doesn't come back to average until about the end of August.
After that, during September to December, it will be behind average. But then, in addition, the details are complicated by how it also zips quickly past (gaining on) our longitude lines near solstice — summer or winter — when it is near +/–23½° latitude.
Here the longitude lines crowd together, if only slightly, giving the maximum effect in June & December, super-imposing upon the preceding scheme.
The **overall result** may be graphed...
See the mnemonic on p. 86b

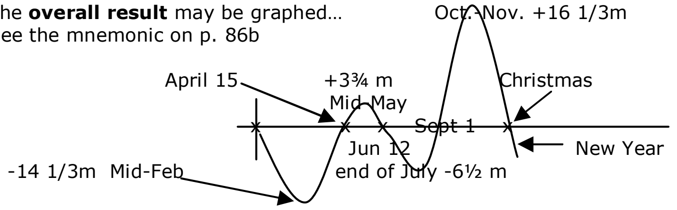

(both hemispheres)

Note that when the astronomers say '+', the Sun is behind itself; we look at the position of the Sun (not directly at it!); it comes up 'late' and we have to add (about a ¼hr) to our guess at the apparent time, to find the clock-time. We say our clocks are 'fast' at predicting the Sun's position, but it is really the Sun which is lagging!
cf Australia is in a "**+**10 hr time-zone" = 10hrs ahead, or fast on Greenwich

---

Rigel-to-Capella is a good S./N. Pair

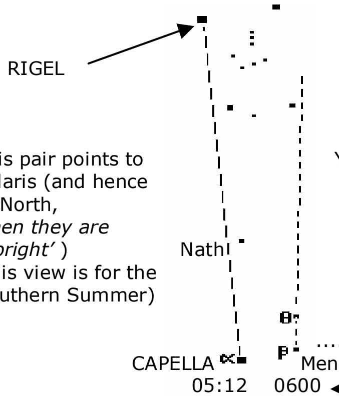

This pair points to Polaris (and hence to North, *when they are 'Upright'*) (this view is for the Southern Summer)

You can also see that the Pot handle points down through the middle of Orion's belt to Meissa and on to Polaris. The Pot 'handle' is actually the scabbard of Orion's sword—seen upside down (SH)

If you cannot see Capella, then Bellatrix to Elnath, right alongside, is also a good guide to Polaris.

Auriga, The Charioteer
....**The six o'clock stars are a useful N./S. pair** — they are upright at 6 a.m. local star time.
← 'Right Ascension', or star-time, in hours of longitude East of the first point of Aries

---

Atmospheric Pressure Changes

The air pressure falls before a front, and rises reasonably rapidly as the front passes. It falls in a trough; rises in a ridge; falls in a heatwave inland; falls with altitude (about one millibar per 10m).
The atmospheric pressure normally fluctuates by a couple of millibars every day in a twice-daily cycle tied to the Sun (3mb in the tropics) and the wind speed may vary with it. The daily pressure on average peaks around 10 a.m. and 10 p.m., being lowest at 4 a.m. and 4 p.m.; This on average, is said to vary the wind speed by a couple of knots, lowest at 9 a.m., peaking at 3 p.m., but all of this is usually masked by other weather changes in the temperate latitudes. Pressure fluctuation is much more significant in the tropics, with any fall (or flattening) more than a few millibars during summer/autumn warning you of a possible hurricane (You then look for long wavelength ocean swells preceding the storm centre). Because we can't feel atmospheric pressure changes, it is very helpful to use a barometer, to detect the longer-term pressure change trend. Test your barometer to see that it does not respond to temperature changes!

Watch the Clouds Cross the Moon

You can detect the upper wind direction most easily by watching clouds move past some fixed point in the sky.
Look during the night but don't forget to look during the day also.
The stars, high buildings, high trees and mountain tops also work well.

The **0, 1, 2, 3, 4** Reference Directions

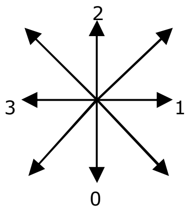

Put these numbers onto the Rise-Noon-Set-Midnight System, so as to follow the Sun.
0 = midnight, South (Southern Hemisphere)
1 = rising, East
2 = noon, North (use South for northern hemisphere)
3 = setting, West
4 = midnight, South — use cycle 4 i.e. call it "0"

This corresponds to measuring anticlockwise! from South! in decimal degrees, that is, in grads!
100 grads = 90° so "1" = 100 grads; "4" = 400 grads or zero grads — add or subtract "4" at any time.
"0.1" in this system = 10 grads; 1/10 of a right angle; **one fist at arm's length**.
Example: In the thick bush, you see your destination but know you will soon lose sight of it. You can't keep a straight line, so you decide to keep track with the 0—1—2—3—4 system and your fist. With each change of direction you update e.g. "0.6, 1.6, 1.0, 0.9..."
For a **back bearing** — when you turn around — **add or subtract 2**.
Add 1.0 when you **turn with** the Sun (**left**), or **subtract 1.0** when you **turn against** the Sun (**right**). (S.H.)

Example: Point to Tahiti

It's 9 a.m. summer-time, the Sun is, say, 20° high and over Tahiti (see time-zone chart).
Firstly, point horizontally toward the Sun (not directly at the Sun!) — that is the direct path to Tahiti.
Secondly, point below the horizon, to locate the actual direction to Tahiti…
Step1. Make a perpendicular from a sunbeam. Perpendicular to the Sun is 'horizontal' in Tahiti…
Step2. Subtract its (present) height (20°) from 90°
The arc from you to Tahiti is 70°
Explanation:
(90° - the height of the Sun), the 'co-height', must be the arc-distance
from you to underneath the Sun. Think of it like this...
If the Sun is 0° away from you, it will be 90° high — you are underneath it.
If the Sun is 10° away from you, it will be 80° high, overhead a point 10° away from you, and so on.
Step3. halve the remainder (70°/2=35°) Now add back on the 20° above the horizon…
"Tahiti is 55° below the Sun" from your point of view. You didn't need a globe or the arc-distance to find it.

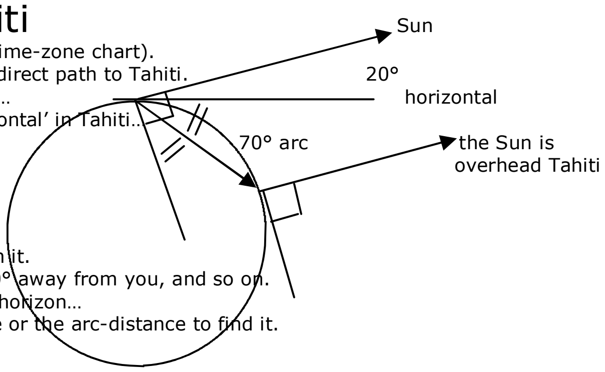

The Sun Oscillates Seasonally between +/- 23½° Latitude

To approximate the Sun's "Declination"—or North Latitude—you approximate its variation as sinusoidal.
Using a 365 day cycle: You approximate 360° as 1° per day, so for ease and accuracy of calculation you can and should approximate the day-number as 30 days per completed month + day of the month.
E.g. **March 21st** is 2 lots of 30 days plus 21 = "81" days or degrees.

Then to use a sinusoidal curve, you need to run from Equinox to Equinox — i.e. starting at March 21, Day 81

Declination = 23.45° x sin (day number of the year – 81)

E.g. 0° in **September 21**, six months or 180 days or degrees later; uses day no. 261.
From solstice to solstice you can use

Declination = 23.45° x –cos (day number of the year + 9)

..since the solstice is around 10 days earlier than the New Year.
Dec 21 solstice, is day no. 11x30 + 21 = 351 "–23°½ in December"
E.g. + 23½ degrees on **June 21**, day no. 171.
From this approximation you can predict noon-height (and, with the right formulae, sunrise & sunset time & direction and E./W. height, of the Sun).

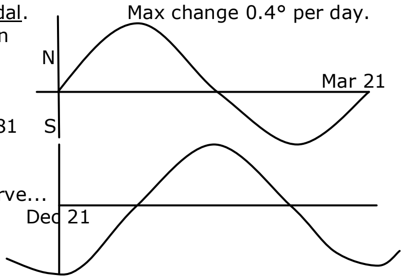

Use the Southern Cross as an Hour Hand

You can visualise the rotation of the Southern Cross as a huge 24-hour clock hand.
It is correct on April Fool's Day i.e. upright at midnight (at 183°T), then it gains two hours a month, or 30°. (Use the correct vertical! — South.)
As a learning aid, mark the 'Southern Cross time' on your calendar every month
e.g. "two hours fast" at the start of May; and "four hours fast" at the end.
Then, anytime you are out at night, you can ask: **"What's the time — by the Cross?"**

Picture the Southern Cross during the Day

Yes it is there behind the blue sky… "If there were an eclipse now, it would show… there!"
When you are navigating by the 'spin' of the Earth, and by the Big Picture, the proper reference direction is the North-South rotation axis,
Which means: Polaris + South Pole Star.
In practice, what we Southerners see, is the Southern Cross.
So if you are imagining: sunrise — noon — set, or Moon position, or a star-compass course, or a shadow-south direction,
you ought to be thinking: "North — Polaris down there; South — South Pole Star up there", then you can picture the Southern Cross as well.

The Moon is Lit By the Sun

When the Moon is **opposite** the Sun, we must see its fully lit face… See diagram of the first man looking at the Moon, below.
e.g. a full Moon rising at Sunset; Setting at Sunrise; Overhead at Midnight or at least highest in the sky.
(A fortnight later…)
When the Moon is **near** the Sun, we see only its dark backside or at most a thin crescent… See the middle man below.
e.g. a crescent Moon is only ever seen near sunrise or sunset. Wait until the Sun is below the horizon — Don't look at the Sun!
At other times the Sun has set, taking the Moon with it, or has risen, outshining the pale crescent nearby.
(A week later…) When the Moon is **halfway** between, we see half of each side… See the third figure...
e.g. half a disk overhead — only at sunrise or sunset; half a disk setting — only at midday or midnight.
A half full Moon is misleadingly called a Quarter Moon… since it takes four weeks to move between the four "quarters".

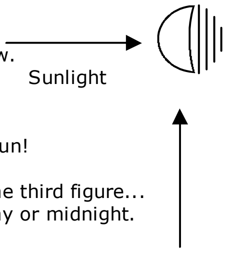

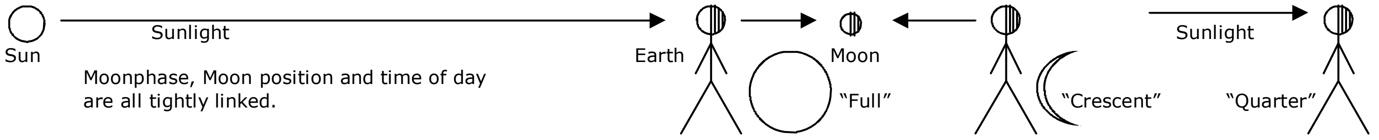

Moonphase, Moon position and time of day are all tightly linked.

Back-Bearing: **Plus or Minus 200° Then Minus or Plus 20°**

(Or — for the very advanced sleepy student: adjust by 20° first (+/-20°) then -/+200°)

Examples: 10° is small, so make it big, 210°, adjust back to 190°
290° is big, so make it small, 090°, then back up to 110°
Advanced: 190° is middling, so boost it first, by 20°, to 210°, then remove 200° and use 10°
210° is middling, so try +20°, then –200° = 030°
When you end up with numbers greater than 360°, remove 360°. When less than zero, add 360°.
Difficult: -10° is small; first convert it, using +360° (= 350°)
350° is big so –200° +20° = 170°
185° + 200° -20° = 365°, so adjust it back to 005°

There is another solution to back-bearings — decimal degrees — but it hasn't taken on yet.

a simpler solution
with a protractor
compass, is to...

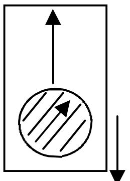

Go this way.

(see page 109b)

Use **Latitude** for a **Distance** Scale (Not 'Long'-itude, Despite the Name)

(Especially on marine charts, atlases and old maps—Newer maps are likely to have the mostly meaningless but well-scaled international map grid.)
Latitude means "distance" from the Equator. Remember the polar "distance" 90° complement to latitude? which measured arc-distance, in degrees!
1° of latitude = 60 nautical miles = 111 km The world is 360° around; or 360x60 = 21,600 nautical miles; or 40,000 km
The circumference of the Earth determines how long a kilometre or a nautical mile is — at least in the somewhat outdated original definitions.
(The N/S circumference of the Earth has proven to be closer to 40,007 km, and the E/W circumference is bloated by the Equatorial bulge.)
One 'kilometre', comes from the decimal system: whereby one circle equals 400 'grads'; one grad has 100 divisions… 40,000 km
Use the latitude scale on the left & right sides of a Mercator map to measure length or distance — but only at that latitude on the map.
But do not use the longitude scales running across the top and bottom of the map — since the spacing of longitude lines should vary with latitude.
Longitude lines must converge towards the Poles, to nothing. They span a variable distance = 1° x cos (latitude), but a Mercator map will (falsely) show a constant separation for longitude lines (this is to make compass courses look straight). A Mercator projection will also show a variable separation for latitude parallels, even though they 'should' be constant. This is to preserve the shape of land areas—The latitude stretching must match the longitude stretching. Therefore any scaling shown on the E/W spacing is only valid for the latitude at which it is subdivided.
When a latitude scale says ½° or 30 n ml or 55 km, it means it, at that latitude, but when a longitude scale says ½°, the distance meaning is missing

Using the Declination of the Sun

- **Noon Height** = 90° – latitude + declination = (from horizontal) Convert "N." to + ; "S." to – e.g. 30° S. = –30
  E.g. Lat 30° S.; Decl. 20° N : Height = 90+30+20=140° or 40° (the two (opposite) heights must add up to 180°)
  i.e. The more the Sun 'declines' away from the Equator, and the more you decline on the other side of the Equator, the more the Sun drops in the sky. Alternatively, the colatitude (90–L) is the maximum height of the sky Equator... Then adjust for the declination.
- **Sunrise/Sunset** — time and direction prediction
  Cosine (angle from North/South) = sine (declination) / cosine (latitude) Try dec =23°.5; lat =37°.1; 55°.67; 62°.6, for a different hexagon
  Cosine (time in degrees from noon) = –Tan (Lat.) x Tan (dec.) 1° of longitude = 4 minutes of time.
  E.g. measure the time between sunrise & set; divide by 2; solve for Lat. (make sure to use horizontal, not horizon.)
- **East/West Sun** — Time and height prediction
  Cosine (E-W time in degrees, from noon) = Tan (dec.) / Tan (Lat.) e.g. 4 hrs each side of noon @ Lat 41° (Northern Tasmanian Summer)
  Sine (E-W height above horizontal) = sin (dec.) / sin (Lat.) e.g. 45° high E./W. from Lat 34 1/3°; +/– 3h21m; Mid-Summer

e.g. take the time when the Sun is E. & W. (opposites — you need to remember to carry 60, not 10, & to interpolate carefully) and then solve for Lat.
See p. 86b for a memorable laugh. Many more formulae and techniques are available, from books on astro-navigation.

Perspective Distortion — Use the Correct Vertical

Vertical is vertical, right? Wrong! To tell the time by the sky, take care.
The higher a pair of points (e.g. stars) gets in the sky, the more you have to choose which vertical is needed for a reference, to judge 'vertical' by.
Each different reference direction will give a different answer to "How vertical are those two points up there?" e.g. Moon cusps, Southern Cross
A side view makes this clearer....
..a straight line which looks horizontal when looking south,
looks diagonal in the East

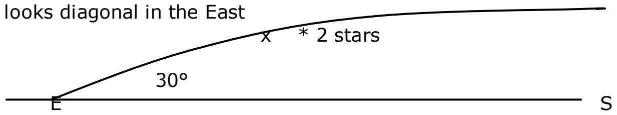

The top view also makes it very clear… a line XY high up in the sky cuts each different vertical at a different angle..

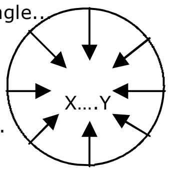

Don't gauge the angle of rising/setting (p9a) at anywhere other than at the rise/set horizon in the rise/set direction.
Don't use the nearest horizon, as your reference for which way is up.
To visualise a constant 'rotation' of the stars, use Polaris or the South Pole Star as the 'centre' of rotation.
I.e. Use N-S longitude lines in the sky rotating in relation to vertical at North/South.

To Determine Upper Wind Direction from the Clouds

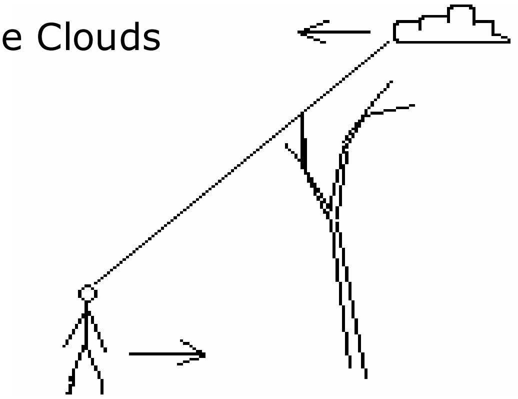

Walk in the reverse direction, while keeping a chosen cloud lined up
on a chosen high point. You must walk across level ground to be accurate.
Mark your starting point, then drag a line in the ground to mark your walking direction.
This method works well for very slowly moving clouds.

Other methods: Watch cloud shadows.
Use the Moon, Venus or the Sun (only when it is very well obscured).
Look in different directions at the moving clouds
until their motion seems neutralised—they are moving directly toward you.
Use binoculars, to magnify the movement (but never look towards the Sun).

The Sunny Skies Quadrant System

To orient yourself on a clear day...
Use the two ends of the dark blue band — where it hits the horizon — and the trend of the middle bit of the band,
in conjunction with the Sun (but don't look directly at the Sun!) and your shadow.

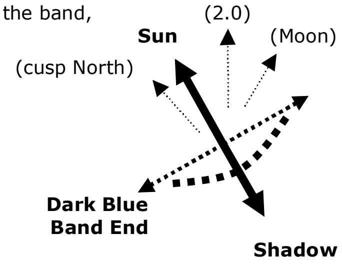

- No matter which way you are facing, you can detect one or two signs to orient yourself by,
  Or at least imagine them if you can't see the Sun just now.
  You don't have to look over your shoulder, which is misleading anyway.

The middle of the blue band is above your shadow, and at right angles to the sunbeams near you.
The whole system rotates, during the day.
If the **Moon** is up, add it into the system,
and **Moon-North** too, which is off to one side of North, and changes slowly during the day.
Combine this system with the fixed systems, of 0, 1, 2, 3, 4 and m.n., rise, a.m., m.d., p.m., set.
E.g. In the diagram, the Sun is to the left of North, so "the Sun is 'afternoon', the Moon is 'before noon'."

Earth
Sun
Stars
Wind
Anticlockwise

The Horizon Rotates to Your Right in the Southern Hemisphere

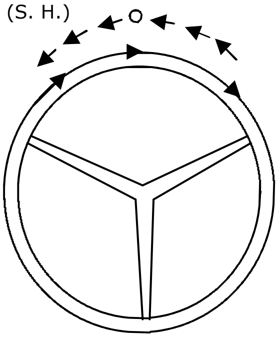

1.Face the High Noon Sun. (Imagine it, if you have to, when it is not midday.) It moves horizontally from Right to Left (S. H.)
2.Reverse your perspective — let the Sun stand still and the horizon move, from Left to Right.
Imagine the horizon sliding horizontally beneath the Sun (and you shifting sideways with it).
3.Now imagine the Sun rising from the eastern horizon — it moves up to your left — then reverse perspective...
Imagine the eastern horizon moving *diagonally down and away* from the Sun, twisting down on the right as it goes.
Face North. Grip the whole horizon on the East and on the West, and imagine a tilted steering-wheel, and rotate it around the Pole Stars. Imagine a "steer right" motion.
The invisible steering wheel faces you diagonally upwards, while the horizon, as if fixed (diagonally) to the steering wheel, starts off horizontal and then rotates out of its horizontal.
4. Keep rotating the horizon, sliding the northern horizon horizontally underneath the high noon Sun, and beyond, until the western horizon comes up diagonally to meet your now "setting" Sun.
You need to do this outside to appreciate it in three dimensions, to have any chance of visualising it easily.
So put your clock back on the Northern wall (see 5e) and say "See, we Southerners have it right, after all!
— the real movement — of the solid stuff — is clockwise"

Tell the Time by the Compass Protractor — Tilt It Toward the Equator

Take a **tilted bearing** of the Sun. The answer comes back as so many degrees to one side of noon-North (or noon-South in the N. H.).
You then use 15° per hour to convert that to hours before or after noon-time ('noon' will be nearer one o'clock during daylight savings time).

Step one: align the compass to True **North**
Step two: **tilt** it upward to the 'Equator' in the sky (the Sun's average path)
Step three: rotate the compass arrow to point toward the Sun
Step four: make the edge of the compass cast no shadow
This avoids looking at the Sun
Step five: measure the degrees away from high-noon North, on a tilt

It is more accurate than using a finger-span
to measure 15° per hour
i.e. Use the compass protractor to make a 24-hour clock face
To use the compass protractor as a 12-hr clockface, see the box opposite.

1. Top View 2. Side View

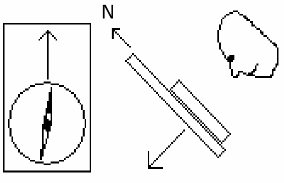

The reason for tilting is because a level sundial/upright gnomon will not give a 15°/hr linear reading (See "The sun passes quickly")

The Southwest Corner of the Great Square is N. S. E. W.

The Great Square of Pegasus only has one really good right angle but it shows N. S. E. W. perfectly.
In the Southern Hemisphere, it is the first star of the Square to rise.

Note: When you hold a star chart over your head
'East and West' reverse!

This is the view for the Southern Hemisphere…

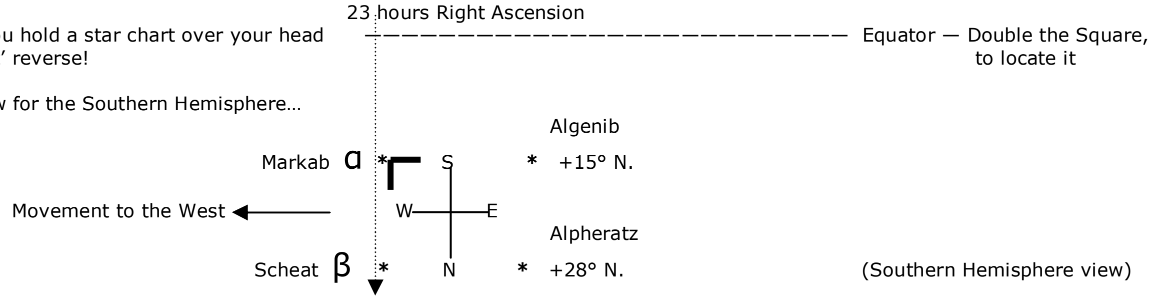

The Moon Rises and Sets 50 Minutes Later Each Night

The Moon moves quickly to the West, but…
It progresses Eastward through the night skies, night by night, by 12½ degrees a day — it 'lags' the Sun or 'moves backwards' during the month.
So that after 29½ days it is back in the same place. It moves about 27 diameters' worth East through the stars, per 24 hours
or about a diameter an hour to the east in relation to the Sun. One 'diameter' is half a degree, and the same as the Sun's.
Sun, Moon, Planets and Stars rise in the Eastern half and set in the Western half, but some are quicker than others. The Moon is the slowest of all.
E.g. (twice-daily) tides get 25 minutes later each tide, on average — half a day + 25 mins apart.
E.g. the 12 month lunar year gains 11 days on the 12-calendar-month year, or "one day per month except for February, on average".
(a calendar month equals 30.4 days; a lunar year equals 354 days. Months 1,3,5,7,8,10,12 have 31 days—1½ extra; ½ extra in 30 day months)
The explanation for the Moon's "slowness" in rotating westwards with all the rest of the sky, i.e. from moonrise to moonset
(it rotates only 14½ degrees per hour, not 15° like the Sun) is that it slowly revolves around the Earth to the East,
in the same clock-sense as the Earth revolves around the Sun,
and so must complete 13 revolutions star-wise, to the East while we see only 12 Moon months.
By the next full Moon, its backside is toward a more easterly star…
For an anti-clockwise rotation system, viewed from the North...

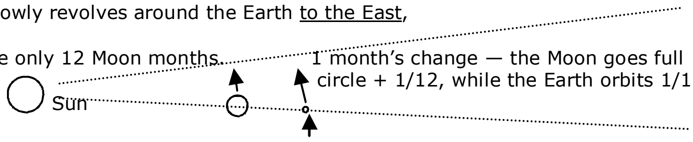

1 month's change — the Moon goes full circle + 1/12, while the Earth orbits 1/12.

Turn 90°? "+/–100° –/+10°" !

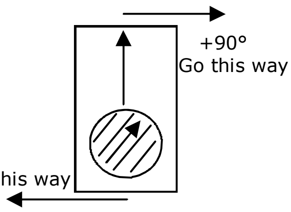

Turning right is turning clockwise; turning left is turning anti clockwise.
Clockwise **increases** the count of degrees East of North. So that requires **+100°** -10°. E.g. 257° changes into 347°
Anticlockwise, when turning left needs **–100°** +10°. E.g. 257° changes into 167°
Or use +/-10° first, then -/+100° if that is easier (near 0°). Adjust by 360° if necessary.
There is no convenient choice as there was with a back-bearing, so you must remember whether to use + or – 100°.
Remember to look at the Sun's clock-sense to judge 'anti-clockwise'
— anticlockwise in the southern hemisphere — to your left. 'With' the Sun implies – 90°
— clockwise — to your right — ('With' the northern hemisphere Sun) implies + 90°.

Hint: When sailing, you can always temporarily head either towards or away from, or side-on to, on the left or the right, a landmark from which you want to take a bearing, without having to change tack. Therefore you can use +/–180° or +/–90° from the boat's compass-heading. In heavy weather this means that you do not need to leave the helm to use a hand-bearing compass, but merely sight along the lines of the boat, or cross-wise along the lines of the cockpit, stern or cabin. Simply remember that the more clockwise heading is bigger than the left-hand direction

Global Perspective — Rotation of the Horizon

Each point on the horizon rotates…

- In a circle
- Around in an Eastwardly sense, like all the real matter in our anticlockwise spiral galaxy
- Around the North-South Earth rotation axis
- Along a fixed sky-latitude circle (i.e. declination circle)
- At a fixed angle of rising/setting (one or the other) in relation to the fixed stars
- Once per day

Example: The "Sugarloafs" were on the left of the Sun at sunrise, but six hours later they had moved to be underneath it, by rotating around Polaris and now, at the end of the day, they have moved way over to its right, and will keep going around for the next 12 hours of night-time.

To further help you visualise the motion, remember that the ground under your feet is twisting you clockwise in relation to the sky, in the SH, so that the stars, etc, near the horizon all move only to the left as they rise or fall. For the Northern Hemisphere use "anti-clockwise… right".

Or Judge the Time By Sight — by Mirroring the Sun Around North (To Make an Hour Hand on a 12 Hr Clockface)

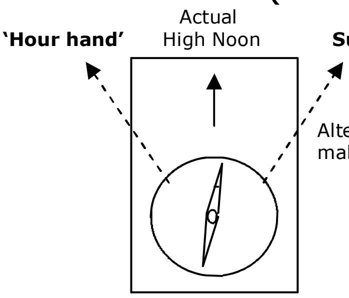

Sun = **'Noon'** position on the Clockface (i.e. ignore real Noon-North after you align the compass properly) (SH)
(Here, the 'hour hand' is showing 2½ hours until 'noon', judging it by sight)
(For the Northern Hemisphere, you have to reverse things… hour hand — Sun — South see p.28b)
Of course this method is really just helping you to GUESS!

Alternative: Line up a pencil tip with Polaris, while making the pencil cast no shadow. With your shadow-side hand, make your index finger ("The Hour Hand") touch the pencil tip, and look identical on the other side of Polaris.
You can make it work on the Moon or a star or a planet if you know when they are North/South of you.
You can make it work for magnetic noon, to avoid using magnetic variation.
It works because the Sun moves at half the speed of an hour hand,
but the doubled-angle shown, doubles that rate of change, to match a clock's hour-hand.
Don't forget to **tilt the compass**, so that its axis lines up with the Earth axis.

The Southern Summer Rectangle is Worth Looking for — since it has four N. S. E. W. guides

1. North: from Procyon, down through the Twins
2. North: from Bellatrix down through Nath
3. West to East: between Bellatrix and Betelgeuse to between Procyon and Gomeissa
4. West to East: Nath to between Pollux and Castor, closer to Pollux

(To remember the two Twins, say "Castor & Pollux, α & β, are North & South")

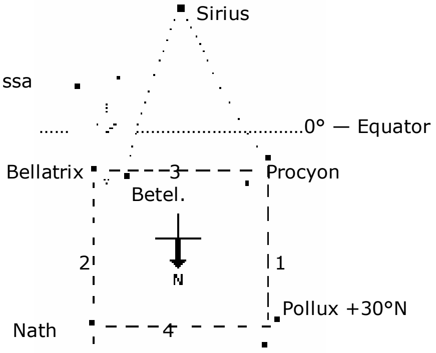

And a Triangle Pointing South

Betelgeuse, Procyon and Sirius.

Clouds and Con-trails and Their Winds

- Fog, which is cloud clinging to the ground, usually oozes downhill. Now, working upwards...
- Topographic cloud banks — may indicate e.g. the plateau edge or a mountain range. I.e. They cling to the topography.
  Also look for eagles soaring along otherwise invisible cliff edges.
- Low clouds — they follow the weather map isobars. Also look for sideslip of flying birds and aeroplanes.
  As a rule of thumb (clenched fist!) low cloud puffs are at least three knuckles wide, alto puffs are two, and cirro puffs are one, from sea level
- Middle ('alto') cloud — the puffs look smaller, the shading, less dense, but no 'iciness' seems apparent. The clouds obscure the Sun's outline somewhat. Their wind-direction is usually different from both higher and lower clouds. This wind will 'steer' low level pressure systems.
  (An 850 mb level chart helps) When a change is due, the middle clouds often betray what low wind direction to expect soon.
- High ('Cirrus') cloud — thinner; smaller puffs if puffy; icier i.e. fleecier, silky, filamentous; transparent, with halos & no shading.
  The Sun's outline remains crisp when seen through cirrus. Don't focus on the Sun!
  The cirrus-level cloud-direction in the subtropics is usually westerly +/−45°.
  Be sure to test this in your latitude. (A 500 mb chart helps) Try to catch glimpses of upper clouds through gaps in the low clouds.
- Jet Con(-densation) Trails — are usually in the high band. Look for them according to a timetable, each day.

The "SWEETHEART" Method of Orientation

We are human, and 'reference directions' are not. So make them more human, more personal, more emotional.
"Home" is not bad. "Sweetheart" is better.
Find something lovely about each cardinal direction. E.g. 'our favourite holiday spot on the coast'
Something you love or hate, want, or fear, about each reference direction…

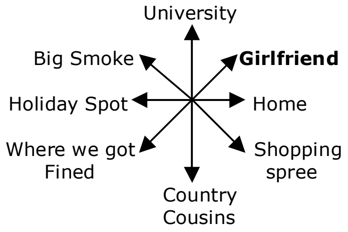

You will find it easier to maintain awareness of bends
and to interpret meaning onto the skyline — what lies beyond and behind that horizon!

During one walk, these highlights are unlikely to change direction much.

How to Shift Your Horizon — Its Timing and Height in relation to the sky.

You can navigate by the "shape" of the sky, back to a spot where you have been before, by noticing the "position" of the moving horizon.
On a round earth you can shift your position and horizon 1°, or 360°, by travelling 1°, or 360°, around the circumference. (111 km, or 40,000 km)
The line where the horizon cuts across the sky, varies with the latitude & longitude & the time of day. Using the Sun as an example…
Note: The "1° for 1°" rule for your own movement, only applies after you neutralise the other motions, so as to be as if the sky is standing still.
E.g. You need to neutralise the Sun's daily motions, by looking at the same time of day — sunrise, noon, 6 a.m., or 12:22 a.m., or whatever.
E.g. You also need to neutralise the Sun's seasonal motion by looking at the same date every year — Christmas morning, June 25, or whatever.
(Leap years can only make a difference of up to 1° — you could improve accuracy by looking only every four years.)
Rule one: for West/East travel, remember that events like Moonrise seem to happen **later to an observer further west**, and…
Earlier in the **E**ast; **four minutes of time (early) for every 1° of longitude (East).** (these rules work in both hemispheres)
E.g. "Christmas Day Sunrise at camp 4 was at 6 a.m. last year, but now I see the Sun already up at 6 a.m., so to get to where we spent Christmas last year, we would have to go further west to make the Sun lower (the horizon higher) at 6 a.m." (Assume you are at the correct latitude)
Rule two: The horizon drops to reveal new sky as you travel onwards — **1° for every 1° of travel** — this works in any direction.
E.g. "The Sun rose to 75° high, but here it reaches only 74½ degrees — so I would need to travel a half-degree toward the Equator for it to be 75° again." (This example has deliberately chosen a North South change.)
Then you can shift your horizon, its timing and its height, by moving around the earth, in order to match a previous timed snapshot of the sky.

Walking in Phase with the Sun, Wind and Moon

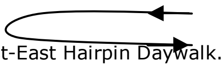

- In a hot climate, walking into the **Sun** is unpleasant, hot and glary, especially in the afternoon, especially if you wear glasses, so... Simple West-East Hairpin Daywalk.
- The morning **wind** can be very uncomfortably cold, and it is best to plan to walk away from it, too.
- Beach walks are more pleasant when you don't have to battle back into the afternoon sea breeze or prevailing wind; so assess the **forecast** before you choose a start point.
- Night walks and late camps are best done in bright **moonlight** — simply predict the phase.
- Spotlighting is the easiest on a waning Moon — since the **dark period** is in the evening, not before dawn.
- Spectacular pink full-Moons only rise at sunset, so it is best to predict such a **moonrise** and plan an early camp with a view.

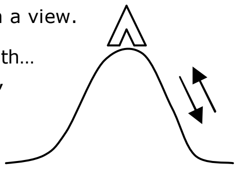

E.g. Up in the afternoon **shade**; down in the morning warmth…
Similar considerations apply to sailing—e.g. aim to make landfall in the pre-dawn. You can identify the harbour, or the coast, by its light-houses, then soon after be able to enter the harbour in daylight.
E.g. Aim never to cross the sandbar on an outgoing tide—with steep breaking waves, in shallow water, getting shallower!
Coordinate Your Route, Pace, Campsites & Timetable To the Prevailing Conditions.

Navigating by the Shape of the Night Sky

Since the star-map — the background to our view of the sky — is unchanging, the 'shape' of the sky involves only the **position of the horizon** across the background of fixed stars, and the **rotation** of that horizon with time.
The sky gives a false impression that the sky itself is spinning, but it is we who are turning upside down! It is our horizon which is changing.

Just as you can neutralise the Sun's motion by looking at the *same time every day* (which will show up only its seasonal and longer term variations) and just as you could neutralise the Moon's motion, by looking at the same phase and Moontime every month (e.g. full moon rise), to show up its unpredictable motion, so too with the stars. You must **look at the same star-time every night**. If, in addition, you look at the stars from the same spot you will notice no change at all in the 'shape of the night sky' or the position of the horizon — star motion is all predictable, with no significant variations.
This makes the stars excellent for navigation — you simply locate the horizon in terms of the heights of the stars, the tilts of star groups, and the star-time of night. E.g. what star appears overhead and at what star time.
Thus, to find that spot again, even much later, you need to make your horizon look identical at the identical star time.
CAUTION: Don't go star-navigating; rather only imagine doing so, until you become an expert. Understanding it is more useful than doing it!
All figures, formulae and theories should be checked before you trust your safety to them.

The Moon is "in Phase" with the Sun — Watch for It

The Moon progresses (12½ degrees per day) contrary to the general rotation, so…
A **new** Moon, first seen after sunset, must grow from a crescent, in the **after sunset** part of the sky, as it separates from the Sun, Eastwards, night by night, more and more so after the new Moon. I.e. a 'waxing' (growing) Moon means Moonlit evenings. After full, the Moon continues Eastwards…
A **full** Moon must diminish, to a crescent, in the **before sunrise** time of night. I.e. a 'waning' Moon leaves the evenings dark, until Moonrise, predawn.
In other words, the predawns are moonlit for the second half of the Moon month, after Moonrise. In the first half, the predawns are dark, after Moonset.
The whole night is dark in the 'no Moon' changeover of months.
During the day, you ask "Is the Moon ahead of or behind the Sun?" I.e. Is the Moon predawn (waning) or post-sunset (waxing)?
Evening spotlighting — is best in the dark before the waning Moon rises — late in the Moon month.
Night-walking—is best in Moonlight after Sunset i.e. on a full Moon, or a waxing, almost full Moon. (early-mid month)
(Think of the Moon as prouder of its nighttime role as it moves away, and ashamed as the Sun gets closer)
Highest and lowest ('Spring') tides occur on or soon after a new or full Moon.
'Neap' tides occur (when the Sun and Moon pull at right-angles to each other) at "half full Moon".
(Shallow waters introduce a lag of a day or three, due to friction and resonance effects.)

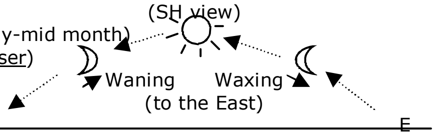

Make it your aim to see these things all happen in phase with the Moon month.

The Clock-Face Method for Direction

Quicker than 360° — it only counts up to 12! — it communicates relative directions easily.
E.g. "3 o'clock high" = "on our right — up in the sky" "9:30 from Rigel" = to the left and just above (that star)
Or, for horizontal directions, " 5 o'clock from the Sun " = looking down-Sun, and slightly to the left "six o'clock true" = South
At 30° per "hour" it is easily compatible with 360° — both are Babylonian systems.
(But don't confuse it with a 24-hour clock face, which is only 15° per hour. Unfortunately our clocks do not use a 24-hour system, which would make them nicely travel around once per day — as fast as the Sun.)
For shifting between clock face directions and 360° directions…
— First convert to or from the hour: Multiply hours by 30, for degrees. Use 12hrs=0 : 0 30 60 **90**; 120 150 **180**; 210 240 **270**; 300 330 **360**
— "5° extra" = "add 10 minutes" e.g. 255° = 0830 hrs (240° = eight hours; 15° = +10+10+10)
— "1° extra" = "add 2 minutes" e.g. 258° = 0836 hrs
When using the clock-face for compass directions, for NE, SE, SW, NW use 1:30, 4:30, 7:30, 10:30.
Translate 1, 2 o'clock as "(roughly) NNE, ENE," etc.
Don't use the minute marks on the clock-face because every 6° is not very user-friendly

Latitude without Numbers

When all you want is the latitude, the shape of the sky can tell you, without numbers, without calculations, without a watch.
E.g. ‘Canopus overhead’, or ‘the Pot sets at 30° off vertical’.

If all you want is to recognise appearances locally, you don’t need to set your Star watch to Greenwich Time, or to know a star’s longitude relative to Greenwich.
E.g. ‘I set my watch to the Dog Star passing North, from home last month’.

If you don’t want calculators and trigonometry, you don’t need to use them.
E.g. ‘The sky is behaving the same as last year, we must be within sight.’

I am about to give you the star time and star longitude information, only for those who wish to work with the easiest benefits from them.
E.g. ‘That must mean that our longitude is 150° E.’

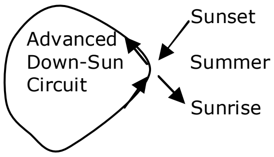
Advanced
Down-Sun
Circuit
Sunset
Summer
Sunrise

This Assumes Southern Hemisphere

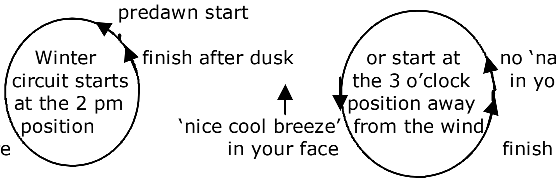
Winter
circuit starts
at the 2 pm
position
predawn start
finish after dusk
‘nice cool breeze’
in your face
or start at
the 3 o’clock
position away
from the wind
no ‘nasty cold wind’
in your face
finish

Note that clockwise circuits and half-day walks must alternate between up-Sun and down-Sun.

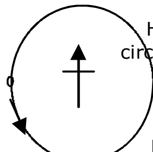

How **not** to circumnavigate a lake — from the 9 p.m. position
The Sun shines in your eyes all day!

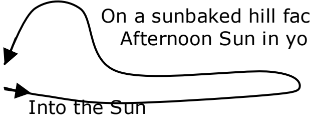
On a sunbaked hill face
Afternoon Sun in your eyes
Into the Sun

How **not** to climb a hill — at the end of the day

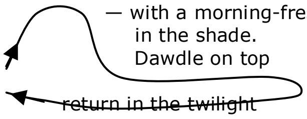
Dawdle on top
return in the twilight

How to climb a hill — with a morning-fresh climb in the shade.

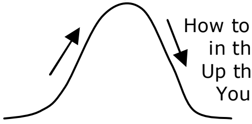

How to climb a mountain in the outback heat —
Up the west face; long lunch on top
You walk always in the shade.

Matching Sky to Sky — Move Only Your Horizon

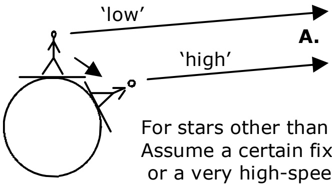

A. Move **towards** a star and it will **climb**
e.g. move toward the pole star
to a higher latitude
and it is higher
For stars other than the pole star
Assume a certain fixed star time
or a very high-speed—to stop the stars moving.

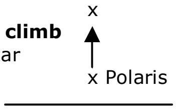

The pole star is too low

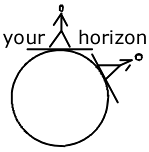

B. Move **sideways** and your horizon **tilts** with you
e.g. a star pair
may not be upright
as expected —
move sideways.
Or perhaps
a star pair
sets before it becomes upright
You need to tilt your horizon
down to the right
by moving to your right..

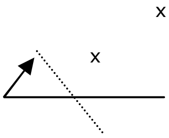

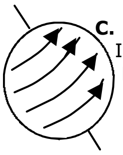

C. Because the Earth spins to the **East**, it leaves the stars behind in the West.
If you move west with the star it will slow down, so…
Move **E**ast and the stars will *get there **E**arlier*
e.g. 2 stars become upright too late?
You need to move easterly.

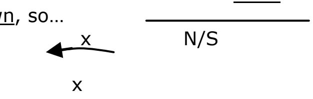

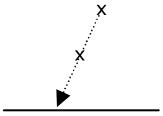

You may need a combination of A., B. and C.

Prevailing Winds on Creek Banks

One of the two Creek banks will often have a mud- or rock- cliff due to prevailing wave-action, with the other bank being a mud flat.
E.g. If prevailing winds are westerly, the west bank will show less erosion.
For more on finding the prevailing wind direction when the wind is not blowing, see section 3: “Nature’s Norths”.
Because prevailing wind is almost continental in scale, navigating by it is a sort of Big Picture technique.

Tips for calm weather

1. Listen for the wind! Sounds will carry better downwind e.g. surf or traffic. Remember this, at night when all else is still and quiet.
2. Watch the angle of mist droplets falling. (Stop still for this one) Or smoke rising.
3. Huff into the air to make a fog.
4. Watch the flimsiest vegetation.
5. Look at the leaves which dangle down from the topmost leaf clusters of a tree.

Mnemonic: “North; Coast”

That’s what you should say, whenever you say ‘North’.
“North (Coast) is over there” — it reminds you to think,
or rather to picture, the invisible picture.
You should make it your business to name the North-Coast town north of you,
before you go, by consulting a roadmap
Then of course, there is the East Coast, and the South Coast and the West Coast.
Next, of course, the same must go for “North-West”
— think of the North-West Coast township, and so on.
You are not functioning properly unless you mean something when you talk!

Go Overseas!

This is a one-off effort, so go to the trouble of getting a globe,
to sort out what is truly “East” and “West” of you,
in a Great Circle straight line.
It will surprise you what East and West should mean to you.

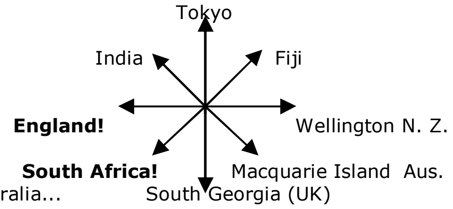
Tokyo
India
Fiji
England!
Wellington N. Z.
South Africa!
Macquarie Island Aus.
South Georgia (UK)

e.g. from Australia...

Mnemonic "The Sun Travels the Tropics"

I.e. Imagine the **tropical countries and oceans** which are **'underneath' the Sun** as it moves East to West during the day — The Equatorial band.
For Australia that means from Central America Across the Pacific Ocean Past New Guinea
On to the Indian Ocean And on toward Africa
When we in Australia see the Sun rise, it is past noon in the Americas. At our sunset, it is approaching noon in Africa.
See the Reference List of Tropical Places (for The Eastern Australian Time Zone) opposite. As a help to global orientation, make it your aim to memorise these time zones in sequence. Check it out with an atlas. See the tear-out clockface diagram in the appendix.
The same sequence holds for any planet, or star near the Equator e.g. for the stars listed 'opposite the Sun' (below), but with different timing.
You can also imagine which country the Sun is overhead while it is night-time near you,
or which country the star opposite the Sun is overhead while it is daytime near you.
Now, with this in mind name the general North/South directions, according to the height of the Sun: To the Sun = "Hot"; To noon = "Hottest"; and the Shadow direction = "Cold"; Shortest shadow = "Coldest".
•Although the Moon, by comparison with the Sun, has a "cold" feel to it, make sure you apply the same list of hot tropical countries to it also.
When you see any Zodiacal body, you should think: "That is a Hot direction; It is over a Hot country." and try to name the country.

Tell the Time of Night by the Star **Opposite the Sun**

Such a star will be **up all night** — whether the night is long in winter, or short in summer — look for it. Write them on your calendar.
This yields the solar time, i.e. with respect to local noontime, but not the daylight-savings or summer-time, so adjust by one hour in summer.

January	Sirius *Castor, Pollux*, Procyon
February	Suhail, Miaplacidus Alphard *Regulus*
March	Pointers (Dubhe-Merak) *Denebola* Phecda Southern Cross
April	Southern Cross Alioth *Spica* Arcturus, Pointers (R.K./Hadar)
May	*Zubenelgenubi*, Kochab Alphecca *Antares*
June	*Antares* Atria Rasalhague Scorpion Sting Kaus Australis Vega
July	Vega *Nunki* Altair (Aquila) Peacock Deneb
August	Deneb Enif (ε Pegasus) Alnair
September	Fomalhaut Middle of Great Square
October	Diphda Mirach Achernar Hamal-Almach
November	Mirfak
December	*Alderbaran* Capella-Rigel Saucepan Six O'clock Stars; Canopus

This list is arranged time-wise
the earlier stars are correct
at the beginning of the month,
the last stars, near the end.
Use each as a 24 hour-hand around the South Pole Star
The Zodiacal stars
(the ones on the Ecliptic)
are in italics.
Only they are truly opposite the Sun.
This chart is **true for**
both Hemispheres, permanently.

Two Different Back-and-Forth-Star-Directions Give You a Position Fix

..Yes, it's just like two bearings (which are not opposite each other) from two landmarks. But because these 'landmarks' are so distant, the accuracy is only to within seeing distance. For better accuracy you would need a theodolite or sextant.
E.g. when two stars look upright, that locates a global star-direction; a second upright pair gives you the second direction, and a position fix.
To fix your latitude. you have to **measure the time-lapse between the two star pairs**. (Simultaneous is simplest.) So you need a plumb-line and also a watch. And for accurate (Greenwich) longitude fixing, you need to work in (Greenwich) **star-time** — otherwise you have only latitude which is fixed. So you need either an almanac, or calculator, or star-watch, or a work-it-out-at-home-later approach.
This is equivalent to fixing which star is exactly overhead and when. To see this, look up past the top of the plumb-line to 'overhead' from the two different directions—one for each pair—to see that they must cross exactly overhead.
The 'fix' is repeated accurately over years, and every night at the same star-time, with some very small predictable drift in location and timing.
No one else on earth can see exactly what you see right now. That's what gives you a unique position fix.
People at the same latitude see something very similar earlier or later than you do.
The best two directions to cross are Northwest/Southeast and Northeast/Southwest. Top-view...
This is because East/West star directions are time-wastingly difficult to determine.
This method is most helpful, of course, at sea, or in the desert, or as a hobby, or when lost.

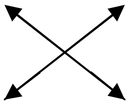

Predict the Full Moon Path According to the Time of the Year

The full Moon is **opposite the Sun**, and is near the **Ecliptic**, i.e. in the Zodiac (the Sun's path through the stars) so that…
When the Sun is in midsummer position—high, and up for a long time—the full Moon will be in a midwinter position—low, and setting again quickly
When the Sun is in the midwinter position—short days—the full Moon will be high and up for a long time — long nights.
The two complement each other, since when one comes up the other goes down, and when one goes down the other comes up.
Both Sun and Moon are on the Ecliptic—not usually on the Equator—but, at the Equinoxes, the Sun is on the Equator, so the full Moon is too.
Near Equinox, of course, the Sun and Full Moon will rise and set in almost the same positions — East and West — and follow similar paths — about 12 hours each, along the Equator. The Equator and the Ecliptic intersect in two opposite points — the equinoctial points.
In other words, you can **use the full Moon to visualise the Ecliptic**, and to guess where the Sun is, i.e. to guess the time of night,
Or you can use the Sun to visualise the Ecliptic, and to guess where the full Moon will be, how high, and for how long.

Finding North from the Moon, by its height & phase

The Moon will follow a Midsummer Sun's path, or a Midwinter Sun's path, or an average height path, according to where on the Ecliptic it lies. Because the phase of the Moon is determined by its relative position to the Sun, the combination of the time of day, the phase, and cusp-Ecliptic gives you an unmistakable sense of where North should be, when you see the Moon. When you see the Moon surprisingly low in the sky, and the cusps confirm that the Ecliptic lies fairly horizontal, you know that it is somewhere near North (SH), following a low and horizontal trajectory, rather than a Moon 'just now rising' on a high trajectory, like a summer Sun in the East.

One Degree is about One in Sixty

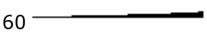

And **n° is about n in 60**. e.g. **10° is 1 in 6** (p96a), and **6° is 1 in 10** (3 fingers wide, or 3½ knuckle-centres, at arm's length, or 'blink your eye').
These are handy rules of thumb for navigators. If you use an isosceles triangle, **the rule is remarkably accurate for most acute angles**.
Example: "Oh I forgot magnetic variation! I've been **5° off**!! Oh well, I will simply adjust sideways.
I have come 12 km; 1/60 of that = 0.2; times five = **1 km off**"
Actually 1° is 1 in 57.3 (1.75:100), predictably slightly more than 1 in 60, but as the angle increases so does the accuracy of the 1:60 rule.
The rule (n° = n across at 60 away) is based on: the third side of the **isosceles triangle** (is twice the sine of half the angle).
(If you don't like numbers, try this other "handy" rule...)
Rule of Finger: point your index finger at something — the fingernail is 1°.
Why is 1° slightly more than one in 60?...
Six equilateral triangles go all around the circle, but, as you know, 6 1/4 radii go around the circle.
This is because the circumference, longer than the hexagon, is 2pi times the radius, or 6.283r
So one 'radian' of arc, which is one radius long, but curved, will be slightly less than 60° i.e. 57.3°.
So if the radius is 60, one sixtieth of a radian will be "one in 60", and slightly less than 1°.
377 of them go around the circle. For 360 degrees around the circumference, the radius needs to be **57.3**.
The equivalent is '5/3 across at 100' (5 at 300, 167 at 1000; above 10°) with the more exact figure for up to 10° = 7/4 in 100

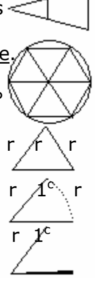

Tropical Countries and Places Under the Sun — Arranged in Time Zones Relative to Eastern Australian Time — 10 hrs ahead of Greenwich

-12hrs.	The mid North Atlantic (is opposite Australia)
-11hrs.	Brasilia 16° Sao Paulo/Río de Janeiro 23½° S.
-10 hrs.	Trinidad and Tobago 10° N.
-9	Cuba/Haiti 20° N.; Bogotá 5° N.
—	Panama 8½ hrs ahead of Australia
-8	Galápagos 0°; Guatemala 15° N.
-7	Gomez 26° S.; Mexico 23½° N.
-6	Pitcairn Island — Easter Island 25 to 27° S. (L.A. 34°N).
—	Tip of Californian Peninsula 23½°N; IDR Gigego 20°N
-5	Mangareva Tropic of Capricorn
-4	Tahiti 17°½ S.
—	Cook Islands 11° S.
-3	Hawaii Island, Tropic of Cancer
-2	Fiji 15° S.
-1	Noumea 21° S. See pp 132f
0	Rockhampton, Tropic of Capricorn 23½° S. (Canberra 36° S.)
1	Alice Springs, Tropic of Capricorn
2	Sulawesi 0°; Manila 15° N.; Port Hedland 20° S.
3	Christmas Island 10° S.; Jakarta 7° S.; Singapore 1°N.
—	Cocos Islands
4	Calcutta Tropic of Cancer 23½° N.
5	Maldives Equator
6	Gulf of Oman Tropic of Cancer
—	Seychelles
7	Port of Aden 12° N.; Mogadishu 3°; Madagascar 23½° S.
8	The north of South Africa on the Tropic; Lake Victoria 0°; Sudan/Egypt on the Tropic; Aswan High Dam
9	Namibian Desert coast on the Tropic; Lake Chad 14° N.
10	Timbuktu 17° N.; Accra 5° N.
11	West Africa coastal bulge 15° N.

Tell North from The Time By Using a Protractor

Reverse the method for telling the time from the compass protractor and North. To make the zero point North...
Make the sundial (protractor) show the correct time (one side of 'noon' time). Simultaneously make sure the sundial face is tilted to be parallel to the Equatorial Plane… I.e. For best results the tilt of the protractor according to the latitude, needs to be accurate.
Consequently, you may find it an interesting exercise to mockup a large (bigger = better) cardboard multi-protractor
— e.g. by marking the sides of a cardboard carton.
Use a plumb-bob to **accurately tilt the protractor, according to your latitude**. E.g. pin the plumb-line to the top, side corner, and mark 90° on the flat face…

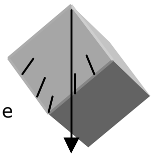

This instrument may be used to find North if you know the local time, or to find the local time if you know North.
One side of the carton is used for the latitude protractor in conjunction with the plumb-bob, a second side can be marked as a sundial protractor.
E.g. use another (long, straight, perpendicular) pin at the top corner of the side facing you, to cast a shadow across a 90 degree scale (— this is for accuracy, rather than make a full 180° or 360°).
Morning and afternoon then require separate scales; Winter-time requires that you face away from the Sun.

For Land-Navigation, Choose the Plumb-Line Star-Pair Method

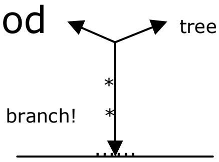

- Use a smooth constant diameter line, preferably thick. Make the plumb-line tall and stand far back. Make the cord either white or black to stand out against the sky, then maybe flash a torch onto it.
- Pour sand around the plumb, to anchor it to the ground. Avoid wind, especially if you hang the line from a tree branch!
- " Blink " the stars "on" and "off" behind the plumb-line, both on its left and on its right, by moving your head.
- Average the "start" & "finish" times. I.e. "It seems to line up now for the first time"… "definitely finished now". One accurate set of measurements is better than several rough ones. One star will gradually move over the top of the other star, and you have to pick exactly when.
- Choose well-separated stars (e.g. Rigel & Capella)
- If you want the timing to give you your longitude immediately, choose a North/South pair (e.g. Rigel & Capella) — (you need to know the star's longitude or its timing elsewhere, i.e. from a known longitude).
- An East/West pair will give you your latitude (without the need for a watch). The observation here is not the time but **whether or not** the given pair will attain to uprightness, after the first glimpse or before the last glimpse. Beyond a certain latitude, a given star pair will always remain tilted, less than vertical. This latter process can be time-consuming. Use a hand-held plumb-line occasionally, to check the progress.

Predict the Moon Phases by " N minus M "

" N " is a number (you calculate) for one particular year (e.g. For 2004 it was "24", in Australia's time zone, for the New Moon)
" M " is the month number (e.g. April is "4")
" N – M " is the day of the month for a certain fixed phase (e.g. in April 2004: 24 – 4 = 20th April 2004 for the New Moon)
April 20 was dark all night. New Moon seems to show the most regularity. The 'new Moon' is more accurately 'no Moon'.
Other phases happen one week later, two weeks later, three weeks later — 'quarter' by quarter — 29½ days for the whole cycle.
(The number, N, in 2004, would be 2; 9; 17; 24, 32 approximately, for 1st; 2nd; 3rd; final and 1st quarters respectively.) You can add 29½ to N.
Then in 2005, the number, N, for the new Moon was 13... (i.e. in April 2005, 13 minus four equals… the ninth of April — no Moon).
And in 2006 I used N = 32; 21 in 2007. The number N for the New Moon must be close to any **New Moon date + M**, e.g. December New Moon+12
Check your diary's calendar each year for the Moon phases; test and tweak a number N which works to within plus or minus one or two days.
The prediction is fairly crude, but with trial and error you can come up with *the best number and phase combination* for that year. For 2010, it is 18.
—usually **11 days less than the year before** (12 for leap year) because the lunar year of 354 days is 11 days shorter than the ordinary solar year
I.e. a good mnemonic is: **the Moon gains one day a month for every month except February** (which has only 28 days).
Hints: *Subtract " M " twice for January and February,* for more accuracy on non-leap years. Alternatively, to cope better with leap years...
Start the year in March (after the short month, February). Call January & February "months 13 & 14", but subtract 1 more for their accuracy.
Note: The "Full Moon" is when the Moon is nearest to "opposite the Sun" and it may not be exactly on the horizon when the Sun is.

Walk across Australia!

(Substitute whatever continent is appropriate)

Mentally line up a wall-poster size map of the whole continent (or bigger) with the track.
When you approach a bend, place the map on the bend, then walk all over it!
It's hard to miss the meaning of the bend if you bother to do this.
Your mind may not believe that it is not still going 'straight ahead' until you force it to catch up with the facts!

English!

"East-around" from North, is West-around from South, in a horizontal clock-sense. An "Easterly" course means East-wards — moving **to** the East.
An "Easterly" motion means East-wards. E.g. an "Easterly spin"-motion = "orbiting in an Easterly sense" = moving E around the N/S spin-axis **But**…
An "East(erly)" **wind** means flowing Westwards, **from** the East. An "Easterly" **current** might be flowing Westwards, from the East, but an easterly tidal "set" is an offset to the East. (Set is how far the tide "sets" you off track. The speed is called the tidal "drift".)
" Due " East means "exactly" East, as when a loan is "due" — "right now!" **Always double-check what someone means**!

North/South and Noontime by **Bisection (of Rise and Set)**

Rise and set are **symmetrical around North/South**, provided you use horizontal, and not the horizon.
In hilly country, instead of ‘rise and set’, you should use **equal height** of the Sun (e.g. in NE and NW).
“Equal height” translates as *equal length shadows* of an *upright*, *straight* stick or cord, onto *flat* ground…………
Or, better, use a bush sextant, p.27c, set to a fixed reading.
There are four ways to bisect two directions to find North/South — so don’t muck it up by 90° or 180°!
The second diagram will work in both hemispheres — but double check the results for sense.
There are also two ways to **bisect the time** — so don’t mix up ‘just before midnight’ with ‘just after noon’.
It is also hard to ‘bisect the time’ to find noon, which of course, is not at 12 noon exactly, nor at 1300 hrs exactly in summer.
To work with hours, you need to **carry 60 minutes**, not 10 or 100! So be careful and crosscheck your figures.
Be accurate or your results will not be worth the effort.
E.g. rise: 0529 hrs; set: 1820 hrs Add… = 2349 hrs Halve… = 1100+(**60** +49)/2 = 1154½ hrs = noon.
Check: total elapsed time: 1820 –0529 = 1251 = 2x0625½; 0529 + 0625½ = 1154½ = correct
1820 – 0625½ = 1154½ = correct

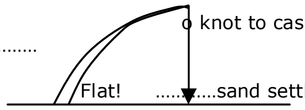

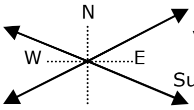

Longitude and Latitude from the Sun

Longitude...
Try timing the Sun for **equal height** each side of noon e.g. NE & NW (SH) to calculate the time of noon.
This can tell you your Greenwich longitude, provided you know the Sun’s longitude, that is, its G. H. A.— from an almanac or calculator.
The Sun shifts predictably at a rate of near 15°/hour westerly. Almanacs make it their business to predict its West longitude exactly.
Of course the Sun’s longitude at the time of your noon is your longitude too — since it is north of you.

Now try timing the Sun in two **opposite directions** at the same heights i.e. East and West.
You could try using the shadow of a plumb-bob on flat and level ground. Mark the shadow carefully.
This will be most accurate in the season when the Sun reaches high — in the summer sky.
By a series of early observations you can guess at when to make a final observation.
Using a theodolite or bubble sextant allows you to graph the rise & fall of the Sun (or a star) and see when ‘same height again’ happens.
See the ‘Bush Theodolite’ for an alternative measuring method.
Latitude (There are easier ways!)...
You can now solve the E/W formulae for your latitude — e.g. try the formula that predicts the timing of East and West for the Sun.

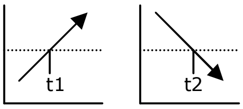

Equal-Height Stars Can Be Useful When the Sea Horizon is Visible

(or with a bubble sextant)

For direction, use the mid-point, of two stars which appear to be of equal height. — a great circle direction
For longitude, the timing is the important bit. — these stars should be approximately north or south of you
For latitude, the test is whether or not the stars do attain to equal height. — these stars will be approximately East or West of you
For a position fix, try to get two pairs, one north-west, the other north-east of you, for instance.
A **bush quadrant** is easy to make with a string tied to a stick! — just pull the string to length to measure or compare star heights.
Rest your hand and chin on something solid if possible. Use your full arm’s length for best accuracy.

You can fix your latitude by measuring the **timing** of **two** separate star-pair events
— the relative timing — e.g. ‘5¼ minutes apart’ (is unique to that latitude).
Simultaneous events are simplest, but separated events are easier for one person to time.
You don’t ‘get on the map’ with this simple method unless you have been there before.
Unless you want a lot of trigonometry no calculations are required,
only the observation of re-recognisable appearances.
You can fix your longitude, too, by measuring any absolute timing — an exact star time — and so get on the map.

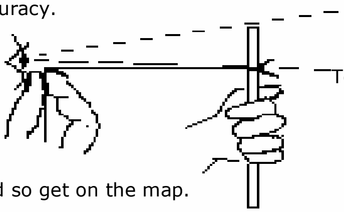

Anticlockwise, to the East, is Your Best Spin-Reference Direction

Because we use our North-South axis as our reference direction, we have to learn the *spin-sense* as something *separate*. See ‘Orientation’ p 35a.
The standard viewpoint for any spiral system (e.g. a weather system, a clock-face) is ‘from the top’, “**as seen from above**”. Horizontal directions are measured “clockwise around” as you look *down* on them. For our spiral galaxy & solar system, the correct viewpoint is therefore ‘from the North’
Looked at ‘from the top’, i.e. from the North, our Galaxy spins anti-clockwise. Or more simply, ‘to the East’. ‘Up North’ is ‘on the left’ of East.
This means that the real motions are East-around the North/South spin axis—opposite to the familiar idea that everything sets in the West.
Because Northerners have their back to the North Pole when they see the Sun Down South of them, they see their horizon spin anti-clockwise
— that is, the hills and trees on the horizon move to the left of the Sun. They are looking “from the top” i.e. from the North further South.
Their Sun only appears to move clockwise (to their right, to the West), because it is their horizon moving—anti-clockwise (to their left, to the East).
To avoid confusion, always make the distinction between real mass movement and apparent virtual rotation. And think: We spin East.
Because we think of ‘clockwise’ and ‘right-hand’ motion as ‘standard’, and setting “in the West” as ‘normal’, it’s a bit of a shock, to rethink!

When you are Down Under, on the South side, looking North at the noon Sun, you see the bottom of our Galaxy, and Solar System, and Earth.
The three ‘North’ spin-axes are all very roughly aligned together, all lying in our “Northern” Hemisphere — not too much tilted from each other or from our own spin-axis. You mentally lift yourself further out (not “up”) into space to look at (but not really ‘down’ on) the spiral motions
Remember that from a ‘standard’ (Northerner’s) point of view, the Southerner’s view is really “a bottom view looking upwards” — from the South half, and you have moved out further South, for a better view — with the South Polestar behind you. You should see clockwise real movements.
Turn around and look Southwards at the Southern sky and you will see as if ‘from the top’—or at least ‘from within’, but looking South. Anticlockwise
Remember to rethink “Up” as “to the *North* Pole of the spin”.

- It should now be easy for you to visualise, in either hemisphere, your Southern horizon spinning anticlockwise.
- Look also at the Moon—visualise its orbit and its spin—‘anticlockwise’ from N.H., but clockwise-around if you are seeing it from e.g. Australia.
- Find the bright planets — can you ‘see’ them moving ‘anticlockwise around the Sun’ in orbit? The Ecliptic marks their path.
- And, go further, visualise the real Sun spinning much like the Earth does, on its own N-S axis. The Sun is North of you in the Southern hemisphere, spinning on its own axis ‘clockwise from below’, or in the Northern hemisphere, South of you ‘anti-clockwise from above’. Use the Ecliptic’s N-S spin axis as a good guide for the Sun’s, or simply use the Earth’s own axis as a rough guide — Don’t look at it, but imagine its sunspots moving.
- Keep going: The Earth hurtles through space in orbit around the Sun ‘to the left, while looking to the Sun, around a N-S axis of spin’ (Down Under, looking North) ‘to the right, relative to the Sun’ (NH) ‘**anticlockwise**, looking South’ (NH)—‘with’ the apparent movement of the Sun, in both cases. That makes the noon horizon move in opposition to the movement of the whole Earth through space. Like the slow ‘inside’ of a spinning ball circling.
- Imagine the whole Earth-orbit in the sky. It’s easy when you get the hang of it. New Zodiac stars seem to emerge (invisibly) on the Sun’s West side, in seasonal procession, because “the Sun is edging Eastward along the Ecliptic in a yearly cycle” i.e. we move to the West side of the Sun and see behind it to the background Zodiac stars. Don’t look at the Sun, but you should be able to ‘see’ this happening as if there were a star chart behind the blue sky! Try just before dawn.
- Now, at night-time, find the centre (thickest part) of the Milky Way near Scorpio/Sagittarius and visualise the whole Solar System orbiting it perpendicular to that, the Sun/Earth/Moon all moving as a unit, to your left (SH) or right (NH) towards Suhail, away from Deneb.

The Milky Way marks the plane of the spiralling galaxy (from which you can determine its N-S axis, tilted differently from ours).

Finding Magnetic Variation by Bisecting Rise and Set

Accuracy is needed.

Measure the compass directions of the morning and evening Sun at equal heights e.g. rise & set on a flat Horizon
Bisect the result. E.g. Rise: 079° Compass Set: 271° Compass
Add: 350° Halve: 175° Use the Opposite: +200–20=355° (See p.18e)
So the Compass North must be 5° to the East of geographic North. **Draw diagrams to avoid getting confused.**
Check all numbers! E.g. 355°–271°=084° 355° to 360°=5° + 079° more=84° Correct.

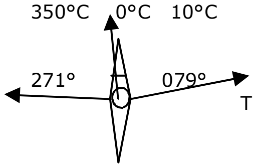

Or 271 – 79 = 192° spanning = 2x96°
79+96 = 175°
= opposite to (375 – 20)°
Correct.
The compass readings are "Least", by 5° (see p 107b)
E.g. 079°C is really 084°T. (see p 14e)

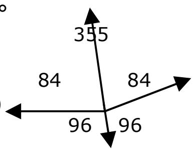

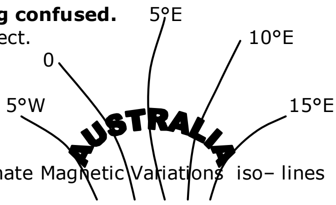

Approximate Magnetic Variations iso– lines

A Bush Theodolite for Overnight Equal Height Observations

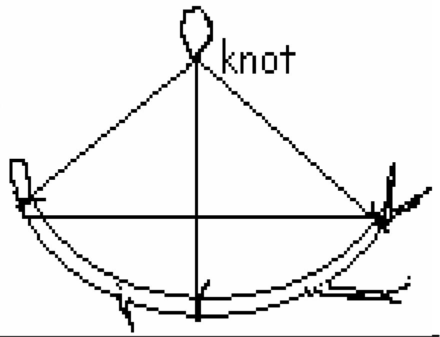

This is for accuracy in calculating... True **North**, Noon **time** and Compass **variation**
And for any equal-height paired observations — Sun/stars/etc.
Find a large bent stick. (Or bend it with a bowstring)
Hold it up by a loop-knot e.g. hang it on a twig.
(If it is early or late i.e. a long shadow, the stick will not need to be bent much.)
Late afternoon, at camp,
Mark where the shadow of the knot (or the other end of the stick) falls on the stick;
take the time; line up the landmark it points to; and find its compass direction.
Early morning, next day,
Hold it up (in reverse); wait until the shadow again falls in its proper place;
get the time, direction and compass reading again.
For a **Bush Bubble Sextant**, arrange a water filled clear tube on a triangle of lashed walking sticks so that there is a trapped bubble in a tube section with an ever so slight convex-upwards bend.
Make the bubble always stay at one marked place when you are using it.

Great-Circle Reference Star Directions

You need some familiar reference points. It's a simple matter of going out at night, waiting around, and tracking down a few of these...
Look for **stars which are closer to the pole than you are**, and see where they **rise and fall vertically**. It happens once or twice a night, per star.
This gives you a series of great-circle star directions, already labeled, because **the 'maximum latitude' is equal to that star's 'declination'**.
E.g. Acrux is about 30° from the pole, or "60° latitude". If you steered towards one of its vertical-motion directions you would end up underneath it.
So as it rises or falls vertically (once each per night, if visible at the time), as it circles the South Pole Star, it lures you to 60°S latitude.
E.g. A star overhead, as the extreme limiting case, is moving East to West, rising vertically from the East and falling vertically to the West, meaning that you are *already at* the maximum latitude of that course, if you had been heading towards that star guide, now travelling E-W also.
You then have to head directly *away from* it (e.g. the next night) to further pursue that course.
Heading away from that star either East or West would take you eventually back towards the Equator, like the way you travelled down there.
The **stars further from the pole, than you are from the Equator,** make a complementary set of reference directions, by using the horizontal direction 'at **right-angles** to where stars **rise and set**'. **The declination is a measure of *how close to the pole* you would pass**.
The stars close to the pole are unavailable. The co-declination labels 'the maximum latitude'.
E.g. Sirius is 17° South, so rising from the horizon, somewhere to the South of East, it would send you to within 17° to the left of the South Pole.
Turn around and it would guide you (part way) to cross the Equator and to within 17° of the left of the North Pole.

Make a Bush Psychrometer

..for absolute humidity and dew-point-temperature measurements

Carry a wettable piece of material — dip it in water — hang it up to dry (in the shade out of the breeze) — time how long it takes.
Measure the air temperature at the same time, and cross correlate the two, to find the absolute humidity. If you calibrate the method at home, by graphing the results for a whole range of different temperature and humidity conditions, you will get to recognise a dry air mass from a moist one.
The piece of material needs to be non-porous, non-rotting and dark coloured — to show the drying process clearly, and to yield repeatable results.
Night-time minimum temperature more or less bottoms out as soon as an air mass starts to dew. The ground then wets down further instead of cooling down further. The normal method of measuring absolute air humidity—the percentage of water vapour—is by evaporatively cooling a water drop in it. It requires wetting a thermometer bulb, fanning it vigorously (trying hard not to break it) and measuring how far you can make the "wet bulb temperature" "dip" below the "dry bulb (air-)temperature". The water temperature won't get all the way down to "dew point"—the temperature at which dew or frost will form in that air mass—unless it already is dripping wet outside. It may get a bit more than half-way, and then only when only dew is expected—DP above 3° C. **If frost is possible,** i.e. when the dew point temperature is 3° or less, **it will dip at most half way down to the dew point**. Why 3°? Well, when the air is very dry, dewing will further reduce the water vapour content of still, cold air, from a dew-point temperature of a few degrees above zero, down to 0°—resulting in 'frozen dew'. **Rule of thumb** for calculating dew-point from dip:
When the DP is 40°C/30°/20°/10°/5°/0° then add to the 'dip' 0/1/2/3/4/5 quarters of the dip, to get down to the real dew point.
If the night is long enough and calm enough to cool down enough, & the sky is clear enough, a DP near zero will result in a frost.
E.g. 15° to 10° = 5° dip, 5°DP. 15° to 13° = 2°dip, 11°DP. 19° to 18° = 1°dip, 17°½DP. (Remember that you can't go far wrong anyway.)

South is a Triply Better Reference Direction than North, in the South

1. The closest geographic Pole in the Southern hemisphere is just over the horizon in Antarctica. Globally seen, all directions pointing South are convergent, whereas directions pointing North are initially divergent, and lengthy and less meaningful... Travelling North you first cross the Equator before it gets colder — so the gut-feel meaning is less clear. The meaningful direction, when you are in South latitudes, is to go further South to get much colder. (It is the same as in the Northern hemisphere, where going further North makes it much colder.)
You may find that in a place like Tasmania for instance, the maps seem to be upside down, because although North is at the top of the map, the logical reference direction is toward the bottom of the map, where the cold place gets colder. You don't think: "This is a warm place. In which direction does it get warmer?".
As a result it is easy to get disoriented by 180°, if you insist on using North as your reference direction. In addition…
2. When you put South at the top, and look South, then the Sun will move to your right, as if seen from the N.H., across the map as time passes, for both hemispheres—and that will make the clock-to-Sun link obvious (like putting a Southern clock on the S. wall to match the Sun's motion). Northerners especially, would find that a more natural way to view any map—as if from further North, looking South, to the Sun moving right.
3. Putting South at the top, with the Sun & Tropics behind you, looking down onto a map, is the mathematically standard viewpoint — like looking from the top of the Galaxy — from North to further South — the land South of you moves Anti-clockwise, to your left, to the East.
You can still orient your map normally and correctly, but there is no law against habitually viewing it from its Northern edge, looking South. Try it.

Find East/West By Delayed Bisection of the Sun and the Full Moon At Rise or Set

To find due East (+/–10°): **Bisect the sunrise and the full Moon rise** — since they are of opposite declination.

For due West (+/-10°): **Bisect the sunset and the full Moon set** — since they set each side of due West.

Obviously, you need to wait about 12 hours for this method — it is really just for your background comfort level about where things are.
There is a slop of 5°, since the Moon's plane of orbit is about 5° away from the Ecliptic.
There is also a slop of about 24 hours, since the full Moon occurs at some time other than Moonrise or Moonset,
plus it is hard to pick an exactly full Moon, and the Moon varies in declination, i.e. North/South position, from day to day.

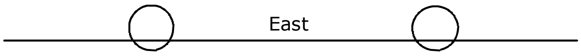

Find North by your (non-digital) Watch-Hands

Point the noon on the watch (**not the 12!**) to the Sun.
Tilt the watch axis toward Polaris as best you can judge.
This tilt matters a lot, but is not usually mentioned!

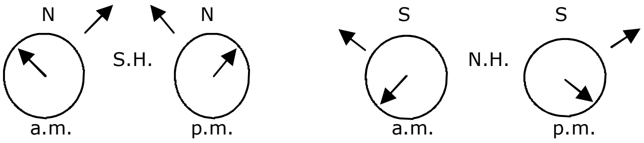

Southern hemisphere: **North is in between the Sun and the hour hand**.
Northern hemisphere: The Sun is in between the hour hand and South.
Why?: Clocks Go Clockwise; at 30° per hour — twice around per day.
For the Southern Hemisphere that means: — twice as fast as the Sun and opposite.
So 'in between' moves neither clockwise nor anticlockwise — it stays N/S. At real noon-time, the hour hand and the Sun cross over.
For the Northern Hemisphere, the hour hand overtakes the Sun at noontime, at South.
It's easy to visualise in fast motion, once you catch the idea — it's like the hare overtaking the tortoise.
Remember to tilt the watch-face up to be parallel to the Sun's orbit, not horizontal, because a sundial shadow is not linear.
You could use the middle of sunrise/sunset times to find the local actual high-noon midday time, which is needed for "point noon at the Sun".
The same method can be applied to the stars, the Moon or the planets — provided you know when the one you want to use is north of you.

Leo the Lion Looks like a Lion or Sphinx Squatting
It Has Four Lines which Show N./S. E./W.

Two in the 'hindquarters'; one along the 'back'; one as the stem of the 'question mark' (Leo's mane).
(You have to turn the '?' around.) For the Southern Hemisphere, turn it all upside down.

This is a good sign in Northern spring/Southern autumn.

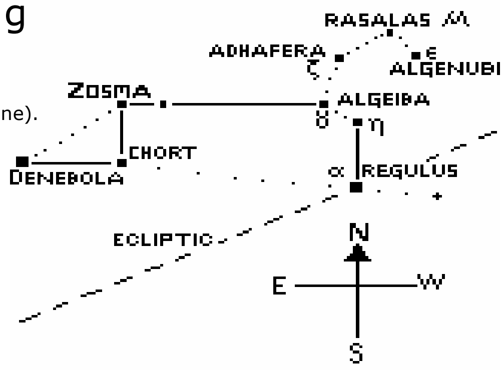

(NH view — LEO is upright)
Direction of movement ⟶

To Predict the Moon's Path, Visualise the Ecliptic

For more clues on how to visualise the Ecliptic, first see the overview at the end of this section.
The Moon stays within 5° of the Ecliptic. That helps you see where the Ecliptic should be, or alternatively,
where the Moon should be. The Full Moon is easy, and has been covered on p. 24d.
A **half-Moon,** for example, is 90° from the Sun; so if one is on the Equator the other is not and vice versa.
E.g. a half Moon in mid-winter will rise and set due E./W. +/- 5° or so.
Secondly, recall that the Earth's spin-axis is tilted 23½ degrees from the solar system.
This is how it would look from Sydney (*), at midday, in Midsummer, i.e. at Southern Summer Solstice…
The two great circles intersect at 23°½. The whole double-circle rotates West over 24 hours. i.e. sunset is at X…
To visualise the Ecliptic (see also p. 30d & 33b), from the Sun...
..Since the Sun moves eastwards seasonally through the stars you can know that
the part of the Ecliptic which the Sun was recently on must 'lead' the Sun to the West.
E.g. after solstice in Southern Summer, the most Southerly part of the Ecliptic is ahead of the Sun
to the West, and so a waning crescent Moon (to the West of the Sun) will be higher in the sky, on average, than the Sun. (SH)
E.g. A crescent Moon at sunset at spring equinox will look extraordinarily high over the sunset glow, because of the 23°½ tilt.

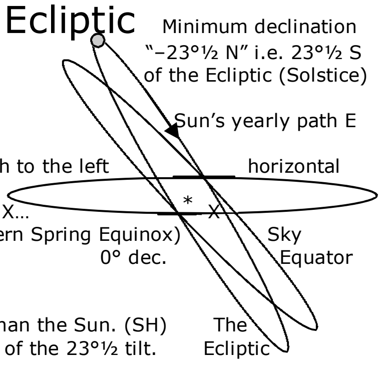

Bearings & Tracks: "If it is Right, it is Big, in my sight"

= Anything on my **Right is Bigger than 'ahead'**. I.e. If I am heading 100°, something on the right has a bearing bigger than 100°, e.g. 200°.
And "Any track on my **Left is too Little to look at**" = Anything on my **Left is Less** than ahead, e.g. 20°.
These mnemonics compare fixed tracks or directions, and their 'Big and Little' measure in degrees, East-around-from-North i.e. clockwise bearings.

[The changeover from 360° to 0° can give you trouble. Where 'ahead' was say 355°, 15° bigger should be 'big' e.g. 370° but will look really small, 10°. Because it is nowhere near 355 in number, there should be no possible confusion, provided you think of 10° as 370°.
'Little' may look really big, if 360° gets in between. E.g. where 'ahead' was 005°, 15° smaller would be –10°, which is the same as 350°, which is nowhere near 5 in number.]

Furthermore, if A sees B on his right, relative to a fixed direction, B sees A on his right,
provided that they use **acute** angles, or they look in **opposite** directions e.g. facing each other.
This could all get **confusing if you don't remember those two provisos**. See p108c for how to put this mnemonic into practice.

Find North from the Combination of the Sun and the Moon

When you are familiar with the path of the Sun and of the Moon, and with their seasonal movements, and with the Equator and Ecliptic, then whenever you see the Moon up during the day, imagine where the Sun's path lies in relation to the Moon's, and vice versa.

This gives you two concentric circles to visualise, which of course allows you to locate Polaris (at the centre of those paths), and from there, North.

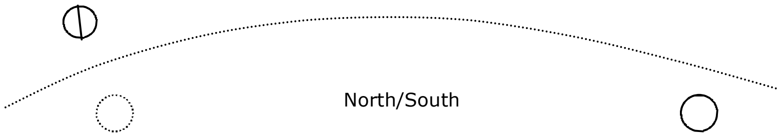

See box d/e below, to know where the Moon should be.

Find Any Other Direction Too — Using the Hour Hand

This method works, in theory (there are easier methods!), for any other body — star, Moon or planet — or for any other direction.
For instance, you may notice when the Sun is due East, and how high it is; or again, when it is in the direction in which you wish to head.
Or again, you may be following a star, and of course it will move and of course you will wish to compensate for its movement.
In these cases you substitute the time of that body being in the chosen direction for "noon", and substitute the chosen direction for "North".
Keep the watch axis pointed to Polaris as best you can judge it.
For example: you know the Sun is West at 4 p.m. At midday it will look like this… (SH version)

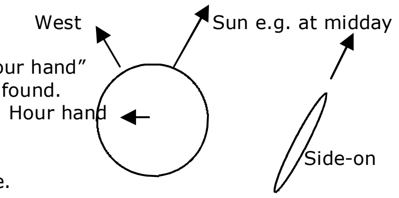

.."Point the four to the Sun; West (or the chosen direction) is between the Sun (or star etc) and the hour hand"
The catch here is that you must imagine the height of the body accurately, before the direction can be found.
This is because the method only shows the rotational position of the body — longitude wise, not its height declination wise — and with the clockface being tilted, the horizontal direction being pointed out is not obvious until you choose a height.
Of course, the accuracy of the method depends also upon aligning the clockface to the Equatorial plane.

String North

Carry a piece of string, about one metre long, to **measure the angular height of the Moon, planet or star** which you intend to use later.
You measure it once, as a matter of preparation, then use it many times. The Moon goes out of date the most quickly, but the stars are permanent

Put a loop knot at one end, to hold the bottom knot, level with the horizon or horizontal. You may have to wait, until the height is clear.
Hold the string at arms' length in a vertical plane to the North-South, which is where the star will reach its highest point.
Put a second knot to show the maximum height. This is equivalent to the 'noon' height for the Sun.

For the Sun you could use the shadow, and hold the string *down* from the horizontal. It will be accurate for a week.
Better still, use the shadow methods soon to be described, to avoid the temptation of looking at the Sun.

Now at *other* times of the day/ night, first hold the string in a vertical plane to reconstruct the expected height, then hold the string horizontally from that same high angle above you. Lastly, twist and turn, so that the horizontal string apparently points to the body used — Moon, star, planet.
You have to **twist yourself until the string is *horizontal*, and at the right *height*, and *points to* the chosen body**. Then you are facing N-S.
Actually the method approximates a slightly curved small-circle star-path with a great-circle straight course.
You can find approximate North with the string, then adjust it slightly by imagining the actual star-path as a small circle around the Pole Star.
The straight string will then appear to point to just above the star, but in your imagination, it curves down to match the guesstimated star-path.
This is the direct sighting equivalent to the 'shortest-shadow South' methods on the next page.

A 4 Week *Position Cycle* of Moonrise/set, to Find East/West

The Moon, like the Sun, lies close to the Ecliptic (in the Zodiac), and completes a cycle along it (within 5°) in 27¼ days (not quite 28 days). So...
The Moon changes position in **a four-week cycle, like *a speeded up year***. It spends a week in each 'seasonal' position: High; mid; low; mid; hi...
The Moon crosses the Equator every fortnight, rising due East and setting due West that day. Find it! Use p9c to find E/W on those days/nights.
In alternate weeks it will rise & set in the Midsummer then Midwinter positions *for the Sun*. Get to know them, and expect each one every 4 weeks.

- Its change of rise/set position day to day is most sensitive near the equinox position, due E/W (be careful), and is least critical a week later.

In between times, the rise/set **positions can be guessed by the day of the week**, by interpolation, as a good direction guide on any day.
The cycle runs fast by only 1 day in 37 1/3 days. So it is a continuous guide to E/W, **for a month at a time**, as it moves through 13 Zodiac signs.

- The best way to **set your Moon-watch is by observation**. If Moon-rise/set is too late at night, then look for Moon-set/rise during the day.

Sunset Full Moon is a good starting point for estimating the nightly-cycle; Sunrise New Moon, for the daytime. Midsummer or Midwinter or Equinoctial full Moon are very good starting points, because the Moon will be in the anti-Sun position. Visualise the Ecliptic (through the Sun and perpendicular to the Moon cusps joined) and its shape in relation to the Equator, and use a compass. Say the full Moon is opposite the Sun on Saturday, but by Monday the Moon sets due West. Every Monday for *up to* 5 weeks the Moonset will be in a recognisable position. E.g. W, N-W, W, S-W, W. On a later Monday, you will need to step back one day and say "every Sunday, for five weeks" the Moon will repeat those four positions. Adjust yourself by one day, *as it shows itself necessary*. You can simply say "It is a Thursday Moon at the moment", expecting to have the change to "a Wednesday Moon" in a fortnight or 2. Keep your finger on the Moon-pulse to keep up-to-date. E.g. Say: "It's a Spring-type Moon next Tuesday".

- To keep track of the weeks, here is a recipe: **Call the new year e.g. 'a Friday Winter Moon'** and adjust 1 day every 37 days, 10 days over 1 full year, or by **4/5 of the completed months**. 7, 8, 9, 10 days for *after* 7-, 8-, 9-, 10-ber, using the Latin names; 0-7 days while *in* the rest, 1-8.
- Perhaps you lose track of a 'Wednesday Moon' for a couple of months, and try to start again. Track the Moon's Motion along the Ecliptic.

The Moon overtakes the Sun at No Moon, gets one season ahead of the Sun within a week, at 'the first quarter', is opposite it in another week ('full'), and is one season behind the Sun in the 3rd quarter. There are 12 and a third months in a year, so the Moon is moving 12 1/3 times more quickly around than the Sun, in its seasonal cycle. **What takes the Sun one month will take the Moon only 2½ days**. (31=2.5x12.3).

Step 0: Visualise the cycle. — I.e. imagine where the **lines** of the Ecliptic and Equator must be in relation to the Sun and Moon.
Step 1: Place the *Sun* on the Ecliptic — I.e. estimate the time of year, and from that, the time before/after the next/last **solstice/equinox**.
Step 2: Try to **visualise where on the Ecliptic the *Moon* must be**, according to its phase.
Step 3: **Interpolate for the exact phase**, and label it a 'Monday Moon' or whatever. Check your guess with a compass if possible.

E.g. There was an eclipse of the Sun on Tues July 22 2009, 1 month after solstice. The 'Winter Moon' must have been 'Sat night'; and for 2009 'Mon'
E.g. 23 November is one month before the SH Summer solstice (2½ days worth of Moon motion through the stars). You can work out, for any phase, which phase and position to expect in 2½ days. For instance, Nov 23 is late Spring, almost Summer, Southern hemisphere (with the Sun South of the Equator and moving further South). The Sun is high and getting higher. Opposite that, is a *Full* Moon position (North of the Equator moving further N). It stays low in the N, getting lower. Or if it is a '*First–Quarter* half-moon' at that time, it will be in the 'late-Summer-almost-Autumn' position (approaching the Equator from the S, midhigh in the sky). Or if the Moon is a *third quarter* half-moon, it will be in an 'almost Spring' position (approaching the Equator from the N). You would therefore know that, taking the last example, over the next 3 weeks, a Waning Crescent/ New Crescent/ Waxing Gibbous/ Past-Full Moon, will rise E/SE/E/NE, and set W/SW/W/NW, in approximately 2½/9/16/23 days.

Sun Path Norths — Summary

- Noon-time: You look at your watch to see how far the Sun is away from apparent-noon-time, then look at how far your shadow must be from shadow-South. Near noon, the (untimed) shortest-shadow methods below aren't accurate, but this timed method can fill in the gap.

- Chord bisection: This works for any body e.g. the Moon.
- Sundial methods: You set the sundial to the right time. E.g. tilt the compass protractor to the equatorial plane.
- Shadow bisection: This can work for the Moon too. Just remember to use a vertical stick on level ground, e.g. a dry puddle.
- Equal-height: You can use an improvised sextant/quadrant/bubble sextant. Don't forget to bisect both the time and the direction, carefully.
- Sun at East/West: You need to be familiar with the height &/or the time when the Sun is East or West.

- Equinoctial sunrise/sunset: You keep track of how far off East/West the sunrise/sunset is, at this time of year.
- Equinoctial rise/set angle: You can extend the Sun's path to noon, rise or set, in a straight line, using the latitude angle.
- Equinoctial East: Where the shadow moves to, with slight adjustments at other times of year.
- The Noon-shadow path is always to due East. This is useful for anytime *near* noon, say +/- 2 hours.
- When the Sun is *very high* near noon, you *need* to see the direction of movement of any point shadow on flat ground — West to East.

Locating Anti-sunrise or Anti-sunset

Imagine a line from your head-shadow, *opposite the Sun's path*, cutting the horizon at the correct rise/set angle p.9a.
(The Sun's path is approximately a great-circle, so now imagine a plane, edge-on along that line, cutting the level ground.
Look for where the shadow would be closest to you. This approximates the Shortest Shadow South…)

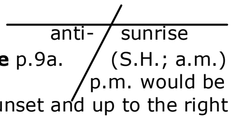

Shortest Shadow Souths ..from calculating, memorising or knowing the angle of high noon shadows.

- Walking-Stick North: You lean your walking-stick, so that at noon it would cast no shadow, and so that right now it casts a shadow at right angles to itself, a direction which is nearly West-East. Adjust (all these methods) slightly for the season, afterwards.
- Body-shadow South: You use your own shadow tip joined at right angles to your shortest-shadow tip (guess at its length and position). This works best in the early morning and late afternoon, where a guess as to your shortest-shadow is sufficient.
- Cardboard Triangle East/West: You cut it to point to high-noon height, like the walking stick & body methods combined.
- Plumb-line North: You tie a string triangle instead, to the right shape, and maybe attach it to your walking stick.
- Notebook North: You flip open one page from horizontal, so that it casts no shadow, and also points straight to high-noon height simultaneously. I.e. the page should be held at the correct angle, mimicking the Sun's maximum height.

You can discover the correct angle by observation, yesterday,
or you can calculate that angle as "latitude degrees minus declination degrees away from vertical"

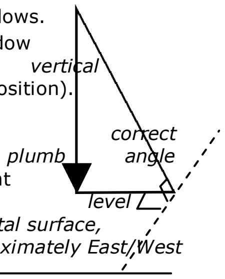

"Star-Time" Gains Four Minutes a Day

3 minutes 56 seconds actually, by appearances; (3m 56.555s if you use the almanac, and 'the first point of Aries', because "Aries" as it is known for short, is not a real star, and it drifts slowly through the Zodiac)
Stars complete one extra orbit — 360° — than the Sun in one year — 366¼ in 365¼ days — so that's about 1° per day which is 4 minutes of time.

By "time" we mean Sun-time "15° per hour on average", in ordinary usage — a slightly varying rate in practice.
By "star-time" we mean "15 and a bit degrees per hour spin rate — a constant rate"

The stars outrun the Sun, night by night, by 1° on average, so the Sun seems to migrate 'backwards'—to the east—through the Zodiac belt of stars.
You could say: **"The Sun drags its future after it!"**, meaning that the Sun sets in the West, but its future star-position shows up just to the East.

You can zero your star-time, by timing when a bright star (preferably one near the Equatorial belt) passes overhead, or north or south of you on a particular night from a known particular longitude. You then calculate star-time from "Sun-Time plus 3m56s per day since then" including part-days in your detailed calculation, at a rate of 10 seconds' time per extra hour — 1 second per six minutes.
Your reference longitude becomes like Greenwich to you — or, alternatively, you can adjust everything back to Greenwich longitude and Greenwich star time. You can also buy quartz controlled star-watches or calculators.

Adjusted Moon-Cusp North: To the Left or Right, By Moving the Sun

"Cusp-North" can be improved, by mentally moving the Sun away from the Ecliptic *onto the sky Equator*. Do the adjustment once a day. Guess the declination of the Sun, from the season; then shift it onto the nearest bit of the Equator and the Sun-to-Moon illumination angle may tilt a little.
Instead of pointing to the pole star for the Ecliptic, in Draco, the new cusp-North will point more to the pole star for the Equator — Polaris.
E.g. the midwinter Sun needs to climb 23½ degrees higher — which would tilt the cusps of the Moon (due to the shift in illumination). Diagram 1.
Caution: The illumination angle of the Moon does not change much when the *Moon* is shifted onto the Equator, and not at all for a half-Moon.
E.g. a half Moon (90° from the Sun) at Equinox could be adjusted by 23½° to bring it onto the Equator (It is like the Sun at Solstice), (diagram 2) but cusp North does not change. Being 90° from the Sun the cusps outline one and the same great-circle.

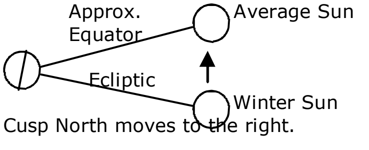

Cusp North moves to the right.

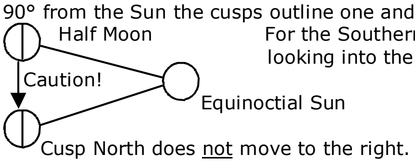

Cusp North does not move to the right.
Beware of perspective tricks

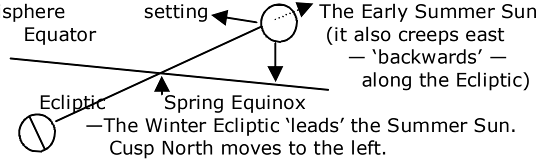

—The Winter Ecliptic 'leads' the Summer Sun.
Cusp North moves to the left.

"By the Nose" Navigation

Our Mental Map is a Nose-Eye View — not at all like the bird's-eye-view given by a layout map. (See "Strip Map Navigation" p.95c)
Oh yes, our mind eventually fills out a (spatially-solid) layout map, but initially we collect a set of (spatially confused) mental photographs from all different perspectives taken from ground level. Our mind doesn't deal in map symbols or contour lines!
Rather it has search images, image enhancement, image recognition, three-dimensional images, and memorised **relative angles**.
Consequently, a more natural method of finding your way is what I call 'Nose Navigation' — almost opposite to mapwork.
Nose navigation uses 'strip maps' (Section 4), relative angles, and a couple of special mnemonics…

"Because (my shadow) seemed to shift right just now; Therefore *I* must have turned left"

Now substitute in the (…): Ayer's Rock/Uluru; That distant cloud; The Sun; The Moon; Compass Needle; East Wind; The RDF Null; That bright star
In other words: "My Nose Turns EVERYTHING Away!" i.e. when I turn my nose, everything else seems to turn oppositely.

This simple rule attacks head-on your insistent illusion that you have been walking "straight" ahead!

And your illogical belief that "Ayer's Rock has moved"! Neither has your shadow 'turned right' just now, simply because your nose 'hasn't moved'!
It is not as easy as you think to apply, e.g. when you are reenacting where you thought you had gone, in your (already confused) head.
The second nose mnemonic is on the opposite page.

Seasonal Adjustments For Equinoctial East

Shadows move due East at Equinox, all-day. At other times of year, when sunrise/sunset is not due East/West, the Shadow path curves slightly at an angle away from sunrise/sunset, an angle which you should become familiar with each week. Start by assuming that the shadow is moving "East"… Verify this **Rule: Shadow Paths Bulge toward the Hot Pole**, even in the tropics, in both hemispheres. (The Sun crosses at Sept 22 & March 21.)
Adjustment: The first guess at **"Hot Pole" Needs to be Adjusted Slightly,** less near noon, little or none nearer equinox, to "un-bulge" the path.

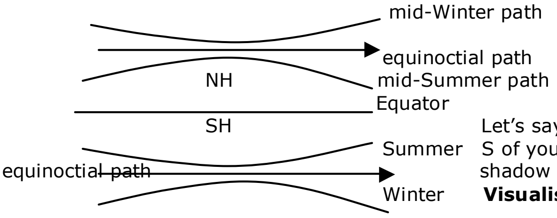

Remember that "the Southern Sun Sets Somewhat South in Southern Summer", so that opposite-Sunrise & opposite-Sunset need to come back South to point out E & W correctly
E.g. You guess East (or West in the a.m.) from the shadow's movement over 1 minute,
(It will roughly be opposite sunset/sunrise direction if you are late/early in the day)
then twist your guess a little back to the true hot pole, and to due East (or West).
Let's say it is midwinter in the Southern Hemisphere, so your shadow is somewhere S of you. The 'hot pole' is the North Pole, so twist 'North', as gauged from your shadow tip, away from the shadow, back toward you (i.e. North, to the hot North pole)
Visualise the bulge, then straighten it out at each end, back toward the hot pole.

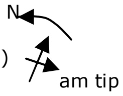

North-Finding from Two Paper Triangles

or even from only one

Cut out a back-to-back triangle shape and fold it along the join. Line it up as shown and 'N' will point to Polaris…

'1' is your South **latitude** (For the Northern Hemisphere use the North Latitude and swap N & S)
'2' is the **polar distance** of the Sun (from the 'depressed' pole, N)
i.e. its co-declination, or 90°-N dec. See p. 18b.
You can set it by observation or calculation; and can fold it to the correct angle, rather than cutting it.
('3' is a right-angle, but '1' and the horizontal is what really matters)
This two-triangle method works best away from noon, but nearer noon simply find the way shadows move
The 'N-S' edge is parallel to the Earth's axis and acts as a hinge around which the Sun revolves.
Sun and shadows move at right angles to that axis.
You only need **one** of the two angles if you find the direction of shadow movement
and swivel 'N-S' to be perpendicular to that. (You don't even need one angle, near noon)

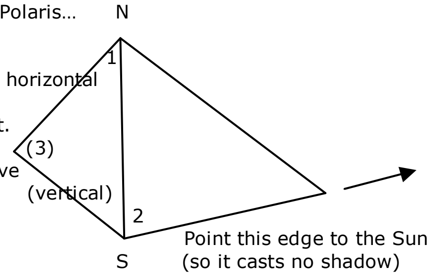

Point this edge to the Sun
(so it casts no shadow)

E.g. Poke a straight stick into the ground, pointing at the Sun, casting no shadow; (Or use any point-shadow rotating around its point)
By waiting a short while, find the direction of movement at a half-way point, in mid-air, i.e. maintaining an equal distance to the tip,
and manoeuvre your 'N-S' axis (index finger), to be perpendicular to how shadows move, then at the correct angle to the Sun and/or Horizontal.

Working the Star-Longitude Numbers

The reference star point, a non–star, is 0°N/S;0°E/W, and is called "The First Point of Aries" or "the Vernal Equinox" where the 'northern spring' Sun crosses the ecliptic (on March 21). Other stars are measured from that (imaginary) star, like latitude and longitude are measured from 0,0 a nameless spot of ocean — each star has an almost fixed longitude in relation to Aries.
N/S 'latitude' is called +/- ' Declination '
E/W 'longitude' is called (a) Right Ascension — R.A. — or 'Star Time' — It is measured **Easterly**, in hours at 15° per 'hour', from "Aries"
Or (b) Hour-Angle — H.A. — It is measured **Westerly**, in degrees, like longitude, but 360° around, from "Aries".
The first, RA, is designed to tell us the time — **when** the star will show — since all the stars move around westward with Aries.
The second, HA, is designed to track the West **longitude** of the star in reference to the Earth, as the star progresses West around.
We need them both, but they are entirely equivalent to each other.
(The only trouble is that a star 'hour' of 15° is slightly faster than a Sun hour of 15° of progress to the West.)
Next you need to know that the first point of Aries was at 100° W. Earth longitude at the beginning of the millennium,
and at every turn of leap-year thereafter, at January 1.0 (midnight). Then it gains at a rate of 1.002737909, or about 1° per day on the Sun, (3m 56s time), so that the stars turn full circle during four seasons. "1°" is adequate accuracy in the short term for us but not for yacht navigators.
E.g. January 31.5= 100+30°+0.5° W., then you add the H.A. (or subtract 15 x the R.A. hours) for a particular star — that's the longitude where the star will be overhead. **For exact numbers consult an almanac**, but I am concerned here with giving you an overview.

Topographic-Level Winds

This is what you feel in your face at ground level, or close to it.
Fog; Smoke; Valley Winds; Sea Breeze; Eddies; Catabatic Winds; Forest Winds; Scree-Slope Winds; Mountain Winds; Glacier Winds; Lake Winds.
The various thermal winds listed will of course be very variable. The thermal winds flow from cooler areas, before rising in the warmer region.
Nevertheless they may betray the slope of the land, for instance at night. For example, fog forms on calm cold nights then drifts slowly downslope.
Being dense, fog will drift down-valley, down-stream, and off-shore. You can see the location of creek outlets from the early morning beach fog.

The ground-winds due to pressure systems, are affected by overall friction and therefore average out slower and less Sun-twisted than upper-winds.
Ground-level winds form waves, foam 'streets', sand ripples & dunes, drag marks, snow ripples & drifts, which all may persist beyond the winds, and they cause lopsided bushes and salt-spray-'cut' bushes. See Section 3 for more details.

Gusts are at different directions to lulls since upper winds penetrate to lower layers.

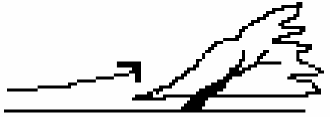

"Nose-Navigation Needs an 'Anti' Mnemonic": "Nor-West, Not-Least"

N.O.R. West: "North **on Right** = Going West" or "Needle on Right = Going **West**" i.e. going into the westerly **half** (top diagram)
N.O.T. L.East: "North **on the Left** = Going East" or "Needle on the Left = Going **East**" = to the easterly **half** (bottom diagram)

This is always true when you are relying on "Needle North". **You don't have to think any further.** But,
It is equivalent to ANTI-CLOCKWISE BEARINGS, which you can't read from a clockwise-marked protractor!
The following numbers refer to an ordinary compass — marked clock-wise, and set to 0° as straight ahead
(not North) so North swings left and right instead… The compass needle might point to "045° NE"
Example: in the rain or fog; the needle swings meaninglessly but your nose is pointing NW
Example: deep in a cave network
Example: on a yacht, with a fixed hand compass The compass protractor numbers say "North-West, 315°"
Example: with a fixed car-compass, not marked anticlockwise but your nose is pointing NE
Example: making a strip map; or using one
Example: using a compass-holder, which fixes 0° as straight ahead Interpret "Anti-clockwise, West from 0"

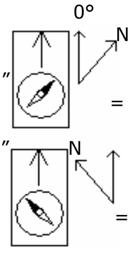

Needle on the Right
use "NORWest"
= North-West N045°W
Needle on the Left
use "NOTLEast"
= North-East N315°W

Revision, Overview and Explanations

Top Down Orientation

The Big Bang **Universe** bloated over billions of years; and broke into bubbles of turbulence; all bits in each bubble orbiting each other; in millions of galaxies, with 20 or so galaxies in our little **local cluster**;
Local galaxies, like the Magellanic Clouds or Andromeda, glow milky (just as 'galact-ose' says 'milk-sugar').
Our galaxy is a great spiral, swirling slowly, anti clockwise from above i.e. Eastwards around its North-South spin-axis — flat like a fried egg — a 'great circle', glowing as the Milky Way. The thickest milkiness, in Sagittarius, marks the Galaxy Centre.
The Milky Way is like the 'Equatorial plane' for the whole spiral, seen from within, with its northern pole star in 'Berenice's Hair' (28° N. 168° W.) and its Southern pole star in 'Sculptor' (28° S. 348° W. 012° E.) and our milky galaxy is, in its own turn, full of spinning, gyrating gyroscopes.
You know how a spinning top will pirouette with a slow-motion gyration when it tilts slightly? Well, precession, as it is called, causes slow changes.
Stars are Suns, especially shiny in our own spiral arm of the galaxy (i.e. closest) and seemingly set solidly in space by their extremely slow motion. So the constellation shapes are chartable for centuries. Their apparent positions are gridded out in relation to the first point of Aries, which in turn is fixed each year by our Northern Spring Equinox—not a star in the sky, but a point which creeps along the Ecliptic, to the West.
Hint: When you are wondering whether a rule holds in both hemispheres or not, ask yourself whether you are down at the level of the spinning Earth or up beyond that in a larger scale. Try to imagine looking at the whole globe from an outside perspective. E.g. the appearance of the Moon — its phase, eclipses, tidal forces, etc — is mostly independent of where we look at it from Earth — and will be the same for both hemispheres.
But it's apparent path — 'noon' direction; rise/set angle; etc — is seen down at Earth level — it changes with the hemispheres.

Closer to Home

The solar system is also spinning flat, anticlockwise, but slanted — tilted like a top, in relation to the Galaxy, and faster.
Its 'top' star is currently in Draco, in the North. Its bottom star, in the South, is the larger Magellanic cloud.
Its 'Equator' is called the 'Ecliptic' — marked by the stars in the Zodiac, and the planets. In our lifetime we won't notice the Ecliptic stars change.
Our Sun circles the centre of the galaxy every 225 million years—heading toward Suhail.
To our sight, because we orbit it Eastwards, i.e. around its N/S axis, not ours, it swings seasonally, and also edges Easterly along the Ecliptic circle, revealing new zodiac stars to the West of it, and placing itself diametrically opposite a series of anti-Solar stars,
so giving us a star calendar to look for.
It also spins easterly, on its own axis, showing its sunspots for some days at a time.
The planets lie planar, 'wandering' (as the Greek word plan-et says), in the Zodiac plane, circling easterly around the Sun, over a timescale of decades, and only ever seen very close to the Ecliptic.
Mnemonic: Son, my very earnest mother 'as just served us 19 platefuls. Sun, Mercury, Venus, Earth, Mars, asteroids, Jupiter, Saturn, Uranus, Neptune, Pluto.
The earth is an inner planet, **circling the Sun**, easterly, so that the stars seem to out-run the Sun in setting West — star-time spins more speedily than solar time. Planets closer to the Sun than we are, are never seen far from the Sun, but outer planets can move right away from it.

The Moon

..moves mostly on the Ecliptic but might miss that mark by as much as 5°. This 5° slant causes a precessional change taking 19 years to cycle.
The Moon moves around the Earth monthly in an easterly ellipse,
and ever eyeing us, each full Moon, is the Rabbit, because the Moon also spins easterly, on its own axis, exactly once a month, so it stays facing us.
The Moon moves slowest to set Westwards, since it migrates eastwards in the sky each day. 12°½. Or **50 minutes later per day**.
And as it pulls the tides with it, they too are later each day, almost an hour, on top of normal locally fixed delays due to friction.
A full Moon rises and sets fully opposite the Sun, give or take 5°.
The Moon is most responsible for tidal forces, but the Sun plays some part too, reinforcing or diminishing the Moon's effect and changing the timing.
Tidal force attracts the nearest water the most, and the farthest water the least, resulting in two tidal bulges...
When the Moon is eccentric from our Equator, there is one big Northern bulge every 24 hours,
and a smaller high tide every 24 hours, as the Earth spins through the bulges.
There are equal bulges every 12 hours on the Equator. In between are the low tides, every 12 hours.
You should be able to tell which bulge is 'nearest' the Moon despite delays due to friction.
Southern latitudes will have one big bulge every 24 hours, and one 'small high tide' every 24 hours, offset by 12 hours from the NH.
Try to correlate the Moon's phase, declination and height with the local tides—timing, inequality, range—e.g. "The Full Moon rises with the tide"

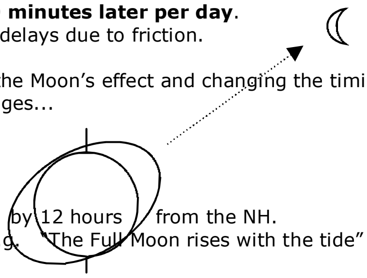

The Moving Horizon

Since it is the Earth which is spinning, not the Universe, it is **the horizon** which rises, in the West — where it eclipses the Sun,
so that the Sun only seems to set in the West.
The horizon slumps in the East, so that all the sky things — Sun, Moon, planets, stars — seem to 'rise' in the East.
Any spots on the Northern half of the horizon, circle to the east around Polaris — clockwise through the sky — i.e. 'seen from below'. Or, seen from above, that is from the North, looking South, spots on the Southern horizon circle anticlockwise around South Pole Star i.e. seen as if 'from the top' looking down on the Earth and its South Pole star. They hit the same stars each night, at the same angle, which is related to latitude,
so that the fixed stars none-the-less seem to trail 'small circle' tracks through our sky, and they always hit their favourite horizon spots when they set, always at the same angle,
and those same stars will rise later in mirror image to that around North/South, giving rise to **north-south-east-west finding by bisection**.
And Equatorial stars will rise/set due East/West.
The southern-hemisphere horizons move to the right, the northern-hemisphere ones to the left. Looking South though, from either hemisphere, it will be seen as the horizon circling anti-clockwise around the South-pole star. How come? Well, the southerners see the bottom of the star-circles, around an elevated pole-star, the northerners see the top of the same circles, around a depressed pole-star.

The Spinning Earth

The earth is **spinning like a top**, on its own axis, easterly, half as fast as a '12 hour' clock.
This gives us **the single most dominant direction clue on the otherwise uniformly round Earth—the North-South spin axis of the Earth.**
Spot its Equatorial Plane by the Pot. The 'top' is always pointed to Polaris. Spot the South Pole Star by the Seven Superstars surrounding it.
The spin generates a magnetic core more or less North/South — Find magnetic North in far northern Canada.
The magnetic field poles generate auroral displays, anywhere near the poles.
The earth is also tilted like a top, at 20 odd degrees.
The tilt gives rise to seasonal Sun height shifts, somewhat like a sinewave (just imagine the Earth's tilted spinaxis not changing as we orbit the Sun) and it turns on un-twinned top-of-the-tide bulges, twice in 24 hours. The Pole Stars are latitude° down/up @ 0°/180° T.
Any tilted top will precess the equinoxes — this generates a precessional gyration of 26,000 years,
but this time, retrogradely, i.e. to the West, along the Ecliptic (3° in 200 years) — i.e. the stars 'move' to the East through our star-map-grid.
Longitude has a lot to do with spin-time, seen from above the spin-axis. E.g. West longitude is how long after Greenwich, a place meets the Sun.
When we want to mark the apparent place of the Sun in the sky, or anything else, at any particular time,
we can just record the latitude and longitude of the point on Earth directly 'under' it i.e. closest to it.
But visualising that in reverse, we simply **project the Equator and poles, latitude and longitude up onto the heavens to grid them out**.

Time is counted in cycles

The larger the distance scale of any motion — pulsation or precession, orbit or rotation — the larger the time scale of the cycle of change.
The Universe has been expanding its size and condensing its matter for many billions of years (and perhaps it will not 'cycle') It appears static to us.
Geological history spans billions of years, like the existence of our spiral arm. In our lives we see only a frozen frame of our spiralling Galaxy. It cycles, but we can't see it. The timescale of the rotation of the galaxy is in hundreds of millions of years.
In our lifespan we see only the slightest change in relative star positions, through telescopes, due to the 'proper' motions of the nearest stars.
It is in the nature of spiral turbulent motion that the larger rotation replicates itself in the smaller scale, but at a faster rate.
Each of the components of our solar system are also rotating and orbiting in the same rotation sense as the Galaxy. (Uranus is an exception)
For example, the Earth is rotating eastwards, orbiting the Sun eastwards, and has a Moon rotating eastwards and orbiting us eastwards.
"Anticlockwise, as seen from the North" is the single best spin you can put upon all that you see—including the N/S Earth spin-axis as 'the single best direction clue on our spherical Earth'. The tilt of the Earth's spin-axis gives us the year, and the Moon's orbit gives us the month, independent of our rotation. The Earth's rotation gives us our day. Star time outstrips solar time slowly. The Southern Cross clocks it correctly at the close of March, and the Centaur soon follows. The first point of Aries gains a touch over the stars, migrating 50 seconds of arc along the Ecliptic per year. The vernal equinox arrives 59 seconds late for its 366th "orbit", after 365 days, requiring 24 leap years in 100 years. 365.242 days per year.
Atomic time, used for astronomy, is independent of all these tilted rotations, orbits and precessions, which are all independent of each other too.

The Precession of the Equinoxes

The actual or 'proper' motions of the stars are only of cosmological interest, because they change relative positions only over centuries.
But history, actually our slowly gyrating Earth-spin axis, rewrites our star charts (**im**properly, as it were) with precessional revisionism, so that stars 'move' slowly easterly through our star-map-reference-grid—an Earth-centred way of seeing the sky — with 26,000 years needed to complete a full circle precessionary Earth-wobble. This artificial motion is called "the precession of the equinoxes"— which means that our reference grid readjusts just as quickly as our earth axis of spin precesses — by five arc minutes in six years. That's why we need an almanac to predict the 'fixed' star positions!! A good analogy would be the slow drift of the north magnetic pole, **if** we set our latitude and longitude by it — that would yield a false or 'improper' 'continental drift' in addition to the proper continental drift. See the appendix at p. 123.
To visualise the Ecliptic, consider the Equator in the heavens to be geostationary, although invisible, but the Ecliptic to be like a sinewave moving daily to the West along it, on the video display unit of the sky—as on a cathode ray oscilloscope—with a period of 24 hrs & an amplitude of 23°½.
The Sun appears fixed somewhere along the Ecliptic, creeping 'backwards' on it (against the setting sky)—to the East—during the whole year
To visualise the precession of 'the Equinoxes', imagine that *the points where the Ecliptic sinewave crosses the* Equator (the x-axis), move 'forwards' ('with' the setting sky) — to the West, very very slowly — during the millennium. At the moment these 'equinoctial points' are between Regulus & Spica (the Autumnal Equinox), and between Diphda & Markab (the Vernal or Spring Equinox, or the 1st point of Aries, (0,0) for the stars).
You might say that the Sun and the Moon and the outer planets all move West through our sky but East through the starry heavens.

Global Weather Patterns

The spinning globe makes the climate zones, and weather belts, follow the latitude bands. When the wind tries to blow, from a High, into a Low, in the Northern Hemisphere/Southern Hemisphere you see huge anticlockwise/clockwise spirals instead, by Coriolis force. E.g. the wind, tries to go straight ahead from the hot spin-equator to the cold spin-pole while the spinning Earth slows down beneath it (it moves slower nearer the Poles).
In the North the wind therefore seems to "move to the right" as we see it, and in the Southern hemisphere "to the left". Circulation is the result.
Note: Winds around a Low go against the Coriolis force. To visualise the pressure system winds at low level, imagine placing two flat-headed screws pointing up, one on Japan and one on Sydney, an ordinary right-handed screw for Japan but **a left-handed spiral screw for the Southern Hemisphere**. Run your fingers over the threads and a right-hand thread is like screwing a screw in or out, a left-hand thread is opposite.
High pressure system winds screw downwards and outwards; Low pressure system winds reverse that—up, converging.

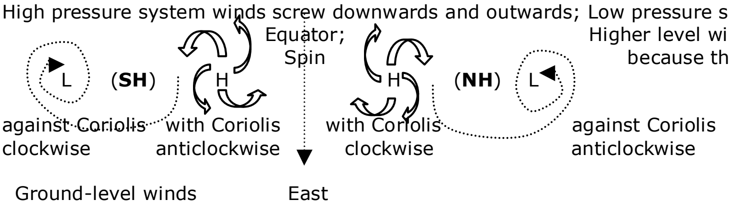

Higher level wind circulation reverses itself from that because the convergence/divergence is reversed.

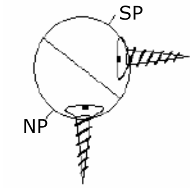

The Spherical Earth

The earth is curved round, like a globe, 360° round — 40,000 km around (by definition of a kilometre) 21,600 nautical miles, by definition.
◆ at 400 grads in a circle that's 100 km/grad ◆ at 360° in a circle that's 111.1 km/degree.
The Equator is a 'great circle'. Longitude circles are 'Great Circles' — passing through opposite poles, and perpendicular to all latitude lines.
Longitude counts 360° around East or West, or 180° East and 180° West from Greenwich.
Its North/South dividing lines mark out 24 one hour time zones of 15° each, and those North-South lines must converge on the poles, so that longitude degrees are of unequal arc-length at different latitudes.
Latitude has 90 equal degrees from the Equator to pole around the surface, which, by definition of the nautical mile, as one minute of arc, makes 60 nautical miles per degree of 'arc-distance'. 21,600 nautical miles must be the same as 40,000 km, so that's 10 km to 5.4 nautical miles. (2 to 1)
Latitude circles are 'small circles'.
We use them to divide the polar regions from the temperate regions (with the Arctic Circle, for instance, marking the limits of 24-hour nights); and to divide the temperate regions from the tropics (for instance with the Tropic of Capricorn, which marks the limit of 'high noon' fully overhead).
'Polar Distance' complements the 'Latitude' and reminds us that arc-distance is real distance.

Moving the Horizon

At the spin-equator, on a globe, the polestars must lie horizontally, and the Equatorial spin-plane must pass East/West overhead.
But move north/south to another latitude, and the earth being round, the horizon — which follows you, like 'up' changing with you — must fall and rise latitude-degrees, so that one pole star falls latitude-degrees out of sight behind you and the other one, ahead, elevates latitude-degrees above the horizon. The Equatorial plane must also fall latitude-degrees, behind you as you move, from being overhead. Other star-latitude (declination) circles in the sky (e.g. star-trail circles) must also rise or fall the same amount — latitude degrees.
Now this same thing happens if you move in **any** direction, say ex degrees away, but for this let's just temporarily freeze the spinning of the globe.
'Horizontal' and 'up' must follow with you exactly ex degrees, and so the horizon now blocks out ex degrees of sky behind you, and reveals ex degrees of new sky ahead of you.
Looking side on, you will see the horizon **tilt** by ex degrees in relation to the old skyscape, lifting behind and falling ahead, of your direction of travel
Looking at the same star-time any night, will "temporarily freeze the nightly spinning of the Earth" for you, for star navigation.

Spherical geometry also means that if a place is 2° or 20° away from you, then it must also be 1° or 10° below your horizon — half the arc distance
E.g. the North Pole on Earth is ½ x (90° minus latitude°) below horizontal, and at 0° True.

END OF SECTION ONE

Navigation is More Than Navigating

Properly done, navigation is your chance to fit yourself in to the experience of walking through the land.
Why are you there walking?... Navigation zeroes in on the land. Do you want to see the landscape?... Navigation opens your eyes.
Don't you want to appreciate ecology and geography?... Navigate by them! Navigation stops you "drifting off" into mindless incompetence.
Do you want to remember the views and special places?... Navigation forces you to improve such memories.
So, achieve confidence when you are alone; tackle more difficult trips; lead others; call yourself "experienced"— navigation is the key.

Common Sense

Navigation skills invite you into the wilderness. The wilderness is a wild place! — a world of dangers. **Don't go**. But if you do...
..at least make sure your navigation skills are up to the task.
Likewise: Don't venture to sea with approximate star navigation techniques until you know exactly what you're doing.
Learn much more about standard astronavigation, before you experiment with new techniques.

Your Subconscious is Superior to Science!

Your natural abilities will do a better job at keeping you oriented. Your brain is a supercomputer.
That makes it subject to the "rubbish in; rubbish out" truism. Your job is to feed relevant data to your subconscious, not to feed, say, a calculator.

As far as important detail goes, your mind will 'go for the jugular'. Exact science will go for unimportant decimal places.

E.g. I can immediately guess "about 8 km away" far more quickly than any triangulation.
E.g. Will any numbers ever tell you how to recognise where you are? Or how to imagine what that peak might look like from behind?

So feed clues into your brain. It is portable, waterproof and teachable.

Don't leave your mental map behind.

Visual Assessment is Your Main Clue — Rely on It More

Since noone navigates well when they "can't see where they are going", we must be relying on visual assessment most of the time.
The idea is to run with that fact of life, and make sure you have a good look when you can
(e.g. from the lookout; while driving up; at the crest of a ridge; at the waterfall; on the bridge; in the riverbed)
Further, bring to conscious focus what is normally subconscious, in what you see.
And aim to remember what you saw. (Draw it; take notes; talk it out; use visual memory skills)

A What-to-Look-for Checklist: How far is it to the skyline? Where are the Sun, sunrise, Moon, wind and North in relation to the skyline?
How does the weather interact with the land — wind, cloud base, Sun shadow, wind shadow, cloud shadows, haze?
Talk the skyline through — name its features. What do you recognise already? What does it remind you of?
What landmarks are visible? What distances to landmarks? What sizes of hills? What is the strike or trend of the ridges, ranges & valleys?
What angles between landmarks? What alignments? What valleys will your current path take you through?
What lies in between? What slopes & shapes are showing, especially for your route? Can you see the path ahead? Is it curved?
What difficulties lie ahead? What easier paths are there? What clear lines of sight are there? What alignments will become visible?
Can anyone see where you have come from? Could you recognise 'here' again? Can you see your destination or departure point?
How many different vegetation types? How high is it to the skyline in each direction? How does what you see mesh with the map?

Integrate: Jetstream, Earth Rotation & Heaven Rotation

Imagine your global motion. Tilt yourself and try (futilely) to oppose the supersonic speed of the Earth's rotation, as if walking upon a treadmill.
Aim for a Big Picture 'feel' for each local direction.
The idea is to go outside and walk through each of these motions until you **know** what it feels like.

(Both hemispheres)

When you are............	**standing still**	**or moving Westwards**	**or Eastwards**	**or Northwards**	**or Southwards**
(read down these lists, each in turn)					
The **Earth** rotates	Eastwards	rotates against you	rotates with you	to your right	to your left
The **Heavens** move	Westwards	slowly forward	quickly upwards	to your left	clockwise
The **high clouds** oppose the heavens outrun the Earth	jet Eastwards	zoom toward you	creep ahead of you	move to your right	to your left
You are being **tilted**	to the East	backwards	forwards	to your right	to your left
Lean oppositely	to the West	forwards	backwards	to your left	to your right

---

Measuring angles at 1 cm per degree

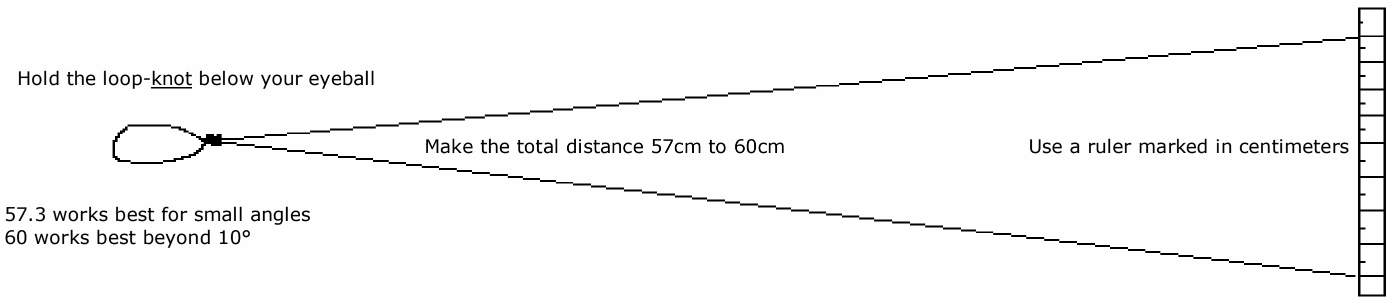
Hold the loop-knot below your eyeball
Make the total distance 57cm to 60cm
Use a ruler marked in centimeters
57.3 works best for small angles
60 works best beyond 10°

Orientation = Direction Sense

Oriental means "the rising" i.e. in the East (i.e. East from Europe!). Occidental means "the setting" i.e. in the West.

Because most ancient temples faced the midyear sunrise and there were no North-pointing magnetic compasses, and since all the world can agree on where equinoctial sunrise is, most of the rest of the world has used East, for most of Time, as their method of "orientation" — as their cardinal direction. Imagine putting East at the top of maps, so the Sun "falls down" the page.

Before magnets, East was the sensible choice for a standard direction.

This is no coincidence: "To the East" is also the watchword for the spin-sense of all our solar system parts. "North" and/or "South" is inferior

The direction of the *axis* of rotation doesn't tell you the *sense* of rotation; But the *direction of rotation*, gives you the axis of rotation *as well.*

East was the more sensible choice. You didn't lose orientation when you crossed the Equator; the Sun still rose in the East.

I recommend that you concentrate on this when changing hemispheres, and even now, to maintain your global 'orientation'.

Orientation means direction sense, direction finding, direction recognition, placing yourself in relation to over-the-horizon landmarks,

Or twisting anything (e.g. your map) to the right orientation.

Then again, it means starting off by getting your bearings straight.

Orienteering means athletes running through the bush.

Disorientation means... possible death.

The Navigator's Top 10 Problems; and Band-aid Solutions

1.	Not staying **alert** to navigation; not doing all you need to do……	**Talk** navigation; love navigation; practise navigation; **interpret** all clues
2.	Few **Clues** — E.g. Rain/Fog/Blizzard/Nighttime/Tall grass or dense vegetation/Look-alike landscapes	**Sit** still, until you can see; **go prepared** e.g. with a contour map; a compass; learn how to keep consistent courses
3.	Getting **Lost**	Remember the **start direction**
4.	Getting **Disoriented**	Be able to reorient; imagine the **Sun** in all situations; work top-down
5.	**Forgetting** the Way, the bends, etc.	Take a **notebook** and pencil; talk it out loud
6.	Losing track of **direction**	**Predict** the trend; use a **compass**; a notebook; integrate all direction clues
7.	Losing track of **distance**	Predict the timing; use a **watch**; use the map scale; and notebook
8.	Inadequate **Map**-Work	**Learn** how to use a map; study it first
9.	Not learning how to navigate	Take this booklet. **Practise.**
	Not knowing what to look for e.g. bush north	Practise being observant, at the expense of other preoccupations.
	Not **understanding** e.g. the lie, contours, heavens	**Teach** what you do know
10.	Choosing a wrong or difficult **route**	Get **directions**
	E.g. a long "shortcut" across a dense gully	Don't 'shortcut'.

Reorientation

Reorientation is not an automatic body sense — it is an ongoing action.

"Orientation" in navigation is a misnomer, since it suggests some sort of initialisation, not to be repeated, rather than a continuous reprocessing.

"Re-orientation" is a more accurate description of what the navigator is involved in. Continuous reorientation.

Whether it is subconscious or conscious, it is something which must be done and redone, or else you are simply not oriented, now.

It is not good enough simply to be "good at navigation (theory)" — you must be putting it into practice, continuously, or else each bend will disorient you.

The essence of reorientation is conscious interpretation of the meaning of what you are doing.

This is, at the best of times, a bit difficult — like mental arithmetic — but it must be done, or else your brain will remain confused.

You may hope that direction sense would "become automatic with practice", but the last step of conscious interpretation must remain conscious.

And that is the process which should become second-nature. You must be aware, or you are fooling yourself into a false sense of security.

Delegating to your subconscious, is abdicating your responsibility. So bring yourself up-to-date, continuously, and...

Do not be overconfident. You are not oriented when you think you know where you are and where other things are—but only when you are right!

There are so many ways to make silly mistakes: confusing alternative systems; mistakes in numbers; forgetting your lessons; optical illusions; instrument errors; flawed theories; trusting approximations; misidentifying a star; making a logical blunder; and, of course... being overconfident.

Lost? Rule One: Stop! ..getting More Lost

STOP — so that you don't get deeper into trouble. ("When you get yourself into a hole, stop digging")

DON'T PANIC — Look for reasons not to panic "I can always walk all night, or sleep under that ledge,
or wait out the weather, or go hungry, or enjoy it…"

PRAY — He really does want to help the helpless.

SIGNAL — Shout 'Help!', 'Cooee', Whistle, Smoky fire, Heliograph, 3 of anything, SOS, CQ, Fire at night, Flashlight

LISTEN — for a reply.

THINK p. 37d

RETHINK — your journey

SIT It Out — = The Simplest Way to Beat the Fog. Gather fuel for a fire. Don't try to move — you may not win.

CAMP — choose a place that will be a good 'base' for further exploration.

SLEEP ON IT

STAY — in the area, where you can be found, e.g. on the track, at a hut or intersection.

PLAN — Implement a pre-**thought-out plan** e.g. Stop, Think, Navigate, Camp, Retrace, EPIRB, Stay.

For a Start: Slow down — almost to a stop. Be slow, cautious and thoughtful, not impulsive.

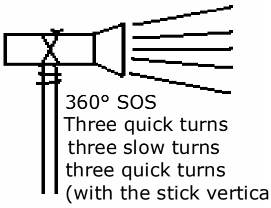

360° SOS
Three quick turns
three slow turns
three quick turns
(with the stick vertical)

Point and Blink = 10% Shift

In the field you simply point and blink—If you focus in the distance, the shift is about 6 degrees, which is 10% sideways.

6° or 1 in 10 is a very useful angle to guesstimate. Point with one eye and finger, then swap eyes, to see what a 6° shift looks like in the distance.

For repeatability you should touch your two index fingertips as you point.

For accuracy, use a pencil tip, and know the exact figure. The instructions below will show you how to hold the pencil to get an exact result.

To put the Point and Blink method on an accurate footing, measure your Inter-Pupillary Distance at Arm's Length.

Stand behind a window, pointing through it at some distant mark on the horizon, with both your finger-tips just touching the windowpane.

Use one eye and the same-side finger; Press your fingerprint onto the window; Swap eyes and fingers, back and forth, moving only your eyelids;

Check your alignment—both your fingertips should appear to exactly overlap when focussed on the distant object. Press your second fingerprint onto the glass; Measure the separation carefully.

Because the pupil is not pinpoint, it matters that you measure the practical separation at the working distance. Measure that working distance carefully— from the corner of your eye to the window when your pointing fingers just touch it.

Compare the two distances; i.e. divide the interpupillary distance by the armslength; **expect about 1:10** or 6°; Find the arcsine of that ratio for better accuracy. Try it out on the length of the Southern Cross. Compare it with 3 fingers, and 4 knuckles, and your binoculars' field of view.

Measure again on a middle-ground object, as when measuring the sideslip of a yacht by pointing at the wake. Measure out the distances as a check

The “Good” Navigator. You.

If you like exploring, you can use that to advantage. Or maybe you are a good photographer, well then you will understand ‘mental photos’.
Use your creative writing streak, in creating memorable names, or else, appreciate the action ‘direct’, carefree and spontaneously.

Perhaps you are the obsessive type — well then, it’s map and compass detail which will suit you.

So whether you are the artistic type, using your visual dominance, or the cool calculating logical sort,
or maybe a ‘thinker’ doing a good job at planning, or a ‘doer’ doing, you already have many useful skills for navigation.
Don’t talk yourself down.

Those handful of useful traits which are already within you, you can develop into an orientation sense.

Think out Loud

Your subconscious will be telling you “That’s odd”.
Your conscience will be whining “Something is wrong”.
Your mind will notice, before you do, “That’s new”.

When you notice something talking inside you, you need to talk back. Out loud. Don’t leave it ‘inside’.

Internal Problem	**Antidote**
“Shouldn’t we be there by now?”	Ask the others.
“Have we missed the junction?”	Discuss it.
“Why is the Sun over there now!”	Name the lurking enemy!

You Need Excuses to Stop and Have a Good Look

- If it is a local peak, say “Let’s look for a minute while we have the opportunity”
- Declare a chocolate stop!
- Take off your pack, adjust the padding. Adjust your socks, air your feet, put on a Band-Aid.
- Get out the maps; Identify the landmarks. Draw the relative angles to other lookouts, for future reference.
- Get out the camera and tripod. Unpack your pack on any pretext — noone can move until you repack!
- Make a sketch, compose a poem. Start an interesting conversation.
- Go for a swim or wash some clothes.
- Cook your main meal at midday & keep the biscuits for the late meal.
- Camp there.

You Can’t Easily Revisit Places Left Behind — unless you first spend memorable times there.

Visualise Familiar but Invisible Landscape in Relation to the Sun

You need to ‘fill in the void’ between the very distant reference, and what you can see, with what you know, but can’t see.
I.e. When you get a clear view, mentally plot and remember the sky-to-land layout for later use. “If the hill disappeared, the hut would be *there.*”
Mentally reconstruct any view which is helpful. E.g. as seen from camp, or from the start, or the peak…
Concentrate on knowing and revisualising landscape direction trends. “If I were at home, I’d see the coast… *across there*”
Do this at various times of day, **paying particular attention to direction *trends* in relation to the Sun’s path**.
After all, that’s what our direction sense consists of—a confidence in the real directions just out of sight, based on what we can see.
Then, when you are ‘bushed in’ and you see the Sun through the canopy you can say…
“Ah! the road is running sideways to the Sun, the valley lies entirely off to the right, I must have swung around left, parallel to the ridge…”

Section 1 dealt with the Sun/noon/North in the global abstract—unrelated to the local detail.
Section 3 deals with the local detail in relation to the Sun.
“Land to Sky” (soon, p 38d) is specially for one time of day, and concerned mainly with particular landmarks, not whole trends.
But now catch up with the overall game—**the Sun’s path, the local skyline, and the local layout are one piece for a substantial local area**, and you need to mentally integrate them, and carry them with you, to avoid disorientation.

Guesstimating Angles at 1 degree per cm

First find how to make your **arm’s-length 57 cm**, e.g. by touching your fists together in front of you.
Next, start measuring in **centimetres**.
Find a 1cm wide fingernail—1°. Check that your fingers are 2° wide, for 2°, 4°, 6°. Are your knuckles 2 cm apart? 2½° for the big ones.
A small fist, with the thumb tucked in is 9°; 10° with the thumb showing; a palm is one ‘point’.
15° and 16° may require experimentation. 20° can be spanned with the fingers.

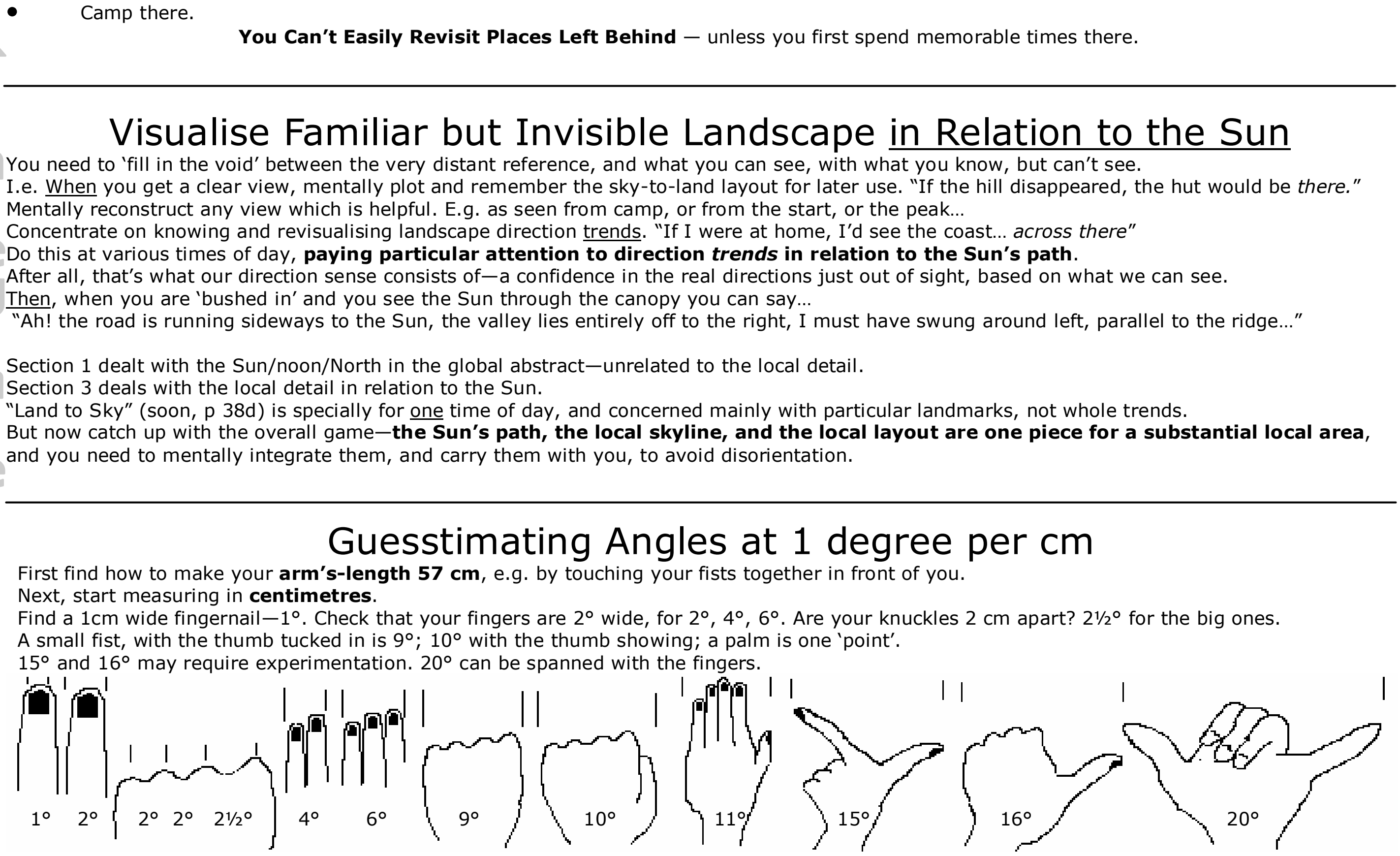

The ‘Blink of eye’ method also yields 6° — point to something with your finger at arm’s length or with something narrower, with one eye, and blink.
Larger horizontal angles can be visualised from between your feet of from overhead.

Initial Orientation is the Single Fundamental — The Get-Back Path

"Which way do I go?" is more important than "Where am I?/How far?/Where is North?"
"How do I get back to a familiar path home?" is the real need. That means you need to orient yourself **before** getting lost.
The single most important direction is **Start to Finish** i.e. Which way did you **leave** the car park? **From** what recognised point?
"Which way have I been **going**?" (if I had to backtrack)
Which direction did you **head off** in — **toward** what identifiable point?
E.g. "Downhill". The direction need not be horizontal. We function on 'paths' more than on straight lines.
E.g. "to the lake". The direction need not be relative to North; It does need to be something you can reverse.
E.g. "with the Moon ahead on the left" "the long way around" "to the opposite side" "away from the mid-morning Sun"

Taking a map and compass is good; A watch is nice; But the only really necessary bit of navigation you must take with you on a walk, to get you back when lost or disoriented, is... what direction you have travelled out.

Initial Orientation is the Single King Principle — and the rest may not be much help without it. See the mnemonic p. 111c for some help.
Which way you went is also the single biggest help for searchers—**tell someone** where you are going!

How I Manage to Get Lost or Disoriented

… **starting off** without paying attention … blindly following the track or the leader … taking a minor junction
… choosing a hard route — thick vegetation … going without a compass, or map … trapped by a peak or clearing (p.45b)
… falling asleep during a journey — a gap in my reckoning … waking up in a new place … not noticing the "end of track" marker
… inside a hospital complex, or a shopping centre; coming out … in look-alike landscape e.g. suburbs; losing the car on city streets.
… getting out of bed the wrong way or entering a town by a last-minute bend … talking; inattention; not noticing bends & junctions; losing track
… misinterpreting "directions"; forgetting them; being given wrong directions … hurrying; "short-cuts"; leaving the track
… on river-flats near river junctions and bends … preoccupied — with map-making! or any other task
… inadequate memory e.g. in caves, the way back, the many bends … stubborn psychological disorientation syndrome
… miscommunication between parties e.g. about how & when to meet … confusion; blunders (see list p. 39b)
… not enough knowledge of what to look for & how to interpret what you do see … visiting the Northern Hemisphere
… minimal input technique of navigation e.g. 'just follow the signs'; overconfidence … misinterpreting the map—especially the scale
… new tracks, not on a map … false identification of landmark features … losing confidence just short of my goal
… unsure of distance covered … navigating by creek junctions or minor dirt roads … blinkered by fog, darkness or thick forest
… wrong turnings; walking in an unintentional circle; overlooking junctions; not looking back … waking up on a dark night, 180° disoriented.
… too much new territory too quickly … negotiating thick bush without keeping orientation

---

Voice Your Thoughts — Don't Suppress Important Clues

When your silent auto-pilot notices something and tries to tell you… Say it out loud. That's the trick to it.
Tell someone. If you are alone, yes, talk to yourself. Bring it to light.
You can't remember the jigsaw bits later, unless you look at them now! **Subconscious thoughts flit away like dreams**, so express them first.

"Oh I didn't expect to reach this so soon!" ? Misidentified
"North seems to have changed. It shouldn't be out there!" Disoriented!
"Hasn't the wind swung? Does anybody agree?" ? Are the others asleep!!
"The shape's not quite right" Most probably because it's wrong!
"The markers have stopped" ? Off the main track

The same goes for run-of-the-mill clues. A forgotten clue is a missing jigsaw piece when you need it, so say them all out loud.
"We're crossing a divide — see — we are now going downhill"
"Creek number... four" "Marker!... Blaze!... Cut Branch!..."

Back-track in Your Mind before You Back-track on Your Steps

Lost?
"Don't Be Afraid to Think"
Spend time rethinking.
Try to reconstruct where you might have erred. Have I overrun my goal, or am I not there yet? Or am I simply off course to left or to right?
Did anything 'not make sense' on the way out?

- Work backwards from now, before you forget.
- Assess how successful back-tracking might prove.
- Don't ask "Where do you think…" but shoot for the details — ask "Why do you think…"
- Each of you should draw a mud-map separately, then come together, to compare recollections. Argue out the sequence.
- Back-track on your presumptions.

Detail the facts with a cool head and many problems will dissolve.

Distance using Apparent Size & 6°

The visual size of people, cars, houses, as measured at arm's length, varies with distance. You know how big they are already.
Blink your eye or otherwise guesstimate what fraction of 6° it makes, and multiply its size up by 10, plus for the fraction, to get the distance to it.
E.g. something one knuckle space wide is 3 times further away than if it were 6° big, and therefore is 30 times further away than its absolute size.

Visual Angle, Size and Distance

Find the angle, and multiply the size up by 60 in 1-per-degree. E.g. If something 1m wide looks 5° across, = it is 12m away. 0.5° means 120m.
Find the distance from the map, and work out the size from the apparent angle, by 1 in 60. Something 10km away, 10° wide is 10/6 km across.
Using 7/4 in 100, for accuracy because the angle is small, something 10km away and 10° across should be 1750m wide.

Side-wise Offset

Blink your eye at something in the distance, and it will shift across the background skyline.
Walk sideways, pacing out the distance, until it lines up with the new skyline mark. Multiply by 10 for the distance to the something.
Alternatively, mark the new position alongside the object, which the 'blink' indicates, then walk up to it and pace out the offset distance & ten-tuple

Don’t Just ‘Get By’ and ‘Make Do’

We all have a handful of techniques we feel comfortable with. We all get lost.

Most “How to” books necessarily are telling you how to battle uphill — against your natural incompetence.
This is doubly true for a “How to Improve” (your navigation) book!
E.g. ‘Improve your memory skills’! Easily said; hard to digest.
Get used to the idea, and therefore embrace it — the idea that you, as a good navigator, should want to extend your familiar range.

Navigation is inherently a problem (of not being able to see around corners; of not having a magnetic sense; of not having excellent memory; etc.).
However humans inherently do have the potential skills to cope. But do you have the guts?

Memories are Made of Landscapes, Not Compass Needles!

What to look for is detailed later, but for now consider...
How to ‘look’ …

Use your compass and map vaguely! Approximately! Lightly!
It is safer to be aware of their meaning, rather than to be accurate but disoriented.

Imagine you had no map, no compass. No signposts. No highways.
“How did the pioneers do it!?” You learn how to not get lost; how to ‘see’ what you are looking at.

Once you get oriented, then you can bury yourself in the instructions on decimal places, for accuracy when you need it.
Otherwise numbers and recipes have an ability to preoccupy your mind and numb your thinking.

Treat your map and compass as shy — self-deprecating — and turn your piercing stare away from them to what they are pointing out to you.

You Need a Road-Map To Climb a Hill!

It is a waste of effort to realise your mistakes at the top of a hill.
“I wish we had brought binoculars” is all too common a regret.
“We must remember next time to bring all our maps”

You have few such opportunities to positively identify landmarks,
and so you should **take the chance to familiarise yourself with what each hill looks like from every other hill.**

But often, it is only your smallest-scale map — the roadmap — which will show you the big layout.

“Didn’t anybody bring the compass??”

It’s up to you to plan ahead, to plan on making the most of your efforts.

‘Land to Sky — Sky to Land’ Navigation

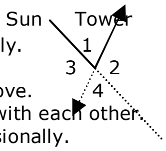

Step Zero: **Leave yourself out**, and your compass, of the scheme of things — so that the whole lifeless picture only changes slowly.
The result is simplicity itself — You see **how the Sun shines across the landscape**. Remember the angles.
Step One: “The Tower is by angle, from the Sun as I see it” Both change only slowly (or rather, seem to) as you move.
The Big Bonus is: You can keep track of all your bends, twists & turns, by the Sun & Land working in conjunction with each other.
You can locate a landmark easily through the vegetation — you know where to expect it when it shows itself occasionally.
i.e. You can always easily locate the tower/peak/etc, from kilometre to kilometre, in relation to the (now slightly shifted) Sun (or Moon, wind, cloud)
Step Two: “The Tower is from my shadow” This allows you to walk away from the Sun, yet still stay oriented to the landmark.
Step Three: “**The Anti-Tower** is ..from the Sun” These allow you to walk away from the landmark,
Step Four: ..from my shadow” yet still “see” where it is.
Step Five: Mnemonic: “The Landscape on the **Left** Lingers Longest in the Light” — its apparent movement as you walk, follows the Sun
& the Shadow’s anticlockwise shift. (SH only) i.e. you don’t need to worry much about time passing, and angles changing.
“But Landscape on the Right… soon goes wrong” — it ‘moves’ against the clock-sense of the Sun. (Southern Hemisphere rules)
i.e. you need to **refresh your angles** more often during the day. (NH: Left Loses the Light; Right is all right)
Step Six: Rule of Thumb: “**15 km away** changes at 15°/hour (i.e. at about the same rate as a low midday Sun) as you walk past it at **4 kph**”.
The Sun shifts at less than 15°/hr when it is rising or setting — so try “20 km away/12°” to match it; and “10 km/25°” if the Sun is very high.
Yes, you should convert the angle to “ESE” or rename it “just S. of sunrise” for later, but **the idea is to use the Sun as it is now, not North.**

Grid-line Walk — Explore in a Square Grid Pattern

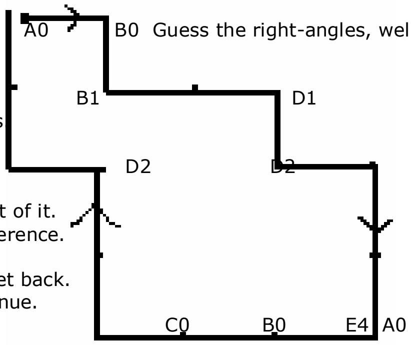

Guess the right-angles, well

This only adds 20% walking effort, but lets you know where you are, and how to get back,
and along a route of your choice, among many choices.
You need to judge a given distance, well, e.g. 100 double-paces; one minute; or 100m; or by pedometer.
Choose a **reference direction** as ‘Up’ or ‘ahead’. Sun-North; Compass Needle North.
Up-weather by clouds; down–clouds. A strong wind. Distant Hill. If needed, keep a straight line by marks.
Choose a **scale** to suit your ambitions, e.g. 100m squares (so that each 10m is an extra 0.1).
Try to cover a whole square at a time. Use 0.5, if needed, or you can cross at a diagonal, 140m long.
Count **forward and back progress** in whole chunks, as A-Z. G.5 means 7½ squares.
After Z comes AA, AB, AC. See p84b for making suitable mnemonics. **Don’t lose count!** That’s the worst of it.
Count **right and left progress** as +/- 0-10, etc. Minus simply means ‘backwards’ or ‘to the left’ of the reference.
When you have definitely decided to head back to base, remember the grid-reference. Stop counting.
Reset the count to Zero. This is easier than trying to count backwards, using subtraction, to get back.
Reverse the reference direction, e.g. to downwind, shadow South, compass South, and continue.
When you arrive at the original grid-count, you should be within sight of home-base.

Disorientation is More Dangerous than Being Lost

…because the same disorientation that will get you lost, may then get you more lost.
Orientation is primarily a mental picture — therefore — Disorientation is a wrong mental picture. It is one and the same principle.
So Disorientation really extends to cover more than direction sense, into bewilderment and logical debilitation!
You can be disoriented, deluded, bewildered and irrational about more than direction or position. E.g. You may misjudge which way is up/out/back.
E.g. you may be mistaken about the time of day, or the existence of a road, or whether the promised weather front has passed you or not.

'Lost' is 'I don't know where I am' which is different from disorientation, since position sense is independent of direction sense.
Again, you can know where you are on the map, but still be bewildered on the ground — not 'fitting in' even if direction-wise oriented.
Then again, you can be oriented, but lost — yet not necessarily worried or in difficulties — because you are oriented.
You can be disoriented unawares — even when you know where you are — i.e. not necessarily aware of the danger of loss of direction sense.
To be both disoriented and lost is most dangerous, but **it is orientation which stops you from getting lost**.

Orientation keeps your finger on the pulse. It stops you getting disoriented. In many ways it gives you a position sense too e.g. that you are in a particular catchment, downwind of a known landmark, up-Sun of the overall course. Orientation of the map puts the map onto the landscape—aligned to reality—so that the map can point out its riches to you. Orientation lets you know where you are, recognise places, & know the way back

---

Logical Blunders — Examples

Mistakes of category: "The last stream was flowing to the left" (Oops!)
"We have crossed the main road" (Now, the main access road is dirt, whereas the main through road is bitumen, but you muddle them up. Or maybe you are unaware of a fork in the main road.)
Mistakes of number: "We have passed two lakes" (which were joined behind the intervening vegetation).
Mistakes of quantity: "We haven't changed direction" (except for one hairpin bend!)
Ironstone may have swung the compass, when you put the map and compass on the ground 'for extra certainty'!
Mistakes of quality: "From the peak to the lake was 090°" — but the false peak you were on, gave you a false sense of accuracy, in the mist!
"See those houses 'across' the river?" — Rather, see the bend on the left — they are on this side!
"The island should poke out from behind the far point. Why isn't it there?" (It is 'the far point', of course)
"But we have crossed the River" No, only a main tributary, or only onto an island.
Gaps in knowledge: "We haven't seen the junction" — because you short-cut past it; around one bush is enough!
Unexpected: E.g. there are sometimes two parallel roads — the old & the new. E.g. a stream sometimes goes underground "Mole Creek".
E.g. A water canal may cross catchments. E.g. "Remember the creek junction at the base of the hill" — The 'hill' was only an island.
Assumption: "That will be the ridge we want up ahead, let's turn off here" — Unproven!
"Ah, we are back to the road!" Yes, but not the same road!

Trust the Untrustworthy

It's like people. 'You can't trust anybody', but you have to trust your life to them sometimes.

A compass has dozens of potential faults, but in a fog, you trust it, implicitly.
You yourself are less trustworthy than a compass.
When all you have outside your own opinions are reasonably reliable external clues (like where the Sun is!) you had better trust your life to them, not to yourself.
For instance, sometimes that 'external clue' is someone else telling you that you are disoriented, when you don't think so — you had better listen!

A disoriented person may not want to trust the 'untrustworthy'. Don't let that be you.
E.g. You may find yourself twisting the map to make it line up with a mountain which doesn't really line up, except to your expectations.
You prefer to believe that the map may be distorted!
It takes practice to believe. I once spent a whole week wondering why all the trees were pointing too much to the sunset side of noon.
I was trusting myself, not the trees, and was simply wrong about where North was. I didn't realise which was the more trustworthy.

Lost? Rule 2: Start! Getting Unlost. Find yourself. Regain confidence.

(After you STOP getting lost — Rule 1) It is time to **START NAVIGATING** — Rule 2. (You can practise this ahead of time, without getting lost.)
E.g. What can I see & recognise? E.g. Can I interpret the wind or sky or trees or drainage for direction?

- First: Mark your spot e.g. tie up a hanky. Mark your Trail—so you can get back here, as 'base'. Retrace your steps, if that is the best option.

Act out a Search & Rescue, looking for your 'lost' starting point. (If you can see your surroundings, you don't have to feel 'lost'.)...

- Explore in all directions systematically.
- Make a map. 'You' in the centre. Locate yourself in the landscape. Climb a high point, to orient yourself; at night too, for lights.

Let others find you...

- Leave notes about your plans. Especially at 'findable' places e.g. a hill top.
- Leave clues so that searchers will not give up for 'lack of clues'. Make some marks visible from the air.

Search out a track...
Ask: "Where should there be a track?" E.g. to... that peak?, a fishing spot?, any low pass through the range?, the main access valley?

Mental Grid Map

This is the square grid method, where you simply walk, along a path of choice, and rather than following the grid lines,
guesstimate your decimal position in the grid scheme.
The beauty of it is that it constantly forces you to orient yourself to a reference *direction*, to notice *distance*, and to be aware of *position*, yet it takes nothing more than mental equipment. You should aim to practice it properly, once, and you may not want to discard it.
It causes less mental strain if there are two or more of you. **Delegate** the left/right to one person; the forward/back to another.
The one person guesstimates distances by one method, their way; The other gauges perhaps a different distance and with different accuracy.
It doesn't really matter if your guesstimation mismatches, provided you remain consistent and imagine a rectangular grid.
In addition you need a feel for triangle lengths, for a 1 in 10 slope, up to a 10 in 10 slope, so you can guesstimate decimals.

~You need to divert 5 in 10 to add more than **10%** to the length walked. E.g. from Z0 to A0.5 takes 112m across 100m squares.
~Maximum extra length is a 45 degree diagonal, at **141%**. E.g. from Z0 to A1 takes 141m. These two set the feel for guessing distances
1 in 9 adds one third. For 6 in 10 to 8 in 10, triple the first figure as the extra %. E.g. from A0 to B0.7 takes 122m.
If in doubt draw a 10x10 square, draw in the diagonals to each of 18 opposite points, and label the lengths as the square root of $100+n^2$
This lets you see the direction of each n in 10 slope, and know its length.
~You could use degrees, up to 45, and call them percent extra distance, with fair accuracy, but there's no need for such figures.

When I simply guessed, by eyeballing distances and directions, it brought me back home without noticeable error, after a few kilometres.

Walk Where the Navigators Talk — if you want to learn navigation

Not out of earshot.
Navigation is a learning experience.
It works better when you can share the process of learning with someone.

Explore

Explore the unexpected — the side tracks. You walk to explore, usually, not to exercise.

It's No Fantasy

Navigational know-how is not "If I were lost I could do this…" but is "What I am doing right now, all the time, will stop me getting lost".
Learning new skills and how to interpret all the clues is the practice of navigation here and now.
Continuously pointing out new clues to yourself, and to others willing to listen, is the process of not losing yourself.

Feed Your Subconscious through Your Conscious Focus

E.g. To improve distance guesstimation, focus on the finest visible distant detail. Notice the haze level, the shadow hue, the sizes you can see.

To train your direction 'sense' (which is entirely lacking in humans) feed it all the clues you can —
Sun & shadows, wind & weather, compass & landmarks, trends & bends… Don't just look at them! Put words and meanings onto what you see.

To extend your memory of a sequence of turnings… use pencil and paper 'memory'.

A working rule of thumb: **Bring to conscious note**
what you hope to keep track of subconsciously.
Don't expect to be able to remember what you haven't noticed or thought about consciously.
So learn what to look for, and how to see it.

Look into the Far Distance

..for Trig. Points; Towers. They are there to be seen, by definition.
..for Roads; Houses
..at night, for City glow; Auroral Lights
..for the dip and strike of the underlying geology
Assess the difficulties ahead, whenever you glimpse them e.g. a steep bank on the other side of a lake. Use binoculars.

Look through the Haze

What are you looking for? You want to pick: intervening ridges, hidden gorges, distant cliffs, overhangs, false peaks, mountains or islands, the Moon, and the trail ahead. Those cloudy lumps are not always clouds. Those invisible peaks are there to be found.

If you don't see them right away… **Wait** ..it is the **changing light conditions** which show up the contours.

E.g. Look at night, when the whole emphasis shifts, and light sources show up.
E.g. Wait for the haze to shift — it moves with the Sun.
E.g. Wait for the haze to thicken — to show up the valleys — weather is changeable
A rain shower may show up a valley; a cold front might clear the air of haze.
E.g. Wait for a cloud shadow or patch of sunlight to drift across — it will disappear into any hollows, and highlight any foreground.

Join the Invisible Big Picture To the Visible Landscape

The Big Picture might be invisible — e.g. the South Pole Star is "Over the South Pole in Antarctica" — but geographic "South" is quite different and is visible — "that knoll there!". Join the two!
The first principle to practise, is to work top-down; but the second and more important need, is to integrate everything.

When you put a continental or global picture onto the landscape, you have a full-scale map of the world in front of you.
e.g. Point just beneath 'that Knoll' and put your imagination into gear...Down over that horizon are icebergs separating from glaciers with a huge splash! Can't you feel the chill? Up there is the night-time aurora. Down there, anti-polar bears look straight up and see the Southern Cross whirling around overhead! And we must be on the south side — the cold dark side — of this hill behind us.

You say: "The destination lies beyond that valley." "Mount Saddleback marks sunrise"
"The track winds around behind this hill." "The Sun skims those treetops"
"Porpoise Bay must be yonder; Melbourne that way; and my girlfriend, 50km in that direction…"
Integrate as you walk — in real time and place

Cyclical Motion Divided into 12ths

Any back-and-forth motions or cycles which are caused by circular orbits or sinusoidal oscillations are easily understood and followed *in units of 30°*.
This is because 30° of rotation will swing *half way across* a radius from top-dead-centre, or half-way up/down from right/left extremities.
E.g. quick rising for 4 twelfths, then *2 twelfths* in the *top half* of the swing, 4 quickly reversing itself, and wallowing for 2 at the bottom.
You may not *see* the 'circular' nature of a one-dimensional oscillation, yet it will still conform to this rule.

Tidal Height	Sun's Seasonal Latitude (or Moon's monthly position cycle; use 2½ days)
1 hour falling slowly from the top	1 month at Northern solstice slowly coming South
2 hours dropping rapidly to midway	2 months rapidly moving back to the Equator — to the South
2 hours dropping rapidly from midway	2 months rapidly moving South of the Equator
1 hour dropping slowly to minimum height	1 month slowly approaching Southern solstice
1 hour at bottom rising slowly	1 month slowly turning back to the North
2 hours rising rapidly to midway	2 months rapidly returning to the Equator — toward the North
2 hours rising rapidly from midway	2 months rapidly moving North of the Equator
1 hour rising slowly to top	1 month slowly approaching the Northern solstice

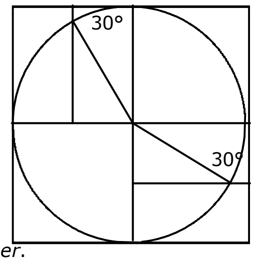

You can simplify that into sixths, saying: "1 sixth *at the extreme*; 2 *on each change*; 1 at the other extreme"
Or say: "*twice as long on the change*". That makes 'the change' sound 'slower' but it actually goes *faster* and *for longer*.

Which Way Did I Come In? Which Way Do I Go on?

Have you ever interrupted your pleasant journey at a roadside picnic area and driven out the wrong way — back the way you came — and not onwards? Or you turn left instead of right when you drive "out" of a local dead-end network, into another dead-end maze?
Or you sat down for a rest on the track and when it's time to go, both directions look identical to you?
It's embarrassing on the road. It's nerve-racking in the bush! You feel struck with insanity. But it's normal enough.

Or when you come back to visit a familiar place, it's on the 'wrong' side of the river! Every time you come back?!

An extra external clue, like the Sun, is needed, and needed to be believed, but it is really needed before you make a fool of yourself.

And you may have to undergo 're-orientation' to dis-illusion yourself.

Don't let it happen to you deep inside a cave system!

Logical Presumption Leads to Utter Delusion

Think about those example 'logical blunders', about your own fallibility, about the dependence of logic upon its basic assumptions.
Being twitted by a logical mistake leaves you with a horrible 'lost' feeling, plus a pathetic 'helplessness' about how to reorient.

These are not simply mistakes of guesstimation, subject to fine-tuning!
Once you subconsciously make a decision about how the world is, then your mind will close to the possibility of error.
False assumption is presumption, so I say "**Presumption Means Delusion**"
Such Left/Right, Yes/No, crossed/uncrossed, 2, or was it 1? -type mistakes lead to utter bewilderment, so I add
"and errors involving discrete categories can be especially indiscreet!"

Being aware of the phenomenon of logical blunders, is half the battle in overcoming them, when they bite.
Keep an Open Mind about the truth of your assumptions!

Trust Less and Prove More

Don't trust yourself At numbers
Or remembering accurately
At judging time or distance
Or walking in a straight line
At identifying landmarks properly,
Or walking back upstream, and so on.
Don't trust A compass to be dead accurate
A map to show North reliably
The weather forecast
Old track markers
A navigational 'fix'
Complicated or finely adjusted instruments. They can easily go out of adjustment, without letting you know.
The point is not so much to **double-check** and **cross-check**, as to: **back-up** with an alternative; **interpret** for sense; understand; and **be wary**.

The Simplest Way to Get Unlost is to Retrace Your Steps

The sooner you go back to where you went wrong, the sooner you can relax. Yes that might be uphill. Yes it might be a long way.
"Oh I wish I had retraced my steps earlier, while I had the chance!"
You still have the chance, you probably just don't want to.
For instance, if you have been pushing on, hoping to find the way, you should be doubly aware of exactly where you have come, since you doubly anticipate the need to retrace, if you fail.
"Short" cuts have a habit of surprising you at just how long they can turn out to be, repeatedly.
A track naturally avoids gullies and obstacles, so a 'shortcut' is almost guaranteed to hit one square on!

Retracing your steps may be something you can't short-cut. It is the shortcut. It short-circuits many other possibilities.

Caution: Sometimes retracing your mistakes will only compound them—more of the same!
You may sort yourself out better if there is an alternative system to try e.g. to walk back upridge if you got lost going downstream.

Approximate Sine-Wave Values 97% accurate

(From school you could have memorised that sin 30° is 0.5 and sin 60° is 0.866, and perhaps sin 45° =0.707 ???)
The first twelfth of an oscillation takes something half-way (0.5) to its peak amplitude (1.0) from the middle ground (0), by the rule of twelfths.
~*The motion is virtually linear for that first 30 degrees*. So to convert those *first* 30 degrees to a sine-value, **Divide by 60**.
I.e. add at a rate of 0.5 in 30, or 1 in 60. E.g. 25°/60 = 5/12 (or 0.417). Sin 25° is actually 0.423, which means we had a −0.006 error.
Sin 15° is 0.259, but we get 0.25 (–.009 error). Sin 20° is 0.342, but we would guess 0.333. That's about the maximum error, 0.01 low
The next twelfth carries the motion from half-way to 86.6%, in a curve; The extra is 36.6% or a slope of 1.22 per degree. Let's use 87,36,1.2
A linear fit cuts the curve twice, and so is close to most of it. If we could add 36%, over 30°, to 0.51, to arrive at 0.87, we get +/− 0.01
~To the ***extra*** degrees, **Add 20% ; Add 1; Call it a percentage** (divide by 100). (*To add 20%, add 1 tenth, 2 times*, e.g. 55 + 5.5 + 5.5 = 66)
E.g. 60 degrees = 30 plus an extra 30; so start with 0.5 and add 0.3 then 0.03, 0.03, 0.01 to get 0.87, which is 0.004 too much.
E.g. 40 gives 0.5 + 0.1 + 0.02 +0.01 = 0.63, instead of 0.643, or a −0.013 error. Sin 45 is about 0.5 plus 0.15 + 0.04 =0.69 or −0.017
To make the next approximation, use 0.87 for the first 60 degrees, then worry about any extra degrees.
The last 12th is the flattish top of the curve, building up to the crest. It climbs only 13.6% from 0.866 to 1.00 in 30 degrees, in a curve.
We start at 0.87 and finish at 1.02 cut the curve twice and best-fit to within 2%.
~For the last 30° degrees **Divide by 2, then 100**, to gain 15%, from 0.87 to 1.02.
E.g. 80 degrees moves to 0.5 then to about 0.87, then an extra 20 degrees or 0.1, yielding 0.97, 0.015 below the real value, 0.985.
75 yields 0.5 + 0.37 + 0.075 = 0.945, 0.021 below the true value 0.966. 70 gives 0.87 + 0.05 = 0.92, 0.02 below 0.940. I.e. +/− 0.02

Self-Motivation is Fundamental

For a sense of direction to become "second nature" you need to **train** your second nature.
Improved navigational ability boils down to… *putting new ideas into practice*.
(Not perfecting old mistakes! Nor leaving the right technique on the shelf)

Don't just know about a good idea; make it yours.
For this you need to want to improve your skills.
to force yourself to practise your weaker points.
to push yourself to develop new habits.
Which is fun.

Enjoy the trip!

How to Feed Your Brain

When and where does the Sun get up?
Get up and look (not at it!). Don't make it an academic question.
You don't need an academic answer, either. You measure it with your eyeball. Perhaps draw it.
That doesn't mean "ignore the details" but "memorise the feel".
You do 'measure' it, but not with numbers.

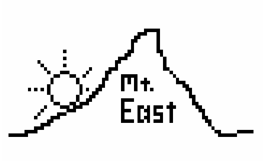

There! This early!

Star navigation? It doesn't have to be a "trigonometrical nightmare".
For us it means: Find the Southern Cross.
Guess the time of night.
Find the Pot. Get to know the stars.
Project their movement forward into the daytime.

Let your brain do the calculations. Show it what to work on.

Wait for a Change in Perspective...

..to reveal the hidden shapes and layout.

As you drive up, keep **re**-examining the changing skyline. Continue the process as you walk. Look for ridges & valleys showing up through the vegetation as well as on the skyline.
E.g. Use parallax to reveal a deeper horizon behind a hidden valley...
Keep your eye on particular bumps on the skyline.
If there is a deeper background, i.e. more distant, then a lump on it will show up by moving with you —
more quickly than a 'foreground' horizon. You may have to look through the deep haze to even locate particular bumps.

Climb a tree, or a hill.
Use binoculars. Look especially for the trail ahead, in relation to the terrain.

It is the **Changing Parallax** which **Shows Up the Landform** and its Relative Distances.

Landmark to Landmark Network Navigation

This is a permanent, unchanging, top-down, whole-region method, **prepared at home**, for a large area, and for particular walks.
1. Choose only **prominent landmarks** from the map. e.g. peaks with transmitters.
Add in major ports, mouths & coastal extremities, lighthouses, around the coast.
Draw the network of triangles between them. ("Which one am I inside?")
Draw the circumscribed circle for each triangle. ("Am I inside or outside?")
Label each line in **kilometres**. ("What does 20km look like?")
Label each horizontal **angle** in degrees. ("Does it look larger (inside the circle) or smaller (outside)?")
Label each line with its **trend**-direction. ("Am I to left or to right of **track**? Which is this river **parallel** to?)
On the right of a line you will be looking to the left of its trend to see one of the landmarks.
2. Keep a **notebook** for local peaks — smaller but locally prominent lookouts.
The idea here is to navigate only by landmarks, not by (favourite) low spots.
You don't have to worry about whether one landmark is visible from another.
The idea is that **even when you can't see a landmark** (as well as when you can),
you can know where it is and how it lies in the land layout, how other trends line up with it and with other visible lines, and *how you fit in*.

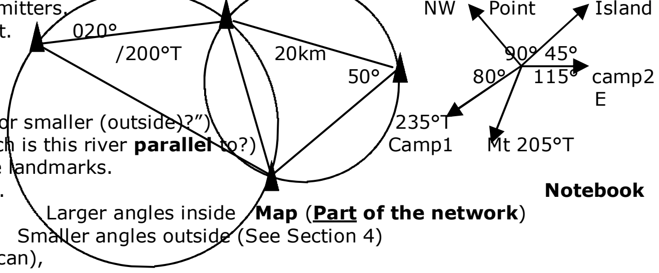

Name Each Clue — with solid ground

"That's Pacific Sunlight shining" — a mid-morning Sun, for Australia
"I have an Anti-American Shadow at the moment" — pointing away from North America
"It's a West Coast Cloud Flow" — you can't lose your orientation
"I love a Tropical Moon" — when the Moon is North of you (SH)
"That wind would blow straight in our back door" — 'Back-Door Blow'
"Downstream from here is uphill from Port Stevens" — 'Steven's Stream'
"This cliff sees the Southern Lights" — 'Cold Cliff'
"Topple-Tree Trunk would take us to True North"

The next step is to add something in line with that — see the next page.

Disorientation — A Killer Syndrome

Pilots have been known to fly upside down, or rigidly leaning to one side according to their mistaken belief about "up"!
If even our natural senses are unreliable we certainly can't afford to believe in any 'sixth sense' of direction, or on any unproven magnetic inbuilt direction-sense in humans! A trudge can dull & distort our senses. Your mind tries valiantly to maintain mental orientation—that is the problem.
Several times each year pilots die, scuba divers get lost in sea caves, cavers fail to find the way up and out, boatmen falsely navigate the fog, and bushwalkers get lost in the bush, fatally. They were not just careless; they were probably trying hard, but in the wrong direction, stubbornly.
Belief is an irrational master. Its force is stronger than reason. The conflict, combined with physical tiredness and sensory deprivation, is stressful.
We prefer to get lost rather than to rethink wrong assumptions!
We become willfully blind to strong contrary clues in view (including signposts, instruments, buoy colours and numbers!).
We magnify weak clues, to justify our delusion. This heightens the stress, because our sub-subconscious knows that we are wrong.
We actually lie to ourselves about what the clues do say! (e.g. we change the number, colour, reading or wording in our mind!)
Such irrational behaviour is explained as defending our psyche from hurt pride and from being devastated by confusion, fear and panic.
Don't laugh it off as "Oh, that's only for the worst case" because it is the less obvious self-delusion which will likely trap you.
Being human includes you — you definitely are vulnerable to being even less sensible than trained pilots!

Casual Outings Catch You Out

..unprepared e.g. without a map or compass or weather forecast.
Casual outings also have a habit (because exploring is interesting) of "I'll go just a little bit further"; and turn into longer than expected full-scale safaris. Exploring is built into the genes of a small fraction of any animal population—it's a group-survival insurance. But don't let it catch you out.

Serious trips go more reliably because you know what is needed, but you rarely prepare well for a casual trip.
e.g. you have no pedometer or notebook and so don't keep a good enough log of distance or direction.
So **Lesson Zero: Make some precaution** (to avoid getting lost) **every time** you wander off.
Another unexpected casual mistake is to decide to leave the track without thinking what precautions may be required.
Try this. Look up and repeat after me: "God, have I forgotten anything? Is there anything I should do first?"
Let someone know the single biggest clue: "I'm going exploring that way",
but then don't forget the unobvious catch—that person might also go exploring & not tell anyone where you went.
(Being disabled by an injury is one thing, but dying of exposure or thirst or wild animals, simply because no-one knows where you are, is another whole added dimension—that is probably the strongest reason for bothering to let someone know.)
So before you go off, pass on any vital information about others, as well as about yourself.

Trust Yourself Less

..when the compass and you disagree, for example.
When the map shows the junction there, but you reckon it should be here, for some reason such as 'but we've come too far!'
Or similar cases, of confusion.

Your reasoning sense has holes in it, big enough to outsmart yourself.

In the end you will usually conclude "Well, I was wrong, due to this very good reason(!), and it was how the map and compass told me it was."
But you don't want to admit it, at the time, until you are proven wrong, yet again.
Your reasoning is not, despite your worship of it, the highest god in the Universe. You need to be able to criticise it freely.

Unreliable as map and compass can be, at times, it is better to trust them more and yourself less.

More is Better — Thought, Time Care

A simple principle to apply is to do a better job at what is marginal.

For instance, "flat" ground nevertheless does have a slope which you can determine with care.
The wind may be gentle but a little extra time could make the difference between success or failure at determining its direction.
You can always double-check a casual decision, e.g. about which way the dry creek-bed flows.
Instead of wandering in an unrecorded meandering fashion, you can mentally log your changes in direction by lining up pairs of marks.
Just before you leave the track, a quick checklist can bring a few necessaries to mind, but only if you take the trouble.

A little extra thought; an adequate time-allocation to each task; more faith; and that bit of extra care—will solve a lot of problems.

Moveable Position vs Fixed Orientation

- It helps to park the car in a cardinal direction, or pitch the tent North/South, since we all rely on such clues to revisualise what we left behind.

- But when you do try to orient yourself to something invisible such as how the car was parked when you walked off, there is a simple trap to avoid
  Don't imagine the car/tent as 'behind' you and twist its imagined *orientation* to suit the Sun or compass;
  Instead change its *position* — *move* it to somewhere other than 'behind'.
  Mentally manipulate its placement, not its orientation only. Or else the 'correct' orientation will confuse you — nothing will make sense.

Don't let the Sun 'change its direction' — that is a tell-tale that you are keeping the reference position fixed *behind* you rather than *moving* it as you swing to left and right.
Because we have no good tools to log changes of position or distance, we tend to neglect that half of the story in favour of maintaining orientation.

- As another hint on how to use a fixed orientation, e.g. North-South, so as not to let it 'change' in your mind, imagine it as a long, huge, high, iron girder bridge overarching you and your car/campsite, in the appropriate direction. Whichever way you imagine it, e.g. through-bolted to the continent, or from one peak to another, make it so solidly fixed to the ground and impressively immovable that you feel so insignificant yourself that you feel *you* must *turn yourself around* in order to pass under it.

Learn Think Look Integrate Integrate

Routine Mind-Shut-Down Procedures

You know how you become dull and passive — when being over-taught, lectured at, just entertained, occupied in repetitive tasks, sitting in front of the television, failing to solve too-difficult maths problems, or being proved to be a perverse idiot against your own opinion…
Automatic self-defence from constant attack... involves shut-down — into a semi-hypnotic trance.

Similarly,
Walking becomes a "trudge" after about an hour without a break, depending on the load, the heat, and how many trudges have preceded on that day, and previous days.
Pain-relieving opiate levels rise. Conversation takes a dive. The path sliding under your feet gets stared at a lot.
Your mind is passive, plus your body is on auto. It is physiologically involved in the hypnotism.

Mental Torpor is a navigator's constant enemy.
Keep the Different-to-Same-Ratio up, for both yourself and your companions, so that you can stimulate each other into full consciousness!

Feed Your Subconscious Subconsciously

E.g. Open your Atlas to a map of the world. Orient it and leave it open, oriented, and in view.
E.g. Place a magnetic compass somewhere in view at home.
E.g. Develop a habit to Orient your watch N-S each night when you take it off, wherever you happen to sleep.
E.g. Sleep with your head to the north each night if possible.
E.g. Clearly visualise what direction your feet are facing before going to sleep each night.
E.g. Imagine your body lying across a map of the continent.

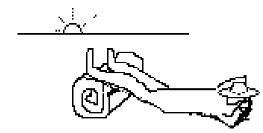

New Habits Open up New Horizons.

"Retracing is Re-preciating"

..This memory-aid is something you have to tell yourself — to avoid blinding yourself and missing the goodness of the walk back.

If you are only titillating your senses for the passing instant, by constantly seeking "new" experiences, you will develop the mind-deadly
"Bin there — Dun that" attitude
Familiarity breeds contempt… "I know the way" doesn't guarantee that you know much else about the way.
The return journey degenerates into a dull, silent, often over-fast, boring, mindless trudge, even if you drove all day to come and see this top-quality location! In effect, you have walked the track twice, but looked at it only once.

Two solutions:
Take an interest in your surroundings, to overcome passive mindlessness.
Questions on your mind overcome boredom. If you have asked questions about a new place, you should be able to answer them on the way back.

Remind yourself of the opportunity to "re-preciate".

Aim to Integrate…

- ..Integrate daytime clues with night-time orientation—use the Moon, the place of sunset, the sunset glow later (which is further South—SH), the twilight Blue Band, Venus, the Ecliptic, the skyline — anything to carry you over from daytime to night-time, during that awkward twilight changeover phase.
- ..Integrate invisible landscape with visible landscape—work out what lies behind each hill, to picture the outside world against that skyline.
- ..Read the "map check-list" and specify what each item means from your point of view—if you didn't have a map, for example.
- ..Integrate the changing, with the unchangeable. E.g. the Sun's daily rotation needs to be followed, by naming its current direction. Same with the track; and your nose; and the wind.
- ..Integrate one walk with another — don't just carry two separate mental maps, add a third overall view.
- ..Integrate each system with each other one. E.g. don't let the map be one "world", while the walk is another, and the night sky a third. Join them into one whole.
- ..Integrate each direction clue with each other direction clue — E.g. "The tower is just before magnetic noon from here" (3 pictures in 1)
  E.g. the boat-radar may fail in the heaviest weather, unable to penetrate the rain squalls—you need a boat-compass too.

It is the **combination** of clues which stops you getting disoriented — if you mistake one, the others will alert you.

Notice and Match Up

'Notice the clues' is the easy half of navigation.
Sometimes it is as simple as seeing that the track is curved.
However, a slight curve due to a river bend can turn you in a hairpin without you realising, even if you *do* notice the curve.
Until you also notice the *extent* of the curve, you are at the mercy of disorientation. Something more than looking is needed.
You need something to orient to, like the wind or clouds or landscape.
Again, it does you little good to have *seen* the clouds steaming by if you forgot to see where they came from.
'Notice *and Interpret*' is the catch-cry, which never stops.
But when the wind stops, will you know what direction you have been going? And 'heading for the mountain' is fine, unless you lose sight of it.
Something more than single-clue orientation and interpretation is needed.

Do Notice; Do Interpret; Do Orient; Then ***Match up***; Link it to something else. And to something else. E.g. the clouds are from the mid-morning Sun; The mountain is on my shadow's port bow; The river has bent from 'towards the coast' around to 'away from the coast'.
For special examples, see '3 in line', opposite

Un-Dis-Oriented

You don’t need to be perfectly and continuously oriented!
In fact it is quite hard to maintain an exact direction-sense, as soon as you turn your head, move a few steps, divert a few degrees, etc.**
You only need to be realistically confident of position & direction.
Hence **it is better to be roughly aware of many clues, than to aim for perfection, say, in location or direction**.
E.g. ‘Moon North’ is approximate, +/-60°, but awfully comforting! — It adds to a complete picture.
E.g. You may not even look at your map again, after studying it once, if by now you already know what it is telling you.
Don’t aim to be perfectly oriented — just undisoriented.

E.g. Hold the conversation on how to judge range; what’s behind the hill; where to find moss; why you can’t yet interpret lichens; and so on… and you can hardly miss seeing the walk.
E.g. Run through the Start—Stop—Rest Mnemonic, p.110, as a memory challenge to all… “What’s next in it?”… which has the benefit of practising all its clues too.
**As an exercise, guess, from a distance, diagonally, where the side of your house or car points to on the horizon, then go and line it up to find out
Or guess which point is opposite that hillock or tree, then establish how accurate your guess was. My results: +/– 10°; +/– 30° when not trying.

Trapped! By Not Looking Back

A fish funnelled into a fish trap may never again find the narrow entrance hidden somewhere in the large perimeter.
Converging tracks and sudden clearings trap bushwalkers too, when they walk through an irreversible change without noticing.
“Where did our small track join onto the main route?” “Where did we hit the fence/road/pipeline/track, and turn along it?”
E.g. stepping into the creekbed, to explore upstream, is a one-way move—you may not recognise it coming back, and could overshoot.

- **Beware of all sudden clearings**.
  “Where did we emerge from the bushes: out into the clearing, onto the peak, onto the riverbank, out onto the beach”?
  Look for footprints — anyone’s, but especially your own.
- Remember the phenomenon — and mark those junctions in your mind! Or even mark the junction on the ground. E.g. place a rock there.
- “Which tributary did we follow down?”
  “Which ridge did we come up on?” More on p.63b
- Log the distance and directions involved at such changes/junctions/clearings, before you walk on and ‘lose track’.

Disoriented? Stop. Say So. Reorient!

Be prepared to voluntarily switch from ‘certainty’ to uncertainty.
Learn to live with fallibility.
As soon as you suspect a disorientation episode, admit it. Say so, out loud. Don’t miss the chance.
Tell your brain off!
“You know you’re just guessing! You made some mistake and won’t admit it! Let’s work this thing out together”

If you don’t know that you are disoriented, that can be worse than being lost but oriented.
You don’t even have to be lost — Even if you know where you are now, if you are disoriented, you may walk on into trouble, rather than be looking to walk out of trouble.

Recall how often you have experienced a hard-to-shake feeling of disorientation.

Lost? Rule 3: Do a Good Job — of walking out or staying put

It may not be safer to stay. Well, if you decide not to stay, don’t be aimless in what you do decide to do…

- E.g. Retrace your steps.
- Keep a straight course — e.g. downstream; along the road; toward the mountain; to night lights; to the traffic noise; a chosen direction
- Be single-minded — purposeful, but not stubborn
- Leave a Trail — you may need to retrace. Yes, it consumes valuable time, but “don’t get more lost”
- Leave clues; notes — for others
- Walk quickly if you must, to cover sufficient ground, but ‘rescuers’ may not be able to keep up!
- Yes, your exit strategy may need to be aborted, but at least nominate an exit strategy for the exit strategy beforetime.

In-Line Integration

Put your preparation into something useful by using the ‘**3-or-more-in-line**’ interpretation method…

“The wind is blowing *from* Port Bluster, *over* Hot Hill, *past* the carpark, *to* us, and on *towards* the Sun.
“The Pacific Moon is shining parallel to the highway”
“The Mountain has Montague behind it, that cleared paddock in front of it, and that hillock in the opposite direction”
“The compass needle points past the Northern airport to the mainland”
“My anti-American shadow is pointing past Penstock to Point West at the moment”
“The stream is flowing from Mt. Big, down over Lake Low towards Mersey Mouth”
“That Tropical Sun is over New Guinea, shining Past Alice Springs, over the intersection to me, and into the teeth of the wind”
“The Nor’Easter clouds are flowing from Noumea, over the Main range, to here”
“That persistent pigeon is cooing from behind camp”

The effect is to turn your reference orientation directions into fixed tracks (see p71bi), making interpretation easy.

Navigation is Mental Work — Stay Alert

The tools to work with may be "map and compass" "line and direction" "time and distance"

But any workman needs, and needs to practise…

Attention Observation Analysis Interpretation Preparation Concentration Memory Visualisation

Specialist knowledge Deft application Reliability Right attitudes A store of safe and useful habits Coordination

In other words: **you get lost by lapses in applying your skills, not through lack of skills**.

Mental Equipment is Very Light to Carry — You already have it; Don't forget to use it!

Instruments Instruct the Intelligent

Don't let the instruments — compass, clock, protractor, map — do the navigating for you.
Rather, let them teach you, how to get by without them.

E.g. practise finding North by everything except the compass; then check your guesses with the compass. Try bush North repeatedly.
E.g. Guess at "adjusted Moon North", then consult the compass.
E.g. with a watch you can practise telling the time by the Moon.
E.g. before you use a protractor, guess the angle.
E.g. gauge the distance by eye, then check it on the map.
E.g. attach a sundial compass to your magnetic compass, for the sole purpose of practising & checking your use of the sundial.
In general, predict and guess before observing, and learn from the results.
You should **always check** an observation for good sense, and check any instrument, or any calculation — but prediction forces you to think **first** (rather than the go-to-sleep-first alternative — which is to always check afterwards).
Control your instruments — don't put yourself onto auto-pilot — or you will hypnotise yourself into blindness.

Look Back Regularly

Look At how you got here — so that you can recognise the way back if you need it — and can make sense of your walk.
At the change from one terrain type to another i.e. at the link between natural sections of a walk — after you walk through the change.
At the directionality of the sky haze down low — the change in thickness and colour.
At the colours of the leaves, lichens & flowers — they change dramatically upsun/downsun, and therefore forwards/backwards.
Look Back… At tracks joining in at an acute angle — which otherwise you won't see until you come back — which will confuse you.
At the bending of the track.
At what you may have dropped or left behind! (Check each other. Anyone carrying loose items should walk in front.)
At the distance come, compared to the distance yet to go.
At the weather clouds stalking up on you from behind.
Can you see your starting point, and get a bearing on it? What is opposite that — i.e. the landmark which now shows you your exact trend to date?

Look Around

Since turning your head while concentrating forwards, taking your eye off the ball or turning a smooth corner are all enough to thwart your direction sense, the anti-dote is to deliberately turn your head! Keep looking around at the wider context, including behind, so you don't lose sight of it.

Imagine an Aerial Photograph of the Continent

Visualise exaggerated shadows lying parallel across your imagined photograph — connect those with your shadow and tree shadows.
Visualise the Sun shining like a flashlight upon your imagined continent map — connect that Sun with the Sun you see (Don't look at it!).
Visualise the cloud shadows moving across it — there's one! See?
Draw the isobars across it, lined up with the low clouds wind direction — feel it in your face.
Superimpose a huge compass needle across the continent map — look at your needle. You should be aligned.

Usually you avoid using any small-scale map, since it is the real and full-scale world you want to see or visualise.
But sometimes, imagining a map helps, e.g.…

1. Your navigational map (i.e. remember what you studied!)
2. The mental continent glued to your compass needle, to help you '(not) swing the map'
3. The continental map you walk over at bends, to help you take in the new orientation
4. The bird's eye/satellite view e.g. when reconstructing where you might have gone wrong and when re-enacting it in miniature
5. Sleeping on a map, with your atlas open, and your compass needle showing
6. The Road Map marked out in Regional Triangles from peak to peak.

Making a M.U.D. M.A.P.-m.a.p.-m.a.p. For a more accurate job, see Mapmaking p101c,103c,106ab,109c.

Mudmap your trip onto paper, because paper remembers better than you do. Use any scrap of paper. Take a pencil and draw you meanderings.
Orient the paper to your expected trend (p51a Rhumb Line & p111c S.T.A.R.T.), so as to give you most room to map your journey.
Mark a dot where you start from — probably on the edge closest to you.
Up-arrow for a reference **D**irection; draw in a long one, Up-Sun is good. e.g. Up-wind; Up-shadow; Up-peak to a prominent landmark; Up-Weather **D**irection. (the Cloud flow); Up-North, by compass needle, or Up-Moon. *Interpret* the reference-orientation-**D**irection(s) before you leave.
U-p**D**-ate the arrow direction(s) or interpretation as needed, e.g. as the Sun shifts slowly, in relation to the wind, or as you walk past the peak.
Maintain a trend, a straight-line 'leg' p73b. Use the reference direction if possible. E.g. use the anti-peak.
Or pick a distant mark and walk up to it somehow; Or line up two points and keep them in line (p85e,87e,89e); Or just guess.
Align the **A**rrow when you stop (See Map Alignment 97a-105a, 103b). Turn **A**round to see where you have come from, to map the directions well.
If all other direction references fail you, **Align the next leg carefully to the last leg**. Be careful, because this will accumulate errors.
A-lignment is the magic ingredient. But it is the extra reference A-rrow which keeps you straight and avoids those accumulated errors.
Interpret the new alignment before you disorient yourself: "Where then is the starting-point now?" — or else you have a map but no sense.
Don't forget to use the map on the way back keeping track of the bends and where the start point is, and the finish point.
Pencil in the **P**rogress **P**arallel to the last leg; Pencil in the last leg *distance*. Turn around again and pencil in the *next* leg *direction.*
Guess at the distance, or pace it as you go, or use a pedometer, or a **stopwatch**. Add helpful notes as needed — p95c, 122 might help.
Repeat the m.a.p. for each new leg.

Avoid Disorientation: OverNavigate

Your aim is surely not: 'how to read a compass' etc. But: 'how to avoid a disorientation episode'
Being 'lost' is no panic, although you try to avoid that. Being 'disoriented' is panic (even when there is no real need).
To avoid the psychological syndrome and the real dangers of being disoriented, and lost, the most sensible attitude is…
Avoid getting disoriented in the first place Maintain orientation
Navigate carefully e.g. learn 'how to read your compass' etc.! Overnavigate — don't rely on one theory, one skill, one faculty
Don't give in to panic, or even to anxiety.

Over-navigating when you can, makes it easier when you can't, e.g. sea-sick or injured, bushed-in, exhausted or confused.

It is a rule in life (Murphy's Law) that things go wrong very quickly, not just one thing at a time, but all at once, compounding the difficulty.
As soon as you realise you are disoriented, you'll probably also realise that you also don't know the time, the position of the sun, the direction of the wind, the direction of the clouds, or whether you can interpret the bushes for noon. Then it begins to rain, go dark, and out come the leeches, and you sprain your ankle too. You could learn this law the hard way, or take my advice and do your 'emergency' navigating before the emergency!

How to Get Lost — Follow the Track— A Track is a Trap-Door

A track will lure you into the middle of the wilds. You end up depending upon it completely.
I suppose tracks know where they are going, but you may not, and they can't think. They are mute guides of the blind.

Antidote: Over-navigate. Do more than you need to. **Don't just follow the track mindlessly. Don't let it do the navigating for you.**

Keep your finger on the pulse e.g. How far have I come? What speed am I averaging? What landmarks can I see?
Have I checked the compass lately? The weather?
STAY ALERT
STAY SKEPTICAL

Reorientation; And a Fresh Start

Step 0. Admit it: "I am dis-oriented!". Say So. Stop. Decide to 'reorient'. Discuss it out loud.
Step 1. RETHINK where you might have gone wrong.
2. IMAGINE the true situation.
3. IMAGINE A BIRD'S EYE VIEW of how it 'must' be. Lay it all out, on the ground, in your re-enactment.
4. USE THE MAP; and all the clues — the time, the clouds, the distance…
5. ACT IT OUT on foot, in miniature. "I left this way, turned…"
6. SAY OUT LOUD a) Well that's where I thought it was b) But it is really over there (e.g. that Sun there).
7. Imagine the stars circling where they ought to; the coasts trending the way the compass tells you; etc.
8. Aim for a NEW START "I don't know what happened, but now I divorce the past".

Only when 'the ground shifts under your feet', do you know that your brain has got hold of the new situation.

How to Leave a Message

- Put it in a prominent place (e.g. a hilltop; a campsite; a rock face) with a prominent marker — bright and colourful for instance.
- Make it helicopter-visible e.g. in a clearing; large
- Use international distress symbols e.g. 'V'; Fire; 3 lots. At sea, a white flag or white stroboscopic (flashing) light.
- Any bit of metal may help you scratch out a message e.g. a zipper runner; a coin; a key; e.g. onto a tree trunk. If not, try a rock, or stick, to scratch with. Next time, have paper and pen with you.
- Use spittle, or urine, to make mud, then cover the mud message, but prominently — e.g. when searchers remove the coat from the tepee they see the message. E.g. make a flat mud pie; press in twigs to form letters; keep the rain off it.
- A freshly dead leaf is like paper, to write on.
- A piece of charcoal makes a handy black chalk.
- Brittle Yakka gum makes a bright orange powder splotch, to attract attention, and can also be used as a somewhat waterproof crayon.

Disorientation in new territory

Familiar territory does *not* disorient you, so you must *have* the mental faculties to 'know where you are' when you can 'see where you are'.
The overload of *novel* surroundings must overwhelm your normal abilities to mentally keep track of small diversions and accumulated progress.
These suggestions aim to make an exploration more 'familiar' and less 'new'. They come from all over this book. I suppose the message is:
'Don't just read about the methods, take extra time and care to put them into practice; extra 'bother'. Slow down; try harder. Remember more'.
Look back *very* regularly, to increase your familiarity of already explored territory. If you had eyes in the back of your head, it should *all* look 'familiar'. Walk back over that territory before continuing, p 68b. It will feel like old familiar ground if you bother. You could find your way through, and back through, an obstacle course at night without a torch, *if* you went forward three legs and back two as a matter of habit or necessity. That's 5 whole journeys. Talk to the obstacles along the way, and use a mnemonic to recognise them, and what you did about them, and how you turned.
Leave visible markers behind you, so looking back is more helpful to orient you. Leave a trail. It boosts your confidence, even if you don't use it.
Look through the foreground into the distance to **maintain contact with reference directions**. Fix what you can see, especially ahead, in your mind, in relation to the Sun, shadow, wind, etc, until you can point to them from behind visual obstacles. Remember more of what you see. Interpret each clue for its relationship with your memories of how things looked. Integrate the bits into a whole. Use some distance measure.
Maintain a straight line if convenient, so the interpretation of where you are is transparent. Use long straight 'legs' so you can interpret your accumulated path more easily. Notice and remember the angles *between* legs especially. Create a mudmap, so you remain oriented to your start position, and to your current position, and to landmarks sighted along the way. Practise these and other techniques, when it doesn't matter, for later.

To Work Harder, Rest Harder

"Mental alertness" may involve

- A siesta nap!
- An early night (not a late-night campfire party or sing-song).
- Regular rests during the walk.

Tiredness, Exhaustion, Constancy — these stop you thinking efficiently, and affect your memory badly too. So do nicotine, alcohol and drugs.

- When you are in difficulties, sleep on the problem, or take your mind off it with a diversion, occasionally.
- For the sake of having stimulating conversation, the whole party ought to guard their mental alertness too.
- Navigation games do more than orient you, they wake you up.

Be alert to your biorhythms—daily, monthly, seasonal—they affect your alertness or make you depressed—all your mental faculties can be affected.

Minds Work Best with Intelligent Questions on Them

- Am I judging the scale on the map properly?
- How many creeks have we crossed? How many more to come?
- What direction will the next creek be flowing in?
- What is the nearest escape route? To where?
- What does this place remind me of?
- Will it frost tonight or not?
- Where would the animals drink, and shelter?
- What is this place's own special uniqueness, and why?

How high and steep is the next climb?
What is our next position-check landmark? How soon?
Could we see Mount X from here?
What side-trips would people make?
Why is it sandy soil?
How fast are those cloud shadows going?

Anticipate, with questions, the things you want answers to.

Look Through the Vegetation, to the Contours

- Bushes cloud your vision. If, for example, you want to 'stick to the ridge', or want to memorise a junction, then look for the more **permanent** clues…
  ..like slope, creeklines, direction trends, terrain, soil & rock types, alignments, distances, context.
  Of course, look at the vegetation too, but not only at the vegetation.
- Trees cloud the skyline. But there are gaps and lines of sight that open up as you pass by. Look through the forest for the ridgelines, slopes, and horizon. As you move sideways watch the parallax—to reveal otherwise hidden contours.
- Look through the trees, to the **most distant**, largest, tallest features to navigate by — you will undoubtedly catch glimpses of these much more often than you think, if you are looking. They may be other trees, so get used to recognising individual trees.
- Examine the height of the sky beneath the canopy.
  E.g. the lowest sky is 'downhill', but **the lowest sky uphill** is toward a pass or gap, & **the highest sky downhill** shows up the ridge.
- Forest-bound **river beds** are often clear of trees and give you a rare distant view.
  Stop in the middle of that rope bridge! And hang around, to look around.

Connections to the Outside World — You Don't Need to Feel Isolated

Ruins connect you to the past. Earthworks and mineworkings also. Even the landscape engraves its history for you to see.
A **GPS** connects you to satellites; a **radio** to a radio transmitter direction; a TV antenna points to the transmitter; a **phone** connects to the grid.
The Sun, Moon, stars, planets + time + vertical can give you an **astro-fix** to connect you to latitude and longitude — globally.
The **magnetic field** connects you to magnetic poles, as do the aurorae. The deep blue polarisation band finds the Sun for you. So does a rainbow.
Navigational antennae link to flight paths. Flight paths indicate airports; pipelines indicate water supplies; electricity lines go to substations.
Paths point to points of interest. Recent tracks betray traffic. Track-markers put you on track; tracks connect to roads; roads to highways; to cities
Nightlights betray townships and cities; traffic noise links you to highways; surf noise, to the beach. Tides tell you about the Sun and the Moon.
River gravel shows up the upstream geology. East West gully shape shows up North; **rivers** connect you to the coast; **ridges** to peaks.
Air pressure here depends upon altitude; and upon the polar front and the pressure systems. Air temperature relates to the size of the landmass.
Absolute humidity tells about the proximity of oceans. **Cirrus clouds** show up the (westerly) upper winds. **Low clouds** and winds remind you of the isobars on the pressure map. Ground winds telegraph the topography. The radio links you to **weather information**. Waves, to distant storms.
Fence lines connect you to farms; boundary lines to survey grids; trig points to **maps**. Smells to camp fires, marshes, farms, abattoirs, toilets, seaweed, water. Stock trails lead to water; seed eating birds lead to waterholes; the plant and insect types are tied to their range and the region.
Vegetation is tied to altitude, latitude and aspect. **Bush north** indicates noon. Bush slant indicates the **prevailing wind**. Recent wind leaves tracks of yesterday. Deep soil temperature is related to altitude. Why feel "lost"?

OverNavigate: Equals Over-Familiarise

Overfamiliarise yourself with your route as you walk it, the first time you walk it.

Familiarity is the navigator's best friend, next time.
I.e. Years later you can more easily retrace your steps. Even when the track has overgrown.
It is an investment opportunity now, not to be wasted.
The extra time and effort will pay dividends, but like all good investments you must have an interest in the future.

E.g. Notice how and where you went wrong — mentally photograph it.
E.g. Lay to heart the most permanent features — the landscape and trends — not the track, its signposts, or bushes, which may change with time, or with a bulldozer.
E.g. Mentally photograph the lie of each junction, as if a fire might burn up the signpost.
E.g. Describe out loud your short-term memory about how you got here, lest that familiarity flits with the passing of the day.
E.g. Have your own 'track-notes' book.

How to Get Lost — Follow the Leader

A leader is like a track. You can follow him mindlessly. He ends up being a seeing guide for the willfully blind.
Ask yourselves, often, this hypothetical question: What would happen if the navigator dropped dead?
Could anyone else navigate competently?
Do others know where we are? On the map?
Could I find my own way back? Where are the cars?
If the 'leader' himself gets disoriented or lost, noone will know where they are!

Antidote: Teach. Learn. Test each other. Involve each other.
Demystify the magic art. Share it. Talk about it.
Listen to the debate. Look at the landmarks. Contribute.
Don't walk fourth in line. (Rotate regularly)

"Here, you find the way — I already know it"

"As Solid As the Rocks in My Head"

"Who am I to argue with the Sun? to say to it 'You should be over there — not where you are!' "

Telling your head off is the sort of attitude you need… when you find yourself walking uphill to the campsite which you know is down by the river!
… when the clues are shouting at you "Something is wrong here!!"

Navigators who are disoriented prefer to believe in such remote but real possibilities as a compass error or a mistake on the map, signpost or buoy, or to believe in the majority opinion being wrong, rather than to believe in the closer and more real possibility: "Perhaps it is ME who has made some mistake".

True, compasses & maps are notoriously prone to error, but your internal feelings, vague recollections, position/time/direction senses are more notoriously unreliable.

Lost? Rule 4: DON'T Do It Again!

Make resolutions. E.g. 'I will always take a compass.' (I tie a key ring compass to my hat string and I always take my hat.)
(I also tie a whistle to my keyring)
E.g. 'I will buy a map/mobile-phone/EPIRB'
E.g. 'I will always leave my itinerary with friends'
E.g. 'I will learn more about navigation; take pencil and paper; practise'
E.g. 'I will never go on a walk unless I have asked: "Which way am I leaving; Going in what direction?" '

Even better… Don't wait to be forced into a sensible habit. Declare your present attitudes to be 'not yet fit for human navigation' and put it right, now.

A Three-Track Mind

Here is a trick to avoid attention-deficit syndrome in your navigating:
Practise concentrating on two things at once.
Your feet find the path by themselves.
Your mouth talks about how good the meal will be.
Your mind sees the scenery without missing it.

Dual tracking your mind is like many other good things — just a little bit too hard to persevere with under normal circumstances, but possible to achieve, with extra dedication.

Don't automatically get lost into only one train of thought.
Slate this down for practice, for the rest of your life.

Interpret!

E.g. If you wear a face mask in these days of extreme UV radiation, or sunglasses, you become aware of when you are walking into the Sun. The obvious interpretation, hanging on to this piece of otherwise subconscious information, is: "I must be travelling toward the Equator". It's a chance to orient yourself to the new orientation. Can you see your shadow ahead of you?—you are getting closer to the pole.
E.g. You might be in a deep dark forest and see a rare bit of sun-dapple piercing the dense canopy. That means you can go over there and see the sunbeams for a direction check.
E.g. The 'crunch crunch' underfoot changes to a 'clop clop' — "Oh, if this is rocky, maybe I've passed a break of slope."
E.g. Exotic plants betray human occupation — you could look for an old access track.
As a general rule: **Look for any asymmetry**, and interpret it — usually for direction. E.g. When swimming with goggles, twinkly bubbles mean "upSun" — dull ones, "down-Sun". The wave and ripple pattern indicate wind direction. Sunbeams show up the down-Sun direction.
E.g. Scuba-diving: The anchor rope shows up-/down- current; alternating current is the swish/swash from the longest wavelength waves—note their direction. Direct current equates with "away from my starting point". Bubbles show "up" (with some adjustment for current if needed).
Interpret the whole story of settlement: Valleys attract farms, dams, fences, boundary roads, houses, access roads. I.e. you see a lonely fence post on a mud flat, and you say "once several paddocks probably converged onto this old watering place—look for old tracks near the fence post". Look for old nail-holes in the post, to find out the direction of the fenceline.

Play the Field-Scientist

Vary your **focus-distance**.
A herpetologist will see snakes at his feet & frogs at 10 metres.
A geologist will see the skyline, and the creekbed outcrops.
An ornithologist will look upwards, a botanist downwards.

Scientists see more, because they know what to look for, and how.
The more they look, the more they see, observe, discover, and learn i.e. the more they learn what to look for.
You can begin the self-reinforcing cycle in yourself by: wanting an explanation, by trying an interpretation.
And, of course, don't forget to look into the topic, into books about it and field guides for instance.
Be interested in all levels, in all scales and directions.

Focussing keenly will improve your navigation.

Orientation is More Fundamental Than Position Sense

We all carry mental recollections of recognisable spots—like a mental photo album.

But try to mentally picture whole stretches of a country—roads, coast, range, river—with a sense of direction & relative location & linkage.

Better is to carry a library of whole mental maps—walks, paths & layouts—not just photos of isolated spots or isolated linear features.

But Best is to keep all those maps in right relation to each other in your mind.

As your ability improves, you focus less on spots and more on the links between spots;
less on position fixing and more on trend analysis;
less on separate techniques, more on integration.
Direction clues are only helpful if you can coordinate them with what else you need to locate.
Positions only make most sense if you can say how they lie, in distance & direction, in relation to other positions.

Think in Pairs, in lines, trends and tracks

Two landmarks lie on a line, of known direction, and fixed track.
We tend to see things individually, and not see the join between them.
So try instead to get to know
the *direction* of 'camp to that moored boat',
the *bearing* of 'Mt Big to Tower Hill',
the *angle between* 'sunset and the headland'
the *trend* of 'Razorback Ridge'
the *pointing* of 'tree one to tree two'
the *line* from 'start to finish'
the *track* from 'Lake Low to River Junction'
the *tilt* of a star-pair
the *join* of two subsections

Needle-North, -South, -East & -West. Identify them on the Landscape

One of the major problems of navigating in bush, is keeping log of estimated side-slip, distance-wise, even if you are oriented, direction-wise. But you can see your sideways movement, see p53a, if you plot known directions onto known points on the skyline, (E.g. Magnetic N,S,E,W require no numbers; E.g. your initial forward bearing is already there waiting to be used). Or better, use points in the middle foreground landscape — about as far away as you intend to walk. E.g. local hillocks, towers, houses—things which will be visible from time to time during the whole walk, the closer the better. Make a note of their directions. Use these spots later to find home or your starting point. When you see them later, in relation to compass, you can visualise how much distance out there you have moved and in what direction. Page 106c “Tom Thumb Parallax” shows you how.

Orientation to Your Start-Finish (Rhumb) Line—by sight or compass

Work out the start-to-finish visual bearing to the destination if it is prominent, or to a landmark or point on the skyline beyond the destination.
Whenever you can see that point, you can see your left-to-right sideways positioning, in relation to the fixed track from start to finish.
A simple way to keep track of your sideways movement, at least in open country, is to choose an approximate rhumb line compass course (from start to finish) and watch it move across the landscape in front of you. E.g. “The follow-me arrow is now pointing more to the *left* of that gulch”.

Orientation to Your Finish Point, by the Parallax Method

If you are making a one-way journey, use the same “parallax method” as above, to keep a log of the estimated sideslip, but use your **destination**, not your start point, for your reference set of bearings. As the walk progresses, you can see yourself closing in on the correct destination.

Exponential Anxiety — the 11:59 syndrome

Your fear of being ‘lost’ rises in inverse proportion to the need for fear, as you approach your goal in unfamiliar terrain.
Will it be there as expected? Why haven’t I got there already? Surely it should have shown by now!
Have I overrun it?? Maybe I just didn’t recognise it? Has it disappeared?
When will this track ever end? Will it end at all?? Will I have to walk all the way back! And so on.

Now, obviously, if you have done your homework well, it will be there, where you thought it was, but maybe not exactly when you thought, so optimistically, you ‘ought to’ arrive.
Distance guesstimation is our weak point.

Don’t add needless anxiety to your background worries. Wait and see it sort itself out.

Anxiety, fear and panic work against good navigation.

Explore Junctions, While You Can

It only adds a few minutes’ pleasant diversion to travel the branch tracks which you don’t want. (Especially if you can dump a heavy pack)
You may never get to find out about them any other way.
This includes natural junctions—of creeks & ridges—and canal junctions.

When you are exploring, you may never retrace those exact steps, so don’t say “I’ll explore that on the way back”.
The other reason not to say that, is that you will be even more tired later.
When you are uncertain of your position, such exploration ought to be considered “essential”, now — not later when you realise your mistake.
Later you may speculate about what it might lead to, but now you can prove it.

In any case, as a matter of principle, explore far enough so that you don’t have to repeat the foray
E.g. walk left, until you are sure you should not go left.

“Meet back here in 10 minutes” is one way to investigate in all directions at once.

Separated? An Agreed Plan Helps

Example: Where did I park the car? Is it stolen or am I lost? Are the others worried yet? What will they expect me to do?
Example: Our driver has driven off! Why? What will we do? What will he do?
Example: She has not returned on time! She was exploring that way.

Try these out...
“We will, for at least one hour and at most two hours, try to rejoin each other before changing to panic-plan”
“Where we were last together, is the agreed rendezvous now”
“After that we check our intended destination for messages”
“We know to communicate via a 3rd party—a 3rd person, a 3rd place, or a message bank… ”
“..the nearest post office/petrol bowzer/pub/police station/emergency service/Fire Brigade/bank/Red Cross/Salvation Army”

Concentrate Outwards

Both introversion and extroversion belong elsewhere.
If you cogitate inwardly, you become blind to the surroundings.
When you chatter away about other things, that's equally distracting.

Don't spoil the special time of being outdoors.

While it is there in front of you, look to see what is to be seen — and talk about that — about the history of the place, or something about it that helps to keep you oriented, to its unique nature.

Learn not to be distracted from the main game.

SPELL Out Numbers — Into Words

The idea is to express, in words, what numbers measure. Humans hate hearing numbers! So translate mathematics into user-friendly words.
e.g. 270° is "West". That's much more meaningful.
10 past One — "Oh, it's midday!" (in your locality, for summer-time) 10.5 km — "almost half-way"
45° South — "sub-tropical, even temperate; halfway to the Pole" 020°T — just to the right of magnetic North.

PAINT your Words — Into Pictures

E.g. To make it even more meaningful, "West" = "Sunset!" (a word-picture)…or even better……………………
"019°" = "Where the Sun is just before it reaches magnetic North"
(This translates numbers into a familiar and visual time of day.)
"Midday" = "High noon!" (How high?) "Half-way" = "Within sight of the Gorge"
"Halfway to the Pole" = "The Sun rises and sets at 45°. The pole star is halfway up"

'Meaning' usually means a 'picture'

Words without meaning are like looking at the brushstrokes of an oil painting through a magnifying glass, missing the painting!

Here!
(Picture it;
Don't look at it!)

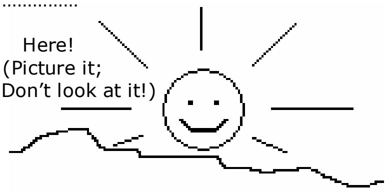

Look at the Ground

- Soil types; rock types (p. 79a)
- Animal runs; animal footprints
- Fallen track markers; toppled-over cairns
- Human footprints e.g. yours, on the way back
  Examine each other's boot treads, then the boot prints. Talk them through — expressing the recognisable peculiarities.
  Practise also on car-tyre treads and prints. (p. 66b)
- If you see a deep trench or a dry mud-hole in Tasmania, that may mean 'an old track' i.e. where constant traffic had refused to let the vegetation grow back.

Look for Foreshortening & Overlap & Parallax

Looking flat on to a shoreline e.g. the other side of the lake, means that any slight inlet or headland is probably a major and very sinuous feature. Focus sharply on it, looking for clues to depth. Don't just run on past it. Move sideways e.g. 20m, to show up the depth, by parallax.

Make Common Sense of the Various Direction Clues

Maybe the navigator has his eyes glued to the compass — but you, you can say "Oh we're just going upwind".
Perhaps you can see six landmarks — but the sensible way to see them is e.g. "We are heading halfway between Mount William and our shadows".
Cloud direction is a good clue, but changeable from hour to hour, so there is no loss in saying "Diagonal to the Sun" which is also changeable.

Don't Degrade the Point of a Clue or System

E.g. The point of a watch, is to tell you where the Sun is. (Don't look at the Sun!)
The purpose of a mnemonic is to remind you to think, not just to remember or to recite.
The reason for a "midnight-rise-noon-set" system is to work with the full-circle—even when you turn corners or turn around. It images the sky
The benefit of a skyline method, is for when you can't see the skyline. You will glimpse it 10 times more often, if you are trying to use it.

The aim of a practical method, is to practise it, not just to know about it!

The Aurorae

The Southern and the Northern Lights are especially frequent during high sunspot years (maximum every 11 years.... 2000, 2011, …).
They are most prominent after midnight. They are best seen in 'high' latitudes (> 40°). One to three days after a solar magnetic storm is best.
The most common aurora is the most easily overlooked — it is dismissed as a 'pale glow' in the North or South, low above the horizon.
But it is not dawn — that is East. It is not from a city — because it changes. Sometimes the sky will be pale green below and rosy above (between the clouds) and pierced by 'searchlights' which move left and right, slowly but visibly. Sometimes 'curtains' appear, shimmer and disappear slowly.
All the rays seem to emanate from somewhere near the North or South Magnetic Pole, and it spans 30° each side or more.
It is an interesting direction guide.
You can estimate current magnetic activity by analysing sunspot activity. **Never look at the Sun**. Instead, throw only an **image** of the Sun onto white paper, as follows... Point your binoculars at the Sun, **without looking through them**! Then, as if you were throwing a minimum shadow of one binocular barrel onto paper, and as if you will **let the paper look at the Sun through the binoculars**, from a distance of say half a metre from the eyepiece, point the big lens at the Sun. Adjust its alignment until the sunbeams go right through it and come out at the other end where the smaller eyepiece lens is. Steady the binoculars properly, and focus the bright, circular, white, image of the Sun so produced, carefully on the paper, by using the focusing mechanism and/or altering the paper distance, so as to reveal tiny pairs of black dots, on the paper. The larger the dots the bigger the storm. Caution! Looking at the Sun will damage your eyes and even looking at the bright image on white paper requires sunglasses.

Orientation to Your Start Point, by the Parallax Method

Step 1: **Start off Aligned**: Make a note of Needle-NSEW, bearings of any other recognisable points, initial-bearing landmark, rhumb-line landmark.
Step 2: **"The Direction Has Changed Its Mountain!"** Instead of saying: "The landmark has changed its direction", first say: "The direction has changed its landmark". I.e. retain in your mind the initial bearing, then, whenever you glimpse that landmark again, you can see how far sideways you have moved off that track—by locating which landscape point that direction now points to. You estimate the sideways disparity across the land, because your compass needle now points to a certain distance to one side of the original mark. Your job is to estimate that distance out there.
Step 3: **"I'm moving through the landscape, out there".** Imagine a thousand of you dotted all over the landscape, all moving parallel, and at the same speed, say, at 4 km per hour. All your initial bearings to all those recognisable spots out there, are moving with you, leaving those spots behind at a rate of 4 km per hour, and now those bearings point to new spots a calculable distance ahead. In reverse, this lets you see how far you have moved sideways to any direction. Bring those distances back to your feet, and see how far away from "start" you are, and how to get back. (At 15 km range, if you are walking at 4 km per hour, you will leave the landscape each side of you, behind at a rate of 15° per hour, but...) From the satellite's point of view, you and your directions are (all) simply moving through the whole landscape at 4 km per hour, no matter at what distance. An alternative viewpoint is to say that every bit of the landscape always moves "equal, parallel and opposite" to your motion, relatively speaking.
Step 4: **Estimate the gap between parallel tracks**—See "Advancing Position Lines" p.108a. I.e. from the initial bearing (a track) to the new one.
—by taking a new bearing & using sin(change in direction) x new range. This converts the distances you can see, to a number of kilometres.
The answer, if you want an answer in numbers, is how far you have to travel at right-angles to the original bearing to restore the original bearing.

Nervous Nellie Navigation — Not Necessary

Anticipate a problem. Do something crisp about it. Clear your mind.

Examples: Continually looking at a compass needle — can be short-circuited — by dividing the route into short legs, p. 73b.
Continually gauging the distance covered — can be short-circuited — by making one prediction before you start, and waiting for it, p92b
Getting lost trying to keep track of many bends — is a symptom of not knowing the overall trend in the first place.
Continually guessing where that junction might have led to, is avoided, by exploring it when you meet it.

You never know when you may be totally distracted from your normal pattern of thinking — it is then that you will need a cleared mind.
Plus, if you habitually allow your mind to move on to the next problem, by dealing with the last problem, you avoid ongoing confusion.
Forethought can see the normal problems coming, and pre-empt them.

Look for the Essentials — the 'Directions'

You can't afford to miss them!
E.g. 'Turn left at the Y-junction' means someone has to be looking for it!
E.g. 'Head uphill after you cross the creek' means: notice the creek bed!
E.g. 'If you hit the main road you have gone too far' means that you were talking about lunch instead!

Directions are worth writing down as soon as you hear them: "First right, third left, 6 km, look for the track"
Try to give directions more sensibly: e.g. "Up Fish River Valley from the Park Boundary"
Try to use fixed directions e.g. "North West" "from the lake to the Peak" "between sunrise & Mount William"
Not variable or relative ones e.g. 'turn left'

A Whistle-Code for Contact Calls

Example: The track splits. It's probably just a detour. "I'll go left; You go right". You keep in contact by whistling, cheerfully & tunefully.
Use your favourite tunes agreed between you...

To signal: "It's me"	Try:	"I did it My Way"
"You Come to me"		"Oh Come All Ye Faithful"
"I'm coming"		"I'm Coming; I'm Coming… ..Old Black Joe"

			The rhythm is...
Radio **Morse**:	CQ = Please Contact Me	= " Dah-dit-dah-dit Dah-dah-dit-dah "	"C is Seasick, Q Forms a Queue" "Calling Calling, All radiOs"
	R = Roger — Received	= " Dit-dah-dit "	"An R dits"
	Wait	= " Dit-dah-dit-dit-dit "	"aWait f'r a bit"
S&R:	Come Back i.e. Recall of Party =	Dah — Dah — Dah — Dah	"Ree-Call-All-Souls"
Other:	International Distress = XXX =	Dah-dit-dit-Dah — 3 times.	"all X's hurt, all ex's hurt, all ex's hurt"
	SOS — real emergency —	di-di-di-dah-dah-dah-di-di-dit — repeated	"A bit of Save Our Souls isn't it?"
			See appendix for the full code.

City Glow

From out in the bush you can usually see the direction of about three cities or towns, by their night-time glow.
This works best in hazy atmosphere — virtually any night! — and especially on overcast nights and when no skylight will show.

You need to check a map before you go — for the nearby townships — then check the theory against the facts,
then you can find it useful when you need it.
E.g. you can hold a straight line at night — relative to any recognised glow.

Don't confuse it with aurorae, sunset afterglow, predawn Sunglow, Milky Way, asteroid belt (Zodiacal lights), headlights, local floodlights, bushfires.

Just remember to get up and look for what you normally don't notice.

Come to your Senses

..And it will be surprising what comes to your senses.
Smell the air — for seaweed, marshes, farms, abattoirs, smoke, smog, civilisation, scent trails, fresh wombat dung, humid air, perfume.
Listen occasionally. To help you hear, cup your hands to your ears. Selectively tune out the high-frequency noises to listen to the low frequencies; and vice versa. Listen through the white noise, and through the silence of night.

For occasional vehicles, trailers bumping, boats thumping, jets overflying, thunder rumbling
For constant traffic rumble of a highway
For ocean rollers thumping on shore. You can also often hear the difference between high and low tide, from the wave noise.
For running water, frogs
For wind noise
For generators, pumps, machinery
For farm stock
For the crunch under your feet, changing

Look **carefully**. "It is amazing what you can see, if you look". It is a most important habit for the navigator and is very rewarding.
E.g. you read what others have scratched over the sign that says "5 hours" and discover "8 hours"; or find old track markers almost overgrown.

TRANSFER Your Pictures into Reality

E.g. "Sunset" = "THERE!" even after sunset!

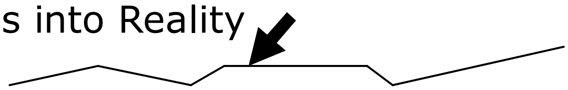

Find the red glow afterwards, further to the left e.g. at 8:30 p.m.

"High noon" = Look at how short your shadow is.
"Within sight" = "Well, can you see it yet?"
"The Pole Star is halfway up" = Do you see how the Cross doesn't set?

Living in the realm of theory is like an absent-minded professor — who can't find his way home, or remember which way he was travelling.

Put That Experience in Your Memory Bank

E.g. "The Pot will go down behind "Pot Hill"."
"I didn't realise how high the Sun can get at 45° latitude!"
"You can see the Gorge from the halfway mark, provided you climb the ridge"
"The Southern Cross keeps correct time at the end of March" — the beginning of Winter. See opposite for the next step…

Look, with Vengeance Aforethought

A bird observer doesn't just stumble into birds, he runs them down.
A native tracker doesn't simply 'notice' the clues, he searches for them.
Your task is to track down navigational clues, then bail them up.

Your enemy is: losing conscious grip on your task i.e. when the trudge gets the better of your brain.

For Practice: Express the unexpressed. It's no good to be only half aware of what is going on.
Interpret the meaning of each clue. It is no good staring at a compass needle, unless you realise where you are going!
Look outwards, to see what you can see.
Change 'a cursory look' into a specific search.
But, what exactly should you be looking for (once you have exhausted your repertoire of clues)??
Answer: Find out! I.e. **Teach yourself**. Don't wait until someone tells you.

Preparation is the Key to Difficult Navigation

Marine navigators in rock-strewn waters, or battling seasickness in heavy weather, or short-handed, know that the only way to cope with seat-of-the-pants reflex navigation is to have **already prepared clear instructions** about danger-bearings; back-bearings; lines of sight and distances to run, so that they don't need to leave the helm.

It is the total picture, beforehand, which orients you, later.

They need to plan the entire passage (so as to make landfall just before dawn, for instance) according to the tidal streams and ocean currents that they expect to meet (since it is the current much more than the winds which affect boat-speed-made-good, even for yachts). Their key desire is to sail with the current, and preferably with the wind and the weather systems, avoiding the storm-season. See "Aiming-off" p96a.

An hour seem such a long time to spare when you are busy at home; But when you get out on foot, relaxed, with hours to walk but no opportunity for desk-work, it seems that every hour not spent in preparing what could have been prepared, was a wasted hour, as far as being useful now.

Helpful Symbols for Standardising your Chart–Work

+	"The Deduced Reckoning Position" from course and distance information ("Ded. reckoning, or D.R.)
△	"The Estimated Position" taking into account currents and leeway to distinguish it from simple D.R.
⊙	A known position, or a Fix

————▶	Lines of possible position should have the arrow on the end e.g. pointing to a landmark
——▶——	A central arrow means a course. i.e. travelling, not pointing. One Arrow signifies the course steered through the water
————▶▶	A Transferred Position Line, i.e. advanced across the chart
——▶▶——	Two arrows is the course made good over the ground
————▶▶▶	(A doubly transferred position line, for a three-way fix)
——▶▶▶——	Three arrows is for the stream or the current.

Look Backwards; Think Forwards.

You sometimes, with only a mud-map to go by, have to follow a straight line by compass or landscape, to a low-lying, invisible feature, e.g. a tarn, hoping to come across it without bypassing it. Well, success will often make you feel that that is the end of the matter; "Mission Accomplished". Until next time!
But look back, if you can, to a recognisable starting point, then take a compass or landscape bearing, backwards to the start.
You then convert that to a forward bearing for next time. Place 2 pebbles on the ground in line, turn around and line up the landscape mark needed
The result is a "track", not just a direction — see p.71b — a 'can't–go-wrong' guide.
If your beginning point is not visible, turn around and take a forward bearing on the skyline feature you think you should have headed towards (and probably did). E.g. "One span to the left of Mt. Visible". The result is only a direction from the start—a guessed-at track—but nearly right.

On p.42d I said only to use bearings from high points to high points, for a general navigational framework.
But add to that, a mental or notebook entry of the bearing from a low destination point back to visible possible starting points, when you arrive successfully, or of the forward course you used — preferably by the landscape itself. They are "How-to-get-to (a low point)" tracks.

(Interpret)

Extend Visible Reality into the Invisible Picture

E.g. "It's boiling hot in Baghdad"
"The Sun is over Townsville"
"From the ridge the water flows down to the Gorge and eventually out to the Murphy Mouth"
"The Earth must be rotating Easterly, from West"

Don't be satisfied with half a meaning.
Until the interpretation leaves you oriented, and able to remember something, you haven't gone far enough.

See p. 71b(1) for the last word on interpreting 270T.

To Get There Quicker, Go Slower

We all prefer to run ahead of necessary preparations, i.e. we prefer instant gratification to wisdom.
We also prefer to get something unpleasant, finished, and quickly, once we have started, e.g. "can we make this bad "shortcut" work out?".

Don't rush!   "More haste, less speed" = More care, less fluster.
Slight panic makes you hurry. Hurry is a recipe for carelessness in navigation, and for blunders.
E.g. over-running the junction
E.g. missing most of the clues
E.g. taking a foolish "shortcut"
Common causes for hurry, other than anxiety, include…
"Let's beat the rain/the dark"
"Let's get back on time/for lunch"
"Let's sort this out quick-smart as to where we are"
"Let's explore one more kilometre before we admit we're lost and turn back"

Search and Rescue Signals — Noises Repeated Regularly

One = We are searching for someone. (Is it you?)

Two = I acknowledge your signal. "Okay; Roger received-and-understood; I heard that"

Three = "Search for me. I need help."

Four = Recall of Search Party

E.g. rifle shots; loud whistles or horns; two boards clapped together

---

End of Section 2.

Nature's Norths

Learn these when the sky — the astronomy Norths — is clear, so that you can **lean** on them later. (Assume S. Hemisphere 1,2,3,5,6)

1. "Dew" North — Dew, moist soil & roads stay wetter in the shade. Frost will persist in the shadiest places, e.g dew South-West.
2. Snow South — Snow remains higher on the shady side of an obstacle.
3. Mud North — Half-dried mud-crusts betray the hot, wind-exposed cracked edges, with the damp mud in the shadowed places.
4. Rock Noon — a whole mountain, or a crack in a rock, can show up the sunny and the shady sides. (See next box)
   — Also, in the evening, after a sunny day, you can feel the colder side of a boulder or gully.
5. **Gully South** — The shady bank of a creek is steeper than the sunny side, at least on average, all around the world.
   The side facing the noon Sun is flatter. The side facing the pole is steeper. This results in roads, railways and buildings etc, being more common on one side—the flatter side—of an East-West river than the other. In other words, creeks tend to cling to the Pole-facing slope of a ridge, as if they migrate equator-wards. Perhaps this is a remnant of a past geological age when the erosion was more marked e.g. hot days eroding the sunny side flat, with ultra-violet light; and constant moisture and icy nights eroding the shady base steeply; therefore suitable mnemonics are: "lying down flat sunbaking" or "steeped in ice" (i.e. South-facing in the S.H.). Lakeshores have steeper or flatter banks according to the same principle.
6. Animal Orientation — "**Barnacle South**" (SH), on dark intertidal rocks — it shows up the wetter i.e. shadier side.
   'Anthills' (Termite mounds) sometimes orient to avoid the heat, so the length of them is North-South.
   Insects prefer the warmer side of a trunk in winter e.g. spider cocoons, ant colonies, boring grubs.
   Some spiders spin their orbs to face noon, like satellite dishes but more vertical; others face sunrise.
   Migratory flight-paths may be North/South &/or upslope/downslope. Animals flee in predictable directions, e.g. inland, uphill, North, upwind e.g. when faced with a fire. They often lie with their backs to a cold wind.
   Siesta South: Farm animals find the midday shade in a paddock. You can smell the urea. You can see the dung piles.
   (7—10 will be expanded soon)
7. **Wind West** — part of the W in the 'WEATHER' mnemonic. It is designed for windswept places when the wind is not blowing.
   The prevailing wind leaves its signature in the visible growth history of the plants, and in the landscape.
8. Wave West — if, in the Roaring Forties, you can see the accumulated effect of ocean rollers.

9. Botanical Asymmetry — where you sort out the competing effects of wind, light, salt, exposure.
10. **Bush Noon** — you never need to be disoriented in the thick bush.
    This works best in temperate latitudes — where the sunlight is directional.

Of course, anything to do with nature, refers to **averaging continual observations**. The most variable signs are the least reliable.
Less Natural Norths

11. Magnetic North; And don't forget magnetic dip — the downward tilt of the magnetic lines of force, varying with latitude.
12. Solar Panel Noon — Observe how we set up solar panels, and verandahs. Nature gives us the Sun, and we all follow it.
13. Satellite Dish Noon — Geo-stationary satellites are, by nature, Equatorial.
14. Microwave dish to microwave dish — Peaks 'naturally' point to each other! See "landmark to landmark network navigation"p42d

Instead of feeling bored, or getting lost, or being 'bushed in', you can have the valuable bonus of feeling at home in the bush, having a hobby, and being oriented to the slope, Sun, vegetation, ecology, weather patterns…
Bush skills are a good investment for future bushwalks. So go for some experimental walks in the bush on overcast, foggy, rainy days.

Don't Forget to Remember, Your Way!

Long-term memory is your best friend — so feed it. You aim for recognition, years later, through familiarity now
e.g. Have you walked the track at night? both ways? twice?
e.g. Collect as much first-and-second-hand information as you can.
e.g. Reread your track-notes occasionally before you lose the memories.
Photographs, Maps, Track guides, Track notes, etc are all permanent on paper! but you aim for permanent memories, not records.
I.e. Mental photographs, Memorised records, What the locals call "knowledge".
Use memory methods so you don't waste your efforts. Use the paper records to help you consolidate your local knowledge.
To that end: Make drawings. Make your own map. Keep a diary. Make your own track directions, names & comments. Tell somebody.
Use biro not pencil. Make a neat copy of your notes and file it.
"Look at…" means "Notice…" means "Take note" means "Take notes" i.e. Take a notebook.

Unlock Memorable Chapters in Your Life

Memory is fundamental in navigation (alongside direction, distance, position sense, visualisation and alertness).
Medium-term memory is how you find your way back, but aim also for long-term recall.
We remember best in definite 'bites' — location by location.

> "Right, team; We have to remember this. Help me. We need to discuss it, describe it, mentally photograph it, draw its skyline, make it a memorable stop, time it, note it, put it in its logical place in sequence."

Go for the Jugular. Zero In on What Matters

Concentrate on the crucial and take-away information
— the crucial "directions" E.g. the sequence of junctions
— the details that have direct orientation value E.g. the visual angle between North and Start-off
— which give relevant memories E.g. the hill-top view laying it all out for you
— track notes e.g. the bearing to follow, the "directions"
— the overall trend — top-down, top priority pieces of information
— the perspectives that fit everything into the context. E.g. the surrounding trends in relation to Sun-North.

E.g. A mental map is better than a 'position' on paper. A map diverts you from the real picture.
Recognition outvotes names; Zero in on the most peculiar visual peculiarity
names outvote position fixes.
The overall details — What you need to know in three years' time, when you try again,
are the **time, distance and difficulty**.
The Sequence (of subsections & features) ought to be remembered... ..rather than e.g. 'bearing 035° magnetic'
Recognising Sections of the walk... ..is more relevant than having a series of map positions.
Surrounding landmarks last longer... ..than "turn left at the red tin-can-marker".

Appreciate the plants and animals; the soils, geology and climate. Instead of "Bin there; Dun that" you aim for familiarity.

The “Lie” of The Land

Think of this as the up-down part of geography. Think “Lie-Down” (versus “Lay-Out” i.e. a map layout).
Think of — Drainage, Watersheds, Catchments & Divides—the accumulated results of Ages of uplift, drainage, erosion & deposition.
— Geomorphology — it determines the shapes
— Ecology — it responds to the shapes, the slopes & drainage
— and how they all fit into each other.
E.g. “This north-facing slope supports warmth-loving vegetation”
“These foothills of the main range drain Westerly”
“The basalt uplift has generated these clayey mud flats”
“This is an isolated hillock within the main valley”
Think of the ‘lie’ as ‘down toward what general direction?’
and ‘rise’ as ‘up toward what encircling range?’
I.e. Put a general trend on major rivers and ridges. Imagine the coast, where all the (now nearly parallel) ridges & rivers run offshore.

Lift your eyes to the “Big Picture” scenery. Interpret it; Memorise its detail; Put it to good use.

Listen to the Pioneers

I.e. listen to the place-names

“The Great Dividing Range”	I suppose this one is obvious
“Blackbird Flat”	This means there is a river nearby
“Stockyard Creek”	A flat clearing is promised
“North Ridge”	This is related to the North Arm of the River
“Huon Valley”	It houses the longest tributary
“Boggy Plains”	Down on the flats!
“Eliza Plateau”	Up on the boggy flats!
“Bright Pass”	Means a change of catchment
“East Pillinger”	It ought to be East of West Pillinger
“Pelion Gap”	A saddle — between catchments
“StrathGordon”	It says “the Valley of” Gordon River
“-Dale”	Also says “Valley”
“Mole Creek”	Burrows underground

Lie Has To Do with Water Drainage

Look for dips in the skyline.
If you find a ‘soak’—where water soaks out—look for a ridge-bound marsh uphill—where water soaks in.
Carry an artificial horizon, to reveal slight down-slopes, e.g. a flat plastic jar lid suspended on three strings, adjusted to horizontal.
Or lay your clear plastic drink bottle on the ground. Maybe even pour out some water to see where it runs off to. Listen for frogs!

Stick to the Track

- If you are waiting for the rescue party
- When a “shortcut” looks inviting
- If you have told someone exactly where you are going
- When you are nursing an ailment
- When sacred or heritage sites are at risk from tramping and erosion, rubbish, pollution, desecration, etc.
- When you are one more tourist in the constantly invading hordes, and “Please keep to the track” is aimed squarely at you
- When you are in a greedy hurry for more of the same elsewhere
- If you can’t find your way out of a wet paper bag.

Reasons for Leaving the Beaten Track

- Tracks often fail to meet the needs of birdwatchers, botanists, photographers and bushwalkers. Remember why you are there. Talk to the ranger about whether the “No Camping” sign refers to drunken car-based parties, or to scientists and bushwalkers too.
- You sometimes want to search out the worst routes! Easy slopes have one sort of ecosystem/niche/microclimate; Difficult terrain has the more unusual biology.
- Interaction with a bare-dirt strip is not exciting. Fan out, Tread Carefully, Go Slowly, Keep your party size small — then you can appreciate the flowers, lizards, lichens. Each step is watchful — which forces you to see these things.
- A fast-track through a nice place robs you of appreciation-time. The experience passes just as quickly as the tramping to get there!
- Designed tracks will often steer you clear of exactly what makes a place interesting to see — cliff-edge views, water-edges, botanical hot-spots, refuge habitats,…

Isolated places often don’t have tracks, and you will need to choose, navigate and negotiate your own route.

Cross-Country Routes

All routes are ‘cross-country’ but some don’t have tracks to follow.
There are two sorts of tracks — Un-natural ones—artificial paths (man & animal made) which form a simple network,
and often stick close to the contours — gently upslope or gently downslope
And — natural routes (following the geomorphology) which cross the contours at right angles and
which form a double network.
I.e. Natural routes are determined by an interleaved network of ridges and rivers...
You are well advised to stick to these routes if only for ease of route finding.
Then there are three sorts of off-track navigation to consider…
— departing from the artificial paths
— departing from the natural routes (to follow a trend to a destination) and
— trying to keep a dead straight line cross-country...

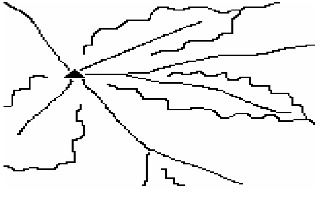

Rock Noon

Rocks blister and crack in the heat. In addition, ultra-violet light hastens all other erosion processes. The top sunny sides erode easily.
Conversely, the shady side is subjected to lichen growth, icing, and moisture, especially at the base.
Differential erosion results in tilted rocks, in asymmetry of cracks and gullies, and in various light and vegetation effects.

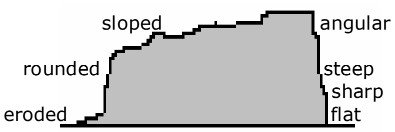

On the hot, rounded side, look for fissures, cracks, furrows, blisters and pock-marking — as signs of heat stress resulting in more rapid erosion.
e.g. Blistered flakes leave a patch of bare rock with no lichens on it.
Such patches are hard to find on the shady side.
Algae grow in the protected fissure under an almost blistered-off flake of rock.
The green stain on the underside of the flake is greener on the sunny side!

Caution: This is **unreliable** unless the rocks show a clearly interpretable pattern.
Of course, **use bedrock** or larger boulders, since they are going to be in their original position — not shifted about regularly.

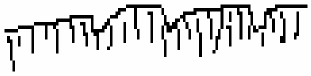

Examine the cracks — The shady side is sharp and steep; the sunny side is rounded.
Just like "Gully South" but in miniature.

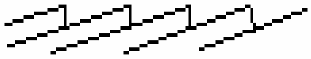

Examine the steps — sharp step-downs indicate a 'dip' in the layers down to noon.

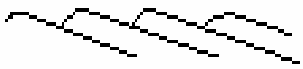

— rounded step-downs show a geological dip away from noon.

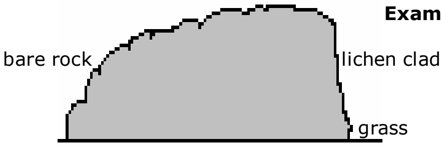

Examine the colours — looking down-Sun you see pale, bare rock (where it has recently blistered).
— looking up-Sun you see multicoloured lichens on a darker background.
— diffuse lichen grows better on crumbly rock. The sunny side will grow such lichens.
— thicker-bodied, textured lichens grow older on the shady, non-eroding side.

Feel the topside vs the backside for a difference in temperature.
Look for green grass and moss in the shade of the rock. These are often far more reliable than the rocks themselves.
Beware that a fire-storm will kill lichens, moss & grass, and will blister the rock, crumble it and split chunks off.
i.e. some of the effects you will notice may be due to a recent fire, e.g. a firestorm from a particular direction.
Rock noon only needs to be reliable enough to tell you which end of your improvised compass points North.

Don't Try to Memorise Details without an Overall Framework

It's a recipe for confusion — to try to keep track of every bend, every hillock, every weather change, every creek bed, every channel.
There are too many of them. They all look alike in close-up.
So start with the map: Learn the **major** rivers and the divides first.
Go for the big names, the overall trends.
Fit the details in by retrospect, after the big picture shows itself.

Don't Try to Remember Anything You Haven't Memorised

i.e. put the effort in first—don't fool yourself. There is a process to memorising something, e.g. establishing a framework first.
And...
Don't try to remember anything you haven't made memorable. Don't battle uphill—make it enjoyable or emotional and it will be easier.
And...Don't write down anything you haven't memorised. First make it memorable, then write it down.
Notes are not an excuse to forget to memorise something!
Like most short-cuts, they are an enticing alternative, but they, like shortcuts, short-change you as well. Memories are precious.
And...Don't try to remember anything you haven't expressed or named. It is a fact of life that we refer to memories by name, in words
And...Don't expect to remember details you haven't understood, or organised, structured, or reduced to a bare minimum—a prompt.
It always helps to interact with, analyse and process anything you want to remember. E.g. investigate it with questions on your mind; criticise it; teach it. It also helps to visually set it out in your notebook with colour, shape, balance and symmetry.

If you relax, because there is a track, or because you have a map, or you have 'directions' for the trip,
and maybe someone else is with you who is familiar with the area, then you cease to build up your own memories and familiarity.
Again you have short-changed yourself.

Write down What You Memorise!

Does that sound like over-kill? It isn't. It's just good sense.

- Memory isn't perfect. So help it along. Do so before you forget.
- You will want to write down your track notes, for years later, anyway.
- Writing helps your memory — it "involves" you more — it stimulates your expression, revises your experience, adds time spent thinking about it, involves your physical participation and your verbal and visual faculties.
- You never know when slight stress (I'm lost!) will distract and divert you for an unknown time.
- To write it all down clears your mind of the previous task — you can concentrate elsewhere.
- On the way back you can easily glance through your notes and then know what to look for.

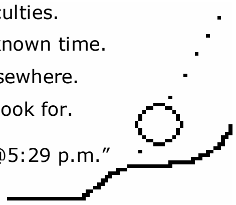

E.g. DRAW a picture of where the Sun sets on the skyline. LABEL it e.g. "Sunset @239°T @5:29 p.m."
Or was it 229°C @5:39? That's why you write it down.

Set aside a notebook for a "TRACK DIARY" (SH)

Major Rivers & Ridges, vs Minor Ones

Obviously, all water, falling on one side of a catchment-dividing ridge, runs to the sea, via a single river mouth; Conversely, river-mouth headlands are dividing ridges. Followed inland, all headlands must reach right around the river catchment—they are the two ends of the dividing ridge. The rivers and ridges form a sort of geometrically inverse pair of systems.
The catchments with the shortest creeks occur between named River mouths, and are called 'coastal'.
I.e. All other coastal land rises to some different, more local, dividing ridge; you may have to undulate up and down along it on the way
Not all creek-lines from a 'major ridge' (a catchment divide), descend all the way i.e. although the water may go all the way to the sea, the creek carrying it may not: the creek might only be a tributary, ending where it empties into a river. Similarly not all ridges running off the catchment-dividing-ridge reach all the way to the coast—some are "dead ends"—they end in a "shoulder", cut off between two creeks joining, or at a river bend. All valley-dividers are ridges, but not all ridges divide whole valleys, or even minor catchments. Those that come to a dead-end before they reach the coast are not dividing ridges, and may be called "minor".
All creek lines do not necessarily start on a dividing ridge: The creek itself might attach to a minor ridge, within the valley area.
Those ridges and rivers that transect or divide whole catchments are called 'major'. The named river ought to be a major line.
Unless you look for them on the map you may not be able to notice anything special about them.
Major ridges & rivers do however tend to stay higher and lower for longer, and often provide you with the **gentlest climbs**,
And they attract roadways.
Main ridges also tend to run parallel to the main valleys and vice versa.

Sort it out on the Map

- You cannot see catchment layouts from ground level!
- If you walk around a slope, half-way up, e.g. following an aqueduct, then you cannot easily know when you cross a major catchment. There are many minor ridges to cross and any one of them could turn into a saddle which divides valleys.
- Around peaks and plateaus, valleys spawn in all directions — it is hard to identify the right ones. High marshes may drain in several directions.
- Near a major river junction, travelling along the dividing shoulder can give a confusing alternation of views, between one valley and then the other.

Catchment Divisions Snake from Coast to Coast

Ridgetops fit parallel to and between the two highest contours marked on your map, and can be very sinuous.
They mark neighbouring catchment edges; & so theoretically you can cross the continent on highest ground without crossing any creek.
Major catchment divisions must reach the coast at a point each side of the river mouth (except, of course, for inland salt lakes).

A Catchment-Division* Separates Whole Valleys

*called a "watershed" in Australia, unlike in USA

Hint 1: Mark catchment divisions on your map with a pink highlighter — pink is opposite to blue-green for creeklines.
Hint 2: Now forget about those new pink delineations on your map. They refer only to the layout, as seen from above.
You need to imagine real, high ridges inbetween deep valleys.
Hint 3: Now forget about those razor-back ridges in your mind. You need to know that they could be remarkably flat paddocks.
— not very eye-catching — look for them carefully.
A catchment division will follow the highest flattest ground and consequently have a few marshes and puddles on it
A metre-high bump in a paddock might separate two raindrops into two different 1000 km journeys in each direction.
So you can't simply "walk downstream" — not until you know where you might end up!

Tracks are like Naughty Children

If you take your eye off them for a moment, they will split as soon as they are out of your sight.

A shortcut is often the result of a junction.
This is because tracks often bend at a T or Y junction.
Shortcutting is constantly encouraged at any bend,
So there will likely be an inviting well-worn shortcut there.
You can always split the party to explore both routes.

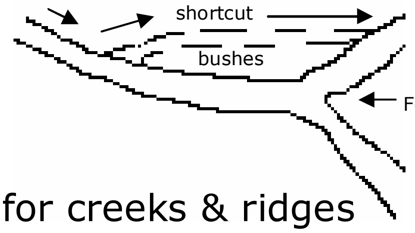

The same is true for creeks & ridges

Unless you stick to the bank of the riverbed, you cannot count on seeing the tributaries, and if you do stick to one side you miss the other side tributaries. Also, where a stream is "braided" into parallel interconnecting branchlets, and you follow one gutter, you are easily blinded to its tributaries joining in from side-valleys.
If you don't stick to the main gutter or crown, you miss seeing junctions on the far side of the river or ridge and can follow a wrong one. But side ridges are harder to see from the crest, just like tributaries are hard to see from the central channel!

Explore Branches While You Can

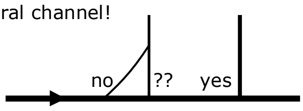

As a matter of principle, it is easier to explore something if you are there, than if you are not!
Be quick to remember that before you walk on past.
Remember… that you are there to enjoy the country. So put down your pack & amble off for a minute or delegate someone to explore.
You may surprise yourself by discovering a helpful signpost at a second nearby junction or a new access point and car park,
Or even, that it is the track you should have taken just now. Bends Breed Shortcuts: A shortcut may look like 'the junction we don't take' which means that you may mistake the real junction anyway.
Avoid endlessly having to wonder later… "But if that other junction turned east, then it might have been the link…" etc.
When you are there, find out for certain.

Convergent Tracks Lead to Civilisation

For example, you get lost, stumble across the track and meet a T-junction — which branch do you take?
You can examine any Y-junction for signs of splitting into two or more minor branches.
The most acute angle of a Y-junction usually points to a focus point (e.g. downstream; car-park)

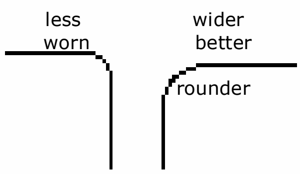

Down-Track vs Up-Track

This one is not obvious, and so it may trap you.
You seem to be following a well-worn path down, until you turn around and try to find your way back up.
It is then that you discover a multitude of other tracks joining into yours at an acute angle!
You see, people walking downhill, downstream, get channeled into the paths of least resistance, converging just like the creek lines do.
But people walking upstream or uphill are faced with choices and tend to spread out into all possibilities, until they find the right one.
The best path is worn in from down traffic, but that may not make it obvious to up-traffic.

Allow extra time when going up-stream, or climbing a steep peak, for eliminating false leads.

Wind West

Assume you are in a wind-swept place in medium to high latitudes e.g. southern Australia, where strong Westerlies prevail.

'W' in the 'WEATHER' mnemonic (p.114b) then reminds you
of the prevailing **W**esterlies in the high 30s, the Roaring 40s, the Furious 50s and the Screaming 60s.

'W' for '**Weather** systems come from the **West'** in temperate latitudes.
and '**Wind** in your face' locates these weather systems for you.
and 'Upper-level **Westerly Wave'** steers these systems to the E/NE/SE in a waving pattern.
A common pattern is to have a warm nor-westerly storm preceding a strong cold sou-westerly change (SH).
and '**Wind-West'** reminds you to look for the lower-level **prevailing Westerlies**, and their effect on the bushes, landscape, etc.

You don't have to be in the Roaring 40s to find a windy place and look for tell-tale signs of the prevailing wind.
— Onshore winds near the coast, being laden with salt, leave unmistakable clues imprinted in the bushes, and on the seashore.
— High plateaus and wide open clearings are usually windswept.
— Gullies produce up-gully breezes and down-gully gusts.
— Hilltops are windy. To get an undistorted picture of the prevailing wind direction, go to a hillock.

To navigate by the local prevailing wind may be more preferable locally, than to navigate by bush north.
The prevailing wind is not as universal or reliable as bush-noon, but it can be easily visible, audible and tangible.

In general, the **tallest trees** are the best tell-tales.
Look for wind-induced asymmetry**, especially at the tree tips** and the **oldest** tree trunks, when you are in a wind-sheltered zone.

Think Wind

Distinguish in your mind the various causes of wind damage…

- Storm Winds — brief but furious, as a weather change hits. They may uproot trees. They are often hot and dry.
- Fire-storms—from the hot dry inland, or sweeping uphill. They can leave a trail of dead black fallen trees lying parallel.
- Cold Change Winds — Cold, Gusty, Strong, Persistent for a day or three; Usually south to southwesterly (S. H.).
- Persistent Prevailing Winds — Westerlies 'Waving' between northwest and southwest in between 'changes'.
- Salt Laden Onshore Winds — They 'cut' the exposed vegetation back to size.
- Cold Blast Winds — These kill off the less hardy competitors.
- Hot Dry Blast Winds — These leave only the waxy/succulent/woody species.
- Sand Blast Winds — They remove vegetation, carve out sandstone sculptures, and remove lichens from the exposed side.
- Windy Season Winds — Ask the locals what to expect at each time of year.

Anything is Memorable

Make it personal	"Where I found the pocket knife!"
anticipated	Study the map beforehand to imagine the sections
active	Do something; involve yourself
important	Draw attention to why it should be remembered
associated	It looks like a … ; sounds like my middle name ; uphill from the mine
creative	Make up an enjoyable name
emotional	Expect to notice, then remember, decision points & anxious times
visual	e.g. symmetric, colourful, shapely
geometric	This triple-trunk matches the triangle of paths here
noticed	Draw attention to it "See how we have just entered tea-tree ?"
peculiar	That tree looks like a traffic policeman!
prominent	The tallest tree here, see?
liked	What aspect about it interests you?
named	e.g. "Midnight Junction"
fitting	"Power Point" where the electricity line crosses the lake
funny	'Willy Willy Waterhole' where we saw the dust-devil
human	"Betty's"

Make Individual Memories Memorable!

The problem is an excess of information, with a shortage of enough meaning to go with each detail.
e.g. hillocks and creeks have endless detail — you can't hope to remember all.
Visual memory is not the problem — pictures are perfect — but recall is difficult.
(And sorting pictures out later, into the original sequence is also hard.)
You need to add memory handles, worthwhile meanings, specific focus, etc.
Hint: Do something memorable there. Dance, Sing, Jump, Shout, Throw stones into the puddle.
Make a cairn. Sketch a tree. Have a meal — Chewing is said to aid memory!
Leave something of 'yours' there — something peculiar — e.g. lean a stick against a tree trunk; scratch the ground.
Insert it into a larger scheme — time & sequence; lie & surroundings; an acrostic poem.
Make a game of it. E.g. "Who can say what's peculiar about this creek?"
Make it 'yours' e.g. Discover it. Climb it. Explore it. Study it. Explain it. Draw it. Map it. Show it to someone. Name it.
Spend time there, looking.

Remembering Trees

Simply spend time looking. It may only be 10 seconds, but that is 2000% more than the usual ½ second!
Look at the branches, the shape, the trunk, until something strikes you
E.g. "Oh, these two branches grow across each other"
E.g. "The bark is right-hand spiral"
E.g. "See that ant-highway going up into the hollow—it might have water in it"
Walk around it, to see the shape from side-on. Characterise it & its context
E.g. "entirely shorter than that old stag's first branch"
E.g. "three tufts on top"

You get better with practice.

Ridges

Ridges have gentler slopes up/down ridge i.e. provided you stick to the crown.
The slope 'makes' rapidly on each flank. It is only the ridge-line which is flatter.
Ridges have nicer views. But these views are easily halved unless you walk right on the crest.
have nice breezes, provided you are right on the crest.
have more open and stunted vegetation plus shallower soil and more exposure to the elements.
are wider and flatter than a creek bed, and their rocks are firmer underfoot and dry.
Their run-off slides off downhill to each side.
These advantages are also lost as soon as you slip off the 'crown' of the ridge.
Off the ridge is steep, vegetated, narrow, loose, windless, hot, humid.
The steepest but best way up a ridge is... up the ridge because (1) it gets even steeper off the ridge
& (2) contouring is 'easier' but it doesn't get you 'up' but only into more difficult terrain.

Ridge Names

▪	Mountain to Mountain	e.g. Mount Ossa to Legges Tor (trace it on the map)
▪	Mountain to Mouth	e.g. Mount Ossa to Arthur River South Head
▪	Mountain to Headland	e.g. Mount Wellington to Wrest Point
▪	Catchment Division	e.g. The Arthur-Pieman Divide
▪	Tributary Divide	e.g. The South Arm Ridge

A ridge may not be "major" but if it divides your walk, you do want to know about it—so make up an appropriate name for it.

Mnemonic: "Ridges Rise"

Whenever you cross a ridge don't forget to ask: "Which way does this ridge rise, in trend terms — to the left or to the right?"
It is like asking which way water is flowing, but the truth can be obscured by local undulations in the ridge.
Ask: "Does this ridge, if followed to the (right), end further inland i.e. upstream, or further downstream?"
It must correlate with the overall corresponding drainage.

Uphill — Which Hill? Upstream — to which Divide?

To Which ridge? Which plateau? Which highest ground? Just as surely as rivers reach the sea, so ridges reach a peak.
Which peak? is the question which orients you.
Until you know which catchments you are climbing out of and in to, you don't have a grasp of where you are.

Sympathy and Understanding with Track Markers

Take time out to explore (e.g. leap-frog fashion) from one track marker to another.
Always guess where a track ought to go.
When you come across a marked trail—maybe poorly marked and/or difficult—the question arises: 'Should we follow it or not?'
Intelligent guesswork helps here

- Where would *I* put a track?
- What is the direction trend of the new trail?
- What is the slope trend of the trail?
- What tracks & routes might you expect?
- What alternative access routes exist?

E.g. uphill, heading West, may well mean: 'to Lake West from the lower car park'

If you leave the track behind, because it is unclear, you may never get back onto it.
But if you work with what you have (a bad track) it may be quicker in the long run.

Old Track Markers

Practise noticing them, before you need to.

- Old axe-cut blazes on tree-trunks — from the original surveyors
- Very old paint flecks in the blazes — search for them
- Blazes and markers on the far side of the tree — look back
- Old nail holes or nails or rust stains — the marker has fallen off
- Fallen markers — on the ground
- Fallen rock cairns — the rocks blend in, but show their pale underbellies
- Arrows on wet rock-faces — the wetness obscures the scratches
- Sawn branches, trunks and logs from when the path was cleared
- A clear line-of-sight through the trees
- Thick young growth — an unused road
- Smooth or shiny rocks with less lichen — worn smooth by feet
- Fallen trunks with a gutter worn across them — betraying foot traffic
- Broken branches of varying ages; bent stems e.g. bracken fronds
- Stumps cut off at ground- or eye-level.

Be Wind-Wise: Listen

Listen directionally — the wind-noise will be mainly from up-wind.
When you think you are navigating by the beach noise, you are probably navigating by the onshore wind, and should keep track of it.
Listen **through** the noise — You may distinguish three frequency bands, as the rumbles emerge from the hiss.
Listen through the **silence** — of night, in the stillness. Up-wind sounds carry a long way.
Listen to the **sounds** (birds, traffic, house, factories) — They carry clearly from up-wind only, especially the high frequencies (whistles)
Listen to the **beach** noise — The 'hiss' will be from a direction more up-wind, and closer, location-wise.
The 'rumble' will be from further away from you and further away from up-wind direction.
Listen to the **thunder** — Air temperature zones create directional effects, so that approaching storms, which are preceded by low level cold bursts of air, have their noises refracted up and out of earshot.
(See "Tips for Calm Weather" p.23d & "Topographic Level Winds" p.25d)

Obvious Tell-Tales: Landscape

- E/W coastal asymmetry e.g. Temperate West Coast ocean beaches might have a kilometre of surf, but East Coast beaches are narrower and calmer. Trade-wind East coasts are surf beaches, but headlands with West facing shorelines will be sheltered.
- Bays end in exposed points. Look for tall, steep, bare sandhills on the weather side. If there are tall sandhills at a narrow point, you will see sandhills from both sides, so ask 'am I seeing the green (sheltered) backside?'
- Sand dunes. There are variations, but bare dunes tend to be cross-wind and vegetated dunes tend to trend down-wind.
- Puddles, Lagoons, Lakes, Creeks, Beaches — Look for wind & wave action effects e.g. dark mudflats up-wind & white sand-steeps downwind. Shallow puddles often occur on the downwind end of a flat drainage.
- Sand ripples show steep backsides downwind, as the wind driven grains topple over the eroding crest.
- Clumps in sand-stretches show a long tail downwind. E.g. behind dead seaweed deposited on the dry beach.
- Grass-clump erosion is the reverse of sand ripples — look for steep bare dirt and exposed roots on the exposed side.
- Camps & campsites, picnic tables & fireplaces, bird nests & animal shelters all avoid the un-sheltered side of vegetation.
- Wind-shadow growth—plants grow taller in the lee of a mountain range, or of a pebble, sand ridge or house.

Obvious Tell-Tales: Windblown Objects

- Flotsam collects on the downwind side of a lake (but on the sheltered side of a coastal point!).
- Trunks fall downwind in storms; branches fall downwind in gales; leaves, nuts and flowers fall downwind in strong winds. Look for ribbon bark hanging in a tree downwind from its source.
- Rubbish (old dead twigs & leaves) collects behind a clump.
- Dead long leaf blades lie cross-wind when blown against an obstacle e.g. a clump...
- Dead long leaf blades lie down-wind when blown from a source e.g. a tussock...
- Broken off branchlets lie with their big spread-out twiggy end downwind. Meadow-grass tufts pulled up by grazing animals also.
- Feathers catch in bushes, usually on downwind branches, and mat against the up-wind side of those branches.

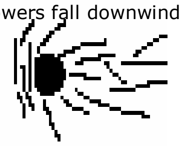

The Difference between Looking at, and Seeing

..has a lot to do with involvement.
e.g. a seed-collector will see the vigour of the plants and their stage of fruiting. Would you even see the plants!
e.g. a hunter will see a route for stalking from behind cover.
e.g. a photographer will see the contrast in light.

Would you as a navigator care to notice footprints, or their absence? What about the flight-paths & contrails, or gems in the river gravel?
Navigation is your opportunity to see more.

Search-Image

It is possible to stare at something obvious and not see it! e.g. the pattern of wind-blown grass stems, or the peculiar geometry of a tree. But after some practice, you can't miss it. So train your brain to the search images before you need them. When you are flustered and hurried you will not feel confident in your abilities to see something difficult, and will likely give up too soon.

We Recognise 'Sameness' But Need to Describe the Differences

A gum-tree is a gum-tree; but this gum tree has three trunks!
A 'bend on the road' is just a bend; but this one bends away from sunrise.
A 'hillside' looks much like all hillsides; but this one is regrowth with wattles dominant.

Don't just gaze at "the view" or "the landscape" or "vegetation". Look past the fuzzy 'sameness' to the unique peculiarities.

Unfortunately that often means: 'Focus on the boring detail'
e.g. 'This hillock has five dead tree trunks, totalling nine main branches' which makes this hillock recognisable and unique.
So. Hint: **When all else fails, COUNT!** (The same goes for botanical species identifications.)

There is often no way around a detailed look. There will not be anything unexpected to hit you in the face — you must keep looking until the interesting peculiarities emerge.
E.g. "No other dirt road junction I have seen has an Acacia sapling growing in each corner."

"Shoulders Stop"

A huge mountain range may end in a "shoulder" i.e. a 'shoulder' amounts only to a 'minor ridge' stopping within the valley.
You will not get to the coast by following it down. (If you are on the very last ridge next to the coast it would be called a headland).

There may also be isolated high ground within a valley, a lot like an island, and this too is doomed to end in only 'shoulders'.
Such high ground is drained by one river only, but via several tributaries.
It will have one, often low, flat connection (a saddle) 'uphill' to the main Divide.

Shoulders are Not Easy (like Ridges) To Walk on

1. They end a spur, and so are likely to be steep, as the ridgeline plunges down to the creek bend or junction below it.
2. They are wetter than ridge tops. Although they shed water, being convex, they are still fed with run-off from above, as on a hillside. So the vegetation can easily become very thick.

Saddle — A Flat between Hills, Sitting Astride a Ridge

They are so named for their resemblance to the two-bumps-two-gutter horse-saddle shape, designed for two legs astride a sway-back.
When a flat area is drained by three creeks and so has three ridges lifting from it, it is called a "monkey saddle".
The third gutter accommodates the tail. This can cause uncertainty-of-choice for navigators!

A "col" is a saddle.
A "pass" is usually a low point in a high ridge i.e. it is a saddle.
A "gap" however may simply be a river gorge.

Every undulation and flattening of a ridge leaves a saddle. In forest, there will be a puddle. It may also attract a campsite or a road.
A saddle may look like a valley crossing—a muddy dip in the path—but the two sides both drop away, if you look carefully through the wall of vegetation.

Weather maps show pressure 'saddles' between two High's and two Low's—p.11d.

You can also picture a saddle as a U-shaped valley going up and over a gap between two hills and continuing down the other side.
Saddles make creek sources, but not every creek source begins as a flat saddle—any ridge or slope may spawn drainage gutters without necessarily flattening out.

Track Markers

— Decode the colour-code. Separate tracks often have separate colours. Reverse track markers may be of a different colour.
— Successive generations of track marking may be evident, or several colour-coded tracks may share one leg of the journey.
— Triple marking, often means "End of Markers" or "End of Track".
— Big cairns mark junctions, campspots, features. Cairns with a stick may also indicate a junction.
— Side-tracks are marked with something extra or different.
— Sticks over the track, or especially crossed sticks say "No Entry. Wrong Way. Go Back".
At an unexpected junction, if you have to step over some branches you have probably missed the intended message.
— Un-needed signposts — often mark the entrance of disused tracks. They may (or may not) be at an unexplained bend in the route.

When You Lose the Track-Markers

Be suspicious! Start by over-navigating beforehand — search for all track markers as a habit, even when the track is well-worn and plainly visible. It gives you the practice for when the track marking does get patchy.
Also, only then will you notice whenever the track takes an unofficial shortcut (e.g. following vehicle tracks, not the walking track).
Or maybe you have taken a wrong turning off the marked track and need to realise it, no matter how good the new track is.
Whenever you lose the markers or whenever you lose the track, explore in these directions: **straight ahead and straight behind**.

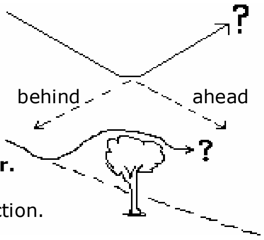

One or two markers may simply have fallen off — so maintain the trend
You may be approaching a main track junction with the first few sidetrack markers removed to conceal the sidetrack you are on.
An unexplained bend, which deteriorates, may be explained by a missed junction/bend...

Where a tree has fallen nearby, try to pick up the trend of the old track, on the far side...
Backtrack to where you went wrong — to the last bend & from the last certain marker.
Next, explore the unexpected directions.
Very often a marker will be placed at a bend, because of the bend, or even because of a Y junction.
Then explore in all directions — "Meet back here in five minutes" might save 50 minutes.

Notice (with Alarm!) When...

..The track itself **deteriorates**. You may be on a false track, made by people doing just what you have done
e.g. you overran a junction or bend; you sidestepped a fallen tree without noticing.
..Boot prints, that were there before, cease. Especially on a 'through' track.
..The track gets **better**! This means that in back-tracking you may miss your small track and follow the other track which joins in there
..The country abruptly **clears** e.g. a beach, a peak, a clearing, a road. There may be no way to find the single small track you used.
By the time you remember to look back, you may have lost the way back out. There may be too *many* paths to choose from.
..Any **barrier** e.g. a beach, a road, a fence, a riverbank — clear or not.
When you hit the barrier you do a right-angled bend, but later you may not remember exactly where you met that barrier.
..Any **catchment change** i.e. crossing a ridge, or even a creek crossing—into "another part of the catchment".

Name the enemy: Be the first to say out loud **"Abaracadabara! Peek behind you! because it's a...(fill in the blank)"** A **B**arrier; A **R**idge/River/Road Junction; A **C**learing; a **D**own-track (or focus point); A **B**etter/worse track; A **R**idge/River crossing; A **Peak**.

Light & Colour Effects

This is one of the delights of nature! Chances are that you will be startled, even if you know what to expect, but only if you look for it. You need to wade out into a uniform reed-bed, preferably when the Sun is side-on to the prevailing wind, then look up- & down- wind. Up-wind, the tussock tips are pointing toward you. You then see between them into the deeper shaded body of the plant.

- The view up-wind is dark & colourful. The best effect is from rich brown or dark green, drooping, chest-high reeds.
- Down-wind the tips lie flat and reflect the sky shine at you — faded & bright.
  The prevailing wind may result in permanent sunbleach on one side of the grass blades.
- On a windy day you will see gusts causing bright patches in a dark background, due to the same flat blade effect.
- To neutralise the directional lighting of the Sun, look down-Sun — to your left & right will show pale vs colourful.
- The combined effect of the Sun and the wind result in four quadrants — darkest up-wind; shiniest downwind; greyest up-Sun.
- This happens even with ordinary green plants — look for it in any low, flat, wind-swept vegetation.
- The prevailing wind doesn't need to be blowing to see its permanent effect on the plants.
- At sunrise in the reed-bed, you should notice a 'funnel-web' effect on your right — like the circular scratches showing up through a sunlit window.
- At other times you may simply notice that any shiny leaves on the left of the Sun tilt left, and on the right of the Sun tilt right.

- Looking up-wind anywhere is greener. Downwind looks more bleached — due to leaf-tip scorch, leaf fading and leaflessness.
- Weather-side sandhills are white. Lee-side sandhills are greener, due to thicker growth.
- Exposed lake banks are steeper & eroded. So, sandy lagoons show white sandy vs dark muddy ends.
- In sandy heaths, downwind shows up as windblown sand patches of erosion. Up-wind shows up as continuous plant cover.
- Rocks in sandy places are sandblasted pale on the up-wind side and darkened with lichens on the downwind side.

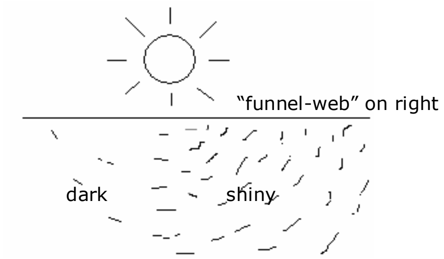

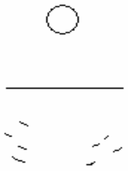

Shiny leaves

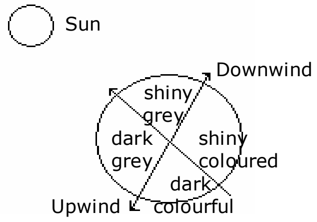

Expect some variations

Place Recognition
You Recognise What You See, By What You Can't See

Every locality fits into a context — the invisible surroundings.
How long did I take to get here? How far from the turnoff? In what direction from home-base? Uphill or down or contour? How many creeks have we crossed up to now?

To know where you are, you can't ignore where you are not.
e.g. 'It was an all-day slog to get there, we finished the day worn out'
is part of where you were, and fundamental to your memory of it.

Landscape recognition is relative.
That involves your senses of time, direction, distance, difficulty, slope, topography, layout, number and network.

Aspects, invisible to the eye, which surround a place, live in your mind's eye.

Description Requires Words
Can You Describe All That You Can Recognise?

E.g. a magpie. What pattern is on the head? What colour legs?
E.g. your brother — what colour eyes? what voice accent?

"Recognition" means "I need to see it before I can recognise it"
"Description" is "I can recall it to my mind and describe it"

If you don't deliberately memorise something you can only recognise it but can't describe it.
So thoughtful examination, notes and practice, can fill in a lot of blanks.
Repeated visits build up a subconscious picture, but you can speed up that familiarisation process by **deliberate scrutiny**.
Expressing the description in explicit words, forces your subconscious 'recognition' powers, into a new dimension.

Hint: Open your mouth; say: "The place where… " and see what comes out spontaneously.
try: "In between…" "Running alongside the… "
"You probably would remember it as the time when… "

Notice When You Cross a Catchment!

It is a major method of navigation
e.g. at night; or in a blizzard; or when you are lost; or in a forest; or descending from high ground; or exploring upstream.
But you need to get used to noticing ridges and to counting creek crossings and to matching the catchments to the map.
Notice also, when you cross from one side to the other side of a catchment.
Ask: Is this riverbed likely to be the lowest, largest channel in the whole catchment?
Look for creek-line vegetation, as sometimes the only clue to the drainage pattern.
Look for any darker patch in the road or track or beach, as a tell-tale sign of a dry creek bed.
Look at the track ruts. If the rocks are exposed, water flows over it. If the ruts flatten out, water drains away from the track there.
A subtle rise, dip or bend in the track can be traced to an obvious ridge or drain crossing it. Look sideways to appreciate the lie.
Assess any puddle, marsh or mud hole — Is it a creek line or a saddle top? — Is it in a dip or on a crest? — Can you trace a gutter?;
Notice any crest; Trace ridges to a nearby peak. I.e. Don't get caught whingeing about the government, instead of noticing the lie.
Ridges divide: they shed water. Creeks collect: they catch water. Valleys drain: they empty into the sea.
Rivers cut the catchment in two.
See "Understanding Contours" for some hints — later in this section. See also p.83b.

The Giveaway! A Change of View — Mnemonic: "Catch the View"

The new panorama is the tell-tale sign that you have changed catchments.
Perhaps you are contouring around a headland and catch the new view around the corner, or you are sidling around a ridge,
or you are breasting the saddle, or reaching the peak, or cresting a paddock… It is the new vista that greets your eyes
which tells you that something has changed. It is also the loss of the old valley view.
Well, maybe you can't see the view, but you may now be able to see the skylight or even blue sky down through the trees below you.
When climbing, you find yourself looking down instead of looking up. Your muscles, of course, tell you the same story.
The deeper the new valley, the more far-reaching, on average, the new view.

You Can't Cross a Creek Twice from the Same Side

If you make two successive creek crossings in the same flow-sense, you have crossed a creek divide into a different creek catchment.

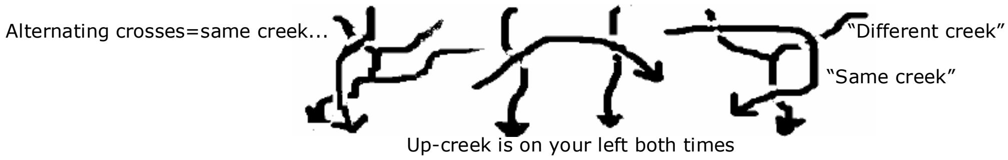
Alternating crosses=same creek...
"Different creek"
"Same creek"
Up-creek is on your left both times

The same rule can hold for ridges, but only those which slope uniformly down from high ground.
Otherwise an unexpected crest slope, instead of being a 'different ridge', may alert you to 'a local up-trend' on a down-sloping ridge.

Mark Your Trail

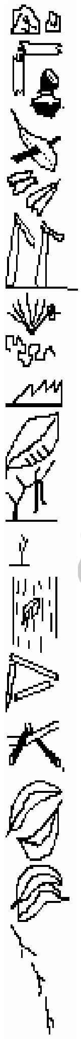
a pair of stones...
a gumnut on a rock
a cut leaf…
dead twig…

Two principles are needed 1. Trail marking takes extra **time** — budget for it.
And this one is widely applicable… 2. It's easy when you know how, so **think** how!

Mark Your Trail Non-Permanently

E.g. when you could otherwise get lost

- Use what is available — pebbles, leaves, tussocks — throughout the area, for ease and consistency.
  e.g. "always look for a pair of stones"; "tie a grass blade on the top shoot of any shrub"
- Leave your mark in a visible place. E.g. "look for freshly broken twigs at the base of each trunk on the path";
  "check the topmost leaves of anything chest-high for a mark"
- Leave your mark in an expected place. E.g. cross-country: Zig-zag in straight lines between prominent features,
  so coming back you know exactly where to look for the next markers, and searchers do too.
  E.g. hillock, to tall tree, to clearing, to outcrop.
- Do something only slightly unnatural: (see diagrams) A gumnut on a rock; a twig through a leaf; a pair of torn leaves;
  grass stems bent in a consistent direction; or tied in a knot; or crumpled; a scissor-cut dead leaf; a scissor-cut fresh
  leaf; bark hanging from a barkless branch e.g. gum bark in a Banksia; a twig drilled into the ground;
  a leaf wedged into bark at eye level; a triangle of twigs; a half broken bent twig hanging in a shrub;
  an opposing pair of gum leaves; a trail of sticks; windblown debris which is reversed; upturn a rock onto leaf litter
- Deliberately plant your boot print when you can. Step on top of older footprints, deliberately.
- Use fresh material so that your trail may be 'aged' according to the fading or drying e.g. of green leaves or mud.
- Position markers so that time, animals, wind & rain will disrupt them (at a recognisable rate). E.g. a teepee of sticks.
- Try a snow cairn; chalk marks on rocks; cotton tied around a branch; a stick dragged in the dirt; a scratch on a rock;
  balls of mud; charcoal or charcoal marks; tiny shreds of paper.
- Choose one method and stick with it, with a backup method in reserve.
- Double-mark a trail e.g. when doubling back, by duplicating it.

Tracking Someone Else

- Look for the **absence** of: spider webs across the path; dew on overhanging branches. To discern a foot-path look for the absence of fallen branches; leaf litter; lichen & moss; old wombat dung on prominent rocks; unbroken sticks; untrampled seedlings; undisturbed soil
- Look for the **presence** of freshly disturbed stones/soil crust/puddles, plus footprints.
- Examine puddles: for wet edges — indicating a recent forward swash; examine for muddy water; look for muddy water in the deepest layers — indicating disturbance some time ago; for wet stems above the water — they have been trodden in and have sprung back
- Overlapping fronds across a path — they 'remember' which way they were last pushed through.
- Don't step on the boot-prints or tracks you are trying to decipher—you often need to look at them again.

Botanical Tell-Tales

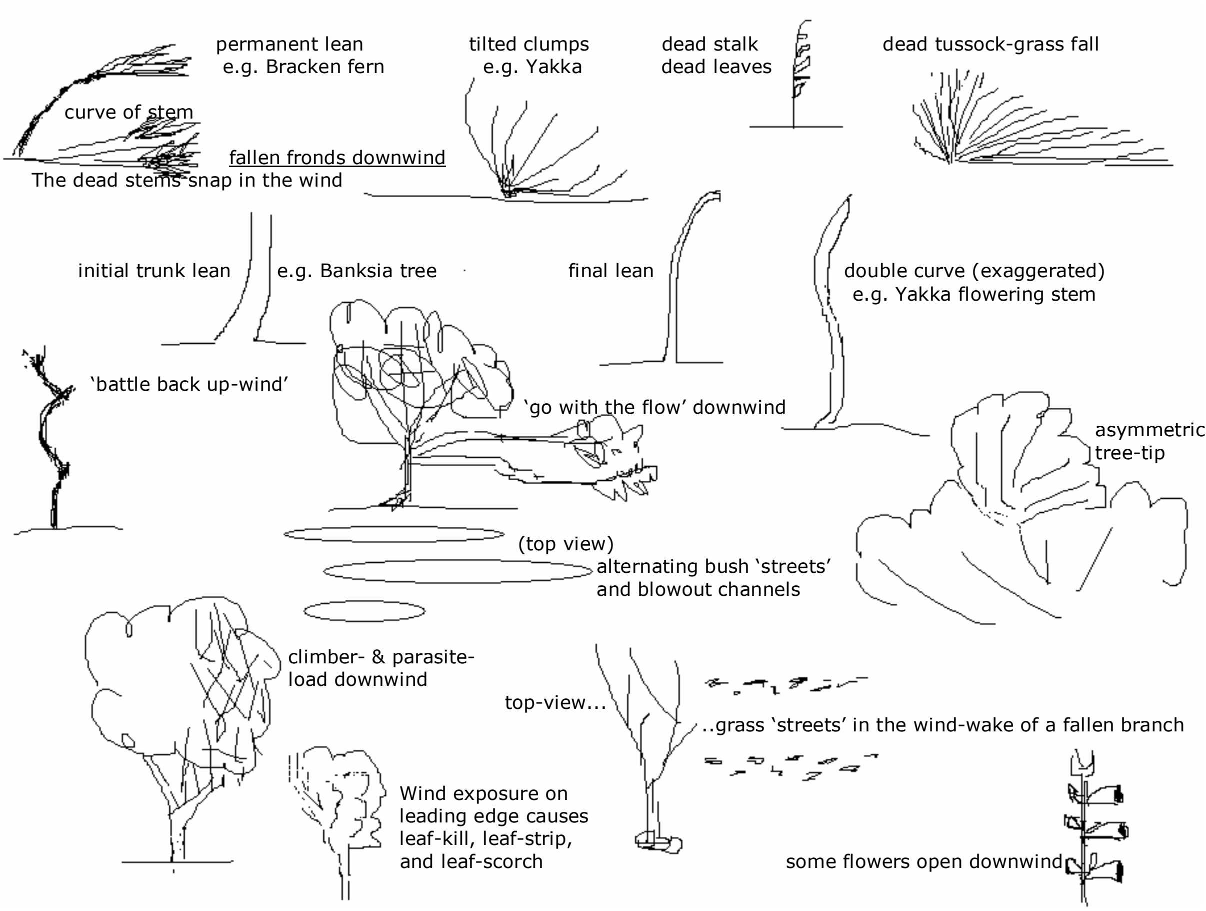
permanent lean
e.g. Bracken fern
curve of stem
fallen fronds downwind
The dead stems snap in the wind
tilted clumps
e.g. Yakka
dead stalk
dead leaves
dead tussock-grass fall
initial trunk lean
e.g. Banksia tree
final lean
double curve (exaggerated)
e.g. Yakka flowering stem
'battle back up-wind'
'go with the flow' downwind
asymmetrical
tree-tip
(top view)
alternating bush 'streets'
and blowout channels
climber- & parasite-
load downwind
top-view...
..grass 'streets' in the wind-wake of a fallen branch
Wind exposure on
leading edge causes
leaf-kill, leaf-strip,
and leaf-scorch
some flowers open downwind

Discuss It

Let others open your eyes: "What can you see about this?" "Where does it remind you of?"
Ask Questions: "How do you know where you are?" "Could you get back here again next year?"
Make a game of it: "What else can you notice from here?"
Express the specifics: e.g. "That boot-print has zed-shaped ripples on the right and ess-shaped ripples on the left.
The heel patten is a smaller version of the front half, as per normal."
e.g. "This tributary lies northwest/southeast and points to the left of the peak."

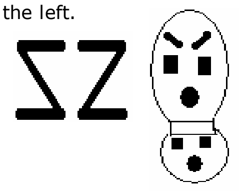

Many aspects of navigation yield to these simple tricks…

discuss it express it out loud make it the topic of conversation
delegate the tasks between yourselves
keep yourself alert
make up your own names
collect the details integrate them

Collect the Details

Identify a Specific Place by its Specific **Collection** of Details

E.g. "A low conical hill in flat terrain south of Mount Cameron"
That is not 'interesting'. The bits are not 'peculiar'. But the collection of the bits is unique.
E.g. Plant and insect species are often so similar to other family members that no one character is 'characteristic' enough.
The species will be identified by a set of characters.
E.g. Is this our old road?? Does anybody recognise it from last year?? but it has…

trend	e.g. toward sunrise
slope	e.g. mostly uphill
straightness	e.g. can't see for more than 20m mostly!
traffic intensity	e.g. many fallen trees across it, and encroaching bushes
length	e.g. it goes on and on and gets nowhere!
vegetation	e.g. set amongst old gums with yakkas under
ends	e.g. it came from the saddle and ends up behind the peak

If you think you recognise it, by the combined characters, you will probably be right.
E.g. Make one name for the big tree on the crest; another to name that crest; another to describe its relative position…
and soon you can't mistake it…"the big tree with the south facing hollow on the port-side Dam-wall Ridge downstream of Tree-fern Crk.
The add-on bits are part of the name of each bit i.e. "the Big-tree Crest" (downstream of Tree-fern Creek) (near the dam wall).

The smallest collection is two, but that is a powerful combination, e.g. a peculiar log, plus a recognisable hollow trunk, is almost unique.
The combination is the peculiarity to remember — a third peculiarity.
The geometry of the combination is a fourth and even more powerful memory aid… "This log points to that hollow".

Include What You Can't See

For instance, have you noticed that there is no bare ground showing? That you can't see farmland? That there are no intervening ridges?

Break of Slope/Make of Slope

The 'make' of slope is the **steepening** as you go downhill ie the "slope" is a down-slope. The best views are at the most abrupt 'make's

The 'break' of slope is the first **flattening** out as you descend. The thickest soils occur below the 'break'.

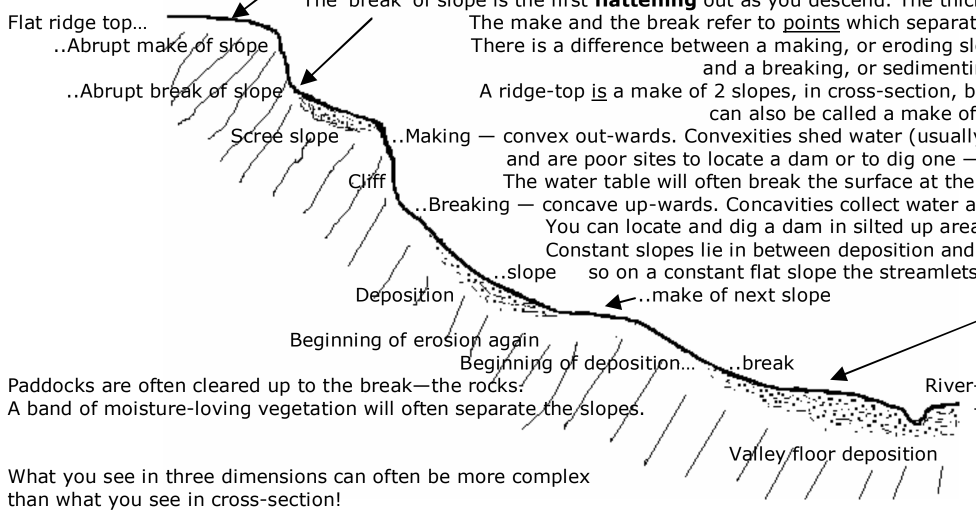

The make and the break refer to points which separate different 'slopes'.
There is a difference between a making, or eroding slope
and a breaking, or sedimenting slope…
A ridge-top is a make of 2 slopes, in cross-section, but any abrupt steepening
can also be called a make of slope.

..Making — convex out-wards. Convexities shed water (usually in all directions) and shed silt
and are poor sites to locate a dam or to dig one — there is bedrock below.
The water table will often break the surface at the more abrupt breaks.

..Breaking — concave up-wards. Concavities collect water and collect silt from above.
You can locate and dig a dam in silted up areas.
Constant slopes lie in between deposition and erosive scouring
so on a constant flat slope the streamlets can 'braid' together

Paddocks are often cleared up to the break—the rocks.
A band of moisture-loving vegetation will often separate the slopes.

River—erosion on the banks = convex
— 'break' at water level; concave
below it

What you see in three dimensions can often be more complex
than what you see in cross-section!
— When convexity changes to concavity there is a **momentary straightening** — a 'break of slope'.
This is also called a 'point of inflection' in the curve.
It's not just a meaningless mathematical quirk; it betrays a set of real changes, resulting from the erosion history.

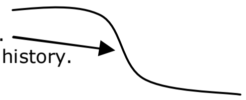

Scree Slopes — Tread Carefully!

The boulder-field at the base of a cliff, gives you a false sense of security!
Each rock seems so solidly jammed in place, but the whole field could 'slump' and slide all the way down, if given too much of a shake.
That is why there are no old trees growing in landslide-prone slopes!

The essence of a scree slope is, like a pile of gravel: It is in a continual state of critical instability, as evidenced by the straight-edged slope (flat but not level), about 36°. 36° doesn't sound like much, but up a sandhill, it puts you on all fours!
The rock surfaces look 'fresh'. Soil doesn't accumulate.
Tread carefully. Distribute your weight on all fours. Tread on bigger rocks.
Choose the cliff-base — it has hand-holds — or the scree-base — it has a solid bottom.

Mark Your Trail More Permanently and Ostentatiously

E.g. when you are lost or in trouble or in doubt

- A teepee
- Mud upon a tussock — quite unnatural; A tussock upon a tussock!
- Hang a piece of clothing from a prominent tree limb
- Dig a hole and make a hump

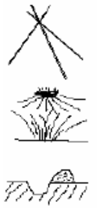

(Direction of travel is ⟶)

Mark Your Trail Invisibly

E.g. so that other walkers who are not lost, don't disrupt your markers, out of a sense of bush etiquette.

- "One mark every 50 paces". E.g. one pointed pebble. You may like to use double-paces.
  It is almost impossible to notice accidentally, but impossible to miss if you look.
- "Every 10 paces, as an option" added to your scheme — to cope with junctions and bends.
- Place a pebble on the ground, on the onward side of an existing cairn. You can recognise it, and scatter it, when you come back through the fog among a maze of similar cairned routes.
- Put a small mark at the base of each track marker (e.g. under blazes-on-a-trunk or ribbons-on-a-branch)
  — you can easily find it, and know that you have been along that track.
- "Half way between (2 markers)" or "10m North (of a juncture)", can remove the marks from obvious view or significance.

A pebble Code — Keep It Simple

All codes have to be memorable to all users — complications never survive the years of disuse.
I.e. use 'one' to 'four' only, preferably only 'two' and 'three'.

Tread the pebbles into the ground for clarity & permanency.
Upturn the pebbles for visibility, and **point the stones** appropriately.

Use the smaller pebbles at the front.

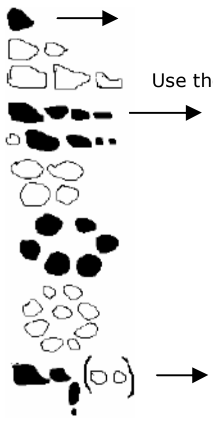

One: "I woz 'ere, going that way"
Two: "Me go that way"
Three: "I'm lost, exploring that way"
Four: "Look for a message, there"
Five: "Someone has seen this message"

Square: "Don't go this way"
Circle: "Oh" for "I'm OK"
Bull's-eye: "I will be back here soon"
Junction: "I have turned off (and returned, then continued)"

Double marking: Close the pebbles up the second time past, until they touch.

Combined Effects of Wind, Salt and Light

When assessing wedge-shaped shrubs, leaning trunks, tilted canopies, etc., try to distinguish wind-cut from salt-cut and from light-lean. But beware, in any natural system, it would be easy to over-generalise from a few known specifics, or falsely extend an inadequate interpretation.

Here is a typical wind-ramp. The leading edge consists of dead twigs.
The vegetation gets greener and thicker and taller downwind.

But you cannot easily tell whether it is the result of salt-laden wind, or strong wind only.

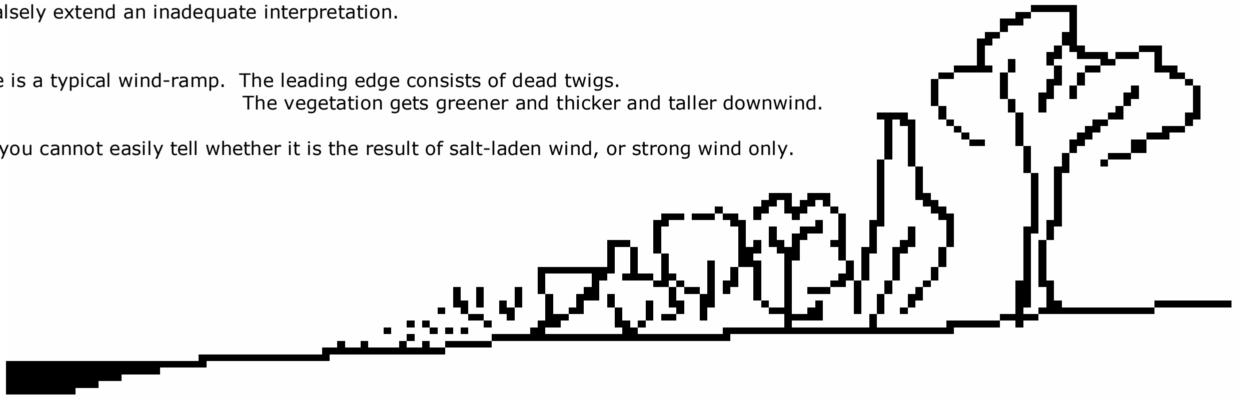

Salt-cut will show you the direction of the prevailing onshore wind only.
Inshore, wind-ramping will show you the direction of the prevailing wind, which may be significantly different in direction.

Botanical Asymmetry

Plants must respond to the wind.
If they did not deliberately push up-wind, we would be walking through a landscape of leaning-over plants.
Therefore it must be possible to 'read' the effect of strong wind written into the branching history of a plant.

The branches of a bush may grow according to the strong wind, but the leaves will still respond according to the asymmetrical light.
The two effects may or may not be independent and you will need to examine them carefully to see the difference.
Look for asymmetry in the green parts of a plant to show up bush-noon and look for asymmetry in the other parts for bush-West.
If there is a macro response to the wind, which will confuse the macro light-response, look for micro signs of a light response, on individual branchlets.

You don't "need to know your plants" beforehand. Instead you examine the species which present themselves.
You assess what responses they make to light, wind, shelter, etc.
E.g. Do the stems colour up in sunlight? If so keep looking at that.
Do the branches throw out more leaves on one side? If not look for something else.

Extend Your Familiar Range Systematically

Explore in every direction, in ever longer legs, radially. Leave recognisable marks alongside recognisable things.
Then explore in concentric circles to make sure that you can recognise individual trees and rocks.
You know you can explore further within the endless sea of scrub, if you have 'familiar friends' which you have seen repeatedly.

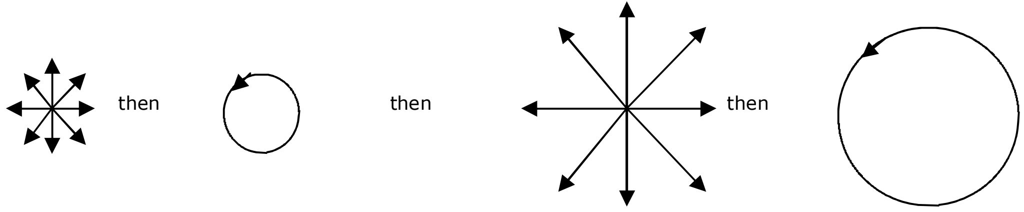

Markers for future recognition e.g. on your own property…

Plant a tree   Graft two branches together   Suspend a forked branch into a forked trunk
Arrange logs e.g. a parallel heap   Arrange rocks e.g. a cairn, or a rock on a stump, or a rock 10 m south of the biggest tree
Dig a trench   Etch a mark   Paint a rock   Leave something solid behind
A stake   A signpost   A numbered metal tag   A path   A Campfire-place or Barbecue

Hints

It's the Campsites You Don't Use, that are Easiest to Forget

> I'm forever asking: "Can anyone remember" whether we can camp between here and there?
> Take notes of potential campsites. It makes your trip-planning more flexible, next time.

In general, take the opportunity to spy out and note down potential future trips — side-tracks, ridges, climbs…

Leap-Frog Walk in New Territory

I like to walk a new track twice, anyway. It is no big difficulty to take your time, once, and achieve better familiarity, by leap-frogging.
Explore 2 legs forward then one leg back, and you cover the ground 3 times on the way out, once more on the way back.
4 legs forward, 2 legs back, and it's like walking a return track 3 times, but it gives you a much more solid handle on each section.
You gain what you want, appreciation and re-preciation, and what you need, familiarity and sequence memory.

The best way is to combine this with a mnemonic for each leg. See AA to ZZ for an excellent method.
Its also a good time to mark each new leg discretely.

Break of Slope — And Route-Finding

- The first signs of an approaching ridge-top when climbing include: the change from deposition (breaking) to erosion (making); i.e. onto outcrops; into exposure to the wind.
- Take a rest at the break of slope — before climbing upwards.
- Get your bearings at the most abrupt make of slope, where the view is clearest.
- Look up- and down-hill for the lowest skylight beneath the trees, to reveal the change of slope.
- Avoid contouring around the break of slope — it is tempting, but — it is steep and unstable.
- In rocky gorges, or on any creek-bank, or around any permanent water, look for a break of slope just beneath water-level.
  It may be safer, than rock-climbing higher up, to get your feet wet, and use such ledges. The slope near the water may be less steep, at least on average on jagged slopes; you can't fall as far; your landing is softer (wet) — if you 'keep as low as possible'.
- On steep gravelly roads, stick to the 'crown', or the 'gutter', to avoid slipping.
  The crown will snake from side to side as the road turns.
- Slopes are awkward to walk on—steeper than the ridge or creek, so,
  when you drift up onto the sides of the creek you want to follow, you will be forced further up — onto the ridge.
  When you drift off the crest of the ridge you want to follow, you will be forced further down — into the creek bed.
  In this way 'minor' detours (e.g. around an outcrop, on the ridge or in the creek bed) may result in major redirection orders!

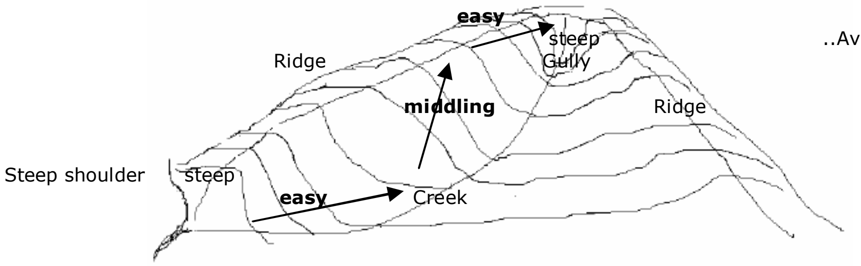

..Avoid the steep bits.

Keeping Track of the **Bends**

Don't just follow your nose — follow the bends!
There have already been some hints on keeping direction sense:"Walk across Australia!" & "Integrate the Jetstream and Earth rotation"
Plus the "Sky to Land" and "Land to Land" navigation systems (p.29e,34d,38d,42d).
Following a track can be mindless. But following the bends in the track means that you understand
— that you have changed direction, and by how much.
We have an ingrained bad habit — that we walk 'away from' the starting point, which is therefore 'behind' us.
Only when we 'turn around' is it 'ahead' of us on the track.
We usually only notice junctions, features & hunger, not distance or direction or exact time.
The commonest mistakes of guesstimation, are failing to notice a track curve,
and **underestimating** exactly how far a gentle but insistent bend will **turn** you around i.e. simple disorientation.

Hints: Notice the bends, with your eye on the sky or the skyline. Establish the habit of looking back.

Desired Equipment

- Compass — look at it often and interpret it
- Polaroid sunglasses — look for the dark blue band
- Radio — use it to keep orientation if you don't have a compass
- Map — orient the map to the landscape, and it acts as a rough compass
  — lay a rhumb-line through the bends, as an average course, as a reference 'track'…

Logging Constant Diversions

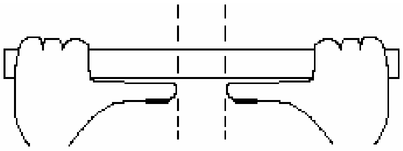

A stick or string helps you measure direction changes…
Held at arm's length, you add or subtract, as you decide to change course left and right.
Thus you can revert to an intended course after a short 'blind' section,
and average out the course changes even without a compass, or being able to see the Sun.
You could use the skyline as a reference direction, or lineup two tree trunks for each change.

A rougher method is to counterbalance your diversions.
If you go left-around the rock; next time, go right-around the tree.

Keeping track of countless bends is one of the few major problems in navigation
— like keeping a log of distance covered, having no location sense, and trying to stay alert.
So aim for an independent cross-check e.g. a landscape fix or a distance log.

Think like a Plant

E.g. Imagine you are a small ground-cover plant. In a desperate attempt to not be eaten or trampled to death you will probably branch and leaf out in all directions to simply survive... So don't expect much botanical asymmetry underfoot.

The idea is to become an amateur ecologist. For instance, you could work out the pollinators available, the dispersal mechanisms used, the biogeographical zones and overlaps, and the associations between species.
Therefore think about the geology, geoform, latitude, altitude and soils.

Temperature & Relative Humidity explain a lot of botany.
E.g. very cold morning winds are also very drying when they warm up later in the day.

Altitude, for instance, affects the species mix, so that with practice you can gauge your elevation above sea level.
You notice when sub-alpine species kick in e.g. Telopea (Waratah), then when true alpine species begin e.g. Richea (Scoparia).
Individual members of genera have altitude preferences and so are helpful to distinguish altitude
e.g. various Coffee berries (Coprosma) or Pink Berry (Cyathodes).
N.B. Since the prime effect of altitude is to drop the temperature, this is really a latitude-dependent temperature effect.
I.e. Going further polewards, the altitude preferences will change.

'**Aspect**' (i.e. South sloping or North facing) will be important in interpreting the presence of particular species.

Living stresses explain simple botanical systems.
E.g. In the desert, late afternoon is the baking-hot-sunlight side (cf noon in forest). Trees will be shiny leafed — to shed excess light (cf searching for more).
So consider the prevailing winds, salt spray, max. and min. temperatures, wet season, disturbance, trampling, fire, clearing, introduced competition or diseases, grazing, salinity, mineralisation.
Pioneer Species — These are the easiest to identify, because they are lonely.
Their preference is the saltiest, hottest, driest, highest, coldest, poorest, windiest, sand-blown, most flood-prone, …
If life becomes too easy, other species will smother them out. Hence, look for those that don't mind the hot baking Sun, etc.

There is a height-**succession** of plantforms grading from **lowest to tallest and deepest rooted**.
Ground lichen; moss; grass; herbs; creepers; shrubs; trees; ' emergents '
The depth of soil is crucial. The taller ones shade out the humbler ones, and their deeper roots seem to starve out the shallow-rooted ones from below. E.g. mosses need no soil and not much light, but will quickly be overgrown if these are provided.
Other factors keeping plants humble are trampling & grazing, fire frequency and dry season. Grasses survive all these stresses well.

Competition between species explains the more diverse vegetation.
E.g. Lichens overrun each other, and leave the rock surface, to become 'leafy' or fibrous or pendulous.
Lush growth betrays the most comfortable living conditions. E.g. rainforest; the underside of a rock in the desert; sheltered microclimates — crowded with species. This is opposite to pioneer species.
E.g. Where excess sunlight is the 'living stress', lush growth will show up on the shady side.

Draw the Skyline, the Ridge-lines, the Outlines
Label it with the 8 main directions

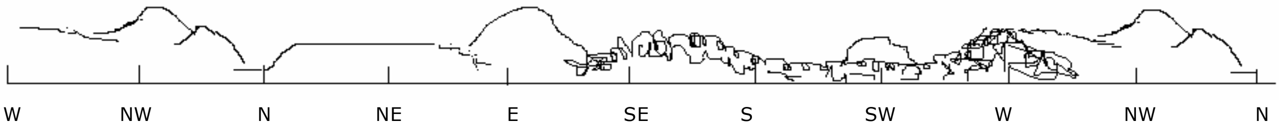

Talk the panorama into words. "Nor-Norwest Hillock, in front of a larger hill on the left, runs down into a gully behind NE– Plateau.
East Mountain is behind NE Plateau and drops away behind the vegetation south of us.
Distant Peak pokes up behind in the SW, and the closer skyline reappears in the West,
falling steadily to the NW Dip before the larger hill."

That took one minute. You do have time to look, notice, talk it through, and even draw it.
Marine navigators know that they have to.
Don't just imagine doing it, or do it silently, or do it half-heartedly — actually do it.
Unlike an avigator who uses a layout map to recognise roads, rivers, ranges, coastlines, etc, from the air,
a walker needs a nose-eye view…
MAPS DON'T DRAW YOU A NOSE-EYE VIEW — SO YOU HAVE TO!

Examine the Skyline BEFORE You Walk up to It

That might mean looking behind you, six hours ago, or to left or right, right now,
because the skyline ahead of you rapidly disappears behind foothills, just when you want a clear view of it!

A clear view of the skyline is uncommon, so you need to take advantage of any skyline view while you have it
— and that is often anywhere else except where you want to look — ahead.

Circuit walks and return trips mean that, if you haven't habitually looked at the surroundings abreast of you & behind you,
you won't be able to recognise what you are walking up to later.

One way to recognise the skyline later, is to use it
e.g. use it to keep track of bends; to keep track of progress; to disclose hidden valleys (by parallax or haze).

Another way is to name what you use e.g. "the hill behind the dam wall from camp 2—Dam Hill"

Mnemonic: "Down in the Mouth"

"Downstream" is "Down to the Coast". "Which coast? Which river mouth?"
Until you have answered that, you are disoriented.

You pass a creek-bed or a flowing stream: "Which way is it flowing?" is only half a question. "Where is it going?" is the other half.
You turn and head 'downhill': You think 'Down' — then say: 'Down in the Mouth — Which mouth?'

If you also ask "Upstream? To which ridge?" there are many possible answers because upstream is divergent.
Nonetheless it is good to know whether a river reaches to the very edge of the valley — whether it is a 'major' tributary.

Which way is the flow?

In dry, flat creek-beds, examine the tilt of the rocks… If any rock ever tilted up into the flow, it would be flipped over.
Examine the silt build up — on the sheltered downstream side of any obstacle.
Look for flood debris — matted against the upstream side of branches.
Consult your artificial horizon, if you have one.

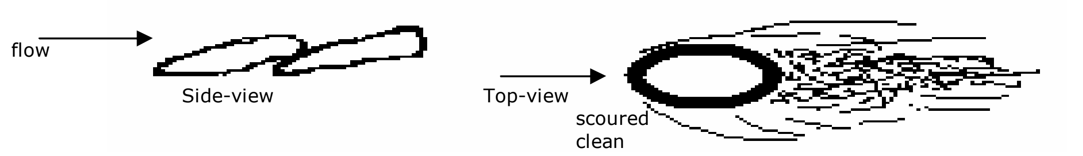

(This is much like grass tussocks in the wind.)

Hint

Don't think "this side" & "the other side". It only makes sense for the one main lowest riverbed.
Think: Up-catchment/down-catchment/cross-catchment, overall, & Port half/Starboard half of the whole catchment in question.
(as seen from the coast, looking upstream)
Think plural: "the other sides"/ "another side" in the network…

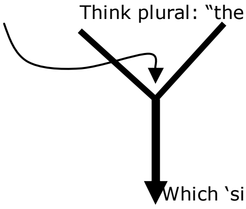

Which 'side' of the river are you now? "I have crossed to another part of the drainage system"
Another possible answer: Examine the map for the contour heights of each channel for the lowest one.
"We are now in the starboard half of the catchment"

The Difference between 'Direction' and 'Track'

'North' is a general orientation reference **direction**.	But	'North from here to there' is a specific route — a 'bearing', 'line' or **'track'**.
'Stick to a direction'/a 'heading' means e.g.'toward sunset'		'This road points toward that Gap' will never change — it is a fixed 'track'.
It can start from any point,	cf.	'Stay 5 miles off the lighthouse' is a (curved) line of position — a track.
and sidesteps around obstacles, without bothering to return to track		'From this stump, past that tree, to the rock' — sticks to a fixed straight line, despite detours.

So the final practical 'interpretation' of, say 270°T.,
is not 'West', not 'sunset', nor 'Pot-set', not 'THERE!', not a drawing in your notebook, Not 'How hot it must be in Baghdad!'
but is a navigational track... "From this place to… that particular tall tree" "Then we will get a line on another point directly West"

I.e. Interpret a 'compass course' , finally, into accurate, achievable 'legs'—'on track'.

Continuous bends

Riverside tracks are especially sinuous, yet slowly so, so that they are deceptive.
Beaches near headlands, and lake edges, are especially insistent in their constant bending.
Contours, constant horizontal angles, and constant distance-off, are all continuously curved. Aqueducts follow contours.
Consequently we are necessarily **continually disoriented**, in following them.
No sooner do we get our bearings straight, than we lose them.
Logically then, you need to orient, not by a direction e.g. the Sun, but if at all possible, to **orient by a 'track'**
i.e. a particular path on the ground — it doesn't have to coincide with your own track, but it often will
e.g. from headland to headland. (The beach will then lie wholly on one side of this shortest distance 'track'.)
A visible, noticeable, landscape track is best e.g. the biggest river valley around.
You now can feel continuously oriented to something fixed, something which you can see & 'feel in your bones'.
As a second best choice, choose a fixed compass course e.g. a rhumb line, and watch where it intersects the landscape.

Follow the bends by anticipating the trends

Before you drive off, or plunge into the twists and turns of endless undulations, examine the map for your average course
— to give you some overall reference track ahead of you. It may be mental, or visual, or by compass.
You can see when you turn to the left or right of it, direction-wise. You can feel it when you drift too far to one side, track-wise.
Interpret each leg meaningfully. E.g. "toward the morning Sun"
(the time of day doesn't matter).
E.g. "Keep Mt Wellington on my port quarter".
E.g. "From the pine trees to the river mouth".
Divide the journey into recognisable sections, if it helps…
Make this a first priority, before studying any details.
Preferably, mark them lightly on the map.

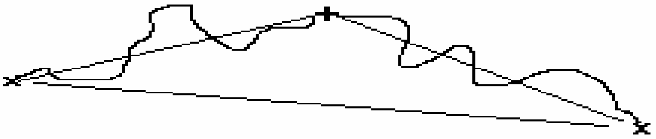

“Mosses Mop up Moisture”

◦ “Mosses grow best on the shaded side of a tree” or a rock, ◦ where there is not enough light for competitors
◦ where there is little soil to hold the moisture, ◦ and it doesn’t dry out ◦ and often there is too much moisture for other plants.
Mosses actually monopolise the moisture
In hot places it is the late afternoon which is driest, so it is **the sunrise side, in a sunny climate**, which is the moister side.

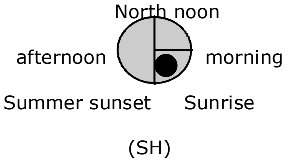

Note: Dead moss is black and brittle.

Examine fallen logs — the shady side has a sheltered dark crevice...

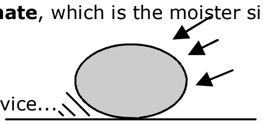

In a hot climate, it may be more accurate to say,
“Mosses grow best on the sheltered side”

Examine ditches — One side might be scorched, the other, lush.
Examine bumps in the ground — One side is greener, with more moss and taller grass.
Examine underneath shrubs — for the moister, mossier, greener side under the bush.
Examine intertidal rocks — for moss-like seaweed growths on the shady, wetter side.

“Lichens Like Light”

Lichens don’t mind being dried out, so sometimes they are opposite to the mosses. However, not all lichens are so robust.
Lichens are fungi which live off algae entrapped and protected within (“Fungi Feed on Food”). They may live on mineralised substrates.
They are therefore very colourful. So the best way to learn lichen-likes is to open your eyes to the colours of each direction…

You may see white blotch lichens looking downSun but not upSun.
Yellow blotches may all be only looking uphill, not downhill. Leafy lichens like the shady enclaves, not the Sun-drenched flats.
Crusty greys and fibrous green might cohabit the underside of branches. “White paint” may only grow an the top of any rocks.
Orange granite might be just above high tide level, etc.
Diffuse lichens colonise the crumbly side of a sunbaked rock. Flat-topped rocks are trampled and won’t have thick lichens.
Look closely at surface colours of rocks and barks and soil-crusts— they may be due to diffuse lichens.
Look on the branches of shrubby growth — the lichens will grow best on one side or the other.

The idea is to start looking locally, before you need them. It’s a nice hobby. And hobbies keep you awake and observant.

“Fern Fronds Face the Front” i.e. the sunny side

Looking down-Sun, you’ll see an array of green topside fronds flat on.
Looking up-Sun, you’ll see more undersides & yellowness & edge-on fronds. It may not always show, but is clear when it does.
Tree ferns retain their lower branches longer on the sunlit side, and so seem to tilt toward the light.
(But in a windy place you will notice that the permanent lean, curve & frond-fall of ferns are downwind.)

Make a Mental Movie

Memorise a spectacular cliff looming closer overhead.
Watch the shape change. See the foreground shift. Notice the colours and contrasts. Look at the landscape flow.
Describe the silhouette. Peer into the deep shadows. Scan the skyline, looking at each interruption or change in trend.
Appreciate the hazy distance.
Follow a complete outline around, naming its shape.

Don’t forget to “record” the sound-track!

Initially try a full production only for beautiful memorable scenes.
With practice you may find that your memories stick for more and more locations.

Take Mental Photographs

Familiarity and Recognition are primarily visual. Your visual impression is adequately vivid, but your recall of it is faulty.
So put handles on the memory and you will be able to revisualise e.g. “where we found the weather balloon”.
To give your memory something to recall, take a detailed look.
The tricks are simply to give it a go, and to notice **the specifics**:

The colours, shapes, skyline, peculiarities, sequences — Talk about them.
Play a game: “Your turn: what else can you add to the picture”
Overlay some geometry on what you see — e.g. the log is parallel to the path and as far from it as the log itself is long.

Now close your eyes and see what you can reconstruct. Open them again and notice where you were wrong or were lacking detail.
Repeat this a few times until you could picture it in bed the next night.

Mental photographs are, after all, **your take-away memories**.
They are what names name; what “directions” talk about, what ‘recognition’ and ‘familiarity’ is based upon.

Take a Mental Photo of Each Change of Track

Where one subsection ends and another starts
— Look forward onto it, with the old terrain in the foreground and the new terrain joining on to it.
Now walk past the join, Look back, take in another view, with the old terrain now in the background.

This adds visual sequencing information to individual memories.

The Inside of a River Bend is Flatter Ground

The inside of a meander may be wider, sandier, marshier or of different vegetation.
On the outside of a meander, cliffs form, through ongoing undercutting. See also p.56a about the steeper North bank of a river (SH).
Consequently plan ahead and work out when to swap sides to find the easiest walking.

If the track becomes steeper underfoot, that would mean that you could be on the outside of a bend.
Check to see whether your heading has changed with the bending river.
If the track becomes flatter underfoot, this may indicate an inside bend (or a river junction)

Rivers Bend! Don't Forget to Look Backwards

Obstacles cause creeks to change direction. Outcrops of bedrock and creek junctions, for instance, cause bends, and this may be shown up ahead of time, by ridgelines pushing in closer, as you follow a riverbed.
Expect to meet a new creek clinging to the base of any ridge or shoulder which crosses your path, and also expect to change trend.

In addition, to that cause of bending, rivers follow contours and so usually are on a curved course anyway.
Furthermore, in the flats, rivers never follow a straight course but will 'meander'.

So if, while following a creek valley, you notice the slope underfoot changing, or the vegetation altering, watch out that the creek doesn't leave you, by bending away from you!

If you find yourself in the middle of a larger than normal river flat, suspect a river junction, and be very careful interpreting the trends.
E.g. you may meet 'the river' 'coming back toward you' when you were trying to go downstream!

So you can easily lose orientation while the winding river-track follows the lie of the land and gradually wends its way around corners!

Look ahead to catch the trend — and interpret it.
Look back, to see the extent of the curve.

Rivers Meander. Ridges Run Straighter

A ridgetop route will often prove shorter than following a winding river.

Keeping a Compass Course, by Legs

You can't see far ahead, but you can line up a rock or a bush which is in the right compass direction, then walk up to it.
It doesn't matter how you get there, as long as you get there.
Repeat the procedure, even if you can see far ahead — because, if you are merely following a compass needle or a distant landmark, you can "side-slip" off track.
Your eyes are freed from the compass needle while you walk.

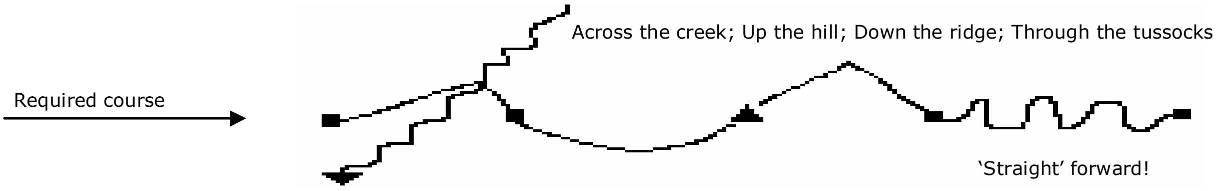
Across the creek; Up the hill; Down the ridge; Through the tussocks
Required course
'Straight' forward!

Turn Smooth Bends into Sharp Jumps

Your brain wants something to "latch on to". A smooth bend is just the opposite! So make a series of straight-line "legs".
You "let go" of the last leg, and make a definite jump, to "latch on to" the next leg, and your subconscious smiles, and remembers.

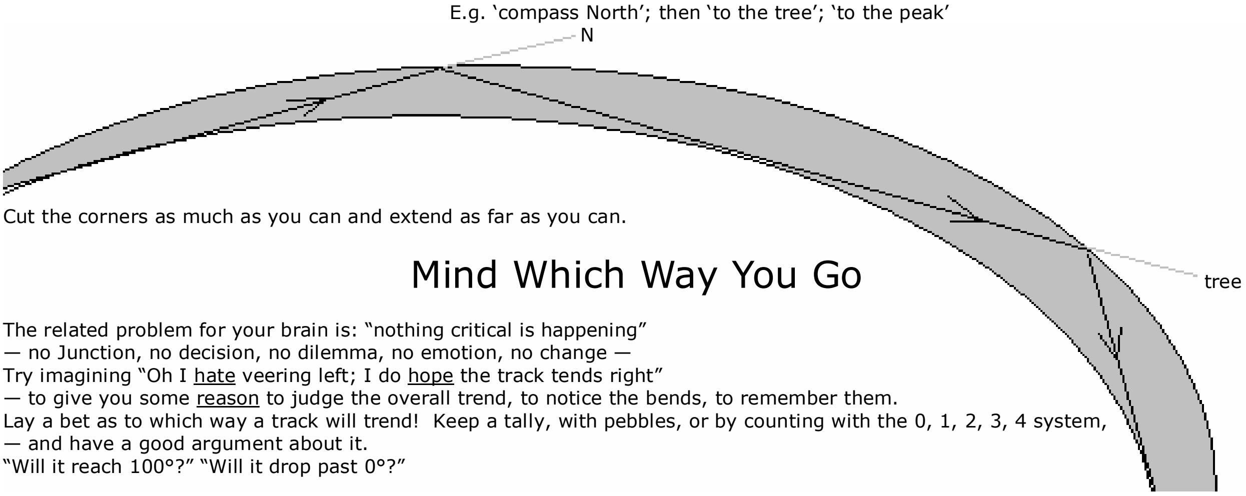

Cut the corners as much as you can and extend as far as you can.

Mind Which Way You Go

The related problem for your brain is: "nothing critical is happening"
— no Junction, no decision, no dilemma, no emotion, no change —
Try imagining "Oh I hate veering left; I do hope the track tends right"
— to give you some reason to judge the overall trend, to notice the bends, to remember them.
Lay a bet as to which way a track will trend! Keep a tally, with pebbles, or by counting with the 0, 1, 2, 3, 4 system,
— and have a good argument about it.
"Will it reach 100°?" "Will it drop past 0°?"

Bush Noon— The Sun is Always out, If You Know How to Look

Trees & shrubs live in the light, day in, day out. You would expect that they should respond to the light, and lean or reach out to the strongest light, and throw out more branches and more leaves to the bright side.
The disposition of branches and leaves stays in view even when the Sun isn't out, and it lasts longer than a cloudy spell.
Although the Sun is a huge factor in explaining foliage asymmetry, be aware of other factors…

- A lake or sea can double the light intensity on one side, pulling the foliage toward it.
- A clearing will let more light in on one side.
- Down slopes allow the trees to see more skylight on the down-side, and less skylight on the up-side.
- Other trees can shield a tree from the Sun. Look especially for recently **fallen** ones on the ground, to explain standing branches
- The tree itself shields its own branches, so that most growth is simply outwards.
- Unrelated factors like gravity, wind, and wood collecting, alter the shape of a tree.

However, if you **average out** all the other influences, you should be left only with the noon-side Sun's influence — "Botanical Noon".
With adequate practice you should never again be "lost in the bush" because it is the thick bush itself which will tell you where North is.
You may have an initial hopeless feeling of "I can't see any meaningful asymmetry!" but the feeling will be temporary (eventually!).

Of course, this works best only where the Sun is always one-sided—in nontropical latitudes.
Hint: If a tree confuses you by pointing in the wrong direction, go over to it to learn why.

False Asymmetry

The slope of the ground
may be forcing the asymmetry,
rather than the differential in the lighting.
In the second diagram, there is asymmetry:
A is ungrassed; B is grassed.

False Symmetry

C and D are both grassed.

True Symmetry

B and C are both grassed.

True Asymmetry

A and D are differentiated from each other.

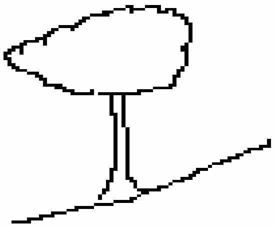

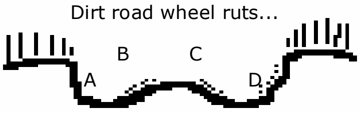

False ("Chalk and Cheese") comparisons abound in Nature. True asymmetry is there too
Your task is to sort it all out.

E.g. Shaded foliage sees mostly blue skylight, which should make it long and lanky, compared to the red-end light from the Sun

E.g. the West side shadow is frostier for longer and the East side shade is cooler in the p.m.

E.g. the North side soil will be warmer deeper than the South side of a thick hedge

Look on the ground beneath a shrub, or log, for asymmetrical shade and shadow effects.

Remembering the Sequence, Needs More Than Logic

Logic is logical and helpful e.g. The bridge was after lunch because we were worried about the crossing over lunch.
The red rock was after that, because the second creek was dry — it had all those beautiful red lichens in it.
Then the funny tree was where we had chocolate, i.e. 3 o'clock stop.
But logic by itself is not 'sexy' enough to be emotive and memorable.
So deliberately add some emotion…
E.g. "lunch", "worried", "beautiful", "red", "funny", "chocolate"
You may need to add a song and dance, a gimmick, a clue to the time or sequence, distance or linkage.
The trouble with a memory gimmick, is that you focus on the gimmick, not on the memory itself.
E.g. If you use an acrostic, remember to feel the sequence of it, as well.

DO Something NUMERICAL

E.g. Lean one stick against "Walking-Stick Cliff" at one o'clock.
Sing two songs at "Sing-Song Pond" at two o'clock.
Make three cairns at three o'clock at the "T(hr)ee Junction"
Make four arrows on "Four Arrow Slope" during four o'clock.
Do it. Don't just imagine yourself doing it.
E.g. Scratch numbers in the ground, in sequence, as you pass points of interest.
e.g. "5FF" (the 5th stop, with the flat floor); "6SS" (six sizzling sausages—the fireplace).

Make a SHAPE; Match the SHAPE

Every or location in a forest track or in open country can be recognised from its **peculiar shapes**. Every bend or fork can be **matched to** the forest, or the shrubs and stones, so both the bend or clearing and the surroundings can be remembered, with joy.
Locations: What does this fallen log **point to**? What is **parallel to** it. *Place* something so it *does* **make a shape**.
How many shrubs do these three trunks enclose? What **face** or **picture** can you make out of the bushes, rock, clearing and the creek?
Do you see that **line** of tussocks; that **clump** of trees half-way up the slope; those parallel gutters; that oval of sand-dunes? Where does it drain? What **direction** between those **paired** stags?
Bends: ~Look for a tree trunk **leaning** over at the same angle as the bend of the track, preferably leaning over the track or bend and in the right direction. The tree stands out from all the others because of that. (Any lean at any angle in any direction will suffice.)
~Look for a branch **pointing** in the new direction, or curling, or **bending** in a way which mimics the track-bend.
~Look for a log **lying parallel** to the new direction, or any ditch, or line of rocks, or outcrop. (It doesn't have to match.)
~Look for a large rock nearby which **fits the shape** of the bend, or orient it until it does. (Matching shapes are always the best)
~Look especially at the tree trunk, or the shrub which blocks the straight line of sight ahead, or the shape of the fallen-over trunk in the way. E.g. "I'll fall into that black hole in the trunk if I keep going straight ahead! I'll have to swerve". "I need to go around that round stump". A tree **right at the bend** will be visible from both directions.
~Look for a forked branch which **matches the angle of a fork** in the track. **A series of parallel branches** at the fenceline. A couple of trunks **leaning towards each other** over the gateway. A tri-parallel set for a creek with two banks.
~Look for a branch which **points downstream**. Watch it for the next 50 years as it grows. A tree recognised is a friend for life.
~At a noticeable uphill pitch, find a leaning tree trunk and imagine climbing up it.
~Some leaning trunks can be interpreted in **reverse**… from top to bottom, to indicate the new direction.
The more you latch on to a peculiar shaped feature and the more you study its peculiarities, the better your recognition will stick.
Try to **double up** on the first peculiarity to clinch the recognisability. A **matching pair** of branches. A second confirmatory log.
You will have to **look backwards** at each bend if you want it to work well for the return trip.
Every bend and feature is a good chance to recognise another couple of trees. You stop at the bend and look around for your "friends".

Land-Form

Work downwards from a 'Big Picture' understanding. Start with a small-scale geology map, showing overall elevations.
Continental-scale domes (anticlines) & depressions (synclines) explain the overall drainage and watersheds.
Fault lines — shown by escarpments & cliffs — and individual ranges, interrupt the overall scheme.
The geological layers may be visible connecting across large distances.
E.g.

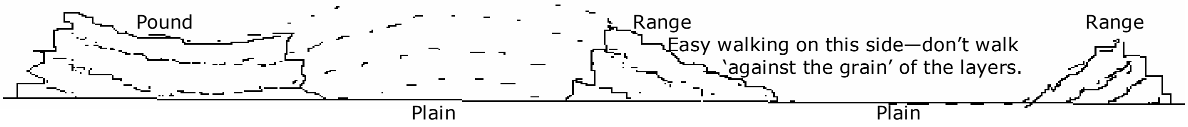
Pound
Range
Easy walking on this side—don't walk 'against the grain' of the layers.
Range
Plain
Plain

Layers become hardened and resistant to erosion when they are bent concavely upwards. They are compressed rather than fissured.
Convex layers are fractured and easily eroded, because they have been stretched. They will have been eroded away.
The present landform is a result of the erosion history on the previous landforms.

Hint

Most peaks have a steep side and a gentle side. Don't forget to make a mental note of that trend in relation to the rest of the land and the other visible peaks
— you may notice a pattern, or simply keep your overall bearings better, or recognise the peak later, from afar.

What Makes a Cliff and an Overhang?

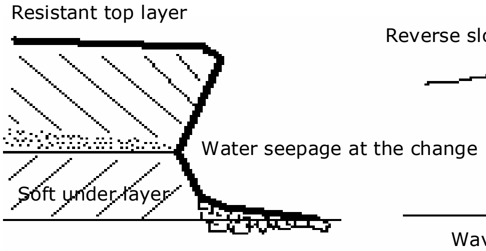

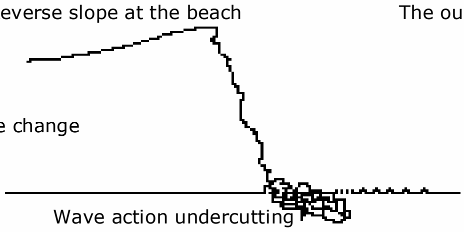

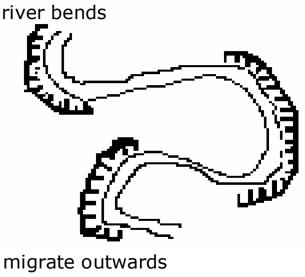

Waterfalls form at hard layers, at fault lines and at escarpments.
Look for outcrops striking across the land.

"End the Bend" i.e. Notice the Un-Bends

A bend is not simply a sharp corner, not just a change in direction: 'end of story!'.
It has a start, a curve, and **a finish, where it straightens out again**.
The curve needs to be **named** as a whole, **measured** as a whole, and thought of as **a whole bend**.
Look for the dividing points between successive curves

The arrows show where a curve to the right changes into one to the left, (a Z bend), & vice-versa (an S bend). These are "points of inflection"

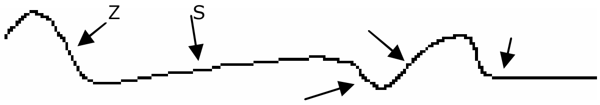

— a straightening out, and remaining straight, is the other possibility.
Incorporate half of the straight section into each adjoining 'bend'.
At the end of each bend, reorient yourself — look behind, at the compass, at the Sun and your shadow. Point to 'home'.

Name the Bend, By the Trend, of its End

Trend:	Left, toward Mt Michael	e.g. "Viewmont Bend"
	Right away from camp	e.g. "The Long Right Leg"
	Left into Lake Jordan	e.g. "Left it down the Drain"
Handedness:	Hard left around the spur	e.g. "The West Bank Left Bank"
	Half left from the creekbed	e.g. "The Left Contour"
	"Clockwise Climb"	
	"Right around the Bend"	Bends will often alternate.
Direction:	"Ending with Sunset in our face — Sunset Slap"	
	"Turning until the Full Moon Rises on our Right — Werewolf Way"	
Compass:	"North Now, to Alaska"	
	"Left-Sou-West"	

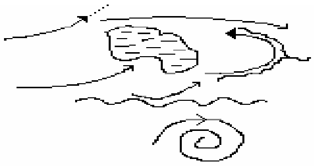

Add some description
to seal the uniqueness e.g... Skyline: "Lone Pine Ridge" "Three Tree Hill"
Detail: "Uphill; On the Left Bank; Just past the river crossing; at 10 a.m."

Notice and Name Perverse Bends

Unexplained bends, or ones which buck the trend, deserve a name.
Most bends are predictable — around obstacles — at creeks — around shoulders & outcrops — shortcutting
— contouring around sloping ground — at Y-junctions — joining with disused overgrown tracks
E.g. Seaside roads will hug the coast, go diagonally downhill but upstream (into a valley) and climb diagonally out of it, downstream — as a compromise between shortness and flatness.
E.g. For the same reason, tracks will often change direction at a creek crossing.

So name the curves that go against the flow…
E.g. "Sinister Switch" (Latin for left) "A Dextrous Diversion" (Latin for right) "Unorthodox Twist" (Greek: not "right" = left)
"A Wandering a Way" "Aimless Avenue" "It shouldn't've gone left!!" "The I'm-going-on-strike Uphill Bend"
"Mind Bender" "Morally Bent"

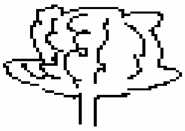

Higher Density of Leafiness on the Noon Side

From underneath, next to the trunk, you can look up in all directions and see which side is leafier.
Looking side-on you may see one solidly green side and the opposite side patchier in its cover.
Trying to look through the foliage to the back of a shrub, you will have most difficulty on the North.

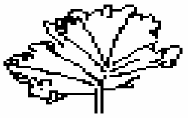

The Lowest Leafy Branches are on the Noon-Side

Looking from the sunny side you see a lower branch or even a wall of greenery, right to the ground.
Low branches on the shady side are often short, dead or only thinly covered with foliage.
Looking toward noon, you see more trunks, dead wood & grey-brown sticks.

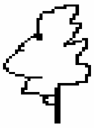

"It's the Shady Side that Shrubbery Shows up"

In practice you don't look for the noon side, but for the absence of leaves low on the pole side.
Looking toward noon, you see ill-developed shapes. Many plants will throw out branches and leaves in profusion in all directions, but you seek out the one direction which is different — the one less densely green — where the light is in short supply. It is **the 'light null' for solar direction finding**.
Look for where no vegetation leans: "No stems, plants, branches point in that direction".

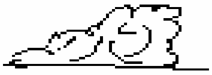

Tall, Steep and Ragged Backside

Shade-side branches often reach for the sky, and are more vertical on that side, making it steep.
The noon side will be more of a smooth continuous wall of greenery from low-down upwards.

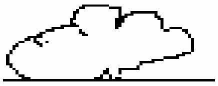

Wedge-Shaped

Looking away from noon, you see smooth and continuous foliage, slanted down toward you, often clinging right to the ground, like a wedge.
Looking toward noon you see flat tall backsides of the bushes.

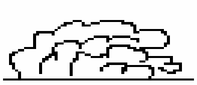

Hollow under the Shady Side

You can stand closer to the trunk, or at least see more of the trunk more clearly from one side.
Down low on the shady side is where the light matters most. There may be space enough for a branch but not light enough. On the sunny side you will see leafiness deep into the low recesses.

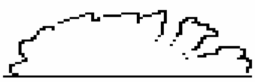

Imperfect Shape on the Shady Side

Many shrubs seem to adopt a perfect dome shape, until you look more closely.

Gaps on the Shady Half.

The backside of shrubs is often tall, untidy, and lacking leafy branches, close to the ground especially.

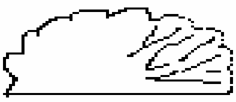

Dead Branches

The overall shape may be perfect, but one side is alive and green, while the other side lacks leaves.
"Dead" South. There is a scale-effect to dead-south: The larger bushes have more depth of foliage.
The more foliage up-sun, the darker down-sun. Only the larger or thicker bushes work best.

Use a Type of Memory which you are Good at

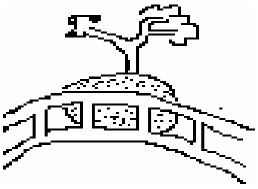

Combine the methods for the most memorable effect.
E.g. pictorial — join images. Make them touch or glue onto, squash or hold hands, follow logically or overlap
— use a stack of pictures to put incidents into sequence.
"After the bridge we came to a red rock and then sat down for a snack at the funny tree"
E.g. ditties — A Capital Tee of Tall Trees at Three O'clock.
E.g. invented "associations"— a vivid or absurd image like a Tasmanian Tiger eating that hill.
Paint it in partly related colours, with shapes and other senses involved, to give extra "handles"/clues.
E.g. Imagine successive scenes in an invented journey/building/furnished room/street scene/pattern points.
Emphasise movement, and links between stages. Make it visual, simple, exaggerated and appropriate.
e.g. "I went inside the cave and talked to the sleepy bear (Ranger) about the bearing to follow, and he said "Seek the morning sun", so coming back out I turned to the east until I met the keeper of the bridge who said "over my dead body lies the path to right; and left". So climbing up over him...
E.g. numbers — One Wooden Bridge, Two Red Rocks, Three Funny Tree… to reflect the hours of the day
Make up a story with rhyme, from e.g. 1, bun for lunch, 2, through the dew, 3, see the sea… 7 rhymes with leaden/Heaven/redden, 11 with Aladdin/pelican/caravan/a raven, 12 with elves/shelve/delve/selves
E.g. shapes — Make a 10-rock pyramid at the 10 o'clock stop; an 11 stick teepee at 11; a 12 leaf hexagon
— e.g. see-saw on a log at 7 and draw a see-saw in your notes—it has 7 corners.

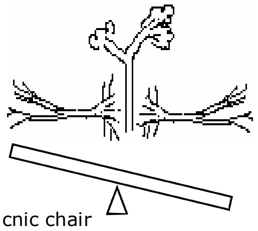

E.g. You can weave a story, of invented images to go with the times of day.
e.g. each digit looks like a... 0 hole/ball/eye, 1 stick/pen, 2 cobra/sphinx/swan, 3 bosoms/camel, 4 yacht/picnic chair
5 measuring scoop/wheelchair, 6 bomb with a fuse/yo-yo, 7 arrow/step, 8 shotgun/glasses, 9 balloon/comma,
10 ball bouncing off a wall/golfball off its tee, 11 jail/stump/goalpost, 12 jail-break/caged lion.
E.g. acrostic — 7 a.m. Z. for Zero Rule Zero: Don't forget to pack the lunch!....7:15 A Bend A in the Creek
7:30 B Bridge across the River at Bend B. 7:45 C We didn't Cee the third bend on the map!
Using this method, elapsed hours happen at Z,D,H,L,P,T,X i.e. every four letters.

Use the Name of the Track for the Acrostic

L ying down tree
E ee! There's a bull!
E e-normous bracket fungus
S wim
P addled Across
A lternative Detour
D
D
O
C
K
S

*Every quarter of an hour, look for things beginning with that letter to remember that kilometre by.
*four letters make up one hour *Use stepped word pairs, 'Lying East-west; Exciting Exit; Extra Searches
*Do Something Alphabetical
E.g. Lie down at L; Scream at E; Examine the fungus at E; SPlash at SP
*Make all your names begin with that letter e.g. Pot-holes, Pines, Possum trees, Possible campsite, Photo
*Write them all down, but try to memorise the best one.
*You can attempt to improve the sequencing, by taking the letters two or more at a time e.g….
LE In the Lee of the lying down tree
EE EEE!
ES E.S.P. told us to find the track next to the E-normous bracket fungus
SP Splash Splash. Had a Swim
*Try also to relate picture to picture in sequence e.g….
Look at the Long horns on the Lying down tree; Enormous bull!; Enormous bracket fungus, dripping with water… Swim!

Acrostics combine individual memories with a clear sequence — firstly by letter, and secondly, by overlapping one pair of letters with the next, and thirdly by overlapping one picture with the next.
You have already memorised many long sequences of letters, e.g. the words of a song, which could be used as a unique acrostic.

Headlands, Bays, Prevailing Winds, Longshore Drift of Sand, Creeks

Headlands are ridges. Bays are valleys.
The ridges determine the headlands, bays, creeks, and — look for this — the position of offshore rocks and islands, especially where a 'point' marks a change in the trend of the coastline.

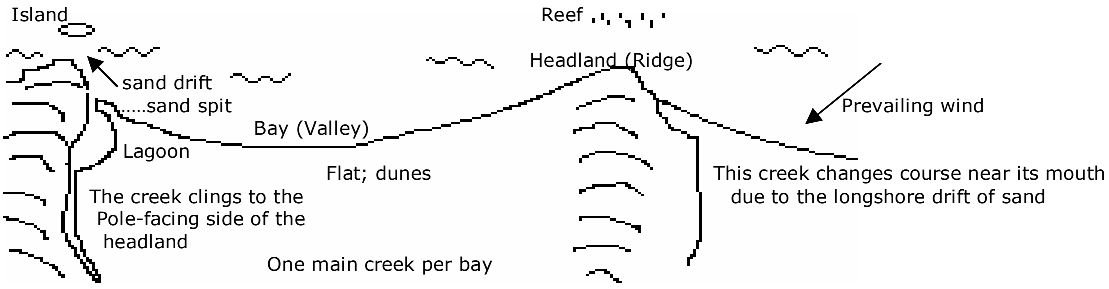
Island
Reef
sand drift
….sand spit
Headland (Ridge)
Prevailing wind
Bay (Valley)
Lagoon
Flat; dunes
This creek changes course near its mouth due to the longshore drift of sand
The creek clings to the Pole-facing side of the headland
One main creek per bay

Beaches — Slope, Particle Size, and Wave Energy

Energetic erosion washes away any small particles.
A steep beach is a high energy beach with big particles.
When it is deep offshore the waves break with high energy.

A flat beach reduces the energy of breaking waves.
A low energy beach is flat with small particles.
Sandbars and shallow sandy offshore seabeds make the waves break over a long-distance.

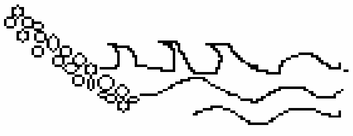

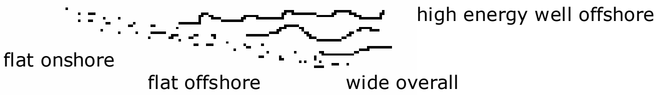
high energy well offshore
flat onshore
flat offshore
wide overall

Much the same seems to happen on land slopes — the flatter slopes have fewer rocks. I.e. the accumulated silt which is not washed away by surface water, covers up the larger rocks.

Sea-levels have been much higher in the past, so that there will almost always be buried boulders just above the present seashore.

Look Through the Trees, to the Sky and Skyline

This is a habit which stops you being befuddled by bends. You focus on the most distant thing you can see…

- E.g. the farthest tree trunk
- E.g. the skyline — ridges, gullies, peaks, light/dark division
- E.g. the clouds — but do stop to examine their movements occasionally!
- E.g. the Sun, the Moon — obviously the best, but they move too, and… don't look at the Sun!
- E.g. the dark blue polarisation band — Use Polaroid sunglasses if you are walking a forest track and can see only the sky.

Don't forget what is abreast of you e.g. a particularly tall tree. If you need to, look behind for some reference point.

Habits take practice to develop. This one is easy, because once you appreciate it, you won't want to give it up.
— It turns many small bends into a few major trends —

Periscope

Climb a tree. Climb a hill. Stand on somebody's shoulders. Or try this…
Attach a mirror (bigger = better) at 45° (try to be exact) to one end of a long stiff stick (e.g. a fishing rod)
— then scan the tree top horizon for landmarks.

Navigating by Trees

Tall trees on high ridges stand out against the sky and are visible for miles around.
Get to know the biggest, tallest, most visible ones.
Individual trees, gum trees especially, are wonderfully fingerprinted, by their branching pattern, outline, position & orientation.
E.g. look for the direction between the two largest branches. This stands out like a weathercock whenever you glimpse that tree again.
E.g. locate the relative positions of a handful of the best trees, so as to cover a wide area.
E.g. try to note the relative direction of pairs of trees.
Most big trees will have large branches to noon. So look for big branches which are *not* pointed to noon.

If you keep looking through the immediate green barrier for more distant (unmoving) skyline trees — you will notice them and recognise them surprisingly often.

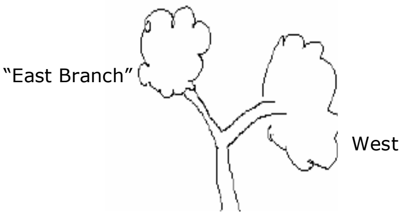
"East Branch"
West

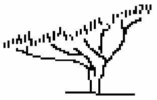

'Solar-Panel' Tilting of the Canopy Top

Some trees may be 'flat-topped', but they are not 'level'. They tilt toward noon. Best seen from side on.

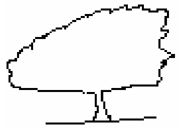

Wedge Shaped

Branches seem to know when to stop growing, and so manage to display
a fairly **uniform**, slanted **flat wall of leaves** toward the outside and especially toward the light.

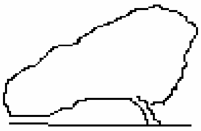

Lopsided Lean

Branches on the backside usually reach higher to bask in the skylight.
Branches on the front side are more horizontal, lower and longer.
Some trees don't seem to try to compensate and simply leave their trunk behind, on one side.

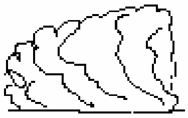

Half-Dome

— an extreme case of a steep back side

NOTE that all these signs work best in open woodland, not in forest understorey, but not always...

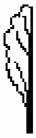

One-Sided Leafiness

E.g. Banksia trees
Such extreme asymmetry often only occurs in the understorey — where light is limited.
E.g. After a bushfire a gum tree revegetates from the trunk — most leafiness survives on the sunny side.

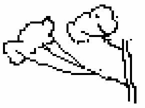

Individual Branchlets

— show similar shapes to whole trees
— tilted, asymmetrical twig patterns, one-sided leafiness, etc
The effect will show up best on the east and west sides of a plant, where the lighting is most asymmetrical.

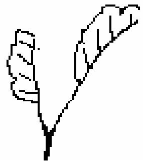

Bare Backbone

Look for it on individual branchlets, on the rear-side of main branches, and sometimes on whole trees.
You could run your finger up the entire length on one side of many shady-side branches.

Group Several Related Images into One

This is a standard method of making your memory ability efficient.
You have less items to remember—they become a few memory prompts, each to a handful of items.
E.g. 4 groups of 4... 6,7,8,9 a.m. / 10,11,12,1 / 2,3,4,5 / 6,7,8,9 p.m.
Write each group out visually shaped e.g. at the points of a square for a group of four
with 2 opposite squares for a group of eight
inside an infinity sign for a pair of items
List the hardest to remember items, first and last.
Group the shapes i.e. the groups, into another visual structure.
Use symmetry wherever possible—it makes it easy to recall.

The Start-Stop-Rest mnemonic (see Section 5. p.110) is a good example of this "clumping" technique—never more than a handful of things to remember, yet it leads you into hundreds of points.
Put the S for Sky at the top, and use a hexagon for the six start up points. Use a clock-face for TIME, and the compass-rose for NSEW.
Balance it all on a triangle, for the 6 a.m. start

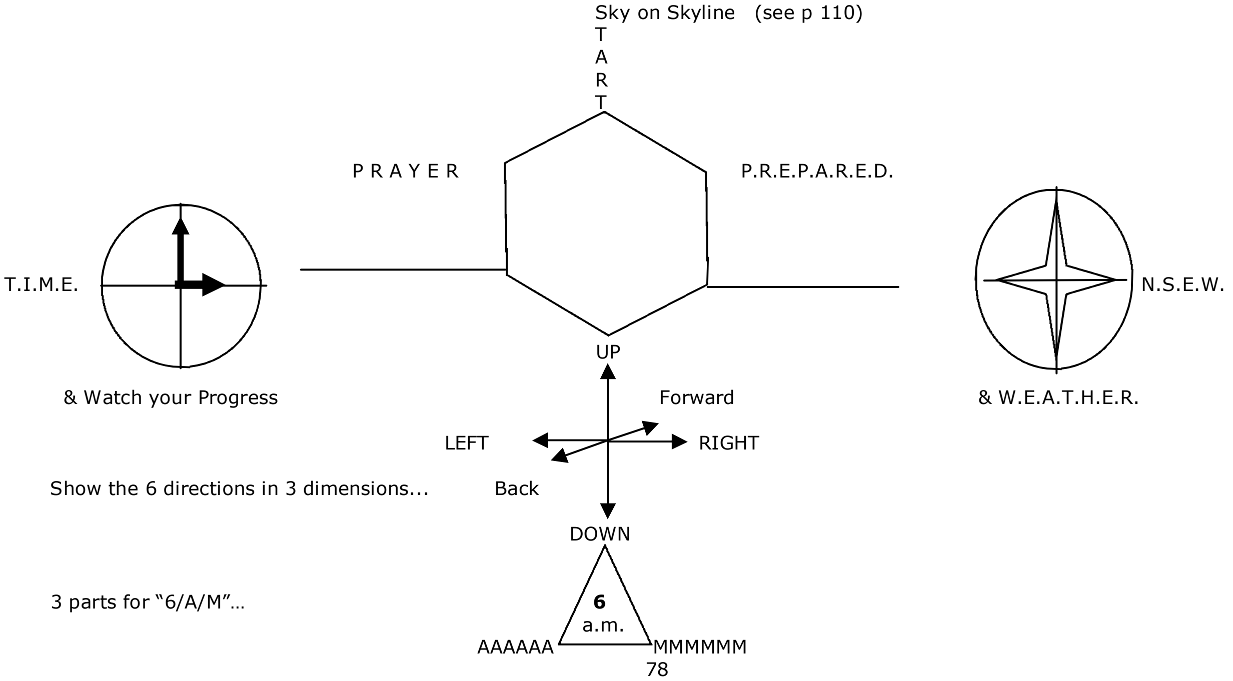

Show the 6 directions in 3 dimensions...

3 parts for "6/A/M"...

Rock Typing

Hardness — how easily eroded. Keep trying to scratch it until you find something just harder than it e.g. your pocket knife.
Colour — what colour silt. Scratch it against harder rock and look at the colour of the crumbs.
Sedimentary — the components are not of local origin. Layers — sedimentary origins. Melted sediments — crystalline bands
Sand content — grano-diorite — pale & silica rich. Clay content — basalt — dark & mineral rich. Mineral crystals in it.
Crystallisation — small crystals mean rapid (surface) cooling; large crystals means slow (deep earth—'plutonic') cooling.
Mineral content — what sort of soils does it produce? Flakiness. Behaviour in fire (Caution!)— it cracks, blisters or explodes! Melting pt.
Density, Magnetism, Fracture Pattern, Transparency, Permeability, Acid Solubility—evaporites and marine deposits may be limey.
Understanding these details is a key memory aid, since it gives you more meaning to look for, to memorise and recognise.
E.g. in basalt areas expect a more diverse & rich flora & fauna.
E.g. in the sandy plains, expect rivers to run underground, trees to have deep tap-roots and to line only the dry watercourses.
E.g. in dry sedimentary areas, look for a harder 'evaporite' layer underneath.
You don't have to identify the mineral, but you should characterise the rock type.

Soil Typing

1. Climate drives soil typing. What is the climate?
2. Latitude drives climate. What is the latitude?
3. High altitude mimics high latitude. What is the altitude?
4. How is the local climate modified? Look especially for temperature and rainfall factors.
   E.g. ranges upwind? A cold landmass? Warm ocean currents up-weather?
5. Evaporation produces salty limey evaporite deposits, whenever the evaporation rate exceeds the rainfall rate. High-temperature and wind, speed up evaporation. Flat inland areas receive less rain. Where rainfall exceeds evaporation, the soils are acidic, leached, and deep. Does the evaporation rate exceed the rainfall rate annually?
6. Climate drives vegetation as well as soil types. Soil types drive vegetation. The three go together. What is the vegetation type?
7. What is the parent bedrock? Granodiorite produces pale, sandy, friable, well-drained, mineral poor soils.
   Basalt produces dark, clay, dense, waterlogged, mineral rich soils.
   Warmth & rainfall aid chemical weathering and erosion rates of the parent bedrock, and so produce deep soils.
8. What are the relative constituents? Shake some soil and water together in a glass jar and leave it to stand until it is layered. Organics & charcoal float. Heavy minerals, pebbles & sand sink. Silts lie on top of them; then clays. Colloids remain suspended. Organic remnants accumulate in soil when it is warm enough to grow plants, but cold enough not to rot them away quickly. The decay of organic materials then provides acids for leaching, provided that high rainfall doesn't dilute the organic acids. Can you see from the sort of soils how cold and wet it has been?
9. What are the depths & layers of soil? It is leaching which transports poorly soluble salts to lower layers, until the water table dilutes the acidity. Subsoil clays also accumulate where leachates are deposited. Evaporites accumulate just below the surface and above the water table. Dry areas have poor soil depth due to slow bed-rock weathering and erosion rates.
10. What is the soil colour? Red indicates partly leached soil — iron and aluminium salts remain. Paler, ash-coloured soils are left behind after further acid leaching. Black soils result from poorly drained unleached soils. The importance is that leaching reduces fertility due to the lack of soluble plant foods. Leaching is aided by warmth (except in the case of limestone which dissolves better in colder climates).

Bush networks

The bush roadmap is made up of a network of interleaved ridges and rivers.
Instead of a major road or a minor road, you have a main ridge or river, shorter ones, and dead-end ones e.g. a 'shoulder'.
A 'cross roads' is where two creeks & two ridges meet — a saddle.
The seashore corresponds to a boundary fence; Peaks and islands act like townships — they are a focus of 'roads' and ring-routes.

Don't Navigate by Networks Only

Networks are fractal — big **creeks** have little creeks running into them; **walking tracks** have side tracks branching off from them; big ridge lines have little **ridge lines** running off from them;
dirt roads have smaller dirt roads which have still smaller dirt roads joining in.
That is why it is difficult to recognise the scale of any map just by looking at the networks on it. The more detailed the map, the more detail there is to show up. The smaller scale the map, the more detail has to be left off.

So there will always be small creeks, unmarked roads, minor tracks and minor ridges to tempt you into thinking
"This one is what is marked on the map".
You will likely find one "exactly" where you look for it, but it may not be the right one.
Don't navigate simply by expectations of network junctions.
E.g. the next creek-line has a 50% chance of being "just as I expected it — a creek-line coming in from the right"
or if not expected, the 50% chance is that you will say "probably just not marked on the map"!
If by chance it is actually correct, it will only encourage you to continue navigating by network,
but you won't always pick the correct one.

Antidotes For Guessing at Network Junctions

- **Distance** covered. Accurate measurements of progress can eliminate some minor junctions which are "almost" where you expect the desired junction to be.
- Trend — of the junction branches. I.e. the **directions** must match your expectations too. E.g. "Each side of North"
- Landmark **bearings**. Try to separate identical looking junctions by cross-checking their position in relation to skyline features.
- Care.
- **Count**. Look-alike branches are nonetheless unique, when numbered.
- Recognition, from last time, if you had bothered to identify its peculiarities. I.e. bother now, for next time.

Branching Patterns — Examine Them for thickness, length, angle, curvature

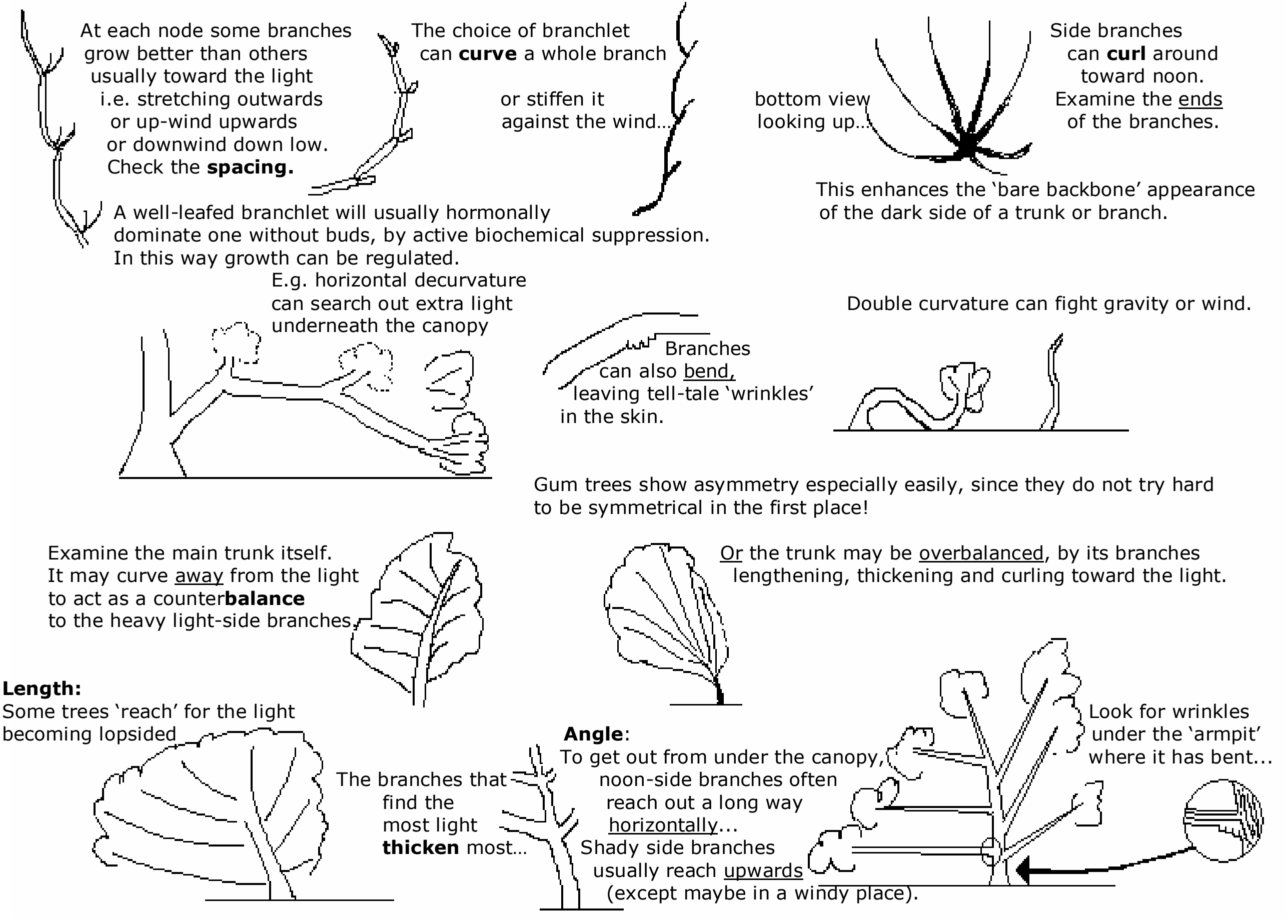

At each node some branches grow better than others usually toward the light i.e. stretching outwards or up-wind upwards or downwind down low. Check the **spacing.**

The choice of branchlet can **curve** a whole branch

or stiffen it against the wind…

bottom view looking up…

Side branches can **curl** around toward noon. Examine the ends of the branches.

This enhances the 'bare backbone' appearance of the dark side of a trunk or branch.

A well-leafed branchlet will usually hormonally dominate one without buds, by active biochemical suppression. In this way growth can be regulated.

E.g. horizontal decurvature can search out extra light underneath the canopy

Branches can also bend, leaving tell-tale 'wrinkles' in the skin.

Double curvature can fight gravity or wind.

Gum trees show asymmetry especially easily, since they do not try hard to be symmetrical in the first place!

Examine the main trunk itself. It may curve away from the light to act as a counter**balance** to the heavy light-side branches.

Or the trunk may be overbalanced, by its branches lengthening, thickening and curling toward the light.

Length:
Some trees 'reach' for the light becoming lopsided

The branches that find the most light **thicken** most…

Angle:
To get out from under the canopy, noon-side branches often reach out a long way horizontally…
Shady side branches usually reach upwards (except maybe in a windy place).

Look for wrinkles under the 'armpit' where it has bent...

Working Names — Invent Them

E.g. Past the 'Little Castle Rock'
Up the 'Yellow-Crowned Tree' trail
Under the U-shaped rock face
Between the Big-Boulder Bluffs

"40 minutes in" is a handy name — when you recognise it, coming back out.

Stupid gawky names help you remember, but even better is to do something stupid to match.
E.g. Gallop like a galah, through the pass, and call it "Gallop Gap with Galahs Galore!"

Standard Name Plus Personal Name

Actual names last longer and are widely known. Use them.
But make up a personalised name too.

E.g. "Oh it is actually called 'Boomer Hill',	but we call it Porridge Hill, because it lies halfway between Oatlands and the Table Tier!
E.g. We call Wollongong "Kaus Australis"	'coz that's where we watched Kaus Australis go straight overhead.
E.g. Kaus Australis becomes	"The Wollongong Star"

Remembering Names

To help you later, when the name is 'on the tip of your tongue',
Count the syllables, consonants, short vowels, long vowels, now.
Rhyme it with something appropriate e.g. Condominium Creek has a pond of minimum seep.
Look at the spelt visual shape… for ascenders, descenders, double letters, symmetry.

Then later run through the alphabet for a guess at the first letter, then at the second, and so on.
Think about anything related, like why you went, what you ate, who you met, what you saw, the weather, the season, the news at the time…
Ask each person for these recollections. Recreate the full scene.

Contour Lines

Contour lines mark horizontal ground.
A simple trick to match up the map contour lines and the landscape, is to **point horizontally** parallel to the ground.
E.g. if you are in a valley, point along both creekbanks, using both hands. The steeper the valley sides, the narrower the contour angle.

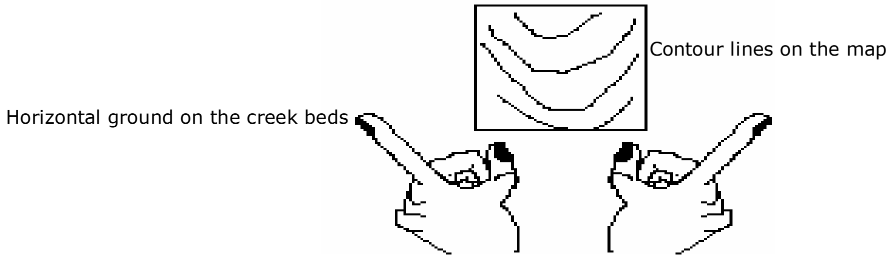
Contour lines on the map
Horizontal ground on the creek beds

Contour lines are like water-marks on the bank of a reservoir, lake or swimming hole, visible after the water-level has dropped.
But bigger like a global flood drying up 10m at a time, leaving water-marks.
like a photograph of it, from above.
like horizontal slices progressively taken off from the top of the mountains, as when you slice clay with a thin wire and you look at what is left.
like building up a model landscape one layer of cardboard at a time, from below, with progressively smaller shapes cut out according to the map contours.
Then you draw the results on a map.
They show the path you have to walk to go 'around the contour' so that you neither climb nor descend.

Contour Navigation: **Practise** It, **Before** you Need It

At night, or in fog, or in a blizzard, you may have no other clues, but the contours, on the map, and underfoot. Try to match them up.
Hint: Don't expect every little bump to show up on the map, but only the average smooth slope.
Hint: Keep good track of compass direction, time & distance as your best crosscheck. Assess the wind. Use an altimeter (,GPS, radar…).
Hint: **Delegate** tasks — map reading; course holding; distance logging; slope assessment; exploration; memory…
Hint: To go quicker, go slower — so as to be sure. Explore the slopes to left and to right. (But don't get lost or disoriented!)
Hold regular conferences to check your judgments.
Hint: Assess slopes by looking sideways, across them. They always seem so steep when you look up!

Boat navigators in a fog, can sail toward land until they reach a predetermined depth, then sail along parallel to the coast, using the depth contours marked on their chart, and some method of finding the depth. Marine depth sounders rely on a good echo, but you can always toss a weighted line over the side to find the actual depth. Make sure to make corrections for the depth of the depth-sounder, the depth of your keel, and the height of the current tide!

Sticking to the Bush Network

- Walking **up-ridge** may perch you on an isolated hillock.
- Walking **down-stream** must get you to salt water; or at least to water, civilisation, flats, roads; but also to thicker vegetation and a loss of view; and sometimes to impassable waterfalls and gorges.
- Walking upridge, uphill, upstream lands you in steeper, drier, more jagged terrain with more open vegetation, a better view and maybe an along-ridge track.
- Walking down-ridge will "take you to the lake", but simply walking downhill may lead you into a heavily vegetated creek-valley which then drains into the lake.

Natural Routes Naturally Converge

That means that, in reverse, they must also **diverge** — which is dangerous, open to mis-navigation.

- Provided you don't cross catchment divides, walking downstream or up-ridge is **convergent** and so is safe for a one-way trip, but **un-safe as a means of retracing your steps back to base**. E.g. it is easy to climb a hill, hard to retrace exactly.
- Walking down-ridge is **divergent — deceptive and dangerous**.
  You cannot easily see ridges peeling off below, and the map won't mark all the creeks you can see.
  If you are trying to retrace your steps or navigate from the map, you will have to be very careful.
- Ridges converge to hilltops — so walking **downridge** leads you widely astray.
- **Upstream** is divergent — you don't know where you will end up, but at least you can find your way back, because creeks converge toward the river mouth.
- Tracks & trails converge to: popular features, bridges, car parks, waterholes, homesteads, campsites, clearings, etc.
  So **leaving a focal point faces you with multiple choices** "How on earth do I get outta here!"

Coping with Divergence

The result of "directions" such as "You can't go wrong. You can't miss it" is that the truer that is, the easier it is to go wrong — after you have found it. Retracing your steps, or finding the desired way out, onward, can become a nightmare.

- Ask directions for the onward bit, not just how to get there.

For backtracking…

- Pay attention on the way in, despite your excitement at arriving. Note the direction of your arrival. Look back regularly. Mark the junctions you used, especially from minor out onto major paths. Retrace your own tracks.

Reverse logic

Be like yachtsmen, who should always explore harbours, shallows and tributaries at low tide "in case we do get stuck".

- Explore upstream — a divergent labyrinth — so that you can walk downstream to return.
  Provided you don't cross water-divides, or overshoot your starting point, "you can hardly go wrong".
  Ahah! There's the hint: Before you set off upstream, peek behind you, to see when to stop "coming back downstream"!
- Explore downridge, and climb back upridge later. Only don't climb down anything you can't climb back up.
- Explore away from a focus, so that you can converge back to it.

Isolated Vegetation Shows the Clearest Asymmetry

E.g. Look in any **clearing** for downwind debris/erosion,
and for up-Sun/down-Sun asymmetry at the edges.
E.g. an isolated tree in a paddock...
E.g. an open creek will give the bushes a clear view across the water...
E.g. a roadway will exaggerate the light asymmetry on one side
and suppress it on the other...
E.g. Look in the wheel ruts of bush tracks for asymmetries due to wind & light…

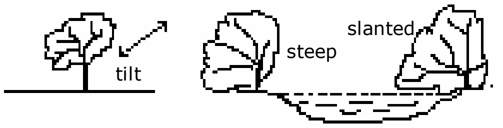

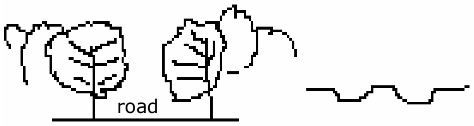

Examine **the tallest trees — The emergent ones** have a clear view of the Sun.

At the **edge** of a clearing, subtract the effect of wind…

Closed Canopy — Look for Gaps

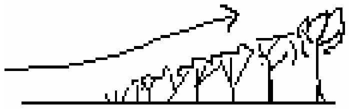

Each tree canopy might look uniform, but be blown back on the upwind side, leaving a gap there, with dead branches, where it *has tried* to fill the gap.
But look for a gap without dead branches and without other explanation
— It might be 'shady south', which gives the tree little incentive to cover.
The tree *could have* spread that far by now, if it had wanted to, but hasn't.

Tall and quick growing saplings **show the most asymmetry…**

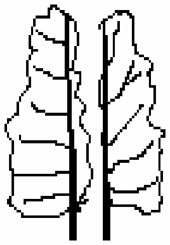

They can be the best tell-tales of all, and the worst!
Because they respond to clearings, slope, water-shine, nearby trees — be careful!
Take extra care to average out their message.
Any one tree gives you too definite an impression!
Back-to-back saplings may show opposite asymmetry, by shading each other...
Treat the overall shape as one tree.
The sapling effect, arises from young forest trees shedding branches more often and more easily than mature trees.

Roadsides & Campsites Reverse Some Effects!

Examine trunks for broken or lopped boughs before interpreting!
The bigger lower branches are easiest and best firewood pickings!

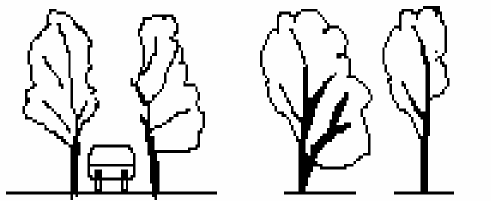

Name the Subsections

Heart-Attack Hill	Bust-me-gall Bend	Starve-gut-Ridge			: Be Vivid
On-the-Double Decline	Cross-the-Creeks Roller-Coaster	Blackberry Bank			: Be Descriptive
Left-hand Loop	Clockwise Climb	Right-around Contour	The Zed Bend		: Be Cheiral
N-E Straight	Into the Sun Slog	West Bank	Downwind Dunes		: Be Directional
Long Leg	Short Stretch	The Middle Mile			: Tell the Distance
Last-legs Leg	Dinnertime Creek	Sunset Strip			: Tell the Time
On the Left of Third Hill	Nine Mile Beach	Five Ways	Fourth Island		: Count
The Bush Blocks	The Rock Band	Turning Point	The Up-Part	The Trip down	: Be Clever
Rubble Ramble	Possum Passage	Flower Flats			: Alliterate
The Birdwatching Bit	The Deer Park	The Gloriously Green Glade			: Be Observant
The Monkey Saddle (three creeks)	The Big Bulge	Round Shoulder	Tight Turn Tongue		: Be Shapely
The Breakaway	The Hither Side of the Hill	Round the Rim	Unto the Brink		: Be Literary
Brown Beach Bay	Wobbly Shacks (= Mount Wobbly behind the shacks)	Zigzag Track			: Be Visual
Perverse Reverse	Inland Cusp	Round the Back	Out to Sea and Back		: Name Whole Curves

Name the Sequence

Fit Each Subsection into a Larger Scheme.

- Name the Links : The Cuvier Track The Mount Barker Road The East West Ridge
- Name the Junctions : Point Pass Steppes Junction Tee -tree Junction
- Contextualise : Beyond Burrowing Crayfish Ridge Cross Catchment Climb
- Overlap Sections : Deer Sally to Sandy's Shore to All-Terrain Trek } these are simultaneous
  Sherwood Forest to Bumpy Flat to Uphill Haul } and overlapped 50%
  I.e. halfway through one section, think about how it is changing character
  into the next obvious section, and anticipate it.
- Time-Distance Clues : Lunchtime Lookout The Approach Thither Heather Tea Junction Ten-Mile Dam
- Lie-Linkage : Downstream from Dusty Waterfall Wind-Shadow Plateau
- Numbers : Fourth Bump Fifth Creek Tee Junction # Three
- Acrostic : Short Steep Step in the Southern Subdivision after the Shallow Segment

Contour Curves: Your Job is to Imagine the Shapes into Life And to Match the Lines to the Landscape

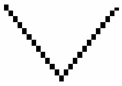

Deep Vee 'Ditches'
— they gather water.

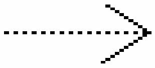

The **sharp** points, point **up**-river. (sharp/up rhymes in p)

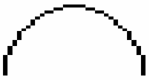

Round 'Shoulders' — they shed water.

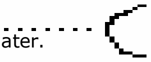

The **round** points point **down**-ridge. (round/down rhymes)

— being higher ground, they separate gullies

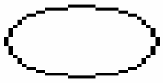

Peak or Depression — you can't easily tell which. Peaks are much more common.

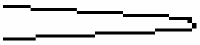

Hairpin 'Ridges' — they can look parallel, like 'a slope', if the edge of the map cuts them off.
Ridges usually show much rounder than this.

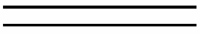

Parallel 'Slopes' —be careful about which way is 'up'. Check that they are not the same height—on a ridge or river

Here is a typical confusing array of parallel lines.
The first two are ambiguous, but seem to lead down into a valley…
The valley is shown as lines 3, 4 & 5…

The three corners here are all sharper than the corresponding ridges.
Downstream is to the left…
Imagine a deep-vee ditch.

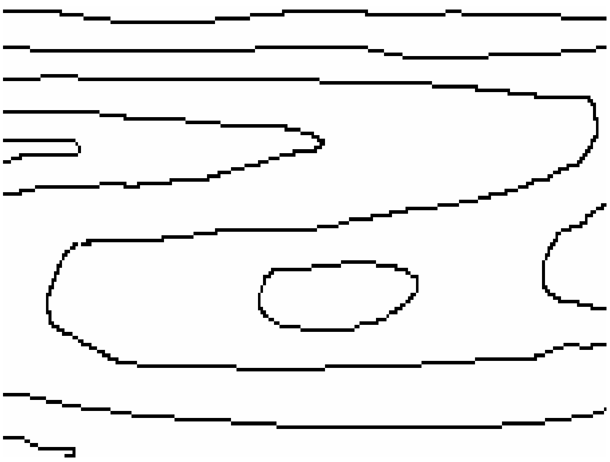

The shoulders and peaks shown here are all much rounder…
..Uphill is → peak ← saddle → ridge or peak ……….
Imagine a circular mound with a spur down to left and right.

Down here, upslope to higher ground is… ↑

Trackless Navigation

When you leave the bush network to cut across the natural routes, you still navigate by them, by noticing when you cross them.
You choose a route in relation to the framework — of ridges, rivers, and the contours that cross them at right angles.
E.g. "I am now going diagonally right, uphill, on the left bank, to the Ridge, then…"
You describe the sort of route you are following, and **start a new subsection whenever that sort of route changes** in quality.
E.g. A dry shallow creekbed marks a change from downhill to uphill.
E.g. The major break of slope—on the "hillside"—marks the change between being on the "hill/crest/ridge" or in the "valley".
Don't forget to ask, as you cross each 'road', what that ridge or creek is doing, to the left & to the right e.g. joining, splitting, rising, falling, dividing major valleys.

The method of Names

The point of this method is to bring your normally subconscious memory into conscious focus.
You put a **name** i.e. an essential description, onto each subsection. This is in addition to naming the creeks and ridges you cross.
That is equivalent to supplying your own signposts & flashing lights to the landscape. But write down the other details also.
As with memorising any journey there are a handful of steps: Notice; Recognise; Describe; Name; Remember; Put it into Sequence.

Bush Networking Essentials

Answer up to six backbone questions about each leg, before deciding on an appropriate name for that subsection…

1. Hill-wise — What Peak? — Is your direction trend in relation to it Uphill or Downhill; towards or away?
2. Hillside-wise — Slope — Choose between Upslope/Downslope/Flat/Contouring.
3. Ridge-wise — Rise — Upridge/Downridge/Crossridge?
4. Drainage-wise — Lie — Upstream/Downstream/Crossstream?
5. Side-wise — using the standard names* — Left bank/Right bank (below the break of slope); Right flank/Left flank (above).
   *As a yachtsman going **up-**stream sees it — Port and Starboard. Use these terms in your names to remind you.
   As a traveler going **up**-ridge sees it.
   I.e. the standard directions are **up**-stream; **up**-ridge; **up**-north; **up**-wind; **up**-hill, **up**-slope.
   Be careful! The 'right bank' is the 'left flank'. You can climb to the right on the left flank (of a ridge which falls to your right).
   Walking diagonally 'downslope' can be 'upstream'.
6. Hand-wise — Tack — Clockwise around or Left around/On the left diagonal or Right diagonal?

Mnemonic: (With your right index finger) touch your **head** (the peak);
forehead (the slope);
bridge of your nose (the break of slope);
down the ridge of your **nose** (the ridge);
nostrils (the drainage);
right **cheek** (the side);
and now look at your **hand** in wonderment (What is my tack?)
Then **point** out some peculiarities...
Now **add the non-essential** — exact measurements e.g. direction/length/time/steepness/count —and descriptions, sequence, context.
(Remember: don't navigate only by networks.) See opposite, for descriptive names.

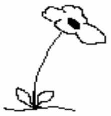

Flowers Respond to Sunlight

Examine the stalk-lean and flower-head tilt
e.g. of daisies.
Look up-Sun and down-Sun to see the striking difference, whenever the Sun is low.

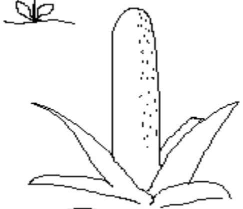

Look for the first flush of flowering on the sunny side of a flower spike
E.g. Yakka (Xanthorrhoea). The mid-morning sunlight direction gets the earliest flowers.

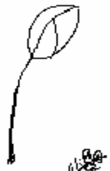

Examine closed flowers to see whether they betray where the Sun last came from.
Many flowers 'follow' the Sun around.

Some leafy stems seem to tilt to the noon, perhaps to point their terminal flowers into the Sun.
E.g. Sea Spurge

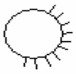

(Top view)
Banksia pistils sprout earlier on the sunny side.

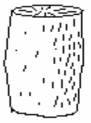

Banksia stamens deepen in colour on the sunny side.

Fruit Distribution and Ripening

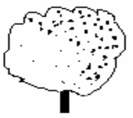

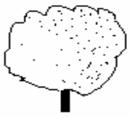

Some bushes will fertilise more flowers in the sunlight because pollinators can see them better.

Fruit are fertilised flowers. So some species 'set' more fruit on the sunny side.
Look in all directions to see which aspect presents the most flowers or fruits to view.

'Ripe' colours are a result of sunlight e.g. apple, peach.
Sun-ripened colours aren't confined to fruits.
Mushrooms are often old enough to have seen some sunlight.
One mushroom in the open may not help you much, but they do come in flushes, so look for a *pattern*.

Remember the Sequence from AA to ZZ

— A Ready Made Acrostic of memorable unique items, with endless extension if needed. I recommend it as the pick of the mnemonics.

You may need to remember each tree in a forest walk, or every twist in a path, or every choice in a maze, or every square in a grid.
AA to ZZ supplies 676 *differently named and sequenced letter pairs*.

If you turn the letter pairs into appropriate *two word phrases*, there is no limit to the useful memory aids.
E.g. I could call this system the Alphabetic Zygote (of 'letters paired' A-Z).
As with all gimmicks, it only works best if you *study and* ***memorise*** **the object *receiving* the label**.
Examine the tree for anything which could attract the label DF (damp foliage; dendritic form; downpointing flowers), or mentally *photograph* the MS munch-stop, or *actually* bird-watch at BW.

It's up to you to use the letter pair in a memorable way. Don't forget to use visual, emotional, sequenced and appropriate words.
Consistent rules help.
On a pure sequence walk, begin with double letters, rather than with A-Z single letters. This makes two word phrases the consistent rule.
In fact begin with *any* double letter, but especially **the initials of the walk**, e.g. BL for Blackmans Lagoon.
Two word Phrases e.g. AA = Ask Again, avoid the ambiguity about which other letter in the word or phrase might be emphasised.
E.g. is Aeroplane, AE or AR or AP or AN? Was 'Anabolic steroids' A or AS?
Add a third word to the phrase as needed, to spell out what it means, e.g. Backwater Island turnoff.

The trick is to make the mnemonic **match** the feature.
E.g. if there is a plague of mosquitoes which you attack with a stick, Aaron's Almond-tree walking stick might be appropriate for AA.
Then the need is to somehow link the previous with the next.
E.g. AB Aaron's Budded walking stick (if you know that story), might focus your attention on the buds in the next gum-tree.
Without a '**theme**' it is difficult to pluck out your memory what 'DF' might have been. But 'Doctor's Emergency' where you stubbed your toe, leads on to 'Doctor's Ferari' when you walk on quickly anyway, and on to 'Doctor's Gasoline' where you topped up with lunch.

If you do a 'grid-line walk', you use A-Z to begin with, and say you move from A0 to C4; you might go through A1, A2, B2, B3, C3, C4.
The memory aid goes A0, then Ahead on 1 leg, hopping, then Across, on 2 legs, Ban-tu, Ban-three, See the three cattle?, Count to 4...
'Minus', in cases of ambiguity, might come out as 'which was 4 metres deep', or 'where I left 4 pebbles behind', or 'and Not 4 blisters' or 'at anything but 4pm'.
When you get into AD you look for a-d **peculiarities** to remember like 'Add 4+6 wattles (to get 46)' or '33AD crucifixion tree' or 'Adder-Death, two score of them'.
As an alternative, name all grid squares with *letters*, e.g. D-G, and look for things to name, such as Dark Green; Discarded Gaiter.

In a maze, you could choose Long words for turning Right; Short words for turning Left, a mixed phrase for Straight Ahead. These match the meaning-words in length of sound.
Maybe you need some **variations**:
One, two, three or four syllables for left, right, through, reverse.
Long first vowel sound vs short vowel sound.
Begin with one *letter* and *end* with the other *sound*, e.g. Amanda, Abib, Ace, Add, Abbey…
Let sounds take priority where needed. E.g. few words start with K, but many start with a hard C. X is hard unless you allow Ex-.

Memorised sequences are most effective if revised almost on the spot, e.g. via a leap-frog walk (2 forward, one back), and on the return.
Equally, other methods are enhanced by adding this mnemonic, such as when marking a trail, writing track notes, or making a mudmap.

Contour Heights: Read the Numbers!

Determine **the Contour Interval** — and interpret which way is 'up' from the **numbers**… Note: On maps, the numbers are printed to be read as if looking **uphill.**

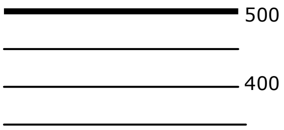

This first diagram is "every 50 m".
And 'up' is toward the top.

Zero means "sea level"
Read the absolute numbers!

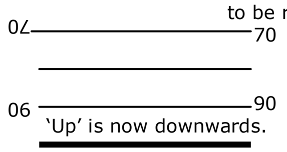

In the second diagram, the contour interval is every '10 m'

'Up' is now downwards.

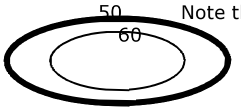

Note that **major** contours may be marked as thicker
e.g. every 50 or 100 m

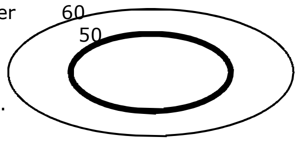

..The first is a peak; the second is a hollow...

Marine charts show spot-depths, 'soundings', in metres and decimeters, below lowest water, with offshore depth-contours.
If the "depth" is underlined, it refers to a "drying height" of intertidal rocks.
At sea, keep a record of the soundings as you sail in to land, preferably in a straight line. Plot them onto tracing paper or mark them on the edge of a ruler, and try to match the pattern to charted depths.

Contour Spacing: Close = Steep Slope

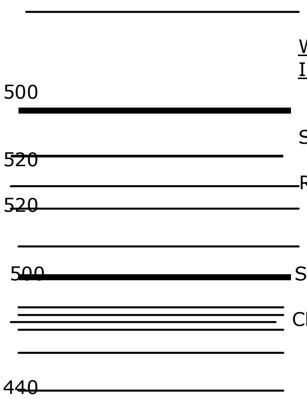

Wide spacing = flat land   Examples: mountain plateaus; valley floors; paddocks; dunes
Irregular contour shapes (meandering all over the map; not shown here) = very flat land

Slope

Ridge — carefully notice when numbers are identical. **The rule "Closer Equals Steeper" only holds for different height levels, not for hairpin contours around ridges or rivers.**
In fact, "close" might equal "flat" i.e. on a ridgetop

Slope

Cliff — Very close = Very steep

Slope; have you worked out downhill?

Walking "Straight" through the Bush

- 'Straight' downstream; upstream; down-ridge; up-ridge; uphill; downhill
- 'Straight' around the contour
- 'Straight' along the path of least resistance
- 'Straight' through the obstacle course, dodging to left and right
- The Line 'of Best fit' — to a chosen trend cross-country

None of these are straight, but most are easier than 'straight across a slope' — crossing innumerable ups and downs!
Most are predictable.
The navigator simply goes last in line with a compass and logs the twists and turns, whatever course that takes.

Why last in line? It is a fact of life that following your nose through the bush quickly disorients the leader of the party, while those following behind maintain their bearings easily.
Also, to clearly show up the trend of a sinuous course, the last person can drag a long smooth string loosely behind him.

Straight around

At a…

- Constant Angle—to the (changing) Sun, Moon, star, planet, landmark e.g. keep the peak "ahead on your left" — a spiral course.
- Constant Distance Off — as judged by the rule of thumb — you keep a prominent landmark appearing to be a constant size — from base to peak. The course is circular, concentric to the landmark. E.g. offshore, around a lighthouse.
- Constant Horizontal Angle — between two prominent features. This gives a cusp-like circular course.

You can retrace these courses, but the going may well be difficult — crossing many streams and ridges.

How to Walk in an Unintentional Circle

Haven't you heard of people walking in a large circle when they are lost?
They thought that they were walking in a straight line, but some hours later they recognise the same scenery!
I've done it; it's no myth. Discover your natural drift...

Blindfold yourself. Try to walk straight across a flat. I bet you can't.
Most people have one slightly shorter leg, usually their left leg.

Now try to walk across a slope, blindfolded of course, in a straight line. I bet you drift.
But once you realise it, you are just as likely to head uphill too much, in overcompensation.

Don't trust your senses in a fog. Find out by trials at home — before you find out the hard way — how far 'off' you are.
Also find out how to properly walk in a straight line in a fog!

Stem and Leaf Colouring

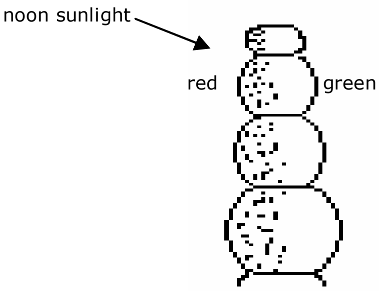

Samphire saltbushes
colour up
on the bright side

Many other plants
show red stems
or leaves etc
Look for a paler side.

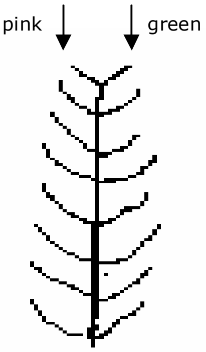

Look from the top
and sway side to side
to find the deeper colour

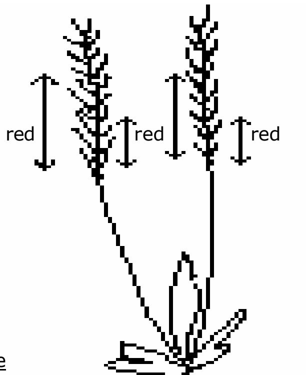

The colour deepens on
the bracts higher up
on the sunny side

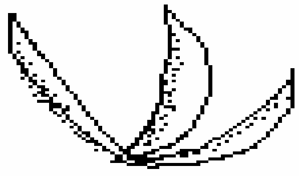

Succulent leaves redden up
on the bright side. Look one
way over the flats & it is
pinker than the other way,
especially just after sunset!

N.B. Older stems may show a reverse trend! The colour fades on the sunny side after having deepened more while young.

Beware: When flower-stalks twist as they grow, the colouring or fading effect is unreliable. You can detect the twisting *by* the fading.

Example: Blackberry stems and leaves redden in sunlight, staying green on the shady side.

New Growth is Colour Differentiated

Looking up- or down-Sun (on a cloudy day) will show up new growth as yellowish or brownish or bluish (depending on the species)
in one direction more than the other.
The sunny side will 'green up' the new leaves more quickly and will look more normal coloured.
But test it out first — it may go the other way — sunlight might promote new growth, before it colours it up.
When new growth is distributed throughout the shrub, the sunny side will often show deeper green (i.e. more of the same pigment)
and be more vari-coloured e.g. some leaves take on a red blush or red margins in response to the Sun.

Revise your Memories

..after a 10 minute distraction
..after an hour, a day, a week, a month, a year e.g. for an acrostic which you wish to remember permanently
..and whenever after, when you begin to have trouble remembering
Choose undistracted, mindless times
e.g. while trudging or being driven, waiting or relaxing, or last thing before sleep, to run over things in your head.
Overpractise your mistakes. Out loud is always better. Take any opportunity to tell someone.

To Remember Formulae such as those at p. 19b

(1) Make up memorable ditties, and refresh your memory, as above, before you lose them.
(2) 'Read' the formulae *for sense* — They are just shorthand for how things *must* behave. If you don't like maths, it's probably because you never understood the common sense behind it, and that is probably because you never had the opportunity to see it, or developed the skill to read formulae.
E.g. '**Sin, Sin, Sin! It's High Latitudes which Declines it**, for both Easterners and Westerners'
for: the three sine ratios in the formula… 'sin (**heigh**t E-W) sin (**latitude**) = sin (**declin**ation)'
i.e. as the latitude increases, the height must decrease, and vice-versa, to preserve any sin (dec) figure as constant.
The formula should return zero height for zero declination, and 90° height for dec=lat.

This next one assumes that you measure the angle away from East-West (rather than from the 'elevated pole', as on p 19b).
You can use 'sine of the variation from E-W' instead of 'cosine of the angle from the pole', because 'co-sine' says 'sine of the co-angle'.
'The sin of the increasing the sunset angle, by cause of increasing latitude, is a sign of decline'.
sin (sunset angle) (multiplied by) cos (lat) = sin (dec)
Read this as: For any specific declination, as lat increases (=the cos reduces), the angle must increase, to preserve a constant sin (dec).
You can also put in dec=0 and expect a 0° answer no matter what the lat.
At the Equator, declination directly determines the direction of set N or S of E-W.

'Why bother with the E-W Sun? Be**cause of its E-W *timing*!** It's **today or too late**' t for tan; d for dec; late for lat;
90° = 6hrs each side of noon cos (time of E-W) = tan D / tan L, *'or' for "o'er" for 'over'*
Lat = 0 throws the calculation into a spin, because any declination other than 0 can't be E or W. Dec = 0 must return 90° for any lat.
Dec = 90 must be impossible (it's N or S!), but so must any dec > lat, since cosine doesn't go above 1. Those stars stay polewards.

'One New Day at a Time? Not quite, **'cause for tomorrow's sunrise time, today's time's not too late**!'
It changes only slowly, as dec changes. cos (time of sunrise) = tan D x —tan L 15° = 1 hr. each side of noon
0,0 must return 90°. Opposite hemispheres must return a large positive cos and hence a smaller angle, and a short day.
Same hemispheres must return a negative cos, which means a long day.
Too much lat, and the tropical bodies disappear altogether, as expected. Too much dec, like a circumpolar star, and it won't even set.

Try **the Equation of Time**, p.17b.
The First FFive months, From FFebruary, Fence-sit Four Features (halF-each side). The next 7 *finish* at four more features.
–1**4**,0,**4**, 0, -**6**,0,1**6**, 0 Makes nice mirror patterns of the two sets of four features. Minus means a fast Sun, up early; a 'slow' clock.
FFFebruary Founders at FFFourteen; Marches up again; April crosses (its hot-cross buns); May turns 4; June double-crosses;
July *finishes* work at six; August rises augustly, to zero; SSSeptember and October get up to SSSixteen;
November and **Dec**ember **dec**line a lot; December crosses at Cross-mass; January Sun decends from the cross, fast, for a new year.

Contour Slopes: Work Out 1 mm and 1 cm Contour Spacing

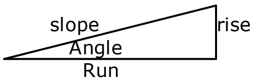

Example: a Rise of 10 m in a Run of 100 m is a Slope of one in 10 and an angle of 5.7°
[Tangent(angle)=rise/run   Angle=Arctan(rise/run)]
rise/run is technically called the "gradient" i.e. an up-slope.

1) Example of 1 mm contour spacing: A 1:100,000 map-scale says: 1mm equals 100m (See 'map scaling factors' p.92a)
with 50m contours   the slope must be 50m every 100m   or 1 in 2   or 26°½.
2) Example, with 1 cm contour spacing: 10 times less slope, in the same example, or one in 20, or 3°

Using these 2 reference spacings (**work them out again for every new map**) fit the actual contour spacing (eg 3mm) between them.
E.g. map scale: 1:25,000 (1mm = 25m); for 10m contours, 10m rise in the 1mm run, on the map (=25m in reality) is,
"1mm contour spacing = 1 in 2.5";   (10mm contour spacing) = 1 in 25;   (20mm contour spacing) = 1 in 50;
So…   3mm means 1 in 7.5

Reference Slopes — To Imagine the Climb with!

One in one; 45°; One up / one across; On all fours!
One Degree is 'one in 57' (or approximately 1 in 60)   So 10° is 'one in 5.7' — maximum for public **roads**
One in Four is 14° — a neat coincidence to memorise.   So One in 40 is 1.4° — maximum for **railways**
36° slope — 8 in 11 — more or less maximum for sand, gravel, scree, or any loose material.
Small slopes may be multiplied with approximate linear accuracy, but in an **inverse** fashion…
Examples: 2° is one in 30 — from the one in 60 rule
1 in eight is 7° — from the one in four rule
5° is one in 11 — either from 57/5 or from 5.7 x 2 or 1/8 of 36°   see also "sine-cos chart" p.106b

The Beaufort Wind Scale — Modified for Bushwalkers

One in 40	contour lines absent or rare; pleasant ambling; wind whistles through the lips; grey hair relatively common
One in 30	car engine quieter than road noise; perambulators plentiful; locomotives absent.
One in 20	car engine noise louder than road noise; walkers plentiful
One in 15	skin moist despite cold conditions; heavy breathing is heard above background noise
One in 10	eyes sting from sweat; bicycles experience difficulty staying upright
One in eight	flies are constant companions; car smoke exits parallel to the ground; mountain bikers walk
One in six	cigarette butts become less frequent; 2 wheel drives mostly absent
One in five	walkers experience difficulty making forward progress
One in four	cannot hear the sound of a cigarette being lit; leaf litter easily lifted off ground
One in three	sturdy walkers stagger and sway; frequent rests; slow progress; loose soil swept away
One in two	large trees tilt at a noticeable angle to the ground; rocks experience difficulty staying in place; walkers remain upright only by zigzagging
One in one	toenails begin to bleed; trail bikes begin to bleed
Two in one	fingernails begin to bleed; small trees are pulled up by their roots
Three in one	contour lines begin to merge; danger of severe falls; you are in danger of being completely… winded!

Walking an Exact Straight Line in Open Country

Assume the visibility is limited, or the terrain difficult, or the vegetation is closed in…

- Lengthen the Whole Party—In Single File. Shout to the leader, if he wanders to left or right. Stop still if you lose sight of the person behind you, i.e. do something to make sure that no one gets lost in the fog.
- Drag a flat object (or a person) on a long string. Tie your end of the string onto a stick, and hold it forward above your head. Alternatively, pivot the stick in your hand at its halfway point, and it will point in the right direction. (Watch that the wind doesn't blow it sideways; Stop the flat object rolling; Stop it sliding downhill; Don't let it drag at an angle)
Try just a slippery fishing line with no object.

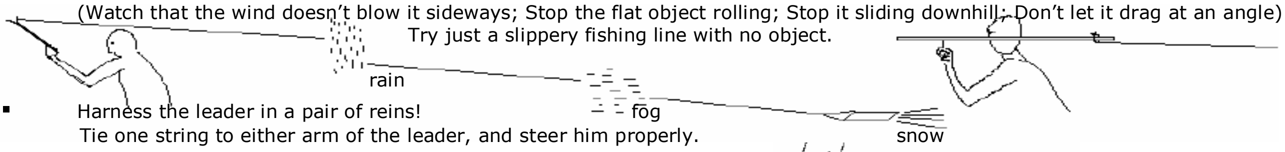

- Harness the leader in a pair of reins!
Tie one string to either arm of the leader, and steer him properly.
- Hold the rear end of the string over a compass.

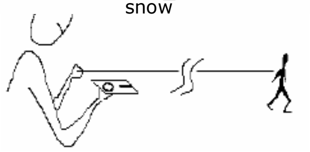

- Leave a trail of visible markers lined up exactly.
You need to look back frequently and drop a new marker.
— stones, sticks, stumps, tree trunks
Leapfrog with only three marker sticks — take them with you.
Leapfrog people!—leave someone behind; send someone ahead; line them up with each other & with a compass if available.
- Scratch a line into the dirt. Look back frequently. Readjust yourself
- Use the radio for radio direction finding.
- Orient by Sunset glow, the wind on your cheeks, the blue band, town lights, distant noise, etc.

To keep to a fixed "track", manoeuvre around obstacles in a rectangular fashion,   or reverse any diversion.

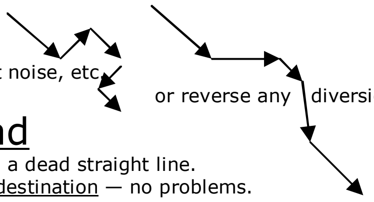

Leap-frog a Straight Line by Looking Ahead

A. A straight line course is easily achieved visually — like keeping a hundred fence posts in a dead straight line.
In clear open country, you simply walk toward a new mark halfway between you and your destination — no problems.
Keep the 2 lined up. When you get halfway choose another halfway point in line. If you can keep sight of the more distant destination, this method means you only use the magnetic compass once, to locate the distant destination—then keep a fixed 'track' visually.
When obstacles force you sideways off track, you can easily regain track, visually.
B. In more difficult circumstances, you cannot see far ahead and usually rely on looking back, as above.
But leaving markers behind you slows you down and keeps you looking around.
This method does not require you to look back, but requires a simple discipline — leapfrogging visible marks exactly ahead of you.
Step one: Choose two distinct marks exactly in line with your desired direction
Step two: **On reaching the first mark, choose a third, still in line. And so on.**
You can do this while following a formed track. The track will bend from the average straight line, but you'll notice that immediately.
I.e. when you don't get to the third mark, and have to start again, you are well aware of the change in trend.

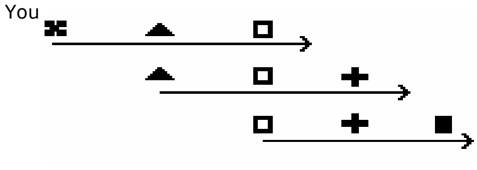

The end result accumulates errors
(despite the averaging process).

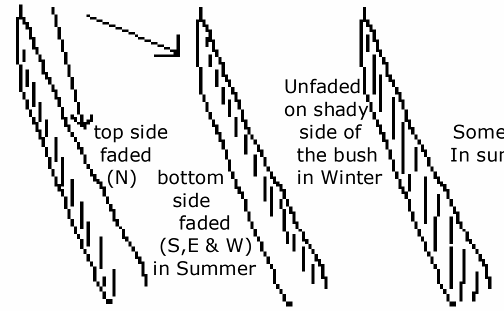

Noon-Sunlight is More Vertical

Examine needle bush (e.g. Hakea) needles, one by one.

Some effects will be more noticeable in winter, and in higher latitudes.
In summer, the 'shady' side is actually bathed in horizontal sunlight twice a day!

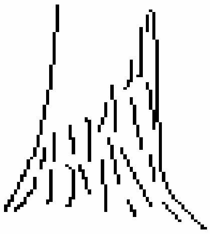

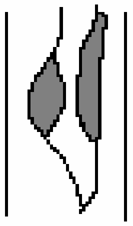

The Shedding of Bark on the Sun-Baked Side

The patchwork of coloured fresh new bark shows that
the bark strips an extra layer deep on the sunny side.
Look also at the coloration of the freshly exposed bark, for asymmetry.
The bark 'stocking' clings higher on the shady side.
It is older, deeper, darker and allows more and older lichens.
Look for the 'curl' of bark ribbons — the pipe-like shapes peeling off —
as there may be more curling shown up on the dry side,
but it depends on the species.

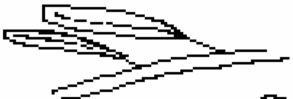

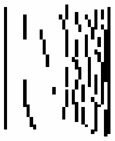

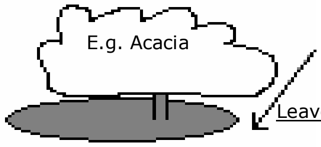

Where to Look for Fading or Lack of it

****Turn leaf blades around** to compare the greenness in each direction.
Find a pair facing the same direction, and twist one of them, then
make a comparison to see which side is more faded.
****Twist an *untwisted vertical reed stem***, to detect the more faded side.
Look at stumps and old posts, tree trunks and fallen sticks.
E.g. Sunlight bleaches the sunny side of bark and dries out any secretions
E.g. The dark side of the trunk will often be darkly stained
e.g. with cider-gum secretions.
Leaves dropped into the shady zone will not fade and dry out so quickly.
The shady zone is offset toward the pole,
because the sunlight penetrates deeper on the noon side.

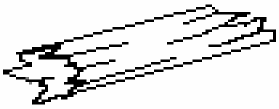

The broken faces of lying timber may be more coloured
and less bleached in the shade.

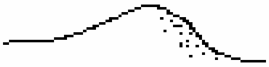

Bumps and slopes, diggings and burrows, may show dead dark moss on the
hot side, or grass which has dried up.

Look for a very subtle pink glow on the shady side of dead timber, which is absent or greyer and paler on the exposed side.
Once your brain sees a hint of it, it becomes easier to see—so persevere.

Bracken fronds die off from their tips, but the fronds that retain their greenness the longest are on the lowest, shady side.
Dead thistle stems fade on the Sunny side.

The topsides of grass blades fade more due to being blown over flat by strong **wind**s, and so are exposed to sunlight, wind & drying.
Upwind edges of green bushes often show a brown wind-scorch.

Fresh diggings (echidna, bandicoot, wombat) expose soil to bleaching. Look at the steepest sides for the least faded soil.
Some fading is due to **oxidization** (exposure to air), some to **dessication** (loss of humidity), some to sun-**bleaching** (exposure to ultra-violet). Again it will pay to study the various effects before you need to rely on them.
Also be aware that diffuse lichens will give a covering colour to dirt, wood, rocks, etc, which may complicate the appearances.

Button-grass flower-stalks fade but darken on their exposed side, perhaps due to fungus. One side is dark grey; the other pale yellow.

Make a Clinometer — a Slope Measurer

One person holds the loop knots apart under slight tension. He barely touches both plumb-bobs onto the ground, to settle them.
Their identical lengths ensure a parallelogram.
The second person alters the middle string until it just touches the plumb-line.
The identical lengths ensure an isosceles triangle.
The 5 cm marks represent whole degrees of slope. Use 6 m of non stretchy cord. Measure and constru .

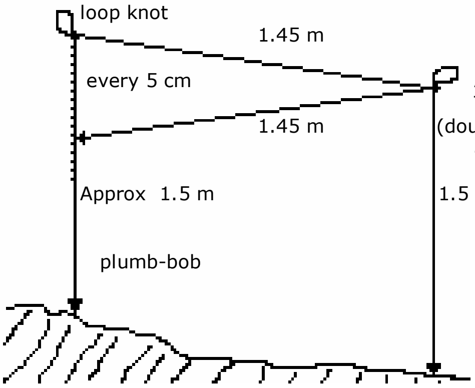

E.g. the slope is 11°½ here…

Accurate from 0° up to 20° or so
The method is based on the isosceles triangle rule
(double the half-angle sine; nearly linear for small angles)
and on 1 in 56. You can use 1.40 to 1.50,
but the lengths must be exactly equal.

The one-man version has two foot-loops.
Spread you feet and hold up the plumb bob.
See which side the plumb leans to.

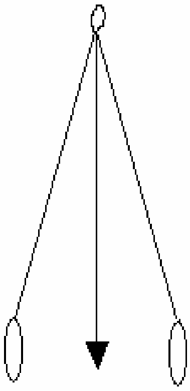

Use it occasionally, to refresh your judgment of slight slopes.
It may also be used to gauge the height of landmarks or stars. (Hang the second loop knot from a sky hook; Sight along the top string; Steady the bobs). Used horizontally, it yields horizontal angles. (Complete the string parallelogram; Sight along the two ‘top’ strings; and double the reading)
Of course this is only for a hobby, for interest’s sake, for mathematical curiosity, not for absolute accuracy!
The more you play, the more you learn.

Just Draw It

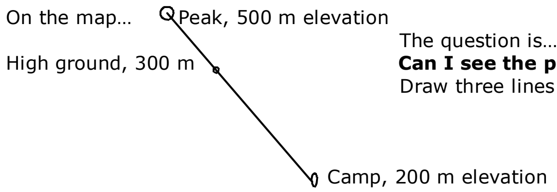

The question is…
Can I see the peak from camp?
Draw three lines of 5, 3, 2 units…

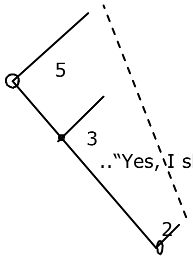

..“Yes, I should see over the ridge to the peak”

Radio Direction Finding — As Good As a Compass

AM stations — Amplitude Modulated — the older style, for medium distance transmissions
— also called Medium Wave; Medium Frequency — MW; MF

Your portable radio has an **internal aerial** for AM stations, which is direction sensitive.
It usually lies lengthwise inside the radio, as a long **ferrite rod** wrapped in a wire coil.
To make that ferrite rod horizontal, you probably need to lay the radio down flat.
Now twist the radio to change its horizontal direction.

When the internal ferrite rod points at the medium wave transmitter,
the reception drops off — you have detected a “null”.

This occurs at two opposite orientations — **pointing away from the transmitter as well**.
(Good reception is by pointing the radio 90° away from that null direction.)

FM stations — Frequency Modulated — the newer style, of shorter distance transmission
— Very high, Ultrahigh frequencies — VHF; UHF — It does not use a ferrite rod.
With a whip aerial, or an extensible aerial, or a long wire aerial,
you can detect a null by pointing the aerial at the transmitter.
This is for horizontally polarised transmissions. Some stations are polarised vertically.

Shortwave stations — Amplitude Modulated — higher frequencies, for longer distance transmission
— SW; High-Frequency, HF — It usually requires an external aerial.

Use a **long wire aerial**; keep the ferrite rod perpendicular to the long aerial.
The null occurs when the long aerial does not point to the transmitter — but it is perpendicular to that direction. (Opposite to FM)
Point the long-wire aerial perpendicular to the transmitter, for a ‘null’.

Weak AM stations can also be used, provided you have a long wire aerial to pull them in. The exact length sometimes matters—you will notice the volume rise and fall in cycles as you wind the aerial back in.

THE RADIO “NULL” ACTS AS A COMPASS NEEDLE — it keeps a constant angle to the distant transmitter tower or repeater.
You swing the radio aerial each side of the null to keep checking its direction.
To avoid having to give the radio constant attention, break your course into separate lines of sight (‘legs’) as with a compass.
Marine navigators know that radio signal direction can be distorted by mountains, or by the signal crossing the coast at an angle, or by reflections off the rigging, but bushwalkers seldom have to worry about that, because they only want a constant reference direction, not an exact bearing.

You could consider setting up a radio beacon so that you can always say: e.g. “The car is that way” or “Home-base is yonder”
End of Section 3.

Distance Guesstimation is a Major Problem for Walkers

Because distance measurement is a major problem, we find it easier to ignore.
We resort to some other method of navigation.
If we simply, out of laziness, give up on logging our pace & distance-already-come, we waste a sensible cross-check.

The first of these three sub-sections will look at the amount of distance you have covered — progress — distance made good…

- Elapsed distance — puts you on the map precisely — unless you are not very sure of your progress! or the map scale!
- Map scale — "It looked so easy on the map!" — Judging time, distance & difficulty is a question of reading the map-scale properly. "How far could we get by lunchtime?". Guessing at a curved distance is always a bit of a bluff. If you hear yourself saying "On the map that's **about** 5 km" stop and measure from the map-scale more accurately!
- Sideways drift — is a matter of trigonometry, when you discover an error in your direction heading
- Up/down progress — is in another dimension. You don't need an altimeter, but it helps. (If you do take a portable altimeter/barometer, make sure to test it to see whether or not it reacts falsely to temperature change—put it in the fridge then in a warm place; then don't forget to keep adjusting it according to the atmospheric pressure changes)

The second sub-section is about the rate of covering distance — pace — and distance-yet-to-go.

The third sub-section is about the distance you can see, but don't cover — estimating the 'range'. How Far to That Landmark?
E.g. You can identify a peak on the map — but only if you are certain enough of the distance from you to it.

Distance = Speed x Time

E.g. 4 km per hour x 3 hours = 12 km
Hint: Measure only elapsed, 'walking' time; Delegate someone to keep track of accumulated time out for stops and diversions.
Hint: Mentally bring forward your starting time e.g. "(as if) we started at 10a.m." (You started at 9, but had two, ½ hour stops)
E.g. foam takes ½ a second to pass your 10 m yacht; speed equals 10/.5 = 20/1= 20 m per second.
E.g. I seem to have drifted 12 km in one day = ½ kph current. E.g. The cloud shadows cover 5km in 2minutes = 150kph wind.

Keep a Log!

That's how mariners keep track—they write it down in the special book. Do you carry a notebook?

Adding Two Speeds

Convert speed to a length (e.g. 4 kph = 4 cm) and direction to an arrow.
Add the two arrows, tail to tip, in either order.
Complete the triangle or quadrangle tail-to-tip.

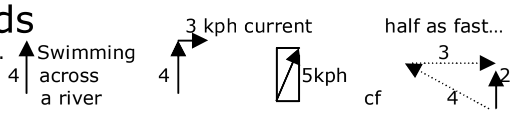

Note that it is faster to swim a river by pointing straight across and letting yourself be swept downstream, than by trying to stem the flow by pointing upstream. Only when the two speeds are perpendicular do they not affect each other e.g. the 4 km per hour is not affected by the current in the first example. Or e.g., If your boat speed is 4kph ahead, and the tidal flow is 5kph against you, your actual movement is backwards! (with any sideways component added in independently).

Example: Three air-speeds ↗ + wind ↘ = actual progress → very fast across the ground, slightly to your right.

I.e. Your speed & direction through the medium (air/water), relative to the ground + the medium's movement, over the ground (the medium carries you when it moves) = your total speed and direction, over the ground.
Lastly, *to analyse apparent wind speed, to avoid confusion, draw only the three winds involved*. I move this way...
First convert your progress over the ground (top arrow) to an opposing apparent wind (right-hand arrow opposing it).
Don't try to add wind to your motion—that's not a sensible question/answer. The medium (air) doesn't carry you when it moves. Answer: I feel the actual ground-wind, a cross-wind (→) in my face, diagonally as shown…

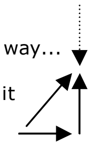

Range: Rule One: Guess

Then don't trust any calculation which doesn't agree with your initial guesstimation.
But remember that distance is deceptive over featureless terrain e.g. water/beach — we underestimate it.
Birds bob their heads to see how much an object shifts against foreground and background. It's worth some practice.

"Range" of Visibility — Two Halves to the Picture

"Extreme visual range" must be calculated in two bites — So Don't Forget To Add!
— e.g. when asking "Should we be able to see Ben Nevis from Brady's Peak?"
— e.g. at sea when you can first see the lighthouse lens, not just its loom.
From a kilometre high you can see 1° of latitude
before the curvature of the earth cuts your line of sight.
— e.g. you can see a 1 km high mountain from two degrees away
if you are on a 1 km high mountain — see diagram…

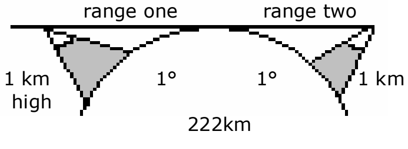

"Range Squared is Dee Height"

This is an easy mnemonic. "Dee" is dee Earth's Diameter i.e. Extreme Range = √(D x H) see p.92c
H is your height above the intervening horizon e.g. the wave tips, or a 250m plateau.
Since the Earth is 40,000 km around, divide by Pi to get **12,700 km diameter**.

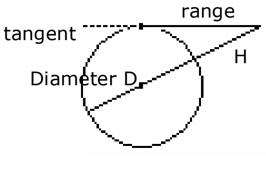

The formula is a theorem in geometry,
about the geometric mean (the tangent) of an extended diameter and the extension length (H).
The formula states: the square (range x range) = the rectangle (not shown: H x D+H).
(i.e. the range substitutes for both parts of the product, 'geometric-' or area-wise)
but since the height H, in reality, is very small, therefore **range^2 = D x H**

To “Read” a Map — Rule One: Don’t Lose Your Place!

Get the geometry fixed in your vision — of where you are on the map e.g. “in the top right grid square, next to that lake”.
And then don’t keep losing your place every time you look at the map again!
..Whether the map is now upside-down, side-on, or whatever.

Rule Two: Imagine Arrows, for the Sun, & Your Movement

Again, visualise them geometrically, in relation to the whole map shape, from one corner of the map, for the Sun, and from where you are on the map, for your movement. When your map is properly oriented, the movement arrow will point in front of you, ahead, matching the trend of your movement onwards.

Rule Three: Get Used To Any New Map Scale

Maps are “Fractal” They show “self-similarity across scale”
i.e. A river system will look the same on a small-scale or large-scale map — there is always finer detail to show.
A road network is the same — the smaller the scale, the more roads are left out, leaving a similar looking network.
A coastline shows such self-similarity. Magnify the details and it still looks like ‘a coastline’.

New Map-Scales are Hard to Adjust to
It takes time and testing to get used to a map scale, because they all look alike no matter what the scale.
To get used to the new ‘feel’ of a map, check up regularly on your map-progress, until you are no longer surprised by how far you have or haven’t come.

Rule Four: Don’t Give Up!

Don’t let the combination of difficulties bluff you!
E.g. You misjudge the scale, misjudge the Sun’s direction by optical illusion, you see turns & bumps not marked on the map, doubt the accuracy of the map, fail to assess the slope accurately,… and so eventually… give up in confusion!
Problems are normal, and **solving them by the map is normal too — don’t give up trying**.

—Also See “Map Checklist” p.113—

“L.M.N.Or P.”

L for Landscape; M for Map; N for Compass Needle: Or for Either/Or (the big point of the mnemonic)**; P for protractor**
Either/Or means: Each component of map&compass work can be used alone and independently for some tasks, if desired.
Plotting compass bearings can seem complicated. This mnemonic is designed to smooth the way.
A navigational compass has a protractor independent of its needle. The trick is to remember that!
The protractor has 2 parts too (p.93b). Note that some compasses have the magnet one-piece—glued onto—the protractor card.
If you get used to using each component separately, you will find ‘map and compass work’ to be simpler.
The trick is to isolate each of “LMN Or**P” in turn, to avoid confusion.** That includes the 2 moving parts of the protractor.
It is hard to use more than two at a time! so know which two you need to use at each stage.
E.g. A ‘prismatic’ compass will let you see both **L**and & numbers (**P**) at the same time—to take an accurate bearing e.g. at sea.
E.g. You can align the **M**ap to the **L**andscape, without ever using a number. Or you can point the **N**eedle to the horizon likewise.
E.g. The **P**rotractor may be used to assess slopes, to measure the Sun’s height, to take a map bearing, to draw an angle on the map, to take a horizontal angle between two landmarks, to sight a vertical angle like a sextant (this may need two people), to find North from Sun-time, or to set a course to the left of the Sun.
So don’t forget to use the **P**rotractor to its full potential, quite independently of that funny little piece of metal rattling around inside it. (**N**).
E.g. An emergency ‘compass’ is just a floating sewing **N**eedle, or a suspended magnet. So the protractor has little to do with orienting the map by the magnetic field, or with heading North by the compass needle.

If your navigational compass has a magnifying glass, a set-square and a graduated ruler, you should hardly get confused by thinking they had something to do with the magnetic field! — so don’t worry when you see some numbers marked around your compass needle — I think you learnt how to use a protractor in elementary school, didn’t you?

Working from land to compass is called “Taking a Bearing” (L+P1,N+P2) You use the protractor to put a number on the direction
..from map to compass is called “Getting a Heading”. (M+P). You then use the needle (P+N) to find then follow that heading (L).
Working from compass to map is called “Plotting a Fix”. (P+M). I.e. You use the protractor, ruler and set square to draw angles.

Maps Mean: Imagine Me! — Making Mental Maps

The meaning of using a map is that **it lines up**
— with what you see; with what you saw; with your mental maps; with what you know; with what you can’t see & don’t know.
E.g. You glimpsed the coastline as you drove in, the lie of the valleys, the drainage of the creeks, and of course you know the highway that you drove in on! You can no longer see them, but it is normal to orient yourself by imagining the invisible.
And then, your mental map is still incomplete, so that you need to use the map to add to your imagination…

“Map To Mind” Navigation

Transfer from the map both visible and invisible features back into your mental model and mental layout.
1. Start with the Invisible Big Pictures — visible on the map but not to you.
Look at the sky and the skyline as you interpret the map for what lies out of sight.
2. Add in the local visible features — visible both on the map and to you.

These two steps work together like lock and key — each is not very useful without the other.
Yet both steps are often neglected and seldom put together as a whole.
“The cars are now behind that notch in the skyline”
“The highway runs behind that hill then turns South-East”
“That low spot marks the river we want”
“Our camp tonight is in line between Mount Socks and the pine plantation yonder”
When you become temporarily bushed in, or blinkered by the rain, or shut in by the dark — then you see whether you can maintain proper orientation — by the invisible — or whether you just feel lost — “I can’t see where I am!”.
If you do not make a mental map and orient yourself by imagining the invisible—you will be disoriented,
even if you “can see where you are”.

♦ Part of the job is to look at the map in order to join your several local mental maps into one regional map.
Remember that a map is a great memory aid, or a great memory bypass — take your choice.

♦ Rather than “put yourself on the map”, put the map in you!

Use the Map Scale — Don't Just Look at It!

Many errors of judgment result from guessing at the map distance, even when the map scale is clearly marked.
The scale is not much good down in the bottom corner of the map, fixed in place! So copy the map scale onto cardboard.
Somehow attach it to the map e.g. into a separate paper-pocket; onto a string; in the map bag.
The string is useful for measuring distance accurately. You place it over a curved path then stretch it out straight to measure it.
Point the corner of the cardboard at the starting position; position the string from there; then stretch the result along the scale.

The Jargon

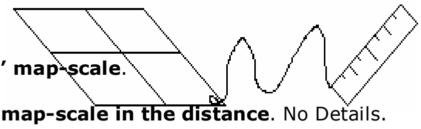

"Large-Scale" is like a "Full-Scale Model" It shows all the details.
Like a 'large-scale simulation' — it leaves out nothing. **Picture a magnified 'Large' map-scale**.
"Small-Scale" is like a "Small-Scale Model". You can't see the small details.
It is like a 'small pilot study' — not full scale. **Imagine a very hard to read small map-scale in the distance**. No Details.
"Full-Scale" is "1 to 1" like the Cosmos. "Half-Scale" is "1 to 2". *"Small-Scale" is* "1 to a million" *like Cosmological diagrams.*
The bigger the second number, the smaller the fraction. The bigger the big number, the bigger the span on the map.

Map-Scaling-Factors are Meant to be Interpreted

1:100,000 means "One unit represents 100,000 units"
Split this into two parts: (a) 100 & (b) 1000 and **use centimetres i.e. 1/100 m** — 1 cm: 100,000 cm
Say (a) 1 cm represents 100 cm i.e. 1 m — **drop off two zeros and call centimetres 'metres'**
And (b) 1 m, from part a, represents 1000 m i.e. 1 km — **drop off another three zeros and call metres 'kilometres'**
So the answer is: "1 cm to 1 km"
1: 1,000,000 1 mm = 1 km I.e. **You can also drop off three (or 6) zeros and call millimetres 'metres' (or km)**
50,000 units makes you calculate '1/50th of a metre' (=2cm), so 1:50,000 must be '2cm to 1km'. Or multiply to 2:100,000.
1: 250,000 Turn this into 4:1,000,000 by multiplying both sides by four. 4mm to a km (since 1mm is 1/1000m)
or (a) ÷100: cm to m (i.e. 4cm to 10,000m no longer cm to cm) (b) ÷10: cm to lots of 10m (c) ÷1000: cm to 10km
the grid should be marked in **10 km squares, of 4 cm each**
Your compass will be marked in centimetres, but perhaps also in special scales e.g. 1 inch to a mile (1: 63,360)

Pace Your Progress — Pre-Dict It, Don't Post-Dict It!

To avoid looking at the map continually, to locate yourself continually, predict the estimated time of arrival **once**, and relax.
We are normally lazy about looking at our watch often, poor at calculating, on the hop, distance covered, and lazy about locating ourselves on the map often. Mental arithmetic is clearly noxious, so we try to avoid it. Working abstractly from time to distance to map, after the event is too hard for us. But we usually only resort to the watch after we begin to doubt our position. (See "the 11:59 syndrome" — we do doubt.)
Something needs to be done about this!

Consequently, the antidote to us refusing to log our distance, is to anticipate the timing instead
i.e. to work more concretely, from map (I can see that) to distance (I can visualise that) to time (I can calculate that).

- Examine the map beforehand — as a first step not a last step. You will have to do the calculation later anyway, on the trot! Why not do it now, calmly.
- Guesstimate the time needed to cover the distance ahead — to the next recognisable milestone. E.g. use 4kph, ¼hr/km
- Look at your watch, as the last step, when you near your goal.

In this way the **landscape 'means' time** to you; "When we reach the bridge, we turn left — and that should be in 25 minutes"
Your guesswork will get better too.
You will find yourself looking at your watch more frequently, with the question: "Should I be there yet?".

Anticipation Is Exciting. Back-Calculation is as exciting as the washing-up.

"Range Squared is: Dee Height" "Dee" is Diameter

i.e. Extreme Range of Visibility = √(D x H) on this side of the horizon. **Don't forget to add the 'beyond the horizon' range**.
Five round-number rules are highlighted. By playing around with some approximations, we can find some rules of thumb…
127 is close to 128 and 125 and 11.3^2; Pi (3.14) is close to √10 (3.16);… I hope these tricks will make the square roots easier

40,000/pi x 40/pi km =	**400** km range from 12.7 km	Notice the two related doubling sequences I have given.
40,000/pi x pi km high =	**200** km range from 3.14km	You can only see twice as far from four times the height
12,500 km x 4/5 km =	**100** km range from 800 m high	125 is five cubed
113^2 x 9^2 / 10^2 =	**100** km range from 810m	34 km from 90 m; 11.2 km from 10 m — 1/3 from 1/9 height
12,500 km x 1/5 km =	**50** km range from 200 m high	"Rule of (lesser) fives: **50 km from 1/5 km**";
12,500 km x 1/20 km =	**25** km range from 50 m high	Multiply or divide H by four to double or halve the range
	12½km range from 12½m	e.g. from the top of a beach

12,700kmx.0127km=**12.7km from 12.7m—**from a ship's deck, to the horizon, when estimating distances across open water.

12,700 km x 1.27 km =	**127** km range from 1.27 km high	1/10 from 1/100. Remember the number 127 must be 400/Pi;
100,000/8 x 10/8 km =	**125** km range from 1.25 km	0.125 is 1/8

12,800 km x 1/2 km =	**80** km range from 500 m high or	"Rule of (greater) fives: **50 miles from 500m**"
12,800 km x 1/2^3 km =	**40** km range from 125 m high	You can also derive 160 km from 2 km; 800 km from 50 km
40,000/√10 x 0.4/√10 =	**40** km range from 127 m high	= "400km from 12.7 km" = 1/10 range from 1/100 the height.
12,800 km x 1/2^5 km =	**20** km range from 31 m high	= 200 from 3.14 above.
12,800 km x 1/2^7 =	**10** km range from 7.8 m high	..if you bother to halve 1000m seven times!
12,800 km x 1/500 =	**5 km range from 2m high**	**From a dinghy/beach**. 1/5 range from 1/25 height cf 50m high.

21,600 nautical ml/pi =	**150** nautical miles from Pi n.ml high	approximately, because 150 squared is 22500.
83^2 n.ml x 1 n.ml. =	83 n.ml. **150km from 1 n.ml.**	60 x 360 / Pi is approximately 83 squared nautical miles
12,300 km x 1 km =	**111 km** range from 1 km height	**1° of latitude, from 1 km high**, approx. "**55** from 1/4" = 250m
111^2 km x 1/9 km	**37 km** range from 111 m high	1/3 of the range from 1/9 of the height

Look Carefully at the Fine Detail on the Map

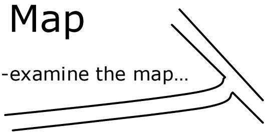

E.g. a kink in the track, just before a junction, may make it look "not like the map!" until… you re-examine the map…
"Oh yes, it is marked like that after all, **I just didn't look closely enough**"
E.g. that maze of fine blue watercourses marked — don't be a-maze-d by it — look at its detail!
There is usually a magnifying glass on the compass.
For a stronger **magnifying glass**, reverse your binoculars. For an even stronger magnifier, remove the eyepiece and use it.

Look Carefully at the Fine Print

E.g. date of compilation ..1864?
E.g. "adjust all longitudes by + 2 minutes of arc"
E.g. yearly magnetic drift
E.g. Contour interval — 10 m or 100 m?

Some Map Details are Not to Scale!

E.g. Two wheel ruts might look like a 100 m wide road on a map (if you took it seriously).
Consequently… 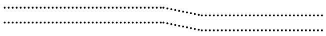 ..may be a 100 m Zig-zag! And…

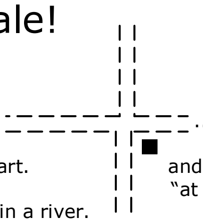

..is not a cross-road and the hut is not "at the junction"

E.g. A "cascade" symbol… ..might equally mean two huge waterfalls, 300 m apart.
E.g. 'Smooth' bends ..probably hide many small extra sinuous ones. E.g. in a river.
Moral: Keep an Open Mind. **Keep your eyes open** for alternative interpretations.

Using the Protractor by Itself

The protractor has two sets of lines — the lines on the movable 0-360 protractor-circle, and the fixed "follow me" arrow, taken together with the edges. Line up one set with horizontal, or North, or whatever; line the other up with the slope or angle to be measured. The angle in degrees clockwise is the space between the zero and the "follow me" arrow, not the other way around.

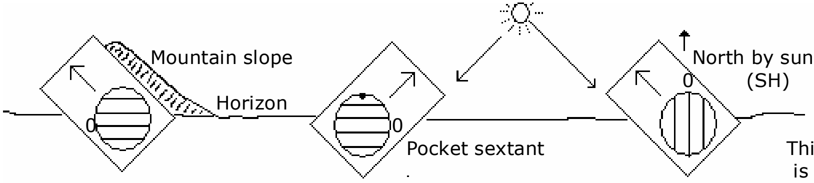

When the angle of the Sun (it moves around at 15°/hr) corresponds with the time of day, you must be facing North/Polaris.
This may be used when the magnetic field is locally anomalous or untrustworthy.
(Tilt the compass axis to Polaris)
Reverse this to get the time from North.

Point the edge of the compass body directly at the Sun until it casts no shadow. (Don't look at the Sun)
If your compass fluid has a bubble in it, it can be used as an artificial horizon.

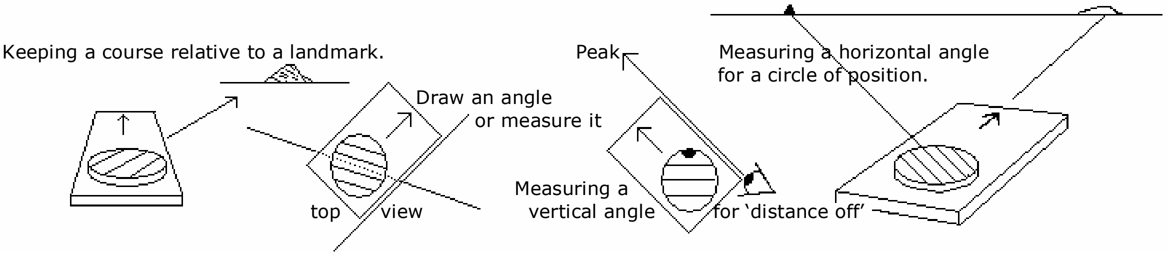
Keeping a course relative to a landmark.
Draw an angle
or measure it
top
view
Peak
Measuring a
vertical angle
Measuring a horizontal angle
for a circle of position.
for 'distance off'

Mental Networks

Our mental maps seem to run on networking — junctions and joins; choices and destinations; routes and 'directions'
Consequently mental maps are prone to logical-type blunders
- Missed junctions e.g. of creeks or paths
- Misidentifications e.g. of segments, features or forks
- Poorly judged measurements of lengths or direction e.g. hairpin bends which aren't noticed
- Mistakes of number

If you reread the section "Logical Blunders" (p.39b) you will notice that most can be seen as a **wrongly imagined network**.
That's one more reason to study the real map, in order to bring your mental map up to speed.
When we function on short term memory of mental branching directions, any shock or daze can be crucially disorienting.

The Most Valued Map

..Is one which you have made yourself. A do-it-yourself job makes it "yours". Practise with a MUDMAP, p46e.

When you copy important detail from other maps onto your map e.g. peak elevations; track notes; waterholes, then your map becomes more valuable than a replacement map.
You add extra details e.g. campsites; walking times; side-tracks & old tracks and suddenly you prefer the old dog-eared version to a brand-new map!
Sometimes it is even better to start with a blank sheet of paper.
But since it does you no good to pinpoint your position on a blank map, unless you are at sea, you proceed to add in the important local details so that you can tell where you are…

To copy a map: **Copy the Grid First**. ..if there is one. Add in many diagonals as guidelines.
The next rule is: **Get the Landmarks in their correct relative position first** **i.e. work top-down.**

The Psychological 'Half-Way Point'

You have to judge and juggle several factors before you decide that you are 'half-way'.

E.g. The "Unknown Territory" Scare-Factor (Ambition versus Safety Margin). In planning your time-budget, you can't relax in a one-way journey through unknown territory, until you know you have a reasonable safety margin in hand

So 2/3 the distance along, might be 'half-way through the test'. Other examples...

The "Point of No Return" — You tend to speed up, once committed to the next goal.
The "Downhill Run" — It is the quicker half.
The "Back-Track Return" — tends to be quicker i.e. the turnaround point can be after lunch.
The Morning Spurt-Afternoon Hurt Syndrome — puts 'halfway' before lunch.
The Side-Track Scheduling — if you don't explore side tracks on the way in, when you are fresh, you may not want to tackle them on the return journey.
The "My Feedbag is That Way" Home-run spurt on the last day.

When You Meet Other Walkers

"When did you leave this morning?" = how far to your next camp.
"Anything about the track that we should know?" E.g. slow sections
"Any deep water crossings" or other hold-ups

Other questions, less related to walking progress and pace, are…

Anyone else ahead of us? (Will there be room for us all?)
Do you carry a track guidebook? (Can I read it?) Do you know the forecast?
Any good water available on the way? Can you recommend a camp site?
Did you pass our tents? (Are they still there!)

Pluck up your courage to ask what you want to know, and be prepared to share what you know.

Advice about Walking Times — Double & Halve!

E.g. Brochures seem to add "return" or "one way" by guesswork sometimes.
So "6 hrs" means 3 to 12 hours until proven otherwise.
E.g. Fit young rangers saying "six hours" may mean "for me; double it for you" or "for you; halve it for me"
E.g. "walking time" may mean "excluding stops, rests and meals" — you could find yourself pushed for time.
E.g. "Not far, you can't miss it, it's easy all the way" could prove disastrous. E.g. ('provided you find the right track')

20 Minutes per Kilometre — Double & Halve — Easygoing Maths

It means **3 km per hour** —an easy-going overall walking pace when you include short rests, difficult vegetation (within reason) and lack of any hurry — tourists on track, walkers off track.

Obviously you need to multiply and divide by 20 to convert kilometres to minutes, or minutes to kilometres.

If the number is small — kilometres — multiply it up to the larger number — minutes — **double** it and add a zero.
If the number is big — minutes — divide it down to the smaller number — kilometres — **halve** it and subtract a zero.
You could also express it as a "half a kilometre every 10 minutes" but the "20" makes the maths easy;
And even simpler, to cope with the zero, express yourself in 'lots of 10 minutes'.
E.g. 5½ kilometres will take about… (double it, to get many minutes)… 11 **lots of 10 minutes**
E.g. "We've been going 50 minutes — (five lots of 10 minutes) — ; that's about… (halve it, to get a few km)… 2½ kilometres"

4 km per hour — Normal Walking Pace — Double and Halve!

You can easily double or halve your pace. Start with 4kph as a sustainable rate of progress, and use the doublings 1,2,4,8…
8 km per hour is someone tall walking in a hurry. 2 km per hour is someone small dawdling. 1kph is for very difficult going.
Coincidentally 4 km per hour is 10 miles in four hours or 2½ miles per hour.
Incidentally also, at 4 kph, a landmark 15 km on your left will change direction at 15° per hour as you walk past it.

Double, the Acute Angle, off Your Nose — For the Distance Off

This double-angle method puts numbers on a very "natural" method of gauging distance…
In the first diagrams you walk from the bottom left to the top, and watch a landmark, ahead on the right, pass you by.
The distance walked = the distance off if you walk far enough to double the original angle (away from straight ahead).
The initial angle must be acute. You need to maintain a straight course throughout.
The simplest case uses the diagonal of a square. 45° changes into 90°
The next simplest example is the equilateral triangle. 60° changes to 120°
Followed by the 30° isosceles triangle. 30° changes to 60°

But the general case is made clear by the geometry of two equal radii in a circle.
It is because the angle at the centre of a circle is double the angle at the circumference.
The angle at the centre is the final angle—double the initial one. top view…

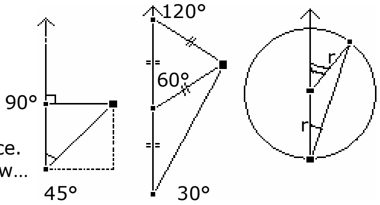

Ahead of you, the angle is always less than 90°.
In the second diagrams, below, the landmark may not be ahead of you and the initial angle may not be acute...

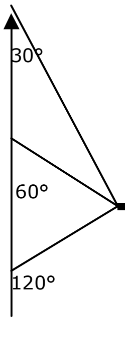

If you walk backwards, from top to bottom, using the top method in reverse, you get...

Halve, Any Angle, from your Stern—For Distance Off,

..i.e. In Retrospect The initial distance off, not the final distance off is what must be travelled.
The initial angle is now the double angle, and need not be acute.
I.e. turn the bow-angle diagrams upside-down…
E.g. 90° becomes 45°; or 120° becomes 60° becomes 30°…

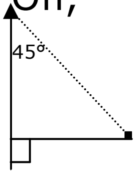

Permanent Landscape is More Trustworthy Than Surface Detail!

When what you see "doesn't match what is on the map" have the sense to distinguish between 100 m deep contours which would take an atom bomb to shift, and walking-track junctions which a fire could reshape.
Professional map-makers and satellite photos rarely make mistakes — in landscape positioning.
But when they add the local details, it is another story.
Time changes such details also. New roads, cleared bushland, second bridges — these will mislead and confuse you.
When your mind says: "The map is wrong!" remember that what the map says — about landscape — is not wrong.

Details Added on to a Map Are Prone to Error

Notorious for errors are: Map North; Map Scale; Roads.
So beware of anything other than a good-quality official government map.

Expect The Map To Be Unhelpful

Many important details will not be shown on the map — you just have to expect that.
You can often trust what is on the map; You can often trust that what you want, won't be on the map.

Maps Do Give You 10 Times More Detail Than You Can Notice

After you get lost and finally sort it all out on the map, you notice… "Oh yes, silly me, those contours are **up**slope not downslope! And they **do** show a creek-line, by those sharp bends! Oh, of course, there is **a peak** on the right, and that fenceline opposite — I missed both of those when I first studied the map. You **can** see Mount Bruce, lined up with the lake, from there, but not at all from over there. And that squiggle in the blue line, is the hair-pin bend that tricked me, but I should have realised that all rivers must drain to the north-west eventually…"

Try to reduce your ten-fold folly to five-fold folly for next time!

Using the Needle Alone — Emergency Magnetic Compass

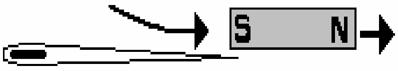

- Magnetise your repair-kit sewing needle; It will probably already be magnetised somewhat. Stroke it with a South Pole toward the sharp end, so that the North-seeking pole is at the pointy end.
- Float the needle in calm water, on the surface tension (no detergent!)
  (a) lay the needle on toilet paper
  (b) lay both on a bowl of clean water
  The paper will wet and sink, leaving the needle afloat. You could also try sticking it through a piece of foam.

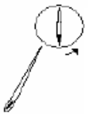

- Hang any magnet horizontally, from hair or fine thread. It helps if you know which end is North-seeking.
  In the field, test it against a Natural North, eg, Bush Noon. At home, test it on a compass. Like poles repel…

Hint

Understand that **magnetic poles should be called 'North-Seeking' and 'South-Seeking'** and that **our "North Pole" is a South-seeking pole!**
(so that North-Pole seeking poles point to it — Unlike poles attract)

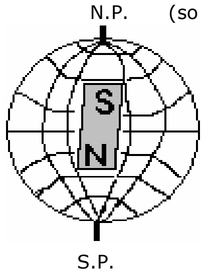

The Earth's internal magnet is as shown (It is slightly off-line with the rotation axis!)

A compass needle on the Earth's surface has a red end labeled 'N' i.e. North-Seeking

The white end is labeled 'S'

Strip Map Navigation — Follow-Your-Nose Navigation

Our mind definitely prefers a strip map — a sequence of "directions" to follow. (None-the-less, make a MUDMAP in conjunction)
Our eyes definitely prefer a strip map — "up" is "straight ahead"! (See "N.O.R.West/N.O.T.Least" "Nose Navigation" p31e)
A strip-map is simply a way of formalising your track notes—every navigator really ought to do it this way…
There are two good ways to make a strip map as you go: with a sheet of graph paper, working from bottom to top, and
in a track notebook, working from top to bottom, but…
Both assume that **progress arrows are up-page**. You could turn your notebook side-on or upside-down to write in it.
In either case you set aside several **vertical** lines for each **type** of information,
and set aside one **horizontal** line for each **subsection** of the journey,
and a second line for descriptions and for comments,
and a line for overall information e.g. trend, drainage, season; totals: time, climbing and distance.
The trick is to use symbols. One vertical line for one sort of symbol e.g. compass needle direction
(See the **Symbol appendix** p 122) Another line for another sort of symbol e.g. slope-direction, and steepness
This way you can cram a lot of information into a small space, in a consistent and regulated manner, and be able to read it!
e.g. …

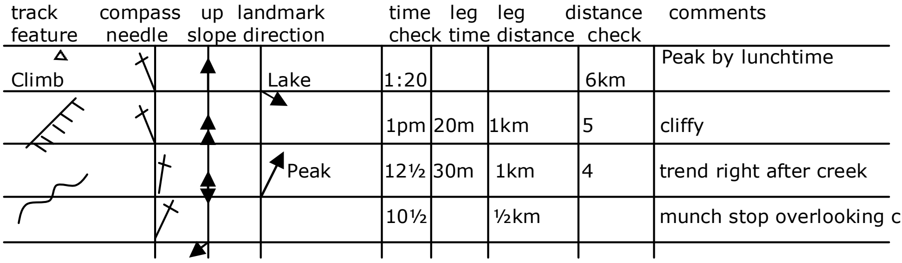

track feature	compass needle	up slope	landmark direction	time check	leg time	leg distance	distance check	comments
Climb			Lake	1:20			6km	Peak by lunchtime
				1pm	20m	1km	5	cliffy
			Peak	12½	30m	1km	4	trend right after creek
				10½		½km		munch stop overlooking creek

e.g...

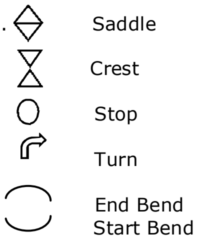

Distance

AIMING OFF by 10° — 10° is **175m per kilometre**

1° is one in 57.3, so one in 5.73 is about **10°**, (and 1 / 5.73 = .175) or **175 in 1000**. (For bigger angles use the 1 in 60 rule)
Alternatively, 2sin5° = 0.174 isosceles chord, yielding "0.174 per one" or 174 per 1000. (167m per 1000m / degree for >10°)
This is easier than working in sixes using "one in 60".

Hint: Express any ratio as "something in 1" or "in 10, 100, 1000"
E.g. '1 in 50' becomes '2 **in 100**' or '200 m in 10 km' (.4 in 20 etc)

Example: "The T-junction is somewhere ahead — about 017°, 5 km. But so that we can't miss it, we must aim off to the right e.g. 027°. We therefore expect to hit the road 875m to the right of the Junction" (5 lots of 175m).
This is how ocean navigators find an island in an ocean: Step 1: sail N/S to the right latitude. Step 2: sail E/W to the right long.
Don't just head straight toward something. If it doesn't show, you won't know which way to turn!
Remember that a 10° error is fairly easy to make when using a compass over rough terrain (or trackless shifting ocean)!

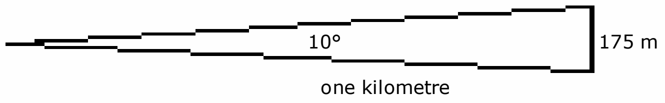

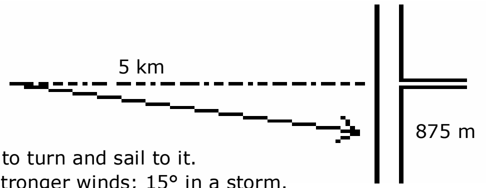

Aiming-off at Sea…

Aim-off to arrive up-current & up-wind of your destination—to make it easy to turn and sail to it.

- Aim-off 6° to counteract leeway when sailing into the wind; 10° in stronger winds; 15° in a storm. Measure the appropriate angle from the wake e.g. 6° is how much your pointing-finger-tip shifts when you change eyes
- Aim-off 6° (1 in 10) to counteract a 10% (1 in 10) cross-current. This refers to the perpendicular component, compared to your speed. The rule works +/−1° up to 36°—up to 60% cross-current. To avoid judging components...
- Aim-off 1°,2°,3°,4°,5°,6°, to counteract a 10% current at 10°,20°,30°,45°,60°,90° to course, either ahead or behind.
- Aim-off to avoid GPS-induced collisions caused by very accurate course-holding. You should plot a course off to one side of the obvious line on the chart, to avoid vessels coming the other way on the same line!

Aim Off 6° or 10%

Blink your eyes at the destination, and when you get there, pace off the 6° offset. Ten-tuple the result for the distance there.

1 Metre per Second is 3.6 kph

This is merely a logical reference point for your calculations. (And slightly faster than "20 minutes/kilometre" (3 km per hour))
60 minutes x 60 secs per min is **3600 seconds in one hour**, at 1 m per second, is 3600 m in one hour, or 3.6 km per hour.

One metre is Up to your belly button A child's double pace One strong fast stride

One second is the time it takes to pronounce "Alligator One" in a normal fast voice—don't let it drag. Follow it with "Parramatta two", "Oodnadatta three", (or "a thousand and four", but "alligator" keeps its speed more consistently when drowsy).
Get used to the rhythm of tapping your finger in time with the seconds — your mind subconsciously remembers the timing.
Take a Watch, Perhaps a Stopwatch
1 m per second is very nearly 2 knots.

"1 Hour Extra per 5 Hours"

You can normally walk at 5 km per hour, on a track, on a short walk.
But you will sustain an average 4 km per hour, when you include fatigue, and rest stops.

"1 Hour Extra per 500m easy climbing"

Gentle up-climbing is tiring—allow this extra time. (We are not concerned here with how long it takes to rock-climb!)
Down-slope walking may speed you up, or slow you down, depending on the slope.
Dense tussock country (e.g. snow grass; button grass) is the equivalent of a climb, in its up-down-up steps. Crossing the drainage can also be as tiring as climbing, in its regular climb-descend-climb-descend path.

Distance by Optical Rangefinder — Accurate up to a Kilometre or Three

We are woeful at guessing distance in that range, so there are focussing devices available which are designed for it.

Range by Parallax and a Piece of String — DIY Rangefinding

Animals shift their heads to see how much an object shifts. Blink you eyes to see the effect; walk sideways to practice it.
For more distant objects, the very slight sideways offset can be magnified then measured...
The landmark you are measuring appears in a slightly different direction from each of two different viewpoints.
Pack a long non-stretch string e.g. a polymer kite-string. Know its length. Mark it every 5 m before you go.
Use it as a giant pair of eyes—a long baseline—& to mark out a large square or rectangle on the ground, one side at a time.
Guess at the right-angles — they don't need to be accurate. Just make sure to use equal tension when you stretch the string.
But use a plumb-bob to mark the corners of the square on the ground accurately — with a pebble.
Use the plumb-line also for accurate sighting of the landmark. (You can pack up the string once you have placed the pebbles.)
Twice as big = four times as accurate (in measuring d) If you have enough room to move, aim for over 10 m sides.
Range = ab/d or a^2/d if a = b; from the similar triangles involved. Measure the shortfall, d, at the front, or the overfall, d, at the baseline.

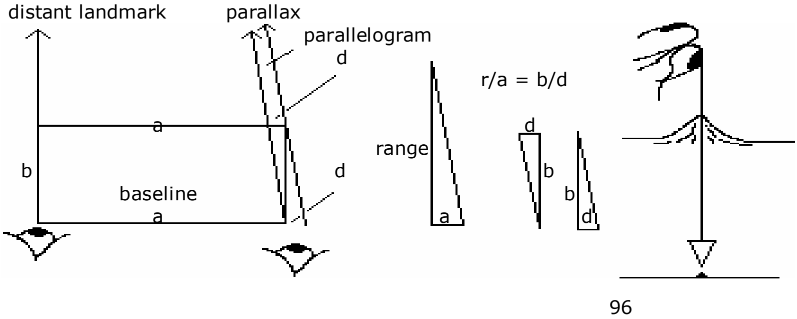

An easy way to remember the formula is to say that the area R d = the area a b because if you shear the ab rectangle, it becomes a long thin parallelogram R long, and d wide, preserving its area.

As a result, the formula becomes...
Range = Area / parallax offset, d
km sq. m mm

The Range=ab/d formula also *makes sense*
The range *should* so depend on a & b and d

"Maps Grip the Universe"

When you swing the map in your hands, around with you as you turn, the whole Universe doesn't follow it! — it doesn't twist!
Instead your map gets disoriented.
Leave the map behind, as you twist! — leave it lined up.
The map is designed to lock into the landscape, to 'map onto' it, and should be used like that, not twisted out of alignment.

You need to swing the map an *equal and opposite* amount to neutralise what you do.
It's called "swinging the map" (when you leave it locked onto true North while you swing to left and to right).

It should not be called that!

It should be called "**NOT Swinging the Map** with you"

Map work includes 'not swinging the map'. See "Magnets Grip Maps". The most common method is to match the map to a compass needle p.103b

Maps mean landmarks — those landmarks there! Open your eyes — and line the map up with the visible landscape. p101a.
i.e. orient the picture of the land — the map — to map onto the actual land.
Line the Grid North up with North e.g. by compass or by Sun.

You may have to twist yourself, or put 'North' at the bottom left, but to always put 'North' up the top, is to divorce your eyes!
Yes, you may have to read names upside down! Yes, it is a bother, to keep changing its 'up' every time you turn.
But it means that you can just lift your eyes from the map to see what it points out. And can know what lies beyond that.
If your map is not consciously oriented, how can you be oriented, to other than what you can see in front of your nose?

Think "+/- 5°" For Compass Bearings; "+/- 10°" For Following a Compass; "+/- 15°" For Using a Compass!

Compass bearings "+/– 5°"
Test yourselves: Ask each person in turn to nominate where "magnetic North" is on the skyline
using one particular compass, handheld. **My results:+/- 5°** **"User Inaccuracy"**
Now try putting that compass on the ground. **My results: +/- 3°** **"Needle Instability, friction" etc**
Try stretching a fishing line over the needle. **My results: +/- 1°** **"We still can't agree!"**
Next try that best method for different compasses. **My results: +/- 3°** **"Calibration and Zero Errors"**
Now try **not** being very careful, with an **untested** compass, at night, when you are tired!

Following a compass blind: Walk on a fixed compass bearing for 100 m or so in a fog i.e. eyes glued to the compass.
Repeat it e.g. turn around & see where you come back to. Give the next person a try at doing better.
I think it matters which eye you use and which hand you hold the compass in.
We don't necessarily walk 'straight ahead' when we try. We don't necessarily point the 'follow me' arrow straight ahead.
We don't do very well at integrating and averaging out any diversions around obstacles.
Slopes play havoc with your sense of a 'straight' line ahead.

Try this…

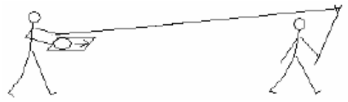

The front person finds the route.
The back person guides the route-finder.
The navigator logs the route directions taken.

Using a compass: You may plot a single line on the map, or as +/- 5°, but mentally allow an extra +/- 10°, i.e. think +/- 15° (for accumulated errors, possible blunders about magnetic variation, etc.). Keep the possibility of error in the back of your mind

---

Caving Strip-Maps — Come in Books

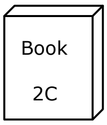

A "cave-directory" booklet records **each separate section of a cave on a separate page**.
E.g. Make your own…

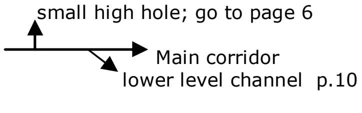

You may need separate booklets for different levels of the cave system.

Use the same symbols as before, for needle, slope, turns, etc.

Use Abbreviations: E1, E2, E3… for Exits; L1, L2, L3… for Light sources; Parentheses (L1) = "Can see L3 from here"

Make up names for each section e.g. "Chamber of Horrors"

Add copious notes e.g. "water level up to ½m deep"

Forethought and cooperation is required to quickly and effectively map a warren or a labyrinth.
Initially you make a muddy map, as you go, but then you add a neat copy into the book later.
(But keep your muddy copies, for next time.)

Up-down Progress

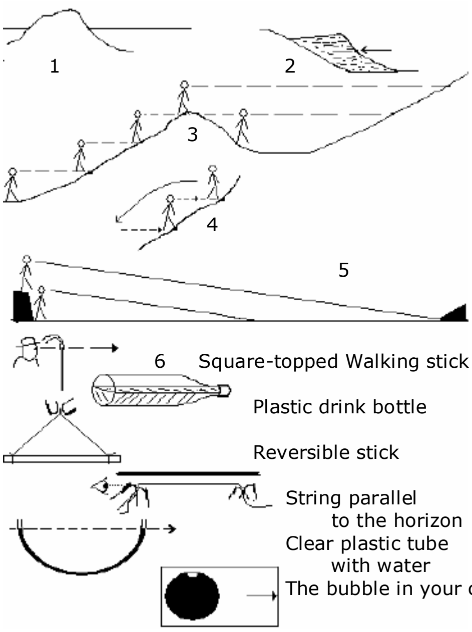
1
2
3
4
5
6 Square-topped Walking stick
Plastic drink bottle
Reversible stick
String parallel to the horizon
Clear plastic tube with water
The bubble in your compass

In the first diagram you can see that you are halfway up the mountain alongside of you — **use the horizon to gauge your altitude**.

Note: the horizon drops below true horizontal as you go higher.

In the second diagram your shadow is obviously halfway up the shadow of the hill you are on — **use the shadow of the mountainside.**

To find your tiny shadow:

Close one eye — to line up more accurately.
Look past the shadow of your head falling onto your hand
Look for a 'halo' — a bright spot directly down-Sun on the plain
Wave your raincoat around
Well, you know where the shadow should be by now!
So you don't even need to see your shadow!

Keep count of height in body-lengths

(3) Walk up to a stone which was horizontal from you before.
(4) Downhill can be counted by leapfrogging with two people.

Half-Way across? Provided you stay in one position...

Looking parallel from half the height gives you half the distance
(5) E.g. when someone is swimming across a lake

Artificial horizons allow you to maintain parallel gaze (6)

E.g. "The edge of the lake is 'two fingers low'

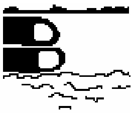

<horizon (or **any distant landmark**)

<lake edge (now, halve your height above water e.g. by squatting down)

Adjust Your Expected Pace

! Three hours up, two hours down — are typical proportions, except in very steep or loose descents — 6 hours up, 6 hrs down.
! You walk more briskly while fresh, in the morning, than after a whole morning's walk. Pace yourself with this in mind.
! Groups travel slower. Allow the slowest person to set the pace, because stragglers may battle or even panic if left behind.
! Retracing your steps is quicker, since it has fewer diversions, less interest, and no unknowns. Allocate less time to return.
! Carrying packs is not necessarily slower, on flats. If your pack is uncomfortable, you won't dawdle, or even want to sit down, and can't stop to look at insects or flowers or to chase birds.
! River crossings chew up enormous amounts of "fussing around" time — organising and reorganising afterwards.
! Beware of smoothed out "routes" marked on maps. The actual terrain may involve many obstacles and consequent slow going and detours. The words "route only" are trying to tell you that.
! Your pace quickens surprisingly when the party falls "into the groove" of walking (without talking, noticing things or relaxing often).
! Path finding can make you nervous especially in unfamiliar territory, which tends to quicken your pace without you knowing it. The same person, following a leader, is relaxed — and slows down.
So when you take the lead and you wonder why the others seem to fall behind, realise that you are too keen.

Don't Overcompensate for Slopes And Meanders

The worst case for slopes is about 36° for rock scree. But that adds only a quarter to the distance shown on the map. It is far more relevant to say "It's the climbing and the awkwardness which will slow us down more".
The worst case for obstacles is grass-tussock negotiation — but that only adds a maximum of 50% to your distance covered.
An ordinary zigzag may only add 10%

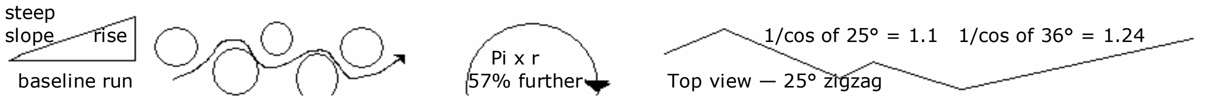

Distance by averaging Maximum & Minimum Guesses

Yes it is worth doing, because it reduces the slop in your individual guesses.
E.g. "More than a kilometre, for sure. Less than 3 km surely! Split the difference & say 2 km

Distance by Comparison with Some Length You Know

E.g. The central business district—1 square mile or whatever
E.g. old-fashioned 'chains' i.e. cricket pitch lengths
E.g. 100 m race; playing fields
E.g. "a day's walk"
E.g. that island which you can see, which the map says is 1 km long.

Distance by Cloud Shadows — Look for Them

- Notice how quickly the cloud shadows move here (use distance/time), then look how slowly they are moving over there
- Time how long they take to get there.
- See how large they look here, and how large there.
- Look how high the cloud base is here, then compare the size of hills in relation to that ground-to-cloud gap, over there.

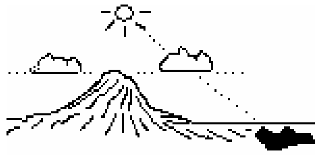

Hint: To see which shadow attaches to which cloud, hold a string, to pass from Sun to shadow (past its cloud) but… Don't look directly at the Sun!

Sound Travels Three Seconds to the Kilometre

This tells you how close the lightning storm is. "(Flash), Alligator One, Alligator Two, Alligator Three, Boom!, One Kilometre!"
If you see a flash to ground, you can find the distance there.
Otherwise, the distance is to the nearest part of the lightning flash overhead.

Align the Track

Line the map up with what you see! Often that is the road ahead, or a well-marked track.
This is a great trick for the car navigator when using a road map — it's so easy to see the line!
It's equally helpful when following any marked track on the map with the map aligned to the landscape — it locates you on the map — on that bit of the map which lines up with the track.

If you were landing a fighter jet onto an aircraft carrier in the fog, you would not want to miss the runway!
Well, think of the real track as the runway on the aircraft carrier, the land as the sea around it, & the tiny map-picture of the track as the jet plane.
Twist and turn your 'jet plane' until it can come in to land properly. Even when you are in the fog!

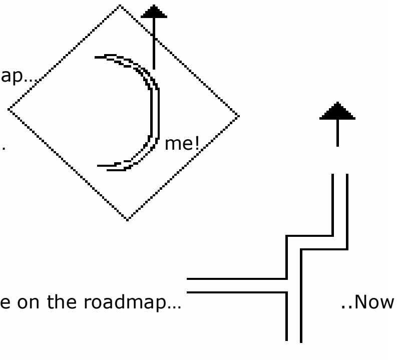

Assuming you have a properly oriented map…

You can find where you are on it…

If you know where you are on the roadmap…

..Now the map is properly oriented.

Compass Errors — Be Impressed! Be Careful!

Electrical currents swing the needle e.g. reading the compass by torchlight; untwisted pairs of DC wires e.g. marine compass lights; Mobile phones; GPS; radio. "Impressed (DC) current" anti-corrosion systems in marinas.

Magnets e.g. other compasses, motors, loudspeakers, clocks, ironstone. Note: Iron-rich dolerite doesn't look rusty.

Iron & Steel e.g. your pack, pocket knife, keys, wrist watch, cutlery, steel-capped boots, glasses, geology hammer, batteries, torch, studs, zips, pens, trig points, shed, fence, car, ironstone, buried machinery, coils of old fencing wire half-buried in the ground, tools, fuel stove, internal springs, reinforced concrete, railway lines, power pylons, pipelines. Navigating in a mineshaft.
Many of these might have permanent magnetism too. Stainless steel is often magnetic.

User error
Parallax — viewing the needle from one side; viewing the scale from side-on.
Misreading — e.g. any scale marked every 2° is easily confused.   Mis-setting the movable scale.
Miscalculations — especially applying the variation the wrong way — adding it instead of subtracting. Confusing left and right.
Out of date chart of deviation. Deviation is the general name for any compass error—apart from expected, normal variation.

Quality of Construction
Needle instability; under-damping; over-damping; motion. Let it settle. Put it on the ground.
Vibration — has unpredictable effects e.g. twisting the needle clockwise. Walking provides regular slow vibration! Motors.
Sticky pivot — Tap the compass   Un-centred pivot
Zero-error — the needle doesn't match the 'zero'   Mis-calibration of the protractor
Needle mis-magnetisation — not quite longitudinal; amateur attempts at re-magnetisation.
Reversed polarity — Poor quality needles can pick up remagnetisation from nearby strong magnets without you knowing it!

Variation The difference between True North and Magnetic North changes with time and place. The map is out of date.
Magnetic drift of the magnetic pole — we only have 'predictions' or 'trends' of an unpredictable variation. Are you up to date?
Latitude change — alters the magnetisation of surrounding fixed iron e.g. a car body or a boat's motor or hull.
Local anomalies e.g. Tasmania is full of ironstone dolerite. Many rocks preserve ancient magnetism.
External magnetic fields e.g. which way a boat moors, or who moors next to it. Competing fields change or capture each other.
Lightning strike after-affects.
Correcting mechanism out of **adjustment**.

Map Folding and Map Holding

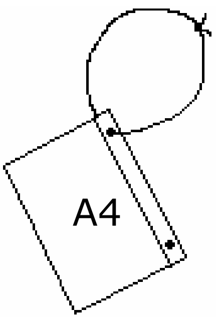

Maps aren't forever — they blow over waterfalls, tear & get unreadable.
Pack a spare
Cover the whole surface with (non-shrinking) clear plastic film
With good-quality laminating plastic & cutting, taping and sealing, maybe it will last forever!

Folds and Corners lose details, and wear through.
Cut along the fold lines cleanly e.g. to A4 size

Use a photocopy of your own map to use in the field, to keep your original map in good condition.
Wet maps lose details when rubbed.
Get an A4 plastic sheet protector for the map. Tape it closed. Hang it around your neck.

Mapwork requires bearing-lines to cross the whole map.
Tape the (A4 or other) sections back together with wide clear tape, leaving wide gaps, as shown.
— you can butt the edges closely when needed; the tape will crease out of the way
— you can refold it at will, as often as you like, and show the two most useful sides out of the many sections.

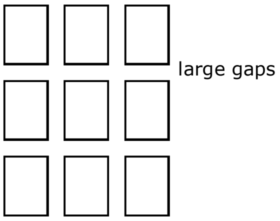

Distance by Double Pace — Test Yours

A 'mile' (or "thousand" in Latin) is a thousand double-paces, on forced march i.e. 5.28 feet (Romans were small people)
My dawdle double-pace is 5.25 feet. What is yours? In metres?
Answer: Test it. Test it again and again. Average it. Each person is different. Relax, don't push yourself.
Test it again uphill, downhill, laden, unladen, tired, in a hurry.
Test it at one double-pace per second — walking by the clock.
It helps to make a 10 m cord. Use a tape measure to make the cord, then use the cord.
E.g. I know I use 12.5 **paces (single steps) per 10 m** i.e. alternating 12 & 13 double places per 20 m.
and I walk at 4.5 kph; 1 1/4 metres per second; eight **seconds per 10 m.** The timing varies more easily than the distance
While watching the seconds tick over on my watch, and striding out in time with them, I manage 1.5m/sec, get tired, and step further out.
You Can Be As Accurate As You Have Been Careful. If you are full-grown, working this out once can set you up for life.

Tie Your Ankles Together!

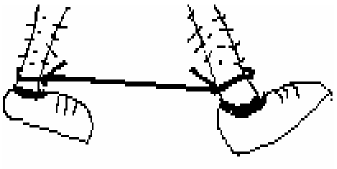

For a well-measured pace (heel to heel) use a 50 cm cord, for instance.
Count double-paces, and don't lose count — see 'finger math'

Test your pedometer

Pedometers work in different ways, so test yours on different terrain. Your knee-lift is higher on uphill sections, and your pace is shorter in rocky terrain. You can make your own pedometer from an old odometer, working off the flex of your boot, your hip, your walking stick, or whatever, or you can roll a small bicycle wheel along the ground, fitted with an odometer.

Don't Lose Count! Use Finger-Math

Finger counting is a silent and **number-free** counting method — you only pronounce the number after you have finished counting. Count automatically, by placing your fingers down, each in turn, onto a hard surface.
A thumb counts for five fingers, and the **left-hand counts 10 per finger**.
When you run out of digits you have 99, so you next say aloud "A hundred" and start over again with no fingers down.
Place each finger down onto a surface, in a **definite sequence**: Right Index ("1"); Middle (two fingers down); Ring (3); Little (4); Lift all four and put down the thumb ("5"); Leave the thumb down and put and leave down the index finger (6); middle (7); ring (8); little (9); Lift all five and put down the left index finger ("10"); Then put down the right index finger again (11), and so on, until you need to put down the left middle finger in place of the other five (two 'tens' down).
With practice, you can count, add, subtract, multiply and divide — almost thoughtlessly — as quickly as with an abacus.
We are concerned here with **counting single or double paces**, e.g. in a fog, and doing so **in lots of 10m** (or 20m or 40m for double paces). I.e. you count up to about 12 or 25 single or double paces mentally, and add "one more lot" to your finger tally.
If you are in a group, you can check each other's counting by stopping where you think every hundred metres is.

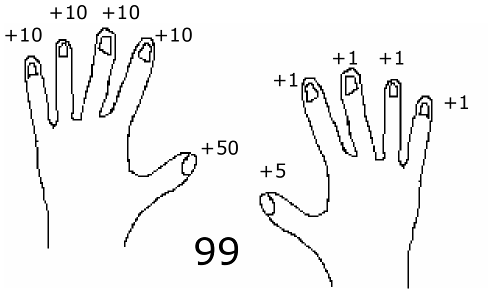

Lift all your fingers off
and nod your head down
for '100', then start again.

100 lots of 10m equals 1km
100 lots of 20 m equals 2 km
100 lots of 40m equals 4 km

or 100 double paces is 1/10 of your mile

Lighting Conditions & Eyesight

Make allowances for optical effects…
Looking up-Sun, especially in the afternoon, you see haze (backlit dust & insects) which gives a false impression of 'distance'.
Looking down-Sun on a clear morning after rain or after a cold change (i.e. minimal haze) things 'look close enough to touch'.
As your eyesight begins to fail in mid-life, everything seems 'distant' but turns out to be surprisingly close
i.e. you have been judging distance by crispness of detail all your life…

Visual Acuity is about One Minute of Arc

..which means that you can spot a 1 m object at 3.6 km on a clear day (since one minute of arc is (1/60)° or '1 in 3600').
I.e. You can see details several thousand times smaller than the distance to them.

This is how seamen can judge distances over open water, to land, even though there are no other clues to go by.
They ask: "Can I see trees? What about tree trunks? Or branches? Or leaves!"

Lighting, Contrast, Haze and Eyesight play havoc with accuracy; But experience is a great teacher.
Feed your mind and memory by deliberately looking through the haze to the finest details you can pick up.
Use binoculars as a cross-check. Have competitions with your family.
Children have better eyesight, but the idea is to judge the distance according to how it looks to you.

Align the Map to the Landmarks

Make the Map Match the Marks — if you know where you are on the map you need only one landmark…

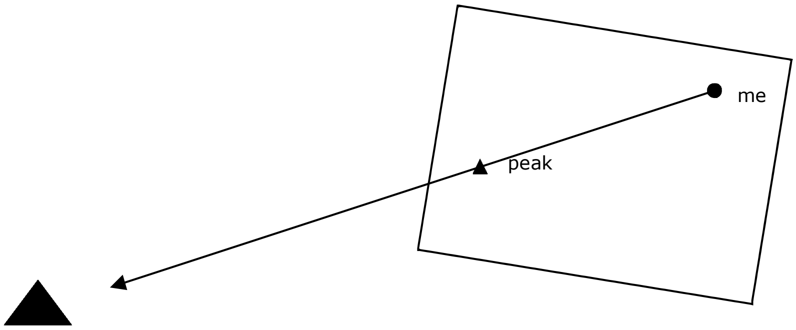

The easiest way to keep the map aligned to the landscape is by then using the direction of the Sun now, a very visual method.
The commonest method uses a magnetic compass to find North.
This one-landmark method can be used whenever you can see something in the distance.

If you don't know where you are on the map you can still orient it to match the landscape, but you will need two or more landmarks identified to do this.

Compass Use — Master It, but Be a Good Master.

- Use it to orient yourself, then put it away!
- Use it to orient the map or to choose a route, then put it away!
- Use it as a protractor. Use it as a ruler. Use it as a map scale measurer. Use it as an artificial horizon. As a set-square.
- Use the set square for a back-course or a side-course without altering the forward-set-course.
- Don't rely on it — use other methods of direction-holding.
- Let it point your eyes outwards, to the horizon.
- Use NORWEST/NOTLEAST, if you have to stare at it — e.g. in a cave.
- Hold it 1½m above rocks; 10m away from cars, sheds, etc; several item 'lengths' away from small items — test it.
- Let it point out a foreground mark as a destination, then follow a visual line of sight
  e.g. 'leapfrog' or manoeuvre by 'legs'
- Don't consult it again until you need to check or choose another point.
- Don't trust it, but trust yourself to it, when you need to.
- CHECK ITS ERRORS. Check it against the map by taking a bearing from a known point to a distant landmark
- Take a pair of forward and reverse bearings—to and from a chosen mark, to check for local magnetic field anomalies.
- Check your errors in using it, e.g. on the move; against someone else.
- Interpret it for orientation value, not just for route-finding.
- Consult it more often than you want to, when you are 'not using it'.
- Check your guesswork as to e.g. 'Gully North' or 'Wind-West' by it.
- Take it.

Layout Map — Blind

- This project requires a pedometer to measure every hundred, metres or paces or something. A stop-watch could do.
- Use graph paper, so that each square on the graph paper represents one 'leg' of 100 (or '1 minute's walking').
  (Hint: Graph paper is a handy thing when making a casual mud-map of your progress, for this very reason.)

Try it at least once. Without triangulation and cross-checks you can build up the layout just by reading the compass.
i.e. Beware: This is a bottom-up method, prone to inaccuracy. After one accurate practice-run, you can do a better casual job.
I achieved over 90% efficiency at coming back to my starting point after a 24 km circuit on my first try!
(In fact I know that the 8% error was mainly only due to a few blunders in marking the wrong direction.
Also, I only measured 'direction' to +/–15° since I was also testing the 'clockface method' of direction — to the nearest ½hr.)

When the compass read "2:30", I needed a reverse clockface on the map, and I needed to mark…

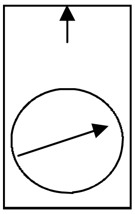

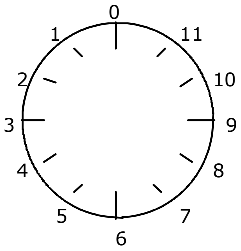

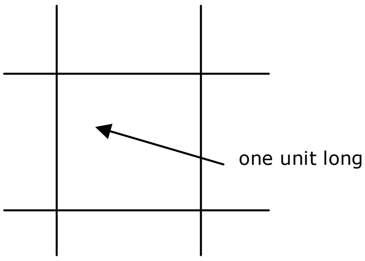

(Hence I made a few left-right blunders which I had no record of!)
You don't need a map for position-sense, if you make your own as you go. Position sense is hard to come by, otherwise.

Line of Sight Can Be Extremely Accurate

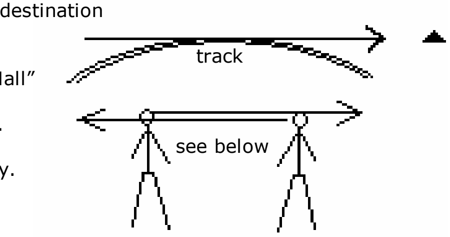

..But is often neglected, for far inferior methods.
E.g. locate yourself on a 'position line' as a means of marking your progress…
E.g. keep a straight line by walking to a point ½ way between you and your next destination
E.g. 100 fence posts can be positioned to look like one when seen end on.
E.g. fishermen can **locate** their offshore fishing spot again
— "The jetty is underneath Mount Lofty, and the tower is behind the Town Hall"
You can sometimes consult a map to fix your position...
E.g. You may notice that **"this part of the track runs directly toward"** a peak.

Un-mappable lines abound. Choose the **nearest** objects in front for best accuracy.
E.g. "That far rock lines up with this bush. **Remember** them!"

Use mapped features and the accuracy is still extraordinary.
E.g. "The peak lines up with the intersection, but if you step only three paces left you can see the difference!"
The alignment doesn't need to be accurate, provided you can memorise the exact inaccuracy e.g. 1° to the left of the hillside.
Two such alignments are needed for a fix. Look around: "What is in front of Mount William? What is behind the lighthouse?"

Two Opposite Landmarks

It **more than doubles** the alignments available to you if you have eyes in the back of your head!
With two people, separate some distance, for accuracy, then **look past each other's heads**.
Walk slowly around in a circle looking at the horizon for possible line-ups.
With one person, **place a rock on the ground then circle it**, scanning the horizon. Place a pebble to mark each feature.
The good thing about this method is that one landmark cannot 'blot out' another — all the visible features are available to you.
You may use this method to stick to a direct 'track' between two landmarks.

A Slow Kilometre Cancels out Three Quick Ones

Let's try to keep up an average of 4 km per hour…
First we travel 2 km in half an hour, but then we hit a slow section and manage only 1 km in the next half-hour (2 kph).
To counteract this we would need to triple our slow speed to 6kph for the next half-hour!
So we then decide to stop and catch our breath for half an hour.
To counteract that we would need to double our desired average speed (to 8kph) for the next ½hr!! Or 6kph for 6km.
i.e. a wasted hour takes several hours of catch-up, to regain.
(2kph for ½ hr + 6kph for ½ hr = 4kph; 0kph for ½ hr + 8kph for ½ hr = 4kph; 0kph for ½hr+ 6kph for 1hr= 6km in 1½hr)

Slow sections chew up the time available, so don't dawdle.
Quick sections can't easily repair the damage, because they don't last long.
You may try to be time-efficient and walk more quickly on the easy sections i.e. hurry along the flats,
but you can't gain much time that way.

To Progress Faster More Easily
Pace Yourself — Slow Down, but Don't Stop

i.e. Pace yourself so that you don't need rest stops, since they ruin your time budget, and it is time which determines pace.
It is better for your overall pace to keep walking slowly than to stop completely.
Go slowly on the hard sections. Pick up your pace, maybe, on long flat easy sections.
Shorten your rest stops e.g. don't remove your packs.
The old rule to cover a lot of ground quickly is: run 50, walk 50 (paces or double paces).
Catch the paradox of it — to keep going faster, go easier.

Approximate Figures for Visual Acuity

At 1:1,000 "I can just count my grandchildren"
At 1:2,000 you can count individual items.
At 1:3,000 you **can't count** them. E.g. people/shrubs/birds
At 1:4,000 you can **see** items, but can't count them.
At 1:5,000 you **can't see** them individually.
At 1:6,000 you can **read unfamiliar letters**
At 1:7,000 you can't **read unfamiliar letters**.
At 1:8,000 you can **read familiar signs**.
At 1:9,000 you can't read familiar signs
At 1:10,000 you cannot see that there are individual letters.
At 1:15,000 you can see single objects e.g. trig points against the sky.

You need to supply figures which work for you.
Multiply the size of the smallest objects you can **count** by 2,000
Multiply the size of the largest objects you **can't** count by 3,000
Multiply the smallest items you can **see** by 4,000.
Multiply the largest details you **can't** see by 5,000.
Multiply the letter-stroke widths you can **read** by 6,000.
Multiply the letter-stroke widths you **can't** read by 7,000.

Of course it's rough! I told you so in Rule 1: Guess!
Adjust the numbers to your eyesight e.g. "I can only count to 1500".
The idea is to build up your subconscious guesswork.
E.g. "I **can count** to 3000 and can count house windows" try 1 m x 3,000 = 3 km away
"I **can't quite see** heads" try 25 cm x 4,000 = 1 km away

Test Your Own Visual Acuity

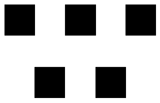

Walk up to a signpost until you can read it. Pace the remaining distance.
Measure the size of its letter strokes.
When do you lose sight of a dollar coin?! Can you count heads in a crowd at 600 m?
Set up a 1 cm test pattern of squares (for others) to count. Try yellow squares too.
Test yourself in different conditions — on an overcast day, in direct sunlight, in glare, on a hazy day, and at a large distance.

Align the Map to Identify Landmarks

Lift your eyes from the properly aligned map back onto the landscape… "There should be a peak in the distance over that way." "Look for an outstation in this direction" "Can you see the transmitter structure??"

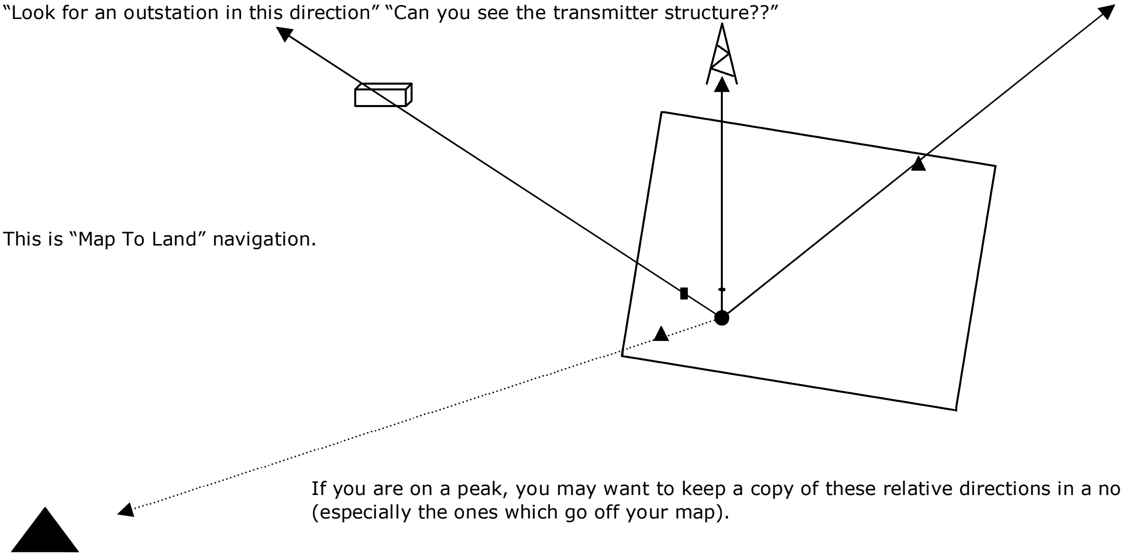

This is "Map To Land" navigation.

If you are on a peak, you may want to keep a copy of these relative directions in a notebook (especially the ones which go off your map).

"Magnets Grip Maps"

This mnemonic is meant to remind you to twist the map to the north — the map should be one piece with the compass needle.

- Place the compass on the map
- **Grip both the map and the compass in one hand**
- Turn the whole map, or yourself, or both — so that the magnetic needle points to the map's magnetic North

The result is that the map is aligned to the landscape.
Keep it that way!
Even while you walk, and twist, and turn… This is called "Swinging the Map".
A one-off effort is called "Orienting the Map".
Swinging all three of you is called "Orienting yourself" e.g. to North.

Align Your Map in the Dark

The commonest way to align the map to the landscape is by a magnetic compass.
The magnet — the compass needle — must 'grip the map', magnetic North to magnetic North, in order to keep the map oriented.
One method is to set the compass protractor to zero — true North — and set the compass edge to grid North/true North (see diagram) and then set the magnet needle to magnetic North — as marked on **the compass rose on the map**
Another method is to line the edge of the compass up with the magnetic North mark on the rose. Then you simply keep the needle point pointing to zero on the protractor.
A third method is to set the compass to the correct compass heading then line the protractor edge up with the desired course on the map, and the needle to the protractor, then follow the compass arrow.
A fourth method is to set the compass to the correct true course on the map, then line the compass needle up with magnetic North on the map, then follow the arrow. But **always line up the two magnetic Norths!**

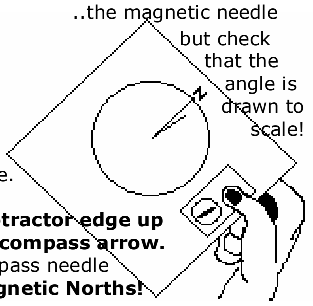

The Magnetic North marked on the map is crying out to be matched up with… ..the magnetic needle but check that the angle is drawn to scale!

A Compass Holder

Rather than hanging the compass around your neck, where you hardly ever see it, you can make it constantly visible.
For map making, it is indispensable.

Version 1: Aluminium wire around your neck and around the compass, like a bib or harmonica holder…
The compass circle is grasped by a circle of non-magnetic wire.

Version 2: Hang a flat map board from your neck, and blue-tack the compass on to the map board…
Drill holes in the four corners of a light board.
One string loop goes around your neck and through the two close corners.
Another string loop goes around your neck and through the other two corners and has an adjusting knot in it (rolling hitch).

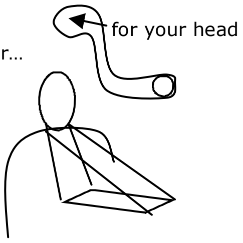

The compass protractor is set to 0° i.e. Use Magnetic North
or it is set to x° i.e. Use True North
To orient the map, or yourself, turn around and face North.
For a back-bearing, turn around and use the white end of the needle to read your direction — where you have come from.
Use the "NORWEST; NOT LEAST" mnemonic to keep you in touch with normal orientation. See p.101c & 106b for how to build up a map, or keep track, blind, using such a compass holder.

"Changeover Lines" of Progress — Relative Apparent Sizes

All visual elements which surround you have an apparent size, each of which changes as you progress.

E.g. "The hut behind us seems so tiny now, poking out from the sea of snow"
E.g. "That hill now seems to loom **high** above the plain" (a vertical angle).

Two apparent sizes will, at some stage, become equal, at e.g. as you move away from one object, which shrinks, and toward another, which expands. Simply notice "the changeover" and you have 'a line of position' to log your progress.

"The hill is the **same size as** the leaning over gum tree" (a vertical angle, a diagonal angle, and a changeover line)

You don't need to remember what size you measured — just compare them for equality!

"The width of the patch of forest is much less than the island is **across**" (horizontal angles)
"But the island alongside us is about the **same width** as the **gap** between the two" (another "changeover point")

You don't get a single point fix, only a line of separation. The third and fourth diagrams each show an example.

A measurement without numbers is like saying "We are halfway there" instead of saying "4.5 km on"
or like waiting until you are "Exactly East of Mount Big" rather than measuring "079°C" at any time
or somewhat like "These two line up, if you put a hand span in the gap"

The closer the objects, the better. The more distant, the more careful you need to be!

Use the Landscape Itself to Measure & Mark Progress

Example: "Where I came out onto the beach,
Sloop Rock was one island to the right of Gannett Island"

A top view shows that this is simply a line of sight, locating a point next to the island.

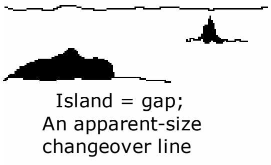

Island = gap;
An apparent-size
changeover line

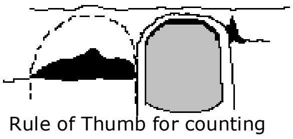

Rule of Thumb for counting

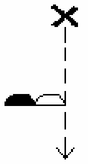

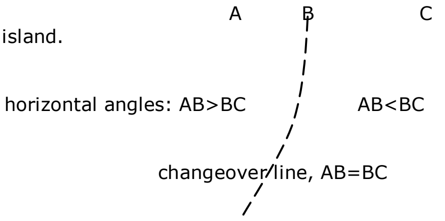

°

Distance-Off by Vertical Angle

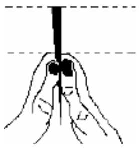

The simplest range finder is a piece of grass held at arm's stretch, (if your fingernails are too small).
Measure the apparent height of something which you can recognise later,
then measure how many millimetres of grass was needed!
You can relocate that distance-off later by its apparent height e.g. distance from a prominent peak;
from a lighthouse seen from offshore; from any recognisable tree; or from a cliff with a clear base to measure up from.
This is how our mind analyses the skyline heights — it is a very 'natural' method.

Yacht navigators will find out the height of a lighthouse from a chart,
measure the angle with a sextant, and calculate the range by trigonometry…

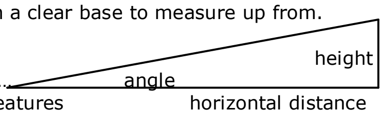

To fix your position requires two separate measurements e.g. two different vertical features
In theory, the plotting looks like two circular arcs intersecting, because any distance-off must give you a circle of possible positions.
The closer the features, the smaller the circles, and the more accurate the fixing of the location.
Confuse your friends with: "1 inch away from the windmill, and 13 mm from the pine tree"…

Safe Distance Off

You might be sailing around a headland or reef, and want to maintain safety…
Without needing to know where you are you can maintain a minimum distance off.
E.g. "no higher than 1° apparent height by vertical angle"
E.g. "keep the horizontal angle acute, between those two points"
E.g. "stay outside the apparent radar range of 5 km"
E.g. "deeper than 20 m"

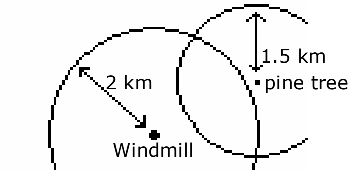

Discard one of the two possible positions.

Circle-of-Position by Horizontal Angle

Here you keep track of the horizontal space between two prominent points; or any noticeable gap; or the apparent width of an island. Use a string stretched across your line of sight, or a stick gripped between two fists.

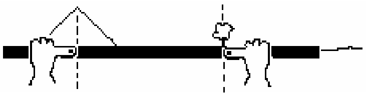

In this horizontal case though, unlike the vertical angle, the 'circle' of position is really only a cusp — part of a circumference, which must pass through the two points, and your position.
If you are closer — a wide angle — the arc is flatter; a narrower angle means you are further away — a larger part of the circle lies on your side of the landmarks.

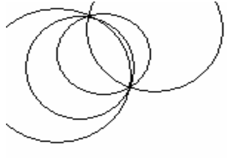

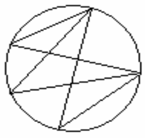

There is a constant angle
at anywhere on the circumference
when looking at the two points.

You don't need to draw this though.
Only memorise or note the angle.
To fix your position make two separate measurements e.g. one vertical angle and one horizontal angle too; or two horizontal angles—using three landmarks (or two vertical angles; or a radar range & direction from a single landmark; or use any other line(s) of position e.g. a compass bearing or line of sight or an advanced position line).
Now you can relocate that buried treasure, that fishing spot, waterhole, rare plant, good campsite… simply through making a couple of notches in your walking stick.

The best range-finder

..is a theodolite/sextant — designed to measure angles very accurately. You can use them horizontally.

Align the Map to Fix Your Position

Match the map to the landscape visually.
(The first result is to show you geographical (true) North.)
(The second result is to show up other landmarks.)
The third result is to show up **where you must be**.
(The fourth result will be to indicate which way to continue.)

I.e. Work **backwards** — from Land to Map.
One method is as follows…
Place a small pebble on the ground.
Place other pebbles in a circle around that,
each one lining up with some landmark.
Now place the map over that centre pebble.
Twist and shift the map until it lines up.
Get down behind the map and visually sight over it.
When everything lines up you must be at the centre of the circle
— over the central pebble.

Your position should **look** obvious!

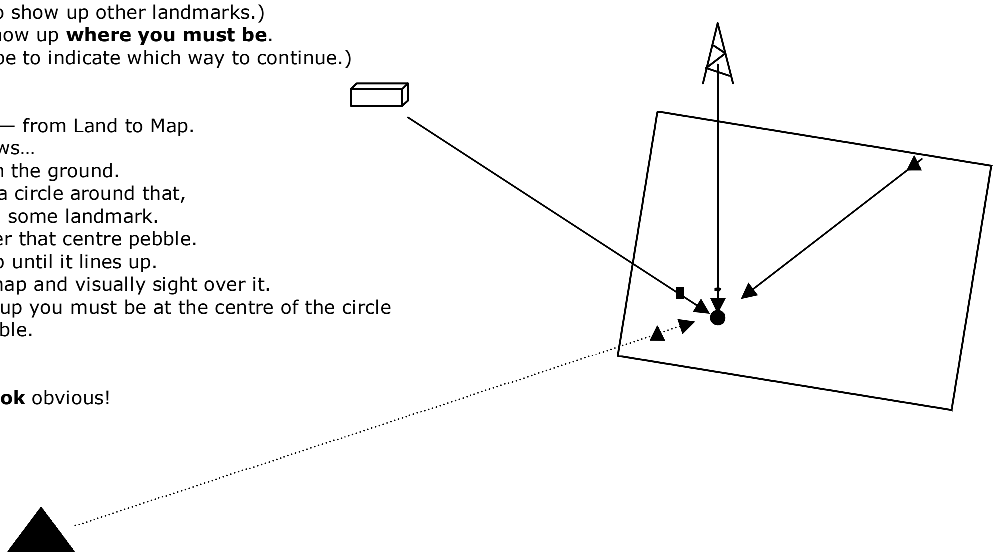

A Horizontal Angle between Features Yields a Position Circle

There is no need to draw this position circle (diagram 1), and it is somewhat awkward to construct, but there are two cases which are well worth drawing (permanently) on the map…

Diagram 2. A right-angle between two peaks — it means that those two peaks mark the **diameter** of the circle — easy to draw.
Diagram 3. The 'circumscribed' circle of three peaks — you can see whether you are inside or outside of that circle.
To find the centre of the circle, use perpendicular bisectors.

The angle seen from the circumference is constant.

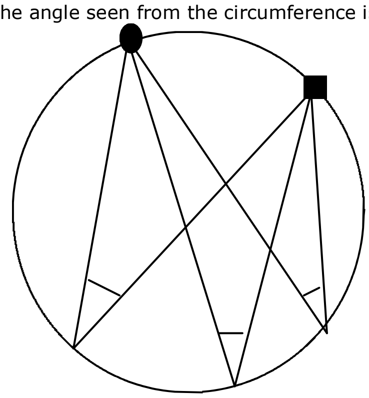

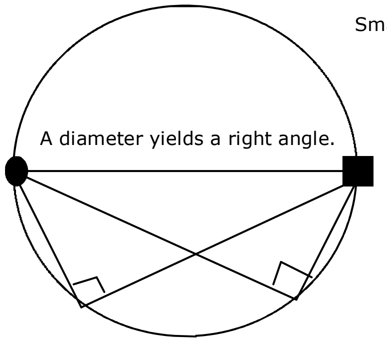

Horizontal angles are **larger** on the **inside**.

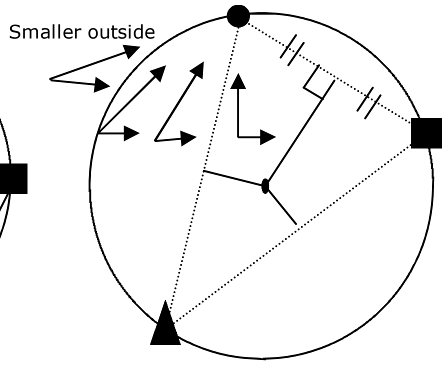

'Two Horizontal Angles' Usually Fixes Your Position

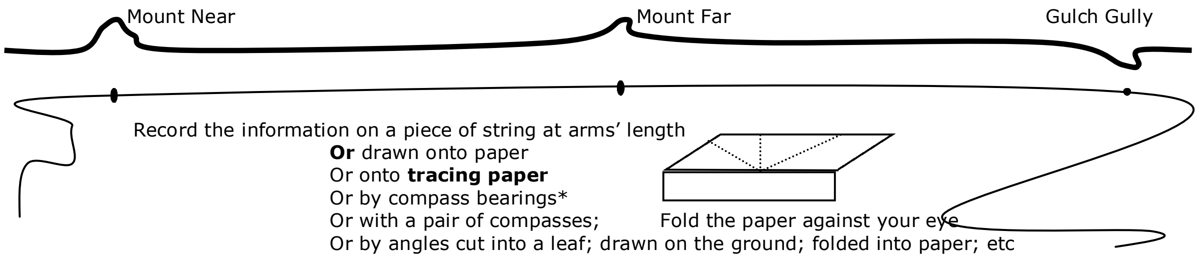

Record the information on a piece of string at arms' length
Or drawn onto paper
Or onto **tracing paper**
Or by compass bearings*
Or with a pair of compasses; Fold the paper against your eye
Or by angles cut into a leaf; drawn on the ground; folded into paper; etc
*you could plot back-bearings, or simply record the three directions for reference.

This is a quick, natural, accurate method which uses no numbers or circles. It is how our mind judges the horizontal layout.
Only one position on the map will produce such a pair of horizontal angles.
If you draw the angles onto tracing paper, you can **place it directly down over the map, to fit it on properly**.
If you make a three-arm trio of compasses with a locknut at the pivot, you can also place this directly onto the map.
Coastal yacht navigators find the method useful, and to measure angles very accurately they use a sextant horizontally.
In theory it plots as two intersecting circles — only one intersection makes sense.
The nearer the landmarks, the smaller the circles, and the better the fix.
Just don't choose three landmarks which are already on one circle along with you!

Place your eye
directly over the pivot

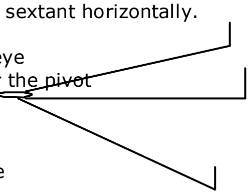

Wrong Right

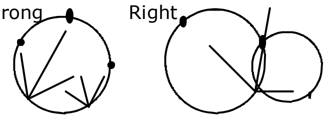

You don't have to draw these circles
Also, if you want to find an angle from the map
you can just read it from the map without drawing any circle

The Sine-Cosine Tally Method of Keeping Track, Blind

This project is only "for experience" — i.e. you are unlikely to want to do it twice! — but you should do it once.
The idea is to **keep a running total of your 'Northing' and 'Easting'** from a starting point.
I scored 97% efficiency in getting back home, on an 11 km circuit, on my first try, with an unreliable home-made pedometer!
It is a 'bottom-up' method — for when you are walking blind — when you cannot work down from the Big Picture.
With pen and paper you record your pedometer readings & directions, and update the numbers, say every hundred double paces.
E.g. "Current Pedometer Reading: 1400; (time-check 1 1/4 mins; pacing check 140 m) efficiency: 90% (slightly indirect path)
Recent Direction 005°; **adjustment needed +90, +8**; new tally: 419, -45; (comments: crossed a wet ditch)"

The chart of 'adjustments' follows this box, with an example.
200% efficiency would mean "I missed the last 100 reading"
50% = "I had a diversion" 110% ="I stopped at 110 — make the next reading 90% efficient"

- ◇ Set it all out with headings and long columns.
- ◇ Your judgment and arithmetic don't need to be accurate, because the errors are likely to cancel (!)
- ◇ Check your pedometer for each type of terrain, by time and by pacing.
- ◇ Avoid blunders e.g. check that the direction and the adjustments are similar between two legs.

It is quite pleasant to stop every minute or so — at every 100 on the pedometer — It keeps you alert, rested and oriented.
Making a full map (see 'Layout Map—Blind) is more satisfying and just as quick.

Mariners can (either draw their tacks, or, using this method) calculate an overall course-made-good from their many compass legs, by first adding up (distance x sin true course) East, and then adding up (distance x cos true course) North, from each individual distance and course.
Their overall distance is $\sqrt{(\mathrm{sum1}^2+\mathrm{sum2}^2)}$ at an angle of arc-tan(sum1/sum2). Then correct for magnetic variation.

Rise and Run Adjustments — Sine and Cos — North and East

Caution: These are COUNTER-Clockwise Bearings! — as seen from a compass holder…
I.e. The needle swings to left and to right while you point straight ahead.
E.g. when the needle points to "20° East of 0°", as in the diagram,
"N(orth) O(n) R(ight)-West" reminds you that you are heading NNW-ish,
not to the East. N20°W is your **anti-**clockwise heading.
Don't let this unusual method of measuring bearings confuse you
about normal clockwise bearings.
There are really only 10 pairs of adjustments used in the whole table
I have made East positive, only to match normal bearings and
Map grids, which use East as positive (West as negative).
Here, **West is negative; East is Positive.**
South is negative; North is positive
Examples: North East is + + N315°W;
South West is – – N135°W;
North West is +– N45°W;
South East is – + N225°W

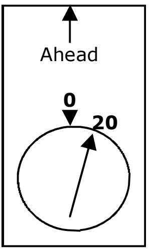

The + and – signs tell you
how to change your tally of North and East.
E.g.
100 paces to the NNW ish, as in the diagrams is...
94 paces of Northing and 34 of Westing
Or +94, –34

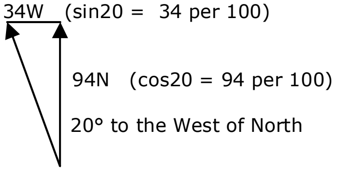

					Change	
Actual direction…	(NE	SE	SW	NW)	N	E
Adjustments…	+ +	– +	– –	+ –	cos	sin
Protractor Readings…	360	180	180	**000**	**100**	**0**
N nnn° W…	355	185	175	**005**	**100**	**9**
	350	190	170	**010**	**98**	**17**
	345	195	165	**015**	**97**	**26**
	340	200	160	**020°**	**94**	**34**
	335	205	155	**025**	**91**	**42**
	330	210	150	**030**	**87**	**50**
	325	215	145	**035**	**82**	**57**
	320	220	140	**040**	**77**	**64**
	315	**225**	**135**	**045**	**71**	**71**
The next figures rerun the previous ones...						
	310	230	130	050	64	77
	305	235	125	055	57	82
	300	240	120	060	50	87
	295	245	115	065	42	91
	290	250	110	070	34	94
	285	255	105	075	26	97
	280	260	100	080	17	98
	275	265	095	085	9	100
	270	270	090	090	0	100

e.g. As a different example, a tally of –640 +770 means that you are 640 paces South
and 770 paces East of your starting point, which means that you are now N230°W from home, according to the table.

Rule of Fingernails

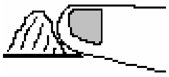

Your pointing fingernail should be about 1°, one in 60, when you stretch your arm.
Your thumbnail at arm's length, held upright, might be one in 50……
A pencil in your fist will be about one in 100.
Check it out with a tape measure and calculator. E.g. 1.2 cm (width) / 69 cm (distance from eye) = one in 57 = 1°
Adjust how you hold your arm (e.g. square in front, diagonally- or fully- on the side) and which fingernail, etc, until you find suitable combinations which you can remember.
But the trick to making this useful in range guesstimations is to guess the distance away first, then estimate the height of a mountain, as follows… 60 km away — use one in 60. One full nail means 1 km showing;
30km away—use 1 in 60 =½km showing; 50km away—use 1 in 50 = 1km showing; 10km away—use 1 in 100 = 100m showing
Reverse logic can give you the distance away, if you can see the base clearly e.g. from sea, and know the full height.

Rule of Tom Thumb

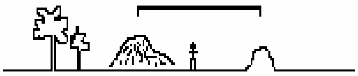

Tom Thumb

At various times you have to judge parallax.
And the trick is to project the needed sideways movement onto the intervening landscape.
Imagine yourself over there on that ridge — Are you as thick as that tree trunk? Are you as tall as that bush?
To make the two landmarks line up, imagine the distance over there in the distance.
"I would have to walk that far sideways over there" = "I would have to walk the same distance sideways over here"

Apparent Size

The size of people, cars, houses, as measured at arm's length, varies with distance. You know how big they are already.
To get a good gauge of very small angles, compare their width & height with a key or knife-blade or toothpick, held edge-on.

Which Way is North?

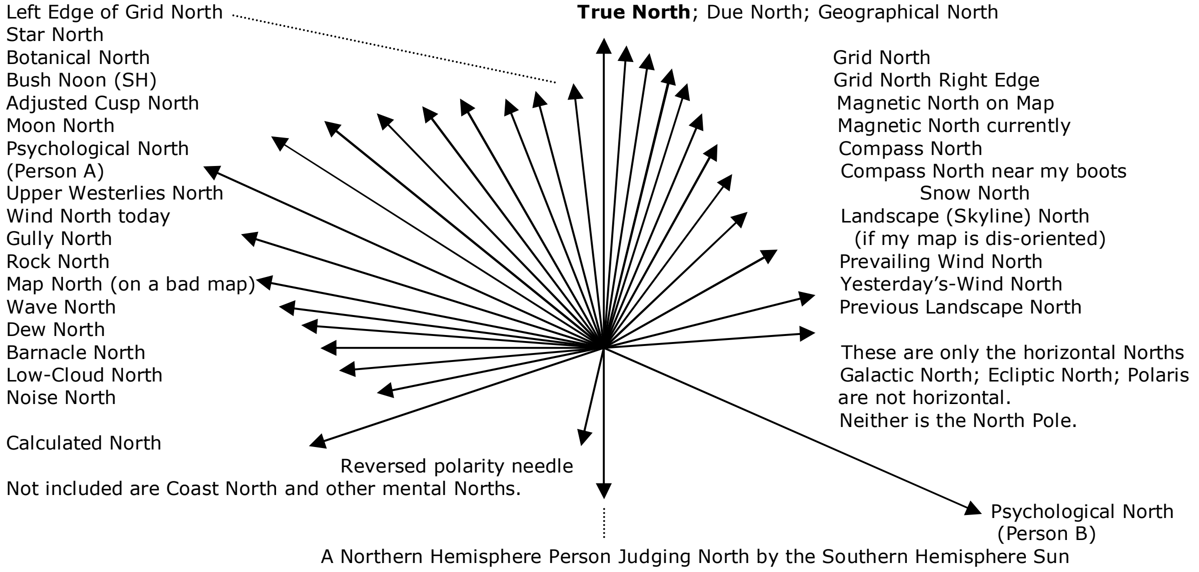
Left Edge of Grid North
Star North
Botanical North
Bush Noon (SH)
Adjusted Cusp North
Moon North
Psychological North
(Person A)
Upper Westerlies North
Wind North today
Gully North
Rock North
Map North (on a bad map)
Wave North
Dew North
Barnacle North
Low-Cloud North
Noise North
Calculated North
Reversed polarity needle
Not included are Coast North and other mental Norths.
True North; Due North; Geographical North
Grid North
Grid North Right Edge
Magnetic North on Map
Magnetic North currently
Compass North
Compass North near my boots
Snow North
Landscape (Skyline) North
(if my map is dis-oriented)
Prevailing Wind North
Yesterday's-Wind North
Previous Landscape North
These are only the horizontal Norths
Galactic North; Ecliptic North; Polaris
are not horizontal.
Neither is the North Pole.
Psychological North
(Person B)
A Northern Hemisphere Person Judging North by the Southern Hemisphere Sun

---

'East-least; West-Best' Rule — Tells You Which Bearing is 'Best'

This rule works only for horizontal directions, measured clockwise.
It is for shifting between True, Grid, Magnetic & Compass bearings 0-360°.
"Least" really means "Less than the other alternative number"   "Best" really means "More than the other choice"
The following wording ("the North which is") is important, since you need to choose correctly between a right and wrong choice!
The "North" which is East **(clockwise)   of another "North"**   ***gives the Least bearing*****.**
Any *"North" to the West* (anticlockwise) of an alternative "North" yields the *Best* & Biggest numbers.
Specifically, it is most used to shift back and forth between magnetic North and true North bearings.
The wording may then be rendered: "Compass error West; compass heading best". Memorise that wording.
Because this mental arithmetic is a difficult ask, you need to draw a diagram to make the conversion clear.
Learn this reminder "Check-up-on-yourself!" diagram…   And visualise it, on the real landscape in front of you.

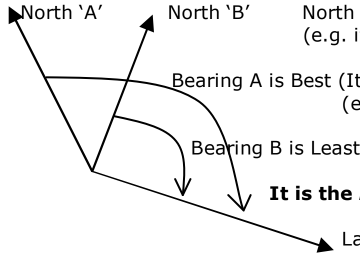
North 'A'
North 'B'
Bearing A is Best (It yields a bigger number — East-around)
Bearing B is Least
Landmark

North 'A' is **West** of North 'B': it yields the **Best** of the two bearings.
(e.g. if A is true North, and B is magnetic North, the 'magnetic variation' is 'To the East'
and "the compass heading is Least")

(e.g. if the magnetic bearing is 085°, the true bearing is bigger… 120°)

North 'B' is **East** of North 'A': it gives the **Least** clockwise angle.

It is the *North* which is "East" or "West"; The *bearings from it* are "least" or "best".
Numbers are meant to make sense — your job is to look for that sense, visually.

E.g. Compass Deviation………If my compass *North* points too far East e.g. by 10°, 050°M = 040°Compass (-*bearing* is Least)
E.g. Grid Convergence Correction…If Grid N is 2° tilted to E of true N, 070°T = 068°G (Grid is Least, since Grid N is East)
E.g. Map to Compass………… 090° on the map means: set your compass to 100°M, when magnetic variation is 10° West (Best)
E.g. Compass to Map………… 205°M as a compass bearing = draw 220°T on the map, when magnetic variation is 15° Least
E.g. *Add* the GridtoMagnetic Convergence clockwise *E* angle to magnetic N to get *Grid* N "MAG=**M**agnetic, **A**dd(the GMC)=**G**rid"

String Map

This is a fun-project. As you un-do each knot later, you relive the journey!
A suitable string is builders' twine — long, thin & strong.
An extraordinary number of code variations are possible, provided you can "read" the different knots afterwards!
Keep it simple, for a start.
Example: Start with a slip knot. Half-hitch it every hundred paces, to keep track of distance.
Tie a special knot at any change — to remember it by.
E.g. a strangler hitch around some bark for "thick forest"!; a thumb knot around a feather for "Water Bird Lake"
Chain the slip loop to keep it going. Or thumb knot the loop to lock it all up.

A simple numerical code is: slip knot; + n half-hitches; chain on; m half-hitches; chain on; ongoing tally…
n might mean something like 'direction' by clockface
m could be 'terrain type'
Here are some variations…

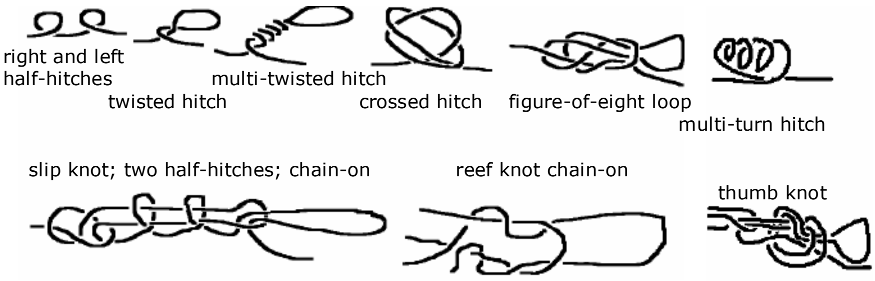
right and left
half-hitches
twisted hitch
multi-twisted hitch
crossed hitch
figure-of-eight loop
multi-turn hitch
slip knot; two half-hitches; chain-on
reef knot chain-on
thumb knot

ProgressCompassCompass

Progressing A Position Line across the Map

Situation: You catch sight of one landmark; plot a back bearing; then travel — say for two hours, West — before you catch sight of a cross-bearing. You can still plot a 'fix'! Example (fictitious): You lose sight of Mount Connor in the Northeast; and later see Ayer's Rock. Your pace is 4kph. From the bearing of Mount Connor, you plot a position line (by a back bearing running back from Mount Connor). You must label this '10 a.m.' Now **shift the whole position line** 2 hours to the West i.e. 8 km, and label it '12md'. To do this choose any two points on it; shift those points 8 km West; draw **a new line parallel to the old**. You are now ready to plot another back bearing — this time from Ayer's Rock — to cross the advanced position line, for a 'pinpoint fix'. Alternatively, you could 'shift' only Mt Connor on the map, first.

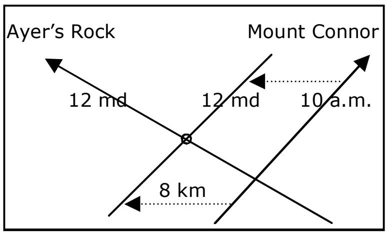

Pinpoint "Fixes" are a Myth!

(Rule: Don't suppress uncertainty, nor reject any unwelcome information!)

Any 'line' of position looks exact. But it comes with unavoidable uncertainty and should be plotted (mentally at least) as a wide band of 'likely position'. Potential inaccuracies abound. E.g. lining up the compass;
reading the compass; plotting that angle;
grid north variation across the map; map distortions;
errors in magnetic variation; protractor inaccuracy;…

You can now advance your "area of uncertainty" across the map, expanding it with each blind move!
'Deduced Reckoning' navigation at sea (i.e. "intelligent guess")
is more a matter of keeping track of **accumulated errors**
due to tides, leeway, inaccurate course-holding/recording/plotting, inaccurate instruments. So too on land.

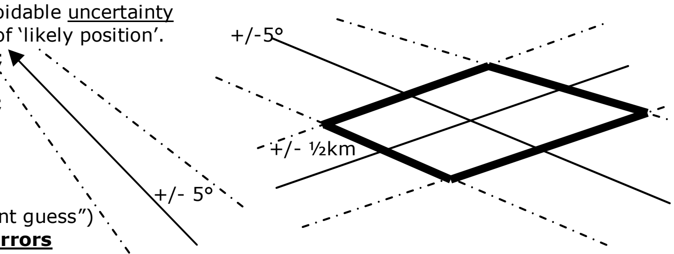

Getting Started in the Dark

The task is to know what direction to head off in, when you can't necessarily see the landscape…
You need a Map & Compass.

TAKE A MAP-BEARING — Find the desired angle, on the map, in relation to the North-South grid……………
— "About 60°, as a **guess**"
— Use the edge of the protractor
— Measure the angle — only the protractor is needed for this!
Line up both sets of parallel lines — mapgrid lines,
and the lines on the needle-holder-circle.
— The protractor will display the angle between the two sets of its parallel lines
"55° E. of N." (It is marked clockwise)
(This is Called Working from **Map to Compass**)

SET THE COMPASS — Using 5° East magnetic variation, and ignoring compass deviation…
Compass bearings should be 'Least'
i.e. 050°C
— adjust the 'Follow Me' Arrow to point to 50° (to the East of North)
Twist the circle on its mounting.
— Don't alter the protractor again!

FOLLOW THE COMPASS — Now turn your attention to the needle
— Line it up with the 'zero' on the protractor, by twisting the whole instrument,
And by using the grid lines on the movable circle as a guide.
— Twist yourself this time, to line up with the 'follow me' arrow. And off you go!

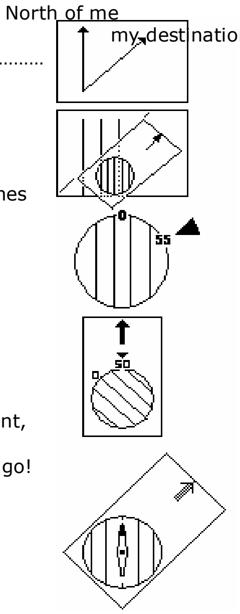

Mnemonic: **Map to Compass** (= Map to Compass Course) Means **M**ap Grid to Compass Zero;
Angle the **Ar**row to the course target;
Put on the correction; **To** Get it, use the East Least West Best Rule;
Compare the Compass Zero to the Compass Needle; **Pas**t the Follow Me Arrow is the Course Target.

Thinking Backwards

"Danger bearings" allow you to stay safe, without knowing your position accurately, but you need to think accurately…
As seen from the lighthouse, you must stay on the right of 315° (and to the left of 060°). That much is clear from the chart. Anything 'to the right', is clockwise from the observer, so for the keeper to keep something (e.g. you) 'on the right' means to keep its bearing bigger than ahead. 'Ahead' is parallel to the 315° track shown, since the lighthouse keeper is wishing to see you 'on its right'. He will be happy to see you on any bearing bigger than 315° and less than 060° i.e. 315 to 360, 0 to 60. From your point of view, that means, that you must *use back bearings,* of 135° and 240°, and then keep those tracks on your right and on your left respectively, as you look in. 'Ahead' is parallel again, but opposite to the track shown on the chart, if you are trying to keep the track 'on your right'. Looking in, moving yourself to the left will place the track further to the right, and make bearings to things on it bigger. That means, in numbers, the lighthouse will bear more than 135° and less than 240°, i.e. It is safe to approach at **anything between 135° and 240°**. Notice that the more/less sense is preserved with such use of backbearings.

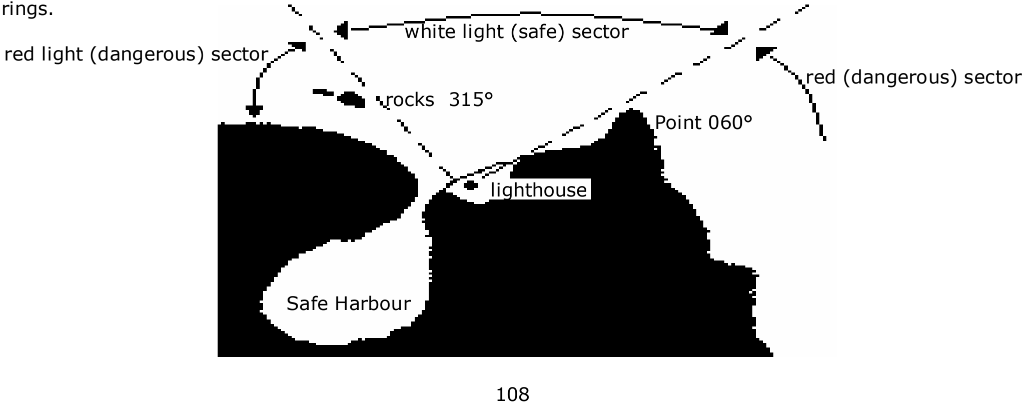

Taking a Bearing from the Landscape

- **Start with a guess** e.g. "That hill is about 060°T from me"
- Point the 'follow me' arrow to the landmark
- Grip and twist the movable protractor circle until the grid marks on it — i.e. the zero — line up with the compass needle
- Read the resulting angle off — between the zero (which is "North" —Compass North in this case), and the landmark.

The result should agree with your guess e.g. "The landmark lies at 065°C" i.e. East of Compass North i.e. clockwise from.
Take a B.E.A.Ring Mnemonic: **B**earings are Tracks; **E**xpect to have to plot a back-bearing;
Arrow to Landmark; **Ring** to Needle and **R**ead It off. (Think: "Arrowing" to remember A)

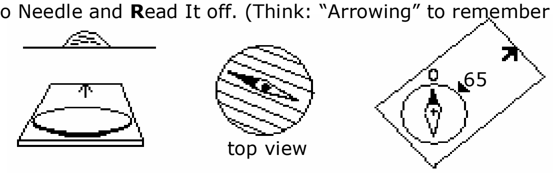

If it helps you sight the landmark, hold your compass string vertically upwards from underneath the tip of the follow me arrow.
The 'Zero' is meant to be used for 'North'! But if you point the follow-me arrow to North and the 'zero' to the landmark, then (a) it is more awkward and (b) it is an anti-clockwise "bearing".

If you are about to use a map, you must now convert the compass bearing to a 'True bearing'.
Firstly mentally change the compass bearing to a magnetic bearing if you are aware of any compass deviation i.e. compass errors. (You usually do not worry about compass deviations when bushwalking.)
Next mentally convert the magnetic bearing to a true bearing using the East-least, West-best rule.
Example: "6° variation West" (The Magnetic *North* is West of True) "Compass Best" so… 065°C = "059° True"
Lastly, physically change the protractor to read 059° when you want to draw on the map, at an angle to the map's True North.

See opposite & above Plotting a ('Compass-to-Map') Back-Bearing

Draw a line on the map from the landmark in an opposite direction. Use the protractor to get the angle right.
If you see a landmark North of you, you must be South of it.
Or, what looks 'South-West of me' must mean that you are 'North-East of it'.
Think backwards! See "Back bearings: +/-200° then -/+20°" p.18e E.g. 059°T from me to it…
Note that I have already converted Compass to True. 239°T from it to me.
One bearing yields **one** line of position (from a back-bearing plot — diagram 1)…
Two bearings from two different directions cross for a 'fix' — diagram 2…
Three bearings don't cross. They form a 'cocked hat' and keep you humble — diagram 3…
In this case, you could be anywhere nearby (which is also true for any "Fix").
Expand your possible positions, don't try to pinpoint a "better" one.
Take more care next time, especially on the closer i.e. the more reliable landmarks.
Check your compass for fixed errors e.g. "always 3° too much".
Choose directions which are quite different from each other, to get the clearest crossing.
Compass to Map Mnemonic: (Compass B.E.A.Ring, then) **Compass to** Grid (or True) Correction;
Make it a Back Bearing; **A**ngle to Grid; **P**encil to Paper—Draw It in. (Think: "Mangle" to remember A)

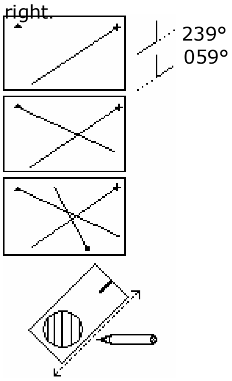

No-Number Back-Bearings

Method 1. Use the white-tipped (South) end of the compass needle, when you take the bearing.
Method 2. Point the "follow me" arrow back toward you, when you take the bearing or follow the compass.
Method 3. Line the protractor zero up with Map South, when you plot the bearing.
Method 4. Point the compass edge to the landmark on the map and draw your back-bearing line backwards from the landmark.
You don't need to alter the compass setting at all (except for the magnetic variation or grid/magnetic convergence angle).
Warning 1. Don't mix back-bearing shortcut-methods, or you may get a 'double negative'!
Warning 2. Check all you do, against reality, for sense.
Warning 3. Don't try to shortcut Compass-to-True conversions at the same time as shortcutting the back-bearing, or you will get confused.

Parallax While You Walk — JUST DRAW IT!

Most **hand-made maps** start with **a known baseline** scale — a straight segment of a known distance.
I.e. "I walked from here to there in a straight line for 1 km" — for this you may need a pedometer.
Or if you are Captain Cook or Matthew Flinders "We sailed on (this) bearing at (this) speed for (this amount of) time"

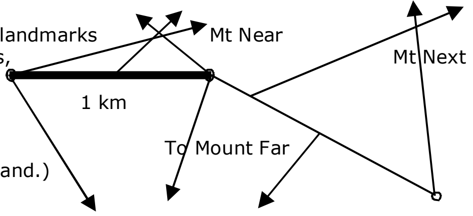

You can then **add cross-bearings at any time** to fix the position of distant landmarks — obviously at the beginning and end of the baseline are the preferred places, perhaps combined with one from the middle…
The longer the baseline the more accurate the fix on distant marks.

You really only need a protractor for this **running base-line triangulation**.
(And on a Mudmap, you just eyeball the direction and draw an arrow in freehand.)

Here is an example of the easiest sort, and the reverse of a double-angle…
After a week at sea, with no sight of the sky to navigate by, I was understandably over-anxious about my position, and "needed" a "fix"! I decided to sail toward an isolated lighthouse-island, in the hope of passing close enough to see it, perhaps, if we weren't too far off course. The winds were hurricane force, of course, and the sky was filled with salt-spray lifted off from the water's surface. It was dusk. Suddenly, above the mist, dead ahead, the lighthouse turned on! Alarmingly high, and close. Too close!
Were we "found" or about to be "lost"!? (See "Aiming Off"!) After the cheers: "What do we do? How far off are we?!"
"Turn at right angles to it; Maintain that course; Take regular bearings until it is diagonal; Keep track of the distance run. Then resume your course. That way we can tell how far away it is."
"How can you do that!" said the bewildered, inexperienced navigator.
I replied: "I'll **just draw it**"

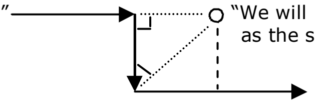

"We will pass as far away as the sideways distance we have run"

A "Start-Stop-Rest" Navigation Mnemonic Overview

START
"Start" has six points, as illustrated in the diagram below…
These mnemonics, e.g. S.T.A.R.T. or S.T.O.P., will be expanded in the rest of this section—they summarise the book.

- ◇ Start with a **Prayer** — "What more do I need to do to avoid getting dead?"
- ◇ **P.R.E.P.A.R.E.D.** — What have I forgotten to do or to take, to practise or remember? Start off oriented before you go!
- ◇ **S.T.A.R.T.** — with your start-off-trend-direction, which is the single most important navigational item. Start off aligned!
- ◇ Start on **T.I.M.E.** — Read your watch, and benchmark your trip. Say: "Time & Tide" as well, because sometimes tide levels are crucial.
  — Check pace & progress. Estimate the Range, because time is all about distance.
- ◇ Start time is **6 a.m.** — It's always time to **look** in **6** directions, plus 2—Time to Look **A**ll **A**round, **A**ppreciatively, plus time to look at the **M**ap
- ◇ Start with **N.S.E.&W.** —4 sets of 4 directions. The final W is for **W.E.A.T.H.E.R.**

The final R signals that you are **Ready** to go. Don't start until you **R**!

S.T.O.P. A **S**top is a **S**ignal—Interpret the Why? of your stop: Does it **S**ignal a change in watershed, catchment or view;
in **T**rend or **T**errain or '**T**ime & **T**ired';
in **O**rientation (embarrassment—if lost, rule one: Stop);
in **P**rogress/**P**osition/**P**lacement—campsite, junctions, features?

REST is best remembered as a mnemonic of 5 RE's…
R.E.-O.R.I.E.N.T. — using these letters as a mnemonic. E.g. **RI** re-Visualise the Invisible Big Picture, and **ENT** Relate yourself to Nature.
RE-RE.L.A.T.E. — re-interpret all directions to relate to each other and to the skyline, as explained below. Now don't forget...
RE-MEMBER — the WAY and the SEQUENCE and especially to NOTE down (in the NOTEBOOK) the NAMES, and DRAW what you need to.
RE-LAX — Play the Navigational Game—teach it—and 5 other games.
There is a final '-art' to 'stop and rest-'
RE-ST-ART 1.1 **re-**pray; 1.2 **re-**prepare; 1.3 **re-**s.t.a.r.t; 1.4 **re-**t.i.m.e.; 1.5 **re-**look (6 a.m.); 1.6 **re-**direct (NSEW), (which includes **re-**weather)
... e.g. 1.3 Re-S.T.A.R.T. after a rest, on the next leg—to relate it to the last & to the *start*-finish trend.

Summary: **"Start"** 6 times**; "Stop, and Re-st-"** (with 4 new **"Re-"**s)

(The Start mnemonic is a good example of the "clumping" technique
—never more than a handful of things to remember, yet it leads you into hundreds of points.)
Put the S for Sky at the top, where it should be.
Use a hexagon for the 6 START up points.
Use a clock-face for TIME,
and the compass-rose for NSEW.
Show the 6 directions in 3 dimensions.
Balance it all on a triangle
—for the "6 a.m. start" , since it has 3 parts.

See the boxes to follow

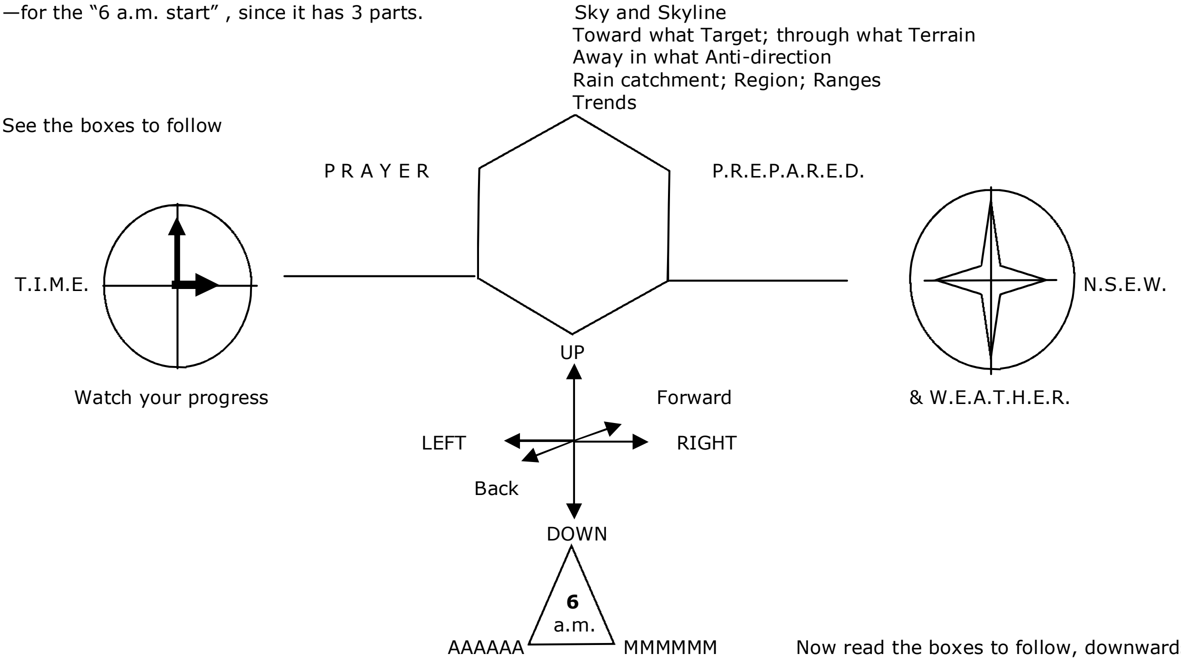

Now read the boxes to follow, downwards...

P.R.E.P.A.R.E.D.—A Before-You-Go Checklist ..so that most of your navigating is already done, from home!

P.ractise Global orientation. Get a blow-up globe. Teach others. Learn (what to learn, look for, do etc.). E.g. predict the Sun.
Learn the Mnemonics, Codes & standards e.g. S&R signals, compass directions
Pace your paces — metres per minute; steps per 10 metres. Left-right bias (blindfold).
R.esearch History. Ruins. Artifacts. Mines. Geology. Get a geological map from the library. Landform. Soils. Climate. Windy season. Prevailing winds. Wet season. Vegetation analysis. Rare species. Biogeography. Geography — global neighbours; time zones; local coasts & river mouths. Catchments. Cardinal features — places N,S,E,W from you. Land lines on map. Place names — their meanings and history.
E.quipment See the next box.
P.lace names See the box after that.
A.stronomical orientation Sleep on a map, and with a compass nearby. Point the map North. Times and directions and heights of Sunrise & set, Moonrise & set. Southern Cross. Pot. Sun at noon, at magnetic North/South, when East, when West, at magnetic East and West. What landscape markers (from the map) for these directions? The star opposite the Sun. Time zone. Local noon — its exact time. The Sun's coordinate predictions. Your longitude. The "Equation of Time". Moon-phase, -number and –day of week. Tide times.
R.eminder Note your intentions—and tell someone where you are going and when you expect to be back.
E.xpectations of the Weather Barometer reading. Get the Weather forecast. Follow the Changes. Buys-Ballot's law. Steering winds.
D.irections & Familiarity Ask someone who has been there. Read track guides. Look at an aerial photograph. Go beforehand! Find your old track notes. Look closely on the way in. Choose a route. Plan your strategies. Coordinate the cars, the car keys, any food drops, shuttles or codes

Equipment — A What-You-Might-Like-to-Take Checklist

(E. in PR**E**PARED, for "Equipment"…)

- **A Map** **A compass** **A clock**
- A notebook and pen — for track notes. A track guide. An aerial photograph. A navigator's handbook.
- For map making — graph paper, paper, pencil, a compass-holder. A map-holder (string around neck to a…), clipboard, stopwatch.
- Polaroid glasses — to orient by the dark blue sky band. Radio — for RDF, weather. Mobile phone. GPS. Camera.
- Binoculars. Pedometer. Rangefinder. Calculator. Altimeter. A sine-cosine table.
- A map cover. A spare map. A navigational protractor. A magnetised needle. A sundial compass. A two-flap cardboard compass.
- Marker tape; Chalk; Scissors (for cutting leaves!). A clear plastic tube. A clear plastic drinking bottle (for an artificial horizon).
- A long piece of string (e.g. for a range-finder). Short pieces of string (e.g. for a plumb-bob). A walking stick. Someone who knows the way.

(P in PRE**P**ARED) **P**lace-Names —Learn Them—You Navigate by the Local Features!

"Look. That's Big Split Rock and that's Little Split Rock. Wild Dog Tier is that range over there, and this plain drains out over there, into the Ouse River You can find your way anywhere around here from those."

Suddenly I was unforgettably oriented, just like he said he was, when his dad told him.
For me, the funny names on the map now meant something worth knowing.

Place-names are not just names you don't know; they name places, which you ought to know.

More accurately, place names 'place' you in the landscape, as well as 'orient' you to it.
A sense of location is almost independent from your sense of direction, and is rarer to come by.
Direction-sense places you in relation to very distant background references i.e. it doesn't 'place' you at all!
Location-sense places you in relation to the immediate foreground — visible landmarks, catchment basins, fixed 'tracks', etc.
Place Names Place You

S.T.A.R.T. off Oriented

— **S**tart-off Direction? (e.g. "upstream") — the **S**ingle most important navigational knowledge — it tells you the return path, and the search path.
Sky/**S**kyline i.e. especially according to the **S**ky (e.g. "toward the Moon; into the sunset") and related to the **S**kyline.
Start-off directions for the parallax method, pp. 51a,53a, e.g. to magnetic NSEW skyline or recognisable middle-ground points, or rhumb-line point and another set of such bearings, taken from the map, as seen from the finish, if you are on a one-way trip.
— **T**oward what **T**arget? Specify a
Track (e.g. "downhill to the lake") or a **T**rend. A rhumb-line bearing is best—from start to finish.
Through what **T**errain? (e.g. especially see what the land-shapes would look like without vegetation)
— Away from What Point? (e.g. "from the lookout carpark), and in what
Anti-direction (e.g. toward that other peak"—a skyline direction)
— **R**ain Catchment — In what major
Region? — as can be gauged by the surrounding
Ranges. (e.g. "on the main spur dividing North from South arm, downstream of…")
— Trend: **R**ead the **T**rend off the map, for starters; then continue to **R**ead the **T**rends & bends and the progress as you go on.
Read (your compass; map; watch) in **R**elation to the **T**rend, i.e. in order to follow the trend, which is the focus of this mnemonic.
Plus: **R**ead any other **T**rends, from any other places, from memory, in relation to the sky. Be able to revisualise them at will.
— Re-s.t.a.r.t. on the next leg, to relate it to the start-finish trend, and to relate it to the previous leg.

T.I.M.E. & Distance — that's the point

TIME to Look at your Watch — PACE YOUR PROGRESS — the point of this T.I.M.E. mnemonic is distance.
— "Watch" your steps—i.e. calculate your pace from the elapsed time;
— "Watch" the Sun without looking at it, because your watch tells you where it should be.
— TIDE times may be necessary.

Instrument Check — watch, stopwatch, pedometer, odometer, rangefinder — for progress
— compass: Double the bow angle, or halve the stern angle p.94c; parallax readings on landmarks.
— (and other instruments — barometer, altimeter, GPS, radio null, radar, depth-sounder)

Map-Scale Check — are you judging the scale well? — What is the next goal? (This M is not for Map in general, but only for distances.)

Estimate the Time of **E**xpected Arrival — guess each milestone ahead of time to set you free from a nervous TIME. Nominate an E.T.A.

Distance

PROGRESS: can also be gauged by lines of sight and changeover lines. Use a stopwatch, or pedometer. Read the scale.

PACE: distance divided by time; 20min/km, double and halve = 3kph; 1mps = 3.6kph; 4kph normal walking pace;
finger-math helps; pace yourself evenly.

RANGE: Use the letters of RANGE & FACULTIES as a mnemonic to remember the different ways of gauging distance in the distance...
Range squared = dee height; Area divided by offset; New direction parallax; Grass-leaf vertical angle, horizontal angles too; Educated guess
Familiar lengths, and Fingernail Rule; Acuity of vision and Apparent sizes; Cloud behaviour; Upper and lower guesses, averaged; Landscape lengths;
Tom Thumb Parallax; Intervening detail & haze alters your perception of distance; Extreme range comes in two bites; Sound travels at 1 km/3secs and
Sidewise offsets with blinking.

150km from 1 nautical mile high; 1° latitude from 1km; 50 miles from 500m; 50km from 1/5 km; 5km from 1/500 km (2 m); 12.7km from 12.7m.

Look in These 6 Directions — in 6 ways — Up and Down (Part of "Start at **6** a.m.")

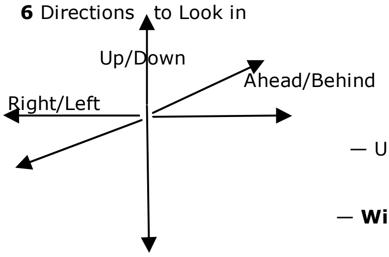

UP — **Astronomy** — Sun Moon Calendar-Stars
— **Sky** — Clouds, Contrails — Blue Band
— What's up with the Weather?
then DOWN —Geology — **Soil** Type; Rock Type
—Clues — Animal **Trails**; Bootprints; Fallen markers

— Up-**Sun** — haze & glare
— Over what tropical country? Then — **Down**-Sun — haze, shadows, colour changes

— **Wind**-wise — Up-Wind — Where on the skyline? — Down-Wind — Around what obstacle?
— Around what pressure system? (Wind in your face)
— From what coast/ocean? Listen; Smell
— Up-Breeze — To what cold surface? — Down-Breeze — To what warm surface?
— What's up-Weather? Where are the steering winds coming from?

— **Land**-wise
— Up-Hill — To what dominant peak?
— Point to the dominant peak, even when it is invisible,
— Down-Hill — antidirection?
turn around to see what lies away from the peak — the "anti-peak",
especially in relation to your shadow — this anti-direction has a lot to
do with a compass back-bearing—a visible line of position, you see
— Up-Ridge — To what hill/hillock? — Down-Ridge — Down which ridge? — To what headland?
— what offshore/underwater topography e.g. shallow reefs; islands?
Since this mnemonic is about direction, try to point to them.
— Up-Slope — To what ridge? — Down-Slope — Into what valley? Name them.
At what angle across the gradient?
— Where is the lowest upslope skyline, under the canopy? — Where is the highest downslope skyline, under the canopy?
— Up-**Stream** — To what distant divide? — **Down**-Stream — What stream? To what distant mouth?
— to what sedimented shelf; or deep-water channel?

Summary: **Above & Below** (twice each) + **Sun & Wind** (above eye-level) + **Land & Water** (below eye-level) = 6 ways for "Up & Down"

---

4 More Slopes

1 in 4, is 14°;

10° (aim off, 1 in 6) is 175 m per km;

1 in 60, is 1° (1 in 57.3 is the exact figure, but it doesn't multiply up linearly after a few degrees, whereas 1 in 60 does well after a few degrees)
& n° is n in 60
(1 in 6 is 10°; 1 in 10 is 6°);
The equivalent is 5/3 in 100, 5 in 300, 167 in 1000, for use with multiplying up with the isosceles triangle angle method.

36°, 8 in 11, is critically unstable.

Multiply them up, inversely.

Look in These 6 Directions — in 6 ways — Ahead and Behind

Look Back Regularly: Is bad weather sneaking up behind you? (p.46c) "Abracadabara Peek behind you" (p.63b)
How much has the track bent around behind you? Any acute track junctions? Can you remember this way back?
Compare Forwards with Backwards: Do you need a back bearing? Are you half-way? Is your starting point visible, to get an anti-track forwards?
Are there any changes in the terrain, to photograph, in your mind? Have you memorised the transition, from the far side?
Look forward — in **time** — the forecast! What is the Estimated Time of Arrival? Look Back — in time — How far have I come?
Look ahead and behind in **U**pper case **Time** — Could you recognise this place much later, in hindsight? — What is the history of this place?

Look AHEAD — Has the **view** changed — into a new catchment? Look BEHIND — Have you left the old view behind — crossed a divide?
— at the terrain, the track showing up ahead, and the difficulties — Do you remember the **way**? Could you recognise it again?
— Look for track **markers**, and for trails ahead — Look behind for hidden, reverse track markers.
— Line up two (new) marks exactly ahead to keep your direction sense. — Look for the old lined up mark left behind.
Is it time to choose another leg, and a **pair of marks** lined up to it? Do I need to find new ones behind me?
Mnemonic: R,S,T-U,V,W,X-Y-Z **R**egularly **R**everse your nose. **S**tart/anti**S**tart bearing & **S**napshot-**S**equence your transitions
time (+**U**pper case Time); The **V**iew; **W**ay (remember it?); **XYZ: X** marks the Track (Markers); the (**Y**&**Z**) Leapfrog Marks =6 ways for Ahead & Behind.

Look in These 6 Directions — in 6 ways — Right and Left

Look RIGHT Through! 1. — the bushes, to the land-form underneath; to the contours & bedrock
2. — the canopy, to the skylight beneath it
3. — the trees, to the most distant marks, including other trees
4. — the treetops, to the sky — the Moon, clouds and the blue band
5. — the foreground, to the most distant skyline horizon, whenever you can glimpse it
6. — the haze, to the skyline details beyond the land horizon—use binoculars too.
E.g. to barely visible mountain peaks, or a second skyline & its parallax, or to islands in the haze, or the rising Moon.
Look LEFT 1. — to the most distant thing you have LEFT BEHIND on each side! i.e. Have you left it behind yet? Is it time to choose another one?
2. — Have you left behind your: handbook; air photos; track guide; track notebook? — No, well, don't just carry them — Look at them!
3. — Have you left anything behind at the rest stop? It's an important habit, to ask this each time you move.
4. — Look to the right and left, for the contours, the natural 'tracks', and the behaviour of the network you are crossing and leaving behind.
5. — Things on the right have big bearings; things on the left are lesser; than straight ahead, even less as you leave them behind.
6. — NORWest & NOTLEast — North needle 'on the left', means you are heading roughly East. (N. On your Right = West-ish)
Summary: **Right Through & Left Behind**.
In summary for all the 6 directions make sure you **Check Each One Out—Thoroughly!** (6 ways each)

..Plus in 2 More Directions — (1.) All, Around, Appreciatively, And At 6 L's

1. **A**ll your senses: Listen, Smell. Look around with your other senses.
2. Look 360° **A**round for the best & worst showings of colour, shine, flowers, greenery, new growth, fern fronds, lichens, trunks, branches, dirt. Compare Ahead with Behind, & Left with Right, & upSun with down-Sun, & up-wind, uphill, etc. for direction-dependent light and colour effects.
3. **A**PPRECIATE — the scenery; the colours; the contrasts; the shapes, outlines and silhouettes; the foreground movements.
4. **A**n Environmental Notice Time (see later box E.N.T. in REORIENT)
5. **A**ny LOCAL PECULIARITIES to remember.
6. **A**nd **A**t — **the six L's...** — **Land**marks — distances, sizes (apparent Largeness), altitudes Use binoculars.
   - — Are any familiar, from other walks you have done elsewhere?
   - — Do the place names make any sense? — interpret them!
   - — **Lay**out — relative angles, compass bearings. Am I inside or outside of major landmark circles/triangles?
   - — **Lie** — slope and terrain, catchment and drainage, contours and altitude
   - — **Lines** Catchment Dividing Ridges and Rivers; Contours at right angles to the slope; Flight Tracks
   - — **Ligns** — 'Same direction' Lines of Sight, or nearly so; and 'Opposite directions' alignments: Am I in between?
     - — the fixed A-lignments . Am I left or right of major 'ligns'?
   - — **Sky-Line** — talk it through; draw it. Cross-relate all directions to each other, and to the skyline

Summary: 'A' for 'a.m.': 6 A's, with 6 L's

..(2.) Look at **the Map**: a Systematic Checklist. (What 'System'??)

The system is to examine **top-down**. From general to specific detail. From large-scale to small-scale. Look for the downward progression in these…
3-D location: Latitude, Longitude, Elevation. Centre of the time-zone. The inset map showing where you are on the continent. Scale & size. The overall width- and breadth- distance of the map coverage. Grid size e.g. 10 km squares. Where you are on the map, geometrically too. The Starting Point. The Relative Location—over the horizon landmarks; coastal towns N.S.E.W. of you; something NE,NW,SE,SW of you. Radio stations. Highways. Light sources. Powerlines. Pipelines. Place names. Visible Landmarks. False peaks. Lines of the land — between landmarks. Recognisable hill-hill reference distances. Recognisable elevations e.g. towers. Grid North; True North; Magnetic North; Magnetic Variation; Yearly Drift in Variation.
Catchments & Divides; Catchment Size; Adjacent Catchments; River Drainage. Land shapes i.e. interpret the slopes, contours, the contour interval.
Trends and Bends — of the coast; rivers; roads; ridges; ranges; the track. Notice any creeks which flow backwards i.e. against the trend, and why.
What rising ground will be visible in each direction? Perimeter features surrounding your walk e.g. a river, a range, a road, the coast, a forest. Escape routes — unplanned exits. What are the ridge-line limits of the sub-catchment which most encloses your possible location? What high points are on it? Possible routes. Clockwise or anticlockwise? Distances. Overall distance of your route. Milestone features. " Wake-Up " landmarks.
Initial bearings to landmarks. Initial angles to landmarks. Initial horizontal angles to landmarks. Final bearings. Rhumb–line trend from start to finish Around what Obstacles? Fine details of the route. Don't forget the fine print — the date and reliability of all the map details; the grid datum for GPS use; the legend of symbols; colour codes; contour line thicknesses.

Mark the Lines-of-the-Land On Your **Map**

6 M's for Map:
Map Checklist (above); Map-Lines (this box); (all below...) Map Alignment; East-Least West-Best Mnemonic; Map & Compass; Map-Making Methods.

Pre-prepare your map — **Mark in the navigational lines** which the mapmaker has left out. Most of these are permanent.

A. Cardinal Tracks — NSEW lines. Across the whole map, from the very prominent peaks; or simply covering a local area, from the local peaks
B. Alignments — e.g. from Peak to Peak; Tower to Town; Point to Point — You may see two in line, or walk in between two landmarks.
C. Directions to major distant features — (for orientation purposes) — e.g. capital cities; coastal towns
D. The Overall 'Track' — the RhumbLine — from start to finish. Mark this lightly; &/or two or three obvious overall 'legs'
E. Circles of Distance Off — e.g. 5, 10, 15 km from Mt. X.
F. Circles of 90° horizontal angle — i.e. make two landmarks the diameter of a circle
G. Circumscribed circles around (i.e. which just include) the triangles in the triangular grid of landmarks
H. Transmitter to Transmitter networks — i.e. a triangular network of major peaks; include the distances and bearings on each line.
I. Flight Paths — if you can get hold of an aeronautical chart — it forms a second triangular network.
J. Catchment divisions.

Mnemonic: Four Straight Lines; Three Circles; Two Sets of Triangles; And One Very Sinuous Set.

Mnemonics for **Map** and Compass Work

◦ Map Alignment

A.L.I.G.N. the **M.A.P:**
A… Align the map to Sun with an imagined **A**rrow, and to your course with **A**nother **A**rrow;
L… to the **L**andscape;
I… To **I**dentify Landmarks and to get a Fix;
G… **G**rip the Magnet to the Map, and the Map to the Universe;
N… **N**ot "Swinging the Map" As You Turn;
Magnetic Needle to **M**agnetic North (in four variations, "0°" to North or arrow to course, Magnetic or True);
Angles Horizontal, fit over the map;
P… Align the **P**ath like landing a **P**lane, or align the map by aligning the **P**ath.

◦ Mnemonic East least, West best

East is Least and West is Best, tells you the North which gives the bearing that is best. i.e. the biggest in number; or lesser in number.

◦ Map & Compass

LMNOrP: Landscape; **M**ap; **N**eedle; **P**rotractor; either one **O**r two at a time, but not all three.

M.A.P. To Com.Pass, = Map to Compass Course, Means…
Map Grid to Compass Zero; **A**ngle the **A**rrow to the course target; **P**ut on the correction;
To Get it, use the East Least West Best Rule;
Compare the Compass Zero to the Compass Needle; **Pas**t the Follow Me Arrow is the Course Target.

Take a B.E.A.Ring: Bearings are Tracks; (**E**xpect to have to plot a back-bearing); **A**rrow to Landmark; **R**ing to Needle and **R**ead it off.

Compass to M.A.P.; Compass bearing, then **Compass**-**to**-True Correction; **M**ake it a Back Bearing; **A**ngle to Grid; **P**encil to Paper—Draw it in

◦ MUDMAP:

Mudmap on paper; **M**ark the start point on one edge; **U**p-arrow for orientation (**U**pdate it); by a Reference **D**irection (Up**D**ate it);
Maintain a trend, a long straight leg; **A**lign the **A**rrow when you stop; **P**encil in the **P**rogress, and a new leg direction.

◦ Other **M**apmaking **M**ethods — Use the letters of COMPASSES:

Copy a Map Top-Down; Own Map is Valued; Mental Map—Map to Mind Navigation; Parallax onto a Running Baseline; Accumulate the Layout Blind; Sine-Cosine Tally of Accumulated Eastings and Northings; String Map; Exploring a Cave System Needs a Book, of Strip Maps.

"N.S.E.& W." — the four cardinal directions, is a mnemonic of fours

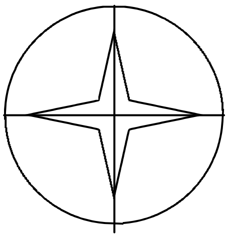

N for "**N**orth Coast" — all about visualising the invisible directions…

Name which coastal towns are N. S. E. W.
"**S**weetheart" features Name your favourite spots NSEW of you
Encircling horizon features — **Horizon** North; **Skyline** South; **Environmental** East; **Landscape** West.
Which-way-am-I-going trend (rhumb-line), from the map, then marked onto the terrain—e.g. to behind that hillock

S for "**S**hortest **Shadow S**outh" — all about the path of the **Sun** (and Moon)—the 1234 system
— and the **S**unny **S**kies quadrant system
— How the Sun, Blue Band trend & ends, and your Shadow, currently lie in relation to…

N Highest **Noon** (How High?)
S Shortest-**Shadow** Souths (SH) & the two-flaps South. Coldest South is the best reference in the SH. String South using the Sun's Shadow.
E Sunrise (Where?) Anti-sunrise from a string-line to the horizon.
W Sunset (When?); Bisect them. Anti-sunset from a string-line in the afternoon.

E for **E**quinoctial **E**ast—i.e. other things astronomical, **E**tc.

Needle-North-Seeking-Northern-Canada—over Red Hot tropics!
South Pole Star (How High?) & Southern Cross (Where?) & Star-Chart octants & Star Compasses & Star Fixes
Ecliptic & Equator & Equinoctial Precession
W for the motion of the Pot, s**e**tting in the W**e**st (rhyme), at the latitude angle

W for **W**ind **W**est, i.e. Natural Directions.

N for Nature's Norths: Rock Noon; Botanical Noon; Solar Panel Noon; Satellite Dish North
S for Snow South; Frost South-East; Steep Gully South; Mud-in-the-Shadow South; Dead South; Siesta South.
E for Moss East-South-East
Wind: Upwind: scorch, erosion & flats, Downwind: shelters, debris & steeps; Noise & colour from up-wind; Rumble & skyshine from downwind.
& **W**eather Directions (& see below)—Where is Wind West? What is Upwind? Upper Winds Steer the Weather (from the West?) What is Upweather?

Summary: 4 directions with 4 variations; Then finish with **W**.E.A.T.H.E.R. After a rest, re-direct yourself.

W.E.A.T.H.E.R. — Check for Change

W. for Wind. Has it changed — in strength, direction, gustiness?
for Wind — in Your Face — Low on Your Left. Low winds bring the rain, clockwise.
for "Weather systems come from the West" — in temperate latitudes — can you make a prediction?
for the upper level steering Westerly-Wave Winds — are they deviating much from average? Are they steering a storm toward you?
for "Wind-West" showing up in the vegetation, for when the prevailing wind is not blowing.
for the prevailing Westerlies (or whatever).
E. for Ephemerals — i.e. the clouds and contrails (don't forget this E.!)
Have they changed? — in height, type, wateriness, thickness, darkness — in directions — steering wind directions? In relation to the Sun?
— Examine them carefully for dissipation (i.e. thinning, soft edges, layering) or for building up (hard edges and heaping).
A. for Atmospheric Pressure Change — imagine the isobars across the land from the forecast.
Sometimes you can see a lowering atmospheric pressure when marsh gas starts to bubble in cold weather i.e. not due to warming.
T. for Temperature — of the air-mass
H. for Humidity — the absolute one (measure the dew point — will it frost?). For the relative humidity try huffing to produce a fog, or onto metal.
Does the air mass humidity & temperature, speed & turbulence let you know which direction the wind must be from?
E. for Expectations — i.e. the weather forecast — test it — is it coming true?
R. for Rain — Will it? R.ight? Are you **R.eady** to move? Grab your R.aincoat and let's go… Keep looking until it is time to S.T.O.P. for a rest…

Oriented? No?
Rule one: Stop! getting more lost.
Rule two: Start! getting unlost. Start navigating!
Rule three: Stay, or Steer a Straight course.
Rule four: Don't do it again!

Position/**P**lacement/**P**rogress—has it changed?
Have you arrived somewhere?
Is it a named feature? A campsite? A junction?

Have you changed view?
Have you changed catchment?
Interpret the Why of your stop...

Trend And **T**errain—Has it changed?
Time & Tired—Does that explain your stop?

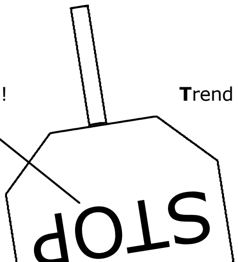

Stop -Signal: Why have you stopped?

REST and R.E.O.R.I.E.N.T yourself

Take 5; a 5 minute rest; REST is a mnemonic of 5's.
RE-RELATE; **RE-ORIENT; RE-member**; **RE-lax;** (= the 4 new RE-'s before you) **RE-start**. = 5 RE's

RE-RELATE to the skyline.
RE-RE for Re-Read the map,
L for Re-Late the Map to the Landscape
A for Astronomy—relate the Sky to the Skyline
T for Trends—imagine the Trend of invisible landscapes, in relation to the Sun, and relate your current trend to a point on the skyline
E for Each direction clue Re-lated to Each other direction clue, especially to the landscape/skyline shapes.
RE.O.RI.ENT — **RE** for Re-Orient the Map to the landscape,
— **O.** for **O**ops! Don't make logical blunders; Have questions on your mind; notice; speak up; integrate; connect; interpret.
Observe the **Overall** trends and bends.
— **R.I.** for **R**e-Visualise the **I**nvisible (Next checklist; and see Section 1 Astronomical Overview; Include the clock-sense mnemonic here)
Re-late yourself to the Big Picture
— **E.N.T.** for an **E**nvironmental **N**otice **T**ime — Re-late yourself to Nature — coming up.

R.I. For Revisualise the Invisible (and then Relate It to the skyline!)

Work down from the Biggest Picture. Work most on the ones you can't yet visualise.

Astronomy — Visualise the paths of Sun/Pot/Moon/Stars; Point to the South Pole Star; North Pole Star; the **Clock-sense** of stars around each Pole

— The star opposite the Sun during the Day & during the night; the Southern Cross during the day; the Ecliptic; the Milky Way

— Earth rotation; Horizon rotation; horizon trails; clock-sense of Horizon rotation; **clock-sense** of low stars' movement around horizon

— the Moon-Sun, and Ecliptic line; The Moon phase explaining the present tidal force. See the overview at the end of section 1.

Global — Latitude; Polar distances; Equators; Earth Poles; Seasonal Sun height; Shadow South — visualise the aurorae; Antarctica;

— North Magnetic Pole in Canada, as a 'South-seeking' Pole; Aurorae; Unequal Tidal bulges. "Go Overseas" p.25e.

— Tropical countries — corresponding to sunrise, noon, noon height, sunset, the country under the Sun now, the time zones

— Light & Dark halves — Sunrise-/sunset-circle countries; Half the arc-distance; How the horizon is shifting due to your movement.

Continental — Weather map isobars—visualise them across the continent — the one in front of your eyes, and on a continental aerial photo.

— Latitudinal wind system and pressure system helices.

— North Coast; N-S-E-W townships; NW-NE-SW-SE features. "Walk across the Continent" p.29e. "Compass-needle Continent" p.9e.

Local — Catchments; Divides; Mouths; Mounts; Radio Transmitters; Night Lights

— Nearby landmarks & places & highways; Home; Car

— Perimeter features; Escape targets & Routes — Bird's-eye layout of your route.

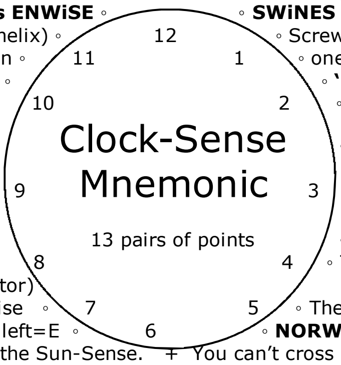

0 The standard viewpoint is "seen from above"; ◦ Clockwise is to your right; Compass Bearings Circle Clockwise;

1 Wrong Way Go Back! **Anti clockwise is ENWiSE** ◦ ◦ **SWiNES** That's how the Sun moves in the northern hemisphere

2 Unscrewing is an anticlockwise motion (on a right-hand helix) ◦ ◦ Screwing in is a 'right-hand' motion (for a right-hand screw)

3 24 hour clockfaces have only 15**°** per hour, like the Sun ◦ ◦ one hour = **30° clockwise**, on the 12 hr clock face method

4 High pressure systems circle **anti**clockwise in the SH ◦ ◦ **'Low' winds** come clockwise, in the SH.

5 Things on the **Left** have a **Lesser Bearing** ◦ ◦Things on the **Right** have a **Bigger Bearing**

6 All the solids rotate **anti-clockwise,** seen from the **N.** ◦ ◦ **Low stars** all around the horizon follow the Sun-sense

The ground under your feet shifts anti-clockwise in NH ◦ ◦ The ground under your feet turns clockwise (SH)

7a&b So the ho**RI**zon **N**ever **H**eads **RI**ght in the **NH** ◦ ◦ The ho**RI**zon **SH**ifts **RI**ght in the **SH**

8 Wind-shift due to ground friction opposes the Sun-sense ◦ ◦ Coriolis forces follow the Sun sense

9 **North lies between the SH Sun and the hour hand** ◦ ◦ The NH Sun is between the hour-hand & South

(actual noon points to Sun; clockface tilted parallel to Equator) (noon-time to Sun; axis to Polaris)

10 The SH Sun seems to move anticlockwise ◦ ◦ The Moon goes backwards, night by night (clockwise, in SH)

11 Nose Nav'n turns everything away **NOTLEast** Nth on the left=E ◦ ◦ **NORWest** — Needle-North on the right = heading West

12 **Shadows** Follow the Sun-Sense. + You can't cross a stream twice in the same flow-sense

---

E.N.T. for Environmental Notice Time —Orient yourself to Nature

Take **a top-down approach**...

Geology; geo-trends; geo-morphs i.e. landforms; geo-Norths

Dividing ridges; dominant peak; upstream catchment size; wind shadow; rain shadow

Season; climate type; prevailing wind; average rainfall

Rocks; bedrock; outcrops; parent rocks e.g. upstream

Soil types; nutrient availability

Soil Drainage — good or poor; Water Sources; groundwater? salinity

Aspect i.e. which direction does the slope face, in relation to high noon?

Living stresses? — salt, minerals, wind, exposure, frost, heat, low humidity, fire, flood, wet roots, pests, diseases, competition, grazing, trampling, disturbance, rock instability.

Vegetation type; the overall density of green, and of cellulose, and the height; dominant species; associations of plant types; diversity; individual species; individual plants. Pollinators available; seed dispersal mechanisms operating.

Mammals; Birds; Reptiles; Insects. Drinking water availability, cover for animals, hunting pressure

The 'Natural Norths' — Gully North; rock North; Bush North; insect North; Dew North; snow North; ripple North; prevailing wind North

---

RE-MEMBER to Note down the Names in the Notebook

Remember the **Way**, and the **Sequence**.

Remember to take a Notebook. Make a neat permanent copy of your notes i.e. in ink, maybe sometime later on, in your master track notebook.

Be Bit-wise!

Name the subsections, junctions & links. Name the peculiarities. In the name, try to include…

The Why, When & Where of your stop.	Link & Sequence information — time, distance, context
Lie of the Land	Words to indicate the direction trend, & curves — left or right
Visual words — colour, shape, alignments, curves	Descriptive words. Collect the details. Enlist the others to describe and name it.

Give yourself something to remember — create an acrostic; craft a creative name; sit down and write up your track notes

Update any acrostic alliteration. And the count e.g. of the creek crossings. Do something numerical. Do something alphabetical.

Be vivid, clever, memorable. Do something Shapely or Geometrical there Do something Memorable there. Do something there.

Draw the shapes, outlines & skylines. Draw the directions of other landmarks from any high viewpoint. Take reference photographs.

Remember most of all to do it all — don't just recite the reminders to do it all!

Mnemonics of Fives: Remember the Way; the Sequence; Make Notes; Invent Names; Draw What You Need to.

Re-lax, and Play — Navigation Games and Activities (5 pairs to recall)

Walk & Talk — a night walk — a fog walk — a walk off the track — climb a hill with a small scale map & binoculars & compass.

"Track My Trail" ("If you want lunch, you'll have to find me — give me a quarter of an hour, then track me down").

Keep a compass course, blind — keep a straight course, blind — on a slope? "Let's all make an acrostic to remember this track by".

Try Orienteering — Try making a map of a maze of trails — then a string map — a strip map — or a sine/cosine tally map.

Look & Learn — "Go to the front of the line if you spot a track marker which the others have missed" In a windy place look for wind clues, etc.

Take it in turns to name a navigational or orientation clue which is visible, until you have to drop out.

"I notice… (fill in the blank). Your turn. What extra can you add (to the description, or to the clues)?"

Guess & Test — Each of you estimate the time of arrival; the direction to the car; the distance to a landmark — then check the map.

Guess the time from the Sun, Moon, Southern Cross, or other stars, then check your watch. Use your compass to check your guesstimation of North from the bush, gullies, rocks, a star-trail, a star-set etc. Predict the tide, or the time & direction of moonrise — then sit down and wait for it. Predict a star-set position. Find your 'shortest shadow' angle today & use it tomorrow.

Make & Use — an odometer from a wheel, a pedometer, a sundial/declination compass, a portable sundial, a string rangefinder, a slope-meter.

Measure — Try some star navigation observations — Rise & set bisection; Use a plumb-line; a bush quadrant; a bush sextant.

Test how accurately you can each use a compass. Test each compass against the others.

& Memorise — yesterday's acrostic. Remember all the mnemonics! Can you add to any of my checklists? Have a star identification evening.

Moon-Phase & Tide-Time Dial

This is a handy tide-predictor, based on the assumption that tides will follow the moon. It shows you where the moon will be, at high tide. It also tells you what the moon is doing if you know the date but haven't seen the moon for a while. The deluxe version can predict tides and phases years ahead of time. It consists of a series of dials rotating on a single pin.

◇ The idea is that many of these concentric cardboard/plastic dials rotate together once properly adjusted, and as quickly as the Moon orbits the Earth. They show the Moon Phase, the Sun, the Earth and the resulting astronomical tide-producing forces.

Each disk is loosely fixed to a neighbour with friction (e.g. a smear of dried silicone glue) or a clip or tape or a pin, to allow adjustment.

Set it up by observation, or by tide-table predictions. Add notes, as you accumulate information about locations, tidal range, etc.
It can be especially useful for travellers, e.g. when fishing an unknown coast.
If the dials are correct for any one tide (spring-tide is best), they should be correct for all the tides thereafter *
* give or take a couple of hours for the smaller effect of the position of the Sun.
The effect is 43% of that of the Moon, and can be estimated with practice—change the hour reading to toward the Sun, or the anti-Sun, a little, if the Moon is near the Sun-anti-Sun line (no more than 45° from). Local shallow water resonance effects will also alter the actual effect of the Sun. When in doubt, around neaps, try averaging the tide-time for one or two days ahead and 1 or 2 days after. The theory is that slosh resonance, or the Moon alone, or the Sun alone, might predict the difficult tides better.

◦ The hanger (outline not fully shown) is hung from the wall and also folds down behind to receive the pin on which the dials rotate.
It doubles as an indicator window, and should mark five consecutive high tides (ignoring a.m./p.m.).
◦ The clockface is 24-hours i.e. 15° per hour clockwise, (markings not fully shown). Think "a.m./p.m." for the two high tides per day, no matter what the reading says.
Each new location visited may need the clockface to be twisted slightly to allow for local tide delays. Keep a note of such settings.
Daylight saving requires a +/– 1hr change — to relocate the Sun forward in summer, back to normal in winter.

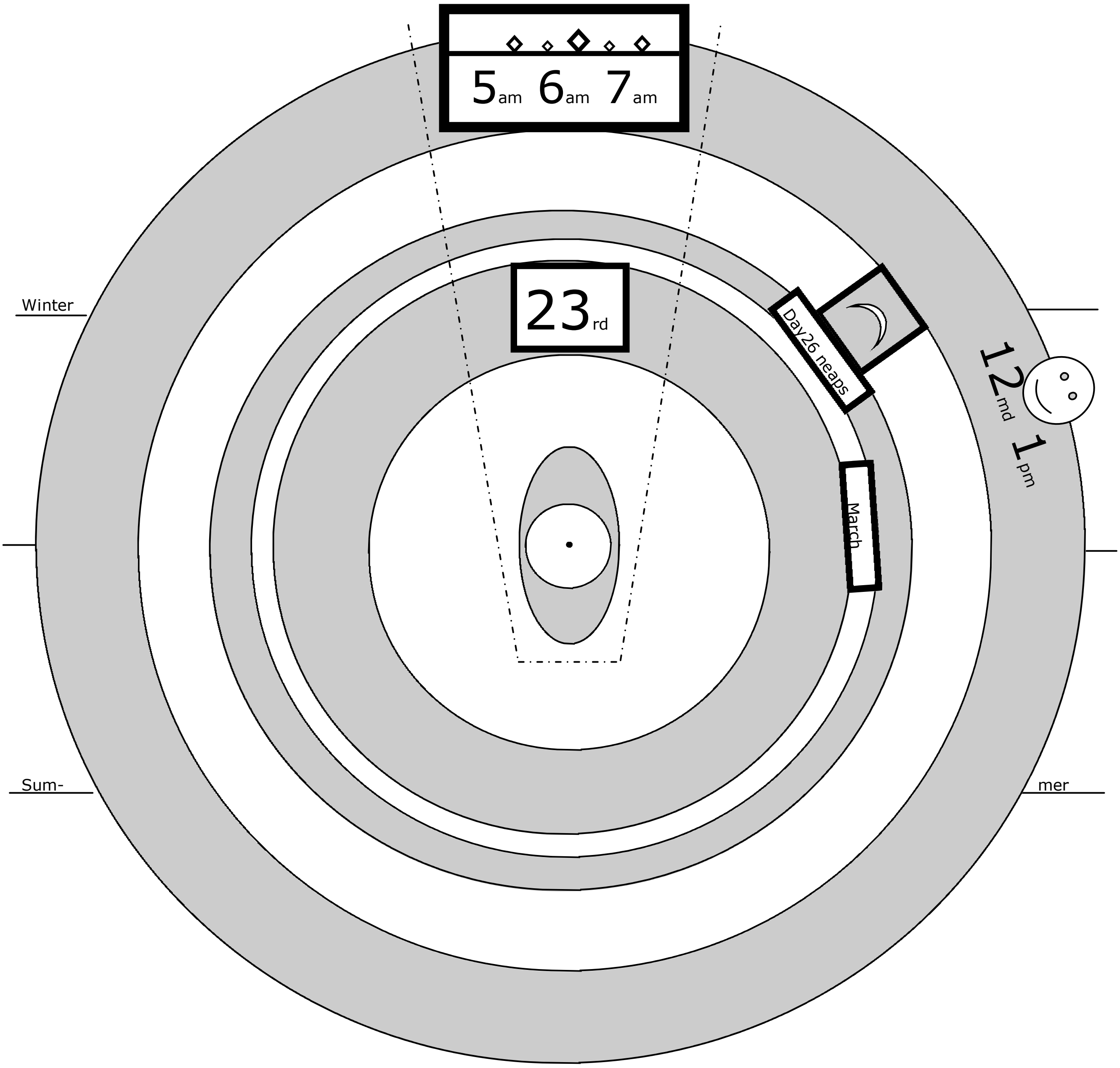

◦ Friction-fix the Moon-window dial (white with grey window) to the hanger, so that it maintains its position. It covers over all the other Moon phases.
◦ The Moon-phase diagrams (29 of them) underneath that, on a separate dial of the same diameter, should look right for the Southern Hemisphere—i.e. the Sun rising in the East, illuminating the Moon; all the phases parallel. Make the Moon-shape highly visible from a distance. Attach the Sun to the No-Moon phase, Day 0, via a transparent link; then set the Sun to the Clockface, to indicate true noon—when the Sun is exactly North.
◦ The optional tide-type ring (showing "neaps"), set to the Moon-phases, allows you to adjust the delay after new/full Moon, before spring tides.
If you also indicate which high tide is dominant—the one nearest the Moon or the one farthest—then that will need adjustment when the Sun crosses the Equator. I suggest that you set the top tide for the tide when the Moon is visible, to give you a visual feel for the tide producing forces.

◦ The date-dial is 29.5(30526805) days around—the complete lunar month—and marked clockwise from 1 to 32 (only one date is shown in the diagram). **Dial up the date, by rotating the Sun**. (For a 12-hour clockface, you would need to double the 12.19077473° per day). Since the next high tide will (if you are looking in the daytime) be on average that afternoon, I suggest you think of the date number as "the middle of the 23rd"—to avoid adjusting the date hand minutely for how far through the day the next tide will be.

◦ The month window is attached to the date dial, at Day 1 of the month, if you include a month setter.
◦ The optional but useful month-setter dial (markings not all shown) starts at March 1, to avoid leap-year adjustments; It makes March 32 = April 1, April 31 = May 1, and so on—spanning approximately 11 days for the year. Restart every March 1, and reset it every month. Calculate the monthly positions by the day number since March 1, avoiding an off-by-one error (17.9°;23.6;41.6;47.3;65.2;83.1;88.8;106.7;112.5;130.4;148.3). Set the month name to the first day of the date dial.
You can also calculate and indicate a one-year increment, 129°.6327779, a leap year, 141°.82355, and a four-year increment, 170°.7218865. To make an adjustment, make the old date indicated by one mark on the month dial, now line up with the increment mark selected, so as to indicate a date in a new month/year/leap year cycle.

◦ An optional Earth ring in the very centre, can show the entire Southern Hemisphere map (not illustrated, or NH on the reverse side), and be set to the right longitude. It is fixed to the hanger, and to the tide-bulge ring...
◦ A second ring just behind the earth can indicate two tidal bulges, one pointing upwards (or otherwise if you want to indicate the local tidal delay compared to the actual ocean bulge).

Mock it up in paper, until you understand the functioning.
Glue those paper discs onto cardboard to trial the thing for a while.
After deciding on improvements, make it neatly, to last for decades.
E.g. Try getting the computer to print the dial markings onto acetate, and glue that to sheets of stiff plastic.

You can experiment with different dial orders, e.g. the date-dial on the outside of the phase-dial.

You can make room for comments,
e.g. by expanding the dial area showing, to say when the moon rises and sets; e.g. "light from midnight onwards"—beneath each Moon-phase diagram
e.g. when and where the Sun rises and sets; Mark three horizontal lines in the background for Summer, Equinoxes and Winter.
I.e. mark them on the backfold of the hanger piece, or on the wall. Try to match midwinter sunrise time to where the "winter horizon" cuts the clockface, which will also suggest the approximate direction of sunrise.
e.g. when and where the first, second and third quarters set in each season; Full Moon opposite the Sun, i.e. 6 months different; Half Moon Waxing is one season ahead of the Sun; Half Moon Waning is one season behind the Sun (as far as which horizon line to use).

You will discover that an invisible connection needs to be made between the outer top window and the inner window and dials, so that they do not rotate with the other dials. One way to do this is to push the pin through a piece of rubber eraser, and glue that rubber to the dial. Another method is to have a square pin, and key the inner dials to it, while using round holes for the rotating dials. A third and simpler method is to have a transparent link from the centre, over the top of the other dials. You do not want the hanger to obscure any diagrams or dials which need to be visible.

The Time and Date dials are the indispensable ones. Instead of dialing up the date, you can use a normal clockface (12md always on top) and an hourhand to point out the next tide; you point the pointer at the date. On a 24hour clock-face it can be made like a + sign, to point out all 4 tides that day.

Here is the simpler alternative version, showing an 11:30 high tide on the 19th…

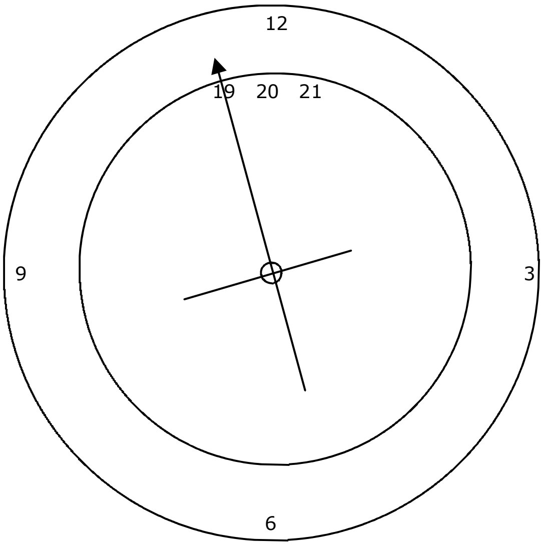

Divide the Star Globe into 8 Octants—and Learn the Stars in Each

The three dividing great circles are the Equator, the Ecliptic, and the Milky Way Galactic Plane.
A star may be North or South of each line, or on one or more of them. The lines on the backside of the globe are shown dashed.
The labels refer to the three lines in the order given. Six of the octants show in the front view with the two in parentheses on the back.
Verify each label before you start. E.g. NSN = North of the Equator, South of the Ecliptic, North of the Milky Way.
Compare with a good star chart e.g. out of a large atlas, as this is only a memory aid.

The Milky Way Crosses the Equator
at the centre of
Procyon-Sirius-Betelgeuse

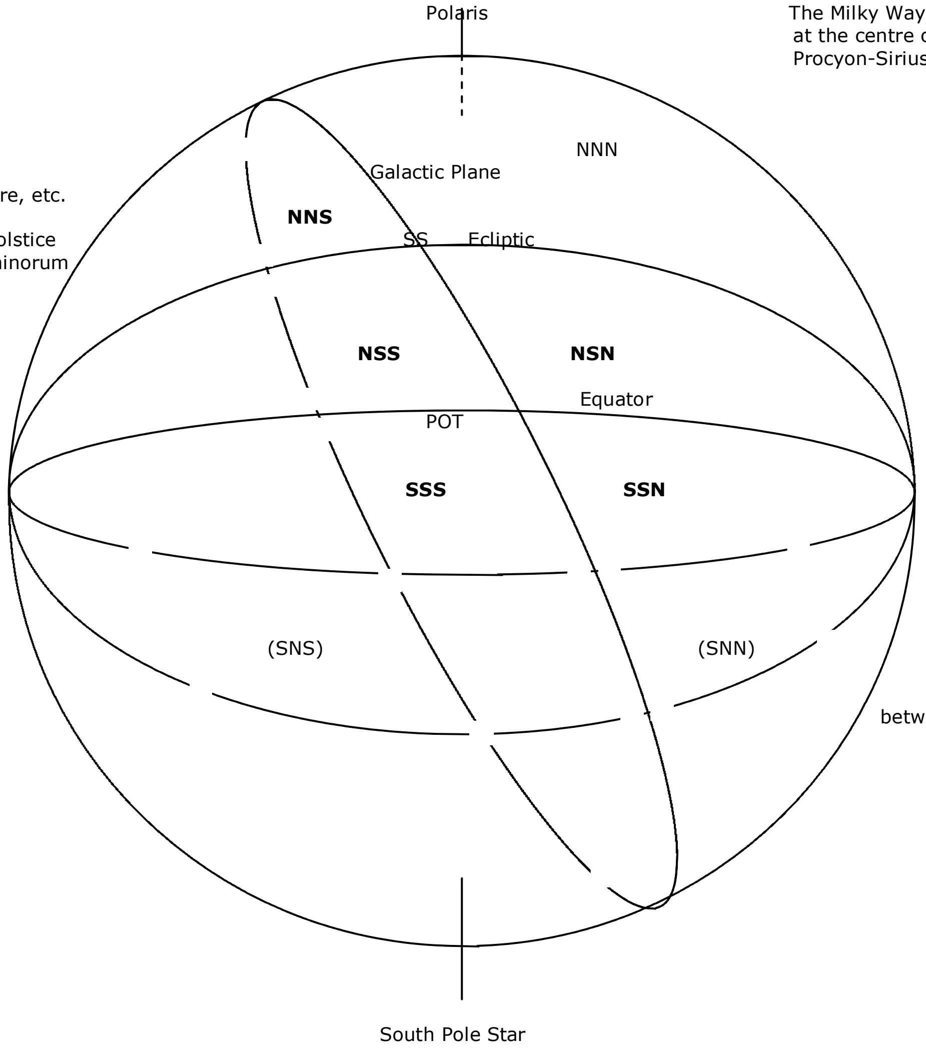

NNS—The Great Square, etc.

Summer (Northern) **S**olstice
Elnath to Gamma Geminorum

NSS
Aldebaran
Betelgeuse
Bellatrix
Menkar

SSS
Pot
Big Dog
Canopus
Diamond Cross
False Cross
Acamar
Achernar (central)
Magellanic Clouds
South Galactic Pole
South Pole Star
Diphda
Ankaa
Fomalhaut
The Crane
Peacock
The Archer
Scorpion's Tail
Atria

NSN
Procyon
Gamma Geminorum

SSN
The Centaur
Antares
Spica
Corvus
Suhail
Alphard

between SSS & SSN
Southern Cross
Southern Pointers

The Southern Figure of Eight—Two Star-Circles—joined by the backbone of Canopus—Achernar—Fomalhaut, a straight line

(not to scale; consult a star chart to verify each star, then find them in the sky)

Fomalhaut
Achernar
Canopus
Sirius |
Orion's belt | these 3 are also straight
Aldebaran |
Pleiades
Aries
Great Square

Milky Way
Cn=Pointers (alpha and beta Centauris)
Southern Cross
Diamond Cross
False Cross

Visible in its entirety in October plus or minus a few months.

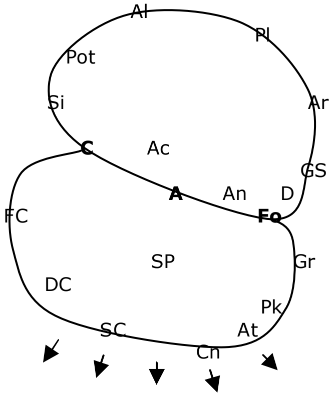

Canopus
Achernar (find Acamar on the way)
Fomalhaut (find Ankaa on the way)
Great Square (find Diphda)
Aries, the Ram
Pleiades, The Seven Sisters
Aldebaran, in the Bull
The Pot, Orion's belt
Sirius, in the Big Dog

Around the South Pole Star
Canopus (find Acamar next)
Achernar (find Ankaa next) *Ach.-Fom.-Pck*
Fomalhaut *form a triangle*
Grus, the Crane, with Alna'ir *to find AlNair*
Peacock then Atria (find Scorpio) *& Ankaa*
Centaur, the two Pointers (find Arcturus)
(find Spica)
Southern Cross (find Corvus, the Crow)
(find Leo)
Diamond Cross, False Cross, Canopus

Divide the Star Globe into 8 Octants—and Learn the Stars in Each

The smaller groups, and the crossover points, are worth learning, because they show you the three lines best.

NNN
Alkaid (central)
Capella
Twins
North Galactic Pole
Big Dipper
Little Dipper
Polaris
Arcturus
Alphecca
Rasalhague
Draco
Vega
Deneb
Most of Leo
(Regulus adjoins NSN)

Summer **S**olstice
Between Aldebaran & Pollux

Autumnal **E**quinox
Between Spica
and Regulus

NNS
Altair
Enif
Great Square
Mirach
Almach
Algol, Mirfak
Aries
Pleiades
Elnath
(Caph, Schedar, Ruchbah)

Winter **S**olstice
Nunki to Sabik

SNN
Zubenelshemali
Sabik
(Zubenelgenubi adjoins SSN)

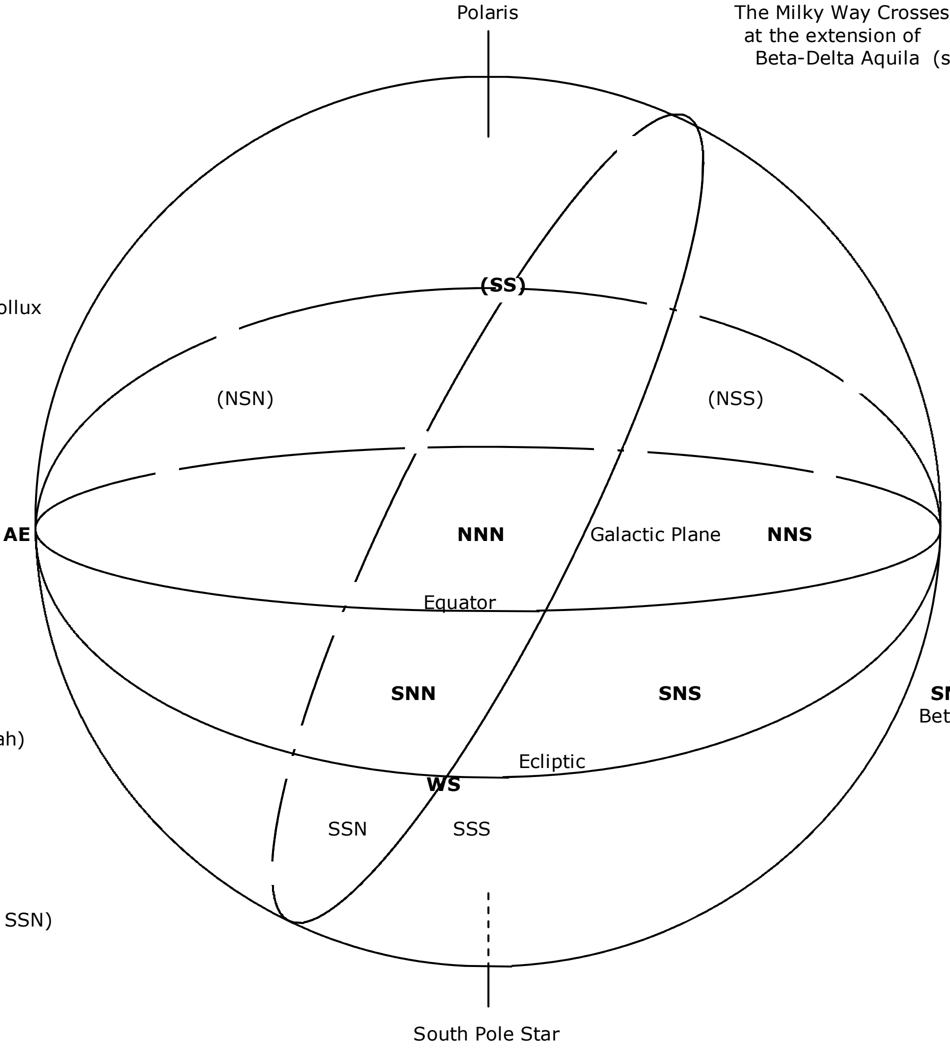

The Milky Way Crosses the Equator
at the extension of
Beta-Delta Aquila (see p.15c)

Spring **E**quinox
between
Diphda & Markab

SNS
Beta Aquarius

The Zodiac i.e. the constellations along the ecliptic. In order of appearance. Not all of them are bright or obvious in shape to look at.

Mnemonic:	Are	Tall	Twins	Cancerous	Leo?	Virginal	Liberated	Scorpions	Shoot	Capricious	Aquarium	Fishes.
i.e.:	Aries,	Taurus,	Gemini,	Cancer,	Lion,	Virgo,	Libra,	Scorpio,	Sagittarius,	Capricornus,	Aquarius,	Pisces
	Ram	Bull		Crab			Scales		Archer	Goat	Water-Carrier	

Morse Code & Rhythm Mnemonics

Dit/Dah/Pause is an extremely flexible system. Adapt it to torchlight, sounds, sewing stitches, body movements… A lot of fun for all the family.
if dit = one time unit; then Dah = three dits long; the di-Dah separator = one time unit; letter separator = three units long; word-separator = 6

	Say…		**Write…**	**Signal…**	Special Meanings (only for reference)
A	an A		. _	di Dah	
B	B bi-bi-bips		_ . . .	Dah di di dit	Fog Horn: I am under tow
C	C is Sea-sick		_ . _ .	Dah di Dah dit	Between ships manoeuvring: I a**c**knowledge your signal
	(Capital letters for a Dah syllable; lower case for a dit)				i.e. You may overtake on the side you indicated
D	Dee did it		_ . .	Dah di dit	Fog Horn: restricted manoeuvrability; deep draught; sailing; etc.
E	egg		.	Dit	I am altering my course to... **Starboard**
F	f' f' Forfeit	f' f' Fore foot	. . _ .	di di Dah dit	
G	Gee Gee git	Go Go get	_ _ .	Dah Dah dit	I intend to overtake you, to Starboard
H	h'-h'-h'-hit		. . . .	di di di dit	I intend to turn completely around to… Fog Horn: Pilot vessel on duty
I	is it		. .	di dit	I am altering my course to... **Port**
J	a J Goes Jay		. _ _ _	di Dah Dah Dah	
K	K for Kay		_ . _	Dah di Dah	
L	an ELL leph'nt /in ELLement /inELLegant		. _ . .	di Dah di dit	(draw out the Ell sounds)
M	My My		_ _	Dah Dah	Fog Horn: "under power, not making way"
N	No en		_ .	Dah dit	
O	Oh Oh Oh		_ _ _	Dah Dah Dah	
P	a P Goes pip		. _ _ .	di Dah Dah dit	
Q	Q Forms a Queue		_ _ . _	Dah Dah di Dah	(to the tune of "Here Comes the Bride"—Queue up to get married!)
R	an R dits		. _ .	di Dah dit	
S	s-s-s	ess ess ess	. . .	di di dit	I am under Astern propulsion
T	Tea		_	Dah	Fog Horn: "under power, making way"
U	it is U		. . _	di di Dah	
V	v' v' v' Vee		. . . _	di di di Dah	(to the tune of Beethoven's V)
W	a White Wine	a World War	. _ _	di Dah Dah	
X	EX mimics P	All ex's Hurt	_ . . _	Dah di di Dah	(drawl the EX, to lengthen it in the first, not in the second)
Y	Why is Y Y?		_ . _ _	Dah di Dah Dah	(say the 'is' quickly)
Z	Zee Goes zi-zit	Zees Go zi-zit	_ _ . .	Dah Dah di dit	I intend to overtake you, to Port
Ü	it is U too	(this pairs to Zed)	. . _ _	di di Dah Dah	(U umlaut or U diaresis. Invent a meaning for it.)

Hint: For Learning, repeat each letter thrice.
By the time the other person recognises the first, the second has finished and he is waiting to check it out against the third.

0	No No No No No		_ _ _ _ _	
1	it Carts Four More Dahs		. _ _ _ _	
2	it is Two Two Two		. . _ _ _	
3	di di di Makes Three		. . . _ _	
4	di di di di Four	pitter patter 4 (your puppy's feet)	. . . . _	
5	fit 'em in a mit	(your five fingers, that is)	. . . . .	
6	Five si– si– si– six		_	
7	Five Seeks seven dits	(=567dits drawled on the 6)	_ _ . . .	
8	8 8 8 bi bits	(8 binary bits in a byte)	_ _ _ . .	
9	9 9 9 9 nips		_ _ _ _ .	
	Long dash	Daaaaaaaash	_______	e.g. (Fog Horn or car horn) before you enter a blind bend!
	Error	it isn't a bit of it	. . . .	Between yachts: "I do not understand/agree with your intention/action"
:	Colon	Now Here Comes in a list	_ _ _ . . .	

Learn the next ones in pairs i.e. swap dit for dah
. a STOP a STOP a STOP (i.e. don't pronounce 'stop' as short) . _ . _ . _
; SEMi SEMi SEMi (i.e. drawl the first syllable) _ . _ . _ .
These 2 shapes are convex to the right
, Go Go comma Go Go Dah Dah di di Dah Dah _ _ . . _ _
? did he Dah Dah did he? did 'e Blah Blah did 'e . . _ _ . .
These 2 come in pairs normally
"" a Quote is a Speech bit di Dah di di Dah dit . _ . . _ .
() Here's the Blah Blah bit Here There's the Blah Blah bit There _ . _ _ . _
These 2 are in the middle of a word ..hyphen and apostrophe
- 'Aitch in the middle Daash ('Hyphen' starts with an 'aitch) _ . . . . _
' it Saves All Those Hard bits di Dah Dah Dah Dah dit . _ _ _ _ .
/ Slaash t' diVide it (Drawl the first vowel) Slaash is a Slaash bit There was the Whole bit _ . . _ .
Á an A Sharp an A You could use it for "Repeat": re Peat Whole re Peat . _ _ . _
É it is E a-cut(e) . . _ . . cute is cut short (You could use it for "Bad Reception": cannot Hear the bits)
Ñ Nya Nya ny' Nya Nya _ _ . _ _ (with thumb to nose, fingers waving) "ny" sound (I use enya for "Do You reCeive Me")

SOS a bit of 'Save Our Souls' isn't it? (no letter breaks; repeat regularly) . . . _ _ _ . . . . . . _ _ _ . . . (Life & death only)
XXX Come quickly Come, Come quickly Come, Come quickly Come All ex's Hurt All ex's Hurt All ex's Hurt (When you Need urgent help)

Signals procedures, developed by hard experience, short circuit a world of miscommunications — stick to them. I.e. learn them too.
I'm Looking For Talk; CQ CQ CQ; Calling Calling All Radios; Hear ye Hear ye All AuditOrs; Tune in Tune in Lend Me y'r Ears
From; D, E; De = French for 'from' Heard of it ; fr'm Dah-di-dit dit
Go Ahead; invitation for anyone to transmit; K Call in Now
Break In B, K Break in to it; Make a Break ("BK DE (identify yourself)" = "let me use this channel")
Starts; Attention; Message Coming Start it Start it Now Start of My bit Now _ . _ . _
End of Paragraph; Pause; new thought; double dash Pause if it is New _ . . . _
Wait a Wait f'r a bit . _ . . . You could invent "Speed up All Your Work" to pair with it.
Over; invitation for your contact to transmit O ver Your Turn stop _ . _ _ . (to start a person-to-person contact: ###DE### Over)
End of Message in traffic; not end of contact; not clear of station the End of My bit; con Clude con Clude it . _ . _.
R; Roger; Received and Understood all of Message
End of Work; but still on air; clear of contact it is the End of Work (to the tune of… here we go Out to Play) . . . _ . _
Clear of all contacts; shutting down; Clear of Station; C, L Clear it Clear it, b' Clear of it
Add your own signals for procedural or other matters. E.g. "Please repeat my message back to me for a check" "Adjust the beam left"
Ä an A an A . _ . _ an A 'm Laut It pairs to C Invent a use for it like "i Don't a-Gree" to be opposite to C, I Agree
Å an A an A too . _ . _ _ an A an Aang Strom
Xhi Khi Khi Khi Khi _ _ _ _ K-H-Like-H soft Kh sound (I use it for "to" i.e. T+O, after DE "from") cf Ree Call All Souls
Ö Oh Oh Oh dit _ _ _ . Goe– Goe– Goethe It pairs to V You could use it to say Sorry

North American Morse is somewhat similar. There is also a Japanese version. Morse is no longer supported as International.

Semaphore Signalling Code & Mnemonics for it

Semaphore is a clockwise system, as viewed by the decoder, and it starts from below.
When the two hands can't be seen, that is the neutral position, used for a… **Pause**
The black square represents the signaller's body; the line(s) are his arm(s).
The numbers 1-9 and 0 are the same as A-I and J. The first 7 single-arm positions are shown…
You can tell the difference, because the signal "numbers coming" precedes a number,
and a "pause" follows the end of the numbers.

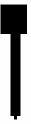

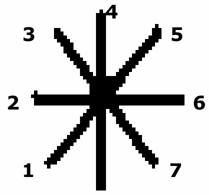

One arm, pointing up, therefore is either D or 4,
Think of God as the 4th Dimension to your life, and remember 'up' as "**4D**"
Or if it sticks in your memory, think perversely, that "**Up is fo(u)r D.own**!?"
Learn the letters in symmetrical pairs.

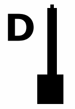

Here is a "**B F**lat" mnemonic. One arm flat is either a B (-flat) or an F (-flat)
"2 be or not 2 B?" and a "**F Sixth** chord" might help you remember the numbers.

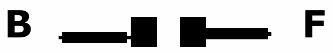

Wave the C to say "Replying". Wave the E to signal "Error"
(cf Morse "C" for "Acknowledge")
The *odd number* positions are diagonals.
1,3,5; 7… **One is an ACE** isn't it? And Dia**g**onal has a **G** in it doesn't it?
a G7 chord to follow the F **F**lat minor **6**
ABC = 123 is easy enough to see, and **5E follows 4D** makes the E position easy

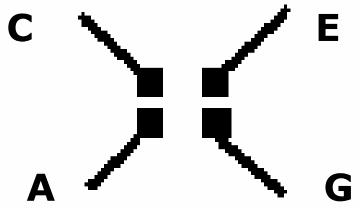

2 Hands…

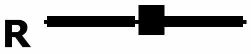

"Roger" is simply an indispensable signal to learn… it means "Yes, I understand, OK, Received"
For Learning: Acknowledge every letter understood, with a Roger

2 Hands Straight:
Think **"Long**, then **Left down = Slash down** or **L**eft **up = L**ift **up"** (referring to **their L**eft arm)…

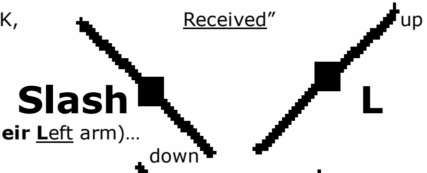

2 hands in a catching formation, above the head…
Shout **"Catch iT…"** and think **"..Tee—numbers coming"**

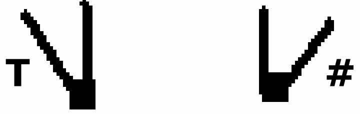

2 Hands in a slip-catching formation, to one side…
Shout: **"HOWZ"**at!

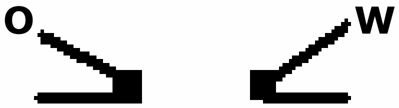

You are meant to wave black and white flags around,
but usually you make do with opening your palms.
Only be careful to position yourself against a good
uniform background.

'aitch rhymes with eight

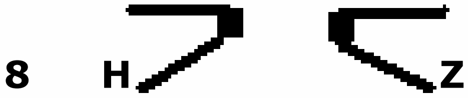

2 hands at a right-angle…
"Ya gotta have the **Right P**y**J**amas" ("PJ"s)
J, the tenth letter, is used for Zero ABCDE FGHIJ

PJ is not in alphabetical order, because in the Semaphore Alphabet J comes after #

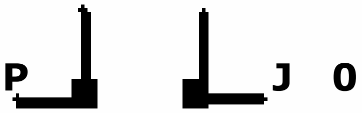

2 hands at a right angle, Vee-shaped, up or down…
Can you see the **shapes** of the **lower case letters n and u** in them?

To attract attention wave U, to say "Message Coming"

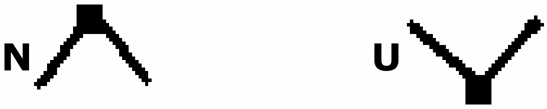

2 hands, like a Roman Numeral Five on its side…
I and **X** are **Roman Numerals too.** The I is used for numeral **9 (IX)** in semaphore
The X looks like **the right-half of the X**

So it sort of goes, left to right "nine, ten"

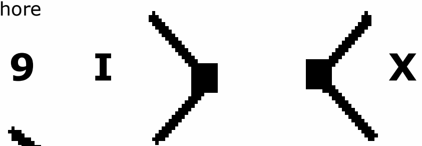

2 hands flapping and banking. I.e. a shallow Vee formation…
"MY Queue**S"** is the clockwise mnemonic
But YQ is not in alphabetic order (like PJ, the only two cases).

You can use Q for "Query" - What did you say/mean?

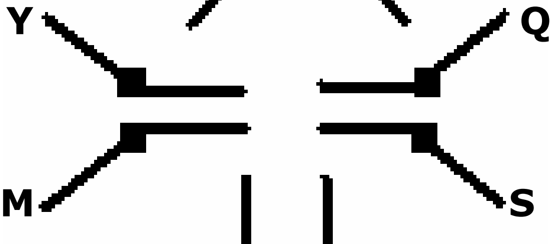

From the sky, down to earth come… **"K**ilo**V**olts" (or **"K**iller**V**olts")

You can follow the clockwise logic of Semaphore if you leave out the JVWXZ till last
Pause, ABCDEFG, HIKLMN, OPQRS, TUY/, #JV, WX, Z

Make each in turn to see the progression.

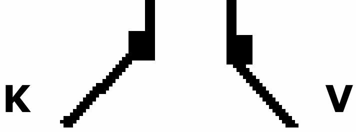

Appendix: More symbols for map-making

Useful for mud-maps, or for track notes.

The first symbol in a strip map column should show the *reason* for the stop. This saves a lot of words and is instant to read.

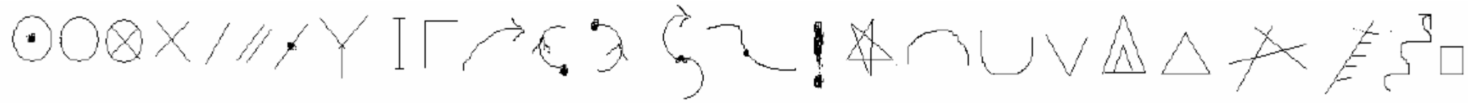

In turn these are:

A Context Entry; Surroundings; Overall Perspective. E.g. catchment and drainage. It's a chance for a general note.
A Stop. Expect a reason like 'Lunch'.
A Location. Expect a description.
Crossing. This might be a crossing of a ridge, river, road, canal, boundary. Expect notes on what is to left and right.
Leg. The first of the 2 main reasons for an entry: either a new subsection (leg) or the location *between* two sections in its own right (next symbols).
Overlap-section between legs. This is a memory aid, to sequence subsections, to remind you how individual sections join to each other.
Change of legs; new subsection; = notes on *the changeover*, rather than on the next section.
Junction (Y for Y junction, but used also for other junctions). This might be a 'natural' junction between ridges or creeks; or an intersection of paths.
Viewpoint; Change of view; First glimpse; Last glimpse (I for 'Eye')
Corner
Bend
Curved section beginning; Arrow shows clock-sense. This one is a Right-hand bend.
Curved section ending; Arrow should show same clock-sense as previous one. This one ends a left-hand bend.
Change from left-curve to right-curve
Break of slope
Feature
Special feature
Ridge; Crest
Saddle;
Valley; creek-bed; Gutter
Tent; Campsite
Peak
Fix
Cliff
Creek
New grid square

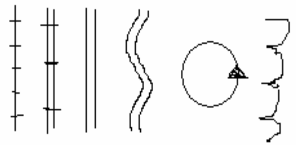

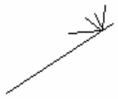

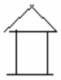

Boundary walks: fenceline; railwayline; road; path; circumnavigation; shoreline
Use symbols which work for you, e.g. Rise or Set direction. Add your own, e.g. Hut

The second symbol ought to show the main direction clue, followed by subsidiary ones as a third symbol

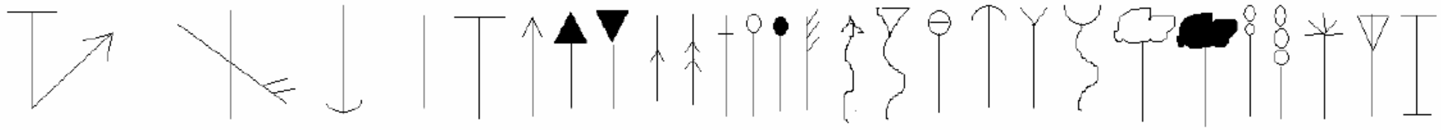

Trending half left from that landmark (write the name of the landmark next to the arrowhead). Useful when following a windy path.
Maintaining half right of downwind;
Walking the crest down ridge — Progress is assumed to be 'Up'wards on the page.
Leg.
Trend direction (for when the terrain is up and down, left and right) ('T' for Trend).
Named landmark; General purpose arrowhead; add the name and the bearing
Peak
Anti-peak direction (e.g. as mentally marked on the skyline opposite the peak)
Up-Slope.
One arrow for 'below average'; Two arrows for 'average'; three arrows for 'above average' — low; medium; high. Use this for other things too.
Compass needle North, modelled after the symbol for North
Up-Sun
Shadow; Down-Sun
Wind source; Upwind. Use the 1-2-3 code for strength. This one is 'above average'.
Down-stream river flow
Up-stream (to a peak) (useful when 'downstream' is behind you)
Moon direction
Up-Ridge; The direction of the highest skyline through the vegetation, if that is all you can see.
Down-valley. The direction of the lowest skyline when that is all you can see.
General Drainage trend (down the valley to the river mouth)
Up-Weather; Low cloud source direction; Cloud Source (when only one set of clouds is visible)
Down-clouds
Alto Clouds source direction (make the circles fluffier than I can show)
Cirro-clouds source direction
Glow at night
Radio-null (from the symbol for an antenna)
Trend of some feature like the coast — write in the details.

1,2,3 for size and importance of junctions…

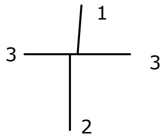

The notebook might be set out as follows: The first line across a narrow notebook is in informal columns; the second is for written notes.

Time ;	**Reason** ;	**Trend** ;	**Directions** ;	**Times** ;	**Distances** ;	**Grid**
Fill in a **number**	Use the top **symbols**	Use the main clue	make a rose of others	elapsed time	total elapsed	e.g. A5
if you don't				leg time	last leg length	
have a watch	Start with an 'Overall'			lost time	lost distance	
				total time out	total diversions	
Sequence is a				net walking time	net progress	
primary need				(Use abbreviations, e.g. NWT, or dedicated columns)		
				(Do your calculations here at leisure)		

Now write *across* the next line(s) the **name; mnemonic; description; comments**
The description of a cross-country leg could include the side or flank, hand, clock-sense, break, steepness, difficulty, curviness, hilliness, perversity, terrain, contours, topography, view, what you can't see, skyline, traffic intensity, track-markers, their age, rock type, soil type, vegetation, biology, colours, peculiarities, reference to another page of the notebook to a diagram.
The more you include in the symbols, the less you need to say, and the easier it is to read later.

Appendix: A List of Star Coordinates.

100 brightest stars
Declination is equivalent to Latitude. –nn = South; +nn = North
Siderial Hour-Angle is equivalent to Longitude West of Zero.

Common Name (s)		Dec.	S.H.A.	Constellation
Miaplacidus		-69.72	221.70	b Carinae
Atria		-69.03	107.83	a Trianguli Australis
Acrux		-63.10	173.35	a Crucis
Rigil Kentaurus/Toliman		-60.84	140.10	a Centauri
Hadar	/Agena	-60.37	149.04	b Centauri
Mimosa		-59.69	168.07	b Crucis
Avior		-59.51	234.37	e Carinae
Aspidiske	/Tureis	-59.28	220.72	i Carinae
Achernar		-57.24	335.56	a Eridani
Gacrux		-57.11	172.21	g Crucis
Peacock		-56.74	053.60	a Pavonis
k Velorum		-55.01	219.46	k Velorum
d Velorum		-54.70	228.82	d Velorum
e Centauri		-53.47	155.02	e Centauri
Canopus	2nd brightest	-52.70	264.01	a Carinae
Muhlifain		-48.97	169.62	g Centauri
Regor		-47.34	237.61	g Velorum
Alna'ir		-46.96	027.95	a Gruis
b Gruis		-46.88	019.34	b Gruis
Suhail	/Al Suhail	-43.43	223.01	l Velorum
Akrab/ Sargas	/Rastaban Sc.	-43.00	095.67	theta Scorpii
Ankaa		-42.31	353.43	a Phoenicis
eta Centauri		-42.16	141.12	eta Centauri
Naos		-40.00	239.10	z Puppis
k Scorpii	(not Akrab)	-39.03	094.38	k Puppis
Shaula		-37.10	096.60	l Scorpii
pi Puppis		-37.10	250.71	pi Puppis
Menkent	(not Menkar)	-36.37	148.33	th Centauri
Kaus Australis		-34.38	083.96	e Sagittarii
e Scorpii		-34.29	107.46	e Scorpii
Ascella /Nushaba		-29.88	074.36	z Sagittarii
Fomalhaut		-29.62	015.59	a Piscis Austrinis
Aludra		-29.30	248.97	eta Canis Majoris
Adhara		-28.97	255.34	e canis Majoris
Antares		-26.43	112.65	a Scorpii
Wezen		-26.39	252.90	d Canis Majoris
Nunki		-26.30	076.18	s Sagitarii
Dschubba		-22.62	119.91	d Scorpii
Diphda	/Deneb Kaitos	-17.99	349.11	b Ceti
Mirzam	/Marzim	-17.96	264.33	b Canis Majoris
Arneb	/Arnebo	-17.82	276.82	a Leporis
Sirius	/Dog (brightest)	-16.72	258.72	a Canis Majoris
Sabik		-15.72	102.41	eta Ophiuci
Spica		-11.16	158.70	a Virginis
zeta Ophiuci		-10.57	110.72	zeta Ophiuci
Saiph		-09.67	273.06	k Orionis
Zubenelschemali		-09.38	130.75	b Librae
Alphard		-08.66	218.10	a Hydrae
Rigel	not Rigil K.	-08.20	281.37	b Orionis
Alnitak	S. belt	-01.95	274.80	z1 Orionis
Alnilam	Mid belt	-01.20	275.94	e Orionis
Mintaka	N. belt	-00.30	277.00	d Orionis

Common Name (s)		Dec.	S.H.A.	Constellation
Polaris	N. Pole Star	+89.26	(322.06)	a Ursa Minoris
Kochab		+74.16	137.32	b Ursa Minoris
Alderamin		+62.59	040.35	a Cephei
Duhbe		+61.75	194.07	a Ursa Majoris
Navi	/Cih	+60.72	345.82	g Cassiopeiae
Ruchbah	/Ksora	+60.24	338.55	d Cassiopeiae
Caph		+59.15	357.70	b Cassiopeiae
Schedar	/Schedir	+56.54	349.87	a Cassiopeiae
Merak		+56.38	194.53	b Ursa Majoris
Alioth		+55.96	166.50	e Ursa Majoris
Mizar	(+Alcor nearby)	+54.93	159.01	z Ursa Majoris
Phecda		+53.69	181.54	g Ursa Majoris
Eltanin	/Etamin	+51.49	090.86	g Draconis
Mirfak		+49.86	308.92	a Persei
Alkaid	/Benetnash	+49.31	153.12	eta Ursa Majoris
Capella		+46.00	280.83	a Aurigae
Deneb	(not Denebola)	+45.28	049.64	a Cygni
Menkalinan		+44.95	270.12	b Aurigae
Almak	/Alamac	+42.33	329.02	g Andromedae
Algol		+40.96	312.96	b Persei
Sadar	/Sadir /Sadr	+40.26	054.45	g Cygni
Vega		+38.78	080.76	a Lyrae
theta Auriga		+37.22	270.07	th Aurigae
Mirach		+35.62	342.57	b Andromedae
Aljanah		+33.97	048.45	e Cygni
Castor	N. twin	+31.89	246.34	a Geminorum
Alpheratz	/Sirrah	+29.09	357.90	a Andromedae
Elnath	/Nath	+28.61	278.43	b Tauri
Scheat	/Schert	+28.08	014.06	b Pegasi
Pollux	S. twin	+28.03	243.67	b Geminorum
Izar		+27.07	138.75	e Bootis
Alphecca	/Gemma	+26.71	126.33	a Coronae Borealis
Hamal		+23.46	328.20	a Arietis
Sheratan	/Sheratain	+20.81	331.33	b Arietis
Zozma	/Zosma	+20.52	191.47	d Leonis
Algeiba		+19.85	205.00	g1 Leonis
Arcturus		+19.18	146.08	a Bootis
Aldebaran	(not Alderamin)	+16.51	291.01	a Tauri
Alhena		+16.40	260.56	g Geminorum
Markab		+15.21	013.82	a Pegasi
Denebola		+14.57	182.73	b Leonis
Rasalhague		+12.56	096.27	a Ophiuci
Regulus		+11.97	207.90	a Leonis
Enif		+09.88	033.96	e Pegasi
Altair		+08.87	062.31	a Aquilae
Betelgeuse		+07.41	271.20	a Orionis
Bellatrix		+06.35	278.71	g Orionis
Procyon		+05.22	245.17	a Canis Minoris
Vernal Equinox, defined as fixed		00.00	000.00	'the first point of Aries'

Figures are from Epoch 2000.0.

Precession of the equinoxes:

Stars drift about 1° of position relative to the (0,0) reference star, in **72 years**, because it moves to the **WNW along the Ecliptic**. See p 33b
The third figure will therefore become off by one after 10-20 years. E.g. Mimosa used to be –59.59; Atria used to be 108.3W
One degree of latitude or declination = 60n.ml. =111km; so a maximum change in **0.1° over 18** years represents 11km, or about 0.6km per year, max.
(The actual change in Declination depends on its relationship to the Northern Spring Equinox point — see below.)
One degree of SHA = 60n.ml. at the Equator; half that at 60° Lat. ; **1° reduction in 80 years** is about 3 seconds later in timing per year, or 1.4 km max, at the Equator.

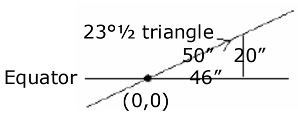

Ecliptic The first point of Aries moves along it at about 50 seconds of arc per year (5/6° in 60 Years)
Declination change of 1/3 minute of arc per year (20 secs of arc)
SHA change of 46 seconds of arc per year $20^2 + 46^2 = 50^2$

If you set your star watch by the actual visible stars each year, you eliminate E-W 'precession' component, and are left with only the drift *in declination*.
So it is of some worth to be able to predict how that rate of change in declination depends on SHA.
Obviously Stars to the North of Aries lose Northern dec, at the full rate. Stars to the South must *gain* Southern Dec.
Stars *opposite* Aries (SHA=180°) do the opposite. Stars exactly in between are *unaffected* (SHA=90° or 270°).
The rate for all cases follows **the cosine of the SHA**. E.g. stars at 60**°** SHA will have 50% of the full dec adjustment (of 1/3 n. ml. per year).
Take care with Polaris, since it is too close to the Pole.

Appendix: Drawing Great Circles

First, bisect a globe.
Two equal halves.

Put the halves back together.
Notice the ellipse.
The backside is dashed.

Draw it symmetrically around the centre.
Use *flat* ends to the ellipse.

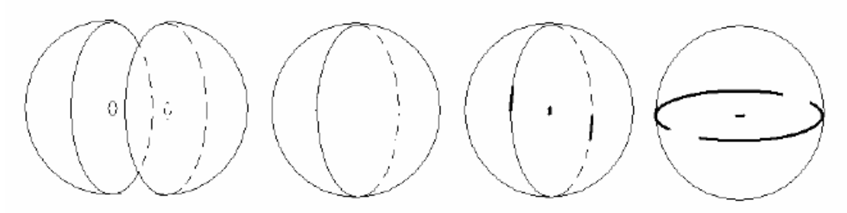

Choose the clearest perspective
E.g. 1. Rising/ setting is 90° away, so put yourself in the centre, surrounded by a 90° circle. Z is used for your Zenith — overhead.
E.g. 2. Look at the Equator side-on to make it and the latitude circles straight.
E.g. 3. Centre on the pole, to show up the polar distance. P is used for the Pole.
E.g. 4. Curve away from the centre. If in doubt, draw the whole ellipse and discard the unwanted parts.
5. A Great-circle can be drawn as a straight line. X is used to mark the position of the Star/ Sun/ Moon/ Planet
6. Draw the 'Vertex' of the course as a tangent to the maximum latitude small-circle.
7. The vertex can be drawn side-on.

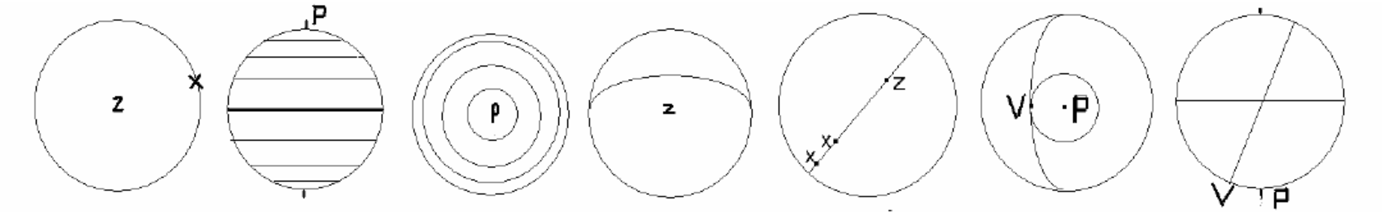

The PZX spherical triangle

8. Deliberately avoid straight lines to show up the curve of the triangles.
8a. ZX is the angular distance from you to the star, or the co-height.
8b. PX is the polar distance or co-declination to a point directly underneath the star.
8c. PZ is the polar distance or co-latitude of your position, 90° — lat.
8A. The angle ZPX, or 't', is a longitude difference — between the SHA of the star and your West longitude. A longitude difference is a time delay.
8B. The angle PZX is a horizontal direction relative to P, towards X. Technically called an 'azimuth'.
8C. The angle PXZ is similar but only meaningful if there were someone under the star, looking toward you.
9. A, B, C and a, b, c are used to label angles and sides of any spherical triangle; It's up to you to know which ones label what.
10. Draw a full 180° to show up the supplementary angles.
11. Draw the Equator cutting East and West of you, 90° away. Mark any right angles, and the complementary angles.

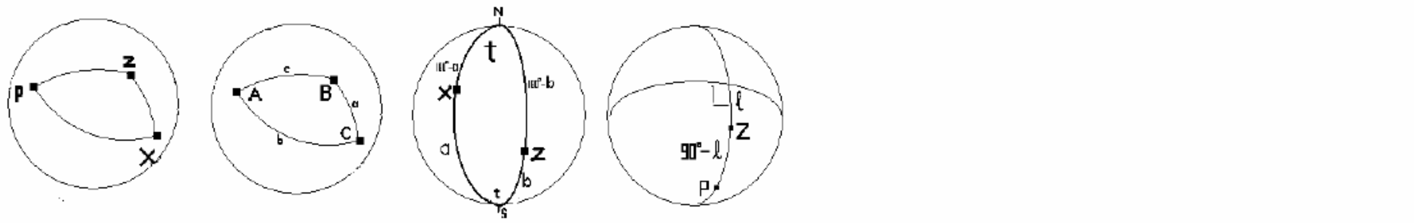

The PZX navigational triangle conforms to a handful of standard **fundamental spherical trigonometrical formulae**…

The Angle **Sum**: a+b and A+B are both greater than, or both smaller than, or both equal to, 180°
The **Sine** Rule: sina/sin A = sin b/sin B = sin c/sin C = constant. Ambiguity (about obtuse angles) can be resolved by the angle sum rule.
The **Polar** Triangle Rule: abcABC can all be swapped for the supplement of the opposite members.
E.g. 'a' becomes '180°—A'; 'B' becomes '180°—b', and the resulting formula is still valid. E.g. the sine rule simply inverts.
The **Cosine** Rule: cos a = cos b cos c + sin b sin c cos A
The Polar Cosine Rule: cos A = — cos B cos C + sin B sin C cos c
The **Four Parts** (in a row) Rule: e.g. AbCa or aBcA in the triangle aBcAbC or aCbAcB. You can find the 4th (outer) part from the first 3.
cos (**inner** side) **cos** (**inner** angle) = s**in** (**in**ner side) c**ot** (**ot**her **side**) — **sin** (**in**ner angle) c**ot** (**o**u**t**er **angle**)
If you want an inner part, use the sine rule and angle sum rule to find the 5th part, then go to the half-angle rules.
The **Five Parts** Rule: sin 1 cos 2 = sin 3 cos 5 + sin 5 cos 3 **cos** 4, where 4 is a side. It gives you any but part number 3, from the other 4.
The Polar Five Parts Rule: sin 1 cos 2 = **—** sin 3 cos 5 + sin 5 cos 3 cos 4, where 4 is an angle
The **Half-Side** Rule, to find the *middle* of 5 parts, when the middle part is a side: halve each side and angle first:
tan (half side) = tan (half sum) cos (half sum) / cos (half difference) ; tan refers to sides, sin and cos to angles
= tan (half difference) sin (half sum) / sin (half difference) but this introduces order-ambiguity because of the sines.
The **Half-angle** Rule to find the middle of 5 parts when it is an angle
cot (half angle) = tan (half sum) cos (half sum) / cos (half difference)
= tan (half difference) sin (half sum) / sin (half difference)

Make North and West positive, and East and South negative, and the formulae automatically work on PZX for angles up to 360°.

The formulae simplify where a side or an angle is 90°, since sin 90°=1 and cos 90°=0.
Napier's Rules…
The rules for **right-angled** triangles and **right-sided** triangles are simplified when the 5 other parts of the triangle laid out **in a pentagon**, and the 3 parts not adjacent to the right angle are changed to their **complement**.
In addition the part opposite a 90° *side* must have an extra **negative** sign.
E.g. if A is the right-*angle*, use (90°—a) as part 3, (90°—B) and (90°—C) for parts 2 and 4; if a is a 90° *side*, use A-90° for part 3.
Then… Sin of any one part = t**a**n t**a**n of **a**djacent parts or = c**o**s c**o**s of **o**pposite parts (in the pentagon of changed parts).
Ambiguity must still be resolved by special rules: Problems occur when there are 2 or more right-angles in the triangle.
'A' and 'a' (opposite parts in the triangle, with matching letter names) are both less than or both more than 90°. Likewise for Bb and Cc.
If one part and any part not opposite it (i.e. not A and a, as above) are on the *same* side of 90°, so are all the other mismatched pairs.
If one part and any part not opposite it (i.e. not A and a) are *not* on the same side of 90°, neither are any of the other mismatched pairs.
These refer to the triangle parts, not the altered pentagon parts.

Appendix: Calendars in the Sky — the first glimpse and last glimpse of stars

~ Firstly, What's wrong with **sunrise/sunset time**?

- To know what time of year it is, you could note the ordinary standard Sun time of sunrise/set at *that* place (e.g. from home).
  —But this is insensitive (maximum shift is about 1 minute a day), useless around solstice, latitude & longitude dependent & requires a watch

~ You need to use apparent *star* time; no watches, to **locate the Sun's position on the Zodiac relative to the stars**.
— i.e. 'what stars are up' at midnight, or at twilight, or what stars are invisibly up at mid-day or sunrise or sunset.

- The Sun creeps Eastward about 1° per day along the Ecliptic (through the middle of the Zodiac belt).
- That quadruples the sensitivity to 4 minutes change a day, steady year-round.
  It is only affected by the Equation of Time, which is a minor variability, 1/4 min a day, max.
- The Sun will 'blot out' each Ecliptic star for about a month at a time.
  That's a rough way of finding the month: 'I can't find The Twins' (July).
- 'Nautical twilight' is when the Sun is 10° below the horizon (about 1 hour, in Tasmania), dark enough for the bright navigational stars to be visible while the horizon still shows clearly enough to be used with a sextant;
  So bright stars 10-15° away from the Sun begin to be visible. That means that a bright Ecliptic star might only be invisible for 3 weeks.
  [Civil twilight is 5° below, requiring lighting; and astronomical twilight is 15° below, dark enough to be called 'night', and to use a telescope]

E.g. 'What stars are up at midnight', uses the 'apparent star time' of the anti-Sun, as it were. You need to bisect sunrise and sunset, and be up then

- Find which star is N-S of you (the 'local apparent star time') at 'local apparent midnight' (= opposite noon — i.e. the anti-Sun stars).

But that is difficult — you need to know the apparent noon (which is longitude dependent), not the clock noon — and it still requires a watch.
So how can you do better?

- You could note **the clock-time of any particular star event**. E.g. follow 'Rigel over Capella' during Summer. E.g. Antares at 10° high.
  This is *simple, and accurate,* but it would only work for that location, no other, since longitude alters the clock-time of appearances.
- You could note **the clock-time *delay*** between any particular star-event and either sunrise or sunset. **This is a good method**, since it pinpoints sunrise or sunset more accurately than saying 'At twilight, when the stars come out, or when they fade out'.
  For best results, choose which star-event directions to use, as explained below, so that people at all latitudes see the same appearances.
  - This pinpoints the star-time—it tells you where the stars are—in relation to the apparent Sun-time (by tying it to sunrise or set).
  - For Rise/Set observations, you need a good horizontal horizon (or other fixed height-of-star) for star-rise/set to be latitude-independent, and location independent; but at home you simply need a repeatable point of view of a given horizon.

~ But to get back onto the main topic: The use of a watch can be eliminated by using twilight 'glimpses'...

  - For Twilight observations, you need dependable eyesight; & a clear sky, with no Moon or light pollution, or haze, to be repeatable.

The idea of a 'glimpse' is that it is positively there or not. And it doesn't need a watch. There are 10 or so varieties...

- If you see a star *before dawn*, it's positively there, or if you can't see it *after dark*, it's not there.
  - Look to the East before dawn and you will see **a new star appearing** for the first time in the season — a first glimpse.
    You will then lose sight of it due to daylight. This is the standard star-calendar event.
  - Look to the West after sunset and a star will have **disappeared** for the rest of the season — a last sighting.
    You glimpse it briefly, before it moves into the sunset glow, and is lost to sight for weeks.
  - Examples: The last glimpse of Spica is just after the September Equinox; Regulus is first seen just ahead of the September Equinox;
    The Equinox is between the two, and both are on the Ecliptic, about 25° each side. That's 25 days for the Sun's movement.
    The first glimpse of Antares is around Christmas time, since it is about 20° ahead of the solstice;
    Aldebaran reappears around the winter solstice; The Twins disappear soon after it.
    The solstice is between these two, 25° away from each.

  You need to check the exact dates of appearance and disappearance because they vary with the exact latitude—this is a semi-local calendar.
- If *a star-pair* has tilted *out* of alignment by the time the stars come out, you simply *don't* see it occur;
  or if it doesn't align before dawn, you can't see it 'happen'.
  - Look pre-dawn for the first date on which you can glimpse **a star pair reach its proper alignment**, before dawn takes over.
  - Look post-sunset for the *last* date on which you can see a star pair attain its right alignment, after which it is lost for a month or so.
- Each day a Western star gets closer to the Western horizon from above. Each day a rising star gets earlier to the Eastern horizon, from below.
  - Look for the first glimpse of a star clearly **rising or setting** *pre-dawn*
  - Look for the last glimpse of a star clearly **rising or setting** *post-sunset*.

  If you see it setting, you know it; if you see it rise, you have seen it happen.
- Wait to see whether a star reaches a known fixed **height** (e.g. not too far above the horizon), to make observation easier.
  - Looking at a reflection of a star in smooth water or oil, will *double* its height 'above horizontal', if you want a method to work anywhere.
- Wait until a star coincides with a known fixed **direction** (e.g. 'over the tower'). This makes it location-specific, but may be useful.

These all have the effect of describing the height or tilt or direction of the horizon 'at twilight'.

- A star **overhead** at twilight will show up as a first/last glimpse. You look for it, but it may not 'happen'. This is latitude-specific.

~ You could now combine the 'glimpse' methods and 'crisp timing' to avoid the indistinctness of twilight (but this is more latitude dependent).
Look 'exactly 1 hour from sunset/rise' to crisp up the timing of the 'events'. Any exact time-span at least that long, would suffice.

~ To clinch the star-time, in a latitude-independent way, **use Great-Circle type directions** ('at right-angles to sunset/sunrise' 'a star near the Sun at twilight' 'opposite it') to ensure that people over a wide range of latitudes more or less agree in their observations.
The standard direction is not sunrise/sunset itself but 'where the Sun is *at twilight*' which is 1 hour to the left (SH); *the brightest part* of the horizon.

- identify star-pairs ***at right-angles to the Sun's* direction at twilight**.
- or examine the rise/set of **stars *near the Sun***, especially those ahead or behind it, in its rising and setting path.
- or look for stars rising, setting or simply being visible, ***opposite* the Sun** at twilight.
- You can use any other 'star direction', but it will only apply to that latitude. The best great-circle directions are between NW/SE & NE/SW.

~ But the sensitivity can be further improved, by **choosing *which* twilight** gives the most daily change…

- The nightly change (averaging 4 mins) is not the same amount through the year.
  It can be reduced to 3 or increased to 5 in a place like Tasmania, as the seasonal day-length varies by several hours.
  The change in sunrise or sunset time by up to a minute a day is overlaid onto the steady star progression.

Look at the time of day when the most rapid change is taking place, since you want the most accurate discrimination:

After mid-summer, look for 'first glimpses' in the pre-dawn ;

- Nights lengthen after the summer solstice — more stars rise in the East than set in the West as the night-time gradually dominates
  At dawn, the Sun lingers a little longer under the horizon allowing more stars to rise, in their creeping ahead of the Sun.
  But at sunset, the Sun goes down earlier, like the stars do, and you don't notice much change as to which stars are up.
- So when the nights lengthen, look **before dawn** and try to see any first glimpse this season.
  Luckily the pre-dawn in Summer/Autumn is mild, but it can be quite early. Use the clock-time delay method 'from sunset', not 'to sunrise'.

After mid-winter, look for 'last glimpses' post-sunset ;

- Conversely, after the winter solstice, the increasing day length dominates over the night sky — more stars set and fewer stars rise.
  You will lose sight of stars setting in the West, quite easily, as the Sun will linger before setting, because of the season,
  but find it hard to see a new star rising in the East for the first time that year, because the Sun pops up early.
- So when the days lengthen, to pinpoint the time of year, look **after sunset** for any last glimpse.
  That's not too late at night in Winter/Spring, and it should not be too cold. Or use the clock-time 'until sunrise', to maximise the daily gap.

The overall idea works in both hemispheres, but the optimal choices of when to look are tied to the seasons, not the date.
Use either method, when near to the solstice, or near the Tropics, or at any time you want to.

- Around mid-winter and mid-summer, you can happily use any of the 10 glimpses, or any of the clock-time methods.
- Near the Tropics, also, there is little differential advantage, compared with temperate latitudes, since day-length is fairly constant

~ Keep a special perpetual calendar (a 365 day chart with no days of the week specified) to note down your observations; with room for related seasonal changes, like bird migrations, frosts, plantings.
All this has little to do with navigation, directly, but gives you practice in *following the seasonal movements* of Sun and stars.

About the Author.

I have a big bushy beard from too much bushwalking and ocean sailing, but now that it has grown too grey, it is time to hang up my boots and write it all down.

My dad was a ’plane navigator during World War 2 and as a child I studied his manuals. Then at University I spent more time in studying naval navigational manuals and in bushwalking off track, finding my way over the snow or canoeing across lakes, than in studying for my exams. I used the library to learn spherical trigonometry, then, as soon as I finished Uni, set out across the trackless ocean. I soon found myself inventing techniques to plug the holes in my normal abilities. I returned home years later, not knowing what I was meant to do in life, but as I lay there on my old bed, I knew that if I was to be a navigator I would like to be one like Captain Cook. He experimented with new techniques, and with old — he described native sea-navigation expertise. How *did* they do it? Even now, no-one seemed to know. I immediately had a vision of one star above another, which kept me busy for years on great-circle navigation methods.

I had honed my skills trying to teach boaties and would-be ocean sailors, but my biggest challenge was to teach my wife and children. We became a feral family for a while and walked every track we came across, here and overseas. Only then did I appreciate how much more there is to orienting yourself *to your surroundings*, than simply having a compass.

My next inspiration came while walking the forest on a foggy day. I simply asked how the trees themselves would hold the secret to which way is which. Years of observation and testing later again, and this is the result. I don’t want it to die with me.

I hope you too can now enjoy the magic of the science and love the art of exploring, even teaching yourself, before you too have to stop still, inside the bounds.

Learn a little before it is too late.

Outline

“Navigation Down-Under” is your complete manual of bush navigation and outdoor orientation for when you go off the well-worn track. A free pocket guide to take with you bush-walking is part of the pack.
Do you know how to predict and find direction from the Moon by memorising just two simple numbers? Can you interpret the Australian bush plants for direction? You will find some never-before published new techniques, as well as all you need to know about traditional methods like Lie-of-the-Land.
If you come from the Northern hemisphere, you’ll need the top-down global approach, and can try some special Southern Hemisphere methods of direction keeping.
Sailors will find here a valuable beginning into celestial navigation.
A lot of it is about how to navigate without a compass, how to stay oriented and how not to get lost. It is an essential course in basic bush awareness and orientation, for surety and safety, plus it also throws out the challenge that you can improve your already existing competence very quickly and easily.
For Scouts or emergency search-and-rescue workers, it is a thorough course, advancing beyond the map and compass. Or perhaps you are a tour-guide and need more than course and distance — You’ll get something more, something more like biogeography or ecology.
Each page is a carefully arranged lucky dip of interesting tidbits.

Keywords

Navigation; Orientation; Celestial Navigation; Bushwalking; Map and Compass; Without a Compass

Section 1. BIG-PICTURE TOP-DOWN pp 3-33

General Introductory Points Big-Picture Rule One:

3 Locate a Distant Reference Direction First
3 Visualisation is your Key Skill
3 Big-Picture Orientation, a Mental Layout Acquired Beforehand
3 Compass Navigation Robs You of a Global Perspective
3 The Big Picture is Invisible — Your Job is to Visualise It!

Global Perspective; Earth

3 Compass Navigation Robs You of a Global Perspective
4 Global Re-orientation — Light-Half/Dark-Half
5 A mnemonic for global orientation: "True Polarity"
6 Mnemonic: "Time Tells Turning"
7 The sky turns 15° per hour —Use your thumb and forefinger
8 Global Orientation: Latitude and Polar Distance
9 The Angle of Rising and Setting—is Determined by Your Lat.
10 Global Orientation: Point to the South Pole…
..half its polar distance below horizontal
11 The North Pole is *Perpendicular to the South Pole*
12 Global Orientation: the South Pole Star, and Polaris
13 Pole Stars—"(Latitude)° above and below, South & North"
Point to 2 Points on the Equator — Halfway between the Poles
15 Perspective Distortion — The Sun Moves in a Straight Line!
16 Global Orientation: Half the Arc-Distance below Horizontal
17 Half the Arc-Distance below Horizontal—2 Explanations
18 Example: Point to Tahiti
19 Use Latitude for a Distance Scale (Not 'Long'-itude)
20 The Horizon Rotates to Your Right in the S. Hemisphere
21 Global Perspective — Rotation of the Horizon
22 How to Shift Your Horizon —Its Timing and Height
23 Latitude without Numbers
24 Mnemonic "The Sun Travels the Tropics"
25 Tropical Countries and Places Under the Sun
— Arranged in Time Zones Relative to Eastern Australian Time
— 10 hrs ahead of Greenwich
26 North/South and Noontime by Bisection (of Rise and Set)
27 Finding Magnetic Variation by Bisecting Rise and Set
Find E/W by Delayed Bisection of Sun & Full Moon At Rise/Set
28 Find North from the Combination of the Sun and the Moon
29 Eastern Australian Standard Time Zone
10 hours ahead of Greenwich
Showing Tropical countries and places underneath the Sun
32,33 Overview
128, 129 Tropical countries and places underneath the Sun

Sun Your Best Reference Direction:

4 Imagine the Sun at High Noon — Always
5 Mnemonic: **N**oon Sun Passes **N**orth of My **N**ose Down Under
6 Notice these, Yesterday — to Guess the Time Today
7 Mnemonic: The Southern Summer Sun Sets Somewhat South
8 Mnemonic: "Sun at High Noon — How High?"
8 On Overcast Days, Average the Brightest Sky over 1/4 hour
9 High Noon — How High?
10 Optical Illusion — Beware!
— Use Your Shadow Not the Sun, For Direction
11 The Dark-Blue Polarisation Band Perpendicular to the Sun
11 The Centre of the Rainbow is Down-Sun
12 Your Shadow Rotates with the same clock-sense as the Sun
Shadow-Rise and Shadow-Set is opposite direction to the Sun
12 Mnemonic: Shadow Rise is on the Right; Set is on the Left
12 Mnemonic: Your Shadow Passes South of Southerners
at Southern Noon
13 The Sun Passes Quickly Past North-South Nearer the Tropics
14 Shortest Shadow South—Opposite Highest Noon.
Always Imagine "Shadow-Noon" on your "Shadow South" Side
15 Locate Polaris (& Find N) by Imagining the Centre of Circling
16 " Noon " is Not at 12 O'clock
17 The "Equation of Time" Diagram
18 The Sun Oscillates Seasonally between +/- 23½° Latitude
19 Using the Declination of the Sun
20 Tell the Time by the Compass Protractor—Tilt to the Equator
21 Or Judge the Time By Sight—by Mirroring the Sun Around N
(To Make an Hour Hand on a 12 Hr Clockface)
22, 23 Walking in Phase with the Sun, Wind and Moon
24 Tell the Time of Night by the Star Opposite the Sun
25a Eastern Australian Standard Time Zone 10 hours ahead
25 Tell North from The Time By Using a Protractor
26 Longitude and Latitude from the Sun
27 A Bush Theodolite for Overnight Equal Height Observations
28 Find North by your (non-digital) Watch-Hands
29 Find Any Other Direction Too — Using the Hour Hand
29c String North
30 Locating Anti-sunrise Shortest Shadow Souths
30 Sun Path Norths — Summary
31 The 2 Flaps Method for N.—Lat. & Dec. Angles Separately
31 Seasonal Adjustments For Equinoctial East
32, 33 Overview
132, 133 Eastern Australian Standard Time Zone
10 hours ahead of Greenwich
Showing Tropical countries and places underneath the Sun

Stars

4 The 4 Bright Stars of Orion Point out Where the Pot is Going
4d The Ecliptic
5 The Two Brightest Stars Point out the South Pole Star
6 The Spinning Earth Makes Circular Paths for the Sun, Moon
7 Stars Leave Star-Trails
8 RISE and SET Directions for a (Local) STAR COMPASS
9 Global Star-Compasses e.g. 'Perpendicular to' rise and set
10 Right Around the Horizon, Stars Move to Your Left, Down S.
11 Identify the Stars That Pass Directly Overhead
12 Ways to Locate the South Pole Star Or the N./S. Meridian
12a&13a Global Orientation: the South Pole Star, and Polaris
13 Pole Stars—"(Latitude)° above and below, South and North"
13 South Pole Star Chart
14 Find the Polar Axis by Star-Trail Bisection
15 The Eagle Clips the Equator, South of Altair
16 Equatorial Stars Rise and Set Due East and West
17 Rigel-to-Capella is a good S./N. Pair
18 Use the Southern Cross as an Hour Hand
18 Picture the Southern Cross during the Day
19 Perspective Distortion — Use the Correct Vertical
20 The Southwest Corner of the Great Square is N. S. E. W.
21 The Southern Summer Rectangle And a Triangle Pointing S.
22 Navigating by the Shape of the Night Sky
23a Latitude without Numbers
23 Matching Sky to Sky — Move Only Your Horizon
24b Stars Opposite the Sun
24 2 Different Back-and-Forth-Star-Directions for a Position Fix
25 For Land-Navigation, the Plumb-Line Star-Pair Method
26 Equal-Height Stars When the Sea Horizon is Visible
27 Great-Circle Reference Star Directions
28 Leo the Lion Looks like a Lion or Sphinx Squatting
It Has Four Lines which Show N./S. E./W.
29 String North Latitude without Numbers
30 "Star-Time" Gains Four Minutes a Day
31 Working the Star-Longitude Numbers
32, 33 Overview
118,119 Appendices: Divide the Star Globe into 8 Octants
—and Learn the Stars in Each
The Southern Figure of Eight
The Zodiac
122 A List of Star Coordinates
124 Drawing Great Circles
125 Calendars in the Sky

Moon

4 Look for the Line Joining Sun, Moon, Planets & Zodiac Stars
— the Ecliptic
6 The Moon; Rule 4: Look for It
8 From the Moon, Locate the Sun-Moon Direct Line, Carefully.
10 *Mentally Rotate the Sun-Moon Line Around Polaris*
12 "Cusp North" — Approximate Direction from the Moon
14 The Rabbit Looks up Ahead at Full Moon Using the Ecliptic
16 The Moon Cusps Push and Pull the Hour Hand Backwards,
Down South
17 Watch the Clouds Cross the Moon
18 The Moon is Lit By the Sun
20 The Moon Rises and Sets 50 Minutes Later Each Night
22 The Moon is "in Phase" with the Sun — Watch for It
24a The Moon Moves Over the Tropics
24 Predict the Full Moon Path According to the Time of the Year
24d Finding North from the Moon, by its height & phase
25 Predict the Moon Phases by " N minus M "
28a Find E/W By Bisection of Sun & Full Moon At Rise or Set
28 To Predict the Moon's Path, Visualise the Ecliptic
29a Find North from the Combination of the Sun & the Moon
29 A 4 Week *Position Cycle* of Moonrise/set, to Find E/West
29e Keeping Track of the Moon's Motion along the Ecliptic
30 Adjusted Moon-Cusp North:
To the Left or Right, By Moving the Sun
32 Overview: The Moon
116,117 Appendix: Moon-Phase & Tide-Time Dial

Wind See also pp. 60a-68a

5 Mnemonic: "Wind? Weather Map!"
7 Mnemonic: "Wind in Your Face"
9 'Low' Clouds Come Clockwise Down under
11 Weather Fronts Distort The (Circular) Wind Circulation
13 Wind-Direction-Change Rule for Buys Ballot's Law
13 Steering Winds
15 Atmospheric Pressure Changes.
17 Watch the Clouds Cross the Moon
19 To Determine Upper Wind Direction from the Clouds
21 Clouds and Con-trails and Their Winds
23 Prevailing Winds on Creek Banks Tips for calm weather
27 Make a Bush Psychrometer
31 Topographic-Level Winds
33 Overview: Global Weather Patterns

52e The Aurorae
53e City Glow

Clockwise See also Map & Compass Work

4 Mnemonic: "Clock"wise Northern Hemisphere SWiNES
6 "VEERING" with the Sun is "Clockwise"
only in the Northern Hemisphere
"BACKING", against the Sun's movement, is "Anticlockwise"
— but only in the Northern Hemisphere.
8 Mnemonic: (when you say) "COMPASS" (say) "CANADA"
10 "Bears Circle Clockwise" Compass Bearings are Clockwise,
A Full Circle, and Eurocentric, Cee?
12 Mariner's "Compass Points" work on Successive Halvings
14 "E. W. N. South" Of What!? Rule: Say What
16 Back-Bearings Killed Sleepy Pilots
18 Back-Bearing: Plus or Minus 200° Then Minus or Plus 20°
20 Turn 90°? "+/-100° -/+10°" !
22 The Clock-Face Method for Direction
24 One Degree is about One in Sixty
26 Anticlockwise is Your Best Spin-Reference Direction
28 Bearings & Tracks: "If it is right, it is big in my sight"
30 "By the Nose" Navigation
32 Overview

Anticlockwise

5 Mnemonic: Down Under, The Sun Moves ANTI-BigBENWiSE
i.e. anticlockwise
7 Southerners Need Anticlockwise Systems
9 A Compass-Needle Continent—Swinging the Continent
11 Interpret a Compass Needle as "RED HOT—WHITE FROST"
13 Four Reference Directions—Name Them "Rise, Noon, Set, Midnight"
15 "Rise-Noon-Set-Midnight" carries its meaning across the Equator!
N.S.E.W. doesn't.
17 The 0, 1, 2, 3, 4 Reference Directions
19 The Sunny Skies Quadrant System
21 The "SWEETHEART" Method of Orientation
23 Mnemonic: "North; Coast"
23 Go Overseas!
27 Walk across Australia! English!
29 South is a Triply Better Reference Direction than North, in the Sth
31 Nose-Navigation Needs an 'Anti' Mnemonic: "Nor-West, Not-Least"
32 Overview
Revision, Overview and Explanations pp. 32,33
Top Down Orientation
Time is counted in cycles
Closer to Home
The Precession of the Equinoxes
The Moon
Global Weather Patterns
The Moving Horizon
The Spherical Earth
The Spinning Earth
Moving the Horizon

Section 2. THE INSIDE STORY—Mental Manoeuvres pp.34-55

Learn

34 Navigation is More Than Navigating Common Sense
36 The "Good" Navigator. You.
38 Don't Just 'Get By' and 'Make Do'
40 Walk Where the Navigators Talk — if you want to learn navigation
40 Explore It's No Fantasy
42 Self-Motivation is Fundamental
44 Routine Mind-Shut-Down Procedures
46 Navigation is Mental Work — Stay Alert
48 To Work Harder, Rest Harder
50 A Three-Track Mind
52 Concentrate Outwards
54 Come to your Senses

Think; Interpret

34 Your Subconscious is Superior to Science!
36 Think out Loud
38 Memories are Made of Landscapes, Not Compass Needles!
40 Feed Your Subconscious through Your Conscious Focus
42 How to Feed Your Brain
44 Feed Your Subconscious Subconsciously
46 Instruments Instruct the Intelligent
48 Minds Work Best with Intelligent Questions on Them
50 Interpret!
52 SPELL Out Numbers — Into Words
52 PAINT your Words — Into Pictures
54 TRANSFER Your Pictures into Reality
54 Put That Experience in Your Memory Bank
55 Extend Visible Reality into the Invisible Picture

Angle:

34 Measuring angles at 1 cm per degree
35 Point and Blink = 10% Shift
36 Guesstimating Angles at 1 degree per cm
37 Distance using Apparent Size;
37 Visual Angle, Size and Distance & 6°; Side-wise Offset
(38, 39 Grid-line Walk & Mental Grid Maps)
40 Cyclical Motion Divided into 12ths
41 Approximate Sine-Wave Values

Look

34 Visual Assessment is Your Main Clue — Rely on It More
36 You Need Excuses to Stop and Have a Good Look
38 You Need a Road-Map To Climb a Hill!
40 Look into the Far Distance Look through the Haze Wait
40 Wait for a Change in Perspective...
42"Retracing is Re-preciating"
44 Look Back Regularly. Look Around
46 Look Through the Vegetation, to the Contours
48 Play the Field-Scientist
50 Look at the Ground
52 Look for Foreshortening & Overlap & Parallax
54 Look, with Vengeance Aforethought

Integrate

3 Visualisation is your Key Skill
34 Integrate: Jetstream, Earth Rotation & Heaven Rotation
36 Visualise Familiar Landscape Trends, Under the Sun
38 'Land to Sky — Sky to Land' Navigation
40 Join the Invisible Big Picture To the Visible Landscape
42 Landmark to Landmark Network Navigation
44 Aim to Integrate…
46 Imagine an Aerial Photograph of the Continent
48 Connections to the Outside World
—You don't need to feel isolated
50 Orientation is More Fundamental Than Position Sense
52 Make Common Sense of the Various Direction Clues
52 Don't Degrade the Point of a Clue or System
54 Preparation is the Key to Difficult Navigation
42e Moveable Position vs Fixed Orientation
43e Name Each Clue — with solid ground
44e Notice and Match Up
45e In-Line Integration
50e Think in Pairs, in lines, trends and tracks
Orientation 35 Orientation = Direction Sense
37 The Get-Back Path is the Single Fundamental
38e Grid-Line Walk
39 Disorientation is More Dangerous Than Being Lost
39e Mental grid Map
41 Which Way Did I Come In? Which Way Do I Go on?
43 Disorientation — A Killer Syndrome
42e Movable Position vs Fixed Orientation
45 Un-Dis-Oriented;
47 Avoid Disorientation: OverNavigate
46e Making a M.U.D. M.A.P.-m.a.p.-m.a.p.
49 OverNavigate: Equals Over-Familiarise
51 Needle N, S, E & W. Identify them on the Landscape
51 Orientation to Start-Finish Rhumb Line—by sight or compass
51 Orientation to Your Finish Point, By the Parallax Method
53 Orientation to Your Start Point, By the Parallax Method
55 Look Backwards, Think Forwards

Disorientation

35 The Navigator's Top 10 Problems; and Band-aid Solutions
37 How I Manage to Get Lost or Disoriented
39 Logical Blunders — Examples
41 Logical Presumption Leads to Utter Delusion
43 Casual Outings Catch You Out
45 Trapped! By Not Looking Back
47 How to Get Lost—Follow the Track—A Track is a Trap-Door
47e Disorientation in new territory
49 How to Get Lost—Follow the Leader
51 Exponential Anxiety — the 11:59 syndrome
53 Nervous Nellie Navigation — Not

Reorientation

35 Reorientation
37 Voice Your Thoughts — Don't Suppress Important Clues
39 Trust the Untrustworthy
41 Trust Less and Prove More
43 Trust Yourself Less
45 Disoriented? Stop. Say So. Reorient!
47 Reorientation; And a Fresh Start
49 "As Solid As the Rocks in My Head"
51 Explore Junctions, While You Can
53 Look for the Essentials — the 'Directions'
55 To Get There Quicker, Go Slower

Lost?

35 Lost? Rule One: Stop! ..getting More Lost
37 Back-track in Your Mind before You Back-track on Your Steps
39 Rule 2: Start! Getting Unlost. Find yourself. Regain confidence
38e Grid-line Walk — Explore in a Square Grid Pattern
39e Mental Grid Map
41 The Simplest Way to Get Unlost is to Retrace Your Steps
43 More is Better — Thought, Time Care
45 Lost? Rule 3: Do a Good Job—of walking out or staying put
47 How to Leave a Message
49 Lost? Rule 4: DON'T Do It Again!
51 Separated? An Agreed Plan Helps
53 A Whistle-Code for Contact Calls
55 Search and Rescue Signals — Noises Repeated Regularly
120 Appendix: Morse Code and Rhythm Mnemonics for it
121 Appendix: Semaphore Code and Mnemonics for it

Section 3. REALLY INTERESTING STUFF
—Bush Navigation pp56-89 See Also 'Wind'

Nature's Norths

56 Nature's Norths
58 Rock Noon
60 Wind West See also Wind, Section 1
62 Think Wind
64 Be Wind-Wise: Listen
64 Obvious Tell-Tales: Windblown Objects
64 Obvious Tell-Tales: Landscape
66 Light & Colour Effects
68 Botanical Tell-Tales
70 Combined Effects of Wind, Salt and Light
70 Botanical Asymmetry
72 Think like a Plant
72 "Mosses Mop up Moisture" "Lichens Like Light"
72 "Fern Fronds Face the Front" i.e. the sunny side
74 Bush Noon— The Sun is Always out, If You Know How to Look
74 False Asymmetry & Symmetry True Symmetry & Asymmetry
76 Higher Density of Leafiness on the Noon Side
76 The Lowest Leafy Branches are on the Noon-Side
76 "It's the Shady Side That Shrubbery Shows up"
76 Tall, Steep and Ragged Backside Wedge-Shaped
76 Hollow under the Shady Side & Imperfect Shape
76 Gaps on the Shady Half Dead Branches
78 'Solar-Panel' Tilting of the Canopy Top
78 Wedge Shaped Lopsided Lean Half-Dome One-Sided Leafiness
78 Individual Branchlets Bare Backbone
80 Branching Patterns
—Examine Them for thickness, length, angle, curvature
82 Isolated Vegetation Shows the Clearest Asymmetry
82 Closed Canopy — Look for Gaps; Saplings
82 Roadsides & Campsites Reverse Some Effects!
84 Flowers Respond to Sunlight Fruit Distribution and Ripening
86 Stem and Leaf Colouring New Growth is Colour Differentiated
88 Noon-Sunlight is More Vertical
88 The Shedding of Bark on the Sun-Baked Side
88 Where to Look for Fading or Lack of it

Memory & Naming Skills See Sect.5, Navigation Mnemonic

56 Don't Forget to Remember, Your Way!
56 Unlock Memorable Chapters in Your Life
56 Go for the Jugular. Zero in on What Matters
58 Don't Try to Memorise Details without an Overall Framework
58 Don't Try to Remember Anything You Haven't Memorised
58 Write down What You Memorise!
60 Anything is Memorable
60 Make Individual Memories Memorable!
60 Remembering Trees
62 The Difference between Looking at, and Seeing
62 Search-Image
62 We Recognise 'Sameness' But Need to Describe the Differences
64 Place Recognition
64 You Recognise What You See, By What You Can't See
64 Description Requires Words
64 Can You Describe All That You Can Recognise?
66 Discuss It
66 Collect the Details
66 Identify a Specific Place by its Specific Collection of Details
66 Include What You Can't See
68 Extend Your Familiar Range Systematically Hints
68 Leap-Frog Walk in New Territory
70 Draw the Skyline, the Ridge-lines, the Outlines
70 Label it with the 8 main directions
72 Examine the Skyline BEFORE You Walk up to It
72 Make a Mental Movie
72 Take Mental Photographs
72 Take a Mental Photo of Each Change of Track
74 Remembering the Sequence, Needs More Than Logic
74 Do Something NUMERICAL
74 Match the SHAPE
76 Use a Type of Memory which you are Good at
76 Use the Name of the Track for the Acrostic
78 Group Several Related Images into One See also p. 110
80 Working Names — Invent Them
80 Standard Name Plus Personal Name
80 Remembering Names
82 Name the Subsections
82 Name the Sequence
84 Revise your Memories To Remember Formulae
86 Remember the Sequence — from AA to ZZ

Lie of the Land

57 The "Lie" of The Land
57 Listen to the Pioneers Lie Has To Do with Water Drainage
59 Major Rivers & Ridges vs Minor Ones Sort it out on the Map
59 Catchment Divisions Snake from Coast to Coast
59 A Catchment-Division Separates Whole Valleys
61 Ridges Ridge Names Mnemonic: "Ridges Rise"
61 Uphill — Which Hill? Upstream — to which Divide?
63 Shoulders Stop. Shoulders are Not Easy (like Ridges) to Walk on
63 Saddle — A Flat between Hills, Sitting Astride a Ridge
65 Notice When You Cross a Catchment!
65 The Giveaway! A Change of View—Mnemonic: "Catch the View"
65 You Can't Cross a Creek Twice from the Same Side
67 Break of Slope/Make of Slope Scree Slopes — Tread Carefully!
69 Break of Slope — And Route-Finding
71 Mnemonic: "Down in the Mouth" Which way is the flow?
71Hint: Don't think "this side" & "the other side".
73 The Inside of a River Bend is Flatter Ground
73 Rivers Bend! Don't Forget to Look Backwards
73 Rivers Meander. Ridges Run Straighter
75 Land-Form. Hint. What Makes a Cliff and an Overhang?
77 Headlands, Bays, Prevailing Winds, Longshore SandDrift, Creeks
77 Beaches — Slope, Particle Size, and Wave Energy
79 Rock Typing Soil Typing

Contours and Slopes

81 Contour Lines
81 Contour Navigation: Practise It, Before you Need It
83 Contour Curves: Your Job is to Imagine the Shapes into Life
And to Match the Lines to the Landscape
85 Contour Heights: Read the Numbers!
85 Contour Spacing: Close = Steep Slope
85 Contour Slopes: Work Out 1 Mm and 1 Cm Contour Spacing
85 Reference Slopes — To Imagine the Climb with!
85 The Beaufort Wind Scale — Modified for Bushwalkers
87 Make a Clinometer — a Slope Measurer
89 Just Draw It

Tracks and Bends

57 Stick to the Track Reasons for Leaving the Beaten Track
57 Cross-Country Routes
59 Tracks are like Naughty Children
59 The same is true for creeks & ridges
59 Explore Branches While You Can
59 Convergent Tracks Lead to Civilisation
59 Down-Track vs Up-Track
61 Sympathy and Understanding with Track Markers
61 Old Track Markers
63 Track Markers When You Lose the Track-Markers
63 Notice (with Alarm!) When...
65 Mark Your Trail Mark Your Trail Non-Permanently
65 Tracking Someone Else
67 Mark Your Trail More Permanently and Ostentatiously
67 Mark Your Trail Invisibly
67 A pebble Code — Keep It Simple

Bends

69 Keeping Track of the Bends Desired Equipment
69 Logging Constant Diversions
71 The Difference between 'Direction' and 'Track'
71 Continuous bends
71 Follow the bends by anticipating the trends
73 Keeping a Compass Course, by Legs
73 Turn Smooth Bends into Sharp Jumps
73 Mind Which Way You Go
75 "End the Bend" i.e. Notice the Un-Bends
75 Name the Bend, By the Trend, of its End
75 Notice and Name Perverse Bend
77 Look Through the Trees, to the Sky and Skylines
77 Periscope Navigating by Trees
79 Bush networks Don't Navigate by Networks Only
79 Antidotes For Guessing at Network Junctions
81 Sticking to the Bush Network Natural Routes Naturally Converge
81 Coping with Divergence Reverse logic
83 Trackless Navigation The method of Names
83 Bush Networking Essentials
85 Walking "Straight" through the Bush Straight around
85 How to Walk In an Unintentional Circle
87 Walking an Exact Straight Line in Open Country
87 Leap-frog a Straight Line by Looking Ahead
89 Radio Direction Finding—As Good as a Compass

Section 4a. DISTANCE pp. 90-108 left-hand pages
See also 'Angle'

Progress

90 Distance Guesstimation is a Major Problem for Walkers
92 Use the Map Scale—Don't Just Look at It The Jargon
92 Map-Scaling-Factors are Meant to be Interpreted
94 Psychological 'Half-Way Point' When You Meet Other Walkers
96 AIMING OFF by 10° — 10° is **175m per kilometre**
96 Aiming-off at Sea; Aim Off 10% or 6°
98 Up-down Progress
100 Distance by Double Pace—Test Yours Tie Your Ankles Together
100 Don't Lose Count! Use Finger-Math
100 Test your pedometer
102 Line of Sight Can Be Extremely Accurate
102 Two Opposite Landmarks
104 "Changeover Lines" of Progress — Relative Apparent Sizes
104 Use the Landscape Itself to Measure & Mark Progress
106 The Sine-Cosine Tally Method of Keeping Track, Blind
106 Rise and Run Adjustments — Sine and Cos — North and East
108 Progressing A Position Line across the Map
108 Pinpoint "Fixes" are a Myth!

Pace

90 Distance = Speed x Time Keep a Log! Adding two Speeds
92 Pace Your Progress — Pre-Dict It, Don't Post-Dict It!
92 Advice about Walking Times — Double & Halve!
92 20 Minutes per Kilometre — Double & Halve — Easygoing Maths
92 4 km per hour — Normal Walking Pace — Double and Halve!
94 1 Metre per Second is 3.6 kph "1 Hour Extra per 5 Hours"
94 "1 Hour Extra per 500 m easy climbing"
96 Adjust Your Expected Pace
96 Don't Overcompensate for Slopes And Meanders
98 Don't Lose Count! Use Finger-Math
100 A Slow Kilometre Cancels out Three Quick Ones
100 To Progress Faster More Easily Pace Yourself
— Slow Down, but Don't Stop
102 Distance-Off by Vertical Angle Safe Distance Off
104 Circle-of-Position by Horizontal Angle
105b A Horizontal Angle between Features Yields a Position Circle
105c 'Two Horizontal Angles' Usually Fixes Your Position
106 Rise and Run Adjustments — Sine and Cos — North and East

Range

24e One Degree is about One in Sixty
34-36e Estimating angles
37e Distance using Apparent Size & 6°
37e Visual Angle, Size and Distance Side-wise Offset
51a & 53a Parallax methods
90 Range: Rule One: Guess
92 "Range" of Visibility — Two Halves to the Picture
92 "Range Squared is Dee Height"
92 "Range Squared is Dee Height"
94 Double, the Acute Angle, off Your Nose — For the Distance Off
94 Halve, Any Angle, from your Stern—For Distance Off
96 Distance by Optical Rangefinder
96 Range by Parallax and a Piece of String DIY Rangefinding
98 Distance by averaging maximum & minimum Guesses
98 Distance by Comparison with Some Length You Know
98 Distance by Cloud Shadows — Look for Them
98 Sound Travels Three Seconds to the Kilometre
100 Lighting Conditions & Eyesight
100 Visual Acuity is about One Minute of Arc
102 Approximate Figures for Visual Acuity
102 Test Your Own Visual Acuity
104 Circle-of-Position by Horizontal Angle
105b A Horizontal Angle between Features Yields a Position Circle
105c 'Two Horizontal Angles' Usually Fixes Your Position
104 The best range-finder
106 Rule of Fingernails Rule of Tom Thumb Apparent Size

Section 4b. MAP and/or COMPASS pp. 91-109 right-pages

Map Use

54e Helpful Chart Symbols 122 Appendix: Symbols
38,39e Grid-line Walk & Mental Grid Maps
46e Making a M.U.D. M.A.P.-m.a.p.-m.a.p.
91 To "Read" a Map — Rule One: Don't Lose Your Place!
91 Rule Two: Imagine Arrows for the Sun, & Your Movement
91 Rule Three: Get Used To Any New Map Scale
91 Rule Four: Don't Give Up!
93 Look Carefully at the Fine Detail on the Map
93 Look Carefully at the Fine Print
93 Some Map Details are Not to Scale!
95 Permanent Landscape is More Trustworthy Than Surface Detail!
95 Details Added on to a Map Are Prone to Error
95 Expect The Map To Be Unhelpful
95 Maps Do Give You 10 Times More Detail Than You Can Notice
97 "Maps Grip the Universe"
99 Align the Track
101 Align the Map the Landmarks
103 Align the Map to Identify Landmarks
105 Align the Map to Fix Your Position
107 Which Way is North?

Magnetic Compass Use See also "Clockwise", Section 1

37e Chart symbols
91 "L.M.N.Or P."
93 Using the Protractor by Itself
95 Using the Needle Alone — Emergency Magnetic Compass
95 Hint Understand that magnetic poles should be called
'North-Seeking' and 'South-Seeking'
97 Think "+/- 5°" For Compass Bearings "+/- 10°"
For Following a Compass; "+/- 15°" For Using a Compass!
99 Compass Errors — Be Impressed! Be Careful!
101 Compass Use — Master It, but Be a Good Master.
103 "Magnets Grip Maps"
103 Align Your Map in the Dark
107 'East-least; West-Best' Rule—Tells You Which Bearing is 'Best'
108b Getting Started in the Dark
108c Thinking Backwards
109a Taking a Bearing from the Landscape
109 Plotting a ('Compass-to-Map') Back-Bearing
109 No-Number Back-Bearings
124 Appendix: Map Symbols

Mapmaking

38e Grid-line walk
39e Mental Grid Map
46e Making a M.U.D.M.A.P.-m.a.p.-m.a.p.
91 Maps Mean: Imagine Me! — Making Mental Maps
91 "Map To Mind" Navigation
93 Mental Networks
93 The Most Valued Map
95 Strip Map Navigation — Follow-Your-Nose Navigation
97 Caving Strip-Maps — Come in Books
99 Map Folding and Map Holding
101 Layout Map — Blind
103 A Compass Holder
106a The Sine-Cosine Tally Method of Keeping Track, Blind
106b Rise and Run Adjustments—Sine and Cos—North and East
107 String Map
109 Parallax While You Walk — JUST DRAW IT!

Section 5. Navigational Mnemonic pp. 110-115
78b Group Several Related Images into One
110 A "Start-Stop-Rest" Navigation Mnemonic
110 P.R.E.P.A.R.E.D.— A Before-You-Go Checklist
111 Equipment—A What-You-Might-Like-to-Take Checklist
111 Place-Names; Learn Them; You Navigate by Local Features
111 S.T.A.R.T. off Oriented
111T.I.M.E. & Distance — that's the point
111 Distance
111 Progress; Pace; Range
112 Look in These 6 Directions — in 6 ways — Up and Down
112 4 More Slopes
112 Look in These 6 Directions — in 6 ways — Ahead and Behind
112 Look in These 6 Directions — in 6 ways — Right and Left
113 ..Plus in Two More Directions
113 (1) All, Around, Appreciatively, And At 6 L's
113 (2) Look at the Map: a Systematic Checklist. What 'System'?
113 Mark the Lines-of-the-Land On Your Map
113 Mnemonics for Map and Compass Work
114 N.S.E.& W.—the 4 cardinal directions, is a mnemonic of fours
114 W.E.A.T.H.E.R. — Check for Change
114 Stop-Signal: Why have you stopped?
114 REST and R.E.O.R.I.E.N.T yourself
115 RI For Revisualise the Invisible & then Relate It to the skyline
115 Clock-Sense Mnemonic
115 ENT for Environmental Notice Time—Orient yourself to Nature
115 REMEMBER to Note down the Names in the Notebook
115 Re-lax, and Play — Navigation Games and Activities

Miscellaneous pp. 34e - 55e
52e The Aurorae
53e City Glow
54e Helpful Symbols for Standardising your Chart–Work

Appendices pp. 116-121
116 Moon-Phase & Tide-Time Dial
118 Divide the Star Globe into 8 Octants
—and Learn the Stars in Each
118 The Southern Figure of Eight
119 Divide the Star Globe into 8 Octants
—and Learn the Stars in Each
119 The Zodiac
120 Morse Code & Rhythm Mnemonics
121 Semaphore Signalling Code & Mnemonics for it
122 Symbols for Track-Notes and Mud-maps
122 A List of Star Coordinates
128, 129 Tropical countries and places underneath the Sun
124 Drawing Great-Circles. The PZX navigational triangle
125 Calendars in the Sky
126,127 About the Author; Outline
Index pp. 128-129
Contents p. 2

Eastern Australian Standard Time Zone

10 hours ahead of Greenwich
Showing Tropical countries and places underneath the Sun

p24a,25a

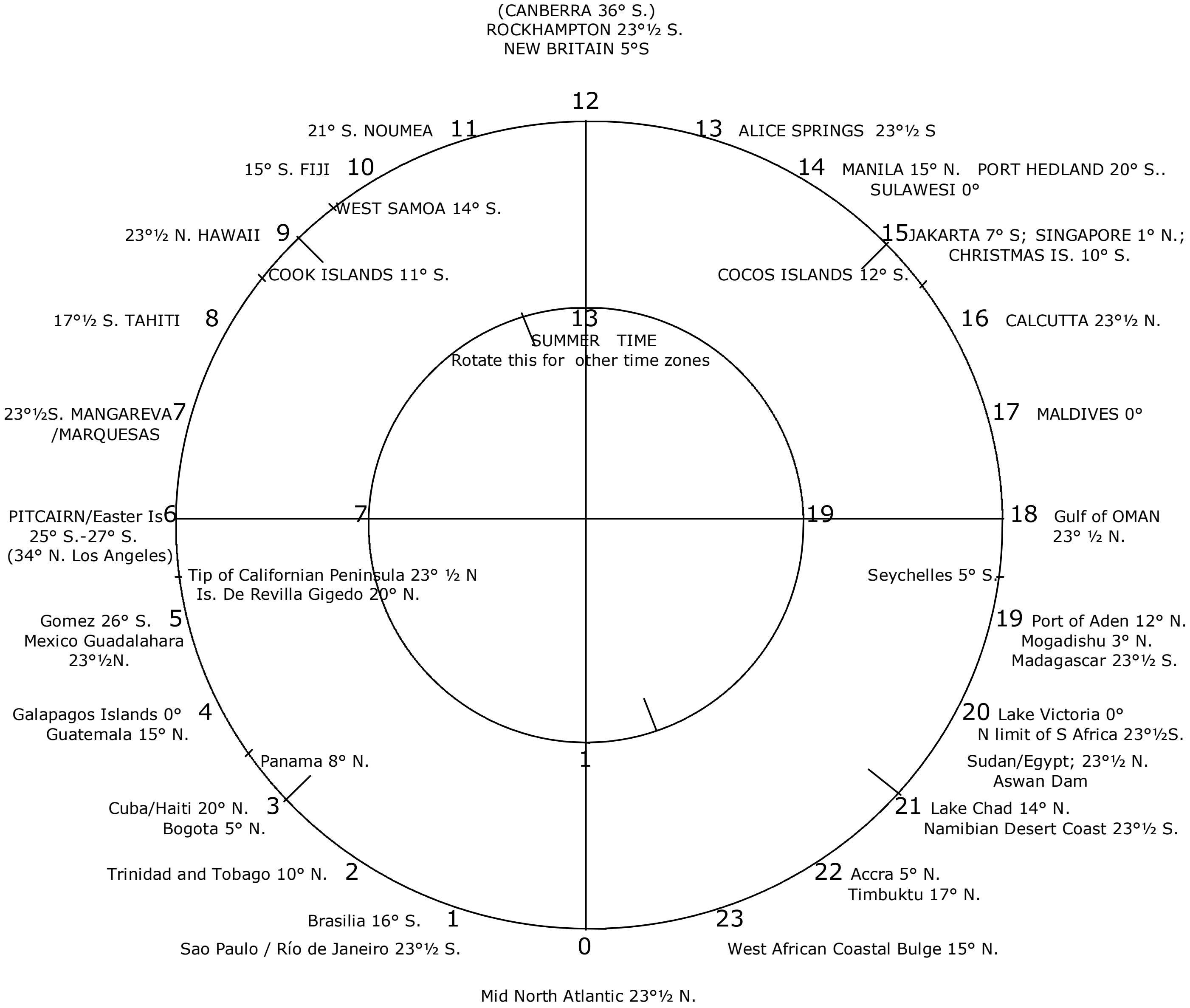
(CANBERRA 36° S.)
ROCKHAMPTON 23°½ S.
NEW BRITAIN 5°S
12
21° S. NOUMEA 11
13 ALICE SPRINGS 23°½ S
15° S. FIJI 10
14 MANILA 15° N. PORT HEDLAND 20° S..
SULAWESI 0°
WEST SAMOA 14° S.
23°½ N. HAWAII 9
15 JAKARTA 7° S; SINGAPORE 1° N.;
CHRISTMAS IS. 10° S.
COOK ISLANDS 11° S.
COCOS ISLANDS 12° S.
17°½ S. TAHITI 8
16 CALCUTTA 23°½ N.
13
SUMMER TIME
Rotate this for other time zones
23°½S. MANGAREVA 7
/MARQUESAS
17 MALDIVES 0°
PITCAIRN/Easter Is 6
25° S.-27° S.
(34° N. Los Angeles)
7
19
18 Gulf of OMAN
23° ½ N.
Tip of Californian Peninsula 23° ½ N
Is. De Revilla Gigedo 20° N.
Seychelles 5° S.
Gomez 26° S. 5
Mexico Guadalahara
23°½N.
19 Port of Aden 12° N.
Mogadishu 3° N.
Madagascar 23°½ S.
Galapagos Islands 0° 4
Guatemala 15° N.
20 Lake Victoria 0°
N limit of S Africa 23°½S.
Panama 8° N.
1
Sudan/Egypt; 23°½ N.
Aswan Dam
Cuba/Haiti 20° N. 3
Bogota 5° N.
21 Lake Chad 14° N.
Namibian Desert Coast 23°½ S.
Trinidad and Tobago 10° N. 2
22 Accra 5° N.
Timbuktu 17° N.
Brasilia 16° S. 1
Sao Paulo / Río de Janeiro 23°½ S.
0
23
West African Coastal Bulge 15° N.
Mid North Atlantic 23°½ N.

Places in capitals are less than 6 hours away from Australia. The others are on the other side of the world.
View this as if from above the North Pole—the Sun ‘moves’ clockwise, but the Earth spins anticlockwise.

Eastern Australian Standard Time Zone

10 hours ahead of Greenwich
Showing Tropical countries and places underneath the Sun

Cut this one out and post it on your wall until you have learned most of it.

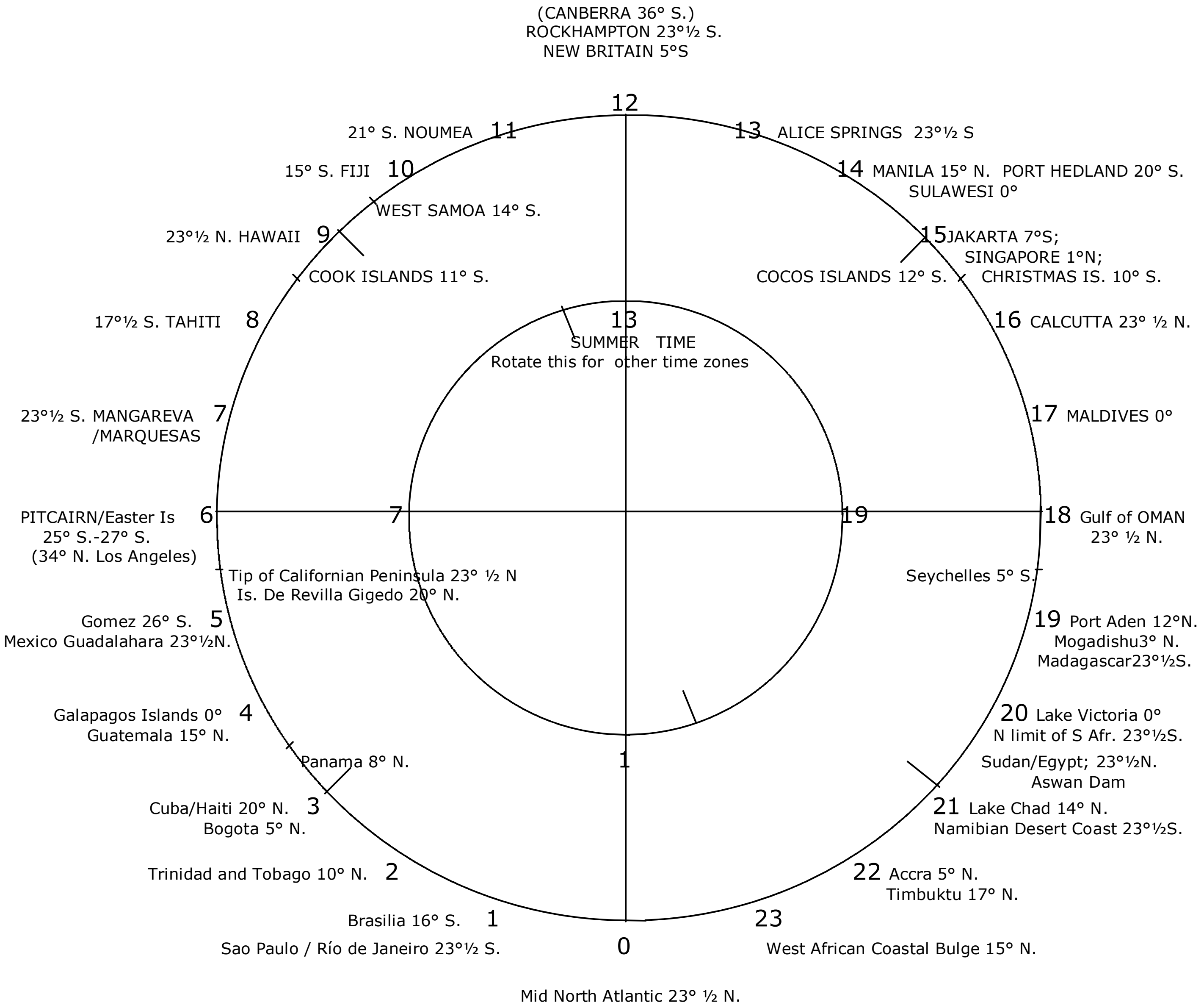
(CANBERRA 36° S.)
ROCKHAMPTON 23°½ S.
NEW BRITAIN 5°S
12
21° S. NOUMEA 11
13 ALICE SPRINGS 23°½ S
15° S. FIJI 10
14 MANILA 15° N. PORT HEDLAND 20° S.
SULAWESI 0°
WEST SAMOA 14° S.
23°½ N. HAWAII 9
15 JAKARTA 7°S;
SINGAPORE 1°N;
COOK ISLANDS 11° S.
COCOS ISLANDS 12° S.
CHRISTMAS IS. 10° S.
17°½ S. TAHITI 8
13
SUMMER TIME
Rotate this for other time zones
16 CALCUTTA 23° ½ N.
23°½ S. MANGAREVA
/MARQUESAS 7
17 MALDIVES 0°
PITCAIRN/Easter Is 6
25° S.-27° S.
(34° N. Los Angeles)
7
19
18 Gulf of OMAN
23° ½ N.
Tip of Californian Peninsula 23° ½ N
Is. De Revilla Gigedo 20° N.
Seychelles 5° S.
Gomez 26° S. 5
Mexico Guadalahara 23°½N.
19 Port Aden 12°N.
Mogadishu3° N.
Madagascar23°½S.
Galapagos Islands 0° 4
Guatemala 15° N.
20 Lake Victoria 0°
N limit of S Afr. 23°½S.
Panama 8° N.
1
Sudan/Egypt; 23°½N.
Aswan Dam
Cuba/Haiti 20° N. 3
Bogota 5° N.
21 Lake Chad 14° N.
Namibian Desert Coast 23°½S.
Trinidad and Tobago 10° N. 2
22 Accra 5° N.
Timbuktu 17° N.
Brasilia 16° S. 1
23
Sao Paulo / Río de Janeiro 23°½ S.
0
West African Coastal Bulge 15° N.
Mid North Atlantic 23° ½ N.

Places in capitals are less than 6 hours away from Australia. The others are on the other side of the world.
View this as if from above the North Pole—the Sun 'moves' clockwise, but the Earth spins anticlockwise.

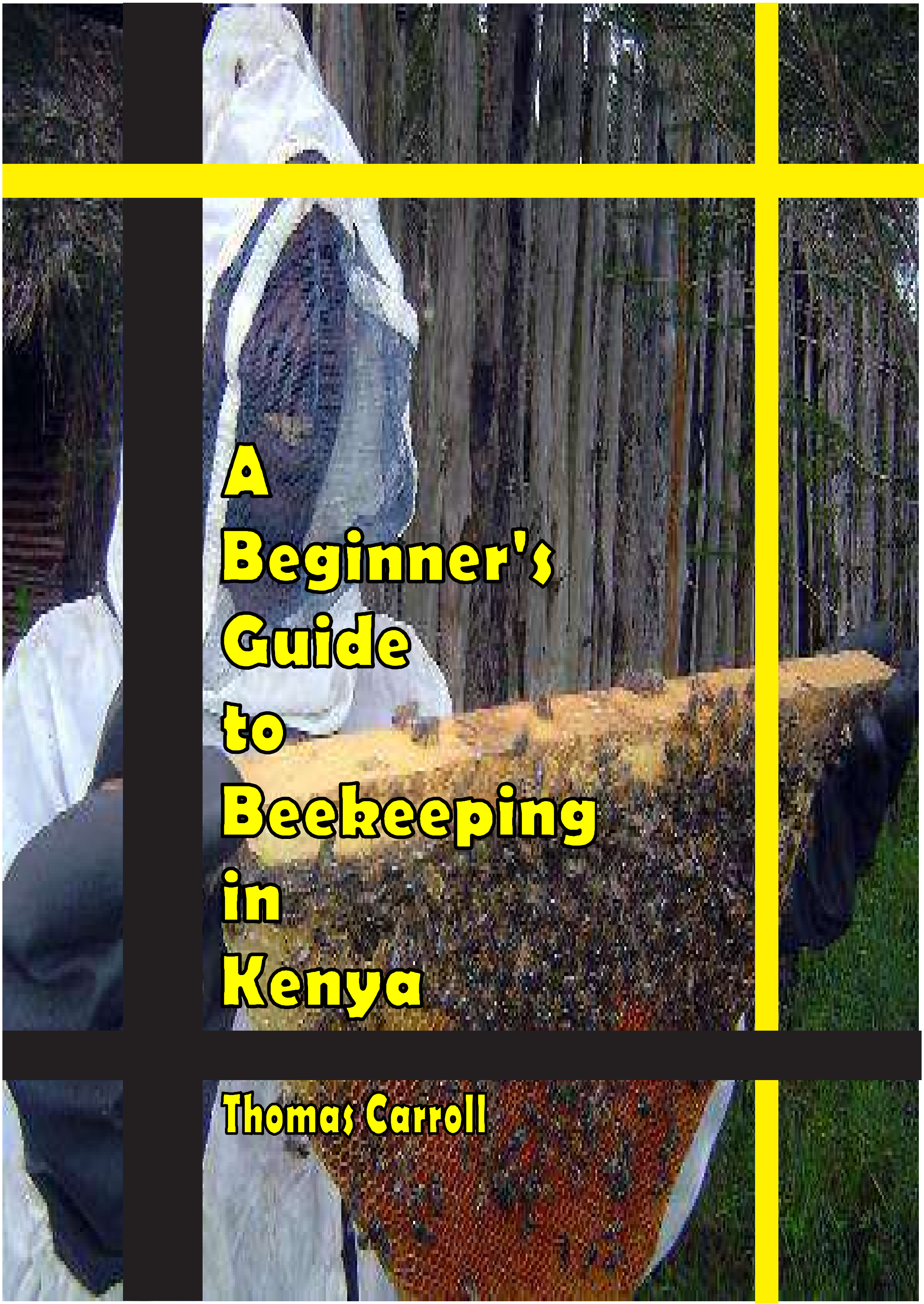

A Beginner's Guide to Beekeeping in Kenya

Thomas Carroll

Contents

ILLUSTRATIONS (ix)

PREFACE (xi)

PART ONE INTRODUCTION TO BEEKEEPING (1)
What is Beekeeping? (1) Getting to Know Bee Species and Races (1)
What Bees Do (2) All about Honey (3) Information on Beeswax (3)
What is Propolis? (3) Residents of the Beehive (3)

PART TWO GETTING STARTED: THE BASICS (6)
Types of Hives (6) Apiary Location (17) Hive Placement (19)
General Apiary Management (22) Beekeeping Equipment and How to Use Them (23)
Handling Bees (25) What If No Bees Enter the Hive? (28)
Causes of Absconding and Its Control (30)

PART THREE THE BEEKEEPING YEAR (31)
Conditions for Maximum Honey Production (31)
Beekeeping Management during the Year (38)
Swarm Prevention and Control (39) Making a Division (40) Feeding (40)
Pests and Diseases and Their Control (41) Major Nectar-bearing Plants in Kenya (43)

PART FOUR THE HONEY HARVEST (47)
Time of Harvesting (47) Harvesting Procedure (47)
Honey Refining (50) Extraction of Beeswax (52)

PART FIVE THE MARKET FOR BEE PRODUCTS (54)
Honey (54) Beeswax (56) Propolis (56)

PART SIX STARTING A BEEKEEPING ENTERPRISE: COSTS AND PROFIT (57)

GLOSSARY OF TERMS (61)

ANNEXES (65)
ANNEX 1 Current Status of the Kenyan Beekeeping Industry (65)
ANNEX 2 More Information on Stingless Bees (67)
ANNEX 3 Hive Recordkeeping (68)
ANNEX 4 Recipes/Uses for Beeswax, Propolis, and Honey (70)
ANNEX 5 Approaches to Beekeeping Extension (74)
ANNEX 6 Sources of Further Information (76)
ANNEX 7 Baraka Agricultural College Short Courses on Beekeeping (81)
ANNEX 8 Beekeeping Resources on the Internet (83)
ANNEX 9 Recommended Reading/References (84)

Illustrations

FIGURES

1. Types or Castes of Bees (**4**)
2. Bee Egg (**5**)
3. Bee Larva (**5**)
4. Bee Pupa (**5**)
5. Construction Drawing of KTBH (**13**)
6. Assembly of KTBH (**14**)
7. Hanging KTBH Posts (**20**)
8. Hanging Patterns for KTBH (**20**)
9. Conditions Required for Maximum Honey Yields (**31**)
10. Seasonality of Honey Production in Kakamega (**35**)
11. Candle Making (**72**)

PHOTOS

1. Wild colony of bees (**6**)
2. The pot hive (**7**)
3. Traditional log hive (**7**)
4. Ugandan basket hive (**8**)
5. Examples of KTBH and Langstroth hive (**8**)
6. Crudely constructed box hive (**9**)
7. The Langstroth frame hive (**10**)
8. Beekeeper in Kitui District (**11**)
9. The Kenya top bar hive (**12**)
10. Top bar (**12**)
11. Working with bees in a KTBH (**15**)
12. Mud/stick hive inside a bee house (**16**)
13. Bee house at Baraka Agricultural College (**17**)
14. Apiary fenced with off-cuts (**18**)
15. KTBH hanging in foreground (**19**)
16. Newly constructed bee house (**21**)
17. Water for bees (**22**)
18. Good gloves (**23**)
19. Farmer in Nakuru wears an inexpensive bee suit (**23**)
20. Smoker (**24**)
21. Lighting the smoker (**24**)
22. Lighting a tin can smoker (**24**)
23. Bee brush and hive tool (**25**)
24. Mick, my son (**25**)
25. Catcher boxes for trapping passing bee swarms (**29**)
26. A tree in flower (**32**)
27. Hive infested with wax moths (**41**)
28. Tree encircled by *mabati* (**42**)
29. Citrus fruit in bloom (**44**)
30. Banana in flower (**44**)
31. Acacia *senegal* in bloom (**45**)
32. Acacia *persiciflora* in flower (**45**)
33. Kei-apple (**46**)
34. *Cordia africana* (**46**)
35. Harvesting white combs (**48**)
36. Harvesting traditional log hive in Tabora, Tanzania (**49**)
37. Harvested honey (**49**)
38. Simple honey refining method for small-scale beekeeper (**50**)
39. Centrifuge (**51**)
40. Uncapping frame full of honey (**51**)
41. Solar wax melter used to extract beeswax (**53**)
42. A sampling of honey products (**54**)
43. Simple box hive for stingless bees (**67**)
44. Hive with shaded glass cover for stingless bees (**67**)
45. An exhibition stand of the Kenyan National Beekeeping Station (**70**)
46. Cake of beeswax for making candles (**71**)
47. Beautiful beeswax candles made using plastic pipe mould (**72**)
48. Bee forage tree seeds (**75**)
49. The teaching apiary at Baraka Agricultural College (**82**)

TABLES

1. Duration of Development Stages (**5**)
2. Beekeepers' Preferences for Different Hive Types (**34**)
3. Plants Suitable for Bees in Molo (**36**)
4. Plants Suitable for Bees in Transmara (**37**)
5. Costs of Starting a Beekeeping Business (**57**)
6. Gross Margin Analysis of Beekeeping (**58**)
7. Gross Margin for a Hectare of Maize (**59**)
8. Hive Recordkeeping (**68**)
9. Marketing Recordkeeping (**69**)

Preface

This book is about beekeeping as practised in Kenya. Wikipedia (a free online encyclopaedia) defines it as “the practice of intentional maintenance of honeybee colonies, commonly in hives, by humans. A beekeeper (or apiarist) may keep bees in order to collect honey and beeswax, or for the purpose of pollinating crops, or to produce bees for sale to other beekeepers. A location where bees are kept is called an apiary”. We also know beekeeping as apiculture, from the Latin *apis* or bee.

The purpose of beekeeping is to produce and sell as much honey as possible. Beekeeping is a means to expand livelihoods and make the best use of available resources, in this case, floral nectar and pollen. A diversified livelihood is a more secure one. Beekeeping also has immense benefits in terms of provision of pollinators, which enhance crop yields. Scientists estimate that one in every three bites of food we eat is a result of active pollination of plants in which bees play a very important role. Bees are environment-friendly and so are beekeepers. Many beekeeping groups initiate tree nurseries as they realise the importance of enhancing the environment to improve honey yields. Beekeepers are also the most active in safeguarding natural environments, as they know honey crops are dependent on local vegetation, mostly trees, as their source of nectar. **Teach people to keep bees and they will plant trees themselves.**

Better beekeeping requires good management of bees and hives. Anyone—young or old, men or women—can keep bees. Beekeeping needs very little land, time, or equipment. It is an ideal activity for small-scale farmers in Kenya and other African countries.

Farmers who want to keep bees generally have two challenges:

1. LACK OF INFORMATION ABOUT BEEKEEPING

This is a common problem: where can beekeepers in Kenya and other African countries get relevant information? Information from other continents such as Europe and America is useful but not always applicable. African bees and conditions are different and require appropriate information on local beekeeping constraints as well as opportunities.

2. A FEAR OF BEES

Many people start beekeeping only to run into problems with excessive stings that result in the farmer getting rid of the bees. This is also a common difficulty in Kenya as Kenyan bees are unpredictable and sometimes very aggressive. Knowledge on how to keep and handle bees properly can overcome a fear of bees.

This book assists with information on beekeeping, which helps overcome fears. It also directs to sources of further information. All beekeepers need to keep learning and improving their craft. Learn from bees themselves, neighbours and fellow beekeepers, the local beekeeping officer/extension worker, and all the books and other resources obtainable.

A Glossary is available after Part Six to assist you with definitions for the more important terms used in this guide.

(N. B. *On exchange rates: KSH refers to the Kenyan shilling. Exchange rates at the time of writing are USD1=KSH72; GBP1=KSH136; EUR1=KSH92.*)

Part One

INTRODUCTION TO BEEKEEPING

WHAT IS BEEKEEPING?

Beekeeping is the art of managing honeybees in order to obtain honey, beeswax, and other bee products for food, income, and sometimes, even medicine. Men and women of any age can carry out beekeeping, but there are taboos against women handling bees in some Kenyan communities.

It is an ideal income-generating activity for groups such as those for women, youth, men, church, etc. We have seen a number of very active youth groups involved in beekeeping. They can generate income without owning land, a major constraint to other lucrative endeavours. Beekeeping requires little space; it complements other farm activities. Beekeeping does not need good soil.

GETTING TO KNOW BEE SPECIES AND RACES

A. HONEYBEES

There are many different species of bees in the world, most of them solitary. A few species of bees are kept to produce honey. In Africa, America, Australia, and Europe, we have *Apis mellifera*, which is a very important species for honey production. Recent research has shown that all races of *Apis mellifera* have origins in Africa. In Kenya, we also have Apis mellifera or the species most familiar to everyone: the honeybee. This book is mostly about the honeybee. Within this species, there are a number of bee races in Kenya that have their particular characteristics and adaptations to our environmental conditions: *Apis mellifera scutellata*, *Apis mellifera monticola*, *Apis mellifera yeminitica* (*nubica*), and *Apis mellifera littorea*.

1. *Apis mellifera scutellata*

These are small with relatively short tongues. The bee is highly aggressive and has great tendency to reproduce (swarm) and abscond (migrate). It is found in plains. Massive flowering attributes to their high reproductive rate, which occurs in plains just before rains. Scutellata is able to nest in a broad range of sites: from cavities to open nests.

2. *Apis mellifera monticola*

This bee is called the mountain bee. *Monticola* are a large dark gentle race with longer hairs than other African bees. The bee inhabits places where clouds, mist, and nocturnal ground frosts obscure the sun. It is the largest bee in Africa. It has a tendency to reduce brood rearing at the first sign of forage decline and may not migrate. It is less productive and less vicious than other species. *Monticola* is found in Meru and Mt Elgon.

3. *Apis mellifera yemenitica* (formally, *Apis mellifera nubica*)

This is the smallest race in Africa. It has the most slender abdomen and the largest yellow abdominal colour band of all African races. It withstands and survives drought conditions by excessive migration. It is found mostly in the northern parts of Kenya.

4. *Apis mellifera littorea*

The bee inhabits lowlands of the Kenyan coast. It does not migrate as much as *scutellata*. It has a tendency to rear brood throughout the year due to availability of forage along the coast.

Bee races I am most familiar with are *A. m. monticola* and *A. m. scutellata*. *A. m. monticola* tend to be in the highlands and more docile. We probably have these around Molo. *A. m. scutellata* are smaller and more aggressive bees found in the lowlands of Kenya. In my immediate vicinity, we have them in Nakuru and Baringo Districts. We may also have hybrids of these two types. Much research has yet to be done in Kenya and Africa as a whole on mapping various bee races and their characteristics.

(Reference: *Beekeepers' Guide Book*, National Beekeeping Station, Nairobi)

B. STINGLESS BEES

There are also species of stingless bees in Kenya. Beekeeping with stingless bees is called meliponiculture. These bees produce honey prized as medicine. The honey is very sweet and liquid compared to *Apis mellifera* honey. Stingless bees may be kept in small hives but are not yet kept commercially in Kenya. There is renewed interest in these bees with the discovery of a new species in Kakamega by a scientist working for the National Museums of Kenya. Beekeepers in other African countries keep stingless bees. There is even a very fascinating webpage about someone in Moshi, Tanzania, who keeps them. Stingless bees are not, however, the main focus of this book; honeybees are.

(N. B. *See Annex 2 for more information on stingless bees in Kenya.*)

WHAT BEES DO

- Bees forage on floral nectar and pollen. They don't compete with livestock for food such as grass.
- Bees help the pollination of flowers, plants, and crops.
- Bees increase the quantity and quality of flowering crops, e.g., coffee, papaw, banana, avocado, macadamia, mangoes, etc.
- Bees fly three to five kms from the hive to collect nectar, so you benefit from other people's flowers.

- Bees produce honey, beeswax, and propolis (used in medicines) as well as other products such as royal jelly, pollen, bee venom, and bee brood. Honey, beeswax, and propolis are the three products currently exploited in Kenya. They are discussed in this book because of their importance.

ALL ABOUT HONEY

- Honey contains sugars (80-85%) that are easily absorbed by the body—good for young, old, and sick. It is made from nectar, a sugary secretion of flowers. Nectar contains 70-80% water. To make honey, bees add enzymes and reduce the water content to that of honey. (Good honey contains less than 19% water.)
- Honey is very good energy food. Use it as sweetener for food (cake, chapati, bread, etc.) and drink. Try some in your tea.
- Honey has medicinal properties. Use it for coughs, ulcers, wounds, and sore throats. (*See recipe for honey cough syrup in Annex 4.*)
- Honey has a high market value. It is used to make local beer and is also a food preservative.
- Honey is always in demand. It is a good source of cash.

INFORMATION ON BEESWAX

- Beeswax is the substance bees use to make their combs. Worker bees secrete it from special wax glands on the abdomen (underside).
- Beeswax is used in cosmetics, soaps, leather and wood preservation, candles, ointments, batik, shoe polish, and foundation sheets for Langstroth hives.
- You can turn honeycombs into wax. Simple instructions and recipes for using the wax are found in Annex 4.
- With a little training, you can make and sell candles, body cream, and shoe polish made with beeswax. Farmers trained by the Baraka Agricultural College/SHDI beekeeping outreach project sell homemade body creams and generate additional income for themselves.

WHAT IS PROPOLIS?

- Propolis is a resin that bees collect from plants. It is black and sticky.
- Bees use propolis to cover the inside of the hive and fill in the cracks.
- It has medicinal value, e.g., as an antibiotic.
- If you chew propolis, it tastes bitter but is good for the throat and chest.
- There is an export market for propolis. Annex 4 has a recipe for making propolis ointment.

RESIDENTS OF THE BEEHIVE

In the beehive, there are three types or castes of bees. It is important for all beekeepers to recognise them. The first caste is called the Queen: a sexually mature female. The second caste is called the Drone: male bees. The third caste is the Worker: an immature female. Details of all three castes plus the brood are given below.

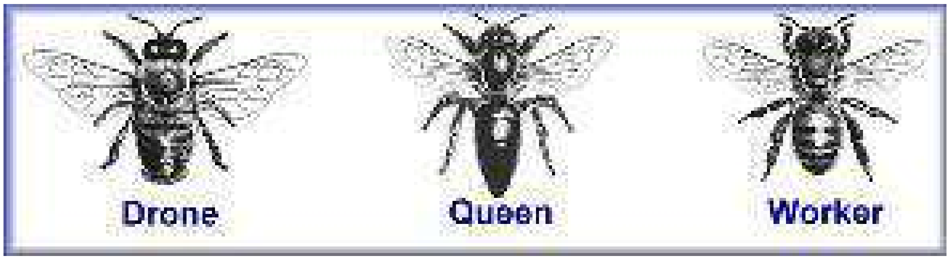

Figure 1: Types or Castes of Bees
Only the worker bee does all the work in the hive!

1. The Queen Bee

The queen is a sexually mature female. There is only one queen in the hive. Her job is to lay eggs that hatch into other bees. She can lay up to 2,000 eggs per day. Queens can live for up to five years, but the most productive period of their life is the first two years.

The queen may be hard to find in the hive, but you can recognise her by her length. She is long and slender and her wings reach only halfway down her back. She is far bigger than the numerous workers but must not be confused with drones, which have a square body and very large eyes.

2. The Drone

There can be several hundred drones in a hive depending on the time of year. In times of food shortage, workers throw drones out of the hive. When a new queen starts life, she mates only once with up to ten drones outside the hive, high up in the air.

Drones have very large eyes used to spot the queen during mating. Their major task is to mate. They have no sting. Drones look large and square. They make a loud buzzing noise when they fly.

3. The Worker

Most bees in the hive are worker bees. As their name suggests, they do all the work. Workers feed the young, feed the queen, guard the hive, and collect honey/pollen/propolis and water. Workers undertake different tasks in the hive depending upon their age after hatching. Worker bees have a sting and use it to defend their nest. Workers can number up to 60,000 in a very strong colony of bees. The lifespan of a worker bee depends on colony activity: during times of high activity, e.g., when collecting nectar to make honey, they can wear out in a few weeks. In the dormant season when there isn't much work, they can live for up to six months.

4. The Young or Brood

The queen lays an egg that looks like a grain of rice. To see eggs, hold combs from the centre of the brood nest (centre of the hive) up to the light. Look carefully at seemingly empty cells; they usually have eggs at the bottom.

The eggs develop into larvae that look like white maggots. The larvae are fed on brood food (produced in glands on the head of nurse worker bees) and some pollen/

honey. (Pollen is the powdery substance produced by the anther of flowers and is rich in protein. It is the coloured substance seen on bees' legs and stored in combs.) Larvae are later sealed in the comb where they turn into adult bees. During the changing process, they are called pupae. Thus there are three stages to adulthood: egg, larva, and pupa.

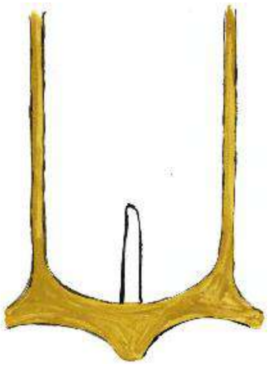

Figure 2: Bee Egg

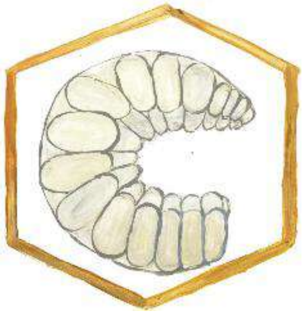

Figure 3: Bee Larva

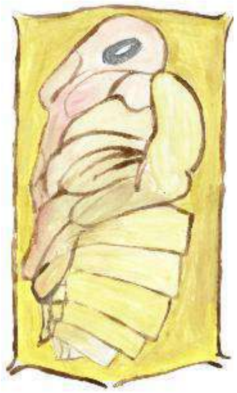

Figure 4: Bee Pupa

Bee Caste	Eggs-Larvae (# of days)	Larvae-Pupae (# of days)	Pupae-Adults (# of days)	Total Days (Egg to Adult)
Queen	3	4.5	7	14.5
Worker	3	4.25	11.25	18.5
Drone	3	6.5	14.5	24.0

Table 1: Duration of Development Stages
Table shows time spent in each stage of development for African bees, specifically *Apis mellifera scutellata*.
(Source: Kigatiira, K. I., *Beekeeping for Beginners*. See Annex 9 for more details.)

Part Two
GETTING STARTED: THE BASICS

TYPES OF HIVES

A hive is the box or some other container where bees live. Without a hive, you cannot be called a beekeeper!

Photo 1: Wild colony of bees
Wild colony of bees nests under a manhole cover.

Different types of hives are used in beekeeping. Some are:

- Pot hives or clay pots (See Photo 2.)
- Log hives (See Photo 3. They are made from hollowed-out tree trunks.
- Basket hives (See Photo 4. They are made from woven sticks smeared with mud.)
- Kenya Top Bar Hive or KTBH (See Photo 5.)
- Langstroth hives (See Photo 5. These use frames.)

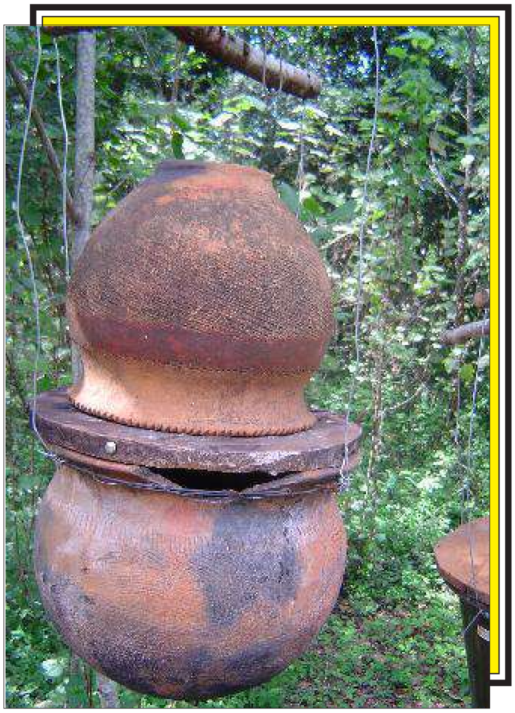

Photo 2: The pot hive
The pot hive has a super or honey chamber on top.

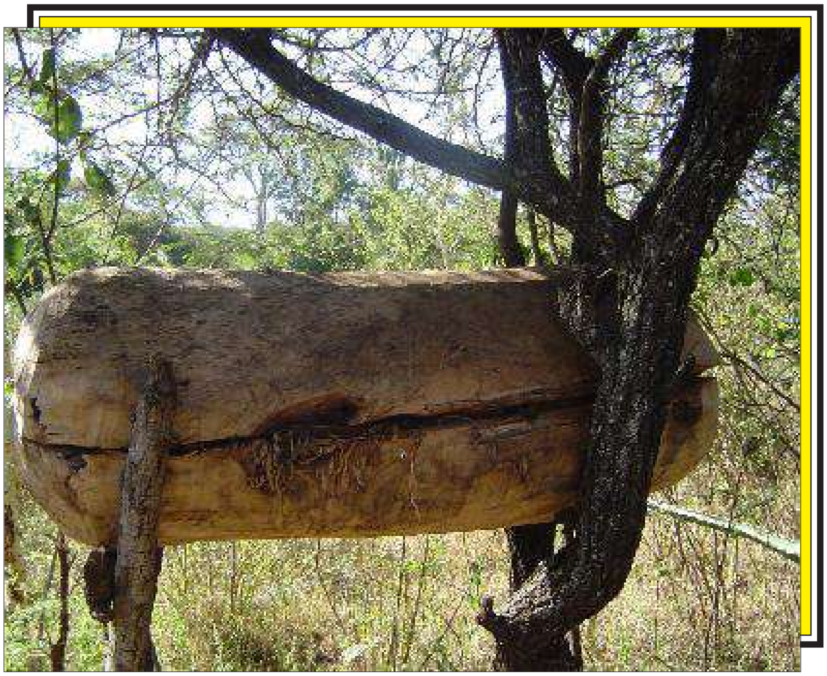

Photo 3: Traditional log hive
A traditional log hive located in Koibatek District.

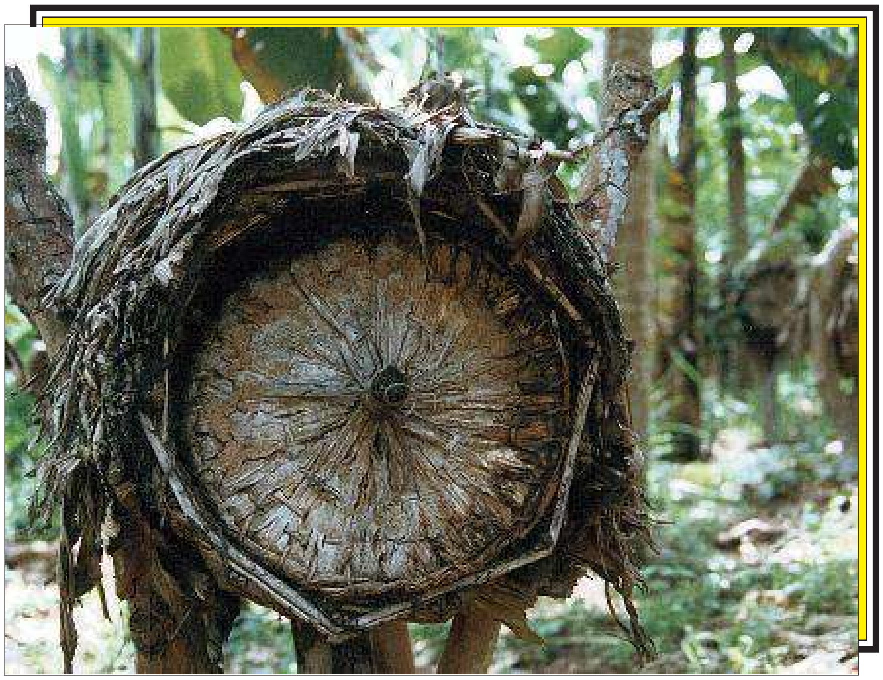

Photo 4: Ugandan basket hive
The lid is removed to harvest honey. Care in harvesting yields top quality honey.

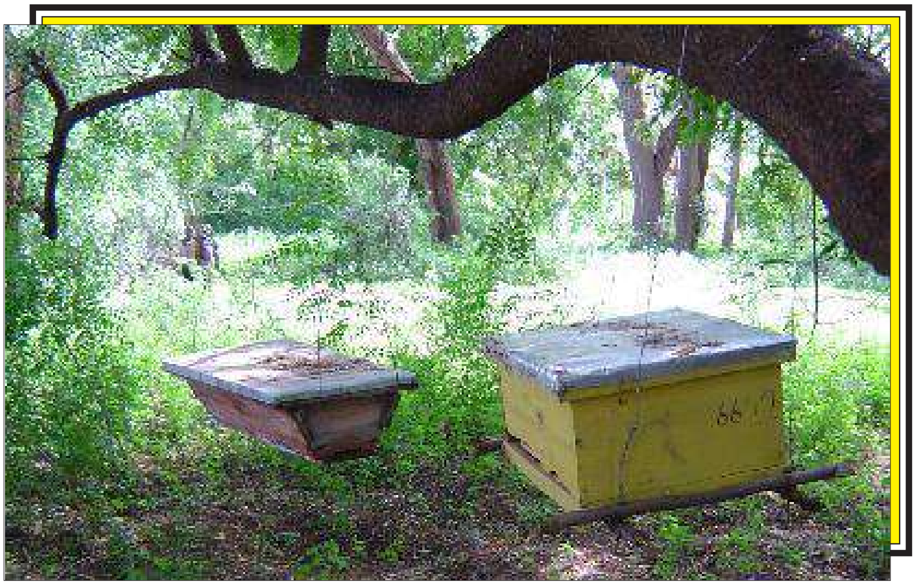

Photo 5: Examples of KTBH (left) and Langstroth hive (right)
Hives in Turkana District. Beginners should go for the simpler and cheaper KTBH.

In this book we will give you measurements of the KTBH and a cheaper mud/stick hive, which also has top bars. These two hives are relatively easy and inexpensive to make. They are very good hives to start with in beekeeping in Kenya. The Langstroth-type frame hive is more complicated and expensive to make/purchase. It has been promoted in Kenya the past few years. Each type of hive has advantages and disadvantages.

ADVANTAGES OF TRADITIONAL HIVES

(e.g., logs, pots, and baskets)

- They are relatively easy and inexpensive to make.
- You can use local materials and traditional knowledge for making them.
- Bees often prefer these hives to more 'modern' types.
- With care in harvesting and handling, honey quality can be very good.

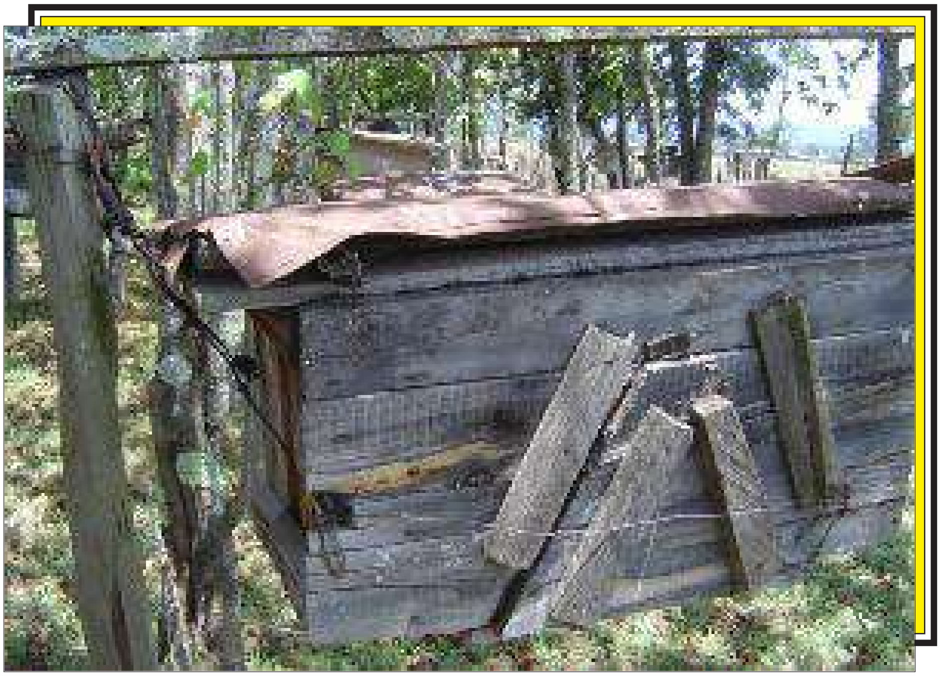

Photo 6: Crudely constructed box hive
This box hive is located in Londiani, Kericho District.

DISADVANTAGES OF TRADITIONAL HIVES

- Bee management is difficult.
- Logs are no longer readily available to make hives in Kenya.
- They can be difficult to harvest.
- Yields and quality of honey can be lower than top bar and frame hives.

ADVANTAGES OF TOP BAR HIVES

- Bee management is possible, making it easy to harvest and manipulate colonies to maximise honey production.
- They are easier and cheaper to make/buy than frame hives.
- They are easier to harvest than traditional hives, creating the potential for better honey quality.
- Honey extraction is easy compared to frame hives, as there is no need for specialised equipment.

DISADVANTAGES OF TOP BAR HIVES

- Combs in the top bar hive are not supported. They can break if not handled carefully.
- Wax is harvested with honey forcing bees to build more wax to replace harvested combs, which results in lower honey yields but more harvested wax.
- The volume of the KTBH, like the traditional hive, is fixed, so the hive can fill very quickly in the honey season. An overcrowded hive can swarm leading to reduced honey crops.

ADVANTAGES OF FRAME HIVES

(e.g., Langstroth)

- It has frames that make the combs very strong especially when transported.
- The honey is extracted by centrifuge, returning wax to bees with potentially much greater honey yields.
- The space of the hive can be added to easily by adding more supers. Additional space is important during honey flow. Supers can be removed when the honey flow is over to contract hive size.
- Using a queen excluder between the bottom brood box and the honey supers means honey is separated completely from the brood and is of high quality.

Photo 7: The Langstroth frame hive

DISADVANTAGES OF FRAME HIVES

- Frame hives are more expensive to make than traditional or top bar hives.
- They are more complicated to make and manage, as they require more spare parts such as frames, foundation starter sheets, and supers, which are costly and not readily available to beekeepers in Kenya.
- Extracted/unextracted honey supers are prone to damage by wax moths, ants, and rodents.
- Frame hives need a centrifugal extractor to extract honey. Extractors are not readily available in Kenya and are expensive to purchase. Beekeepers are often forced to cut honey from frames just as in top bar hives.
- Frame hives require a greater level of investment (time and money) as well as beekeeping skills to make them pay off.

In recent years, Langstroth frame hives have been promoted in Kenya. There has been considerable hype around their use with many people automatically believing they are the best hives. Frame hives are not necessarily better than either traditional or top bar hives. Frame hives have the potential to be better in terms of higher yields and better quality honey **if** managed well! Unfortunately in many cases, too much emphasis is placed on the technology and not on building beekeeping skills and capacities.

Photo 8: Beekeeper in Kitui District
Beekeeper poses in front of Langstroth hives.
Donor support provides many such hives in Kenya.

As a beginner, you can easily start beekeeping in Kenya using the intermediate technology top bar hive such as the KTBH—unless someone donates a Langstroth to you! Once you have beekeeping experience, understand why you should purchase a frame hive, and have money to do so, and then use the frame hive by all means, but you should be aware that there are no automatic increased yields from having a Langstroth. The hive is only as good as you manage it. I have produced up to 40kgs per KTBH while many people who purchased Langstroths have produced little or nothing! So understand bees and bee management and you will produce good crops of honey no matter the type of hive you have. When you have better skills and knowledge, move to the more expensive hive, which can then give better yields.

This book deals primarily with top bar hives because they are the easiest for beginners in Kenya who often do not have a lot of money to spend for starting up. The principles of beekeeping are, however, the same for both top bar and frame hives and issues of bee management are very similar.

THE KENYA TOP BAR HIVE

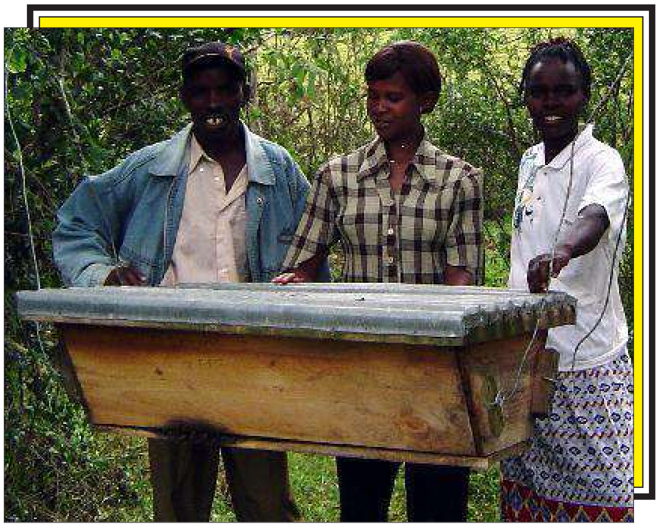

Photo 9: The Kenya top bar hive
The Kenya top bar hive or KTBH was developed in Kenya and is now used around the world.

The Kenya Top Bar Hive (KTBH) was developed in Kenya and is a good option for many Kenyan beekeepers and other beekeepers from around the world. KTBH are also used outside Africa, e.g., the United States of America. Its big advantage is its simplicity and low cost. You also get to produce beautiful comb honey for home consumption or for sale. Figure 5 shows the construction drawing of a KTBH. Remember to follow the measurements—the top bars, in particular!

Very important: You must copy **exactly** the measurements we give you. When it comes to bee management, it is important that all your hives are the same size so you can move top bars from one hive to another. This allows better bee management. In particular, watch the dimensions of the top bar: 3.2cm wide and 48.3cm long. The ridge at the centre of the top bar is coated with beeswax to give bees a line to follow when they build their comb. The smell of beeswax also attracts bees to enter a new hive.

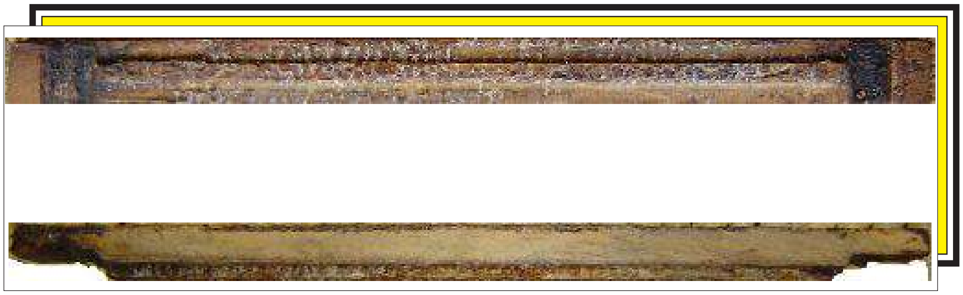

Photo 10: Top bar
There are two views provided for the top bar: top view (top) and side view (bottom).
Notice the ridge protruding from centre in side view.

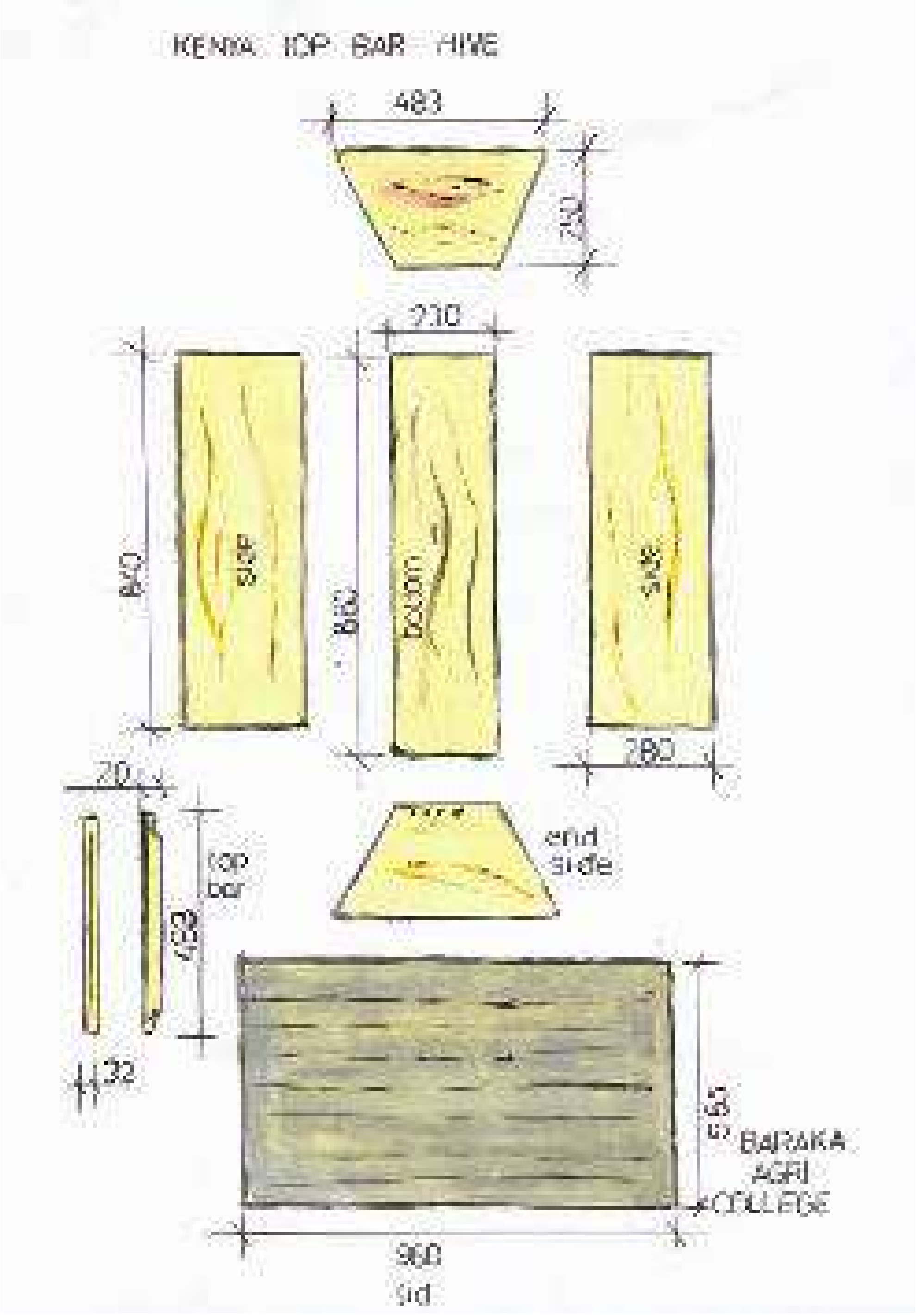
KENYA TOP BAR HIVE
483
250
230
840
side
860
bottom
side
280
20
top bar
483
32
end side
565
BARAKA AGRI COLLEGE
960
lid

Figure 5: Construction Drawing of KTBH
Measurements are in millimetres.
(Source: Baraka Agricultural College Workshop)

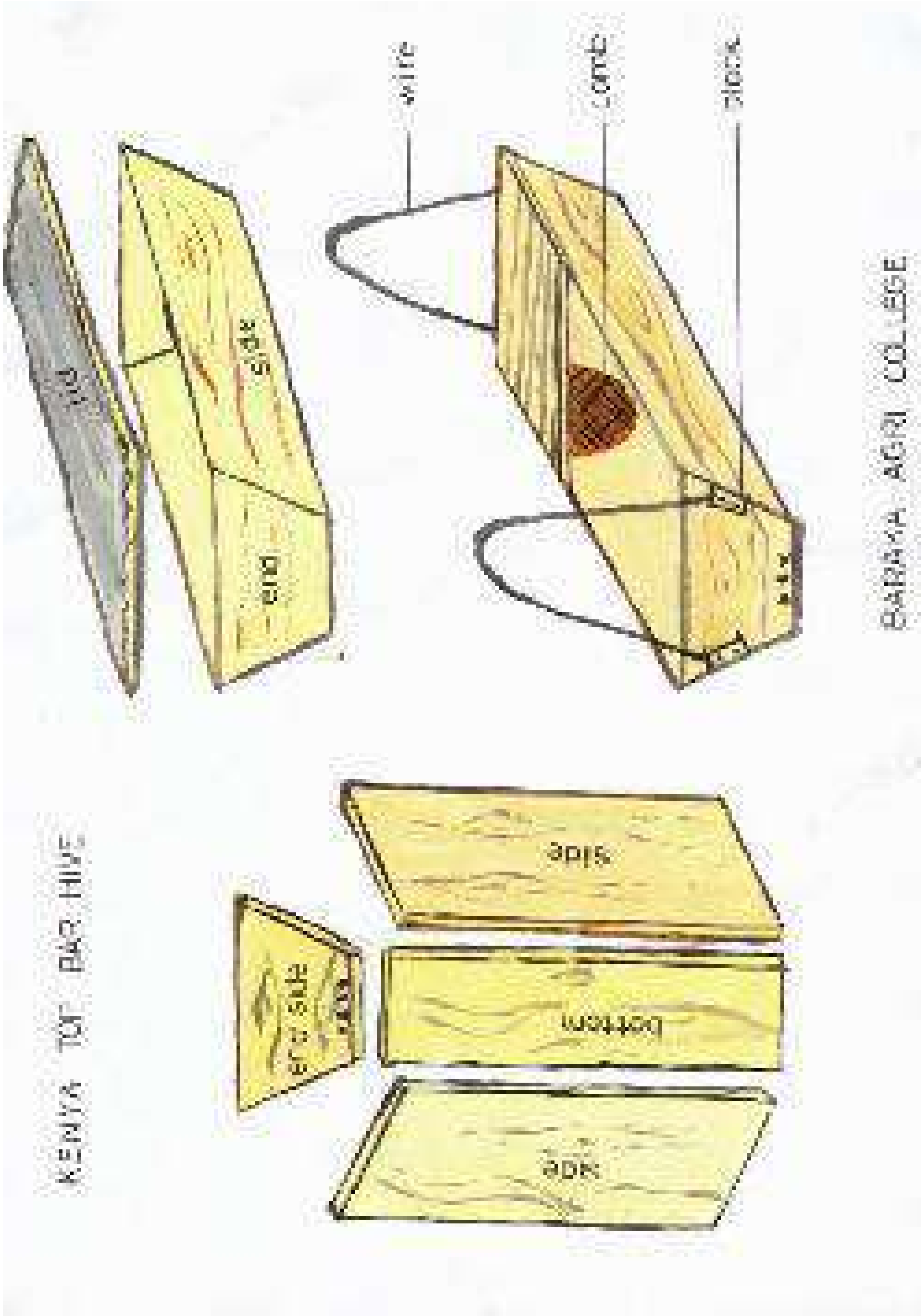
KENYA TOP BAR HIVE
lid
side
end
wire
comb
block
BARAKA AGRI COLLEGE
end side
side
bottom
side

Figure 6: Assembly of KTBH
(Source: Baraka Agricultural College Workshop)

TIPS FOR MAKING A GOOD KTBH

- Use timber ¾ inch thick (19mm) to make the hive.
- Utilise dry timber to prevent cracking and warping.
- Make use of whatever timber is inexpensive and available locally. Pine, cypress, and *Grevillia robusta* are commonly used, but if you have your own wood, use it.
- Get the right size for the hive body of the KTBH though some drawings show a slightly longer hive. The measurements given are those used at Baraka Agricultural College.
- Follow the measurements of the top bars exactly. It is **very important**. Top bars are made a specific measurement to ensure that bees build one comb per top bar making combs easy to inspect.
- Top bars should be the right width and fit the hive body well.
- Don’t forget to wax top bars to attract bees. Use a paintbrush to apply melted beeswax onto protruding ridge at the centre of the top bar. The strip of wax guides bees to build straight combs that are easy to inspect and harvest.
- All parts of the hive should fit together properly. There should be no holes.

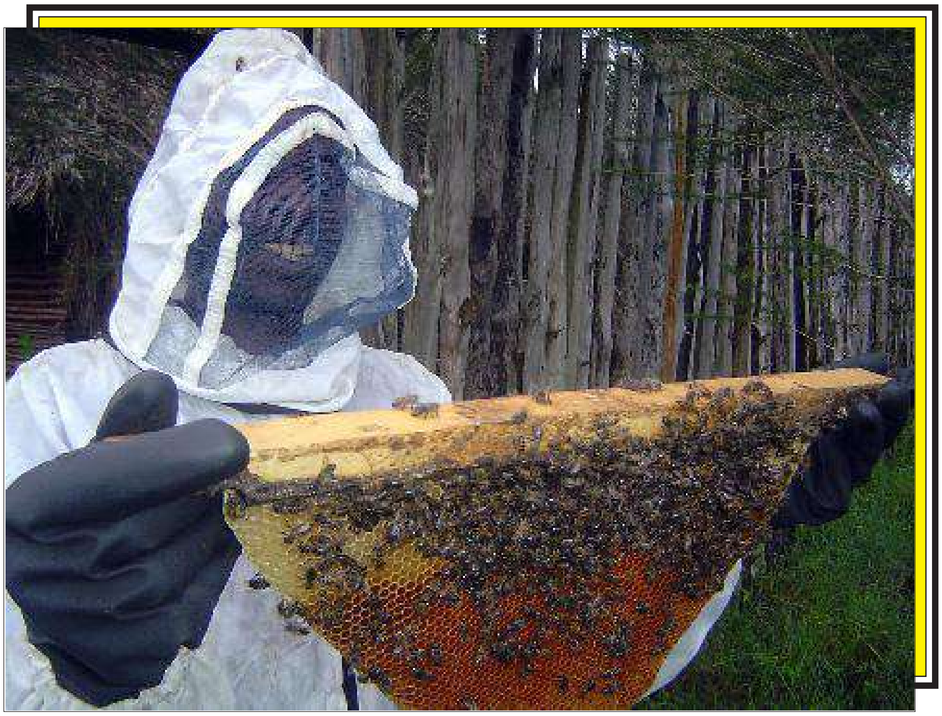

Photo 11: Working with bees in a KTBH

Notice one comb per top bar, which allows modern bee management by the ability to move combs without breaking them.

THE MUD/STICK HIVE

A mud/stick hive is the answer to all of you who say you have no money to start beekeeping. No excuses allowed. Beekeeping is about skills and knowledge. Even with very little money, you can start beekeeping by making your own equipment.

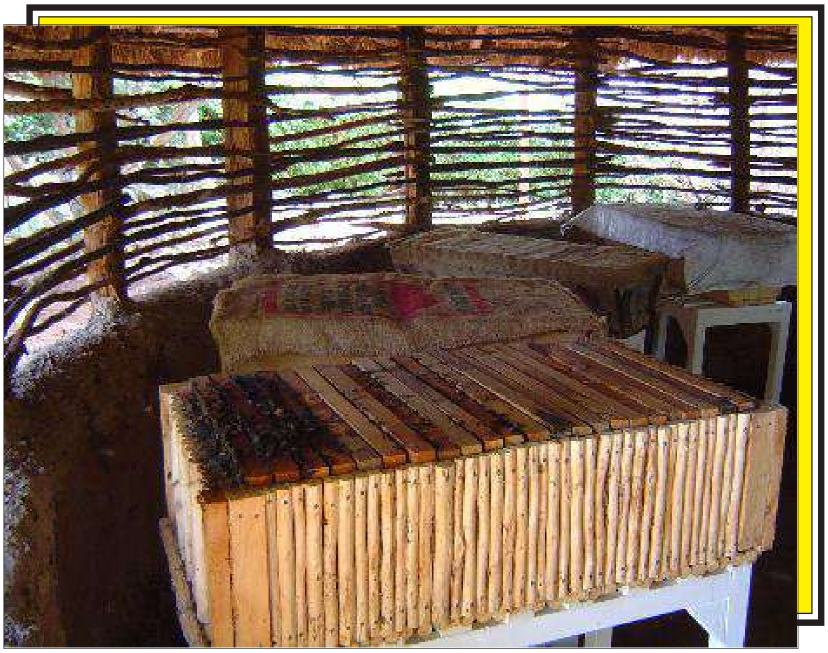

Photo 12: Mud/stick hive inside a bee house
The bee house is also made from mud and sticks.
The hive body is made from sticks with gaps between them filled with mud.

For the mud/stick hive, use the same dimensions (length, width, and height) as the body of the KTBH, but make the body square. For the body, use about 50 long and 30 short straight sticks. First, make a frame of sticks for the bottom. You can also make a timber frame and add sticks as shown in Photo 11. To this frame, tie/nail on other sticks. Use wires or nails to do this. If you bind them tightly, you can have a very strong frame. Plaster the inside of the hive with mud and cow dung and leave to dry. Top bars fit along the top of the hive just like the KTBH. Remember to make some small holes at one end of the hive as the entrance. As a roof, you can use *mabati* (tin). The mud/stick hive works best, however, under shelter, e.g., a bee house, because of potential damage from heavy rains. If inside one, just use a sack for a cover.

BARAKA
DEMONSTRATION
APIARY
BEE HOUSE
VANTAGES: SHADE FOR HIVES,
CURITY FROM THIEVES/PESTS, EFFICIENT
ACE UTILISATION, APPROPRIATE FOR
SMALL FARMS.
PES OF HIVES
OP BAR.
TRADITIONAL FIXED COMB.
FRAME HIVE LANGSTROTH
B D U

Photo 13: Bee house at Baraka Agricultural College
Baraka Agricultural College is in Molo.

TIPS FOR MAKING A GOOD MUD/STICK HIVE

- Shelter hive from heavy rains.
- **Important**: make top bars the same measurement as that of the KTBH, which allows transfers from one type of hive to the other.

The two hive designs mentioned are an improvement of the traditional log and basket hives because with top bars, combs are moveable, which allows management of bees. Top bar hives are not as expensive or difficult to manage as Langstroth hives making them more suitable for use by small-scale farmers.

You can also be innovative and use other materials to make the hive body. In Zambia, beekeepers use a concrete KTBH hive, which works very well. The body is made from concrete and set on a stand. Top bars are added on top of the hive and covered with sacking material. The Zambian concrete hive is placed under grass shade cover to keep out rains and the hot sun.

APIARY LOCATION

An apiary is a place where hives are kept. Keep a limit of not more than 20-25 hives per apiary depending on availability of bee forage. Bees forage in a radius of about three to five kilometres from the apiary, so if you want to keep more than 20 hives, find another site three or more kilometres away from the existing one. Plant a good high hedge around your apiary using a shrub such as kei-apple. As you wait for the fence to grow, you can use off-cuts—waste timber from sawmill—to make a fence. The hedge separates bees from people and animals, which is important in Kenya as our bees can be aggressive. Apiary siting is very important on small farms to ensure safety and comfort living with bees.

Photo 14: Apiary fenced with off-cuts
The fence, made of waste timber, is to screen bees away
from people and animals. (Nakuru)

- Choosing a good site to hang your hives is very important.
- If you choose a poor site, bees may sting people and animals.
- If the site is insecure, hives and honey may be stolen.
- If you live in a hot area, your hives will need shade as well as water.
- If you live in a cool area such as the highlands, only minimal shade is required or the bees will be cold and damp.
- Get a good balance between light and shade. A bright apiary without direct sunshine appears to be best. Bees are most active then.

An ideal site to establish an apiary would be:

- Away from humans and livestock, dwelling areas, roads, and public areas.
- Safe from strong direct sunshine, windy areas, and theft.
- Accessible to nearby water and bee plants, i.e., provide water and bee forage shrubs, trees and/or crops.

- Away from swampy areas and smelly places.
- Near a good source of nectar, e.g., forests, trees, or nectar-bearing crops.

So now you have made a hive or two. Did you follow instructions carefully? Are your top bars the right measurement? Have you chosen the site for the apiary? If you are still not sure, ask someone who knows. Remember: once bees enter hives, it will be more difficult to change things. The bees may become aggressive or the hive will be heavy with brood and honey. We now come to the procedure for putting hives in the apiary.

HIVE PLACEMENT

A. HANGING HIVES

(N. B. *Not recommended for mud/stick hive, as it is not strong enough. See below for placement.*)

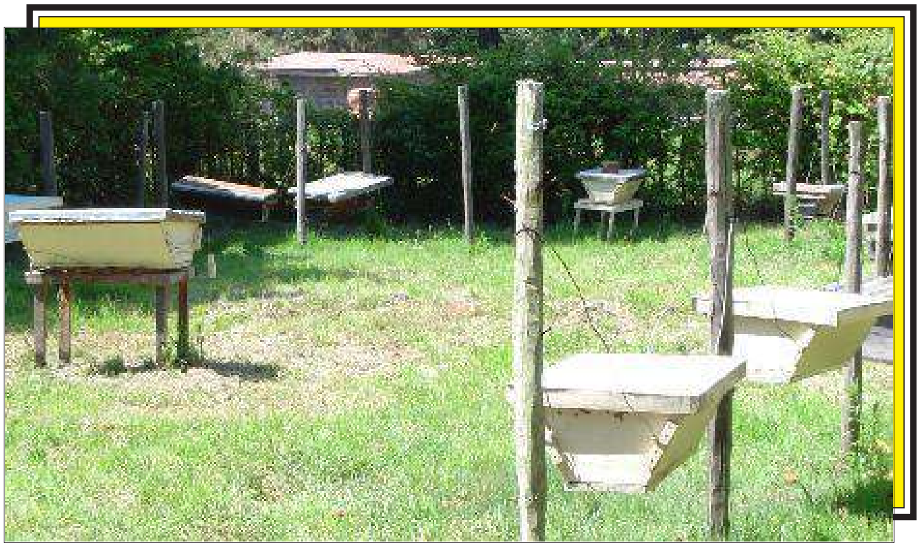

Photo 15: KTBH hanging in foreground
Another one is on a stand, background left.

- Use two strong and heavy posts, each of about 2.5 to 3 metres long.
- Dig two holes about ¾ metre deep and 2 metres apart—or take two strides.
- Pack soil and stones around posts. Make sure posts are very firm. Think ahead to when the hive will be heavy with honey. If they are not firm, they will fall over later.
- **Remember**: once bees enter the hive, it will be difficult to make changes!
- Now that holes are dug and posts are in position, use wires to hang the hive between the two posts.
- **Remember**: put wire around the back of the posts as shown in Figure 7.
- **Remember**: hang the hive at waist height and keep it level to ease work and not strain your back.

Take time to do things properly and you will be happy later on!

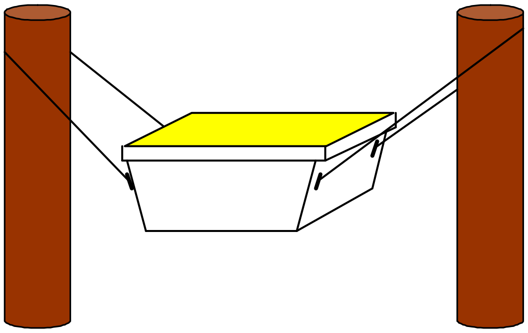

Figure 7: Hanging KTBH Posts
The posts are two metres apart and the hive about one metre from the ground.

You can also hang hives between two trees or a post and a tree. Follow the same instructions, but remember to use strong heavy nails. When hanging more than one hive, you can follow some of the patterns below:

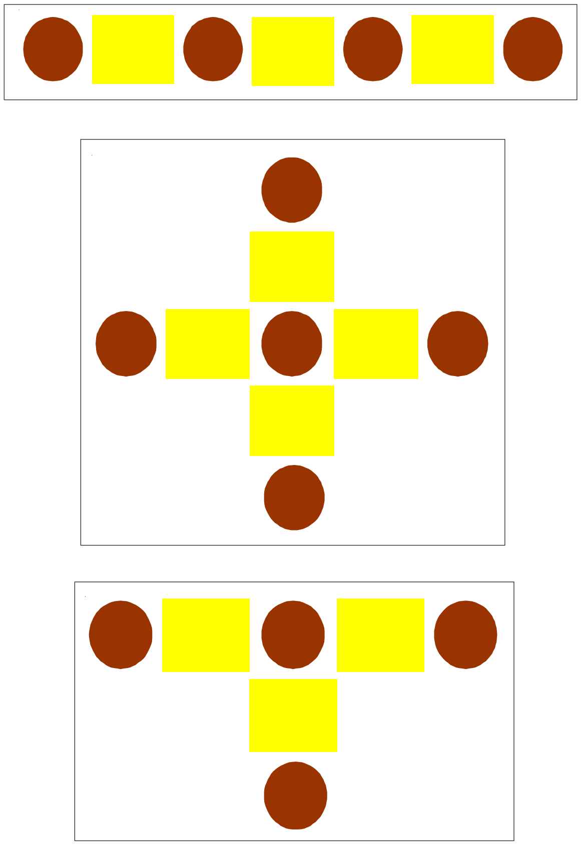

Figure 8: Hanging patterns for KTBH
Above are the line pattern (top), cross pattern (middle), and T-pattern (bottom).
(□ = hive, ○ = hanging post)

B. HIVE PLACEMENT

- If you don't want to hang hives, you can place them on stands. It is best to use this method for the mud hive, as it is not strong enough to hang.
- **Remember**: the stand should be sturdy and high enough for the hive to be at waist level.
- Put legs of the stand in cans of used engine oil to prevent pests from getting into the hive.
- Hives on stands are more prone to attack by the honey badger. (See list of pests and diseases in Part Three.)

(N. B. *Putting a number of hives under a shelter or into a bee house is also an option. Just make sure that bees have a way of getting in and out of their hives. A bee house is useful because you can lock it to prevent thieves from stealing honey or children from interfering with hives. A bee house can be a simple mud hut with holes in the walls for bees to come and go. From my experience with bee houses, get a balance between shade and light: no strong direct sunshine but also plenty of light. The balance of light and shade will depend where you are. In a cold highland area, you need less shade than in the hot lowlands.*)

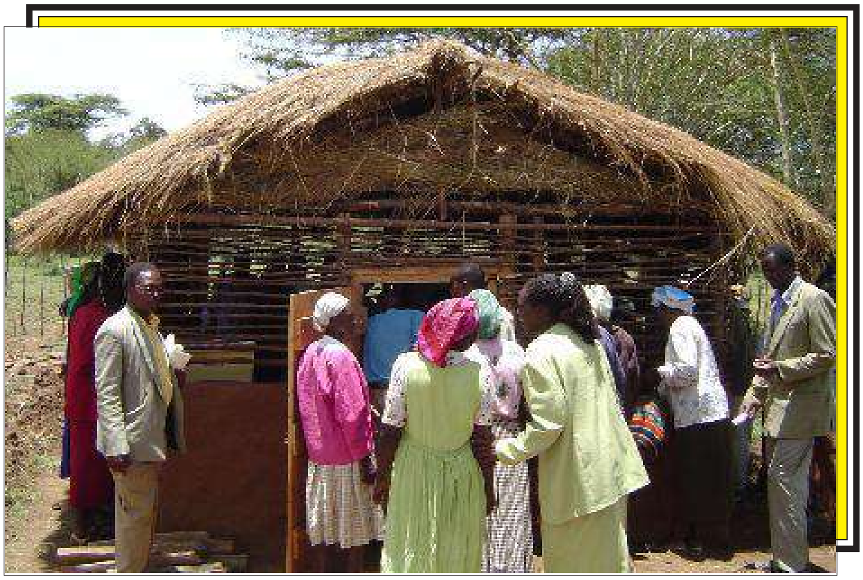

Photo 16: Newly constructed bee house
Farmers in Koibatek and Ministry of Livestock staff inspect the structure.

GENERAL APIARY MANAGEMENT

- Keep apiary clean: cut grass short and trim branches that reach hives to prevent pests, e.g., ants, from crawling into hives and disturbing bees.
- Grease wires holding hives to keep off crawling pests, particularly ants.
- Do not make noise around the apiary as it can agitate bees.
- Supplement the bee's source of nectar by cultivating certain plants around the apiary, e.g., bananas, sunflower, mangoes, citrus, coffee, eucalyptus, pawpaws, passion fruits, croton, acacia, bottlebrush, *Calliandra*, etc. Trees will also give shade to bees and provide a screen between bees and nearby people and animals.
- Remove old combs from the hive to prevent wax moth damage.
- Do not spill honey near the apiary. It will attract pests and cause bees to fight with their neighbours, i.e., bees from adjacent hives.
- Inspect hives regularly to monitor bee progress and ensure no pests are attacking them. (See list of pests and diseases in Part Three.)
- When bees are making honey, inspect hives once a week by opening them up. This is important, so you ensure bees have enough space to store honey. With the KTBH hive, remove honey to make space; with the Langstroth, add an extra super. At other times, inspect hives once a month. You can also simply look at hives almost daily without opening them.
- Observing bees for a few minutes can tell you a lot. Are the bees collecting pollen? If they are, it indicates that bees are feeding the young. It also shows they are healthy and have a laying queen.
- Provide a permanent source of water for bees near the apiary. This is important as lack of water can contribute to absconding in the dry season.

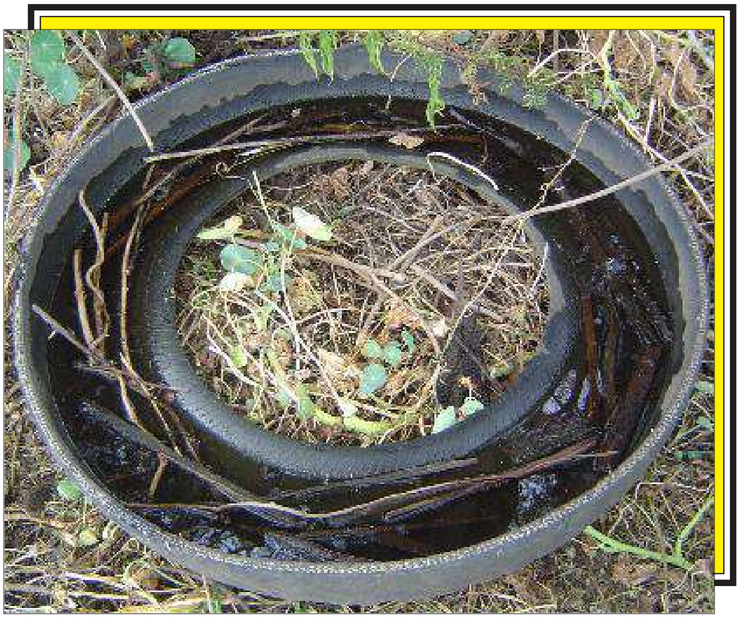

Photo 17: Water for bees

Provide water slightly away from apiary and not in their flight path to avoid bees defecating in the water, which spreads disease. Place sticks for bees to stand.

BEEKEEPING EQUIPMENT AND HOW TO USE THEM

1. Hive

The hive is for housing bees. (KTBH and other types have already been mentioned.) It is the most important piece of beekeeping equipment. Without a hive, you cannot be a beekeeper!

2. Bee Suit

The bee suit consists of the following items that give protection from stings:

- Veil: for covering head and face
- Overall: for covering the rest of the body
- Gloves: for covering hands
- Gumboots: for covering feet

Photo 18: Good gloves
Good gloves are essential when working African bees.

(N. B. *Farmers are able to improvise their own bee suits and make them out of sacking material, which cost about KSH300 or US$4.17 to assemble.*)

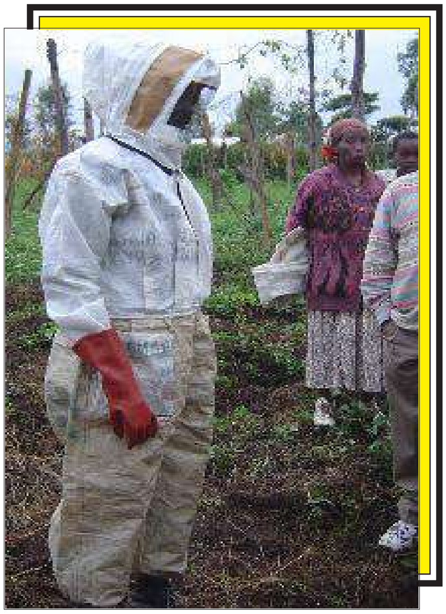

Photo 19: Farmer in Nakuru wears an inexpensive bee suit
It is made from used sacks combined with rubber gloves.

3. Smoker

The smoker is a vital piece of equipment as smoke is used to control bees and keep them from becoming aggressive when handled. When lighting a smoker, use dried grass to light it. When burning well, put in some damp grass, so that it produces a thick cool smoke. Do not allow the smoker to spark or flame, which will burn bees. An alternative smoker fuel is dry cow dung. Be careful when extinguishing the smoker that you do not start a fire!

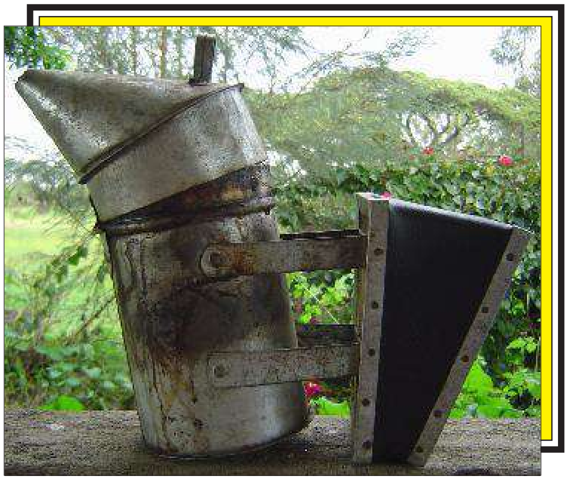

Photo 20: Smoker

(N. B. *A simple smoker can be made from a tin can punched with holes and with a wire handle attached. The beekeeper blows smoke onto bees with his/her breath.*)

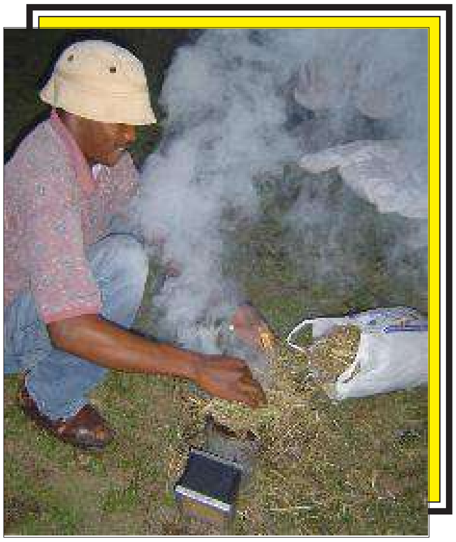

Photo 21: Lighting the smoker
Use cool smoke from damp grass.
Do not use toxic materials,
e.g., rubber or plastics.

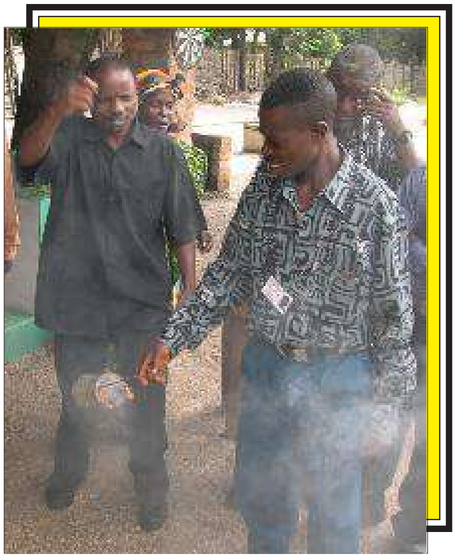

Photo 22: Lighting a tin can smoker
Beekeeping trainees in DR Congo light an
alternative smoker. In beekeeping, all
equipment can be improvised at low cost.

The hive, bee suit, and smoker are the most important pieces of equipment. The following items can also help:

4. Catcher Box

It is used for catching swarms and transferring bees from one place to another. It is a miniature hive.

5. Hive Tool

It is for lifting top bars. You can improvise the hive tool by using a strong knife.

6. Bee Brush

It is for brushing bees from combs when harvesting honey or at any other time you need to remove bees. You can use a feather as an alternative.

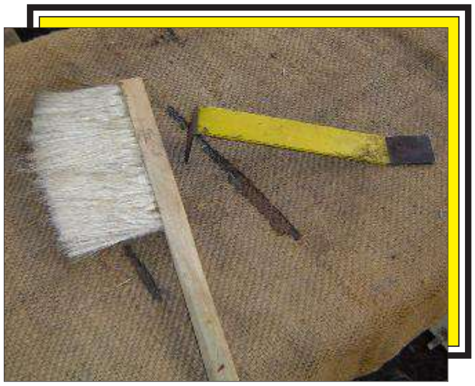

Photo 23: Bee brush (left) and hive tool (right)

A harvesting bucket is also essential as well as a knife for cutting combs when harvesting honey.

HANDLING BEES

Photo 24: Mick, my son

My eight-year-old son works African bees in Nakuru. He wears a full suit, gloves, and gumboots. If a young boy can handle bees without fear, adults should be able to do the same. When you understand bees, fear goes away!

Many people are afraid of bees because they sting. Many of us in Kenya hear of stories where bees attack and even kill people and livestock. (African bees can be very aggressive. Handle them carefully.) Like other forms of livestock, bees must be handled with respect and care. If managed properly, bees will not cause any problems.

Some bee types tend to sting less than others. A beekeeper who frequently inspects his/her hives can easily tell those bees, which are better and more docile. He/she can then eliminate more aggressive bees and breed from docile ones. (Not all African bees are aggressive. I have worked bees in Northern Somalia, which were very docile and didn't require a veil! Our bees in Molo are also generally docile.)

With frequent handling, bees appear "used" to inspection so less aggressive. A beekeeper will come to know the character of his bees. If you have many hives, number them and keep records of the bees' behaviour at each inspection. Eliminate aggressive colonies.

When handling bees:

- Always wear a bee suit and take the time to put it on properly. Many people do not do this and get stung when hives are open. If you don't know how to put on a bee suit, get someone to help you—maybe your experienced friend.
- Avoid wearing woollen clothes as they agitate bees, which can be stuck in them.
- Avoid drinking alcohol and using strong smelling soaps or sprays all of which may aggravate bees.
- Always use a smoker when handling bees. Smoke makes bees suck honey from the combs. It calms them down and masks alarm smells the bees use for communication. Do not let the smoker go out during the operation or the bees can become aggressive. Keep plenty of smoker fuel handy as you work. It is always better to have two smokers alight in case one goes out.
- Start with the least aggressive colonies always. This allows you to work in peace with the pleasant colonies first.
- Work gently and quietly. Do not knock or bang the hive as this makes bees angry.
- Always handle bees in the evening between 5.30pm and darkness. (Be careful initially. Once you know your bees and their aggressive tendencies, you will know the best time to handle them.) If bees become aggressive during evening hours, they have a chance to cool down before the following morning. They seem to be less aggressive in the cool of evening. For bees you've never handled before or for very aggressive bees, take extra precaution in handling them at dusk. Use a torch to see them.
- When handling, avoid crushing bees or making sudden movements. Work carefully and with confidence. Remain calm even if bees become aggressive. If bees appear to be getting out of control, close up the hive and try another day. If bees get into your veil, remain calm. Walk to a safe distance before rectifying the problem.

- Work hives with two or more people at a time. One person can lift out combs while the other uses the smoker. This allows better control of bees.
- Do not stand in front of hive entrance when examining hive. Bees flying in and out may become agitated to find their way blocked. Always cut down disturbance to bees in every way you can.
- Advise any onlookers to move away quietly if stung, covering their eyes. No running about waving arms as this can annoy bees.
- Remove bee stings from skin as soon as possible using a hive tool or your nail to scrape it off. Trying to pull it out tends to squeeze in more venom. Use smoke to cover the scent of a sting. When a bee stings, the scent attracts other bees to sting you again if you do not use smoke.
- On finishing the job, close up the hives. **Do not go directly to where you are to remove your bee suit.** Take a route via bushes or tall maize, sugarcane, etc., if these are in the vicinity. Rub against leaves to rid yourself of bees, which might be following or on you.
- If you are with a fellow beekeeper, check each other and see that there are no bees on yourselves before removing bee suits.

As you gain beekeeper experience, you will be able to judge the mood of bees more accurately and handle them calmly. For the first few times, it is better to get a friend to help you. I have a short video online to demonstrate how I handle bees in Nakuru. You can view it from the following address: www.apiconsult.com/video.htm.

Warning: Some people get an allergic reaction to bee stings. The normal response in most people is localised swelling and pain. Bee sting allergy symptoms often begin with a dry cough. The eye area begins to itch and swell. As symptoms progress, the patient begins to sneeze and wheeze and develops itchy skin irritation. These symptoms are warning signs of a dangerous condition called anaphylaxis. Anaphylactic symptoms include sudden anxiety and weakness, breathing difficulty, chest tightness, very low blood pressure, loss of consciousness, and shock. Anaphylactic shock can occur within minutes and result in death. A patient in shock needs immediate medical treatment.

People who are allergic to bee stings should avoid situations in which they are likely to get stung. Anaphylactic shock is treated with an injection of epinephrine (also called adrenaline), a hormone that stimulates the heart and relaxes the airways. You may combine it with an injection of antihistamine, which counteracts the histamine produced by immune cells during an allergic reaction. Those known to have severe bee sting allergies should carry a self-injection kit, including antihistamine tablets, for emergency treatment; however, they should still seek medical care after self-medication.

For most people, a few stings may actually be beneficial. Bee venom is used to treat arthritis and can be extracted from bees commercially. Contact ICIPE Nairobi for further information on bee venom. (See address in Annex 6.)

(N. B. *The health information on bee sting allergies is intended to supplement the care provided by your doctor. It is not intended or implied to be a substitute for professional medical advice.*)

WHAT IF NO BEES ENTER THE HIVE?

It often happens that bees do not enter the hive for quite a while. The new hive you have worked so hard to build stays empty. An empty hive does not produce any honey! A survey of beehives in Kakamega and Kirinyaga revealed that hive occupation varied between 40 to 70%. A recent survey in Kitui by DANIDA (2005) showed occupation rates of between 30 to 75%. At a farmers' research meeting at Baraka Agricultural College in 2006, three different groups of farmers working separately said low hive occupation rates was their biggest threat to beekeeping.

When this occurs, check the following:

- Have you waxed your top bars?
- Is the hive clean and pest-free?
- Are there any ants or rats nesting in it? If so, clean it out and re-grease wires. Bees do not like a dirty hive.
- Is it the swarming season? Ask a friendly beekeeper in your area. Swarming occurs when bees reproduce themselves and fly out of the old hive looking for a new home. During the swarming season, it is easier for hives to be occupied quickly. In Kerugoya town in 1995, bees occupied two hives at our extension office before we even had time to hang them.
- Have you chosen a good site for hanging the hive?

What will happen if it were not the swarming season? If it was but bees still did not enter the hive, what could you do? The beekeeper can put bees into the hive. There are different ways to do this:

A. CATCH A WILD SWARM OF BEES

For this operation, it is best to seek help from an experienced friend. The swarm has a better chance of staying in its new hive during a nectar flow. Don't waste time with swarms smaller than a person's head as they are more trouble than they are worth.

1. First, prepare your hive by smearing it with some melted beeswax, so it smells nice for bees.
2. Use your smoker and bee suit. Some swarms may be hungry and difficult to manage. Always be careful with strange bees!
3. Look for a swarm of bees clustering on a branch where you can catch them. Wear your bee suit and smoke bees very gently so as not to disturb them. (Don't smoke if you can avoid it.) Shake bees into a catcher box or similar container, e.g., cardboard box. If the queen falls into the box, the rest of the bees will follow. Wait 20 minutes or so. If bees return to the original site, try again.
4. Once you have the bees, leave the box in a shady place until evening. Make sure they do not become hot in the container, so covering with a damp cloth helps.
5. When evening comes, take home the bees. Shake them into your empty hive.
6. If you have other hives, give bees a comb with some uncapped honey and a brood comb with eggs from another hive to encourage them to stay.

(N. B. *At any stage in this operation, the bees may decide they don't like being disturbed and fly away. African bees frequently abscond.*)

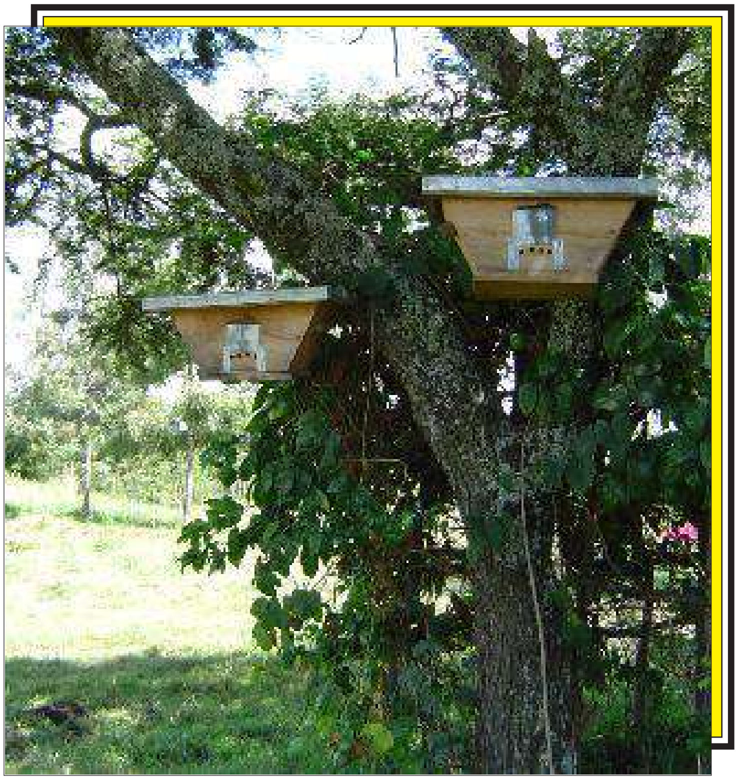

Photo 25: Catcher boxes for trapping passing bee swarms
This location trapped 17 swarms in 1 season alone. (Nakuru District)

B. TRANSFER BEES INTO YOUR HIVE

You can transfer bees into a top bar hive from a wild nest or from a traditional hive with fixed combs, which are not moveable. To do this, cut out combs one by one and tie them to the top bars of your hive. Get help on this one as it is not easy for the beginner.

C. MAKING A DIVISION OF AN EXISTING COLONY

You can make a division of an existing colony, explained later under the section, "The Beekeeping Year".

(N. B. *In many countries, you can purchase a starter bee colony called a nucleus or package of bees: a queen and worker bees with or without wax combs on frames. In the USA, you start most new beehives by ordering a package of live bees, usually in a three-pound size [1.3kgs]. Bee breeders supply these packages. In Kenya, we currently don't have bee breeders or suppliers of nucleus hives so the beekeeper is on his/her own at the moment in getting hives occupied.*)

CAUSES OF ABSCONDING AND ITS CONTROL

Absconding happens when bees leave the hive completely. It is very frustrating for the beekeeper that has done so much to get the hive occupied.

BELOW ARE SOME CAUSES OF ABSCONDING AND ITS PREVENTION

- Lack of food and water for bees. Leave food for bees when harvesting. If necessary, feed them with sugar syrup. Provide a source of water in the apiary.
- Frequent attacks by pests. Check hives frequently. Grease hanging wires to control ants.
- Unfavourable weather conditions, e.g., cold and rain. Make sure the bees have enough food.
- Careless handling by beekeeper, e.g., breaking combs, over-smoking, etc. Be more careful in handling them.
- Genetic character of bees. Some bees are prone to absconding, so don't breed those with this trait.
- Excessive heat on the hive—no shade in very hot sun. Put hives in a well-shaded place when in hot areas.
- Extreme cold and damp. Don't put hives under dense trees in high altitude areas. Give them a sunny place with minimal shade.

Part Three
THE BEEKEEPING YEAR

CONDITIONS FOR MAXIMUM HONEY PRODUCTION

We have seen how to get started in beekeeping: how to make the hive, choose the apiary site, hang the hive, and get bees. Now let's look at the bigger picture: how bees change throughout the year. You will get to know about the bee colony cycle. It is important for the beekeeper to understand this, so he/she can manipulate the bee colony to produce larger amounts of honey.

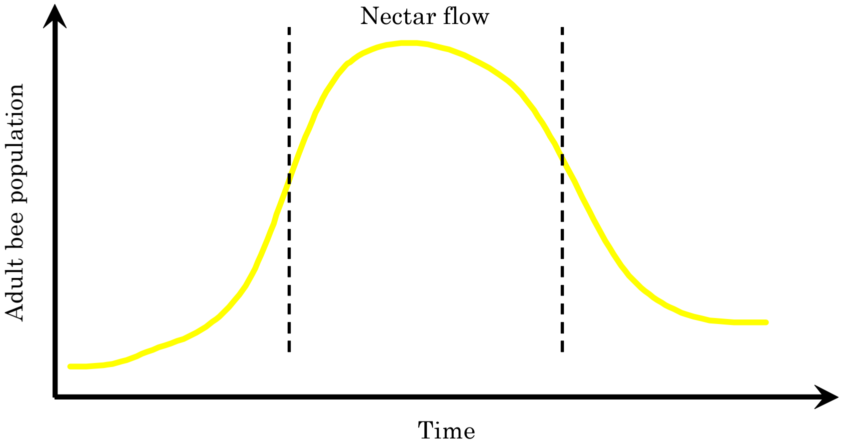

Figure 9: Conditions Required for Maximum Honey Yields

The chart shows that, if peak colony strength occurs at the time of nectar flow, a good honey crop results, providing nectar flow is good and weather conditions are right.

Your chief aim as a beekeeper is to get peak colony strength at the right time, i.e., during main nectar flow. Imagine this: peak strength of the colony comes when flowers are finished, which means bees will sit in the hive with little work to do. Manipulating the colony to be at peak strength at the right time is fundamental to good beekeeping. Good flying conditions, i.e., dry weather, for bees are also important during nectar flow to get maximum yields. This factor is, however, beyond your control as beekeeper.

Figure 9 also shows that colony size changes over time. When food is plentiful, there are more bees because workers feed the queen bee more food and she lays more eggs. When food is scarce, workers feed the queen less and the number of bees in the colony shrinks.

The bees' behaviour is very sensitive to their environment. When there is plenty of food, bees make more honey for eating when there is less food. The beekeeper shares in this stock of food. When harvesting, always remember to leave some for the bees. Don't be too greedy!

EXAMPLE ONE: KIRINYAGA

As with most parts of Kenya, Kirinyaga District has huge variation in climate. It stretches from the snow-capped peaks of Mt Kenya to the semi-arid Mwea Division. In between, there are tundra, bamboo, montane forest, tea, and coffee zones. Conditions for bees vary widely in all these areas. Things as climate and vegetation can change totally over a few kilometres. As with the changing environment, yield potentials and harvesting seasons for bees differ broadly. It is, therefore, very important for a beekeeper to be knowledgeable of the environment where he keeps bees because circumstances change dramatically over short distances.

In Kirinyaga, there are two harvesting seasons: March/April and September/ October. The yield of honey obtained in Kerugoya averaged 7.5kgs/hive per harvest or 15kgs/hive per year. The main nectar-bearing trees in the area are coffee (*Coffea spp*), bananas (*Musa spp*), *Grevillia robusta*, avocados (*Persea americana*), macadamia (*Macadamia tetraphylla*), mangoes (*Mangifera indica*), croton, pawpaw (*Carica papaya*), and beans (*Phaseolus vulgaris*). Flowering maize is an important source of pollen.

Hive occupation can vary depending on the time of year. Swarming season was observed to be in July. Bee swarms have been noted to migrate from Mwea Division to areas of higher altitude and vice versa, based on the annual period.

Mwea Division

Mr. Kariuki is new to Mwea Division and has bought three acres of land. He has never owned a beehive. At one of the beekeeping field days, Mr. Kariuki observed that most apiaries were sited under trees. He said he had no trees on his new farm but asked which tree species he should plant to act as bee forage. He was given the following advice:

Grevillia robusta: *mukima* (*Gikuyu*), silky oak (English). This is the most abundant tree in the district. It grows very fast and produces a lot of nectar. It flowers from March to April thus ensuring honey flow in May and June.

Photo 26: A tree in flower

It is also a source of fodder and timber as well as coffee shade, mulch and green manure. Wildings and seedlings propagate the tree.

Acacia mellifera: *muthigira* (*Gikuyu*), hook thorn (English). It is a low shrub propagated by direct sowing, seedlings, coppicing, and wilding. It flowers before the rains. The flowers produce quality honey. Other uses are as fodder (pods, twigs, leaves, and flowers), live fence, and fuel wood.

Others of the same genus are *A. lahai* (*Mugaa*), *A. seyal* (*Mugaa*), *A. abyssinica* (*Mugaa*), *A. brevispica* a.k.a. wait-a-bit-thorn (*Mwikunya*), and *A. Gerardo* (*Muthi*).

Eucalyptus saligna: *muringamu*, Sydney blue gum. The tree grows at an altitude of 1200 to 2400m except in arid areas and those infested with termites. Seedlings and direct sowing propagate it. It flowers in June and July. You should not plant the tree near crops but in a woodlot.

Other useful trees and shrubs are:

- *Croton megalocarpus* (*mukinduri*): flowers in March to April, useful for boundary marking, timber, and shade.
- *Azadirachta indica* (*mwarubaini*): flowers in March to April with seeds in June; useful as medicine, insecticide, shade, fertiliser, and de-wormer.
- *Calliandra calothyrsus* (*calliandra*): used for shade, windbreaks, and fodder. The shrub flowers in February to March.
- *Callistemon citrinus* (bottlebrush): very fast growing; both useful and ornamental.
- *Cajanus cajan* (njugu): drought-resistant food crop with other uses as nitrogen fixation agent and fodder (leaves and pods).
- *Kigelia africana* (*muratina* or sausage tree): flowers in January to February. The unripe fruit is poisonous. It does well in arid areas. The Kikuyu do not plant it near homesteads as they used it to mark gravesites. It is a slow-growing tree, a non-prolific seeder with poor germination rate.

Cultivated bee-forage trees include:

- Papaw (*Carica papaya*)
- Bananas (*Musa sp*)
- Beans (*Phaseolus sp*)
- *Muembe* (*Mangifera indica*)
- *Mubera* (*Psidium guajava*)
- *Mukandamia* (*Macadamia tetraphylla*)

EXAMPLE TWO: KAKAMEGA

The following is a brief description of beekeeping practices, potential yields, flowering trees, and seasonality of honey production around Kakamega.

There are three different types of hives used in the area, namely:

- Traditional log hives
- Kenya top bar hives (KTBH)
- Basket hives

Log hives are made from trees species such as *munga*, *likoro*, *munuku*, and *mukomari* (Luhya terms).

KTBH are reputed to be very productive if managed well, but they may get too hot in the dry season and too cold in the wet season, so you need to insulate them. Using grass for heat retention in the wet season works well.

Basket hives are least favoured by beekeepers.

Factors	**Basket**	**KTBH**	**Log**
Quantity Harvested	3	2	1
Ease of Harvest	3	1	2
Heat Properties	2	3	1
Least Expensive	1	3	2
Rain Resistance	3	1	2
Durability	3	2	1
Total Score	**15**	**12**	**9**

Table 2: Beekeepers' Preferences for Different Hive Types
(The ranking was as follows: 1 = best, 2 = second best, 3 = last.)

(N. B. *Table 2 shows a good lesson for those of us who promote improved hive designs. The log hive comes out the best with the KTBH second and the basket hive last. If the KTBH is not managed to maximise its advantages [e.g., moveable combs], it can be outperformed by the log hive. Advantages of the KTBH are many, e.g., ability to manipulate bees for increased honey production and ease of harvesting that allows collection of comb honey, which obtains a significantly higher price in the market. But if you don't manage a KTBH or Langstroth and it produces crude honey, it is actually worse than a log hive because of its higher purchase cost.*)

The swarming season in Kakamega is reported to be in November.

Hives are normally harvested only once a year. Bees produce honey from November to April, but production is highest between December and March. (See Figure 10.) You can get 30 to 40kg/hive of crude honey harvesting in February/March. Harvesting can take place more than once during this period. As the weather becomes cooler when rains start in March/April, bees become net consumers of honey. If you don't harvest before the rains, yields will go down.

Honey production is reliant on the flowering of forest trees and other plants, including *Leucaena leucocephala*, bananas, *isungusa* (Luhya), *isirimoi* (Luhya), and *iludolio* (Luhya). The flowering of *Croton megalocarpus* (*musine*, Luhya) is an indicator of when to harvest.

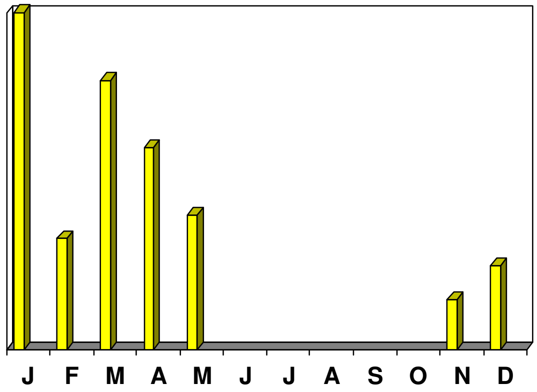

Figure 10: Seasonality of Honey Production in Kakamega

(Reference for Kakamega section: *Beekeeping in Kakamega Core Target Area*, ABLH Report No.15 of 1996.)

EXAMPLE THREE: MOLO

Molo is an example of beekeeping in a highland area where potential is lower than other areas due to high altitude. With good management of bees combined with good weather conditions, however, yields of up to 20kg/hive are achievable—very good yield for this area. In poor seasons and under poor management, 3kg/hive yields have been obtained.

Beekeeping in the area has lagged behind for many years. The following are major problems:

- There are few flowering plants and shrubs to produce nectar.
- Vast acreages have been turned into large-scale farming, thus reducing natural flowering vegetation.
- There is lack of knowledge on the management of beehives.
- Cold climate in the area makes bees abscond the hives.
- Farmers are reluctant to get into beekeeping probably because of the above factors.

Plant Name	**Flowering Time**
Dombeya goetzentii	September/October
Bottle brush (*Callistemon citrinus*)	June/July/August
Sagebrush	August/September/October
Dahlia	February to April
Maize	August to September (pollen only)
Eucalyptus saligna	February to April
Vernonia	July/August/September
Pawpaw	August/September
Wild rape	July/August
Croton megalocarpus	March/April
Fuchsia (ladies' eardrop)	year-round except April/May and June
Daisy	February/March and August/September

Table 3: Plants Suitable for Bees in Molo

Please note that the above table is only a guide. Depending on the season, plants may flower late or not at all. The honey harvest season in Molo is towards end of September/October from the *dombeya* tree, which flowers around this time. Beekeeping at Baraka Agricultural College in Molo has improved since I came in 1994. We planted extensively a fuchsia hedge known as ladies' eardrop around the college grounds. Over the years, the problem of absconding bees has dropped as bees now have forage all year round. Honey yields from Baraka hives have also increased. This proves that planting bee forage has a very positive impact on beekeeping. In particular, if communities can come together and plant collectively, the impact can be much greater.

EXAMPLE 4: TRANSMARA

This is a good beekeeping area. There are many trees and shrubs, which have not been cultivated, that grow in the district. Natural vegetation remains relatively undisturbed. Farmers are organised into groups and they own most of the hives, not individuals. An umbrella body called Transmara Association of Beekeepers coordinates them.

Beekeepers usually harvest three times a year: March, July, and November. From each harvest, about 10kgs/hive of honey is expected. The main honey flow is in November. Apart from these times, you may also harvest honey in other months.

Plant Name	**Flowering Time**
Oskett	May to September
Ogilai	May to September
Olmolongina	May to September
Oltorioi	June to August
Orkikeorgos	June to August
Olkinyei	November to February
Okilenyai	November to February
Olkel	November to February
Olchartuyian	June to August/November to February
Olmesuli	June to August/November to February
Olea africana (wild olive)	June to August/November to February
Thurnbegia atela	June to August/November to February
Scutia myrtina	June to August/November to February
Cardia moncica	June to August/November to February
Acacia seiberiana	June to August/November to February

Table 4: Plants Suitable for Bees in Transmara
Maasai names may not be spelled correctly.

Thanks to Baraka Agricultural College for the Molo and Transmara sections.

EXAMPLE 5: NANDI HILLS

Some of the major flowering/honey-producing trees and plants in Nandi Hills are:

- Avocado: March/April
- *Dombeya*: April, September to December
- *Grevillia robusta*
- Banana: all year round
- Pawpaw: all year round
- Beans: April/May and September/October
- Weeds: always
- Coffee
- Sydney blue gum
- *Tebesuet* (Nandi)
- *Croton macrostachyus*: November to February

Honey Harvesting

April to June is the best season. It is also possible to harvest honey from October to February.

A Sample Harvest Season
(*J. Muemah,* beekeeper)

"In April, I harvested 33kgs of crude honey from one hive. After two months, in June, I harvested a further 15kgs of crude honey from the same hive, which gave a total of 48kgs of crude honey. After refining the honeycombs, I got 24kgs refined honey and 1.5kgs of beeswax. From the 24kgs of honey, I filled 48 500g jars for sale. I sold each of the 48 jars at KSH120, which gave me revenue of KSH5,760 (US$80) from just one hive. This yield per hive is, however, not always obtained.

"Another time, I harvested 16 colonies and only got 67kgs of crude honey. Eight of the colonies were strong and eight were weak. This is why it is very important to manage bees properly. With proper knowledge and management of all colonies, it is possible to have them give maximum yield".

(N. B. *The above example is an interesting one. Eight colonies in the apiary were strong and eight were weak. This situation is common. In many apiaries, only a few colonies out of many give significant yields. Beekeepers should aim to ensure that all colonies are strong and producing honey if the enterprise is to maximise profits. Get rid of weak and unproductive colonies and divide productive colonies to replace them. Combine this selection of the best bees with good bee management.*)

For more information on beekeeping in Nakuru District, see www.apiconsult.com, our website where you can find regular updates on what is happening in my apiary through past editions of the Apiconsult e-newsletter. Click on the news and articles section.

BEEKEEPING MANAGEMENT DURING THE YEAR

From the above examples, we realise that conditions for bees vary widely throughout the country. You can also see that management of bees depends on where you find them. It is very important that the beekeeper gets to know the area in which he/she lives. What plants and trees do bees use? When do they flower? When are the swarming seasons? Which trees/plants give the best honey and which give bitter honey? When are the correct times of year to expect honey? Which are signs of harvesting? What factors such as rainfall and temperature affect plant flowering and nectar secretion?

Collect and write down information. It is very important in helping you become a good beekeeper. Be observant always of your bees and the environment in which you live. Such knowledge will allow you to make timely and sound decisions on bee management. In essence, all beekeepers in Kenya must become their own researchers! Research (systematic investigation to establish facts) should not only be seen as the preserve of academics but also that of ordinary people, including beekeepers. See Annex 5 for more information on Participatory Action Research.

Most beekeepers in Kenya do not manage their bees at all. Every time a beekeeper is seen to go near a hive with a bee suit, he/she is said to be going for harvest. "*Unakwenda kuvuna?*" (Are you going to harvest?)

In this book are highlighted simple and achievable management practices. If you have invested money in more expensive equipment, e.g., Langstroth, then you also need to manage bees better to get good returns on your investment.

Traditional hives with a fixed comb make management very difficult. A top bar hive makes management far easier as does the frame hive Langstroth. Seek further advice on management of bees and read as much as you can on the subject. Outlined below are some simple practices that you can carry out to help you improve your honey yield.

SWARM PREVENTION AND CONTROL

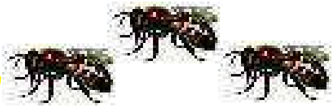

Swarming is what happens when bees divide themselves to reproduce. If swarming occurs when bees are building up in numbers or during nectar flow (when there is an abundance of flowers around), the beekeeper will not get much honey because the bee population in the hive will have reduced dramatically when they are required to collect nectar to make honey. (Look again at Figure 9). The beekeeper should try, therefore, to keep his/her bees from swarming. It is impossible to prevent all swarming, but with proper management of bees, it can be minimised and controlled.

You can take some simple steps to help stop swarming:

If possible examine the bees every 7 to 10 days during the honey period, when flowering season is about to start, until after harvest. Make sure that the queen has enough room to lay eggs by making extra space around the brood nest. If the queen runs out of cells to lay eggs in, the bees will want to swarm. Make extra space by exchanging empty combs with those filled with honey around the brood nest. This is important as African bees build up quickly and can easily become crowded in the hive and want to swarm. Give the queen plenty of space to lay eggs and space for bees to store honey.

If the bees are building queen cells (long thumb-shaped cells protruding from the edge of combs) to make a new queen then you know they are going to swarm. To prevent swarming you can do one of the following:

- Destroy all queen cells in the colony wanting to swarm and switch hive locations with a weaker colony. Foraging bees of the strong colony will return to the original site of the hive and strengthen the weaker colony. As the weaker colony becomes stronger, the strong colony becomes weaker. This may control the swarming urge of the strong colony.

- Destroy all queen cells and give brood combs without bees to the weak colony. This may get rid of the swarming urge in the strong colony.
- Artificially swarm the bees. (See "Making a Division" below.)

MAKING A DIVISION

Break down all queen cells except one.

Transfer comb with the queen cell, one other comb of brood, and two combs of food (honey and pollen) into a catcher box/new hive. Include bees on all combs. Shake in bees from other combs as well. Remember to put brood combs in the middle and honeycombs on either side to insulate the brood nest.

If possible move the catcher box/new hive to a new site at least two kilometres from the old site. (It is not always practical to move the new colony two kilometres away. The division can still work if you move it a short distance, but you can expect adult foragers to return to the old site and weaken the new colony. Give extra bees to compensate for those that will return to the old site.) These bees will become a new colony. The rest of the bees will continue working and a new queen will hatch out in the new colony.

Most of the adult bees will remain in the old hive and continue making honey; however, avoid making divisions during honey season because it reduces honey production. Make divisions after honey flow to increase colony numbers. You can make a division without queen cells as long as the new colony has eggs in the combs transferred. They will be able then to make a queen cell and raise a new queen. Bees can make a queen from worker eggs or worker larvae less than three days old (i.e., very small larvae).

FEEDING

Feeding bees is not recommended in Kenya unless you are sure as to why you are feeding them and the benefits you are likely to get. You waste money to buy sugar for feeding bees when you need it for yourself and your family. People often feed when it is unnecessary and they get no return for their money. Bees can also abscond and your money is wasted.

If want to feed bees, however, the best time to do so is:

- Before nectar flow (stimulatory feeding) when the queen bee will lay eggs and the colony will build up in numbers before honey flow. Start six to eight weeks before the date of flowering because it takes about 40 days from egg to adult foraging bee.
- In times of food shortage, e.g., drought or an excessively wet and cold period.

Feeding at such times may prevent bee migration and keep bees alive.

To feed bees, dissolve white sugar in water by heating gently in a *sufuria* or pot, stirring all the time.

To stimulate bees: use 1:1 solution by volume of sugar and water.
To feed bees to store: use 2:1 solution by volume of sugar and water.

Place the sugar syrup in a feeder box in the hive when it cools. Place some twigs/ grass on the surface of the floating syrup to prevent bees from drowning.

Warning: Never spill sugar syrup or honey around the apiary as it can start robbing: bees attacking each other to raid stores of honey. Weak colonies are prone to robbing by strong ones.

PESTS AND DISEASES AND THEIR CONTROL

For the most part, beekeepers in Kenya need only worry about the following major pests:

- Wax moths
- Safari and other ants
- Honey badgers
- Human beings

(N. B. *Other pests such as birds and wasps can be problematic in some areas of Kenya.*)

Wax Moth

- Destroys wax comb and damages hive parts.
- Control by having a strong colony.
- Remove old combs that bees are unable to cover in time of food scarcity when colony size shrinks.
- Avoid holes and cracks in top bars and hive body where wax moths can lay eggs.

The wax moth is a very difficult and destructive pest to control particularly with frame hives when supers have to be stored after honey extraction.

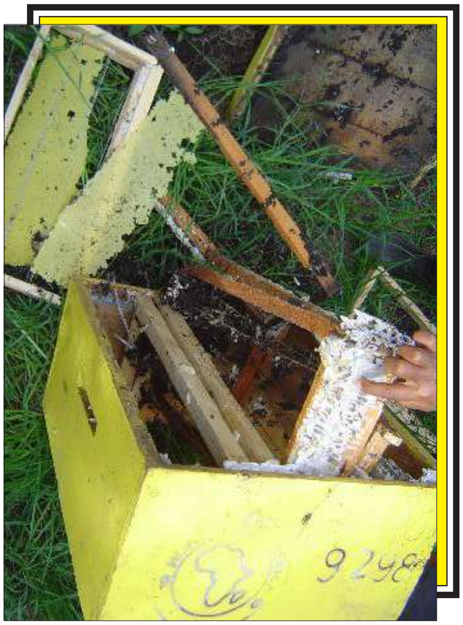

Photo 27: Hive infested with wax moths
Wax moths infested this hive after bees absconded during the dry season.

Safari and Other Ants

- Hang hives and grease hanging wires regularly.
- Spread ashes around posts holding hives.
- Keep grass short and branches from touching hives.
- Make hive parts fit together without gaps.
- If hives are on stands, place legs of stands in tins of old engine oil.

Termites

- Treat posts used for hanging.

Hive Beetles

- Use holes instead of slits for the entrance.
- Keep strong colonies.
- Hand pick or destroy them if found in hive. African hive beetles, particularly the small hive beetle (*Aethina tumida*), are not serious pests in Kenya but are now causing problems for beekeepers in the USA, Europe, and other parts of the world where they have been introduced accidentally.

Birds (Honey Guides)

- Don't leave brood combs exposed. Scare birds away.

Honey Badger

- Hang hives securely to prevent the badger from knocking them down. Hives placed on stands are prone to attack by the honey badger.
- In Transmara, log hives are suspended two metres from the ground by means of a twisted bark rope to prevent damage from the honey badger.
- In Kitui, beekeepers put *mabati* around tree trunks bearing hives to prevent the badger from climbing them.

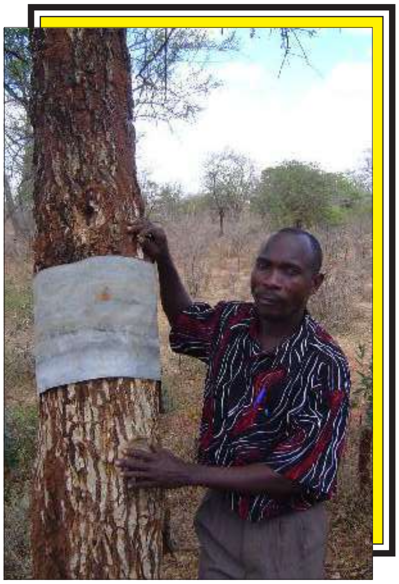

Photo 28: Tree encircled by *mabati*

The *mabati* is to prevent honey badgers from climbing and attacking bee colonies. When the badger encounters the tin, he cannot grip and falls down!

Pirate Wasp

- Attacks and eats bees.
- Place a dish of water below hive entrance, which has a mirror in the bottom. Wasps fly in and drown.

Bee Louse

- Negligible damage.

Man

In parts of Kenya, people cause a lot of damage to hives. They often steal hives and honey, making it difficult to keep hives in areas where security is poor. Children often antagonise bees by throwing stones at beehives, which often results in bees stinging people and livestock. Always keep your hives where you can supervise them. The problem is worse with modern hives kept close to the ground.

Note on Varroa Mites (*Varroa destructor*)

Varroa is a parasitic mite, which causes bee colonies to die out. This pest causes havoc to bee populations in many parts of the world, including Africa. Varroa mites are now in many southern African countries and may spread to Kenya in the future. Kenya currently is varroa-free and we should try to keep it that way. Bee importations into Kenya are prohibited and should never happen to prevent the accidental introduction of this destructive mite or other bee pests and diseases into our country.

MAJOR NECTAR-BEARING PLANTS IN KENYA

I have written about many trees, shrubs, and crops useful for bees in Kenya elsewhere in this guide, but below is a consolidation of the major ones. Planting multipurpose bee forage trees is a good idea for farmers as you can benefit from bee forage plants in other ways such as fencing posts and firewood. Suitability of tree species for planting varies widely in Kenya due to variation in climate. To find out which are the best species to plant, seek advice from your local Beekeeping Officer or Forest Officer.

Major nectar-bearing plants in Kenya: acacia spp, banana, citrus, dombeya, eucalyptus (gum), croton spp, jacaranda, kei-apple (*Albera caffra*), sunflower, vetch, rape, and coffee.

It helps if you have what people call a flowering calendar for the area. Make one by noting the major nectar-bearing trees in your area and the times they flower. Record this information carefully. It is then easy to predict the flowering of trees and the dates. (They may vary from year to year.)

Photo 29: Citrus fruit in bloom

Photo 30: Banana in flower
It is a good source of nectar in banana-growing areas.

Photo 31: Acacia *senegal* in bloom
Acacia species are very important nectar sources in arid areas of the country.

Photo 32: Acacia *persiciflora* in flower
This is in Nakuru.

Photo 33: Kei-apple
It makes a great fence and is a good forage plant for bees.

Photo 34: *Cordia africana.*
It is very attractive to bees.

Part Four

THE HONEY HARVEST

TIME OF HARVESTING

Harvest honey during dry spells, i.e., January, February, March, July, August, September, November, and December. The harvest time in each area differs, so check the right time in your area. In areas where there are dominant bee plants like coffee, sunflower, etc., you should harvest after the flowers wither.

Regular inspection of hives during nectar flow will ensure that the beekeeper harvests as soon as honey is ready. Do not harvest unripe honey. (See below.)

Ideal harvesting time of the day is from 5.30pm to 7.30pm.

HARVESTING PROCEDURE

- Ensure you are sting-proof by putting on protective clothing.
- You will need a smoker in good working condition and a plastic bucket with lid—must be clean and dry—for storing honey.
- Smoke the entrance of the hive with about eight to 10 puffs and then gently lift the lid and smoke again. Leave hive for a minute or two before opening lid to allow smoke to affect bees. Smoke causes bees to engorge themselves with honey making it difficult for them to bend and sting. (They become too full!)
- Very gently tap top bars with a hive tool. A hollow sound will indicate where there is no comb.
- Remove top bar from the hive that has no comb attached, so you can examine the rest of bars in the hive. Honeycombs are usually at the end of the hive opposite the entrance. Select combs that are ¾ or more sealed or capped full of honey. (These combs are said to be ripe or have a low moisture content, <19%, which ensures that honey will not ferment later when bottled. Leave combs with brood and pollen for future production of honey.)

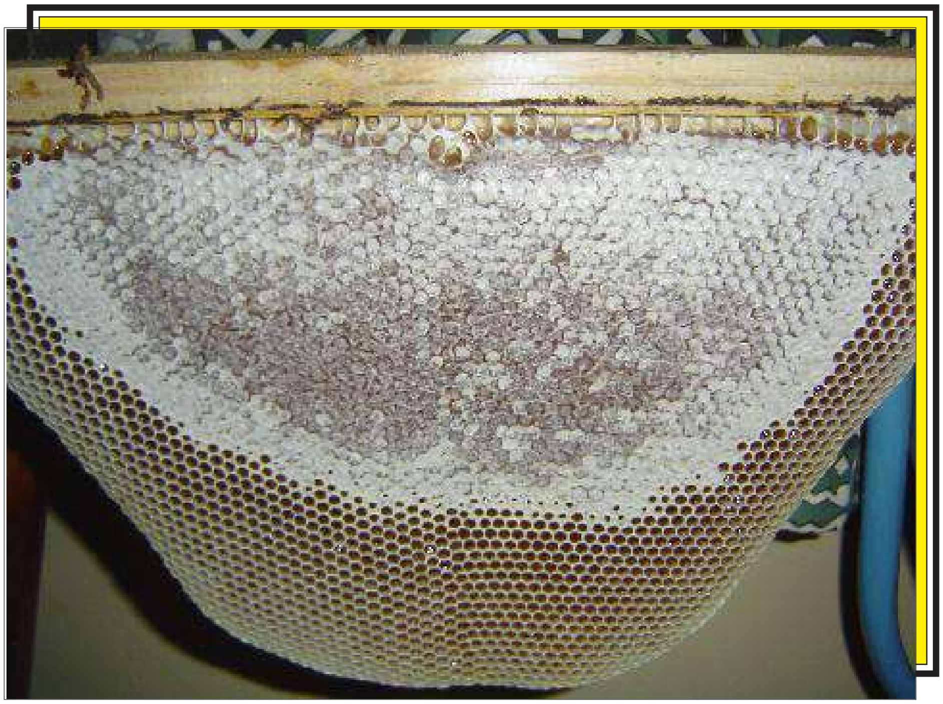

Photo 35: Harvesting white combs
Harvest white combs that are at least ¾ capped, as shown above, for the best honey.

When harvesting a comb:

- Brush bees gently from the comb using a bee brush. You can cut the harvested comb from the top bar to fall into the bucket.
- Replace lid of bucket to prevent bees from entering with the honey.
- Return top bar, minus comb, to the hive.
- As an alternative, place the whole comb and top bar (after brushing bees off) in another empty hive or catcher box where you can take it away later for comb honey. Fix spare top bar in place of the one removed.

Gentle smoking is a continuous process during harvest time to control bees. It is important to avoid smoking the honey directly or excessively because it can damage honey flavour. After harvesting, replace the first bar and cover hive with the lid. Make a final smoke before you leave to keep bees away from the harvester and to prevent them from following him/her all the way home. Remember to move through a bushy area first to get rid of bees.

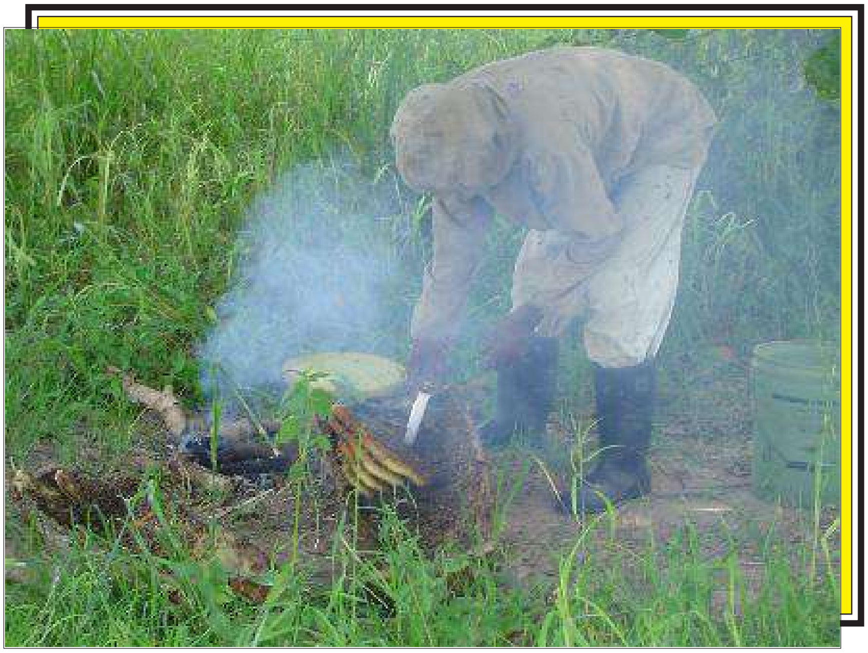

Photo 36: Harvesting traditional log hive in Tabora, Tanzania
The hive is lowered to the ground for harvest and the two halves of the log separated.

(N. B. *Two people are better than one when harvesting or carrying out any bee operations. Two smokers in operation are also better than one to ensure a continuous supply of smoke for subduing bees. Sufficient smoke is very important when working aggressive bees.*)

Photo 37: Harvested honey
Honey harvested from the log hive is placed in clean bucket to ensure its high quality.

HONEY REFINING

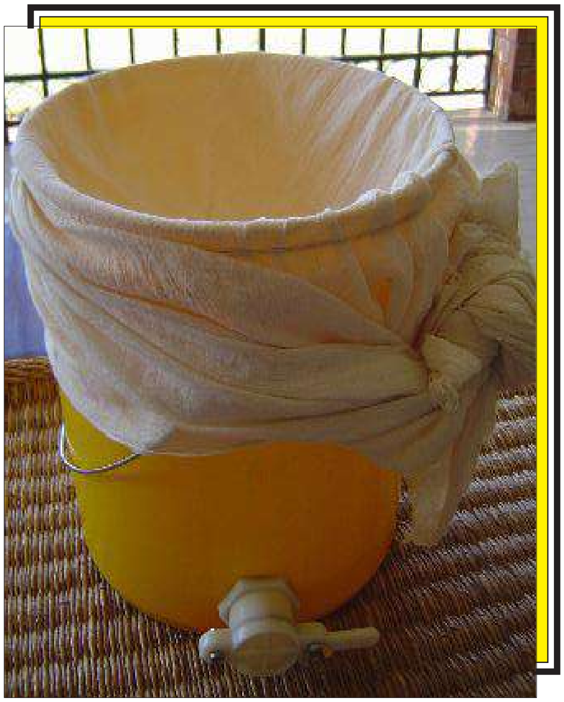

Photo 38: Simple honey refining method for small-scale beekeeper

STEPS

1. Remove wax capping from combs using a knife to cut off the capping.
2. Break combs into smaller particles and sieve them through a net or nylon fabric into a plastic container (See Photo 36.) The sieving process can take a few days. Cover sieved honey with a lid and keep in a dry room away from bees.
3. You can place the bucket in sunshine for two or three hours to heat honey gently so it flows freely.

A word of warning: **NEVER** boil honey as this destroys its flavour and medicinal characteristics!

Once honey has drained through the cloth and settled at the bottom of the bucket (usually after 2 days or so), use a jug to pour it into honey jars for sale. For wax remaining behind on the straining cloth, squeeze out any remaining honey and process leftover wax. (See below.)

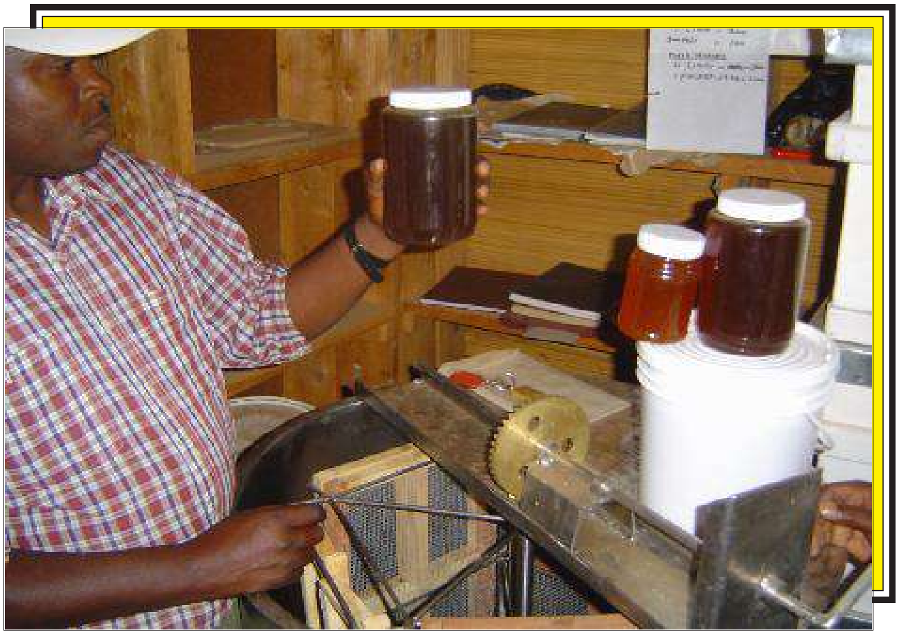

Photo 39: Centrifuge
Centrifuge that handles honey from the KTBH (in wire baskets) as well as frames from the Langstroth hive.

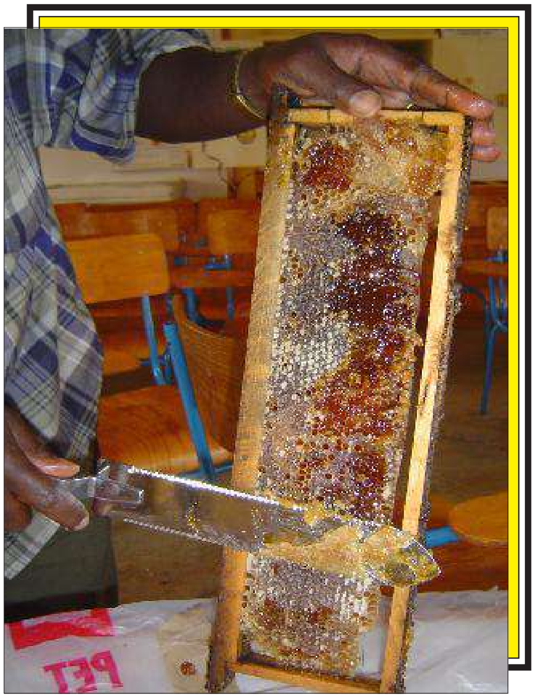

Photo 40: Uncapping frame full of honey
It is done with an uncapping knife in preparation for extraction with a centrifugal extractor.

EXTRACTION OF BEESWAX

Beeswax is a product of the beehive. (For every 100kgs of honeycomb, you can get about eight to 10kgs of beeswax.) Honeybees secrete beeswax in the form of thin scales. 12- to 17-day-old worker bees produce them through glands on their ventral (stomach) surface. Honeybees use beeswax to build honeycomb cells in which they raise their young and store honey and pollen.

To produce wax, bees must consume about eight times as much honey by mass. Estimates are that bees fly 150,000 miles to yield one pound of beeswax or 530,000 km/kg.

PROCEDURE TO EXTRACT WAX FROM COMBS

1. Mix combs and water in a *sufuria* (aluminium pot) and heat. Wax melts at about 62 to 64° Celsius, so there is no need to boil. Boiling damages the wax and can be dangerous. Overheated wax can burst into flames. Do not use iron, brass, zinc, or copper containers for heating wax as it can discolour the finished product.
2. Pour melted combs and water into an extraction bag. You can use cotton for sieving. (You can also use the small bags maize seeds come in after you clean thoroughly.)
3. Smear sides of a second *sufuria* with soapy water to prevent wax from sticking to its sides.
4. Filter wax into the second *sufuria*. Use two sticks (such as two top bars) to squeeze the bag containing melted combs to extract wax. The yellow wax will come out along with water; waste will remain in the filter bag. If the combs contained bee brood, you can feed these to poultry as they would be cooked by then.
5. After filtering, wax separates from water and floats to the top.
6. Remove wax after leaving it to cool in the *sufuria*, with lid on to keep away dust, for 12 hours.
7. Scrape dirt from the bottom of wax cake when cooled.
8. Store wax blocks in a cool dry place. Never store near pesticides/chemicals as it may absorb them.

Your wax block is now ready for sale or for further use. Wax currently sells in Kenya at a price of about KSH100 to 150 per kilogramme or more depending on the demand.

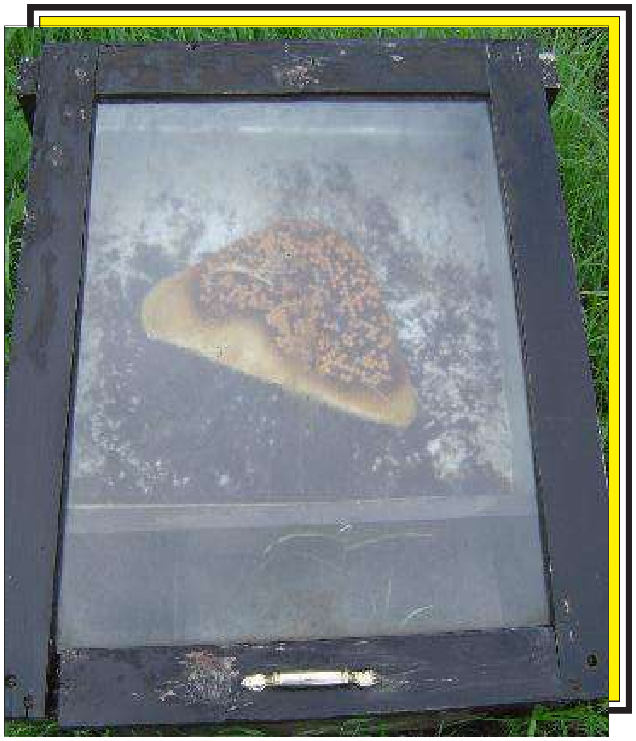

Photo 41: Solar wax melter used to extract beeswax
It is most efficient for use with new combs/cappings instead of brood comb.

Part Five

THE MARKET FOR BEE PRODUCTS

OUT OF THIS WORLD RECIPES MADE WITH BARAKA WILD EQUATORIAL FOREST HONEY
EAT HONEY!

Photo 42: A sampling of honey products.
Most people enjoy honey!

HONEY

There is currently a very high demand for honey both in Kenya and overseas. Kenya is now licensed to export honey to the European Union. Market opportunity exists for beekeepers but lack of production is at present the main constraint in Kenyan beekeeping. Good quality honey can fetch a high price. Any farmer who has the good fortune and skill to obtain honey should have no trouble selling it. In general, if you present honey to the consumer in a clean, unspoiled condition, the price will be higher. Beekeepers should aim for the highest grade of honey to maximise returns from beekeeping. Honey in Kenya is sold in the following grades:

CRUDE HONEY

This is a mixture of ripe and unripe honey. At harvesting time, the wax, honeycomb, and bee-and-brood comb are all mixed into one container. This container is often an old tin. Crude honey is used mainly for brewing local beer because quality requirements are not very strict. The demand for this type of honey is high.

SEMI-REFINED HONEY

Semi-refined honey is generally the liquid honey that remains when you skim wax off the top of crude honey. Honey sinks to the bottom as it is heavier. Semi-refined honey still contains particles of wax and other debris such as bees' legs. It can be stored for the beekeeper's own use or it can be refined further and packed for sale. It gains a higher price than crude honey.

REFINED HONEY

Refined honey is clean. You strain it to remove all particles of beeswax and other materials. Remember: refined honey is unchanged, it is only strained. Nothing else is added so it is still the pure honey that bees made in the hive.

CHUNK HONEY

Whole combs of capped honey can be harvested carefully from the beehive. You can cut up pieces of the comb and put them into jars of liquid honey. This gives the consumer a feeling that the honey is real and not adulterated with sugar. Chunk honey can fetch a higher price than refined honey.

COMB HONEY

Honeycombs of capped honey that have a nice white capping can be cut up, placed on small trays, and covered with cling film. These are very marketable in Kenya and command a very high price in the market, particularly in affluent Nairobi suburbs and other towns. This product should be the ultimate aim of all beekeepers with access to these markets. This product is priced per gram.

PACKAGING AND LABELLING

Package honey in either plastic or glass jars, which should be clean and dry. 454g jars are available in Kenya from Nairobi suppliers as are plastic trays for selling comb honey. (See Annex 6 for contacts of suppliers.)

Labelling of honey must include:

- net weight of honey,
- name and address of producer,
- country of origin (Kenya), and
- description of contents, e.g., "pure honey"

Initially you can make a simple label by hand or on a computer, which you can then photocopy. You can have a more attractive label printed later, but this is costly and requires many jars to spread the cost.

Some people use Kasuku/Kimbo plastic containers or tree top bottles (700ml) to sell honey. Metal paint tins (*mikebe ya rangi*) are not good. Honey is acidic and can eat the metal and spoil the honey.

WHERE TO SELL HONEY

You can sell your bottled and labelled honey directly to shops. Build up a name for supplying the right quality, in the right quantity, at the right price, on time. Many beekeepers develop a reputation for beekeeping and people flock to their homes to buy honey directly. If you are enterprising, you can even set up a shop of your own selling bee products. You could also buy and sell honey from other beekeepers.

A shop or kiosk located in the right place such as the market or along a busy route can attract many customers.

You can also consider the option of forming a beekeeping association to market products of farmers in your area. Examples are the Kakamega Forest Beekeepers Association and the Transmara Association of Beekeepers. You can also consider the option of starting a cooperative. An example of a successful beekeeping cooperative is Ruai Beekeeping Cooperative Society in Naru Moru, which markets eight tonnes of honey per annum collectively. (See Annex 6.)

The advantage of farmers marketing together is fixed costs, e.g., transport, can be spread over a larger quantity of honey reducing cost. Collective selling gives farmers access to higher priced markets such as Nairobi.

Whatever type of honey you produce, you can easily find a market for it in Kenya and the price you receive will compare favourably with that available elsewhere in the world.

BEESWAX

Most people in Kenya throw away wax combs upon harvesting or after honey extraction. Beekeepers do not know its value. Local villages use beeswax in very limited ways, e.g., for shoe repairs by cobblers. Some companies such as Bata (shoe company) and Kiwi (shoe polish company) purchase beeswax, which they often obtain from *miti ni dawa* (honey beer) brewers. You leave the wax after brewing beer from crude honey, which contains honeycombs.

Baraka Agricultural College buys clean beeswax cakes from other buyers. (See Annex 6 for college contacts.) You can also use your beeswax to make some of the recipes in Annex 4.

PROPOLIS

Export markets for propolis exist. People use it as medicine, selling it as capsule, ointment, or tincture (dissolved). You can chew propolis raw as medicine for the throat. It is on sale in this form in very limited quantities in Nairobi. When harvesting, simply scrape off propolis and store in an airtight container. You can also try making medicine from propolis, e.g., propolis ointment. Try the recipe in Annex 4.

Part Six

STARTING A BEEKEEPING ENTERPRISE: COSTS & PROFITS

I have based these figures on 20 KTBH hives per farmer. On the other hand, I have calculated equipment costs from the Baraka Agricultural College Workshop bee equipment price list. All figures are valid for 2007.

The farmer can add significant value if he/she sells honey in jars or bottles. This is realistic in many parts of Kenya where there is a very strong local demand for honey with high prices. In other parts of Kenya where local production is high, bulk sale of semi-refined honey at a lower price to middlemen is the norm. In remote areas of the country, honey prices tend to be low but hive numbers are much higher, particularly in semi-arid areas where individuals can own hundreds of traditional fixed comb hives. An average production figure of 20kgs per occupied hive is achievable for most farmers; however, this can increase to 30kgs or more depending on the area and the management of bees, especially if you control swarming and maximise honey production. An occupation rate of 80% means 16 hives out of 20 are in production, which is a good occupation rate by Kenyan standards. That said, there is no reason why the farmer cannot achieve 100% occupation rates with some effort.

In this example, we take an average farmer with 20 KTBH hives, with an 80% occupation rate, producing an average of 20kgs per occupied hive. The farmer harvests and sells honey in bulk at a price of KSH100/kg with no value addition. This example is not the best or worst but an average situation.

Expenditure	Cost (KSH)
20 Kenya top bar hives @ 2,000/hive	40,000
Bee suit @ 3,270	3,270
Smoker @ 800	800
25 Hanging posts @ 150/post	3,750
Grease and miscellaneous items	500
Total	**48,320**

Table 5: Costs of Starting a Beekeeping Business

Estimated honey production from one colony = 20kgs
Total honey production from 16 occupied hives = 320kgs
Estimated price of 1kg honey in 2007 = KSH100/kg
Estimated income from 16 occupied hives = KSH32,000
Net income year one = KSH32,000 – 48,320 = (16,320)

A negative of KSH16,320.

Equipment will have a lifespan of up to ten years or more depending on how one cares for them, so in subsequent years, capital costs will be paid off and net income will be much higher. The question is how do poor farmers finance capital costs to get started? The important thing in beekeeping is that a farmer can start by using simple hives made from local materials. It is not necessary or advisable for you to take a loan to get started unless you are absolutely sure you can make the hives produce. This certainty comes from experience. Start slowly. Learn as you go along. Once you understand what you are doing, expand. Many farmers in Kenya who purchased expensive frame hives on loan had ended up repaying through other means as the hives failed to produce anything—very demoralising. In particular, with the recent promotion of the Langstroth hive, people are misled by the belief that it will automatically produce honey and lots of it. **Be warned: this is not the case!** It will only yield well with good management.

GROSS MARGIN ANALYSIS

The following is an analysis of beekeeping versus a major source of livelihood in Kenya: maize production. The table below looks at the same beekeeping enterprise above minus capital costs. It also looks at the enterprise at different production levels: low, medium, and high. Low is 10kgs per occupied hive/annum. Medium is 20kgs per occupied hive/annum. High is 30kgs per occupied hive/annum. The sale price of honey is estimated at KSH100/kg. Through packaging, one can easily add value and increase the price of honey. Many farmers in Kenya can sell honey in raw form for KSH150/kg or more to their neighbours, e.g., beekeepers in the Lare area of Nakuru District.

Description	**Production Level**		
Production	Low (10kgs)	Medium(20kgs)	High(30kgs)
Yield per hive x 16	160	320	480
Price per kg honey	100	100	100
Gross Output (a x b)	16,000	32,000	48,000
Variable Costs			
Grease and miscellaneous	500	500	500
Total Variable Costs	**500**	**500**	**500**
Gross margin/16 occupied hives	15,500	31,500	47,500

Table 6: Gross Margin Analysis of Beekeeping

Description	Production Level		
Production	Low	Medium	High
Yield of maize (90kg-bag)	20	50	87
Price of maize	1,300	1,300	1,300
Gross Output (a x b)	26,000	65,000	113,100
Variable Costs			
Total variable costs	18,000	32,500	60,714
Gross margin/hectare	8,000	32,500	52,386

Table 7: Gross Margin for Hectare of Maize

(Source: Baraka Agricultural College)

From the above comparison, a beekeeping enterprise of 20 KTBH compares very favourably with returns from a hectare of maize, particularly at low levels of efficiency in maize production. In addition, a beekeeper has the option of adding value to his/her honey to increase returns while a maize farmer is more at the mercy of market forces where gluts in production frequently depress prices. The bee enterprise takes up very little space in comparison to maize and the time input is much less. The most important but often overlooked inputs for successful beekeeping are skills and knowledge, so develop them very well.

GLOSSARY OF TERMS

Abdomen - third posterior section of a bee’s body.

Absconding - when a colony of bees abandons the hive or nest site due to infestation of pests or lack of food. (It happens with African bees).

Allergic - in beekeeping, refers to the severe bodily reaction to bee venom following a sting.

Anther - the part of a flower’s stamen that produces pollen. The male part of the flower.

Apiary - bee hives grouped together.

Bee space - gap large enough for bees to walk and work, e.g., the space between two parallel combs or between a comb and the wall of the hive.

Beeswax - material produced by honeybees for building combs.

Brood nest - the centre area of the combs of a colony where brood is reared.

Capped honey - ripe honey in the comb, sealed with a wax capping.

Cappings - thin wax covers of cells full of honey, particularly after they have been removed from the surface of the comb prior to honey extraction.

Caste - different types of mature honeybees (and other social insects).

Comb - beeswax structure of two plates of cells arranged back to back on a midrib and used to rear brood or store honey and pollen.

Drone - male honeybee. Drones undertake no work within the hive. Their sole function is to fertilise the queen.

Feeder box - box used to feed bees artificially with sugar syrup.

Foundation - thin sheet of beeswax usually embossed with a hexagonal pattern of cells placed in each wooden frame that serves as a base upon which honeybees build their comb.

Frame hive - hive containing frames. The honeybees are encouraged to build their combs within these frames. The frames enable combs to be lifted from the hive for examination.

Gland - an organ that produces and releases one or more substances.

Hive - artificial structure for housing a colony of bees. It is also a verb to denote the process of introducing a swarm into a hive to house it permanently.

Langstroth hive - a design of frame hive. The inventor, Rev L. Langstroth, recognised the importance of bee space and this allowed him to design the movable frame hive.

Larva - an insect in the first stage of development after hatching from the egg. (plural: larvae)

Migration - seasonal movements of whole honeybee colonies, leaving no brood behind in the nest.

Nectar - Sweet liquid secreted by flowers. Watery solution of various sugars.

Participatory Action Research - PAR has emerged in recent years as a significant methodology for intervention, development, and change within communities and groups. It is now promoted and implemented by many international development agencies and university programmes as well as countless local community organisations around the world. PAR involves people working together to solve their problems through learning in a systematic way. As the name implies, the output of the research is action.

Pollen - fine dust-like substances, which are the male reproductive cells of flowering plants collected by bees as a source of protein.

Propolis - an aromatic, resinous, antibiotic plant material collected by bees for blocking holes, waterproofing, and generally protecting the hive from intruders and diseases.

Pupa - final stage of development of the bee in which it is sealed into its cell and from which it will emerge it as an adult. (plural: pupae)

Queen - female parent of the colony. The only sexually developed female.

Race - biological definition of race is a categorisation of organisms with differing characteristics while maintaining enough similarities to be a part of a common genus and species. In this context, it is synonymous with subspecies.

Ripe honey - capped honey in the comb, i.e., bees have added enzymes and dehydrated it to an acceptable moisture level below 19% before sealing.

Species - in biology, a species is a group of related organisms that, more or less, share a distinctive form and are capable of interbreeding.

Stamen - male reproductive organ of a flower. It consists of a stalk on the end of which is the anther.

Starter - narrow strip of wax foundation fastened in a frame or under a top bar for the bees to start building comb in the desired position.

Super - any hive box placed above the brood nest or chamber. Usually contains combs in which bees will store honey.

Sustainable livelihood - occupation or employment enabling someone to provide for his/her basic needs and to be secure in its continuation in the future.

Swarming - when a honeybee colony becomes large enough to divide into two.

Top bar hive - intermediate technology hive in which bees are encouraged to build their combs suspended from bars placed across the top of the hive.

Unripe honey - honey in cells not fully processed by hive bees, i.e., when still thin and not capped.

Worker bee - female honeybees that make up the bulk of the colony and undertake all the work of the colony except for mating and egg laying. Workers are sterile females.

ANNEX 1
Current Status of the Kenyan Beekeeping Industry

BRIEF HISTORY OF BEEKEEPING IN KENYA

In the late 1960s, the Kenyan Ministry of Agriculture, the International Bee Research Association (IBRA), and Oxfam collaborated on the development of an appropriate hive for Kenya. The result of this joint effort is the Kenyan Top Bar Hive (KTBH). In the 1970s, the Ministry of Agriculture began introducing the KTBH to Kenyan farmers supported by the Canadian International Development Agency (CIDA). Beekeeping cooperatives were set up in many parts of Kenya to facilitate marketing of bee products. All of these marketing structures collapsed later. Many have made numerous attempts to develop beekeeping in the country since the 1960s, but production of honey is still very low in relation to its potential. In recent years, people have promoted Langstroth frame hives as the answer to Kenya's beekeeping problems but with limited success. Neither the Langstroth frame hive nor KTBH have had much impact on Kenyan beekeeping, as about 80% of all beehives in Kenya are still traditional type hives. In spite of the production potential, Kenya does not export honey. The European Union has licensed Kenya to export to their region.

BACKGROUND TO THE KENYAN BEEKEEPING INDUSTRY

Kenya relies heavily on agriculture like other East African countries. 75% of its people live in rural areas and 60% of these live in absolute poverty. Kenya is a nation of small holders with over five million small-scale farmers and pastoralists. Cutbacks in public services and the free market philosophy of recent years have hit rural communities very hard. As this is unlikely to change, the future of such rural communities will depend on developing their capacities from within to meet the challenge. Beekeeping is an opportunity to harvest and add value to a local resource (floral nectar) to generate wealth and employment and beat poverty.

80% of Kenya's land area is either semi-arid or arid land. Beekeeping is suitable for these areas due to prevalent plant species, e.g., acacia trees, and also due to limited livelihood options available in these areas. About 80% of Kenyan honey is produced in arid and semi-arid areas. (Muya 2004)

Majority of Kenyan beekeepers still use traditional systems of beekeeping: simple fixed combs, mostly hollow log hives, in spite of over 30 years of beekeeping extension carried out by the government and non-governmental organisations (NGOs) to promote improved hives, e.g., the Kenya Top Bar Hive, which is an intermediate technology hive.

In the last five years, there has been a major push by some NGOs and private companies, supported by major donors, to introduce the Langstroth frame hive used in Europe and America. There is conflicting information between the actual impact of these hives and the claim of success by its promoters. Information collected in a number of recent beekeeping studies indicates limited impact on enhanced production of bee products and on improvement of livelihoods. Conflicting information on different production technologies results in confused farmers who receive different messages from different extension agents.

In relation to bee product marketing, research indicates that the Kenyan honey market is underdeveloped due to low volumes. Volumes and quality have not yet been reached for export though Kenya has been licensed to export honey to the European Union since 2003. The shortage of bee product production also means that local demands are not met sufficiently. Significant quantities of imported honey, mostly from Tanzania, abound in the Kenyan market, but some people repackage them as Kenyan honey. Opportunities exist for the development of fair trade and organic honey export markets due to Kenya's largely pollution-free environment and disease-free bees.

Some quarters report that local production of honey has improved in recent years due to reduced official honey imports; however, one can explain it by an increased cross-border trade in honey marketed as Kenyan honey in local markets.

Other issues of importance affecting the Kenyan beekeeping industry include:

- Lack of a clear government beekeeping policy but which is now under development. Different agencies take different approaches to beekeeping with no national coordinating body resulting in a disorganised and fragmented industry.

- Environmental degradation, mostly forest destruction/charcoal burning that results in reduced beekeeping potential in Kenya.

- Little or no beekeeping research in Kenya.

- Fragmentation and disorganisation of beekeepers. Beekeepers are difficult to reach due to poor infrastructure resulting in difficulties in bulking honey for marketing.

OPPORTUNITIES IN KENYAN BEEKEEPING

- There is a strong local demand for honey with current production unable to meet local demand.
- There is a ready export market with a strong international demand for honey and bee products. The European Union allows Kenya to export to their region.
- The bee population is relatively disease-free.
- There are considerable existing skills and knowledge in beekeeping.
- Kenya is a big country with space for a much larger beekeeping industry.

ANNEX 2
More Information on Stingless Bees

Apart from honeybees (*Apis mellifera*), there are other types of bees in Kenya that collect nectar and make honey. These are stingless bees, which usually live in holes in the ground. There are many species of stingless bees (*Meliponinae*) in Kenya and these are yet unexploited on a commercial basis. In Brazil and India, these bees are utilised because of their ability to make good quality medicinal honey. People prize this honey as a cure for baldness and impotence. These bees can yield up to five kgs per season. Stingless bees are the best pollinators as they penetrate deeper into the flower. Baraka Agricultural College in Molo and the National Beekeeping Station at Lenana in Nairobi have started keeping stingless bees. Beekeepers can keep them in small hives. They store honey in small pots that you can harvest daily with a clean syringe!

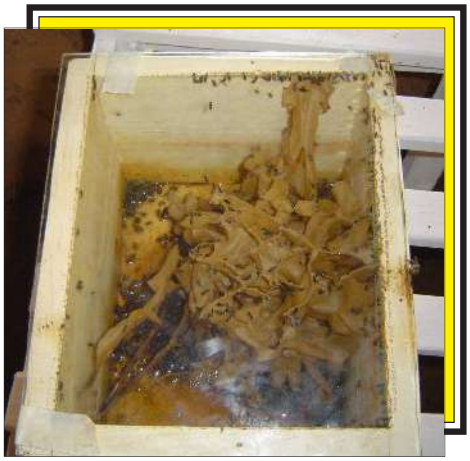

Photo 44: Hive with shaded glass cover.
Bees require darkness.

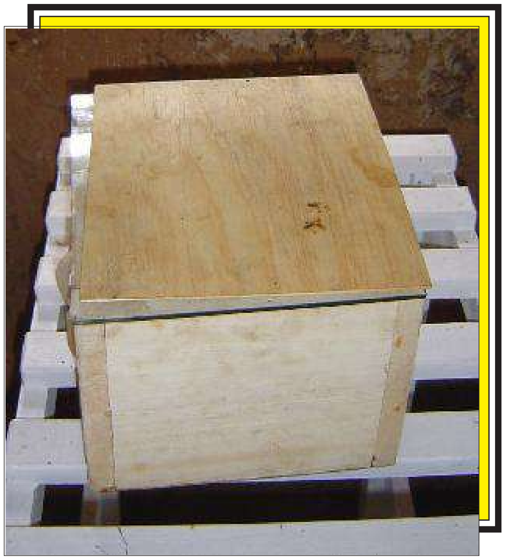

Photo 43: Simple box hive for stingless bees.
(Baraka Agricultural College, Molo)

ANNEX 3

Hive Recordkeeping

Hive No.	Occupation Date	Date of Inspection and Comments	Harvesting Date	Amount Crude Honey (kgs)	Amount Refined Honey (kgs)	Amount Beeswax (kgs)	Remarks

Table 8: Hive Recordkeeping

You can copy the previous table in a hardback notebook, which you should fill out after each hive inspection. You can also photocopy it. In the remarks section, you can specify the inspection date and any information of use such as the temperament of the bees, i.e., if they are docile or difficult to handle. This information is useful when selecting bees in the future. You can record any management activities here such as: (a) amount of food fed to bees (if any), (b) the cost, (c) action taken to control swarming, etc.

Below is a table for marketing records.

Date	**Amount of Honey Sold (kgs)**	**Price per kg of Honey**	**Total Revenue**

Table 9: Marketing Recordkeeping

ANNEX 4

Recipes/Uses for Beeswax, Propolis, and Honey

Photo 45: Kenyan National Beekeeping Station exhibition stand
The stall displayed many excellent bee products.

BEESWAX FURNITURE/WOOD POLISH

Ingredients
200g Beeswax
100g Turpentine
50g Baby oil

Procedure
Grate beeswax into flakes. Gradually add turpentine to soften wax. Add oil and mix. Store in a tin with a tight-fitting top or in a jar.

Tip for sewing shoes, leather, and other thick materials:

Pull thread through small block of beeswax. The wax stiffens and smoothes the thread making it easier to sew.

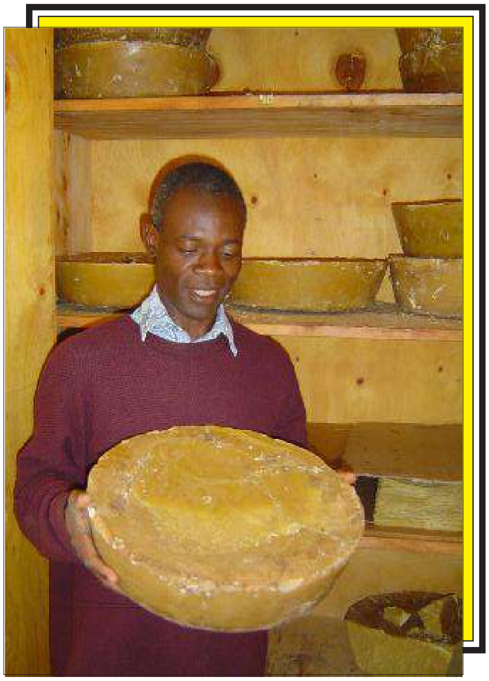

Photo 46: Cake of beeswax for making candles
(Baraka Agricultural College, Molo)

BEESWAX CANDLES

Requirements
Beeswax
Wick (use cotton thread)
Bottle top
Plastic pipe that fits into bottle top
Cooking oil

Procedure
Smear inside of plastic pipe with soapy water or cooking oil. Thread wick through pipe. Punch hole in the middle of the bottle top. Thread wick through hole and knot it.

Melt beeswax indirectly, i.e., use a double pan (one *sufuria* sitting in hot water inside another *sufuria*). Do not boil the wax.

Cover bottom end of pipe to prevent wax flowing away by pushing it down into bottle top. You can also stick end of pipe in the ground. Tie wick to suspension wire. Pour melted wax into mould making sure wick is in the middle of the candle. Wait for wax to cool before filling the rest of the mould with molten wax. Leave candle to solidify. (It should come out easily from mould because you had greased it with cooking oil.)

Candle is now ready for use. Vary wick to get a good candle that burns well. Use many plastic pipes to make many candles at the same time.

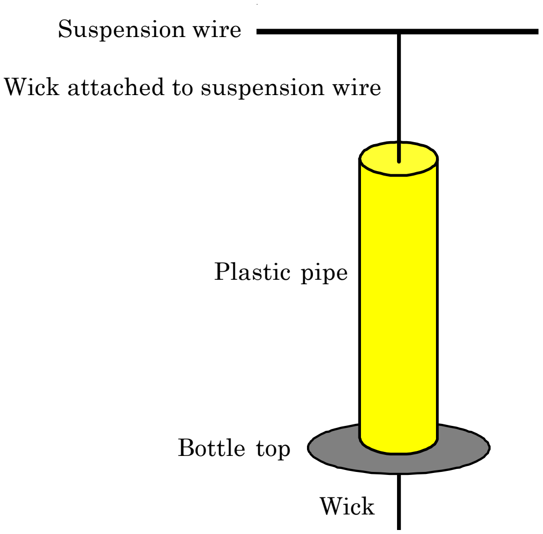
Suspension wire
Wick attached to suspension wire
Plastic pipe
Bottle top
Wick

Figure 11: Candle Making

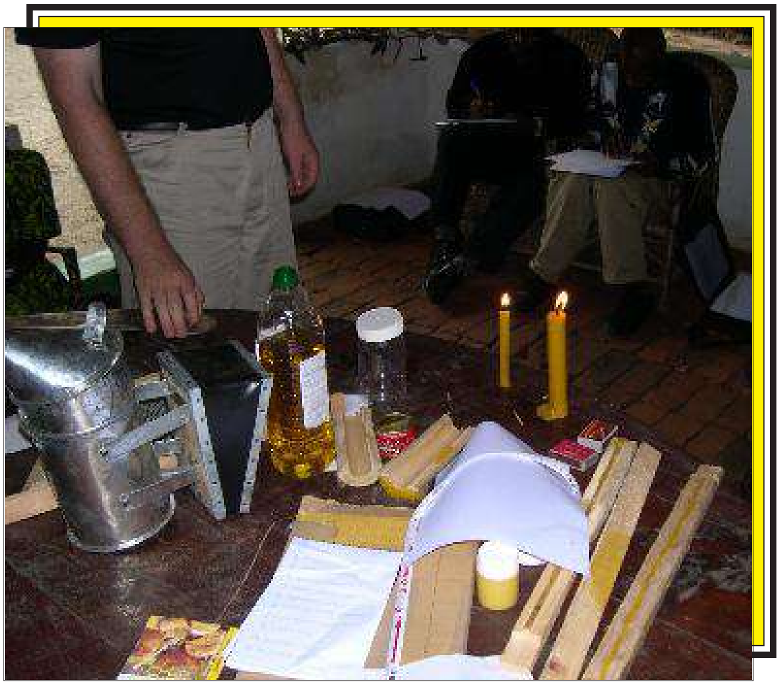

Photo 47: Beautiful beeswax candles made using plastic pipe mould

PROPOLIS HEALING OINTMENT

The first stage is to make propolis granules:

Step 1
1. Scrape off the propolis from hives putting scrapings in a tin.
2. Add enough water to cover scrapings and heat.
3. Wax will melt and float to surface.
4. Propolis will stick to bottom of can.
5. Stir often to help release wax.
6. Remove from heat and let cool.

Step 2
1. Remove waxy layer formed on water surface.
2. Pour off water and save coloured mass beneath.

Step 3
1. Place can in cool place to harden propolis andmake it brittle. Freezer is best.
2. Dry and store in sealed container.

Ingredients
1 tablespoon Beeswax
4 tablespoons Mineral oil (e.g., baby oil)
1 tablespoon Propolis granules
1 tablespoon Honey

Procedure
Melt ingredients in double boiler. Stir until it cools. It should have the consistency of cream. You can pack in small containers and sell. Use cream as a skin ointment for sores, rashes, blisters, etc. Many farmers' groups we work with are now making and selling beeswax and propolis creams and making good money from this activity. For more information, contact Baraka Agricultural College. (See contacts in Annex 5.)

HONEY COUGH SYRUP

Ingredients
6 tablespoons Pure honey
Juice of 2 Lemons

Mix ingredients together well before taking in spoonfuls for coughs.

Warning: Be careful when working with hot wax and other ingredients used to make these recipes. I cannot be held responsible for adverse effects resulting from the use of methods or ingredients mentioned in this publication.

ANNEX 5

Approaches to Beekeeping Extension

In this section, we outline experiences of Baraka Agricultural College that has been involved in beekeeping extension services to farmers since 2002. An Irish development agency, Self-Help Development International (SHDI) funded the extension project. Since 2002, the college has been working with beekeeping groups to develop beekeeping mostly in Rift Valley Province but also in other parts of Kenya/Eastern Africa.

INNOVATIVE ASPECTS OF THE BAC/SHDI PROGRAMME

1. Participatory Action Research (PAR) Approach to Beekeeping

As a country, Kenya covers a very large area presenting a huge diversity of climates and vegetation within its borders. Beekeeping extension workers can only make general recommendations. There is enormous need for beekeepers to adapt and try out ideas for themselves at local level. The basis of the PAR approach is the importance of learning from what we do in a systematic way. Every aspect of life can benefit from this kind of approach. The Baraka/SHDI project encourages farmers to innovate/solve their own problems in beekeeping. This approach comes in light of an inadequate government beekeeping extension service as well as little or no formal research in Kenya beekeeping. In November 2006, farmers researched reasons why they have low hive occupation rates, so they can make decisions on what to do about their problem.

The PAR approach to beekeeping development takes into account and builds on farmers' indigenous technical knowledge.

Photo 48: Bee forage tree seeds
BAC/SHDI farmer researchers collected and documented these seeds.

2. Bee Multiplication

There is considerable shortage of bees and many hives are unoccupied. The project works with farmers to ensure bees occupy hives. This is done through training and support in colony division and swarm trapping. The project has started training farmers in queen rearing in response to one of Kenya's biggest beekeeping problems: low hive occupation rates.

3. Farmer-to-Farmer Trainers

The project works with beekeeping groups in different geographic areas. A cluster of groups has a farmer beekeeping extension worker that the project trained and supported. In turn, he/she trains other farmers in beekeeping. This has proved a very effective beekeeping promotion method and builds local capacity in beekeeping.

4. Artisan Training

The Baraka Agricultural College/SHDI project trains local carpenters and tailors in making beekeeping equipment. This has assisted beekeeping development as beekeepers can access beekeeping equipment at affordable prices. This approach also encourages self-reliance amongst local communities.

(N. B. *For further information on the Baraka Agricultural College/SHDI project, please contact the college. See Annex 6.*)

ANNEX 6
Sources of Further Information

KENYA

You can contact the following organisations as a starting point for further information on beekeeping in Kenya. Many others are involved in beekeeping; the number changes all the time. This list is not exhaustive and is just to establish a place for further networking.

Apiculture Division
Ministry of Livestock and Fisheries Development, Hill Plaza Building
P.O. Box 34188–00100
Nairobi
Telephone: (254) (0)20 2722601/2722637
Fax: (254) (0)20 2728609
Remarks: Contact for information and advice or get in touch with your local District Beekeeping Officer. There's one in every district.

Baraka Agricultural College
P.O. Box 52
Molo
Telephone: (254) (0)51 721091
Email: baraka@sustainableag.org
Website: www.sustainableag.org
Remarks: Contact for bee equipment, advice, training courses, honey, and beeswax marketing

The Kenya Honey Council
P.O. Box 271-00606
Sarit Centre, Nairobi
Telephone: (254) (0)20 4183120
Email: info@kenyahoneycouncil.org
Website: www.kenyahoneycouncil.org
Remarks: An umbrella body representing different stakeholders in the Kenya beekeeping sector.

National Beekeeping Station
Apiculture and Emerging Livestock Division
Ministry of Livestock and Fisheries Development Lenana
P.O. Box 34188-00100
Nairobi
Telephone: (254) (0)20 564302
Remarks: Check out for its library, bee equipment, and advice.

The Kenya Beekeepers Association
c/o The National Beekeeping Station
P.O. Box 34188-00100
Nairobi

Self-Help Development International Kenya
2nd Floor, Catholic Diocese of Nakuru Building
Stadium Road off Kenyatta Avenue
Telephone: (254) (0)51 2212291
Email: kenya@shdi.org
Remarks: The agency currently promotes beekeeping in the Gilgil area.

Ruai Beekeeping Cooperative Society
P.O. Box 8
Naru Moru
Remarks: This is a great example of collective marketing of bee products in Kenya. It markets up to 8 tonnes of honey per annum.

ICIPE (International Centre for Insect Physiology and Ecology)
P.O. Box 30772-00100
Nairobi
Telephone: (254) (0)20 8632000
Fax: (254) (0)20 8632001/8632002
Remarks: It has a commercial insect section dealing with beekeeping
E-mail: dg@icipe.org

African Union
Maendeleo Ya Wanawake House
Nairobi
Remarks: It has an interesting collection of beekeeping books in its library.

General Plastics Limited
P.O. Box 10032
Nairobi
Enterprise Road off Mombasa Road
Industrial Area near Hillock Inn
Telephone: (254) (0)20 530032/3/4/5
Remarks: It supplies plastic jars and lids for packing honey.

Honey Care Africa Limited
Muringa Avenue, Jamhuri Park
Nairobi
Telephone: (254) (0)20 574448
Remarks: It promotes Langstroth hives. It also buys and markets honey.

African Beekeepers Limited
Industrial Area
P.O. Box 3752-00506
Nairobi
Mobile: (254) (0)722 700226
Remarks: It manufactures equipment and markets honey.

Kerio Valley Development Authority
KVDA Plaza
P.O. Box 2660
Eldoret
Telephone: (254) (0)53 2063361
Email: kvda@kenyaweb.com

Strengthening Informal Sector Training and Enterprise (SITE)
Telephone: (254) (0)20 2718155
Jabavu Road, Nairobi
Remarks: It deals in training and support to beekeeping and bee equipment.

REST OF AFRICA

Bee Natural Products Ltd
PO Box 5318
Pan Africa House
Plot 3 Kimathi Avenue
Kampala, Uganda
Telephone: (256) (0)41 234676/ 253844
Mobile: (256) (0)77 907045
Fax: (256) (0)41 259052
Website: www.beenaturalproducts.com

Tanzanian Ministry of Natural Resources and Tourism
P.O. Box 9372
Dar-es-Salaam, Tanzania
Telephone: (255) (0)22 130091
Email: fordev@africaonline.co.tz
Remarks: It has very good publications on beekeeping in English and Swahili.

The National Honey Show Limited
P.O. Box 71326
Dar-es-Salaam, Tanzania
Telephone: (255) (0)22 2126280
Email: epictanzania@email.com
Remarks: It organised the first Tanzanian honey show in Dodoma in 2006.

Tawiri Njiro Wildlife Research Centre
Ministry of Natural Resources and Tourism, Forestry and Beekeeping Division
c/o P.O. Box 246
Dar-es-Salaam, Tanzania
Email: fordev@africaonline.co.tz
Remarks: Tawiri runs beekeeping courses and carries out beekeeping research.

Alternative Contacts:
P.O. Box 661
Arusha, Tanzania
Telephone: (255) (0)27 7677
Telefax: (255) (0)27 2548240
E-mail: tawiri@habari.co.tz or info@tawiri.org

Plant Protection Research Institute South Africa
Private Bag x134
0001 Pretoria, South Africa
Website: www.arc.agric.za
Remarks: It has a very good book, *Beekeeping in South Africa*.

TUNADO (The Ugandan National Apiculture Development Organisation)
c/o Bee Natural Products Limited (See above.)

INTERNATIONAL

Bees for Development
Troy, Monmouth NP25 4AB
United Kingdom
Telephone: (44) 016007 13648
Fax: (44) 016007 16167
Website: www.beesfordevelopment.com
Remarks: It produces a very nice magazine. People may pay for subscriptions with beeswax. It also sells beekeeping books.

Bee Support
Email: micha_el@dds.nl
Remarks: It undertakes promotion of beekeeping for development.

Hives Save Lives
Saltdean, Brighton BN51 9AB
United Kingdom
Website: www.hivessavelives.com
Remarks: This organization has current projects on Uganda.

Bees Abroad UK
Website: www.beesabroad.org.uk
Remarks: A UK-based charity supporting beekeeping development. Bees Abroad also has projects in Kenya.

International Bee Research Association (IBRA)
18 North Road, Cardiff CF1 3DY
United Kingdom
Website: www.ibra.org.uk
Remarks: It has a good library on beekeeping around the world and publishes the *Journal of Apicultural Research* and *Bee World*.

CTA/Agromisa
P.O. Box 41
Wageningen, The Netherlands
Remarks: It has booklets on agriculture and bees.

Bee Vital

Website: www.BeeVitalPropolis.com
Remarks: A UK-based company that purchases propolis.

ANNEX 7
Baraka Agricultural College Short Courses on Beekeeping

The following one-week short courses in beekeeping are currently on offer from Baraka Agricultural College. It offers courses both in Swahili and English, depending on the needs of the group. For further information on course timings and costs, please get in touch with the Short Courses Coordinator at Baraka Agricultural College. See contact details below.

An Introduction to Beekeeping (Establishing a Beekeeping Enterprise)

This course is about starting a beekeeping enterprise in Kenya. The college aims the course at beekeeping beginners, but experienced beekeepers can also benefit by attending. The course covers an introduction to bee biology and behaviour (African bees, *Apis mellifera scutellata*), beekeeping equipment, apiary establishment, bee and hive management, harvesting, and handling and marketing honey.

Processing of Bee Products

This course covers in detail how to process and add value to bee products, including honey, beeswax, propolis, bee brood, and pollen. Local and international markets/ marketing for these products is also covered. One spends much time in practical lessons, i.e., making a variety of value-added recipes, which one can market and sell in Kenya and overseas. Recipes include a variety of candles and creams, polishes, and medicinal/nutritional products from the beehive.

Making Bee Equipment

Making bee equipment is a specialised course aimed at artisans and those interested in making their own beekeeping equipment to quality standards. Both beekeeping staff and Baraka Agricultural College workshop staff, who manufacture beekeeping equipment, facilitate the course. The course covers making a Kenya Top Bar Hive, Langstroth hive, smoker, bee suits, hive tools, and bee brushes. They also give details on how to make a simple honey extractor. Much of the course held at Baraka Agricultural College workshop is actually making the equipment.

Bee Breeding and Multiplication

This is a new course on offer for 2007. The course is designed to assist beekeepers/ beekeeping extension workers occupy the many empty beehives in Kenya. Research shows hive occupation rates as low as 30% in some areas of the country. The course covers details of bee genetics, selection, and breeding. Simple methods of bee multiplication such as swarm attractants, making divisions, and queen rearing are covered in detail, including the Demaree, Miller, Jenter, and Grafting methods of rearing queens.

Further details of Baraka Agricultural College beekeeping courses are available from the college website: www.sustainableag.org. You can also contact the Short Courses Coordinator at telephone number: (254) (0)51 721091; or send email to shortcourses@sustainableag.org.

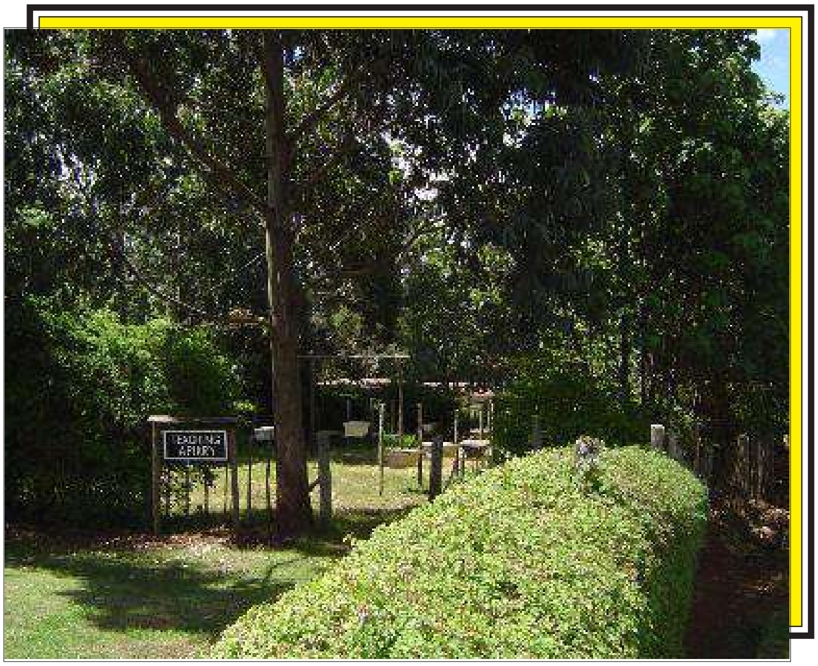

Photo 49: The teaching apiary at Baraka Agricultural College

ANNEX 8

Beekeeping Resources on the Internet

There are a number of websites, which are of interest to beekeeping in Kenya and Africa. They offer a large variety of resources and information. It is well worth checking out internet resources. There is a list of beekeeping resources suitable for Africa on the links page of Apiconsult. See www.apiconsult.com/beekeeping-links.htm. This page is updated regularly.

In addition, the Apiconsult website has a discussion forum on African beekeeping where beekeepers on the African continent share information. Please join the discussion and share your experiences.

As of November 2006, I consider the following websites useful:

www.apiconsult.com

www.apiservices.com (loads of information on world beekeeping)

www.beesfordevelopment.org (UK-based site promoting beekeeping in development)

www.gsu.edu/~biojdsx/main.htm (beekeeping in top bar hives)

www.honeybadger.co.za (South African site on beekeeping)

www.honeycareafrica.com (beekeeping with Langstroth hives in Kenya)

www.topbarhive.com (information on top bar hives in the USA)

en.wikipedia.org/wiki/Top-bar_hive (online encyclopaedia with information on top bar hives and general beekeeping to which you can edit and add your own ideas and information to the articles presented)

There are many more.

A good way to generate more beekeeping links is to search with a search engine such as Google (www.google.co.ke) on “beekeeping Africa” or “top bar hives”, etc.

Happy surfing!

Encyclopedia of INSECTS

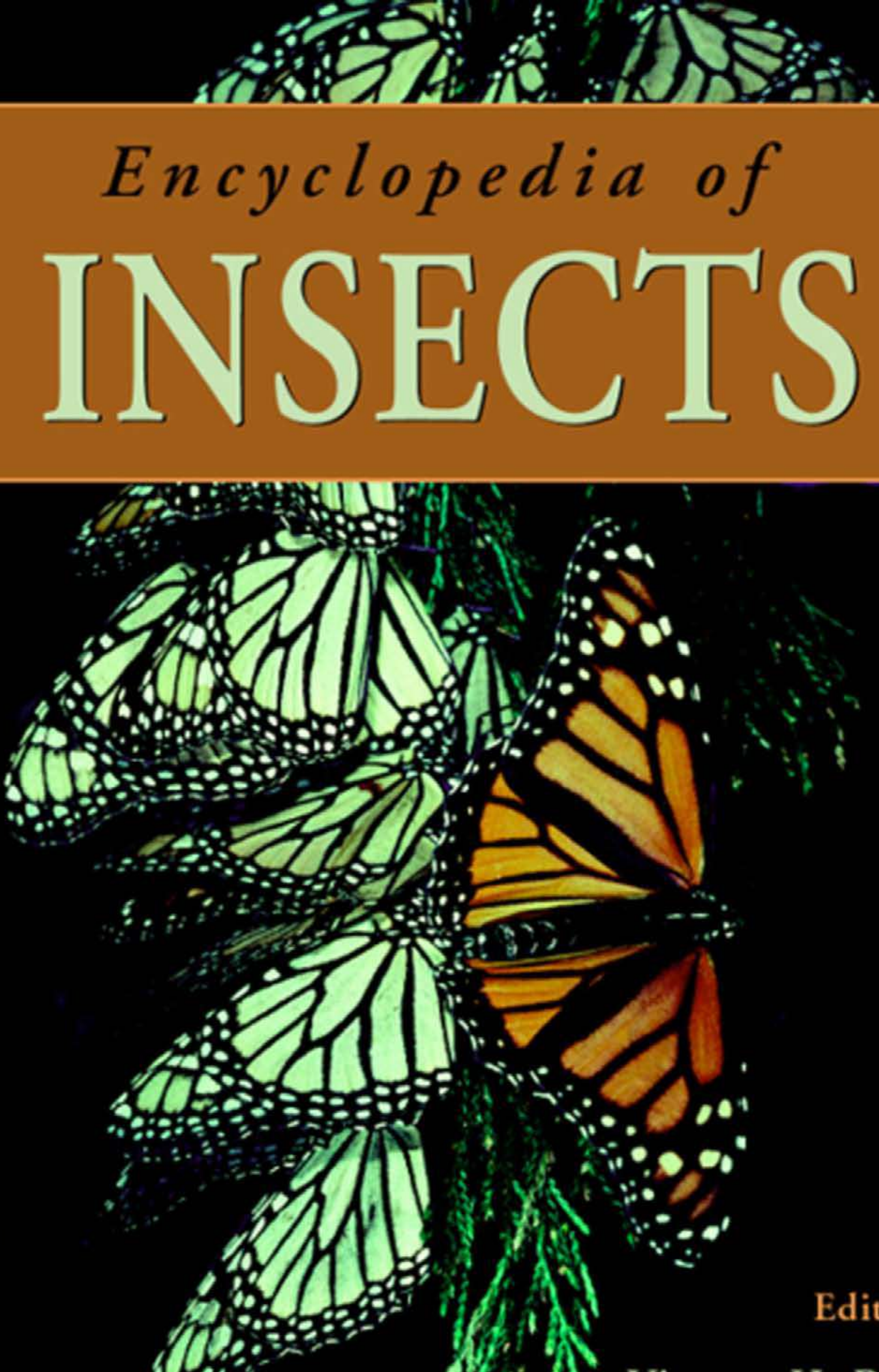

Editors

Vincent H. Resh

Ring T. Cardé

A

Acari

see *Mites; Ticks*

Accessory Glands

Diana E. Wheeler
University of Arizona

The accessory glands of reproductive systems in both female and male insects produce secretions that function in sperm maintenance, transport, and fertilization. In addition, accessory glands in females provide protective coatings for eggs. Accessory glands can be organs distinct from the main reproductive tract, or they can be specialized regions of the gonadal ducts (ducts leading from the ovaries or testes). Typically, glandular tissue is composed of two cell types: one that is secretory and the other that forms a duct. The interplay between male and female secretions from accessory glands is a key element in the design of diverse mating systems.

ACCESSORY GLANDS OF FEMALES

Management of Sperm and Other Male Contributions

Sperm management by females involves a wide range of processes, including liberation of sperm from a spermatophore, digestion of male secretions and sperm, transport of sperm to and from the spermatheca, maintenance of stored sperm, and fertilization.

Accessory gland secretions can have digestive functions important in sperm management. First, digestive breakdown of the spermatophore can free encapsulated sperm for fertilization and storage. Second, male contributions can provide an important nutritional benefit to their mates. Female secretions can digest the secretory components of male seminal fluid to facilitate a nutritive role. In addition, females can digest unwanted sperm to transform it into nutrients. Third, female secretions in some species are required to digest sperm coverings that inhibit fertilization.

Transfer of sperm to and from the spermatheca is generally accomplished by a combination of chemical signals and muscular contractions. Secretions of female accessory glands in some species increase sperm motility or appear to attract sperm toward the spermathecae. Transport of fluid out through the wall of the spermatheca may also create negative pressure that draws in sperm.

Sperm can be stored for some length of time in spermathecae, with the record belonging to ant queens that maintain sperm viability for a decade or more. Secretions of spermathecal glands are poorly characterized, and how sperm is maintained for such extended periods is not known. Spermathecal tissue seems to create a chemical environment that maintains sperm viability, perhaps through reduced metabolism. A nutritional function is also possible.

Transport of sperm out of storage can be facilitated by the secretions of the spermathecal gland, which presumably activate quiescent sperm to move toward the primary reproductive tract. One potential function of female accessory glands that has been explored only slightly is the production of hormone-like substances that modulate reproduction functions.

Production of Egg Coverings

Female accessory glands that produce protective coverings for eggs are termed colleterial glands. Colleterial glands have been best characterized in cockroaches, which produce an oothecal case surrounding their eggs. Interestingly, the left and right glands are anatomically different and have different products. Separation of the chemicals permits reactions to begin only at the time of mixing and ootheca formation. Other protective substances produced by glands include toxins and antibacterials.

Nourishment for Embryos or Larvae

Viviparous insects use accessory glands to provide nourishment directly to developing offspring. Tsetse flies and sheep keds are dipterans that retain single larvae within their reproductive tracts and provide them with nourishment. They give birth to mature larvae ready to pupate. The gland that produces the nourishing secretion, rich in amino acids and lipids, is known as the milk gland. The Pacific beetle roach, *Diploptera punctata*, is also viviparous and provides its developing embryos with nourishment secreted by the brood sac, an expanded portion of oviduct.

ACCESSORY GLANDS OF MALES

Accessory glands of the male reproductive tract have diverse functions related to sperm delivery and to the design of specific mating systems.

Sperm Delivery

Males of many insects use spermatophores to transfer sperm to females. A spermatophore is a bundle of sperm contained in a protective packet. Accessory glands secrete the structural proteins necessary for the spermatophore's construction. Males of the yellow mealworm, *Tenebrio molitor*, have two distinct accessory glands, one bean-shaped and the other tubular (Fig. 1). Bean-shaped accessory glands contain cells of at least seven types and produce a semisolid material that forms the wall and core of the spermatophore. Tubular accessory glands contain only one type of cell, and it produces a mix of water-soluble proteins of unknown function. Spermatophores are not absolutely required for sperm transfer in all insects. In many insects, male secretions create a fluid medium for sperm transfer.

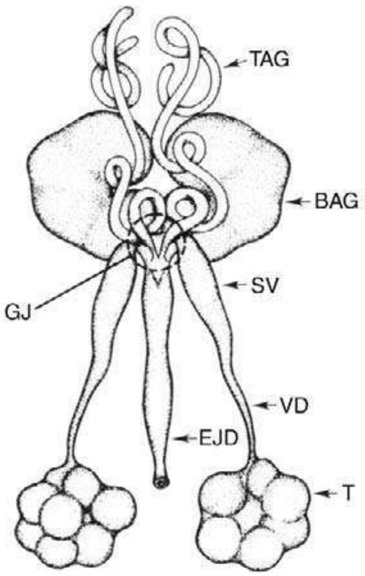

FIGURE 1 Male reproductive system of *T. molitor*, showing testes (T), ejaculatory duct (EJD), tubular accessory gland (TAG), and bean-shaped accessory gland (BAG). [From Dailey, P. D., Gadzama J. M., and Happ, G. M. (1980). Cytodifferentiation in the accessory glands of *Tenebrio molitor*. VI. A congruent map of cells and their secretions in the layered elastic product of the male bean-shaped accessory gland. *J. Morphol.* **166**, 289–322. Reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.]

Effects on Sperm Management and on the Female

The effects of male accessory gland secretions in the female are best known for the fruit fly, *Drosophila melanogaster*, in which the function of several gene products has been explored at the molecular level. Since insects have a diversity of mating systems, the specific functions of accessory gland secretions are likely to reflect this variation.

In *Drosophila*, the accessory glands are simple sacs consisting of a single layer of secretory cells around a central lumen (Fig. 2). Genes for more than 80 accessory gland proteins have been identified so far. These genes code for hormonelike substances and enzymes, as well as many novel

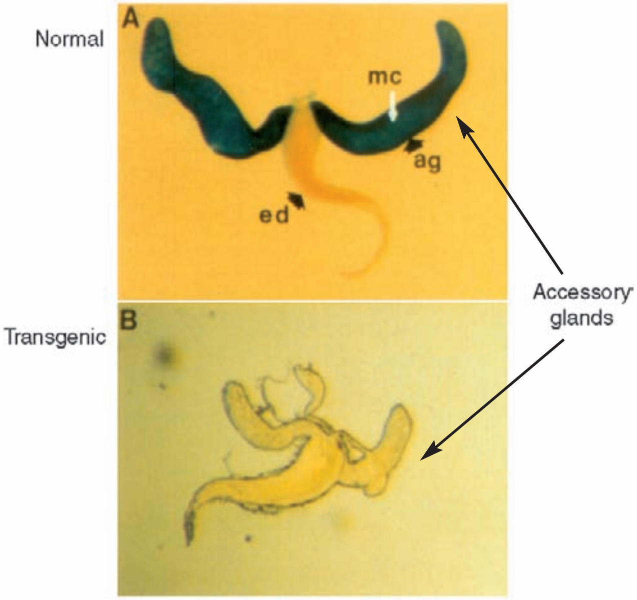

FIGURE 2 Accessory gland of *D. melanogaster*. (A) The cells in this normal accessory gland express b-galactosidase driven by a promoter of a gene for an accessory gland protein. (B) A transgenic accessory gland, cells expressing the gene have been selectively killed after eclosion. These flies were used to explore the function of accessory gland secretions. In transgenic males, accessory glands are small and translationally inert. [From Kalb, J. M., DiBenedetto, A. J., and Wolfner, M. F. (1993). Probing the function of *Drosophila melanogaster* accessory glands by directed cell ablation. *Proc. Natl. Acad. Sci. USA* **90**, 8093–8097. Copyright 1993, National Academy of Sciences, U.S.A.]

proteins. The gene products or their derivatives have diverse functions, including an increased egg-laying rate, a reduced inclination of females to mate again, increased effectiveness of sperm transfer to a female's spermatheca, and various toxic effects most likely involved in the competition of sperm from different males. A side effect of this toxicity is a shortened life span for females. Other portions of the reproductive tract contribute secretions with diverse roles. For example, the ejaculatory bulb secretes one protein that is a major constituent of the mating plug, and another that has antibacterial activity.

See Also the Following Articles

Egg Coverings • *Spermatheca* • *Spermatophore*

Further Reading

Chen, P. S. (1984). The functional morphology and biochemistry of insect male accessory glands and their secretions. *Annu. Rev. Entomol.* **29**, 233–255.

Eberhard, W. G. (1996). "Female Control: Sexual Selection by Cryptic Female Choice." Princeton University Press, Princeton, NJ.

Gillott, C. (1988). Arthropoda—Insecta. *In* "Accessory Sex Glands," (Adiyodi and Adiyodi, eds.). Vol. 3 of "Reproductive Biology of Invertebrates," pp. 319–471. Wiley, New York.

Happ, G. M. (1992). Maturation of the male reproductive system and its endocrine regulation. *Annu. Rev. Entomol.* **37**, 303–320.

Wolfner, M. F. (2001). The gifts that keep on giving: Physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila. Heredity* **88**, 85–93.

Acoustic Behavior

see *Hearing*

Aedes Mosquito

see *Mosquitoes*

Aestivation

Sinzo Masaki
Hirosaki University

Aestivation is a dormant state for insects to pass the summer in either quiescence or diapause. Aestivating, quiescent insects may be in cryptobiosis and highly tolerant to heat and drought. Diapause for aestivation, or summer diapause, serves not only to enable the insect to tolerate the rigors of summer but also to ensure that the active phase of the life cycle occurs during the favorable time of the year.

QUIESCENCE

Quiescence for aestivation may be found in arid regions. For example, the larvae of the African chironomid midge, *Polypedilum vanderplanki,* inhabit temporary pools in hollows of rocks and become quiescent when the water evaporates. Dry larvae of this midge can "revive" when immersed in water, even after years of quiescence. The quiescent larva is in a state of cryptobiosis and tolerates the reduction of water content in its body to only 4%, surviving even brief exposure to temperatures ranging from +102°C to –270°C. Moreover, quiescent eggs of the brown locust, *Locustana pardalina,* survive in the dry soil of South Africa for several years until their water content decreases to 40%. When there is adequate rain, they absorb water, synchronously resume development, and hatch, resulting in an outburst of hopper populations. The above-mentioned examples are dramatic, but available data are so scanty that it is difficult to surmise how many species of insects can aestivate in a state of quiescence in arid tropical regions.

SUMMER DIAPAUSE

Syndrome

The external conditions that insects must tolerate differ sharply in summer and winter. Aestivating and hibernating insects may show similar diapause syndromes: cessation of growth and development, reduction of metabolic rate, accumulation of nutrients, and increased protection by body coverings (hard integument, waxy material, cocoons, etc.), which permit them to endure the long period of dormancy that probably is being mediated by the neuroendocrine system.

Migration to aestivation sites is another component of diapause syndrome found in some species of moths, butterflies, beetles, and hemipterans. In southeastern Australia, the adults of the Bogong moth, *Agrotis infusa,* emerge in late spring to migrate from the plains to the mountains, where they aestivate, forming huge aggregations in rock crevices and caves (Fig. 1).

Seasonal Cues

Summer diapause may be induced obligatorily or facultatively by such seasonal cues as daylength (nightlength) and temperature. When it occurs facultatively, the response to the cues is analogous to that for winter diapause; that is, the cues are received during the sensitive stage, which precedes the responsive (diapause) stage. The response pattern is, however, almost a mirror image of that for winter diapause (Fig. 2). Aestivating insects themselves also may be sensitive to the

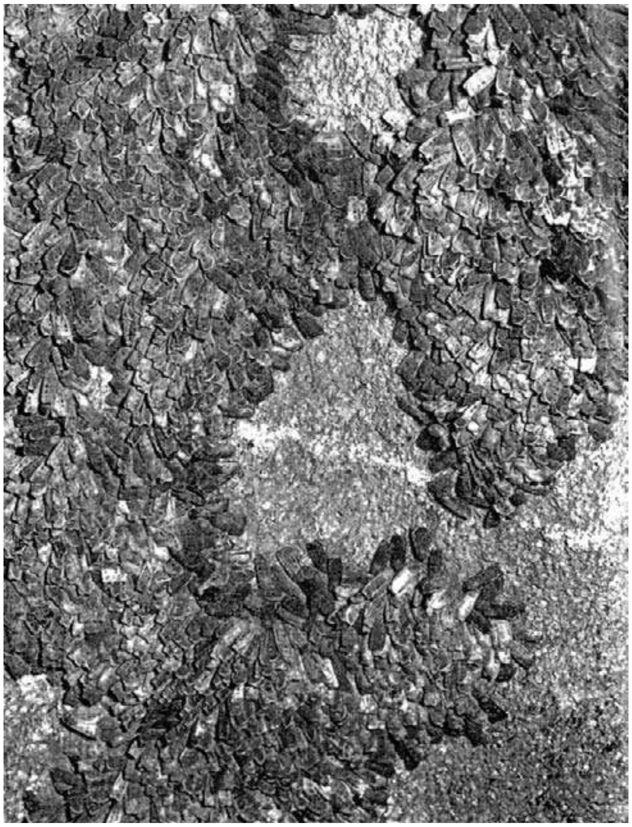

FIGURE 1 Bogong moths, *Agrotis infusa,* aestivating in aggregation on the roof of a cave at Mt Gingera, A. C. T., Australia. [Photograph from Common, I. (1954). *Aust. J. Zool.* **2,** 223–263, courtesy of CSIRO Publishing.]

seasonal cues; a high temperature and a long daylength (short nightlength) decelerate, and a short daylength (long nightlength) and a low temperature accelerate the termination of diapause.

The optimal range of temperature for physiogenesis during summer diapause broadly overlaps with that for morphogenesis, or extends even to a higher range of temperature. Aestivating eggs of the brown locust, *L. pardalina,* can terminate diapause at 35°C and those of the earth mite, *Halotydeus destructor,* do this even at 70°C. The different thermal requirements for physiogenesis clearly distinguish summer diapause from winter diapause, suggesting that despite the superficial similarity in their dormancy syndromes, the two types of diapause involve basically different physiological processes.

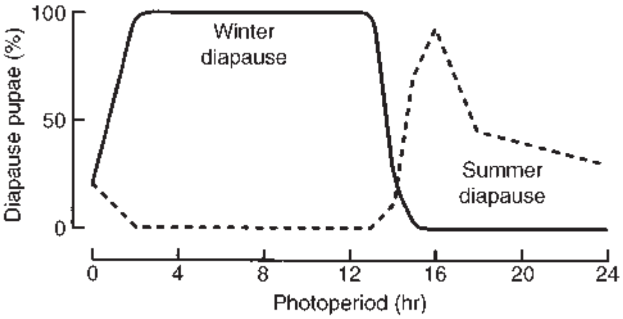

FIGURE 2 Photoperiodic response in the noctuid *M. brassicae* controlling the pupal diapause at 20°C. Note the different ranges of photoperiod for the induction of summer diapause (dashed line) and winter diapause (solid line). [From Furunishi *et al.,* 1982, reproduced with permission.]

See Also the Following Articles

Cold/Heat Protection • Diapause • Dormancy • Migration

Further Reading

Common, I. F. B. (1954). A study of the biology of the adult Bogong moth, *Agrotis infusa* (Boisd.) (Lepidopera: Noctuidae), with special reference to its behaviour during migration and aestivation. *Austral. J. Zool.* **2,** 223–263.

Furunishi, S., Masaki, S., Hashimoto, Y., and Suzuki, M. (1982). Diapause response to photoperiod and night interruption in *Mamestra brassicae* (Lepidoptera: Noctuidae). *Appl. Entomol. Zool.* **17,** 398–409.

Hinton, H. E. (1960). Cryptobiosis in the larva of *Polypedilum vanderplanki* Hint. (Chironomidae). *J. Insect Physiol.* **5,** 286–300.

Masaki, S. (1980). Summer diapause. *Annu. Rev. Entomol.* **25,** 1–25.

Matthée, J. J. (1951). The structure and physiology of the egg of *Locustana pardalina* (Walk.). *Union S. Afr. Dept. Agric. Sci. Bull.* **316,** 1–83.

Tauber, M. J., Tauber, C. A., and Masaki, S. (1986). "Seasonal Adaptations of Insects." Oxford University Press, New York.

Africanized Bees

see *Neotropical African Bees*

Agricultural Entomology

Marcos Kogan
Oregon State University
Ronald Prokopy
University of Massachusetts

The study of all economically important insects is the object of the subdiscipline "economic entomology." Agricultural entomology, a branch of economic entomology, is dedicated to the study of insects of interest to agriculture because they help increase crop production (e.g., pollinators); help produce a commodity (e.g., honey, silk, lacquer); cause injury leading to economic losses to plants grown for food, feed, fiber, or landscaping; cause injury to farm animals; or are natural enemies of agricultural pests and, therefore, considered to be beneficial. Study of all fundamental aspects of the ecology, life history, and behavior of insects associated with agricultural crops and farm animals falls within the realm of agricultural entomology. These studies provide the foundation for the design and implementation of integrated pest management (IPM) programs (Fig. 1).

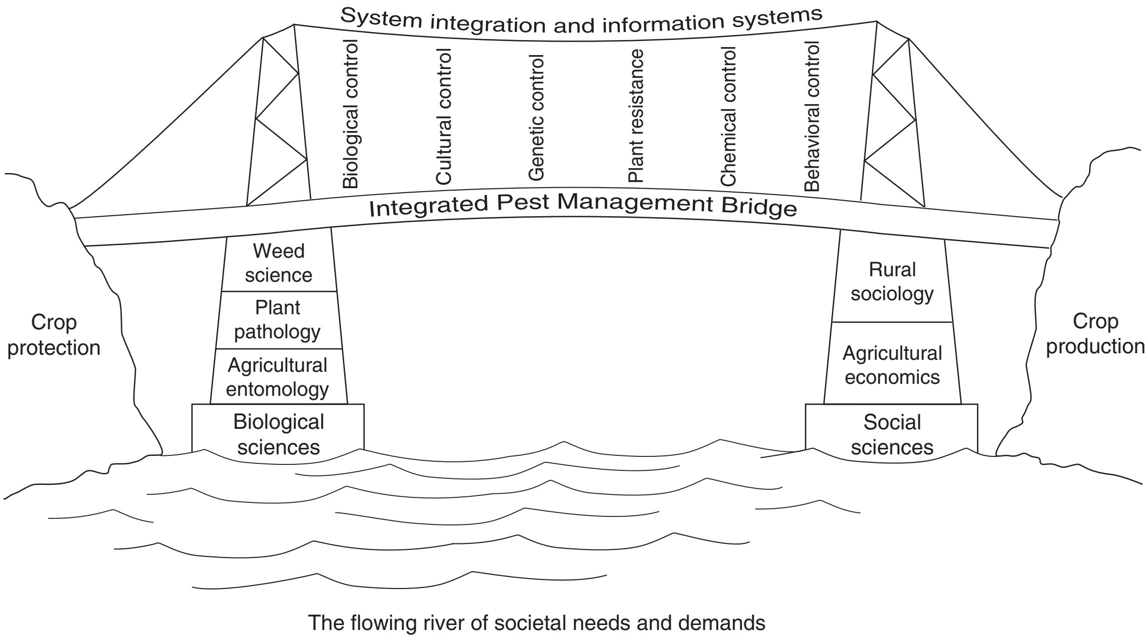

FIGURE 1 A bridge metaphor: agricultural entomology is conceived as one of the main pillars, together with plant pathology and weed science, of supporting the "integrated pest management bridge." The bridge connects two-way "traffic" between crop production and crop protection. The other pillar is provided by the social sciences of economics and sociology. The main tension cables, which are system integration and information systems, hold the vertical lines that together give stability to the bridge; these are the tactical components of IPM. Under the bridge runs the "river" of ever shifting societal needs and demands.

ECONOMIC ENTOMOLOGY

Insects are regarded by some as the main competitors of humans for dominance on the earth. Humans depend on insects for pollination of many crops, for production of honey and silk, for the decomposition of organic matter and the recycling of carbon, and for many other vital ecological roles. But it is the negative impact of insect pests that has been of greatest concern to humans. There are no reliable estimates of aggregate losses caused by insects as vectors of pathogens and parasites of humans and domestic animals, as agents causing direct damage to dwellings and other human-made structures, and as pests of crop plants and farm animals, but the amounts run to probably hundreds of billions of dollars annually. Losses caused by insects and vertebrate pests worldwide in the production of only eight principal food and cash crops (barley, coffee, cotton, maize, potato, rice, soybean, and wheat) between 1988 and 1990 have been estimated at $90.5 billion.

In the late 1800s and early 1900s, entomology became established in many academic and research institutions as a discipline equal in rank with botany and zoology. The diversity of insects and their economic importance was the justification for ranking the study of a class of animals (Insecta) as being equivalent to the study of two kingdoms of organisms (plants and animals other than insects). Through the first half of the twentieth century, there was a schism between basic and applied (or economic) entomology. Since then, common use of the expression "economic entomology" has declined, being replaced by designations of its principal branches, such as agricultural entomology, forest entomology, urban entomology, and medical and veterinary entomology. A detailed historical account is beyond the scope of this article, but Table I provides a chronology of some landmarks in the development of agricultural entomology through the ages.

The realm of agricultural entomology includes all basic studies of beneficial and pest insects associated with agricultural crops and farm animals. This article deals mainly with crops, but the general principles and concepts are equally applicable to farm animals. The starting point of such studies is a correct identification of the insect species, in accordance with the science known as biosystematics.

BIOSYSTEMATICS

Scientific nomenclature is a powerful tool for obtaining information about the basic biology of closely related species within a genus. When systematic studies have been extended beyond the naming of species (taxonomy) and contain detailed information on geographic distribution, host records, and biology of one or more species in a genus, it is often possible to extrapolate the information to other closely related species of that genus. Although details of the biology must be ascertained for each individual species, biosystematics offers a blueprint to follow when dealing with a new pest. For example, the genus *Cerotoma* (Coleoptera: Chrysomelidae) contains 10 to 12 species distributed from southern Brazil to the northeastern United States. All seem to be associated with herbaceous plants in the family Fabaceae (bean family). The biology of two of the species, *C. trifurcata* in North America and *C. arcuata* in South America (Fig. 2), has been studied extensively. Based

TABLE I **Some Landmarks in the Historical Development of Agricultural Entomology**[a]

Significant events	Years ago from 2000	Date
Beginnings of agriculture	10,000	8000 B.C.E.
First records of insecticide use	4,500	2500 B.C.E.
First descriptions of insect pests	3,500	1500 B.C.E.
Soaps used to control insects in China	900	1100
Beginning of scientific nomenclature—10th edition of Linnaeus, *Systema Naturae*	242	1758
Burgeoning descriptions of insects	100–200	18th and 19th centuries
First record of plant resistance to an insect	169	1831
Charles Darwin and Alfred Wallace jointly present paper on the theory of evolution	142	1858
First successful case of biological control: the cottony cushion scale, on citrus, in California, by the vedalia beetle	112	1888
First record of widespread damage of cotton in Texas by the cotton boll weevil	106	1894
First record of an insect resistant to an insecticide	86	1914
First edition of C. L. Metcalf and W. P. Flint's *Destructive and Useful Insects*	72	1928
Discovery of DDT and beginning of the insecticide era	61	1939
First report of insect resistance to DDT	54	1946
Term "pheromone" coined by P. Karlson and P. Butenandt, who identified first such substance in the silkworm moth	45	1959
First edition of Rachel Carlson's *Silent Spring*	48	1962
Expression "integrated pest management" first appears in the press	32	1968
Rapid development of molecular biology	20	1980s
Release of Bt transgenic varieties of cotton, corn, and potato	5	1990s

[a] Based in part on Norris *et al.* (2003).

on information for these two species, it is possible to infer that the other species in the genus share at least some of the following features: eggs are laid in the soil adjacent to growing leguminous plants; larvae feed on nitrogen-fixing root nodules and pupate in soil inside pupal cases; first-generation adults emerge when seedlings emerge, and second-generation adults emerge when plants are in full vegetative growth, feeding first on foliage and, later on, switching to feeding on developing pods. The biosystematic information on the genus allows students of agricultural entomology in South, Central, or North America to understand, at least in general terms, the role of any other species of *Cerotoma* within their particular agroecosystem.

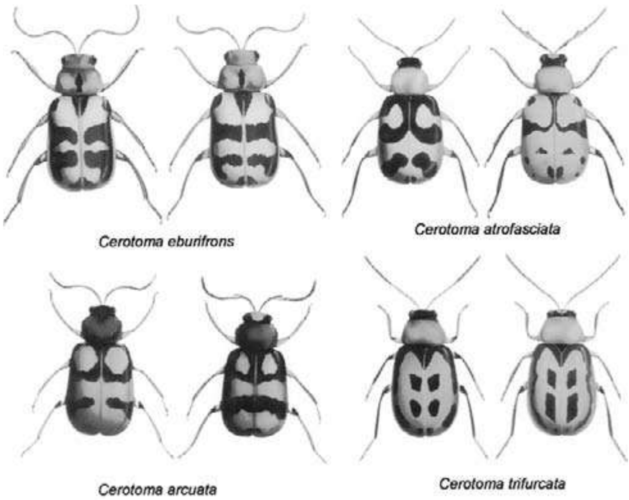

FIGURE 2 Morphological diversity and biological similarities in the genus *Cerotoma*: four of the dozen known species are illustrated by male and female specimens. The species are clearly distinguishable by morphological characters, but they have similar life histories and behaviors. (From unpublished drawings by J. Sherrod, Illinois Natural History Survey.)

The flip side of this notion is recognition that closely related and morphologically nearly undistinguishable (sibling) species may have many important biological differences. Examples of the critical need for reliable biosystematics studies are found in the biological control literature. The present account is based on studies conducted by Paul DeBach, one of the leading biological control specialists of the twentieth century. The California red scale, *Aonidella aurantii*, is a serious pest of citrus in California and other citrus-producing areas of the world. Biological control of the red scale in California had a long history of confusion and missed opportunities because of misidentification of its parasitoids. The red scale parasitoid *Aphytis chysomphali* had been known to occur in California and was not considered to be a very effective control agent. When entomologists discovered parasitized scales during foreign exploration, the parasitoids were misidentified as *A. chrysomphali* and therefore were not imported into California. It was later discovered that the parasitoids were in fact two different species, *Aphytis lingnanensis* and *A. melinus*, both more efficient natural enemies of the California red scale than *A. chrysomphali*. Now *A. lignanensis* and *A. melinus* are the principal red scale parasitoids in California. Further biosystematics studies have shown that what was once thought to be single species, *A. chrysomphali*, parasitic on the California red scale in the Orient and elsewhere, and accidentally established in California, is in fact a complex including at least seven species having different biological adaptations but nearly indistinguishable morphologically.

Knowledge of the name of a species, however, is not an indication of its true potential economic impact or pest status. A next important phase in agricultural entomology is, therefore, the assessment of benefits or losses caused by that species.

PEST IMPACT ASSESSMENT

The mere occurrence of an insect species in association with a crop or a farm animal does not necessarily mean that the species is a pest of that crop or animal. To be a pest it must cause economic losses. The assessment of economic losses from pests is the subject of studies conducted under conditions that match as closely as possible the conditions under which the crop is grown commercially or the animals are raised. Much of the methodology used in crop loss assessment has been established under the sponsorship of the Food and Agriculture Organization (FAO) of the United Nations as a means of prioritizing budget allocations and research efforts. Key data for these studies relate to the determination of the yield potential of a crop. The genetic makeup of a crop variety determines its maximum yield in the absence of adverse environmental factors. This is known as the attainable yield. To determine the attainable yield, the crop is grown under nearly ideal conditions; the actual yield is what occurs when the crop is grown under normal farming conditions. The difference between attainable and actual yields is a measure of crop loss (Fig. 3).

To assess crop losses and attribute the losses to a specific cause (e.g., the attack of a pest) requires setting up experiments to isolate the effect of the pest from all other constraints. Methodologies vary with pest category—whether the pests are insects, vertebrates, plant pathogens, or weeds, for example. The quantitative relationship between crop losses and pest population levels is the basis for computing the economic injury level for the pest. The economic injury level is a fundamental concept in IPM.

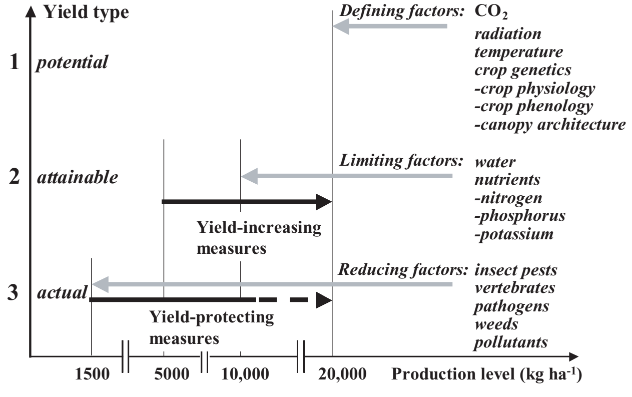

FIGURE 3 Factors impacting the yield potential of a generic crop. (Adapted from information on a Web site originated at IMI/University of Miami, Summer Institute.)

LIFE HISTORY AND HABITS

Once the identity and pest status of a species have been well established, it becomes essential to extend the informational base on the life history and habits of the species to the conditions under which the crop is grown. Economically important life history traits include information on developmental threshold temperatures and temperature-dependent developmental rates. These data are used in modeling the phenology of the pest. Other essential studies include the orientation, feeding, host selection, and sexual behavior of the species. Many of these studies provide the foundation for strategic planning in IPM and for the development of target-specific control tactics. For example, the study of sexual behavior involves the definition of the role of pheromones in mating and the identification of those pheromones. These, in turn, may be used for monitoring pest incidence and abundance or in mating disruption, both valuable components of IPM systems for many crops. The study of host selection behavior often leads to the identification of kairomones, equally important in IPM development.

PHENOLOGY

The life cycle of different insect species varies greatly, although all insects undergo the basic stages of development from egg to reproductive adult (or imago). Depending on the length of the life cycle, there is considerable variation in the number of generations per year, a phenomenon called voltinism. A univoltine species has one generation per year; a multivoltine species may have many generations per year. The range of variation in the Insecta is evident when one considers that the 17-year periodical cicada has one generation every 17 years, whereas whiteflies or mosquitoes may complete a generation in about 21 days. Under temperate climate conditions, generations often are discrete, but under warmer subtropical conditions they frequently overlap. The definition of temporal periodicity in an organism's developmental cycle is called phenology. The relationship between the phenology of the crop and the phenologies of its various pests is of interest in agricultural entomology. Figure 4 shows an example of such a relationship for soybean grown under conditions typical for the midwestern United States.

POPULATION AND COMMUNITY ECOLOGY

Population and community level studies are within the scope of insect ecology. Although the species is the focal biological entity for agricultural entomology, for management purposes it is essential to understand population and community level processes. Populations are assemblages of conspecific individuals within a defined geographical area (e.g., a crop field, a river valley, a mountain chain). Many insects have a large reproductive capacity. As calculated by Borror, Triplehorn, and Johnson, a pair of fruit flies (*Drosophila*), for

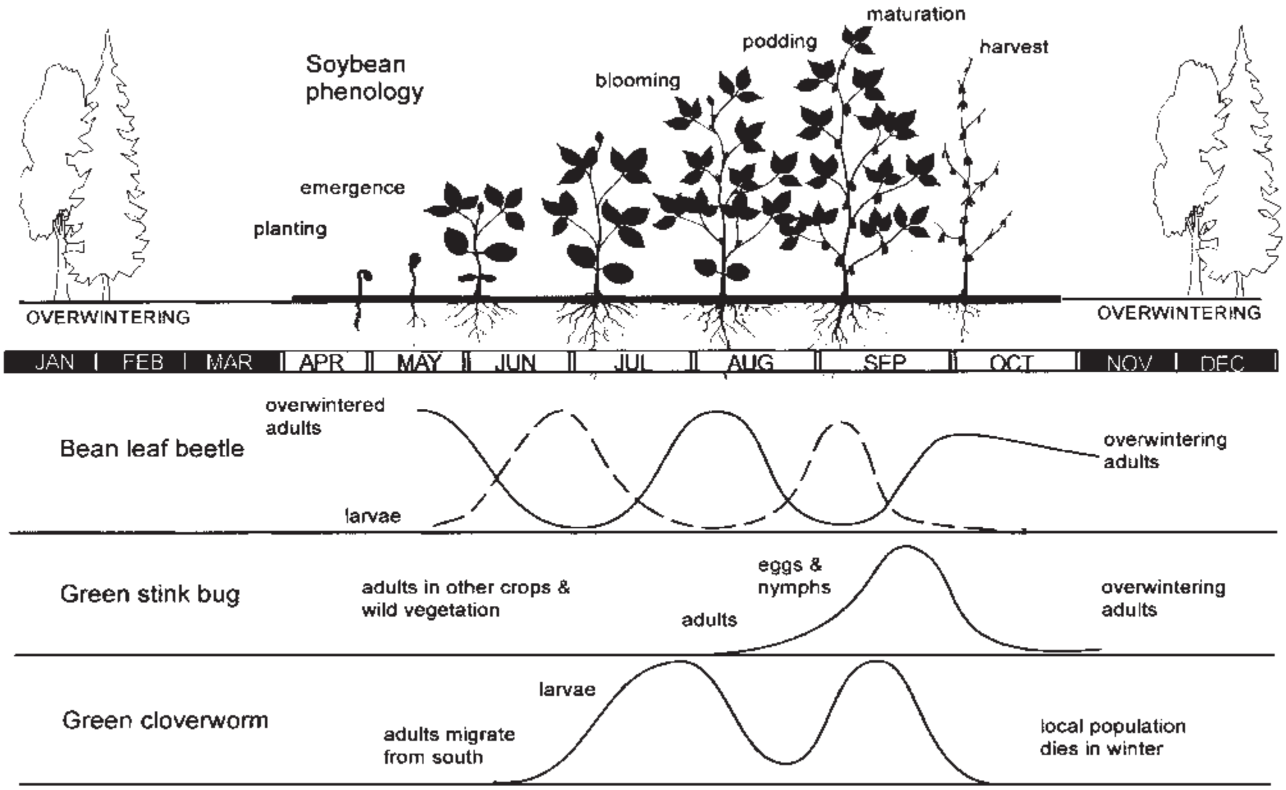

FIGURE 4 Crop phenology and pest phenology: relationship between the phenology of soybean in the midwestern United States and three of its most common insect pests, the bean leaf beetle, *C. trifurcata* (Coleoptera: Chrysomelidae); the green stink bug, *Acrosternum hilare* (Hemiptera: Pentatomidae); and the green cloverworm, *Hypena scabra* (Lepidoptera: Noctuidae).

example, produces 100 viable eggs, half of which yield females that in turn will lay 100 eggs and so on for 25 possible generations in one year; by the end of the year, the 25th generation would contain 1.192×10^{41} flies, which, if packed tightly together, 60,000 to a liter, would form a ball of flies 155 million km in diameter or a ball extending approximately from the earth to the sun. Obviously, such unlimited population growth does not occur in nature. Normally, populations are regulated by the combined actions of both physical (or abiotic) and biological (or biotic) factors of the environment. An understanding of the mortality factors that help regulate insect populations is one of the most active areas of research in agricultural entomology.

The set of species coexisting in an area and interacting to varying degrees form what is known as an ecological community. In a crop community, the crop plants and the weeds that persist within the crop field or grow along the borders are the primary producers. The animals within the crop community maintain dynamic trophic relationships: some feed on living plants, others on the decaying plants, and still others on animals. Those that feed on the plants are the herbivores, or primary consumers. Pests are primary consumers on the crop plants. Parasitoids and predators are the secondary consumers. Those that feed on the pests are beneficial natural enemies. Finally, decomposers and detritivores feed on decaying organic matter. All biotic components of the community are interconnected by "food webs." An understanding of food webs and trophic interactions in crop communities is important because it provides a basis for interpreting the nature of disturbances in crop ecosystems. Disturbances in trophic relations may lead to outbreaks of pest organisms and the need for control actions.

LINKS TO IPM SYSTEMS DEVELOPMENT

With the advent of integrated pest management and its success in the last third of the twentieth century, it has become difficult to separate agricultural entomology from IPM. In entomology, the two fields of endeavor are inextricably interconnected. A reliable database of biological information provides the means to design and develop IPM strategies. For example, there is growing interest in methods of enhancing biological control through habitat management. The technique requires information on source–sink relationships among pests and natural enemies across crop plants, neighboring crops, natural vegetation, and especially managed vegetation in the form of cover crops and field hedges. Theoretically, diversification of the crop ecosystem leads to an increase in natural enemies and to greater stability of the system. The complexity of interactions, however, makes it difficult to interpret conflicting results of experiments designed to test working hypotheses. The analysis of within-field and interfield movement, the host selection behavior of phytophagous and entomophagous insects, multitrophic interactions among community members, and the dynamics of populations, all under the scope of agricultural entomology, are only a few of the many components of the knowledge base necessary to develop advanced IPM systems.

The advent of the World Wide Web has had a major influence on accessibility to basic information on agricultural entomology. Most major agricultural research centers have developed Web pages that organize information and make it available to students worldwide. More importantly, the dynamic nature of the Web offers the opportunity to provide weather-driven modeling capabilities that greatly increase the scope and applicability of studies about the phenology and population dynamics of major pest organisms. Two sites that offer such capabilities are http://www.orst.edu/Dept/IPPC/wea/ and http://www.ipm.ucdavis.edu/PHENOLOGY/models.html.

Entomologists in the late 1800s and early 1900s studied the biology of insect pests in great detail. Articles and monographs published during that period remain valuable sources of information. These early entomologists recognized that deep knowledge of the life history of an insect and its habits could provide insights useful for the control of agricultural and other pests. The advent of organosynthetic insecticides in the mid-1940s created the illusion that pest problems now could be solved forever. Many entomologists redirected their efforts to testing new chemicals and neglected basic insect biology studies. The failure of insecticides to eradicate pests and the environmental problems engendered by the misuse of these chemicals led to the advent of IPM. For IPM to succeed, entomologists have had to return to the basics and again refocus their efforts on the study of insect biology. Agricultural entomology has come full circle as new generations of entomologists endeavor to refine knowledge of the group of animals that remain humans' most serious competitors.

See Also the Following Articles

Biological Control • History of Entomology • Integrated Pest Management • Phytophagous Insects • Plant–Insect Interactions • Population Ecology

Further Reading

Borror, D. J., Triplehorn, C. A., and Johnson, N. F. (1992). "An Introduction to the Study of Insects." Saunders College Publishers and Harcourt Brace College Publishers, Philadelphia and Fort Worth.

DeBach, P. (1964). "Biological Control of Insect Pests and Weeds." Reinhold, New York.

Evans, H. E. (1984). "Insect Biology: A Textbook of Entomology." Addison-Wesley, Reading, MA.

Huffaker, C. B. and Gutierrez, A. P. (1999). "Ecological Entomology." Wiley, New York

Jones, D. P. (1973). Agricultural entomology. *In* "History of Entomology." (R. F. Smith, T. E. Mittler, and C. N. Smith, eds.), pp. 307–332. Annual Reviews, Palo Alto, CA.

Norris, R. F., Caswell-Chen, E. P., and Kogan, M. (2003). "Concepts in Integrated Pest Management." Prentice Hall, Upper Saddle River, NJ.

Oerke, E. C. (1994). "Crop Production and Crop Protection: Estimated Losses in Major Food and Cash Crops." Elsevier, Amsterdam and New York.

Price, P. W. (1997). "Insect Ecology." Wiley, New York.

Schowalter, T. D. (2000). "Insect Ecology: An Ecosystem Approach." Academic Press, San Diego, CA.

Alderfly

see *Megaloptera*

Amber

George Poinar Jr.
Oregon State University

Amber is a fossilized resin ranging from several million to 300 million years of age. This material is a gold mine for the entomologist because it contains a variety of insects preserved in pristine, three-dimensional condition. Fossils in amber provide evidence of lineages dating back millions of years (Table I). External features are preserved so well that taxonomists can make detailed comparisons with living taxa to follow evolutionary development of genera and even species. Amber has a melting point between 200 and 380°C, a hardness of 2 to 3 on the Moh's scale, and a surface that is insoluble to organic solvents. Aside from providing direct evidence of an insect taxon at a particular time and place, amber insects give clues to past distributions and phylogeny, as well as indirect evidence of plants and vertebrates and the establishment of symbiotic associations, and clues for reconstructing ancient landscapes.

TABLE I Significant Amber Deposits in the World

Deposit	Location	Approximate age (million years)
Baltic	Northern Europe	40
Burmese	Burma (Myanmar)	100
Canadian	Alberta, Manitoba	70–80
Chinese	Fushun Province	40–53
Dominican	Dominican Republic	15–45
Hat Creek	British Columbia, Canada	50–55
Lebanese	Middle East	130–135
Mexican	Chiapas	22–26
New Jersey	Northeastern United States	65–95
Siberian (Taimyr)	Russian arctic	78–115
Spanish	Alava, Basque country	100–115

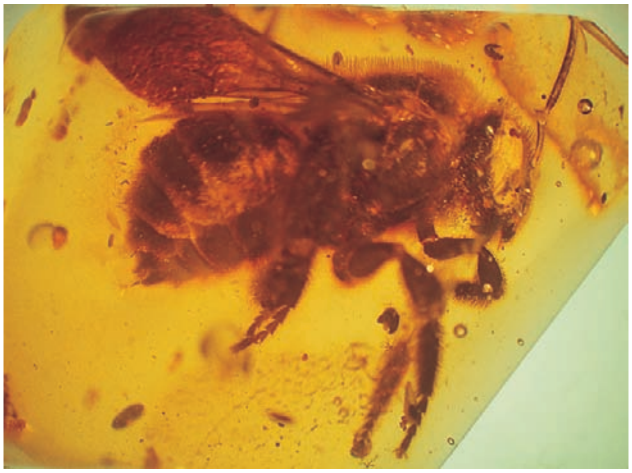

FIGURE 1 Origins of the honey bee lineage are provided by this primitive bee, which possesses characters of both modern *Apis* and extinct apids. Its presence in Baltic amber suggests a European origin of *Apis,* thus challenging the current view that honey bees originated in Asia.

USE OF AMBER IN TRACING INSECT LINEAGES

As a result of the excellent preservation of amber insects, specific genera (the majority, if not all, amber insects are extinct at the species level) can be recognized and compared with modern ones. In this way, lineages can be traced back tens of millions of years. An example is a small parasitic wasp of the genus *Aphelopus* (Hymenoptera: Dryinidae) trapped in Lebanese amber. This genus is still extant, and the fossil demonstrates a lineage that has survived for 130 to 135 million years.

The origin of genera can also be obtained from amber insects owing to their high degree of preservation. A recent example from Baltic amber, which contains a variety of bees, deals with the origin of the common honey bee. This fossil contains basic features characteristic of the genus *Apis* as we know it today, including pollen-collecting apparati on the hind legs and a barbed stinger (Fig. 1). This appears to be one of the most primitive *Apis* ever discovered, thus indicating a time (40 mya) and place (northern Europe) for the origin of the honey bee.

PROVIDING INDIRECT EVIDENCE OF OTHER ORGANISMS

There are size and habitat limitations to the types of organism that can be trapped in amber. For example, many plants would not normally leave flowers or leaves in the resin, and when they did, the remains would likely be difficult to identify. Vertebrates might leave hairs, feathers, or scales but these structures would also be difficult to identify. However, arthropods that are specific to certain hosts (e.g., ticks and mammals) can provide clues to other organisms that existed at that time. This use of fossils relies heavily on the principle of behavioral fixity, which asserts that, at least at the generic level, the behavior of a fossil organism would have been similar to that of its present-day descendants.

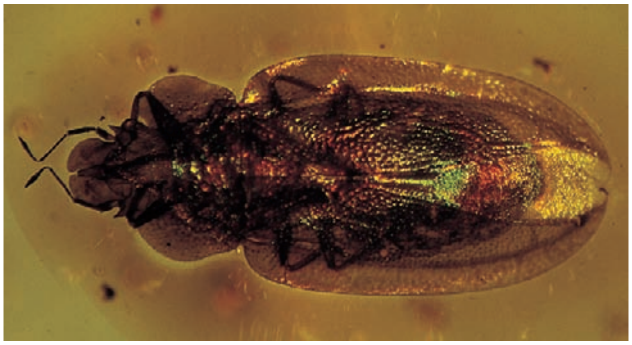

FIGURE 2 The unique morphological features (smooth, flattened body) of this fossil palm bug (*Paleodoris lattini*) in Dominican amber not only characterize it systematically but also provide clues to its lifestyle of living in confined spaces between the unopened fronds of pinnately leafed palms.

Many insects form specific associations with plants. Such associations can often be deduced by the morphological features of the insect (functional morphology). One extremely flattened hemipteran in Dominican amber (Fig. 2) that was identified as a palm bug displayed characters similar to those of an existing species in the same subfamily. The extant species lives between the closed leaves of royal palms (*Roystonea* spp.) in Cuba. This fossil provided indirect evidence that pinnately leafed palms, quite likely an extinct species of *Roystonea,* existed in the original amber forest. Other plant-specific insects, such as fig wasps and palm bruchids, provide evidence of figs and palms in the original ecosystem.

Insects that require a blood meal to complete their development can also be used as indirect evidence of a vertebrate group. Evidence of birds in the original Dominican amber forest is implied by the presence of a female *Anopheles* mosquito in amber because extant species of this subgenus normally attack birds. The presence of other vertebrate groups is implied by fleas (Siphonaptera), horseflies (Diptera: Tabanidae), biting midges (Diptera: Ceratopogonidae), and other bloodsucking arthropods such as ticks.

PROVIDING INDIRECT EVIDENCE OF SPECIFIC HABITATS

Amber insects can provide evidence of specific habitats. Diving beetles (Coleoptera: Dytiscidae), caddisflies (Trichoptera), and damsel flies (Odonata) all provide evidence of aquatic habitats. The *Anopheles* mosquito belongs to a group that normally oviposits in ground pools. Other insects can provide evidence of phytotelmata (standing water in plant parts), wood, moss, bark, and detritus.

PALEOSYMBIOSIS

Because of the sudden death of captured organisms in amber, symbiotic associations may be preserved in a manner unlikely to occur with other types of preservation. Also, the fine details of preservation may reveal morphological features characteristic

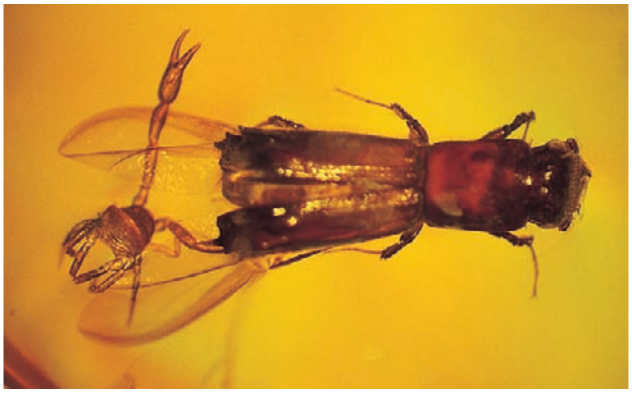

FIGURE 3 Documentation of paleophoresis is provided by a pseudoscorpion grasping the tip of the abdomen of a platypodid beetle in Dominican amber. Similar rider–carrier associations occur today, suggesting that this behavior is mandatory for survival of the pseudoscorpion.

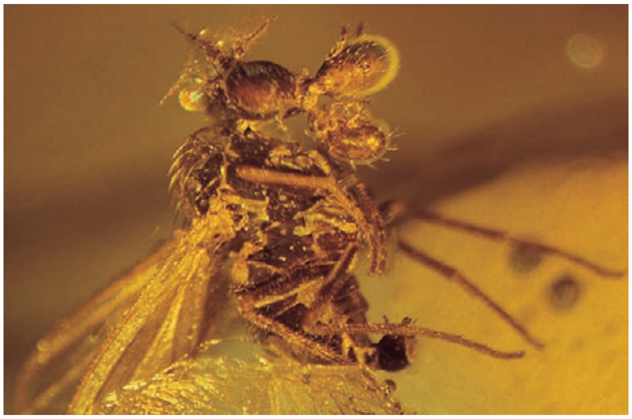

FIGURE 4 Paleoectoparasitism is shown by two thrombidid mites attached to the mouthparts of a long-legged fly (Diptera: Dolichopodidae) in Baltic amber.

of symbiotic associations. Cases of paleosymbiosis in amber include inquilinism, commensalism, mutualism, and parasitism.

Paleoinquilinism involves two or more extinct organisms living in the same niche but neither benefiting nor harming each other. Numerous insects form inquilinistic associations under tree bark, and many pieces of amber contain flies and beetles common to this habitat.

Phoresis (one organism transported on the body of another organism) is probably the most typical type of paleocommensalism in amber. This usually involves mites and pseudoscorpions being carried by insects. The arachnid benefits by being conveyed to a new environment, where the food supply is likely to be better than the last one. The carrier generally is not harmed and only serves as a transporting agent. An example of this category in Dominican amber consists of pseudoscorpions being carried by platypodid beetles (Coleoptera: Platypodidae) (Fig. 3). The method of attachment of the pseudoscorpion to the beetle was the same then as it is today. In fact, these ancient records lead scientists to believe that such behavior is mandatory for the survival of the pseudoscorpions that live in beetle tunnels and require effective dispersal mechanisms for survival.

In paleomutualism, both organisms benefit and neither is harmed. Amber bees carrying pollen provide evidence of insect–plant mutualism in which the bee obtains a food supply and the plant is pollinated. An example of insect–insect mutualism is demonstrated by a rare fossil riodinid butterfly larva in Dominican amber. Specialized morphological features of this *Theope* caterpillar indicative of a symbiotic association are balloon setae and vibratory papillae in the neck area, and tentacle nectary organ openings on the eighth abdominal tergite. Extant caterpillars in this genus have similar features and are associated with ants. The tentacle nectary organs provide nourishment for the ants, whereas the vibratory papillae (which beat against the head capsule and make an audible sound) and balloon setae (which emit a chemical signal) are used to attract ants when the caterpillar is threatened by an invertebrate predator or parasite. This fascinating association between butterfly larvae and ants was established at least 20 mya.

Paleoparasitism is very difficult to verify in the fossil record. There are many records of amber insects (especially wasps and flies) whose descendants today are parasitic on a wide range of organisms, but to discover an actual host–parasitic association is quite rare.

Paleoectoparasitism is the most obvious of all parasitic associations found in amber. The ectoparasite is often still attached to its host, and systematic studies can be conducted on both organisms. In amber, ectoparasites are usually parasitic mites, such as the larvae of Thrombididae attached to the mouthparts of a fly in Baltic amber (Fig. 4). These larval mites were feeding on the host's hemolymph, and their mouthparts are still in place. After molting to the nymphal stage, the parasites would leave the fly and become free-living predators. Large infestations could kill the host. These mites are not to be confused with phoretic ones, which are simply carried around by insects.

Paleoendoparasitism is extremely difficult to verify because internal parasites are rarely preserved as fossils. However, some parasites attempt to leave their hosts when they encounter resin. Mermithid nematodes (Mermithidae: Nematoda) and hairworms (Nematomorpha) that have nearly completed their development and are almost ready to emerge from their host will often reveal their presence (Fig. 5). Under normal conditions, they would enter soil or water and initiate a free-living existence.

BIOGEOGRAPHICAL STUDIES

In many instances amber insects provide evidence of a more extensive distribution in the past for various insect genera and families as well as indicating a warmer climatic regime in many parts of the world. Perhaps the most spectacular examples of this phenomenon are insects discovered in amber sites located far from their descendants' current habitat. Examples from

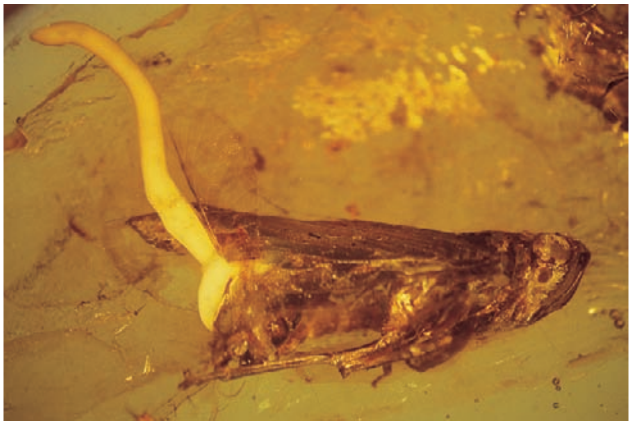

FIGURE 5 Paleoendoparasitism in amber is exemplified by a mermithid nematode (Nematoda: Mermithidae) emerging from the body of a planthopper (Homoptera: Fulgoroidea) in Baltic amber. Such records set minimum dates for the establishment of host–parasite associations.

Dominican amber include *Mastotermes* termites (Isoptera: Mastotermitidae) and *Leptomyrmex* ants (Hymenoptera: Formicidae) that obviously were part of the insect fauna some millions of years ago in the Caribbean but occur nowhere in the New World today. Both genera are represented today by a single relict species in the North and East Australian Region.

A North American example is the presence of the tropical arboreal ants of the genus *Technomyrmex* in Hat Creek amber in British Columbia, Canada, living 50 mya, hundreds of kilometers north of their present-day range. These tropical ants in Eocene Hat Creek amber provide evidence that the climate in that region of the world shifted from tropical to temperate. Other examples of past distributions involve the palm bug shown in Fig. 2, which has no present-day descendants in the Dominican Republic, with only a single living Cuban species in the subfamily. Similarly, there are no members of the genus *Theope* in the Dominican Republic or the Greater Antilles today, all living representatives being restricted to Mexico, and Central and South America. Further evidence of climatic shifts over time are clear with many of the Baltic amber insects, many of whose descendants occur in the Old World tropics today. The primitive honey bee shown in Fig. 1 evolved under subtropical conditions that characterized most of northern Europe in the Eocene. Thus, it is not surprising that most of the species and varieties of the genus *Apis* live only under tropical conditions today.

RECONSTRUCTING ANCIENT LANDSCAPES

Every amber fossil tells a story and is a piece of a jigsaw puzzle that can be used to reconstruct the natural environment at the time the amber was being produced. The challenges are to identify the inclusions, determine their biology and ecology by researching the habits of their extant descendants, and then make inferences regarding the original environment. There will always be gaps in the puzzle because there are many lifeforms that are too large to become entrapped in amber or have a lifestyle that does not normally bring them into contact with the sticky resin. However, the habitat that existed in that ancient world can, in large part, be reconstructed by studying select insects that can be typified as phytophagous, soil-loving, bark inhabitants, or parasites, and identifying the associated predators, vertebrates, and special habitats.

See Also the Following Articles

Biogeographical Patterns • Fossil Record

Futher Reading

Boucot, A. J. (1990). “Evolutionary Paleobiology of Behavior and Coevolution.” Elsevier, Amsterdam.

Poinar, G. O. Jr., (1992). “Life in Amber.” Stanford University Press, Stanford, CA.

Poinar, G. O., Jr., and Milki, R. (2001). “Lebanese Amber. The Oldest Insect Ecosystem in Fossilized Resin.” Oregon State University Press, Corvallis.

Poinar, G. O., Jr., and Poinar, R. (1994). “The Quest for Life in Amber.” Addison-Wesley, Reading, MA.

Poinar, G. O., Jr., and Poinar, R. (1999). “The Amber Forest.” Princeton University Press, Princeton, NJ.

Ross, A. (1998). “Amber.” Harvard University Press, Cambridge, MA.

Weitschat, W., and Wichard, W. (2002). “Atlas of Plants and Animals in Baltic Amber.” Pfeil, Munich.

Anatomy: Head, Thorax, Abdomen, and Genitalia

David H. Headrick
California Polytechnic State University
Gordon Gordh
U.S. Department of Agriculture

Anatomy is a subdiscipline of morphology concerned with naming and describing the structure of organisms based on gross observation, dissection, and microscopical examination. Morphology and anatomy are not synonyms. Morphology is concerned with the form and function of anatomical structure; because anatomy is an expression of organic evolution, morphology seeks to investigate possible explanations for organic diversification observed in nature. Before 1940 insect morphology focused on naming and describing anatomical structure. The need for this activity has not diminished, as much about insect anatomy remains to be revealed, described, and understood. This article focuses on the anatomical structures of the three major tagmata of the insect body: head, thorax, and abdomen, and on the external genitalia. A hypothetical ground plan for major structures is given, followed by themes in anatomical variation based on adaptation observed in the Insecta.

CONTEXT OF ANATOMICAL STUDY

Terms of Orientation and Conventions

Terms to describe orientation are not intuitive for insects. Most orientation terms are derived from the study of the human body—a body that stands upright—and their application to insects causes confusion. Some standard terms used with insects include anterior (in front), posterior (behind), dorsal (above), ventral (below), medial (middle), and lateral (side). Anatomical description usually follows in the same order, hence, we begin our discussion with the head, move on to the thorax and then the abdomen, and finish with the genitalia. Description of the relative placement of anatomical features can be cumbersome, but they are critical elements in the study of anatomical structure because relative position is one of the three basic tenets of homology, including size and shape, and embryology.

Measures of Success

The design of the insect body can be described as successful for many reasons: there are millions of species, they range in size over four orders of magnitude, their extensiveness of terrestrial and aquatic habitat exploitation (the diversity of resources), and once a successful form has been developed, there appears to be relatively little change over evolutionary time (Fig. 1). The basic insect design allows for adaptation to a variety of environmental requirements. The success of the design is rooted in the nature of the main material used for its construction.

The Building Material

When we look at an insect, it is the integument that we see. Structurally, the integument is a multiple-layered, composite organ that defines body shape, size, and color. The ultrastructure of the integument is composed of living cells and the secretory products of those cells. Each layer is of a different thickness and chemical composition, and each displays physical properties different from those of the surrounding layers. Perhaps more importantly, the integument also is the organ with the greatest diversity of structure and function.

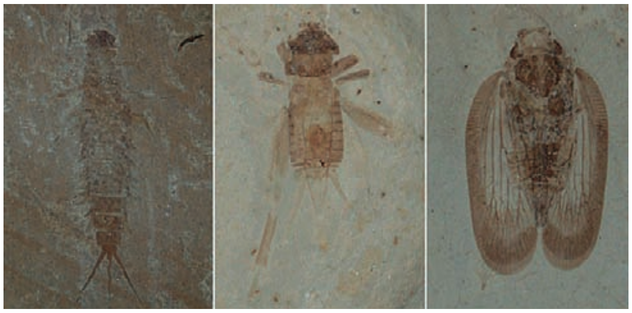

FIGURE 1 Fossil insects are easily recognizable today, indicating an early establishment of a successful design. Left to right: *Heplagenes* (Late Jurassic 150 mya, Liaoning, China); cricket (Eocene, 50 mya, Green River formation, Utah); fulgorid (Eocene, 50 mya, Green River formation, Utah).

There are two common misconceptions about the integument. First, some believe chitin is responsible for integument hardness. Actually, there is proportionally more chitin found in the soft and flexible membranous parts of the integument than in the hard, sclerotized plates. Integument hardness is attributed to an increased number of cross-linkages between protein chains contained in the integument layers. Second, some believe that the integument is rigid and that growth is incremental and limited to expansion during molting; yet some endopterygote insects are able to grow continuously between molts.

The integument determines the shape of the insect body and its appendages. One of the most captivating features of insects is their seemingly infinite variation in body shapes—everything from a simple bag (Hymenoptera grub) to a mimic of orchid flowers (Mantidae). Similarly, appendage shape is exceedingly plastic. Terms such as "pectinate," "flabbate," and "filiform" are among more than 30 terms taxonomists have proposed to describe antennal shapes. Leg shapes are similarly highly variable and express functional modifications. Among these shapes are "cursorial," "gressorial," "raptorial," "fossorial," and "scansorial." Again, these modifications of shape reflect the function of structure. Finally, wing shapes are highly variable among insects and are determined by body size and shape as well as by aerodynamic considerations.

Tagmata

Most people recognize the three tagmata—head, thorax, and abdomen—as characteristic of insects. The way they appear is rooted in a division of responsibilities. The head is for orientation, ingestion, and cognitive process; the thorax for locomotion; and the abdomen for digestion and reproduction. But even casual observations reveal further divisions of these body regions.

Segmentation of Tagmata

Two types of segmentation are evident among arthropods, primary and secondary. Primary segmentation is characteristic of soft-bodied organisms such as larval holometabolans. The body wall in these organisms is punctuated by grooves or rings that surround the anterior and posterior margin of each somite. These rings represent intersegmental lines of the body wall and define the limits of each somite. Internally, the grooves coincide with the lines of attachment of the primary longitudinal muscles. From a functional standpoint, this intrasegmental, longitudinal musculature permits flexibility and enables the body to move from side to side.

More complex plans of body organization exhibit structural modifications. Secondary segmentation is characteristic of hard-bodied arthropods, including adult and nymphal insects. Secondary body segmentation is an evolutionarily

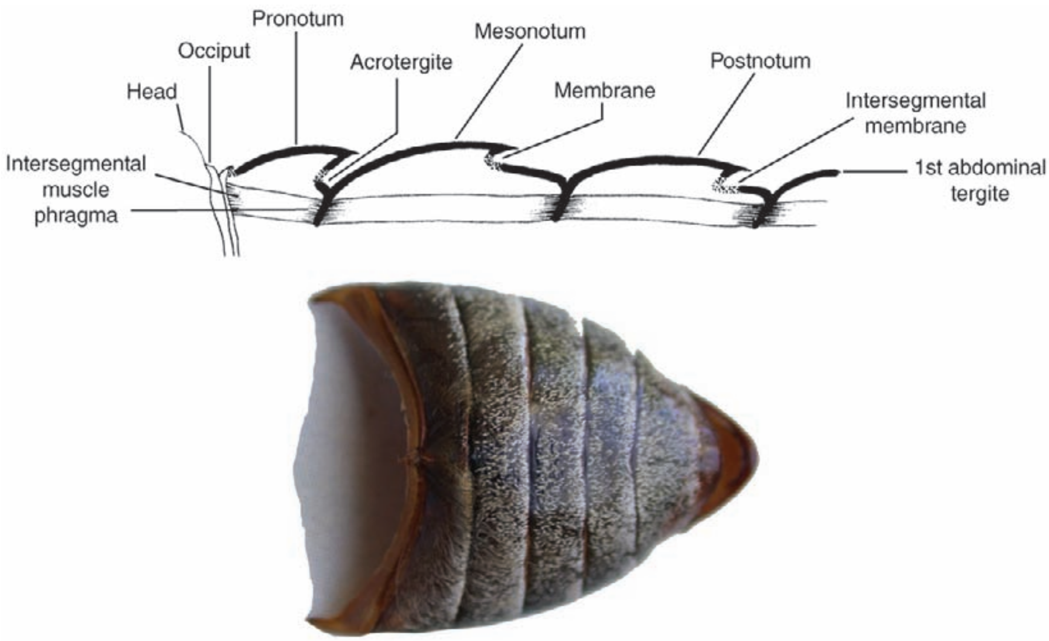

FIGURE 2 Secondary segmentation. Top: diagram of sagittal section of dorsal sclerites of thorax. Bottom: ventral view of abdominal sternites showing overlap due to secondary segmentation (Coleoptera: Scarabaeidae).

derived anatomical feature. The musculature we see in secondary segmentation is *inter*segmental, or between segments (Fig. 2). The acquisition of secondary segmentation represents a major evolutionary step in the development of the Arthropoda. The soft-bodied arthropod has primary segmentation and muscles that are *intra*segmental, or within each segment. Movement of the body and its parts is relatively simple because the body wall is flexible. However, when the body wall becomes hardened, flexibility is restricted to the articulation between hardened parts or the extension provided by intersegmental membranes. The arthropod is, in a metaphorical sense, clad in a suit of armor; most movement is possible only if soft and flexible membranes are positioned between inflexible (hardened) body parts. Exceptions may be seen in the indirect flight mechanism of pterygote insects.

In all probability secondary segmentation evolved many times, and it probably continues to evolve in response to specific problems confronting insects today. Secondary segmentation is most evident and most readily appreciated in the insect abdomen. It is less apparent in the thorax and almost totally obscured in the head.

Sclerites

The hardening of the body wall contributes significantly to the external features observed in insects. Sclerites are hardened areas of the insect body wall that are consequences of the process of sclerotization. Sclerites, also called "plates," are variable in size and shape. Sclerites do not define anatomical areas and do not reflect a common plan of segmentation. Sclerites develop as *de novo* hardening of membranous areas of the body wall, as *de novo* separations from larger sclerotized areas of the body, and in other ways.

The hardened insect body displays many superficial and internal features that are a consequence of hardening. Understanding the distinction between these conditions and the terms applied to them is critical in understanding insect anatomy and its application in taxonomic identification. These features are of three types. First, sutures (Latin, *sutura* = seam), in the traditional sense of vertebrate anatomists, provide seams that are produced by the union of adjacent sclerotized parts of the body wall. On the insect body, sutures appear as etchings on the surface of the body and form lines of contact between sclerites. Second, sulci (Latin, *sulcus* = furrow) represent any externally visible line formed by the inflection of cuticle. Biomechanically, a sulcus forms a strengthening ridge. In contrast, lines of weakness are cuticular features that are used at molting. Lines of weakness are frequently named as if they were sutures, but they should not be viewed as such. For instance, the ecdysial cleavage line is a line of weakness that is sometimes considered to be synonymous with the epicranial suture. The two features are similar in position and appearance, but structurally they may have been derived from different conditions. Finally, apodemes (Greek, *apo* = away; *demas* = body) are hardened cuticular inflections of the body wall that are usually marked externally by a groove or pit. Structures called apophyses (Greek, *apo* = away; *phyein* = to bring forth) are armlike apodemes. Apodemes have been defined as a hollow invagination or inflection of the cuticle and an apophysis as a solid invagination. Functionally, apodemes strengthen the body wall and serve as a surface for muscle attachment.

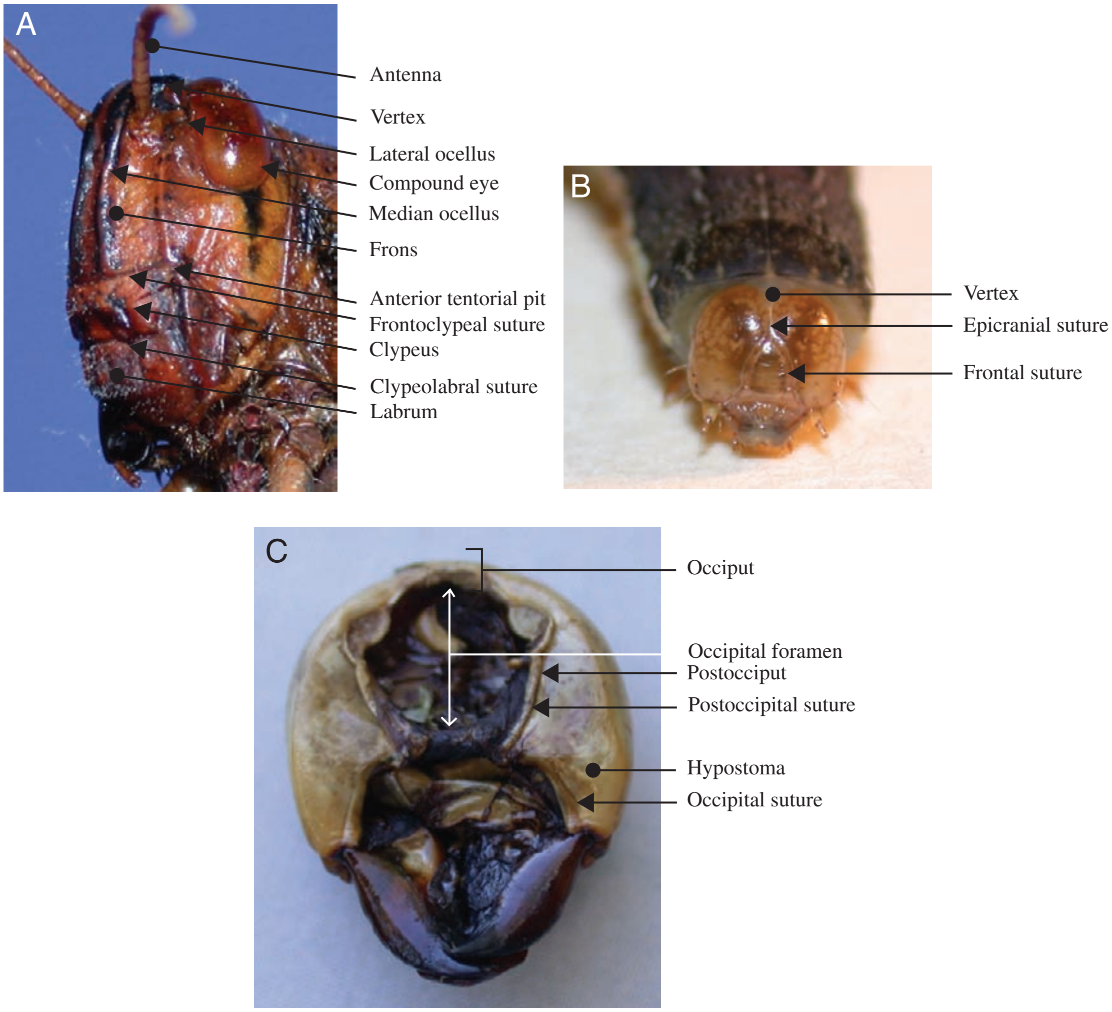
A
Antenna
Vertex
Lateral ocellus
Compound eye
Median ocellus
Frons
Anterior tentorial pit
Frontoclypeal suture
Clypeus
Clypeolabral suture
Labrum
B
Vertex
Epicranial suture
Frontal suture
C
Occiput
Occipital foramen
Postocciput
Postoccipital suture
Hypostoma
Occipital suture

FIGURE 3 (A) Anterior view of the head of a grasshopper (Orthoptera: Acrididae). (B) Larval pterygote head showing epicranial and frontal sutures (Lepidoptera: Noctuidae). (C) Posterior aspect of the head (Orthoptera: Stenopelmatidae).

Sclerites receive different names depending upon the region of the body they are located. Tergites (Latin, *tergum* = back) are sclerites that form a subdivision of the dorsal part of the body wall (tergum). Latrotergites are sclerites that form as a subdivision of the lateral portion of the tergum. Sternites (Latin, *sternum* = breast bone) are sclerites that form as a subdivision of the ventral part of the body wall (sternum), or any of the sclerotic components of the definitive sternum. Pleurites (Greek, *pleura* = side) are sclerites in the pleural region of the body wall that are derived from limb bases.

HEAD

The head is a controversial area for anatomical nomenclature, but it provides some of the best examples of evolutionary trends in anatomy. Most insect morphologists believe that the head of modern insects represents the fusion of several segments that were present in an ancestral condition. However, the number of segments included in the ground plan of the insect head has been a contentious issue among morphologists for more than a century. Any argument that attempts to explain head segmentation must take into account comparative anatomical, embryological, and paleontological evidence, and must examine modern forms of ancestral insects.

Ground Plan of the Pterygote Head

Given the difficulty in homologizing anatomical features of the head, we describe regions associated with landmarks of a ground plan or an idealized hypognathous insect head. In terms of modern insects, the Orthoptera probably come closest to displaying all the important landmark sutures and sclerites that form the head (Fig. 3A).

The vertex (Latin, *vertex* = top; pl., vertices) is the apex or dorsal region of the head between the compound eyes for insects with a hypognathous or opisthognathous head. This definition does not apply to prognathous heads because the primary axis of the head has rotated 90° to become parallel to the primary axis of the body. The vertex is the area in which ocelli are usually located. In some insects this region has become modified or assumes different names.

The ecdysial suture (coronal suture + frontal suture, epicranial suture, ecdysial line, cleavage line) is variably developed among insects. The suture is longitudinal on the vertex and separates epicranial halves of the head (Fig. 3B). Depending on the insect, the ecdysial suture may be shaped like a Y, a U, or a V. The arms of the ecdysial suture that diverge anteroventrally, called the frontal sutures (frontogenal sutures), are not present in all insects (Fig. 3B). Some of these complexes of sutures are used by insects to emerge from the old integument during molting.

The frons is that part of the head immediately ventrad of the vertex (Fig. 3A). The frons varies in size, and its borders are sometimes difficult to establish. In most insects the frons is limited ventrally by the frontoclypeal suture (epistomal suture), a transverse suture located below the antennal sockets. As its name implies, the suture separates the dorsal frons from the ventral clypeus (Fig. 3A).

The face is a generalized term used to describe the anteromedial portion of the head bounded dorsally by the insertion of the antennae, laterally by the medial margins of the compound eyes, and ventrally by the frontoclypeal suture. In some insects the area termed the face is coincident with some, most, or all of the frons.

The clypeus (Latin, shield) is a sclerite between the face and labrum (Fig. 3A). Dorsally, the clypeus is separated from the face or frons by the frontoclypeal suture in primitive insects. Laterally, the clypeogenal suture demarcates the clypeus. Ventrally, the clypeus is separated from the labrum by the clypeolabral suture (Fig. 3A). The clypeus is highly variable in size and shape. Among insects with sucking mouthparts the clypeus is large.

The gena (Latin, cheek; pl., genae) forms the cheek or sclerotized area on each side of the head below the compound eye and extending to the gular suture (Fig. 3). The size of the gena varies considerably, and its boundaries also often are difficult to establish. In Odonata the gena is the area between compound eye, clypeus, and mouthparts. The postgena (Latin, *post* = after; *gena* = cheek; pl., postgenae) is the portion of the head immediately posteriad of the gena of pterygote insects and forms the lateral and ventral parts of the occipital arch (*sensu* Snodgrass) (Fig. 3). The subgenual area is usually narrow, located above the gnathal appendages (mandible and maxillae), and includes the hypostoma (Figs. 3 and 4) and the pleurostoma. The pleurostoma is the sclerotized area between the anterior attachment of the mandible and the ventral portion of the compound eye. The hypostoma is posteriad of the pleurostoma between the posterior attachment of the mandible and the occipital foramen. The subgenal suture forms a lateral, submarginal groove or sulcus on the head, just above the bases of the gnathal appendages (Fig. 4). The subgenal suture is continuous anteriorly with the frontoclypeal suture in the generalized pterygote head. Internally, the subgenal suture forms a subgenal ridge that presumably provides structural support for the head above the mandible and maxillae. In some instances, the subgenal suture is descriptively divided in two. The part of the suture that borders the proximal attachment of the mandible to the head (Fig. 4) is called the pleurostomal suture (the ventral border of the pleurostoma). The posterior part of the subgenal suture from the mandible to the occipital foramen is called the hypostomal suture (the ventral border of the hypostoma).

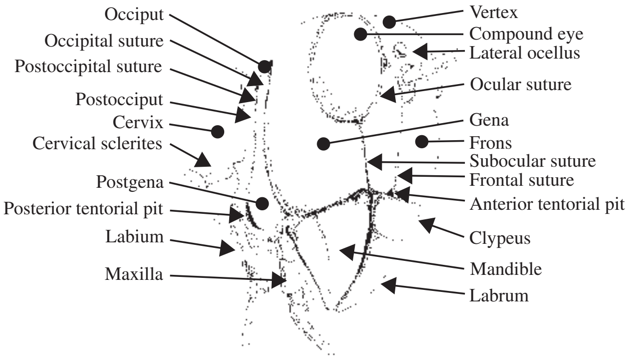

FIGURE 4 Generalized view of an insect head. [This and other line drawings after Snodgrass, R. E. (1935). "Principles of Insect Morphology," McGraw-Hill Co.].

Head Size and Shape

The size and shape of the head and its appendages reflect functional adaptations that can be used to explain biological details of the insect—the realm of morphology as opposed to anatomy.

SIZE Upon casual observation, the size of any given insect's head appears to be in proportion to the size of its body. A head that is disproportionately small or large relative to body size suggests that some adaptation has taken place that serves a functional need. Proportional head size varies considerably in the Insecta. Some fly families have very tiny heads in relation to their body size (e.g., Diptera: Acroceridae). Among Orthoptera, grass-feeding species typically have larger heads than herbaceous-feeding species. The large head is filled with powerful adductor muscles because grasses (monocots) are more difficult to chew than dicotyledonous plants. Furthermore, the postseedling stages of grasses are nutrient poor, meaning that more grass must be bitten, chopped, or ground to provide adequate nutrition.

SHAPE Head shape varies considerably among insects. Many unusual shapes seem to be influenced by behavior and may be used to illustrate examples of structural form and function. The functional importance of head shape may be difficult to determine in preserved specimens. A few hours of observation with living insects can provide considerable insight into the importance of shape. Globular heads are seen in some insects, including the burrowing crickets (e.g., stenopelmatines and gryllids). This form of head is adapted for pushing soil. Hypercephalic heads are seen in the males of some Diptera (Sepsidae, Diopsidae, Drosophilidae, and Tephritoidea) and Hymenoptera (Pteromalidae and Eurytomidae); the broad heads of the males are featured in various aspects of courtship behaviors.

Topographical Features of the Head

Morphologists experience considerable difficulty in defining regions and determining homologies of structure on the insect head. We cannot unambiguously characterize topographical features of the insect head because more than a million species are involved in the definition, and they show incredible diversity in head anatomy. Shape alone is not adequate or suitable because there are many head shapes, and often a head shape can be derived independently in several unrelated lineages. Some head shapes are influenced by behavior.

AXIAL POSITION The posture or orientation of the head in its resting position relative to the long axis of the body can be important in providing definitions of the anatomical features of the head. Axial position in insects typically falls into three basic categories: hypognathous, prognathous, and opisthognathous.

In general zoological usage, the word "hypognathous" (Greek, *hypo* = under; *gnathos* = jaw) serves to designate animals whose lower jaw is slightly longer than the upper jaw. In entomological usage, "hypognathous" refers to insects with the head vertically oriented and the mouth directed ventrad. Most insects with a hypognathous condition display an occipital foramen near the center of the posterior surface of the head. The hypognathous condition is considered by most insect morphologists to represent the primitive or generalized condition. The hypognathous position is evident in most major groups of insects and can be seen in the grasshopper, house fly, and honey bee. Other conditions are probably derived from ancestors with a hypognathous head.

In general zoological usage "prognathous" (Greek, *pro* = forward; *gnathos* = jaw) refers to animals with prominent or projecting jaws. In entomological usage, the prognathous condition is characterized by an occipital foramen near the vertexal margin with mandibles directed anteriad and positioned at the anterior margin of the head. When viewed in lateral aspect, the primary axis of the head is horizontal. Some predaceous insects, such as carabid beetles and earwigs, display the prognathous condition. In other insects, such as cucujid beetles and bethylid wasps, the prognathous position may reveal a solution to problems associated with living in concealed situations such as between bark and wood or similar confined habitats.

In general zoological usage, "opisthognathous" (Greek, *opisthos* = behind; *gnathos* = jaw) refers to animals with retreating jaws. In entomological usage, the opisthognathous condition is characterized by posteroventral position of the mouthparts resulting from a deflection of the facial region. The opisthognathous condition is displayed in many fluid-feeding Homoptera, including leafhoppers, whiteflies, and aphids.

SUTURES OF THE HEAD Head sutures are sometimes used to delimit specific areas of the head, but there are problems. Establishing homology of sutures between families and orders is difficult. From a practical viewpoint, standards have not been developed for naming sutures among insect groups. Some names are based on the areas delimited (e.g., frontoclypeal suture); other sutures are named for the areas in which the suture is found (e.g., coronal suture). Sutures frequently have more than one name (e.g., frontoclypeal suture and epistomal suture are synonymous).

The compound eye is an important landmark on the insect head. An ocular suture surrounds the compound eye and forms an inflection or an internal ridge of the integument (Figs. 3, 4). The ocular suture is not present in all insects and is difficult to see in some insects unless the head is chemically processed for microscopic examination. When present, the ocular suture probably provides strength and prevents deformation of the compound eye.

A subocular suture extends from the lower margin of the compound eye toward the subgenal suture. In some species the subocular suture (Fig. 4). may extend to the subgenal suture; in other species it may terminate before reaching another landmark. This suture is straight and commonly found in the Hymenoptera, where it may provide additional strength for the head.

POSTERIOR ASPECT OF THE HEAD The entire posterior surface of the head is termed the postcranium (Fig. 3). The surface may be flat, concave, or convex, depending on the group of insects. The occiput (Latin, back of head) of pterygote insects is the posterior portion of the head between the vertex and cervix (Latin, neck). The occiput is rarely present as a distinct sclerite or clearly demarcated by "benchmark" sutures. When present, the occiput signifies a primitive head segment. In some Diptera the occiput forms the entire posterior surface of the head. In other insects it forms a narrow, horseshoe-shaped sclerite.

The occipital suture (hypostomal suture *sensu* MacGillivray) is well developed in orthopteroids, but it is not present in many other groups of pterygote insects (Fig. 3C). When present, the occipital suture forms an arched, horseshoe-shaped groove on the back of the head that ends ventrally, anterior to the posterior articulation of each mandible. Internally, the occipital suture develops into a ridge, providing strength for the head.

The postoccipital suture is a landmark on the posterior surface of the head and is typically near the occipital foramen (Fig. 3C). The postoccipital suture forms a posterior submarginal groove of the head with posterior tentorial pits marking its lower ends on either side of the head. Some morphologists regard this suture as an intersegmental boundary (labium) between the first and second maxillae.

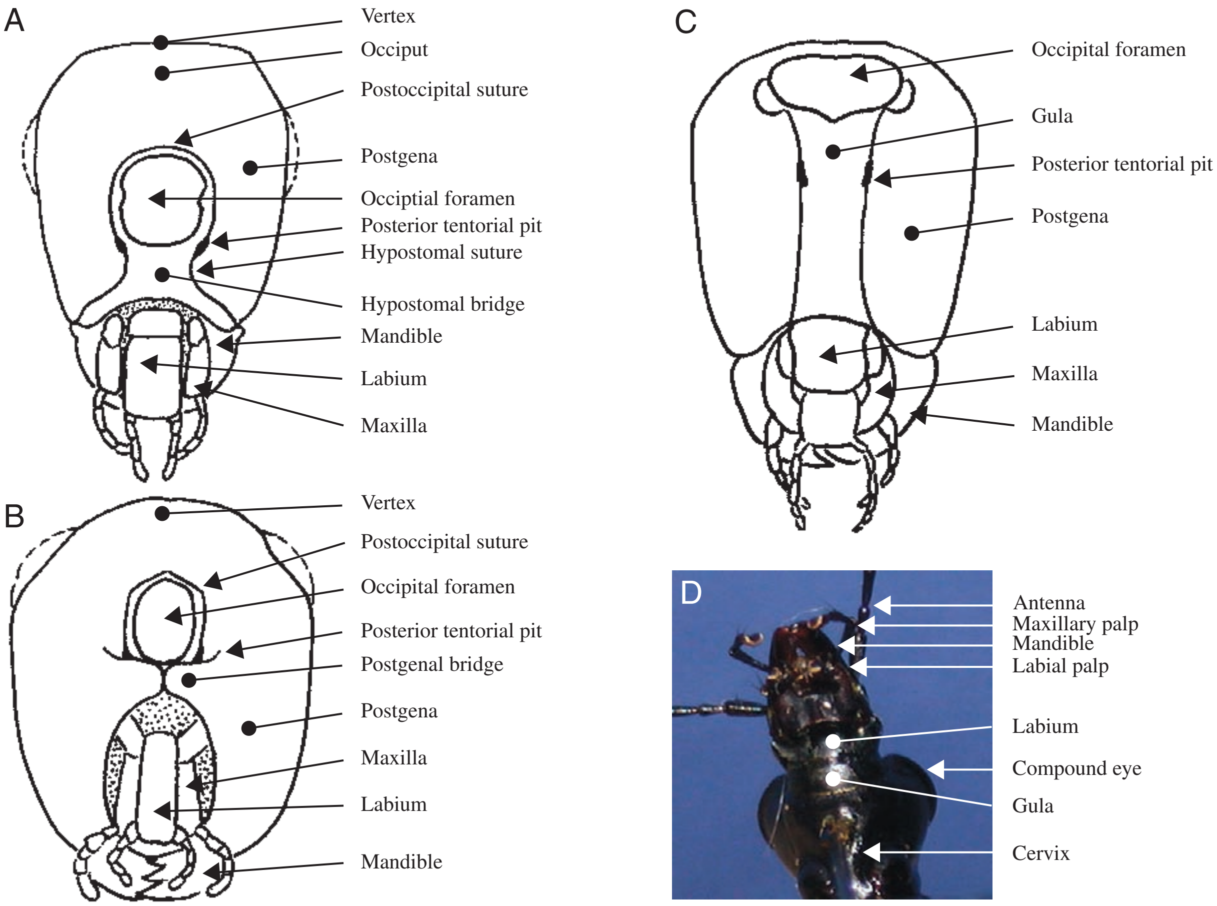
A
Vertex
Occiput
Postoccipital suture
Postgena
Occiptial foramen
Posterior tentorial pit
Hypostomal suture
Hypostomal bridge
Mandible
Labium
Maxilla
B
Vertex
Postoccipital suture
Occipital foramen
Posterior tentorial pit
Postgenal bridge
Postgena
Maxilla
Labium
Mandible
C
Occipital foramen
Gula
Posterior tentorial pit
Postgena
Labium
Maxilla
Mandible
D
Antenna
Maxillary palp
Mandible
Labial palp
Labium
Compound eye
Gula
Cervix

FIGURE 5 Occipital closures (diagrammatic): (A) hypostomal bridge, (B) postgenal bridge, (C) gula and (D) ventral view of gula of adult ground beetle (Coleoptera: Carabidae).

Internally, the postoccipital suture forms the postoccipital ridge that serves as an attachment for the dorsal prothoracic and cervical muscles of the head. The absence of the postoccipital suture in pterygote insects is a derived condition.

The postocciput of pterygotes forms the extreme posterior, often U-shaped sclerite that forms the rim of the head behind the postoccipital suture. The postocciput is interpreted as a sclerotic remnant of the labial somite in ancestral insects.

In pterygotes such as Orthoptera the occipital foramen and the mouth are not separated. More highly evolved insects have developed sclerotized separations between the mouthparts and the occipital foramen. At least three types of closure have been identified (Fig. 5): the hypostomal bridge, the postgenal bridge, and the gula. An understanding of these structures provides insight into the operation of the head and suggests evolutionary trends in feeding strategies.

The hypostomal bridge is usually developed in adult heads displaying a hypognathous axial orientation. The bridge is formed by medial extension and fusion of hypostomal lobes (hypostoma) (Fig. 5A). The hypostomal bridge is the ground plan condition of closure for the posterior aspect of the head, but it is not restricted to primitive insects. The hypostomal bridge is found in highly developed members of the Heteroptera, Diptera, and Hymenoptera. In Diptera the hypostomal bridge also has been called the pseudogula.

The postgenal bridge is a derived condition from the hypostomal ground plan and is developed in adults of higher Diptera and aculeate Hymenoptera. The bridge is characterized by medial extension and fusion of the postgenae, following a union of the hypostoma (Fig. 5B). The posterior tentorial pits retain their placement in the postoccipital suture.

The gula (Latin, gullet; pl., gulae) is developed in some Coleoptera, Neuroptera, and Isoptera. Typically, the gula is developed in heads displaying a prognathous axial orientation and in which posterior tentorial pits are located anteriad of the occipital foramen. (Fig. 5C, D). The median sclerite (the gula) on the ventral part of a prognathous head apparently forms *de novo* in the membranous neck region between the lateral extensions of the postocciput. The gula is a derived condition that is found in some but not all prognathous heads.

Endoskeletal Head Framework

Although the hardened integument of the head forms a structurally rigid capsule, this design is insufficient to solve the problems associated with muscle attachment and maintaining structural integrity during chewing. Thus, insects have evolved a tentorium (Latin, tent; pl., tentoria): a complex network of internal, hardened, cuticular struts that serve to reinforce the head. The tentorium forms as an invagination of four apodemal arms from the integument in most pterygote insects. The tentorium strengthens the head for chewing, provides attachment points for muscles, and also supports and protects the brain and foregut.

Anatomically, the tentorium consists of anterior and posterior arms. In most insects, the anterior arms arise from facial inflections located just above the anterior articulations of the mandibles. Externally, the arms are marked by anterior tentorial pits positioned on the frontoclypeal or subgenal (pleurostomal) suture (Fig. 4). Internally, the anterior region may form a frontal plate. Posterior arms originate at the ventral ends of the postoccipital inflection. They are marked externally by posterior tentorial pits (Fig. 4). The posterior arms usually unite to form a transverse bridge or corpus tentorii (internally) across the back of the head. Dorsal arms (rami), found in many insects, arise from the anterior arms. They attach to the inner wall of the head near antennal sockets. The dorsal arms are apparently not an invagination of cuticle, because pits do not mark them externally.

Mandible Articulation and Musculature

The hypothetical ancestor of insects is thought to possess a mandible with one point of articulation. Later, insects acquired a second point of articulation. The basis of this assumption comes from a survey of the Hexapoda in that the modern Apterygota have a monocondylic mandible and the Pterygota have a dicondylic mandible.

The term condyle (Greek, *kondylos* = knuckle) refers specifically to a knoblike process. For the mandible, the condyle is the point of articulation with the head. On the head itself is an acetabulum (Latin, *acetabulum* = vinegar cup), a concave surface or cavity for the reception and articulation of the condyle (Fig. 6).

The dicondylic mandible is the derived condition and is found in the Lepismatidae and Pterygota. The dicondylic mandible has secondarily acquired an articulation point anterior to the first point in the monocondylic mandible. These attachments form a plane of attachment. In the monocondylic mandible there is no plane of attachment, and the mandibles move forward or rearward when the muscles contract. The two points of articulation create a plane of movement that restricts the direction of mandible movement.

THORAX

The thorax represents the second tagma of the insect body. The thorax evolved early in the phylogenetic history of insects. In most Paleozoic insects the thorax is well developed and differentiated from the head and abdomen, and the three distinct tagmata probably developed during the Devonian.

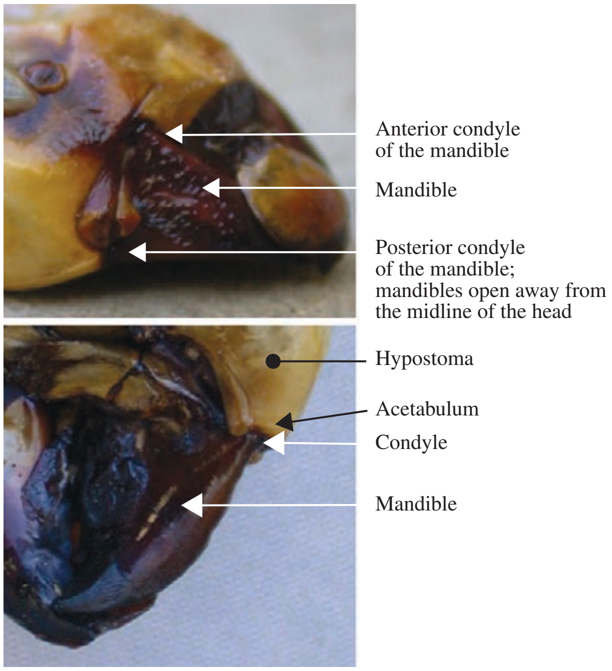
Anterior condyle of the mandible
Mandible
Posterior condyle of the mandible; mandibles open away from the midline of the head
Hypostoma
Acetabulum
Condyle
Mandible

FIGURE 6 Mandible articulation. Top: lateral view; bottom: posterior view (Orthoptera: Stenoplematidae).

In terms of insect phylogeny, the thorax of Apterygota is strikingly different in shape compared with the head or abdomen. Of modern apterygotes only the Collembola display taxa in which thoracic tagmatization and segmentation are not obvious.

Apparently, the primary, functional role of the thorax has always been locomotion, since the primary modifications of the thorax have been for locomotion (first walking, and then flight). Modification for locomotion probably developed before other morphological adaptations, such as metamorphosis. Diverse independent and interdependent mechanisms for locomotion have evolved throughout the Insecta, including walking, flight, and jumping. Active participation in flight by insects is unique among invertebrates.

Anatomy of the Thorax

The cervix is the connection between the head (occipital foramen) and the anteriormost part of the thorax (pronotum) (Fig. 2). Typically, the area between the head and pronotum is membranous. The ground plan for the insect cervix contains two cervical sclerites on each side of the head that articulate with an occipital condyle of the head and the prothoracic episternum. Musculature attached to these sclerites increases or decreases the angle between the sclerites, and creates limited mobility of the head.

The thorax of modern insects consists of three segments termed the prothorax, mesothorax, and metathorax. The last two collectively are called the pterothorax (Greek, *ptero* = wing or feather) because extant insects bear wings on these segments only. The individual dorsal sclerites or terga of the thoracic segments are also known as nota (Greek, *notos* = back; sing, notum). The nota of Apterygota and many immature insects are similar to the terga of the abdomen with typical secondary segmentation. The nota of each thoracic segment are serially distinguished as the pronotum, mesonotum, and metanotum.

The size and shape of the prothorax are highly variable. The prothorax may be a large plate as in Orthoptera, Hemiptera, and Coleoptera, or reduced in size forming a narrow band between the head and mesothorax as in Hymenoptera. The prothorax is usually separated or free from the mesothorax. The sclerites are separated by a membrane that may be large and conspicuous in more primitive holometabolous insects such as Neuroptera and Coleoptera, or reduced in size in more highly evolved holometabolous insects such as Diptera and Hymenoptera.

The pterothorax includes the thoracic segments immediately posteriad of the prothorax. In winged insects the relationship between thoracic segments involved in flight can be complicated. In contrast, the thorax of larval insects and most wingless insects is relatively simple. The mesothorax and metathorax of these insects are separated by membrane. Adult winged insects show a mesothorax and metathorax that are consolidated (i.e., more or less united) to form a functional unit modified for flight.

The development of the pterothoracic segments varies among winged insects. When both pairs of wings participate equally in flight, the two thoracic segments are about the same size. This condition is seen in Odonata, some Lepidoptera, and some Neuroptera. When one pair of wings is dominant in flight, the corresponding thoracic segment is commensurately larger and modified for flight, whereas the other thoracic segment is reduced in size. This condition is seen in Diptera and Hymenoptera, where the forewing is large and dominant in flight. The reverse condition is seen in the Coleoptera, where the hind wing is large and dominant in flight.

In more closely related insect groups, such as families within an order, that are primitively wingless or in which wings have been secondarily lost in modern or extant species, many modifications to the thorax occur. Many wingless forms can be attributed to environmental factors that promote or maintain flightlessness. For instance, island-dwelling insects are commonly short winged (brachypterous), or wingless, whereas their continental relatives are winged, presumably because for island species, flight increases the likelihood of being carried aloft, moved out to sea, and subsequently lost to the reproductive effort of the population. The anatomical consequences of flightlessness can be predictable; in the Hymenoptera, short wings bring a disproportionate enlargement of the pronotum and reduction in size of the mesonotum and metanotum.

Sutures and Sclerites of Wing-Bearing Segments

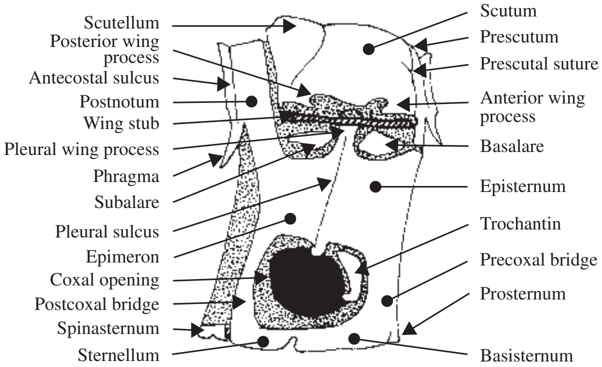

FIGURE 7 Diagram of the pterygote pterothorax.

The wing-bearing segments of the thorax are subdivided into a myriad of sclerites that are bounded by sutures and membranous areas. These sutures and sclerites are the product of repeated modification of the thorax in response to various demands placed on the insect body by the environment. Similar modifications have occurred independently in many groups of insects; some modifications are unique. Generalizations are difficult to make, given the large number of sutures and sclerites, coupled with the number of insects that there are to consider.

Dorsal Aspect

The nota of the pterothorax are further subdivided into the prescutum, scutum, and scutellum; again, serially distinguished as mesoscutum and mesoscutellum, and metascutum and metascutellum (Fig. 7). Additionally, there are sclerites anterior and posterior to the notum, as discussed shortly.

The prescutum is the anterior portion of the scutum, laterally bearing prealar bridges separated by the prescutal suture from the mesoscutum. The scutum is the largest dorsal sclerite of the notum and is bounded posteriorly by the scutoscutellar suture, which divides the notum into the scutum and scutellum. The scutellum is generally smaller than the scutum. In Heteroptera it is a small triangular sclerite between the bases of the hemelytra. In Coleoptera the scutellum is the small triangular sclerite between the bases of the elytra. In Diptera and Hymenoptera the scutellum is relatively large, forming a subhemispherical sclerite, sometimes projecting posteriad. The posteriormost sclerite of the notum is the postnotum, separated from the notum by secondary segmentation. In some insects there is a postscutellum (metanotal acrotergite) that forms the posteriormost thoracic sclerite of the metanotum, or the posteriormost sclerite of the thorax. In Diptera the postscutellum appears as a transverse bulge below the scutellum.

The acrotergite and postnotum deserve further explanation. Again, the anteriormost sclerite is an acrotergite, the anterior precostal part of the notal plate. The postnotum is an intersegmental sclerite associated with the notum of the preceding segment. The postnotum bears the antecosta, a marginal ridge on the inner surface of the notal sclerite corresponding to the primary intersegmental fold. The postnotum also usually bears a pair of internal projecting phragmata. The antecostal suture divides the acrotergite from the antecosta, the internal ridge marking the original intersegmental boundary. Thus, when the antecosta and acrotergite are developed into larger plates and are associated with the notum anterior to them, they are referred to as a postnotum. The final structure associated with the dorsal aspect of the pterothorax is the alinotum (Greek, *ala* = wing; *notos* = back; pl., alinota). The alinotum is the wing-bearing sclerite of the pterothorax.

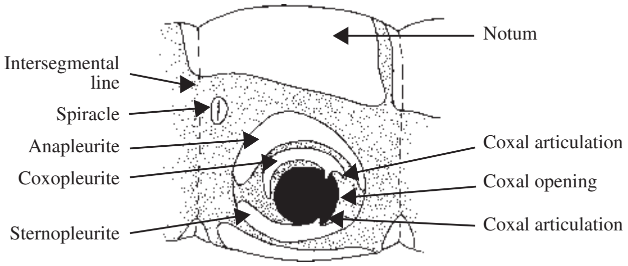

FIGURE 8 Pleural aspect of the apterygote thorax: diagrammatic.

Wing Articulation

The thoracic components necessary for wing movement include the prealar bridge, anterior notal wing process, and posterior notal wing process. The components of the wing itself that articulate with the thoracic components are the humeral and axillary sclerites; they form the part of the wing closest to the body and are not treated in this article.

The prealar bridge is a heavily sclerotized and rigid supporting sclerite between the unsclerotized membrane of the pterothorax and the pleuron; it supports the notum above the thoracic pleura. The prealar bridge is composed of cuticular extensions from the anterior part of the prescutum and antecosta. The anterior notal wing process is the anterior lobe of the lateral margin of the alinotum supporting the first axillary sclerite (Fig. 7). The posterior notal wing process is a posterior lobe of the lateral margin of the alinotum that supports the third axillary sclerite of the wing base (Fig. 7).

Lateral Aspect

The pleuron (Greek = side; pl., pleura) is a general term associated with the lateral aspect of the thorax. Adults, nymphs, and active larvae all display extensive sclerotization of the pleural area. Sclerites forming this part of the body wall are derived from the precoxa, subscoxa, or supracoxal arch of the subcoxa.

PLEURAL REGIONS OF THE THORAX

Apterygota and Immature Plecoptera The anapleurite is the sclerotized area above the coxa (supracoxal area) (Fig. 8). The coxopleurite is a sclerotized plate situated between the coxa and the anapleurite (Fig. 8). It bears the dorsal coxal articulation, the anterior part of which becomes the definitive trochantin. The sternopleurite, or coxosternite, is the definitive sternal sclerite that includes the areas of the limb bases and is situated beneath the coxa (Fig. 8).

Pterygota The basalare is a sclerite near the base of the wing and anterior to the pleural wing process (Fig. 9). The basalare serves as a place of insertion for the anterior pleural muscle of the wing. The subalare is posterior to the basalare and the pleural wing process (Fig. 9). It too serves as a place for insertion of the wing's posterior pleural muscle. The tegula is the anterior most independent sclerite associated with the wing base. The tegula is typically scalelike, articulates with the humeral sclerite, and protects the wing base from physical damage. The tegula is absent from Coleoptera and from the metathorax of most orders. The pleural wing process is located at the dorsal end of the pleural ridge and serves as a fulcrum for the movement of the wing (Fig. 9). The parapteron is a small sclerite, articulated on the dorsal extremity of the episternum just below the wings (Fig. 7).

The pleural suture is an easily visible landmark on the pterothoracic pleura (Fig. 9). It extends from the base of the wing to the base of the coxa. The pleural ridge is formed internally by the pleural suture and braces the pleuron above the leg. The episternum is a pleural sclerite anterior to the pleural suture and sometimes adjacent to the coxa (Fig. 9); the episternum is typically the largest lateral thoracic sclerite between the sternum and the notum. The epimeron is the posterior division of a thoracic pleuron adjacent to the coxa and posterior to the pleural suture (Fig. 9); it is typically smaller than the episternum and narrow or triangular. The episternum and the epimeron of many insects have become subdivided into several secondary sclerites bounded by sutures. The simplest condition shows the episternum divided into a dorsal anepisternum and a ventral katepisternum (Fig. 9). Similarly, the epimeron is divided into an anepimeron and katepimeron. The trochantin is a small sclerite at the base of the insect leg of some insects (Figs. 7, 9). Some workers theorize that the trochantin may have developed into the pleural wall. The trochantin is often fused to the episternum or absent.

The precoxal bridge is anterior to the trochantin, usually continuous with the episternum, frequently united with the

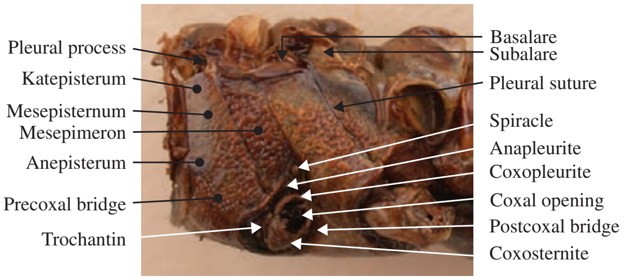

FIGURE 9 Lateral aspect of the pterygote thorax (Orthoptera: Acrididae).

basisternum, but also occurs as a distinct sclerite (Fig. 9). The postcoxal bridge is the postcoxal part of the pleuron, often united with the sternum behind the coxa (Fig. 9). The sclerite extends behind the coxa and connects the epimeron with the furcasternum. The meron is a lateral, postarticular basal area of the coxa and is sometimes found disassociated from the coxa and incorporated into the pleuron. The meron is typically large and conspicuous in panorpid and neuropteran insects. In Diptera the meron forms a separate sclerite in the thoracic pleuron.

Ventral Aspect

The ground plan of the sternum (Greek, *sternon* = chest; pl., sterna) consists of four sclerites, including an intersternite (spinasternite), two laterosternites (coxosternites), and a mediosternite (Fig. 10). The mediosternite and the laterosternite meet and join, and the line of union is called the laterosternal sulcus (pleurosternal suture) (Fig. 10).

Paired furcal pits are found in the laterosternal sulcus (Fig. 10). A transverse sternacostal sulcus bisects the ventral plate and thereby forms an anterior basisternite and posterior furcasternite (Fig. 10). The basisternite (basisternum) is the primary sclerite of the sternum (Fig. 10). It is positioned anterior to the sternal apophyses or sternacostal suture and laterally connected with the pleural region of the precoxal bridge. The furcasternite (furcasternum) is a distinct part of the sternum in some insects bearing the furca (Fig. 10). The spinasternum is a “spine-bearing” intersegmental sclerite of the thoracic venter, associated or united with the preceding sternum. The spinasternum may become part of the definitive prosternum or mesosternum, but not of the metasternum. The sternellum is the second sclerite of the ventral part of each thoracic segment, frequently divided into longitudinal parts that may be widely separated (Figs. 7, 10).

ABDOMEN

The abdomen is more conspicuously segmented than either the head or the thorax. Superficially, the abdomen is the least specialized of the body tagma, but there are notable exceptions such as the scale insects. The abdomen characteristically lacks appendages except cerci, reproductive organs, and pregenital appendages in adult Apterygota and larval Pterygota.

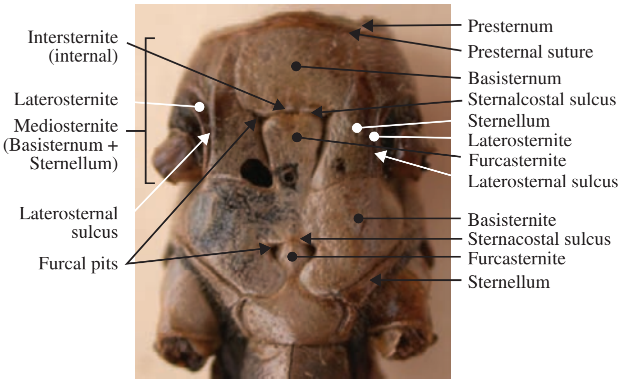

FIGURE 10 Ventral aspect of the thorax (Orthoptera: Acrididae).

Ground Plan of the Abdomen

The ground plan abdomen of an adult insect typically consists of 11 to 12 segments and is less strongly sclerotized than the head or thorax (Fig. 11). Each segment of the abdomen is represented by a sclerotized tergum, sternum, and perhaps a pleurite. Terga are separated from each other and from the adjacent sterna or pleura by a membrane. Spiracles are located in the pleural area. Modification of this ground plan includes the fusion of terga or terga and sterna to form continuous dorsal or ventral shields or a conical tube. Some insects bear a sclerite in the pleural area called a laterotergite. Ventral sclerites are sometimes called laterosternites. The spiracles are often situated in the definitive tergum, sternum, laterotergite, or laterosternite.

During the embryonic stage of many insects and the postembryonic stage of primitive insects, 11 abdominal segments are present. In modern insects there is a tendency toward reduction in the number of the abdominal segments, but the primitive number of 11 is maintained during embryogenesis. Variation in abdominal segment number is considerable. If the Apterygota are considered to be indicative of the ground plan for pterygotes, confusion reigns: adult Protura have 12 segments, Collembola have six. The orthopteran family Acrididae has 11 segments, and a fossil specimen of Zoraptera has a 10-segmented abdomen.

Anamorphosis is present among some primitive ancestral hexapods such as the Protura—they emerge from the egg with eight abdominal segments and a terminal telson. Subsequently, three segments are added between the telson and the last abdominal segment with each molt. In contrast, most insects undergo epimorphosis in which the definitive number of segments is present at eclosion. Given the extent of variation in abdominal segmentation, morphologists conventionally discuss the abdomen in terms of pregenital, genital, and postgenital segmentation.

Abdominal Anatomy

Typically, the abdominal terga show secondary segmentation with the posterior part of a segment overlapping the anterior

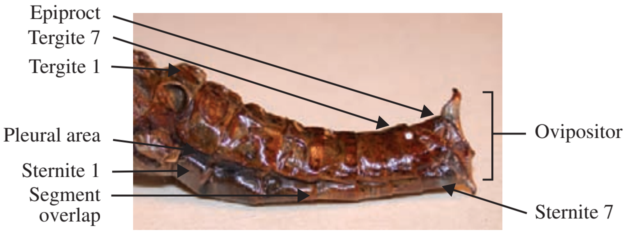

FIGURE 11 Insect abdomen (Orthoptera: Acrididae).

part of the segment behind it (Fig. 11). Such overlap prevents damage or injury to the animal while it moves through the environment, particularly in confined spaces.

The pregenital segments in male insects are numbered 1 through 8; the pregenital segments in female insects are numbered 1 through 7 (Fig. 11). Among the Apterygota, male genitalia in Collembola are positioned between segments 5 and 6 and in Protura between segments 11 and the paraproct. Genital segments of Pterygota include segment 9 in males and segments 8 and 9 in females. Postgenital segments of pterygote insects are 10 and 11 in females and 9 and 10 in males.

In general there is little modification of the pregenital sclerites. A notable exception is found in the Odonata. Male Odonata do not have an intromittent organ on segment 9. Instead, the male moves the abdominal apex forward and deposits sperm in a reservoir along the anterior margin of the third abdominal sternum. Other modifications of the pregenital sclerites are not related to sexual behavior. Some of these modifications are glandular.

Modification of the genital sclerites from the ground plan is frequently observed among insects. Adult Pterygota are characterized by a well-developed reproductive system, including organs of copulation and oviposition. This duality of function has resulted in considerable differentiation of associated segments and contributed to difference of opinion regarding homology of genitalic parts. Among pterygote insects the male genitalia are generally positioned on segment 9. The ninth sternum is called a hypandrium (Greek, *hypo* = beneath; *aner* = male; Latin, *-ium* = diminutive) in many insects, including Psocoptera. In Ephemeroptera, the tenth sternum is called a hypandrium. Fusion of segments 9 and 10 in Psocoptera results in a structure called the clunium (Latin, *clunais* = buttock).

The gonopore (Greek, *gone* = seed; *poros* = channel) of the female reproductive system serves as the aperture through which the egg passes during oviposition. The gonopore usually is located on segment 8 or 9. Enlargement of sternum 8 in some female insects is called a subgenital plate.

Modification of postgenital sclerites is frequently observed and seems to be a functional response to adaptations associated with copulation and oviposition. Some modifications include fusion of the tergum, pleuron, and sternum to form a continuous sclerotized ring. The phenomenon is notable in apterygota and pterygote insects.

The eleventh abdominal segment forms the last true somite of the insect body. Frequently, this segment is found in embryonic stages of primitive insects even when it cannot be observed in postemergent stages. When the eleventh segment is present, it forms a conical endpiece that bears an anus at the apex, flanked laterally by cerci (Greek, *kerkos* = tail) (Fig. 12). The dorsal surface of the eleventh segment is called an epiproct (Greek, *epi* = upon; *proktos* = anus); the ventrolateral surface is called a paraproct (Greek, *para* = beside; *proktos* = anus) (Fig. 12). A longitudinal, medial, membranous area connects the paraprocts ventrally. Primitive groups of extant insects such as Thysanura and Ephemeroptera, and some fossil groups such as Paleodictyoptera, display a conspicuous, long, median filament that apparently projects from the apex of the epiproct. This is called the appendix dorsalis or caudal style. The appendage appears annulated and similar in shape to the lateral cerci, but the function of the appendix is unknown. The twelfth abdominal segment is called the periproct in Crustacea, and it forms a telson in some embryonic insects. The periproct appears in adult Protura and naiadal Odonata.

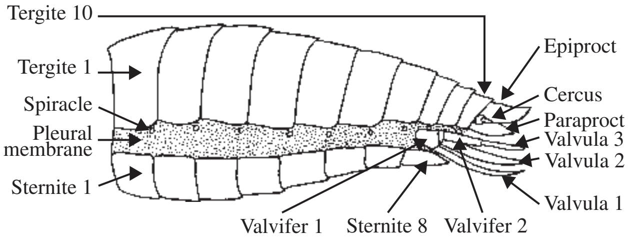

FIGURE 12 Abdominal segmentation: diagrammatic.

Abdominal Appendages

Presumably, the hypothetical ancestor of the Insecta was a myriapod with one pair of appendages for each body segment. Among contemporary insects the head appendages are represented by the antennae, mandibles, and the first and second maxillae. Thorax appendages are represented by legs, whereas the wings are considered to be secondary in origin. In most Apterygota, paired abdominal appendages are apparent. In most true insects embryological appendages are formed and lost before eclosion. The appendages found in embryos apparently represent ancestral conditions that are not expressed in postembryonic stages of modern insects. In modern insects, most pairs of appendages have been lost, and the irregular distribution of the remaining appendages makes a summary evaluation difficult. Abdominal appendages do not resemble the structure of thoracic legs of any insect.

Appendages are common among some entognathous hexapods, and some ancestral forms display unique abdominal appendages. Collembola are highly specialized entognathous Hexapoda. The abdomen of Collembola bear saltatorial appendages, which gives the group its common name of springtail, and a ventral tube, the collophore, which is the basis of the ordinal name.

The collophore (Greek, *kolla* = glue; *pherein* = to bear) is found on the first abdominal segment of Collembola. The collophore forms a ventromedial tube that is eversible with hydrostatic pressure and is drawn inward with retractor muscles. Some morphologists believe the collophore represents the fusion of paired, lateral appendages of an ancestor. An early explanation of the collophore function noted it was an organ of adhesion. The collophore also is used as a grooming organ in some Collembola. The collophore is connected to secretory glands in the head, and the median longitudinal channel on the venter of the thorax extends from the head to the base of the collophore.

OTHER APPENDAGES Protura maintain short, cylindrical appendages on each of the first three abdominal segments. Each of these arises from membranous areas between the posterolateral angles of the terga and sterna. The position suggests a pleural origin.

APPENDAGES OF PTERYGOTA The aquatic neuropteran larva *Sialis* has long, tapering, six-segmented appendages on each of the first seven body segments. These appendages articulate to pleural coxopodites. Similar appendages are found on the abdomen of some aquatic coleopteran larvae.

The tenth abdominal segment is present in most larval and adult Holometabola. As noted earlier, it is sometimes fused with segment 11. Segment 10 displays paired appendicular processes called pygopodia in Trichoptera, Coleoptera, and Lepidoptera. Pygopods form terminal eversible appendages in some beetle larvae. Pygopodia are bilaterally symmetrical, with eight podia, or feet, per side. Control of the podia is apparent because they are not always everted or inverted. Podia are withdrawn into the segment and have a common or median stalk. Each podium has several rows of equally spaced acanthae that apparently serve as holdfasts. Functionally, the acanthae enable the larvae to attach to and move on different substrates. When the larva walks on a flat substrate, the pygopodia are retracted into the body. When the larva walks on the edge of a leaf, the pygopodia are everted and used as holdfasts.

The larval prolegs of terrestrial Lepidoptera and Symphyta are not well developed, but they are adapted to grasping substrates. These structures are considered to be serially homologous with legs, but they also are referred to by some as adaptive structures with no relation to legs.

The adult pterygote abdomen has appendages that are not generally observed. These appendages are grouped for discussion based on the segments of the abdomen on which they are found.

Pregenital appendages are rare among insects. Adult whiteflies have a curious structure on sternum 8 that propels honeydew away from the body. Genitalia are segmental appendages and are treated in the next section. Postgenital appendages include cerci (Latin, circle), which are thought to represent primitive appendages because they are found in the Apterygota (except Protura) and many Pterygota. Cerci originate on abdominal segment 11 in a membranous area between the epiproct and the paraproct (Fig. 12). In insects that have lost segment 11, the cerci appear to originate on segment 10. Cerci occur in all orders among the Hemimetabola except for hemipteroids; among the Holometabola, they are found only in the Mecoptera and Symphyta.

Cerci are highly variable in size and shape and function. They are longer than the body in Thysanura, and in some Orthoptera cerci may be indistinct. Cerci resemble forceps in Japygidae and are annulated in Dictyoptera. In Dictyoptera they detect air currents, are sensitive to sound, and may be chemoreceptive. Some Ephemeroptera use cerci to propel themselves through water. Japygidae and Dermaptera probably use cerci to subdue prey. In some groups such as Embioptera and Orthoptera, cerci are sexually dimorphic and may serve a role in copulation.

There are some features on the insect body that appear as appendages but are not. Urogomphi (Greek, *oura* = tail; *gomphos* = nail; sing., urogomphus) are fixed or mobile cuticular processes on the apical abdominal segment of some coleopteran larvae. They may or may not be homologous with cerci, or other true appendages.

GENITALIA

The examination of the reproductive anatomy of different insect orders helps to develop an appreciation for the evolutionary trends in the formation of the external genitalia. The male genitalia are derived from the ninth abdominal segment. The female genitalia are derived from the eighth and ninth abdominal segments. In the female, the aperture through which the egg passes is called a gonopore. The gonopore serves as a boundary between the external and internal genitalia and is usually independent of the anus. Exceptions include some flies, such as the Tephritidae, where a common lumen termed a cloaca serves for excretion, copulation, and oviposition.

There is usually a single, medially located gonopore. The Dermaptera and Ephemeroptera are ancient groups of hemimetabolous insects. Both orders display a condition in which the lateral oviducts do not combine to form a median oviduct. Instead, the lateral oviducts independently connect with paired gonopores on the conjuctival membrane along the posterior margin of the seventh abdominal segment.

Many female insects with a genitalic opening on the posterior margin of the eighth abdominal segment display an appendicular ovipositor (Fig. 13). The ovipositor is a structure that develops from modified abdominal appendages or segments. It functions in the precise placement of eggs. It is commonly assumed that insects that do not show an ovipositor have ancestors that had an ovipositor. Thus, the structure has been lost during the course of evolutionary adaptation to a particular lifestyle.

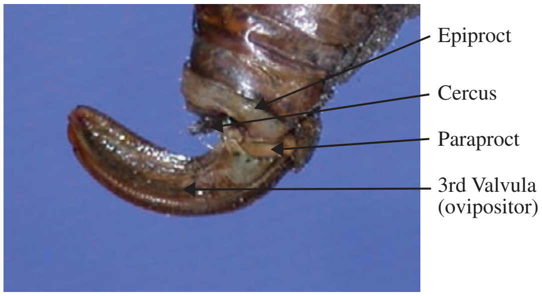

FIGURE 13 Appendicular ovipositor (Orthoptera: Tettigoniidae).

Female insects with a genitalic opening on the posterior margin of the ninth abdominal segment typically display a rudimentary or suppressed appendicular ovipositor. These insects lack special provisions for egg placement, but sometimes they reveal other abdominal modifications intended to facilitate oviposition.

Female Genitalia

Morphologists often use the Thysanura as a starting point for developing a generalized model to explain the evolution of the external reproductive system of pterygote insects. The thysanuran abdomen has basal sclerotized plates called coxopodites on which styli are attached. These plates are serially homologous along the abdomen, and the pregenital plates are regarded as identical with the genital plates. The plates located on segments 8 and 9 are considered to be genital plates. The styli associated with these segments are called gonapophyses. There are four gonapophyses on segments 8 and 9 (i.e., a pair of styli on each segment). The gonapophyses are medially concave and directed rearward. The basal sclerite is called a gonocoxa, and in some Thysanura it may be fused with the style.

The primitive pterygote with a gonopore on segment 8 has an appendicular ovipositor that consists of three components. A basal apparatus corresponds to the basal plate or primitive gonocoxite of the thysanuran abdominal appendage. The second part is the first valvifers (on the eighth sternum), and second valvifers (on the ninth sternum) are responsible for providing support and points of articulation for the tube through which the egg passes (Fig. 14). Interpolated between the first and second valvifers is a small sclerite called a gonangulum, which articulates with the second gonocoxite and tergum 9. The gonangulum is present in Odonata and Grylloblatoidea. It apparently is fused with the first valvifer in Dictyoptera and Orthoptera. In the remaining orders these structures are highly variable.

The shaft of the ovipositor consists of two pairs of elongate, closely appressed sclerites called the first and second valvulae (Fig. 14). The first pair of valvulae is positioned on the eighth abdominal sternum. The second pair of valvulae is located on the ninth abdominal sternum and is dorsal in position. Third valvulae are positioned on the posterior end of the second valvifers. These valvulae usually serve as a sheath for the shaft of the ovipositor (Figs. 13, 14).

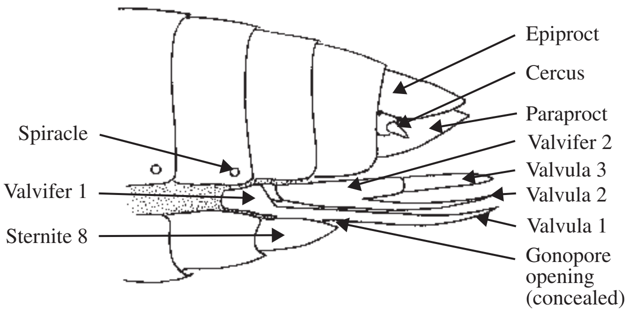

FIGURE 14 Female genitalia (diagrammatic), based on orthopteran female.

Male Genitalia

The primary function of the male genitalia in insects is insemination of the female. Methods of achieving insemination that involve special functions of the external genitalia include clasping and holding the female, retaining the connection with the female gonopore, the construction of spermatophores, and the deposition of spermatophores or semen into the female genital tract; in some insects the injection of semen takes place directly into the female body (traumatic insemination of some Hemiptera). Other functions of the male genitalia include excretion and various sensory functions.

The genitalia of male insects exhibit such an enormous variety of shapes and constituent parts, often further complicated by structural rotation or inversion of all or some of the parts, that determination of a ground plan is virtually impossible. Examination of ancient orders shows highly variable and specialized conditions. In general, the coxites of the eighth segment in most apterygotes are reduced and without gonapophyses, and they are absent altogether in the Pterygota. Thus, the male external genitalia are derived from the ninth abdominal coxites.

Again, the Thysanura have genitalia that closely resemble that of the pterygote orders: a median intromittent organ or phallus, and paired lateral accessories (the periphallus of Snodgrass). The phallus is a conical, tubular structure of variable complexity (Fig. 15). Primitive insects may not display differentiated parts, and the entire structure may be long, sclerotized, and tapering apicad. In a ground plan

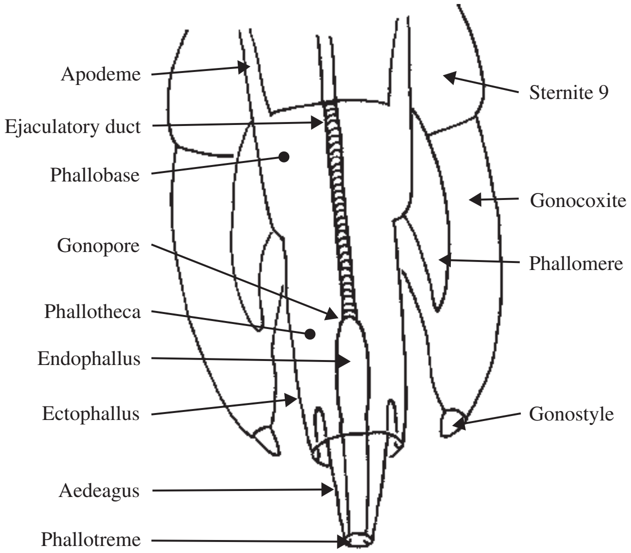

FIGURE 15 Male genitalia (diagrammatic).

condition for pterygote insects, there is a sclerotized basal portion termed the phallobase and a distal sclerotized portion called the aedeagus (Fig. 15). The phallobase in insects is characterized by highly variable development: sometimes sclerotized and supporting the aedeagus, sometimes forming a sheath for the aedeagus. The phallobase often contains an apodeme, which may provide support or a point for muscle attachment. The phallobase and aedeagus are joined by a membranous phallotheca (Fig. 15). The external walls of the phallobase and aedeagus are called the ectophallus (Fig. 15). The gonopore is positioned at the apex of the ejaculatory duct and is concealed within the phallobase. The gonopore is connected to the apex of the aedeagus via a membranous tube called the endophallus (Fig. 15). In some insects the endophallus may be everted through the aedeagus. The circular aperture at the apex of the aedeagus is called the phallotreme (Fig. 15). In some insects the endophallus and the gonopore may be everted through the phallotreme and into the female's bursa copulatrix. Genital lobes referred to as phallomeres form at the sides of the gonopore in the ontogeny of some insects. Usually the phallomeres unite to form the phallus.

See Also the Following Articles

Body Size • Integument • Legs • Mouthparts • Muscle System • Segmentation

Further Reading

Chapman, R. F. (1982). "The Insects. Structure and Function." 3rd ed. Hodder & Stoughton, London.

DuPorte, E. M. (1957). The comparative morphology of the insect head. *Annu. Rev. Entomol.* **2,** 55–77.

Gordh, G., and Headrick, D. H. (2001). "A Dictionary of Entomology." CAB Internation, Wallingford, Oxon, U.K.

Hinton, H. E. (1981). "The Biology of Insect Eggs." 3 vols., Pergamon Press, Oxford, U.K.

Matsuda, R. (1969). "Morphology and Evolution of the Insect Abdomen with Special Reference to Developmental Patterns and Their Bearings upon Systematics." Pergamon Press, Oxford, U.K.

Snodgrass, R. E. (1935). "Principles of Insect Morphology." McGraw-Hill, New York and London.

Tuxen, S. L. (1970). "Taxonomist's Glossary of Genitalia in Insects." 2nd ed. Munksgaard, Copenhagen.

Anopheles Mosquito

see *Mosquitoes*

Anoplura

see *Phthiraptera*

Antennae

Catherine Loudon

University of Kansas

Antennae are segmented appendages that function primarily as chemosensory and mechanosensory structures. An insect typically has a single pair of antennae located on its head. Antennae in juvenile insects are often very different in morphology from antennae in adult insects, typically being larger or more elaborate in the adult stage. Adult antennae may be sexually dimorphic, appearing very different in the males and females. Antennae are absent in the wingless insects belonging to the order Protura and may be extremely reduced in size in some holometabolous larvae.

STRUCTURE

The overall shape of most insect antennae is elongate and cylindrical (Fig. 1, top), although elaborations into plumose, lamellate, or pectinate forms have arisen many times in different insect lineages (Fig. 1, bottom). An elongate, cylindrical morphology, probably the ancestral condition for insect antennae, is found in fossil insects and many other arthropods. There are three parts to an insect antenna: the scape, the pedicel, and the flagellum. The scape is the first segment (most proximal) of the antenna, and it is attached to the head by a rim of flexible, intersegmental cuticle. Thus, the scape (and the rest of the antenna) can move with respect to the head. All the antennal segments are similarly joined to each other by thin, flexible cuticle.

The movements of an antenna are controlled in part by one or two pairs of muscles that attach inside the head (such

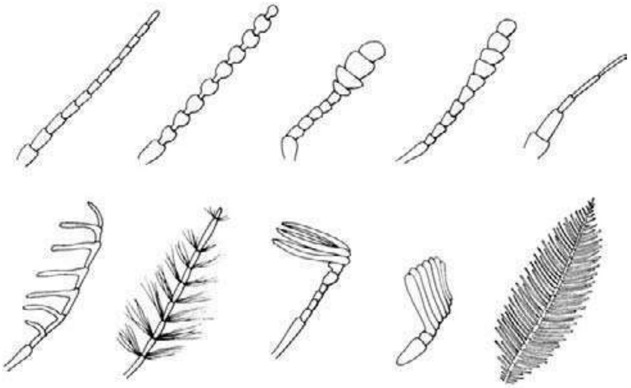

FIGURE 1 Insect antennae exhibit a variety of shapes including elongate morphologies (top) and those with lateral elaborations (bottom). [After Romoser, W. S., and Stoffolano, J. G., Jr. (1998). "The Science of Entomology," WCB/McGraw-Hill, Boston, and Loudon, C., *et al.* (1994). *J. Exp. Biol.* **193,** 233–254, published by McGraw-Hill, with permission of the McGraw-Hill Companies.]

as on the tentorium) with the other end attached inside the scape. An additional pair of muscles runs from the scape to the next segment of the antenna, the pedicel. The combined action of these two sets of muscles is capable of moving an antenna in almost any direction. The final (most distal) segment of the antenna, the flagellum, is the most variable in morphology among insects. The only insects that have intrinsic muscles in the flagellum (joining adjacent segments) are members of the wingless orders Collembola and Diplura. In all other insects (the majority), there are no muscles in the flagellum. Many specialists prefer "annulus" or "subsegment" to "segment" for an individual part of a flagellum in this latter group of insects, because "segment" is reserved for parts with their own musculature. Movements of an annulated flagellum without intrinsic musculature may still occur, such as the spreading and closing of the lamellae or lateral extensions in an antenna (Fig. 1, bottom), but these movements are driven by changes in the pressure of the hemolymph (blood) inside the antenna and thus are hydraulic rather than muscular.

In most insects, circulation of hemolymph through an antenna is facilitated by muscular pumping by an accessory heart located in the head near the base of the antenna. This antennal heart pumps the hemolymph into a blood vessel that discharges the hemolymph at the distal end of the antenna. The return flow of the hemolymph back to the head (and the general open circulatory system of the insect) is not inside a blood vessel. The lumen of an antenna also contains tracheae and nerves, which branch into any lateral extensions of the flagellum. Sensory neurons that send action potentials in response to chemical or physical stimuli sensed by the antennae terminate in the deutocerebrum of the brain. The deutocerebrum is also the site of origin for the motor neurons that stimulate the muscles associated with the antennae.

GROWTH AND DEVELOPMENT

Antennal growth and development in holometabolous insects (those that undergo complete metamorphosis) differs greatly from that in other insects. In holometabolous insects, adult antennae form from imaginal disks, which are clumps of undifferentiated cells that will develop into adult structures. The antennal imaginal disks may appear in the embryonic (fly) or late larval (moth) stage of the immature insect. Properties of the antennal imaginal disks determine to a large extent the chemical stimuli to which an adult will respond, as is seen from experiments in which antennal imaginal disks were cross-transplanted between larvae, which were then reared to adulthood and assayed.

In hemimetabolous and apterygote (wingless) insects, the nymphs are very similar in overall form and habit to the adults, and their antennae resemble smaller, shorter versions of the adult antennae. As with all external structures that are replaced at each molt, a new antenna is formed inside the old antenna. The primary morphological change that occurs at each molt is that the flagellum lengthens with the addition of more segments or annuli, either at the distal end (orders Collembola and Diplura), the proximal end (most other insects), or along the length of the flagellum (some members of the orders Orthoptera and Odonata).

Antennae are serially homologous to mouthparts and legs, reflecting the ancestral condition of a single pair of appendages per body segment shared by arthropods and related groups. Common developmental features between legs and antennae can be seen, for example, in the action of the homeotic gene called *Antennapedia,* which results in the substitution of leglike appendages for antennae on the head when expressed ectopically in mutant *Drosophila.* Leglike appendages appearing in the antennal location in adult insects have also been observed after regeneration of antennae following injury during the larval stage (Fig. 2).

FUNCTION

The primary function of antennae is the assessment of the chemical and physical characteristics of the environment. Detection is made with innervated chemosensory and mechanosensory organs that are arrayed on the antennae. A single antenna usually has sensory organs of several types, with different properties. Most of the chemosensory organs are located on the flagellum and often take the form of microscopic chemosensory hairs (sensilla) each only 1 or 2 μm in diameter. Some antennae, such as the feathery pectinate antennae of silkworms (*Bombyx mori*), have tens of thousands of sensilla, which are capable of very thoroughly sampling the air that passes in the small spaces between them. A cockroach antenna may have hundreds of thousands of sensilla. The chemicals that may be detected by chemoreceptors on the

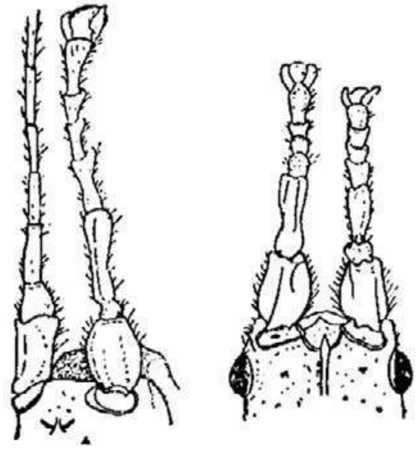

FIGURE 2 Left: head of an adult Indian stick insect (*Carausius morosus*) with a normal antenna on the left and a regenerated antenna with leglike morphology on the right. Right: head of adult *C. morosus* with two regenerated antennae with leglike morphology. [After Fig. 78 in Wigglesworth, V. B. (1971). "The Principles of Insect Physiology." Chapman & Hall, London, with the kind permission of Kluwer Academic Publishers.]

antennae are usually biological in origin and airborne (volatiles), although (depending on the insect species) the sampled chemical compounds are sometimes in a liquid or associated with a solid surface. The chemicals intercepted by antennae may alert the insect to the presence of prospective mates, food, suitable places to lay eggs, or predators.

The physical stimuli detected by mechanoreceptors on the antennae may be used by the insect to indicate air speed during flight, to detect vibrations of the air, or to detect solid boundaries in its environment by touch. While a single mechanosensory hair will send information to the brain about the local physical conditions existing at its microscopic location, an antenna also has mechanosensory organs that evaluate the physical forces acting on the antenna as a whole. These mechanosensory organs, located near the base of the antenna, include Johnston's organ, Böhm bristles, hair plates (groups of mechanosensory hairs), and campaniform sensilla (thin flexible patches of cuticle that are innervated). Johnston's organ is located in the pedicel and responds to changes of location or vibrations of the whole antenna. In contrast, the Böhm bristles, located near the scape–pedicel boundary, send information to the brain about the antennal position, rather than its movements. The variety of mechanosensory organs associated with the first two segments of the antennae are believed to act together to inform a flying insect about its air speed, because greater flying speed will cause greater deflection of the antennae by the air rushing past. Contact chemosensory hairs, so called because the chemical compounds are usually detected when the insect is touching a liquid or solid surface with the antennae, often have mechanosensory capabilities as well and are usually located near the distal ends of antennae.

The function of the antennal sensory organs will be affected by their arrangement on the antennae. For example, sensory organs on the distal tip of a very long antenna will permit chemical or physical sampling of the environment far from the body of the insect. Close packing of sensory hairs will decrease the airflow in their vicinity, and hence will modify both the chemical and physical sampling of the environment by those hairs. The function of the antennae will also be dependent on the behaviors of the insect that will affect the airflow around the antennae, such as flying, wing fanning, postural changes, or oscillating the antennae. A structure projecting into the environment is liable to collect debris that might interfere with its sensory function; both antennal grooming behaviors and modifications of leg parts against which an antenna is scraped are common in insects (Fig. 3). In some insects, antennae are modified for nonsensory functions such as clasping mates during copulation (fleas and collembolans), holding prey items (beetle larvae), or forming a temporary physical connection between an underwater air reservoir and the atmosphere (aquatic beetles).

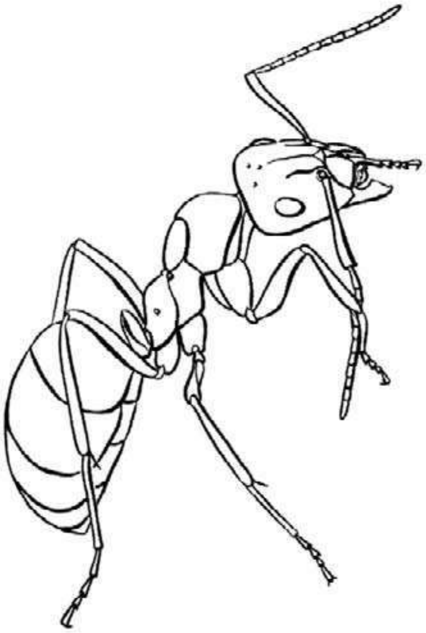

FIGURE 3 A worker ant (*Formica polyctena*) cleans one of its antennae by dragging it across the specialized comb of right foreleg. (Reprinted by permission of the publisher from THE INSECT SOCIETIES by Edward O. Wilson, Cambridge, MA.: The Belknap Press of Harvard University Press, Copyright 1971 by the President and Fellows of Harvard College. Original drawing by Turid Hölldobler.)

See Also the Following Articles

Chemoreception • Imaginal Discs • Mechanoreception • Pheromones

Further Reading

Hansson, B. S., and Anton, S. (2000). Function and morphology of the antennal lobe: New developments. *Annu. Rev. Entomol.* **45,** 203–231.

Heinzel, H., and Gewecke, M. (1987). Aerodynamic and mechanical properties of the antennae as air-current sense organs in *Locusta migratoria.* II. Dynamic characteristics. *J. Comp. Physiol. A* **161,** 671–680.

Kaissling, K. E. (1971). Insect olfaction. *In* Handbook of Sensory Physiology (L. M. Beidler, ed.), Vol. IV of "Chemical Senses," Part 1, "Olfaction," pp. 351–431. Springer, Verlag, Berlin.

Keil, T. A. (1999). Morphology and development of the peripheral olfactory organs. *In* "Insect Olfaction" (B. S. Hansson, ed.), pp. 5–48. Springer-Verlag, Berlin.

Loudon, C. and Koehl, M. A. R. (2000). Sniffing by a silkworm moth: Wing fanning enhances air penetration through and pheromone interception by antennae. *J. Exp. Biol.* **203,** 2977–2990.

Pass, G. (2000). Accessory pulsatile organs: Evolutionary innovations in insects. *Annu. Rev. Entomol.* **45,** 495–518.

Schneider, D. (1964). Insect antennae. *Annu. Rev. Entomol.* **9,** 103–122.

Schneiderman, A. M., Hildebrand, J. G., Brennan, M. M., and Tumlinson, J. H. (1986). Transsexually grafted antennae alter pheromone-directed behaviour in a moth. *Nature* **323,** 801–803.

Steinbrecht, R. A. (1987). Functional morphology of pheromone-sensitive sensilla. *In* "Pheromone Biochemistry" (G. D. Prestwich and G. J. Blomquist, eds.), pp. 353–384. Academic Press, London.

Zacharuk, R. Y. (1985). Antennae and sensilla. *In* "Comprehensive Insect Physiology Biochemistry and Pharmacology" (G. A. Kerkut and L. I. Gilbert eds.), Vol. 6, pp. 1–69. Pergamon Press, New York.

Ants

Nigel R. Franks
University of Bristol

The ants comprise a single family, the Formicidae, within the superfamily Vespoidea and the order Hymenoptera. There are 16 extant subfamilies of ants with a total of 296 extant genera. Some 9000 to 10,000 species of ants have been described, and it is estimated that there may be 15,000 species of ants alive in the world today. The earliest known fossil ants are from the Cretaceous (ca 105–110 mya), but ants probably did not become common until the Eocene (ca 45 mya).

EVOLUTION AND ECOLOGICAL SUCCESS

Ants are now extremely successful ecologically. There may even be an equal biomass of ants and humanity in the world today. They dominate, at their size scale, many terrestrial ecosystems from latitudes north of the boreal tree line to such southern climes as Tierra del Fuego, Chile. In certain tropical forests the contribution of ants to the biomass is spectacular. In Brazilian rain forests, for example, the biomass of ants has been estimated as approximately four times greater than the biomass of all of the vertebrates combined.

One of the reasons ants are so successful is that their colonies have extremely efficient divisions of labor: they evolved factories millions of years before we reinvented them. Another reason is that they can modify their immediate environment to suit themselves, much as we do. Leafcutter ants (*Atta*), for example, evolved agriculture tens of millions of years before humanity developed agronomy. Furthermore, leafcutter ants also use antibiotics and symbiotic bacteria to protect the crop of fungi they grow on the leaves they collect. By contrast, weaver ants (*Oecophylla*) fashion homes from living leaves by sowing them into envelopes, using their larvae as living shuttles and the silken thread they produce as glue. Ants can also dominate areas by mobilizing large numbers of well coordinated foragers; indeed, an ant colony's foragers can be so numerous and well organized that they give the impression of being everywhere at once.

Ants can also be important as seed distributors and as seed harvesters, in the turnover of soils, and in the regulation of aphid numbers and the minimization of outbreaks of defoliating insects. Economically important pest species include the imported fire ant (*Solenopsis invicta*) in North America and leafcutter ants (such as *Atta*) in the neotropics. There are also many ecologically destructive "tramp" ants or invasive species that have been distributed to alien habitats by human commerce.

Ants and plants often have closely coupled ecological relationships. Certain plants even encourage ants by producing rewards such as energy-rich elaiosomes on their seeds to encourage seed dispersal, nutritious Beltian bodies and extrafloral nectaries to entice ants to visit their leaves and shoots (hence to remove the plant's natural enemies while there), or even by supplying preformed homes (domatia) to invite ants directly to inhabit and thus better protect them. Although many ants are hunter-gatherers, very many species tend aphids for the excess honeydew they excrete. By "milking" aphids in this way, ants can in effect become primary consumers of plant products and by thus operating at a lower trophic level they can build up a larger biomass than obligate carnivores would be able to do. Yet most ants mix their diet by also consuming animal protein; for example, they will devour their own aphid milk cows if the latter become sufficiently abundant.

Arguably, the best evidence of the ecological success of ants is that their worst enemies are other ants.

EUSOCIALITY, SOCIAL ORGANIZATION, AND SOCIAL DIVERSITY

Except for a few species that have secondarily lost the worker caste, all ants are eusocial: they have an overlap of adult generations, cooperative brood care, and reproduction dominated by a minority of the colony's members. Typically, an established ant colony consists of one or more queens (each of which may have mated with one or more winged males on a nuptial flight), an all-female set of wingless workers, and the colony's brood of eggs, larvae, and pupae. The majority of queens mate only before they establish a colony. Thereafter, they store the sperm they have received.

All ants have haplodiploid sex determination. This property probably had a major role in the evolution of their eusociality through kin selection. Males are haploid, having only a single set of chromosomes, and thus the sperm that individual males produce is genetically homogeneous. Hence, the (diploid) daughters of the same mother and father are unusually closely related to one another, a circumstance likely to have favored the evolution of female workers. Nevertheless, there can be continuing conflicts within colonies between the workers and the queen (or queens) over the sex ratios they produce and which colony members produce the males. Queens can choose to produce either unfertilized (haploid) eggs destined to become males or fertilized (diploid) eggs. The latter may develop into workers or potential new queens (gynes) generally depending on how much food they receive as larvae. The workers may or may not be sterile. Fertile workers produce viable (unfertilized) haploid eggs that can develop into males. Hence, there can be conflict both among the workers and between the workers and the queen over whose sons the colony produces. Indeed, in many species of ants with only small numbers of workers in their mature colonies, there are dominance hierarchies among the workers, who fight one another over egg production. Sometimes the queen moves with active aggression against the most dominant worker to curtail its production of sons in favor of her own. In addition, even when workers are sterile and serve one, singly mated queen, they may prefer to raise more of the queen's daughters, to whom they are more closely related, than the queen's sons. For all these reasons, the study of ants has had a major impact in recent pioneering evolutionary biology because these insects provide test cases by which the evolutionary resolution of the tension between cooperation and conflict can be explored. It is clear, though, that the apparent social cohesion of ant colonies is often partly an illusion.

Among ants, there is a diversity of mating systems and social organizations. So even though it is tempting to think of the typical ant colony as having a single, singly mated queen and occupying a single nest site, the diversity of social systems among the ants is in fact huge. For example, many ant species consist of facultatively multiqueened (polygynous) colonies. Indeed, roughly half of European ant species exhibit polygyny, and there seems to be no reason to regard this as an unusual proportion. Some ant colonies are founded by solitary queens; some by groups of unrelated queens that may later fight over who will be the one to succeed. Other colonies simultaneously occupy multiple nests (polydomy), a habit often associated with polygyny, while others exhibit colony fission, with both daughter colonies usually being monogynous. Most persistent polygyny is associated with the secondary adoption of queens. Unusual social systems include queenless ants, workerless ants (inquilines), and slave-making ants. In certain queenless species, the workerlike females produce other diploid females through a parthenogenetic process called thelytoky. By contrast, certain inquilines have dispensed with the worker caste, and queens infiltrate and exploit established colonies of other species. Slave making may occur both intraspecifically and interspecifically. Interspecific slave making is also associated with nonindependent colony foundation in which slave-maker queens infiltrate established colonies of their host species, kill the host queen or queens, and produce workers that are reared by currently available host workers. The slave-maker workers raid other neighboring host colonies to capture large larvae and pupae. Such raids thus replenish the stocks of slave workers, which do all the foraging and brood rearing for the slave makers. There are also ant species in which there are polymorphic queens, others in which there are polymorphic males, and many in which there are polymorphic workers.

One of the outcomes of eusociality is that established colonies can be well defended by the workers against enemies. Thus, ant colonies are relatively K-selected; that is, they are selected to hold onto resources and to persist for long periods rather than being ephemeral, here-today-gone-tomorrow, r-strategists. Associated with this trait is the extreme longevity of ant queens. It is estimated that they can live 100 times longer than other solitary insects of a similar size. Worker populations in mature, well-established monogynous colonies range from a few tens of millions to 20 million, and certain so-called supercolonies consist of a huge network of linked nests each with many queens. One supercolony of *Formica yessensis* in Japan may have as many as 300 million workers. Given such longevities and densities, it is clear that ants may also prove to be important model systems for understanding the spread of disease or the evolution of mechanisms to minimize the spread of disease among viscous populations of close kin. It is even possible that polygyny and multiple mating (polyandry) have evolved, at least in part, to promote genetic heterogeneity within colonies and thus help to minimize disease risks.

DIVISION OF LABOR

The relatively large biomass of ants in many ecosystems can be attributed not just to the way in which the ants interact with other organisms but to the way in which they interact with their nestmates in general and, in particular, to efficiencies that accrue from divisions of labor. One of the most dramatic traits associated with the division of labor among the workers is physical polymorphism, which is the presence of different physical worker forms within the same colony. In the African army ant, *Dorylus wilverthi,* for example, the smallest workers at 0.12 mg dry weight are only 1% of the dry weight of the largest workers (soldiers), and this relatively great size range is exceeded in certain other species (e.g., in *Pheidologeton diversus,* the smallest workers have a dry weight that is about 0.2% that of the largest majors). It is not just the size range that is impressive in such species but also the degree of polymorphism among the workers. Darwin, writing in *The Origin of Species,* seemed well aware not only of the phenomenon but also of its implications. Indeed, one of Darwin's most penetrating insights in his 1859 masterpiece was his suggestion that sterile forms evolved in social insects because they are "profitable to the community" and that "selection may be applied to the family, as well as to the individual." He further suggested that once such colony-level selection had begun, the sterile forms could be molded into distinct castes "Thus in [the army ant] *Eciton,* there are working and soldier neuters, with jaws and instincts extraordinarily different" (Fig. 1a, b).

Such worker polymorphism is now known to be associated with the differential growth rates of different putative tissues and body parts during the preadult stages. Indeed, the study of ants made a major contribution to the development of the concept of allometric growth (Fig. 1c, d). Notably polymorphic genera include the army ants *Eciton* and *Dorylus,* leafcutter ants (*Atta*), carpenter ants (*Camponotus*), and members of the genera *Pheidole* and *Pheidologeton.* Indeed, *Camponotus* and *Pheidole* are the two most species-rich ant genera.

However, genera with polymorphic workers are in the minority. Approximately 80% of ant genera consist entirely of species with monomorphic workers, most of the remaining genera consist of species in which there are at most only two easily recognizable worker morphs, and only about 1% of genera have species in which three or more worker morphs can be relatively easily recognized within colonies.

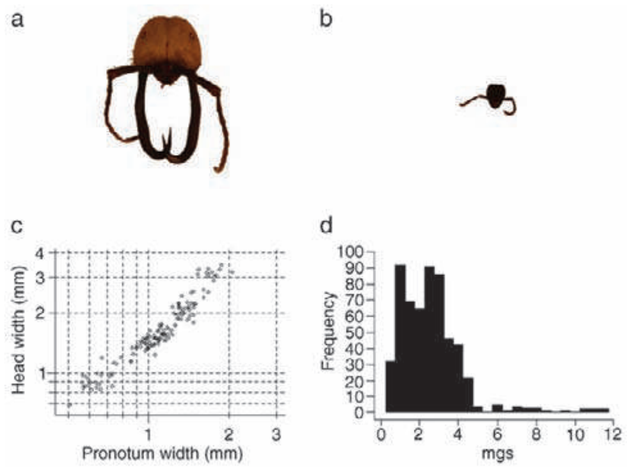

FIGURE 1 The army ant, *Eciton burchelli.* (a) Head of major worker. (b) Head of minor worker. (c) Head width vs ponotum width allometry for workers. (d) Frequency–dry weight histogram for a large sample of workers. The allometrical relationship has a slope greater than 1, so larger workers (such as majors) have disproportionately large heads. The size frequency distribution is skewed to the right so relatively few of these very large majors are produced. (Drawings © Nigel R. Franks.)

Polymorphism among the workers is mostly associated with extreme physical specialization. Thus, *Eciton* majors have ice-tong-like mandibles and are specialist defenders of the colony against would-be vertebrate predators or thieves (Fig. 1a). It has been shown that colonies of *Pheidole pallidula* can produce more defensive majors in response to stresses induced by conspecific competitors. Majors are not always for defense: large-headed majors in *Pheidole* and *Messor* serve as specialist grinders of harvested seeds. Even among such polymorphic species, however, the majority of workers belong to castes of generalists, which give their colonies an ability to respond rapidly to changes in the environment. Such generalists show behavioral flexibility not possible with the extreme morphological specialization of certain physical castes. Nevertheless, divisions of labor also occur within the majority generalist caste. Such workers typically specialize in different tasks at different times during their lives. This is known as temporal polyethism, in contrast to physical polyethism.

The sophisticated divisions of labor in monomorphic ants are being investigated. In *Leptothorax albipennis,* the workers show very little size variation, and colonies consist of, at most, a few hundred such workers living in flat crevices between rocks. Such crevices can be only 2 or 3 cm wide and deep and may have an internal cavity height of only 1 or 2 mm. Individual workers could easily roam all around such nests within a minute, but instead they have spatial fidelity zones; that is, they remain faithful to certain parts of the nest and the segregated tasks within such areas for months on end. The workers can even reconstruct their own spatial fidelity zones relative to one another if, and when, their colony is forced to emigrate to a new nest site because of the destruction of the old site. In this (and many if not all) ant species, younger workers tend to work deep within the nest at its safe center, tending the queen and the eggs. As they get older, workers tend to move progressively out from the center of the nest, and toward the end of their lives they eventually engage in the most dangerous task of foraging in the outside world, where they are likely to meet predators and other hazards. However, the correlation between age and task is often very weak, and in an increasing number of species it has been shown that the division of labor among monomorphic workers is extremely flexible. Workers can respond to the removal of other workers by reverting to tasks that they did earlier in their lives or, if need be, they may begin foraging even when they are very young. Thus, though age may influence what workers do, it is unlikely to be the organizing principle of the division of labor in many species. Rather, it seems that workers are continuously monitoring their workloads and the delays they experience while waiting to interact with their nestmates and will flexibly change their tasks accordingly to maximize their productivity.

COMMUNICATION AND PHEROMONES

Ants have diverse systems of communication, but by far the most important medium for signaling involves the chemicals known as pheromones. Ants can deposit chemical trails to recruit nestmates to discoveries of food. Many ants can also produce highly volatile chemicals to signal alarm when they encounter dangerous predators or other hazards. Different ants in different subfamilies use a remarkable diversity of glandular structures even just to produce recruitment pheromones. These may be produced from cloacal glands, Dufour's glands, the hindgut, poison glands, pygidial glands, rectal glands, sternal glands, or even tibial glands on the back legs. Furthermore, many pheromones appear to be complex mixtures of many chemical compounds.

Pheromones can be effective in minute quantities; it has been estimated that one milligram of the trail substance of the leafcutting ant, *Atta texana,* if laid out with maximum efficiency, would be sufficient to lead a colony three times around the world.

Nestmate recognition is another important aspect of communication in ants. A pleasing metaphor for the ant colony is a factory inside a fortress. Ant colonies are dedicated to the production of more ants; but workers need to "know" that they are working for their natal colony, and colonies also need to be well defended against other ants and against infiltration by other arthropods, which might tap into their resources. Ant colonies employ colony-specific recognition cues as one of their defense systems. These are often in the form of cuticular hydrocarbons that can be spread throughout the colony both by grooming and trophallaxis (the latter is usually associated with liquid food exchange). Slave-making ants circumvent the recognition cues of their slaves by capturing them as larvae and pupae—these captives are not yet imprinted on their natal colony odor but later become imprinted on the odor of the

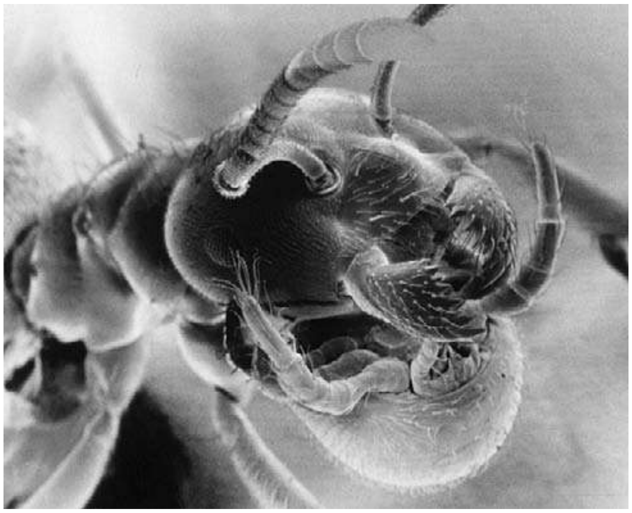

FIGURE 2 Scanning electron micrograph of a worker of *Lasius flavus* with a kleptoparasitic mite, *Antennophorus grandis,* gripping on its head. The mite steals food when two workers exchange nutritious liquids during trophallaxis. (Photomicrograph © Nigel R. Franks.)

colony that kidnapped them after they have metamorphosed into adult workers. Sometimes colony-specific odors also can be influenced by chemicals picked up from the colony's environment. Nevertheless, countless species of arthropods from mites to beetles have infiltrated ant colonies. For example, more than 200 species of rove beetle (Staphilinidae) are associated with New World army ants alone, and other groups such as mites are probably even more species rich. Often these infiltrators are called "guests" simply because their relationships with their host ant colony and to its resources are unknown (Fig. 2).

SELF-ORGANIZATION, COLLECTIVE INTELLIGENCE, AND DECISION MAKING

A rapidly developing approach to the study of ants and other social insects is the application of self-organization theories. Here self-organization can be defined as a mechanism for building spatial structures and temporal patterns of activity at a global (collective or colony) level by means of multiple interactions among components at the individual (e.g., worker) level. The components interact through local, often simple, rules that do not directly or explicitly code for the global structures. The importance of studies of such self-organization is that they can show how very sophisticated structures can be produced at the colony level with a fully decentralized system of control in which the workers have no overview of the problems they are working to solve.

A simple and very intuitive example of how ants use self-organization is found in their ability to select short cuts. Certain ants can select the shortest paths to food sources. Indeed, where there is a short and a long path to the same food source, the decision-making mechanism can be surprisingly simple. The ants that happen to take the shorter path get there and back more quickly than the ants that happen to take the longer path. All the ants lay attractive trail pheromones, and such pheromones are reinforced more rapidly on the shorter path simply because that path is shorter and quicker. In such cases, individual ants do not directly compare the lengths of the two paths, but the colony is able to choose the shorter one. Sometimes the shorter path is used exclusively, while at other times a small amount of traffic may continue to use the longer path. Having some traffic that continues to use the longer path is likely to be costly in the short term, but it may represent a beneficial insurance policy if the shorter path becomes blocked or dangerous. Self-organization also has a major role in such phenomena as brood sorting, rhythms of activity within nests, and building behaviors. This new approach may help to answer, at least in part, the age-old challenge of how ant colonies are organized.

See Also the Following Articles

Caste • Colonies • Nest Building • Pheromones • Sex Determination • Sociality

Further Reading

Bolton, B. (1994). "Identification Guide to the Ant Genera of the World." Harvard University Press, Cambridge, MA.

Bourke, A. F. G., and Franks, N. R. (1995). "Social Evolution in Ants." Monographs in Behavioral Ecology. Princeton University Press, Princeton, NJ.

Camazine, S., Deneubourg, J.-L., Franks, N. R., Sneyd, J., Theraulaz, G., and Bonabeau, E. (2001). "Self-Organization in Biology." Princeton University Press, Princeton, NJ.

Detrain, C., Deneubourg, J.-L., and Pasteels, J. (1999) "Information Processing in Social Insects." Birkhäuser Verlag, Basel, Switzerland.

Grimaldi, D., and Agosti, D. (2000). A formicine in New Jersey Cretaceous amber (Hymenoptera: Formicidae) and the early evolution of the ants. *Proc. Natl. Acad. Sci. USA* **97,** 13678–13683.

Hölldobler, B., and Wilson, E. O. (1990). "The Ants." Belknap Press, Cambridge, MA.

Keller, L., and Genoud, M. (1997). Extraordinary lifespans in ants: A test of evolutionary theories of ageing. *Nature* **389,** 958–960.

Passera, L., Roncin, E., Kauffmann, B., and Keller, L. (1996). Increased soldier production in ant colonies exposed to intraspecific competition. *Nature* **379,** 630–631.

Schmidt-Hempel, P., and Crozier, R. H. (1999). Polyandry versus polygyny versus parasites. *Philos. Trans. R. Soc. (Lond) (B)* **354,** 507–515.

Sendova-Franks, A. B., and Franks, N. R. (1994). Social resilience in individual worker ants and its role in division of labour. *Proc. R. Soc. (Lond) (B)* **256,** 305–309.

Aphids

John T. Sorensen

California Department of Food and Agriculture

Aphids are remarkable, evolutionarily exquisite creatures, and among the most successful insects. Aphid evolution has been shaped through nutrient-driven selection and by the host plants on which they feed, and aphids have responded by developing intricate life cycles and complex polymorphisms. These sap-feeding hemipterans have coped with a hostile world through developing an exceptionally high reproductive rate and passive wind-borne dispersal, a strategy in which individuals are quite expendable, but survival and prosperity of their genes are guaranteed. Because of their intriguing evolutionary adaptations, aphids were among our most worthy competitors as humans entered the agricultural era.

MAJOR GROUPS AND HOST AFFILIATIONS

Aphids, as the superfamily Aphidoidea, belong to the Sternorrhyncha within the Hemiptera, a group they share with Aleyrodoidea (whiteflies), Psylloidea (jumping plant lice), and Coccoidea (scale insects and mealybugs). Aphidoidea has three families: Adelgidae (adelgids), Phylloxoridae (phylloxorids), and Aphididae (aphids), although some workers place the Adelgidae and Phylloxoridae in a separate superfamily, Phylloxoroidea. Adelgids and phylloxorids are primitive "aphids" and older groups, each with about 50 species. They differ from Aphididae by having an ovipositor and by reproducing by means of ovipary. Adelgids are restricted to conifers (Pinaceae), and some form characteristic galls (e.g., *Adelges piceae,* balsam woolly adelgid). Phylloxorids, which may also form galls, occur on plants of the Salicaceae (willow family), Fagaceae (oak family), Juglandaceae (walnut family), and Rosaceae (rose family). An exceptional species, *Daktulosphaira vitifolae,* grape phylloxera, feeds on grapes (Vitaceae), damaging European grape cultivars unless they are grafted to resistant rootstocks developed from American grape species.

Aphids originally evolved on woody plants in the Northern Hemisphere and are functionally replaced by whiteflies and psyllids in the Southern Hemisphere. As a group, they evolved and began their diversification with angiosperms, over 140 mya during the lower Cretaceous. While most fossil aphid groups became extinct during the Cretaceous–Tertiary boundary, most modern aphid groups radiated during the Miocene. Aphids have siphunculi, which vary by group from being mere pores on the abdominal surface to being very elongate tubes. They also have a cauda, which varies by group from rounded and hardly noticeable to knobbed or long and fingerlike. Aphids lack an ovipositor and are viviparous, bearing young parthenogenetically.

Aphid taxonomy is difficult; their subfamily classification has been argued and confused with nearly as many classifications as aphid taxonomists. Remaudiere and Remaudiere's 1997 classification, followed here, recognizes about 25 aphid subfamilies, with tribal groupings for about 600 genera and 4700 species of aphids. Many aphid lineages coevolved with, and radiated among, their host plant groups. Often during their phylogenetic history, however, aphid groups opportunistically switched to radically unrelated host groupings, driven by developmental requirements but tempered by evolutionary constraints.

Many aphid subfamilies are small, but several are larger and important: Chaitophorinae (e.g., *Sipha flava,* yellow sugarcane aphid), on Salicaceae and Gramineae (grass family); the closely related Myzocallidinae (e.g., *Therioaphis trifolii* f. *maculata,* spotted alfalfa aphid), Drepanosiphinae (e.g., *Drepanaphis acerifoliae,* painted maple aphid), and Phyllaphidinae (e.g., *Phyllaphis fagii,* beech aphid), often considered to be one subfamily and usually on dicotygledonous trees, but also Fabaceae (legume family) and bamboo; Lachninae (e.g., *Essigella californica,* Monterey pine aphid), mostly on Pinaceae, but also Fagaceae, Rosaceae, and Asteraceae (composite family) roots; and Pemphiginae (e.g., *Pemphigus bursarius,* lettuce root aphid), often on roots and host alternating to dicotyledonous trees forming galls. Other noteworthy subfamilies include Pterocommatinae, on Salicaceae; Greenideinae, on Fagaceae; Mindarinae, on Pinaceae; and the host-alternating Anoeciinae and Hormaphidinae, the latter causing galls.

The largest and most evolutionarily recent subfamily, Aphidinae, has two large, diverse, and agriculturally important tribes. The first tribe, Macrosiphini (e.g., *Aulacorthum solani,* foxglove aphid), is diverse in genera, which often lack attendance by ants but may alternate hosts. The second tribe, Aphidini, is diverse in species but less so in genera; these are often attended by ants. Tribe Aphidini has two important subtribes. Subtribe Rhopalosiphina (e.g., *Rhopalosiphum padi,* bird cherry–oat aphid) host alternates between Rosaceae to Gramineae or Cyperaceae (reed family). Subtribe Aphidina (e.g., *Aphis fabae,* bean aphid) host alternates mostly among Rosidae and Asteridae and is home to genus *Aphis,* which alone contains well over 1000 species.

NUTRITION-DRIVEN EVOLUTION: LIFE CYCLES AND POLYMORPHISM

Aphid life cycles are complex and may be either monoecious or dioecious, involving holocycly or anholocycly. Because of this, aphids have evolved many specialized morphs; a multitude of confusing, often synonymous names have risen among aphid biologists, but these are minimized here.

In the simple and generalized monoecious holocyclic aphid life cycle (Fig. 1A), a single host plant species is used throughout the year and sexual morphs are produced in the fall, usually in response to decreasing daylength. The males and oviparae mate, producing genetically recombinant eggs that overwinter on the host plant and often experience high mortality. In the spring, the fundatrix emerges from the egg, matures parthenogenetically, and gives live birth to nymphs that become viviparae and continue in that reproductive mode through the summer. If the aphid group produces plant galls, the fundatrix is responsible for their production.

The viviparae may be apterae or alatae (Fig. 2), but in some groups (e.g., Drepanosiphini, some Myzocallidinae) all viviparae are alatae. The parthenogenetic reproduction of

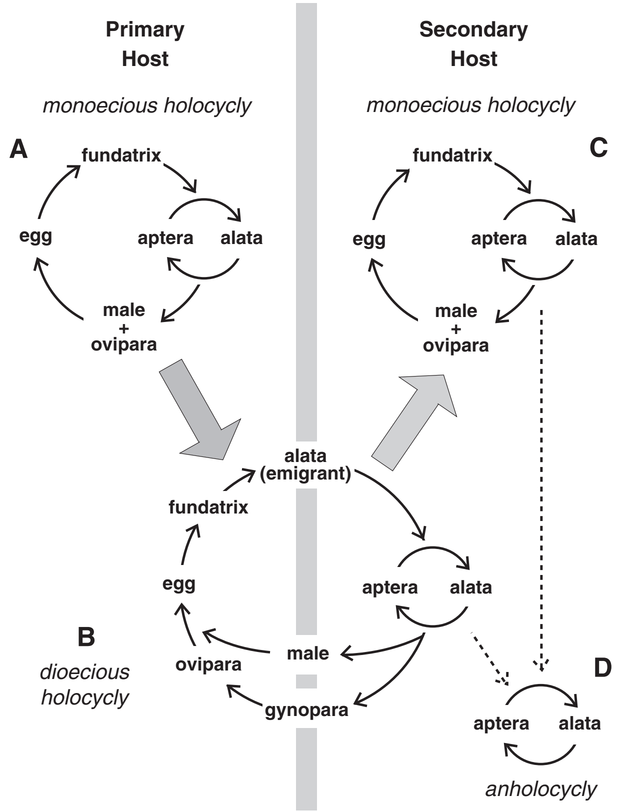

FIGURE 1 Evolutionary development of generalized aphid life cycles. Initially, aphids developed monoecious holocycly (A) on an ancestral woody primary host, where aestivation occurred because sap amino acids were unavailable during summer growth cessation. Next, multiple subfamilies independently evolved dioecious holocycly (B), where viviparae moved to summer-growing herbaceous secondary hosts but returned to their ancestral host in autumn. In some aphids, secondarily monoecious holocycly (C) developed on the secondary host when the primary host was lost. Often in warm areas, where selection for an overwintering egg is not imposed, some populations of dioecious and secondarily monoecious holocyclic aphids may lapse into facultative anholocycly (D) on their secondary hosts; this condition may become obligate anholocycly if the ability to produce sexuals is lost.

viviparae allows very rapid buildup of numbers and collapse of generation time. When a viviparous nymph is born, it has the embryos of both its daughters and granddaughters within it, creating a "telescoping" of generations. Apterae have lost their wings and associated musculature to optimize reproduction. They produce more offspring per female than do alatae, which must invest resources in their flight apparatus. However, alatae produce progeny earlier in life than do apterae, giving their relatively reduced number of offspring a better generational turnaround time than apterae mothers can. Apterae are selectively produced when the host plant is a good source of nutrients. Once an aphid population has built, either inducing a crowding effect among apterae or stressing its host to the level of impacting nutrient levels, the population usually switches to produce alatae, which migrate to better situations. However, the risks of successful migration

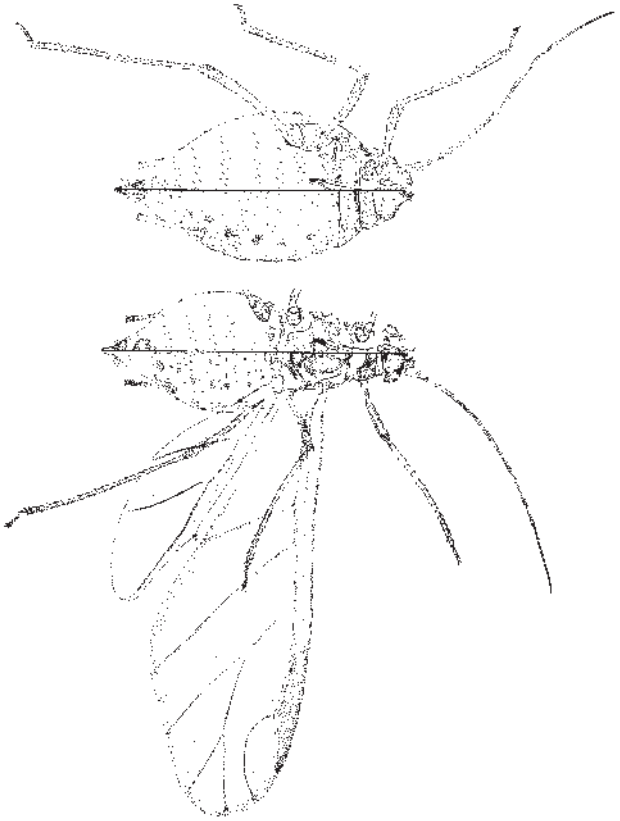

FIGURE 2 Aptera (top) and alata (bottom) viviparae of *M. persicae.* Shown in split images with ventral (upper half) and dorsal (lower half) aspects with heads to the right. [Drawings by Tokuwo Kono, modified from Kono, T., and Papp, C. S. (1977). "Handbook of Agricultural Pests, Aphids, Thrips, Mites, Snails, and Slugs." California Department of Food and Agriculture, Sacramento.]

are great, especially for monophagous aphids that feed on uncommon hosts, because the flight of alatae is wind-borne and relatively passive. Alatae can be blown over 1600 km, often across an ocean, and survive the trip. Upon successfully alighting on their proper host and feeding for a short time, alatae begin autolysis of their flight musculature, precluding further flight but self-cannibalistically providing nutrients for their offspring. The production of viviparae continues until fall conditions trigger production of the sexuals.

A second, more complicated dioecious life cycle (Fig. 1B) has independently evolved among several different aphid groups that show seasonal alternation between differing hosts. This dioecious cycle probably evolved in response to the seasonally inadequate supply of nitrogen-based nutrients, especially amino acids, on their primary host. The phloem sap that aphids feed on has limited nitrogen availability, and nitrogen is the limiting nutrient in aphid development. Woody deciduous plants normally translocate amino acids in quantity only during the spring, when they are foliating, and in the fall, when leaf senescence breaks down leaf protein and nitrogen is translocated to the roots for overwinter storage. Aphids groups evolving on and restricted to such plants face a nitrogen deficit during the summer, when active plant growth ceases and phloem sap is low or devoid of nitrogen. Such groups (e.g., *Periphyllus* spp.) may develop an aestivating nymph that halts growth until fall. Other aphid groups (e.g., Aphidinae) whose ancestors originated on deciduous woody plants, have evolved to leave those primary hosts during the late spring, after the nitrogen flush associated with foliation has ceased. In doing so, their spring alatae, as emigrants, migrate to herbaceous secondary hosts that actively grow and transport nitrogen during the summer. In the fall, however, as their secondary hosts die back, the aphids return to their woody primary host by producing migrating males and gynoparae. There, the aphid's sexuals, its males and oviparae, capture that host's fall nitrogen flush and mate to lay their overwintering eggs in anticipation of the spring nitrogen flush. Depending on the aphid or its group, the secondary hosts may vary from quite specific to a broad number of botanical groups; but the primary hosts are often specific to a plant genus. Most aphid lineages have adapted specific types of secondary hosts, such as grasses (e.g., *Metopolophium dirhodum,* rose-grain aphid), roots (e.g., *Smynthurodes betae,* bean root aphid), other woody plants (e.g., *Hormaphis hamamelidus*), or herbs (e.g., *Macrosiphum rosae,* rose aphid). Some aphids specialize on secondary hosts of a particular environmental ecotype; for example, *Rhopalosiphum nymphae,* waterlily aphid, uses aquatic plants in many plant families.

Some aphid lineages (e.g., *Schizaphis graminum,* greenbug) have evolved beyond dioecious holocycly, entirely leaving their primary host to remain on their secondary host, in secondarily monoecious holocycly (Fig. 1C). However, an important form of year-round residence on the secondary host occurs in warmer climates, where populations do not require an egg for overwintering survival. Under such conditions, otherwise holocyclic dioecious or monoecious populations may lapse facultatively into anholocycly on their secondary hosts (Fig. 1D). If such populations remain anholocyclic long enough, they may eventually evolve into obligate anholocycly by losing the ability to produce sexual morphs, despite undergoing environmental conditions that normally trigger their production. Depending on the aphid lineage and its adaptation to its host(s) or their alternation, nearly all aphid morphs may be winged or wingless, but the morph's wing condition is specific to the aphid group.

APHID BEHAVIOR

Aphids feed by inserting their rostrum-borne stylets into a plant and ratcheting them between plant cells, seldom penetrating any until they enter the phloem sieve tubes and extract sap. Stylet advancement is lubricated by saliva containing a pectinase that loosens the bonding between plant cells. The saliva forms a stylet sheath that is left in the plant when the stylets are withdrawn. To cope with a sap diet, aphid guts have specialized groups of cells, mycetomes, containing rickettsia-like symbiotic bacteria, mycetocytes, which aid in synthesis of nutrients. These bacteria, which are passed from mother to embryonic daughter, have coevolved with aphid lineages, differing among them.

Whereas aphids largely rely on a high reproductive rate and great dispersive ability to maximize survival in a hostile environment, morph-specific behaviors exist to promote genetic survival of the individual or its clone. Behavior of alatae optimizes dispersion and finding a successful host. When alatae initially take to flight, they are attracted to the short wavelengths of light that predominate in a clear or cloudy sky, and fly up toward them. After flying a while, however, they come to prefer the longer light wavelengths reflected from plants, and they descend, moving to them. In some species, alatae have shown a preference for colors characteristic of their host plant's leaves. Generally alatae are attracted to yellow, a predominant hue in growing or senescent plants, which are better nitrogen sources. Upon alighting on a plant, they briefly probe below the epidermis with the rostrum to locate specialized secondary plant compounds that are of no nutritional value but are specific to the aphids' given host. If these feeding triggers are not found, the alatae move on.

In contrast, apterae usually move only when necessary to procure a better feeding site or if a predator or parasite molests them. Ants tend apterae in many aphid groups in a form of facultative mutualism; in some relationships, ants actively "farm" their aphid "cattle" by moving them among locations. Generally, however, aphid groups with elongate siphunculi are less likely to be tended by ants. In exchange for the aphid's sugary honeydew waste, the ants protect them from predators, such as coccinellid, lacewing, and syrphid fly larvae, or specialized aphid parasites, such as chalcidoid and braconid wasps. When stroked by the ant's antennae, the aphid will raise the tip of its abdomen, extruding a honeydew drop, which may be retracted if not accepted by the ant. If an ant does not accept honeydew after a while, the aphid will revert to its normal behavior of flicking the honeydew drop away with its hind leg or cauda, to prevent an accumulation of honeydew from fouling the aphid colony.

Aphids communicate by chemicals and sound. Parasites and predators are often foiled by the use of an aphid alarm pheromone, such as *trans*-β-farnesene. When molested, aphids exude microdroplets of alarm pheromone from their siphuncular pores, and in response adjacent aphids quickly withdraw their stylets from their host and drop to the ground. Shortly thereafter, the fallen aphids visually orient to vertical lines or structures and move toward them in an attempt to climb the plant stem. Aphid sexual pheromones are also used as male attractants by oviparae, being released from specialized pores on their hind tibiae. Sound communication is used by *Toxoptera* spp., which have a stridulatory mechanism consisting of a row of short pegs on the legs, which are rubbed against filelike ridges on the lower epidermis of the abdomen, just below the siphunculi. When disturbed, colonies of *T. aurantii* emit an audible high piercing stridulatory sound, to which their apterae respond.

Fundatrices of gall-forming species use species-specific patterns of feeding or probing behavior to induce characteristically shaped galls on their specialized hosts, in which their progeny can safely develop. The fundatrix of *P. bursarius* climbs the developing leaf petiole on *Populus nigra* about halfway and probes its rostrum around the petiole to create an array of punctures oriented perpendicularly to the petiole shaft. This induces a swollen globular gall with a slit oriented perpendicularly to the petiole shaft. In contrast, on the same host, a *Pemphigus spyrothecae* fundatrix probes the petiole shaft in an upward spiral array of punctures, yielding a corkscrew-shaped petiole gall. Not only do plant galls provide a protective encasement for aphid development, but aphids of even nongalling species do better on galled tissue, probably because of a local increase in plant nutrients in that tissue.

Many aphid species have some lower degree of sociality, especially among apterae, which is expressed as a gregariousness within colonies and probably confers better protection or response to attacks by natural enemies. Alatae of *Drepanosiphum platanoides,* sycamore aphid, are more likely to be distributed in a clustered manner among sycamore leaves, in groups in which the tips of their antennae and legs touch among the aphids. Some aphids have evolved a higher degree of sociality, however. The tribe Cerataphidini of the Hormaphidinae has genera in which species produce a soldier morph with enlarged forelegs, which defend their relative clones differentially. Soldiers discriminate between soldiers and nonsoldiers but do not attack soldiers of their own species. The investment in soldier production by the colony is related to areas needing defense, such as a gall's surface.

AGRICULTURAL IMPORTANCE

While aphids are among the most serious agricultural problem insects, only about 250 species are considered to be agricultural pests. Pest aphids may affect only a very specific host (e.g., *Brachycorynella asparagi,* asparagus aphid), or group of related hosts, such as crucifers (Brassicaceae) (e.g., *Brevicoryne brassicae,* cabbage aphid). Some, however, are quite polyphagous (e.g., *Aphis gossypii,* cotton aphid; *Myzus persicae,* green peach aphid), with an extremely wide host range. Some common polyphagous pest aphids represent sibling species complexes that are morphologically identical but differ in karyotype. They comprise anholocyclic clones, or biotypes, that differ in host preferences, ability to transmit diseases, or resistance to pesticides.

Aphids cause damage in several ways. They can build to high population densities and damage plants directly, by removing enough sap to cause withering and eventual plant death. If not washed off, aphid excrement, or honeydew, can build up enough on plants to serve as a medium for the growth of sooty molds, impairing photosynthesis and plant development, and eventually promoting other fungal diseases. Salivary secretions of some aphids are phytotoxic, causing stunting, leaf deformation, and gall formation. Even if the feeding effects of aphids are not apparent, they may affect plant hormone balances, changing host metabolism to their advantage, thus essentially hijacking the plant's physiological functions.

The aphid vectoring of stylet-borne and circulative plant viruses is the most serious problem to agriculture posed by aphids. Stylet-borne viruses occur on the aphid's epidermis and are not aphid specific. These viruses are acquired quickly and transmitted during the aphid's probing of the plant's epidermis. They are nonpersistent, however, and the aphid's infectiousness is lost upon molting. Circulative viruses, in contrast, live internally in the aphid's gut. The aphid must feed for a while to acquire these viruses, which require an incubation period before they can be successfully transmitted. They are persistent, however, and once infected, the aphid remains a vector throughout its life. The virus–aphid–plant linkage is fairly specific for circulative viruses, and a given virus is transmitted by only one or few aphid species. Virus-infected plants often show an aphid-attractive yellowing and have increased free amino acids, so aphids benefit by virus transmission.

APHID CONTROL IN AGRICULTURAL CROPS AND HOME GARDENS

Agricultural control of aphids best uses an integrated pest management (IPM) strategy, where species are identified and tactics reflect the allowable tolerance level on a crop. Within fields, aphids may be monitored by means of yellow water pans or sticky traps, which attract them. In some agricultural regions, especially seed-growing areas with plant virus sensitivities (e.g., the Netherlands, Idaho), specialized agencies run aerial trapping networks in which large suction traps are used to detect alates and forecast population levels. Proper aphid IPM emphasizes sustainable control, maximizing organically compatible methods to minimize effects on nontarget species, such as biological control agents, or vertebrates. IPM tactics include cultural control methods, such as minimizing weed or ant populations that promote aphids, using ultraviolet-light-reflecting or colored films near plants to repel alates, or interplanting pollen and nectar plants among crop rows to promote aphid natural enemies. Biological control agents include small wasps (e.g., *Aphidius* sp.) that parasitize aphids and disperse well within populations. Predators, which as immatures voraciously consume aphids, can be released. These include lacewings (e.g., *Chrysopa* spp.), aphid midges (e.g., *Aphidoletes* spp.), and ladybird beetle larvae (e.g., *Hippodamia convergens*). Predators may, however, disseminate when released as adults. One can apply entomopathic fungi (e.g., *Beauveria bassiana*), whose spores attach to the aphid's exoskeleton, penetrate it, and kill the aphid. Insect growth regulators applied by spray act through various means to prevent maturation of aphids. These may act in conjunction with biological control agents if the latter fail to provide adequate control. Use of chemical poisons in aphid IPM should be minimized because of the effect on nontarget species. While poison use may sometimes be necessary, heavy usage promotes insecticidal resistance in aphids, as well as secondary resurgence of aphid populations, once biological control agents have been hampered. Chemical poisons range from less toxic pyrethroids to more toxic organophosphates. They may be applied directly as contact insecticidal sprays or dusts, or indirectly as plant systemic insecticides that are ingested with the plant's sap. Cultivation of aphid-resistant crop varieties is also important.

In home gardens and yards, nontoxic controls should be emphasized. Aphid detection involves inspection of buds, stems, fruits, and the underside of leaves, where the insects are most likely to congregate. Effective control can simply involve frequently hosing aphids off plants with water, being careful to hit the leaf undersides. Spray applications of a mixture of garlic and water may repel aphids. Sprays of cuticle-disrupting insecticidal soaps, which cause fatal desiccation, often give control. Under overhanging trees, problems from aphid sooty molds on driveways, patios, and walkways are best controlled by hosing the surfaces. Control for aphid galls or leaf distortion on deciduous trees can be problematic, sometimes requiring the winter application of a dormant oil to kill overwintering eggs. Ultimately, elimination of the tree to may be required to solve problem, so tree species in yards should be carefully selected and placed, in view of their potential aphid pests.

See Also the Following Articles

Ants • Biological Control • Rostrum • Sternorrhyncha

Further Reading

Blackman, R. L., and Eastop, V. F. (1984). "Aphids on the World's Crops: An Identification Guide." Wiley, New York.

Blackman, R. L., and Eastop, V. F. (1994). "Aphids on the World's Trees: An Identification and Information Guide." CAB International, Wallingford, U.K.

Dixon, A. F. G. (1985). "Aphid Ecology." Blackie & Son, Glasgow, U.K.

Minks, A. K., and Harrewijn, P. (1987). "Aphids, Their Biology, Natural Enemies and Control," Vols. A, B , C. Elsevier, Amsterdam.

Moran, N. (1992). The evolution of aphid life cycles. *Annu. Rev. Entomol.* **37,** 321–348.

Remaudiere, G., and Remaudiere, M. (1997). "Catalogue of the World's Aphididae." INRA Editions, Paris.

Stern, D. L. (1995). Phylogenetic evidence that aphids, rather than plants, determine gall morphology. *Proc. R. Soc. Lond. B.* **260,** 85–89.

Stern, D. L., and Foster, W. A. (1996). The evolution of soldiers in aphids. *Biol. Rev. Cambridge Philos. Soc.* **71,** 27–79.

Apis Species

(Honey Bees)

Eva Crane

International Bee Research Association

Honey bees (genus *Apis*) are social insects in the family Apidae, order Hymenoptera; they are among the Aculeata (i.e., those having stingers). They evolved after the separation of the Americas and Australia from Eurasia/Africa and are native only in the Old World. The genus *Apis* probably first appeared in the Eocene, about 55 mya. Tropical species *A. dorsata* and *A. florea* existed by the end of the Oligocene 25 mya, and cavity-nesting *A. mellifera* and *A. cerana,* which can also live outside the tropics, were separate species by the end of the Pliocene about 2 mya. Therefore, the highly advanced cavity-nesting species have existed only perhaps a tenth as long as the open-nesting species, which were confined to the warmer tropics. The most important species to humans is *A. mellifera,* which has been introduced all over the world for use in beekeeping.

THE GENUS *APIS*

Known Species

The genus *Apis* contains 11 known species. *A. mellifera* (Fig. 1) is the source of most of the world's honey. It is native throughout Africa, the Middle East, and Europe except for the far north regions. All other *Apis* species are native to Asia. *A. cerana,* which is kept in hives in the temperate zone as well as the tropics, is smaller than *A. mellifera,* and it makes smaller colonies. Other Asian species that build a multiple-comb nest in a cavity are *A. koschevnikovi* and *A. nuluensis* reported in Borneo, and *A. nigrocincta* in Sulawesi.

Other *Apis* species native in parts of the Asian tropics build a single-comb nest in the open. The most important to humans is *A. dorsata,* a bee much larger than *A. cerana. A. laboriosa,* which is even larger, lives in parts of the Himalayas too high for *A. dorsata.* Much smaller than *A. cerana, A. florea* is widespread below around 500 m and can live in drier areas than *A. dorsata.*

Mating, and How Reproductive Isolation Is Achieved

Honey bees mate in flight; the process has been studied in detail in *A. mellifera,* and involves three stages. A queen flies out when only a few days old, and drones that are flying in the area, attracted by the pheromones she produces, follow her. If a drone succeeds in clasping the queen with his legs,

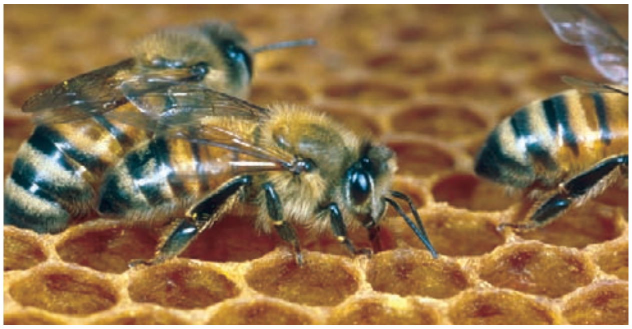

FIGURE 1 Worker honey bees *(Apis mellifera)* on honeycomb. (Photograph courtesy of P. Kirk Visscher.)

his endophallus is everted and mating occurs. When they separate, part of his genitalia remains in the queen, and he falls away and dies. She may mate more than once (usually on the same flight), and the semen she receives is stored in her spermatheca for use throughout her egg-laying life.

The main component of the pheromone attracting drones to the queen seems to be the same for all *Apis* species (9-oxo-*trans*-decenoic acid). In an area with more than one species, reproductive isolation can be achieved if the drones of different species fly at different times of day.

APIS MELLIFERA

Colony Life

The reproduction of individual bees takes place in the colony, and each colony normally contains a single mated female (the queen), many nonreproductive females (workers) and, during the reproductive season, a smaller number of reproductive males (drones). Colonies reproduce by swarming during a season when much food is available. The workers rear several young queens, each in a special queen cell. The old queen and perhaps half the workers of the colony leave as a swarm, which finds a new nest site. One of the young queens mates and heads the parent colony; the others are killed.

Many aspects of the beekeeping cycle and social behavior of honey bee colonies have been studied in detail (see Further Reading).

In the tropics, temperatures are never too low for plants to flower or for bees to fly, and colony activity is governed by rainfall rather than temperature. There are two seasonal cycles in the year, so colonies do not grow as large, or store as much honey, as they do in temperate zones. If the stores of a colony of *A. mellifera* become low in a dearth period, the colony may leave its hive and fly to a nearby area where plants are starting to bloom, rebuilding its combs in a nest site there. Such movements are referred to as absconding or migration, and preventing them is an important part of beekeeping in tropical Africa.

Subspecies and Their Distribution

During the Ice Ages, geographical features in Europe such as mountains confined *A. mellifera* to several separate areas, where they diversified into a number of subspecies or races. The most important in world beekeeping, and their native areas, are *A. mellifera ligustica* (Italian) in northwestern Italy south of the Alps, *A. mellifera carnica* (Carnolian) in the eastern Alps and parts of the Balkans, *A. mellifera caucasica* (Caucasian) in Georgia and the Caucasus mountains between the Black Sea and the Caspian Sea, and also *A. mellifera mellifera* north of the Alps.

The first introductions of *A. mellifera* from Europe to new continents, after 1600, enabled future beekeeping industries to build up and flourish in many countries. Some of the subsequent introductions of *A. mellifera* carried diseases or parasites not previously present, and these caused much damage. From the late 1800s, after the movable-frame hive was devised, there was great interest in breeding more productive honey bees, and colonies of many races were transported from the Old World to other continents. Italian bees, especially, could store much honey in warm regions with consistently good nectar flows. During the 1900s, scientists introduced exotic species and races of honey bees into Europe for experimental purposes, but none is known to have survived in the wild.

Moving honey bees to new areas in tropical or subtropical environments can have wide-reaching effects. In 1956 a number of honey bee queens were transported from southern Africa to Brazil in an attempt to improve the beekeeping in that South American country, where bees of European origin performed poorly. Through an accident, a few of the African queens escaped with swarms, and this led to hydridization with bees of European origin. The consequent “Africanized” bees had characteristics that enabled them to become dominant over the “European” bees already in the American tropics, and they spread rapidly, reaching the Amazon by 1971, the north coast by 1977, Mexico by 1986, and then several southern U.S. states.

In warm regions, many native plants may be pollinated by small bees (Apoidea) whose populations are reduced if colonies of the larger *A. mellifera* are introduced, which in turn can endanger the reproduction of such native plants. This problem has been reported in Australia and Brazil.

A. mellifera is now used in beekeeping in almost every country in the world.

APIS CERANA AND RELATED SPECIES

Of the subspecies of *A. cerana,* the Asian hive bee, *A. cerana indica* is present from Yunnan in China through India to the Philippines. *A. cerana cerana* is in much of China, also the Himalayas, Afghanistan, and the Russian Far East, and *A. cerana japonica* in Tsushima Island and Japan.

After *A. mellifera* was introduced in eastern Asia, *A. cerana* became restricted to areas with native flora. Then in 1985–1986 colonies of *A. cerana* were taken from one of the Indonesian islands to Irian Jaya, also Indonesian but part of New Guinea. The bees reached Papua New Guinea by 1987 and islands in the Torres Strait by 1993. By 2000 they were found (and killed) in Brisbane, Australia, and strenuous efforts are being taken to prevent any further entry and spread of this bee in Australia because it would probably carry the varroa mite, a pest that is of serious economic importance to beekeepers.

APIS DORSATA AND RELATED SPECIES

The large single-comb nests of the giant honey bee, *A. dorsata,* built in the open, are still the most important source of honey in India and some other tropical Asian countries. *A. dorsata* is present in most of the Indo-Malayan region, from the Indus River in the west to the eastern end of the

Indonesian chain of islands, and from the Himalayas to Java in the south. *A. breviligula* is in the Philippines, and *A. binghami* in the Celebes. In the high Himalayas, *A. laboriosa*—a species even larger than *A. dorsata*—nests up to 3000 m, whereas *A. dorsata* rarely nests above 1250 m.

APIS FLOREA AND RELATED SPECIES

The area of the smaller *A. florea* extends as far northwest as Iran. It has also been reported around the Persian Gulf in Iran and Iraq, and in the Arabian peninsula. It reached parts of this last area, and also Sudan, by the aid of humans, and people may also have facilitated its spread along the coast west of the Indus Valley. It is characteristically found in hot dry areas at altitudes below 500 m; in some localities it is the only honey bee that could survive. Its eastern range does not extend as far as that of *A. dorsata,* possibly because *A. florea* could not cross wide sea channels.

A. andreniformis, rather similar to *A. florea,* occurs in southern China, Myanmar (Burma), Palawan in the Philippines, Thailand, Indonesia, Laos, and Vietnam. It is likely that some early statements about *A. florea* in these areas refer instead to *A. andreniformis.*

See Also the Following Articles

Beekeeping • *Caste* • *Hymenoptera* • *Neotropical African Bees* • *Pollination and Pollinators*

Further Reading

Crane, E. (1990). "Bees and Beekeeping: Science, Practice and World Resources." Heinemann Newnes, Oxford, U.K.

Crane, E. (1991). "The World History of Beekeeping and Honey Hunting." Duckworth, London.

Graham, J. M. (ed.) (1992). "The Hive and the Honey Bee." rev. ed. Dadant and Sons, Hamilton, IL.

Hepburn, H. R., and Radloff, S. E. (1998). "Honeybees of Africa." Springer-Verlag, Berlin.

Michener, C. D. (2000). "The Bees of the World." John Hopkins University Press, Baltimore, MD.

Otis, G. W. (1991). A review of the diversity of species within *Apis. In* "Diversity in the Genus *Apis*." (D. R. Smith, ed.), Ch. 2. Westview Press, Boulder, CO.

Rinderer, T. E. (ed.) (1986). "Bee Genetics and Breeding." Academic Press, New York.

Ruttner, F. (1988). "Biogeography and Taxonomy of Honeybees." Springer-Verlag, Berlin.

Aposematic Coloration

Mathieu Joron

Leiden University

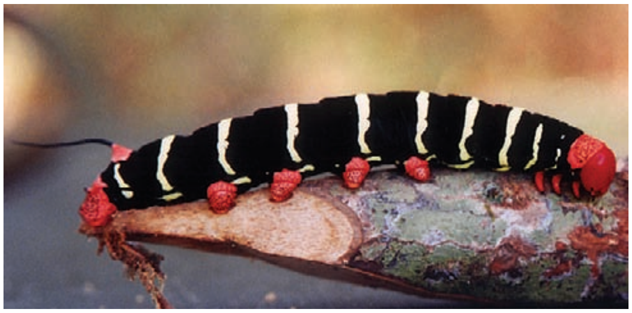

FIGURE 1 *Pseudosphinx tetrio* hawk moth caterpillar from the Peruvian Amazon, showing a combination of red and black, classical colors used by aposematic insects. These larvae feed on toxic latex-sapped trees in the Apocynaceae. Length 14 cm. (Photograph © M. Joron, 1999.)

Insects attract collectors' attention because they are extremely diverse and often bear spectacular colors. To biologists, however, bright coloration has been a constantly renewed puzzle because it makes an insect a highly conspicuous prey to prospective predators. Charles Darwin understood that bright colors or exaggerated morphologies could evolve via sexual selection. However, he felt sexual selection could not account for the conspicuous color pattern of nonreproductive larvae in, for example, *Pseudosphinx* hawk moth caterpillars (Fig. 1). In a reply to Darwin about this puzzle, Alfred R. Wallace proposed that bright colors could advertise the unpalatability of the caterpillars to experienced predators. Indeed, prey that are not edible to predators are predicted to gain by exhibiting conspicuous and very recognizable colors; experienced predators can then correctly identify and subsequently avoid attacking such prey. E. B. Poulton later developed this idea, expanded it to other warning signals (i.e., sounds or smells), and coined the term "aposematism" to describe this phenomenon (from the Greek "away" and "sign").

Aposematic color patterns are found everywhere throughout the insects, from black- and yellow-striped stinging wasps to black and red, bitter-tasting lady beetles, or brightly colored, poisonous tropical butterflies. Although warning coloration has involved fascination, empirical and theoretical studies for some time, the puzzle of aposematism still motivates much debate today. First, although there is little doubt that bright coloration is often an antipredatory strategy, how aposematism evolves is far from clear. This is because brightly colored mutants in a population of cryptic (camouflaged) prey are more exposed to predators. How can a warning coloration evolve in a prey if the very first mutants exhibiting such coloration in the population are selected against? Second, the reasons for the brightness and conspicuousness of warning colors are not always clear and may be multiple. Are aposematic colors "road signs" that help predators learn better to differentiate inedible from edible prey, or are bright colors more easily memorized and associated to bad taste by predators? Did yellow and red colors, often borne by poisonous insects, evolve because of innate biases against these colors in the predators' brains, or are more complex cognitive, behavioral, frequency-dependent, or coevolutionary mechanisms involved in the evolution of warning patterns?

Finally, why are warning patterns highly diverse in the insect world, whereas all toxic prey would gain by bearing the same color, thus reducing the probability of being sampled by a naïve predator?

WHAT TO ADVERTISE

"Aposematism is quite simply the correlation between conspicuous signals, such as bright coloration, and prey unprofitability," Candy Rowe wrote in 2001. But why should some prey become unprofitable in the first place, while others do not?

Unprofitability is difficult to define, and even more difficult to measure. It is certainly contextually defined, because the propensity of an animal to eat something is highly dependent on its level of hunger and its ability to use the prey for energy once eaten. Palatability (i.e., the predator's perception of prey profitability) greatly determines whether the predator will or will not eat the prey. Predator–prey coevolution led predators in part to rely on proximal perception to gauge the prey profitability. In particular, taste sensitivity may well have evolved in predators as an assessment of food toxicity: indeed, predators usually consider toxic chemicals to be distasteful. Some insects have external defenses such as horns, or spines, many of which cause irritation. Such physical defenses may be coupled to venom, as with the irritant hairs of many caterpillars or hymenopteran stingers. These insects may be otherwise perfectly profitable, and some predators evolve ways around the physical defenses, such as bee-eaters that are able to remove a bee's stinger and venom sac. Other insects have passive chemical defenses that predators discover upon consumption, such as chemicals in the hemolymph or sequestration glands of lubber grasshoppers or monarch butterflies. Such insects usually develop extra signals such as powerful smells, at least when handled, to advertise their toxicity before being consumed.

Toxicity is not the only way an insect can be unprofitable to predators. Difficulty in capturing prey (due to fast escape, erratic flight, breakable wings, etc.), or difficulty in handling prey (due to toughness or a hard cuticle) are other ways that insects can bring no net reward to the predators that spend energy chasing them, even if the chase results in the prey being seized. However, multiple unprofitability traits might be important in the evolution of warning signals.

Predators can have three kinds of response to a prey depending on their perception of prey profitability. If consuming a prey leads the predator to be more likely to attack similar prey in the future, perhaps even using the prey's appearance as a search image, the prey is called palatable. In feeding experiments, this usually leads birds to attack nearly 100% of the palatable prey offered. Of course, the predator may satiate after consuming a number of prey, and consequently the propensity to attack may decrease at high prey densities. In contrast, if experienced predators are less likely to attack similar prey, the prey is called unpalatable. Of course, predators' memorizing capacity, and the strength of the prey unpalatability, may all influence how fast information regarding prey is acquired and how long it is retained. However, a distasteful prey will inevitably lower the predators' instantaneous propensity to attack this prey further, an effect analogous to an immediate satiation. Finally, eating the prey may have no effect on the predator's subsequent behavior, which means that the prey is effectively neutral. This category is mainly derived from theory; there is little evidence that it exists in nature.

Variations in unpalatability among prey species, along what is called the "(un)palatability spectrum," affect the rate at which predators modify their behavior with experience. Predator's perceived toxicity is likely to be a sigmoid function of actual toxin concentration per unit prey mass, meaning that little of the palatability spectrum may fall into intermediate perceptions between "unpalatable" and "fully palatable." Although how predators learn is still under debate, experiments and theory suggest that they respond to a large extent to the (perceived) concentration of nasty chemicals they can tolerate per unit time.

The distastefulness of insects is generally linked to the host plants they utilize. Indeed, many distasteful or defended insects are herbivorous; most defended nonherbivorous insects are Hymenoptera. Some plant families, like the Solanaceae and the Passifloraceae, which are hosts to many chemically defended insects, contain alkaloids and cyanogens, respectively, as secondary metabolites. Some insects, like monarch butterflies (*Danaus plexippus*) that feed on *Asclepias* plants (milkweeds, Asclepiadaceae), sequester the compounds of such plants and store them; these insects thus avoid the toxic effects of the toxic compounds altogether. In soft-bodied insects (e.g., larvae), toxins are usually stored near the teguments or in special glands, ready to release their contents upon handling. The toxicity of insects that extract and sequester plant chemical compounds is dependent on the concentration of these compounds in the host plant. Sawfly larvae (Hymenoptera: Tenthredinidae), for example, reflex-bleed drops of hemolymph when touched; the unpalatability of such larvae is shown to be directly dependent on the glucosinolate concentration of their crucifer host plant over 24 h before "bleeding."

Other insects, however, synthesize their toxins *de novo,* like many chrysomelid beetles; they probably use the same enzymatic machinery that serves (or has served, in their ancestors) to detoxify the plant's secondary compounds. Although many of these species still use precursors derived from their food plant, these insects are usually less dependent on the plant's toxicity to develop their own noxious compounds. Some groups like ithomiine or heliconiine butterflies also get toxin precursors in their adult diet.

Whatever route to distastefulness is taken, we observe a general correlation between clades of distasteful insects and toxicity in host plant families. In butterflies, the distasteful Troidinae (Papilionidae) tend to feed on Aristolochiaceae, monarchs (Nymphalidae: Danainae) usually feed on milkweeds (Asclepiadaceae), longwing butterflies (Nymphalidae: Heliconiinae) feed on Passifloraceae, and clearwings (Nymphalidae: Ithomiinae) mainly on Solanaceae and Apocynaceae. In contrast, butterfly clades feeding on chemical-free monocotyledonous plants, like browns (Nymphalidae: Satyrinae) on grasses, or owl-butterflies (Nymphalidae: Brassolinae) on palms or Marantaceae, did not evolve distastefulness. Thus toxicity in insects may frequently have evolved as a mere by-product of adaptation to utilize new kinds of food, particularly toxic plants. The costs of detoxification or toxin production could be covered by the benefits of invading competition-free hosts, perhaps assisted by the increased survival afforded by chemical protection.

DISGUSTING, BRIGHT, SIMPLE, AND CONTRASTED: WHY AND HOW TO ADVERTISE

Why should unprofitable prey advertise? Instead of parading with gaudy colors, why should all prey not try to escape predators' detection altogether through camouflage? Although the initial steps to aposematism are not obvious, the advantage of aposematic signals once established is clear. Indeed, numerous studies have shown that most predators are able to learn and recognize, and subsequently avoid, prey they associate with a bad experience.

Some distasteful prey, such as the transparent ithomiine butterflies found in the forest understory in tropical America, are not particularly conspicuous. The rampant mimicry found in this group of inconspicuous butterflies demonstrates that predators are able to learn and avoid such prey (although other stimuli, e.g., motion, might also be important). Still, most distasteful insects are brightly colored. Why should aposematic signals usually be conspicuous, and use simple color patterns of red, yellow, or black? Most of the answer is likely to be found in the cognitive behavior of the predators that selected for such colors. Several hypotheses have been put forward to explain the correlation between bright colors and unprofitability in insects. Bright contrasted colors are thought to be (1) easier to learn, (2) more difficult to forget, and (3) as different as possible from edible prey, thereby facilitating the avoidance of recognition errors. All these mechanisms are supported by experimental data to some extent (e.g., Fig. 2). Because both predators and aposematic prey benefit from correct identification, aposematic colors are believed to take advantage of any bias in the predator's cognition system. Likewise, predators in turn gain in being biased in the same direction as that taken by the prey. Therefore, prey signaling and predators' cognition are likely to have coevolved, which, incidentally, makes experimental evidence for any of the foregoing hypotheses generally hard to establish independently. Hypothesis 3 is the most likely to involve interactions between perception and cognition in the predators, leading to fast evolution of the prey's signals.

Many aposematic insects simultaneously send signals of different kinds, and some argue that such "multimodal" warning signals may reveal unconditioned biases that are absent when each sensory modality is examined alone. Assuming that predators would rely solely on color and not behavior, motion, or sounds is perhaps simplistic, and it is sometimes argued that multiple signals could even be a prerequisite for the evolution of warning coloration. In fact, the reason for the apparent importance of multimodality probably lies again in the coevolutionary history of predators and their prey, which shapes innate biases. Predators are generally good entomologists for the potential prey they encounter often, and predators integrate various sensory modalities to make decisions regarding a particular action. Most aposematic insects are mimicked by edible species (Batesian mimics) that parasitize the warning function of the signal. The presence of these Batesian mimics reduces the reliability of the warning signal and means lost prey for the predators. Model species may therefore escape being mimicked by evolving new dimensions for signaling, in addition to the established one (i.e., in different sensory dimensions).

Many warningly colored insects live in groups, which enhances the warning function of their signal for three reasons (Fig. 3). First, predators tend to associate and retain noxiousness

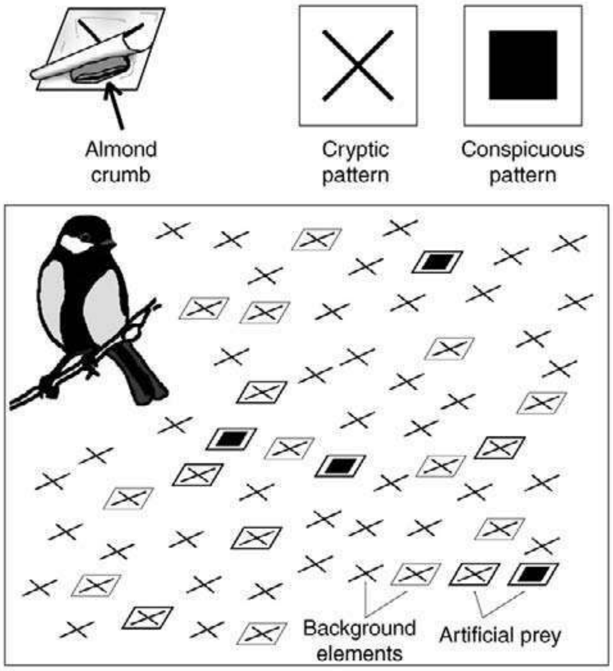

FIGURE 2 Schematic outline of the "novel world" experimental design developed by R. Alatalo and J. Mappes (University of Jyväskylä, Finland) to study the mechanisms of predator's selection on conspicuous prey. Great tits, *Parus major,* are trained to forage in a room covered with small black symbols (e.g., crosses) on a white background, some of which are actual prey. Almond crumbs are placed between two 1-cm^2 pieces of paper glued together that bear a symbol on the outside. Black-squared prey items stand out conspicuously on the black-crossed background and represent potential warning signals, whereas black-crossed prey items are cryptic. The novelty of all symbols ensures that innate or previously learned prey recognition does not interfere with the predator's response during the experiment. This setup also partly resolves one drawback of garden experiments, where the local food abundance for predators is artificially increased, making the searching costs, search images, and other predatory behavior unrealistic. By playing on the palatability of the prey items, it is possible to monitor how the birds learn to avoid the conspicuous signal. Mimics can also be incorporated in the environment at varying frequencies to study the dynamics of Batesian mimicry.

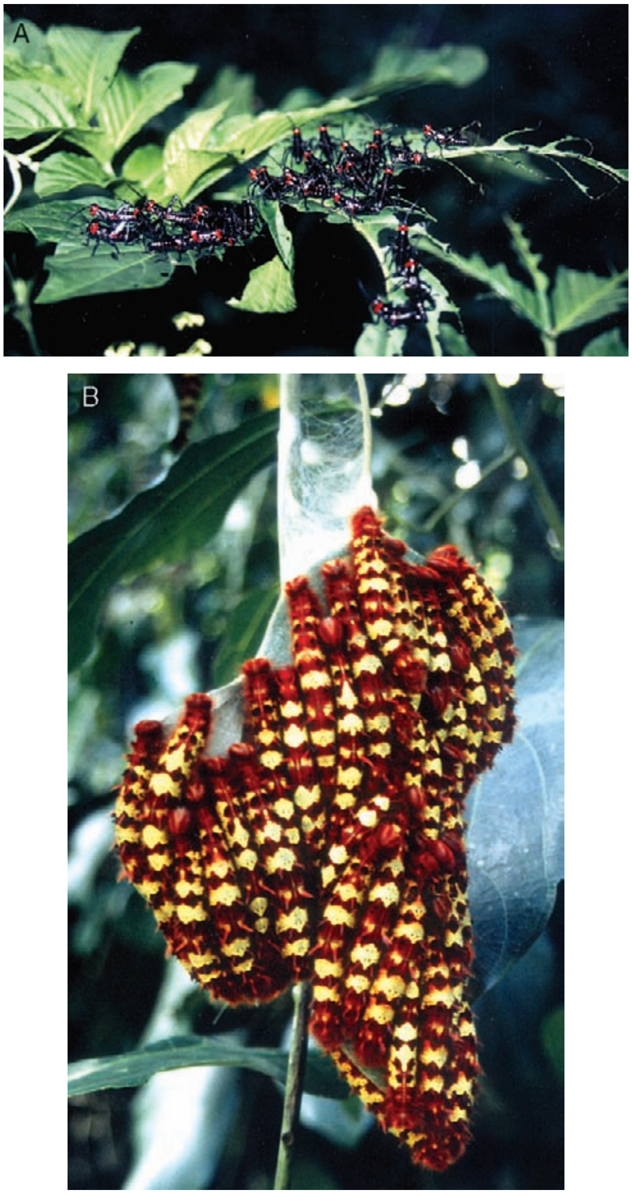

FIGURE 3 Examples of gregarious warningly colored insects. (A) Gregarious *Chromacris* lubber grasshoppers nymphs (Orthoptera: Romaleidae) feeding on a toxic *Solanum mite*. Although it is not a bright color, black is often used as a warning color by insects, presumably because it increases conspicuousness by contrast against green foliage. (B) Gregarious *Morpho* sp. caterpillars spending the day in a dense cluster. They disperse at night for feeding. Clusters of aposematic prey often create an emergent, enhanced pattern presumably perceived as a supernormal stimulus by the predators and therefore better memorized. Caterpillars about 10 cm long. (Photographs from the Peruvian Amazon, © M. Joron, 1997.)

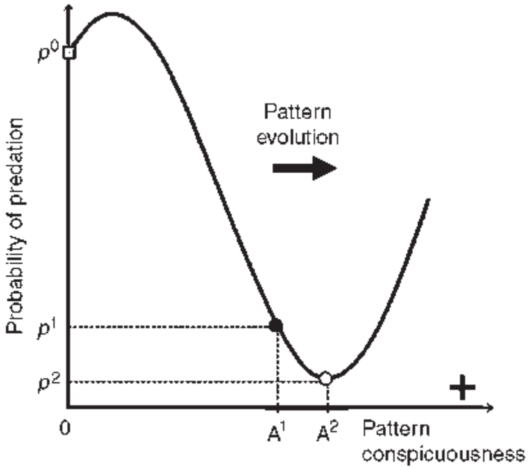

FIGURE 4 Peak shift mechanism applied to prey conspicuousness. The curve describes a fitness function that relates pattern conspicuousness to the probability of suffering predation. Predators are assumed to have knowledge of prey ranging from crypsis (open square: zero conspicuousness, tastefulness, high probability of predation) to aposematism (solid circle: high conspicuousness, A^1; distastefulness, low probability of predation, p^1). From this knowledge, predators extrapolate the palatability of prey with higher conspicuousness they have not yet experienced, hence the curve beyond the solid circle. In particular, slightly more conspicuous prey (open circle) elicit a stronger rejection response than the known aposematic prey (solid circle), and therefore suffer an even lower probability of predation, p^2. Consequently, selection should cause the conspicuousness of the distasteful prey to shift gradually from A^1 to A^2. In contrast, cryptic distasteful prey cannot drift away from crypsis (open square) toward higher levels of conspicuousness because they are more likely to be preyed upon when they become more conspicuous. An initial phenotypic jump is necessary to get to levels of conspicuousness where birds can categorize the prey as warningly colored, avoiding confusion with the normal staple of cryptic tasteful prey.

and a warning signal more quickly when presented with multiple copies of the same signal. Second, all unpalatable prey get an individual advantage in living in groups by the immediate avoidance, similar to the immediate satiation discussed earlier, caused in the individual predator that avoids the group altogether. Third, at the population level, clustering spatially also reduces the number of individual predators the population of prey is exposed to and has to educate, which again enhances the benefit of signaling. In short, it pays to be sitting right next to the toxic individual that is being sampled by a predator, because this is where the probability of predation is lowest, hence the advantage to living in groups. However, many solitary prey also exhibit warning coloration, and gregariousness does not necessarily evolve prior to aposematism.

One common mechanism leading to fast divergence in signals throughout the animals, known as peak shift, hinges here on the coupling of prey coloration, predator experience, and predator innate aversion. Prey can be placed on a conspicuousness axis (Fig. 4), with cryptic edible prey at one end (close to 0 conspicuousness) and incipient aposematic prey or new color pattern mutants at some distance down the axis. More conspicuous prey usually elicit stronger aversion in educated predators, which extrapolate the idea that stronger conspicuousness should mean stronger noxiousness. Therefore, stronger signals (away from edible prey appearance) lead to a supernormal response in the predators that thereby select for increased conspicuousness in the prey. This mechanism is a special case of a runaway process and could be an important route to the evolution of aposematic prey that are bright and contrasted. It is thought that the coupling of such cognitive biases with the ability to learn leads to the selection of more strongly exaggerated warning colors and patterns in noxious prey than is expected in purely nonlearning predators.

EVOLUTIONARY ROUTES TO APOSEMATIC COLORATION

The Problems

There are obvious benefits to bearing warning colors in a population of warning-colored prey. As noted by early naturalists like A. R. Wallace and later E. B. Poulton, experienced predators avoid warningly colored prey, and presumably the number of prey killed during the predators' education is lower than in the absence of signaling. These benefits are clear at the group level but are not so clear at the individual level, because the first warningly colored individuals in a population of cryptic (and noxious) prey suffer strongly increased predation. Indeed, novel warningly colored prey not only suffer increased detection by prospective predators, but also elicit *no* avoidance in the predators. Consequently, there is strong positive frequency dependence, putting novel rare warning signals at a disproportionate disadvantage against an established strategy (crypsis, or another already established warning signal).

How could warning signals evolve at all if the first mutants using this strategy are killed? Laboratory experiments using the "novel world" design (Fig. 2) show rather unequivocally that aposematic patterns cannot evolve gradually in unpalatable prey. Indeed, small increases in visibility in cryptic prey increased attack rates without enhancing learning. Similarly, deviant phenotypes in established warning patterns suffered stronger predation. Finally, rare conspicuous prey suffered disproportionate predation, even when presented in groups. Therefore, a gradual increase in conspicuousness towards aposematism seems unlikely. This means that the evolving population must undergo a sudden jump, both in phenotype (to get a pattern that predators categorize as a different item) and in numbers beyond a threshold frequency (to allow the local predators to learn about the new pattern). Once the new pattern has achieved the minimum frequency and phenotypic thresholds, positive frequency dependence helps the new mutant to spread in the population. Peak shift or other processes can then occur, increasing the conspicuousness or adding other components to the signal. How can these evolutionary leaps be achieved—or circumvented—by an incipient aposematic prey?

Deterministic Evolution via Immediate Benefits

NEOPHOBIA A new aposematic form could in theory escape the disadvantage of being rare and novel by causing immediate avoidance without having to be tasted at all by the predator. Indeed, predators are somewhat reluctant to sample novel-looking prey, particularly if novelty is associated with bright colors. This phenomenon is called "neophobia," a kind of diet conservatism in predators. Neophobia could arise from various foraging biases, such as the formation of search images in the brains of predators as they search for edible-looking prey and ignore other prey, or via cultural inheritance, as with nestlings that tend to prey upon what they were fed by their parents. Neophobia is sometimes presented as a potential route toward aposematism. However, it does not really resolve the frequency dependence problem, because it is essentially a transient phenomenon involving no information acquisition by predators. Therefore as soon as numbers grow, however slightly, neophobia tends to vanish. Neophobia should best be classified as a predator's bias, like other innate biases against colors, smells, or sound, evolved by predators in response to their prey environment. Such biases are likely to channel the ultimate form taken by the aposematic signal (to the benefit of both preys and predators), but it is unlikely that they cause its evolution in the first place.

INDIVIDUAL ADVANTAGE One obvious way around initial obstacles is not to be killed by predators' attacks. Then, prey could both educate the predators and be avoided in subsequent encounters. Indeed, most birds taste-test their prey before ingesting them, and many aposematic prey have noxious compounds in their outer parts, making it possible to be tasted but not injured by predators. For instance, ithomiine and danaine butterflies concentrate alkaloid in their wings. Day-flying pericopine moths let a voluminous and bitter hemolymph froth out of their body, likely tasted (or smelled) by a predator before it has profoundly injured the moth. Moreover, most unpalatable butterflies have very elastic bodies, which resist crushing. Strong smells that predators take as a warning for bad taste or toxicity, like those of stinkbugs, are another way by which prey can gain immediate advantage without having to be effectively tasted by the predators.

PREY ALREADY CONSPICUOUS Another way by which prey can overcome the difficulty of evolving conspicuous color is not to suffer any cost (i.e., avoid the necessity of a phenotypic leap) as a result of increased conspicuousness. Indeed, most flying insects are rather conspicuous in flight and rely on their difficulty of capture to escape predation. They may not suffer any cost to bearing conspicuous colors, and indeed many butterflies, if not most, irrespective of their palatability, display bright patches of colors on the upper side of their wings, visible in flight, while having cryptic underwings making them inconspicuous when sitting. Such bright dorsal colors might initially evolve as sexual signals in male–male or male–female interactions long before unpalatability evolves. Once noxiousness has evolved, predators can learn an already conspicuous pattern without making recognition errors because of the resemblance to the palatable prey they have as search images. In a way, conspicuous flying insects can be said to be "preadapted" to evolve warning colors. But such patterns can then also change or drift according to predators' biases. In particular, already bright color patterns can be enhanced toward brighter coloration through processes like peak shift, as described earlier. According to James Mallet, examples of this mechanism are the unpalatable *Taenaris* and *Hyantis* (Nymphalidae: Morphinae), which have evolved strikingly conspicuous warning spots via the enhancement of some of the less conspicuous eyespots that are still found on the undersides of their palatable relatives, the well-known blue *Morpho* butterflies.

MÜLLERIAN MIMICRY The easiest way to avoid the cost of rarity and conspicuousness altogether is to jump to an aposematic pattern already present in the habitat and known by the local predators. The shared appearance between several defended prey is called Müllerian mimicry, and it is likely that most aposematic species evolved via this route. Indeed, mimicry rings usually include a large number of Müllerian species (all of which are noxious). Of these, only one evolved the pattern first, followed by the other species that colonized an already protected pattern. This pattern of evolution is detectable by examining the biogeography and phylogeny of the species in question. For example, *Heliconius erato* and *H. melpomene* are Müllerian mimics throughout their distribution range. However, the *H. melpomene* was shown to have much younger color pattern races, with a clearly distinct genealogy, than *H. erato,* suggesting that *H. melpomene* is a Müllerian mimic that adopted the established color patterns of *H. erato.*

Population Processes: Kin Selection, Drift, and Shifting Balance

Because many unpalatable prey are indeed gregarious, it is easy to conclude that gregariousness allows the evolution of aposematism. The evolution of aposematism through gregariousness relies on the predator rejecting the whole group after sampling only one or few individuals. This extrapolation from one prey to the whole group is analogous to a superfast learning in the predator, which can be enhanced by conspicuousness. However, it also pays for aposematic preys to live in groups, thereby increasing their apparent density to the local predators. It is therefore not clear whether gregariousness or aposematism should evolve first to trigger the evolution of the other. Groups of gregarious larvae (Fig. 3) are usually family groups, suggesting that kinship might allow a new mutation quickly to get to a locally high frequency in such little-dispersing insects through kin selection. However, one should be aware that relatedness per se is not what favors the local rise in frequency of the gene here, but simply the local founding event by one or few family groups.

In fact, many adult aposematic adult insects are either not gregarious at all or do not aggregate in family groups. Besides, some of the most gregarious insect larvae come from the joint oviposition of several unrelated females. Although these examples could have arisen after the initial evolution of warning color through kin selection, it is more parsimonious to infer that non-kin-selection arguments can also explain the evolution of aposematism. Drift alone, particularly, followed by positive frequency dependence, is a good candidate mechanism (and in fact kin founding is only a special case of genetic drift). Indeed, when the ratio of predators to prey decreases in a locality, selection for antipredatory strategies is greatly diminished, allowing the exploration of other color pattern possibilities by the local population. Using release–recapture techniques of different warningly colored forms of *H. cydno* in Ecuador, D. Kapan showed that selection was relaxed when the butterflies were released in larger numbers. Therefore, the prey population could move via genetic drift above the required threshold, after which the new warning color invades the population. Positive frequency dependence has the interesting property that although it hinders the initial evolution of new patterns, it hinders the removal of any pattern once it has been established. If genetic drift in prey populations matches the fluctuations of selection pressures in time and space, new local aposematic patterns can be established frequently in different locations. These are essentially the first and second steps of the shifting balance theory of S. Wright. Competition between geographically adjacent warning color types then allows one pattern to spread to neighboring populations, like the traveling waves of color races documented in South America for *H. erato* or *H. melpomene.*

CONCLUSIONS

Despite the advantages of bearing a warning coloration established in the locality, the evolution of aposematism is not straightforward because proximal mechanisms seem to represent obstacles to its initial evolution. However, aposematic patterns are extremely diverse at all geographical and taxonomic levels, and this major discrepancy between theory and nature clearly suggests that positive frequency-dependent arguments are not as restrictive against the rise of novel warning colors. Similarly, predator generalization, which should not allow gradual shift of cryptic prey toward bright warning colors, does not seem to be efficient in restricting the rise of new conspicuous patterns.

In fact, both population dynamics and psychological arguments might well explain such spectacular diversification. First, positive frequency dependence would allow new local forms to be established through drift, relayed by other processes involving predator's cognitive biases. Second, the initial steps toward warning color are determined largely by which cognitive biases in the predators are exploited. That is, the initial pathway taken toward the evolution of warning coloration probably profoundly affects the aposematic phenotype that eventually evolves. Similarly, positive frequency dependence prevents deviations from the evolutionary pathway that is taken. In short, although aposematism is not expected predictably to evolve via Fisherian selection, it is such a powerful strategy once evolved that it is possibly inevitable in a contingent and varying world, where the nature and the height of the initial obstacles to its evolution fluctuate. It may thus follow a ratchetlike pattern of evolution, where more routes may lead toward aposematism than routes away from it.

See Also the Following Articles

Chemical Defense • Crypsis • Mimicry • Monarchs

Further Reading

Alatalo, R. V., and Mappes, J. (1996). Tracking the evolution of warning signals. *Nature* **382,** 708–710.

Edmunds, M. (1974). "Defence in Animals. A Survey of Anti-predator Defences." Longman, New York.

Endler, J. A. (1988). Frequency-dependent predation, crypsis, and aposematic coloration. *Philos. Trans. R. Soc. Lond. B* **319,** 505–524.

Guilford, T. (1988). The evolution of conspicuous coloration. *Am. Nat.* **131,** S7–S21.

Lindström, L., Alatalo, R. V., Lyytinen, A., and Mappes, J. (2001). Strong antiapostatic selection against novel rare aposematic prey. *Proc. Nat. Acad. Sci. U.S.A.* **98,** 9181–9184.

Mallet, J., and Joron, M. (1999). Evolution of diversity in warning color and mimicry: Polymorphisms, shifting balance and speciation. *Annu. Rev. Ecolo. System.* **30,** 201–233.

Mallet, J., and Singer, M. C. (1987). Individual selection, kin selection, and the shifting balance in the evolution of warning colors: The evidence from butterflies. *Biol. J. Linn. Soc.* **32,** 337–350.

Poulton, E. B. (1890). "The Colours of Animals." Trübner, London.

Rowe, C. (ed.). (2001). Warning signals and mimicry. Special issue of *Evolutionary Ecology* [1999, vol. 13, no 7/8]. Kluwer, Dordrecht, The Netherlands.

Sillén-Tullberg, B. (1988). Evolution of gregariousness in aposematic butterfly larvae: A phylogenetic analysis. *Evolution* **42,** 293–305.

Sword, G. A., Simpson, S. J., El Hadi, O. T. M., and Wilps, H. (2000). Density dependent aposematism in the desert locust. *Proc. R. Soc. Lond. B Biol. Sci.* **267,** 63–68.

Wallace, A. R. (1879). The protective colours of animals. *In* "Science for All" (R. Brown, ed.), pp. 128–137. Cassell, Petter, Galpin., London.

Apterygota

Apterygota is a subclass of the class Insecta in the phylum Arthropoda. It contains two orders, the Archaeognatha and the Thysanura.

Aquatic Habitats

Richard W. Merritt
Michigan State University

J. Bruce Wallace
University of Georgia

Less than 3% of the world's total water occurs on land, and most of this is frozen in polar ice caps. Streams and rivers are one of the more conspicuous features of the landscape; however, their total area is about 0.1% of the land surface, whereas lakes represent about 1.8% of total land surface. Some authors have questioned whether insects have been successful in water because aquatic species represent only a small portion of the total hexapod fauna. However, 13 orders of insects contain species with aquatic or semiaquatic stages, and in five of these (Ephemeroptera, Odonata, Plecoptera, Megaloptera, and Trichoptera) all species are aquatic with few exceptions (Table I). Few aquatic insects spend all of their life in water; generally any insect that lives in water during a portion of its development is considered to be "aquatic." Usually, but not always, for most "aquatic" species, it is the larval stage that develops in aquatic habitats, and the adults are terrestrial (Table I). The pupae of some taxa undergoing complete metamorphosis (i.e., holometabolous) remain within the aquatic habitat; in others the last larval instar moves onto land to pupate, providing the transition stage from the aquatic larva to the terrestrial adult.

The success of insects in freshwater environments is demonstrated by their diversity and abundance, broad distribution, and their ability to exploit most types of aquatic habitat. Some species have adapted to very restricted habitats and often have life cycles, morphological, and physiological adaptations that allow them to cope with the challenges presented by aquatic habitats. One aquatic environment in which insects have not been very successful is saltwater habitats, although some 14 orders and 1400 species of insects occur in brackish and marine habitats; only one group occurs in the open ocean. One of the most widely accepted attempts to explain why more insects do not live in marine environments is that successful resident marine invertebrates evolved long before aquatic insects and occupy many of the same niches inhabited by freshwater insects. Thus, marine invertebrates, such as crustaceans, have barred many insects from

TABLE I Occurrence of Life Stages in Major Habitat Types for Aquatic and Semiaquatic Representatives of Insect Orders (A, adult; L, larvae; P, pupae)

Order	Terrestrial	Freshwater
Collembola	A, L	A, L
Ephemeroptera	A	L
Odonata	A	L
Heteroptera	A	A, L
Orthoptera	A, L	A, L
Plecoptera	A	L
Coleoptera	A, L, P	A, L
Diptera	A, P	L, P
Hymenoptera	A	A, L, P
Lepidoptera	A	L, P
Megaloptera	A, P	L
Neuroptera	A, P	L
Trichoptera	A	L, P

Modified from Ward, J. V. (1992). "Aquatic Insect Ecology," Vol. 1, "Biology and Habitat." Wiley, New York.

marine habitats by competitive exclusion. Problems with osmoregulation have been given as another reason for the paucity of saltwater species; however, one of the two multicellular animals found in the Great Salt Lake is a member of the order Diptera (see later: Unusual Habitats), providing evidence that some insects display a strong ability to osmoregulate.

The first aquatic insects are believed to have inhabited flowing water as early as the Permian and Triassic. It was not until the late Triassic and early Jurassic that evidence of abundant lentic, or still-water, fauna arose, accompanied by rapid diversification of water beetles, aquatic bugs (Heteroptera), and primitive Diptera. On the basis of several lines of evidence including osmoregulation, fossil evidence, secondary invasions to water of many taxa, and great variation in gill structure among and within orders, some authors have suggested that the first insects may have lived in water rather than in terrestrial habitats. However, the general consensus is that an aquatic origin for insects seems unlikely and that aquatic insects may not have shown up until 60 to 70 million years later than their terrestrial counterparts.

Freshwater systems are often divided into standing (lentic) and flowing (lotic) waters. Although such a division is useful for indicating physical and biological differences, habitat diversity can vary tremendously within these two broad categories, and some of the same taxa may be found in both lentic and lotic habitats, depending on the physiological constraints of a given habit. Many factors influence successful colonization of aquatic insects to a given habitat; however, most of these would fall under four broad categories: (1) physiological constraints (e.g., oxygen demands, respiration, osmoregulation, temperature effects), (2) trophic considerations (e.g., food acquisition), (3) physical constraints (e.g., coping with harsh habitats), and (4) biotic interactions (e.g., predation, competition). However, these categories are so interrelated that detailed analysis of each factor separately is very difficult.

HABITAT, HABIT, AND TROPHIC CLASSIFICATION SYSTEM

The classification system used here for lotic and lentic habitats stresses the basic distinction between flowing water (i.e., streams, rivers) and standing water (i.e., ponds, lakes, swamps, marshes) habitats (Table II). This separation is generally useful in describing the specific microhabitats (e.g., sediments, vascular hydrophytes, detritus) in which aquatic insects may be found. Both stream/river currents and lake shoreline waves often create erosional (riffle-type) habitats and may resemble each other in their physical characteristics, whereas river floodplain pools and stream/river backwaters create depositional (pool-type) habitats that may resemble lake habitats as well (Table II). Within a given habitat, the modes by which individuals maintain their location (e.g., clingers on surfaces in fast-flowing water, sprawlers on sand or on surfaces of floating leaves, climbers on stem-type surfaces, burrowers in soft sediments) or move about (e.g., swimmers, divers, surface skaters) have been categorized (Table III). The distribution pattern resulting from habitat selection by a given aquatic insect species reflects the optimal overlap between habit and physical environmental conditions that comprise the habitat, such as bottom type, flow, and turbulence. Because food in aquatic habitats is almost always distributed in a patchy fashion, the match between habitat and habit is maximized in certain locations. This combination will often result in the maximum occurrence of a particular species.

In view of the complex physical environment of streams, it is not surprising that benthic invertebrates have evolved a diverse array of morphological adaptations and behavioral mechanisms for exploiting foods. Throughout this article we will follow the functional classification system originally described by K. W. Cummins in 1973, which is based on the mechanisms used by invertebrates to acquire foods (Table IV). These functional groups are as follows:

- **Shredders,** which are insects and other animals that feed directly on large pieces of organic matter (e.g., decomposing leaves and fragments of wood >1 mm in size) and their associated fungi and bacteria, and convert them into fine particulate organic matter (FPOM) through maceration, defecation, and physical degradation;
- **Collector-filterers,** which have specialized anatomical structures (e.g., setae, mouth brushes, fans, etc.) or silk and silklike secretions that act as sieves to remove fine particulate matter less than 1 mm in diameter from the water column;
- **Collector-gatherers,** which gather food, primarily FPOM, that is deposited within streams or lakes;
- **Scrapers,** which have mouthparts adapted to graze or scrape materials (e.g., periphyton, or attached algae, and the associated microbes) from rock surfaces and organic substrates;
- **Predators,** which feed primarily on other animals by either engulfing their prey or piercing prey and sucking body contents.

These functional feeding groups refer primarily to modes of feeding or the means by which the food is acquired, and the food type per se (Table IV). For example, shredders may select leaves that have been colonized by fungi and bacteria; however, they also ingest attached algal cells, protozoans, and various other components of the fauna along with the leaves.

LOTIC HABITATS

Streams vary greatly in gradient, current velocity, width, depth, flow, sinuosity, cross-sectional area, and substrate type, depending on their position in the landscape with respect to geology, climate, and the basin area they drain. Anyone who has spent much time around or wading in streams is aware that these can be extremely diverse habitats, often manifesting great

TABLE II Aquatic Habitat Classification System

General category	Specific category	Description
Lotic–erosional (running-water riffles)	Sediments	Coarse sediments (cobbles, pebbles, gravel) typical of stream riffles.
	Vascular hydrophytes	Vascular plants growing on (e.g., moss, *Fontipalis*) or among (e.g., pondweed, *Potamogeton pectinatus*) coarse sediments in riffles.
	Detritus	Leaf packs (accumulations of leaf litter and other coarse particulate detritus at leading edge or behind obstructions such as logs or large cobbles and boulders) and debris (e.g., logs, branches) in riffles.
Lotic–depositional (running-water pools and margins)	Sediments	Fine sediments (sand and silt) typical of stream pools and margins.
	Vascular hydrophytes	Vascular plants growing in fine sediments (e.g., *Elodea,* broad-leaved species of *Potamegeton, Ranunculus*).
	Detritus	Leaf litter and other particulate detritus in pools and alcoves (backwaters).
Lentic–limnetic (standing water)	Open water	On the surface or in the water column of lakes, bogs, ponds.
Lentic–littoral (standing water, shallow-water area)	Erosional	Wave-swept shore area of coarse (cobbles, pebbles, gravel) sediments.
	Vascular hydrophytes	Rooted or floating (e.g., duckweed, *Lemna*) aquatic vascular plants (usually with associated macroscopic filamentous algae).
	Emergent zone	Plants of the immediate shore area (e.g., *Typha,* cattail), with most of the leaves above water.
	Floating zone	Rooted plants with large floating leaves (e.g., *Nymphaea,* pond lily), and nonrooted plants (e.g., *Lemna*).
	Submerged zone	Rooted plants with most leaves beneath the surface.
	Sediments	Fine sediments (sand and silt) of the vascular plant beds.
	Sediments	Fine sediments (fine sand, silt, and clay) mixed with organic matter of the deeper basins of lakes. (This is the only category of "lentic–profundal.")
Lentic–profundal (standing water, basin)	Freshwater lakes	Moist sand beach areas of large lakes.
Beach zone	Marine intertidal	Rocks, sand, and mud flats of the intertidal zone.

After Merritt, R. W., and Cummins, K. W. (1996). "An Introduction to the Aquatic Insects of North America." Kendall/Hunt, Dubuque, IA.

TABLE III Categorization of Aquatic Insect Habits: That Is, Mode of Existence

Category	Description
Skaters	Adapted for "skating" on the surface where they feed as scavengers on organisms trapped in the surface film (e.g., Heteroptera: Gerridae, water striders).
Planktonic	Inhabiting the open-water limnetic zone of standing waters (lentic; lakes, bogs, ponds). Representatives may float and swim about in the open water but usually exhibit a diurnal vertical migration pattern (e.g., Diptera: Chaoboridae, phantom midges) or float at the surface to obtain oxygen and food, diving when alarmed (e.g., Diptera: Culicidae, mosquitoes).
Divers	Adapted for swimming by "rowing" with the hind legs in lentic habitats and lotic pools. Representatives come to the surface to obtain oxygen, dive and swim when feeding or alarmed; may cling to or crawl on submerged objects such as vascular plants (e.g., Heteroptera: Corixidae, water boatman; Coleoptera: adult Dytiscidae, predaceous diving beetles).
Swimmers	Adapted for "fishlike" swimming in lotic or lentic habitats. Individuals usually cling to submerged objects, such as rocks (lotic riffles) or vascular plants (lentic) between short bursts of swimming (e.g., Ephemeroptera: Siphlonuridae, Leptophlebiidae).
Clingers	Representatives have behavioral (e.g., fixed retreat construction) and morphological (e.g., long, curved tarsal claws, dorsoventral flattening, ventral gills arranged as a sucker) adaptations for attachment to surfaces in stream riffles and wave-swept rocky littoral zones of lakes (e.g., Ephemeroptera: Heptageniidae; Trichoptera: Hydropsychidae; Diptera: Blephariceridae).
Sprawlers	Inhabiting the surface of floating leaves of vascular hydrophytes or fine sediments, usually with modifications for staying on top of the substrate and maintaining the respiratory surfaces free of silt (e.g., Ephemeroptera: Caenidae; Odonata: Libellulidae).
Climbers	Adapted for living on vascular hydrophytes or detrital debris (e.g., overhanging branches, roots and vegetation along streams, submerged brush in lakes) with modifications for moving vertically on stem-type surfaces (e.g., Odonata: Aeshnidae).
Burrowers	Inhabiting the fine sediments of streams (pools) and lakes. Some construct discrete burrows that may have sand grain tubes extending above the surface of the substrate or the individuals may ingest their way through the sediments (e.g., Ephemeroptera: Ephemeridae, burrowing mayflies; Diptera: most Chironominae, Chironomini, bloodworm midges). Some burrow (tunnel) into plants stems, leaves, or roots (miners).

After Merritt, R. W., and Cummins, K. W. (1996). "An Introduction to the Aquatic Insects of North America." Kendall/Hunt, Dubuque, IA.

TABLE IV General Classification Systems for Aquatic Insect Trophic Relations

	Subdivision of function group			
Functional group[a]	**Dominant food**	**Feeding mechanism**	**Examples of taxa**	**General particle size range of food (μm)**
Shredders	Living vascular hydrophyte plant tissue	Herbivores—chewers and miners of live macrophytes	Trichoptera: Phryganeidae, Leptoceridae	> 10^3
	Decomposing vascular plant tissue and wood—coarse particular organic matter (CPOM)		Trichoptera: Limnephilidae	
			Plecoptera: Pteronarcyidae, Nemouridae Diptera: Tipulidae, Chironomidae	
Collectors		Detritivores—filterers or suspension feeders	Trichoptera: Hydropsychidae	
	Decomposing fine particular organic matter (FPOM)		Diptera: Simuliidae	< 10^3
		Detritivores—gatherers or deposit (sediment) feeders (includes surface film feeders)	Ephemeroptera: Ephemeridae, Baetidae, Ephemerellidae Diptera: Chironomidae	
Scrapers	Periphyton—attached algae and associated material	Herbivores—grazing scrapers or mineral and organic surfaces	Trichoptera: Glossomatidae Coleoptera: Psephenidae Ephemeroptera: Heptageniidae	< 10^3
Predators (engulfers)	Living animal tissue	Carnivores—attack prey, pierce tissues and cells, suck fluids	Heteroptera: Belostomatidae, Odonata,	> 10^3
	Living animal tissue	Carnivores—ingest whole animals (or parts)	Plecoptera: Perlidae, Perlodidae Coleoptera: Dytiscidae, Megaloptera Trichoptera: Rhyacophilidae	> 10^3

After Merritt, R. W., and Cummins, K. W. (1996). "An Introduction to the Aquatic Insects of North America." Kendall/Hunt, Dubuque, IA.

[a]General category based on feeding mechanism.

differences over short distances. In the upper reaches of a catchment or drainage basin, small streams often display a range of habitats characterized by areas that are shallow, with fast flow over pebbles, cobbles, and boulders. There are also areas with steep gradients, cascades, or waterfalls when the underlying substrate is bedrock. There also may be areas of slow velocity in pools of deeper water.

In many streams draining forested watersheds, pools are found. Pools are depositional areas during normal flow as organic and inorganic particles settle to the substrate, and a similar settling process often occurs in side channels or backwater areas of streams. Pools are also created upstream of large instream pieces of wood, which may form obstructions known as debris dams. Because pools are generally characterized by reduced water velocity, many of the small particles normally suspended in fast flows settle to the bottom. In many low-gradient streams, including large rivers, bottom substrate often consists of silt, sand, and gravel-sized particles that are frequently moved by the force of the flowing water. In such systems, large pieces of woody debris entering the river from bank erosion or from adjacent floodplain or upstream areas may represent an important habitat for invertebrate colonization.

Substratum characteristics are often perceived as a major contributor to the distribution of many invertebrates; however, many other factors, including water velocity, food, feeding habits, refuge, and respiratory requirements, can be associated with specific substrates. Substratum particle size is influenced by several items, including geology, physical characteristics of the rock, past and present geomorphic processes (flowing water, glaciation, slope, etc.), climate and precipitation, and length of time over which the processes occur. These in turn influence landform, which exerts a major influence on various hydrological characteristics of aquatic habitats. Unlike many lentic environments, in lotic systems the velocity of moving water is sufficient to pass the water around the body of an insect and turbulence provides reaeration; thus, dissolved oxygen is rarely limiting to stream inhabitants. Local transport and storage of inorganic and organic materials by the current may be either detrimental (e.g., scouring action) or beneficial (as a food source). For example, most aquatic insects in flowing waters are passive filter feeders and depend on the water current for delivery of their food. Scouring flows may therefore remove from the streambed large organic particles (e.g., leaves) as well as smaller ones, creating temporary reductions in food supplies. In contrast, moderately rapid flows may facilitate feeding of some scraper or grazer insects by preventing excessive sedimentation buildup on the surfaces on which they feed.

Some Insects and Their Adaptations to Erosional Habitats

Adaptations of aquatic insects to torrential or "rapid flow" habitats include the dorsoventral flattening of the body, which serves two purposes: it increases the organism's area of contact with the surface substratum, and it offers a mechanism by which animals can remain in the boundary layer when water velocity diminishes, thereby reducing drag under subsequent exposure to high velocities. However, this second idea may be an oversimplification. Indeed, some authors have suggested that the dorsoventrally flattened shape may actually generate lift in the insect. Examples of animals inhabiting stones in torrential habitats include a number of mayflies (Ephemeroptera) belonging to the families Heptageniidae (Fig. 1A) and Ephemerellidae; some Plecoptera, such as Perlidae (Fig. 1G); some Megaloptera (i.e., Corydalidae) (Fig. 1D); and caddisflies (Trichoptera), such as Leptoceridae (*Ceraclea).*

In addition to body shape, many mayflies and stoneflies have legs that project laterally from the body, thereby reducing drag and simultaneously increasing friction with the substrate. Most of these taxa are either scrapers or gatherers on surfaces of stones or predators on other aquatic insects. Undoubtedly, the diverse physical forces encountered in aquatic environments, especially streams, influence the array of morphologies found among aquatic insects.

In some caddisflies (e.g., Glossosomatidae), the shape of the case rather than the insect is modified. The larvae of Glossosomatidae in their tortoiselike cases are frequently seen grazing on the upper surfaces of stones in riffle areas. Another curious caddisfly grazer on stone surfaces is *Helicopsyche,* whose larvae construct coiled cases of sand grains shaped like snail shells. Both glossosomatids and helicopsychids reach their greatest abundances in sunny cobble riffles, where they feed on attached periphyton or algae. Another lotic insect that relies on a rather streamlined case is the limnephilid caddisfly *Dicosmoecus* (Fig. 1C).

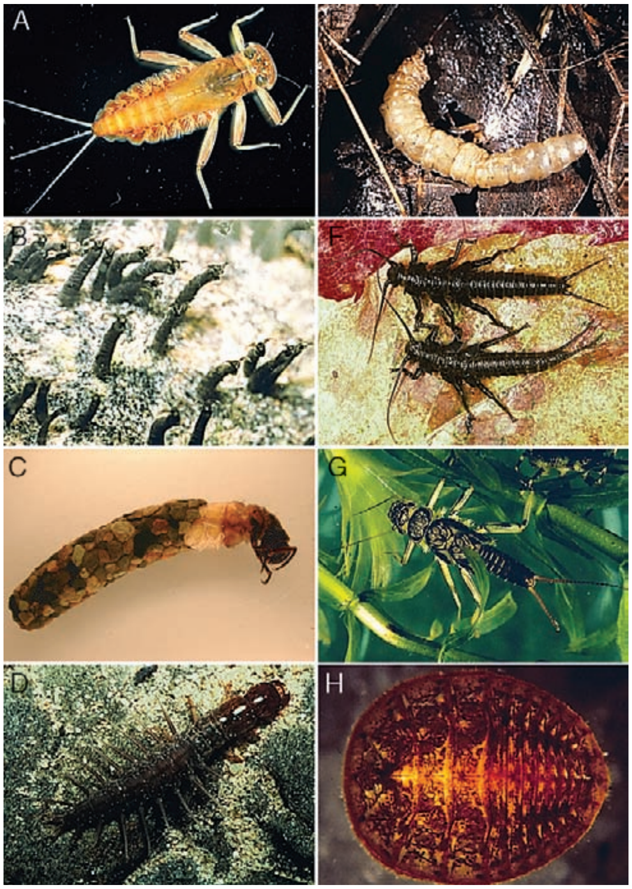

FIGURE 1 Typical insects inhabiting lotic environments. (A) Ephemeroptera: Heptageniidae (*Rhithrogena*). (Photograph by H. V. Daly.) (B) Diptera: Simuliidae (*Simulium*), (C) Trichoptera: Limnephilidae (*Dicosmoecus*), (D) Megaloptera: Corydalidae (*Corydalus*), (E) Diptera: Tipulidae (*Tipula*), (F) Plecoptera: Pteronarcyidae (*Pteronarcys*), (G) Plecoptera: Perlidae, (H) Coleoptera: Psephenidae (*Psephenus).*

Larvae of the dipteran family Blephariceridae are unusual in that they possess hydraulic suckers. A V-shaped notch at the anterior edge of each of the six ventral suckers works as a valve out of which water is forced when the sucker is pressed to the substrate. The sucker operates as a piston with the aid of specialized muscles. In addition, a series of small hooks and glands that secrete a sticky substance aid sucker attachment as the larvae move in a zigzag fashion, releasing the anterior three suckers, lifting the front portion of the body to a new position, and then reattaching the anterior suckers before releasing and moving the posterior ones to a new position. These larvae are commonly found on smooth stones in very rapid velocities and are usually absent from stones covered with moss and from roughened stones that interfere with normal sucker function. Several other aquatic insects have structures that simulate the action of suckers. The enlarged gills of some mayflies (e.g., *Epeorus* sp. and *Rhithrogena* sp.: Fig. 1A) function as a friction pad, and *Drunella doddsi* has a specialized abdominal structure for the same purpose. Some chironomids have "pushing prolegs" represented by circlet of small spines that function as a false sucker when pressed to the substrate. Mountain midge larvae (Deuterophlebiidae) possibly use a similar mechanism to attach their suckerlike prolegs. Most of these animals are primarily grazers on thin films of epilithon (algae, associated fine organic matter, and microbes) found on the surface of stones.

Flowing water usually carries many organic (and inorganic) particles and a number of insects exploit these suspended particles. Filter-feeding collectors (Table IV) exploit the current for gathering food with minimal energy expenditure. For example, certain filtering collectors exploit locations where flows converge over and around substrates, thus allowing the animals to occupy sites of greater food delivery. Examples include caddisfly larvae belonging to the families Hydropsychidae and Brachycentridae. Silk is used for attachment by a number of caddisflies (e.g., Hydropsychidae, Philopotamidae, and Psychomyiidae), which build fixed nets and retreats (Fig. 2A). Although the Philopotamidae are found in riffle habitats, their fine-meshed, tubelike nets are usually found in crevices or undersides of stones in low velocity microhabitats (Fig. 2C). The nets of the caddisfly, *Neureclipsis,* are limited to moderately slow (< 25 cm s^{-1}) velocities and the large (up to 20 cm long), trumpet-shaped nets (Fig. 2D) are used to capture small animals drifting downstream. *Neureclipsis* larvae are often very abundant in some lake outflow streams where drifting zooplankton are abundant.

Some case-making caddisflies (e.g., *Brachycentrus* sp.) also use silk for attaching their cases to the substrate in regions of moderately rapid flow. Many chironomid larvae construct fixed silken retreats for attachment or silken tubes that house the larvae, with a conical catchnet spun across the lumen of the tube. Periodically, the larva devours its catchnet with adhering debris that has been swept into the burrow by the water currents. Meanwhile, other chironomid larvae such as *Rheotanytarsus* spp. construct small silk cases that are attached to the stream substratum with extended hydralike arms. The arms project up in the current and are smeared with a silklike secretion to capture particles.

Larval blackflies (Simuliidae, Fig. 1B) use a combination of hooks and silk for attachment. The thoracic proleg resembles that of chironomids and deuterophlebiids, described earlier, and the last abdominal segment bears a circlet of hooks, which it uses to anchor itself to substrates. The larva moves forward, inchwormlike, spins silk over the substrate, and attaches the proleg and then the posterior circlet of hooks to the silken web. Most blackfly larvae possess well-developed cephalic fans, which are used to filter small particles from suspension. These attached larvae twist their bodies longitudinally from 90° to 180° with the ventral surface of the head and fans facing into the current. The fusiform body shape of blackfly larvae reduces turbulence and drag around their bodies, which are often located in regions of relatively rapid flow. Blackfly pupae are housed in silken cases that are attached to the substrate.

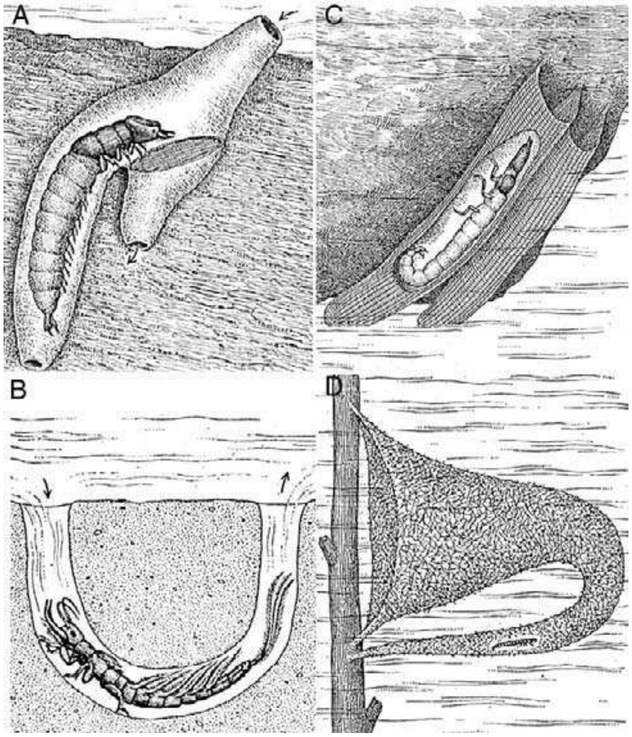

FIGURE 2 Representative lotic insects in their environment: (A) Caddisfly larva (*Macrostenum*) in its retreat grazing on materials trapped on its capture net, (B) mayfly larva of *Hexagenia* (Ephemeridae) in its U-shaped burrow, (C) tubelike nets of philopotamid caddisfly larvae (Philotamidae) on the lower surface of a stone, (D) the caddisfly larva and cornucopia-shaped net of *Neureclipsis* (Polycentropodidae). [Habitat drawings modified and taken from Wallace, J. B., and Merritt, R. W. (1980). Filter-feeding ecology of aquatic insects. *Annu. Rev. Entomol.* **25,** 103–132, (B); Merritt, R. W., and Wallace, J. B. (1981). Filter-feeding insects. *Sci. Am.* **244,** 131–144 (A, C, D).]

Although unidirectional current is the basic feature of streams, most lotic insects have not adapted to strong currents, but instead have developed behavior patterns to avoid current. Very few lotic insects are strong swimmers, probably because of the energy expenditure required to swim against a current. Downstream transport or drift requires only a movement off the substrate to enter the current. Streamlined forms, such as the mayflies *Baetis* spp., *Centroptilum, Isonychia* spp., and *Ameletus* spp., are capable of short rapid bursts of swimming, but most lotic insects move by crawling or passive displacement. One characteristic of these latter mayflies is the possession of a fusiform, or streamlined, body shape: examples include several Ephemeroptera such as *Baetis, Centroptilum,* and *Isonychia,* as well as a number of beetle (Coleoptera) larvae. A fusiform body shape reduces resistance in fluids, and within the mayflies the shape is often associated with excellent swimming abilities.

The benthic fauna in streams often can be found in cracks and crevices, between or under rocks and gravel, within the boundary layer on surfaces, or in other slack-water regions. Another method of avoiding fast currents is living in debris accumulations consisting of leaf packs and small woody debris. This debris offers both a food resource and a refuge for insects and contains a diverse array of aquatic insects including stoneflies such as Peltoperlidae and Pteronarcyidae (Fig. 1F), caddisflies such as Lepidostomatidae and some Limnephilidae, as well as dipterans such as chironomids and tipulid crane flies (Fig. 1E).

In some streams with unstable sandy or silt substrates, woody debris can represent a "hot spot" of invertebrate activity. Wood debris provides a significant portion of the stable habitat for insects in streams when the power of the flowing water is insufficient to transport the wood out of the channel. In addition to the insect component using wood primarily as a substrate, there is often a characteristic xylophilous fauna associated with particular stages of wood degradation. These include chironomid midges and scraping mayflies (*Cinygma* spp. and *Ironodes* spp.) as early colonizers, and larvae and adults of elmid beetles. In western North America, an elmid (*Lara avara*) and a caddisfly (*Heteroplectron*) are gougers of firm waterlogged wood, chironomids are tunnelers, and the tipulids, *Lipsothrix* spp., are found in wood in the latest stages of decomposition. Woody debris is most abundant in small, forested watersheds, but it is also an important habitat in larger streams with unstable beds. In the southeastern coastal plain of the United States and in low gradient mid- and southwestern streams and rivers with unstable bottom substrate, woody debris or "snags" often represent the major habitat for aquatic insect abundance and biomass. High populations and biomass of filter-feeding animals such as net-spinning caddisflies (*Hydropsyche* spp., *Cheumatopsyche* spp., and *Macrostenum*) (Fig. 2A), and blackflies occur in these streams and rivers. In addition to filter feeders, other groups such as odonates, mayflies, stoneflies, elmid beetles, nonfiltering caddisflies, and dipteran larvae can be locally abundant on large pieces of woody debris. Invertebrate shredders and scrapers promote decomposition of outer wood surfaces by scraping, gouging, and tunneling through wood. In fact, wood gouging habits of some net-spinning caddisflies have been blamed for the failure of submerged timber pilings that had been supporting a bridge!

Sand and silt substrates of rivers and streams are generally considered to be poor habitats because the shifting streambed affords unsuitable attachment sites and poor food conditions. An extreme example of this instability is the Amazon River, where strong currents move the bedload downstream, resulting in dunes of coarse sand up to 8 m in height and 180 m in length, thus largely preventing the establishment of a riverbed fauna. However, sandy substrates do not always result in poor habitat for all aquatic insects: some sandy streams are quite productive. Blackwater (i.e., high tannic acid concentrations from leaf decomposition) streams of the southeastern United States have extensive areas of sand, with some of insects, such as small Chironomidae (< 3 mm in length), exceeding 18,000/m^{-2} in abundance. Their food is derived from fine organic matter, microbes, and algae trapped in the sandy substrate. Numerically, the inhabitants of sandy or silty areas are mostly sprawlers or burrowers, with morphological adaptations to maintain position and to keep respiratory surfaces in contact with oxygenated water. At least one insect, the mayfly *Ametropus,* is adapted for filter feeding in sand and silt substrates of large rivers. *Ametropus* uses the head, mouthparts, and forelegs to create a shallow pit in the substrate, which initiates a unique vortex (flow field in which fluid particles move in concentric paths) in front of the head and results in resuspension of fine organic matter as well as occasional sand grains. Some of these resuspended fine particles are then trapped by fine setae on the mouthparts and forelegs. Many predaceous gomphid (Odonata) larvae actually burrow into the sediments by using the flattened, wedge-shaped head and fossorial (adapted for digging) legs. The predaceous mayflies *Pseudiron* spp. and *Analetris* spp. have long, posterior-projecting legs and claws that aid in anchoring the larvae as they face upstream. Some mayflies (e.g., Caenidae and Baetiscidae) have various structures for covering and protecting gills, and others (e.g., Ephemeridae, Behningiidae) have legs and mouthparts adapted for digging. The predaceous mayfly *Dolania* spp. burrow rapidly in sandy substrates and have dense setae located at the anterior–lateral corners of the body as well as several other locations. The larva uses its hairy body and legs to form a cavity underneath the body where the ventral abdominal gills are in contact with oxygenated water.

Dense setae also are found in burrowing mayflies belonging to the family Ephemeridae that are common inhabitants of sand and silt substrates. They construct shallow U-shaped burrows and use their dorsal gills to generate water currents through the burrow (Fig. 2B), while using their hairy mouthparts and legs to filter particles from the moving water. Hairy bodies seem to be a characteristic of many animals dwelling on silt substrates, which include other collector mayflies such as *Caenis, Anepeorus,* and some Ephemerellidae.

Many dragonflies (e.g., *Cordulegaster* spp., *Hagenius* spp., and Macromiidae) have flattened bodies and long legs for sprawling on sandy and silty substrates. Some caddisflies, such as *Molanna,* have elongate slender bodies but have adapted to sand and silt substrates by constructing a flanged, flat case. They are camouflaged by dull color patterns and hairy integuments that accumulate a coating of silt. The eyes, which cap the anterolateral corners of the head, are elevated over the surrounding debris. The genus *Aphylla* (Gomphidae) is somewhat unusual in that the last abdominal segment is upturned and elongate, allowing the larvae to respire through rectal gills while buried fairly deep in mucky substrate. Some insects burrow within the upper few centimeters of the substratum in depositional areas of streams. This practice is found among some dragonflies and a number of caddisflies, including *Molanna,* and various genera of the families Sericostomatidae and Odontoceridae.

Specialized Flowing Water Habitats

The hyporheic region is the area below the bed of a stream where interstitial water moves by percolation. In gravelly substrates or glacial outwash areas, it may also extend laterally from the banks. In some situations an extensive fauna occurs down to one meter in such substrates. Most orders are represented, especially taxa with slender flexible bodies or small organisms with hard protective exoskeletons. Some stoneflies in the Flathead River of Montana spend most of their larval period in this extensive subterranean region of flow adjacent to the river. Stonefly larvae have been collected in wells over 4 m deep, located many meters from the river. Rivers draining glaciated regions where there are large boulders and cobble appear to have an exceptionally well-developed hyphoreic fauna.

Other specialized flowing water habitats include the madicolous (or hygropetric) habitats, which are areas in which thin sheets of water flow over rock. These often approach vertical conditions (e.g., in waterfalls) and have a characteristic fauna. Among common animals in these habitats are caddisflies, including several microcaddisflies (Hydroptilidae), Lepidostomatidae, beetles such as Psephenidae, and a number of Diptera larvae belonging to the Chironomidae, Ceratopogoniidae, Thaumaleidae, Tipulidae, Psychodidae, and some Stratiomyiidae.

Thermal (hot) springs often have a characteristic fauna, which is fueled by algae and bacteria adapted to high temperatures. The common inhabitants include a number of dipteran families such as Chironomidae, Stratiomyiidae, Dolochopodidae, and Ephydridae, as well as some coleopterans. A number of these survive within rather narrow zones between the thermal spring and cooler downstream areas.

LENTIC HABITATS

Lentic or standing-water habitats range from temporary pools to large deep lakes and include marshes and swamps, as well as natural (i.e., tree holes, pitcher plants) and artificial (i.e., old tires, rain barrels) containers. The available habitats and communities for insects in a pond or lake were defined in Table II. These habitats include the littoral zone, which comprises the shallow areas along the shore with light penetration to the bottom and normally contains macrophytes (rooted vascular plants). The limnetic zone is the open-water area devoid of rooted plants, whereas the deeper profundal zone is the area below which light penetration is inadequate for plant growth, water movement is minimal, and temperature may vary only slightly between summer and winter. The aquatic and semiaquatic insect communities inhabiting these zones are known as the pleuston (organisms associated with the surface film), plankton and nekton (organisms that reside in the open water), and benthos (organisms associated with the bottom, or solid–water interface). Nektonic forms are distinguished from plankton by their directional mobility, and the latter are poorly represented in lentic waters by insects; the majority of insects found in standing-water habitats belong to the benthos. Their composition and relative abundance is dependent on a variety of factors, some of which are integrated along depth profiles. The overall taxonomic richness of benthic insect communities generally declines with increasing depth.

Among the aquatic communities of lentic habitats, the following orders of aquatic and semiaquatic insects are commonly found within the littoral, limnetic, and profundal zones: the springtails (Collembola), mayflies (Ephemeroptera), true bugs (Heteroptera), caddisflies (Trichoptera), dragonflies (Anisoptera) and damselflies (Zygoptera), true flies (flies, gnats, mosquitoes, and midges) (Diptera), moths (Lepidoptera), alderflies (Megaloptera), and beetles (Coleoptera). Not all these groups occur in lakes, and many are associated with ponds or marshes; examples of typical lentic insects are shown in Figs. 3 and 4.

The Pleuston Community

The unique properties of the water surface or air–water interface constitute the environment of the pleuston community. The Collembola, or springtails, are small in size, have a springing organ (furcula), and a water-repelling cuticle that enables them to be supported by and move across water surfaces. Among the true bugs, the Gerridae (water striders) and related families, the Veliidae (broad-shouldered water striders) and Hydrometridae (water measurers), are able to skate across the water. Adaptations for this habit include retractable preapical claws to assist in swimming, elongate legs

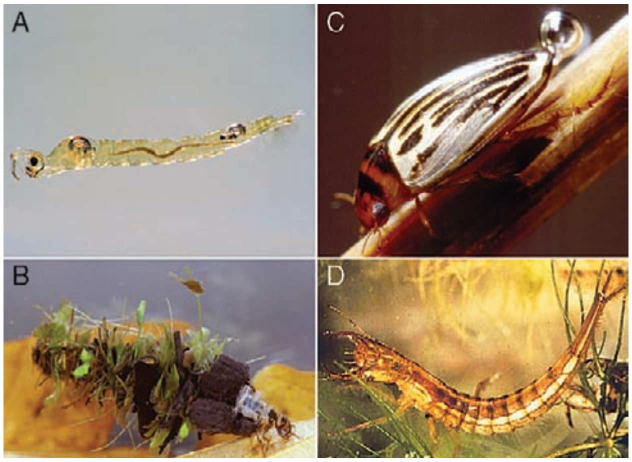

FIGURE 3 Typical insects inhabiting lentic environments. (A) Diptera: Chaoboridae (*Chaoborus*), (B) Trichoptera: Limnephilidae (*Limnephilus*), (C) Coleoptera: Dytiscidae (*Agabus*), (D) Coleoptera: Dytiscidae. (Photographs in A, B, and C by M. Higgins.)

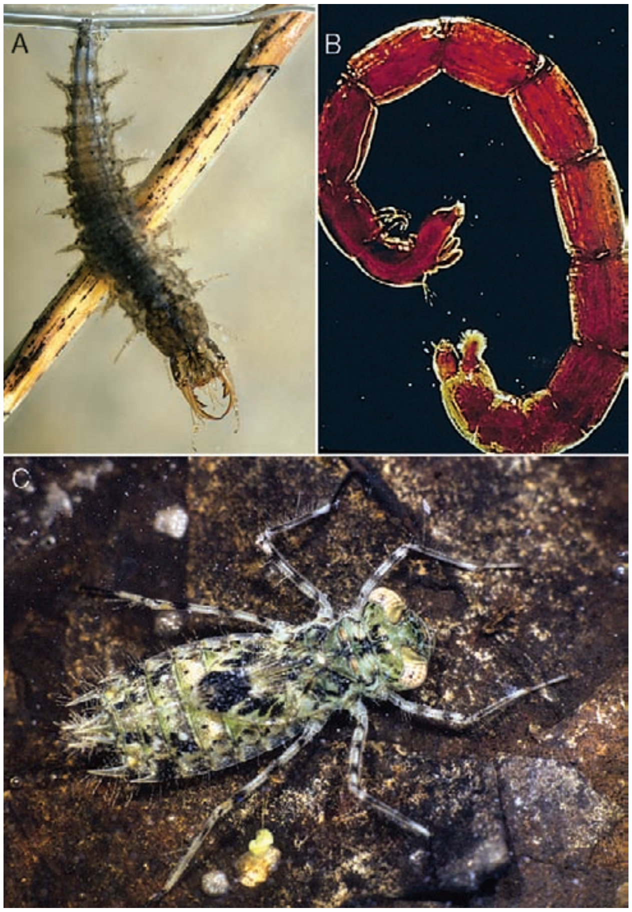

FIGURE 4 Typical insects inhabiting lentic environments (A) Coleoptera: Hydrophilidae (*Hydrochara).* (Photograph by M. Higgins.) (B) Diptera: Chironomidae (*Chironomus*), (C) Odonata: Libellulidae (*Pantala).* (Photograph by M. Higgins.)

and body to distribute the insect's weight over a large area of the surface film, and hydrofuge (nonwettable) hairpiles for support on the surface. Some gerrids also are capable of detecting surface vibrations caused by potential prey. Adult whirligig beetles (Gyrinidae) live half in and half out of water with each eye divided into upper and lower halves, permitting vision simultaneously in both the air and the water; glands keep the upper portion of the body greased to repel water. The middle and hind legs of adult gyrinids are paddle shaped, enabling them to be one of the most effective swimming invertebrates. Among the Diptera, only the mosquitoes (Culicidae) may be considered to be permanent members of the pleuston of lentic waters. The larvae and pupae of most species use the underside of the surface film for support. Larval *Anopheles* lie horizontally immediately beneath the air–water interface, supported by tufts of float hairs on each. Larvae of most other genera (*Aedes, Culex, Culiseta*) hang upside down, with an elongated terminal respiratory siphon penetrating the surface film. Feeding adaptations associated with pleuston specialization include predation by the Hemiptera and Coleoptera, to practically all functional feeding modes by different mosquito larvae, including collecting-filtering and gathering, scraping, and shredding (Table IV).

The Nekton and Plankton Communities

The nekton are swimmers able to navigate at will (e.g., Coleopera, Hemiptera, some Ephemeroptera), whereas plankton are floating organisms whose horizontal movements are largely dependent on water currents. The phantom midge *Chaoborus* sp. (Chaoboridae) (Fig. 3A) is normally regarded as the only planktonic insect and is abundant in many eutrophic (nutrient-rich) ponds and lakes. The tracheal system in these larvae is reduced to kidney-shaped air sacs that function solely as hydrostatic organs, and the larvae slowly descend or rise by adjusting the volume of the air sacs. *Chaoborus* remains in benthic regions during the day but moves vertically into the water column at night. These journeys are dependent on light and oxygen concentrations of the water. The larvae avoid predation by being almost transparent, and they have prehensile antennae that are used as accessory mouthparts to impale zooplankton (Fig. 3A). The only other group of insects that may be considered to be planktonic are the early chironmid instars, which have been reported in the open water column.

Among the Heteroptera, nektonic species are in the Notonectidae (back swimmers), Corixidae (water boatman), and Belostomatidae (giant water bugs), all of which are strong swimmers. Many of these rise to the water surface unless continously swimming or clinging to underwater plants. Notonectids have backs formed like the bottom of a boat and navigate upside down. They hang head downward from the surface or dive swiftly, using their long hind legs as oars. On the underside of the body, they carry a silvery film of air, which can be renewed at regular intervals, for breathing while submerged. Two genera of backswimmers (*Anisops* and *Buenoa*) use hemoglobin for buoyancy control, and this adaptation has enabled these insects to exploit the limnetic zone of fishless lentic waters, where they prey on small arthropods. They have been considered for use as biological control agents for mosquito larvae in some areas of North America. In contrast to notonectids, corixids always swim with the back up, using their elongate, flattened oarlike legs. Although some water boatmen are predators, they are the only group of semiaquatic Heteroptera that have members that are collectors, feeding on detritus and associated small plant material. The Belostomatidae are strong swimmers, but probably spend most of their time clinging to vegetation while awaiting prey, rather than actively pursuing their food in the open water. They are masters of their environment and capture and feed on a variety of insects, tadpoles, fish, and even small birds. The eggs of many belostomatids are glued to the backs of the males by the females and carried in this position until nymphs emerge, a remarkable adaptation for protection of the eggs.

Although most aquatic beetles (Coleoptera) are associated with the substrate, members of the Dytiscidae (predaceous diving beetles) and the Hydrophilidae (water scavenger beetles) are often found swimming in the water column and together constitute the majority of all species of water beetles. The dytiscids are mainly predators in both the adult and larval stage (Fig. 3C, D), while adult hydrophilids are omnivorous, consuming both living and dead materials. The larvae of hydrophilids are predaceous. (Fig. 4A). To respire, hydrophilid adults, having their largest spiracles on the thorax, break the surface film with their antennae; dytiscids, having their largest spiracles on the abdomen, come up tail-end first, as do the larvae of both families. Overall, there are actually few truly nektonic insects, and most of them pass through the limnetic zone when surfacing for emergence. This may be, partly, because with no resting supports in the limnetic zone, maintaining position requires continuous swimming or neutral buoyancy. The vast majority of lentic insects occur in shallow water with emergent plants and are considered to be part of the benthos.

The Benthos Community

Benthos, derived from the Greek word for bottom, refers to the fauna associated with the solid–water interface and includes insects residing on the bottom or associated with plant surfaces, logs, rocks, and other solid substrates. In lentic habitats, many insects fall into this category as mentioned earlier, particularly the Chironomidae, which often represent over 90% of the fauna in the profundal (deep-water) zone of lakes and ponds. These inhabitants are mostly burrowers that feed on suspended or sedimented organic materials and are capable of tolerating low dissolved oxygen or even anaerobic conditions. Chironomid larvae build U- or J-shaped tubes with both openings at the mud–water interface. Body undulations cause a current of water, providing conditions under which oxygen and particulate food can be drawn through the tube. Some midge larvae found in sediments (mainly *Chironomus* sp.) are bright red and are known as bloodworms (Fig. 4B). The red color is caused by the respiratory pigment hemoglobin, which enables a larva to recover rapidly from anaerobic periods because the pigment takes up oxygen and passes it to the tissues more quickly than is possible by diffusion alone.

Other members of the benthos of deeper waters include the mayfly, *Hexagenia* (Ephemeridae), which inhabits the silt and mud of nearshore lake bottoms and has legs modified for digging to construct U-shaped burrows (Fig. 2B). Mayfly numbers have been increasing because of improved water quality standards for lakes and streams. Exceptions to the main constituents of the profundal zone are some immature mayflies, stoneflies, and caddisflies that have been collected at depths from 30 to 100 m in Lake Superior, Michigan. Also, a stonefly, *Utacapnia lacustra* (Capniidae), occurs at depths of 80 m in Lake Tahoe, Calfornia–Nevada, and completes its entire life cycle at this depth, never needing to surface.

Several orders of aquatic insects reach their greatest abundance and diversity in the shallow littoral zone of ponds and lakes as benthos typically associated with macrophytes (macroalgae and rooted vascular plants). The occupants are burrowers, climbers, sprawlers, clingers, swimmers, and divers (Table III) and include the Ephemeroptera, Heteroptera, Odonata, Trichoptera, Megaloptera, Lepidoptera, Coleoptera, and Diptera. The same groups occupy marshes and some swamps, which generally tend to be shallow, with an associated plant zone across the entire surface. Mayflies belonging to the families Baetidae and Siphlonuridae are generally swimmers, clingers, and climbers in vegetated ponds and marshes and mainly feed by means of collecting-filterering or -gatherering (Table IV). Heteroptera include the water scorpions (Nepidae), which have long slender respiratory filaments and are well concealed by detritus and tangled plant growth because of their sticklike appearance. These sit-and-wait predators capture organisms that frequent their place of concealment. Other families of Heteroptera adapted for moving through vegetation in ponds are the Pleidae or pygmy back-swimmers and creeping water bugs, the Naucoridae.

The Odonata, particularly the Gomphidae, are all predators and usually conceal themselves by either burrowing in substrate, sprawling among fine sediment and detritus, or climbing on vascular plants. Sprawlers are more active hunters and include the Libellulidae (Fig. 4C) and Corduliidae. Numerous setae give them a hairy appearance to help camouflage the larvae, and color is protective in patterns of mottled greens and browns. Most Zygoptera (damselflies) and the dragonfly (Anisoptera) family Aeshnidae are mainly climbers or clingers, lurking in vegetation or resting on stems of aquatic plants. The larvae stalk their prey, and both dragonfly and damselfly larvae have a unique lower lip (the labium) armed with hooks, spines, teeth, and raptorial setae that can extend to seize prey and then bring it back into the mouth, holding the food while it is being eaten. The food of larval odonates consists of other aquatic insects such as midges, semiaquatic bugs, and beetles, as well as small fish. Predators of larval odonates include aquatic birds, fish, and large predaceous insects.

In the order Megaloptera, which includes the hellgrammites or dobsonfly larvae of streams, only the predaceous larvae of the alderfly (*Sialis*) is common in ponds and lakes. They are generally found in sand or mud along the margins, but occasionally in deeper water, and they prey on insect larvae and other small animals. The only aquatic family in the related order Neuroptera is the Sisyridae (the spongilla flies), and these are found feeding on freshwater sponges that occur in some streams and the littoral zones of lakes and ponds. The larvae, which occur on the surface or in the cavities of the host, pierce the sponge cells and suck the fluids with their elongated mouthparts.

Although most caddisflies are observed living in lotic waters, several families of caddisflies are either associated with temporary ponds in the spring, aquatic vegetation in permanent ponds, lakes and marshes, or wave-swept shore lines of lakes. The Hydropsychidae (net spinners), Helicopsychidae (snail case makers), Molannidae, and Leptoceridae are often found along wave-swept shorelines of lakes, and their feeding habits range from those of scrapers and collector-filterers to predators. The Phryganeidae and several genera within the Limnephilidae are climbers, clingers, and/or sprawlers among vegetation in temporary and permanent ponds and marshes; generally, they are shredders of vascular hydrophytes and other decaying plants. The cases of lentic caddisfly families vary with the environment they are found in. Some cases consist of narrow strips of leaves put together in spiral form around a cylinder (Phryganaeidae: *Phryganea* sp.), others consist of plant materials such as leaves and bark arranged transversely to produce a bulky cylindrical case (Limnephilidae: *Limnephilus*) (Fig. 3B).

Both aquatic and semiaquatic moths (Lepidoptera) occur in lentic habitats, and several genera form close associations with vascular hydrophytes. Larvae of the family Pyralidae (*Parapoynx* sp.) spend the first two instars on the bottom and feed on submerged leaves of water lilies, whereas older larvae generally become surface feeders. Silk spun by the caterpillars is often used to build protective retreats, and pupation usually takes place in silken cocoons or silk-lined retreats. Larval habits of aquatic and semiaquatic moths include leaf mining, stem or root boring, foliage feeding, and feeding on flower or seed structures. One semiaquatic lepidopteran called the yellow water lily borer (the noctuid *Bellura gortynoides*), mines the leaves as a young caterpillar and then bores into the petioles of lilies as an older caterpillar. Within the petiole, larvae are submerged in water and must periodically back out to expose the posterior spiracles to the air before submerging again. The larvae swim to shore by undulating their bodies and overwinter under leaf litter in protected areas.

In addition to the water scavenger and predaceous diving beetles that may occur as nekton swimming through the water column, larvae and adults of other beetles are considered to be part of the benthos of ponds and marshes. These include the Haliplidae (crawling water beetles), which are clingers and climbers in vegetation, and the Staphylinidae (rove beetles), which are generally found along shorelines and beaches, as well as in the marine intertidal zone. The Scirtidae (marsh beetles) are generally found associated with vascular hydrophytes but also are a prominent inhabitant of tree holes. The aquatic Chrysomelidae (leaf beetles) occur commonly on emergent vegetation in ponds, especially floating water lily leaves. The larvae of one genus, *Donacia,* obtain air from their host plant by inserting the sharp terminal modified spiracles into the plant tissue at the base of the plant. Water lilies can be heavily consumed by larvae and adults of the chrysomelid beetle, *Galerucella* sp., and some of the aquatic herbivorous beetles belonging to the family Curculionidae (weevils) include pests of economic importance such as the rice water weevil (the curculionid *Lissorhoptrus*).

The Diptera is clearly one of the most diverse aquatic insect orders, inhabiting nearly all lentic habitats and representing all functional feeding groups and modes of existence. Although the benthic Chironomidae may reach their highest densities in the profundal zone of eutrophic lakes and ponds, they also are largely represented in the littoral zone associated with submergent and emergent plants, where they often graze on the algae attached to leaf surfaces or are vascular plant miners. Other dipteran families that occur in the littoral or limnetic zone, along with their specific habitat, habit (mode of locomotion, attachment, or concealment), and functional feeding mode are summarized in Table V. Among these, a few are of particular interest because of their high diversity and/or abundance in these habitats, namely the crane flies (Tipulidae), the shore and brine flies (Ephydridae), and the marsh flies (Sciomyzidae). The Tipulidae, the largest family of Diptera, are found along the margins of ponds and lakes, freshwater and brackish marshes, and standing waters in tree holes. A few littoral species inhabit the marine intertidal zone. To these are added the large numbers of species that are semiaquatic, spending their larval life in saturated plant debris, mud, or sand near the water's edge or in wet to saturated mosses and submerged, decayed wood. Ephydridae larvae have aquatic and semiaquatic members and occupy several different lentic habitats ranging from salt water or alkaline pools, springs, and lakes to burrowers and miners of a variety of aquatic plants in the littoral margins of these freshwater lentic habitats. All larvae utilize a variety of food, but algae and diatoms are of particular importance in their diet. The Sciomyzidae share some of the same habitat with the shore and brine flies, particularly fresh- and saltwater marshes, and along margins of ponds and lakes among vegetation and debris. The unique aspect of their larval life is that they are predators on snails, snail eggs, slugs, and fingernail clams. The aquatic predators float below the surface film and maintain buoyancy by frequently surfacing and swallowing an air bubble. Prey may be killed immediately or over a few days.

MARINE HABITATS

As noted earlier, insects have been largely unsuccessful in colonizing the open ocean, except for some members of the heteropteran family Gerridae. Most marine insects live in the intertidal zone (i.e., between high and low tide marks), especially on rocky shores or associated with decaying seaweed on sandy beaches (Table I). Although several orders have representatives in the intertidal zone, only a few orders, notably the Diptera, Coleoptera, and Collembola, have colonized these habitats in any numbers. The harsh physical environment of this area has forced these groups to occur buried in sand or mud and to hide in rock crevices or under seaweed.

UNUSUAL HABITATS

Because of adaptive radiation over evolutionary time, insects have colonized virtually every aquatic habitat on earth. Therefore, it is not surprising that these organisms are found

TABLE V Summary of Ecological Data for Benthic Aquatic and Semiaquatic Diptera Larvae Inhabiting Lentic Habitats

Family	Habitat	Habitat	Functional feeding mode
Ceratopogonidae (biting midges, "no-see-ums")	Littoral zone (including tree holes and small temporary ponds and pools)	Generally sprawlers, burrowers or planktonic (swimmers)	Generally predators some collector-gatherers
Chironomidae (nonbiting midges)	All lentic habitats including marine, springs, tree holes	Generally burrowers, sprawlers (most are tube builders); some climber-clingers	Generally collector-gatherers, collector-filterers; some shredders and scrapers
Corethrellidae	Limnetic and littoral margins	Sprawlers	Predators
Psychodidae (moth flies)	Littoral detritus (including tree holes)	Burrowers	Collector-gatherers
Ptycopteridae (phantom crane flies)	Vascular hydrophytes (emergent zone), bogs	Burrowers	Collector-gatherers
Tipulidae (crane flies)	Littoral margins, floodplains (organic sediment)	Burrowers and sprawlers	Generally shredders, collector-gatherers
Dolichopodidae	Littoral margins, estuaries, beach zones	Sprawlers, burrowers	Predators
Stratiomyidae (soldier flies)	Littoral vascular hydrophytes; beaches (saline pools, margins)	Sprawlers	Collector-gatherers
Tabanidae (horseflies, deerflies)	Littoral (margins, sediments and detritus); beaches, marine and estuary	Sprawlers, burrowers	Predators
Canacidae (beach flies)	Beaches—marine intertidal	Burrowers	Scrapers
Ephydridae (shore and brine flies)	Littoral (margins and vascular hydrophytes)	Burrowers, sprawlers	Collector-gatherers, shredders, herbivores (miners), scrapers, predators
Muscidae	Littoral	Sprawlers	Predators
Scathophagidae (dung flies)	Vascular hydrophytes (emergent zone)	Burrower-miners (in plant stems), sprawlers	Shredders
Sciomyzidae (marsh flies)	Littoral—vascular hydrophytes (emergent zone)	Burrowers, inside snails	Predators or parasites
Syrphidae (flower flies)	Littoral (sediments and detritus), tree holes	Burrowers	Collector-gatherers

in the most unusual of aquatic habitats. The title of most versatile aquatic insect must be shared among members of the dipteran family Ephydridae, or shore flies. Shore flies can breed in pools of crude petroleum and waste oil, where the larva feed on insects that become trapped on the surface film. Other species of this family (*Ephydra cinera*), known as brine flies, occur in the Great Salt Lake, Utah, which has a salinity six times greater than that of seawater. Larva maintain water and salt balance by drinking the saline medium and excreting rectal fluid that is more than 20% salt. Another related family of flies, the Syrphidae, or "rat-tailed maggots," occur in sewage treatment lagoons and on moist substrates of trickling filter treatment facilities. Both families have larvae with breathing tubes on the terminal end, which permits the larvae to maintain contact with the air while in their environment. Some Stratiomyiidae, or soldier flies, live in the thermal hot springs of Yellowstone National Park with temperatures as high as 47°C! Other members of this family inhabit the semiaquatic medium of cow dung and dead corpses. A few species of insects have invaded caves and associated subterranean habitats, as mentioned earlier (see Lotic Habitats).

Another unsual aquatic habitat that several insect orders occupy is referred to phytotelmata or natural container habitats and include tree holes, pitcher plants, bromeliads, inflorescences, and bamboo stems. Synthetic container habitats, such as old tires, cemetery urns, rain gutters, and similar natural habitats such as hoofprints also harbor similar insects. Some of these habitats are extremely small and hold water only temporarily, but nevertheless can be quite diverse. The most common order found in these habitats is the Diptera with more than 20 families reported. Over 400 species of mosquitoes in 15 genera alone inhabit these bodies of water and some of these species are important vectors of disease agents.

Insect communities inhabiting pitcher plants (*Sarracenia purpurea*) in North America are exemplified by a sarcophagid or flesh fly (*Blaesoxipha fletcheri*), a mosquito (*Wyeomyia smithii*), and a midge (*Metriocnemus knabi*). The relative abundance of these pitcher plant inhabitants is related to the age, inasmuch as each of the three species consumes insect remains that are in different stages of decomposition. Specifically, the larvae of the flesh fly feed on freshly caught prey floating on the pitcher fluid surface. The mosquito larvae filter feed on the decomposed material in the water column, and the midge larvae feed on the remains that collect on the bottom of the pitcher chamber. Temporary habitats are important because they are populated by a variety of species, often with unique morphological, behavioral, and physiological properties.

See Also the Following Articles

Cave Insects • Marine Insects • Mosquitoes • Soil Habitats • Swimming

Further Reading

Allan, J. D. (1995). "Stream Ecology." Kluwer, Dordrecht, the Netherlands.

Brönmark, C., and Hansson, L.-A. (1998). "The Biology of Lakes and Ponds." Oxford University Press, Oxford, U.K.

Cummins, K. W. (1973). Trophic relations of aquatic insects. *Annu. Rev. Entomol.* **18,** 183–206.

Cushing, C. E., and Allan, J. D. (2001). "Streams: Their Ecology and Life." Academic Press, San Diego, CA.

Hynes, H. B. N. (1970). "The Ecology of Running Waters." University of Toronto Press, Toronto, Ont., Canada.

McCafferty, W. P. (1981). "Aquatic Entomology." Jones & Bartlett, Boston.

Merritt, R. W., and Cummins, K. W. (eds.). (1996). "An Introduction to the Aquatic Insects of North America." Kendall/Hunt, Dubuque, IA.

Merritt, R. W., and Wallace, J. B. (1981). Filter-feeding in aquatic insects. *Sci. Am.* **244,** 131–144.

Resh, V. H., and Rosenberg, D. M. (1984). "The Ecology of Aquatic Insects." Praeger Scientific, New York.

Ward, J. V. (1996). "Aquatic Insect Ecology," Vol. 1, "Biology and Habitat." Wiley, New York.

Williams, D. D. (2002). "The Ecology of Temporary Waters." Blackburn Press, Caldwell, NJ.

Williams, D. D., and Feltmate, B. W. (1992). "Aquatic Insects." CAB International, Wallingford, Oxon, U.K.

Arachnida

see *Scorpions; Spiders*

Archaeognatha

(Bristletails)

Helmut Sturm
University Hildesheim, Germany

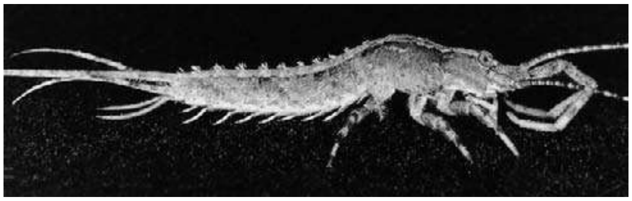

FIGURE 1 A male archaeognathan (*Machilis germanica*), body length ca. 12 mm, lateral view; for details see Fig. 2.

The Archaeognatha (Microcoryphia; part of the subdivided order Thysanura) are apterygote insects with a body size between 6 and 25 mm and a cylindrical shape (Fig. 1). The eyes are large and contiguous, and there are two lateral and one median ocelli (small single eyes with a single beadlike lens). The flagellate (whiplike) antennae extend one-half to three times the length of the body. The mouthparts are ectognathous (freely visible) and the mandibles are linked with the head by a monocondylic joint (i.e., one point of attachment). Some authors believe that this feature distinguishes the Archaeognatha from all other ectognathous Insecta. The seven-segmented maxillary palps are longer than the legs. The thoracic tergites II + III are in lateral view strongly arched, and the two or three tarsal segments of the legs are rigidly united. Some taxa have additional scopulae (dense brushes of specialized hairs) on the distal end of the third tarsal segment. There are mostly pairs or double pairs of eversible vesicles on the coxites of the abdomen (Fig. 3). On each of the abdominal coxites II to IX, styli (pointed, nonarticulated processes) are present (Fig. 3).

Females have two long gonapophyses on each of the abdominal segments VIII + IX, forming the ovipositor. The penis of the males on abdominal segment IX varies in length, and in Machilidae it is fitted with paired parameres on abdominal segments IX or VIII + IX. The three filiform and scaled caudal appendages (one long filum terminale and two laterally inserted cerci) are directed backward. Tergites, cerci, and coxites are always scaled. The molts continue in adult stages. Many species are petrophilous (living on and under stones). The order comprises about 500 species in two families (Machilidae and Meinertellidae).

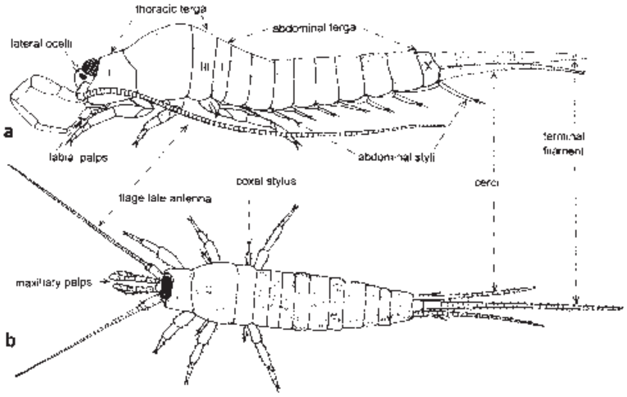

FIGURE 2 General structure of Archaeognatha, semidiagrammatic. (a) Lateral view. (b) Dorsal view, color pattern of dorsal scales intimated. (Reprinted from *Deutsche Entomologische Zeitschrift* **48,** p. 4, © 2001 by Wiley-VCH, with permission.)

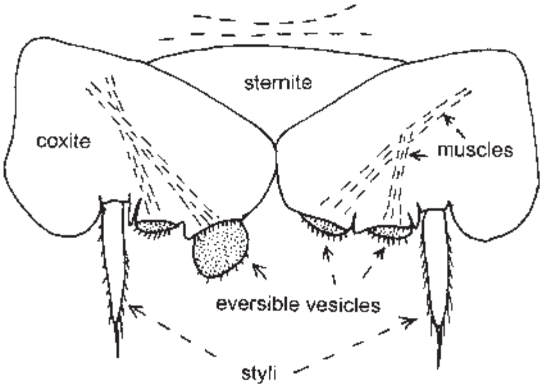

FIGURE 3 Abdominal coxites III of *M. germanica,* ventral view. The eversible vesicles can be exserted by increasing the inner pressure and retracted by muscles.

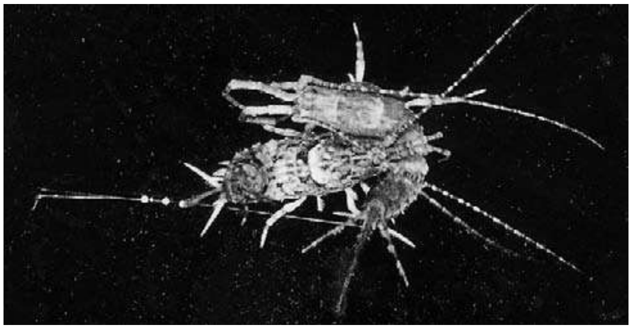

FIGURE 4 Mating position of *M. germanica,* dorsal view. The male has drawn out a secreted thread, deposited three sperm droplets on the thread, and taken up a U form. The ovipositor of the female is touching one of the sperm droplets.

FOSSIL RECORD, SYSTEMATICS, AND BIOGEOGRAPHY

The fossil record of Paleozoic and Mesozoic apterygotes is poor, and many of the fossils of ectognathous representatives cannot be clearly assigned to extant orders. From the Mesozoic, the only archaeognathan fossil is *Cretaceomachilis libanensis* from the lower Cretaceous of Lebanon. For the Cenozoic period, there are many fossils of Archaeognatha, most being amber inclusions. For example, from Baltic amber (ca. 35 mya) seven species of Machilidae are known. All extant forms and the fossils from the Cretaceous and the Tertiary can be included in the superfamily Machiloidea. This group includes two families, the more primitive Machilidae (46 genera and some 325 species), with three subfamilies (Machilinae, Petrobiinae, Petrobiellinae), and the more derived Meinertellidae (19 genera and some 170 species).

The Machiloidea are distributed worldwide. Only the Meinertellidae occur in South America, the Caribbean, South Africa, Australia, and Melanesia. Both Machilidae and Meinertellidae occur in the United States.

BEHAVIOR AND ECOLOGY

The mating behavior of archaeognathans is unique. There are three different modes of sperm transfer. In the most widespread and unique mating behavior, a carrier thread is used. In *Machilis germanica,* for example, the male approaches the female and drums on her with his long maxillary palps. The female shows “willingness” to mate by moving toward the male and bending up the tip of her abdomen. The male then attaches a secreted thread to the ground with his parameres. While the thread is being drawn out, the male secretes three to five sperm droplets onto the thread (Fig. 4). The male curves simultaneously around the female, preventing her from moving forward. Finally, the male places the sperm droplets onto the ovipositor of the female. The indirect transmission of sperm droplets, which are deposited on a thread stretched between parameres and the ground, is unique within the animal kingdom. At least two other possibilities of indirect sperm transfer are known. In *Petrobius* (Machilidae), the sperm are moved directly from the penis onto the ovipositor of the female and in all Meinertellidae sperm are transferred by stalked spermatophores deposited on the ground.

Archaeognatha are found in habitats with very different climates. Representatives of the genus *Allopsontus* (Machilidae) are found up to 5000 m in the Himalayan region. In contrast, two meinertellid species live in the Amazonian forest. Some genera (e.g., *Petrobius*) are found near the seacoast. In tropical forests, meinertellid genera are found on the leaves of bushes and trees. Most Archaeognatha feed on green algae, lichens, and dead leaves. Several species of spiders probably are their principal predators.

Their protection against enemies is probably provided by three main factors: (1) the presence of long appendages with sensilla (filum terminale, cerci, antennae); (2) a dense scale cover on the relatively thin and flexible tergites and coxites; and (3) their ability to jump, which is fully developed in all free-living stages and in all recent representatives, and probably is their most effective defense.

PHYSIOLOGY

The cuticle of the Archaeognatha is generally thin and flexible and bears a multitude of sensory setae and other sensory organs. The hypopharynx (a mouthpart between maxillae and labium) is well developed. The coxal vesicles absorb fluids; their number varies from none to two within a coxite. In all machilid males, coxite IX bears the penis (length from $\frac{3}{4}$ of the coxite to $1\frac{1}{2}$ in Machilidae, in Meinertellidae ca. $\frac{1}{2}$ of the coxite length). In females, the paired gonapophyses on coxites VIII and IX form the ovipositor. It is of different length and has generally a specific chaetotaxy in each species.

DEVELOPMENT

The eggs (diameter 0.7–1.3 mm) are deposited into crevices of rocks or bark, and this stage lasts 60 to 380 days. Developmental stage I (first free-living larva) has a distinct prognathy, a strong forward projection of parts of the maxillae (laciniae), which aids in emergence from the egg, and rod-like setae on the head and the terga. These features disappear after the first molt. From developmental stage III onward, a scale cover is present in all terga and on abdominal coxites and caudal appendages. The molts continue during adult life.

See Also the Following Article

Zygentoma

Further Reading

Bitsch, C., and Bitsch, J. (1998). Internal anatomy and phylogenetic relationships among apterygote insect clades (Hexapoda). *Ann. Soc. Entomol. France (N.S.)* **34,** 339–363.

Mendes, L. F. (1990). An annotated list of generic and specific names of Machilidae (Microcoryphia, Insecta) with identification keys for the genera and geographical notes. *In* "Estudios, Ensaios e Documentos." *Centro Zool. Inst. Invest. Tropicall,* Lisbon.

Sturm, H., and Machida, R. (2001). Archaeognatha. *In* "Handbook of Zoology," *Vol. IV,* Part 37. Walter de Gruyter, Inc., Berlin/New York.

Wygodzinsky, P. W., and Schmidt, K. (1980). Survey of the Microcoryphia (Insecta) of the northeastern United States and adjacent provinces of Canada. *Ann. Mus. Novitates* **2701,** 1–17.

Arthropoda and Related Groups

James H. Thorp
University of Kansas

More than 75% of all described species in terrestrial, freshwater, and marine ecosystems are in the phylum Arthropoda. No other invertebrate phylum, with the possible exception of the nematodes, approaches their economic and ecological importance. This article briefly reviews all subphyla and classes of Arthropoda, with its nearly one million described species, and provides more details about major arthropod taxa not covered in other entries in the Encyclopedia. Aspects of arthropod evolutionary relationships, diversity, anatomy, physiology, and ecology are discussed.

OVERVIEW OF THE PHYLUM ARTHROPODA

The Arthropoda is a phylum more diverse than any other living or extinct animal taxon. Counted among this immense assemblage are beetles, butterflies, silverfish, centipedes, scorpions, mites, sea spiders, crabs, sow bugs, and barnacles, and many other common names too numerous to mention. Arthropods are the numerically dominant metazoan on land and rank among the most prominent benthic (bottom-dwelling) and planktonic members of freshwater and marine ecosystems. They colonize virtually every conceivable habitat—from the equator to the poles, from high mountains to deep ocean trenches, and from rain forests to deserts and hot springs—and fill all trophic niches above the level of primary producer. Parasitism, especially ectoparasitism, is common in some groups, but most species are free-living. They range in size from tiny gall mites (80 μm) to Japanese spider crabs with leg spans of 3.6 m. While some arthropods are vectors for human disease-causing organisms and others are major agricultural competitors with humans, they are also vital to the functioning of most ecosystems and a boon to humans in many ways. In addition to deriving nutrition from some arthropods (e.g., directly or indirectly from bees, crabs, lobsters, and shrimp), humans probably could not survive without arthropods.

The name "Arthropoda" is from the Greek, meaning "jointed foot." The presence of jointed appendages is the primary feature distinguishing arthropods from other phyla. Advantages provided by these appendages, a metameric or segmented body, and a hard skeleton are the three most important reasons for the phylum's success. Arthropods are segmented like annelid worms, but the evolutionary trend has been to fuse several metameres into body regions (tagmata) with specialized functions. Spiders have two tagmata, insects have three, and many crustaceans have two; however myriapods (millipedes and centipedes) lack tagmata. Arthropods have chitinous and proteinaceous exoskeletons that are frequently strengthened with calcium salts. A modest, nonchitinized endoskeleton of inwardly projecting apodemes aids muscular attachment. To allow for continued somatic growth, the exoskeleton is shed periodically during ecdysis, a relatively strenuous and often dangerous process. Modifications of the exoskeleton have permitted arthropods to fly, swim, run, and burrow effectively.

Except for the molluscan cephalopods (e.g., the octopus), arthropods surpass all invertebrates in internal organ complexity. Although they are a phylum with a coelom, this structure no longer serves as a hydrostatic skeleton (as in annelids) but persists only as a cavity surrounding reproductive and/or excretory organs. The principal body cavity is instead the hemocoel, which is derived from the circulatory system. The open circulatory system consists of a dorsal heart, blood sinuses, and one or more discrete vessels. Hemocyanin and hemoglobin are the principal oxygen-carrying blood pigments. Respiration is achieved through the skin surface in some small species, but with gills in most aquatic organisms and tracheae and/or book lungs in terrestrial species. Excretory and osmoregulatory organs vary in type in accordance with the typical environmental moisture and salt content, as do the primary excretory products (ammonia in water and usually either uric acid or guanine on land). Cilia are absent externally and internally. The neural system is highly developed, with brain centers and complex sensory organs; indeed, next to vertebrates and cephalopods, the arthropod brain is the most complex on earth.

Most species reproduce sexually and are primarily dioecious (i.e., with an individual being a single gender), although parthenogenesis occurs in many taxa. Courtship and brood care are uncommon but are found in some members of all subphyla. Rather than possessing the spiral cleavage typical of many other protostomates, arthropods usually develop by superficial cleavage of a cytoplasmic layer above a yolky sphere. Larvae or discrete juvenile stages are common in terrestrial and aquatic taxa, but aquatic larvae never resemble the trochophore larvae that characterize related phyla.

EVOLUTIONARY RELATIONSHIPS WITH OTHER PHYLA

Arthropods were traditionally linked with the phylum Annelida in the phylogenetic clade Articulata because both are metameric phyla, but more recent molecular analyses provide no support for a particularly close relationship. Instead, arthropods seem to be linked with other phyla that must shed their cuticle during ecdysis to grow. This clade of "Ecdysozoa" also includes the phyla Tardigrada, Onychophora, Nematoda, and Nematomorpha, along with the more distantly related Priapulida and Kinorhyncha. At a greater phylogenetic distance from Ecdysozoa is the other major protostomate clade consisting of the phyla Rotifera, Annelida, Mollusca, Bryozoa (Ectoprocta), Brachiopoda, and Phoronida. Pentastomida is listed by some authors as a separate phylum with links to Arthropoda through the clade Ecdysozoa, but here it is included as a class within the arthropod subphylum Crustacea.

Within Ecdysozoa, the three closest phyla are Arthropoda, Tardigrada, and Onychophora. This conclusion is based in part on molecular studies using 18*S* rRNA. Fossil evidence from the mid-Cambrian (~520 mya) suggests that onychophoran-like limbs developed in aquatic invertebrates and may have served as a preadaptation for terrestrial life. Similarities in morphology and physiology also seem to link these phyla. For example, tardigrades possess striated muscles, paired ventral nerve cords, and a large hemocoel. In addition to these characteristics, onychophorans have the following arthropod-like features: a tracheal respiratory system, mandible-like mouth appendages, cardiac ostia, an excretory system comparable to the green gland of crustaceans, one pair of antennae, and similar defensive secretions produced by repugnatorial glands. Both Onychophora and Tardigrada, however, have some decidedly non-arthropod-like characteristics (e.g., nonjointed legs). Onychophora, Myriapoda, and Hexapoda are grouped by some systematists into Uniramia, a single phylum of arthropod-like animals having a single branch (ramus) of body appendages. According to this theory Uniramia is phylogenetically isolated from Crustacea and Chelicerata, but all are in the superphylum Arthropoda.

TAXONOMIC DIVERSITY AND INTRAPHYLETIC AFFILIATIONS

Arthropoda is treated here as a monophyletic clade of genetically diverse but evolutionarily linked species. Some zoologists, however, maintain that this alleged phylum is actually an artificial, polyphyletic grouping of similar taxa evolving multiple times from different prearthropod ancestors. Much of this debate has centered on evolutionary relationships between the phyla Arthropoda and Onychophora. Classified within Arthropoda are one extinct subphylum (sometimes called super class), the Trilobitomorpha (trilobites), and four living subphyla: Chelicerata (spiders, mites, horseshoe crabs, and sea spiders), Myriapoda (millipedes and centipedes), Hexapoda (springtails, bristletails, beetles, flies, true bugs, etc.), and Crustacea (crayfish, barnacles, water fleas, pill bugs, etc.). Sometimes the number of extant subphyla is reduced to three (Chelicerata, Uniramia, and Crustacea) or even two groups (Chelicerata and Mandibulata).

Molecular studies of arthropod phylogeny present a reasonably clear picture of relationships among three of the four living subphyla. Chelicerates are evolutionarily distinct from insects and crustaceans, and they differ from all other living arthropods in lacking a tagma for either a "head and trunk" or a "head, thorax, and abdomen." Instead, they possess an anterior prosoma without a distinct head and a posterior opisthosoma. Another major clade evident from gene sequences is Mandibulata, composed of the other three extant subphyla. Morphological observations of appendages would seem then to link Myriapoda and Hexapoda into a group (Uniramia) of taxa with only one branch to each appendage and distinct from the biramous Crustacea, but molecular evidence is inconclusive on this point. In some gene sequence trees, myriapods are tightly linked with insects, while other molecular analyses show the millipedes and centipedes as deeply entangled within other genetic branches.

Accurate estimates of both relative and absolute diversities of arthropods are often problematic because of the enormous species richness, large number of unexplored habitats, greater emphasis on studies of economically important taxa, and increasingly serious lack of qualified taxonomists. For those reasons, the literature is replete with divergent estimates of the total number of species in most groups, especially the insects and mites. Table I lists the classes of Arthropoda and includes estimates of taxonomic diversity.

SUBPHYLUM TRILOBITOMORPHA

Trilobites probably played a crucial role in the evolution of living arthropods. Members of the now extinct subphylum Trilobitomorpha began roaming primeval seas in the Precambrian, reached their zenith in the late Cambrian with 4000–10,000 species, and then slowly went extinct around 230 to 275 mya. They were flattened, bilaterally symmetrical arthropods with bodies divided by longitudinal and transverse grooves unlike living arthropods (Fig. 1). Most adults were 2 to 7 cm in length, but giants of 50 cm are known. Their organ systems probably resembled those in modern arthropods. Smaller species were probably planktonic suspension feeders, whereas most species and all larger species were probably benthic deposit feeders or facultative predators. Unlike most other arthropods, trilobites

TABLE I Estimates of Arthropod Diversity, with Comments on Certain Arthropod Taxa and Two Related Phyla[a]

Taxon	Estimated number of species (families)	Biological features
Phylum Arthropoda	1–5 million	
Subphylum Trilobitomorpha	4,000–10,000	Extinct marine trilobites
Subphylum Chelicerata	77,000–1 million	Originally marine, but subsequent evolution has primarily been in terrestrial mites
Class Merostomata	5	Marine horseshoe crabs and extinct sea scorpions
Class Arachnida	~76,000 (550)	Spiders, scorpions, and mites
Class Pycnogonida	1,000 (8+)	Sea spiders
Subphylum Myriapoda	13,000 (140+)	Terrestrial millipedes, centipedes, and others
Class Chilopoda	3,000 (20)	Predaceous centipedes
Class Symphyla	160 (2)	Small (1–8 mm), mostly herbivorous; live in forest litter
Class Diplopoda	10,000 (120)	Millipedes
Class Pauropoda	500 (5)	Minute (< 1.5 mm) dwellers in forest litter
Subphylum Hexapoda	1–4 million	Insects, springtails, bristletails, etc.
Class Ellipura	6,000+	Wingless, entognathous (orders Protura and Collembola, or springtails)
Class Diplura	800 (9)	Blind, wingless inhabitants of forest litter; entognaths
Class Insecta	1–4 million species	Winged and wingless insects; all adults with six pairs of legs
Subclass Apterygota	600 (5)	Primitive, wingless insects (order Thysanura with bristletails and silverfish)
Subclass Pterygota	~1 million+ (906)	Mostly winged insects (grasshoppers, true bugs, beetles, flies, butterflies, ants, etc.)
Subphylum Crustacea	~50,000	Shrimp, crabs, waterfleas, barnacles, copepods, etc.
Class Cephalocarida	9 (2)	Primitive; live in soft marine sediments
Class Malacostraca	29,000 (103)	Crabs, water scuds, isopods, mantis shrimp, etc.
Class Branchiopoda	1,000 (29)	(= Phyllopoda) water fleas and brine, clam, pea, and tadpole shrimp
Class Ostracoda	6,650 (46+)	Seed shrimp enclosed in a bivalved chitinous carapace
Class Mystacocarida	11	Interstitial species living in shallow water or intertidally
Class Copepoda	8,000 (97)	Dominant crustaceans in zooplankton; a few parasites of marine fish and invertebrates
Class Branchiura	125 (1)	Fish lice (ectoparasites)
Class Pentastomida	100 (7)	Highly modified parasites of tetrapod vertebrates
Class Tantulocarida	10+ (9)	Deep-water parasites of crustaceans; some sources estimate diversity up to 1200 species
Class Remipedia	9+	Ancient, vermiform crustaceans found in marine caves; some estimate diversity as high 1200 species
Class Cirripedia	1,000 (20)	True barnacles and small groups of parasitic taxa
Phylum Onychophora	70 (2)	Velvet worms; most confined to tropical habitats
Phylum Tardigrada	800 (17)	Water bears in aquatic and moist terrestrial habitats

[a] Estimates of species richness are for living taxa only, except for the subphylum Trilobitomorpha. The reliability of these estimates varies widely among taxa.

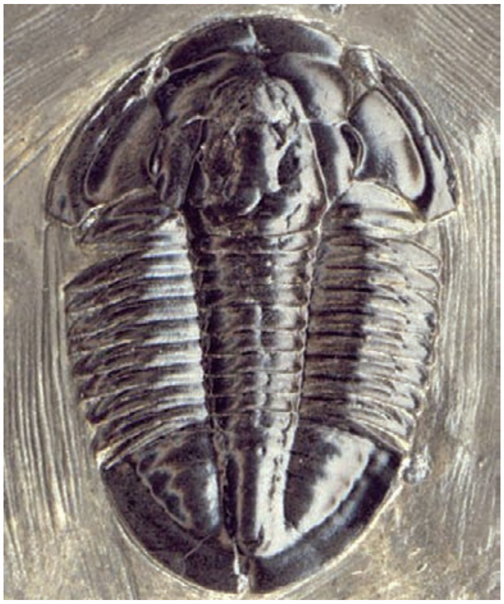

FIGURE 1 *Asaphiscus wheeleri,* an extinct trilobite in the subphylum Trilobitomorpha. (Photograph courtesy of Sam M. Gon, III.)

proceeded gradually through a life with three larval stages, 14 or more juvenile steps in the first year of life, and multiple adult stages lasting a maximum of 3 more years.

SUBPHYLUM CHELICERATA

Few invertebrates on land or sea are so often miscast in a sinister role as the spider, which along with mites, sea spiders, and horseshoe crabs comprises the 77,000 described species in the subphylum Chelicerata. Indeed, most arachnids and all other chelicerates either are harmless to humans or are actually quite helpful in their roles as predators and parasites of insects or as decomposers of terrestrial litter. Although the subphylum evolved in the sea, the ocean now supports only the five species of horseshoe crabs (class Merostomata), a thousand or so species of sea spiders (class Pycnogonida), and a few mites.

Class Merostomata: Horseshoe Crabs

During the full or new moons of late spring and early summer when tides are the highest, vast numbers of horseshoe crabs

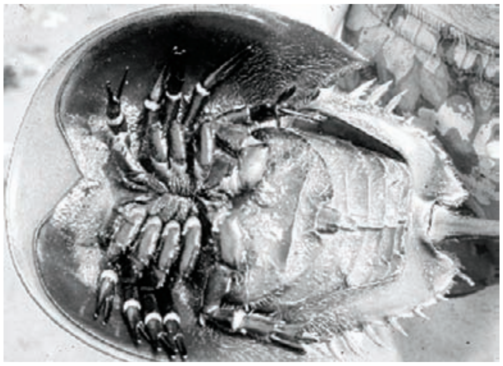

FIGURE 2 Representative of the subphylum Chelicerata: ventral view of an adult horseshoe crab, *Limulus polyphemus* (class Merostomata).

(all *Limulus polyphemus;* Fig. 2) come ashore in eastern North America to breed in bays and estuaries. The five Asian and North American species of this class are remnants of this strictly marine class. Although formidable looking and up to 60 cm in length, horseshoe crabs are harmless to humans. When not breeding, these chelicerates reside on or in soft bottoms in shallow water. They are scavengers and predators on clams and worms. Although they can swim weakly by flapping their book gills (modified abdominal appendages), their primary locomotion is walking. Extinct merostomates (order Eurypterida) may have been the largest arthropods ever to have evolved (nearly 3 m long) and seem to have given rise to terrestrial arachnids. Despite their name, horseshoe crabs are not closely related to true crabs (subphylum Crustacea).

Class Arachnida

During the middle Paleozoic, chelicerates made the rigorous transition from water to land; only later did over 5000 species of arachnid mites adopt a secondary aquatic existence. Arachnids then rapidly radiated in form and species richness in association with their predaceous and parasitic exploitation of insects. One crucial factor in this success has been the diverse uses of silk by spiders, pseudoscorpions, and some mites. Although some degree of "arachnophobia" afflicts many people, relatively few of the 76,000 described species are directly harmful to humans because of their venom, link with diseases and allergies, or competition for plant resources. More than balancing their negative attributes is the substantial role in biocontrol of insect pests.

SCORPIONS, SPIDERS, AND HARVESTMEN

True Scorpions The 1500 to 2000 species of true scorpions (order Scorpiones) are elders of the arachnid clan (Fig. 3). In addition to the true scorpions, several other arachnid orders are called "scorpions": false scorpions (Pseudoscorpiones with 2000 species), wind scorpions (Solifugae with 900 species), whip scorpions (Uropygi with 85 species), and tailless whip scorpions (Amblypygi with 70 species). All are much smaller than true scorpions, but are also typically carnivorous.

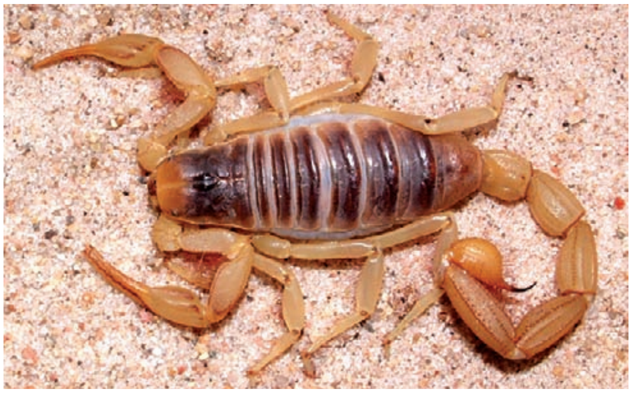

FIGURE 3 Desert hair scorpion, *Hadrurus arizonensis.* (Photograph by Jim Kalisch, courtesy of University of Nebraska Department of Entomology.)

True Spiders (Order Araneae) Most arachnids lack biting mouthparts and must, therefore, partially digest prey tissue before sucking it into their bodies. Prey are subdued with poison injected by fangs present on each chelicera. Arachnids reproduce with indirect fertilization (without a penis), often after elaborate courtship rituals. Their leglike pedipalps are used by males to transfer spermatophores. Eggs are wrapped in a protective silken cocoon, and brood care is common. Silk is produced normally by caudal spinnerets and by a small platelike organ (cribellum) in cribellate spiders only. Uses for this silk include cocoons, egg sacs, linings of retreats, and capture webs. Locomotion is typically by walking or jumping, but aerial dispersal through the process of "ballooning" with long silken threads is common in most spiderlings and adults of some smaller taxa. Most spiders are terrestrial and are found anywhere insects are located. All are carnivorous, and ecological divergence in prey type and capture method has led to the wide evolutionary radiation. In addition to insects, spiders attack other spiders, small arachnids, and a few other prey taxa including small vertebrates. Several spiders are poisonous to humans, such as the black widow and the brown recluse.

Harvestmen The order Opiliones includes arachnids known as "daddy longlegs," a name reflecting its enormously long walking legs. They are also called "harvestmen" because some species undergo a seasonal population explosion each autumn around the farm harvest. They have "repugnatorial glands" that produce an acrid secretion to repel predators. The 5000 species are more closely related to mites than to true spiders. Most are tropical, but taxa are known from colder subarctic and alpine zones. Opilionids frequent humid forest floors, being less arboreal than true spiders. Although carnivory on small arthropods and worms is common, harvestmen are notable as the only arachnids other than mites that consume vegetation.

MITES AND TICKS At least 30,000 species of arachnid mites and ticks have been described in the order (or subclass) Acari. The major habitat of mites is on land, where they are either free-living or parasites of plants and animals, but lakes, streams, and even hot springs support 5000 taxa, with the marine fauna being less diverse.

Many acarines are ectoparasitic in larval and/or adult stages. Animal parasites attack mammals (including humans and domestic animals), birds, reptiles, amphibians, fish, aquatic and terrestrial insects, other arachnids, and some other invertebrates, including echinoderms, mollusks, and crustaceans. They are vectors for several human diseases, and some (e.g., chiggers) have annoying bites. Many people develop allergies to mites living on household dust. On the other hand, microscopic mites commonly consume dead tissue and oily secretions on human faces, and they are used to control harmful insects and mites. Their feeding habits and role in spreading viruses make them severe pests of natural and agricultural plants. A great diversity of mites, however, are free-living, mostly in forest and grassland litter, where they feed directly on litter or on microorganisms decomposing detritus. Many mites prey on other mites, nematodes, and small insects.

Class Pycnogonida: Sea Spiders

The body shape and gangling legs of sea spiders call to mind their terrestrial namesakes. Most of the thousand species live in shallow benthic zones at higher latitudes. They are predominantly predators of hydroids, bryozoans, and polychaetes, but some consume microorganisms, algae, and even detritus. Food is either macerated with chelae or externally predigested and then sucked into the digestive tract with a proboscis.

Sea spiders have a barely perceptible head and a body comprising four pairs of long legs joined by a narrow, segmented trunk. Most are small (1–10 mm), but some deep-water behemoths reach 6 cm in body length with a 75-cm leg span. Their eyes are mounted on an tubercle to give them a 360° arc of vision.

SUBPHYLUM MYRIAPODA

Myriapoda (“many feet”) is a subphylum of elongate arthropods with bodies divided into a head and trunk with numerous segments, most of which have uniramous appendages; no pronounced tagmatization is evident. Myriapods range in length from 0.5 to 300 mm and are primarily terrestrial. Most live in humid environments, commonly in caves. Some have invaded arid habitats, but few are aquatic. Four classes are recognized: Diplopoda (millipedes), Chilopoda (centipedes), Pauropoda, and Symphyla, with 10,000, 3000, 500, and 160 species, respectively. The last two are minute dwellers of the forest floor that consume living or decaying vegetation. Symphylans look somewhat like centipedes but the adults have 14 trunk segments and 12 pairs of limbs; the posterior end of the trunk has two conical cerci and spinning glands. Members of the class Pauropoda are soft-bodied, blind myriapods with 9 to 11 leg-bearing trunk segments and branched antennae.

Class Chilopoda: Centipedes

Centipedes are dorsoventrally flattened with 15 to 173 segments, each with one pair of legs (Fig. 4A). Poisonous forcipules (fangs) enable centipedes to kill and consume insects, other centipedes, annelids, mollusks, and sometimes small vertebrates; under most circumstances, the poison is not lethal to people. The body is partially hung beneath the legs to increase stability and to allow hind legs to step over front ones, which allows the insects to run swiftly in search of prey or to escape predators. Centipedes are found in most terrestrial environments including the desert fringe; the latter is surprising given their chitinous, noncalcified exoskeleton, which is relatively permeable to water.

Class Diplopoda: Millipedes

Millipedes have a somewhat cylindrical body with 11 to 90 segments (which are really fused “diplosegments”) and two pairs of legs per segment. Segmental plates are constructed to prevent “telescoping” as the body bulldozes through forest litter, while still allowing the animal to roll up or coil when threatened (Fig. 4B). Millipedes are slow moving and herbivorous by nature, eating decaying leaves and wood. They lack poisonous fangs and instead repel predators with volatile poison produced by repugnatorial glands. Millipedes are relatively long-lived, with some surviving 7 years.

SUBPHYLUM HEXAPODA

Hexapoda (“six feet”) includes a tremendous diversity of winged insects (class Insecta, subclass Pterygota) and many fewer wingless insect (subclass Apterygota) and noninsect classes (Diplura and Ellipura, orders Collembola and Protura).

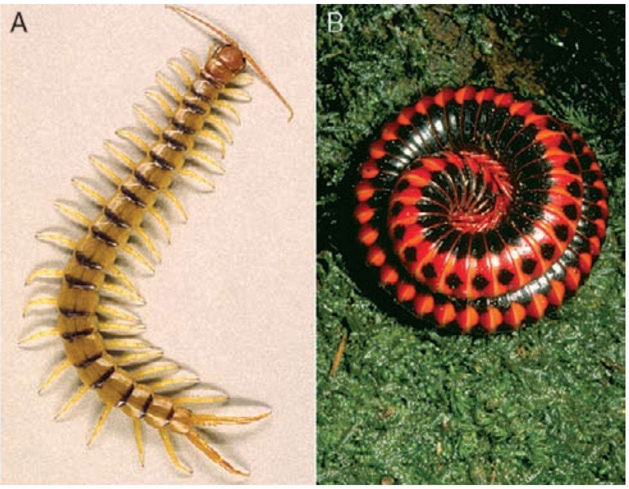

FIGURE 4 Members of the subphylum Myriapoda. (A) Centipede (class Chilopoda). (Photograph by Jim Kalisch, courtesy of University of Nebraska Department of Entomology.) (B) Millipede (class Diplopoda), coiled in a defensive posture. (Photograph by D. R. Parks.)

Hexapods have three major body regions (head, thorax, abdomen) and six thoracic legs.

Entognathous Hexapods: Collembola, Protura, and Diplura

Entognathous hexapods include two small taxa (class Diplura and Ellipura, order Protura) living in moist forest litter and a large group of springtails (class Ellipura, order Collembola) with at least 4000 species in terrestrial and semiaquatic environments. Most springtails live in moist terrestrial environments, but some colonize the surface film of quiet fresh and marine waters. They occur at densities much higher than almost any other invertebrate in soil litter. Unlike insects, springtails have only six abdominal segments, and cleavage of their eggs is total. Their name is derived from their ability to spring forward several centimeters when a forked structure (the furcula) flexed under the abdomen is rapidly uncocked. They have indirect fertilization, the young closely resemble adult Collembola, and adults continue molting throughout their lives (2–50+ molts). Springtails feed on decomposing organic matter or on microorganisms at the water surface. Proturans are completely terrestrial, their antennae have nearly atrophied away, and their front legs function somewhat like antennae. Diplurans are primitive hexapods whose ancestors may have given rise to both Protura and Collembola, and they are more closely related to insects than are ellipurans. Diplurans are blind and have two prominent abdominal cerci.

Class Insecta: Winged and Wingless Insects

The million or so species in the subclass Pterygota include all winged invertebrates and some insect species that have secondarily lost wings during evolution. They include two orders of ancient winged insects (Ephemeroptera and Odonata) and some 25 to 30 (depending on the classification system) orders of modern folding-wing insects. Most have 11 abdominal segments. The head features two antennae and compound eyes. Respiration is generally with internal tracheae, but aquatic species may use external, tracheate gills or other means to obtain sufficient oxygen. Fertilization is usually direct, distinct developmental stages are common, and molting generally stops with attainment of reproductive maturity. Their most prominent features are two pairs of wings, but a great many insects (e.g., fleas) lack wings or have dispensed with either the hind (e.g., flies) or fore pair (beetles). Among their beneficial attributes are pollination of most flowering plants, production of honey and silk, predation on harmful insects, decomposition of animal wastes and carcasses, and facilitation of ecological processes at all trophic levels above primary producer. Negative attributes include transmission of diseases, annoying bites, and damage to crops, stored food, ornamental plants, forests, and wooden structures.

All insect species that did not evolve from a winged hexapod and whose adults all lack wings are in the subclass Apterygota, order Thysanura. This small group of 600 or so primitive species includes bristletails, silverfish, and rock jumpers. These are small to medium-sized insects (5–25 mm) without compound eyes. They have an 11-segmented abdomen with a prominent caudal filament between two terminal cerci. Fertilization is indirect, and molting continues after the reproductive state has been reached (unlike insects). No pronounced metamorphosis is evident from subadult to adult stages. Thysanurans are swift, agile runners (probably to avoid predators) and are omnivorous scavengers of animal and plant matter. Most live in litter of forests and grasslands, but silverfish also infest houses, where they can extensively damage clothing and books.

SUBPHYLUM CRUSTACEA

Crustaceans surpass all other invertebrates in their direct contribution to human diets (from crabs, shrimp, lobsters, and crayfish) and are vitally important to many ecosystems, especially planktonic food webs. Unfortunately, they also foul boat hulls (barnacles) and destroy wooden piers in coastal waters (burrowing isopods).

Distinguishing characteristics of adults include the following: five-pairs of cephalic appendages (two mandibles, four maxillae, and two antennae), two to three tagmata, a chitinous cuticle often elaborated as a shieldlike carapace, more than 11 abdominal segments, and jointed, biramous appendages. Evolutionary trends involved specialization of mouthparts, body segments, and appendages for locomotion, sensory reception, and reproduction. Respiration is typically with gills, and hemocyanin is the principal respiratory pigment. Excretion of ammonia generally occurs through modified nephridia. Sexes are mostly separate, but hermaphroditism is common. Development always includes triangular nauplius larvae (with six appendages and a median eye), which are commonly planktonic. Many crustaceans have a relatively sophisticated behavioral repertoire and communicate visually, tactilely, and chemically.

Most of the roughly 50,000 species are marine, but crustaceans are ubiquitous in freshwater habitats and a few species have colonized saline lakes and terrestrial environments. Crustaceans are most often scavenging predators or have a generally omnivorous diet. They range in length from minute to truly gigantic (0.25 mm to 360 cm).

Class Malacostraca: Shrimp, Crayfish, and Others

About 60% of crustacean species are malacostracans, including all species consumed by humans (mostly decapods) and many important benthic crustaceans in marine and freshwater ecosystems (e.g., amphipods and isopods).

Most orders and the 29,000 species in this class are dominated by crustaceans that live near the bottom. Ecologically important exceptions include planktonic krill (order Euphausiacea), a vital prey of baleen whales. Parasitism is relatively rare and is confined mostly to ectoparasites of fish, crustaceans (both mostly in the large order Isopoda), and marine mammals (small component of the large order Amphipoda).

Class Branchiopoda: Water Fleas, Brine Shrimp, and Others

Nearly 25% of all freshwater crustaceans are branchiopods ("gill feet"), and almost all crustaceans in saline lakes are in this class; only 3% of the species occur in oceans. The class contains about 800 species of water fleas (in four orders of cladocera), 200 species of fairy and brine shrimp (order Anostraca), and a few taxa of clam (order Spinicaudata), pea (Laevicaudata), and tadpole shrimp (Notostraca). Most taxa other than cladocerans are restricted to ephemeral pools, and most branchiopods are suspension feeders. Branchiopods are noted for producing dormant embryos resistant to adverse environments.

Class Ostracoda: Seed Shrimp

The 6650 species of seed or mussel shrimp are minute crustaceans characterized by a protective, dorsal, bivalved carapace of chitin heavily impregnated with calcium carbonate. Freshwater species rarely exceed 3 mm in length, but marine taxa can reach 30 mm. Ostracodes are especially prevalent in freshwater habitats (particularly benthic areas) but are common in marine environments. A few genera have adapted to damp humus habitats of the forest floor. Ostracods are typically suspension feeders on benthic and pelagic detritus, and almost all are free-living.

Class Copepoda

In relatively permanent freshwater and marine environments, the 8000 species of copepods are the most important planktonic crustaceans, and other species contribute to the benthic fauna. Herbivory on microalgae prevails, but raptorial feeding on other zooplankton is common. Although freeliving copepods predominate, bizarre forms, barely recognizable as copepods, have evolved as parasites of fish and invertebrates.

Class Cirripedia: Barnacles

The name "barnacle" evokes a rugged image of the sea in the minds of many people, but few recognize this taxon's kinship with familiar crustaceans like shrimp and crabs because the barnacle's body is hidden inside calcareous plates and free-living species are sedentary. The 1000 species in the class Cirripedia include free-living species that live directly on a hard surface or are raised on a stalk (peduncle). This surface may be an inanimate object (e.g., rocks, floating refuse, ship hulls) or the outer layer of a living whale, turtle, invertebrate, or other larger organism. Barnacles usually cement the head to hard surfaces and employ setose legs to capture plankton. Other species, barely recognizable as barnacles, are ecto- and endoparasites of echinoderms, corals, and other crustaceans, especially crabs.

Pentastomida and Other Crustaceans

The remaining 264 to 1500 species of crustaceans are divided among six classes. Of these, only the ectoparasitic fish lice (class Branchiura), with 125 species, were recognized more than a few decades ago. Three classes closely resemble copepods—Cephalocarida, Mystacocarida, Tantulocarida—with the first two living in marine sediments and the last being ectoparasites of deep-water crustaceans. The class Tantulocarida is usually listed with 10 to 20 species, but some scientists believe that the true diversity is greater than 1000. Members of the class Remipedia are presently restricted to tropical underwater caves. Their long bodies with abundant lateral appendages call to mind segmented polychaete worms.

CLOSELY RELATED PHYLA: TARDIGRADA AND ONYCHOPHORA

Tardigrada: Water Bears

Tardigrades are called "water bears" because of their slow lumbering gait and relatively massive claws on lobopodous legs. The permeability of the cuticle limits tardigrades to aquatic habitats (often interstitial), the surface film of terrestrial mosses, and damp soil. About 800 species have been described, but many more undoubtedly exist in unexplored habitats. Faced with inhospitable microhabitats from environmental changes, both terrestrial and aquatic species may undergo cryptobiosis, where the body becomes dehydrated and metabolism is greatly reduced until favorable conditions return. Some water bears have been "resuscitated" from this state after decades! These normally dioecious organisms can also reproduce by parthenogenesis. Tardigrades typically feed on plants cells pierced by a pair of mouth stylets and sucked into the gut, but a few species are carnivorous.

Onychophora: Velvet Worms

Velvet worms are giants compared with tardigrades, for some individuals reach a length of 15 cm, but they share many characteristics with this sister phylum of Arthropoda. They are generally nocturnal and move by extending their legs, with hydrostatic pressure generated by muscular contraction within the legs.

Though most are herbivores or omnivores, many species consume small arthropods in a rather unique manner. They attack their prey and also defend themselves from predators by expelling a sticky, proteinaceous substance that entangles the target. This glue is produced by slime glands within oral papillae. The onychophoran then bites and secretes toxins to kill and partially liquefy the victim.

See Also the Following Articles

Biodiversity • Fossil Record • Insecta, Overview

Further Reading

Anderson, D. T. (ed.) (1998). "Invertebrate Zoology." Oxford University Press, Melbourne, Australia. (See especially Chaps. 10–14, 18.)

Bliss, D. E. (ed.-in-chief). (1982–1985). "The Biology of Crustacea," Vols. 1–10. Academic Press, New York.

Parker, S. P. (ed.). (1982). "Synopsis and Classification of Living Organisms," Vol. 2, McGraw-Hill, New York.

Pechenik, J. A. (2000). "Biology of the Invertebrates." 4th ed. McGraw-Hill, New York (Particularly Chaps. 14 and 15.)

Thorp, J. H., and Covich, A. P. (eds.) (2001). "Ecology and Classification of North American Freshwater Invertebrates." 2nd ed. Academic Press, San Diego, CA. (See especially Chaps. 16–23.)

Auchenorrhyncha

(Cicadas, Spittlebugs, Leafhoppers, Treehoppers, and Planthoppers)

C. H. Dietrich
Illinois Natural History Survey

The hemipteran suborder Auchenorrhyncha is the group of sapsucking insects comprising the modern superfamilies, Cercopoidea (spittlebugs, Fig. 1), Cicadoidea (cicadas, Fig. 2), Membracoidea (leafhoppers and treehoppers, Fig. 3), and Fulgoroidea (Fig. 4) Together, these groups include over 40,000 described species. Morphologically, Auchenorrhyncha differ from other Hemiptera in having the antennal flagellum hairlike (aristoid), the rostrum (modified, beaklike labium) arising from the posteroventral surface of the head, a complex sound-producing tymbal apparatus, and the wing-coupling apparatus consisting of a long, downturned fold on the

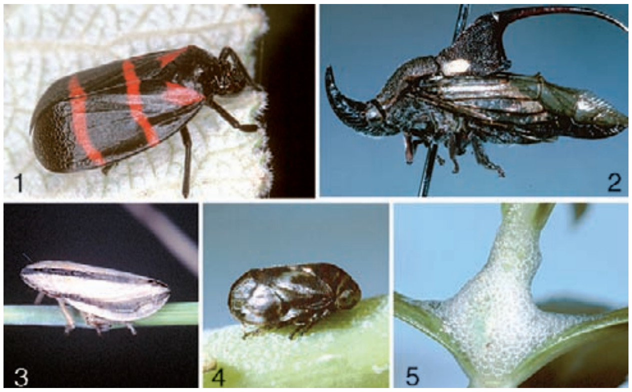

FIGURE 1 Cercopoidea: spittlebugs and froghoppers: (1) *Tomaspis* sp. (Cercopidae), Mexico, (2) *Machaerota* sp. (Machaerotidae), Vietnam, (3) *Paraphilaenus parallelus* (Aphrophoridae), Kyrgyzstan, (4) *Clastoptera obtusa* (Clastopteridae), Illinois, U.S.A., (5) spittle mass of *P. spumarius* nymph, Illinois, U.S.A.

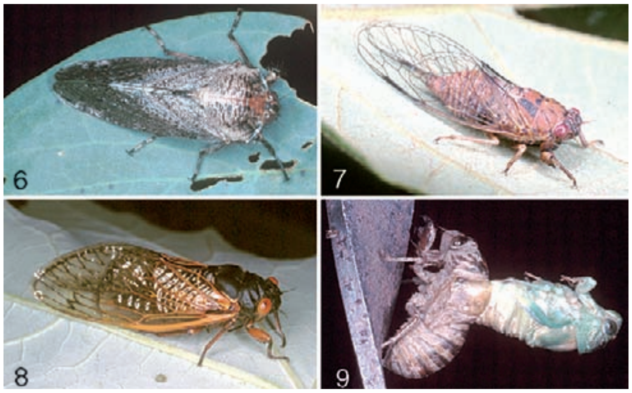

FIGURE 2 Cicadoidea: cicadas: (6) a hairy cicada, *Tettigarcta crinita* (Tettigarctidae), Australia, (7) *Melampsalta calliope* (Cicadidae), Illinois, U.S.A., (8) a periodical cicada, *Magicicada cassini,* with a 13-year life cycle, Illinois, U.S.A., (9) a dog day cicada, *Tibicen* sp., molting into the adult stage, Illinois, U.S.A.

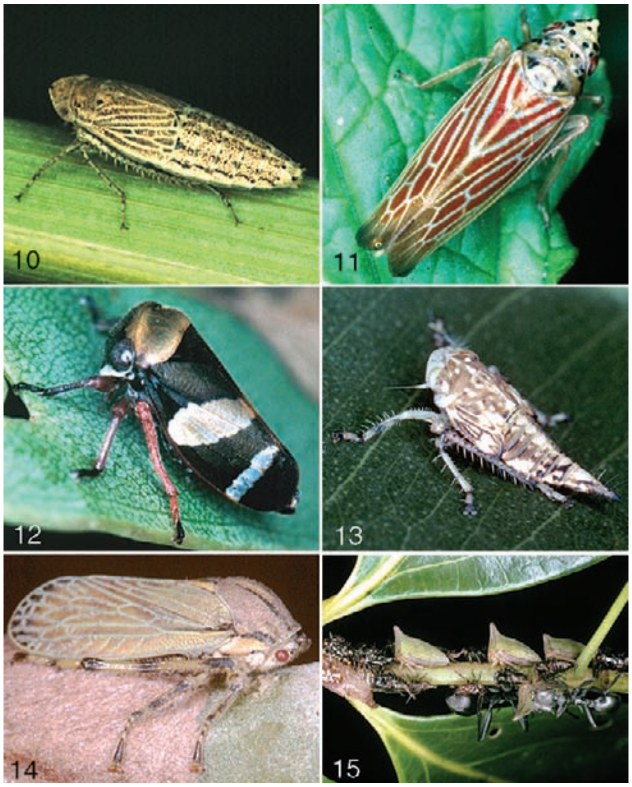

FIGURE 3 Membracoidea: leafhoppers and treehoppers: (10) a brachypterous, grass-feeding leafhopper, *Doraturopsis heros,* Kyrgyzstan, (11) *Pagaronia triunata* (Cicadellidae), California, U.S.A., (12) *Eurymeloides* sp. (Cicadellidae), Australia, (13) fifth instar of *Neotartessus flavipes* (Cicadellidae), Australia, (14) female *Aetalion reticulatum* (Aetalionidae) guarding egg mass, Peru, (15) ant-attended aggregation of treehopper adults and nymphs (Membracidae: *Notogonia* sp.), Guyana.

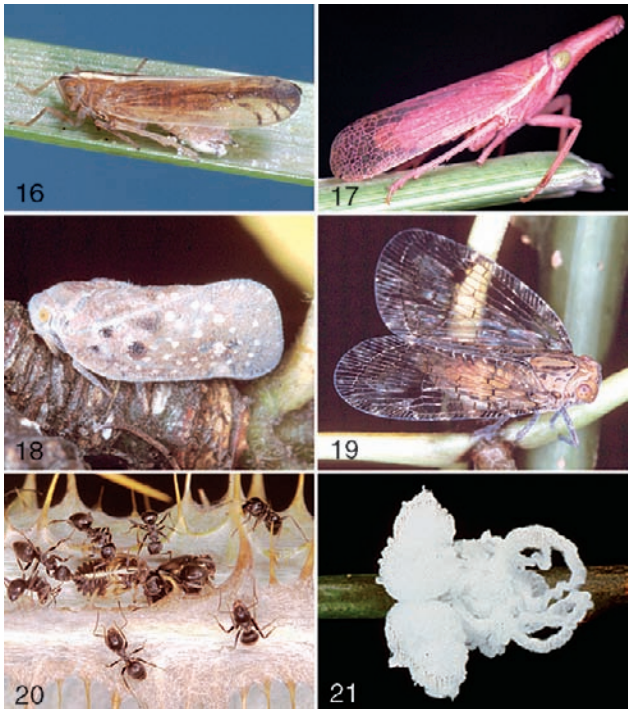

FIGURE 4 Fulgoroidea: planthoppers: (16) female *Stenocranus* sp. (Delphacidae) covering oviposition site with wax, Illinois, U.S.A., (17) *Chanithus scolopax* (Dictyopharidae), Kyrgyzstan, (18) *Metcalfa pruinosa* (Flatidae), Maryland, U.S.A., (19) *Biolleyana* sp. (Nogodinidae), Mexico, (20) *Tettigometra* sp. (Tettigometridae) nymphs tended by ants, Greece, (21) unidentified planthopper nymph completely covered with wax filaments, Guyana.

forewing and a short, upturned lobe on the hind wing. Auchenorrhyncha are abundant and ubiquitous insects, distributed worldwide in nearly all terrestrial habitats that support their host plants, but they are particularly diverse and speciose in the tropics. Some are important agricultural pests, injuring plants either directly through feeding and oviposition, or indirectly through the transmission of plant pathogens.

PHYLOGENY AND CLASSIFICATION

Nomenclature

The monophyly of the four existing superfamilies of Auchenorrhyncha has long been accepted, but controversy persists regarding the relationships of these lineages to each other and to various other fossil and extant hemipteran lineages. Consequently, no single classification scheme has gained universal acceptance, and the nomenclature of the various groups is presently unstable. Traditionally, Auchenorrhyncha were treated as one of three suborders of the order Homoptera. Fossil evidence, as well as phylogenetic analyses based on DNA sequences of extant taxa, suggest that Heteroptera (true bugs; Hemiptera, *sensu stricto*) arose from within Homoptera and, possibly, from within Auchenorrhyncha. Thus, many recent workers have combined Homoptera and Heteroptera into a single order. This order is usually referred to as Hemiptera (*sensu lato*), but some entomologists advocate using the ordinal name Rhynchota to avoid confusion with the more restricted definition of Hemiptera (Heteroptera) widely used in the literature. Some recent workers have further proposed dividing the Auchenorrhyncha into two suborders: Clypeorrhyncha for the lineage comprising Cicadoidea, Cercopoidea, and Membracoidea, and Archaeorrhyncha for Fulgoroidea. The older names Cicadomorpha and Fulgoromorpha, respectively (usually treated as infraorders within suborder Auchenorrhyncha), are more commonly used for these two groups. For convenience, and because the phylogenetic status of the group has not been elucidated satisfactorily, Auchenorrhyncha is retained here as the subordinal name with the caveat that this group may represent a paraphyletic assemblage rather than a monophyletic group. The current classification of families is presented in Table I.

TABLE I Classification of the Hemipteran Suborder Auchenorrhyncha (synonyms and common names in parentheses) Excluding Extinct Taxa

- Auchenorrhyncha (Cicadinea)
  - Infraorder Cicadomorpha (Clypeorrhyncha, Clypeata)
    - Superfamily Cercopoidea (spittlebugs, froghoppers)
      - Aphrophoridae
      - Cercopidae
      - Clastopteridae
      - Machaerotidae
    - Superfamily Cicadoidea (cicadas)
      - Cicadidae (Platypediidae, Plautillidae, Tettigadidae, Tibicinidae)
      - Tettigarctidae (hairy cicadas)
    - Superfamily Membracoidea (Cicadelloidea)
      - Aetalionidae (Biturritiidae)
      - Cicadellidae (Eurymelidae, Hylicidae, Ledridae, Ulopidae, leafhoppers)
      - Melizoderidae
      - Membracidae (Nicomiidae, treehoppers)
      - Myerslopiidae (Cicadellidae, in part)
  - Infraorder Fulgoromorpha (Archaeorrhyncha, planthoppers)
    - Superfamily Fulgoroidea
      - Acanaloniidae
      - Achilidae
      - Achilixiidae
      - Cixiidae
      - Delphacidae
      - Derbidae
      - Dictyopharidae
      - Eurybrachidae
      - Flatidae
      - Fulgoridae (lanternflies)
      - Gengidae
      - Hypochthonellidae
      - Issidae
      - Kinnaridae
      - Lophopidae
      - Meenoplidae
      - Nogodinidae
      - Ricaniidae
      - Tettigometridae
      - Tropiduchidae

Fossil Record

Auchenorrhyncha arose in the Paleozoic, first appearing in the fossil record in the Lower Permian (280 mya) and, judging from the abundance of forms described from Permian strata, they diversified explosively. These early auchenorrhynchans had adults with well-developed jumping abilities and somewhat resembled modern leafhoppers and spittlebugs, but nymphs (juveniles) associated with these insects were bizarrely flattened or biscuitlike, with short legs, foliaceous lobes on the head, thorax, and abdomen (similar to those of some modern Psyllidae) and elongate mouthparts, suggesting a sessile, cryptic lifestyle. The fulgoromorphan and cicadomorphan lineages (Table I) apparently diverged by the middle Permian. By the late Permian, Fulgoroidea appeared and Cicadomorpha (*sensu lato)* had diverged into the Pereboreoidea, comprising three extinct families of large cicada-like insects, and the smaller Prosboloidea, from which the three modern cicadomorphan superfamilies apparently arose. Cicadomorphans with a greatly inflated frontoclypeus (Clypeata in the paleontological literature = Clypeorrhyncha) did not appear until the Mesozoic. Prior to that, the head of Cicadomorpha resembled that of modern Psyllidae in having the frontovertex extended ventrad on the face to the antennal ledges and the lateral ocelli situated close to the eyes. This change in head structure is thought to have been associated with a shift from phloem to xylem feeding. Xylem feeding was apparently the predominant feeding strategy of the group throughout the Mesozoic, but in the late Cretaceous or early Tertiary the major lineages of phloem-feeding leafhoppers and treehoppers, which predominate in the recent fauna, arose. In these insects, the frontoclypeus became more flattened, probably because of the reduction in size of the cibarial dilator muscles. This was presumably in response to a shift from feeding on xylem, which is under negative pressure, to phloem, which is under positive pressure. Cicadoidea and Cercopoidea first appeared in the Triassic, and Membracoidea in the Jurassic. With the exception of Tettigarctidae, which arose in the late Triassic and is now confined to Australia, extant families of these groups do not appear in the fossil record until the Cretaceous or early Tertiary. Most Auchenorrhyncha from Baltic and Dominican amber of the Tertiary age are virtually indistinguishable from modern forms.

LIFE HISTORY

Courtship

Adult male and female Auchenorrhyncha locate each other by means of species-specific acoustic courtship signals. These signals are produced by specialized organs at the base of the abdomen called tymbals, present in both sexes (except female cicadas). A few cicadas and planthoppers are also able to use the stridulatory surfaces of their wings to produce sound. The loud, sometimes deafening, calls of many male cicadas are well known. In noncicadoids, the courtship calls are usually inaudible, being transmitted through the substrate, and distinct tympana are absent. The calls of some leafhoppers and planthoppers, audible only with special amplifying equipment, are among the most complex and beautiful of any produced by insects. Males move from plant to plant, signaling until they receive a response from a female. In addition to intensification of the vibrational signals, precopulatory behavior in some species may involve the male buzzing or flapping the wings, tapping the female with the legs, or repeatedly walking around or over the female. Copulation involves insertion of the male aedeagus into the female vulva at the base of the ovipositor and may last from a second or less to several hours, depending on the species. Females of most species seem to mate only once, while males often mate several times.

Oviposition and Nymphal Development

Females lay eggs singly or in batches, usually either by inserting them into plant tissue or by depositing them on plant surfaces [Figs. 3(14), 3(15), and 4(16)]. In some groups, eggs are deposited in the soil or in litter. Egg batches may be covered with plant debris, wax filaments, or secretions produced by various internal glands. Eggs may or may not undergo diapause depending on the species and climate. After hatching, the juveniles [nymphs, Figs. 2(9), 3(13), 3(15), 4(20), and 4(21)] undergo five molts prior to reaching the adult stage. In most species the nymphs feed on aboveground parts of host plant, but in cicadas, Cercopidae, a few fulgoroid families, and a few leafhopper genera, the nymphs are subterranean root feeders. Formation of galls, common among aphids and psyllids, is known in only one Auchenorrhyncha species (a leafhopper). Nymphal development requires from a few weeks to several years (in cicadas), depending on the species. Some species exhibit parental care behavior (see later).

BEHAVIOR AND ECOLOGY

Feeding and Digestion

Adult and nymphal Auchenorrhyncha feed by inserting the two pairs of feeding stylets (modified mandibles and maxillae) into the host plant tissue, injecting saliva, and ingesting fluid. Unlike Sternorrhyncha, in which the stylets pass between the cells of the host tissue (intercellular feeding), Auchenorrhyncha stylets usually pierce the cells (intracellular feeding). After selecting an appropriate feeding site based on visual and chemical cues, the insect presses the tip of the labium onto the plant surface and inserts the feeding stylets. Just prior to, and during probing of the plant tissue with the stylets, the insect secretes sheath saliva that hardens on contact with air or fluid to form an impervious salivary sheath surrounding the stylets. The sheath forms an airtight seal that prevents leakage of air or fluid during feeding. Stylet probing continues until a suitable tissue is found (xylem, phloem, or mesophyll, depending on the species), after which feeding can commence. During feeding, watery saliva is injected into the plant to aid digestion and to prevent clogging of the stylet opening. This is also the mechanism by which the insect may infect the plant with pathogens (see later). Feeding may last from a few seconds to many hours at a time, depending on the auchenorrhynchan species and the quality of the plant tissue. During feeding, droplets of liquid excretion are ejected from the anus, several droplets per second in some xylem feeders.

Plant sap is a nutritionally imbalanced food source; phloem is high in sugar and xylem is, in general, nutrient poor and extremely dilute. Auchenorrhyncha have acquired various adaptations that enable them to convert the contents of plant sap into usable nutrients. Most Cicadomorpha have part of the midgut modified into a filter chamber that facilitates rapid removal of excess water. Fulgoroidea lack a distinct filter chamber but have the midgut tightly coiled and partially or completely enclosed in a sheath of specialized cells that apparently absorb solutes from the gut contents. A broad array of transovarially transmitted (i.e., from the mother through her eggs to her offspring) prokaryotic endosymbionts have also been identified in various Auchenorrhyncha species. The roles of these endosymbionts have not been fully elucidated, but presumably they function in the conversion of the nutritionally poor plant sap on which the insects feed into essential vitamins, amino acids, and sterols. The symbionts are housed either intracellularly in specialized fat body cells called mycetocytes, intracellularly in the fat body, or in the gut epithelium. Several distinct mycetomes, consisting of groups of mycetocytes, are often present. In Cicadomorpha, each mycetome may house up to six different kinds of endosymbiont. In Fulgoroidea, only a single kind of endosymbiont is housed in each mycetome.

Host Associations

Nearly all Auchenorrhyncha are plant feeders; the few known exceptions (e.g., Fulgoroidea: Achilidae and Derbidae) feed on fungi as nymphs. Auchenorrhynchans use a wide variety of plants including mosses, horsetails, ferns, cycads, conifers, and angiosperms, but the vast majority of species feed on flowering plants. Most species appear to be restricted to a single genus or species of plants. Many species, particularly among the xylem-feeding groups, normally use a few or a single plant species but are capable of feeding and developing on a variety of alternate hosts if the preferred host is not available. A few xylem-feeding species have extremely broad host ranges. For example, the meadow spittlebug, *Philaenus spumarius,* with over 500 documented food plants, has the broadest known host range of any herbivorous insect. Phloem- and mesophyll-feeding species, comprising the majority of Auchenorrhyncha, tend to have narrower host ranges than xylem feeders, and many species appear to use a single plant family, genus, or species. Host associations appear to be conservative in some auchenorrhynchan lineages. Delphacidae and Cicadellidae (Deltocephalinae) include large numbers of grass- and sedge-specialist species and are among the dominant herbivores in grasslands. Most of the major lineages of Auchenorrhyncha do not exhibit a distinct preference for any particular plant taxon and usually include both host-generalist and host-specialist species. Some species alternate hosts during different stages in the life cycle or in different seasons. For example, nymphs of many leafhoppers and treehoppers develop on herbs, but the adult females oviposit on a woody host.

Migration

Most species of Auchenorrhyncha are relatively sedentary, completing their life cycle within a small area. Although most species have well-developed wings and are strong fliers, few seem to move more than a kilometer from their birthplace. Many species, particularly those inhabiting grasslands and deserts, are submacropterous or brachypterous [short winged, Fig. 3(10)] and, thus, incapable of sustained flight. Some of these species occasionally produce macropterous (long-winged) females that move to new patches of suitable habitat. Other species produce both short- and long-winged forms either simultaneously or in alternate generations. The proportion of macropterous to brachypterous forms often varies in response to population density. Some Auchenorrhyncha species undergo annual migrations that may cover hundreds of kilometers. Not coincidentally, many of these accomplished migrants are important agricultural pests. Among the best studied of these are the brown planthopper (*Nilaparvata lugens)* and the potato leafhopper (*Empoasca fabae).* Neither of these species can normally overwinter in high latitudes. Populations build up in the tropical or subtropical parts of their range and migrate to higher latitudes each spring. They are assisted in their migratory flights by convection and favorable winds, and the initiation of migratory behavior is apparently triggered by favorable atmospheric conditions. Sporadic incidents of very-long-range migrations have also been documented. In one such incident in 1976, swarms of *Balclutha pauxilla* (Cicadellidae), probably originating from a source population in Angola, descended on Ascension Island, 2700 km away in the mid-Atlantic.

Thermoregulation

Most Auchenorrhyncha species appear to regulate their body temperature behaviorally, by seeking out microhabitats in which the ambient temperature remains within a narrow range and moving among alternate microhabitats as conditions change. In some cicadas, physiological mechanisms are also involved. Some species are facultatively endothermic, producing metabolic heat to facilitate calling, courtship, and other activities. This is usually accomplished by vibrating the flight or tymbal muscles until the body temperature rises to an optimal level. Some desert cicadas cool themselves by evaporation of excess water released through pores on the thorax and abdomen. In this way they are able to remain active at ambient temperatures that would kill other insects.

Defense and Escape

Because they are among the most abundant phytophagous insects in many habitats, Auchenorrhyncha are an important food source for numerous vertebrate and invertebrate predators (see next section: Natural Enemies). Species of Auchenorrhyncha exhibit myriad strategies for avoiding predation. These range from relatively simple behaviors, such as dodging around to the opposite side of a leaf or branch as a predator approaches, or hiding under a leaf sheath, to complex mutualistic associations and mimicry. Adults of many species are strong flyers and nearly all (except cicadas) are also excellent jumpers. Juvenile (nymphal) cicadas, spittlebugs, treehoppers, and some planthoppers are incapable of jumping and have adopted other strategies for avoiding predators. All cicada nymphs and many spittlebug and planthopper nymphs are subterranean; thus, their exposure to most predators is minimal. Spittlebug nymphs live within masses of froth and machaerotid nymphs live in calcareous tubes cemented to the host plant. The free-living nymphs of most other auchenorrhynchans appear to rely on cryptic coloration and body forms to escape detection by visual predators such as birds. For example, many treehopper nymphs are strongly flattened with the ventral surfaces of the body concave, enabling them to lie flat against the bark or leaf surfaces of their host plant. Others resemble plant parts such as bud scales or leaflets. Many planthopper nymphs secrete copious quantities of wax [Fig. 4(21)], with which they coat themselves and, often, surrounding parts of their host plants. The wax may prevent parasites and predators from grasping the nymphs, allowing them to leap away. Adults of some species mimic various venomous arthropods such as ants, wasps, robber flies, assassin bugs, and spiders. Some bear horns or spines on the pronotum [Membracidae, Fig. 3(15)] or scutellum [Machaerotidae, Fig. 1(2)] that make them physically difficult for some vertebrate predators to swallow. Many adult cercopids and membracids have conspicuous (aposematic) color patterns, presumably indicating that they are unpalatable. Others have the forewing apices marked with false eyespots, and a few (e.g., Fulgoroidea: Eurybrachidae) have prolongations resembling antennae; the head and thorax of such species often bear transverse lines resembling abdominal segmentation. Adults of various planthopper species mimic lizards, flowers, and lichens. Another strategy involves complex mutualistic associations with ants and other social hymenopterans. Ant mutualism has been documented in numerous lineages of Fulgoroidea and Membracoidea and occurs universally in some groups [e.g., tettigometrid planthoppers Fig. 4(20) and eurymeline leafhoppers]. In such groups, the nymphs usually form aggregations that are tended by ants. The aggressive worker ants drive off predators and receive gifts of honeydew, a sugary excretion, from the nymphs. Ant mutualism may have facilitated the development of subsocial behavior in some groups (see Membracoidea section under Diversity).

Natural Enemies

Auchenorrhyncha are preyed upon by insectivorous vertebrates such as birds and lizards, as well as by invertebrate predators such as spiders, ants, assassin bugs, wasps, and robber flies. Auchenorrhyncha are also attacked by various parasitoids such as dryinid and chalcidoid wasps, epipyropid moths, pipunculid flies, strepsipterans, and nematodes. Because they feed on plant sap, cicadomorphans are not usually susceptible to infection by viral, bacterial, or protozoan pathogens. Thus, entomopathogenic fungi, which do not need to be ingested to infect insects, are the most important pathogens of Auchenorrhyncha.

Economic Importance

Although the vast majority of species of Auchenorrhyncha are benign, the group contains some of the most destructive pests of agriculture. Among the most important are the brown planthopper, sugarcane planthopper (*Perkinsiella saccharicida*), corn planthopper (*Peregrinus maidis*), meadow spittlebug, beet leafhopper (*Neoaliturus tenellus*), potato leafhopper, corn leafhopper (*Dalbulus* spp.), African maize leafhopper (*Cicadulina* spp.), green rice leafhopper (*Nephotettix* spp.), and various grape leafhoppers (*Arboridia* and *Erythroneura* spp.).

Auchenorrhyncha injure plants directly through feeding or oviposition or, more often, indirectly through the transmission of plant pathogens. Economic injury to plants involving cicadas, which occurs rarely, is mainly due to oviposition, although some species occasionally inflict feeding damage (e.g., on sugarcane). Spittlebugs injure plants primarily through feeding and through transmission of xylem-limited bacterial pathogens. Species of Cercopidae are the most significant pests of forage grasses in pastures in Latin America and are also destructive of sugarcane. Interestingly, much if not most of the economic damage done by spittlebugs is due to native spittlebug species colonizing nonnative hosts (e.g., introduced forage grasses, clovers, etc.). Presumably, such plants lack natural resistance to spittlebugs and are more susceptible to injury.

Leafhoppers and planthoppers are among the most significant groups of vectors of plant pathogens, transmitting viruses, bacteria, and mycoplasma-like organisms. Over 150 species are known vectors of economically important plant pathogens. The insects usually acquire the pathogen by feeding on an infected plant, but some pathogens may be transmitted transovarially from mother to offspring. Phloem-limited viral and mycoplasma-like pathogens typically multiply within the vector and enter the plant when the insect injects saliva during feeding. Some xylem-limited bacterial pathogens (e.g., *Xylella*) are apparently unable to travel from the gut to the salivary glands and require regurgitation from the foregut during vector feeding to infect the plant. Annual losses to maize, rice, and sugarcane attributed to pathogens spread by leafhoppers and planthoppers are estimated in the hundreds of millions of dollars. Xylem-feeding cicadelline leafhoppers are also the main vectors of *Xylella fastidiosa,* which causes X diseases of stone fruits (*Prunus* spp.), Pierce's disease of grape, citrus variegated chlorosis, and alfalfa dwarf.

Some Auchenorrhyncha species are considered to be beneficial. Cicadas are used as food by several human cultures. The use of Auchenorrhyncha in biocontrol of weeds has also begun to be explored. For example, a Neotropical treehopper species (*Aconophora compressa)* has been introduced into Australia for control of *Lantana* (Verbenaceae).

Control

Control of auchenorrhynchan pests has traditionally involved the use of conventional contact insecticides, but overuse of chemical insecticides has led to the development of resistance in many pest species and has suppressed populations of their natural enemies. Modern integrated pest management has promoted greater use of resistant plant varieties, cultural control (e.g., removal of litter to reduce numbers of overwintering individuals), and biological control by means of parasitoids and pathogens, as well as more judicious use of pesticides.

CAPTURE AND PRESERVATION

Auchenorrhyncha are most commonly collected by sweeping vegetation with a heavy canvas net. Many species are also attracted to lights. Vacuum collecting is effective for collecting from dense grassy vegetation where many species reside. A gasoline-powered leaf blower fitted with a vacuum attachment can be used to suck the insects from dense vegetation. A fine-mesh insect net bag taped to the end of the intake nozzle will capture the specimens. Other effective collecting methods include Malaise trapping and insecticidal fogging of forest canopy. Auchenorrhycha may be killed in a standard insect killing jar containing potassium cyanide or ethyl acetate, or by freezing.

Specimens for morphological study are usually mounted dry on pins or point mounts. Point mounts should be glued to the right side of the thorax. To identify the species of a specimen, it is often necessary to examine the male genitalia. To do this, the abdomen is removed and soaked in 10% potassium hydroxide solution for several hours (or boiled in the same solution for a few minutes) to clear the pigment. The abdomen is then rinsed in clean water containing a small amount of glacial acetic acid, rinsed again in pure water, and immersed in glycerine. After examination, the cleared abdomen is stored in a glass or plastic microvial pinned beneath the rest of the specimen. Auchenorrhyncha may also be preserved indefinitely in 80 to 95% ethanol, but this causes some green pigments to fade to yellow.

DIVERSITY

Cercopoidea

Cercopoidea (froghoppers and spittlebugs, Fig. 1) are characterized by the following combination of morphological characters: head with frontoclypeus inflated; median ocellus absent; ocelli on crown distant from margin; pronotum extended to scutellar suture; body clothed with fine setae; hind coxae conical, tibia without rows of setae but often with one or more conspicuous spines; male subgenital plate present. The superfamily comprises four families Aphrophoridae, Cercopidae, Clastopteridae, and Macherotidae. The first Cercopoidea (Procercopidae) appear in the fossil record during the Lower Jurassic. These insects retained a median ocellus and apparently lacked the dense setal covering of modern cercopoids. Aphrophoridae and Cercopidae did not appear until the middle Cretaceous; Clastopteridae and Machaerotidae apparently arose during the Tertiary.

Approximately 2500 species and 330 genera of Cercopoidea have been described. The classification has not been revised in over 50 years, and the phylogenetic status of most cercopoid genera and higher taxa remains unknown. Cercopidae [Fig. 1(1)], the largest family, differs from Aphrophoridae [Fig. 1(3)], the next largest, in having the eyes slightly longer than wide and the posterior margin of the pronotum straight (instead of emarginate). The small families Machaerotidae and Clastopteridae differ from other Cercopoidea in having a well-developed appendix (distal membrane) on the forewing. Machaerotidae [Fig. 1(2)] differ from Clastopteridae [Fig. 1(4)] in having two or more r-m crossveins in the forewing and in lacking an outer fork on the radial vein of the hind wing.

Production of "spittle" is a unique characteristic of Cercopoidea [Fig. 1(5)]. Nymphs of Machaerotidae produce the froth during molts, while in other families nymphs live permanently surrounded by the froth. The lateral parts of nymphal abdominal segments are extended ventrally into lobes, which form an open or closed (in machaerotids) ventral cavity, filled with air. The nymphs introduce bubbles of air into their liquid excretion by bellowslike contractions of this device; periodically the tip of the abdomen is extended through the surface of spittle mass to channel air into the cavity. The same air supply is used for breathing via spiracles that open into the ventral cavity. The froth is stabilized by the action of the secretory products manufactured in the highly specialized Malpighian tubules of the nymphs and mixed into the main watery excreta. Wax secreted by plates of epidermal glands on the sixth through eighth abdominal terga (Batelli glands) may also help stabilize the froth.

The function of the spittle mass is not completely understood. It is usually assumed that it protects the insect from predators and desiccation. Cercopoid nymphs are sessile and live within the spittle mass (or, in Machaerotidae, inside fluid-filled tubes). In some species, nymphs tend to aggregate, forming large spittle masses containing hundreds of individual nymphs. Nymphs of Cercopidae apparently feed on roots, whereas aphrophorid and clastopterid nymphs occur on aboveground parts of their host plants. Nymphs of the Machaerotidae live immersed in liquid inside tubes cemented to the twigs of their host plants. The tubes are constructed from calcium carbonate and other salts secreted by the midgut and an organic matrix secreted by the Malpighian tubules. Adult cercopoids do not produce spittle and are free living. They cannot run, and often use only the front and middle legs to walk, dragging the extended hind legs. Consequently, they rely mostly on their strong jumping and flying abilities for movement.

Species of Cercopoidea are often restricted to particular habitats, but many if not most seem to be capable of utilizing a variety of host plants. Many species seem to prefer actinorrhizal and other nitrogen-fixing hosts, presumably because the xylem sap of such plants contains more amino acids and is more nutritious. Cercopoidea is a predominantly tropical group, occurring mostly in wet and mesic habitats. Nevertheless, the genus *Clastoptera,* has radiated extensively in north temperate North America, and *Aphrophora* comprises numerous arboreal species throughout the Holarctic. Cercopidae are primarily grassland insects, feeding on grasses and other herbs. The family Aphrophoridae includes both grass-feeding and arboreal species. Machaerotidae and Clastopteridae are primarily arboreal.

Members of the superfamily Cercopoidea occur worldwide. Cercopidae and Aphrophoridae are pantropical in distribution, with relatively depauperate faunas in the Holarctic. Machaerotidae are restricted to the Oriental and Australian regions. Clastopteridae are mostly New World animals, but one small genus, possibly misclassified, occurs in the oriental region. Most tribes are restricted to either the New or the Old World, and phyletic diversity seems to be highest in the oriental region. A few genera (e.g., *Philaenus* and *Aphrophora*) are widespread, partly as a result of human activities, but most are restricted to a single biogeographic realm.

Cicadoidea

Cicadoidea (cicadas, Fig. 2) are distinguished from other extant Auchenorrhyncha in having fossorial front legs (in nymphs) and three ocelli grouped in a triangle on the crown of the head; in addition they lack the ability to jump. They are conspicuous insects because of their large size (1.5–11 cm) and the loud courtship calls of the males. Most authorities recognize two families: Cicadidae and Tettigarctidae. Tettigarctidae [Cicadoprosbolidae in the paleontological literature, Fig. 2(6)], which differ from Cicadidae in having the pronotum extended to the scutellum and lacking distinct tympana, are a relict group with two extant species in southern Australia and Tasmania and several fossil taxa dating to the Lower Jurassic. Cicadidae [Fig. 2(7–9)], which do not appear in the fossil record until the Paleocene, comprise two main (possibly polyphyletic) groups, those with the tymbals (sound-producing organs) concealed and those with exposed tymbals. These two groups are sometimes given status as separate families, Cicadidae (*sensu stricto)* and Tibicinidae, respectively. Together these groups comprise approximately 1300 extant species. Phylogenetic analyses of the major lineages are in progress and it is likely that the classification of the superfamily will be substantially revised in the near future.

Although cicadas almost always lay eggs on aboveground parts of their host plant, the nymphs drop to the ground soon after hatching and use modified (fossorial) front legs to burrow into the soil, excavating a subterranean feeding chamber adjacent to a root. They feed on the xylem of the roots of perennial plants, coating themselves and lining their burrows with "anal liquid" that appears to be similar to that produced by cercopoid nymphs. Development in most species requires from 2 to 6 years (13 or 17 years in the periodical cicadas of temperate North America). Larger nymphs of some species inhabiting wet habitats construct towers of mud that facilitate aeration of the burrow. Mature nymphs emerge from the ground and climb onto a vertical surface prior to molting into the adult stage [Fig. 2(9)]. As far as is known, all cicadas feed on xylem sap; hence the frontoclypeus is strongly inflated owing to the presence of strong cibarial dilator muscles. Like the Cercopoidea, cicadas do not walk or run well; instead they rely on flight to move over distances greater than a few centimeters. In some cicada species, males are sedentary, often forming aggregations and calling loudly in choruses to attract females. In others, the male calls are less audible, and males fly frequently from place to place in search of females. Male and female cicadas have auditory organs (tympana) at the base of the abdomen. Unlike other Auchenorrhyncha, female cicadas (except *Tettigarcta*) do not produce acoustic signals. Tettigarctidae differ from other cicadas in producing only substrate-borne signals (in males and females).

Cicadoidea are the most ecologically uniform of the Auchenorrhyncha superfamilies. Nymphs of all species are subterranean root feeders, and adults feed on the aboveground parts of their host plants. Most cicada species tend to be associated with particular habitats, and many seem to be host plant specific. Sympatric species often call at different times of day or mature during different seasons, thus temporally partitioning their habitat. The cicada faunas of deserts and savannas are particularly rich in genera and species, but tropical rain forests also harbor a great diversity of species.

Cicadoidea occur worldwide but, like the other two cicadomorphan superfamilies, are largely a tropical group. A few genera (e.g., *Cicada, Cicadetta*), occur on several continents, but most are restricted to a single biogeographic realm. Most species appear to have fairly narrow geographic ranges. The high degree of endemism in many groups has proven useful in studies of biogeography, particularly in the geologically complex island areas of the oriental and Australian regions.

Membracoidea

Membracoidea (leafhoppers and treehoppers, Fig. 3), by far the most speciose of the auchenorrhynchan superfamilies, are characterized morphologically by the narrow costal space of the forewing, the large, transversely articulated metathoracic coxae, the elongate hind femora, the longitudinal rows of enlarged setae on the hind tibiae, and the presence of scutellar apodemes. The superfamily includes Cicadellidae (leafhoppers), a paraphyletic taxon that apparently gave rise to a lineage comprising the three currently recognized families of treehoppers (Melizoderidae, Aetalionidae, and Membracidae). A fifth family, Myerslopiidae, consists of two genera of small, flightless, litter-dwelling insects found only in New Zealand and Chile and thought to represent a distinct, relatively primitive lineage. Together, these groups comprise nearly 25,000 described species, currently grouped into about 3500 genera.

Membracoidea first appeared in the Jurassic, represented by the extinct family Karajassidae. These early membracoids were leafhopperlike insects with inflated faces (indicative of xylem feeding), and they retained a median ocellus and more primitive wing venation (forewing with CuA_1 free distally), but nevertheless had acquired the rows of enlarged setae on the hind tibia characteristic of modern leafhoppers. The first Cicadellidae appeared in the Lower Cretaceous. Treehoppers (Aetalionidae and Membracidae) make their first appearance in Tertiary age Mexican and Dominican amber.

The largest family, Cicadellidae [Fig. 3(10–13)], is characterized by the presence of four rows of enlarged, spinelike setae on the hind tibia, a peg-and-socket joint between the hind coxae, and the production of brochosomes. Membracidae [Fig. 3(15)], the next largest family, differ from Cicadellidae in having three or fewer rows of enlarged setae on the hind tibia, the male genital capsule with a lateral plate, and the pronotum enlarged, usually extended posteriorly over the scutellum and frequently bearing spines, horns, or other ornamentation. Like Membracidae, Aetalionidae [Fig. 3(14)]. have three or fewer setal rows on the hind tibia but differ in having the front femur fused to the trochanter, in having the scutellum completely exposed, and in having digitiform processes on the female genital capsule. Melizoderidae also resemble Membracidae but differ in having parapsidal clefts on the mesonotum. Myerslopiidae, thought to be the most primitive membracoid family, are bizarre, flightless insects with elytra-like forewings, vestigial ocelli, and a triangular mesocoxal meron resembling that of Cercopoidea. The phylogenetic status and relationships among the major lineages are only beginning to be understood.

Cicadellidae are unique among insects in producing brochosomes, which are minute proteinaceous granules synthesized in a specialized segment of the Malpighian tubules. After each molt, leafhoppers spread brochosomes over external surfaces of the body in an act known as anointing. Rows of modified setae on the legs of leafhoppers are used to distribute the brochosomes during anointing and subsequent acts of grooming. The brochosome coating of nymphal and adult leafhoppers makes the integument extremely hydrophobic and protects leafhoppers from becoming entrapped in drops of water and their own often copious excreta.

Ant mutualism and parental care behavior are widespread among treehoppers [Membracidae and Aetalionidae, Fig. 3(15)]. Females of many species guard their eggs [Fig. 3(14)] and sometimes remain with the nymphs throughout their development. In the treehopper tribes Hoplophorionini and Aconophorini, ant mutualism was lost but parental care was retained. In these groups, females are often able to drive off invertebrate predators by buzzing the wings and/or using the hind legs to kick the intruder off the plant. Acoustic alarm signals produced by the nymphs trigger the mother's defensive response. Female *Aconophora* coat the stem of the host plant on either side of their egg masses with a sticky secretion that traps predators and parasitoids.

Most species of Membracoidea seem to have fairly narrow host and habitat requirements, and this has probably contributed to their remarkable diversity. Particularly notable are the large leafhopper faunas of temperate and tropical grasslands, where they are, by far, the most speciose component of the grass-feeding herbivore fauna. Many leafhopper species in deserts and dry grasslands are flightless or only occasionally produce winged individuals. This trait has presumably reduced gene flow among populations and facilitated speciation in some lineages. In temperate forests of the Northern Hemisphere, the leafhopper subfamily Typhlocybinae has diversified extensively through specialization on individual tree genera and species. In tropical forest canopies, the treehopper family Membracidae and the leafhopper subfamilies Idiocerinae and Typhlocybinae are particularly diverse. In Australia, the endemic fauna has radiated extensively on *Eucalyptus.* The North American treehopper tribe Smiliini has radiated extensively on oak (*Quercus* spp.).

Membracoidea are distributed worldwide. Among the five currently recognized families, Cicadellidae and Membracidae occur on all continents except Antarctica. Aetalionidae have a disjunct neotropical/oriental distribution, Melizoderidae are restricted to South America, and Myerslopiidae occur only in New Zealand and Chile. Most species and genera are restricted to a single continent; many tribes and subfamilies are also restricted to particular continents.

Fulgoroidea

Fulgoroidea (planthoppers, Fig. 4) differ from other Auchenorrhyncha in having the frons occupying most of the facial part of the head and usually with distinct longitudinal carinae, tegulae usually present at the base of the forewings, the second segment (pedicel) of the antenna enlarged and (usually) bearing conspicuous placoid sensilla, the forewing anal veins confluent basad of the claval margin, and longitudinal carinae usually present on the head, pronotum, scutellum, and legs. Most have two ocelli dorsolaterally on the head, anterad of the compound eyes, but some Cixiidae also have a medial ocellus on the face. Fulgoroidea first appear in the fossil record in the middle Permian, and Cixiidae appear in the Jurassic. Other modern fulgoroid families apparently arose during the Cretaceous or early Tertiary. Twenty families are currently recognized, comprising approximately 1400 genera and 12,000 species. Fulgoroid families are distinguished from each other based mainly on the shape of the head, the spination of the hind tarsi, and the venation of the forewing. Fulgoroidea are the most morphologically variable of all auchenorrhynchan superfamilies, ranging from 1 mm to over 9 cm in length and exhibiting extensive variation in head shape, wing venation, and genital morphology.

Unlike Cicadomorphans, nymphs of Fulgoroidea apparently do not coat themselves with specialized Malpighian tubule secretions. Instead, they produce wax from specialized glands on the abdominal terga and other parts of the body. The wax forms a hydrophobic coating and may conceal some insects from predators. Adult females of many fulgoroid families also produce wax, with which they coat their eggs [Fig. 4(16)]. In certain tropical fulgoroid species, adults of both sexes produce strands of wax up to 75 cm in length. Aggregation behavior with or without ant mutualism has been documented for nymphs and adults in a few fulgoroid families, but egg guarding is known only in Tettigometridae.

In contrast to the ecologically similar Cicadoidea, the Fulgoroidea are the most ecologically diverse superfamily of Auchenorrhyncha. Nymphs of Derbidae and Achilidae live under bark or in litter, feeding on fungi, while nymphs of Cixiidae, Hypochthonellidae, and Kinnaridae are subterranean root feeders. At least four families include cavernicolous (cave-dwelling) species. Ant mutualism has been documented in several fulgoroid families and seems to occur universally among Tettigometridae [Fig. 4(20)], nymphs of which usually inhabit ant nests. Nymphs of most remaining families and nearly all adults feed on the aboveground parts of vascular plants and most seem to be host specialists. Planthopper species usually feed on woody dicotyledonous plants, but most Delphacidae are grass or sedge specialists. Several species of Delphacidae feed on emergent plants in marshes and are capable of walking on the surface of the water. Delphacidae primarily inhabit temperate and tropical grasslands, and diverse faunas of Issidae, Dictyopharidae (Orgeriinae), and Tettigometridae occur in deserts.

Fulgoroidea occur throughout the temperate and tropical regions of the world but are most diverse in the tropics. The Old World tropics harbor the greatest numbers of described families, genera, and species, but the neotropical fauna is less well studied and may be comparable in diversity. The holarctic fauna is rich in Delphacidae and Issidae, but most other families are poorly represented or absent. Tettigometridae, Ricaniidae, Gengidae, Hypochthonellidae, and Meenoplidae are apparently restricted to the Old World. Some genera, particularly in Cixiidae and Delphacidae, are also cosmopolitan in distribution, but most appear to be restricted to a single biogeographic realm.

See Also the Following Articles

Host Seeking, for Plants • Phytophagous Insects • Phytotoxemia • Plant Diseases and Insects • Prosorrhyncha • Sternorrhyncha

Further Reading

DeLong, D. M. (1971). The bionomics of leafhoppers. *Annu. Rev. Entomol.* **16,** 179–210.

Denno, R. T., and Perfect, T. J. (eds.) (1994). "Planthoppers: Their Ecology and Management." Chapman & Hall, New York.

Hamilton, K. G. A. (1982). "The Insects and Arachnids of Canada," Part 10, "The Spittlebugs of Canada, Homoptera: Cercopidae." Agriculture Canada, Ottawa, Ontario.

Marmarosch, K., and Harris, K. F. (eds.). (1979). "Leafhopper Vectors and Plant Disease Agents." Academic Press, New York.

Metcalf, Z. P. (1960). "A Bibliography of the Cercopoidea (Homoptera: Auchenorhyncha [sic])." Waverly Press, Baltimore.

Metcalf, Z. P. (1960–1962). "General Catalogue of the Homoptera," fascicle VII, "Cercopoidea." North Carolina State College, Raleigh.

Metcalf, Z. P., and Wade, V. (1966). "A Catalogue of the Fossil Homoptera (Homoptera: Auchenorhyncha [sic])." Waverly Press, Baltimore.

Moulds, M. (1990). "Australian Cicadas." New South Wales University Press, Kensington, NSW, Australia.

Nault, L. R., and Rodriguez, J. G. (1985). "The Leafhoppers and Planthoppers." Wiley, New York.

Oman, P. W., Knight, W. J., and Nielson, M. W. (1990). "Leafhoppers (Cicadellidae): A Bibliography, Generic Check-list and Index to the World Literature 1956–1985." CAB International Institute of Entomology, Wallingford, Oxon, U.K.

Ossiannilsson, F. (1949). Insect drummers. *Opusc. Entomol. Suppl.* **10,** 1–145.

Shcherbakov, D. E. (1996). Origin and evolution of Auchenorrhyncha as shown by the fossil record. *In* "Studies on Hemiptera Phylogeny" (C. W. Schaefer, ed.). Entomological Society of America, Lanham, MD.

Sogawa, K. (1982). The rice brown planthopper: Feeding–physiology and host–plant interactions. *Annu. Rev. Entomol.* **27,** 49–73.

Wood, T. K. (1993). Diversity in the New World Membracidae. *Annu. Rev. Entomol.* **38,** 409–435.

Autohemorrhage

Many insects voluntarily discharge blood in response to a threat. This behavior, autohemorrhaging, may serve as a physical deterrent (e.g., by enveloping a potential predator or by exposing the predator to noxious substances). Many beetles in the Chrysomelidae, Meloidae, and Lampyridae are well known for this behavior, which is also called reflex bleeding. For example, the blood of blister (meloid) beetles exposes potential predators to the noxious substance cantharidin. Species apparently release blood through an increase in hydrostatic pressure. When the hydrostatic pressure returns to normal levels, much of this blood is withdrawn into the insect's hemocoel, and so little is actually lost from the insect.

Autotomy

Autotomy is a defensive response to attack involving the amputation or active breaking of a body part along a breakage plane and usually involves loss of a leg. Many invertebrates (e.g., crayfish, daddy-long-legs), including insects such as crickets, grasshoppers, and walkingsticks, and many vertebrates (e.g., salamanders) exhibit this ability. For example, walkingsticks (Phasmida) have weakened areas at the trochanter that break under stress, such as when an appendage is grasped by a predator. If the insect is not an adult, regeneration occurs at the next molt. The amputated leg of the walkingstick twitches after being detached, which may divert the predator's attention away from the attacked insect. A grasshopper, when held by a hind leg, can voluntarily discard that limb by intense muscular contraction and rupture it at the trochanter–femur junction; autotomy can also be induced by mechanical pressure or electrical shock. The individual may benefit from the loss of limb by surviving the potential predator but loss of balance, reduced ability to forage for food, and reduced ability to escape from the next predator result as well.

Sting autotomy, the self-amputation of the stinger and its glands, occurs in many social Hymenoptera as part of colony defense, especially against vertebrates, and may occur because of the size and shape of the sting barbs. Chemical cues released by the detached venom apparatus may enable other attacking individuals to orient themselves to the predator.

See Also the Following Article

Defensive Behavior

Further Reading

Personius, K. E., and Chapman, R. F. (2002). Control of muscle degeneration following autotomy of a hindleg in the grasshopper, *Barytettix humphreysii. J. Insect Physiol.* **48,** 91–102.

B

Bee

see *Apis Species*

Beekeeping

Eva Crane
International Bee Research Association

Beekeeping is the establishment and tending of colonies of social bees of any species, an activity from which the beekeeper obtains a harvest or reward. This reward is usually honey, but it may be some other bee product, or bees themselves (e.g., queens, or colonies for pollination). In beekeeping, each colony is usually in a hive, but some beekeeping is done with honey bees that build their nests in the open. Beekeeping is also done with certain nonsocial bees that are reared for pollinating crops.

TECHNIQUES OF MODERN MOVABLE-FRAME HIVE BEEKEEPING WITH *APIS MELLIFERA*

Most of the world's beekeeping is done with *A. mellifera.* In past centuries, these bees were kept primarily for the production of honey and beeswax. Beekeeping is still done mainly to produce honey, but there are also other specialized types of operation. These include the rearing of queens or package bees for other beekeepers who are producing honey. Another type of beekeeping provides colonies of bees to pollinate crops, since in many areas of large-scale agriculture the native pollinators have been destroyed. Since the 1950s specialized beekeeping has also been developed for the production of royal jelly, pollen, and bee venom.

Each type of beekeeping requires the management of colonies to stimulate the bees to do what the beekeeper wants—for instance, to rear more young house bees to produce royal jelly, or more foragers to pollinate crops. During the 1900s, effective methods were developed for the commercial production of substances other than honey: bee brood, bee venom, beeswax, pollen, propolis, and royal jelly.

Honey Production

A colony of honey-storing bees collects nectar from which it makes honey. Nectar is not available continuously, and to store much honey a colony of bees needs many foraging bees (over, say, 10 days old) whenever a nectar "flow" is available within their flight range. Bees may fly 2 km if necessary, but the greater the distance, the more energy they expend in flight, and the more nectar or honey they consume. Thus, it is often cost-effective for the beekeeper to move hives to several nectar flows in turn during the active season.

Figure 1 shows a movable-frame hive with two "deep" boxes. The hive in Fig. 2 also has two deep boxes for brood (i.e., immature bees, eggs, larvae, and pupae), and a shallow box for honey that is less heavy to lift. Any number of honey boxes (also called supers) may be added to a hive, but these are always separated from the brood boxes by a queen excluder, to keep the honey free from brood. Some empty combs in these supers may stimulate honey storage, but supers are not added far in advance of their likely use by the bees.

It is essential that hives and frames have standard dimensions and that an accessory (spacer) be used to ensure that frames are always exactly the correct distance apart.

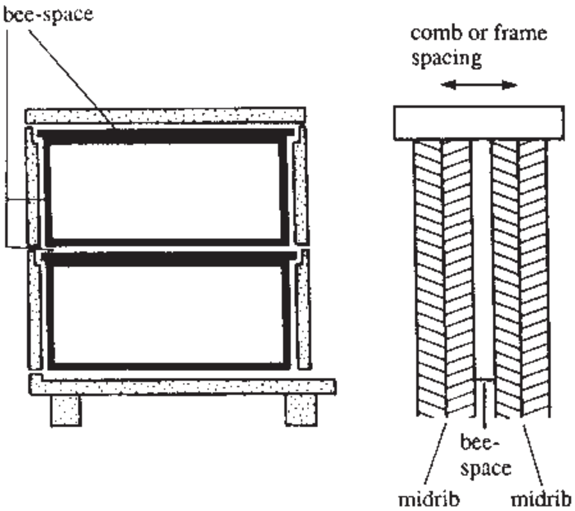

FIGURE 1 Vertical section through a movable-frame hive, showing a brood comb in each box, and the bee spaces. [After Crane, E. (1990). "Bees and Beekeeping: Science, Practice and World Resources." Heinemann, London.]

Queen Production

Large-scale operations are done in five steps, which provide specific conditions for the successive developmental stages of the immature bees that will develop into mated and laying queens.

1. The larvae from which queens will be reared are taken from worker cells of a colony that is headed by a breeder queen selected for chosen genetic characters.

2. Very young larvae are transferred into cell cups mounted mouth down on wooden bars in a "cell-starter colony" that has been queenless for 2 to 4 h. This colony is made up of many (young) nurse bees, and little or no other brood for them to rear. Its bees build the cell cups further and feed the young larvae, and the colony can "start" the rearing of 45 to 90 queen larvae a day.

3. As the larvae grow larger, they receive more food and are better cared for if the number of nurse bees per larva is high. So it is usual to put about 15 cells in each of a number of colonies, where they are separated from the colony's queen by a queen excluder.

4. When the bees have finished feeding the larvae, they seal each immature queen in her cell. The only requirements of an immature queen during the next 7 days are appropriate conditions of temperature and humidity, and these are provided in an incubator. Each queen must emerge from her cell as an adult in a separate cage, for protection from attacks by other queens already emerged.

5. Finally each queen is placed in a "mating hive" containing a few hundred or more workers but no other queen. These hives are taken to a mating apiary, which contains a few strong colonies that include many drones (i.e., males) of a selected strain of honey bees. The apiary is located as far as possible from hives that might contain other drones; a distance of 15 km is likely to be safe, but it varies according to the terrain. When the queen is a few days old, she flies out and mates with drones, and a few days later starts to lay eggs. This shows that she is ready to head a colony.

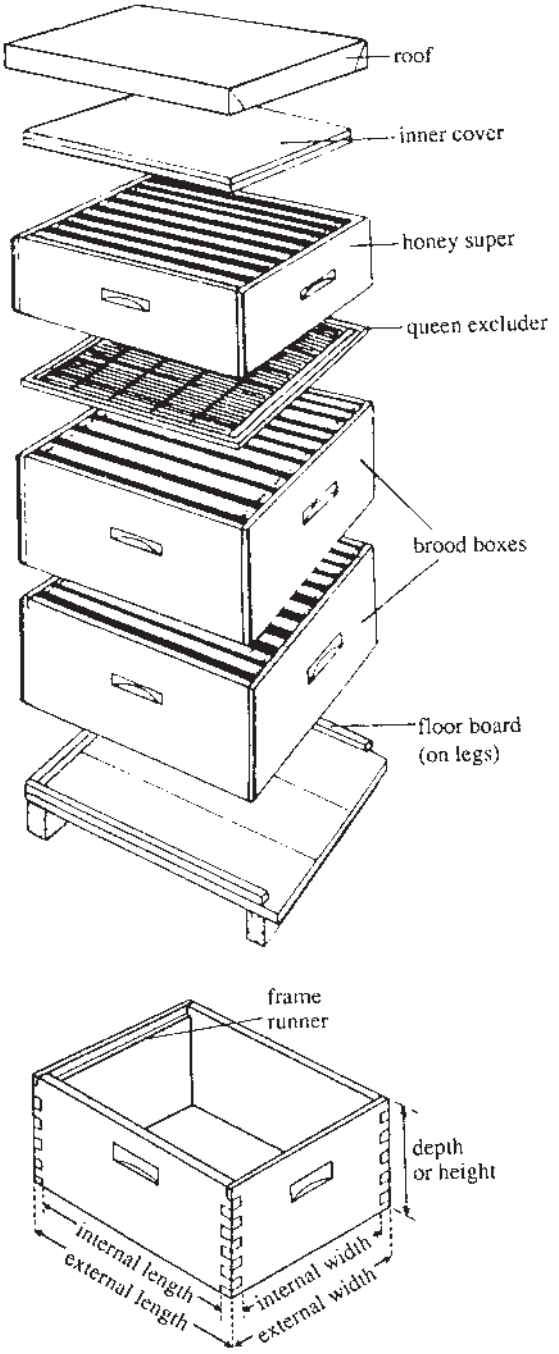

FIGURE 2 Top: exploded view of a movable-frame hive showing the component parts. Bottom: empty hive box showing one of the frame runners. [After Crane, E. (1990). "Bees and Beekeeping: Science, Practice and World Resources." Heinemann, London.]

Package Bee Production

The term “package bees” is used for a number of young worker bees (usually approximately 1 kg) hived with a newly reared and mated queen; these bees together have the potential to develop into a honey-producing colony.

Package bees are normally produced at relatively low latitudes where spring comes early, and are sold at higher latitudes where it is difficult to keep colonies over the winter; many northern beekeepers find it more cost-effective to kill some or all of their colonies when they harvest the season’s honey, and to buy package bees next spring. (If they overwinter no colonies, they can follow another occupation for 6 months or more; at least one beekeeper spends the Canadian winter beekeeping in New Zealand, where it is then summer.) The site where the packages are produced should be earlier weatherwise, by 2 months or more, than the site where the bees are used. A package bee industry is most likely to be viable where a single country stretches over a sufficient north–south distance (at least 1000 km, and up to 2000 or even 2500 km). But in New Zealand, package bees are produced at the end of the bees’ active season and sent by air to Canada, where the season is just starting.

Package bees are prepared as follows. First, all the bees are shaken off the combs of three or four colonies into a specially designed box, taking care that the queens are left behind. The bees are then poured through the “spout” of the box into package boxes, each standing on a weighing machine, until their weight is either 1 or 1.5 kg, as required. Each box is given a young mated queen in a cage, and a can of syrup with feeding holes. (Enough bees are flying around to return to their hives and keep the colonies functional.) For transport, the package boxes are fixed by battens in groups of three or four, slightly separated; they may travel 2400 km, and the truck needs special ventilation. Air transport, though possible, presents various difficulties.

Crop Pollination

Colonies taken to pollinate crops should be strong, with many foraging bees, and also much unsealed brood (to stimulate the bees to forage for pollen), and space for the queen to lay more eggs. Hives should not be taken to the crop before it comes into bloom, or the bees may start foraging on other plants and continue to do so when the crop flowers. If the hives are in a greenhouse, four to eight frames of bees in each may be sufficient, but the beekeeper must check regularly that the bees have enough food; alternatively, each hive may be provided with two flight entrances, one into the greenhouse and one outside. Beekeepers who hire out hives of bees for crop pollination need to have a sound legal contract with the crop grower; they should also be aware of the risks of their bees being poisoned by insecticides.

In addition to honey bees, certain native bees are especially efficient in pollinating one or more crop species, and several species are managed commercially for pollination. The following are quite widely used for the crops indicated: *Andrena* spp. for sarson and berseem in Egypt and India; *Bombus* spp. for tomato and red clover in Finland and Poland; *Megachile* spp. for alfalfa in Chile, India, South Africa, and the United States; *Nomia melanderi* for alfalfa in the United States; *Osmia* spp. for alfalfa in France; and *Xenoglossa* spp. for apple in Japan, Poland, and Spain, also cotton and curcurbits in the United States.

Special Features of Beekeeping in the Subtropics and Tropics

The subtropics (between 23.5° and 34°N, and 23.5° and 34°S) include some of the most valuable world regions for honey production. Like the temperate zones, they have an annual cycle with a distinct seasonal rhythm and a well-marked summer and winter; however, the climate is warmer and the winters are mild, so the bees can fly year-round. All the major honey-exporting countries include a belt within these subtropical latitudes: China, Mexico, Argentina, and Australia.

Between the Tropics of Cancer and Capricorn (23.5°N and S), the situation is different. The seasons (and honey bee colonies) undergo two cycles in the year because the noonday sun is overhead twice a year. So colonies do not generally grow as large as at higher latitudes, nor do they store as much honey. When forage becomes scarce, a colony may cease brood rearing, then fly as a unit to a nearby area where plants are coming into bloom; this flight is referred to as absconding or migration. So one beekeeper may lose colonies, while beekeepers in the other area put out bait hives to receive the swarms.

Beekeeping in the tropics using traditional hives has been well studied, and many development programs have been carried out to introduce more advanced methods. Francis Smith pioneered successful movable-frame hive beekeeping in tropical Africa.

In the tropics, bee diseases are of less importance than at higher latitudes, but bees in torrid zones may be subject to attack by more enemies, certain birds, mammals, and insects. Tropical honey bees therefore defend their nests more vigorously than temperate-zone honey bees. For instance, tropical African honey bees *(A. mellifera)* are easily alerted to sting and, as a result of rapid pheromone communication between individuals, they may attack *en masse.* People in tropical Africa have grown up with the bees and are accustomed to them. But after 1957, when some escaped following introduction to the South American tropics, they spread into areas where the inhabitants had known only the more gentle European bees, and those from tropical Africa were given the name “killer bees.” But once beekeepers in South America had learned how to handle the new bees, they obtained much higher honey yields than from the European bees used earlier.

OTHER ASPECTS OF MODERN HIVE BEEKEEPING

World Spread of *A. mellifera*

In the early 1600s the bees were taken by sailing ship across the Atlantic from England to North America. They would have been in skeps (inverted baskets made of coiled straw), which were then used as hives. The first hives were probably landed in Virginia. The bees flourished and spread by swarming, and other colonies were taken later. By 1800 there were colonies in some 25 of the areas that are now U.S. states, and by 1850 in a further 7. The bees were kept in fixed-comb hives (skeps, logs, boxes).

The bees may possibly have been taken from Spain to Mexico in the late 1500s, but they reached other countries later: e.g., St. Kitts-Nevis in 1720, Canada in 1776, Australia in 1822, and New Zealand and South America in 1839. They were taken later to Hawaii (1857) and Greenland (1950).

In Asian countries where *A. cerana* was used for beekeeping, *A. mellifera* was introduced at the same time as movable-frame hives. Some probable dates of introduction were 1875–1876 in Japan, 1880s in India, 1896 in China, and 1908 in Vietnam.

Between 1850 and 1900 there was widespread activity among beekeepers in testing the suitability of different races of *A. mellifera* for hive beekeeping. The most favored race was Italian *(A. m. ligustica),* named from Liguria on the west coast of Italy, south of Genoa.

Origination and World Spread of Movable-Frame Beekeeping

The production of a movable-frame hive divided the history of hive beekeeping into two distinct phases. This new hive type was invented in 1851 by Reverend Lorenzo Lorraine Langstroth in Philadelphia. He was familiar with the Greek movable-comb hive (discussed later under Traditional Movable-Comb Hive Beekeeping) and with some rectangular hives devised in Europe that contained wooden frames for the bees to build their combs in. These hives, however, had only a very small gap between the frames and the hive walls, and the bees built wax to close it. In 1853 Langstroth described how he had often pondered ways in which he "could get rid of the disagreeable necessity of cutting the attachments of the combs from the walls of the hives." He continued, "The almost self-evident idea of using the same bee-space [as between the centerlines of combs in the frames] in the shallow [honey] chambers came into my mind, and in a moment the suspended movable frames, kept at a suitable distance from each other *and from the case containing them,* came into being" (author's italics). Framed honey combs were harvested from an upper box, and the brood was in the box below. A queen excluder between the boxes prevented the queen from laying eggs in the honey chamber.

The use of hives based on Langstroth's design spread rapidly around the world, dimensions often being somewhat smaller in countries where honey yields were low. Some dates for their first known introduction are 1861, United Kingdom; 1870, Australia; 1878, South Africa; 1880s, India; and 1896, China.

Beekeeping with *A. cerana* in Movable-Frame Hives

Bees of most races of *A. cerana* are smaller than *A. mellifera;* they also build smaller colonies and are less productive for the beekeeper. Unlike *A. mellifera, A. cerana* does not collect or use propolis. *A. cerana* was the only hive bee in Asia until *A. mellifera* was introduced in the late 1800s; it had been kept in traditional hives (logs, boxes, barrels, baskets, pottery) since the first or second century A.D. in China and probably from the 300s B.C. in the upper Indus basin, now in Pakistan.

The movable-frame hives used for *A. cerana* are like a scaled-down version of those for *A. mellifera.* Colony management is similar, except that the beekeeper needs to take steps to minimize absconding by the colonies. In India 30 to 75% of colonies may abscond each year. To prevent this, a colony must always have sufficient stores of both pollen and honey or syrup, and preferably a young queen. Special care is needed to prevent robbing when syrup is fed. Colonies must also be protected against ants and wasps.

The bees at higher latitudes are larger, and in Kashmir (altitude 1500 m, and above) *A. cerana* is almost as large as *A. mellifera* and fairly similar to it in other characteristics; for instance, the colonies do not abscond.

Honey Bee Diseases, Parasites, Predators, and Poisoning

The main brood diseases of *A. mellifera,* with their causative organisms, are American foulbrood (AFB), *Paenibacillus larvae;* European foulbrood (EFB), *Melissococcus pluton;* sacbrood, sacbrood virus (Thai sacbrood virus in *A. cerana*); and chalkbrood, *Ascosphaera apis.* Diseases of adult bees are nosema disease, *Nosema apis;* amoeba disease, *Malpighamoeba mellificae;* and virus diseases. Parasites are tracheal mite, *Acarapis woodi;* varroa mites, *Varroa jacobsoni,* and *V. destructor;* the mite *Tropilaelaps clareae;* bee louse (Diptera); *Braula* spp.; and the small hive beetle, *Aethina tumida.*

Disease or parasitization debilitates the colonies, and diagnosis and treatment require time, skill, and extra expense. Most of the diseases and infestations just listed can be treated if colonies are in movable-frame hives, and in many countries bee disease inspectors provide help and advice. Colonies in fixed-comb hives and wild colonies cannot be inspected in the same way, and they can be a long-term focus of diseases. But by far the most common source of contagion is the transport into an area of bees from elsewhere.

The parasitic *Varroa* mite provides an example. It parasitized *A. cerana* in Asia, where the mite and this bee coexisted. In the Russian Far East, it transferred to introduced *A. mellifera,* whose developmental period is slightly longer, allowing more mites to be reared. Because colonies could then die from the infestation, the effects were disastrous. In the mid-1900s, some infested *A. mellifera* colonies were transported to Moscow; from there, mites were unwittingly sent with bees to other parts of Europe, and they have now reached most countries in the world.

Since the 1950s it has been increasingly easy to move honey bees (queens with attendant workers, and then packages of bees) from one country or continent to another. One result has been that diseases and parasites of the bees have been transmitted to a great many new areas, and to species or races of honey bee that had little or no resistance to them.

The development of large-scale agriculture has involved the use of insecticides, many of which are toxic to bees and can kill those taken to pollinate crops. In California alone, insecticides killed 82,000 colonies in 1962; in 1973 the number was reduced to 36,000, but in 1981 it had risen again, to 56,000. More attention is now paid to the use of practices that protect the bees, including selecting pesticides less toxic to beneficial insects, using pesticides in the forms least toxic to honey bees (e.g., granular instead of dust), spraying at night when bees are not flying, spraying only when the crop is not in flower, and using systemic insecticides and biological pest control. Possible actions by the beekeeper are less satisfactory: moving hives away from areas to be treated, or confining the bees during spraying by placing a protective cover over each hive and keeping it wet to reduce the temperature.

By 1990, legislation designed to protect bees from pesticide injury had been enacted in 38 countries, and a further 7 had established a code of practice or similar recommendations.

TRADITIONAL FIXED-COMB HIVE BEEKEEPING

A. mellifera in the Middle East, Europe, and Africa

Humans have obtained honey and wax from bees' nests in the Middle East, Europe, and Africa since very early times. Beekeeping with *A. mellifera* was probably initiated in an area when the human population increased so much that it needed more honey or wax than was available at existing nest sites, or when some change occurred that reduced the number of nest sites—for instance, when trees were felled to clear land for agriculture.

In the Middle East, population increase was linked with the development of civilizations. The earliest known hive beekeeping was done in ancient Egypt, and similar traditional beekeeping is still carried out in Egypt. In Abu Ghorab, near Cairo, an Old Kingdom bas-relief from around 2400 B.C. shows a kneeling beekeeper working at one end of hives built into a stack; smoke is used to pacify the bees, and honey is being transferred into large storage pots. Over time, the use of horizontal cylindrical hives spread throughout the Mediterranean region and Middle East, and also to tropical Africa, where hollow log hives were often fixed in trees, out of reach of predators.

In the forests of northern Europe, where honey bees nested in tree cavities, early humans obtained honey and wax from the nests. When trees were felled to clear the land, logs containing nests were stood upright on the ground as hives. As a result, later traditional hives in northern Europe were also set upright. In early types such as a log or skep, a swarm of bees built its nest by attaching parallel beeswax combs to the underside of the hive top. If the base of the hive was open as in a skep, the beekeeper harvested honey from it. Otherwise harvesting was done from the top if there was a removable cover, or through a hole previously cut in the side.

Skeps used in northwestern Europe were made small so that colonies in them swarmed early in the active season; each swarm was housed in another skep, and stored some honey. At the end of the season, bees in some skeps were killed with sulfur smoke and all their honey harvested; bees in the other skeps were overwintered, and their honey was left as food during the winter.

A. cerana in Asia

In eastern Asia the cavity-nesting honey bee was *A. cerana,* and it was kept in logs and boxes of various kinds from A.D. 200 or earlier. But farther west in the upper Indus basin horizontal hives rather similar to those of ancient Greece are used, and it has been suggested that hive beekeeping was started in the 300s B.C. by some of the soldiers of the army of Alexander the Great, who settled there after having invaded the area.

Stingless Bees (Meliponinae) in the Tropics

In the Old World tropics, much more honey could be obtained from honey bees than from stingless bees, and the latter were seldom used for beekeeping. But in the Americas, where there were no honey bees, hive beekeeping was developed especially with the stingless bee, *Melipona beecheii,* a fairly large species well suited for the purpose. It builds a horizontal nest with brood in the center and irregular cells at the extremities, where honey and pollen are stored. The Maya people in the Yucatan peninsula in Mexico still do much beekeeping with this bee. The hive is made from a hollowed wooden log, its ends being closed by a wooden or stone disk. To harvest honey, one of the disks is removed to provide access to honey cells; these are broken off with a blunt object, and a basket is placed underneath the opening to strain the honey into a receptacle below. Many similar stone disks from the 300s B.C. and later were excavated from Yucatan and from the island of Cozumel, suggesting that the practice existed in Mexico at least from that time.

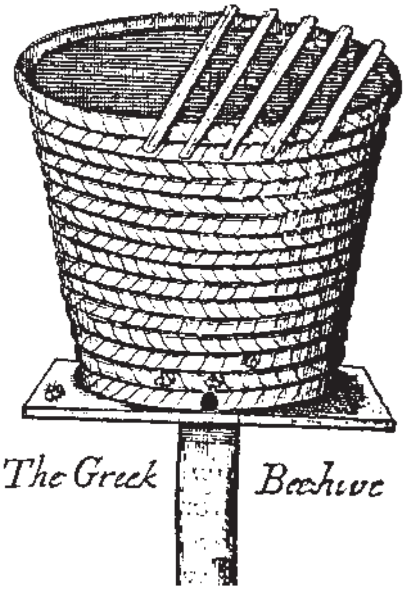

FIGURE 3 Sir George Wheler's drawing of a Greek top-bar hive. [After Wheler, G. (1682). "A Journey into Greece." W. Cademan and others, London.]

Nogueira-Neto in Brazil developed a more rational form of beekeeping with stingless bees. In Australia the native peoples did not do hive beekeeping with stingless bees, but this has recently been started.

TRADITIONAL MOVABLE-COMB HIVE BEEKEEPING

Movable-comb hive beekeeping was a crucial intermediate step between fixed-comb beekeeping, which had been done in many parts of the Old World, and the movable-frame beekeeping used today.

In a book published in 1682 in England, Sir George Wheler recounted his journeys in Greece and provided details of the hives he saw there (Fig. 3). He described the wooden bars shown lying across the top of the hive as "broad, flat sticks" and said that the bees built a comb down from each top-bar, which "may be taken out whole, without the least bruising, and with the greatest ease imaginable." So it was a movable-comb hive. The Greek beekeepers must have placed the bars at the bees' natural spacing of their combs. They made a new colony by putting half the bars and combs from a hive into an empty one; the queen would be in one of the hives, and the bees in the other would rear a new queen.

In the mountain range that separates Vietnam from China, some of the native peoples use a movable-comb hive for *A. cerana;* it is not known how old this